QoS-Aware Call Admission Control
in
Wideband CDMA Wireless Networks

BY

ALEXANDER D. OLIVER

A thesis submitted to the School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
July, 2003

Your file Votre référence
ISBN: 0-612-86161-9
Our file Notre référence
ISBN: 0-612-86161-9

Abstract

This thesis proposes a novel QoS-aware Call Admission Control (CAC) scheme for radio access in wideband wireless UMTS networks. It features an efficient CAC algorithm coupled with a QoS class separation mechanism based on the transmitted power of each individual mobile terminal.

Three inter-relating components have been introduced to extend a currently existing UMTS uplink admission control scheme. First, a measurement-based module monitors the load of the system during a moving time interval. This gives a better estimation of the actual load and removes the existing model's limitation of estimating the current load by taking into consideration only the instantaneous load of the system.

Second, a power-prediction module simulates the state of the system *k* steps ahead and partly bases the admission decision on what the future state of the system would look like if this particular call were actually accepted.

Third, the above two modules have been integrated in a call admission control algorithm with a power sharing structure. This structure aims at achieving differential treatment between the classes. Thus under light load conditions, traffic from all QoS types is accepted. In case there are sufficient resources, the classes are allowed to borrow resources from each other. Under high load conditions, however, the per-class resource allocations are strictly observed. This is integrated with the *k*-step power prediction component in that if it is discovered that in *k* steps ahead any of the classes will need to exceed their share (thus violating the QoS guarantees), then they will be denied admission.

The interworking of the three components is especially useful in cellular systems, which are characterized by rapidly changing medium conditions and hence have varying resources.

The modules have been implemented in a 3G system-level simulator and have been seamlessly integrated in a realistic UMTS scenario to demonstrate improvement in both functionality and efficiency.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Hossam Hassanein for his valuable advice, guidance and support.

Gratefully acknowledged is the financial assistance provided by Queen's University, as well as the Canadian Institute for Telecommunications Research.

I also extend a warm thanks to my family for their love and encouragement. Thank you for believing!

List of Acronyms

2G	2nd Generation Wireless Communication Systems (GSM)
2.5G	2.5 Generation Wireless Communication Systems (GPRS)
3G	3rd Generation Wireless Communication Systems (UMTS)
4G	4th Generation Wireless Communication Systems
3GPP	3rd Generation Partnership Project
bps	Bits per Second
CAC	Call Admission Control
CDMA	Code Division Multiple Access
CN	Core Network
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
GGSN	Gateway GPRS Support Node
GMSC	Gateway Mobile Services Switching Center
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
HLR	Home Location Register
IMT-2000	International Mobile Telecommunications 2000
ISDN	Integrated Services Digital Network
MBAC	Measurement-Based Admission Control
Mcps	Megachips per second
ME	Mobile Equipment
MSC/VLR	Mobile Services Switching Center/Visitor Location Register
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RAB	Radio Access Bearer
RAN	Radio Access Network
RLC	Radio Link Control
RNS	Radio Network Subsystem
RRM	Radio Resource Management
SGSN	Serving GPRS Support Node
SIR	Signal to Interference Ratio
SLA	Service Level Agreement
TDD	Time Division Duplex

TDMA	Time Division Multiple Access
TTI	Transmission Time Interval
UE	User Equipment (the UMTS equivalent to a Mobile Terminal)
UMTS	Universal Mobile Telecommunication System
USIM	UMTS Subscriber Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
VBR	Variable Bit Rate
WCDMA	Wideband Code Division Multiple Access

Contents

Chapter 1 - Introduction **1**

1.1 Research Motivation 6

1.2 Thesis Scope 8

1.3 Thesis Organization 8

Chapter 2 - Background and Related Work **10**

2.1 Measurement-Based Admission Control 10

2.2 Power Prediction 14

2.3 Call Admission Control and QoS Enforcing Mechanism 17

2.3.1 UMTS Admission Control Approaches 17

2.3.1.1 Measurement of the Air Interface Load 19

2.3.1.1.1 Uplink Load Factor 19

2.3.1.1.2 Load Estimation Based on Wideband Received Power 21

2.3.1.1.3 Load Estimation Based on Throughput 22

2.3.1.1.4 Comparison of the Uplink Load Estimation Methods 23

2.3.1.2 New RAB Load Increment Estimation 24

2.3.1.2.1 Load Increment Estimation Based on Received Power 25

2.3.1.2.2 Load Increment Estimation Based on Throughput 26

2.3.1.2.3 Comparison of the Load Increment Estimation Methods 26

2.3.2. QoS Enforcer 27

Chapter 3 - Implementation Details **31**

3.1 MBAC Implementation 31

3.2 Power Prediction Implementation 35

3.3 QoS-aware CAC and Power-sharing Structure Implementation 38

Chapter 4 - Performance Evaluation **44**

4.1. Model Description and Simulation Setup 45

4.1. MBAC Results 48

4.1.1 Higher Throughput Results 51

4.1.2 Equal Throughput Results 53

4.1.3. Equal Throughput Results 57

4.1.4 Higher Throughput Results 59

4.2. Power Prediction Results 60

4.3. QoS Enforcer 65

4.3.1 Power Staring Structure Functionality 65

4.3.2 Interworking of Power Prediction and Power Sharing 67

4.4. Summary of Results 72

Chapter 5 - Conclusions and Future Work **74**

Bibliography **77**

Appendix A - UMTS Background **80**

Appendix B - OPNET Modeler Overview **90**

Vita **102**

List of Figures

Figure 3.1: E_b/N_o values of a single RAB over a short period of time 32

Figure 3.2: MBAC Implementation for calculating the average total load through the E_b/N_o average 34

Figure 3.3: Mobility Prediction Logic 37

Figure 3.4: Preemption logic in RNC 40

Figure 3.5: RAB acceptance functionality 42

Figure 3.6: Algorithm for computing uplink capacity 42

Figure 3.7: Capacity Adjustment Function 43

Figure 4.1: Topology Snapshot 47

Figure 4.2: Aggregate and Average Uplink Eb/No Value Curves 49

Figure 4.3: Total Uplink Traffic using HTCAC and OCAC 50

Figure 4.4: Total Uplink Traffic using HTCAC and OCAC (Enlarged) 51

Figure 4.5: Blocking Ratio for HTCAC and OCAC (Higher Throughput Results) 52

Figure 4.6: Dropping Ratio for HTCAC and OCAC (Higher Throughput Results) 53

Figure 4.7: Blocking Ratio for HTCAC and OCAC (Equal Throughput Results) 54

Figure 4.8: Dropping Ratio for HTCAC and OCAC (Equal Throughput Results) 55

Figure 4.9: Total Uplink Traffic using HTCAC and OCAC 56

Figure 4.10: Total Uplink Traffic using HTCAC and OCAC (Enlarged) 56

Figure 4.11: Blocking Ratio for HTCAC and OCAC (Equal Throughput Results) 58

Figure 4.12: Dropping Ratio for HTCAC and OCAC (Equal Throughput Results) 58

Figure 4.13: Blocking Ratio for HTCAC and OCAC (Higher Throughput 59

Results)
Figure 4.14: Dropping Ratio for HTCAC and OCAC (Higher Throughput Results) 60
Figure 4.15: Blocking Ratio with and without Power Prediction 62
Figure 4.16: Dropping Ratio with and without Power Prediction 62
Figure 4.17: Total Uplink Traffic with and without Power Prediction 63
Figure 4.18: Dropping Probability versus Mobility Factor 64
Figure 4.19: Power Sharing 66
Figure 4.20: Power Sharing (2) 68
Figure 4.21: Effects from Using Mobility Prediction 70
Figure 4.22: Power Sharing with and without Prediction 71
Figure A.2: UMTS Network Architecture 82
Figure B.1: The three domains in OPNET 95
Figure B.2: Transceiver Pipeline Logic 100

Chapter 1

Introduction

The times we live in are distinguished by a remarkable variety of cellular standards operating throughout the globe. Users who need to stay connected cannot travel from one country to another without investing in multiple devices to comply with the different standards. Furthermore, the currently existing technologies are increasingly unable to satisfy the constantly growing requirements for augmented data rates and an enriched service mix.

A new standard has been developed to address both issues. First, it is designed to bring users closer to the concept of global roaming by providing a unified, inclusive approach to exchanging wireless information. Second, unlike present-day cellular networks, which are predominantly limited to voice services, it is dedicated to also offer data services with high bit rates and flexible capabilities. This emerging standard, now about to hit the market in Europe, is known as Third Generation (3G).

To better understand the technological significance of 3G technologies, we must put them the context of cellular communications in general, and provide some background about their predecessors – First and Second Generation systems (1G and 2G, respectively).

A cellular communications system is composed of both mobile and fixed components. The fixed side is made up of low-power wireless transceivers (base stations), interconnected through a fixed network. This is referred to as the backbone network. Base stations consist of an antenna and a controlling unit. They manage the air interface of their respective coverage areas, which are called "cells." The cell is the system's basic geographic unit and is typically represented as a hexagon. Cell sizes vary depending on the landscape and the number of users. The mobile side consists of mobile terminals that communicate with the base stations via wireless links. The most popular mobile terminal is the cell phone, although an increasing number of other types of wireless devices fall under this category.

Mobile terminals have two modes of communication with the base station, depending on whether they are sending or receiving information. These are called uplink and downlink, respectively. As uplink and downlink transmissions must be accommodated in the same (limited) spectrum, some form of governance needs to be imposed on their resource usage, as well as the way they are separated from each other. Traditional methods of separating uplink from downlink have been putting them on different frequencies, allocating different time slots to them or a combination of both. Thus, for example, 1G technologies utilized frequency division multiplexing (FDMA), and 2G technologies are time division systems (TDMA). The Global System for Mobile Communications (GSM), the most popular system nowadays, is hybrid and utilizes both FDMA and TDMA features.

One of the main conceptual issues in the design and operation of a cellular network is how to handle the mobility of the users. The handover process was developed to allow the seamless transition of a terminal from one cell over to another. Handover occurs when the cellular network automatically transfers a call from one base station to another because the first one can no longer maintain communication with the mobile.

UMTS (Universal Mobile Telecommunications System) is a 3G mobile technology developed within the International Telecommunications Union's IMT-2000 framework. One of several initially proposed 3G flavors, UMTS has been selected as the preferred technology for the majority of 3G operators. With data rates of up to 2Mbps, it is designed to provide services like mobile video telephony, quick access to information, fast download of data, and high-bandwidth Internet connectivity for people on the move.

The most essential UMTS components have already been drafted and standardized, and the basic operational requirements and the system architecture are already well understood. At the same time, a number of issues are not subject to standardization and are designated by the IMT-2000 as implementation-dependent. This provides an opportunity for competitive product differentiation between the different network manufacturers and operators. Thus there is plenty of room for innovations and enhancements in many areas within 3G. One such area is Call Admission Control, or CAC.

Admission control, part of Radio Resource Management (RRM), is one of the components responsible for the optimal utilization of the air interface resources. More specifically, it is a set of Radio Access Network (RAN) algorithms that determine whether a new Radio Access Bearer[1] (RAB) can be established or not.

Together with load control, power control and handover control, admission control helps guarantee the Quality of Service (QoS), maintain the planned coverage area, and maximize the system throughput. This needs to be done in an environment where the nature of the traffic is characterized by different bit rates, different services, and a mix of quality requirements. This means that implementing an intelligent, optimized admission control mechanism, can significantly increase the operability and the performance of a cellular network.

In second-generation cellular systems, like GSM, the decision whether a new call can be accepted or not can be considered a straightforward task since it only depends on the available number of channels in the cell. This is purely a fixed hardware limitation and depends on the number of channels and the frequencies assigned to each cell during the network planning process[2].

With the introduction of third generation cellular systems like UMTS, the two problems have changed their roles. Because of the use of WCDMA as the underlying technology, the task of frequency allocation no longer needs to be dealt with by network planning. This is due to the fact that Wideband Code Division Multiple Access is used, which, unlike the 2G's TDMA and FDMA, allows all ongoing connections to continuously use the whole wideband spectrum. As explained below

[1] Hereafter, the term "RAB" will be used instead of a "call" to signify a radio connection. It more accurately captures the nature of 3G communications because a UMTS link can involve not only traditional voice calls but also multimedia transmissions and web-browsing sessions.

[2] Network planning is the process of designing the fixed network infrastructure in terms of number of cell sites, cell site locations, number and architecture of communication nodes, etc. It is the complement of radio resource management, which is the way radio resources are dynamically managed in order to meet the instantaneous demands of users moving around the network.

and further in this study, this makes the system interference-limited and invalidates the number of connections as an accurate specifier of the actual capacity of the cell. This is especially true in the case of 3G systems, which, in addition to the traditional voice service, are required to support multimedia traffic, including data and video. (For a more detailed description of UMTS, please refer to Appendix A.)

In UMTS, the acceptance of a new connection depends on the Signal-to-Interference Ratio (SIR) values achievable by each existing connection once the new one is activated. These values are functions of the transmitted powers which, due to power control mechanisms, depend on the mobile user positions. Since the power available at both the base stations (referred to as Node Bs as the 3GPP-accepted[3] terminology stipulates) and the mobile terminals (called UEs for "User Equipment") is limited, the smaller the distance between the UE and the Node B, the larger the number of users that can be served in the cell. Thus the cell capacity is no longer fixed but also depends on the spatial distribution of the users and their individual services. This effect is called "cell breathing" and has important implications for the admission control process. The near-far problem is another important effect that admission control has to take into account. This is a situation in which a mobile close to the base station is received at a higher power than a mobile farther away, thus blocking the reception of the mobile which is far away. Combined with another important effect – soft capacity (explained in Chapter 2) – these issues make admission control a very complex problem.

[3] 3GPP, or the Third Generation Partnership Project, is an international consortium of companies and organizations responsible for drafting and standardizing the technical specifications a 3rd Generation Mobile System based on evolved GSM core networks and radio access technologies. The consortium has until now drafted and released a complete set of specifications, which is the basis for the UMTS test systems currently launched in Europe.

1.1 Research Motivation

As suggested in the preceding section, the function of admission control is to maximize the capacity, while providing QoS guarantees and maintaining the stability of the network. Thus, its optimization is of vital importance to both manufacturers and operators alike.

In an environment characterized by a limited available spectrum, the available resources must be utilized as efficiently as possible. In this respect, admission control has to make sure that the radio access network operates at a point where there are no unused radio interface resources. At the same time, the guaranteed QoS for the ongoing calls must always be maintained. This means that CAC should never admit more calls than it can support. The coexistence of these two tasks is complicated by the volatile nature of the wireless medium. Undesired effects like thermal noise, shadowing and fading[4] make the total available resources vary with time.

Therefore the basis of any good admission control mechanism should be a carefully drafted trade-off between the blocking ratio (the probability that the connection will not be admitted in the first place) and the dropping ratio, which is the probability that the connection would be forcefully terminated because the system has run out of resources.

Because admission control, like most RRM strategies, has not yet been standardized and is thus an open issue, addressing the problem has attracted a great deal of

[4] Thermal noise is a measure for the amount of external interference generated by various sources in and around the cell. Shadowing refers to the slow variations in propagation conditions due to terrain changes caused by objects like trees and foliage. Fading and fast-fading result from fluctuations of the received signal over short periods of time as a result of multipath propagation. Multipath propagation refers to a radio signal, which takes two or more paths to the antenna because it is reflected off buildings or other obstructions.

research effort. This has been especially true within the framework of CDMA schemes where various admission control schemes have been widely proposed and studied. Some of them have been mathematically derived; others have been simulated or implemented in testbeds. Nevertheless, few studies in the open literature regarding admission control approach realistic scenarios aligned with the actual 3G standards and specifications. Most consider general CDMA cases and never take into account the actual specifics of a full-fledged UMTS system.

Therefore, the aim of this thesis is to propose a viable, realistic, UMTS-tailored, admission control mechanism while closely following all applicable 3GPP standards and requirements. The measurement-based and the power prediction modules have been adapted and have been both put in the context of UMTS. One of the most popular currently available CAC proposals[5] is extended to provide the basis for an enhanced, QoS-based, admission control scheme.

At the same time, it should be noted that the concepts presented and the extensions developed within the framework of this thesis are not only UMTS-specific but can be applied in other systems which have CDMA as their underlying technology. Thus, for example, other possible systems that can see improvement using the ideas presented herein are Multi-carrier CDMA, TDM/CDMA and others.

[5] The proposal used and extended is based on the work of two leading Nokia researchers who have greatly contributed to the UMTS development and standardization efforts. Their book "WCDMA for UMTS" has been one of the main sources of information for the development of the OPNET Modeler UMTS model, which is the simulation package used in this thesis.

1.2 Thesis Scope

The scope of this thesis is limited to investigating admission control in the uplink direction of the radio access network. *Uplink* is the communication path from a mobile terminal to a Node B, as opposed to *downlink* communication from a Node B to a mobile terminal. The distinction between the uplink and the downlink is made since their implementation and performance are vastly different and radio resource management operates largely independently for both directions. This study also assumes that frequency division duplex (FDD) mode is used. FDD, in contrast with TDD (time division duplex), is a communication method that separates uplink and downlink channels by allocating them different frequencies, rather than different time intervals. Finally, for the sake of simplicity, only three of the four UMTS classes are used in this thesis. Extending this to a four-class scenario should be straightforward.

1.3 Thesis Organization

A top-down approach is used to organize this thesis. First, a general introduction to the field is presented in Chapter 2. The three main components of interest are presented and background information is given about each of them. Then, Chapter 3 gives a detailed description of the modifications and enhancements that have been implemented to create an improved QoS-based uplink admission control scheme. The performance of the proposed strategy is investigated with simulations using a system level simulator in Chapter 4. A conclusion and some suggestions for future work make up Chapter 5. Finally, Appendix A gives an overview of UMTS, its fundamental building blocks and the mechanisms that manage their behavior, including WCDMA and RRM.

Chapter 2

This thesis proposes a framework to enhance existing UMTS networks on three accounts. First, it introduces a measurement-based component; second, this measurement-based component is integrated with a power prediction module; and third, the proposed framework feeds the results obtained to a call admission control algorithm with a QoS-enforcing mechanism.

Research in all three areas has been active over the past few years. Various schemes have been proposed and simulated. Each has claimed that it helps achieve an improvement in some aspect of the system behavior. However, the contexts of these proposals have been rather heterogeneous, and their implementations have been more or less case- or system-specific. Although it is logical that some of the techniques would work in a 3G scenario as well, it is hard to make that claim without the proper research and simulation. The aim of this thesis is to capitalize on the possibilities this niche provides.

2.1 Measurement-Based Admission Control

The first component, measurement-based admission control (MBAC), is one of two possible approaches to admission control. It came to replace the older, more rudimentary approach, parameter-based admission control. MBAC gained popularity when the first attempts for QoS provisioning in wireline networks were made [1].

The Asynchronous Transfer Mode (ATM) is a prominent example. It has been shown that for many new applications, some characteristics of the traffic behavior could not be provided. Without these traffic parameters, parameter-based admission control schemes would either not work or would overestimate the required bandwidth, causing low network utilization [2].

Measurement-based admission strategies, on the other hand, do not calculate the worst-case behavior of the existing connections but use actual system measurements. A decision rule is applied to predict how the performance will be affected by the admission of a new flow, based on the measured data. Measurements are combined with the characterization the new traffic provided by the admission request in order to decide whether the new RAB can be accepted or not.

A key technique that will be borrowed from MBAC is the use of some type of moving average for the value of the actual system load. This is needed because it is often difficult to correctly evaluate the amount of resources needed for all connections, especially if they are of a variable bit rate (VBR) nature. This problem is even more sharply posed in wireless systems like WCDMA, where the amount of resources used by a connection depends on the interference actually generated. The generated interference, in turn, depends on a combination of factors like the bit rate, the requested bit error rate (BER), and the (often changing) distance between the UE and the Node B. When any of these factors changes, so does the amount of resources the connection requires. At the same time, to make things even more complicated, the total capacity of the system is variable as well. This is due to the wireless medium dynamics that is inherent to UMTS. Varying levels of interference for example can cause the total resources to periodically shrink. The combination of these changes in the wireless medium, the user mobility and the bursty nature of packet traffic produce a dynamic system that simply cannot be described using static parameters.

The specifics of the moving average MBAC implementation will be explained in subsequent chapters. It should also be mentioned that multiple flavors and variations of MBAC routines have been proposed and studied. Examples are in [3] – [5].

The following list summarizes some of the major measurement-based approaches. Although the list is by no means comprehensive, the described algorithms represent a broad sample of existing MBAC proposals [6].

Measured Sum (MS)

The Measured Sum algorithm admits a new flow of rate *r* if the sum of the rate of this flow and the estimated rate of existing flows is less than a utilization target times the link bandwidth. Therefore, the new flow is admitted if

$$\hat{v} + r < \upsilon\mu,$$

where $\hat{v}$ is the measured load of existing traffic, μ is the link bandwidth, and υ is a user-defined utilization target, designed to limit the maximum link load. Upon admission of the new flow the load estimate is increased by $\hat{v} = \hat{v} + r$. A time window estimator is used to derive the estimated rate of existing flows.

Hoeffding Bounds (HB)

The Hoeffding Bounds algorithm computes the equivalent bandwidth for a set of flows using the Hoeffding bounds. A new flow is admitted if the sum of the peak rate of the new flow and the measured equivalent bandwidth is less than the link utilization. An exponential averaging measurement mechanism is used to produce the load estimate.

Tangent at Peak (TP)

This algorithm is based on the tangent at the peak of an equivalent bandwidth curve computed from the Chernoff Bounds. It admits a new flow if the following condition is met:

$$np(1-e^{-sp})+e^{-sp}\hat{v}\le\mu,$$

where n is the number of admitted flows, p is the peak rate of the flows, s is the space parameter of the Chernoff Bound, $\hat{v}$ is the estimate of the current load, and μ is the link bandwidth. This algorithm uses a point sample measurement process.

Measure CAC (MC)

The Measure admission control algorithm is based on deviation theory. It admits a new flow if the sum of the peak rate of the flow and the estimated bandwidth of existing flows is less than the link bandwidth. The estimated bandwidth takes as input a target loss rate and makes use of the scaled cumulative generating function of the arrival process.

Aggregate Traffic Envelopes (TE)

This algorithm uses the measurements of the maximal traffic envelopes of the aggregate traffic, capturing variability on different time scales. Both the average and the variance of these envelopes, as well as their target loss rate, are used as input into the admission algorithm.

In general, MBAC proposals range from simple and ad hoc routines to complicated mathematical derivations. MBAC research is typically focused on equations and, to a lesser extent, the measurement algorithms used to decide whether or not to accept an incoming flow. The algorithms presented differ in both the underlying theory and the specific measurement and admission control equations they use. However, a comparative performance study of MBAC approaches finds out an interesting fact. It appears that although MBAC is clearly better than ordinary, parameter-based, non-MBAC approaches, most MBAC algorithms, under nominal conditions, achieve almost identical levels of performance, regardless of their level of complexity. That is, for a given level of QoS, they all achieve very similar levels of utilization [7]. Therefore, we argue that while this thesis will make use of a simple form of MBAC, the results obtained should be commensurate with the results that would have been obtained with a more sophisticated scheme. This is in fact an important point to make and is very appealing in real systems like UMTS where complexity is already a big issue.

2.2 Power Prediction

The next component under consideration, power prediction, has also been focus of research, which, naturally emerged only after the advent of wireless technologies. Power prediction is closely related to mobility prediction as the transmission power in cellular systems always depends on the location of the mobile with respect to the base station. The aggregate transmission power within a cell is a very important parameter to observe, as it is equivalent to the total interference in the current cell. As was stated in the preceding sections, the more interference is generated in a cell, the fewer connections can be successfully maintained. Therefore, it is one of the primary aims of radio resource management to minimize the transmission power so that it is just enough to maintain the contracted QoS level, and not even a fraction more. This issue is discussed in more detail in the Power Control section of the UMTS description.

Various mobility prediction methods have been studied in the context of designing mobile reservation protocols. Examples can be found in [7]–[11]. Table 2.1 below summarizes some of the major mobility prediction proposals. It should be noted that despite the substantial research efforts in this direction, the actual user mobility patterns are not yet well understood [11].

Proposed Scheme	**Prediction approach and user mobility model**	**Type of improvement and prediction**
Mobile Motion Prediction Algorithm	• Prediction is based on the user's movement history. • Movements consist of regular and random components, which can either be matched with circle/track patterns or simulated by the Markov chain model.	• Migration of mobile-floating agents for service pre-connection, resource pre-assignment and data pre-fetching. • Prediction is highly accurate with regular movements but accuracy decreases linearly with increasing random component.
Hierarchic Position-Prediction Algorithm	• Prediction is based on the user's movement history and instantaneous signal strength measurements of surrounding cells. • User's movement can be mapped into previous mobility patterns, with matching operations such as insertion, deletion and changing. • Intercell movements can be also estimated by current location, velocity and cell geometry.	• Setting up and reserving resources along a mobile's path, and planning quick handovers between the base stations. • Minimizing the occurrence of location registration and update procedures. • Prediction remains reasonably accurate (75%) despite the influence of random movements.
Profile Based Next-cell Prediction Algorithm	• Prediction is based on the user's movement history and the classification of locations. • Mobility can be purely random, mostly random, purely deterministic, and mostly deterministic. • The type of location can be office, corridor and common room.	• Providing advance reservation and adaptation in resource management. • Prediction is highly accurate with fully predictable movements, 80% accurate for typically observed movements and 70% for fully random movements.

Shadow Cluster Concept	• Prediction is based on the user's movement history. • Movements simulate highway traffic with various constant speeds traveling in forward and backward directions.	• Estimation of resource requirement and decision of call admission. • The percentage of dropped calls reduces from 14% to 1% (without and with mobility prediction, respectively), with a minimal deduction of bandwidth utilization from 30% to 25% (without and with prediction, respectively).
Per-user Profile Replication Scheme	• Prediction is based on the user's movement history. • Mobility model is derived from statistical analysis of actual call traffic traces, vehicle and airplane traffic data, and government transportation surveys. • Simulated movements can be random walks and repetitive roundtrips.	• Improving the efficiency of location management. • Compared with a basic hierarchical model, this scheme can serve more than 90% of calls by local database lookups, with the penalty of a slight increase in bandwidth requirement and twice the memory requirements.

Table 2.1: A brief description of some mobility management proposals

Mobility prediction, however, can serve several different purposes depending on the type of system it is used in [11]. In mobile reservation systems, the main goal is to predict the handover time from one cell to another so that adequate resources in the new cell can be reserved. This may entail blocking of new connections in order to reserve some free capacity for the connection that will be handed over. The logic behind this is that it is usually much more annoying for a user to lose an ongoing call than not to be able to establish a new one in the first place. In 3G systems, the task of mobility prediction is greatly simplified by the ability of the neighboring Node B to estimate accurately the handover time well in advance. This is made possible by centralizing the administration of multiple adjacent Node Bs by one radio network controller (RNC) and the principle of soft handover, which allows a mobile in transition to be simultaneously connected to both the base station in the cell it is departing and the base station in the cell it is entering.

Mobility prediction in 3G systems plays a slightly different role. Here the required transmission and reception power levels directly determine the air interface capacity, since they determine the transmitted interference. This is true both for the uplink and the downlink. Thus mobility prediction in 3G can assist in the admission control process by providing an estimate for the total amount of interference that would be generated in the system after the mobiles have moved in a certain quasi-random pattern.

Finally, it should be noted that power control plays an important role in providing optimal system operation. It tries to minimize transmission powers of both mobiles and Node Bs so that the available system capacity would be maximized.

2.3 Call Admission Control and QoS Enforcing Mechanism

The core of the QoS Enforcer module is the admission control algorithm. Taking into account the QoS information from the request under consideration, as well as the continuous data coming from MBAC and the power prediction modules, admission control makes the decision as to whether to grant the request or to reject it. Therefore, in order to give details on how the QoS mechanism works, we first need to give an account of the actual admission control algorithm.

2.3.1 UMTS Admission Control Approaches

If the air interface loading is allowed to increase excessively, the coverage area of the cell will be reduced below the planned values, and the quality of service of the existing connections will not be guaranteed [12]. Before admitting a new connection, admission control needs to check that the admission will not sacrifice the planned coverage area or the quality of the existing connections. In UMTS, the admission control algorithm is executed every time a radio access bearer is set up or modified. This thesis will not consider modifications to the radio access bearer (RAB), i.e., run-time changes in the QoS specification of the flow, since it is out of the scope of the research goals.

The CAC functionality is located in the RNC where the load information from several cells can be obtained. The admission control algorithm estimates the load increase that the establishment of the radio bearer would cause in the radio network. This has to be estimated separately for the uplink and the downlink directions. The requesting bearer can be admitted only if both uplink and downlink admission control admit it, otherwise it is rejected because of the excessive interference it would produce in the network. The limits for admission control are set by radio network planning.

In the open literature, there are two competing approaches as to how the admission control problem should be addressed. These are Wideband power-based CAC and Throughput-based CAC. There have been a number of publications about both strategies. Power-based admission control for example is studied in [13]–[15]. Throughput-based admission control is presented in [12], [16] and [17].

The call admission control algorithm can be broken down into two sub-steps that need to be accomplished so that the final decision can be made. These steps are the measurement of the current air interface load and estimating the load increment that the new request will bring into the system.

2.3.1.1 Measurement of the Air Interface Load

Within the framework of the two CAC strategies, and as far as the uplink direction is concerned, there exist two major approaches for load estimation that can be used in WCDMA networks: load estimation based on wideband received power, and load estimation based on throughput.

Before we explore these two issues, however, we must present an important concept, which is essential for the task of measuring the air interface load. This is the load factor which is a concept that inter-relates specific aspects from radio dimensioning, capacity planning and resource estimation.

2.3.1.1.1 Uplink Load Factor

The load factor η is used to indicate the total load in the air interface. It is a measure for the spectral efficiency of a WCDMA cell. For example, if the load is said to be 60% of the (current) CDMA capacity, this means that the load factor $\eta = 0.60$. Because of the asymmetric traffic characteristic of 3G systems, there is a separate load factor for the uplink and downlink directions. As this thesis is dealing with the uplink case, only the uplink load factor, η_{UL}, will be discussed.

Apart from η, which is a measure of the aggregate cell load in the uplink or the downlink, another notion exists to quantify the load that each individual connection brings into the system. This is the individual load factor. The sum of the individual load factors of all active uplink connections yields $\eta_{UL.}$ As the individual load factor is a core concept in the functioning of any UMTS admission control protocol, we next present the mathematical foundations on which load factor calculations are based. We first define E_b/N_o, the energy per user bit divided by the noise spectral density. This is simply a ratio of the energy carried by each user (information) bit to the external noise in the air interface.

$$(E_b/N_o)_j = \text{Processing gain of user } j \,.\, \frac{\text{Signal of user } j}{\text{Total received power (exc. own signal)}} \qquad (2.1)$$

This can be written as

$$(E_b/N_o)_j = \frac{W}{R_j.v_j} . \frac{P_j}{I_{total} - P_j}$$

where W is the chip rate, P_j is the received signal power from user j, v_j is the activity factor of user j, R_j is the bit rate of user j, and I_{total} is the total received wideband power including thermal noise power in the base station. Table 2 in section 2.3.1.1.3 gives more insight into these parameters. Solving for P_j gives

$$P_j = \frac{1}{1 + \frac{W}{(E_b / N_o)_j . R_j \,.\, v_j}} + I_{total}$$

If we define $P_j = L_j \,.\, I_{total}$ we can obtain the load factor L_j for one connection

$$L_j = \frac{1}{1 + \frac{W}{(E_b / N_o)_j . R_j \,.\, v_j}} \qquad (2.2)$$

Using the individual load factor for every connection is a convenient way to describe this connection's particular contribution towards the total load level. Clearly, the sum of the individual (uplink) load factors yields the total load factor η_{UL}. This does not include the external interference and background noise, which have to be accounted for separately. Load factors are pivotal to admission control and will be repeatedly used further on in the discussion. The notion of using a load factor is equally applicable to both power-based and throughput-based strategies.

2.3.1.1.2 Load Estimation Based on Wideband Received Power

This approach makes use of the fact that the received power levels can very conveniently be measured by the base station. Based on those measurements, the uplink load factor can be obtained.

The received wideband power, P_{RX_total}, also equivalent to the total interference I_{total}, can be divided into the powers (interference levels) of own cell (intra-cell) users I_{own}, other cell (inter-cell) users, I_{oth}, and background and receiver noise, P_N:

$$P_{RX\_total} = I_{total} = I_{own} + I_{oth} + P_N \tag{2.3}$$

The uplink load level can be derived from the equation below. As shown in [18], the stage of calculating and summing the individual load factors can be skipped as the power-based approach directly yields the final result:

$$\eta_{\mathrm{UL}} = \frac{I_{own} + I_{oth}}{P_{RX\_total}} \tag{2.4}$$

This case heavily relies on hardware mechanisms inside the Node B to perform the actual measurements.

2.3.1.1.3 Load Estimation Based on Throughput

An alternative way for estimating the uplink load factor is by summing of the individual load factors of the users that are connected to this base station. The derivation takes into account the external interference and is shown in Equation 2.5 below:

$$\eta_{UL} = (1+i) \cdot \sum_{j=1}^{N} L_j = (1+i) \cdot \sum_{j=1}^{N} \frac{1}{1+\frac{W}{(E_b/N_o)_j \cdot R_j \cdot v_j}} \qquad (2.5)$$

The terms and variables that make up this equation are listed in Table 2 below.

Parameter	Definition	Recommended value
N	Number of users per cell	
v_j	Activity factor of user j at physical layer	**0.5 + overhead** for speech, **1.0** for data
E_b/N_o	Signal energy per bit divided by noise spectral density that is required to meet a predefined QoS level (e.g. bit error rate). Noise includes both thermal noise and interference.	Dependent on service, bit rate, multipath fading channel, mobile speed, et.
W	WCDMA chip rate	**3.84** Mcps
L_j	Load factor for one connection	Dependent on service
R_j	Bit rate of user j	Dependent on service
i	Other cell to own cell interference ratio as seen by the base station receiver	Macro cell with omnidirectional antennas: **55%**

Table 2.2: Parameters used in the throughput-based load estimation equation

According to the current proposal, load estimation uses the instantaneous measured values for E_b/N_o, i and v and the number of users N to estimate the instantaneous air interface load. This is one of the major weaknesses of the scheme. This is because the estimation is reliable and accurate only if it spans a certain interval of time. In throughput-based estimation, interference from other cells is not directly included in the load but needs to be taken into account with the parameter *i*. Because a single-cell scenario is considered in this thesis, the value of *i* will always be assumed to be zero.

2.3.1.1.4 Comparison of the Uplink Load Estimation Methods

Table 2.3 compares the above load estimation methods. For comparative purposes, a 2G (GSM) load-estimation method (based on the number of connections) is also included.

Criterion	Wideband received power	Throughput	Number of connections
What to measure	Wideband received power I_{total} per cell	Uplink E_b/N_o and bit rates R for each connection	Number of connections
What needs to be assumed or measured separately	Thermal noise level (external interference power)	Other-to-own interference ratio	Load caused by one connection
Other-cell interference	Included in the measurement of the wideband received power	Specified explicitly in *i*	Assumed explicitly when choosing the maximum number of connections
Soft capacity	Yes, automatically	Not directly, possible via RNC	No
Other interference sources	Reduced capacity	Reduced coverage	Reduced coverage

Table 2.3: Comparison of uplink load estimation methods

In the wideband power-based approach, interference in the adjacent cells is directly included in the load estimation because the measured wideband power includes all interference that is received in that carrier frequency by the base station. If the loading in the adjacent cells is low, then a higher load can be attained in the current cell. This is the principle of soft capacity.

The problem with wideband power-based load estimation is that the measured power can include interference from adjacent frequencies. This could originate from another operator's mobile located very close to the base station antenna. The Node B receiver cannot separate the interference from an own carrier and from other carriers by the wideband power measurements. Therefore, the interference-based method can overestimate the load of the own carrier.

Throughput-based load estimation does not take the interference from adjacent cells or adjacent carriers directly into account. If soft capacity is required, information about the loading of the adjacent cells can be obtained within the RNC. If the loading in the adjacent cells is high, the coverage area of the cell is affected.

The third load estimation method in the table is simply based on the number of concurrent connections. This approach is used in second-generation networks where all connections use fairly similar low bit rates and no high bit rate connections are possible. In 3G systems, the mix of different bit rates, services and quality requirements prevents the use of this approach. It is not reasonable to assume that the load caused by one 2-Mbps user is the same as that caused by one speech user.

2.3.1.2 New RAB Load Increment Estimation

Once the current load of the system has been estimated, the load increment that the addition of the new connection would bring into the system needs to be calculated. In keeping with the taxonomy assumed in the preceding section, the possible alternatives can be again divided into two broad categories – power-based and throughput-based. Here, however, the differences are less prominent and the two cases share some significant similarities.

2.3.1.2.1 Load Increment Estimation Based on Received Power

In power-based CAC, the use of the total power received by the base station is used as the primary uplink admission control decision criterion. In the uplink, a new user is admitted if the new resulting interference level is lower than the threshold value:

$$I_{total\_old} + \Delta I > I_{threshold}$$

Where I $_{total\_old}$ is the total interference before the admission of the new bearer, ΔI is the interference the new connection would bring into the system if admitted, and $I_{threshold}$ is the network planning-set limit.

The change in interference uplink power can be obtained by Equation (2.6). The equation is based on the assumption that the power increase is the derivative of the old uplink interference power with respect to the uplink load factor, multiplied by the load factor of the new user ΔL.

$$\begin{aligned} &\frac{\Delta I}{\Delta L} \approx \frac{dI_{total}}{d\eta} \Leftrightarrow \\ &\Delta I \approx \frac{dI_{total}}{d\eta} \Delta L \Leftrightarrow \\ &\Delta I \approx \frac{P_N}{(1-\eta)^2} \Delta L \Leftrightarrow \\ &\Delta I \approx \frac{I_{total}}{1-\eta} \Delta L \end{aligned} \tag{2.6}$$

In the equation, the load factor of the new user ΔL is the estimated load factor of the new connection. The equation was already introduced in Section 2.3.1.1.1 but is repeated here for convenience:

$$\Delta L = \frac{1}{1 + \frac{W}{(E_b / N_o)_j . R_j . v_j}} \qquad (2.7)$$

where W is the chip rate, R_j is the bit rate of the new user, E_b/N_o is the assumed E_b/N_o of the new connection, and v is the assumed activity factor of the new connection. As the aim of this study is to explore packet switched data, the voice activity factors of all connections will be set to 1.

At the same time, the value of the assumed E_b/N_o of the new connection (called "target E_b/N_o" from now on) will be dynamically calculated at the time of admission. There is a direct correspondence between the target E_b/N_o, the requested bit rate and the requested BER.

2.3.1.2.2 Load Increment Estimation Based on Throughput

In throughput-based CAC, the new requesting user is admitted into the radio access network if

$$\eta_{UL} + \Delta L \leq \eta_{UL\_threshold}$$

where η_{UL} is the uplink load factor before the admittance of the new connection and is estimated as described in Section 2.3.1.1.3. The load factor of the new user ΔL is calculated as in Equation (2.7), that is, just in the same way as the load factor in the power-based strategy.

2.3.1.2.3 Comparison of the Load Increment Estimation Methods

As was the case with the load estimation methods, the two load increment calculation strategies are based on different techniques. Each has its potential benefits and disadvantages. What is common to both, however, is the calculation of the load factor of the new connection ΔL, obtained through Equation (2.7).

The logic behind this equation is central to the research in this thesis. The equation comprises elements from both the power-based domain, and the throughput-based domain. On the one hand, the term E_b/N_o signifies the required strength of the transmitted signal in terms of the actual energy the mobile uses to transmit one user bit of information at the requested bit error rate. It is measured in dB or dBm and is readily convertible to Watts[6]. On the other hand, the actual bearer throughput is specified through the use of the rate R (measured in bps). Thus if any aspect or term in this equation is manipulated, it will have an effect on both CAC strategies.

2.3.2. QoS Enforcer

The QoS Enforcer module is the mechanism responsible for maintaining the power sharing structure and guaranteeing the contracted share of each class.

There are several levels at which QoS can be provisioned in the UMTS air interface, and CAC is one of them. According to the 3GPP recommendations, at least the following functions are needed: QoS monitor, QoS translator, QoS policer, QoS shaper, Resource Manager, and Admission controller [20].

There are two ways for admission control to help support QoS. First, RAB re-negotiation can be implemented to dynamically change the QoS characteristics of the

[6] dBm stands for Decibels below 1 Milliwatt and is a general measurement of power loss in decibels using 1 milliwatt as the reference point. A signal received at 1 mW yields 0 dBm, and a signal at 0.1 mW is equivalent to 10 dBm. The UMTS standard for the maximum UE transmission power at any one moment is approximately 24 dBm, or 250 mW.

flow. Although this function is performed by the admission control functionality, it is in fact logically part of congestion control. It is out of the scope of this thesis.

The second mechanism for CAC to support QoS provisioning is to allocate resources to as many connections as possible, treat each traffic class according to the corresponding traffic contract, and keep the system from congestion (i.e., never over-allocate resources). In order for us to understand better the mechanism for giving each class different treatment, the concept of power sharing needs to be introduced.

Power sharing is based on the concept that the total resources can be distributed among the traffic classes as needed. Thus, the sum of the capacities allocated to all classes equals the total system capacity. The principle of power sharing is exemplified in the ability of any of the traffic classes, regardless of QoS, to be granted all resources if there is no other traffic in the cell. Thus one class would be allowed to borrow capacity from another in case the first class is not using it. This would allow the available resources to be utilized more efficiently.

At the same time, admission control needs to guarantee that in case there is other traffic in the cell, each class will get at least its designated share. The need for such functionality is especially vivid in case there is enough traffic to saturate the whole capacity of the system and the flows that make up this traffic belong to all traffic classes. Adopting an approach like this will provide fairness by preventing a traffic class from monopolizing the whole capacity. The allocated shares and the way they are distributed among the classes depend on the operator's policies on QoS and the corresponding service level agreements (SLAs).

Higher-priority connections need to be guaranteed that regardless of the current load, they will always be guaranteed a certain share of the total air capacity. In case there are no other traffic classes competing for access in the same cell, high-priority flows can be allocated the whole air interface capacity. Depending on the particular network operator's QoS policy, high-priority RABs may or may not be forcefully terminated (or preempted) if they exceed their allocated share. Lower-priority connections, on the other hand, will always be preempted if they exceed their allocated share.

For the purposes of this thesis, the admission/rejection functionality presented in [12] is used. It has been extended to include support for three traffic classes with dynamically adjustable share boundaries and forceful termination of oversubscribing flows.

2.4. Summary

This chapter described the framework in which the enhanced CAC scheme is implemented. The measurement-based (MBAC) component was introduced and its applicability was discussed. It is important to note here that a simple MBAC approach is used as a means to illustrate a concept. Naturally, other MBAC schemes are possible, which can potentially produce better results. Power prediction is again addressed in one out of many possible ways. While a simple Brownian movement is assumed, other approaches can be taken. The concept of call admission control was introduced and some insight into the logic behind it was given. As this is an implementation-dependent issue, there could have been several alternatives here as well. In this sense, the scheme presented is an illustration of the viability of different combinations of components, as well as a testament to the flexibility of the existing opportunities. The specific implementation details and performance results are presented in the chapters to follow.

Chapter 3

Implementation Details

This chapter gives a detailed overview of the proposed algorithms and the way they are implemented. The MBAC implementation is presented first and a moving average algorithm is described. Power prediction is exemplified in a mobility prediction algorithm based on a random mobility model. QoS-aware CAC is discussed next and the pre-emption logic governing the behavior of the power sharing structure is presented. The functions presented are part of the admission control module and would be physically located in the Radio Network Controller.

3.1 MBAC Implementation

As explained earlier, the need for an MBAC aspect in the proposed admission control scheme arises from the fact that the current proposal [12] is based on the use of the instantaneous measured values of the run-time system parameters to estimate the current load in the cell. By referring to Equation (2.7), the most important such parameter is the E_b/N_o value. The other parameters, the number of users N, the activity factor v, and the other-to-own interference level i, will be kept constant throughout the simulation. This is needed so that the E_b/N_o value would be isolated as the only parameter that would be manipulated. In this way we can be sure that whatever effects or behavioral changes appear in the system, they would stem from none other than the different handling of E_b/N_o.

At the same time, it should again be emphasized that the aggregate E_b/N_o value in a cell is proportional to the required transmission power, which means that is has a direct correspondence to the total interference generated in that cell. Therefore, any admission control scheme needs to be extra careful when estimating the exact value that should be used whenever a new RAB requests access to the medium. As was made clear in the preceding sections, using the instantaneous value of E_b/N_o is not an appealing strategy due to the varying nature of the air interface. This point is well illustrated in Figure 3.1 below, which shows the varying nature of the E_b/N_o curve for a typical scenario.

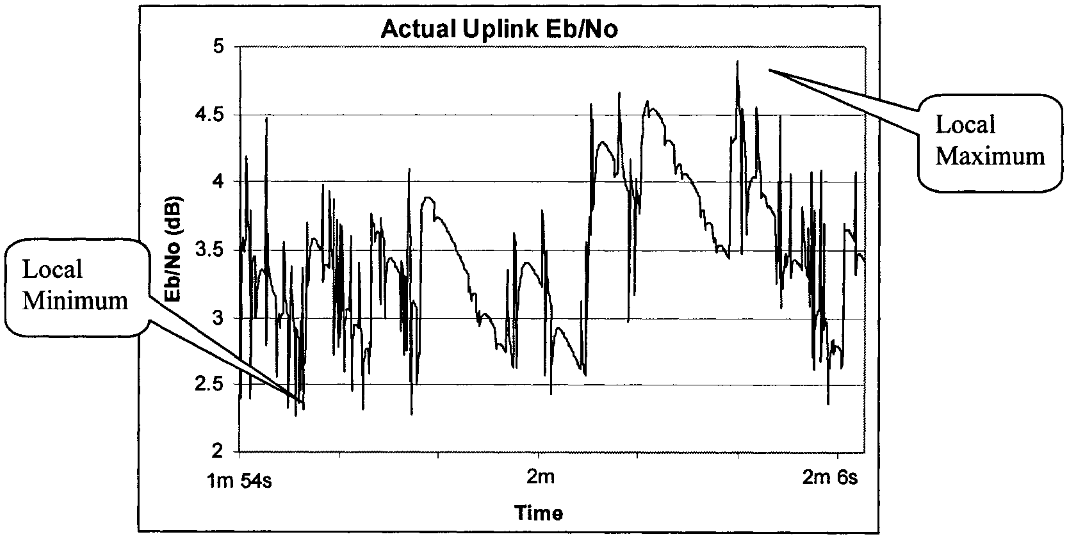


Figure 3.1: E_b/N_o values of a single RAB over a short period of time[7]

[7] The spike/decay pattern easily visible in this graph is due to the operation of power control, to which the E_b/N_o value is directly related. Whenever a packet is successfully transmitted, transmission power is decreased by a small fraction, ΔP. After the successful transmission of another packet, power is again decreased by the same amount, and so on. However, when a packet is in error, transmission power is ramped up by ten times ΔP. This explains the big spikes after periods of gradual decay.

The figure shows that the value of E_b/N_o varies from 2.3 dB at t = 1m 55s (at the local minimum) to 4.8 at t = 2m 4s (the local maximum). This is almost 109% change of the E_b/N_o value only in the course of 9 seconds. Depending on whether the instantaneous value of E_b/N_o happens to be around 2.5 or around 4.5, we can have two cases. First, if the current reading is below the average actual load, the CAC module is passed information that is overly optimistic. This can lead to the inappropriate admission of a new radio access bearer, and the system may not have enough resources to support this extra traffic. Second, if the current reading is above the actual average load, the load estimation is too pessimistic and this could well mean the rejection of a requesting bearer when this bearer could have been supported. To support this claim, a simple experiment was run in a single-cell, single-user scenario where continuous traffic was transmitted from a mobile terminal to an Ethernet-based server at a data rate of 80 Kbps and a BER of 0.01. At connection time, the requested E_b/N_o value was 4.07 dB, which was about 10.6% of the available air interface capacity. At the same time, if the requested E_b/N_o value were only 1 dB lower or 1 dB higher, i.e., 3.07 dB or 5.07 dB, respectively, the actual percentage of resources needed would have been calculated by CAC as 8.3% and 12.9%, respectively. With the presence of many high-bandwidth users, an inaccuracy of this scale will imminently lead to one or more flows being wrongfully accepted or rejected. In other words, some of the time, the system would be characterized by higher than necessary blocking, and some of the time there would be excessive dropping. To the network operator both cases translate into decreased levels of service, inefficiencies, and lost revenue.

To avoid situations like the ones described above, it would be beneficial to use some form of average to account for sharp, unrepresentative, variations of the observed parameters. Therefore, to give a more accurate representation of the actual conditions in the cell, a moving E_b/N_o average has been implemented. It was demonstrated in [8] that most MBAC approaches achieve similar results. Therefore, a simple moving average is used.

A two-dimensional FIFO queue has been implemented to hold the last n time-value pairs consisting of the Transmission Time Interval (a short TTI explanation is given under footnote 9 in section 3.1) and its corresponding E_b/N_o value. Given the particular traffic parameters of the cell, which this thesis examines, it has been found that an optimal value for n is when it is equivalent to around 15 sec. This means that the current uplink load estimate is based on the load during the last 15 sec. Therefore, sharp variations in traffic volumes take more time to be registered. To compensate for this, an additional "smoothness" limitation must be imposed on the traffic generation parameters. In other words, to claim that this research has any realistic value, it should be highly unlikely for the used parameters to allow little or no traffic during a time interval of, for example, 1 minute, and then a steep increase in the number of requested connections during the next 10 seconds. In fact, it is well known that such irregularities seldom arise, as the incoming calls in any standard switching center arrive according to a Poisson distribution.

The code block in Figure 3.2 shows the MBAC function, which estimates the average total uplink load.

```
    current_UL_load_estimate

    I. Calculate the Eb/No average per user for the last n sec
       using a global two-dimensional time-value array

1.    for (each 20 ms interval for the last n sec) do // have to probe each TTI
2.        for (all active connections) do
3.                aggregate_eb_no += eb_no_array[connection++][time++]
      end for
      end for
4.        average_eb_no = aggregate_eb_no/(number_of_values)

    II. Calculate total current average uplink load
5.        for (all active connections) do
6.          current_average_UL_load +=
7.              1 / (1 + (uplink_proc_gain[conn] / (activity_factor[conn] * average_eb_no)));
            end for
8.  return current_UL_load
end UP_load_estimate
```

Figure 3.2: MBAC Implementation for calculating the average total load through the E_b/N_o average

Lines 1-3 sum the E_b/N_o values of all active connections for the last *n* seconds. Line 4 calculates the E_b/N_o average per user and stores it in the *average_eb_no* parameter. This parameter is then used in lines 5-7 which implement Equation (2.7). The variable *uplink_proc_gain* has been calculated earlier. As shown in Figure 3.6, it is equal to the WCDMA chip rate *W* divided by the data rate *R*. The average total uplink load is returned by the function in line 8.

3.2 Power Prediction Implementation

The required power levels in a wireless system are to a large extent dependent on the physical position of the UE with respect to the Node B. Therefore the basis of any power prediction algorithm should be mobility prediction. This thesis adopts a simple random mobility model, which is often used in the performance evaluation of wireless systems. The model does not reflect a real situation representative of actual system dynamics but rather aims at creating a stressful environment where the viability of the proposed schemes can be tested.

The model assumes that mobile terminals move independently of each other. They move in a random Brownian movement and make stops of varying duration. In the course of their movement, some of the UEs can gradually start drifting away from the base station and some may get closer to it. In order for the mobiles that are moving away to keep their QoS levels, their transmission powers need to be ramped up. While this would be nothing more than the normal operation of power control, the result would be an increase of the aggregate interference level in the cell.

Two important effects can arise out of this situation. First, if it turns out that because of its movement a mobile will reach a power outage[8], it must not be accepted by call admission control in the first place. This is especially probable under loaded conditions when the coverage of the cell has shrunk because of the effect of cell breathing. When this happens, the mobile may end up located in a spot where it can neither connect to its home Node B, nor can it be handed over to another cell. Second, in managing the total cell resources, we also need to be concerned with the power shares we have allocated to each traffic class. If it turns out that because of its movement, a mobile will cause its traffic class to exceed its allocated power share, we do not accept this UE either.

[8] Power outages are situations when the communication between the UEs and the Node B is disrupted because the UE transmission signal is not strong enough to reach the base station antenna. Outages can occur either if the mobile is too far away from the Node B, or the load in the cell has decreased the cell coverage.

The function in Figure 3.3 below executes the mobility prediction part of the scheme. It generates a list of 10 possible future scenarios and uses the load estimation functions used elsewhere in the admission control scheme to evaluate the admissibility of each of those scenarios. If they are inadmissible in more than 2 out of those 10 cases, the connection request is refused. If the request is admissible, the function checks if in any of the scenarios, accepting the connection would cause a violation in the stipulated proportions between the classes. If the probability of this is low (if it would happen in up to 2 out of our 10 cases) then the connection is safe to accept. If the probability is higher, then the RAB is blocked. The admissibility information is passed to the *uplink_capacity_compute* function which, taking into account information from the function *current_UL_load_estimate*, makes a decision whether the uplink capacity is sufficient or not.

```
future_UL_load_estimate
        I. Calculate 10 random possible locations for all mobiles in time t = 1, 2, 3, ..... n [sec]
            X_UE_coordinate = rand(i)
            Y_UE_coordinate = rand(j)
        II. Based on these locations, calculate new target $E_b/N_o$ values
            eb_no_target = f(R, BER, X_UE_coordinate, Y_UE_coordinate)
                    // The $E_b/N_o$ values are also a function of
// the requested data rate and the requested bit error rate
III. Given the values obtained in step II, calculate the required total capacity needed
            for (all active connections) do
              futire_UL_load += 1 / (1 + (uplink_proc_gain[conn] / (activity_factor[conn] *
future_eb_no[conn])));
           end for
        IV. Check if at any point of time the proportion between the classes would be violated,
        accounting for the new call as well.
       //This is accomplished only after executing
       function uplink_capacity_adjust listed below
    for (all QoS classes) do
        if uplink_capacity[qos] > allocated_share[qos] then
           accept RAB  //call function RAB_admit listed below
        end if
    end for
return future_UL_load
end future_UL_load_estimate
```

Figure 3.3: Mobility Prediction Logic

Lines 1-2 generate random coordinates and assign them to the mobiles to create a set of virtual locations for those mobiles. Based on those locations, line 3 calculates the corresponding E_b/N_o values which will need to be maintained for successful communication. The next three lines (4-6) calculate the required total capacity that would be needed to support transmission. Given the result of this calculation, lines 7 and 8 check if at any point of time the proportion between the classes would be violated, while accounting for the new call as well. Line 9 calls *RAB_admit*, a routine which takes care of parameter and resource adjustment when a bearer is admitted.

It should be mentioned that the mobility prediction function effectively combines power sharing with mobility prediction. This is done by imposing another admission criterion. If in *k* steps the ratio between the classes would be violated, then do not accept the call. If the ratio would remain within limits, then it is safe to admit the connection.

3.3 QoS-Aware CAC and Power-sharing Structure

The power-sharing scheme assumes that there is a pool of resources that can be allocated to the different classes. This pool is the total power budget, measured in terms of the total interference created by the aggregate transmission powers of the UEs.

Each of the classes is guaranteed a certain minimal fraction of the total power budget. However, any class, provided that it is the sole user of the system, would be allowed to take the whole capacity. In case RABs from other classes appear, connections belonging to the first class would be dropped so that the bearers from the new class are admitted until this class's minimal budget is reached.

A ratio of the aggregate power levels of the different classes is maintained to distribute the power allocations. This thesis assumes that the following distribution exists and should be maintained between the class allocations:

$$\sum PWR_{CLASS\ 1} / \sum PWR_{CLASS\ 2} = \sigma_{1\_2}, \text{ and } \sum PWR_{CLASS\ 2} / \sum PWR_{CLASS\ 3} = \sigma_{2\_3}$$

Assume that σ_{1_2}=5/3 and σ_{2_3}=3/2. Thus, under high load conditions, class 1 will have 50% of the resources, class 2 will have 30%, and class 3 will have 20%. In the event of not having enough flows of a certain class, traffic of other classes can take over the unutilized capacity. For instance, when there is no class 1 traffic, class 1 would have 0% of the resources, class 2 may have 35%, and class 3 may have 65%. When enough class 1 flows re-appear, then the necessary number of class 2 or class 3 calls will be dropped to restore the minimal power allocations.

It should be noted that other schemes are possible as well. For example, connection renegotiation can be implemented instead of preemption. Thus, if there is not enough capacity to meet the requirements of a higher, oversubscribed class, this class may be relegated to use the QoS service conditions of a lower class. Another option is to buffer the requests and admit them according to the current power shares in the cell. However, if preemption is chosen, there are several options as well. A definite possibility is to implement a scheme such as the one presented here. In this way a certain degree of fairness will be provided at the expense of the possibility of preempting an oversubscriber at any point of time. As a modification to this scheme, another approach may be implemented in which oversubscribers can only be forcibly released if they belong to a lower QoS class than the class requesting access. Which of these schemes will be implemented depends on the specific customer contracts and service level agreements. To make the system more flexible, a network operator may wish to have all these flavors implemented with an option to switch between them on an as needed basis.

Figure 3.4 shows the RAB dropping scheme, which makes possible the forceful connection release (preemption) and the freeing of resources.

```
Admission_control
1.  for every RAB SETUP request do
2.    I. Calculate UL Capacity
              //call function uplink_capacity_compute listed below
3.    II. Calculate DL Capacity
              //have to check downlink as well since each UL RAB adds overhead on the DL
              // If there is enough capacity, grant the request.
4.     if (uplink and downlink capacity sufficient) then
5.  Admit RAB and allocate resources  //call function RAB_admit listed below
      else     // Capacity insufficient either on uplink or downlink or both
              // Try to pre-empt other requests...
6.       for each QoS classes do
         //starting from the lowest priority, skipping class of currently requesting RAB,
              //look for RABs that can be preempted
7.          if (currently rejected class underutilized)
8.               AND (any of the other classes overutilized)
9.                  while (NOT done_preempting) do
10.               I. Recompute uplink and downlink capacity
                    // call function uplink_capacity_compute listed below
11.               II. Account for the new request and subtract the
                    capacity from the RABs that will be pre-empted
12.                 if (uplink and downlink capacity sufficient) then
13.                   1. Admit RAB and allocate resources  //call function RAB_admit (below)
14.                   2. Exit while loop
                      else
15.                       // Keep searching for RABs to pre-empt as
    // the ones currently found are not enough to accommodate new call
            end while
      if (done_preempting) then
16.      for (all QoS classes) do
17.         I. Send RAB RELEASE messages to all users that need to be preempted
18.           II. Perform housekeeping - cancel timers, update preemption buffers, etc.
         end for
      end if
19.     else  //Pre-emption was not successful because not enough RABs were found.
20.           1. Discard new request
21.     2. Send RAB ASSIGNMENT FAILURE message to SGSN
end else
end for
end Admission_control
```

Figure 3.4: Preemption logic in RNC

For every connection request (line 1), this routine computes the uplink and downlink capacity (lines 2-3). Downlink capacity needs to be computed even for uplink-only bearers because uplink connections add overhead on the downlink as well. If there is enough uplink and downlink capacity (line 4), the connection is admitted (line 5). If the available capacity is not enough to support the addition of the new call, a loop is created (line 6) to go through the already admitted calls and find a RAB to be pre-empted. The loop starts from the lowest priority RABs and skips the class of the RAB currently requesting access. Lines 7-9 make ensure that we only enter the loop if the requesting class is underutilized and an over-utilized class will be preempted. Lines 10-11 re-compute the uplink and downlink capacity, accounting for the new request and subtracting the capacity from the RABs that will be pre-empted. If a preemption-suitable RAB is found (line 12), the requesting bearer is admitted and the lower-priority connection is dropped (line 13). The loop is exited in line 14. In case no suitable connection is found, the whole procedure is repeated with another traffic class (line 15). In case the pre-emption succeeded, lines 16-18 send RAB Release messages to all RAB(s) to be pre-empted and perform some low-level system tasks like canceling timers, updating preemption buffers, etc. In case no pre-emption-suitable RAB was found, the new request is discarded and the corresponding signaling messages are sent (lines 19-21).

The code fragment in Figure 3.5 outlines the steps necessary for the admission of a new radio access bearer.

```
RAB_admit

1.  Adjust total used bandwidth to reflect newly admitted RAB
2.  Adjust used aggregate bandwidth per QoS class
      // call function uplink_capacity_adjust listed below
3.  Update RAB list in case pre-emption is used
4.  Send Radio Link Addition Request message to Node B
5.  Send RAB Setup message to UE.
  end RAB_admit
```

Figure 3.5: RAB acceptance functionality

Lines 1-2 adjust two system parameters: the total used bandwidth and the bandwidth user per QoS class. Then, a system-wide RAB list is updated in case the admitted RAB will need to be pre-empted later on (line 3) and a message is sent to the Node B to request radio link establishment (line 4). Finally, a RAB Setup message is sent to the mobile terminal to notify it about the successful establishment of the call.

The function in Figure 3.6 calculates the load increment that each new connection would bring into the system if it were actually admitted.

```
uplink_capacity_compute

Compute the uplink processing gain based on requested data rate
Compute the Eb/No target from the requested BER
Compute load increment as done in Equation (2.5)
Calculate current system load
    //call function current_UL_load_estimate
Calculate future system load
    //call function future_UL_load_estimate
Add load increment to current system load
Make admission decision based on the past and the present
    if (total_uplink_capacity < uplink_loading_factor) AND
    (future_total_uplink_capacity < uplink_loading_factor)
    Set uplink_capacity to SUFFICIENT
  else
    Set uplink_capacity to INSUFFICIENT
end uplink_capacity_compute
```

Figure 3.6: Algorithm for computing uplink capacity

First, the uplink processing gain is obtained by dividing the total WCDMA bandwidth (3.84 Mcps) by the requested data rate (line 1). Line 2 computes the E_b/N_o target, which is a function of the requested data rate and the requested bit error rate. Then, line 3 computes the load increment by implementing Equation (2.5). Lines 4-5 compute the current and the future system load, and are respectively based on the current parameters and mobility prediction. In line 6, the load increment is added to the current system load. Then the admission decision is made, based on both the current and the future system load (lines 7-10).

Figure 3.7 below shows the function, which takes care of adjusting the uplink capacity depending on whether a request has just been granted or a radio bearer has been released.

```
uplink_capacity_adjust
1.   if (Request Granted) then
2.        I. Increment the total uplink load
3.        II. Increment uplink load per QoS class
4.   else          //Request released or pre-empted
5.     I. Decrement the total uplink load
6.     II. Decrement uplink load per QoS class
   end if
end uplink_capacity_adjust
```

Figure 3.7: Capacity Adjustment Function

If a RAB assignment request has been granted (line 1), the total uplink load, as well as the load per QoS class is incremented (lines 2-3). If a connection has been released or pre-empted, the total uplink load and the load per QoS class are decremented.

Chapter 4

Performance Evaluation

In this chapter, a performance study of the proposed framework is conducted. The results are analyzed and compared to the performance of the popular proposal by Holma and Toskala [13]. The proposal in [13] is based on the 3GPP recommendations and provides a thorough, albeit somewhat simplistic, "feature-stripped" solution. This base version shall henceforth be referred to as the Holma and Toskala Call Admission Control, or "HTCAC", and the currently proposed scheme – as the Optimized CAC, or "OCAC."

This chapter is divided into several parts dedicated to the performance improvements that each of the three components creates on its own. In this way it is clearly shown what the specific gains from each module are. Then, an all-inclusive result section illustrates the interworking of the separate components and the additional functionality obtained by their combination.

Section 4.1 gives an overview of the simulation model used, and the corresponding scenario parameters. The obtained results are shown and investigated in the next section. A summary of the results is provided in Section 4.4.

4.1. Model Description and Simulation Setup

OPNET Modeler 9.1 has been used[9] as the simulation environment in this study. A specialized open source UMTS model designed in accordance to the 3GPP requirements and jointly developed by OPNET Technologies, Inc. and Telcordia has been used and extended.

A single-cell scenario has been created in which a number of mobile stations send high rate uplink traffic to an Ethernet-based server. Each UE transmits data streams (RABs) belonging to all three traffic classes. Thus there are three types of users with respect to their data rates – 32 Kbps users, 64 Kbps users and 128 Kbps users. The total capacity requested by all RABs is designed in such a way that most of the time the requested resources will be around 20% more than what is actually available. In this way there will often be rejected traffic, as there will often be the need to apply the mechanisms which guarantee the precise class allocations. While it is true that most systems would not normally work under such conditions, it is not extremely uncommon for some of the cells to become overloaded in a similar way. To increase the request/rejection dynamics, connection durations are short to medium so that more RABs get to request establishment. At the same time, inter-request times vary to avoid uniformity and patterns in behavior. The UEs move in a random manner with varying speeds, and stop in one place for varying time intervals (dwelling time) before picking a new direction.

The following table summarizes the simulation parameters used.

[9] An extensive overview of OPNET Modeler is given in Appendix B.

Simulation Parameter	Value	Unit
Deployment scheme	Single cell, single Node B	
Cell radius	1500	meters
User speed	0 – 120	Km/h
User dwelling time	3 – 20	sec
Inter-request time	0.5 – 1.5 (uniform distrib.)	sec
Upload file size	4, 8, 16	Kbytes
User bit rates	32, 64, 128	Kbps
Uplink RLC mode[10]	Unacknowledged	
Transmission Time Interval[11]	20	ms
Activity factor for all RABs	1	

Table 4.1: Simulation Parameters Used

Two rounds of simulations were run to obtain the results presented in this chapter – one with the base version, and one with the extended version. Naturally, all parameters and settings were kept the same.

For the set of simulations belonging to the power prediction section, a special mobility factor has been introduced. This allows us to control the mobility characteristics of the different users. Two parameters are controlled by the mobility factor: the movement speed and the movement trajectory. A user with a low mobility factor will move slowly and will stay relatively close to the Node B. A user with a high mobility factor will move at a higher rate of speed and will often go near and even beyond the borders of the cell. Users move in a random fashion, regardless of their mobility factor.

[10] Unacknowledged RLC means that the Radio Link Control (RLC) layer does not handle flow-control issues such as data retransmission or duplicate detection and leaves this responsibility to the TCP layer.

[11] The Transmission Time Interval or TTI is an important concept in the operation of the radio interface. The TTI indicates the inter-arrival time of data blocks from the MAC layer to the physical layer, and is always a multiple of the length of one radio frame (10 ms). The default TTI value is 20 ms.

A snapshot of the topology of the simulated network is presented in Figure 4.1. The figure depicts a single hexagonal cell with a number of UEs (enough to use up all available capacity), a Node B, an RNC, a core network node (SGSN), and an Ethernet-based server. The scenario features WCDMA links between the UEs and the Node B and ATM links between the Node B and the RNC, as well as between the RNC and the SGSN. IP over ATM is employed for these links, which is in accordance with the 3GPP technical specifications. Standard 100 Mbps Ethernet connects the UMTS gateway with the external network. IPv4 QoS headers are used to distinguish between the different QoS classes and preserve the UMTS class separation. The trajectories of the mobile nodes are not shown.

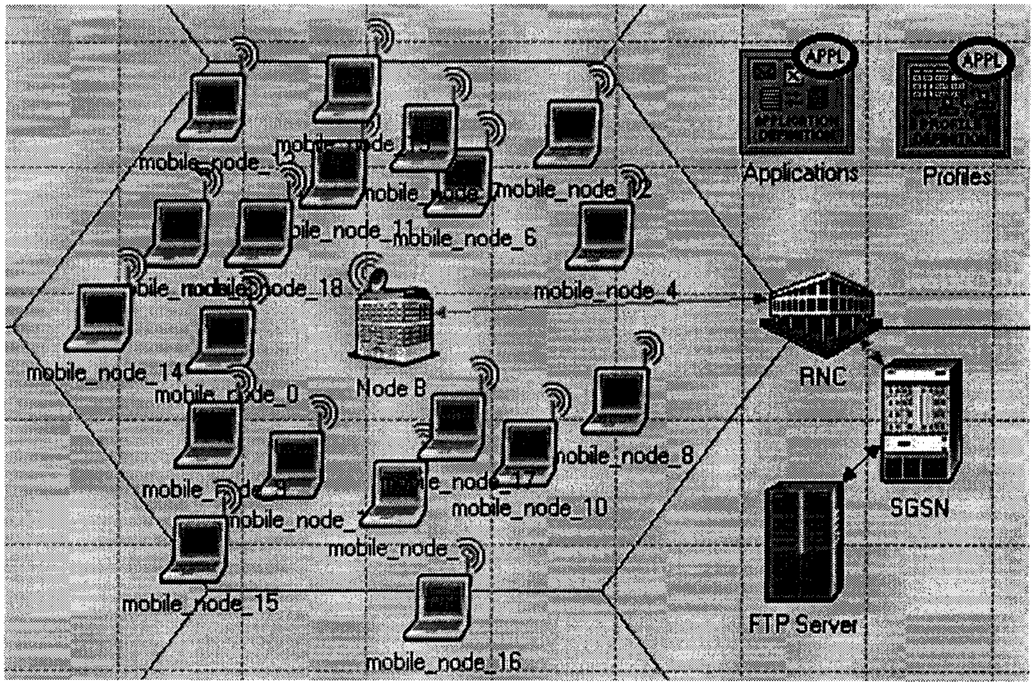


Figure 4.1: Topology Snapshot

4.1. MBAC Results

Three parameters will be of particular interest to us when discussing admission control efficiency: throughput, blocking and dropping.

We define throughput as the amount of uplink traffic that is successfully transmitted in the air interface for a given time. The blocking ratio is the ratio of rejected connections to the total number of connections, which have requested establishment, and the dropping ratio is the result of the number of dropped connections divided by the total number of active connections.

$$\text{Blocking ratio} = \frac{\text{Total number of rejected connections}}{\text{Total number of connections that have requested access}}$$

$$\text{Dropping ratio} = \frac{\text{Total number of dropped connections}}{\text{Total number of admitted connections}}$$

Before we introduce the numerical results, we demonstrate the operation of the moving load average. As described in Chapter 3, the aggregate E_b/N_o values of all active connections are observed and a current average is dynamically generated. Then this average is used by call admission control to calculate the actual system load.

Figure 3.2 depicted the actual E_b/N_o graph of a single RAB. It was observed that the nature of the E_b/N_o curve is quite jittery. The same is true for the aggregate E_b/N_o curve, which is a composite curve obtained by adding together the E_b/N_o curves of all active RABs. The aggregate E_b/N_o curve is presented in figure 4.2 below.

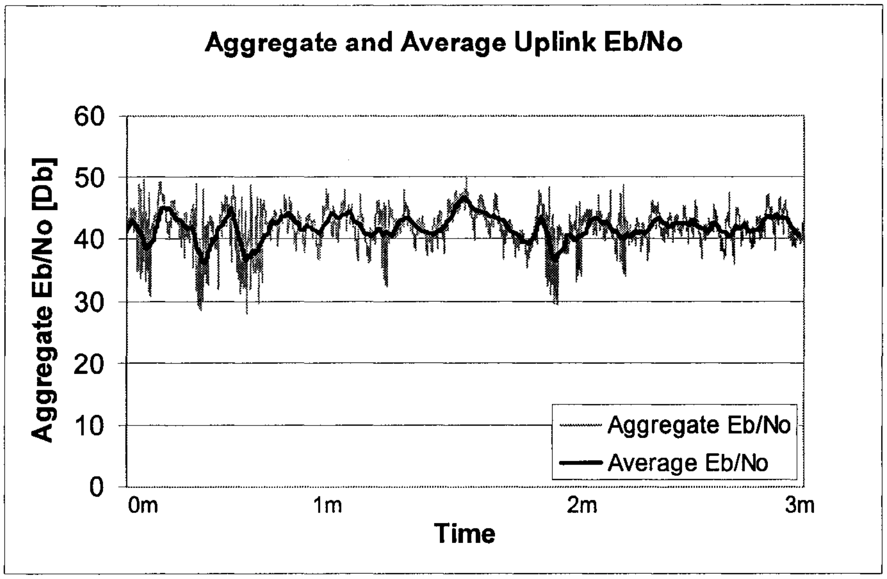


Figure 4.2: Aggregate and Average Uplink Eb/No Value Curves

It can be seen from the figure that the aggregate E_b/N_o values can vary from around 30 to 50 db within a matter of seconds. This is a significant deviation, which leads to inefficiencies. The figure also shows a second curve representing the average E_b/N_o. We claim that using this curve gives a more accurate measurement of the actual system load. Within the framework of this study, HTCAC uses the Aggregate E_b/N_o curve, and OCAC uses the Average E_b/N_o curve.

The next figure plots the corresponding system throughput for the same three-minute period. It should be noted that there is no direct (or visible) correspondence between the E_b/N_o value and the effective throughput because multiple E_b/N_o values can correspond to the same data rate depending on the spatial location of the user and the signal quality requirements of the connection. The figure shows the uplink throughput for both HTCAC and OCAC.

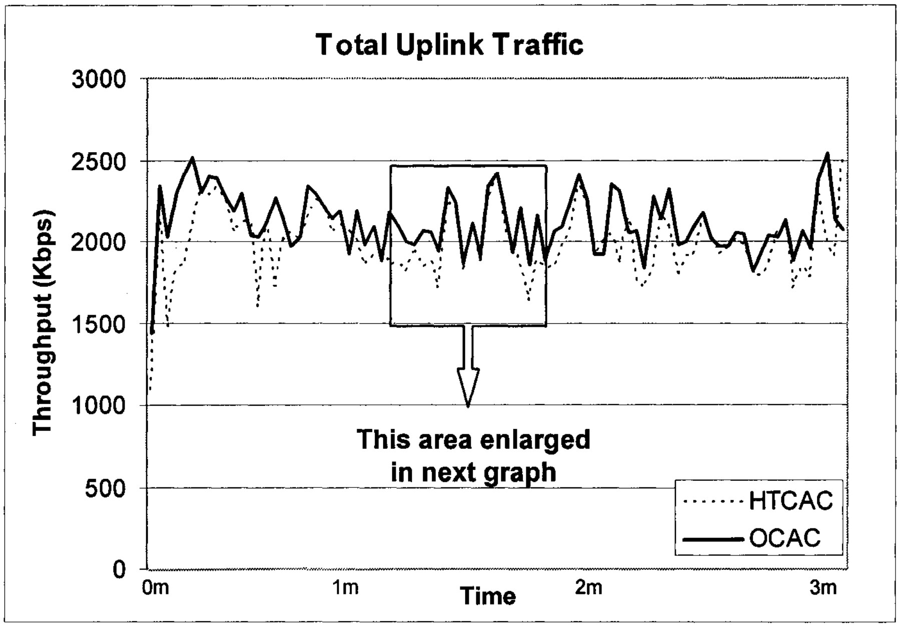


Figure 4.3: Total Uplink Traffic using HTCAC and OCAC

It should be noted that for the purposes of this analysis, time periods in the order of several minutes or more are too long to be of interest. This is because all events and behavioral patterns under consideration take place within several tens of seconds. For example, the E_b/N_o value can significantly change in the course of 10–15 seconds; a group of users traveling at 120 km/h can move comparatively far away from a Node B in less than 30 seconds; and switching an external interferer on or off would immediately change the propagation conditions of the medium.

For all these reasons, from now on, smaller time intervals will be examined. This will in addition reveal more detail than what can be observed in the above figure. Below is the enlarged version of the marked 30-second time interval.

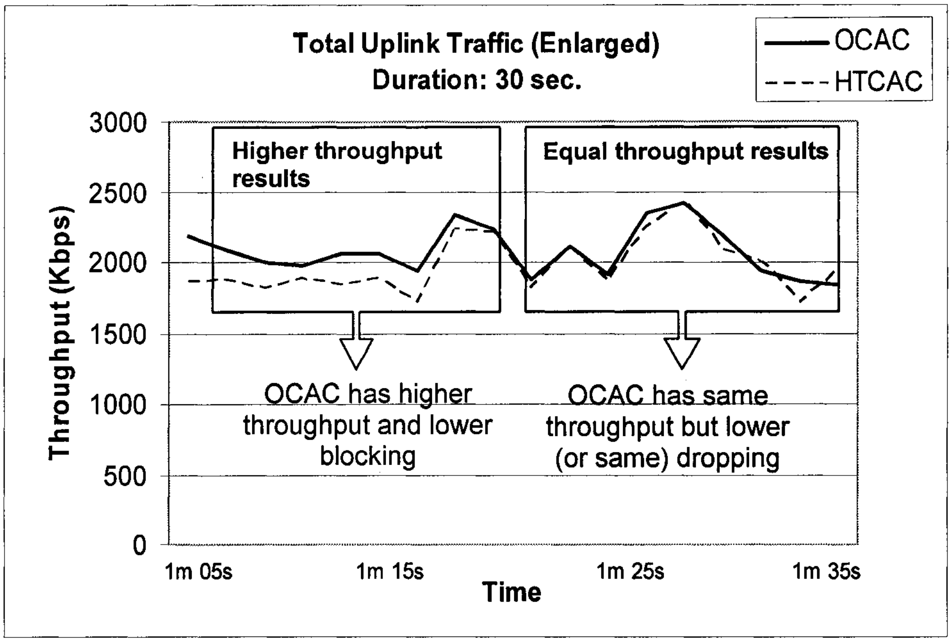


Figure 4.4: Total Uplink Traffic using HTCAC and OCAC (Enlarged)

For functional convenience, we will logically divide the above plot into two parts – left and right, as marked above. It is important to differentiate between these two parts, as they are manifestations of two different types of behavior.

4.1.1 Higher Throughput Results

A. Throughput

One of HTCAC's inefficiencies is clearly seen in the left part of the graph. This is where taking the instantaneous parameters of the system has falsely led HTCAC to believe that there are less resources than there really are. As shown in the graph, this leads to a smaller throughput than can actually be supported by the system. Thus the first gain achieved by OCAC is an increased data throughput.

B. Blocking

Because HTCAC underestimates the available capacity, it wrongfully prevents a number of RABs from being established. Thus, as shown in Figure 4.5, the blocking ratio in HTCAC is higher than the blocking ratio in OCAC.

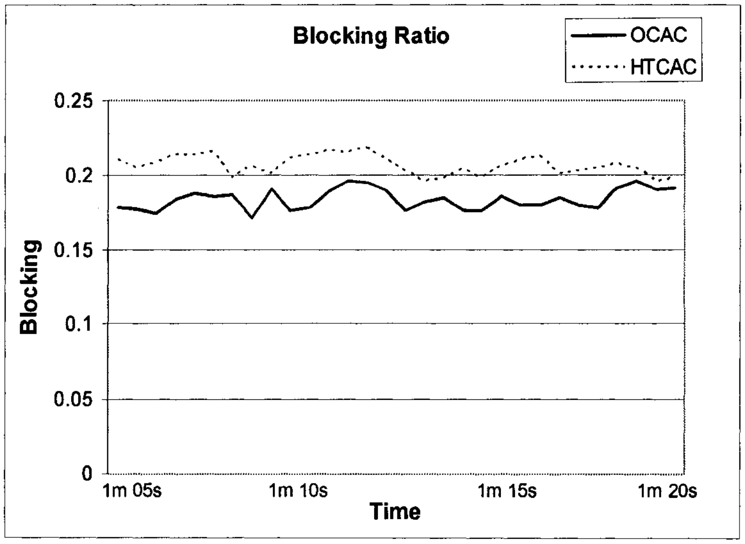


Figure 4.5: Blocking Ratio for HTCAC and OCAC (Higher Throughput Results)

C. Dropping

Figure 4.6 shows that in terms of dropping, in the left part of the graph, OCAC and HTCAC exhibit very similar behavior. A dropping rate close to zero means that although OCAC has allowed more traffic, it has not overestimated the available resources.

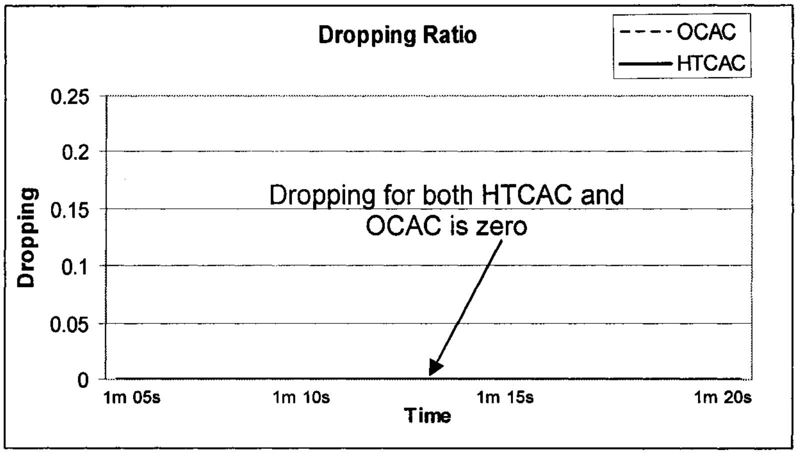


Figure 4.6: Dropping Ratio for HTCAC and OCAC (Higher Throughput Results)

4.1.2 Equal Throughput Results

The right part of the plot suggests that after the middle of the observed interval (at around 1 minute 20 seconds), HTCAC starts achieving the same efficiency as OCAC. While this is certainly so in terms of throughput, the situation is different when blocking and dropping are considered.

A. Throughput

The effective throughput measured in the right part of the graph is similar for both HTCAC and OCAC. There is no gain as far as throughput is concerned.

B. Blocking

In this particular example, blocking for HTCAC and OCAC is the same. Two rounds of simulations with different random seed values were run so that the two curves would not coincide.

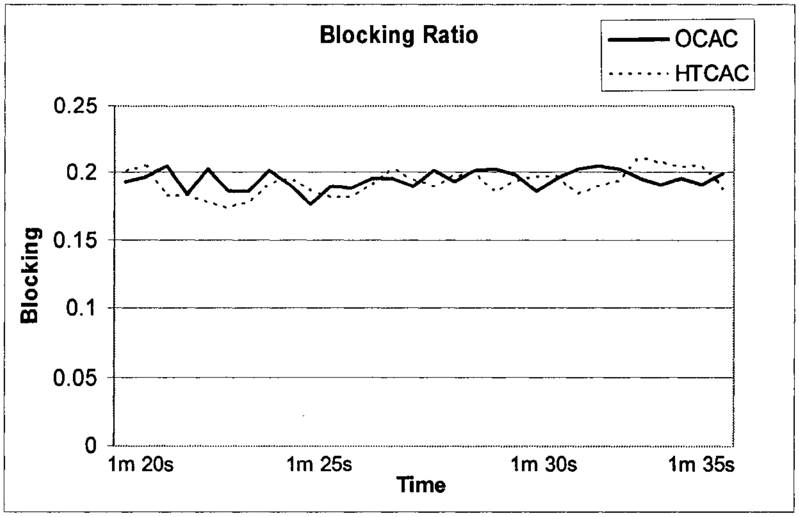


Figure 4.7: Blocking Ratio for HTCAC and OCAC (Equal Throughput Results)

It has to be noted that despite the similar throughput, blocking for HTCAC could sometimes be lower. This is because HTCAC may overestimate the amount of available resources and accept more connections. This, however, comes at the expense of increased dropping as described below.

C. Dropping

There are two possible scenarios for the dropping rate in the right part of the graph. This depends on whether HTCAC has been "lucky" or not. First, HTCAC could have overestimated the available resources, thus optimistically admitting more traffic than the system has resources for. In this case, the effective throughput is the same as in OCAC and blocking is lower. However, this comes at the expense of having to drop some of the admitted connections. Thus the HTCAC dropping rate would be higher than the OCAC dropping rate.

A second possibility is for the instantaneous load value used by HTCAC not to be an overestimation but to be close or equal to the actual system load. In this case HTCAC simply manages to "guess" the real system load. Therefore the dropping rate is the same with both schemes.

In this particular case, HTCAC manages to guess right and happens to use a sequence of values that are close to the actual air interface load. Thus the HTCAC dropping rate is commensurate with the dropping rate of OCAC. As in the previous section, we find both of them to be again equal to zero.

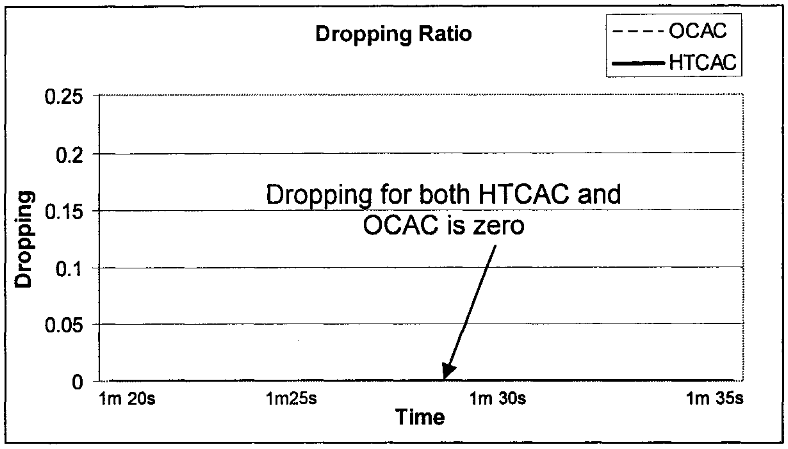


Figure 4.8: Dropping Ratio for HTCAC and OCAC (Equal Throughput Results)

The next set of graphs presents an alternative case where HTCAC does not manage to guess the actual amount of used resources but overestimates it instead. A 1-minute interval from another simulation run is presented, and a 30-second fragment is again closely examined.

Figure 4.9 below shows the throughput achieved by the two schemes. As in the previous case, we can observe that on average OCAC allows a higher throughput although at times the effective data rates achieved by the two approaches are identical.

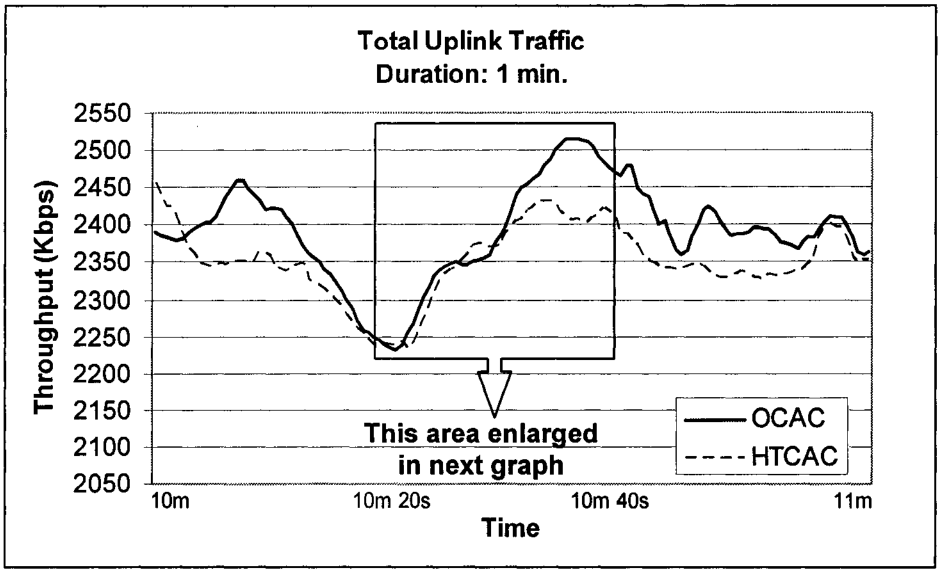


Figure 4.9: Total Uplink Traffic using HTCAC and OCAC

We shall again divide the graph into two parts in order to separate the two types of behaviors. Figure 4.10 illustrates this separation in an enlarged graph spanning the 30-second interval marked above.

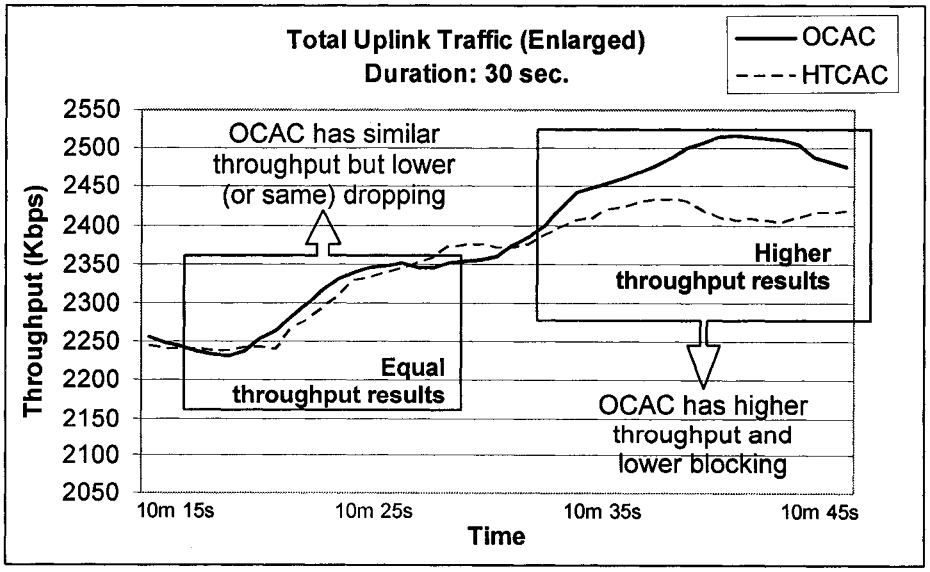


Figure 4.10: Total Uplink Traffic using HTCAC and OCAC (Enlarged)

4.1.3. Equal Throughput Results

A. Throughput

It is clear from the left part of the graph that for the first 15 seconds, throughput is the same (or similar) for both HTCAC and OCAC. As in the previous example, this could be the result of HTCAC either underestimating the actually used resources (thus overestimating the amount of available capacity) or just happening to guess right. Here however, HTCAC is not as fortunate. This fact can be established by examining the blocking and dropping rates below.

B. Blocking

Figure 4.11 compares the dropping rates of HTCAC and OCAC for the interval of interest. It is clear from the graph that HTCAC has a lower blocking rate. This means that more connections are accepted by the HTCAC mechanism. Nonetheless, the same amount of traffic is actually transmitted in the air interface (as seen in figure 4.10). This is due to the high price HTCAC's optimistic admission policy carries – an increased dropping rate.

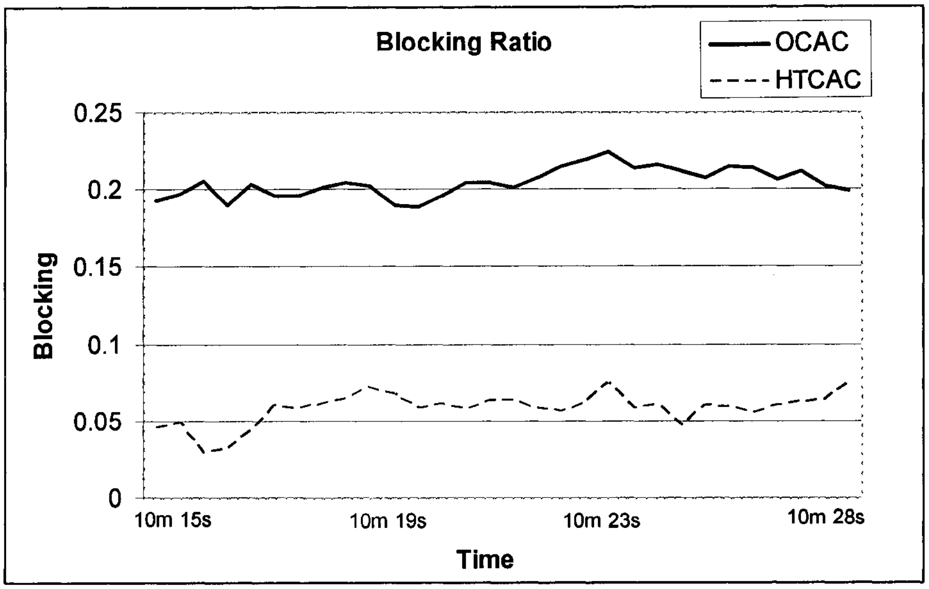


Figure 4.11: Blocking Ratio for HTCAC and OCAC (Equal Throughput Results)

C. Dropping

Figure 4.12 shows the dropping rates of the two schemes. Because HTCAC admits more RABs than there is capacity for, some of these RABs will be dropped. Thus HTCAC exhibits a higher dropping rate, as shown in the picture.

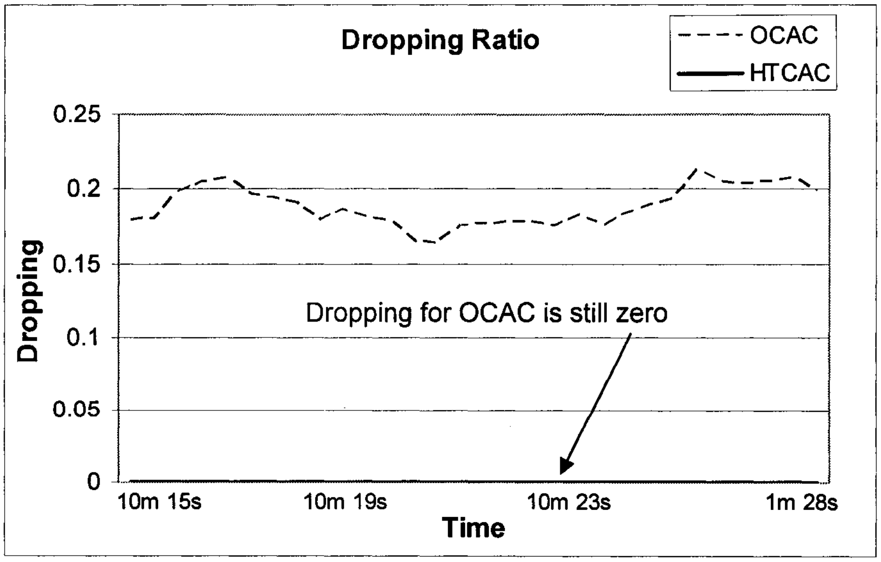


Figure 4.12: Dropping Ratio for HTCAC and OCAC (Equal Throughput Results)

As mentioned earlier, it is more important to have lower dropping than lower blocking since from the user perspective it is more undesirable to lose an ongoing call or browsing session than not to be able to establish one in the first place.

4.1.4 Higher Throughput Results

This example has already served its purpose in illustrating the case where HTCAC overestimates the actual system load. This behavior is exemplified in the already examined left part of Figure 4.10. However, for the sake of completeness, analysis for the right part is presented below.

A. Throughput

For the second part of the graph, the actual throughput allowed by OCAC is higher than the throughput allowed by HTCAC.

B. Blocking:

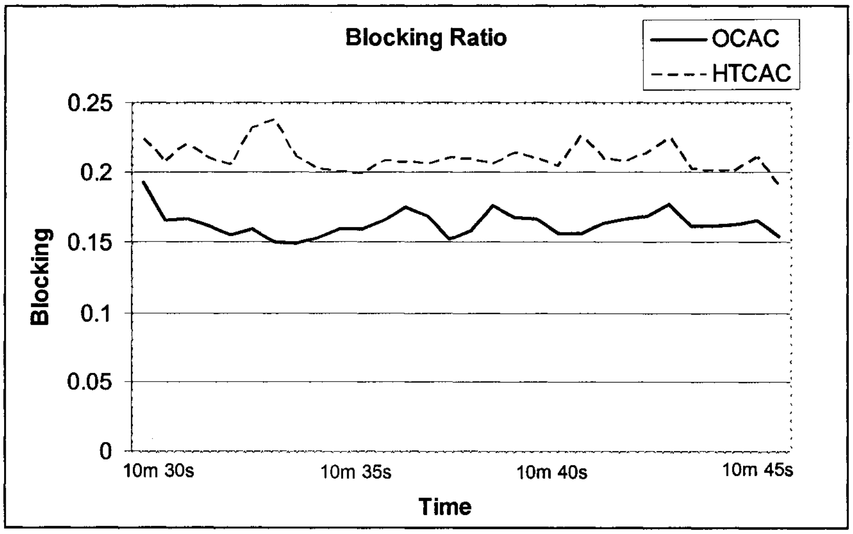


Figure 4.13: Blocking Ratio for HTCAC and OCAC (Higher Throughput Results)

OCAC has a lower blocking rate as it accepts more traffic than HTCAC. This is shown in figure 4.13.

C. Dropping:

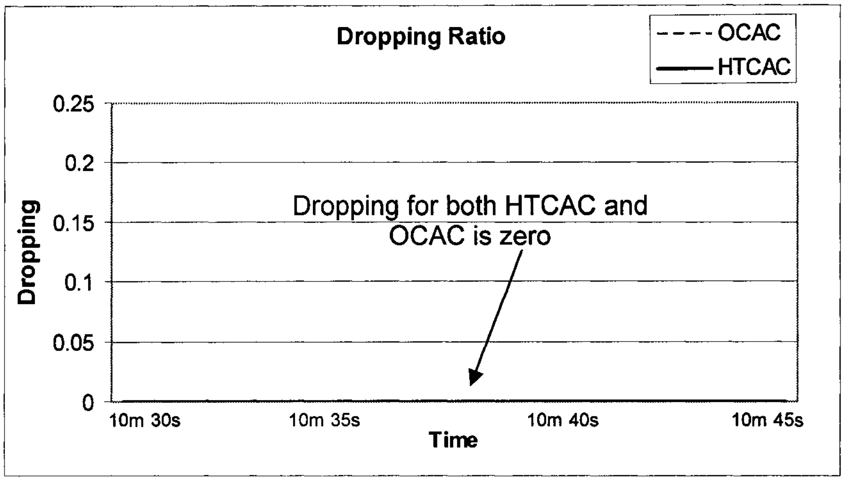


Figure 4.14: Dropping Ratio for HTCAC and OCAC (Higher Throughput Results)

Dropping is zero for both schemes. This is shown in Figure 4.14 above.

4.2. Power Prediction Results

The power prediction (or mobility prediction) functionality has been described in preceding chapters. In summary, what it strives to accomplish is to identify connections that are likely to either cause a power outage or a violation of the per-class allocations. This depends on the current load, the requested capacity, as well as the user mobility. Generally, the more resources a connection requests, or the more mobile a user is, the higher the probability that this connection would fail due to power-related problems. Therefore, power prediction should reject the request in the first place.

This may have two possible effects. First, rejecting connections that would otherwise be accepted could increase the blocking probability. However, as this would result in a decreased dropping rate and it is more desirable to have lower dropping than lower blocking, the higher blocking may well be justified and there would be an improvement. Second, if the capacity that would have been allocated to the connections which are rejected by power prediction is granted instead to other non-problematic connections, both the blocking rate and the throughput would remain the same. The dropping rate would be lower and hence constitutes an improvement over the original version as well. Which of the two alternatives would be the case depends on the number of users requesting access, as well as the overall mobility patterns in the cell. For example, if there are just enough users to take up the capacity of the cell (or less than that), chances are that there would be an increase of the blocking rate. Alternatively, if the cell's resources are grossly insufficient for all the users requesting access, the blocking rate would remain the same, as the capacity not given to some connections would be given to others.

Results from two scenarios are presented in this section: one scenario where mobility and power prediction is disabled and another where it is enabled. Figure 4.15 plots the blocking rates for both scenarios over an interval of 30 seconds. Clearly, this represents the first case where blocking is higher.

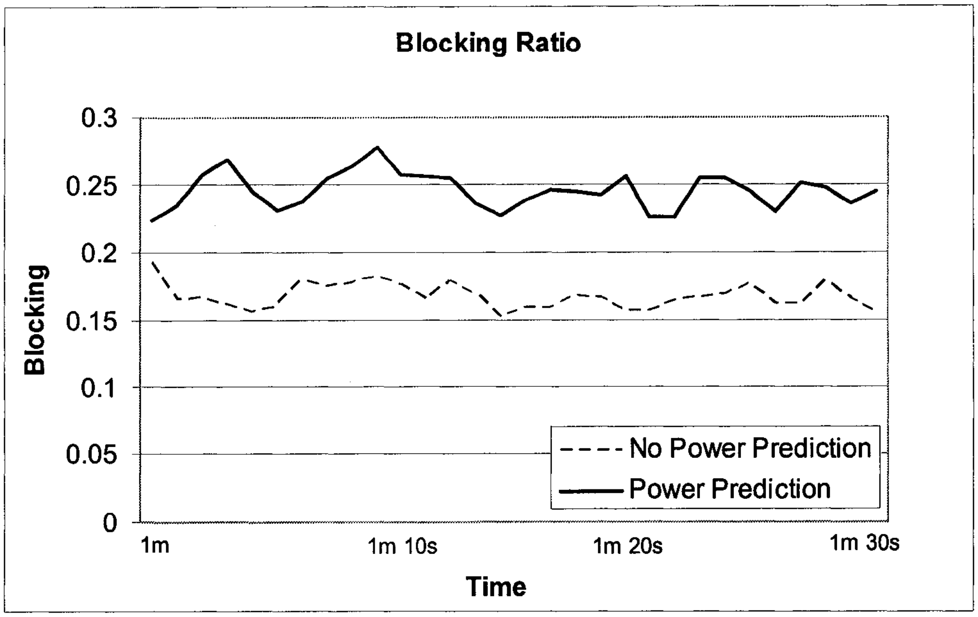


Figure 4.15: Blocking Ratio with and without Power Prediction

Figure 4.16 below shows the corresponding dropping rates of the same scenarios for the same time interval. The results have been generated by randomly assigning a different mobility factor to each connection.

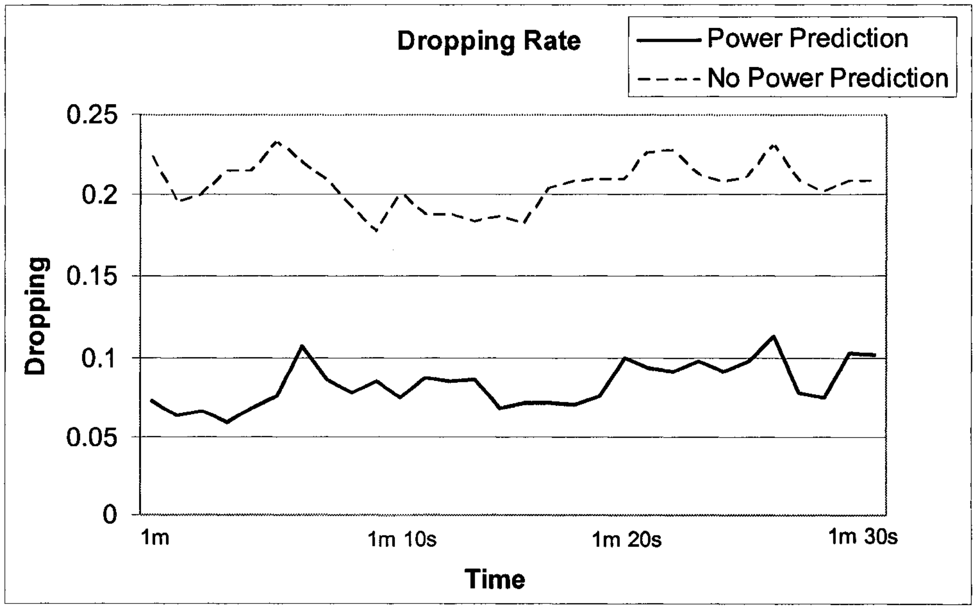


Figure 4.16: Dropping Ratio with and without Power Prediction

As expected, the dropping rate with mobility and power prediction is significantly lower than the case without mobility and power prediction. This shows that despite its simplicity, the prediction scheme improves the network's efficiency. Naturally, if the mobility factor mix were modified (for example to comply with the actual mobility patterns in the network), the specific values would differ. However, what is important is the actual improvement of the case with mobility prediction over the case with no mobility prediction.

To have a better idea about the behavior of the scheme, the figure below shows the effective throughput of the air interface in the two scenarios. It can be seen that in terms of throughput the two schemes have similar performance.

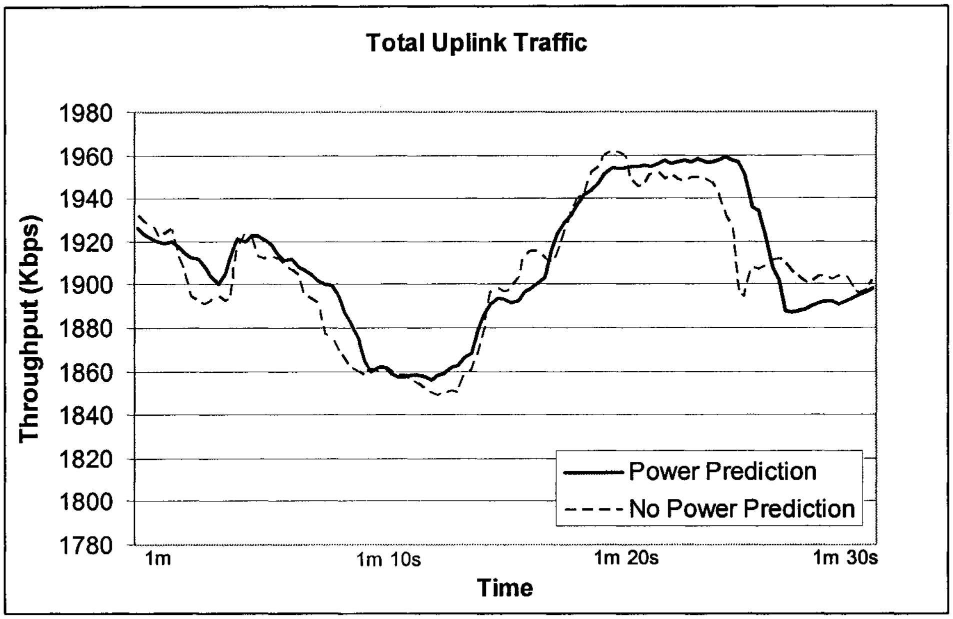


Figure 4.17: Total Uplink Traffic with and without Power Prediction

It should be noted here that the above throughput results are for real-time data. It is assumed that upon connection termination (dropping), all already transmitted real time information is lost. With this provision in mind, it could be stated that the throughput readings for both cases are comparable. Therefore, with everything else being equal, the power prediction call admission control scheme is superior to the non-enhanced scheme in that it has better dropping characteristics.

If it is assumed, however, that this is a non-real-time transmission and the receiving end has a means of buffering the incoming information (which is most likely not true if the receiver is a UE as well) then the actual throughput would be higher when no prediction is employed. Thus a conclusion can be drawn that the particular scheme to be employed largely depends on the type of traffic we are dealing with. A dynamic scheme adjustment would be beneficial in this case.

Another interesting fact to observe in the graph is that the overall throughput is noticeably lower than the throughput in the previous cases. This is because the UEs are more mobile and reach the periphery of the cell more often. The cell-breathing effect is in place. To better illustrate this point, the next graph shows the measured correlation between the dropping rate expected (predicted) by the mobility prediction mechanism and the mobility factor.

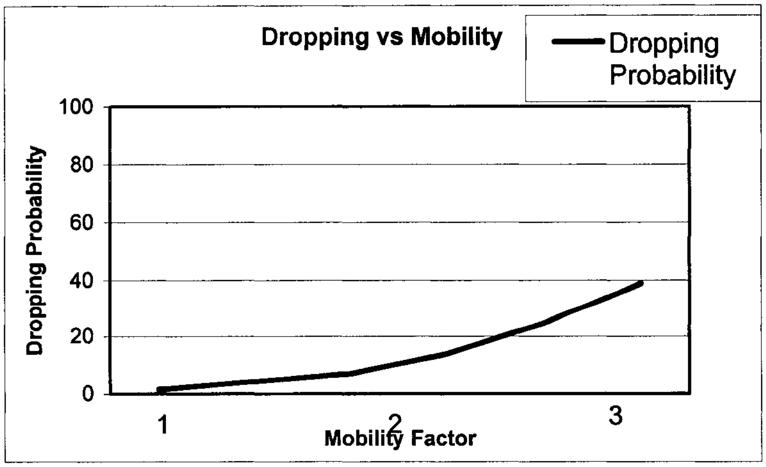


Figure 4.18: Dropping Probability versus Mobility Factor

The graph above attests to the fact that power outages or violations of the per-class allocations are more likely to occur to fast-moving, high data rate users than for slow, low-bandwidth connections. Both power outages and violations of the per-class allocations can happen to any mobile that has moved far enough from the Node B but are particularly likely to happen if the considered connection is transmitting at a high data rate. This effect is a clear manifestation of the cell-breathing phenomenon. Normally, the network's first line of defense against this would be to hand over the user to another base station/cell. However, this may not always be possible due to insufficient resources in the other cell as well. Thus dropping is unavoidable and a forceful termination of the connection occurs.

4.3. QoS Enforcer

This section presents the results obtained by combining the power prediction routine with the admission control power sharing structure. We begin by examining the functionality of the power sharing structure itself.

4.3.1 Power Sharing Structure Functionality

Figure 4.19 shows a scenario, which illustrates the dynamics of the per-class allocations in the system. With per-class relative resource allocations controlled by σ_{1_2}=5/3 and σ_{2_3}=3/2. Minimum guaranteed resource allocations used are then: 50% to Class 1, 30% to Class 2, and 20% to Class 3. If at any moment of time, there is no traffic belonging to other traffic classes, any of the three classes is entitled to the whole capacity of the network. To better visualize the dynamics behind the capacity sharing and class allocations, a special scenario has been created where RABs belonging to different classes request admission at 30-second intervals from each other. The functionality of this scenario as far as power sharing is concerned is presented in Figure 4.18 below.

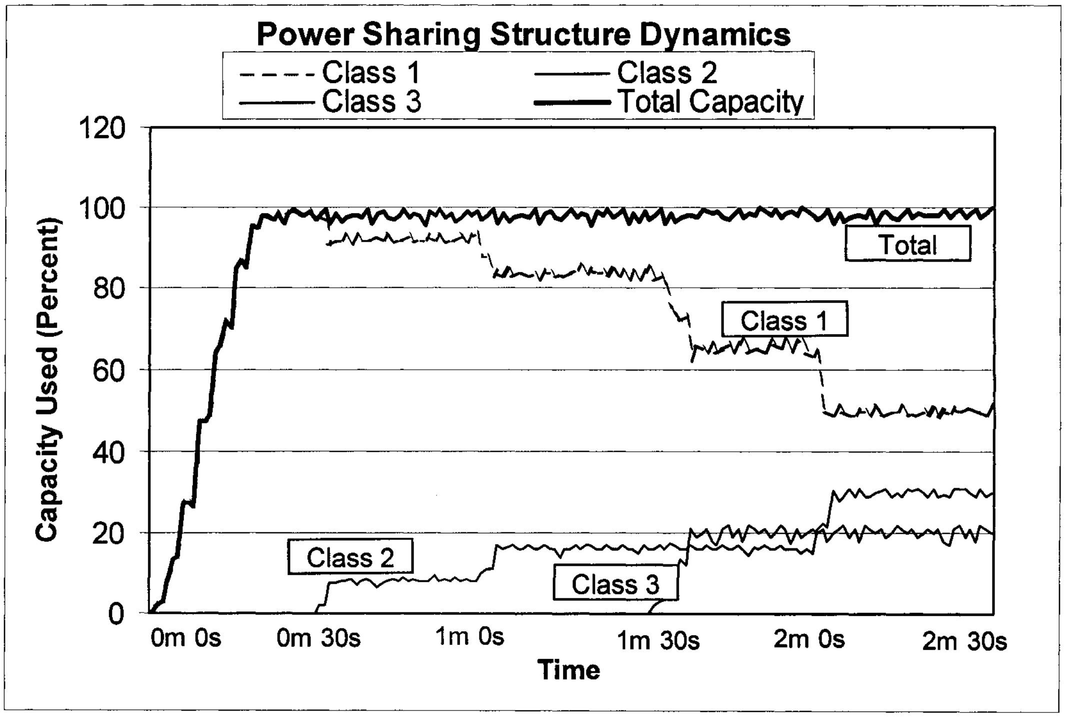


Figure 4.19: Power Sharing Operation

The plot shows that in the beginning the only traffic in this cell belongs to Class 1. Being the only user of the medium, it is allowed to consume all resources. Thus during the first 30 seconds the resource usage of Class 1 goes as high as 100%. However, when a Class 2 connection is admitted after 30 *sec*, some Class 1 bearers are dropped to accommodate the Class 2 RABs. This is because Class 1, although the highest priority class, is the only oversubscriber in the system. This functionality is shown in the interval covering the second half of the first simulated minute. The total capacity continues to be 100% but is no longer determined by connections belonging to one class only. It is now the sum of the resources used by Class 1 and Class 2 bearers. At the beginning of the second simulated minute another Class 2 connection appears and more Class 1 RABs are dropped. This is again due to Class 1 being the only oversubscriber despite its high priority. When Class 3 connections get active at time = 1 *min* 30 *sec*, more resources are taken away from Class 1 until the 50%, 30% and 20% allocations for Classes 1, 2 and 3, respectively, are attained.

4.3.2 Interworking of Power Prediction and Power Sharing

Power prediction aims at preventing violations of per class allocations. Using power sharing graphs like the ones in the preceding section can help visualize the functionality of power prediction and show the difference it makes for the system. Figure 4.20 below presents a special case with no power prediction employed. In the scenario considered here, Class 2 high data rate users move and stay away from the base station causing a disturbance in the per class allocations. To further illustrate the impact of the scenario, a group of high-data rate users with the highest mobility factor are used in the simulation.

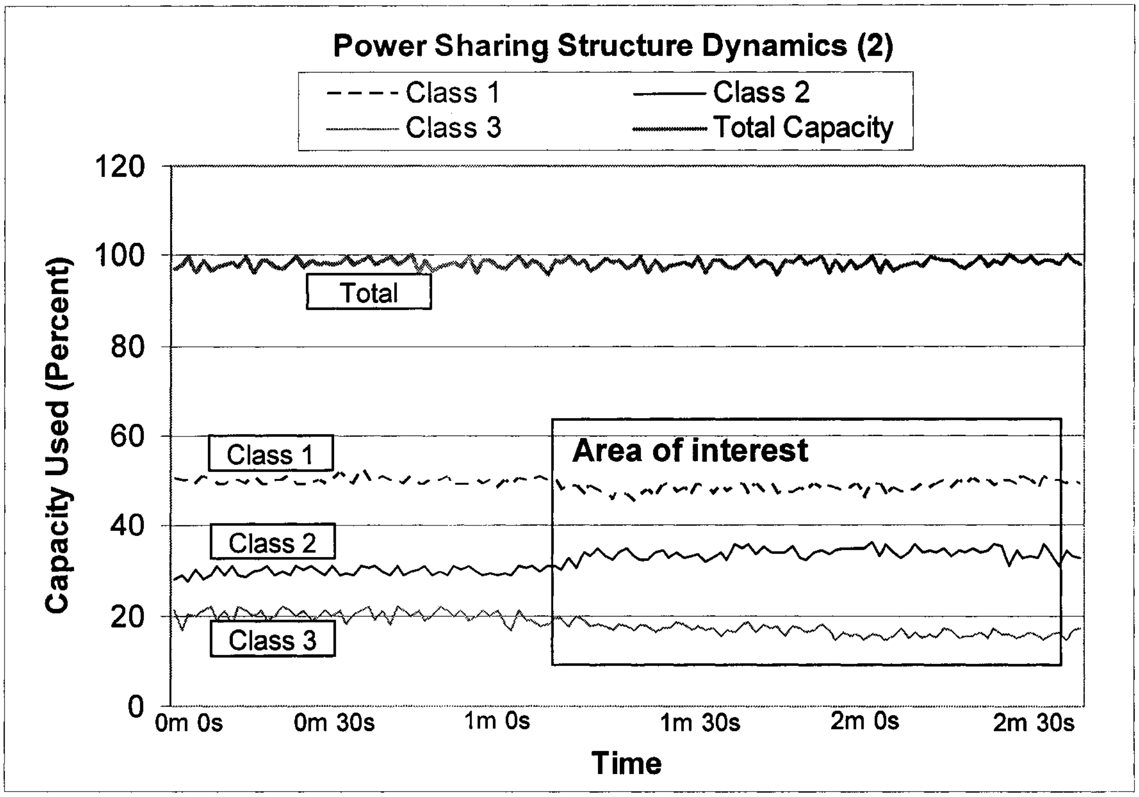


Figure 4.20: Power Sharing with no Prediction

Once a high data rate user starts moving away from the Node B, its power control mechanism ramps up its transmission power. Because of the settings used in the simulation, this begins to happen to a number of terminals at the beginning of the second simulation minute. The combined effect of multiple mobiles having to increase their transmission powers leads to a number of other UEs being "out-shouted." This is especially possible if these UEs are located close to the Class 2 mobiles or are near the borders of the cell. Figure 4.20 shows how such an effect occurs. Some Class 1 and Class 3 connections experience signal deterioration to such an extent that communication between them and the Node B can no longer be maintained. These connections are terminated as evidenced by the slightly lowered capacities for both Class 1 and Class 3. At the same time, an increase in the capacity allocated to Class 2 can be observed.

Although the scenario described in the preceding paragraph is a definite possibility, our simulations show that an alternative development – a power outage – is more likely to occur. In a power outage, it is the Class 2 mobiles that suffer the most from their increased power needs resulting from the large distances to the Node B. In another simulation, where outages, and not allocation violations, occur, Class 2 connection terminations are registered, leaving Class 1 or Class 3 bearers unaffected. The total capacity remains the same because the freed resources are swiftly allocated to other connections requesting establishment Thus the capacity graph looks exactly like the left part of Figure 4.20 (the interval 0 – 60 sec).

Figure 4.21 below shows the advantage of mobility and power prediction. When it is employed, outages and allocation violations are eliminated (or decreased).

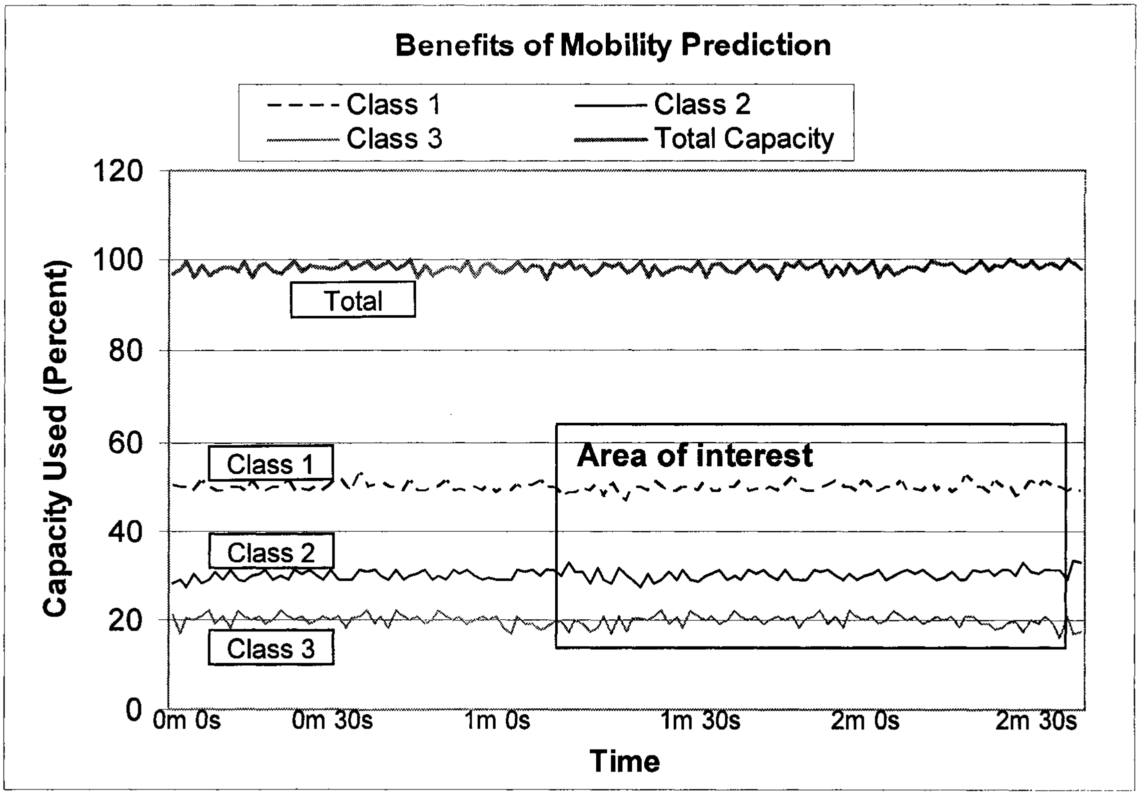


Figure 4.21: Effects of Using Mobility Prediction

Figure 4.22 shows the superimposed graph of the two plots presented above. It clearly outlines the differences between the case with mobility prediction and with no mobility prediction.

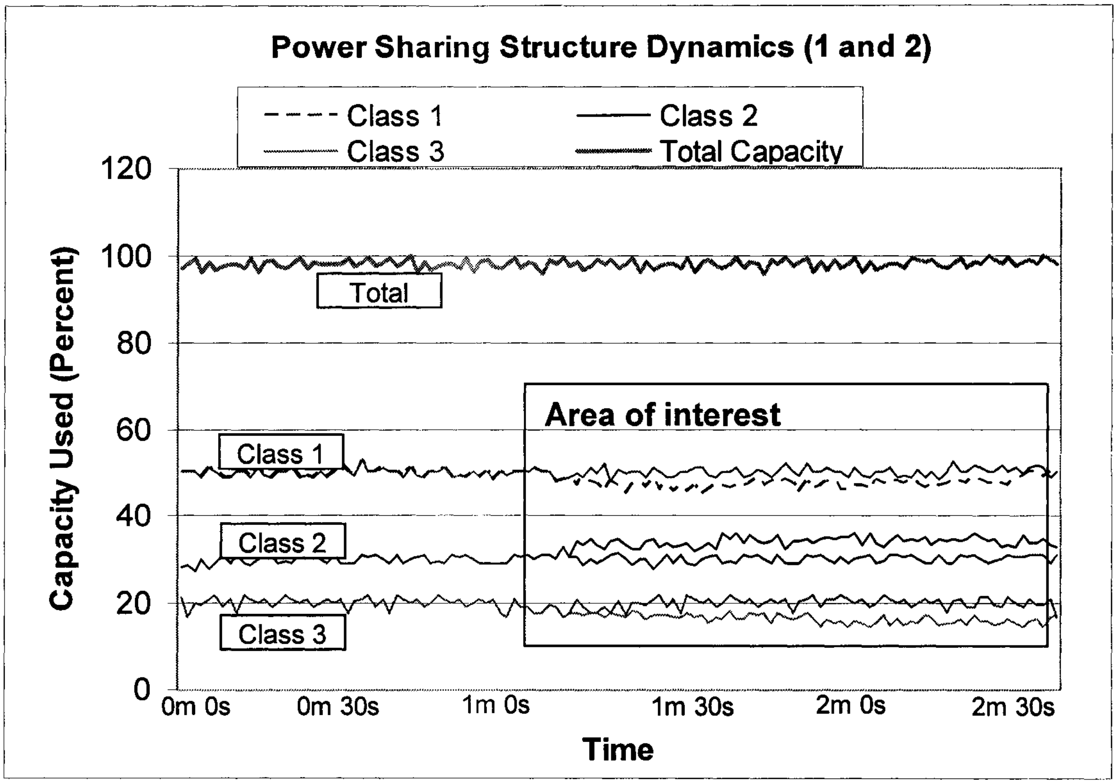


Figure 4.21: Power Sharing with and without Prediction

The differences between using and not using mobility prediction vary depending on scenario conditions like mobility and data rate. On average, the relation in Table 4.2 holds true for our simulations. The table illustrates the relationship between the mobility factor (introduced in Section 4.1) and the probability of a power outage or a share allocation violation.

Mobility factor	Likelihood of Power Outage		Likelihood of Allocation Violation	
	No Mobility Prediction	With Mobility Prediction	No Mobility Prediction	With Mobility Prediction
1	0%	0%	0%	0%
2	8 – 15%	0 – 5%	1 – 3%	0 %
3	30 – 40%	10 – 15%	10 – 12%	0 – 3%

Table 4.2: Impact of Mobility on Outages / Allocation Violations

As can be expected, the more mobile UEs are, the higher their chances for encountering mobility-related problems like power outages and violations of the allocated per class shares. The table presents data collected with and without mobility prediction. The comparison shows that employing prediction significantly decreases the chances of undesired mobility-related effects. Depending on the specific mobility patterns in the network, mobility prediction can be tuned to be more or less aggressive or can be customized for each traffic class.

4.4. Summary of Results

It was shown in this chapter that the introduction of the components discussed in this thesis leads to a number of behavioral improvements. Given the same amount of resources, better performance has been achieved by managing the available capacity in an optimized manner.

An MBAC module helps obtain a better estimation of the actual system load and thus allows for an improved allocation of resources. It has been demonstrated that by implementing some additional optimizations, significant improvements can be achieved. We have observed throughput increases in the range of 80 – 120 Kbps, which could mean that up to a dozen additional users could be accommodated in a cell that would otherwise reject them. At the same time, blocking is up to 15% lower and dropping is decreased by over 15%.

A power sharing structure has been formalized and relationships have been established between the three traffic classes. Clear policies for handling class behaviors in times of light and heavy load have been introduced. While this may seem to be a natural element to implement in a wireless system, it is indeed a novel feature due to the fact that second generation networks only have one traffic class (voice) and no differentiation is possible. Power prediction is used to better manage the inter-class resource allocations and prevent problems before they happen. Despite the non-sophisticated approach taken to power prediction, it has been shown that a UMTS system can benefit from basing its admission decision on some form of projected future outlook. As evidenced by the performance discussion, power prediction can help identify connections, which are likely to cause a mobility-related problem in the future, most often through a combination of a request for a large amount of resources and high mobility. Because such connections are likely to either cause a power outage or a violation of the per-class allocations, they are not admitted in the first place. As a result, the blocking rate may be up to 15% higher, but the dropping rate may decrease by 10-15%.

Chapter 5

Conclusions and Future Work

This thesis proposed a novel QoS-aware Call Admission Control (CAC) scheme for radio access in wideband wireless UMTS networks. It features an efficient CAC algorithm coupled with a QoS class-separation mechanism based on the transmitted power of each individual mobile terminal. A number of improvements have been proposed and implemented in this study. Their major impact on the performance of UMTS is reflected in increased throughput and lower blocking and dropping.

Higher throughput means higher cost-efficiency for the network operator. More customers can be supported (and billed) using the same initial investment. Lower blocking translates into fewer "all circuits are busy" messages which can be quite annoying to users, especially in times when they are using their wireless equipment for mission-critical purposes. Decreased dropping is an even bigger gain for the wireless customer because a frequent call termination could easily cause a customer to switch to another provider. This is especially true in view of the low cost that such an action would incur.

These all are significant issues, particularly in today's telecommunications world where companies are making every possible effort not only to attract new clients but also to retain their already existing customer base. Furthermore, the presented optimizations are relatively simple as opposed to the overall complexity of a UMTS network. Therefore, there should not be much resistance towards implementing solutions of this caliber.

Of course, this is an initial evaluation of all the possibilities that exist to improve the efficiency of a UMTS network. It presents only a fraction of all the optimizations and improvements that could be introduced. As far as the particular components of this thesis are concerned, a number of issues may be further researched to gain a better understanding of what would work best.

The first issue, which needs to be verified, is the theory that most MBAC algorithms achieve similar levels of performance, regardless of their level of complexity. This statement is made in [7] and its validity in wireless environments has to be confirmed. This thesis has shown that even a simple form of MBAC, such as the one actually used, would produce results that are accurate enough. While we have shown that MBAC improves the performance of the system, there may be more sophisticated measurement-based schemes which could yield better results.

Power prediction logic is another area that almost by default offers room for improvement. The mobility prediction algorithm used aims not so much at creating a realistic mobility scenario but at generating a stressful environment. Logically, if the proposed approach works in stressful conditions, it has good chances of being useful in a real mobility prediction scheme. However, it can probably be optimized by implementing techniques to take into account the user's mobility history or using statistical traces of actual inter-cell movements.

As suggested in the performance evaluation section, more research is necessary to determine the relationship between the simulation parameters and the reasons why power allocation violations occur as opposed to power outages (and vice versa). Once the cause is determined, it may be possible to implement a scheme, which would turn per class allocation violations into outages or outages into violations. Arguably, it is preferable to have outages rather than violations because outages only affect the problematic connections themselves and do not unfairly to punish other users.

In conclusion, UMTS, as any other highly-complex system, offers a multitude of aspects, which could be improved and modified. This is even more so in light of the open architecture, which the 3GPP consortium has embraced when developing the UMTS technical specifications. Thus network operators and developers are given a chance to implement proprietary solutions and differentiate their services and products. Indeed, there is no doubt that equipment manufacturers and industry researchers are constantly seeking ideas for optimizations like the ones presented in this thesis.

Some ideas may be successful and will be implemented in future systems and some will need improvement and tuning after their initial deployment. Others may fail because of poor planning or inappropriate design. Still, it is clear that 3G will come to be. It will pave the way for fourth generation systems, or 4G, ideas and speculations about which have been tossed in the public arena for quite some time now. New services will be devised and killer applications will be frantically sought. As usual, there are as many unknowns as there are possible paths for technology to embark on. However, one thing is certain:

The future is bright for data, if we get it right [27].

Bibliography

[1] Y. Lai and S. Tsai, "Some Fair Measurement-based Admission Controls," *Global Telecommunications Conference (GLOBECOM) Proceedings*, vol. 4, 2001.

[2] K. Shiomto, N. Yamanaka and T. Takahashi, "Overview of measurement-based connection admission control methods in ATM networks," *IEEE Communication Surveys*, first quarter, 1999.

[3] S, Floyd, "Comments on measurement-based admissions control for controlled-load services," *Technical report, Lawrence Berkeley Laboratory*, July 1996.

[4] R. Gibbens and F. Kelly, "Measurement-Based Connection Admission Control," *15th International Teletrafic Congress*, June 1997.

[5] E. Knightly and J. Qiu, "Measurement-based admission control with aggregate traffic envelopes," *IEEE ITWDC '98*, Italy, September 1998.

[6] S. Jamin and L. Breslau, "Comments on the Performance of Measurement-Based Admission Control Algorithms," *Proceedings of IEEE Infocom*, March 26-30, 2000.

[7] L. Breslau and S. Shenker, "Measurement-based admission control: what is the research agenda?," *Seventh International Workshop on QoS, IW QoS*, 1999, pp. 3 -5

[8] C. Oliveira, J. B. Kim and T. Suda, "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks," *IEEE Journal on Selected Areas in Communications*, pp. 858–874, August 1998.

[9] A. K. Talukdar, B. R. Badrinath and A. Acharya, "Inte-grated services networks with mobile hosts: Architecture and performance," *Wireless Networks*, pp. 111–124, 1999.

[10] T. Liu, P. Bahl and I. Chlamtac, "Mobility Modeling, Location Tracking, and Trajectory Prediction in Wireless ATM Networks," *IEEE Journal on Selected Areas in Communications*, Volume 16, Issue 6, Aug. 1998, pp 922 -936

[11] J. Chan and A. Seneviratne, "A Practical User Mobility Prediction Algorithm for Supporting Adaptive QoS In Wireless Networks," *IEEE International Conference on Networks ICON'99, Brisbane*; Sep. 1999.

[12] E. S. Elmallah and H. S. Hassanein, “A Power-aware Admission Control Scheme for Supporting the Assured Forwarding Model in CDMA Cellular Networks,” *27th Annual IEEE Conference on Local Computer Networks (LCN)*, Nov. 2002.

[13] H. Holma and A. Toskala, “WCDMA for UMTS Radio Access for Third Generation Mobile Communications,” John Wiley & Sons, 2001.

[14] E. Dahlman, P. Beming, J. Knutsson, F. Ovesjo, M. Persson and C. Roobol, “WCDMA-the Radio Interface for Future Mobile Multimedia Communications,” *IEEE Transactions on Vehicular Technology*, Vol. 47, No. 4, Nov 1998, pp. 1105 - 1118.

[15] C. Y. Huang and R. D. Yates, “Call Admission in Power Controlled CDMA Systems,” *Vehicular Technology Conference Proceedings, 1996*. IEEE 46th, Vol. 3, May 1996, pp. 1665 -1669.

[16] J. Knutsson, P. Butovisch, M. Persson and R. D. Yates, “Evaluation of Admission Control Algorithms for CDMA Systems in a Manhattan Environment,” *Vehicular Technology Conference Proceedings, 1996*. IEEE 46th, Vol. 1, pp. 392-396.

[17] H. Holma and J. Laakso, “Uplink admission control and soft capacity with MUD in CDMA,” *Vehicular Technology Conference, 1999,* IEEE VTS 50th, Vol. 1, pp. 431-435

[18] W. S. Jeon and D. G. Jeong, "Call admission control for CDMA mobile communications systems supporting multimedia services,” *IEEE Transactions on Wireless Communications*, Vol. 1, No 4, Oct. 2002.

[19] J. Laiho, A. Wacker and T. Novosad, “Radio Network Planning and Optimization for UMTS,” John Wiley & Sons, 2002.

[20] 3GPP TSG SA WG2, “Functions for UMTS QoS Provisioning,” *TDoc C-99-057*, 15 – 19 March 1999 Stockholm, Sweden.

[21] 3GPP Technical Report TR 23.907, “QoS Concept and Architecture”, http://www.3gpp.org/, Sep. 1999.

[22] 3GPP web site, http://www.3gpp.org/, Releases 1999, 2001 and 2002.

[23] N. Spencer, “An overview of digital telephony standards,” *The Design of Digital Cellular Handsets (Ref. No. 1998/240)*, Mar. 1998.

[24] IEEE web site, http://www.ieee.org/, April, 2003.

[25] M. P. J. Baker and T. J. Mouslsley, “Power control in UMTS Release '99,” *IEEE First International Conference on 3G Mobile Communication Technologies, 2000 Proceedings*, Publ.. No. 471, pp 36 -40.

[26] 3GPP Technical Specification TS 25.922, “Radio resource management strategies”, http://www.3gpp.org/, Release Sep. 1999.

[27] A. Kobl, “The Evolution from GSM to UMTS,” PhD Thesis at the Vienna University of Technology, 2001.

Appendix A

UMTS Background

The Universal Mobile Telecommunications System is a Third Generation wireless protocol that is part of the International Telecommunications Union's IMT-2000 vision of a global family of 3G mobile communications systems. UMTS is expected to deliver low-cost, high-capacity mobile communications, offering data rates up to 2Mbps.

A.1 UMTS QoS Classes

UMTS has been designed to support a variety of QoS requirements that are set by end users and end-user applications. 3G services will vary from conventional voice telephony to more complex data applications including voice over IP, video conferencing over IP, web browsing, email and file transfer. 3GPP has identified four different main traffic classes for UMTS according to the nature of traffic: conversational, streaming, interactive and background [20].

According to the 3GPP specifications, the main distinguishing factor between these classes is how delay sensitive the traffic is: the conversational class is meant for traffic which is very delay sensitive while the background class is the most delay insensitive traffic class. In the initial phases of UMTS, the conversational and streaming classes will mainly be used to carry real-time traffic flows, while the interactive and background classes will transmit scheduled non-real-time packet data. The UMTS QoS classes are summarized in Table A.1.

Traffic class	**Conversational Class**	**Streaming class**	**Interactive class**	**Background class**
Fundamental characteristics	• Preserve time relation (variation) between information entities of the stream • Conversational pattern (stringent and low delay)	• Preserve time relation (variation) between information entities of the stream	• Request response pattern • Preserve payload content	• Destination is not expecting the data within a certain time • Preserve payload content
Example of the application	• Voice • Video telephony • Video games	• Streaming multimedia	• Web browsing • Network games	• Background download of emails

Table A.1: UMTS QoS Classes

A.2 Network Architecture

The UMTS network architecture will be an evolution of the GSM and GPRS[12] networks, and will therefore have a number of similarities with their architecture. It consists of two parts: the UMTS Terrestrial Radio Access Network (UTRAN) and the Core Network (CN). UTRAN provides the air interface for UMTS terminals and the core network is responsible for switching and routing of calls and data connections to external networks like the Public Switched Telephone Network (PSTN) and the Integrated Services Digital Network (ISDN). The UMTS system architecture is presented in Figure A.2. A complete description of the network architecture and the

[12] The General Packet Radio Service (GPRS) is an enhancement to the GSM system in that it supports data packets. GPRS enables continuous flows of IP data packets over the system for applications such as Web browsing and file transfer. GPRS differs from GSM's short messaging service (SMS) which is limited to messages of 160 bytes in length. GPRS is an example of a transitionary technology, known as 2.5G

interfaces between the logical network elements can be found in the 3GPP Technical Specifications [21].

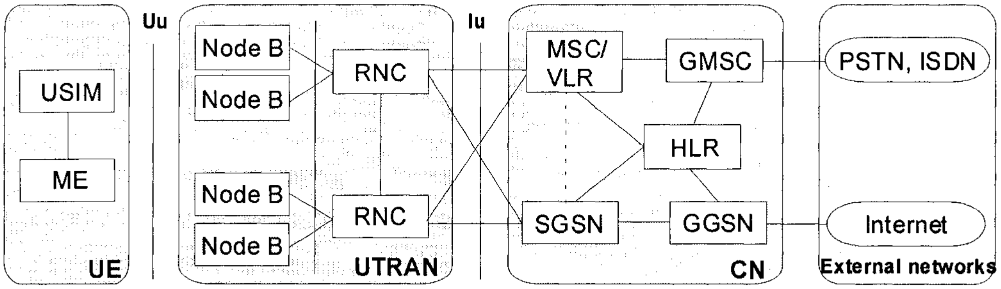


Figure A.2: UMTS Network Architecture

A.3 User Equipment

The UE consists of two parts – the Mobile Equipment (ME) and the UMTS Subscriber Identity Module (USIM).

The ME is the radio terminal used for radio communication over the Uu interface (the WCDMA radio interface). The USIM is a smart card that holds the subscriber identity, performs authentication algorithms, and stores authentication and encryption keys, as well as subscription information that is needed at the terminal.

A.4 UMTS Radio Access Network

The UTRAN also consists of two distinct elements – the Node B and the Radio Network Controller (RNC).

The Node B is the physical unit for radio transmission and reception with cells. The Node B consists of a set of antennas (providing antenna diversity which compensates for multipath propagation) and a controlling element that performs the air interface processing. This includes channel coding, interleaving, rate adaptation and spreading. The Node B is also responsible for softer handovers and inner closed-loop power control. It logically corresponds to the GSM Base Station.

The RNC owns and controls the radio resources in its domain (which covers all Node Bs connected to it). It is responsible for functions like outer loop power control, handover control, congestion control and admission control in the cells it controls. It also and manages code allocation for new radio links, as well as packet scheduling.

A.5 UMTS Core Network

UMTS is based on an evolved GSM core network and integrates circuit and packet switched traffic. The modules that make up the core network are the Home Location Register (HLR), the Mobile Services Switching Center/Visitor Location Register (MSC/VLR), the Gateway MSC (GMSC), the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN).

The home location register is a database in charge of the management of mobile subscribers. It holds the subscriber and location information enabling the charging and routing of calls towards the MSC or SGSN, where the mobile station is registered at that time. The mobile switching center constitutes the interface between the radio system and the fixed networks. The MSC performs all necessary functions in order to handle the circuit switched services to and from the mobile stations. A mobile station roaming in an MSC area is controlled by the visitor location register in charge of this area. The Gateway MSC is the switch at the point where the UMTS Public Land

Mobile Network (PLMN) is connected to external circuit switched networks. All incoming and outgoing circuit switched connections go through the GMSC.

The Serving GPRS support node has similar functionality to that of MSC/VLR, but it is used for packet switched services. Gateway GPRS support node has the same functionality for the packet domain as the GMSC has for the circuit domain.

A.6 WCDMA Air Interface

A.6.1 Multiple Access

The basis for any mobile system is its multiple access scheme, that is, the way the radio spectrum is divided into channels and the common transmission medium is shared between users. Wideband Code Division Multiple Access (WCDMA) is the multiple access method selected by IMT-2000 as the underlying UMTS air interface technology.

WCDMA is conceptually different from traditional technologies like Time Division Multiple Access (TDMA) or Frequency Division Multiple Access (FDMA). Like both of them, WCDMA is a contentionless scheme but it achieves this in a fundamentally different way. Traditionally, eliminating contentions has been done in a fixed fashion by allocating each user a static part of the transmission capacity, or in a demand-assigned fashion, in which scheduling takes place only between the users that have something to transmit. The fixed-assignment technique is used in FDMA and TDMA. With FDMA, the total system bandwidth is divided into several frequency channels which are allocated to users. With TDMA, one frequency channel is divided into time slots which are allocated to users, and the users only transmit during their assigned timeslots. FDMA and TDMA can be used separately or in a combination. An example of the combined operation of TDMA and FDMA is GSM [23]. Examples of demand-assignment contentionless protocols are token bus and token ring LAN standards (IEEE 802.4 and 802.5) [23].

For the sake of completeness, it should be mentioned that a typical contention protocol is ALOHA, where all simultaneously transmitted signals collide [24]. Conflicts are resolved by retransmission after a random amount of time which causes deteriorated utilization, especially under heavily-loaded conditions.

CDMA and WCDMA achieve contentionless operation by using a spread spectrum technique. It overlaps every transmission on the same carrier frequency by assigning a unique code (spreading sequence) to each radio access bearer. This spreading sequence is used to encode to the user information in a way that it will appear as noise to other users. Upon reception by the Node B, the encoded (spread) signal is decoded (despread) using the corresponding spreading code which has been communicated to the Node B ahead of time. Thus the original user bit stream is recovered and multiple users are allowed to transmit simultaneously. The human analogy for this would the ability a person to discern his or her own language in a room full of people speaking many other languages.

This latest UMTS version (release '99) uses a direct spread chip rate of 3.84 Mcps (Megachips per second) and a nominal bandwidth of 5 MHz. The model supports one of WCDMA's two duplex modes: Frequency Division Duplex (FDD) and Time Division Duplex (TDD). FDD has been chosen as the mode of preference for UMTS. In FDD mode, uplink and downlink transmissions use different frequency bands. The radio frame has a length of 10 ms and is divided into 15 equally-sized slots. Spreading factors vary from 256 to 4 for an FDD uplink and from 512 to 4 for an FDD downlink. With these spreading factors, data rates of up to 2 Mbps are attainable.

A.7 Radio Resource Management

Radio resource management comprises a set of algorithms that control the use of the WCDMA radio resources. RRM functionality is partially located in UEs, Node Bs but it is mainly in the RNCs. Unlike 2G networks where RRM is based on hard limits, UMTS deals with varying capacity and flexible services which make 3G RRM much more complex. RRM functionality is aimed at guaranteeing quality of service, offering high capacity and to maintaining the planned coverage area. Thus, RRM optimization is an important part of WCDMA systems when trying to achieve efficient UTRAN performance. This issue is discussed in more detail in Section 1. The basic RRM functions can be classified as follows: power control, handover control, admission control, packet scheduling, and load control.

A.7.1 Power Control

The primary use of power control is to maximize the total air interface capacity by controlling interference. In the uplink, the requirement for power control is complex and the problem arises because of the multiple access interference. Since all users transmit by using the same bandwidth at the same time, they interfere with one another. Due to the attenuation, the signal received by the base station from a UE close to the base station will be stronger than the signal received from another UE located at the cell boundary. Hence, the close users will out-shout the distant users, thus effectively blocking a large part of the cell. This is called the near-far effect, mentioned earlier. To eliminate this problem, all signals, irrespective of distance, should arrive at the base station with the same mean power (assuming that all signals carry the same amount of information, that is, all signals have the same information rate). A solution to this problem is power control, which, by adjusting the UE transmission powers, attempts to achieve a constant received mean power for each user.

By controlling the UE transmission power, power control provides protection against two other undesired effects, which cause variations in the received signal strength: shadowing, and fast-fading. The power control mechanism adjusts the UE transmission power so that it is always the minimum required to maintain a given signal-to-noise ratio (SIR) for the required QoS. In this way, each user contributes to the interference to the least extent possible. There is a direct correspondence between the SIR and E_b/N_o value used throughout the thesis.

Power control is employed on a per connection basis. Typically, a slow outer-loop power control and a fast closed-loop power control are used [25]. Without an accurate power control, WCDMA systems cannot operate. Closed-loop power control compensates for the rapid signal fluctuation at the receiver. The receiver estimates the received SIR, and issues a power control command to the sending end to either increase or decrease its transmission power.

The frequency of fast power control in WCDMA is 1.5 kHz in both uplink and downlink. Because different SIR values correspond to different bit error rates in different radio environments and conditions, outer-loop power control is needed to map the desired BER into the required SIR target. The SIR-target is independently adjusted for each connection based on the BER measurements during the connection. The frequency of the outer loop power control is typically 10–100 Hz.

A.7.2 Handover Control

Handover or handoff is the action of switching an ongoing connection between radio channels in one cell to another without interruption. In WCDMA, handovers can generally be divided into soft handovers and hard handovers [26].

Soft handover is a necessary element of CDMA systems to avoid excessive interference from the neighboring cells. In soft handover, a mobile station is simultaneously connected to more than one Node B. Soft handover increases the performance at the cost of a greater number of connections. When the signal strength of a base station (the so-called pilot signal) exceeds a certain threshold within the UE, the UE enters the soft handover state. When the signal strength drops below drop threshold, the base station is removed from the set of cells that form a soft handover.

Hard handovers in WCDMA refer, for instance, to inter-frequency and inter-system handovers. Inter-frequency handovers are used to balance the load between the carriers if

a base station uses several of them. Inter-system handovers can be made for quality or coverage reasons, for instance, between WCDMA and GSM.

A.7.3 Load Control

The task of load control is to ensure that the system remains stable and does not become overloaded. Thus, the basic purpose of load control is the same as that of admission control. The main conceptual difference is that load control is a continuous process, whereas admission control is carried out as a single event. Load control performs its task by measuring the amount of uplink interference and the total downlink transmission power. If an overload situation is encountered, load control returns the system quickly and controllably back to the normal state (load). Load control performs both preventive actions like the reduction of E_b/N_o targets, and, in rare cases, more extreme actions such as the dropping of connections.

A.7.3 Admission Control

Admission control is responsible for estimating the current load of the system and taking a decision whether a requesting RAB should be accepted or not. Admission control is presented in detail in Chapter 2 and beyond.

Appendix B

OPNET Modeler Overview

OPNET Modeler provides a comprehensive development environment for modeling and performance evaluation of communication networks and distributed systems. It is a system level simulator which mirrors the behavior of protocols, device models and real-life networks. It is a planning tool which facilitates changes by illustrating how the networked environment will actually perform.

OPNET Modeler is used by more than 2500 Customer organizations worldwide. Table B.1 below lists a small fraction of the blue chip companies which are currently using Modeler in their research and development activities.

Manufacturers	Enterprises	Service Providers
3Com Corporation Boeing Cisco Ericsson Intel Lucent Technologies Nokia Nortel Networks	Citicorp FBI Microsoft NASA Oracle Radio Shack	AT&T, AT&T Wireless British telecom Deutsche Telecom France Telecom Sprint TELUS MCI (formerly WorldCom)

Table B.1: Some of the blue chip companies using OPNET

A simulator of Modeler's caliber can be used to achieve a wide variety of practical goals. For example, a Modeler user can have a fairly high degree of certainty when answering questions like how much would replacing an Ethernet hub with an FDDI hub increase throughput, how much would a customized protocol improve efficiency in a cellular network or how much would adding 100 users in a cell increase response time. The following section gives more detail into the most common scientific uses of OPNET.

B.1 Typical Applications of OPNET

OPNET can be used as a platform to develop models of a wide range of systems. Some examples of possible applications are listed below with specific mention of supporting features:

- Standards-based LAN and WAN performance modeling – detailed library models provide major local-area and wide-area network protocols. Configurable application models are also provided by the library, or new ones can be created.
- Internetwork planning – hierarchical topology definitions allow arbitrarily deep nesting of subnetworks and nodes so that large networks can be efficiently modeled.
- Research and development in communication architectures and protocols – OPNET allows the specification of general logic schemes and provides extensive support for communications-related applications. Finite state machines provide a natural representation for protocols.
- Distributed sensor and control networks – OPNET allows the development of sophisticated application-level models, as well as underlying communications protocols and links. Customized performance metrics can be computed and

recorded. Scripted and stochastic inputs can be used to drive the simulation model, and processes can dynamically monitor the state of objects in the system via formal interfaces provided by statistic wires.

- Resource sizing – accurate, detailed modeling of a resource's request-processing policies is required to provide precise estimates of its performance when subjected to peak demand (for example, a packet switch's processing delay can depend on the specific contents and type of each packet as well as its order of arrival). Queuing capabilities of Proto-C provide easy-to-use commands for modeling sophisticated queuing and service policies; library models are provided for many standard resource types.
- Mobile packet radio networks – specific support for mobile nodes, including predefined or adaptive trajectories; predefined and fully customizable radio link models; geographical context provided by OPNET network specification environment.
- Satellite networks – specific support for satellite nodes, including automatic placement on specified orbits, a utility program for orbit generation, and an orbit visualization and orbital-configuration animation program.
- Intelligence and tactical networks – support for diverse link technologies; modeling of adaptive protocols and algorithms in Proto-C; notification of network component outages and recoveries; scripted or stochastic modeling of threats. Radio link models support determination of friendly interference and jamming.

B.2 Key System Features

OPNET Modeler (henceforth referred to as OPNET) offers an extensive set of features designed to support general network modeling as well as simulation of particular network aspects. This section provides a brief enumeration of some of the most important capabilities of OPNET.

- Object orientation – Systems specified in OPNET consist of objects, each with configurable sets of attributes. Objects belong to classes which provide them with their characteristics in terms of behavior and capability. Classes can also be derived from other classes, or modified for specific support for particular applications.
- Hierarchical models – OPNET models are hierarchical, naturally paralleling the structure of actual communication networks.
- Graphical specification – Whenever possible, models are entered via graphical editors. These editors provide an intuitive mapping from the modeled system to the OPNET model specification.
- Flexibility to develop detailed custom models – OPNET provides a flexible, high-level programming language with extensive support for communications and distributed systems. This environment allows realistic modeling of all communications protocols, algorithms, and transmission technologies.
- Automatic generation of simulations – Model specifications are automatically compiled into executable, efficient, discrete-event simulations implemented in the C programming language. Advanced simulation construction and configuration techniques minimize compilation requirements.

- Application-specific statistics—OPNET provides numerous built-in performance statistics that can be automatically collected during simulations. In addition, modelers can augment this set with new application-specific statistics that are computed by user-defined processes.
- Interactive analysis – All OPNET simulations automatically incorporate support for analysis via a sophisticated interactive debugger.
- Cosimulation –OPNET can be connected with external simulators to see how the models in those simulators interact with OPNET models. The external models can represent anything from network hardware to end-user behavior patterns.

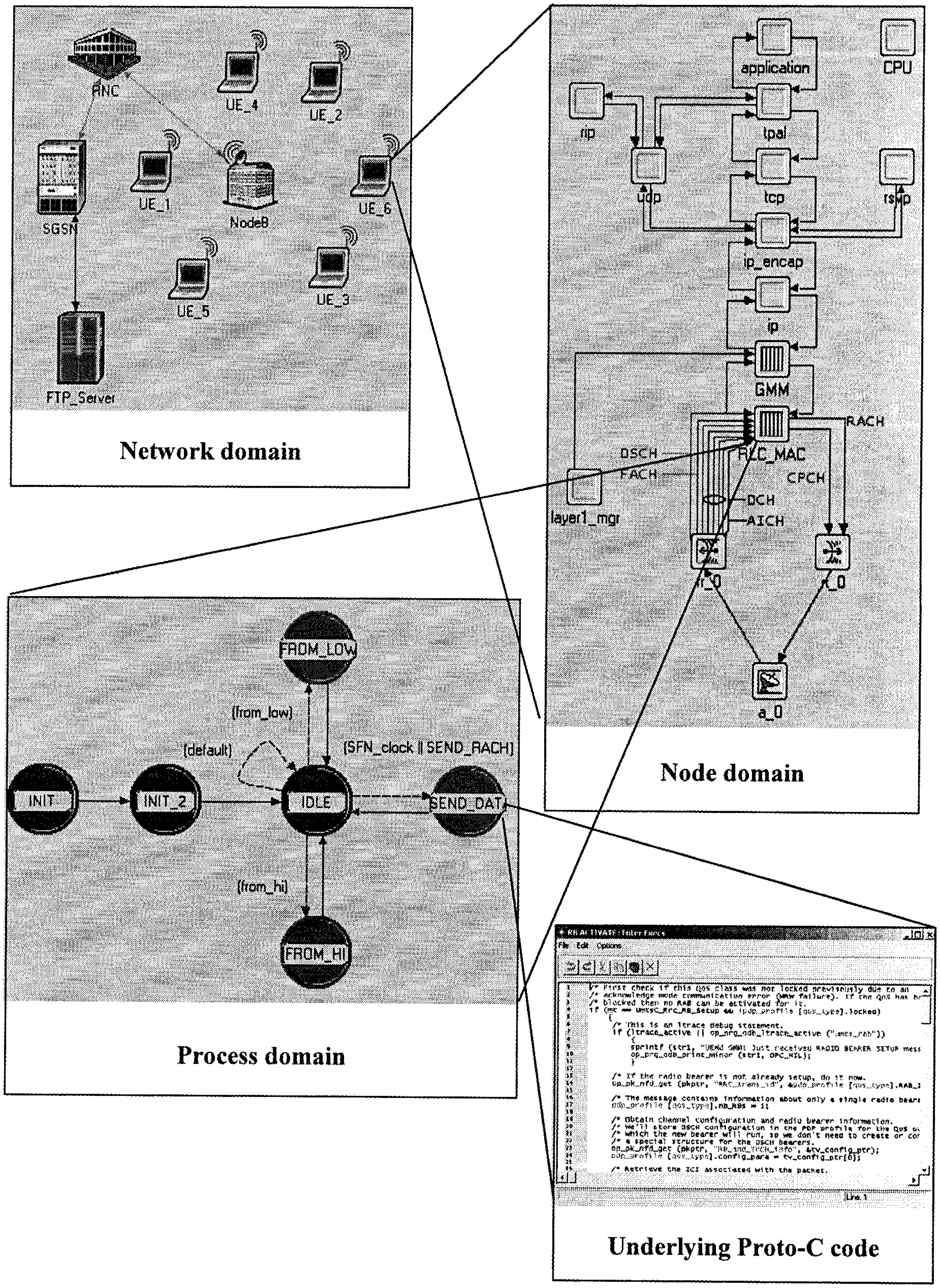
B.3 Simulator Structure

One of OPNET's most convenient features is its hierarchical structure. Objects can be specified and manipulated on three different levels – network, node and process levels. These are referred to as the modeling domains of OPNET, since they essentially span all hierarchical levels of a model.

Domain	**Modeling Focus**
Network	Network topology described in terms of subnetworks, nodes, links, and geographical context.
Node	Node internal architecture described in terms of functional elements and data flows between them. A node consists of modules which represent the internal capabilities of a node such as data creation, transmission, reception, storage, internal routing and queuing.
Process	Behavior of processes (protocols, algorithms, applications), specified using finite state machines and extended high-level language.

Table B.2: OPNET modeling domains

Figure B.1 below gives a graphical illustration of the relation between the three modeling domains.



Network domain

Node domain

Process domain

Underlying Proto-C code

```
/* First check if this QoS class was not locked previously due to an
/* acknowledge mode communication error (AMW failure). If the QoS has be
/* blocked then no RAB can be activated for it.
if (mt == UmtsC_Rrc_RB_Setup && !pdp_profile [qos_type].locked)
    {
    /* This is an ltrace debug statement.
    if (ltrace_active || op_prg_odb_ltrace_active ("umts_rab"))
        {
        sprintf (str1, "UEMd GMM: Just received RADIO BEARER SETUP mess
        op_prg_odb_print_minor (str1, OPC_NIL);
        }

    /* If the radio bearer is not already setup, do it now.
    op_pk_nfd_get (pkptr, "RRC_trans_id", &pdp_profile [qos_type].RAB_I

    /* The message contains information about only a single radio beare
    pdp_profile [qos_type].nb_RBS = 1;

    /* Obtain channel configuration and radio bearer information.
    /* We'll store DSCH configuration in the PDP profile for the QoS o
    /* which the new bearer will run, so we don't need to create or cor
    /* a special structure for the DSCH bearers.
    op_pk_nfd_get (pkptr, "RB_and_TrCH_info", &tv_config_ptr);
    pdp_profile [qos_type].config_para = tv_config_ptr[0];

    /* Retrieve the ICI associated with the packet.
```

Figure B.1: The three domains in OPNET

The figure clearly shows that the farther down the hierarchy we go, the more detailed the functional elements become. Thus UE_6 which is a wireless node in the network domain, is broken down into its building elements in the node domain. It is shown to consist of a TCP/IP stack, a receiver, a transmitter and an antenna. In turn, the MAC layer of this stack consists of a finite state machine with six states and conditional transitions between these states. The behavior of each of those states (the corresponding algorithms and protocols) is governed by the underlying C/C++ code.

B.4 Data Collection and Simulation

The objective of most modeling efforts is to obtain measures of a system's performance or to make observations concerning a system's behavior. OPNET supports these activities by creating an executable model of the system. Provided that the model is sufficiently representative of the actual system, OPNET allows realistic estimates of performance and behavior to be obtained by executing simulations. Several mechanisms are provided to collect the desired data from a simulation. These are output vectors, output scalars, and animations.

Output vectors represent time-series simulation data. They consist of a list of entries, each of which is a time-value pair.

Scalar statistics are individual values that represent a metric of interest. They are often derived from vector statistics, as an average, peak value, final value, or other statistic. Typically, a single value for each scalar statistic is recorded during a simulation; when many simulations are run, their scalar outputs are combined to form a graph.

Animations are graphical depictions of certain system parameters such as packet flows, node movements or state transitions.

B.5 Statistics Collection Modes

There are three types of statistic collection modes – all values, sample and bucket mode.

In the all values mode, every data point is collected from a statistic. In the the sample mode, data is collected according to a user-defined time interval or sample count. For example, the user could specify that data be collected every 10th simulation second or every 10th data point. When the bucket mode is used, all the data points are collected over the time interval or sample count and are processed according to a user-defined parameter-- max, min, sum, count, sample average or time average. The bucket collection mode has been used for the purposes of this study.

B.6 Event-driven Simulation

OPNET is an event-driven simulator. Simulation time advances only when an event with a later time is taken from a global structure called event list. No simulation time occurs during an invocation of a process model. No time elapses during transitions between states.

B.7 Traffic Modeling

There are three methods of modeling network traffic in OPNET. Each of them gives a different level of detail which is a function of the number of explicitly modeled events in the system (as opposed to the number of approximated or analytically deducted events). The types of modeled traffic are explicitly modeled traffic and conversational pair traffic. The user can use any of these methods separately or in a combination to create an accurate model of the load on a network. For the purposes of this thesis, only explicit traffic has been used.

B.7.1 Explicitly Modeled Traffic

When traffic is modeled explicitly, every single event pertaining to the traffic generation, transmission, routing, signal loss, reception, etc. is explicitly modeled. This is done separately for every data packet and allows the user to achieve a high level of precision by specifying in great detail the nature of the generated traffic.

The biggest, and possibly the only disadvantage of modeling traffic explicitly is that because an enormous number of network events are generated, simulation run times tend to be rather long, especially for large, complex systems.

B.7.2 Conversational Pair Traffic

This is a method of analytical modeling of the traffic. Mathematical formulae are used to handle traffic flows as a whole, and not on a per-packet basis. This type of modeling generates very few events and thus allows shorter simulation runtimes to be achieved. All that is needed from the user is to specify a source and one or more destinations for the traffic.

Conversation Pair Traffic is either manually entered imported from applications which monitor the traffic patterns of real systems (such as NetMetrix, NetFlow or NetScout).

B.8 Radio Modeling

Radio is a broadcast technology and depends on dynamically changing parameters. The simulation must evaluate the possible connectivity between a transmitter channel and every receiver channel for each transmission.

The network level characteristics factored into these calculations are the locations of the source and destination nodes, the distance between the nodes, and the direction the radio signal travels from the source node to the destination node.

If the nodes are mobile or satellite nodes, these position-related parameters may change during a simulation.

B.8.1 Radio Links

A radio link is not statically represented by an object, as are point-to-point and bus links. Radio links can exist between any radio transmitter–receiver channel pair and are dynamically established during simulation. The possibility of a radio link between a transmitter channel and a receiver channel depends on many physical characteristics of the components involved, as well as time-varying parameters, which are modeled in the so-called Transceiver Pipeline Stages. In OPNET simulations, parameters such as frequency band, modulation type, transmitter power, distance, and antenna directionality are common factors that determine whether a radio link exists at a particular time or can ever exist

B.8.2 Transceiver Pipeline

The transceiver pipeline implements the functionality normally attributed to the physical layer in real systems. The pipeline models the transmission of packets across a communications channel with varying parameters. It is divided into multiple stages, each modeling a particular aspect of the channel. Its ultimate goal is to determine whether or not a packet can be received at the link's destination by handling functions like noise factoring, BER estimation and error detection and correction.

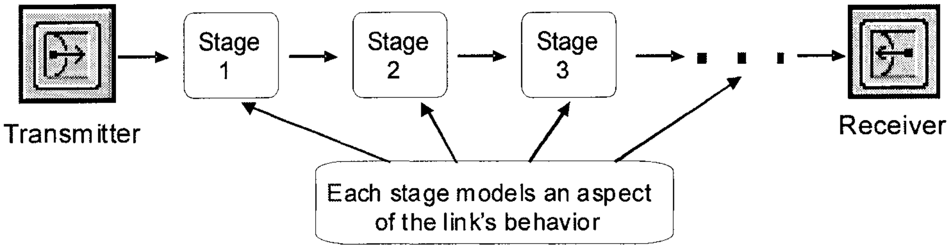


Figure B.2: Transceiver Pipeline Logic

The table below lists the pipeline stages on both the transmitter's and the receiver's side.

Radio Transmitter	Radio Receiver
• Receiver Group • Transmission Delay • Link Closure (LOS) • Channel Match • Tx Antenna Gain • Propagation Delay	• Rx Antenna Gain • Received Power • Background Noise • Interference Noise • Signal-to-Noise Ratio • Bit Error Rate • Error Allocation • Error Correction

Table B.3: Pipeline stages associated with the transmitter and receiver

B.9 Summary

In summary, OPNET Modeler is a powerful simulation environment which allows the user to design and study communication networks, devices, protocols, and applications with a great deal of flexibility.

Modeler's object-oriented approach mirrors the structure of actual networks and network components, so the system intuitively maps to the desired model. Modeler supports a wide variety of network types and technologies, including wireless and cellular, and allows even the most difficult questions to be answered with confidence.