AN ADAPTIVE FILTERING ARCHITECTURE FOR QOS-BASED HETEROGENEOUS NETWORKS

by

FLAVIO DE ANGELIS

A dissertation submitted to the Graduate Faculty in Engineering
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy, The City University of New York.

2005

UMI Number: 3187426

UMI Microform 3187426

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This manuscript has been read and accepted for the
Graduate Faculty in Engineering in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Prof. Ibrahim Habib

Date — Chair of Examining Committee

Dean Mumtaz Kassir

Date — Executive Officer

Prof. Leonid Roytman (CUNY, CCNY EE Dept.)

Prof. Georgios Ellinas (CUNY, CCNY EE Dept.)

Prof. Mohamed Zharan (CUNY, CCNY EE Dept.)

Prof. Andrew Campbell (Columbia University EE Dept.)

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

AN ADAPTIVE FILTERING ARCHITECTURE
FOR QOS-BASED HETEROGENEOUS NETWORKS

by

Flavio De Angelis

Adviser: Professor Ibrahim Habib

The convergence of mobility, Internet, and multimedia services is, finally, seen possible through Third Generation (3G) networks. This thesis is motivated by the need to address two key issues that this evolution will arise: i) the difficulty of guaranteeing the quality of service (QoS) requirements of multimedia services, while maximizing the efficiency of the IP network assets, and ii) the difficulty of maintaining these requirements over the time-varying bandwidth of the wireless channel.

The aim of this thesis is to propose and analyze an adaptive filtering architecture that solves these problems, by enabling an "intelligent" interaction between the multimedia applications and the network. The proposed architecture can be implemented as a part of a "middleware" layer that interfaces the applications with the underlying network infrastructure. It also calls for the deployment of adaptive filters at the output ports of each node of the network. Each filter is capable of selecting both the coding rate and format of the multimedia traffic based upon the dynamic conditions of the underlying heterogeneous network, e.g., core and radio access networks' links. In the proposed

solution, a service provider can offer two distinct grades of service (GoS) to which users may subscribe: Premium, or Economy. These grades of services offer the same multimedia contents but at different levels of qualities. A quality level is defined by a set of parameters such as call blocking and dropping rates, coding rate and format of the media. The proposed architecture will then maximize the utilization of the networks' assets (i.e., links' capacities, buffers, switching ports) and avoid congestion while maintaining the QoS requirements (e.g. bit error rate, packet transfer delay, percentages of packets lost and delayed) of each type of media (e.g., video, voice and data). Some of the value-added benefits include increased number of admitted users, reduced percentage of lost and delayed packets, lower packet transfer delays, reduced call dropping and blocking probabilities, higher throughput-based and power-based cell loading. These improvements indicate better utilization of the network's assets that translates into additional revenues for the provider without incurring additional capital expenditures.

Acknowledgments

First, I would like to express my deepest gratitude to my adviser Professor Ibrahim Habib for his continuous support, encouragement and invaluable academic advice. I sincerely appreciate his tremendous efforts, countless hours of detailed technical discussions and close supervision. I have been greatly influenced by his methodology in the field of scientific research. His technical expertise and knowledge have been instrumental in conducting my research and completing my thesis.

I would like also to thank the Faculty of the Department of Electrical Engineering at The City College of New York for their support. Special thanks go to the Examining Committee for their help, efforts, and support with the thesis. I also like to express my appreciation for my fellow graduate students who have made the years of my stay at the University pleasant and enjoyable.

Next, special thanks are due to my family - my father Bruno, my mother Franca and my sister Ombretta - for their support all the time. They were always there and gave me the love and support when I needed them the most.

Finally, my deepest thanks are to my girlfriend Noemi for helping me a lot in so many ways. Without her, this work would not have been possible.

To Noemi

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In today's competitive telecommunication market, service providers for *Third Generation* (3G) networks such as *Universal Mobile Telecommunication Systems* (UMTS) are under increasing pressure to maximize profitability by decreasing operating costs and increasing revenues, simultaneously.

One way to decrease operating costs is by migrating 3G networks from the current architecture, which is implemented as two separate infrastructures dedicated to each media and type of service, to a more cost-efficient *Internet Protocol* (IP) packet-switched infrastructure. In this future architecture, radio access channels, implemented with *Wideband Code Division Multiple Access* (W-CDMA) technology [1, 10] instead of being dedicated to voice and video will be shared by voice, video and data.

One way to increase revenues is by enhancing the number of subscribers that such heterogeneous IP-based network (e.g., core and radio access networks' links) will be able to support while offering customized multimedia services tailored to meet subscribers' demands.

These two concerns share the need for *Quality of Service* (QoS) to ensure that the requirements of the users are satisfied.

1.1 Motivation

This thesis is motivated by the need to address two key challenges that the deployment of such heterogeneous network arise:

i) How to guarantee the QoS requirements [2, 23, 52], (e.g. *Bit Error Rate* (BER), packet transfer delay, percentage of lost or delayed packets) of multimedia services delivered to the mobile users while increasing the utilization of the IP network assets [3] (e.g. bandwidth, buffers, or ports);

ii) How to maintain efficiently these QoS requirements over the limited [4, 5] and time-varying bandwidth of the wireless channel without overloading the radio access nodes.

For the first issue, the *Internet Engineering Task Force* (IETF) has proposed distinct approaches [6, 7, 8] to reserve network resources in order to provide guaranteed services. This concept is also called a "connection oriented IP network", e.g. *Multi Protocol Label Switching* (MPLS). The key principle is that network resources are allocated dynamically in response to traffic demands or QoS requirements from the applications. Hence, the network guarantees the committed QoS parameters as long as the application does not violate its declared traffic parameters (e.g., coding rate and format). However, in this approach, network resources are assigned without any form of "intelligent control" of the applications based upon network specifics (e.g., maximizing throughput, managing traffic-loads, or maximizing revenues by increasing assets' utilizations). Therefore, as traffic load increases, availability of the allocated resources may temporarily fluctuate due to the multiplexing effect resulting in prolonged episodes of packets lost or delayed.

This can determine a degradation of the multimedia applications and a reduction in the effective utilization of the IP resources.

For the second issue, a resource reservation is even more challenging because of the unpredictable and time-varying characteristics (e.g. shadowing, attenuation, fading, and interference) [9] of the radio channel and the users' mobility. The W-CDMA technology offers an advantage that could be used in order to effectively utilize the limited bandwidth radio channel. This is the option of assigning variable air interface bit-rates through the *Variable Spreading Factor* (VSF) method [10], or the *Multi Code* (MC) method [11]. However, new IP enhanced services such as video and audio streaming, web browsing or voice-over IP consume a significant amount of the network resources. If the radio channel does not have enough capacity to deliver the amount of traffic to be downloaded to the mobile users, packets would be delayed or discarded (the buffer size at the link layer is finite) and resources would be wasted. This scenario could result in re-transmissions, leading to congestion and further degradation in the network throughput [12], as well as violating the QoS requirements.

1.2 Thesis Aims

The aim of this thesis is to propose an approach that treats the previous issues with a different philosophy in order to overcome the abovementioned drawbacks. Our proposal is an "intelligent" interaction between the applications and the network resources that will allow service providers to utilize effectively the network resources, and increase their revenues without violating QoS parameters. The achievement of these objectives requires the fulfillment of two different conditions. On one hand, applications must be able to "declare", first, a "set of acceptable QoS parameters" and, then, accept some form of

"flexible" guarantees instead of hard guarantees. On other hand, the network must be able to "intelligently" control the rate of the content delivered from the application according to the available resources [13, 14, 15, 16].

Our proposed solution allows mobile users to download traffic within a range of acceptable quality levels instead of just a fixed, static one, which could lead to block the request of the application if the network resources were not available. Quality levels are defined by a set of parameters such as coding rate and format of the media streams delivered. The network is able to control adaptively the downloaded traffic by using a software module, the proposed *Multimedia Adaptation Filter* (MAF) that is installed at the output ports of each node of the 3G network. These filters can manipulate the quality levels of the media streams by operating in an "adaptive middleware" layer interfacing the applications with the network.

The amount and type of traffic downloaded to the mobile users is tuned, or adapted, according to the different conditions in the downstream links [17] of the heterogeneous networks (e.g. 3G core and radio access networks) and the users' *Service Level Agreements* (SLA). In our solution, the service provider can offer two distinct *Grades of Service* (GoS) to which users may subscribe: *Premium* (P), or *Economy* (E). These grades of services offer the same contents but at different levels of qualities. Subscribers to the Premium service pay more for their connections to receive a higher range of quality levels.

The filtering operation is performed via a two steps approach: first, the coding rate of the media stream (e.g., AMR, W-AMR, and MPEG-4) is scaled without changing the format of the media. Secondly, if necessary, the coding rate of the media is further

reduced via transcoding, i.e. changing the format of the media [18, 19, 20]. To illustrate, consider the case of voice packets generated from a *Wideband Adaptive Multi-rate* (W-AMR) speech codec. A possible action of the filter is to scale down the generated bit-rate from 24 kbps to 16 kbps. Alternatively, transcoding can be applied to the same flow changing the coding format from W-AMR into *Adaptive Multi-rate* (AMR) format (the AMR codec rates range from 12 kbps to 4 kbps). Similarly, in the case of video packets that are coded using *MPEG-4* video codec, the filter may scale the codec rate from 512 kbps down to 64 kbps. Alternatively, transcoding can be applied to the same media changing the coding format from *MPEG-4* to *H.263* video codec (video codec rate is changed from 64 kbps to 32 kbps).

To enable the actions of the filters implemented in both core and radio access networks, it is important to design adaptive resource management algorithms. The QoS filtering algorithms that this thesis proposes and analyzes are an example of such adaptive algorithms [21, 22]. Our proposed solution is based upon utilizing *Genetic Algorithms* (GAs) to solve multi-objective optimization functions that adaptively select the coding rate and format for each type of traffic flow subject to different constraints in the heterogeneous networks such as links' capacity, buffers state, total power, and number of available spreading codes. In the core network, a traffic flow is defined as an abstract of aggregate traffic of the same type of media (e.g., audio, video, or data) that is downloaded to a group of users with the same grade of service (Premium or Economy). In the access network, a traffic flow represents the single media downloaded to a distinct mobile user.

To summarize, the aim of this thesis is to construct multi-objective functions that control the filters located in both the core and the radio access network nodes. The action performed by the filters is essentially to meet the following requirements:

(i) Avoid congestion on the end-to-end path including the radio link;

(ii) Provide preferential treatment to Premium class subscribers over Economy class users;

(iii) Ensure that the QoS parameters of each traffic flow (e.g., bit error rate (BER), transfer delay, percentage of packets lost, and delayed) are maintained, regardless of the class of the user;

(iv) Maximize the utilization of network assets (e.g., links' capacity, buffers, ports, throughput-based and power-based cell loading);

(v) Minimize the power transmitted to each user in order to prolong the user equipment battery life;

(vi) Increase the revenues by accommodating a higher number of users and by offering customized services tailored to meet subscribers' demands.

1.3 Thesis Structure

The rest of the thesis is organized as follows. Chapter 2 presents the main components and features of the heterogeneous (core and radio access networks) 3G architecture. Next, a brief review of the current state-of-the-art of QoS in 3G networks is developed. This includes, first, a short introduction to the QoS terminology, principles and concepts, then, a description of both UMTS and IP QoS models. The chapter ends by introducing new ultimate solutions for QoS support, which are based on intelligent interactions between applications and network.

Following this, in Chapter 3, our proposed adaptive filtering architecture for QoS-based 3G networks and beyond is described. The major component of the architecture, the filter, is then detailed in Section 3.2. This includes a description of the main filtering parameters, and types of media formats that the filters can manipulate. A discussion about the deployment issues of the filters with transcoding capabilities is also added. The chapter ends by showing how the filter is expected to treat the traffic flows directed to Premium and Economy class users.

Chapter 4 defines the QoS filtering algorithm for controlling the filters in the core network nodes. The chapter begins by describing the proposed resource manager scheme at a core network node. This is followed by an example of possible assignments of the filtering levels, coding rates and coding formats for Premium and Economy class traffic flows. The major contribution of this chapter is the definition of the multi-objective function, which is used to control the actions of the proposed filter and the class-based queuing system at the output port of a core network node. Following this, an overview of techniques to solve an optimization problem is described and, the proposed solution, Genetic Algorithms, is detailed. This chapter ends by providing several experiments and simulation results that show the significant improvement of the proposed approach in the system performance.

Chapter 5 defines the QoS filtering algorithm for controlling the filters in the radio access network nodes. The chapter begins by providing a description of the radio access network with filtering capabilities followed by an example of possible assignments of the filtering parameters and traffic models for Premium and Economy class traffic flows. Next, two different options for allocating radio resources in W-CDMA networks are

detailed: variable spreading code and multi code. The main contribution of this chapter is the definition of the optimization function, which is used to control the filter and the rate schedulers of the W-CDMA network. To test the proposed algorithm in terms of increasing the revenues of the service provider, a revenue model and pricing mechanisms for Premium and Economy class users are also defined. Finally, simulation experiments and results are provided by emulating a microcellular urban environment.

The thesis ends with Chapter 6, which presents a summary of the thesis, thesis major contributions and future work.

Chapter 2

Quality of Service in Heterogeneous 3G Networks and Beyond

This chapter, first, presents the main components of the architectural model of the 3G-UMTS network; then, provides an overview of QoS terminology, principles and concepts. Next, current techniques for managing QoS in IP-based 3G networks are briefly discussed. A more comprehensive overview of architectures supporting QoS is given in references [23] and [24]. The chapter ends by introducing new challenging solutions to provide QoS support in 3G networks and beyond.

2.1 3G-UMTS Networks

Internet protocol is a universal network layer protocol for the Internet, and is becoming a promising universal network-layer protocol over all wireless systems as well. Network operators are working to define a Third Generation network evolution toward all IP architecture [25], in which the architecture is migrated from a double (circuit and packet switched) backbone to a unique packet switched infrastructure.

Fig. 2.1 represents the architectural model of the 3G network that is referenced in this thesis to enable the delivery of IP multimedia content. As shown in the figure, edge media content servers are connected to the IP *Core Network* (CN) of a *Public Land Mobile Network* (PLMN) that is made up of routers interconnected via fiber links. In

particular, in a 3G-UMTS environment [26, 27] the edge routers of the core network are network nodes called *Gateway GPRS Support Node* (GGSN) and *Serving GPRS Support Node* (SGSN). They are in charge of several tasks among whose is the routing functionality. The SGSN manages users' mobility and supports packet communication towards the access network, while GGSN maintains the connections towards the external (high-speed) switched data networks such as internet. As shown in the figure, the UMTS CN transfers packets from the GGSN to the SGSN using the Gn interface. The *Home Location Register* (HLR) contains permanent data of the subscribers and is also responsible of Mobility Management related procedures. Edge servers are located in a *Virtual Private Network* (VPN) that belongs to the PLMN provider.

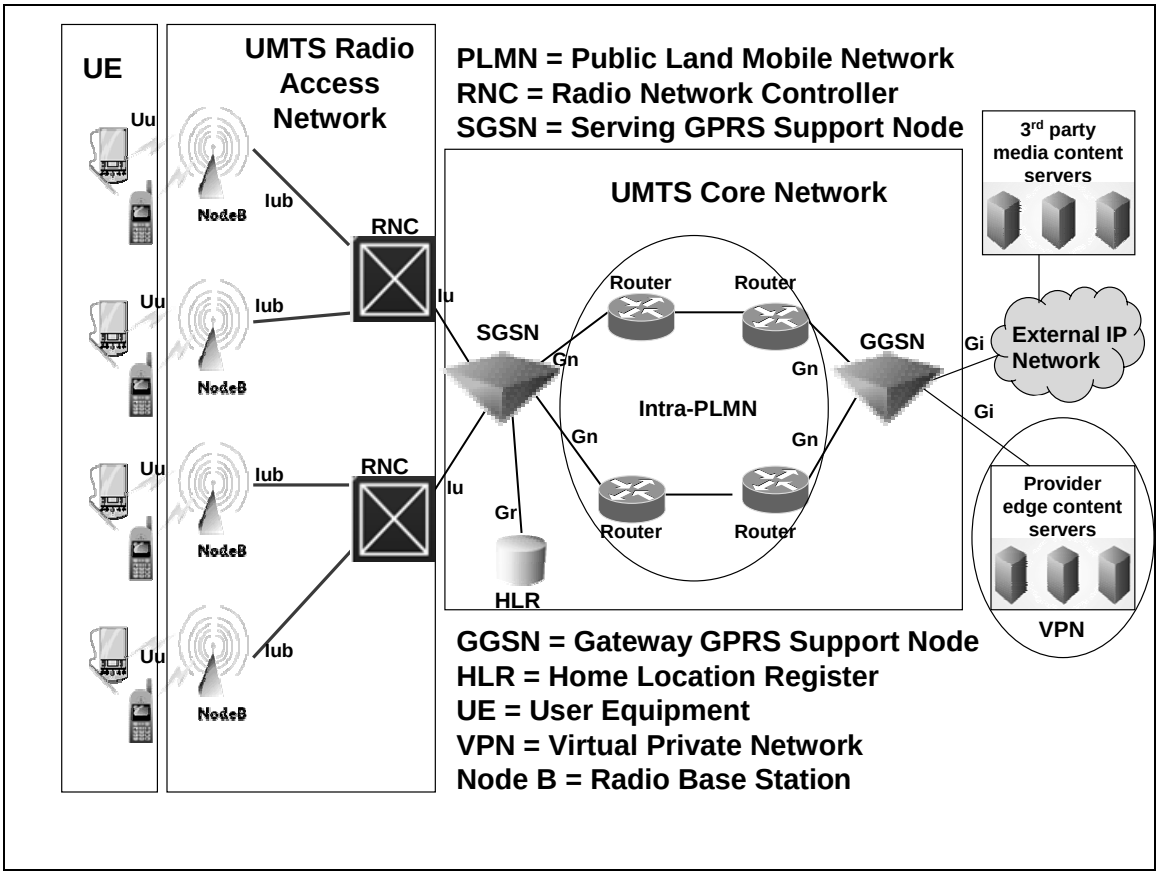

Fig. 2.1 Third-Generation systems architecture

Different types of contents are downloaded from 3rd party content media servers to the edge servers via a *Content Delivery Network* (CDN) that overlays the external IP network. The Edge server and the external IP networks are connected to the GGSN via Gi interface. From the edge servers, contents are downloaded via fiber links to the *UMTS Terrestrial Radio Access Network* (UTRAN) that is made up of *Radio Network Controllers* (RNC) and base stations (Node-B). As depicted in the figure, the UTRAN is connected to the SGSN through Iu PS interface. The RNC carries out all the functions that are related with the allocation of radio resources and QoS management functionalities such as admission control, congestion control, physical channel rate and power allocation. The *User Equipment* (UE) or *Terminal Equipment* (TE) represents the 3G network terminal that can be connected to one or more base stations. The UE downloads multimedia content from the UTRAN via the Uu, and Iub interfaces. In addition of being able to place voice calls, the mobile user or *Mobile Station* (MS) has the opportunity in this network, to establish communication links with the Internet to exchange multimedia contents, downloads e-mails, browse the web, etc.

2.2 QoS in IP-Based 3G UMTS Networks

This section, first presents a general definition of the concept of quality of service, then, introduces this concept in the 3G UMTS networks and the IP environment.

2.2.1 Overview of QoS

Generally speaking, QoS is the ability of a network element (e.g. an application, a host or a router) to provide some level of assurance for consistent multimedia services delivery. This level of assurance is based upon concepts of *bandwidth* (e.g. system level

bit-rate, and application level bit-rate), *timeliness* (e.g. packet transfer delay, jitter, and response time), *reliability* (e.g. bit error rate, packet loss, percentage of packets delayed), *service availability* (e.g. call blocking and dropping rate), *perceived quality* and *cost* (e.g. per-use, or per-unit). Basically, QoS parameters can be classified according two main categories: network-based and application-based parameters. Table 2.1 summarizes the main network-based QoS parameters, whereas Table 2.2 shows the main application-based QoS parameters.

Table 2.1 Examples of possible network-based QoS parameters

CATEGORY	PARAMETER	DESCRIPTION
Timeliness	**End-to-End transfer delay**	**Time taken for a packet to be transmitted from source to destination**
	Jitter	**Variation in delay or response time**
	Response time	**Round-trip time from request transmission to reply receipt**
Bandwidth	**System level rate**	**Bandwidth required or available in bits per second**
Reliability	**BER**	**Bit error rate**
	Percentage of packets lost	**Proportional of data that does not arrive as sent**
	Percentage of packets delayed	**Proportional of data that does not arrive as expected**

Table 2.2 Examples of possible application-based QoS parameters

CATEGORY	PARAMETER	DESCRIPTION
Quality levels	**Coding rate**	**Bandwidth required from the application**
	Coding format	**Compression method of the media for the application**
Availability	**Call dropping rate**	**Percentage of calls dropped**
	Call blocking rate	**Percentage of calls blocked**
Cost	**Per-usage cost**	**Cost per unit time or per unit of data**
	Flat-rate cost	**Cost per subscription**

A user that initiates a multimedia call expects the application to provide a predefined level of quality [28]. Application QoS parameters can be usually specified in terms of the coding rate and format of the media (e.g. audio, video, and data) downloaded. From a

user perspective, coding rates and formats of the media are mapped to a higher set of perceived QoS parameters such as picture details, picture color accuracy, frame width, video rate, video smoothness, voice quality (wideband or narrowband voice signal), text summary and so on. Users pay a certain amount of money (cost per unit or per subscription) to receive both a targeted quality level and service availability (call dropping and blocking rate).

Based upon the application characteristics (e.g., peak data rate, burst length), application requirements are mapped to lower network level QoS parameters. These parameters can be specified in terms of network performance requirements such as latency, bandwidth, jitter, and packet loss rate that must be guaranteed during a multimedia call establishment. Before any multimedia stream is transmitted, network QoS parameters must be communicated to the resource management entities of all involved system components. Then, the required resources must be admitted, reserved and allocated along the path between the sender and the receiver. These basic steps are performed during multimedia call establishment.

2.2.2 UMTS QoS Model and Layered Architecture

Most of the services that are featured in 3G-UMTS can be divided into four main QoS classes primarily differentiated based on their ability to tolerate transfer delay and BER. As defined in [52], *Conversational* and *Streaming* classes preserve time relation between information entities of the stream and have a guaranteed rate from the network. They are suitable to carry real time traffic since they define an upper limit on transfer delay within their QoS profiles. The main difference between them is the maximum transfer delay value. Conversational traffic is subject to the strict human perception in a

voice over IP talk or video conferencing while real time Streaming traffic has slightly flexible delay requirements. Therefore, when strict guarantees are necessary, as in real time voice and video applications, using the conversational traffic class can be a good choice. On the contrary, the streaming class is slightly relaxed in terms of delays.

On the other hand, *Interactive* and *Background* classes are mainly intended to represent conventional Internet applications (e.g., web browsing, telnet, ftp, and email). Since, background class is designated for background traffic, e.g. email or file downloading the destination is not expecting data within a certain time. Therefore, this class is a best effort one that offers no qualitative or quantitative guarantees. Interactive class traffic essentially follows a request-response pattern and can only justly provide qualitative guarantees. Interactive traffic classes have higher priorities than background ones in terms of resource assignment to ensure responsiveness.

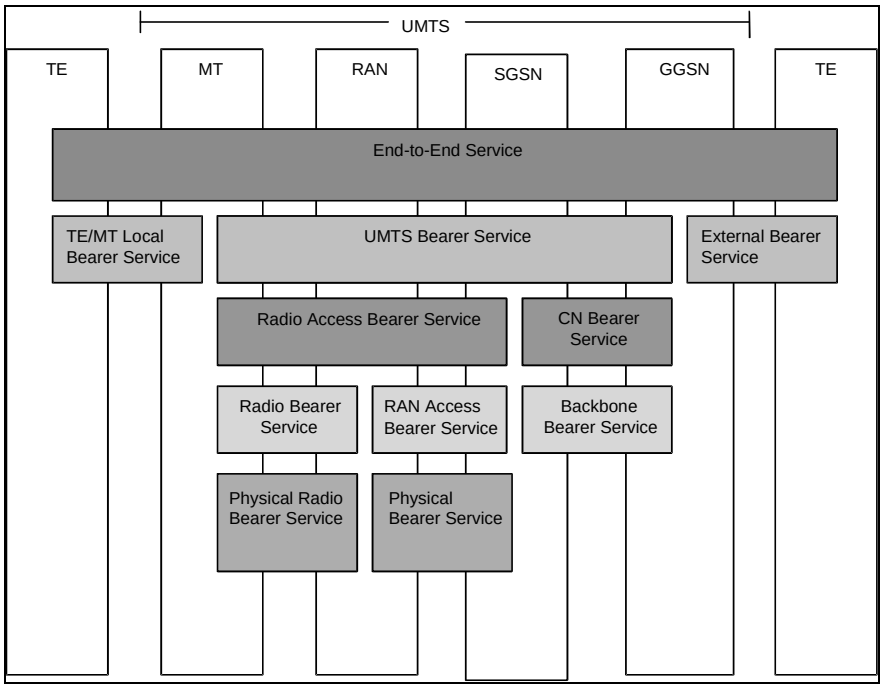

Fig. 2.2 UMTS end-to-end QoS layered architecture

The end-to-end QoS [2] in a 3G network is performed through the layered architecture illustrated in Fig. 2.2, where each bearer service is a combination of one or more other bearers in the lower layers. A *Bearer Service* (BS) defines all aspects between communicating end points to enable the provision of a contracted QoS. These aspects include control signaling, user plane transport and QoS management functionality. The different bearers are set up in conjunction with the *Packet Data Protocol* (PDP) context activation procedure. UMTS bearer service, offered by the UMTS operator, is in charge of the QoS management inside UMTS network, i.e., from mobile station to core network.

The UMTS bearer service, consisting of the radio access bearer and the core network bearer, realizes QoS in the UMTS network. The Radio access bearer service itself consists of a radio bearer service and an Iu-bearer service.

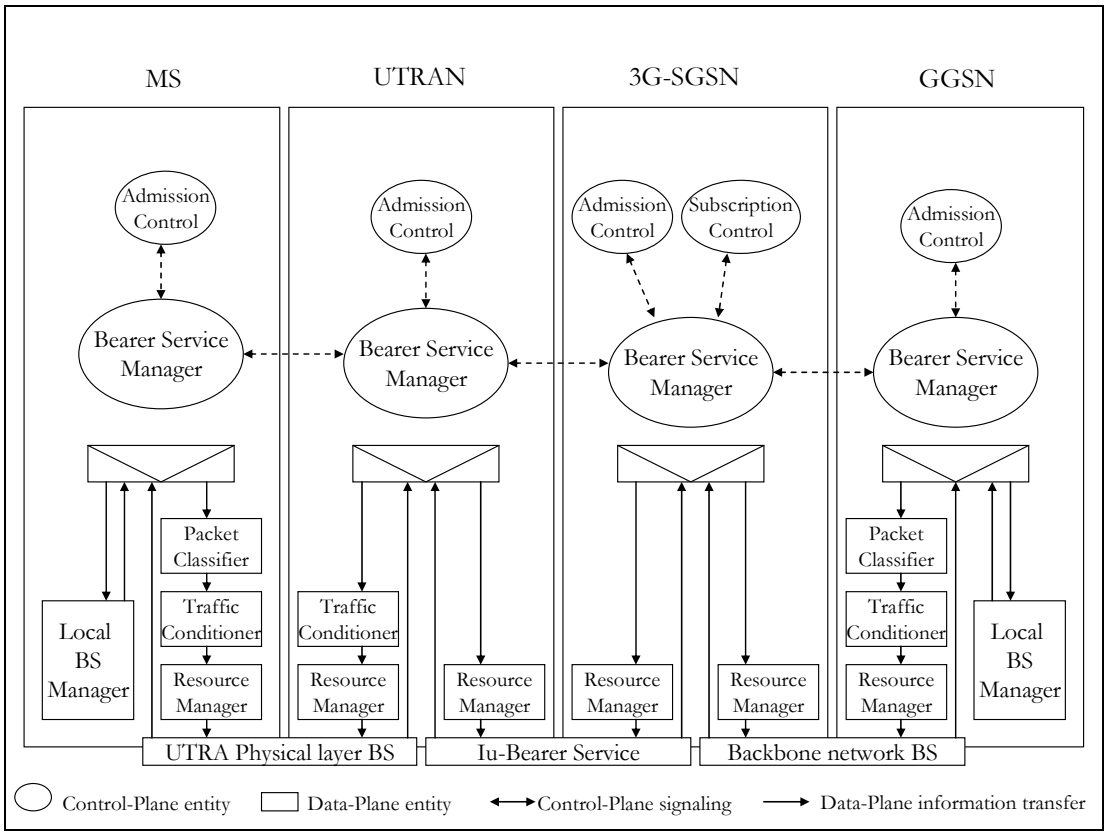

Fig. 2.3 QoS components in a UMTS end-to-end QoS architecture

The Radio Link Control's User-plane (RLC-U) layer between the RNC and the MS supports the radio bearer service. The Iu-bearer service provides transport services between RNC and SGSN. The core network bearer service provides transport services within the UMTS core network, e.g., between a SGSN and a GGSN. The various bearers can be realized in a functional architecture to provide end-to-end QoS for each PDP context [2]. An adaptation of architecture to realize this end-to-end support is shown in Fig. 2.3. The main control-plane and data-plane components of this architecture are discussed in the following.

a) *Admission Control* — This module maintains information about available and allocated resources, determines whether resources are available for each bearer service requests, and reserves resources if they are available. The admission control module in SGSN has the responsibility to finally accept or reject the PDP context activation and the requested QoS parameter values. GGSN and UTRAN verify whether they can locally support the bearers associated with the QoS profile in the PDP context.

b) *Bearer Service Manager* — A bearer service manager coordinates control plane signaling to establish, modify, and maintain the particular bearer service for which it is responsible. Each BS manager is typically a collection of multiple component managers that handle individual bearers, including UMTS, radio, Iu, and core network. A particular BS manager may perform QoS attribute translation. The UMTS BS manager in the MT provides functionality to translate UMTS QoS attributes into those supported by a local bearer and request local bearer services. The UMTS BS manager in the 3G-SGSN translates UMTS bearer attributes into RAB, Iu, and CN bearer attributes and requests the respective managers to provide the corresponding bearer services. Similarly, the UMTS

BS manager in the GGSN translates UMTS attributes into CN and external bearer attributes and requests the respective BS managers to provide those services. Each of the BS managers may consult the admission control entity before providing the bearer service. The BS manager in 3G-SGSN may also consult a subscription control entity.

c) *Resource Manager* – Each of the resource managers is responsible for managing access to resources in accordance with the particular bearer service. Thus, a resource manager provides support for QoS attributes required for a bearer service, and it may achieve this by such means as scheduling, bandwidth management, and power control.

d) *Traffic Conditioner* – The role of a traffic conditioner is to provide conformance of input traffic to the specification agreed in the bearer service. A traffic conditioner may achieve this by traffic shaping or traffic policing. In the MT, the traffic conditioner ensures uplink traffic conformance, whereas in the GGSN and UTRAN, it provides downlink traffic conformance.

e) *Packet Classifier* – The packet classifier in the MS assigns packets received from the local bearer service manager to the correct UMTS bearer based on such information as the DSCP, trans-port layer port numbers, security parameter index, etc. In GGSN, the packet classifier assigns packets received from the external bearer service manager to the appropriate UMTS bearer.

This is actually where the UMTS QoS model, described above, is in effect. However, recent releases of UMTS aim at providing user terminals with different IP-based services and extend the 3G network features with Internet Engineering Task Force protocols and QoS models that support wide range of multimedia services [29]. In the following section, these models are detailed.

2.2.3 Internet Protocol (IP) QoS Models

As originally designed, the Internet supports only "best-effort" delivery of packets across network links. In this context, traffic is processed as quickly as possible, but there is no guarantee as to timeliness or actual delivery. Only an additional higher level end-to-end protocol such as *Transmission Control Protocol* (TCP) provides end-to-end reliability by using congestion window and packets retransmission.

As demands for quality of service arise, a number of architectures have been proposed by various organizations in the communications industry in order to achieve QoS provisioning. In particular, the research effort in the area of IP QoS has been carried out by the IETF according to three main approaches: the *Integrated Services* (Int-Serv) model, the *Differentiated Services* (Diff-Serv) model and the MPLS model. The following sub-sections provide main characteristics and features of these models.

2.2.3.1 Integrated Services Model

Int-Serv [6] is a bandwidth reservation technique that builds virtual circuits across the IP-based networks. Bandwidth requests come from applications running in hosts. Once a bandwidth reservation is made, the bandwidth cannot be reassigned or preempted by another reservation or by other traffic. Bandwidth is reserved for each flow by using the *Resource Reservation Protocol* (RSVP). Network nodes along the path between sender and receiver must coordinate with one another to set up an RSVP path, and then remember state information about the flow. The Int-Serv model defines three classes of service:

a) *Guaranteed* – With bandwidth, bounded delay, and no loss guarantee.

b) *Controlled Load* – Approximating best effort service in a lightly loaded network

c) *Best-effort* – Similar to what internet currently provides under a variety of load conditions, from light to heavy.

However, the RSVP approach is now considered too unwieldy for the Internet, where millions of flows may exist across a router, but appropriate for smaller enterprise networks.

2.2.3.2 Differentiated Services Model

Diff-Serv [7] is a *Class of Service* (CoS) model that differentiates traffic by user, service requirements, and other criteria; then, it marks packets so that network nodes can provide different levels of service via priority queuing or bandwidth allocation, or by choosing dedicated routes for specific traffic flows. A policy management system controls service allocation. Diff-Serv takes a stateless approach that minimizes the need for nodes in the network to remember anything about flows. It is not as good at providing QoS as the stateful (Int-Serv) approach, but more practical to implement across the Internet. Network elements simply respond to the packets' markings without the need to negotiate paths or remember extensive state information for every flow. In addition, applications do not need to request a particular service level or provide advance notice about where traffic is going. Diff-Serv model is scalable because it handles aggregate flows and minimize the signaling, thus reducing the complexity of per-flow soft state at each node. The Diff-Serv service class is specified in the Diff-Serv field of each IP packets. Three types of Per Hop forwarding behaviors are defined

a) *Expedited Forwarding* (EF) – Provides minimal delay, jitter, and packet loss, and it guarantees the required bandwidth. Packets that violate traffic profile requirements are dropped.

b) *Assured Forwarding* (AF) – Classifies IP packets into four traffic classes and three levels of drop precedence. In case of congestion, high drop precedence packets are more likely to be dropped than low drop precedence packets.

c) *Best Effort* (BE) – Provides no guarantees.

2.2.3.3 Multi Protocol Label Switching Model

MPLS [8] provides connection-oriented services by building *Label Switched Paths* (LSPs) across networks and rapidly forwarding IP packets across the network through these paths. By labeling packets with an indicator of the LSP they are to traverse, it is possible to eliminate the overhead of inspecting packets at every network device along the way. MPLS adds labels to packets that indicate forwarding behavior, but packets travel across predefined circuits. MPLS is generally more sophisticated and complex than Diff-Serv, but provides better QoS capabilities. LSPs are similar to virtual circuits in ATM and frame relay networks, and *Traffic Engineering* (TE) approaches can be used to create LSP that delivers a required level of service. For example, a path can be created that provides high bandwidth and low delay for "premium" customers who are willing to pay for it. In another example, multiple paths can be defined between two endpoints to provide load balancing and backup service in the event of a line failure. LSPs are set up using signaling protocols, such as RSVP-TE or *Constraint Routed Labels Distribution Protocol* (CR-LDP).

2.3 Challenges in QoS

The key principle of all the QoS models described in the previous section is that network resources are allocated (statically or dynamically) in response to fixed demands

or QoS requirements from the applications. However, in case the resources to satisfy the demands from the applications cannot be observed the request for the connection is blocked. Similarly, in case the network dynamic traffic-loads dictate that only a small amount of resources is temporarily available, the network cannot continue to guarantee this level of quality that is pre-defined to be acceptable for the application. In these scenarios, multimedia applications would be degraded and network resources wasted.

A new ultimate solution for QoS support [30] is to allocate network resources with a form of "intelligent control" of the applications. This can be achieved by implementing an adaptation mechanism in a "middleware" layer interfacing the applications with the network. There are, basically, two different strategies for locating the adaptation mechanism on the end-to-end path from the source to the destination: 1) at the source, and 2) at the network nodes along the end-to-end path. In the following sub-sections, main characteristics and functionalities as well as pros and cons of the two strategies are briefly described.

2.3.1 Source Adaptation

In the first strategy, the control of the applications is performed by implementing the adaptation mechanism at the source. Examples of schemes that are utilizing this strategy can be found in the references [31, 32, 33, 34, 48]. This solution requires the periodic transmission from the receivers of quality reports on the traffic received. These reports include the measured bandwidth, bit error rate, packet transfer delay, jitter, percentage of packets lost and delayed. The source evaluates the reports in order to identify significant changes in the quality of the end-to-end path. If these changes are large enough to justify a change in the adaptation process, a new set of application-based parameters (e.g. coding

rate) are selected. In the case of improvement of the communication path, the source improves the quality levels (see Table 2.1) of the transmitted traffic. On the contrary, in the case of degradation of the communication path, the source reduces the quality levels down until the congestion dissolves.

This strategy can be easily deployed over a packet-switching network because the feedback mechanism is controlled by signaling messages at the transport layer (e.g., Real Time Transport Control Protocol (RTCP)). However, the disadvantages are not negligible.

First, both the source and the receiver have to be altered in order to implement the feedback-based adaptation mechanism. While this is a moderate task for the receiver, it is an extremely costly and timely process for legacy Internet servers. Content providers will be extremely reluctant to adopt this solution. Moreover, since the source must evaluate the feedback from many receivers, the amount of resources required substantially increases, particularly because of the volume of information that needs to be processed.

In addition, in cases where the end-to-end distance is fairly large, the delay introduced through the reporting packets can severely affect the performance of the solution. This can be more profound when wireless access is involved, where the frequency and the severity of changes in the wireless channel are high. If the report reaches the source too late to inform about a potential congestion problem, the source will already have flooded the congested network, failing to adapt effectively to the link quality fluctuations. An increase in the end-to-end delay quickly decreases the effectiveness of this solution since the source can no longer react in time. This allows the stream to flood the bottleneck link and to force a significant amount of dropped packets.

Moreover, putting the adaptation mechanism in the source precludes the utilization of this strategy in multicasting scenarios. In fact, in this case, the same multimedia stream with identical applications parameters will reach all receivers.

In conclusion, the simplicity of the source adaptation solution makes it efficient for simple cases with small variability, but cannot stand as a viable solution under extremely variable conditions or when multicasting is involved.

2.3.2 Filtering Adaptation

In the second strategy, the control of the applications is performed by implementing the adaptation mechanism within the end-to-end path, at the intermediate nodes identified as the most appropriate for performing the most effective adaptation. Examples of schemes that are utilizing this strategy can be found in the references [13, 14, 18, 35, 36, 37]. The intermediate nodes manipulate the multimedia streams coming from the source along the end-to-end path. The application-based parameters can be updated as the multimedia stream traverses the intermediate nodes.

There are several advantages that come with the adoption of this solution. First, the flexibility of locating the adaptation mechanism at different positions in the end-to-end path is an important advantage of the filtering solution over the other one. In fact, the filtering adaptation can be located at the most critical positions in the end-to-end path. On the contrary, the stationary location of the adaptation mechanism in the source leads to inefficiencies under some topological scenarios. For example, the source adaptation cannot be applied for multicasting scenarios. On the contrary, the filtering solution can be used to satisfy the requirements that are needed in these scenarios [15, 38].

Moreover, the filtering solution has the ability to react to changes with minimal delay and with near perfect accuracy. A monitoring mechanism can continuously scan the performance of the bottleneck link and immediately react to any significant changes. In particular, the filtering solution is able to adapt the multimedia streams efficiently, even in situations where unpredictable and highly variable wireless links are involved. When the adaptation point is strategically located close to the bottlenecked link, the reaction time is kept to minimum and it is unrelated to the distance from the source.

In addition, modifying intermediate nodes and adding simple filtering functionality is significantly less of a burden than changing a legacy content provider, which is requested in the source adaptation.

Naturally, the filtering solution also presents some disadvantages. First, having an intermediate intercepting the stream raises security issues. The third party operating the network nodes with the adaptation mechanism must be trusted by the receiver and the source. Moreover, the complexity of the network nodes adaptation is significantly higher than that of the other solution and requires nodes with powerful CPUs and memory.

Finally, the filtering adaptation approach demands adaptation awareness from some or many of the nodes on the end-to-end path, especially in multicasting scenarios.

In conclusion, the filtering adaptation approach performs exceptionally well in all situations and is ideal for both multicasting and wireless access scenarios. However, it typically induces increased complexity.

The next chapter presents our proposed filtering approach.

Chapter 3

An Adaptive Filtering Architecture

This chapter, first, proposes the development of an adaptive filtering architecture, which spans both core and radio access networks, to address the issues outlined in Chapter 1 and Chapter 2; then, it provides a description of the main filtering parameters and types of media formats that the filters can manipulate. The chapter ends by defining two distinct grades of services (Premium and Economy), and by showing how the filter is expected to preferentially treat the traffic flows directed to the Premium class subscribers versus Economy ones, based on the amount of money that users are willing to pay for their services.

3.1 QoS Filtering Architecture

In the proposed QoS-based filtering architecture, illustrated in Fig. 3.1, multimedia adaptation filters are installed at the output ports of every node of the 3G network. These filters control the bit rate values at which *traffic flows* are downloaded from the edge servers to mobile users according to the different conditions of the wireless network, or more generally according to the dynamic conditions of the downstream links of the heterogeneous network.

In the core network, a traffic flow represents the aggregate traffic of similar media streams (e.g. voice, audio, or data streams) that are downloaded to a group of users who

share some common attribute. This traffic flow, for example, can be characterized along a route by assigning a single LSP in each downstream link. On the contrary, in the access network nodes, a traffic flow represents a single media stream downloaded to a specific user. Each filter located in the output ports of the core network nodes manipulates in the same way all media streams contained in the same traffic flow. On the contrary, in the access network nodes, each filter would be controlling a single media stream directed to a specific user.

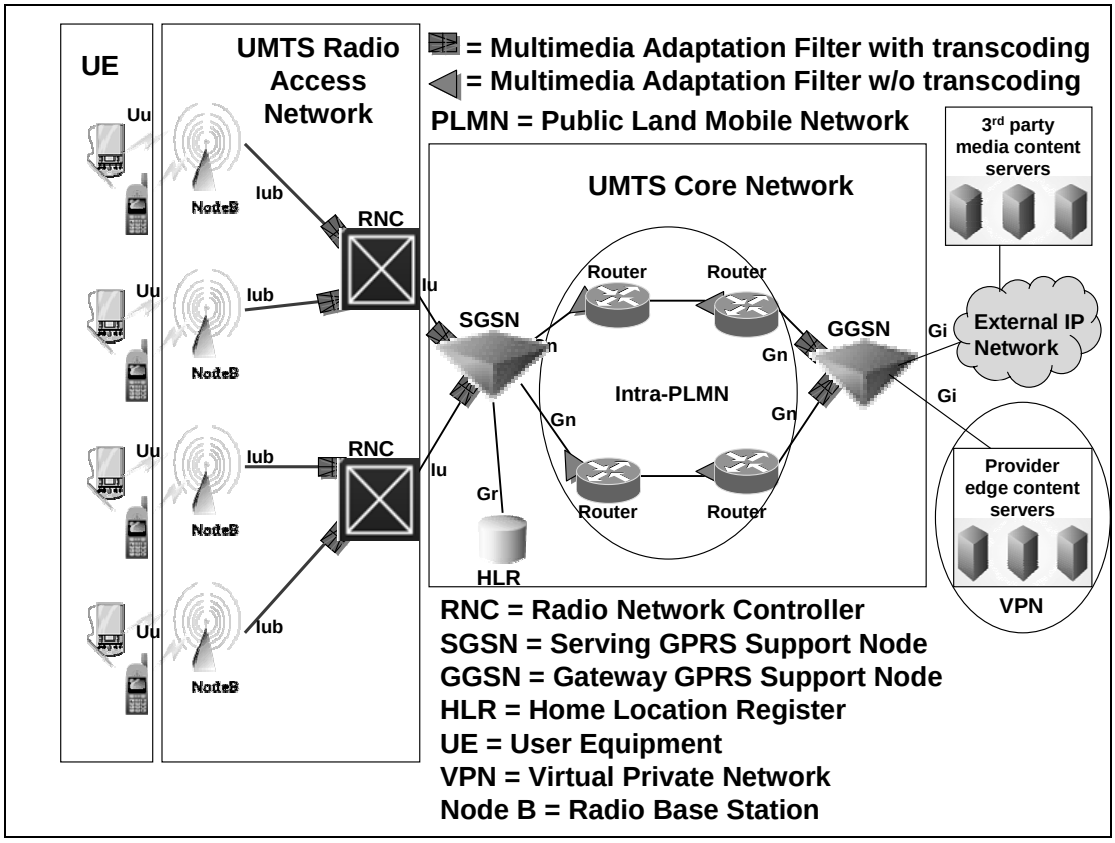

Fig. 3.1 The QoS filtering architecture

The multimedia adaptation filters are installed at the output ports of every node of the 3G network because every node along the end-to-end path is equally important in reacting to local congestion and delivering the QoS support to the application layer. Moreover, every node is important to utilize "efficiently" the available resources of the

core or radio access networks such as ports, links' capacity, buffers, total power, and spreading codes. In particular, the filters implemented in the nodes close to the edge multimedia storage servers (GGSN), in the wireless/wired interface (SGSN) and in the vicinity of the wireless access (RNCs) are also equipped with transcoding capabilities. The transcoding capability (see Section 3.2) allows the filter to further reduce the downloading bit rate of the multimedia content by changing the compression method of the media streams.

In the RNCs, transcoding is implemented to react to heavy congestion episodes that can happen due to the time-varying conditions of the radio channels allocated to mobile users. For example, shadowing or large scale path attenuation can determine a noticeable reduction in the usable wireless bandwidth and this usually lasts over the transmission of a large number of packets. During this time, since the amount of bandwidth variation is relatively large, transcoding can be applied to avoid prolonged delays in the download of the media contents.

In the SGSN node, transcoding can be activated to preserve the QoS requirements of the traffic flows when traffic loads in downstream links change due to the users' mobility in inter-RNC handovers. Inter-RNC handover takes place as the mobile device, during the same connection, moves from a cell covered by one RNC to an adjacent cell of a different RNC. The critical point is that, during the handoff, the connection can be blocked in the new cell if there is no sufficient spare capacity to support the connection. In this case, transcoding is adopted to reduce the probability of dropping the multimedia call by tuning the amount of traffic to be downloaded to the available bandwidth.

Finally, in the gateway node (GGSN), where traffic from external wired network is likely heterogeneous [39] with higher bit rates and different compression methods, transcoding is applied to adapt that amount and type of content to the requirements of the types of media supported by the UMTS network.

The multimedia adaptation filters are software modules that are installed at the output ports of the network nodes together with a control module and a data collection module.

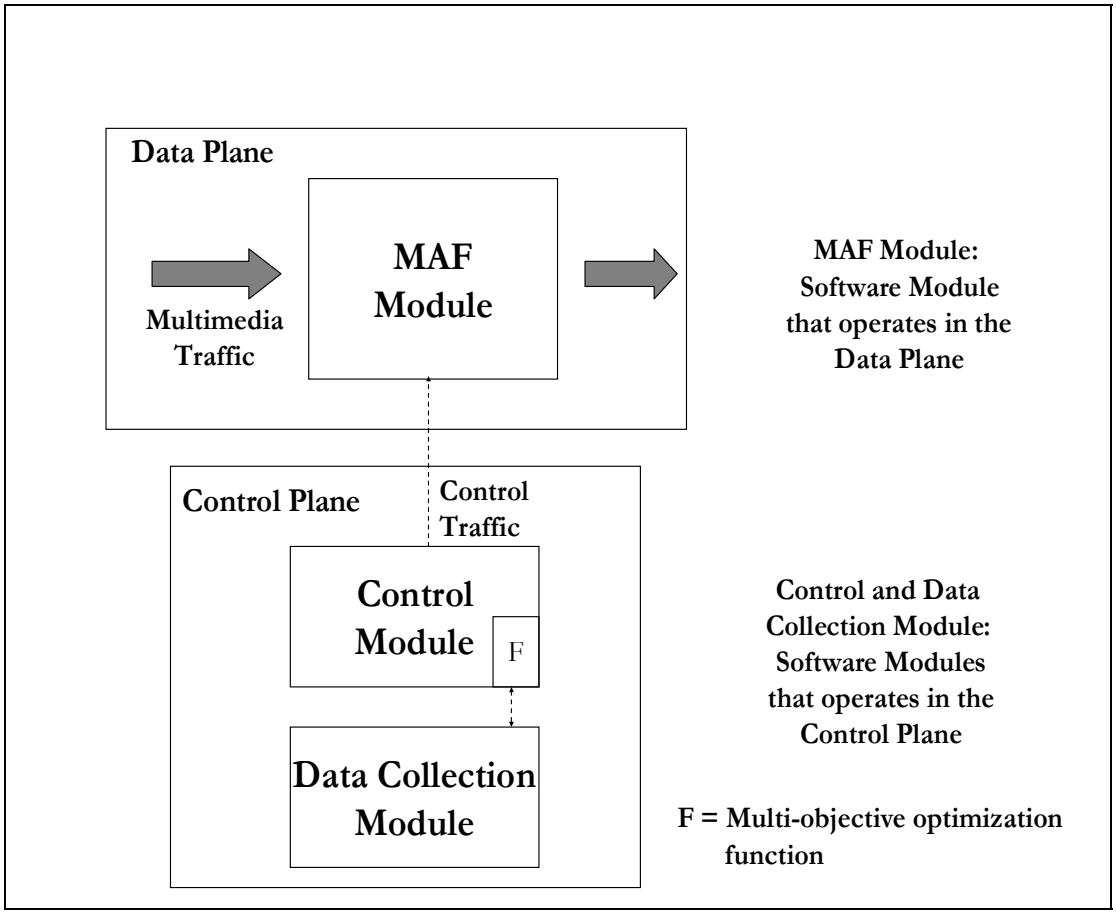

Fig. 3.2 Data and control plane

As shown in Fig. 3.2, the filter module operates in the data plane by manipulating the media content. The control and data collection modules are also software modules but they operate in the control plane. The control module contains a multi-objective

optimization function F that controls all the actions performed by the filter in the data plane. The data collection module gets necessary information required for the operation of the optimization function.

In essence, by implementing the proposed filters, the network operator is able to "intelligently" tune, or adapt, the amount of traffic downloaded that depends upon the available resources of the core or radio access networks. The following section provides a detailed description of the multimedia adaptation filter and its main filtering parameters.

3.2 Multimedia Adaptation Filters

A multimedia adaptation filter (MAF) is a software module that selects a specific *quality level* or *filtering level* q_j for the *j*th downloaded traffic flow. In a general case, each quality level is selected from a set of *M* values according to the filter granularity and processing power

$$q_j \in \{1, 2, 3, ..., M\}. \quad \textbf{(1)}$$

The maximum quality level "*M*" indicates the minimum filtering action taken by the filter, which means no filtering is done. The minimum level "*1*" indicates the maximum filtering action taken by the filter, which means a traffic flow with the minimum grade of quality.

The choice of a quality level establishes specific *coding rate,* CR_j, and *coding format,* CF_j, for the correspondent traffic flow.

$$CR_j \in \{CR^{[1]}, CR^{[2]}, CR^{[3]}, ..., CR^{[M]}\}, \quad \textbf{(2)}$$

$$CF_j \in \{CF^{[1]}, CF^{[2]}, CF^{[3]}, ..., CF^{[V]}\}.$$

(3)

The values of the coding rates indicate the bit-rate values used for downloading the media of the traffic flow, whereas the coding format denotes the compression method used by the same media. The highest coding rate, $CR^{[M]}$, is associated with the maximum bit-rate at which the media codec can operate while the minimum coding rate, $CR^{[1]}$, corresponds to a guaranteed bit-rate.

$$CR_{guar} < CR_j < CR_{\max}.$$

(4)

The highest compression method, $CF^{[V]}$, indicates the format with the highest quality, whereas the lowest compression method, $CF^{[1]}$, corresponds to the format with the lowest quality. As summarized in Table 3.1, by increasing the quality levels, the *grade of quality* of the downloaded media improves in term of higher coding rates and better coding formats.

Table 3.1 Relationship between quality levels, coding rates, coding formats and grades of quality

Quality levels q_j	Coding rates CR_j	Coding formats CF_j	Grades of quality
1	**$CR^{[1]}$**	**$CF^{[1]}$**	**Low**
2	**$CR^{[2]}$**	**$CF^{[1]}$**	**…**
…	**…**	**…**	**Medium**
…	**…**	**…**	**…**
M	**$CR^{[M]}$**	**$CF^{[V]}$**	**High**

The filter without transcoding capabilities is expected to reduce the coding rates but cannot change the compression method. Reducing the coding rates without transcoding requires short processing time and simple packet processing because it is performed by simply inspecting the content of the packets' headers, and accordingly dropping the

packets based upon the results of the QoS filtering algorithm. To illustrate, traffic flows that include video streams compressed with coding format MPEG-4 (see Section 3.2.2 for a list of referenced coding formats) transport IP packets that contains three types of video frames: intra-frames (I), predictive (P), and interpolative (B). The coding rate of a traffic flow can be reduced when IP packets with P or/and B frames are dropped. This operation can be performed in all nodes of the network. On the contrary, transcoding requires longer processing time and more complex packet processing because the content of the packets' payload needs to be first de-compressed from its original format and then re-compressed in the new format. This operation is not implemented in the intra-PLMN routers because it requires access "on the fly" to the application layer.

The filters are implemented in the output port of the heterogeneous network nodes together with corresponding optimization functions algorithms. These adaptive algorithms, which will be defined in Chapter 4 and Chapter 5, control "intelligently" all the operations of the filters. Since the values of the filtering levels and associated coding rate and format for each traffic flow, are chosen by these "local" optimization functions there is no a centralized management scheme to control the filters. Hence, different policies in different nodes could be also allowed.

Moreover, a signaling overhead needs to be introduced, due to the implementation of the filters, to carry the information exchanged between the optimization functions (see Sections 4.2 and 5.3) in different nodes. The signaling overhead contains information about the filtering levels, coding rates and formats for the media included in each traffic flow, the number of media streams included in the flow, and the class of users (see Section 3.4) to which the traffic flow is directed. For example, this set of information

could be added either in the object fields of the signaling protocol RSVP message, as proposed in [40], or in the Diff-Serv field of each IP packet, in a Diff-Serv approach.

3.2.1 Deployment Issues of Filters with Transcoding Capabilities

This sub-section discusses the deployment issues of the multimedia adaptation filters. Basically, two possible scenarios arise: i) the filters are implemented inside the nodes; ii) the filters are implemented in an external server. In the first case, both the filtering and optimization functions (see Sections 4.2 and 5.3) are implemented in a dedicated processor inside the node. An IP packet carrying the content is forwarded to the application layer for transcoding. Recall that video and voice packets are transferred using *Real Time Streaming Protocol* (RTSP), *Real-Time Transport Protocol* (RTP) [41], and *User Datagram Protocol* (UDP), whereas data packets are transferred using *Hyper Text Transfer Protocol* (HTTP) and TCP protocol. The transcoding software module changes the coding format and forwards the newly encoded RTP (video or voice) or HTTP (data) packet to the IP layer, which then sends the packet to the output port for encapsulation in layer two, and subsequent transmission on the link. In the second case, both the filtering and optimization functions are implemented in an external proxy server [42] that is installed after the output port of the node. The optimization function, in this case, sends control messages from the server to the node to select values for the weights of the queues in the node.

3.2.2 Media Streams and Codecs

The types of media streams that can be downloaded are *voice* (or *speech*) (s), *video* (v), and *data* (d) (text, still images, bitmap and vector graphics). These types of media streams and their coding formats are assumed to be those defined within the

specifications of *Third Generation Partnerships Project* (3GPP) [43] and represented in Table 3.2. In this thesis, these media coding formats are used, although 3G-UMTS networks could support media with other formats.

Adaptive Multi-Rate (AMR) and *Wideband Adaptive Multi Rate* (W-AMR) codecs are used to encode voice traffic. AMR codec [44] supports 8 different speech codecs modes with peak encoding bit rates ranging from 4.75 kbps to 12.2 kbps and DTX (Discontinuous Transmission) mode (1.95 kbps). On the contrary, W-AMR [45] codec operates on bit-rates between 6.6 and 23.85 kbps (9 different speech codecs modes). The W-AMR codec gives superior speech quality.

Table 3.2 Media Codecs

Media types	Coding formats CF[1]	Coding formats CF[2]
Speech	AMR	AMR-Wideband
Video	H.263	MPEG-4
Still images	GIF	JPEG
Text	XML	XHTML

ISO/IEC *MPEG-4* and ITU *H.263* codecs are used for video traffic. MPEG-4 [46] and H.263 [47] video codecs have different transmission modes with average bit rates ranging from 10 Mbps to 64 kbps and from 64 kbps to 16 kbps, respectively. Distinct coding rates are obtained by generating scalable sub-streams (base and/or enhancement layers) that correspond to a spatial, temporal and/or *Signal to Noise Ratio* SNR scalability for the video frames. MPEG-4 also supports an object-based representation of the video scene by allowing separate *Audio Visual Objects* (AVO) [48] and a fine grained scalability called *Fine-Granular-Scalability* (FGS) [49].

Graphic Interchange Format (GIF), *Joint Photographic Expert Group* (JPEG), *Extensible Markup Language* (XML) and *Extensible Hypertext Markup Language*

(XHTML) codecs are used to encode bitmap or vector graphics and text for data traffic, respectively. Different coding rates are related to the number, resolution, color, and size of still images or how the text is summarized in web pages [42]. Hence, Web pages can be filtered to loose redundant objects [50, 51] or to translate the markup language that is interpreted by the displaying device.

Note that still images and text are treated by the filter as a single media without differentiating their coding rates, although each has a different coding format.

3.3 End-to-End Transfer Delay

A key QoS parameter that it is important to consider for the media content delivery via the heterogeneous network is the *end-to-end transfer delay*. It is assumed that the media types (voice, video, and data) that can be included in a traffic flow are downloaded according to the pre-requisites of a specific QoS class of service defined within the specifications of 3GPP: conversational, streaming, and interactive [52]. A traffic flow that contains voice streams is related to the conversational class of service, while a traffic flow with video streams represents the streaming service. Finally, text and still images are downloaded according to the prerequisites of the interactive services.

Given an assigned routing path (see Fig. 3.1) to download multimedia contents from the edge servers to the mobile users, and predefined end-to-end delay constraints [52], [53] minimum and maximum transfer delays for the *j*th traffic flow can be estimated, in each downstream link, as follows

$$\Delta_{\min} < \Delta_j < \Delta_{max}, \tag{5}$$

where minimum, $\Delta_{\min}$, and maximum Δ_{max}, transfer delay values are obtained by dividing the end-to-end transfer delay constraints [52, 53] within the core or radio access networks by the average number of hops along the downstream path.

One of the goals of the proposed architecture is to download the multimedia contents in each downstream within the required transfer delay constraints defined in (5). The end-to-end transfer delay may be violated, along the end-to-end path, if the requested bandwidth was temporarily not available in some downstream links. In these cases, the filtering process can be applied, in some downstream links, in order to maintain the transfer delay. Content reduction causes a temporary decrease in the quality levels of the downloaded media but avoids prolonged delays and packet loss for downloading the media.

3.4 Grades of Service (GoS): Premium and Economy

The choice of a specific quality level with its corresponding coding rate, and format that can be chosen for the *j*th traffic flow depends on the type of media contained in the flow and the type of user (or group of users) to which the flow is directed. In this thesis, two different types of users or "grades of services" are defined: *Premium* (P) and *Economy* (E). The two grades of services will offer users the same contents but at different levels of qualities. To illustrate, for Premium services, web pages can be browsed, at least, at 32 kbps, while for Economy class users this rate goes down to 16 kbps. Similarly, Premium service subscribers can receive a service like delivery of video content with MPEG-4 video format, at most, at 512 kbps, while for Economy the same service is with H.263 format at 128 kbps.

In order to receive content with a superior grade of quality in terms of higher coding rates and better coding formats for the media downloaded, Premium class subscriber will pay a higher tariff (see Sections 4.5 and 5.5) than an Economy class subscriber. They also will obtain other advantages that are achieved by using our filtering approach. For example, the proposed algorithm will assure that, in case of congestion, the filtering levels for Premium class traffic flows will be reduced much more gracefully than Economy class ones. Moreover, the algorithm will guarantee that the percentages of packets (voice, or video) lost and delayed (data) for Premium class users will be much less than those for Economy class users. Finally, Premium class users will have higher priority to be admitted in the network and lower probability to have their calls dropped in handover scenarios in comparison with Economy class users.

The most significant feature of the adaptive filtering architecture is that the range of selectable coding rates and types of formats for each traffic flow provides the network operator with a "flexible" means to provide different classes of users with different levels of quality while maximizing the network throughput according to the dynamic conditions of the links. Users will receive multimedia content within their agreed-upon contractual range of levels of quality that are pre-defined in their users' profiles. By selecting dynamically the levels of quality, the network operator is able to "intelligently" tune, or adapt the amount of traffic downloaded that depends upon the available resources of the downstream links in the core and radio access networks.

In this chapter, a general description of the filtering architecture and parameters was given. In the next two chapters, a more detailed description of the adaptive algorithms,

which are used to control the filtering operations in the core and radio access networks nodes, will be provided.

Chapter 4

A Filtering Algorithm in the Core Network Nodes

The aim of this chapter is to construct an adaptive algorithm that controls the filters located in the core network nodes. The chapter begins by presenting the scheme of the proposed Resource Manager in a core network node. The main components are the filter and a class-based queuing system. Next, an example of possible assignments of the filtering levels, coding rates and coding formats for Premium and Economy class traffic flows downloaded in the core network is provided. Following this, the multi-objective optimization function is defined, in order to control the actions of both the filter and a class-based queue. The function is solved by using a randomized search technique, called Genetic Algorithms. A revenue model and pricing mechanisms are also defined to compute the revenues gained by the service provider. The chapter ends by describing four different simulation experiments and showing the simulation results. A core network node without filtering is compared with other with filtering. The results show the significant improvements in the system performance of the proposed approach.

4.1 Resource Manager in a Core Network Node

As described in the previous section, the multimedia adaptation filters implemented in the core network are installed at the output ports of the SGSN, GGSN and intra-PLMN routers. This section describes the block diagram of the Resource Manager scheme for an

output port at a SGSN node. In this case, the multimedia adaptation filter and control module (optimization function) are implemented inside the node (see Section 3.2.1). As shown in Fig. 4.1, the proposed Resource Manager includes a traffic classifier, the multimedia adaptation filter, a class-based queueing packet scheduler, a traffic dispatcher, a control module and a data collection module. Once all routing operations are performed within the node, traffic is directed to the appropriate output port where the filter is installed.

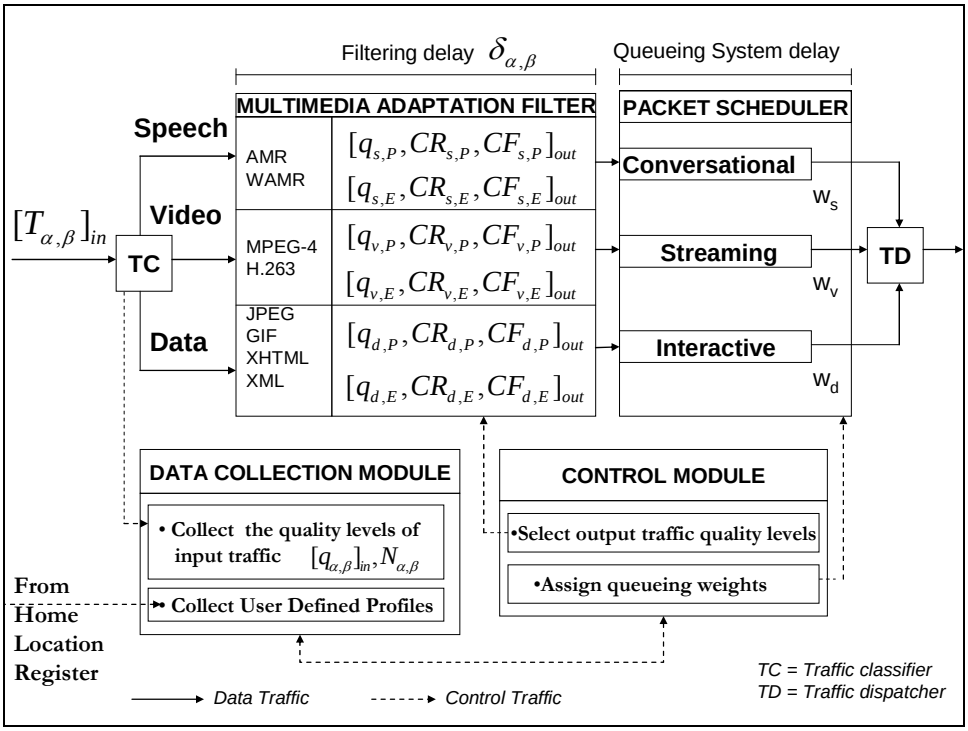

Fig. 4.1 Resource manager scheme for an output port at a SGSN node

As shown in Fig. 4.1, $[T_{\alpha,\beta}]_{in}$ represents the *j*th flow of the output port, characterized by media-type α and downloaded to the same category β of users. Recall that in the core network, a traffic flow represents the aggregate traffic of similar media streams that are downloaded to a group of users who share some common attribute. Therefore, the filter manipulates in the same way all media streams contained in the same traffic flow.

The *j*th flow of the output port is associated with a filtering vector $[q_{\alpha,\beta}, CR_{\alpha,\beta}, CF_{\alpha,\beta}]_{in}$ and with the current number $N_{\alpha,\beta}$ of media streams included in the same flow. During a certain time, the multi-scale filter manipulates the *j*th flow by assigning, if it is necessary, a new output vector

$$[q_{\alpha,\beta}, CR_{\alpha,\beta}, CF_{\alpha,\beta}]_{in} \Rightarrow [q_{\alpha,\beta}, CR_{\alpha,\beta}, CF_{\alpha,\beta}]_{out}. \quad \textbf{(6)}$$

If output and input vector parameters are different, an additional filtering delay $\delta_{\alpha,\beta}$ is introduced and has to be considered together with queuing and transmission delays. The filtering delay increases when transcoding is applied. Since the end-to-end delay constraints are different for traffic flows that are associated with distinct QoS classes of services (see Section 3.3), weights $[w_s, w_v, w_d]$ for each queue (Conversational, Streaming, and Interactive) are introduced. It is assumed, for simplicity, that the queuing and transmission delays of each traffic flow can be approximated by an *M/M/1* queuing system for every class-based queue.

A class-based queueing packet scheduler allocates dynamically a link-sharing bandwidth $w_\alpha \cdot 100\%$ (expressed as a percentage of the overall link bandwidth C) to each class of service over a predefined control period [54]. The link-sharing structure, illustrated in Fig. 4.2, specifies how the bandwidth is divided for a downstream link among the traffic flows. The link-sharing allocations are dynamic because change in response to current conditions on the network, according to our proposed algorithm.

The control module in Fig. 4.1 includes the optimization function that is implemented using GAs. Its details are provided in Section 4.2. The collection module collects *User Defined Profiles* (UDP). A UDP, which is stored in the HLR, contains the

range of quality levels, coding rates and formats that can be assigned for each media-type directed either to Premium or Economy class users.

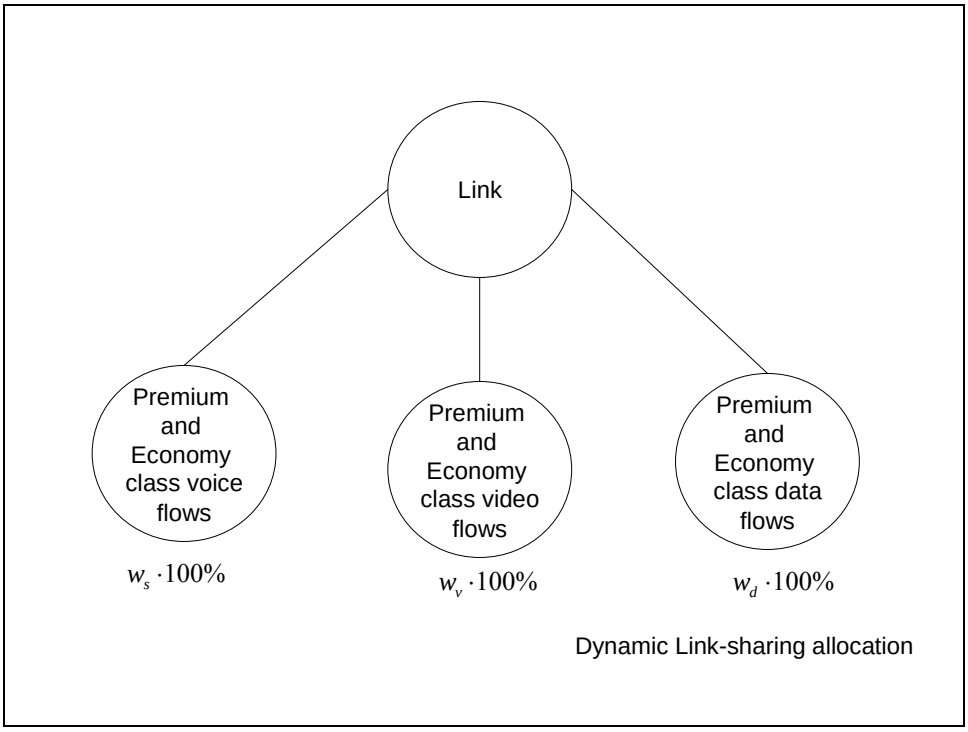

Fig. 4.2 Dynamic link-sharing structure

This information, which will be provided in the next sub-section, defines the solution space for the optimization function. The control module collects also the values of the minimum and maximum transfer delays for traffic flows downloaded in the output link.

4.1.1 A Possible Range of Quality Levels

As mentioned in Section 3.4, the choice of a specific quality level that can be chosen for the *j*th traffic flow depends on the type of media contained in the flow and the type of user group to which the flow is directed. Table 4.1 shows a possible assignment of filtering levels, coding rates and coding formats for Premium and Economy class traffic flows that are downloaded in the core network. These values will be used in the experiments provided at the end of this chapter. Still images and text are treated by the

filter as a single media without differentiating their coding rates; however, each has a different coding format (e.g., GIF or JPEG for images, XML or XHTML for data).

Table 4.1 Possible combinations of quality levels, coding rates, and formats for different media of Premium and Economy class flows downloaded via the core network

Speech			Video			Data		
q_j	CR_j (kbps)	CF_j	q_j	CR_j (kbps)	CF_j	q_j	CR_j (kbps)	CF_j
Economy class flows								
1	4	AMR	1	32	H.263	1	16	GIF, XML
2	8	AMR	2	64	MPEG-4	2	24	GIF, XML
3	10	AMR	3	96	MPEG-4	3	32	GIF, XML
4	12	AMR	4	128	MPEG-4	4	64	GIF, XML
Premium class flows								
2	8	AMR	2	64	MPEG-4	3	32	GIF, XML
5	16	AMR-W	4	128	MPEG-4	4	64	GIF, XML
6	20	AMR-W	5	384	MPEG-4	5	128	JPEG, XHTML
7	24	AMR-W	6	512	MPEG-4	6	256	JPEG, XHTML

As shown in Table 4.1, in case a traffic flow is directed to Economy class users the selectable filtering levels range from one to four (M=4). On the contrary, for a Premium traffic flow with voice media streams, the selectable filtering levels are two, five, six, and seven (M=7). In case of a Premium traffic flow containing video media streams, admissible filtering levels are two, four, five, and six (M=6). Finally, three, four, five and six are the filtering levels in case the Premium class traffic flow contains data streams (M=6). Since higher quality levels are associated with higher coding rates and better coding formats, the traffic flows downloaded to Premium class users have a higher quality range than those of the Economy class users.

A Premium class subscriber will pay more than an Economy class subscriber but will receive content with a superior grade of quality.

When a filtering process is applied, the media content is reduced during the process. Content reduction causes a temporary decrease in the quality of the downloaded media but avoids prolonged delays for downloading the media. Fig. 4.3 shows an example for downloading 32 Mb video content to a group of Economy class users via the core network.

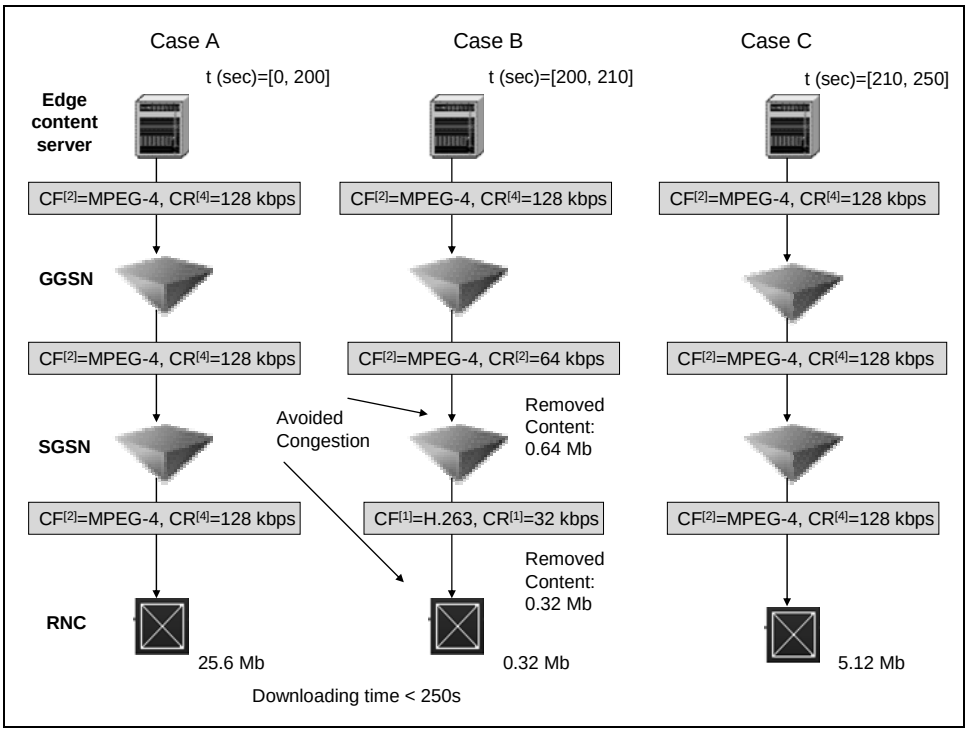

Fig. 4.3 Delivery of video content (32 Mb) to a group of Economy class users

When the process starts (case A; from t = 0s to t = 200s) the available capacity along the end-to-end path is enough to maintain the highest coding rate (128 kbps) and coding format (MPEG-4) for the traffic flow directed to mobile users group. During this time, end-to-end transfer delay and bandwidth requirements are satisfied and all downloaded content (25.6 Mb) reaches the radio network controller. During case B (from t = 200s to t = 210s), the available capacity in some downstream link do not permit the delivery of the information with the original QoS parameters. For instance, a degradation of radio

channel conditions could result in this scenario. In this case, the coding rate and format are reduced to avoid possible congestion and to keep delay and bandwidth within admissible range. At the same time, some content is removed (0.96 Mb) to respect the downloading time (< 250s). At last, in case C (from t = 210s to t = 250s), when conditions improve, content downloading turns back to the initial case.

4.2 The Optimization Function in a Core Network Node

This section defines the optimization function that is used to control the actions of both the filter and the queuing system. It operates on a control period (Δt) that determines the granularity level of the control actions performed by the optimization function. The shorter the control period, the finer the control is over the resources of the network, but at the expense of increasing the processing complexity of the function. As explained in the previous section, each flow is characterized by the same type of media streams ($\alpha \in \{s, v, d\}$) and is downloaded to a group of users ($\beta \in \{E, P\}$). In this section, only the parameters α and β are used to characterize a flow in order to explain the operations performed on each distinct flow. An incoming traffic flow, $[T_{\alpha,\beta}]_{in}(t)$, is characterized by the vector $[q_{\alpha,\beta}, CR_{\alpha,\beta}, CF_{\alpha,\beta}]_{in}(t)$. The actions performed by the optimization function determine after a predefined control period, Δt, an output traffic flow, $[T_{\alpha,\beta}]_{out}(t+\Delta t)$, and a weights vector, $\underline{w}(t+\Delta t)$. The output traffic flow is characterized by the vector $[q_{\alpha,\beta}, CR_{\alpha,\beta}, CF_{\alpha,\beta}]_{out}(t+\Delta t)$ whereas $\underline{w}(t+\Delta t)$ includes the weights of the class-based queuing system: $w_s(t+\Delta t)$, $w_v(t+\Delta t)$ and $w_d(t+\Delta t)$. Define the number of actual media streams contained within each flow to be $N_{\alpha,\beta}(t)$ and the

available capacity on an output link to be $C_{av}(t)$. The data collection module includes the information of the transfer delay constraints $(\Delta_{\alpha,\beta}^{\min}, \Delta_{\alpha,\beta}^{\max})$ and the range of filtering levels, coding rates and formats represented in Table 4.1. Finally, let $[q_{\alpha,\beta}, CR_{\alpha,\beta}, CF_{\alpha,\beta}]_{\min}$ be the vector with the minimum values for the filtering level, coding rate and format contained in Table 4.1.

The objective here is to find the output traffic flow and the weights vector that satisfy the following constraints

a) $w_s(t+\Delta t) + w_v(t+\Delta t) + w_d(t+\Delta t) = 1,$

b) $\Delta_{\alpha,\beta}^{\min} \le \Delta_{\alpha,\beta}(t+\Delta t) \le \Delta_{\alpha,\beta}^{\max},$

c) $R_{tot}(t+\Delta t) = \sum_{\alpha=s,v,d} \sum_{\beta=P,E} N_{\alpha,\beta}(t) \cdot CR_{\alpha,\beta}(t+\Delta t) \le C_{av}(t),$

d) $[q_{\alpha,\beta}, CR_{\alpha,\beta}, CF_{\alpha,\beta}]_{\min} \le [q_{\alpha,\beta}, CR_{\alpha,\beta}, CF_{\alpha,\beta}]_{out}(t+\Delta t)$
$\le [q_{\alpha,\beta}, CR_{\alpha,\beta}, CF_{\alpha,\beta}]_{in}(t),$

and maximize the multi-objective function F, defined as follows

$$F = F_{Util} \cdot F_R \cdot F_\Delta.$$

(7)

The first component of the objective function, F_{Util}, encourages solutions to use the largest portion of the available bandwidth $C_{av}(t)$ in order to maximize the capacity utilization. This function is represented by

$$F_{Util} = \begin{cases} R_{tot}(t+\Delta t) / C_{av}(t), & if \quad R_{tot}(t+\Delta t) \le C_{av}(t) \\ 0, & otherwise. \end{cases}.$$

(8)

The second component, F_R, controls the assignment of the filtering level or quality level $(q_{\alpha,\beta})_{out}(t+\Delta t)$ and, consequently, associated coding rate $(CR_{\alpha,\beta})_{out}(t+\Delta t)$ and format $(CF_{\alpha,\beta})_{out}(t+\Delta t)$, for each output traffic flow. The function is given by

$$F_R = \prod_{\alpha=s,v,d} F_D^{\alpha} \cdot \left(\sum_{\alpha=s,v,d} F_R^{\alpha,P} + \sum_{\alpha=s,v,d} F_R^{\alpha,E} \right) / 6 ,$$

(9)

where

$$F_R^{\alpha,P} = \begin{cases} \left[\dfrac{(q_{\alpha,P})_{out}(t+\Delta t) - (q_{\alpha,P})_{\min}}{(q_{\alpha,P})_{in}(t) - (q_{\alpha,P})_{\min}} \right]^2, & if \quad (q_{\alpha,P})_{\min} \le (q_{\alpha,P})_{out}(t+\Delta t) \le (q_{\alpha,P})_{in}(t) \\ 0, & otherwise, \end{cases},$$

(10)

$$F_R^{\alpha,E} = \begin{cases} \sqrt{\dfrac{(q_{\alpha,E})_{out}(t+\Delta t) - (q_{\alpha,E})_{\min}}{(q_{\alpha,E})_{in}(t) - (q_{\alpha,E})_{\min}}}, & if \quad (q_{\alpha,E})_{\min} \le (q_{\alpha,E})_{out}(t+\Delta t) \le (q_{\alpha,E})_{in}(t) \\ 0, & otherwise, \end{cases},$$

(11)

$$F_D^{\alpha} = \begin{cases} 0, & if \quad \dfrac{(q_{\alpha,P})_{in}(t) - (q_{\alpha,P})_{out}(t+\Delta t)}{(q_{\alpha,P})_{in}(t)} > \dfrac{(q_{\alpha,E})_{in}(t) - (q_{\alpha,E})_{out}(t+\Delta t)}{(q_{\alpha,E})_{in}(t)} \\ 1, & otherwise. \end{cases}.$$

(12)

The choice of the filtering levels is determined in a different way for traffic flows downloaded to Premium and Economy class users. The function $F_R^{\alpha,P}$ is quadratic and controls the output filtering levels for the traffic flows downloaded to Premium class users. The function $F_R^{\alpha,E}$ is radical and determines the filtering levels for the Economy class traffic flows. Both functions encourage an output filtering level that is as close as

possible to the input filtering level, but the strength that they utilize is different. The quadratic function is used for Premium traffic flows because it encourages high filtering levels more strongly than the radical function. Consequently, Premium class traffic flows tend to be degraded less than Economy class traffic flows. The reduction in the filtering level of a certain traffic flow depends on the assignments of the filtering levels of the output traffic flow in comparison with the filtering levels of the incoming traffic flow. This is defined according to the following formula:

$$D_{\alpha,\beta}(t+\Delta t) = \frac{(q_{\alpha,\beta})_{in}(t) - (q_{\alpha,\beta})_{out}(t+\Delta t)}{(q_{\alpha,\beta})_{in}(t)}.$$

(13)

The filtering reduction can be related to the reduction in the bit-rate of the media content. As the filtering level is decreased so is the bit-rate of the content; however, the correspondence is not one-to-one proportionally. For example, consider an input traffic flow with voice streams that is coded at filtering level 4, which corresponds for an Economy class user to a bit-rate of 12 kbps. The output traffic flow filtering level has been reduced to a level 2, which corresponds to a bit rate of 8 kbps. This means a 50 % of filtering reduction; however, this does not mean that the bit-rate has been reduced by 50 %. The content has been certainly reduced but the percentage, 34 % in this example, depends upon the relationship between coding rates and filtering levels. This correspondence has been previously provided in Table 3.1.

In case the number of media streams $N_{\alpha,P}(t)$ contained in the Premium class traffic flows, is equal or higher than the number of streams $N_{\alpha,E}(t)$ included in the Economy class traffic flows, both functions (10) and (11) cannot guarantee that the filtering

reduction $D_{\alpha,P}(t+\Delta t)$ for Premium class flows will be lower than the reduction $D_{\alpha,E}(t+\Delta t)$ for the Economy class flows. The function F_D assures that this condition is always respected.

The last term of the objective function, F_Δ, ensures that the transfer delay of each traffic flow is not violated. It is assumed that each class-based queue can be modeled as a simple *M/M/1* queuing system. Hence, queuing and transmission delay are calculated based upon this approximation. The total transfer delay of a traffic flow for a downstream link can be written as

$$\Delta_{\alpha,\beta}(t+\Delta t)=\frac{1}{\frac{w_\alpha(t+\Delta t)\cdot C_{av}(t)}{L}-\frac{N_{\alpha,E}(t)\cdot CR_{\alpha,E}(t+\Delta t)+N_{\alpha,P}(t)\cdot CR_{\alpha,P}(t+\Delta t)}{L}}+\delta_{\alpha,\beta} \quad \textbf{(14)}$$

where $\delta_{\alpha,\beta}$ represents the filtering delay that could be introduced by the filter while L is the average length of media packets. The value of the filtering delay depends on the choice of coding rates and coding formats of the input and output traffic flows. If both coding formats and coding rates of input and output traffic flows change, a transcoding delay δ_2 is introduced. In case, only coding rates change, a lower filtering delay δ_1 is considered. No filtering delay is added when input traffic parameters are maintained. This delay can be represented as follows:

$$\delta_{\alpha,\beta}=\begin{cases}\delta_2, & if \quad [CR_{\alpha,\beta},CF_{\alpha,\beta}]_{in}(t)\neq[CR_{\alpha,\beta},CF_{\alpha,\beta}]_{out}(t+\Delta t)\\ \delta_1, & if \quad [CR_{\alpha,\beta}]_{in}(t)\neq[CR_{\alpha,\beta}]_{out}(t+\Delta t),[CF_{\alpha,\beta}]_{in}(t)=[CF_{\alpha,\beta}]_{out}(t+\Delta t)\\ 0, & if \quad [CR_{\alpha,\beta},CF_{\alpha,\beta}]_{in}(t)=[CR_{\alpha,\beta},CF_{\alpha,\beta}]_{out}(t+\Delta t).\end{cases} \quad \textbf{(15)}$$

The delay function is given by

$$F_{\Delta} = F_w \cdot \prod_{\alpha=s,v,d} \prod_{\beta=P,E} F_{\Delta}^{\alpha,\beta}$$

(16)

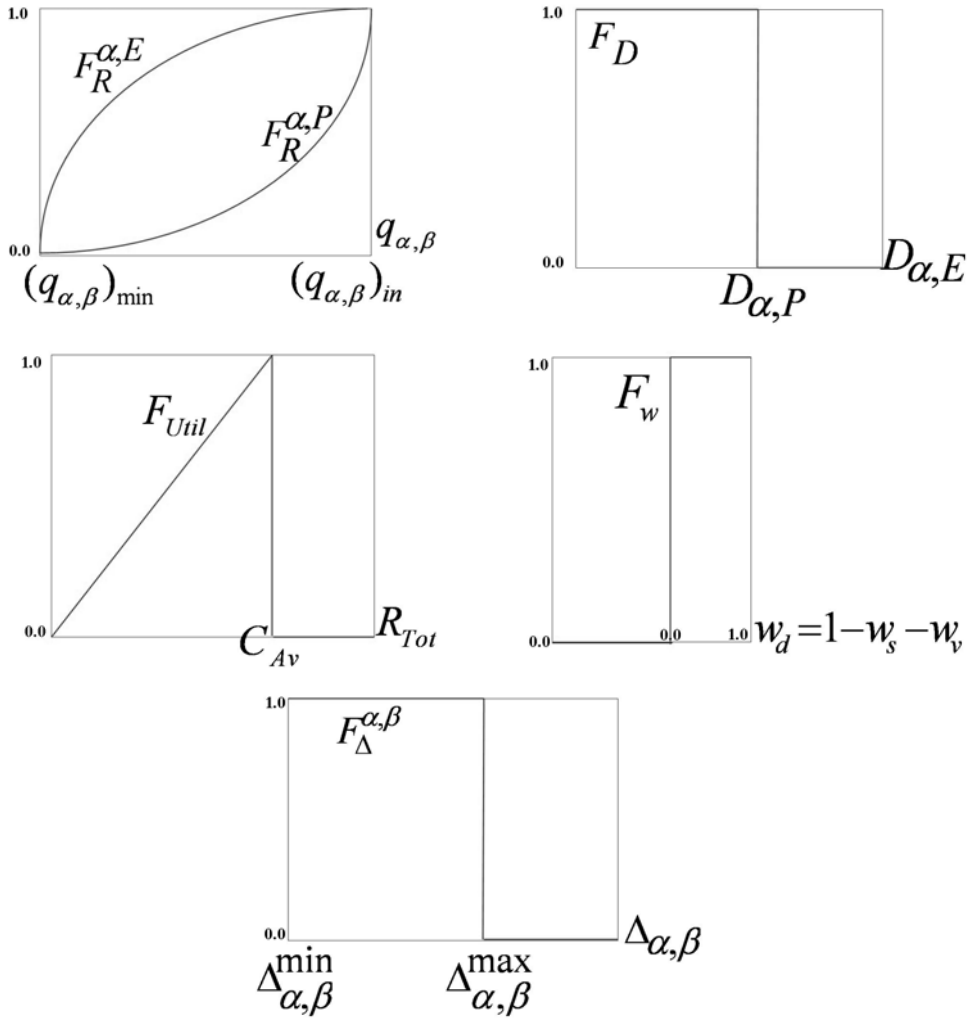

Fig. 4.4 Multi-objective function components

The function $F_{\Delta}^{\alpha,\beta}$ guarantees that the transfer delay of each traffic flow is maintained within pre-defined constraints while the function F_w ensures that the constriction a) for the queuing weights is always satisfied

$$F_{\Delta}^{\alpha,\beta} = \begin{cases} 1 & if \quad \Delta_{\alpha,\beta}^{\min} \le \Delta_{\alpha,\beta}(t+\Delta t) \le \Delta_{\alpha,\beta}^{\max} \\ 0 & otherwise \end{cases}$$

(17)

$$F_w = \begin{cases} 1, & if \quad \sum_{\alpha=s,v,d} w_\alpha(t+\Delta t) = 1 \\ 0, & otherwise. \end{cases}$$

(18)

Fig. 4.4 is a graphical representation of the multi-objective components defined above. The next two sections, first, provide an overview of techniques that can be used to solve a generic optimization problem, then, present the details of the technique chosen in order to optimize our specific function: Genetic Algorithms (GAs).

4.3 Overview of Search Techniques

An optimization problem can be solved analytically or by using search techniques. Using *Analytic Approaches*, optimal solutions can be found by differentiating the objective function with respect to each variable. Hence, these techniques are not suitable in our case since the objective function the product of step functions that are not differentiable. Considering search techniques, many methods are found: *Hill Climbing*, *Enumerative*, *Random Search Algorithms* and *Randomized Search Techniques*. Hill Climbing finds an optimum solution by following the local gradient of the function. They are deterministic and assume that the problem space being searched is continuous in nature. This is not true in our problem because the problem space is discontinuous and non-unique. Moreover, they only find the local optimum in the neighborhood because they rely on a single point to search throughout the space. Enumerative methods consist of looking at the function values at every point in the solution space but if this is large, the computational task becomes massive and sometimes intractable. Random Search

Algorithms only perform random walks, recording the best optimum values discovered so far. They do not use knowledge gained from previous results and thus do not learn. On the contrary, Genetic Algorithms (GAs) that belong to the last cited search method, use random choices to guide themselves but these are not directionless. They use knowledge gained by the previous results, combining them with randomizing features. They also look at many different areas of the problem space at once by generating systematically a population of points instead of a single one. Therefore, they are more efficient to solve non-unique and multimodal search spaces. GAs have been used to address diverse practical optimization problems related to dynamic resource management, see [55] and [56].

4.4 Genetic Algorithms (GA) Details

The Genetic Algorithms (GAs) are utilized in order to solve the optimization problem described in Section 4.2. As described in [57], the use and implementation of a GA requires the determination of the six main components described in the following sub-sections.

4.4.1 Chromosome Representation

Each *chromosome* is a way of representing the variable(s) in the function to be optimized. It is a string made up of a sequence of genes from a certain alphabet. An alphabet could consist of binary digits, floating point numbers, integers, symbols, etc.

In our optimization problem, the chromosome is a vector whose elements are simply the queueing weights $\underline{w}(t)$ and the quality levels $(q_{\alpha,\beta})_{out}(t)$ of the output traffic flow $[T_{\alpha,\beta}]_{out}(t)$. The queueing weights, which range from zero to one, are represented by a

floating point notation with accuracy of 10^{-4} whereas the quality levels are represented by integers.

4.4.2 Initialization

The GA starts to operate on an initial population of chromosomes, and thereafter generates better ones to find the best solution. The most common method is to randomly generate chromosomes for the entire population. However, since GAs can iteratively improve existing solutions, the beginning population can be seeded potentially good solutions.

In our problem, if the number of traffic flows is high, the solution space of the optimization problem contains many areas where the value of the objective function is zero (no solution). This is due to several constraints taken into consideration in the optimization problem and the numerous allowable combinations of the weights. The queueing weights range between 0 and 1 with a floating-point accuracy of 10^{-4}. To prevent the GA from being trapped in these areas, it is generated an initial population with some possible solutions that include appropriate combinations of the weights. The initial population consists of chromosomes

$$[\, f_{s,P} \quad f_{v,P} \quad f_{d,P} \quad f_{s,E} \quad f_{v,E} \quad f_{d,E} \quad w_s \quad w_v \quad w_d \,] \tag{19}$$

that include all the possible combinations of filtering levels ($4^6 = 4096$) $f_{s,P} \in \{2,5,6,7\}$, $f_{v,P} \in \{2,4,5,6\}$, $f_{d,P} \in \{3,4,5,6\}$, $f_{s,E} \in \{1,2,3,4\}$, $f_{v,E} \in \{1,2,3,4\}$, $f_{d,E} \in \{1,2,3,4\}$, with appropriate choices for the weights as follows:

$$\begin{cases} w_s = \dfrac{N_{s,P} \cdot CR_{s,P} + N_{s,E} \cdot CR_{s,E}}{R_{tot}} \\ w_v = \dfrac{N_{v,P} \cdot CR_{v,P} + N_{v,E} \cdot CR_{v,E}}{R_{tot}} \\ w_d = \dfrac{N_{d,P} \cdot CR_{d,P} + N_{d,E} \cdot CR_{d,E}}{R_{tot}} \end{cases} \quad \textbf{(20)}$$

where $R_{tot} = \sum_{\alpha=s,v,d} \sum_{\beta=P,E} N_{\alpha,\beta} \cdot CR_{\alpha,\beta}$

By this initial population, the possibilities of obtaining no solutions (objective function zero) are reduced in case the number of flows $N_{\alpha,\beta}$ is high and some solution exists, simultaneously.

4.4.3 Selection or Reproduction

The selection of chromosomes to produce successive generations plays an important role in the genetic algorithm. A probabilistic selection is performed based upon the individual's fitness such that the better individuals have an increased chance to be chosen. To solve our optimization problem, normalized geometric selection [62] is used.

4.4.4 Genetic Operators

Genetic operators provide the basic search mechanism of the GA. The operators are used to create new chromosomes based on existing solutions in the population. There are two basic types of operators: *crossover*, and *mutation*. Crossover takes two chromosomes and produces two new ones by partially exchanging information between them using cross site at random. Mutation alters one chromosome by changing the value of a string position.

In our problem, a combination of arithmetic, heuristic, and simple crossovers is used. A combination of uniform and not uniform mutations is utilized [62].

4.4.5 Evaluation

Each chromosome is evaluated to obtain its fitness score. The *fitness* function evaluates the closeness of a certain chromosome to the optimum solution. This function should be carefully designed to describe the optimization problem.

In our problem, the evaluation function is given by the optimization function defined in (7).

4.4.6 Termination

The GA moves from generation to generation selecting and reproducing chromosomes until a termination criterion is met.

To solve the optimization problem in Section 4.2, two strategies are used in conjunction with each other: *convergence* and *time-out* criteria. The convergence criterion is applied when there is a lack of improvement in the best solution over a specified number of generations. Time out case happens when a maximum number of generations are reached. This corresponds to a maximum control period allotted to GA to solve the optimization problem.

4.4.7 Mechanics of Genetic Algorithms

The mechanics of a simple genetic algorithm are summarized in the following procedure:

1. Initialize a population of chromosomes. (Initialization)

2. Evaluate each chromosome in the population. (Evaluation)

3. Select parent chromosomes probabilistically according to their fitness. (Selection)

4. Create new chromosomes by mating parent ones; apply mutation and recombination as the parent chromosome mate. (Genetic operators)

5. Evaluate the new chromosomes using the designated fitness function. (Evaluation)

6. Delete old weak members of the population to make room for the new strong chromosomes. (Selection)

7. If time is up, or convergence is achieved, stop and return the best chromosome; otherwise go back to step 4 (Termination)

In our case, the operations of GA are illustrated in Fig. 4.5.

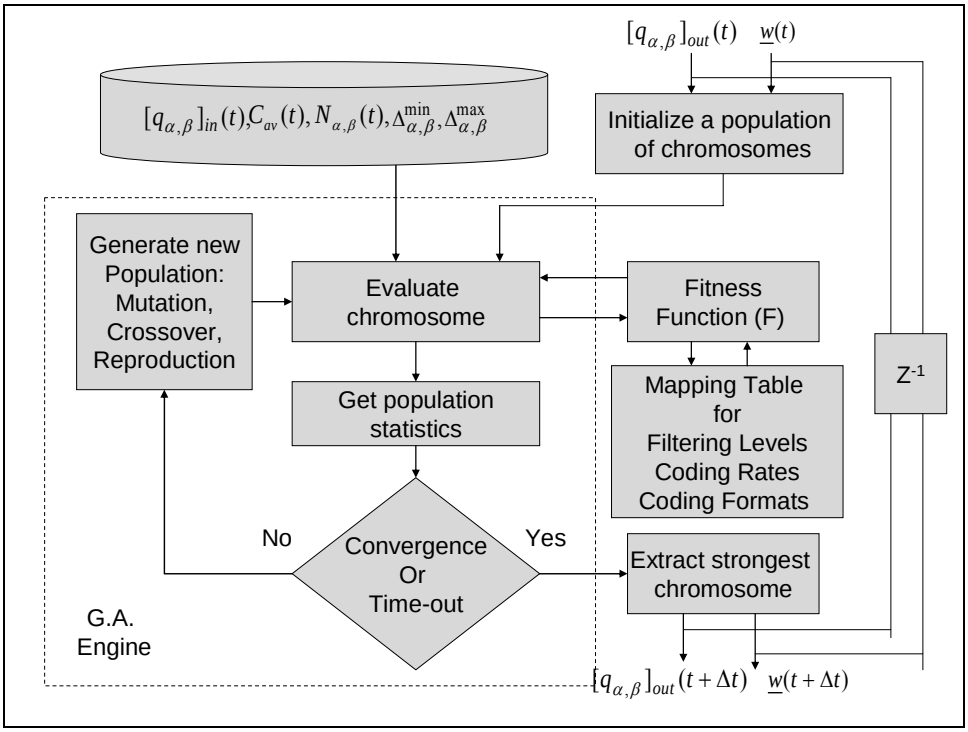

Fig. 4.5 Genetic Algorithm block diagram

Once the stopping criterion is reached, the algorithm sends new values to the filter $[q_{\alpha,\beta}]_{out}(t+\Delta t)$ and to the queue $\underline{w}(t+\Delta t)$. These are the values that have maximized the objective function. They are maintained for the duration of one control period until a new

cycle is ended. The filtering action and consequently the content reduction are applied to the traffic flows whenever the value of the output filtering level $[q_{\alpha,\beta}]_{out}(t+\Delta t)$ is different from that of the input $[q_{\alpha,\beta}]_{in}(t)$ .In the meantime, a new solution is also being used to initialize part of the initial population at the next control period. This feedback is performed to ensure that solutions are held in case there is no change in the input parameters.

4.5 A Revenue Model and Pricing Mechanisms

This section describes the revenue model and the pricing policies used to evaluate the total revenues gained by the service provider. From a user perspective, it is important that all the QoS requirements of the media streams are satisfied, and a preferential treatment is provided to the users that pay more. From a service provider perspective, it is important that the links' throughput and the total revenues are maximized.

In computing the revenues, first, the issue of pricing [58] must be taken into account. Among the pricing policies [59], flat rate pricing is the most common mode of payment today for bandwidth services, due to its simplicity and cost-effectiveness for the user. Although flat rate pricing is advantageous to subscribers, it does not offer any motivations for users to adjust their services' demands. On the other hand, usage-based pricing regulates usage by imposing a fee based on the amount of data actually sent. These two pricing mechanisms are static because prices are independent of current state or condition of the network. A pricing policy may be also dynamic in the sense that prices vary as a result of network's load, and congestion status.

In this thesis, the pricing mechanism is based on the amount of traffic downloaded, similar to the usage-based pricing mechanisms, but it also depends on the class of users requesting the service [60], and the current format assigned by the filter to the media, similarly to a dynamic pricing mechanism. In fact, since the filters are capable of reducing the levels of quality of media streams, by changing their coding formats according to the dynamic conditions of the network, the pricing mechanism will take into account a decrease in the tariff when this event occurs. A price differentiation per class of users is also necessary because filters provide a privileged treatment to the media streams downloaded by users who receive higher levels of quality.

In essence, the proposed pricing mechanism takes into account different factors such as the class of users requesting the service, the bit rates and the types of format at which media streams are downloaded. The following simple revenue model [61] is used:

$$\Phi = \sum_{\alpha=s,v,d} \sum_{\beta=P,E} N_{\alpha,\beta} \cdot T_{\alpha,\beta} \cdot CR_{\alpha,\beta}.$$

(21)

The revenue measure Φ describes the total revenues generated due to the traffic downloaded by both Premium and Economy class users during a control period. In this model, $T_{\alpha,P}$ and $T_{\alpha,E}$ are the tariffs that Premium and Economy class users pay during a control period (Δt) in order to download each type of media stream with filtering level $q_{\alpha,\beta}$, coding rate $CR_{\alpha,\beta}$, and coding format $CF_{\alpha,\beta}$. Without loss of generality, the tariffs in a control period are assumed to be expressed in dollars/Mbps. In the following, it is defined a pricing strategy in three different scenarios.

In scenario 1, the tariff for Premium class users depends upon the selected filtering levels (see Table 4.1) assigned to the downloaded media streams. On the other hand, the tariff for Economy class users does not depend on the choice of the filtering levels but only on the amount of traffic downloaded as follows:

$$T_{\alpha,P} = \begin{cases} 0.20 \ \ \$/Mbps, & q_{\alpha,P} > 4 \\ 0.15 \ \ \$/Mbps, & otherwise, \end{cases}$$

(22)

$$T_{\alpha,E} = 0.10 \ \ \$/Mbps, \ \ \forall q_{\alpha,E}.$$

(23)

In scenario 2, the tariffs for both Premium and Economy class users depend on the amount of traffic downloaded and on the filtering levels of the downloaded media streams as follows:

$$T_{\alpha,P} = \begin{cases} 0.20 \ \ \$/Mbps, & q_{\alpha,P} > 4 \\ 0.10 \ \ \$/Mbps, & otherwise, \end{cases}$$

(24)

$$T_{\alpha,E} = \begin{cases} 0.10 \ \ \$/Mbps, & q_{s,E} \geq 2; q_{v,E} \geq 2; q_{d,E} > 2 \\ 0.05 \ \ \$/Mbps, & otherwise. \end{cases}$$

(25)

Finally, in scenario 3, the tariffs for both Premium and Economy class users do not depend neither on the filtering levels of the downloaded media streams, nor on the class of services, but only on the amount of traffic downloaded as follows:

$$T_{\alpha,P} = 0.10 \ \ \$/Mbps, \ \ \forall q_{\alpha,P}.$$

(26)

$$T_{\alpha,E} = 0.10 \quad \$/Mbps, \quad \forall q_{\alpha,E}.$$

(27)

Scenario 3 represents a pure usage-based pricing mechanism whereas, scenarios 1 and 2 include also a price differentiation based upon the different classes of users as well as the actual filtering levels of the media streams. In scenarios 1 and 2, Premium class users pay a higher tariff because the filter provides a preferential treatment to their media streams over those directed to Economy class users.

4.6 Core Network Node Simulator

This section explains the simulations experiments that have been carried out to evaluate the performance of the proposed algorithm for a node (SGSN) of the core network. It is assumed that the core network contains three nodes (routers), i.e., two hops. The filtering delays and the packet transfer delay constraints, which are considered for each traffic flow, are provided in Table 4.1. The transfer delay constraints for each traffic flow are obtained by dividing the maximum packet transfer delay in the core network (Table 4.2) by the number of hops.

The simulator emulates a core network node (SGSN) with a 15 Mbps output link where multimedia contents are downloaded. It is written in Matlab and the Genetic Algorithm optimization tool (Gaot), which is described in [62], is used to solve the optimization problem. The maximum time (Δt) allotted to GA to solve the optimization problem is fixed to 1 sec. In fact, the genetic algorithm takes less than this maximum period to satisfy the convergence criterion.

Table 4.2 Transfer delay constraints and filtering delays for each class of service or traffic flow in the core network

	Conversational Class of service	Streaming class of service	Interactive class of service
Traffic flows	**Voice**	**Video**	**Data**
Delay characteristics	**strict and low**	**Bounded**	**Tolerable**
Transfer delay constraints	**(0 ms, 10 ms)**	**(0 ms , 15 ms)**	**(0 ms , 40 ms)**
Maximum transfer delay in the Core Network	**20 ms**	**30 ms**	**80 ms**
Filtering delays	**δ_1=5 ms, Δ_2=9 ms**	**δ_1=5 ms, δ_2=9 ms**	**δ_1=5 ms, δ_2=9 ms**

To simplify the simulations, it is assumed that each of the incoming traffic flows is associated with the highest quality level described in Table 4.1.

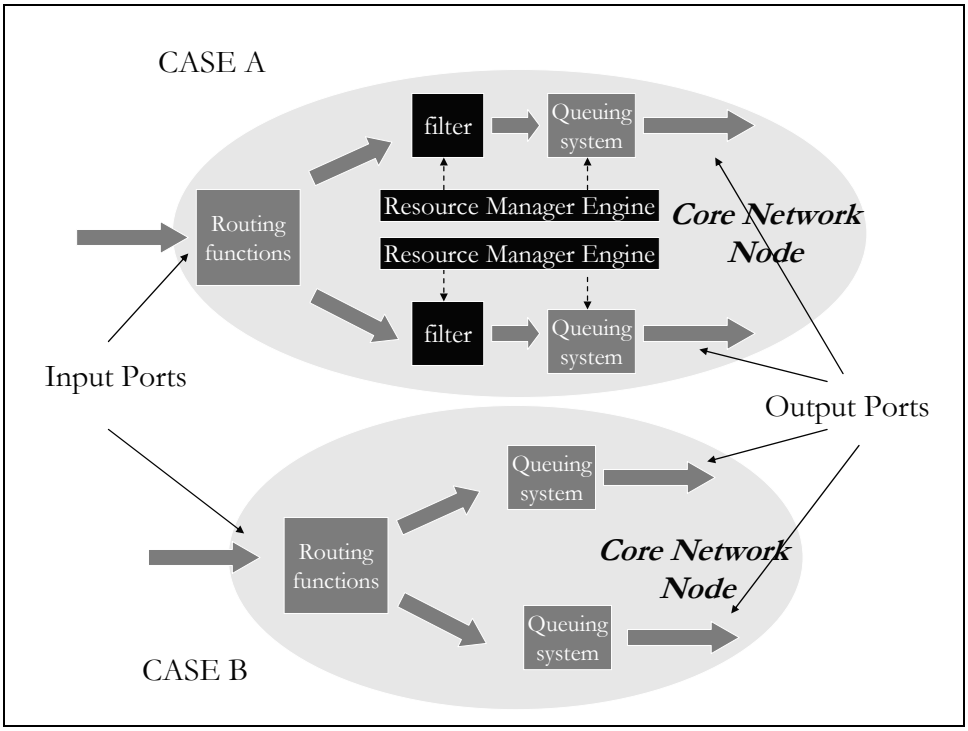

Fig. 4.6 Architecture of a node without filter (Case A) and a node with filter (Case B)

The quality levels of the incoming traffic flows are fixed for our simulations, though in general these could be time-dependant. Moreover, it is considered an average packet length of L=1000 bits for all different traffic flows even if the packet size depends on the type of traffic flows. Voice and video traffic sources generate streams with a constant bit-

rate that corresponds to the highest coding rates in Table 4.1. For example, voice traffic streams of Premium class users are generated with a constant bit rate of 24 kbps. Video traffic streams of Premium class users are generated with a constant bit rate of 512 kbps. Data traffic flows are modeled according to a Poisson distribution with an average rate corresponding to the highest coding rate in Table 4.1. To test the algorithm, a single node case without filtering (Case A), is compared with another with filtering (Case B), see the architecture in Fig. 4.6. In the following sections, different experiments are provided to study the aspects and features of the proposed algorithm in different scenarios.

4.7 Experiment 1

In this experiment, each user requests a multimedia call containing all types of media: voice, video and data. First, the link is gradually loaded with only Premium class calls until the last user is blocked. Then, this experiment is repeated taking into consideration all possible combinations of Premium and Economy class users. Each simulation continues until the last user is blocked. This happens when the QoS requirements are violated. Performance parameters of interest include system capacity (maximum number of users that can download content via the tested link), the links' utilization, transfer delay of each traffic flow, and the average filtering reduction in the quality (coding rates and formats) of the media content for the traffic flows downloaded by the Premium and Economy class users.

4.7.1 System Capacity

This sub-section evaluates the effectiveness of the proposed algorithm to enhance the capacity of the core network in terms of the maximum number of users that could be

admitted to download multimedia content via the tested link without violating the pre-defined QoS requirements. The results are shown in Fig. 4.7.

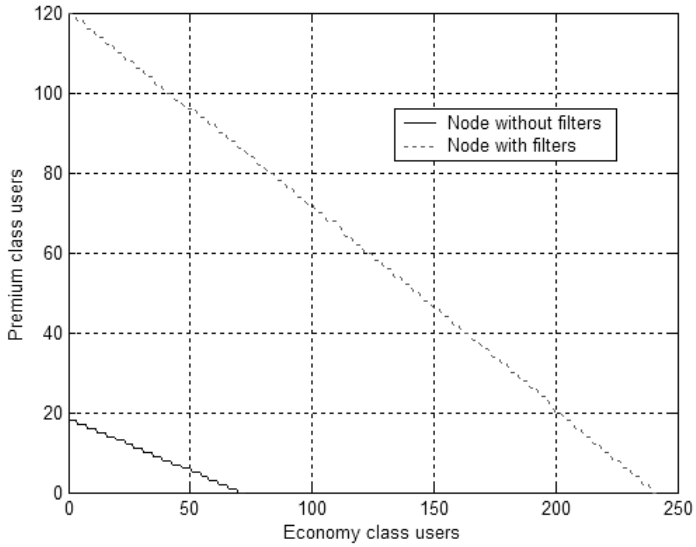

Fig. 4.7 Admission region of Premium and Economy class users

Without filtering, the total number of acceptable combinations of Economy and Premium class users is 700; whereas with filtering it increases by a factor of 20 to about 14000. Clearly, there is a major gain in terms of increasing the network capacity. For example, the maximum number of Economy class users has increased from 73 to 242, almost a 200% improvement. Similarly, the maximum number of Premium class users has increased from 18 to about 120, another impressive 500% gain. This is achieved without violating the pre-defined constraints of transfer delay, as well as maximizing the link utilization. These results are shown in the next two sub-sections.

4.7.2 Link Capacity Utilization

Fig. 4.8 compares the utilization of the available capacity C_{av} of the link for all admissible combinations of users. In a node without filters, case A, neither the coding rates nor the format of the media could be altered. In case B, the filter can alter the bit-

rate that is requested to download each traffic flow. Values for capacity utilization are shown in detail in Fig. 4.9.

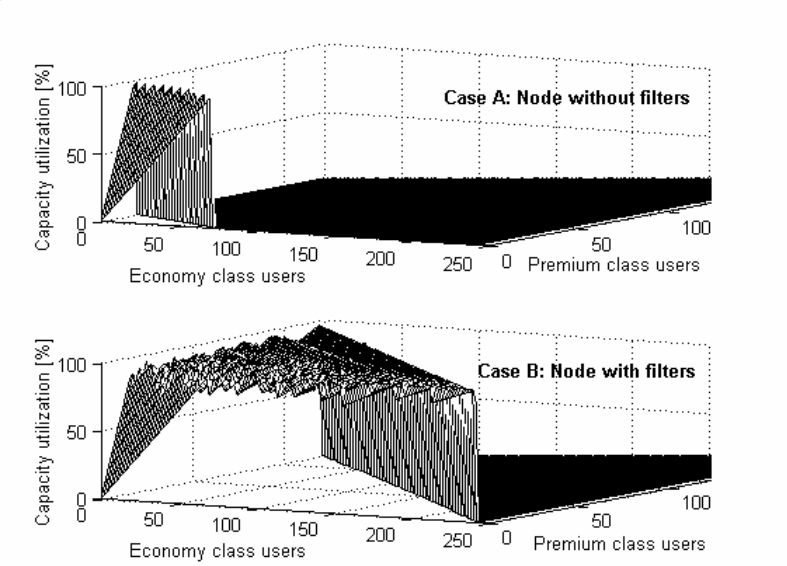

Fig. 4.8 Capacity utilization against admitted users for a node without filters (Case A) and a node with filters (Case B)

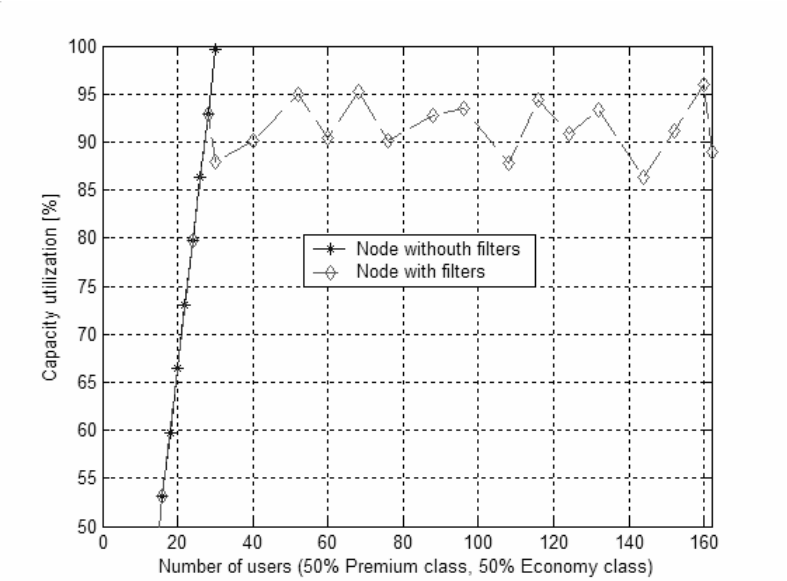

Fig. 4.9 Capacity utilization versus admitted users (50% Premium class, 50% Economy class).

This graph is a cross section of the 3D graphs shown in Fig. 4.8, and shows the capacity utilization against the number of users when the percentage of Premium and

Economy class admitted users is maintained equally at 50% each. Up to 28 users, the available capacity is enough to support the requested bit-rates of all downloaded traffic flows without filtering. As the system overloads, only the filtering case can support the users while maintaining high utilization, and without violating the delay constraints.

4.7.3 Packet Transfer Delay

The values of the transfer delays of each traffic flow (voice, video, and data) are provided in Fig. 4.10.

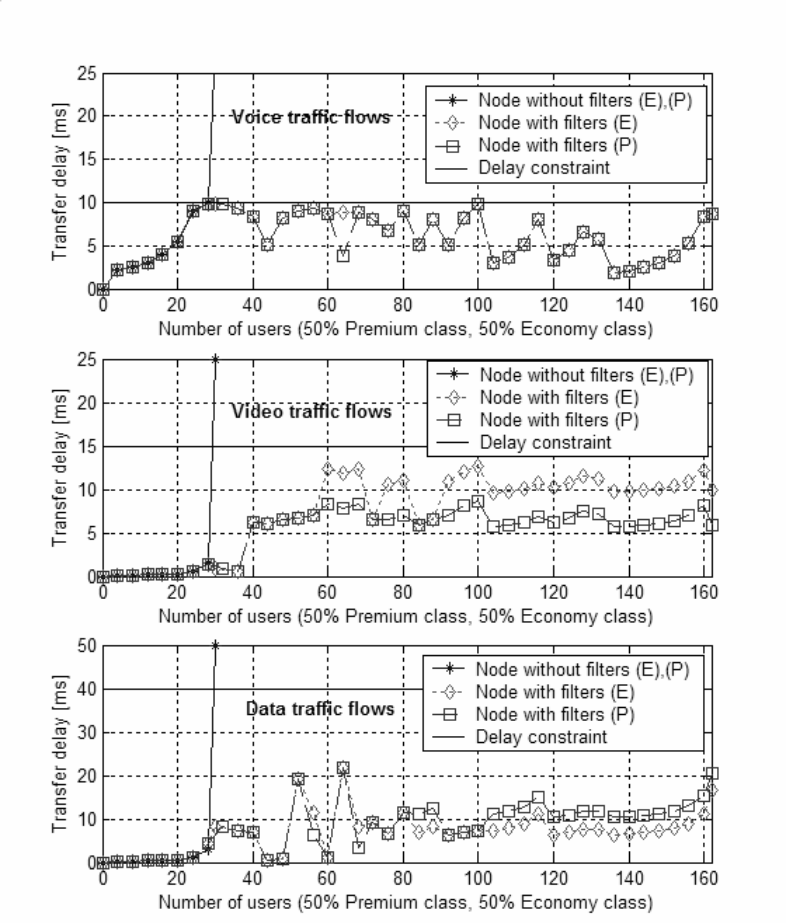

Fig. 4.10 Transfer delay for Premium and Economy class traffic flows (voice, video, and data) against admitted users (50% Premium class, 50% Economy class).

In the graph, the percentage of Premium and Economy class users is still maintained at 50% each, as in Fig. 4.9. For a node with filtering, the transfer delays for both Premium and Economy class traffic flows are depicted. Note that since different filtering actions are taken on Premium and Economy traffic flows, the filtering delays and consequently the transfer delays for the traffic flows are different. In the range of users (up to 28), the transfer delay of all traffic flows is maintained within the pre-defined constraints. As the system overloads, the filter starts to reduce the coding rates of some media types. This action is performed to avoid violating the pre-defined delay constraints (Fig. 4.10). In case of a node without filtering, the additional users are admitted because there is enough bandwidth but delay constraints are no longer respected although the capacity utilization continues to increase (Fig. 4.9). Hence, the impact of that is that the users will perceive "slower" downloads. On the other hand, the filter manages to satisfy the pre-defined constraints of each traffic flow (10ms for voice, 15ms for video, and 40ms for data) at the price of decreasing the amount of data to be transferred.

4.7.4 Quality Levels Reduction

In order to evaluate the effect of the filtering actions taken by the filter on the Premium and Economy class traffic flows, an *Average Filtering Reduction Level* for Premium and Economy traffic flows is defined. The filtering reduction is calculated by averaging the function in equation (13) with respect to all types of media. This is provided in the following equation:

$$D_{\beta} = \frac{1}{3} \cdot \sum_{\alpha=s,v,d} D_{\alpha,\beta} .$$

(28)

Fig. 4.11 shows the impact of the reduction in the filtering levels (or coding rates) on different traffic flows in three distinct scenarios.

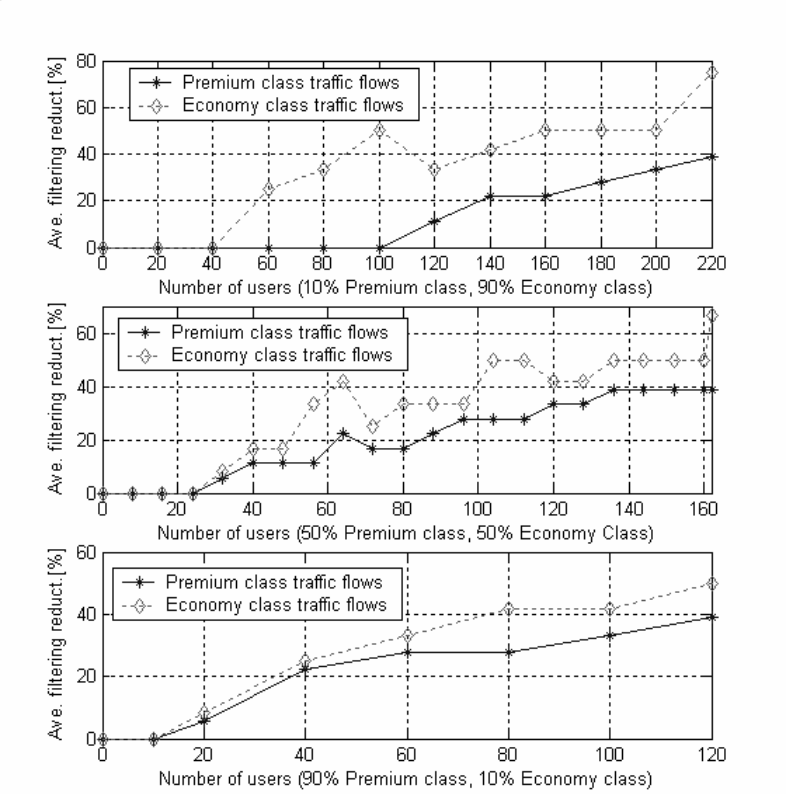

Fig. 4.11 Average filtering reduction level for Premium and Economy class traffic flows versus admitted users.

In each one, different percentages of Premium and Economy class users are used. In all scenarios, the rate reduction for Premium class traffic flows is always less than that of the Economy ones. The maximum value of the reduction for traffic belonging to Premium users is 40%, whereas for Economy class users this value is 75%. An average reduction of 75% for Economy class users means that those users will have all three media streams with the filtering level equal to one, which corresponds to the their minimum guaranteed coding rates. Furthermore, as this thesis mentioned in Section 4.2, the filtering reduction is not directly proportional to the content reduction. In this case, if the filtering levels are replaced by the coding rates in equations (13) and (28), an average

reduction of 73% for the multimedia content is obtained. For Premium class users, an average filtering reduction of 40% means that there will be approximately 55% of content reduction. In this case, the filtering level for voice streams is seven, for video streams is two, whereas for data streams is three. If the percentage of Premium class users is lower than the percentage of Economy class users, functions (10) and (11) work well in term of degrading the coding rates of the Economy class traffic flows much more than Premium class flows, and function (12) is not in effect within the optimization function (7). This means that, in this case, the filtering reduction for Premium and Economy class users would be still the same whether the function (12) was included in the optimization function or not. If the percentage of Premium class users increases, the filtering degradation for both Premium and Economy class traffic flows becomes much closer. Hence, function (12) tends to play a more significant factor in the optimization function. In fact, as this thesis indicated in Section 4.2, if the number of media streams contained in the Premium class traffic flows is equal or higher than the number of streams included in the Economy class traffic flows, function (12) assures that the filtering reduction for Premium class flows will be lower than that of the Economy class flows. The results in Fig. 4.11 confirm that function (12) is in effect in the case when the number of Premium class users is 50% or 90%. In fact, without this function the degradation in the Premium class users could not be maintained lower than that of the Economy class users. Accordingly, the optimization function (7) works better in case the number of Premium class users is lower than Economy class users. This situation is represented in the first graph of Fig. 4.11 where 10% of Premium and 90% of Economy class users are admitted.

All results presented in Fig. 4.11 show the average filtering reduction of traffic flows versus the increasing number of Premium and Economy class users. A different way to evaluate the effect of the filtering actions is to graph the average filtering reduction of the traffic flows versus the increasing percentages of Premium or Economy class users.

Fig. 4.12 and Fig. 4.13 represent the filtering reduction for Premium and Economy traffic flows versus different percentages of Premium and Economy class users when a fixed total number of traffic flows is loaded in the tested link. In Fig. 4.12, the total number of traffic flows is set to 24, 30, and 40, whereas in Fig. 4.13 the number of flows is fixed to 60 and 100. As expected, in all scenarios the filtering reduction of Premium class traffic flows is always maintained lower than Economy class ones. Moreover, in all scenarios the filtering reduction of both traffic flows increases as the percentage of Premium class users that are admitted increases.

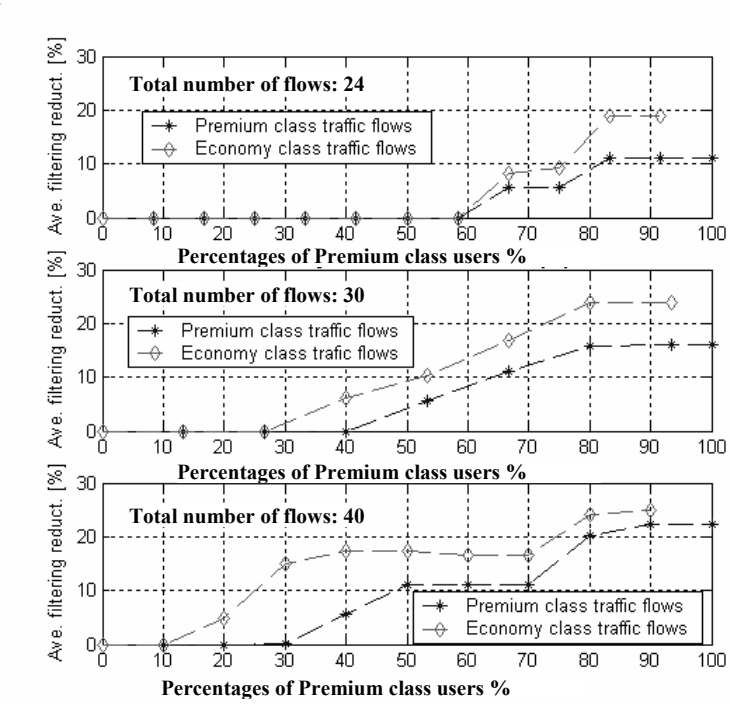

Fig. 4.12 Average filtering reduction for Premium and Economy traffic flows versus percentages of Premium and Economy class users with a fixed number of flows (24, 30, and 40) for a node with filters.

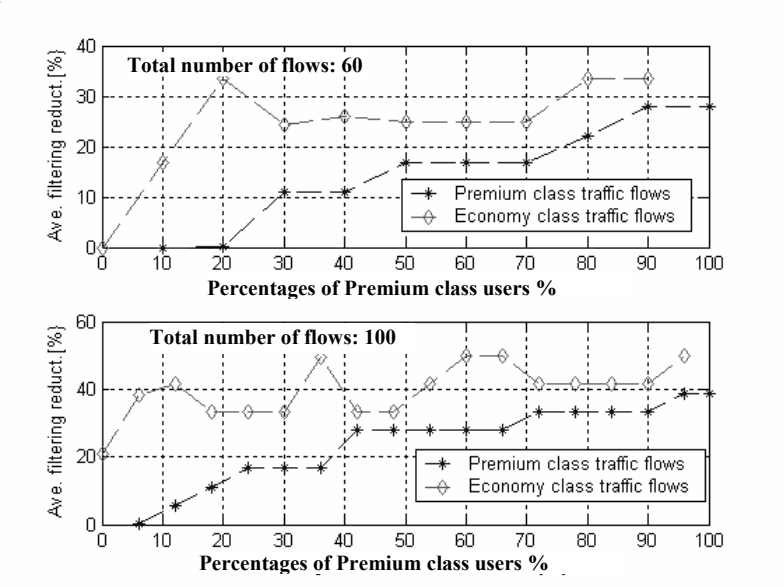

Fig. 4.13 Average filtering reduction for Premium and Economy traffic flows versus percentages of Premium and Economy class users with a fixed number of flows (60 and 100) for a node with filters.

If the total number of traffic flows increases (e.g., from 24 to 100 flows), the filtering reduction of traffic flows is activated when lower percentages of Premium class users are admitted (e.g., from 65% to 5%). To illustrate, if 24 traffic flows are loaded in the link, the filtering reduction of Premium traffic flows begins when 65% of Premium class users are admitted (about 16 Premium class traffic flows out of 24 total flows). On the contrary, if 100 traffic flows are loaded, the filtering reduction begins when 5% of Premium class users are admitted (5 Premium class traffic flows out of 100 total flows).

Finally, when the percentage of Premium class users is low the filtering reduction for Economy class traffic flows is much more than that of Premium class ones. On the contrary, as the percentage of Premium class users increases the filtering reduction for Premium and Economy class traffic flows tends to be closer.

4.7.5 Selection of Coding Rates

If the quality levels of traffic flows are reduced, the coding rates of the media included in the traffic flows are lowered, accordingly. This sub-section shows how the coding rates of traffic flows that include all types (voice, video, and data) of media are lowered.

Fig. 4.14 shows the coding rate values for all types of Premium and Economy traffic flows versus the number of users.

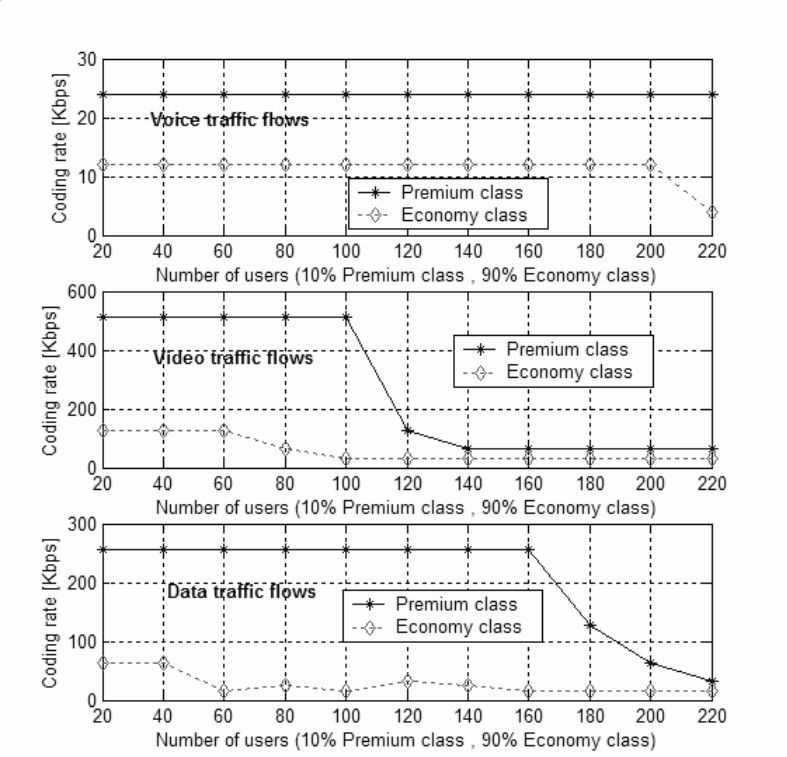

Fig. 4.14 Coding rate for Premium and Economy class traffic flows against total admitted users (10% Premium, 90% Economy).

The percentages of Premium and Economy class users are fixed to 10% and 90%, respectively, because this scenario can be considered as the most realistic case. As shown in Fig. 4.14, for small traffic-loads all types of media streams included in both Premium and Economy traffic flows can be downloaded at the highest possible coding rates.

Premium class users can download video streams at a coding rate of 512 kbps, data streams at a rate of 256 kbps and voice streams at a rate of 24 kbps. For a node without filtering, the coding rates are fixed and only 4 Premium class users can be admitted. With filtering, however, up to 10 Premium users can download video streams at the highest coding rate. In case of data streams, up to 16 users can download text and still images at the maximum coding rate. The coding rate of voice streams is decreased at a lesser rate than that of the other streams. This is because the QoS requirement for the transfer delay of voice streams is very close to the value of the filtering delay. When the filtering delay is added to the queuing and transmission delays, the resultant total transfer delay is often higher than the QoS requirement for maximum tolerable transfer delay, thus the filtering action on the voice streams is not allowed by the algorithm.

In order to maintain the highest possible coding rates to the Premium class users, Economy traffic flows have to be degraded not just more frequently, but also more aggressively. However, in case a small number of users are admitted and the percentage of Economy class users is much larger than the percentage of Premium class users, the coding rate reduction for each Economy class user will not be so aggressive because it can be distributed fairly among a larger number of Economy class users in comparison with the small percentage of actual Premium class users. If more users are admitted, Premium class users, also, will start to observe a decrease in the quality of their media because the available capacity will not be enough to satisfy the highest coding rates. At the same time, Economy traffic flows will start to be more intensively degraded. A network operator may decide to operate the optimization function such that the number of admitted users is less than what is actually possible. This mode of operation will limit the

maximum percentage of degradation that the users will see. This is an important trade-off between revenues (maximum number of admitted subscribers) and users' satisfaction with the service provided.

4.7.6 Transcoding Activation

According to the proposed filtering approach, the filtering operation is performed via a two steps approach: first, the coding rate of the traffic flows is reduced without transcoding; and then, "if necessary", the coding rate is further reduced via transcoding. As mentioned, the transcoding operation determines the change of the coding format for the media included in the traffic flow. This sub-section examines the transcoding activation. Fig. 4.15 shows the offered traffic-loads that trigger the activation of rate reduction via transcoding.

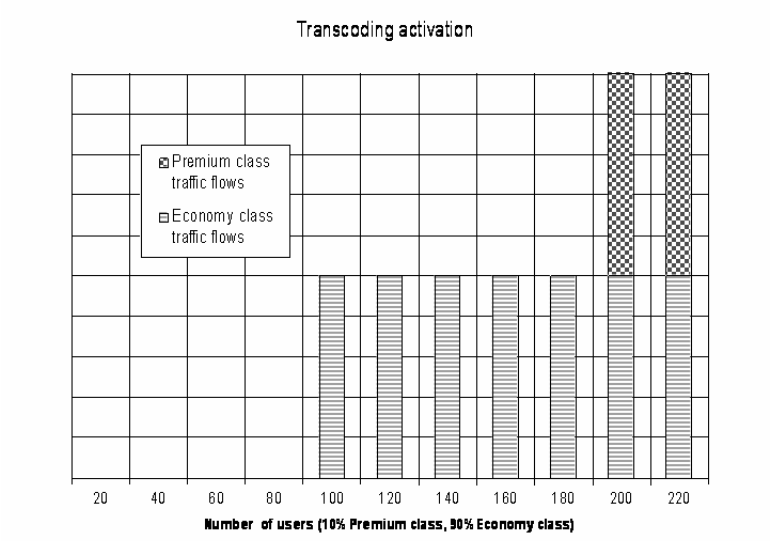

Fig. 4.15 Transcoding activation against total admitted users (10% Premium class, 90% Economy class).

The same scenario (10% Premium and 90% Economy class users) as in the previous sub-section is evaluated. Clearly, transcoding is activated only at high traffic-loads (100 users and above is the equivalent to more than 90% utilization).

Results in the next experiments (see Section 4.8.2) will show that the most favorable operating region of the provider has a maximum of 100 users due to the saturation of revenues. Hence, transcoding would be rarely activated inside the core network; whereas, it would be more activated in the RNCs (see Section 5.7.4) due to the fact that the radio channel impairments and users' mobility could lead to episodes where the capacity of the channel would drop suddenly, thus causing congestion and consequently activation of the transcoding.

4.8 Experiment 2

In this experiment, each user requests a multimedia call containing all types of media: voice, video and data. The link is first loaded with only Premium class calls (traffic flows). For each number of admitted calls, the total revenues, gained on average in a control period, are computed using the revenue model in Section 4.5. The simulation continues as the revenues calculated, via admitting new calls, keep increasing. This experiment is, then, repeated taking into account fixed percentages of Premium and Economy class calls until the link is loaded with only Economy class calls. For each percentage of Premium and Economy class calls, the simulation continues by admitting new calls, until the revenues stop increasing. This means that, in each simulation, additional users are not accepted if the revenues gained by admitting those users do not improve.

For further discussions, the thesis considers a theoretical case of a node without filtering in which the link is completely loaded with Premium and Economy class calls and the ratio between the total coding rate assigned to download Premium and Economy class media streams is maintained about as 4 to 1. This ratio is calculated by using the

coding rates of the media streams in case no filtering is performed (see Table 4.1). The computation of the maximum revenues Φ and the number of admitted Premium N_P and Economy N_E class users against the different percentages of Premium, x_P, and Economy, x_E, class calls is achieved, for the theoretical case, by using the following equations (see Appendix A):

$$\Phi = \Phi_P + \Phi_E = \frac{C_{av}(Mbps)}{\left(1 + \frac{x_E}{x_P} \cdot \frac{\sum_{\alpha=s,v}^{d} CR_{\alpha,E}^{w.f.}(Mbps)}{\sum_{\alpha=s,v}^{d} CR_{\alpha,P}^{w.f.}(Mbps)}\right)} T_{\alpha,E}^{w.f.}(\$/Mbps) + \frac{C_{av}(Mbps)}{\left(1 + \frac{x_P}{x_E} \cdot \frac{\sum_{\alpha=s,v}^{d} CR_{\alpha,P}^{w.f.}(Mbps)}{\sum_{\alpha=s,v}^{d} CR_{\alpha,E}^{w.f.}(Mbps)}\right)} T_{\alpha,E}^{w.f.}(\$/Mbps)$$

(29)

$$N_P = \frac{C_{av}(Mbps)}{\frac{x_E}{x_P} \cdot \sum_{\alpha=s,v}^{d} CR_{\alpha,E}^{w.f.}(Mbps) + \sum_{\alpha=s,v}^{d} CR_{\alpha,P}^{w.f.}(Mbps)}$$

(30)

$$N_E = \frac{C_{av}(Mbps)}{\frac{x_P}{x_E} \cdot \sum_{\alpha=s,v}^{d} CR_{\alpha,P}^{w.f.}(Mbps) + \sum_{\alpha=s,v}^{d} CR_{\alpha,E}^{w.f.}(Mbps)}$$

(31)

where $CR_{\alpha,P}^{w.f.}$, and $CR_{\alpha,E}^{w.f.}$ are the coding rates allocated to the Premium and Economy class calls, while $T_{\alpha,P}^{w.f.}$ and $T_{\alpha,E}^{w.f.}$ are the tariffs for Premium and Economy class users if no filtering is done.

This experiment is performed to verify the effectiveness of the proposed algorithm to increase the revenues of the service provider by accommodating an increased number of admitted users. An interesting result to report is that the performance improvement of the system (measured by the gain in the number of admitted users at a certain utilization

factor) is not simply bounded by the maximum available link throughput. It is, rather, limited by the additional revenue gained by admitting more users. The increase in the revenue saturates at a certain offered traffic-load. Hence, it is not worth it, from a service provider perspective, to admit additional users above this traffic-load despite the fact that the filtering algorithm results indicate otherwise. This traffic load determines the most "favorable operating region" for both service provider and subscribers. Other performance parameters of interest include per user average costs for downloading media streams. All mentioned results are shown in the following sub-sections.

4.8.1 Maximum Revenues and Operating Region

The results for the maximum revenues and the maximum number of admitted users (calls), which maximize the revenues (operating region), versus different percentages of Economy and Premium class calls are shown in Fig. 4.16, and Fig. 4.17, respectively.

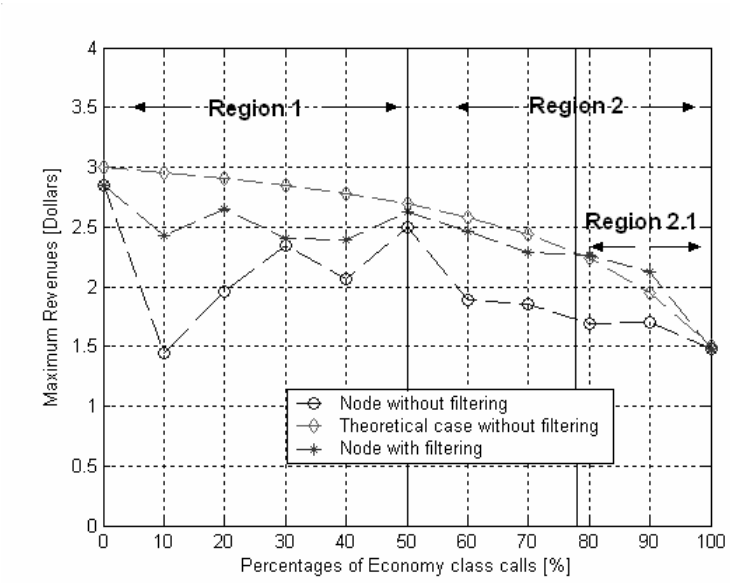

Fig. 4.16 Maximum revenues generated, on average, in a control period against the percentages of Premium and Economy class calls for a node without filtering, for a theoretical case without filtering, and for a node with filtering (Tariffs in Scenario 1).

The revenues are calculated using the tariffs (22), and (23), defined in Section 4.5 for scenario 1. As shown in Fig. 4.16, the highest revenues of a provider ($ 2.85 for the node with or without filtering; $ 3 for the theoretical case) are achieved when the link capacity is completely loaded with Premium class calls. In this case, the maximum number of accepted Premium class users is 18 for the node with or without filtering, and 18.75 for the theoretical case (see Fig. 4.17). On the contrary, if only Economy class calls load the link the revenues are halved ($ 1.48 for the node with or without filtering; $ 1.5 for the theoretical case). The maximum number of Economy class users is 73 for the node with or without filtering, and 75 for the theoretical case.

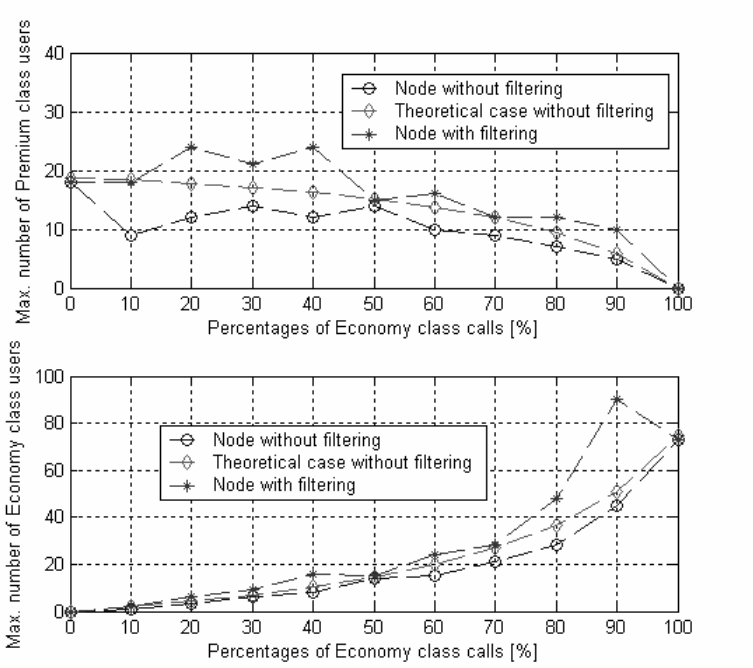

Fig. 4.17 Maximum number of admitted Premium and Economy class users against the percentages of Economy and Premium class calls for a node without filtering, a theoretical case without filtering, and a node with filtering.

Hence, because of the differentiated pricing mechanism adopted in scenario 1, the revenues are higher when the capacity is allocated only to Premium class calls. Moreover, consider averaging the maximum revenues in Fig. 4.16 and the number of

admitted users in Fig. 4.17 with respect to the percentages of Premium and Economy class calls. Without filtering the resultant revenues would be $ 1.96 (the number of admitted users would be 32), whereas with filtering would increase to $ 2.36 (the number of admitted users would increase to 48), almost a 21% improvement (50% improvement in the number of admitted users). In the theoretical case, the resultant revenues would be $ 2.53, only 7% higher than those with filtering.

As shown in Fig. 4.16, the maximum revenues in a node with filters are always higher than those without filtering if Economy and Premium class calls are loaded in the link. To understand how the filtering algorithm can increase the maximum revenues, Fig. 4.16 is analyzed in three different regions. For higher percentages of Premium class calls (see Region 1 in Fig. 4.16), the revenues with filtering are higher because the capacity utilization is maintained higher through the filtering action as new calls are admitted. In fact, since Premium class calls consume a larger amount of bandwidth than Economy class ones (about 4 times more) it is difficult, without filtering, to load efficiently the link because additional Premium class calls cannot be admitted. To illustrate, consider the link loaded with 90% of Premium and 10% of Economy class calls. Without filtering, up to 9 Premium and 1 Economy class users can be accepted (Fig. 4.17) and the maximum provider revenues are about $ 1.5 (Fig. 4.16). With filtering, up to 18 Premium and 2 Economy class users could be admitted (Fig. 4.17) and the maximum revenues are increased almost to $ 2.5 (Fig. 4.16), that is a 67% improvement. With a node with filtering a larger capacity is utilized, leading to higher maximum revenues.

If the percentage of Economy class calls is higher than that of Premium ones (see Region 2 in Fig. 4.16), higher maximum revenues are achieved because the filtering

reduction in the media of Premium class calls is much less than that of Economy ones. To illustrate, consider the link loaded with 80% of Economy and 20% of Economy class calls. Without filtering, only up to 28 Economy and 7 Premium class users can be accepted and the maximum revenues are about $ 1.69. With filtering, up to 48 Economy and 12 Premium class users could be admitted and the maximum revenues are increased to $ 2.26 that is a 34% improvement. This is due to the fact that, as the number of admitted calls increases, more bandwidth is allocated, through the filtering algorithm, to Premium class calls that are more lucrative for the provider.

Finally, when the percentage of Economy class calls is higher than 78% (see Region 2.1 in Fig. 4.16), the maximum revenues in the filtering case are not only higher than those for the node without filtering, but are also higher than those for the theoretical case. Considering a link loaded with 90% of Economy and 10% of Premium class calls, the maximum revenues are $ 1.44 without filtering, $ 1.95 in the theoretical case and $ 2.2 in the filtering case. In this case, by admitting additional users and providing a preferential treatment to the Premium class calls (the coding rates of Economy class calls are reduced much more than those of Premium class calls), the provider can obtain maximum revenues even higher than an ideal case in which the capacity utilization is 100%.

4.8.2 Saturation of Revenues Gain

This sub-section elaborates the limitations on the gain achieved by increasing revenues via admitting more users (calls). The simulation in this case is not stopped at the maximum value of the revenues but continues until the QoS requirements of each media streams are maintained within the pre-defined constrains. Fig. 4.18 represents the revenues gained by a provider versus an increasing number of calls in case the percentage

of Premium and Economy class calls is fixed to 10% and 90%, respectively. This classification is chosen because it presents a realistic scenario in which most users are indeed Economy class subscribers. The graph is showed to compare the revenues of a provider for a node with and without filters when all different pricing mechanisms defined in Section 4.5 are applied. Using the tariffs (22), and (23), defined in scenario 1, and the tariffs (24), and (25), in scenario 2, the increase in the revenues ends at 100 calls, although the graph show that additional 120 calls could be admitted. When more than 100 users download multimedia contents, the filtering algorithm maintains the QoS requirements but the revenues do not increase. Hence, there is no advantage for a service provider to admit any more users above this limit despite the fact the filtering algorithm would maintain the QoS parameters even for a larger number of calls.

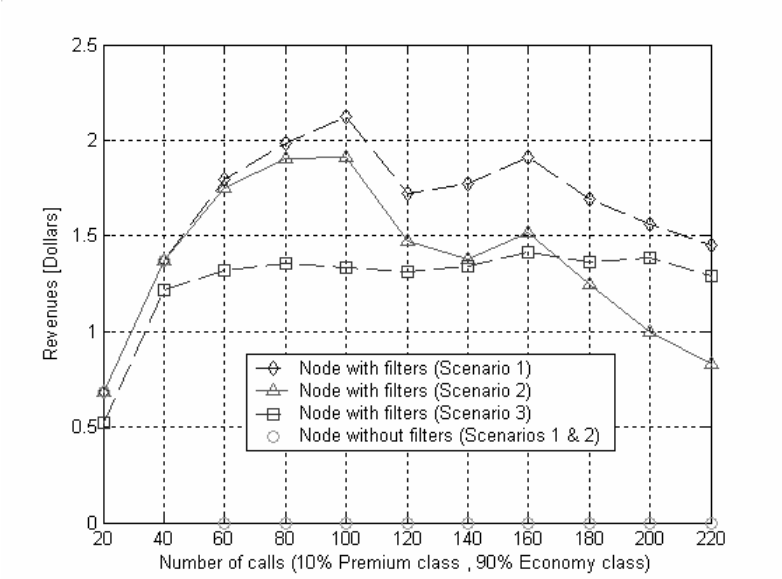

Fig. 4.18 Revenues generated on average in a control period versus admitted calls (10% Premium class, 90% Economy class) for a node with filters (Tariffs in Scenarios 1 , 2 & 3) and a node without filters (Tariffs in Scenarios 1 & 2).

For the filtering case and the tariff plan of scenario 1, the revenues reach $ 2.2 for 100 users. However, the maximum revenues for the no filtering case are $ 1.4 at

approximately 40 calls, since the network cannot admit more than 40 users. The revenue gain is approximately 50%, whereas the number of admitted users has more than doubled. Hence, the range between 40 and 100 calls is a favorable operating region for both users and the operator.

In this region, the revenues increase because the filtering levels of traffic flows directed to Premium class users, who pay a higher tariff than Economy class users in scenarios 1 and 2, are still not reduced or are reduced at a significant lesser rate than those of the flows directed to the Economy class users. As more than 100 calls are admitted, the filtering degradation of Premium and Economy class traffic increases, and the revenues, calculated with the pricing policy in scenarios 1 and 2, decreases.

On the contrary, using the tariffs (26), and (27) in scenario 3, the revenue is maintained constant at $ 1.4 (the capacity utilization is 94%) when the filter is activated. This is due to the fact that Premium and Economy class users pay the same tariff for the traffic downloaded. Hence, if the provider defines an appropriate pricing policy that takes into account the different users' demands and priorities, an additional revenue gain can be achieved by enabling the proposed filtering architecture.

All results presented in Fig. 4.18 show the revenues versus the increasing number of Premium and Economy class calls. A different way to evaluate the saturation of the gain in the revenues is to graph the revenues versus the increasing percentages of Premium or Economy class calls. These results are obtained by loading the link with different percentages of Premium and Economy class calls while maintaining the number of total calls fixed.

Fig. 4.19 shows the saturation of revenues (five different scenarios) versus different percentages of Premium and Economy class calls when the total number of admitted calls is set to 24, 30, 40, 60, and 100. For each scenario, although the pre-defined QoS requirements of each call are maintained within the pre-defined constrains with the filtering algorithms, the revenues gain saturates when the percentage of admitted Premium class calls reaches a certain value. The revenues are calculated using the tariffs (22), and (23) in Section 4.5.

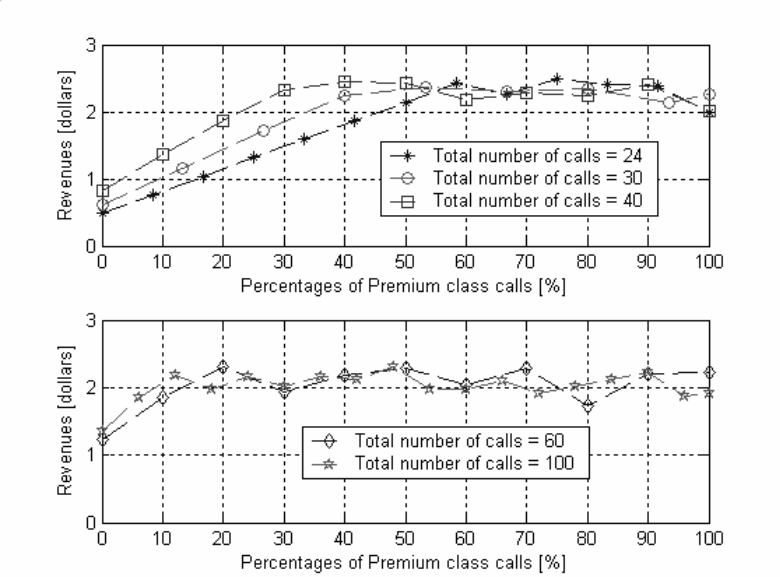

Fig. 4.19 Revenues generated on average in a control period versus percentages of Premium and Economy class calls with a fixed number of total calls (24, 30, 40, 60, and 100) for a node with filters.

As shown in Fig. 4.19, the revenues increase when higher percentages of Premium class calls are loaded until a certain value from which the revenues gain saturates. In this case, the revenues do not decrease as in Fig. 4.18, but tend to a stable value because the total number of calls is fixed.

The percentage of Premium class calls in which the revenues gain starts to saturate is higher if the total number of calls is smaller. To illustrate, for a total number of 60 calls

the saturation in the revenues gain starts at 0.2% of Premium class calls, while for 24 total calls starts at 0.6% of Premium ones.

4.8.2.1 Confidence Intervals

In order to prove that the shown results are reliable, confidence intervals were also calculated. The confidence intervals are computed for a confidence level of 99%. To illustrate, the thesis considers the revenues calculated in Fig. 4.19 for a total number of 40 calls when 40%, 70%, and 90% of Premium class calls are loaded. For 40% of Premium class calls, the confidence intervals are 2.468 ± 0.062, while for 70% are 2.273 ± 0.034. Finally, for 90% of Premium class calls confidence intervals are $2.404 \pm 1.034 \cdot 10^{-14}$. Note that these intervals are calculated by iterating the simulation 100 times. The calculated confidence intervals indicate that the algorithm performs well in term of providing similar solutions in different iterations.

4.8.3 Revenues per User

This sub-section is focused on providing results on the revenues per users gained by the service provider. Fig. 4.20 represents the revenues versus an increasing number of calls in case the percentage of Premium and Economy class calls is fixed to 10% and 90%, respectively. In scenarios 1 and 2 (see tariffs in Section 4.5), since the filtering levels of Premium class traffic flows are not reduced, Premium class users pay on average $ 0.16 during a control period. If more than 10 Premium class users are admitted, there is a significant decrease in the revenues per Premium class user, due to the filtering reduction (see Fig. 4.11). This leads to a reduction in the total revenues of the operator. On the other hand, the revenues per Economy class user are equal to $ 0.02 when the filter is still not activated (from 0 to 40 calls).

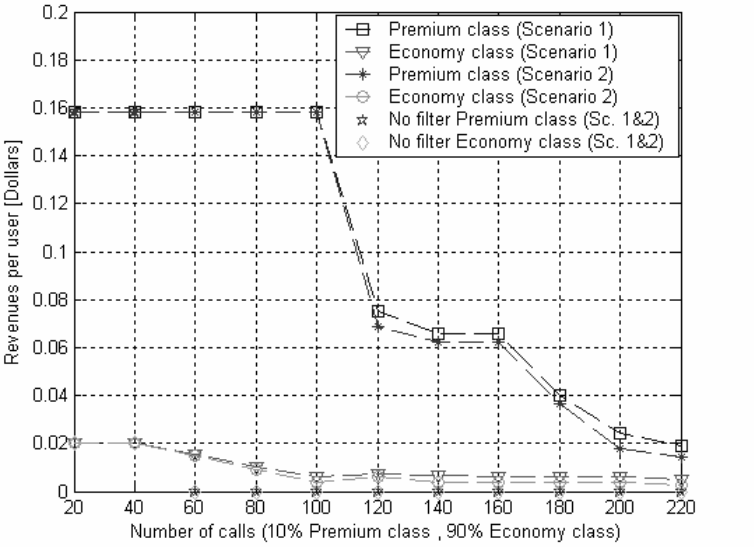

Fig. 4.20 Revenues per Premium and Economy class user on average in a control period versus admitted calls (10% Premium class, 90% Economy class) for a node with and a node without filters (Tariffs in Scenarios 1 & 2, Section 4.5).

If the link becomes overloaded (more than 40 calls), the average filtering levels of Economy class traffic flows are reduced and the revenues per Economy class user are lowered. In the range between 40 and 100 calls, although the revenues per Economy class user decrease as more users are admitted, the revenues per Premium class user do not change, leading to an increase in the total revenues of the operator because Premium users pay a higher tariff.

4.9 Experiment 3

This experiment is considered in order to verify the performance of the optimization algorithm in terms of reducing the number of calls that can be dropped due to user mobility. The experiment is illustrated in Fig. 4.21. It is assumed that the SGSN has three output links in the downstream path. Each link is used for downloading content to a group of Premium and Economy mobile users in a specific coverage area. Each user requests a multimedia call containing all types of media: voice, video and data. The users

move between neighboring coverage areas (inter-RNC handover). Therefore, the number of calls loaded in each output link changes in time according to the current location of the users.

When the users move to the next coverage area, their multimedia calls can be dropped if the available bandwidth is not enough to satisfy the constraints of the optimization function. In the experiment, it is assumed that the number of users that roam from one area to the next is modeled according to a Poisson distribution. In particular, Premium class users roam at an average rate of 0.1 users per second, while Economy class users roam at an average rate of 0.3 users per second.

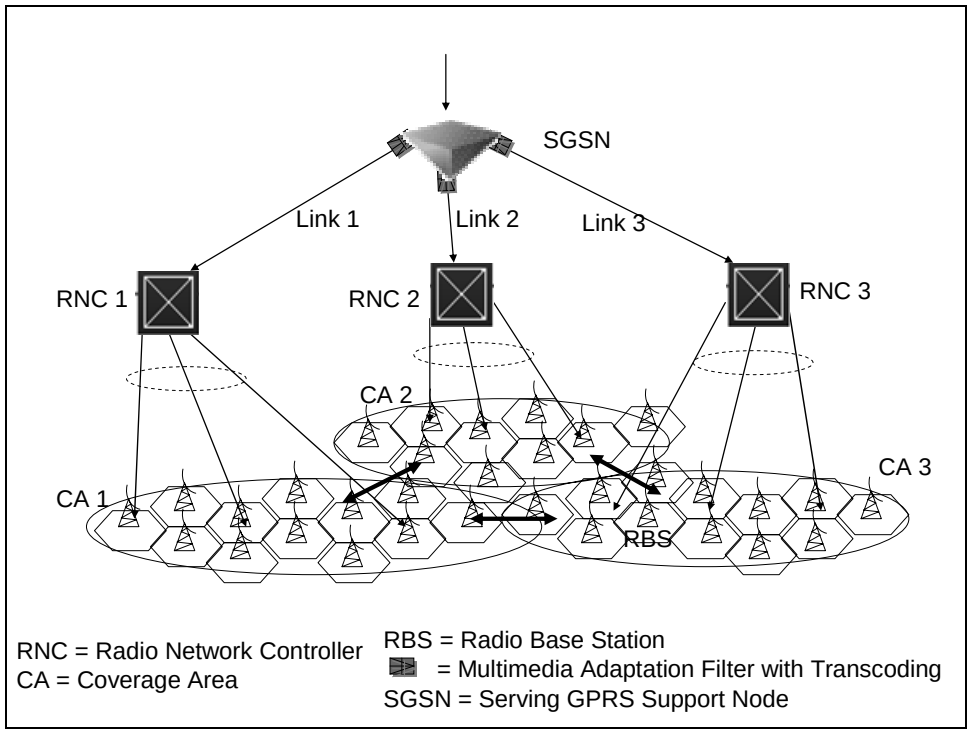

Fig. 4.21 Network architecture for inter-RNC handover

It is assumed an initial number of users in each coverage area to be 40 users (4 Premium, and 36 Economy). Our choice of users corresponds to a heavy traffic load (80

%) in case without filtering and to a medium traffic load (40 %) in case with filtering. The simulation time was 2000 sec.

4.9.1 Call Dropping Rate

Fig. 4.22 shows the total number of calls currently loaded in the three links against the time, for a node with filtering and a node without filtering. As shown in the graph, some Premium and Economy class calls are dropped during the simulation time due to the users' mobility. For a node without filters, 6 Premium and 68 Economy class calls are dropped. On the contrary, for a node with filters, only 1 Premium and 32 Economy class calls are dropped. Hence, the filtering algorithm decreases the number of calls dropped when users roam from one area to another. Fig. 4.23 represents the call dropping rate for mobile users versus simulation time.

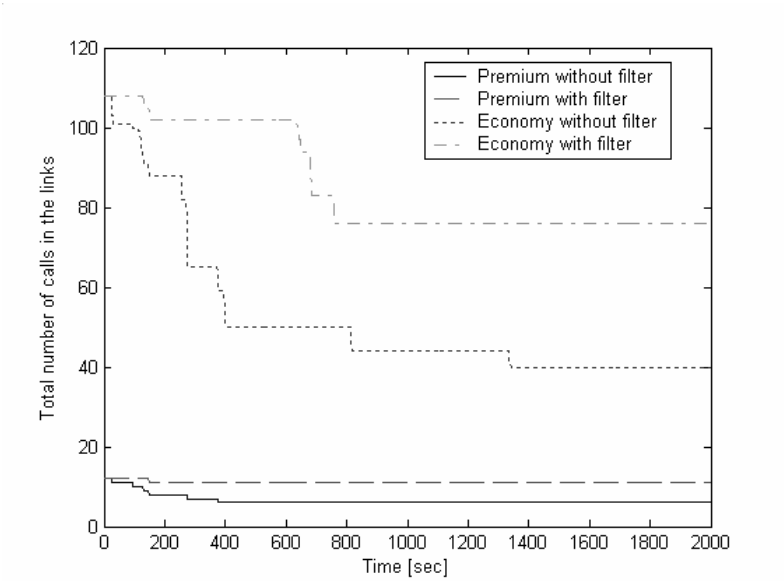

Fig. 4.22 Total number of calls in the three links versus simulation time for a node with filtering and a node without filtering.

Call dropping rate for Premium and Economy class users is always maintained lower in the case of a node with filtering than that for a node without filtering.

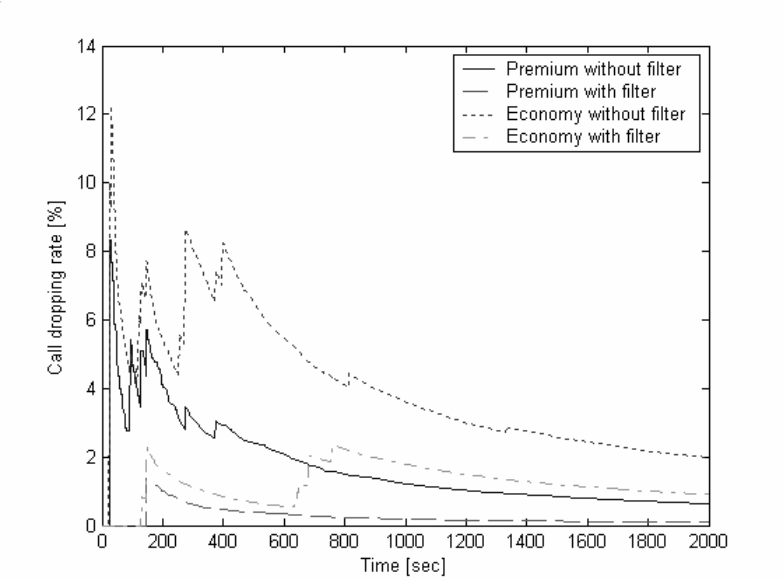

Fig. 4.23 Call dropping rate of Premium and Economy class users versus simulation time for a node with filtering and a node without filtering.

At the end of the simulation time, the call dropping rates for Premium and Economy users are 0.63% and 1.98% respectively, for a node without filtering. On the contrary, Premium and Economy class calls are dropped at rates of 0.09% and 0.91% respectively for a node with filtering. Hence, the filters in the core network reduce the probability of loosing multimedia services when users move between different radio network controllers (inter-RNC handovers).

4.10 Experiment 4

The experiment is to calculate the call blocking rates for Premium and Economy class users and to study the impact of the control period on the performance of the algorithm. In this experiment, each user requests a call for downloading a single type of media. Therefore, voice, video and data calls are generated separately. All types of media calls are generated with the same percentage (33% voice, 33% video, and 33% data) according to a Poisson distribution. Moreover, the simulations have been performed

using 10% Premium class calls and 90% Economy class calls. This classification represents a realistic scenario in which most users are indeed Economy class subscribers. Each call lasts 10 minutes, whereas the control period is set to 1 sec. A new call is admitted if constraints a), b), c), and d), previously defined in Section 4.2, are satisfied for the existing calls as well as for the new ones. In the case that both Premium and Economy class calls arrive simultaneously, then the Premium class calls have a higher priority for admission.

4.10.1 Call blocking rate

Fig. 4.24 compares the call blocking rates between the case of no-filtering (Case A), and that with filtering (Case B).

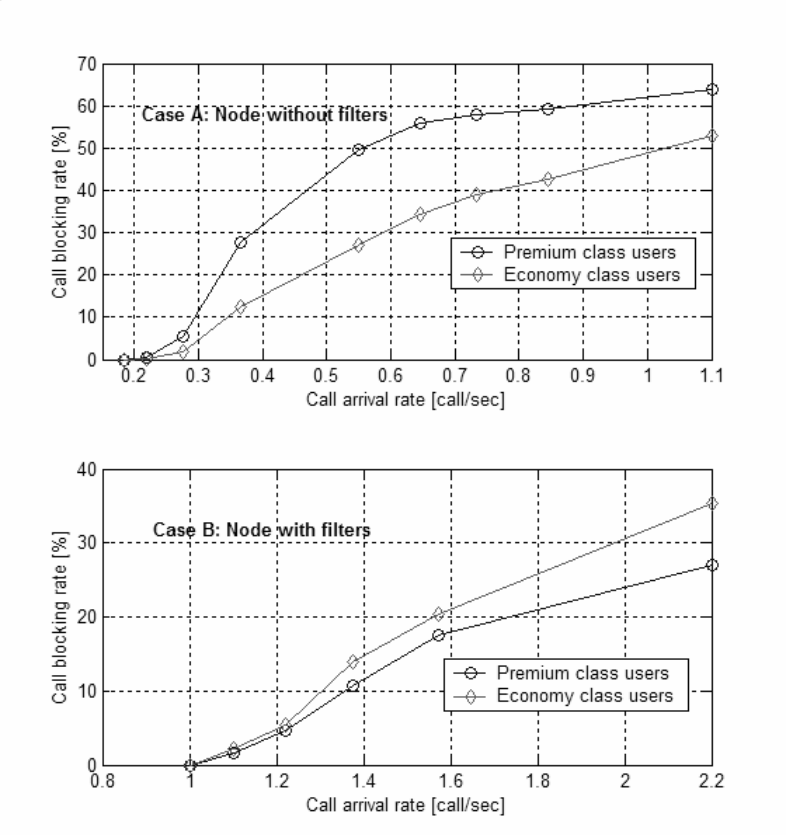

Fig. 4.24 Call blocking rate for Premium and Economy class users against call arrival rate. (10% Premium class, 90% Economy class).

As shown in the figure, call blocking rates are determined separately for each of the Premium and Economy class users against the call arrival rates (10% Premium class, 90% Economy class). Two important results are obtained in this experiment. First, the call blocking rates for either the Premium or the Economy class users are lower in case B than those of case A. To illustrate, consider a target call blocking rate of 1% for Premium class users, the correspondent calls' arrival rate is approximately 1.03 calls per second for the filtering case. At this arrival rate the call blocking rate for Premium class users, without filtering, rises approximately to 60%. Similarly, if it is considered a target call blocking rate of 5% for Economy class users, the correspondent calls' arrival rate is 1.2 calls per second with filtering. At this arrival rate the call blocking rate of Economy class users, without filtering, is more than 55%.

Secondly, the call blocking rate of Premium class users is higher than that of the Economy class users in case A, while it is lower in case B. To illustrate, at an arrival rate of 1.1 calls per second, about 63% of Premium class calls and 53% of Economy class calls are blocked without filtering. With filtering, 1.7% of Premium class calls and 2.1% of Economy class calls are blocked. Therefore, filtering ensures that the call blocking rate for the Premium class users is indeed lower than that of the Economy class users. In case of filtering, Premium class users will pay more for their service but their calls have a higher priority of admission over Economy class users.

4.10.2 Impact of the Control Period

Fig. 4.25 represents the call blocking rate with filtering for both Premium and Economy class users versus the control period (Δt). To illustrate, the call blocking rate is calculated for an arrival rate of 1.1 calls per second. Fig. 4.25 indicates that the call

blocking rate of Premium class users decreases and the call blocking rate of Economy class users increases with the increase in the control period. For large control periods, the new arrival calls have to wait for the algorithm to solve the admission criteria in the next control period. During this interval, several Premium and Economy class calls can both request admission to the network. The probability to block Economy class calls is, of course, higher than that of the Premium class calls. For small control periods, the algorithm is activated more frequently and is thus able to solve the optimization problem in a more prompt fashion.

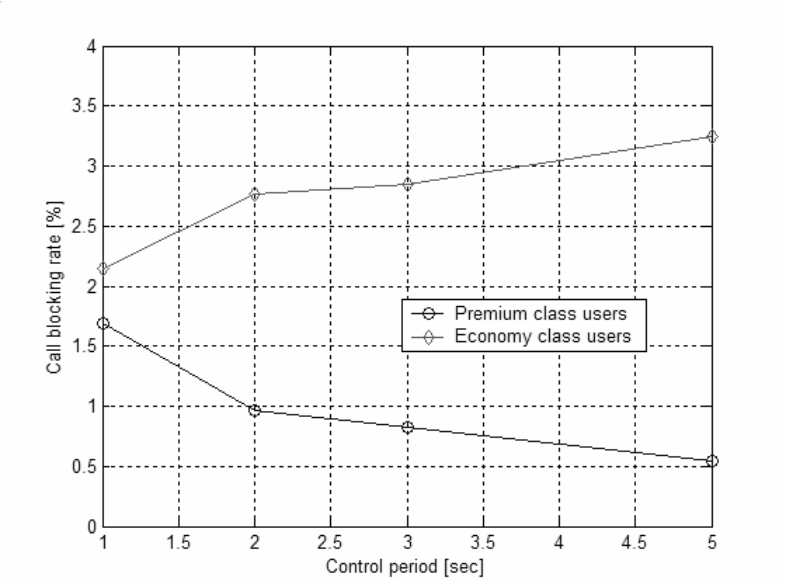

Fig. 4.25 Call blocking rate for Premium and Economy class users against control period at an arrival rate of 1.1 calls per second. (10% Premium class, 90% Economy class).

In this scenario, Premium class calls are blocked with higher percentages because some Economy class calls have been accepted in the previous control periods.

Chapter 5

A Filtering Algorithm in the Radio Access Network Nodes

This chapter aims at constructing an adaptive algorithm that controls the multimedia adaptation filters located in the nodes of the radio access network. The chapter begins by presenting the functional model of the radio access network. This includes two main components: the filter, which is implemented at the output ports of each RNC, and the air interface rate schedulers, which are included in the base station and user equipments. Two different methods are, then, illustrated to allocate resources in the W-CDMA air interface: variable spreading factor and multi code. Following this, the multi-objective optimization function is defined, in order to control the actions of both the filter and the rate schedulers. The optimization function defined in the previous chapter (see Section 4.2) is not appropriate to control the filters in the access nodes because it does not directly take into account the major features of the wireless environment (e.g., interference, shadowing, or attenuation) and is not able to select separately filtering parameters (the coding rates and formats) for the media streams downloaded to distinct users. In fact, in the radio access network, a traffic flow represents the media stream (e.g. voice or video or data streams) that is downloaded to a single user that has subscribed to a specific grade of service (Premium, or Economy). Therefore, a filter located in the

output ports of the access network nodes is able to manipulate in a different way the media streams directed to the mobile users.

The last sections of the chapter are focused on describing the experiments that have been performed and the simulation results. A microcellular pedestrian model emulates the wireless environment. To test the algorithm, a radio network controller without filtering capabilities is compared with other with filtering.

5.1 Radio Access Network with Filtering Capabilities

This section presents the functional model of the radio access network with filtering capabilities. In the simplified network shown in Fig. 5.1, an RNC is connected to a reference base station BS_0, and the *k*th mobile user is attached via an UE to the reference base station. Different traffic flows are downloaded from the RNC to the mobile users connected to BS_0. To illustrate, the figure depicts the *j*th traffic flow $[T_j]$, which contains the media α ($\alpha \in \{s, v, d\}$) downloaded via the base station BS_0 to the *k*th user (*k=1, 2, 3, …, N*). Once all routing operations are performed within the RNC, the traffic flow $[T_j]$ is directed to the appropriate output port, which is connected to the base BS_0. The port of the RNC connected to BS_0 includes, mainly, the proposed filter, a *Forward Error Control* (FEC) block and a set of transmission buffers. The base station BS_0 and the equipment of the *k*th user include rate schedulers for spreading and de-spreading operations and antennas for power transmission. The FEC block is used to introduce controlled redundancy in the original traffic flow in order to improve the reliability (target BER) of the wireless communication system.

The RNC, also, includes a *Radio Resource Manager* (RRM) that controls the operations of the filter and the assignment of the radio resources. In particular, the RRM is composed by two different modules: the control module and the data collection module. The control module instructs the filter to select specific filtering parameters (coding rate and format) for each downloaded traffic flow and, at the same time, manages the air interface resources by assigning spreading codes and transmitted powers to each user. The values of the parameters (e.g., selected coding rates, coding formats, spreading codes, and power levels) that are sent by the control module to control the filter and the rate schedulers are computed by optimizing our proposed multi-objective function that will be described in Section 5.1.

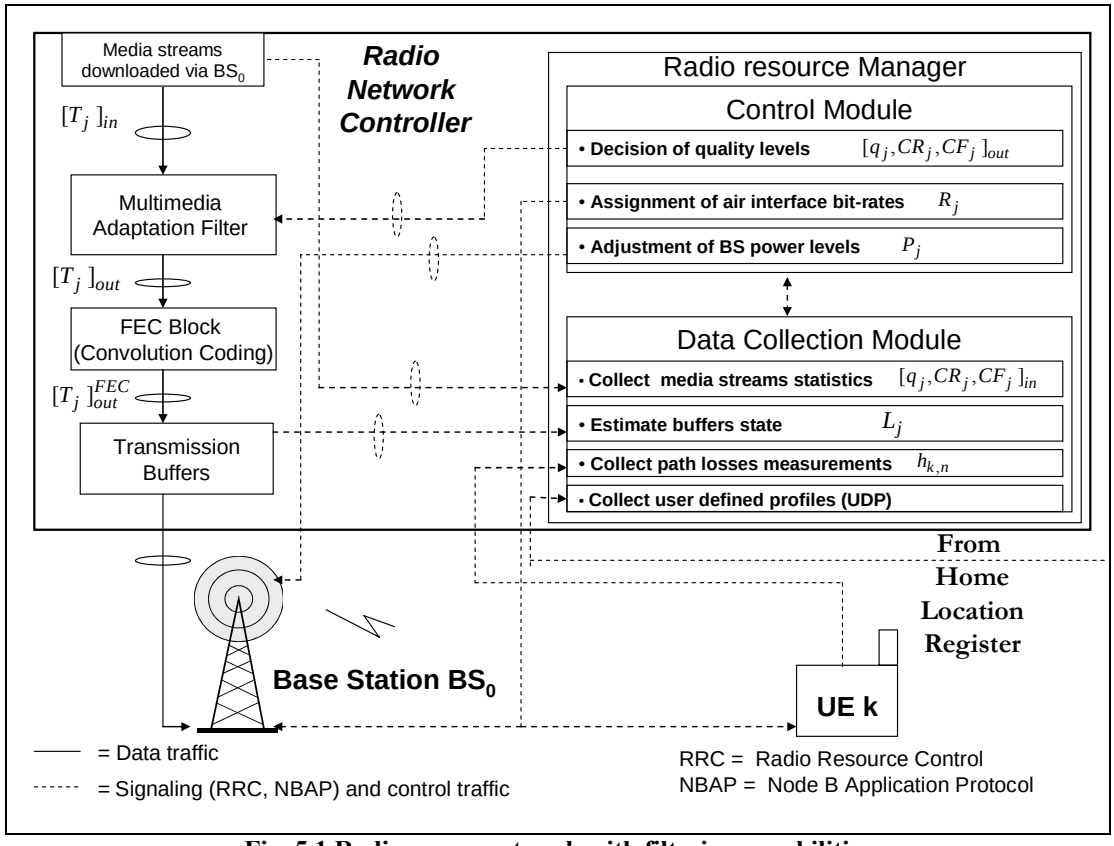

Fig. 5.1 Radio access network with filtering capabilities

The information (e.g., traffic flows statistics, buffers state, path losses, and users' defined profiles) needed to solve the optimization function is gathered by the data collection module. As shown in Fig. 5.1, the traffic flow directed to the output port $[T_j]_{in}$ is associated with a filtering vector $[q_j, CR_j, CF_j]_{in}$. This information is gathered by the data collection module of the RRM. During a certain time, the filter manipulates the *j*th flow by assigning, if it is necessary, a new output vector

$$[q_j, CR_j, CF_j]_{in} \Rightarrow [q_j, CR_j, CF_j]_{out}. \quad \textbf{(32)}$$

After the filtering operation, the coding rate of the output traffic flow $[T_j]_{out}$ is, then, further adjusted by adding redundant bits. The resultant traffic, which includes the redundant bits added by the FEC block, is denoted $[T_j]_{out}^{FEC}$. Since the FEC scheme is predefined and the number of redundant bits is fixed, the assignment of the coding rate $(CR_j)_{out}$ for each flow determines the channel coding rate $(r_j)_{out}$ for the traffic $[T_j]_{out}^{FEC}$. The relationship between coding rate and channel coding rate is basically given by [63]

$$(CR_j)_{out} = \frac{m}{n} \cdot (r_j)_{out} \quad (n > m) \quad \textbf{(33)}$$

where *n*-*m* represents the number of redundant bits added by the FEC block to each transmitted block *n*. Once redundant bits are added, packets are segmented into *Transport Blocks* (TBs) and loaded in the transmission buffers. If spreading codes and power levels are available, one or more TBs are then encapsulated in the radio frame and sent during a *Time Transmission Interval* (TTI) from the reference base station to the mobile user. Section 5.2 will explain how spreading codes and power levels can be assigned by the

radio resource manager in order to satisfy a targeted BER. The following sub-section shows a possible assignment of coding rates and formats for the media downloaded to Premium and Economy class users.

5.1.1 A Possible Range of Quality Levels and Traffic Models

The choice of a specific quality level with its corresponding coding rate, and type of coding format that can be chosen for the *j*th traffic flow depends on the type of media contained in the flow and the type of user to which the flow is directed. Table 5.1 shows a possible assignment of quality levels, coding rates, and formats for traffic flows downloaded to Premium and Economy class users via the radio access network.

Table 5.1 Possible combinations of quality levels, coding rates and formats for different media of Premium and Economy class traffic flows downloaded via the radio access network

Voice			Video			Data		
q_j	CR_j (kbps)	CF_j	q_j	CR_j (kbps)	CF_j	q_j	CR_j (kbps)	CF_j
Economy class flows								
1	0, 1.95	AMR	1	μ=16 σ=1	H.263	1	μ=8	GIF, XML
2	0, 4.75	AMR	2	μ=32 σ=2	MPEG-4	2	μ=16	GIF, XML
3	0. 7.95	AMR	3	μ=64 σ=4	MPEG-4	3	μ=32	GIF, XML
4	0, 18.25	AMR-W	4	μ=128 σ=8	MPEG-4	4	μ=64	JPEG, XHTML
Premium class flows								
2	0, 4.75	AMR	2	μ=32 σ=2	MPEG-4	2	μ=16	GIF, XML
3	0. 7.95	AMR	3	μ=64 σ=4	MPEG-4	3	μ=32	GIF, XML
4	0, 18.25	AMR-W	4	μ=128 σ=8	MPEG-4	4	μ=64	JPEG, XHTML
5	0, 23.85	AMR-W	5	μ=256 σ=16	MPEG-4	5	μ=128	JPEG, XHTML

This information defines the *User Defined Profile* (UDP) and is loaded in the data collection module (see Fig. 5.1) whenever a user initiates a call. As expected, mobile

users that subscribe to Premium class services receive multimedia content with a higher range of quality levels. As shown in the table, quality levels for the media directed to Economy class users range from 1 to 4, whereas those downloaded to Premium ones range from 2 to 5. It is assumed that each type (voice, video, and data) of traffic flow that is downloaded in the radio access network is generated according to a specific traffic model.

The traffic generated for voice traffic flows is simulated according to the ON-OFF model [64]. In the ON state, traffic is produced at the peak rate of the codec; no traffic is generated in the OFF state. The time intervals (in T_{voice} units = 20ms) spent in ON and OFF states are exponentially distributed with p and q mean residence times, respectively.

The traffic generated for a video traffic flow is thought as the aggregated of M independent ON-OFF mini-sources [65]. A mini-source in the ON state produces traffic at the constant rate of V bit/s; no traffic is generated in the OFF state. The time intervals (in T_{video} units = 40 ms) spent in ON and OFF states are geometrically distributed with p and q mean residence times, respectively. Parameters p, q, and V of each mini-source are obtained as

$$p = \frac{1}{aT_{video}}\left(1 + \frac{\mu^2}{M\sigma^2}\right) \quad \textbf{(34)}$$

$$q = \frac{1}{aT_{video}}\left(1 + \frac{M\sigma^2}{\mu^2}\right) \quad \textbf{(35)}$$

$$V = \frac{\mu}{M} + \frac{\sigma^2}{\mu}$$

(36)

Finally, a data traffic source oscillates between two possible states: *packet call* and *reading time* [66]. In the packet call state, a source produces a number of packets geometrically distributed with mean value N_d and the message inter-arrival time is geometrically distributed with mean value T_{pk}. In the reading time state (length geometrically distributed with mean value $T_{reading}$), no traffic is generated. The source activity factor is:

$$\phi = \frac{N_d T_{pk}}{N_d T_{pk} + T_{reading}}$$

(37)

Each message has a length in bytes l_{w_byte} following a Pareto normal distribution with cut-off. The average packet length and coding rate are given respectively by

$$\overline{l_{w\_byte}} = \frac{\alpha_{sh} \cdot l_{w\_byte\_\min} - l_{w\_byte\_\max}\left(\frac{l_{w\_byte\_\min}}{l_{w\_byte\_\max}}\right)}{\alpha_{sh} - 1}$$

(38)

$$\mu = \frac{8\overline{l_{w\_byte}}\phi}{T_{pk}}$$

(39)

where α_{sh} is denoted shape factor.

Once a specific quality level is selected for the traffic flow to be downloaded, the traffic flow is generated according to the specific traffic model described above and the coding rate and coding format illustrated in Table 5.1.

To illustrate, if the quality level 4 is assigned for a voice traffic flow directed to Premium class users, the coding rate will oscillates between 23.85 kbps and zero and the format will be W-AMR. In case the filter scales the quality level from 4 to 3, the voice traffic flow will oscillates between 18.25 kbps and zero. In the same way, if the filter selects the quality level 3 for a video traffic flow downloaded to Economy class users, the coding rate will have an average rate and a standard deviation of 64 and 4 kbps, respectively. The format will be MPEG-4.

5.2 Management of Air Interface Resources

Since recent 3G partnership projects [66, 67] indicate that W-CDMA is the prevailing air interface in 3G radio access cellular networks, the attention of this thesis is focused on such wireless environment. In order to allocate radio resources, W-CDMA technology has basically two different possible options: variable spreading factor and multi code methods. Basically, the VSF method allows the allocation of one spreading code to each user, whereas the MC method permits the assignment of multiple codes to a single user. Both methods are illustrated in the following two sub-sections.

The VSF method does not allow the allocation of different portions of bandwidth to distinct media downloaded to the same mobile user. However, since all simulations described in Section 5.6 are performed using the VSF method on the condition that each user download a single media (voice, video, or data users), this method is also described.

5.2.1 Variable Spreading Factor Option

In this case, the *k*th user downloads the media (voice, video, or data) included in the traffic flow *j*. In downlink VSF CDMA, the received bit energy to noise density ratio (E_b/N_0) for user *k* is given by [68]

$$\left(\frac{E_b}{I_0}\right)_k = \frac{P_k h_{k,0}}{(1-\alpha_k)(P_{tot,0}-P_k)h_{k,0} + N_0 W + \sum_{n=1}^{K} P_{tot,n} h_{k,n}} G_k \quad \textbf{(40)}$$

where $G_k = \frac{W}{R_k}$ is the spreading gain for user *k*, *W* is the system bandwidth while R_k is the air interface bit-rate to the user *k*. The air interface bit-rate is defined as

$$R_k = q\frac{R_c}{S.F} \quad \textbf{(41)}$$

where R_c is the chip rate (R_c = 3.84 Mcps), *S.F.* is the spreading factor and *q* depends on the modulation used. Since in UMTS networks a QPSK modulation is used in downlink, *q* is equal to 2.

The parameter $h_{k,n}$ represents the path loss including the shadowing effect from the base station BS_n to user *k*, α_k is the orthogonality factor (α_k = 1 is perfectly orthogonal), N_0 is the background noise density, and *K* is the number of the neighboring cells interfering to the reference cell 0. Moreover, P_k and $P_{tot,0}$ denote the required transmit power for user *k* and total transmission power from the reference cell BS_0, respectively. The relationship between P_k and $P_{tot,0}$ is given by

$$P_{tot,0} = \sum_{k=1}^{N} P_k$$

(42)

where N is the total number of users in the reference cell. It will be used γ_k to denote the required $\left(\frac{E_b}{I_0}\right)_{req,k}$ level. The other-cell interference factor is also defined as follows

$$\lambda_k = \sum_{n=1}^{K} \frac{P_{tot,n} h_{k,n}}{P_{tot,0} h_{k,0}}$$

(43)

It denotes the ratio of the inter-cell interference to the total power received from the own cell to user k. In this thesis, it is assumed that $P_{tot,0} = P_{tot,n}$ so that the inter-cell interference depends only on the path losses, including the shadowing effect.

In a W-CDMA system, power control is a fundamental way to reduce excessive inter and intra cell interferences and prolong battery life. As with most existing work, power control is assumed to be perfect: the assigned power to each user is adjusted to achieve exactly the required BER level. However, in practical systems, power control imperfections may occur and cause some misadjustment of received power.

From Equation (40) (see Appendix B), the minimum transmitted power to user k, in case γ_k is respected, is [69]:

$$P_k = \frac{g_k}{h_{k,0}} \frac{N_0 W}{1 - \sum_{k=1}^{N} g_k (1 - \alpha_k + \lambda_k)} \cdot \left[h_{k,0} (1 - \alpha_k + \lambda_k) \sum_{k=1}^{N} \frac{g_k}{h_{k,0}} + 1 - \sum_{k=1}^{N} g_k (1 - \alpha_k + \lambda_k) \right]$$

(44)

where $g_k = \frac{\gamma_k}{G_k + \gamma_k(1-\alpha_k)}$.

The air interface bit-rate R_k is the bit-rate that is allocated to download traffic via the radio channel *k*. In UMTS networks, two types of air interface channels are defined in downlink: dedicated and shared. Dedicated channels, which have a fixed spreading factor, are utilized for non-bursty traffic such as audio and video applications or for large file transfer application. On the contrary, shared channels are utilized for bursty traffic such as web browsing since they have dynamically varying spreading factor that is informed to the terminal on a TTI basis.

Since our objective is to effectively utilize the radio channel limited bandwidth by adapting the amount of traffic downloaded (i.e., selecting in every control period a specific level of quality), shared channels with variable spreading factors are utilized to download all types of traffic flows. The air interface bit-rate R_k and power levels P_k of the shared channels are allocated every control period (Δt) so that the following constraints are satisfied:

a) Throughput-based load factor: $\eta_{thr} = \sum_{k=1}^{N} g_k(1-\alpha_k+\lambda_k) < 1$;

b) Power-based load factor: $\eta_{pow} = \frac{P_{\max,0}}{P_{tot,0}} < 1$;

c) Number of available spreading codes: $\sum_{k=1}^{N} R_k \le n \cdot R_c$.

The control period defines the time period that is allotted to our proposed algorithm to find the values of bit rates and powers. It is composed by multiples TTIs. If the previous constraints cannot be satisfied for all radio channels, some user will not be

permitted to transmit information ($R_k = 0$ and $P_k = 0$) during that control period. Therefore, the RRM is in charge of deciding whether a specific user is allowed to transmit or not during this period and, in the former, which are the air interface bit rate (spreading factor) and power to be used for this transmission.

In case the *k*th user is allowed to transmit, the air interface bit-rate can be chosen among a set of values as follows:

$$R_k^{\min} < R_k < R_k^{\max}. \tag{45}$$

The minimum air interface bit rate, $R_k^{\min}$, is 30kbps [26] and corresponds to a maximum spreading factor $SF_k^{\max}$ of 256. On the contrary, the maximum bit-rate $R_k^{\max}$ ($SF_k^{\min}$) depends on the type of traffic flow and the class of user to which the traffic is directed.

Table 5.2 Admissible range of air interface rates or spreading factors for each traffic flow

Traffic flows	Min Bit Rate (kbps) R^{Min} Max Spreading Factor SF^{Max}	Max Bit Rate (kbps) R^{Max} Min Spreading Factor SF^{Min}
Premium class voice traffic flow	**R = 30, SF = 256**	**R = 120, SF = 64**
Economy class speech traffic flow	**R = 30, SF = 256**	**R = 60, SF = 128**
Premium class video traffic flow	**R = 30, SF = 256**	**R = 480, SF = 16**
Economy class video traffic flow	**R = 30, SF = 256**	**R = 240, SF = 32**
Premium class data traffic flow	**R = 30, SF = 256**	**R = 480, SF = 16**
Economy class data traffic flow	**R = 30, SF = 256**	**R = 240, SF = 32**

In our work, since the quality levels for the traffic flows directed to Premium class users are higher than those for downloading flows to Economy ones (see Table 5.1), $R_k^{\max}$ is expected to be higher for the radio channels allocated to Premium class users. Similarly, since video and data traffic flows consume higher bandwidth than voice flows, $R_k^{\max}$ is also expected to be higher for the radio channels allocated to download video and

audio packets. In Table 5.2, a possible range of admissible air interface rates or spreading factors is illustrated.

5.2.2 Multi Code Option

In a generic Multi Code CDMA, the received bit energy to noise density ratio (E_b/N_0), if the *i*th code is assigned to the user *k*, is given by [70].

$$\left(\frac{E_b}{I_0}\right)_i = \frac{P_i h_{k,0}}{(1-\alpha_i)(P_{tot,0}-P_i)h_{k,0} + N_0 W + \sum_{n=1}^{K} P_{tot,n} h_{k,n}} G \quad \textbf{(46)}$$

where $G = \frac{W}{R_b}$ is the processing gain, while R_b is the "basic" rate (30 kbps) of the multi code CDMA

The parameter $h_{k,n}$ represents the path loss including the shadowing effect from the base station BS_n to user *k*, α_i is the orthogonality factor of *i*th code (α_i = 1 is perfectly orthogonal), N_0 is the background noise density, and *K* is the number of the neighboring cells interfering to the reference cell 0. Moreover, P_i and $P_{tot,0}$ denote the required transmit power for *i*th code and total transmission power from the reference cell BS_0, respectively. The relationship between P_i and $P_{tot,0}$ is given by

$$P_{tot,0} = \sum_{i=1}^{M} P_i \quad \textbf{(47)}$$

where *M* is the total number of codes used by the reference cell. It will be used γ_i to denote the required $\left(\frac{E_b}{I_0}\right)_{req,i}$ level. We also define the other-cell interference factor as follows

$$\lambda_i = \sum_{n=1}^{K} \frac{P_{tot,n} h_{k,n}}{P_{tot,0} h_{k,0}}$$

(48)

It denotes the ratio of the inter-cell interference to the total power received from the own cell to *i*th code. To simplify, $P_{tot,0} = P_{tot,n}$ so that the inter-cell interference depends only on the path losses, including the shadowing effect.

From Equation (46), the minimum transmitted power to transmit *i*th code to the user k, in case γ_i is respected, is:

$$P_i = \frac{g_i}{h_{k,0}} \frac{N_0 W}{1 - \sum_{i=1}^{M} g_i (1 - \alpha_i + \lambda_i)} \cdot \left[(1 - \alpha_i + \lambda_i) \sum_{i=1}^{M} g_i + 1 - \sum_{i=1}^{M} g_i (1 - \alpha_i + \lambda_i) \right]$$

(49)

where $g_i = \frac{\gamma_i}{G + \gamma_i (1 - \alpha_i)}$.

The air interface bit-rate R_b and the power level P_i are allocated to the user *k*, if *i*th code is assigned. The total air interface rate that can be allocated to the user *k* for downloading the media α is defined as follows

$$R_{\alpha,k} = m_{\alpha,k} R_b$$

(50)

where $m_{\alpha,k}$ is the number of codes assigned to the user *k* to download the same media. Therefore, in the multi-code CDMA, multiple codes and power levels are assigned to the different users every control period (Δt) so that the following constraints are satisfied:

a) Throughput-based load factor: $\eta_{thr} = \sum_{i=1}^{M} g_i(1-\alpha_i+\lambda_i) < 1$;

b) Power-based load factor: $\eta_{pow} = \frac{P_{\max,0}}{P_{tot,0}} < 1$;

c) Number of total codes: $\sum_{k=1}^{N} \sum_{\alpha=s,v,d} m_{\alpha,k} \le M_{\max}$.

In case the kth user is allowed to transmit, the air interface bit-rate can be chosen among a set of values as follows:

$$R_{\alpha,k}^{\min} < R_{\alpha,k} < R_{\alpha,k}^{\max}. \quad \textbf{(51)}$$

The minimum air interface bit rate, $R_{\alpha,k}^{\min}$, is R_b and corresponds to the assignment of one code. On the contrary, the maximum bit-rate $R_{\alpha,k}^{\max}$ corresponds to the assignment of $m_{\alpha,k}$ codes and depends on the type of traffic flow and the class of user to which the traffic is directed, as illustrated in Section 5.2.1.

5.3 The Optimization Function in a Radio Access Network Node

The control module of the radio resource manager, in Fig. 5.1, is used to control the actions of both the filter and the air interface rate schedulers for downloading the traffic flows from the RNC to the mobile users connected to the reference base station BS_0.

This section defines the optimization function that is used to calculate the values for the filtering parameters, air interface rates and power levels that are sent from the control module to the filter and rate schedulers, respectively. The optimization function operates on a control period (Δt). The shorter the control period, the finer the control is over the resources of the wireless network, but at the expense of increasing the processing complexity of the function. The control period is composed by multiple Time Transmission Intervals. In each TTI, radio frames are sent from the reference base station to the mobile users if codes and power levels are assigned.

As explained in the Section 5.1, the *j*th traffic flow is characterized by the media stream α ($\alpha \in \{s, v, d\}$) that is downloaded to the *k*th user (k=1,2,.., N). In this section, the parameters α and k are used to characterize a flow in order to explain the operations performed on each distinct flow. The incoming traffic flow $[T_{\alpha,k}]_{in}(t)$ is characterized by the vector $[q_{\alpha,k}, CR_{\alpha,k}, CF_{\alpha,k}]_{in}(t)$. The actions performed by the optimization function determine after a predefined control period, Δt the output traffic flow $[T_{\alpha,k}]_{out}(t+\Delta t)$ and the air interface rate $R_{\alpha,k}(t+\Delta t)$.

The output traffic flow is characterized by the vector $[q_{\alpha,k}, CR_{\alpha,k}, CF_{\alpha,k}]_{out}(t)$, whereas the air interface rate is associated to the correspondent power level $P_{\alpha,k}(t+\Delta t)$ as shown in Section 5.2. The data collection module gets all information required for the operation of the function. This information is given by the average size $\overline{L_{\alpha,k}}(t-\Delta t, t)$ of each buffer in the previous control period, the maximum coding rate of each input media stream in the previous control period, the measurements of path losses for each mobile user and user defined profiles (UDP). A UDP, which is stored in the Home Location

Register, contains the range of quality levels, coding rates and formats represented in Table 5.1 that can be assigned for each media-type directed either to Premium or Economy class users. It also contains information about the QoS parameters' constraints such as maximum transfer delay $(\Delta_{\alpha,k}^{\max})$ and percentages of packets lost or delayed.

Assuming an MC method, our objective is to find the output traffic flows, the air interface rates and powers for traffic flows downloaded via the reference base station BS_0 such that the constraints a), b), and c) defined in Section 5.2.2 are satisfied and the following multi-objective function F is maximized:

$$F = F_Q + F_R + F_\Delta \tag{52}$$

The first component of the objective function, F_Q, controls the assignment of the multimedia quality levels $(q_{\alpha,k})_{out}(t+\Delta t)$, and associated coding rate $(CR_{\alpha,k})_{out}(t+\Delta t)$, and format $(CF_{\alpha,k})_{out}(t+\Delta t)$ for each media α downloaded to the user k.

This function is represented by

$$F_Q = \frac{1}{6(N_P + N_E)}\left(\sum_{k=1}^{N_P}\sum_{\alpha=s,v,d} F_Q^{\alpha,k,P} + \sum_{k=1}^{N_E}\sum_{\alpha=s,v,d} F_Q^{\alpha,k,E}\right) \tag{53}$$

where

$$F_Q^{\alpha,k,P} = \begin{cases} \left[\dfrac{(q_{\alpha,k})_{out}(t+\Delta t)}{(q_{\alpha,k})_{in}(t)}\right]^2, & if \quad (q_{\alpha,k})_{\min} \le (q_{\alpha,k})_{out}(t+\Delta t) \le (q_{\alpha,k})_{in}(t) \\ 0, & otherwise, \end{cases} \tag{54}$$

$$F_Q^{\alpha,k,E} = \begin{cases} \sqrt{\dfrac{(q_{\alpha,k})_{out}(t+\Delta t)}{(q_{\alpha,k})_{in}(t)}}, & if \quad (q_{\alpha,k})_{\min} \le (q_{\alpha,k})_{out}(t+\Delta t) \le (q_{\alpha,k})_{in}(t) \\ 0, & otherwise, \end{cases}$$

(55)

In Equation (53), N_P, and N_E denote the number of Premium and Economy class users, located within BS_0 during the control period, respectively. The choice for the quality levels is determined in a different way for traffic flows downloaded to Premium and Economy class users. The function $F_Q^{\alpha,k,P}$ is quadratic and controls the output quality levels for the traffic flows downloaded to Premium class users. The function $F_Q^{\alpha,k,E}$ is radical and determines the quality levels for the Economy class traffic flows. Both functions encourage an output quality level that is as close as possible to the input quality level, but the strength that they utilize is different. The quadratic function is used for Premium class traffic flows because it encourages high quality levels more strongly than the radical function. Consequently, quality levels for Premium class traffic flows will be maintained higher than those for Economy class traffic flows.

The second component F_R, controls the assignment of the air interface bit-rates to download media to different users. The function is given by

$$F_R = \frac{1}{6}\left(\frac{1}{N_P} \sum_{k=1}^{N_P} \sum_{\alpha=s,v,d} F_R^{\alpha,k} + \frac{1}{N_E} \sum_{k=1}^{N_E} \sum_{\alpha=s,v,d} F_R^{\alpha,k} \right)$$

(56)

where

$$F_R^{\alpha,k} = \begin{cases} \dfrac{R_{\alpha,k}(t+\Delta t)}{R_{\alpha,k}^{\max}}, & if \quad 0 \le R_{\alpha,k}(t+\Delta t) \le R_{\alpha,k}^{req}(t+\Delta t) \le R_{\alpha,k}^{\max} \\ 0, & otherwise. \end{cases}$$

(57)

In the equation (57), $R_{\alpha,k}^{req}(t+\Delta t)$ represents the air interface bit-rate that is requested in the next control period to download, in each TTI, TBs of the *j*th traffic flow with the delay target period $\Delta_{\alpha,k}^{\max}$. It is defined as follows

$$R_{\alpha,k}^{req}(t+\Delta t) = \max\left\{ R_{\alpha,k}^{req1}(t+\Delta t), .., R_{\alpha,k}^{reqh}(t+\Delta t), .., R_{\alpha,k}^{req\frac{\Delta t}{TTI}}(t+\Delta t) \right\}$$

(58)

where

$$R_{\alpha,k}^{reqh}(t+\Delta t) = \min\left(R_{\alpha,k}^{\max}, \frac{L_{\alpha,k}^{h-1} + z_j \cdot \max\{r_{\alpha,k}(t-\Delta t, t)\} \cdot TTI}{\Delta_{\alpha,k}^{\max}} \right) \quad 1 \le h \le \frac{\Delta t}{TTI}$$

(59)

$$L_{\alpha,k}^{h} = \begin{cases} \overline{L_{\alpha,k}}(t-\Delta t, t) & h = 0 \\ L_{\alpha,k}^{h-1} + \left(z_{\alpha,k} \cdot \max\{r_{\alpha,k}(t-\Delta t, t)\} - R_{\alpha,k}^{reqh}(t+\Delta t) \right) \cdot TTI & 1 \le h \le \dfrac{\Delta t}{TTI} \end{cases}$$

(60)

In equation (59), $R_{\alpha,k}^{reqh}(t+\Delta t)$ represents the bit-rate requested to transfer *TBs* in the *h*th *TTI* of the next control period. It is computed by dividing the *j*th buffer size estimated at the *h*th *TTI* of the next control period by the target delay period $\Delta_{\alpha,k}^{\max}$ [71]. The buffer size at the *h*th *TTI* depends on both the estimated buffer size $L_{\alpha,k}^{h-1}$ at the previous *(h-1)*th *TTI* and the estimated channel coding rate $r_{\alpha,k}(t+\Delta t)$ for the traffic flow $[T_{\alpha,k}]_{out}^{FEC}(t+\Delta t)$

in the next control period. It is assumed that the estimated rate is calculated by multiplying the maximum value of the channel coding rate in the previous control period, $\max\{r_{\alpha,k}(t-\Delta t,t)\}$, by a traffic coefficient denoted $z_{\alpha,k}$. The value assigned to $z_{\alpha,k}$ depends on the maximum delay tolerance of the media included in the traffic flow. The packet transfer delay is stringent for voice *TBs* ($z_{s,k}$ = 1), bounded for video *TBs* ($z_{v,k}$ = 0.8), and tolerable for data *TBs* ($z_{d,k}$ = 0.6). The choice of a high value for the traffic coefficient (e.g., $z_{\alpha,k}$ = 1) determines a higher estimated value for the bit-rate that is requested to download the *TBs*, thus reducing the probability of violating the delay requirements. Hence, an immediate result is that this parameter can be used to control the percentages of packets lost or delayed for each media-type.

As illustrated in equation (57), $F_R^{\alpha,k}$ encourages an air interface bit rate that is as close as possible to the requested one, regardless of the class of the user and the type of media. However, since N_P is expected lower than N_E in a realistic scenario, the term $\frac{1}{N_P}\sum_{k=1}^{N_P}\sum_{\alpha=s,v,d} F_R^{\alpha,k}$ supports the allocation of the requested bit-rates more strongly than the term $\frac{1}{N_E}\sum_{k=1}^{N_E}\sum_{\alpha=s,v,d} F_R^{\alpha,k}$. Therefore, the bit-rates that are requested to transfer media to Premium class users are more aggressively encouraged than those used to send media to Economy class users.

Moreover, as shown in equation (57), the value of $R_{\alpha,k}$ is encouraged to be as close as possible to the requested bit rate $R_{\alpha,k}^{req}(t+\Delta t)$ instead of the maximum predefined rate $R_{\alpha,k}^{\max}$. Two reasons justify this choice. Firstly, the function tries to avoid allocating

additional air interface bit rates that are not necessary to satisfy the targeted delay requirements, regardless of the class of the user and type of media. This is because the assignment of additional bit-rates to some users increases the interference power to other users and, then, the possibility to not satisfy the requested bit-rates of these users.

Secondly, this function attempts to encourage strongly the assignment of requested air interfaces bit-rates that are closer to the maximum allowable bit-rates. To illustrate, consider a user downloading a media under bad radio channel conditions. If the requested rate is not assigned, the number of *TBs* buffered from one control period to another increases, leading to an increase of the same requested rate. As the value of the requested bit rate approaches the maximum allowable rate, transfer delay requirements start to be violated. In this case, the algorithm encourages the assignment of the air interface rates for this radio channel more strongly than that for other radio channels so that the QoS requirements may be recovered as soon as possible.

The third component F_{Δ}, encourages a solution in which transfer delay to download *TBs* is less than the maximum transfer delay $\Delta_{\alpha,k}^{\max}$. It is given by

$$F_{\Delta} = \frac{N_E}{N_P} \sum_{k=1}^{N_P} \sum_{\alpha=s,v,d} F_{\Delta}^{\alpha,k} + \sum_{k=1}^{N_E} \sum_{\alpha=s,v,d} F_{\Delta}^{\alpha,k} \quad \textbf{(61)}$$

where

$$F_{\Delta}^{\alpha,k} = \begin{cases} 1 & 0 \le \Delta_{\alpha,k}(t+\Delta t) \le \Delta_{\alpha,k}^{\max} \\ 0 & otherwise \end{cases} \quad \textbf{(62)}$$

The transfer delay $\Delta_{\alpha,k}(t+\Delta t)$ is calculated as follows:

$$\Delta_{\alpha,k}(t+\Delta t)=\max\left(\Delta_{\alpha,k}^{1}(t+\Delta t),\ldots,\Delta_{\alpha,k}^{h}(t+\Delta t),\ldots,\Delta_{\alpha,k}^{\frac{\Delta t}{TTI}}(t+\Delta t)\right)$$

(63)

where

$$\Delta_{\alpha,k}^{h}(t+\Delta t)=\frac{L_{\alpha,k}^{h-1}+r_{\alpha,k}(t+\Delta t)\cdot TTI}{R_{\alpha,k}(t+\Delta t)}$$

(64)

$$L_{\alpha,k}^{h}=\begin{cases}\overline{L_{\alpha,k}}(t-\Delta t,t) & h=0\\ L_{\alpha,k}^{h-1}+\left(r_{\alpha,k}(t+\Delta t)-R_{\alpha,k}(t+\Delta t)\right)\cdot TTI & 1\le h\le\dfrac{\Delta t}{TTI}\end{cases}$$

(65)

In Equation (64), $\Delta_{\alpha,k}^{h}(t+\Delta t)$ represents the transfer delay to download *TB*s of the *j*th traffic flow at the *h*th *TTI* of the next control period. The transfer delay is estimated by dividing the *j*th buffer size at the *h*th *TTI*, $L_{\alpha,k}^{h}$, by the air interface bit-rate $R_{\alpha,k}(t+\Delta t)$ that is selected by the algorithm. The buffer size at the *h*th *TTI* depends on both the buffer size at the previous *TTI*, $L_{\alpha,k}^{h-1}$, and the amount of bits that arrive during the *h*th TTI. The amount of bits is determined by the filtering parameters selected by the algorithm. As shown in Equations (60) and (65), the initial buffer size $L_{\alpha,k}^{0}$ is estimated by averaging the buffer size in the previous control period over all *TTI*s.

The function $F_{\Delta}^{\alpha,k}$ encourages a solution in which the maximum transfer delay is respected, regardless of the type of user and type of media. Nevertheless, since in the

realistic case N_P is less than N_E, the term $\frac{N_E}{N_P}\sum_{k=1}^{N_P}\sum_{\alpha=s,v,d} F_{\Delta}^{\alpha,k}$ supports the fulfillments of the transfer delay requirements more strongly than $\sum_{k=1}^{N_E}\sum_{\alpha=s,v,d} F_{\Delta}^{\alpha,k}$.

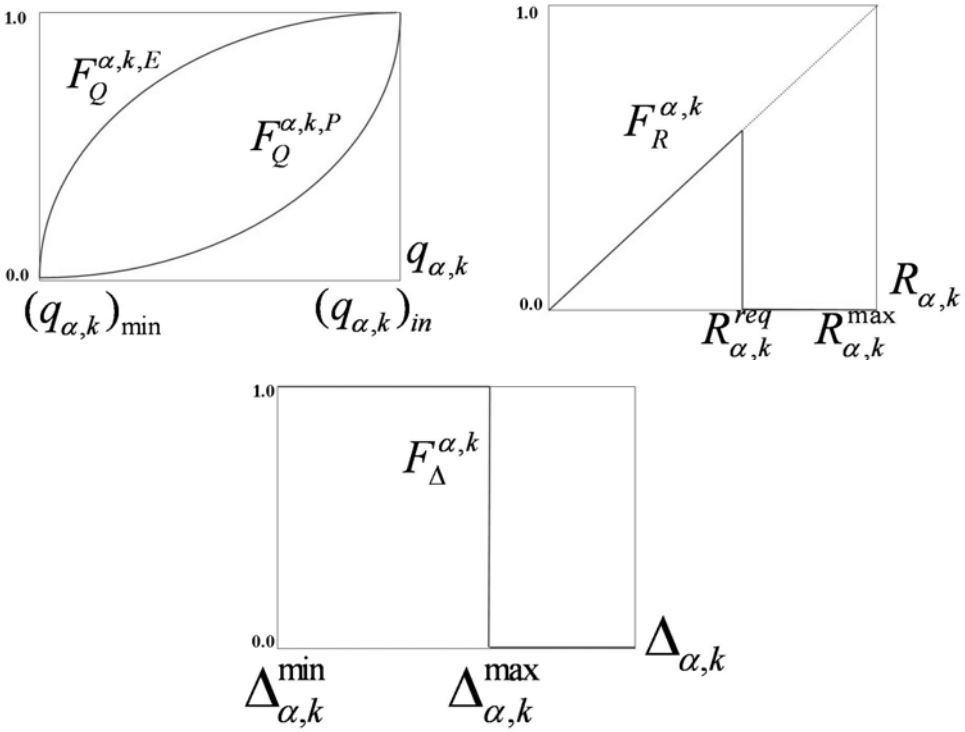

Fig. 5.2 Multi-objective function components

Fig. 5.2 is a graphical representation of the multi-objective components defined above. To perform the optimization, the Genetic Algorithms (GAs) are used as in the previous chapter. Details of the GAs are provided in the following section.

5.4 Genetic Algorithm (GA) Details

The main features and components of the Genetic Algorithms were comprehensively introduced and presented in Section 4.4. Therefore, this section will be focused on the description of the specific parts of the GA used to solve the optimization problem in the previous section.

The fitness function was composed in (52), so let us compose the chromosome. In this case, the chromosome is a vector whose elements are the quality levels of the output traffic flows and the air interface bit-rates to download the three types of media to all Premium and Economy class users, like this:

$$[q_{s,1}, R_{s,1}, q_{v,1}, R_{v,1}, q_{d,1}, R_{d,1}, ..., q_{s,j}, R_{s,j}, q_{v,j}, R_{v,j}, q_{d,j}, R_{d,j}, ...$$
$$...., q_{s,N_E+N_P}, R_{s,N_E+N_P}, q_{v,N_E+N_P}, R_{v,N_E+N_P}, q_{d,N_E+N_P}, R_{d,N_E+N_P}]$$

(66)

Recall that power levels are computed if the air interface bit-rates are assigned, as indicated in Section 5.2. The chromosome represents one point in the search space encoding one solution to our function *F*. In the GA, the fittest chromosomes of any population tend to reproduce and survive to the next generation, thus improving successive generations. In this case, the chromosome is represented by a binary vector and its length is $N_E \sum_{\alpha=s,v,d} (n_\alpha^{q,E} + n_\alpha^{R,E}) + N_P \sum_{\alpha=s,v,d} (n_\alpha^{q,P} + n_\alpha^{R,P})$, where $n_\alpha^{q,P}$, $n_\alpha^{R,P}$, $n_\alpha^{q,E}$, and $n_\alpha^{R,E}$ are the number of bits used to encode q and R for the media α downloaded to Premium and Economy class users, respectively. In the implementation of the GA, $n_\alpha^{q,P} = n_\alpha^{q,E} = 2$ bits are used for all media (4 quality levels are shown in Table 5.1). On the contrary, $n_\alpha^{R,P}$ and $n_\alpha^{R,E}$ depend on the type of media. For traffic flows with speech packets, $n_s^{R,P} = n_s^{R,E} = 2$ bits are used, for video and data traffic flows $n_v^{R,P} = n_v^{R,E} = n_d^{R,P} = n_d^{R,E} = 3$ bits (see the admissible values of air interfaces bit rates illustrated in Table 5.1). The complexity of the algorithm depends on the number of admitted users and the number of bits used to encode the two variables. Some results will be shown in Section 5.7.8.

The operation of GA is depicted in Fig. 5.3 and is summarized in Section 4.4. GA includes some operators like reproduction, crossover and mutation that are applied to successive populations of chromosomes to create new better ones. In our algorithm, it is used convergence as a stopping criterion. The convergence is reached if the fitness score of the *n*th previous best-of-generation does not exceed 1% of the fitness score of the current best-of-generation. In our simulation, n = 40. Note that initial population is generated randomly and contains 20 chromosomes. Once the convergence criterion is reached, the algorithm sends new values for the allocation of filtering parameters and air interface bit rates for the jth traffic flow to the filter, and the rate schedulers. These are the values that have maximized the objective function. They are maintained for the duration of one control period until a new cycle is ended.

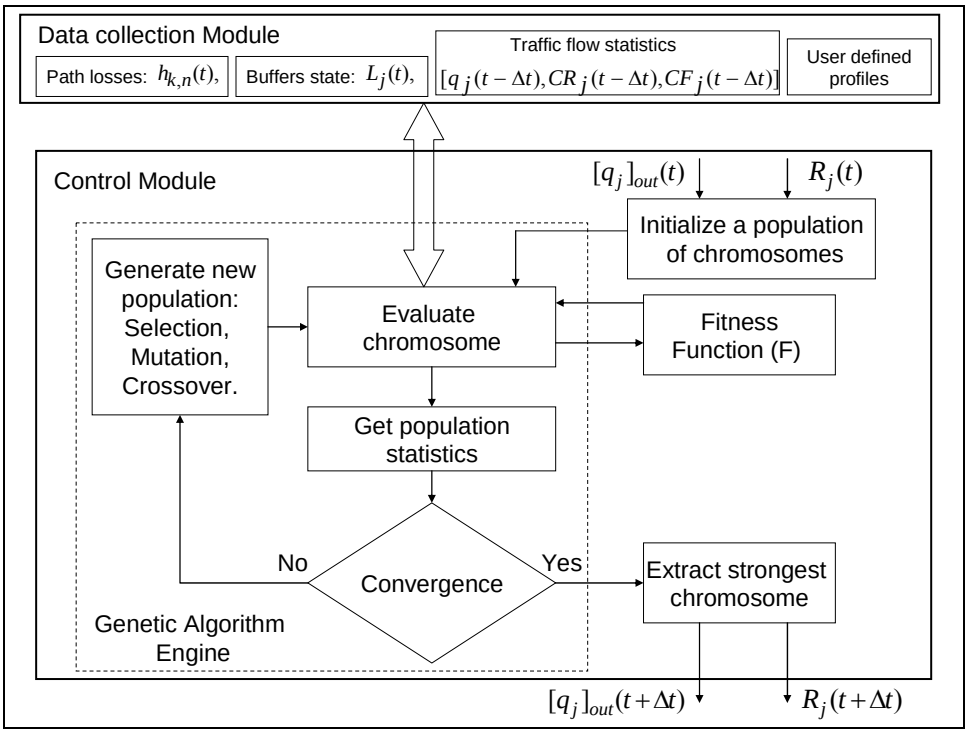

Fig. 5.3 Block diagram of GA

5.5 A Revenue Model and Pricing Mechanisms

This section describes the revenue model and the pricing policies that were used to evaluate the total revenues gained by the service provider when mobile users download content in the radio access network. In computing the revenues, it is assumed that the pricing mechanism takes into account different factors. Firstly, because the filters provide a preferential treatment to streams directed to Premium class users over those directed to Economy class ones, Premium calls users pay higher tariffs than Economy class ones. Secondly, because the grade of quality of downloaded media depends on the quality levels selected by the filter, users pay more if media streams are downloaded at higher quality levels. Thirdly, because the percentages of packets lost or delayed are also an important QoS measurement, users pay less if download media at higher percentages of packets lost or delayed. Finally, because the grade of quality may change every Δt, tariffs are periodically updated. Based upon these factors, we define a revenue model as follows

$$\Phi(\Delta t) = \sum_{k=1}^{N_P} \sum_{\alpha=s,v,d} T^P_{q_{\alpha,k}}(\Delta t) + T^P_{\chi_{\alpha,k}}(\Delta t) + \sum_{k=1}^{N_E} \sum_{\alpha=s,v,d} T^E_{q_{\alpha,k}}(\Delta t) + T^E_{\chi_\alpha}(\Delta t)$$

(67)

The revenue measure $\Phi(\Delta t)$ describes the revenues generated due to all media types downloaded by Premium and Economy class users via BS_0 during Δt. As shown in Equation (67), each Premium and Economy class user pays two different types of tariffs for downloading a media-type. The first type of tariffs $T^P_{q_{\alpha,k}}(\Delta t)$, $T^E_{q_{\alpha,k}}(\Delta t)$ is related to the quality levels $q_{\alpha,k}$ selected every Δt by the filter, as follows

$$T^P_{q_{\alpha,k}}(t) = \begin{cases} T^P_\alpha & \$, \quad q_{\alpha,k} = 2 \\ T^P_{q_{\alpha,k}-1}(t) \cdot (1+\alpha_T) & \$, \quad otherwise, \end{cases}$$

(68)

$$T^E_{q_{\alpha,k}}(t) = \begin{cases} T^E_\alpha & \$, \quad q_{\alpha,k} = 1 \\ T^E_{q_{\alpha,k}-1}(t) \cdot (1+\alpha_T) & \$, \quad otherwise, \end{cases}$$

(69)

where the coefficient α_T represents the percentage of tariff increment between two consecutive quality levels, whereas T^P_α and T^E_α are the tariffs that Premium and Economy class users pay to download the specific media (speech, video, or data) at the minimum quality levels (quality level 1 for Economy class user and quality level 2 for Premium class users). The tariff that users pay increases if the traffic flow is downloaded with a higher quality level. Since Premium class users download traffic flows at higher quality levels $T^P_\alpha > T^E_\alpha$.

The other type of tariffs $T^P_{\chi_\alpha}(\Delta t)$ and $T^E_{\chi_\alpha}(\Delta t)$ depends on the percentages of media packets lost or delayed $\chi^P_{\alpha,k}$ and $\chi^E_{\alpha,k}$ during Δt, as follows

$$T^P_{\chi_{\alpha,k}} = \begin{cases} \dfrac{\chi^P_{\alpha,\max} - \chi^P_{\alpha,k}}{\chi^P_{\alpha,\max}} T^P_\alpha & \$, \quad \chi^P_{\alpha,k} < \chi^P_{\alpha,\max} \\ 0 & \$, \quad \chi^P_{\alpha,k} \geq \chi^P_{\alpha,\max} \end{cases}$$

(70)

$$T^P_{\chi_{\alpha,k}} = \begin{cases} \dfrac{\chi^E_{\alpha,\max} - \chi^E_{\alpha,k}}{\chi^E_{\alpha,\max}} T^E_\alpha & \$, \quad \chi^E_{\alpha,k} < \chi^E_{\alpha,\max} \\ 0 & \$, \quad \chi^E_{\alpha,k} \geq \chi^E_{\alpha,\max} \end{cases}$$

(71)

where $\chi^P_{\alpha,\max}$ and $\chi^E_{\alpha,\max}$ are the maximum percentages of packets lost or delayed. As indicated in the equations above, users will pay the highest tariffs if the average

percentage of packets lost (delayed) is equal to zero. As expected, if the average percentage of packets lost (delayed) increases the tariff will decrease.

5.6 Radio Access Network Node Simulator

This section describes the simulations experiments carried out to evaluate the performance of the proposed algorithm in a node (RNC) of the radio access network. The simulator emulates an RNC connected to seven base stations: the reference one, BS_0, and other six ($K = 6$) neighboring ones. The base stations are centrally coordinated in cells that compose a hexagonal regular layout. Mobile users are distributed uniformly over each cell space. The simulator is written in C++ and the GALib tool described in [72], is used to solve the objective function defined in (52).

To test the algorithm, an RNC without filtering capabilities is compared with another with filtering. Different traffic flows (voice, video, or data) are downloaded from the RNC to the mobile users via the reference base station. In the case without filtering, two different scenarios are considered. In scenario 1, traffic flows are downloaded with the highest coding rates and formats described in Table 5.1 (quality levels 5 and 4 for Premium and Economy class traffic flows, respectively). In scenario 2, Premium and Economy class traffic flows are downloaded with coding rates and formats that correspond to quality levels 4 and 3, respectively.

In the filtering case, traffic flows are generated with the quality levels as in scenario 1, but the quality levels can be adaptively changed by our proposed filter. These traffic flows are generated with the abovementioned coding rates and formats according to the traffic models described in Section 5.1.1.

The main traffic parameters are illustrated in Table 5.3.

To simplify the simulations, it is assumed that the reference base station is allowed to assign one spreading code to each user using the VSF method described in Section 5.2.1.

Table 5.3 Traffic environment parameters

Time intervals for voice traffic flows (T_{units}= 20 ms)	**Exponential (ON: p=3 s), (OFF: q=3 s)**
Number of mini-sources for video traffic flows, *M*	**2**
Time intervals for video mini-sources (T_{units}= 40 ms)	**Geometric (ON: p in (34)), (OFF: q in (35)), (a=3.6 s^{-1})**
Time intervals in reading state for data traffic flows (T_{units}= 40 ms)	**Geometric ($T_{reading}$ = 3 s)**
Packet size for data traffic flows	**Pareto with cut-off ($\overline{l_{w\_byte}}$ in (38)), (α_{sh} =1.1)**
Number of packets in data traffic flows	**Geometric (N_d = 3000)**
Time intervals in packet call state for data traffic flows	**Geometric (T_{pk} = 0.04 s)**

Therefore, each user that initiates a call can request only a conversational (voice users), streaming (video users), or interactive service (data users) with a specific grade of service: Premium or Economy.

It is also assumed that the media types (audio, video, or data) included in a traffic flow are downloaded according to the QoS requirements illustrated in Table 5.4.

Table 5.4 QoS requirements

QOS PARAMETERS	VALUES
Targeted BER for voice traffic flows	**10-2**
Targeted BER for video traffic flows	**5 10-3**
Targeted BER for data traffic flows	**10-3**
Maximum delay for packets of voice traffic flows	**80 ms**
Maximum delay for packets of video traffic flows	**250 ms**
Maximum delay for packets of data traffic flows	**400 ms**
Max. percentage of lost voice packets (P)	**1.5 %**
Max. percentage of lost voice packets (E)	**3 %**
Max. percentage of lost video packets (P)	**5 %**
Max. percentage of lost video packets (E)	**10 %**
Max. percentage of delayed data packets (P)	**15 %**
Max. percentage of delayed data packets (E)	**30 %**

The main QoS parameters are maximum packet transfer delay, maximum percentage of packet lost (delayed), and targeted BER. In particular, packets for video and voice users are discarded if not transmitted within 250 ms and 80 ms [52], respectively. Instead,

packets directed to data users have a "virtual" deadline (considered equal to 400 ms [53]): if the packet is not transmitted within its deadline, it is transmitted anyway, but it is counted among those packets sent with an excessive delay. In the course of a simulation run, 40 simulation iterations are conducted. Statistics collected from the simulation are averaged per type (voice, video, and data) of user and grade (Premium and Economy) of service over all simulation iterations. The duration time for a simulation run is fixed to 120 sec., while 120 ms is the value for the control period of the genetic algorithm. Main system parameters are listed in Table 5.5.

Table 5.5 System parameters

Modulation Scheme	**QPSK**
W-CDMA spreading method	**Variable spreading coding**
Convolution coding for voice traffic flows	**1 / 3**
Convolution coding for data and video traffic flows	**1 / 2**
System Bandwidth, W	**5 MHz**
Chip Rate, R_C	**3.84 Mcps**
Time Transmission Interval	**10 ms**
Genetic Algorithm Control Period	**120 ms**
Simulation Length	**120 sec**
Number of Iterations	**40**

Firstly, the experiment is performed to test possible coexisting voice, data and/or video users that can be admitted in BS_0 without violating QoS requirements. As a realistic case, in all simulations the percentages of Premium and Economy services are maintained at 20 % and 80 %, respectively. First, we gradually load BS_0 with only voice users until the last combination of Premium and Economy class users is blocked. A combination of Premium or Economy class users is accepted if the average percentages of packets lost (delayed) are less than the maximum values defined in Table 5.4. As shown in the table, Premium class users expect to receive media with lower percentages of packets lost or delayed than those of Economy class users. To illustrate, the maximum percentage of

packets lost for downloading voice packets to Premium and Economy class users is 1.5 % and 3%, respectively. In Table 5.4, targeted BER are also represented. As expected, targeted BER for data packets must to be less than video and voice packets. Before showing the results of the experiments, which have been carried out in two different scenarios, the cellular model is described.

5.6.1 Microcellular Pedestrian Model

The experiments are performed on a microcellular environment. The coverage area is small and shaped by neighboring constructions. The main assumptions are relatively short radio paths (200m to 1000m), low base station antenna (3m to 10m).

The total path loss $h_{k,n}$ from the base station BS_n to user k is modeled as a product of two variables:

$$h_{k,n} = s_{k,n} \times a_{k,n},$$

(72)

where $s_{k,n}$ is the variation in the received power, due to shadow fading, and $a_{k,n}$ is the large scale propagation loss. In each control period, the shadowing variable $s_{k,n}$ is calculated according to the following formula:

$$s_{k,n} = C(d_k) \cdot s^*_{k,n} + \sqrt{1-\left(C(d_k)\right)^2} \cdot N(0,8)$$

(73)

where $C(d_k)$ is a spatial correlation coefficient, $s^*_{k,n}$ represents the shadowing calculated in the previous control period, d_k is the distance covered by the kth user during a control period, and N(0,8) is a Gaussian variable with mean zero and standard deviation

8. The spatial correlation coefficient $C(d_k)$ is defined as a function of the distance d_k covered by the kth user, as follows

$$C(d_k) = \exp\left(-\frac{d_k}{d_{corr}} \ln 2 \right) \tag{74}$$

To calculate $a_{k,n}$ a simplification of the empirical model in [73] is considered. The model was established through measurements taken in downtown San Francisco.

Table 5.6 Path-Loss formulas for high rise environment

ROUTE	PATH LOSS FORMULA
Line Of Sight (LOS)	$81.14 + 39.4\log(f_g) - 0.99\log(h_b) + (15.8 - 5.73\log(h_b))\log(D)$ $D < 0.3Km$ $(43.38 - 32.1\log(D)) + 45.7\log(f_g) + 18.07\log(h_b) + (32.10 + 13.9\log(h_b))\log(D)$ $D \geq 0.3Km$
Non Line Of Sight (NLOS)	$143.21 + 29.74\log(f_g) - 0.99\log(h_b) + (47.23 + 3.72\log(h_b))\log(D)$
D = BS- MU separation (km), h_b = BS antenna height, f_g = central frequency	

This urban area consists of mostly tall buildings each having more than ten stories. Hence, there exists a clear distinction between the *Line-Of-Sight* (LOS) and *Non-Line-Of-Sight* (NLOS) propagation. Table 5.6 summarizes the path loss formulas while Fig. 5.4 illustrates the considered model.

Mobile users are distributed uniformly over each cell space. To simplify the calculation of the other-cell interference factor, it is assumed that the neighboring cells contain the same number of users included in the reference cell.

The motion of each user is characterized as follows:

1. Speed is uniformly distributed between 0 and 4 km/h (pedestrian users).

2. When a user reaches the boundary, the user is reflected within the cell with a shift of 90 degree so that the number of users is always maintained.

3. A constant motion direction in the cell is originally generated with an angle uniformly distributed from *0* to 2π .

A summary of the main physical environment parameters is listed in Table 5.7.

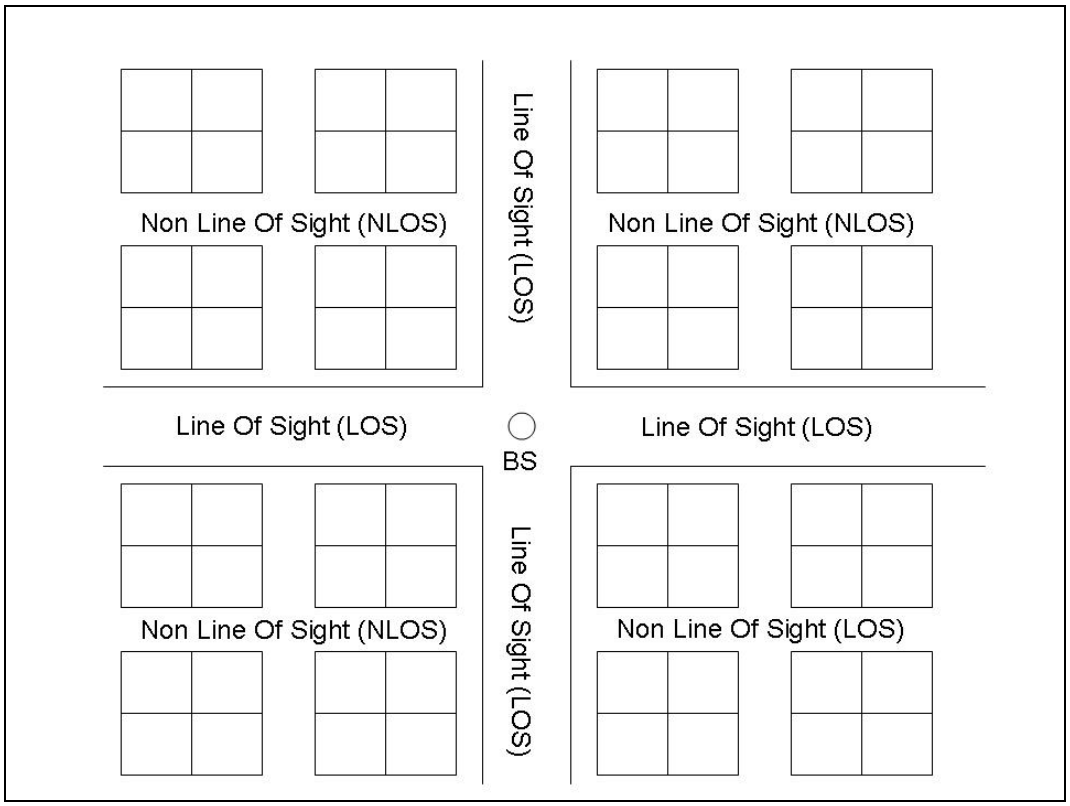

Fig. 5.4 Urban model

Table 5.7 Physical Environment parameters

Orthogonality factor, α_j	**0.9**
Base station maximum transmit power, $P_{tot,n}$	**40 W**
Number of neighboring Cells, K	**6**
Mobile station antenna height, h_m	**1.5 m**
Base Station antenna height, h_b	**7 m**
Shadowing attenuation (dB)	**Mean = 0, Standard deviation = 8**
spatial correlation distance, d_{corr}	**15 m**
Streets width	**15 m**
Cell Radius	**1 Km**
Mobile Speed	**0-4 km/h (pedestrian model)**

5.7 Experiment 1

In this experiment, all simulations are performed maintaining the percentages of Premium and Economy class of users or "grades of service" to 20 % and 80 %, respectively. Each simulation consists in loading the reference base station with a certain number of mobile users and collecting statistics. The simulations are performed taking into consideration all possible combinations of voice, video and/or data users that satisfy the QoS requirements. From one simulation to other, the number of users is increased and the simulations continue until the last combination of user is blocked on the condition that the QoS requirements in Table 5.4 are not satisfied.

Performance parameters of interest include maximum number of admitted users (BS_0 capacity), average percentage of lost (for audio and video users) and delayed (for data users) packets, average packet transfer delay, and average multimedia quality levels. Further parameters are throughput-based and power-based cell loading, and convergence time of the GA. These results will be shown in the following sub-sections.

5.7.1 System Capacity

The results for the system capacity, in terms of the maximum number of users that could be admitted in the base station BS_0 without violating the QoS requirements in Table 5.4, are shown in Fig. 5.5. Without filtering in scenario 1 (a), the total number of acceptable combinations of voice, video, and/or data users (20 % Premium, 80 % Economy) is 13, whereas in scenario 2 (b) it increases by a factor of 5.6 to 74. However, with filtering (c) this number has increased to 95 by a factor of 7.3.

Therefore, there is a gain in terms of increasing the capacity of the base station. To illustrate, the maximum number of admitted video, data, or voice users for an RNC

without filtering, is 5, 10, and 30, in scenario 1 (a), whereas 10, 30, and 60, in scenario 2 (b), respectively. However, the number of admitted users has increased to 15, 30, and 75 for an RNC with filtering (c). This gain is achieved while reducing the percentages of lost and delayed packets, and the average packet transfer delay, as illustrated in the following sub-section.

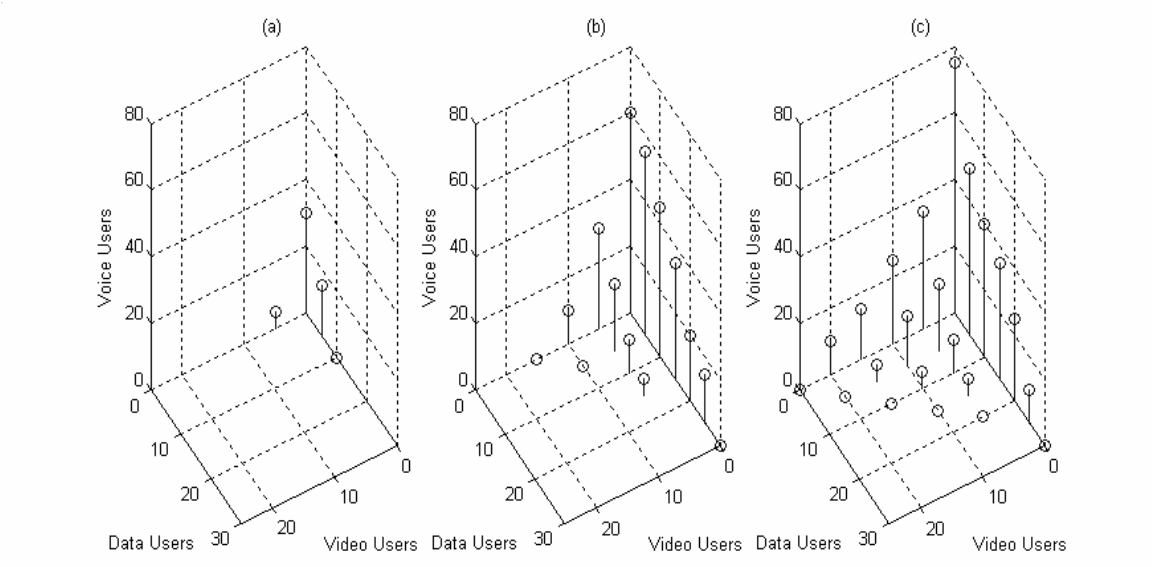

Fig. 5.5 Admission region of voice, video, and data users (20% Premium, 80% Economy), for an RNC without MAF in scenario 1 (a), for an RNC without MAF in scenario 2 (b), and for an RNC with MAF (c)

5.7.2 Percentages of Packets Lost and Delayed

Fig. 5.6 compares the percentages of lost (video and voice) and delayed (data) packets versus the number of voice (a), video (b), or data (c) users, accordingly. As expected, the highest percentages are obtained for an RNC without MAF in scenario 1. The RNC with the proposed MAF always outperforms that without filtering in scenario 1, regardless of the type of users and the grade of service. To illustrate, with 25 voice users, the percentages of packets lost for Premium services have decreased from 1.23 % to 0.02 %, whereas those for Economy ones from 0.74 % to 0.11 %.

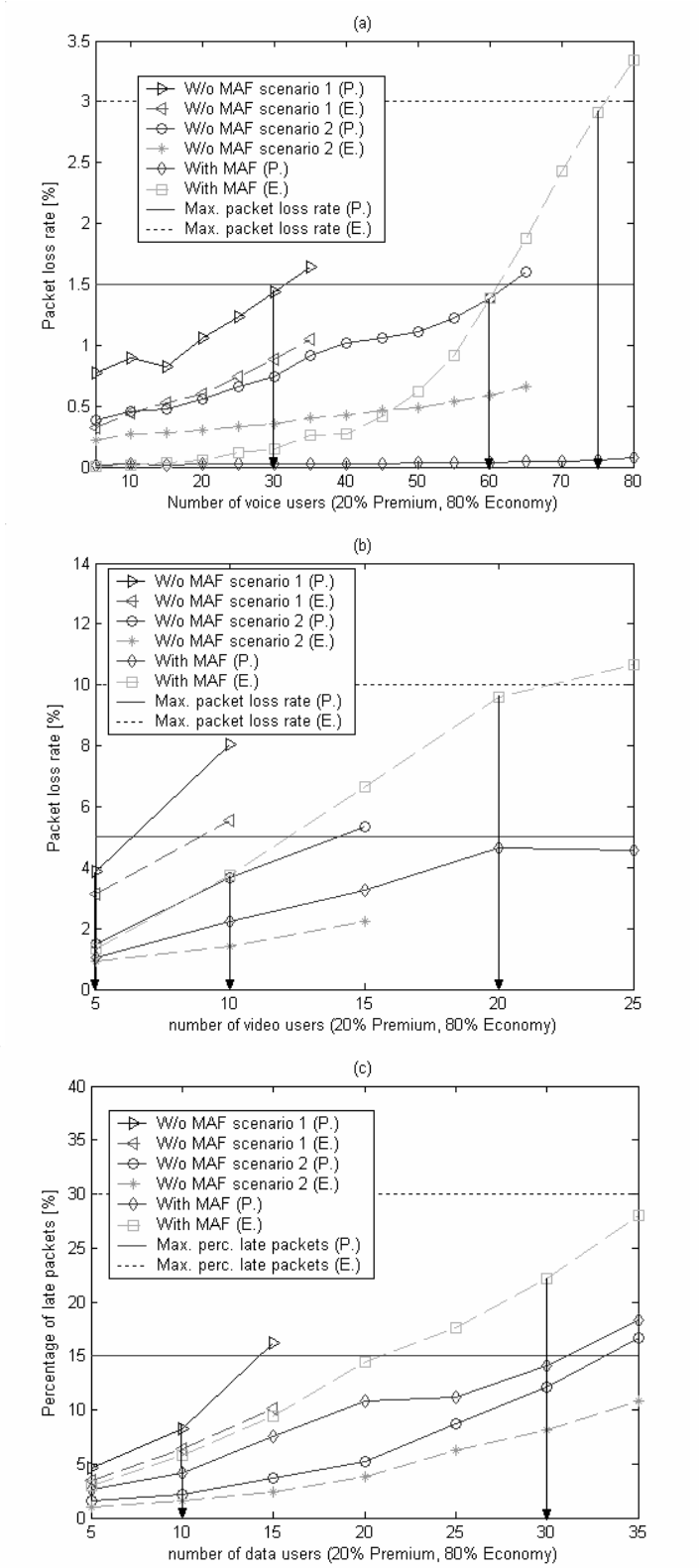

Fig. 5.6 Percentage of packets lost (voice, and video) or delayed (data) versus voice (a), video (b), or data (c) users (20% Premium, 80% Economy), for an RNC without MAF in scenario 1, for an RNC without MAF in scenario 2, and for an RNC with MAF

Similarly, with 5 video users, these percentages have decreased from 3.85 % to 1.02 % and from 3.11 % to 1.33 % for Premium and Economy services, respectively. Equally, the percentages of packets delayed have decreased from 8.25 % to 4.18 % and from 6.36 % to 5.74 %, respectively, with 10 data users. It also should be noted that for an RNC without filtering, in both scenarios 1 and 2, the percentages of packets lost and delayed for Economy services are less than those for Premium ones. Accordingly, for the RNC without filters, the maximum number of accepted combinations of users is determined by the constraints in the QoS parameters for the Premium class traffic flows, while, in a node with filtering, it is limited by those for the Economy ones. In fact, since the radio channels exhibit randomly time-varying impairments and the RRM of the RNC without filtering does not provide priorities between different grades of service, it is more probable that Premium class traffic flows, which require higher air interface bit rates to be downloaded, will perceive higher percentages of packets lost or delayed. On the contrary, the RRM with our proposed optimization function (52) permits to achieve the best performances for Premium services and a good performance for Economy services. However, as shown in the graph, the RNC with filtering is outperformed by that without filtering in scenario 2 for some combinations of admitted users. To illustrate, the percentages of packets delayed for Premium and Economy class data users are lower in an RNC without filtering in scenario 2 that those for a node with filtering. This is due to the fact that the traffic flows in scenario 2 are generated with lower quality levels (see Fig. 5.8).

5.7.3 Average Packets Transfer Delay

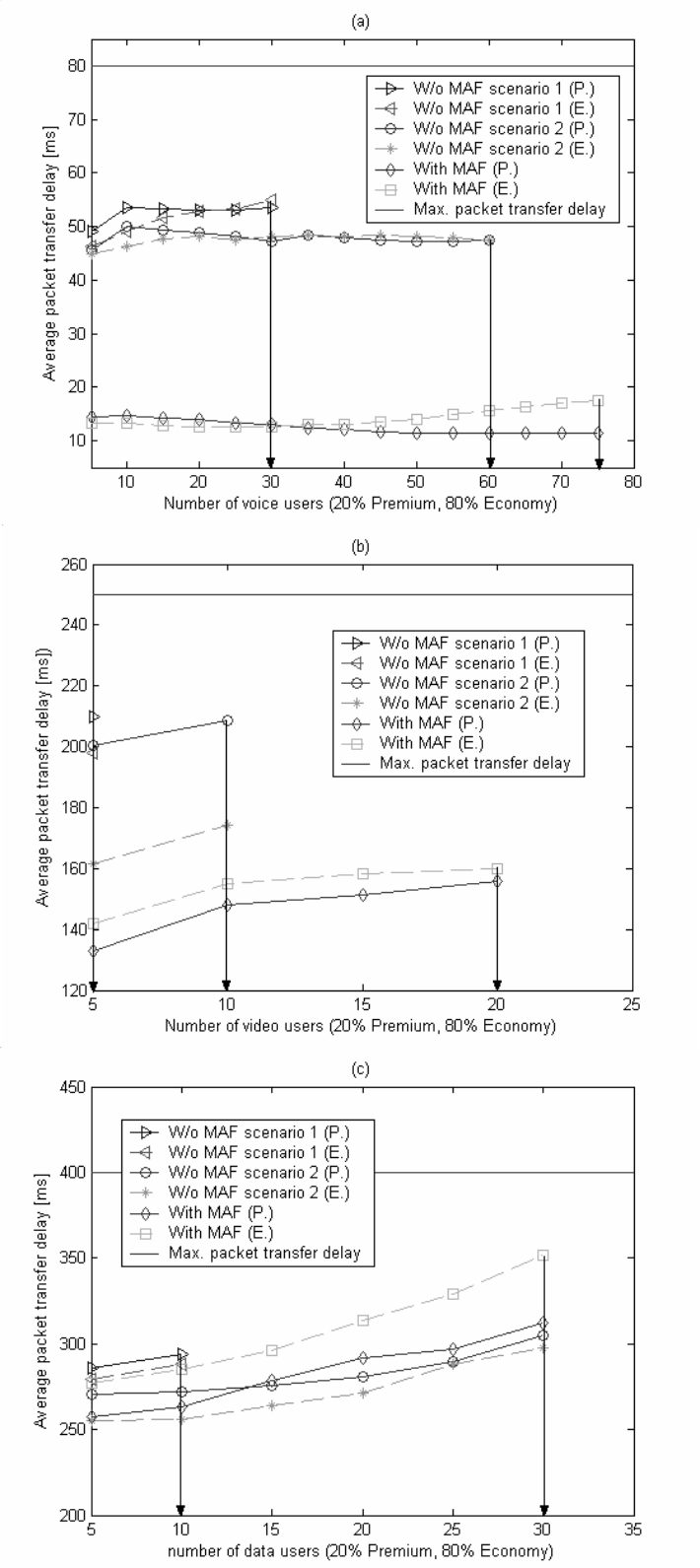

Fig. 5.7 Average packet transfer delay versus voice (a), video (b), or data (c) users (20% Premium, 80% Economy), for an RNC without MAF in scenario 1, for an RNC without MAF in scenario 2, and for an RNC with MAF

The values of average packet transfer delay for voice, video, or data users (20 % Premium, 80 % Economy) are provided in Fig. 5.7.

The highest packet transfer delays are obtained, as expected, for an RNC without filtering in scenario 1. In fact, in an RNC without MAF under bad radio channel conditions, the available air interface bit-rate to a user can be decreased, but coding rate and format of traffic flow cannot be changed, accordingly. The impact of that is that several *TBs* will be delayed in the transmission buffer and users will perceive "slower" downloads. On the contrary, the proposed RRM maintains the transfer packet delay low by tuning the amount of traffic to be downloaded with the current available air interface bit rates. To illustrate, with 25 voice users, the average packet transfer delay for Premium (Economy) class traffic flows is reduced from 53.17 ms (53.22 ms) to 13.38 ms (12.53 ms). Similarly, with 10 data users, the average packet transfer delay for Premium (Economy) class traffic flows is reduced from about 294 ms (287 ms) to 263.2 ms (284.8 ms). Similar values can be obtained for the average packet transfer delay directed to video users. However, as shown in Fig. 5.7.c, in scenario 2, the average data packets delays are lower than those for a node with filtering. This is due to the fact that packets directed to data users are not discarded but delayed if the QoS requirements are not satisfied.

5.7.4 Quality Levels

In order to evaluate the effect of the filtering actions on the traffic flows, the average quality levels for Premium and Economy services versus voice, video, and data users are represented in Fig. 5.8.

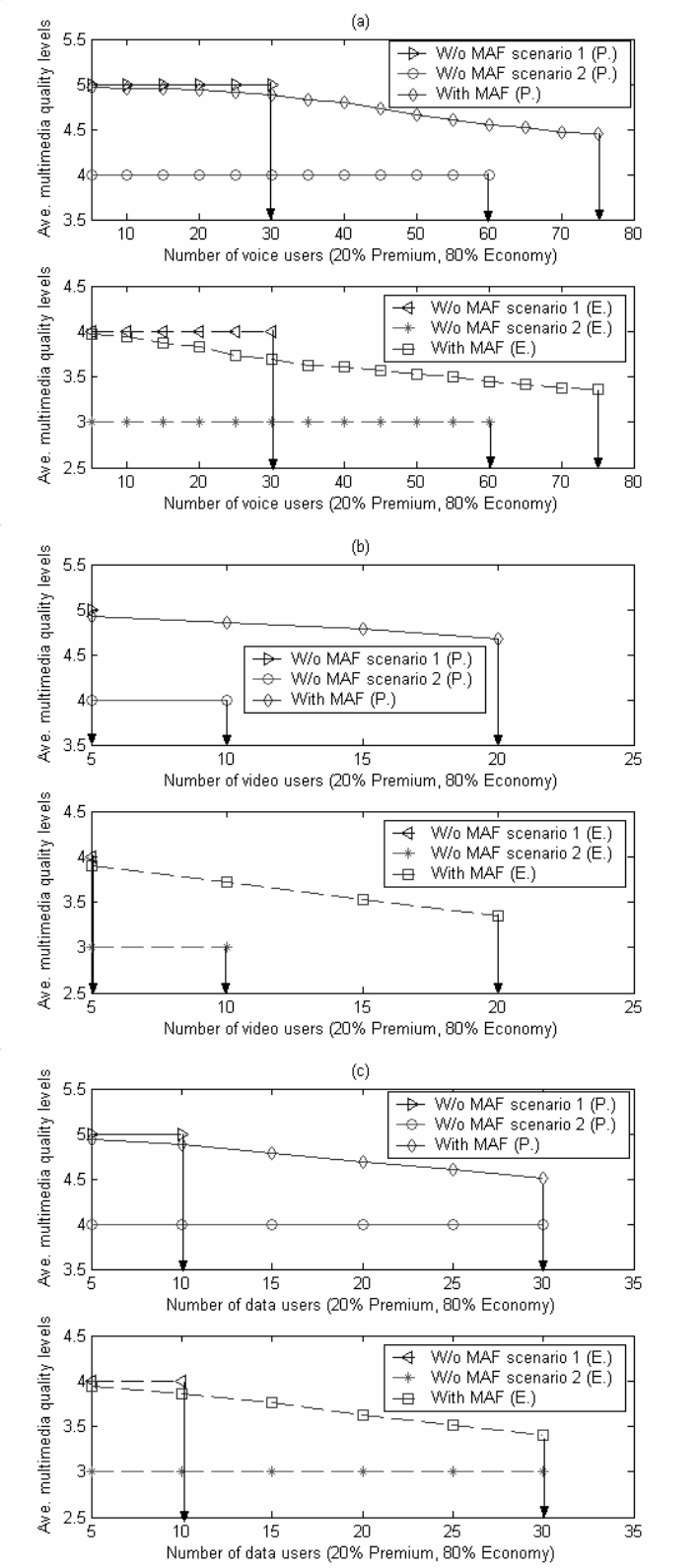

Fig. 5.8 Average multimedia quality levels of traffic flows versus voice (a), video (b), or data (c) users (20% Premium, 80% Economy), for an RNC without MAF in scenario 1, for an RNC without MAF in scenario 2, and for an RNC with MAF

For an RNC without MAF in scenario 1, traffic flows are downloaded with the highest quality levels but, as shown before, mobile users perceive slow downloads with higher percentages of packets lost.

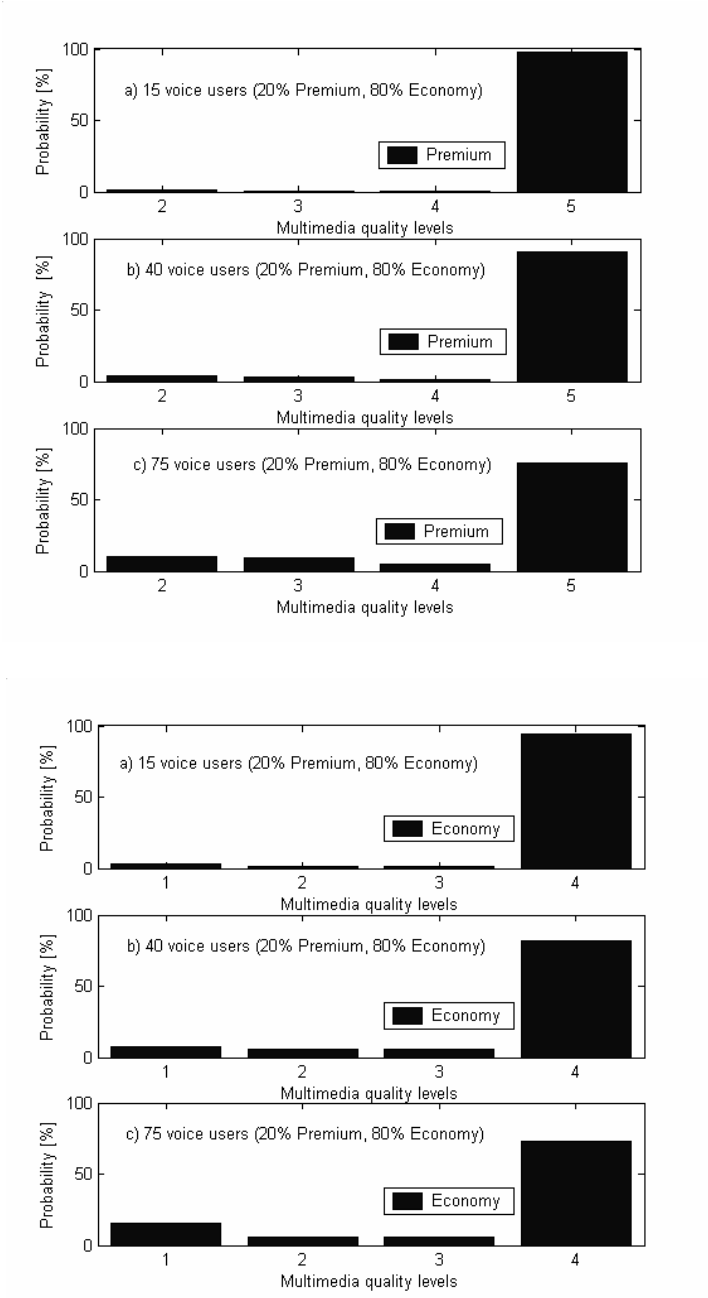

Fig. 5.9 Probability density function of the quality levels for Premium and Economy flows with 15 (a), 40 (b), and 75 (c) voice users (20% Premium, 80% Economy), for an RNC with MAF

On the contrary, the RNC without MAF in scenario 2, in some cases, outperforms in terms of average packet transfer delay or percentages of packet delayed the RNC with filtering, but Premium and Economy class traffic flows are generated with lower quality levels: 4 and 3. As shown in the graph, average quality levels in the proposed RNC are always higher than those in scenario 2. In fact, for Premium class traffic flows they are maintained between 4 and 5, whereas for Economy class traffic flows between 3 and 4. To show how the quality levels are reduced by the filter, the probability density for the quality levels of Premium and Economy class traffic flows is represented in Fig. 5.9. Three different combinations of admitted voice users (20 % Premium, 80 % Economy) are considered: 15 (a), 40 (b) and 75 (c). As expected, quality levels are reduced more frequently when the number of users increases. To illustrate, with 15 voice users, probabilities of selecting quality levels 5, 4, 3, 2 (4, 3, 2, 1) for Premium (Economy) traffic flows are about 97.9 %, 0.2 %, 0.5 %, and 1.4 % (94.3 %, 1.5 %, 1.5 %, and 2.7 %), respectively. If the number of users increases to 75, the correspondent probabilities become about 75.6 %, 4.4 %, 9.3 %, and 10.6 % (72.7 %, 5.8 %, 5.9 %, and 15.6%). From these results, it also should be noted that the average quality levels for Premium services are reduced more gracefully than those for Economy services as the number of voice users increases. This proves that functions (54) and (55) work well in term of providing preferential treatment to Premium class traffic flows over Economy ones. The same considerations can be done for the average quality levels of traffic flows directed to data and video users.

5.7.5 Throughput-Based and Power-Based Cell Loading

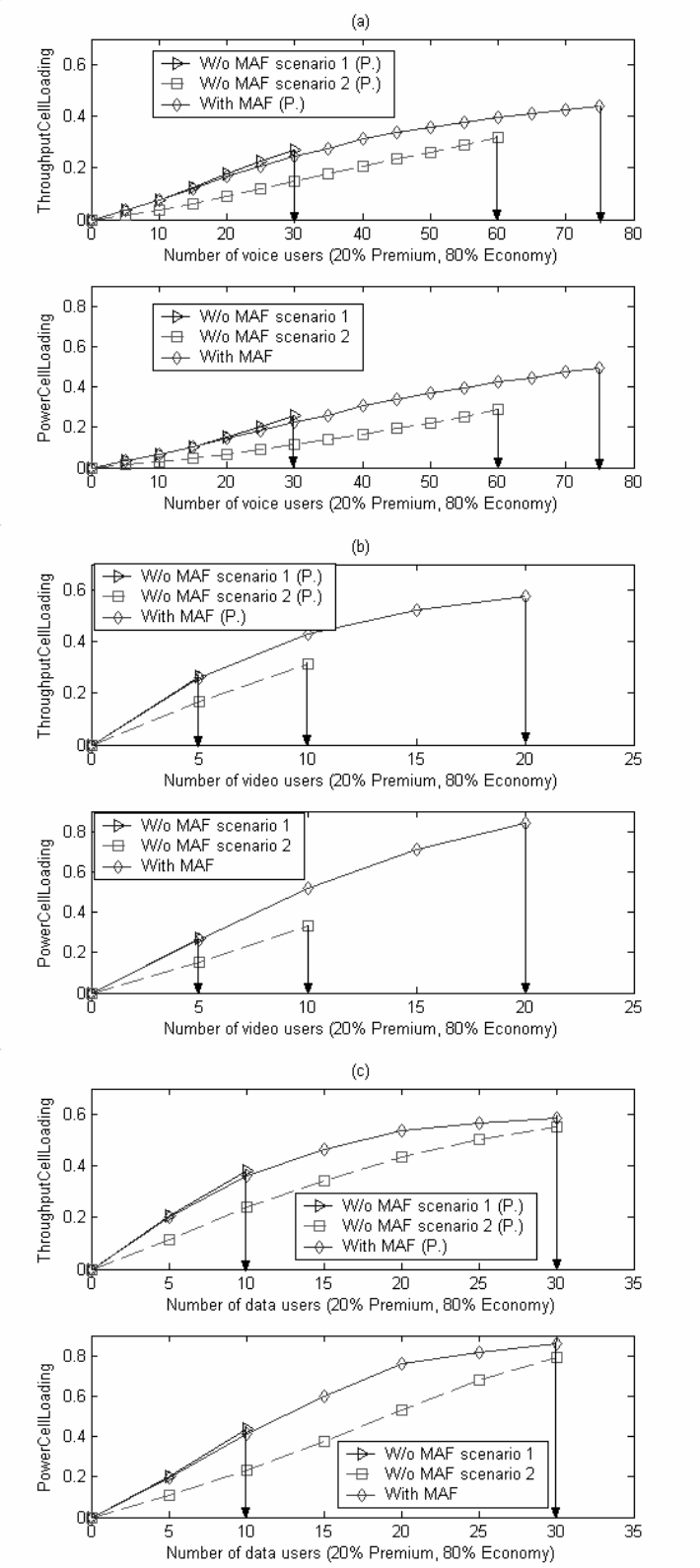

Fig. 5.10 Power-based and throughput-based cell loading versus number of voice (a), video (b), or data (c) users (20% Premium, 80% Economy) for an RNC without filters (scenarios 1 & 2) and an RNC with filters

The proposed algorithm is required to meet the QoS requirements for each traffic flow while utilizing efficiently the radio resources. In Fig. 5.10, throughput-based and power-based cell loading, which are calculated using Equations a) and b) in Section 5.2.1, are represented versus admitted users.

Recall that the RRM for an RNC without filtering (in both scenarios 1 and 2) does not provide priorities between traffic flows but allocates resources in order to maximize the throughput of the base station. Since throughput-based and power-based cell loading for the RNC without MAF in scenario 1 are comparable to those of an RNC with filtering, the filtering algorithm performs well in term of decreasing the percentages of packets lost (delayed), while offering better performances to Premium services and maximizing throughput-based and power-based based cell loading. On the contrary, throughput and power based cell loading for an RNC in scenario 2 are significantly lower. Therefore, although the RRM in scenario 2 outperforms in some cases the RNC with filtering (see Figs. 4 and 5), the total amount of traffic downloaded to mobile users in scenario 2 is significantly lower.

5.7.6 Total Revenues

The results for the total revenues gained by the service provider versus number of voice users are shown in Fig. 5.11. The revenues are calculated using the model in (67) defined in Section 5.5. The tariffs that users pay, in each control period, in order to download media with different levels of quality are calculated using the pricing mechanisms in (68), (69), (70), and (71), defined in Section 5.5. In this experiment, T_s^P and T_s^E were fixed to 0.0002 $ and 0.0001 $, respectively. Moreover, three different values were assigned to the coefficient α_T : 150 %, 70 %, and 30 %. As expected, the

revenues in a RNC with and without filtering increase if the percentage of tariff increment between two consecutive levels of quality, α_T , raises from 30 % to 150 % .

Moreover, as shown in Fig. 5.11, the revenues gained in a RNC with filtering are always higher than those in a RNC without filtering in both scenarios 1 (a) and 2 (b), regardless of the number of voice users admitted and the selected value for the coefficient α_T .

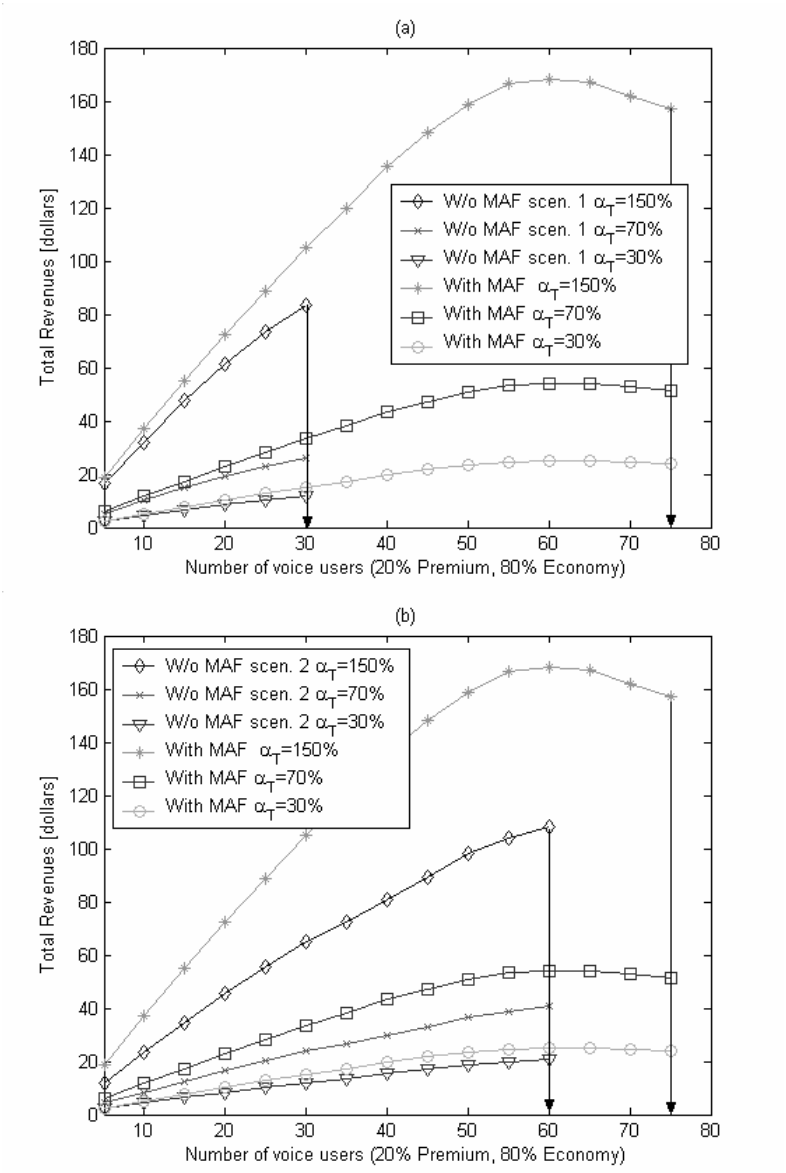

Fig. 5.11 Total revenues versus number of voice users (20% Premium, 80% Economy) (a) for an RNC with and without filters (scenario 1) (b) for an RNC with and without filters (scenario 2)

The increase in the revenues is basically obtained in three ways:

a) By accommodating an increased number of users. To illustrate consider Fig. 5.11 (a). In a RNC without filtering (scenario 1) the service provider can obtain revenues up to 83.4 $ (α_T = 150 %), 26.3 $ (α_T =70 %), or 11.7 $ (α_T =30 %). On the contrary, in a RNC with filtering the service provider can increase the revenues, by accommodating a higher number of users, up to 180 $ (α_T = 150 %), 54 $ (α_T =70 %), or 25.13 $ (α_T =30 %).

b) By downloading media with higher levels of quality to the same number of users. To illustrate, consider a combination of 60 voice users in Fig. 5.11 (b). In a RNC without filtering (scenario 2) the service provider can obtain 108 $ (α_T = 150 %), 40.63 $ (α_T =70 %), or 21.03 $ (α_T =30 %). With the same number of voice users, the service provider can increase the revenues to the same values as above because the levels of quality of the media downloaded are higher.

c) By downloading media with less percentage of packets lost. Consider a combination of 25 voice users in Fig. 5.11 (a). In a RNC without filtering (scenario 1) the service provider can obtain 73.22 $ (α_T = 150 %), 23.02 $ (α_T =70 %), or 10.03 $ (α_T =30 %). With the same number of voice users, the service provider can increase the revenues, by decreasing the percentages of packet lost, to 3.22 $ (α_T = 150 %), 23.02 $ (α_T =70 %), or 10.03 $ (α_T =30 %).

Another important result to report is that in an RNC with filtering the increase in the revenues ends at 60 voice users although the graph shows that additional 15 users could be admitted. When more than 60 users download packets, the filtering algorithm

maintains the QoS requirements but the revenues do not increase. Hence, there is no advantage for a service provider to admit any more users above this limit despite the fact the filtering algorithm would maintain the QoS parameters even for a larger number of calls.

5.7.7 Revenues per User

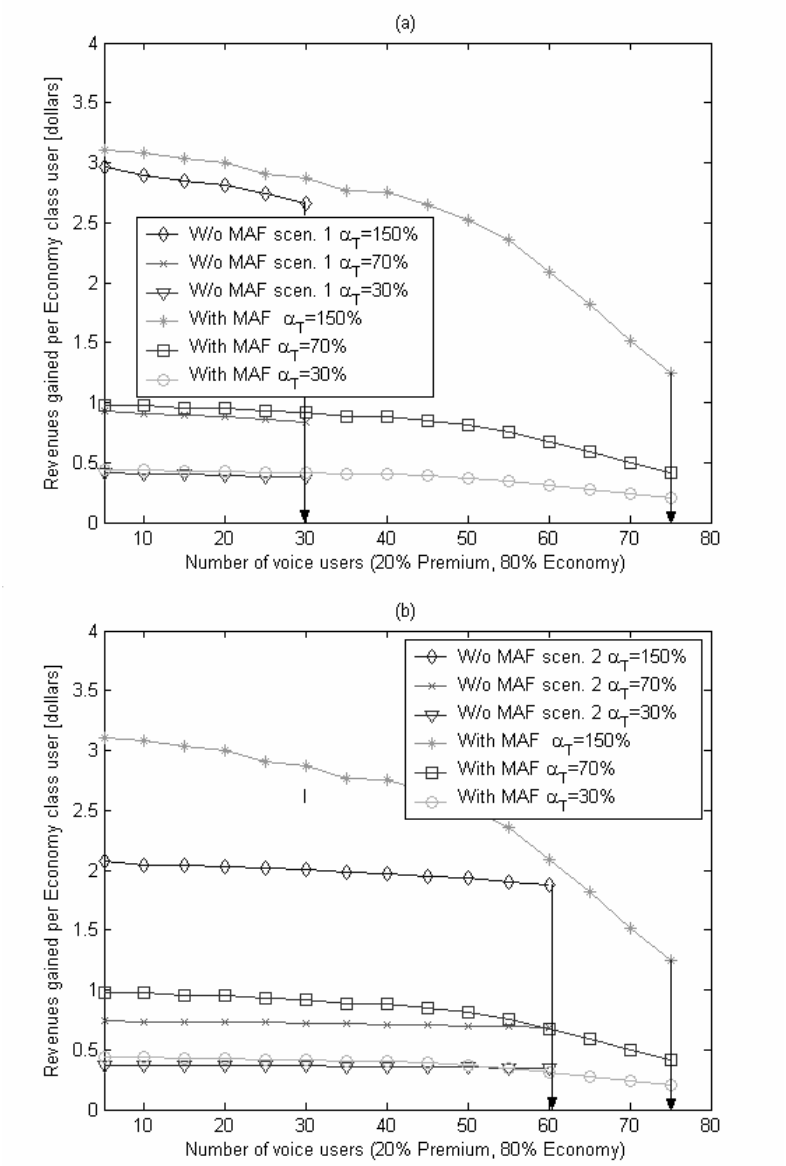

Fig. 5.12 Revenues gained per Economy class user versus number of voice users (20% Premium, 80% Economy) (a) for an RNC with and without filters (scenario 1) (b) for an RNC with and without filters (scenario 2)

In this sub-section, results for the revenues gained by the service provider per Premium and Economy class users are shown in Fig. 5.12 and Fig. 5.13, versus number of admitted voice users. As shown in Fig. 5.12 in a node (RNC) with filtering Economy class users pay less if the number of admitted users increases, regardless of the selected value for the coefficient α_T .

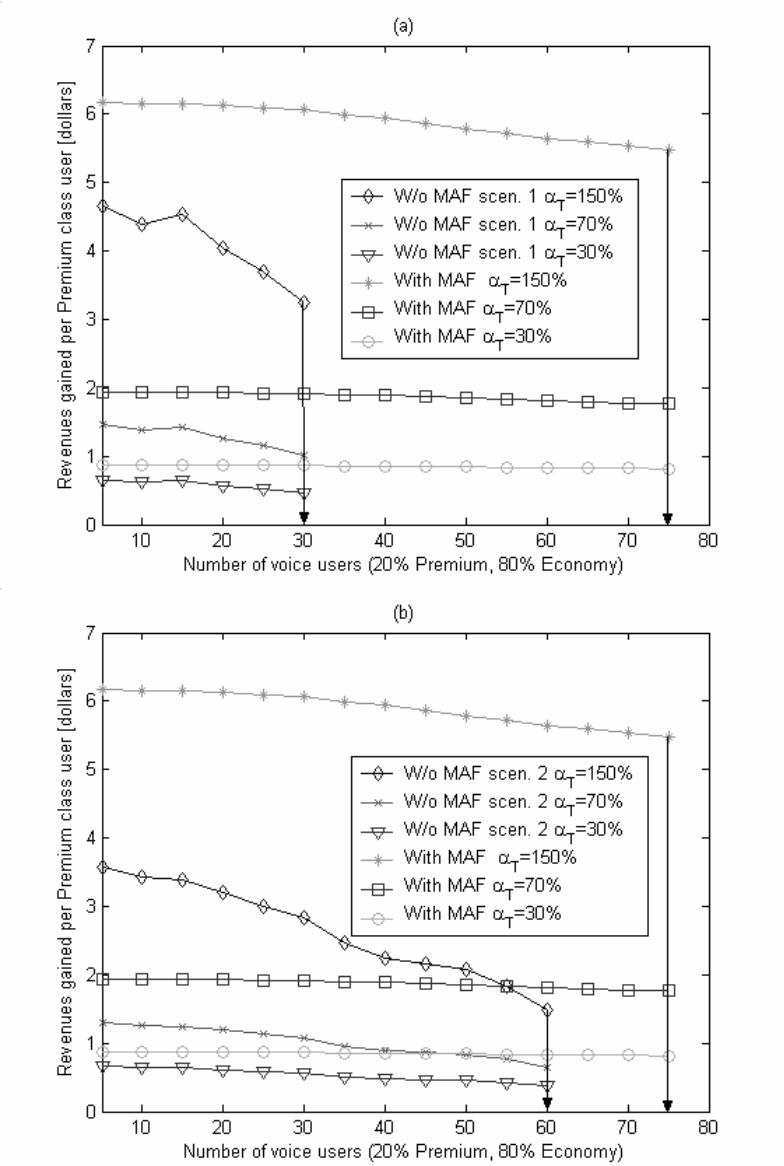

Fig. 5.13 Revenues gained per Premium class user versus number of voice users (20% Premium, 80% Economy) (a) for an RNC with and without filters (scenario 1) (b) for an RNC with and without filters (scenario 2)

In fact, both the percentage of packets lost and the average quality levels for Economy class users tend to decrease if the number of admitted users increases. Therefore, under high traffic load, Economy class users will receive content with a lower quality but they will pay less. On the contrary, if the number of admitted voice users increases, Premium class users will pay almost the same (see Fig. 5.13). In fact, since the proposed filtering algorithm manages to maintain the QoS parameters (percentage of packets lost and the average quality level) for traffic flows downloaded to Premium class users, regardless of the number of admitted users, Premium class users will continue to pay the same tariff. On the contrary, in a node (RNC) without filtering (scenarios 1&2), if the number of admitted users increases, Premium class users will pay a lower tariff, while Economy class users will pay the same tariff. In fact, since the radio channels exhibit randomly time-varying impairments and the RNC without filtering does not provide priorities between traffic flows downloaded to Premium and Economy class users, it is more probable that Premium class traffic flows, which require higher bandwidth and more stringent QoS parameters, will perceive higher percentages of packets lost and superior reductions of the quality levels.

Based upon these results, it may be concluded that in a node with filtering, the network resources are more profitable for the service provider because they are reserved for downloading traffic flows directed to users that are willing to pay more for their service. This is another factor that has to be considered, together with those discussed in the previous section, in order to increase the revenues of a service provider.

5.7.8 Genetic Algorithm Convergence

An important parameter to consider for evaluating the performance of the GA is the convergence time. This parameter is represented in Fig. 5.14 versus the number of voice, video, or data users.

As mentioned in Section 5.4, the stopping criterion used by our GA engine is based on a convergence condition. A requirement for the effectiveness of the solutions of the optimization problem is that the control period, which is the maximum period allotted to solve the objective function, needs to be longer than the convergence time. From Fig. 5.14, it is clear that the algorithm converges within our defined control period (120 ms), regardless of the number and type of admitted users. Note that the convergence time is calculated by running the C++ emulator on a Pentium IV processor with a processing power of 3.5 GHz.

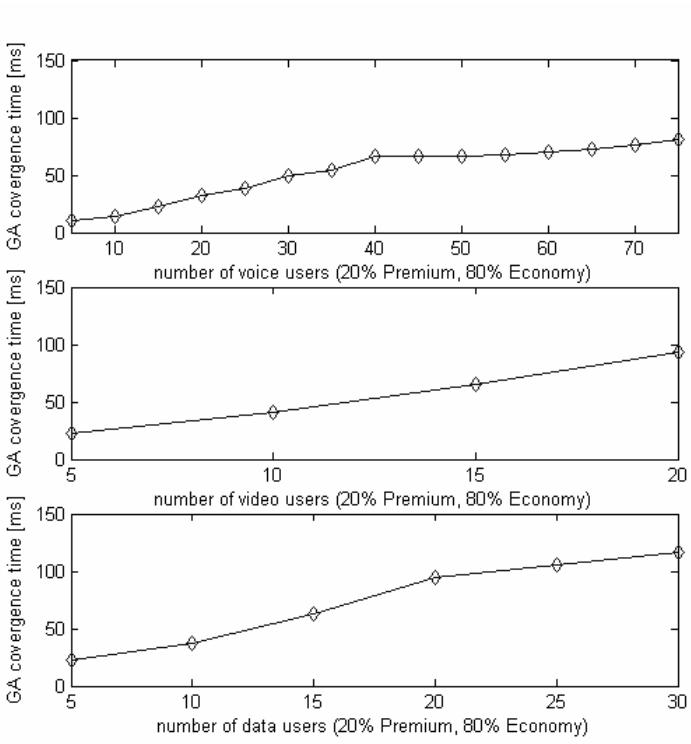

Fig. 5.14 Convergence time of the Genetic Algorithm versus different numbers of voice, video, or data users (20% Premium, 80% Economy)

As expected, the computational complexity of solving an optimization problem and, consequently, its convergence time increase as the number of variables (users) and bits to encode them increase

5.8 Experiment 2

In this experiment, it is considered a reference base station in which 20 voice users (20 % Premium and 80 % Economy) initiate a call. The simulation is repeated while varying the control period from 40 ms to one full second. At each run, a constant control period during the simulation is used. As usual, to obtain reasonable average statistics each run is iterated several times.

5.8.1 Impact of the Control Period

The average data recorded are illustrated in Fig. 5.15 and Fig. 5.16. In particular, Fig. 5.15 shows cell throughput (a) and packet loss rate (b), while Fig. 5.16 depicts average packet transfer delay (a), and average multimedia levels (b) versus the control period. As the control period decreases, the percentage of packets lost and the average transfer delay decrease, while the cell throughput and the average quality levels increase. In fact, at small Δt, the algorithm is activated more frequently and this makes it able to solve any problem as soon as it appears. Hence, air interface bit rates and quality levels are updated more frequently according to the changes in the wireless conditions and traffic burst. It should be noted that for a small control period (40 ms) the percentage of packets lost and the average transfer delay increase instead of decreasing. In fact, since the estimation of the input traffic is performed on the traffic arrived in the previous control period, if the period is too small the solutions of the algorithm are not precise.

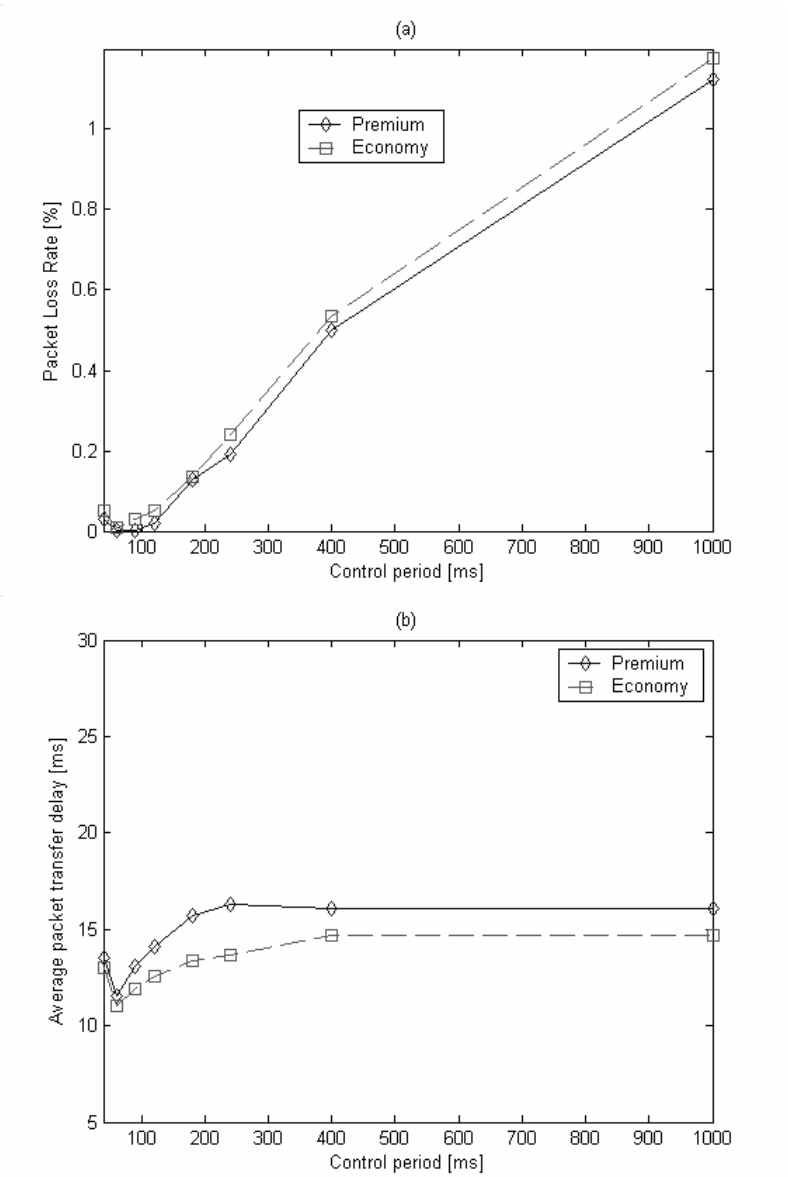

Fig. 5.15 Packet loss rate (a), average transfer delay (b), versus control period for an RNC with MAF connected to a reference cell including 20 voice users (20 % Premium, 80 % Economy).

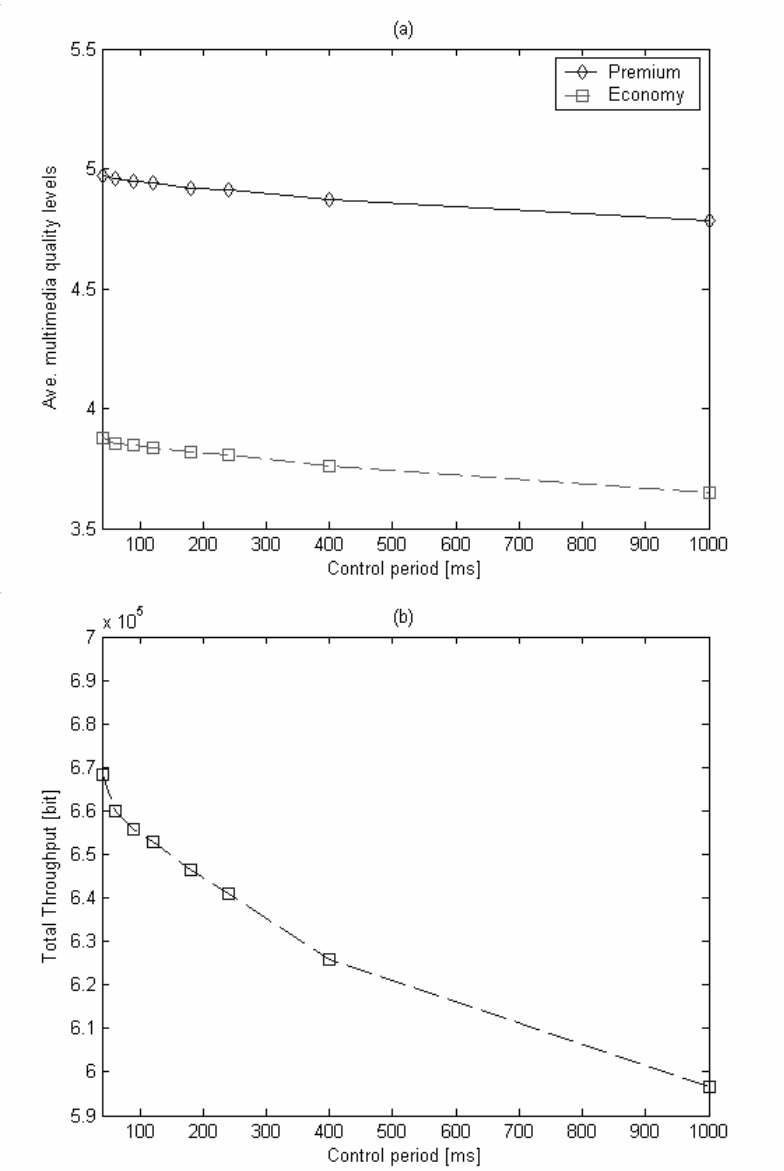

Fig. 5.16 Average multimedia quality levels (a), and total throughput (b) versus control period for an RNC with MAF connected to a reference cell including 20 voice users (20 % Premium, 80 % Economy).

Chapter 6

Conclusions and Future Directions

In this chapter, the conclusions of the thesis are presented. This chapter begins with an overview of the arguments of the thesis. Following this, the main contributions of the thesis are presented. Finally, the thesis concludes by providing some indicators for future work.

6.1 Summary of Thesis

The thesis began in Chapter 1 with an outline of key problems involved in the deployment of heterogeneous (core and radio access networks) IP-based networks. The chapter argued that today's network technologies (e.g., Diff-Serv, MPLS, etc.), which use dynamic or static approaches to reserve network resources in response to predefined traffic demands or QoS requirements from the applications, do not provide any form of intelligent control of the applications. The proposed solution was formulated to solve the issues. An "intelligent" interaction between the applications and the network resources could allow service providers to utilize effectively the network resources, and increase their revenues without violating QoS parameters.

Chapter 2, first, presented the main components of the architectural model of the 3G network; then, it provided an overview of fundamental QoS principles, concepts, and terminology. Next, current techniques for managing QoS in IP-based 3G networks were

described. Finally, two adaptation strategies based on intelligent interactions between applications and network were presented as new ultimate solutions for QoS support.

Chapter 3 presented an outline of the author's contribution to the end-to-end QoS research, an adaptive filtering architecture for QoS-adaptive heterogeneous networks. According to the proposed architecture, multimedia adaptation filters were implemented in the nodes of the heterogeneous networks in order to adapt the amount and type of traffic downloaded from content servers to the mobile users along the end-to-end path. The major features and parameters of the filter were introduced and discussed. The thesis also introduced two different grades of service to which mobile users could subscribe: Premium and Economy.

Chapter 4 presented the author's first major contribution to define a multi-objective function able to control the actions of the proposed filter and class-based queuing system within the nodes of the core network. Experiments were conducted to test the proposed approach. Results proved that significant improvements could be obtained by implementing both the filter and optimizations function in the nodes.

Finally, Chapter 5 showed the author's second major contribution to characterize a multi-objective function able to manage the operations of the filter and air interface rate schedulers within the nodes of the radio access network. Both multi-objective functions, defined in Chapter 4 and Chapter 5, were optimized using a randomized search techniques called Genetic Algorithms. Experiments were also conducted in order to verify the effectiveness of the approach to increase the number of admitted users in the radio access network while maximizing the throughput-based and power-based cell loading.

6.2 Conclusions

This thesis presented an adaptive filtering architecture that enables the delivery of IP multimedia services in 3G networks by adaptively controlling, along the end-to-end path, both the allocated bandwidth and the transfer delay of traffic flows downloaded from content servers to mobile users. The filtering architecture included the installation of filters at the output ports of each node of 3G heterogeneous (core and radio access networks) network.

The primary objectives of the proposed architecture was to guarantee the service level agreements of two different classes of users (Premium and Economy) while enhancing the utilization of the networks' assets, avoiding congestion, and maintaining the QoS requirements for each type of media.

To achieve these goals, users were allowed to download multimedia traffic flows within a range of pre-defined coding rates and formats instead of just a single fixed one. The filtering architecture provided the network operator with the flexibility to control not just the coding rate of the content but also its type as well. The filters selected the optimal coding rate and format for each traffic flow according to the dynamics of the links' traffic-loads and users' requests.

The most contribution of this thesis was the definition of two multi-objective optimization functions, which were used to control the actions of the filters. These functions have been solved using Genetic Algorithms in order to meet the objectives outlined above while satisfying the constraints of the heterogeneous networks (e.g., links' capacity, buffers state, total power, and number of spreading factors).

Performance evaluation results showed that the proposed filters provided significant gains in term of reducing the call blocking and dropping probabilities (two important measures of the quality of the service that the user receives), providing multimedia services with QoS guarantees without wasting links capacities (i.e., decreasing cost of service to operators), increasing the number of accepted users for the same utilization factor (i.e., increasing revenues to operators), and finally preventing congestion episodes by avoiding the downloading of large amounts of data when capacity was unavailable. Interesting results reported were that i) the increase in the revenues saturated at a certain number of users beyond which admitting any more users was not cost-effective for the network operator, and ii) the revenue gain was feasible as long as the network provider offered customized classes of service and differentiated pricing mechanisms.

6.3 Future Work

Possible future research work is included in the following sub-sections.

6.3.1 End-to-End Performance Analysis

In this thesis, we did not study the performance on an end-to-end basis. It is important to extend this work to analyze the impact of different types of filters on an end-to-end basis. Possible issues to investigate include quantifying the advantages, if any, of applying different filtering policies depending upon the location of the filters in the network topology. Additionally, it will be interesting to quantify the impact of filtering on higher level transport protocols, e.g. TCP. Finally, a key issue is to study, on an end-to-end basis, the interaction between adaptive filtering and multicast routing.

6.3.2 Adaptive Reconfiguration of the Network Resources

Another important issue is to develop and study efficient algorithms that aim at enhancing the utilization of the network assets by re-configuring these assets to areas of the network where temporarily demands of short-term capacities raise. For example, in situations where users in an airport or on a train require large capacity but only for few hours or during rush hours traffic, it may be effective, in these scenarios, to re-allocate capacities from an under-utilized area of the network to these over-loaded areas through rapid provisioning and reconfiguration techniques.

6.3.3 Deployment Issues of Receivers with Heterogeneous Capabilities

Deployment issues that are worth investigating include utilizing techniques like software radio for automatic reconfiguration of the receiver's air interface technologies and applications. In the next future, a receiver will be able to select different radio interface functionalities, such as modulation techniques, error correction schemes, and antenna patterns, according to the different wireless conditions and environments. These functionalities instead of being implemented in hardware will be performed via software modules, called Digital Signal Processors (DSP). It will be interesting to extend this work to analyze the deployment and impact of reconfigurable receivers within an adaptive filtering architecture.

Appendix A

Appendix A is intended to show how the maximum revenues Φ (Equation (29)) and the number of admitted Premium N_P and Economy N_E class users (Equations (30), (31)) are calculated, for the theoretical case illustrated in Section 4.8, versus the different percentages of Premium, x_P, and Economy, x_E, class calls and the available link capacity $C_{av}(Mbps)$.

Define the portions of the available link capacity assigned for Premium and Economy class calls to be $B_P(Mbps)$ and $B_E(Mbps)$. These parameters are given by:

$$B_P(Mbps) = N_P \cdot \sum_{\alpha=s,v}^{d} CR_{\alpha,P}^{w.f.}(Mbps)$$

(i)

$$B_E(Mbps) = N_E \cdot \sum_{\alpha=s,v}^{d} CR_{\alpha,E}^{w.f.}(Mbps)$$

(ii)

Moreover, define the ratio between the total coding rate assigned to download Premium and Economy class media streams to be *b*, as follows

$$b = \frac{\sum_{\alpha=s,v}^{d} CR_{\alpha,P}^{w.f.}(Mbps)}{\sum_{\alpha=s,v}^{d} CR_{\alpha,E}^{w.f.}(Mbps)}$$

(iii)

Using the previous equation and dividing Equation (i) by Equation (ii), we obtain:

$$\frac{B_P(Mbps)}{B_E(Mbps)} = \frac{N_P}{N_E} \cdot b$$

(iv)

Since $\frac{N_P}{N_E}$ is equal to $\frac{x_P}{x_E}$, we can replace the previous equation with

$$\frac{B_P(Mbps)}{B_E(Mbps)} = \frac{x_P}{x_E} \cdot b$$

(v)

By solving the following system of linear equations with respect to $B_P(Mbps)$ and $B_E(Mbps)$

$$\begin{cases} B_P(Mbps) + B_E(Mbps) = C_{av}(Mbps) \\ \dfrac{B_P(Mbps)}{B_E(Mbps)} = \dfrac{x_P}{x_E} \cdot b \end{cases}$$

(vi)

we obtain

$$B_E(Mbps) = \frac{C_{av}(Mbps)}{\frac{x_P}{x_E} \cdot b + 1}$$

(vii)

$$B_P(Mbps) = \frac{C_{av}(Mbps)}{\frac{x_E}{x_P} \cdot \frac{1}{b} + 1}$$

(viii)

From Equations (i), (ii), (iii), (vii), and (viii) we can determine Equations (29), (30) and (31).

Appendix B

This Appendix B is intended to calculate from Equation (40), the minimum transmitted power to user k such that the targeted $\left(\frac{E_b}{I_0}\right)_{req,k} = \gamma_k$ for all N users included in the reference BS_0 is respected. From Equations (40) and (48) it yields:

$$\frac{P_k h_{k,0}}{(1-\alpha_k)(P_{tot,0}-P_k)h_{k,0} + N_0 W + \lambda_k P_{tot,0} h_{k,0}} G_k = \gamma_k$$

(ix)

Using the previous equation P_k is determined, as follows:

$$P_k = \frac{\gamma_k}{G_k + \gamma_k(1-\alpha_k)}\left[(1-\alpha_k+\lambda_k)P_{tot,0} + \frac{N_0 W}{h_{k,0}}\right]$$

(x)

and $g_k = \frac{\gamma_k}{G_k + \gamma_k(1-\alpha_k)}$ is defined.

Using Equations (48) and (x), we also write

$$P_{tot,0} = \sum_{k=1}^{N} P_k = \sum_{k=1}^{N} g_k \left[(1-\alpha_k+\lambda_k)P_{tot,0} + \frac{N_0 W}{h_{k,0}}\right]$$

(xi)

From the previous relationship, it is determined $P_{tot,0}$ as follows:

$$P_{tot,0} = \frac{\sum_{k=1}^{N} \frac{g_k}{h_{k,0}} N_0 W}{1 - \sum_{k=1}^{N} g_k (1 - \alpha_k + \lambda_k)}$$

(xii)

By replacing $P_{tot,0}$ in Equation (x), it yields:

$$P_k = g_k \left\{ (1 - \alpha_k + \lambda_k) \left[\frac{\sum_{k=1}^{N} \frac{g_k}{h_{k,0}} N_0 W}{1 - \sum_{k=1}^{N} g_k (1 - \alpha_k + \lambda_k)} \right] + \frac{N_0 W}{h_{k,0}} \right\}$$

(xiii)

Equation (49) can be easily determined from (xiii).

References

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

[2] 3GPP, "End-to-End QoS Concept and Architecture", *Technical Specification TS 23.207*

[3] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proceedings IEEE*, vol. 83, no. 10, pp. 1374-1396, Oct. 1995

[4] M. Naghshineh, M Schwartz, and A. Acampora, "Issues in wireless access broadband networks," *Proceedings of Winlab '95*, 1995

[5] M. Naghshineh, and M. Willebeek-LeMair, "End-to-End QoS Provisioning in Multimedia Wireless/Mobile Networks Using an Adaptive Framework," *IEEE Communications Magazine*, vol. 35, no. 11, pp. 72-81, Nov. 1997

[6] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the internet Architecture: An Overview", *IETF RFC 1633*, July 1994

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

[8] E. Rosen, et alt., "Multi Protocol Label Switching Architecture", *IETF RFC 3031*, Jan 2001

[9] S. Rappaport, "Wireless Communications", Prentice Hall, New Jersey, 1996

[10] T. Oianpera and R. Prasad, "Wideband CDMA for Third Generation Communications", Artech House, 1998

[11] I. C. Lin and R. D. Gitlin, "Multi-code CDMA Wireless Personal Communications Networks," *IEEE International Conference in Communications'95* , Seattle, WA, pp. 1060-64, June 1995

[12] A. Campora, "Wireless ATM: A perspective on issues and prospects," *IEEE Transactions on Personal Communications*, vol. 3, no. 4, pp. 8-17, Aug. 1996

[13] L. Delgrossi, C. Halstrick, D. Hehmann, R. G Herrtwich, O. Krone, J. Sandvoss, and C. Vogt, "Media scaling for audiovisual communication with the Heildelberg transport system," *Proceeding of ACM Multimedia '93*, Aug. 1993

[14] A. Campbell, G. Coulson, and D. Hutchinson, "Transporting QoS adaptive flows" *ACM/Springer Verlag Multimedia Systems Journal*, vol. 6, no.3, pp. 167-178, May 1998

[15] N. Yeadon, F. Garcia, D. Hutchinson, and D. Shepherd, "Filters: QoS Support Mechanisms for Multi-peer Communications," *IEEE Journal of Selected Areas in Communications*, vol. 14, no. 7, pp. 1245-1262, Sept. 1996

[16] A. Balachandran, A. Campbell, and M. Kounavis, "Active filters: Delivering scaled media to mobile devices," *Proceedings of 7th International Workshop on NOSSDAV*, May 1997

[17] M. Mirhakkak, N. Schult, and D. Thomson, "Dynamic Bandwidth Management and Adaptive Applications for a Variable Bandwidth Wireless Environment", *IEEE Journal of Selected Areas in Communications*, vol. 19, no. 10, pp. 1984-1997, Oct. 2001

[18] A. Kassler, and A. Schorr, "Generic QoS Aware Media Stream Transcoding and Adaptation" *Proceeding of the Packet Video Workshop 2003*, April 2003

[19] A. Vetro, H. Sun, and Y. Wang, "Object-based Transcoding for Adaptable Video Content Delivery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 387-401, Mar. 2001.

[20] T. Shanableh, and M. Ghanbari, "Heterogeneous Video Transcoding to Lower Spatio-Temporal Resolutions and Different Encoding Formats," *IEEE Transactions on Multimedia*, vol. 2, no. 2, pp. 101-110, June 2000

[21] A. Kassler, A. Neubeck, and P. Schulthess, "Real-Time Filtering of Wavelet Coded Videostreams for Meeting QoS Constraints and User Priorities," *Proceedings of the Packet Video Workshop*, May 2000

[22] A. Campbell, N. Yeadon, D. Hutchinson, and C. Aurrecoechea, "A Dynamic QoS Management Scheme for Adaptive Digital Video Flows," *Proc. 4th Int. Workshop on NOSSDAV*, October 1993

[23] C. Aurrecoechea, A. T. Campbell, and L. Hauw, "A Survey of QoS Architectures," *ACM Multimedia System Journal*, Special Issue on QoS Architecture, May 1998

[24] D. Chalmers and M. Sloman, "A survey of Quality of Service in Mobile Computing Environments", *IEEE Communications Surveys*, pp. 2-10,2nd Qtr., 1999

[25] P. Frene, D. Rasseneur, and P. Tournassoud, "Mobile evolution towards full IP multimedia", *Alcatel Telecommunication Review 2000*

[26] Holma, and Toskala. "WCDMA for UMTS: Radio Access for Third Generation Mobile Communication", 2nd edition, John Wiley & Sons, 2001

[27] H. Kaaranen, A. Ahtiainen, L. Laitinen, S. Naghian, and V. Niemi, "UMTS Networks", John Wiley & Sons, 2001.

[28] K. Nhrstedt and R. Steimetz, "Resource Management in Networked Multimedia Systems", *IEEE Computer*, pp.52-63, May 1995

[29] 3GPP, "IP Multimedia (IM) Subsystem - Stage 2", *Technical Specifications TS 23.228*

[30] X. Gao, G. Wu, and T. Miki, "End-to-End QoS Provisioning in Mobile Heterogeneous Networks", *IEEE Wireless Communications Magazine*, pp. 24-34, June 2004

[31] H. Liu and M. El Zarki, "Adaptive source rate control for real-time wireless video transmission", *ACM Mobile Networks and Applications*, vol. 3, no. 1, pp. 49-60, Jan 1998

[32] C. Y. Hsu, A. Ortega, M. Khansari, "Rate Control for Robust Video Transmission over Burst-Error Wireless Channels", *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 5, pp. 756-773, May 1999

[33] L. Galluccio, F. Licandro, G. Morabito, and G. Schembra, "An Analytical Framework for the Design of Intelligent Algorithms for Adaptive-Rate MPEG Video Encoding in Next-Generation Time-Varying Wireless Networks", *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 369-384, Feb. 2005

[34] R. Rejaie, M. Handley, D. Estrin, "Quality Adaptation for Congestion Controlled Video Playback over the Internet", *ACM Computer Communication Review*, vol. 29, no. 4, pp. 189-200, October 1999

[35] M. Margaritidis, G. C. Polyzos, "MobiWeb: Enabling Adaptive Continuous MediaApplications over Wireless Links", *IEEE International Conference in 3G Wireless 2000*; San Francisco, June 2000.

[36] V. Bharghavan and V. Gupta, "A Framework for Application Adaptation in Mobile Computing Environments" *Proceedings of the IEEE Compsoc'97*, Nov. 1997

[37] O. Angin, A. Campbell, M. E. Kounavis, and R. Liao "The mobiware Toolkit: Programmable Support for Adaptive Mobile Networking" *IEEE Personal Communications*, vol. 5, no. 4, pp. 32-43, Aug. 1998

[38] J. C. Pasquale, G. Polyzos, E. W. Anderson, and V. P. Kompella, "The Multimedia Multicast Channel", *Proceedings of the 3rd NOSSDAV*, La Jolla, CA, Nov.1992

[39] T. Warabino, S. Ota, D. Morikawa, M. Ohashi, H. Nakamura, H. Iwashita, and F. Watanabe, "Video Transcoding Proxy for 3Gwireless Mobile Internet Access," *IEEE Communications Magazine,* vol. 38, no. 10, pp. 66-71, Oct. 2000

[40] J. Zhou, "Heterogeneous Multicasting based on RSVP and QoS filters," *ICCT'98,* Oct. 98, Beijing

[41] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: a Transport Protocol for Real-Time Applications." *Internet Engineering Task Force RFC 1889*, January 1996

[42] J. Smith, R. Mohan, and C. S. Li, "Transcoding internet content for heterogeneous client devices" *IEEE International Conference on Circuits and Systems*, vol. 3, pp. 509-602, June 1998

[43] 3GPP, "Transparent End-to-End Packet switched Streaming Services (PSS), Protocols and codecs", *Technical Specification TS 26.234*

[44] E. Ekudden, R. Hagen, I. Johansson, and J. Svedberg, "The adaptive Multi-Rate Speech Coder" *IEEE Proceedings of Workshop on Speech Coding*, Porvoo, June 1999

[45] 3GPP, "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description", *Technical Specifications TS 26.171*

[46] MPEG-4 ISO/IEC, 1994, "Coding of Moving Pictures and Associated Audio – MPEG-4 Overview", *CD 14496-2*

[47] ITU-Telecom Standardization Sector of ITU, "Video Coding for Low Bit Rate Communication", *ITU-T Recommendation H.263*, March 1996

[48] T. Ahmed, A. Mehaoua, R. Boutaba, "Adaptive Packet Video Streaming Over IP Networks: A Cross Layer Approach", *IEEE Journal of Selected Areas in Communications*, vol. 23, no. 2, pp. 385-401, Feb. 2005

[49] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, pp. 301–317, Mar. 2001

[50] B. Housel, and D. Lindquist, "WebExpress: A System for Optimizing Web Browsing in a Wireless Environment", *MOBICOM'96 Demonstration Session*, November 1996

[51] M. Liljeberg, T. Alanko , M. Kojo, H. Laamanen, and K. Raatikainen, "Optimizing World Wide Web for Weakly Connected Mobile Workstations: An Indirect Approach", *Proceedings of the 2nd International Workshop on Services in Distributed and Networked Environments,* 1995

[52] 3GPP, "QoS Concept and Architecture", *Technical Specification TS 23.107*

[53] M. C. Chuah, and O. Yue, "Engineering Wireless Multimedia Networks for QoS Differentiation," *Technical Report*, CUHK, Sept. 2001

[54] S. Floyd, and V. Jacobson, "Link-sharing and Resource Management Models for Packets Network," *IEEE Vehicular Technology Conference '03*, Orlando, Oct. 2003

[55] M. Moustafa, I. Habib, and M. Naghshineh, "Wireless resource management using genetic algorithm for mobile equilibrium," *Elsevier Computer Networks*, vol. 37, no. 5, pp. 631-643, Elsevier Science, Nov. 2001.

[56] M. Moustafa, I. Habib, and M. Naghshineh, "Efficient Radio Resource Control in Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 2385-2395, Nov 2004

[57] D. Goldberg, "Genetic Algorithms in Search, Optimization, and Machine Learning", Reading MA, Addison-Wesley, 1989

[58] S. Shenker, D. Clark, D. Estrin, and S. Herzog, "Pricing in computer networks: Reshaping the research agenda," *ACM Computer Communication Review*, vol. 26, no. 2, pp. 19-43, April 1996

[59] L.A. DeSilva, "Pricing for QoS–Enabled Network: A Survey" *IEEE Communications Surveys*, May 2000

[60] A. Gupta, D.O. Stahl, and A.B. Whinston, "The Economics of Network Management" *Communications of the ACM*, vol.42, no. 9, Sept. 1999

[61] C. Lindemann, M. Lohmann, and A. Thummler, "Adaptive Call Admission Control for QoS/Revenue Optimization in CDMA Cellular Networks" *ACM Journal on Wireless Networks (WINET)*, vol. 10, pp. 457-472, 2004

[62] C. Houck, J. Jones, and M. Kay, "A genetic for function optimization: A Matlab Implementation," *NCSU-IE TR 95-09*, 1995

[63] S. Haykin, "Communication Systems", John Wiley & Sons, 2001

[64] P. Brady, "A Statistical Analysis of On-Off Patterns in 16 conversations," *Bell Systems Technical Journal*, vol. 47, no.1, pp. 73-91, 1968

[65] C. Blondia, O. Casals, "Performance Analysis of Statistical Multiplexing of VBR Sources," *IEEE Proceeding of INFOCOM'92*, pp. 828-838, 1992

[66] 3GPP, "Universal Mobile Telecommunication System (UMTS); Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS," *ETSI TR 101 112, version 3.2.0*

[67] 3GPP2, "Introduction to cdma2000 standards for spread spectrum systems", *3GPP2 C.S0001-A*, 2000

[68] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley, "On the capacity of a cellular CDMA system," *IEEE Transactions on Vehicular Technology*, vol. 40, pp. 303-312, May 1991

[69] S. Shin, K. Lee, and K. Kim, "Performance of the Packet Data Transmission Using the Other-Cell-Interference Factor in DS/CDMA Downlink," I*EEE Wireless Communications and Networking Conference 2002*, Orlando, March 2002

[70] G.M. Su, Z. Han, A. Kwasinski, M Wu, K.J. Liu, and N. Farvardin, "Distortion Management of Real-time MPEG-4 FGS Video over Downlink Multi-code CDMA Networks", *IEEE International Conference in Communications*, vol. 5, pp. 3071-3075, June 2004.

[71] P. Das and J. Y. Khan, "Quality of Service Based Resource Management for Packet Switched Data Over the WCDMA Uplink," *IEEE International Conference in Communications 2004*, Paris, France

[72] M. Wall, "GAlib: A C++ Library of Genetic Algorithm Components", *http://lancet.mit.edu/ga*

[73] D. Har, H. Xia, and H. Bertoni, "Path-Loss prediction model for microcells", *IEEE Transactions on Vehicular Technology*, vol. 48, no. 5, pp. 1453-1462, Sept. 1999.

Publications

1. F. De Angelis, and I. Habib, "An Adaptive Filtering Algorithm for Delivering IP Multimedia Services in W-CDMA Cellular Networks", paper submitted to **IEEE Transactions on Wireless Communications**.

2. F. De Angelis, and I. Habib, "Multimedia Content Delivery using Adaptive Filtering in Time-Varying W-CDMA Networks," paper accepted to **IEEE Global Telecommunications Conference (Globecom2005)**, St. Louis, USA, Dec. 2005.

3. F. De Angelis, S. Giannetti, G. Giambene, and I. Habib, "Scheduling for Differentiated Traffic Types in HSDPA Cellular Systems," paper accepted to **IEEE Global Telecommunications Conference (Globecom2005)**, St. Louis, USA, Dec. 2005.

4. F. De Angelis, I. Habib, F. Davide, and M. Naghshineh, "Increasing Revenues via Adaptive Filtering in Wired/Wireless Networks," paper to appear in **Elsevier Computer Networks Journal,** 2005.

5. F. De Angelis, and I. Habib, "Adaptive filtering for Delivering Multimedia Services in W-CDMA Cellular Networks," Proceedings of the **IEEE International Conference on Wireless Networks, Communications, and Mobile Computing (WirelessCom2005)**, Hawaii, USA, June, 2005.

6. F. De Angelis, P. Altieri, G. Giambene, and I. Habib, "Differentiated QoS Provision for Multimedia Traffic in Wifi Systems," Proceedings of the **IEEE International Conference on Communications (ICC2005)**, Seoul, Korea, May 2005.

7. F. De Angelis, I. Habib, F. Davide, and M. Naghshineh, "Intelligent Content Aware Services in 3G Wireless Networks", **IEEE Journal of Selected Areas in Communications (JSAC)**, vol. 23, no. 2, pp. 221-234, Feb. 2005. (Special edition on Intelligent Services and Applications in Next Generation Networks; invited paper).

8. F. De Angelis, I. Habib, F. Davide, and M. Naghshineh, "The Impact of Revenues on Delivering Differentiated IP Multimedia Services in Wired/Wireless Networks", Proceedings of the **3rd Workshop on QoS in Multi-service IP Networks (QoS-IP 2005)**, pp. 405-418, Catania, Italy, Feb. 2005.

9. F. De Angelis, I. Habib, F. Davide, and M. Naghshineh, "Improving the Delivery of Multimedia Services using an Adaptive Filtering Strategy", Proceedings of the **IEEE Global Telecommunications Conference (Globecom2003)**, vol. 6, pp. 3478-3482, San Francisco, USA, Dec. 2003.

10. F. De Angelis, I. Habib and, F. Davide, "QoS Filtering in Integrated Wired/Wireless Networks using Genetic Algorithms," Proceedings of the **IEEE Vehicular Technology Conference (VTC2003-Fall)**, vol. 5, pp. 3478-3482, Orlando, USA, Oct. 2003.

11. P. Barbano, F. De Angelis, and I. Habib, "Dynamic Re-Configuration in 3G Networks based on Adaptive Time/Frequency Filtering", Proceedings of the **IEEE Vehicular Technology Conference (VTC2003-Fall)**, vol. 3, pp. 1944-1948, Orlando, USA, Oct. 2003.