Mathematical Modelling of End-to-End Packet Delay in Multi-hop Wireless Networks and their Applications to QoS Provisioning

A thesis submitted for the degree of Doctor of Philosophy

by

Yu Chen

Communications and Information Systems Research Group

Department of Electronic and Electrical Engineering

University College London

November 2013

Statement of Originality

I, Yu Chen confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Date:

“*Time* is money.” — Benjamin Franklin (1706 – 1790)

“Je n’ai pas *le temps*.” (I do not have *time*.) — Évariste Galois (1811 – 1832)

Abstract

This thesis addresses the mathematical modelling of end-to-end packet delay for Quality of Service (QoS) provisioning in multi-hop wireless networks. The multi-hop wireless technology increases capacity and coverage in a cost-effective way and it has been standardised in the Fourth-Generation (4G) standards.

The effective capacity model approximates end-to-end delay performances, including Complementary Cumulative Density Function (CCDF) of delay, average delay and jitter. This model is first tested using Internet traffic trace from a real gigabit Ethernet gateway.

The effective capacity model is developed based on single-hop and continuous-time communication systems but a multi-hop wireless system is better described to be *multi-hop* and *time-slotted*. The thesis extends the effective capacity model by taking *multi-hop* and *time-slotted* concepts into account, resulting in two new mathematical models: the multi-hop effective capacity model for multi-hop networks and the Mixed Continuous/Discrete-Time (MCDT) effective capacity model for time-slotted networks.

Two scenarios are considered to validate these two effective capacity-based models based on ideal wireless communications (the physical-layer instantaneous transmission rate is the Shannon channel capacity): 1) packets traverse multiple wireless network devices and 2) packets are transmitted to or received from a wireless network device every Transmission Time Interval (TTI). The results from these two scenarios consistently show that the new mathematical models developed in the thesis characterise end-to-end delay performances accurately.

Accurate and efficient estimators for end-to-end packet delay play a key role in QoS provisioning in modern communication systems. The estimators from the new effective capacity-based models are directly tested in two systems, faithfully created using realistic simulation techniques: 1) the IEEE 802.16-2004 networks and 2) wireless tele-ultrasonography medical systems. The results show that the estimation and simulation results are in good agreement in terms of end-to-end delay performances.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Izzat Darwazeh for his first-class supervision, motivation, support and friendship. Izzat is a rigorous professor in research and has a solid background in mathematics and electronic engineering. His instinct to the theoretical problems and suggestion to practical applications inspired me hugely, and pushed and encouraged me to grasp the fundamentals of my research area. Izzat is also a patient supervisor; when I got stuck in some problems during my PhD study, Izzat always asked me to keep calm and talked with me, giving me many useful and inspiring thoughts and suggestions. There is no doubt that without his guidance I would not have finished this thesis.

I would like to thank Dr. Yang Yang for his supervision from September 2008 to February 2010. Yang is a flexible and open-minded supervisor. In my first and second years of PhD study, I was encouraged by Yang to attend several mathematics and physics courses in UCL, which helped me bridge the gap between mathematics and engineering.

I would like to thank Dr. Manoj Thakur for his suggestion of developing a realistic simulation testbed (appeared in Chapter 6), Dr. Ioannis Andreopoulos for his professional explanation of video encoding and decoding techniques (appeared in Chapter 7) and Professor Robert Istepanian and Dr. Nada Y Philip of Kingston University for collaborating in the work described in Chapter 7 and for providing the relevant ultrasound medical video.

My colleagues and friends have made my life at UCL enjoyable and memorable. I would especially thank Ali Anvari, Bing Xia, Bo Tan, Bowen Cao, Dai Jiang, Fang Zhao, Fei Qin, George Matich of SELEX Galileo. George Smart, Guangxiang Yuan, Haixia Chen, Hui Guo, Jianling Chen, Jie Chen, Jie Xiong, Jingjing Huang, John Atkinson, Kevin Chetty, Lei Wang, Lulu Wang, Marcus Perrett, Mingchu Tang, Penglei Li, Ryan Grammenos, Spyridon Papadopoulos, Ting Wang, Tongyang Xu, Weilong Wang, Wuxiong Zhang, Xiaoyu Han, Xuewu Dai, Yemi James, Yi Fang and Yi Zheng

During the course of this work I have been funded by an EU project under ICT, Framework 7, Cooperating Objects Network of Excellence (CONET), Study Assistance Scheme (SAS) and Izzat's funding for the last two years. I want to take this opportunity to thank Prof. Pedro Marron of Universität Duisburg-Essen for leading the CONET project, Dr. Nataliya Popova and Dr. Marco Mendes of Schneider Electric for leading a work package with us at UCL, Prof. and China Pro-Provost Zhengxiao Guo for suggesting the SAS and especially Prof. Izzat Darwazeh for securing funding for my PhD.

Finally, this dissertation is dedicated to my parents and girlfriend for their love, sacrifice and support.

Contents

List of Figures

List of Tables

List of Symbols and Operators

$A'(n;T_s)$	arrival rate at slot n in an MCDT model
$A'(t)$	arrival rate at time t in a continuous-time model
$A(t)$	cumulative arrival in bits over the time interval $[0,t)$ in a continuous-time model
$A(t;T_s)$	cumulative arrival in bits over the time interval $[0,t)$ in an MCDT model
C_n	service time of the n packet
$D(E_n)$	end-to-end delay of the packet arrives at instant E_n
$D(n;T_s)$	delay experienced by the packet arrives at the beginning of slot n in an MCDT model
$D(t)$	delay experienced by the packet arrives at time t in a continuous-time model
$D_n^{(q)}$	queueing delay of the n^{th} packet
$D_\infty/D(\infty)$	end-to-end delay when a system reaches its steady state
D_n	end-to-end delay of the n^{th} packet
$D_{\max}$	maximum delay bound in a continuous-time model
$E[\cdot]$	expectation operator of the event $\{\cdot\}$
E_n	time instant of the n^{th} packet arrival
H	hop number in a multi-hop routing path
$K_{\max}$	maximum delay bound in an MCDT model
L_n	n^{th} packet size in bits

$Q(t)$	queue length at time t in a continuous-time model
$Q(t;T_s)$	queue length at time t in an MCDT model
$S'(n;T_s)$	actual bit service in an MCDT model
$S'(t)$	actual bit service in a continuous-time model
$S(t)$	cumulative actual bit service over the time interval $[0,t)$ in a continuous-time model
$S(t;T_s)$	cumulative actual bit service over the time interval $[0,t)$ in an MCDT model
T_n	n^{th} inter-arrival time
T_s	slot time
V_n	n^{th} packet arrival instant
$\Lambda_X(\theta)$	asymptotic log-moment generation function of a stochastic process $\{X(t), t \geq 0\}$
$\Pr\{\cdot\}$	probability of the event $\{\cdot\}$
$\Psi(t)$	service characterisation curve in a continuous-time model
$\Psi(t;T_s)$	service characterisation curve in an MCDT model
$\alpha^{(c)}(\theta)$	effective capacity function in a continuous-time model
$\alpha^{(c)}(\theta;T_s)$	effective capacity function in an MCDT model
ϵ	arbitrary probability
$\epsilon_{\max}$	predefined probability
$\gamma^{(c)}(\mu)/\gamma$	probability of non-empty buffer in a continuous-time effective capacity model
$\gamma^{(c)}(\mu;T_s)/\gamma_m$	probability of non-empty buffer in an MCDT effective capacity model
γ_{eb}	probability of non-empty buffer in the effective bandwidth-effective capacity model
$\log(x)$	natural logarithm of a number x
$\mathbb{N}$	natural number $\{0,1,2,3,\cdots\}$

$\mathbb{R}$	real number $(-\infty, \infty)$
$\mathbb{R}_0^+$	non-negative real number $[0, \infty)$
$\mathbf{1}\{Q[n] > 0\}$	indicator of buffer emptiness. 1 if the buffer is non-empty and 0 otherwise.
$\mathbf{1}\{d_i \leq x\}$	indicator of delay that is less than or equal to x. 1 if $d_i \leq x$ and 0 otherwise.
μ	constant arrival rate
μ_X	mean of the random variable X
$\sigma^{(c)}$	delay error term of a service characterisation curve in a continuous-time model
$\sigma_m^{(c)}$	delay error term of a service characterisation curve in an MCDT model
σ_X	standard deviation of the random variable X
$\sup_t S(t)$	least upper bound of a set $S(t)$
$\theta^{(c)}(\mu)$	QoS exponent of a connection in a continuous-time effective capacity model
$\theta^{(c)}(\mu; T_s)$	QoS exponent of a connection in an MCDT effective capacity model
θ_{eb}	QoS exponent of a connection in the effective bandwidth-effective capacity model
θ_{lb}	QoS exponent of a connection in the packetised effective capacity model with a leaky bucket
$\tilde{S}(t)$	cumulative bit service over the time interval $[0, t)$ in a continuous-time model
$\tilde{S}(t; T_s)$	cumulative bit service over the time interval $[0, t)$ in an MCDT model
f_m	maximum Doppler rate
$g[n]$	channel gain at the n^{th} slot

k	normalised delay
$p^{(c)}(\mu)$	success probability of a connection in a continuous-time model
$p^{(c)}(\mu; T_s)$	success probability of a connection in an MCDT model
p_a	arrival traffic correlation index
p_d	departure traffic correlation index
$r[n]$	capacity at the n^{th} slot
r_{AWGN}	average channel capacity
y_n	$(n+1)^{\text{st}}$ slot boundary

List of Acronyms

3G	Third-Generation
3GPP	3rd Generation Partnership Project
4CIF	4×CIF
4G	Fourth-Generation
AMC	Adaptive Modulation and Coding
ARQ	Automatic Repeat reQuest
BER	Bit Error Rate
BWA	Broadband Wireless Access
CCDF	Complementary Cumulative Density Function
CDF	Cumulative Density Function
CIF	Common Intermediate Format
DBVP	Delay Bound Violation Probability
DiffServ	Differentiated Services
EB-EC	Effective Bandwidth-Effective Capacity

EDF	Earliest Deadline First
ESS	Extended Service Set
FER	Packet Error Rate
FIFO	First-In First-Out
FPS	Frames per second
FTP	File Transfer Protocol
HDF	Highest DBVP First
HSDPA	High-Speed Downlink Packet Access
HSPA	High Speed Packet Access
IntServ	Integrated Services
IP	Internet Protocol
ISP	Internet Service Provider
LDP	Large Deviation Principle
LoS	Line-of-Sight
LTE	Long Term Evolution
LTE-A	Long Term Evolution-Advanced
M-QoS	Medical Quality of Service
MANET	Mobile Ad hoc Networks
MCDT	Mixed Continuous/Discrete-Time
MMR	Mobile Multi-hop Relay

NLOS	Non-Line-Of-Sight
OFDMA	Orthogonal Frequency-Division Multiple Access
OSI	Open Systems Interconnection
PDF	Probability Density Function
PGF	Probability Generating Function
PMF	Probability Mass Function
PSNR	Peak Signal-to-Noise Ratio
QCI	QoS Class Identifier
QCIF	Quarter CIF
QoS	Quality of Service
RAN	Radio Access Network
SINR	Signal-to-Interference-and-Noise ratio
SNR	Signal-to-Noise Ratio
SSIM	Structural Similarity
TFP	Traffic Forwarding Policy
TTI	Transmission Time Interval
UAS	Uniform Arrival and Service

UMTS	Universal Mobile Telecommunications System
VoIP	Voice over Internet Protocol
WiMAX	Worldwide Interoperability for Microwave Access
WMN	Wireless Mesh Network

Chapter 1

Introduction

Multi-hop wireless technology enables communications between base stations and subscriber stations through one or more relay stations. An example of multi-hop wireless networks is shown in Fig 1.1. In contrast to the deployments of wired backhaul links, the technology provides a cost-effective way for increasing network capacity and coverage. As a result, such a technology is supported by many mainstream wireless standards, such as the Long Term Evolution-Advanced (LTE-A) release 10 [1], the IEEE 802.16j [2] and the IEEE 802.11s [3].

In practise, providing Quality of Service (QoS) guarantees is a fundamental component of broadband wireless networks because network operators or service providers aim to deliver satisfactory Internet service to end users. The *end-to-end packet delay* is a QoS metric and is especially important for delay-sensitive applications, such as Voice over Internet Protocol (VoIP), interactive gaming and video conferencing.

This first chapter of the thesis describes the work motivation (Section 1.1), the main contributions (Section 1.2) and the thesis structure (Section 1.3), and lists author publications derived from research work (Section 1.4).

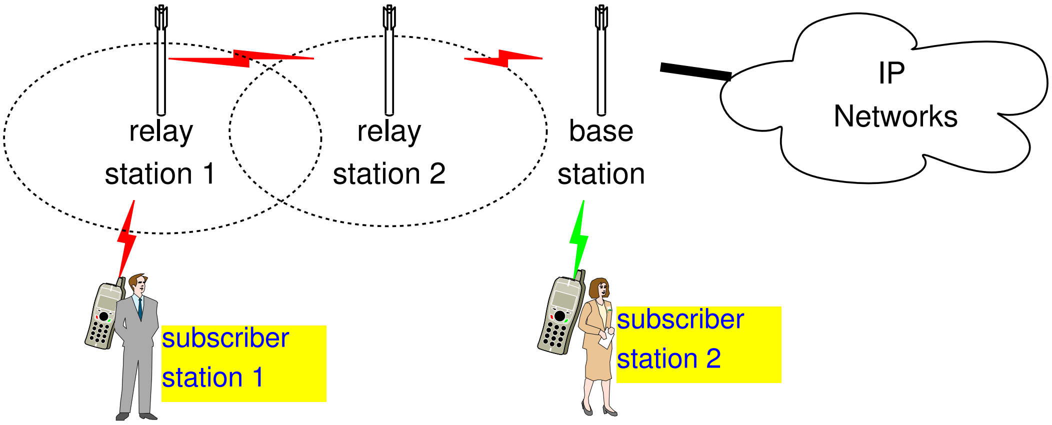


Figure 1.1: An example of multi-hop wireless networks

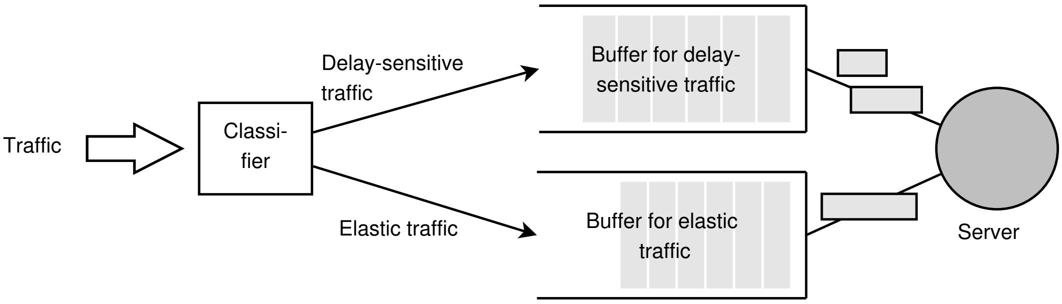


Figure 1.2: An example of traffic classification

1.1 Motivation

Usually, a large number of traffic flows are carried across networks simultaneously. Each flow requires servicing according to specified QoS requirements. Each flow passes several network devices (routers or switches) along its route from the source node to the destination node.

The general idea that serves as a basis for all QoS support methods is as follows: the arrival rate is usually unchangeable; the service rates (provided by the processors and interfaces of network devices) are non-uniform among different traffic classes; it is possible to introduce multiple service classes and to ensure that the service rate of each class is based on its QoS requirements. Fig. 1.2 shows a simplest case that all flows are divided into two classes: 1) delay-sensitive traffic (real-time or synchronous traffic) and 2) elastic traffic that can tolerate significant delays (asynchronous traffic).

The most important factors that directly influence end-to-end delay are the *arrival rate* and the *service rate*. The exact relation between them can be established by the effective capacity model, which is proposed by Wu and Negi [4] in 2003. Unlike the conventional definition of capacity, the effective capacity assumes the existence of buffers, which causes delays, and may be considered as the capacity that is constrained by delay. The effective capacity model translates the effects of a time-varying service rate to end-to-end packet delay distribution or Complementary Cumulative Density Function (CCDF) of packet delay.

The effective capacity model is suitable for *single-hop* wireless systems, the effective capacity model motivates us to investigate the following problems/questions:

- The end-to-end delay of any packet increases after passing a device. What are the exact effects of the number of devices, which a flow has to pass through, on the end-to-end delay distribution of the flow?
- Some communication systems (like most of Third-Generation (3G) and Fourth-Generation (4G) systems) transmit packets every Transmission Time Interval (TTI). Is the effective capacity model still valid to describe delay-constrained behaviours in such systems?

In addressing these problems, research led to a set of contributions listed below and resulted in associated publications listed in Section 1.4.

1.2 Main Contributions

This thesis focuses on the theoretical advancements of the effective capacity model to provide accurate characterisations of delay performances in various system models, which allows accurate and simple design of QoS supported systems. The main contributions of this work are listed below:

- **Validation of the effective capacity model using real Internet data in a wired network** (Chapter 3 based on paper 7 in Section 1.4): a publicly-available Internet traffic trace over gigabit Ethernet is downloaded. The trace contains information on arrival time and size of every packet over the collection period. The Lindley equation [5] is a general form of describing the evolution of delay processes. Hence, in this thesis, the use of the Lindley equation to analyse the trace is reported. The results show that the tail distribution of packet delay is indeed exponentially distributed.

- **Proposal of the multi-hop effective capacity model** (Chapter 4 and paper 2 in Section 1.4): on the basis of the single-hop effective capacity model, the multi-hop effective capacity is developed to characterise multi-hop delay performance. New mathematical formulae are derived, including CCDF of delay, average delay and jitter. The model is tested by computer simulations. The results show that the mathematical formulae and simulation results are in good agreement, thus validate the newly proposed model.

- **Proposal of the Mixed Continuous/Discrete-Time (MCDT) effective capacity model** (Chapter 5 based on papers 8, 5 and 4 in Section 1.4): the conventional effective capacity model was developed based on continuous-time models. For time-slotted communication systems, such as packet-switched networks, a better representation will be using a MCDT description. In this thesis, the conventional effective capacity model are revised, resulting in the new MCDT effective capacity. New mathematical formulae are derived, including of CCDF of delay, average delay and jitter. The CCDF of packet delay is characterised by two functions, namely the probability of non-empty buffer and the success probability of a connection. A set of simulation experiments are established. The simulation results of delay distributions with estimates obtained from the conventional and revised models are compared. The results indicate that the model proposed in thesis always gives better estimation and suggest that the conventional model of [4] is inaccurate for wireless time-slotted communications.

- **Design of a cross-layer multi-hop IEEE 802.16-2004 simulation platform** (Chapter 6 based on papers 6 and 3 in Section 1.4): a new Simulink cross-layer multi-hop simulation platform that integrates network layer, link layer and the IEEE 802.16-2004 physical-layer standard is developed. Simulation results for 1) three-hop scenarios under different traffic load conditions and 2) single-hop scenarios with non-Automatic Repeat reQuest (ARQ) traffic are reported. The simulation results and estimation results from the estimators of Chapters 4 and 5 are in good agreement, which suggests the availability of the estimators for estimating end-to-end delay distributions in real systems.

- **Design of a wireless medical tele-ultrasonography system** (Chapter 7 based on paper 9 in Section 1.4): end-to-end delay is an important consideration in the design of wireless tele-ultrasonography systems because these types of applications are delay-sensitive and have critical delay constraints. A cross-layer simulation platform is built to 1) represent a scenario of remote medical ultrasound video streaming over the IEEE 802.16-2004 networks and 2) transmit a real ultrasound medical video. The simulation results and estimation results from the estimator of Chapter 5 are in good agreement, thus this example verifies the theoretical and modelling development of the thesis by experimentation.

1.3 Organisation of the Thesis

Following this introductory chapter, the remainder of the thesis is organised as follows.

Chapter 2 first provides an overview of three concepts that are related to the thesis areas of investigation title and its research topic: the multi-hop wireless network; packet delay and end-to-end delay modelling. Specifically, section 2.2 discusses the technical background of the multi-hop wireless technology; its implementation types; its advantages and disadvantages and standards that use this technology. Section 2.3 explains the definition of a packet delay and its statistical characterisations. Section

2.4 introduces various modelling techniques of end-to-end delay, including the network calculus theory, the effective capacity model and the Lindley equations.

Chapter 3 analyses a real Internet traffic trace over gigabit Ethernet to validate the effective capacity model. An Ethernet is modelled as a G/G/1 queueing network in Section 3.2. Section 3.3 explains 1) the use of the Lindley equation and the traffic trace for deriving end-to-end packet delay; 2) the basis analysis of the trace, including the traffic load and packet size distributions and 3) the use of regression method to fit the CCDFs of end-to-end packet delay. The empirical and fitted CCDFs of end-to-end packet delay from the trace are plotted in Section 3.4.

In Chapter 4, a multi-hop wireless network is modelled as a queueing network in Section 4.2. The multi-hop effective capacity model is formally proposed in Section 4.3. A simulation platform and an estimation algorithm based on the model (multi-hop effective capacity-based model) are developed in Section 4.4. Simulation results and estimation results are then compared and discussed in Section 4.5.

In Chapter 5, a general wireless time-slotted communication system is modelled as a MCDT queueing model. In Section 5.3, the conventional effective capacity model is adapted to the MCDT effective capacity model. A simulation platform and an estimation algorithm based on the model (MCDT effective capacity-based estimator) are developed in Section 5.4. Simulation results and estimation results are then compared and discussed in Section 5.5.

Chapter 6 presents an application: estimation of the end-to-end delay performances in Worldwide Interoperability for Microwave Access (WiMAX) systems. Section 6.2 explains the building blocks of the simulation platform that is based on the IEEE 802.16-2004 standard. The accuracy of the multi-hop effective capacity-based estimator and the MCDT effective capacity-based estimator is tested in this platform. Simulation results and estimation results are then compared and discussed in Section 6.3.

Chapter 7 tests presents another application: estimation of the end-to-end frame delay in wireless tele-ultrasonography medical system. Section 7.2 introduces a wireless tele-ultrasonography medical system and its system model. In Section 7.3, the details of the simulation platform are described. Results are illustrated and discussed with respect to Medical Quality of Service (M-QoS) metrics in Section 7.4.

In Chapter 8, the thesis is summarised in Section 8.1 and future research directions are suggested in Section 8.2.

1.4 List of Publications

The contributions presented in Chapter 1.2 have led to two journal papers (one under review) and seven conference papers. All publications are listed in chronological order:

1. **Yu Chen**, Jia Chen and Yang Yang, "Multi-hop Delay Performance in Wireless Mesh Networks", MONET 2008, Vol. 13 (published before the PhD started)

2. **Yu Chen**, Yang Yang and Izzat Darwazeh, "A Cross-Layer Analytical Model of End-to-end Delay Performance for Wireless multi-hop Environments", IEEE GLOBECOM 2010, Miami, USA, December 2010

3. **Yu Chen** and Izzat Darwazeh, "End-to-end Delay Performance Analysis in IEEE 802.16j Mobile multi-hop Relay (MMR) Networks", IEEE ICT 2011, Ayia Napa, Cyprus, May 2011

4. **Yu Chen** and Izzat Darwazeh, "Effective Capacity (EC) Model in Fixed-length Packet-switching Systems", LCS, September 2011

5. **Yu Chen** and Izzat Darwazeh, "Poster Abstract: Effective Capacity Model in the Discrete Time Domain", EWSN 2012, Trento, Italy, Feb 2012

6. **Yu Chen** and Izzat Darwazeh, "An Estimator for Delay Distributions in Packet-based Wireless Digital Communication Systems", IEEE WCNC 2013, Shanghai, China, April 2013

7. **Yu Chen** and Izzat Darwazeh, "Quality of Service (QoS) Analysis of an Internet Traffic Trace Over gigabit Ethernet", IEEE ICT 2013, Casablanca, Morocco, May 2013

8. **Yu Chen** and Izzat Darwazeh, "Mixed Continuous/Discrete Time Effective Capacity Model for Wireless Slotted Communication Systems", the IEEE transaction on Wireless Communications (submitted March 2013)

9. **Yu Chen**, Nada Philip, Robert Istepanian and Izzat Darwazeh, "End-to-end Delay Distributions in Wireless Tele-ultrasonography Medical Systems", IEEE GLOBECOM 2013, Atlanta, USA, December 2013 (accepted)

Chapter 2

Multi-hop Wireless Networks, End-to-End Packet Delay and Modelling of Packet Delay

2.1 Introduction

Congestion and queues are generic features of packet-switched networks. As shown in Fig. 2.1, the packet-switching principle assumes the presence of buffers at the input/output interface of each packet switch/router. Packet buffering at times of network congestion is the main mechanism for supporting and smoothing bursty data traffic, ensuring better throughput for these types of networks when compared to circuit-switched networks [6]. This feature was the main reason for the popularity of packet-switched networks in general and Internet Protocol (IP) networks in particular for connecting computers.

The drawback of using buffers is that they also introduce unpredictable and variable delays when packets traverse across networks; the main source of problems for delay-sensitive traffic. Before the late 1990s, data networks mainly transferred "elastic" (non time-sensitive) traffic in traditional applications such as File Transfer Protocol

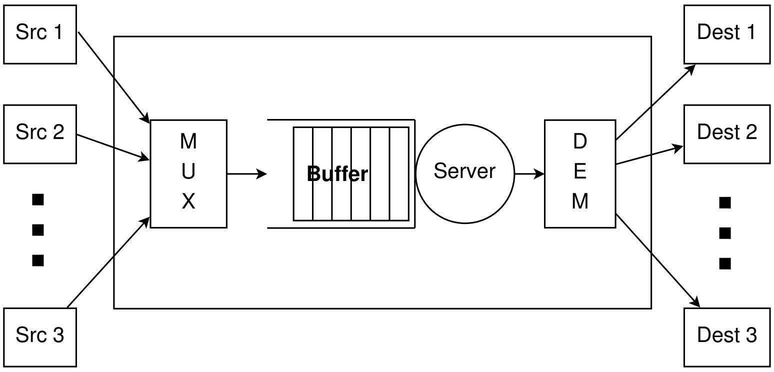


Figure 2.1: Buffer inside a switch or router

(FTP), e-mail and web browsing. The rapid growth of popularity in time-sensitive applications, especially Voice over Internet Protocol (VoIP), interactive gaming and video conferencing, stimulated the development and deployment of tools that can to some extent compensate for such negative effects of queueing.

Quality of Service (QoS) metrics reflect negative effects of congestion in packet-switched networks and are usually related to three network performance metrics [6, 7]:

1. **Delay:** Packets are delivered to destination nodes with delays, which may vary from packet to packet (one measure is *jitter*). As suggested by the thesis title, the random behaviour of end-to-end packet delay is the focus of this work;

2. **Throughput (or delivered data rate):** Different traffic types require different data rates. Data rate is measured on some time interval as a result of dividing the volume of successfully transmitted data by the interval duration.

3. **Loss:** Packets fail to reach their destinations. Losses can occur in the physical layer, link layer and/or network layer [8].

This chapter first provides an overview of the multi-hop wireless technology in Section 2.2. The concept of a packet delay and its statistical characterisations are discussed in the next Section 2.3. The basic techniques of end-to-end delay modelling are presented in Sections 2.4. Literature review for specific topics is further introduced at the beginning of each chapter from Chapter 3 to Chapter 7.

2.2 Multi-hop Wireless Networks

The main distinguishing characteristic of a multi-hop wireless network is its capability of self-organisation and multi-hop communications [9]; these introduce *hierarchical* network architectures [10]. Multi-hop wireless networks have two categories:

1. Mobile Multi-hop Relay (MMR) networks and

2. Wireless Mesh Networks (WMNs).

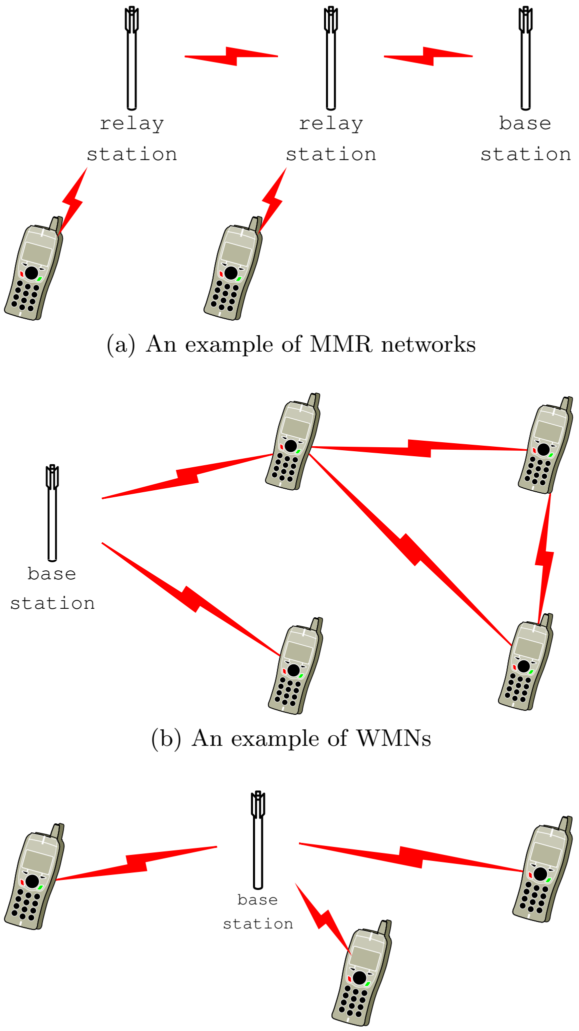
Figs. 2.2a and 2.2b show two examples of MMR networks and WMNs.

In MMR networks, the network architecture consists of base stations, relay stations and subscriber stations. A relay station is not directly connected to wired infrastructure and has the minimum functionality necessary to support multi-hop communications. The important aspect is that subscriber station to subscriber station communication paths have to include a base station or a relay station.

In WMNs, traffic can be routed through other subscriber stations and can also occur directly between them. Nodes are comprised of mesh routers and mesh clients and thus routing process is controlled not only by base stations or relay stations but also by subscriber stations. Each node can forward packets on behalf of other nodes that may not be within direct wireless transmission range of their destination. A system that has a direct connection to backhaul services outside the mesh network is termed a mesh base station. All the other systems are called mesh subscriber stations. Commercial deployment scenarios of the wireless mesh technology were studied in [10].

2.2.1 Advantages and Challenges of the Multi-hop Wireless Technology

Compared to centralised networks that only support single-hop communications (Fig. 2.2c shows an example of centralised networks), the multi-hop wireless technology brings many benefits:



(a) An example of MMR networks

(b) An example of WMNs

(c) An example of centralised networks

Figure 2.2: Examples of different network architectures

- *Flexibility in routing traffic:* The capability of choosing the right routing path would greatly save bandwidth resources by localising the traffic (similar to the Internet routing protocols).

- *Capacity enhancement:* The cell edge area and dead zone area (caused by the radio blockage) usually have lower Signal-to-Interference-and-Noise ratio (SINR) values than other areas inside the cell. Therefore, deploying relay nodes or mesh routers in such area overcomes this problem and improve the system performance.

- *Further capacity gain:* The multi-hop architecture introduces the space diversity so it enables cooperative multiple-input multiple-output [11] technology. Other advanced techniques can be incorporated into multi-hop architectures to further enhance capacity, such as cognitive radio [12] and adaptive precoding [13].
- *Scalability:* Nodes act as repeaters to transmit data from nearby nodes to peers that are too far away to reach [14], resulting in a network that can span large distances, especially over rough or difficult terrain.
- *Reliability:* Resilience to unit failure with the provision of backup paths.
- *Cost-saving:* Deploying relay stations or mesh routers is more economically effective than deploying base stations, making the multi-hop wireless technology an ideal solution for network operators (especially for those cash-strapped Internet Service Providers (ISPs) and carriers) to roll out robust and reliable wireless broadband service access with acceptable up-front investment [14].
- *Energy efficiency:* The multi-hop architecture can significantly reduce the energy consumption in situations where users want to make a lot of high data-rate connections. When there is no traffic, relay stations or mesh routers may further save energy by switching to the low-energy stand-by mode [15].
- *Potential for very high spectral efficiencies:* In physical mesh realisation, directional antennas and smart antennas could be used to improve the transmission in multi-hop wireless networks [16, 17].

However, the multi-hop wireless technology also introduces some challenges:

- *QoS guarantees:* The end-to-end delay is associated with channel conditions and number of hops, both of which are difficult to be guaranteed.
- *Security in WMNs:* Until now, there has been no centralised trusted authority to distribute a public key in WMNs due to its distributed system architecture

[18]. Without a convincing security solution, the wireless mesh technology may not be able to succeed because customers may lack the incentive to subscribe to unreliable services.

- *Timing synchronisation in time division multiple access (TDMA) Systems [19]:* TDMA systems have a high requirement for the time synchronisation, but in a distributed multi-hop network, accurate synchronisation within the global network is difficult to achieve.

- *Support for routing functionality and propriety signalling*: The unique characteristics of multi-hop wireless networks suggest that they demand a specific solution to the support for routing functionality and propriety signalling, which is still an open question [20].

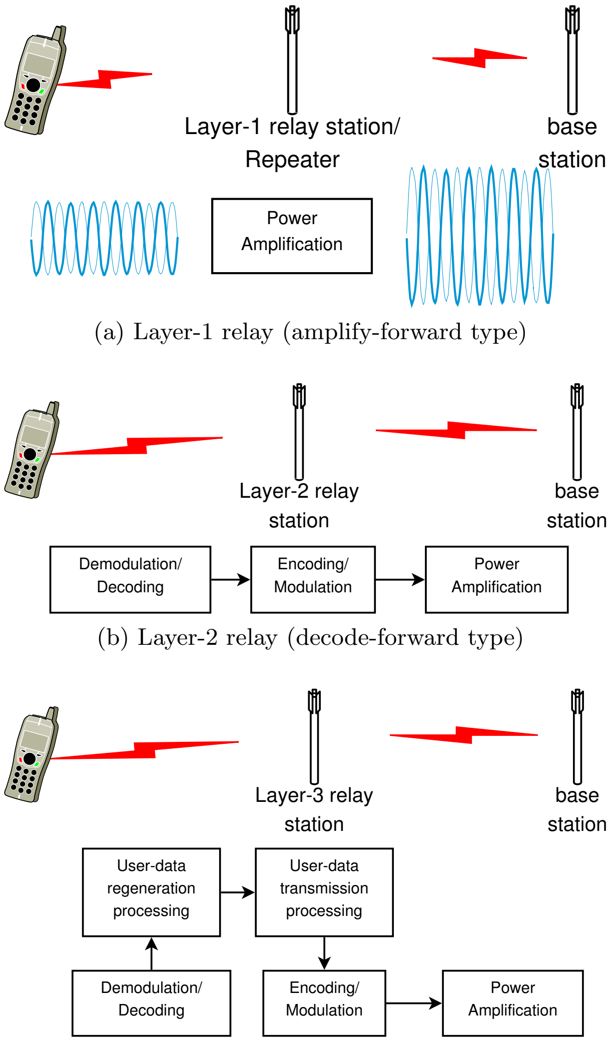
2.2.2 Implementation Types of the Multi-hop Wireless Technology

On the basis of the Open Systems Interconnection (OSI) 7-layer model, there are three types of the multi-hop wireless technology [21]:

1. The layer-1 relay (the relay station is called a booster or repeater): As shown in Fig. 2.3a, it is an amplify-forward type of relay technology. radio frequency signals received on the downlink from the base station are amplified and transmitted to the mobile station. In a similar manner, radio frequency signals received on the uplink from the mobile station are amplified and transmitted to the base station.

2. The layer-2 relay: As shown in Fig. 2.3b, it is a decode-forward type of relay technology. radio frequency signals received on the downlink from the base station are demodulated and decoded and then encoded and modulated again before being sent on to the mobile station.

3. The layer-3 relay: As shown in Fig. 2.3c, it also performs demodulation and decoding of radio frequency signals received on the downlink from the base sta-



(a) Layer-1 relay (amplify-forward type)

(b) Layer-2 relay (decode-forward type)

(c) Layer-3 relay

Figure 2.3: Types of relay networks

tion, but then goes on to perform processing (such as ciphering and user-data concatenation/segmentation/reassembly) for retransmitting user data on a radio interface and finally performs encoding/modulation and transmission to the mobile station.

The layer-1 relay is the simplest type of relay so it makes for low-cost implementation and short processing delays. However, the layer-1 relay amplifies inter-cell interference and noise together with desired signal components thereby deteriorating the received SINR and reducing the throughput-enhancement gain [21].

The layer-2 and layer-3 relays perform the demodulation and decoding processing at relay stations, which overcomes the drawback in layer-1 relay of deteriorating received SINR caused by amplification of inter-cell interference and noise. Therefore, these two types can achieve a better throughput-enhancement. On the other hand, these two types require more processing resources, including modulation/demodulation, encoding/decoding processing, and radio-control functions (like mobility control, retransmission control by Automatic Repeat reQuest (ARQ), and user-data concatenation/segmentation/reassembly), all of which increase manufacture cost and processing delay. In contrast to the layer-2 relay, the layer-3 relay is more flexible because it processes packets in layer 3 so that it is capable of performing the traffic routing function. However, such a mechanism will further increase processing delay.

2.2.3 Multi-hop Wireless Technology in Standards

In this section, mainstream wireless standards that uses the multi-hop wireless technology are briefly introduced; standards include Long Term Evolution-Advanced (LTE-A), the IEEE 802.16j and the IEEE 802.11s.

2.2.3.1 LTE-A

The LTE-A is standardised by the 3rd Generation Partnership Project (3GPP) Release 10 [1]. The LTE-A is a software upgrade for the Long Term Evolution (LTE) networks; and this standard requires peak download rates over 1gigabit, which fully supports the 4G requirements based on the International Telecommunication Union – Radiocommunication Sector (ITU-R). There are several key improvements and additions of new features to the LTE, among which is the MMR technology.

A number of potential deployment scenarios are of interest to major operators [22]. Although not all of them were prioritised for Release 10 relay specification, the discussion is helpful for scenario identification of future relay and related technologies. The potential usages of the MMR technology are as follows [21]:

1. Extending the coverage area to mountainous and sparsely populated regions (rural area and wireless backhaul scenarios): For extending coverage to such areas, the MMR technology is a more economical solution than deploying fixed-line backhaul links. The MMR technology should also be effective for providing temporary coverage when earthquakes or other disasters strike or when major events are being held (emergency or temporary coverage scenario), i.e., for situations in which the deployment of dedicated fixed-line backhaul links is difficult.

2. Urban scenarios: In urban areas, the MMR technology can be an effective solution when the installation of utility poles or laying of cables inside buildings become difficult in some countries and regions. Furthermore, pico base stations and femtocells can be used for urban hot spots, dead spots, and indoor hot spots.

3. Group mobility scenario: relay stations can be installed on vehicles like trains or buses to reduce the volume of control signals from moving mobile stations.

2.2.3.2 IEEE 802.16j

The IEEE 802.16 is a series of wireless broadband standards with several releases of these standards to support new technologies. The commercialised brand name of the IEEE 802.16 standard is Worldwide Interoperability for Microwave Access (WiMAX) [23]. The WiMAX technology offers high data rates over a relatively large coverage area and its recent developments from both academia and industry are surveyed in [24].

As one specific standard of the IEEE 802.16 standard family, the IEEE 802.16j was released in 2009 to support the MMR operation in answer to the current deployment issues:

1. Coverage limitations or low SINR at cell edge caused by significant signal attenuation at high spectrum

2. Poor signal reception due to shadowing or even coverage holes

3. Limited spectrum

The standard allows for the MMR architecture to work with traditional cellular architectures and to offer substantial reduction in unit price and better spectral utilisation, thereby providing competitive system performance [25]. The basic idea is that the WiMAX Mobile Devices in unfavourable locations can communicate with base stations through the intermediate relay stations at high data rates. relay stations may be fixed, nomadic or mobile. The main targeted usage scenarios of the MMR technology are fixed infrastructure, in-building coverage, temporary coverage and coverage on mobile vehicle usage [26].

2.2.3.3 IEEE 802.11s

The IEEE 802.11s is within the standard family of the IEEE 802.11 (sometimes known as Wi-Fi). The IEEE 802.11s has the protocols that build interoperable wireless links and multi-hop paths between multiple access points and is still in the revision stage on July 2013.

The basic ideas of the IEEE 802.11s is that 1) it uses 802.11-based physical-layer devices and link-layer protocols for providing the functionality of an Extended Service Set (ESS) mesh network, and 2) it includes some amendments for mesh networking, defining how wireless devices can interconnect to create an ad-hoc network. An example of the architecture of the IEEE 802.11s wireless mesh networks is shown in Fig. 2.4. As seen from the figure, the IEEE 802.11s contains three types of entities [27]:

1. **mesh point:** a mesh point supports the mesh services (i.e., the mesh formation as well as the operation of the mesh network, including the path selection and frame forwarding);

2. **mesh access point:** a mesh access point is an access point with the mesh point functionality, thus providing both mesh services and access point services;

3. **mesh portal point:** a mesh portal point is a portal with the mesh point functionality, thus interfacing the mesh network to other external networks. Although

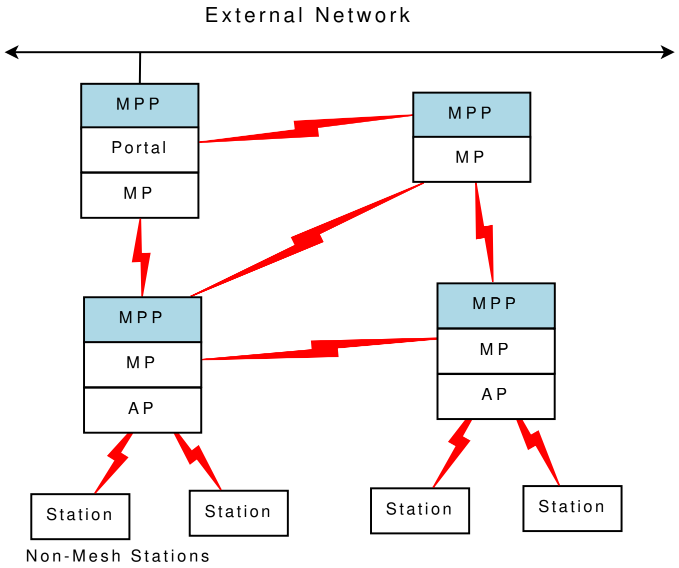


Figure 2.4: An IEEE 802.11s wireless mesh network [20]

there is only one mesh portal point in the figure, the architecture should allow the existence of multiple mesh portal points.

Mesh points, mesh access points and mesh portal points are interconnected via peer-to-peer mesh links, while each station and mesh access point pair are connected via downlink/uplink. The IEEE 802.11s does not change any behaviour of non-mesh stations so mesh access points look like normal access points to non-mesh stations. The mesh portal point provides a MAC bridging functionality between a mesh network and non-802.11 external networks.

2.3 End-to-End Packet Delay

End-to-end packet delay is the time taken for a packet to be transmitted across a network from source to destination; its formal definition is given below:

End-to-end packet delay Denote by E_n the time instant of the n^{th} packet arrival (the *last bit* or the *most significant bit* of the packet has been received) at a network and denote by V_n the time instant of the n^{th} packet departure (the *last bit* or the *most*

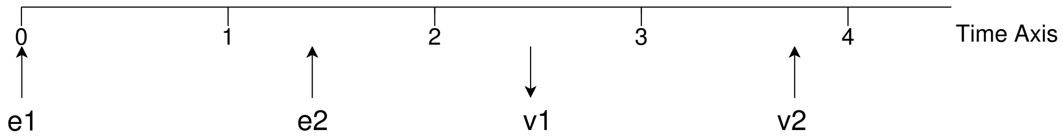


Figure 2.5: An example of packet arrivals and departures

significant bit of the packet has been transmitted) from the network. The end-to-end delay of the n^{th} packet D_n (or the delay of the packet arrives at instant E_n, $D(E_n)$) can be mathematically calculated from

$$D_n = D(E_n) = V_n - E_n \tag{2.1}$$

An example of packet arrivals and departures is shown in Fig. 2.5. It worth noting that

- The notation of $D(E_n)$ is commonly used in fluid traffic models (see the next Section 2.4.1);
- D_∞ or $D(\infty)$ denotes the end-to-end delay when a system reaches its steady state;
- Packet is an object concept. It can refer to a IP packet, a video frame (see Chapter 7) and so on.

In most communication systems, end-to-end packet delay are random variables because packet arrivals and departures are stochastic processes.

An *end-to-end delay process* $\{D_n, n \geq 1\}$ is an ordered sequence of packet end-to-end delays from the 1^{st} packet; and it is a stochastic process. Consider a *realisation* $\{d_1, d_2, d_3, \cdots, d_N\}$ of an end-to-end delay process with N samples.

A realisation is a single instance of a stochastic process. Any arguments that apply to a sample path also apply to the process as a whole [28].

through out this thesis, if a set of capital letters with indexes refers to a stochastic process, the set of the same small letters with indexes refers to a realisation of this stochastic process.

Fig. 2.6a shows the end-to-end delays of the first 100 packets from an Internet traffic trace (this topic will be discussed in more detail in Chapter 3).

The basic probability concepts considered in this thesis are:

1. **Probability Density Function (PDF) of end-to-end delay** The PDF of the random variable D_∞ , denoted by $f_{D_\infty}(x)$ is the function that gives the likelihood of D_∞ taking the value x, for any real number x:

$$f_{D_\infty}(x) = \frac{d}{dx}\Pr\{D_\infty \leq x\} \tag{2.2}$$

2. **Complementary Cumulative Density Function (CCDF) of end-to-end delay** The CCDF of a random variable D_∞ is the function that gives the probability of D_∞ being greater than a real number x. By definition, its value can be computed as the integral of the PDF from x to ∞:

$$\Pr\{D_\infty > x\} = \int_x^\infty f_{D_\infty}(t)dt \tag{2.3}$$

3. **Empirical CCDF of end-to-end delay** Let $\{d_1, d_2, d_3, \cdots, d_N\}$ be a realisation of an end-to-end delay process with N samples. The empirical CCDF of end-to-end delay is a function of x, which equals the decimal fraction of the observations that are greater than x:

$$\hat{\Pr}\{D_\infty > x\} = \frac{\text{number of elements in the realisation} \leq x}{N} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{d_i \leq x\} \tag{2.4}$$

where $\mathbf{1}\{d_i \leq x\}$ is the indicator of event $\{d_i \leq x\}$, defined as

$$\mathbf{1}\{d_i \leq x\} = \begin{cases} 1 & \text{if } d_i \leq x \\ 0 & \text{if } d_i > x \end{cases} \tag{2.5}$$

Fig. 2.6b shows a CCDF of the first 100 packets.

4. **End-to-end delay distribution** In this thesis, it is the same concept as the CCDF of end-to-end delay.
5. **Stationary delay process** The stationarity indicates that the packet delay statistics are not time dependent, i.e., every packet delay has the identical distribution;
6. **Ergodic delay process** The ergodicity indicates that the sample mean of a realisation converges to the expectation of any packet delay.

Furthermore, the *descriptive statistics* of end-to-end delays are more frequently used in practise [6] and it includes

1. **Average end-to-end delay**, $\hat{\mu}_D$: the value is expressed as the sum of all delays $\{d_1, d_2, d_3, \cdots, d_N\}$, divided by the total number of all measurements (N):

$$\hat{\mu}_D = \frac{\sum_{i=1}^{N} d_i}{N} \tag{2.6}$$

In this thesis, estimates are typically written by adding a circumflex over the symbols.

2. **Jitter**, σ_D: the value represents the average deviation of delays from the average delay:

$$\hat{\sigma}_D = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \hat{\mu}_D)^2} \tag{2.7}$$

3. **Maximum delay bound** $D_{\max}$: it is the value that packet delays must not exceed with a predefined probability, $\epsilon_{\max}$:

$$\Pr\{D_\infty > D_{\max}\} \leq \epsilon_{\max} \tag{2.8}$$

To elaborate on the third item of maximum delay bound, suppose the maximum delay and the predefined probability are specified to be 100 ms and 0.95, respectively. To obtain an assessment that serves as evidence of the quality of the network operation, it is sensible to use the concept of the empirical CCDF of end-to-end packet delay (2.4) and introduce the following requirement:

$$\hat{\Pr}\{D_\infty > 100 \text{ ms}\} \leq 0.95 \tag{2.9}$$

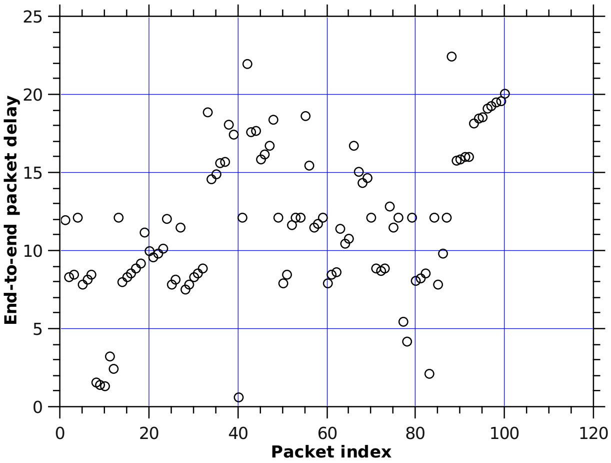
Depending on the application type, it is possible to use a specific set of descriptive statistics. For example, consider music broadcasting over the Internet. Since this service is not interactive, it tolerates significant delays for individual packets, sometimes of several minutes. However, delay variation for music broadcasting must not exceed 100 to 150 ms; otherwise, the playback quality is significantly degraded [6]. As a result, the QoS requirements for this case must include limitations on the average delay variation.

2.3.1 QoS Requirements of Packet Delay in Standards

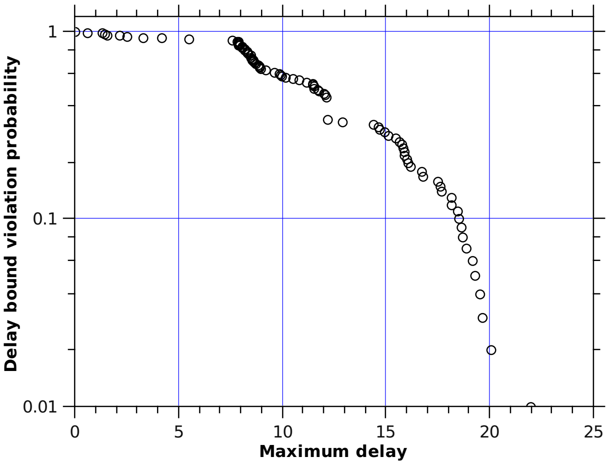
In this section, the QoS requirements of packet delay in the LTE, the IEEE 802.16 and the IEEE 802.11 are briefly introduced.

2.3.1.1 LTE

The LTE standard provides QoS provisioning from the ground up. The concept of Traffic Forwarding Policy (TFP) denotes a set of pre-configured traffic handling attributes within a particular user plane network element, for example, a Radio Access Network (RAN)-TFP includes several attributes such as a link layer protocol model (acknowl-



(a) Packet queueing delay of the first 100 packets from an Internet traffic trace



(b) Empirical CCDF of the first 100 packets

Figure 2.6: Packet delays and empirical CCDF of packet delay

edged or unacknowledged), a power setting and a default uplink maximum bit rate; a gateway-TFP includes a default downlink maximum bit rate. A RANs or a gateway usually support a number of TFPs. Furthermore, QoS Class Identifier (QCI) is associated with a TFP. Table 2.1 lists the set of standardised QCIs and their characteristics [29].

Table 2.1: Standardised QCIs for LTE [22]

QCI	**Packet delay (ms)**	**Packet loss**	**Services**
1	100	10^{-2}	Conversational voice
2	150	10^{-3}	Conversational voice (live streaming)
3	50	10^{-3}	Real-time gaming
4	300	10^{-6}	Non-conversational video (buffered streaming)
5	100	10^{-3}	IP Multimedia Subsystem (IMS) signalling
6	300	10^{-6}	Video (buffered streaming)
7	100	10^{-3}	Voice, video (live streaming) and interactive streaming
8/9	300	10^{-6}	TCP-based (e.g., WWW, e-mail), FTP and P2P

2.3.1.2 IEEE 802.16

The IEEE 802.16 standard defines several traffic types and each type has its specific QoS requirements. Table 2.2 lists all service classes and their requirements [30].

2.3.1.3 IEEE 802.11e

Since the original IEEE 802.11 lacks the support of QoS, the IEEE 802.11e is a supplement to enhance QoS performance [31].

In general, the IEEE 802.11e defines two ways of characterising QoS, namely, prioritised and parametrised QoS. Prioritised QoS is expressed in terms of relative delivery priority. Parametrised QoS, on the other hand, is a strict QoS requirement that is expressed in terms of quantitative values, such as data rate, delay bound and jitter bound.

2.4 Modelling of End-to-End Packet Delay

A switch or router of Fig. 2.1 can be simply modelled as a queueing system like Fig. 2.7. The system can be generally described using Kendall's notation [32, 33] (it is the standard notation used to describe and classify a queueing node) as $A/S/c/K/D$, in which

Table 2.2: 802.16e QoS service classes, specifications and applications [23]

QoS Category	**QoS Specification**	**Applications**
Unsolicited Grant Service (UGS)	maximum sustained rate, maximum latency tolerance, jitter tolerance	VoIP
Real-Time Polling Service (rtPS)	minimum reserved rate, maximum sustained rate, maximum latency tolerance, traffic priority	Streaming Audio or Video
Extended Real-Time Polling Services (ErtPS)	minimum reserved rate, maximum sustained rate, maximum latency tolerance, jitter tolerance, traffic priority	Voice with silence suppression (VoIP)
Non-Real-Time Polling Service (nrtPS)	minimum reserved rate, maximum sustained rate, traffic priority	File Transfer Protocol (FTP)
Best-Effort Service (BE)	maximum sustained rate, traffic priority	Data Transfer, Web Browsing and so on

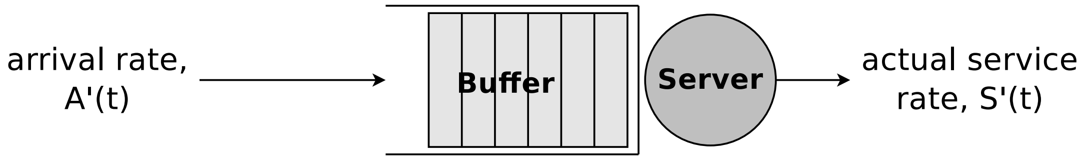


Figure 2.7: A queueing model

- A denotes an inter-arrival distribution;
- S denotes a service-time distribution;
- c denotes the number of servers;
- K denotes the buffer size (the maximum bits that a buffer can hold);
- D denotes a queue discipline.

The study of the behaviours of queueing models is called *queueing theory*. The first paper on queueing theory was published in 1909 by Danish mathematician Agner Krarup Erlang [34]. Since then, the queueing theory has been continuously studied and has extensive results [33]. Moreover, the network calculus theory is a new branch of queueing theory pioneered by R.L.Cruz [35, 36]. The theory gives a theoretical framework for analysing performance guarantees in computer networks and includes the effective capacity model, which characterise end-to-end delay distributions in wireless networks.

In this section, different types of data flow models are first introduced in Section 2.4.1. Calculating end-to-end delays based on data flow models are presented in Section 2.4.2. The network calculus theory and the effective capacity model are explained in Section 2.4.3 and 2.4.4, respectively. In Section 2.4.5, the Lindley equation is introduced. This equation will be extensively used throughout the thesis.

2.4.1 Data Flow Models

Consider there are two types of time domain, i.e., continuous time and discrete time. Data flow models can be classified into two models:

1. **continuous-time model**: $t \in \mathbb{R}_0^+ = [0, \infty)$;

2. **discrete-time model**: the model is time-slotted and is observed only at slot boundaries. $t \in \mathbb{N} = \{0, 1, 2, 3, ...\}$.

In real systems, there is always a minimum granularity (bit, word, cell or packet), therefore discrete time could always be assumed [37]. However, it is often computationally simpler to consider continuous time.

Continuous-time models could be characterised by the following stochastic processes with their formal definitions used in this thesis:

1. **arrival process** $\{A'(t), t \geq 0\}$: $A'(t)$ is the arrival rate at time t,

2. **cumulative arrival process** $\{A(t), t \geq 0\}$: $A(t)$ is the total number of bits arrived over the time interval $[0, t)$, i.e., $A(t) = \int_0^t A'(\tau) d\tau$,

3. **service process** $\{\tilde{S}'(t), t \geq 0\}$: $\tilde{S}'(t)$ is the service rate (the server is *capable* to serve) at time t,

4. **cumulative service process** $\{\tilde{S}(t), t \geq 0\}$: $\tilde{S}(t)$ is the number of bits that the server is *capable* to serve over the time interval $[0, t)$, i.e., $\tilde{S}(t) = \int_0^t \tilde{S}'(\tau) d\tau$,

5. **actual service process** $\{S'(t), t \geq 0\}$: $S'(t)$ is the actual service rate (the server *actually* served) at time t,

6. **cumulative actual service process** $\{S(t), t \geq 0\}$: $S(t)$ is the number of bits that is *actually* served by the server over the time interval $[0, t)$, i.e., $S(t) = \int_0^t S'(\tau)d\tau$.

and for discrete-time models, the following processes are used

1. **arrival process** $\{A'[n], n \geq 1\}$: $A'[n]$ is the number of bits arrived during slot n and is a constant because of the assumption of the constant bit arrival rate,

2. **cumulative arrival process** $\{A[n], n \geq 1\}$: $A[n]$ is the total number of bits arrived from slot 1 to slot n, i.e., $A[n] = \sum_{i=1}^{n} A'[i]$,

3. **service process** $\{\tilde{S}'[n], n \geq 1\}$: $\tilde{S}'[n]$ is the number of bits that the server is *capable* to serve during slot n,

4. **cumulative service process** $\{\tilde{S}[n], n \geq 1\}$: $\tilde{S}[n]$ is the number of bits that the server is *capable* to serve from slot 1 to slot n, i.e., $\tilde{S}[n] = \sum_{i=1}^{n} \tilde{S}'[i]$,

5. **actual bit service process** $\{S'[n], n \geq 1\}$: $S'[n]$ is the number of bits that is *actually* served by the server during slot n,

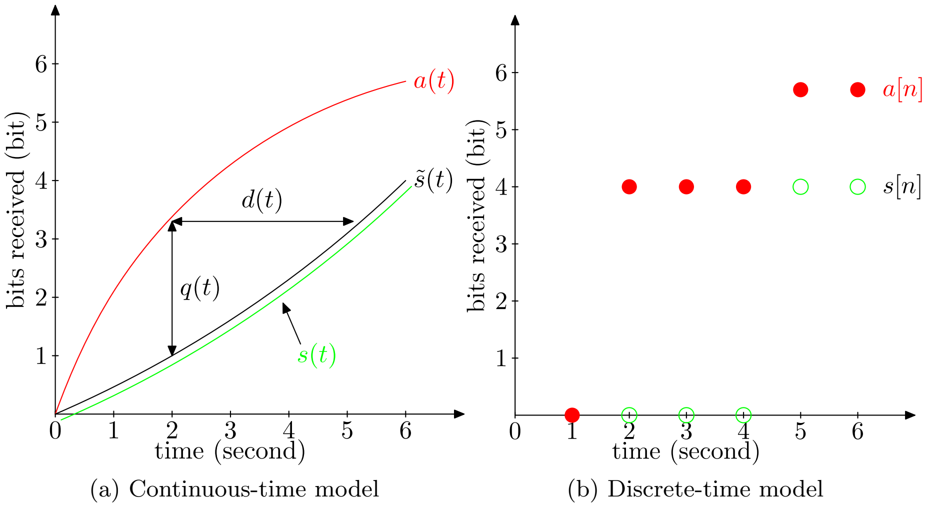
6. **cumulative actual service process** $\{S[n], n \geq 1\}$: $S[n]$ is the number of bits that is *actually* served by the server from slot 1 to slot n, i.e., $S[n] = \sum_{i=1}^{n} S'[i]$.

Fig. 2.8 shows realisations of cumulative processes in these two data flow models. Furthermore, there are two types of traffic source models [37]:

1. **fluid traffic model**: packet sizes are infinitesimally small or packets are arrived bit by bit;

2. **packetised traffic model**: packets have non-negligible sizes.

Fluid or packetised traffic models can exist in continuous-time or discrete-time models.



(a) Continuous-time model

(b) Discrete-time model

Figure 2.8: Examples of data flow models

2.4.2 End-to-End Packet Delay in Fluid and Packetised Models

As shown in Fig. 2.8a, if it is the *fluid traffic* model and the queue discipline is First-In First-Out (FIFO), the end-to-end delay of the packet arrives at t, $d(t)$, is the horizontal difference between the realisations of a cumulative arrival process $\{a(t), t \geq 0\}$ and a cumulative actual service process $\{s(t), t \geq 0\}$:

$$d(t) = \inf_{\tau \geq 0} \{a(t) \leq s(t+\tau)\} \tag{2.10}$$

and $q(t)$ is the vertical difference.

The *packetised* traffic model is a more realistic model than the fluid traffic model because packet sizes are usually non-negligible in practical situations. This model further allows us to study queueing delays of packets. In telecommunication and data networks, the end-to-end packet delay D_n usually consists of *four* elements [38]:

1. **Transmission delay** $D_n^{(t)}$: the time taken for a packet to be transmitted at the transmitter

2. **Radio propagation delay** $D_n^{(r)}$: the time taken for a packet to reach its receiver

3. **Signal processing delay** $D_n^{(s)}$: the time taken for a packet to be decoded at the

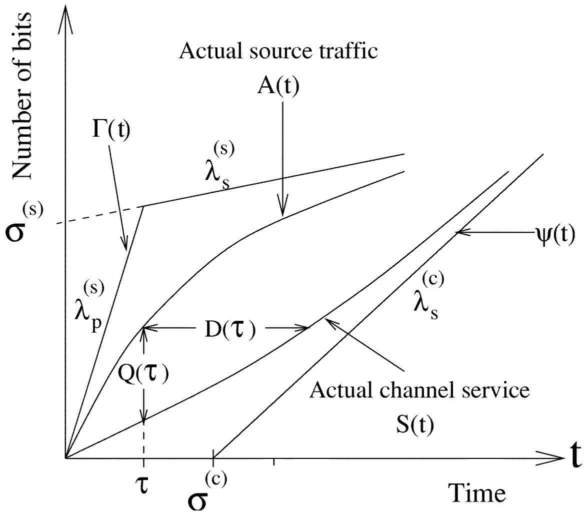


Figure 2.9: An example of traffic and service characterisation

receiver

4. **Queueing delay** $D_n^{(q)}$: the time take for a packet to wait in buffers or the time between the packet arrival instant and the instant when the first bit or the least significant bit was sent.

Their mathematical relation can be simply expressed as

$$D_n = D_n^{(t)} + D_n^{(r)} + D_n^{(s)} + D_n^{(q)} \tag{2.11}$$

Assuming that the radio transmission delay and signal processing delay are small enough to neglect, (2.11) becomes

$$D_n = D_n^{(t)} + D_n^{(q)} \tag{2.12}$$

2.4.3 Network Calculus Theory

Pioneered by R. Cruz [35], the network calculus theory is a analysis tool for the QoS guarantees in wired networks, such as the Asynchronous Transfer Mode (ATM) networks and the Internet.

The traffic characterisation in the network calculus theory requires that the amount of data (i.e., bits as a function of time) produced by a source conform to an upper bound, called the traffic envelope $\Gamma(t)$. Similarly, the service characterisation $\Psi(t)$ for guaranteed service in the network calculus theory is a guarantee of a minimum service rate. Functions $\Gamma(t)$ and $\Psi(t)$ are specified in terms of certain traffic and service parameters, respectively.

The examples of traffic and service characterisations are provided as below, starting with the concept of the leaky bucket. A leaky bucket is a device that shapes the arrival rate $A'(t)$. The bucket of this device is initially empty and it can hold up to $\sigma^{(s)}$ bits. The bucket also has a hole and leaks at a rate of $\lambda_s^{(s)}$ bit per second ($\lambda_s^{(s)}$ is termed as the sustainable rate) when it is not empty.

- **Traffic envelop** $\Gamma(t)$: a traffic envelope characterises the source behaviour in the following manner: over any window of size t, the cumulative arrival process does not exceed $\Gamma(t)$ (see Fig. 2.9). For example, the Usage Parameter Control (UPC) parameters used in ATM systems [39] are specified by

$$\Gamma(t) = \min\{\lambda_p^{(s)}t, \lambda_s^{(s)}t + \sigma^{(s)}\} \tag{2.13}$$

where $\lambda_p^{(s)}$, $\lambda_s^{(s)}$ and $\sigma^{(s)}$ are the peak data rate, the sustainable rate and the leaky-bucket size, respectively [40].

- **Service characterisation** $\Psi(t)$: Like $\Gamma(t)$ is the upper bound of the cumulative arrival process, a network service characterisation $\Psi(t)$ gives a lower bound of the cumulative actual service process. $\Psi(t)$ has the property that $\Psi(t) \leq S(t)$ for any time t. Both $\Gamma(t)$ and $\Psi(t)$ are negotiated during the admission control and resource reservation phase. An example of a network service characterisation is the service specification (R-SPEC) curve used for guaranteed service on the

Internet [41, 40]:

$$\Psi(t) = \left(\lambda_s^{(c)}(t - \sigma^{(c)})\right)^+ \tag{2.14}$$

where $(x)^+ = \max\{x, 0\}$, $\lambda_s^{(c)}$ is the constant service rate, and $\sigma^{(c)}$ is the delay error term (due to propagation delay, link sharing and so on).

As shown in Fig. 2.9, the traffic envelop consists of two segments; the first segment has a slope equal to the peak source data rate $\lambda_p^{(s)}$, while the second segment has a slope equal to the sustainable rate $\lambda_s^{(s)}$, with $\lambda_s^{(s)} < \lambda_p^{(s)}$. $\sigma^{(s)}$ is the axis intercept of the second segment. $\Gamma(t)$ has the property that $A(t) \leq \Gamma(t)$ for any time t. The service characterisation curve $\Psi(t)$ also consists of two segments: the horizontal segment indicates that no packet is being serviced due to delay errors while the second segment has a slope equal to the service rate $\lambda_s^{(c)}$.

2.4.3.1 Min-Plus Operation

The *min-plus operation* associates $A(t)$, $\tilde{S}(t)$ with $S(t)$ [42, 37, 43]:

$$S(t) = \min_{0 \leq s \leq t} \{A(t-s) + \tilde{S}(s)\} \tag{2.15}$$

This operation is extensively used in the network calculus theory. Moreover, if we replace min in (2.15) into addition, and addition into multiplication, (2.15) will be transformed into the conventional convolution operation:

$$f * g(t) = \int_0^t f(t-s)g(s)ds, \quad 0 \leq t \leq \infty \tag{2.16}$$

2.4.4 Effective Capacity Model

Unlike wireline links that typically have bounded sources and capacities (see Section 2.4.3), wireless channels have low reliability and time-varying capacities, which may cause severe QoS violations. Therefore, it is better to characterise wireless channels from a statistical view point.

The effective capacity model is an analysis tool for modelling the end-to-end delay in a stochastic manner and it was proposed by Wu and Negi in 2003 [4]. After the advent of the effective capacity model, numerous efforts were made to

1. calculate the effective capacity function in various communication systems and channels [44, 45, 46, 47, 48, 49],
2. use the effective capacity function to optimised systems [50, 51, 52, 53, 54, 55, 56],
3. develop applications based on the model and estimator [57, 58, 59, 60, 61, 62, 63, 64, 65, 66],
4. revise the model for more realistic communication scenarios [57].

The following property is needed in the propositions for all effective capacity-based models. Let $\{X(t), t \geq 0\}$ be a stochastic process. The asymptotic log-moment generation function of $X(t)$ is used in the large deviation theory [67] and it is given by

$$\Lambda_X(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E[e^{-\theta \int_0^t X(\tau)d\tau}] \tag{2.17}$$

Property 2.4.1 *1. The asymptotic log-moment generation function of $X(t)$, $\Lambda_X(\theta)$, is finite for all $\theta \in (-\infty, \infty)$*

2. $\Lambda_X(\theta)$ is differentiable for all $\theta \in (-\infty, \infty)$

The conventional effective capacity model appeared in [4] is introduced in Section 2.4.4.1. The Effective Bandwidth-Effective Capacity (EB-EC) model and the packetised effective capacity model with a leaky bucket are based on the work in [57] and these are introduced in Sections 2.4.4.2 and 2.4.4.3, respectively.

2.4.4.1 Conventional Effective Capacity Model

In simple words, the effective capacity model associates effective capacity functions and constant arrival rates with CCDFs of packet end-to-end delay.

Let $r(t)$ be the instantaneous channel capacity at time t and $\{r(t), t \geq 0\}$ is stationary. The effective capacity function of $-r(t)$ is defined as

$$\alpha^{(c)}(\theta) = \frac{\Lambda_{-r}(\theta)}{\theta} = \lim_{t\to\infty} \frac{1}{t} \log E[e^{-\theta \int_0^t r(\tau)d\tau}] \tag{2.18}$$

Consider a queue of infinite buffer size supplied by a constant data rate source (the constant data rate is μ). If the property 2.4.1 holds for $\{r(t), t \geq 0\}$ and there is a unique solution θ^* for the equation

$$\alpha^{(c)}\left(\frac{\theta}{\mu}\right) = \mu \tag{2.19}$$

then the probability of the packet delay $D(\infty)$ exceeding a delay bound $D_{\max}$ can be estimated by [68]

$$\Pr\{D(\infty) \geq D_{\max}\} \approx e^{-\theta^* D_{\max}} \tag{2.20}$$

From (2.20), the tail distribution of packet queueing delay is exponentially bounded.

Let ${\alpha^{(c)}}^{-1}(\mu)$ be the inverse function of $\alpha^{(c)}(\theta)$. The following equation holds based on (2.19)

$$\frac{\theta}{\mu} = {\alpha^{(c)}}^{-1}(\mu) \tag{2.21}$$

Given a traffic load μ, the solution θ^* of (2.19) can be obtained from (2.21):

$$\theta^* = \mu\left({\alpha^{(c)}}^{-1}(\mu)\right) \tag{2.22}$$

Since the solution θ^* is a function μ, θ^* is denoted as $\theta^{(c)}(\mu)$ in this thesis.

Furthermore, it is found that for small $D_{\max}$, the following equation is more accurate

[69] than (2.20) is:

$$\Pr\{D(\infty) \geq D_{\max}\} \approx \gamma^{(c)}(\mu)e^{-\theta^{(c)}(\mu)D_{\max}} \tag{2.23}$$

where $\gamma^{(c)}(\mu)$ is regarded as the probability of the packet delay being greater than zero when the queueing system has reached the steady state, i.e.,

$$\gamma^{(c)}(\mu) = \Pr(D(\infty) > 0) \tag{2.24}$$

(2.23) is the mathematical modelling of packet delay in the effective capacity model. In (2.23), $\gamma^{(c)}(\mu)$ is called the *probability of non-empty buffer* and $\theta^{(c)}(\mu)$ is called the *QoS exponent of a connection.*

2.4.4.2 Effective Bandwidth-Effective Capacity Model

In this section, the *EB-EC model*, which is a effective capacity-based end-to-end delay distribution modelling with variable bit-rate sources, is introduced. Specifically, the arrival process $\{A'(t), t \geq 0\}$ is

- a type of fluid traffic,
- stationary and
- PDF of the arrival rate $A'(t)$ (in bit) is *exponentially decaying* with its effective bandwidth function known

> In simple words, the effective bandwidth function characterises the asymptotic behaviour of a queueing system. The concept of effective bandwidth was first proposed independently in [70, 71, 72] and was initially developed for high-speed digital networks. The general framework of the theory, including the computation of the effective bandwidth for Markov processes and other general processes and the associated calculus, was carried out in [73, 74, 75,

76, 77, 78, 79]. The effective bandwidth theory has been used to analyse various traffic sources [76, 69, 80], queueing models [81, 75, 82] and networks [83, 84].

For convenience, the definition of the effective bandwidth is replicated here. Assume that the arrival process $\{A'(t), t \geq 0\}$ is stationary and the asymptotic log-moment generating function of the arrival process, defined as

$$\Lambda_{A'}(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E[e^{\theta \int_0^t A(\tau)d\tau}] \tag{2.25}$$

exists for all $\theta \geq 0$. Then, the effective bandwidth function of the arrival process is defined as

$$\alpha^{(s)} = \frac{\Lambda_{A'}(\theta)}{\theta}, \forall \theta > 0 \tag{2.26}$$

From the previous section, the effective capacity function of a wireless capacity process $\{r(t), t \geq 0\}$, $\alpha^{(c)}(\theta)$, is given by (2.18).

Further assume that the log-moment generating functions of the channel $\Lambda_{-r}(\theta)$ and the arrival process $\Lambda_{A'}(\theta)$ satisfy Property 2.4.1. If the following equation

$$\alpha^{(s)}(\theta) = \alpha^{(c)}(\theta) \tag{2.27}$$

has a unique solution θ_1^*, the distribution of end-to-end delay $D(\infty)$ satisfies

$$\lim_{D_{\max}\to\infty} \frac{\log \Pr\{D_\infty > D_{\max}\}}{D_{\max}} = -\theta_{eb} \tag{2.28}$$

where $\theta_{eb} = \theta_1^* \alpha^{(s)}(\theta_1^*)$

For small $D_{\max}$, the CCDF of packet delay can be expressed as

$$\Pr\{D(\infty) \geq D_{\max}\} \approx \gamma_{eb} e^{-\theta_{eb} D_{\max}} \tag{2.29}$$

where γ_{eb} is called the *probability of non-empty buffer* and θ_{eb} is called the *QoS exponent of a connection.*

2.4.4.3 Packetised Effective Capacity Model with a Leaky Bucket

In the previous sections 2.4.4.1 and 2.4.4.2, fluid traffic model is assumed. This section introduces the packetised effective capacity model with a leaky bucket, which is for

- packetised traffic model
- constrained by a leaky bucket

The effective capacity function of a wireless capacity process $\{r(t), t \geq 0\}$ is characterised by $\alpha^{(c)}(\theta)$; the log-moment generating functions of the channel $\Lambda_{-r}(\theta)$ satisfies Property 2.4.1. Given a traffic flow having maximum packet size $L_{\max}$ and constrained by a leaky bucket with bucket size $\sigma^{(s)}$ and token generating rate $\lambda_s^{(s)}$, the distribution of end-to-end delay $D(\infty)$ satisfies

$$\lim_{D_{\max}\to\infty} \frac{\log \Pr\{D_\infty > D_{\max}\}}{D_{\max} - L_{\max}/\lambda_s^{(s)} - \sigma^{(s)}/\lambda_s^{(s)}} = -\theta_{lb} \tag{2.30}$$

where θ_{lb} is the unique solution of the following equation

$$\alpha^{(c)}\left(\frac{\theta}{\lambda_s^{(s)}}\right) = \lambda_s^{(s)} \tag{2.31}$$

2.4.5 Lindley Equation

The Lindley equation describes the evolution of queueing systems. It has two versions: the first one is for generating realisations of delay processes (introduced in Section

2.4.5.1); the second one is for generating realisations of queue length processes (introduced in Section 2.4.5.2).

2.4.5.1 Lindley Equation for Delay Processes

The first Lindley equation for the waiting times of the customers was proposed in 1952 by D. V. Lindley [5]. Since the equation is applied to any G/G/1/∞/FIFO queueing model, it is often regarded as a *general form* for

- delay processes
- continuous-time or discrete-time models
- packetised traffic models only

Let T_n, C_n and $D_n^{(q)}$ denote the random variables of the n^{th} inter-arrival time (the time difference between the n^{th} packet arrival instant and $(n+1)^{\text{st}}$ packet arrival instant), the service time of the n packet and the queueing delay value of the n^{th} packet, respectively.

For a G/G/1/∞/FIFO queueing model, $D_n^{(q)}$ is given by the first version of the Lindley equation:

$$D_n^{(q)} = \max\{0, D_{n-1}^{(q)} + C_{n-1} - T_{n-1}\} \tag{2.32}$$

Eq. (2.32) is a recursive equation, stating the relation between the queueing delay of the n^{th} and $(n-1)^{\text{st}}$ packets with effect of the service time of the $(n-1)^{\text{st}}$ packet and the $(n-1)^{\text{st}}$ inter-arrival time. If T_n or C_n is a random variable, $D_n^{(q)}$ is a random variable.

On the basis of (2.12), the end-to-end delay of the n^{th} packet is given by

$$D_n = C_n + D_n^{(q)} \tag{2.33}$$

2.4.5.2 Lindley Equation for Queue Length Processes

In a time-slotted queueing system, a *queue length process* $\{Q[n], n \geq 1\}$ is an ordered sequence of queue lengths from the slot 1; and it is a stochastic process. The second Lindley equation is for

- queue length processes
- discrete-time models only
- fluid or packetised traffic models

Denote by $Q[n]$ the queue length at the $n + 1^{\text{st}}$ slot boundary. Q_n is a random variable. For a G/G/1/∞ queueing model, $Q[n]$ is given by the second version of the Lindley equation:

$$Q[n] = \max\{0, Q[n-1] + A'[n] - \tilde{S}'[n]\} \tag{2.34}$$

Similar to the attributes of (2.32), (2.34) is also a recursive equation that is used for calculating queue length processes. Moreover, it is worth noting that in contrast to the first Lindley equation, (2.34) does not require the FIFO assumption.

In summary, suppose that we are interested in an end-to-end delay process $\{D_n, n \geq 1\}$ in a queueing system. The effective capacity-based models answers the question of what the probability $\Pr\{D_n > D_{\max}\}$ will be while the Lindley equation answers the question of what a realisation of the delay process $\{d_n, n \geq 1\}$ will be.

Furthermore, the three effective capacity-based models (introduced in Sections 2.4.4.1, 2.4.4.2 and 2.4.4.3) are suitable for different cases. The conventional effective capacity model is the simplest version but serves as the basis of the other two effective capacity-based models. For fluid traffic model, the EB-EC model is a general model with a relatively weak constraints on the arrival process and the service process. For packetised traffic model, the model in Section 2.4.4.3 is not as general as the EB-EC model for fluid traffic model in the sense that the traffic source must be constrained by a leaky bucket.

2.5 Conclusions

The concept of QoS is tied with packet-switching communication systems and QoS metrics include packet delay, through and packet loss ratio. When delay-sensitive traffic is carried in systems, mechanisms for providing end-to-end delay guarantees are needed.

In general, this chapter surveyed three concepts that relates to this thesis, namely *multi-hop wireless networks* (Section 2.2), *packet delay* (Section 2.3) and *mathematical modelling of packet delay* (Section 2.4).

The multi-hop wireless technology is a natural combination of centralised networks and ad-hoc networks. This technology answers many tough issues encountered in wireless communication systems, like dead zone problem and the deployment in the harsh terrain. In Section 2.2, the architecture of multi-hop wireless networks from the system and network perspective was introduced. Its advantages, challenges were thoroughly investigated, together with its three types of implementations. This section was finalised by an introduction of the multi-hop wireless technology in the LTE-A, the IEEE 802.16j and the IEEE 802.11s standards.

The concept of end-to-end packet delay and its use in assessment of network performance was carefully studied in Section 2.3. This section was also ended with an introduction of the QoS provisioning in the LTE, the IEEE 802.16e and the IEEE 802.11 standards.

Mathematically modelling the end-to-end delay was shown in the last section (Section 2.4). The system model, the data flow models and the network calculus theory were first introduced. As the extensions of the network calculus theory that model end-to-end packet delay distributions, three basic models are explained: 1) the effective capacity model, 2) the effective bandwidth-effective capacity model and 3) the packetised effective capacity model with a leaky bucket. These three models will be used for our model development from Chapters 3 to 7. The Lindley equation for describing the evolution of end-to-end packet delay and queue length was introduced in Section 2.4.5.1. This equation will be used in the next chapter to validate the the packetised effective capacity model with a leaky bucket (the third basic model) and in the Chapters 6 and 7 for developing realistic simulation platforms.

Chapter 3

End-to-End Packet Delay Analysis of an Internet Traffic Trace over Gigabit Ethernet

3.1 Introduction

Modelling end-to-end packet delay in Internet Protocol (IP) networks depends on the statistical nature of the packet inter-arrival and packet size distributions. Self-similar [85, 86], long-range dependent [87], heavy-tail distributed [88] and Poisson[89] models are assumed to characterise traffic statistics. Some complicated simulators and the Monte-Carlo method are used to evaluate Internet performances [90]. However, such research is too theoretical to be of practical use.

The packetised effective capacity model with a leaky bucket (see Section 2.4.4.3) shows that the tail distributions of end-to-end packet delay are exponentially bounded. On the other hand, the Lindley equation (see 2.4.5.1) characterises end-to-end packet delay in any G/G/1/∞/FIFO queueing systems [5]. The use of the Lindley equation to analyse the Internet traffic traces was first reported by Park et al. in 2005 [91]. This equation has been used to analyse data/voice services of CDMA 2000 systems [92].

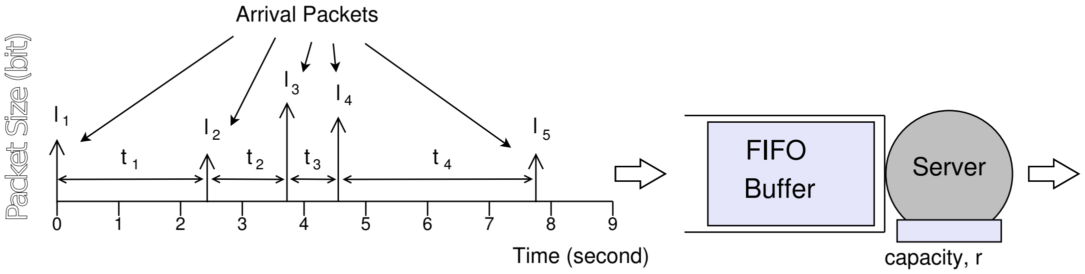


Figure 3.1: System model of a gigabit Ethernet system

In this chapter, the Lindley equation is applied to an Internet traffic over gigabit Ethernet to 1) analyse the end-to-end packet delay performances and 2) validate the effective capacity model. The traffic trace is publicly downloadable from the University of Massachusetts Amherst (UMASS) trace repository [93]. Most results discussed in this chapter have been published in the ICT 2013 conference proceedings [94].

In Section 3.2, a gigabit Ethernet gateway is modelled as a queueing system. The methodology used in this chapter is explained in Section 3.3. This section includes the explanation of the fitting distribution technique, the Internet traffic trace used in this chapter and calculating end-to-end packet delay using the Lindley equation. The results and discussion are presented in Section 3.4. Section 3.5 concludes this chapter.

3.2 System Model

A gigabit Ethernet system can be modelled as a G/G/1/∞/FIFO queueing model, as shown in Fig. 3.1. Packets are randomly generated from higher layers and are first stored in a buffer and later served by *one* server. The capacity of the server is fixed to a constant value c; in a gigabit Ethernet system, such a value equals 1 gigabit. The buffer size is assumed to be infinite. Packets in the buffer are served in a fair manner, which is the First-In First-Out (FIFO) discipline.

Let V_n and L_n denote the random variables of the n^{th} packet arrival instant and of the size of the n^{th} packet, respectively. Fig. 3.1 also shows an example of a sequence of packets arriving at the queueing system. The arrivals are a sequence of unequal-height

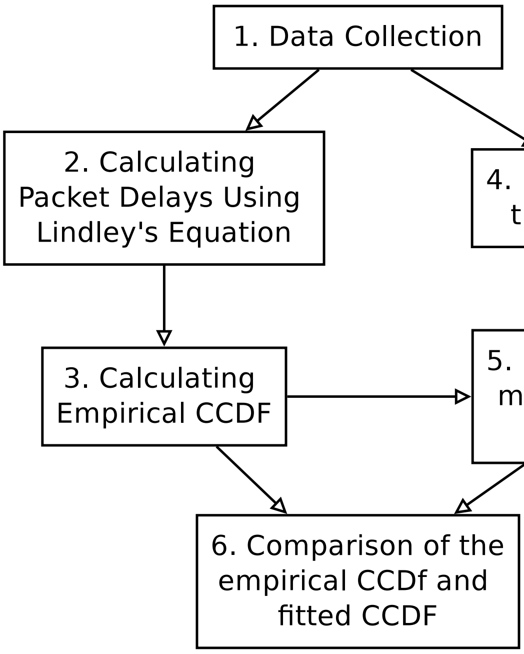

1. Data Collection
2. Calculating Packet Delays Using Lindley's Equation
4. Basic Analysis of the Traffic trace
3. Calculating Empirical CCDF
5. Using regression method to fit the CCDF
6. Comparison of the empirical CCDf and fitted CCDF

Figure 3.2: Steps of validating the effective capacity model

impulses, which are randomly scattered in a time line; the heights of impulses indicate respective packet sizes.

Let T_n and C_n denote the random variables of the n^{th} inter-arrival time (the time difference between the n^{th} packet arrival instant and $(n+1)^{\text{st}}$ packet arrival instant) and the service time of the n packet, respectively. Their values can be calculated from the following equations

$$T_n = V_{n+1} - V_n \tag{3.1}$$

$$C_n = \frac{L_n}{c} \tag{3.2}$$

For any G/G/1 queueing systems, the inter-arrival and service distributions are assumed to be general.

3.3 Method

The *distribution fitting method* [95] is the fitting of a probability distribution to a series of data and will be used to validate the effective capacity model. Fig. 3.2 shows the detailed steps of such a method.

Steps 1, 2 and 3 are a procedure of plotting empirical distribution. Step 1 involves data gathering, basic data manipulation, which will enhance data quality, and evaluation, including some measures of data. Step 2 uses Lindley's equation to obtain a series of packet delays. Step 3 is to plot empirical distributions from the series of packet delay. The basic data analysis is carried out in step 4. In step 5, the empirical Complementary Cumulative Density Function (CCDF) obtained in Step 3 is fitted by the regression method. The fitted distribution giving a close fit is supposed to lead to good predictions.

3.3.1 Internet Traffic Trace over Gigabit Ethernet

An Internet traffic trace over gigabit Ethernet from UMASS trace repository [93] was downloaded. This trace is from a fibre gigabit Ethernet connection entering UMASS on 14th, November, 2004. The monitoring infrastructure consists of passive taps that redirect the signal from the underlying Gigabit fibre link. The tapped signal passes through regeneration equipment and is finally fed into an Endace® Data Acquisition and Generation (DAG) card in a special-purpose PC. The DAG card strips off the TCP/IP headers of the packets, affixes an accurate time stamp on the header-record and writes it to a file.

The total size of the file is 382 MB. Wireshark was used to filter out the information of arrival instants and packet sizes out from the file and then sort packets based on their arrivals. After processing the file, 11,976,472 packet arrival instants and sizes were obtained over a period of 81.63 seconds.

For the purpose of testing stationarity, the trace is first divided into *twelve* groups so each of which has 1,000,000 packets except the last group of 976,473 packets. In each group except the last one, the information of packet arrival instants and packet sizes are as follows:

$$\mathbf{v} = \{v_1, v_2, v_3, \cdots v_{1000000}\} \tag{3.3}$$

$$\mathbf{l} = \{l_1, l_2, l_3, \cdots l_{1000000}\} \tag{3.4}$$

Eqs. (3.3) and (3.4) are the information needed to carry out analysis in this chapter.

3.3.2 Calculating End-to-End Packet Delay

The Lindley equation associates packet inter arrivals and service times with end-to-end packet delays. On the basis of (3.1) and (3.2), the sequences of packet arrival instants (3.3) and packet sizes (3.4) are first translated to the sequences of inter arrivals $\mathbf{t}$ and service times $\mathbf{c}$:

$$\mathbf{t} = \{(v_2 - v_1), (v_3 - v_2), \cdots (v_{1000000} - v_{999999})\} \tag{3.5}$$

$$\mathbf{c} = \{\frac{l_1}{c}, \frac{l_2}{c}, \cdots \frac{l_{999999}}{c}\} \tag{3.6}$$

Denote by $d_n^{(q)}$ and d_n the queueing delay and the end-to-end packet delay of the n^{th} packet. Assume the queueing delay of the first packet is 0. For a G/G/1/∞/FIFO queueing model, $d_n^{(q)}$ is given by the first version of the Lindley equation (see Section 2.4.5.1 for further discussion):

$$d_n^{(q)} = \max\{0, d_{n-1}^{(q)} + c_{n-1} - t_{n-1}\} \tag{3.7}$$

By recursively using (3.7), a sequence of packet queueing delay is obtained by

$$\mathbf{d}^{(q)} = \{d_1^{(q)}, d_2^{(q)}, d_3^{(q)}, \cdots d_{1000000}^{(q)}\} \tag{3.8}$$

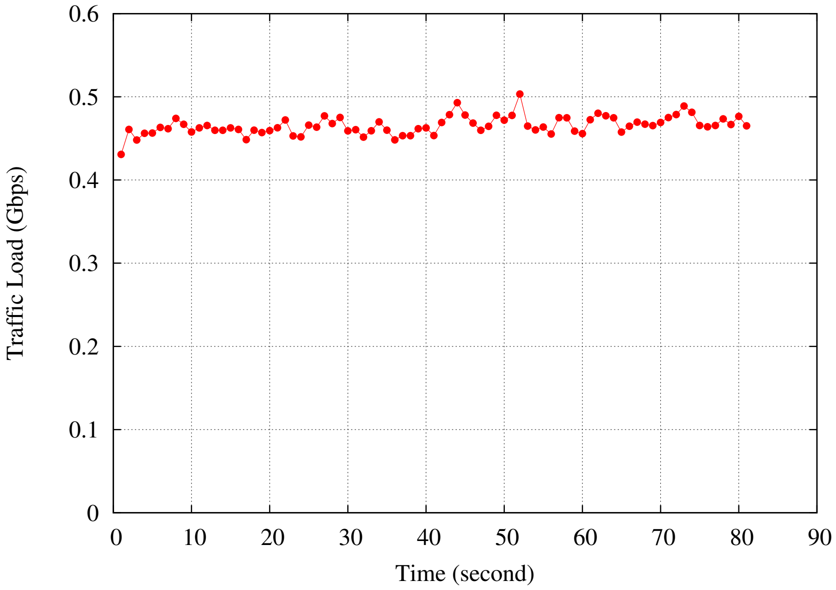


Figure 3.3: Traffic load over time

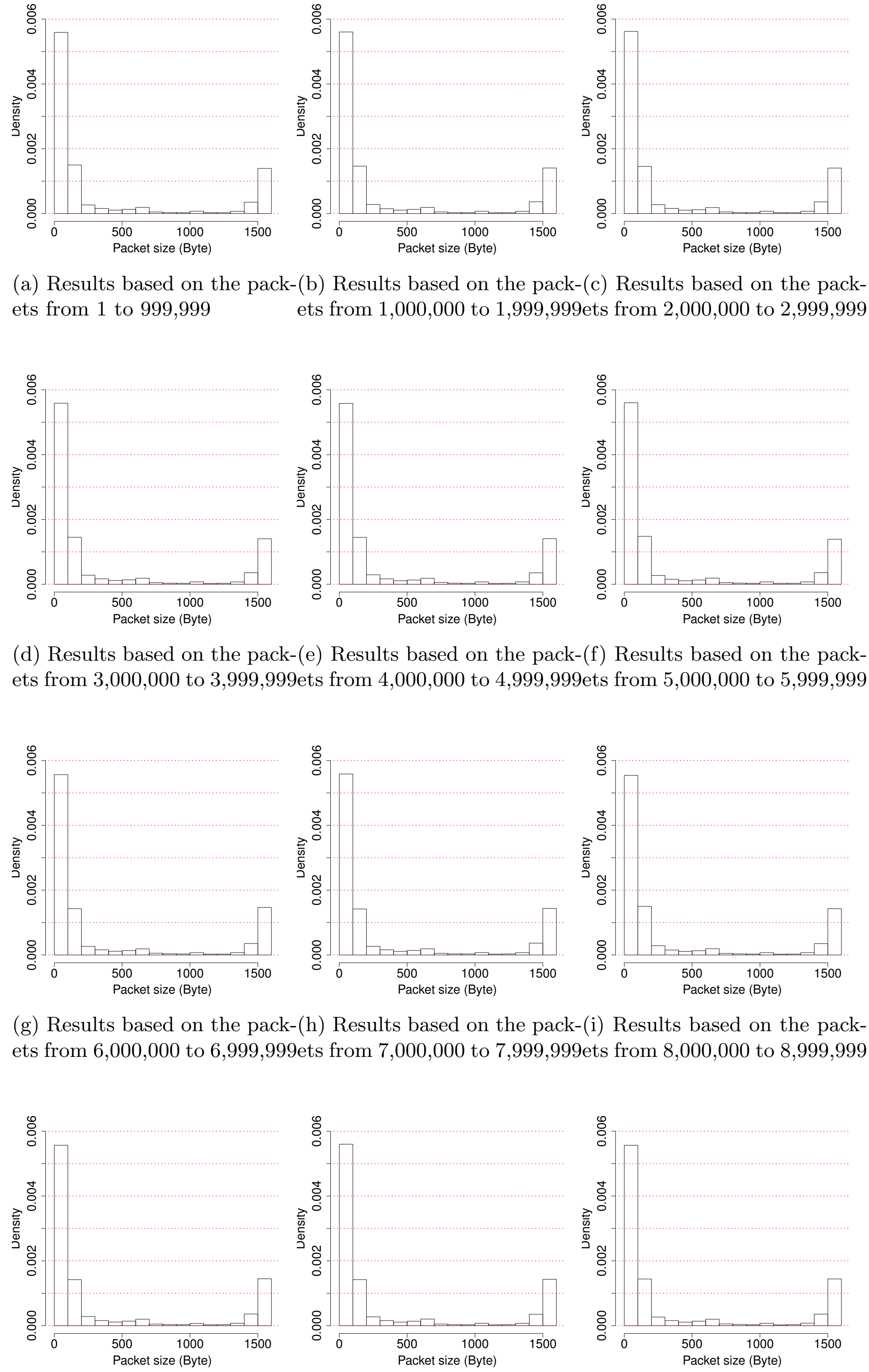
Finally, the sequence of end-to-end packet delay **d** is given by

$$\mathbf{d} = \{(d_1^{(q)} + c_1), (d_2^{(q)} + c_2), (d_3^{(q)} + c_3), \cdots (d_{1000000}^{(q)} + c_{1000000})\} \tag{3.9}$$

3.3.3 Basic Analysis of the Trace

Fig. 3.3 shows traffic loads per second over the collection period (81.63 seconds). It is clear that the traffic loads are fairly constant during this period. For the entire duration, the average inter-arrival time and the standard deviation of the inter-arrival time are 6.90 μs and 13.85 μs, respectively, which indicates that the traffic does not have Poison-distributed inter-arrival times.

On the basis of (3.4), the histograms of packet sizes in twelve groups are shown in Fig. 3.4. The Y-axes in the figure are the frequency density (the height of a rectangle is equal to the frequency divided by the width of the interval). In the trace, the minimum packet size is 64 bytes and the maximum packet size $L_{\max}$ is 1518 bytes, which conform to the standard of Ethernet II. Moreover, these two types of packets (64-byte and 1518-byte packets) are mostly common in the trace. Such observations conform to the



(a) Results based on the packets from 1 to 999,999

(b) Results based on the packets from 1,000,000 to 1,999,999

(c) Results based on the packets from 2,000,000 to 2,999,999

(d) Results based on the packets from 3,000,000 to 3,999,999

(e) Results based on the packets from 4,000,000 to 4,999,999

(f) Results based on the packets from 5,000,000 to 5,999,999

(g) Results based on the packets from 6,000,000 to 6,999,999

(h) Results based on the packets from 7,000,000 to 7,999,999

(i) Results based on the packets from 8,000,000 to 8,999,999

(j) Results based on the packets from 9,000,000 to 9,999,999

(k) Results based on the packets from 10,000,000 to 11,999,999

(l) Results based on the packets from 11,000,000 to 11,976,472

Figure 3.4: Histograms of packet sizes in 12 groups

Table 3.1: Descriptive statistics of end-to-end packet delay in 12 groups of the Internet traffic trace

Group	Packet range	Avg. E2E delay (μs)	Jitter (μs)
1	1 – 999,999	12.21	6.84
2	1,000,000 – 1,999,999	12.37	6.94
3	2,000,000 – 2,999,999	12.22	6.84
4	3,000,000 – 3,999,999	12.32	6.87
5	4,000,000 – 4,999,999	12.37	6.87
6	5,000,000 – 5,999,999	12.35	6.98
7	6,000,000 – 6,999,999	12.56	7.01
8	7,000,000 – 7,999,999	12.55	7.02
9	8,000,000 – 8,999,999	12.40	6.91
10	9,000,000 – 9,999,999	12.43	6.98
11	10,000,000 – 10,999,999	12.55	7.06
12	11,000,000 – 11,976,4721	12.44	6.98

finding for wide-area Internet traffic reported by Thompson et al [96].

3.3.4 Regression Method

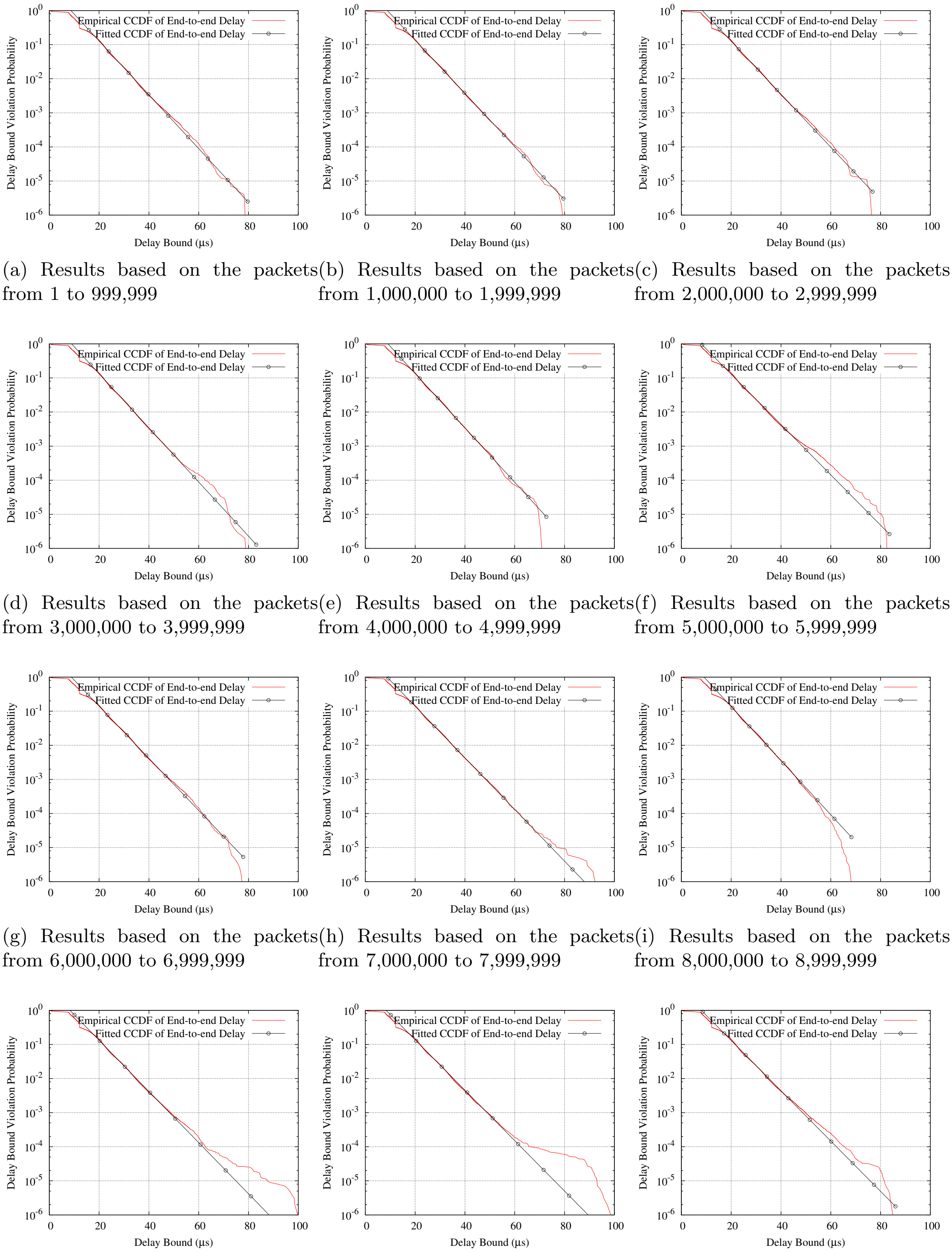
It is shown in the packetised effective capacity model with a leaky bucket (see Chapter 2, Section 2.4.4.3) that the tail distributions is exponentially bounded. Therefore, when the maximum delay bound $D_{\max}$ is large enough, the logarithm of CCDF of queueing delay can be linearised to

$$Y = \theta_{lb} D_{\max} + B \tag{3.10}$$

where $Y = \ln(\Pr\{D_\infty > D_{\max}\})$ and B is the y-intercept of (3.10). Using the *least squares* approach, one finds the parameters θ_{lb} and B from a linear regression of Y on $D_{\max}$ so the CCDF is fully defined.

3.4 Results and Discussion

Fig. 3.5 shows the empirical and fitted CCDFs of end-to-end packet delay from twelve groups. The x-axes are delay bound (the unit is microsecond) and the y-axes are Delay Bound Violation Probability (DBVP) in log scale. The empirical results are shown



(a) Results based on the packets from 1 to 999,999

(b) Results based on the packets from 1,000,000 to 1,999,999

(c) Results based on the packets from 2,000,000 to 2,999,999

(d) Results based on the packets from 3,000,000 to 3,999,999

(e) Results based on the packets from 4,000,000 to 4,999,999

(f) Results based on the packets from 5,000,000 to 5,999,999

(g) Results based on the packets from 6,000,000 to 6,999,999

(h) Results based on the packets from 7,000,000 to 7,999,999

(i) Results based on the packets from 8,000,000 to 8,999,999

(j) Results based on the packets from 9,000,000 to 9,999,999

(k) Results based on the packets from 10,000,000 to 11,999,999

(l) Results based on the packets from 11,000,000 to 11,976,472

Figure 3.5: Empricial CCDFs of packet end-to-end delay in 12 groups

Table 3.2: Estimated θ_{lb} in 12 groups of the Internet traffic trace

Group	Packets	$\hat{\theta}_{lb}$
1	1 – 999,999	181729.12
2	1,000,000 – 1,999,999	179675.90
3	2,000,000 – 2,999,999	178971.23
4	3,000,000 – 3,999,999	182850.23
5	4,000,000 – 4,999,999	183696.01
6	5,000,000 – 5,999,999	169973.27
7	6,000,000 – 6,999,999	176069.23
8	7,000,000 – 7,999,999	174413.12
9	8,000,000 – 8,999,999	182666.81
10	9,000,000 – 9,999,999	172803.58
11	10,000,000 – 10,999,999	170361.33
12	11,000,000 – 11,976,4721	169831.43

in red solid lines while the fitted CCDFs are plotted in black lines with circle marks. From the figures, the empirical and estimated CCDFs are well matched especially when a delay bound is greater than 20 μs and less than 50 μs, indicating that the tail distributions of end-to-end packet delay are all exponentially bounded. Tables 3.1 and 3.2 list descriptive statistics of end-to-end delays and estimated θ_{lb}s in twelve groups, respectively.

3.5 Conclusions

In this chapter, the packetised effective capacity model with a leaky bucket (see Chapter 2, Section 2.4.4.3) was validated via a publicly available Internet traffic trace from a gigabit Ethernet system. The trace was downloaded from UMASS trace repository and it contains the information of packet arrival instances and packet length sizes.

The distribution fitting technique was adopted and its detailed steps were explained in Section 3.3. The traffic trace was first divided into twelve groups for the purpose of testing stationarity. In Section 3.3.2, the Lindley equation was applied to obtain twelve sequences of end-to-end packet delay from the trace. The basic analysis of the trace was carried out in Section 3.3.3, including twelve packet size histograms. The results show that the majority of packets have either extremely small or large packet sizes, which is a typical pattern of the Internet traffic. In Section 3.3.4, the procedure of obtaining fitted CCDF based on the empirical CCDF and the least square approach was described.

In Section 3.4, twelve empirical and fitted CCDFs of end-to-end packet delay were plotted. It was found that tails of the empirical CCDFs of packet queueing delay are exponentially bounded, validating the packetised effective capacity model with a leaky bucket. Furthermore, as shown in Tables 3.1 and 3.2, the end-to-end delay process from the trace may be wide-sense stationary because across all twelve groups,

1. the average end-to-end delay values and jitter values fluctuate slightly,

2. the estimated values of θ_{lb} are fairly consistent.

Finally, the conclusion drawn in this chapter is based on a trace from a wired network. However, the empirical results presented in chapter may suggest that tail distributions of packet delay in wireless networks will be exponentially distributed.

Chapter 4

Multi-hop Effective Capacity Model for Multi-hop Wireless Networks

4.1 Introduction

The modelling techniques mentioned in Chapter 2, Section 2.4 is based on single-hop queueing systems/communication systems; the effective capacity model was validated in Chapter 3 for the Internet traffic over gigabit Ethernet. However, a multi-hop wireless network can be modelled as a series of queues.

The modelling of end-to-end packet delay in multi-hop systems is complicated by many interrelated factors that jointly determine the final results, and these factors includes 1) hop numbers, 2) the arrival distribution at each queue, 3) the service distribution at each queue, and 4) the dependence between nodes [97].

Conventionally, the studies of end-to-end packet delay in multi-hop networks have been carried out by using the classic queueing theory. In the literature, queue-length behaviours in queueing networks are extensively studied on the basis of Jackson networks [98, 99]. Methods of calculating average end-to-end delay in queueing networks are presented in [100, 101]. However, due to the high complexity of queues behaviour in different nodes within wireless multi-hop networks, the realistic analysis of delay performance using the classic queueing theory becomes intractable [38]. For this reason, only simplified cases are considered, such as one-dimension-and-linear networks [102] and average end-to-end packet delay [101].

Apart from the classic queueing theory, the network calculus theory is a recent-developing theory, providing deep insights into flow problems, and the properties of delay and buffer dimensioning associated with networking [37, 103]. The theory mostly considers bounded sources and services so end-to-end delay characterisations are too loose to be useful [104, 105, 43, 106]. Furthermore, the effects of hop numbers are studied in [107, 108, 109].

In this chapter, the Effective Bandwidth-Effective Capacity (EB-EC) model (see Chapter 2, Section 2.4.4.2) is extended to model the end-to-end delay distributions, average delay and jitter in multi-hop wireless networks. Most results discussed in this chapter have been published in the GLOBECOM 2010 conference proceedings [110].

The rest of this chapter is organised as follows: Section 3.2 introduces the system model of a multi-hop wireless network. The multi-hop effective capacity model and analysis of multi-hop delay performance are given in Section 4.3. Analytical and simulation results are then compared and discussed in Section 4.5. Finally, Section 4.6 concludes the chapter.

4.2 System Model

A system model with a three-hop routing path is shown in Fig. 4.1. Each node can be modelled as a queueing system (the box at the bottom of Fig. 4.1). To make the system model more like a routing path, rather than the one-dimension-and-linear routing path, two traffic correlation indices, p_a and p_d are introduced. For example, arrival traffic of each node is comprised of two parts: one is from its previous node and the other is from its local area or other sources that are outside the routing path. Therefore, the

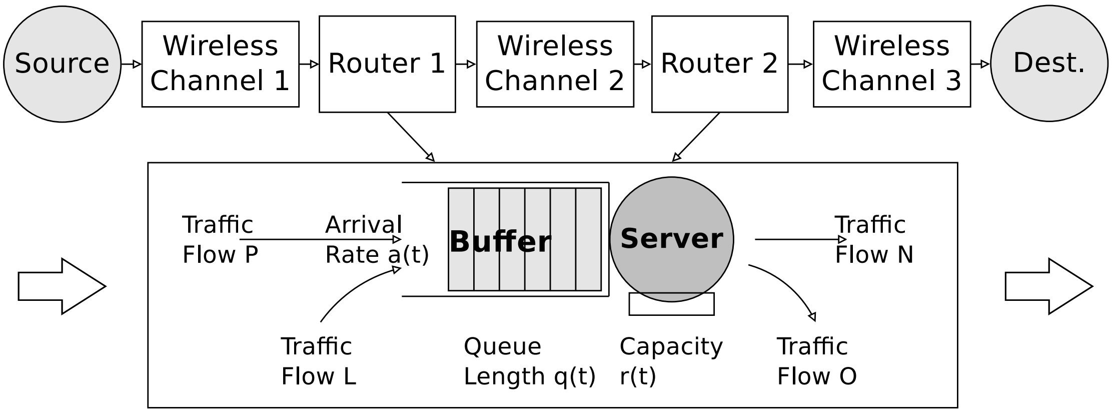


Figure 4.1: A 3-hop System Model

arrival traffic correlation index p_a is defined as

$$p_a = \frac{\text{Traffic Flow P}}{\text{Traffic Flow P + Traffic Flow L}} \tag{4.1}$$

Similarly, traffic inside the buffer of each node has two destinations: one is the next node and the other is outside the routing path. The departure traffic correlation index p_d is defined as

$$p_d = \frac{\text{Traffic Flow N}}{\text{Traffic Flow N + Traffic Flow O}} \tag{4.2}$$

Finally, packets from different sources are mixed and buffered in a single buffer and are served based on the First-In First-Out (FIFO) policy. The buffer size in each node is assumed to be infinite. Each node uses dedicated frequency bands to transmit and receive signals so the channel collision is eliminated in our model.

4.3 Multi-Hop Effective Capacity Model

4.3.1 Effective Bandwidth-Effective Capacity Model

Since the arrival rate is not the constant, the EB-EC model (see Chapter 2, Section 2.4.4.2) is chosen. The EB-EC model shows that when the queueing system reaches its steady state, the Complementary Cumulative Density Function (CCDF) of the packet delay $D(\infty)$ can be approximated by

$$\Pr(D(\infty) \geq D_{\max}) \approx \gamma_{eb} e^{-\theta_{eb} D_{\max}}, \tag{4.3}$$

where γ_{eb} and θ_{eb} are probability of non-empty buffer and the QoS exponent of a connection.

4.3.2 Multi-hop Effective Capacity Model

For the H-hop scenario, suppose that wireless nodes are ordered by the sequence in which the packet traverses along the H-hop path and are numbered from 1 to H. The following lemma is an instant result:

Lemma 4.3.1 *The Probability Density Function (PDF) of a packet delay $D_i(\infty)$ at the i^{th} node is*

$$f_i(t) = \gamma_i \theta_i \exp(-\theta_i t) + (1 - \gamma_i)\delta(t), \{i \in \mathbb{N} : i \leq H\}, \tag{4.4}$$

where γ_i and θ_i are the probability of non-empty buffer and the QoS exponent of a connection at the i^{th} node.

Proof At the i^{th} node, the probability of $D_i(t)$ exceeding a delay bound $D_{\max}$ satisfies

$$\Pr\{D_i(t) > D_{\max}\} \approx \gamma_i e^{-\theta_i D_{\max}} \tag{4.5}$$

The Cumulative Density Function (CDF) of a packet delay is

$$\Pr\{D_i(t) \leq D_{\max}\} = 1 - \Pr\{D_i(t) > D_{\max}\} \tag{4.6}$$

So by definition, the PDF of a packet delay at the i^{rmth} node is given by

$$f_i(t) = \frac{d(1 - \gamma_i \exp(-\theta_i x))}{dx} \tag{4.7}$$

$$= \gamma_i \theta_i \exp(-\theta_i t) + (1 - \gamma_i)\delta(t) \tag{4.8}$$

■

Assume that packet delay values in each node are independent. By using Lemma 4.3.1, the following proposition is given to characterise end-to-end delay distribution:

Proposition 4.3.2 *The H-hop CCDF is estimated by*

$$\Pr(\sum_{i=1}^{H} D_i > x) = 1 - \int_0^x f_1(t) * f_2(t) ... * f_i(t) dt \tag{4.9}$$

where "" stands for convolution, and $f_i(t)$ is the PDF of node h.*

For a proof of Proposition 4.3.2, see Appendix A.1.

Since proposition 4.3.2 involves multiple integrals, making delay bound violation probability uneasy to compute numerically, the proposition below introduces a discrete-form equation which speeds up the calculation.

Proposition 4.3.3 *The discrete-form equation is the same as (4.9) in Proposition 4.3.2 when*

$$\forall i, j \in \{i, j \in \mathbb{N} : i, j \leq H\}, if i \neq j, then \theta_i \neq \theta_j$$

, and is

$$\Pr(\sum_{i=1}^{H} D_i > x) = \sum_{i=1}^{H} \left[\prod_{j=1, j \neq i}^{H} (1 + \frac{\gamma_j \theta_i}{\theta_j - \theta_i}) \right] \gamma_i e^{-\theta_i x} \tag{4.10}$$

For a proof of Proposition 4.3.3, see Appendix A.2.

Note that in the case when $i \neq j$, but $\theta_i = \theta_j$ (means two distributions are identical), we could easily mitigate the problem by slightly changing either the value of θ_i or the

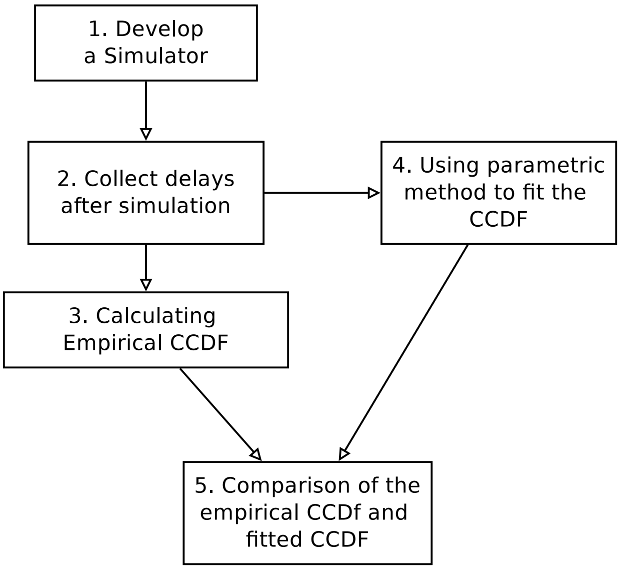

1. Develop a Simulator
2. Collect delays after simulation
3. Calculating Empirical CCDF
4. Using parametric method to fit the CCDF
5. Comparison of the empirical CCDf and fitted CCDF

Figure 4.2: Steps of validating the effective capacity model

value of θ_j.

Finally, the average delay and jitter are derived as:

Corollary 4.3.4 *The average delay and jitter can be expressed as*

$$E[D] = E\left[\sum_{i=1}^{H} D_i\right] = \sum_{i=1}^{H} \frac{\gamma_i}{\theta_i} \tag{4.11}$$

$$\sigma = \sqrt{\mathrm{Var}\left(\sum_{i=1}^{H} D_i\right)} \tag{4.12}$$

$$= \sqrt{\sum_{i=1}^{H}\left(\frac{2\gamma_i}{\theta_i^2} - \left(\frac{\gamma_i}{\theta_i}\right)^2\right)} \tag{4.13}$$

For a proof of Corollary 4.3.4, see Appendix A.3.

4.4 Method and Simulation Platform

The *distribution fitting method* will be used to validate the multi-hop effective capacity model. The detailed steps of this method is shown in Fig. 4.2.

Steps 1, 2 and 3 are the procedure for plotting empirical distribution. Step 1

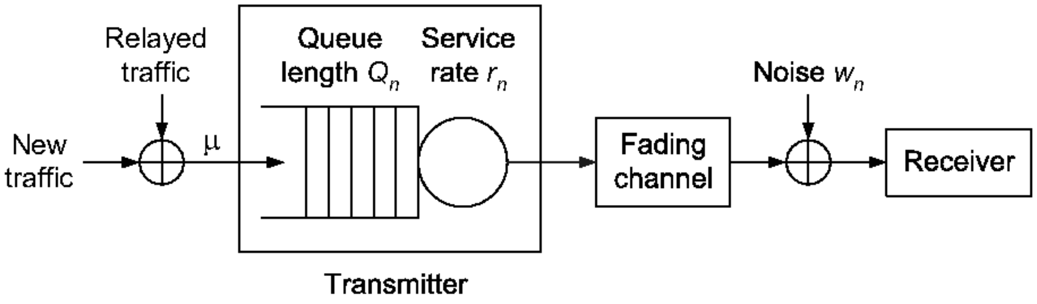


Figure 4.3: System model for point-to-point packet transmissions

(Sections 4.4.1 and 4.4.2) involves simulator development with the simulator being able to record end-to-end packet delay. Steps 2 and 3 are to plot empirical distributions from the collected packet delays from the simulator. In step 4 (Section 4.4.3), the parametric method is use to generate the fitted CCDFs of packet delays. The empirical CCDFs and fitted CCDFs are compared. The fitted distribution giving a close fit is supposed to lead to good predictions.

4.4.1 Simulation Platform

MATLAB is used to develop the simulation platform based on the system model of Fig. 4.1 and the Monte-Carlo method to evaluate the delay performance over a wireless 3-hop path. Furthermore, the arrival traffic correlation index is assumed to equal the departure traffic correlation index, i.e., $p_a = p_d = p$.

The source node generates fixed-size packets at a constant rate towards different destinations, and other nodes generate packets (at every time slot, the number of packets generated is a random variable and has the Bernoulli distribution) to imitate the process of traffic coming from the node themselves and other sources outside the network. Each node gets packets destined to itself, and the destination node collects packets from the source node and analyses the CCDF of delay, average delay and jitter.

The system model at each node is shown in Fig. 4.3 (the system model is identical to the simulation platform in [4]). The average Signal-to-Noise Ratio (SNR) is fixed in each simulation run and the instantaneous channel gain $g[n]$ is perfectly known at the transmitter side. The transmission rate $r[n]$ at each time slot could be approximated by the Shannon capacity:

$$r[n] \approx B \log_2(1 + |g[n]|^2 \times \mathrm{SNR}_{\mathrm{avg}}) \tag{4.14}$$

By using sophisticated modulation and coding schemes, it is possible to approach the Shannon capacity of a wireless channel [111, 112].

The equivalent AWGN channel capacity is defined as,

$$r_{\mathrm{AWGN}} = B \log_2(1 + \mathrm{SNR}_{\mathrm{avg}}) \tag{4.15}$$

$$\Rightarrow B = \frac{r_{\mathrm{AWGN}}}{\log_2(1 + \mathrm{SNR}_{\mathrm{avg}})} \tag{4.16}$$

Thus, by substituting B in (4.14) with (4.15), we have

$$r[n] = \frac{r_{\mathrm{AWGN}} \log_2(1 + |g[n]|^2 \times \mathrm{SNR}_{\mathrm{avg}})}{\log_2(1 + \mathrm{SNR}_{\mathrm{avg}})} \tag{4.17}$$

The first-order auto-regressive (AR(1)) model is used to generate correlated random channel gain ($g[n]$) sequence as the Rayleigh fading channel,

$$g[n] = \kappa \times g_{n-1} + \nu[n], \tag{4.18}$$

where $\nu[n]$ is the complex random variable, having the normal distribution $N(0, 1)$ in real and imaginary parts. The coefficient κ is determined by [4]

$$\kappa = 0.5^{T_s/T_c}, \tag{4.19}$$

where T_s is the sampling rate, and T_c is the coherent time that is a function of maximum Doppler rate f_m.

Table 4.1: Simulation parameters

Parameters	**Values**
Channel Model	Rayleigh Distr.
Average SNR, $\mathrm{SNR}_{\mathrm{avg}}$	5 dB and 15dB
AWGN channel capacity, c_{AWGN}	100kbps
Maximum Doppler rate, f_m	30Hz
Traffic Load, μ	75 and 85 kbps
Time Slot, T_s	$1/\mu$
Traffic Correlation	0.25 and 1
Hop Number, H	3

Assume the radio ray hitting the mobile at an angle β. The vehicle motion with respect to the incoming ray introduces a Doppler frequency shift [113]

$$f_m = \frac{v \cos \beta}{\lambda} \tag{4.20}$$

Considering $\lambda = c/f_0$, (4.20) becomes

$$v = \frac{f_m}{f_0 \cos \beta} \cdot c \tag{4.21}$$

4.4.2 Simulation Settings

For simulation purpose, we use different traffic loads (light traffic load (μ = 75kbps) and heavy traffic load (μ = 85kbps)), traffic correlation (p = 0.25 and p = 0.75) and Signal-to-Noise Ratio (low SNR (SNR = 5dB) and high SNR (SNR = 15dB)) cases to evaluate the performance. The maximum Doppler rate is set to be 30Hz (if the centre frequency is 2.4 GHz (e.g. devices using ISM bands are mostly operating at 2.45 GHz), the speed of the node is 3.75m/s). We take 100kbps as the channel capacity under AWGN channel condition [114]. Table 4.1 lists simulation parameters used in the chapter.

4.4.3 Multi-hop Effective Capacity-based Estimator

On the basis of (4.10), (4.11) and (4.12), CCDF of delay, average delay and jitter can be estimated if the estimated values of γ_i and θ_i in each hops are known. Suppose each node is observed every time slot (the duration of a slot is T_s seconds) for a total N slots over a period of (NT_s) seconds. For the node i, three quantities are recorded at the $n^{\text{th}}(n \geq 1)$ slot boundary: the indicator of whether the buffer is non-empty $\mathbf{1}\{Q_i[n] > 0\}$ (1 if the buffer is non-empty and 0 otherwise), the queue length $Q_i[n]$ and the number of bits arrived $A'_i[n]$. After the observation, we have one *realisation* of an indicator process $\{\mathbf{1}\{Q_i[n] > 0\}, n \geq 1\}$, one of a queue length process $\{Q_i[n], n \geq 1\}$ and one of an arrival process, each of which have N samples, i.e.,

$$\mathbf{1}\{\mathbf{q}_i > 0\} = \{\mathbf{1}\{q_i[1] > 0\}, \mathbf{1}\{q_i[2] > 0\}, \mathbf{1}\{q_i[3] > 0\}, \cdots, \mathbf{1}\{q_i[N] > 0\}\} \tag{4.22}$$

$$\mathbf{q}_i = \{q_i[1], q_i[2], q_i[3], \cdots, q_i[N]\} \tag{4.23}$$

$$\mathbf{a}'_i = \{a'_i[1], a'_i[2], a'_i[3], \cdots, a'_i[N]\} \tag{4.24}$$

In the end, the same estimation technique in [4] is applied to estimate γ_i and θ_i:

$$\hat{\mu} = \frac{1}{NT_s} \sum_{j=1}^{N} a'_i[j] \tag{4.25}$$

$$\hat{\gamma}_i = \frac{1}{N} \sum_{j=1}^{N} 1\{q_i[j] > 0\} \tag{4.26}$$

$$\hat{\theta}_i = \frac{\hat{\gamma} \cdot \hat{\mu} \times N}{\sum_{j=1}^{N}(q_i[j])} \tag{4.27}$$

and CCDF of delay, average delay and jitter can be estimated by

$$\Pr \hat{D} > x) = \sum_{i=1}^{H} \left[\prod_{j=1, j \neq i}^{H} (1 + \frac{\hat{\gamma}_j \hat{\theta}_i}{\hat{\theta}_j - \hat{\theta}_i}) \right] \hat{\gamma}_i e^{-\hat{\theta}_i x} \tag{4.28}$$

$$\widehat{E[D]} = \sum_{i=1}^{H} \frac{\hat{\gamma}_i}{\hat{\theta}_i} \tag{4.29}$$

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{H} \left(\frac{2\hat{\gamma}_i}{\hat{\theta}_i^2} - \left(\frac{\hat{\gamma}_i}{\hat{\theta}_i} \right)^2 \right)} \tag{4.30}$$

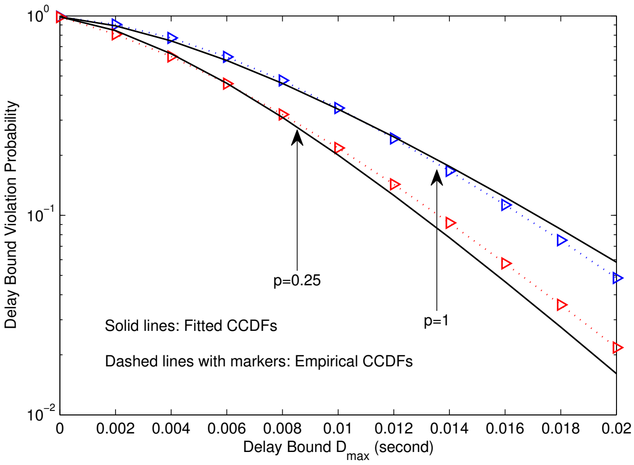
4.5 Results and Discussion

Fig. 4.4 shows the simulation results and estimation results based on Chapter 4.4.3 under various conditions. The X-coordinates are delay bounds (the unit is milli seconds), and the Y-coordinates are delay bound violation probabilities. The analytical and simulation results are shown in solid lines and in dashed lines with markers, respectively. Moreover, results of different traffic correlations are pointed out by upper arrows with traffic correlation indices underneath.

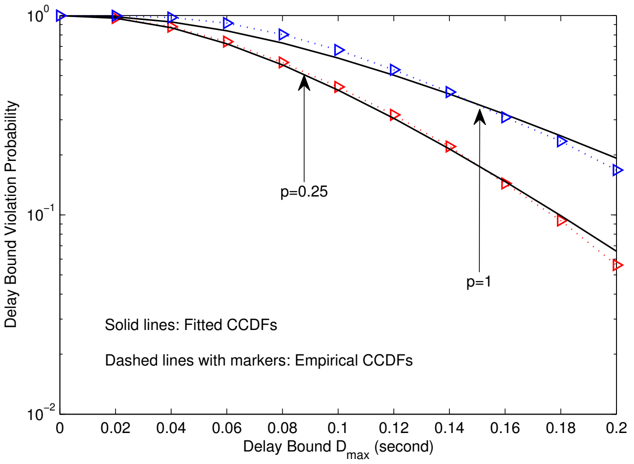
As seen from Fig. 4.4, the estimation results based on Chapter 4.4.3 give accurate estimates of CCDF of packets in all simulation scenarios, indicating that the multi-hop effective capacity model is suitable for modelling end-to-end packet delay in wireless mesh networks. Table 4.2 lists the estimated γ_h and θ_h at each node.

Table 4.3 lists the average delay and jitter from simulation results and estimation results from Chapter 4.4.3. It can also be further concluded that the estimation algorithm developed in this chapter performs well under different conditions.

In another perspective, as partially mentioned in [114], traffic loads and traffic correlation indices and SNR affect the delay performance. The answer to the performance being degraded by increasing traffic load is quite intuitive, and it is because in the same time period, more packets arrives at the system when the traffic load condition is heavier, leading to the severer queueing delay if the service rate is unchanged (as we have in our case). The reason for the degraded performance by decreasing SNR is intuitive



(a) Light Traffic Load ($\mu = 75$kbps)



(b) Heavy Traffic Load ($\mu = 85$kbps)

Figure 4.4: Empirical and fitted CCDFs of end-to-end packet delay

as well. It is because in the same time period, less packets are served, leading to the severer queueing delay. The reason for the effect of traffic correlation (although have not been proved yet) may be attributed to the input traffic pattern. In other words, more burstiness of input traffic (by making traffic correlation indices r bigger) would more easily turn its buffer into idle state and more easily waste the service, which in turn worsens the delay performance.

Table 4.2: Estimated γ_h and θ_h from the multi-hop effective capacity-based estimator

Wireless Situation		**3-hop effective capacity model**					
Traffic load (kbps)	r	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
75	0.25	0.69	0.81	0.81	317.92	370.86	371.70
75	1	0.69	0.80	0.83	323.36	266.00	234.76
85	0.25	0.96	0.98	0.98	32.38	27.39	28.44
85	1	0.96	0.98	0.98	32.38	19.53	17.35

Table 4.3: Descriptive statistics of end-to-end delays

Wireless Situation		**Average delay (ms)**		**Jitter (ms)**	
Traffic load (kbps)	r	Sim.	Est.	Sim.	Est.
75	0.25	6.6	6.5	5.2	4.8
75	1	8.7	8.7	5.9	6.3
85	0.25	99.5	100	55.2	59.3
85	1	136.9	136.8	64.9	83

4.6 Conclusions

In this chapter, the multi-hop effective capacity model was developed based on the single-hop EB-EC model to characterise delay performances in multi-hop wireless networks. Specifically, CCDF of a packet delay, average delay and jitter over multi-hop wireless path were derived in Section 4.3.

The distribution fitting technique was adopted and its detailed steps were explained in Section 4.4. The procedure of building a simulation platform was described in Section 4.4.3. An estimator based on this multi-hop effective capacity model was developed in Section 4.4.3.

The simulation and estimation results were shown in Section 4.5. Impacts of delay performance from different traffic correlations, traffic loads and SNR cases were also considered and discussed. The results showed that our estimator provides good estimates of CCDF of packets, average delay, and jitter, and gave a key insight into the QoS provisioning for multi-hop wireless networks.

Chapter 5

Mixed Continuous/Discrete-Time Effective Capacity Model for Wireless Slotted Communication Systems

5.1 Introduction

The conventional effective capacity model (see Chapter 2, Section 2.4.4.1) is based on continuous-time system models. The model 1) shows tail distributions of packet queueing delay to be *exponentially* bounded and 2) approximates the Complementary Cumulative Density Function (CCDF) of packet queueing delay by two functions, namely the probability of non-empty buffer and the QoS exponent of a connection.

Unlike Ethernet systems described in Chapter 3, some real communication systems are *time-slotted*, for example, frames in Universal Mobile Telecommunications Systems (UMTSs) or High Speed Packet Access (HSPA) systems are transmitted every Transmission Time Interval (TTI). Such systems may be modelled as a discrete-time queueing model [115, 116]. Studies of delay distribution approximations in discrete-time queueing models have conventionally been carried out by using the classic queueing theory [117, 118, 119, 120, 121]. Recent research includes the Bayesian inference for a $Geo/G/1$ discrete-time queue [122, 123] and the maximum likelihood inference for discretely observed Markov jump processes [124]. These approaches are *accurate* in estimating delay distributions but are rather computationally *inefficient*.

The aim of this chapter is to re-investigate the effective capacity model for slotted systems. In this chapter, Mixed Continuous/Discrete-Time (MCDT) models are used and such models are discrete-time system models with consideration of slot time (this modelling technique is used in [125]). In discrete-time queueing models, the possible values of packet queueing delay can only be non-negative integers. Similarly, the possible values of packet queueing delay in MCDT models can only be non-negative integers multiplied by slot time. Consequently, tail distributions of packet queueing delay in mixed continuous/discrete time models is formally shown to be *geometrically* bounded. A mixed continuous/discrete time effective capacity model is proposed to approximate the CCDF of packet queueing delay by two functions, namely the probability of non-empty buffer and the success probability of a connection. The mathematical formulae of average packet delay and jitter are also derived in this chapter.

Finally, we develop a discrete-time simulation platform and conduct a set of simulation experiments. The empirical CCDFs of packet queueing delay , average delay and jitter are compared with that from the conventional and revised effective capacity models. The results indicate that the revised MCDT model, unlike the conventional effective capacity model, always give close estimations for different channel models. Most results discussed in this chapter were submitted to the IEEE transaction on wireless communications [121, 126, 127].

The remainder of this paper is organised as follows: Section 5.2 introduces modelling of wireless slotted communication systems as MCDT queueing models. In Section 5.3, the conventional effective capacity model is adapted to the MCDT effective capacity model. Section 8.2.1.1 discusses the MCDT Effective Bandwidth-Effective Capacity

Table 5.1: Notations and definitions

Notation for MCDT models	**Notation for CT models**	**Definition**
$A(t;T_s)$	$A(t)$	cumulative bit arrival over the time interval $[0,t)$
$A'(n;T_s)$	$A'(t)$	bit arrival at time t or during slot n
$\alpha^{(c)}(\theta;T_s)$	$\alpha^{(c)}(\theta)$	Effective Capacity function
$D(n;T_s)$	$D(t)$	delay experienced by the packet arrives at time t or at the beginning of slot n
$\gamma^{(c)}(\mu;T_s)/\gamma_m$	$\gamma^{(c)}(\mu)/\gamma$	probability of non-empty buffer
$K_{\max}$	$D_{\max}$	delay bound
k		normalised delay bound
$\Psi(t;T_s)$	$\Psi(t)$	service characterisation curve
$p^{(c)}(\mu;T_s)$ or p_m		success probability of a connection
$Q(t;T_s)$	$Q(t)$	queue length at time t
$S(t;T_s)$	$S(t)$	cumulative actual bit service over the time interval $[0,t)$
$S'(n;T_s)$	$S'(t)$	actual bit service
$\sigma_m^{(c)}$	$\sigma^{(c)}$	delay error term of a service characterisation curve
T_s		slot time
$\tilde{S}(t;T_s)$	$\tilde{S}(t)$	cumulative bit service over the time interval $[0,t)$
$\tilde{S}'(n;T_s)$	$\tilde{S}'(t)$	bit service at time t or during slot n
$\theta^{(c)}(\mu;T_s)$	$\theta^{(c)}(\mu)/\theta$	QoS exponent of a connection
y_n		$(n+1)^{\text{st}}$ slot boundary

(EB-EC) model and delay distribution characterisation. Analytical and simulation results are then compared and discussed in Section 5.5. Finally, Section 4.6 concludes the chapter. The notations for MCDT models and their counterparts for continuous-time (CT) models are listed in Table 5.1 in alphabetical order.

5.2 System Model

Fig. 5.1 shows the data link layer with interfaces to adjacent layers of a transmitter/receiver pair of a wireless communication system. On the transmitter side, the packet generator in the network layer generates packets and pushes them to the queue in the data link layer. The physical interface retrieves packets from the queue and sends them to the receiver over the wireless channel. The transmitter part can be represented by a queueing model, which is shown in Fig. 7.2. The queue size is assumed to be infinite and the queue discipline is First-In First-Out (FIFO). If the queueing model is

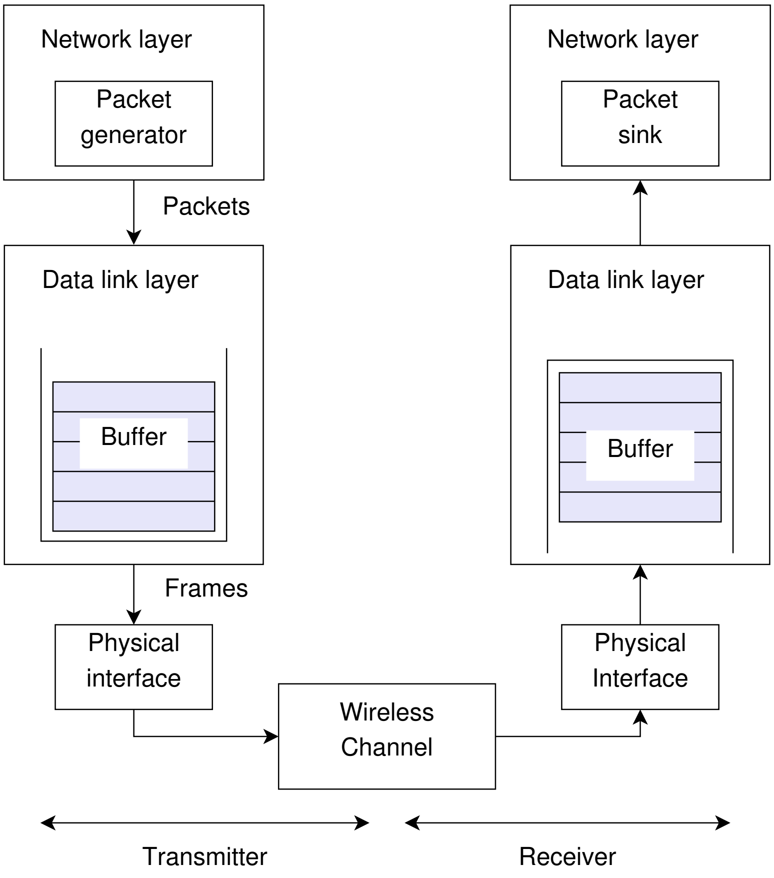


Figure 5.1: A transmitter/receiver pair of a wireless communication system

continuous-time, packet arrivals and departures can take place at any time. An example is shown in Fig. 5.3.

When the system of Fig. 5.1 is time-slotted (the transmitter/receiver pair operates every time interval), the transmitter part is better described as a discrete-time queueing model. We take slot 1 as the first slot and define the n^{th} slot boundary y_n as the beginning of slot $(n + 1)$ so y_0 is the first slot boundary. A discrete-time queueing model operates at every slot boundary. By the convention for discrete-time queueing models [28], packet arrivals occur after a slot boundary while departures take place before a slot boundary. Fig. 5.4 shows an example.

In this chapter, we use a *mixed continuous/discrete time queueing model.* Such a model is an extension of the discrete-time queueing model, in which all events, including

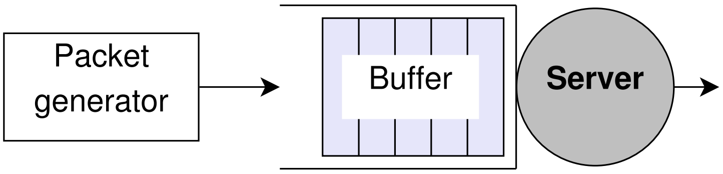


Figure 5.2: A queueing model

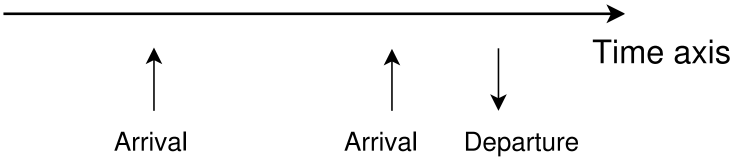


Figure 5.3: Packet arrivals and departures in a continuous-time queueing model

arrivals and departures, are measured in *time units* rather than *slots*. In detail, the extension is achieved by using the slot time T_s defined as the duration of a slot with units of seconds. As direct results, slot n is the time interval $[(n-1)T_s, nT_s)$ and the slot boundary y_n is equivalent to nT_s second.

Finally, uncertainties of the wireless channel between the transmitter and receiver reduce transmission reliability. We assume an ideal transmission, i.e., the instantaneous channel capacity is based on Shannon's channel capacity. We follow [4] and further assume a fluid traffic model (in a fluid traffic model packet lengths are infinitesimally small) of a constant bit arrival rate μ.

5.3 Mixed Continuous/Discrete-Time (MCDT) Effective Capacity Model

The continuous-time queueing models and mixed continuous/discrete time queueing models have differences and commonalities. The end-to-end delay and queue length in these two models have different attributes. They are discussed in Section 5.3.1 and their differences are the main motivation for developing a new model based on mixed continuous/discrete time queueing models. Section 5.3.2 illustrates a common attribute between both models, i.e., end-to-end delay characterisation, which bridges

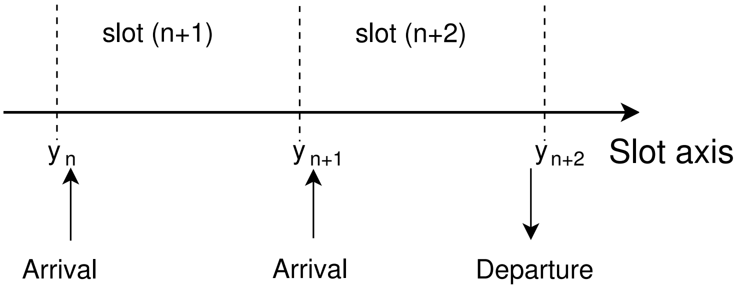


Figure 5.4: Packet arrivals and departures in a discrete-time queueing model

the gap between the conventional effective capacity model and the new model. In the last section (Section 5.3.3), we propose our mixed continuous/discrete-time effective capacity model.

5.3.1 End-To-End Delay and Queue Length in Continuous-Time Queueing Models

In a continuous-time queueing model, $D(t)$ and $Q(t)$ are denoted as the end-to-end delay of the packet arrives at time t and the queue length at time t, respectively. Their values are closely related to the following stochastic processes:

1. arrival process $\{A'(t), t \geq 0\}$: $A'(t)$ is the bit arrival rate at time t (since we assume a constant bit arrival rate in Section 5.2, $A'(t)$ equals μ),

2. service process $\{\tilde{S}'(t), t \geq 0\}$: $\tilde{S}'(t)$ is the bit service rate (the bit rate that the server is capable to serve) at time t,

3. actual service process $\{S'(t), t \geq 0\}$: $S'(t)$ is the actual bit service rate (the actual bit rate that is actually served by the server) at time t,

4. cumulative arrival process $\{A(t), t \geq 0\}$: $A(t)$ is the total number of bits arrived over the time interval $[0, t)$, i.e., $A(t) = \int_0^t A'(\tau)d\tau$,

5. cumulative service process $\{\tilde{S}(t), t \geq 0\}$: $\tilde{S}(t)$ is the number of bits that the server is capable to serve over the time interval $[0, t)$, i.e., $\tilde{S}(t) = \int_0^t \tilde{S}'(\tau)d\tau$,

6. cumulative actual service process $\{S(t), t \geq 0\}$: $S(t)$ is the number of bits that is actually served by the server over the time interval $[0, t)$, i.e., $S(t) = \int_0^t S'(\tau)d\tau$.

As shown graphically in Fig. 5.5, $D(t)$ is the horizontal difference between $\{A(t), t \geq 0\}$ and $\{S(t), t \geq 0\}$ and its possible values can be any non-negative real number:

$$D(t) \in [0, \infty) \tag{5.1}$$

$Q(t)$ is the vertical difference.

For a mixed continuous/discrete queueing model, we use $D(n; T_s)$ and $Q(t; T_s)$ to denote the end-to-end delay of the packet arrives just after the beginning of slot n and the queue length at time t, respectively. We also have

1. arrival process $\{A'(n; T_s), n \geq 1\}$: $A'(n; T_s)$ is the number of bits arrived during slot n and is a constant because of the assumption of the constant bit arrival rate,

2. service process $\{\tilde{S}'(n; T_s), n \geq 1\}$: $\tilde{S}'(n; T_s)$ is the number of bits that the server is capable to serve during slot n,

3. actual service process $\{S'(n; T_s), n \geq 1\}$: $S'(n; T_s)$ is the number of bits that is actually served by the server during slot n.

The cumulative arrival process $\{A(t; T_s), t \geq 0\}$, cumulative bit service process $\{\tilde{S}(t; T_s), t \geq 0\}$ and cumulative actual bit service process $\{S(t; T_s), t \geq 0\}$ have the same definitions as $\{A(t), t \geq 0\}$, $\{\tilde{S}(t), t \geq 0\}$ and $\{S(t), t \geq 0\}$.

As shown in Fig. 5.6, $D(n; T_s)$ is still the horizontal difference between $\{A(nT_s), n \geq 0\}$ and $\{S(nT_s), n \geq 0\}$; $Q(t)$ is still the vertical difference and its value changes only at slot boundaries.

In Fig. 5.6, the realisations of a cumulative bit arrival process and a cumulative actual bit service process are *staircase* functions. Therefore, the possible values of

$D(n;T_s)$ can be any non-negative integers multiplied by T_s:

$$D(n;T_s) \in \{0, T_s, 2T_s, 3T_s, \cdots\} \tag{5.2}$$

The change of the possible values of end-to-end delay (from (5.1) to (5.2)) motivates us to re-investigate the conventional effective capacity model.

5.3.2 End-to-End Delay Characterisation

According to [4], for a given $0 < \epsilon \leq 1$, the end-to-end delay characterisation is defined as the probability bound on a delay bound $D_{\max}$ satisfies

$$\sup_t \Pr\{D(t) > D_{\max}\} \leq \epsilon \tag{5.3}$$

where $\sup_t$ is the least upper bound of a set. Since $D(t)$ is the horizontal difference between $A(t)$ and $S(t)$, and $A'(t)$ is a constant, the end-to-end delay characterisation is associated with the service characterisation.

A service characterisation curve is defined as

$$\Psi(t) = (\mu(t - \sigma^{(c)}))^+ \tag{5.4}$$

where $\sigma^{(c)}$ is the delay error term [4]. During a busy period (the buffer is non-empty), if the beginning of the period is set as time 0 and let $\sigma^{(c)}$ equal the delay bound $D_{\max}$, we have

$$\sup_t \Pr\{S(t) < \Psi(t)\} = \sup_t \Pr\{\tilde{S}(t) < \Psi(t)\} \tag{5.5}$$

$$= \sup_t \Pr\{D(t) > D_{\max}\} \tag{5.6}$$

The first equation holds because when the queue is busy, the bit service process is the same as the actual bit service process. The second equation holds because both the events of $\{D(t) > D_{\max}\}$ and events of $\{S(t) < \Psi(t)\}$ indicate the shaded area of Fig. 5.5, in other words, they are the same events.

Similarly, for a given $0 < \epsilon \leq 1$, the end-to-end delay characterisation in mixed continuous/discrete time queueing models is defined as the probability bound on a delay bound $K_{\max}$ satisfies

$$\sup_{n} \Pr\{D(n;T_s) > K_{\max}\} \leq \epsilon \tag{5.7}$$

Since $\{0, T_s, 2T_s, \cdots\}$ are the possible values of $D(n;T_s)$, they should also be the possible values that delay bounds $K_{\max}$ can choose from.

A new service characterisation curve for mixed continuous/discrete time queueing models is defined as

$$\Psi(t;T_s) = (\mu(T_s\lfloor t/T_s \rfloor - \sigma_m^{(c)}))^{+} \tag{5.8}$$

$\Psi(t;T_s)$ is a staircase function.

During a busy period in a mixed continuous/discrete time queueing model (shown in Fig. 5.6), if the beginning of the period is set as time 0 and let $\sigma_m^{(c)}$ equal the delay bound $K_{\max}$, the following equations are can be derived

$$\begin{aligned} &\sup_{nT_s} \Pr\{S(nT_s;T_s) < \Psi(nT_s;T_s)\} && (5.9)\\ &= \sup_{nT_s} \Pr\{\tilde{S}(nT_s;T_s) < \Psi(nT_s;T_s)\} && (5.10)\\ &= \sup_{nT_s} \Pr\{D(n;T_s) > K_{\max}\} && (5.11) \end{aligned}$$

using the same reasons for (5.6).

The consistency between (5.6) and (5.11) suggests the eligibility of adapting the conventional effective capacity model for mixed continuous/discrete time queueing models.

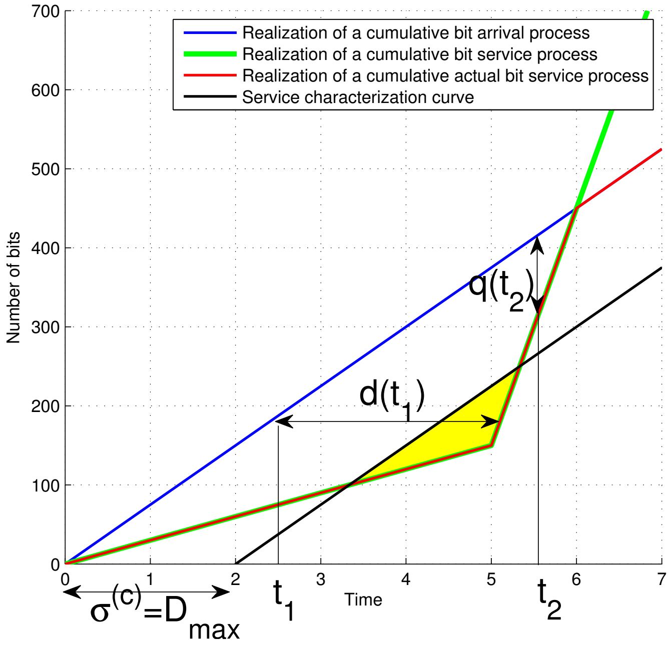


Figure 5.5: Graphical interpretations of $D(t)$ and $Q(t)$ in a continuous-time queueing model

5.3.3 Mixed Continuous/Discrete-Time Effective Capacity Model

Assume that the bit service process $\{\tilde{S}'(n; T_s), n \geq 0\}$ is stationary and the asymptotic log-moment generating function of $\tilde{S}(t; T_s)$, defined as

$$\Lambda(-\theta; T_s) = \lim_{n\to\infty} \frac{1}{n} \log E[e^{-\theta \tilde{S}(n;T_s)}] \tag{5.12}$$

exists for all $\theta > 0$. The effective capacity function is defined as

$$\alpha^{(c)}(\theta; T_s) = \frac{-\Lambda(-\theta; T_s)}{\theta}, \text{ for all } \theta \tag{5.13}$$

Consider a mixed continuous/discrete time queue supplied by a packet generator of a constant bit arrival rate μ. If there is a unique solution $\theta^{(c)}(\mu; T_s)$ of the equation

$$\alpha^{(c)}(\theta; T_s) = \mu \tag{5.14}$$

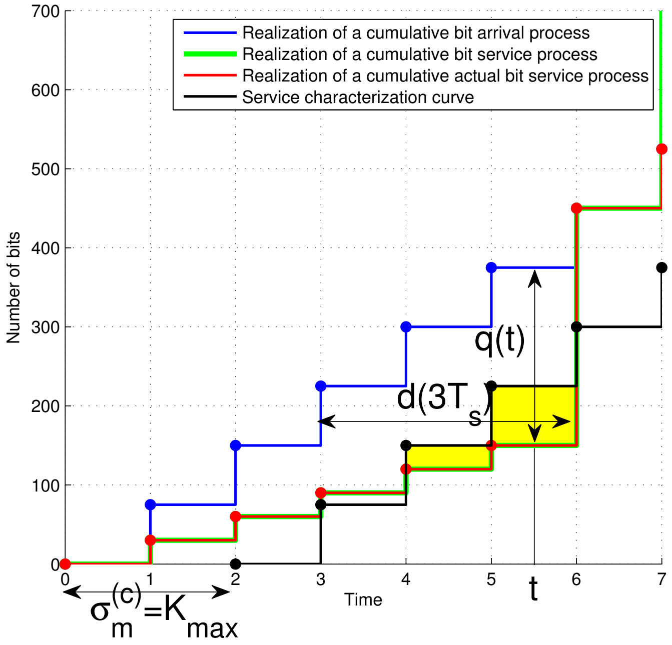


Figure 5.6: Graphical interpretations of $D(n; T_s)$ and $Q(t; T_s)$ in a mixed continuous/discrete time queueing model

the probability of $D(n; T_s)$ exceeding a delay bound $K_{\max}$ satisfies

$$\sup_n \Pr\{D(n; T_s) > K_{\max}\} \approx \gamma^{(c)}(\mu; T_s) e^{-\theta^{(c)}(\mu; T_s) K_{\max}}, \tag{5.15}$$

$$K_{\max} \in \{0, T_s, 2T_s, \cdots\} \tag{5.16}$$

$\gamma^{(c)}(\mu; T_s)$ is again called the *probability of non-empty buffer* because

$$\gamma^{(c)}(\mu; T_s) = \sup_n \Pr\{D(n; T_s) > 0\} \tag{5.17}$$

$$= \sup_t \Pr\{Q(t; T_s) > 0\} \tag{5.18}$$

and $\theta(c)(\mu; T_s)$ is the *QoS exponent of a connection.*

If we substitute $e^{-\theta^{(c)}(\mu; T_s)}$ and $K_{\max}$ in (5.16) with $(1 - p^{(c)}(\mu; T_s))^{1/T_s}$ and kT_s, we have

$$\sup_n \Pr\{D(n;T_s) > kT_s\} \approx \gamma^{(c)}(\mu;T_s)(1 - p^{(c)}(\mu;T_s))^k, \tag{5.19}$$

$$k \in \{0, 1, 2, \cdots\} \tag{5.20}$$

The tail distribution of packet end-to-end delay in (5.19) is geometrically bounded because the distribution of $D(n;T_s)$ given $D(n;T_s) > 0$ is approximately a geometric distribution. So $p^{(c)}(\mu;T_s)$ in (5.19) is called the *success probability of a connection.* $\gamma^{(c)}(\mu;T_s)$ and $p^{(c)}(\mu;T_s)$ are functions of the constant bit arrival rate μ and the slot time T_s. They both define our proposed mixed continuous/discrete time effective capacity model. For the rest of paper, we will use γ_m and θ_m, which are shorthand for $\gamma^{(c)}(\mu;T_s)$ and $p^{(c)}(\mu;T_s)$, respectively.

Finally, the average delay and jitter in a MCDT queueing system are derived as:

Proposition 5.3.1 *The average delay and jitter can be expressed as*

$$\text{average delay} = \mu_D = E[D] = \frac{\gamma_m}{p_m} \cdot T_s \tag{5.21}$$

$$\text{jitter} = \sigma_D = \sqrt{E[D-\mu]^2} = \sqrt{\left(\frac{2(1-p_m)\gamma_m}{p_m^2} + \frac{\gamma_m}{p_m} - \left(\frac{\gamma_m}{p_m}\right)^2\right)} \cdot T_s \tag{5.22}$$

For a proof of Proposition 5.3.1, see Appendix A.4.

5.4 Method and Simulation Platform

The same methodology explained in Chapter 4, Section 4.4 is adopted in this chapter. Furthermore, apart from the multi-hop architecture, the simulation platform is identical to the one in Chapter 4.

Table 5.2: Simulation parameters

Parameter	Value
Average channel capacity (Kbps), r_{AGWN}	100
Average SNR (dB)	15
Channel model	Rayleigh or Rice Distribution
Constant bit arrival rate (Kbps), μ	75
K factor: K	3
Maximum Doppler rate (Hz), f_m	5 or 10
Slot time (ms), T_s	5 or 10
Total packets generated in each simulation	1 Million

5.4.1 Simulation Settings

Two different slot times are used: 1ms (1ms is used in [4]) and 10ms (10ms is one option of TTI in the UMTS standard [128]); two types of fading channels are used: Rayleigh fading channel that represents non-line-of-sight communications and Rician fading channel that represents line-of-sight communications; two different Doppler rates are used: 5Hz and 10Hz (suppose the carrier frequency is 2.4GHz and the angle of arrival is 0 degree. 5Hz and 10Hz maximum Doppler rates correspond to 6.25m/s and 12.5ms/ of node speed, respectively). The other simulation parameters are listed in Table 7.2 in alphabetical order. We have eight simulation runs in total and collect realisations of the delay process and queue length process from each simulation.

5.4.2 MCDT Effective Capacity-based Estimator

Suppose we observe a wireless slotted communication system at every slot boundary for a total N slots over a period of (NT_s) seconds. At the $n^{\text{th}}(n \geq 1)$ slot boundary, we record two quantities: the indicator of whether the buffer is non-empty $\mathbf{1}\{Q(nT_s;T_s) > 0\}$ (1 if the buffer is non-empty and 0 otherwise), the queue length $Q(nT_s;T_s)$. Therefore, we have one *realisation* of an indicator process $\{\mathbf{1}\{Q(nT_s;T_s) > 0\}, n \geq 1\}$ and one of a queue length process $\{Q(nT_s;T_s), n \geq 1\}$, each of which have N samples, i.e.,

$$\mathbf{1}\{\mathbf{q} > 0\} = \{\mathbf{1}\{q[1] > 0\}, \mathbf{1}\{q[2] > 0\}, \mathbf{1}\{q[3] > 0\}, \cdots, \mathbf{1}\{q[n] > 0\}\} \tag{5.23}$$

$$\mathbf{q} = \{q[1], q[2], q[3], \cdots, q[N]\} \tag{5.24}$$

The indicator process and the queue length process are two stochastic processes and are assumed to be stationary and ergodic.

It is shown in [129] that if the bit service process $\{\tilde{S}'(n;T_s), n \geq 1\}$ is stationary and ergodic and $E[\tilde{S}'(n;T_s) > \mu]$, then the queue length process $\{Q(nT_s;T_s), n \geq 1\}$ converges in distribution to a random variable $Q(\infty;T_s)$ when n goes to ∞ and will eventually agree with a stationary and ergodic process. Furthermore, $\mathbf{1}\{Q(nT_s;T_s) > 0\}$ is 1 when $Q(nT_s;T_s) > 0$ and is 0 otherwise. Hence, a stationary and ergodic queue length process suggests a stationary and ergodic indicator process.

In Sections 5.4.2.1 and 5.4.2.2, we adapt the conventional estimator of [4] to estimate γ_m and θ_m of the mixed continuous/discrete time EC model. The computational complexity performances of the conventional estimator and the revised estimator are compared in Section 5.4.2.3.

5.4.2.1 Estimator for γ_m

The following equation is a standard result if an indicator process is stationary and ergodic:

$$E[\mathbf{1}\{Q(T_s;T_s) > 0\}] = E[\mathbf{1}\{Q(nT_s;T_s) > 0\}] \tag{5.25}$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{Q(iT_s;T_s)\} \tag{5.26}$$

The expectation of $\mathbf{1}\{Q(T_s;T_s)>0\}$ is equivalent to $\Pr\{Q(T_s;T_s)>0\}$ because

$$E[\mathbf{1}\{Q(T_s;T_s)>0\}] = \Pr\{Q(T_s;T_s)=0\}\cdot 0 \tag{5.27}$$

$$+\Pr\{Q(T_s;T_s)>0\}\cdot 1 = \Pr\{Q(T_s;T_s)>0\} \tag{5.28}$$

From (5.18), γ_m is unbiasedly estimated by

$$\hat{\gamma}_m = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{q_i>0\} \tag{5.29}$$

5.4.2.2 Estimator for p_m

If a queue length process is stationary and ergodic, the delay process $\{D(n;T_s), n\geq 1\}$ is also stationary and ergodic. Consequently, the CCDF of $D(n;T_s)$ is identical to that of $D(1;T_s)$ and (5.19) becomes

$$\sup_n \Pr\{D(n;T_s)>kT_s\} = \Pr\{D(1;T_s)>kT_s\} \tag{5.30}$$

$$\approx \gamma_m(1-p_m)^k, k\in\{0,1,2,\cdots\} \tag{5.31}$$

In (5.31), the distribution of $D(1;T_s)$ given $D(1;T_s)>0$ approximates a geometric distribution, so the expectation of $D(1;T_s)$ given $D(1;T_s)>0$ is the reciprocal of p_m. The expectation of $D(1;T_s)$ can be easily derived as

$$E[D(1;T_s)] = \frac{T_s\gamma_m}{p_m} \tag{5.32}$$

On the other hand, because of the property of stationarity and ergodicity, the expectation of queue length $E[Q(T_s;T_s)]$ is the sample mean of a realisation with infinitely long samples, i.e.,

$$E[Q(T_s;T_s)] = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}q_i \tag{5.33}$$

Hence, the unbiased estimator for $E[Q(T_s;T_s)]$ is given by

$$E[\widehat{Q(T_s;T_s)}] = \hat{q} = \frac{1}{N}\sum_{i=1}^{N} q_i \tag{5.34}$$

Furthermore, Little's law gives a simple relation among the bit arrival rate and the expectations of queue length and queueing delay by [130]

$$E[Q(T_s;T_s)] = \mu \cdot E[D(1;T_s)] \tag{5.35}$$

By combining (5.32), (5.34) and (5.35), we have an unbiased estimator for p_m:

$$\hat{p}_m = \mu \frac{T_s \hat{\gamma}_m}{\hat{q}} \tag{5.36}$$

5.4.2.3 Computational Complexities of the Conventional Estimator and the Revised Estimator

Given the same realisations $\mathbf{s}$ of (5.23) and $\mathbf{q}$ of (5.24), the conventional estimators for γ and θ in [4] are estimated as follows:

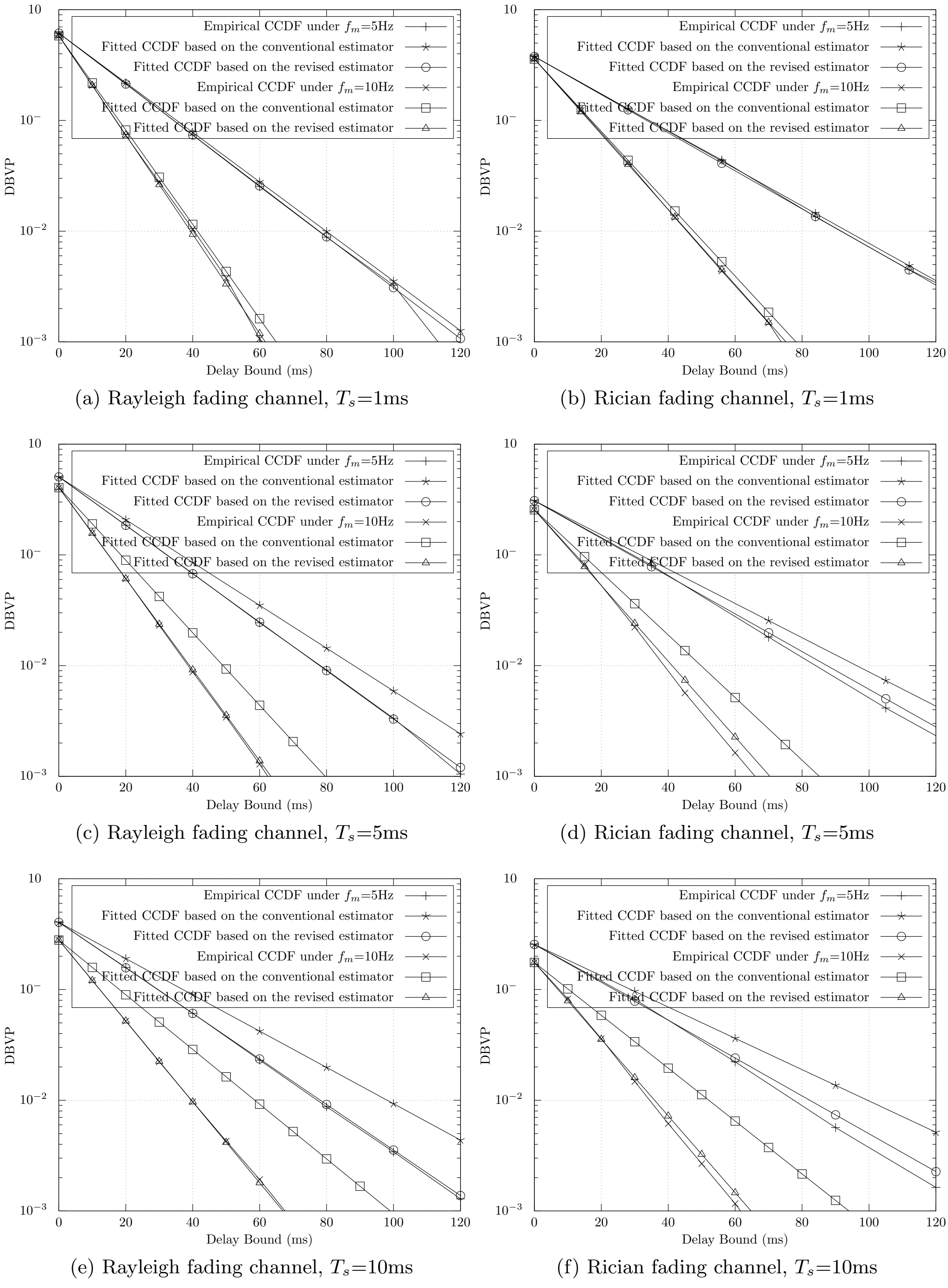
$$\hat{\gamma} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{q_i > 0\} \tag{5.37}$$

$$\hat{\theta} = \mu \frac{\hat{\gamma}}{\hat{q}} \tag{5.38}$$

Since (5.29) and (5.37) are exactly the same and the ratio of (5.36) to (5.38) is T_s, we conclude that the computational complexities of the conventional estimator and the revised estimator are the same.

5.5 Results and Discussion

Fig. 5.7 shows results of CCDFs of packet end-to-end delay in various values of T_s (1ms, 5ms and 10 ms). The figures of the left column are results under Rayleigh fading channel



(a) Rayleigh fading channel, T_s=1ms

(b) Rician fading channel, T_s=1ms

(c) Rayleigh fading channel, T_s=5ms

(d) Rician fading channel, T_s=5ms

(e) Rayleigh fading channel, T_s=10ms

(f) Rician fading channel, T_s=10ms

Figure 5.7: Empirical and fitted CCDFs of end-to-end packet delay under different fading channels

Table 5.3: Estimated $\gamma(\mu)$, $\theta(\mu)$, $\gamma_m(\mu;T_s)$ and $p_m(\mu;T_s)$ from the revised and conventional estimator

Wireless Situation			**Revised Estimator**		**Conventional Estimator**	
Channel	f_m (Hz)	T_s (ms)	$\hat{\gamma}_m$	$\hat{p}_m$	$\hat{\gamma}$	$\hat{\theta}$
Rayleigh	5	1	0.61	0.05	0.61	51.55
Rayleigh	5	5	0.51	0.22	0.51	44.52
Rayleigh	5	10	0.40	0.38	0.40	37.72
Rayleigh	10	1	0.58	0.10	0.58	98.05
Rayleigh	10	5	0.40	0.38	0.40	75.44
Rayleigh	10	10	0.28	0.57	0.28	56.88
Rician	5	1	0.38	0.04	0.38	38.83
Rician	5	5	0.31	0.18	0.31	35.62
Rician	5	10	0.26	0.33	0.26	32.55
Rician	10	1	0.36	0.08	0.36	75.26
Rician	10	5	0.26	0.33	0.26	65.10
Rician	10	10	0.18	0.55	0.18	54.98

and the ones of the right column are results under Rician fading channel. Furthermore, each figure contains results of 5Hz and 10Hz maximum Doppler rates. x-axes are delay bound in millisecond and y-axes are the delay bound violation probability in log scale. Empirical CCDFs are obtained from the realisations of delay processes.

From the figures, the fitted CCDFs from the revised estimator are always closer to the empirical CCDFs than that from the conventional estimator. Specially under the 10Hz maximum Doppler rate in Fig. 5.7e, the estimation result given 60ms delay bound from the revised estimator achieves one order of magnitude improvement comparing with the estimate from the conventional estimator. This finding indicates that the MCDT effective capacity model is more suitable for modelling packet delay in time-slotted wireless communication systems. Table 5.3 lists the estimated $\gamma^{(c)}(\mu)$s, $\theta^{(c)}(\mu)$s, γ_ms and θ_ms in all simulation scenarios.

Table 5.4 lists the average delay and jitter from simulation results, estimation results from conventional and revised estimations in all simulation scenarios. It can also be concluded that the MCDT effective capacity model developed in this chapter gives better characterisation of delay performance than the conventional effective capacity model.

Table 5.4: Descriptive statistics of end-to-end packet delay

Wireless Situation			**Average delay (ms)**	**Jitter (ms)**		
Channel	f_m (Hz)	T_s (ms)	Sim.	Sim.	Est.	Est. [4]
Rayleigh	5	1	11.92	17.43	17.90	17.56
Rayleigh	5	5	11.37	17.94	19.53	18.02
Rayleigh	5	10	10.73	18.51	21.30	18.61
Rayleigh	10	1	5.93	8.98	9.26	8.94
Rayleigh	10	5	5.37	9.26	10.65	9.31
Rayleigh	10	10	4.92	9.98	12.20	9.98
Rician	5	1	9.73	20.24	20.16	19.91
Rician	5	5	8.67	18.96	20.29	19.19
Rician	5	10	7.85	18.10	20.51	18.50
Rician	10	1	4.77	9.98	10.20	9.96
Rician	10	5	3.93	9.05	10.26	9.25
Rician	10	10	3.20	8.48	10.30	8.61

Furthermore, the differences between estimates from both estimators when T_s=10ms are much more significant than that when T_s=1ms. Such a finding is intuitive because when $T_s = 0$, a mixed continuous/discrete time queueing model becomes a continuous-time queueing model and both estimators will eventually produce the same estimate.

5.6 Conclusions

In this chapter, we model wireless slotted communication systems as MCDT queueing models. Such a model is an extension of the discrete-time queueing model, which further considers the effect of slot time. In Section 5.3, the conventional effective capacity model was adapted to MCDT queueing models. An MCDT effective capacity model was proposed to approximate the CCDF of packet end-to-end delay by two functions, namely the probability of non-empty buffer and the success probability of a connection. An estimation algorithm for these two functions was developed in this chapter as well.

The distribution fitting technique (the same as one in Chapter 4 Section 4.4) was adopted. The procedure of building a simulation platform was described in Section 4.4.3. An estimator based on this MCDT effective capacity model was developed in Section 4.4.3.

In Section 5.5, several simulations with different slot times, wireless channel models and maximum Doppler rates were carried out. The simulation results showed that the revised estimator was accurate in estimating CCDFs and was much more accurate than the conventional estimator when slot times are long (10ms slot time). This is attributed to the fact that the conventional estimation algorithm is developed based on continuous-time queueing models, which do not behave in a time-slotted manner.

In summary, the MCDT effective capacity model is an appropriate tool for analysing delay performances in wireless slotted communication systems.

Chapter 6

Estimation of End-to-End Delay Distributions in the IEEE 802.16-2004 Networks

6.1 Introduction

The simulation platforms built in Chapters 4 and 5 were based on ideal point-to-point communications. The cross-layer simulation concept is a new way to consider the Quality of Service (QoS) in networks [131]. The cross-layer design shares information between not necessarily adjacent levels; it is used to achieve optimistic performance [132, 133, 134]. In the literature, research on network performance in the IEEE 802.16j Mobile Multi-hop Relay (MMR) networks is limited to the analysis of network throughput and average packet delay [135, 136, 25, 137, 138, 139, 140].

In this chapter, a new realistic cross-layer simulation platform is developed using Simulink and it integrates the IEEE 802.16-2004 physical layer, a simplified link layer and network layer to introduce the concept of multi-hop architectures.

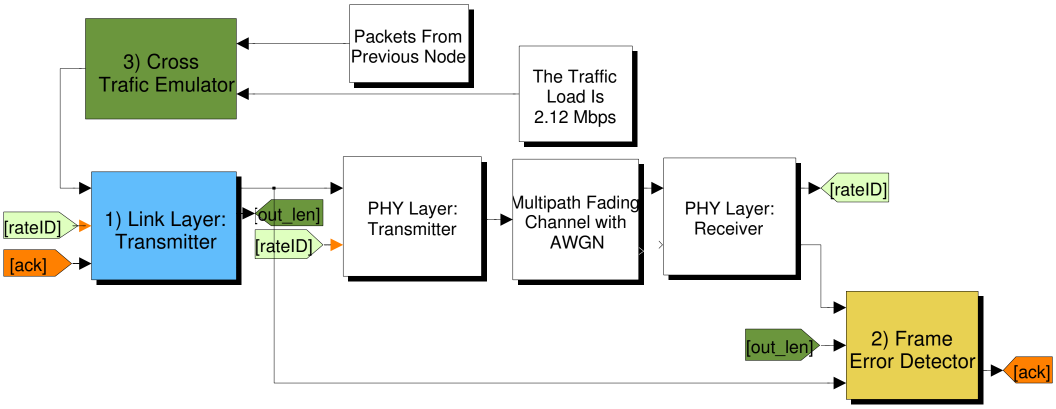


Figure 6.1: Cross-layer simulation platform based on the IEEE 802.16-2004

Simulink, developed by the MathWorks, Inc. is a data flow graphical programming language tool for modelling, simulating and analysing multi-domain dynamic systems. Other simulators also have sufficient communication structures but Simulink is chosen due to its wide availability.

Such a simulation platform enables users to answer the essential question posed in this thesis: is it possible to use the multi-hop effective capacity model and Mixed Continuous/Discrete-Time (MCDT) effective capacity model to estimate end-to-end delay performances in realistic communication scenarios? Most results discussed in this chapter have been published in the ICT 2011 conference proceedings and the WCNC 2013 conference proceedings [65, 141].

The rest of the chapter is structured as follows: Section 6.2 explains the important building blocks of the simulator and how the Lindley equation is used. Results for multi-hop scenarios and single-hop scenarios are illustrated and discussed in Sections 6.3. Section 6.4 concludes the chapter.

6.2 Simulation Platform

One major difference between the pure physical-layer simulation and cross-layer simulation is the assumption of traffic patterns. In detail, the pure physical-layer simulation assumes there are always binary bits coming from the upper-layer, whereby the objective of such simulation is to investigate Bit Error Rate (BER). The cross-layer simulation uses a queue that provides memory to store extra bits until such bits can be transmitted. By using this method, from the physical-layer simulation view point, the assumption that there is a continuous stream of binary bits will be invalid (except for the special case when the buffer in the link layer is assumed to be of infinite capacity). Therefore the physical-layer has to be idle for statistically distributed time periods when there are no bits to serve. Conversely, from the link-layer point of view, the network may take different forms of the packet inter-arrival time distribution and the packet departure distribution. Therefore, the objective of the cross-layer simulation is to investigate QoS performances.

The simulation platform is developed based on the layer-3 type of implementations (discussed in Chapter 2, Section 2.2.2) and the system model of Fig. 4.1 to resemble the IEEE 802.16j (discussed in Chapter 2, Section 2.2.3.2). However, it contains some modifications:

1. Each node uses dedicated frequency bands to transmit and receive signals;

2. The scalable OFDMA and the Multiple-Input Multiple-Output (MIMO) technologies are not implemented although they are supported in the IEEE 802.16j standard;

3. There will no packet overhead insertion in each layer.

The operations performed at each node are shown in Fig. 6.1 and details in each layer are explained in Sections 6.2.1, 6.2.2 and 6.2.3.

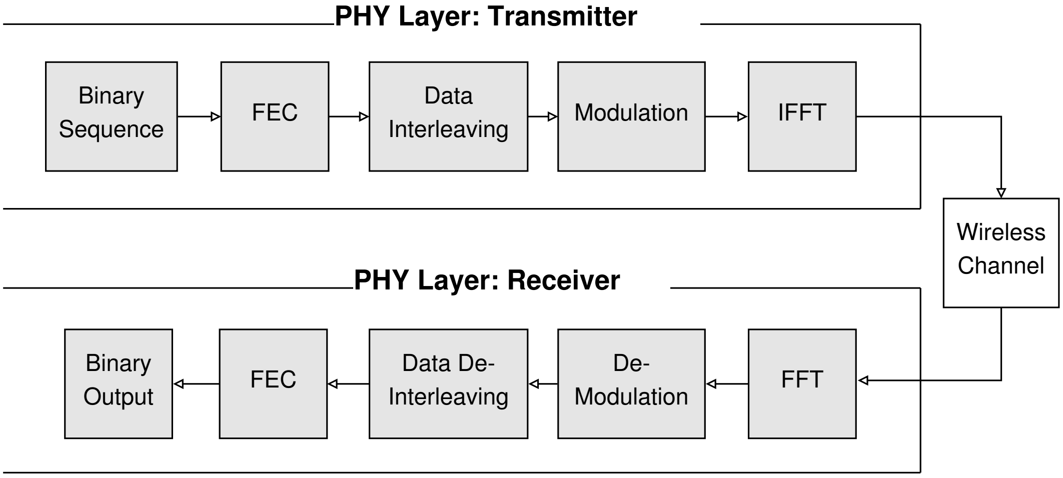


Figure 6.2: The IEEE 802.16j Physical-layer Implementation

6.2.1 Physical-layer Implementation of the IEEE 802.16-2004

Fig. 6.2 shows the physical-layer building blocks based on the IEEE 802.16-2004 standard. The standard uses Orthogonal Frequency-Division Multiple Access (OFDMA) as the primary channel access mechanism to mitigate the Non-Line-Of-Sight (NLOS) effect. Other contemporary technologies are also used, including Forward Error Correction (FEC), consisting of a Reed-Solomon (RS) outer code concatenated with a rate-compatible inner convolutional code (CC), Data interleaving and Adaptive Modulation and Coding (AMC) scheme. The IEEE 802.16-2004 simulation model is available in the Simulink library[142] and is integrated in our simulator.

6.2.2 Link-layer Implementation

As shown in Fig. 6.1, the Automatic Repeat reQuest (ARQ) scheme is used to ensure a reliable transmission because bit errors are unavoidable in real-world communications and so are packet errors. There are several ways to detect an erroneous packet in a computationally-efficient manner, such as using Cyclic Redundancy Check (CRC). However, an ideal assumption is assumed in our simulation, i.e., the ACK indicator is computed by comparing the received packet and the original packet transmitted.

Since the AMC scheme is in use, the service rate at slot n, which is denoted by $r[n]$, is calculated based on the feedback on the estimated Signal-to-Noise Ratio (SNR) of a wireless channel at slot $(n-1)$ that is estimated at the receiver side. Therefore, $r[n]$ is time-varying.

Finally, the Lindley equation for queueing length processes (2.34) (see Chapter 2, Section 2.4.5.2) is adopted to calculate the sequences of queue length and departure bits. When ARQ scheme is switched on, the following algorithm is implemented (this algorithm and the next one are written in Matlab code because any Simulink model is able to execute Matlab codes).

```
if (ack[n-1] == 1)
    s[n] = min(q[n-1] + a[n], r[n]);
    q[n] = max(0, q[n-1] + a[n] - r[n]);
else
    s[n] = min(q[n-1] + a[n] + s[n-1], r[n]);
    q[n] = max(0, q[n-1] + a[n] + s[n-1] - r[n]);
end
```

When ARQ scheme is switched off (the non-ARQ scheme is mainly for multimedia services that are delay-sensitive but accept some packet loss in data streams [143]), an alternative algorithm is implemented:

```
s[n] = min(q[n-1] + a[n], r[n]);
q[n] = max(0, q[n-1] + a[n] - r[n]);
```

6.2.3 Layer-3 Relay Implementation

Since packets departed from each node have two possible destinations:

1. the next node along the routing path
2. other nodes outside the routing path

Table 6.1: Simulation parameters

Parameter	Value
Bandwidth (MHz)	3.5
Channel model	Rician three-path distribution
Delay vector (ms)	[0 0.4 0.9]*1e-3
Gain vector (dB)	[0 -5 -10]
3-hop Scenario	
Average SNR (dB)	15
Average traffic load (Mbps)	2.08 or 2.60
K-factor	3
Maximum Doppler rate (Hz)	10
OFDM symbol time: T_s (ms)	0.072
Simulation time (seconds)	15
Traffic correlation index: p	25% or 75%
Single-hop Scenario	
Average SNR (dB)	9 or 12
Constant inter-arrival time (slot) t_c	1
K-factor	0.5
Maximum Doppler rate: f_m (Hz)	0.5
OFDM Symbol time: T_s	4.96 ms
Simulation time (second)	30
Traffic load (Mbps)	2.12

The cross-layer emulator in Fig. 6.1 acts as a routing decision maker that decide if a packet is to be relayed to its following node or not. In the mean time, each node will generate a certain amount of packets that emulates either local traffic or traffic from other sources outside the routing path.

6.2.4 Simulation Settings

6.2.4.1 Multi-hop Scenarios

Consider every channel between relay stations has a strong Line-of-Sight (LoS) component. In the simulation, the Rician fading channel is assumed. Four simulation scenarios are investigated: 1) light traffic load and weak traffic correlation ($\mu = 2.12$Mbps and $p = 0.25$), 2) light traffic load and strong traffic correlation ($\mu = 2.12$Mbps and $p = 0.75$), 3) heavy traffic load and weak traffic correlation ($\mu = 2.65$Mbps and

$p = 0.25$) and 4) heavy traffic load and strong traffic correlation ($\mu = 2.65$Mbps and $p = 0.75$). Other values of simulation parameters are listed in Table 6.1.

6.2.4.2 Single-hop Scenarios

According to the deployment details described in [27, 144], the wireless bandwidth in simulation is 3.5 MHz, a physical-layer frame contains 69 OFDM symbols and its duration is 4.968ms. Packets in the link-layer buffer are served and new packets arrive at the buffer every 4.968 ms. The packet generator generates a constant input rate with a constant packet length of 10,000 bits.

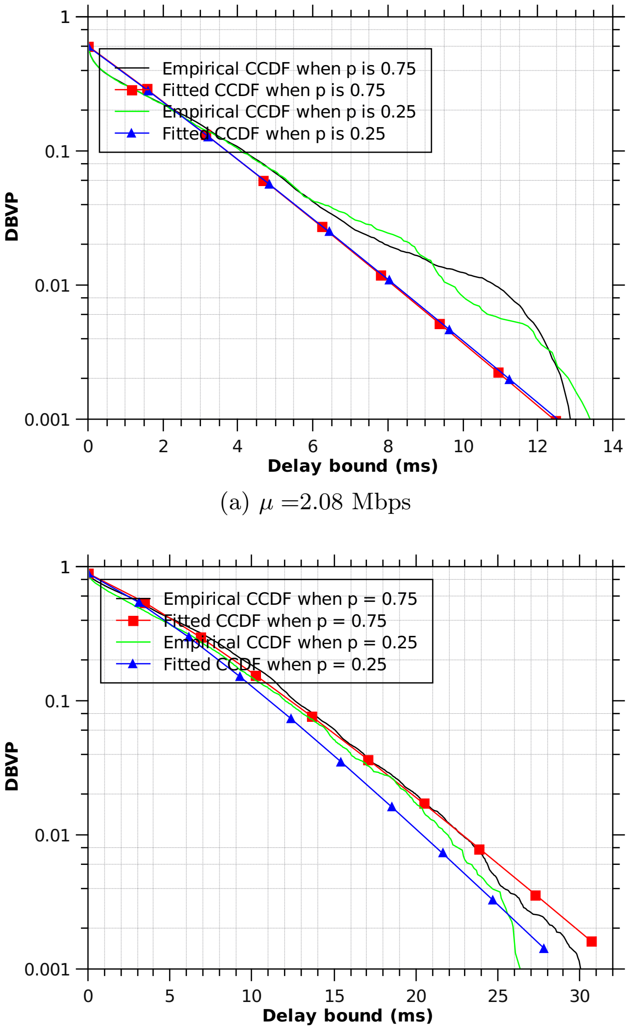
Two scenarios are simulated: the average SNR is 9dB and the average SNR is 12 dB (values of the receiver SNR assumptions are proposed in Table 266 of the IEEE 802.16e amendment of the standard [145]). A three-path Rician fading channel model is assumed to represent Line-of-Sight (LoS) communications. Table 6.1 lists simulation parameters used in the paper.

6.3 Results and Discussion

The results in the multi-hop and single-hop IEEE 802.16-2004 network are presented in Sections 6.3.1 and 6.3.2, respectively. For multi-hop scenarios, the estimator developed in Chapter 4, Section 4.4.3 is adopted to estimate delay performances. For single-hop scenarios, the estimator developed in Chapter 5, Section 5.4.2 is adopted.

6.3.1 Multi-hop Scenarios

Fig. 6.3 show the simulation and analysis results under different traffic loads. The X-coordinates are Delay Bounds (the unit is milliseconds), and the Y-coordinates are Delay Bound Violation Probabilitys (DBVPs). Simulation results are shown in red solid lines, while estimation results are shown in blue dashed lines. Moreover, results of different traffic correlations are grouped and pointed out by upper arrows with traffic



(a) μ =2.08 Mbps

(b) μ =2.60 Mbps

Figure 6.3: Empirical and fitted CCDFs of end-to-end packet delay

correlation indices underneath. Table 6.3 summarises the simulation and estimation results of average delay and jitter under different conditions.

On the basis of the results in figures and tables, simulation results well match the estimation results, showing the accuracy of the multi-hop effective capacity model. Additionally, the results indicate that the delay performance will be affected by tuning the value of traffic load or traffic correlation index, which is thoroughly discussed in [114].

Table 6.2: Estimated parameters from the multi-hop effective capacity-based and MCDT effective capacity-based estimators

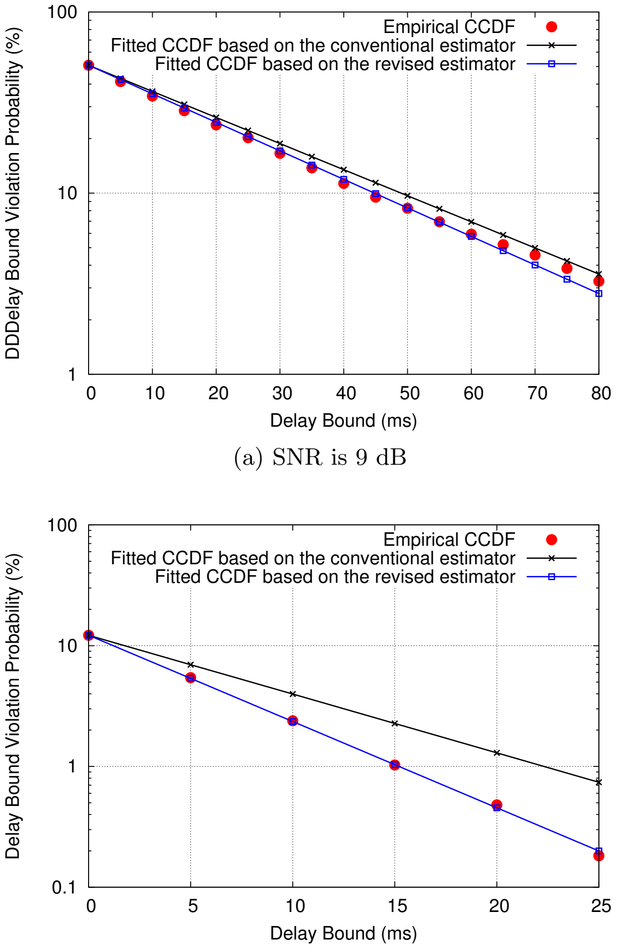
Wireless Situation			**3-hop effective capacity model**					
3-hop scenario								
μ (Mbps)	r	SNR (dB)	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
2.08	0.25	15	0.17	0.29	0.32	872.33	604.41	569.58
2.08	0.75	15	0.16	0.30	0.33	857.87	604.92	585.01
2.60	0.25	15	0.49	0.53	0.53	388.90	287.47	322.65
2.60	0.75	15	0.50	0.53	0.54	387.28	256.84	274.42
			MCDT effective capacity model					
Single-hop scenario								
			$\hat{\gamma}_m$			$\hat{p}_m$		
2.12	1	9	0.51			0.17		
2.12	1	12	0.12			0.56		

Table 6.3: Descriptive statistics in different wireless situations

Wireless Situation			**Avg. E2E delay**		**Jitter**	
3-hop scenario						
μ (Mbps)	p	SNR (dB)	Sim. (ms)	Est. (ms)	Sim. (ms)	Est. (ms)
2.08	0.25	15	1.24	1.24	2.16	1.85
2.08	0.75	15	1.25	1.25	2.16	1.85
2.60	0.25	15	4.76	4.76	5.17	4.67
2.60	0.75	15	5.32	5.32	5.40	5.22
Single-hop scenario						
2.12	1	9	15.22		25.23	24.59
2.12	1	12	1.08		3.51	3.55

6.3.2 Single-hop Scenarios

Fig. 6.4 shows simulation and estimation results when the SNR of the wireless channel is 9 dB and 12 dB. The X-coordinate is the queueing delay (the unit is milli seconds) and the Y-coordinate is the Delay Bound Violation Probability (the unit is percentage). There are three lines in each figure, the simulation result is shown in solid circles, the estimation result obtained from revised estimation algorithm is plotted in a line with squares and the line with cross symbols is from the conventional algorithm. It is clear that for both simulation scenarios, the revised estimation algorithm gives results that are very close to those obtained from full cross-layer simulations, and are more accurate than those estimated by Wu's estimation algorithm.



(a) SNR is 9 dB

(b) SNR is 12 dB

Figure 6.4: Empirical and fitted CCDFs of end-to-end packet delay

Table 6.3 compares simulation results with estimation results regarding the average end-to-end delay and jitter, which validates the accuracy of the revised estimation algorithm. Besides, the BER results are 7.7e-3 and 3.4e-3 for 9 dB scenario and 12 dB scenario, respectively. in this chapter, we only measure BERs and refer the discussion of the effects of BER performance to [146, 147, 148].

6.4 Conclusions

In this chapter, a realistic cross-layer simulation platform was designed. The platform was implemented using Simulink, however, it is generic nature would allow it to be equivalently implemented using other simulation/mathematical modelling software. The simulator implements the key building blocks in physical, link and network layers and is used to 1) evaluate the link-layer end-to-end delay performance and verify the performances of the multi-hop effective capacity model and MCDT effective capacity model.

Two scenarios were considered: multi-hop communication systems with ARQ and single-hop communication systems with non-ARQ. In the first scenario, different traffic loads and traffic correlations were conceived. In the second scenario, different channel conditions were conceived.

In each scenario, delay values of every packet were collected to obtain Complementary Cumulative Density Function (CCDF) of a packet delay, average delay and jitter. The simulation results of delay performances and estimation results from the methods in Chapters 4.4.3 or 5.4.2 are in good agreement, indicating that it may be appropriate to use the multi-hop effective capacity model and MCDT effective capacity model to characterise delay performances for relay wireless communication systems.

Chapter 7

Estimation of End-to-End Frame Delay Distributions in Wireless Tele-ultrasonography Medical Systems

7.1 Introduction

Ultrasonography systems are commonly used in medical diagnosis for various medical conditions [149]. What is termed as wireless tele-ultrasonography medical systems are the systems that transmit the ultrasonic images over wireless telecommunication networks. The nature of the images and the application necessitates specialised system design parameters including Quality of Service (QoS).

In wireless tele-ultrasonography medical systems, the concept of Medical Quality of Service (M-QoS) was introduced in [150]. M-QoS was defined as the "*augmented requirements of critical mobile health-care applications with respect to traditional wireless QoS requirements*" and considers diagnostic image quality, frame rate and end-to-end delay [150]. Table 7.1 shows an example of M-QoS metrics for a tele-ultrasonography scenario. The major drive for defining end-to-end delay constraints is because in video streaming applications, if a video frame does not arrive on time, the play out process will pause, which is a highly undesired effect.

Earlier work in this area focused on video quality enhancement using Q-learning approach [150, 151, 152] or cross-layer adaptation to wireless transmission medium [153, 154, 155]. Evaluations of average end-to-end delay and jitter were carried out in [156, 157, 158]. However, the work mentioned above did not study end-to-end delay distributions, which provide complete information of end-to-end delay.

In this chapter, it is the first time that the effective capacity technique is adopted to characterise and estimate end-to-end delay distributions in wireless medical systems. The estimator developed in Chapter 5, Section 5.4.2 is directly applied to such systems. To validate the accuracy of the estimation algorithm, a cross-layer simulation platform is built to represent a wireless tele-ultrasonography medical system. The platform includes link-layer functions and physical-layer implementations that follow the fixed Worldwide Interoperability for Microwave Access (WiMAX) standard. Furthermore, a real ultrasound medical video from a portable ultrasound machine is MPEG-2 encoded and is transmitted in the platform. The simulation results from the platform are compared with the estimation results from the estimation algorithm. Most results discussed in this chapter were accepted to be presented in the GLOBECOM 2013 conference [159].

The rest of the chapter is structured as follows: Section 7.2 introduces a wireless tele-ultrasonography medical system and its system model. In Section 7.3, the details of the simulation platform is described. Results are illustrated and discussed with respect to M-QoS metrics in Section 7.4. Section 7.5 concludes the chapter.

7.2 System Model

A wireless tele-ultrasonography medical system is shown in Fig. 7.1. Such a system is composed of an expert site, a wireless communication media and a patient site. Ultra-

Table 7.1: m-QoS for a tele-ultrasonography scenario [146]

m-QoS metrics	Acceptable values
end-to-end delay	< 350 ms
Frames per second (FPS)	> 5
Image quality (PSNR)	> 36 dB
Image quality (SSIM)	> 0.9
Ultrasound frame sizes (x × y pixels)	4×CIF (4CIF) (704×576) Common Intermediate Format (CIF) (352×288) Quarter CIF (QCIF) (144×176)

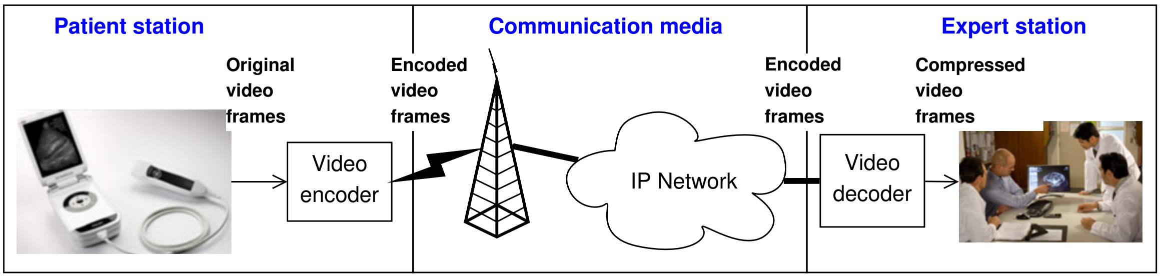


Figure 7.1: Wireless tele-ultrasonography medical system

sound videos are generated on *patient sites* and are transmitted through communication networks to *expert sites* for diagnosis. Video encoder and decoder are used to compress raw videos and make them suitable for transmission. Further details on this system are described in [160, 161].

The communication media block in Fig. 7.1 may be modelled as a queueing model. Fig. 7.2 shows a system model of a wireless tele-ultrasonography medical system model. As seen in the figure, raw video frames are a sequence of fix-sized frames that are generated from the video generator every t_c seconds; the video encoder encodes raw video frames and output encoded video frames (compressed variable-length binary sequences) every t_c seconds. The queue discipline is assumed to be First-In First-Out (FIFO) and queue size is assumed to be infinite.

Consider High Speed Packet Access (HSPA) and WiMAX systems, in which physical-layer frames are transmitted every Transmission Time Interval (TTI). Let T_s denote the duration of a slot (or a TTI). For such systems, the queueing model in Fig. 7.2 is better described as being Mixed Continuous/Discrete-Time (MCDT) (MCDT queueing

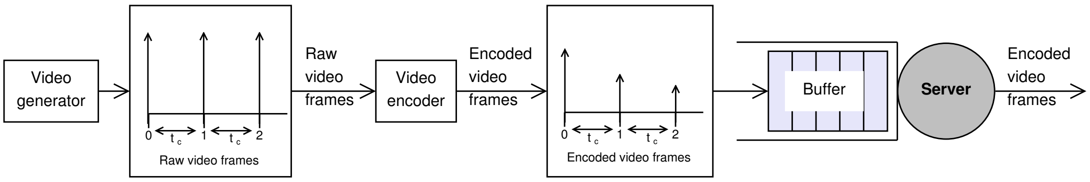


Figure 7.2: System model of a wireless tele-ultrasonography medical system

systems are discrete-time systems with consideration of slot time T_s and described in Chapter 5, Section 5.3). By convention, frames arrive after slot boundaries and depart before slot boundaries [28].

7.3 Simulink Simulation Platform

Fig. 7.3 shows a cross-layer Simulink simulation platform based on the system model of Fig. 7.2. Since it is a single-hop scenario, the network layer can be safely ignored and the platform only implements three layers, namely, physical layer, link layer and application layer. The physical-layer implementation is identical to the implementation described in Chapter 6.2.1.

Similar to Chapter 6.2.2, the link-layer implementations include a link-layer queue and radio-control functions (retransmission control by Automatic Repeat reQuest (ARQ), and video frame segmentation/reassembly). The encoded video frames are stored in a link-layer buffer. By following the assumptions made in Section 7.3, the queue discipline is set to be FIFO and the queue size is set to be large enough so that queue overflow is eliminated.

In the implementation, an extra process, which is different from that in Section 6.2.2, is required to solve the following two questions: 1) what bits should be sent and 2) when the bits should be sent.

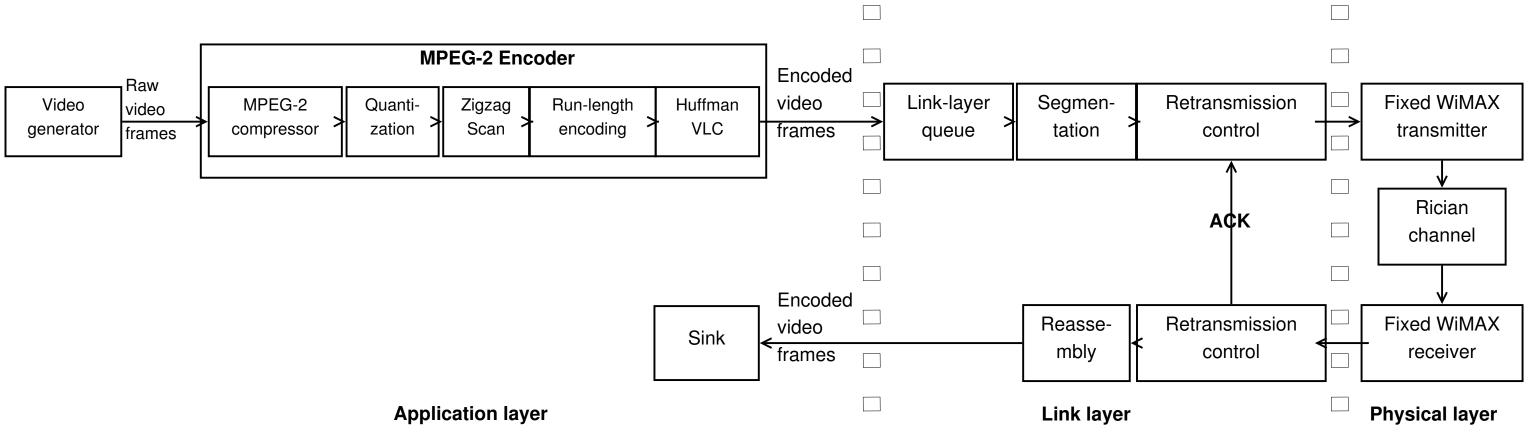


Figure 7.3: Overview of a cross-layer simulation platform

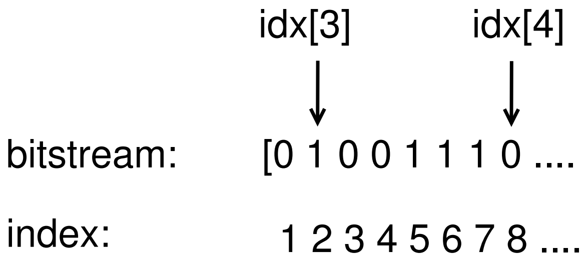


Figure 7.4: Indexing operation

The first problem is solved by a vector *bitstrem*, and the second is addressed by a vector *idx*. Fig. 7.4 shows an example vector *bitstrem* that contains the information of a medical video. The values of idx[n] and idx[n-1] are used to indicate which bits should be sent at slot n, i.e., bitstream[idx[n-1] + 1, idx[n]]. The values of idx[n] and idx[n-1], together with queue length and output bit length, are calculated from the algorithm (this algorithm is written in Matlab code) below:

```
if (ack[n-1] > 0)
    s[n] = min(q[n-1] + a[n], r[n]);
    q[n] = max(0, q[n-1] + a[n] - r[n]);
    idx[n] = idx[n-1] + s[n-1];
else
    s[n] = min(q[n-1] + a[n] + s[n-1], r[n]);
    q[n] = max(0, q[n-1] + a[n] + s[n-1] - r[n]);
    idx[n] = idx[n-1];
end
```

MPEG-2 encoder is used for video compression. Its implementations are shown in Fig. 7.3 and are based on the ISO/IEC 13818-2 standard [162].

7.3.1 Simulation Settings

The raw video captured from a portable ultrasound machine has a frame rate of 25 FPS. According to the system model described in Section 7.3, the encoded video (the

Table 7.2: Simulation parameters

Parameter	**Value**
MPEG-2 encoder	
Chroma sub-sampling format	4:2:0
Frame type pattern	I P P P P
Motion vector search	Logarithmic type
Scales for I and P-frames	5 or 7
Fixed WiMAX network	
Average SNRs (dB)	7, 12 or 19
Bandwidth: BW (MHz)	3.5
Channel model	Rician three-path channel
Delays of three paths (ms)	[0 0.4 0.9]*1e-3
Gains of three paths (dB)	[0 -5 -10]
K-factor	0.5
Maximum Doppler rate (Hz)	0.5
Simulation time (second)	30
TTI or slot time: T_s (ms)	4.968 (69 OFDM symbols)

video frames are fed into the link-layer block) also has a frame rate of 25 FPS so t_c is 40 ms in this case.

Three wireless environments are used: average SNRs of 7 dB (low SNR), 12 dB (medium SNR) and 19 dB (high SNR). Together with two different scale factors, there are six simulation scenarios. Parameters for the MPEG-2 encoding and the fixed WiMAX network are listed in Table 7.2.

7.4 Results and Discussion

The basic video analysis is carried out and presented in Section 7.4.1. In Section 7.4.2, the simulation results are contrasted with those obtained from the estimator (see Chapter 5, Section 5.4.2).

7.4.1 Analysis of Video Qualities

The raw video is of resolution 320 x 240 (nearly a CIF). Properties of the raw video are summarised in the second column of Table 7.3.

Table 7.3: Video properties, qualities and compression ratios among raw videos and compressed videos

	Raw video	**Scale = 5**	**Scale = 7**
Data compression ratio	1:1	118.62:1	132.31:1
Frame rate (fps)	25	—	—
Frame size (pixels×pixels)	320×240	—[1]	—
PSNR (dB)	∞	41.66	39.06
SSIM	1	0.977	0.964
Total frames	750	—	—
Video length (second)	30	—	—

[1] Same as the raw video

Two quantisation scale factors (5 and 7) for I and P-frames are used. The scale factor has a trade-off between quality and compression. The average Peak Signal-to-Noise Ratio (PSNR) values, Structural Similarity (SSIM) indices and data compression ratios using these two scale factors are listed in Table 7.3. It is shown that 1) both of their image qualities meet the requirements of M-QoS and 2) the scale factor of 7 results in better compression but at the expense of worse quality.

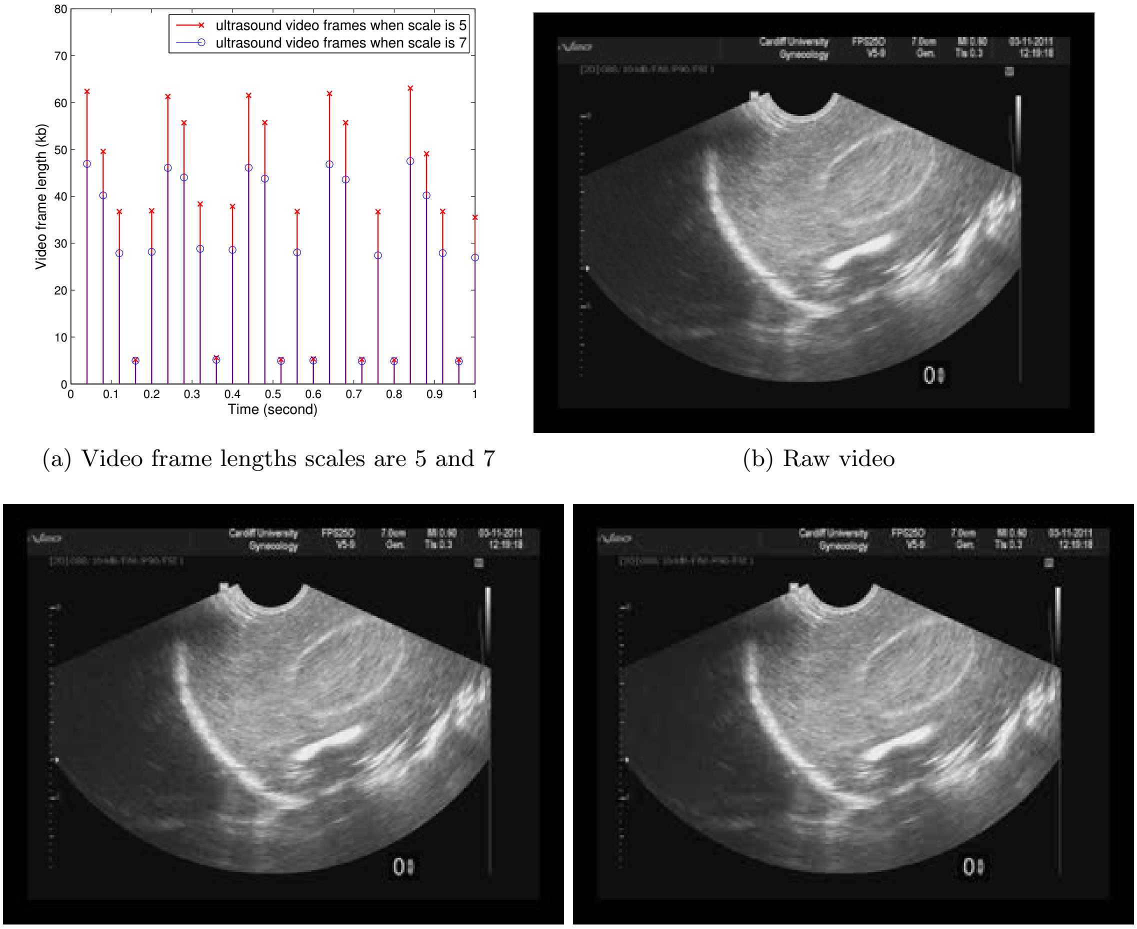
Fig. 7.5a shows video frame lengths when Scales are 5 and 7. As seen from the figure, the frames when Scale is 5 are always larger than that when Scale is 7.

Figs. 7.5b, 7.5c and 7.5d show the comparative visual results for the first frame of three videos. The frame from the raw video is reported in Fig. 7.5b; the MPEG-2 decoded frames are shown in Figs. 7.5c and 7.5d. By visual inspection, it can be seen that Fig. 7.5c presents a higher visual quality than Fig. 7.5d because Fig. 7.5c is decoded from the encoded video with the scale factor of 5.

7.4.2 Performance of the Estimator

Since there are 750 frames in the ultrasound video, 750 samples of frame delays will be obtained in each simulation.

Table 7.4 shows delay performances in each simulation scenario. It is seen that when the SNR is 7 dB and the scale factor is 5, the average end-to-end delay fails



(a) Video frame lengths scales are 5 and 7

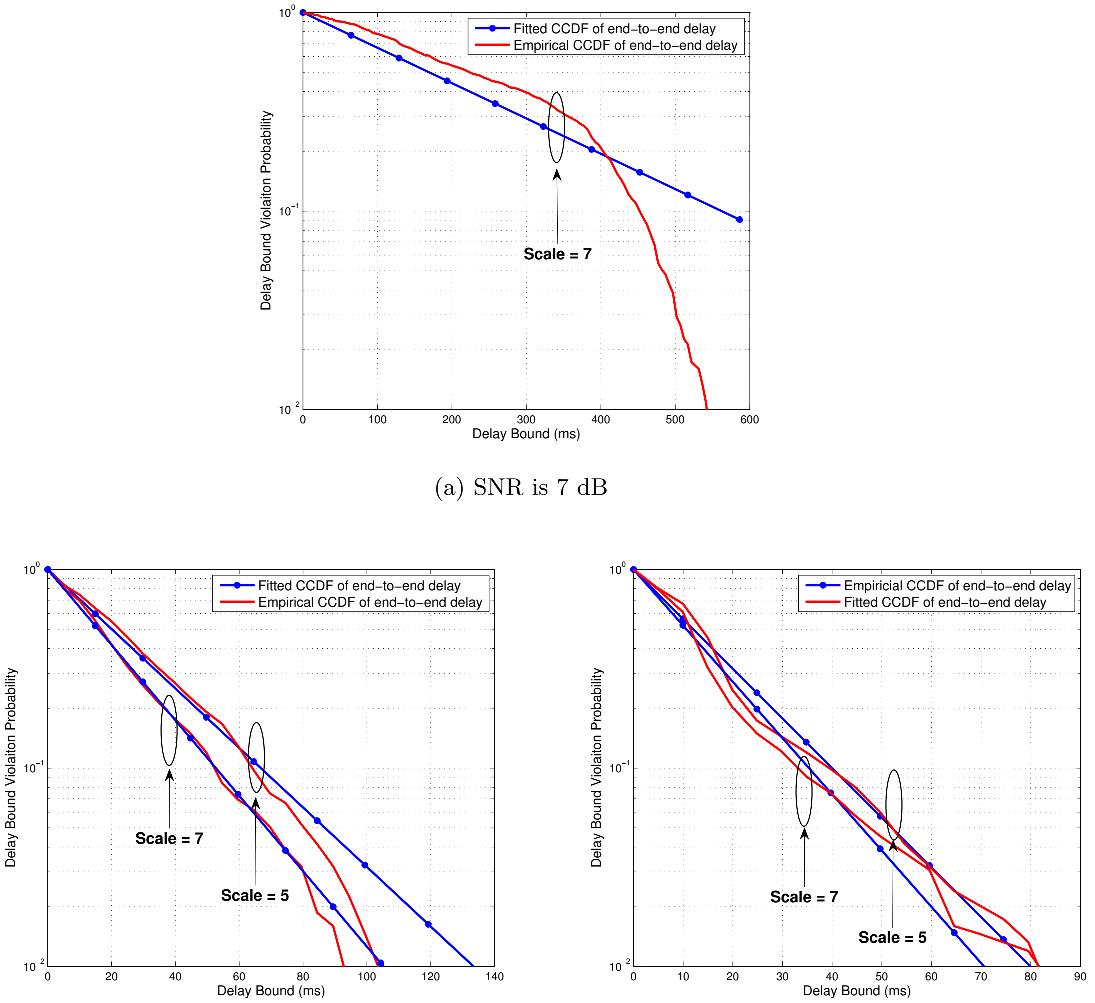
(b) Raw video

(c) Scale = 5

(d) Scale = 7

Figure 7.5: Video frame lengths and comparative visual results of medical video images ((c) and (d) after MPEG-2 decoding)

to meet the 350 ms delay requirement. In such a situation, the system is not stable because the bit rate coming to the system is higher than the system capacity, making the system overloaded. The average end-to-end delays of the other five scenarios are well below the threshold of 350 ms. However, the average delay when the SNR is 7 dB and the scale factor is 7 is considerably high and 30% of video frames experienced delay more than 350 ms. The system in this situation is heavily loaded due to the low system capacity when the SNR is low. Table 7.4 also lists estimated average delays and jitters in different scenarios. On the basis of the results from the five stable system scenarios, it can be concluded that the estimation algorithm provides good estimates



(a) SNR is 7 dB

(b) SNR is 12 dB

(c) SNR is 19 dB

Figure 7.6: Empirical and fitted CCDFs of end-to-end frame delay

under all wireless situations.

Considering the system under the SNR of 7 dB and the scale factor of 5 is unstable, empirical and fitted Complementary Cumulative Density Functions (CCDFs) of end-to-end delay for the rest of the five simulation scenarios are plotted only in Fig. 7.6. The X-axes are delay bounds (the unit is ms), and the Y-axes are Delay Bound Violation Probabilitys (DBVPs). The simulation results of CCDFs of end-to-end delay are shown in red lines, while the fitted results are shown in blue lines with circle marks. Results are grouped into three figures under three SNR values and in each figure, results of

Table 7.4: Descriptive statistics of end-to-end frame delays

Wireless Situation		**Average end-to -end delay (ms)**	**Jitter (ms)**		**DBVP of 350ms end-to-end delays**
SNR (dB)	Scale	Sim.	Sim.	Est.	9.4e-1
7	5	2287.27	1449.23	2299.16	3.0e-1
7	7	204.53	148.14	216.63	0
12	5	7.65	14.20	13.78	0
12	7	5.51	11.16	10.59	0
19	5	6.9	8.94	7.04	0
19	7	6.94	9.31	7.08	0

Table 7.5: Estimated parameters from the MCDT effective capacity-based estimator

Wireless Situation		**MCDT effective capacity model**	
SNR (dB)	Scale	$\hat{\gamma}_m$	$\hat{p}_m$
7	5	1	0.0022
7	7	1	0.0226
12	5	1	0.2654
12	7	1	0.3225
19	5	1	0.5216
19	7	1	0.5196

different scale factors are pointed. Estimated parameters of the probability of non-empty buffer $\hat{\gamma}_m$ and the successful probability of a connection $\hat{p}_m$ in each scenario are included in Table 7.5 for reference. Note that all γ_ms equal 1. It is because when frame sizes are non-negligible, the frame transmission delays will be non-negligible, resulting in non-zero end-to-end frame delays.

As seen from the figures, all simulation results are no smooth when compared to the one million packet simulation in Chapter 5; the value of 750 frame delay samples is not a large number that could reduce the bumpiness of results. Apart from the issue of smoothness, the estimation results show similar trends against simulation results with the exception of case when SNR is 19 dB. It may be explained by taking the fact that when the system is heavily loaded, the system needs more time to reach stability.

7.5 Conclusions

This chapter presented an application of using effective capacity technique for wireless tele-ultrasonography medical systems.

A cross-layer simulation platform was built to represent a wireless tele-ultrasonography medical system. A real ultrasound medical video from a portable ultrasound machine was used as a data source. The video was first MPEG-2 encoded and then transmitted via a platform that integrates link-layer functions and fixed WiMAX physical-layer implementations.

The estimator was tested in six simulation scenarios of different channel conditions and quantisation scale factors. The results showed that in most cases the estimation results are close to simulation results in terms of average delay, jitter and CCDF of packet end-to-end delay.

The effective capacity-based estimation algorithm is appropriate to estimate end-to-end delay distributions for wireless ultrasound video streaming. Since the MPEG-2 video compression algorithms and codecs combine spatial image compression and temporal motion compensation, different types of raw videos will have different compression ratios. The images of the ultrasound video used in this chapter are highly correlated, which results in higher compression ratios comparing with normal videos. However, the effective capacity model and the estimation algorithm do not specifically depend on any types of videos. Therefore, the finding in this chapter further suggest the potential capability of the estimation algorithm for other types of video streaming applications with different end-to-end delay distributions.

Finally, the work reported in this chapter although limited to study basics on ultrasound medical videos, the estimator would be applicable to different systems where there is required bounded end-to-end delay. The estimation results would guide the designer of such systems so that required QoS provisions are obtained.

Chapter 8

Conclusions

Multi-hop wireless technology, Quality of Service (QoS) provisioning and Transmission Time Interval (TTI) are three concepts wildly used in modern communication systems. End-to-end packet delay is a key QoS metric; the effective capacity model is a delay-constrained capacity that relates the wireless capacities to the end-to-end link-layer packet delay distributions. The thesis extends the effective capacity model to characterise delay distributions in multi-hop wireless networks and time-slotted networks that transmit packets every TTI.

8.1 Summary of the Thesis

Chapter 1 introduced the research topic of this thesis and the motivations behind the research. In Chapter 2, technical and research overviews of *three* concepts that directly relate to our research topic were given, namely, the multi-hop wireless networks, end-to-end packet delay and modelling of end-to-end packet delay.

As the basic model of the thesis, the effective capacity model gives characterises packet delay performances in wireless single-hop networks, i.e., mathematical formulae of Complementary Cumulative Density Function (CCDF) of delay, average delay and jitter of delay. Specifically, the model states that under certain conditions, the tail distributions of packet delay are *exponentially* bounded; the CCDF of packet delay can

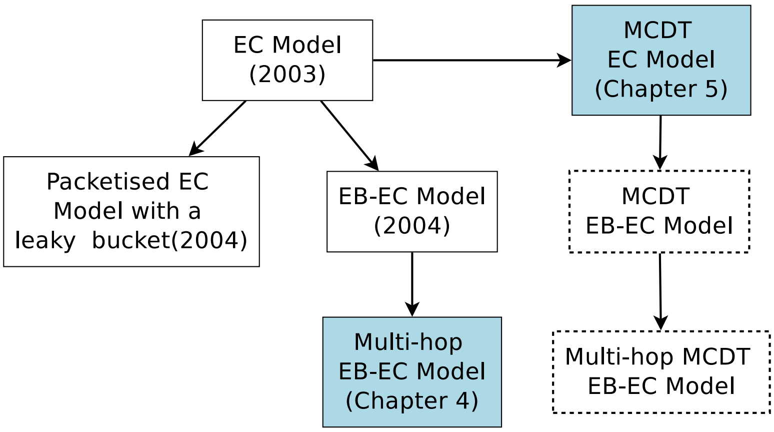

EC Model (2003)
MCDT EC Model (Chapter 5)
Packetised EC Model with a leaky bucket(2004)
EB-EC Model (2004)
MCDT EB-EC Model
Multi-hop EB-EC Model (Chapter 4)
Multi-hop MCDT EB-EC Model

Figure 8.1: Theoretical progression of the effective capacity model; extensions of this thesis in shaded boxes

be characterised by two functions, namely, the probability of non-empty buffer and the QoS exponent of a connection. Since the model is limited to constant rate and fluid traffic sources, it was extended to effective-bandwidth sources (Effective Bandwidth-Effective Capacity (EB-EC) model) and packetised traffic sources from a leaky bucket (packetised effective capacity model with a leaky bucket).

The packetised effective capacity model with a leaky bucket was first tested using an publicly available Internet traffic trace over gigabit Ethernet from the University of Massachusetts Amherst (UMASS). The Lindley equation was used to analyse the trace and produced a sequence of packet delays, which further produced the empirical CCDF of packet delay. The results showed that the tail distribution is indeed exponentially bounded, which validates the EB-EC model in a wired network.

All extensions of the model and theoretical progressions of the effective capacity model are shown in Fig. 8.1. The extensions developed in this thesis are shown in light-blue boxes.

The development of the multi-hop effective capacity model was in Chapter 4. The model gives mathematical formulae of CCDF of end-to-end delay, average delay and jitter in multi-hop wireless environments. For the H-hop scenario (packets traverse H nodes along a route), the CCDF of end-to-end delay can be characterised by $2H$

functions: H functions of the probability of non-empty buffer and H functions of the QoS exponent of a connection. A cross-layer (link layer and physical layer) multi-hop simulator was built and the capacity of the physical layer is based on Shannon's capacity. The empirical CCDFs and the fitted CCDFs of end-to-end delay based on the model are in good agreement.

The Mixed Continuous/Discrete-Time (MCDT) effective capacity model was developed in Chapter 5 and it formulates the CCDF of end-to-end delay, average delay and jitter in wireless time-slotted communication systems. One distinct attribute of such systems is that the systems transmits packets every TTI so they are better represented as MCDT models (discrete-time models with consideration of slot time). The conventional effective capacity model was developed based on continuous-time models, making themselves not suitable for MCDT models. The new MCDT effective capacity model characterise the CCDF of packet end-to-end delay by two functions, namely the probability of non-empty buffer and the success probability of a connection. A cross-layer (link layer and physical layer) discrete-time simulator was built and the capacity of the physical layer is based on Shannon's capacity. The empirical CCDFs and the fitted CCDFs of end-to-end delay based on the MCDT model are in good agreement, which is not the case for fitted CCDFs based on the conventional effective capacity model.

In Chapter 6, a realistic cross-layer simulation platform is built using Simulink and it integrates the IEEE 802.16-2004 physical layer and a simplified link layer that supports the multi-hop architecture. Such a simulation platform is used to validate the multi-hop effective capacity model and the MCDT effective capacity model in a more practical context. Delay performances from the simulation and estimation results were in good agreement, indicating that it may be appropriate to use the multi-hop effective capacity model and MCDT effective capacity model to characterise delay performances for multi-hop wireless communication systems.

One critical factor in tele-ultrasonography medical services is the end-to-end frame delay. Chapter 7 showed an example of using the effective capacity technique to estimate delay distributions in wireless tele-ultrasonography medical systems. The cross-layer simulation platform developed in Chapter 6 was further improved to enable deterministic bits transmission. Furthermore, a real ultrasound medical video from a portable ultrasound machine is MPEG-2 encoded and is transmitted through the platform. The simulation results showed that in most cases the estimation results are close to simulation results in terms of delay distributions, indicating it is appropriate to use the MCDT EB-EC model to estimate end-to-end delay performances for wireless medical video streaming.

8.2 Future Work

In this section, future research directions are pointed out.

8.2.1 Theoretical Advancements of the Effective Capacity Model

As indicated in Fig. 8.1, two aspects of the effective capacity model that have not been developed theoretically are the MCDT EB-EC model and the multi-hop MCDT EB-EC model. These two models are sketched in Sections 8.2.1.1 and 8.2.1.2, respectively.

8.2.1.1 MCDT Effective Bandwidth-Effective Capacity model

In Chapter 5, it is formally shown that in a time-slotted wireless communication system, if the limits of the asymptotic logarithmic moment generating function of variable arrival rate $A'(n;T_s)$

$$\Lambda_A(\theta;T_s) = \lim_{n\to\infty} \frac{1}{n} \log E\left[e^{\theta \sum_{i=1}^{n} A'(i;T_s)}\right] \tag{8.1}$$

exists and the asymptotic logarithmic moment generating function of negative variable service rate $-\tilde{S}'(n;T_s)$

$$\Lambda_{\tilde{S}}(\theta;T_s) = \lim_{n\to\infty} \frac{1}{n} \log E\left[e^{-\theta \sum_{i=1}^{n} \tilde{S}'(i;T_s)}\right] \tag{8.2}$$

exists, the CCDF of end-to-end delay in single-hop MCDT EB-EC model may be approximated by

$$\begin{gathered}\Pr(D(\infty) \geq kT_s) \approx \gamma(\mu;T_s)(1-p(\mu;T_s))^k, \\ k \in \{0,1,2,\cdots\}\end{gathered} \tag{8.3}$$

This model may be validated via simulations using cross-layer (physical layer and link layer) simulators in which the physical-layer capacity is based on

- the Shannon' capacity or
- the IEEE 802.16-2004 standard (see Chapter 6)).

8.2.1.2 Multi-hop MCDT EB-EC model

In an H-hop scenario, nodes are ordered by the sequence in which a packet traverses from the source to the sink, and are numbered from 1 to H (same as the system model of Fig. 4.1). Assume that the nodes in the routing path are time-synchronised and the slot time T_s in each node is the same. From (8.3), the Probability Mass Function (PMF) of queueing delay at i-th ($1 \leq i \leq H$) node is expressed as

$$p_i(k) = \Pr(D_i = kT_s) = \gamma_i p_i (1-p_i)^{k-1} u(k-1) + (1-\gamma_i)\delta(k) \tag{8.4}$$

Further assume that 1) the delay experienced by a specific packet in any nodes are independent and 2) in each node, the limits of the asymptotic logarithmic moment generating functions of variable arrival rate $\Lambda_A(\theta;T_s)$ and variable service rate $\Lambda_{\tilde{S}}(-\theta;T_s)$

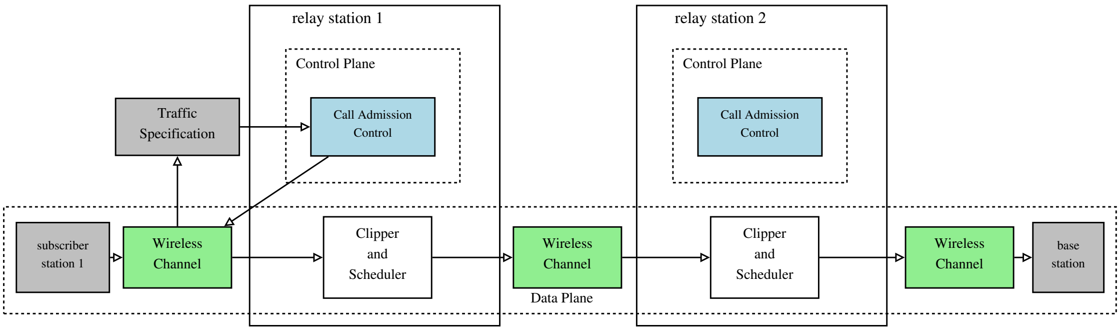


Figure 8.2: Network architecture for QoS provisioning in multi-hop wireless networks

exist. The H-hop CCDF is approximated by

$$\Pr(\sum_{i=1}^{H} D_i > kT_s) = 1 - \sum_{k=0}^{n} p_1(k) * p_2(k) \cdots p_H(k) \tag{8.5}$$

where "*" stands for convolution.

This model may be validated via simulations using cross-layer (physical layer and link layer) simulators in which the physical-layer capacity is based on

- the Shannon' capacity or
- the IEEE 802.16-2004 standard (see Chapter 6)).

8.2.2 Qos Provisioning in Multi-hop Wireless Networks

Fig. 8.2 illustrates an example of the *network-centric QoS provisioning architecture* in a 3-hop communication link. In such communication networks, functions of handling QoS can be broadly divided into two planes, namely a control plane (controls the status of every call session) and a data plane (deals with data which are physically stored in the buffer). Components within these two planes are introduced as follows:

1. **Control plane**

   - *call admission controller*: Each time a call setup request is initiated, the call admission controller determines whether or not the new call can be accepted while guaranteeing the QoS of established calls. In some service models, the call admission controller is also responsible for computing and allocating an equivalent bandwidth and buffer based on the traffic specification of the call;

2. **Data plane**

   - *Clipper*: The clipper decides which packet to drop when congestion occurs;
   - *Scheduler*: The scheduler decides the order of packet transmission.

Latency-based scheduler, clipper and call admission controller have been extensively studied in the literature [163, 164, 165, 166] and they use the concept of *maximum delay bound*. It is natural to utilise the simple estimators developed in this thesis for these components in multi-hop wireless networks. Such an area requires studies and would be expected to lead to optimised multi-hop network design.

Overall, this thesis addressed mathematical development, modelling and simulation-based verification of the important QoS metric of delay, in wireless networks. It is hoped that the work described in this thesis will be helpful to future wireless network designer and operators in a move to satisfy the ever increasing hunger for bandwidth with the ever increasing demand of "excellent quality".

Appendix A

Appendix

A.1 Proof of Proposition 4.3.2

$$
\begin{aligned}
&\Pr(\sum_{h=1}^{H} D_h \leq z) = \Pr(\sum_{h=1}^{H-1} D_h + D_H \leq z) \\
&= \iint\limits_{x+y\leq z} f_{\sum_{h=1}^{H}}(x,y)\,dx\,dy \\
&(\sum_{h=1}^{H-1} D_h \text{ and } D_H \text{ are 2 } r.v., \text{ and are denoted as } X \text{ and } Y) \\
&= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{z-y} f_{\sum_{h=1}^{H}}(x,y)\,dx \right] dy \\
&= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{z} f_{\sum_{h=1}^{H}}(t-y,y)\,dt \right] dy (\text{ let } x = t-y) \\
&= \int_{-\infty}^{z} \left[\int_{-\infty}^{+\infty} f_{\sum_{h=1}^{H}}(t-y,y)\,dy \right] dt \\
&= \int_{-\infty}^{z} f_{\sum_{h=1}^{H-1}}(t) * f_H(t)\,dt (\text{ Independence Assumption }) \\
&= \cdots = \int_{0}^{z} f_1(t) * f_2(t) ... * f_H(t)\,dt
\end{aligned}
\tag{A.1}
$$

By substituting z with x and using the property that $\Pr(\sum_{h=1}^{H} D_h > x) = 1 - \Pr(\sum_{h=1}^{H} D_h \leq x)$, we have proposition 3.2. ■

A.2 Proof of Proposition 4.3.3

The *mathematical induction* is used to prove (4.10)) is true. When $h = 1$,

$$\Pr(\sum_{h=1}^{1} D_h > x) = \Pr(D_1 > x) = \gamma_1 \exp(-\theta_1 x) \tag{A.2}$$

which is the same as (4.3).

Next we want to show that if $\Pr(\sum_{h=1}^{H} D_h > x)$ is true for an arbitrary hop number, H, $\Pr(\sum_{h=1}^{H+1} D_h > x)$ is also true. It is equivalent to prove $\Pr(\sum_{h=1}^{H} D_h \leq x)$ for all H greater than 1 is true, which is more clearer to demonstrate in this case. Thus, assume that for an arbitrary hop number, H, $\Pr(\sum_{h=1}^{H} D_h \leq x)$ is also true. We then derive $\Pr(\sum_{h=1}^{H+1} D_h \leq x)$ by using Proposition 4.3.2. Let $F_h(x) = \Pr(\sum_{h=1}^{H} D_h \leq x)$

$$\begin{aligned}
&\Pr(\sum_{h=1}^{H+1} D_h \leq x) = F_{H+1}(x) \\
&= \int_{-\infty}^{x} f_{H+1}(t) * f_{\sum_{h=1}^{H}}(t)dt = f_{H+1}(t) * \int_{-\infty}^{x} f_{\sum_{h=1}^{H}}(t)dt = f_{H+1}(t) * F_h(t) \\
&= (1-\gamma_{H+1})F_h(x) + \int_{0}^{x} F_h(x-\tau)\gamma_{H+1}\theta_{H+1}e^{-\theta_{H+1}\tau}d\tau = (1-\gamma_{H+1})F_h(x) + \\
&\int_0^x \left(1 - \sum_{h=1}^{H}\left[\prod_{i=1,j\neq h}^{H}(1+\frac{\gamma_j\theta_i}{\theta_j-\theta_i})\right]\gamma_i e^{-\theta_i(x-\tau)}\right)\cdot\left(\gamma_{H+1}\theta_{H+1}e^{-\theta_{H+1}\tau}\right)d\tau \\
&= (1-\gamma_{H+1})F_h(x) + \frac{\gamma_{H+1}\theta_{H+1}}{-\theta_{H+1}}e^{-\theta_{H+1}\tau}\Big|_0^x - \\
&\sum_{h=1}^{H}\left[\prod_{i=1,j\neq h}^{H}(1+\frac{\gamma_j\theta_i}{\theta_j-\theta_i}\gamma_{H+1}\theta_{H+1}\gamma_i)\right]\frac{e^{-\theta_i x}}{\theta_{H+1}-\theta_i}\cdot e^{-(\theta_{H+1}-\theta_i)\tau}\Big|_0^x \\
&= (1-\gamma_{H+1})F_h(x) + \gamma_{H+1}(1-e^{-\theta_{H+1}x}) + \\
&\sum_{h=1}^{H}\left[\prod_{i=1,j\neq h}^{H}(1+\frac{\gamma_j\theta_i}{\theta_j-\theta_i}\gamma_{H+1}\theta_{H+1}\gamma_i)\right]\frac{(e^{-\theta_{H+1}x}-e^{-\theta_i x})}{\theta_{H+1}-\theta_i} \\
&= 1 - \sum_{h=1}^{H+1}\left[\prod_{i=1,j\neq h}^{H+1}(1+\frac{\gamma_j\theta_i}{\theta_j-\theta_i})\right]\gamma_i e^{-\theta_i x}
\end{aligned} \tag{A.3}$$

which means $F_{H+1}(x)$ holds.

Therefore, $F_H(x)$ and (4.10) are true for all H starting with 1. ■

A.3 Proof of Corollary 4.3.4

$$E[D] = E\left[\sum_{h=1}^{H} D_h\right] = \sum_{h=1}^{H} E[D_h] = \sum_{h=1}^{H} \int_{-\infty}^{+\infty} t f_h(t) dt = \sum_{h=1}^{H} \frac{\gamma_h}{\theta_h} \tag{A.4}$$

$$\sigma = \sqrt{\mathrm{Var}\left(\sum_{h=1}^{H} D_h\right)} = \sqrt{\sum_{h=1}^{H} \mathrm{Var}\left(D_h\right)} = \sqrt{\sum_{h=1}^{H} (E[D_h^2] - E^2[D_h])} \tag{A.5}$$

$$= \sqrt{\sum_{h=1}^{H} \left(\int_0^{\infty} t^2 f_h(t) dt - \left(\frac{\gamma_h}{\theta_h}\right)^2\right)} = \sqrt{\sum_{h=1}^{H} \left(\frac{2\gamma_h}{\theta_h^2} - \left(\frac{\gamma_h}{\theta_h}\right)^2\right)} \tag{A.6}$$

■

A.4 Proof of Proposition 5.3.1

From (5.19), the PMF of queueing delay is expressed as

$$p(k) = \Pr(D = kT_s) = \gamma_m p_m (1 - p_m)^{k-1} u(k-1) + (1 - \gamma_m)\delta(k) \tag{A.7}$$

Suppose Ω is the sample space of D and let X be a function defined as

$$X(\omega) = \omega / T_s \tag{A.8}$$

then, X is the normalised discrete random variable of D and its sample space is $\{0, 1, 2, \dots\}$

Moreover, the mean and standard deviation of X and mean and standard deviation of D have the following relations.

$$\mu_X = \mu_D / T_s \tag{A.9}$$

$$\sigma_X = \sigma_D / T_s \tag{A.10}$$

By using a standard result, i.e.,

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x} \quad \text{when } |x| < 1$$

The Probability Generating Function (PGF) of X is given by

$$\begin{aligned}\Pi_X(s) = E[s^X] &= \sum_{k=0}^{\infty} s^k p(k) \\ &= -\gamma_m + \frac{\gamma_m p_m}{(1-p_m)} \cdot \frac{1}{1-(1-p_m)s}\end{aligned}$$

According to the behaviour of PGF, the normalised mean and jitter of X are given

$$\mu_X = E[X] = \Pi'_X(1) = \frac{\gamma_m}{p_m} \tag{A.11}$$

$$\begin{aligned}\sigma_X &= \sqrt{E[X-\mu]^2} = \sqrt{\Pi''_X(1) + \mu_X - \mu_X^2} \\ &= \sqrt{\left(\frac{2(1-p_m)\gamma_m}{p_m^2} + \frac{\gamma_m}{p_m} - \left(\frac{\gamma_m}{p_m}^2\right)\right)}\end{aligned} \tag{A.12}$$

By substituting μ_X in (A.9) with (A.11) and σ_X in (A.10) with (A.12), we have Proposition 5.3.1. ■

References

[1] E. Seidel, “White paper on lte advance,” Nomor Research, Tech. Rep., July 2008.

[2] “Ieee standard for local and metropolitan area networks part 16: Air interface for broadband wireless access systems amendment 1: Multihop relay specification,” *IEEE Std 802.16j-2009 (Amendment to IEEE Std 802.16-2009)*, pp. 1–290, 2009.

[3] G. Hiertz, D. Denteneer, S. Max, R. Taori, J. Cardona, L. Berlemann, and B. Walke, “Ieee 802.11s: The wlan mesh standard,” *IEEE Wireless Communications*, pp. 104–111, Feb 2010. [Online]. Available: http://www.comnets.rwth-aachen.de

[4] D. Wu and R. Negi, “Effective capacity: a wireless link model for support of quality of service,” *Wireless Communications, IEEE Transactions on*, vol. 2, no. 4, pp. 630 – 643, July 2003.

[5] D. V. Lindley, “The theory of queues with a single server,” *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 48, pp. 277–289, 1952.

[6] N. Olifer and V. Olifer, *Computer Networks: Principles,Technologies And Protocols For Network Design*. Wiley India Pvt. Limited, 2006.

[7] A. Tanenbaum, *Computer networks*, ser. Computer Science. Prentice Hall PTR, 2003.

[8] J. F. Kurose and K. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

[9] P. Mach and R. Bešt'ák, "Wireless mesh and relay networks," in *Research in Telecommunication Technology 2006*, 2006.

[10] R. Bruno, M. Conti, and E. Gregori, "Mesh networks: commodity multihop ad hoc networks," *Communications Magazine, IEEE*, vol. 43, no. 3, pp. 123 – 131, march 2005.

[11] A. Sadek, W. Su, and K. Liu, "Multinode cooperative communications in wireless networks," *Signal Processing, IEEE Transactions on*, vol. 55, no. 1, pp. 341–355, 2007.

[12] J. Mitola and J. Maguire, G.Q., "Cognitive radio: making software radios more personal," *Personal Communications, IEEE*, vol. 6, no. 4, pp. 13–18, 1999.

[13] C. Masouros and T. Ratnarajah, "Utilization of primary-secondary cross-interference via adaptive precoding in cognitive relay assisted mimo wireless systems," in *Communications (ICC), 2011 IEEE International Conference on*, 2011, pp. 1–5.

[14] Y. Zhang, J. Luo, and H. Hu, *Wireless Mesh Networking: Architectures, Protocols and Standards*, 1st ed. Auerbach Publications, Dec. 2006.

[15] I. Haratcherev and A. Conte, "Practical energy-saving in 3g femtocells," in *Communications Workshops (ICC), 2013 IEEE International Conference on*, 2013, pp. 602–606.

[16] P. Piggin, B. Lewis, and P. Whitehead, *Mesh networks in Fixed Broadband Wireless Access.* Radiant Networks PLC, Essex UK, 2003.

[17] S. Bellofiore, J. Foutz, R. Govindaradjula, I. Bahceci, C. A. Balanis, A. S. Spanias, J. M. Capone, and T. M. Duman, "Smart antenna system analysis, integration and performance for mobile ad hoc networks (manets)," 2005.

[18] I. F. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: a survey," *Computer Networks (Amsterdam, Netherlands: 1999)*, vol. 47, no. 4, pp. 445–487, Mar. 2005.

[19] (2014) Network Time Synchronization Bibliography. http://www.eecis.udel.edu/m̃ills/biblio.html.

[20] . Sonia, . Raouf, . Youssef, and . Brent, "Routing protocols in wireless mesh networks: challenges and design considerations," 2006.

[21] M. Iwamura, H. Takahashi, and S. Nagata, "Relay technology in lte-advanced," *NTT DoCoMo Technical Journal*, vol. 12, no. 2, pp. 29–36, 2010.

[22] "3GPP R1-082975: Application scenarios for LTE-Advanced relay," China Mobile, Vodafone, Huawei, Tech. Rep., 2008.

[23] WiMAX.com. (2010) Wimax reaches "mass-market," devices to triple in 2010. Online, Available: http://www.wimax.com/wimax/wimax-reaches-qmass-marketq-devices-to-triple-in-2010.

[24] M. Ma, *Current Technology Developments of WiMax Systems.* Springer, 2009.

[25] V. Genc, S. Murphy, and J. Murphy, "Performance analysis of transparent relays in 802.16j mmr networks," in *Proc. WiOPT 2008*, 2008, pp. 273–281.

[26] J. Sydir, "Harmonized contribution on 802.16j (mobile multihop relay) usage models," July 2006.

[27] B. G. Lee and S. Choi, *Broadband Wireless Access and Local Networks: Mobile WiMAX and WiFi*, 1st ed. Artech House Publishers, May 2008.

[28] M. Woodward, *Communication and computer networks: modelling with discrete-time queues.* Pentech, 1993.

[29] G. T. . V8.3.1, “Technical specification, policy and charging control architecture (release 8),” Tech. Rep.

[30] WiMAX, “Mobile wimax – part i: A technical overview and performance evaluation,” 2006.

[31] Q. Ni, L. Romdhani, and T. Turletti, “A survey of qos enhancements for ieee 802.11 wireless lan,” *Wireless Communications and Mobile Computing*, vol. 4, no. 5, pp. 547–566, August 2004.

[32] D. G. Kendall, “Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain,” *The Annals of Mathematical Statistics*, vol. 24, no. 3, pp. 338–354, 1953.

[33] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*, 4th ed. New York, NY, USA: Wiley-Interscience, 2008.

[34] A. K. Erlang, “The Theory of Probabilities and Telephone Conversations,” *Nyt Tidsskrift for Matematik*, vol. 20, no. B, pp. 33–39, 1909.

[35] R. Cruz, “A calculus for network delay. i. network elements in isolation,” *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 114 –131, Jan 1991.

[36] R. L. Cruz, “A calculus for network delay. ii. network analysis,” vol. 37, no. 1, pp. 132–141, 1991.

[37] J. L. Boudec and P. Thiran, *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*, 1st ed. Springer, Aug. 2001.

[38] D. Bertsekas and R. Gallager, *Data Networks.* Prentice-Hall, 1987.

[39] T. Athanasiadis, Y. Avrithis, and S. Kollias, “The atm forum. traffic management specification version 4.0 (1996),” Tech. Rep., 1996.

[40] R. Guérin and V. Peris, *Quality-of-service in Packet Networks: Basic Mechanisms and Directions.* IBM Watson Research Center, 1998.

[41] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, “Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification,” RFC 2205 (Proposed Standard), Internet Engineering Task Force, Sep. 1997, updated by RFCs 2750, 3936, 4495, 5946.

[42] C. Chang, *Performance Guarantees in Communication Networks*, 1st ed. Springer, Apr. 2000.

[43] A. Burchard, J. Liebeherr, and S. D. Patek, “A min-plus calculus for end-to-end statistical service guarantees,” vol. 52, no. 9, pp. 4105–4114, 2006.

[44] K. Angrishi and U. Killat, “Analysis of a real-time network using statistical network calculus with effective bandwidth and effective capacity,” pp. 1–15, 2008, measuring, Modelling and Evaluation of Computer and Communication Systems (MMB), 2008 14th GI/ITG Conference.

[45] S. Akin and M. C. Gursoy, “Effective capacity analysis of cognitive radio channels for quality of service provisioning,” vol. 9, no. 11, pp. 3354–3364, 2010.

[46] A. Balasubramanian and S. Miller, “The effective capacity of a time division downlink scheduling system,” *Communications, IEEE Transactions on*, vol. 58, no. 1, pp. 73–78, 2010.

[47] E. A. Jorswieck, R. Mochaourab, and M. Mittelbach, “Effective capacity maximization in multi-antenna channels with covariance feedback,” vol. 9, no. 10, pp. 2988–2993, 2010.

[48] Q. Wang, D. Wu, and P. Fan, “Effective capacity of a correlated rayleigh fading channel,” *Wireless Communications and Mobile Computing*, vol. 11, no. 11, pp. 1485–1494, Nov. 2011.

[49] J. S. Harsini and M. Zorzi, “Effective capacity for multi-rate relay channels with delay constraint exploiting adaptive cooperative diversity,” vol. 11, no. 9, pp. 3136–3147, 2012.

[50] L. Musavian and S. Aissa, “Quality-of-service based power allocation in spectrum-sharing channels,” in *Proc. IEEE Global Telecommunications Conf. IEEE GLOBECOM 2008*, 2008, pp. 1–5.

[51] ——, “Adaptive modulation in spectrum-sharing systems with delay constraints,” in *Proc. IEEE Int. Conf. Communications ICC ’09*, 2009, pp. 1–5.

[52] ——, “Cross-layer analysis of cognitive radio relay networks under quality of service constraints,” in *Proc. IEEE 69th Vehicular Technology Conf. VTC Spring 2009*, 2009, pp. 1–5.

[53] ——, “Effective capacity of delay-constrained cognitive radio in nakagami fading channels,” vol. 9, no. 3, pp. 1054–1062, 2010.

[54] L. Musavian, S. Aissa, and S. Lambotharan, “Adaptive modulation in spectrum-sharing channels under delay quality-of-service constraints,” vol. 60, no. 3, pp. 901–911, 2011.

[55] J. Tang and X. Zhang, “Quality-of-service driven power and rate adaptation for multichannel communications over wireless links,” vol. 6, no. 12, pp. 4349–4360, 2007.

[56] X. Li, X. Dong, and D. Wu, “On optimal power control for delay-constrained communication over fading channels,” vol. 57, no. 6, pp. 3371–3389, 2011.

[57] D. Wu and R. Negi, “Downlink scheduling in a cellular network for quality-of-service assurance,” vol. 53, no. 5, pp. 1547–1557, 2004.

[58] ——, "Effective capacity channel model for frequency-selective fading channels," in *Proc. Second Int Quality of Service in Heterogeneous Wired/Wireless Networks Conf*, 2005.

[59] ——, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," vol. 54, no. 3, pp. 1198–1206, 2005.

[60] S. Ren and K. B. Letaief, "Optimal effective capacity for cooperative relay networks with qos guarantees," in *Proc. IEEE Int. Conf. Communications ICC '08*, 2008, pp. 3725–3729.

[61] A. Abdrabou and W. Zhuang, "Stochastic delay guarantees and statistical call admission control for ieee 802.11 single-hop ad hoc networks," vol. 7, no. 10, pp. 3972–3981, 2008.

[62] D. Wu and R. Negi, "Power control and scheduling for guaranteeing quality of service in cellular networks," *Wireless Communications and Mobile*, no. 352, 2008.

[63] G. Femenias, J. Ramis, and L. Carrasco, "Using two-dimensional markov models and the effective-capacity approach for cross-layer design in amc/arq-based wireless networks," vol. 58, no. 8, pp. 4193–4203, 2009.

[64] Q. Wang, D. O. Wu, and P. Fan, "Delay-constrained optimal link scheduling in wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 9, pp. 4564–4577, Nov. 2010.

[65] Y. Chen and I. Darwazeh, "End-to-end delay performance analysis in ieee 802.16j mobile multi-hop relay (mmr) networks," pp. 488–492, 2011, telecommunications (ICT), 2011 18th International Conference on.

[66] X. Li, *Radio Access Network Dimensioning for 3g Umts*, ser. Advanced Studies Mobile Research Center Bremen. Vieweg+Teubner Verlag, 2011.

[67] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer, Mar. 1998.

[68] Z.-L. Zhang, D. Towsley, and J. Kurose, "Statistical analysis of the generalized processor sharing scheduling discipline," vol. 13, no. 6, pp. 1071–1080, 1995.

[69] G. Choudhury, D. Lucantoni, and W. Whitt, "Squeezing the most out of atm," *Communications, IEEE Transactions on*, vol. 44, no. 2, pp. 203 –217, Feb 1996.

[70] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," vol. 9, no. 7, pp. 968–981, 1991.

[71] F. P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Syst.*, vol. 9, no. 1-2, pp. 5–15, 1991.

[72] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type uas channel," *Queueing Syst. Theory Appl.*, vol. 9, no. 1-2, pp. 17–28, Oct. 1991.

[73] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," vol. 39, no. 5, pp. 913–931, 1994.

[74] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass markov fluids and other atm sources," vol. 1, no. 4, pp. 424–428, 1993.

[75] C.-S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin, "Effective bandwidth and fast simulation of atm intree networks," in *Proceedings of the 16th IFIP Working Group 7.3 international symposium on Computer performance modeling measurement and evaluation*, ser. Performance '93. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1994, pp. 45–65.

[76] A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," vol. 1, no. 3, pp. 329–343, 1993.

[77] W. Whitt, “Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues,” *Telecommunication Systems*, vol. 2, pp. 71–107, 1993, 10.1007/BF02109851.

[78] P. W. Glynn and W. Whitt, “Logarithmic asymptotics for steady-state tail probabilities in a single-server queue,” *Journal of Applied Probability*, vol. 31, pp. pp. 131–156, 1994.

[79] X. Yu, I. L.-J. Thng, and Y. Jiang, “Measurement-based effective bandwidth estimation for long range dependent traffic,” in *Proc. IEEE Region 10 Int. Conf. TENCON Electrical and Electronic Technology*, vol. 1, 2001, pp. 359–365.

[80] C. Li, A. Burchard, and J. Liebeherr, “A network calculus with effective bandwidth,” vol. 15, no. 6, pp. 1442–1453, 2007.

[81] A. W. Berger and W. Whitt, “Extending the effective bandwidth concept to networks with priority classes,” *IEEE Communications Magazine*, vol. 36, no. 8, pp. 78–83, 1998.

[82] C.-S. Chang and T. Zajic, “Effective bandwidths of departure processes from queues with time varying capacities,” in *INFOCOM ’95. Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Bringing Information to People. Proceedings. IEEE*, apr 1995, pp. 1001 –1009 vol.3.

[83] F. Ishizaki and G. U. Hwang, “Cross-layer design and analysis of wireless networks using the effective bandwidth function,” vol. 6, no. 9, pp. 3214–3219, 2007.

[84] C. Ortiz, J.-F. Frigon, B. Sanso, and A. Girard, “Effective bandwidth evaluation for voip applications in ieee 802.11 networks,” pp. 926–931, 2008, wireless Communications and Mobile Computing Conference, 2008. IWCMC ’08. International.

[85] W. Willinger, M. Taqqu, W. E. Leland, and D. V. Wilson, "Selfsimilarity in high-speed packet traffic: Analysis and modeling of ethernet traffic measurements," *Statistical Sci*, vol. 10, pp. 67–85, 1995.

[86] V. Paxson and S. Floyd, "Wide area traffic: the failure of poisson modeling," *Networking, IEEE/ACM Transactions on*, vol. 3, no. 3, pp. 226 –244, jun 1995.

[87] D. P. Heyman, "Sizing backbone internet links," *Oper. Res.*, vol. 53, no. 4, pp. 575–585, Jul. 2005.

[88] R. Adler, R. Feldman, and M. Taqqu, *A Practical Guide to Heavy Tails: Statistical Techniques and Applications.* Birkhäuser Boston, 1998.

[89] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "Internet Traffic Tends *Toward* Poisson and Independent as the Load Increases," in *Nonlinear Estimation and Classification*, C. Holmes, D. Denison, M. Hansen, B. Yu, and B. Mallick, Eds. New York: Springer, 2002, pp. 83–109.

[90] P. Heegaard, B. Helvik, and R. Andreassen, "Application of rare event techniques to trace driven simulation," in *Simulation Conference, 2005 Proceedings of the Winter*, DEC. 2005, p. 10.

[91] C. Park, F. Hernndez-Campos, J. S. Marron, and F. D. Smith, "Long-range dependence in a changing internet traffic mix," *Computer Networks*, vol. 48, no. 3, pp. 401–422, 2005.

[92] J.-H. Kim, H.-J. Lee, S.-M. Oh, and S.-H. Cho, "Performance modeling and evaluation of data/voice services in wireless networks," *Wirel. Netw.*, vol. 14, no. 2, pp. 233–246, Mar. 2008.

[93] "Umass trace repository," http://traces.cs.umass.edu/index.php/Network/Network, 2004.

[94] Y. Chen and I. Darwazeh, "Quality of service (qos) analysis of an internet traffic trace over gbps ethernet," in *Telecommunications (ICT), 2011 18th International Conference on*, 2013.

[95] H. Cramér, *Mathematical Methods of Statistics*, ser. Princeton landmarks in mathematics and physics. Princeton University Press, 1999.

[96] K. Thompson, G. J. Miller, and R. Wilder, "Wide-area internet traffic patterns and characteristics," vol. 11, no. 6, pp. 10–23, 1997.

[97] R. M. Loynes, "On the waiting-time distribution for queues in series," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 27, no. 3, pp. pp. 491–496, 1965.

[98] J. R. Jackson, "Networks of waiting lines," *Operations Research*, vol. 5, no. 4, pp. 518–521, 1957.

[99] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, "Open, closed, and mixed networks of queues with different classes of customers," *J. ACM*, vol. 22, pp. 248–260, April 1975.

[100] M. Reiser and S. S. Lavenberg, "Mean-value analysis of closed multichain queuing networks," *J. ACM*, vol. 27, pp. 313–322, April 1980.

[101] N. Bisnik and A. A. Abouzeid, "Queuing network models for delay analysis of multihop wireless ad hoc networks," *Ad Hoc Networks*, vol. 7, no. 1, 2009.

[102] M. Xie and M. Haenggi, "Towards an end-to-end delay analysis of wireless multihop networks," *Ad Hoc Networks*, vol. 7, no. 5, 2009.

[103] Y. Q. L. Y. J. S. Jiang, Y., "Fundamental calculus on generalized stochastically bounded bursty traffic for communication networks," *Computer Networks*, vol. 53, no. 12, pp. 2011–2021, 2009, cited By (since 1996) 3.

[104] J. Kurose, “On computing per-session performance bounds in high-speed multi-hop computer networks,” in *Proceedings of the 1992 ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, ser. SIGMETRICS ’92/PERFORMANCE ’92. New York, NY, USA: ACM, 1992, pp. 128–139.

[105] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn, “Effective envelopes: statistical bounds on multiplexed traffic in packet networks,” in *Proc. IEEE Nineteenth Annual Joint Conf. of the IEEE Computer and Communications Societies INFOCOM 2000*, vol. 3, 2000, pp. 1223–1232.

[106] P. Jurcík, R. Severino, A. Koubaa, M. Alves, and E. Tovar, “Real-time communications over cluster-tree sensor networks with mobile sink behaviour.” IEEE Computer Society, 2008.

[107] F. Ciucu, A. Burchard, and J. Liebeherr, “Scaling properties of statistical end-to-end bounds in the network calculus,” vol. 52, no. 6, pp. 2300–2312, 2006.

[108] A. Burchard, J. Liebeherr, and F. Ciucu, “On q(h log h) scaling of network delays,” in *Proc. INFOCOM 2007. 26th IEEE Int. Conf. Computer Communications. IEEE*, 2007, pp. 1866–1874.

[109] M. Fidler, “An end-to-end probabilistic network calculus with moment generating functions,” in *Proc. 14th IEEE Int. Workshop Quality of Service IWQoS 2006*, 2006, pp. 261–270.

[110] Y. Chen, Y. Yang, and I. Darwazeh, “A cross-layer analytical model of end-to-end delay performance for wireless multi-hop environments,” in *Proc. IEEE Global Telecommunications Conf. GLOBECOM 2010*, 2010, pp. 1–6.

[111] C. Berrou, A. Glavieux, and P. Thitimajshima, “Near Shannon limit error-correcting coding and decoding: Turbo-codes,” in *Proceedings IEEE International*

Conference on Communications, vol. 2. Geneva, Switzerland: IEEE, May 1993, pp. 1064–1070.

[112] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: turbo-codes," vol. 44, no. 10, pp. 1261–1271, 1996.

[113] M. Schwartz, *Mobile Wireless Communications.* Cambridge University Press, 2005.

[114] Y. Chen, J. Chen, and Y. Yang, "Multi-hop delay performance in wireless mesh networks," *MONET*, vol. 13, no. 1-2, 2008.

[115] H. Kobayashi and A. Konheim, "Queueing models for computer communications system analysis," vol. 25, no. 1, pp. 2–29, 1977.

[116] H. Takagi, *Queueing Analysis: Discrete-time systems*, ser. Queueing Analysis: A Foundation of Performance Evaluation. North-Holland, 1993.

[117] D. Cox, "The statistical analysis of congestion," *Journal of the Royal Statistical Society*, vol. 118, pp. 324–335, 1955.

[118] M. L. Chaudhry and U. C. Gupta, "Queue-length and waiting-time distributions of discrete-time gi^{X}/geom/1 queueing systems with early and late arrivals," *Queueing Syst.*, vol. 25, no. 1-4, pp. 307–324, 1997.

[119] S. H. Chang and D. W. Choi, "Performance analysis of a finite-buffer discrete-time queue with bulk arrival, bulk service and vacations," *Comput. Oper. Res.*, vol. 32, pp. 2213–2234, September 2005.

[120] N. Tian and Z. G. Zhang, "The discrete-time gi/geo/1 queue with multiple vacations," *Queueing Syst. Theory Appl.*, vol. 40, pp. 283–294, Apr. 2002.

[121] Y. Chen and I. Darwazeh, "Effective capacity (EC) model in fixed-length packet-switching systems," 2011.

[122] P. L. Conti, “Large sample bayesian analysis for geo/g/1 discrete-time queueing models,” *The Annals of Statistics*, vol. 27, no. 6, pp. pp. 1785–1807, 1999.

[123] ——, “Bootstrap approximations for bayesian analysis of geo/g/1 discrete-time queueing models,” *Journal of Statistical Planning and Inference*, vol. 120, no. 12, pp. 65 – 84, 2004.

[124] M. Bladt and M. Srensen, “Statistical inference for discretely observed markov jump processes,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 67, no. 3, pp. pp. 395–410, 2005.

[125] K. Rovers, J. Kuper, M. van de Burgwal, A. Kokkeler, and G. Smit, “Mixed continuous/discrete time modelling with exact time adjustments,” in *Wireless Communications and Mobile Computing Conference (IWCMC), 2011 7th International*, july 2011, pp. 1111 –1116.

[126] Y. Chen and I. Darwazeh, “Poster abstract: Effective capacity model in the discrete time domain,” in *EWSN 2012*, 2012.

[127] ——, “Mixed continuous/discrete time effective capacity model for wireless slotted communication systems,” *the IEEE transaction on Wireless Communications (submitted)*, 2013.

[128] F. Muhammad, *An Introduction to Umts Technology: Testing, Specifications and Standard Bodies for Engineers and Managers.* Brown Walker Press, 2008.

[129] J. Walrand, *An introduction to queueing networks.* Prentice Hall, 1988.

[130] J. D. C. Little, “A proof of the queuing formula: L=λW,” *Operations Research*, vol. 9 (3), p. 383387, 1961.

[131] R. Dhaou, V. Gauthier, M. I. Tiado, M. Becker, and A.-L. Beylot, “Cross layer simulation: Application to performance modelling of networks composed of manets and satellites.” in *Network Performance Engineering*, ser. Lecture Notes

in Computer Science, D. D. Kouvatsos, Ed. Springer, 2011, vol. 5233, pp. 477–508.

[132] N. Cranley, T. Debnath, and M. Davis, "An experimental investigation of parallel multimedia streams over ieee 802.11e wlan networks using txop," in *Communications, 2007. ICC '07. IEEE International Conference on*, 2007, pp. 1740–1746.

[133] J. Wang, M. Venkatachalam, and Y. Fang, "System architecture and cross-layer optimization of video broadcast over wimax," vol. 25, no. 4, pp. 712–721, 2007.

[134] S. Deb, S. Jaiswal, and K. Nagaraj, "Real-time video multicast in wimax networks," in *Proc. INFOCOM 2008. The 27th Conf. Computer Communications. IEEE*, 2008, pp. 1579–1587.

[135] M. Etoh and T. Yoshimura, "Advances in wireless video delivery," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 111–122, 2005.

[136] Z. Tao, A. Li, K. H. Teo, and J. Zhang, "Frame structure design for ieee 802.16j mobile multihop relay (mmr) networks," in *Proc. IEEE Globecom '07*, 2007, pp. 4301–4306.

[137] K. Pentikousis, J. Pinola, E. Piri, and F. Fitzek, "An experimental investigation of voip and video streaming over fixed wimax," in *Proc. 6th Int. Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops WiOPT 2008*, 2008, pp. 8–15.

[138] R. Fei, K. Yang, S. Ou, S. Zhong, and L. Gao, "A utility-based dynamic bandwidth allocation algorithm with qos guarantee for ieee 802.16j-enabled vehicular networks," in *Proc. SCALCOM-EMBEDDEDCOM'09*, 2009, pp. 200–205.

[139] M. Hu, H. Zhang, T. A. Le, and H. Nguyen, "Performance evaluation of video streaming over mobile wimax networks," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, 2010, pp. 898–902.

[140] L. Al-Jobouri, M. Fleury, and M. Ghanbari, "Cross-layer scheme for wimax video streaming," in *Proc. 3rd Computer Science and Electronic Engineering Conf. (CEEC)*, 2011, pp. 86–91.

[141] Y. Chen and I. Darwazeh, "An estimator for delay distributions in packet-based wireless digital communication systems," in *Wireless Communications and Networking Conference (WCNC)*, 2013.

[142] (2009) Communications blockset: IEEE 802.16-2004 OFDM PHY link, including space-time block coding. http :// www.mathworks.com / products / commblockset / demos.html.

[143] Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos, "Error resilient video coding techniques," vol. 17, no. 4, pp. 61–82, 2000.

[144] A. Kumar, *Mobile broadcasting with WiMAX: principles, technology, and applications*, ser. Focal Press media technology professional. Focal Press, 2008.

[145] the IEEE, "IEEE standard for local and metropolitan area networks part 16: Air interface for fixed and mobile broadband wireless access systems amendment 2: Physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1," Tech. Rep., 2006.

[146] A. Sadka, *Compressed video communications.* John Wiley & Sons, 2002.

[147] I. E. Richardson, *H.264 and MPEG-4 Video Compression: Video Coding for Next Generation Multimedia*, 1st ed. Wiley, Aug. 2003.

[148] Y. Xu and Y. Zhou, "H.264 video communication based refined error concealment schemes," vol. 50, no. 4, pp. 1135–1141, 2004.

[149] (2006) The history of ultrasound: A collection of recollections, articles, interviews and images. www.obgyn.net.

[150] R. S. Istepanian, N. Y. Philip, and M. G. Martini, “Medical qos provision based on reinforcement learning in ultrasound streaming over 3.5 g wireless systems,” *Selected Areas in Communications, IEEE Journal on*, vol. 27, no. 4, pp. 566–574, 2009.

[151] M. G. Martini, R. S. H. Istepanian, M. Mazzotti, and N. Y. Philip, “Robust multilayer control for enhanced wireless telemedical video streaming,” vol. 9, no. 1, pp. 5–16, 2010.

[152] A. Alinejad, N. Y. Philip, and R. S. Istepanian, “Cross-layer ultrasound video streaming over mobile wimax and hsupa networks,” *Information Technology in Biomedicine, IEEE Transactions on*, vol. 16, no. 1, pp. 31–39, 2012.

[153] A. Panayides, M. Pattichis, C. Pattichis, and A. Pitsillides, “A tutorial for emerging wireless medical video transmission systems [wireless corner],” *Antennas and Propagation Magazine, IEEE*, vol. 53, no. 2, pp. 202–213, 2011.

[154] C. J. Debono, B. W. Micallef, N. Y. Philip, A. Alinejad, R. S. H. Istepanian, and N. N. Amso, “Cross-layer design for optimized region of interest of ultrasound video data over mobile WiMAX,” vol. 16, no. 6, pp. 1007–1014, 2012.

[155] A. Panayides, Z. Antoniou, Y. Mylonas, M. Pattichis, A. Pitsillides, and C. Pattichis, “High-resolution, low-delay, and error-resilient medical ultrasound video communication using h. 264/avc over mobile wimax networks,” 2012.

[156] Y.-Y. Tan, N. Philip, and R. S. Istepanian, “Fragility issues of medical video streaming over 802.11 e-wlan m-health environments,” in *Engineering in Medicine and Biology Society, 2006. EMBS’06. 28th Annual International Conference of the IEEE.* IEEE, 2006, pp. 6316–6319.

[157] C. N. Doukas, I. Maglogiannis, and T. Pliakas, “Advanced medical video services through context-aware medical networks,” in *Engineering in Medicine and Bi-*

ology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE. IEEE, 2007, pp. 3074–3077.

[158] A. Panayides, M. Pattichis, and C. Pattichis, "Wireless medical ultrasound video transmission through noisy channels," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE.* IEEE, 2008, pp. 5326–5329.

[159] Y. Chen, N. Philip, R. Istepanian, and I. Darwazeh, "End-to-end delay distributions in wireless tele-ultrasonography medical systems," in *Global Telecommunications Conference (GLOBECOM 2013), 2013 IEEE*, 2013.

[160] R. S. Istepanian, E. Jovanov, and Y. Zhang, "Guest editorial introduction to the special section on m-health: Beyond seamless mobility and global wireless healthcare connectivity," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 8, no. 4, pp. 405–414, 2004.

[161] S. Garawi, R. S. Istepanian, and M. A. Abu-Rgheff, "3g wireless communications for mobile robotic tele-ultrasonography systems," *Communications Magazine, IEEE*, vol. 44, no. 4, pp. 91–96, 2006.

[162] I. T. Union, "Iso/iec 13818-2 mpeg-2," *Geneva, ITU H*, vol. 262, 1995.

[163] D. Ferrari and D. C. Verma, "A scheme for real-time channel establishment in wide-area networks," vol. 8, no. 3, pp. 368–379, 1990.

[164] L. Georgiadis, R. Guerin, and A. Parekh, "Optimal multiplexing on a single link: delay and buffer requirements," vol. 43, no. 5, pp. 1518–1535, 1997.

[165] J. Liebeherr, D. E. Wrege, and D. Ferrari, "Exact admission control for networks with bounded delay services," Charlottesville, VA, USA, Tech. Rep., 1994.

[166] L. R. Dennison and D. Chiou, "Latency-based scheduling and dropping," US patent US 7 626 988 B2, 12, 2009.