Joint Radio Resource Management in Heterogeneous Wireless Networks

Kamil Hussien Suleiman

This thesis is submitted in fulfillment of the academic requirements
for the degree of Master of Science in Electrical Engineering
in the Faculty of Engineering and The Built Environment
University of Cape Town
2008

As the candidate’s supervisor, I have approved this dissertation for submission.

Name: Full name and title of thesis advisor

Signed by candidate

Signed: ______________________

Date: ________________________

Declaration

I declare that this thesis is my own work. Where collaboration with other people has taken place, or material generated by other researchers is included, the parties and/or materials are indicated in the acknowledgements or are explicitly stated with references as appropriate.

This work is being submitted for the Master of Science in Electrical Engineering at the University of Cape Town. It has not been submitted to any other university for any other degree or examination.

Name Signed by candidate

Date

Abstract

The complementary nature of wireless access technologies has resulted in the concept of the integration of overlaid wireless access networks to create a robust and ubiquitous system called heterogeneous wireless networks. In such a system, any mobile user running any application can connect to any of the available access technologies. The envisaged heterogeneous wireless networks will create a more efficient and cost-effective system for service providers and better services to network users. However, the integration of different access technologies is a considerable technological challenge.

Designing a good joint radio resource management (JRRM) scheme for a heterogeneous wireless network is part of the integration challenges. The design goal of a JRRM is optimising the trade-off between resource utilisation, quality of service provisioning, fairness, and simplicity of design, amongst others. The diversity of wireless access networks and the resulting complexity of interworking different access networks leave many issues of the JRRM open. The fact that many analytical as well as simulation models of heterogeneous wireless networks are not based on fairly realistic assumptions often affects the practicality of research outputs.

In this work, we design JRRM taking into consideration different features of component access technologies of a heterogeneous wireless environment. We also consider different important realistic features of access networks, such as asymmetry of their overlap and the variability of network resources according to traffic conditions. We use the network simulator (ns-2) to validate our design.

Acknowledgements

I would like to acknowledge members in Chan Group and Dlodlo Group for the supportive learning interactions.

Friends Abiy Getachew, Kaleb Gizaw, Kende Mekonnen, Mesfin Tadesse, Nebiha Bedru, and many others, have one way or the other, made things comfortable and interesting. I would like to deeply express my thanks to all of them.

Table of Contents

Declaration ... iii

Abstract ... iv

Acknowledgements ... v

Table of Contents ... vi

List of Figures ... viii

Glossary ... x

1 Introduction ... 1

1.1 Towards Heterogeneous Wireless Networks ... 1

1.2 Joint Radio Resource Management in Heterogeneous Wireless Networks ... 5

1.3 The Need for Advanced Joint Radio Resource Management ... 6

1.4 Scope and Contributions of this Work ... 7

1.5 Outline of the Thesis ... 7

2 Literature Review on Radio Resource Management ... 8

2.1 Serving Discipline Schemes ... 8

2.2 Joint Radio Resource Management in Heterogeneous Wireless Networks ... 14

3 Proposed Scheme for Serving Discipline ... 18

3.1 Description ... 18

3.2 Simulation Setup ... 20

3.3 Definitions of Simulation Parameters ... 24

3.4 Results and Analysis ... 24

4 Joint Radio Resource Management: System Architecture ... 31

4.1 A High Level Description ... 31

4.2 Horizontal Radio Resource Management ... 32

4.3 Vertical Radio Resource Management ... 32

4.3.1 Call Arrival ... 32

4.3.2 Necessary Handoff ... 33

4.3.3 Desirable Handoff ... 33

4.4 Coordinating the Components of the Joint Resource Management Scheme ... 34

5 Joint Radio Resource Management: Decision Making ... 36

5.1 Description 36
5.2 Simulation Setup 36
5.3 Designs, Results and Analysis 40
5.3.1 Simple JRRM...40
5.3.2 Considering Widening Opportunities for Future Desperate Users 44
5.3.3 Load Balancing...46
5.3.4 Considering Resource Variation in Dynamic Resource Allocation 50
6 Conclusions and Future Work 55
References 56

List of Figures

FIGURE 1. OVERLAYS FORMED BY DIFFERENT WIRELESS ACCESS TECHNOLOGIES. 3
FIGURE 2. AN EXAMPLE OF COMPLETE PARTITION. 9
FIGURE 3. BANDWIDTH ALLOCATION IN THE DP SCHEME. 10
FIGURE 4. BANDWIDTH ALLOCATION IN THE DTBR SCHEME. 11
FIGURE 5. A COMPARISON OF SERVING DISCIPLINE SCHEMES. 20
(A) THE DRBR SCHEME (B) THE MODIFIED DTBR SCHEME. 20
FIGURE 6. SIMULATION NETWORK TOPOLOGY TO STUDY SERVING DISCIPLINES. 21
FIGURE 7. AGGREGATE THROUGHPUT COMPARISON. 25
FIGURE 8. BLOCKING PROBABILITY COMPARISON. 26
FIGURE 9. RT BLOCKING PROBABILITY COMPARISON. 26
FIGURE 10. NRT BLOCKING PROBABILITY COMPARISON. 27
FIGURE 11. DROPPING PROBABILITY COMPARISON. 28
FIGURE 12. NRT DROPPING PROBABILITY COMPARISON. 28
FIGURE 13. RT DROPPING PROBABILITY COMPARISON. 29
FIGURE 14. PERCENTAGE OF ACTIVE RT USERS OVER TOTAL NUMBER OF ACTIVE USERS. 30
FIGURE 15. LOCATION OF HRRM AND VRRM IN A HETEROGENEOUS 31
WIRELESS NETWORK. 31
FIGURE 16. THE FLOW CHART OF THE VRRM. 33
FIGURE 17. THE CAPACITY AND THRESHOLDS OF DIFFERENT NETWORKS. 35
FIGURE 18. NETWORK TOPOLOGY CONSISTING OF UMTS CELLS AND WLAN 37
HOTSPOTS. 37
FIGURE 19. SIMULATION TOPOLOGY CONSISTING OF A UMTS CELL AND A WLAN HOTSPOT. 38
FIGURE 20. THE FLOW CHART OF THE SIMPLE JRRM. 41
FIGURE 21. THE OVERLAP OF TWO ACCESS NETWORKS. 42
FIGURE 22. THROUGHPUT COMPARISON. 43

FIGURE 23. BLOCKING PROBABILITY COMPARISON. 43
FIGURE 24. THE OVERLAP OF TWO NETWORKS IN DIFFERENT FASHIONS. 44
(A) PERFECT OVERLAP (B) PARTIAL OVERLAP 44
(C) COMPLETE OVERLAP OF A SMALLER NETWORK IN A LARGER ONE. 44
FIGURE 25. THROUGHPUT COMPARISON. 45
FIGURE 26. BLOCKING PROBABILITY COMPARISON. 46
FIGURE 27. THROUGHPUT COMPARISON. 48
FIGURE 28. BLOCKING PROBABILITY COMPARISON. 48
FIGURE 29. ACHIEVABLE THROUGHPUT OF IEEE802.11 AS A FUNCTION OF 50
NUMBER OF CLIENT USERS. 50
FIGURE 30. THROUGHPUT COMPARISON. 51
FIGURE 31. BLOCKING PROBABILITY COMPARISON. 52
FIGURE 32. THROUGHPUT COMPARISON. 53
FIGURE 33. BLOCKING PROBABILITY COMPARISON. 53
FIGURE 34. DROPPING PROBABILITY COMPARISON. 54

Glossary

Call blocking: rejecting the service request of a user who attempts to start a new session.

Call dropping: interrupting an ongoing session. In the context of this thesis, this term applies only to handoff calls.

Heterogeneous wireless networks: a system of an overlay of wireless access networks of different technologies that are integrated so that mobile terminals can connect to any of those networks.

Joint radio resource management: the process in which radio resource, typically bandwidth, is allocated to network users in a heterogeneous wireless network environment for their connection sessions to maximize resource utilization as well as user satisfaction.

1 Introduction

1.1 Towards Heterogeneous Wireless Networks

With the increasing need for more efficient and user-friendly communication services and the subsequent strong business drive, wireless technologies are rapidly growing in both capacity and quality of service. The ease of connection, freedom of movement and low cost of subscription make radio access networks preferable to landline access networks, although there are disadvantages such as lower data rate.

Currently, the most-used radio access technologies are cellular networks. The first and second generations of cellular networks were circuit switched and hence their sole service was voice communication. The first-generation cellular system (1G) was short-lived and was replaced by the second-generation cellular system (2G), with the main improvement being a migration of speech channels from analog to digital. The most popular 2G standard is the Global Mobile System for Mobile Communications (GSM). The growing interest of mobile users in non-voice communications like short message service (SMS) and the need for improved data rate led to the evolution of GSM towards General Packet Radio Service (GPRS) and Enhanced Data Rates for GSM Evolution (EDGE), which incidentally were given the names 2.5G and 2.75G, respectively. GPRS and EDGE are able to provide reduced forms of Internet access: a notable example is the wireless access protocol (WAP).

However, GPRS and EDGE rely on GSM's infrastructure, which was originally designed for voice communication. This led to a significant limitation on data services. In 1999, the International Telecommunication Union (ITU) defined the International Mobile Telecommunications-2000 (IMT-2000) as the global standard for third-generation (3G) wireless communications. 3G wireless networks enable mobile subscribers to access various multimedia services (voice, data, video, etc) with greater network capacity through improved spectrum efficiency. Although 3G networks outperform previous counterparts in many parameters, the most outstanding development is a migration towards a completely packet-switched system. Wideband Code Division Multiple Access (WCDMA) is a popular evolution of GSM/GPRS for successfully becoming a 3G technology.

Despite the fact that the latest cellular technologies enable users to enjoy non-voice services, they are predominantly being used for voice calls. Even though highly robust standards have been established, technologies earlier than 3G still dominate the cellular network market significantly. One of the reasons for slow deployment of the latest standards is the relatively expensive license for the frequency band they require (eg. a pair of 5MHz bands for WCDMA). Therefore, cellular networks in general do not fulfill subscribers' expectations, especially for Internet connection.

There are multiple other forms of radio access technologies as well. The 802.11 family of wireless local area networks is a rapidly growing means of Internet connection. 802.11 networks offer broadband access to users at a relatively cheap cost as they operate in a frequency band that does not require licensing. However, users are limited to local area roaming during sessions. Satellite networks are also alternative means of communications for reasons such as very wide coverage area. However, they are less popular because of business issues such as the expensive deployment capital required, as well as technical issues such as high packet delay.

As mentioned above, diverse radio access technologies exist today, each offering a different option in terms of its coverage area, data rate, cost, etc. Additionally, the trend shows that existing technologies are being improved and new ones are emerging to offer better performance and cost to end-users and service providers. Various access networks are often deployed in a similar geographical area and form an overlay of networks, as shown in Figure 1. However, individual wireless networks typically have their own subscribers and a user running a certain application can connect only to a given network. If the user moves out of the coverage area of the radio access network, the user will be disconnected and service will be interrupted. A more ubiquitous service can be achieved if all the available wireless access networks are integrated.

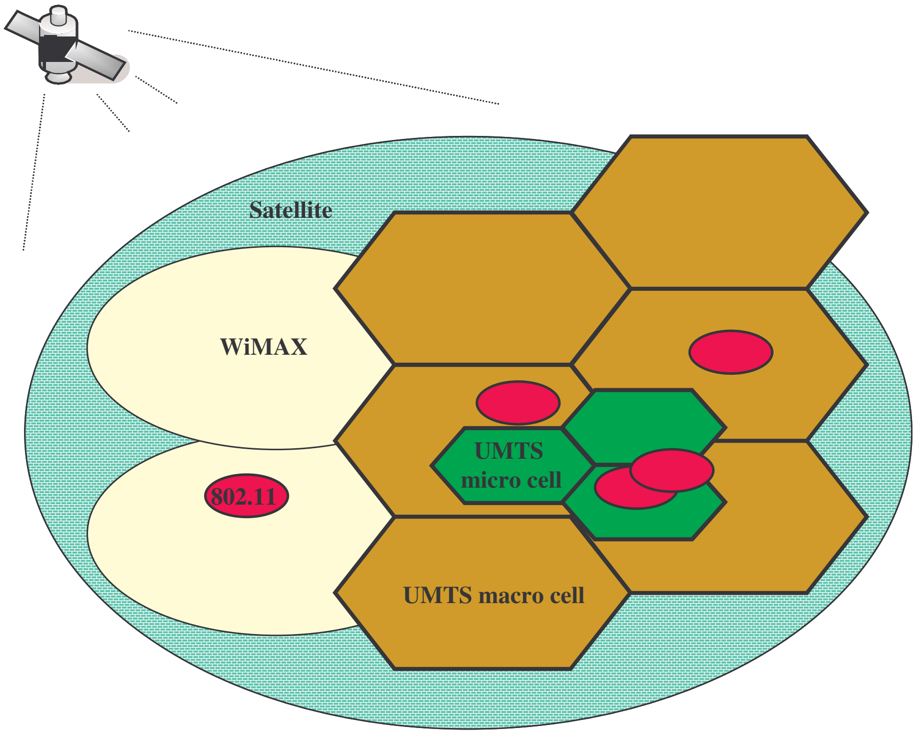

Figure 1. Overlays formed by different wireless access technologies.

The idea of benefiting from integrating the different wireless access technologies has led to the concept of beyond-3G wireless networks known as heterogeneous wireless networks. In such a heterogeneous environment, any user running any mobile application on any device can connect to any of the available radio access networks. This feature will not only bring a much more ubiquitous wireless network access, but will also create the capability of matching the best radio access network with the intended user during session initiation. The other new feature will be the so-called vertical handoff: switching of a user on an ongoing session from one radio access technology to another. In this work, we use the terms “next generation wireless networks/systems” and “heterogeneous wireless networks” interchangeably.

In order to realize the envisioned next generation wireless networks, many steps including standardization of some features of individual radio access technologies are crucial. The fact that radio access networks are evolving to offer packet-switched services enables users of any application to connect to a wider range of radio access networks. Also, many of the most recent radio access technologies like UMTS have the capability to use the IP backbone as their core networks. This is an important step since in the widely popular architecture of the next generation wireless networks called loose coupling, the different wireless access networks adopt the IP backbone as their core network [6]-[9]. The global roaming of mobile terminals will be supported by the IP backbone [19].

The next generation wireless networks also demand several new functionalities. Advanced mobility schemes should be implemented to enable users to roam seamlessly from one radio access network to another during a session. There are also other issues related to authentication and security. Our research also focuses on one such issue: radio resource management.

An overlay of two or more different wireless access technologies must exist before heterogeneity can be realized. A hierarchical cellular network is a good example of overlay network. Hierarchical cellular networks are formed when the coverage areas of different tiers of the same access technology but with different cell size overlap and are integrated. Service providers deploy such a hierarchy to explore advantages in delivering service according to user characteristics [25]. For example, connecting highly mobile users to wider tiers and vice versa would reduce the number of handoffs undergone by the highly mobile users and hence results not only in less signaling, but also increases the stability of the overall system.

Despite the fact that hierarchical cellular networks are not heterogeneous wireless networks, technological advance in hierarchical cellular networks is likely to have a significant impact on heterogeneous wireless networks, as the two scenarios have much in common.

Currently, there is a strong restriction of some types of mobile terminals to certain types of access networks. For example, people would most probably think of using a laptop to connect to a WLAN and a cell phone to a cellular network; the reverse is not straightforward to consider.

Nevertheless, electronics has also advanced to enable more than one radio interface to be implemented on the same mobile terminal [26]. Some mobile terminals have already been introduced to the market. Apple's iPhone and Nokia's E60 are good examples that enable users to connect to at least three radio interfaces including GSM, WCDMA, and IEEE802.11. Given that users have the business eligibility (eg. subscription), they are able to initiate connectivity through any of these access technologies. This is an important advance towards the next generation wireless networks.

The above capability of connecting to a choice of multiple access networks requires manual configuration. The “intelligence” of the system to connect/switch the mobile terminal to a preferred access network that is best in terms of overall performance is a requirement for realising heterogeneous wireless networks.

1.2 Joint Radio Resource Management in Heterogeneous Wireless Networks

When a mobile user attempts to connect to a wireless network and has been authenticated, there must be enough network resource available to provide the service requested to the user. This resource might be a vacant resource (which is not currently being used), a result of preemption of a less prioritized user, or a result of degradation of quality of service (QoS) of some ongoing sessions. Still, availability of enough network resource does not guarantee accepting the user. The user should be granted the service only if providing the available network resource to it is a better decision in terms of overall performance than reserving it for a higher priority user or purposes such as network management signaling.

The network resources can be queue buffers, power, bandwidth, etc. However, when dealing with resource allocation problems, many people carrying out research on wireless networks focus on bandwidth to achieve a better scheme for radio spectrum utilization. While some of the frequency bands are free to be used unlicensed, some are licensed and expensive. In both cases the frequency spectrum is a limited resource.

As applications are growing to demand more and more bandwidth, on the one hand research is being carried out into increasing the effective data rate for a given frequency band - examples include research on turbo coding in CDMA networks and carrier sensing in IEEE802.11 networks. On the other hand, some work, including this paper, concentrates on how to utilize the bandwidth effectively to minimize wastage of resources, while at the same time dealing with reservation issues with regard to prioritizing certain service requests. This study concerns what is referred to as radio resource management (RRM). An example of such schemes is radio resource control (RRC) in Universal Mobile Telecommunications Systems (UMTS) [1]. There are also homogeneous systems such as IEEE802.11 (a/b or g) that do not implement any

RRM scheme. Users are admitted to the system without any limit and the bandwidth is equally shared among the active users.

Extensive work has been carried out to achieve a good RRM scheme in homogeneous wireless networks where a single radio access network technology consisting of at most multiple cells is managed independently of other access network technologies.

Joint radio resource management is the process in which radio resource, typically bandwidth, is allocated to network users in a heterogeneous wireless network environment for their sessions to maximize resource utilization as well as user satisfaction. The term “joint” in the above definition indicates that radio resources are managed jointly among different radio access technologies in the heterogeneous wireless network scenario. [37].

1.3 The Need for Advanced Joint Radio Resource Management

Radio resource management (RRM) schemes for homogeneous networks have been amply discussed and consequently, many approaches have been proposed in the literature and some schemes have been implemented in the market. In general, the design goal of such schemes is to optimize the trade-off between resource utilization and quality of service during a request by a user to connect to a network. The decision is made independently of other access networks in the vicinity of the user as the user has access only to a single wireless network.

However, the heterogeneity of the next generation wireless networks requires new approaches, as users will have choices from multiple wireless access networks. In fact, they will even be able to connect to multiple wireless access networks simultaneously. Therefore, the aim of joint radio resource management (JRRM) is to implement admission decision policy that optimizes the *overall trade-off* between resource utilization and quality of service in a heterogeneous network scenario. This means admission decisions should take into account a wide range of multiple networks and users that have access to them. As heterogeneous wireless networks are in the development stage, many essential features are little explored and JRRM is no exception. Although researchers have made significant contributions, much more is needed to maximize the benefits from the real features of component networks as well as the global picture of the heterogeneous topology by using realistic mathematical and/or simulation models or test beds. Our work is a design approach to JRRM by studying the impact of the protocols that component networks use on the overall system performance.

1.4 Scope and Contributions of this Work

The scope of the work in this thesis is two-fold; firstly we study serving disciplines as part of joint radio resource management (JRRM). Here, we focus on the general principle in which a policy deals with connection requests given a certain amount of bandwidth whether the network scenario under consideration is homogeneous or heterogeneous. After comparing different approaches, we propose a serving discipline scheme to be employed in our JRRM. Secondly, we discuss the design of our JRRM scheme for heterogeneous wireless networks. Our JRRM scheme will take into account the multiple access networks and the dynamics of the network resources and traffic statistics and will employ our serving discipline to make decisions during connection requests.

Our work is an attempt to show the impact of a realistic assumption of network resources, traffic nature, and their interactions on the design of a JRRM protocol. For example, we exemplify the complexity of the design due to the imperfection of the overlap of coverage of different wireless access technologies. We will use the Network Simulator (commonly known as NS-2) for simulating a heterogeneous wireless network environment consisting of IEEE802.11 and Universal Mobile Telecommunications Systems (UMTS) networks.

1.5 Outline of the Thesis

The rest of the thesis will be organized as follows. In Chapter 2, we will present a literature review on radio resource management both in homogeneous and heterogeneous wireless networks. In Chapter 3, we will propose our serving discipline that we will apply as a sub-functionality of the JRRM. In Chapter 4, we will present our proposed JRRM system architecture. In Chapter 5, we will present the proposed JRRM decision-making policy and its analysis. Finally, in Chapter 6, we will conclude the paper and list possible directions towards research on JRRM.

2 Literature Review on Radio Resource Management

Radio resource management (RRM) is a general term that applies both to homogeneous and heterogeneous wireless scenarios. In homogeneous scenarios, RRM is generally discussed with a contextual meaning that is equivalent to serving discipline. However, in heterogeneous scenarios, beyond the concept of serving discipline, it involves more complex functionalities as it considers diversity of access technologies and network coverage. To distinguish the RRM in homogeneous and heterogeneous scenarios, we use the terms "serving discipline" and "joint radio resource management", respectively.

2.1 Serving Discipline Schemes

Upon a connection request (traditionally called a call arrival), a serving discipline scheme decides whether or not to admit the call, predicting the impact of doing so on the system. In simple terms, serving disciplines aim at maximizing resource utilization and user satisfaction.

As bandwidth is a limited resource, implementing a highly efficient serving discipline is a major objective as it enables operators to reduce costs and increase revenue. At the same time, users should be guaranteed services in acceptable limits of QoS parameters (call-dropping probability, data rate, etc). An acceptable limit of a QoS parameter can be determined both by the user's profile and the type of application (real-time, non-real-time, etc).

Fairness is also an important component of QoS. For example, considering the simultaneous arrival of two users with different priority levels and demanding an equal amount of service (eg. data rate), there should not be a possibility of admitting the lower priority user while rejecting the higher priority one.

Often, maximizing QoS is done at the cost of reducing resource utilization, and vice versa. For example, reservation of large amounts of bandwidth for higher priority users will increase the satisfaction of those users. But the reservation might cause underutilization. Another important consideration is simplicity. This is because complexity, especially at the moment of call arrival, is not desired. Therefore, a good serving discipline seeks to optimize the trade-off between resource utilization, QoS, and simplicity.

The simplest serving discipline is first come-first served. However, it offers no priority of service. Another instance of a simple serving discipline is the one presented in [14], which is based on complete partitioning (CP) as illustrated in Figure 2. In this scheme, the bandwidth capacity C is divided into segments where each type of call can be admitted only to a segment assigned to it. The major problem associated with this algorithm is the waste of bandwidth. A call request to a full segment is rejected even if there is enough bandwidth in a segment assigned for lower or equal priority service calls. The implementation of a pair of uplink and downlink channels in cellular systems like UMTS is an example of such a discipline.

Figure 2. An example of complete partition.

In [5], a “movable” boundary system was proposed based on CP to support voice and data calls. In this algorithm, the boundary for the partition is adjustable to deal with traffic changes in the system. The above scheme was extended in [17] using the so-called guard channel policy to differentiate between new and handoff voice calls. As shown in Figure 3, according to this scheme, called ‘dynamic partition’ (DP), K1 out of the C channels of a cell are reserved for voice calls (new/handoff) and K2 channels are reserved for data calls. The remaining (C - K1 - K2) channels are shared in a fair manner by both voice and data calls. New voice calls can only use K3 out of the K1 channels.

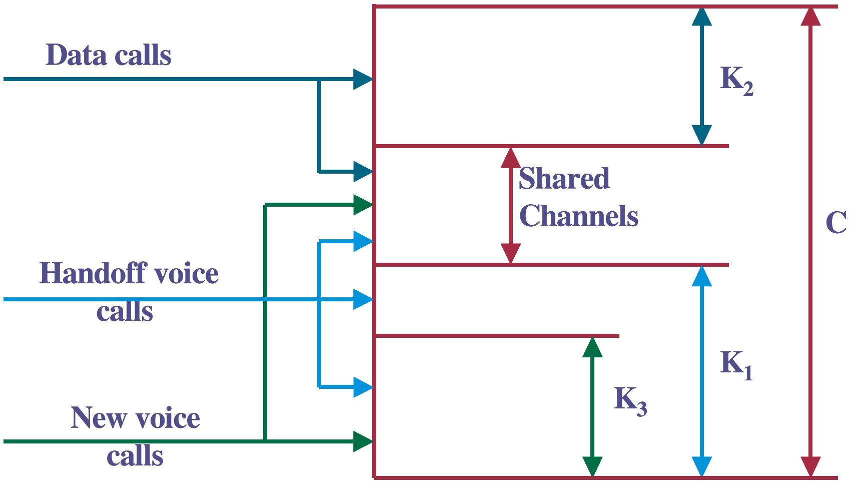

Figure 3. Bandwidth allocation in the DP scheme.

According to the DP algorithm, when a new voice call arrives, a channel is sought in K3. If a channel is available there, it will be admitted. Otherwise, a channel is sought in the shared area. It will be blocked if there is no channel in both K3 and the shared area. When a handoff voice call arrives, a similar search is done in the voice-only area and then in the shared area. It will be dropped only if no channel is available both in the voice-only and shared area. A similar decision is made for a data call arrival by first searching in the data-only area and then in the shared area. With reference to the DP algorithm, the system suffers from inefficiency in the cases where there is no bandwidth available for a voice call (new/handoff) while there is enough bandwidth in the data-only area. This is a waste of radio resource since data calls should receive lower priority than voice calls.

As with the DP algorithm, many other research contributions are also based on a hybrid of partitioning and sharing. For example, the work in [30] divides the channel into voice, data, and shared channels in a way similar to that of the DP algorithm. The most significant difference, however, is that if there is no bandwidth available to a data call, packets will be queued in a limited amount of buffer in the hope that a channel will be released for it within a very short time.

In [15] and [16] the authors proposed the dual threshold bandwidth reservation (DTBR) scheme for voice/data cellular networks. The authors in [17] extended the DTBR scheme to support variable (or elastic) data bandwidth requirements and studied its performance. This scheme is based on complete sharing (CS) and the authors showed that it offers higher network utilization than the DP algorithm while meeting guaranteed QoS. The authors in [17] also showed that one of the good features of the DTBR algorithm is that a higher priority call is never rejected while there is a bandwidth available for a lower priority call. In this work, we will repeatedly mention the DTBR scheme, as we believe it is a good scheme based on multiple parameters that we take into account.

In the DTBR scheme, the C channels of the cell are divided into three regions by two thresholds K1 and K2 (K1 > K2), as shown in Figure 4. When the network occupancy level L is lower than the threshold K2, both voice and data calls can be admitted into the system. When L is greater than K2, no data call can be admitted into the system. When L is greater than K1, no data call or new voice call can be admitted to the system. A handoff voice call will be dropped only if there is no channel available.

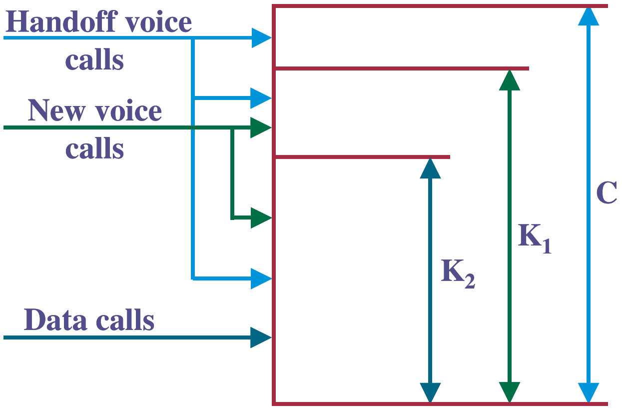

Figure 4. Bandwidth allocation in the DTBR scheme.

In general, non-real-time calls, for instance data calls, are flexible with regard to the amount of bandwidth they require. This allows a generous amount of bandwidth degradation of ongoing calls to receive new ones.

In [20], a bandwidth adaptation scheme based on *per flow* degradation was proposed by defining a term called “degrade profile”. In the cases where the bandwidth of an access network is full and the ongoing calls are using bandwidth more than the minimum of their requirement, degradation is implemented upon arrival of new or handoff calls. In order to admit calls, this scheme degrades the longest calls in the system with the hope that those flows have greater probabilities of quitting the system and leaving fewer degraded connections. The authors in [21] proposed a bandwidth adaptation scheme based on *per class* degradation. Here, in order to admit a call, the lowest possible priority class calls are degraded. Ongoing higher priority class calls are not affected by arrival of lower priority calls. The performance analysis in [21] showed that the per class adaptation scheme is better in terms of fairness (by treating flows of the same class equally) and simplicity. But the per flow scheme is better in terms of resource utilization.

The authors in [17] extended the DTBR scheme by considering the possibility of bandwidth degradation of elastic data calls. According to this scheme, data calls, unlike voice calls, are flexible to a limited extent in the amount of bandwidth they require. Therefore, they carried out analysis based on a model in which voice calls strictly request 1 unit of bandwidth while data calls request 2 units of bandwidth but are flexible to receive less bandwidth to a minimum of 1 unit. This mitigates blocking/dropping of calls by enabling degradation of ongoing or new calls. Here, upon arrival of a new voice call, if bandwidth is not available for it, the scheme checks whether the occupancy level would be less than K1 (refer to Figure 4) by equally degrading all ongoing data calls as little as possible to a minimum of 1 unit. If the check is positive, the call will be admitted at the cost of degrading the data calls; otherwise, it will be blocked. If a handoff voice call arrives while the occupancy level is C, the scheme checks if the occupancy level would be less than C given that the ongoing data calls are degraded. If the check is positive, the call will be admitted at the cost of degrading the data calls; otherwise, it will be dropped. The same procedure is carried out upon arrival of data calls while occupancy level is greater than K2, but attempts are made to admit the call at the cost of degrading the bandwidth of the incoming and all the ongoing data calls. The degradation in the DTBR scheme can be categorized as a per class degradation scheme.

Departure of calls is also an important issue in radio resource management. Departure of calls from an access network can happen either because of end of sessions or because of handoffs. The optimization problem here is how to allocate efficiently the released bandwidth to ongoing calls or reserve it.

According to the per flow scheme in [20], when a call departs from the system, an attempt is made to assign the released bandwidth to the shortest call. According to the work in [21], when a call departs, the per class scheme allows the highest priority class to use more bandwidth.

According to the DTBR scheme in [17], when a call departs from a cellular network, the released bandwidth is shared by the ongoing calls until each data call gets a maximum bandwidth. However, since fixed bandwidth is allocated to voice calls, no change of bandwidth is made for them.

Some efforts have also been made to ensure fairness. In [30] for example, a call admission control scheme was proposed with the aim of maximizing fairness in terms of call blocking probability and call dropping probability. The work argues that data traffic faces a higher call blocking probability than voice traffic, since data traffic requires more bandwidth than voice traffic. For this reason, data traffic users are rarely admitted. The paper proposed a method to overcome this unfairness by allocating a buffer for the underprivileged data traffic. It also applied the biased coin method [31] to achieve long-term fairness based on short-term fairness level. According to this method, a traffic class (example data) that is significantly underprivileged will deliberately be privileged in the future.

We argue that in the above biased coin approach, although we can achieve class-wise fairness, we do so by worsening unfairness among individual calls. For example, the scheme can privilege a data traffic user because data traffic users were underprivileged in the past. But this means a voice user is paying the cost of what voice users (probably other individuals) have unfairly used in the past.

More complex serving-discipline algorithms have also been proposed. A few of the most interesting ones are based on fuzzy logic, auctioning, and genetic algorithms. Due to the scope of the thesis, we will not discuss them.

In summary, as users arrive randomly, an intelligent admission control scheme optimizes the trade-off between resource utilization and QoS by effectively predicting future demands. Long-term prediction of system conditions should be based on cumulative effect of repeated measurements. Therefore, serving disciplines should be both reactive and predictive in terms of deciding dynamic parameters like acceptance thresholds.

2.2 Joint Radio Resource Management in Heterogeneous Wireless Networks

Although not standardized yet, joint radio resource management (JRRM) is believed to be a vital functionality in the next generation wireless networks. Many researchers are involved in JRRM although they use different terminologies. The following terms are found in the literature to refer to the same functionality: joint radio resource management (JRRM) [34], [35] common radio resource management (CRRM) [23], [24], and joint call admission control (JCAC) [3], [10].

Much of the work focuses on how decisions are made by JRRM schemes. Early work on JRRM includes the proposal in [36] to deal with policy-enabled handoffs across heterogeneous wireless networks. The work is an effort to allow users to express policies on what is the "best" wireless system at any moment, and make tradeoffs among network characteristics and dynamics such as cost, performance, and power consumption. Similarly, the work in [18], [20], [21], and [22] presents an adaptive policy-based access management system to support heterogeneous wireless networks.

In [27], a multi-layer predictive admission control policy (MLPAC) was proposed to minimize global blocking probability while guaranteeing a hard constraint on handoff dropping probability. The work is an extension of the overflow scheme [28], [29] used in two-layer hierarchical cellular systems (HCS) to heterogeneous wireless networks. In the overflow scheme, when a connection is rejected by a micro-cell, its admission is considered in the overlaying macro-cell.

Much research is also specifically dedicated to JRRM in cellular/802.11 heterogeneous environment [3], [20], [23], [32], and [33]. This is due to the fact that cellular networks are consistently the dominant means of wide area wireless network access and 802.11 networks are the most popular means of local area wireless network access. Due to the difference in the service they offer, especially the fact that cellular networks enable wide area mobility and the IEEE802.11 family of networks in general offer high data rate for low cost, a significant beauty of cellular/802.11 interworking is hoped to be revealed.

In [32] JRRM was studied for data and voice QoS support in cellular/WLAN interworking. The work proposes an admission strategy that prefers the cellular networks for voice service and WLAN for data services. For instance, whenever a new voice call arrives at a cellular/WLAN overlay area, it will first try to gain admission to the cellular network. If there is not enough resource in the cellular network even after degrading ongoing data calls to their minimum required bandwidth, then it will overflow to the WLAN to request admission. The reason behind this is that connecting voice calls in the cellular networks minimizes the frequency of vertical handoff and consequently reduces the impact of processing overhead and handoff latency. Besides, the fine QoS guarantee provided by cellular networks meets the requirement by voice calls. When a new data call arrives in such an area, it will try to gain admission to the WLAN. If the admission request is rejected by the WLAN due to lack of sufficient bandwidth, it will not try the cellular network for fear it may severely affect the capability of the cellular network in carrying voice traffic.

The main problem associated with the above scheme is the unnecessary data call blocking during the times when there is sufficient resource in the cellular network and insufficient resource in the WLAN. The integrated system should instead utilize the bandwidth of the cellular network during such cases to optimize efficiency as well as user satisfaction.

The authors in [3] presented a JRRM scheme for an integrated UMTS/WLAN network. Because of the fact that WLAN technologies are short-range networks that operate in unlicensed frequency bands, they offer lower cost to users than UMTS does. Thus, the vertical handoff from UMTS to WLAN was defined as a desirable handoff. On the other hand, the mobile terminal's connection has to remain seamless as a user connected to the WLAN roams out of the WLAN domain. Thus, the handoff from the WLAN to the UMTS was defined to be a necessary handoff. The work in [3] gives a higher priority to necessary handoffs than to desirable handoffs.

Departure of calls in heterogeneous wireless networks should be handled in a different way from that of homogeneous networks. The methods applied in the case of homogeneous wireless networks aim only to allow the bandwidth released in an access network by call departures to be used by the ongoing calls. Those methods do not account for avoiding high difference in network loads (i.e. creating load balance) between the different access networks that might result from departures of calls.

In [32] and [33] a policy framework for load balancing in cellular/WLAN integrated networks was presented. The authors in [42] presented a load balancing mechanism based on cooperative games. In [11], we proposed a load balancing system architecture and resource allocation policy.

The extra signaling for vertical handoffs function should be kept as low as possible so that it does not consume a lot of network resources as well as mobile terminal's power. Further, in [2] it was shown that frequent bandwidth switching among different bandwidth levels can be worse than a large QoS degradation ratio. Therefore, although we wish to achieve ideal load balance at any moment, we have to limit the frequency of load balancing.

Bandwidth efficiency, blocking probability, dropping probability, and fairness are not the only performance metrics of a good JRRM. And hence, JRRM should be studied from different perspectives. In [25] an extensive effort was made on the JRRM framework. Among other issues, the work deals with controlling the operational cost of JRRM modules. In [10] and [11] we proposed a JRRM system architecture in which we focused on how the different sub-functionalities of the JRRM should be implemented in the different parts of the integrated system. The aim of the work is to create a hierarchy that will avoid unnecessary signaling and processing at certain stages of the JRRM that are specific to an access network technology. We will describe it in Chapter 4 with some amendments and a few notional changes.

Still, many practical issues remain open. The diversity of wireless access networks and the resulting complexity of interworking scenarios between different access networks leave many challenges unobserved at research level. The fact that many of the analytical as well as simulation models of heterogeneous wireless networks are not based on fairly realistic assumptions often affects the practicality of the research outputs.

For example, the algorithms presented in [38]-[40] are based on the assumption that multiple radio access technologies are deployed with collocated cells (i.e. perfect overlap of cells). The papers suggest that such type of overlap is technically feasible and may also save installation cost. We argue that although these opinions are right, assumption of collocated cells is very simplistic. Currently, cellular network vendors generally deploy base stations at different places, with different cell sizes. The same is true for WLAN owners. Therefore, there should be a

JRRM algorithm that accommodates a more realistic scenario in which radio access technologies may overlap in a random fashion or may not even overlap at some places.

Another instance of unrealistic assumption is a constant amount of network capacity assumed in many papers such as [27] and [41]. In reality, the capacity of a network generally varies according to traffic and environmental conditions. For example, high interference generally results in a decrease in the achievable throughput of the system. The MAC protocol in IEEE 802.11 networks is also an interesting example. As the number of active client users increases, contention of packets increases, which reduces the achievable throughput of the network.

3 Proposed Scheme for Serving Discipline

3.1 Description

In our model, two types of calls arrive at a wireless access network; real-time (RT) and non-real-time (NRT), which categorize calls such as voice communication and downloading a web document, respectively. Applications are becoming more and more diverse and systems are also converging; hence, one should expect the arrival of any type of application with any demand to arrive in the system under consideration. RT and NRT calls may as a consequence demand any data rate characterized by exponential distribution, which is widely accepted as a distribution for telecommunications traffic.

A real-time application user demands a constant data rate for the entire session, while a non-real-time user can be served with a more flexible and varying data rate as long as the amount of granted bandwidth is not less than a certain minimum acceptable data rate.

The unlimited traffic demand expectation is not a mere assumption; it matters in the design issues of a radio resource management scheme. For example, assuming such an unlimited demand, the assumption of the minimum acceptable data rate of NRT users can have a significant impact on the overall performance of the system. Let us consider a situation where the occupancy level is high and the serving discipline allows any type of call to be admitted. Intuitively, if the minimum acceptable data rate for NRT calls is too low, more NRT and fewer RT users will have the chance to be admitted to the system, and vice versa.

As we mentioned in Chapter 2, we believe that the approach in the dual threshold bandwidth reservation (DTBR) scheme with elastic data model is a good scheme in terms of network utilization and user satisfaction. However, it assumes that data users are not as sensitive to service interruption as voice users are. Therefore, it suggests that handoff data users should be treated with the same priority given to new data users, while handoff voice users should receive the maximum priority.

Let us define the following acronyms:

PND: priority given to new data calls

PHD: priority given to handoff data calls

PNV: priority given to new voice calls

PHV: priority given to handoff voice calls

As shown in Figure 5(a), according to the DTBR scheme, PND = PHD <= PNV <= PHV.

However, we believe that data call users are also sensitive to service interruption. There are even situations in which the data service interruption can be worse than voice call interruption. For example, many people would prefer their voice communication with someone to be interrupted than the interruption of a file download that has already progressed for a long time. Hence, there are people as in [30] who believe the opposite of the beliefs expressed in the design of the DTBR scheme on this matter. As the subject of giving a higher priority between handoff data calls or handoff voice calls is controversial, in this work we decide to give them equal priority. Therefore, in our work, the priorities are in the order: PND < = PNV <= PHV = PHD.

Let us define the following terms for a system:

C: the capacity or the maximum achievable throughput of the system.

L: the occupancy level L of the system given that an incoming call is admitted.

As depicted in Figure 5(b), our serving discipline can be described as follows. The capacity of the system is divided into three slots using two thresholds K1 and K2 (where K1 >= K2). When a new NRT call arrives, it will be admitted if the occupancy level L is less than K2. Otherwise, L is calculated by assuming that all ongoing NRT calls and the incoming call are degraded equally by as little as possible and to a minimum of their minimum acceptable data rate. If this calculated value of L is less than K2, then the call will be admitted; otherwise, it will be rejected. When a new RT call arrives, it will be admitted if L is less than K1. Otherwise, L is calculated by assuming that all ongoing NRT calls are degraded equally by as little as possible and to a minimum of their minimum acceptable data rate. If this calculated value of L is less than K1, then the call will be admitted; otherwise, it will be rejected.

A similar procedure is carried out as above during handoff RT and handoff NRT call arrivals, with the only difference being that they are admitted as long as L is less than or equal to C.

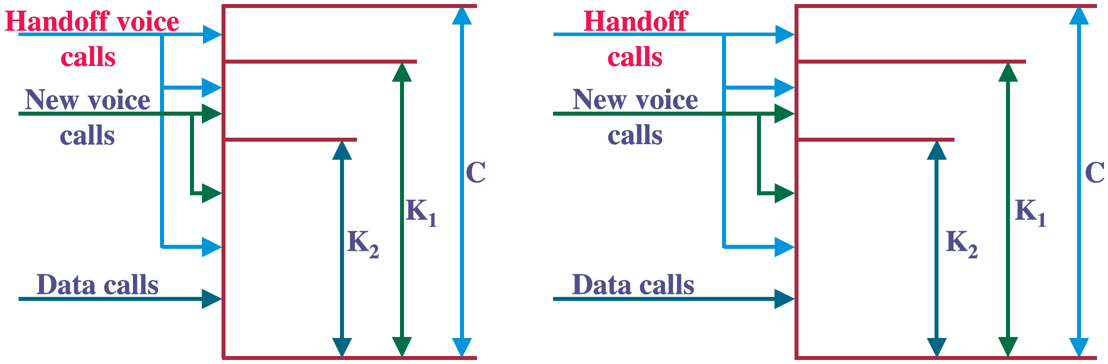

Figure 5. A comparison of serving discipline schemes.

(a) The DRBR scheme (b) The modified DTBR scheme.

Therefore, our serving discipline scheme is very similar to the DTBR scheme except that it treats handoff NRT calls with a priority as high as that of handoff RT calls. Our assumption of traffic mainly differs from that in [17] due to the fact that we consider unlimited bandwidth demand while in [17] data users request 2 units and voice users request 1 unit of bandwidth.

As handoff calls usually constitute a very small proportion of call arrivals, we believe that giving handoff NRT calls a priority as high as that of handoff RT calls does not significantly reduce the satisfaction of RT users in a system that is generally stable (a system in which aggregate users' demand is not much more than system capacity). This is because, in a system condition where L is greater than K1, it is only handoff calls that may be admitted to the system, but the departure event from the system includes calls accepted as new as well as handoff calls. In such a condition, rate of departure will be greater than rate of admission. As a result, it is very unlikely that a handoff RT call is dropped because of the admission of a handoff NRT call. This is because the admission of a handoff NRT call affects a future admission of a handoff RT call only if the departure of calls during the time between these two events does not leave a sufficient capacity for the handoff RT call. The next sections are dedicated to proving this argument through simulation.

3.2 Simulation Setup

In all of the simulations, we used the Network Simulator (commonly called ns-2) version 2.29. We used the module “mobility package for NS-2” contributed by National Institute of Standards and Technology (NIST), USA, in which the UMTS feature is based on enhanced UMTS radio access network extensions for ns-2 (EURANE).

To compare serving-disciplines, we built a UMTS radio access network in which new and handoff users request service from a UMTS network. The EURANE currently can simulate only a single cell with no clear boundary. However, building a single UMTS cell topology as shown in Figure 6 is enough to simulate the effects of new and handoff calls. We created the cell boundary by writing a code to limit the location of the arrival and connection events.

The specific access technology simulated is High-Speed Uplink Packet Access (HSUPA) category 6.

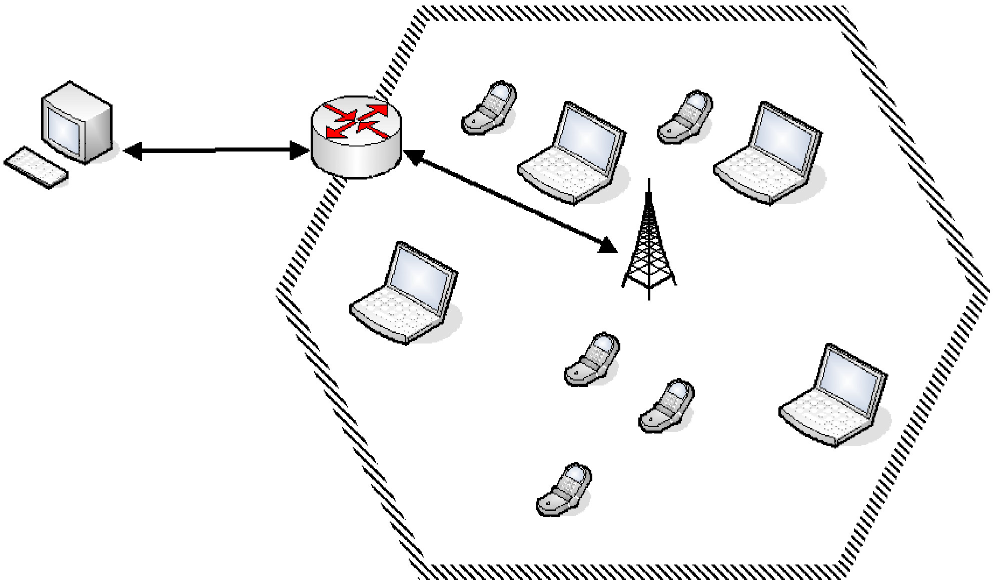

Figure 6. Simulation network topology to study serving disciplines.

The other feature that we added was limiting the achievable throughput of the system. According to the module, there is no limit on the achievable throughput of the system: as the sending rate of the users increases, the system throughput increases linearly and indefinitely. In order to solve this problem, we defined the capacity of the system by limiting the bandwidth of the wired link that connects the base station to the gateway router to create a bottleneck, which in a sense defines the system capacity.

Two types of traffic, real-time and non-real-time are considered in the simulation. A real-time application demands a constant data rate for the entire connection session, while a non-real-time user can be served with a more flexible and varying data rate. We have taken the following assumptions:

1. Real-time and non-real-time users arrive with equal probability.
2. The inter-arrival time between two consecutive users is exponentially distributed.
3. Data are sent only one-way, from mobile terminals to the base station outside the wireless domain.
4. Once admitted, a real-time user is served with a constant data rate while the data rate of non-real-time users can be increased or decreased throughout the session, depending on bandwidth availability.
5. During a call request, the call duration of the session is found by generating a random number characterized by exponential distribution. A real-time application call ends only when either this call duration ends or the mobile terminal roams out of the network coverage area. But, for non-real-time users, because data are sent faster or more slowly depending on resource availability, the call duration is recalculated whenever granted data rate changes to make it shorter or longer.
6. The initial position of the users is randomly distributed in the WLAN hotspot.
7. Mobility pattern of users is a random waypoint in the area of a rectangle, slightly bigger than the cell of the UMTS network to represent roaming of nodes inside the cell and rarely leaving the cell.
8. We do not deal with channel effects such as interference and fading.

Each data point in the results was obtained by averaging over a simulation time of 2 hours. Table 1 summarizes relevant parameters considered in the simulation.

Table 1. Simulation parameters.

Simulation Parameter	**Value**
Capacity of the system	5.76 Mbps
Guard Channel	1Mbps
Cell side length	1Km
Ratio of RT users to NRT users	1 : 1
Percentage of handoff calls to all calls (This applies both to RT and NRT)	10%
Average data rate requested by real-time users	100000 bps
Average data rate requested by non-real-time users	50000 bps
Minimum acceptable data rate by non-real-time users	50000 bps
Average call duration	120s
Movement of nodes	Random waypoint
Minimum speed of nodes	0 m/s
Minimum speed of nodes	20 m/s

More information is available in the simulation code on the accompanying CD.

3.3 Definitions of Simulation Parameters

For clarity, let us define the following parameters that we applied in the experiments.

Aggregate arrival rate: the rate at which any type of network users (real-time, non-real-time, new, handoff) arrive (attempt to connect) in a network under consideration. In our homogeneous scenario, it is the rate at which users arrive in the UMTS cell. But in the heterogeneous scenario that we will discuss in Chapter 5 where a WLAN hotspot and a UMTS cell are overlaid, aggregate arrival rate refers to the total rate at which users arrive in both the UMTS cell and the WLAN hotspot.

Aggregate throughput: the total amount of data rate at which all users are served by a network under consideration. In our homogeneous scenario, it is the data rate at which the UMTS cell serves users. But in the heterogeneous scenario that we will discuss in Chapter 5, it is the data rate at which both the UMTS cell and the WLAN hotspot serve users.

Call blocking probability: the ratio of number of blocked new calls to the total number of new call arrivals.

Call dropping probability: the ratio of number of dropped handoff calls to the total number of handoff call arrivals.

RT call blocking probability: the ratio of number of blocked new RT calls to the total number of new RT call arrivals.

NRT call blocking probability: the ratio of number of blocked new NRT calls to the total number of new NRT call arrivals.

RT call dropping probability: the ratio of number of dropped RT handoff calls to the total number of handoff RT call arrivals.

NRT call dropping probability: the ratio of number of dropped NRT handoff calls to the total number of handoff NRT call arrivals.

3.4 Results and Analysis

Here, we compare the DTBR scheme with our scheme, which we will call the modified DTBR scheme throughout. Our aim is to show that the modified DTBR scheme outperforms the

DTBR scheme in terms of dropping probability while achieving similar performance in terms of other parameters.

Figure 7 shows the throughput performance comparison between the DTBR scheme and the modified DTBR scheme. We observe very similar performance between the two schemes.

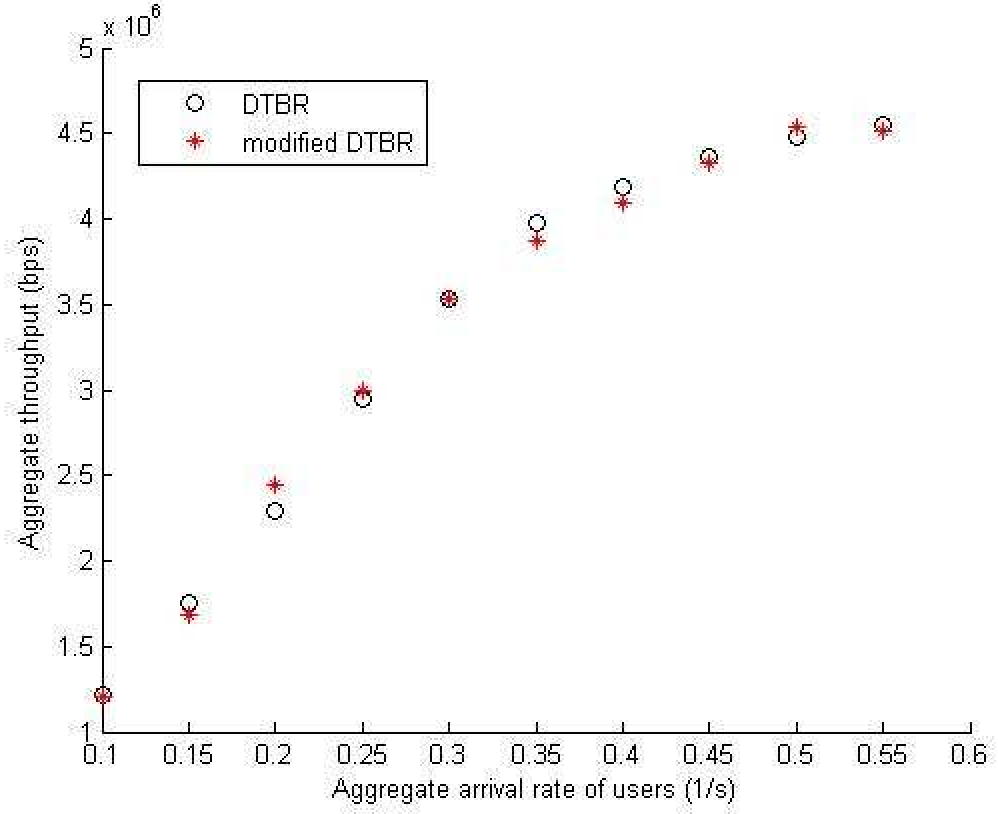

Figure 7. Aggregate throughput comparison.

In Figure 8, the blocking probability plots of the two schemes only show inconsistent variations in performance. We believe that those variations are attributed to the randomness of events in the simulation: they do not prove the superiority of one scheme over the other. The same is true with the blocking probability comparisons where the blocking probability is found from the events involving only RT or NRT user arrivals as shown in Figures 9 and 10, respectively.

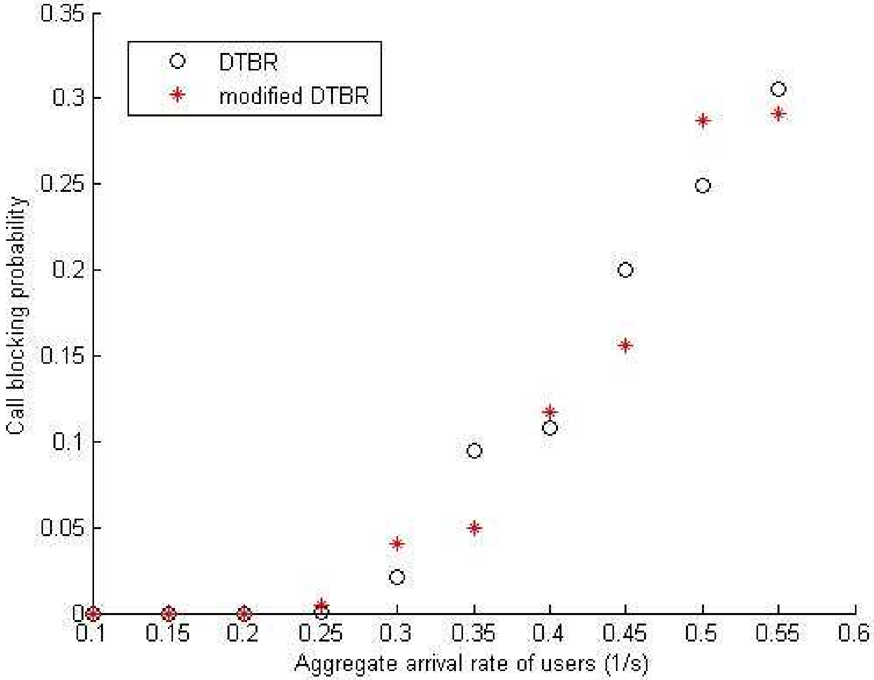

Figure 8. Blocking probability comparison.

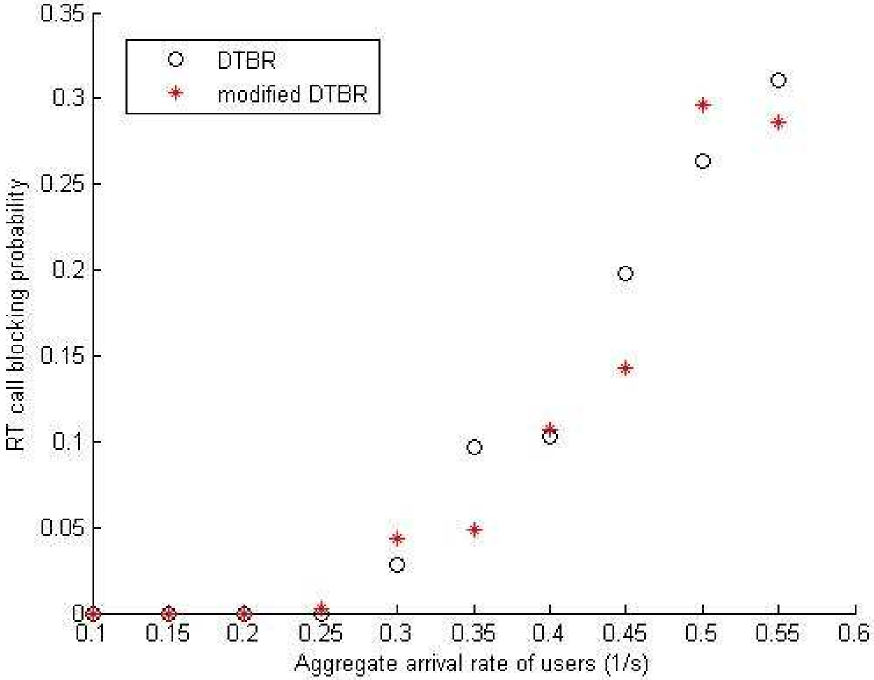

Figure 9. RT blocking probability comparison.

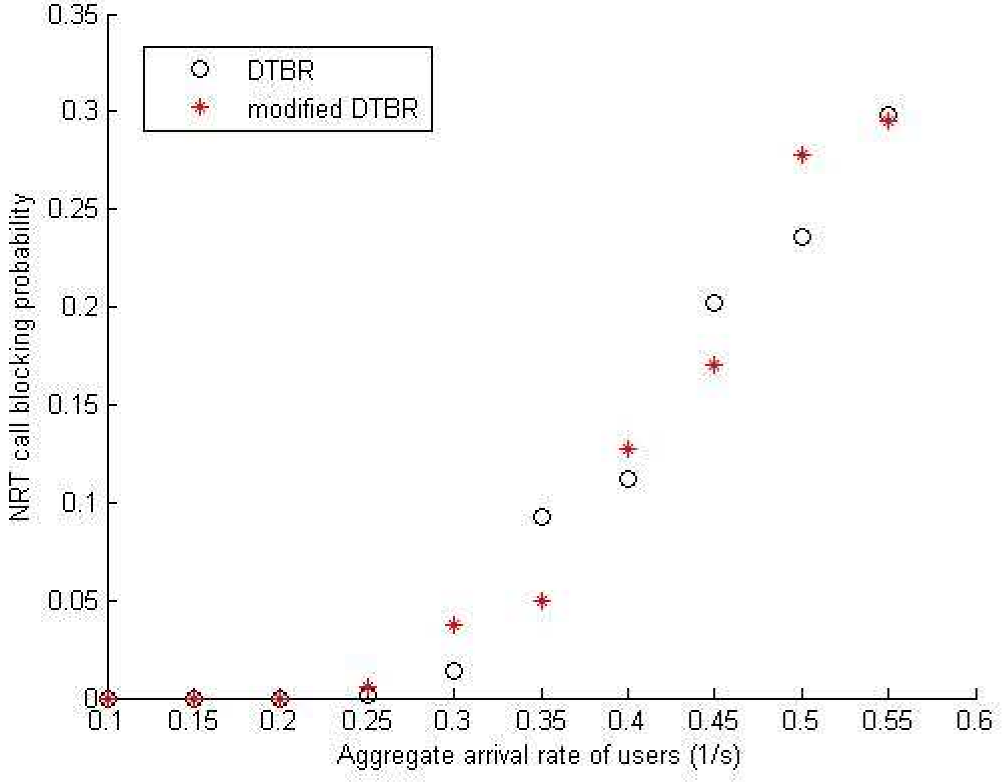

Figure 10. NRT blocking probability comparison.

We observe interesting differences when we see the plots of dropping probability in Figure 11. The modified DTBR scheme shows a big improvement. This is easily explained by looking at Figure 12 in which dropping probability was found by considering only NRT handoff events. Due to the fact that our scheme significantly reduces NRT dropping probability, and NRT handoff events on average constitute half of the handoff events, the overall dropping probability is also reduced. Meanwhile, the RT dropping probability remains similar in both schemes in Figure 13.

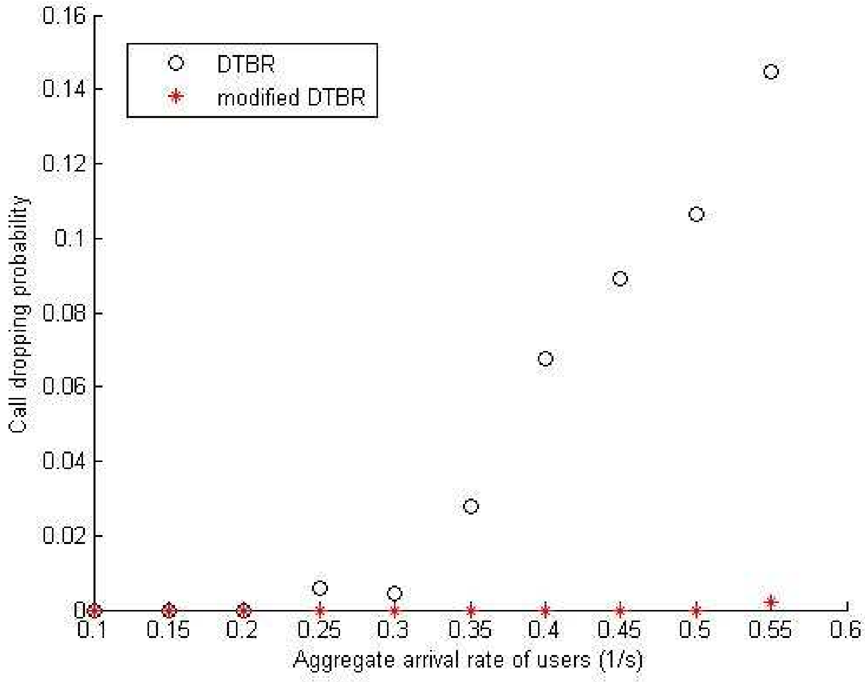

Figure 11. Dropping probability comparison.

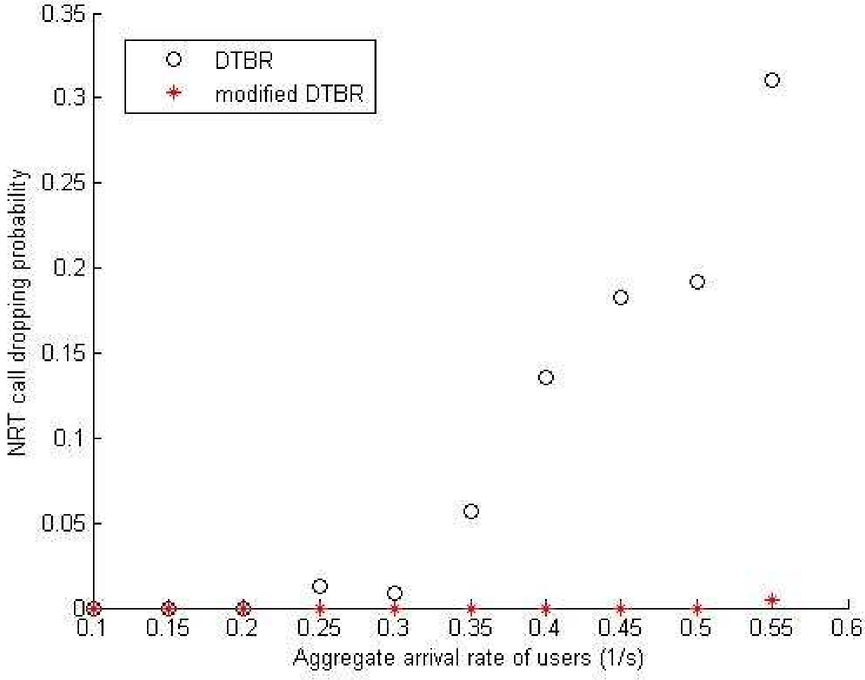

Figure 12. NRT dropping probability comparison.

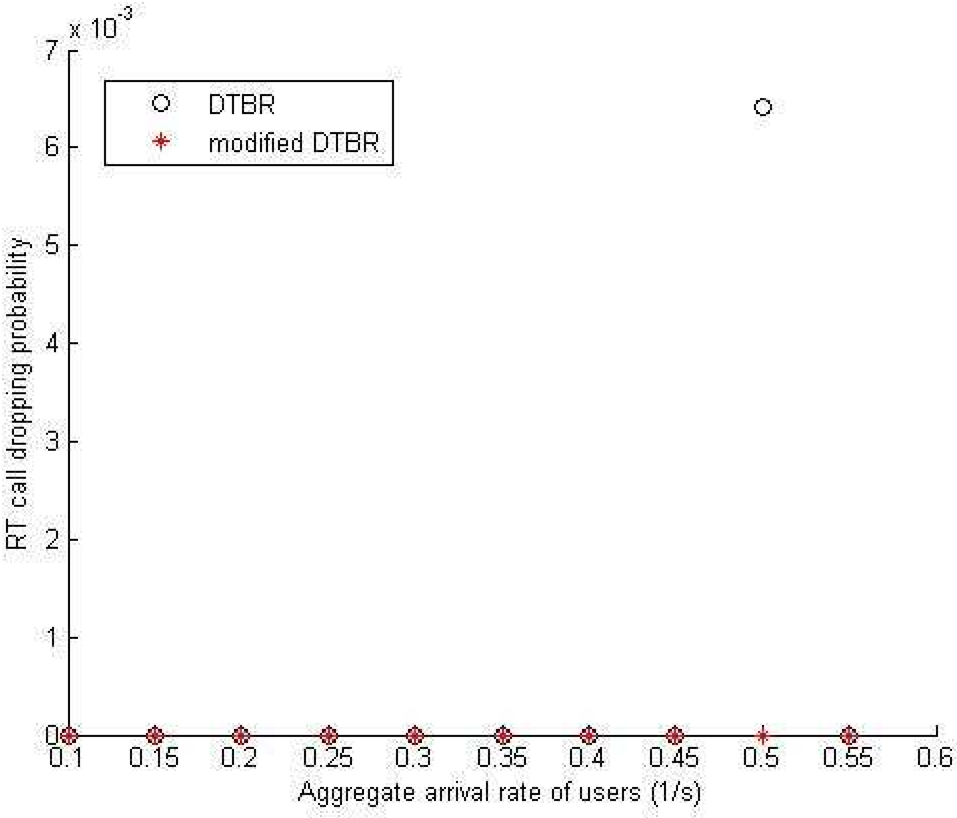

Figure 13. RT dropping probability comparison.

Finally, we also measured the average service mix of users to ensure that our dropping probability improvement is not achieved at the cost of greatly affecting the service mix. As shown in Figure 14, we can see that the service mix also does not show a noticeable change.

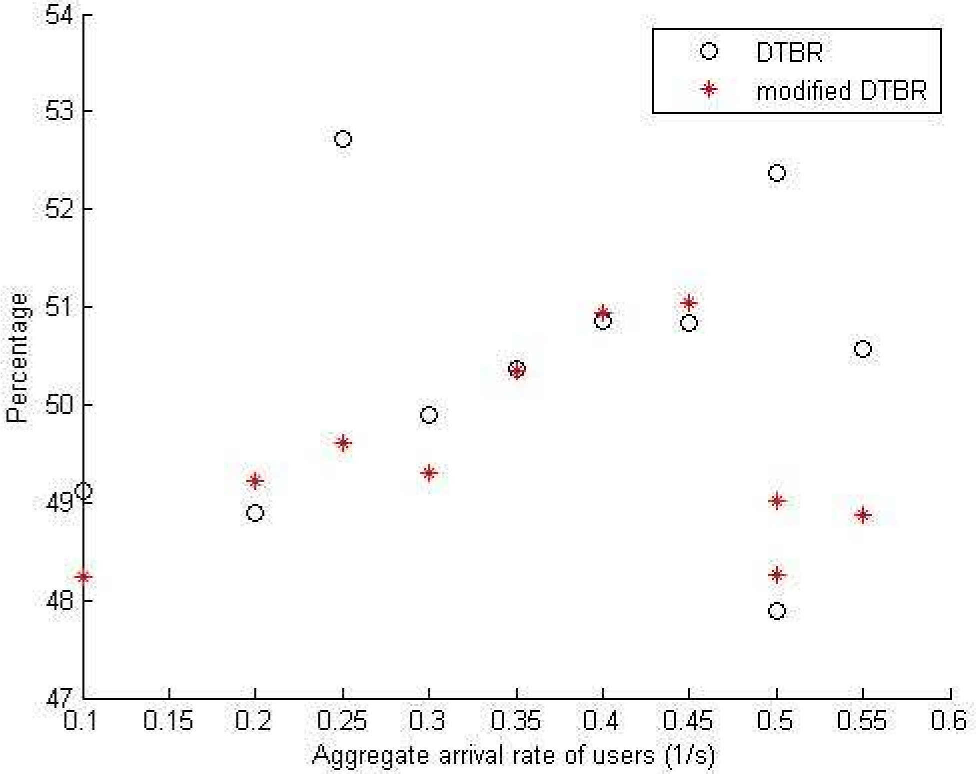

Figure 14. Percentage of active RT users over total number of active users.

Based on the results so far, we conclude that our scheme reduced dropping probability of handoff users in addition to maintaining the achievements of the dual threshold bandwidth reservation (DTBR) scheme in terms of resource utilization, blocking probability, simplicity and fairness.

4 Joint Radio Resource Management: System Architecture

In [10], we proposed a hierarchical joint radio resource management approach that is aimed at defining the location of the entities that perform the different parts of the joint radio resource management. The main goals of the work are simplicity, scalability, and robustness of the JRRM. We will describe it in this chapter with some amendments and a few notional changes.

4.1 A High Level Description

In the hierarchical JRRM, the task of the JRRM protocol is divided into two hierarchies, namely, horizontal radio resource management (HRRM) and vertical radio resource management (VRRM). The HRRM controls intra-domain admission policy in an access network. It is deployed in a decentralized mode. In each access network there will be an HRRM that works for it. The VRRM controls inter-domain admission policy to distribute calls among the different access networks that are in a similar geographical area. Hence, it is centralized in the backbone of the network. The HRRM and VRRM communicate with each other at necessary times. The aim of this approach is to avoid processing unnecessary steps due to a unified JRRM that involves too much signalling. Figure 15 shows how the JRRM components can be located in a heterogeneous wireless network environment.

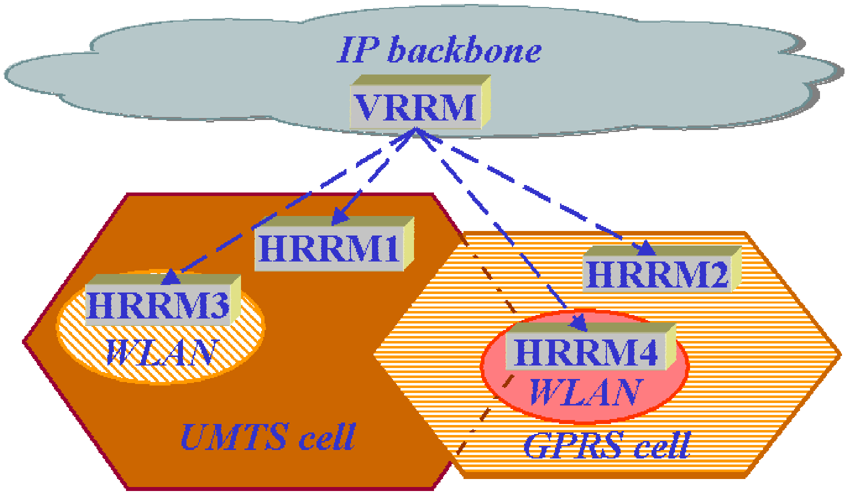

Figure 15. Location of HRRM and VRRM in a heterogeneous wireless network.

4.2 Horizontal Radio Resource Management

Each access network is managed by an HRRM protocol that is assigned to it. The function of the HRRM is similar to the RRM that pre-4G wireless networks employ. However, the proposed HRRM needs to be able to carry out a few additional functions in order to comply with the heterogeneous environment. It replies to a query by the VRRM (to be explained later) whether or not it is possible to admit a new call and what the current status of the access network is. It also periodically reports in its own time to the VRRM about the status of the access network.

The information that the HRRM provides to the VRRM is determined according to the serving discipline employed. The thresholds of the serving discipline scheme are determined based on service providers' preferences as to which category of calls they want to serve more.

4.3 Vertical Radio Resource Management

Three conditions may trigger the VRRM algorithm. We will discuss them in detail in the following subsections.

4.3.1 Call Arrival

When a new call arrives, the VRRM requests traffic information from all HRRMs in access networks that are in the vicinity of the user. From the replies, the VRRM learns which access networks are able to admit the call and what the preference of each network is for the type of the service request (voice/data). It then compares the replies from the HRRMs to find the most suitable network in the current situation. One criterion for suitability of a network is that the less the occupancy level the more suitable it is to admit the call. This will reduce load imbalance among the different access networks. However, the criteria of selecting the most suitable access network vary depending on the diversity of the networks in the heterogeneous environment. We will discuss this further in Chapter 5. If there are multiple access networks equally suitable to admit the call, then the VRRM will randomly choose one of them to admit it. The flow chart in Figure 16 shows how the VRRM operates when a new call arrives.

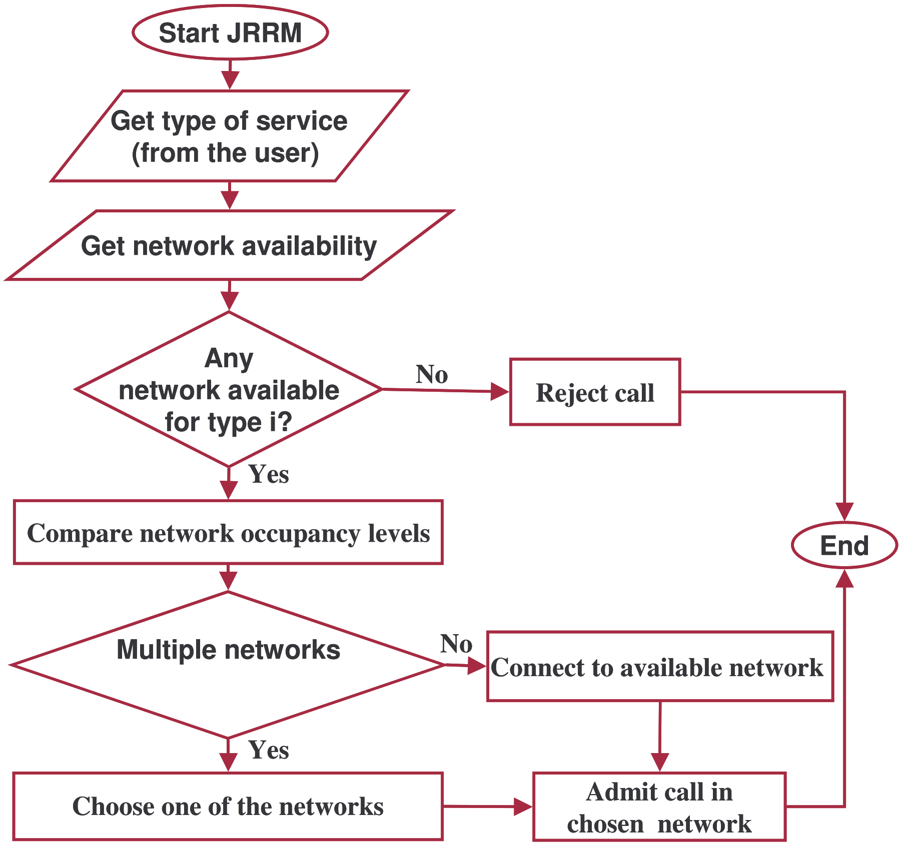

Figure 16. The flow chart of the VRRM.

4.3.2 Necessary Handoff

Not all access networks are necessarily available to a service type at a certain location. Therefore, when the end user roams through the heterogeneous network, there is a chance that the access network to which the mobile terminal (MT) is connected is not available at certain locations. In this case, the HRRM will know that a vertical handoff is necessary and it informs the VRRM about this situation. The VRRM will then choose the best access network to switch the call.

4.3.3 Desirable Handoff

Even though the VRRM considers load balance when it admits new calls, load imbalance occurs due to departures of calls. The VRRM periodically examines the traffic conditions of the different access networks using the report that it receives from individual HRRMs. When the load imbalance approaches a certain threshold, the VRRM will decide to choose some MTs from the heavily loaded networks and switch them to lightly loaded ones.

The period over which the load reports are sent by HRRMs to the VRRM should not be very short so that the consequent signaling would not be high. It should not be very long either. This is because it is desirable that the system should quickly react to changes in the traffic conditions [4].

However, the VRRM should avoid unnecessary vertical handoffs to reduce signalling, which consumes network resources and the power of mobile terminals. Thus, if network load imbalance occurs at an acceptable level, vertical handoffs will not be implemented. The JRRM instead places more emphasis on attempting to create balance during the arrival of new calls.

4.4 Coordinating the Components of the Joint Resource Management Scheme

When a call arrives at a heterogeneous wireless network, the VRRM issues a query to each HRRM asking whether or not the access network can admit the call and what the occupancy level L of the access network is. Here, the capacity of a network is seen as divided into multiple slots, and L is expressed in terms of the slot, which the occupancy level has reached. The reply from the HRRMs will have the form (b, L) where b takes the values 1 (for “yes the call can be admitted”) and 0 (for “no the call cannot be admitted”). The decision by the HRRM on whether or not a call can be admitted to an access network is made based on the serving discipline employed, namely, the modified dual threshold bandwidth reservation scheme. For example, in Figure 17, for a new data call arrival, the replies from the HRRMs in the WLAN, GPRS, and UMTS will be (1, 0), (0, 1), and (0, 1), respectively. Thus, the VRRM will choose the WLAN to admit the call because it is the only access network that can admit it.

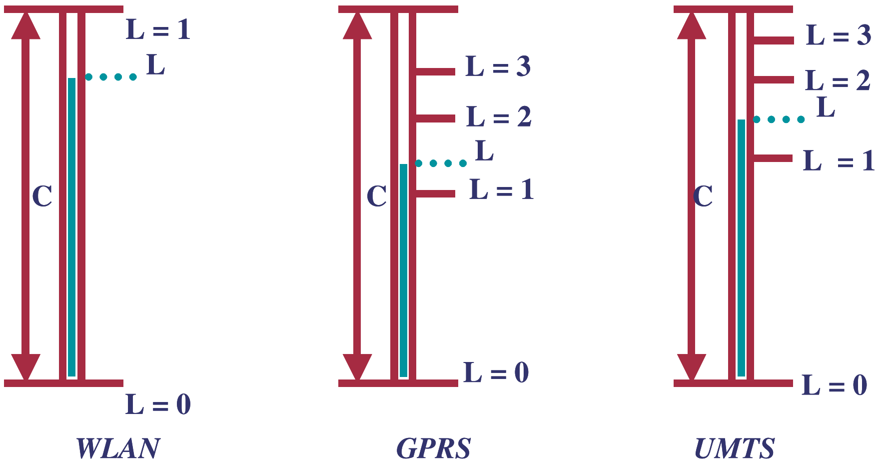

Figure 17. The capacity and thresholds of different networks.

5 Joint Radio Resource Management: Decision Making

In this chapter, we discuss the JRRM from the decision-making point of view, such as the conditions in which admission of a certain type of call arrival is possible or a vertical handoff is triggered.

5.1 Description

In this work we discuss some practical issues that appear in designing the JRRM. The design of our JRRM is based on a combination of many considerations, including occupancy level in individual access networks, nature of overlap of networks, load balancing, and nature of individual access networks. In order to appreciate the significance of those considerations, we will start by designing a simple JRRM and will improve it by adding more and more features one after the other, each time including results and analysis. In each case, we aim at improving system performance in terms of throughput and blocking probabilities. The dropping probability is kept to a desired amount according to the amount of bandwidth reserved for handoff users. Therefore, it does not vary much.

5.2 Simulation Setup

In our simulation, we make an effort to capture the advantages and disadvantages of different approaches in designing the JRRM for a heterogeneous network topology consisting of WLAN and UMTS radio access technologies, as shown in Figure 18.

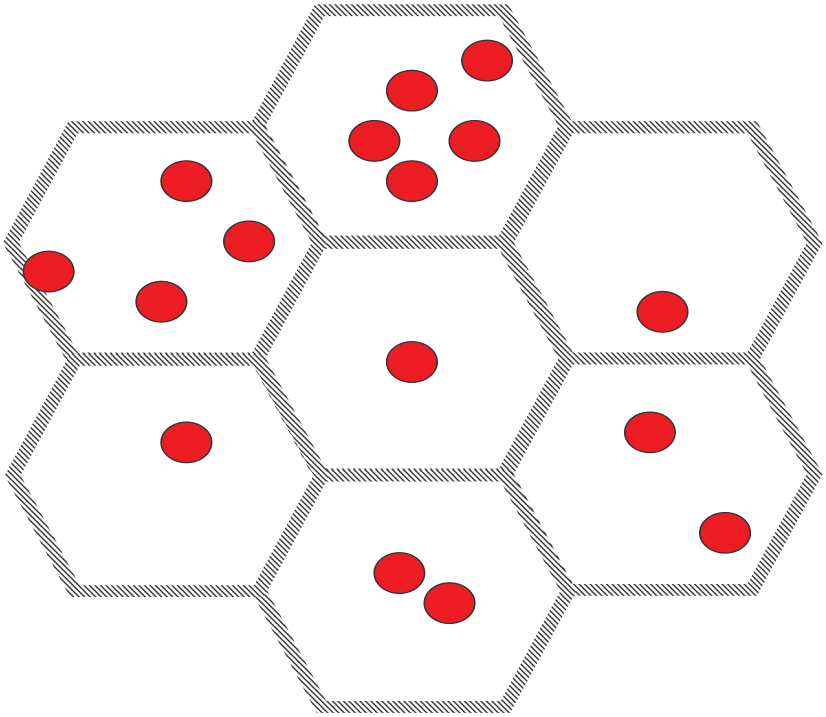

Figure 18. Network topology consisting of UMTS cells and WLAN hotspots.

Our actual simulation focuses on the events that happen in the centre hexagonal cell of the UMTS and the centre circular WLAN hotspot. Unfortunately, the Enhanced UMTS Radio Access Network (EURANE) does not support multiple UMTS cells. Therefore, we created a hexagonal boundary based on calculating the position of mobile terminals. Since our simulation requires only a heterogeneous wireless network topology to implement our radio resource management scheme, we created a simulation topology consisting of a single UMTS cell and a WLAN hotspot as shown in Figure 19. The events that occur in the overlay of the UMTS cell and the WLAN hotspot are sufficient for us to show the effect of handoff calls to/from neighbouring cells of the UMTS cell, in addition to new calls arriving in the overlay network.

The specific technologies simulated are HSUPA category 3 and IEEE 802.11b for the UMTS and the WLAN, respectively.

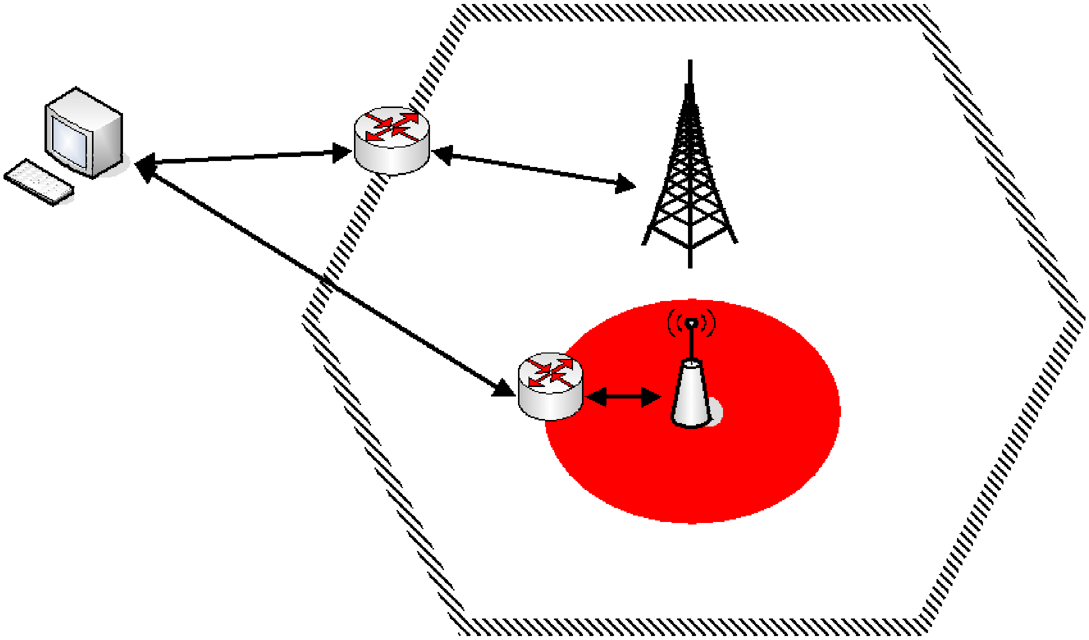

Figure 19. Simulation topology consisting of a UMTS cell and a WLAN hotspot.

Mobile terminals that are located in the WLAN region have a choice to connect either to the WLAN or to the UMTS. If a WLAN user arrives in the handoff region of the WLAN, the RRM will decide whether or not to connect it to the UMTS with a handoff priority. Here, the user will be connected to the UMTS cell only if there is enough resource available for it. Otherwise, it will wait for the release of resource hoping that this happens before it leaves the WLAN hotspot. If the user unfortunately leaves the WLAN area before resource is released, then it will be dropped. If a user connected to the UMTS arrives in the UMTS handoff region, it will be deleted by the simulation (imagining that it is being taken care of by a neighbouring cell) as the detailed event is out of the boundary of the statistics that we intend to collect.

Upon a request to start a network session, a user who can geographically access a WLAN hotspot and a UMTS cell will be connected either to a WLAN hotspot or to a UMTS cell or will be rejected based on our radio resource management scheme. If the accessible network is only the UMTS network, the user will be accepted by the UMTS network or will be rejected.

Users arrive in the WLAN hotspot more frequently than in the other areas that the UMTS cell covers. We assume this because the places where WLAN hotspots are available are in general densely populated by connection users (eg. a university campus or a residential area) and UMTS cells cover places of both high and low population densities. In our simplified model, the ratio of users that arrive in the WLAN area to the rest of the UMTS cell coverage is 3, which is roughly equal to the ratio of the bandwidth capacity of the WLAN to that of the UMTS. This value also applies to the ratio of number of WLAN subscribers to that of UMTS subscribers in the case where we simulate a non-integrated system. For a given mean arrival rate, inter-arrival time between two consecutive arrivals is exponentially distributed.

Each data point in the results was obtained by averaging over a simulation time of 2 hours. This was done by running a 2 hours and 10 minutes simulation time and discarding the first 10 minutes, to produce more stable result. Table 2 summarizes relevant parameters considered in the simulation.

Table 2. Simulation parameters.

Simulation Parameter	**Value**
Capacity of the HSUPA	1.46 Mbps
Physical layer data rate of IEEE 802.11	11Mbps
HSUPA cell side length	1Km
IEEE 802.11 hotspot radius	100m
Ratio of RT users to NRT users	1 : 1
Percentage of handoff calls to all calls (This applies both to RT and NRT)	0.25%
Average data rate requested by real-time users	100000 bps

Average data rate requested by non-real-time users	50000 bps
Minimum acceptable data rate by non-real-time users	50000 bps
Average call duration	120s

5.3 Designs, Results and Analysis

5.3.1 Simple JRRM

Our simple JRRM can be described as follows. When a call arrives in the heterogeneous environment, if it is in the coverage area of a single network, then a check is performed as to whether or not there is resource available to accept it in the network. As we discussed in Chapter 4, this is done by HRRMs in the access network. If resource is available, then it will be admitted in the access network. Otherwise, it will be rejected. If it is in the coverage area of multiple access networks, then a check is performed for resource availability in each access network. If there is no resource in any of the access networks to accept it, then it will be rejected. If there is only one network with enough resource for the call, then it will be accepted in this network. If multiple networks have enough resource to accept it, then it will be accepted randomly to one of them, with all of those networks having equal probability of accepting it. The flow chart in Figure 20 shows decision making during arrival of a call, whether new or handoff.

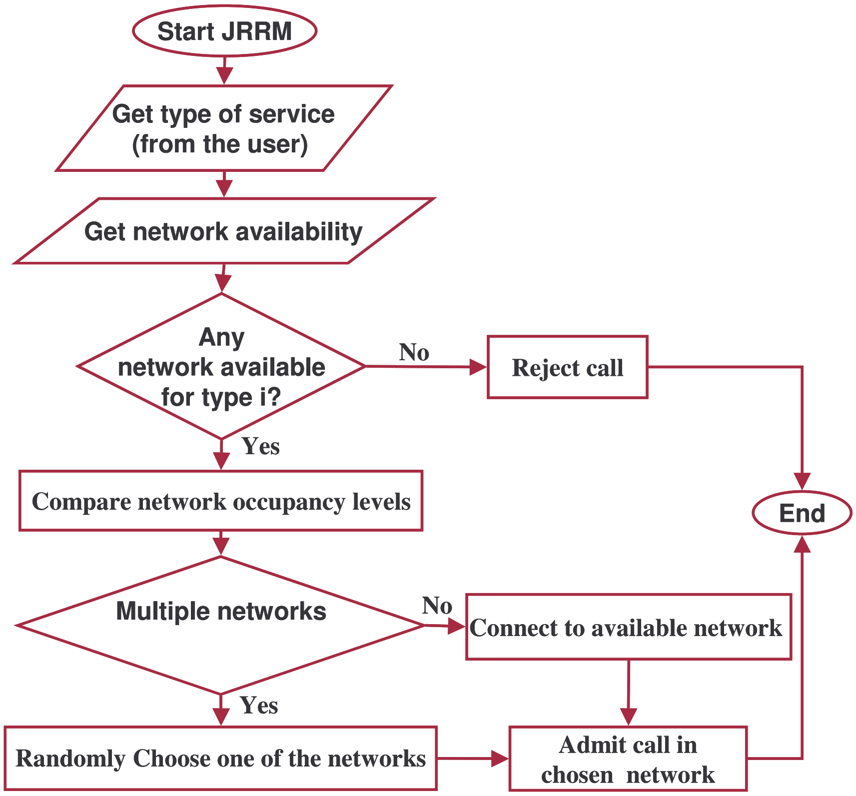

Figure 20. The flow chart of the simple JRRM.

Imagine implementing the modified dual threshold bandwidth reservation scheme discussed in Chapter 3 in the heterogeneous wireless network with coverage areas depicted in Figure 21. For example, for a new data call arrival in the overlay area of the two networks, assuming both of the access networks are in the occupancy level in which they can accept the call, the simple JRRM implementation would decide to randomly admit the call to either of the two access networks.

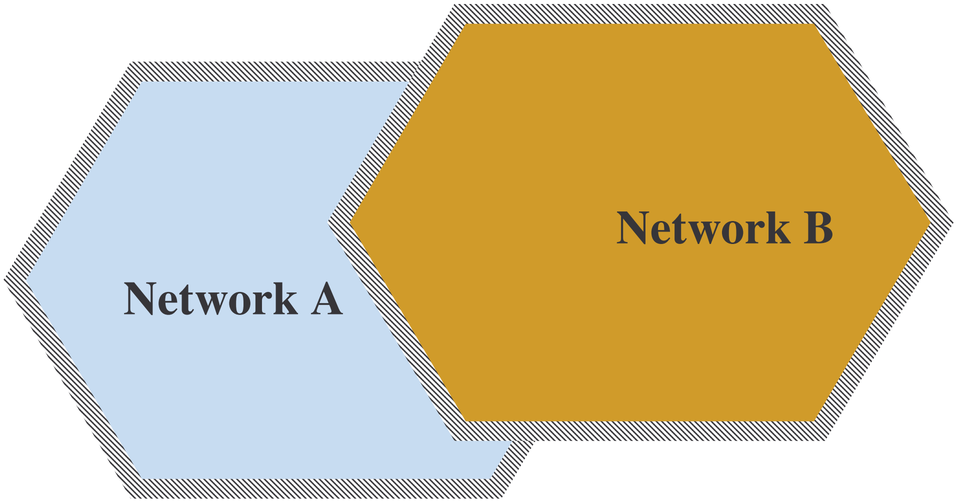

Figure 21. The overlap of two access networks.

While this approach seems to create at least a wider opportunity for a new call than in the case of non-integrated networks, it is not always true. For instance, assume network A in Figure 21 has a capacity double the data rate of network B. And for simplicity assume that in both the networks, the reserve for handoff users is 0 Mb, and users arrive randomly in any position in the coverage area of either or both of the access networks. Then, implementing the simple JRRM in the integrated system would fill network B at a traffic rate much lower than a situation in which the system is not integrated and the individual access networks have separate subscribers proportional to their capacity. That means, with an aggregate user arrival rate that is much lower than the system capacity, calls have a higher blocking probability in the case of the simple JRRM than in that of the non-integrated system.

In Figures 22 and 23, the lower throughput and the higher blocking probability of the simple JRRM as compared with the non-integrated system for low aggregate user arrival rate show this concept. As the difference is not too much (specifically to our simulation setup), this is not very clearly revealed in the case of throughput comparison. But we see a clearer difference when we compare the blocking probabilities.

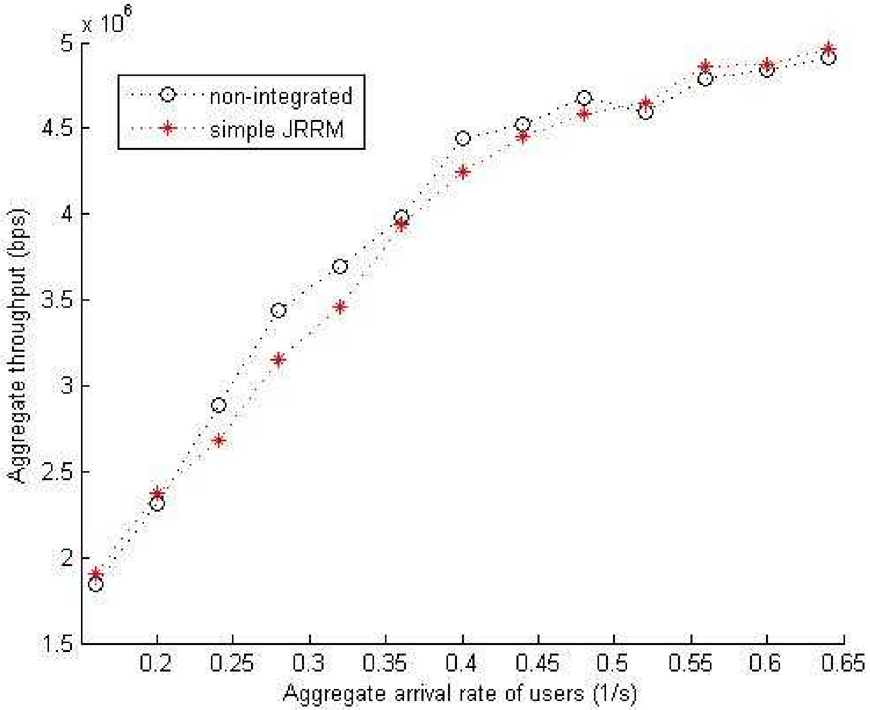

Figure 22. Throughput comparison.

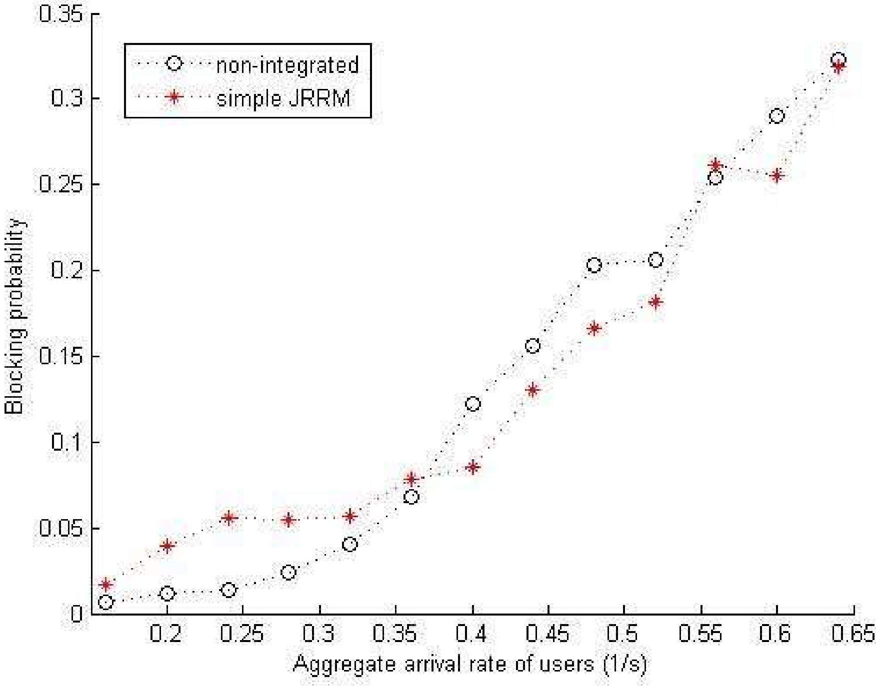

Figure 23. Blocking probability comparison.

5.3.2 Considering Widening Opportunities for Future Desperate Users

As discussed in the previous section, when there is a choice, random selection of networks in general causes high imbalance in cases where the capacities of the different access networks are different. Therefore, we believe that the choice of networks should consider widening opportunities for future users.

Figure 24 depicts three different cases of a wireless overlay. In the case of Figure 24(a), the two access networks are overlaid with perfect overlap. Here, we can choose to apply simple JRRM, because as long as there is enough resource in one of the networks, an incoming call will be admitted.

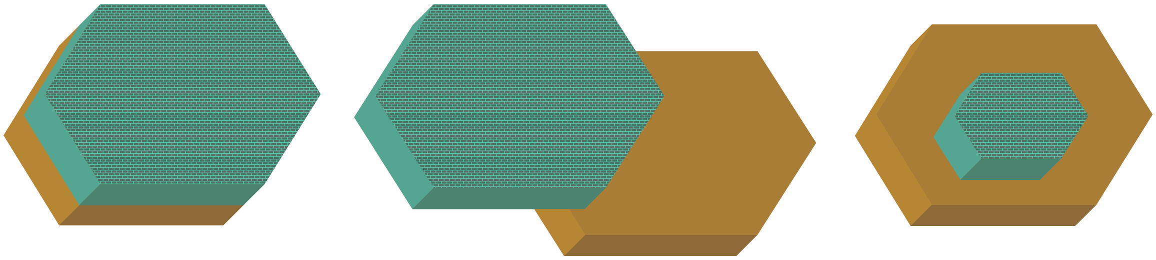

Figure 24. The overlap of two networks in different fashions.

(a) Perfect overlap (b) Partial overlap

(c) Complete overlap of a smaller network in a larger one.

However, an imperfect overlap, as in the case of Figure 24(b) and (c), requires the widening of opportunities for incoming calls in non-overlapping areas. For example, in the case of Figure 24(b), whenever there is a choice of network, we suggest that the user should be admitted to Network A or B with a probability proportional to the capacity of the two access networks. However, in the case of Figure 24(c), whenever there is a choice, we admit the call into the smaller network. The overlay type in Figure 24(c) is similar to our simulation scenario and hence we applied a decision of this nature.

While the throughput comparison in Figure 25 shows only very little difference, the blocking probability comparison in Figure 26 proves the above concept. The differences do not appear in very low average arrival rates because during such cases, both schemes admit almost all users. Also, at very high arrival rates, such differences will not appear because users are blocked mainly because resource is unavailable consistently, no matter which scheme we implement.

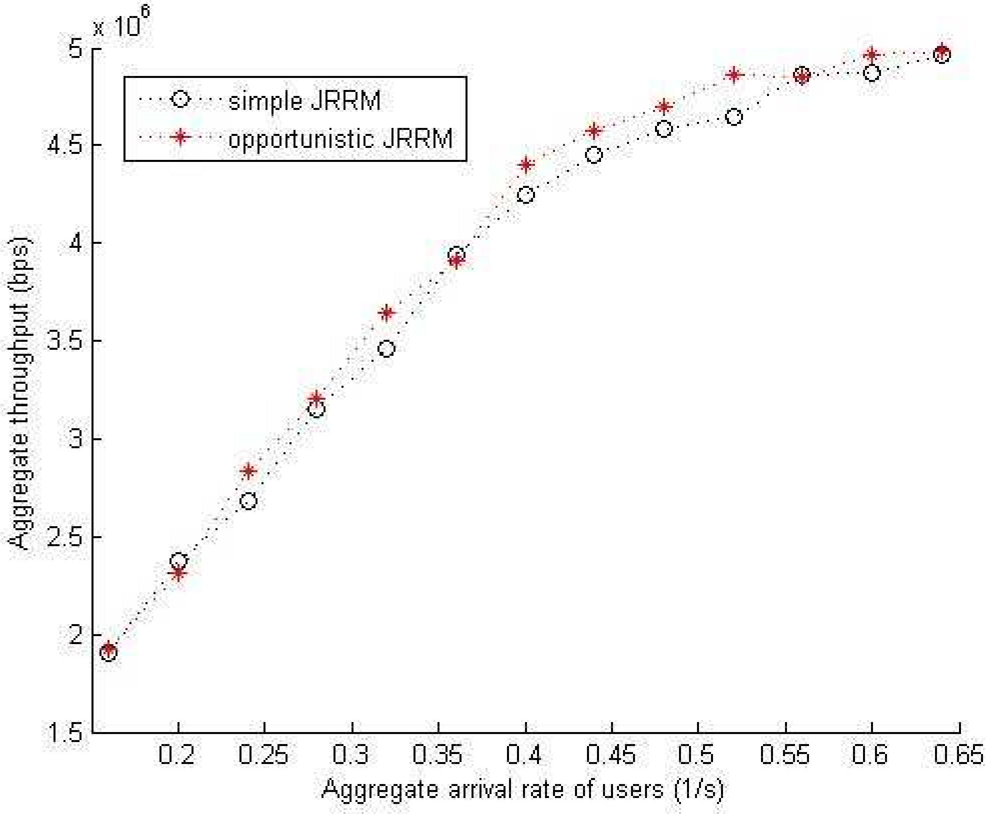

Figure 25. Throughput comparison.

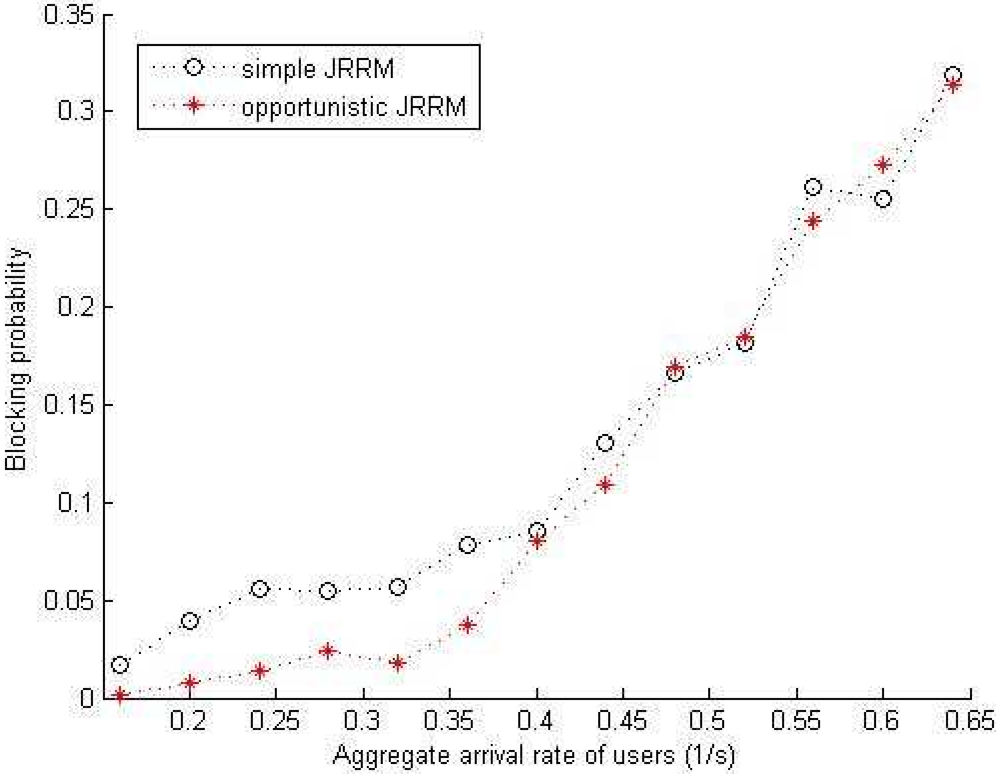

Figure 26. Blocking probability comparison.

5.3.3 Load Balancing

The JRRM in the previous sections does not include a load balancing scheme. Load balancing is an attempt to maximize the availability of resources to future users that arrive anywhere in the heterogeneous environment by switching ongoing calls from one access network to another. Load imbalance among access networks in a heterogeneous environment occurs for many reasons despite attempts that are made by the JRRM to create balance whenever a call arrives. For example, the location of an incoming call might not allow for creating balance, often because there is imperfect overlap and not all of the access networks are accessible to the call. Also, departure of calls often results in imbalance.

Therefore, there should be an explicit task of the JRRM dedicated to load balancing, so that at stable traffic conditions there will be higher probability of an incoming call to find enough space. However, this might not always be possible as the policy of individual access networks might vary, in addition to the fact that not all wireless access networks are geographically accessible to all active users.

Attempts are often made to distribute load and service mix among all wireless access networks. However, this is not always the most effective thing to do for an efficient overall system. For example, in the three different scenarios of overlap in Figure 24, different approaches of load balancing should be implemented for greater efficiency of the system.

In Figure 24(a), the load balancing task will not have an influence in reducing call blocking probability. In Figure 24(b), we believe that attempts should be made to maintain the load in each access network proportional to its capacity.

In Figure 24(c), however, attempts should be made to switch as many active users as possible from the larger network to the smaller one. Intuitively, this is because it leads to maximizing efficiency by making available as much capacity as possible to users that arrive in the area only covered by the wider access network area. As the simulation topology has an overlap similar to the one in Figure 24(c), we implemented such a load balancing approach.

Figures 27 and 28 show the performance comparisons between a design with and without adding the load balancing feature in addition to all the other features described so far.

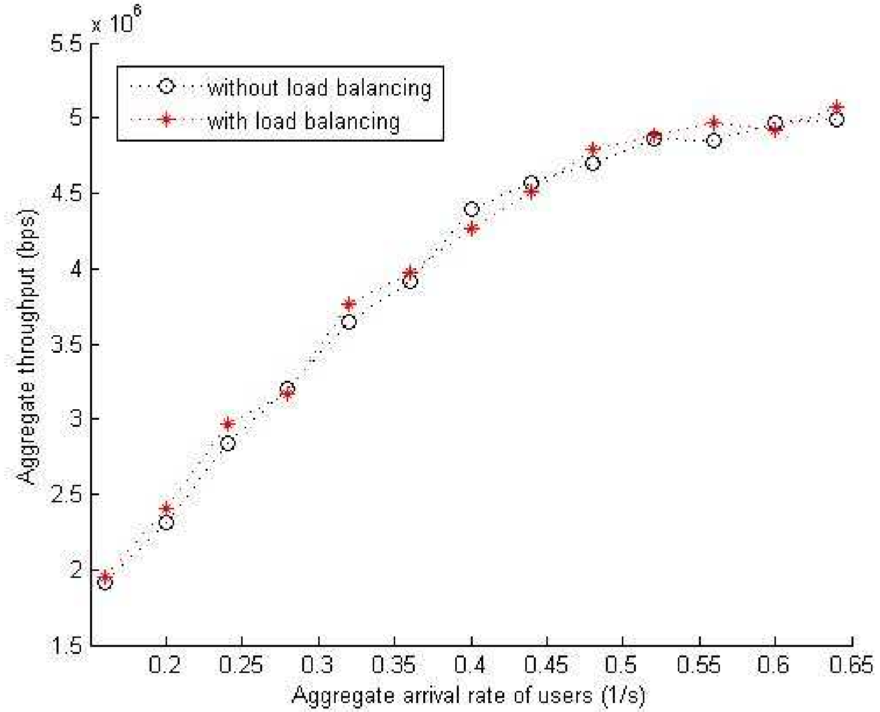

Figure 27. Throughput comparison.

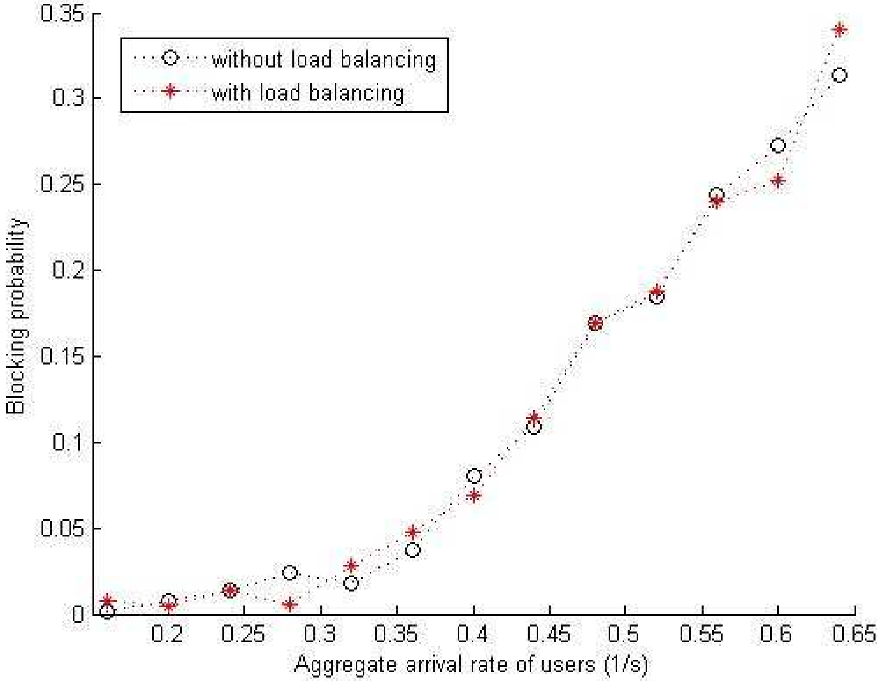

Figure 28. Blocking probability comparison.

Our argument of load balancing, however, is not revealed from Figures 27 and 28, as we see neither higher throughput nor lower blocking probability. We believe that this is not because our argument is invalid; it is because the two access technologies react differently to the traffic load.

The UMTS network uses dedicated uplink channels, namely E-DCH, for each user to send packets. Therefore, from the MAC protocol perspective, the capacity of the UMTS network is constant. However, the MAC protocol of the IEEE802.11 networks performs differently according to the number of users, and hence the capacity varies significantly as shown in Figure 29, which is a repetition of a plot found in [12] and [13]. The achievable throughput of the IEEE802.11 network is almost a monotonically decreasing function of the number of active client nodes. This is because, as the number of users increase, the throughput performance of career sense multiple access (CSMA) implemented at the MAC layer of the 802.11 networks reduces due to increased packet backoff times.

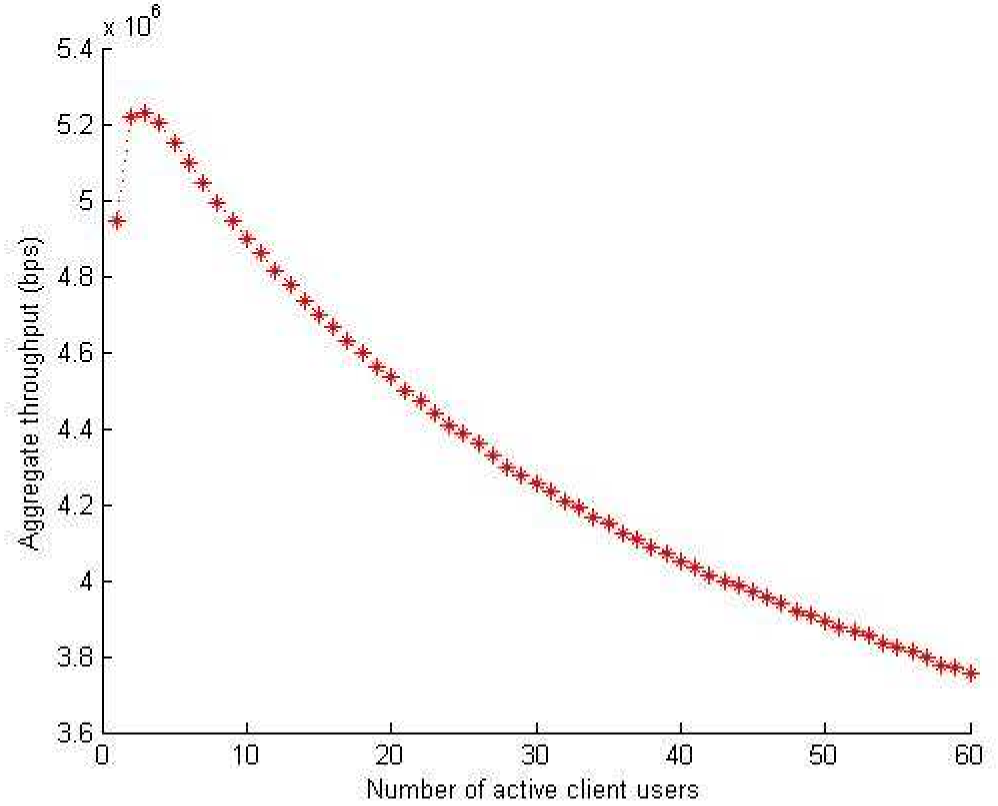

Figure 29. Achievable throughput of IEEE802.11 as a function of number of client users.

5.3.4 Considering Resource Variation in Dynamic Resource Allocation

In order to account for the reaction of the IEEE 802.11 network, we suggest adding the following feature. For a mobile terminal located in the overlay area, if the data rate demand of the application is low, it should be connected to the UMTS; otherwise, it should be connected to the WLAN. For a certain number of active users, such a decision decreases the average number of users in the WLAN and increases that of the UMTS. This in a sense increases the capacity of the WLAN, while the UMTS remains unaffected. Note that, as we mentioned before, the ns-2 does not have physical layer features like interference. Therefore, our comparison is only based on the interactions between protocols at higher layers.

In the simulation, we took data rate demands less than or equal to 30kbps as "low", and anything higher as "high" to consider the variation of the capacity of the IEEE802.11 network in the design of the JRRM. The results are shown in Figures 30 and 31.

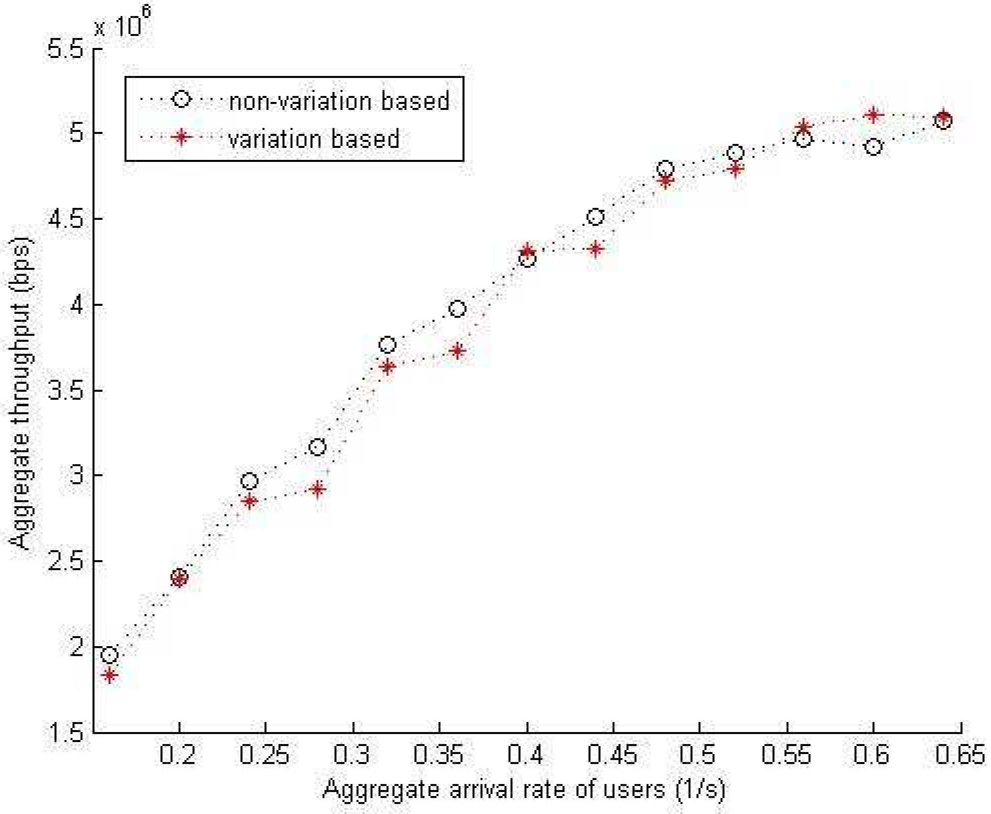

Figure 30. Throughput comparison.

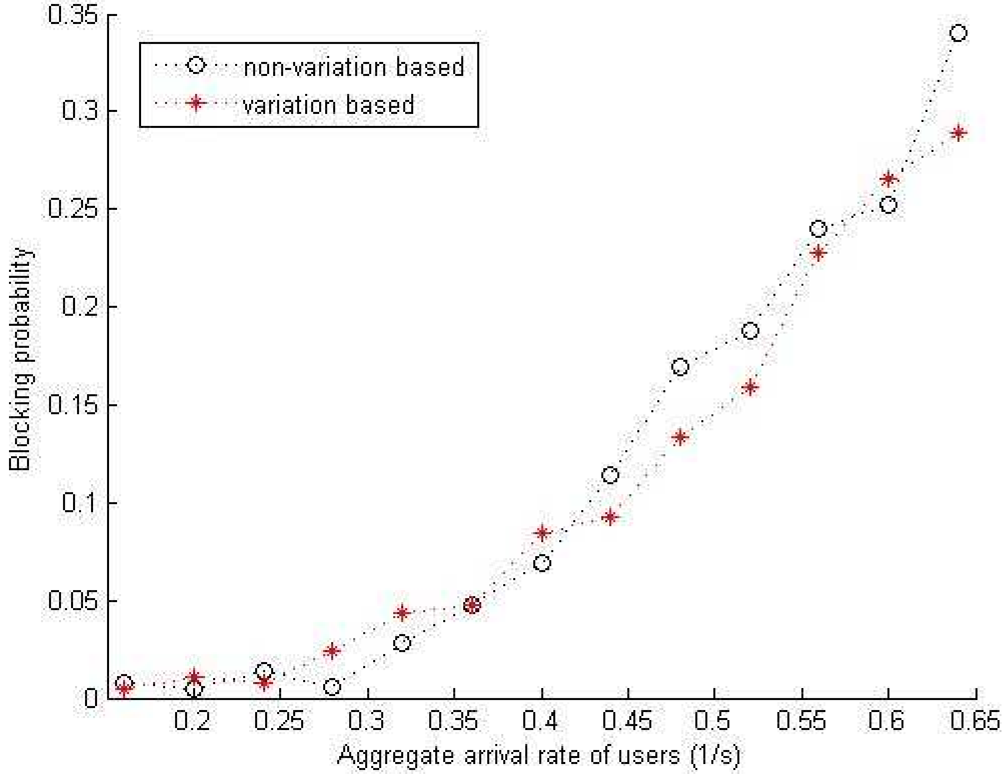

Figure 31. Blocking probability comparison.

Here, while the throughput performance does not show a clear trend of the strength or weakness of the new feature, the blocking probability shows a more prominent improvement.

So far, we were adding one feature on top of the other. As the contribution of each individual feature is small, the result is also affected by the randomness of the scheduled events, so it has often been difficult to conclude something from the visual comparisons. However, a much clearer improvement is revealed when we compare the simple JRRM with the final JRRM, which is a cumulative result of multiple considerations.

As we see from Figure 32, the throughput improvement is consistent at high user arrival rates. From Figure 33, we notice that the blocking probability shows a trend that enables us to conclude an improvement in both low and high user arrival rates.

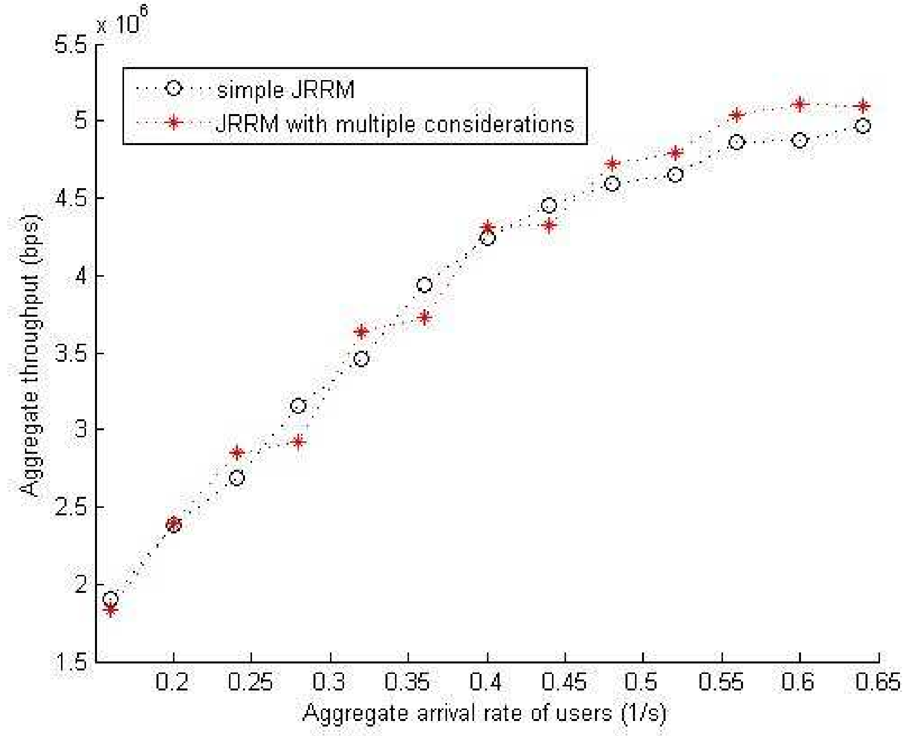

Figure 32. Throughput comparison.

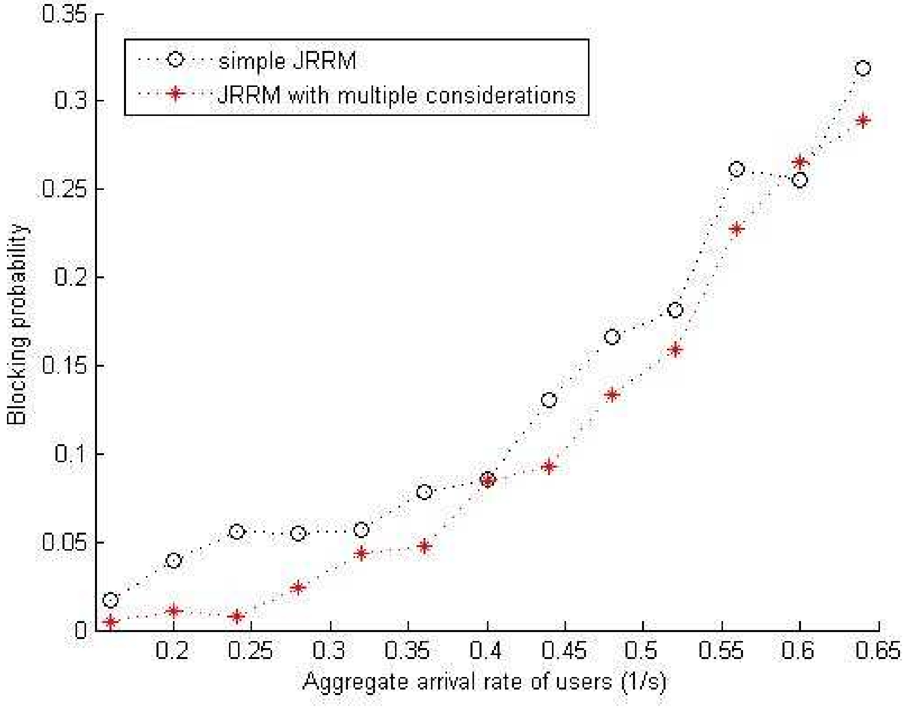

Figure 33. Blocking probability comparison.

As we mentioned before, our main target is to improve system performance in terms of throughput and blocking probability. The dropping probability is fixed according to the amount of bandwidth that is reserved for handoff users. For this reason and also because handoff dropping events throughout the simulation time are relatively very few, we do not expect noticeable difference between the dropping probability performance of the different approaches discussed so far as shown in Figure 34.

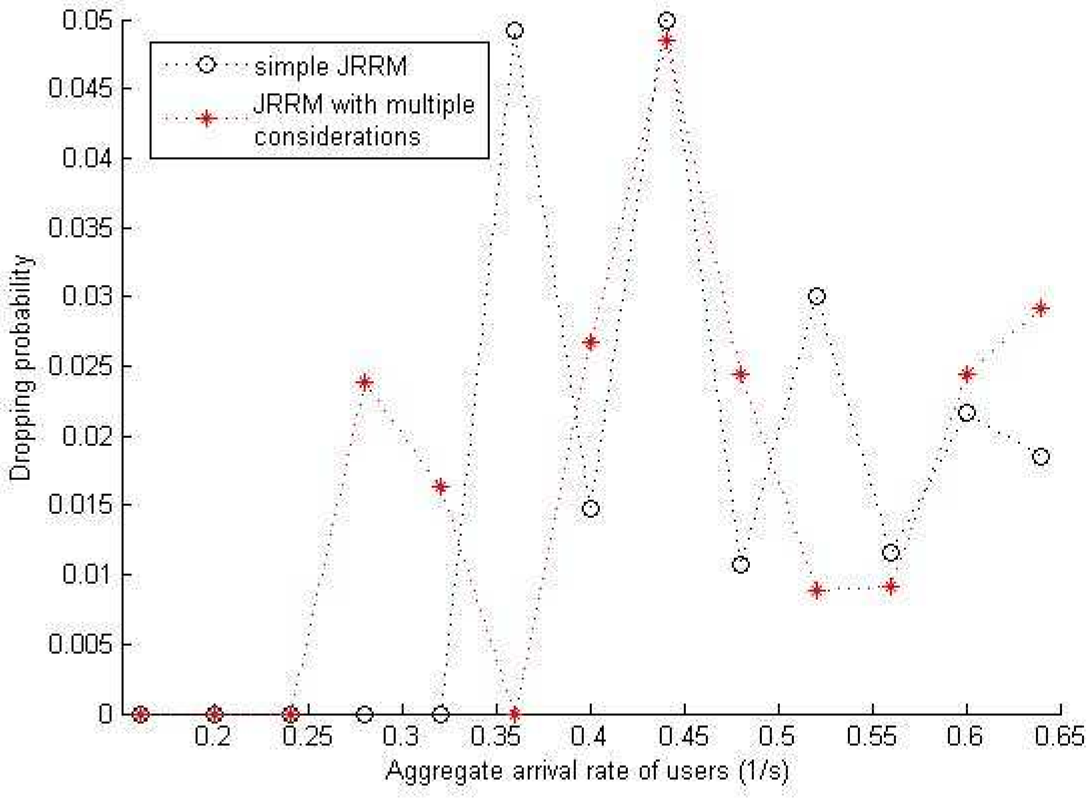

Figure 34. Dropping probability comparison.

6 Conclusions and Future Work

Joint radio resource management (JRRM) in heterogeneous wireless networks is a functionality that performs much more complex tasks than its homogeneous counterpart. As the scenarios of the next generation wireless networks are mainly studied in a modeling environment, much effort is needed to come up with a JRRM protocol that takes so many considerations of significance in the real heterogeneous scenario.

The achievable throughput is highly variable in some radio access networks. This variability depends on many factors determined by the radio access technology, its configuration, environmental conditions, and user properties. Therefore, for the same mix of users and radio access technologies, connecting certain users to certain radio access technologies can contribute to optimizing availability of resources.

Future potential work on the JRRM includes devising a mathematical model to describe events in the heterogeneous wireless scenario. The model should be able to capture the impact of the diversity of access networks and user types, the type of overlap of different access networks, the user mobility pattern, etc on the performance of the JRRM. Another promising work is finding a solution to simulate the impact of the physical layer protocols on the performance of the JRRM.

References

[1] 3GPP TS 25.331 v1.1.0, "RRC Protocol Specification".

[2] C. T. Chou and K. G. Shin, "Analysis of combined adaptive bandwidth allocation and admission control in wireless networks," Proc. IEEE INFOCOM, pp. 676-684, June 2002.

[3] O. E. Falowo and H. A. Chan, "Joint call admission control for integrated UMTS-WLAN network," Proc. SATNAC, pp. 83-88, Sept. 2005.

[4] M. Ghaderi, R. Boutaba, and G. W. Kenward, "Stochastic admission control for quality of service in wireless packet networks," in Proc. 4th IFIP Networking Conference, pp. 1309-1320, Waterloo, Canada, May 2005.

[5] Y. -R. Huang, Y. -B. Lin, and J. -M. Ho, "Performance analysis for voice/data integration on a finite-buffer mobile system," IEEE Trans. Veh. Technol., vol. 49, pp. 367-378, Mar. 2000.

[6] W. Kellerer and H. Vogel, "A communication gateway for infrastructure-independent 4G wireless access," IEEE Comm. Mag., pp. 126-131, Mar. 2002.

[7] A. K. Salkintzis, C. Fors, and R. Pazhyannur, "WLAN-GPRS Integration for Next-Generation Mobile Data Networks," IEEE Wireless Communications, October 2002.

[8] A. Sanmateu, "Seamless mobility across IP networks using mobile IP," Computer Networks, 2002.

[9] P. Vidales, G. Mapp, F. Stajano, J. Crowcroft, and C. J. Bernardos, "A practical approach for 4G systems: deployment of overlay networks," TRIDENTCOM, Trento, Italy, pp. 172-181, Feb. 2005.

[10] K. H. Suleiman, H. A. Chan, and Mqhele E. Dlodlo, "A hierarchical approach to joint call admission control in heterogeneous wireless networks," IST-Africa, South Africa, May 2006.

[11] K. H. Suleiman, H. A. Chan, and Mqhele E. Dlodlo, "Load Balancing in the Call Admission Control of Heterogeneous Wireless Networks," IWCMC, Vancouver, Canada, pp. 245 – 250, July 2006.

[12] R. Vijayakumar, T. Javidi, and M. Liu, "From saturation to nonsaturation: A study on 802.11 networks," Dept. of Electrical Eng., University of Michigan, Tech. Rep. CSPL-363, 2005.

[13] K. H. Suleiman, T. Javidi, M. Liu, and S. Kittipiyakul, "The impact of MAC buffer size on the throughput performance of IEEE 802.11," submitted to IEEE Transactions on Mobile Computing, Jan. 2008.

[14] Y. C. Lai, and S. F. Tsai, "A fair admission control for large bandwidth multimedia applications," Proc. GLOBECOM, Sam Antonio, TX, pp 317-322, Nov. 2002.

[15] L. -Z. Li, B. Li, B. Li, and X. –R. Cao, "Performance analysis of bandwidth allocations for multi-service mobile wireless cellular networks," Proc. IEEE WCNC, pp. 1072-1077, Mar. 2003.

[16] L. Yin, B. Li, Z. Zhang, and Y. –B. Lin, "Performance analysis of a dual threshold reservation (DTR) scheme for voice/data integrated mobile wireless networks," Proc. IEEE WCNC, pp. 258-2262, Sept. 2000.

[17] B. Li, L. Li, B. Li, K. M. Sivalingam, and X. –R. Cao, "Call admission control for voice/data integrated cellular networks: performance analysis and comparative study," IEEE Journal on Selected Areas in Communications, vol. 22, pp. 706-718, May 2004.

[18] K. Muray, R. Mathur, and D. Pesch, “Adaptive policy based access management in heterogamous wireless networks,” Proc IEEE WPMC, Vol. 1, pp. 325-329, Oct. 2003.

[19] C. Perkins, “IP mobility support for IPV4,” IETF, RFC 3344, Aug. 2002.

[20] X. G. Wang, G. Min, J. E. Mellor, K. Al-Begain, and L. Guan, “An adaptive QoS framework for integrated cellular and WLAN networks,” Elsevier Journal on Computer Networks, pp. 167-183, 2005.

[21] X. G. Wang, G. Min, J. Mellor, and K. Al-Begain, “A QoS based bandwidth management scheme in heterogeneous wireless networks,” International Journal on Simulation: Systems, Science, and Technology, pp. 9-17, June 2004.

[22] F. Yu, V. Wong, and V. Leung, “A new QoS provisioning method for adaptive multimedia in cellular wireless networks,” Proc IEEE INFOCOM, pp. 2131-2142, Mar. 2004.

[23] N. Vulic, S. Heemstra de Groot, and I. Niemegeers, “Common radio resource management for WLAN-UMTS integration at radio access level,” Proc. IST Mobile and Wireless Summit, Dresden, Germany, June 2005.

[24] J. Perez-Romero, O. Sallent, R. Agusti, P. Karlsson, A. Barbaresi, L. Wang, F. Casadevall, M. Dohler, H. Gonzalez, and F. Cabral-Pinto, “Common radio resource management: functional models and implementation requirements,” Proc. IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), vol. 3, pp. 2067-2071, Sept. 2005.

[25] A. –E. M. A. Taha, “A framework for radio resource management in heterogeneous wireless networks,” PhD thesis, Queen’s University, Ontario, Canada, Sept. 2007.

[26] A. Koukab, Y. Lei, and M. J. Declercq, “A GSM-GPRS/UMTS FDDTDD/WLAN 802.11a-b-g multi-standard carrier generation system,” IEEE Journal of Solid-State Circuits, vol. 41, pp. 1513-1521, July 2006.

[27] R. B. Ali and S. Pierre, “An efficient predictive admission control policy for heterogeneous wireless bandwidth allocation in next generation mobile networks,” IWCMC, Vancouver, Canada, pp. 635-640, July 2006.

[28] S. S. Rappaport, H. Lon-Rong, “Microcellular communication systems with hierarchical macro cell overlays: traffic performance models and analysis,” Proc. IEEE, vol. 82, pp. 1383-1397, 1994.

[29] C. Chang, C. J. Chang, and K. -R. Lo, “Analysis of a hierarchical cellular system with reneging and dropping for waiting new and handoff calls,” IEEE Transactions on Vehicular Technology, vol. 4, no. 48, pp. 1080-1091, 1999.

[30] Y. H. Hwang and S-K. Noh, “A call admission control scheme for heterogeneous service considering fairness in wireless networks,” Proc. Fourth Annual ACIS International Conference on Computer and Information Science (ICIS’05), Washington, DC, USA, pp. 688-692, 2005.

[31] L. J. Wei, “The adaptive biased coin design for sequential experiments,” Journal of Annals of Statistics, vol. 6, pp. 92-100, Jan. 1978.

[32] W. Song, H. Jiang, W. Zhuang, and X. Shen, “Resource management for QoS support in cellular/WLAN interworking,” IEEE Network, vol. 19, pp. 12-18, Sept./Oct. 2005.

[33] W. Song, W. Zhuang, and Y. Cheng, "Load balancing for cellular/WLAN integrated networks," IEEE Network, vol. 21, pp. 27-33, Jan./Feb. 2007.

[34] L. Giupponi, R. Agustı´, J. Pe´rez-Romero, and O. Sallent, “A framework for JRRM with resource reservation and multiservice provisioning in heterogeneous networks,” Mobile Networks and Applications (MONET), vol. 11, pp. 825-846, Dec. 2006.

[35] X. Liu, V. O. K. Li, and P. Zhang, “Joint radio resource management through vertical handoffs in 4g networks,” Proc. IEEE Global Telecommunications Conference (GLOBECOM), San Francisco, USA, pp. 1-5, Nov. 2006.

[36] H. J. Wang, “Policy-Enabled Handoffs Across Heterogeneous Wireless Networks,” Proceedings of the Second IEEE Workshop on Mobile Computer Systems and Applications (WMCSA), New Orleans, USA, Feb. 1999.

[37] O. Sallent, J. Pérez-Romero, R. Agustí, L. Giupponi, C. Kloeck, I. Martoyo, S. Klett, and J. Luo, “Resource auctioning mechanisms in heterogeneous wireless access networks,” IEEE 63rd Vehicular Technology Conference (VTC), Melbourne, Australia, vol. 1, pp. 52-56, May 2006.

[38] O. E. Falowo and H. A. Chan “Adaptive bandwidth management and joint call admission control to enhance system utilization and QoS in heterogeneous wireless networks,” EURASIP Journal on Wireless Communications and Networking, 2007.

[39] A. Pillekeit, F. Derakhshan, E. Jugl, and A. Mitschele-Thiel, “Force-based load balancing in co-located UMTS/GSM networks,” Proc. The 60th IEEE Vehicular Technology Conference (VTC), Los Angeles, Calif, USA, vol. 6, pp. 4402-4406, Sept. 2004.

[40] J. P´erez-Romero, O. Sallent, and R. Agust´ı, “Policy-based initial RAT selection algorithms in heterogeneous networks,” Proc. The 7th IFIP International Conference on Mobile and Wireless Communications Networks (MWCN), Marrakesh, Morroco, Sept. 2005.

[41] E. Stevens-Navarro, A. H. M. Rad, and V. W. S. Wong, “On optimal admission control for multi-service cellular/WLAN interworking,” Proc. Global Telecommunications Conference (GLOBECOM), Washington, DC, pp. 5042-5047, Nov. 2007.

[42] I. M. Suliman and C. Pomalaza-R´aez, “Radio resource allocation in heterogeneous wireless networks using cooperative games,” Proc. The Nordic Radio Symposium, Oulu, Finland, Aug. 2004.

Appendix

The Accompanying CD-ROM

The following files are on the accompanying CD-ROM that includes the thesis document and related materials:

1. The thesis document in PDF format.
2. The installation file for the “mobility package for NS-2”.
3. A “read me” text file which is an instruction on how to run the ns-2 simulation and analysis codes.
4. Simulation code written in TCL programming language.
5. Analysis code written in AWK programming language to extract data from the output of the simulation.