**A NOVEL QoS AWARE INTEGRATED
CHARGING AND RESOURCE
CONTROL FRAMEWORK FOR NEXT
GENERATION WIRELESS
NETWORKS**

By

Walid Ibrahim

A thesis submitted to the Faculty of Graduate Studies in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Department of Systems and Computer Engineering
Faculty of Engineering
Carleton University
Ottawa, Canada
August 2002

© 2002, Walid Ibrahim

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-79430-X

Canada

The undersigned recommend to the Faculty of Graduate Studies and
Research acceptance of the thesis

**A NOVEL QoS AWARE INTEGRATED
CHARGING AND RESOURCE
CONTROL FRAMEWORK FOR NEXT
GENERATION WIRELESS
NETWORKS**

Submitted by

Walid Ibrahim

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

**Rafik Goubbran, Chair Department of
Systems and Computer Engineering**

Patrick O'Kelly, External Examiner

John W. Chinneck, Thesis Co-Supervisor

Shalini Periyalwar, Thesis Co-Supervisor

Carleton University
August 2002

ABSTRACT

Powered by the magnificent success of the Internet, wireless networks are currently undergoing an evolution toward their third generation. The third generation wireless services will include Internet access to interactive multimedia and video conferencing as well as traditional services such as voice, email and web access. Creating a system that successfully combines both multimedia and wireless technologies requires the solution of many challenging technical issues. One of the most important issues is the creation of a charging and resource control scheme suitable for wireless environment that charges customers fairly according to the Quality of Service (QoS) that they receive.

Existing wireless charging and resource control schemes are not suitable for next generation networks since they have been designed and optimized to support circuit switched voice and low bitrate data calls only. At the same time, several deficiencies have been identified in the currently proposed wireless and wireline QoS based resource control schemes.

In this research we propose a novel QoS aware **I**ntegrated **C**harging and **R**esource **C**ontrol (ICRC) framework that can efficiently manage next generation wireless networks. The proposed framework charges customers fairly using a new QoS-based charging model. It improves the resource allocation efficiency by allocating resources to customers based on their QoS requirement. It gives service providers the flexibility to select and/or implement ICRC policies that best fit their needs. Finally, the proposed framework utilizes a new reservation-less handoff scheme.

ACKNOWLEDGMENTS

I am deeply indebted to my supervisors, Professor John Chinneck, and Dr. Shalini Periyalwar for their encouragement and friendship throughout this research. Their continuous support, patience, and guidance have been enormous and have been a major factor in the success of this work.

My family has been very supportive to me throughout the years. I am very grateful to my wife Hoda, for her moral support and constant encouragement. She and my son, Hazem were very patient throughout many days and weekends where they missed having lots of fun because I always had some work to do.

I am so grateful to my mother, mother in law, and grand parents. Without their prayers, I wouldn't have finished my work. I would like also to thank my sisters, and sisters and brothers in law for their encouragement and support throughout the last few years.

I would like to dedicate this thesis to the memory of my father, Ezzat Hassan Ibrahim. His dedication to raising me in the best way he could is the reason I have finished this research.

TABLE OF CONTENTS

1. INTRODUCTION.....	1
1.1 Motivation.....	1
1.2 Objectives	3
1.3 Thesis Contents.....	5
2. BACKGROUND	7
2.1 Microeconomics Overview	7
2.1.1 Utility Function	7
2.1.2 Competitive Market	11
2.1.3 Network Efficiency vs. Economic Efficiency	12
2.1.4 Fairness	13
2.1.5 Auctions.....	16
2.2 The Evolution Toward High Speed Wireless Internet	18
2.2.1 QoS in Wireless Networks.....	20
2.2.2 Wireless QoS Classes.....	21
2.3 Subjective QoS	23
3. REVIEW OF THE STATE OF THE ART	26
3.1 Charging in Wireline Data networks	26
3.1.1 Static Pricing Policies.....	27
3.1.2 Dynamic Pricing Policies.....	32
3.1.3 Shadow Pricing.....	41
3.2 Charging in Wireless Data Networks	41
3.3 Charging Scheme Evaluation Criteria	42
3.4 Resource Control in Wireless Data Networks.....	45
3.4.1 Class Based Resource Allocation	46

3.4.2	Graceful Degradation of Service	47
3.4.3	Adaptive Bandwidth Allocation in Wireless Networks	48
3.4.4	QoS Provisioning Framework	51
3.4.5	Rate-Based Borrowing Scheme	52
3.5	Conclusions.....	53
4.	PROBLEM STATEMENT	54
5.	INTEGRATED CHARGING AND RESOURCE CONTROL FRAMEWORK	60
5.1	Framework Charging Model	66
5.2	Customer's QoS Profile	71
5.2.1	How to Use the QoS Profile.....	75
5.2.2	QoS Profile Benefits	78
5.3	Market Model	82
5.4	Framework Policies and Tuning Parameters	84
5.4.1	Performance Tuning Parameters.....	87
5.5	Framework Processes	90
5.5.1	Framework Manager	90
5.5.2	Price Adjustment	92
5.5.3	Admission Control Process.....	95
5.5.4	Resource Allocation Process.....	100
5.5.5	Charging Process.....	100
5.6	Assessing the ICRC FRAMEWORK	102
5.6.1	Framework Overhead.....	102
5.6.2	Predictability.....	103
5.6.3	Flexibility.....	104
5.6.4	Practicality	105
5.6.5	Control	105
5.6.6	Choice	106

5.6.7	Customer Interaction	106
5.6.8	Technical Knowledge Requirement	107
6.	CHARGING AND RESOURCE CONTROL POLICIES	108
6.1	Revenue Based Policy	108
6.1.1	Problem Formulation	109
6.1.2	Price Adjustment Scheme	110
6.1.3	Admission Control Scheme	113
6.1.4	Resource Allocation Process.....	118
6.2	Utility Based Policy	123
6.2.1	Utilitarian Criterion Based Resource Allocation.....	124
6.2.2	Equality Criterion Based Resource Allocation	126
6.2.3	Price Adjustment Process.....	128
6.2.4	Admission Control Process.....	129
6.2.5	The Policy Optimality	130
6.3	Market Based Policy	130
6.3.1	Resource Allocation Process.....	131
6.3.2	Price Adjustment Process.....	134
6.3.3	Admission Control Process.....	135
6.3.4	The Policy Optimality	135
7.	EXPERIMENTAL SETUP.....	136
7.1	Traffic and Mobility Generation Module	137
7.2	Resource Control Module	141
7.2.1	IBM Optimization Library (OSL)	145
7.3	Performance Metrics	147
7.4	Compared Polices	151
8.	RESULTS AND COMPARISONS.....	154
8.1	Call Blocking Percentage	154

8.2	Resource Utilization Level.....	158
8.3	Bandwidth market price	159
8.4	Call Dropping Percentage	161
8.5	Customer Satisfaction.....	163
8.6	Average Network Revenue	166
8.7	Comprehensive Network Satisfaction	168
8.8	Tuning parameters Effect	169
8.8.1	Minimum Admission Threshold.....	170
8.8.2	Minimum Bandwidth market price.....	173
8.8.3	Maximum Blocking Percentage	177
8.8.4	The Objective Minimum Percentage Improvement.....	179
8.9	Customer's Budget Effect	181
8.10	Exact vs. Heuristic Revenue-Based Resource Allocation.....	182
8.11	Validation.....	183
8.12	Conclusions.....	186
9.	CONCLUSIONS	189
9.1	Summary of Contributions	191
9.2	Future Research	192

LIST OF FIGURES

<i>Number</i>	<i>Page</i>
Figure 2-1: Utility Function for Elastic Traffic	9
Figure 2-2: Utility Function for Real-time Applications	10
Figure 2-3: Stepwise Utility Function	10
Figure 3-1: Effective Bandwidth Calculation	36
Figure 5-1: The Framework Deployment in a UMTS Network	65
Figure 5-2: MAC Layer Scheduler	68
Figure 5-3: Video QoS Profile Graphical User Interface.....	77
Figure 5-4: ICRC Framework Hierarchy	85
Figure 5-5: Framework Processes	91
Figure 5-6: Finding the Optimal BW market price to Maximize a Policy Objective.	94
Figure 5-7: Call Admission Flowchart for Maximizing a Policy Objective	99
Figure 5-8: Zero Dropping Resource Allocation Scheme.	101
Figure 6-1: Video QoS Profiles	111
Figure 6-2: Network Revenue Changes with the Bandwidth Market Price	111
Figure 6-3: Optimal Price Adjustment Heuristic Scheme	114
Figure 6-4: Revenue Based New Call Admission Flowchart	117
Figure 6-5: Heuristic Resource Allocation Scheme.....	122
Figure 6-6: Utility Criterion Based Resource Allocation	125
Figure 6-7: Equality Based Resource Allocation.	127
Figure 6-8: Market Based Resource Allocation.	133
Figure 7-1: Simulation Environment.....	136
Figure 7-2: Resource Control Class Diagram	142
Figure 7-3: Call State Machine	144
Figure 7-4: A Failed Call Setup Request Sequence Chart	145

Figure 7-5: Successful Call Lifetime Sequence Chart	146
Figure 8-1: Call Blocking Percentage.....	155
Figure 8-2: The Rate-Based Call Blocking Components	157
Figure 8-3: Resource Utilization Level	158
Figure 8-4: Bandwidth market price.....	160
Figure 8-5: Market Policy Bandwidth market price	161
Figure 8-6: Call Dropping Percentage.....	162
Figure 8-7: Admitted Calls Average Satisfaction Level.....	163
Figure 8-8: Comprehensive Grade of Service.....	165
Figure 8-9: Network Generated Revenue per Unit of Resource.....	167
Figure 8-10: Comprehensive Network Satisfaction.....	169
Figure 8-11: Revenue Based Policy: Call Blocking Percentage.....	171
Figure 8-12: Revenue Based Policy: Average Revenue	172
Figure 8-13: Revenue Based Policy: Admitted Calls Average Satisfaction Level	172
Figure 8-14: Utility Based Policy: Resource Utilization	174
Figure 8-15: Utility Based Policy: Blocking Percentage	175
Figure 8-16: Utility Based Policy: Average Revenue.....	176
Figure 8-17: Utility Based Policy: Admitted Calls Average Satisfaction Levels	176
Figure 8-18: Utility Based Policy: Call Blocking Percentage	177
Figure 8-19: Utility Based Policy: Admitted Calls Average Satisfaction Levels	178
Figure 8-20: Utility Based Policy: Average Revenue.....	178
Figure 8-21: Revenue Based Policy: Call Blocking Percentage.....	179
Figure 8-22: Revenue Based Policy: Average Revenue	180
Figure 8-23: Revenue Based Policy: Average Admitted Calls Satisfaction Levels	180
Figure 8-24: Satisfaction Levels for Various Customer's Budgets	181
Figure 8-25: Revenue Based Policy: Optimization Procedure vs. Generated Revenue ..	182
Figure 8-26: Utility Based Policy: Minimum Bandwidth Market Price vs. Bandwidth Market Price.....	185

LIST OF TABLES

<i>Number</i>	<i>Page</i>
Table 2-1: European Path Toward 3G	19
Table 2-2: American Path Toward 3G	19
Table 2-3: UMTS QoS Classes	22
Table 2-4: IEEE WBA Recommended QoS Parameters	22
Table 3-1: Charging Schemes Evaluation	45
Table 5-1: Example End-to-End Delay Tariff Parameter	69
Table 5-2: Example Frame Error Rate Tariff Parameter	71
Table 5-3: A QoS Profile Example	76
Table 5-4: A QoS Profile for a Guaranteed Poor Satisfaction Level	80
Table 5-5: Price Satisfaction Profile	82
Table 5-6: QoS Profile Definition	84
Table 5-7: ICRC Framework Evaluation	107
Table 6-1: Customer's Utility	124
Table 6-2: Market Based Resource Allocation Sequence	132
Table 7-1: Call Type Probabilities	139
Table 7-2: Framework Default Tuning Parameters	141
Table 8-1: Call Blocking Percentage Confidence Intervals	157
Table 8-2: Bandwidth Utilization Level Confidence Intervals	159
Table 8-3: Bandwidth Market Price Confidence Intervals	162
Table 8-4: Customer's Average Satisfaction Level Confidence Intervals	164
Table 8-5: Comprehensive QoS Confidence Intervals	166
Table 8-6: Network Revenue Confidence Intervals	168
Table 8-7: Comprehensive Network Satisfaction Confidence Intervals	170

GLOSSARY

3GPP.	Next generation Partnership Project
AMR.	Adaptive Multi-Rate
ARQ.	Automatic Repeat Request
BER.	Bit Error Rate
CDMA.	Code Division Multiple Access
ICRC.	Integrated Charging and Resource Control
ITU.	International Telecommunication Union
ETED.	End-To-End-Delay
ETSI.	Telecommunications Standards Institute
FER.	Frame Error Rate
FEC.	Forward Error Correction
GSL.	Gnu Scientific Library
GSM.	Global System for Mobile communication
LP.	Linear Programming
MILP.	Mixed Integer Linear Programming
MINLP.	Mixed Integer Non-Linear Programming
MPEG.	Moving Picture Experts Group
OSL.	Optimization Solution and Library
PMP.	Paris Metro Pricing
QoS.	Quality of Service
QoS-CM.	Quality of Service Charging Model
RNC.	Radio Network Controller
TDMA.	Time Division Multiple Access
UMTS.	Universal Mobile Telecommunication System
WCDMA.	Wideband Code Division Multiple Access

Chapter 1

INTRODUCTION

1.1 Motivation

As demands increase for wireless wideband communication services, such as high-speed Internet access and high-quality video and audio transmission, next-generation mobile communication systems must be designed to efficiently support different kinds of applications with diverse bandwidth and QoS requirements. Current wireless networks, however, are not optimally designed for efficiently and effectively supporting the introduction of these new types of traffic. The system is built on a design specifically intended for voice terminals but is already being used for other data oriented applications as well. In order to make the system more efficient for future multimedia terminals, several steps should be taken. More specifically, charging and resource control schemes must be reevaluated, and new more efficient versions must be implemented as required.

On the application side, instead of using traditional multimedia applications with hard QoS guarantee requirements, a new class of applications called adaptive applications will

be used. Adaptive applications are capable of adjusting their resource usage depending on availability. For instance, an audio application could reduce its transmission rate during times of high congestion by using a lower coding rate or an alternate coding technique.

In the case of non-adaptive applications, a new or a handoff call is now forced to block or terminate respectively if the admitting cell does not have enough resources to satisfy the incoming call QoS requirements. However, in an adaptive multimedia paradigm, the resources reallocation algorithm could reduce the amount of resources allocated to active calls, thereby freeing more resources to accommodate the incoming call. Moreover, a QoS negotiation algorithm could also be used to adjust the amount of resources allocated to an incoming call according to the system resources utilization level.

During its lifetime, an adaptive application adjusts its traffic inputs or QoS demands in response to feedback signals from the network about the current state of network resources. It is important to mention that adaptive applications will work well only when customers cooperate and do not modify their applications to neglect the feedback signals. Adaptive applications become less effective when customers have no benefit from responding to the feedback signals.

For the same application, different customers are believed to have different QoS perception levels. For instance, an 8 Kbps voice connection might be very satisfactory for one customer and unacceptable for another. Currently, the resource allocation schemes treat customers similarly despite the expected variance in their satisfaction levels.

Treating customers equally will either waste scarce network resources by allocating more resources to already satisfied customers or will provide some customers with unacceptable network services.

At the same time, treating customers equally could lead to the loss of some business opportunities for the service providers. Some customers would have paid a higher price in order to receive a higher QoS satisfaction level. Future resource control schemes should give customers the opportunity to convey their QoS preferences. The network resources should then be allocated in accordance to the customer's preferences given that higher QoS services will cost more.

1.2 Objectives

The objective of this work is the development of a novel QoS aware ICRC framework that could be efficiently used for next generation wireless networks. The framework should allow customers to convey their different satisfaction levels in terms of QoS requirements. Customers should be able to define one or more satisfaction levels per application. Adaptive resources control and charging schemes are required to deal with the natural real-time changes in the wireless resources availability and the massive number of handoff calls expected from the tiny cell size of the next generation wireless networks.

Since customers will be invited to select their QoS preferences, the framework should give customers a strong incentive to reveal their true QoS requirements. The strong incentive is required to prevent customers from overbooking the scarce wireless resources as well as to avoid overcharge to the customers.

Given that the wireless resources are very limited, the framework should maintain a minimal transmission overhead while providing customers with means to express their QoS requirements. The number and the size of the resource negotiation and control messages should be minimized.

It is strongly believed that dropping a handoff call is much more annoying to customers than blocking a new one. As a result, the framework should treat handoff and new call admission requests differently. Higher priority should be given to the handed-off calls. The framework design should also incorporate the adaptive application concept to minimize the impact from the massive number and variety of handoff calls.

Finally, the ICRC framework should also give service providers the flexibility to choose the ICRC policy that best fits their marketing strategy. We believe that the ICRC framework should utilize various optimization techniques to fulfill the policy objective established by the service provider. The service provider should have the ability to adjust a set of optimization parameters to tune the resource control objective at any time according to market conditions.

1.3 Thesis Contents

In the following chapter, a background of microeconomics and its applications in telecommunication industry is provided. The definitions of the utility function and resource allocation fairness and optimality are given. The relationship between customers and service providers is described based on competitive market concepts. An overview of the evolution towards multimedia wireless networks is illustrated. The wireless QoS requirements and challenges are discussed.

Chapter 3 highlights the state of the art in the charging techniques currently used in data networks. It draws attention to the differences between dynamic and fixed charging schemes and the usage of charging schemes to control network congestion. Resource allocation policies currently proposed for the next generation wireless networks are also illustrated.

The problem statement is the focus of Chapter 4. In this chapter, the various deficiencies in the currently used and proposed resources allocation schemes are discussed.

Chapter 5 includes the details of the thesis contributions. A novel customer's QoS profile and generic charging technique is proposed. The framework tuning parameters and the market model are illustrated along with the common features of the framework processes.

To demonstrate the framework flexibility, three different ICRC policies are proposed in Chapter 6. For each policy, the functionality of the framework processes is described in detail. The policies heuristic and optimal schemes are described in this Chapter too.

The simulation environment is described in Chapter 7. The simulation assumptions and traffic parameters are given along with third party libraries used for random number generation (RNG) and optimization problem solving.

In Chapter 8, the performance of the proposed framework is validated. The simulation results are discussed and compared with two of the currently used charging and resource control policies.

Chapter 9 concludes the thesis with a summary of the work, a list of the contributions, and possible extensions to the provided solution are outlined.

Chapter 2

BACKGROUND

In this Chapter, the background material related to the thesis is presented. An overview of microeconomics resource allocation concepts such as utility functions, fairness and optimality is given followed by an overview of the wireless evolution toward its third generation and the wireless QoS aspects.

2.1 Microeconomics Overview

In this section, some of the microeconomics concepts that have been introduced in the resource allocation and charging literature are highlighted. Concepts like utility functions, competitive markets, optimality, and fairness criteria are described.

2.1.1 Utility Function

Utility is the variable whose relative magnitude indicates the direction of preference. The statement “basket A is preferred to basket B” and the statement “basket A has more utility than basket B” mean the same thing [72]. In other words, the utility functions describe

how customers rank the performance of an application depending on the delivered service [81].

In multi-service networks, utility, ideally, should be expressed as a function of actual QoS parameters, such as delay, packet loss, number of frames, resolution, and sound quality. However, in most real networks, such quality measures are virtually impossible to predict in advance and are closely dependent on factors such as traffic models and network topology. Although there are applications currently having a rich customer interface that allows customers to provide some application-specific performance information to network services, it has been shown in [40] that this information is poorly correlated with network-layer performance.

For example, the cell loss rate for an MPEG video stream may correlate only weakly with packet loss and application quality. Therefore, utility is often expressed as a function of the amount of resources allocated to the customer. In this fashion, the utility still indicates a customer's sensitivity to changes in QoS.

A utility function depends mainly on two factors: the application it represents, and customers' perception. One can expect that the e-mail utility function should be different from Voice over IP or streaming video utility functions. At the same time, for the same type of application, the satisfaction levels differ from one customer to another, and hence the utility functions defined by them.

In the literature, data applications are usually classified based on their elasticity to the amount of allocated resources. Applications such as electronic mail, file transfer, and telnet sessions are rather tolerant of end-to-end delay (ETED) and throughput. This class of applications is called elastic applications. A concave bandwidth utility function, Figure 2-1, can be used to describe the utility as a function of the available bandwidth for this class of applications [13].

Figure 2-1: Utility Function for Elastic Traffic

Real-time applications are generally delay sensitive and have hard real-time requirements. For this class of applications, degradation in bandwidth may severely degrade the performance. A reasonable model for this class of applications is a nearly single step utility-bandwidth function as shown in Figure 2-2.

Figure 2-2: Utility Function for Real-time Applications

Another group of applications cannot adjust their functionality to any arbitrary bandwidth offered by the network. These applications could have a vocoder that operates at specific code rates only or different video compression algorithms. Such applications may have a stepwise utility function as shown in Figure 2-3.

Figure 2-3: Stepwise Utility Function

2.1.2 Competitive Market

Competitive market model is a method to model the competition between players (consumers and producers) over scarce resource based on supply and demand [33]. The competitive market model consists of scarce resources and two types of agents, consumers and producers. A resource is an item or service which is valued by agents in the economy. Since it is scarce, there is never enough of the resource to satisfy all the agents all the time. For this reason, allocation decisions must be made. Consumers require resources to satisfy wants. Producers create or own the resources sought by consumers.

Agents come together at a market, where they buy and sell resources. Prices are set with respect to supply and demand. The price increases if the demand is greater than the supply and decreases when the demand is less than the supply. When they are equal, the market and price are in equilibrium. This moment is referred to as “clearing the market” and the resulting allocation is Pareto optimal.

2.1.2.1 Pareto optimality

Pareto optimality, developed by “Vilfredo Pareto”, is the most widely accepted criterion of economic efficiency. A distribution of a given quantity of goods is Pareto optimal, and thus efficient, if and only if there is no feasible alternative distribution of those goods in which at least one person is better off and no one is worse off [72].

Given a set of resources and entities, it is easy to see that there are many different possible Pareto-optimal allocations. To be more specific, the Pareto-optimal allocations can be

further classified based on fairness definitions. A resource allocation algorithm should find the Pareto optimal allocation that fulfills the fairness criterion adopted by the network. Later in the thesis, we will show that some of the proposed schemes are Pareto optimal (Section 6.2.5 and 6.3.4).

2.1.3 Network Efficiency vs. Economic Efficiency

In focusing on customer preferences, two different notions of efficiency have been defined in [67]:

- Network efficiency refers to the utilization of network resources, such as bandwidth and buffer space.
- Economic efficiency refers to the relative valuations the customers attach to their network service.

If a network can maintain an acceptable level of service while minimizing the resources needed to provide this service, it is said that its operation is network efficient. For example, by statistically multiplexing bursty transmissions, the bandwidth required can often be reduced from that of a pure circuit-switched approach while still meeting the delay requirements of the applications.

On the other hand, if no customer currently receiving a particular QoS values it less than another customer who is being denied that QoS, it is said that the operation is economically efficient. For example, if one customer is willing to pay \$X per second for

un-delayed access to a 1 Mbps link, and a second customer is willing to pay only $\$X/2$, and if only one of them can be accommodated, then in an economically efficient network the bandwidth will be allocated to the first customer.

2.1.4 Fairness

In telecommunication networks, a resource allocation scheme seeks to allocate a limited set of resources to a set of individuals with demands. For example, assume a network with X customers each seeking to transmit data between some pair of nodes. The resource allocation scheme must decide how to allocate a portion of the available bandwidth to each customer. A common approach to such resource allocation problems is to allocate resources based on some fairness figures. It is important to note that fairness in resource allocation becomes an issue only when there are unsatisfied demands and customers have to compete for their share.

2.1.4.1 Max-Min Fairness

Although several definitions of fairness arise from various disciplines, in the networking world the most popular notion is indeed that of max-min fairness (also referred to as the classical notion of fairness) [10]. In max-min fairness, the available link bandwidth is equally shared among connection which are bottlenecked at the link.

Max-min fairness is a widely used technique for resource allocation in cases where some customers' demands are smaller than others. It operates as follows:

- Resources are allocated in increasing order of demand.
- Customers are never allocated shares higher than their demand.
- All customers with unsatisfied demands are allocated equal shares.

Initially all customers get at least as much as the small customer demands. The remaining resources are evenly distributed among the customers with unsatisfied demands. It follows that those customers with unsatisfied demands cannot increase their share without decreasing the share of a customer with an already small one.

2.1.4.1.1 Weighted Max-Min Fairness

Fairness should not necessarily imply equal distribution of resources to all those customers with unsatisfied demands. Under certain criteria, it may be justifiable to allocate more resources to some customers than to others; this leads to a generalization of max-min fairness [41]. Weighted max-min fairness generalizes the concept of max-min fairness for the case where customers have different rights to resource allocation. Each customer i is associated with a weight w_i which reflects its right to a relative resource share. The weighted max-min fair allocations are calculated as follows:

- Resources are allocated in increasing order of demand normalized by weight.
- A customer is never allocated a share higher than its demand.
- Customers with unsatisfied demands are allocated shares in proportion to their weights.

2.1.4.2 Social Welfare

Social welfare is a special type of fairness based on adjusting the customers' utility functions. Two types of welfare fairness are commonly used in the literature: Utilitarian Criterion and Equality Criterion [32].

- **Utilitarian Criterion**

Utilitarian Criterion, sometimes referred to as utility maximization, is a Pareto-Optimal allocation that results in the greatest sum of the customers' utilities

- **Equality Criterion**

Equality Criterion is a Pareto-Optimal allocation that results in an equal level of utility for all customers.

The Utilitarian Criterion and Equality Criterion are utilized later in Section 6.2.1 and 6.2.2.

2.1.4.3 Proportional Fairness

Proportional fairness is another type of fairness that is utilized by the CDMA-HDR (High Data Rate) scheduler [53]. In wireless networks, the customer's channel condition is a primary factor in determining the data rate that the customer can attain. The CDMA-HDR scheduler takes advantage of the wireless channel variability, which results in variations of the requested rate over a period of time. Instead of serving customers based on first in first out scheme, the scheduler gives higher priority to customers with good channel

conditions. The scheduler priority is based on a combination of the signal to interference ratio as well as the duration of time since the customer had last been served.

The scheduler serves customers that are near their peaks in terms of the requested rates. Occasionally, the customers may not be served for periods of milliseconds when their requested rates are lower. By selecting the optimal time to transmit data to customers, the scheduler provides customers with average data rate higher than if they were served on a first in first out basis.

For example, a particular customer has signal conditions that support an average of 614.4 Kbps. The changes in the environment surrounding the customer cause the channel conditions to oscillate between low and high data rates with an average of 614.4 Kbps. The scheduler serves the customer only when his channel conditions can support 614.4 Kbps or higher.

2.1.5 Auctions

An auction is a method of allocating scarce resources. A method that is based upon competition where a seller wishes to obtain as much money as possible, and a buyer wants to pay as little as necessary. An auction offers the advantage of simplicity in determining market-based prices. It is efficient in the sense that an auction usually ensures that resources are allocated to those who value them most highly and ensures also that sellers receive the collective assessment of the value.

An auction is a very good way to sell resources in a competitive environment where the buyers frequently know more than the seller about the value of the item. Usually the seller does not want to suggest a price first out of fear that his ignorance will prove costly, and so holds an auction to extract payment he might not otherwise realize.

There are different ways to classify auctions. There are open auctions as well as sealed-bid auctions. There are auctions where the price ascends and auctions where the price drops at regular intervals. Generally, experts agree that there are four major one-sided (i.e. only bids are permitted, but not asks) auction formats: English, Dutch, First-Price sealed-bid, and Uniform Second-Price (Vickrey auction) [60].

2.1.5.1 Uniform Second-Price Auction

In Uniform Second-Price format [69] [58], the bids are sealed and each bidder does not know the bids or the bidding strategies of other bidders. The item is awarded to the highest bidder at a price equal to the highest unsuccessful bid. In other words, a winner pays less than the highest bid. If, for example, bidder A bids \$10, bidder B bids \$15, and bidder C offers \$20, bidder C would win, however he would only pay the price of the second-highest bid, namely \$15. When auctioning multiple units, all winning bidders pay for the items at the same price, which equals the highest losing price.

In a Uniform Second-Price auction no one is discouraged out of fear that he will pay too high a price. Aggressive bidders receive sure and certain awards but pay a price closer to

market consensus. The price that the winning bidder pays is determined by competitor's bids alone and does not depend upon any action the bidder undertakes.

In this research we are interested most in the general form of the uniform Second-Price auction in which the M highest bidders win and pay a uniform price equal to the $M+1^{st}$ highest bid [58].

2.2 The Evolution Toward High Speed Wireless Internet

While 1st generation analog wireless networks are still deployed in some areas, the 2nd generation dominates today's wireless market. The digital 2nd generation networks (e.g. GSM, TDMA, and cdmaOne,) provide service providers with higher capacity and network coverage, and customers with better voice quality.

Although they enjoyed tremendous success over the last few years, the 2nd generation networks have been designed and optimized for circuit switching voice calls only, with minor circuit switching low bitrate (9.6 and 14.4 Kbps) data access. At the same time, the incredible success of the Internet and the increasing popularity of multimedia applications (e.g. videoconferencing, video on demand and voice streaming) have increased the demand for higher bitrate and packet switching services in wireless networks.

In order to achieve that, a new generation of wireless networks has been proposed. The new third generation networks will offer at least 144 Kbps for high-speed customers, 384

Kbps for low-mobility outdoor customers, and up to 2 Mbps for indoors customers. At the same time, 2.5-generation technologies (e.g. GPRS and EDGE) have been proposed to fulfill the immediate need for higher bitrate services. Those technologies offer rates up to 384 Kbps.

Table 2-1 and Table 2-2 illustrate the peak data rates for European and American technologies as they have evolved to the third generation.

	2 G	2.5 G			3G
Technology	GSM	GPRS	HSCSD	EDGE	UMTS
Air Interface	TDMA	TDMA	TDMA	TDMA	WCDMA
Max bitrate	9.6 Kbps	21.4 Kbps	14.4 Kbps	384 Kbps	2 Mbps
Switching	Circuit	Packet	Circuit	Both	Both

Table 2-1: European Path Toward 3G

	2 G	2.5 G		3 G
Technology	CdmaONE	1xRTT	1xEV	3xRTT
Air Interface	CDMA	CDMA	CDMA	WCDMA
Max bitrate	14.4 Kbps	114 Kbps	2.4 Mbps	2 Mbps
Switching	Circuit	Both	Both	Both

Table 2-2: American Path Toward 3G

2.2.1 QoS in Wireless Networks

Wireless networks differ from wireline networks in the way they guarantee the QoS level provided to admitted terminals throughout the call lifetime. In wireline networks, the resource control complexity arises from the need to allocate finite network resources among multiple traffic flows while satisfying their QoS requirements. However, using appropriate admission control, resource reservation, and resource scheduling, a wireline network can assure a QoS to multiple data terminals. Furthermore, once the call has passed the admission control test, the QoS offered will not change during the call lifetime. Therefore wireline networks usually offer “guaranteed” QoS services to their customers.

Things, however, are different in mobile wireless networks. In the mobile wireless environment, even in the absence of any interference, the quality of the radio channel may undergo dramatic and rapid changes due to phenomena such as multipath reflections and fading as the terminal moves. As a result, the radio signal arriving at the receiver shows time varying distortion and an increase in the bit error rate (BER). This period of fading lasts for a random time duration that depends on the speed of the mobile as well as the existence of physical objects (e.g. high buildings) in the environment. If any terminal encounters this situation and the BER becomes higher than the threshold specified in the QoS contract, it should be handed off to another channel with better signal quality or even to another base station.

Moreover, in mobile wireless networks the availability of wireless resources at the connection setup time does not necessarily guarantee that the resources will be available throughout the connection lifetime. Due to the limitations of the radio spectrum, future wireless systems will use micro/pico-cellular architecture in order to provide a higher system capacity. Because of the small coverage area of the micro/pico-cells, handoff events will occur at much higher rates than in today's networks. It is highly possible that a terminal can be handed-off to new base station that does not have enough resources to accommodate the arriving call. At that time, if the terminal cannot adapt its bandwidth requirement and accept the smaller bandwidth offered by the new base station, the call will be terminated.

Because of these limitations, it is not practical to provide guaranteed QoS services to wireless applications [80], [24]. Instead, a "*soft QoS services* " approach is needed, where QoS management negotiates a range of acceptable values with adaptive multimedia applications.

2.2.2 Wireless QoS Classes

Four QoS parameters have been used to classify the QoS services offered by the wireless network: bitrate, frame error rate (FER), end-to-end delay, and delay jitter. Based on these parameters, the 3GPP committee has identified four QoS classes for next generation wireless networks as shown in Table 2-3 [1], [2].

Traffic Class	Class Description	Example	Relevant QoS
Conversational	<ul style="list-style-type: none"> - Preserves time relation between entities making up the stream - Conversational pattern based on human perception - Real-time 	<ul style="list-style-type: none"> - Voice over IP - Video conferencing 	<ul style="list-style-type: none"> - Low jitter - Low delay
Streaming	<ul style="list-style-type: none"> - Preserves time relation between entities making up the stream - Real-time 	<ul style="list-style-type: none"> - Real-time video 	<ul style="list-style-type: none"> - Low jitter
Interactive	<ul style="list-style-type: none"> - Bounded response time - Preserves the payload content 	<ul style="list-style-type: none"> - Web browsing - Database retrieval 	<ul style="list-style-type: none"> - Round trip delay time - Low BER
Background	<ul style="list-style-type: none"> - Preserves the payload content 	<ul style="list-style-type: none"> - Email - File transfer 	<ul style="list-style-type: none"> - Low BER

Table 2-3: UMTS QoS Classes

The QoS parameters recommended by the IEEE wireless broadband access (WBA) for each of these classes are shown in Table 2-4 [44].

Medium	APPLICATION	DATA RATE	ONE-WAY DELAY	JITTER	FER
Audio	Conversational voice	4-13 Kbps	<150 msec preferred <400 msec limit	< 1 msec	< 3%
Audio	Voice messaging	4-13 Kbps	< 1 sec for playback < 2 sec for record	< 1 msec	< 3%
Audio	High quality audio streaming	32-128 Kbps	< 10 sec	< 1 msec	< 1%
Video	Videophone	32-384 Kbps	< 150 msec preferred <400 msec limit		< 1%
Video	Video streaming	32-384 Kbps	< 10 sec		< 1%
Data	Web-browsing	Unspecified	< 4 sec /page	N.A	Zero

Table 2-4: IEEE WBA Recommended QoS Parameters

The Third Generation Partnership Project (3GPP) has selected the adaptive multi-rate (AMR) speech codec [27] as the mandatory codec for next generation wireless networks

[3]. The 3GPP has also selected the ITU-T H.263 [51], [52] and the MPEG-4 [46]-[48] as the standard codecs for next generation video transmission [28].

2.3 Subjective QoS

How customers think about the QoS they receive from a communication network is very different from how the network defines this QoS. Customers usually use subjective parameters to judge the perceived QoS. When listening to some form of audio media, they judge quality based on different parameters such as how good it sounds, whether there are scratches, pops, and hisses or clear sound, whether the playback is smooth or choppy. When watching video, they judge the resolution of the images, if there are any imperfections in them, and if the playback is fluid or has inappropriate starts and stops. In the case of any interactive media such as a conversation or videoconference, latency and ETED become extremely important.

On the other hand, network engineers use objective parameters to define the service quality. They deal with statistics about the stream throughput, the error rate suffered by the transmitted data in terms of the number of bits that are corrupted and the number of packets that are lost. They also define QoS based on the stream delay characteristics, both the latency from the time a packet enters the network to the time it is delivered and the variability of delay between each packet..

There is an obvious disconnect between these two perspectives of network QoS. Since both describe the same thing, QoS mapping techniques are usually used to relate one to the other. One of the more popular approaches in the literature to the problem of mapping subjective to objective QoS is the use of opinion rankings [7]. Since customer's QoS perception is often poorly understood and each individual customer has unique preferences, the ITU has defined Mean Opinion Scores (MOS) as an accurate way to represent customer subjective QoS.

This is accomplished by doing careful subjective tests [49], [50] on a MOS for the particular application as a function of the QoS parameters. For example, the ITU P.800 is a subjective MOS of voice quality. It is derived from an evaluation of various pre-selected voice samples over different transmission media, replayed under controlled conditions, to a mixed group of men and women. This mixed group of men and women score the voice quality of voice samples by rating them from 1 (worst) to 5 (best). Scores are then weighted to derive a single MOS score rating.

This method is effectively a measurement-based perceptual model. One advantage of this approach is that the available selections can be described relatively in terms any customer would easily understand (e.g. excellent, good, fair). However, it makes the assumption that an accurate representation of an average customer can be generated, which is not entirely possible. There will always be cases where a specific customer's perception of a

given service will be very different from the aggregate preference, and in these situations this method will be inadequate, but typically it should be acceptable.

For example, in [49] several tests have been conducted to measure the subjective quality of audio codecs in the range of 6-56 Kbps data rate. Each codec is tested with several audio samples include speech, single instrument a, pop, classic, and complex. Testes rate the perceived audio quality for each codec according to ITU BS.562.3 recommendation.

Chapter 3

REVIEW OF THE STATE OF THE ART

3.1 Charging in Wireline Data networks

Charging and accounting play a very important role in the success of today's high-speed multiservice networks deployment. Charging policies are needed for a network to recover its costs in a competitive and fair way from the diverse population of customers [30]. It is important to emphasize that setting prices for network services has always been a marketing and strategic decision rather than an engineering concern. It is widely believed that an appropriate charging policy will provide incentives for customers to behave in ways that improve overall utilization and performance of the network [16].

Service providers usually charge customers according to the following general equation [30]

$$\text{Charge} = \alpha \cdot \beta \cdot T + FC + \sum_i SC_i \quad \text{Equation 1}$$

Where

- β is the amount of resources used.
- T is the call duration.
- α is the tariff parameter, which represents the cost per unit of resource used per unit time.
- FC is the fixed cost and is normally attributed to connection set-up or cost of negotiation, e.g., the switch resource used during connection set-up.
- SC_i is the additional charge for facilities (such as connection holding and automatic answering) that a typical network operator might offer to its customers.

The subscript “ i ” represents the service identity since there may be a number of subscription costs included in a bill.

Based on the tariff parameter (α) the charging policy could be classified as dynamic “($\alpha(\tau)$ ” , in which prices fluctuate as a result of some network condition, or static, in which prices are independent of network load.

3.1.1 Static Pricing Policies

The price in this category is set in the contract between the customer and the service provider. The time scale of this price change is much longer than the session duration. Prices normally do not change simply because of instantaneous congestion within the network, but rather due to long-term observation of network usage and market conditions.

For example, the service provider will give advance notice of lower prices to stay competitive, or raise prices to meet increasing costs.

The static pricing parameters can be modified by other session characteristics. Three price modifiers, which are commonly used in static pricing, are time-of-day, destination, and usage. With the time-of-day modifier, the price depends on the calendar time, such as the time of day (peak or off-peak), day of the week (weekday or weekend), or public holiday. It is a form of congestion pricing based on long-term observations. With the destination modifier, the price depends on the distance between the sender and the receiver. It can be based on the number of hops traversed or simply the location of the receiver. With the usage modifier, the price depends on the amount of usage, i.e. the duration of a session or the amount of data transmitted. This is to encourage long sessions in order to minimize the overhead of session setup process, or to encourage high volume customers.

Several static pricing policies have been proposed in the literature. Three of them are: flat rate, Paris Metro, and priority pricing.

3.1.1.1 Flat Rate Pricing

In a flat rate policy, a fee is charged for a predefined amount of bandwidth to access the network. This fee is fixed and varies neither with the actual usage of the bandwidth nor with the state of congestion on the network. The main appeal of flat rate pricing for customers and service providers is its simplicity, which reduces risks and administrative costs. Predictable fees reduce risks for customers, and to some extent for service

providers, by providing them with known expectations for payments, which facilitates planning and budgeting. In addition, flat rates avoid the potentially considerable administrative costs of tracking, allocating, and billing for usage.

However, the foreseeable consequence of a flat rate pricing policy is the emergence of a “tragedy of the commons”, which arises when a common resource is degraded by over-usage. Well-known examples are ocean fisheries, urban roads, air and water. Whenever a public good does not belong to some legal entity empowered to manage the resource through usage restrictions and/or fees, there are inadequate incentives for individual customers to restrict usage to the socially optimal level. Consequently, the public good deteriorates, and the public suffers a loss relative to the potential social benefits of the commons [39].

Using a flat rate policy, each customer logs in without the incentive to limit his usage, since he has already paid the fixed fee to access the network. He is only concerned with his costs and benefits. Eventually, congestion occurs and another potential customer is made to wait. The result is that resources are allocated by time rather than through pricing among customers who have paid the fixed access fee. It is important to note that the impatient customers who will voluntarily leave the congested network are not necessarily the low-value customers of the network. The implication of this is the loss of social welfare.

In a nutshell, a flat rate pricing policy uses a fixed announced price list for each service offered by the network. The price list for each service depends on static factors such as distance and expected day/night traffic and it has nothing to do with the actual traffic. It is very simple to implement, however it doesn't utilize the network properly and it forces the customers to overbook the network resources in order to guarantee a certain QoS level.

3.1.1.2 Paris-Metro Pricing

To improve the performance of the flat rate pricing and to avoid the complexity of the dynamic charging policies, the Paris Metro Pricing (PMP) policy has been introduced in [73]. Until some years ago the Paris Metro system offered transportation in first and second-class cars with identical quality and number of seats but a first class ticket costs twice as much as a second-class ticket. The system operates in equilibrium with the first class cars being less crowded than the second class.

The Paris Metro system inspired engineers at AT&T Labs to enhance the flat rate pricing policy by adding new classes of services to their network. In a PMP network, each link is divided into a maximum of four channels (classes). In the basic version of PMP, each channel is then assigned a fixed price and a fraction of the link capacity. Channels with the same associated price construct a virtual sub-network. Customers have the choice of selecting the network to use given that QoS will be higher in the more expensive sub-

network. To improve stability and to fit customer's preference for simplicity, the assigned capacity and prices are assumed to be constant for long intervals.

Although there are several prices in the network, PMP is still considered a static pricing policy because prices do not change with changes in the resource utilization. Like flat rate pricing, PMP is simple and easy to implement. However, PMP does not guarantee QoS even to customers using the highest price sub-network. It also reduces the link statistical multiplexing gain by dividing the link capacity into several channels.

3.1.1.3 Priority Pricing

A novel static pricing policy called "priority pricing" has been introduced in [38]. In this policy, the service price is set in such a way that the performance penalty received for requesting a less than optimal service class is offset by the reduced price of the service. At the same time, they do not make optimal service classes so expensive that even performance sensitive customers do not use them.

In their research, Gupta et al. argue that there are mainly two decisions that have to be made by each gateway or router in the network. When the line is free and there are packets in the queue, the gateway must select the next packet to transmit. When a packet arrives and there is no room in the queue, the gateway must decide which packet to discard. For the simplest multiservice discipline, they introduced two priority levels. Thus, on each packet there is a priority service flag and a priority no-drop flag. There are then four service classes, corresponding to the four possible combinations of flag settings.

The gateway gives preference in service to those packets with the priority flags on. Within a priority class, service is provided in a FIFO order. Thus, when the line is idle, the earliest arrived packet that has the service priority flag on is serviced. If there are no such packets, then the earliest arrived packet without the service priority flag is serviced. Packets are discarded in the following order: both priority flags off, service priority flag on and no-drop flag off, service priority flag off and no-drop flag on, and finally both priority flags on. A per byte charge is then applied based on the priority flag set in the packet header.

Simulation results [38] show that under priority pricing, customers are motivated to choose the appropriate service class, and the multiservice discipline enables the performance sensitive applications to achieve better performance. However, the scheme requires a very complicated billing subsystem. In order to charge customers fairly, the billing subsystem should count and check the header for each transmitted packet. This scheme is also unfair when the line is idle because, although both the high performance and the low performance customers will receive the same service, the higher performance customer will be charged more for setting the priority flags on.

3.1.2 Dynamic Pricing Policies

Dynamic policies are more complex and in general require more sophisticated accounting and processing; consequently, they tend to be more costly to implement than static pricing policies. Their biggest advantage is that they offer flexibility to react to changes in offered

traffic and are better equipped to track the optimal prices to be charged by the network at any given time. A common criticism of dynamic pricing mechanisms is that they often require a level of computational complexity that may be impractical [23]. Even more important, there are substantial barriers to customer acceptance of dynamic pricing: customers may find such policies difficult to understand and to budget for. This explains why virtually all pricing policies now in place for multiservice networks are of a static nature.

The usage-fee price in this category could be calculated based on the demand on the network resources as in usage-based policies or the congestion level as in congestion-sensitive policies.

3.1.2.1 Usage Based Policies

Usage-based pricing controls the resource usage by imposing a fee based on the amount of data actually sent. It imposes usage costs regardless of whether the network is congested or not.

3.1.2.1.1 Expected Capacity

A novel pricing policy based on customer's expected capacity has been introduced in [15]. In this policy, the customer's expected capacity during both normal and heavy traffic periods is defined in a contract with the service provider. The customer's expected capacity could be defined in several ways, such as a minimum steady rate or the maximum data object size. In general, it can be defined in any way that the provider and

consumer agree on. The provider then must provision the network resources to carry the expected capacity of his subscribers during both normal and heavy traffic periods.

Unlike other charging schemes, the expected capacity charging scheme charges customers based on the expected capacity specified in the service contract and not on the actual capacity. Charging customers based on their expected capacity reduce the scheme overhead by eliminating any need for traffic measurements. From the customer perspective, charging customers based on their expected capacity meets the needs of many customers for fixed or predictable costs.

Based on the expected traffic contract, customer's packets will be tagged either inside (in) or outside (out) of the expected capacity envelope. To implement a pure expected capacity scheme, this tagging always occurs, whether or not there is congestion anywhere in the network. At any point of congestion, the packets that are tagged as being out are selected to receive a congestion pushback notification. If there is no congestion, of course, there is no discrimination between in and out packets; all are forwarded uniformly.

Although the expected capacity scheme reduces the billing overhead significantly by eliminating the need for traffic measurements, it is still not suitable for the next generation wireless networks. The scheme is only fair for customers who have an excellent knowledge about their applications traffic patterns and can accurately predict their usage during both normal and heavy traffic periods over a long period of time. If the customer does not have enough knowledge or cannot predict his traffic pattern over a long period of

time, he will end up paying for capacity that he will not ever use or experience a less than expected QoS service.

3.1.2.1.2 Effective Bandwidth

The effective bandwidth associated with a stationary source has been defined in [55] as follows: Assume that $X[0,t]$ is the amount of traffic received from a source in an interval of length t . Assume also that $X[0,t]$ has stationary increments, then the effective bandwidth of the stream is defined as:

$$\alpha(s,t) = \frac{1}{st} \log E[e^{sX[0,t]}]$$

where s, t denotes the system parameters which depend on the characteristics of the multiplexed traffic and the link parameters (capacity and buffer size).

Several charging schemes have been proposed based on effective bandwidth calculations. In [79], each connection is charged with a price a per unit of time, a price per bit b and an optional connection price c . The parameters a and b depend on the traffic characteristics negotiated during call setup following the concept of effective bandwidth.

At connection set-up the customer declares a mean rate m for the new connection. It is assumed that the customer has a prior knowledge of the source characteristics. The a and b parameters are chosen such that the straight line $a + b \cdot x$ defines the tangent of the effective bandwidth function at the point $x = m$ (Figure 3-1). If the customer actually

sends traffic with a mean rate M different from m he is punished with a charge higher than that he would have to pay when he had declared the mean rate to be M .

Figure 3-1: Effective Bandwidth Calculation

This scheme has the following drawbacks:

- The customer has to declare the mean rate. So customers who are not able to predict the mean rate of their application are punished.
- The network has to measure the mean rate M for each connection, which consumes a significant processing power.
- A customer who is not familiar with the effective bandwidth concept will not understand the reason and the extent of the punishment.

3.1.2.2 Congestion Sensitive Policies

Congestion-sensitive policies adjust the resources usage price dynamically based on the current resources utilization level. This price could then be used as a feedback signal.

During heavy traffic periods, the price feedback signal should discourage adaptive customers from sending more traffic. While during light traffic periods, the feedback signal should encourage these customers to send any traffic they have ready to transmit.

By associating a cost measure with network loading, all customers can be signaled with the prices necessary to recover the cost of the current network load. Price-sensitive customers can choose whether or not to transmit their traffic according to the individual willingness to pay the current price. Customers who value network service more than the current price will choose to transmit, while those who value it less will wait for a lower price. When the network is lightly loaded then the price will be close to zero, and all customers can input traffic.

3.1.2.2.1 Responsive Pricing Scheme

Customers of current data networks respond to multiple forms of feedback signals. Customers decide whether or not to use a particular network, perhaps based on the network's charging structure or their previous experience with it. Furthermore, if customers observe that the network is usually heavily loaded at certain times of the day and lightly loaded at others, they may schedule their network usage accordingly.

The responsive pricing scheme has been introduced in [67] in which the service price is used as a new feedback control signal. The responsive pricing scheme measures the buffer occupancy at the gateway. This occupancy is converted into a price per packet,

which is then transmitted back to each active adaptive application. The applications then decide on how many packets to send during the next interval.

The overhead in this scheme includes the measuring of the network congestion level, adjusting the resource price based on the measured congestion level, and negotiating the amount of resource allocated to each customer based on the newly calculated resource price. The resource negotiation may take several iterations until the customers and the charging scheme agree on a balanced resource price, which in turn consumes a considerable amount of time and signaling overhead.

3.1.2.2.2 Smart Market Scheme

Smart Market is a congestion-sensitive pricing policy proposed in [66]. Under this policy, the actual price for each packet is determined based on the current state of network congestion. The policy assumes that the network is not congested most of the time; at such times the price for usage is set to zero. However, when the network is congested, packets are queued, delayed, and dropped. Instead of using FIFO queuing, the smart market uses priority based queuing algorithms. Customers are expected to bid a price for their packets. Packets are then prioritized based on the bids the customer associated with their packet. In order to make the policy incentive-compatible, customers are not charged the price they bid, but rather the bid of the highest bid packet that is not admitted to the network (second price auction format).

Simulation results [66] show that this mechanism provides the right incentive for customers to reveal their true priority. The basic idea of the second price auction is that the price a customer pays is not determined by the priority he sets, but by the bid of the first packet that is rejected from the network. This means that the customer has no incentive to misrepresent the true valuation of his packets. The price for the packet admission set by the smart market is effectively the value of the packets that are not admitted to the network. If the total value of those packets exceeds the cost of expanding the network in order to handle them, then it is appropriate to expand the network capacity.

Despite its advantages, the smart-market scheme is not suitable for wireless environments for the following reasons:

1. It assumes that most of the time the network is not congested, which is not true in the case of a wireless limited resource environment.
2. Attaching a bid to a packet transmitted by the customer introduces unnecessary overhead and consumes lots of scarce network bandwidth.
3. Processing each packet to determine its priority requires a huge amount of processing power.
4. Since it depends heavily on customer's interaction, only customers with a good understanding of the resource allocation and charging policies, and their applications requirements could benefit from this policy.

3.1.2.2.3 Auction Based Pricing

Instead of attaching a bid with each transmitted packet as in the smart market technique, the auction-based pricing technique [68] combines a budget function provided by each customer with an auction based admission control and bandwidth allocation. The budget function represents the upper bound on the bandwidth price the client is willing to pay, as a function of time. In order to simplify the customer budget function, the proposed scheme assumes that the customer's budget function increase linearly with the call lifetime.

The admission control process allocates bandwidth to calls based on the competitiveness of their budget function. During the peak periods, the service provider runs an auction that uses the budget functions of the competing calls to determine the amount of bandwidth allocated to each call.

Although the amount of overhead has been reduced significantly in the auction-based policy compared to the smart market policy, introducing the budget function concept increases the system complexity and requires intelligent customers with good understanding of the whole system. Moreover, using a non-decreasing time based budget may cause a large number of calls to be dropped during their early stage.

3.1.3 Shadow Pricing

To highlight the concept of the shadow price consider the case where there is unemployed labor in the economy. The cost of using that labor to society is virtually zero. By employing it no sacrifice is made in terms of other goods produced. The shadow price of labor at this time is zero, even though the workers, if employed, would have to be paid a wage. Alternatively suppose that labor is in short supply. In this case, employing a worker may cost a firm only the going wage, but the cost to society of that firm employing that worker is the production the worker could have produced in his previous occupation. The shadow price of labor in this case reflects the benefit that would result from relaxing the labor constraint by one unit.

As mentioned above, the marginal cost of carrying data is zero when the network is uncongested. However, in case of congestion, the charge that customers should pay is a shadow price that reflects the cost to other customers through increased congestion. A charging scheme based on the shadow price concept has been introduced in [56].

3.2 Charging in Wireless Data Networks

Second generation wireless service providers charge customers equally. Customers are charged based on a per minute access time scheme that gives them a limited amount of access time for a predefined fixed price. Service providers may also charge customers an

extra price if they exceed their maximum allowable access time. Service providers give customers the choice to select unlimited access for a considerable higher price.

Although charging and accounting schemes are very important for the success of the network operation, no multiservice charging schemes have been proposed yet for the next generation wireless networks. Third generation service providers may charge data customers based on the currently used per minute of connection time charging scheme or use a new scheme such as per Kbyte of transmitted traffic or a flat rate scheme.

3.3 Charging Scheme Evaluation Criteria

Since several charging schemes have been proposed for telecommunication networks, it is important to set evaluation criteria to compare them. Two evaluation criteria have been introduced in [63] and [75]. The most relevant criteria to this research are as follow:

1. Overhead and Complexity:

How much network and computing resources does the scheme consume? Overhead might not be an important issue in wireline networks however it is very important for wireless networks. At the same time, computational complexity is very critical especially for handheld devices.

2. Predictability

Most customers and service providers appear to want a charging scheme that offers “no surprises”. Is the tariff known before a connection is made? How easy is it to predict the long-term charge? This could change if the charging scheme involves a periodic re-negotiation process with the network operator.

3. Flexibility

Can the charging scheme be extended or modified to accommodate the charging objective and requirements of the operators? Can the charges be carried out easily and in a way that is understandable, fair and equitable to all interested parties?

4. Practicality

How easy and obvious is the implementation of the charging scheme?

5. Control

In a competitive environment the charging scheme is an important factor in the operability of a network. It alone will determine how the network is utilized and the ability of the operator to recover costs and stay in business. Therefore charging schemes must help controllability of the network.

6. Choice

Does the charging scheme offer real choices to the customers in terms of different modes of operation according to the customer willingness to pay more or less? For

example, can the customer select a low quality of service and get a cheaper deal for the same service?

7. Customer interaction

Does the scheme require any customer interaction? If yes, how often? It is hard to believe that someone will keep watching the dynamically changing price of a call waiting for the right moment to make his call. It is also hard to believe that a customer will adjust the required QoS or the call price according to the changes in network state. On the other hand, videoconference applications may automatically renegotiate their QoS parameters as congestion builds up. A videoconference session, depending on the customer configuration, might accept some sort of resolution degradation or even switching from color to black and white when the transmission price is high.

8. Technical Knowledge Requirements

Does the customer need to know about the application QoS requirement and traffic statistics? Does he need to know about the underlying network resources and the effect of each resource on his application?

Table 3-1 evaluate the wireline data networks charging schemes based on the above evaluation criteria.

	STATIC SCHEMES		DYNAMIC SCHEMES		
	FLAT RATE	PRIORITY	EXPECTED CAP.	RESPONSIVE	SMART MARKET
Overhead	Very low	High	Low	Very high	Very high
Predictability	High	High	High	Low	Low
Flexibility	Low	Low	Low	High	High
Practicality	High	Low	High	Low	Low
Control	Low	Low	Low	High	High
Choice	None	Little	Little	Little	Much
Interaction	None	Low	None	High	High
Knowledge	None	Low	High	High	High

Table 3-1: Charging Schemes Evaluation

3.4 Resource Control in Wireless Data Networks

In order to manage QoS in multimedia wireless terminals, the call admission control and the resource allocation schemes should answer the following questions:

1. What admission control procedure should be used to limit the number of admitted calls in order to guarantee QoS?
2. How should physical resources be allocated to different terminals in a fair and efficient way such that the dropping and the blocking probabilities for both new and handoff calls for all traffic types are minimized?

In the remainder of this section, we will highlight some of the resource control schemes currently proposed for the third generation wireless networks.

3.4.1 Class Based Resource Allocation

A class-based resource allocation scheme has been described in [8]. This scheme differs in the way that non real-time applications (e.g. ftp, e-mail) use the bandwidth that is not used by real-time applications (e.g. video, audio). The total number of channels “ Z ” available at the base station is divided into two parts, one for real-time applications “ X ” and the other for non real-time applications “ Y ”.

- Complete Partitioning (CP)

In this scheme, the base station capacity is completely partitioned where real-time applications use up to X channels and $Y = Z - X$ channels are assigned to the data applications

- Real-time Complete Access (CA)

In this scheme, real-time connections can use up to the total BS capacity ($X \leq Z$) with preemptive priority over the non-real-time applications. At any given time, channels unutilized by real-time applications are equally shared by all non real-time applications.

- Real-time Restricted Access (RA)

In RA scheme, real-time connections can use up to X channels with preemptive priority over non-real-time connections and $Y = Z - X$ channels are dedicated to non-real-time applications. Similar to the CA scheme, the channels unutilized by real-time applications are made available to non-real-time applications at anytime.

Experimental results [8] show that the CA and the RA scheduling mechanisms provide better QoS than the CP scheme. This is because channels that are not utilized by either real-time or non real-time traffic are wasted in the CP mechanism. The results show also that the RA scheme performs better than the CA scheme in that the desired QoS of real-time calls can be achieved by allocating them a capacity smaller than the total base station capacity. This result in more bandwidth being available to non real-time connections. In other words, by making the total base station capacity available to real-time connections, a QoS better than the objective can be achieved at the expense of an unacceptable QoS to non real-time connections.

3.4.2 Graceful Degradation of Service

In order to decrease the new and handoff call blocking probability, the resource allocation algorithm should try as much as it can to accept the admitting call even if it does not have enough resources to accommodate it. The resource manager can keep the blocking probability under a certain limit by degrading the QoS provided to active customers.

In the case of bandwidth, the resource manager will degrade the QoS by discarding some of the application packets. In order to degrade the QoS gracefully, the resource manager should know which packets it should discard for each application. For example, should it discard every other packet, or discard packets in a burst fashion? The answer for this question depends heavily on the nature of the application. One approach to providing the resource manager with the required information is to keep a packet-discard profile for

each application type at the base station. In case of congestion, the network controller could access the applications records and discard packets in accordance.

The major drawback of this solution is that it requires the resource manager to be application aware. The resource manager must create a profile for each application currently running in the network. It should also update its database every time a new application is introduced, which is not applicable in future multiservice networks.

Another approach to solving this problem has been proposed in [78]. In this approach the customer specifies, during setup time, a preferred way in which his data should be discarded in case of congestion.

As mentioned above, bandwidth degradation is required in order to admit more incoming calls. However, bandwidth degradation translates to poor QoS which may not be acceptable to customers. An analytical model to compute the service degradation balance point has been introduced in [22]. The model measures the positive revenue generated by the increase in the carried traffic (number of calls) and the negative revenue generated by the bandwidth degradation. It then calculates the payoff generated by bandwidth degradation.

3.4.3 Adaptive Bandwidth Allocation in Wireless Networks

An adaptive admission and bandwidth allocation scheme for multiservice wireless networks has been presented in [74]. In this scheme, multiservice applications have been

classified as class-I (real-time) and class-II (non real-time). The scheme assumes that the following information will be available during a call setup: the call type, the call desired and minimum amount of resources, and the maximum acceptable connection dropping probability.

During call setup, the scheme attempts to allocate the desired amount of bandwidth in the cell where the new connection is generated. If the desired amount of bandwidth is not available, the new connection is rejected. In case of class-I calls, the scheme goes one step further by reserving a fraction of the required bandwidth in all neighboring cells. The amount of resources allocated in the neighboring cells is calculated as a function of the requested amount of bandwidth or as a function of the number of class-I calls.

The scheme gives higher priority to handoff calls over new calls. A new call is admitted only if there is available bandwidth equal to the desired amount of bandwidth, whereas a handoff call is admitted if only the minimum amount of resources is available

This scheme has the following major drawbacks: first, reserving resources for class-I traffic in all neighboring cells is a huge waste of the scarce network resources. Second, the scheme does not give a solution for the handoff of class-II connections. It simply assumes that class-II connection can wait until there are enough resources in the new cell. Third, the scheme does not consider the effect on the resource reservations in the neighboring cells on the new call blocking probabilities in those cells, especially in case of unbalanced traffic loads.

Another bandwidth adaptive framework has been proposed in [59], [61] and [62]. The framework manages the system resources using admission control and resource allocation and reallocation algorithms. The authors introduce the degradation period ratio (DPR) QoS parameter, which represents the fraction of a call's lifetime during which the bandwidth allocated to the call is below a predefined value. The aim of the research was to minimize the number of calls with lower than predefined “target bandwidth”.

The Call Admission Control (CAC) algorithm predicts the number of calls in a given cell after an estimation period T based on the expected number of handoff calls from neighboring cells and the expected number of remaining calls in the given cell. The CAC then calculates the cell overload probability, which is the steady state probability that there is at least one call that has lower than the target bandwidth. If the cell overload probability is expected to be greater than the predetermined QoS value after T time units, then the originating call is rejected.

The aim of the resource allocation scheme is to maximize the number of calls with at least the target bandwidth at any instant. A resource reduction scheme applies to the case where a new or a handoff call arrives in the given cell. Depending on the situation, it allocates suitable bandwidth to the incoming call and adjusts the amount of bandwidth allocated to existing calls, if necessary. In the case of an outgoing handoff call or call completion in the given cell, a resource expansion scheme expands the bandwidth allocated to calls with lower than target bandwidth to the target bandwidth or more.

3.4.4 QoS Provisioning Framework

A QoS provisioning framework has been proposed in [19], [20], [21], and [54]. The framework classifies calls as real-time and non-real-time. Based on a location prediction scheme, the real-time calls are further classified as local or departing calls.

During the call setup time, the call specifies its average bandwidth requirements, minimum bandwidth requirements, and whether it is a real-time or non-real-time call. If the admitting call is classified as a departing real-time call, the scheme uses the received signal strength approach to predict the three closest neighboring cells to the call. The call is admitted only if the call average bandwidth could be reserved in both the current and the predicted destination cells.

Local real-time calls are admitted based on the bandwidth availability in the current cell only. If the amount of available bandwidth is not enough to admit the real-time call, one or more ongoing non-real-time calls might be forced to be buffered so that the released bandwidth could be used to admit the real-time call.

If bandwidth is available, a fixed amount of bandwidth is allocated to each non-real-time call on a time-sharing basis. If the system is fully loaded, the framework may admit non-real-time calls by stealing some channel capacity from inactive real-time calls.

Once the admission control admits a call, the call packets are then marked as real-time or non-real-time and passed to the packet sorter. The packet sorter looks in to the packet and

puts it in either a real-time packet queue or non-real-time packet queue. The packet scheduler gives real-time packets a higher priority by allocating more time slices to real-time queues.

3.4.5 Rate-Based Borrowing Scheme

A rate-based borrowing scheme [26] is another form of service degradation schemes in which the QoS of the existing calls are degraded in order to admit a new or a handoff call. Two techniques are used to control the call dropping and blocking probability. A borrowing technique is used to minimize the call blocking probability by allowing the incoming call to borrow a fraction of the existing calls' allocated bandwidth in order to reach the minimum operating threshold. The same as the adaptive bandwidth and the QoS provisioning scheme mentioned above, the scheme requires that each call specify its type (i.e. real-time or non-real-time) and the call maximum and minimum bandwidth during the setup time.

During the call setup time, if there are not enough resources, the scheme borrows a fraction from each existing call, if possible, in order to provide the admitting call with the minimum operating bitrate. The fraction of bitrate borrowed from each call depends on the difference between the call maximum and minimum bitrates. The bitrate is only borrowed from VBR calls since the maximum and minimum bitrates are equal in the case of CBR calls. As soon as bandwidth becomes available, the borrowed bandwidth is returned to the degraded calls.

In the adaptive bandwidth scheme [74], this scheme reserves a pool of bandwidth at each cell to be used by handoff calls only. However instead of changing the pool size adaptively based on network conditions, a fixed pool size is reserved in this scheme. The fixed reserved pool of bandwidth is only used if there is not enough bandwidth for a real-time handoff call to meet the minimum bandwidth needed. Non-real-time calls have no access to the reserved pool. The authors claim that this scheme performs as well as the adaptive pool scheme without any need for inter-cell communication.

3.5 Conclusions

Charging and resource control schemes are not tied together for third generation wireless networks. Most the resource control schemes proposed for third generation wireless networks have focused only on minimizing the call dropping and blocking probabilities and ignored the economic side of the resource control such as the service provider revenue.

Integrate adaptive charging and resource control to provide different soft QoS levels and control both new and handoff calls admission control are areas that are not yet explored. Furthermore, the currently proposed resource allocation schemes do not take advantage of the customer QoS perception to provide soft QoS multimedia services.

Chapter 4

PROBLEM STATEMENT

Selecting the proper charging scheme plays a crucial role in the success of telecommunication network deployment. Charging policies are used by the service providers to recover the network costs in a competitive and fair way from the diverse population of customers. By adjusting the charging scheme, a service provider can attract new customers, compete with other service providers, and introduce new services and periods of promotions.

However, no QoS-based charging schemes have been proposed yet for the next generation wireless networks. Second generation wireless networks charge customers equally. Customers are charged based on a per minute access time scheme that gives them a limited amount of access time for a predefined fixed price. Service providers may also charge customers an extra price if they exceed their maximum allowable access time. At the same time, the wireline multiservice charging techniques [11], [17], [29], [33], [36], [57], [67], [79], are not suitable for wireless networks because of the disadvantages discussed in Section 3.1.

Therefore, new efficient QoS-based charging schemes specifically designed for the next generation networks are needed. These schemes should charge customers dynamically based on the network conditions and the amount of resources they have asked for and score well against the charging scheme evaluation criteria introduced in Section 3.3.

Much of the work on QoS and multimedia resource control has been focused on wireline networks [6], [12], [14], [18], [35], [63], [84], [86] and does not address the additional complications imposed by the wireless environment. This work cannot be simply applied to wireless networks due to the fundamental differences between wireless and wireline networks such as the higher BER, lower bandwidth, and the non-uniform transmission quality. Moreover, mobile terminals behave very differently from stationary terminals and impose several essential QoS requirements on both the networks and the end-systems.

Although a number of multimedia resource control schemes have been proposed for the next generation wireless networks, those schemes have several deficiencies in the way they allocate scarce network resources, handle the handoff calls, and charge customers. Moreover, these schemes are rigid in nature and do not give the service provider enough flexibility to improve the network performance or achieve different resource control objectives (e.g. maximizing profit or maximizing fairness).

To support heterogeneous QoS applications in a limited resources environment such as wireless networks, the charging and resource control schemes should be as efficient as possible. By efficient we mean that resources should be allocated on demand. At any

point in time, unnecessary resource allocation should be avoided. The framework should also give customers a strong incentive to reveal their true resource requirements.

In order to avoid unnecessary resource allocation, next generation resource allocation schemes should take the customer's QoS perception into consideration when allocating resources to each active call. Allocating the same amount of resource to each customer regardless of the differences in their QoS perception wastes scarce network resources since some customers may be more than satisfied with fewer resources.

Unfortunately, most of the resource control schemes currently proposed for multiservice wireless networks differentiate between customers based on the type and the nature of their running applications only (e.g. real-time vs. non-real-time). Although some of the schemes mentioned in Section 3.4 require that each call specify its preferred and minimum bitrates during the call setup time, this information is assumed to be constant for all customers running the same application type (i.e. calls with the same type have the same preferred and minimum bitrates) and it is not used to measure the customer's satisfaction level.

The graceful degradation schemes mentioned in Section 3.4.2 allow customers to provide a service degradation profile during the call setup time. The degradation profile is used to inform the resource control scheme about the preferred way to discard the application packets during a resource crisis, and has nothing to do with the customer's satisfaction level.

Although allocating resources based on the customers QoS perception can theoretically improve the resource control efficiency, in practice this approach may provide no enhancement or may even worsen efficiency. Providing customers with the ability to specify their QoS requirements could persuade them to overbook the network resources in order to guarantee their maximum QoS satisfaction.

To prevent customers from overbooking the network resources, the charging and resource control schemes should give customers a strong incentive to reveal their true QoS requirements. For example, the graceful degradation scheme gives customers the opportunity to specify their discard profile; however, it does not give them the right motivation to convey the actual discard profile. Consequently, customers may easily convey a false discard profile that minimizes the number of discarded packets in order to maintain the maximum possible satisfaction level.

In addition to resource limitations, active calls in the next generation wireless networks are expected to experience several handoffs during their lifetime, which in turn increases the probability of a premature call termination (call dropping probability). The handoff techniques currently proposed for next generation wireless networks minimize the call dropping probability by utilizing different resource reservation techniques. These techniques, as discussed in section 3.4, include fixed and adaptive bandwidth reservation.

It is a known fact that reserving a portion of the cell bandwidth exclusively for handoff calls will lower network utilization and increase call blocking probability [25], [76], [82],

[71]. Developing an efficient handoff handling technique capable of maintaining an acceptably low call dropping probability without wasting the scarce network resources is another challenging task facing the next generation resource control framework.

Finally, most of the resource control schemes proposed for the next generation wireless networks are designed and optimized toward a single objective only, which is usually minimizing the call dropping probability and the call blocking probability if possible. These schemes are not capable of considering other objectives such as maximizing the service provider revenue or optimizing the resource allocation scheme to maximize the customer's satisfaction level. The next generation charging and resource control schemes should be flexible enough to allow service providers to select and tune the performance of the framework to achieve different resource control objectives.

In summary, the goal of this thesis is to develop a novel QoS aware integrated charging and resource control framework that has the following characteristics:

- *Integrates the charging and resource control schemes.*
- *Provides a charging scheme that:*
 - *Charges customers dynamically based on the network conditions and the amount of resources customers have asked for.*
 - *Provides strong incentives to customers to reveal their true QoS requirements.*
 - *Scores well against the evaluation criteria in Section 3.3.*

- *Is flexible in allowing the network operators to pursue different resource control policies (e.g. maximizing network revenue, or maximizing fairness).*
- *Enhances resource usage adaptation based on resources availability and customer's QoS preferences.*
- *Provides soft QoS services in terms of customer's QoS perception.*

Chapter 5

INTEGRATED CHARGING AND RESOURCE CONTROL FRAMEWORK

To address the specific areas discussed in the previous chapter, we propose a novel QoS aware ICRC framework. The framework improves the efficiency of the resource allocation process by considering the customer's QoS preferences while allocating resources to active calls. Instead of allocating the same amount of resource to each customer regardless of the differences in the customer's QoS perceptions, the framework allocates resources to active calls based on the QoS parameters specified during the call setup time.

In order to obtain access to the customer's QoS preference, the framework introduces the QoS profile concept (described in detail in Section 5.2). Using the QoS profile customers can specify their QoS preferences in terms of QoS parameters offline. The QoS profile is then submitted to the framework during the call setup time. The QoS profile allows customers to specify QoS parameters corresponding to four different satisfaction levels, namely the excellent, good, fair, and poor satisfaction levels.

As mentioned previously in Chapter 4, allowing customers to specify their QoS requirements may persuade them to act selfishly and overbook the scarce network resources by provisioning their QoS profiles with higher QoS parameters than they really need. The framework motivates customers to reveal their true QoS requirements by charging them differently, based on the QoS level they request (provisioned QoS parameters).

Charging customers differently has the potential of increasing the network and economic efficiency since some existing customers would pay a higher price for better service, while other customers would ask for a lower QoS level in favor of a lower price. By allowing customers to select their QoS requirements and charging them in accordance, service providers can provide more fair and competitive services. The level of customer happiness will also improve since customers will be getting the service that best fits their needs.

In wireless networks, there are mainly two types of resources: signal power and network bandwidth. The signal power and the customer's signal quality and location determine the maximum data rate the customer may receive. In 3G wireless systems such as 1XEV-DO [9] and HSDPA [31], the forward link is a "fat pipe" with transmission at full power to every user. In this case, power control does not apply, and the user data rate is determined by matching the modulation and code rate to the perceived signal quality, which may vary with distance, speed, etc. The data rate a user can receive at some given instance may be

greater than, equal to, or less than the minimum rate needed by the application. The system exploits the user's favorable signal quality conditions and transmits only to the users who can achieve the desired rate.

In this research, it is implicitly assumed that users in the system can be given their desired rate on the basis of scheduling the users at instances when they can achieve the desired rate. Power control is not considered since a "fat-pipe" transmission is assumed. A similar simplifying assumption is made for the reverse link as well. Future research may consider power controlled systems as discussed in Section 9.2.

To charge the customers equally based on the amount of resources they consume the framework assigns a price per Kbps of the network bandwidth that changes dynamically based on the network condition (bandwidth market price). The bandwidth market price is defined as follows:

Definition 1:

The bandwidth market price is the current price to transmit one Gbit of traffic using one kbps of the network bandwidth.

Customers are charged based on a new dynamic charging scheme that depends on the bandwidth market price and the amount of bandwidth a customer needs to reach a specific QoS satisfaction level. The new charging model is described in detail in Section 5.1.

Since the admission control decision and the amount of bandwidth allocated to each customer depend on the bandwidth market price, customers may need to know the current bandwidth market price before they initiate their calls. Given that all the wireless networks broadcast network specific information using a public access channel (e.g. paging channel), the framework may add the current bandwidth market price to the rest of the broadcast information.

During the call setup time, the call's QoS parameters are mapped to the equivalent amount of network resources. QoS parameter mapping techniques are outside the scope of this research. Readers may refer to [42], [83], and [85] for more information about QoS parameter mapping in data networks. For simplicity, the framework simulation program only maps the call's bitrate QoS parameter to the equivalent amount of bandwidth. Resources required to satisfy the call's other QoS parameters such as signal power, scheduler priority and extra amount of bandwidth for ARQ and FEC are not considered in this research.

Instead of reserving a fraction of the network resources exclusively for handoff calls, the framework minimizes the call dropping probability by degrading the QoS level provided to existing active customers to free enough resources to accommodate the handed-off call. The framework compensates the degraded calls by reducing the charging rate to match the actual amount of resources allocated to each call. Adjusting the amount of resources

allocated to active customers depends on the ICRC policy selected by the service provider. Examples of ICRC policies are described in detail in Chapter 6.

Unlike other charging and resource control frameworks, the proposed framework gives the service provider a high degree of network management flexibility. By supporting several ICRC policies, the framework allows service providers to select or create the policy that best fits their marketing strategy.

In addition to a range of policies, the framework provides network operators with a set of tuning parameters to tune the policy behavior. Tuning parameters such as minimum admission threshold, minimum bandwidth market price, minimum revenue increase, and blocking and dropping percentages can be used. For example, if a service provider sets the framework policy to maximize the network revenue, he could use the call dropping and/or the blocking percentage to insure that maximizing the network revenue does not increase the percentage of dropped and/or blocked calls beyond a certain threshold.

The framework is designed using a distributed approach to reduce the process complexity and the inter-cell communication. The distributed approach also helps the service providers to deploy and upgrade the framework gradually without affecting the rest of the network. In a distributed system, the framework processes should be able to make their decisions based solely on the information available at the cell site. Since the computational power required to run the framework might be available only at a central

site (e.g. an RNC node in a UMTS network Figure 5-1), a number of framework instances should be instantiated to maintain the one framework per cell site relationship.

In the following section the framework charging model is described in detail followed by the QoS profile concept Section 5.2. The framework market model formulation is introduced in Section 5.3. The framework processes and tuning parameters are the focus of Section 5.4. The framework processes are described in Section 5.5 followed by the framework assessment in Section 5.6.

Figure 5-1: The Framework Deployment in a UMTS Network

5.1 Framework Charging Model

In a multiservice wireless network, different customers will have different QoS requirements. As a result, the amount of network resources allocated to each customer will be different too. In order to charge the customers fairly based on the amount of network resource they consume, we introduce a QoS based Charging Model (QoS-CM) that takes into consideration the QoS parameters specified by the customers in their QoS profiles. The QoS-CM is a generic charging model that has a tariff component for each QoS parameter supported by the resource control framework.

$$T = T_B + T_E + T_F \quad \text{¢/Gbit of transmitted traffic} \quad \text{Equation 2}$$

Where T_B denotes the bitrate tariff, T_E denotes the ETED tariff, and T_F denotes the FER tariff.

This tariff formula allows the framework to support two grades of services. A basic service would support only the customer bitrate. More advanced services would support the bitrate plus one or both of the ETED and the FER.

The bitrate tariff component is directly proportional to the bitrate assigned to the customer (br Kbps) and the bandwidth market price (μ ¢/Gbit/Kbps). The bitrate tariff increases linearly with the increase in the bitrate assigned to the customer or the increase in the bandwidth market price.

$$T_B = \mu \cdot br$$

Equation 3

In addition to the bitrate, real-time applications with tight delivery requirements may ask for a guaranteed ETED. For the forward link traffic, we are only interested in the delay portion from the time a data packet arrives at the base station until it is delivered to the mobile terminal. This delay consists of three parts: processing, scheduling, and propagation delays. The propagation delay depends on the distance between the mobile terminal and the base station. Considering a macro cell environment, the delay introduced by this component can be neglected since radio waves cover a distance of around 300 km in one millisecond.

The processing delay is the deterministic time taken by both the base station and the mobile terminal to encode/decode and interleave the data packets. The processing delay depends on the codecs used and on the interleaving schemes. The scheduling delay is mainly the time a packet spends at the MAC layer waiting for a time slot(s) to be transmitted. The transmission time slots are granted by the MAC layer scheduler (Figure 5-2).

A multimedia aware MAC layer differentiates between different traffic classes based on their delay requirements. For real-time applications, the MAC layer scheduler may reserve a periodic time slot on the access channel. Once the time slot has been reserved, no other calls may contend for it. The scheduler may also grant transmission time slots on a priority packet-by-packet basis. The priority scheme often relies on the time-to-live

parameter assigned to each packet. The shorter the packet's time-to-live parameter the higher the priority it receives from the scheduler.

Figure 5-2: MAC Layer Scheduler

The framework charges customers with bounded delay time an extra premium (T_E) to offset the social cost they impose on the other active calls by reserving a transmission time slot or asking for a higher scheduling priority. An exact calculation of T_E requires pricing the MAC layer time slots and calculating the number of time slots consumed by each call. For simplicity, we assume T_E is calculated as a percentage of the basic service tariff.

$$T_E = \alpha \cdot T_B = \alpha \cdot \mu \cdot br$$

Equation 4

The fraction α is a service provider configurable parameter. The service providers may adjust α based on their marketing strategy. A service provider may waive the guaranteed ETED charge completely by setting α to zero. When it is used, the value assigned to α should be adjusted such that it is inversely proportional to the call ETED as shown in Table 5-1. The numbers shown in Table 5-1 are arbitrary numbers given as an example to highlight the effect of specifying a maximum ETED QoS parameter on the call total cost.

End to End Delay (msec)	α
ETED < 150	0.40
ETED < 200	0.30
ETED < 250	0.25
ETED < 300	0.20
ETED < 350	0.15

Table 5-1: Example End-to-End Delay Tariff Parameter

Like ETED, applications with bounded FER requirements may ask for a guaranteed maximum FER. In the wireless environment, a forward error correction (FEC), automatic repeat request (ARQ), or a hybrid FEC/ARQ scheme could be used to control the call FER. In order to detect or correct a corrupted packet, more redundant bits are added to

each transmitted packet. If required, more network resources could be allocated to retransmit any corrupted packets.

The customer signal quality and hence the amount of resources consumed to guarantee a FER percentage depends to a great extent on the customer location and the surrounding environment (e.g. fading and interference). The customer signal quality is affected by the amount of interference received from neighboring customers. As the interference level increases the customer signal quality decreases and the amount of allocated resources needed to guarantee a FER level increases.

Therefore, we think that charging customers based on the amount of resources consumed to guarantee a maximum FER is unfair. It is unfair to charge customers more because they are far from the base station or to overcome the interference they get from other customers. As a solution the service provider should charge calls with guaranteed FER requirement an extra percentage of the basic service tariff that depends only on the required FER.

$$T_F = \gamma \cdot T_B = \gamma \cdot \mu \cdot br$$

Equation 5

The fraction γ is also a service provider configurable parameter. The service provider may adjust γ to work toward their marketing strategy. The value assigned to γ should be adjusted such that it is inversely proportional to the call FER as shown in Table 5-2.

Frame Error Rate	γ
FER = 0%	0.4
FER < 1%	0.3
FER < 2%	0.2
FER < 3%	0.15

Table 5-2: Example Frame Error Rate Tariff Parameter

The numbers shown in Table 5-2 are arbitrary numbers given as an example to highlight the effect of specifying a maximum FER QoS parameter on the call total cost.

The generic call tariff can now be defined as:

$$T = \mu \cdot br + \alpha \cdot \mu \cdot br + \gamma \cdot \mu \cdot br$$

$$T = br \cdot \mu \cdot (1 + \alpha + \gamma)$$

Equation 6

Where $\mu \cdot (1 + \alpha + \gamma)$ denotes the QoS market price or the tariff to transmit 1 Gbit of traffic using one Kbps on the network bandwidth.

5.2 Customer's QoS Profile

As mentioned earlier, next generation wireless networks will be dominated by adaptive multimedia applications. Adaptive multimedia applications deliver various QoS satisfaction levels based on the amount of allocated resources. Most of the existing QoS based resource allocation schemes assume that customers running the same application

have the same QoS perception. This assumption could hold for non-adaptive applications, however it does not hold in case of adaptive ones. For the same adaptive application, different customers could have different QoS perceptions. For example, some customers could perceive a 10 Kbps voice channel as good service while others perceive it as only fair service.

At the same time, the value of information is variable across and within the applications. Sometimes an audio file download might be more valuable than an email. Sometimes business emails will not be as valuable as a personal message. Even the same email can have a different value depending on the subscriber's environment.

In order to gather the customer's QoS perception, we introduce the customer QoS profile concept. The QoS profile provides the customers with a way to initialize the call QoS parameters based on their QoS perception. Customers can use the QoS profile to initialize the call's QoS parameters such as bitrate, ETED, and FER. By adjusting the QoS parameters, customers can build QoS classes that comply with the classes defined in Section 2.2.2 or create their own new ones. Note that, a "customer" may imply an individual or an organization when the QoS profile could be built for a group of people within the organization.

The value assigned to each QoS parameter depends on the customer's running application, QoS perception and the QoS classes supported by both the base station and the customer's mobile terminal. For example, a high QoS customer with an AMR-WB

codec supported terminal should set the call bitrate to one of the bitrates supported by the AMR-WB codec (e.g. 8.85, 14.25, 18.25, or 21.05).

The lack of ability to predict call cost before it is started is one of the major reasons why customers prefer fixed charging schemes over dynamic ones. To improve the call cost prediction of the framework charging schemes, customers are asked to assign a “*budget*” to the QoS profile.

Definition 2:

The customer’s budget is the maximum price per Gbit of transmitted traffic the customer is willing to pay in order to achieve a specified QoS satisfaction level.

By setting the QoS parameters and assigning a budget to the QoS profile, the customer enters a compulsory agreement with the service provider to pay a maximum price equal to the assigned budget per Gbit of transmitted traffic. Since different customers ask for different QoS parameters and hence different amounts of network resources, it is unfair to differentiate between customers based on their absolute “*budget*”. Instead, the framework differentiates between customers based on their “*bid*”.

Definition 3:

The customer's bid is the maximum "basic service" price the customer is willing to pay to transmit one Gbit of traffic using one Kbps of the network bandwidth.

The customer's bid is used internally by the framework and it is invisible to the customer. Based on the framework call tariff (Equation 6), the framework calculates the customer's basic service budget as follows:

$$\begin{aligned} bud &= bud_b + \gamma \cdot bud_b + \alpha \cdot bud_b \\ bud_b &= \frac{bud}{(1 + \alpha + \gamma)} \end{aligned} \quad \begin{array}{l} \text{¢/Gbit} \\ \text{Equation 7} \end{array}$$

where bud_b denotes the basic service budget.

The customer's bid is then calculated as follow:

$$bid = \frac{bud_b}{br} = \frac{bud}{br \cdot (1 + \alpha + \gamma)} \quad \frac{\text{¢}}{\text{Gbit} \times \text{Kbps}} \quad \text{Equation 8}$$

For example, if a customer is willing to pay 900 ¢/Gbit of transmitted traffic and asks for 16 Kbps bitrate, 2% FER and 200 msec ETED, the customer will have a bid =

$$\frac{900}{16 \times (1 + 0.3 + 0.2)} = 37.5 \frac{\text{¢}}{\text{Gbit} \times \text{Kbps}}.$$

Another customer with 800 ¢/Gbit budget and 12 Kbps bitrate, 0 FER and 300 msec ETED requirements will have a bid =

$$\frac{800}{12 \times (1 + 0.2 + 0.4)} = 41.67 \frac{\text{¢}}{\text{Gbit} \times \text{Kbps}}.$$

The customer bid is compared against the bandwidth market price to decide on the amount of bandwidth allocated to each customer. If the customer bid is higher than the bandwidth market price, the framework grants the customer the amount of bandwidth asked for. Otherwise the call is blocked or dropped.

In order to gain access to the network resources, the customer should adjust his bid to match or exceed the current bandwidth market price. The customer can increase his bid by increasing his budget or by lowering his QoS requirements. Assuming the customer has a limited budget, the customer should then lower his QoS requirements such that the required amount of bandwidth multiplied by the bandwidth market price is still within the limited budget.

To be more competitive in acquiring the network resources, customers are allowed to create multiple QoS entries in their profile. Each entry has the QoS parameter settings and the budget corresponding to one of the customer's satisfaction levels. In this research we limit the number of QoS entries to four corresponding to the excellent, good, fair and poor satisfaction levels. An example is shown in Table 5-3.

5.2.1 How to Use the QoS Profile

Each adaptive application comes with a default QoS profile. The default QoS profile has a recommended setting for the QoS parameters and budget for each satisfaction level. Customers, however, can modify the default setting to reflect their own QoS perception or

to create a more competitive QoS profile. A customer could also use multiple QoS profiles to run the same application with different QoS settings on different occasions.

Customer QoS Profile					
QoS Satisfaction	QoS Parameters			Budget	Bid
	Bitrate	Delay	FER		
Excellent	12.2 Kbps	150 msec	1%	1000	48.22
Good	10.2 Kbps	150 msec	1%	900	51.90
Fair	10.2 Kbps	250 msec	2%	800	54.09
Poor	7.4 Kbps	300msec	3%	600	60.06

Table 5-3: A QoS Profile Example

To simplify the setting of the QoS profile parameters, the next generation wireless terminals may provide customers with a simple graphical user interface that allows them to manually adjust their QoS parameters. The wireless terminal may also playback a sample clip based on the selected QoS parameters. Figure 5-3 shows an example of a customer's graphical user interface. Different customers may provide different interfaces.

Although customers have the freedom to select the budget associated with each satisfaction level, they should follow two rules to create a competitive QoS profile:

1. The budget assigned to a QoS satisfaction level should be lower than or equal to the budget assigned to the next higher satisfaction level. This rule insures that the customer will pay less for a lower QoS satisfaction level.
2. The budget should be such that the bid associated with a satisfaction level is higher than the bid associated with the next higher satisfaction level.

The figure shows a graphical user interface for configuring a Video QoS Profile. It includes five vertical sliders for adjusting parameters: Frame size, Frame rate, Resolution, Number of colors, and Budget. Each slider is controlled by up and down arrow buttons. Below the sliders are buttons for 'Play', 'Stop', 'OK', 'Apply', and 'Cancel'. To the right, a 'Satisfaction Level' section contains four radio button options: 'Excellent', 'Good', 'Fair', and 'Poor'.

Figure 5-3: Video QoS Profile Graphical User Interface

For the example shown in Table 5-3, assume that the customer has a maximum budget of 1000 ¢/Gbit of transmitted traffic. In this case, the customer should assign his maximum budget (1000 ¢/Gbit) to the excellent satisfaction entry. The customer should then select the good satisfaction level budget such that it is less than or equal to 1000 ¢/Gbit to satisfy

the first rule. At the same time, the good satisfaction level budget should lead to a bid higher than the excellent QoS bid in order to satisfy the second rule.

Since the excellent QoS bid = $\frac{1000}{12.2 \times (1 + 0.4 + 0.3)} = 48.22 \text{ ¢/Gbit/Kbps}$, the customer's

good budget should be greater than $48.22 \times 10.2 (1 + 0.4 + 0.2) = 786.95 \text{ ¢/Gbit}$. The customer could then select any value in the $[786.95, 1000]$ range for the good satisfaction level budget. The higher the selected budget the more competitive the customer profile is.

Since customers may not follow the above rules while setting their QoS profiles, the QoS profile can be validated either at the user side or at the network side. In case of the user side, after the QoS profile is modified the customer's handheld terminal may check the validity of the QoS profile. The handheld device may display a warning message if a minor error is found in the modified QoS profile. The handheld device may also mark the modified profile as invalid if a major error is discovered. If the handheld device does not have the capabilities to validate the modified QoS profiles, the framework may still validate the received QoS profile during call setup time and take the proper action based on the severity of the discovered errors.

5.2.2 QoS Profile Benefits

The QoS profile concept adds several benefits to the proposed framework. For instance, increasing the associated bid as the satisfaction level decreases allows customers to

compete for a lower QoS level if they cannot afford the amount required to maintain the current QoS satisfaction level. For the QoS profile in Table 5-3, if the bandwidth market price increases from 45 to 55 ¢/Gbit/Kbps, the customer will still have a chance to secure the good satisfaction level until the bandwidth market price increases beyond 61.27 ¢/Gbit/Kbps of transmitted traffic.

In section 3.1.1.2 we mentioned that the PMP scheme does not guarantee the QoS offered even to customers using the highest price subnet. At the same time, PMP does not give customers the choice of accessing different QoS subnets if they are not satisfied with the QoS provided by the current subnet. Unlike PMP, using the QoS profile approach gives customers simultaneous access to four different satisfaction levels. Based on the bandwidth market price and the budget assigned with each satisfaction level, the customer will be assigned the proper amount of resources and satisfaction level.

To reduce the control and negotiation messaging overhead, the framework allows the customer's QoS profile to be transmitted only once during call setup. Based on the network conditions and the call's QoS profile, the framework dynamically adjusts the amount of resources allocated to a call throughout its lifetime without exchanging any further control or negotiation messages. Moreover, by only signaling changes to the default QoS profile once during the call setup time, only a minor signaling effort is needed.

Furthermore, the QoS profile allows the proposed framework to provide both guaranteed and soft QoS services. In general, the framework can guarantee a certain QoS level to the customer if the customer bid for that level is always higher than or equal to the bandwidth market price. The customers may also relax the maximum price constraint by assigning an ∞ budget to a certain QoS level (Table 5-3: poor level). In that case, the customer puts no limit on the maximum price he is willing to pay to get the poor QoS level, i.e. the call will never drop. The customer is committed to pay whatever the market price is for that QoS level.

Customer QoS Profile					
QoS Satisfaction	QoS Parameters			Budget	Bid
	Bitrate	Delay	FER		
Excellent	12.2 Kbps	150 msec	1%	1000	48.22
Good	10.2 Kbps	150 msec	1%	900	51.90
Fair	10.2 Kbps	250 msec	2%	800	54.09
Poor	7.4 Kbps	300msec	3%	∞	∞

Table 5-4: A QoS Profile for a Guaranteed Poor Satisfaction Level

The customers could also adjust their QoS profile to minimize the blocking and dropping probabilities. The customer can guarantee a zero dropping probability by relaxing the bid constraint for any QoS satisfaction level. The same concept could be used to minimize the call blocking probability. However, to minimize the blocking probability the customer

should relax the bid associated with the minimum admission threshold defined by the service provider.

In addition to the service QoS satisfaction, the QoS profile can also be used to reveal the customers price satisfaction. A customer can do that by assigning the same QoS parameters to all the satisfaction levels. Instead of assigning the maximum price the customer is willing to pay for the excellent satisfaction level, the budget is assigned to the lower poor satisfaction level. A lower bid could then be assigned to each higher QoS satisfaction level (Table 5-5). By doing so, the customer will get the service at the lowest price possible.

A customer could also use the QoS profile to inform the system about the preferred way the offered service should be degraded in case of a resource crisis. For the QoS profile shown in Table 5-3, in case of resource shortage, the customer asks the framework to first degrade the ETED QoS parameter from 150 msec to 300 msec and the FER to 2% instead of 1 %. The framework could then reduce the assigned bandwidth to 10.2 Kbps given that the ETED is improved back to 200 msec and the FER increased to 3%. After that the framework could further reduce the bitrate to 7.4 Kbps and the ETED delay to 350 msec and keep the FER at the 3% level.

5.3 Market Model

The framework utilizes a competitive market model to highlight the interaction between customers and the service provider. The model has two agents: service providers represented by the resource control framework and customers represented by their applications. Applications require resources to satisfy their QoS requirements. The service provider owns the network resources. With the help of the resource control framework, the service provider seeks to maximize their objective by selling their resources in the market.

Customer QoS Profile					
Price Satisfaction	QoS Parameters			Budget	Bid
	Bit Rate	Delay	FER		
Excellent	12.2 Kbps	150 msec	3%	600	31.73
Good	12.2 Kbps	150 msec	3%	750	39.66
Fair	12.2 Kbps	150 msec	3%	900	47.59
Poor	12.2 Kbps	150 msec	3%	1000	52.88

Table 5-5: Price Satisfaction Profile

Customers are selfishly competing for the amount of available resources. Each customer tries to maximize his QoS satisfaction level by reserving the maximum amount of resource possible. The market is fair since every customer pays the same price for the

same service. Customers who get more resources, performance, or reliability from the network should expect to pay more.

To control the amount of resources allocated to each call, each resource in the network is assigned a market price. The bandwidth market price changes dynamically as a result of a call admission, handoff, or termination. Adjusting the bandwidth market price dynamically to reflect changes in the network condition plays a crucial role in how the framework achieves optimality. Improper setting of the bandwidth market price could lead to lower network revenue, higher blocking and dropping probability, and/or lower customer satisfaction levels.

We assume that the network has a maximum bandwidth capacity $= bw^{\max}$. The bandwidth has a market price $= \mu$. The network supports a set of applications ψ and has a set of customers Φ . Each customer $i \in \Phi$ has an application type $j \in \Psi$. Each application j has a QoS profile. The QoS parameters and the bid assigned with each QoS level by the customer for an application are defined as shown in Table 5-6.

The total bandwidth used $= bw^u = \sum_i br_i$, where br_i is the amount of bandwidth assigned to customer i . Hence, the available bitrate is denoted by $bw^f = bw^{\max} - bw^u$.

The call dropping percentage (dr) is defined as the percentage of the dropped handoff calls in the last N handed-off calls that had arrived at the base station. The call blocking

percentage (bl) is defined as the percentage of the blocked new calls in the last N new calls that had arrived at the base station.

SATISFACTION LEVEL	BITRATE	ETED	FER	BID
Excellent	br_i^{xl}	$eted_i^{xl}$	fer_i^{xl}	bid_i^{xl}
Good	br_i^{gd}	$eted_i^{gd}$	fer_i^{gd}	bid_i^{gd}
Fair	br_i^{fr}	$eted_i^{fr}$	fer_i^{fr}	bid_i^{fr}
Poor	br_i^{pr}	$eted_i^{pr}$	fer_i^{pr}	bid_i^{pr}

Table 5-6: QoS Profile Definition

To highlight the effect of the budget assigned in the customer's QoS profile on the satisfaction level they obtain, the customer admission threshold (adm_i) is defined as the maximum satisfaction threshold the customer can afford, based on the bandwidth market price during the admission phase.

5.4 Framework Policies and Tuning Parameters

The ICRC framework represents an abstract concept that is used to build different types of ICRC policies as shown in Figure 5-4. Having the flexibility to support several types of ICRC policies is very crucial for service providers to cope with changes in the market, attract new customers, and/or introduce new services. A service provider may create a

new ICRC policy that minimizes the call blocking probability, maximizes the number of customers having a certain level of QoS satisfaction, or that allocates resources to each customer such that all customers have the same QoS perception. Three different policies are described in detail in Chapter 6 to demonstrate the framework flexibility and the ability to support several ICRC policies.

Figure 5-4: ICRC Framework Hierarchy

The ICRC policy consists of four main processes: admission control, resource allocation, price adjustment, and charging. The four processes interact with each other to achieve a specific objective set by the service provider. We define the ICRC policy as follows:

Definition 4:

A framework ICRC policy is a combination of an admission control scheme, a resource allocation scheme, a price adjustment scheme and a charging scheme that work together towards a defined objective set by the service provider.

In addition to the QoS profile and the QoS-CM introduced earlier, the framework enforces a list of rules that each new policy has to comply with. Service providers have the freedom to implement new ICRC policies as long as their policies follow the following framework general rules:

- 1- Each customer conveys his QoS profile during call setup time.
- 2- Resources are allocated to customers based on their QoS profile.
- 3- Resources are allocated such that all customers have a QoS perception within their maximum and minimum thresholds.
- 4- Incoming handoff calls are minimized by degrading the amount of resources allocated to existing customers, no resource reservation is required.
- 5- The bandwidth market price is adjusted dynamically based on network conditions.
- 6- Customers are charged based on the bandwidth market price and the amount of resource they consume.
- 7- The amount of resource allocated to each customer is adjusted such that no one is charged more than his maximum budget.

5.4.1 Performance Tuning Parameters

The framework offers service providers with a list of tuning parameters. After selecting or implementing the ICRC policy that fits their needs, the service providers may tune the policy performance further by adjusting one or more of the following performance tuning parameters:

- 1- Minimum admission threshold.
- 2- Minimum bandwidth market price.
- 3- Minimum improvement percentage of the objective function value.
- 4- The maximum dropping percentage.
- 5- The maximum blocking percentage.

The minimum admission threshold (adm^{min}) is used to control the number and quality of calls admitted to the network. The service provider can set the minimum admission threshold to either poor, fair, good, or excellent. A new call admission request is rejected if the admission threshold of the call is lower than the system minimum admission threshold ($adm_i < adm^{min}$). The minimum admission threshold tuning parameter is usually used with the objective of controlling the QoS satisfaction level provided to the existing calls.

For the same bandwidth market price, raising the minimum admission threshold will mean that fewer calls can afford the new minimum admission threshold. By adjusting the minimum admission threshold the service provider can control the number of calls

admitted to the network and hence the value of the framework objective function and the satisfaction level received by active calls.

The minimum bandwidth market price ($\mu^{\min}$) is the minimum price the service provider can accept per unit of resource. The service provider sets this value to prevent the framework from dragging the bandwidth market price too low in order to accommodate a new call or maximize the customer's satisfaction level.

During the call admission phase, the admission control scheme may compare the current value of the objective function with the expected value if the call is admitted. The new call is admitted only if the expected value is higher than the current one by at least the minimum percentage threshold ($obj^{\min}$). The service provider may use this value to protect the QoS provided to existing calls by admitting only calls that increase the network revenue significantly. Using a percentage value instead of an absolute value makes the parameter more effective as the value of the objective function increases and helps in the load balancing. As the value of the objective function increases, the absolute increase in the objective function value that the admitting call should provide increases; and hence, the number of admitted calls decreases.

The maximum call dropping percentage ($dr^{\max}$) is used to help the framework make the decision to either keep or drop an existing or a handed-off call. An existing call will be dropped if the bandwidth market price is increased beyond the call poor satisfaction

threshold bid. Most of the time, the bandwidth market price is adjusted in order to accommodate a new or handoff call, or to reallocate resources after an active call is terminated or handed-off to another cell. A handed-off call could be dropped if the bandwidth market price at the new target cell is higher than the call poor satisfaction bid.

In any of these cases, if the dropping percentage is lower than the maximum value specified by the service provider, the call might be dropped. Otherwise, the framework will try to readjust the bandwidth market price to avoid violating the maximum dropping percentage threshold. A service provider may set the maximum dropping percentage to zero in order to prevent the framework from dropping calls regardless of the effect on the policy objective.

The maximum call blocking percentage ($bl^{\max}$) is also used to help the admission control process make the decision to either admit or block a new call. A new call could be blocked for one of the following reasons:

1. The new call admission threshold is lower than the system minimum admission threshold.
2. Admitting the call will cause an active call to drop.
3. Admitting the call will degrade the value of the policy objective function.

If any of these conditions occurs and the system blocking percentage is lower than the maximum threshold set by the service provider, the call will be blocked. Otherwise the framework will try to adjust the bandwidth market price in order to admit the new call.

5.5 Framework Processes

As mentioned earlier each ICRC policy consists of four main processes: price adjustment, admission control, resource allocation, and charging. In order to manage the policy selection, handle the customer's QoS profile, and setup the framework tuning parameters, the framework adds another process called "*framework manager*" as shown in Figure 5-5. The five processes are working together to accomplish the objective set by the service provider. In the next sections, the common features of each process are described. Each process is described later in detail according to a particular framework policy.

5.5.1 Framework Manager

The framework manager process is the framework interface to the external world. It provides the service providers with a set of interface functions to add a new policy, select the running policy and set the framework tuning parameters. It also communicates with other wireless system components to receive the call admission, handoff, and termination requests.

Figure 5-5: Framework Processes

Upon receipt of a new/handoff admission request, the process stores the call's important parameters such as the call status, QoS profile, identification number, the current amount of allocated resources, and the average satisfaction level in the call database. The call database is used as a convenient way to share information between the framework processes about the current managed calls. The call database is accessible by all the framework processes and it is updated on a regular basis to reflect the changes in the call conditions when a new or a handoff call is admitted to the cell or an existing call is terminated or handed-off to a neighboring cell.

The framework manager is also responsible for the execution of the ICRC policy. It coordinates the execution of the other processes in order to achieve the framework objective. It also keeps track of the framework important data such as the current

objective function value, call dropping and blocking percentage, and the current bandwidth market price and utilization level.

5.5.2 Price Adjustment

The price adjustment process adjusts the bandwidth market price based on the framework policy and tuning parameters. At any point in time, an infinite number of values could be used as the bandwidth market price. Finding the optimal price is the first step toward framework optimality. We define the optimal bandwidth market price as follows:

Definition 5:

An optimal bandwidth market price (μ^{opt}) maximizes the value of the framework objective function without violating any of the framework control constraints.

From the definition, the optimal bandwidth market price depends on the system policy. Searching for the optimal bandwidth market price that maximizes the network revenue is different from searching for the one that maximizes customer's satisfaction levels. Finding the optimal price could take several iterations. During each iteration, the price adjustment process may increase or decrease the bandwidth market price until the optimal bandwidth market price is found. For example, the stopping condition for a revenue-based

local optimal search could be that the newly generated revenue is lower than the previous value. A flowchart of the market price adjustment algorithm is given in Figure 5-6. Adjusting the increment and decrement value and the search stopping condition are described later in detail based on the framework policy.

In addition to optimizing the framework objective, the price adjustment process insures that the bandwidth market price follows the changes in the network utilization. The main concept is to increase the bandwidth market price when there is excess resource demand, reduce it when there is a shortfall in demand and keep it the same when demand and supply are in balance.

Binding the bandwidth market price to the resource utilization level helps in load balancing and improves the overall utilization of the network resource. During low utilization periods, a low bandwidth market price will encourage active customers to improve their QoS satisfaction level by buying more resources from the network. At the same time, it allows the admission control process to admit more new calls that would otherwise be rejected because of their low budget.

During high utilization periods, increasing the bandwidth market price will free resources by forcing active calls to lower QoS satisfaction levels if they cannot afford the current satisfaction level under the new price. Increasing the bandwidth market price also controls the utilization level by limiting the number of calls admitted to the network.

Figure 5-6: Finding the Optimal BW market price to Maximize a Policy Objective.

5.5.3 Admission Control Process

The main function of the admission control process is to look at the call setup requests received at the base station and decide if a new or handoff call should be accepted or rejected. The admission control process is very crucial for QoS multi-service based networks. Accepting a new call could improve the policy objective function, however it could affect the QoS satisfaction level received by active calls (social cost). On the other hand, rejecting a new call increases the call blocking percentage and reduces the customer's overall QoS satisfaction.

Most of the existing admission control schemes admit new calls only if there is enough resources available to accommodate them. However, using adaptive QoS applications, the amount of resource allocated to active calls could be adjusted to free more resources to accommodate new calls. Consequently, the admission control process behaves differently depending on the amount of available resources.

Since admitting new calls may perhaps reduce the amount of resource allocated to other calls and hence the QoS they receive, new rules should be introduced to measure the trade-off between the objective function gain from admitting the new calls and the social cost imposed on the active calls. For instance, if the reduction in the QoS received by active calls after admitting a new one outweighs the increase in the network revenue, the new call is blocked.

In general, the admission control process makes its decision based on the network condition, the QoS profile of the admitting call, and the objective of the framework policy. The process compares the current value of the policy objective function and the projected value if the new call is accepted. The admission control algorithm may admit the new call only if the value of the framework policy is improved and no other constraints are violated.

Property 1:

Reject the new call request if the new value of the policy objective function is lower than the current value.

$$obj^{new} \geq obj^{cur}$$

Equation 9

In addition to the policy objective function value, the admission control blocks any new call if it does not have enough budget to achieve a predefined QoS satisfaction threshold. The admission control process compares the bandwidth market price with the bid associated with a predefined QoS satisfaction threshold. If the satisfaction level bid is lower than the bandwidth market price, the call is rejected.

Property 2:

Reject a new call request if it cannot afford a minimum predefined QoS satisfaction threshold.

$$adm_i \geq adm^{\min}$$

Equation 10

Given that blocking a new call is much less harmful than forcing an active call to drop, admitting a new call should not force any active call to premature termination.

Property 3:

Reject the new call request if admitting the new call will force any of the active calls to premature termination.

Since an active call is dropped only if the new optimal market price is higher than the call poor threshold bid, the admission control algorithm insures that admitting a new call will not increase the bandwidth market price higher than the lowest poor threshold bid of an active call.

$$bid_i^{pr} > \mu^{opt}$$

$$\forall i \in \Phi$$

Equation 11

At the same time, the admission control process admits a handoff call if degrading the QoS perception level of the existing calls to the minimum perception level could free enough resources for the handoff call to reach its satisfaction level.

Property 4:

A handoff call is only blocked if the amount of available resource is not enough to reach the minimum QoS satisfaction threshold even after degrading the active calls QoS level to their minimum threshold.

$$\sum_i br_i^{pr} > br^{\max} \quad \forall i \in \Phi \quad \text{Equation 12}$$

Since an active call is dropped only if the new optimal market price is higher than the call poor threshold bid, the admission control algorithm insures that admitting a new call will not increase the bandwidth market price higher than the lowest poor threshold bid of an active call.

$$bid_i^{pr} > \mu^{opt} \quad \forall i \in \Phi \quad \text{Equation 13}$$

The call admission process is summarized in Figure 5-7

Figure 5-7: Call Admission Flowchart for Maximizing a Policy Objective

5.5.4 Resource Allocation Process

After calculating the new optimal bandwidth market price, the resource allocation process allocates the available amount of resource among active calls. The amount of resource allocated to each call depends on three factors:

- Resource allocation scheme
- Bandwidth market price
- Call QoS profile.

Different resource allocation schemes could be used. The selection of the appropriate scheme is governed by the framework policy. Most of the resource allocation schemes can insure zero call dropping percentage by allocating the poor QoS threshold to each active call first (Figure 5-8). The scheme could then use the remaining amount of resource to maximize the network utilization or maximize the number of calls having a predefined QoS satisfaction threshold. The resource allocation process could also allocate the available resource to impose certain fairness criteria.

5.5.5 Charging Process

The charging process is responsible for generating the customer's charging records and sending them to the billing subsystem. In order to calculate the customer tariff the process should have access to the customer's QoS profiles, the amount of bandwidth allocated to each customer, and the current resource market price. The charging process then calculates the customer's tariff based on the charging module introduced in Section 5.1.

Figure 5-8: Zero Dropping Resource Allocation Scheme.

5.6 Assessing the ICRC FRAMEWORK

In this section we assess the proposed framework based on the criteria introduced in Section 3.2.

5.6.1 Framework Overhead

In wireless networks, a major design constraint is the amount of spectrum consumed by the control and the acknowledgement messages transmitted between the control system and mobile terminals. Designing a sophisticated resources control system that has a high degree of interaction with the mobile terminals could lead to lower network capacity.

Utilizing a distributed design reduces the system overhead by eliminating any need for interconnection, synchronization or update control messages between neighboring cells or between a cell and a central control unit.

The number of control messages between customer terminals and the proposed framework is reduced substantially by bringing the QoS profile concept into play. QoS profiles allow customers to convey their QoS requirements in advance during the call setup time. The framework then adjusts the amount of resource allocated to each call based on the bandwidth market price and the customer QoS profiles. There is no need to renegotiate the QoS parameters or to send QoS control messages back and forth between the framework and the mobile terminal throughout the call lifetime.

The framework expects to receive the customer QoS profile once during the call setup time. Customers could change their QoS profile at any time, however these changes will take effect only during the next session. Customers may also create multiple QoS profiles per application. For example, a customer could have a low satisfaction QoS profile to run the application during the daytime when the network is expected to be congested, and another high satisfaction QoS profile after peak hours. Further reductions in the number of control messages can be obtained if the framework stores the last QoS profile used by the customer and receives only the updated changes to that profile.

5.6.2 Predictability

Although the bandwidth market price changes dynamically with the network conditions (dynamic pricing), customers and service providers are still able to predict the service price and network revenue. By attaching a bid to each QoS satisfaction level, a customer insures that the maximum price he will pay is the amount he specifies in his QoS profile. Depending on the network conditions, a customer may get better than expected service for a low price but he will never pay more than his budget.

At the same time, service providers can also predict their revenue. Service providers can adjust the minimum bandwidth market price and select the price adjustment and resource allocation scheme that fulfills their marketing strategy.

5.6.3 Flexibility

The proposed framework provides high degree of flexibility to both customers and service providers. In addition to the flexibility in choosing the QoS parameter settings and the bids associated with each QoS satisfaction level, customers have the flexibility to select between guaranteed or soft QoS service. They also have the flexibility to adjust the QoS profile to minimize their blocking and dropping probabilities.

The flexibility of the resource control framework is very important for the service providers. It allows them to react quickly to changes in the market, keep their existing customers satisfied, and attract new customers. The proposed framework allows service providers to select among different resource allocation and price adjustment schemes. New resource allocation and price adjustment schemes can be incorporated in the framework easily. As well, the framework allows the service providers to tune the performance of the framework by adjusting the tuning parameters. For instance, a service provider may provide periods of promotion by adjusting the minimum bandwidth market price.

Having a distributed framework adds another level of flexibility. A service provider can simply use the distributed nature of the framework to introduce new services or offer promotions to certain regions without affecting the charging and the resource control schemes currently used in other regions. It also simplifies the network deployment and upgrading process.

5.6.4 Practicality

The proposed framework is practical and straightforward to implement and to manage. A few changes should be made in call setup protocol and in the supported applications. The initialization phase of each supported application should be modified to gain access to the customer QoS profile. The application should then use the modified call setup protocol to transmit the QoS to the control system.

For simplicity, during deployment the framework could support a limited number of applications and control policies. New application's support and control policies can be incorporated in the framework later if required. Using a distributed design allows the service provider to insure the correctness and effectiveness of the new control schemes on a limited number of cells before full deployment.

5.6.5 Control

The proposed framework provides service providers with two levels of control. The first level is to select the ICRC policy that best fits their marketing strategy. The second level of control is to tune the performance of the selected policy further by adjusting the framework tuning parameters. For example, a service provider may control the number of admitted customers by adjusting the framework minimum admission threshold. Although some ICRC policies may adjust the bandwidth market price based on the bandwidth demand level, the service provider can still prevent the policy from dragging the

bandwidth market price below the framework minimum resource market price. The service provider may also control the new call's dropping and blocking percentages by tuning the corresponding parameters.

5.6.6 Choice

The framework offers real choices to the customers in terms of different modes of operation according to the customer's budgets and their willingness to pay. A customer could select a low QoS to attain of a cheaper deal. In fact, the QoS profile provides customers with a great deal of flexibility to plan their QoS preferences and their spending.

5.6.7 Customer Interaction

As mentioned before, the framework requires minimal interaction between the mobile terminals and the control system during the call setup time. Customers have no access to the bandwidth market price and are not allowed to negotiate the QoS parameters during the call lifetime.

Customers are not expected to change their QoS profiles too often. Each adaptive application is assumed to have a few default QoS profiles to run the application during low and high network traffic periods. Customers could use the default QoS profile, change the default settings, or create more QoS profiles to be used in different occasions.

However, it is highly unlikely that customers will change the QoS profile each time before launching their applications.

5.6.8 Technical Knowledge Requirement

Special technical background is not an essential requirement for customers to receive good and fair services. The QoS profile should be easy to understand and to modify if required. It should not contain any parameters related to the underlying media (e.g. cdma2000, UMTS, wcdma, etc.)

Ideally, customers should express their QoS perception in terms of the application's QoS parameters such as resolution, screen size, and sound quality. However, for simplicity and ease of implementation we assume that each application comes with an agent that maps the application QoS parameters to the corresponding network QoS parameters (amount of resources).

Alternatively, with the help of the default QoS profiles and proper documentation, customers could learn quickly how to specify and change the amount of resource corresponding to their QoS satisfaction levels.

Overhead	Predictability	Flexibility	Practicality	Control	Choice	Interaction	Knowledge
Low	Medium	High	High	High	Much	Low	Medium

Table 5-7: ICRC Framework Evaluation

Chapter 6

CHARGING AND RESOURCE CONTROL POLICIES

Flexibility is one of the major advantages of the proposed ICRC framework. The framework may contain one or multiple ICRC policies. Service providers can easily apply the policy that best suits their marketing strategies. In the following sections, three of these policies are discussed in detail namely the revenue-based policy (6.1), the utility based policy (6.2), and the market based policy (6.3). Throughout each section the policy different charging and resource control schemes are represented along with several optimization aspects such as heuristic and global maximization techniques.

6.1 Revenue Based Policy

A revenue-based policy is an obvious candidate for next generation wireless networks. This policy aims to maximize the service provider revenue regardless of the customers QoS satisfaction. Network resources are allocated to high budget calls despite the social cost imposed on the other active calls.

Although it sounds unfair to the average consumers, this policy is very attractive to business and cooperate customers with high-QoS requirements. During high traffic periods, customers with high budget will act selfishly by guiding the bandwidth market price higher in order to satisfy their QoS requirements. The policy does not offer QoS guaranteed services. However, a customer could increase his chances to receive a certain QoS satisfaction level by maximizing the bid assigned to that satisfaction level.

6.1.1 Problem Formulation

The objective of the framework is to maximize the service provider revenue.

$$\text{Max} \quad Z = \sum_i \mu \cdot br_i \cdot (1 + \alpha + \gamma) \quad \text{Equation 14}$$

subject to the following constraints:

1. The bandwidth market price should be greater than the minimum price specified by the service provider

$$\mu \geq \mu^{\min} \quad \text{Equation 15}$$

2. Each resource has a maximum capacity

$$\sum_i br_i \leq br^{\max} \quad \text{Equation 16}$$

3. Each call has four bids

$$\sum_i \mu \leq bid_i^{xl} \text{ or } bid_i^{gd} \text{ or } bid_i^{pr} \text{ or } bid_i^{fr} \quad \text{Equation 17}$$

4. Each call has a maximum satisfaction level

$$br_i \leq br_i^{xl} \quad \forall i \in \Phi \quad \text{Equation 18}$$

5. Each call has a lower bound on the amount of resource allocated.

$$br_i \geq 0 \quad \forall i \in \Phi \quad \text{Equation 19}$$

6.1.2 Price Adjustment Scheme

The scheme utilized by the price adjustment process plays an essential role in achieving the policy objective (Equation 14). The main target of the scheme is to find the bandwidth market price that maximizes the service provider revenue. Due to the non-linearity of the objective function, finding the global optimal price can be very challenging.

In non-linear environments, it is very likely that a search for the optimal price will lead to a local optimal solution. For example, consider a simple case where a network has only one video call with QoS profile shown in Figure 6-1-a. Assume that the bandwidth market price is 9 ¢/Gbit/Kbps when another video call is admitted. The new video call has a QoS profile as shown in Figure 6-1-b. The price adjustment process has only three choices. It can either increase, decrease, or leave the resource market price unchanged. Figure 6-2 shows the effect of changing the bandwidth market price on the network generated-revenue.

Satisf. Lvl	Bit Rate	Budget	Bid
Excellent	192 kpbs	2000	10.4
Good	128 kpbs	1600	12.5
Fair	96 kpbs	1400	14.6
Poor	64 kpbs	1000	15.6

(a)

Satisf. Lvl	Bit Rate	Budget	Bid
Excellent	256 kpbs	2200	8.6
Good	192 kpbs	1900	9.9
Fair	128 kpbs	1500	11.7
Poor	64 kpbs	900	14.1

(b)

Figure 6-1: Video QoS Profiles

Figure 6-2: Network Revenue Changes with the Bandwidth Market Price

From this simple example, it is clear that increasing the bandwidth market price is not always the right way to increase the network revenue. Although a higher market price generates more per unit revenue, it might discourage customers from buying more resources which leads to both lower network utilization and generated revenue. On the

other hand, a lower market price will lead to a lower per unit revenue and higher overall network utilization.

The aim of the price adjustment scheme is to find the optimal bandwidth market price that maximizes the network revenue by solving the tradeoff between network utilization and bandwidth market price. Since there is infinite number of values that could be used, a linear search for the optimal market price would be very time consuming and impossible to complete in a bounded time.

6.1.2.1 Enumerative Optimal Price Adjustment Scheme

We have developed a heuristic price adjustment scheme to find a good bandwidth market price in a bounded time. Assume that the amount of assigned bandwidth to each customer ($br_i, \forall i \in \Phi$) is constant. Consequently, the network revenue (Equation 14) will be directly proportional to the bandwidth market price. In this case, the optimal market price will be the maximum price the service provider could use without losing his customers.

However, since customers have limited budgets, the network revenue will follow the increase in the bandwidth market price until one of the customer's budget constraints is activated (Equation 17). Any increase in the bandwidth market price beyond this price will reduce the network utilization, and hence the generated revenue.

In other words, the price adjustment scheme may increase the bandwidth market price until it hits one of the customers' budget constraints. With any price increase after this

point, the revenue will fall immediately to a local minimum and start to rise again until the next budget constraint is activated. Therefore, we can identify the points where the bandwidth market price equals one of the customers' budget limits as local maximum. Any infinitesimal increase in the bandwidth market price above the local maximum point will lead to a local minimum as one of the calls will not be able to afford its current satisfaction level any more.

Since the local maximum points are well defined, the heuristic algorithm can then identify the global maximum by selecting the local maximum point with the highest generated revenue. In the worst case, the heuristic algorithm can identify the global maximum in a polynomial time of order $4N$, where N is the number of active customers (Figure 6-3).

The search for the optimal market price is constrained by the framework tuning parameters. It neglects any local maximum with bid that is lower than the service provider minimum market price $\mu^{\min}$. The search also stops if the new market price will force either the call dropping or blocking percentage above the maximum threshold.

6.1.3 Admission Control Scheme

Upon the arrival of a new/handoff call admission request, the admission control task is to decide either to accept or reject the admission request. In a competitive market setting, in addition to the network conditions, the admission control decision depends heavily on the admitting call QoS profile. To increase the chances of admission, a customer should

either maximize the admission threshold budget or reduce the required amount of resource.

Figure 6-3: Optimal Price Adjustment Heuristic Scheme

In the literature, admission control schemes have been used to limit the number of calls admitted to the network in order to guarantee the QoS offered to active calls or to minimize the handoff dropping probability. In contrast with these schemes, the objective of the revenue-based admission control is to maximize the network revenue regardless of the customers' service satisfaction.

The basic concept of the revenue based admission control scheme is to compare the current network revenue with the expected generated revenue if the call admission request is accepted. The admitting call is rejected if the new generated revenue is lower than whatever the network is currently generating. During high network utilization periods, admitting a new call significantly affects the QoS satisfaction level received by existing customers. Therefore, the service provider could control the admission process further by setting a minimum revenue increase percentage the new generated revenue must exceeds in order to accept the new call.

6.1.3.1 Admitting a New Call

In a competitive market setting, making the admission control decision based on the current bandwidth market price is unfair. For example, if there is enough resource to accommodate the admitting call (i.e. $br^f > br_i^{th}$), and the call admission threshold is lower than the policy minimum admission threshold, the admitting call will be rejected. However, since there is enough resource to accommodate the new call, it might be

beneficial to the service provider to reduce the market price to allow the admitting call to achieve the required minimum admission threshold ($adm_i \geq adm^{\min}$).

With the cooperation of the price adjustment process, the admission control process checks the network performance with the admission request assumed accepted or rejected. The price adjustment process provides the admission control process with the projected optimal bandwidth market price if the call is accepted. The admission process then uses the new optimal price to make its decision.

The admission control scheme first compares the network revenue generated using the new bandwidth market price with the network revenue currently generated. If the new generated revenue does not exceed the minimal revenue increase threshold, the new call is blocked. The scheme then compares the call admission threshold with the minimum admission threshold. If the call admission threshold is lower than minimum admission threshold for the call ($adm_i < adm^{\min}$) the call is also blocked.

Moreover, if admitting a new call forces the dropping percentage over the maximum limit, the admission control scheme will block the new call. Finally, before blocking a new call the admission process checks whether the blocking percentage exceeds the maximum allowable limit. If so, the call is admitted unless the blocking percentage conflicts with the dropping percentage. The revenue-based call admission logic is summarized in Figure 6-4.

Figure 6-4: Revenue Based New Call Admission Flowchart

6.1.3.2 Admitting a Handoff Call

Since dropping an ongoing call is more annoying than blocking a new one, the admission control scheme treats handoff calls differently. The scheme does not reserve any resources for handoff calls. Instead, it relies on readjusting the amount of resource allocated to active calls in order to free an amount of resources equal to at least the

handoff call poor satisfaction threshold. Regardless of the network revenue and the minimum admission threshold, a handoff call admission request is accepted if the target cell has enough resource to guarantee at least the poor satisfaction threshold to the existing active calls in addition to the handed-off one (i.e. $br^{\max} \geq \sum_i br_i^{pr}$).

6.1.4 Resource Allocation Process

After finding the new optimal bandwidth market price, the resource allocation process calculates the amount of resources designated for each customer based on the new bandwidth market price and the customer's QoS profiles. This is an optimization problem. We investigate two approaches to determine the amount of resource that should be allocated to each customer. In the first approach, the resource allocation problem is modeled as a mixed integer linear programming (MILP) problem and solved using a third party mathematical programming solver. A heuristic resource allocation scheme is utilized in the second approach.

6.1.4.1 Mathematical Programming Scheme

In order to obtain the optimal resource allocation that maximizes the network revenue, the revenue maximization problem introduced in section 6.1.1 is modeled as a mixed integer linear programming (MILP) problem. The problem is reduced from a mixed integer non-linear programming (MINLP) to a MILP by substituting the bandwidth market price

variable by the optimal market price obtained by the price adjustment process. The MILP is modeled as follow:

Objective function:

$$\text{Max } Z = \sum_i \sum_{j=pr}^{xl} \mu \cdot br_i^j \cdot x_{i,j} \quad \text{Equation 20}$$

Subject to:

$$\sum_{j=pr}^{j=xl} x_{i,j} = 1 \quad \forall i \in \Phi \quad \text{Equation 21}$$

$$\sum_{j=pr}^{j=xl} x_{i,j} \cdot bid_i^j \geq \mu \quad \forall i \in \Phi, j \in (pr, fr, gd, xl) \quad \text{Equation 22}$$

$$\sum_{j=pr}^{j=xl} x_{i,j} \cdot br_i^j \geq br_i^{pr} \quad \forall i \in \Phi, j \in (pr, fr, gd, xl) \quad \text{Equation 23}$$

$$x_{i,j} \text{ is binary} \quad \forall i \in \Phi, j \in (pr, fr, gd, xl) \quad \text{Equation 24}$$

The first constraint insures that only one satisfaction level is selected per call. The second constraint insures that the bid associated with the selected satisfaction level is greater than or equal to the bandwidth market price. The third constraint is used to minimize the call dropping percentage by setting the allocated resource lower bound to the poor satisfaction level. Finally, the fourth constraint states that the variables are binary. This constraint along with the first constraint is essential for the optimization solver to select only one satisfaction level per call.

Solving the MILP model has the advantage of finding the optimal resource allocation that maximizes the network revenue. However, based on the number of binary variables, finding the optimal solution might be very time consuming and therefore impractical in real-time systems.

6.1.4.2 Heuristic Resource Allocation Scheme

The aim of the heuristic resource allocation scheme (Figure 6-5) is to allocate resources to active calls in a fast and efficient manner. For each call, the resource allocation scheme identifies the new maximum affordable satisfaction level by comparing the new bandwidth market price with the call's QoS satisfaction bids. The customer's new resource demand is the amount of resource associated with the new maximum affordable satisfaction level. The scheme then calculates the new total resource demand by adding up the customer's individual demands.

The scheme functionality depends on the amount of available resource at the base station with respect to the total resource demand. If the available amount of resource at the base station is higher than the total resource demand (i.e. $br^{\max} \geq \sum_i br_i$), the scheme simply allocates to each customer the maximum amount of resource he could afford. Otherwise, the scheme has to decide on the optimal amount of resource it should allocate to each customer in order to maximize the network utilization and hence the generated revenue.

In order to reduce the time consumed in identifying the maximum affordable amount of resource and calculating the optimal amount of resource allocated to each customer, we developed a sorted best-fit heuristic resource allocation scheme. The scheme starts by creating an array of satisfaction level entries; four per call corresponding to the poor, fair, good, and excellent satisfaction levels for each customer. Each entry has the following fields: the satisfaction level score, the amount of resource needed to reach that level, and the associated bid. The scheme then initializes the amount of allocated resources and the satisfaction level assigned to each customer as well as the total amount of used resource to zero.

For each satisfaction level entry, the scheme calculates the amount of resource required by the call to reach that satisfaction level by subtracting the amount of resource currently assigned to the call from the amount of resource associated to the entry. The scheme checks if there is enough available resource to satisfy the call entry. If the amount of free resource is not enough the entry is discarded. The scheme then checks if the customer's current satisfaction level is higher than the entry satisfaction level. If yes, the entry is discarded. Otherwise, the scheme compares the entry bid with the bandwidth market price. If the entry bid is lower than the bandwidth market price the entry is also discarded. Otherwise, the entry is added to a candidate list.

Figure 6-5: Heuristic Resource Allocation Scheme

After filling the candidate list, the scheme selects the candidate entry that requires the highest amount of resource. It then updates the amount of resources allocated to the call, the call satisfaction level, the amount of resource available at the base station, and starts a new iteration. The scheme stops if the newly created candidate list has no entries.

In order to minimize the call dropping percentage, the resource allocation scheme can be modified to initialize each active call with enough resource to reach the poor satisfaction level. The scheme could also minimize the blocking percentage by initializing the admitting call at the poor satisfaction level as well. Both the optimal and the heuristic resource allocation are compared the in Section 8.10.

6.2 Utility Based Policy

The utility-based policy is another example of an ICRC policy that can be applied through the proposed framework. This policy allocates network resources with the objective of maximizing the customers' utility (satisfaction levels). The customer's utility is the value corresponding to the customer's satisfaction level as shown in Table 6-1. The numbers in Table 6-1 are arbitrary numbers selected to show the increase in the customer's utility due to the increase in the satisfaction level. The policy supports both utilitarian criterion and equality criterion social welfare resource allocation.

SATISFACTION LEVEL	UTILITY
Excellent	$u^{xl} = 4$
Good	$u^{gd} = 3$
Fair	$u^{fr} = 2$
Poor	$u^{pr} = 1$

Table 6-1: Customer's Utility

6.2.1 Utilitarian Criterion Based Resource Allocation

The utilitarian criterion aims to maximize the total customer utility.

$$\text{Max } Z = \sum_i u_i \quad \forall i \in \Phi, u_i \in (u^{xl}, u^{gd}, u^{fr}, u^{pr}) \quad \text{Equation 25}$$

Where u_i denotes the customer satisfaction level.

The flowchart of the utilitarian criterion resource allocation scheme is depicted in Figure 6-6. Upon the arrival of a call admission request, the resource allocation process calculates the projected Pareto optimal resource allocation assuming that the call admission request is accepted. In order to do that the process utilizes the same heuristic resource allocation scheme introduced in section 6.1.4.2 with the following two modifications. First, instead of comparing the entry bid with the bandwidth market price, the entry bid is compared with the minimum bandwidth market price accepted by the service provider (i.e. $bid_i \geq \mu^{\min}$).

Figure 6-6: Utility Criterion Based Resource Allocation

The second modification is that, after creating the candidate list, instead of selecting the candidate entry that requires the highest amount of resource, the scheme selects the candidate entry that requires the lowest amount of resource to reach the entry satisfaction level. This modification is required to maximize the number of calls having the maximum possible QoS satisfaction level.

6.2.2 Equality Criterion Based Resource Allocation

The equality criterion based policy treats customers equally. It attempts to allocate resource to customers such that all of them have the same QoS satisfaction level. The resource allocation process (Figure 6-7) allocates resources to customers in several iterations. During each iteration, the process sets a target satisfaction level and allocates resource to customers aiming that each call will reach the preset target satisfaction level.

Starting with the poor satisfaction level, the process first calculates the total amount of resource needed such that all the active calls can reach the preset level. The total required amount of resource is calculated by adding up the amount of resource required by each call to reach the preset satisfaction level. A call is only considered if the bid associated with the preset level is not lower than the minimum bandwidth market price (i.e. $bid_i \geq \mu^{\min}$).

Figure 6-7:Equality Based Resource Allocation.

The process then compares the calculated required amount of resource with the available amount of free resource. If there is enough resource, the network updates the amount of resource allocated to each call, the calls' satisfaction levels, the amount of free resource, and advances the target satisfaction level.

If the amount of free resource is not enough, the process then has to select the subset of calls that will reach the preset level. Selecting this subset of calls could be based either on a fair or a utility-based scheme. In the case of a fair scheme, the resource is allocated first to the call with the highest associated bid to insure that the resource is allocated to the call that most appreciates the QoS satisfaction level received. In other words, calls are prioritized based on the bid they have associated with the preset level.

On the other hand, the utility-based resource allocation scheme allocates the resource such that the number of calls enjoying the target satisfaction level is maximized. Instead of allocating resource to the call with the highest bid, the resource is allocated first to the call that requires the lowest amount of resource to reach the target satisfaction level.

6.2.3 Price Adjustment Process

After defining the projected Pareto optimal resource allocation, the price adjustment process task is to recover the projected bandwidth market price. The price adjustment process utilizes a straightforward scheme which loops through all active calls and

identifies the call's projected QoS satisfaction level and associated bid. The scheme then declares the lowest recovered bid as the network projected bandwidth market price.

6.2.4 Admission Control Process

The admission control process uses the projected bandwidth market price and resource allocation to decide whether to accept or reject a new call request. The new call admission request is rejected in any of the following cases:

1. Accepting the new call will force one or more of the active calls to terminate.
2. The projected bandwidth market price is lower than the network minimum market price set by the service provider.
3. The new call admission threshold is lower than the minimum admission threshold.
4. The new value of the network objective function is lower than the current value (i.e. the new customers' utility summation is lower than the current value).

The above rules can be bypassed to control the call blocking percentage. In cases 3 and 4 above, if rejecting the new call admission request will increase the new call blocking percentage above the maximum threshold, the new call is accepted regardless of its admission threshold or the value of the policy objective.

In case of handoff admission request, the admission control process activates the first rule only. Handoff calls are admitted regardless of the new objective function value or the new bandwidth market price. A handoff call is admitted if there is enough resource to guarantee at least poor QoS satisfaction level for all of the active calls in addition to the handed-off call.

6.2.5 The Policy Optimality

The utilitarian and the equality criterion based resource allocation schemes are optimal. During low traffic periods, the resource allocation schemes are optimal since all the active calls will reach the excellent satisfaction levels. During high traffic periods, the resource allocation is Pareto optimal since no call can improve its QoS satisfaction level without lowering the QoS satisfaction level received by another active call.

6.3 Market Based Policy

The last example policy to be explored is the market-based policy. In this policy, the service provider creates a second price auction to determine the Pareto optimal bandwidth market price and amount of resource allocated to each customer. Customers are represented by their QoS profiles and, as usual, define up to four bids corresponding to their poor, fair, good, and excellent QoS satisfaction levels.

The market based policy does not set an objective for the resource allocation policy, instead it leaves the resource allocation and the market price adjustment to be determined by the market demand and supply conditions.

6.3.1 Resource Allocation Process

The resource allocation scheme creates an array of entries. Each call is represented by four entries corresponding to the call's four satisfaction levels. Each entry consists of the following fields: the call and the satisfaction level identification numbers, the satisfaction level associated bid, and the amount of resource the call needs to reach the entry satisfaction level.

The array is then sorted in a descending order based on the bid field. If more than one entry has the same bid, those entries are then sorted further in ascending order based on the satisfaction level field (i.e. resources are allocated to the lower QoS satisfaction level entries first). If more than one entry has the same bid and satisfaction level, those entries are sorted further in ascending order based on the amount of resource field. Consider the two QoS profiles given in Figure 6-1, the resource allocation process sorts the entries as shown in

Table 6-2.

Starting with the first entry, the resource allocation process checks if the amount of free resource is higher than the amount of resource field. If the amount of free resource is

sufficient, the scheme updates the call's projected satisfaction level and the amount of free resource. The process then proceeds by selecting next entry in the array (Figure 6-8).

Call id	Satisfaction level	BID	Amount of resource.
B	POOR	7	6 Kbps
A	POOR	7	8 Kbps
B	FAIR	6.8	2 Kbps
A	FAIR	6.5	2 Kbps
B	GOOD	6.5	2 Kbps
A	GOOD	6.4	4 Kbps
A	EXCELLENT	6.3	2 Kbps
B	EXCELLENT	5.8	4 Kbps

Table 6-2: Market Based Resource Allocation Sequence

The scheme terminates if the current selected entry has a bid lower than the minimum bandwidth market price, all the active calls reach the excellent satisfaction level, or the amount of remaining free resource is not enough to improve the satisfaction level of any of the active calls.

Although the service providers can relay on the customers' competitive and selfish behavior to generate a reasonable bandwidth market price, they could still use the auction reserve price concept to guarantee at least the minimum bandwidth market price. The auction reserve price is the seller minimum acceptable price.

Figure 6-8: Market Based Resource Allocation.

Since the customer's bids are assumed to increase with the decrease of the satisfaction level, the resource allocation process allocates resource to satisfy the calls poor satisfaction level first. However, if a customer specified an excellent satisfaction level bid higher than the value specified by another customer for the poor satisfaction level, the first customer will reach his excellent satisfaction level before the second customer even reaches the poor satisfaction level.

In order to guarantee call continuation (and minimize the call dropping probability), the resource allocation process may allocate enough resource to active calls such that they can achieve the poor satisfaction level threshold before starting the resource auction. A second price auction is then held for the remaining amount of resource. The effect of guaranteeing the handoff call continuation over the bandwidth market price is discussed later in Section 8.3.

6.3.2 Price Adjustment Process

In order to recover the optimal bandwidth market price, the price adjustment scheme makes a small modification to the second price auction rules. Instead of using the first unused bid as the auction price, the scheme searches for the last used entry and uses the bid field of the next lower entry as the projected bandwidth market price. This modification is required since some of the array intermediate entries might not be used because the amount of free resource is insufficient

6.3.3 Admission Control Process

Like the utility-based admission control process, market-based admission control uses the projected resource allocation and market price to decide whether to accept or reject a call admission request. A new call admission request is rejected in the following cases:

1. The admitting call's bids and required amount of resource are high enough that it may force one or more active call to terminate.
2. The call projected satisfaction level threshold is lower than the network minimum admission threshold.
3. The call highest bid is lower than the minimum bandwidth market price.

The admission control process may bypass the second and third rules to guarantee the predefined maximum blocking percentage or the continuation of a handoff call. As usual, a handoff call admission request is accepted if the resource allocation process can guarantee at least the poor satisfaction level to all the active calls along with the handed-off call.

6.3.4 The Policy Optimality

The same as the utility-based policy, the market based policy is also optimal. During light traffic periods customers could reach their excellent satisfaction levels. During heavy traffic periods, resource is allocated to high-bid calls first. Low-bid calls will not be able to improve their satisfaction level without lowering the satisfaction level of the high-bid calls. Hence, the policy is Pareto optimal.

Chapter 7

EXPERIMENTAL SETUP

A software program has been built to simulate the behavior of the proposed framework. The simulation program consists of two modules: traffic generation and resource control (Figure 7-1). Using two separate modules insures that the traffic generation module is not affected by the implementation of the resources control module.

Figure 7-1: Simulation Environment

The network simulation runs in two phases. During the first phase, the traffic generation module generates the calls traffic data (e.g. the call inter-arrival time, type, holding time,

QoS profile, etc.) and records them in traffic files. During the second phase the resource control module retrieves the traffic data from the stored traffic files and creates and configures the traffic calls in accordance. These modules are described in detail in the following sections.

7.1 Traffic and Mobility Generation Module

The traffic and mobility generation module utilizes a maximally equidistributed combined Tausworthe random number generator to simulate the actual traffic generated at a next generation wireless cell. The combined Tausworthe generator has been proposed in [64] and recommended for software random number generation in [65]. The implementation of the random number generation algorithm is provided by the GNU scientific library (GSL) [34].

The GSL library provides a large collection of random number generators that can be accessed through a uniform interface. Additional interface functions are available for transforming uniform random numbers into samples from continuous or discrete probability distributions such as the Gaussian, lognormal or Poisson distributions.

We model the call arrivals as a Poisson process with mean arrival rate λ and exponential inter-arrival time (time until next new call) with mean $= 1/\lambda$. In order to evaluate the performance of the proposed framework under different traffic loads, we have generated several traffic load scenarios by changing the value of λ from 1 call/min up to 5 calls/min.

Assuming medium mobility calls, the probability that the arrived call is a handoff call is 0.5 [74]. Changing the call arrival rate affects the number of calls simultaneously served by the framework. The average number of calls simultaneously served by the framework ranges from 5 to 25 calls.

The framework supports five different types of application:

- Narrowband audio (e.g. normal voice calls)
- Wideband audio (e.g. audio streaming)
- Narrowband video (e.g. video phone)
- Wideband video (e.g. video streaming, news on demand)
- Data (e.g. WWW traffic).

In [26] and [74], traffic types are assumed to have the same probability (i.e. the probability that the generated call is a narrow band voice call equals the probability that the new call is a wideband video call equals 0.2). However, we assume that in the near future the voice calls will have a higher market share than the wideband video calls as shown in Table 7-1.

For voice and video calls the call holding time is modeled as an exponential distribution with mean $\mu = 180$ sec. The WWW traffic is modeled as a session of packet calls based on the traffic model proposed in [29]. Each packet call represents the amount of traffic generated to download one web page. The time between packet calls represents the time

the customer takes to read the downloaded page. The packet call size is modeled as Pareto distribution with cutoff. The Pareto shape parameter is 1.2. The packet call has a mean size of 25 Kbytes, minimum size of 4.5 Kbytes and maximum size of 2 Mbytes. The number of packet calls per session is a Geometric distribution with mean 5. The time between packets calls is also Geometric with mean 120 seconds.

Call Type	Probability
Narrowband voice	0.3
Wideband voice	0.2
Narrowband video	0.2
WWW traffic	0.2
Wideband video	0.1

Table 7-1: Call Type Probabilities

In addition to call arrival and holding time generation, the traffic module associates a random QoS profile to each call. The bitrates associated with each QoS satisfaction level are selected randomly based on the call type and the IEEE recommended QoS parameters for wireless broadband applications (Table 2-4). During the bitrate selection we insure that the bitrate assigned to each satisfaction level is higher than or equal to the bitrate assigned to the next lower satisfaction level ($br_i^{xl} \geq br_i^{gd} \geq br_i^{fr} \geq br_i^{pr}$).

In order to have a realistic QoS satisfaction level selection model, the random selection of the audio QoS satisfaction levels is guided by the QoS MOS shown in [49], while

selection of the video QoS satisfaction levels is guided by the customer's MOS and the video profiles and levels presented in [50], [51], [52].

We assume that the maximum price customers are willing to pay (bud_i^{xl}) follows their income distribution. Customers with high personal income are expected to place higher bids than customer with low personal income. Since the Lognormal distribution has been widely used to model personal income [37] [77], we assume that the bid assigned to the customers' excellent satisfaction level (bud_i^{xl}) follows the same lognormal distribution. For the other satisfaction levels, the budget selection follows the two rules set in Section 5.2.1.

In the simulation we assume that customers are competing only for the network bandwidth. We assume that the average bandwidth market price is 5 ¢/Gbit/Kbps. Hence, we model the customers' excellent QoS bid as a lognormal distribution with a mean equals 5 and standard deviation equals 1.

The proposed ICRC policies are simulated using default QoS tuning parameters. The tuning parameters are then changed individually to study the effect of each parameter on the performance of the proposed policies. The default tuning parameters are shown in Table 7-2.

7.2 Resource Control Module

The resource control module is a software implementation of the proposed resource control framework. The module has a multithreads event driven architecture, and consists of several OO classes (Figure 7-2).

Tuning Parameter	Value
Minimum admission threshold	Poor
Minimum bandwidth market price	0
Minimum improvement percentage of the objective function value	0%
Maximum dropping percentage	0%
Maximum blocking percentage	100%

Table 7-2: Framework Default Tuning Parameters

The module main execution thread starts by spawning three threads: the statistics thread, the resource control thread, and the traffic replay thread. The statistics thread is responsible for gathering and formatting the calls and cells statistics records. The thread wakes up after a configurable period of time and dumps the gathered data to output files.

The module can have only one ICRC policy running at a time. The ICRC policy to apply is selected by passing the proper parameter at the command line. The resource control

object contains the price adjustment, admission control, and resource allocation algorithms required to build a specific ICRC policy.

Figure 7-2: Resource Control Class Diagram

Upon starting, the resource control thread creates two pipes (read and write) in order to communicate with the traffic calls and waits until it receives a message from a new or a

handed-off call. The thread populates the address of the read pipe so any new call knows where to send a setup request message.

As its name indicates, the trafficReplay thread replays the traffic recorded earlier by the traffic and mobility generation module. It reads the traffic files and utilizes the retrieved data to create and initialize data calls. The thread then adds the call to the call database and spawns a new thread to start the call. After creating and initializing a call, the thread sleeps for the call's interarrival time.

After the call thread is started, it sends a setup request synchronous message to the resource control thread. If the admission request is accepted, the resource control thread will allocate enough resource to the call and send a setup request acknowledgment to the call. The call then updates its state to active and sleeps until its holding time has expired or it receives a resource update or termination message from the resource control thread. Each call object has a state that is governed by a state machine (Figure 7-3).

If the call has more than one session (e.g. WWW call), at the end of each session, the resource allocation process de-allocates the call resources and the call updates its state to inactive and sleeps until the start of the following session. At the start of the new session the call sends an activation message to the resource control thread. If the activation request is accepted, the resource control process acknowledges the activation message and the call updates its state to active again. If the activation request is rejected, the call

stays inactive for an exponential random time with mean 30 sec and sends its activation message again.

Figure 7-3: Call State Machine

The interaction between the call thread and the resource control thread is done through the exchange of either synchronous or asynchronous messages. Figure 7-4 shows the message

sequence chart for a failed call setup request, while Figure 7-5 shows the message sequence chart for a successful one.

Figure 7-4: A Failed Call Setup Request Sequence Chart

7.2.1 IBM Optimization Library (OSL)

In order to allocate resource to calls optimally, the resource allocation objects, based on the framework policy, might have to solve a mixed integer nonlinear programming (MINLP) optimization problem. The proposed framework utilizes heuristic algorithms to solve the resource allocation optimization problem and validates the results by solving the same problem using the IBM Optimization Solutions Library (OSL). For instance, the

MINLP resource allocation problem introduced in Section 6.1.1 is reduced to an MILP problem by fixing the resource market price. The MILP is then solved using the OSL solver as shown in Section 6.1.4.1.

Figure 7-5: Successful Call Lifetime Sequence Chart

OSL is IBM's library of high-performance mathematical subroutines for use by applications that solve optimization problems. OSL provides a collection of tools for solving linear programming (LP), quadratic programming (QP), and mixed integer programming (MIP) problems. The MIP solver is capable of handling either a linear or quadratic objective function.

7.3 Performance Metrics

Wireless charging and resource control schemes have been assessed using limited performance metrics such as call blocking and dropping probabilities and resource utilization level. In addition to those metrics, new performance metrics are introduced in this Section.

The ICRC policy performance is assessed using the following metrics:

- Average new call blocking and handoff call dropping percentages
- Admitted calls average satisfaction level
- Comprehensive grade of service
- Average revenue per unit of resource
- Comprehensive network satisfaction.
- Average resource utilization level

The *new call blocking percentage* is measured as the average percentage of calls blocked in the last “X” new calls. The *handoff call dropping percentage* is measured as the

average percentage of calls dropped in the last “Y” handoff calls. For this simulation both X and Y have been set to 1000 calls.

To calculate the *admitted calls average satisfaction level*, the total time the calls spend at the excellent, good, fair, and poor levels is measured. A weight of four is assigned to the period of time the calls have spent at the excellent satisfaction level. Weights of three, two, and one are assigned to the good, fair, and the poor times respectively. The *admitted calls average satisfaction level* is then measured by dividing the total weighted time by the total call time.

$$S_{adm} = \frac{\sum_j 4 \times t_j^{xl} + 3 \times t_j^{gd} + 2 \times t_j^{fr} + t_j^{pr}}{\sum_j t_j^{xl} + t_j^{gd} + t_j^{fr} + t_j^{pr}} \quad \forall j \in \Phi \quad \text{Equation 26}$$

where t_j^{xl} , t_j^{gd} , t_j^{fr} and t_j^{pr} denotes the time an admitted call j spent at the excellent, good, fair, and poor satisfaction levels.

The *comprehensive grade of service* is measured by adding the blocking and dropping call satisfaction to the average satisfaction level achieved by the admitted calls. To do that, an average satisfaction level of “b” is assigned to each blocked call while an average satisfaction level of “d” is assigned to each dropped call regardless of the satisfaction level it had received in the past. Both “b” and “d” are service provider configurable parameters that depend on the average customer reaction to blocked and dropped calls. A proper setting for “b” and “d” could be obtained by conducting a customer satisfaction

level survey. For the simulation a -2 satisfaction level has been assigned to each dropped call while a -1 satisfaction level has been assigned to each blocked call.

The *comprehensive grade of service* “ GoS_{comp} ” is then calculated as:

$$GoS_{comp} = \frac{S_{adm} \times N + b \times M + d \times L}{N + M + L} \quad \text{Equation 27}$$

where N denotes the number of admitted calls, M denotes the number of blocked new calls, and L denotes the number of dropped handoff calls.

The network revenue is accumulated throughout the simulation time. To calculate *the average revenue rate per unit of resource*, the accumulated revenue is divided by the total simulated time and the total amount of network resource available at the cell side.

In wireless networks, finding the optimal cell size is one of the critical network design issues. Adjusting the cell size affects the number of calls the cell will serve (i.e. the cell traffic load). Since the amount of resource per cell is fixed, increasing the traffic load may increase the cell’s generated revenue and the call blocking and dropping percentages, thereby reducing the customers’ satisfaction level.

In order to help the service provider find the proper traffic load and hence cell size, we introduce the *Comprehensive Network Satisfaction* (S_{comp}) as a new performance metric.

The S_{comp} penalizes the network-generated revenue (service provider satisfaction) for any

reduction in the customer's satisfaction level. Since the maximum grade of service the system can achieve is 4, we define the normalized comprehensive grade of service as $GoS'_{comp} = GoS_{comp} / 4$. The *Comprehensive Network Satisfaction* is then calculated as follows:

$$S_{comp} = Rev_{avg} \times GoS'_{comp} = Rev_{avg} \times GoS_{comp} / 4 \quad \text{Equation 28}$$

A separate statistic is also collected to highlight the effect of the customer's QoS profiles on the customer's perceived QoS satisfaction level. Calls are classified into four groups based on their admission threshold. Calls with excellent admission threshold are assigned to the first group. Calls with good, fair, poor admission threshold are assigned to groups two, three, and four respectively. The admitted calls average QoS satisfaction level is then calculated for each group separately.

The performance metrics 95% and the 99% confidence interval are measured based on the following equation [5]:

$$\mu = M \pm z \cdot \frac{\sigma}{\sqrt{N}} \quad \text{Equation 29}$$

Where M denotes the metric mean, σ denotes the metric standard deviation, N is the number of samples and “ z ” is the confidence interval coefficient. The value of the coefficient “ z ” depends on the calculated confidence interval (e.g. $z = 1.96$ and 2.576 for 95% and 99% confidence intervals respectively) [5]. To improve the metrics confidence

intervals, the traffic model is used to generate a total of 100,000 new and handoff call requests in the course of a single simulation run. The metric's mean is sampled periodically after each 1000 call requests (i.e. $N = 100$).

7.4 Compared Policies

The simulation program is also used to simulate the behavior of two other state of the art policies, namely the wireless rate-based borrowing (Section 3.4.5) and the wireline responsive (Section 3.1.2) pricing and resource control policies. The simulation results obtained for these policies are then compared with the simulation results obtained from the proposed ICRC policies.

The rate-based borrowing policy is selected for comparison for two reasons. First, it is one of the latest policies proposed for the third generation wireless networks. Second, it is similar to the proposed framework in various aspects such as making the resource allocation and admission control decisions based exclusively on local base station information, handles several types of calls, differentiates between new and handoff calls, and provides a solution to control both the call blocking and dropping probabilities.

The rate-based borrowing policy is purely a resource control scheme and does not address the customers charging issues. Therefore, we had to enhance it further by adding a simple fixed charging scheme to the simulated policy. A fixed charging scheme is selected

because the original rate-based borrowing policy does not adjust the resource allocation dynamically based on the resource price.

We assume that customers are charged 4 ¢/Gbit/Kbps of transmitted data. We also assumed that 5% of the base station bandwidth is reserved exclusively for handoff calls and the borrowing scheme can borrow up to 50% of the difference between the call's preferred and minimum bitrates [26].

Despite the simplicity and the straightforward implementation of the fixed pricing based resource control policies, we have selected a dynamic pricing based policy to represent the wireline resource control policies because of their several advantages highlighted in Section 3.1.

The responsive pricing scheme dynamically adjusts the resource price based on the network congestion level. If the network is not congested, the congestion price is set to zero and customers are charged based on a fixed price per Gbit of transmitted data. If the network is congested, the responsive pricing scheme increases the price and negotiates the amount of resource allocated to each customer until the congestion is eliminated. The responsive pricing policy utilizes simple resource price based admission control and resource allocation schemes. The admission control scheme admits the new or the handoff call if the call bid is higher than the current resource price. The resource allocation scheme allocates resources to active calls based on the current resource price. Active customers are allocated either the preferred amount of resources they have asked

for or the maximum amount of resources they can afford. The policy does not provide special handling for handoff calls.

Both the rate-based and the responsive policies drop an active call if its allocated amount of resources has been reduced below its minimum specified amount. The call's minimum and preferred amount of resources are assumed to be available during the call setup time.

Although the same traffic data is used to simulate the performance of the three proposed policies along with the rate-based and the responsive policies, most of the call's QoS profile information is not used by the rate-based and the responsive pricing policies during the call setup time. Only the call's excellent and poor bitrates and the excellent budget are used. The call's excellent and poor satisfaction level bitrates are used to represent the call's preferred and minimum amount of resource. The call's excellent budget is also used to represent the arriving call budget.

Chapter 8

RESULTS AND COMPARISONS

In this chapter the performance of the proposed framework is validated. The simulation results are discussed and compared with two currently used charging and resource control policies.

8.1 Call Blocking Percentage

Introducing the QoS profile concept gives the proposed framework a clear advantage over both the base-rate and the responsive policies in terms of managing the call blocking percentage as shown in Figure 8-1. For those two policies, if the bandwidth market price is higher than the admitting call excellent satisfaction bid the call is usually blocked. Instead of blocking the call, the framework may admit it if one of the bids associated with a lower satisfaction level is higher than the bandwidth market price. Thus, the proposed policies achieve a much lower call blocking percentage than both the rate-based and the responsive pricing policies.

The revenue based policy blocks more calls than the utility and the market based policies since it guides the bandwidth market price higher than them both. Since it reduces the bandwidth market price as the traffic load increases, the utility-based policy increases the number of blocked calls as the traffic load increases mainly to maximize the call's satisfaction level.

Figure 8-1: Call Blocking Percentage

Most of the time, a new call is blocked if it does not have enough budget or there is not enough resource at the base station to satisfy its QoS requirements. For revenue-based and the utility-based policies, the admission control schemes also check if admitting the new call will reduce the value of the policy objective function. Therefore, both the utility-based and the revenue-based policies block more calls than the market-based policy.

The market-based policy generates the lowest call blocking percentage due to the fact that customers are encouraged to increase their bids as the satisfaction level decreases and

resources are allocated to the highest bid first. If customers comply with the QoS profile rules, the market-based policy will allocate resources to satisfy the active calls poor satisfaction level first which eventually reduces the call blocking percentage.

On the other hand, both the rate-based and the responsive pricing policies block a much higher percentage of new calls (up to 42% at 5 calls/min traffic load). The responsive pricing policy blocks calls only if the arrived call's bid is lower than the bandwidth market price. The responsive policy bandwidth market price consists of a fixed and a congestion price component. The fixed price component is responsible for 16 % the total blocking percentage while the congestion price component is responsible for the rest.

In the case of the rate-based policy, calls are blocked for two reasons. First, the call does not have enough budget to access the network. Second, there is not enough bandwidth to accommodate the arrived call. As a result, the rate-based policy call blocking percentage has two components, namely price based and resource based blocking percentages as shown in Figure 8-2.

Providing a heterogeneous traffic multimedia traffic using existing resource control schemes, in general, leads to higher call blocking percentage than in circuit switched voice network. This is because existing resource control schemes are not designed to support the high bitrate requirements of the multimedia calls.

The proposed policies reduce the call blocking percentage significantly because they:

1. Free more resource by adjusting the amount of resource allocated to existing calls.
2. Admit a new call if the new free resource is enough for the admitting call to reach the minimum admission threshold

Figure 8-2: The Rate-Based Call Blocking Components

The 95% and 99% confidence intervals for the call blocking percentage at 5 calls/min traffic load are shown in Table 8-1.

POLICY	STDEV	95% CONF. INTER.	99% CONF. INTER.
Responsive	0.78%	41.86 ± 0.18%	41.86 ± 0.24%
Rate-Based	1.28%	41.93 ± 0.3%	41.93 ± 0.4%
Revenue-Based	0.46%	13.92 ± 0.109%	13.92 ± 0.14%
Utility-Based	0.34%	13.29 ± 0.08%	13.29 ± 0.105%
Market-Based	0.45%	6.78 ± 0.11%	6.78 ± 0.14%

Table 8-1: Call Blocking Percentage Confidence Intervals

8.2 Resource Utilization Level

The simulation results show that the network resource utilization level varies based on the applied policy and traffic load (Figure 8-3). The rate-based borrowing policy generates the lowest resource utilization mainly due to the 10% of the bandwidth reserved exclusively for handoff calls. By raising the congestion price dynamically with an increase in the resource utilization level, the responsive policy generates the second lowest resource utilization level. The market-based policy generates the highest resource utilization level.

Figure 8-3: Resource Utilization Level

By maintaining the lowest call blocking percentage, the market-based policy generates the highest resource utilization level. Since allocating more resource generates more revenue, the revenue-based policy utilizes the network resources efficiently. The utility-based policy generates a slightly lower resource utilization level than the revenue-based policy.

The 95% and 99% confidence intervals for the bandwidth utilization level at 5 calls/min traffic load is shown in Table 8-2.

POLICY	STDEV	95% CONF. INTER.	99% CONF. INTER.
Responsive	0.94%	80.89 $\pm$ 0.22%	80.89 $\pm$ 0.29%
Rate-Based	0.82%	68.35 $\pm$ 0.19%	68.35 $\pm$ 0.25%
Revenue-Based	0.82%	87.45 $\pm$ 0.19%	87.45 $\pm$ 0.25%
Utility-Based	0.65%	82.56 $\pm$ 0.15%	82.56 $\pm$ 0.20%
Market-Based	0.69%	89.22 $\pm$ 0.16%	89.22 $\pm$ 0.22%

Table 8-2: Bandwidth Utilization Level Confidence Intervals

8.3 Bandwidth market price

The changes in the bandwidth market price as the traffic load increases from 1 call/min to 5 calls/min are depicted in Figure 8-4. By searching for the highest possible revenue, the revenue-based policy raises the bandwidth market price higher than any other policy. The utility-based policy is the only policy that reduces the bandwidth market price with the increase in the network traffic load. At higher traffic loads, the variety of call QoS profiles increases, and hence the probability that an active low bid call may ask for the lowest amount of resources to reach the next satisfaction level also increases. Consequently, the utility-based policy will allocate the required amount of resource to that call and move the bandwidth market price lower. The service provider may use the minimum bandwidth

market price tuning parameter to prevent the price adjustment process from dragging the bandwidth market price below a certain threshold.

Figure 8-4: Bandwidth market price.

At the same time, with more call variety, the probability of high budget calls arriving also increases. Since the market policy utilizes a second price auction for resource allocation, the bandwidth market price is expected to move higher with the increase in the arriving call bids. However, since a zero call dropping percentage is assumed, the price adjustment process may not follow the expected increase in the call's bids in order to prevent forcing any active call below the poor satisfaction level. Therefore, the bandwidth market price is controlled by the bid associated with the active calls poor QoS satisfaction level rather than the highest bid (Figure 8-5).

As expected the simulation results confirm that the rate-base policy has a fixed bandwidth market price (4 ¢/Gbit/Kbps) regardless of the network traffic load. The simulation results confirm also that the responsive pricing policy increases the bandwidth market price as the traffic load increases due to the increase in the network congestion level.

Figure 8-5: Market Policy Bandwidth market price

The 95% and 99% confidence intervals for the bandwidth market price at 5 calls/min traffic load is shown in Table 8-3.

8.4 Call Dropping Percentage

The simulation results show that the proposed framework can significantly reduce the call dropping percentage as shown in Figure 8-6. It does so by degrading the QoS provided to existing active calls and admitting handoff calls if there are enough resources to satisfy their minimum QoS requirements.

POLICY	STDEV	95% CONF. INTER.	99% CONF. INTER.
Responsive	0.037	4.87 ± 0.0087	4.87 ± 0.011
Rate-Based	0.0	4.0 ± 0.0	4.0 ± 0.0
Revenue-Based	0.008	5.32 ± 0.0019	5.32 ± 0.0025
Utility-Based	0.0056	3.76 ± 0.0013	3.76 ± 0.0017
Market-Based	0.03	4.61 ± 0.007	4.61 ± 0.009

Table 8-3: Bandwidth Market Price Confidence Intervals

By reserving 10% of the cell bandwidth exclusively for handoff calls the rate-based policy reduces the average handoff blocking percentage to a maximum of 5% at 5 calls/min traffic load. On the other hand, given that the responsive pricing policy does not provide any special handling for handoff calls, its average call dropping percentage increases dramatically to 30% as the traffic load increases to 5 calls/min.

Figure 8-6: Call Dropping Percentage

8.5 Customer Satisfaction

As mentioned earlier, the comprehensive grade of service depends heavily on the values the service provider defines for the call blocking and dropping satisfaction. In this simulation we assume a “-1” satisfaction is assigned to each blocked call while a “-2” satisfaction is assigned to each dropped call. The admitted calls average satisfaction level and the comprehensive grade of service simulation results are shown in Figure 8-7 and Figure 8-8 respectively.

In both cases, the utility-based policy generates the highest average satisfaction level possible given the limited amount of network resource. By blocking a high percentage of new calls, both the rate-based and the responsive policies provide admitted calls with a higher average satisfaction level, just slightly lower than the utility-based policy.

Figure 8-7: Admitted Calls Average Satisfaction Level

The 95% and 99% confidence intervals for the customer's average satisfaction level at 5 calls/min traffic load is shown in Table 8-4.

POLICY	STDEV	95% CONF. INTER.	99% CONF. INTER.
Responsive	0.01	3.52 ± 0.0024	3.52 ± 0.0031
Rate-Based	0.015	3.4 ± 0.0035	3.4 ± 0.0046
Revenue-Based	0.031	2.2 ± 0.0072	2.2 ± 0.0095
Utility-Based	0.01	3.56 ± 0.0023	3.56 ± 0.003
Market-Based	0.035	2.72 ± 0.008	2.72 ± 0.0108

Table 8-4: Customer's Average Satisfaction Level Confidence Intervals

When the call blocking and dropping satisfaction levels are considered in the comprehensive grade of service, the responsive policy generates the lowest average satisfaction level. The responsive policy generates a lower average satisfaction level than the rate-based policy due to its high call dropping percentage.

While maximizing the network revenue, the revenue-based policy provides admitted calls with the worst average satisfaction level. When the comprehensive grade of service is considered, the revenue-based policy slightly outperforms the responsive pricing.

As mentioned before, by utilizing a second price based resource allocation scheme, the market-based policy allocates resources to the highest bid first. As a result, resources are usually allocated to satisfy the call poor satisfaction level first, and thus few calls are

blocked, but many of the admitted calls achieve only a poor satisfaction level, so the average call satisfaction level is reduced.

Most of the time, reducing the call blocking percentage increases the number of calls sharing the limited network resources and hence reduces the admitted calls average satisfaction level. Since it has the lowest call blocking percentage, the market-based policy provides admitted calls with the second worst average satisfaction level. At the same time, the market-based policy generates the second highest comprehensive grade of service.

The 95% and 99% confidence intervals for the comprehensive GoS at 5 calls/min traffic load is shown in Table 8-5.

Figure 8-8: Comprehensive Grade of Service

8.6 Average Network Revenue

The simulation results show that the network revenue depends heavily on the network traffic load. One scheme may outperform another scheme during light traffic periods while the same scheme may perform much worse during heavy traffic periods. Another scheme may generate good revenue during light traffic periods and continues its good performance as the traffic load increases (Figure 8-9).

POLICY	STDEV	95% CONF. INTER.	99% CONF. INTER.
Responsive	0.01	1.72 ± 0.0125	1.72 ± 0.0164
Rate-Based	0.019	2.28 ± 0.0043	2.28 ± 0.0058
Revenue-Based	0.035	1.98 ± 0.008	1.98 ± 0.0108
Utility-Based	0.015	3.25 ± 0.0035	3.25 ± 0.0046
Market-Based	0.042	2.59 ± 0.0099	2.59 ± 0.013

Table 8-5: Comprehensive QoS Confidence Intervals

As expected, the revenue-based policy generates more revenue than all other schemes during both light and heavy traffic loads. By blocking a small number of calls and allocating resource to the highest bid first, the market-based policy generates the second highest network revenue.

As mentioned earlier, using a fixed price scheme could lead to the loss of some business opportunities for the service providers since some customers may pay more for the same received service. By utilizing a fixed pricing policy and reserving 10% of the network

bandwidth exclusively for handoff calls, the rate-based policy generates the lowest network revenue.

Figure 8-9: Network Generated Revenue per Unit of Resource

During light traffic periods, the utility-based policy generates higher revenue than the responsive and the rate-based policies. As the traffic load increases, the responsive pricing scheme starts picking up and exceeds the revenue generated by the utility-based policy due to the increase in the congestion-pricing component.

The 95% and 99% confidence intervals for network revenue at 5 calls/min traffic load are shown in Table 8-6.

POLICY	STDEV	95% CONF. INTER.	99% CONF. INTER.
Responsive	0.086	5.07 ± 0.02	5.07 ± 0.026
Rate-Based	0.063	3.5 ± 0.015	3.5 ± 0.019
Revenue-Based	0.066	5.84 ± 0.016	5.84 ± 0.02
Utility-Based	0.024	4.1 ± 0.0056	4.1 ± 0.0074
Market-Based	0.057	5.24 ± 0.013	5.24 ± 0.018

Table 8-6: Network Revenue Confidence Intervals

8.7 Comprehensive Network Satisfaction

As described earlier, the comprehensive network satisfaction performance metric is very important for network design. The service provider can use it to decide on the optimal traffic load (cell size) and the ICRC policy to apply. The simulation results for S_{comp} are depicted in Figure 8-10.

The simulation results show that the rate-based policy has the lowest comprehensive network satisfaction. During light traffic periods, the responsive policy generates more revenue than the rate-base policy. As a result, the responsive policy comprehensive network satisfaction is higher than for the rate-based policy during light traffic periods. As the traffic load increases the responsive policy performs as poorly as the rate-based policy due to the increase in the responsive policy call dropping percentage.

The three proposed policies perform very similarly to each other. Due to their low call blocking and dropping percentages, they provide much better performance than the rate-based and the responsive policies when the comprehensive grade of service is assessed. In terms of traffic load, the proposed policies performances improve as the traffic load increases until it hits the 4 calls/min mark and then it starts to deteriorate.

Figure 8-10: Comprehensive Network Satisfaction

The 95% and 99% confidence intervals for comprehensive network satisfaction at 5 calls/min traffic load are shown in Table 8-7.

8.8 Tuning parameters Effect

To illustrate the effect of the tuning parameters on the framework performance, the simulation results obtained by changing the minimum admission threshold, the maximum

blocking percentage, and the minimum bandwidth market price tuning parameters are discussed below.

POLICY	STDEV	95% CONF. INTER.	99% CONF. INTER.
Responsive	0.031	2.17 ± 0.007	2.17 ± 0.007
Rate-Based	0.031	2.0 ± 0.0073	2.0 ± 0.0095
Revenue-Based	0.018	2.89 ± 0.0043	2.89 ± 0.0057
Utility-Based	0.008	3.33 ± 0.0018	3.33 ± 0.0025
Market-Based	0.019	3.4 ± 0.004	3.4 ± 0.0058

Table 8-7: Comprehensive Network Satisfaction Confidence Intervals

8.8.1 Minimum Admission Threshold

The Minimum Admission Threshold is the minimum satisfaction threshold an admitting call should attain in order to be admitted by the admission control process. Service providers can use this parameter to improve the admitted calls average satisfaction level by preventing low budget calls from gaining access to the network resources. The following simulation results show the effect of increasing the minimum admission threshold from poor to excellent on the performance of the revenue-based policy. The revenue-based policy is selected since it provides admitted calls with the lowest average satisfaction level.

The simulation results show that the call blocking percentage increases as the minimum admission threshold increases due to the fact that many calls will not be able to afford the minimum admission threshold. The revenue based policy call blocking percentage increases from 14% to 57% at 5 calls/min traffic load as the minimum admission threshold increases from poor to excellent. (Figure 8-11).

Figure 8-11: Revenue Based Policy: Call Blocking Percentage

Although increasing the minimum admission threshold will block low budget calls and provide more resources to the high budget calls, the average network revenue still deteriorates as the minimum admission threshold increases (Figure 8-12). The average network revenue deteriorates as a result of the reduction in the bandwidth market price. The bandwidth market price decreases because high satisfaction level thresholds in general have lower bids than the low satisfaction level ones.

Figure 8-12: Revenue Based Policy: Average Revenue

As expected, the simulation results show that the admitted customers average satisfaction levels increase with the increase in the minimum admission threshold (Figure 8-13). The reduction in the resource utilization level as the minimum admission threshold increases leaves enough room for existing calls to improve their QoS satisfaction levels.

Figure 8-13: Revenue Based Policy: Admitted Calls Average Satisfaction Level

8.8.2 Minimum Bandwidth market price

The minimum bandwidth market price tuning parameter is the minimum price the service provider will accept per unit of bandwidth. It prevents the ICRC policy from dragging the bandwidth market price lower than a minimum threshold set by the service provider. The following simulation results show the effect of increasing the minimum bandwidth market price from 3.0 to 5.0 on the performance of the utility-based policy. The utility-based policy has been selected to illustrate the effect of this tuning parameter since the simulation results show that the utility-based policy generates the lowest bandwidth market price and network revenue.

The simulation results show that increasing the minimum bandwidth market price from 3.0 to 5.0 reduces the average resource utilization level achieved by the utility-based policy (Figure 8-14). Increasing the minimum bandwidth market price reduces the amount of resource consumed by individual calls and hence the resource utilization level. The utilization level drops by around 20% from its value when the minimum bandwidth market price increases from 3.0 to 5.0 at 1 call/min traffic load. As the traffic load increases, the unutilized bandwidth is used to admit more new calls. As a result, the utilization level drops by only 6.5% from its value when the minimum bandwidth market price increases from 3.0 to 5.0 at 5 calls/min traffic load.

Figure 8-14: Utility Based Policy: Resource Utilization

The utility based policy blocks new calls for two reasons. First, there is not enough bandwidth to accommodate a new call. Second, a call admission threshold is lower than the network minimum admission threshold. During low traffic periods, since there is always enough bandwidth to accommodate new calls, calls are mainly blocked due to the second reason. By increasing the minimum bandwidth market price, more calls will not be able to afford the minimum admission threshold. As a result, the call blocking percentage increases with the increase in the minimum bandwidth market price.

As the traffic load increases, the network utilization level increases as well, and call blocking due to the lack of network resources becomes more significant (first reason). By increasing the minimum bandwidth market price, some calls will not be able to afford their current satisfaction level and they will be pushed to a lower satisfaction level. Hence, the amount of bandwidth consumed by admitted calls decreases and more

bandwidth becomes available to accommodate new calls which result in a lower call blocking percentage. The call blocking percentage simulation results are depicted in Figure 8-15.

Figure 8-15: Utility Based Policy: Blocking Percentage

Increasing the minimum bandwidth market price improves the network average generated revenue (Figure 8-16). Increasing the minimum bandwidth market price from 3.0 to 5.0 shows no effect on the revenue at 1 call/min traffic load because the increase in the minimum bandwidth market price is balanced by the reduction in the resource utilization level. When the traffic load increases, the increase in the minimum resource price starts to improve the network-generated revenue since the resource utilization level starts to increase by admitting more calls.

Figure 8-16: Utility Based Policy: Average Revenue

As predicted, by preventing the price adjustment process from dragging the bandwidth market price lower than a minimum threshold, low budget calls are not able to achieve a higher QoS satisfaction level. Therefore, the average admitted call satisfaction level deteriorates with the increase in the minimum bandwidth market price (Figure 8-17).

Figure 8-17: Utility Based Policy: Admitted Calls Average Satisfaction Levels

8.8.3 Maximum Blocking Percentage

The maximum blocking percentage is used to limit the call blocking percentage to a certain value as shown in Figure 8-18. The following simulation results show the effect of limiting maximum call blocking percentage to 2% on the performance of the utility-based policy. The utility-based policy has been selected to highlight the effect of limiting the call blocking percentage (i.e. increasing the number of admitted calls) on the admitted calls average satisfaction level.

Figure 8-18: Utility Based Policy: Call Blocking Percentage

Limiting the call blocking percentage to a maximum threshold affects both the generated network revenue and the admitted calls average satisfaction level. As mentioned earlier, reducing the maximum call dropping percentage increases the number of calls sharing the limited network resources, and hence reduces the admitted call average satisfaction level as shown in Figure 8-19.

Figure 8-19: Utility Based Policy: Admitted Calls Average Satisfaction Levels

At the same time, the network revenue is slightly reduced, since the admission control scheme is forced to admit lower budget calls in order to guarantee the maximum call blocking percentage (Figure 8-20).

Figure 8-20: Utility Based Policy: Average Revenue

8.8.4 The Objective Minimum Percentage Improvement

Service providers may use the minimum percentage improvement of the objective function value to protect the average satisfaction level provided to admitted calls by blocking any call that does not increase the objective function value significantly. The following simulation results show the effect of increasing the minimum percentage improvement of the network revenue from 0% to 2% on the performance of the revenue-based policy. The revenue-based policy has been selected to highlight the effect of this tuning parameter on both the network revenue and the admitted calls average satisfaction level.

Figure 8-21: Revenue Based Policy: Call Blocking Percentage

Figure 8-21 shows that the revenue-based policy call blocking percentage increases significantly from 14% to 40% at 5 calls/min traffic load when the minimum percentage improvement of network revenue is raised from 0% to 2%. At the same time, the

network-generated revenue is slightly reduced by 3.4 % for the same change (Figure 8-22).

Figure 8-22: Revenue Based Policy: Average Revenue

Increasing the minimum percentage improvement of network revenue improves the admitted calls average satisfaction level significantly as fewer calls now share the limited network resources (Figure 8-23).

Figure 8-23: Revenue Based Policy: Average Admitted Calls Satisfaction Levels

8.9 Customer's Budget Effect

As mentioned earlier, in order to study the effect of the customer budget on the QoS satisfaction level achieved, customers are classified into groups based on their admission threshold. The simulation results (Figure 8-24) show that, at 5 calls/min traffic load, the revenue-based and the market policies allocate resource to video calls in direct proportion to the customer's admission threshold. Customers with a high admission threshold enjoy a higher average satisfaction level than customers with lower admission thresholds. Since resources are not allocated based on the traffic type, the same behavior is expected for other traffic types (e.g. voice and data).

Figure 8-24: Satisfaction Levels for Various Customer's Budgets

On the other hand, in the utility policy, calls are treated equally regardless of their admission threshold. The resource is allocated in iterations. During each iteration, the resource allocation process selects the call that requires the smallest amount of resource in

order to reach a higher satisfaction level. Resources are then allocated to the selected call regardless of its budget.

8.10 Exact vs. Heuristic Revenue-Based Resource Allocation

The proposed revenue-based policy resource allocation scheme is simulated using both an exact MIP solver optimization library (OSL [43]) and the heuristic scheme described in Section 6.1.4.2. The simulation results show that the exact MIP solver generates slightly higher revenue, 1% - 7% based on the traffic load, than the heuristic scheme. However, the exact solver library has an unacceptable average processing time of 1.404 seconds compared to the heuristic average processing time of 0.00003 seconds. The exact solution time depends on the number of active calls. The higher the number of calls the longer the time consumed by the MIP solver to find the optimal solution using a depth first branch and bound solution technique.

Figure 8-25: Revenue Based Policy: Optimization Procedure vs. Generated Revenue

Although the exact solver generates more revenue than the heuristic scheme, it is much more expensive to use the exact solver since it is much slower and requires more computational power than the heuristic scheme. Therefore, we think that the heuristic scheme is more suitable for practical implementation.

8.11 Validation

To validate the simulation program, the simulation results are usually compared to the real system data or the simulation results published in the literature. Unfortunately, we were unable to validate the simulation model using real data since third generation networks are not deployed yet. At the same time we were unable to validate the simulation results with simulation in the literature results for the following reasons:

- The rate-based policy utilizes a traffic model incompatible with the third generation wireless recommended data (e.g. 30 Mbps average aggregate cell bandwidth and supports video 6 Mbps video streaming)
- The rate-based policy provides simulation results for call blocking and dropping performance metrics only.

Instead, several tests have been conducted to validate the correctness of the simulation program. First, the traffic files have been examined to insure that the traffic data follows the configured traffic model and the customer's QoS profiles follow the rules introduced in Section 5.2.1.

The simulation results obtained are also checked for consistency. The simulation results of the rate-base policy call blocking probability (Figure 8-2) show that the percentage of new calls blocked due to a budget shortage does not depend on the traffic load. At any traffic load, the percentage of customers having an excellent bid lower than the resource market price remains constant. On the other hand, the percentage of call blocking due to the lack of resource increases with the increase in the traffic load as the number of calls competing for the same amount of resources increases.

The simulation results of the market policy bandwidth market price (Figure 8-5) show that when the no-drop constraint is activated, the price adjustment process does not increase the bandwidth market price with the increase in the traffic load in order to avoid forcing an existing or handoff call to terminate. When the no-drop constraint is relaxed (i.e. the admission control policy could drop any existing or handoff call), the bandwidth market price starts to increase as the traffic load increases. This is because more customers are now competing for the same amount of bandwidth and the bandwidth is allocated to the highest bid customer first.

Figure 8-11 shows that, for the same traffic load, increasing the framework minimum admission threshold parameter for the revenue-based policy increases the call blocking percentage since more calls will have an admission threshold lower than the system minimum admission threshold.

Figure 8-26 shows that the utility-based policy average resource market price is always higher than the minimum resource market price selected by the service provider. At the same time, Figure 8-18 shows that when the framework maximum call-blocking percentage constraint is activated, the system call blocking percentage is always lower than or equal to the configured value.

Figure 8-26: Utility Based Policy: Minimum Bandwidth Market Price vs. Bandwidth Market Price

Figure 8-21 shows that, for the same traffic load, increasing the network revenue minimum improvement threshold increases the call blocking percentage. This is because as the network revenue minimum improvement threshold increases an admitting call should have enough budget such that the new network revenue after the call is admitted is higher than the current network revenue plus the new minimum revenue improvement threshold.

8.12 Conclusions

The simulation results show that the proposed framework gives the service providers the flexibility to design policies which achieve different objectives. For example, by selecting the proper combination of resource allocation, admission control, and price adjustment schemes, the service provider could generate network revenue that cannot be achieved using existing resource control schemes.

The simulation results show also that selecting a policy in order to achieve a predefined objective may worsen other performance metrics. For instance, improving the network average revenue leads to lower customer satisfaction levels and a higher call blocking percentage. However, by combining the revenue, customer's satisfaction level, and the call blocking and dropping performance metrics (adjusted generated revenue), the proposed policies show in general better performance than the fixed and dynamic policies proposed in the literature.

The simulation results verify also that the service providers can tune the performance of a selected policy further by adjusting the tuning parameters. For example, if a service provider selects the revenue-based policy and notices that the customer average satisfaction levels are not acceptable, he might then sacrifice some of the generated revenue in order to improve the satisfaction levels by increasing the minimum admission threshold (Figure 8-12, Figure 8-13).

The simulation results show also that the responsive pricing policy outperforms the rate-based policy in several respects. The responsive policy generates more revenue, provides admitted customers with a higher average satisfaction level, and blocks a smaller percentage of calls than the rate-based policy. However, the responsive pricing policy is unsuitable for third generation wireless networks due to its unacceptable 31% call dropping percentage.

Although the rate-based policy was able to reduce the call dropping percentage to a maximum of 5.4% at 5 calls/min traffic load, it has two major disadvantages. First, it did not provide a low call blocking probability; it blocks on average 41% of the new admitting calls. Second, it generates much lower network revenue compared to the other policies. The revenue generated by the rate-based policy is 15% lower than the next higher policy and 41% lower than the revenue-based policy.

The simulation results show also that the revenue-based policy generates the highest network revenue at the cost of providing the admitted calls with the lowest average satisfaction level. On the other hand, by guiding the bandwidth market price lower, the utility-based policy provides admitted calls with the highest average satisfaction level and generates the second lowest network revenue with respect to the other policies.

Among all the simulated policies the market-based policy is the only one that does not have a major disadvantage. It has the lowest call blocking percentage. It also generates the

second highest network revenue and provides admitted customers with a better average satisfaction level than the revenue-based policy.

In a nutshell, the proposed framework policies manage the network resources more efficiently than both the rate-based and the responsive pricing policies. The proposed policies generate higher revenue and provide admitted calls with higher average satisfaction level. More importantly, the proposed policies reduce both the call blocking and dropping percentages significantly.

Furthermore, the proposed framework gives service providers the flexibility to select the ICRC policy that best fits their needs. Service providers may select the market-based policy and rely on the customer's selfish behavior to drive the network performance. They may also select either the revenue-based or the utility-based policy and use the framework tuning parameters to tune the performance of the selected policy to improve the average satisfaction level provided to admitted customers or improve the generated revenue.

Chapter 9

CONCLUSIONS

- This thesis proposed a novel QoS aware integrated charging and resource control framework as a solution for the wireless networks resource control problems described in Chapter 4.
- As shown in Chapter 6, the proposed framework has the flexibility to support different ICRC policies. The framework allows service providers to select or implement the ICRC policy that best fits their marketing strategy.
- Chapter 8 shows that the policies supported by the proposed framework outperform other currently proposed wireless and wireline resource control policies in various aspects.
- The simulation results in Chapter 8 prove that improving the resource control efficiency allows the proposed framework to significantly minimize the call dropping and blocking percentages, increase the network generated revenue, and/or improve the customer's satisfaction level.

- By adjusting the amount of resources allocated to existing active calls, the reservation-less handoff scheme significantly reduces the call dropping probability, but does degrade the customer QoS marginally.
- The reservation-less handoff scheme does not affect the resources utilization level, however it slightly reduces the network revenue and the customer's average satisfaction level. The network revenue is affected because the price adjustment process may have to lower the bandwidth market price to prevent a premature termination of an ongoing call.
- Service provider's satisfaction and customer's satisfaction contradict each other. Increasing the network revenue decreases the customer's average satisfaction level and vice versa.
- Increasing the resource market price is not always the proper approach to increasing the service provider revenue. Based on the network conditions, reducing the resource market price may improve the network utilization level and generate higher revenue.
- Dynamic pricing schemes generate more revenue than fixed pricing ones.
- Adjusting the bandwidth market price dynamically based on the customer's budget generates more revenue than adjusting the resource market price based on the resource utilization level.

- Charging customers differently based on their QoS parameters encourages customers to reveal their true QoS requirements.
- The proposed framework scores well against the assessment criteria introduced in Section 3.2.

9.1 Summary of Contributions

The main contributions of this thesis are summarized as follows:

- Developed a new technique that allows customers to efficiently convey their QoS requirements (QoS profile).
- Developed a novel QoS aware integrated charging and resource control framework and demonstrated its superiority to existing and proposed policies via simulation. The framework has several advantages:
 - a. Provides soft QoS services in terms of customer's QoS perception.
 - b. Allows resource usage adaptation based on the traffic conditions and customer's QoS perception.
 - c. Provides an effective new set of performance control (tuning) parameters.
 - d. Is flexible enough to allow service providers to design new policies and tune existing ones.
- Proposed a new QoS based charging model that:

- a. Charges customers dynamically based on the network conditions and the amount of resource they have asked for.
 - b. Persuades customers to reveal their true QoS requirements.
 - c. Increases the customer's confidence in the dynamic charging scheme by using their budgets as the maximum limit for the call price.
- Developed three new ICRC policies: revenue based, utility based, and market based.
 - Developed a reservation-less handoff handling technique that gracefully degrades ongoing call's quality to accommodate handoff calls.
 - Introduced new performance metrics for multimedia systems operating on soft QoS.

9.2 Future Research

In general, there is very limited research in next generation wireless networks charging techniques. Research for charging within integrated networks (e.g., cellular with W-LANs, ad hoc networks) and for asymmetry bandwidth allocation are examples for areas open for further research.

In this work only one type of wireless resource has been considered, namely the cell bandwidth. Future research should add the cell limited power budget to the resource control schemes. Joint power control and bandwidth assignment techniques might be required since adjusting a call power level may affect the amount of bandwidth that should be allocated to that call in order to achieve a predefined minimum BER level.

Future work is required to study the performance of the proposed framework under different handoff and movement patterns. Unbalanced handoff patterns where the handoff-call arrival and departure rates are different (sink and source cells) should be considered.

Further research may focus on the development of new optimal resource control schemes by taking advantage of the flexibility and the efficiency of the new customer QoS aware framework rules proposed in this work. It may also focus on the analysis of different fairness aspects such as fairness across the same type of traffic and across different traffic types.

The inability to predict the call cost before the call is started is one of the major disadvantages of dynamic charging schemes. Further research is required to study the improvement in the customer's confidence when the dynamic charging scheme charges them up to their declared budget.

REFERENCES

- [1] 3G TR 23.907, “QoS Concept”, version 1.2.0, May 1999
- [2] 3GPP TS 23.107, “QoS Concept and Architecture”, version 3.7.0, Jan. 2002.
- [3] 3GPP TS 26.071, “Mandatory Speech Codec speech processing functions”, version 3.1.0, 1999.
- [4] 3GPP TS 26.171, “AMR speech codec, wideband; General description”,
- [5] D.R. Anderson, D.J. Sweeney, T.A. Williams, “Statistics for Business and Economics”, 7th ed. International Thomson Publishing, 1998
- [6] The ATM Forum, Technical Committee, “Traffic Management Specification”, Version 4.0, af-tm-0056.000, April 1996.
- [7] R.T. Apteker; J.A. Fisher; V.S. Kisimov; H. Neishlos, “Distributed Multimedia: User Perception and Dynamic QoS,” High-Speed Networking and Multimedia Computing, San Jose, CA, February 1994, pp. 226-34.
- [8] P. Bahl, I. Chlamtac, and A. Farago, “Resource Assignment for Integrated Services in Wireless ATM Networks,” International Journal of communication Systems, April 1998, pp. 29-41
- [9] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, “CDMA/HDR: A bandwidth efficient high-speed wireless data service for Nomadic users,” IEEE Communications Magazine, July 2000, pp. 70-77
- [10] D. Bertsekas and R. Gallager, Data Networks, Prentice Hall, NJ, 1987.
- [11] A. Bouch and M.A. Sasse “It ain't what you charge it's the way that you do it: A customer perspective of network QoS and pricing,” Proc. IFIP/IEEE International Symposium on Integrated Network Management (IM'99), May 1999
- [12] A. Campbell, G. Coulson, F. Garcia, D. Hutchison, and H Leopold “Integrated Quality of Service for Multimedia Communications,” In IEEE Infocom 93, San Francisco California, March 1993.

- [13] Z. Cao and E.W. Zegura. "Utility Max-Min: an Application-Oriented Scheme," Infocom'99. March 1999.
- [14] W. Chen, C. Huang, and S. Ding "A Scheme for QOS Control in ATM Switching Systems," In IEEE ICC 95, Seattle Washington, June 1995.
- [15] D.S Clark, "Internet cost allocation and pricing," in Internet Economics, Eds McKnight and Bailey, MIT press, 1997
- [16] R. Cocchi, S. Shenker, D. Estrin and L. Zhang, "Pricing in computer networks: motivation, formulation, and example," IEEE/ACM Transaction on Networking, vol. 1, issue 6, 1993
- [17] C. Courcoubetis, V. Siris, G.D Stamoulis, "Charging Issues for Services in Broadband Networks," Proceedings of the OECD Conference on the Economics of the Information Society, Istanbul, December 1995.
- [18] R.L. Cruz, "Quality of Service Guarantees in Virtual Switched Networks". IEEE Journal of Selected Areas on Communication, no.13, vol. 6, August 1995.
- [19] S.K. Das, M. Chatterjee, and N.K. Kakani, "QoS Provisioning in Wireless Multimedia Networks," Wireless Communications and Networking Conference (*WCNC '99*), New Orleans, Louisiana, pp. 1493-1497, October 1999.
- [20] S.K. Das, R. Jayaram, N. K. Kakani and S.K. Sen, "A call admission and control scheme for quality-of-service (QoS) provisioning in next generation wireless networks", Wireless Networks vol. 6, issue 1, 2000, pp.17-30.
- [21] S.K. Das, R. Jayaram, S.K. Sen, "An optimistic quality-of-service provisioning scheme for cellular networks", Proc. of IEEE International Conference on Distributed Computing Systems (ICDCS), Maryland, Baltimore, pp. 536-542, May 1997.
- [22] S.K. Das and S.K. Sen, "Quality-of-Service Degradation Strategies in Multimedia Wireless Networks," IEEE VTC '98, Ottawa, Canada, May 1998.
- [23] L. Dasilva, "Pricing for QoS-Enables Networks: A Survey," IEEE Communications Surveys, 2000.

- [24] N. Davies, "The impact of mobility on distributed systems platforms", Proceedings of the IFIP / IEEE International Conference on Distributed Platforms, Dresden, 1996 pp 18-25
- [25] E. Del Re, R. Fantacci, and G. Giambene, "Handoff and Dynamic Channel Allocation Techniques in Mobile Cellular Networks," IEEE. Trans. on VTC, vol. 44, no. 2, May 1995.
- [26] M. El-Kadi, S. Olariu, H. Abdel-Wahab, "A Rate-Based Borrowing Scheme For QoS Provisioning in Multimedia Wireless Networks," IEEE trans. on parallel and distributed systems, vol. 13, no. 1, Jan 2002.
- [27] ETSI EN 301 703, "Adaptive Multi-Rate (AMR); Speech processing functions, General description", GSM 06.71, version 7.0.2, 1998.
- [28] ETSI TS 126 235 "Packet switched conversational multimedia applications; Default codecs", 3GPP TS 26.235 version 4.0.0 Release 4
- [29] ETSI TR 101 112, "Selection procedures for the choice of radio transmission technologies of the UMTS," UMTS 30.03 version 3.2.0
- [30] M. Falkner, M. Devetsikiotis and I. Lambadaris, "An Overview of Pricing Concepts for Broadband IP Networks," IEEE Communications Surveys, 2000.
- [31] M. Frodigh, S. parkvall, C. Roobol, P. Johansson, and P. Larsson, "Future Generation Wireless Networks," IEEE Personal Communications magazine, October 2001, pp. 10-17.
- [32] E.W. Fulp and D. Reeves, "The Fairness and Utility of Pricing Network Resources Using Competitive Markets", Computer Networks Journal, Feb. 2000.
- [33] E.W. Fulp, M. Ott, D. Reininger and D.S. Reeves, "Paying for QoS: An Optimal Distributed Algorithm for Pricing Network Resources", Proc. of the IEEE Sixth International Workshop on Quality of Service, pp 75-84, 1998
- [34] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi "GNU Scientific Library Reference Manual," Edition 1.1, for GSL Version 1.1, 9 January 2002, http://sources.redhat.com/gsl/ref/gsl-ref_toc.html.

- [35] G. Gallassi, G. Rigolio, and L. Fratta "ATM: Bandwidth Assignment and Bandwidth Enforcement Policies," In IEEE Globecom 89, Dallas Texas, November 1989.
- [36] A. Girard, "Revenue Optimization of Telecommunication Networks", IEEE Transaction on Communication, vol. 41, pp. 583-591, April 1993.
- [37] J. Greenwood and T. Mukoyama, "The Effect of Income Distribution on the Timing of New Product Introduction", Department of Economics, University of Rochester, Rochester, November 2001
- [38] A. Gupta, D.O. Stahl, A.B. Whinston, "Priority Pricing of Integrated Services Networks, Internet Economics", Eds McKnight & Bailey, MIT press, 1997.
- [39] A. Gupta, D.O. Stahl, and A.B. Whinston, "The Internet: A Future Tragedy of the Commons," Conference on Interoperability and the Economics of Information Infrastructure, July 6-7, 1995
- [40] J. Hirshleifer, "Price Theory and Applications", third edition, Prentice-Hall, 1984
- [41] Y. T. Hou, H. Tzeng, and S.S. Panwar, "A Weighted Max-Min Fair Rate Allocation for Available Bitrate Service," In Proc. of the IEEE GLOBECOM, pp. 492 - 497, 1997.
- [42] J.F. Huard, and A.A Lazar, "On QoS mapping in Multimedia Networks," 21th IEEE COMPSAC, August, 1997, Washington, D.C.
- [43] M.S. Hung, W.O. Rom, and A.D. Waren, "Optimization with IBM OSL," June 1993, ISBN 0-89426-244-0
- [44] IEEE 802.16 broadband wireless access working group, "Quality of service classes for BWA", July 22, 1999.
- [45] ISO/IEC 14496-2, "Information technology - Coding of audio-visual objects - Part 2: Visual", 1999
- [46] ISO/IEC JTC1/SC29/WG11 N2724, "MPEG-4 Applications", Seoul, March 1999.
- [47] ISO/IEC JTC1/SC29/WG11 N2325, "Overview of MPEG-4 Profiles and Levels", July 1998

- [48] ISO/IEC JTC1/SC29/WG11 N4030, "Overview of the MPEG-4 Standard", March 2001.
- [49] ISO/IEC JTC1/SC29/WG11/N2425 "MPEG-4 Audio Verification Test Results: Audio on Internet," October 1998.
- [50] ITU-R BT.500-10, "Methodology for the subjective assessment of the quality of television pictures", 2000
- [51] ITU-T Recommendation H.263, "Video coding for low bitrate communication".
- [52] ITU-T Recommendation H.263 (annex X), "Annex X, Profiles and levels definition".
- [53] A. Jalali, R. Padovani, R. Pankaj, "Data Throughput of CDMA-HDR, a High Efficiency Data Rate Personal Communication Wireless System," VTC'2000, pp.1854-1858.
- [54] N. K. Kakani, S.K. Das, S.K. Sen, and M. Kaippallimalil, "A Framework for Call Admission Control in Next Generation Wireless Networks," Proc. of the First ACM International Workshop on Wireless Mobile Multimedia (WoWMoM '98), Dallas, pp. 101-110, Oct 1998
- [55] F.P. Kelly, "Notes of Effective Bandwidth," In Stochastic Networks: Theory and Applications, Oxford University Press, 1996. 141-168.
- [56] F.P. Kelly, A.K. Maulloo, and D.K.H. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability," Journal of the Operational Research Society 49, pp. 237-252, 1998.
- [57] F.P. Kelly, "Charging and rate control for elastic traffic", European Transactions on Telecommunications, vol. 8, pp 33-37, 1997
- [58] H. Kikuchi, "(M+1)st-Price Auction," Proc. of The Fifth International Conference on Financial Cryptography '01, IFCA, pp.291-298, February 2001.
- [59] S. Kim, T. Kwon, and Y. Choi, "Call Admission Control for Prioritized Adaptive Multimedia Services in Wireless/Mobile Networks," IEEE Vehicular Technology Conference (VTC '00 spring), Tokyo, May 2000.

- [60] P. Klemperer, "Auction Theory: A Guide to the Literature," *Journal of Economic Surveys*, no 13, 1999, pp.227-286.
- [61] T. Kwon, Y. Choi, C. Bisdikian and M. Naghsineh, "Call admission control for adaptive multimedia in wireless/mobile networks," *Proc. of first ACM international workshop on wireless mobile multimedia*, Dallas, TX USA, Oct. 1998, pp. 111-116
- [62] T. Kwon, S. Das, I. Park and Y. Choi, "Bandwidth Adaptation Algorithms with Multi-Objectives for Adaptive Multimedia Services in Wireless/Mobile Networks," *ACM WoWMoM '99*, Seattle, Aug.1999
- [63] J. Kurose "Open Issues and Challenges in Providing Quality of Service Guarantees in High-Speed Networks," *Computer Communication Review*, 23(1), January 1993.
- [64] P. L'Ecuyer, "Maximally Equidistributed Combined Tausworthe Generators", *Mathematics of Computation*, 1996, pp. 203-213
- [65] P. L'Ecuyer, "Random Number Generation", Chapter 4 of the *Handbook on Simulation*, Jerry Banks Ed., Wiley, 1998, pp. 93-137.
- [66] J.K. Mackie-Mason, "A Smart Market for Resource Reservation in a Multiple QoS Network," 25th Annual Telecom Policy Research Conference, Alexandria, VA, September 28, 1997.
- [67] J.K. MacKie-Mason, L. Murphy, and J. Murphy, "The Role of Responsive Pricing in the Internet", *Journal of Electronic Publishing*, U. of Michigan Press, May 1996.
- [68] G. Malewicz and A.A. Shvartsman, "An Auction-Based Flexible Pricing Scheme for Renegotiated QoS Connections and Its Evaluation", *Proc. of Seventh International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'99)*, 1999.
- [69] P. Milgrom, "Auctions and bidding: a primer", *Journal of Economic Perspectives*, 1989, pp. 3-22.
- [70] M. Naghshineh, and A.S. Acampora, "QoS Provisioning in Micro-Cellular Networks Supporting Multimedia Traffic", *Proc. IEEE Veh. Technol. Conf.*, 1995, pp. 1075-1084

- [71] S. Nelakuditi, R.R. Harinath, S. Rayadurgam and Zhi-Li Zhang, "Revenue Based Call Admission Control for Wireless Cellular Networks", International Conference on Personal Wireless Communications, Feb. 1999.
- [72] W. Nicholson, "Microeconomic Theory, Basic Principles and Extensions", The Dryden press, 1989
- [73] A. Odlyzko, "Paris Metro Pricing: The Minimalist Differentiated Solution," Proc. of the IEEE/IFIP Int. Workshop on Quality of Service - IWQoS'99, June 1999.
- [74] 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 comm., vol. 16, no. 6, Aug. 1998.
- [75] E. Scharf, "First Steps towards the Development and Assessment of Effective ATM Charging Algorithms", IEE Colloquium on Broadband Charging, November 1996.
- [76] G. Senarath and D. Everitt, "Performance of handover priority and queuing systems under different handover request strategies for microcellular mobile communication systems," Proc. 45th IEEE VTC, 1995, pp. 897-901.
- [77] R. D. Sinclair, "What does the World Income Distribution Look Like?" The Maxwell School of Syracuse University, Department of Economics, November 2000, version 2.2
- [78] S. Singh, "Quality of Service Guarantees in Mobile Computing," Journal of Computer Communications, vol. 19, pp. 359-371, 1996.
- [79] D. Songhurst and F. Kelly, "Charging Schemes for multiservice networks," ITC15. Amsterdam, pp. 781-790, 1997.
- [80] M. Srivastava, and P. P. Mishra, "On Quality of Service in Mobile Wireless Networks", Proc. IEEE Veh. Technol. Conf., 1997, pp. 147-158
- [81] A. Takayama, "Mathematical Economics", Cambridge University Press, 1985
- [82] N. D. Tripathi, J.H Reed, and H.F. VanLandingham, "Handoff in Cellular Systems", IEEE Personal Communications Dec 1998.

- [83] T. Yamazaki and J. Matsuda, "On QoS Mapping in Adaptive QoS Management for Distributed Multimedia Applications," Proc. ITC-CSCC'99, vol.2, pp. 1342-1345, July, 1999
- [84] Z. Wang and J. Crowcroft, "Quality of Service Routing for Supporting Multimedia Applications". IEEE Journal of Selected Areas on Comm., no. 14, vol. 7, Sept. 1996.
- [85] R.M. Wood, G. Fankhauser, R. Kreula and E. Monni, "QoS Mapping for Multimedia applications in Wireless ATM Environment", ACTS Mobile Summit 97, Aalborg, October 1997.
- [86] G.M. Woodruff and R. Kositpaiboon "Multimedia Traffic Management Principles for Guaranteed ATM Network Performance," IEEE Journal on Selected Areas in Communications, no. 8, vol. 3, April 1990.