Resource Allocation for Cellular/WLAN Integrated Networks

by

Wei Song

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The next-generation wireless communications have been envisioned to be supported by heterogeneous networks using various wireless access technologies. The popular cellular networks and wireless local area networks (WLANs) present perfectly complementary characteristics in terms of service capacity, mobility support, and quality-of-service (QoS) provisioning. The cellular/WLAN interworking is thus an effective way to promote the evolution of wireless networks. As an essential aspect of the interworking, resource allocation is vital for efficient utilization of the overall resources.

Specially, multi-service provisioning can be enhanced with cellular/WLAN interworking by taking advantage of the complementary network strength and an overlay structure. Call assignment/reassignment strategies and admission control policies are effective resource allocation mechanisms for the cellular/WLAN integrated network. Initially, the incoming calls are distributed to the overlay cell or WLAN according to call assignment strategies, which are enhanced with admission control policies in the target network. Further, call reassignment can be enabled to dynamically transfer the traffic load between the overlay cell and WLAN via vertical handoff. By these means, the multi-service traffic load can be properly shared between the interworked systems.

In this thesis, we investigate the load sharing problem for this heterogeneous wireless overlay network. Three load sharing schemes with different call assignment/reassignment strategies and admission control policies are proposed and analyzed. Effective analytical models are developed to evaluate the QoS performance and determine the call admission and assignment parameters. First, an admission control scheme with service-differentiated call assignment is studied to gain insights on the effects of load sharing on interworking effectiveness. Then, the admission scheme is extended by using randomized call assignment to enable distributed implementation. Also, we analyze the

impact of user mobility and data traffic variability. Further, an enhanced call assignment strategy is developed to exploit the heavy-tailedness of data call size. Last, the study is extended to a multi-service scenario. The overall resource utilization and QoS satisfaction are improved substantially by taking into account the multi-service traffic characteristics, such as the delay-sensitivity of voice traffic, elasticity and heavy-tailedness of data traffic, and rate-adaptiveness of video streaming traffic.

Acknowledgments

First and foremost, I wish to express my sincere appreciation to my supervisor, Dr. Weihua Zhuang, who has guided me all the way to this Ph.D. thesis. Being an excellent mentor, she has patiently led me to the academic research world step by step. I still remember clearly our first talk and her advice about study and life. The more I experience, the more deeply I understand the wise words. Her continuous and genuine support is indispensable for me to finish the Ph.D. study. I have learned so much from her rigorous research attitude, innovative thinking, and efficient work style.

Furthermore, I would like to thank Dr. Junshan Zhang, Dr. Xinzhi Liu, Dr. Sagar Naik, and Dr. Pin-Han Ho for serving on my thesis advisory committee. Also, I would thank Dr. Guangzhe Fan as a delegate. They have devoted precious time reading my thesis and helping me improve the quality. Their insightful comments and invaluable suggestions are really appreciated.

Special thanks go to Dr. Jon W. Mark and Dr. Xuemin Shen of the Centre for Wireless Communications (CWC). They create such an outstanding research lab and cultivate an open and free academic environment. I benefit greatly from their solid and broad knowledge, thoughtful views, and constructive advice. I am also very grateful to the other fellow colleagues of CWC for all stimulating discussion, fruitful collaboration, and generous sharing of expertise. My deep gratitude goes to Ms. Lin X. Cai and Ms. Fen Hou, who have been around me sharing happiness and tough times with me for the past four years.

Many thanks are also due to the University of Waterloo, a young and energetic star gaining global excellence. I am proud of having been part of it.

This thesis is devoted to my dear parents and brother for their everlasting love and support.

To my dear parents and brother

Contents

1 Introduction **1**

1.1 The Cellular/WLAN Interworking 1

1.2 Resource Allocation for Quality-of-Service Provisioning 3

1.3 Motivation and Research Contributions 6

1.4 Outline of the Thesis 8

2 Literature Review and Background **9**

2.1 The Cellular/WLAN Integration Architecture 9

2.2 Vertical Handoff Management 14

2.3 Call Assignment and Admission Control 19

2.3.1 Call assignment strategy 20

2.3.2 Call admission control 21

2.4 Multi-Service Provisioning 22

2.4.1 Conversational class 23

2.4.2 Interactive class 23

2.4.3 Streaming class 28

2.5 Summary 31

3 System Model and Research Topics **32**

3.1 The Cellular/WLAN Integrated Network 33

3.2 Multi-Service Traffic Model 35
3.3 Location-Dependent User Mobility Model 37
3.4 Research Topics 39
3.5 Summary 42

4 Admission Control with Service-Differentiated Assignment 43
4.1 Capacity Model 43
4.2 Admission Scheme with Service-Differentiated Assignment 47
4.2.1 Assignment strategy with service differentiation 47
4.2.2 Admission control policies 49
4.3 Performance Analysis of the Proposed Scheme 52
4.3.1 QoS evaluation for voice service 53
4.3.2 QoS evaluation for data service 58
4.4 Numerical Results and Discussion 65
4.5 Summary 69

5 Randomized Assignment for Distributed Implementation 70
5.1 Decentralization with Randomized Assignment 70
5.2 Determination of Assignment Parameters Based on MGFs 73
5.3 Numerical Results and Discussion 79
5.3.1 Accuracy validation of QoS evaluation approaches 79
5.3.2 Dependence of utilization on assignment parameters 87
5.3.3 Impact of user mobility and traffic variability 89
5.4 Summary 91

6 Size-Based Assignment with SRPT Scheduling 93
6.1 Assignment and Scheduling for Heavy-Tailed Data Calls 94

6.2 Performance Analysis of the Proposed Scheme 98
6.2.1 Analytical model for QoS evaluation 98
6.2.2 Determination of data call size threshold 103
6.3 Numerical Results and Discussion 104
6.3.1 Accuracy validation of the analytical model 104
6.3.2 Impact of data call size threshold 106
6.3.3 Performance improvement with the proposed scheme 109
6.3.4 Overload protection via SRPT scheduling 117
6.4 Summary 118

7 Multi-Service Load Sharing with Two-Way Reassignment 119
7.1 Video Streaming Service 120
7.2 Multi-Service Load Sharing Scheme 122
7.2.1 Initial call assignment with service differentiation 122
7.2.2 Two-way call reassignment via dynamic vertical handoff 126
7.3 Simulation Results and Discussion 128
7.4 Summary 133

8 Conclusions and Further Work 134
8.1 Major Research Results 134
8.2 Further Work 136

A WLAN Capacity Analysis 140

Bibliography 145

List of Abbreviations 158

List of Notations 161

List of Tables

4.1 Search algorithm for admission regions. 53
4.2 System parameters. 65
7.1 System parameters for simulations. 129
A.1 WLAN parameters. 144

List of Figures

2.1 Integration architectures for UMTS/GPRS and IEEE 802.11 WLANs. . 10
2.2 Structure of interactive data sessions. 24
2.3 Video frame size based on GBAR model. 30

3.1 System model for a cellular/WLAN integration network. 34
3.2 Modeling of user mobility within a cell/WLAN cluster. 39

4.1 WLAN throughput vs. voice and data traffic load. 46
4.2 Limited fractional guard channel policy for voice calls in the cell. 51
4.3 State transition rate diagram for voice calls in the cell. 56
4.4 State transition rate diagram for data calls in the cell. 62
4.5 Acceptable data traffic load vs. data admission region of WLAN. . . . 66
4.6 Acceptable data traffic load vs. voice admission region of WLAN. . . . 68

5.1 Analytical and simulation results of voice call QoS. 80
5.2 Analytical and simulation results of data call QoS. 81
5.3 Analytical and simulation results of mean data response time. 83
5.4 Analytical results of voice call QoS. 84
5.5 Analytical results of data call QoS in the WLAN-covered area. 85
5.6 Analytical results of data call QoS in the cellular-only area. 86
5.7 Acceptable data traffic load vs. fraction of voice traffic to WLAN. . . . 88

5.8 Acceptable data traffic load vs. fraction of data traffic to WLAN. . . . 89
5.9 Fraction of voice to WLAN vs. voice admission region of WLAN. . . . 90
5.10 Acceptable data traffic load vs. fraction of data traffic to WLAN. . . . 91

6.1 Data call QoS under PS service discipline. 96
6.2 Analytical and simulation results with varying λ_d. 105
6.3 Analytical and simulation results in heavy-tailed cases. 106
6.4 Voice and data call QoS vs. Φ_d under different load conditions. 109
6.5 Voice and data call QoS vs. Φ_d with different Weibull factors W_{L_d}. . . . 111
6.6 Performance of load sharing schemes vs. data traffic load λ_d. 113
6.7 Performance of load sharing schemes vs. Weibull factor W_{L_d}. 115
6.8 Mean data response time under SRPT or PS. 117

7.1 Session information collection with 3GPP PSS. 122
7.2 Flowchart of the multi-service load sharing scheme. 127
7.3 Voice call blocking probability of the WLAN-first scheme. 130
7.4 Data call QoS of different load sharing schemes. 131
7.5 Video streaming QoS of different load sharing schemes. 132

A.1 Average service rate for packets from one data flow. 143

Chapter 1

Introduction

Motivated by the ever-increasing demand for wireless communication services, the past decade has witnessed rapid evolution and successful deployment of wireless networks. It is widely accepted that next-generation wireless networks will be heterogeneous in nature with multiple wireless access technologies. While the heterogeneity poses new challenges to achieve interoperability among different wireless networks, their complementary characteristics can be exploited with the interworking to enhance service provisioning. The popular cellular networks and wireless local area networks (WLANs) are two most promising technologies, and the cellular/WLAN interworking has attracted much research attention from both the industry and academia.

1.1 The Cellular/WLAN Interworking

Cellular networks are originally designed to provide high-quality voice service with wide-area coverage. The first generation (1G) cellular networks are upgraded to the second generation (2G) with digital technologies. The 2G systems, e.g., the global system for mobile communications (GSM), are further extended with packet switching for more efficient data transmission. For example, the data transmission rate is increased from

9.6 kbit/s with GSM to around 100 kbit/s with the general packet radio service (GPRS). Currently, the third generation (3G) augmented with multimedia service support has been commercialized, such as the universal mobile telecommunication system (UMTS) and cdma2000©[1]. The UMTS system supports a data rate up to 2 Mbit/s with greater capacity and improved spectrum efficiency. However, the deployment cost remains high due to expensive radio spectrum and implementation complexity.

On the other hand, WLANs have also achieved great success and provide higher data rates at a much lower cost. For example, the most popular WLAN standard IEEE 802.11b operates at the license-exempt industrial, scientific, and medical (ISM) frequency band from 2.4 GHz to 2.483 GHz. It extends the physical (PHY) layer of the original 802.11 standard based on direct sequence spread spectrum (DSSS) and supports a data rate up to 11 Mbit/s. The subsequent revisions 802.11a and 802.11g employ orthogonal frequency-division multiplexing (OFDM) and offer a maximum rate of 54 Mbit/s at the unlicensed 5 GHz and 2.4 GHz bands, respectively. However, designed as a wireless extension to the wired Ethernet, a WLAN can only cover a small geographic area. For instance, an 802.11b access point (AP) can communicate with a mobile within up to 60 m at 11 Mbit/s and up to 100 m at 2 Mbit/s with omnidirectional antennas.

We can see that the two types of networks present complementary strength in terms of mobility support, data rate, and implementation cost. Cellular/WLAN interworking can provide mobile users with both ubiquitous connectivity and high-rate data service in hotspots. Indeed, the cellular/WLAN interworking is an efficient way to accelerate the evolution toward next-generation wireless networks. The standardization for 3G/WLAN interworking is now in progress by the 3rd Generation Partnership Project (3GPP) and the 3rd Generation Partnership Project 2 (3GPP2) from a cellular network operator's perspective. Six interworking scenarios are defined in 3GPP TR 22.934 [1]

[1] cdma2000© is a registered trademark of the Telecommunications Industry Association (TIA-USA).

to implement the 3GPP/WLAN interworking step by step. The interworking requirements, architecture, and procedures (e.g., network selection, authentication, charging, etc.) have been specified in 3GPP TR 23.234 [2]. Nonetheless, the specification on quality-of-service (QoS) provisioning is still limited to very high-level discussion, such as in 3GPP TR 23.836 [3].

1.2 Resource Allocation for Quality-of-Service Provisioning

It is known that mobile wireless networks exhibit some features distinct from wired networks. First, a wireless channel becomes time-varying and location-dependent due to radio propagation characteristics. The achievable channel capacity may be substantially degraded by impairments such as large-scale path loss and small-scale fading resulting from multipath time delay spread and Doppler frequency dispersion. To improve channel capacity, frequency reuse is enabled in cellular networks. QoS provisioning in cellular networks is further complicated by co-channel interference and user mobility. Resource allocation plays a key role in effectively provisioning QoS guarantee and efficiently utilizing the scarce radio resources. There has been extensive research on resource allocation for homogeneous wireless networks, which involves various aspects from the packet level to call level (connection level), such as packet scheduling and medium access control (MAC), flow and congestion control, QoS routing, and call admission control (CAC).

In the cellular/WLAN integrated network, resource allocation becomes much more challenging due to network heterogeneity. The resource allocation techniques need to be adapted to this heterogeneous networking environment and address many emerging new problems. To achieve a high utilization efficiency, the resources in the integrated systems should be jointly considered in the allocation. Also, the unique characteristics

of the integrated network should be taken into account.

- **Heterogeneous wireless access environment:** Aimed at different applications, the cellular networks differ from WLANs at the physical layer, medium access and link control layers. For example, both UMTS and cdma2000© employ code-division multiple access (CDMA). Schedulers located at base stations (BSs) coordinate multiple connections to access the shared wireless channel. For instance, a two-phase request-grant access is used for the multiple-access uplink (from the mobile to the base station). A mobile first sends a transmission request to the base station through a contention channel. The base station acknowledges the successful request and reserves necessary resource for data transmission to follow. The mobile then starts transmission using the allocated resource. The centralized control and reservation-based resource allocation, together with proper admission control to limit the traffic load, enable fine QoS provisioning in cellular networks. Nonetheless, due to the expensive frequency spectrum and implementation complexity, the deployment cost is very high. It is still rather challenging to support bandwidth-intensive services.

  In contrast, the channel access of most popular WLANs is contention-based random access, e.g., the mandatory distributed coordination function (DCF) of IEEE 802.11 WLAN. DCF adopts carrier sense multiple access with collision avoidance (CSMA/CA) and binary exponential backoff. The access point also needs to compete for access with other mobiles, different from a base station. As a result, the service provisioning is in a best-effort manner without QoS assurance. Although many QoS enhancement mechanisms are proposed for WLANs, such as admission control [4], QoS provisioning capability of WLANs is still very limited in comparison with that of cellular networks.

- **Hierarchical overlay network structure:** After three-generation evolution, cellular networks have widely entrenched infrastructure, which provides almost

ubiquitous connectivity and supports user mobility levels from fast highway vehicles to stationary users in an indoor environment. In contrast, WLANs are usually deployed disjointly in hotspot local areas, where the traffic intensity is typically much higher than surrounding areas. Thus, the cellular/WLAN interworking results in an overlay structure. Both cellular access and WLAN access are available to mobiles within WLAN-covered areas. The incoming traffic load should be properly shared between the overlay cell and WLAN for QoS enhancement, congestion relief, cost reduction, etc.

- **Multi-service traffic load:** It is expected that multi-service support will be an essential requirement for future wireless networks. Different services usually require different QoS deliveries. Real-time service such as voice and video are sensitive to delay, while the main concern for delay-tolerant data service is the throughput. Multiple services can take a good advantage of the complementary strength of the two networks. For example, benefiting from the centralized infrastructure, cellular networks can effectively serve real-time traffic with stringent QoS requirements. On the other hand, WLANs can be a good choice for elastic data service, which may experience traffic asymmetry at the uplink and downlink. The contention-based access of WLANs enables a virtually time division duplexing (TDD) mode to efficiently handle the load asymmetry and flexibly adapt to traffic elasticity. From the above observations, we can see that the service type is an important factor in resource allocation for cellular/WLAN interworking.

- **Location-dependent user mobility:** User mobility has been extensively studied for resource allocation in hierarchical cellular networks where microcells are overlaid with macrocells [5]. However, many design principles are not applicable to cellular/WLAN interworking, although there exists a similar overlay structure. For WLANs deployed in indoor hotspots such as offices and hotels, the low user mobility within these areas results in a heavy-tailed residence time [6]. Conse-

quently, a uniform mobility model becomes invalid for a large cell. As the user residence time in a cell or WLAN directly affects channel holding time, this new characteristic further complicates the resource allocation.

1.3 Motivation and Research Contributions

Previously, the research on cellular/WLAN interworking focuses on relatively high-level issues such as integration architecture [7–9]. The interworking is considered from the perspectives of access control and security, mobility management, billing, etc. The objectives are to minimize modification to current network standards, reuse existing network infrastructure, and reduce implementation complexity at the same time. In particular, vertical handoff management has attracted substantial research attention. The handoff between wireless networks of different access technologies is referred to as *vertical handoff*, in contrast to *horizontal handoff* within a homogeneous wireless network, e.g., between base stations of cellular networks or access points of WLANs. Many vertical handoff algorithms are proposed to achieve seamless and fast handoff between the cell and the WLAN [10, 11].

However, there are still not many research efforts devoted to the resource allocation for cellular/WLAN interworking. This research area is actually very important as the interworking only becomes really meaningful if the overall resources are efficiently utilized. Many new challenges need to be addressed as outlined in Section 1.2. In particular, with an overlay structure, the incoming traffic load can be properly shared between the integrated systems. Calls originating in the overlay area should be first assigned to the covering cell or WLAN. Admission control then complements the assignment strategy accordingly by limiting the traffic load admitted to each network. Due to heterogeneous underlying network support, the call assignment strategy can significantly affect user QoS experience and resource utilization of the integrated network. Moreover, the desired load sharing can be enhanced with call reassignment, i.e.,

ongoing calls are dynamically transferred between the coupled cell and WLANs via vertical handoff. That is, the resources should be reallocated based on system dynamics for QoS improvement or higher utilization. There have been some research works in this area such as [12, 13]. Nonetheless, they may neglect the location-dependent user mobility, or address only single service. In fact, user mobility and traffic characteristics are key factors for efficient resource allocation. In this research, we focus on resource allocation issues for the cellular/WLAN integrated network via effective call assignment/reassignment and admission control. Unique characteristics of the integrated network have been taken into account, such as the network heterogeneity, location-dependence of user mobility, and multi-service traffic characteristics and QoS requirements.

Specially, we have investigated the cellular/WLAN integrated network to identify the characteristics to be considered in the resource allocation [14]. A practical system model is established to capture the key characteristics so as to facilitate the evaluation of interworking performance [15, 16]. In particular, it captures the hierarchical overlay structure, location-dependent user mobility characteristics, and complementary QoS provisioning features of the two networks. Further, we have developed effective call assignment strategies and admission control policies for voice and data services in the integrated network [17–20]. The overall resource utilization can be improved by exploiting the traffic characteristics such as delay-sensitivity of voice traffic and heavy-tailedness of elastic data traffic. The assignment and admission parameters can be properly determined with the proposed analytical approaches. Also, we have evaluated the impact of user mobility and traffic characteristics on interworking effectiveness. Moreover, call reassignment via dynamic vertical handoff has been studied to effectively complement the initial call assignment to achieve the desired load sharing [20, 21]. Last, we have extended the research on load sharing to a multi-service scenario, including conversational voice service, interactive data service, and video streaming service [21, 22].

The multi-service traffic load is jointly considered in the call assignment/reassignment, so that the overall resources of the integrated network are efficiently utilized for QoS provisioning.

1.4 Outline of the Thesis

This thesis is organized as follows. Chapter 2 presents a brief literature survey for important research issues on cellular/WLAN interworking, such as integration architecture, vertical handoff management, call assignment and admission control. Also, we review some background on multi-service provisioning. The system model is given in Chapter 3, which also highlights the research topics of this thesis. Chapter 4 proposes and analyzes an admission control scheme with service-differentiated call assignment. In Chapter 5, the admission scheme is extended with randomized call assignment to enable distributed implementation. Moreover, an effective QoS evaluation approach is developed to determine the assignment parameters and evaluate the impact of user mobility and traffic variability on resource utilization. In Chapter 6, an even higher utilization is achieved by exploiting the heavy-tailedness of data call size with a call assignment strategy based on a size threshold and the efficient *shortest remaining processing time* (SRPT) scheduling discipline. Chapter 7 investigates multi-service load sharing for conversational voice service, interactive data service, and video streaming service. Two-way call reassignment is considered to complement initial call assignment. Also, as call reassignment may involve large bandwidth variation with the heterogeneous accesses, rate adaptation is applied to video streaming calls. Finally, Chapter 8 gives conclusions of this research and outlines possible further work.

Chapter 2

Literature Review and Background

The cellular/WLAN interworking is a promising step in the evolution toward next-generation wireless networks. In the literature, there are many research works addressing the interworking issues such as integration architecture, vertical handoff management, and QoS provisioning.

2.1 The Cellular/WLAN Integration Architecture

Typically, WLANs can be deployed by a cellular network operator, a commercial WLAN operator, an authority of a public hotspot (e.g., airports and hotels), or a business enterprise for internal use [8]. The interworking mechanisms vary with the ownership, provisioned services, and target penetration level between the two networks. Basically, more exposure is possible to integrate WLANs owned by a cellular network operator itself. Many integration architectures have been proposed for cellular/WLAN interworking. According to the inter-dependence between the two networks, the integration architectures are classified into two categories in [23], i.e., tight coupling and loose coupling. The interworking can be very tight with an integration at the radio access network (RAN) level or less tight with the integration in the cellular core network (CN).

The interworking can also be loose with the two networks integrated beyond the core network and usually through an external Internet protocol (IP) network.

Tight-coupling architecture

In the tight coupling, the WLAN is connected to the cellular network as a cellular RAN. The integration point of the WLAN to the UMTS network can be the gateway GPRS support node (GGSN), the serving GPRS support node (SGSN), or the radio network controller (RNC). Figure 2.1 illustrates a simplified architecture of interworking UMTS/GPRS networks with 802.11 WLANs. The lines with tags "a", "b", and "c" are examples of tight coupling at different integration points.

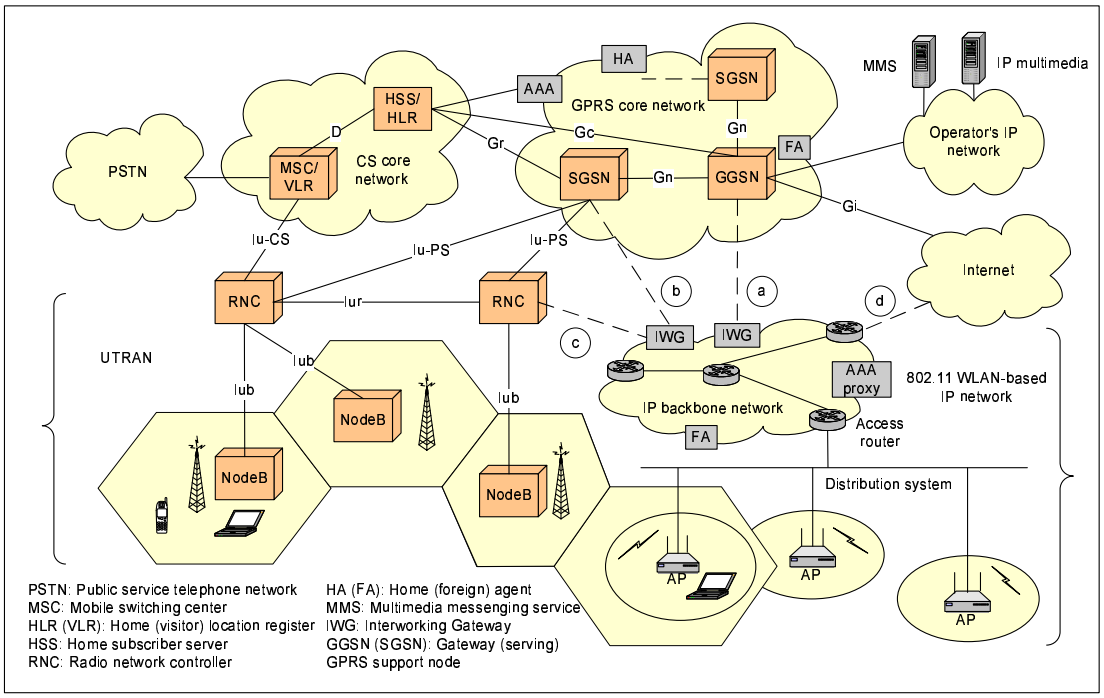


Figure 2.1: Integration architectures for UMTS/GPRS and IEEE 802.11 WLANs: (a) Tight coupling at GGSN; (b) Tight coupling at SGSN; (c) Tight coupling at RNC; (d) Loose coupling via an external IP network.

The integration architectures proposed in [24, 25] are typical tight-coupling examples with the WLAN interfaced at GGSN. The *802.11 gateway* in [24] and *SGSN emulator SGSN'* in [25] meet the UMTS core at the *Gn* interface if the WLAN is deployed by the UMTS operator or at the *Gp* interface if by another independent operator. Under the same GGSN, the WLAN and UMTS RAN are different routing areas (RA), while their associated mobiles have IP addresses assigned from the same pool. Then, roaming across the WLAN and UMTS results in `Inter-SGSN RA Update` without IP address change. Following the UMTS mobility management, the mobile's location is maintained by the home subscriber server (HSS), while the packet data protocol (PDP) context is tracked via the 802.11 gateway or SGSN'. Packets that arrive at the GGSN from external networks can then be tunneled to the mobile.

As proposed in [26–28], the tight coupling can also connect the WLAN to SGSN as an alternative cellular RAN via the *Iu-ps* interface or *Gb* interface for legacy GPRS network. In [26], an interworking gateway named GPRS interworking function (GIF) is implemented to hide WLAN particularities. In [28], an *RNC emulator* is used to connect the WLAN at the SGSN, similar to the *SGSN emulator SGSN'* in [25]. The UMTS layer-3 procedures can be followed for mobility management and session management. With the WLAN viewed as a typical RA by the SGSN, seamless mobility is achieved by means of `Intra-SGSN RA Update` instead of `Inter-SGSN RA Update` for the case of interworking at GGSN.

The integration is even tighter by coupling at the RNC level via the *Iub* or *Iur* interfaces. In [29], an interworking unit is proposed to handle integration-specific and radio procedures. For interworking at the *Iub* interface, WLAN-related signaling can be carried via the UMTS interface or over the WLAN. In contrast, the interworking unit working at the *Iur* interface emulates part of the functionality of a drift RNC associated with a serving RNC. As existing *Iur* interface does not support all control procedures related to call establishment like the *Iub* interface, the WLAN cannot

function independently of the UMTS network.

The preceding tight coupling can reuse the existing cellular infrastructure to a large extent, while the cellular radio is simply replaced with WLAN radio. User roaming across the two domains is mainly based on cellular mobility management, which enhances inter-domain mobility support. Particularly, the cellular infrastructure is mostly reused with the interworking at the RAN level and the handoff latency is expected to be minimal. The main disadvantage of tight coupling is the high implementation complexity. An interworking unit equipped with a cellular-compatible interface is necessary to expose the WLAN to the cellular core or even the cellular RAN. Moreover, as today's 3G cellular networks are designed to support low-rate data traffic, the 3G core may become potential bottleneck with the injection of high-rate WLAN traffic. Relatively, the interworking at GGSN may induce less congestion since a large part of data traffic bypasses the cellular core. Nonetheless, the network elements of the cellular core or RAN should be upgraded to support the extra traffic from WLANs. Furthermore, the cellular protocols tailored for highly mobile users in hostile outdoor environments may not operate properly for WLANs [30]. Thus, radio resource management (RRM) and radio resource control (RRC) need to be modified and adapt to the involved networks. With tight coupling, it is possible to apply joint resource allocation to optimize overall system capacity [31]. A best radio interface available can be selected with a proper decision algorithm. Therefore, the tight-coupling architecture with the above characteristics is ideal to integrate WLANs deployed by cellular operators.

Loose-coupling architecture

For the loose-coupling approach, the two networks are integrated beyond the core networks and usually through an external IP network such as the Internet (shown with the line "d" in Figure 2.1). Among this category, typical examples include the second architecture proposed in [24] and that in [26], the *peer network architecture* in [25],

the *Mobile IP approach* in [28], the *gateway approach* in [32], and the *operator WLAN system* (OWLAN) in [33].

It is well recognized that wireless networks are evolving toward the *all-IP* direction, as shown in the specification [34] for the 3G cellular network cdma2000©. Many IP-based technologies are introduced in the 3G core network, e.g., Mobile IP [35] for mobility management and authentication, authorization, and accounting (AAA) [36] framework for user access control. The 3G core networks evolve to function more like an IP backbone. For the WLAN as a wireless extension of wired Ethernet, since only the physical layer and link layer are specified, it is a natural choice to adopt popular IP-based protocols for higher layers. The loose-coupling architecture usually employs the pervasive IP technology to glue together the cellular network and WLANs, and follows the *de facto* standards of the Internet community such as the Mobile IP and AAA framework. As a result, within the two networks, different mechanisms can be implemented independently to handle user mobility, authentication, etc. Seamless roaming across the two networks can be achieved with the aid of Mobile IP. Also, with the flexible AAA framework, 3G-specific authentication mechanisms can be reused, while independent WLAN service providers can implement their preferred authentication methods such as the popular standards in the Internet community. Although imposing minimal modification to the WLAN, the loose coupling requires the cellular network to be augmented with extra functionalities such as Mobile IP and AAA support.

On the other hand, the loose coupling is relatively inefficient because of a long signaling path, redundant processing in the two networks, and a large number of network elements involved in management operations. For instance, because the mobility signaling has to traverse a long path across two separated domains, a relatively high handoff latency is induced. To overcome the inefficiency, it is necessary to apply extra techniques such as cross-layer control and context management. Mobile IP enhancement mechanisms such as regional registration and dynamic home agent assignment can be

applied to reduce handoff latency. The signaling procedure can also be simplified by coupling the authentication/authorization procedure with mobility management [37].

In fact, the future wireless network is expected to be a converged network, in which a common IP-based core network is shared by a variety of access networks [38]. The access technology-specific functions only propagate till the gateway to the shared core network. As such, the access heterogeneity terminates within the access network and homogeneous management is provided for mobility, security, QoS, etc. Although heterogeneous systems in the future converged network are expected to share a common core network as in the tight coupling, the shared core may not necessarily be the cellular core. Similar to the loose coupling, the common core network can be a separate external IP backbone. Being consistent with the converging trend and allowing for independent deployment and flexible implementation, the loose-coupling architecture has been preferably adopted in many research works [7, 24, 39] to address different interworking problems. Nonetheless, the tight coupling with enhanced performance tends to be the next logical step toward seamless cellular/WLAN interworking [40].

2.2 Vertical Handoff Management

Mobility management consists of two aspects, i.e., location management and handoff management. Location management continuously tracks the mobile's location, while handoff management maintains ongoing connections when switching attachment points. Generally, the handoff process is divided into three stages, i.e., initiation, decision, and execution. Depending on the decision entities, there are mobile-controlled handoff, mobile-assisted handoff, and network-controlled handoff.

Horizontal handoff within a homogeneous wireless network is inherently supported as a functionality of the mobility management. In the 3G cellular network UMTS, there is GPRS mobility management (GMM) for the link-layer and network-layer mobility. Tunneling protocols are used in the cellular core to support roaming. In cdma2000©

system, Mobile IP is introduced to provide network-layer transparency to IP-based applications (IP mobility) under the same packet data serving node (PDSN) and between different PDSNs. In contrast, current UMTS specifications only support IP mobility under the same GGSN node. In a three-state evolution specified in [41], Mobile IP is being considered for inter-UMTS or inter-technology IP mobility. An overview of the mobility management in UMTS and cdma2000© can be found in [42].

The mobility management in WLANs is much simpler since they are designed for local areas. In 802.11, a distribution system (e.g., an 802.3-type Ethernet) connects multiple basic service sets (BSSs) into an extended service set (ESS). Each BSS is under control of an AP in the infrastructure mode. Mobility across the BSSs within an ESS is handled by the APs involved. The layer-2 inter-access point protocol (IAPP) specified in 802.11f facilitates user roaming between APs of different vendors. When IP connectivity is provided in the WLAN, IP micro-mobility protocols can be further introduced to support IP mobility.

On the other hand, vertical handoff in heterogeneous wireless networks need to address many new challenges posed by network heterogeneity. In a loosely coupled cellular/WLAN network, the vertical handoff can be mobile-assisted or mobile-controlled, while tight coupling offers mobile-assisted or network-controlled vertical handoff with enhanced performance but high complexity [40].

Monitoring and measurement of network conditions

In order to make an intelligent handoff decision, timely information must be retrieved from candidate networks. Traditional metrics measured include received signal strength (RSS), signal-to-noise ratio (SNR) or bit error rate (BER), packet loss rate, etc. In the heterogeneous cellular/WLAN interworking environment, the information collection becomes much more challenging. Especially, the loose coupling may result in a large overhead and long latency for the information exchange between the two networks.

In [43], the network information is collected from mobiles via power control and link adaptation signaling and are managed with local databases. The data can be retrieved from the database upon request and transferred to the mobile through extended cell broadcast or in-band signaling in a piggyback fashion. In [32], the mobile's dual network interfaces are always enabled active for control messages. In this way, the mobile keeps receiving periodic advertisements from both networks indicating network conditions such as link performance, channel utilization, and traffic load.

Handoff decision algorithms

While the cellular network provides ubiquitous connectivity with wide-area coverage, WLANs are only deployed disjointly in hotspot areas. The cellular/WLAN interworking then results in an overlay structure, which offers both cellular access and WLAN access to dual-mode mobiles in WLAN-covered areas. A similar topology exists in hierarchical cellular networks, in which small-sized microcells are overlaid with large macrocells. However, the cellular network and WLANs differ intrinsically in the physical layer, medium access and link control layers. It is necessary to differentiate the *downward vertical handoff* from a cell to a WLAN and *upward vertical handoff* from a WLAN to a cell. Further, vertical handoff may originate from QoS enhancement or load balancing considerations other than maintaining connectivity. Hence, not only can vertical handoff proceed when a mobile moves out of the cell/WLAN border, but also back-and-forth vertical handoff can take place when a mobile moves within the cell/WLAN. Vertical handoff algorithms need to decide whether and when to perform a handoff to minimize the unnecessary handoff and the impact of ping-pong effect, and where to direct the handoff (i.e., radio selection) in case of multiple access interfaces.

Many works on the vertical handoff decision are based on metrics such as RSS [44, 45], SNR [46], and user moving speed [47]. Due to network heterogeneity, such traditional metrics in the two networks are rather disparate and should be used in a way

different from that for horizontal handoff. The handoff algorithm proposed in [48] uses the number of continuous WLAN beacon signals whose strength falls below a predefined level. The handoff thresholds are differentiated according to handoff direction and traffic delay-sensitivity. In [49], two handoff decision algorithms are compared with respect to a user satisfaction function. The first algorithm requires handoff to the WLAN whenever it becomes available, while no handoff is allowed in the second algorithm if the mobile is engaged in real-time or streaming sessions. It is observed that the second algorithm outperforms the first for both the corporate and on-the-move mobility models.

Moreover, there are advanced vertical handoff decision algorithms, which simultaneously consider different factors such as network characteristics, service type, user mobility, network conditions, user preference, and cost. The handoff decision problem is first formulated to introduce these factors and define the objectives, e.g., satisfaction of user requirement and maximization of revenue. Then, the decision problem needs to be solved efficiently with techniques such as fuzzy logic. In [50], the handoff decision is formulated as a fuzzy multiple attribute decision making (MADM) problem. To rank the candidate handoff targets, there are different MADM solutions such as the multiplicative exponent weighting [51], simple additive weighting [52], and the technique for order preference by similarity to ideal solution (TOPSIS) [52]. The vertical handoff decision in [53] is based on an integrated algorithm of analytic hierarchy process (AHP) and grey relational analysis (GRA). AHP quantitatively weights decision alternatives by hierarchical and pairwise comparison, while GRA ranks network alternatives efficiently through building a grey relationship with an ideal option.

Handoff execution procedures

Given a handoff decision, the handoff should be performed in a fast, smooth, and seamless way [54] to minimize handoff latency and packet loss during handoff. The most

popular network-layer solution is Mobile IP [35]. By introducing mobility agents and IP tunneling, upper-layer applications are provided transparency to IP address changes due to user movement. However, as the original Mobile IP protocol suffers from the triangle routing problem, a long handoff latency may be involved when the visited foreign network and home network are far apart. Hence, Mobile IP is more suitable for macro-mobility with infrequent movement and often between different administrative domains (inter-domain). Many Mobile IP variants for intra-domain micro-mobility are proposed to reduce handoff latency by means of localizing signaling via regional/hierarchical registration (tunneling-based) or host-specific routing (routing-based) [55].

In a case of vertical handoff in heterogeneous networks, there are specific techniques to enhance handoff performance, such as pre-handoff, bicasting, and link-layer triggering. Pre-handoff is started for necessary handoff preparation, such as `Binding Update` with home agent, authentication with new AAA server/proxy, resource reservation, etc. The pre-handoff activation can be based on RSS threshold [56], measurement region [57], and detection of a better network available [48]. Cross-layer techniques such as link-layer triggering can also enhance the handoff performance. The multi-layer hierarchical Mobile IPv6 scheme in [58] takes advantage of the layer-2 inter-access point protocol (IAPP). By means of the notification message from the associated AP, the mobility agent is promptly aware of the appearance of a mobile in a new subnet.

In addition to network-layer approaches, there are also transport-layer and application-layer solutions. The transport-layer scheme proposed in [59] supports UMTS and WLAN vertical handoff via stream control transmission protocol (SCTP). Although mobility management at the transport layer enables network-independence, more functions need to be introduced to end systems. A typical application-layer handoff solution is based on session initiation protocol (SIP) [60, 61]. SIP is a key signaling protocol for IP multimedia subsystem (IMS) of UMTS, in which real-time multimedia services are supported within a packet-switched domain. The UMTS-WLAN vertical

handoff can take advantage of SIP to facilitate handoff-associated negotiation for QoS, AAA and charging (AAAC) support. In general, application-layer solutions induce less modification to existing protocols and infrastructures of the 3G network and WLANs. Nonetheless, relatively longer handoff latency may be incurred with the lower-layer network attachment and SIP location update [60].

2.3 Call Assignment and Admission Control

With the cellular/WLAN interworking, there is ubiquitous cellular coverage, while both cellular access and WLAN access are available in the overlay area. Initially, an incoming new call should be properly assigned to either the covering cell or WLAN. The selected target network decides whether to accept or reject the call based on its admission control policy. If there is no sufficient free bandwidth to admit the call in the preferred network, the call can overflow to the other network or just leave the system. Moreover, if enough resources are released from call completion or outgoing handoff in the preferred network, an overflow call can be reassigned to its preferred network. The call reassignment is also referred to as *take-back* in some literature.

Complementing the aforementioned call assignment/reassignment strategies, the admission control policies in the target networks need to be properly designed to limit the admissible traffic load and provide QoS assurance. For example, as handoff dropping is more undesirable than new call blocking, handoff calls should be prioritized over new calls in the admission control policy, e.g., by means of reserving guard channels, queueing handoff calls, and so on. In addition, the admission control policy of wireless overlay networks needs to differentiate calls in different areas, since the accessible resources vary with locations.

According to the open systems interconnection basic reference model (OSI Reference Model), the call assignment and admission control are fundamental functions of the network layer. Due to heterogeneous wireless access technologies at the link layer

of the cellular network and WLANs, it is imperative to adapt these control functions to the available wireless links. Particularly, with the cellular/WLAN interworking, an incoming call should be properly routed and admitted to an underlying integrated system so that the requested QoS is supported efficiently. Different from wired networks, the wireless link capacity is highly time-varying and location-dependent. Hence, effective capacity evaluation is essential for admission decision. To ensure sufficient link capacity for QoS satisfaction, only when the admission request is accepted by the underlying link layer can the incoming call be admitted to a specific network.

In the literature, there has been extensive study on call assignment and admission control for hierarchical cellular networks with a similar two-tier overlay topology. While large macrocells (higher tier) provide wide-area coverage, small-size microcells (lower tier) further improve the capacity in urban areas by reducing the frequency reuse distance. As such, two tiers of coverage are provided by macrocells and microcells.

2.3.1 Call assignment strategy

User mobility is a most widely used factor in call assignment. Considering the small and possible disjoint coverage of microcells, incoming new calls from highly mobile users are preferably assigned to macrocells, while ongoing calls are not handed over to available microcells. Moreover, calls from fast moving users may not be allowed to overflow to the lower-tier microcells, as they may experience an intolerably high handoff rate and handoff failure probability. Although the overflow can promote the overall utilization of the resources in both tiers, user-perceived QoS may be unacceptable. Thus, the call assignment can be based on a user speed or residence time threshold, which depends on factors such as QoS requirements, handoff rate constraint, traffic balancing between the tiers, and so on. In [62], the speed threshold is dynamically adjusted with traffic load. When the traffic load is light, the speed threshold is decreased to keep more users in macrocells and undergo fewer handoffs. In contrast, the speed threshold is increased

under a heavier load so as to carry more mobiles in microcells of higher capacity. Similarly, the handoff rate control scheme proposed in [63] dynamically adjusts the speed threshold according to the handoff rate and call blocking probability.

Channel occupancy is another important factor for call assignment or network selection. In [64], partitioned buffer is used to buffer new calls, handoff calls, and overflow calls. The traffic load can be estimated by comparing the queue lengths of the overlay macrocell and microcell. Calls are preferably assigned to the macrocell or microcell with a shorter queue length, which indicates a lighter traffic load. Thus, the traffic load is balanced between the tiers to improve channel utilization. Moreover, service type is also essential for call assignment, since microcells at the lower tier usually support a higher data rate. High-rate service requests can be assigned to the lower tier with preference, while low-rate services are admitted to the upper tier if possible. In [65], two service classes are considered, which differ in bandwidth requirement. Calls requiring a larger bandwidth are preferably served at the lower tier, where a larger amount of resources are usually available. Access to the upper tier is only attempted as a consequence of call blocking or forced termination at the lower tier.

In [12], optimal and adaptive call assignment strategies are proposed for data service in hierarchical overlay networks. Both user velocity and amount of data to transfer are taken into account. The objective is to minimize the expected number of users in the system and expected load seen by an incoming user. The idea behind the strategies is to have admitted users depart from the system faster and thereby free more bandwidth for remaining users. Optimal thresholds for velocity and data amount are derived, which depend on not only the cell capacity, mean call arrival rate and average data size, but also the distributions of user velocity and data size.

2.3.2 Call admission control

Furthermore, there are some comprehensive works on call admission control with different assignment strategies, overflow/take-back policies, and new/handoff traffic dif-

ferentiation. The analytical model proposed in [66] studies a speed-sensitive assignment strategy and two-way traffic overflow/take-back. Nonetheless, the model delays take-back until microcell border crossing to render a tractable analysis. Moreover, it considers only one service type, i.e., traditional voice telephony, and no guard channels are reserved for handoff calls. The analysis shows that the tiered network with both overflow and take-back presents the best performance in comparison with schemes without overflow or with only one-way overflow and take-back.

Similarly, the velocity-based scheme in [67] considers bidirectional call overflow but no take-back of overflow traffic. In this model, guard channels are reserved in both macrocells and microcells for handoff calls. In [5], mobility changes are taken into account. Overflow and take-back are allowed only for slow mobiles, while fast mobiles are assigned only to macrocells. In [65], a cut-off priority is applied by reserving certain macrocell channels for calls in the area with only macrocell coverage. Thus, these calls are privileged over calls from areas with also microcell coverage.

In [13], an optimal joint session admission control scheme is proposed for multimedia traffic. It is based on a semi-Markov decision process (SMDP) to maximize the overall network revenue with QoS constraints. Saturated traffic (i.e., there is always backlogged data in the service queue of an active session) is considered for all service classes, whose QoS requirements are differentiated by packet delay, saturation bandwidth, and signal-to-interference ratio (SIR). However, the specific traffic details and user-perceived QoS metrics are neglected.

2.4 Multi-Service Provisioning

As an essential requirement for future wireless networks, multi-service support is also an important motivation for cellular/WLAN interworking. Four service classes are defined for UMTS systems in [68], i.e., the conversational, streaming, interactive, and background classes. The main distinguishing factor of the classification is the delay

sensitivity. The conversational class is highly delay-sensitive, while the background class is the most delay tolerant.

2.4.1 Conversational class

The conversational class is characterized by a two-way conversational communication pattern. A typical example is voice telephony, which has been extensively studied and widely deployed.

2.4.2 Interactive class

The interactive class comprises non-real time services with a request-response pattern, such as Web browsing, voice messaging, and file transfer. A main QoS criterion is the transfer delay (also known as response time) to measure the responsiveness, e.g., how fast a Web page is successfully downloaded and appears after it has been requested, or equivalently the call throughput, which is the ratio of file size over the transfer delay [69]. Although the transfer delay should be bounded to maintain fluent interactions, the delay requirement is much less stringent than that of conversational services [70]. A transfer delay of $2-4$ seconds per page is the proposed bound for Web browsing and a desirable target is 0.5 seconds.

Interactive sessions exhibit the on-off dynamics [71] shown in Figure 2.2. If the download of a Web page or a data file is viewed as a packet data call, an interactive session consists of a sequence of data calls (the "on" phases). After downloading a Web page, the user may take a "reading time" (the "off" phase), denoted by S_r, before requesting the next page, and finish the session after reading a number of pages. The reading time is assumed to be exponentially distributed, and the number of data calls in a session (denoted by M_d) is geometrically distributed.

For interactive data services such as Web browsing and file transfer, it is observed that the packet-level traffic presents asymptotic self-similarity and high variability over

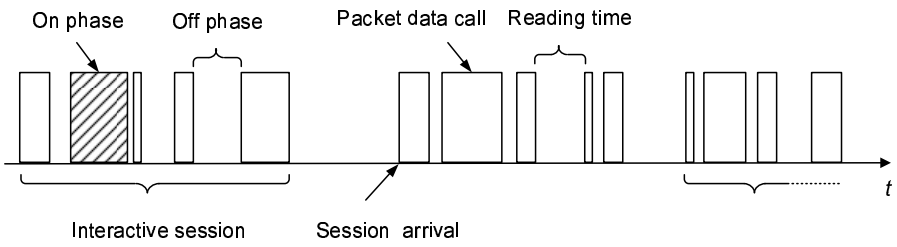


Figure 2.2: Structure of interactive data sessions.

a wide range of time scales [72]. This property is attributed to the heavy-tailed file size and burstiness induced by traffic control mechanisms such as closed-loop congestion control [73]. Although the complex packet-level traffic characterization may prevent feasible performance analysis, the QoS metrics of interest are actually more dependent on higher flow-level or session-level behaviors and less relevant to packet-level dynamics. For example, the mean response time depends on flow fluctuation and bandwidth sharing manner among in-progress flows [69]. As an essential call-level traffic characteristic, the heavy-tailedness of data file size has been extensively studied in the literature. There are many models to capture the statistics of real file size with well-known heavy-tailed distributions such as log-normal, Weibull, and Pareto distributions.

A non-negative random variable X with cumulative distribution function (CDF) $F(x)$ is said to be heavy-tailed if

$$\lim_{x\to\infty} e^{\lambda x}\overline{F}(x) = \infty, \qquad \text{for all } \lambda > 0 \tag{2.1}$$

where $\overline{F}(x) = 1 - F(x)$ is the survival function, also known as reliability function and complementary CDF. In particular, X is said to have a long right tail if

$$\lim_{x\to\infty} \frac{1 - F(x+y)}{1 - F(x)} = 1, \qquad x \geq 0, \quad y \geq 0. \tag{2.2}$$

It indicates that if X exceeds a large value x, it is likely to be greater than any larger value $x + y$ as well. All long-tailed distributions are heavy-tailed, but the converse is

false as it is possible to construct heavy-tailed distributions that are not long-tailed. Moreover, all commonly used heavy-tailed distributions have subexponentiality. All subexponential distributions are long-tailed, but examples can be constructed of long-tailed distributions that are not subexponential. The distribution of X is subexponential if

$$\overline{F^{*2}}(x) \sim 2\overline{F}(x), \qquad \text{as } x \to \infty \tag{2.3}$$

where $\overline{F^{*2}}(x) = 1 - F^{*2}(x)$ with F^{*2} being the 2-fold convolutions of probability distributions. For two independent, identically distributed (i.i.d.) random variables X_1 and X_2 with common distribution function $F(\cdot)$, the convolution of F with itself, F^{*2}, is defined as

$$F^{*2}(x) = \Pr[X_1 + X_2 \leq x] = \int_{-\infty}^{\infty} F(x-y)F(y)\, dy. \tag{2.4}$$

Then, the subexponentiality implies that, for any $n \geq 1$,

$$\overline{F^{*n}}(x) \sim n\overline{F}(x), \qquad \text{as } x \to \infty \tag{2.5}$$

where the n-fold convolution F^{*n} is defined similar to (2.4). The probabilistic interpretation of this is that, for a sum of n i.i.d. random variables $X_1, ..., X_n$,

$$\Pr[X_1 + ... + X_n > x] \sim \Pr[\max(X_1, ..., X_n) > x], \qquad \text{as } x \to \infty \tag{2.6}$$

which is often known as the principle of the single big jump. In other words, if X_i is the i^{th} claim of an insurance portfolio, the tails of the distribution of the sum and of the maximum of the first n claims are asymptotically of the same order [74]. The survival functions $\overline{F}(x)$ of subexponential distributions go to 0 more slowly than exponentially in the case of an exponential distribution [75]. In particular, a heavy-tailed distribution is said to be power-tailed, if the survival function decays according to a power law, i.e.,

$$\overline{F}(x) = \Pr[X > x] \sim x^{-\alpha}, \qquad \alpha > 0, \quad x \geq 0, \quad \text{as} \quad x \to \infty. \tag{2.7}$$

Power-tailed distributions are heavy-tailed but the reverse is not true, as their survival functions do not necessarily decay as slowly as a power-law function. The Pareto

distribution is both heavy-tailed and power-tailed, while the log-normal distribution is heavy-tailed but not power-tailed. A Weibull distribution is only heavy-tailed with a shape parameter within the range of (0, 1), but not power-tailed.

In [76], the data file size L_d is modeled by a Weibull distribution, whose probability density function (PDF) is given by

$$f_{L_d}(x) = \frac{\alpha_d}{\beta_d}\left(\frac{x}{\beta_d}\right)^{\alpha_d - 1} e^{-(x/\beta_d)^{\alpha_d}}, \qquad 0 < \alpha_d \leq 1, \quad \beta_d > 0, \quad x > 0 \tag{2.8}$$

where α_d is the shape parameter and β_d is the scale parameter. The PDF of the Weibull distribution is denoted by $W_b(x, \alpha_d, \beta_d)$ for simplicity. The mean of L_d is given by $\mathrm{E}[L_d] \triangleq \overline{L}_d = \beta_d\, \Gamma(1 + \frac{1}{\alpha_d})$, where $\Gamma(\cdot)$ is the Gamma function. The exponential distribution is actually a special case of the Weibull distribution with $\alpha_d = 1$, while the Weibull distribution is heavy-tailed if $0 < \alpha_d < 1$. The smaller the α_d value, the heavier the tail that occurs in a given Weibull distribution. To assess the degree of heavy-tailedness, *Weibull factor* is introduced in [74], which is defined as

$$W_{L_d} = x\frac{d}{dx}\Big[\ln(-\ln(1 - F_{L_d}(x)))\Big] = \frac{\frac{d}{dx}[\ln(-\ln(1 - F_{L_d}(x)))]}{\frac{d}{dx}[\ln(-\ln(1 - F_{exp}(x)))]} \tag{2.9}$$

where $F_{L_d}(\cdot)$ is the CDF of L_d and $F_{exp}(\cdot)$ is the CDF of an exponential distribution. For a Weibull distribution, the Weibull factor actually equals the shape parameter α_d.

In the specification of 3GPP TS 30.03 [71] for the evaluation of UMTS systems, the data call size is modeled by a truncated Pareto distribution with a PDF

$$f_{L_d}(x) = \begin{cases} \dfrac{\gamma_d \cdot (l_d)^{\gamma_d}}{x^{\gamma_d + 1}}, & l_d \leq x < u_d \\ \displaystyle\int_{u_d}^{\infty} \frac{\gamma_d \cdot (l_d)^{\gamma_d}}{s^{\gamma_d + 1}}\, ds, & x = u_d \end{cases} \tag{2.10}$$

where γ_d $(1 < \gamma_d < 2)$ is the shape parameter, $[l_d, u_d]$ is the size range, and the mean data call size is given by

$$\mathrm{E}[L_d] \triangleq \overline{L}_d = \frac{\gamma_d \cdot l_d - u_d \cdot \left(\frac{l_d}{u_d}\right)^{\gamma_d}}{\gamma_d - 1}. \tag{2.11}$$

It is known that performance analysis tends to be extremely difficult with heavy-tailed distributions involved in the system model. For example, a Pareto distribution has an infinite mean if the shape parameter is less than 1 and an infinite variance if the shape parameter is less than 2. To render effective and tractable analysis, it is proposed in [77] to fit a large class of heavy-tailed distributions (including Pareto and Weibull distributions) with hyper-exponential distributions. Hyper-exponential distributions are a special class of phase-type distributions, which are a very general mixture of exponential distributions and have been used to approximate general distributions. In particular, for a distribution with a coefficient of variance (CV) larger than 1, a hyper-exponential distribution can be used since the CV of a hyper-exponential distribution is always larger than 1. As observed from real measurements, the data file size usually has a typical CV larger than 1, and thereby can be well approximated with a hyper-exponential distribution.

An important feature of heavy-tailedness is the so-called "*mice-elephants*" phenomenon [78]. With respect to the data call size, it implies that most data calls have a quite short length while a small fraction of data calls have an extremely large size. To reduce the number of parameters and render tractable analysis, the data call size can also be approximated by a two-stage hyper-exponential distribution [79, 80], whose PDF is defined as

$$f_{L_d}(x) = \frac{b}{b+1} \cdot \frac{1}{\frac{1}{b} \cdot \overline{L}_d} e^{-\frac{b}{\overline{L}_d} x} + \frac{1}{b+1} \cdot \frac{1}{b \cdot \overline{L}_d} e^{-\frac{1}{b} \cdot \frac{1}{\overline{L}_d} x}, \qquad b \geq 1, \quad x > 0 \qquad (2.12)$$

where the parameters b and $\overline{L}_d$ can be obtained by the first and second moments fitting. In particular, b can completely characterize the "*mice-elephants*" feature. A larger value of b corresponds to a data call size with a higher variability. Furthermore, since the hyper-exponential distribution consists of a linear mixture of exponentials, the analytical study involving (2.12) can be extended to higher-order hyper-exponential distributions with more exponential components, which can more accurately approach the original heavy-tailed distribution. Hence, hyper-exponential approximation can

not only provide analytical tractability but also well capture the essential properties of heavy-tailed distributions.

2.4.3 Streaming class

The streaming class includes many appealing services such as video streaming and becomes very popular in wireless networks [81]. A primary feature of the streaming class is that the content is played back at the receiver during the delivery. Instead of satisfying a low delay bound as for conversational services [82], streaming services need to maintain a continuous steady flow for smooth playback at the receiver. A playout buffer is introduced at the receiver and the playout starts after a pre-roll time (S_p). The playout buffer can counter against traffic burstiness and also absorb delay jitter resulting from network bandwidth variations. If a frame to play has not been completely delivered to the buffer at a fetch time, underflow occurs and the playback halts. Before the playback restarts, the receiver rebuffers for certain time S_b to accumulate enough data that can be played for a duration S_f [83]. We can see that two key QoS metrics are the pre-roll time and rebuffer time. A reasonable start-up pre-roll time should be less than 10 seconds as specified in [70]. To measure the playback smoothness, another QoS metric referred to as *underflow ratio* [84] is defined as

$$U_s = \frac{S_b}{T_s} \tag{2.13}$$

which is the ratio of the total rebuffer time over video clip duration T_s and should be bounded to guarantee acceptable service quality.

According to the statistics of streaming media stored on the Web [85], the median of video clip duration is about 2 minutes and very likely to be heavy-tailed. The video

clip duration can be modeled by a truncated Pareto distribution with a PDF

$$f_{T_s}(t) = \begin{cases} \dfrac{\gamma_s \cdot (l_s)^{\gamma_s}}{t^{\gamma_s+1}}, & l_s \le t < u_s \\ \displaystyle\int_{u_s}^{\infty} \frac{\gamma_s \cdot (l_s)^{\gamma_s}}{s^{\gamma_s+1}}\, ds, & t = u_s \end{cases} \tag{2.14}$$

where γ_s is the shape parameter and $[l_s, u_s]$ is the duration range.

As video clips capture moving images at fixed frame rates to display fluent motion, the video traffic is inherently long-range dependent and adjacent frames are highly correlated. Compression algorithms are necessary to remove the redundancy in order to store and transmit video sequences efficiently. In UMTS systems, H.263 is the mandatory codecs for packet-switched video streaming. Advanced video coding (AVC), MPEG-4 Part 10, also known as H.264, is recommended for higher-quality video [86]. Predicative coding is used in H.263 and MPEG-4 to remove temporal redundancy [87]. Intracoded frames (compressed versions of raw frames independent of other frames) are interleaved with forward/bidirectionally predicted frames (referring preceding and/or succeeding frames). The encoded frame size and in turn the encoding bit rate have large variability, which also depends on the video content such as texture details, scene change, and object motion speed.

In [88], the video source is modeled with a gamma-beta autoregressive (GBAR) process. Let L_n denote the frame size of the n^{th} frame, which is approximated by a stationary stochastic process with a marginal gamma distribution $G_a(\alpha_s, \eta_s)$ with PDF

$$f_L(x) = \frac{(x/\eta_s)^{\alpha_s-1}}{\alpha_s \Gamma(\alpha_s)} e^{-x/\eta_s}, \qquad \alpha_s,\ \eta_s > 0, \quad x > 0 \tag{2.15}$$

where α_s and η_s are the shape and scale parameters, respectively, the mean and variance are given by $\overline{L} = \alpha_s \eta_s$ and $\sigma_L^2 = \alpha_s \eta_s^2$, respectively. Then, L_n is modeled by a GBAR process with

$$L_n = BL_{n-1} + A \tag{2.16}$$

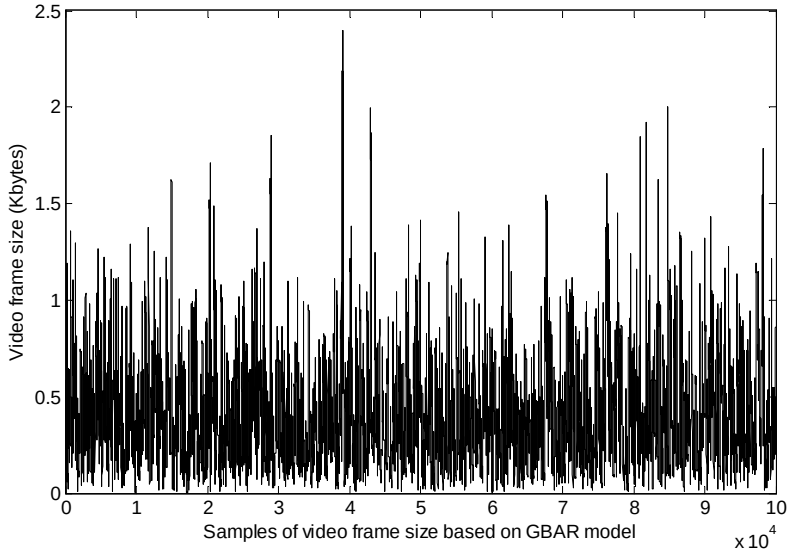


Figure 2.3: Video frame size based on GBAR model.

where A also follows a gamma marginal distribution $G_a(\alpha_s - \beta_s, \eta_s)$ and B has a beta marginal distribution $B_e(\beta_s, \alpha_s - \beta_s)$ with parameters β_s and $\alpha_s - \beta_s$. The PDF of a beta distribution with parameters p and q is given by

$$f(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}, \qquad p,\ q > 0, \quad 0 < x < 1. \tag{2.17}$$

The parameters α_s and η_s can be estimated from the mean and variance of collected statistics of frame sizes, while β_s can be obtained from the autocorrelation function of the frame sequence L_n, which is given by

$$r_L(k) = \left(\frac{\beta_s}{\alpha_s}\right)^k, \qquad k = 0,\ 1,\ 2, \cdots . \tag{2.18}$$

Figure 2.3 shows the sample frame sizes generated from the GBAR process with $\overline{L} = 400$ bytes, $\sigma_X = 272.8$ bytes, and $\frac{\beta_s}{\alpha_s} = 0.984$. It can be seen that the video traffic generated at a constant frame rate is very bursty due to highly variable frame sizes.

The traffic variation is similar to the alternate on-off in voice traffic, although there is no complete "silence" to sustain a continuous video stream.

2.5 Summary

The cellular/WLAN interworking issues such as integration architecture and vertical handoff management have been well explored in previous works. Nonetheless, not many research efforts have been devoted to resource allocation aspects, which are vital to efficiently utilize the overall resources of the integrated network for QoS provisioning. For this heterogeneous wireless overlay network, an essential resource allocation technique is the aforementioned call assignment with admission control and call reassignment via vertical handoff. The complementary strength and overlay structure can be exploited to properly share the incoming traffic load between the interworked systems. This technique has been extensively studied and used in hierarchical cellular networks with a similar overlay topology. However, it is still an open issue for cellular/WLAN interworking as the network heterogeneity introduces many new challenges. The unique characteristics of the cellular/WLAN integrated network should be carefully addressed in resource allocation. In particular, many previous works in this area neglect multi-service provisioning, which has become an essential requirement for wireless networks. Actually, multi-service support is a key motivation for cellular/WLAN interworking as the two networks present complementary strength in serving different services. This research will investigate the impact of multi-service provisioning on the interworking.

Chapter 3

System Model and Research Topics

In this study, we consider the interworking of a 3G cellular network and WLANs with physical and MAC specification similar to that of IEEE 802.11 standards. The cellular network provides ubiquitous coverage but a relatively small bandwidth. Fine-grained QoS is enabled for multiple services with a centralized infrastructure and CDMA-based multiple access. On the other hand, WLANs occupy a larger license-exempt frequency band and the channel access is based on a random access protocol. Although WLANs are originally designed for best-effort service, there has been extensive research on QoS enhancement to WLANs. It is reasonable to consider WLANs equipped with better QoS support such as effective admission control.

As observed in Chapter 2, most recent research on cellular/WLAN interworking focuses on issues such as integration architecture and vertical handoff management. Resource allocation is another essentially important aspect of the interworking. The heterogeneous overlay network necessitates effective mechanisms for call assignment/reassignment together with admission control and vertical handoff. Many previous works in this area do not address multi-service support or neglect the unique characteristics of the integrated network. The system model under consideration captures the multi-service traffic characteristics and location-dependent user mobility.

3.1 The Cellular/WLAN Integrated Network

Currently, most widely deployed WLANs employ simple contention-based protocols for medium access control. In the original IEEE 802.11 standard, a per-node queue with per-node backoff is used for channel contention and collision resolution. This per-node based principle penalizes heavy-loaded nodes with many flows (e.g., the access point). It is unfair and ineffective to support multimedia traffic with various QoS requirements. On the other hand, the MACAW [89] uses per-flow queue with per-flow backoff for channel contention. By this means, a node with multiple flows is viewed as multiple virtual nodes, each having one flow. A similar principle is adopted in IEEE 802.11e, where each node has multiple service queues for different access categories. Given the advantage of per-flow contention in multi-service support, we consider per-flow contention-based WLANs, which follow 802.11 DCF for contention and collision resolution. Actually, although contention-free channel access can provide hard QoS guarantee more than differentiated QoS, the interworking performance with contention-free WLANs may not be better due to inefficient polling schemes [90].

As WLANs operate at license-exempt frequency bands, a large bandwidth is available to support a high date rate, e.g., up to 11 Mbit/s in IEEE 802.11b. In contrast, current widely deployed 3G cellular networks support a relatively low data rate. For example, the UMTS system (Release 1999) can provide a data rate up to 2 Mbit/s for low-mobility applications (up to 10 km/hr) [91]. There are also some enhancement technologies such as the high speed packet access (HSPA), which can promote the downlink packet rate of UMTS access network up to 14 Mbit/s. However, these broadband wireless technologies are still not widely applied to the cellular networks in operation. Also, the deployment of microcells or picocells in hotspots is not so cost-effective as WLAN deployment. Hence, we focus on the interworking of WLANs and 3G cellular networks with a much smaller cell capacity.

As shown in Figure 3.1, the 3G cellular network provides ubiquitous connectivity

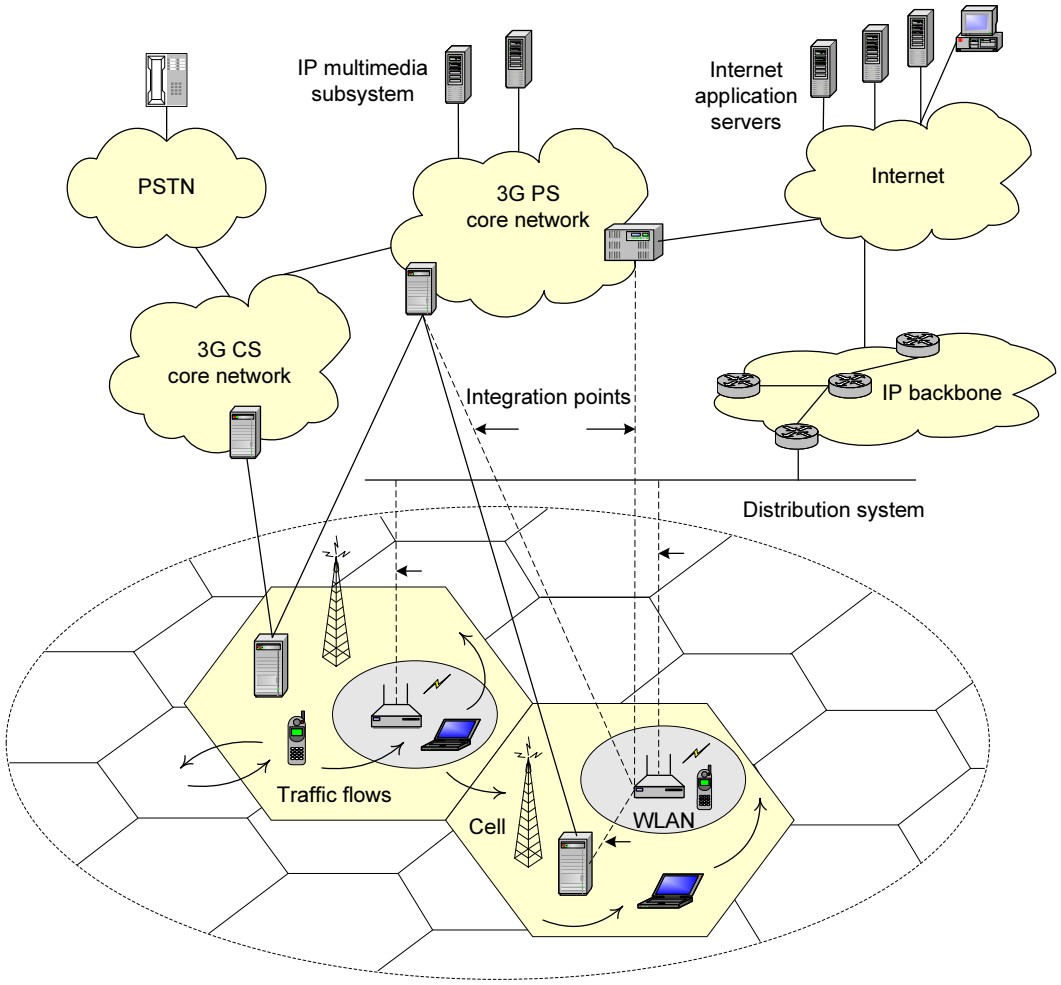


Figure 3.1: System model for a cellular/WLAN integration network.

over wide-area coverage, while WLANs are deployed disjointly in hotspot local areas. To begin with a simple topology, we consider in this study that there is one overlay WLAN in a cell and the WLAN is overlaid with one and only one cell. The cell and its overlay WLAN are referred to as a *cell/WLAN cluster*. The mobile devices are dual-mode and equipped with network interfaces to both the cellular network and WLAN [8]. Thus, both cellular access and WLAN access are available to dual-mode mobiles within the WLAN-covered areas, which are referred to as *double-coverage areas*. Since the two

networks operate at different frequency bands, the two network interfaces can be active simultaneously to assist vertical handoff [92]. In contrast, the areas with only cellular coverage are referred to as *cellular-only areas*.

To maximize the interworking gain, joint resource allocation can be applied to efficiently utilize the shared resources in the overlay network for QoS provisioning [31]. The complementary characteristics and dynamics of both networks should be considered so as to enhance their strength and compensate for the restriction. As discussed in Section 2.1, tight coupling is ideally suitable for cellular operators to integrate self-deployed WLANs. A central controller can access sufficient timely information of neighboring cells/WLANs, such as in-progress traffic load, link performance, and channel occupancy. Hence, joint resource allocation is intrinsically enabled with a tight coupling. A large interworking gain is achievable at the cost of a relatively high complexity.

On the other hand, loose coupling is very popular due to the low implementation complexity and deployment flexibility. In a loosely coupled cellular/WLAN network, basic network information can still be exchanged in between via effective control signaling mechanisms. The signaling messages should be appropriately designed and compressed to avoid an excessive load. Also, the relatively long latency across the two networks needs to be overcome to ensure timely update of network information. On the whole, joint resource allocation is also feasible for loose coupling. In view of the above observation, this study is not restricted to a specific tight-coupling or loose-coupling integration architecture.

3.2 Multi-Service Traffic Model

As discussed in Section 2.4, four service classes are defined for UMTS systems, namely, the conversational, streaming, interactive, and background classes. Both the conversational class and streaming class are meant for real-time services. Typically, some conversational-class services may require a constant bandwidth to satisfy very demand-

ing QoS constraints. The streaming class can accept a range of bandwidth adaptation, whereas the call duration is independent of the occupied bandwidth. A smaller bandwidth does not lead to a longer transmission time but quality degradation [93]. The interactive class and background class include traditional Internet applications such as Web browsing, file transfer, Telnet, and E-mail. As non-real time services, they are tolerant of elastic bandwidth. The call duration depends on the data size and occupied bandwidth. As the background class (e.g., E-mail) is of a best-effort service nature, we only consider the first three classes requiring certain QoS assurance.

In this study, voice service is considered as a representative service of the conversational class. Voice calls are assumed to arrive as a Poisson process, having an exponentially distributed duration T_v with mean in the order of several minutes.

For interactive data service, the session arrivals are also assumed to be Poisson, as the sessions are invoked independently by a large number of independent users. Each interactive session exhibits the on-off dynamics shown in Figure 2.2. We also take into account the heavy-tailedness of data call size. To investigate the impact of traffic variability on interworking performance, we use a two-stage hyper-exponential distribution given in (2.12). The "*mice-elephants*" feature of heavy-tailed distributions is captured in a simple way to render tractable analysis. Further, a Weibull distribution in (2.8) is used when the heavy-tailedness needs to be explored more accurately. The simulation study in Chapter 7 considers the traffic model specified for the evaluation of UMTS systems [71]. In particular, the data call size is characterized by a truncated Pareto distribution as given in (2.10).

For the streaming class, we consider video streaming service as a representative service, which is the most challenging in comparison with data or audio streaming. Depending on whether the video clips to be streamed are encoded on-line or pre-stored in the media server, there is live streaming or stored on-demand streaming [94]. In this study, we consider on-demand streaming with primarily unidirectional traffic flow,

which comprises a significant fraction of streaming traffic. Session information (such as video clip duration, total amount of data to stream, and bit variability) is known *a priori* and can be exploited for QoS provisioning. To cope with the large bandwidth variation due to heterogeneous accesses, encoding bit rate adaptation is incorporated to protect against buffer underflow and maintain steady streaming flow. Here, we model the video source with a GBAR process proposed in [88]. Given a constant frame rate (in frames/s), denoted by f_s, the frame sizes of a video clip are modeled with the GBAR process defined in (2.16). It properly captures the high variability and correlation of the frames in a video clip. The video clip duration is assumed to follow a truncated Pareto distribution given in (2.14), which has been validated by statistics of streaming media stored on the Web [85].

3.3 Location-Dependent User Mobility Model

It is known that most WLANs are deployed in indoor environments like cafés, offices, and airports. Users within these areas are mostly static or only maintain a pedestrian-level mobility. Thus, it becomes not reasonable to apply a homogeneous mobility model for mobiles within the coverage of a large cell. For example, when a user drives to the office, its mobility level may change from a high vehicular speed on the highway to being almost static in the office. To statistically characterize user mobility, user residence time should vary with the location, which is either within the cellular-only coverage or double coverage.

With a statistical equilibrium assumption, we focus on a single cell with an overlay WLAN, i.e., a cell/WLAN cluster. As shown in [6], the indoor deployment and low user mobility result in a heavy-tailed user residence time within a WLAN. To avoid the complexity of directly applying heavy-tailed distributions in performance analysis, the user residence time within a WLAN, denoted by T_r^w, is modeled with an approximate

hyper-exponential distribution [79], whose PDF is given by

$$f_{T_r^w}(t) = \frac{a}{a+1} \cdot \frac{1}{\frac{1}{a} \cdot \frac{1}{\eta^w}} e^{-a\eta^w t} + \frac{1}{a+1} \cdot \frac{1}{a \cdot \frac{1}{\eta^w}} e^{-\frac{\eta^w}{a} t}, \qquad a \geq 1, \quad t > 0 \tag{3.1}$$

where the mean and squared coefficient of variance are respectively

$$\mathrm{E}[T_r^w] = (\eta^w)^{-1}, \quad \frac{\mathrm{Var}[(T_r^w)]}{\mathrm{E}^2[T_r^w]} \triangleq C_{v,T_r^w}^2 = 2a + \frac{2}{a} - 3. \tag{3.2}$$

This model well captures the "*mice-elephants*" property of heavy-tailedness. A large fraction $\frac{a}{a+1}$ of the users stay within the WLAN for a mean time $\frac{1}{a} \cdot \frac{1}{\eta^w}$, while the other $\frac{1}{a+1}$ of the users have a mean residence time of $a \cdot \frac{1}{\eta^w}$. Increasing the parameter a results in T_r^w with higher variability.

On the other hand, the user residence time in the area of a cell with only cellular access, denoted by T_r^c, is assumed to be exponentially distributed with parameter η^c. Users moving out of the cellular-only area enter neighboring cells with a probability p^{cc} and enter the coverage of the overlay WLAN in the target cell with a probability $p^{cw} = 1 - p^{cc}$. Therefore, the residence time of users admitted in the cell follows more complicate phase-type distributions as shown in Figure 3.2. Let T_{r1}^c and T_{r2}^c denote the cell residence time of a call from the cellular-only area and that of a call from the double-coverage area, respectively. The moment generating functions (MGFs) of T_{r1}^c and T_{r2}^c are derived from Figure 3.2 as

$$\Phi_1(s) = \sum_{i=1}^{\infty} (p^{cw})^{i-1} p^{cc} \frac{\eta^c}{\eta^c - s} [\psi(s)]^{i-1} \tag{3.3}$$

$$\Phi_2(s) = \sum_{i=1}^{\infty} (p^{cw})^{i-1} p^{cc} [\psi(s)]^{i} \tag{3.4}$$

where $\psi(\cdot)$ is the MGF of $T_r^c + T_r^w$, given by

$$\begin{aligned} \psi(s) &= \mathrm{E}\left[e^{s(T_r^c + T_r^w)}\right] \\ &= \frac{\eta^c}{\eta^c - s} \left[\frac{a}{a+1} \cdot \frac{a\eta^w}{a\eta^w - s} + \frac{1}{a+1} \cdot \frac{\frac{1}{a}\eta^w}{\frac{1}{a}\eta^w - s}\right]. \end{aligned} \tag{3.5}$$

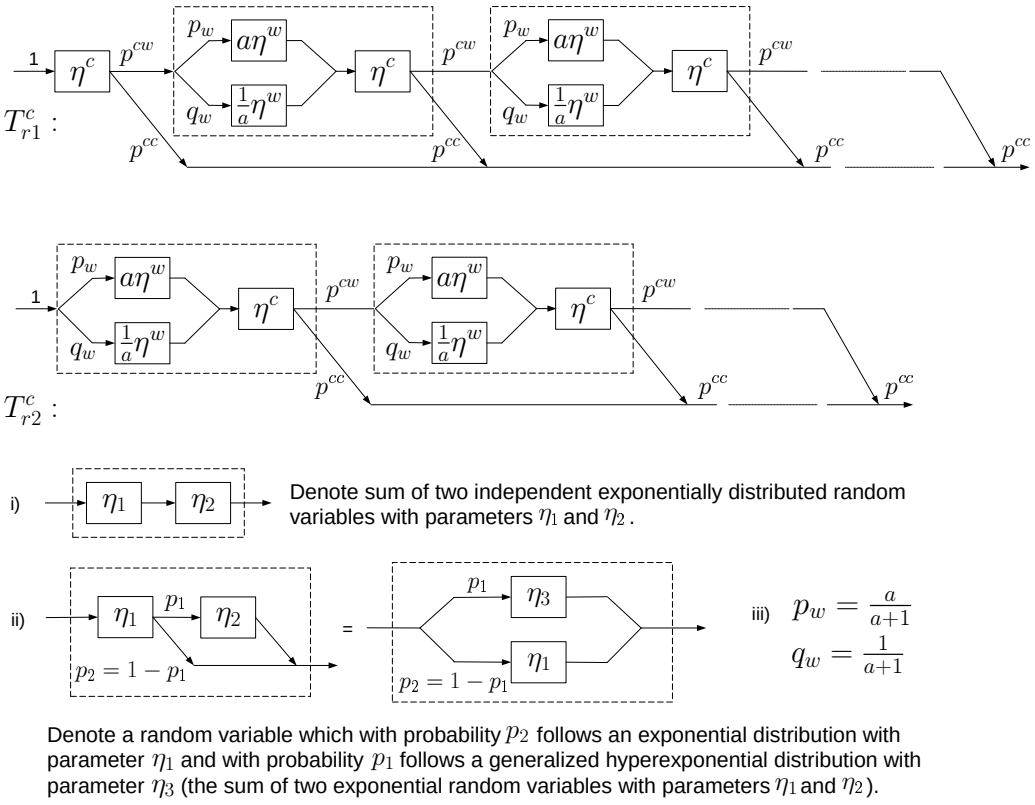


Figure 3.2: Modeling of user mobility within a cell/WLAN cluster.

3.4 Research Topics

This research is to contribute toward efficient resource allocation for a cellular/WLAN integrated network shown in Figure 3.1. In such a heterogeneous wireless overlay network, efficient resource allocation is vital for a high resource utilization and effective QoS provisioning to multiple services. Many challenges are introduced by the heterogeneous accesses, multi-service traffic, and location-dependent mobility. As discussed in Section 2.3, call assignment and admission control is an essential resource allocation technique to share the traffic load between interworked systems. As the cellular network and WLANs differ in capacity, mobility support, and QoS provisioning, call assignment

and admission decision can have a substantial impact on overall QoS satisfaction and resource utilization.

Most previous works in this area only address single service and cannot effectively exploit the interworking gain. As an essential requirement for wireless networks, multi-service support induces many challenging issues such as traffic characterization, service-specific QoS mechanisms, and resource sharing among multiple services. Further, this call assignment and admission control problem is complicated by the heterogeneous QoS provisioning capability of underlying networks. While fine-grained QoS is enabled in the cellular network with the centralized control and reservation-based resource allocation, QoS provisioning of WLANs is rather limited due to contention-based random access. Nevertheless, not enough research attention has been directed to exploiting the complementary QoS provisioning with multi-service traffic load. Finally, the most popular speed-sensitive assignment strategy for hierarchical cellular networks is not applicable to cellular/WLAN interworking. This is because the indoor deployment of most WLANs results in very low mobility in WLANs. Moreover, the user residence time within a WLAN becomes heavy-tailed and user mobility within a large cell is location-dependent, as shown in Figure 3.2. The unique mobility characteristics also affect call assignment/reassignment strategies and admission control policies. In more detail, this research will study the following topics:

- **Initial call assignment strategy:** Given the overlay structure in the cellular/WLAN integrated network, both cellular access and WLAN access are available in the double-coverage area. According to the call assignment strategy, an incoming call is initially assigned to its preferred cell/WLAN to request admission. From both the network and mobile user's perspectives, the initial call assignment should take into account various factors such as system capacity, service type, traffic characteristics, QoS requirements, user mobility and network bandwidth occupancy. Also, it is crucial to jointly consider the multi-service traffic load in

the assignment, so that the shared resources of the two networks are allocated as a whole to exploit the interworking gain. On the other hand, we also need to investigate distributed call assignment strategy to render feasible implementation in a loosely coupled network.

- **Call admission control policy:** Following initial call assignment, an incoming call is accepted or rejected according to the admission control policy of the target system (either a cell or a WLAN). The maximum admissible traffic load should be limited by admission control. Sufficient resources can then be reserved to satisfy QoS requirements of admitted traffic, such as bounded voice delay, mean data response time, and underflow ratio of video streaming. Also, the constraints on new call blocking probability and handoff call dropping probability need to be met to maximize resource utilization.

  With cellular/WLAN interworking, there are downward vertical handoff from the cell to the overlay WLAN and upward vertical handoff from the WLAN to the overlay cell. As seen in Figure 3.1, downward vertical handoff is optional since dual accesses are available in the double-coverage area. Hence, the admission control policy needs to properly differentiate new calls, horizontal handoff calls, and downward/upward vertical handoff calls. On the other hand, in contrast to the cellular-only area, the WLAN offers a resource backup for the double-coverage area to support a higher traffic density. To maintain fair access, it is also necessary to differentiate calls of different areas in the admission control policy.

  To ensure QoS satisfaction to admitted traffic, admission parameters should be determined properly based on the system capacity region, which is dependent on traffic characteristics, QoS requirements, and also underlying scheduling schemes. The system capacity can be improved by employing efficient scheduling schemes with affordable complexity. For example, the processor sharing (PS) and shortest remaining processing time (SRPT) disciplines can take a good advantage of the

data traffic elasticity and heavy-tailedness.

- **Call reassignment via vertical handoff:** In addition to initial call assignment, ongoing traffic load can also be dynamically transferred between the interworked systems by means of vertical handoff, for purposes such as QoS enhancement, congestion relief, load balancing, and so on. The call reassignment can be activated by user movement and performed at WLAN border crossing. Also, depending on network states, call reassignment can be enabled within the WLAN whenever sufficient spare capacity becomes available in the preferred network.

  As large bandwidth variations are possible with handover between heterogeneous accesses, it is also imperative to apply QoS adaptation corresponding to call reassignment. The adaptive QoS delivery is especially favorable to enhance service quality with available bandwidth. Nonetheless, a higher signaling overhead may be involved with call reassignment. Therefore, call reassignment is preferably considered for long-lived voice and video streaming calls, as the mean transfer delay of interactive data calls are bounded within a short duration of seconds.

3.5 Summary

In this chapter, our system model and research topics are outlined. We consider a heterogeneous wireless overlay network integrating a 3G CDMA cellular network and contention-based WLANs. Resource allocation is essentially important to efficiently utilize the overall resources in the integrated network. In particular, this research will explore the topics such as call assignment/reassignment and admission control for effective load sharing between the interworked systems. Multi-service QoS provisioning can be enhanced by exploiting the complementary strength of underlying networks. Also, the location-dependent user mobility should be properly addressed in the resource allocation. As discussed in Section 3.4, these research topics are closely related and should be investigated in a comprehensive way.

Chapter 4

Admission Control with Service-Differentiated Assignment

As discussed in Chapter 3, the cellular network and WLANs are complementary in terms of QoS support for different services. To maximize the overall resource utilization, it is necessary to differentiate the service type in assigning an incoming call to the overlay cell and WLAN. The target network decides whether to accept or reject the incoming call according to its admission control policy. The admission parameters should be properly determined so that a maximum traffic load can be admitted with QoS satisfaction.

4.1 Capacity Model

To ensure QoS satisfaction to admitted traffic, the traffic load in the cell or the WLAN should be properly limited within the corresponding capacity region. In this work, we consider both conversational voice service and interactive data service, which are provisioned different QoS support in the cellular network and WLANs.

CDMA-based cell capacity

Consider a CDMA cellular system with integrated voice and data services. Suppose voice traffic is delivered with dedicated channels (DCH), while data traffic can be transported over the downlink shared channels (DSCH). As data services such as Web browsing may lead to load asymmetry for the uplink and downlink, the capacity of the more congested downlink is analyzed in the following. The numbers of admitted users are limited to bound the interference level and satisfy user QoS requirements for the ratio of bit energy to noise and interference power spectral density, $\frac{E_b}{N_0}$. The downlink capacity in terms of the maximum numbers of simultaneously admitted voice and data users can be evaluated using a cell load factor [95], defined as

$$\eta_{DL} = \sum_{i=1}^{n_v^c} \frac{\rho + f_{DL}}{\frac{W_c}{\left(\frac{E_b}{N_0}\right)_v \alpha_v R_{b,v}^c} + \rho} + \sum_{i=1}^{n_d^c} \frac{\rho + f_{DL}}{\frac{W_c}{\left(\frac{E_b}{N_0}\right)_d R_{b,d}^c} + \rho} \tag{4.1}$$

where n_v^c and n_d^c are the number of voice and data users, respectively, ρ is the orthogonality factor, f_{DL} is the ratio between intercell interference and total intracell power measured at the user receiver, W_c is the total cell bandwidth, $R_{b,v}^c$ ($R_{b,d}^c$) is the bit rate of voice (data) users, α_v the activity factor of voice users, and $\left(\frac{E_b}{N_0}\right)_v$ and $\left(\frac{E_b}{N_0}\right)_d$ are the $\frac{E_b}{N_0}$ requirements of voice and data users, respectively. Then, the power limitation of the base station is equivalent to bounding the cell load factor by

$$\eta_{max} = 1 - \frac{P_p + P_N X_n}{P_{T,max}} \tag{4.2}$$

where

$$X_n = \sum_{i=1}^{n_v^c} \frac{L_{p,i}}{\frac{W_c}{\left(\frac{E_b}{N_0}\right)_v \alpha_v R_{b,v}^c} + \rho} + \sum_{i=1}^{n_d^c} \frac{L_{p,i}}{\frac{W_c}{\left(\frac{E_b}{N_0}\right)_d R_{b,d}^c} + \rho} \tag{4.3}$$

with $L_{p,i}$ being the path loss for the i^{th} user, P_p the power devoted to common control channels, P_N the background noise power, and $P_{T,max}$ the maximum transmitted power of the base station. From (4.1) - (4.3), we can derive the cell capacity region in terms

of (n_v^c, n_d^c) vectors, in which the $\frac{E_b}{N_0}$ requirements of voice and data users are satisfied with the limited transmission power of the base station.

WLAN capacity

Suppose there are n_v^w voice flows and n_d^w data flows admitted in a WLAN[1]. Data transmission follows the request to send (RTS)-clear to send (CTS)-DATA-ACK handshaking for channel access, while voice flows follow the basic access method due to the small payload size of voice packets.

Let λ_v^p denote the mean rate (frames/slot) of packet arrivals from a voice source. For data applications such as Web browsing, the data file to be transmitted is usually pre-stored in the application server. There is always traffic during the lifetime of a data call, referred to as *saturated case*. Also, the downlink data traffic can be regulated by applying rate control at the access point [96]. Let $\lambda_d^p(n_v^w, n_d^w)$ denote a data flow's packet input rate with n_v^w voice flows and n_d^w data flows admitted in the WLAN. For the saturated case, $\lambda_d^p(n_v^w, n_d^w)$ can be considered as equivalent to the achievable packet service rate for a data flow, denoted by $\xi_d^w(n_v^w, n_d^w)$. Similarly, the service rate for packets from a voice flow is denoted by $\xi_v^w(n_v^w, n_d^w)$. For simplicity, voice and data packet sizes are assumed to be constant.

It is observed in [97] that there exists an optimal operating point for the WLAN in the unsaturated case, beyond which the packet delay increases dramatically and the throughput drops quickly. When the packet service rate is larger than the arrival rate (network stability constraint) and the collision probability is small enough (e.g., less than 0.1), the service queue of a flow is almost empty and the packet delay is sufficiently small (say, less than 30 ms) to meet the requirement of real-time voice service. Therefore, we derive the WLAN capacity region in terms of the feasible set of

[1]Each voice call in the WLAN has two voice flows from and to the mobile, while each data call has a one-way data flow to the mobile.

(n_v^w, n_d^w) vectors to satisfy the stability constraints that

$$\xi_v^w(n_v^w, n_d^w) > \lambda_v^p \tag{4.4}$$
$$\xi_d^w(n_v^w, n_d^w) > \lambda_d^p(n_v^w, n_d^w)$$

where the packet service rates $\xi_v^w(\cdot)$ and $\xi_d^w(\cdot)$ can be obtained with the analytical approach given in Appendix A.

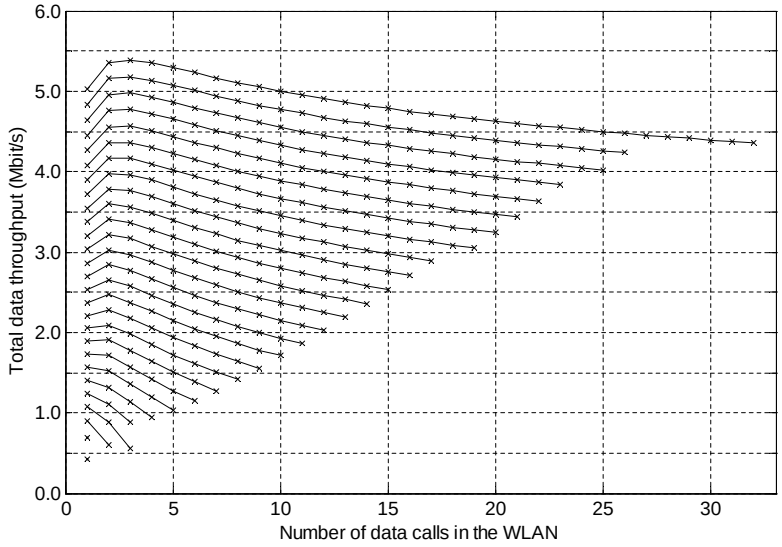


Figure 4.1: WLAN throughput with different numbers of voice and data calls.

It has been observed in many works that the WLAN capacity is dependent on the traffic characteristics and contention status. Following the WLAN parameters in Table A.1, Figure 4.1 shows the data throughput when different numbers of voice and data calls are accommodated in the WLAN. The curve on the top illustrates the case without voice calls, while the curves below are given by increasing the number of admitted voice calls one by one. It can be seen that the maximum achievable throughput when there is no voice call in service is around 5.4 Mbit/s over a 11 Mbit/s physical channel. When more data calls contend for access, the achievable throughput

is degraded due to a larger overhead from severer contention. As observed from the gap between adjacent curves, the data throughput is reduced by around 112 kbit/s to admit one more voice call, although the packet stream out of the voice codec only has a mean rate of 8 kbit/s in this example. As a result, the numbers of calls admitted in the WLAN are rather limited to maintain a small collision probability and satisfy the delay requirement of voice service. Meanwhile, a high throughput is achievable for each data call.

4.2 Admission Scheme with Service-Differentiated Assignment

For a new call originating within the double-coverage area, a decision needs to be made on whether to assign the incoming call to the covering cell or the WLAN. If the admission request is rejected by the preferred network, it can overflow to the other network to request admission. Moreover, when a mobile moves from the cellular-only area to the double-coverage area, its associated call can be handed over to the WLAN for load balancing, QoS enhancement, or cost reduction. If there is not sufficient spare capacity in the WLAN to accommodate the handoff call, the call can stay in the cellular network. Due to heterogeneous underlying technologies, the admission choice can have a significant impact on overall resource utilization and QoS satisfaction. A higher utilization and better QoS assurance becomes achievable, if the resources of both networks are allocated by jointly considering factors such as network capacity, traffic characteristics, user mobility, and QoS support capability.

4.2.1 Assignment strategy with service differentiation

It is known that the evolution of cellular networks has been motivated by voice telephony service, which is very mature and still dominates the operators' revenue. The

centralized infrastructure enables dedicated resource allocation, which can provide hard QoS guarantee to services requiring constant bandwidth. The large cell size and ubiquitous cellular coverage can reduce handoff frequency and in turn the impact of handoff latency on delay-sensitive real-time traffic. In contrast, as WLANs suffer from a large overhead for contention and collision resolution, only a very limited number of voice calls can be admitted. The achievable throughput may also be severely jeopardized to support the real-time voice calls. Further, the small and disjoint coverage of WLANs has an adverse effect on voice calls as frequent vertical handoff may be involved.

On the other hand, interactive data calls can accept elastic bandwidth. A larger bandwidth leads to a faster departure from the system. The large WLAN bandwidth can be efficiently utilized by elastic data traffic to improve the multiplexing gain. A data call is very likely to complete within the WLAN and does not need to hand over to the cell when the mobile moves out of the WLAN coverage. As such, the data traffic load is effectively relieved from the cell. In view of the above consideration, we develop the following call assignment/reassignment strategies with service differentiation.

- In the cellular-only area, both new and handoff calls can only request the cell for admission. The cell accepts or rejects an incoming call according to its admission control policy.

- A new voice call originating in the double-coverage area first attempts to get admission to the cell. If rejected by the cell according to its admission control policy, the call overflows to the WLAN to request admission. Only when rejected by both the cell and the WLAN will the call leave the system. A similar strategy is applied to new data calls in the double-coverage area except that data calls first try the WLAN for admission.

- When a mobile moves from the cellular-only area into the WLAN coverage, its associated voice calls are not handed over from the cell to the WLAN. No reas-

signment strategy is applied to avoid QoS degradation induced by vertical handoff and the inefficient real-time service support of the WLAN. In contrast, ongoing data calls served by the cell will attempt to hand over to the WLAN. A handoff data call is admitted if there is sufficient spare capacity in the WLAN to accommodate the data call without QoS violation to existing traffic. Otherwise, the data call remains served by the cell.

4.2.2 Admission control policies

An important aspect of multi-service support is to properly share the total bandwidth among different services. For contention-based WLANs, the resources are actually shared in a complete sharing (CS) manner by multiple services. The CS resource sharing penalizes services with a larger bandwidth requirement and privileges services requiring a smaller bandwidth and those with aggressive traffic [98]. Admission control can provide certain QoS protection by restricting the bandwidth occupancy of each service, which actually corresponds to partitioning the shared resources in a sense. This conservative policy may lower the resource utilization.

In the cellular network, the centralized control enables reservation-based resource sharing, which can compensate for the limitation of WLANs in service differentiation. It is known that voice calls have stringent delay requirement, while interactive data traffic is much more delay-tolerant and accepts elastic bandwidth. A restricted access mechanism [99] can be used to share the cell bandwidth between voice and data services. Voice service is offered preemptive priority over data service and only occupies up to a minimum amount of bandwidth to meet its QoS requirements. As data traffic can adapt to elastic bandwidth, all the bandwidth unused by voice traffic is shared equally by active data calls. By this means, the restricted access mechanism achieves a highest utilization in comparison with complete sharing and complete partitioning (CP) [100]. Also, one service is offered certain QoS protection against traffic overload of the other.

On the other hand, handoff calls should be prioritized over new calls as handoff dropping is more undesirable than new call blocking. Moreover, calls from different areas of the overlay network should be differentiated in accessing the available bandwidth. In the double-coverage area, a rejected call can overflow to the other network to request admission, whereas a call in the cellular-only area is cleared from the system if rejected by the cell. Thus, we consider a limited fractional guard channel policy shown in Figure 4.2 to provide new and handoff traffic in the cellular-only area a priority over new traffic in the double-coverage area. Because call blocking and dropping probabilities are very sensitive to the amount of reserved bandwidth, the guard bandwidth for higher-priority traffic is here fractional other than an integer number of guard channels. The voice admission region of the cell is given by $(N_v^c, G_{v1}^c, G_{v2}^c)$, where N_v^c is the maximum number of voice calls allowed in the cell, G_{v2}^c $(\leq N_v^c)$ is a real number indicating the guard bandwidth reserved for voice calls in the cellular-only area and inaccessible to new voice calls in the double-coverage area, while G_{v1}^c $(\leq G_{v2}^c)$ indicates the guard bandwidth reserved only for handoff voice traffic in the cellular-only area. For instance, a new voice call in the cellular-only area is accepted if the number of voice calls in a cell (k_v^c) satisfies $k_v^c \leq \lfloor N_v^c - G_{v1}^c \rfloor - 1$, or with a probability $1 - (G_{v1}^c - \lfloor G_{v1}^c \rfloor)$ when $k_v^c = \lfloor N_v^c - G_v^c \rfloor$. Correspondingly, data admission region of the cell is given by $(N_d^c, G_{d1}^c, G_{d2}^c)$, so that data traffic is also prioritized according to mobile location and new/handoff call differentiation.

In Section 4.1, we derive the WLAN capacity region in terms of the maximum numbers of voice and data calls that can be simultaneously accommodated in the WLAN. Accordingly, the data packet service rate $\xi_d^w(n_v^w, n_d^w)$ (in bit/s) is obtained for each vector (n_v^w, n_d^w) in the capacity region. It is observed that $\xi_d^w(n_v^w, n_d^w)$ decreases dramatically with a larger n_v^w, which implies the inefficient voice support of WLANs. To avoid the inefficiency region, a best operating point can be selected to limit the maximum numbers of voice and data calls admitted in the WLAN by N_v^w and N_d^w,

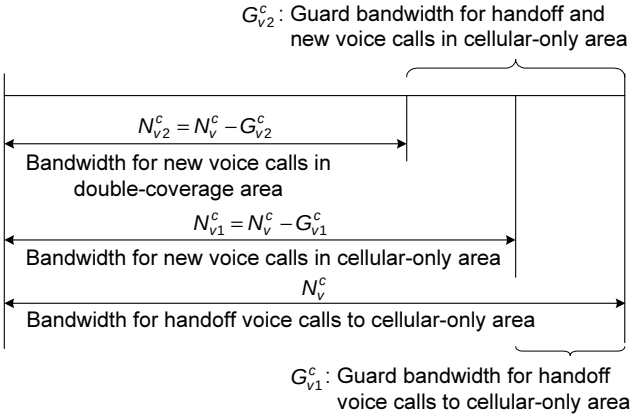


Figure 4.2: Limited fractional guard channel policy for voice calls in the cell.

respectively. As vertical handoff from the cell to the WLAN is optional, we introduce two other admission parameters G_v^w and G_d^w to denote the WLAN bandwidth reserved for new traffic. For this admission scheme under study, $G_v^w = N_v^w$ and $G_d^w = 0$, as no voice calls are handed over to the WLAN while the handoff proceeds for a data call if there is spare capacity in the WLAN.

Given different QoS support and resource sharing policies in the underlying networks, the admission regions of the cell and the WLAN should be configured properly to satisfy the corresponding QoS requirements. In this work, call blocking and dropping probabilities are required to be bounded by Q_{PB} and Q_{PD}, respectively. Also, the mean response time of data calls is constrained to be not greater than Q_T to guarantee the interactive service. Let B_{v1}^c (B_{v2}^c) denote the voice call blocking probability of the cell in the cellular-only (double-coverage) area, B_{d1}^c (B_{d2}^c) the data call blocking probability of the cell in the cellular-only (double-coverage) area, B_v^w (B_d^w) the voice (data) call blocking probability of the WLAN, D_v^c (D_d^c) the voice (data) call dropping probability of the cell, and $\overline{T}_d^c$ ($\overline{T}_d^w$) the mean response time of data calls carried by the cell (WLAN). Due to user mobility and the overlay structure, user QoS experience is jointly dependent on the servicing cell and WLAN. The admission regions of the cell

and the WLAN should be determined subject to the following constraints:

$$B_v^w \cdot B_{v2}^c \le Q_{PB}, \quad B_{v1}^c \le Q_{PB}, \quad D_v^c \le Q_{PD} \tag{4.5}$$
$$B_d^w \cdot B_{d2}^c \le Q_{PB}, \quad B_{d1}^c \le Q_{PB}, \quad D_d^c \le Q_{PD}, \quad \overline{T}_d^c \le Q_T, \quad \overline{T}_d^w \le Q_T.$$

It can be seen in Section 4.4 that the admission regions can significantly affect the overall resource utilization.

4.3 Performance Analysis of the Proposed Scheme

To properly determine the admission regions, i.e., (N_v^w, N_d^w) for the WLAN, $(N_v^c, G_{v1}^c, G_{v2}^c)$ and $(N_d^c, G_{d1}^c, G_{d2}^c)$ for the cell, we propose a search algorithm given in Table 4.1. First, we derive the cell and WLAN capacity regions based on the analysis in Section 4.1 and Appendix A, respectively. Then, we determine the admission parameters for voice calls $(N_v^w, N_v^c, G_{v1}^c, G_{v2}^c)$ to satisfy the QoS constraints that $B_v^w B_{v2}^c \le Q_{PB}$, $B_{v1}^c \le Q_{PB}$, and $D_v^c \le Q_{PD}$. Next, the admission parameters for data calls $(N_d^w, N_d^c, G_{d1}^c, G_{d2}^c)$ are obtained to maximize the acceptable data traffic load λ_d, so that $B_d^w B_{d2}^c \le Q_{PB}$, $B_{d1}^c \le Q_{PD}$, $D_d^c \le Q_{PD}$, $\overline{T}_d^c \le Q_T$, and $\overline{T}_d^w \le Q_T$. As seen from steps 6 and 11, the QoS metrics in terms of call blocking/dropping probabilities and mean data response time need to be evaluated effectively in each search round. The evaluation is actually quite complex because multiple dimensions are involved with the coupling between the cell and the WLAN, resource sharing between voice and data services, and differentiation of new and handoff traffic in different areas. The following approach employs proper decomposition and statistical averaging techniques to simplify the analysis. It is also applicable to analyze other call assignment strategies such as the randomized strategy discussed in Chapter 5 and the *WLAN-first* scheme in [16], in which voice and data calls always first try to get admitted to the available WLAN.

Table 4.1: Search algorithm for admission regions.

1: Derive the cell capacity region in terms of vectors (n_v^c, n_d^c) to satisfy the $\frac{E_b}{N_0}$ requirements as shown in Section 4.1
2: Derive the WLAN capacity region in terms of vectors (n_v^w, n_d^w) to meet the stability constraints as shown in Appendix A
3: $N_{v,max}^w = \max(n_v^w)$: $(n_v^w, n_d^w) \in$ WLAN capacity region
4: $N_{v,max}^c = \max(n_v^c)$: $(n_v^c, n_d^c) \in$ cell capacity region
5: **for** $N_v^w = 0, ..., N_{v,max}^w$ **do** // Evaluation for voice traffic.
6: By bisection search, determine minimum N_v^c $(\leq N_{v,max}^c)$ and (G_{v1}^c, G_{v2}^c) so that $B_v^w \cdot B_{v2}^c \leq Q_{PB}$, $B_{v1}^c \leq Q_{PB}$, and $D_v^c \leq Q_{PD}$
7: $N_{d,max}^w = \max(n_d^w)$ with $n_v^w = N_v^w$
8: **for** $N_d^w = 0, ..., N_{d,max}^w$ **do** // Evaluation for data traffic.
9: Initialize $\lambda_{d,min}$ and $\lambda_{d,max}$
10: $\lambda_d \leftarrow (\lambda_{d,min} + \lambda_{d,max})/2$ // Denote the mean arrival rate of data calls by λ_d.
11: By bisection search, determine $(N_d^c, G_{d1}^c, G_{d2}^c)$ and the acceptable mean data call arrival rate λ_d which satisfy $B_d^w \cdot B_{d2}^c \leq Q_{PB}$, $B_{d1}^c \leq Q_{PD}$, $D_d^c \leq Q_{PD}$, $\overline{T}_d^c \leq Q_T$, and $\overline{T}_d^w \leq Q_T$
12: **if** Solutions for $(N_d^c, G_{d1}^c, G_{d2}^c)$ exist **then** // The given traffic load λ_d is acceptable.
13: $\lambda_{d,min} \leftarrow \lambda_d$; $\lambda_d \leftarrow (\lambda_{d,min} + \lambda_{d,max})/2$
14: **else**
15: $\lambda_{d,max} \leftarrow \lambda_d$; $\lambda_d \leftarrow (\lambda_{d,min} + \lambda_{d,max})/2$
16: **end if**
17: **if** The acceptable λ_d converges **then**
18: Exit loop
19: **end if**
20: Record the maximum acceptable λ_d
21: **end for**
22: **end for**
23: Output (N_v^w, N_d^w), $(N_v^c, G_{v1}^c, G_{v2}^c)$, and $(N_d^c, G_{d1}^c, G_{d2}^c)$ which maximize the acceptable λ_d with QoS satisfaction

4.3.1 QoS evaluation for voice service

First, we evaluate the QoS metrics of voice traffic in the WLAN. Let λ_{v1} and λ_{v2} denote the mean arrival rate of new voice calls in the cellular-only area and the double-coverage area, respectively. Then, the voice traffic load offered to the WLAN comprises new calls

from the double-coverage area with a mean rate λ_{nv}^{w} and handoff calls from the overlay cell with a mean rate λ_{hv}^{cw}. According to our call assignment strategy, $\lambda_{nv}^{w} = \lambda_{v2}\, B_{v2}^{c}$. The channel holding time of voice calls in the WLAN is $T_v^w = \min(T_v, T_r^w)$, whose PDF can be derived from (3.1) as

$$f_{T_r^w}(t) = \frac{a}{a+1} E_x(a\eta^w + \mu_v) + \frac{1}{a+1} E_x(\eta^w/a + \mu_v) \tag{4.6}$$

where $\mu_v = \mathrm{E}^{-1}[T_v]$ and $E_x(\cdot)$ is the PDF function of an exponential distribution, defined as

$$E_x(\lambda) = \lambda e^{-\lambda t}, \qquad \lambda > 0, \quad t > 0. \tag{4.7}$$

The mean channel holding time is then

$$\mathrm{E}[T_v^w] = \frac{a}{a+1} A_I(a\eta^w, \mu_v) + \frac{1}{a+1} A_I(\eta^w/a, \mu_v) \triangleq (\mu_v^w)^{-1} \tag{4.8}$$

where $A_I(\cdot)$ is defined as

$$A_I(\nu_1, \nu_2) = \frac{1}{\nu_1 + \nu_2}, \qquad \nu_1 > 0, \quad \nu_2 > 0. \tag{4.9}$$

For tractability, both new and handoff call arrivals to the WLAN are assumed to be Poisson. Then, voice calls in the WLAN can be modeled by an $M/G/K/K$ queueing system. As the steady-state probabilities of an $M/G/K/K$ queue are insensitive to the service time distribution, the probability of having k_v^w voice calls in the WLAN is obtained as

$$\pi_v^w(k_v^w) = \pi_v^w(0) \prod_{i=1}^{k_v^w} \frac{\lambda_v^w(i)}{i \cdot \mu_v^w}, \qquad k_v^w = 1, ..., N_v^w \tag{4.10}$$

where $\lambda_v^w(i) = \lambda_{nv}^w + \lambda_{hv}^{cw}$ if $i \le N_v^w - G_v^w$, and $\lambda_v^w(i) = \lambda_{nv}^w$ when $N_v^w - G_v^w + 1 \le i \le N_v^w$. The voice call blocking probability of the WLAN is then $B_v^w = \pi_v^w(N_v^w)$, and the rejection probability for handoff voice calls from the cell is $D_v^w = \sum_{i=N_v^w - G_v^w}^{N_v^w} \pi_v^w(i)$.

Due to the limited fractional guard channel policy and varying mobility within the cell, QoS evaluation for the cell is more complicate. We model voice calls in the cell with a two-dimensional Markov process. The state (k_{v1}^c, k_{v2}^c) denotes the numbers of

existing voice calls in the cellular-only area and the double-coverage area, respectively, where $0 \leq k_{v1}^c + k_{v2}^c \leq N_v^c$. The state transition rate diagram is shown in Figure 4.3. For presentation clarity, the diagram is divided into several areas and only example transitions are shown in each area between a tagged state and its neighboring states.

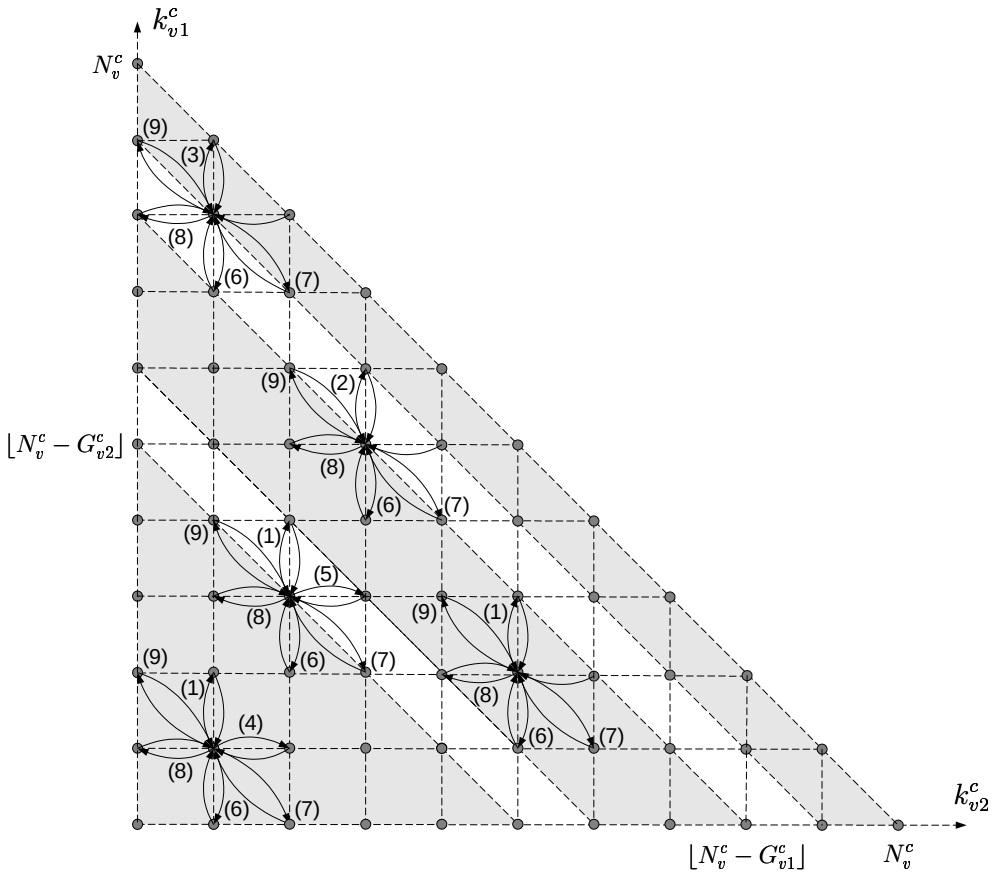
Under the Poisson assumption for call arrivals, the mean arrival rate of voice calls in the cellular-only area is $\lambda_{v1} + \lambda_{hv}^{cc} + \lambda_{hv}^{wc}$, where λ_{hv}^{cc} and λ_{hv}^{wc} denote the mean rates of handoff voice calls from neighboring cells and the overlay WLAN, respectively. Let λ_{nv2}^c denote the mean arrival rate of new voice calls in the double-coverage area. For the admission scheme under study, $\lambda_{nv2}^c = \lambda_{v2}$, i.e., all new voice calls in the double-coverage area first try the cell to get admitted. According to the limited fractional guard channel policy, the transition rate from (k_{v1}^c, k_{v2}^c) to $(k_{v1}^c + 1, k_{v2}^c)$ is given by

$$
\begin{aligned}
&(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c + 1, k_{v2}^c): \qquad (4.11)\\
&(1)\ \lambda_{v1} + \lambda_{hv}^{cc} + \lambda_{hv}^{wc},\ \text{if } k_{v1}^c + k_{v2}^c \leq \lfloor N_v^c - G_{v1}^c \rfloor - 1\\
&(2)\ [1 - (G_{v1}^c - \lfloor G_{v1}^c \rfloor)]\, \lambda_{v1} + \lambda_{hv}^{cc} + \lambda_{hv}^{wc},\ \text{if } k_{v1}^c + k_{v2}^c = \lfloor N_v^c - G_{v1}^c \rfloor\\
&(3)\ \lambda_{hv}^{cc} + \lambda_{hv}^{wc},\ \text{if } \lceil N_v^c - G_{v1}^c \rceil \leq k_{v1}^c + k_{v2}^c \leq N_v^c - 1.
\end{aligned}
$$

Similarly, the transition rate from (k_{v1}^c, k_{v2}^c) to $(k_{v1}^c, k_{v2}^c + 1)$ is given by

$$
\begin{aligned}
&(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c, k_{v2}^c + 1): \qquad (4.12)\\
&(4)\ \lambda_{nv2}^c,\ \text{if } k_{v1}^c + k_{v2}^c \leq \lfloor N_v^c - G_{v2}^c \rfloor - 1\\
&(5)\ [1 - (G_{v2}^c - \lfloor G_{v2}^c \rfloor)]\, \lambda_{nv2}^c,\ \text{if } k_{v1}^c + k_{v2}^c = \lfloor N_v^c - G_{v2}^c \rfloor.
\end{aligned}
$$

On the other hand, call completion or handoff also result in a state transition. For a voice call in the cellular-only area, it may depart from the cell due to call completion or handoff to neighboring cells. As both voice call duration and user residence time in the cellular-only area are exponentially distributed, the cell may transit from state (k_{v1}^c, k_{v2}^c) $(k_{v1}^c \geq 1)$ to $(k_{v1}^c - 1, k_{v2}^c)$ if a voice call ends, hands over to a neighboring



Transition rates from a tagged state (k_{v1}^c, k_{v2}^c) :

Figure 4.3: State transition rate diagram for voice calls in the cell.

cell, or gets admitted to the overlay WLAN[2]. Thus, the transition rate from (k_{v1}^c, k_{v2}^c)

[2]In the proposed admission scheme, there is no voice handoff to the WLAN. This case can be modeled by setting $G_v^w = N_v^w$, which results in a 100% rejection probability.

to $(k_{v1}^c - 1, k_{v2}^c)$ is approximated by

$$(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c - 1, k_{v2}^c): \qquad (6)\ k_{v1}^c \cdot [\mu_v + (p^{cc} + p^{cw} \cdot (1 - D_v^w))\eta^c],\ \text{if } k_{v1}^c \geq 1. \quad (4.13)$$

In contrast, the transition from (k_{v1}^c, k_{v2}^c) $(k_{v1}^c \geq 1)$ to $(k_{v1}^c - 1, k_{v2}^c + 1)$ is induced by user movement into the WLAN coverage. The transition rate is approximated by

$$(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c - 1, k_{v2}^c + 1): \qquad (7)\ k_{v1}^c \cdot p^{cw}\eta^c D_v^w,\ \text{if } k_{v1}^c \geq 1. \quad (4.14)$$

Moreover, a call departure in the double-coverage area may incur transitions from (k_{v1}^c, k_{v2}^c) $(k_{v2}^c \geq 1)$ to $(k_{v1}^c, k_{v2}^c - 1)$ or to $(k_{v1}^c + 1, k_{v2}^c - 1)$ with rates

$$\begin{aligned} &(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c, k_{v2}^c - 1): && (8)\ k_{v2}^c \cdot \mu_v,\ \text{if } k_{v2}^c \geq 1 \\ &(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c + 1, k_{v2}^c - 1): && (9)\ k_{v2}^c \cdot (\mu_v^w - \mu_v),\ \text{if } k_{v2}^c \geq 1 \end{aligned} \quad (4.15)$$

where μ_v^w is given by (4.8), which is the total departure rate of a voice call leaving the double-coverage area due to either call completion or user movement. While voice call completion results in the transition from (k_{v1}^c, k_{v2}^c) to $(k_{v1}^c, k_{v2}^c - 1)$ with rate $k_{v2}^c \cdot \mu_v$, the transition rate from (k_{v1}^c, k_{v2}^c) to $(k_{v1}^c + 1, k_{v2}^c - 1)$ is approximated by $k_{v2}^c \cdot (\mu_v^w - \mu_v)$.

By solving the balance equations of this two-dimensional Markov process, we can obtain the steady-state probability of (k_{v1}^c, k_{v2}^c), denoted by $p_v^c(k_{v1}^c, k_{v2}^c)$. Then, the probability of having k_v^c voice calls in the cell is given by

$$\pi_v^c(k_v^c) = \sum_{i=0}^{k_v^c} p_v^c(i, k_v^c - i), \qquad k_v^c = 0, 1, ..., N_v^c. \quad (4.16)$$

The voice handoff dropping probability is then $D_v^c = \pi_v^c(N_v^c)$. The blocking probabilities of the cell for new voice calls in the cellular-only area and the double-coverage area are respectively

$$B_{v1}^c = (G_{v1}^c - \lfloor G_{v1}^c \rfloor) \cdot \pi_v^c(\lfloor N_v^c - G_{v1}^c \rfloor) + \textstyle\sum_{i=\lceil N_v^c - G_{v1}^c \rceil}^{N_v^c} \pi_v^c(i) \quad (4.17)$$

$$B_{v2}^c = (G_{v2}^c - \lfloor G_{v2}^c \rfloor) \cdot \pi_v^c(\lfloor N_v^c - G_{v2}^c \rfloor) + \textstyle\sum_{i=\lceil N_v^c - G_{v2}^c \rceil}^{N_v^c} \pi_v^c(i). \quad (4.18)$$

As seen from the preceding analysis, the call blocking/dropping probabilities are directly dependent on handoff traffic load. With an equilibrium assumption for the system, each cell is statistically the same as any other one. Then, the mean rate of incoming handoff voice calls from neighboring cells λ_{hv}^{cc} is equal to that of outgoing handoff voice calls, which can be obtained by

$$\lambda_{hv}^{cc} = \sum_{i=0}^{N_v^c} \sum_{j=0}^{N_v^c - i} i \cdot p^{cc} \eta^c \cdot p_v^c(i,j). \tag{4.19}$$

Similarly, the mean rate of potential handoff voice calls into the WLAN coverage is given by $\lambda_{hv}^{cw} = \sum_{i=0}^{N_v^c} \sum_{j=0}^{N_v^c - i} i \cdot p^{cw} \eta^c \cdot p_v^c(i,j)$. On the other hand, a voice call admitted to the WLAN may complete within the WLAN, or it may need to hand over to the cell if it is not finished when the mobile moves out of the WLAN coverage. This handoff probability is derived as follows:

$$\begin{aligned} H_v^{wc} &= \Pr[T_v > T_r^w] = \int_0^\infty f_{T_r^w}(t)\ dt \int_t^\infty \mu_v e^{-\mu_v \tau} d\tau = \int_0^\infty f_{T_r^w}(t) e^{-\mu_v t}\ dt \\ &= \frac{a}{a+1} \cdot \frac{a\eta^w}{a\eta^w + \mu_v} + \frac{1}{a+1} \cdot \frac{\frac{1}{a}\eta^w}{\frac{1}{a}\eta^w + \mu_v}. \end{aligned} \tag{4.20}$$

Then, the mean arrival rate of handoff voice calls from the WLAN to the cell is $\lambda_{hv}^{wc} = [(1 - B_v^w)\lambda_{nv}^w + (1 - D_v^w)\lambda_{hv}^{cw}] \cdot H_v^{wc}$. With the inter-dependence between handoff traffic load and steady-state probabilities, the QoS metrics need to be evaluated recursively.

4.3.2 QoS evaluation for data service

For interactive data services such as Web browsing, the mean data response time is bounded within seconds to guarantee fluent interaction. In contrast to voice calls with a duration of minutes, data calls arrive and depart in a much smaller time scale. In an extreme case that there are no voice call arrivals or departures during a data call duration, the QoS evaluation for data service can be decomposed from voice service. In particular, this limiting behavior for a Markov process is referred to as *nearly complete*

decomposability [101]. Thus, we can first analyze the data performance conditioned on the number of voice calls in the system and then obtain the QoS approximation by averaging over the steady-state probabilities of voice calls.

Consider data service in the WLAN. Let (k_v^w, k_d^w) denote the current state of the WLAN with k_v^w voice calls and k_d^w data calls. Assume data call arrivals to the WLAN are a Poisson process with a mean rate λ_d^w. Then, the new and handoff data calls to the WLAN can be viewed as two virtual service classes with Poisson arrival rates $\frac{b}{b+1}\lambda_d^w$ and $\frac{1}{b+1}\lambda_d^w$, respectively, and exponentially distributed service requirements with mean $\frac{1}{b} \cdot \overline{L}_d$ and $b \cdot \overline{L}_d$, respectively [102]. From the WLAN capacity analysis, we can obtain the average data service rate $\xi_d^w(k_v^w, k_d^w)$ for each data call at state (k_v^w, k_d^w). As shown in [97], when the WLAN works in the proper operating range, the packet collision probability is quite small and each packet sees an approximately constant service rate. Hence, taking into account both call completion and handoff out of the WLAN, the departure rates of the two virtual data classes can be derived from (2.12) and (3.1) as

$$\mu_{d1}^w(k_v^w, k_d^w) = \left[\frac{a}{a+1}A_I(a\eta^w, b\nu_d^w(k_v^w, k_d^w)) + \frac{1}{a+1}A_I(\eta^w/a, b\nu_d^w(k_v^w, k_d^w))\right]^{-1} \tag{4.21}$$

$$\mu_{d2}^w(k_v^w, k_d^w) = \left[\frac{a}{a+1}A_I(a\eta^w, \nu_d^w(k_v^w, k_d^w)/b) + \frac{1}{a+1}A_I(\eta^w/a, \nu_d^w(k_v^w, k_d^w)/b)\right]^{-1} \tag{4.22}$$

where $\nu_d^w(k_v^w, k_d^w) = \frac{\xi_d^w(k_v^w, k_d^w)}{\overline{L}_d}$ and $A_I(\cdot)$ is defined in (4.9). The offered data traffic load at state (k_v^w, k_d^w) is then

$$\rho_d^w(k_v^w, k_d^w) = \frac{\frac{b}{b+1} \cdot \lambda_d^w(k_d^w)}{k_d^w \cdot \mu_{d1}^w(k_v^w, k_d^w)} + \frac{\frac{1}{b+1} \cdot \lambda_d^w(k_d^w)}{k_d^w \cdot \mu_{d2}^w(k_v^w, k_d^w)} \tag{4.23}$$

where $\lambda_d^w(k_d^w) = \lambda_{nd}^w + \lambda_{hd}^{cw}$ for $k_d^w \leq N_d^w - G_d^w$, and $\lambda_d^w(k_d^w) = \lambda_{nd}^w$ when $N_d^w - G_d^w + 1 \leq k_d^w \leq N_d^w$, with λ_{nd}^w and λ_{hd}^{cw} being the mean arrival rates of new and handoff data calls to the WLAN, respectively. For the call assignment strategy under study, $\lambda_{nd}^w = \lambda_{d2}$

and $G_d^w = 0$. Therefore, the probability of having k_d^w data calls in the WLAN is

$$\pi_d^w(k_d^w) = \sum_{i=0}^{N_v^w} \pi_v^w(i) \prod_{j=1}^{k_d^w} \pi_d^w(0)\rho_d^w(i,j), \qquad k_d^w = 0, 1, ..., N_d^w \tag{4.24}$$

where $\pi_v^w(\cdot)$ is the steady-state probabilities of voice calls in the WLAN given by (4.10). Thus, data call blocking probability of the WLAN is $B_d^w = \pi_d^w(N_d^w)$, while data call rejection probability is $D_d^w = \sum_{i=N_d^w - G_d^w}^{N_d^w} \pi_d^w(i)$. According to the Little's law, the mean response time of data calls carried by the WLAN can be obtained as

$$\overline{T}_d^w = \sum_{i=0}^{N_v^w} \pi_v^w(i) \sum_{j=1}^{N_d^w} \frac{j \cdot \prod_{l=1}^{j} \pi_d^w(0)\rho_d^w(i,l)}{\lambda_{nd}^w \cdot (1 - B_d^w) + \lambda_{hd}^{cw} \cdot (1 - D_d^w)}. \tag{4.25}$$

Next, data calls in the cell are modeled by a two-dimensional Markov process shown in Figure 4.4. Similar to the analysis for voice traffic, it captures the variability of user mobility in the cell and the limited fractional guard channel policy. Consider a tagged state (k_{d1}^c, k_{d2}^c) with k_{d1}^c and k_{d2}^c denoting the numbers of data calls in the cellular-only area and the double-coverage area of the cell, respectively. Due to call arrivals in the cellular-only area, the transition rate from (k_{d1}^c, k_{d2}^c) to $(k_{d1}^c + 1, k_{d2}^c)$ is

$$\begin{aligned}
&(k_{d1}^c, k_{d2}^c) \rightarrow (k_{d1}^c + 1, k_{d2}^c): \qquad (4.26)\\
&(1)\ \lambda_{d1} + \lambda_{hd}^{cc} + \lambda_{hd}^{wc},\ \text{if } k_{d1}^c + k_{d2}^c \leq \lfloor N_d^c - G_{d1}^c \rfloor - 1\\
&(2)\ [1 - (G_{d1}^c - \lfloor G_{d1}^c \rfloor)]\,\lambda_{d1} + \lambda_{hd}^{cc} + \lambda_{hd}^{wc},\ \text{if } k_{d1}^c + k_{d2}^c = \lfloor N_d^c - G_{d1}^c \rfloor\\
&(3)\ \lambda_{hd}^{cc} + \lambda_{hd}^{wc},\ \text{if } \lceil N_d^c - G_{d1}^c \rceil \leq k_{d1}^c + k_{d2}^c \leq N_d^c - 1
\end{aligned}$$

where λ_{hd}^{cc} and λ_{hd}^{wc} are the mean arrival rates of handoff data calls from neighboring cells and the overlay WLAN, respectively. The transition from (k_{d1}^c, k_{d2}^c) to $(k_{d1}^c, k_{d2}^c + 1)$ is incurred by call arrivals to the cell in the double-coverage area. Let λ_{nd2}^c denote the mean arrival rate of new data calls to the cell from the double-coverage area. For the admission scheme under study, $\lambda_{nd2}^c = \lambda_{d2}\, B_d^w$, as only data calls rejected by the WLAN

will overflow to the cell. Thus, the transition rate from (k_{d1}^c, k_{d2}^c) to $(k_{d1}^c, k_{d2}^c + 1)$ is

$$\begin{aligned}&(k_{d1}^c, k_{d2}^c) \rightarrow (k_{d1}^c, k_{d2}^c + 1): \qquad (4.27)\\ &(4)\ \lambda_{nd2}^c,\ \text{if } k_{d1}^c + k_{d2}^c \leq \lfloor N_d^c - G_{d2}^c \rfloor - 1\\ &(5)\ [1 - (G_{d2}^c - \lfloor G_{d2}^c \rfloor)]\, \lambda_{nd2}^c,\ \text{if } k_{d1}^c + k_{d2}^c = \lfloor N_d^c - G_{d2}^c \rfloor.\end{aligned}$$

The state transitions due to departure events are more complex with the hyper-exponentially distributed data call size and WLAN residence time. First, state (k_{d1}^c, k_{d2}^c) can transit to $(k_{d1}^c - 1, k_{d2}^c)$ when a data call completes within the cell, hands over to a neighboring cell, or hands over to the overlay WLAN with sufficient spare capacity to admit the call. Given that a data call finds sufficient spare capacity in the WLAN with a probability $(1 - D_d^w)$, the conditional user residence time in the cellular-only area, denoted by $\tilde{T}_r^c$, can be modeled by an exponential distribution with parameter $\tilde{\eta}^c = [p^{cc} + (1 - D_d^w)p^{cw}]\eta^c$. On the other hand, given k_v^c voice calls and k_d^c data calls carried by the cell, the time that a data call stays with the cell before it completes, denoted by $\tilde{T}_d^c(k_v^c, k_d^c)$, is approximately hyper-exponential with a PDF

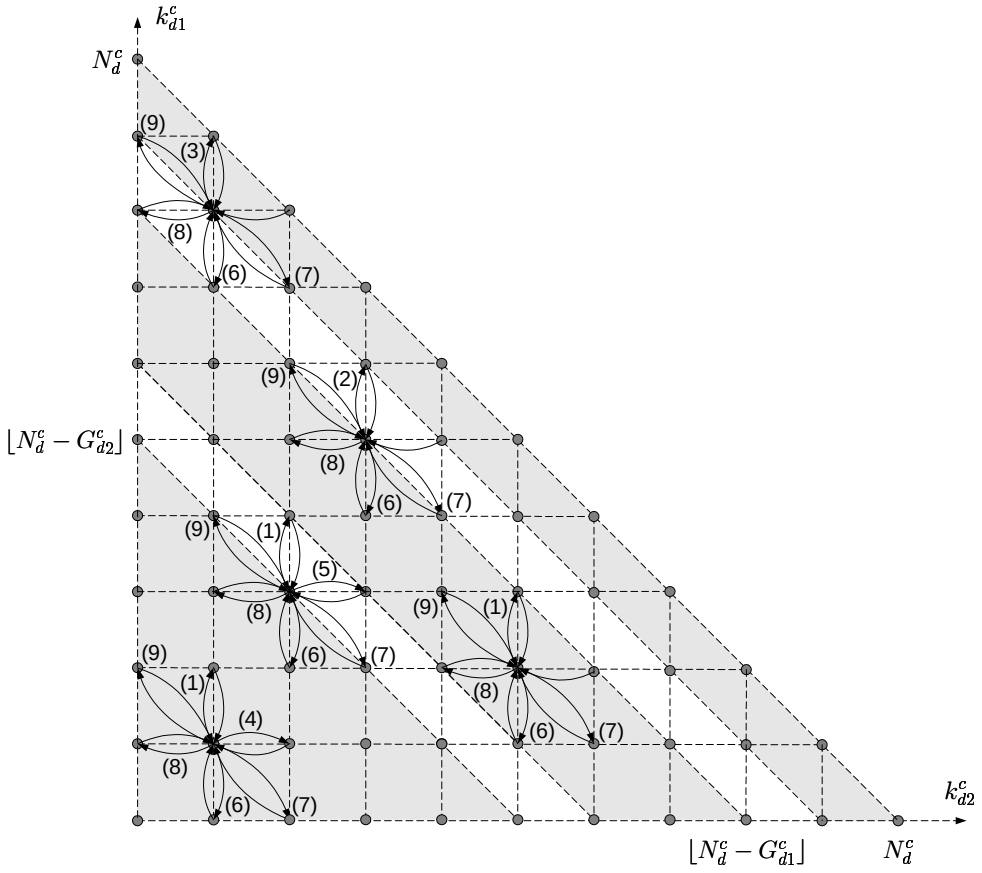
$$f_{\tilde{T}_d^c(k_v^c, k_d^c)}(t) = \frac{b}{b+1} E_x(b\nu_d^c(k_v^c, k_d^c)) + \frac{1}{b+1} E_x(\nu_d^c(k_v^c, k_d^c)/b) \qquad (4.28)$$

where $E_x(\cdot)$ is defined in (4.7) and $\nu_d^c(k_v^c, k_d^c) = \frac{R_{b,d}^c}{\overline{L}_d}$ with $R_{b,d}^c$ being the data service rate of the cell in (4.1). Given the state transition from (k_{d1}^c, k_{d2}^c) to $(k_{d1}^c - 1, k_{d2}^c)$, the conditional channel holding time of data calls in the cell, denoted by $T_{d1}^c(k_v^c, k_d^c)$, is $\min[\tilde{T}_r^c, \tilde{T}_d^c(k_v^c, k_d^c)]$ with mean

$$\mathrm{E}[T_{d1}^c(k_v^c, k_d^c)] = \frac{b}{b+1} A_I(\tilde{\eta}_c, b\nu_d^c(k_v^c, k_d^c)) + \frac{1}{b+1} A_I(\tilde{\eta}_c, \nu_d^c(k_v^c, k_d^c)/b) \triangleq [\phi_{d1}^c(k_v^c, k_d^c)]^{-1}. \qquad (4.29)$$

Thus, having k_v^c voice calls in the cell, the state transition rate from (k_{d1}^c, k_{d2}^c) to $(k_{d1}^c - 1, k_{d2}^c)$ is

$$(k_{d1}^c, k_{d2}^c) \rightarrow (k_{d1}^c - 1, k_{d2}^c): \qquad (6)\ k_{d1}^c \cdot \phi_{d1}^c(k_v^c, k_{d1}^c + k_{d2}^c),\ \text{if } k_{d1}^c \geq 1. \qquad (4.30)$$



Transition rates from a tagged state (k_{d1}^c, k_{d2}^c):

(1) $\lambda_{d1} + \lambda_{hd}^{cc} + \lambda_{hd}^{wc}$

(2) $[1 - (G_{d1}^c - \lfloor G_{d1}^c \rfloor)] \cdot \lambda_{d1} + \lambda_{hd}^{cc} + \lambda_{hd}^{wc}$

(3) $\lambda_{hd}^{cc} + \lambda_{hd}^{wc}$

(4) λ_{nd2}^c

(5) $[1 - (G_{d2}^c - \lfloor G_{d2}^c \rfloor)] \cdot \lambda_{nd2}^c$

(6) $k_{d1}^c \cdot \phi_{d1}^c(k_v^c, k_{d1}^c + k_{d2}^c)$

(7) $k_{d1}^c \cdot p^{cw} \eta^c D_d^w$

(8) $k_{d2}^c \cdot \mu_d^c(k_v^c, k_{d1}^c + k_{d2}^c)$

(9) $k_{d2}^c \cdot [\phi_{d2}^c(k_v^c, k_{d1}^c + k_{d2}^c) - \mu_d^c(k_v^c, k_{d1}^c + k_{d2}^c)]$

Figure 4.4: State transition rate diagram for data calls in the cell.

The transition from (k_{d1}^c, k_{d2}^c) to $(k_{d1}^c - 1, k_{d2}^c + 1)$ indicates that a data call attempts to hand over to the WLAN but rejected by the WLAN. So it remains served by the

cell but moves into the double-coverage area. Thus, the transition rate is

$$(k_{d1}^c, k_{d2}^c) \rightarrow (k_{d1}^c - 1, k_{d2}^c + 1): \qquad (7)\ k_{d1}^c \cdot p^{cw}\eta^c D_d^w, \text{ if } k_{d1}^c \geq 1. \tag{4.31}$$

Given data call completion within the double-coverage area, the cell state transits from (k_{d1}^c, k_{d2}^c) to $(k_{d1}^c, k_{d2}^c - 1)$ with a rate

$$(k_{d1}^c, k_{d2}^c) \rightarrow (k_{d1}^c, k_{d2}^c - 1): \qquad (8)\ k_{d2}^c \cdot \mu_d^c(k_v^c, k_{d1}^c + k_{d2}^c), \text{ if } k_{d2}^c \geq 1 \tag{4.32}$$

where $\mu_d^c(k_v^c, k_{d1}^c + k_{d2}^c) = \mathrm{E}^{-1}[\tilde{T}_d^c(k_v^c, k_{d1}^c + k_{d2}^c)]$. When the cell carries k_v^c voice calls and k_d^c data calls, for a data call in the double-coverage area and served by the cell, the channel holding time before it completes within this area or moves out of this area with unfinished service, denoted by $T_{d2}^c(k_v^c, k_d^c) = \min[T_r^w, \tilde{T}_d^c(k_v^c, k_d^c)]$, follows a hyper-exponential distribution with a PDF

$$\begin{aligned} f_{T_{d2}^c(k_v^c,k_d^c)}(t) = {} & \tfrac{a}{a+1}\tfrac{b}{b+1}E_x(a\eta^w + b\nu_d^c(k_v^c, k_d^c)) + \tfrac{a}{a+1}\tfrac{1}{b+1}E_x(a\eta^w + \nu_d^c(k_v^c, k_d^c)/b) \\ & + \tfrac{1}{a+1}\tfrac{b}{b+1}E_x(\eta^w/a + b\nu_d^c(k_v^c, k_d^c)) + \tfrac{1}{a+1}\tfrac{1}{b+1}E_x(\eta^w/a + \nu_d^c(k_v^c, k_d^c)/b). \end{aligned} \tag{4.33}$$

Then, the total transition rate from (k_{d1}^c, k_{d2}^c) to $(k_{d1}^c, k_{d2}^c - 1)$ or $(k_{d1}^c + 1, k_{d2}^c - 1)$ is $k_{d2}^c \cdot \phi_{d2}^c(k_v^c, k_{d1}^c + k_{d2}^c)$, where

$$\phi_{d2}^c(k_v^c, k_{d1}^c + k_{d2}^c) = \mathrm{E}^{-1}[T_{d2}^c(k_v^c, k_{d1}^c + k_{d2}^c)]. \tag{4.34}$$

To take advantage of the data traffic elasticity, data calls in the cell are served under the PS discipline, i.e., all active data calls equally share the bandwidth unused by ongoing voice calls. With admission control in place, the system is similar to an $M/G/1/K - PS$ queue, whose steady-state probabilities are insensitive to service requirement [103]. Although the time that the cell stays at state (k_{d1}^c, k_{d2}^c) before transiting to $(k_{d1}^c + 1, k_{d2}^c - 1)$ is not exponentially distributed, we can approximate the transition rate by

$$(k_{d1}^c, k_{d2}^c) \rightarrow (k_{d1}^c{+}1, k_{d2}^c{-}1): \qquad (9)\ k_{d2}^c \cdot [\phi_{d2}^c(k_v^c, k_{d1}^c{+}k_{d2}^c) - \mu_d^c(k_v^c, k_{d1}^c{+}k_{d2}^c)], \text{ if } k_{d2}^c \geq 1. \tag{4.35}$$

By numerically solving the balance equations of the Markov process in Figure 4.4, we can obtain the steady-state probability of (k_{d1}^c, k_{d2}^c) given k_v^c voice calls carried by the cell, denoted by $\tilde{p}_d^c(k_{d1}^c, k_{d2}^c|k_v^c)$. Then, the probability of having k_d^c data calls in the cell is given by

$$\pi_d^c(k_d^c) = \sum_{i=0}^{N_v^c} \pi_v^c(i) \sum_{j=0}^{k_d^c} \tilde{p}_d^c(j, k_d^c - j|i), \qquad k_d^c = 0, 1, ..., N_d^c \tag{4.36}$$

where $\pi_v^c(\cdot)$ is the steady-state probabilities of voice in the cell given by (4.16). The data handoff dropping probability is then $D_d^c = \pi_d^c(N_d^c)$. The blocking probabilities of the cell for new data calls in the cellular-only area and the double-coverage area (B_{d1}^c and B_{d2}^c, respectively) can be obtained in a way similar to that of (4.17) and (4.18). Moreover, the mean data response time can be obtained by the Little's law as

$$\overline{T}_d^c = \sum_{i=0}^{N_v^c} \pi_v^c(i) \sum_{j=0}^{N_d^c} \sum_{k=0}^{N_d^c - j} \frac{(j+k) \cdot \tilde{p}_d^c(j, k|i)}{(1 - B_{d1}^c)\lambda_{d1} + (1 - B_{d2}^c)\lambda_{nd2}^c + (1 - D_d^c)(\lambda_{hd}^{cc} + \lambda_{hd}^{wc})}. \tag{4.37}$$

Similar to the analysis for voice traffic, we can obtain the mean rates of handoff data calls at the equilibrium state. The mean arrival rate of handoff data calls out of the WLAN is given by

$$\lambda_{hd}^{wc} = \sum_{i=0}^{N_d^w} i \cdot \eta^w \cdot \pi_d^w(i). \tag{4.38}$$

The mean arrival rate of handoff data calls between neighboring cells and that from the cell to the overlay WLAN can be respectively obtained as

$$\begin{aligned} \lambda_{hd}^{cc} &= \sum_{i=0}^{N_v^c} \pi_v^c(i) \sum_{j=0}^{N_d^c} \sum_{k=0}^{N_d^c - j} j \cdot p^{cc} \eta^c \cdot \tilde{p}_d^c(j, k|i) \\ \lambda_{hd}^{cw} &= \sum_{i=0}^{N_v^c} \pi_v^c(i) \sum_{j=0}^{N_d^c} \sum_{k=0}^{N_d^c - j} j \cdot p^{cw} \eta^c \cdot \tilde{p}_d^c(j, k|i). \end{aligned} \tag{4.39}$$

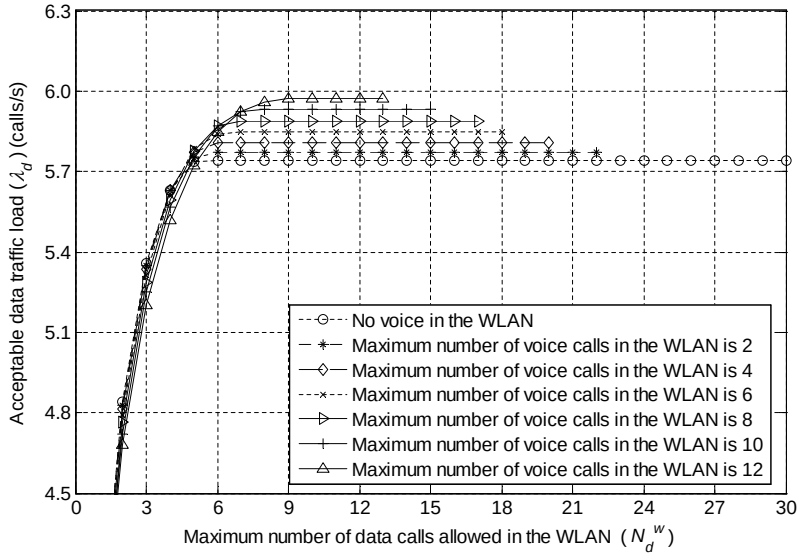
Again, the inter-dependence among handoff arrival rates and steady-state probabilities of data calls necessitates recursive computation to obtain the QoS metrics.

Table 4.2: System parameters.

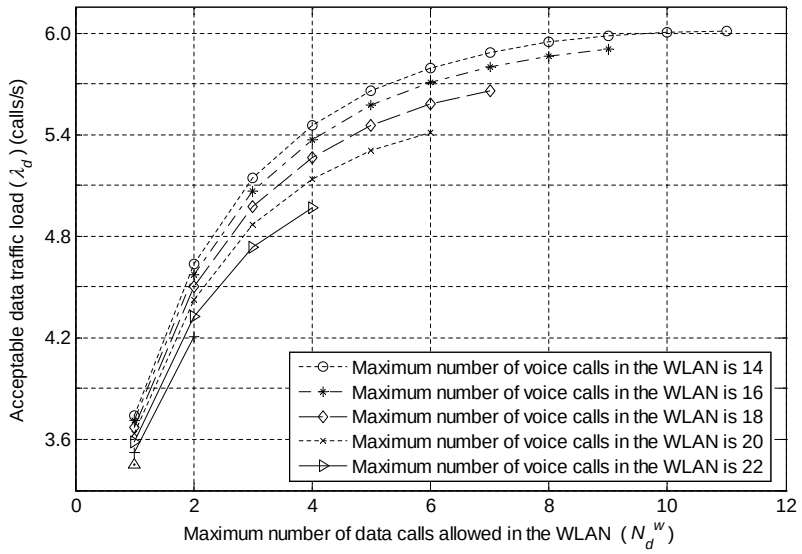
Parameter	**Value**	**Parameter**	**Value**
$(\eta^c)^{-1}$	10 min	$(\eta^w)^{-1}$	14 min
p^{cc}	0.76	p^{cw}	0.24
W_c	3.84 Mchips/s	C^w	11 Mbit/s
$(\mu_v)^{-1}$	140 s	$R^c_{b,v}$	12.2 kbit/s
λ_{v1}	0.12 calls/s	λ_{v2}	0.18 calls/s
Q_{PB}	0.01	Q_{PD}	0.001
Q_T	4.0 s	$\overline{L}_d$	64 kbytes
ρ	0.4	f_{DL}	0.55
α_v	0.43	$P_{T,max}$	43 dB
P_p	33 dB	P_N	-106 dB
$\left(\frac{E_b}{N_0}\right)_v$	4.60 dB	$\left(\frac{E_b}{N_0}\right)_d$	4.65 dB

4.4 Numerical Results and Discussion

In the section, we discuss some important observations obtained from the numerical results. Given in Table 4.2 are the system parameters. Figure 4.5 illustrates how the acceptable data traffic load (mean data call arrival rate $\lambda_d = \lambda_{d1} + \lambda_{d2}$) varies with the maximum number of data calls allowed in the WLAN (N_d^w) when the maximum number of voice calls allowed in the WLAN (N_v^w) is fixed to different values. It is observed from Figure 4.5 that the acceptable data traffic load increases with N_d^w when N_d^w is relatively small. This can be explained as follows. Generally, with the coupling between the cell and its overlay WLAN, both the time that a data call is carried by the cell and the WLAN contributes to the total transfer delay. However, as shown in the WLAN capacity analysis, only a limited number of data calls can be admitted to ensure a small collision probability and guarantee the delay requirement of voice traffic. As a result, a high throughput is offered to each data call and most data calls can finish within the WLAN. Therefore, when fewer data calls are allowed in



(a)



(b)

Figure 4.5: Acceptable data traffic load (λ_d) versus maximum number of data calls allowed in the WLAN (N_d^w) under QoS constraints.

the WLAN by choosing a smaller N_d^w, more data calls still need to be accommodated by the cell to bound the data call blocking and dropping probabilities. That is, the WLAN resources are not fully utilized to balance the data traffic load from the cell. Therefore, it is desirable to increase N_d^w and admit into the WLAN as many data calls as allowed by the WLAN capacity region. The mean transfer delay is still well bounded, because a high throughput is available for each admitted data call with the large WLAN bandwidth.

As illustrated in Figure 4.5, the increase of acceptable data traffic load with N_d^w becomes unnoticeable when N_d^w is large (say, more than 15). Indeed, with a larger N_d^w, more data traffic is assigned to the WLAN and balanced from the cell. Nonetheless, when the data traffic load is further increased, more guard bandwidth needs to be reserved for new and handoff data calls in the cellular-only area. Because the cell bandwidth is much smaller than the WLAN bandwidth, the cell becomes the bottleneck of the integrated system. Hence, the acceptable data traffic load cannot be continuously increased without QoS violation to data calls in the cellular-only area. The acceptable data traffic load is almost the same with large values of N_d^w.

For each curve in Figure 4.5, there is a maximum data traffic load acceptable with a certain value of N_d^w. From these curves, we can obtain Figure 4.6, which shows the relationship between the acceptable data traffic load and the maximum number of voice calls allowed in the WLAN (N_v^w). It is observed that there exists a value of N_v^w (i.e., 14 in the example) which maximizes the acceptable data traffic load. With this configuration, N_v^w is less than the WLAN capacity for voice service, which is 29 in this example. That is, voice traffic in the double-coverage area should be restricted not to occupy all the WLAN bandwidth. This is attributed to the cellular/WLAN coupling and voice/data resource sharing. First, since a larger value of N_v^w indicates that more voice traffic in the double-coverage area is assigned to the WLAN and relieved from the cell, more cell bandwidth can be used for data traffic in the cellular-only area,

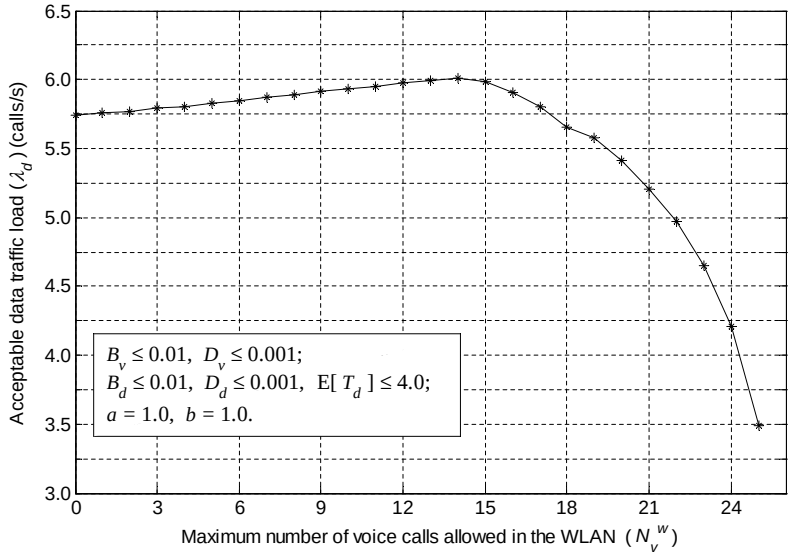


Figure 4.6: Acceptable data traffic load (λ_d) versus maximum number of voice calls allowed in the WLAN (N_v^w) under QoS constraints for voice call blocking/dropping probabilities (B_v and D_v, respectively), data call blocking/dropping probabilities (B_d and D_d, respectively), and mean data response time ($\mathrm{E}[T_d]$).

and the overall data transfer time is reduced (load balancing effect). This leads to a larger acceptable data traffic load. Second, when N_v^w is further increased to approach the WLAN capacity, the acceptable data traffic load decreases. When more voice calls are admitted to the WLAN, the number of data calls that can be simultaneously accommodated by the WLAN decreases and the data service rate drops. As a result, the maximum number of data calls allowed in the cell (N_d^c) needs to be increased so that the overall data call blocking and dropping probabilities meet the corresponding constraints. Due to the much smaller cell bandwidth, an increased traffic load assigned to the cell results in a longer data transfer time. When this penalty incurred by voice support in WLANs overwhelms the advantage of the load balancing effect, the acceptable data traffic load begins to decrease.

4.5 Summary

In this chapter, we have proposed and analyzed an admission control scheme with service-differentiated call assignment for cellular/WLAN interworking. Considering the complementary mobility and QoS support of the coupled networks, voice calls are preferably assigned to the cell, while data calls first try the available WLAN for admission. New and handoff calls in different areas are prioritized with limited fractional guard channel policies. To compensate for the limited QoS differentiation capability of WLANs, restricted access mechanism is applied in the cell for resource sharing between voice and data services. The PS service discipline further takes a good advantage of data traffic elasticity to improve utilization. The main contributions of this work are as follows:

- An analytical model based on two-dimensional Markov processes is developed to evaluate QoS metrics in terms of call blocking/dropping probabilities and mean data response time. The model properly captures the unique characteristics of an integrated cell/WLAN cluster, such as the location-dependent user mobility, highly variable data call size, and user residence time within the WLAN.

- The admission regions can be determined by applying the QoS evaluation in a search algorithm. It is observed from the numerical results that the overall resource utilization closely depends on the configuration of admission regions. In a best case, the maximum number of voice calls allowed in a WLAN is less than the WLAN capacity for voice service. That is, voice traffic in the double-coverage area should be restricted not to occupy all the WLAN bandwidth. Indeed, because data traffic is adaptive to elastic bandwidth, it can take a good advantage of the low mobility and large bandwidth in the double-coverage area.

Chapter 5

Randomized Assignment for Distributed Implementation

In Chapter 4, we introduce an admission scheme with service-differentiated call assignment for cellular/WLAN interworking. The admission parameters are determined in such a way to maximize the overall resource utilization. Actually, the rationale behind is to properly distribute the multi-service traffic load to the integrated cell and WLAN so as to effectively exploit their complementary strength. In this chapter, we further generalize the admission scheme with randomized assignment to enable distributed implementation. Moreover, a more effective analytical approach is developed for QoS evaluation by means of moment generating functions (MGFs). Based on the analytical model, we further investigate the impact of mobility and traffic variability on the determination of assignment parameters.

5.1 Decentralization with Randomized Assignment

With the heterogeneous QoS support of the underlying integrated network, voice and data traffic in the double-coverage area should be properly directed to the cell and the

WLAN. Due to the heterogeneity, especially when the two systems are loosely coupled, it is challenging for a central controller to timely obtain updated information of both systems (e.g., the numbers of ongoing calls in the cell and overlay WLANs) to make an optimal decision for each admission request. Also, there is a high control overhead since signaling messages have to traverse a long path involving many network elements. To reduce signaling overhead, frequent information exchanges may not be affordable; but outdated network information is adverse to decision accuracy. Consequently, in a loosely coupled cellular/WLAN network, distributed call assignment and admission control is more practical, although the decision may not be optimal in terms of maximizing resource utilization with QoS guarantee.

Instead of applying complex criteria for call assignment, we consider a simple admission scheme to investigate the dependence of resource utilization on load sharing and the impact of mobility and traffic characteristics. An incoming voice (data) call in the double-coverage area requests admission to the cell with a probability θ_v^c (θ_d^c), while it requests admission to the WLAN with a probability $\theta_v^w = 1 - \theta_v^c$ ($\theta_d^w = 1 - \theta_d^c$). The assignment parameters θ_v^c and θ_d^c (or θ_v^w and θ_d^w) are determined for a given traffic load and broadcast to the associated mobiles. Then, a mobile can make a decision on its own according to these parameters and send the admission request to the corresponding target network. With the simplicity, the proposed admission scheme can be implemented in a distributed manner. Also, as the network elements involved in the admission decision are limited to be as few as possible, the signaling overhead is reduced. The cellular network and the integrated WLANs only need to exchange information and update the above assignment parameters with traffic variation. It is especially suited for the cases that it is not affordable to base each admission decision on the system states of both networks. By controlling the assignment probabilities, the incoming traffic load is properly shared by the integrated cell and WLAN. Although the proposed scheme enables simple implementation, it may not fully exploit the performance gain achievable

from the interworking.

This assignment strategy can be extended and applied in combination with other call assignment criteria. For example, the cellular network also differs from WLANs in pricing rates. The service cost in the WLAN is generally much lower than that in the cellular network. A mobile user may prefer to get admitted to the WLAN for the low cost or to the cellular network if the service quality is more important. Suppose that an incoming call requests admission to the WLAN with a probability γ^w, while it requests admission to the cell with a probability $\gamma^c = 1 - \gamma^w$. Let $\omega_i,\ i = 1, 2, ..., r$, denote the relative weights of r different criteria and θ_i^w (θ_i^c) the probability of selecting the WLAN (cell) as the target when the i^{th} criterion is considered. Then, we have

$$\gamma^w = \sum_{i=1}^{r} \omega_i \cdot \theta_i^w, \qquad \gamma^c = \sum_{i=1}^{r} \omega_i \cdot \theta_i^c \tag{5.1}$$

where

$$\sum_{i=1}^{r} \omega_i = 1, \quad \theta_i^w = 1 - \theta_i^c, \qquad i = 1, 2, ..., r. \tag{5.2}$$

For example, when the overall utilization maximization and user preference are considered, the probabilities of selecting the WLAN and the cell as the admission target are respectively given by

$$\gamma^w = \omega_1 \cdot \theta_1^w + \omega_2 \cdot \theta_2^w, \qquad \gamma^c = \omega_1 \cdot \theta_1^c + \omega_2 \cdot \theta_2^c = 1 - \gamma_w \tag{5.3}$$

where $\theta_1^w = \theta_d^w$ and $\theta_1^c = \theta_d^c$ for an incoming data call; $\theta_1^w = \theta_v^w$ and $\theta_1^c = \theta_v^c$ for a voice call; θ_2^w and θ_2^c can be configured according to whether the user prefers the low cost and high data rate of the WLAN or the guaranteed real-time service quality of the cell; the weights ω_1 and ω_2 are based on the relative importance of the two criteria.

5.2 Determination of Assignment Parameters Based on MGFs

Similar to the admission scheme discussed in Chapter 4, the assignment parameters θ_v^w and θ_d^w (or corresponding θ_v^c and θ_d^c) are determined to properly distribute the voice and data traffic load to the overlay cell and WLAN. First, the voice traffic load should be measured and estimated, as voice calls fluctuate in a larger time scale. Given the voice traffic load, the assignment parameters can then be determined to maximize the acceptable data traffic load (λ_d) with QoS satisfaction. That is,

$$\max_{(\theta_v^w,\, \theta_d^w)} \lambda_d \tag{5.4}$$

s.t.

$$\theta_v^w B_v^w + \theta_v^c B_{v2}^c \le Q_{PB}, \quad B_{v1}^c \le Q_{PB}, \quad D_v^c \le Q_{PD}$$

$$\theta_d^w B_d^w + \theta_d^c B_{d2}^c \le Q_{PB}, \quad B_{d1}^c \le Q_{PB}, \quad D_d^c \le Q_{PD}, \quad \overline{T}_d^c \le Q_T, \quad \overline{T}_d^w \le Q_T$$

where $\theta_v^w B_v^w + \theta_v^c B_{v2}^c$ is the blocking probability in the double-coverage area for new voice calls, and $\theta_d^w B_d^w + \theta_d^c B_{d2}^c$ is that for new data calls. The analytical model proposed in Section 4.3 is also applicable to the QoS evaluation for this case. The mean arrival rates of new voice and data calls to the cell from the double-coverage area are respectively

$$\lambda_{nv2}^c = \theta_v^c \cdot \lambda_{v2}, \qquad \lambda_{nd2}^c = \theta_d^c \cdot \lambda_{d2}. \tag{5.5}$$

Similarly, the mean arrival rates of new voice and data calls to the WLAN from the double-coverage area are respectively

$$\lambda_{nv}^w = \theta_v^w \cdot \lambda_{v2}, \qquad \lambda_{nd}^w = \theta_d^w \cdot \lambda_{d2}. \tag{5.6}$$

Nonetheless, as the QoS evaluation for the cell is based on two-dimensional Markov processes, the computation complexity increases with the size of state space. In this work,

we circumvent the computation complexity of solving large-scale balance equations by means of moment generating functions (MGFs).

In general, suppose X and Y are two independent random variables with X being exponentially distributed with parameter λ. Then,

$$\Pr[X > Y] = \int_0^\infty f_Y(y) \int_y^\infty \lambda e^{-\lambda x} dx\, dy = \Psi_Y(-\lambda) \tag{5.7}$$

where $f_Y(\cdot)$ and $\Psi_Y(\cdot)$ are the PDF and MGF of Y, respectively. Letting $Z = \min(X, Y)$, the PDF of Z is given by $f_Z(z) = f_X(z)[1 - F_Y(z)] + f_Y(z)[1 - F_X(z)]$, where $f_X(\cdot)$, $F_X(\cdot)$, and $F_Y(\cdot)$ denote the PDF and CDF of X, and the CDF of Y, respectively. Then, the mean of Z is

$$\mathrm{E}[Z] = \mathrm{E}[X] - \int_0^\infty f_Y(y) \frac{1}{\lambda} e^{-\lambda y} dy = \frac{1}{\lambda} - \frac{1}{\lambda} \Psi_Y(-\lambda). \tag{5.8}$$

Due to the location-dependent mobility within a cell, calls in the cellular-only area and the double-coverage area differ in channel holding time. Depending on the WLAN state, the average channel holding time of voice calls in the cellular-only area can be derived from (3.3) and (5.8) as

$$\begin{aligned} \mathrm{E}[\min(T_v, T_{r1}^c)] &= \frac{1}{\mu_v} - \frac{1}{\mu_v} \Phi_1(-\mu_v) \\ &= \frac{1}{\mu_v} - \frac{1}{\mu_v} p^{cc} \frac{\eta^c}{\eta^c + \mu_v} \frac{1}{1 - p^{cw} \psi(-\mu_v)} \triangleq \frac{1}{\mu_{v1}^c} \end{aligned} \tag{5.9}$$

when there is not sufficient spare capacity in the WLAN for a voice call; and it is $1/(\mu_v + \eta^c)$ when the incoming voice call can be admitted to the WLAN. Similarly, for voice calls in the double-coverage area, the average channel holding time is $\frac{1}{\mu_v} - \frac{1}{\mu_v} \psi(-\mu_v)$ if there is free room for one more voice call in the WLAN or otherwise

$$\begin{aligned} \mathrm{E}[\min(T_v, T_{r2}^c)] &= \frac{1}{\mu_v} - \frac{1}{\mu_v} \Phi_2(-\mu_v) \\ &= \frac{1}{\mu_v} - \frac{1}{\mu_v} p^{cc} \frac{\psi(-\mu_v)}{1 - p^{cw} \psi(-\mu_v)} \triangleq \frac{1}{\mu_{v2}^c} \end{aligned} \tag{5.10}$$

where $\Phi_2(\cdot)$ is given by (3.4). To simplify analysis, we take an average for the mean service rates of voice calls in the cellular-only area and the double-coverage area, which are respectively given by

$$\tilde{\mu}_{v1}^c = D_v^w \mu_{v1}^c + (1 - D_v^w)(\mu_v + \eta^c) \tag{5.11}$$

$$\tilde{\mu}_{v2}^c = D_v^w \mu_{v2}^c + (1 - D_v^w)\frac{\mu_v}{1 - \psi(-\mu_v)}. \tag{5.12}$$

Since voice traffic admitted to the cell from the cellular-only area and the double-coverage area has different average channel holding time approximated by $(\tilde{\mu}_{v1}^c)^{-1}$ and $(\tilde{\mu}_{v2}^c)^{-1}$, respectively, the cell can be viewed as a multi-service loss system [104]. A product-form steady-state distribution exists and is insensitive to service time distributions, provided that the resource sharing among services is under coordinate convex policies. This requires that transitions between states come in pairs. For loss systems with trunk reservation (e.g., the guard channel policy), the insensitivity property and product-form solution are destroyed due to one-way transitions at some states. A recursive method is proposed in [98] to approximate the steady-state distribution, which is shown to be accurate for a wide range of traffic intensities and when the service rates (such as $\tilde{\mu}_{v1}^c$ and $\tilde{\mu}_{v2}^c$) do not greatly differ from each other. Moreover, call blocking probabilities are *almost* insensitive to service time distributions. Hence, we use the recursive approximation to obtain the probability of having k_v^c voice calls in the cell as

$$\pi_v^c(k_v^c) = \pi_v^c(0) \prod_{i=1}^{k_v^c} \left[\frac{\lambda_{v1}^c(i)}{i \cdot \tilde{\mu}_{v1}^c} + \frac{\lambda_{v2}^c(i)}{i \cdot \tilde{\mu}_{v2}^c} \right], \qquad k_v^c = 1, ..., N_v^c \tag{5.13}$$

where

$$\lambda_{v1}^c(i) = \begin{cases} \lambda_{v1} + \lambda_{hv}^{cc} + \lambda_{hv}^{wc}, & i \le \lfloor N_v^c - G_{v1}^c \rfloor \\ \left[1 - (G_{v1}^c - \lfloor G_{v1}^c \rfloor)\right]\lambda_{v1} + \lambda_{hv}^{cc} + \lambda_{hv}^{wc}, & i = \lceil N_v^c - G_{v1}^c \rceil \\ \lambda_{hv}^{cc} + \lambda_{hv}^{wc}, & \lceil N_v^c - G_{v1}^c \rceil + 1 \le i \le N_v^c \end{cases} \tag{5.14}$$

$$\lambda_{v2}^c(i) = \begin{cases} \lambda_{nv2}^c, & i \le \lfloor N_v^c - G_{v2}^c \rfloor \\ \left[1 - (G_{v1}^c - \lfloor G_{v1}^c \rfloor)\right]\lambda_{nv2}^c, & i = \lceil N_v^c - G_{v2}^c \rceil. \end{cases} \tag{5.15}$$

The voice call blocking and dropping probabilities of the cell can then be obtained from $\pi_v^c(k_v^c)$, $k_v^c = 0, 1, ..., N_v^c$.

Given the inter-dependence between the QoS metrics and handoff traffic load, the mean arrival rates of handoff calls out of the cell can be derived recursively. The handoff arrival rates are dependent on handoff probabilities, which are the probabilities that at least one more handoff is required before call completion. The handoff probability of voice calls in the cellular-only area to neighboring cells, denoted by H_v^{cc}, can be obtained according to (5.7) as

$$H_v^{cc} = p^{cc} \cdot \Pr[T_v > T_{r1}^c] = p^{cc}\Phi_1(-\mu_v). \tag{5.16}$$

Similarly, the handoff probability of voice calls in the cellular-only area to the overlay WLAN, denoted by H_v^{cw}, is given by $H_v^{cw} = p^{cw}\Phi_1(-\mu_v)$. Then, the handoff traffic between neighboring cells (λ_{hv}^{cc}), and that between the cell and the overlay WLAN (λ_{hv}^{cw} and λ_{hv}^{wc}) can be obtained by solving the following equations

$$\lambda_{hv}^{cc} = H_v^{cc}\Big[\lambda_{v1} \cdot (1 - B_{v1}^c) + (\lambda_{hv}^{wc} + \lambda_{hv}^{cc})(1 - D_v^c) + \lambda_{nv2}^c \cdot (1 - B_{v2}^c)H_v^{wc}\Big] \tag{5.17}$$

$$\lambda_{hv}^{cw} = H_v^{cw}\Big[\lambda_{v1} \cdot (1 - B_{v1}^c) + (\lambda_{hv}^{wc} + \lambda_{hv}^{cc})(1 - D_v^c) + \lambda_{nv2}^c \cdot (1 - B_{v2}^c)H_v^{wc}\Big] \tag{5.18}$$

$$\lambda_{hv}^{wc} = H_v^{wc}\Big[\lambda_{nv}^w \cdot (1 - B_v^w) + \lambda_{hv}^{cw} \cdot (1 - D_v^w)\Big]. \tag{5.19}$$

To evaluate the QoS metrics of data calls in the cell, two other important aspects need to be properly addressed. First, with the restricted access mechanism, the data call service rates become dependent on both voice and data calls in the cell, as all bandwidth unused by current voice traffic is shared equally by active data calls. Second, the high variability of data call size should be properly dealt with in the QoS evaluation. Under the assumption of nearly complete decomposition of data traffic from voice, when there are j voice calls and k data calls carried by the cell, the cell operates like a symmetric

queue [105] for data calls with

$$\phi(k) = k \cdot \tilde{\mu}_d^c(j,k), \qquad \gamma(l,k) = \delta(l,k) = \frac{1}{k} \tag{5.20}$$
$$l = 1, 2, ..., k, \qquad k = 1, 2, ..., N_d^c$$

where $\phi(k)$ ($\phi(k) > 0$ if $k > 0$) is the total service rate when there are k customers (data calls) in the queue in positions $l = 1, 2, ..., k$; $\tilde{\mu}_d^c(\cdot)$ is the service rate dedicated to each customer; $\gamma(l,k)$ is the fraction of the service rate directed to the customer in position l ($\sum_{l=1}^{k} \gamma(l,k) = 1$); $\delta(l,k+1) = \gamma(l,k+1)$ (symmetric condition) is the probability that an arriving customer moves into position l. A data call carried by the cell may depart due to a handoff to another cell or WLAN. This departure is independent of the queuing position of the data call and behaves like in a multi-server loss system without waiting room. In addition, a data call may also depart from the cell due to call completion. Since the remaining bandwidth unused by current voice calls is shared equally by existing data calls in a PS manner, a fair share of the total service rate is dedicated to each data call irrelevant to its queueing position. A data call arrival or completion affects the amount of resources allocated to each data call, but each data call still keeps a fair share. Therefore, $\delta(l,k)$ and $\gamma(l,k)$ are independent of the queueing positions (i.e., l) of data calls and satisfy the symmetric condition. Hence, data service in the cell can be modeled by a symmetric queue with multiple classes. The service rate $\tilde{\mu}_d^c(\cdot)$ in (5.20) needs to be extended to each class as follows.

Similar to the QoS evaluation for data traffic in Section 4.3.2, data calls admitted in the cell are differentiated into two virtual classes with exponentially distributed service requirements with mean $\frac{1}{b} \cdot \overline{L}_d$ and $b \cdot \overline{L}_d$, respectively [102]. Then, data service in the cell is modeled by a symmetric queue serving multiple classes. Given j voice calls and k data calls in the cell, as in (5.11) and (5.12), the service rates of the two virtual classes

of data calls in the cellular-only area can be approximated by

$$\tilde{\mu}_{d1}^{c1}(j,k) = D_d^w \mu_{d1}^{c1}(j,k) + (1 - D_d^w)\big[b \cdot \nu_d^c(j,k) + \eta^c\big] \tag{5.21}$$

$$\tilde{\mu}_{d2}^{c1}(j,k) = D_d^w \mu_{d2}^{c1}(j,k) + (1 - D_d^w)\big[\nu_d^c(j,k)/b + \eta^c\big] \tag{5.22}$$

where $\nu_d^c(j,k) = \frac{R_{b,d}^c}{\overline{L}_d}$ with $R_{b,d}^c$ given by (4.1) and

$$\mu_{d1}^{c1}(j,k) = \frac{b \cdot \nu_d^c(j,k)}{1 - \Phi_1(-b \cdot \nu_d^c(j,k))}, \qquad \mu_{d2}^{c1}(j,k) = \frac{\nu_d^c(j,k)/b}{1 - \Phi_1(-\nu_d^c(j,k)/b)}. \tag{5.23}$$

Similarly, the service rates of the two virtual classes of data calls admitted to the cell from the double-coverage area can be obtained as

$$\tilde{\mu}_{d1}^{c2}(j,k) = D_d^w \mu_{d1}^{c2}(j,k) + (1 - D_d^w)\frac{b \cdot \nu_d^c(j,k)}{1 - \psi(-b \cdot \nu_d^c(j,k))} \tag{5.24}$$

$$\tilde{\mu}_{d2}^{c2}(j,k) = D_d^w \mu_{d2}^{c2}(j,k) + (1 - D_d^w)\frac{\nu_d^c(j,k)/b}{1 - \psi(-\nu_d^c(j,k)/b)} \tag{5.25}$$

where

$$\mu_{d1}^{c2}(j,k) = \frac{b \cdot \nu_d^c(j,k)}{1 - \Phi_2(-b \cdot \nu_d^c(j,k))}, \qquad \mu_{d2}^{c2}(j,k) = \frac{\nu_d^c(j,k)/b}{1 - \Phi_2(\nu_d^c(j,k)/b)}. \tag{5.26}$$

For symmetric queues such as processor-sharing queues and multi-server queues without waiting room (i.e., loss systems), a product-form stationary queue occupancy distribution exists and is applicable to arbitrarily distributed service requirements [105]. Hence, given k_v^c voice calls in the cell, the equilibrium distribution of the symmetric queue for data traffic in the cell is given by

$$\tilde{\pi}_d^c(k_d^c|k_v^c) = \tilde{\pi}_d^c(0|k_v^c) \prod_{i=1}^{k_d^c} \left[\frac{\frac{b}{b+1}\lambda_{d1}^c(i)}{i \cdot \tilde{\mu}_{d1}^{c1}(k_v^c, i)} + \frac{\frac{1}{b+1}\lambda_{d1}^c(i)}{i \cdot \tilde{\mu}_{d2}^{c1}(k_v^c, i)} + \frac{\frac{b}{b+1}\lambda_{d2}^c(i)}{i \cdot \tilde{\mu}_{d1}^{c2}(k_v^c, i)} + \frac{\frac{1}{b+1}\lambda_{d2}^c(i)}{i \cdot \tilde{\mu}_{d2}^{c2}(k_v^c, i)} \right] \tag{5.27}$$

where $\lambda_{d1}^c(\cdot)$ and $\lambda_{d2}^c(\cdot)$ are the mean arrival rates of data calls from the cellular-only

area and the double-coverage area, respectively, given by

$$\lambda_{d1}^{c}(i) = \begin{cases} \lambda_{d1} + \lambda_{hd}^{cc} + \lambda_{hd}^{wc}, & i \leq \lfloor N_d^c - G_{d1}^c \rfloor \\ \left[1 - (G_{d1}^c - \lfloor G_{d1}^c \rfloor)\right]\lambda_{d1} + \lambda_{hd}^{cc} + \lambda_{hd}^{wc}, & i = \lceil N_d^c - G_{d1}^c \rceil \\ \lambda_{hd}^{cc} + \lambda_{hd}^{wc}, & \lceil N_d^c - G_{d1}^c \rceil + 1 \leq i \leq N_d^c \end{cases} \tag{5.28}$$

$$\lambda_{d2}^{c}(i) = \begin{cases} \lambda_{nd2}^{c}, & i \leq \lfloor N_d^c - G_{d2}^c \rfloor \\ \left[1 - (G_{d1}^c - \lfloor G_{d1}^c \rfloor)\right]\lambda_{nd2}^{c}, & i = \lceil N_d^c - G_{d2}^c \rceil. \end{cases} \tag{5.29}$$

Let $\pi_d^c(\cdot)$ denote the steady-state probability of data calls in the cell. Then, $\pi_d^c(k_d^c) = \sum_{i=0}^{N_v^c} \pi_v^c(i)\tilde{\pi}_d^c(k_d^c|i)$, $k_d^c = 0, 1, ..., N_d^c$. The data call blocking and dropping probabilities and mean data response time can be obtained from π_d^c as in Section 4.3.

5.3 Numerical Results and Discussion

In this section, we first validate the accuracy of the QoS evaluation approaches based on Markov processes and MGFs. Further, we investigate the impact of traffic and mobility variability on the determination of assignment parameters and resulting resource utilization. The same system parameters are used as given in Table 4.2.

5.3.1 Accuracy validation of QoS evaluation approaches

In Section 4.3, we propose a QoS evaluation approach based on two-dimensional Markov processes. Possible approximation errors may be induced due to the location-dependent user mobility model within the cell, traffic prioritization by the limited fractional guard channel policies, and correlation between voice and data traffic. Here, we conduct computer simulation to verify the analysis accuracy. Figure 5.1 - Figure 5.3 compare the analytical results based on two-dimensional Markov processes and computer simulation results under different user mobility and traffic conditions. Figure 5.1 illustrates the results for voice call blocking and dropping probabilities. With a larger mobility

variability parameter a, the user residence time in the WLAN deviates more from the exponential distribution and has a higher variability. It can be seen in Figure 5.1 that the analytical results agree well with the simulation results for different values of a.

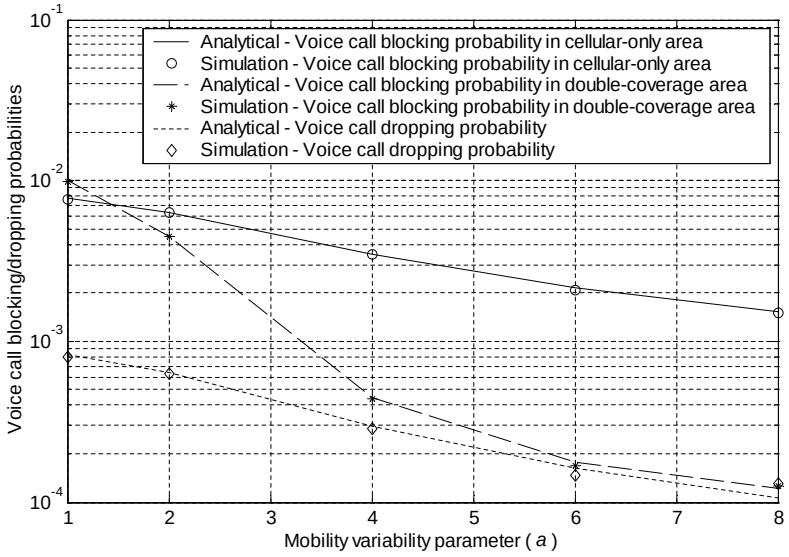
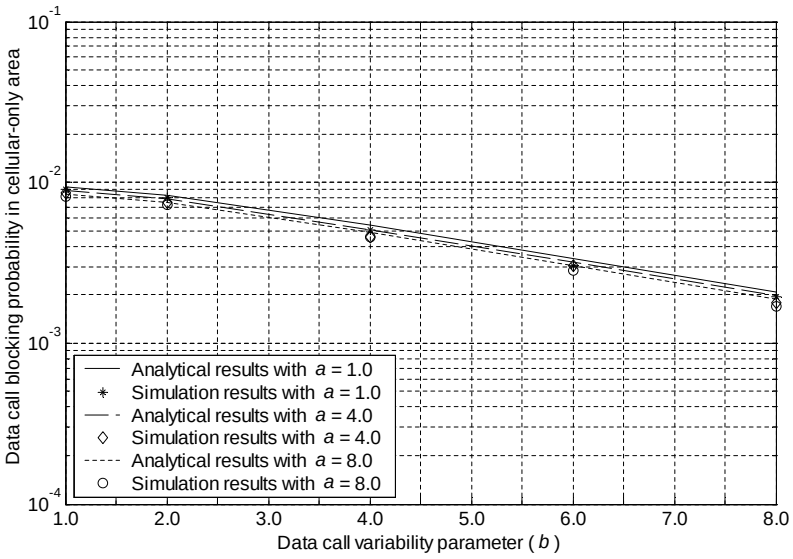
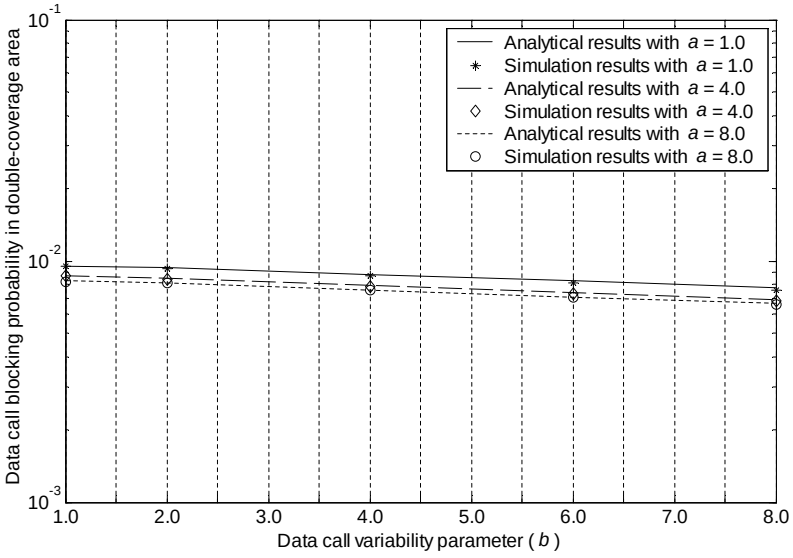


Figure 5.1: Analytical and simulation results of voice call blocking and dropping probabilities.

Figure 5.2 - Figure 5.3 show the analytical and simulation results for data call QoS metrics such as data call blocking/dropping probabilities and mean data response time. It is observed that the analytical results are very close to the simulation results. The gap is much less than 10%, although it increases slightly with the data call variability parameter b for mean data response time in the cell ($\overline{T}_d^c$). The error is induced by the assumption of nearly complete decomposability to decouple the analysis for data calls from voice. To take advantage of the data traffic elasticity, data calls are served under the PS discipline. The mean data response time is insensitive to data call size distribution if the total service capacity is fixed. However, the insensitivity is generally lost with a varying capacity, and the call-level QoS improves with a higher variability

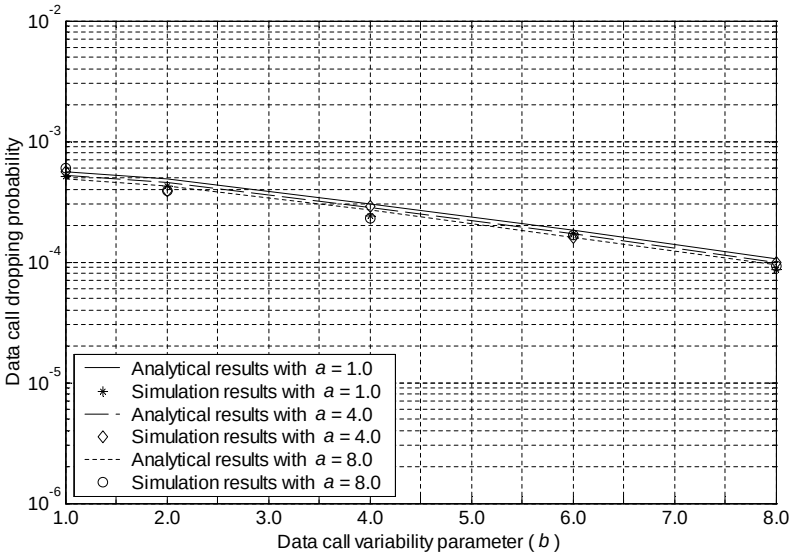


(a)



(b)

Figure 5.2: Analytical and simulation results of data call blocking and dropping probabilities. (a) Blocking probability of new data calls in the cellular-only area (B_{d1}^{c}). (b) Blocking probability of new data calls in the double-coverage area ($\theta_d^w B_d^w + \theta_d^c B_{d2}^c$).

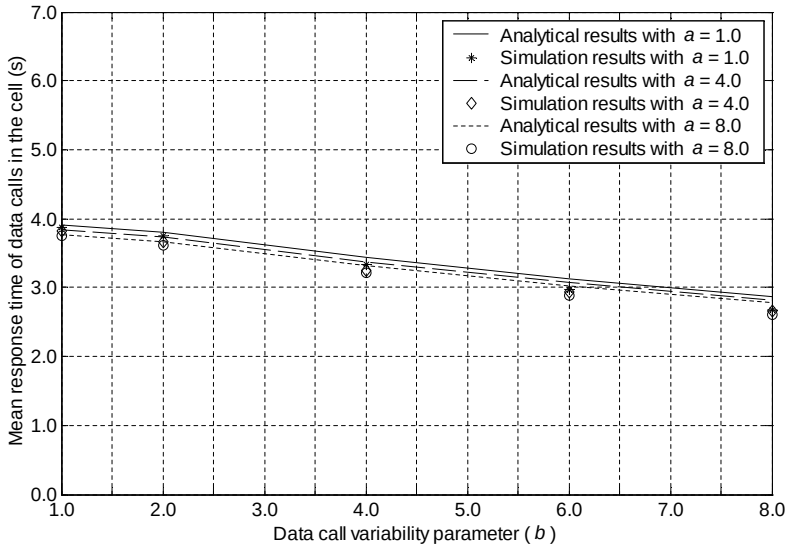


(c)

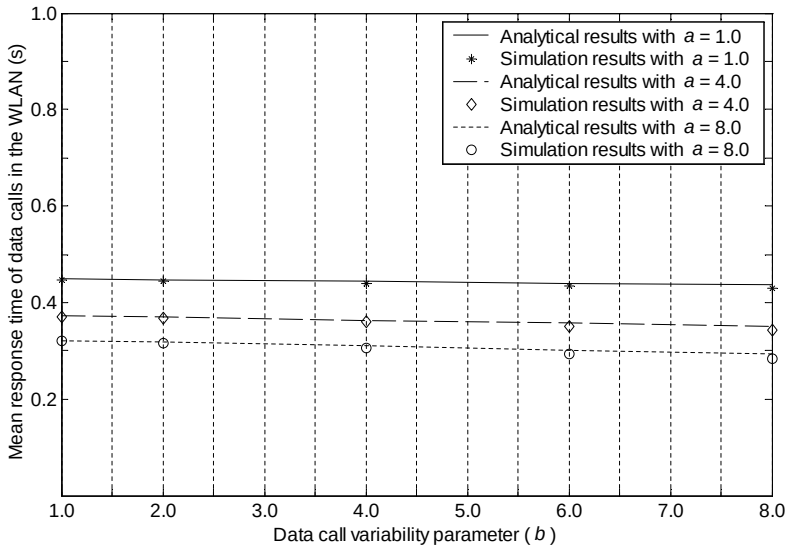
Figure 5.2: Analytical and simulation results of data call blocking and dropping probabilities. (c) Data call dropping probability (D_d^c).

for data call size [106]. With the proposed admission scheme, the bandwidth available to data traffic actually fluctuates with voice call arrivals and departures. As a result, the mean data response time $\overline{T}_d^c$ is overestimated with a larger value of b. Nonetheless, the analysis is still quite accurate especially when data calls arrive and depart in a much smaller time scale than voice calls.

In this work, we simplify the analytical model based on two-dimensional Markov processes by means of MGFs. Figure 5.4 - Figure 5.6 compare the analytical results of the two approaches. As shown in Figure 5.4, the analytical results for voice call blocking and dropping probabilities match well and are tightly bounded by the corresponding requirements. The approach based on MGFs evaluates the QoS metrics with closed-form approximation, which makes it possible to have the assignment parameters adaptive to time-varying traffic load.



(a)



(b)

Figure 5.3: Analytical and simulation results of mean data response time. (a) Mean response time of data calls in the cell ($\overline{T}_d^c$). (b) Mean response time of data calls in the WLAN ($\overline{T}_d^w$).

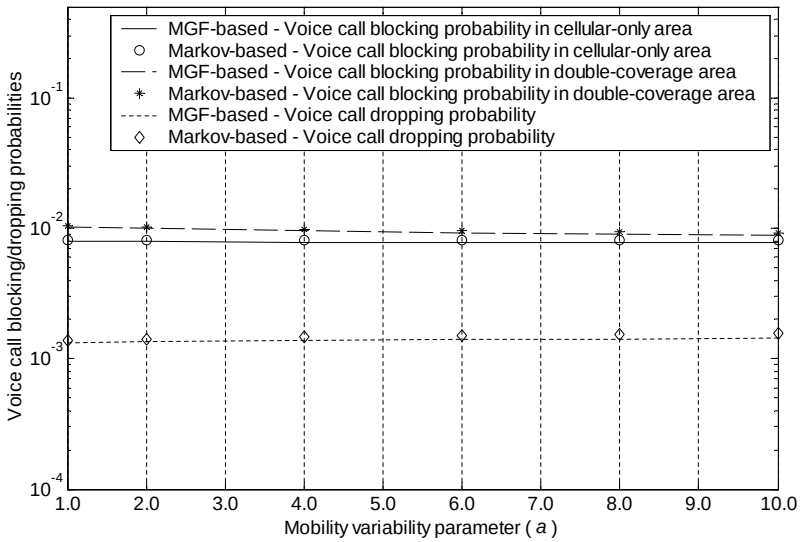
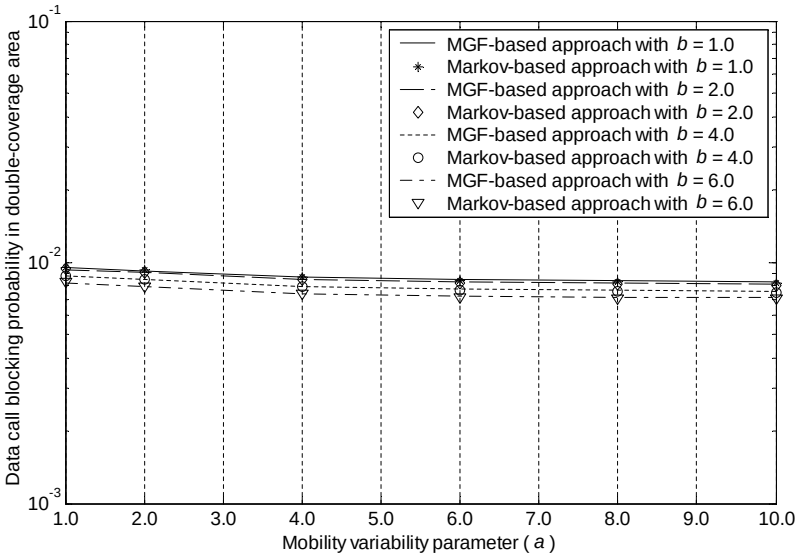
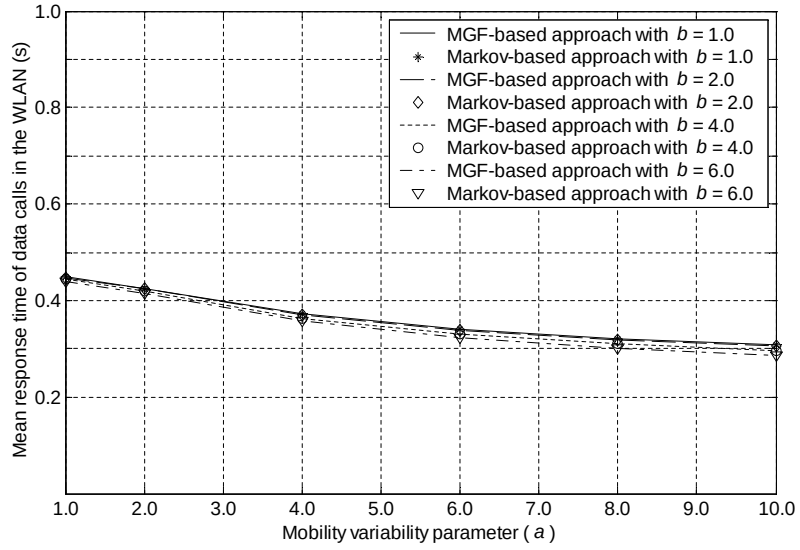


Figure 5.4: Analytical results of Markov-based and MGF-based approaches for voice call blocking and dropping probabilities.

Figure 5.5 - Figure 5.6 show the analytical results of the two evaluation approaches for data call blocking/dropping probabilities and mean data response time. Similarly, we can see that the MGF-based analytical approach presents an accurate evaluation but with a much lower complexity. Moreover, it is observed that the data call variability parameter b can significantly affect the data call QoS. As shown in Figure 5.5, the QoS of data calls in the WLAN-covered area is improved with b not so much as that in the cellular-only area. With the large WLAN bandwidth and conservative admission control in place, the traffic variability can be well absorbed as a high throughput is available to admitted data calls. On the other hand, when there is a higher data call variability, not only is more traffic load relived from the cell by the WLAN, but also most data calls in the cell have a shorter channel holding time and depart from the system faster. As a result, the data call QoS in the cell is significantly improved.

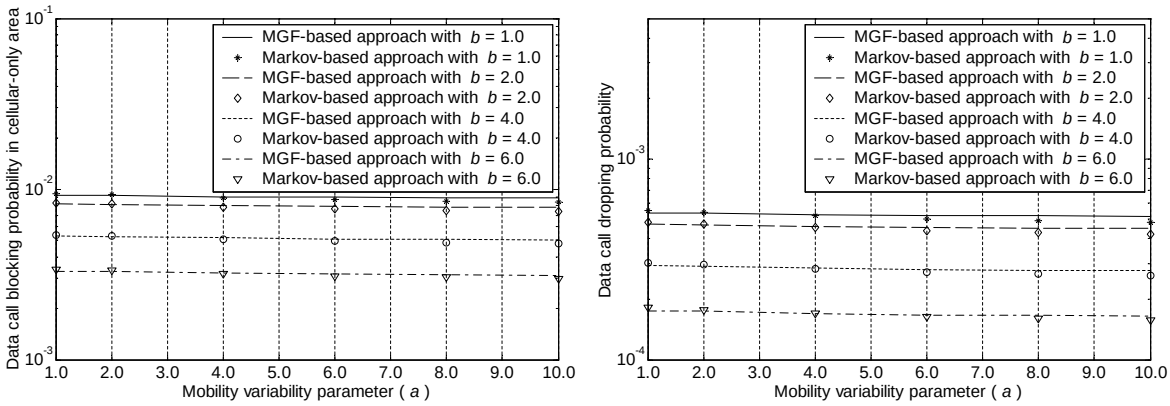


(a)



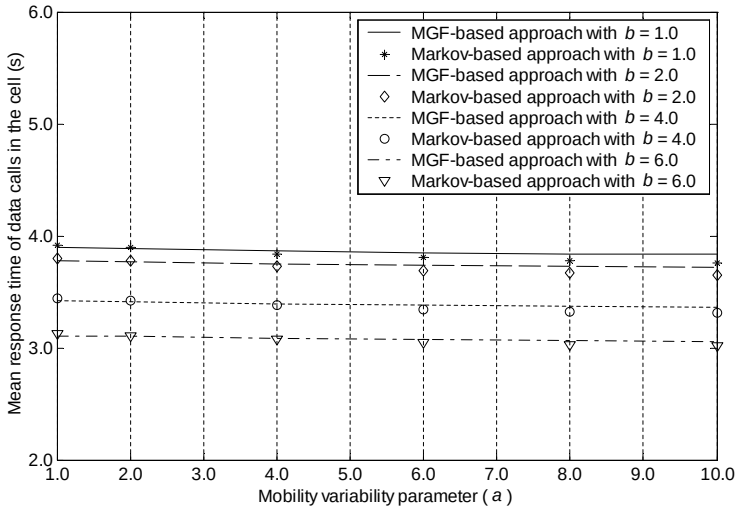
(b)

Figure 5.5: Analytical results of data call QoS in the WLAN-covered area. (a) Blocking probability of new data calls in the double-coverage area ($\theta_d^w B_d^w + \theta_d^c B_{d2}^c$). (b) Mean response time of data calls in the WLAN ($\overline{T}_d^w$).



(a)

(b)



(c)

Figure 5.6: Analytical results of data call QoS in the cellular-only area. (a) Blocking probability of new data calls in the cellular-only area (B_{d1}^{c}). (b) Dropping probability of handoff data calls in the cellular-only area (D_{d}^{c}). (c) Mean response time of data calls in the cell ($\overline{T}_{d}^{c}$).

5.3.2 Dependence of utilization on assignment parameters

The MGF-based approach given in Section 5.2 evaluates QoS metrics accurately and effectively. The assignment parameters θ_v^w and θ_d^w can be determined by applying the MGF-based QoS evaluation in a search algorithm similar to that given in Table 4.1. The corresponding voice and data admission regions can also be obtained for the given θ_v^w and θ_d^w. The best configuration should maximize the traffic load acceptable to a given cell/WLAN cluster.

Figure 5.7 shows the dependence of the acceptable data traffic load (λ_d) on the admission parameter θ_v^w, which is the probability that an incoming voice call in the double-coverage area requests admission to the WLAN. It can be seen that there exist optimal values of θ_v^w that maximize the acceptable λ_d. Here, the voice traffic load is fixed for investigation simplicity. Hence, a maximum λ_d indicates a maximum resource utilization. As discussed in Section 4.4, this results from the load sharing of voice and data traffic within the overlay cell and WLAN. On one hand, when more voice traffic in the double-coverage area is directed to the WLAN with a larger θ_v^w, more cell bandwidth is available for data calls. The cell with a smaller bandwidth is actually the bottleneck of the whole integrated system for data traffic. Hence, with the load balancing of the WLAN, the congestion of the cell and in turn the whole system is effectively relieved. On the other hand, with a larger θ_v^w, the maximum number of voice calls allowed in the WLAN (N_v^w) should also be larger to meet the voice call blocking/dropping probability requirements. However, the WLAN is very inefficient in supporting voice traffic. The small coverage of the WLAN also leads to frequent vertical handoffs between the cell and the WLAN, which may degrade voice quality and increase call dropping possibility. The break of a call into more service stages is also detrimental to multiplexing gain. Although the voice traffic load to the cell is reduced to an extent, the number of data calls that can be accommodated by the WLAN is also significantly reduced. As a result, the total acceptable traffic load starts to decrease when θ_v^w is larger than a threshold.

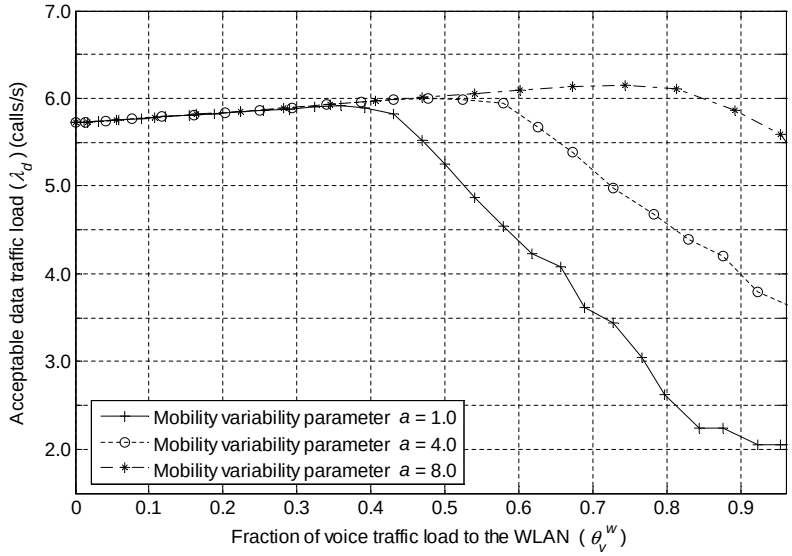


Figure 5.7: Acceptable data traffic load (λ_d) versus the fraction of voice traffic carried by the WLAN (θ_v^w) with different mobility variability parameters (a).

Figure 5.8 shows the variation of the acceptable data traffic load (λ_d) with θ_d^w, i.e., the probability that a data call in the double-coverage area requests admission to the WLAN, which is correlated with θ_v^w. With a smaller θ_v^w to carry a less voice traffic load in the WLAN, θ_d^w can be larger to admit more data calls and provide enough bandwidth for each admitted call. In Figure 5.7, before θ_v^w reaches the point for a maximum acceptable data traffic load, θ_v^w is less than 0.74, and N_v^w less than 16 is sufficient to meet the bounds for voice call blocking/dropping probabilities. For these cases, a high data call throughput is achievable since the number of voice calls allowed in the WLAN is quite restricted, and θ_d^w can be as large as 90% to carry almost all the data traffic in the double-coverage area. Within this region, a larger θ_v^w alleviates more voice traffic from the cell but does not affect much the data service in the WLAN, which results in a larger acceptable λ_d. On the other hand, when θ_v^w and corresponding N_v^w are further increased, θ_d^w is even smaller and the WLAN cannot carry a large portion

of the data traffic load in the double-coverage area. As a result, the bottleneck effect of the cell becomes evident. Thus, as shown in Figure 5.8, the acceptable λ_d decreases with a smaller θ_d^w. In conclusion, the effectiveness of the WLAN as a complement to the cell may be greatly jeopardized with a small θ_d^w and a very large θ_v^w. To maximize the utilization, θ_v^w should be large enough to balance voice traffic load from the cell and also small enough to avoid an inefficient utilization of the WLAN for voice support.

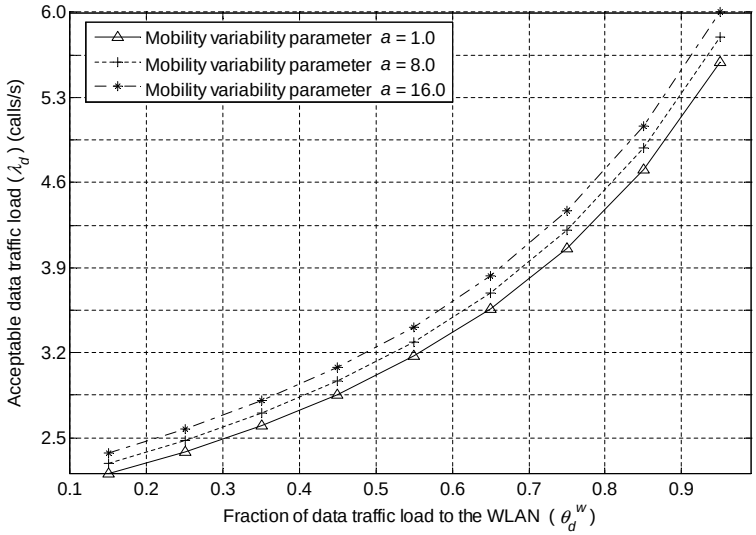


Figure 5.8: Acceptable data traffic load (λ_d) versus the fraction of data traffic carried by the WLAN (θ_d^w) with different mobility variability parameters (a).

5.3.3 Impact of user mobility and traffic variability

The curves in Figure 5.7 and Figure 5.8 are obtained with different mobility variability parameters (a). It is observed that the acceptable data traffic load (λ_d) is larger with a larger value of a. That is, a higher utilization is achievable when the variability of user mobility in the double-coverage area is higher. As illustrated in Figure 5.7, when $a = 1.0$, 4.0, and 8.0, the highest utilization is achieved with $\theta_v^w = 0.36$, 0.48, and 0.74,

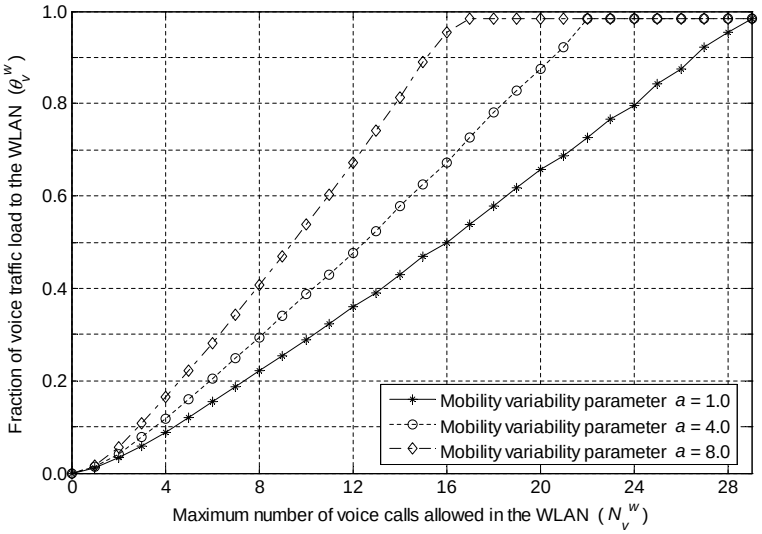


Figure 5.9: Fraction of voice traffic carried by the WLAN (θ_v^w) versus the maximum number of voice calls allowed in the WLAN (N_v^w) with different mobility variability parameters (a).

respectively. A larger parameter a indicates that more users staying within the WLAN for a shorter time. Then, more voice calls may have a smaller channel holding time and occupy the WLAN bandwidth for a less time. As shown in Figure 5.9, given a fixed N_v^w (i.e., the maximum number of voice calls allowed in the WLAN), when the parameter a is larger, a larger fraction of voice calls in the double-coverage area can be carried by the WLAN and relieved from the cell. Therefore, the data call throughput in the cell is higher and more traffic is acceptable with QoS satisfaction.

The data call variability also affects the assignment parameters and resource utilization. As shown in Figure 5.10, more data traffic is acceptable with a larger value of b, which indicates a higher variability of data call size. For a fixed mean of data call size, a larger value of b indicates that more data calls have a smaller size and less have an extremely larger size. Hence, more data calls have a shorter channel holding time and can be carried by the WLAN with a high throughput. Also, more data calls

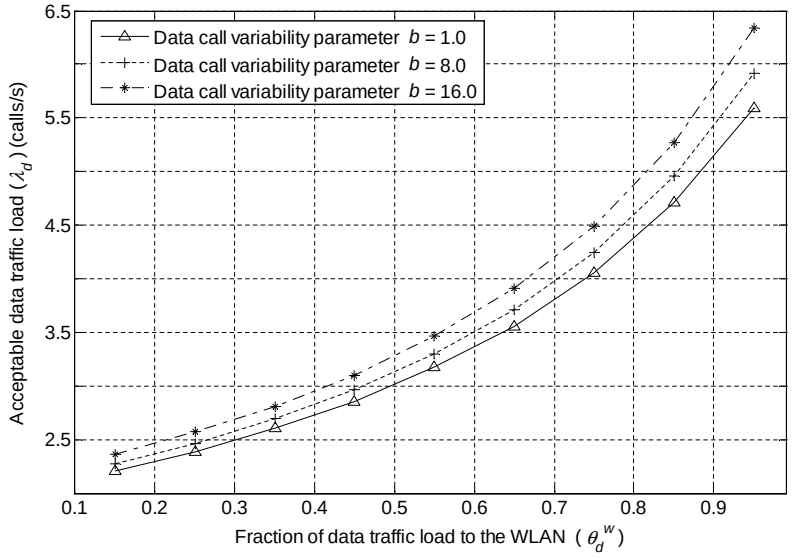


Figure 5.10: Acceptable data traffic load (λ_d) versus the fraction of data traffic carried by the WLAN (θ_d^w) with different data call variability parameters (b).

are likely to complete service within the WLAN and do not need to hand over to the cell when users move out of the WLAN coverage. Therefore, the WLAN bandwidth is more effectively utilized to reduce the traffic load to the cell of small bandwidth. With the MGF-based approach, we can take into account the variability of user mobility and data traffic in determining the assignment parameters.

5.4 Summary

In this chapter, we have proposed a new admission control scheme with randomized call assignment for cellular/WLAN interworking. An incoming call in the double-coverage area is assigned to the overlay cell and WLAN according to properly defined assignment probabilities. The main contributions of this work are as follows:

- The proposed scheme enables distributed control to render feasible implementa-

tion in a loosely coupled network. The control overhead is reduced by avoiding frequent signaling exchanges to update states of both networks.

- An MGF-based analytical approach is developed to effectively and accurately evaluate the QoS metrics such as call blocking/dropping probabilities and mean data response time. The assignment parameters can be determined with the analytical approach to achieve a high utilization.

- We have investigated the impact of user mobility and data traffic variability on resource utilization. It is observed that the high data traffic variability can be exploited to improve the interworking effectiveness.

Chapter 6

Size-Based Assignment with SRPT Scheduling

From the preceding study, we can see that traffic load sharing is essentially important to maximize the interworking effectiveness. As discussed in Chapter 4 and Chapter 5, the voice and data traffic load is distributed to the coupled systems via call assignment and admission control, which properly differentiates upward/downward vertical handoff calls, horizontal handoff calls, new calls in the cellular-only area and the double-coverage area. The impact of user mobility on load sharing has also been investigated in Chapter 5. Nonetheless, similar to most previous works, the research attention is directed to vertical handoff calls crossing WLAN borders. Actually, call reassignment can also be performed via dynamic vertical handoff within the overlay area. In [107], there is some initial study on dynamic session transfer in hierarchical integrated networks as an analogy to task migration in distributed operating systems. However, not many analytical works consider the dynamic vertical handoff within the overlay area triggered by network states instead of user mobility. As the dynamics of both integrated systems are involved, the load sharing problem becomes very complex, particularly for a multi-service scenario.

In this work, we propose a new load sharing scheme for voice and elastic data traffic in the cellular/WLAN integrated network. The main contributions are as follows:

- In the proposed scheme, voice traffic load is preferably admitted into the cell by means of both initial call assignment and call reassignment via dynamic vertical handoff. A large multiplexing gain is achieved by pooling the free bandwidth in the two systems to effectively serve elastic data traffic.
- The heavy-tailedness of data call size is exploited by an assignment strategy based on a size threshold. Further, the system capacity is improved by using the efficient *shortest remaining processing time* (SRPT) scheduling for data calls in the cell.
- The system performance is evaluated accurately with an analytical approach. It characterizes the heavy-tailedness of data traffic and dynamic vertical handoff triggered by network states. The data call size threshold can be determined with the analytical model.

6.1 Assignment and Scheduling for Heavy-Tailed Data Calls

As shown in Section 4.2, it is very inefficient to support real-time services in the WLAN due to excessive control overhead. In contrast, the cellular network has strength in real-time service provisioning. The large cell size and ubiquitous cellular coverage can reduce handoff frequency and in turn the impact of handoff latency on the delay-sensitive real-time traffic. Thus, an incoming voice call is preferably distributed to the cell, and overflows to the WLAN only if there is not sufficient spare capacity in the cell. In this work, we further consider call reassignment for voice calls via dynamic vertical handoff from the WLAN to the cell, which can be performed whenever the cell has spare capacity to accommodate more voice calls. As such, voice calls are more

concentrated in the cell and provisioned fine QoS guarantee. The bandwidth unused by voice traffic in the two systems can then be pooled to serve data calls. The rationale behind the idea can be understood by viewing the integrated cell and WLAN as two coupled queueing systems with service rates C_1 and C_2, respectively. By exploiting the cellular/WLAN interworking and vertical handoff, the performance of the two coupled systems within the overlay area can approach that of one queue with a larger service rate ($C_1 + C_2$), which maximizes the multiplexing gain [108].

On the other hand, admitted calls in the WLAN are served under the PS discipline with the contention-based access. The service queue with PS discipline exhibits unique characteristics, which should be considered in the resource management and may significantly affect the utilization. Based on queueing analysis for $M/G/1/K-PS$ queues, Figure 6.1 can be obtained to illustrate the dependence of QoS on the offered traffic load factor (ρ_d) and number of admitted data calls (N_d). It can be seen that the mean response time T_d increases relatively slowly with ρ_d, when the system is underloaded with $\rho_d \ll 1$. For a moderately large value of N_d, the data call blocking probability B_d is very small and T_d is almost independent of N_d. However, when overload occurs with $\rho_d \geq 1$, T_d increases fast and almost linearly with ρ_d and N_d, while B_d converges fast to the limit $\frac{\rho_d - 1}{\rho_d}$ with a moderately large value of N_d [109]. Hence, admitting more calls is not effective to reduce the blocking probability in overload but may significantly degrade the perceived performance. It is important to ensure that the system operates in a normal load condition, so that the blocking probability is bounded and a sufficiently high throughput is maintained for admitted calls [109].

Because voice calls are preferably distributed to the cell via initial call assignment and call reassignment via dynamic vertical handoff, the average cell bandwidth available to data traffic is relatively low when the voice traffic load is high. It is necessary and feasible to serve data calls in the cell with a more efficient service discipline. In this work, we consider the shortest remaining processing time (SRPT) discipline, which is

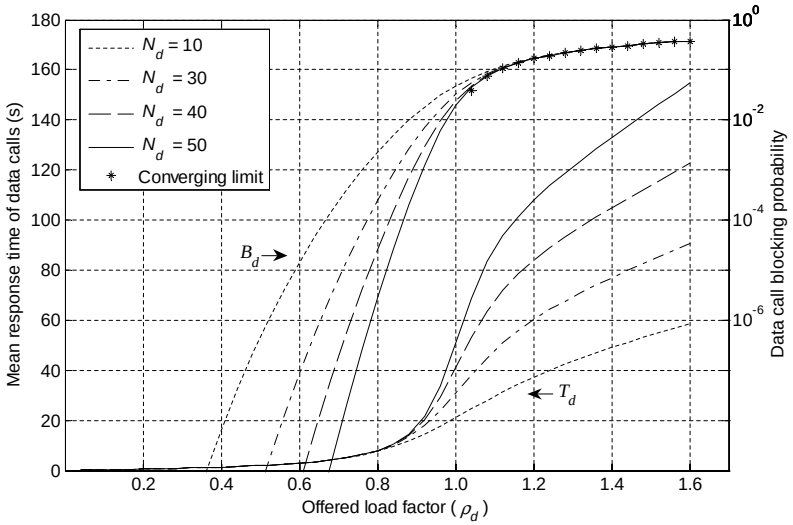


Figure 6.1: Data call QoS in terms of mean response time (T_d) and data call blocking probability (B_d) under PS service discipline versus offered load factor (ρ_d) and number of admitted data calls (N_d).

optimal in terms of minimizing the mean response time. Under the SRPT, only one call with the least remaining data to transmit is scheduled first and receives service at an instant. Given an incoming data call with a size smaller than the remaining data size of the call in service, the ongoing call is preempted and waits in the queue, while the new call is served subsequently. In contrast, under the PS, each ongoing call shares an equal quantum of service. As such, smaller-size calls under the SRPT will not be stuck in the system for such a long duration as when the bandwidth is shared with data calls of a larger size.

It is known that the SRPT can significantly outperform the PS when the call size is heavy-tailed and the load is high. It may be suspected that the improvement of SRPT over PS comes at the expense of a longer response time for calls with a larger data size. Thus, the SRPT is often thought to be unfair as it favors short calls and penalizes long calls. An argument for this claim is the Kleinrock conservation law [110], which holds

for service disciplines not making use of the size but is not necessarily true for size-based disciplines such as the SRPT. It is proved in [111] that, for any load condition and any continuous heavy-tailed size distribution with finite mean and variance, at least 99% of the data calls have a smaller response time under the SRPT than under the PS. These 99% of calls actually do significantly better, and the unfairness of SRPT diminishes with the heavy-tailed property. In addition, the control overhead of SRPT such as for preemption is also not higher than that of PS [111]. In practical systems, the PS may be implemented in a round-robin manner and each call is preempted after receiving one quantum of service. In contrast, the preemption of SRPT only occurs when a new call of a smaller size arrives, which involves less preemption overhead. As the SRPT is applied at the call level instead of the packet level, the implementation complexity and cost should be affordable.

In this study, we consider some specific elastic data applications such as Web browsing and file transfer. They usually preserve a request-response pattern and are primarily unidirectional from application servers to user terminals. The Web documents or data files are pre-stored in a Web server or file server. It is possible to know the data call size *a priori* from session signaling. For example, a session description protocol (SDP) offer/answer mechanism has been proposed as an Internet draft for file transfer [112]. By introducing a set of new SDP attributes, it is possible to deliver some meta information of the file (such as content type and size) before the actual transfer. On the other hand, cross-layer design has become very popular and essential in the wireless domain to address the unique challenges such as the scarce radio resources and highly error-prone transmission conditions. The information exchange across different protocol layers can further improve the system performance.

Hence, in our load sharing scheme, we exploit the meta information of data calls that can be passed to the network layer. In particular, a data call is distributed to the cell if the call size is not greater than a threshold Φ_d and the cell bandwidth available

to data traffic is at least R_d^c. Otherwise, that data call is assigned to the WLAN. By properly determining the size threshold (to be discussed in Section 6.2.2), we can improve the overall resource utilization without degrading the user QoS experience.

6.2 Performance Analysis of the Proposed Scheme

In this section, we analytically evaluate the QoS metrics such as voice/data call blocking probabilities and mean data response time, based on which we can determine the data call size threshold (Φ_d).

6.2.1 Analytical model for QoS evaluation

As discussed in Section 4.1, data calls in the WLAN share the available bandwidth in a PS manner. Under the PS, the mean response time is insensitive to the call size distribution if the overall service capacity is fixed. Nonetheless, due to the random access in the WLAN, the bandwidth available to data traffic actually fluctuates not only with voice call arrivals/departures but also with the contention status. The insensitivity of mean response time is generally lost in case of a varying capacity [106]. For data calls of a heavy-tailed size and high variability, the call-level performance even improves over the case with an exponentially distributed data call size. However, with admission control in place, the insensitivity can be retained for a high load condition, where proper resource allocation and load control are critical to prevent QoS violation. In a light load case, the call blocking probability is usually sufficiently low and all admitted calls are provided satisfactory QoS. Hence, we assume that the QoS of data traffic in the WLAN is insensitive to the heavy-tailed call size distribution. The insensitivity assumption is validated by the numerical results given in Section 6.3, although conservative control is possible for a light load due to QoS underestimation.

Similar to the previous study, we assume that voice and data call arrivals to the

double-coverage area are independent Poisson processes with mean rates denoted by λ_v and λ_d, respectively. Since data calls are assigned to the integrated cell and WLAN based on the data call size and bandwidth occupancy, the data call arrivals to the cell and the WLAN are still Poisson processes with mean rates denoted by λ_d^c and λ_d^w, respectively. In addition, given the insensitivity assumption for data service in the WLAN, we can model the integrated cell/WLAN cluster with a three-dimensional Markov process, in which the state (i, j, k) denotes the numbers of voice and data calls in the WLAN (i and j, respectively) and the number of voice calls in the cell (k). The steady-state probability is denoted by $\pi(i, j, k)$. Based on the bandwidth occupancy of voice traffic in the cell and the overall data call size distribution given in (2.8), the mean data call arrival rate to the cell can be derived as

$$\lambda_d^c = \lambda_d \cdot \delta_d^c \cdot \chi_d^c \tag{6.1}$$

$$\delta_d^c = \int_0^{\Phi_d} f_{L_d}(x)\, dx, \quad \chi_d^c = \sum_{(i,j)} \sum_{k:\ C_d^c(k) \geq R_d^c} \pi(i, j, k)$$

where δ_d^c is the fraction of data calls with a size not greater than Φ_d, $(1 - \chi_d^c)$ is the probability that such a data call is blocked by the cell due to congestion, and $C_d^c(k)$ is the maximum cell capacity available to data traffic when there are k voice calls in progress. Similarly, the mean data call arrival rate to the WLAN can be obtained as

$$\lambda_d^w = \lambda_d \cdot \left[\delta_d^c \cdot (1 - \chi_d^c) + (1 - \delta_d^c)\right] = \lambda_d \cdot \left(1 - \delta_d^c \cdot \chi_d^c\right). \tag{6.2}$$

The corresponding state transition rates of the three-dimensional Markov process are given at the top of next page, where N_v^c and N_v^w are the maximum numbers of voice calls admitted in the cell and the WLAN, respectively, $N_d^w(i)$ is the maximum number of data calls allowed in the WLAN with i voice calls in progress[1], $\xi_d^w(i, j)$ is the mean

[1] N_v^c, N_v^w, and $N_d^w(i)$ are obtained from the admission regions of the cell and the WLAN, i.e., the feasible sets of vectors (n_v^c, n_d^c) and (n_v^w, n_d^w), respectively. Here, $N_v^c = \max(n_v^c)$, $N_v^w = \max(n_v^w)$, and $N_d^w(i) = \max(n_d^w)$, given $n_v^w = i$.

$$
\begin{aligned}
&(i,j,k) \rightarrow (i,j,k+1): \ \lambda_v, && \text{if } i \le N_v^w,\ j \le N_d^w(i),\ k \le N_v^c - 1 \\
&(i,j,k) \rightarrow (i,j,k-1): \ k \cdot \mu_v, && \text{if } i = 0,\ j \le N_d^w(i),\ 1 \le k \le N_v^c \\
&(i,j,k) \rightarrow (i+1,j,k): \ \lambda_v, && \text{if } i \le N_v^w - 1,\ j \le N_d^w(i+1),\ k = N_v^c \\
&(i,j,k) \rightarrow (i-1,j,k): \ (i+k) \cdot \mu_v, && \text{if } 1 \le i \le N_v^w,\ j \le N_d^w(i),\ k = N_v^c \\
&(i,j,k) \rightarrow (i,j+1,k): \ \lambda_d^w, && \text{if } i \le N_v^w,\ 0 \le j \le N_d^w(i) - 1,\ k \le N_v^c \\
&(i,j,k) \rightarrow (i,j-1,k): \ j \cdot \xi_d^w(i,j)/g_d^w, && \text{if } i \le N_v^w,\ 1 \le j \le N_d^w(i),\ k \le N_v^c
\end{aligned}
\tag{6.3}
$$

service rate provided to each data call when there are i voice calls and j data calls in the WLAN, and g_d^w is the mean size of data calls flowing to the WLAN. Note that the transition rate from state (i,j,k) to state $(i-1,j,k)$ consists of two components. One is due to the completion of the i voice calls in the WLAN with a mean rate of $i \cdot \mu_v$, and the other is due to the completion of the k voice calls in the cell with a mean rate $k \cdot \mu_v$. When one of the k voice calls in the cell completes and makes room for a new voice call, one of the i voice calls in the WLAN can be handed over to the cell. According to the proposed scheme and the overall data call size distribution given in (2.8), g_d^w can be derived as

$$
g_d^w = \frac{(1-\chi_d^c)\displaystyle\int_0^{\Phi_d} x f_{L_d}(x)\ dx + \int_{\Phi_d}^{\infty} x f_{L_d}(x)\ dx}{\delta_d^c \cdot (1-\chi_d^c) + (1-\delta_d^c)}. \tag{6.4}
$$

The first term in the numerator of (6.4) corresponds to data calls of a size not greater than Φ_d, which are blocked by the cell due to congestion with a probability $(1-\chi_d^c)$ and overflow to the WLAN. The second term in the numerator accounts for the data calls that have a size larger than Φ_d and are assigned to the WLAN to request admission. The denominator is a normalization constant for the size distribution of data calls flowing to the WLAN.

Due to the interdependence between i and k as shown in the state transition rates of (6.3), the size of the state space does not explode with the third dimension of the

Markov process, i.e., the number of voice calls in the cell. The steady-state probabilities $\pi(i, j, k)$ can be obtained by solving a very sparse linear system of balance equations. Then, the voice call blocking probability B_v is given by

$$B_v = \sum_{\substack{(i,j):\ i \le N_v^w \\ j > N_d^w(i+1)}} \pi(i, j, N_v^c). \tag{6.5}$$

That is, an incoming voice call is blocked if there are N_v^c voice calls in the cell and not sufficient spare capacity is available for one more voice call, and if the WLAN is also congested with i voice calls and j data calls, which means that, with the j data calls already in progress, the admission of one more voice call in the WLAN will result in delay violation to the admitted i voice calls.

As illustrated in Figure 6.1, when overload occurs, the mean response time under the PS increases dramatically with the offered load and the number of admissible calls (N_d), while the call blocking probability converges and cannot be reduced by increasing N_d. In contrast, in an underload case, the call blocking probability is sufficiently small with a reasonably large value of N_d and the mean response time is almost independent of N_d. Similar phenomenon is observed for the SRPT discipline. Hence, the QoS of data calls can be assured by maintaining an underload condition for data traffic in the cell. This can be achieved by properly determining the size threshold Φ_d. Then, the data call blocking probability B_d can be obtained as

$$B_d = \left[\delta_d^c \cdot (1 - \chi_d^c) + (1 - \delta_d^c)\right] B_d^w = (1 - \delta_d^c \cdot \chi_d^c) \cdot B_d^w \tag{6.6}$$

where B_d^w is the data call blocking probability of the WLAN and is given by

$$B_d^w = \sum_{\substack{(i,j):\ i \le N_v^w \\ j+1 > N_d^w(i)}} \sum_{k=0}^{N_v^c} \pi(i, j, k). \tag{6.7}$$

That is, the admission of a new data call should not degrade the WLAN capacity so much that the bandwidth requirement of ongoing voice calls cannot be satisfied. From

the Little's law, the mean response time of data calls served in the WLAN can be obtained as

$$\overline{T}_d^w = \frac{1}{\lambda_d^w \cdot (1 - B_d^w)} \sum_{\substack{(i,j):\ i \le N_v^w \\ j \le N_d^w(i)}} \sum_{k=0}^{N_v^c} j \cdot \pi(i,j,k). \tag{6.8}$$

On the other hand, the mean response time of data calls admitted to the cell can be obtained approximately from the $M/G/1-SRPT$ queue. This is because data call arrivals to the cell is still a Poisson process with a mean rate λ_d^c given in (6.1). The data call blocking probability is negligibly small if an underload condition is guaranteed by the threshold Φ_d. The average bandwidth allocated to data calls is

$$\overline{C}_d^c = \sum_{\substack{(i,j):\ i \le N_v^w \\ j \le N_d^w(i)}} \sum_{k=0}^{N_v^c} C_d^c(k) \cdot \pi(i,j,k). \tag{6.9}$$

Then, based on the formulas in [113], the mean response time is approximated by

$$\overline{T}_d^c = \int_0^{\Phi_d} \frac{1}{\delta_d^c} f_{L_d}(x)\ \Gamma_d^c(x)\ dx \tag{6.10}$$

where $\frac{1}{\delta_d^c} f_{L_d}(x)$ $(0 < x \le \Phi_d)$ is the PDF of the data call size in the cell, and $\Gamma_d^c(x)$ is the conditional response time for a data call of size x, given by

$$\Gamma_d^c(x) = \int_0^y \frac{dt}{1 - \rho_d^c(t)} + \frac{\lambda_d^c \Big[\int_0^y t^2 g_{L_d}(t)\ dt + y^2(1 - G_{L_d}(y)) \Big]}{2\big[1 - \rho_d^c(y)\big]^2} \tag{6.11}$$

$$\rho_d^c(y) = \lambda_d^c \int_0^y t \cdot g_{L_d}(t)\ dt, \quad y = \frac{x}{\overline{C}_d^c} \tag{6.12}$$

$$g_{L_d}(t) = \frac{1}{\delta_d^c} W_b(t, \alpha_d, \beta_d / \overline{C}_d^c), \quad 0 < t \le \Phi_d / \overline{C}_d^c. \tag{6.13}$$

Here, $g_{L_d}(\cdot)$ denotes the PDF of a bounded Weibull distribution and $G_{L_d}(\cdot)$ the corresponding CDF. In contrast to the data call size distribution $W_b(x, \alpha_d, \beta_d)$ given in (2.8), the scale parameter β_d is proportionally modified with $\overline{C}_d^c$ to switch the unit from data call size to service time.

For comparison purpose, when data calls in the cell are served under the PS discipline, the mean response time can be approximated by [103]

$$\overline{T}_d^c = \frac{(\overline{\rho}_d^c)^{N_d^c+1}(N_d^c\,\overline{\rho}_d^c - N_d^c - 1) + \overline{\rho}_d^c}{\lambda_d^c \cdot \left[1-(\overline{\rho}_d^c)^{N_d^c}\right](1-\overline{\rho}_d^c)}, \quad \overline{\rho}_d^c = \rho_d^c(\Phi_d/\overline{C}_d^c) \tag{6.14}$$

where $\overline{\rho}_d^c$ is the average load factor of data traffic in the cell, which can be obtained from (6.12), and N_d^c is the maximum number of data calls allowed in the cell. Taking into account the size-based assignment for data traffic, the overall mean response time of data calls can be evaluated by

$$\overline{T}_d = \frac{\delta_d^c\,\chi_d^c \cdot \overline{T}_d^c + \left[\delta_d^c \cdot (1-\chi_d^c) + (1-\delta_d^c)\right](1-B_d^w) \cdot \overline{T}_d^w}{\delta_d^c\,\chi_d^c + \left[\delta_d^c \cdot (1-\chi_d^c) + (1-\delta_d^c)\right](1-B_d^w)}. \tag{6.15}$$

6.2.2 Determination of data call size threshold

In the proposed scheme, voice calls are preferably distributed to the cell for high efficiency and fine QoS. Data traffic should be properly balanced between the two systems correspondingly. Based on the observations in Section 6.1, there are some important principles to follow in determining the data size threshold Φ_d.

First, an underload condition should be ensured for data traffic in the cell. That is, the data load factor in the worst case, denoted by $\hat{\rho}_d^c$, is less than 1. Similar to (6.12), $\hat{\rho}_d^c$ can be obtained as

$$\hat{\rho}_d^c = \lambda_d^c \int_0^{\Phi_d/R_d^c} t \cdot \frac{1}{\delta_d^c} W_b(t, \alpha_d, \beta_d/R_d^c)\; dt \tag{6.16}$$

where R_d^c is the minimum cell bandwidth available to data traffic, and $\frac{1}{\delta_d^c}W_b(t, \alpha_d, \beta_d/R_d^c)$, $0 < t \le \Phi_d/R_d^c$, denotes the PDF of a bounded Weibull distribution with shape parameter α_d and scale parameter β_d/R_d^c. Moreover, data calls with a smaller size usually expect a shorter response time than those with a larger size. As data calls in the cell have a smaller size than most of those in the WLAN, our second principle is to guarantee that $\overline{T}_d^c \le \overline{T}_d^w$. The mean response time $\overline{T}_d^w$ and $\overline{T}_d^c$ are given by (6.8) and (6.10), respectively.

Last, when determining the size threshold Φ_d, we should make a good trade-off between user-perceived QoS such as mean data response time and GoS in terms of call blocking probabilities. An appropriate threshold Φ_d^* can be determined to satisfy the following condition:

$$B_d(\Phi_d) < B_d(\Phi_d^*) \Rightarrow \overline{T}_d(\Phi_d) > \overline{T}_d(\Phi_d^*), \quad \forall\ \Phi_d \neq \Phi_d^*. \tag{6.17}$$

That is, the size threshold Φ_d should be chosen so that the mean response time $\overline{T}_d$ is minimized without increasing the data call blocking probability B_d. As such, the resource utilization is improved without degrading the QoS performance. As $\overline{T}_d$ and B_d can be evaluated analytically with the model given in Section 6.2, Φ_d^* can be determined with various search techniques such as the golden section search [114]. Also, as $\overline{T}_d$ may be sensitive to Φ_d in the range around Φ_d^*, the above principles need to be applied conservatively to guarantee system stability. In Section 6.3.2, we will discuss more details about the dependence of $\overline{T}_d$ and B_d on Φ_d and the determination of Φ_d.

6.3 Numerical Results and Discussion

In this section, we first validate the analytical model given in Section 6.2, and then investigate the impact of data call size threshold on user-perceived QoS. Last, the performance of the size-based load sharing scheme is compared with those of the admission schemes with service-differentiated assignment and randomized assignment, discussed in Chapter 4 and Chapter 5, respectively. System parameters in Table 4.2 are used in the following numerical analysis. For investigation simplicity, we fix the voice traffic load by taking the mean voice call arrival rate λ_v at 0.45 (calls/s).

6.3.1 Accuracy validation of the analytical model

Figure 6.2 shows the analytical and simulation results of voice and data call blocking probabilities (B_v and B_d, respectively) and mean response time of data calls ($\overline{T}_d$) when

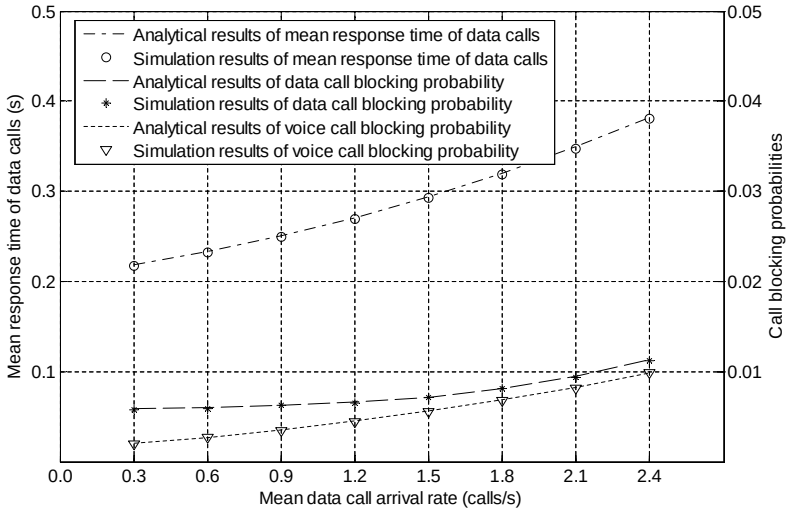
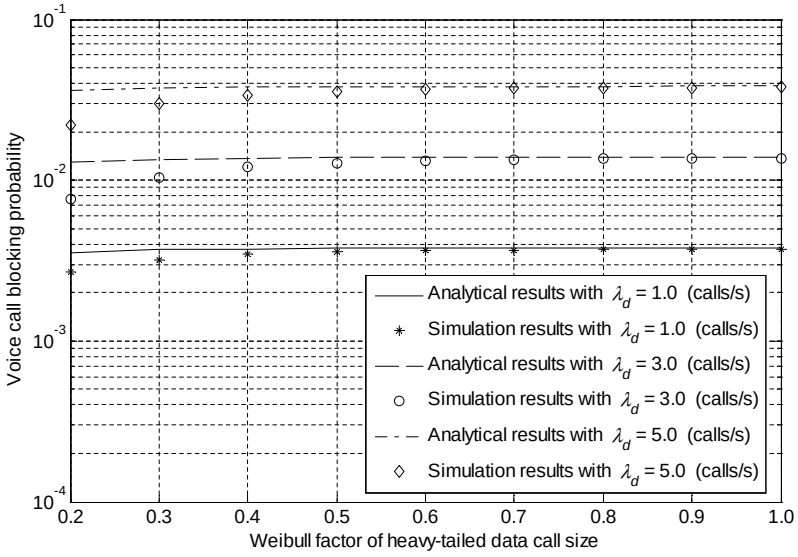


Figure 6.2: Analytical and simulation results of voice and data call blocking probabilities (B_v and B_d, respectively) and mean data response time ($\overline{T}_d$) versus mean data call arrival rate (λ_d) with an exponentially distributed data call size ($W_{L_d} = 1.0$).

the data call size is exponentially distributed, i.e., the Weibull factor $W_{L_d} = 1$. It can be seen that the analytical results match well the simulation results under different load conditions (λ_d). Figure 6.3(a) - Figure 6.3(c) further illustrate the cases with a heavy-tailed data call size, i.e., $0 < W_{L_d} < 1$. Similarly, the analytical results agree with the simulation results, except that the voice and data call blocking probabilities are slightly overestimated when $W_{L_d} \leq 0.3$. This is due to the increase of heavy-tailedness with a small W_{L_d}. In our analytical model given in Section 6.2, we assume that the QoS of data calls in the WLAN is insensitive to the data call size distribution under the PS service discipline. Due to the varying WLAN capacity, the insensitivity is impaired and the call-level QoS may improve when a greater variability is induced with the heavy-tailed call size [106]. Nonetheless, the insensitivity is expected to retain when the call blocking probabilities are sufficiently small. For example, as seen in Figure 6.3, with a relatively light traffic load and a smaller data call blocking probability, the gap

between the analytical results and simulation results when $W_{L_d} \leq 0.3$ is much smaller. As the system is usually designed to ensure call blocking probabilities in the order of 10^{-3} - 10^{-2}, the analytical model in Section 6.2 is valid for the following performance analysis.

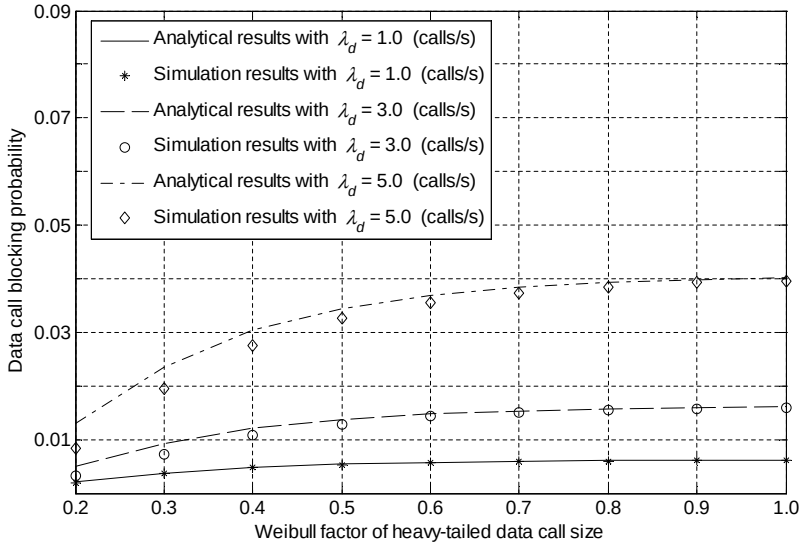


(a)

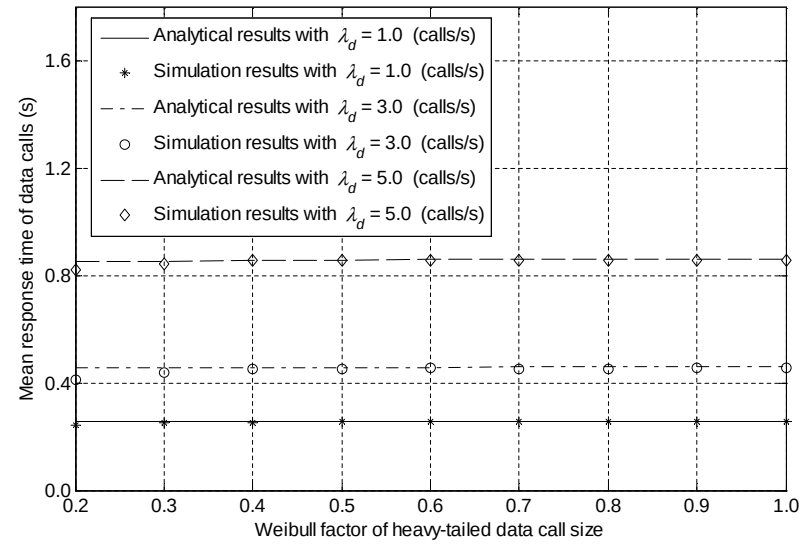
Figure 6.3: Analytical and simulation results of voice and data call QoS versus Weibull factor (W_{L_d}) for a heavy-tailed data call size with $\lambda_d = 1.0$, 3.0, and 5.0 (calls/s), respectively. (a) Voice call blocking probability (B_v).

6.3.2 Impact of data call size threshold

As discussed in Section 6.2.2, the data size threshold Φ_d should be properly determined to improve user-perceived QoS. Figure 6.4(a) - Figure 6.4(c) show the impact of Φ_d on voice and data call blocking probabilities (B_v and B_d, respectively) and mean data response time ($\overline{T}_d$) in different load conditions (λ_d). It is observed that B_v, B_d, and $\overline{T}_d$ only slightly decrease with Φ_d when Φ_d is relatively small. After a certain threshold (say,



(b)

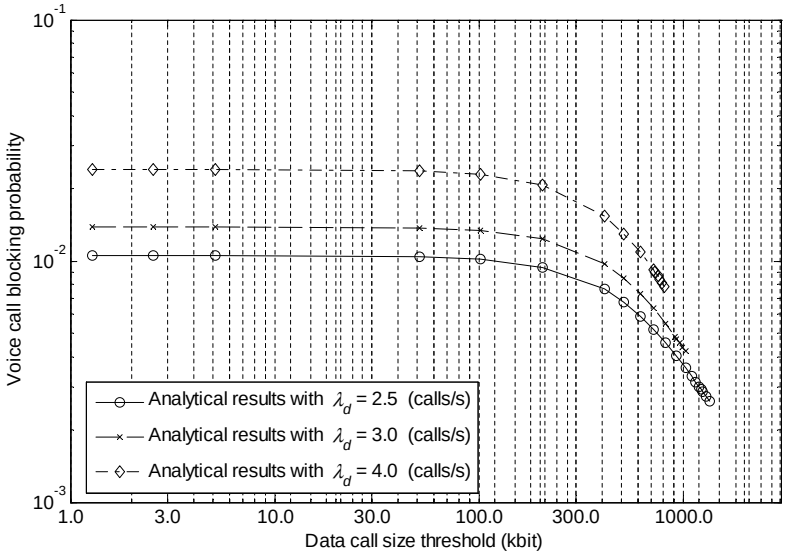


(c)

Figure 6.3: Analytical and simulation results of voice and data call QoS versus Weibull factor (W_{L_d}) for a heavy-tailed data call size with $\lambda_d = 1.0$, 3.0, and 5.0 (calls/s), respectively. (b) Data call blocking probability (B_d). (c) Mean data response time ($\overline{T}_d$).

$\Phi_d = 102.4$ kbits), B_v and B_d begin to decrease faster with Φ_d. When Φ_d is sufficiently large (e.g., $\Phi_d \geq 640.0$ kbits), $\overline{T}_d$ even increases exponentially with Φ_d. The phenomena observed in Figure 6.4 can be explained as follows. First, the explosive increase of $\overline{T}_d$ with a large value of Φ_d is due to congestion in the cell. As seen from (6.2), more data traffic load is assigned to the cell when Φ_d is larger. Due to a small cell bandwidth and high occupancy by voice calls, the data call performance is degraded substantially if the cell is overloaded. On the other hand, when Φ_d is relatively small, the decrease of $\overline{T}_d$ with Φ_d is attributed to the fact that the cell bandwidth unused by voice traffic can be efficiently utilized by small-size data calls under the SRPT. When Φ_d is sufficiently small to meet the underload condition, the larger the value of Φ_d, the more the data calls of a small size that can be assigned to the cell. Under the SRPT, the small-size data calls in the cell will not stay in the system for such a long duration as in the case where the bandwidth is shared among data calls of a large size in a PS manner.

Figure 6.5(a) - Figure 6.5(c) further demonstrate the impact of Φ_d with different heavy-tailedness degrees of data call size. We vary the shape parameter α_d ($0 < \alpha_d \leq 1$) in (2.8) and select the scale parameter β_d accordingly to keep the same mean value $\overline{L}_d$. Let the Weibull factor $W_{L_d} = \alpha_d$ denote the degree of heavy-tailedness. The smaller the value of W_{L_d}, the heavier the tail of the distribution of data call size. It can be seen in Figure 6.5(a) and Figure 6.5(b) that voice and data call blocking probabilities (B_v and B_d, respectively) decrease with Φ_d more slowly if W_{L_d} is smaller. When the data size threshold Φ_d is relatively small, B_v even increases with a larger W_{L_d}. On the other hand, as shown in Figure 6.5(c), the mean data response time $\overline{T}_d$ first slowly decreases with Φ_d until a sufficiently large Φ_d leads to an explosive increase of $\overline{T}_d$ due to system overload. In contrast to Figure 6.4(c) with an exponentially distributed data call size, the reduction of $\overline{T}_d$ with Φ_d is more evident in the heavy-tailed case. For a smaller W_{L_d} (say, 0.2), $\overline{T}_d$ decreases more slowly and can achieve an even smaller lower bound. This is due to the mice-elephants property of heavy-tailed distributions. A smaller W_{L_d}



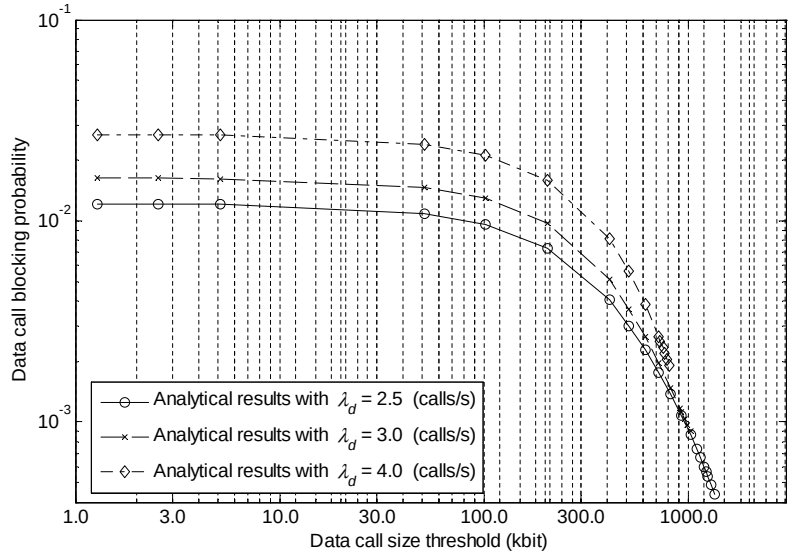
(a)

Figure 6.4: Voice and data call QoS versus data size threshold (Φ_d) with an exponentially distributed data call size ($W_{L_d} = 1.0$) and different load conditions of $\lambda_d = 2.5,\ 3.0$, and 4.0 (calls/s), respectively. (a) Voice call blocking probability (B_v).

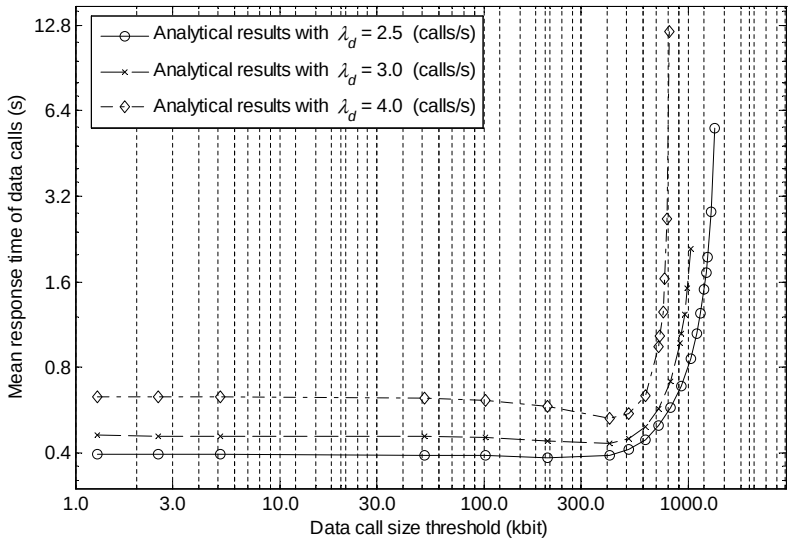
(i.e., a higher level of heavy-tailedness) implies that there is a larger fraction of even shorter data calls and that less data calls have a much larger size. Given the same size threshold Φ_d, more data calls can then be efficiently served under the SRPT in the cell. As a result, a smaller $\overline{T}_d$ is achievable with an appropriate size threshold. From Figure 6.4 and Figure 6.5, we can conclude that the load conditions and traffic characteristics should be properly incorporated in determining the data size threshold.

6.3.3 Performance improvement with the proposed scheme

To evaluate the effectiveness of the size-based load sharing scheme, we compare its performance with those of the admission schemes with service-differentiated assignment and randomized assignment, discussed in Chapter 4 and Chapter 5, respectively. Fig-

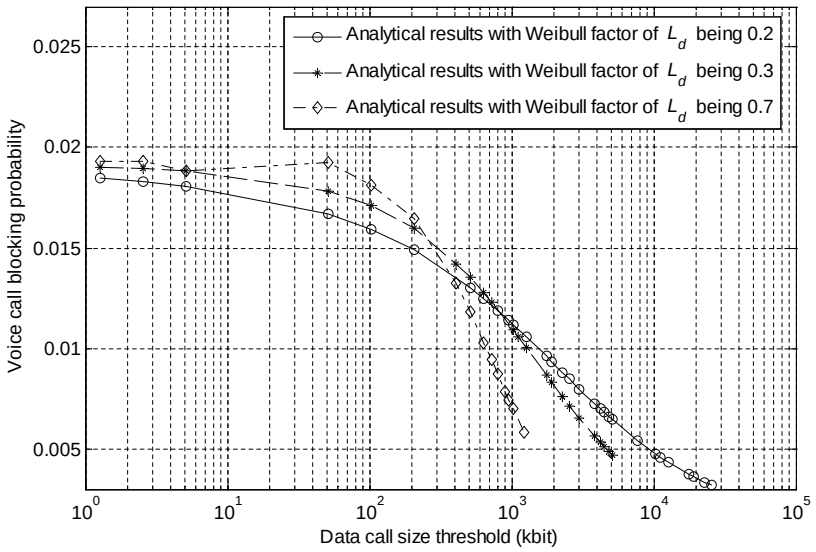


(b)



(c)

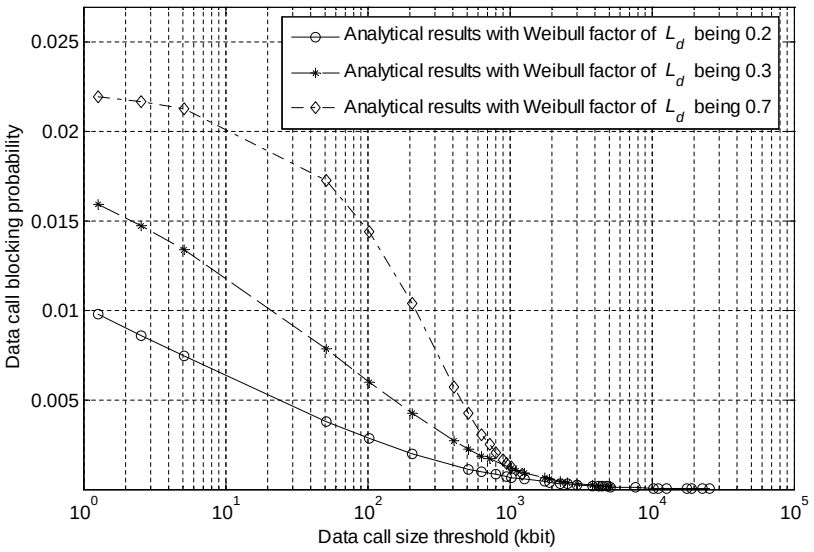
Figure 6.4: Voice and data call QoS versus data size threshold (Φ_d) with an exponentially distributed data call size ($W_{L_d} = 1.0$) and different load conditions of $\lambda_d = 2.5,\ 3.0$, and 4.0 (calls/s), respectively. (b) Data call blocking probability (B_d). (c) Mean data response time ($\overline{T}_d$).



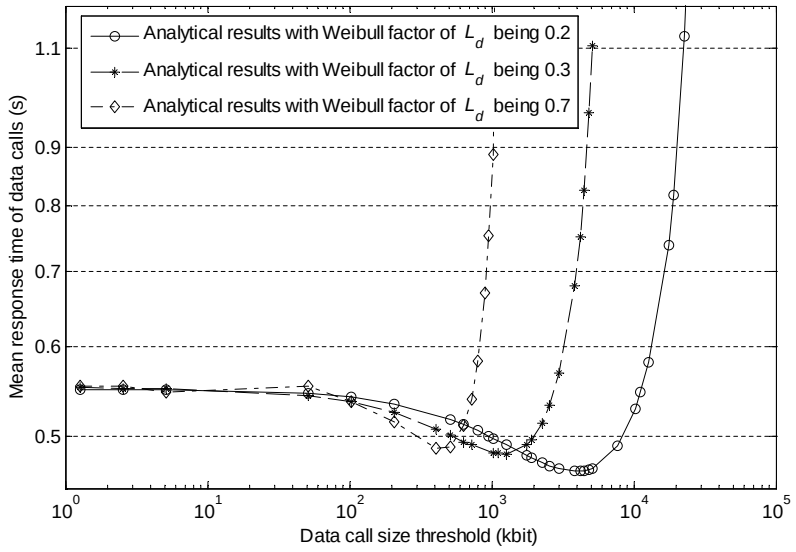
(a)

Figure 6.5: Voice and data call QoS versus data size threshold (Φ_d) with mean data call arrival rate $\lambda_d = 3.6$ (calls/s) and different heavy-tailedness of data call size, i.e., $W_{L_d} = 0.2,\ 0.3$, and 0.7, respectively. (a) Voice call blocking probability (B_v).

ure 6.6(a) - Figure 6.6(c) show the performance of the three schemes in terms of voice and data call blocking probabilities (B_v and B_d, respectively) and mean data response time ($\overline{T}_d$), respectively. Significant performance improvement is observed with the size-based scheme. For example, in the case of $\lambda_d = 3.6$ (calls/s), B_v of the size-based scheme is 71.4% smaller than that of the randomized scheme, while B_d is reduced by more than 85.6% and $\overline{T}_d$ is 46.8% lower. A performance gain of 45.4% and 74.8% is achieved by the size-based scheme with respect to the service-differentiated scheme for B_v and B_d, respectively, although $\overline{T}_d$ of the two schemes is very close. In some cases, $\overline{T}_d$ of the scheme with service-differentiated assignment is even slightly lower than that of the size-based load sharing scheme. However, this low mean data response time of the service-differentiated scheme is achieved at the expense of much higher call block-



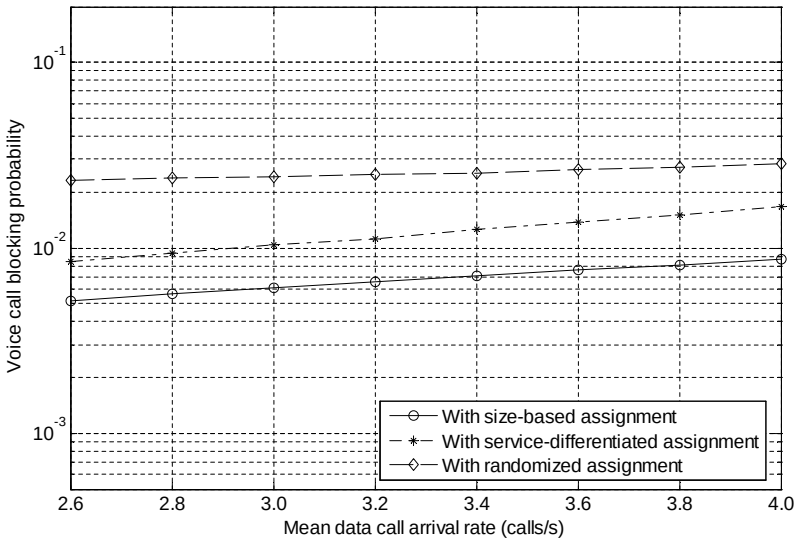
(b)



(c)

Figure 6.5: Voice and data call QoS versus data size threshold (Φ_d) with mean data call arrival rate $\lambda_d = 3.6$ (calls/s) and different heavy-tailedness of data call size, i.e., $W_{L_d} = 0.2,\ 0.3$, and 0.7, respectively. (b) Data call blocking probability (B_d). (c) Mean data response time ($\overline{T}_d$).

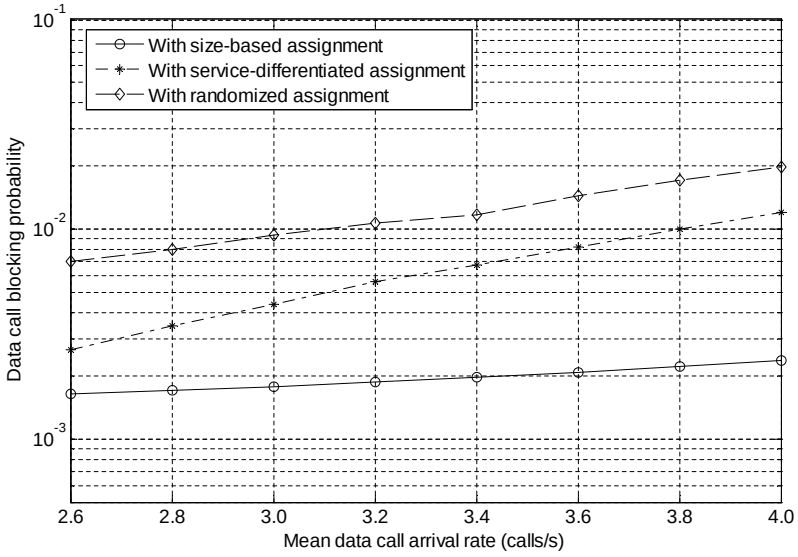
ing probabilities B_v and B_d. The size-based load sharing scheme still outperforms the other two schemes.



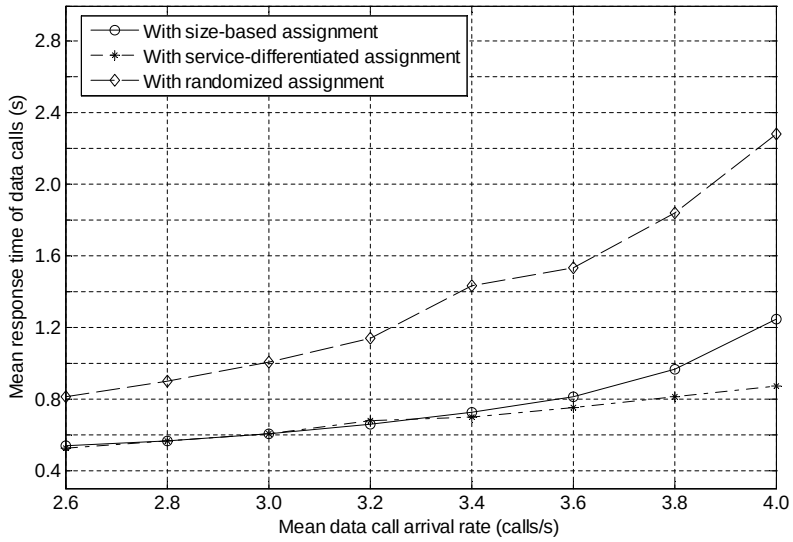
(a)

Figure 6.6: Performance of different load sharing schemes versus mean data call arrival rate (λ_d) with an exponentially distributed data call size ($W_{L_d} = 1.0$). (a) Voice call blocking probability (B_v).

Figure 6.7(a) - Figure 6.7(c) show the performance of the three schemes with different Weibull factors W_{L_d}, i.e., different heavy-tailedness degrees of the data call size. It can be seen that an even larger performance gain is achievable with the size-based scheme for B_d and $\overline{T}_d$ when W_{L_d} is smaller, i.e., the data call size is distributed with a heavier tail. For example, when $W_{L_d} = 0.2$, B_d of the size-based scheme is more than 95% smaller than those of the other two schemes, while the reduction is around 87.7% when $W_{L_d} = 0.8$. Similarly, when W_{L_d} decreases from 0.8 to 0.2, the reduction of $\overline{T}_d$ with respect to the randomized scheme increases from 49.6% to 79.7%. In comparison with the service-differentiated scheme, the size-based scheme reduces $\overline{T}_d$ by 7.7% when



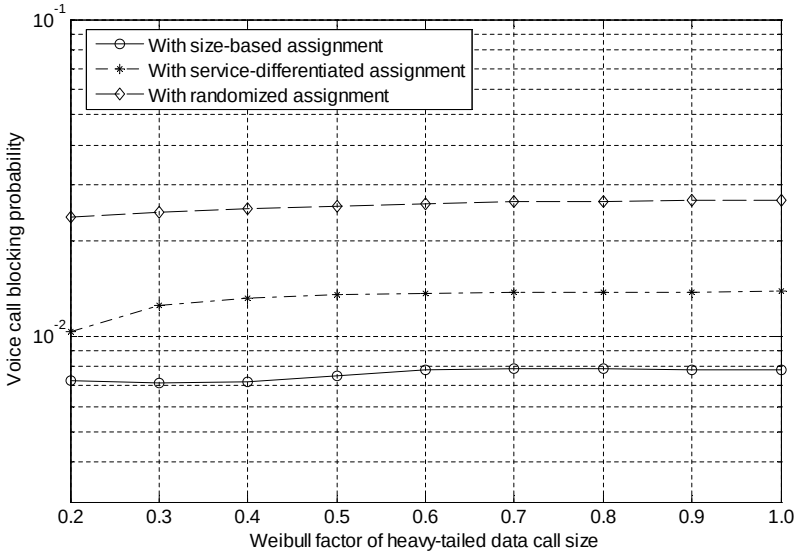
(b)



(c)

Figure 6.6: Performance of different load sharing schemes versus mean data call arrival rate (λ_d) with an exponentially distributed data call size ($W_{L_d} = 1.0$). (b) Data call blocking probability (B_d). (c) Mean data response time ($\overline{T}_d$).

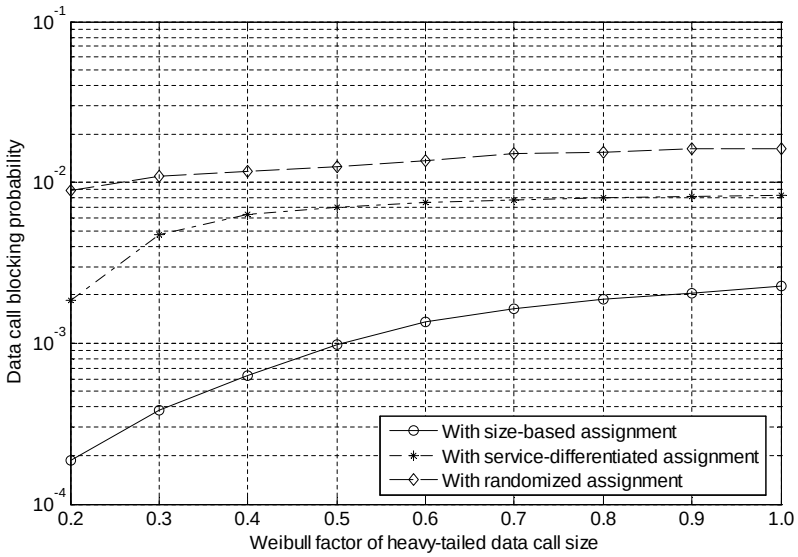
$W_{L_d} = 0.8$ and by 32.8% when $W_{L_d} = 0.2$. In addition, the reduction of $\overline{T}_d$ with W_{L_d} is due to the much higher call blocking probabilities, which restrict the total admissible traffic load to share the bandwidth.



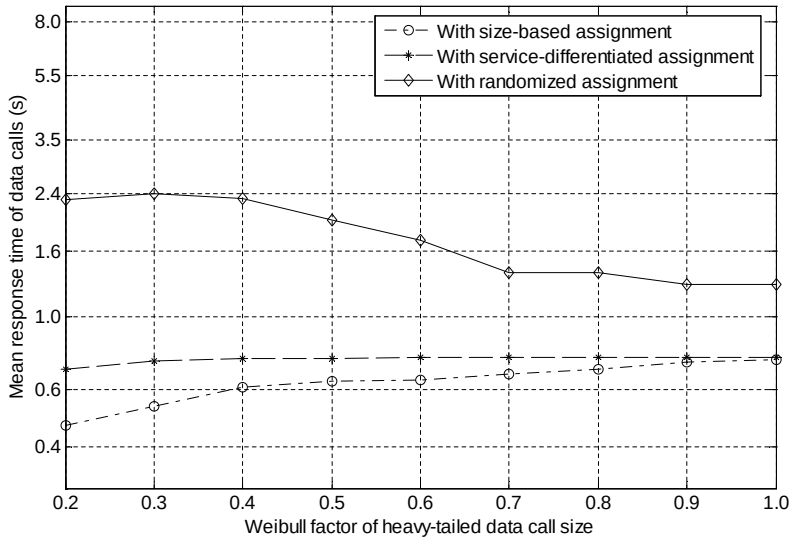
(a)

Figure 6.7: Performance of different load sharing schemes under a heavy-tailed data call size versus Weibull factor (W_{L_d}) with mean data call arrival rate $\lambda_d = 3.6$ (calls/s). (a) Voice call blocking probability (B_v).

The significant performance gain observed in Figure 6.6 and Figure 6.7 lies in the fact that the size-based load sharing scheme not only takes advantage of the complementary QoS of the integrated systems in initial call assignment, but also exploits vertical handoff in call reassignment to maximize the multiplexing gain. Moreover, the data size threshold can be appropriately determined with the approach given in Section 6.2, which effectively takes into account the load conditions and heavy-tailedness of data call size. Nonetheless, the size-based scheme requires that the data call size be known *a priori* via session signaling. The signaling and control overhead for dynamic



(b)



(c)

Figure 6.7: Performance of different load sharing schemes under a heavy-tailed data call size versus Weibull factor (W_{L_d}) with mean data call arrival rate $\lambda_d = 3.6$ (calls/s). (b) Data call blocking probability (B_d). (c) Mean data response time ($\overline{T}_d$).

vertical handoff may increase the implementation complexity.

6.3.4 Overload protection via SRPT scheduling

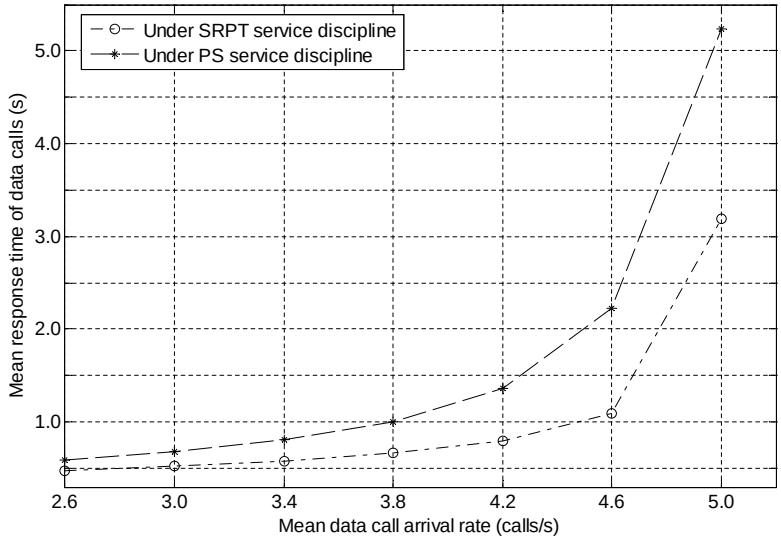


Figure 6.8: Mean response time ($\overline{T}_d$) of data calls with $W_{L_d} = 1.0$ under SRPT or PS service discipline applied in the cell.

As discussed in Section 6.1, data calls in the cell are served under the SRPT, which can be enabled by the centralized resource allocation and benefit the system with the best performance achievable. The advantage of SRPT is particularly more evident in system overload when it is more challenging for the cell of small bandwidth to provide QoS guarantee. Figure 6.8 compares the mean data response time $\overline{T}_d$ when the SRPT or PS are applied respectively to serve data traffic in the cell. It can be seen that, when system overload occurs, $\overline{T}_d$ under the SRPT is significantly reduced in contrast to that under the PS. At the same time, both service disciplines exhibit similar voice and data call blocking probabilities. It is known that there exists a trade-off between $\overline{T}_d$ and call blocking probabilities. That is, when more calls are admitted and share a

given bandwidth, $\overline{T}_d$ increases although call blocking probabilities decrease. Hence, the observation of a significantly reduced $\overline{T}_d$ and close call blocking probabilities implies a higher resource utilization under the SRPT.

6.4 Summary

In this chapter, we have proposed a new load sharing scheme for voice and elastic data services in the cellular/WLAN integrated network. It takes into account both initial call assignment and call reassignment via dynamic vertical handoff. The heavy-tailed data call size is effectively exploited with the size-based call assignment and SRPT scheduling. An effective analytical model is developed to determine the data size threshold. It is observed from numerical results that the new size-based scheme significantly outperforms two previously proposed schemes with service-differentiated assignment and randomized assignment.

Chapter 7

Multi-Service Load Sharing with Two-Way Reassignment

In this chapter, we extend our previous works on load sharing for cellular/WLAN interworking to a multi-service scenario. Video streaming is further considered in addition to conversational voice service and interactive data service. The main contributions of this work are as follows:

- A multi-service load sharing scheme is proposed for voice, data, and video streaming traffic in the integrated network. The service-differentiated call assignment strategy in Chapter 4 is extended to make a better use of the high WLAN throughput with scalable layered video traffic. Further, two-way call reassignment via vertical handoff is considered to migrate long-lived voice and video streaming calls between the coupled cell and WLAN.

- Encoding bit rate adaptation is applied to video streams to cope with transmission rate variation and protect against buffer underflow. Simulation analysis is performed to evaluate the impact of rate-adaptiveness of video streaming on interworking performance.

7.1 Video Streaming Service

As discussed in Section 2.4.3, there are many appealing services in the streaming class such as video streaming, which becomes very popular in wireless networks. For example, the 3GPP packet-switched streaming (PSS) is one of the most mature streaming standards in the wireless communications industry [115]. Although video streaming is also real-time service similar to the conversational class, it is essential for video streaming to preserve playback continuity and smoothness instead of maintaining a stringent low delay. Hence, a certain range of bandwidth adaptation is acceptable for streaming services, while some conversational-class services may require a constant bandwidth. Similar to interactive data services such as Web browsing, video clips for on-demand streaming are pre-stored in the media server. The saturated elastic traffic can be dynamically adapted to available bandwidth. Session-related information (such as video clip duration, total amount of data to stream and bit variability) can be known *a priori* and exploited for QoS provisioning.

In order to guarantee streaming service quality, it is imperative for the elements involved in a streaming session to work cooperatively, which are the network resource controller, media server for storing and distributing video clips, and user equipment for video playback [116]. From the perspective of the network resource controller (e.g., the base station), it is not efficient to allocate dedicated resources to video sessions with bursty traffic. The bandwidth allocation should be adapted (e.g., via scheduling scheme) to feedback from user equipment and network measurement. At the receiver client side, adaptive media playout (AMP) can be applied to improve the streaming service quality. By default, a video stream is played back at the same rate as the encoding frame rate. Nonetheless, the playout rate can be adjusted within a small range without being noticeable [117]. Even for the challenging audio sequence, although an increase or decrease of the sampling rate results in an audible change to the pitch, time scale modification technique can shorten or elongate the audio stream while preserving

the pitch. Subjective tests have shown that it is often unnoticeable to slow down the playout rate of video and audio up to 25%. Thus, the initial data consumption rate can be reduced to shorten the pre-roll time without increasing the underflow probability.

Although the above adaptive techniques at the network resource controller and user equipment can counter against traffic burstiness and bandwidth variation, high implementation complexity may be involved. In this work, we only consider rate adaptation at the media server to adapt the encoding bit rate of video streams to the transmission rate. For stored video streaming, multirate encoding can be done prior to transmission and the media server switches the bitstreams of different rates (bitstreaming switching) according to the available transmission rate. Another solution is stream thinning, which removes nonreferenced video frames to adjust video encoding rate and in turn the total amount of data to transmit [115]. MPEG-4 introduces a fine-granular scalability (FGS) video encoding scheme, in which the video sequence is encoded into a base layer and an enhancement layer. Then, a target encoding bit rate is achieved by truncating the enhancement layer.

The encoding rate adaptation is especially crucial for networks with large bandwidth variations, e.g., due to heterogeneous access technologies, network congestion, channel fading, etc. An appropriate rate adaption can protect against buffer underflow and provide interrupt-free playback. The average decoding rate of video streams also needs to be maximized for good video quality. At the same time, rate variation between consecutive video segments should be minimized to maintain smooth video quality [118]. To achieve the above objective, the rate adaptation should take into account information such as playback buffer occupancy at the receiver client, available network bandwidth, radio link conditions, and potential handoff. Take the 3GPP PSS as an example. Figure 7.1 shows the protocol stack of media streaming services in UMTS. Video encoded frames are segmented into packets of the real-time transport protocol (RTP), which delivers the media streams over the user datagram protocol (UDP) with the aid of the

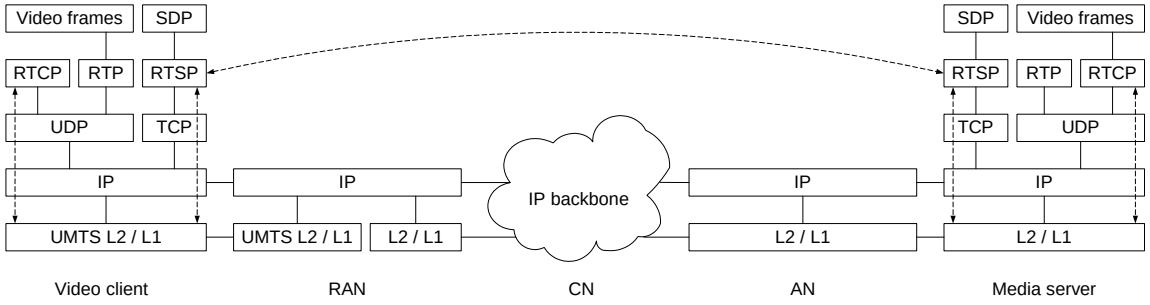


Figure 7.1: Session information collection with 3GPP PSS.

real-time transport control protocol (RTCP). The session description protocol (SDP) specifies the presentation description such as media encoding information, while the real-time streaming protocol (RTSP) provides application-layer signaling for the establishment and delivery of streams between the media server and client. The RTSP signaling enables feedback from the client for wireless link characteristics, maximum playback buffer size, desired media content (in time) to be sustained in the buffer, etc. More dynamics of the playback buffer (e.g., the remaining playout time of the buffered media) can also be derived from RTCP signaling to assist encoding rate adaptation.

7.2 Multi-Service Load Sharing Scheme

As discussed in Section 2.4, the conversational, streaming, and interactive services differ in traffic characteristics and QoS requirements. The multi-service traffic load in the overlay area should be properly shared between the integrated systems to maximize the interworking effectiveness.

7.2.1 Initial call assignment with service differentiation

As discussed in Section 4.2, new voice calls in the double-coverage area are preferably assigned to the cell so as to benefit from the efficient voice support. On the other hand,

for on-demand video streaming service and data service such as Web browsing, the video clips or data files are pre-stored in the media server or Web server. The saturated elastic traffic can be adapted to available bandwidth. Moreover, interactive data sessions exhibit the on-off dynamics shown in Figure 2.2, whereas video traffic is much burstier due to compression algorithms of video codecs. Although the mean response time of data calls needs to be bounded, video streaming has a more stringent requirement for underflow ratio. According to the WLAN capacity analysis in Section 4.1, the number of admissible saturated calls is very limited to satisfy the stability constraint. As a result, a high throughput is achievable to admitted calls. The high throughput can guarantee a very low mean response time to data calls, and can also support scalable video traffic with a high bit rate for good video quality. Since the mean response time only needs to be sufficiently small to ensure fluent interaction, the high WLAN throughput offers more merits to video streaming. Hence, it is reasonable to assign video streaming calls first to the WLAN to request admission, while interactive data sessions with elastic traffic and less stringent QoS requirements are preferably assigned to the cell. Thus, a small collision probability is maintained in the WLAN for QoS guarantee to real-time voice and video streaming calls.

In this work, we assume that FGS encoding [119] is applied to compress a video stream to any arbitrary bit rate R_s over a range $[R_l, R_u]$, given by

$$R_s = R_l + \theta_s \cdot (R_u - R_l) = R_B + \theta_s \cdot R_E \tag{7.1}$$

$$R_l = R_B, \qquad R_u = R_B + R_E, \qquad 0 \le \theta_s \le 1 \tag{7.2}$$

which consists of a base layer of bit rate R_B and an adaptive enhancement layer of bit rate $\theta_s \cdot R_E$. The encoding bit rate R_s can be adapted to available network bandwidth (transmission rate) and client buffer occupancy. As the cell bandwidth is preferably allocated to voice traffic, video streams in the cell are delivered with base-layer encoding to admit more voice calls. On the other hand, rate adaptation is necessary and feasible for video streams in the WLAN to effectively utilize the high throughput. A simple

rate adaptation algorithm is proposed in [120]. The rationale behind the algorithm is that video bit rate is adjusted in a manner to maintain the playback buffer occupancy (in media time) above a given threshold, referred to as *target protection time* in 3GPP standards. The rate adaptation is performed for video segments of a fixed playback duration. Nonetheless, it becomes invalid for the WLAN because the transmission rate to serve a video stream fluctuates with contending calls during the streaming of a video segment. There may be a conflict to use instantaneous playback buffer occupancy and average transmission rate in deriving the target encoding rate. Also, the adaptation strategy is rather conservative as it only attempts to avoid rate upshift prior to attaining the buffer occupancy threshold. Hence, we extend the algorithm to evaluate the impact of rate-adaptive video streaming service on cellular/WLAN interworking.

Let Q^* denote the target protection time for playback buffer occupancy. The encoding bit rate $R_s(t_k)$ of a video stream at time t_k is determined according to current playback buffer occupancy $Q(t_k)$ and available transmission rate $\xi_s^w(t_k)$, which depends on the numbers of active voice, data, and video streaming calls. As $\xi_s^w(\cdot)$ varies with traffic load fluctuation, we estimate the expected interval between adjacent call arrivals or departures, denoted by Ω_t. To maintain the target buffer occupancy, it is required that

$$\frac{\xi_s^w(t_k)\,\Omega_t}{R_s(t_k)} + Q(t_k) - \Omega_t \geq Q^* \tag{7.3}$$

$$\text{i.e.,} \quad \xi_s^w(t_k) \geq R_s(t_k)\,\left[1 + \frac{Q^* - Q(t_k)}{\Omega_t}\right]. \tag{7.4}$$

If the current buffer occupancy $Q(t_k)$ is below the target level Q^*, the encoding bit rate should be degraded to

$$R_s(t_k) \leq \xi_s^w(t_k) \Big/ \left[1 + \frac{Q^* - Q(t_k)}{\Omega_t}\right] \tag{7.5}$$

in order to avoid possible buffer underflow. In contrast, if the desired protection level is achieved, the encoding rate can also be upshifted to improve the media presentation quality. According to the media information obtained during session establishment,

application requirements are mapped to QoS attributes (e.g., bounded underflow ratio) and corresponding bandwidth requirement in the range of $[\Psi_l, \Psi_u]$. If $Q(t_k)$ is further above the target level Q^* so that $1 + \frac{Q^* - Q(t_k)}{\Omega_t} < 0$, we have $R_s(t_k) = R_u$ when $\xi_s^w(t_k) \geq \Psi_u$. If $\Psi_l \leq \xi_s^w(t_k) \leq \Psi_u$, $R_s(t_k) = R_B + \theta_s \cdot R_E$, where θ_s should be properly chosen to bound the underflow ratio. It varies with the specific video codec configuration and video clip content, which is out of the scope of this work. Finally, the adapted bit rate $R_s(t_k)$ should fit within the feasible range $[R_l, R_u]$ by setting

$$R_s(t_k) \Leftarrow \min\{R_u, \max\{R_l, R_s(t_k)\}\}. \tag{7.6}$$

In Chapter 6, we have observed that data call QoS under the SRPT can be significantly improved over that under the PS. In this work, we further consider the on-off dynamics of interactive sessions shown in Figure 2.2. In the cellular network, the remaining bandwidth unused by conversational and streaming sessions can be allocated to data traffic to exploit data traffic elasticity. Then, the maximum number of interactive sessions admitted in the cell (denoted by N_d^c) should be restricted to bound the mean data call transfer delay. To enable tractable analysis, we assume that the total bandwidth available to data traffic is equally shared among active data calls of ongoing interactive sessions. That is, data calls are served under the PS discipline. The application of the SRPT in this scenario is left for further work.

To determine N_d^c adaptively with time-varying traffic load, we periodically estimate the mean arrival rate of interactive sessions to the cell and the average cell bandwidth available to data traffic, denoted by $\overline{\lambda}_d^c(t_k)$ and $\overline{C}_d^c(t_k)$, respectively, for the k^{th} control period. Given the interactive session structure in Figure 2.2, the distribution of data call interarrival time is approximately exponential when the session arrival intensity is large [121]. The mean data call arrival rate is then $\overline{M}_d\, \overline{\lambda}_d^c(t_k)$, where $\overline{M}_d$ is the average number of data calls in an interactive data session. Hence, we have the offered data traffic load to the cell

$$\rho_d^c(t_k) = \frac{\overline{M}_d\, \overline{\lambda}_d^c(t_k) \cdot \overline{L}_d}{\overline{C}_d^c(t_k)}. \tag{7.7}$$

As shown in Section 6.1, when the system is overloaded with $\rho_d^c \geq 1$ (the denotation of the control period in the following is omitted for simplicity), the mean response time of data calls $\overline{T}_d^c$ increases almost linearly with N_d^c at a rate $\frac{\rho_d^c}{\overline{M}_d \overline{\lambda}_d^c} = \frac{\overline{L}_d}{\overline{C}_d^c}$. Thus, the largest N_d^c satisfying the delay bound Q_T is given by

$$N_d^c = Q_T \frac{\overline{C}_d^c}{\overline{L}_d}. \tag{7.8}$$

For an underload case with $\rho_d^c < 1$, N_d^c is configured to be the largest value satisfying $\overline{T}_d^c \leq Q_T$. Following an $M/G/1/K - PS$ queueing model, $\overline{T}_d^c$ can be obtained as

$$\overline{T}_d^c = \frac{(\rho_d^c)^{N_d^c+1}(N_d^c \rho_d^c - N_d^c - 1) + \rho_d^c}{\overline{M}_d \overline{\lambda}_d^c \cdot (1 - (\rho_d^c)^{N_d^c})(1 - \rho_d^c)}. \tag{7.9}$$

7.2.2 Two-way call reassignment via dynamic vertical handoff

In addition to the above initial call assignment with service differentiation, ongoing calls can also be reassigned to the overlay cell or WLAN via dynamic vertical handoff. It is known that interactive data calls are short-lived as the mean response time is bounded within several seconds to maintain fluent interaction. Hence, we only consider the long-lived voice and video streaming calls in the reassignment. Figure 7.2 illustrates the flowchart of load sharing with call assignment/reassignment.

First, call completion or handoff out of the cell may release sufficient free bandwidth to accommodate more voice calls in the cell. Then, certain ongoing voice calls carried by the WLAN can be handed over to the cell. Second, video streaming calls may overflow to the cell if the WLAN is rather congested at the initial admission time. However, video streams can only be delivered with the base-layer encoding in the cell due to bandwidth limitation. Thus, if call departures in the WLAN result in spare capacity for more streaming calls, vertical handoff is triggered to reassign some ongoing streaming calls from the cell to the WLAN. Video streams with the largest remaining data to transfer are selected first for the migration. The involved streaming control signaling

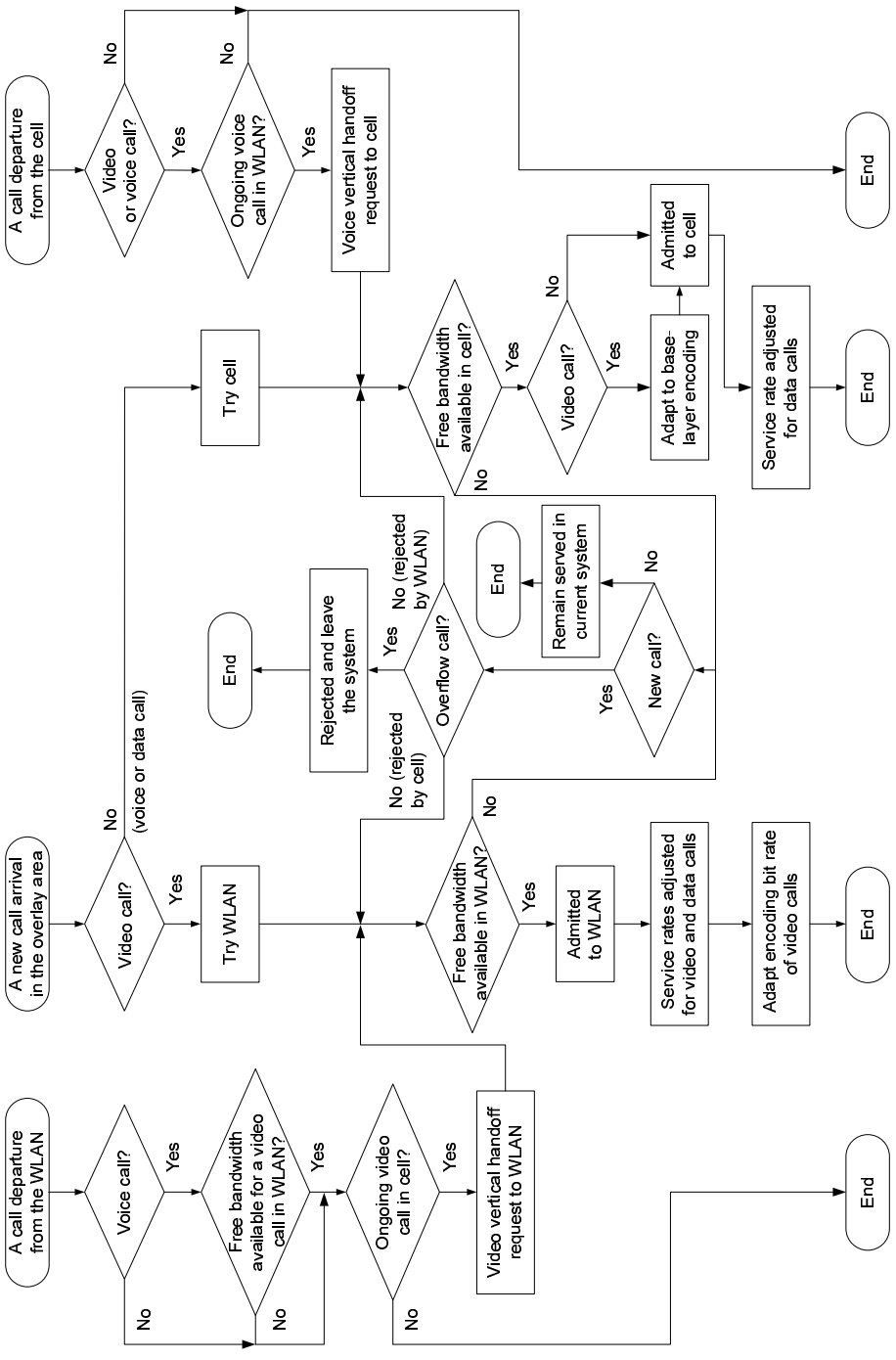


Figure 7.2: Flowchart of the multi-service load sharing scheme.

can provide basic information about the streaming call and the states of the client and network. For example, the remaining video clip size to transfer can be derived from the RTCP signaling. In this way, the real-time voice and video streaming calls are served in the cell and WLAN, respectively, with preference.

7.3 Simulation Results and Discussion

In this section, we evaluate the performance of the proposed multi-service load sharing scheme, which employs service-differentiated initial call assignment and two-way call reassignment via vertical handoff. The performance is compared with that of the WLAN-first scheme analyzed in [16]. For the WLAN-first scheme, all service calls are preferably assigned to the available WLAN during admission. Only calls rejected by the WLAN overflow to the overlay cell to request admission.

As seen in Section 2.4, video streaming service deals with very complex video traffic varying with specific codecs and contents. The rate adaptation and rebuffering due to buffer underflow further complicates the performance evaluation. Moreover, the two integrated systems and three service classes involve too many analysis dimensions to evaluate the QoS analytically. Hence, the following results are mainly based on computer simulation. As observed in [122, 123], it is extremely slow for simulations with a heavy-tailed workload to converge to steady states. This is especially true for power-tailed cases, such as Pareto distributions with a shape parameter less than 2. To improve the simulation confidence, we follow the approaches suggested in [123] and use truncated Pareto distributions given in (2.10) and (2.14) instead of unbounded ones. Also, separate random seeds and generation streams are used for different random variables to guarantee independence. Given in Table 7.1 are the simulation parameters, which are selected by referring to the evaluation specifications for UMTS [71, 124] and cdma2000© systems [125], video statistics in [85, 87], and the video source model in [88].

Figure 7.3 shows the voice call blocking probability of the WLAN-first scheme.

Table 7.1: System parameters for simulations.

Symbol	**Value**	**Definition**	**Symbol**	**Value**	**Definition**
C^c	2.0	Cell bandwidth (Mbit/s)	C^w	11.0	Physical channel rate of WLAN (Mbit/s)
$\overline{T}_v$	180.0	Mean voice call duration (s)	$R^c_{b,v}$	12.2	Encoding bit rate of voice calls (kbit/s)
f_s	10	Constant frame rate of video clips (frames/s)	$\overline{L}$	400.0	Mean video frame size (byte)
σ_L	272.8	Standard deviation of video frame size (byte)	β_s/α_s	0.984	Autocorrelation factor of video frame size
$\tilde{T}_s$	2.0	Median video clip duration (min)	l_s	1.0	Minimum video clip duration (min)
u_s	60.0	Maximum video clip duration (min)	γ_s	1.000161	Shape parameter for distribution of video clip duration
Q_U	0.1	Bound for average underflow ratio of video streaming calls	S_p	10.0	Maximum pre-roll time of video streaming calls (s)
S_f	5.0	Rebuffer video data in media duration (s)	Q^*	5.0	Target protection time for playback buffer (s)
R_l	32.0	Minimum video encoding bit rate (base layer) (kbit/s)	R_u	256.0	Maximum video encoding bit rate (kbit/s)
$\overline{M}_d$	5	Mean number of data calls in an interactive session	$\overline{S}_r$	5.0	Mean reading time in an interactive session (s)
$\overline{L}_d$	25.0	Mean data call size (kbyte)	σ_{L_d}	101.1	Standard deviation of data call size (kbyte)
l_d	4.5	Minimum data call size (kbyte)	u_d	2000	Maximum data call size (kbyte)
γ_d	1.1	Shape parameter of data call size distribution	Q_T	4.0	Bound for mean response time of data calls (s)

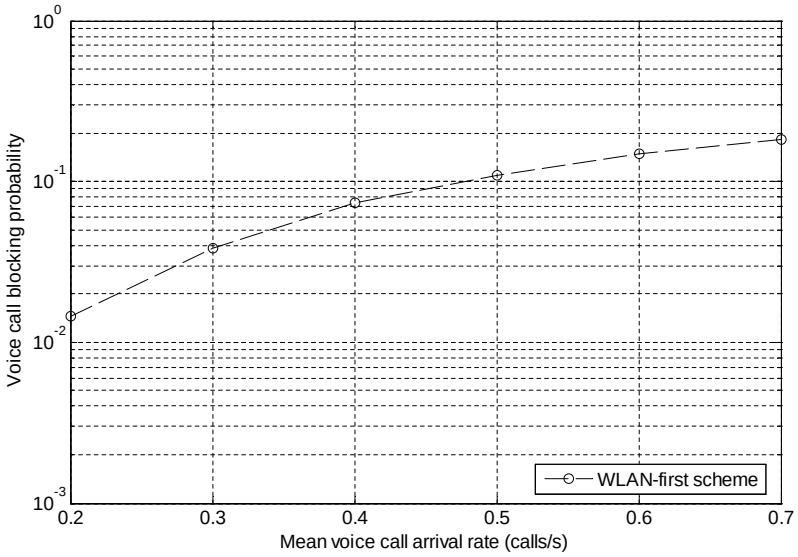


Figure 7.3: Voice call blocking probability of the WLAN-first scheme.

In each simulation round, around 10^6 voice call arrivals are generated. For the new multi-service load sharing scheme, no voice call blocking is observed, which indicates a negligibly small blocking probability. Actually, we see from the simulation results that the blocking probabilities of data and video streaming calls are also significantly reduced. The performance improvement validates the effectiveness of the new scheme, in which multi-service calls are jointly considered and properly distributed to the overlay cell and WLAN. Two-way call reassignment via dynamic vertical handoff further enhances the service quality by reassigning overflow traffic to the preferred network whenever spare capacity is available. Nonetheless, a higher implementation complexity can be induced as network state updates may be necessary to determine a good handoff timing.

Figure 7.4 compares the mean response time and blocking probability of data calls. For the new load sharing scheme, as no session blocking is observed among the 10^6 interactive data session arrivals of each simulation round, only the blocking probability

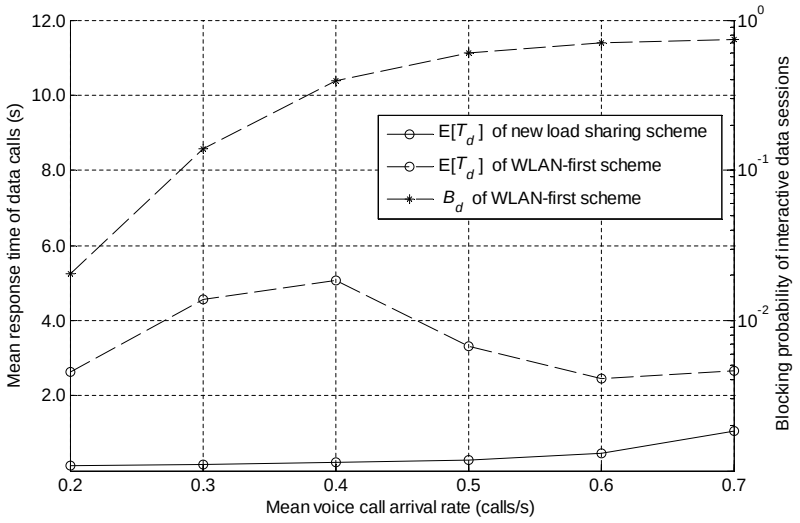
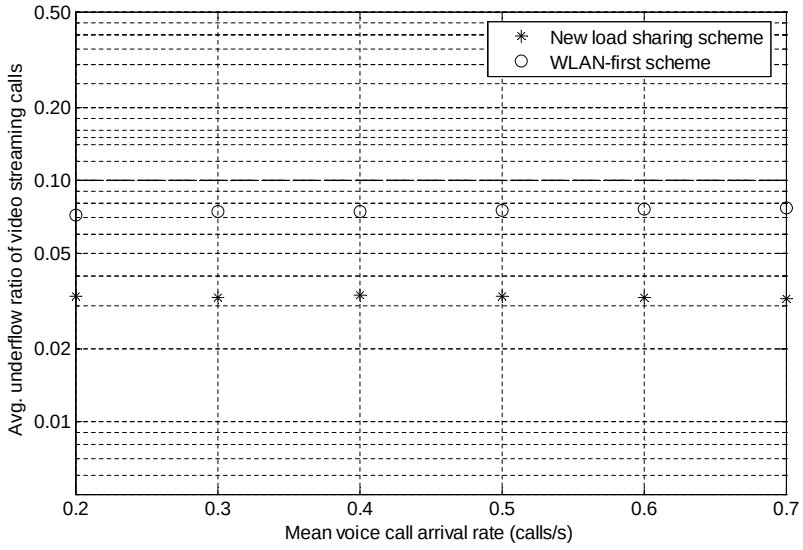


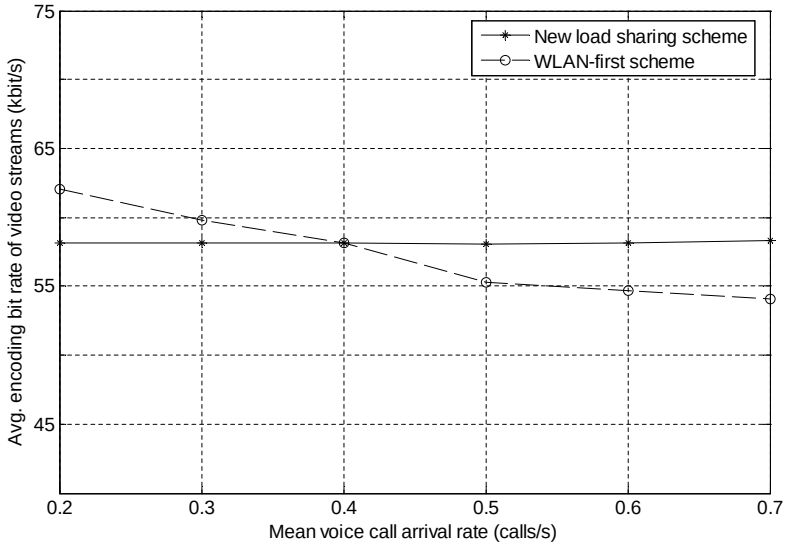
Figure 7.4: Mean data response time (E[T_d]) and blocking probability (B_d) of interactive data sessions of the new multi-service load sharing scheme and the WLAN-first scheme.

of the WLAN-first scheme is shown, which is much higher than the negligibly small blocking probability of the new scheme. Moreover, the new scheme guarantees a much lower mean response time for data calls. Different from the call assignment strategy discussed in Section 4.2, the new scheme preferably assigns data calls to the cell instead of the WLAN in presence of video streaming calls. Although both services have elastic traffic, video streaming calls are much more bandwidth demanding. Data calls with relatively loose QoS requirements can more effectively utilize the free bandwidth unused by real-time traffic in the cell of small bandwidth. The assignment of data calls to the cell is also constructive to maintaining a small collision probability in the WLAN. More spare capacity is then available in the WLAN to serve video streaming calls, which can fully utilize the high throughput to improve video presentation quality to a large extent.

Figure 7.5 illustrates the video streaming QoS in terms of average underflow ratio and encoding bit rate, where the latter is an indicator of video quality. As expected, the average underflow ratio of video streams is well bounded with the extended rate



(a)



(b)

Figure 7.5: Video streaming call QoS of the new multi-service load sharing scheme and the WLAN-first scheme. (a) Average underflow ratio of video streaming calls. (b) Average encoding bit rate of video streams.

adaptation algorithm. A larger average video encoding rate is achieved with the new scheme for a high traffic load, although it is slightly smaller than that of the WLAN-first scheme in a light load condition. This is because the high traffic load can result in severe contention in the WLAN without proper load sharing. The jeopardized WLAN throughput will not only degrade video stream quality but also lead to a large call blocking probability. Then, more video streaming calls overflow to the cell and are only provisioned a base-layer encoding rate. Consequently, in high load conditions, the overall average encoding bit rate of video streams is lower with the WLAN-first scheme. For the light load cases, the higher encoding rate of the WLAN-first scheme is achieved at the expense of a much larger blocking probability for video streaming calls. As the WLAN-first scheme cannot effectively share the multi-service traffic load between the cell and the WLAN, the numbers of real-time calls admitted in the WLAN need to be very restricted to maintain a small collision probability.

7.4 Summary

In this chapter, we have studied the load sharing for voice, data, and video streaming services in the cellular/WLAN integrated network. The proposed multi-service load sharing scheme takes into account the essential traffic characteristics such as data traffic elasticity and burstiness of video sources. The multi-service traffic load is effectively shared between the integrated systems by means of service-differentiated initial call assignment and two-way call reassignment via dynamic vertical handoff. The overall system performance is improved by the new scheme in comparison with that of the WLAN-first scheme. Call blocking probabilities and mean data response time are significantly reduced, while the average encoding bit rate of video streams is increased in high load conditions.

Chapter 8

Conclusions and Further Work

In this chapter, we summarize the main research results and discuss further work.

8.1 Major Research Results

The objective of this research is to investigate and develop resource allocation schemes for the cellular/WLAN integrated network to efficiently utilize the overall resources. An essential resource allocation technique for this heterogeneous wireless overlay network is load sharing, which effectively distributes multi-service traffic load across the integrated systems by means of initial call assignment with admission control and call reassignment via vertical handoff. In this way, efficient resource allocation is achievable by taking advantage of the overlay structure and complementary strength of underlying networks. Specifically, the main research results are summarized as follows:

- System modeling for cellular/WLAN interworking: A system model is developed for the integrated network to capture the essential characteristics and enable tractable analysis. Especially, the contention-based WLAN is modeled accurately to facilitate the investigation. Important QoS provisioning features of the WLAN are properly characterized and considered in developing resource alloca-

tion strategies.

- Analytical QoS evaluation approaches: To properly determine the admission and assignment parameters, the QoS metrics such as call blocking/dropping probabilities and mean data response time are evaluated analytically based on Markov processes and further simplified with moment generating functions. The analytical approaches take into account the unique characteristics of the integrated cell/WLAN cluster, such as the location-dependent user mobility, heavy-tailed data call size and highly variable user residence time within the WLAN.

- Call assignment and admission control schemes for load sharing: After carefully examining the characteristics of the heterogeneous overlay network, we have proposed and analyzed three load sharing schemes, in which admission control is coupled with different call assignment strategies. The traffic load is distributed to the integrated cell and WLAN by matching a service request with the better network support. The overall resources are jointly considered and allocated to efficiently support the traffic load.

  For the admission scheme with service-differentiated call assignment, conversational voice calls are preferably assigned to the cell so as to benefit from the strength of the cellular network in real-time service. In contrast, interactive data calls are first assigned to the WLAN to fully utilize the high throughput with elastic traffic. The overall resource utilization can be improved by properly determining the admission regions. Further, a randomized assignment strategy is studied to achieve the desirable load sharing by controlling the assignment probabilities. It enables distributed implementation, although the performance may be compromised slightly. Moreover, the heavy-tailedness of data call size is effectively exploited by a size-based assignment strategy and SRPT scheduling discipline. As a result, the interworking performance is significantly improved.

- Enhancement of load sharing by call reassignment via vertical handoff: In addition to initial call assignment at admission time, the load sharing can be further enhanced by call reassignment via vertical handoff. First, we only consider call reassignment at WLAN border crossing and focus on vertical handoff triggered by user mobility. It is observed that the variability of user mobility has an impact on the overall resource utilization. In the load sharing scheme with size-based call assignment and SRPT scheduling, we further study dynamic vertical handoff to reassign overflow voice calls in the WLAN to the preferred cell whenever there is sufficient spare capacity in the cell. By this means, the traffic is efficiently supported in the preferred network. Also, the free bandwidth of the integrated systems is pooled to achieve a larger multiplexing gain. Finally, two-way call reassignment via dynamic vertical handoff is investigated for a multi-service scenario with conversational voice service, interactive data service, and rate-adaptive video streaming service. Video streaming calls can also be reassigned to the WLAN so that the high throughput is utilized to improve video presentation quality. As seen from the numerical and simulation results, call reassignment plays an important role in complementing initial call assignment to achieve the desired load sharing.

8.2 Further Work

Although the complementary strength of the cellular network and WLANs has promoted their interworking, the network heterogeneity also poses many research challenges to resource allocation. This research has investigated the call assignment and admission control issues to take advantage of the cellular/WLAN interworking for load sharing. There are still many open issues to extend the research in the following aspects:

- Although the overlay structure exists in a real integrated network, the network

topology is actually much more complex. There may be multiple WLANs within a cell and the overlapped WLANs can provide a relatively larger continuous coverage. As a result, horizontal handoff between adjacent WLANs should be taken into account. When a mobile moves out of the current WLAN, it becomes not mandatory to hand over the associated calls to the overlay cell as neighboring WLANs are also possible choices. Hence, call reassignment strategy needs to be adapted to consider this handoff decision, which can affect the load sharing similar to the vertical handoff studied in this thesis.

- It is known that the widely deployed WLAN products are based on IEEE 802.11b, which has rather limited capability in QoS provisioning. Nonetheless, many state-of-art techniques can introduce extended features to WLANs [4]. For example, IEEE 802.11e supports differentiated services with enhanced distributed coordination function (EDCF) in the distributed mode. Also, controlled hybrid coordination function (HCF) is augmented with admission control and scheduling to enhance QoS guarantee in a centralized access mode. Hence, better QoS support is expected for real-time services in upgraded WLANs. On the other hand, 3G cellular networks are also evolving toward broadband access by employing technologies such as Evolution-Data Optimized (EV-DO) for cdma2000© and High-Speed Packet Access (HSPA) for UMTS. The data transmission (especially, at the downlink) can be improved with techniques such as adaptive modulation and coding and fast packet scheduling at the base station. With a much larger network capacity, IP multimedia services such as video streaming becomes increasingly popular in the mobile wireless domain [115], whereas the revenue from traditional dominating voice and data services will shrink gradually. To support the highly variable and bandwidth-intensive multimedia services, this heterogeneous interworking environment should address many new challenges. The resource allocation strategy needs to be revised accordingly, so that consistent QoS

is maintained for smooth presentation.

- Load sharing within cellular/WLAN interworking has been studied in this thesis for a multi-service scenario including voice, data, and video streaming services. The proposed multi-service load sharing scheme can be extended further by taking advantage of the unique characteristics of video streaming service. A higher resource utilization is achievable if more information can be properly used in matching service demand with network support. For example, taking into account the heavy-tailed data call size, the proposed load sharing scheme with size-based call assignment and SRPT scheduling significantly improves the system performance. For video streaming service, session-related information such as video clip length and variability of encoding bit rate can be derived from control signaling. Therefore, we can also enhance the multi-service load sharing performance by appropriately using the extracted information.

- Nowadays, there is an increasing level of decentralization in wireless networks. Particularly, in the heterogeneous cellular/WLAN interworking environment, it can be more cost-effective to distribute the intelligence of a centralized resource controller among mobile nodes, which can be cooperated in a mesh mode. A virtual hierarchy can be formed dynamically among mobile nodes with multihop relay techniques. In comparison with the single-hop infrastructure, multihop relay can extend coverage and enhance system capacity at a low cost. Mobile multihop relay mode (MMR) is being drafted in IEEE 802.16j for wireless metropolitan networks (WirelessMAN), which is also termed as WiMAX for worldwide interoperability for microwave access by the MiMAX Forum[1]. Also, IEEE 802.11s is being specified to amend 802.11 WLAN with mesh networking, so that the 802.11 PHY/MAC layers can support both broadcast/multicast and unicast delivery over

[1]The MiMAX Forum is an industry group to promote and certify compatibility and interoperability of broadband IEEE 802.16-based WirelessMAN products.

self-configuring multi-hop topologies. In fact, multihop relay is even more favorable for cellular/WLAN interworking. As WLANs are usually deployed in disjoint hotspots, the relay between feasible access points can eliminate the necessity of single-hop connection to the cellular base station. Legacy single-mode mobile terminals can also benefit from the relay of dual-mode mobile terminals [126].

Under such a multihop mesh infrastructure for cellular/WLAN interworking, load sharing becomes critical and extremely challenging to provide QoS guarantee. Strategies for distributed control can be applied at mobile nodes to determine the preferred access, depending on information locally available such as broadcast messages from a higher-level controller or peer nodes and real-time measurements. Without global timely information and overall optimized planning, the distributed control may lead to transient network instability and inconsistent service quality. Also, the relay path routing may have a substantial impact on the traffic load to mesh points and in turn affects the decision for load sharing. Hence, the load sharing strategy should work cooperatively with relay path routing to take full advantage of the heterogeneous networking and multihop relay.

Appendix A

WLAN Capacity Analysis

Following the notation in Section 4.1, this appendix gives the details of deriving the WLAN capacity region. The feasible set of (n_v^w, n_d^w) vectors can be obtained to satisfy the stability constraints that $\xi_v^w(n_v^w, n_d^w) > \lambda_v^p$ and $\xi_d^w(n_v^w, n_d^w) > \lambda_d^p(n_v^w, n_d^w)$. Therefore, we need to properly evaluate the packet service rates $\xi_v^w(n_v^w, n_d^w)$ and $\xi_d^w(n_v^w, n_d^w)$.

When a packet from a voice flow transmits in a slot, a collision will happen if any other voice/data flow transmits in the same slot. The collision probability is given by

$$p_v = 1 - (1 - \rho_v^p \tau_v)^{n_v^w - 1}(1 - \rho_d^p \tau_d)^{n_d^w} \tag{A.1}$$

where[1]

$$\rho_v^p = \frac{\lambda_v^p}{\xi_v^w(n_v^w, n_d^w)}, \quad \rho_d^p = \frac{\lambda_d^p(n_v^w, n_d^w)}{\xi_d^w(n_v^w, n_d^w)} \tag{A.2}$$

and τ_v and τ_d are the transmission probability of a voice flow and a data flow in a slot, respectively, given by [127]

$$\tau_v = \frac{2(1-2p_v)(1-p_v^{m+1})}{W(1-(2p_v)^{m'+1})(1-p_v) + W(1-2p_v)2^{m'}(p_v^{m'+1} - p_v^{m+1}) + (1-2p_v)(1-p_v^{m+1})} \tag{A.3}$$

$$\tau_d = \frac{2(1-2p_d)(1-p_d^{m+1})}{W(1-(2p_d)^{m'+1})(1-p_d) + W(1-2p_d)2^{m'}(p_d^{m'+1} - p_d^{m+1}) + (1-2p_d)(1-p_d^{m+1})} \tag{A.4}$$

[1]For the saturated case, $\rho_d^p = 1.0$.

with $W = \mathrm{CW}_{\min} + 1$ and $\mathrm{CW}_{\min}$ being the initial backoff window, which is 31 in IEEE 802.11, m the retransmission limit, and m' the maximum backoff stage. Similarly, the collision probability for a data flow transmitting in a slot is

$$p_d = 1 - (1 - \rho_v^p \tau_v)^{n_v^w} (1 - \rho_d^p \tau_d)^{n_d^w - 1}. \tag{A.5}$$

For data traffic, the time durations of a successful and collided transmission from a data flow are respectively given by

$$T_{sd} = T_{DIFS} + T_{RTS} + T_{SIFS} + T_{CTS} + T_{SIFS} + T_{D\_DATA} + T_{SIFS} + T_{ACK} \tag{A.6}$$

$$T_{cd} = T_{DIFS} + T_{RTS} + T_{CTS\_TO} \tag{A.7}$$

where T_{D_DATA}, T_{RTS}, T_{CTS}, T_{ACK}, T_{DIFS}, T_{SIFS}, and T_{CTS_TO} are the transmission time of a `DATA` frame from a data flow, `RTS` frame duration, `CTS` frame duration, transmission time of an `ACK` frame, DCF interframe space (DIFS), short interframe space (SIFS), and waiting time for a `CTS TIMEOUT`, respectively. On the other hand, the time duration of a successful transmission from a voice flow is

$$T_{sv} = T_{DIFS} + T_{V\_DATA} + T_{SIFS} + T_{ACK} \tag{A.8}$$

with T_{V_DATA} being the transmission time of a voice `DATA` frame. The time for a collided transmission from a voice flow, denoted by T_{cv}, depends on the traffic types of the collided frames. A target voice frame may collide with only voice frames with a probability $q_{vv} = (1 - \rho_d^p \tau_d)^{n_d^w} \left[1 - (1 - \rho_v^p \tau_v)^{n_v^w - 1}\right] / p_v$. The target voice frame may also collide with at least one data frame with a probability $q_{vd} = \left[1 - (1 - \rho_d^p \tau_d)^{n_d^w}\right] / p_v$. We then have

$$T_{cv} = (T_{DIFS} + T_{V\_DATA} + T_{ACK\_TO}) \cdot q_{vv} + T_{cd} \cdot q_{vd} \tag{A.9}$$

where T_{ACK_TO} is the waiting time for an `ACK TIMEOUT`.

Based on an analytical method similar to that in [128, 129], we further have

$$\frac{1}{\xi_v^w(n_v^w, n_d^w)} = \big[(n_v^w - 1)\rho_v^p + 1\big]T_{sv} + n_d^w \rho_d^p T_{sd} + \overline{W}_v + \frac{1}{k}\big[((n_v^w - 1)\rho_v^p + 1)\overline{T}_{cv} + n_d^w \rho_d^p \overline{T}_{cd}\big] \tag{A.10}$$

$$\frac{1}{\xi_d^w(n_v^w, n_d^w)} = n_v^w \rho_v^p T_{sv} + \big[(n_d^w - 1)\rho_d^p + 1\big]T_{sd} + \overline{W}_d + \frac{1}{k}\big[n_v^w \rho_v^p \overline{T}_{cv} + ((n_d^w - 1)\rho_d^p + 1)\overline{T}_{cd}\big] \tag{A.11}$$

where k is the average number of voice/data flows involved in a collision, $\overline{W}_v$ ($\overline{W}_d$) is the average backoff time of a voice (data) flow, and $\overline{T}_{cv}$ ($\overline{T}_{cd}$) is the average collision time of a frame from a voice (data) flow. According to the contention procedure, $\overline{W}_v$ and $\overline{W}_d$ can be derived as [129]

$$\overline{W}_v = \frac{1}{2(1-p_v)(1-2p_v)}\Big[W(1-(2p_v)^{m'+1}) + W2^{m'}(p_v^{m'+1} - p_v^{m+1})(1-2p_v) - (1-2p_v)(1-p_v^{m+1})\Big] \tag{A.12}$$

$$\overline{W}_d = \frac{1}{2(1-p_d)(1-2p_d)}\Big[W(1-(2p_d)^{m'+1}) + W2^{m'}(p_d^{m'+1} - p_d^{m+1})(1-2p_d) - (1-2p_d)(1-p_d^{m+1})\Big]. \tag{A.13}$$

The average time that a frame from a voice flow is involved in a collision is given by

$$\overline{T}_{cv} = \sum_{i=1}^{m} i\, T_{cv} \cdot p_v^i \cdot (1-p_v). \tag{A.14}$$

Similarly,

$$\overline{T}_{cd} = \sum_{i=1}^{m} i\, T_{cd} \cdot p_d^i \cdot (1-p_d). \tag{A.15}$$

Note that k is set to 1 in [128] and 2 in [129]. Actually, k can be evaluated more accurately as follows. At any time slot, there are approximately $\bar{n} = n_v^w \rho_v^p + n_d^w \rho_d^p$ flows with backlogged traffic, and the transmission probability of a voice or data flow can be approximated by $\bar{\tau} = \frac{\tau_v + \tau_d}{2}$. Denote the probability that i flows transmit in a slot as

$p(i) = \binom{\bar{n}}{i}\bar{\tau}^i(1-\bar{\tau})^{(\bar{n}-i)}$. Then, the average number of voice and data flows involved in a collision is given by

$$k = \frac{\sum_{i=2}^{\bar{n}} i \cdot p(i)}{\sum_{i=2}^{\bar{n}} p(i)} = \frac{\bar{n}\bar{\tau} - p(1)}{1 - p(0) - p(1)}. \tag{A.16}$$

Based on (A.1), (A.3), (A.4), (A.5), (A.10), and (A.11), the voice and data packet service rates $\xi_v^w(n_v^w, n_d^w)$ and $\xi_d^w(n_v^w, n_d^w)$ can be derived recursively. The feasible set of (n_v^w, n_d^w) vectors satisfying the stability constraints are also obtained correspondingly.

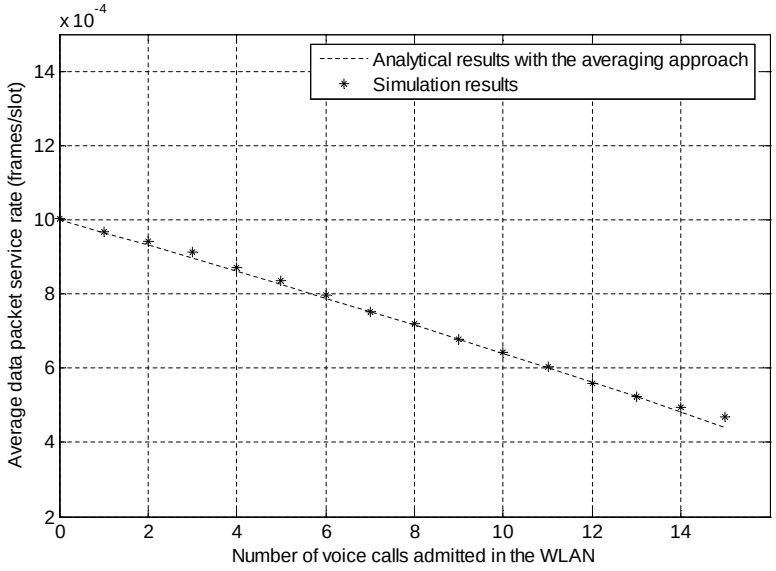


Figure A.1: Average service rate for packets from one data flow $\xi_d^w(n_v^w, n_d^w)$ with $n_d^w = 10$ and n_v^w varying within the WLAN capacity region.

To verify the accuracy of the analysis, we compare our analytical results with computer simulation results. Since quite a few bugs are found in the WLAN MAC implementation of ns-2 simulator [130], we develop a discrete event-driven simulator with C/C++. In each simulation round, more than 10^6 voice and data packets are generated so that the statistics on packet service rates are collected after the simulated system attains the equilibrium state. The results of multiple simulation rounds are averaged

Table A.1: WLAN parameters.

Parameter	**Value**	**Parameter**	**Value**
m'	5	m	7
$\mathrm{CW}_{\min}$	31	Slot	20 μs
T_{SIFS}	10 μs	T_{DIFS}	50 μs
T_{RTS}	13.6 slots	T_{CTS}	12.4 slots
T_{ACK}	10.2 slots	λ_v^p	0.0004 frames/slot
Voice payload	50 bytes/packet	Data payload	1000 bytes/packet

to remove randomness effect. Given in Table A.1 are the WLAN parameters, which are selected by referring to the most popular IEEE 802.11b. As an example, Figure A.1 compares the analytical and simulation results of data packet service rate when n_d^w is fixed to 10 and n_v^w varies in the analytically derived WLAN capacity region. It can be seen that the analytical results agree well with the simulation results, which validates the accuracy of our analytical model.

Bibliography

[1] 3GPP, "Feasibility study on 3GPP system to wireless local area network (WLAN) interworking," 3GPP TR 22.934 V7.0.0, June 2007.

[2] 3GPP, "3GPP system to wireless local area network (WLAN) interworking; system description (Release 6)," 3GPP TS 23.234 V7.5.0, Mar. 2007.

[3] 3GPP, "Quality of service (QoS) and policy aspects of 3GPP - wireless local area network (WLAN) interworking (Release 7)," 3GPP TR 23.836 V1.0.0, Dec. 2005.

[4] H. Zhu, L. Ming, I. Chlamtac, and B. Prabhakaran, "A survey of quality of service in IEEE 802.11 networks," *IEEE Wireless Commun. Mag.*, vol. 11, no. 4, pp. 6–14, Aug. 2004.

[5] K. Maheshwari and A. Kumar, "Performance analysis of microcellization for supporting two mobility classes in cellular wireless networks," *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 321–333, Mar. 2000.

[6] S. Thajchayapong and J. Peha, "Mobility patterns in microcellular wireless networks," *IEEE Trans. Mobile Comput.*, vol. 5, no. 1, pp. 52–63, Jan. 2006.

[7] M. M. Buddhikot, G. Chandranmenon, S. Han, Y.-W. Lee, S. Miller, and L. Salgarelli, "Design and implementation of a WLAN/cdma2000 interworking architecture," *IEEE Commun. Mag.*, vol. 41, no. 11, pp. 90–100, Nov. 2003.

[8] A. K. Salkintzis, "Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks," *IEEE Wireless Commun. Mag.*, vol. 11, no. 3, pp. 50–61, June 2004.

[9] M. Bernaschi, F. Cacace, G. Iannello, S. Za, and A. Pescape, "Seamless internetworking of WLANs and cellular networks: architecture and performance issues in a Mobile IPv6 scenario," *IEEE Wireless Commun. Mag.*, vol. 12, no. 3, pp. 73–80, June 2005.

[10] Q. Zhang, Z. Guo, and W. Zhu, "Efficient mobility management for vertical handoff between WWAN and WLAN," *IEEE Commun. Mag.*, vol. 41, no. 11, pp. 102–108, Nov. 2003.

[11] C. Guo, Z. Guo, Q. Zhang, and W. Zhu, "A seamless and proactive end-to-end mobility solution for roaming across heterogeneous wireless networks," *IEEE J. Select. Areas Commun.*, vol. 22, no. 5, pp. 834–848, June 2004.

[12] T. E. Klein and S.-J. Han, "Assignment strategies for mobile data users in hierarchical overlay networks: performance of optimal and adaptive strategies," *IEEE J. Select. Areas Commun.*, vol. 22, no. 5, pp. 849–861, June 2004.

[13] F. Yu and V. Krishnamurthy, "Optimal joint session admission control in integrated WLAN and CDMA cellular network," *IEEE Trans. Mobile Comput.*, vol. 6, no. 1, pp. 126–139, Jan. 2007.

[14] W. Song, H. Jiang, W. Zhuang, and X. Shen, "Resource management for QoS support in cellular/WLAN interworking," *IEEE Network*, vol. 19, no. 5, pp. 12–18, Sept.-Oct. 2005.

[15] W. Song, H. Jiang, W. Zhuang, and A. Saleh, "Call admission control for integrated voice/data services in cellular/WLAN interworking," in *Proc. IEEE ICC*, vol. 12, June 2006, pp. 5480–5485.

[16] W. Song, H. Jiang, and W. Zhuang, "Performance analysis of the WLAN-first scheme in cellular/WLAN interworking," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1932–1952, May 2007.

[17] W. Song and W. Zhuang, "QoS provisioning via admission control in cellular/wireless LAN interworking," in *Proc. IEEE BROADNETS*, Oct. 2005, pp. 543–550.

[18] W. Song, Y. Cheng, W. Zhuang, and A. Saleh, "Improving voice and data service provisioning in cellular/WLAN integrated networks by admission control," in *Proc. IEEE GLOBECOM*, Nov. 2006, WLC34-3.

[19] W. Song, Y. Cheng, and W. Zhuang, "Improving voice and data services in cellular/WLAN integrated network by admission control," *IEEE Trans. Wireless Commun.*, to appear, available at http://bbcr.uwaterloo.ca/∼wzhuang/papers/Zhuang_TW-Apr-06-0146R2.pdf.

[20] W. Song and W. Zhuang, "Multi-service load sharing for cellular/WLAN interworking," submitted to IEEE Transactions on Wireless Communications.

[21] W. Song and W. Zhuang, "Resource allocation for conversational, streaming, and interactive services in cellular/WLAN interworking," in *Proc. IEEE GLOBECOM*, Nov. 2007.

[22] W. Song and W. Zhuang, "Multi-class resource management in a cellular/WLAN integrated network," in *Proc. IEEE WCNC*, Mar. 2007, pp. 3070–3075.

[23] ETSI, "Requirements and architectures for internetworking between HIPERLAN/2 and 3rd generation cellular systems," ETSI TR 101 957, Aug. 2001.

[24] M. M. Buddhikot, G. Chandranmenon, S. Han, Y.-W. Lee, S. Miller, and L. Salgarelli, "Integration of 802.11 and third-generation wireless data networks," in *Proc. IEEE INFOCOM*, vol. 1, Apr. 2003, pp. 503–512.

[25] V. K. Varma, S. Ramesh, K. D. Wong, M. Barton, G. Hayward, and J. A. Friedhoffer, "Mobility management in integrated UMTS/WLAN networks," in *Proc. IEEE ICC*, vol. 2, May 2003, pp. 1048–1053.

[26] A. K. Salkintzis, C. Fors, and R. Pazhyannur, "WLAN-GPRS integration for next generation mobile data networks," *IEEE Wireless Commun. Mag.*, vol. 9, no. 5, pp. 112–124, Oct. 2002.

[27] M. Jaseemuddin, "An architecture for integrating UMTS and 802.11 WLAN networks," in *Proc. 8th IEEE Int'l Symp. on Computers and Commun. (ISCC)*, vol. 2, July 2003, pp. 716–723.

[28] S.-L. Tsao and C.-C. Lin, "Design and evaluation of UMTS-WLAN interworking strategies," in *Proc. IEEE VTC*, vol. 2, Sept. 2002, pp. 777–781.

[29] N. Vulic, I. Niemegeers, and S. H. de Groot, "Architectural options for the WLAN integration at the UMTS radio access level," in *Proc. IEEE VTC*, vol. 5, May 2004, pp. 3009–3013.

[30] R. Pichna, T. Ojanperä, H. Posti, and J. Karppinen, "Wireless Internet - IMT-2000/wireless LAN interworking," *Journal of Communication and Networks*, vol. 2, no. 1, pp. 46–57, Mar. 2000.

[31] J. Luo, R. Mukerjee, M. Dillinger, E. Mohyeldin, and E. Schulz, "Investigation of radio resource scheduling in WLANs coupled with 3G cellular network," *IEEE Commun. Mag.*, vol. 41, no. 6, pp. 108–115, June 2003.

[32] J.-C. Chen and H.-W. Lin, "A gateway approach to mobility integration of GPRS and wireless LANs," *IEEE Wireless Commun. Mag.*, vol. 12, no. 2, pp. 86–95, Apr. 2005.

[33] J. Ala-Laurila, J. Mikkonen, and J. Rinnemaa, "Wireless LAN access network architecture for mobile operators," *IEEE Commun. Mag.*, vol. 39, no. 11, pp. 82–89, Nov. 2001.

[34] 3GPP2, "cdma2000 wireless IP network standard," P.S0001-B v2.0, Sept. 2004.

[35] C. Perkins, Ed., "IP mobility support for IPv4," IETF RFC 3344, Aug. 2002.

[36] C. de Laat, G. Gross, L. Gommans, J. Vollbrecht, and D. Spence, "Generic AAA architecture," IETF RFC 2903, Aug. 2000.

[37] W. Song, W. Zhuang, and A. Saleh, "Architectures for integrating wireless LAN and cellular networks," *Int. J. Wireless and Mobile Computing*, to appear, available at http://bbcr.uwaterloo.ca/∼wzhuang/papers/wei-IJWMC.pdf.

[38] K. Sabnani, "Converged networks of the future," in *NSF Wireless/Mobile Planning Group Wksp.*, Aug. 2005.

[39] G. M. Koien and T. Haslestad, "Security aspects of 3G-WLAN interworking," *IEEE Commun. Mag.*, vol. 41, no. 11, pp. 82–88, Nov. 2003.

[40] G. Lampropoulos, N. Passas, L. Merakos, and A. Kaloxylos, "Handover management architectures in integrated WLAN/cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 7, no. 4, Fourth Quarter 2005, http://www.comsoc.org/livepubs/surveys.

[41] 3GPP, "Combined GSM and Mobile IP mobility handling in UMTS IP CN," 3GPP TR 23.923 V3.0.0, June 2000.

[42] A.-C. Pang, J.-C. Chen, Y.-K. Chen, and P. Agrawal, "Mobility and session management: UMTS vs. cdma2000," *IEEE Wireless Commun. Mag.*, vol. 11, no. 4, pp. 30–43, Aug. 2004.

[43] M. Siebert, M. Schinnenburg, and M. Lott, "Enhanced measurement procedures for vertical handover in heterogeneous wireless systems," in *Proc. IEEE PIMRC*, vol. 1, Sept. 2003, pp. 166–171.

[44] M.-H. Ye, Y. Liu, and H.-M. Zhang, "The mobile IP handoff between hybrid networks," in *Proc. IEEE PIMRC*, vol. 1, Sept. 2002, pp. 265–269.

[45] H. Bing, C. He, and L. Jiang, "Performance analysis of vertical handover in a UMTS-WLAN integrated network," in *Proc. IEEE PIMRC*, vol. 1, Sept. 2003, pp. 187–191.

[46] R. Inayat, R. Aibara, and K. Nishimura, "A seamless handoff for dual-interfaced mobile devices in hybrid wireless access networks," in *Proc. 18th Int'l Conf. on Adv. Information Networking and Applications (AINA)*, vol. 1, Mar. 2004, pp. 373–378.

[47] H. Badis and K. A. Agha, "An efficient mobility management in wireless overlay networks," in *Proc. IEEE PIMRC*, vol. 3, Sept. 2003, pp. 2500–2504.

[48] H. S. Park, S. H. Yoon, T. H. Kim, J. S. Park, M. S. Do, and J. Y. Lee, "Vertical handoff procedure and algorithm between IEEE 802.11 WLAN and CDMA cellular network," *Lecture Notes in Computer Science (LNCS)*, no. 2524, pp. 103–112, 2003.

[49] E. Vanem, S. Svaet, and F. Paint, "Effects of multiple access alternatives in heterogeneous wireless networks," in *Proc. IEEE WCNC*, vol. 3, Mar. 2003, pp. 1696–1700.

[50] E. Stevens-Navarro and V. W. Wong, "Comparison between vertical handoff decision algorithms for heterogeneous wireless networks," in *Proc. IEEE VTC*, vol. 2, May 2006, pp. 947– 951.

[51] K. Yoon and C. Hwang, *Multiple Attribute Decision Making: An Introduction*. Sage Publications, 1995.

[52] W. Zhang, "Handover decision using fuzzy MADM in heterogeneous networks," in *Proc. IEEE WCNC*, vol. 2, Mar. 2004, pp. 653–658.

[53] Q. Song and A. Jamalipour, "Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques," *IEEE Wireless Commun. Mag.*, vol. 12, no. 3, pp. 42–48, June 2005.

[54] X. Gao, G. Wu, and T. Miki, "End-to-end QoS provisioning in mobile heterogeneous networks," *IEEE Wireless Commun. Mag.*, vol. 11, no. 3, pp. 24–34, June 2004.

[55] A. T. Campbell, J. Gomez, S. Kim, C.-Y. Wan, Z. R. Turanyi, and A. G. Valko, "Comparison of IP micromobility protocols," *IEEE Wireless Commun. Mag.*, vol. 9, no. 1, pp. 72–82, Feb. 2002.

[56] A. E. Xhafa and O. K. Tonguz, "Reducing handover time in heterogeneous wireless networks," in *Proc. IEEE VTC*, vol. 4, Oct. 2003, pp. 2222–2226.

[57] F. Du, L. M. Ni, and A. H. Esfahanian, "HOPOVER: a new handover protocol for overlay networks," in *Proc. IEEE ICC*, vol. 5, May 2002, pp. 3234–3239.

[58] C. W. Lee, L. M. Chen, M. C. Chen, and Y. S. Sun, "A framework of handoffs in wireless overlay networks based on Mobile IPv6," *IEEE J. Select. Areas Commun.*, vol. 23, no. 11, pp. 2118–2128, Nov. 2005.

[59] L. Ma, F. Yu, V. C. M. Leung, and T. Randhawa, "A new method to support UMTS/WLAN vertical handover using SCTP," *IEEE Wireless Commun. Mag.*, vol. 11, no. 4, pp. 44–51, Aug. 2004.

[60] G. Wu, P. J. M. Havinga, and M. Mizuno, "Wireless Internet over heterogeneous wireless networks," in *Proc. IEEE INFOCOM*, vol. 3, Nov. 2001, pp. 1759–1765.

[61] N. Banerjee, W. Wu, and S. K. Das, "Mobility support in wireless Internet," *IEEE Wireless Commun. Mag.*, vol. 10, no. 5, pp. 54–61, Oct. 2003.

[62] K. L. Yeung and S. Nanda, "Channel management in microcell-macrocell cellular radio systems," *IEEE Trans. Veh. Technol.*, vol. 45, no. 4, pp. 601–612, Nov. 1996.

[63] S. A. Ghorashi, F. Said, and A. H. Aghvami, "Handover rate control in hierarchically structured cellular CDMA systems," in *Proc. IEEE PIMRC*, vol. 3, Sept. 2003, pp. 2083–2087.

[64] K.-R. Lo, C.-J. Chang, C. Chang, and C. B. Shung, "A combined channel assignment strategy in a hierarchical cellular systems," in *Proc. IEEE 6th Int'l Conf. on Universal Personal Commun.*, vol. 2, Oct. 1997, pp. 651–655.

[65] F. Santucci, W. Huang, P. Tranquilli, and V. K. Bhargava, "Admission control in wireless systems with heterogeneous traffic and overlaid cell structure," in *Proc. IEEE VTC*, vol. 3, Sept. 2000, pp. 1106–1113.

[66] B. Jabbari and W. F. Fuhrmann, "Teletraffic modeling and analysis of flexible hierarchical cellular networks with speed-sensitive handoff strategy," *IEEE J. Select. Areas Commun.*, vol. 15, no. 8, pp. 1539–1548, Oct. 1997.

[67] W. Shan, "Performance evaluation of a hierarchical cellular system with mobile velocity-based bidirectional call-overflow scheme," *IEEE Trans. Parallel Distrib. Syst.*, vol. 14, no. 1, pp. 72–83, Jan. 2003.

[68] 3GPP, "Quality of service (QoS) concept and architecture," 3GPP TS 23.107 V7.0.0, June 2007.

[69] J. W. Roberts and L. Massoulié, "Bandwidth sharing and admission control for elastic traffic," ITC Specialist Seminar, Oct. 1998.

[70] 3GPP, "Services and service capabilities," 3GPP TS 22.105 V8.4.0, June 2007.

[71] 3GPP, "Selection procedures for the choice of radio transmission technologies of the UMTS," 3GPP TS 30.03 V3.2.0, Apr. 1998.

[72] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[73] S. B. Fredj, T. Bonald, A. Proutière, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: a study of congestion at flow level," in *Proc. ACM SIGCOMM*, Aug. 2001, pp. 111–122.

[74] R. Daris and L. Torelli, "Some indices for heavy-tailed distributions," in *Proc. 31st Int'l ASTIN Colloquium*, June 2000, pp. 45–54.

[75] M. Fischer, D. Masi, D. Gross, and J. Shortle, "Loss systems with heavy-tailed arrivals," *The Telecommunications Review*, no. 15, pp. 95–99, 2004.

[76] K. M. Rezaul and A. Pakštas, "Web traffic analysis based on EDF statistics," in *Proc. 7th Annual PostGraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNet)*, June 2006.

[77] A. Feldmann and W. Whitt, "Fitting mixtures of exponentials to long-tail distributions to analyze network performance models," *Perform. Eval.*, vol. 31, no. 3-4, pp. 245–279, Jan. 1998.

[78] N. Benameur, S. B. Fredj, F. Delcoigne, S. Oueslati-Boulahia, and J. W. Roberts, "Integrated admission control for streaming and elastic traffic," in *Proc. 2nd Int'l Wksp. on Quality of Future Internet Services*, Sept. 2001, pp. 69–81.

[79] H. C. Tijms, *Stochastic Models - An Algorithm Approach.* John-Wiley and Sons, 1994, p. 359.

[80] O. J. Boxma, A. F. Gabor, R. Nunez-Queija, and H.-P. Tan, "Performance analysis of admission control for integrated services with minimum rate guarantees," in *Proc. 2nd Conf. on Next Generation Internet Design and Engineering (NGI)*, Apr. 2006, pp. 41–47.

[81] W. Montes, G. Gomez, R. Cuny, and J. F. Paris, "Deployment of IP multimedia streaming services in third-generation mobile networks," *IEEE Wireless Commun. Mag.*, vol. 9, no. 5, pp. 84–92, Oct. 2002.

[82] M. Lundevall, B. Olin, J. Olsson, N. Wiberg, S. Wanstedt, J. Eriksson, and F. Eng, "Streaming applications over HSDPA in mixed service scenarios," in *Proc. IEEE VTC*, vol. 2, Sept. 2004, pp. 841–845.

[83] E. B. Rodrigues and J. Olsson, "Admission control for streaming services over HSDPA," in *Proc. AICT/SAPIR/ELETE*, vol. 00, July 2005, pp. 255–260.

[84] C. Johansson, H. Nyberg, and P. de Bruin, "Streaming services in GSM/EDGE-radio resource management concepts and system performance," in *Proc. IEEE VTC*, vol. 3, Oct. 2001, pp. 1765–1769.

[85] M. Li, M. Claypool, R. Kinicki, and J. Nichols, "Characteristics of streaming media stored on the Web," *IEEE/ACM Trans. Networking*, vol. 5, no. 4, pp. 601–626, Nov. 2005.

[86] 3GPP, "Transparent end-to-end packet-switched streaming service (PSS); protocols and codecs," 3GPP TS 26.234 V7.3.0, June 2007.

[87] F. H. P. Fitzek and M. Reisslein, "MPEG-4 and H.263 video traces for network performance evaluation," *IEEE Network*, vol. 15, no. 6, pp. 40–54, Nov.-Dec. 2001.

[88] D. P. Heyman, "The GBAR source model for VBR videoconferences," *IEEE/ACM Trans. Networking*, vol. 5, no. 4, pp. 554–560, Aug. 1997.

[89] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: a media access protocol for wireless LAN's," *ACM SIGCOMM Computer Communication Review*, vol. 24, no. 4, pp. 212–225, Oct. 1994.

[90] A. K. Salkintzis, G. Dimitriadis, D. Skyrianoglou, N. Passas, and N. Pavlidou, "Seamless continuity of real-time video across UMTS and WLAN networks: challenges and performance evaluation," *IEEE Wireless Commun. Mag.*, vol. 12, no. 3, pp. 8–18, June 2005.

[91] J. F. Huber, D. Weiler, and H. Brand, "UMTS, the mobile multimedia vision for IMT-2000: A focus on standardization," *IEEE Commun. Mag.*, vol. 38, no. 9, pp. 129–136, Sept. 2000.

[92] M. Stemm and R. H. Katz, "Vertical handoffs in wireless overlay networks," *Mobile Networks and Applications*, vol. 3, no. 4, pp. 335–350, 1998.

[93] S. Racz, M. Telek, and G. Fodor, "Call level performance analysis of 3rd generation mobile core networks," in *Proc. IEEE ICC*, vol. 2, June 2001, pp. 456–461.

[94] D. V. Sole and A. C. Auge, "Session information based admission control strategy for streaming services over all-IP 3G networks," in *Proc. IEEE PIMRC*, vol. 3, Sept. 2004, pp. 1807–1811.

[95] J. Pérez-Romero, O. Sallent, R. Agusti, and M. A. Diaz-Guerra, *Radio Resource Management Strategies in UMTS*. New York: Wiley, 2005.

[96] H. Zhai, X. Chen, and Y. Fang, "A call admission and rate control scheme for multimedia support over IEEE 802.11 wireless LANs," *Wireless Networks*, vol. 12, no. 4, pp. 451–463, July 2006.

[97] H. Zhai, X. Chen, and Y. Fang, "How well can the IEEE 802.11 wireless LAN support quality of service?" *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 3084–3094, Nov. 2005.

[98] P. Tran-Gia and F. Hübner, "An analysis of trunk reservation and grade of service balancing mechanisms in multiservice broadband networks," in *Proc. IFIP Workshop TC6*, 1993, pp. 83–97.

[99] M. Naghshineh and A. S. Acampora, "QoS provisioning in micro-cellular networks supporting multiple classes of traffic," *Wireless Networks*, vol. 2, no. 3, pp. 195–203, Aug. 1996.

[100] R. N. Queija, J. L. van den Berg, and M. R. H. Mandjes, "Performance evaluation of strategies for integration of elastic and stream traffic," in *Proc. 16th Int'l. Teletraffic Congress*, 1999, pp. 1039–1050.

[101] R. N. Queija, "Processor-sharing models for integrated-services networks," Ph.D. dissertation, Eindhoven University of Technology, Jan. 2000.

[102] M. Marsan, "Performance analysis of hierarchical cellular networks with generally distributed call holding times and dwell times," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 248–257, Jan. 2004.

[103] F. Delcoigne, A. Proutière, and G. Régnié, "Modeling integration of streaming and data traffic," *Perform. Eval.*, vol. 55, no. 3-4, pp. 185–209, Feb. 2004.

[104] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer-Verlag, 1995.

[105] F. P. Kelly, *Reversibility and Stochastic Networks*. New York: Wiley, 1979.

[106] R. Litjens and R. J. Boucherie, "Elastic calls in an integrated services network: the greater the call size variability the better the QoS," *Perform. Eval.*, vol. 52, no. 4, pp. 193 – 220, May 2003.

[107] S. Lincke-Salecke, "Load shared integrated networks," in *Proc. 5th European Personal Mobile Commun. Conf. (EPMCC)*, Apr. 2003, pp. 225–229.

[108] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Prentice Hall, 1989.

[109] S. B. Fredj, S. Oueslati-Boulahia, and J. W. Roberts, "Measurement-based admission control for elastic traffic," in *Proc. 17th Int'l. Teletraffic Congress*, Dec. 2001, pp. 161–172.

[110] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. New York: John Wiley and Sons, 1975.

[111] N. Bansal and M. Harchol-Balter, "Analysis of SRPT scheduling: investigating unfairness," *ACM SIGMETRICS Performance Evaluation Review*, vol. 29, no. 1, pp. 279–290, June 2001.

[112] M. Garcia-Martin, M. Isomaki, G. Camarillo, and S. Loreto, "A session description protocol (SDP) offer/answer mechanism to enable file transfer," Internet draft, June 2007.

[113] L. E. Schrage and L. W. Miller, "The queue M/G/1 with the shortest remaining processing time discipline," *Operations Research*, vol. 14, no. 4, pp. 670–684, Jul.-Aug. 1966.

[114] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C, The Art of Scientific Computing (second edition)*. Cambridge: Cambridge University Press, 1999.

[115] P. Frojdh, U. Horn, M. Kampmann, A. Nohlgren, and M. Westerlund, "Adaptive streaming within the 3GPP packet-switched streaming service," *IEEE Network*, vol. 20, no. 2, pp. 34–40, Mar.-Apr. 2006.

[116] A. Kyriakidou, N. Karelos, and A. Delis, "Video-streaming for fast moving users in 3G mobile networks," in *Proc. 4th ACM Int'l Wksp. on Data Engineering for Wireless and Mobile Access (MobiDE)*, June 2005, pp. 65–72.

[117] B. Girod, J. Chakareski, M. Kalman, Y. Liang, E. Setton, and R. Zhang, "Advances in network-adaptive video streaming," in *Proc. Tyrrhenian Int'l. Wksp. on Digital Communications (IWDC)*, Sept. 2002.

[118] P. de Cuetos and K. W. Ross, "Adaptive rate control for streaming stored fine-grained scalable video," in *Proc. 12th Int'l Wksp. on Network and Operating Systems Support for Digital Audio and Video*, May 2002, pp. 3–12.

[119] M. van der Schaar and H. Radha, "A hybrid temporal-SNR fine-granular scalability for Internet video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 318–331, Mar. 2001.

[120] L. S. Lam, J. Y. B. Lee, S. C. Liew, and W. Wang, "Adaptive rate control for streaming stored fine-grained scalable video," in *Proc. 18th Int'l Conf. on Advanced Information Networking and Applications (AINA)*, May 2004, pp. 346–351.

[121] T. Bonald, A. Proutière, G. Régnié, and J. W. Roberts, "Insensitivity results in statistical bandwidth sharing," in *Proc. 17th Int'l. Teletraffic Congress*, Dec. 2001, pp. 125–136.

[122] G. S. Fishman and J. B. F. Adan, "How heavy-tailed distributions affect simulation-generated time averages," *ACM Trans. Modeling and Computer Simulation*, vol. 16, no. 2, pp. 152–173, Apr. 2006.

[123] M. C. Weigle, "Improving confidence in network simulations," in *Proc. 37th Conf. on Winter Simulation*, Dec. 2006, pp. 2188–2194.

[124] 3GPP, "Physical layer aspects of UTRA high speed downlink packet access (Release 4)," 3GPP TR 25.848 V4.0.0, Mar. 2001.

[125] 3GPP2, "cdma2000 evaluation methodology," 3GPP2 C.R1002-0 Version 1.0, Dec. 2004.

[126] H.-Y. Wei and R. D. Gitlin, "Two-hop-relay architecture for next-generation WWAN/WLAN integration," *IEEE Wireless Commun. Mag.*, vol. 11, no. 2, pp. 24–30, Apr. 2004.

[127] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement," in *Proc. IEEE INFOCOM*, vol. 2, June 2002, pp. 599–607.

[128] O. Tickoo and B. Sikdar, "A queueing model for finite load IEEE 802.11 random access MAC," in *Proc. IEEE ICC*, vol. 1, June 2004, pp. 175–179.

[129] L. X. Cai, X. Shen, J. W. Mark, L. Cai, and Y. Xiao, "Voice capacity analysis of WLAN with unbalanced traffic," *IEEE Trans. Veh. Technol.*, vol. 55, no. 3, pp. 752–761, May 2006.

[130] H. L. Vu and T. Sakurai, "Accurate delay distribution for IEEE 802.11," *IEEE Communications Letters*, vol. 10, no. 4, pp. 317–319, Apr. 2006.

List of Abbreviations

3G	The third-generation cellular network
3GPP	The 3rd Generation Partnership Project
3GPP2	The 3rd Generation Partnership Project 2
AAA	Authentication, authorization, and accounting
AMP	Adaptive media playout
AN	Access network
AP	Access point
AVC	Advanced video coding
BER	Bit error rate
BS	Base station
CAC	Call admission control
CDF	Cumulative distribution function
CDMA	Code-division multiple access
CN	Core network
CP	Complete partitioning
CS	Complete sharing
CSMA/CA	Carrier sense multiple access with collision avoidance
CTS	Clear to send
CV	Coefficient of variance
DCF	Distributed coordination function

DCH	Dedicated channel
DIFS	DCF interframe space
DSCH	Downlink shared channel
DSSS	Direct sequence spread spectrum
EDCF	Enhanced distributed coordination function
FGS	Fine-granular scalability
GBAR	Gamma-beta autoregressive process
GGSN	Gateway GPRS support node
GPRS	General packet radio service
GSM	The global system for mobile communications
HCF	Hybrid coordination function
HSS	Home subscriber server
IAPP	Inter-access point protocol
IMS	IP multimedia subsystem
IP	Internet protocol
ISM	Industrial, scientific, and medical frequency band
MAC	Medium access control
MADM	Multiple attribute decision making
MGF	Moment generating function
OFDM	Orthogonal frequency-division multiplexing
PDF	Probability density function
PDP	Packet data protocol
PDSN	Packet data serving node
PSS	Packet-switched streaming
QoS	Quality-of-service
RA	Routing area
RAN	Radio access network

RNC	Radio network controller
RSS	Received signal strength
RTCP	Real-time transport control protocol
RTP	Real-time transport protocol
RTS	Request to send
RTSP	Real-time streaming protocol
SCTP	Stream control transmission protocol
SDP	Session description protocol
SGSN	Serving GPRS support node
SIFS	Short interframe space
SIP	Session initiation protocol
SIR	Signal-to-interference ratio
SNR	Signal-to-noise ratio
SRPT	Shortest remaining processing time service discipline
TDD	Time division duplexing
UDP	User datagram protocol
UMTS	Universal mobile telecommunication system
WLAN	Wireless local area network

List of Notations

a	Variability parameter of hyper-exponential-distributed user residence time in the WLAN 38
α_d	Shape parameter of Weibull-distributed data call size 26
α_s	Shape parameter of marginal gamma distribution for video frame size 29
α_v	Activity factor of voice calls 44
b	Variability parameter of hyper-exponential-distributed data call size 27
B_d	Overall data call blocking probability 101
B_{d1}^c (B_{d2}^c)	Data call blocking probability of the cell in the cellular-only (double-coverage) area 51
β_d	Scale parameter of Weibull-distributed data call size 26
B_v	Overall voice call blocking probability 101
B_{v1}^c (B_{v2}^c)	Voice call blocking probability of the cell in the cellular-only (double-coverage) area 51
B_v^w (B_d^w)	Voice (data) call blocking probability of the WLAN 51
χ_d^c	Probability that a data call is not blocked by the cell due to congestion .. 99
$\mathrm{CW}_{\min}$	Initial backoff window of the WLAN 141
δ_d^c	Fraction of data calls with a size not greater than Φ_d 99
D_v^c (D_d^c)	Voice (data) call dropping probability of the cell 51
$\frac{E_b}{N_0}$	Ratio of bit energy to noise and interference power spectral density 44
$\left(\frac{E_b}{N_0}\right)_d$	$\frac{E_b}{N_0}$ requirement of data calls in the cell 44
$\left(\frac{E_b}{N_0}\right)_v$	$\frac{E_b}{N_0}$ requirement of voice calls in the cell 44
η_{DL}	Load factor of cell downlink 44
η_{max}	Upper bound of load factor of cell downlink 44
$(\eta^c)^{-1}$	Mean user residence time in the cellular-only area of a cell 38

η_s	Scale parameter of marginal gamma distribution for video frame size	29
$(\eta^w)^{-1}$	Mean user residence time in a WLAN	38
f_{DL}	Ratio of intercell interference and total intracell power at the user receiver	44
f_s	Video frame rate	37
γ_d	Shape parameter of Pareto-distributed data call size	26
γ_s	Shape parameter of Pareto-distributed video clip duration	29
g_d^w	Average size of data calls in the WLAN	100
G_{v1}^c (G_{d1}^c)	Guard bandwidth reserved for handoff voice (data) calls in cellular-only area	50
G_{v2}^c (G_{d2}^c)	Guard bandwidth reserved for voice (data) calls in cellular-only area	50
G_v^w (G_d^w)	WLAN bandwidth reserved for new voice (data) calls	51
H_v^{cc}	Handoff probability of voice calls in cellular-only area to neighboring cells	76
H_v^{cw}	Handoff probability of voice calls in cellular-only area to the overlay WLAN	76
H_v^{wc}	Handoff probability of voice calls from the WLAN to the overlay cell	58
L	Video frame size	29
λ_d^c (λ_d^w)	Mean rate of data call arrivals to the cell (WLAN)	99
$\lambda_d^p(\cdot)$	Mean rate of packet arrivals from a data source	45
λ_{hd}^{cc}	Mean rate of handoff data calls from neighboring cells to the cell	60
λ_{hd}^{cw}	Mean rate of handoff data calls from the cell to the overlay WLAN	59
λ_{hd}^{wc}	Mean rate of handoff data calls from the WLAN to the overlay cell	60
λ_{hv}^{cc}	Mean rate of handoff voice calls from neighboring cells to the cell	55
λ_{hv}^{cw}	Mean rate of handoff voice calls from to the cell to the overlay WLAN	54
λ_{hv}^{wc}	Mean rate of handoff voice calls from the WLAN to the overlay cell	55
λ_{nd2}^c	Mean rate of new data call arrivals to the cell from double-coverage area	60
λ_{nd}^w	Mean rate of new data call arrivals to the WLAN	59
λ_{nv2}^c	Mean rate of new voice call arrivals to the cell from double-coverage area	55
λ_{nv}^w	Mean rate of new voice call arrivals to the WLAN	54
λ_{v1} (λ_{v2})	Mean arrival rate of new voice calls in cellular-only (double-coverage) area	53
λ_v^p	Mean rate of packet arrivals from a voice source	45
λ_v (λ_d)	Mean arrival rate of voice (data) calls in the double-coverage area	99
L_d	Data call size	26
$L_{p,i}$	Path loss for the i^{th} call in the cell	44
m	Retransmission limit of the WLAN	141
m'	Maximum backoff stage of the WLAN	141

M_d	Number of data calls in an interactive data session	23
n_d^c	Number of data calls in the cell	44
n_d^w	Number of data flows in the WLAN	45
N_v^c (N_d^c)	Maximum number of voice (data) calls allowed in the cell	50
n_v^c	Number of voice calls in the cell	44
N_v^w (N_d^w)	Maximum number of voice (data) calls allowed in the WLAN	50
n_v^w	Number of voice flows in the WLAN	45
Ω_t	Estimated mean interval between adjacent call arrivals/departures	124
P_N	Power of background noise in the cell	44
p^{cc}	Probability of users moving out of cellular-only area to neighboring cells	38
p^{cw}	Probability of users moving out of cellular-only area to overlay WLAN	38
p_d	Collision probability of packets from a data flow in the WLAN	141
$\Phi_1(\cdot)$	Moment generating function of T_{r1}^c	38
$\Phi_2(\cdot)$	Moment generating function of T_{r2}^c	38
Φ_d	Data call size threshold for call assignment	97
$\pi_d^c(\cdot)$	Steady-state probability of data calls in the cell	64
$\pi_d^w(\cdot)$	Steady-state probability of data calls in the WLAN	60
$\pi_v^c(\cdot)$	Steady-state probability of voice calls in the cell	57
$\pi_v^w(\cdot)$	Steady-state probability of voice calls in the WLAN	54
$\pi(\cdot)$	Steady-state probability of voice and data calls in a cell/WLAN cluster	99
P_p	Power devoted to common control channels of the cell	44
$\psi(\cdot)$	Moment generating function of $T_r^c + T_r^w$	38
$[\Psi_l, \Psi_u]$	Range of effective bandwidth requirement for video streaming calls	125
$P_{T,max}$	Maximum transmission power of the cell base station	44
p_v	Collision probability of packets from a voice flow in the WLAN	140
Q_{PB}	Upper bound for call blocking probability	51
Q_{PD}	Upper bound for call dropping probability	51
$Q(\cdot)$	Current playback buffer occupancy of a video streaming call	124
Q_T	Upper bound for mean response time of data calls	51
Q^*	Target protection time (in media duration) for playback buffer occupancy	124
R_B	Base-layer bit rate of video clips	123
R_E	Enhancement-layer bit rate of video clips	123
$[R_l, R_u]$	Range of encoding bit rate for video streams	123

$R^c_{b,d}$	Bit rate of data calls in the cell	44
$R^c_{b,v}$	Bit rate of voice calls in the cell	44
R^c_d	Cell bandwidth reserved for data traffic	98
ρ	Cell orthogonality factor	44
$r_L(\cdot)$	Autocorrelation function of GBAR process for video frame size	30
R_s	Encoding bit rate of a video stream	123
S_b	Rebuffer time of a video streaming call	28
S_f	Rebuffer data in media time in a video streaming call	28
S_p	Pre-roll time of a video streaming call	28
S_r	Reading time in an interactive data session	23
T_{ACK}	Transmission time of an `ACK` frame in the WLAN	141
T_{ACK_TO}	Waiting time for a `ACK TIMEOUT` in the WLAN	141
T_{cd}	Time duration of a collided transmission from a data flow in the WLAN	141
$\overline{T}_{cd}$	Average collision time of a frame from a data flow in the WLAN	142
T_{CTS}	Time duration of a `CTS` frame in the WLAN	141
T_{CTS_TO}	Waiting time for a `CTS TIMEOUT` in the WLAN	141
T_{cv}	Time duration of a collided transmission from a voice flow in the WLAN	141
$\overline{T}_{cv}$	Average collision time of a frame from a voice flow in the WLAN	142
T_{DIFS}	DCF interframe space of the WLAN	141
T_{RTS}	Time duration of an `RTS` frame in the WLAN	141
T_{sd}	Time duration of a successful transmission from a data flow in the WLAN	141
T_{SIFS}	Short interframe space of the WLAN	141
T_{sv}	Time duration of a successful transmission from a voice flow in the WLAN	141
τ_d	Transmission probability of a data flow in a WLAN time slot	140
τ_v	Transmission probability of a voice flow in a WLAN time slot	140
$\overline{T}_d$	Mean response time of data calls	103
$\overline{T}^c_d$ ($\overline{T}^w_d$)	Mean response time of data calls carried by the cell (WLAN)	51
T_{D_DATA}	Transmission time of a `DATA` frame from a data flow in the WLAN	141
T_{V_DATA}	Transmission time of a `DATA` frame from a voice flow in the WLAN	141
θ_s	Adaptive factor for the enhancement-layer encoding of video clips	123
θ^c_v (θ^c_d)	Probability that a voice (data) call in the double-coverage area requests admission to the cell	71

θ_v^w (θ_d^w)	Probability that a voice (data) call in the double-coverage area requests admission to the WLAN 71
T_{r1}^c	Cell residence time of a call from the cellular-only area 38
T_{r2}^c	Cell residence time of a call from the double-coverage area 38
T_r^c	User residence time in the cellular-only area of a cell 38
T_r^w	User residence time in the WLAN 37
T_s	Video clip duration 28
T_v	Voice call duration 36
U_s	Underflow ratio of video streaming calls 28
W	Size of initial backoff window of the WLAN 141
W_c	Total cell bandwidth 44
$\overline{W}_d$	Average backoff time of a data flow in the WLAN 142
W_{L_d}	Weibull factor of Weibull-distributed data call size L_d 26
$\overline{W}_v$	Average backoff time of a voice flow in the WLAN 142
$\xi_d^w(\cdot)$	Mean service rate of the WLAN for packets from a data flow 45
$\xi_s^w(\cdot)$	Available transmission rate for a video streaming call in the WLAN 124
$\xi_v^w(\cdot)$	Mean service rate of the WLAN for packets from a voice flow 45