Analytical Modelling of Scheduling Schemes under Self-similar Network Traffic

Traffic Modelling and Performance Analysis of Centralized and Distributed Scheduling Schemes

Submitted in accordance with the requirements of the University of Bradford for the degree of Doctor in Philosophy

by

Lei LIU

Department of Computing

School of Computing, Informatics and media

University of Bradford

2010

Abstract

High-speed transmission over contemporary communication networks has drawn many research efforts. Traffic scheduling schemes which play a critical role in managing network transmission have been pervasively studied and widely implemented in various practical communication networks. In a sophisticated communication system, a variety of applications co-exist and require differentiated Quality-of-Service (QoS). Innovative scheduling schemes and hybrid scheduling disciplines which integrate multiple traditional scheduling mechanisms have emerged for QoS differentiation. This study aims to develop novel analytical models for commonly interested scheduling schemes in communication systems under more realistic network traffic and use the models to investigate the issues of design and development of traffic scheduling schemes.

In the open literature, it is commonly recognized that network traffic exhibits self-similar nature, which has serious impact on the performance of communication networks and protocols. To have a deep study of self-similar traffic, the real-world traffic datasets are measured and evaluated in this study. The results reveal that self-similar traffic is a ubiquitous phenomenon in high-speed communication networks and highlight the importance of the developed analytical models under self-similar traffic.

The original analytical models are then developed for the centralized scheduling schemes including the Deficit Round Robin, the hybrid PQGPS which integrates the traditional Priority Queueing (PQ) and Generalized Processor Sharing

(GPS) schemes, and the Automatic Repeat reQuest (ARQ) forward error control discipline in the presence of self-similar traffic.

Most recently, research on the innovative Cognitive Radio (CR) techniques in wireless networks is popular. However, most of the existing analytical models still employ the traditional Poisson traffic to examine the performance of CR involved systems. In addition, few studies have been reported for estimating the residual service left by primary users. Instead, extensive existing studies use an ON/OFF source to model the residual service regardless of the primary traffic. In this thesis, a PQ theory is adopted to investigate and model the possible service left by self-similar primary traffic and derive the queue length distribution of individual secondary users under the distributed spectrum random access protocol.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACK	Acknowledgement
ARQ	Automatic Repeat reQuest
BR	bit-by-bit Round Robin
CR	Cognitive Radio
DCF	Distributed Coordination Function
DDoS	Distributed Denial of Service
DiffServ	Differentiated Service
DIFS	Distributed Inter-Frame Space
DoS	Denial of Service
DRR	Deficit Round Robin
EBA	Empty Buffer Approximation
fBm	fractional Brownian motion
FCFS	First Come First Serve
FIFO	First In First Out
FQ	Fair Queueing
GPS	Generalized Processor Sharing
HOL	Head-of-Line
LAN	Local Area Network
MAC	Medium Access Control
MMPP	Markov Modulated Poisson Process
MIMO	Multi-Input Multi-Output
PGPS	Packet-by-packet Generalized Processor Sharing
PHB	Per-Hop Behaviours

PQ	Priority Queueing
QoS	Quality-of-Service
RR	Round Robin
R/S	Rescaled Range
SCFQ	Self Clock Fair Queueing
SR-ARQ	Selective Repeat Automatic Repeat reQuest
SSSQ	Single-Server-Single-Queue
VBR	Variable Bit Rate
VT	Variance Time
WFQ	Weighted Fair Queueing

Acknowledgements

It is a great pleasure to thank many people who made this thesis possible. Without their help, this beautiful thing will not happen.

First and foremost, my eternal gratitude goes to my supervisor, Dr. Min. With his inspiration, his patience, and his great efforts to explain things clearly, he helped to provide a good basis for the present thesis. Throughout my thesis-writing period, he provided encouragement, sound advice, good company, and lots of excellent ideas. I would have been lost without his guidance.

Dr. Jin is a mentor, and also a warm friend. I wish to express my sincere thanks to his understanding, encouragement, and personal guidance for both my academic and social life.

I am deeply grateful to my student colleagues for providing a stimulating and fun environment where we study and grow. I am indebted to my friend, Jia Hu who accompanies with me through the hard times.

I owe my loving thanks to my parents. They have lost a lot due to my study abroad. Without their understanding and endless supports, it would have been impossible for me to finish the work. They let me own a happy family, and always be my stronghold.

Last but not least, I would like to express my sincere thanks to Department of Computing, School of Computing, Informatics and Media, University of Bradford for providing the facilities and supports during my PhD programme.

List of Publications

Lei Liu, Xiaolong Jin, Geyong Min and Keqiu Li, “Performance Modelling and Analysis of Deficit Round Robin Scheduling Scheme with Self-Similar Traffic”, *Journal of Concurrency and Computation: Practice and Experience*, 2010. Accepted.

Lei Liu, Xiaolong Jin, Geyong Min, “Performance Analysis of an Integrated Scheduling Scheme in the Presence of Bursty MMPP Traffic”, *Journal of Systems and Software*, 2010. Accepted.

Lei Liu, Xiaolong Jin and Geyong Min, “Modelling an Integrated Scheduling Scheme under Bursty MMPP Traffic”, in proc. The 23rd International Conference on Advanced Information Networking and Applications (*WAINA’09)*, pp. 212-217, 2009.

Lei Liu, Xiaolong Jin, Geyong Min and Keqiu Li, “An Analytical Model of Deficit Round Robin Scheduling Mechanism under Self-Similar Traffic”, in proc. The International Conference on Scalable Computing and Communications (ScalCom’09), pp. 319-324, 2009.

Lei Liu, Xiaolong Jin and Geyong Min, “Analytical Modelling and Performance Evaluation of Multi-Buffer ARQ systems under Prioritized Self-similar Traffic”, in proc. The IEEE International Communications Conference (ICC’10), 2010.

Lei Liu, Xiaolong Jin, Geyong Min and Jia Hu, “Modelling and Analysis of Dynamic Spectrum Access in Cognitive Radio Networks with Self-Similar Traffic”, in proc. The IEEE Global Communications Conference (GlobeCom’10), 2010. Accepted.

Chapter 1

Introduction

With the rapid development of highly sophisticated devices, high-speed transmission is pervasive and has enriched every aspect of modern communication networks. High transmission speed enables various powerful and fancy network-based applications in daily life. These applications require diversified Quality-of-Service (QoS) and necessitate the differentiated service (DiffServ) [1, 2, 3, 4], a novel computer networking architecture that classifies and manages multiple network traffic as well as provides distinct service. Scheduling scheme is a promising mechanism to make service differentiation possible in communication networks. Thanks to a variety of scheduling schemes, allocating the precious resources fairly or in a weighted manner to the contending consumers according to individual requirements is no longer a significant barrier. As a consequence, tremendous research efforts have been devoted to investigating and measuring the performance of diversified scheduling schemes in communication networks due to their popularity and great importance.

Having understood how the contending traffic flows of communication systems are handled by the scheduling schemes, it is inadequate to accurately investigate, measure and predict the performance of a scheduling system. Traffic pattern plays an important role and has significant impact on the design, performance and management of communication systems. Hence, modelling the characteristics of realistic traffic flows is a critical and challenging task. Many traffic models have been developed. For example, the Poisson model which was firstly proposed by

French mathematician Poisson describes a stochastic process in which the events/instances take place continuously and independently from each other. The Poisson model has been substantially used in emulating the traffic arrival process of a variety of communication networks [5], especially in modelling telephone calls arriving at a switch system [6] and website page requests [7].

Thereafter, many other models have been proposed to discriminate the various traffic characteristics in real-world networks. For example, Markov-Modulated Poisson Process (MMPP) [8] can qualitatively model the time-varying arrival rate and captures the important correlation between the inter-arrival times. Moreover, Constant Bit-Rate and Variable Bit-Rate models [9, 10, 11, 12] have been adopted in modelling voice and video traffic under different working conditions. However, the aforementioned traffic models have limitations which make them inappropriate in modelling traffic of complex high-speed networks because the traffic flows in most existing communication networks, e.g., Ethernet, Internet and 802.11b wireless networks have been proven to exhibit the self-similar nature [9, 10, 13, 14, 15]. The existence of the traffic self-similarity is inevitable in performance analysis of communication networks. Self-similarity exhibits the phenomenon of a part of the object that is exactly or approximately similar to itself. In mathematics, self-similarity is defined as a part of the object that shows the same statistical properties at different scales. In communication, self-similar traffic is able to exhibit dependencies over a wide range of time scales. This is contrasted with the traditional short-term traffic models, such as Poisson traffic. Empirical studies of measurement of traffic traces have shown that self-similarity has serious impact on the performance of communication networks [9, 13].

1.1 Motivations

The provisioning of DiffServ QoS requirements has become an increasingly pressing demand in the last decade. Performance analysis of communication networks is not only a technique issue but also a commercial consideration. The service providers in the modern world are concerned whether the QoS will meet the requirements of their customers within limited resources. To this end, the issue of how to allocate the resource effectively draws much attention. Hence, performance evaluation of the designed systems through analytical modelling has risen from such needs.

Analytical models of various scheduling schemes enables to predict the behaviours and performance of targeted systems. It is the most efficient way to cut the expense, save time and manage the precious resource effectively. However, it is a very challenging task to analyze the behaviours of individual buffers which are scheduled under different schemes. This is because scheduling schemes allow the individual buffers to share the unique server such as Generalized Processor Sharing (GPS) [16], Deficit Round Robin (DRR) and Fair Queueing (FQ) scheduling schemes [17]. The ongoing interaction among traffic flows scheduled by the above scheduling schemes cannot be directly isolated. Thus, it brings barriers in modelling these scheduling systems. Scheduling schemes, such as Priority Queueing (PQ), GPS, and DRR, are able to classify the buffers into different priority levels. Applications with stringent QoS requirements could be separated and served with the high priority by the aid of such a scheduling mechanism. The inappropriate models for traffic arrivals may lead to unexpected and even incorrect results. Hence, characterizing the nature of the realistic traffic is the foundation of accurately

modelling and evaluating the performance of selected systems. Since the packetized traffic pervasively exhibits the self-similar nature, conventional traffic models do not apply to self-similar traffic. The design and development of analytical models for various scheduling schemes in the presence of self-similar traffic is an open, fertile and attractive area of research.

1.2 Aims and Objectives

The proliferation of multimedia applications has brought new challenges on QoS requirements, such as fair service and differentiated QoS. On the other hand, the burgeoning popularity and importance of modelling the characteristics of self-similar traffic in communication networks has drawn much attention. In order to meet these demands, this thesis aims to develop new and cost-effective tools for investigating and evaluating the performance of innovative and integrated scheduling systems subject to self-similar traffic. The main objectives of this study are:

- To measure and evaluate the self-similar nature by applying Hurst parameter estimators on realistic and simulation traffic traces.
- To develop efficient and cost-effective analytical tools for DRR scheduled systems subject to self-similar traffic.
- To develop an analytical model for the hybrid PQGPS system which combines the advantages of PQ and GPS scheduling mechanisms for service differentiation.
- To develop a performance model of multi-buffer Automatic Repeat reQuest (ARQ) system under the prioritized self-similar traffic and use

this model to configure the delay bound in order to decrease the queueing loss and transmission loss probabilities.

- To derive an analytical model of a typical Cognitive Radio (CR) [12, 18] network which focuses on analyzing the queueing performance of secondary users and the effective service shared by them.

To achieve the goals, this study develops novel approaches to isolate the contending traffic flows from the ongoing interactions. The accuracy of the developed models are corroborated through extensive comparisons between the analytical and simulation results.

1.3 Thesis Organization

This thesis is inspired by the foreseen advantages and pressing demand of investigating the performance of various scheduling systems by taking self-similar traffic into account. In this subsection, the architecture of this thesis is outlined. Self-similar traffic is employed as the input of the analytical models, and hence is introduced first. Next, the models of DRR scheduling as a single case, and PQGPS scheduling as a hybrid case are presented. Afterwards, an analytical model of multi-buffer ARQ system is addressed. Finally, the performance analysis of the CR network which employs decentralized spectrum access of coordinating the transmission is investigated.

- Chapter 2 reviews the existing studies in three categories. Firstly, traffic modelling of communication networks and systems is presented. Traffic pattern has significant impact on the accuracy and appropriateness of

analytical models. Next, this chapter focuses on the introduction of scheduling algorithms which handles the traffic in communication networks. Finally, a very promising concept of Cognitive Radio is presented. A lot of research efforts have been drawn to this subject. The corresponding studies in open literature have also been reviewed.

- Chapter 3 begins with the introduction of autocorrelation, one of the measurements used in testifying the self-similar phenomenon. Autocorrelation is a mathematical tool for revealing the repeat patterns in a long time scale. The autocorrelation of a stochastic process presents the correlation of the process at different time epochs. Autocorrelation is required to express the self-similar nature of network traffic. The degree of the self-similarity which is persisted in the realistic traffic can be characterized by Hurst parameter. In this chapter, three representative estimators of Hurst parameter, which are variance time, rescale range and periodogram are introduced in detail. Then, these estimators are applied to measure the normal traffic traces and the traffic traces that contain the labelled attacks in order to evaluate the traffic self-similarity in terms of the Hurst parameter.

- Chapter 4 introduces an analytical model of DRR scheduling systems subject to self-similar traffic. DRR is a promising fair scheduling mechanism owing to its low complexity and excellent ability of achieving a good degree of fairness in terms of throughput. A decoupling approach is developed to eliminate the queueing effect between the contending traffic

flows, and hence each traffic flow is isolated from the original system. The merits of the proposed model are demonstrated through an application.

- The hybrid PQGPS scheduling system is able to provide prioritized service meanwhile support fair service to traffic flows. Chapter 5 introduces the hybrid PQGPS scheduling scheme and presents a novel analytical model for the integrated PQGPS scheduling system subject to self-similar traffic.

- In Chapter 6, a performance model of the prioritized multi-buffer ARQ system which takes differentiated QoS and error control strategy into account is developed. With the increasing number of requests to access the service medium and receive differentiated services, the reliability of transmission has become a critical factor which has significant impact on the performance of communication systems. In this part, the schematic structure of the multi-buffer ARQ system is presented. Having understood the basis of the nested PQ scheduling mechanism, the approaches to obtaining the desired loss probabilities are introduced. The model is corroborated through the comparison between the analytical and simulation results, and the advantages of the developed analytical model are demonstrated through an application.

- Chapter 7 demonstrates the analytical modelling of the functionality of CR techniques and a typical CR network. CR as a promising technique has been proposed to enable the utilization of the licensed frequency bands when they are vacant. Instead of the most widely employed ON-OFF source to model the service of secondary users, this chapter adopts a

decomposition method which derives the residual service left by the primary user by taking self-similar traffic into account. With the residual service, secondary user can be isolated from the CR network. Further, the queueing performance of each secondary user is derived by obtaining the corresponding effective service.

- Chapter 8 summarizes the thesis and looks into the future work.

1.4 Contributions

This thesis focuses on developing the analytical models of representative scheduling systems subject to the self-similar traffic including DRR scheduling which is a single scheduling; PQGPS scheduling which combines two scheduling schemes; Prioritized ARQ which employs nested PQ scheduling; a typical CR network which adopts the spectrum access resource allocating scheme. The contributions of this thesis are listed as follows:

- Three estimators of Hurst parameter are reviewed. Estimators are applied on the realistic traffic traces provided by MIT Lincoln Lab [19] to reveal the self-similar nature existing in communication networks. Particularly, DDoS attacked traffic trace is measured by the Hurst parameter estimators. The results show that even the attacked traffic still exhibits the self-similar nature. However, there are significant changes on the degree of self-similarity and the autocorrelation between the traffic data before, during and after the attack.

- A novel analytical model for deriving the upper and lower bounds of the queue length distributions of individual traffic flows in DRR scheduled systems subject to self-similar traffic is developed. To eliminate the ongoing interactions between the contending traffic flows, a priority conversion approach is proposed to derive the upper bound of queue length distribution for each traffic flow. The idea is inspired by converting the DRR scheduled system into a PQ system, by doing so, the low priority part of the modulated PQ system receives less service than that of the original DRR system. Correspondingly, an approach which modulates the arrival rate of one of the traffic flows handled by DRR is proposed in order to convert the original complex system into SSSQ systems

- It is very challenging to directly analyze the queueing performance of each traffic flow in a hybrid PQGPS scheduling system. To overcome the constraint, a two-step decomposition approach is developed. By doing this, the complex hybrid system can be divided into a collection of SSSQ systems.

- An analytical model by taking the transmission reliability into account through the ARQ strategy. Particularly, there has been no model reported for ARQ under self-similar traffic. The performance of individual queues in the multi-buffer ARQ system is derived by isolating the corresponding queue from the original complex system under the condition that the isolated queueing system is statistically equivalent to the corresponding original queue. Then, the loss probabilities of individual queues of the

multi-buffer ARQ system can be obtained by examining the isolated queueing system.

- Performance modelling of a CR network is developed. Firstly, the residual service left by the primary user is obtained. The derived service is actually the effective service shared by all the secondary users. It is more practical than using ON-OFF sources of assuming the service of secondary users. Afterwards, the secondary users are further isolated from their interactions by obtaining the corresponding effective service. Finally, based on the individual effective service, the queue length distribution of individual secondary users can be calculated.

- The developed analytical models are applied to demonstrate their merits in order to enhance the performance, investigate the configuration of the weights or build a cost-effective communication system.

Chapter 2

Literature Review and Preliminaries

In today's society, daily life has been significantly altered by the booming network based applications, such as in medicine, education, manufacturing and entertainment industry. Particularly, with the growing utilization of multimedia devices in the real world, diversified communication systems have become more necessary than ever before. However, the growing proliferation of network devices has caused the congestion of networks and communication systems, which significantly degrades their performance. In addition, the increasing number of applications which request network resources is rising day by day. Hence, the service providers seek feasible ways of investigating, examining and predicting the performance of their own communication systems simultaneously. By doing this, it enables the providers to offer the service which meets the QoS demands of applications while controlling the cost.

To achieve the goal of investigating the performance of communication systems, simulation and analytical modelling are the commonly used approaches. A simulation system is a computerized or programmed virtual system which emulates the behaviours of a communication system. Simulation usually runs over time in order to study the impact of the defined interactions and activities. In other words, simulation is iterative in the predefined regulations and constraints, by doing which one can learn, revise and understand the analogical actions and behaviours of the corresponding practical systems. On the other hand, an analytical model is a mathematical characterization of performance metrics of the practical system on

specified aspects. It is employed to promote understanding of the practical system normally based on mathematical derivations.

Compared to the simulation approach, analytical modelling is time and cost efficient because it takes a significant amount of time to obtain the results from simulations. Scarcity of simulation time may lead the system to an unstable status and hence unexpected endings of the performance results. On the contrary, one can achieve the goal of measuring the performance of designed systems through analytical models momentarily. As a consequence, analytical modelling is preferred as a cost-effective tool of investigating the performance of communication systems. This is not to say that analytical modelling is omnipotent but it is an efficient .

However, there is a critical issue which should be taken into account. In order to properly model a practical system, traffic patterns should be considered. In the real world, the appropriateness and accuracy of modelling communication systems lie on the processes which are chosen to describe system inputs. Thus, the more close to the activities of the real-world traffic, the more efficient the model is.

2.1 Traffic Modelling

Traffic modelling is an important means of illustrating the nature and characteristics of traffic flows in practical communication networks [20, 21, 22]. It provides the researchers a mathematical tool for emulating the behaviours of network traffic, and hence makes the modelling of system inputs possible. With the development of traffic models, there have been a number of representative cases which can characterize the inputs of the corresponding systems in a good degree, such as Poisson, MMPP and self-similar models.

2.1.1 Poisson Traffic Model

The Poisson traffic model is widely employed as inputs in various communication networks [23]. It is a stochastic counting process in which events take place continuously and independently of one another. In statistics, the time intervals between events in a Poisson process follow the exponential distribution [24]. This is the key to model and generate the Poisson traffic. The cumulative distribution function of exponential distribution is given by [24]

$$P(X < x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{2.1}$$

where λ is the mean arrival rate of a Poisson process, and $1/\lambda$ is the mean time period between two arrivals. x is a stochastic variable.

The prominent use of Poisson model is to emulate the call arrivals to a phone call centre i.e. a switchboard [23]. By employing Poisson model, it is possible to predict the probability that a call may be blocked. In other words, the probability of failure connection attempts can also be calculated. For instance, a cell phone service provider is able to configure the device to control the service level that customer calls get a busy signal within a reasonable rate.

2.1.2 MMPP Traffic Model

Compared to Poisson processes, another preferred traffic model, the so-called ON-OFF source [25] which characterizes the traffic bursty properties, is widely involved in the development of analytical models [26]. According to such traffic models, packets arrive only during the ON state/period. Just the opposite, the traffic

source is idle when it is in OFF state/period, which means no data is generated. In open literature, there are two important instances of ON-OFF source:

- Exponentially distributed inter-arrival time: the interval of time between the adjacent arrivals during ON periods follows exponential distribution. In other words, in each ON periods, the input traffic is Poisson process.
- Deterministic inter-arrival time: the intervals between the arrivals when the source is in ON state are constant/fixed.

As aforementioned, the traffic in modern communication networks is commonly composed of a variety of patterns. Pervasive studies have shown interests in Markov Modulated Poisson Process (MMPP) [8] as inputs of bursty traffic. MMPP is raised from the ON-OFF source. Compared to the two states of ON-OFF source, the amount of states of MMPP could be any finite number. In each state, the arrival is a Poisson process with a different arrival rate. MMPP switches the state from one to another based on the related transition rate and forms a Markov chain which has the property that the next state depends only on the current state. The popularity of MMPP lies on its capability of capturing traffic burstiness and qualitatively modelling time-varying arrival rate and important correlation between inter-arrival times [8]. Moreover, MMPP is controllable and trackable. Hence MMPP is widely adopted for characterising the point processes which have randomly varied arrival rates over time periods especially in communication modelling [27, 28]. The pioneer works subject to MMPP were first induced by Naor and Yechiali [29], and later by Neuts [30]. Afterwards, a huge number of studies and publications have focused on the topic. A more sophisticated study on MMPP has

been given by Fischer and Meier-Hellstern [8]. They have presented approaches to calculate the queue length distribution of a queueing system subject to the MMPP arrival with exponential service time and determined service time. Moreover, the waiting time of such queueing system has been discussed.

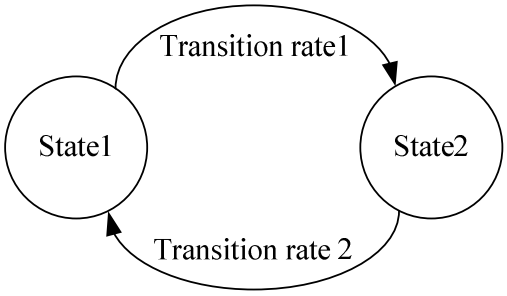


Figure 2.1: Two-state MMPP.

In this section, a two-state MMPP is formulized for simplicity. Figure 2.1 shows a typical two-state MMPP and how the two states switch from one to the other. It is clear that the state switches based on the corresponding transition rate. The terms $\lambda_i, \delta_i \quad i = 1,2$ are the mean arrival rates and transition rates of the corresponding state. The MMPP is readily parameterized by a rate matrix Λ and infinitesimal generator Q [8]

$$\Lambda_j = \begin{bmatrix} \lambda_1^j & 0 \\ 0 & \lambda_2^j \end{bmatrix}, \quad Q_j = \begin{bmatrix} -\delta_1^j & \delta_1^j \\ \delta_2^j & -\delta_2^j \end{bmatrix}. \tag{2.2}$$

The inverse of the transition rate denotes the length of the time period that the related state persists before next switching. Then the mean arrival rate of the aforementioned MMPP can be given by

$$\overline{\lambda^j} = \frac{\lambda_1^j \delta_2^j + \lambda_2^j \delta_1^j}{\delta_1^j + \delta_2^j}. \tag{2.3}$$

2.1.3 Self-Similar Traffic Model

Many researchers have shown that the self-similar nature is a ubiquitous phenomenon which is consistent in most kinds of communication network traffic [9, 14, 31, 32, 33, 34]. However, MMPP is incapable of characterizing the traffic self-similarity. Therefore, it is inevitable to introduce a novel traffic model which illustrates self-similarity. With the increasing knowledge on practical traffic, researchers have revealed the importance of self-similar studies, after it was first exposed by Leland, Taqqu and Willinger in Ethernet traffic [13]. They demonstrate the finds that Ethernet Local Area Network (LAN) traffic is statistically self-similar, and none of the commonly used traffic models at that time was appropriate to feature this behaviour. Their experimental work on datasets collected from practical network shows the evidence of traffic self-similarity. Later on, the self-similar nature of traffic was widely observed and studied in other communication networks, such as ad hoc networks [35], IP networks [36] and popular 802.11 wireless LAN networks [14]. All the mentioned studies have applied the similar traffic measurement approach on the datasets collected from the corresponding network circumstance. Based on the observation of Hurst parameter of these datasets, the traffic in these networks has shown self-similar nature. According to the facts of practical experiments in realistic systems, it is undeniable that the self-similar nature of traffic has significant impact on the performance of communication networks owing to its scale-invariant burstiness and large-lag correlation nature. As a promising traffic model, fractional Brownian motion (fBm) [9, 37, 38] process is preferred in this

thesis due to the reasons that it is self-similar; it has stationary increments; and the process exhibits long-range dependence.

The cumulative fBm process in terms of time t is formulized [9] as

$$A(t) = mt + \sqrt{am}\bar{Z}(t), \tag{2.4}$$

where m is the mean arrival rate of the process, a is defined as the variance coefficient of fBm process, and $\bar{Z}(t)$ is a centred fBm process (i.e. $\mathrm{E}\bar{Z}(t) = 0$). The variance of $\bar{Z}(t)$ is given by

$$\bar{v}(t) = \mathrm{Var}\,\bar{Z}(t) = t^{2H}, \tag{2.5}$$

where H is the Hurst parameter which is ranged $H \in [0.5,1)$. Hurst parameter is an important element of characterizing self-similar processes which indicates the degree of self-similarity. Moreover, it reflects and implies the burstiness of the traffic [39]. The covariance of $\bar{Z}(t)$ can then be described by

$$\overline{Cov}(t_1,t_2) = \frac{1}{2}\left(\bar{v}(t_1) + \bar{v}(t_2) - \bar{v}(t_1 - t_2)\right) = \frac{1}{2}\left(t_1^{2H} + t_2^{2H} - (t_1 - t_2)^{2H}\right). \tag{2.6}$$

Based on Equations (2.5-2.6), the variance function of the fBm process $A(t)$ is given as

$$v(t) = am\bar{v}(t) = amt^{2H}, \tag{2.7}$$

alone with its covariance function as

$$Cov(t_1,t_2) = \frac{1}{2}am\left(t_1^{2H} + t_2^{2H} - (t_1 - t_2)^{2H}\right). \tag{2.8}$$

2.2 Scheduling Algorithms

Growing network applications have exhausted the precious resource of communication systems. Network users seek the possibility of plundering service from other individuals in order to fulfil their own QoS demand. Therefore, research attentions have been drawn to the design and implementation of novel and efficient scheduling schemes [1, 16, 17, 36, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49] which aid the transmission and resource allocation in high-speed networks. In the history of the studies of scheduling algorithm, it was initially proposed to enable the processes or data flows of accessing the system resources or media, e.g. communication bandwidth. The scheduling algorithms make the system possible for achieving a target QoS with reference to various purposes and regulate the service of the system. Particularly, contemporary communication systems are often required to execute more than one process at a time, and may transmit multiple flows simultaneously. Hence, the powerful scheduling algorithm arises and is employed as an alternative to the traditional First Come First Served (FCFS)/ First In First Out (FIFO) mechanism.

An effective and efficient scheduling scheme should cover the concerns of server utilization, throughput, queueing delay for packet in queues and the proportional fairness. More importantly, with the development of communication techniques and diverse requirements of applications, the increasing demand of providing priority service, guaranteed service and fair service, which may refer to service differentiation in the unique scheduling system, has appeared.

2.2.1 Round Robin

Owing to the increasingly pressing demands of a range of QoS requirements, the studies on design of novel scheduling schemes have emerged endlessly. To achieve the goal of fairness, the pioneer work backdated to 1987 which switches the service to traffic flows by employing a Round Robin (RR) manner has been proposed by Nagle [50]. The service discipline of RR assigns time slices or service quantum to each traffic flow in circular order. RR mechanism will pass the empty queue without compensating the service in future rounds. This scheduling mechanism prevents the traffic flows from increasing its service opportunity, which results in the delay of other flows. By doing this, the aggressive traffic flow merely increases its own backlog. However, there are few reported practical routing/switching systems which adopt the scheduling mechanism proposed by Nagle. Because the scheduling is not actually fair when it handles the traffic flows with variable length packets [42]. It is obvious that the traffic flow with the larger packet size receives more service in one round. Hence the fair service allocation of RR scheduling is broken under this situation.

2.2.2 Fair Queueing

Thereafter, Demers, Keshav and Shenkar [17] extended the work based on the modification of Nagle's algorithm, called fair queueing. This fair queueing algorithm emulates the bit-by-bit Round Robin (BR) in packet level. The switching systems which employ such a scheduling algorithm are possible to allocate the resource to each traffic flow fairly regardless of the packet length [46]. But, the scheduling system has to maintain a virtual BR scheduling algorithm in order to determine which flow should be served next. The fair queueing algorithm sustains

the complexity of $O(\log(n))$ [17, 42], where n is the number of active traffic flows. The flaw of fair queueing algorithm is that fair queueing becomes inefficient in terms of complexity with the increase of the number of active traffic flows. As a consequence, it is not feasible to implement such a mechanism in high speed networks where a large number of traffic flows are usually involved.

Afterwards, the studies focus on reducing the complexity of fair queueing algorithm but maintain the characteristic of fairness. A significant work, called self-clocked fair queueing (SCFQ) algorithm [49] which reduce the computational complexity of calculating virtual finishing time to $O(1)$ by employing a new virtual time function. Compared to fair queueing algorithm, the virtual time which stamps the finishing time of a packet is referenced to its own queueing system instead of creating the need for maintaining and computing the corresponding virtual time from a hypothetical BR system. Although SCFQ reduces the computational complexity of computing the virtual finishing, those of dequeue and enqueue are still $O(\log(n))$. As a consequence, SCFQ retains the $O(\log(n))$ sorting bottleneck.

Although the traditional round robin has its own weaknesses and disadvantages, there are a number of scheduling algorithms which are developed to achieve the goal of fairness by extending the functions of round robin. Scheduling mechanisms such as Fair Round Robin [44], Elastic Round Robin [43], Frame-Based Proportional Round Robin [45] and Ordered Round Robin [48]. The extended and modified RR scheduling algorithm overcomes the aforementioned weaknesses to a certain extent while maintains the advantage of low complexity.

2.2.3 Generalized Processor Sharing

Besides the variants of RR scheduling algorithm, Generalized Processor Sharing scheduling (GPS) [16, 51, 52] has attracted much attention due to its perfect fairness of allocating the resource to multiple traffic follows. Particularly, GPS enables the QoS differentiation, and is capable of providing guaranteed services to individual traffic flows, meanwhile allowing them to share the excess service capacity from each other. More important, the service assigned to each traffic flow in GPS scheduled systems is controllable by adjusting the weight of the corresponding traffic flow. GPS is an efficient, flexible and fair service discipline, however, it is not possible to implement in the present switching systems since it assumes fluid traffic which is infinitesimal units instead of the packet. Although GPS is a theoretically ideal scheduling scheme, it is useful as a benchmark against which realizable scheduling mechanisms can be measured [16, 51, 53]. Stiliadis and Varma [53] have compared the inherent latency of representative scheduling algorithms and verified that GPS has zero latency which makes it an idealized service discipline. In addition, there is a number of scheduling schemes track the performance and service discipline of GPS closely, such as Weighted Fair Queueing (WFQ) [24, 53, 54] also known as Packet-by-packet Generalized Processor Sharing (PGPS) which is a generalization of that GPS allows the guaranteed bandwidth service at packet level.

Before these scheduling schemes can be implemented in practical systems, a recurring request of service providers is willing to measure and evaluate the performance of scheduling schemes. In this thesis, the analytical models subject to self-similar traffic under various scheduling schemes are developed. So far, a number of widely used traffic models have been introduced. In the following part,

the scheduling schemes for which the models are developed in this thesis will be demonstrated.

2.2.4 Deficit Round Robin

Initially, a promising modified round robin scheduling scheme is presented, which possesses excellent fairness in terms of throughput, the so called Deficit Round Robin [42, 55, 56]. DRR is able to provide fair service to non-empty traffic flows with the low complexity of $O(1)$ [42, 55] by comparing to the aforementioned FQ scheduling system. DRR is equipped with deficit counters for each traffic flow. The deficit counter contains the remainders of service quantum which restricts and controls the service volume to be received for each traffic flow in one round. The unused service volume which is recorded by deficit counter of a traffic flow in the current round will be granted to the corresponding traffic flow for next round [42]. In this way, the service which is deserved by any traffic flow cannot be plundered by other traffic flows. It can be readily seen that the service received by each non-empty queue only depends on the corresponding quantum. As a result, it can be readily extended to a weighted service scheduling scheme by appropriately configuring the service quantum assigned to individual traffic flows. In other words, the service provided to each traffic flow is trackable and controllable. Additionally, DRR overcomes the barrier of the fairness of scheduling crashes when the packet size varies [42, 56]. By contrast with FQ scheduling, DDR is still fair without knowing the mean packet size of each traffic flows.

Due to the advantages of DRR scheduling, it has been widely implemented in commercial high-speed routers [57]. Figure 2.2 presents the schematic diagram of a typical DRR scheduling system. As aforementioned, the server switches to each

traffic flow and assigns new service quantum in a conventional round robin manner in every round. This means that the volume of the deficit counter of a traffic flow at the beginning of each round is equal to the summation of its quantum and the volume of its remaining service from the pervious round. The number of packets of an active flow that can be served in each round is determined by the value of its deficit counter. The term active flow is the one with at least one packet waiting for service in the corresponding buffer.

Distinguished from the original round robin scheduling, the server in DRR scheduling system begins to serve the next active flow in each round under either of two conditions:

- The current queue is empty.
- There is not enough service volume left to serve one more packet of the present flow.

This service process is illustrated in Figure 2.2. By doing this, the greedy traffic with the large packet size only punishes itself for the reason that the remainder of service of each traffic flow is reserved by them only.

Existing studies on developing analytical models of communication systems which are handled by diverse scheduling schemes have shown various aspects of considerations. Many research efforts have been made to analyze the performance of DRR systems on various communication networks. For example, Stiliadis and Verma [53] corroborated an upper bound or the packet latency in DRR. They developed a tool for analyzing the maximum delay when a packet waits until its service. Their study has shown that the delay which is an upper bound relies on the assumption on the potential maximum packet size of individual flows. However, if

the packet size is variable, it is not easy to see the maximum packet size of each flow. Later on, Kanhere and Sethu [55] developed a tighter upper bound for the scheduling delay of each flow in DRR. The recent study of Lenzini, Migozzi and Stea [56] further provided an upper bound on the respect of the starting up latency of DRR, namely, the maximum delay time of a head-of-line packet. These existing studies related to DRR have focused on the delay/latency of packets caused by inherent scheduling decisions. The accuracy of the obtained upper bounds of latency depends on the maximum packet size of individual traffic flows, which may be significantly diverse in a variety of communication systems [55].

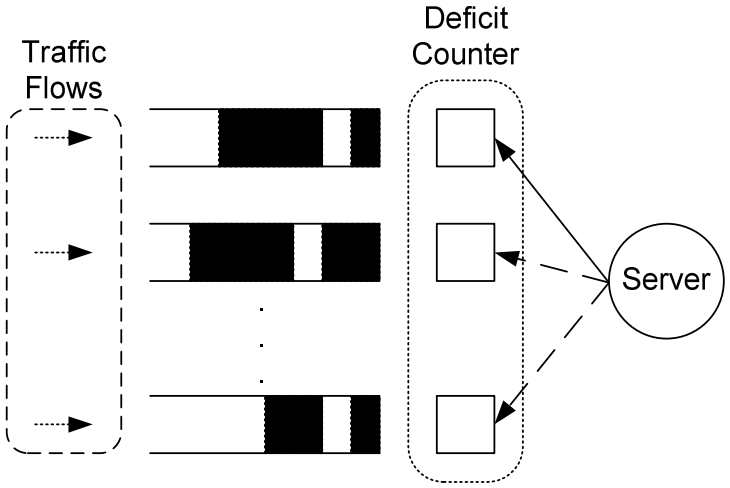


Figure 2.2: DRR scheduling.

2.2.5 Hybrid PQGPS

Recently, the provisioning of DiffServ has emerged as a stressing demand [1, 4, 58]. Various network applications induce the necessity of QoS differentiation. For instance, voice application users cannot tolerate too much data loss in their communications, while the lag in video communications can significantly degrade user-perceived QoS. To meet the need, a hybrid scheduling scheme, namely, Priority

Queueing-Generalized Processor Sharing (PQGPS) scheduling which integrates the fundamental PQ and GPS scheduling schemes, has emerged. Individually, PQ scheduling is capable of guaranteeing high transmission speed for time-sensitive traffic flows which provides best-effort service to non-critical ones [24, 59]; on the other hand, the advantages of GPS scheduling relies on its appealing features of providing fair and guaranteed services to individual traffic flows, meanwhile allowing them to share the excess service from each other [16, 51, 60, 61]. Particularly, the service assigned to each traffic class in GPS scheduling systems is controllable by adjusting the weight assigned to it. Analytical models of communication systems handled by PQ and GPS scheduling schemes are reported in the open literature [59, 62, 63, 64]. Especially in recent years, performance analysis of PQ or GPS systems are subject to both shot rang dependent traffic (e.g., Poisson traffic) and self-similar traffic has drawn many efforts. Ashour and Le-Ngoc [59] has analyzed the issues on the priority queueing of long-range dependent traffic. Similar concerns of priority queueing analysis for self-similar traffic in high-speed networks have been studied by Quan and Chung [64]. Besides, a number of analytical models for GPS have been proposed. Borst, Mandjes and van Uitert analysed GPS queues with heterogeneous traffic, light-tailed and heavy-tailed input [62, 63]. Particularly, Jin and Min [65] have developed a flow-decomposition approach of predicting the performance of GPS under homogeneous self-similar traffic. Yet, there are still few reported works on the analytical models of hybrid PQGPS systems, especially taking the traffic self-similarity into account.

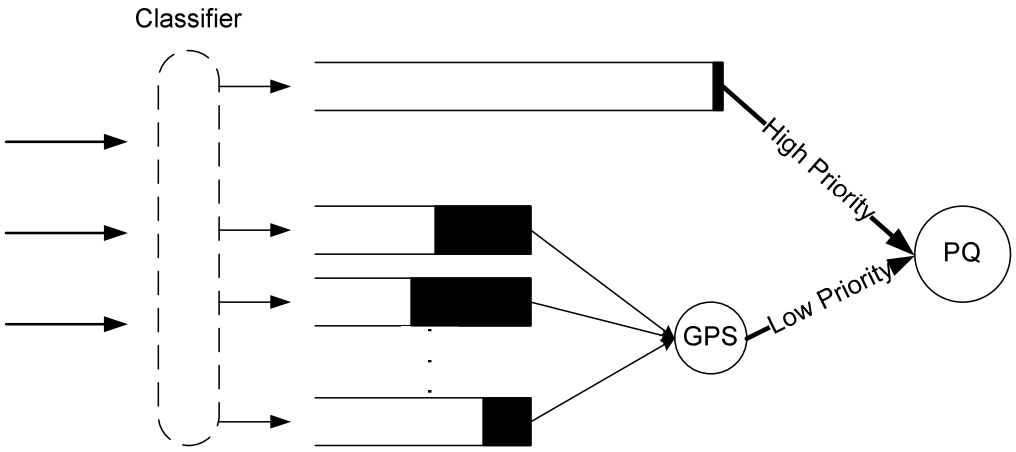


Figure 2.3: Hybrid PQGPS scheduling.

Figure 2.3 shows the schematic of how PQGPS is able to classify the traffic and provide differentiated service respectively.

As an integrated scheduling scheme, PQGPS [66] succeeds as it incorporates the advantages of both PQ and GPS. The targeted hybrid system serves the applications that have the most stringent QoS requirements and allocate the residual resource fairly to others according the pre-assigned GPS weights.

Until now, a single scheduling, DRR and a hybrid scheduling, PQGPS have been introduced. The next subsection presents an application of nested PQ scheduling for supporting the implementation of ARQ error control strategy in a prioritized multi-buffer system.

2.2.6 Automatic Repeat reQuest

ARQ forwarding error control strategy [67, 68] is pervasively deployed in practical communication systems. ARQ aims to provide the algorithm of retransmitting damaged or lost packets in wireless communication networks.

However, with the development of high-speed network techniques, the original stop-and wait ARQ strategy cannot meet the demand of high efficient retransmission as it has to wait for positive or negative Acknowledgements (ACKs). Figure 2.4 shows how the stop-and-wait ARQ [69] works during the transmission.

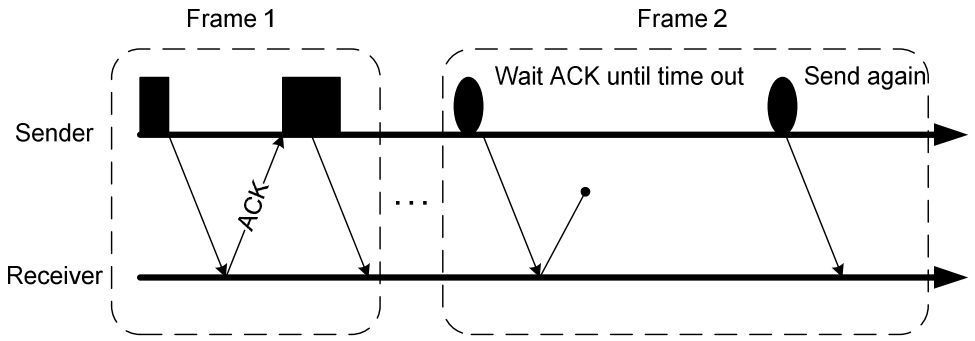


Figure 2.4: Stop-and-Wait ARQ.

From Figure 2.4, it is clear that the sender has to wait for the Acknowledgement (ACK) to process the next transmission. If the ACK is positive, the sender continues the transmission as shown in Frame1 of Figure 2.4. Otherwise, the sender retransmits the previous data. It is even worse if the retransmission is triggered according to ACK timer expiry, which means that the sender retransmits the data after waiting for a fixed time interval without acknowledged, as shown in Frame 2 of Figure 2.4.

As the inefficiency of the stop-and-wait algorithm, a modified ARQ scheme, namely, go-back-N [70] ARQ has been proposed. This specific instance of ARQ strategy employs sliding window protocol with window size N to continuously process a number of packets. The receiver keeps tracking the packet it expects to receive, and send all ACKs back to the sender. Figure 2.5 depicts the transmission

and retransmission process according to a go-back-N strategy with window size $N = 3$.

It can be readily seen that Packet 5 is not successfully received by the receiver. The retransmission is initialized upon the receiving of negative ACK. Go-back-N retransmits all the packets after Packet 5 within the previous (i.e., Packet 6 in Figure 2.5) window and fill the vacancies (i.e., Packet 7). Then, the retransmission frame is Packets 5-7 in this case.

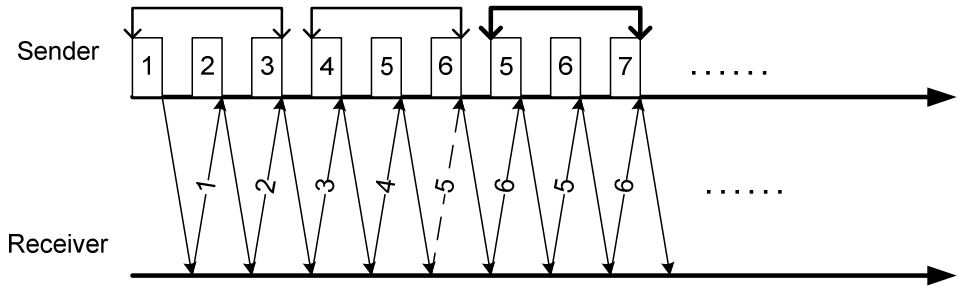


Figure 2.5: Schematic diagram of go-back-N strategy with window size *N=3*.

Compared to stop-and-wait ARQ, go-back-N is a more efficient strategy because the packets are sent on the link instead of stopping and waiting for an ACK to process the followings. However, it can be observed from Figure 2.5 that go-back-N is inevitable to transmit the packets multiple times. Especially, if one packet is lost or damaged in a frame, all following packets have to be retransmitted even when they are received without error. To improve this flaw, Selective Repeat ARQ (SR-ARQ) [71, 72, 73, 74] has been developed.

SR-ARQ is considered as the most efficient retransmission strategy, since it only retransmit the damaged or lost packets [74]. From Figure 2.6, it is noticeable

that SR-ARQ maintains the windows on the base of both the sender and receiver bases which have identical window size. The receiver process keeps a track of the sequence number of the earliest packet it has not received, while it sends that number along with every corresponding ACK. In addition, the receiver fills its window with the subsequent packet and replies to the sender with ACKs and sequence number of the most primitive missing packet. The sender will retransmit the packet which is ACKed by the receiver once it has sent all the packets in the sender window. By doing this, it is not hard to see that the sender continuously sends packets without waiting and only retransmits the damaged or lost packet which is ACKed by the receiver.

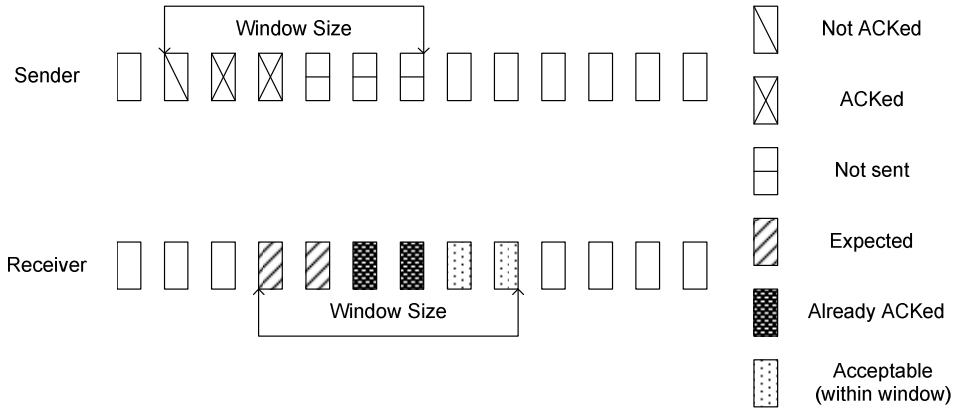


Figure 2.6: SR-ARQ.

As a practical application of scheduling systems, transmissions and retransmissions of packets according to any error control strategy need to be coordinated by schedulers. Communication systems which employ ARQ strategy are the most popular in real-world applications. Existing studies on such systems were focused on single arrival buffer ARQ systems [68, 71, 75]. Other studies [76, 77, 78, 79] on this topic evaluate the performance of ARQ systems in which packets are retransmitted under the delay constraint subject to conventional Poisson traffic for

ease of calculations and modelling. Larsson and Johansson [67] studied an ARQ system with multiple inputs, where all retransmitted packets are fed back to the unique ARQ buffer. The flaw of such systems is that retransmitted packets are treated equally regardless of their various delay constraints and differentiated QoS requirements.

A prioritized communication system which employs the SR-ARQ system is of our interest. Figure 2.7 depicts the aforementioned system which classifies the traffic into different priority sessions. In each session, an ARQ buffer is equipped to store the retransmitted packets for the corresponding arrivals.

It is not hard to see that the inherent mechanism which handles the transmission is a nested PQ scheduling scheme. At the top level, the two sessions which are composed of the corresponding arrival buffer and ARQ buffer are handled by PQ. By doing this, the service assigned to traffic flows is prioritized. Further, the retransmitted packets have a higher priority to be served over those in the corresponding buffer, which implies a prioritized relationship between the arrival buffer and its ARQ buffer in each session.

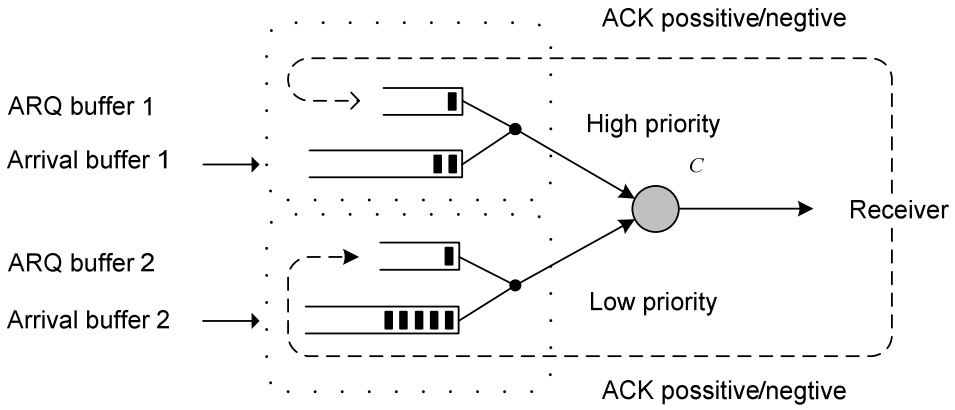


Figure 2.7: Prioritized multi-buffer ARQ system.

2.3 Spectrum Access in Cognitive Radio Networks

Scheduling schemes are used in wireless networks for handling the channel sharing and packets switching. In recent years, with the growing number of wireless applications, the usable spectrum is approaching its full capacity. However, much of licensed spectrum is idle almost at any time everywhere according to the authorities [80]. Figure 2.8 presents the measurements of spectrum occupancy based on six reports [81, 82, 83, 84, 85, 86], which is provided by Shared Spectrum Company (SSC). It is obvious that the spectrum occupancy is low in each frequency band.

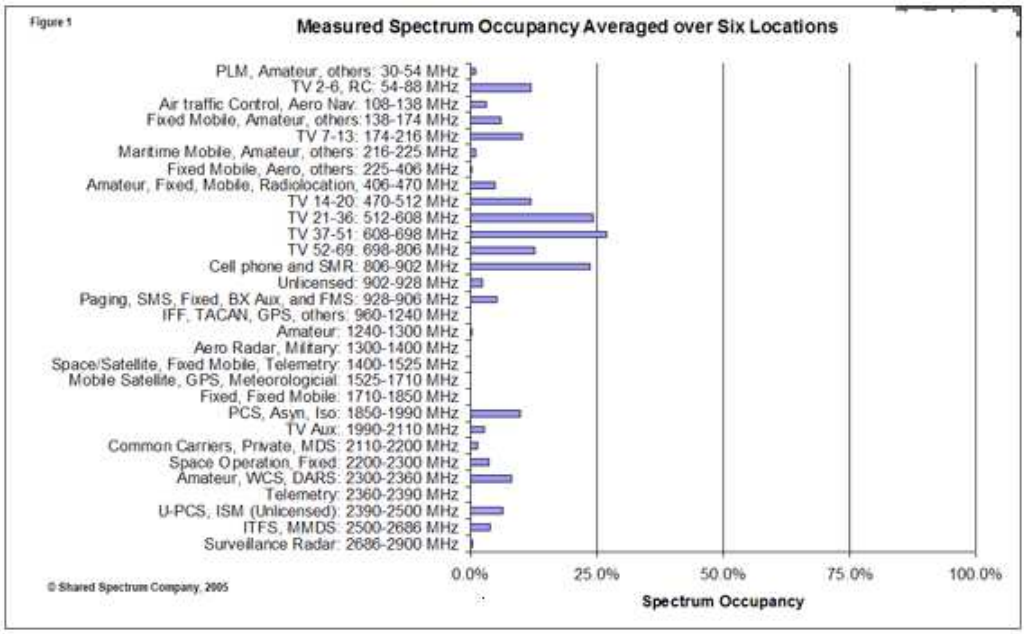


Figure 2.8: Overview of spectrum occupancy provided by SSC [81, 82, 83, 84, 85, 86].

The maximum occupancy rate is around 25% (i.e., TV 37-51:608-698 MHZ). Especially, there are particular bands which are almost idle all the time (e.g., 960-1240 MHZ, 1400-1525 MHZ and 2360-2390 MHZ). In addition, Figure 2.9 which is provided by the same company shows the spectrum occupancy at different locations.

From this figure, one can observe that the spectrum occupancy is low even in New York which is almost the busiest and most crowded city on this planet. The fact tells a different story. This finding implies the inefficient spectrum management policy in the present wireless communication networks.

CR technology has emerged as a promising means to utilize the licensed spectrum bands when they are vacant. It enables to share the excess service of licensed users with unlicensed users. In CR networks, primary users (i.e., licensed users) should not be interfered by secondary users (i.e., unlicensed users or CR users) during their transmission. On the contrary, secondary users must release their spectrum access when primary users are sensed or detected. CR technique induces dynamic spectrum access which helps secondary users discover the spectrum holes (also known as white space) and evades primary users.

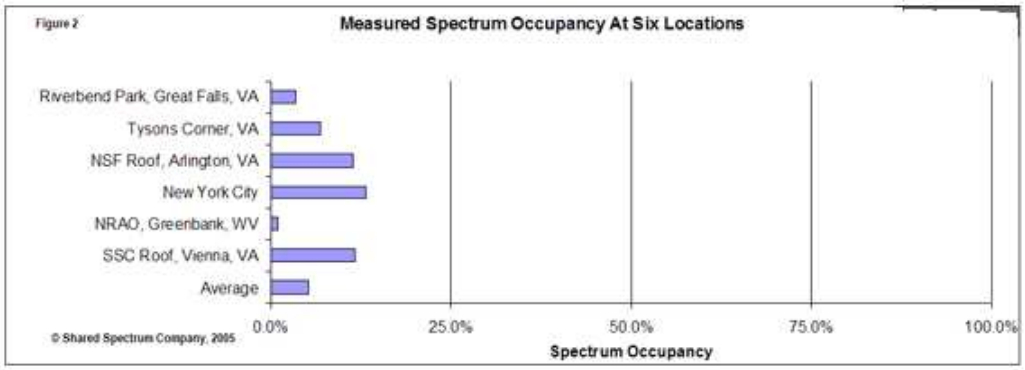


Figure 2.9: Average spectrum occupancy at different locations provided by SSC [81, 82, 83, 84, 85, 86]

Recently, significant research efforts have been made on studying CR networks on all aspects. Excellent studies on improving the efficiency of sensing spectrum provide the insight and potential possibility of reducing sensing errors along with low energy consumption [87, 88]. Besides minimizing sensing errors, many studies (e.g., [89]) have been focused on the issue of improving the efficiency

of opportunistic spectrum sharing that could achieve the goal of maximizing downlink throughput without interfering with primary users. Hamdi, Zhang and Letaief [89] conducted pioneer work on opportunistic spectrum sharing in multi-input multi-output (MIMO) networks. In addition, Musavian and Aissa [15] present the capacity and power allocation for spectrum sharing in fading channels. Further, the studies of CR technique have been regarded in other communication networks, such as cellular networks [10]. In spite of the works on either sensing efficiency or sharing efficiency, queueing performance of users in CR networks is one the most import issues. To the best of our knowledge, it is hard to find the related works on the analysis of queueing behaviour of individual in CR networks. Especially, the scarcity of analytical models in presence of self-similar traffic regarding CR technique is still an open issue.

It is not hard to see the increasing demand of predicting and evaluating the performance of the communication networks or systems. Packets transmission relies on the scheduling adopted in networks. Hence, a powerful scheduling mechanism is a means of allocating the network resource effectively. Therefore, performance analysis of the scheduling systems is key measurement and evaluation of the corresponding networks. As aforementioned, there have been a lot of studies that focus on performance modelling for scheduling systems. However, it has been addressed in the previously that the vulnerabilities due to the inefficient traffic model which is inadequate of characterizing the self-similar nature exist. As a result, analytical models of innovative and hybrid scheduling mechanisms are developed in the following chapters.

Chapter 3

Measurement of Self-Similar Traffic

3.1 Introduction

Traffic measurement is an important means which is widely employed by service providers and analysts to make appropriate decisions and plan future developments [90]. In analytical models of practical communication systems, identification of the traffic pattern is always of great importance. The facts and experimental works show that traffic patterns have great impact on performance prediction, design and implementation of practical communication networks. Additionally, characterizing and featuring traffic properties are keys of effectively modelling designated communication systems.

Traffic measurement is not limited to the identification of traffic patterns or calculation of the traffic intensity of a specific arrival process. It provides the basis and fundamentals of closely modelling the corresponding practical traffic. For instance, by measuring the traffic of a call centre, researchers had found that a Poisson process can qualitatively model the telephone calls arriving to a switching system [6], since the inter-arrival times between consecutive calls can be well characterized by an exponential distribution. Without the knowledge of traffic measurement, it is difficult to establish analytical models of realistic traffic, and hence the communication systems. However, with the development of communication techniques, a large number of advanced applications have merged. As a consequence, the traffic arrivals of such applications are no longer as simple as

those in a call centre. Paxson and Floyd [7] have spotted the failure of using Poisson traffic and even the distribution of inter-arrival times between packets differs from the exponential distribution. As a result, more sophisticated traffic models which are capable of characterising the nature of arrivals in contemporary communication networks have drawn lots of attention. Particularly, the self-similar nature of network traffic has been found to consist in various and almost all communication networks, such as LAN [13], 802.11 wireless network [14] and Variable-Bit-Rate (VBR) video traffic [33].

In this new era, traffic patterns play an exceeding role analytical modelling. According to the aforementioned studies, contemporary communication network traffic exhibits dependencies over large time scales. It is shown that performance analysis of modern communication systems without taking traffic self-similarity into account may lead to the unexpected results [13, 91]. Hence, the analytical models subject to self-similar traffic are essential and promising.

What is the self-similarity? Self-similarity is a widespread and important phenomenon in nature. It is a remarkable ingredient of fractal processes reported during recent years. Back in 1920, the physicist, Lewis Fry Richardson, an expert on fluid turbulence, had spotted that bigger eddies have smaller eddies and so on. The observation has shown a phenomenon that an object is exactly or approximately similar to a part of itself. In other words, one or more parts of the object have the same shape as the whole. In statistics, the parts of the process present the same statistical properties at different scales [92]. There are many examples in nature which exhibit the self-similar phenomenon. Since Leland, Taqqu and Willinger [13] have reported the self-similar nature of Ethernet traffic, this phenomenon has been

pervasively examined in many kinds of communication network traffic. Measurements of traffic traces are a means which leads to the wide recognition of self-similarity in network traffic.

This chapter is organized as follows. Firstly, it presents the definition of self-similarity in mathematics. Next, four different Hurst parameter estimators will be presented. Particularly, this study reveals that the DoS/DDoS attack traffic traces exhibit self-similar nature.

3.2 Measurement of Hurst Parameter

3.2.1 Definitions

At the beginning, the so-called autocorrelation is essential and introduced. In statistics, autocorrelation [13] describes the correlation between the values at different time epochs of a designated stochastic process. It is an efficient mathematical tool for revealing repeating patterns which may be buried under noise. Let $\{X_t : t \in N\}$ be an arbitrary time-series. $ACR(k)$ denotes as the autocorrelation function, and then can be given by

$$ACR(k) = \frac{\mathrm{E}\left[(X_t - \mu)(X_{(t+k)} - \mu)\right]}{v^2} \tag{3.1}$$

where E(·) is the expectation, μ and v^2 are the mean and variance of the process respectively, k is the non-overlapping lag. $ACR(k)$ is ranged from $[-1,1]$ where 1 indicates perfect correlation, and -1 indicates anti-correlation. The correlation only depends on the non-overlapping lag k of the pair of values for second-order

stationary processes. In other words, if k remains unchanged, $ACR(k)$ is constant no matter the position of the value in the time series.

Having understood the properties of autocorrelation, the mathematical definition of self-similarity can be given accordingly. Time series X_t is said to be self-similar, if it meets the following condition [13, 91, 92]

$$ACR(k) \sim Ak^{-\beta_1}, \tag{3.2}$$

where $A > 0$ and $\beta_1 \in (0,1)$. It is worth noting that $\sim$ indicates asymptotically equal. As aforementioned, Hurst parameter is critical in measuring self-similar processes and can be calculated by

$$H = 1 - \frac{\beta_1}{2}. \tag{3.3}$$

In other definition, a time series X_t is said to be self-similar, if the following equation holds [93]

$$f(\psi) \sim A'|\psi|^{-\beta_2}, \tag{3.4}$$

where $A' > 0$, $\beta_2 \in (0,1)$, ψ is the frequency, and $f(\psi)$ is the spectral density function that can be calculated by autocorrelation function $ACR(k)$ and variance v^2 of the time series [93].

$$f(\psi) = \frac{v^2}{2\pi} \sum_{k=-\infty}^{\infty} ACR(k) e^{ik\psi}, \tag{3.5}$$

where $i = \sqrt{-1}$. Then the parameter β_2 has a relationship with Hurst parameter by

$$H = \frac{1 + \beta_2}{2}. \tag{3.6}$$

The degree of self-similarity is characterized by Hurst parameter if it is ranged from $(0.5,1)$ otherwise, the time series lose the property of self-similarity.

3.2.2 Hurst Parameter Estimators

Previously, the Hurst parameter is well-defined mathematically. However, measuring the Hurst parameter is still problematic. Each of current estimators for Hurst parameter has its advantage and weakness. The estimators that are introduced in this study are chosen for diversified purposes. Rescaled Range (R/S) estimator as a conventional means to calculate the Hurst parameter has been adopted in early years. Variance Time estimator is quickly converged as the amount of data collected. Periodogram estimator measures the Hurst parameter through frequency transforms.

A. Rescaled range estimator

R/S estimator has been proposed in early 1960s by Harold Edwin Hurst. The origin of R/S estimator is to provide a means of evaluating variability changes with the length of the time-ranged being concerned of a series and soon applied on the analysis of fractional Gaussian noises in water resources research [94]. In mathematical, let $R(n)$ be the range of the rescaled series, which is defined by

$$R(n) = \max(W_1, W_2, ...W_n) - \min(W_1, W_2, ...W_n), \tag{3.7}$$

where n is the block length, and W is defined as

$$W_k = (X_1 + X_2 + ... + X_k) - k\overline{X}(n), k = 1,2,3...n, \tag{3.8}$$

in which $\{X_t : t \in N\}$ is defined in Section 3.2.1 and μ is the mean of this series. S of "R/S" stands for the stand deviation of the rescaled series W and it is denoted in this thesis by $S(n)$. $S(n)$ can be calculated by the following equation [13, 93].

$$S(k) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(X_i - \overline{X}(k))^2}, k = 1,2,3,...n. \quad (3.9)$$

Then, according to Hurst's found, it is given that

$$\mathrm{E}[R(n)/S(n)] \sim A_H n^H, \quad (3.10)$$

where A_H is a positive finite constant independent of n. Hence, if log-log plot is applied on Equation (3.10), we have

$$\log(\mathrm{E}[R(n)/S(n)]) \sim \log A_H + H \log n. \quad (3.11)$$

It is not hard to see that the Hurst parameter H is the slope of the line which presents Equation (3.11). R/S estimator is preferred in many studies, such as the work done by Leland, et al [13] on examining the self-similar nature of Ethernet traffic. However, there are arguments on employing R/S estimator to calculate Hurst parameter. The issue mainly lies on how to choosing the values of n. Small or large values of n may lead the unexpected estimation results of Hurst parameter.

B. Variance time estimator

Hurst effect can be observed in most of time series. The accuracy of estimating the Hurst parameter of a designated series relies on a large number of collected data. Variance Time (VT) estimator [13, 91, 92, 93], also known as aggregated variance, which will be introduced in this section has low computational complexity. The principle of VT estimator is to aggregate the original time series X_t, $\{X_t : t \in N\}$ into a newly formed series by the block size m. The total number of blocks is $B = N/m$ where N is the length of the collected data. The aggregated process is composed of mean values over each block, which can be obtained by

$$X_b^{(m)} = \frac{1}{m}(X_{bm-m+1} + X_{bm-m+2} + ... + X_{bm}), b = 1,2,3...B. \qquad (3.12)$$

Then, according to the study in the literature [32], we have

$$\text{var}(X^{(m)}) \sim A_v m^{-\beta_3}, \qquad (3.13)$$

where A_v is a finite positive value which is independence of block size m. β_3 has a relationship with Hurst parameter by $H = (1-\beta_3)/2$. Still, log-log is applied on Equation (3.13), then it is given that

$$\log \text{var}(X^{(m)}) \sim -\beta_3 \log m + \log A_v. \qquad (3.14)$$

By selecting different block size m, a number of points can be drawn by plotting $\log \text{var}(X^{(m)})$ against $\log m$. These points are converged on a line which is present by Equation (3.14) with its slope $-\beta_3$.

C. Periodogram estimator

Periodogram estimator is a more refined data analysis approach for Hurst parameter. It is based on the analysis of spectral density of a series data set. This estimator is pervasively employed for estimating the Hurst parameter for Gaussian sequences. According to Equations (3.4) and (3.5), the Hurst parameter can be obtained if the spectral density function $f(\psi)$ can be calculated. Then the periodogram of a given time series of length N can be defined by [95]

$$I(\psi) = \frac{1}{2\pi N}\left|\sum_{j=1}^{N} X_j e^{ij\psi}\right|, \qquad (3.15)$$

where ψ and i are the frequency and the imaginary part defined in Section 3.2.1, respectively. Having understood the definition of self-similarity according to the spectral density, it can be given

$$I(\psi) \sim A'|\psi|^{-\beta_2}. \quad (3.16)$$

As aforementioned, the Hurst parameter can be obtained through Equation (3.6) by applying log-log on Equation (3.16).

3.2.3 Traffic Measurement and Results

In this subsection, the aforementioned three estimators are applied to calculate the Hurst parameter. As a start, the traffic traces that will be used are collected by MIT Lincoln lab for Defence Advanced research Projects Agency (DARPA) [96] which is a department of developing many technologies, including networking [19]. Afterwards, this study aims to investigate the self-similar nature of attacked traffic traces.

A. Sample data collected by Lincoln Lab

MIT Lincoln lab establishes a simulation network which collected the off-line network traffic and audit logs. The traffic traces provided by Lincoln lab are provided five weeks. The designated data sets which are preferred in this thesis are outside sniffing and inside sniffing data collected on Thursday and Friday of the first week. These four data sets are attack free, which means the simulation of the established network by Lincoln lab emulates the normal behaviours and activities of network users.

The outside sniffing and inside sniffing data are collected during continuously 22 hours a day. Figures 3.1-3.4 are plotted by the number of packets against the time epoch in second. A time epoch presents when the value of packet number is record. For example, if a value of the number of packets is record at 03:10:10 am, the time epoch is $3*60*60+10*60+10=11410$.

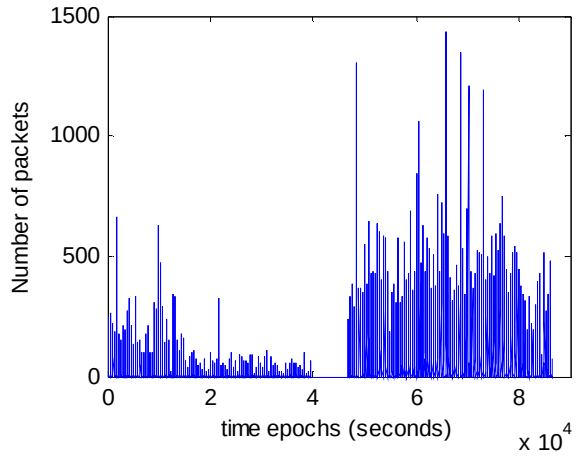


Figure 3.1: Outside sniffing dataset collected on Friday of the first week.

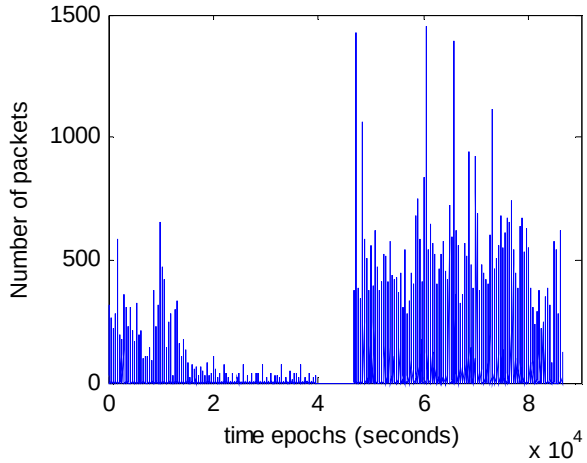


Figure 3.2: Inside sniffing dataset collected on Friday of the first week.

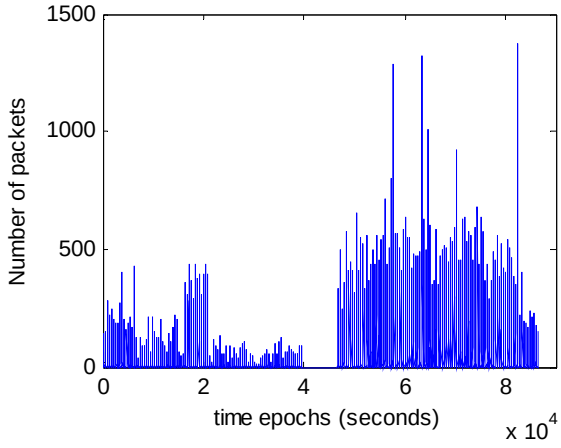


Figure 3.3: Outside sniffing dataset collected on Thursday of the first week.

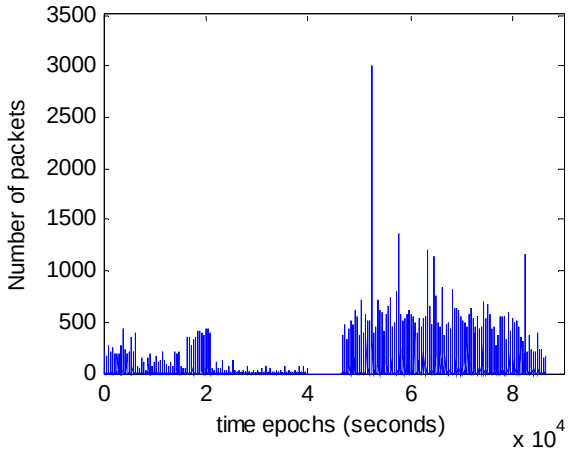
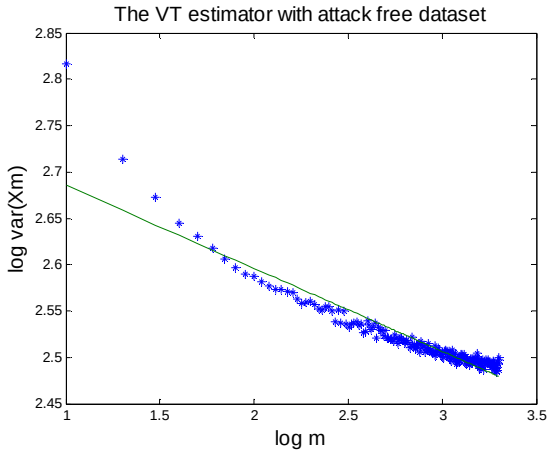


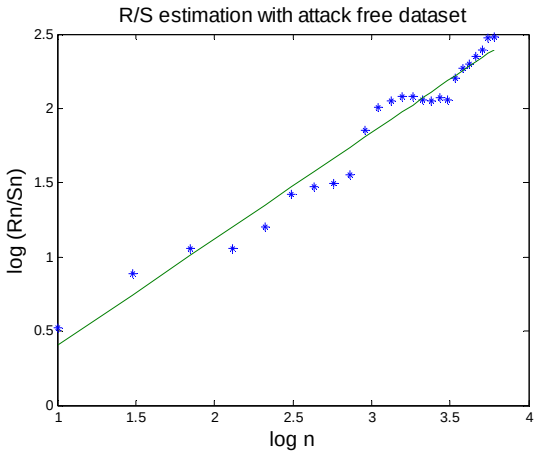
Figure 3.4: Inside sniffing dataset collected on Thursday of the first week.

The gap between time epoch intervals (40000,50000) is the time for server maintaining and resetting. From Figures 3.1-3.4, it is not hard to see that the network is busy between time interval (50000,80000) and has low activities during interval (20000,40000) for each day.

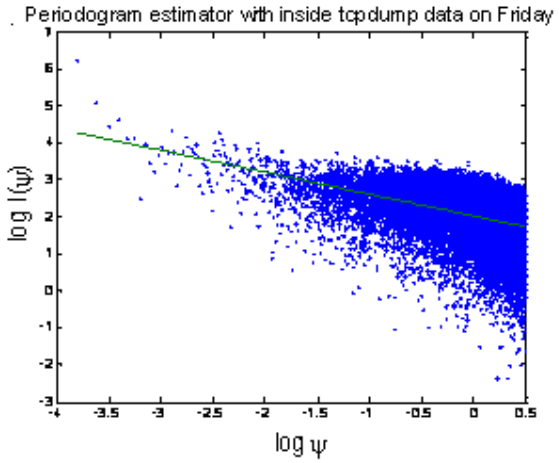
If the estimators are applied on these two representative datasets, the Hurst parameter can be calculated (shown in Table 3.1).



(a) VT estimator



(b)R/S estimator



(c) Periodogram estimator

Figure 3.5: Hurst parameter approximations by using inside sniffing dataset collected on Friday.

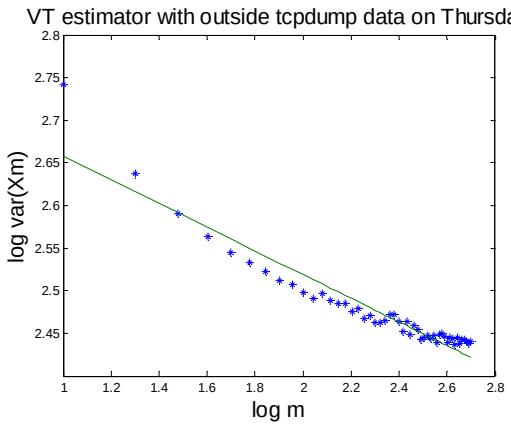
Table 3.1
Hurst parameter estimations of inside sniffing dataset collected on Friday

Estimators	Result of H
VT(m=100)	0.767
R/S	0.743
Periodogram	0.749

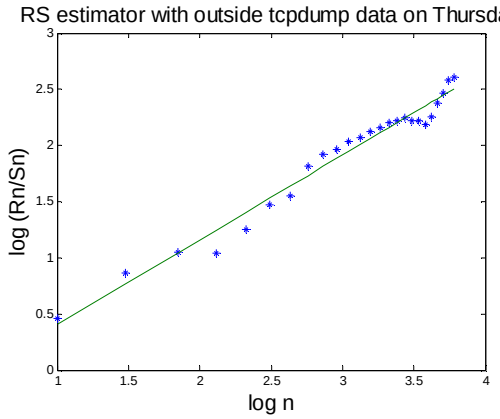
The lines in Figure 3.5(a-c) are mentioned in Section 3.2.2, and the slopes the these lines are the parameters β_3, H, β_2, respectively. Table 3.1 presents the Hurst parameter approximations according to the aforementioned estimators. It is clear that all three estimators can measure the desired value in a good degree. The estimated Hurst parameters suggest that the traffic trace collected on Friday by Lincoln lab presents self-similar phenomenon. In addition, the experiments show that the traffic generated by the normal behaviours and activities of network users comply with self-similar nature, which is critical to researchers who employ self-similar traffic as the inputs or arrivals in their works.

Figure 3.6 shows the Hurst parameter approximations by using the outside sniffing dataset collected on Thursday by Lincoln Lab. It is readily to see that the phenomena which are observed in Figure 3.5 still hold in this part. It is worth noting that the result calculated by the periodogram estimator is larger than those of VT and R/S estimators. Hurst parameter approximations from various estimators may have difference when the actual Hurst parameter of the traffic approaches to the bounds.

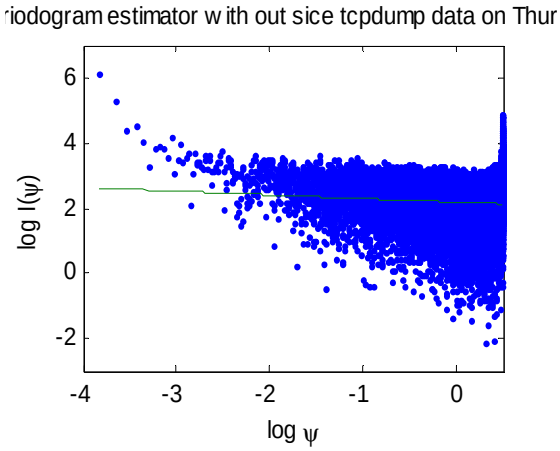
From the derivation of each Hurst parameter estimator, it is not hard to see that the VT and R/S estimators have low computational complexity. In addition, the number of the plotted points which is required by fitting the line is much less than that of the periodogram estimator. It is an enabler of the evaluation and measurement of the Hurst parameter of real-time traffic in communication networks. However, the advantage of the periodogram estimator lies its capability of accurately calculate the Hurst parameter of the traffic where hidden periodic events or signals occur. Based on the observations of Hurst parameter estimations, the VT estimator is preferred in this thesis.



(a) VT estimator



(b)R/S estimator



(c) Periodogram estimator

Figure 3.6: Hurst parameter approximations by using outside sniffing dataset collected on Thursday.

Table 3.2
Hurst parameter estimations of outside sniffing dataset collected on Thursday

Estimators	Result of H
VT(m=100)	0.895
R/S	0.879
Periodogram	0.93

The above datasets which are used to examine the Hurst parameter are attack free. It can bee seen that such kind of traffic which is generated by the normal behaviours and activities of the network users exhibits self-similar nature. Further, the traffic trace which contains attack or malicious activities is investigated. Does the attacked traffic still exhibit the self-similar nature? This question will be addressed in what follows.

B. Datasets with labelled attack

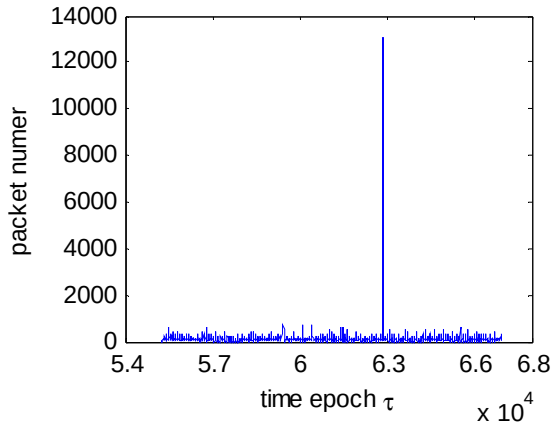
Among the recognized attacks, Denial-of-Service (DoS) attack and its variant Distributed Denial-of-Service (DDoS) attack are the worst of its kind [97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107]. Recent years, DoS/DDoS has been developed as a highly distributed attack tool arising from its original appearance. The attackers may scan millions of legitimate users for vulnerable weaknesses and seek for potential victims [108, 109]. A significant symptom during DDoS attack is that a large amount of malicious traffic is pumped out by the users which are manipulated by the intruders or attackers to overwhelm the victims.

In this chapter, the Hurst parameter estimators are applied on traffic traces that include representative DDoS attack scenarios. The first attacked dataset is still provided by Lincoln lab. The attack scenario is performed over multiple audit sessions which have been classified into 5 attack phases. The first four phase are the

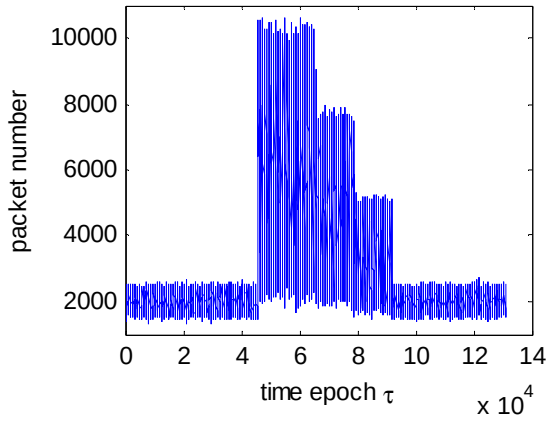
preparations for launching attack, e.g., probing of live IP, breaking via vulnerability and injecting daemons. The fifth phase is 5-seconds-length attack. The attack tool which is used in this scenario to launch the attack is Trojan mstream DDoS software. When this daemon is injected into the machines, it generates and sends the DDoS attack packets to victims.

Figure 3.7(a) shows the Lincoln lab DoS attacked raw data, from which one can easily observe that the attack duration around time epoch 63000. VT, R/S and periodogram estimators are applied on this sample data, and obtain the approximations of Hurst parameter as 0.59, 0.67 and 0.63, respectively. The results suggest that the attacked dataset provided by Lincoln lab exhibits self-similar nature.

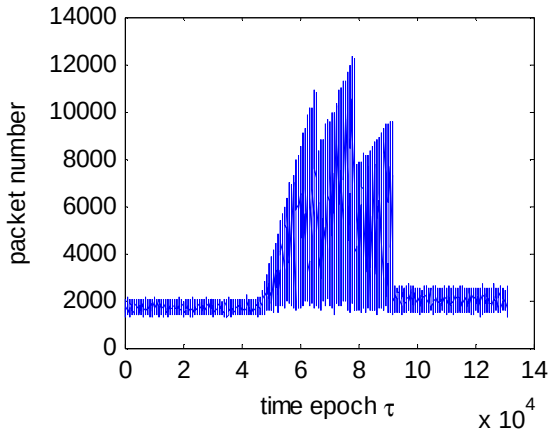
However, the attack duration of the Lincoln lab Dos data is only 5 seconds, thus it is hard to observe the changes of Hurst parameter during or after the attack phase. Therefore, a simulation environment which is composed of 10 nodes/end users is established. All nodes emulate the normal activities before the attack begins. When the attack is initiated, the zombie nodes which are manipulated by the intruder will send malicious flood traffic while legitimate nodes remain their behaviours. In our simulation, 30% of the total nodes are randomly configured as zombies to pump out malicious traffic. The attack traffic traces are generated according to the three attacking scenarios which are constant intensity, ramp up behaviour and pulse. Moreover, the normal traffic and the malicious traffic are generated by conditionalized random midpoint displacement algorithm [37, 38].



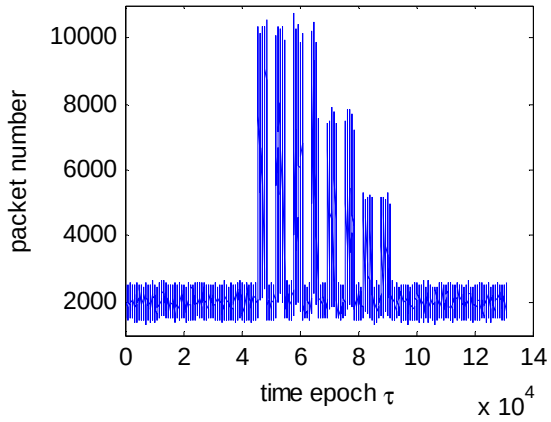
(a) MIT Lincoln Lab DoS attack data



(b) DDoS attack data simulation-constant intensity



(c) DDoS attack data simulation-ramp up

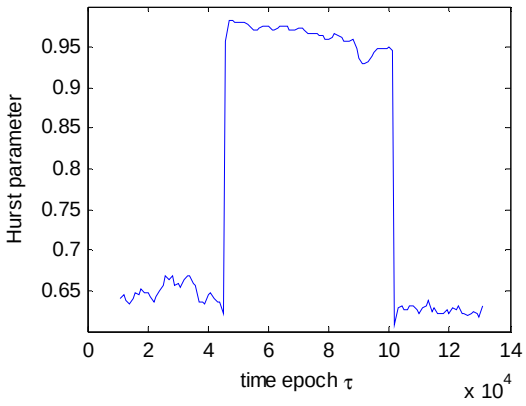


(d) DDoS attack data simulation-pulse

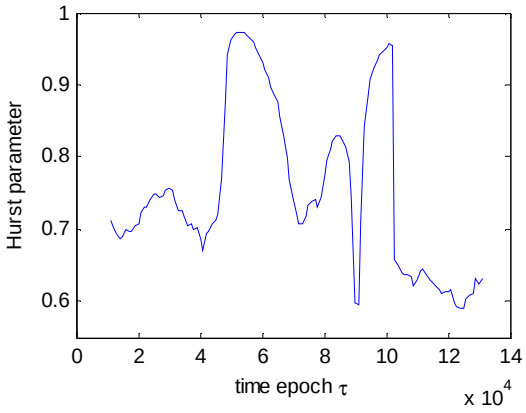
Figure 3.7: Sample datasets generating from test-bed of MIT and simulations

Figure 3.7(b) is a typical data which is collected from constant intensity attack scenario. Zombies in this scenario randomly start sending flood traffic after receiving the attack command from the intruder. In addition, the flood traffic sending by zombies will be randomly terminated. This is because the malicious traffic may be affected by user behaviours or unforeseen circumstance during the attack, which result in the stochastic attack duration and ending phase. Zombies in this scenario send the malicious traffic with a constant intensity (i.e., traffic rate). The dataset collected for Hurst parameter estimation is actually a superposed traffic of all the nodes. It is practical and can be readily implemented in real-world system by virtual of a firewall or classifier.

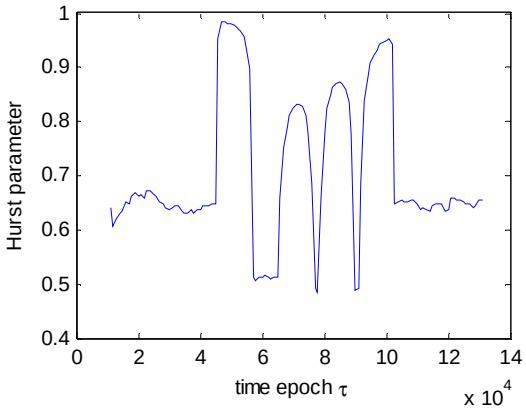
The second scenario which is shown in Figure 3.7(c) configures the initial intensity of each zombie to be 10% of the constant intensity scenario. This intensity will be increase by 10% of the constant intensity every 30 seconds until it reaches the peak. Such kind of configuration of the attack intensity emulates the behaviour of smart intruders who are trying to evade the filters or scanning.



(a) With constant intensity dataset



(b) With ramp-up behaviour dataset



(c) With pulse behaviour dataset

Figure 3.8: Hurst parameter approximation for attacked traces

The traffic of pulse attack scenario has similar shape as that of human being's pulse. The attack duration and starting time of each zombie is still random after receiving attack command. Compared to the previous two scenarios, the attack intensity adjusts like ON-OFF switch. When that attack switches to the ON state, the attack intensity is set to constant intensity attack with full power, otherwise 0.

Next, the Hurst parameter is evaluated when every 1000 time epochs added into the corresponding traffic trace. From Figure 3.8(a-c), it can be observed that the Hurst parameter is stable or weak stationary before the attack. Then the Hurst parameter significantly changes during the attack and remain stationary when the attack is over. Although the Hurst parameter fluctuates drastically, it is sill within the range $(0.5, 1)$. Hence the attacked traffic traces maintain the self-similar nature.

More importantly, the Hurst parameter estimations of the three scenarios have suggested that the superposed traffic still exhibits the self-similar nature, which is a critical support of performance analysis of queueing systems subject to multi self-similar arrivals.

3.3 Summary

Modelling of traffic pattern plays a crucial role of the performance analysis of communication systems. Recently, traffic self-similarity has been shown to be a ubiquitous phenomenon in most communication networks. This chapter has dedicated to the estimation and measurement on self-similar nature of traffic by means of the Hurst parameter. Three commonly used Hurst parameter estimators haven been reviewed and applied on the traffic datasets, which are Variance Time, Rescaled Range and Periodogram estimators.

The results of the estimated Hurst parameter show that the normal traffic in practical networks present self-similar characteristic. Furthermore, the Hurst parameter of Denial-of-Service (DoS) attacked traffic traces has been examined. The dataset collected by Lincoln lab with labelled attack has been proven to exhibit self-similar nature. Due to the insufficient attack duration, the Hurst parameter estimators are applied on the simulation datasets. These datasets depict the traffic which is attacked by selected DoS scenarios. By examining the Hurst parameter, it can be observed that self-similarity consists in the attacked traffic, though the Hurst parameter has changed significantly during the attack period.

Through the measurements, it is ready to see that self-similar nature pervasively exists in network traffic, no matter the normal traffic or attacked traffic. The research works and studies based on traffic self-similarity are reasonable and hence reliable.

Chapter 4

Analytical Modelling of a Deficit Round Robin Scheduling System in the Presence of Self-similar Traffic

4.1 Introduction

A variety of applications in communication networks commonly share the links for their respective transmission of data, voice traffic, or video traffic, in which traffic flows are usually contending for the unique service. On the other hand, the contending traffic flows belonging to the same category are expected to receive the identical service opportunity. As a consequence, allocation of the precious bandwidth to all traffic flows in a fair and efficient manner becomes a desirable QoS demand.

To address the issue on fair resource allocation and complexity at the mean time, a promising modified round robin mechanism, namely, DRR is addressed in this chapter, which is able to achieve excellent fairness in terms of throughput and has very low complexity, $O(1)$ [42]. The unused service volume of a traffic flow in the current round will be stored as deficit for next round, which is a key feature of DRR. By doing this, the service capacity which should be deserved by one traffic flow cannot be plundered by other traffic flows. Fairly handling packets of variable size from different traffic flows without knowing their mean packet size is a

significant advantage of DRR. DRR scheduling mechanism has been widely implemented in practical commercial high-speed routers, i.e., Cisco 12000 series internet routers. Consequently, many research efforts have been made to analyze the performance of DRR systems on various communication networks.

This chapter aims to analytically investigate the queueing performance of the DRR scheduling scheme. A modulation approach and a apriority transfer approach are developed to calculate the desired upper and lower bounds, respectively. Specifically, the priority of the designated flow is downgraded and it is thus more likely to share its service with the other flow. In other words, the traffic flow which gives out its priority receives less service as compared to that in the original system. In this way, the upper bound of the corresponding queue length is obtained. On the other hand, the lower bound is obtained by applying the modulation approach which "smoothes" the arrival rate of one flow and hence the other traffic flow can receive more service than expected. Hence the lower bound of queue length of this traffic flow can be readily calculated. Next, the accuracy of developed bounds is validated through the comparison between the simulation experiments and the analytical results under various representative scenarios in this chapter. Further, to illustrate its utilization and application, the analytical model is employed to investigate the impact of packet size on queue length. Finally, the mode is applied to evaluating the effects of weight combination of traffic flows on the queueing performance.

The rest of this chapter is organized as follows. In Section 4.2, the designated DRR system and nested stochastic processes which are adopted to address the variable packet size will be introduced. The analytical model of obtaining the upper and lower bounds of queue length will be demonstrated in Section 4.3. Then, the

validation of the model under various setting scenarios will be presented in Section 4.4. In Section 4.5, the model is applied to investigate the effects of weight combinations of traffic flows on queue length. Finally, Section 4.6 is the summary of this chapter.

4.2 System Description and Nested Process

This section introduces the targeted DRR scheduling system at the first place. Next, the modelling issue of variable packet sizes by employing a nested stochastic process is addressed.

4.2.1 DRR Scheduling System

Figure 4.1 presents the schematic diagram of a DRR scheduling system. According to the DRR scheduling mechanism, an active traffic flow $F_i, i=1,2$ is assigned a pre-defined quantum size $quantum_i, i=1,2$ accordingly, and then the maximum volume of service is the number of units denoted and stored by the corresponding deficit counter, i.e., $DC_i, i=1,2$. It is worth noting that the minimum service unit is bit and each packet is usually composed of a number of bits. By the DRR scheduling mechanism, the remaining quantum which is not enough to cover the head-of-line packet in buffers will be saved in DC_i for the next round. If the queue of the current flow is empty, DC_i is set to be zero and the system immediately switches to the next active flow. To keep tracking the active flows, DRR maintains a list which contains the tags of active flows. A tag in the list is removed if the corresponding queue is empty. Otherwise, the tag will be added to the end of the list upon the packet arrival. The inputs for the system are denoted by

$fBm1$ and $fBm2$, respectively. Figure 4.1 clearly demonstrates that the packets feed into the system are of different size.

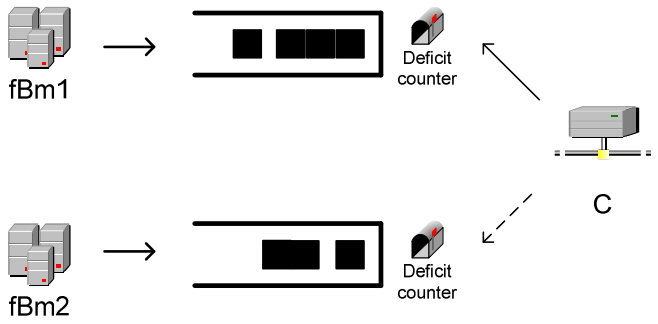


Figure 4.1: DRR scheduling system subject to self-similar traffic.

4.2.2 Nested Stochastic Processes

Firstly, there is a fact that packets of network traffic are normally made up of smaller units, namely, bits. If a packet fBm arrival process is formulized by $A_i(t)$

$$A_i(t) = m_i t + \sqrt{a_i m_i}\,\overline{Z}(t), \tag{4.1}$$

where m_i and a_i are the mean arrival rate and the variance coefficient of the corresponding fBm process. $\overline{Z}(t)$ is a centred fBm process which is defined in Chapter 2. With different packet sizes in terms of bit, a new arrival process can be obtained from $A_i(t)$. The derivation of such a process, denoted by $A_i'(t)$ is actually a nested stochastic process and can also be approximated as fBm [110]. If the mean and variance of the distribution of packet sizes are denoted by μ' and v', respectively. The mean arrival rate and variance of $A_i'(t)$ can then be given by

$$m_i' = m_i \mu', \tag{4.2}$$

$$v_i^{'}(t) = a_i m_i \mu'^2 + m_i v' . \quad (4.3)$$

Based on Equation (4.3), the variance coefficient of $A_i'(t)$ can be readily denoted as

$$a_i' = a\mu' + v'/\mu' . \quad (4.4)$$

Note that the Hurst parameter of $A_i'(t)$ remains the same as the $A_i(t)$ [110]. To model the nested process, the distribution of the packet size plays a critical role. Previous studies have concluded that the exponential distribution can be used to approximate the packet size distribution of Internet traffic [111], which makes it practical in analytical works which take packet size into account. In addition, exponentially distributed packet size has finite mean and variance, and hence makes the derivation of analytical models tractable. For these reasons, the exponential distribution is adopted to model the packet size in this thesis. And hence, it is an enabler of using nested process to address variable packet size issue in the developed analytical model.

4.3 Analytical upper and lower bounds of queue length distribution

This section presents the approaches to analyze the performance of the DRR scheduling system. It is a challenge job to find an analytical approach to directly and exactly obtaining the queueing performance of individual traffic flows, for the reason that the interaction between the flows scheduled by DRR. As a result, a modulation method which converts the original DRR scheduled system into a collection of Single-Server Single-Queue (SSSQ) systems is presented. Consequently, the inter-relationship between the traffic flows has been eliminated. Theoretically, the queue of the modulated flow is expected to keep empty in the original DRR system through the modulation of the mean arrival rate of one traffic flow. By doing this, the queue

length of the modulated flow has no impact on the other flow. Subsequently, the lower bound of the queue length distribution can be obtained by solving a SSSQ system. On the other hand, the upper bound is calculated by converting the DRR system into a priority Queueing (PQ) system, which implies that one traffic flow receives less service than expected by giving up its priority.

4.3.1 Modulation Approach

An active flow i is expected to receive the service no less than $C \cdot \phi_i / F$, where C is the total service capacity of the original DRR system, ϕ_i is the quantum size assigned to flow i in each round, F is the frame size which is defined as $F = \sum_{j=1}^{n} \phi_j$, and n ($n = 2$) is the number of traffic flows scheduled by DRR. This minimum service capacity is referred as the guaranteed service capacity in the rest of this chapter.

Since the complexity of its dynamic bandwidth sharing policy among active flows, there is hardly a practical method of obtaining the original service for individual traffic flow handled by DRR scheduling. Alternatively, an approach is proposed to modulating the arrival in order to avoid the impact caused by the interaction between the contending traffic flows. Finally, the goal of decomposing the complex DRR system into SSSQs can be achieved. Specifically, it takes two steps to obtain the SSSQs from the original DRR system through modulation.

For the first step, the procedure of modulating the service rate of traffic flow in order to eliminate the interacting impact is addressed. To demonstrate the approach, the dynamic queue of the DRR system is defined by

$$Q_1(t) = A_1(t) - \left(\phi_1 C / F + \max\{(\phi_2 C / F - A_2(t)), 0\}\right). \quad (4.5)$$

The above equation shows the relationship between traffic flow 1, i.e., $A_1(t)$ and traffic flow 2, i.e., $A_2(t)$. If the queue of flow 2 is always empty, the service received by flow 1 can be readily denoted as $C - \min\{A_2(t), \phi_2 C / F\}$. In this way, it is possible to isolate flow 1 from the original system by a modulated service. As a result, the queue length of this SSSQ can be readily given as

$$\tilde{Q}_1(t) = A_1(t) - \left(C - \min\{A_2(t), \phi_2 C / F\}\right). \quad (4.6)$$

From the comparison between Equation (4.5) and Equation (4.6), it is clear that the modulated queue is a lower bound which is formulized by Equation (4.7)

$$\tilde{Q}_1(t) \le Q_1(t). \quad (4.7)$$

Now, the two steps of obtaining the lower bound of queue length will be presented, which are shown in Figure 4.2. It is readily to see that first step modulates the mean arrival rate of flow 2 with its guaranteed service only. By doing this, it is capable of eliminating its queueing effects on flow 1. The goal of the modulation approach is to find a smoothed arrival process with expected mean arrival rate m_2^s with which the queue is always empty subject to the modulated arrival under the corresponding guaranteed service which is $g_2 = C \cdot \phi_2 / F$. To achieve the goal of modulation on arrival rate, an iterative approach is employed. According the aforementioned modulation theory, the expected modulated arrival rate can be obtained by examining an SSSQ system. Following the method presented in [112], the distribution of the queue length of an SSSQ system subject to fBm arrivals can be obtained by

$$P(Q > x) \approx e^{-\alpha^2/2}, \quad (4.8)$$

where α here is the minimum value of function $Y(t)$ which is given by

$$Y(t) = \frac{x + (sc - m)t}{\sqrt{amt^{2H}}}, t > 0. \tag{4.9}$$

To distinguish the previous definition, *sc* is defined as the service capacity of a general SSSQ system. a, m, and H characterize the fBm arrival process which is the input of this general SSSQ system. x is the queue length. Function $Y'(t)$ can be obtained by differentiating Equation (4.9), and then solve $Y'(t) = 0$. Hence we have

$$t_{dts} = \frac{Hx}{(C - m_2)(H - 1)}, \tag{4.10}$$

t_{dts} is one value of t in Equation (4.9) when $Y(t)$ attains its minimum, which indicates the most probable time scale with which overflow occurs [112].

In what follows, the corresponding parameters of Equation (4.9) are replaced in order to calculate the queue length distribution of modulated flow 2. Thus it is given that

$$Y^s(t) = \frac{x + (g_2 - m_2^s)t}{\sqrt{a_2 m_2^s t^{2H_2}}}, t > 0. \tag{4.11}$$

By combing Equations (4.10) and (4.11), the distribution that queue length greater than 1 subject to the modulated flow 2 is given by

$$P(Q_2 > 1) = \exp\left(-\frac{1}{2} \frac{(1 + (g_2 - m_2^s)t_{dts})^2}{a_2 m_2^s t_{dts}^{2H}} \right). \tag{4.12}$$

Then the modulated arrival rate m_2^s is obtained such that the corresponding queue is empty (i.e., $P(Q > 1) \to 0$).

In the second step of modulation approach which is shown in Figure 4.2, the modulated process with arrival rate m_2^s is fed back to the DRR system. As the actual service capacity assigned to flow 2 in the original DRR system is no less than its

guaranteed service capacity, g_2, the corresponding queue subject to the modulated process is always empty.

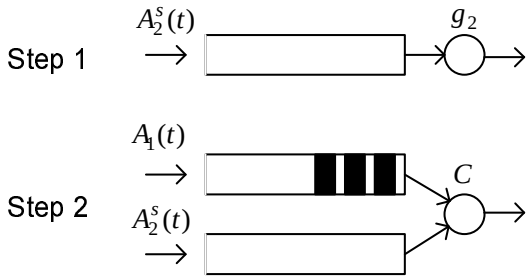


Figure 4.2: The modulation approach to obtain the lower bound

From Figure 4.2, Step 2, it is clear that the total queue is exclusively composed of that of flow 1 in the modulated DRR system. In other words, the queueing impact of the modulated flow 2 is negligible. Therefore, it is reasonable to use the total queue length of the modulated DRR system to approximate that of flow 1. Consequently, the queueing performance of flow 1 can be obtained by solving an SSSQ system subject to the superposition of the original arrival process of flow 1 and the modulated arrival process of flow 2 under the capacity of the DRR system, C. This is actually a lower bound since flow 1 receives more service in the modulated system than that in the original system. The desired distribution of queue length of flow 1 can be obtained according to Equations (4.9) and (4.10), if the arrival processes of flows 1 and 2 in the modulated system can be merged as one. To deal with the merging issue of self-similar traffic flows, Fan and Georganas [113] showed that the superposition of two self-similar traffic flows still holds the self-similar nature. More specifically, the mean arrival rate of the superposed traffic is the sum of the mean arrival rates of the original traffic flows. The Hurst parameter is equal to the larger one of the original traffic flows; and the variance is the sum of

those of the original flows. According to their study, the descriptors of the self-similar process superposed from $A_1(t)$ and $A_2(t)$ are given by

$$m_{total} = m_1 + m_{2,} \tag{4.12}$$

$$H_{total} = \max(H_{1,} H_2), \tag{4.13}$$

$$a_{total} m_{total} = a_1 m_1 + a_2 m_2. \tag{4.14}$$

So far, the derivations of obtaining the lower bound of the distribution of queue length are based on packet level. Next, packet size is taken into account by means of a nested process presented in Section 4.2.2. Then, we combine Equations (4.2), (4.3), (4.4), (4.9) and (4.10) as follows

$$Y^p(t) = \frac{x\mu' + (C - m_2^s - m_1)\mu' t}{\sqrt{a_1 m_1 \mu'^2 + m_1 v' + a_2 m_2^s \mu'^2 + m_2^s v'}}, t > 0. \tag{4.15}$$

Finally, the lower bound of queue length distribution of flow 1 can be calculated by replacing α in Equation (4.8) with the minimum value of $Y^p(t)$. Following the same way, the lower bound of queue length distribution of flow 2 can be obtained by modulating the arrival rate of flow 1.

4.3.2 Priority Conversion Approach

The approach of obtaining the lower bound has been demonstrated in the previous subsection. This subsection deals with the upper bound of the queue length distribution of each flow in the DRR system. Understanding that the guaranteed service capacity for a traffic flow scheduled by DRR scheduling is the minimum service it can receive, it is obvious that a rough upper bound for the queue length distribution of one flow in the DRR system can be obtained by solving an SSSQ system subject to its original input under the guaranteed service capacity. However,

such an upper bound is rather loose. Alternatively, a priority conversion approach which gives out the deserved service of one traffic flow to another is developed. According to the DRR scheduling mechanism, it provides a guaranteed service capacity to each flow and allocates the exceeding service to active flows. Moreover, DRR is a fair scheduling mechanism inherent. If the DRR system is converted into a PQ system, the lower priority part of this PQ system obviously receives less service than that of the original DRR system. In this way, the upper bound of the queue length distribution of the individual flows can be obtained.

Figure 4.3 reveals that the DRR system is converted into a PQ system by setting the high priority to flow 2. Then, the upper bound of the queue length distribution of flow 1 is calculated. According to the Empty Buffer Approximation (EBA) method which is widely adopted to model the lower priority part in a two-queue PQ system [114], the queue of the system is almost exclusively composed of the lower priority traffic and hence it can be used to approximate that of its lower priority part. Hence, the queue length distribution of the low priority part in the converted PQ system is derived by combing the nested process regarding the packet size as

$$P(Q > x) = \left(2\pi(1+\sqrt{U}\,(x\mu')^{1-H})^2\right)^{-1/4} \exp\left(-\frac{1}{2}U(x\mu')^{2-2H}\right), \qquad (4.16)$$

where U is defined as

$$U = \frac{\left((C\mu' - (m_1+m_2)\mu')(1-H)\right)^{2H}}{(\sum_{i=1}^{2} a_i m_i \mu'^2 + m_i \nu')(1-H)^2 H^{2H}}. \qquad (4.17)$$

Based on the priority approach, the queue length distribution of lower priority part of the converted PQ system is actually the upper bound of that of the original DRR system. Similarly, the upper bound corresponding to another flow can

be calculated by exchanging the roles of flow 1 and flow 2 in the previous derivations.

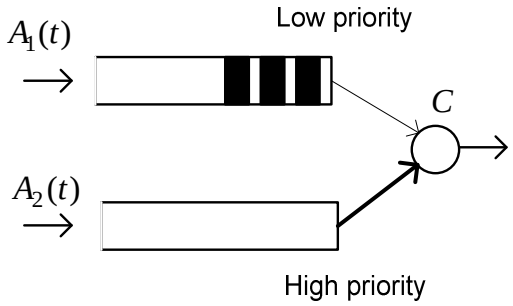


Figure 4.3: Priority conversion of the original system.

4.3.3 Further Discussion

In practice, there are several extreme cases in a DRR scheduling system. The following will address these cases and demonstrate how to analyze the corresponding queueing performance.

As a start, a simple extreme case which has $\exists\, i \in \{1, ..., n\},\ \phi_i / F \to 1$, where ϕ_i is the quantum of flow i, F is the frame size defined in previous subsection, and n is the number of flows is introduced. It is readily to see that the guaranteed service for flow i is approximately equals to 1. This is equivalent to a priority queueing system that set flow i with the highest priority. Hence, the queueing performance can be obtained by solving a PQ system.

As a fact that traffic flows scheduled by DRR commonly share the unique server, there is another extreme case where the arrival rate of one traffic flow is greater that its guaranteed service . Next, a two-queue DRR system is addressed in

such a case. Without loss of generality, we assume that the mean arrival rate of flow 1 is less than its guaranteed service i.e., $m_1 < C\phi_1 / F$ and that of flow 2 exceeds the corresponding guaranteed service i.e., $m_2 > C\phi_2 / F$. For obtaining a stable system, $m_1 + m_2 < C$ is required. It is obvious that under this setting, flow 2 with arrival rate m_2 has to take the residual service of flow 1 to clear its backlogs. On the other hand, flow 1 is not able to share service from flow 2, as its queue is seldom empty. Therefore, flow 1 can be directly isolated from the DRR system to be an SSSQ system subject to its original arrival with the corresponding guaranteed service. Consequently, the queue length distribution of flow 1 can be investigated by virtue of this SSSQ system.

For the other flow (i.e., flow 2), there is not a better way available to address its performance rather than the bounding approach. The analytical upper bound of its queue length distribution can be obtained by following the priority conversion approach and setting flow 1 as the high priority part. The modulation approach to obtaining the lower bound is still effective. However, it may not as accurate as normal cases where the arrival rates of both traffic flows are less than the corresponding guaranteed service. To bridge this gap, an alternative service capacity $C_2' > C_2$ is employed, where C_2 is the statistically equivalent service capacity received by flow 2 in the original system. C_2 is obtained by splitting the excess service of flow 1. We have validated that flow 1 receives $C\phi_1 / F$ when its queue is not empty. Hence, flow 2 may only share excess service from flow 1 when its queue is empty. If $P_1(0)$ denotes the probability that the queue of flow 1 is empty, the maximum excess service that flow 2 can share from flow 1 is $P_1(0)C\phi_1 / F$. However,

the queue of flow 2 is empty when the system is idle. $P_s(0)$ is defined as the probability that the system is idle. Consequently, we have $P_2(0) \geq P_s(0)$ and

$$(P_1(0) - P_s(0))C + (1 - P_1(0) - P_s(0))\phi_2 C / F > C_2 . \quad (4.18)$$

The service capacity C_2' can be obtained, i.e., the left side of Equation (22), with which flow 2 is able to be isolated from the original DRR system and hence derive the desired lower bound.

4.4 Model Validation

The developed analytical bounds of the queue length distributions of the DRR system are validated through extensive comparisons between analytical and simulation results. The simulator of the DRR scheduling system is programmed in C. The self-similar traffic are generated according to randomlized midpoint algorithm [91]. The results are obtained the developed discrete event simulation. Packet sizes are modelled by an exponential distribution with mean 10 bits and variance 100.

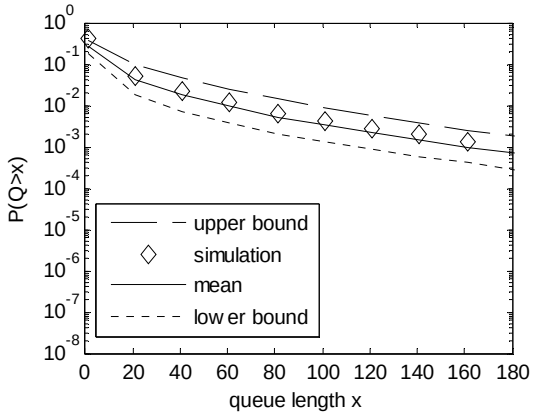
Table 4.1
The parameter settings of Cases 1, 2, and 3.

	c	m_1	m_2	a_1	a_2	H_1	H_2	ϕ_1	ϕ_2
Case 1	1800	95	70	1	2	0.7	0.7	400	300
Case 2	1000	50	35	1	2	0.8	0.8	599	399
Case 3	1800	100	65	1	1	0.8	0.8	1200	600

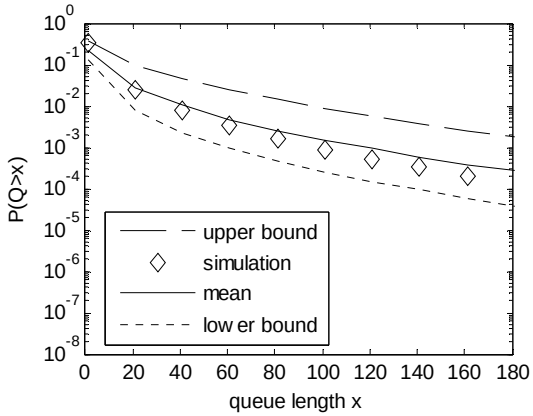
In what follows, the analytical and simulation results under three representative cases are presented, respectively, where the dashed, dotted, and solid lines denote upper bounds, lower bounds, and their geometric means. The parameter settings of these cases are presented in Table 4.1, where c denotes the service rate in

bits per second (i.e., $c = C\mu'$); the unit of arrival rate is packets per second and the unit of quantum is bit.

Figure 4.4 shows the comparison between the analytical and simulation results of individual traffic flows of the DRR system in Case 1. It is clear that the simulation results lies within the upper and lower bounds and the gap between the two bounds is tight. Through extensive experiments, the geometric mean of the upper and lower bounds can approximate the corresponding queue length.



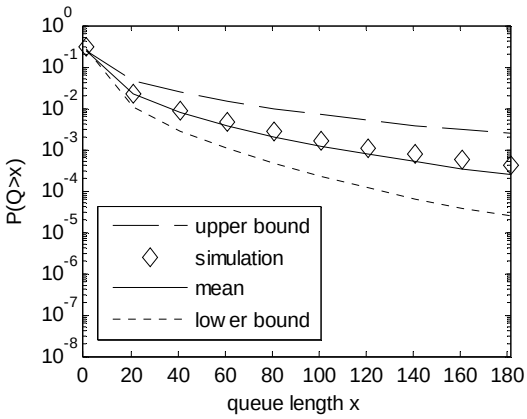
(a) Queue 1



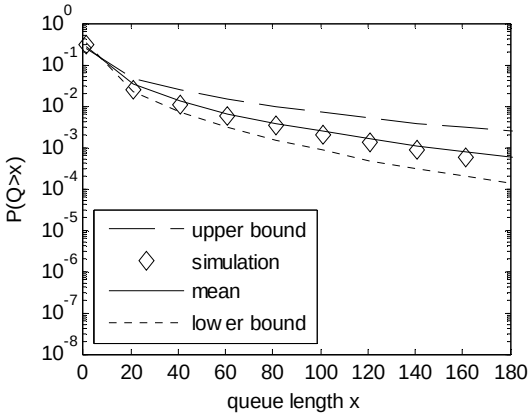
(b) Queue 2

Figure 4.4: The analytical and simulation results of the length distributions of the two queues in Case 1.

Figure 4.5 depicts the analytical and simulation results in Case 2. Compared to Case 1, the service rate, quantum size and arrival processes are changed. Particularly, the Hurst parameter of the arrival processes is altered. As the Hurst parameter has a significant impact on the performance of queueing systems, this change provides us with a good insight into the effectiveness of the developed model. In Figure 4.5, it is obvious that the phenomenon observed in Case 1 holds in Case 2 as well. In particular, the geometric mean of the upper and lower bounds is still a good approximation to queue length.



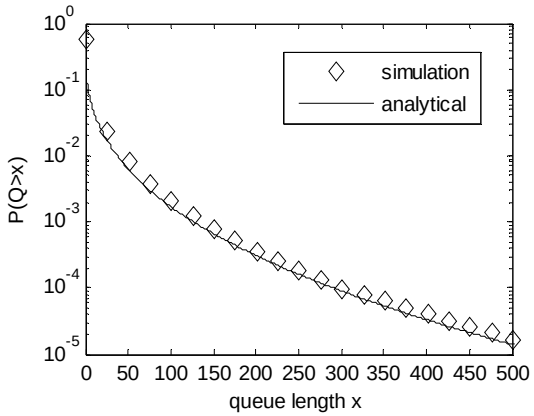
(a) Queue 1



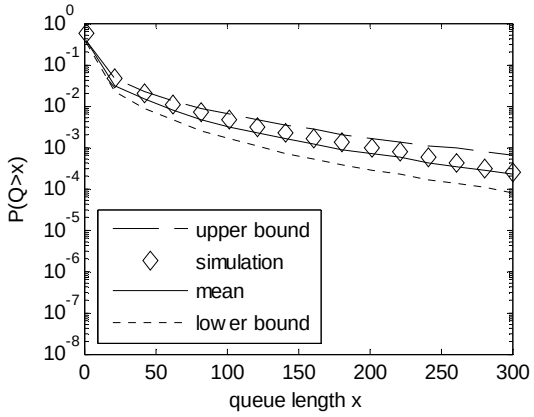
(b) Queue 2

Figure 4.5: The analytical and simulation results of the length distributions of the two queues in Case 2.

Case 3 is an extreme case which was discussed in Section 4.3.3. Figure 4.6 shows the analytical and simulation results. Particularly, Figure 4.6(a) represents only the simulation result of the queue length distribution of flow 1 and its approximation because of the reason specified in Section 4.3.3. For flow 2, it can be observed that the gap between the upper and lower bounds of its queue length distribution is narrow and the geometric mean as an approximation has an excellent agreement with the simulation result.



(a) Queue 1



(b) Queue 2

Figure 4.6: The analytical and simulation results of the queue length distribution of the two queues in Case 3 (Extreme case).

Based on the results shown in these figures, a conclusion that the proposed analytical model possesses a good degree of accuracy under different cases can be made.

4.5 Applications of the Model

In this section, the developed analytical model is adopted to evaluate the queueing performance under various combinations of the weights assigned to different traffic classes. Next, the impact of packet size on the performance of the DRR system is investigated.

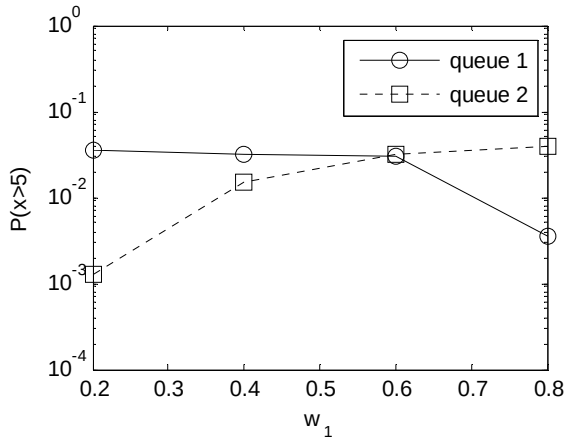


Figure 4.7: The impact on the queue length under various combinations of the weights of the DRR system.

Let us first define $w_i = \phi_i / F$ **(**i =1,2) as the weight of individual traffic flow i in the DRR system where $w_1 + w_2 = 1$. Then, the queue length distribution of both flows can be examined under various combinations of the weights of individual flows. Figure 4.7 presents the queue length distributions, i.e., $P(Q > 5)$, of both flows as the weight of flow 1 increases from 0.2 to 0.8. It is worth noting that the queue length distribution of flow 1 slightly decreases before the turning point

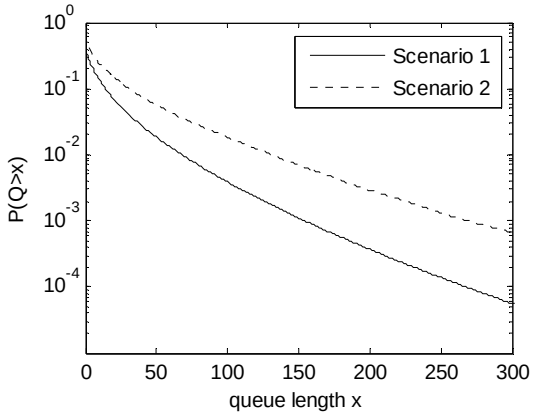
$w_1 = 0.6$ and then drops sharply. This is because the DRR system serves flow 1 under the extreme case when w_1 is small and the mean arrival rate of flow 1 excesses its guaranteed service capacity. After the turning point $w_1 = 0.6$, the mean arrival rate of flow 1 becomes less than the guaranteed service capacity and thus its queue length drops.

In the following, the impact of different mean packet sizes on the DRR system with the fixed mean arrival rates of traffic flows is evaluated. There are two cases. The parameter setting of the two cases is given in Table 4.2.

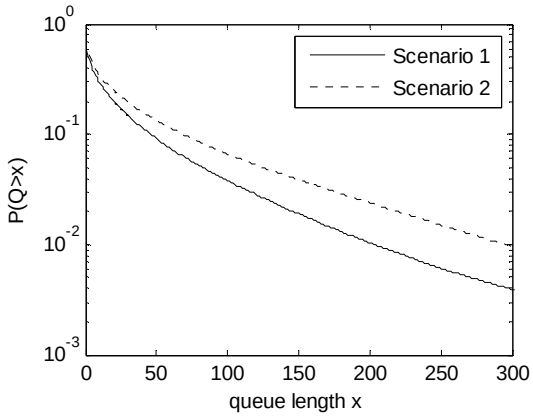
Table 4.2
The parameter settings of Cases 1 and 2.

	c	m_1	m_2	a_1	a_2	H_1	H_2	ϕ_1	ϕ_2	μ_1'
1	180	90	80	1	1	0.7	0.7	550	450	1
2	180	45	80	1	1	0.7	0.7	550	450	2

The packet size of both traffic flows follows an exponential distribution with mean 1. For the second case, the mean packet size of flow 1 (i.e. μ_1') is doubled meanwhile keep the mean arrival rate of the nested process of flow 1 unchanged. Figure 4.8 shows that the queue length of flow 1 in Case 2 increases because of the increase of packet size. Moreover, this phenomenon also exists for traffic flow 2 although its packet size remains. This is because that the excess service of flow 1 that can be shared in Case 1 is less than that in Case 2. Thus, the queue length of flow 2 increases in the case with the less service capacity. The above results demonstrate that the developed model provides us a cost-efficient tool to investigate the impact of the mean packet size and the combinations of the weights of individual flows.



(a) Queue 1



(b) Queue 2

Figure 4.8: The comparison of two cases with different mean packet sizes.

4.6 Summary

Fairly allocating precious bandwidth or network resource to a set of traffic flows sharing a common communication link or server is an increasingly pressing demand in contemporary communication networks. Among variants of the scheduling mechanisms which aim to fair service allocation, DRR is very promising due to its capability of serving packet flows with nearly perfect fairness in terms of throughput. Particularly, DRR is an effective scheduling mechanism because it is

able to address the unfairness caused by the variable packet size. However, it is challenging to directly isolate traffic flows of the complex DRR system into a collection of SSSQ systems because of the interactions. This chapter has developed a novel analytical model which proposes a modulation approach as well as a priority conversion method to derive both lower bound and upper bound for the queue length distribution of each traffic flow handled by the DRR scheduling scheme. Particularly, in the developed model, the effect of variable packet sizes on performance analysis of DRR systems has been fully taken into account. Through the comparison between analytical and simulation results, it has been validated that the developed model is effective and accurate to predict the queueing performance of DRR systems in the presence of self-similar traffic. Finally, the analytical model is applied to investigate the queue length under various combinations of the weights of individual flows and examine the impact of the packet size on the queueing performance.

Chapter 5

An Analytical Model of the Hybrid PQGPS Scheduling System Subject to Self-similar Traffic

5.1 Introduction

Service providers are interested in techniques that are able to handle diversified services required by different users. To this end, the provision of the differentiated QoS has emerged due to an increasingly pressing demand [2, 3, 115, 116, 117]. As a result, the DiffServ structure is proposed to meet the needs of classifying and managing network traffic which requires diversified QoS. To achieve the goal of DiffServ, a variety of scheduling mechanisms have been developed and further applied to communication networks which provide diversified QoS [17, 60, 115, 116].

Recently, an integrated scheduling scheme, referred to as PQGPS, which combines the fundamental PQ and GPS scheduling mechanisms, has attracted many research interests [66]. The PQ [59] scheduling mechanism is pervasively employed to guarantee high speed transmission for time-sensitive traffic flows while provide the best-effort service to non-critical ones. On the other hand, GPS, as an ideal traffic scheduling mechanism, has been extensively studied owing to its capability of handling the traffic flows which expect to receive differentiated services [66, 118].

Particularly, the service capacities received by individual traffic flows scheduled by GPS are controllable by appropriately configuring the weights assigned to them. Compared to the fundamental scheduling mechanisms (e.g., PQ or GPS), the hybrid PQGPS scheme is promising owing to its ability of providing the prioritized service to real-time applications as well as offering the fairness among other applications.

Inspired by the DiffServ and traditional scheduling mechanisms, the hybrid PQGPS scheduling scheme incorporates the advantages of the PQ and GPS scheduling mechanisms. This chapter focuses on examining the performance of the hybrid system subject to self-similar traffic.

To analyze the performance of each traffic flow scheduled by the complex PQGPS scheduling mechanism, a flow-decomposition method is proposed, which is able to isolate each traffic flow from the original hybrid PQGPS scheduled system. Specifically, the original system is first decomposed into an SSSQ system and a GPS system. Further, a bounding approach is adopted to divide the GPS system into a collection of SSSQ systems. Finally, the performance of each traffic flow of the PQGPS system can be obtained by examining the corresponding SSSQ system.

The rest of this chapter is organized as follows. Section 5.2 presents the hybrid PQGPS system and parameterizes its inputs. In Section 5.3, the decomposition approach for converting the PQGPS system into SSSQ systems is demonstrated. To obtain the performance of each flow handled by GPS, a bounding approach will be introduced. Section 5.4 compares the analytical and simulation results under various scenarios to validate the accuracy of the model. This chapter is concluded in Section 5.5.

5.2 Preliminaries

This section will introduce the hybrid PQGPS system and parameterized the input processes of each traffic flow. Related techniques of examining the queue length of an SSSQ system subject to an fBm process are given because the hybrid PQGPS system is converted into SSSQ systems eventually.

5.2.1 System Description

As shown in Figure 5.1, the hybrid system is composed of a unique server and three traffic flows. The traffic flow which has stringent QoS requirements is denoted by $A_1(t)$. To meet the need of the stringent QoS requirements, $A_1(t)$ is served with strict high priority. On the other hand, $A_2(t)$ and $A_3(t)$ are handled by the GPS scheduling mechanism in the low priority part of the PQGPS system.

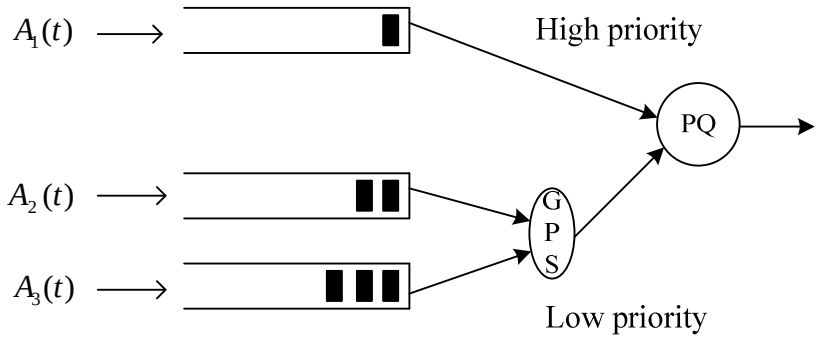


Figure 5.1: The PQGPS system.

In this chapter, the overall service capacity of the hybrid system is set to be c, and the effective service of the GPS system, which is the residual service left by the high priority part is denoted as c_{gps}. The traffic flows fed to the hybrid system

are modelled by fractional Brownian motion (fBm) processes which can effectively capture the self-similar nature [9].

5.2.2 Traffic Parameterization

The cumulative arrival processes, denoted as $A_i(t), i = 1,2,3$, can be formulized as follows:

$$A_i(t) = m_i t + Z_i(t), \tag{5.1}$$

where m_i is the mean arrival rate of the corresponding traffic flow and $Z_i(t) = \sqrt{a_i m_i}\,\overline{Z}(t)$. a_i is the variance coefficient of the fBm process. It is defined by $a = \mathbf{Var}(A(t))/\mathrm{E}(A(t))$ (i.e. the ratio of the variance and the mean). $\overline{Z}(t)$ is a centred (i.e., $\mathrm{E}\overline{Z}(t) = 0$) fBm with a variance function as,

$$\bar{v}_i(t) = \mathrm{Var}\,\overline{Z}(t) = t^{2H_i}, \tag{5.2}$$

where $H_i, i = 1,2,3$ denotes the Hurst parameter of the corresponding traffic flows and a covariance function is given as

$$\overline{Cov}_i(t_1, t_2) = \frac{1}{2}\left(\bar{v}_i(t_1) + \bar{v}_i(t_2) - \bar{v}_i(t_1 - t_2)\right) = \frac{1}{2}\left(t_1^{2H_i} + t_2^{2H_i} - (t_1 - t_2)^{2H_i}\right). \tag{5.3}$$

The variance function of each traffic flow can then be given by

$$v_i(t) = a_i m_i \bar{v}_i(t) = a_i m_i t^{2H_i}. \tag{5.4}$$

Its covariance function can be readily obtained as

$$Cov_i(t_1, t_2) = \frac{1}{2} a_i m_i \left(t_1^{2H_i} + t_2^{2H_i} - (t_1 - t_2)^{2H_i}\right). \tag{5.5}$$

5.2.3 Queue Length Distribution

This section introduces the approach to obtain the queue length distribution of an SSSQ subject to self-similar traffic. This is because the hybrid PQGPS system is eventually divided into a collection of SSSQ systems. In this study, let $A(t)$ denote the input of a general SSSQ system with a constant service capacity C. The total service received by the traffic flow during time interval (s,t) is $C(t-s)$, and the amount of traffic arrives within the same period can be denoted by $A(s,t)$. Consequently, the backlog of this queueing system at time instance t can be described as

$$Q(s,t) = \sup_{s \le t} \{A(s,t) - C(t-s)\}. \tag{5.6}$$

Mannersalo and Norros [119] have proposed an approach to model the bounds of queue length distribution based on Large Derivation Principle [120]. Based on their findings, the upper and lower bounds of queue length distribution are given by

$$P(Q > x) \le e^{-Y(\hat{t})/2}, \tag{5.7}$$

$$P(Q > x) \ge \Phi\left(\left|\sqrt{Y(\hat{t})}\right|\right) = \frac{1}{\sqrt{2\pi}} e^{-Y(\hat{t})/2}. \tag{5.8}$$

$\Phi(\cdot)$ is the residual distribution function of the standard Gaussian distribution [119, 121]. $\hat{t}$ minimizes $Y(t)$

$$Y(t) = \frac{(-x + (C-m)t)^2}{Cov(t,t)}, \tag{5.9}$$

where x is the queue length.

Through extensive studies, the geometric mean of the upper and lower bounds has been proven to be an effective and accuracy approximation to the queue length distribution [64], and can be given by

$$P(Q > x) \approx \frac{1}{\sqrt[4]{2\pi\left(1 + Y(\hat{t})\right)^2}} e^{-Y(\hat{t})/2}. \tag{5.10}$$

5.3 Analytical Modelling of Hybrid PQGPS Scheduling System

In general, modelling the hybrid PQGPS scheduling system is considerably complex due to the interdependent relationships among its traffic flows. In this section, the decomposition approach which can divide the hybrid PQGPS system into a group of isolated SSSQ systems is presented. Specifically, the hybrid system is first divided into an SSSQ system subject to $A_1(t)$ and a GPS system subject to $A_2(t)$ and $A_3(t)$ at PQ level. Afterwards, the GPS system is further decomposed in to two SSSQ systems. It is very challenging to directly isolate the traffic flows scheduled by GPS. Therefore, this section presents the method of obtaining the analytical upper and lower bounds of each GPS buffer by adopting a decomposition approach and a modulation approximation, respectively. Finally, the performance of the original PQGPS system can be evaluated by examining the decomposed SSSQ systems.

5.3.1 Queue Decomposition at PQ level

According to Figure 5.1, the buffer subject to the arrival process $A_1(t)$, let's say queue 1, and the GPS system in the low priority part subject to arrival processes $A_2(t)$ and $A_3(t)$, let's say GPS session 1 and GPS session 2 respectively, are

handled by the PQ policy. At this level, the hybrid system can be readily decomposed into two parts. According to the EBA approach [122], queue 1 can be directly isolated from the original hybrid system as an SSSQ system with the service capacity c and the original input, since queue 1 has the strict high priority over the GPS system and the queueing effect of the low priority part on the high priority part of a PQ system is negligible. Consequently, the queue length distribution of queue 1 can be obtained by solving a SSSQ system. The approximation method is provided in Section 5.2.

In the hybrid PQGPS system, the GPS part only receives the residual service left by the high priority part. It is critical to model the effective service capacity of the low priority part. The queue length of a PQ system is almost exclusively composed of that of the low priority part, and hence the total queue length of the GPS system can be used to approximate that of the hybrid system. From Equations (5.7)-(5.10) in Section 5.2.3, it is easy to see that the queue length distribution is simply decided by the corresponding $Y(t)$. By extending Equation (5.9) the computation of multi-input case is given by $Y^n(t)$

$$Y^n(t) = \frac{\left(-x + (c - \sum_{i=1}^{n} m_i)t\right)^2}{\sum_{i=1}^{n} Cov_i(t,t)}. \tag{5.11}$$

According to EBA, the queue length of GPS part can be used to approximate that of the original system. Hence it is given

$$\frac{\left(-x + (c - \sum_{i=1}^{3} m_i)t\right)^2}{\sum_{i=1}^{3} Cov_i(t,t)} = \frac{\left(-x + (c_{gps} - \sum_{i=2}^{3} m_i)t\right)^2}{\sum_{i=2}^{3} Cov_i(t,t)}. \tag{5.12}$$

where c_{gps} is the effective service capacity of the GPS subsystem. By solving Equation (5.12), it can be given

$$c_{gps} = \sum_{i=2}^{3} m_i + \left(c - \sum_{i=1}^{3} m_i\right)\left(\frac{\sum_{i=2}^{3} a_i m_i}{\sum_{i=1}^{3} a_i m_i}\right), \tag{5.13}$$

Upon the first step decomposition at PQ level, the hybrid PQGPS system has been equivalently separated into an SSSQ system with service capacity c and a classical two-session GPS system with effective service capacity c_{gps}, respectively. The queue length distribution of the original high priority traffic flow can then be obtained by examining the decomposed SSSQ system. In the following subsection, the GPS subsystem is being further decomposed.

5.3.2 Queue Decomposition of GPS Scheduled System

With the effective service capacity, c_{gps} obtained by the decomposition at the PQ level, the GPS subsystem is isolated from the original hybrid system. Performance analysis on the GPS system in a stochastic setting is considerably hard because the service received by each GPS session depends on both the traffic arrival rate and queue length of the other GPS session. For this reason, it is hardly to find any accurate method for deriving the excess service received by each GPS session. As a consequence, a bounding approach is employed to circumvent this difficulty. The bounding approach is widely recognized for eliminating the queueing effect of interacted traffic flows which shares a unique server [41, 123]. According to this approach, a two-session GPS system can be analyzed as two independent queueing systems by applying a decoupling method to each GPS session.

A. Lower bound approximation

Without loss of generality, GPS session 1 is used as an example to demonstrate the derivation of the lower bound of the queue length distribution by employing a modulation approach. Its key idea is to "smooth" the arrival of GPS session 2 and hence enable GPS session 1 to receive more service than that in the original hybrid system. There are two representative scenarios for the designated GPS system listed in this section:

Scenario 1 $m_2 < \mu_1 c_{gps}, m_3 < \mu_2 c_{gps}$;

Scenario 2 $m_2 < \mu_1 c_{gps}, m_3 > \mu_2 c_{gps}$, or $m_2 > \mu_1 c_{gps}, m_3 < \mu_2 c_{gps}$.

The first scenario is as an unbiased one while the second scenario is a biased one. From the definition of scenario 1, it is ready to see that the mean arrival rates of the GPS sessions (i.e., GPS scheduled traffic flows) do not exceed the corresponding guaranteed service capacity.

The method is carried out at two steps. At the first step, the mean arrival rate of GPS session 2 is modulated in order that the corresponding queue is empty when it only receives the guaranteed service capacity $\mu_2 c_{gps}$. The mean arrival rate of the modulated process is denoted as m_3'. Then the modulated traffic is fed to replace the corresponding original traffic flow of GPS session 2. In the modulated system, one can easily find that the effective service capacity received by GPS session 1, c'_{e_gps1} is greater than that of the original system. Consequently, the queue length of GPS session 1 which is calculated with c'_{e_gps1} is a lower bound. Since the modulated service rate of GPS session cannot contribute to the backlog with its guaranteed service capacity, the buffer of GPS session 2 in the original system is empty. Because the effective service capacity received by GPS session $j, j = 1,2$ satisfies

$c_{e\_gpsj} > \mu_j c_{gps}$. The new system with the modulated arrival rate of session 2 is shown in Figure 5.2.

At the second step, the modulated system can be converted to an SSSQ system by superposing the two traffic flows. Since the buffer of modulated GPS session 2 is empty, the total queue length of the modulated system is exclusively composed of that of GPS session 1. Hence, the obtained queue length distribution is a lower bound approximation.

The modulated arrival rate of GPS session 2, m_3', which could achieve the desired queue length probability can be obtained by using Equations (5.9) and (5.10). Fan and Georganas [113] suggested that the superposition of independent self-similar processes is still a self-similar process with the mean rate of the sum of those of the individual processes, the variance of the sum of these processes and the largest Hurst parameter among those of the processes. Subsequently, the lower bound can be obtained by solving an SSSQ system. The lower bound of GPS session 2 can be calculated by exchanging the roles of GPS session 1 and 2 above.

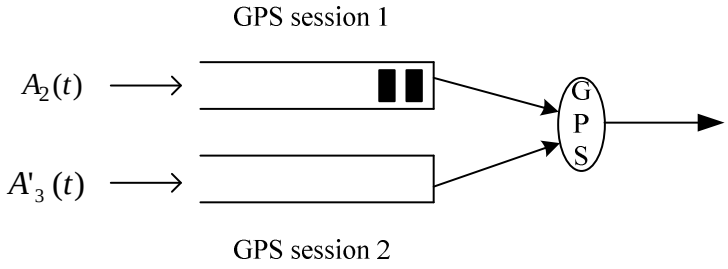


Figure 5.2: The modulated GPS system.

In the biased scenario, $m_2 + m_3 < c_{gps}$ is required for a stable system. Without loss of generality, let $m_2 < \mu_1 c_{gps}, m_3 > \mu_2 c_{gps}$. In this situation, the queue length of GPS session 2 will typically not shrink except that GPS session 1 is empty. As a result, GPS session 1 cannot receive excess service capacity GPS session2 and is served only with its own guaranteed service capacity. Hence, the approximation of queue length distribution of GPS session 1 can be given by solving an SSSQ system with $\mu_1 c_{gps}$ subject to the original arrival rate $A_2(t)$. The lower bound of queue length distribution of GPS session 2 can be obtained by follow the procedure demonstrated for the unbiased scenario. For $m_2 > \mu_1 c_{gps}, m_3 < \mu_2 c_{gps}$, it is can be obtained by replacing the corresponding parameters.

B. Upper bound approximation

This section provides an upper bound approximation by employing a priority conversion approach which converts the GPS system into a PQ system. By doing this, a fair queueing system becomes an unfair queueing system. Hence the upper bound of one GPS session can be derived by assigning this session with a low priority in the PQ system. The session in the low priority part definitely receives less service capacity than that in the original GPS system, since it only receives the residual service capacity of the PQ system. The residual service capacity of the PQ system can be obtained using Equation (5.13). Here, the larger one of the derived residual service capacity and its corresponding guaranteed service capacity is chosen as the effective service capacity for obtaining the upper bound under the biased scenario. The detail of decomposing a PQ system has been introduced in Section 5.3.1. By replacing the appropriate parameters, the queue length distribution of the

low priority part of the converted system can be obtained, which is actually the upper bound.

5.4 Model Validation

The appropriateness of an analytical model relies on its effectiveness and accuracy of analyzing and predicting the real system. This section compares the analytical and simulation results though extensive experiments under different settings. As aforementioned in Section 5.3, two settings of the system, unbiased and biased, are given as

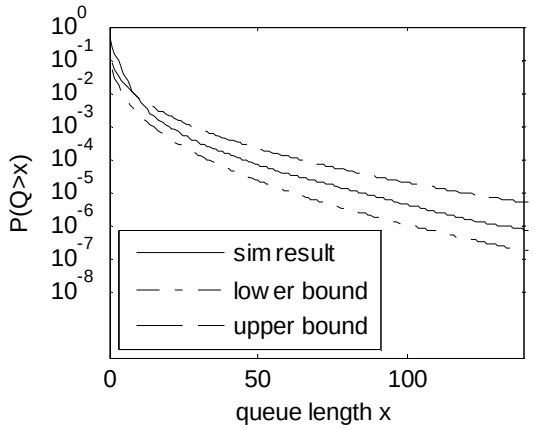
Unbiased setting:

$$c = 230, \mu_1 = 5/9, \mu_2 = 4/9, m_1 = 48, a_1 = 1, H_1 = 0.8,$$
$$m_2 = 85, a_2 = 1, H_2 = 0.8, m_3 = 70, a_3 = 1, H_3 = 0.8$$

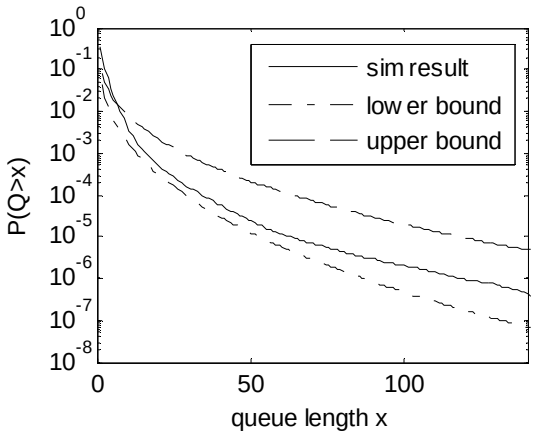
Biased setting:

$$c = 230, \mu_1 = 6/9, \mu_2 = 3/9, m_1 = 48, a_1 = 1, H_1 = 0.8,$$
$$m_2 = 100, a_2 = 1, H_2 = 0.8, m_3 = 70, a_3 = 1, H_3 = 0.8$$

For the reason that the high priority queue is served with the overall service capacity of the hybrid system and the traffic load is low, the priority queue is almost empty. Hence, the comparison between the analytical and simulation results of the high priority queue is not demonstrated.



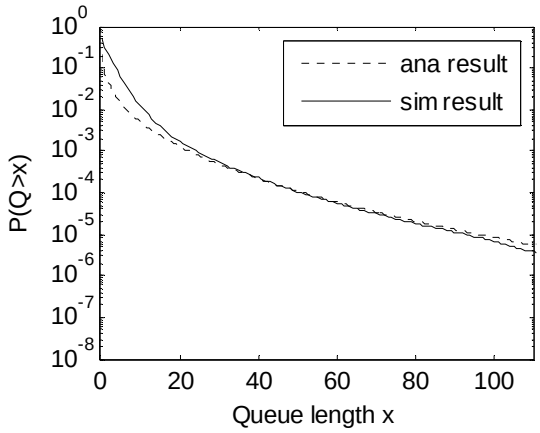
(a) GPS session 1



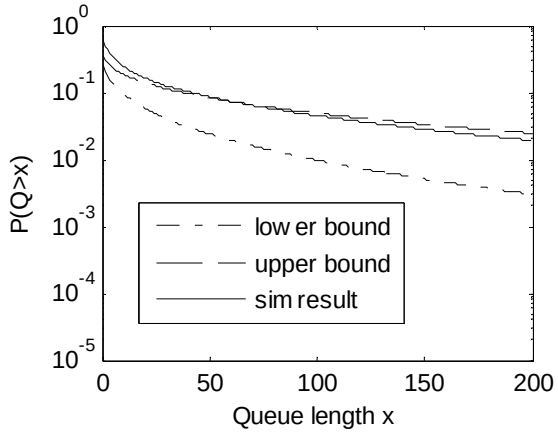
(b) GPS session 2

Figure 5.3: Bounding approach under the unbiased scenario

Under the unbiased setting scenario, the mean arrival rates of the GPS sessions do not exceed the corresponding guaranteed service capacity. From Figure 5.3, it is not hard to see that the simulation results are asymptotically located within the upper and lower bounds. The gap between the upper bound and the simulation result is very small for both GPS sessions. The lower bound can be varying by the precision of Equation (10). In our experiments, let $P(Q > x) < 0.05$.



(a) GPS session 1



(b) GPS session 2

Figure 5.4: Bounding approach under the biased scenario.

In the biased setting scenario, the mean arrival rate of one GPS session exceeds its guaranteed service capacity. For the reason of the stability of the system, $m_2 + m_3 < c_{gps}$ is required. Since GPS session 1 can hardly receive the excess service capacity of the other session, it is only served with the guaranteed service

capacity. Hence, the performance metrics can be well estimated as depicted in Figure 5.4(a) On the other hand, GPS session 2 cannot clear the backlog with its guaranteed service capacity. Thus, Figure 5.4 shows that the queue length distribution of GPS session 2 is much larger than that of GPS session 1. In addition, Figure 5.4(b) shows that the upper bound of GPS session 2 is very tight.

5.5 Summary

With the rapid development of network based applications, the provisioning of differentiated QoS has become a pressing demand. In addition, traffic self-similarity has been proven to be a ubiquitous phenomenon in nearly all kinds of communication networks, and it has attracted many research efforts. To meet the needs of differentiated QoS requirements, a promising hybrid PQGPS scheduling mechanism has merged, which integrates the advantages of PQ and GPS scheduling schemes.

In this chapter, an analytical model has been developed to derive the queue length distribution of individual traffic flows scheduled by the hybrid PQGPS scheduling. In order to isolate each traffic flow from the complex PQGPS system, the original system is first divided into an SSSQ system and a GPS system. Further, a modulation approach and a priority conversion approach are proposed to obtain the analytical bounds of the queue length distribution. By comparing the analytical and simulation results, the developed model has been validated under two representative scenarios. From the validation, it can be concluded that the developed model can be used as an efficient tool to estimate and predict the queue length distribution of the PQGPS system which can meet the QoS requirements of individual traffic flows.

Chapter 6

Performance Analysis of a Multi Buffer ARQ Systems under Prioritized Self-similar Traffic

6.1 Introduction

In practical networks, various unpredictable errors occur during data transmission. Packet loss which is caused by unforeseen circumstances significantly degrades the performance and reliability of communication systems [124]. For this reason, the Automatic Repeat reQuest (ARQ) based error control strategy has been proposed and widely deployed to retransmit damaged or lost packets in wireless communication networks owing to its reliability [68, 71, 75]. SR-ARQ has been proposed and regarded as the theoretically most efficient forwarding error control strategy, because it retransmits only the negatively acknowledged packets. Therefore, it can achieve the larger throughput than the pure stop-and-wait ARQ strategy (refer to Section 2.2.6).

In realistic communication networks, packets may be discarded during transmission, even if the ARQ strategy has been implemented. Loss analysis of real-time traffic has emerged as an impressing issue and drawn many research efforts [72, 125]. In general, the actual delay experienced by a packet which transmits over wireless communication links is composed of two parts. The first part, called

queueing delay, is the sojourn time of the packet waiting for service in the queue. In other words, it is the time interval between the arrival of the packet and its transmission over the wireless link. The second part, namely transmission delay, is the time interval between the beginning of the first transmission and the instant of the successful transmission of one packet.

The service time of each packet with stringent QoS requirements is relatively small as compared to the corresponding allowed delay bound. This is to guarantee that there is enough time for multiple transmissions and feedback before the deadline of the packet transmission. This chapter develops an analytical model of a multi-buffer ARQ system subject to prioritized self-similar traffic for the provisioning of both differentiated QoS [1, 4, 58] and reliable data transmission by employing the ARQ error control strategy. Specifically, the developed analytical model focuses on the loss probabilities of individual buffers of the multi-buffer ARQ system. Each buffer is isolated from the complex system through a decomposition approach. Consequently, the original system is divided into a group of SSSQ systems, which are statistically equivalent to the performance of the corresponding queues in the original system. Therefore, the loss probabilities of individual queues of the multi-buffer ARQ system can be obtained by examining the resulting SSSQ systems. The validity and accuracy of the obtained loss probabilities have been demonstrated through extensive comparison between analytical and simulation results under different working conditions. Finally, the developed model is further adopted to investigate the effects of the service capacity and delay bound of multi-buffer ARQ on system performance.

The remainder of this chapter is organized as follows. Section 6.2 describes the system model and preliminaries. In Section 6.3, a decomposition approach which is used to partition the original multi-buffer ARQ system and obtain effective service capacity of the resulting subsystems is presented. Further, the queueing and transmission loss probabilities of individual decomposed subsystems can be derived. The comparison between the analytical and simulation results under various settings, which validates the accuracy of the developed model, will be given in Section 6.4. Section 6.5 demonstrates its application to design and implement prioritized service systems with ARQ error control effectively and financially. Finally, the chapter is summarised in Section 6.6.

6.2 System Description

This section presents the multi-buffer ARQ system. Figure 6.1 shows a schematic diagram of such a system. It is clear that the integrated system is composed of an arrival buffer and an ARQ buffer to store the original arrived packets and retransmission packets, respectively. The two subsystems are handled using the PQ scheduling mechanism, because practical switching systems are commonly expected to provide service of two different priority levels to traffic flows, namely, high priority and low priority. Without loss of generality, the two subsystems in Figure 6.1 are denoted as Subsystem 1, which consists of ARQ buffer 1 and Arrival buffer 1, and Subsystem2, which is composed of ARQ buffer 2 and Arrival buffer 2 in this chapter, respectively. The strict high priority is set to Subsystem 1. Within each subsystem, the ARQ buffer is served with high priority over the corresponding arrival buffer. By doing this, the packets in ARQ buffers are not likely to bear queueing delay.

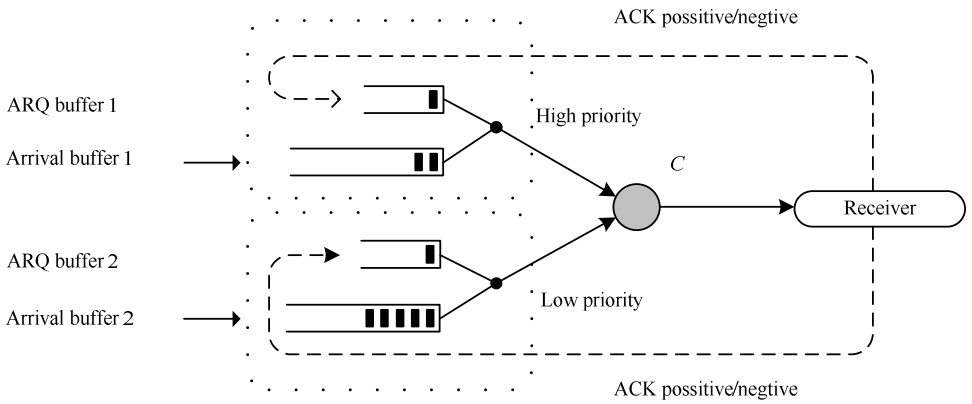


Figure 6.1: **The multi-buffer ARQ system.**

The subsystem only retransmits the packets with a negative ACK, if an error or damage occurs to a transmitting packet. According to the SR-ARQ strategy, all the packets which have positive ACKs which denote that the designated packets have been successfully received by the destination have no impact on the developed system. Thus, the traffic flow of the corresponding ARQ buffer is only composed of the retransmitted packets. For the modelling purpose, the overall service capacity of the wireless link is denoted as C, as shown in Figure 6.1. The arrival processes of individual subsystems, i.e., Subsystem 1 and Subsystem 2, are formulized as $A_i(t), i = 1,2$, which are fBm processes defined by

$$A_i(t) = m_i t + \sqrt{a_i m_i}\,\overline{Z}(t), \tag{6.1}$$

where m_i is the mean arrival rate of the corresponding process and a_i is the variance coefficient. The variance of the designated process is

$$\overline{v}_i(t) = Var\ \overline{Z}(t) = t^{2H_i}, \tag{6.2}$$

where H_i is the Hurst parameter of the related process, and the corresponding covariance it given by

$$\overline{Cov}(t_1,t_2)=\frac{1}{2}\left(\bar{v}(t_1)+\bar{v}(t_2)-\bar{v}(t_1-t_2)\right)=\frac{1}{2}\left(t_1^{2H}+t_2^{2H}-(t_1-t_2)^{2H}\right). \quad (6.3)$$

Hence, the variance function of $A_i(t)$ can be described by

$$v_i(t)=a_i m_i \bar{v}_i(t)=a_i m_i t^{2H_i}. \quad (6.4)$$

Then the covariance function of the corresponding fBm process can be given by

$$Cov(t_1,t_2)=\frac{1}{2}am\left(t_1^{2H}+t_2^{2H}-(t_1-t_2)^{2H}\right). \quad (6.5)$$

Let $R_i(t), i=1,2$ characterize the cumulative retransmitted traffic flow of Subsystem i up to time t. To obtain a stable system, $\sum_i A_i(t)+\sum_i R_i(t)<Ct$ is required. In this chapter, the failure probability of an arbitrary packet on its ith transmission is denoted by FP_i. In addition, the channel conditions are considered constant. It has been shown that $FP_i, \forall i$ are constant, if without packet integration, and hence denoted by p [126].

It is worth noting that the traffic flows feed to the ARQ buffers are split from the corresponding original arrival processes. From Figure 6.2, for each subsystem, In a large scale, the original process has been split into two, retransmission process fed to ARQ buffer with probability FP_i and successfully transmitted one without retransmission with probability $1-p$. To cope with the split self-similar process, Fan and Georganas [113] have studied the merging and splitting of self-similar traffic flows by the independent splitting operation. More importantly, they verified that the process which is merged by self-similar processes is still self-similar. On the other hand, the splitting processes still exhibit self-similar nature. Specifically, the mean arrival rate of the merged process is the sum of the original processes, and the

Hurst parameter is equal to the largest one of the original processes. For the process splitting, the Hurst parameters of the split processes are identical and its value is equal to that of the original process. The mean arrival rate of each split process is subject to the independent split rate, which is p in this chapter. Merging and splitting of the self-similar processes enable us to investigate the traffic flows which are fed to the ARQ buffers.

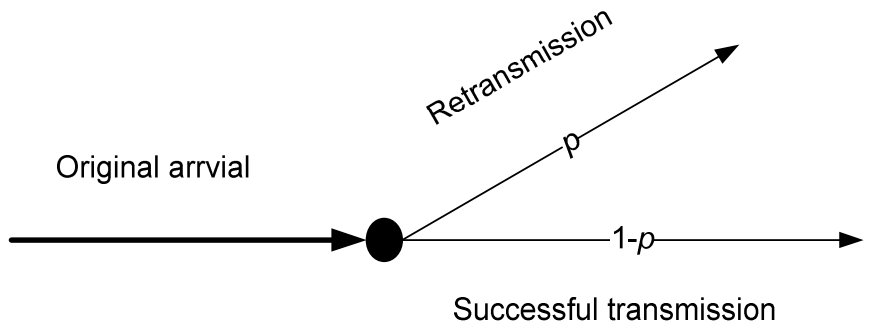


Figure 6.2: Splitting process subject to probability p

The developed multi-buffer ARQ system is fed with self-similar traffic flows and packets are discarded due to the violation of the delay bound D which characterizes the stringent QoS requirement of real time traffic. In this chapter, the arrival buffer is infinite or sufficiently large in order to guarantee that no packet is lost due to buffer overflow.

6.3 Loss Probability of Individual Buffers

This section demonstrates the key of multi-buffer ARQ modelling. The analysis of loss probabilities of individual buffers is of much concern. The loss of

packets is only caused by the violation of aforementioned delay bound D. In this part, the delay experienced by a packet is composed of the queueing delay and transmission delay before its successful transmission. The packets will be discarded if their actual delays exceed their delay bounds. By observing Figure 6.1, it is ready to see that the packets in arrival buffers will be only dropped if their sojourn time which is referred to the aforementioned queueing loss exceeds the initialized delay bounds, while those packets in ARQ buffers depend on both of their sojourn time and the number of retransmissions which is named transmission loss.

6.3.1 Queueing Loss

If there is an error free channel, queueing loss occurs only if the packet cannot be sent before the allowed delay bound. In other words, the total delay that a packet experiences consists only of its waiting time in buffer, which is mainly determined by the length of the corresponding queue it sees. Hence the queue length of a buffer within time interval (s,t) can be denoted as [9]

$$Q(s,t) = \sup_{s \le t} \{A(s,t) - c_e(t-s)\}, \tag{6.6}$$

where c_e denotes the effective service capacity received by the relevant buffer. The term effective service capacity denotes the capacity that is actually received by the corresponding queue. In other words, the decomposition approach seeks a service capacity with which the queueing performance of the corresponding queue is statistically equivalent to that of the original system. For ease of the demonstration, let c_e be known for the following equations. The derivation and calculation of c_e will be presented in Section 6.3.3.

Next, the mathematical expression of the packet loss due to the violation of the delay bound is given by

$$P(d > D) = \sup_{t} P(A(t) - c_e t > D c_e), \tag{6.7}$$

where d is the sojourn time of an arbitrary packet. There is not a practical approach which can be employed to obtain the result of Equation (6.7) directly. Instead, a bounding approach is employed to obtain the loss probability. The upper and lower bounds of the loss probability can be given as [61]

$$P(d > D) \le e^{-Y(\hat{t})/2}, \tag{6.8}$$

$$P(d > D) \ge \Phi\left(\left|\sqrt{Y(\hat{t})}\right|\right), \tag{6.9}$$

where $\Phi(\cdot)$ is the residual distribution function of the standard normal distribution. There exists a time instant $\hat{t}$ which maximizes function $Y(t)$ which is given by

$$Y(t) = \frac{\left(\int_0^{t+D} c_e(\tau) d\tau - mt\right)^2}{Cov(t,t)}. \tag{6.10}$$

In the open literature [61], the geometric mean of the upper and lower bound as the approximation has been proven effective. Therefore, the approximation of the queueing loss probability of an SSSQ system is given by

$$P(d > D) \approx \frac{1}{\sqrt[4]{2\pi\left(1 + Y(\hat{t})\right)^2}} e^{-Y(\hat{t})/2}. \tag{6.11}$$

6.3.2 Transmission Loss

Transmission loss is referred to the loss caused by the time delay which is the sum of the transmission and queueing delay of a packet exceeds the corresponding delay bound. As shown in Figure 6.1, the packets in the ARQ buffer have the strict high priority over those in the related arrival buffer in each subsystem. It guarantees that these packets have more opportunities to be retransmitted and hence reduces

transmission loss. In each subsystem, the ARQ buffer and arrival buffer are handled by PQ scheduling mechanism. According to the EBA, the backlog of a PQ system is almost exclusively composed of those of the low priority part [127]. Therefore, it can be reasonably assumed that the ARQ buffer is almost surely empty. Hence, the sojourn time of a retransmitted packet in an ARQ buffer is negligible.

The most significant impact on the number of packet retransmission is the elapsed sojourn time experienced by packets since they arrive in the system. The maximum number of retransmissions of packets depends only on the total remaining time until packets expired. Since the packets in ARQ buffers do not bear any queueing delay, the maximum number of retransmissions allowed for a packet is $n = \lfloor (D-d)c_e \rfloor$. The distribution of the sojourn time of an arbitrary packet in the arrival buffer can be calculated by

$$P(d > t) \approx \frac{1}{\sqrt[4]{2\pi\left(1+Y(\hat{t}+t)\right)^2}} e^{-Y(\hat{t}+t)/2}, \tag{6.12}$$

According to Equation (6.12), the probability density function of the number of retransmissions can be presented as

$$P(N=n) = \begin{cases} P(d > D-(n+1)/c_e) - P(d > D-n/c_e), & \textbf{if } 0 \le n < \hat{n} \\ P(d>0) - P(d > D-\hat{n}/c_e), & \textbf{if } n = \hat{n} \end{cases}, \tag{6.13}$$

where $\hat{n} = \lfloor Dc_e \rfloor$ is the maximum number of transmissions that the delay bound allows at a packet arrival instance. The failure probability of exactly n transmissions is p^n. Consequently, the approximation of estimating the transmission loss is given as follows

$$P_{trans} = \sum_{i=0}^{\hat{n}} p^{i+1} P(N=i), \tag{6.14}$$

6.3.3 Service Capacity Decomposition

The aforementioned approaches of obtaining both queueing and transmission probabilities are subject to SSSQ systems. Hence there is an issue that all queues in the multi-buffer ARQ system which is presented in Figure 6.1 share the unique server. The service allocated to each queue is interacted. To isolate every queue from the original complex system, a decomposition approach is further applied to obtain the effective service capacities for every single queue. By doing this, the interaction of individual service can be eliminated, and hence the complex system has been divided into a group of SSSQ systems. In the following, the decomposition approach is presented in two steps; at the first place, two subsystems are isolated from the original system; then, each subsystem is further decomposed into SSSQ systems.

Following the PQ scheduling mechanism, the overall service capacity of the original system can be readily divided into effective service capacities of each subsystem by employing EBA [127] approach. According to EBA, the queueing effect of the low priority part on the high priority part is negligible, since the total queue length is almost composed of that of the low priority part. Therefore, the effective service capacity of the high priority subsystem can be readily give by $c_{eH} = c$. Moreover, the total queue length distribution of a PQ system can be used to approximate that of its low priority part. The total queue length distribution of a PQ system is determined by the following factor [114],

$$U(k_x) = \frac{((c - m_{A_1(t)} - m_{R_1(t)} - m_{A_2(t)} - m_{R_2(t)})k_x - x)^2}{\sum_{i=1}^{2} Cov_{A_i(t)}(k_x, k_x) + Cov_{R_i(t)}(k_x, k_x)}, \qquad (6.5)$$

where $k_x = \arg\min_k U(k)$ i.e., $U(k)$ attains its minimum when $k = k_x$; In the same way, the queue length of the low priority subsystem $P_{low}(X > x)$ is determined by the following equation

$$V(k'_x) = \frac{((c_{eL} - m_{A_2(t)} - m_{R_2(t)})k'_x - x)^2}{Cov_{A_2(t)}(k'_x, k'_x) + Cov_{R_2(t)}(k'_x, k'_x)}, \tag{5.16}$$

where $k'_x = \arg\min_{k'} V(k')$. Actually, the inputs of the ARQ buffers are the superposition of multiple retransmitted processes, which are split from the corresponding original traffic flows according to the error rate and distinct from each other by the number of the retransmissions experienced by their packets. In support of splitting a self-similar traffic, Fan and Georganas [113] have proven that a flow split from self-similar traffic by independent splitting operations is still self-similar and its Hurst parameter is the same as that of the original traffic. Upon this fact, it is obvious that all the retransmitted processes have the same Hurst parameter as the input traffic, since they are all split from it. Moreover, Fan and Georganas have showed that the merging of self-similar processes with the same Hurst parameter is still a self-similar process with this Hurst parameter. Therefore, the inputs of the ARQ buffers can be estimated by superposing the corresponding retransmitted processes. The mean arrival rate of an entire retransmitted process, i.e., the one input into an ARQ buffer, can be given by

$$m_{R(t)} = m_{A(t)} p \frac{1 - p^{\hat{n}}}{1 - p}, \tag{6.17}$$

where p and $\hat{n}$ have been defined in Section 6.3.2 as the failure probability and the maximum number of transmission that the delay bound allows respectively. The covariance function $Cov_{R(t)}(t,t)$ can be derived based on Equation (6.16) as

$$Cov_{R(t)}(t,t) = p\frac{1-p^{\hat{n}}}{1-p}Cov_{A(t)}(t,t). \tag{6.18}$$

Consequently, the EBA method can be applied to obtain the effective service capacity of the low priority subsystem, c_{eL}, by combining Equations (15) and (16), i.e., $U(k_x) = V(k_x')$:

So far, the effective service capacities for both subsystems have been obtained. In what follows, the effective service capacities for each buffer are further derived. As expatiated in Figure 6.1, in each subsystem, the packets in the ARQ buffer are served with strict high priority over the corresponding arrival buffer. The effective service capacities of Arrival buffer 1 and ARQ buffer 1 can be derived by employ the approach which is used to decompose the overall capacity by replacing the appropriate parameters. According to the inherent characteristic of the PQ scheduling mechanism, the effective service capacity of ARQ buffer 1 can be given by $c_{eR_1(t)} = c_{eH}$. Afterwards, the effective service capacity of Arrival buffer 1, $c_{eA_1(t)}$, can be calculated by

$$\min_s\left(\frac{((c_{eH} - m_{A_1(t)} - m_{R_1(t)})s - x))^2}{Cov_{A_1(t)}(s,s) + Cov_{R_1(t)}(s,s)}\right) = \min_{s'}\left(\frac{((c_{eA_1(t)} - m_{A_1(t)})s' - x)^2}{Cov_{A_1(t)}(s',s')}\right). \tag{6.19}$$

Consequently, the queueing loss of a single subsystem can be obtained from Equations (6.8)-(6.11) by replacing the corresponding parameters with appropriate values, and the transmission loss can be estimated by Equations (6.11)-(6.14).

6.4 Model Validation

The simulator is programmed in C by employing the discrete event simulation algorithm. The arrival processes for both subsystems are generated by the same traffic generator which has been addressed in Chapter 4. This part will present 4 groups of parameter settings which are used to validate the developed model. The significance of an analytical model is essentially determined by its ability to accurately predict and measure the actual performance behaviour of the designated communication systems. The following examines the accuracy of the developed model in estimating both the queueing and transmission loss of each subsystem. Specifically, the developed model is validated via comparing the analytical and simulation results obtained from four representative cases. For each case, the geometric mean of the aforementioned upper and lower bounds of loss probabilities is plotted, which is presented in Section 6.3.1. The overall service capacity of the multi-buffer ARQ system be $c = 100$ packets/second The arrival traffic flows are generated by the widely adopted conditionalized random midpoint displacement algorithm [37, 38] owing its relatively low time complexity and the ability of producing real-time fBm traffic without prior knowledge of trace length. The detailed parameter settings of all cases are given in Table 6.1:

Table 6.1
The parameter settings of Cases 1, 2, 3, and 4.

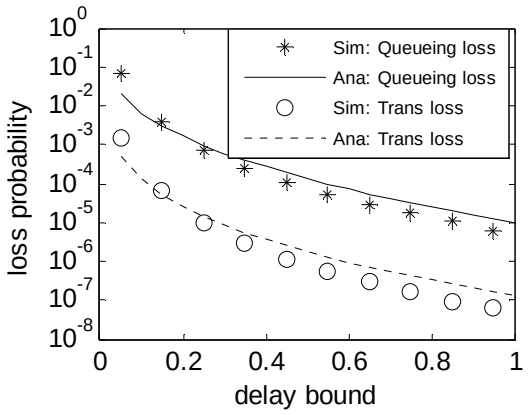
	$m_{A_1(t)}$	$m_{A_2(t)}$	a_1	$a2$	H	p
Case 1	75	10	1	1	0.8	0.1
Case 2	10	65	1	1	0.8	0.1
Case 3	75	10	1	1	0.7	0.1
Case 4	80	10	1	1	0.75	0.05

6.4.1 Case 1

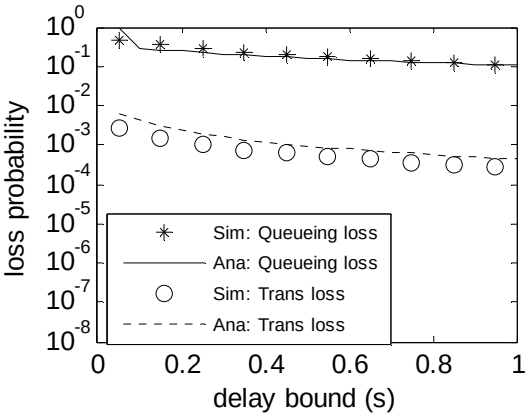
In the first case, the mean arrival rate of the fBm traffic of the high priority subsystem is considerably large, as compared to that of the low priority subsystem and the overall service capacity of the system. Therefore, the queueing loss will occur in the arrival buffer of the high priority subsystem in view of the fact that the sojourn time is not negligible.

In Figure 6.2(a), the loss probabilities of the high priority subsystem are plotted against delay bound. The error rate for each transmission is a constant, given as $FP_i = p = 0.1$. It is obvious that the asymptotic trends of both queueing and transmission loss are well approximated. The curves tend to gentleness with the increase of delay bound. Especially, the curves gradually become to be flat in Figure 6.2(b). That means only increasing delay bound cannot dramatically improve packet loss after a reasonable threshold of delay bound.

Figure 6.2(b) presents the simulation and analytical results of the queueing and transmission loss of the low priority subsystem. It is not hard to observe that the queueing and transmission loss predicted by the analytical model are very close to the simulation results. Because the low priority subsystem only receives the residual service capacity, the sojourn time of each packet in the arrival buffer of this subsystem can be very large. As a consequence, the curve of the queueing loss is close to 1 and then becomes smooth as delay bound increases.



(a) The high priority subsystem



(b) The low priority subsystem

Figure 6.3: The simulation and analytical results of loss probabilities in Case 1.

6.4.2 Case 2

For the second group of parameter settings, the mean arrival rate of the high priority subsystem traffic is much smaller than the overall capacity. Meanwhile, as the high priority subsystem is served with the overall service capacity, its queueing time is negligible. Thus, nearly all the packets can be successfully transmitted within the delay bound. As compared to Case 1, the backlog of the high priority subsystem

in this case is almost empty in simulation. As a consequence, there will be no empirical loss probability curves of the high priority subsystem. From the perspective of the analytical model, the predicted curves do exist. However, the values are too small to be presented in Figure 6.3 due to the limit of the vertical axis. From Figure 6.3, it can be observed that the analytical and simulation result of the low priority subsystem are matched in a good degree.

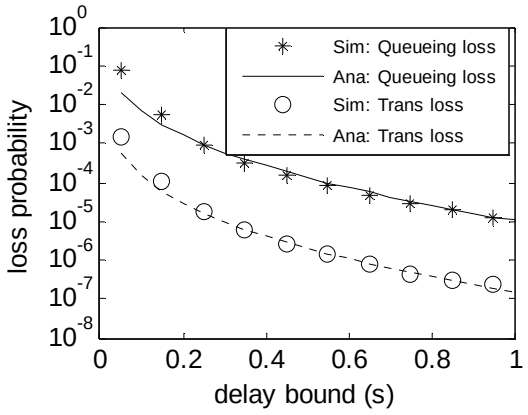


Figure 6.4: The simulation and analytical results of loss probabilities of the low priority subsystem in Case 2.

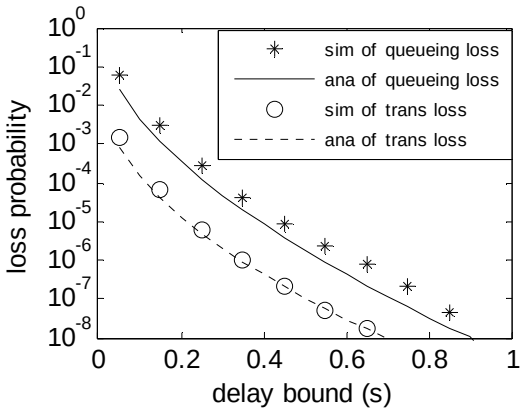
Comparing Cases 1 and 2, it is obvious that the loss probabilities of the low priority part in Case 2 are smaller than those of Case 1. This is caused by the input rate of the high priority part, which determined how much service capacity can be dispatched to the low priority part. From Figures 6.2 and 6.3, it can be noted that the analytical model for estimating the queueing and transmission loss of both subsystems is efficiency in the above two cases.

6.4.3 Case 3

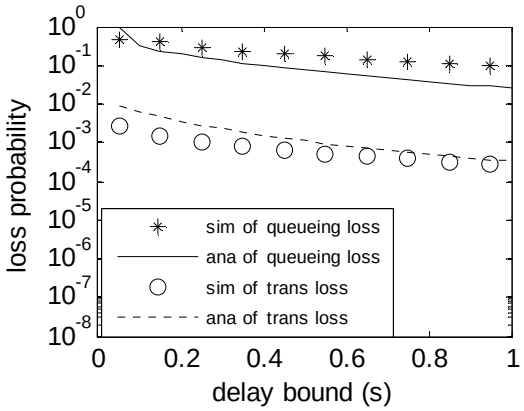
In this case, the Hurst parameter is altered for further examining the applicability of the developed model, because in this case the curves of both subsystems can be obtained.

It is clear that the aforementioned findings observed from Case 1 also hold in this case. Besides, a new phenomenon can be found in Figure 6.4(a). The arrival intensity of the system remains the same as Case 1. However, the queueing and transmission loss probabilities of the high priority subsystem decrease faster against the increasing delay bound than those of Case 1. This observation can be readily explained: The burstiness of self-similar traffic decreases as the Hurst parameter becomes small. As a consequence, it reduces the periods of large queue build ups and leads to small queues. Thus, the sojourn time in the arrival buffer decreases and more residual time is available for transmission. Finally, it results in lower loss probabilities than those of Case 1.

On the other hand, it is worth noting that the queueing and transmission loss of low priority does not change significantly in contrast with those of Case 1. This makes sense, because the low priority part only receives the residual service capacity of the entire system. The traffic intensity remains high in contrast with the effective service capacity of the low priority part. Because of this reason, the sojourn time of the packets in this subsystem changes insignificantly. This results in the slightly changed loss depicted in Figure 6.4(b).



(a) The high priority subsystem



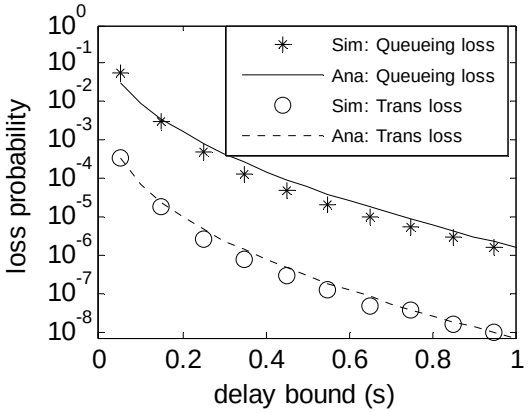
(b) The low priority subsystem

Figure 6.5: The comparison between the analytical and simulation results of loss probabilities in Case 3.

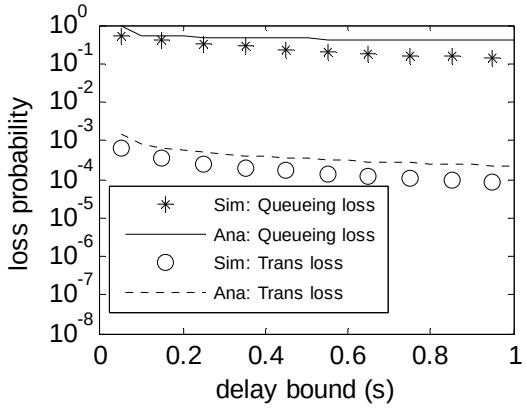
6.4.4 Case 4

It has been validated that the developed model is still efficient and accuracy with a different Hurst parameter. This case alters mean arrival rate, Hurst parameter and failure probability/error rate. The high priority subsystem still has a large mean arrival rate rather than a small one for the reason mentioned in Case 3. Figure 6.5 verified that the phenomena observed in the previous cases also exist in Case 4. From all of the above four cases, it can be observed that the analytical results are

very close to the simulation results. This observation suggests that the developed analytical model has a good degree of accuracy and applicability in estimating the queueing and transmission loss probabilities of the subsystems handled by the PQ scheduling. In addition, the findings show that the queueing loss of the high priority subsystem rises as the Hurst parameter increases, which implies the enhancement of the burstiness of self-similar traffic. Moreover, increasing the delay bound beyond a certain value does not significantly reduce the packet loss in all representative cases.



(a) The high priority subsystem



(b) The low priority subsystem

Figure 6.6: The comparison between the analytical and simulation results of loss probabilities in Case 4.

6.5 Applications of the Model

Applying a proper analytical model on the corresponding practical communication system is an effective and efficient way of investigating and measuring the performance of the designed systems. In the most recent decade, the provisioning of a cost-effective transmission channel is not only a technical issue, but also a financial consideration. In a practical system, packet loss significantly degrades its performance. Therefore, a cost-effective system may be built effectively by employing an appropriate analytical model to assess its potential performance. By deploying the developed model, packet loss of a multi-buffer ARQ system can be analyzed, which can aid DiffServ requirements.

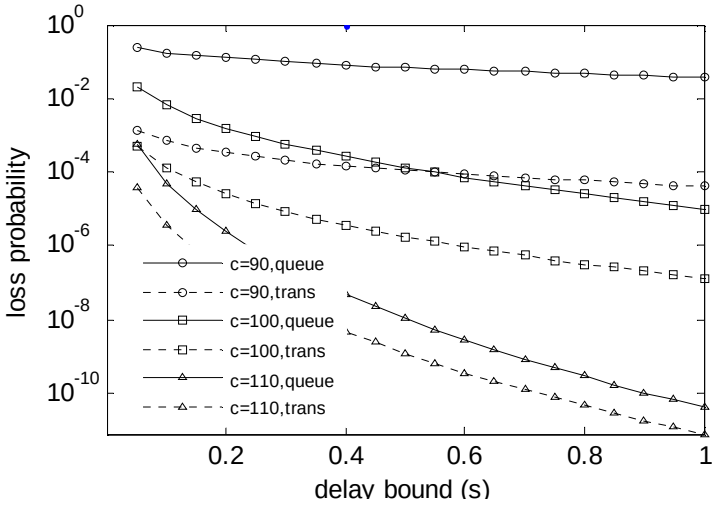


Figure 6.7: Queueing and transmission loss probabilities against delay bound in the cases with different service capacities.

A wireless communication link adopting SR-ARQ is expected to provide DiffServs to critical and non-critical traffic flows. In an extreme case, most of the traffic patterns are critical, which means that these traffic cannot tolerate too much delay, while the residual service capacity is provided to non-critical traffic. The

mean arrival rate of the critical traffic is 75, and that of the non critical traffic is 10. From Figure 6.7, it can be seen that the transmission and queueing loss of this system under various service capacities are considerably different from each other.

The delay bound of each packet is initially set to be 0.4 and the QoS requires that the total loss probability is less than 10^{-6}. It is obvious that service capacity $c = 110$ meets the requirement. Since increasing the service capacity causes the high cost, and increasing the delay bound may low down the loss probability but low cost, it is reasonable to increase the delay bound of the packets before enlarging the service capacity. The trend of increasing the delay bound against the loss probability can be examined according to Figure 6.8. However, it is obvious that the loss

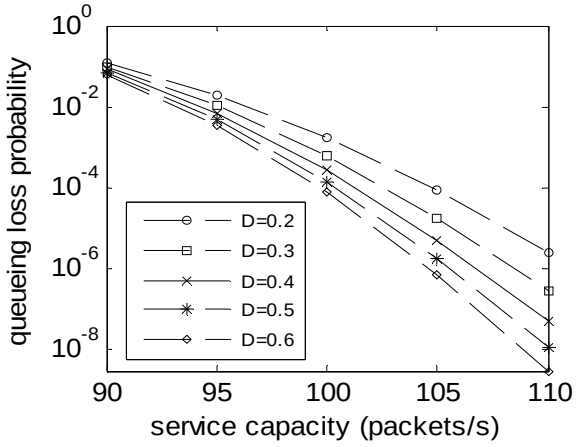


Figure 6.8: Queueing loss probabilities with various delay bounds against service capacity.

probabilities cannot be significantly reduced after delay bound is beyond a certain threshold. Figure 6.9 shows how much the queueing loss probability of the high priority part decreases against delay bound. It can be noted that the queueing loss almost keeps unchanged when the delay bound exceeds 0.4 for Cases 1, 2 and 4.

Usually, a trade-off between QoS and cost is mainly concerned in building a practical system. As compared to increasing service capacity, augmenting delay bound is considerably economical. A reasonable maximum value of the delay bound can be examined in advance, e.g. Figure 6.9. Then the service capacity can be adjusted to meet the QoS requirements with a considerably low cost.

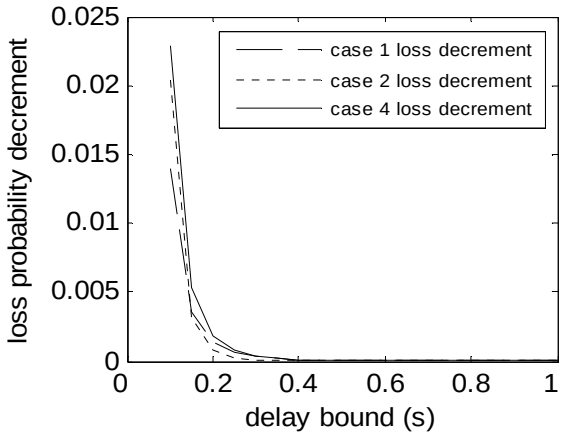


Figure 6.9: Loss probability decrement against delay bound of Cases 1, 2 and 4.

6.6 Summary

The provisioning of reliable data transmission in high-speed wireless communication channels has become an increasingly pressing demand because of the rapid development of wireless techniques. On the other hand, various modern network based applications require differentiated services due to their distinct QoS requirements. With the demands of providing DiffServ while guaranteeing the reliable data transmission, an analytical model of a multi-buffer ARQ system subject to prioritized self-similar traffic has been developed. PQ scheduling is used to distinguish traffic flows into different priority classes.

This chapter analyzes the loss probabilities of individual buffers of the multi-buffer ARQ system. Specifically, the effective service capacity of each subsystem is obtained by decomposing the overall link capacity by adopting the EBA method. Further, the effective service capacity of each subsystem and every buffer of both subsystems is derived. Then, each buffer has been isolated from the original complex system based on their individual effective service capacities. Finally, the desired loss probabilities can be readily obtained by examine the corresponding SSSQ systems.

Through comparison between the results obtained from simulation experiments and the analytical model, it can be observed that the developed model exhibits a good degree of accuracy in predicting the queueing and transmission loss of the multi-buffer ARQ systems. Afterwards, the analytical model is applied on an example to explain how it can predict and analyze an appropriate delay bound and a reasonable service capacity in implementing a cost-effective multi-buffer ARQ system.

Chapter 7

Performance Modelling of Dynamic Spectrum Access in Cognitive Radio Networks with Self-Similar Traffic

7.1 Introduction

The growing proliferation of wireless devices requires more spectrum usage. A plausible fact shows that the usable unlicensed spectrum is approaching its full capacity and will be exhausted before long. On the other hand, the authority of US Federal Communications Commission (FCC) [80] reveals a different story, that is, much of the licensed spectrum is almost idle in most circumstances. This contradiction implies the inefficient spectrum management policy in the presence of wireless communication networks [128].

For the above reasons, an urgent issue, namely, how to utilize the licensed spectrum bands when they are vacant, has been brought to our attention. To fulfil the demand, CR [12, 18, 129, 130, 131] technology has been proposed to mitigate the problems resulting from the inefficient usage of frequency spectrum. In a CR network, contending users are divided into two categories: primary users (also called licensed users) and secondary users (also called unlicensed users or CR users). Primary users should not be interfered by secondary users during their transmission.

On the contrary, secondary users must release their spectrum occupation when primary users are detected.

This chapter develops a performance model of a typical CR network subject to self-similar traffic, which employs dynamic spectrum access protocol. Dynamic spectrum access is different from all the aforementioned scheduling mechanisms on allocating the resource to contending users in the networks. Before the dynamic spectrum access is induced, the Medium Access Control (MAC) is introduced as preliminary. Theoretically, MAC [132, 133] is designed as an effective methodology to allow the contending users on a Wired/Wireless LAN to share their interconnecting resource. According to whether or not a coordinator is required for the protocol, MAC can be classified into two categories, namely, distributed MAC and centralized MAC [134]. Distributed MAC is also called decentralized MAC in some open literature [130, 135]. Scheduling mechanisms which were discussed in previous chapters are all centralized access methodology, since it is a necessity to have a coordinator which is able to regulate and decide the order of accessing the resource among contending users. However, this chapter develops an analytical model of a wireless communication system which employs distributed access control protocol for packet transmission.

This chapter begins with modelling the residual service which may be received and shared by secondary users in a CR network. It is important because residual service left by primary users enable service providers to measure the service which can be assigned to potential secondary users. Existing studies commonly assume the residual service to be ON-OFF processes directly. Such a modelling approach does reduce the complexity of calculations and depicts the fluctuating

nature of the residual service. However, the existing performance studies of CR networks based on such an assumption may be problematic, as they do not take the traffic pattern of primary users into account. As reiterated, the fact throughout this thesis is that self-similar nature consists in traffic flows of almost all communication networks. To fill the gap, this chapter aims to provide an analytical tool for estimating the residual service left by the primary user by taking its traffic pattern into account and predict queueing behaviours of individual secondary users of the CR network. It is worth noting that the channel capacity considered in this chapter follows general Gaussian distributions instead of being a constant, because the fading effect has significant impact on channel capacity and results in its variability and randomicity [136, 137, 138].

The remainder of this chapter are organized as follows: Section 7.2 introduces the functionality and how packets are being serviced through the designated CR network. Performance analysis of the CR networks is demonstrated at two steps in Section 7.3. Firstly, the residual service left by the primary user is modelled by a stochastic process. Afterwards, the approach of obtaining the queue length distribution of individual secondary users is given. In Section 7.4, the developed model is validated under various representative scenarios. Finally, this chapter is summarized in Section 7.5.

7.2 System Description

In this section, essential preliminaries for the analytical model regarding the CR technique will be demonstrated.

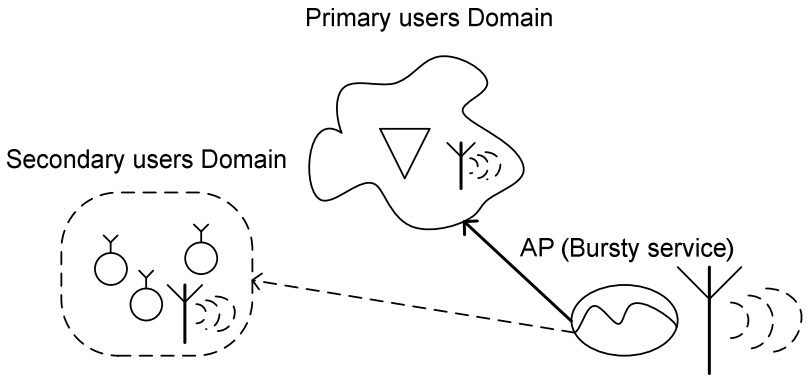


Figure 7.1: A CR network model in local domain.

From Figure 7.1, the designated CR network is composed of a primary user and a group of secondary users equipped with a unique server (i.e., AP) which has bursty service capacity. Two classes of users contend for the respective transmission. Secondary user network which contains only the secondary users can not plunder service from the primary user network. The secondary users are assumed to be able to perfectly sense and evade the existence of the primary user in order that the secondary users are able to contend for transmission immediately after the channel is available and no primary user is detected. A dynamic spectrum access scheme is employed in the CR network. It allows and regulates the users to transmit packets by sharing wireless channels. The dynamic spectrum access scheme introduced in this chapter mainly contains two parts. Firstly, it restricts the high priority to the primary user. Once the packet of the primary user senses the channel idle, it transmits right after it reaches the head of the queue. Secondly, the dynamic spectrum access coordinates the transmission among the secondary users.

The head-of-line (HOL) packets of individual secondary users are responsible for monitoring the channel status. When the channel is sensed idle and

persists for a period of Distributed Inter-Frame Space (DIFS), which is a fixed time interval, the secondary users starts backoff progress. Backoff is almost the most recognized solution to resolve contention between different users which are excepting to access the medium. It is essential because there is not a controller which manages the order of transmission of different users in distributed MAC. Different users send packets simultaneously will result in the collision which has impact on the performance of the network, especially throughput. By employing backoff, the collision could be minimized when multiple secondary users with HOL packets sense the channel idle at the same time. Specifically, the exponential backoff algorithm [139] is employed for resolving collisions. The method requires each user to randomly select a number which is uniformly distributed in the range $(0, W-1)$, where W is defined as the contention window in this chapter. $W = W_{\min}$ is the initial value of any packet for its first transmission attempt. For each transmission failure of one packet after its fist attempt, W is doubled, up to a predefined maximum value $W_{\max} = 2^b W_{\min}$, here b is the number of transmission failures. The packet will be sent when the corresponding backoff reduces to 0. It is worth noting that the backoff countdown process deactivates when a transmission is detected on the channel, and resumes when the channel is sensed idle again after a DIFS. In this study, the transmission failure of a packet occurs only when this packet is collided. It is not needed for the primary user to uphold a backoff algorithm for the reason that it transmits the HOL packet only on the condition that the channel is idle.

7.3 Performance Modelling

7.3.1 Modelling of the residual service

The dynamic spectrum access manages the service among the network users. It is not hard to see that the maximum service assigned to secondary users lies on the behaviour of the primary user. In addition, the service actually received by each secondary user is interconnected. To eliminate the interactions in the CR network, each user is finally isolated from the original network step by step.

Firstly, all the secondary users are isolated from the CR network by the residual service left by the primary user. The obtained residual service guarantees that the performance of the secondary users in the isolated system is statistically equivalent to that of the original one. In detail, the service discipline which handles the primary and secondary users is inherently equivalent to PQ scheduling scheme. Accordingly, the primary queue is almost certainly empty of the reason that the high priority queue is served with the entire capacity of the system and the total queue of a PQ system is almost composed of that the low priority part. To this end, a reasonable assumption can be made, that the output process of the primary user, denoted by $r'(t)$ is identical to its original arrival process. Hence, the residual service capacity left by the primary user can be given

$$\tilde{C}_s(t) = \tilde{C}(t) - r'(t) = \tilde{C}(t) - A_p(t), \tag{7.1}$$

where $\tilde{C}(t)$ is the bursty service capacity of the original network. $A_p(t)$ denotes the original traffic flow of the primary user. It is defined as

$$A_p(t) = m_p t + \sqrt{a_p m_p}\, Z(t), \tag{7.2}$$

where m_p and a_p are the mean arrival rate and the variance coefficient of the corresponding traffic flow. $Z(t)$ has two properties: 1) it is a normalized fractional Brownian motion and has stationary increments; 2) The first and the second moments of $Z(t)$ are 0 and t^{2H_p} for all t, H_p is the Hurst parameter of the primary traffic flow. The variance function of this formulized fBm process is given by

$$v_p(t) = a_p m_p \bar{v}_p(t) = a_p m_p t^{2H_p}. \tag{7.3}$$

Then, the covariance of fBm process can be obtained by

$$CV(t_s, t_q) = \frac{1}{2} am(t_s^{2H} + t_q^{2H} - (t_s - t_q)^{2H}). \tag{7.4}$$

In the virtual system shown in Figure 7.2, all secondary users are treated as a unique arrival, i.e., the secondary network. Hence, the virtual system is composed of two arrival sessions. The session of high priority part is the primary user, while the other session of low priority is the secondary network.

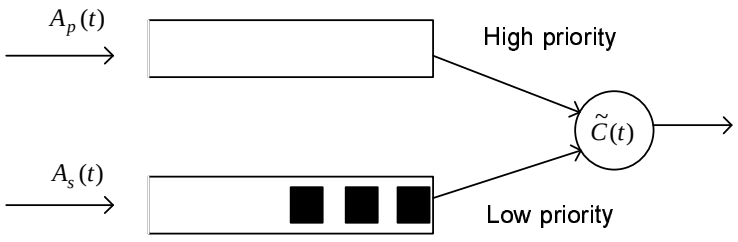


Figure 7.2: A virtual PQ system

It is worth noting that the overall service capacity is characterized by a Gaussian stochastic process and defined as

$$\tilde{C}(t) \sim (n, v), \tag{7.5}$$

where n and v are the mean and variance of the service process, and the variance function this Gaussian process is $a_g nt$ where a_g is the variance coefficient.

According to the PQ scheduling mechanism, the primary user receive the overall service capacity and hence can be readily isolated from the original system. However, the queueing performance of the primary user is not being derived because of the under utilized spectrum fact aforementioned.

On the other hand, the queueing behaviour of the secondary users is of concern. It is not hard to see the actual service of the secondary network, $\tilde{C}_s(t)$ is the reminder of extracting $A_p(t)$ from $\tilde{C}(t)$. $\tilde{C}_s(t)$ can be modelled by a Gaussian process with mean $n - m_p$, variance $v + a_p m_p$ and variance function $a_p m_p t^{2H_p} + a_g nt$ for the following reasons: 1) $\tilde{C}_s(t)$ has stationary increments since it can be characterized by its mean and variance; 2) $\tilde{C}_s(t)$ can be justified by the Central Limit Theorem.

It is obvious that the traffic flow of the secondary network is not involved in the procedure of obtaining the residual service capacity above. In other words, the residual service capacity is efficient no matter how the secondary network is organized or which protocol is employed for aiding transmission. However, the virtual traffic flow of the secondary network is defined as $A_{sn}(t)$ for corroborating the efficiency of the decomposition approach and the accuracy of the obtained service capacity.

7.3.2 Service Capacity of Individual Secondary Users

The second critical issue in this chapter is how to analyze the performance of individual secondary users in a CR network. With the residual service capacity which is obtained in Section 7.3.1, the primary user and the secondary network have been isolated from each other. In what follows, an approach to obtaining the service capacity of each secondary user is presented. With the derived effective service capacity, each secondary user can be isolated from the secondary network and forms a SSSQ system with the corresponding effective service. Consequently, the queueing performance of each secondary user can be obtained by examining the corresponding SSSQ system.

In this study, the traffic flows of the secondary users follow identical distribution, which is denoted as

$$A_{su}(t) = m_{su}t + \sqrt{a_{su}m_{su}}Z(t), \tag{7.6}$$

where m_{su} and a_{su} are the mean arrival rate and variance coefficient, respectively. The Hurst parameter of the traffic flow is denoted by H_{su}. For ease and clarity of explanations, the following definitions are given

p'	channel idle probability
p_0	the probability that a station (secondary user) is empty
τ'	transmission rate (saturation)
τ	transmission rate (non-saturation)
p_c	collision probability
σ	mean service time of a packet
σ'	time interval between the starts of two consecutive decrements of the backoff counter
m	the number of collisions before a successful transmission
n_{su}	the number of users in secondary network
$W_{\min}$	the minimum contention window
T_{SIFS}	short interframe space

T_{DIFS}	DCF interframe space
T_{ACK}	time to transmit an ACK message
T_{header}	time to transmit the MAC header and PHY header
T_{body}	time to transmit the packet body/payload

For each secondary user (user and station are interchangeable in this part), it can only establish the transmission when the corresponding queue is non-empty, the transmission probability of a secondary user can be given by

$$\tau = \tau'(1-p_0), \tag{7.7}$$

where τ' denotes the probability that a user transmits under the saturated traffic condition. According to the famous Bianchi model [139], the expression of τ' is given by

$$\tau' = 2(1-2p_c)^{-1}\Big((1-2p_c)(W_{\min}+1) + p_c W_{\min}(1-2^b {p_c}^b)\Big). \tag{7.8}$$

The corresponding definitions can be found in the above list. p_0, the empty probability of a station is equal to the server utilization of the corresponding SSSQ because the buffer is infinite. On the other hand, the channel idle probability, p', is given by

$$p' = (1-\tau)^{n_{su}}, \tag{7.9}$$

because the channel is only idle when none of the users transmits. Likewise, collisions occur then one of n_{su} users is transmitting and at least one of the remaining $n_{su}-1$ users transmits simultaneously. Hence, the collision probability can be expressed as

$$p_c = 1-(1-\tau)^{n^s-1}. \tag{7.10}$$

The actual service capacity received by individual secondary users is the reverse of its mean service time, namely,

$$C_{su} = 1/\sigma . \tag{7.11}$$

The mean service time is the interval from the instance that a packet becomes an HoL packet to the point that it is successfully transmitted. Mathematically, it can be denoted as

$$\sigma = T_{access} + T_{trans}, \tag{7.12}$$

where T_{access} is the mean channel access delay of an HoL packet. The service time is assumed to be exponentially distributed. Hence, the actual received service capacity by a single secondary user is a Poisson process with variance function $C_{su}t$. Provided that a packet is successfully transmitted after experiencing a number of m collisions, its access delay consists of the delay from m unsuccessful transmissions and delay from $(m+1)$ backoff stages. Therefore, we have

$$T_{access} = \sum_{m=0}^{\infty} \left(mT_c + \sigma' \sum_{i=0}^{m} \frac{W_i - 1}{2} \right) p^m (1-p), \tag{7.13}$$

where T_c is the mean duration of a single collision. Hence, mT_c is the time wasted on m failed transmissions. T_c is given by

$$T_c = T_{header} + T_{body} + T_{ACK} + T_{SIFS} + T_{DIFS} . \tag{7.14}$$

Let p_s be the probability that there is a successful transmission among the remaining $(n_s - 1)$ users when the tagged user is in the backoff status. Hence, we have

$$p_s = (n_s - 1)\tau(1-\tau)^{n_s-2}. \tag{7.15}$$

Since the channel is idle with probability p_0, a successful transmission occurs with probability p_s, and a collision happens on the channel with probability $(1 - p_0 - p_s)$, σ' is given by

$$\sigma' = p_0 T_s + p_s T_{trans} + (1 - p_0 - p_s) T_c, \tag{7.16}$$

where T_{trans} denotes the successful transmission time of a packet

$$T_{trans} = T_{SIFS} + T_{DIFS} + T_{ACK} + T_{header} + T_{body}. \tag{7.17}$$

7.3.3 Queue Length Distribution

Given the effective service capacity of individual secondary users, the queue length distribution of each secondary user can be readily obtained by examining the corresponding SSSQ systems. Naturally, the queue length accumulating process can be expressed as

$$QL(t) = \sup_{j \le t}(A_s(j,t) - \tilde{C}_s(j,t)), t \in (-\infty, \infty). \tag{7.18}$$

Accordingly, Equation (7.19) always holds

$$P\{QL(t) > x\} \ge P\{(A_s(0,t) - \tilde{C}_s(0,t)) - (A_s(0,j) - \tilde{C}_s(0,j))\}. \tag{7.19}$$

Consequently, for a stable system it has

$$\begin{aligned} P\{QL(t) > x\} &= P\{QL(0) > x\} \\ &= P\{\sup_{t \le 0}[\tilde{C}_s(t) - A_s(t)] > x\} \\ &\ge \max_{t \ge 0} P\{A_s(t) - \tilde{C}_s(t) > x\} \end{aligned} \tag{7.20}$$

For a Gaussian distributed variable Y, the tail probability can be given by the following equation

$$P\{Y > y\} = \Phi\left[\frac{y - m_y}{v_y}\right], \tag{7.21}$$

where m_y and v_y are the mean and variance of Y. Combing Equations (7.20) and (7.21), a lower bound of the queue length distribution can be given as

$$P\{QL > x\} \geq \max_{t \geq 0} \Phi\left[\frac{(\hat{C} - \hat{m})t + x}{\sqrt{v_{\hat{C}}(t) + v_{\hat{m}}(t)}}\right]. \tag{7.22}$$

where $\hat{C}$ and $\hat{m}$ are the mean of the service capacity and arrival rate, respectively. $v_{\hat{C}}(t)$ and $v_{\hat{m}}(t)$ are the variance functions of the corresponding process. By employing a further approximation [9],

$$\Phi(y) \approx (2\pi)^{-1/2}(1+y)^{-1}\exp(-y^2/2), \tag{7.23}$$

and solving Equation (7.22) and (7.23), the lower bound of the queue length distribution can be finally given by

$$P\{QL > x\} \geq (2\pi)^{-1/2}(1+x)^{-1}\exp\left(-\frac{((\hat{C} - \hat{m})g + x)^2}{2(v_{\hat{C}}(g) + v_{\hat{m}}(g))}\right), \tag{7.24}$$

where $g = H_s(1-H_s)^{-1}(n - m_p - m_s)^{-1}$ obtains the maximum in Equation (7.22).

The following gives an analytical upper bound of the queue length distribution which has been introduced in [119] and also regarded as a basic approximation:

$$P\{QL' > x\} \le \exp\left(-\frac{(x-(\hat{C}-\hat{m})t^*)^2}{2(v_{\hat{C}}(t^*)+v_{\hat{m}}(t^*))} \right), \tag{7.25}$$

where $t^* > 0$ and minimizes

$$F(t) = \frac{(x-(\hat{C}-\hat{m})t)^2}{v_{\hat{C}}(t)+v_{\hat{m}}(t)}. \tag{7.26}$$

Based on experiments, the geometric mean of the above lower and upper bounds can be used as the approximation of the queue length distribution.

7.4 Model Validation and Performance Analysis

This section validates the developed model through comparing analytical and simulation results. At the first place, it is of great importance to examine the residual service capacity left by the primary user. Because, the derivation of queue length distribution of each secondary user is based on the residual service capacity. Secondly, further comparisons between the analytical and experimental results for the secondary user are demonstrated.

7.4.1 Validation of the Residual Service Capacity

The simulation is developed by employing C. The traffic for both primary and secondary users are generated by adopting the same conditionalized midpoint algorithm as demonstrated in the previous chapters. To verify that the residual service obtained is accurate, a unique virtual traffic flow is set as the input of the secondary network. It has been demonstrated in Section 7.3.1 that the secondary network traffic is not involved in the derivation of the residual service capacity. Hence, the analytical and experimental results of the secondary network can be

directly compared by examining an SSSQ system subject to the virtual traffic flow and the derived residual service capacity.

The parameter settings for the three cases of the primary and virtual secondary traffic flows are shown below:

Case 1 $m_p = 10\ p/s, m_s = 80\ p/s, a_p = 1, a_s = 1, H_p = 0.7, H_s = 0.7.$

Case 2 $m_p = 10\ p/s, m_s = 80\ p/s, a_p = 1, a_s = 1, H_p = 0.8, H_s = 0.8.$

Case 3 $m_p = 20\ p/s, m_s = 70\ p/s, a_p = 1, a_s = 1, H_p = 0.7, H_s = 0.7.$

The mean service capacity of the original channel is set to be 100 packets/second (p/s), and the variance is 100.

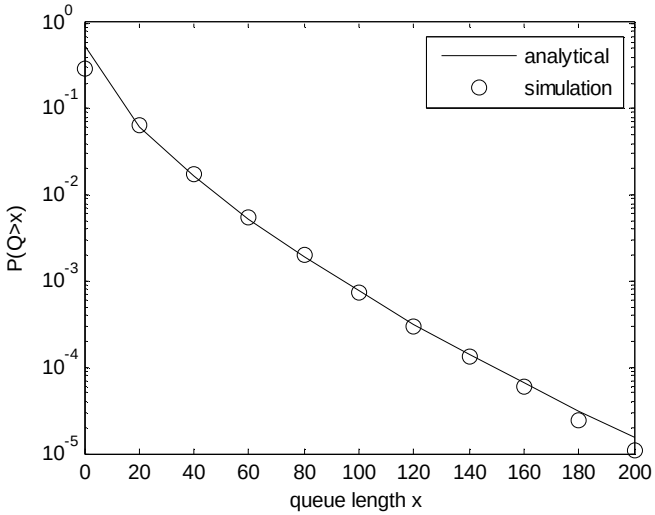


Figure 7.3: The analytical and simulation results on the distribution of queue length in Case 1.

In the first case, the mean arrival rate of the primary traffic is 10% of the total service capacity and that of the secondary traffic 80 p/s. Note that 10% is the widely reported proportion of the arrival arte of primary users to the service capacity of the

channel. Figure 7.3 shows that the analytical results of queue length distribution of the secondary traffic flow well matches the simulation result. Moreover, it can be seen from the same figure that t he experimental result is asymptotically exact to that of the simulation, as queue length increases. These observations validate that the residual service received by the secondary network, which is derived by the proposed approach is effective and accurate.

In the second case, the setting of the arrivals varies that of Case 1 on Hurst parameters. Hurst parameters of both primary traffic and secondary traffic are increased. This is an important alteration, since the Hurst parameter reveals the degree of self-similarity of the traffic and has impact on the variance of the traffic. Firstly, the phenomena which observed in Case 1 sill hold in this case. Compared to Case 1, it is worth noting that the proportion of the arrival rates to the service capacity has not changed in this case, however, the queue length increases significantly. This demonstrates the effect of the Hurst parameter on the queueing performance.

In case 3, the proportion of the total arrival to the original service capacity remains. However, the arrival rates of the primary and secondary users are re-configured. Specifically, the load of the primary traffic is increased from 10% to 20%, and meanwhile decreases the arrival rate of the secondary traffic to 70 p/s. From Figure 7.4, it can be found that the analytical result still shows excellent agreement with the simulation one. The phenomena which were observed in Case 1 still hold in this case.

By observing the analytical and simulation results of all three cases, there are two findings: 1) The variable service process that models the dynamic residual

service left by the primary user is truly effective; 2) The approach that is adopted to calculate the queue length distribution under variable service capacity is accurate.

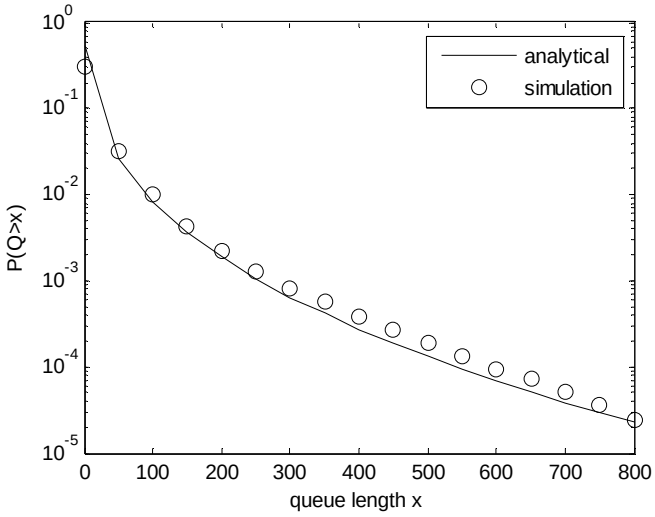


Figure 7.4: The analytical and simulation results on the distribution of queue length in Case 2.

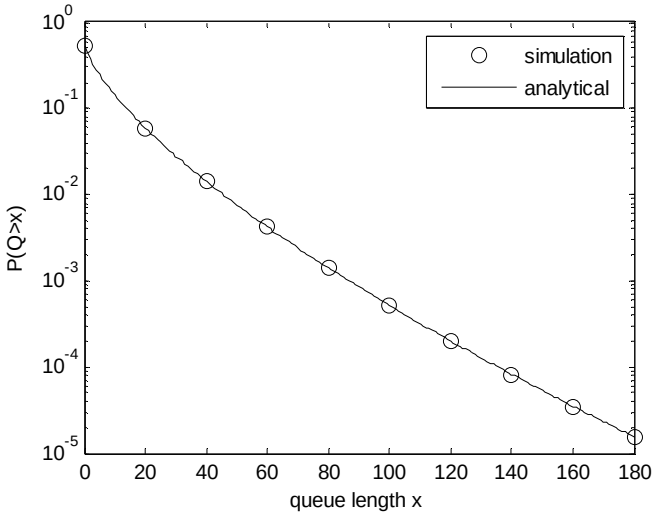


Figure 7.5: The analytical and simulation results on the distribution of queue length in Case 3.

7.4.2 Validation of Queueing Performance of Secondary Users

Because the secondary network has no impact on deriving the residual service, this service is valid to any secondary network no matter how they are organized. This study applies the residual service to the secondary network which employs the dynamic spectrum access scheme which is introduced in the previous sections to coordinate the data transmission of secondary users. The parameters are shown in Table 7.1.

Two cases demonstrate the validity of the developed model. In the first case there are 5 stations/secondary users contending for the residual service capacity. The traffic flows of the secondary users are identical and given as and the channel bit rate is 22 Mbits/s. The arrival of each secondary user is $m^s = 160\,p/s, a^s = 1, H^s = 0.75.$

Table 7.1
ODCF System Parameter Settings

Packet size	8184 bits
MAC header	272 bits
PHY header	128 bits
ACK	112 bits + PHY header
Propagation Delay	1 μs
Slot Time Length	50 μs
DIFS period	128 μs
SIFS period	28 μs
Minimum Contention Window	32
Maximum Contention Window	1024

The traffic flow of the primary user has the same variance coefficient and the Hurst parameter, while the mean arrival rate is 5% of channel capacity in terms of

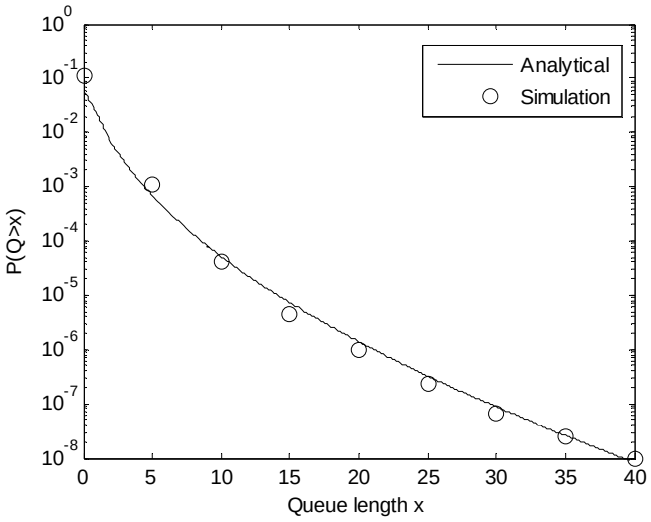


Figure 7.6: Analytical and simulation results of the case with 5 stations.

packet. This study did not give the queue length of the primary user for the reason that its payload is too low comparing to the channel capacity. On the other hand, Figure 7.6 presents the queue length of each secondary user. Apparently, the analytical result agreed with the simulation result even when the order of magnitude drops to 10^{-8} .

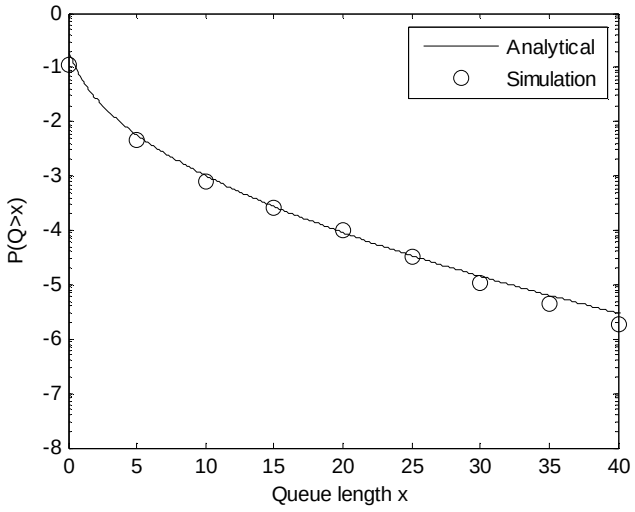


Figure 7.7: Analytical and simulation results of the case with 10 stations.

For the secondary case, the number of the contending users is increased to 10. Still, the traffic flow setting is given as $m^s = 58\,p/s, a^s = 1, H^s = 0.75$. The channel bit rate is down to 11 M bits/s which is also a most employed value of practical wireless networks. The arrival rate of the primary user is 60 p/s.

By examining Figure 7.7, it can be readily found that the analytical result shows excellent agreement with the simulation result. This implies the appropriateness and efficiency of the developed approach of obtaining the service of the secondary user network. In addition, it can be deemed that the developed model for studying the queue length distribution of individual secondary users in CR networks is feasible and accurate.

7.5 Summary

Recently, there has been a growing interest in studying the Cognitive Radio (CR) technique, which provides a means to address the issues resulting from the increasing demand of spectrum usage and the inadequate available spectrum. A lot of research efforts have been made on various topics regarding CR networks, such as, spectrum sensing, opportunistic spectrum sharing and maximized downlink throughput.

However, little work has been reported on analyzing the residual service left by the primary user and the queueing performance of secondary users in present of self-similar traffic. This chapter has developed an analytical model for addressing the performance of CR networks, where self-similar traffic is adopted as the input of both primary and secondary users and the variable channel capacity is modelled as a stochastic Gaussian process by taking the fading nature of wireless channels into

account. Specifically, the primary and secondary users are isolated from the original CR network respectively by employing a service decomposition approach. Subsequently, the queue length distribution of each secondary user is obtained by further deriving their corresponding effective service that is equivalent to the actually received service in the CR network. Finally, the performance of the complex secondary network was addressed by examining the simple SSSQ systems subject to the original inputs and with the obtained effective service.

Extensive experiments have validated the developed approaches to obtain the residual service left by the primary user and the effective service actually received by individual secondary users, as well as the proposed model for investigating the queue length distribution of individual secondary users. The analytical results show excellent agreements with simulation ones. Therefore, it can be concluded that the developed model can be adopted as an efficient tool for studying the performance of CR networks in the presence of self-similar traffic.

Chapter 8

Conclusions and Future Work

With the growing proliferation of network based multimedia applications, the provisioning of differentiated QoS in communication networks has become an increasingly pressing demand. The differentiated QoS has induced tremendous research efforts on traffic scheduling mechanisms. This chapter summarizes the major work reported in the thesis with the focus on the novel analytical models which have been developed to investigate and measure the performance of various scheduling systems. Further, the trend of the development of the resource allocation scheme and future direction of studying the scheduling algorithms will be discussed.

8.1 Conclusions

Scheduling mechanisms are pervasively implemented in practical communication systems to enable the provisioning of DiffServ QoS. In this thesis, novel analytical tools for performance analysis of centralized and decentralized scheduling systems in the presence of self-similar traffic have been demonstrated. The appropriateness and accuracy of the developed models have been validated by the comparison between analytical and simulation results. The simulators of the designated scheduling systems are programmed in C. The fBm traffic traces are generated by using the conditionalized random midpoint displacement algorithm. The major achievements are summarized as below:

In Chapter 3, the Hurst parameter estimators are applied to investigate the practical traffic datasets provided by MIT Lincoln Lab. The measurement results has verified that the traffic exhibits self-similar nature. Further, the estimators have been employed to investigate the traffic datasets which include Denial-of-Service (DoS) attacks. Through the experiments and observations, we have the following findings: 1) The attacked traffic datasets still featured self-similar nature. 2) The Hurst parameter increases in the presence of the huge amount of malicious traffic. Particularly, three commonly adopted attack mechanisms are taken into account. The traffic shapes which depict the datasets of different DoS attack mechanisms are different, i.e. constant rate, rump up behaviour and pulse. The same results regarding the Hurst parameter has been observed when the estimators are applied on these datasets.

In Chapters 4, 5, 6, analytical models for centralized scheduling mechanisms have been developed. As a start, DRR scheduling is studied, which is a promising scheduling mechanism owing to its nearly perfect fairness in terms of throughput and low computational complexity. The variable packet size has been fully taken into account in developing the analytical model and deriving the distribution of queue length for individual traffic flows. Further, the impact of the variable packet size on the performance of the DRR system has been examined by employing the developed model. Finally, the configuration of the weights of individual flows has been investigated in order to meet the desired QoS requirements.

Subsequentially, an analytical model for a hybrid PQGPS scheduling system has been developed. This hybrid scheduling mechanism integrates the PQ scheduling and GPS scheduling. This study is inspired by three Per-Hop Behaviours (PHBs) in

the DiffServ structure, i.e., expedite forwarding, assured forwarding and best effort forwarding. The multimedia applications which have stringent QoS requirements should be marked for expedite forwarding PHB. The time-insensitive applications could use assured forwarding, and the non-critical applications belong to class handled by best effort forwarding. PQGPS scheduling is a novel algorithm which can provide guaranteed and high-speed service in high priority level meanwhile support the assured forwarding and best effort service by configuring the weights in GPS scheduling system. The developed model of the hybrid scheduling system is an efficient tool for analyzing the performance of each traffic flow in the PQGPS system. Specifically, a bound approach has been developed to isolate the individual queues from the original system and hence convert the complex system into a group of Single-Server Single-Queue (SSSQ) systems. By doing this, the performance of the PQGPS system can be readily obtained by examining the obtained SSSQ systems.

Next, a nested PQ scheduling mechanism which is capable of supporting differentiated QoS with transmission error control is constructed. The transmission over the medium is not perfect all the time. The unforeseen circumstance may result in the transmission error. Therefore, a mathematical model has been developed for a multi-buffer Automatic Repeat reQuest (ARQ) system. The novelty of the multi-buffer ARQ system lies in its capability of providing prioritized service to different traffic flows, meanwhile retransmitting the error packets within the corresponding priority level. The developed model is an efficient tool for analyzing the loss probabilities for each buffer. In more detail, each subsystem which is composed of an arrival buffer and an ARQ buffer in the same priority level is separated by obtaining the effective service capacity. Furthermore, the ARQ buffer is isolated

from the arrival buffer by decomposing the effective service capacity of the subsystem. Eventually, the loss probabilities have been calculated. The developed model is validated by comparing the analytical results and simulation results.

Finally, an analytical model has been given for addressing the performance of CR networks, where self-similar traffic is adopted as the inputs and the variable channel capacity is modelled as a stochastic Gaussian process by taking the fading nature of wireless channels into account. The challenge of modelling a CR network lies in its medium access discipline. The primary users must not be disturbed during their transmission, and the secondary users only receive the residual service and contend for the chances of transmission. Unlike the existing studies, in this study the primary traffic is modelled as an fBm process. Then, we derive a bursty residual service, instead of directly using ON-OFF sources for secondary users. In more detail, the CR network has been divided into an SSSQ system and a secondary network which contains all secondary users. Afterwards, the effective service capacity is further derived in terms of the mean service time for each packet. Finally, the desired queueing performance has been obtained. Through the comparison between the analytical and simulation results, it has been demonstrated that the derived residual service capacity of the CR network is efficient regardless of how the secondary network is organized. Hence, the obtained residual service capacity is not limited to the CR network presented in this thesis, which makes it an efficient tool for analyzing the performance of primary and secondary users in CR networks.

8.2 Future work

Modelling of traffic scheduling mechanisms is till an open issue. To the best of our knowledge, there are a large number of constrains in modelling the exact behaviours in practical communication networks.

For the work presented in Chapter 5, we will try to extend the decomposition approach for GPS scheduling to $n, n \geq 3$ buffers in order to make a more general approach to analyzing the queueing performance of the complex system. The main theoretical basis which inspires us is the sharing policy of the excess service. The analytical model is developed based on allocating the excess service for two session GPS system, and hence it leads the way to $n, n \geq 3$ sessions scenario.

As aforementioned, it is onerous to find a perfect mathematical model for characterizing wireless channels in diversified environments. The actual channel capacity is affected by various reasons, such as energy consumption, medium and unforeseen interferences. A more general and practical channel model is preferred. In addition, performance modelling of CR networks with multi-input and multi-output is still an open issue.

For other aspects, our analytical models which are presented in this thesis assume an infinite buffer. However, if the buffer is finite, which is true for most existing practical communication systems such as routers and switches, the developed models can be enhanced following the clue of extending or modifying the queue length distribution and combining the knowledge of the G/G1/k queue.

References

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," *RFC 2475*, 1998.

[2] K.-D. Wu and W. Liao, "On service differentiation for multimedia traffic in multi-hop wireless networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 5, pp. 2464-2472, 2009.

[3] S. Androulidakis, T. Douokoglou, G. Patikis, and D. Kaqklis, "Service differentiation and traffic engineering in IP over WDM networks," *IEEE Communications Magazine*, vol. 46, no. 5, pp. 52-59, 2008.

[4] K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the differentiated servcies field in the IPv4 and IPv6 headers," *RFC 2474*, 1998.

[5] J. Matsumoto and Y. Watanabe, "Individual traffic characteristics queueing systems with multiple Poisson and overflow inputs," *IEEE Transactions on Communications*, vol. 33, no. 1, pp. 1-9, 1985.

[6] C.-C. Chan and S. V. Hanly, "Calculating the outage probability in a CDMA network with spatial Poisson traffic," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 1, pp. 183-204, 2001.

[7] V. Paxson and S. Floyd, "Wide-area traffic: the failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226, 1995.

[8] W. Fischer and K. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook," *Journal of Performance Evaluation*, vol. 18, no. 2, pp. 149-171, 1993.

[9] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, no. 3-4, pp. 387-396, 1994.

[10] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary user behavior in cellular networks and implications for dynamic spectrum access," *IEEE Communications Magazine*, vol. 47, no. 3, pp. 88-95, 2009.

[11] L. Zhang, Y. Xin, and Y.-C. Liang, "Weighted sum rate optimization for cognitive radio mimo broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 8, no. 6, pp. 2950-2959, 2009.

[12] Q. Zhao and B. M. Brian, "A survey of dynamic spectrum access," *IEEE Signal Processing magazine*, vol. 24, no. 3, pp. 79-89, 2007.

[13] W. E. Leland, M. S. Taqqu, and W. Willinger, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1-15, 1994.

[14] C. Oliveira, K. J. Bae, and T. Suda, "Long-range dependence in IEEE 802.11b wireless Lan traffic: an empirical study," in proc. of 18th Annual Workshop on Computer Communications (CCW'2003), pp. 17-23, 2003.

[15] L. Musavian and S. Aissa, "Capacity and power allocation for spectrum-sharing communications in fading channels," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 148-156, 2009.

[16] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated servcies netowrks: the single-node case," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 344-357, 1993.

[17] A. Demers, S. Keshav, and S. Shenkar, "Analysis and simulation of fair queueing algorithm," *Journal of Internetworking: Research and Experience*, vol. 1, pp. 3-26, 1990.

[18] I. F. Akyildiz, W.-Y. Lee, and K. R. Chowdhury, "Crahns: Cognitive radio ad hoc networks," *Journal of Ad Hoc networks*, vol. 7, no. 5, pp. 810-836, 2009.

[19] "MIT Lincoln laboratory DARPA intrusion detection evaluation data sets," http:// www.ll.mit.edu/ mission/ communications /ist/corpora /ideval/data /2000/LLS_DDOS_1.0.html.

[20] Z. Sun, D. He, L. Liang, and H. Cruickshank, "Internet QoS and traffic modelling," *IEE Proceedings on Software*, vol. 151, no. 5, pp. 248-255, 2004.

[21] C. H. Liew, C. K. Kodikara, and A. m. Kondoz, "MPEG-encoded variable bit-rate video traffic modelling," *IEE Proceedings on Communications*, vol. 152, no. 5, pp. 749-756, 2005.

[22] V. S. Frost and B. Melamed, "Traffic modelling for telecommunications networks," *IEEE Communications Magazine*, vol. 32, no. 3, pp. 70-81, 1994.

[23] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized hand off procedures," *IEEE Transactions on Vehicular Technology*, vol. 35, no. 3, pp. 77-92, 1986.

[24] L. Kleinrock, *Queueing Systems: Volume I - Theory*. New York: Wiley Interscience, 1975.

[25] P. D. Mitchell, T. C. Tozer, and C. Grace, "Bandwidth assignment scheme for ON-OFF type data traffic via satellite," *Electronics Letters*, vol. 37, no. 19, pp. 1191-1193, 2001.

[26] D. K. Kim, "Capacity estimation for an SIR-based power controlled CDMA system supporting ON-OFF traffic," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 4, pp. 1094-1101, 2000.

[27] T. C. Wong, J. W. Mark, and K. C. Chua, "Delay performance of voice and MMPP video traffic in cellular wireless ATM network," *IEE Proceedings on Communications*, vol. 148, no. 5, pp. 302-309, 2001.

[28] S. Y. Yousef, C. M. Strange, and J. A. Schormans, "ATM modelling: parameterisation of 4-phase MMPP model for admission control of superposed traffic sources," *Electronics Letters*, vol. 33, no. 10, pp. 829-830, 1997.

[29] P. Naor and U. Yechiali, "Queueing problems with heterogenous arrivals and service," *Operations Research*, vol. 19, no. 3, pp. 722-734, 1971.

[30] M. F. Neuts, "A queue subject to extroneous phase changes," *Journal of Applied Probability*, vol. 3, pp. 78-119, 1971.

[31] J. Beran, "Statistical methods for data with long-range dependence," *Statistical Science*, vol. 7, no. 4, pp. 404-427, 1992.

[32] J. Beran, *Statistics for long-memory processes*: Chapman and Hall, 1994.

[33] J. Beran, R. Sherman, and M. S. Taqqu, "Long-range dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications*, vol. 43, no. 234, pp. 1566-1579, 1995.

[34] X. Jin and G. Min, "Performance modelling of hybrid PQ-GPS systems under long-range dependent network traffic," *IEEE Communications Letters*, vol. 11, no. 5, pp. 446-448, 2007.

[35] F. Rossetto and M. Zorzi, "A low-delay MAC solution for MIMO ad hoc networks," *IEEE Transactions on Communications*, vol. 8, no. 1, pp. 130-135, 2009.

[36] E. Hossain and V. K. Bharagava, "A centralized TDMA-based scheme for fair bandwidth allocation in wireless IP networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 11, pp. 2201-2214, 2001.

[37] T. Dieker, "Simulation of fractional Brownian motion." Amsterdam: Vrjie Universityeit, 2002.

[38] I. Norros, P. Mannersalo, and J. L. Wang, "Simulation of fractional Brownian motion with conditionalized random midpoint displacement," *Annals of Advanced Performance*, vol. 2, no. 1, pp. 77-101, 1999.

[39] R. Ritke, X. Hong, and M. Gerla, "Contradictory relationship between Hurst parameter and queueing performance," *Telecommunication Systems*, vol. 16, no. 1-2, pp. 159-175, 2001.

[40] M. Lelarge, "Asymptotic behavior of generalized processor sharing queues under subexponential assumptions," *Queueing Systems: Theory and Applications*, vol. 62, no. 1-2, pp. 51-73, 2009.

[41] F. Lo Presti, Z.-L. Zhang, and D. Towsley, "Bound, approximations and applications for a two-queue GPS system," *Technical Report 95-109*, 1995.

[42] M. Shreedhar and G. Varghese, "Efficient fair queueing using deficit round robin," *IEEE/ACM Transactions on Networking*, vol. 4, no. 3, pp. 375-385, 1996.

[43] S. S. Kanhere, H. Sethu, and A. B. Parekh, "Fair and efficient packet scheduling using elastic round robin," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 3, pp. 324-336, 2002.

[44] X. Yuan and Z. Duan, "Fair round-robin: a low complexity packet scheduler with proportional and worst-case fairness," *IEEE Transactions on Computers*, vol. 58, no. 3, pp. 365-379, 2009.

[45] A. Sarkar, P. P. Chakrabarti, and R. Kumar, "Frame-based proportional round-robin," *IEEE Transactions on Computers*, vol. 55, no. 9, pp. 1121-1129, 2006.

[46] A. G. Greenberg and N. Madras, "How fair is fair queueing," *Journal of the ACM*, vol. 39, no. 3, pp. 568-598, 1992.

[47] R. Urgaonkar and M. J. Neely, "Opportunistic scheduling with reliability guarantees in cognitive radio networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 6, pp. 1-13, 2009.

[48] J. Yao, J. Guo, and L. N. Bhuyan, "Ordered round-robin: an efficient sequence preserving packet scheduler," *IEEE Transactions on Computers*, vol. 57, no. 12, pp. 1690-1703, 2008.

[49] S. Golestani, "A self-clocked fair queueing scheme for broadband applications," in proc. of IEEE INFOCOM'94, pp. 636-646, 1994.

[50] J. Nagle, "On packet switches with infinite storage," *IEEE Transactions on Communications*, vol. 35, no. 4, pp. 435-438, 1987.

[51] Z.-L. Zhang, D. Towsley, and J. Kurose, "Statistical analysis of the generalized processor sharing scheduling discipline," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1071-1080, 1995.

[52] M. Ashour and T. Le-Ngoc, "Multi-scale analysis of generalised processor sharing queues with long-range dependent traffic inputs and variable service rates," *IET Communications*, vol. 3, no. 6, pp. 992-1004, 2009.

[53] D. Stiliadis and A. Varma, "Latency-rate servers: a general model for analysis of traffic scheduling algorithms," *IEEE/ACM Transactions on Networking*, vol. 6, no. 5, pp. 611-624, 1998.

[54] A. Banchs and X. Perez, "Distributed weighted fair queueing in 802.11 wireless Lan," in proc. of International Conference on Communications (ICC 2002), 2002.

[55] S. S. Kanhere and H. Sethu, "On the latency bound of deficit round robin," in proc. of 11th ICCCN, pp. 548-553, 2002.

[56] L. Lenzini, E. Mingozzi, and G. Setea, "Performance analysis of modified deficit round robin schedulers," *Journal of High Speed Networks*, vol. 16, no. 4, pp. 399-422, 2007.

[57] Cisco, "Understanding and configuring MDRR/WRED on the Cisco 12000 series internet routers," *http://www.cisco.com/ warp/ public/ 63/ mdrr_wred_overview.html*.

[58] J. Babiarz, K. Chan, and F. Baker, "Configuration guidelines for DiffServ servcie classes," *RFC 4594*, 2006.

[59] M. Ashour and T. Le-Ngo, "Priority queueing of long-range dependent traffic," in proc. of GLOBECOM'04, pp. 3025-3029, 2003.

[60] C. Oottamakorn, S. Mao, and S. S. Panwar, "On generalized processor sharing with regulated multimedia traffic flows," *IEEE Transactions on Multimedia*, vol. 8, no. 6, pp. 1209-1218, 2006.

[61] X. Jin and G. Min, "An analytical model for generalized processor sharing scheduling with heterogeneous network traffic," in proc. of 22 ACM Symposium on Applied Computing, pp. 198-202, 2007.

[62] S. Borst, M. Mandjes, and M. van Uitert, "GPS queues with heterogeneous traffic classes," in proc. of 21st IEEE Conference on Computer and Communications, pp. 74-83, 2002.

[63] S. Borst, M. Mandjes, and M. van Uitert, "Generalized processor sharing with light-tailed and heavy-tailed input," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 821-834, 2003.

[64] Z. Quan and J. Chung, "Priority queueing analysis for self-similar traffic in high-speed networks," in proc. of 2003 IEEE International Conference on Communications, 2003.

[65] X. Jin and G. Min, "Flow-decomposition for performance prediction of GPS mechanism under homogeneous self-similar traffic," in proc. of 2007 IEEE International Conference on Networking, Sensing and Control (ICNSC'07), pp. 769-774, 2007.

[66] S. I. Maniatis, E. G. Nikolouzou, and I. S. Venieris, "QoS issues in the converged 3G wireless and wired networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 44-53, 2002.

[67] P. Larsson and N. Johansson, "Multi-user ARQ," in proc. of IEEE Vehicular Technology Conference, pp. 2052-2057, 2006.

[68] L. Le and E. Hossain, "An analytical model for ARQ cooperative diversity in multi-hop wireless networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1786-1791, 2008.

[69] J. Li and Y. Q. Zhao, "Resequencing analysis of stop-and-wait ARQ for parallel multichannel communications," *IEEE/ACM Transactions on Networking*, vol. 17, no. 3, pp. 817-830, 2009.

[70] C. Primentel and R. L. Siqueira, "Analysis of the go-back-N protocol on finite-state Markov Rician fading channels " *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2627-2632, 2008.

[71] M. A. Kousa, A. K. Elhakeem, and H. Yang, "Performance of ATM networks under hybrid ARQ/FEC error control scheme," *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 917-925, 1999.

[72] Z. Quan and J. Chung, "Asymptotic loss of real-time traffic in wireless mobile networks with selective-repeat ARQ," *IEEE Communications Letters*, vol. 8, no. 9, pp. 638-640, 2004.

[73] W. Luo, k. Balachandran, S. Nanda, and K. K. Chang, "Delay analysis of selective-repeat ARQ with applications to link adaptation in wireless packet data systems," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1017-1029, 2005.

[74] S. Lin and M. Miller, "The analysis of some selective-repeat ARQ schemes with finite receiver buffer," *IEEE Transactions on Communications*, vol. 29, no. 9, pp. 1307-1315, 1981.

[75] A. Chockaligam and M. Zorzi, "Adaptive ARQ with energy efficient backoff on Markov fading links," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1445-1449, 2008.

[76] A. Brandt and M. Brandt, "On a two-queue priority system with impatience and its applications to a call centre," *Methodology and Computing in Applied Probability*, vol. 1, no. 2, pp. 191-210, 1999.

[77] A. Brandt and M. Brandt, "On the two-class M/M/1 system under preemptive resume and impatience of the prioritized customers," *Queueing Systems*, vol. 47, no. 1-2, pp. 147-168, 2004.

[78] B. D. Choi and B. Kim, "M/M/1 queue with impatient customers of higher priority," *Queueing Systems*, vol. 38, no. 1, pp. 49-66, 2001.

[79] F. Iravani and B. Balcioglu, "On priority queues with impatient customers," *Queueing Systems*, vol. 58, no. 4, pp. 239-260, 2008.

[80] FCC, "Report of the spectrum efficiency working group," FCC Spectrum Policy Task Force 2002 http://www.fcc.gov/sptf/reports.html.

[81] M. A. McHenry and S. Chunduri, "Spectrum Occupancy Measurements, Location 3 of 6: National Science Foundation Building Roof, April 16, 2004, Revision 2," Shared Spectrum Company Report 2005

[82] M. A. McHenry, D. McCloskey, and J. Bates, "Spectrum Occupancy Measurements, Location 6 of 6: Shared Spectrum Building Roof, Vienna, Virginia, December 15-16, 2004," Shared Spectrum Company Report 2005

[83] M. A. McHenry, D. McCloskey, and G. Lane-Roberts, "Spectrum Occupancy Measurements, Location 4 of 6: Republican National Convention, New York City, New York, August 30, 2004 - September 3, 2004, Revision 2," Shared Spectrum Company Report 2005

[84] M. A. McHenry and K. Steadman, "Spectrum Occupancy Measurements, Location 2 of 6: Tyson's Square Center, Vienna, Virginia, April 9, 2004," Shared Spectrum Compancy Report 2005

[85] M. A. McHenry and K. Steadman, "Spectrum Occupancy Measurements, Location 1 of 6: Riverbend Park, Great Falls, Virginia," Shared Spectrum Company Report 2005

[86] M. A. McHenry and K. Steadman, "Spectrum Occupancy Measurements, Location 5 of 6: National Radio Astronomy Observatory (NRAO), Green Bank, West Virginia, October 10 -11, 2004, Revision 3," Shared Spectrum Company Report 2005

[87] Y.-F. Chen and N. Beaulieu, "Performance of collaborative spectrum sensing for cognitive radio in the presence of Gaussian channel estimation errors," *IEEE Transactions on Communications*, vol. 57, no. 7, pp. 1944-1947, 2009.

[88] D. Datla, R. Rajbanshi, A. M. Wyglinski, and G. J. Minden, "An adaptive spectrum sensing architecture for dynamic spectrum access networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 4211-4219, 2009.

[89] K. Hamdi, W. Zhang, and K. B. Letaief, "Opportunistic spectrum sharing in cognitive mimo wireless networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 4098-4109, 2009.

[90] A. Penttinen, *Traffic Modelling and Measurements*: Helsinki University of Technology, 1999.

[91] A. J. Field, U. Harder, and P. G. Harrison, "Measurement and modelling of self-similar traffic in computer networks," *IEE Proceedings on Communications*, vol. 151, no. 4, pp. 355-363, 2004.

[92] B. Mandelbrot, "How long is the coast of Britain? Statistical self-similarity and fractional dimension," *Science*, vol. 156, no. 3775, pp. 636-638, 1967.

[93] R. Clegg, "A practical guide to measuring the Hurst paramter," in proc. of 21st UK Performance Engineering Workshop, 2006.

[94] B. B. Mandelbrot and J. R. Wallis, "Computer experiments with fractional gaussian noise," *Water Resources research*, vol. 5, pp. 228-267, 1969.

[95] J. Geweke and S. Porter-Hudak, "The estimation and application of long memory time series models," *Journal of Time Series Analysis*, vol. 4, no. 4, pp. 221-238, 1983.

[96] Defense Advanced Research Projects Agency http://www.darpa.mil.

[97] W. H. Allen and G. A. Marin, "The LoSS technique for detecting new Denial of Service attacks " in proc. of IEEE Southeast Conference, pp. 302-309, 2004.

[98] Z.-Q. Gao and N. Ansari, "Differentiating Malicious DDoS attack traffic from normal TCP flows by proactive tests," *IEEE Communications Letters*, vol. 10, no. 11, pp. 793-795, 2006.

[99] J. Kong, M. Mirze, J. Shu, C. Yoedhana, M. Gerla, and S. Lu, "Random flow network modelling and simulations for DDoS attack mitigation," in proc. of International Conference on Communications, pp. 487-491, 2003.

[100] G. A. Marin, "Network Security Basics," *IEEE Security & Privacy*, vol. 3, no. 6, pp. 68-72, 2005.

[101] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage, "Inferring Internet Denial-of-Service activity," *ACM Transactions on Computer Systems*, vol. 24, no. 2, pp. 115-139, 2006.

[102] S. Ranjan, R. Swaminathan, M. Uysal, A. Nucci, and E. Knightly, "DDos-shield: DDoS-resilient scheduling to counter application layer attacks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 1, pp. 26-39, 2009.

[103] A. Sarika, A. Saumay, and G. Bryon, "DDoS attack simulation monitoring and analysis," CS 590D: Security topics in networking and distributed systems final project report, Purdue University, West Lafayette, IN 2004

[104] K. Y. Tau, C. S. Lui, F. Liang, and Y. Yam, "Defending against distributed denial-of-service attacks with Max-Min fair server centric router throttles," *IEEE/ACM Transactions on Networking*, vol. 13, no. 1, pp. 29-42, 2005.

[105] Y. Xiang and W.-L. Zhou, "Protecting web applications from DDoS attacks by an active distributed defence system," *International Journal of Web Information Systems*, vol. 2, no. 1, pp. 37-44, 2006.

[106] Y. Xie and S.-Z. Yu, "Monitoring the application-layer DDoS attacks for popular websites," *IEEE/ACM Transactions on Networking*, vol. 17, no. 1, pp. 15-25, 2009.

[107] Y. Yuan and K. Mills, "Monitoring the macroscopic effect of DDoS flooding attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 4, pp. 324-335, 2005.

[108] "TCP SYN flooding and IP spoofing attacks," CERT Advisory CA-1998-01 http://www.cert.org.

[109] "Smurf IP Denial-of-Service attacks," CERT Advisory CA-1998-01 http://www.cert.org/advisories/CA-98.01.html.

[110] J. Ansell, A. Bendell, and S. Humble, "Nested renewal processes," *Advances in Applied Probability*, vol. 12, no. 4, pp. 880-892, 1980.

[111] G. Dan, V. Fodor, and G. Karlsson, "On the effects of the packets size distribution on FEC performance," *Computer Networks*, vol. 50, no. 8, pp. 1104-1129, 2006.

[112] D.-Y. Eun and N. B. Shroff, "A measurement analytic approach for QoS estimation in a network based on the dominate time scale," *IEEE/ACM Transactions on Networking*, vol. 11, no. 2, pp. 222-235, 2003.

[113] Y. Fan and N. D. Georganas, "On merging and splitting of self-similar traffic in high-speed networks," in proc. of the 12th international conference on computer communication on Information highways : for a smaller world and better living: for a smaller world and better living, 1996.

[114] X. Jin and G. Min, "Performance analysis of priority scheduling mechanisms under heterogeneous network traffic," *International Journal of Computer and System Sciences*, vol. 73, no. 8, pp. 1207-1220, 2007.

[115] L. Mamatas and V. Tsaoussidis, "Differentiating services with noncongestive queueing (NCQ)," *IEEE Transactions on Computers*, vol. 58, no. 5, pp. 591-604, 2009.

[116] A. Nyandoro, L. Libman, and M. Hassan, "Service differentiation using the capture effect in 802.11 wireless LANs," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 2961-2971, 2007.

[117] G. Tan and S. A. Jarvis, "A payment-based incentive and service differentiation scheme for peer-to-peer streaming broadcast," *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, no. 7, pp. 940-953, 2008.

[118] D. Nandita, J. Kuri, and H. S. Jamadagni, "Optimal call admission control in generalized processor sharing (GPS) schedulers," in proc. of IEEE INFOCOM'01, pp. 468-477, 2001.

[119] P. Mannersalo and I. Norros, "A most probable path approach to queueing systems with general Gaussian input," *Journal of Computer Networks*, vol. 40, no. 3, pp. 399-411, 2002.

[120] A. Weiss, "An introduction to large deviations for communication networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 6, pp. 938-952, 1995.

[121] P. Mannersalo and I. Norros, "Approximate formulae for Gaussian priority queues," in proc. of 17th ITC, pp. 991-1002, 2001.

[122] X. Jin and G. Min, "Modelling and analysis of priority queueing systems with multi-class self-similar network traffic: A novel and efficient queue-decomposition approach," *IEEE Transactions on Communications*, vol. 57, no. 5, pp. 1444-1452, 2009.

[123] F. Lo Presti, Z.-L. Zhang, and D. Towsley, "Bounds, approximations and applications for a two-queue GPS system," in proc. of IEEE INFOCOM'96, pp. 1310-1317, 1996.

[124] Q. Huang, K.-T. Ko, and V. B. Iversen, "Approximation of loss calculation for hierarchical networks with multiservice overflows," *IEEE Transactions on Communications*, vol. 56, no. 3, pp. 466-473, 2008.

[125] Y. J. Liang, J. G. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: effect of burst losses and correlation between error frames," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 7, pp. 861-874, 2008.

[126] J. J. Metzner, *Reliable data communications*: New York: Academic, 1994.

[127] A. W. Berger and W. Whitt, "Effective bandwidths with priorities," *IEEE/ACM Transactions on Networking*, vol. 6, no. 4, pp. 447-460, 2003.

[128] J. T. Macdonald, "A survey of spectrum occupancy in Chicago," Illinois Institute of Technology: Technical Report 2007

[129] R. Chiang, G. Rowe, and K. Sowerby, "A quantitative analysis of spectral occupancy measurements for cognitive radio," in proc. of the 65th IEEE Vehicular Technology Conference (VTC Spring 2007), pp. 3016-3020, 2007.

[130] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP

framework," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589-600, 2007.

[131] C. Huang and S. A. Jafar, "Degrees of freedom of the mimo inference channel with cooperation and cognition," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4211-4220, 2009.

[132] *IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, 1997.

[133] L. Kleinrock and F. A. Tobagi, "Part I-carrier sense multiple-access modes and their throughput-delay characteristics," *IEEE Transactions on Communications*, vol. 23, no. 12, pp. 1400-1416, 1975.

[134] Y. Li, S. Mao, S. Panwar, and S. Midkiff, "On the performance of distributed polling service based medium access control," *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4635-4645, 2008.

[135] A. Abdrabou and W.-H. Zhuang, "Stochastic delay guarantees and statistical call admission control for IEEE 802.11 single-hop ad hoc networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 10, pp. 3972-3981, 2008.

[136] M. A. Kamath, B. L. Hughes, and X. Yu, "Gaussian approximations for the capacity of mimo Raleigh fading channels," in proc. of the 36th Asilomar conference on Signals, Systems and Computers, pp. 614-618, 2002.

[137] M. R. Mckay, P. J. smith, H. A. Suraweera, and I. B. Collings, "Accurate approximations for the capacity distribution of ofdm-based spatial multiplexing," in proc. of the IEEE International Conference on Communication (ICC'07), pp. 5377-5382, 2007.

[138] P. Smith and M. Shafi, "On a Gaussian approximation to capacity of wireless mimo systems," in proc. of the IEEE International Conference on Communication (ICC'02), pp. 406-410, 2002.

[139] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535-547, 2000.