# A Novel Pricing Approach to Support QoS in 3G Networks

Prepared by:
Gavole Vitalis Ozianyi

Supervised by:
Neco Ventura

Department of Electrical Engineering
University of Cape Town
2006

This dissertation is submitted to the University of Cape Town
in fulfilment of the academic requirements
for the Degree of Master of Science in Engineering

20 January 2006

# Declaration

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

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

Signed by candidate

Signature removed

Vitalis G. Ozianyi

20-01-2016

Date

# Acknowledgements

I would like to express my sincere gratitude to the following individuals and organisations for their assistance during the course of this project.

Mr. Neco Ventura, for his supervision and guidance throughout the project.

David Waiting and Eugene Golovins, to whom I am grateful for their advice and constructive feedback.

My fellow colleagues in the Communications Research Group (CRG) and Speech Technology (STAR) group at UCT, for the useful discussions during various presentations.

My parents, sisters and brothers, for their love and constant support in every aspect of my life. Thank you.

# Synopsis

The provision of services in communication networks requires network infrastructure that is often expensive and requires maintenance. Network bandwidth and other resources are needed for the delivery of the services, and when the demand for resources exceeds the supply network congestion will occur. Network congestion compromises the quality of service (QoS) performance of multi-service networks, since the transport characteristics of some applications (e.g., real-time multimedia) would not be met.

Pricing in communication networks is commonly used to achieve congestion control. By pricing, network operators and service providers aim at facilitating responsible use of limited network resources to improve user satisfaction and lead to the maximisation of profits. The optimum tariff rates used for charging of mobile services are affected by factors like the market forces affecting the industry. Generally, the tariff rates increase with the guaranteed or the achieved QoS level.

3G networks are designed to offer more bandwidth and guaranteed QoS, thus users would be able to use real-time multimedia and other QoS-sensitive applications. If legacy pricing schemes will be retained in these networks, users will not get the incentives to utilise the enhanced capacity.

This study investigates a pricing approach that introduces service profiles into DiffServ-enabled 3G networks, in which the tariff rates depend on the degree of congestion in the network. The study proposes a hybrid pricing scheme that incorporates characteristics of different pricing proposals to achieve the goals of pricing of 3G and other IP network services. The pricing and QoS control algorithm for the platinum and the gold profiles is developed. It will be shown that the concept of resource sharing between profiles results in improved network performance in terms of efficient resource utilisation, improved user satisfaction and maximisation of profits.

An evaluation framework that supports the assesment of connection admission control, profile management and traffic control mechanisms is designed and implemented. This framework allows QoS performance of different profiles to be analysed, with specific focus on network throughput under varying levels of bandwidth, different queue priority and resource sharing.

# Contents

# List of Figures

# List of Tables

# Glossary

This section defines some of the commonly used terms and abbreviations that appear throughout this document.

**1G** The first generation of analog mobile phone technologies, including AMPS, TACS and NMT.

**2G** The second generation of wireless communication systems (including GSM, CDMA IS-95 and D-AMPS IS-136) using digital transmission and advanced control techniques to improve the performance of voice communications, provide special features and limited digital messaging capabilities, such as GSM.

**3G** The third generation of mobile phone technologies covered by the ITU IMT-2000 family. It allows greater bandwidth and opens the way to increased data-over-wireless solutions.

**3GPP** The 3rd Generation Partnership Project is a global collaboration between 6 partners: ARIB, CWTS, ETSI, T1, TTA, and TTC. The group aims to develop a globally accepted 3rd-generation mobile system based on GSM.

**Access Point (AP)** An entity that bridges information between the wireless medium and the distribution medium on behalf of its associated stations. Access points are only used in infrastructure wireless LANs.

**CAPEX** Capital Expenses include all the costs necessary to develop and make available the elements of the network. This is basically the costs of equipment for an operator i.e., hardware plus software and all costs directly relative to the network element (installation, upgrade etc).

**DiffServ** An architecture for providing different types or levels of service for network traffic. One key characteristic of diffserv is that flows are aggregated in the network, so that core routers only need to distinguish a comparably small number of aggregated flows, even if those flows contain thousands or millions of individual flows.

**GPRS** The General Packet Radio Services is a GSM Packet Based bearer for the delivery of data services. With GPRS charges are based on the amount of information downloaded rather than the duration of the connection.

**GSM** The Global System for Mobile Communications is one of the leading digital cellular systems. It uses narrowband TDMA, which allows eight simultaneous calls on the same radio frequency.

**IntServ** Integrated Services is a model used for providing traffic forwarding service levels in IP/MPLS networks. It allows for microflows to be created with reserved resources (such as bandwidth) and other traffic handling characteristics. Traffic is pushed into these microflows in the direction of the required destination. The disadvantage of IntServ is that the microflows must be explicitly traced and reserved, and this can cause scalability problems.

**IP** Internet Protocol provides for the transmission of datagrams from a source to a destination. The source and destination are hosts identified by fixed-length IP addresses.

**MMS** Multimedia Message Service, a method of transmitting graphics, video clips, sound files text messages over wireless networks using the WAP protocol.

**OPEX** Operation expenses is typically the cost of operation for the operator (e.g. maintenance, sometimes software updates etc).

**PHB** Per-hop-behaviour is a description of the externally observable forwarding treatment applied at a differentiated services-compliant node to a behavior aggregate.

**QoS** Quality of Service is the idea that transmission rates, error rates, and other characteristics of a communications network can be measured, improved, and, to some extent, guaranteed in advance.

**SLA** Service Level Agreement is a contractual agreement between a service provider and a subscriber specifying the QoS parameters that the subscriber can expect to receive.

**UMTS** Universal Mobile Telecommunications Service, part of the IMT-2000 initiative, is a 3G standard supporting a theoretical data throughput of up to 2 Mbps.

**VoIP** Voice over IP is the two-way transmission of voice information over a packet-switched TCP/IP network. (Also known as "IP telephony".)

**WAP** The Wireless Application Protocol is a secure specification that allows users to access information instantly via handheld wireless devices such as mobile phones, pagers, two-way radios, smartphones and communicators.

# Chapter 1

# Introduction

## 1.1 Background Information

Mobile communication networks are made of sophisticated and expensive switching devices that are interconnected using different media. The maintenance and upgrading of the networks is a continuous process. Service providers invest a lot in installing and maintaining the networks so as to provide communication services to their customers. Users expect communication services that are of the best possible quality and which are available whenever needed. The convergence of telecommunication and data networks, as well as the increasing demand for multimedia applications creates the need for a properly managed service provisioning environment. Charging for the delivery of communication services is a way of introducing responsibility and accountability between the users, network operators and service providers. The charging information generated in the network is composed of different components that correspond to the various service providers, content providers and operators who participate in the delivery of the end-to-end service to the user. Different charging schemes have been proposed and used for different network services [1].

The driving force in the choice of a particular pricing /charging[1] scheme is its simplicity and the maximising of revenue for the network operators. In terms of simplicity, the flat rate charging scheme is at the top of the list [2]. The network

[1] The terms pricing and charging are used interchangeably in this document. The definition of these terms are given in appendix A.1

overhead as a result of the use of the flat rate charging scheme is close to zero. In flat rate charging, both the users and the service providers can predict their expenses and incomes respectively [3]. However, this charging scheme only thrives in a network offering best-effort service (e.g. the Internet); hence it has some shortfalls, such as failing to provide an atmosphere for quality of service (QoS) guarantee [1]. There is an imbalance between the benefit of heavy utility versus light utility users of flat-rate networks and there is a possibility of wastage of network resources, since charges incurred by users are independent of the volume of traffic transferred or the duration of use.

The disadvantages of the flat-rate scheme are partly overcome by usage-based charging schemes. With usage based charging, the users pay according to the measured level of service usage [2]. In voice telephony (both mobile and fixed), the computation of the charge considers the duration of the call and the distance between the source and destination [4]. The general trend in mobile communications is to charge data services according to the volume of traffic transferred, bandwidth allocations and the achieved QoS [5]. This type of charging makes the costs incurred by the users to depend on the volume of data transferred. In mobile networks, packet switching (PS) protocols were introduced in the 2nd generation of the Global System for Mobile communications (GSM). QoS limitations, common with the packet switched protocol, were noted with the introduction of the GPRS system. The 3G network design has dealt with QoS challenges [6, 7]. QoS provisioning in mobile networks was necessary to meet the demand for a variety of new services. The introduction of multimedia applications in 3G and future generation networks poses a great challenge to the network operators in the choice of the appropriate charging scheme. The ideal charging scheme would give necessary incentives that encourage the users to utilise the full capacity of the network and at the same time it would enable the network operators and the service providers to maximise their revenue.

The benefits of usage-based charging can be summarised as follows:

- The charges incurred by users reflect the usage of the network resources.
- It helps overcome some of the drawbacks of the flat rate charging scheme.
- It can be structured in a way that gives incentives for users to limit their traffic and contribute to congestion control (e.g., time-of-day pricing) [5]

- When combined with QoS provisioning, improved user satisfaction encourages service providers to explore new market scenarios.

## 1.2 Problem Description

The developers of services and network management systems in the mobile domain tend to adopt from existing fixed networks. Whereas the adoption works perfectly for some services, it does not map exactly when it comes to charging. The evolution of mobile communication networks for 2G to 3G has facilitated the introduction of new services (e.g. real-time multimedia transport), which require guaranteed network performance characteristics [6]. New pricing and charging mechanisms are required for billing the new services. Although the adequacy of using the charging schemes mentioned above for 2G networks may not be challenged, it is clear that 3G and future generation network services should be charged differently. The reasons for this proposition are as follows:

- The rapid growth in the number of users of mobile services combined with the continued introduction of new applications render legacy charging methods insufficient. Research has shown that users are often willing to use network services at different tariff rates [8]. The conditions under which users are ready to pay more for the same service are of great importance to network operators. The priority (importance) attached to the service by the users can easily influence their willingness to pay a higher fee, especially when there is a guarantee that the services will be delivered during adverse periods (e.g. when congestion is high). The charging mechanisms currently used for mobile services do not provide service providers with a method of fully exploiting such scenarios; neither do they empower users to optimise on similar situations.

- The situation described above becomes conspicuous when dealing with traffic from different applications. The network characteristics of different applications are diverse, and they dictate how the applications should be handled on the network. For example multimedia applications require higher priority at congested nodes; on the other hand file transfer applications can tolerate considerable delay. The current charging schemes for packet switched services do not consider the network characteristics of the applications and hence they are insufficient for use in multi-service networks like the 3G [9].

- When congestion occurs in the network, the service quality is bound to degrade unless adequate mechanisms are put in place to counter the problem. Charging has been used as a congestion control mechanism, especially for mobile and fixed telephone services, in what is commonly known as the time-of-day (ToD) pricing [5]. Time-of-day assumes that network congestion is greater during the day than at night, hence the former is charged higher than the latter. Even though ToD pricing has been successful for some telecommunication services, it does not scale properly to data services and multimedia traffic patterns, since network congestion tends to vary on short time frames.

- The interaction between QoS and pricing creates a complex system in which the service providers aim at maximising their revenue while users want to get the best service at the lowest possible cost [8]. The present mobile communications architecture does not provide the two parties with appropriate interfaces and solutions for achieving a balance between their diverse interests.

- In present generation networks the charging function requires accounting information, the pricing tariff and the profile of the user. In this case the profile is basically a static service level agreement (SLA) [10] between the user and the service provider, and hence it would remain the same for long periods, e.g. months. In 3rd and future generation networks it would be important for users to change their profiles as frequent as they wish [9]. This means that more information will have to be transported to the charging function at short intervals.

- The differentiated services (DiffServ) architecture has been proposed for QoS provisioning in IP networks, to which 3G is part [6]. In DiffServ, traffic differentiation is performed, and resource allocations to different classes of service are defined; however, resource guarantees are not available to competing traffic sources. With this approach traffic sources or users whose traffic falls in the same DiffServ class are not assured of service when the aggregate volume of traffic exceeds what the core network routers are configured to forward [9].

## 1.3 Thesis Objectives

This study aims to explore the importance of using a charging model that considers the QoS requirements and the network transport cost of providing services to the users. The design of the proposed scheme intends to utilise the charging functionalities that have been proposed and/or deployed in 3G networks (e.g. the multi-level charging architecture that has been proposed in the 3GPP) [4]. A new pricing or rating scheme is proposed and investigated on a test-bed.

This thesis aims to show the need of a QoS-aware charging model as necessitated by the convergence of mobile communication networks and the Internet, and the evolution of mobile networks towards an all IP system [11]. The study aims to prove that the overlaying of structured network service profiles onto a network using the DiffServ architecture [12] for QoS provisioning and the use of connection admission control (CAC) will facilitate the achievement of individual QoS guarantees. The performance of the proposed QoS provisioning approach will be investigated on an evaluation framework.

This thesis investigates the benefits of dynamic pricing schemes, and aims to prove that a well structured multi-tariff charging scheme, whose tariff levels are related to appropriate QoS levels will achieve a 3G network management system that is flexible, economically efficient, socially fair, and which achieves high levels of network efficiency. The system will enhance interaction between the users and the network, enabling users to influence the cost vs. QoS relationship of the service. This would address the pricing requirements of multi-service QoS-enabled networks.

Using design architecture, this study aims to prove that the provision of interfaces for users to modify their network service profiles will enable them to easily influence the cost vs. QoS relationship, and that the incorporation of network profile management agents can be used to enforce this relationship. This thesis also aims to prove that the proposed scheme will lead to the achievement of the goals mentioned above, thus improving user satisfaction, and hence providing an environment for the network operators and service providers to maximise their revenue.

## 1.4 Scope and Limitations

This study considers the pricing of the UMTS bearer service resources, i.e. the radio access bearer (RAB) and the core network (CN) bearer service resources [6]. It explores the management of the IP bearer service, for the support of interworking between the UMTS and external IP networks that use DiffServ for QoS provisioning. However, little attention is given to the technical details involving the implementation of DiffServ. In the design of the proposed charging scheme, the policy based admission control for the allocation of resources in 3G networks is considered. In the evaluation framework, the provision of QoS is done using a combination of traffic control measures and connection admission control. The pricing algorithm and QoS evaluations are not only applicable to 3G networks, but they also apply to other wireless networks, and fixed networks that use the Internet protocol.

This thesis does not deal with the allocation of radio channels for the support of the RAB services. In addition to this, only relative values are used in describing the amounts of the different resources that are needed to support QoS in the network.

This study does not consider the charging requirements of the different applications and services, which might be delivered to the users over the 3G network. Since it is common for the network operator to be in charge of billing the users (the users are normally presented with a comprehensive charge for the services) [4], the effect of the proposed charging scheme on this arrangement is considered. However, as indicated earlier, this study only considers the utilisation of resources required for the transport of user data in the UMTS network. The details of content charging or the apportioning of revenue to various providers and operators is not handled explicitly.

## 1.5 Theses Outline

The remaining sections of this document are structured as follows:

- Chapter 2 gives a review of pricing and charging schemes (e.g., flat-rate and usage-based), both proposed and those commonly used, for the Internet and 3G networks. Special attention is given to dynamic and QoS-based pricing schemes.

- In chapter 3, the focus is on the architectural design of the proposed charging scheme. The architecture encourages the use of a QoS-aware pricing scheme that is achieved through a dynamic pricing strategy, for the support of QoS in 3G networks. The design of the end-to-end QoS provisioning architecture, together with the interfaces to facilitate user-network interaction is presented in this chapter. The network management architecture of the proposed charging scheme, incorporating several network agents is given. A discussion of the feasibility of implementing the proposed charging scheme on 3G networks and an overview of the practical challenges that this methodology faces finalise this chapter.

- In chapter 4, the design and implementation of the framework on which the proposed scheme was evaluated is presented. The design and implementation demonstrate the operation of the entities of the proposed scheme.

- Chapter 5 illustrates the tests and evaluations that were performed on the testbed. The results that were obtained and observations are presented, and a discussion of each evaluation step is given. This study uses the results to prove the feasibility of the proposed charging scheme.

- Chapter 6 concludes the work done in this thesis and gives recommendations for future work to be done on the proposed charging scheme for the support of QoS billing in 3G and future generation networks.

# Chapter 2

# Literature Review

This chapter provides a detailed overview of the literature pertaining to QoS, pricing. charging and billing in Internet and mobile communication networks. It provides the motivation for a network congestion dependent pricing scheme for a system in which users are able to specify their QoS requirements and willingness to pay for the services. The chapter emphasises the benefits of a hybrid billing scheme that uses a dynamic pricing strategy. while enhancing interaction between users and the system, and compares it to legacy pricing schemes for mobile services. The relationship between QoS and network pricing is elaborated.

## 2.1 The Role of Pricing and Charging in Communication Networks

It was mentioned in Chapter 1 that network infrastructure is required for the provision of communication services. New technologies that improve network efficiency and capacity have been deployed to meet the increasing demand for new and better communication services. Network operators and service providers recover the incurred capital and maintenance costs through imposing levies for the use of services offered on mobile networks [1]. The demand for communication services/resources is an important factor in the utilisation of network capacity. Generally, the demand varies with periods of the day. When the demand for resources exceeds the network capacity, congestion would occur; hence congestion control and management is required. Charging is used for regulating the use of network resources through applying price premiums to some network services and increasing the tariffs to levels that only a few users are ready to pay [1, 5].

Time-of-day (ToD) pricing is a network pricing strategy that is used to manage congestion in networks offering telephony and other services. In ToD, the prices for network services are set to a high value during the day and to a lower value at night. ToD relies on the fact that the demand for communication services is generally higher during the day than at night, thus the high daytime tariffs are meant to deter some users from using the network. It is normal for many users to attempt to minimise their communication costs, thus price-sensitive users would use the network when network prices are low [13]. Reduced demand for network resources leads to sustainable congestion levels.

The demand for network resources by users can be classified into two categories, i.e. heavy-utility and light-utility usage [3]. Heavy-utility users tend to use the network for longer periods as compared to the light-utility users. If network services were offered at no cost, some users (e.g. the heavy-utility users) would remain connected to the network "indefinitely". The capacity of every network is limited, and when call admission control (CAC) is used for congestion control, new users are granted access to the network only when capacity is available. It is through charging that the heavy-utility users are forced to withdraw from the network to minimise their communication costs, thus creating capacity for other users. Charging in this case is used to promote social fairness among users.

In addition to the fostering of fairness among users, charging places responsibility on the network operator to provide high quality services to users and independent service providers. Users are sensitive to the quality of the chargeable services they receive [14]. By creating a relationship of responsibility and accountability between the users, network operators and service providers, charging obligates service providers to provide services whose quality tallies with the market price. Properly structured billing schemes have facilitated the formation of business relationships in the communications industry, and this has led to innovative ideas (e.g. mobile roaming) [15]. Mobile roaming enables users to stick to the same number and mobile device while moving across networks owned by different operators.

## 2.2 Understanding Billing in Mobile Networks

The services offered on mobile networks can be considered to exist at two levels, i.e. the network transport level and the service level. From the users' point of view, these levels are inseparable. Users are interested in the actual service (e.g., multimedia content) that is delivered to their devices. The pricing schemes used for content charging are aimed at stimulating high levels of content service access, whereas network transport (bearer) pricing schemes are designed to meet diverse requirements. Since revenue generation in mobile networks is based on network resource usage, the network transport (bearer) service requires appropriate components for collecting and processing of charging information.

Different network bearer pricing schemes have been researched; however, only a few have been implemented on practical commercial systems. The applicable pricing scheme usually influences charging and billing on the system. For the sake of the reader, a discussion of some terms related to billing (e.g., pricing, accounting and charging) is given in appendix A.1 and more background information on billing in mobile and fixed networks is given in appendix A.2.

In practical implementations, the choice of a pricing scheme for a given network is related to the traffic management scheme used in the provision of communication services [3]. Flat-rate pricing, which is commonly used for the public Internet, generally co-exists with best-effort transport. Best-effort means that all applications using the network are treated equally at network nodes (e.g., routers and switches); hence they suffer equal delay and packet loss in the event of network congestion [16]. In networks that offer performance guarantee or QoS, usage-based pricing schemes are normally used to meet the requirements of users, network operators and content providers. Comprehensive information on user, operator and content provider requirements for pricing and charging schemes is given in appendix A.3.

## 2.3 Review of Pricing and Charging

### 2.3.1 Flat-rate Pricing

As mentioned above, flat-rate (or fixed) pricing is common with the Internet. In flat-rate pricing users pay a periodical subscription fee to gain access to the network and its services for the subscribed period (e.g., one month). In telecommunication networks this translates to making unlimited calls[1], sending unlimited short messages, or unlimited mobile Internet access especially for low bandwidth services (e.g., text browsing). Flat-rate charges are not based on the level of resource usage [13]; however, depending on the capacity of the communication link, the value of the subscription fee is usually different. A broadband connection offering 2 Mbps would be priced higher than a 56 Kbps modem link. When flat-rate pricing is used, systems for the accounting of resource usage are not required [1], hence it is simple to implement. The use of periodical subscription charges enables users to predict their communication costs, and service providers can predict their income [3]. It should be noted that flat-rate pricing has thrived in best effort service networks, which means that it lacks support for QoS provisioning.

QoS provisioning is required for the guarantee of user satisfaction, and it is needed by new and more demanding services, which enable network operators and service providers to generate more revenue [2]. Since flat-rate pricing lacks means for achieving congestion control, flat-rate priced networks that are under provisioned for the number of subscribers are likely to offer poor service [5]. There are times when users need to choose between getting services at high QoS levels or at a lower cost. For example heavy-utility users would prefer receiving network services at lower costs. This requirement is met by pricing schemes that provide logical channels that offer services at different QoS levels and prices.

### 2.3.2 Paris-Metro Pricing

Paris-metro pricing (PMP) is a proposal that provides options for users to trade network prices for quality and vice versa. In PMP, the Internet is partitioned into

[1] Fixed price charging is used for local telephone services in some countries. Refer to appendix A.2 for more information

logical sub-networks that are allocated different levels of non-sharable network resources [3, 9, 17]. Each logical sub-network operates on a best effort basis and is priced differently. Still flat-rate pricing is used for each sub-network; however, users would choose one of the logical sub-networks for the transmission of their traffic, and this implicitly defines the service level [1]. PMP aims at utilising users' sensitivity to network price levels in achieving better service in the sub-networks that are priced higher. It is expected that price sensitive users would choose sub-networks with lower prices. The more expensive sub-networks are expected to attract fewer users, hence they should be less congested.

PMP is expected to give flexibility to users since they can choose either the low-priced channels and incur less costs or the higher priced channels and receive better quality services. The simplicity of flat-rate charging is retained due the use of flat-rate pricing for all the sub-channels.

Using the evaluation criteria given in appendix A.4, the PMP scheme is found to have the following shortfalls:

- It fails to support individual user's QoS guarantees since each sub-network operates on a best-effort basis.
- Since the higher priced sub-networks are expected to attract fewer users, the overall utilisation of resources would be low; hence they would become over provisioned and less efficient. On the other hand, When more QoS-sensitive users switch to high-priced sub-networks, poor service will be experienced in those sub-networks due to over-loading.

Although PMP may facilitate the achievement of load balancing on the network, it does not use pricing to support congestion control.

### 2.3.3 Responsive Pricing

Pricing schemes that use the state of network congestion to set the applicable network prices would perform better in current communication networks. Responsive pricing belongs to this category of pricing schemes. The network prices are relayed to users as price signals. Responsive pricing assumes that users are rational with regard to price signals [18]. In the event of high network utilisation, network resources become scarce and congestion is bound to occur. When this happens, network prices are increased so that price sensitive users are forced to reduce their traffic. On the other hand, when there is low network utilisation, the prices are reduced enabling the community of adaptive users to increase their traffic.

Adaptive users who do not tolerate packet losses but are able to delay transmission are termed as elastic. On the other hand, inelastic users require strict delay guarantees, but can tolerate some degree of packet loss [1].

When elastic users transmit during off-peak periods, network utilisation is smoothed out; hence the overall network efficiency is improved. By responding to price signals, users can gain control over their transmission decisions.

Responsive pricing introduces increased signalling information between users and the network; hence more network resources would be used by the billing system. This is especially the case when congestion is determined along the end-to-end path of the source and destination nodes. Network congestion is known to be more pronounced at the network edges than in the core. Thus using the congestion levels at the edge of the network to set the prices would be reliable.

### 2.3.4 Edge Pricing

Edge pricing is based on approximations of network congestion at the edge of the network. The charges are based on the use of the network edge, rather than along the expected path of the source and destination of the communication session [19]. They are locally computed based on simple expected values of congestion. Approximations of the expected congestion are done using various methods (e.g. the time-of-day pricing concept [18]). Time-of-day pricing uses medium time frames, hence it enables edge pricing to support congestion control on the affected time frames.

In edge pricing, session data can be captured locally so that charging does not involve exchanging billing data with other networks. Although edge pricing minimises the measurement requirements for accounting and billing [5], it lacks visibility of routing via external networks; hence it doesn't consider the costs of the traffic to the affected networks. Edge pricing is not sufficient for multi-service platforms with diverse QoS requirements, since it does not permit users to prioritise traffic from some applications. Prioritising traffic from critical applications is necessary when using networks with limited resources and where QoS guarantees are offered.

### 2.3.5 Priority Pricing

Priority pricing [9, 20] is designed for networks that provide priority service classes, which are priced according to the priority received by traffic at various network nodes. The priority associated with different classes is implemented at network nodes that experience congestion. Packets belonging to high priority classes are charged at a higher price, and they receive preference at router queues. Users are expected to indicate the value of their traffic by choosing the appropriate priority level. The performance penalty received for using a less-than-optimal class is offset by the reduced cost of the service, whereas the monetary penalty incurred for using higher cost, higher quality service classes is offset by the improved performance. The priority pricing strategy is the only way of deterring users from choosing the highest priority class as default.

Priority pricing provides users with the flexibility needed in multi-service networks (e.g., multi-media applications can be given higher priority than file transfer applications). Like in Paris Metro pricing, priority pricing lacks means of preventing price-insensitive users from clogging the high priority classes. Some form of admission control is needed to limit the number of users who can simultaneously use the high priority and other classes.

Since priority pricing directly relates to traffic control and management, it can be used in creating QoS profiles that are priced differently. By using class of service-based QoS provisioning schemes like DiffServ, the QoS profiles can be priced differently to create more options for users.

## 2.4 Quality of Service in 3G and IP Networks

Over-provisioning of network resources both in the access and core network segments was suggested as a means of achieving QoS provisioning in communication networks. However, due to the development of new bandwidth intensive network applications, over-provisioning is not a long-term solution. This technical challenge led research into ways of providing QoS on multi-service networks (e.g. 3G networks).

3G networks (e.g. the UMTS) are designed to have more capacity, i.e. high bandwidth in the radio access network (RAN) and the core network. This would enable the end-users to access mobile Internet services at bandwidths ranging from a few Kbps to 2 Mbps [6]. In spite of the increased capacity, 3G networks require QoS provisioning to meet the requirements of the different mobile applications. Since 3G networks are shifting towards an all-IP platform, QoS provisioning schemes developed for IP networks can be adopted into 3G networks. For quick reference, information on QoS standardisations for 3G networks is available in appendix B.

### 2.4.1 Integrated Services

The integrated services (IntServ) architecture [21] is one of the QoS provisioning schemes that have been widely researched. In IntServ, end-to-end resource reservations are made for each connection establishment attempt. IntServ relies on the resource reservation protocol (RSVP) [22], which uses traffic control mechanisms (i.e. packet classification, admission control and packet scheduling). The packet classifier determines the QoS class (and perhaps the route) for each packet. For each outgoing interface, the packet scheduler achieves the promised QoS. During reservation setup, an RSVP QoS request is passed to the admission control and the policy control modules. The admission control module determines whether the node has sufficient available resources to supply the requested QoS. Policy control determines whether the user has administrative permission to make the reservation. If both checks succeed, parameters are set in the packet classifier and in the link layer interface (i.e., in the packet scheduler) to obtain the desired QoS.

RSVP mechanisms provide a general facility for creating and maintaining distributed reservation states across a mesh of multicast or uni-cast delivery paths. The use of end-to-end resource reservation by IntServ is the means that can truly provide resource assurance in a network; however, the need to maintain state information for each flow at every network node will result in scalability problems when applied to large networks [18, 9].

DiffServ [12] is an alternative architecture to IntServ. It has received wider acceptance than IntServ due to its simplicity and scalability to large networks.

### 2.4.2 Differentiated Services

The DiffServ architecture uses the concept of edge and core routers. The edge routers are configured to handle complex traffic control algorithms (e.g., packet classification and marking, traffic conditioning etc.), while the core routers would forward packets using per-hop-behaviours based on simple identification means. DiffServ achieves scalability by aggregating a traffic classification state that is conveyed by means of IP-layer packet marking using the DiffServ code point (DSCP). Network resources are allocated to traffic streams by service provisioning policies that govern how traffic is marked and conditioned upon entry to a DiffServ-capable network, and how that traffic is forwarded within the network [12]. Fig. 2.1 shows a logical view of a DiffServ packet classifier and traffic conditioner. This function is normally located in the edge routers of the DiffServ network.

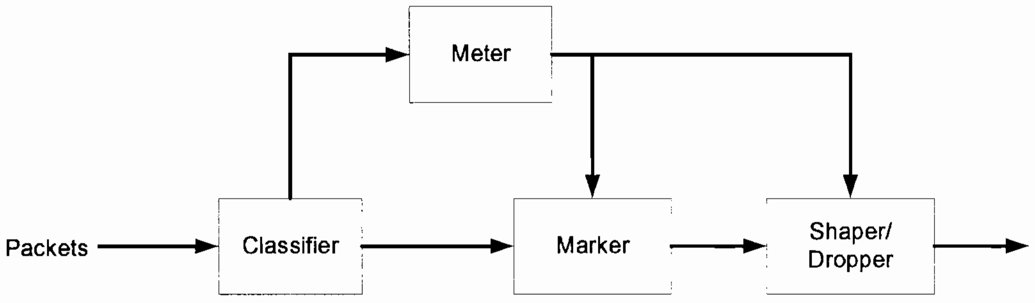


Figure 2.1: Logical view of a DiffServ packet classifier and traffic conditioner

In the DiffServ implementation, applications with similar network requirements, in terms of QoS attributes, are filtered into the same class of service (CoS). The

QoS received by each CoS is specified in service level agreements (SLA) between the users and the network operator. Packets belonging to the same CoS are given the same treatment in the core network, hence preferential treatment (e.g., traffic handling priority) can only be given to aggregate traffic in the same CoS. Without traffic differentiation at the flow level, when the traffic volume in a given CoS exceeds the permitted capacity all the flows in the CoS will experience poor QoS.

Controlling the amount of traffic that is offered to the network by a given class is an effective method of avoiding overloading of the CoS. The use of admission control or connection admission control (CAC) for supporting QoS is common in telecommunication networks (e.g., GSM and UMTS). CAC can be built upon DiffServ to support QoS provisioning in IP networks.

### 2.4.3 Admission Control

Admission control (AC) refers to a set of measures taken by the network to balance between the QoS requirements of new connections and the current network utilisation without affecting the grade of service of existing connections [23, 24]. It is used to control the network load by restricting the access to the network and hence improve the level of QoS guarantee. Admission control approaches can be categorised in a number of ways such as parameter-based approaches versus measurement-based approaches and edge/end-point admission control versus hop-by-hop admission control [18]. Parameter-based approaches assume some traffic pattern and try to maintain the aggregated resource consumption below the total capacity. Measurement-based admission control relies on the measurement of current network load and therefore responds faster to the network status and consequently improves the network utilisation.

As being done in the IntServ/RSVP architecture, admission control is traditionally performed on a hop-by-hop basis. Each intermediate network element along the path has to decide whether the new request can be accommodated or not and reserves resources accordingly. Since IntServ is un-scalable, admission control would be more beneficial if used with DiffServ. However, adding admission control functionality to all the core elements violates the DiffServ principle of leaving the core simple. End-point/edge admission control that pushes the admission control functionality to the edge of the network is more suitable in this case.

The CAC strategy is to limit the number of admitted connections. A new connection request will be rejected/dropped if the network lacks sufficient resources to support it.

By exploiting the scalability aspect of the DiffServ architecture, the features of CAC, and network pricing schemes that influence user behaviour, a network billing system that meets the QoS requirements of different applications and users can be developed. Tianshu et.al [18] investigate the integration of congestion-sensitive QoS-pricing schemes and admission control for multi-service networks. They use the time-of-day (ToD) pricing concept in setting prices for the access networks and a dynamic pricing strategy is used in the core network. The ToD prices are updated weekly or monthly. The access network providers are faced with dynamic prices in the core network. They set a price threshold for the use of the core network based on some rule (e.g., the service provider's profit must be greater than or equal to zero).

Although the scheme proposed in [18] promises a scalable network pricing scheme that provides price predictability to the users, it takes great freedom away from the users since the access network provider is the one who chooses the traffic class that optimises his profits. The class chosen might not meet the requirements of all users.

Current communication network management models give users limited choice in the determination of the tariff rates for the services they use. The subscribers are mainly given two tariff periods, i.e. the peak (daytime) period and the off-peak (night) period. It has been proved that some users are willing to use specific services at tariff rates that are higher than the normal rates [8]. This is associated to the importance attached to the services by the users. In a network that offers different levels of QoS, users would want to use important services at a guaranteed high level of QoS (guaranteed delivery even during periods of congestion). In order to facilitate this, networks must provide interfaces through which users can indicate their willingness to use specific services at particular tariff rates and QoS levels.

# Chapter 3

# Design Considerations

## 3.1 The Motivation for Dynamic QoS-Based Billing in 3G Networks

The need for QoS-aware billing schemes was highlighted in Chapter 1, and a review of literature on QoS and billing was given in Chapter 2. This chapter presents the technical details of the charging scheme proposed in this thesis, i.e. the Dynamic QoS-Based Charging Model (DQBCM). Charging schemes for communication networks differ in the way the pricing or rating of network resources is done. In networks where pricing is used to achieve congestion control, the prices are normally raised to counter increasing congestion levels. Congestion dependent pricing schemes have been designed using different approaches. In the current operation of communication networks the congestion is estimated over large time frames such as periods of the day (e.g. the time-of-day pricing); however, it is important that congestion estimates be done on short time frames to reflect the actual demand for network resources. This would truly support congestion dependent pricing [9].

Internet access can be achieved through the use of different network access technologies (e.g. GPRS, WLAN, WiMAX, ISDN, ADSL, etc). Each of these technologies supports different data rates. In charging for the Internet access, the applicable access fee is commonly pegged on the achieved data rates and QoS values. This practice is used in mobile data communications, whereby the tariff rate applied to a given packet data protocol (PDP) context depends on the authorised QoS parameters, with higher tariff rates being used to guarantee the required QoS [25]. 3G networks are designed to offer more resources both in the core network and the radio access networks. The introduction of new mobile services is expected to load the network, hence the projected data levels will utilise the capacity of the networks. However, if the current tariff structures are retained, users of mobile communication services will not get the right incentives to utilise the full capacity of the networks.

The demand for network resources keeps on varying, leading to periods when the network experiences an abundance of resources, e.g., bandwidth. During such periods users of applications that require extra bandwidth could have their communication sessions modified so as to authorise more resources for them. In such a scenario, current tariff structures would lead to charges that are several times greater than current average values.

Mobile networks have become multi-service networks that offer access to a variety of applications (e.g. multimedia content and other value added services). The network transport requirements for these applications range from loose bandwidth requirements to tight delay tolerances. The requirements of the users of the services also range from the expectation of a guaranteed high quality network service (that often comes at a higher cost) to the preference of lower prices for the use of the network.

The success of a new pricing scheme greatly depends on user acceptance. Users prefer pricing schemes that use strategies they are familiar with. This explains why network operators often adopt legacy pricing schemes for new services. An online survey conducted in this research revealed interesting views among users of mobile services. The online survey was open to all users and the aim was to collect views regarding the justification of the tariff rates used for mobile Internet services. The other objective was to gather suggestions from users on which pricing schemes should be used for particular services. Three pricing schemes were presented to the users, i.e. flat-rate, usage-based and pricing schemes that relied on other factors (e.g. QoS and how important users regard certain services)[1].

The outcome of the online survey is given in table 3.1. Mixed views were obtained

[1] Further details on these pricing schemes are given in appendix C.7.

Table 3.1: Online Survey on User Views Regarding Pricing

| ***Description*** | ***Result*** |
|---|---|
| Number of Participants | 27 |
| Mobile users | 27 |
| Mobile Internet users | 16 |
| Justified tariffs | 2 |
| *Support usage-based charging for:* | |
| Web Traffic | 12 |
| FTP and SMTP | 16 |
| Streaming Video and Audio | 11 |
| Video Phone | 11 |
| *Support Flat-rate charging for:* | |
| Web Traffic | 14 |
| FTP and SMTP | 10 |
| Streaming Video and Audio | 8 |
| Video Phone | 5 |
| *Support Special tariff charging for:* | |
| Web Traffic | 0 |
| FTP and SMTP | 3 |
| Streaming Video and Audio | 9 |
| Video Phone | 11 |

Table 3.2: User willingness to pay for different services

| ***Service / User rating*** | ***Low*** | ***Standard*** | ***Moderate*** | ***High*** |
|---|---|---|---|---|
| Web traffic | 12 | 5 | 5 | 5 |
| FTP and E-mail | 10 | 8 | 4 | 5 |
| Streaming Audio | 6 | 12 | 6 | 3 |
| Streaming Video | 4 | 10 | 6 | 7 |
| Video Phone | 4 | 5 | 13 | 5 |

regarding how much users were willing to pay for different services. The users were required to rank their willingness to pay for the services in the following categories: low, standard, moderate and high (refer to appendix C.7). The different user views are given in table 3.2.

Statistics from the online survey indicated that most users of regular mobile services (e.g. voice telephony) also use mobile Internet services. Most users do not justify the tariff rates used for mobile services, i.e. they believe the rates are very high. The support for usage-based charging schemes for mobile Internet services is relatively higher than that of flat-rate charging and users believe that multimedia services (e.g. video phone) should be charged using special tariff schemes.

Regarding user willingness to pay for different mobile services, most users prefer low and standard tariffs for all services. Some users indicated willingness to pay moderate and high rates for specific services, especially multimedia services such as video phone and streaming video.

When the tariff rates for some services are not constant, users may hesitate to take up those services due to the uncertainty in the final cost. In order to encourage the uptake of the services, appropriate interfaces that will enable the users to interact with the network in real-time, so as to indicate their willingness to use the services when the tariff rates are raised, are required. This will not only allow users to select options that give them access to services at an affordable cost, but also it will enable them to optimise on the QoS levels. The DQBCM is designed to support a multi-service network where tariff rates are changed according to the current demand for resources.

The DQBCM varies tariff rates according to the changing state of network congestion. A QoS-aware billing scheme is needed in any network that provides service level agreements (SLA), which are an inherent part of the DiffServ architecture. UMTS networks are multi-service networks that offer services that need some level of resource guarantee. In order to encourage the usage of the full capacity available on these networks, pricing can be used as an incentive. The DQBCM uses a pricing algorithm that not only meets the requirements of congestion control, but also gives incentives to encourage the usage of the network during periods of low resource utilisation. By using network profile management agents and interfaces that allow users to modify their service usage profiles, the DQBCM facilitates network management conditions that are needed to meet the expectations of the users.

## 3.2 DQBCM Classes of Service and Profiles

Three CoS are used in the design of the DQBCM scheme. The conversational and the streaming multimedia CoS of the UMTS are combined to form a single multimedia CoS. The other CoS are the interactive and the background CoS. In implementing the DiffServ architecture, services in the multimedia CoS, which requires minimum delay and jitter, receive the expedited forwarding (EF) PHB. Services belonging to the interactive and background CoS do not tolerate packet losses; hence they receive the Assured Forwarding (AF) PHB. The available network resources are allocated to each CoS according to observed usage patterns, i.e. a CoS with popular services, in terms of network resource usage and revenue generation would receive a greater percentage of the network resources. The allocations define the amount of resources (e.g. bandwidth and buffer space) that will be available to traffic in that CoS when the network capacity is fully utilised. During periods when the network has an abundance of resources, any CoS may "borrow" the excess resources, if they are required by its applications.

When resources belonging to a CoS are borrowed by another CoS, service request blocking will occur if the lending CoS attempts to admit a new connection when the network capacity is fully utilised.

In a network system where the price for services changes dynamically, it is important to receive indication from users about their willingness to use or to continue using the services under the new conditions. The rate at which network service conditions (e.g. prices for resources) change would be very high; hence expedited response from the users will certainly not be achieved. To overcome this problem, and to achieve the service management flexibility that this research investigates, the DQBCM introduces network service profiles. Three profiles, i.e. platinum, gold and silver are introduced as an overlay onto the QoS classes. By selecting a service usage profile, the users indicate their willingness to use the network under the conditions specific to the profile. Information about the profiles selected by the users will be available in a database that is accessed by the profile management agent, which would relay it to other network entities when needed. Figure 3.1 illustrates the role of service profiles in the DQBCM system.

The classes of service discussed in this research are designed to offer services at different prices and QoS levels, which the users may select depending on different conditions (e.g. the service being requested and their sensitivity to varying network prices). Unlike the interactive and the background CoS, the multimedia CoS will only accommodate the platinum and the gold profiles. The reason for this is that best effort transport, which is offered by the silver profile is not appropriate for the multimedia CoS. 3G mobile devices have enhanced features, such as buffering capabilities that can be used by some applications in the streaming multimedia CoS to compensate for the effects of network delay and jitter. This property is considered in the design of the gold profile for streaming multimedia services in the DQBCM scheme.

The platinum profile targets users whose main requirement is predictable high levels of QoS, even during periods when the network is congested, but they are insensitive to changes in the price of network resources, the gold profile is intended for users who are very sensitive to fluctuations in the prices of network services, but they are using services that can tolerate substantial changes in the level of QoS, the silver profile is designed for users whose main intention is to stay online, especially for long durations.

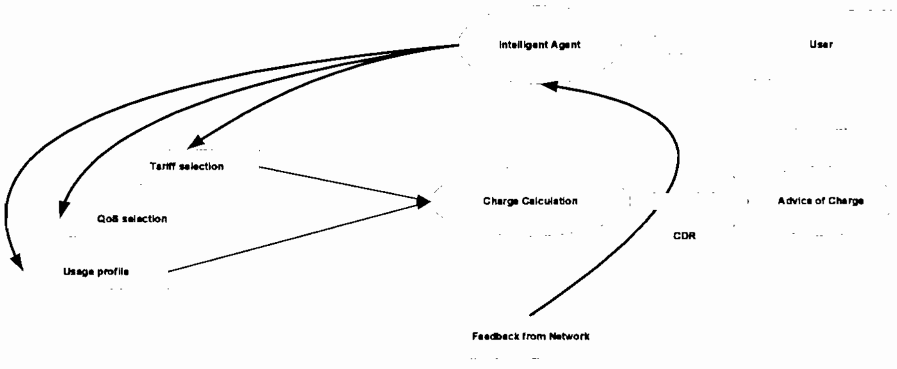


Figure 3.1: The role of service profiles in the DQBCM system

### 3.2.1 Platinum Profile

The platinum profile offers the highest level of QoS in a given CoS. The characteristics of the platinum profile are as follows:

- Guaranteed high QoS level, even during periods of network congestion

- Prices for network resources vary with the demand for network resources

The QoS control algorithm for the platinum profile ensures that even during periods when network congestion is high the active flows will receive the same level of QoS as during periods when there is an abundance of network resources. This profile is best suited for real-time multimedia and other applications that require guaranteed network performance conditions. Since the tariff rates for charging services in the platinum profile will vary with the changing network congestion state, users of this profile are considered to be less sensitive to variations in network prices.

The platinum profile considers that in any competitive commercial system, there would be some users that are willing to pay extra so as to improve their benefit. This profile targets such users; hence by allocating enough resources to the platinum profile, the network would enable the operators to maximise their income [8].

The network resource pricing algorithm for this profile works as follows: the admission control function (refer to section 3.3) is configured to admit new PDP sessions until the number of active sessions becomes equal to $N_p$. $N_p$ represents the allowed maximum number of active sessions in the platinum profile.

Pricing of resources in the platinum profile relies on feedback from the admission control function. The tariff rates to be used for a given period are affected by the demand of network resources, which is represented by the number of active sessions ($I_p$). In a practical implementation (e.g. the UMTS) the amount of network resources allocated to different PDP sessions would be different. For this reason, the value of $I_p$ is a hypothetical figure that could represent the actual number of connections that are being served, or the relative amount of network resources that have been assigned to the active PDP sessions. Similarly, $N_p$ could represent the maximum number of connections that can be served by the platinum profile or the maximum amount of network resources that can be allocated to PDP sessions in the platinum profile. Equation 3.1 presents the formula for varying the price of network resources in this profile.

$$T_p = \alpha(I_p)^n \tag{3.1}$$

$\alpha = \frac{\Upsilon}{N_p}$ is a pricing constant for the platinum profile, and it relates to value of $N_p$ and factors affecting revenue targets for the network (e.g. CAPEX and OPEX). $n$ specifies how fast the tariff rate increases.

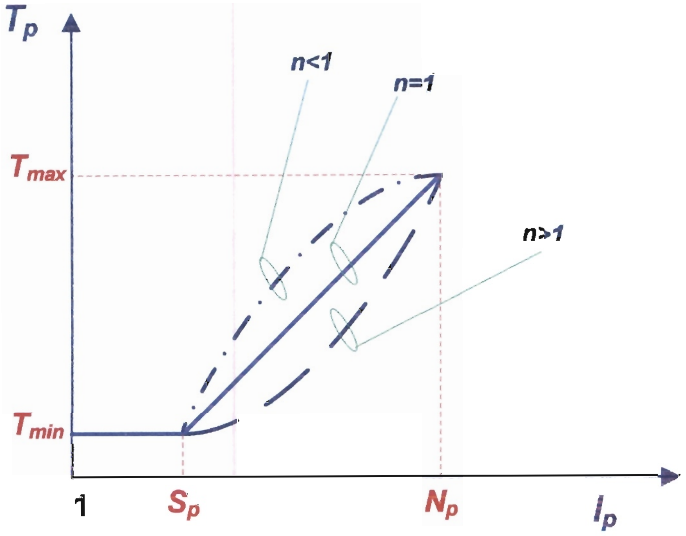


Figure 3.2: Platinum profile tariff rates

The prices for resources in the platinum profile will also be affected by the number of active sessions ($I$) that have been admitted to the affected CoS, and ideally it would depend on the number of active sessions on the network. Since the platinum profile offers higher QoS than the other profiles, the price for using this profile would be relatively higher than the other profiles. There is a level of congestion below which the prices for network resources in the platinum profile would be equal to those of the gold profile. This considers the fact that when the level of congestion in the network is low, traffic from the platinum profile would not get preferential treatment over traffic in the gold profile. This fulfils the requirements of demand based pricing [9]. Figure 3.2 gives a graphical representation of the variation of platinum profile tariffs with the value of $I_p$.

Thus, the differential pricing becomes effective when the value of $I_p$ equals $S_p$. At this point it is assumed that the active flows in the platinum profile are receiving preferential treatment on the network. This implies that tariff rates for the platinum profile will be at the minimum ($T_{min}$) for $I_p \leq S_p$, and at the maximum ($T_{max}$) for $I_p = N_p$. Thus, Eq. 3.1 becomes:

$$T_p = T_{min} + \alpha(I_p - S_p)^n \tag{3.2}$$

which is applicable for $S_p \leq I_p \leq N_P$, where $\alpha$ can be determined by considering the maximum tariff level for the platinum profile as given in Eq. 3.3.

$$T_{max} = T_{min} + \alpha(N_p - S_p)^n \tag{3.3}$$

Hence $\alpha = \frac{T_{max} - T_{min}}{(N_p - S_p)^n}$. After substituting $\alpha$, the formula for charging communication sessions in the platinum profile is as given by Eq. 3.4. It is applicable for all values of $I_p \in [S_p; N_p]$. $T_p = T_{min}$ for $I_p \leq S_p$ and $T_p = T_{max}$ for $I_p = N_p$.

$$T_p = T_{min} + \frac{T_{max} - T_{min}}{(N_p - S_p)^n}(I_P - S_p)^n \tag{3.4}$$

The platinum profile has great advantages to the network, for example it would enable the network operators to maximise profits when the number of subscribers using this profile is high. It is expected that with the popularity of multimedia applications, most of the users of real-time multimedia applications will prefer to use the platinum profile due to the guaranteed QoS. The guaranteed QoS is simply expressed as an equal partition of the total resource pool ($R_p$) distributed among the defined maximum admissible platinum profile sessions ($N_p$), i.e. $Q_p = \frac{R_p}{N_p}$, and it is constant. By lowering network prices during periods of low congestion, the platinum profile would give incentives for users to use the network, hence improving network efficiency.

### 3.2.2 Gold Profile

This profile is characterised by constant prices for a unit measure of a specific service. This simply means that if the cost of transmitting $W$ megabytes of data is $p$ cents, the price would remain constant, i.e. $p$ cents, regardless of the level of network congestion. This profile provides a trade-off between constant predictable network prices and the QoS level that is attainable during periods of network congestion. Users of the gold profile are considered to be sensitive to variations in the price of network services. On the other hand, they are tolerant to changes in the level of QoS.

The characteristics of this profile are:

- Constant price for network services - this is regardless of the level of network congestion.
- Network congestion affects the QoS level. This implies that the gold profile offers a limited level of QoS to its flows. When congestion sets in, the QoS parameters of each flow would be downgraded. The degree of congestion in the network is abstracted using the number of active flows ($I_g$) in the profile, hence as the value of $I_g$ approaches $N_g$, the QoS level will degrade.
- The admission control strategy (refer to section 3.3) limits the maximum number of sessions that can be admitted to this profile to $N_g$.

The QoS received by active flows in the gold profile is inversely proportional to the value of $I_g$, and it would be at the minimum ($Q_{min}$) when $I_g = N_g$. The instantaneous QoS value for the active flows can be written as:

$$Q_g = Q_{max} - \beta.(I_g - 1)^m \tag{3.5}$$

where $Q_{max} = \frac{R_g}{1}$, i.e. the whole gold resource pool is used by one session and $m$ defines how rapid the QoS level degrades with congestion. The minimum resource allocation per flow is represented by Eq. 3.6, hence the constant $\beta$ is given in Eq. 3.7.

$$Q_{min} = R_g - \beta.(N_g - 1)^m = \frac{R_g}{N_g} \tag{3.6}$$

$$\beta = \frac{N_g.R_g - R_g}{N_g(N_g - 1)^m} = \frac{R_g}{N_g(N_g - 1)^{m-1}} \tag{3.7}$$

Substituting $\beta$ into Eq. 3.5, the instantaneous QoS formula for each gold session becomes:

$$Q_g = R_g[1 - \frac{1}{N_g(N_g - 1)^{m-1}}.(I_g - 1)^m].k_Q \tag{3.8}$$

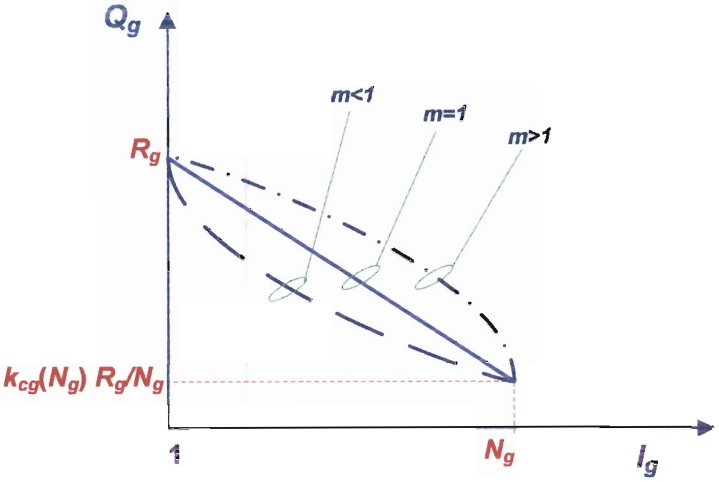


Figure 3.3: Gold profile QoS variation

Here we introduce the correction coefficient $k_Q$, as a decreasing function of the number of active sessions $I_g$, used to prevent the occurrence of congestion. This simply means that the QoS allocations for the gold profile sessions will decrease with an increase in the number of active sessions. Figure 3.3 illustrates the variation in the gold profile QoS. The optimal constant tariff is found by assuming an equal price for a resource unit for both the gold and platinum profiles; $\frac{T_g}{Q_{gmin}} = \frac{T_{pmax}}{Q_p}$ *money/resource*, where $Q_{gmin}$ is the minimum tariff for the gold profile flows and $T_{Pmax}$ is the maximum tariff for the platinum profile flows. Substituting $Q_{gmin} = \frac{R_g}{N_g}.\kappa(N_g)$ and $Q_p = \frac{R_p}{N_p}$, the tariff rate for the gold profile would be:

$$T_g = T_{max}\frac{R_g}{R_p}\frac{N_P}{N_g}\kappa(N_g) \tag{3.9}$$

An example of services that are suitable for the gold profile are high priority background data transfers, essential web-browsing applications like e-commerce that involve database searches etc. The cost of using services in this profile would either be on a volume of traffic or on duration of use basis. The gold profile retains the current trends in QoS provisioning and charging of mobile Internet services, i.e., when the level of congestion in the network increases, the QoS achievable in this profile degrades up to a pre-set minimum level [26]. The admission control scheme ensures that the minimum QoS does not fall below the set threshold. For the multimedia CoS, the gold profile would be appropriate for streaming applications, since buffering of data can be used to compensate for network effects like delay and jitter, hence the degradation in QoS during periods of congestion can be tolerated.

### 3.2.3 Silver Profile

The silver profile is designed to cater for price-sensitive users, whose basic aim is to stay connected to the network for long periods. It uses flat-rate pricing and provides network services using best-effort transport, hence it does not offer QoS guarantee to its users. Connection admission control is used to limit the number of active sessions to a maximum value $N_s$. This ensures that the active sessions do not experience total service failure due to excessive congestion in the network. This profile essentially accommodates the current trend in Internet communications, where best-effort transport offers no QoS guarantee to the users [16].

The use of flat-rate pricing in the silver profile will attract high service usage levels, which means that the network utilisation efficiency would be high. Lack of QoS guarantee for this profile makes it suitable for applications that tolerate poor network conditions (e.g. bandwidth fluctuations, low levels of network bandwidth and large delays).

### 3.2.4 Resource Sharing between Profiles

The design architecture of the DQBCM permits resource sharing between different profiles. The platinum and gold profiles are used to illustrate the resource sharing algorithm; the profiles are considered to share a common pool of resources equal to $R$. The idea of resource sharing allows the maximisation of network (bandwidth) utilisation, by admitting extra sessions in the platinum profile or enhancing QoS of currently active sessions in the gold profile. Resource sharing prevents the wastage of resources that occurs when static resource allocations to traffic classes are made [9].

Resource sharing is achieved by relocating unused network resources from one profile to another. If the platinum profile is the recipient of the relocated resources, the effective value of $N_p$, i.e. $N_p'$ will be set to a higher figure. This would lead to a condition where $N_p < I_p$ and $N_p \leq N'_p$ , which means that the platinum profile would admit extra communication sessions. When this occurs, the network operators and independent service providers will be able to serve more users and maximise on revenue generation. On the other hand, when a low priority profile receives the relocated resources, the QoS of its active flows will be boosted and improved network performance will enhance user satisfaction.

The relocation of resources from one profile to another would be done according to the following algorithm:

- When the platinum (higher priority) profile is the recipient, the low priority profile (e.g. gold) must be experiencing an over-provision of network resources, i.e. the flows in the gold profile should be achieving close to the maximum QoS level. Thus, the relocation of resources would not have an adverse effect on its QoS performance.
- When the platinum (higher priority) profile is the donor, the number of active sessions in the platinum profile should be less than the maximum admissible value, i.e. $I_p < N_p$. Relocation of resources to the gold or silver profiles does not permit them to admit extra sessions, instead the relocated resources are used to improve the QoS level of the active flows. If a session establishment request is received by the platinum profile, the relocated resources will be retrieved to enable the admission of the session.

It is assumed that the users of the platinum profile have sharing priority over the gold profile adherents. However, the expansion of the platinum profile resources is done as long it does not affect the minimum required QoS for the gold profile flows. For the platinum profile we always have fixed level of the QoS ($Q_p$) equal to $\frac{R_P}{N_p}$. Substituting $N_p = N - N_g$, and $R_P = R - R_{gmax}$; i.e. $R_p = R - Q_{min}.N_g$, the resource allocations for each flow in the platinum profile becomes

$$Q_p = \frac{R - Q_{min}N_g}{N - N_g} \tag{3.10}$$

where $R$ is the total resource amount; $Q_{min}$ is the minimum required resource allocation per gold profile session; $N_g$ is the maximum number of the sessions supported by the gold profile and $N$ is the total supported sessions number for the whole resource pool. All these parameters are specified as input values for the pricing algorithm. The following inequalities must be true: $N_g < N$, $Q_{min} < Q_P$. The tariff rate for charging sessions in the platinum profile would become

$$T_p = T_{min} + \frac{T_{max} - T_{min}}{(N - N_g - S_p)^n}.(I_p - S_p)^n \tag{3.11}$$

A graphical representation of the platinum profile tariffs with resource sharing is given in Fig. 3.4. There are also several restrictions for admitting $\Delta I_p$ extra sessions in the platinum profile over the guaranteed number of sessions $(N - N_g)$. If $I_p + \Delta I_p > N - N_g$, then $I_p + \Delta I_p \leq \frac{R - I_g.Q_{min}}{Q_p}$. Substituting in Eq. 3.10, we obtain

$$\Delta I_p \leq \frac{(R - I_g.Q_{min}).(N - N_g)}{R - N_g.Q_{min}} - I_p \tag{3.12}$$

i.e, extra sessions can be admitted in the platinum profile if the QoS level of the currently active sessions in the gold profile is satisfactory.

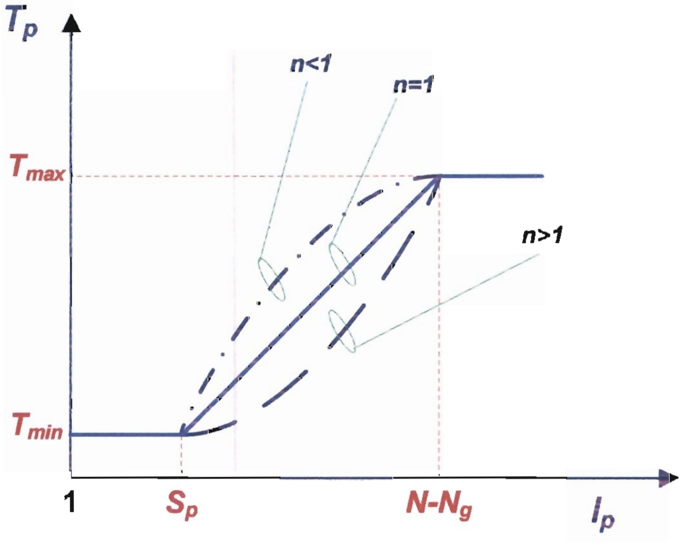


Figure 3.4: Platinum profile tariffs - with resource sharing considerations

Once the resources of the platinum service profile have been appropriately allocated to the active sessions, the remaining part of the bandwidth pool can be shared between the active flows of the gold profile. The QoS achieved by each flow can be represented as:

$$Q_g = \frac{R - R_p}{I_g} \tag{3.13}$$

substituting for $R_p$ it becomes:

$$Q_g = \frac{R - I_p Q_p}{I_g} \tag{3.14}$$

after further substitutions we get:

$$Q_g = \frac{R(N - N_g) + I_p(Q_{min} N_g - R)}{I_g(N - N_g)} \tag{3.15}$$

which is applicable for $1 \leq I_g \leq N_g$. If the constraint coefficient $k_Q \leq 1$ is introduced for controlling congestion, then:

$$Q_g = \frac{R(N - N_g) + I_p(Q_{min}.N_g - R)}{I_g.(N - N_g)}.k_Q \tag{3.16}$$

In the simplified interpretation, the congestion phenomena is abstracted as the increasing number of active sessions $I_p + I_g$. $k_Q$ can be determined experimentally.

The idea is to construct a table containing the arbitrary values of the sum $(I_p + I_g)$ and corresponding values of the two dimensional packet loss probability function $L[k_Q(I_p + I_g)]$, $k_Q \geq 1$, which can be collected from a series of measurements performed on the identical pattern of transmitted data during a session. After that, $k_Q$ is selected, which results in the minimal $L$ for each $(I_p + I_g)$. The graph in Fig. 3.5 illustrates the available QoS for the gold profile in respect to the number of sessions $I_p$ and $I_g$.

From Eq. 3.9, the tariff rate for the gold profile with resource sharing considerations becomes,

$$T_g = T_{max} \frac{Q_{min}(N - N_g)\kappa(N_g)}{R - Q_{min} N_g} \tag{3.17}$$

## 3.3 Admission Control Policy

The DQBCM admission control policy defines the algorithm for managing the admission of new sessions to the network. Admission control is performed at the

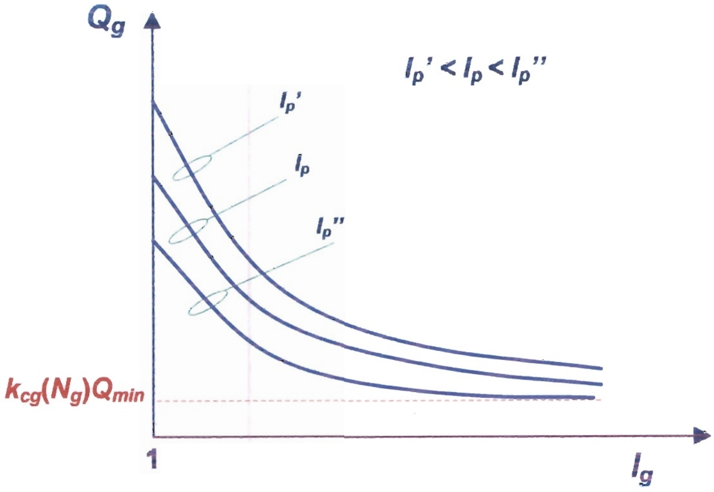


Figure 3.5: Gold profile QoS - with resource sharing

service profile level to ensure that the maximum admissible number of sessions ($N$) is not exceeded. The admission control function (ACF) is in charge of all session admission operations, i.e. monitoring the number of active sessions ($I$) in each profile and reporting to various entities that require this information.

The ACF mainly deals with new connection attempts; however, when users change their service usage profiles in the course of service delivery, it ensures that the request is handled without causing unexpected effect to the QoS received by flows in the affected profiles. This involves verifying that the condition $I \leq N$ is satisfied for the requested profile, before and after the profile change.

To meet the QoS specifications of different profiles, admission control will be done for each profile, i.e. platinum, gold and silver in every CoS. By defining the maximum number of communication sessions each profile should handle, the CoS will in effect achieve the desired QoS levels. This would translate to the whole system when the effect each CoS has on the others is taken into account. The admission control strategy guarantees a minimum level of QoS performance for active flows in the gold and silver profiles; flows in the platinum profile are guaranteed of steady high QoS at all times.

In the admission control strategy, when $I = N$, new session establishment attempts will be rejected. New session activation requests would be allowed after existing sessions terminate or get deactivated. This requires monitoring of the activity of all active sessions and defining conditions when an active session would be considered inactive; hence the admission control and traffic management function would close it to free network resources. The freed resources would be used to admit a new session request in the profile, or another profile would be allocated the resources so as to boost the QoS for its flows.

In the UMTS network, the DQBCM admission control function would be located in the GGSN. The DQBCM ACF would co-ordinate with the policy decision function (PDF) and the application function (AF) to ensure the authorisation of packet data protocol (PDP) context requests is done according to availability of network resources. When a PDP context activation request is received by the ACF it is relayed to the UMTS PDF. The PDF will process the request and authorise the allocation of the resources requested for the PDP context.

The resources for the PDP context would be authorised when the profile it falls in has capacity to service the request. To achieve this, the PDF and the ACF would keep track of the number ($I$) of admitted active sessions in each profile and all CoS and the sum of the authorised PDP QoS parameter values for all sessions in the profiles and CoS . The total amount of network resources allocated in a profile will be determined from the sum of the PDP context parameter values of its sessions. If all sessions in a profile were assumed to be allocated an equal amount of resources, the product of $I$ and the values of the QoS parameters of one session would give the total amount of the resources allocated in that profile.

By using the number of active PDP context sessions to represent the demand for network resources and the congestion level in the network, the admission control policy facilitates the achievement of congestion control in the network. With information on resources allocation for all profiles and CoS, the PDF will indicate the possibility of a relocating extra (unused) resources from one profile to another.

## 3.4 Network Agents and User Profile Management

### 3.4.1 User Service Profile Management

Users always want to know the cost of the services prior to actual usage. They also need to be informed of changes in pricing that may occur during the course of service delivery. If this were done, the network would expect fast response from the users on the action to be taken when a dispute arises. However, the user's action would not be quick enough to provide an expedited response to the network management system. Thus, the user profile management system is designed to bridge between the users and the network management system.

As mentioned earlier, users can select different profiles at the beginning of a communication session. Session establishment attempts would only succeed if the number of active sessions in the selected profile hasn't exceeded the value set by the network operator. The user profiles would be stored in a database; on mobile networks the Home Location Register (HLR) would be used. Permanent entries in the database will include a unique user identity, while the temporary entries will include a temporarily assigned network access identifier and the service usage profiles for each CoS. A default service usage profile would be assigned to users when they are registered on the network for the first time. When the user accesses the network for the first time, the default profile would be considered. After profile changes are made, the last profile selected by the user will be considered as the default profile. When users change their service profiles the database would be updated accordingly.

### 3.4.2 Network Agents

The complexity associated with a dynamic billing (e.g. informing users whenever a price change occurs) requires means that shield the user from involvement in decisions that require fast action. Network agents incorporated in the DQBCM system absorb the complexity and present the users with friendly interfaces to use in specifying the intended service usage. As a bridging entity between users and network management functions, i.e. the pricing, admission control and traffic control functions, the network agents get user input and translate it into rule sets for use by the different network management functions.

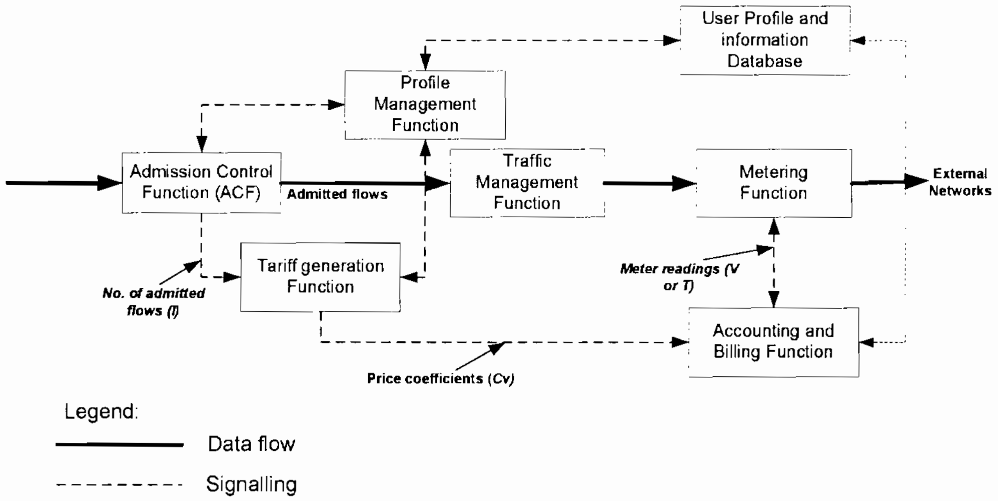

User Profile and information Database
Profile Management Function
Admission Control Function (ACF)
Admitted flows
Traffic Management Function
Metering Function
External Networks
Tariff generation Function
No. of admitted flows (I)
Meter readings (V or T)
Accounting and Billing Function
Price coefficients (Cv)
Legend:
Data flow
Signalling

Figure 3.6: Networks agents and user profile management

Service usage profiles enable users to provide the information required by network agents. For example, network agents would initialise the traffic control state of the traffic control and management system using information stored in the user profile database. The initialisation would enable the traffic control and management function to perform classification of traffic into the right profiles, and the admission control function would accurately perform connection admission control on session establishment requests made by users. When users change their service usage profiles, the network agents will update the rule sets for the network management functions. Fig. 3.6 illustrates the flow of information between users, network agents and network management functions. Interaction between different functional blocks of the system will also be facilitated by the network agents.

## 3.5 Pricing and Charging

The platinum and gold profiles require tariffs for use in charging of the network services. The formulae for the tariff rates for these profiles were discussed earlier. Each CoS of the network has a characteristic pricing constant associated with it. CoS that require higher QoS parameters (e.g. queue priority) would use a higher valued pricing constant. In the DQBCM model, the multimedia CoS is configured to receive the highest queue precedence, thus its flows would get high queue priority to ensure low packet delay and elimination of jitter. It would only make sense to charge the multimedia CoS at a comparatively higher tariff than the interactive and background CoS.

The enforcement of the tariff rates would have the platinum profile in the multimedia CoS charged higher than the platinum profiles in the interactive and background CoS; the same principle applies to the gold profile. The interactive CoS would be configured to offer a higher QoS performance than the background CoS. This is in line with the QoS requirements of the two CoS, as given in appendix B.2.1.

Charging for the use of resources in the platinum and the gold profiles requires metering and accounting procedures. The volume of traffic (e.g., bytes) can be used as the accounting metric, hence individual users would be charged according to the level of resource usage at the tariff rates that apply to specific profiles and the actual session when the services were delivered.

## 3.6 DQBCM QoS Provisioning Mechanism

The advantages of the DiffServ architecture for QoS provisioning in IP networks, together with the proposal by the 3GPP of having an all-IP mobile network were considered in the design of the DQBCM charging scheme. With an all-IP UMTS network, major routing nodes in the network can be used to enforce the traffic control functions of the DiffServ architecture. According to Fig. 3.7, the SGSN and the GGSN can be used in implementing some of the functions of the DQBCM scheme. The pricing strategy of the DQBCM is designed to support the provision of QoS in the UMTS Core Network and the RAN, and for inter-working with QoS-enabled external IP networks.

The GGSN is the appropriate DiffServ boundary node for traffic going to the external IP networks. It supports traffic classification, marking, conditioning and other important functions of the DQBCM scheme (e.g. policy-based admission control). The policy enforcement point (PEP) of the UMTS network is located within the GGSN [27, 28]; this will enable the GGSN to perform the enforcement

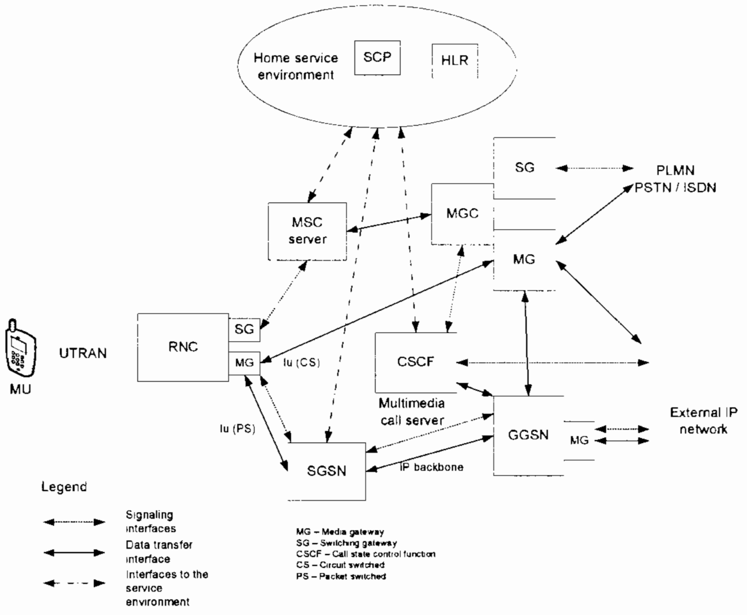


Figure 3.7: A simplified release 2000 all-IP UMTS architecture (Adopted from [11]).

of resource authorisations of the DQBCM scheme. In an all-IP UMTS network, the SGSN can be used as a DiffServ boundary node. The packet classification, policing, marking and conditioning would thus be done here. With this approach, the GGSN is left to perform other functions of the DQBCM scheme (e.g. user profile management, tariff determination and pricing, and admission control).

## 3.7 Technical Considerations of the Architecture

The challenges facing the practical implementation of the DQBCM can be classified as user involvement and system integration. The system integration challenges include support for current charging practices (e.g. prepaid billing). User involvement includes profile modification and the notification of network price changes.

### 3.7.1 Support for Prepaid Services

Prepaid billing is facilitated by online charging, hence charging information affects the real-time delivery of the service. For example when the user's credit balance gets exhausted, service delivery would be terminated. Credit control mechanisms (e.g. Diameter credit control [29, 30]) are used to facilitate the online charging process. In the DQBCM design, credit control measures like the Diameter application would be used to facilitate online charging and support prepaid billing. For the platinum profile the credit reservation would be based on the maximum possible charge that may be incurred for transporting the application over a given period (e.g., one minute). The maximum charge would be incurred when $I_p \geq N_p$. When content is purchased from independent service providers, the credit reservation would include the additional charge for accessing an application after credit reservation for the cost of the content has been done.

### 3.7.2 User Involvement

Mobile units or user equipments (UE) are becoming advanced in terms of memory, processor speed, screen resolution etc. Software for interpreting network management decisions can be stored on the UE, hence reducing the use of network resources in the transmission of control information between network entities and the UE. The software would have definitions for control and error codes that are exchanged between network entities and the UE. For example, when a user's request for services in a given profile cannot be granted, the system would send an error code to the UE. The software will interpret the error code and display a message that is understood by the user.

The UE should provide the users with appropriate interfaces for profile modification. Menu driven or web-based interfaces can be used for this purpose. As a way of encouraging users to optimise on the benefits of different profiles, profile modification should not incur major charges. Advanced optimisation of the benefits of different service profiles can be achieved by providing users with information on the congestion state of different profiles and the varying tariff patterns. Users of the platinum profile would find information on the tariff patterns to be useful in cases where cost minimisation is necessary.

## 3.8 Limitations of the Testbed Implementation

This section emphasises the technical differences between the network targeted by the DQBCM design, i.e. 3G network and the testbed network. The testbed design and implementation was done on a WLAN and a LAN; hence the network access technology (CSMA/CA and CSMA/CD) [31] is different from that used in 3G networks (e.g. WCDMA) [32]. The other technological differences include radio resource management schemes, the number of mobile nodes, the available network resources, network security considerations and the different type of services offered. The implementation of the DQBCM system must address some complex issues; however, the testbed only investigates some of the architectural features. Recommendations are given on how other design elements of the DQBCM system may be approached. However, the design is a proposal open to further investigation and development.

### 3.8.1 Mobile Hand-off and Roaming

Hand-off would have implications on charging and service availability, including QoS. Admission control is used to guarantee service availability to the mobile node and sustainability of QoS to nodes being served in the new network. In the design of the emulation framework for the DQBCM system, mobility issues are not considered; hence this is left as an open issue for further research.

### 3.8.2 Network Services

3G networks offer a wide range of services (e.g., voice telephony, location based services and MMS) that belong to different CoS. In the emulation testbed, the implementation of CoS is achieved by defining characteristics of a traffic generator that is used. The goal of the traffic control is to manage resource allocation to different CoS and profiles and evaluate the QoS performance with varying network congestion levels.

### 3.8.3 Accounting, Charging and Billing

In the testbed. a minimum implementation of accounting is done. The aspects of charging and billing are not investigated. Research in this thesis is focused on pricing and its effects on network operations. The testbed investigates pricing only to the level of tariff determination as the level of congestion in the network (platinum profile) changes.

# Chapter 4

# The Architectural Design of the Evaluation Platform

## 4.1 Choice of Platform

Evaluating the performance of a mobile network billing scheme requires a real-life system offering a range of services to users. However, with the availability of statistics on usage patterns of the different services that are offered on mobile communication networks, modules of the billing system can be developed and tested. Simulators for mobile billing schemes are not readily available, hence network emulators need to be developed for testing specific functions of the billing scheme. The use of network emulators has the following advantages:

- Network emulators run real protocol stacks, which greatly improves their accuracy and reliability

- They have processing overhead that is characteristic of real systems. This not only improves their accuracy and reliability, but also it guarantees high performance if the real systems would use specialised equipment

The reasons given above led to the choice of building a custom emulator for evaluating important aspects of the proposed pricing scheme.

## 4.2 Objectives of the Implementation

The following were the main objectives of the emulation testbed:

- To create network service profiles and interfaces that would enable users to change their service usage profiles and to test the profile management function
- To develop an architecture for traffic control for the classes of service (CoS) and the profiles in each CoS
- To develop and test the admission control function
- To develop and test the tariff generation function

## 4.3 Network Management Components

The emulation testbed was implemented in two phases. The first phase consisted of a network management part, a radio access network component and three mobile units. The network management component resides in two i386 machines, each running Debian Linux. One of the machines was used as the media access gateway (AG) while the other was used as the network access controller (AC). In the second phase, the network traffic generators were added to the system, thus replacing the mobile units and eliminating the media access gateway.

The radio access network was implemented using a D-Link access point directly connected to the AG. Each of the mobile units had a WLAN interface card.

## 4.4 Network Topology

Figure 4.1 illustrates the conceptual topology that forms the basis of the evaluation platform. The topology is composed of several mobile units (MU), which provide access to network services and the interface for modifying of service usage profiles. The AG performs MU registration and enforces network authorisation to identify the different MUs. The AC performs the network management functions

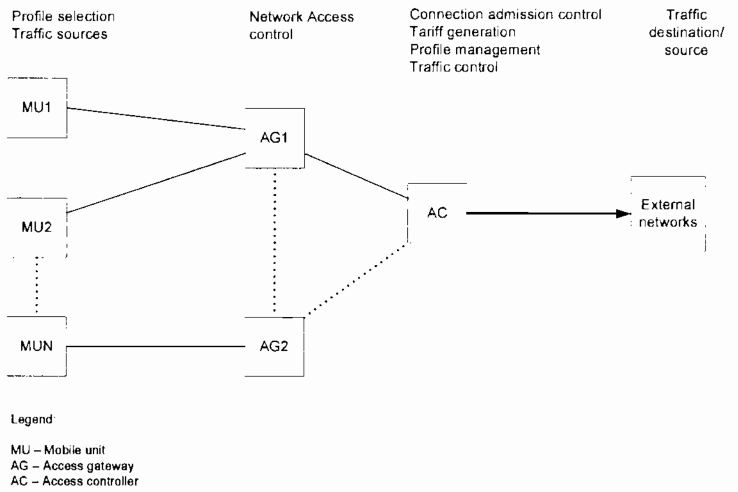


Figure 4.1: Conceptual topology of the implementation

of the DQBCM system, which include authentication and authorisation of the MUs, admission control, tariff generation, service profile management and traffic control. The layout of the first phase of the implementation is shown in Fig. 4.2, while the second phase is shown in Fig. 4.3.

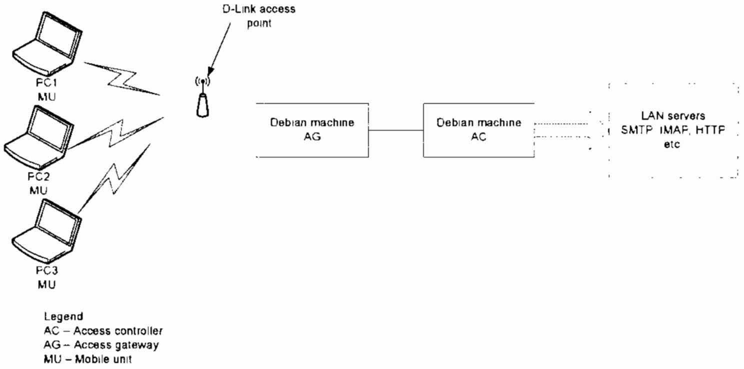


Figure 4.2: Layout of the First implementation phase

Figure 4.2 illustrates that the mobile units provided the users with access to services on a LAN. Access to the services was managed by the AC. The radio access network was secured by a firewall system configured on the AG. The mobile units allowed the author to test the web interface, where users could modify their service usage profiles. The evaluation of the tariff generation process required a

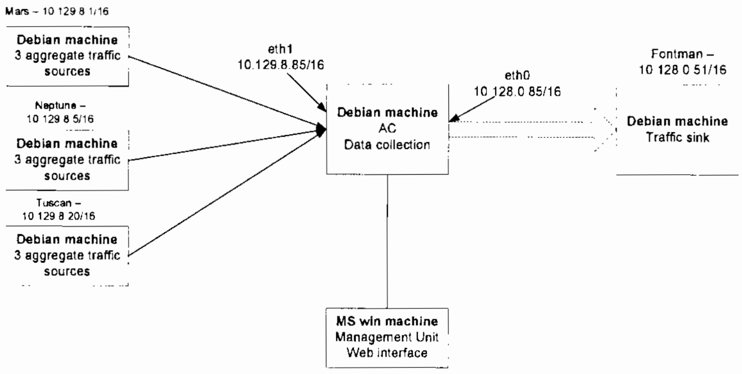


Figure 4.3: Layout of the second implementation phase

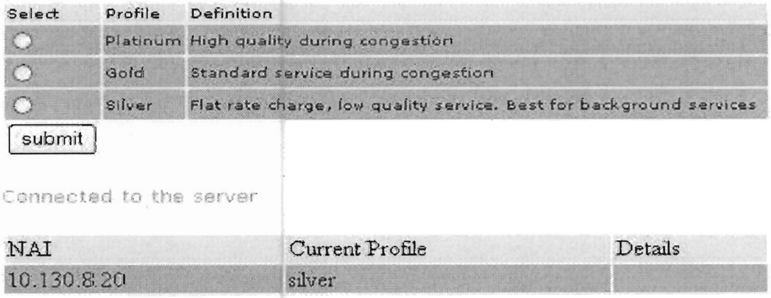
flexible and controllable set of traffic sources, which the mobile units could not provide. For this reason, the second phase of the implementation was carried out.

In the second phase, the mobile units, the access point and the AG were phased out and network traffic generators were introduced. As shown in Fig. 4.3, the traffic sources were implemented using three Linux PCs. Two additional network entities were introduced in the second phase. The first was a network management station, running Microsoft Windows. The station provided access to the web interface for emulating the modification of service usage profiles. The other entity was a traffic sink. The functions of the AC remained the same; however, the authentication and authorisation functions became redundant.

### 4.4.1 Mobile Units

In many networks, authentication and authorisation (AA) of the user equipment (UE) is required prior to gaining access to the network. In this testbed, the AA procedure facilitates the functioning of the traffic control and the service profile management agents. The Ethernet media access control (MAC) address was used to uniquely identify every MU that was registered on the testbed network. For evaluating the QoS performance achieved by the MUs when using different service profiles, the author considered using standard Internet applications, (e.g. SMTP, IMAP and HTTP) that were running on various servers on a LAN. However, it was impossible to predict the network usage pattern of these applications on the LAN, a feature that was necessary for evaluating the performance of the admission control and the tariff generation functions of the AC. For this reason, the author decided only to use the mobile units to test the profile modification interface and the AA procedure.

The MUs were issued with a network access identifier (NAI) during the registration process. The NAI was an IP address that was dynamically issued by the AG, and its validity was achieved through binding it to the MAC address of the MU. Users would change their service usage profiles through a standard web page hosted on the AC. On the page they would view and change their current profiles. Figure 4.4 shows the profile modification interface for one CoS; silver is the current profile for the user whose NAI is 10.130.8.20.



| Select | Profile | Definition |
|---|---|---|
| ○ | Platinum | High quality during congestion |
| ○ | Gold | Standard service during congestion |
| ○ | Silver | Flat rate charge, low quality service. Best for background services |

Connected to the server

| NAI | Current Profile | Details |
|---|---|---|
| 10.130.8.20 | silver | |

Figure 4.4: Profile modification page

Since the MUs did not exhibit the flexibility that was required for evaluating the traffic control and tariff management functions of the AC, network traffic generators (refer to section 4.4.5) were introduced.

### 4.4.2 Access Point

The role of the access point (AP) is to provide the radio network coverage of the WLAN. D-Link APs are normally ready to use after a small setup procedure. For the purpose of network security, access to the core network[1] through an AP needs to be controlled. In the testbed, the AP was connected to a media access gateway as shown in Fig. 4.2. By default, the AP forwards the data link (MAC) address of the mobile units. This property was useful in facilitating the AA enforcement at the AG.

[1] Core network, in this context, refers to the part of the network that requires access control

### 4.4.3 Media Access Gateway

The AG performed the registration of the MUs by assigning them unique IP addresses from a pre-defined pool of addresses. Each IP address was bound to the respective MU's MAC address, and the pair was sent as an AA binding message to the AC for authentication and authorisation. The transmission of the binding message was done through a UNIX socket that was established between the AG and AC whenever a new MU was detected.

An AA reply was received by the AG after the binding information had been processed, and this determined if the MU would be authorised to use the network or not. Authorisation of the MU was done by enabling its IP and MAC addresses in the access list of the firewall on the AG. Figure 4.5 illustrates the signalling information flow during the MU registration process.

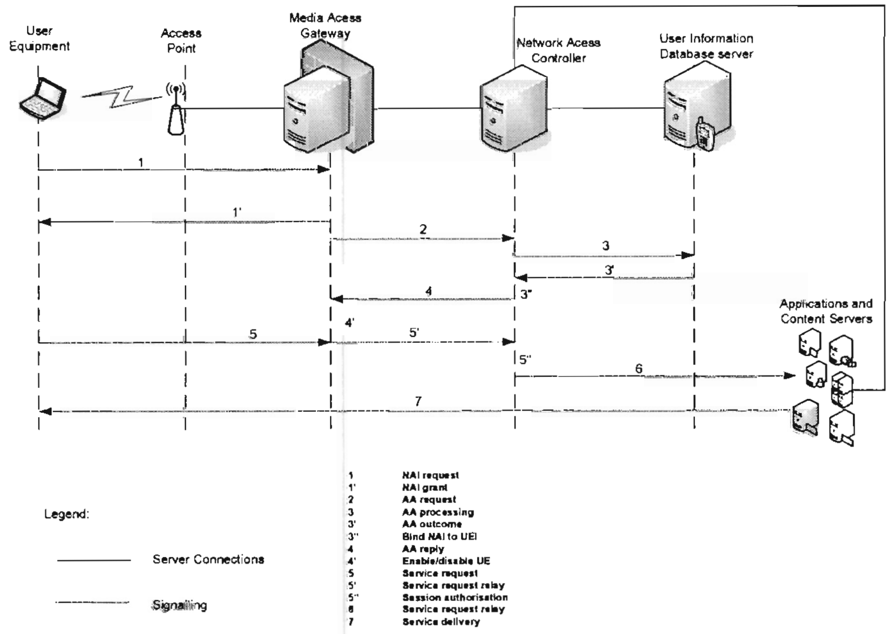


Figure 4.5: Signalling information flow in the testbed

#### Design Requirements

The design requirements for the media access gateway are as follows:

- Hardware requirements:
  - An i386 machine with two Ethernet network interface cards
  - An access point
- Software requirements
  - Debian Linux[2]

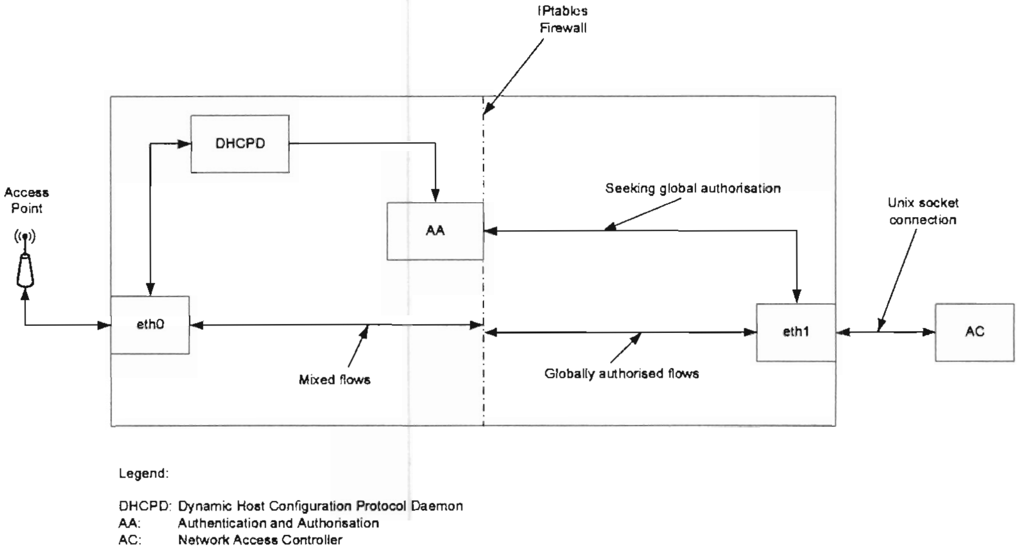

IPtables Firewall
DHCPD
AA
Seeking global authorisation
Access Point
Unix socket connection
eth0
eth1
AC
Mixed flows
Globally authorised flows
Legend:
DHCPD: Dynamic Host Configuration Protocol Daemon
AA: Authentication and Authorisation
AC: Network Access Controller

Figure 4.6: Topological design of the media access controller

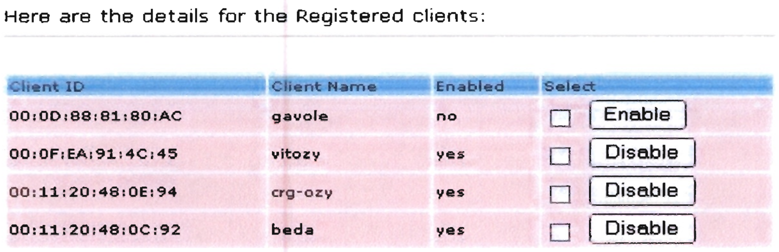
The AG setup is illustrated in Fig. 4.6. The AG initiated the registration of MUs through the *detect* and *enable* processes. The authorisation enforced by the AG is known as global authorisation.

### 4.4.4 Network Access Controller

The network access controller (AC) performs the following functions: authentication and authorisation of MUs, connection admission control, traffic management and control (e.g. bandwidth management), user profile management and tariff generation.

[2] Detailed design requirements are given in appendix B.3

On receiving an AA request from the AG, the AC attempts to authorise the MU by verifying its MAC address and the corresponding *enabled* status in a database. Figure 4.7 shows sample details of MUs that are currently registered on the network.



Here are the details for the Registered clients:

| Client ID | Client Name | Enabled | Select |
|---|---|---|---|
| 00:0D:88:81:80:AC | gavole | no | □ Enable |
| 00:0F:EA:91:4C:45 | vitozy | yes | □ Disable |
| 00:11:20:48:0E:94 | crg-ozy | yes | □ Disable |
| 00:11:20:48:0C:92 | beda | yes | □ Disable |

Figure 4.7: Details of registered clients

### Admission Control

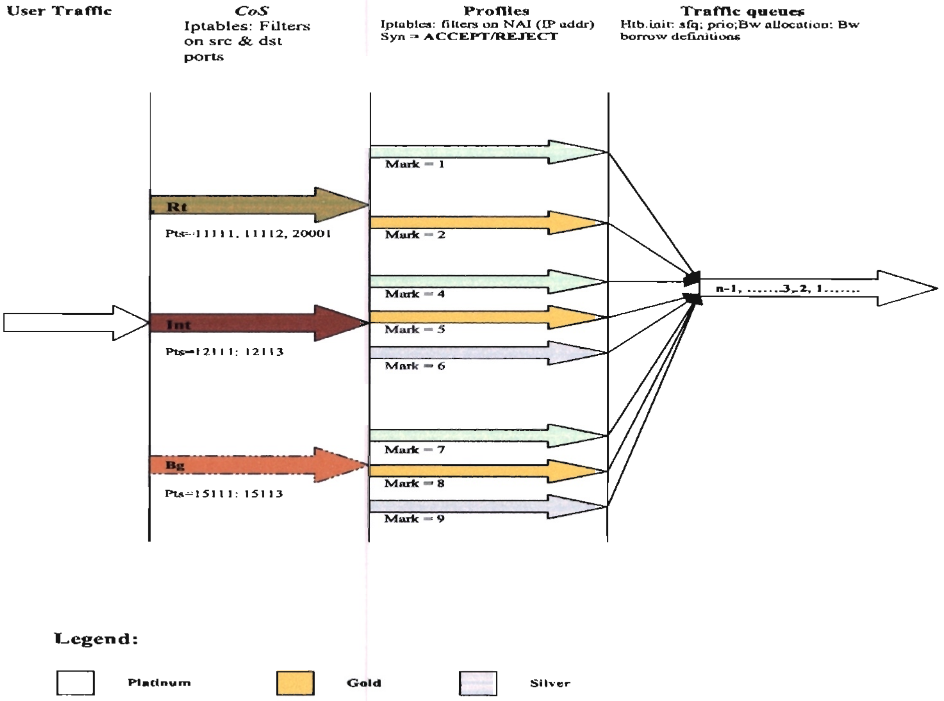


Figure 4.8: The testbed traffic control structure

Global authorisation allows the users to access the network and request services.

When a communication session establishment attempt is made, the admission control function (ACF) initiates a session authorisation process. It relies on the CoS and profile information that is defined in the traffic control architecture of the system. As discussed earlier, the ACF authorises new session establishment attempts until the value $I$ equals the value of $N$ for the affected profile.

In the AC block, the ACF was implemented by monitoring the number of TCP established between the MUs and servers on the LAN (Fig. 4.2). When the value of $I$ was less than $N$, the $SYN$[3] packet of each new connection was allowed to proceed to the destination, otherwise the ACF blocked it, thus preventing the establishment of new sessions.

The ACF operated at the profile level in each CoS. The CoS was identified by the destination and source port numbers in the TCP headers. The profile for the affected user was identified by the source or destination IP address (NAI) in the packet header. The implementation structure of the admission control and traffic control function is illustrated in Fig. 4.8.

### Traffic Control and Profile Management

Traffic control and management on the DQBCM testbed is done according to the DiffServ architecture [12]. However, the DiffServ architecture is not implemented on the testbed. Traffic control on the testbed considers peer-to-server (p2s) and server-to-server (s2s) applications only, hence packet classification into respective CoS is implemented using a static setup.

Packet classification was done using the *IPtables* package. A profile management function on the AC keeps track of profile changes made by users. A profile trigger function (PTF) responds to profile change requests from users, and it sends the user's identity plus the profile information to the profile management function, which would update the profile database. The information sent by the PTF includes: the user's NAI (IP address of the MU), the current (old) service profile, the requested (new) service profile and the CoS affected by the profile change.

When a profile change targets a profile whose capacity is fully used, i.e. $I = N$ in the target profile, enforcing the profile change would make $I > N$ hence

[3] $SYN$ packets are used to set up new TCP connections

compromising the QoS of the flows in the profile. In order to avoid this, profile change enforcement is done after verifying capacity availability to support new flows in the target profile.

The components of the traffic control function include a DiffServ function and the profile management function. As mentioned earlier the DiffServ function was not implemented on the testbed; instead traffic control was achieved using modules that are available in the Linux kernel [33, 34, 35, 36]. The testbed traffic control function was sufficient for evaluating the DQBCM performance.

Network resource management was implemented using the traffic control mechanism that has been partly described above, with the addition of a configurable open source script known as *htb.init* [34]. This script works on the principle of the hierarchical token bucket [36] and Linux traffic control. After classifying packets into respective CoS and profiles, packet marking is used to identify them. In a real DiffServ implementation, the ToS byte of each packet would be used to uniquely mark packets belonging to different profiles and CoS.

In the traffic control configuration, static allocation of bandwidth to each profile and CoS was defined. Resource (bandwidth) sharing between different profiles and CoS ensures that unused bandwidth in a CoS or profile is allocated to another CoS or profile where the demand is high. In evaluating the performance of different CoS and profiles, differential allocation of bandwidth and queue priority to the CoS and profiles was done. The full details of the traffic control evaluation are given in the next chapter.

### Pricing and Tariff Generation

The pricing agent is required to generate tariffs for the platinum profile. It probes the admission control function for the number of active sessions ($I_p$). Determination of the tariff rates is done separately for each CoS. The formula to calculate the pricing coefficients is given in Eq. 3.4. The pricing function was implemented on the AC by configuring the ACF to write to a pricing file where the pricing function read the values of $I_p$.

Once the tariff rates have been calculated, accounting of network resource usage is done to determine the usage for which the user should be charged. In the testbed, a portion of the required accounting procedure was implemented. This dealt with the initiation of the accounting session at the time the mobile user was granted global authorisation. It should be noted that, for the platinum profile, a new accounting session must be initiated whenever a price update occurs. This would correspond to the initiation of a new CDR sessions in 3G systems as a result of change in QoS parameters, tariff periods and other factors.

### 4.4.5 Traffic Generators

The network traffic generator [37] used in the testbed generated UDP and TCP traffic, and it had options for defining the payload size, source and destination port numbers and the target host IP address. These characteristics were considered adequate for the tests outlined in the next chapter.

At the traffic sources, the packet size, destination port numbers, and the number of concurrent connections for each source were varied to emulate traffic for different CoS. The aggregate traffic volume consisted of a variable number of independent connections. Each connection represented an application level traffic source of packets belonging to a specific CoS. The tests aimed at estimating the throughput and packet loss ratio achieved by each source under different network conditions as given in section 5.1.

### 4.4.6 Traffic Sink

The traffic sink listened for connections from the traffic sources. Once the connections were established, it received packets and echoed them back to the traffic sources. It listened for connections on specific port numbers, and provided the flexibility that was needed to test the network access controller functions.

# Chapter 5

# QoS Performance Evaluation

The architectural design of the evaluation framework was presented in Chapter 4. Several tests were performed using the framework in order to determine the QoS performance of the classes of service and the profiles. The results from the tests are analysed and discussed in this chapter. The evaluations show that the DQBCM scheme would lead to a practical system that meets user satisfaction and enables network operators and service providers to maximise their revenue.

## 5.1 Network Traffic Control and QoS Performance

In this section, a detailed report on the tests that were conducted on the evaluation platform to investigate the performance of entities used for network traffic control (e.g. resource management between classes of service and the profiles) is given. Fig. 4.3 illustrates the layout of the framework used in performing the evaluations. It shows the three computers that were used as clients to send traffic to the traffic sink using different destination port numbers.

The aim of the setup was to emulate sources of multimedia, interactive and background CoS traffic. The traffic source setup configuration details are given in table 5.1. Seven sets of tests (four on CoS evaluation and three on profile evaluation) were performed to evaluate the traffic control system. A discussion of the results is presented.

Table 5.1: Configuration details for different CoS

| *Client name* | *IP Address* | *CoS* | *Destination port no.* | *Packet size (bytes)* |
|---|---|---|---|---|
| Mars | 10.128.8.1 | Interactive | 11111 | 300-1000 |
| Neptune | 10.128.8.5 | Multimedia | 12111 | 100 |
| Tuscan | 10.128.8.20 | Background | 15111 | 500-1000 |

### 5.1.1 QoS Performance of Classes of Service

QoS performance tests for the CoS were done as follows: the evaluation framework was configured to allocate certain levels of bandwidth to each CoS. In the first two tests, the framework depicted two networks; one with an abundance of resources and the other had scarce resources. This was achieved by setting bandwidth allocation using the traffic control tools on the network access controller (AC). In the network with abundant resources, each CoS was allocated 170 Kbps of bandwidth, and the network with scarce resources each CoS was allocated 18Kbps. To achieve different packet forwarding preferences, the multimedia CoS was assigned the highest queue priority, the interactive CoS received the second preference, while the background CoS was assigned the lowest queue priority.

The traffic load sent on the network by the clients was increased in steps and each test was run for approximately two minutes. The number of connections reflects the traffic load on the network. The packets sent by each client were counted and the number of packets received at the traffic sink was recorded too. Table 5.2 shows the network performance results of a network with an abundance of resources, while table 5.3 presents the network performance results of a network with scarce resources.

A graphical illustration of the network performance under the abundance and the scarce resource availability is given in Fig. 5.1. Two further tests were performed with differential bandwidth allocation to the CoS. The multimedia CoS was allocated 28 Kbps, the interactive CoS received 16 Kbps and the background CoS received 10 Kbps. In the first set of these tests, the CoS were not allowed to borrow unused resources from each other, i.e. bandwidth sharing was not enabled, while in the second case bandwidth sharing was enabled. Table 5.4 illustrates

Table 5.2: Network with abundant resources (Relatively over-provisioned)

| *CoS* | *No. of connections* | *Total Packets send* | *Total Packets received* | *Performance Index* |
|---|---|---|---|---|
| Multimedia | 1 | 21682 | 21593 | 0.9959 |
| Interactive | 1 | 4288 | 4274 | 0.9967 |
| Background | 1 | 3743 | 3731 | 0.9968 |
| Multimedia | 10 | 14413 | 14413 | 1 |
| Interactive | 10 | 2843 | 2843 | 1 |
| Background | 10 | 2521 | 2520 | 0.9996 |
| Multimedia | 50 | 20173 | 20173 | 1 |
| Interactive | 50 | 4006 | 4006 | 1 |
| Background | 50 | 3540 | 3540 | 1 |
| Multimedia | 100 | 21276 | 21276 | 1 |
| Interactive | 100 | 4270 | 4242 | 0.9934 |
| Background | 100 | 3792 | 3761 | 0.9918 |
| Multimedia | 500 | 20232 | 19974 | 0.9872 |
| Interactive | 500 | 4165 | 4036 | 0.9690 |
| Background | 500 | 3967 | 3593 | 0.9057 |
| Multimedia | 1000 | 20260 | 19202 | 0.9478 |
| Interactive | 1000 | 4159 | 4030 | 0.9690 |
| Background | 1000 | 4462 | 3588 | 0.8041 |

the results of the network performance without bandwidth sharing, while table 5.5 presents the results obtained with bandwidth sharing. The graphical representation of the two cases is given in Fig. 5.2.

### Analysis

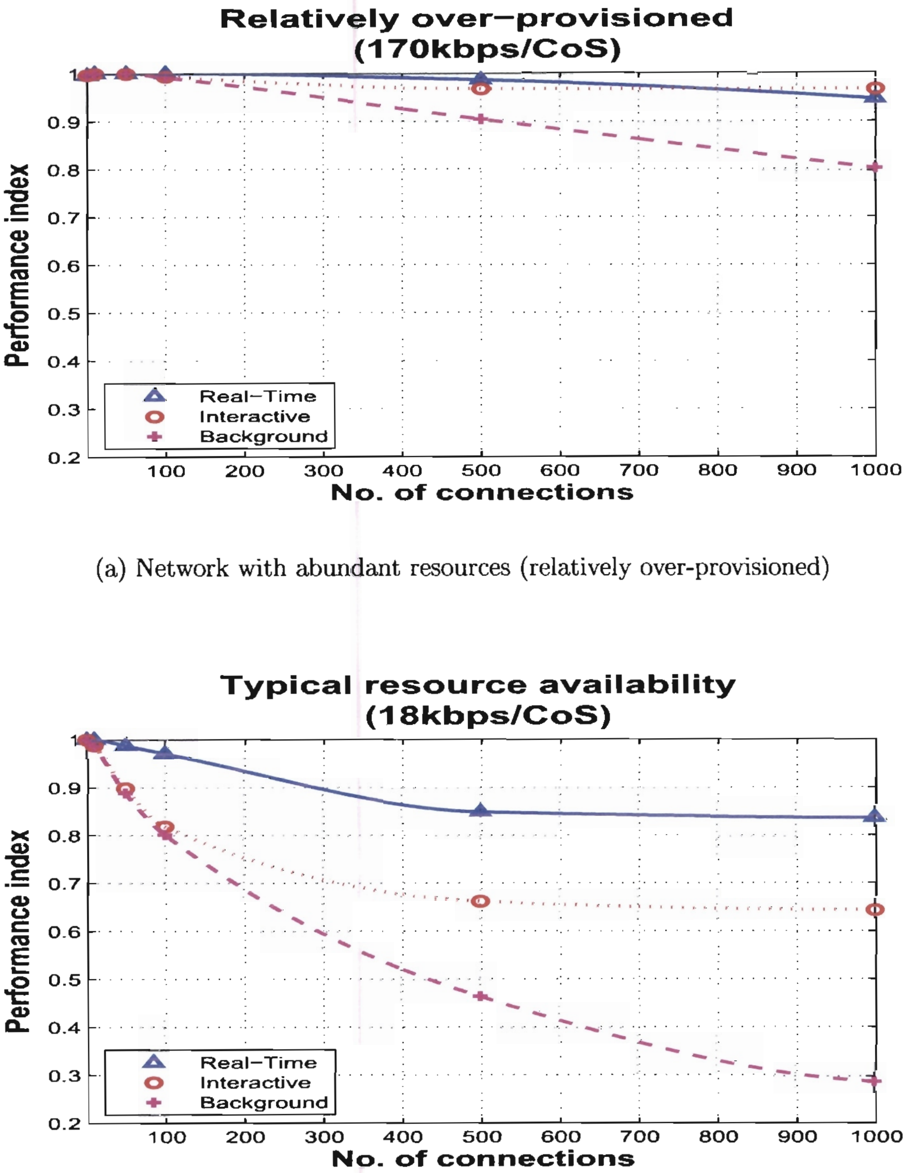
The results presented clearly indicate that network performance depends on resource availability. The relatively over-provisioned network depicted a good performance level, while the typical resource availability (scarce resources) case indicates a rapid degradation in the network performance as the number of connections (network load) increases. With the highest queue priority, the multimedia CoS shows better performance. The background CoS, which was set to the lowest queue priority suffered the worst performance degradation as the network load

Table 5.3: Network with scarce resources, depicts the typical resource availability on communication networks

| *CoS* | *No. of connections* | *Total Packets send* | *Total Packets received* | *Performance Index* |
|---|---|---|---|---|
| Multimedia | 1 | 2080 | 2080 | 1 |
| Interactive | 1 | 405 | 405 | 1 |
| Background | 1 | 356 | 356 | 1 |
| Multimedia | 10 | 2072 | 2072 | 1 |
| Interactive | 10 | 409 | 404 | 0.9878 |
| Background | 10 | 363 | 358 | 0.9862 |
| Multimedia | 50 | 2079 | 2079 | 0.9875 |
| Interactive | 50 | 444 | 399 | 0.8986 |
| Background | 50 | 395 | 351 | 0.8886 |
| Multimedia | 100 | 2201 | 2137 | 0.9709 |
| Interactive | 100 | 509 | 415 | 0.8153 |
| Background | 100 | 459 | 368 | 0.8017 |
| Multimedia | 500 | 2909 | 2473 | 0.8501 |
| Interactive | 500 | 728 | 483 | 0.6634 |
| Background | 500 | 919 | 428 | 0.4657 |
| Multimedia | 1000 | 2777 | 2329 | 0.8387 |
| Interactive | 1000 | 695 | 448 | 0.6446 |
| Background | 1000 | 1396 | 403 | 0.2887 |

was increased.

When differential bandwidth allocation is introduced (refer to Fig. 5.2), the multimedia CoS shows better performance than the interactive and background CoS, which achieve reduced performance as compared to the case when each CoS received 18 Kbps. With resource (bandwidth) sharing, the multimedia CoS still achieves good performance and the interactive and background CoS achieve improved performance. Rapid performance degradation occurs when the number of connections exceeds 100.



(a) Network with abundant resources (relatively over-provisioned)

(b) Network with scarce resources, i.e. the typical network resource availability on communication networks

Figure 5.1: Network performance under (a) Abundant and (b) scarce resources

### Discussion

The results presented above show that network bandwidth is a critical resource that influences the QoS performance of networks carrying multi-service traffic.

Table 5.4: Differential resource allocation between CoS - without bandwidth sharing

| *CoS* | *No. of connections* | *Total Packets send* | *Total Packets received* | *Performance Index* |
|---|---|---|---|---|
| Multimedia | 1 | 2093 | 2093 | 1 |
| Interactive | 1 | 231 | 231 | 1 |
| Background | 1 | 130 | 130 | 1 |
| Multimedia | 10 | 3642 | 3535 | 0.9706 |
| Interactive | 10 | 421 | 394 | 0.9359 |
| Background | 10 | 235 | 217 | 0.9234 |
| Multimedia | 50 | 2909 | 2909 | 1 |
| Interactive | 50 | 364 | 320 | 0.8791 |
| Background | 50 | 224 | 176 | 0.7857 |
| Multimedia | 100 | 2738 | 2677 | 0.9777 |
| Interactive | 100 | 392 | 298 | 0.7602 |
| Background | 100 | 263 | 165 | 0.6274 |
| Multimedia | 500 | 3512 | 3067 | 0.8733 |
| Interactive | 500 | 598 | 346 | 0.5786 |
| Background | 500 | 682 | 185 | 0.2713 |
| Multimedia | 1000 | 3564 | 3119 | 0.8751 |
| Interactive | 1000 | 617 | 454 | 0.4117 |
| Background | 1000 | 1231 | 353 | 0.2868 |

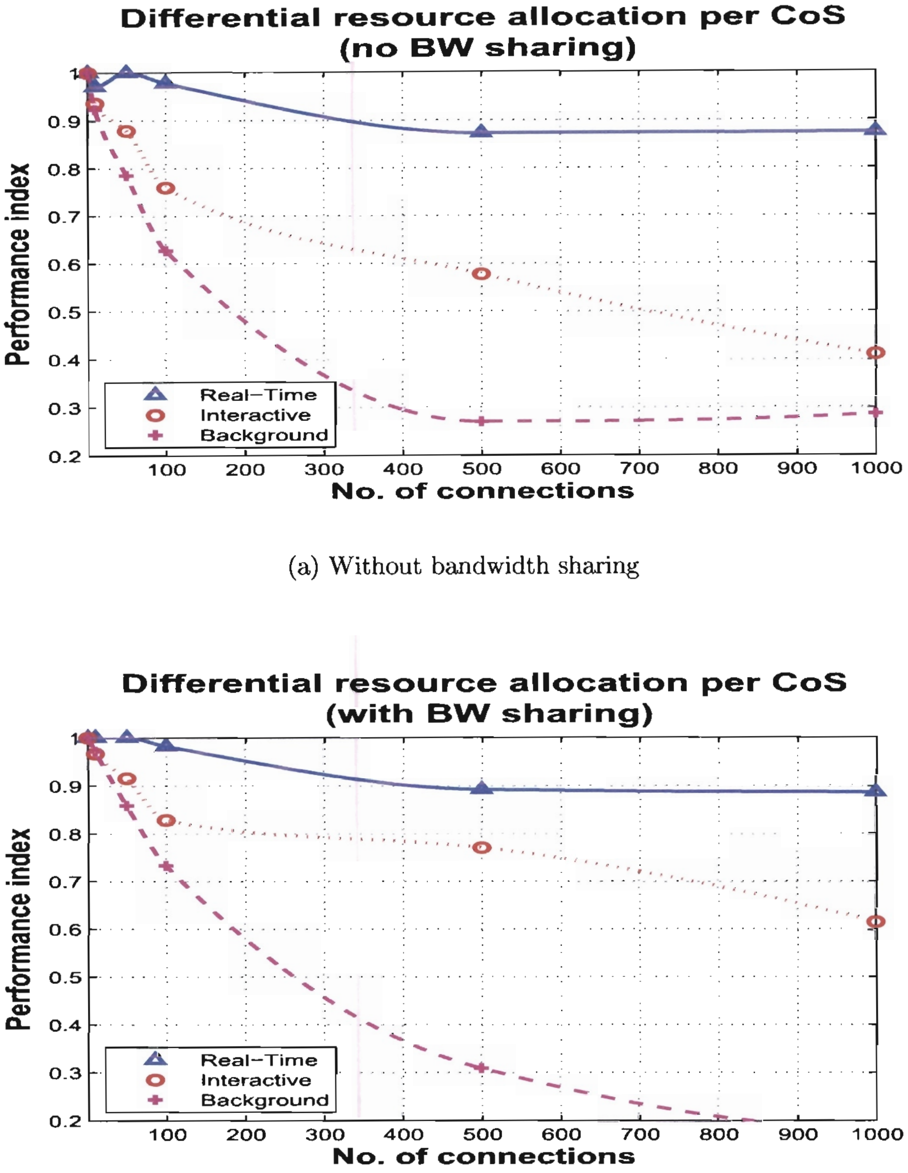
The CoS with a higher allocation of bandwidth is always expected to perform better than one with less bandwidth allocation; the trend depicted in table 5.2 and Fig. 5.1 (a) when the number of connections exceeds 500 defies this expectation, but it can be associated to tolerances within the evaluation framework. The setting of high queue priority for the multimedia CoS was done according to the DiffServ convention, where traffic in the multimedia CoS requires low delay guarantee. Thus lower packet loss due to buffer overflows was expected in the multimedia CoS as shown by the above results.

Table 5.5: Differential resource allocation between CoS - with bandwidth sharing

| *CoS* | *No. of connections* | *Total Packets send* | *Total Packets received* | *Performance Index* |
|---|---|---|---|---|
| Multimedia | 1 | 2016 | 2016 | 1 |
| Interactive | 1 | 229 | 229 | 1 |
| Background | 1 | 132 | 131 | 1 |
| Multimedia | 10 | 2609 | 2609 | 1 |
| Interactive | 10 | 304 | 294 | 0.9671 |
| Background | 10 | 175 | 170 | 0.9714 |
| Multimedia | 50 | 4378 | 4378 | 1 |
| Interactive | 50 | 535 | 490 | 0.9159 |
| Background | 50 | 319 | 274 | 0.8589 |
| Multimedia | 100 | 4203 | 4128 | 0.9822 |
| Interactive | 100 | 554 | 459 | 0.8285 |
| Background | 100 | 357 | 262 | 0.7329 |
| Multimedia | 500 | 4037 | 3603 | 0.8925 |
| Interactive | 500 | 656 | 406 | 0.7713 |
| Background | 500 | 725 | 225 | 0.3103 |
| Multimedia | 1000 | 4062 | 3603 | 0.8870 |
| Interactive | 1000 | 648 | 399 | 0.6157 |
| Background | 1000 | 1227 | 227 | 0.1850 |

### 5.1.2 QoS Performance of Individual Profiles

QoS evaluation for individual profiles was done using one CoS with three profiles. Each profile was identified using the IP address of the traffic source. Traffic from the three profiles was sent to the same destination port number and the packet size was the same; the traffic characteristics are summarised in table 5.6. The layout used in performing the profile evaluation tests is similar to Fig. 4.3. The tests were performed in steps and the traffic load in each profile was increased by increasing the number of connections from each traffic source. The profile performance tests were done in three sets. In the first set all profiles were assigned equal bandwidth, i.e. 18 Kbps. In the second and third set of tests, the platinum profile was assigned 28 Kbps, the gold profile received 16 Kbps, while the silver profile got 10 Kbps. In the DQBCM design, all profiles are set to the same



(a) Without bandwidth sharing

(b) With bandwidth sharing

Figure 5.2: Differential bandwidth allocation between CoS

queue priority, thus traffic from the three clients was expected to experience same treatment on the network, i.e. equal delay and packet drop.

Table 5.7 shows the network performance characteristics of the three profiles when

Table 5.6: Traffic characteristics for the three profiles

| *Profile name* | *Source IP address* | *Destination port no.* | *Packet size (bytes)* |
|---|---|---|---|
| Platinum | 10.129.8.1 | 11111 | 1000 |
| Gold | 10.129.8.5 | 11111 | 1000 |
| Silver | 10.129.8.20 | 11111 | 1000 |

evaluated with no effective bandwidth control, and Fig. 5.3 gives the graphical presentation of the evaluations.

Table 5.7: Equal bandwidth allocation to the profiles - no bandwidth control

| *Profile* | *No. of Connections* | *Packets send* | *Packets received* | *Performance Index* |
|---|---|---|---|---|
| Platinum | 1 | 630 | 630 | 1 |
| Gold | 1 | 630 | 630 | 1 |
| Silver | 1 | 630 | 630 | 1 |
| Platinum | 10 | 807 | 805 | 0.9975 |
| Gold | 10 | 807 | 805 | 0.9975 |
| Silver | 10 | 807 | 805 | 0.9975 |
| Platinum | 50 | 751 | 701 | 0.9334 |
| Gold | 50 | 751 | 709 | 0.9441 |
| Silver | 50 | 751 | 709 | 0.9441 |
| Platinum | 100 | 817 | 725 | 0.8874 |
| Gold | 100 | 817 | 725 | 0.8874 |
| Silver | 100 | 817 | 725 | 0.8874 |
| Platinum | 500 | 844 | 597 | 0.7073 |
| Gold | 500 | 1073 | 597 | 0.5564 |
| Silver | 500 | 1089 | 597 | 0.5482 |
| Platinum | 1000 | 796 | 549 | 0.6897 |
| Gold | 1000 | 1025 | 549 | 0.5356 |
| Silver | 1000 | 1549 | 549 | 0.3544 |

The network performance characteristics of the profiles with differential bandwidth allocation and no bandwidth sharing are presented in table 5.8, while the performance characteristics with bandwidth sharing are given in table 5.9. The

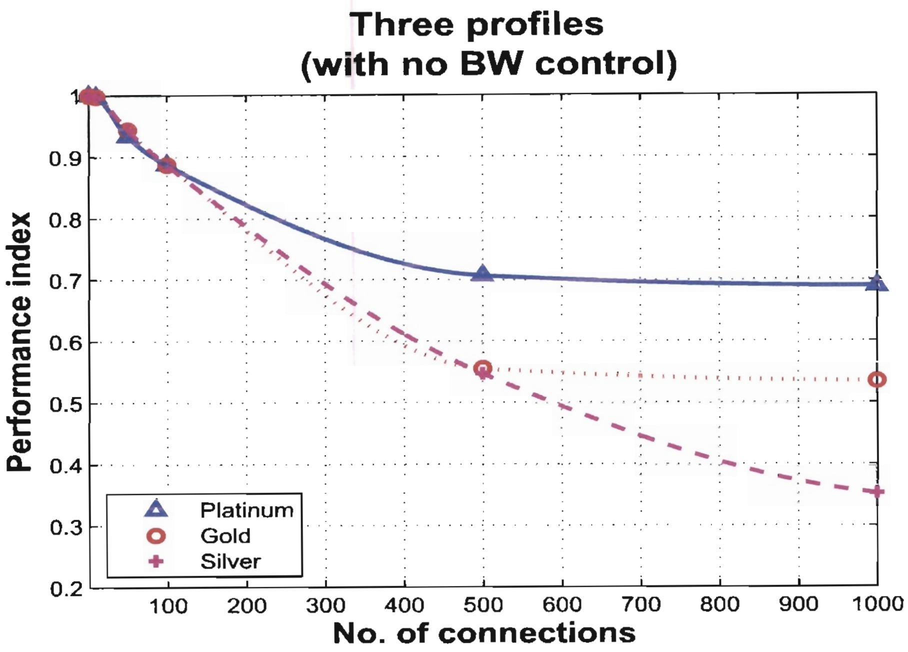


Figure 5.3: Performance index at the profile level - no bandwidth control

graphical presentation of the performance characteristics of the profiles is given in Fig. 5.4.

### Analysis

The graph in Fig. 5.3 shows that the three profiles achieve equal performance index until the number of connections exceeds 100, when some traffic sources achieve better network performance than the others. In this case the platinum source[1] performs better than the gold and silver, the gold and silver sources on the other hand show an alternating network performance pattern.

When differential bandwidth allocation is introduced, the platinum profile performances better than the gold and silver profiles. A general improvement in the performance index of the three profiles is shown in Fig. 5.4 (b), which is attributed to bandwidth sharing.

[1] The term source is used instead of profile because there is no characteristic differentiation on traffic from the different sources.

Table 5.8: Differential bandwidth allocation to profiles - without bandwidth sharing

| *Profile* | *No. of Connections* | *Packets send* | *Packets received* | *Performance Index* |
|---|---|---|---|---|
| Platinum | 1 | 938 | 938 | 1 |
| Gold | 1 | 542 | 542 | 1 |
| Silver | 1 | 342 | 342 | 1 |
| Platinum | 10 | 1138 | 1136 | 0.9982 |
| Gold | 10 | 658 | 656 | 0.9970 |
| Silver | 10 | 418 | 412 | 0.9856 |
| Platinum | 50 | 890 | 852 | 0.9573 |
| Gold | 50 | 534 | 492 | 0.9213 |
| Silver | 50 | 358 | 312 | 0.8715 |
| Platinum | 100 | 1124 | 1036 | 0.9217 |
| Gold | 100 | 688 | 596 | 0.8663 |
| Silver | 100 | 472 | 376 | 0.7966 |
| Platinum | 500 | 1083 | 840 | 0.7756 |
| Gold | 500 | 960 | 484 | 0.5042 |
| Silver | 500 | 804 | 308 | 0.3831 |
| Platinum | 1000 | 1083 | 848 | 0.7830 |
| Gold | 1000 | 960 | 488 | 0.5083 |
| Silver | 1000 | 1304 | 308 | 0.2362 |

### Discussion

When all profiles are allocated equal bandwidth, traffic from each profile will compete for network resources equally. If the packet sending rate for all profiles is the same, no profile will have resources from another profile relocated to it, since at anytime the demand for resources by all profiles will be the same. Thus the network performance index of all profiles is expected to be the same as depicted in Fig. 5.3 when the number of connections are less than 100. However, as the load increases, the network will drop packets randomly; hence some profiles might achieve better performance than others. The achieved performance would not be predictable during congestion, as illustrated in Fig. 5.3, when the number of connections exceeds 100.

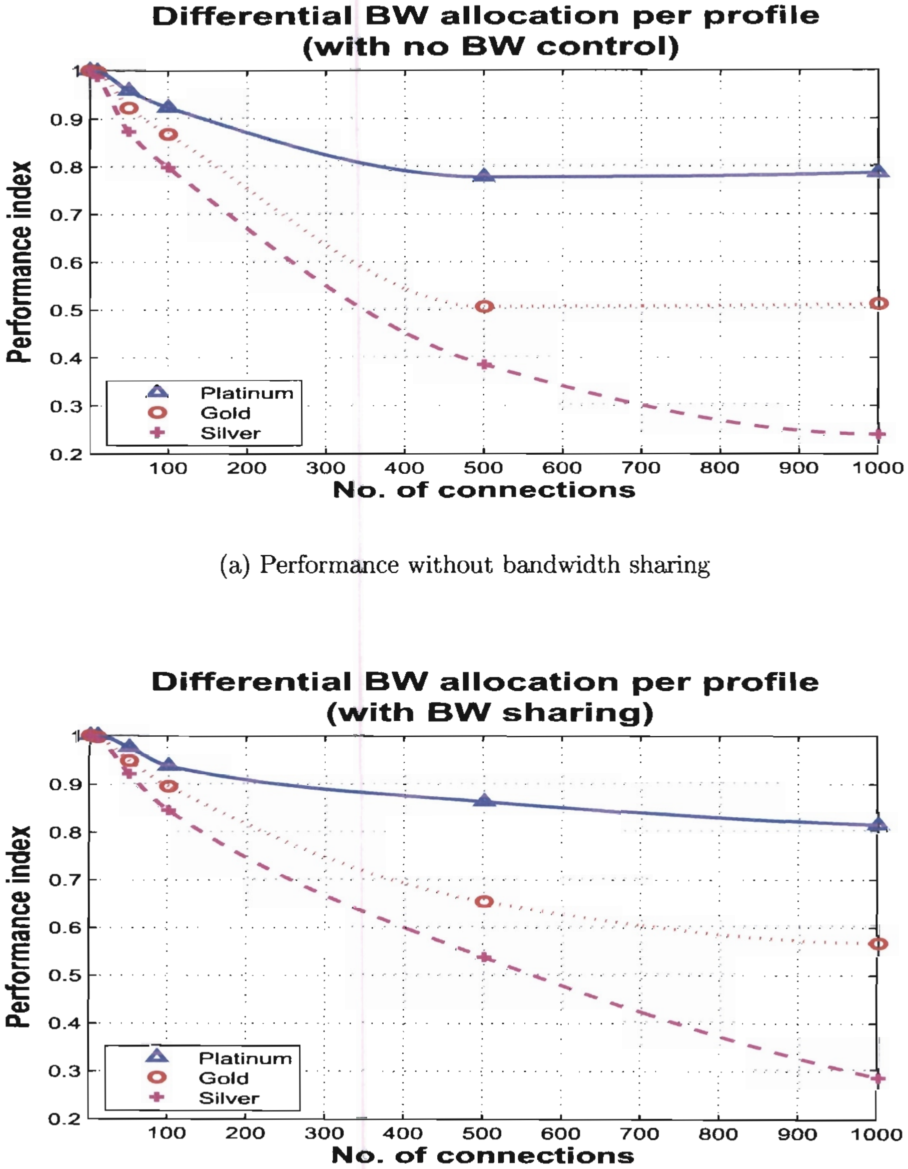
As illustrated in Fig. 5.4, profiles with higher levels of QoS resources (e.g. band-

Table 5.9: Differential bandwidth allocation between profiles - with bandwidth sharing

| *Profile* | *No. of Connections* | *Packets send* | *Packets received* | *Performance Index* |
|---|---|---|---|---|
| Platinum | 1 | 921 | 921 | 1 |
| Gold | 1 | 537 | 537 | 1 |
| Silver | 1 | 341 | 341 | 1 |
| Platinum | 10 | 946 | 946 | 1 |
| Gold | 10 | 553 | 551 | 0.9964 |
| Silver | 10 | 354 | 352 | 0.9944 |
| Platinum | 50 | 1346 | 1312 | 0.9747 |
| Gold | 50 | 801 | 759 | 0.9476 |
| Silver | 50 | 522 | 480 | 0.9195 |
| Platinum | 100 | 1436 | 1344 | 0.9359 |
| Gold | 100 | 867 | 775 | 0.8939 |
| Silver | 100 | 588 | 496 | 0.8435 |
| Platinum | 500 | 1791 | 1544 | 0.8621 |
| Gold | 500 | 1371 | 895 | 0.6528 |
| Silver | 500 | 1060 | 568 | 0.5358 |
| Platinum | 1000 | 1303 | 1056 | 0.8104 |
| Gold | 1000 | 1091 | 615 | 0.5637 |
| Silver | 1000 | 1384 | 392 | 0.2832 |

width) achieve better QoS performance as compared to profiles with less bandwidth. As discussed earlier, network operators can improve their revenue by attracting more users to profiles that offer better service (e.g., the platinum profile, whose tariff rates increase with the number of communication sessions) especially during congested periods.

It is clear from the graphs in Fig. 5.4 (a) and (b) that resource sharing leads to improved network performance for all profiles. During the two minute period when traffic for each profile was sent across the network. times when each profile had unused resources were expected. During these times. the system relocates the unused resources to other profiles that had packets to send. This behaviorbehaviour is what led to improved QoS performance of the three profiles, thus the resource utilisation efficiency was improved.



(a) Performance without bandwidth sharing

(b) Performance with bandwidth sharing

Figure 5.4: Performance index at the profile level with differential bandwidth allocation

Comparing Fig. 5.2 and Fig. 5.4, it can be seen that enforcing bandwidth differentiation at the profile level achieves QoS improvement for the affected profiles in a CoS; bandwidth differentiation between different CoS still achieves improved performance, especially when the network is not congested. The silver profile clearly depicts this when its performance improves due to bandwidth sharing before congestion sets in, but it performs poorly under extremely congested conditions.

## 5.2 Admission Control

As discussed in section 4.4.4, the ACF should verify the availability of network capacity in the requested service profile before a service request can be admitted. The service usage profile management tests performed in section 5.3 illustrate the importance of this configuration in the ACF. In testing the admission control function, the setups in Fig. 4.2 and Fig. 4.3 were used. Using the setup in Fig. 4.2, an attempt to establish a connection (e.g. remote shell using *ssh*) from a mobile unit to a server on the the LAN was made. The value of $N_p$ for the profile affected by the connection was set to 4, i.e. $N_p = 4$. This value of $N_p$ would permit the establishment of four *ssh* connections.

When an attempt to establish the connections was made, the first four succeed. An attempt to exceed the pre-set maximum value of admissible connections failed. When the value of $N_p$ was increased, more connections could be setup. Decreasing the value of $N_p$ after the maximum number of admissible connections, i.e. 4 in this illustration had been made did not have an effect on the established connections, i.e. it did not lead to the termination of an on-going *ssh* session. The setup in Fig. 4.3 was used to test the establishment of TCP and UDP connections. The value of $N_p$ was set to 50 or 100 and after synchronising the clocks of the three traffic sources they were initialised to make connections to the traffic sink. The synchronising of the clocks of the traffic sources was done using the *rdate* tool available in Linux and the automated initialisation of the traffic sources was achieved using the *cron* facility, available in Linux.

Two sets of tests were performed using TCP and UDP traffic. Both tests were done by varying the value of $N_p$ and configuring each source to establish as many connections as possible. The performance results when using TCP traffic are given in table 5.10 and a graphical representation is given in Fig. 5.5. The

Table 5.10: Admission control performance using TCP traffic

| Client Name | Maximum Connections in profile ($N_p$) | Admitted Connections per Client |
|---|---|---|
| Mars | 30 | 10 |
| Neptune | 30 | 2 |
| Tuscan | 30 | 18 |
| Mars | 50 | 33 |
| Neptune | 50 | 6 |
| Tuscan | 50 | 12 |
| Mars | 75 | 42 |
| Neptune | 75 | 5 |
| Tuscan | 75 | 29 |
| Mars | 100 | 48 |
| Neptune | 100 | 24 |
| Tuscan | 100 | 33 |
| Mars | 150 | 64 |
| Neptune | 150 | 72 |
| Tuscan | 150 | 25 |
| Mars | 200 | 100 |
| Neptune | 200 | 65 |
| Tuscan | 200 | 57 |
| Mars | 250 | 118 |
| Neptune | 250 | 35 |
| Tuscan | 250 | 113 |

admission control function was designed to support TCP traffic only, hence it was not expected to work with UDP traffic.

### Analysis

The results in table 5.10 and Fig. 5.5 show that the admission control function admitted varying number of connections from the three clients, and no specific client occupied the full capacity of a given profile. It is also clear from table 5.10 that the ACF had performance tolerances, since the number of admitted connections exceeded the predefined maximum number during some occasions.

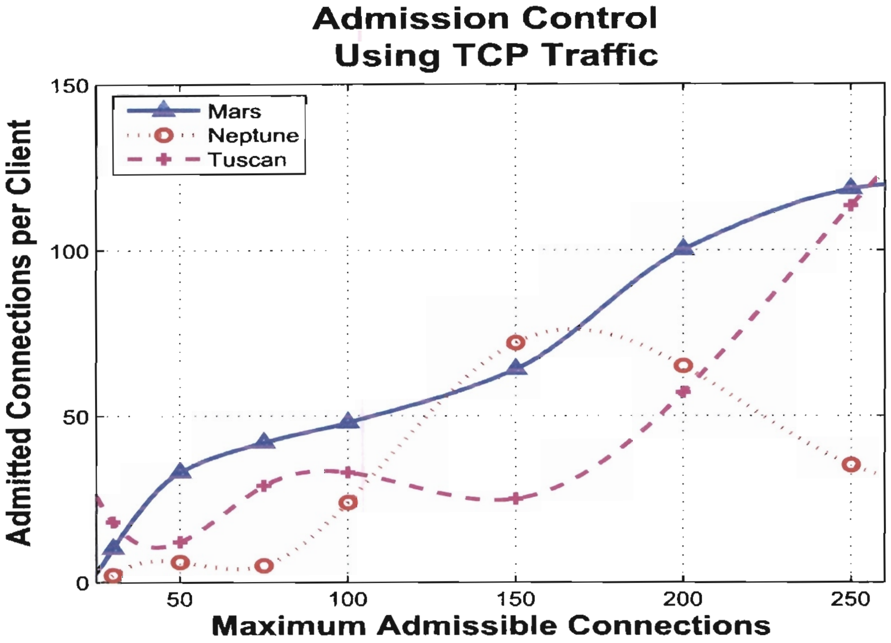


Figure 5.5: Admission control performance with TCP traffic

### Discussion

The results above indicate that at the profile level, where the ACF operates, preference is not given to specific clients. The clients, which represent mobile units for different users, compete to establish connections in the same way they would compete for network resources. This is visible from the pattern of graphs in Fig. 5.5, where different clients are able to establish varying number of connections and at some point each of the three clients establishes the highest number of connections.

## 5.3 Profile Management

When users post their service usage profiles, the profile management agent updates the profile database and effects the changes in the traffic control system. The effectiveness of the profile management agent was determined by evaluating the profile management process. The different parts of the profile management agent, i.e. user interface, profile trigger function and the profile enforcement server process were evaluated.

The user interface was web-based and its evaluation was based on observations of its reliability to enable the users to select a preferred profile. Evaluation of the profile trigger function and the profile enforcement process determined the success and failure of performing profile modification as requested by the user.

The profile management function worked in relation to the admission control and the traffic management functions. Evaluation of the profile management performance centred on determining that user profile change requests were processed and enforced without causing unexpected degradation on QoS for on-going sessions.

The results of this evaluation were mainly observational, i.e. service profile changes that were requested when the user was not involved in a data transfer session (out-of-session) were enforced by updating the user profile database, while service profile changes involving an on-going session (in-session) involved verifying the availability of capacity in the requested profile.

The server output illustrating the process of events when a user changes from one profile to another is given in appendix C.6.

### Discussion

Enforcement of out-of-session profile changes achieved 100% success.

Performing profile changes for on-going communication sessions showed that the maximum limit of the admissible flows $N$ in a profile could be exceeded. This was as a result of the ACF's reliance on the TCP $SYN$ packet for performing admission control. On-going sessions do not require re-establishment of a connection; hence TCP $SYN$ packets are not used. Thus before on-going flows are admitted to a new profile, the ACF should be modified to perform verification of capacity availability in the target profile. The modification involves receiving a real-time update from the ACF of the number of sessions that had been admitted to the requested profile. This step was not implemented on the testbed, thus future work can be done in this area.

## 5.4 Pricing and Accounting

### 5.4.1 Pricing

In the DQBCM system, pricing of resources in the platinum profile depends on feedback from the ACF. For the silver and gold profiles, the tariff rates would be independent of the demand of resources in the network; however, for the platinum profile, the tariff rates would vary with the number of admitted connections as reported by the ACF. The performance of the pricing function greatly depends on the admission control function.

Since the role of the pricing function mainly involves computation of tariff rates using the information obtained from the ACF, the amount of system and network resources it would consume would be little. Network resources would only be needed if the pricing updates need to be sent to an accounting function that is located remotely.

### 5.4.2 Accounting

The level of accounting performed on the emulation testbed relied on the IP accounting module of the Linux operating system. The author did not investigate the performance of accounting for the DQBCM system; however, it is noted that accounting is a crucial part of charging, and as mentioned in chapter 2, it is through accounting that the system can associate resource usage with the users concerned. Depending on the accounting metric (e.g. duration in seconds or volume of data transmitted) that is used, the accounting function would use the tariff rate that is advertised by the pricing function to generate the CDRs for the active sessions. When new prices are advertised for the platinum profile, the accounting function would initiate new CDR sessions.

In a system with many users and services, the accounting function would consume a substantial amount of system resources and its accuracy and speed would greatly influence the performance of the billing system.

## 5.5 Comparative Evaluation of the DQBCM Pricing Proposal

The features of the Dynamic QoS-Based Charging Model (DQBCM) are summarised in table 5.11. They are as follows:

- The scheme is compliant with IP networks - it targets 3G and next generation networks (NGN), which are shifting to an all-IP setup.
- Accounting for resource usage is required for both the gold and the platinum profiles.
- The connection admission control (CAC) controls congestion in the network and the DiffServ architecture facilitates QoS provisioning through traffic control and management.
- Individual QoS is achieved in the platinum and the gold profiles.
- High network efficiency is achieved through the lowering of the platinum profile tariffs during periods when the network has abundant resources - this encourages usage of the network during off-peak hours.
- The characteristics of different profiles provide social fairness since the users can select the easy to understand profiles according to their preferences - the silver profile would be affordable to low income users.
- The traffic control strategy works on short time frames, since it is designed for inter-working with the UMTS policy based admission control which is used for every connection attempt.

Table 5.11: Comparative evaluation of the proposed pricing scheme

| **Evaluation Index / Pricing Scheme** | **Flat** | **Priority** | **DQBCM** |
|---|---|---|---|
| Compliance | IP | IP | IP |
| Billing Measures | No | Yes | Yes |
| Congestion control and Traffic management | No | Yes(rel) | Yes (CAC) |
| Individual QoS | No | No | Yes |
| Network efficiency | Low | High | High |
| Economic efficiency | Low | High | High |
| Social fairness | Yes | No | Yes |
| Time frame | Long | Short | Short |

# Chapter 6

# Conclusions

## 6.1 Summary and Discussion

Using the results of the online survey, it can be concluded that users have a general understanding of the pricing of different services that are offered on mobile networks, and their willingness to pay higher charges for multimedia services is an indication that the DQBCM system is likely to receive acceptance by the users.

The DQBCM addresses various challenges of pricing schemes for mobile and other IP networks. Through the use of service profiles, the challenge of user involvement is addressed; by selecting appropriate profiles, users are able to influence the QoS received and also the affordability of the services they use. The storage of profile information on network database systems enables the network management system to make decisions on behalf of the users without directly involving them in speedy decision making processes. Users are only required to specify their service profiles at their convenient time (e.g. prior to service usage). Users are presented by friendly interfaces for profile modifications and the use of network agents facilitates the functions of all entities of the DQBCM network.

The DQBCM service profiles create a hybrid pricing scheme that achieves improved network efficiency. This is achieved through resource sharing between profiles, which ensures excess network resources in one profile are to be used by traffic in another profile that may be experiencing congestion. If the platinum profile is the recipient of the shared resources, it would be able to admit extra user sessions. In this case the system would serve more users (sessions) at the peak tariff rate, hence generating more revenue. On the other hand, when the gold profile receives the excess resources, the QoS received by its active flows would be improved due to the availability of more resources.

The DQBCM fosters fairness amongst users of multi-service networks. Users of the platinum profile benefit greatly from their ability to meet varying tariff rates by using services at guaranteed QoS levels. On the lower end the silver profile users, who would generally be on a restricted budget, are able to receive services from the network due to appropriate resource allocations to the silver profile. The gold profile achieves a balance between steady tariff rates and changing QoS levels.

The DQBCM system achieves congestion control by performing admission control on all session establishment attempts. The admission control strategy is based on limiting the maximum number of active sessions in each profile to a sustainable value. In the platinum profile, the admission control influences the pricing of resources, thus by combining admission control and usage-based pricing as in the DQBCM, the network would control usage of network resources amongst users. Users would only remain on the network when they can afford the existing network prices; when some users drop-off from the network, new users get a chance to use the network.

The use of the number of active sessions to abstract the congestion level in the network would simplify the configuration of network equipment to achieve the desired admission control limits. The network operator would choose the value of $N$ that supports the desired network performance for each profile. The accuracy of this approach depends on the assumption that all flows in a profile receive an equal allocation of bandwidth and other resources.

The DQBCM can be extended to other access networks such as the conventional Internet access via ISPs where service level agreements are enforced. However, the effectiveness of this pricing scheme may only be proven through extensive market analysis and deployment on a commercial system. The authentication and profile management system can be used in WLAN environments when Internet access services are to be controlled. In this case, access control can be achieved using the authentication and authorisation scheme developed in this project. The media access gateway functionality can be implemented in a conventional residential gateway device. This would facilitate the deployment of large networks with remotely located media access gateways that are controlled using one network access controller.

## 6.2 Future Work

Since the success of the Dynamic QoS-Based Charging Model depends on user acceptance, it is recommended that trials on communication networks where services are offered to users should be done to reveal the user perception and the performance of the scheme. The users of the DQBCM network should be given comprehensive details of the pricing scheme, including QoS and tariff characteristics of each profile, the effects and benefits of using particular profiles for different applications. This will enable users to make right profile choices.

In the architectural design of the DQBCM scheme, it was assumed that all flows in a profile are allocated equal amounts of QoS resources. In actual communication networks, this would not be the case, thus further investigation is needed to support network systems in which different flows in a profile are allocated varying amounts of network resources.

The QoS performance results from the evaluation framework were based on network throughput only; however, other network performance characteristics (e.g., packet delay and jitter) can be used to evaluate the performance of the profiles of the pricing scheme.

The dependence of the pricing function on the admission control function for the platinum profile would make the tariff rates to follow the value of $I_p$ in real-time. This would cause a high network overhead in the system, hence the pricing function should be modified to advertise new tariff rates at predefined intervals of $I_p$. This will reduce the rate at which new charging detail record (CDR) sessions would be initiated and terminated. This considers that a slight change in the value of $I_p$ would not have a considerable effect on the level of QoS received by the active flows.

On 3G networks, an estimation of network prices for the platinum profile can be presented to the user by displaying a bar code similar to the typical indication remaining battery strength on a mobile terminal. The estimation can be averaged from the trend in network prices over a specific period (e.g. 5 to 30 minutes). This would enable the platinum user to make informed choices on when to use the network. For the gold profile, the estimated network congestion state can be relayed to the user in a similar manner.

# Bibliography

[1] M. Falkner, M. Devetsikiotis, and I. Lambadaris, "An Overview of Pricing Concepts for Broadband IP Networks," *IEEE Communication Surveys*, vol. 3, pp. 2–13, Second Quarter 2000.

[2] N. Blefari-Melazzi, D. D. Sorte, and G. Reali, "Usage-based Pricing Law to Charge IP Network Services with Performance Guarantees," *IEEE International Conference on Communications*, vol. 25, pp. 2652–2656, Apr. 2002.

[3] D. J. Songhurst, *Charging Communication Networks - From Theory to Practice.* Elsevier Science B. V., first ed., 1999.

[4] M. Koutsopoulou, A. Kaloxylos, A. Alonistioti, L. Merakos, and K. Kawamura, "Charging, Accounting and Billing Management Schemes in Mobile Telecommunication Networks and the Internet," *IEEE Communication Surveys and Tutorials*, vol. 6, pp. 50–58, First Quarter 2004.

[5] J. Cushnie, D. Hutchison, and H. Oliver, "Evolution of Charging and Billing Models for GSM and Future Mobile Internet Services," tech. rep., Hewlet Packard, HP Laboratories Bristol, HPL-IRI-2000-4, July 2000.

[6] S. Dixit, Y. Guo, and Z. Antoniou, "Resource Management and Quality of Service in Third-Generation Wireless Networks," *IEEE Communications Magazine*, vol. 39, pp. 125–133, Feb. 2001.

[7] N. Baghaei and R. Hunt, "Review of Quality of Service Performance in Wireless LANs and 3G Multimedia Application Services," *Computer Communications*, vol. 27, pp. 1684–1692, November 2004.

[8] T. T. Ahonen, *m-Profits - Making Money from 3G Services.* John Wiley and Sons, 2002.

[9] A. J. O'Donnell and H. Sethu, "Congestion control, differentiated services, and efficient capacity management through a novel pricing strategy," *Computer Communications*, vol. 26, pp. 1457–1469, 2003.

[10] D. Grossman, "New Terminilogy and Clarifications for Diffserv," *IETF RFC 3260*, Apr. 2002.

[11] L. Bos and S. Leroy, "Towards an All-IP-Based UMTS System Architecture," *IEEE Network*, vol. 15, pp. 36–45, Jan/Feb 2001.

[12] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," *RFC 2475*, Dec. 1998.

[13] M. Karsten and et. al, "Charging for Packet-Switched Network Communications - Motivation and Overview," *Computer Communications*, vol. 23, pp. 290–302, February 2000.

[14] S. Yaipairoj and F. C. Harmantzis, "Dynamic Pricing with "Alternatives" for Mobile Networks," *Proc. IEEE WCNC*, pp. 671–676, March 2004.

[15] M. Peirce and D. O'Mahony, "Flexible Real-time Payment Methods for Mobile Communications," *IEEE Personal Communications*, vol. 6, pp. 44–55, December 1999.

[16] S. Shin, M. B. H. Weiss, and H. Correa, "A Progressive Analysis of Internet Market: From best effort to quality of service," *Telecommunications Policy*, vol. 5-6, pp. 363–389, June-July 2004.

[17] A. Odlyzko, "Paris Metro Pricing for the Internet," *Proc. ACM Conference on Electronic Commerce*, pp. 140–147, 1999.

[18] T. Li, Y. Iraqi, and R. Boutaba, "Pricing and Admission Control for QoS-enabled Internet," *Computer Networks*, vol. 46, pp. 87–110, Sep. 2004.

[19] S. Shenker. D. Clark, D. Estrin, and S. Herzog, "Pricing in Computer Networks: Reshaping the Research Agenda," *ACM Computer Communications Review*, pp. 19–43, Apr. 1996.

[20] R. Cocchi, D. Estrin, S. Shenker, and L. Zhang, "A Study of Priority Pricing in Multiple Service Class Networks," *ACM, Communications architecture and protocols*, pp. 123–130, 1991.

[21] R. Blake, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview," *IETF RFC 1633*, June 1994.

[22] R. Braden, L. Zand, S. Berson, S. Herzog, and S. Jamin, "Resources Reservation Protocol (RSVP)," *IETF RFC 2205*, Sept. 1997.

[23] 3GPP, "Vocabulary for 3GPP Specifications (Release 5)," *Tech. Ref. 3GPP TR 21.905 V5.8.0*, Sep 2003.

[24] S. Malomsoky, S. Racz, and S. Nadas, "Connection admission control in UMTS radio access networks," *Computer Communications*, vol. 26, pp. 2011–2023, Nov. 2003.

[25] 3GPP, "Technical Specification Group services and System Aspects; Telecommunication management; Charging management; Packet Switched (PS) domain charging (Release 6)," *Tech. Spec. 3GPP TS 32.251 V6.2.0*, Mar. 2005.

[26] 3GPP, "Technical Specification Group Services and System Aspects; Quality of Service (QoS) concept and architecture (Release 6)," *Tech. Spec. 3GPP TS 23.107 V6.3.0*, June 2005.

[27] 3GPP, "Technical Specification Group Core Network; End-to-end Quality of Service (QoS) signalling flows (Release 6)," *Tech. Spec. 3GPP TS 29.208 V6.3.0*, Mar. 2005.

[28] 3GPP, "Technical Specification Group Core Network; Policy Control Over Go interface (Release 6)," *Tech. Spec. 3GPP TS 29.207 V6.2.0*, Dec. 2004.

[29] 3GPP, "Technical Specification Group services and System Aspects; Telecommunication management; Charging management; Diameter charging applications (Release 6)," *Tech. Spec. TS 32.299 V6.2.0*, Mar. 2005.

[30] H. Hakala, L. Mattila, J. Koskinen, M. Sutra, and J. Loughney, "Diameter Credit-Control Application," *IETF RFC 4006*, August 2005.

[31] P. Brenner, "A Technical Tutorial on the IEEE 802.11 Protocol," tech. rep., Breesecom Wireless Communications, 1997.

[32] J. D. Vriendt, P. Laine, C. Lerouge, and X. Xu, "Mobile Network Evolution: A revolution on the move," *IEEE Communications magazine*, vol. 40, pp. 104–111, April 2002.

[33] B. Hubert and et. al, *Linux Advanced routing and Traffic Control HOWTO*, July 2002.

[34] L. Bulej, "Htb.init - Script to set HTB traffic control." Internet - http://freshmeat.net/projects/htb.init, 2004.

[35] M. Lin, H. Luo, and L. F. Chang, "A Linux-based EGPRS real-time test bed software for wireless QoS and Differentiated Service studies," *IEEE-ICC*, vol. 25, pp. 1039–1044, April 2002.

[36] M. Dereva, "Htb Linux queueing discpline manual - User guide." Internet - http://luxik.cdi.cz/ devik/qos/htb, 2002.

[37] R. Sandilands, "Network Traffic Generator." Internet - http://www.sourceforge.net/projects/traffic, Dec. 2002.

[38] E. A. Harrington, "Voice/Data Integration Using Circuit Switched Networks," *IEEE Transactions on Communication*, vol. 28, pp. 781–793, June 1980.

[39] 3GPP, "Technical Specification Group Services and System Aspects; Telecommunication management; Charging management; Charging architecture and principles (Release 6)," *Tech. Spec. 3GPP TS 32.240 V6.1.0*, Mar. 2005.

[40] J. Gozdecki, A. Jajszcyk, and R. Stankiewicz, "Quality of Service Terminology in IP Networks," *IEEE Communications Magazine*, vol. 41, pp. 153–159, March 2003.

[41] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, "A Framework for QoS-based Routing in the Internet," *IETF RFC 2386*, Aug. 1998.

[42] Y. Guo, Z. Antoniou, and S. Dixit, “IP Transport in 3G Radio Access Networks: An MPLS-based Approach,” *IEEE WCNC*. vol. 3, pp. 11–17, Mar. 2002.

[43] A. Viswanathan, N. Feldman, Z. Wang, and R. Callon, “Evolution of Multiprotocol Label Switching,” *IEEE Communications Magazine*, vol. 36, pp. 165–173, May 1998.

# Appendix A

# Background Information on Billing

## A.1 Terms Used in Billing

### A.1.1 Accounting

Accounting is the process of apportioning charges between the home environment, serving network and the user [4, 23].

### A.1.2 Charging

Charging is a function within the telecommunications network whereby information related to a chargeable event is collected, formatted and transferred in order to make it possible to determine usage for which the charged party may be billed [4, 23, 25]. The different terms in the definition of charging are:

A *function* is a collection of well defined procedures for doing a given task [23].

A *Chargeable event* is an event that requires reporting to the billing function so that the user's account can be updated accordingly [23].

*Charged party* refers to the subscriber, operator or other provider who is liable for paying the bill [23, 25].

### A.1.3 Billing

Billing refers to the functions where charging data records (CDRs) generated by the charging functions are transformed into bills requiring payment [23, 25].

### A.1.4 Pricing

Pricing is the process of determining the price for a service unit from both costs and market analysis, and the tariff model to be adopted (e.g., flat-rate or usage based) [3]. In pricing, a monetary value is associated with a measurable quantity of the service.

## A.2 Billing in Mobile and Fixed Networks

The convergence of the Internet and the telecommunications world has enabled the independent application/service providers to out-source the process of controlling and charging for the access and use of their applications to the network operators [4]. Fig. A.1 illustrates the location of application and content servers with respect to a 3G network.

Fig. A.2 illustrates the functions involved in the process of charging for the use of the network bearer resources, while Fig. A.3 is an illustration of the steps undertaken in charging for content delivery using a prepaid mobile system. In Fig. A.2, the relationship between various functions of a billing system, i.e. pricing, accounting, charging and billing are shown. Information in Fig. A.3 shows that user approval is needed prior to content delivery, and the content and network bearer charges are inseparable from the users' view.

Fixed-line networks have traditionally offered voice telephony as the dominant service for many decades. Voice telephony is mainly charged according to the duration of the call (i.e., usage-based charging) and the distance between the calling and the called parties. In some countries (e.g., the U.S.A) local telephone calls incur flat-rate charges [5]. The use of duration as an accounting metric in charging voice telephony services is influenced by circuit switching that is used in the routing of telephone calls. In circuit switching, a fixed amount of network

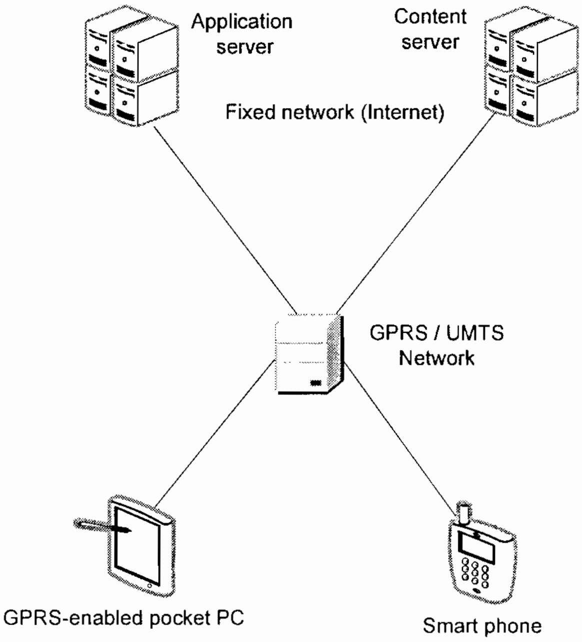


Figure A.1: General view of content access on mobile networks

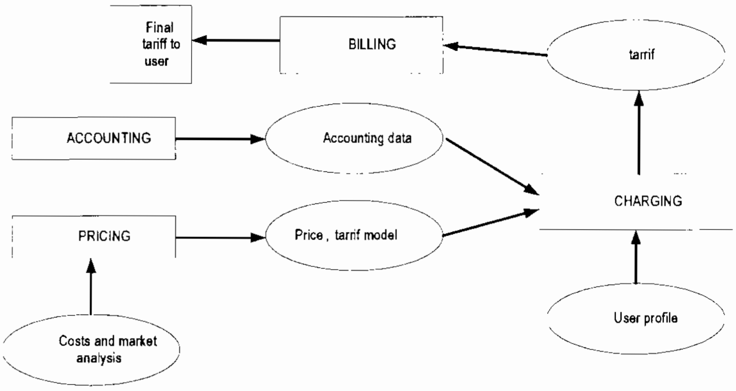


Figure A.2: Charging for the bearer service (Adopted from D.J. Songhurst [3])

resources is allocated to a connection, and the resources are held for the duration of the call [38].

In mobile networks, voice telephony services are generally charged on a time and usage basis [5]. Charging information is generated by the Mobile Switching Centre (MSC) for 1G and 2G networks and the UMTS Mobile Switching Centres

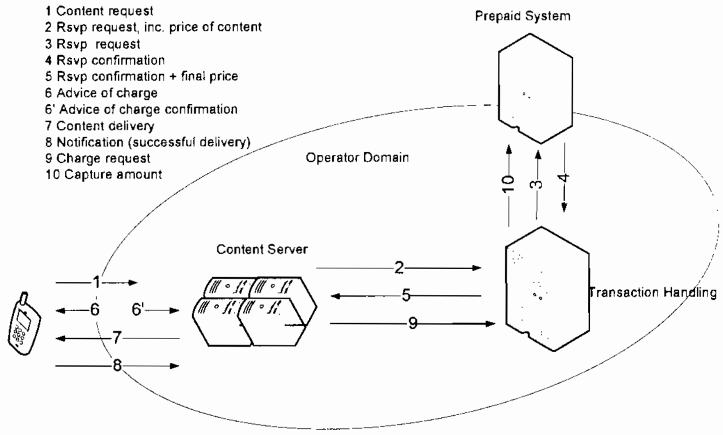


Figure A.3: Charging for content and application services (Adopted from a Siemens magazine on Communications)

(UMSC) for UMTS networks. Mobile networks offer data and multimedia services. Since the introduction of packet switched data services, charging practices in mobile networks have greatly changed. The use of the duration of service as the only accounting metric was insufficient for packet data services. Since traffic from some Internet applications is bursty in nature, the level of network resource usage is not constant; hence it was necessary to apply volume of traffic accounting and QoS based schemes for charging of mobile network services.

## A.3 Charging Requirements

The goals of charging of commercial telecommunication services are: to generate revenue and enable network operators and service providers to operate profitably and to provide incentives for users to appropriately control their traffic demands, which results in efficient network operation [3]. These goals are achieved by addressing the various charging requirements. The requirements can be classified as user requirements, operator requirements and system and network operation requirements [13].

### A.3.1 User Requirements

#### Predictability of Charges

Users want to know before hand the cost of the service [3]. If an exact *a priori* specification of communication charges cannot be given, a rough estimate of the charges should be possible. If a higher price than previously announced has to be charged, the users' explicit approval is required [13], thus the users need to be informed about changing prices.

#### Convenience and Ease of Use

The charging scheme should provide the user with crucial information about the service in a manner that is easy to understand. Standard interfaces should be in place to enable the users to control their communication costs through the selection of appropriate contracts and tariffs. Finally, users prefer receiving a comprehensive and detailed bill that explains important components of the aggregate charge [3, 4, 13].

#### Traceability and Accuracy of Charging

Users need to know the cost of each session and application they used. Connection charges should indicate the choice of service, tariffs and the measured usage parameters (e.g., QoS) for different sessions [13].

### A.3.2 Network Operator and Service Provider Requirements

#### Revenue Generation

Revenue generation is achieved through subscription and usage charges [3]. Subscription charges lead to predictable revenue, whereas usage-based charges are influenced by the variable network costs and other factors.

#### Technical Feasibility

The simplicity of the charging approach ensures that the implementation would be done with low input. Since usage-based charging imposes several overheads due to the required accounting and the processing of the accounting information into user bills, the exchange of information between network nodes for the purpose of storage and processing must be minimised [13].

#### Flexibility

The charging scheme should be adjustable to meet the needs of all types of customers, to work efficiently with user applications, and to serve as a basis for charges that are defined at the service level [3].

#### Support for a Variety of Business Models

Different operators may use different charging schemes; hence the usage parameters for charging should be standardised in order to produce a flexible scheme that will support inter-operability between multiple domains. The charging architecture should accommodate various pricing models (e.g., flat-rate, usage-based and QoS-based) to fulfil innovative business models in addition to traditional ones. One of the most recent requirements is the support of both prepaid (online) and postpaid (off-line) charging mechanisms [29].

### A.3.3 System Requirements

#### Stability and Reliability of the Service

Whenever a guarantee of QoS is offered to the user, the network is required to meet the agreement throughout the communication session. In case of QoS degradation or service disruption, an appropriate refund mechanism should be applied [13]. The charging mechanism in place should help in improving the probability of network service availability even during periods of peak demand. Once a flow is admitted, the reliability of the service measured in terms of the QoS guarantee should be noticeable.

**Flexibility**

This is required so as to enable the users to specify their willingness to pay; in some scenarios it is either the senders or the receivers who are expected to pay for the service. The system requires explicit approval from the paying user before the delivery of billable services.

## A.4 Evaluation Criteria for Pricing Schemes

Different pricing schemes can be evaluated using the criteria given below [1]:

- Compliance with existing technologies - Pricing schemes that are compatible with existing technologies (e.g. ATM and IP) are easier to adapt than those requiring significant changes to the underlying network.
- Measurement requirements for billing and accounting - this relates to the implementation complexity of the pricing schemes.
- Support for congestion control and traffic management - it rates the applicability of the pricing scheme in congestion control by exploiting the price-sensitivity of users, where it induces them to limit their traffic during congested periods.
- Support for individual QoS guarantees - it analyses whether the scheme supports QoS guarantees for individual users.
- Degree of network efficiency - this refers to the expected utilisation levels of the network. High utilisation levels are beneficial to network operators since income levels will be high (i.e. only when usage-based charging is used). To the user, high utilisation levels may lead to denial of service. If QoS guarantee mechanisms are not used (e.g., best-effort transport), heavily utilised networks will provide poor service.
- Impact on social fairness - this criterion indicates whether the pricing scheme prevents some users from accessing the network purely as a result of their inability to pay high tariff rates (e.g., during congested periods).

- Pricing time frame - this is the period over which the tariff rates are likely to change. Short time frames are desirable when pricing is used for congestion control. However, the complexity of implementation and the users' reaction to rapidly changing rates must be considered.

## A.5 3G Network Services

The 3G system is often associated with multimedia services such as VoIP, videophone, video games etc [8]. The delivery of services over mobile networks is initiated via packet data protocol (PDP) context sessions that are triggered either by the end-users or systems belonging to the independent service providers and the Public Land Mobile Network operator. The three main PDP context types are mobile-to-mobile, mobile-to-PDN, and PDN-to-mobile; they are illustrated in Fig. A.4.

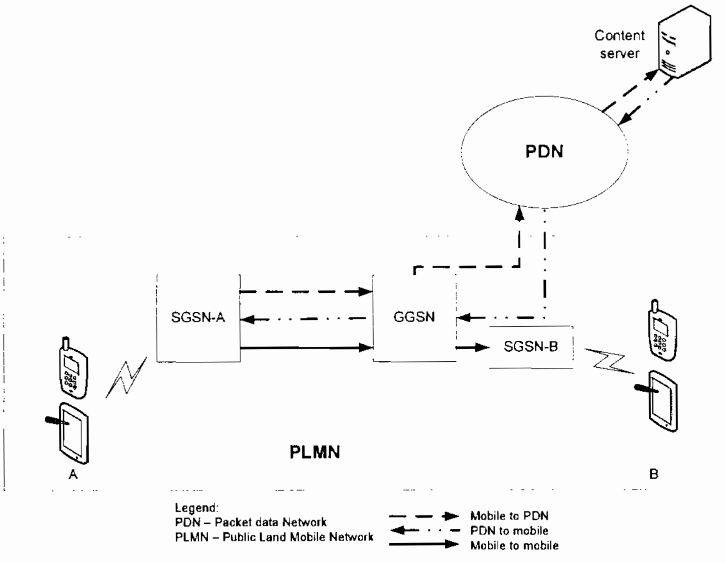


Figure A.4: PDP context types

Mobile-to-mobile PDP contexts involve the exchange of information like video and audio clips, or video conversation and VoIP sessions between the mobile users. Charging for these services is done by mobile operators whose networks are used for service delivery [25].

A mobile-to-PDN PDP context is initiated by a mobile unit that intents to exchange data with a content/application server across the packet data network. The PDP context would be triggered in SGSN-A and the GGSN, shown in fig.A.4, and it will be flagged as a MU triggered context.

A PDN-to-mobile PDP context involves a packet switched domain context from a content server to a MU. The GGSN receiving the packets generates the G-CDR and the SGSN serving the MU generates the S-CDR.

Each of the PDP context types illustrated in Fig. A.4 can be associated with a given set of 3G services. As mentioned earlier, the mobile-to-mobile PDP context could involve services like MMS (e.g., audio and video clips and images), video conversation and VoIP. The mobile-to-PDN PDP context includes services like email, e-commerce, web browsing and file transfer (FTP). The PDN-to-mobile PDP context could include services like news updates on traffic, sports and weather and advertisements that are enriched with multimedia content.

## A.6 Charging of Mobile Network Services

The charges incurred for the use of mobile network services are comprised of elements corresponding to the different providers of the end-to-end service. The network bearer service charges are directly linked to the network operators' CAPEX and OPEX. When introducing a new service to the market, the network operators and service providers use an introductory price that is as close as possible to the optimal price. The optimal price will ensure that users quickly pick up the service and continue to use it at the same price for long periods, and at the same time it enables the network operators and service providers to generate the required levels of revenue [8].

When the service delivered involves access to content or applications from independent providers, the charge levied will include the cost of the content plus the cost of transmitting the information across the network. The tariff rates used for the network bearer service is influenced by many factors (e.g., the QoS level achieved by the PDP context). The rates are generally proportional to the achieved QoS level, which implies that higher tariff rates are used when the achieved QoS is high.

Due to interconnection between networks owned by different operators, the Internet and telecommunications media providers, the delivery of mobile services often involves traversing networks other than that owned by the home operator. In such a scenario, every involved network operator or service provider would charge for the use of their portion of the network.

### A.6.1 Accounting and Charging for the Network Bearer Resource Usage

Accounting for network resource usage enables network operators to determine the level of resource consumption by individual users. In billing management, the accounting function associates metered values with characteristics applicable to a session. A session in this context refers to a period over which charging parameters (e.g., prices for network resources) of the system remain unchanged. In packet switched mobile networks (e.g., GPRS and UMTS) the SGSN and the GGSN perform real-time monitoring of resource usage in order to detect the relevant chargeable events. The duration of PDP context sessions and the volume of data transferred during the sessions are part of the details captured by the metering modules [25]. In off-line charging, the collected charging information is processed in order to generate charging detail record (CDR) files that are then transferred to the Billing Domain. A CDR is opened at PDP context activation, and the volume of data transferred during the context is counted separately in the up-link and down-link directions. When a change of charging conditions occurs, the volume count is added to the CDR and a new count is started. The generation of a new CDR can be triggered by the following conditions: QoS change, tariff period change and CDR closure. CDR closure is triggered either by the end of a PDP context within the GSNs or by the operation and maintenance (OM) procedures, e.g. data volume limit, time limit, maximum number of charging condition changes (QoS/tariff change), etc [39].

In off-line charging, resource usage accounting is reported to the Billing Domain (BD) after the resource usage has occurred, thus the charging information does not affect the real-time delivery of the service. In online charging, a subscriber account located in an Online Charging System, is queried prior to granting permission to use the requested network resources, thus the charging information can affect the real-time delivery of the service and hence a direct coordination of the charging mechanism with the control of the network resource usage is required [39].

There are two main ways of accounting for the use of mobile network resources; the duration of a PDP context session and the volume of data transferred during a session. These accounting metrics give a precise measure of the level of resource usage. Some network services can be associated to the method of accounting for resource usage (e.g. a voice call of a certain duration, the transport of data of a given volume, or the submission of multimedia of a certain size).

When using duration as an accounting metric, the determination of the charges to be incurred for the use of network resources requires accurate correlation between the timing periods and changes affecting other charging characteristics (e.g. the QoS level or the tariff period). The charge calculation can be represented by $aT + c_1$, where $T$ is the duration of resources usage, while $a$ and $c_1$ are charging coefficients related to the resource consumption. From the users viewpoint, this charging method is suitable for real-time services with delay guarantees (e.g., interactive speech or video) [3]. Volume of traffic accounting requires correlation between the amount of data transferred and the charging characteristics that are bound to change during service delivery. The charges incurred can be represented by $bV + c_2$, where $V$ is the volume of data transferred; $b$ and $c_2$ are charging coefficients. When charging is done by duration and volume, the charging formula would be of the form $aT + bV + c$. When flat-rate charging is used, only subscription charges would be levied, i.e. there would be no connection charges.

# Appendix B

# Quality of Service

## B.1 Common QoS Terms for IP Networks

In the telecommunications context, the term *service* refers to the capability to exchange information through a telecommunications medium, provided to a customer by a service provider. In the IP context, service is defined by the ITU as, "a service provided by the service plane to an end user and which utilises the IP transfer capabilities and associated control and management functions, for the delivery of users information specified by the service level agreements" [40]. *Quality* is an assessment of whether the service satisfies the users expectations. The QoS aspect that depends on the network performance level is known as *intrinsic* QoS. QoS from the users' point of view is generally known as *perceived* QoS.

### B.1.1 QoS Parameters

There are four primary QoS parameters, i.e. reliability, delay, jitter and bandwidth [7].

- *Bandwidth* - is the bit-rate available for transferring user data.
- *Delay* - is the delay experienced by packets traversing the network; it can occur at a particular node or it can be end-to-end.
- *Jitter* - this is the inter-packet delay variations.

- *Reliability* - ratio of the packets delivered to the ones sent.

### B.1.2 Class of service

Class of service (CoS) is a set of parameters available with a specific service [40]. The IETF defines CoS as "the definitions of the semantics and parameters of a specific type of QoS" [41]. In 3G networks, a QoS class identifies a bearer service, which is associated with a set of bearer service characteristics [28].

### B.1.3 Service Level Agreement

Service level agreement (SLA) is a term used in DiffServ to describe the service contract that specifies the forwarding service a customer should receive [10]. In general, SLA is a negotiated agreement between the service provider and the customer regarding the levels of service characteristics and the associated set of metrics[1] [40].

## B.2 End-to-end QoS in 3G networks

The GPRS core network is generally suited for providing best effort service, hence it is able to support packet data services that do not require strict QoS guarantees (e.g., web browsing). Since newer services (e.g., real-time multimedia) require QoS guarantees, the 3rd generation partnership project (3GPP) and other standardisation bodies formulated ways of supporting QoS in mobile networks [6, 27, 28]. Although the IP technology lacks in-built QoS provisioning capabilities, its simplicity and advantages have led to the idea of an all-IP UMTS network [6, 11].

Due to interconnection with the Internet, effective provisioning of end-to-end QoS on 3G networks must involve the Internet. Fig. B.1 shows a basic layout of the UMTS network showing the flow of traffic from the Internet to a mobile user equipment. When implementing QoS provisioning, the aim is to allocate the available network resources to applications according to the demand of resources by

[1] ITU definition of SLA

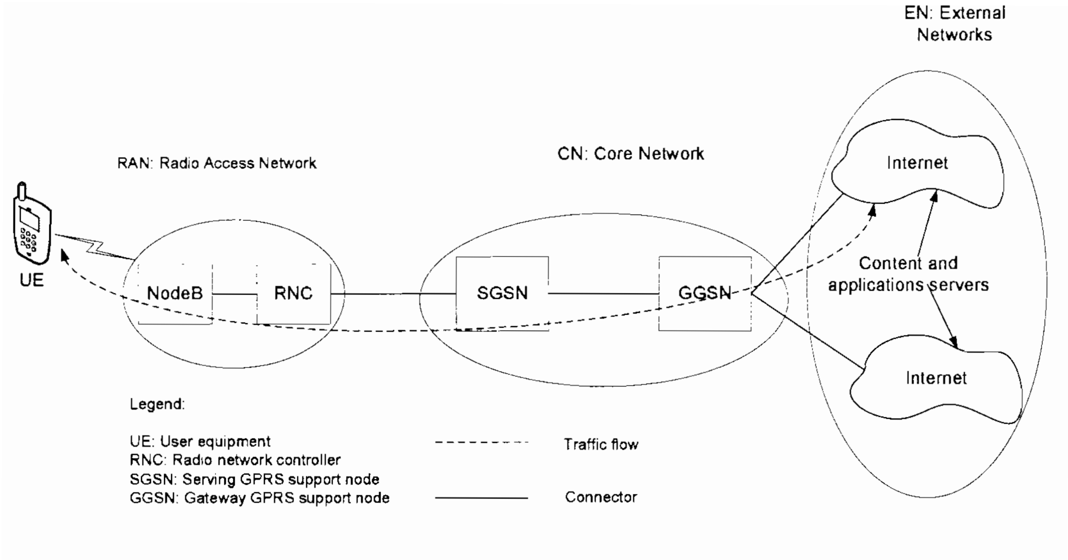

EN: External Networks
RAN: Radio Access Network
CN: Core Network
UE
NodeB
RNC
SGSN
GGSN
Internet
Content and applications servers
Internet
Legend:
UE: User equipment
RNC: Radio network controller
SGSN: Serving GPRS support node
GGSN: Gateway GPRS support node
Traffic flow
Connector

Figure B.1: Basic layout of the UMTS network, illustrating data flow from the Internet to a user equipment

those applications. QoS provisioning schemes manage the available resources on the network using different approaches. In IP networks, three schemes have been widely researched; the integrated services (IntServ) architecture [21], which relies on end-to-end resource reservations, the multi-protocol label switching (MPLS) [42, 43] and the differentiated services (DiffServ) architecture [12].

The DiffServ architecture, which has found wider acceptance as a result of its simplicity and scalability to large networks [18, 9], offers a small number of service classes that are assigned different priorities. Packets from applications that belong to a given class of service are forwarded according specific Per-Hop Behaviours (PHB) in the core network.

### B.2.1 QoS Requirements for 3G Applications and Services

Four overriding UMTS QoS classes have been standardised by the 3GPP, i.e. the conversational, streaming, interactive and the background classes of service. These classes are mainly distinguished by their tolerance to packet delay and jitter.

Table B.1: QoS requirements of 3G applications

| APPLICATION | REQUIRED RELIABILITY | TOLERABLE DELAY | TOLERABLE JITTER | BANDWIDTH REQUIREMENT |
|---|---|---|---|---|
| E-mail & File transfer | Low | High | - | High |
| Web access | Low-medium | Medium | - | Medium |
| Control | Null | Low | - | - |
| Voice and Video telephony | Low-medium | Low | Low | Medium-high |
| Streaming Video and Audio | Low-medium | Low-medium | Low-medium | Medium-high |

### Conversational CoS

This class relates to real-time person-to-person communication such as videophone, audio conversation etc. The acceptable transfer delay limit is very strict [7, 26].

### Streaming CoS

This class involves applications where a user views non-real-time video or listens to non-real-time audio. It involves a one way transport scheme. The examples of applications in this class include: video-on-demand, live MPEG, web-radio, and multi-casts. The requirements of the streaming CoS include the preservation of time delay between packets (i.e. jitter). There are no requirements on low delay, which makes it possible to use buffering capabilities in the end-user equipment [7, 26].

### Interactive CoS

The interactive CoS applies to cases where the end-user (e.g., a human or a machine) is online and requesting data from remote equipment (e.g., a server). Examples include web-browsing, database retrieval, polling for statistics, and automated database enquiries. The round-trip delay and preservation of the payload content (i.e., low bit error rate) are the fundamental QoS characteristics of this class [7, 26].

### Background CoS

This applies to the case when an end-user, typically a computer, sends and receives files in the background. Example applications are E-mails, SMS, FTP downloads etc. The main QoS requirement is data integrity, thus the content should be transferred transparently (with low bit error rate) [7, 26].

## B.3 Evaluation Framework Design Requirements

### Media Access Gateway

The design requirements for the media access gateway are as follows:

- Hardware requirements:
  - A personal computer (PC)
  - Two Ethernet network interface cards, that are mounted on the PC
  - An access point ( a D-Link access point was used)
- Software requirements
  - Linux operating system software (Debian Linux was used in this project)
  - IP-tables - this is a software package that is used for configuring network firewalls on Linux
  - DHCP - this package provides dynamic host configuration of network interface cards of client computers, i.e. issuing of IP addresses
  - Bash - this is a shell scripting package that allows flexible interaction with the Linux system

The Linux kernel configuration requirements are as follows:

- Packet sockets
- Unix domain sockets
- TCP/IP networking
- Networking support
- Drivers for the two network cards
- IP advanced router - this enables the system to route packets between the AP and the AC
- DHCP support
- IP-tables support
- MAC address match support

### Network Access Controller

Hardware requirements

- a personal computer
  - two Ethernet network interface cards
- Software requirements
  - Linux operating system software (Debian Linux was used in this project)
  - IP-tables package
  - Bash
  - Postgresql - an Open Source database program
  - PHP - a web application development module
  - Apache - an Open Source web server

The Linux kernel configuration requirements are as follows:

- Packet sockets
- Unix domain sockets
- TCP/IP networking
- Networking support
- Drivers for the two network cards
- IP-conntrack support - this is a connection tracking module that is required for the admission control function
- IP advanced router - this enables the system to route packets between the AG and the external networks
- IP policy routing
- IP-tables support
- IP use of netfilter mark as routing Key
- IP fast network address translation or full NAT support
- IP use of TOS value as routing key
- Packet filtering support
- QoS support
- HTB packet scheduler

# Appendix C

# Source Code for System Procedures

## C.1 Mobile Unit Registration Process[1]

### C.1.1 MU Registration Process Source Code I - detect

*File name: detect*

### C.1.2 MU Registration Process Source Code II - enable

*File name: enable*

### C.1.3 MU Registration Process Source Code III - Socket client

*File name: socket_client.c*

### C.1.4 MU Registration Process Source Code IV - gettimeofday

*File name: gettimeofday.c*

[1] The source code is available on the accompanying CDROM. The file names for each process are specified in the respective sections in this appendix.

### C.1.5 MU Registration Process Source Code V - Socket server and AA database

*File name: socket_server.c*

### C.1.6 MU Registration Process Source Code VI - clid.h

*File name: clid.h*

## C.2 Connection Admission Control - Interactive CoS, Platinum profile

*File name: trackingInt_plat*

## C.3 User Interface to Profile Change

### C.3.1 User Interface to Profile Change - Page 1

*File name: bgprofmgt.php*

### C.3.2 User Interface to Profile Change - Page 2

*File name: bgprof.php*

### C.3.3 User Interface to Profile Change - Page 3

*File name: bgprof_upt.php*

## C.4 Profile Management Function

### C.4.1 Profile Management Function - Profile Trigger Function

*File name: tcclient.c*

### C.4.2 Profile Management Function - Profile Change Server Process

*File name: tcserver.c*

### C.4.3 Profile Management Function - tclid.h

*File name: tclid.h*

## C.5 Traffic Control

### C.5.1 Traffic Control function - Packet classification and marking

*File name: tcp_ control*

## C.6 Profile Change Enforcement

- Changing from the platinum to the silver profile in the interactive CoS

```
IP: 10.129.8.20 new profile: intsilver old profile: intplatinum
iptables -D int -s 10.129.8.20 -j intplatinum
iptables -I int -s 10.129.8.20 -j intsilver
iptables -D int -d 10.129.8.20 -j intplatinum
iptables -I int -d 10.129.8.20 -j intsilver
```

- Changing from the gold to the silver profile in the interactive CoS

```
IP: 10.129.8.5 new profile: intsilver old profile: intgold
iptables -D int -s 10.129.8.5 -j intgold
iptables -I int -s 10.129.8.5 -j intsilver
iptables -D int -d 10.129.8.5 -j intgold
iptables -I int -d 10.129.8.5 -j intsilver
```

## C.7 Online Survey on Pricing of Mobile Services

### C.7.1 Flat-rate charging

In the flat rate charging scheme, the user pays a standard subscription fee for a given period; commonly a month. The eventual use of the service would then be unlimited i.e. a subscriber who uses the service for, say, one hour a day would pay same subscription fee as one who uses it 24 hours a day. Varying subscription fee levels are levied according to the bandwidth (speed) of the user's link to the Internet

### C.7.2 Usage-Based Charging

In this scheme, the user incurs an extra charge every time he/she uses the service. The charges are calculated either per duration e.g. seconds, of usage or for the volume of data e.g. bytes, transfered for every session. There are cases where the charge levied might vary depending on various parameters, the common being the time of the day. daytime charges are generally higher than other times.

### C.7.3 Special-Tariff Charging

This category would be viewed as advanced and complex. When Quality of Service considerations are put into the picture, then issues like service level agreements (SLA) come into play. This is where thresholds on the level of network performance are set. The network operator would agree to meet the set conditions and the subscriber would be willing to pay some agreed charge for using the service under those conditions. Multimedia applications like video conferencing, Voice over IP and streaming services fall in the category of Internet services that would require steady level of network performance.

### C.7.4 Network Charge Ratings

This section gives ratings of the charges endorsed to various mobile Internet services. The base of the ratings is the charge incurred in making a one minute call from a mobile phone to another mobile phone on the same mobile providers network. This is equated to the cost of transferring 1 Megabytes of data on the mobile network. For instance, a rating of *standard* means the cost of transferring 1Mb of data at 144Kbps should be equal to the cost of a one minute mobile-to-mobile call on the same operator network.

- *Low* - Less than the charge for a single minute call on same operator network.
- *Standard* - Same as the charge for a minute's call on same operator network.
- *Moderate* - More than twice but less than five times the cost of a minute's call on the same operator network.
- *High* - More than five times the cost of a minute's call on the same operator network.

# Appendix D

# Accompanying CDROM

The accompanying CDROM is located on the inside back cover of this document. The contents of the CDROM are as follows:

- A soft copy of this thesis in PDF format
- The lyx source files used in generating this document
- The source code files of the evaluation framework
- Relevant publications used during the research of this thesis.