Intelligent Content Aware Services in 3G Wireless Networks

Flavio De Angelis, *Student Member, IEEE*, Ibrahim Habib, *Senior Member, IEEE,*
Fabrizio Davide, and Mahmoud Naghshineh, *Fellow Member, IEEE*

***Abstract*—In this paper, we address the problem of optimizing the delivery of multimedia services with different quality of service (QoS) requirements to mobile users. We assume that the network provides two distinct classes of service (CoS) to which users may subscribe: Premium, or Economy. Subscribers to the Premium service pay more for their connections but receive a higher level of quality measured by a set of parameters such as call blocking probability, coding rate, and format of the multimedia services. By optimizing the delivery of the multimedia services, we mean that the network guarantees that all users receive their agreed-upon contractual level of quality while maximizing the links' throughput, avoiding congestion, and maintaining the QoS requirements for each type of media (e.g., video, voice and data). Our proposed solution is based upon utilizing Genetic Algorithms (GAs) to solve a multi-objective optimization function that adaptively selects the downloading bit-rate for each type of traffic flow subject to the constraints of the optimization function. A traffic flow is an abstract of aggregate traffic of the same type of media (e.g., voice, video, or data) that is downloaded to a group of users who share some common attribute such as the same class of service. The optimization function is adaptive in the sense that the selected downloading bit-rates are time-dependant according to the dynamics of the links' traffic-loads, and users' requests. It is implemented on every output port of each node in the network. The function is used to control a filter that changes the coding rate of each media-type and, if necessary, performs transcoding of one, or more, media-types (i.e., video, voice, or data). Simulation results show significant improvement in terms of increasing the number of admitted users while maintaining the QoS requirements as well as target call blocking rates. 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.**

***Index Terms*—Genetic algorithms, IP, multimedia streams, quality of service, QoS filtering, 3G, wireless networks.**

Manuscript received November 1, 2003; revised June, 1 2004. This work was presented in part at VTC'03 and GLOBECOM'03.

F. De Angelis and I. Habib are with the Electrical Engineering Department, City University of New York, NY 10016 USA (e-mail: dean@ee-mail.engr.ccny.cuny.edu; habib@ccny.cuny.edu).

F. Davide is with Telecom Italia Learning Services, viale Parco dei Medici, 61, 00147 Rome, Italy.

M. Naghshineh is with IBM Technology Group, Yorktown Heights, NY 10598 USA.

I. Introduction

Delivering IP services with guaranteed quality of service requirements [1]-[3] to mobile users is a challenging area of research. This is due to two main reasons: 1) the difficulty of guaranteeing the QoS requirements while maximizing the efficiency of the IP network resources [4], and 2) the difficulty of maintaining those QoS requirements over the limited bandwidth wireless channel [5], [6] that exhibits time-varying impairments such as fading, shadowing and attenuation [7].

New IP enhanced services such as video and audio streaming, web browsing or voice-over IP consume a significant amount of the network resources (i.e., bandwidth, buffers). These resources would be wasted if the wireless channel did not have enough capacity to deliver the downloaded traffic to the mobile users. This scenario could happen because of the unpredictable time-varying characteristics of the wireless channel. Obviously, this scenario could result in re-transmissions, leading to congestion and further degradation in the network throughput [8], as well as violating the QoS requirements. Hence, it is important to devise a means of adaptively controlling the rates at which media streams are downloaded [9]-[13] according to the dynamic conditions of the wireless network, or more generally according to the dynamic conditions of the downstream links [14].

To illustrate, consider downloading a web page with images, voice, and video contents to a group of mobile users. If the wireless channel conditions do not permit the delivery of the downloaded information to the mobile users, then this information will create congestion episodes in the access nodes, causing significant decrease in the throughput of all downstream links. Furthermore, the users will perceive significant delays in the speed of their download. Hence, it is important to design an "adaptive" resources management algorithm that can avoid possible congestion episodes and, accordingly, adapt both the amount of the multimedia content as well as the rate at which this content is downloaded. The QoS filtering algorithm that we propose and analyze in this paper is an example of such adaptive algorithms [15], [16].

The proposed algorithm is used to control the actions of filters that are located in every node of the network. Accordingly, the filters are tuned to select sub-optimal bit-rate values that are required to download each traffic flow. This is done via a two steps approach: first, the coding rate of the traffic flows is reduced [17] without transcoding; and then, if necessary, the coding rate is further reduced via transcoding [18]-[20]. For example, a possible action of one filter could be to change the bit-rate of a voice traffic flow with a Wideband Adaptive Multi-rate (W-AMR) coding format by scaling the coding rate from 24 kbps to 16 kbps, or transcoding it from W-AMR into AMR coding format (coding rate is from 12 kbps to 4 kbps). Similarly, in the case of a video traffic flow with MPEG-4 coding format, a filter may scale the coding rate from 512 kbps to 64 kbps, or trans-code the same flow from MPEG-4 to H.263 coding format (coding rate is changed from 64 kbps to 32 kbps). 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. 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. Since all proposed filters are able to change coding formats, and rates, by selecting several filtering levels of a media stream, we define them as "multi-scale". The action performed by each multi-scale filter is essentially to meet the following requirements

1. Avoid congestion on the end-to-end path including the radio link.
2. Provide preferential treatment to Premium class subscribers over Economy class users.
3. Ensure that the QoS parameters of each traffic flow are maintained, regardless of the class of the user.
4. Maximize the utilization of each link and the number of users that can be admitted to the network.

The rest of the paper is organized as follows. In Section II, we present our QoS filtering architecture. Next, in Section III, we define the multi-objective optimization function. In Section IV, the Genetic Algorithm is presented. Simulations results are provided in Section V and the conclusions are given in Section VI.

II. THE QOS FILTERING ARCHITECTURE

In the network architecture shown in Fig. 1, different types of media contents are stored in a number of media servers. Each server is connected to high-speed routers in the core network. These routers are inter-connected in some mesh network configuration via high-speed fiber links. The radio access network is made up of radio network controllers and base stations. We depict, however, only the one-way communication from the media servers to the mobile end-users (downstream).

In our proposed architecture, the multi-scale filters are installed at the output ports of every router, server, as well as the radio network controllers (RNC) in the wireless network. This is needed in order to react to local congestion and maintain the QoS requirements in every router along the end-to-end path. Our results show that the filters in the core network do not need to be equipped with transcoding capabilities (see section V.1). On the other hand, the filters in the RNCs would support transcoding in order to react to the time-dependant variations in the cellular network capacities due to users' mobility, as well as radio channel impairments.

The filters control the traffic flows and adapt them according to the different conditions of network. In the core network, a traffic flow represents the aggregate traffic of similar media streams (e.g. voice or video 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 Label Switched Path (LSP) in each downstream link. Each filter located in the output ports of the routers 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.

In this paper, we intend to construct an algorithm that controls the multi-scale filters located in the core network. A multi-scale filter is an entity that manipulates adaptively all types of incoming media streams downloaded to the mobile users. The types of media streams that can be downloaded in the network are video, speech, text and still images. Each media-type can assume the different coding formats that are represented in Table I and are used in 3G-UMTS (Universal Mobile Telecommunications System) networks [21].

The filter selects a specific filtering level $f^{[k]}$ for the k-th downloaded traffic flow. Each filtering level is selected from a set of M values according to the filter granularity and processing power

$$f^{[k]} \in \{1, 2, 3, ..., M\}. \quad (1)$$

The maximum filtering 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 the minimum bit-rate for downloading a flow as well as the minimum quality of the image, voice or video media. This is illustrated in Table II.

The choice of a filtering level establishes a specific coding rate $R^{[k]}$ and coding format $CF^{[k]}$ for the correspondent media streams

$$R^{[k]} \in \{R_1, R_2, R_3, ..., R_M\}, \quad (2)$$

$$CF^{[k]} \in \{CF_1, CF_2\}. \quad (3)$$

As indicated in Table II, by increasing the filtering level, the quality of the downloaded media improves in term of higher coding rates and better coding formats, so the highest filtering level M is associated with the highest level of quality.

The type of coding format defines the compression method used to encode each media-type. The values of coding rates determine the bit-rate values for downloading the media content included in the k-th traffic flow. The minimum coding rate corresponds to a guaranteed bit-rate, while the highest coding rate is associated with the maximum downloading bit-rate

$$R_{guar} < R^{[k]} < R_{\max}. \quad (4)$$

The choice of a specific filtering level with its corresponding coding rate, type of coding format that can be chosen for the k-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. In our work, we define two different types of users or "classes of service": *Premium* (P) and *Economy* (E).

In Table III, we show a possible assignment of filtering levels, coding rates and coding formats for Premium and Economy class traffic flows. 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).

As shown in Table III, 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 filtering 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. Moreover, in case of congestion, traffic flows that belong to Premium class users will be degraded much more gracefully that those of the Economy class users. Furthermore, Premium class users have also higher priority to be admitted in the network in comparison with Economy class users.

The most significant feature of the filtering strategy is that the range of selectable coding rates and types of formats for each traffic flow provide 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 and stored in a management database.

An important QoS parameter that we need to consider for the media content delivery is the end-to-end transfer delay. We assume that the media types included in a traffic flow are downloaded according to the pre-requisites of a specific QoS class of service defined within the specifications of Third Generation Partnerships Project (3GPP): *Conversational, Streaming, and Interactive* [2]. 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 to download multimedia contents and predefined end-to-end delay constraints [2], we can estimate, for each downstream link, minimum and maximum transfer delays for each k-th traffic flow

$$\Delta_{\min} < \Delta^{[k]} < \Delta_{max}. \quad (5)$$

Fig. 2 represents the block diagram of the Resource Manager. It includes a traffic classifier, the multi-scale filter, a class-based queuing system, a traffic dispatcher, an optimization function and a management database. Once all routing operations are performed within the node, traffic is directed to the appropriate output port where the filter is installed.

As shown in Fig. 2, $[T_{\alpha,\beta}^{[k]}]_{in}$ represents the k-th flow of the output port, characterized by media-type α and downloaded to the same category β of users. This flow is associated with a filtering vector $[f^{[k]}, R^{[k]}, CF^{[k]}]_{in}$ and with the current number $N^{[k]}$ of media streams included in the same flow. During a certain time, the multi-scale filter manipulates the k-th flow by assigning, if it is necessary, a new output vector

$$[f^{[k]}, R^{[k]}, CF^{[k]}]_{in} \Rightarrow [f^{[k]}, R^{[k]}, CF^{[k]}]_{out}. \quad (6)$$

If output and input vector parameters are different, an additional filtering delay $\delta^{[k]}$ 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, we introduce weights $[w_s, w_v, w_d]$ for each queue (Conversational, Streaming, and Interactive). We assume, 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.

The optimization function in Fig. 2 is implemented using GAs. Its details are provided in the next Section. All

information related to the feasible range of coding rates and formats for each type of flow is stored in the management database where each user profile is recorded. This information, which is provided in Table III, defines the solution space for the optimization function. Furthermore, the management database contains the values for the minimum and maximum transfer delay for each traffic flow downloaded in the output link.

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.

In Fig. 3, we show an example for downloading 32 Mb video content 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.

III. THE OPTIMIZATION FUNCTION

The optimization function is used to control the actions of both the multi-scale 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 we explained in the previous Section, *k* is the identification number of the traffic flows (in this case from one to $\alpha \times \beta$). Each flow is characterized by the same type of media streams ($\alpha \in \{s, v, d\}$). Each flow is downloaded to a group of users ($\beta \in \{E, P\}$). In this Section, we will use only the parameters α and β 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 $[f_{\alpha,\beta}, R_{\alpha,\beta}, CF_{\alpha,\beta}]_{in}(t)$. 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 $[f_{\alpha,\beta}, R_{\alpha,\beta}, CF_{\alpha,\beta}]_{out}(t+\Delta t)$ 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 management database includes the information of the transfer delay constraints $(\Delta^{\min}_{\alpha,\beta}, \Delta^{\max}_{\alpha,\beta})$ and the range of filtering levels, coding rates and formats represented in Table III. Finally, let $[f_{\alpha,\beta}, R_{\alpha,\beta}, CF_{\alpha,\beta}]_{\min}$ be the vector with the minimum values for the filtering level, coding rate and format contained in Table III.

The optimization function F is defined as the product of three different objective functions

$$F = F_{Util} \cdot F_R \cdot F_{\Delta}. \tag{7}$$

The objective here is to find the output traffic flow and the weights vector which maximize the function F and satisfy the following constraints:

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

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

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

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

where $R_{tot}(t+\Delta t)$ represents the throughput for the downstream link.

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

$$F_{Util} = \begin{cases} R_{tot}(t+\Delta t)/C_{av}(t), & if \quad R_{tot}(t+\Delta t) \le C_{av}(t) \\ 0, & otherwise. \end{cases} \tag{8}$$

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

$$F_R = \prod_{\alpha=s,v,d} F_D^{\alpha} \cdot \Big(\sum_{\alpha=s,v,d} F_R^{\alpha,P} + \sum_{\alpha=s,v,d} F_R^{\alpha,E} \Big) / 6 \qquad (9)$$

where

$$F_R^{\alpha,P} = \begin{cases} \left[\dfrac{(f_{\alpha,P})_{out}(t+\Delta t) - (f_{\alpha,P})_{\min}}{(f_{\alpha,P})_{in}(t) - (f_{\alpha,P})_{\min}} \right]^2, & if \quad (f_{\alpha,P})_{\min} \le (f_{\alpha,P})_{out}(t+\Delta t) \le (f_{\alpha,P})_{in}(t) \\ 0, & otherwise, \end{cases} \qquad (10)$$

$$F_R^{\alpha,E} = \begin{cases} \sqrt{\dfrac{(f_{\alpha,E})_{out}(t+\Delta t) - (f_{\alpha,E})_{\min}}{(f_{\alpha,E})_{in}(t) - (f_{\alpha,E})_{\min}}}, & if \quad (f_{\alpha,E})_{\min} \le (f_{\alpha,E})_{out}(t+\Delta t) \le (f_{\alpha,E})_{in}(t) \\ 0, & otherwise, \end{cases} \qquad (11)$$

$$F_D^{\alpha} = \begin{cases} 0, & if \quad \dfrac{(f_{\alpha,P})_{in}(t) - (f_{\alpha,P})_{out}(t+\Delta t)}{(f_{\alpha,P})_{in}(t)} > \dfrac{(f_{\alpha,E})_{in}(t) - (f_{\alpha,E})_{out}(t+\Delta t)}{(f_{\alpha,E})_{in}(t)} \\ 1, & otherwise. \end{cases} \qquad (12)$$

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:

$$D_{\alpha,\beta}(t+\Delta t) = \frac{(f_{\alpha,\beta})_{in}(t) - (f_{\alpha,\beta})_{out}(t+\Delta t)}{(f_{\alpha,\beta})_{in}(t)}. \qquad (13)$$

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 III.

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. We assume that each class-based queue can be modeled as a simple M/M/1 queuing system and we calculate queuing and transmission delay based upon this approximation. The total transfer delay of a traffic flow for a downstream link can be written as

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

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:

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

The delay function is given by

$$F_\Delta = F_w \cdot \prod_{\alpha=s,v,d} \prod_{\beta=P,E} F_\Delta^{\alpha,\beta}. \qquad (16)$$

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

$$F_{\Delta}^{\alpha,\beta} = \begin{cases} 1, & if \quad \Delta_{\alpha,\beta}^{\min} \le \Delta_{\alpha,\beta}(t+\Delta t) \le \Delta_{\alpha,\beta}^{\max} \\ 0, & otherwise, \end{cases} \tag{17}$$

$$F_{w} = \begin{cases} 1, & if \quad \sum_{\alpha=s,v,d} w_{\alpha}(t+\Delta t) = 1 \\ 0, & otherwise. \end{cases} \tag{18}$$

To perform the optimization, the Genetic Algorithms (GAs) are used. In the following Section, we provide the details of the GA.

IV. GENETIC ALGORITHM DETAILS

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. However, these techniques are not suitable in our case since the objective function includes the product of step functions that are not differentiable.

Considering search techniques, we find many methods: *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) [22] 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. Genetic Algorithms have been used to address diverse practical optimization problems related to dynamic resource management, see [23] and [24].

The basic elements of a GA are a chromosome and a fitness function. In our case, a chromosome is a vector whose elements are simply the filtering levels $(f_{\alpha,\beta})_{out}(t)$ of the output traffic flow $[T_{\alpha,\beta}]_{out}(t)$ and the queuing weights $\underline{w}(t)$. Each chromosome is usually represented by a binary vector. In [25] it has been shown that natural representations are more efficient and produce better solutions, thus we used a floating-point representation [26]. A fitness function is the objective function defined by equation (7) that needs to be optimized. GAs includes some operators like reproduction, crossover and mutation that are applied to successive populations of chromosomes to create new better ones. The operation of GA is depicted in Fig. 4 and is summarized in [27].

Once the stopping criterion is reached, the algorithm sends new values to the filter $[f_{\alpha,\beta}]_{out}(t+\Delta t)$ 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 $[f_{\alpha,\beta}]_{out}(t+\Delta t)$ is different from that of the input $[f_{\alpha,\beta}]_{in}(t)$. 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. In case the number of traffic flows is high, the solution space of the optimization problem contains many areas where the value of the fitness function is zero (i.e., no solution). This is due to the several constraints taken into consideration in the optimization problem and the numerous allowable combinations of the weights. The queuing weights range between zero and one with a floating-point accuracy of 10^{-4}. To prevent the GA from being trapped in these areas, we generate an initial population with some possible solutions that include appropriate combinations of the weights. The initial population consists of chromosomes $[f_{s,P} \;\; f_{v,P} \;\; f_{d,P} \;\; f_{s,E} \;\; f_{v,E} \;\; f_{d,E} \;\; w_s \;\; w_v \;\; w_d]$ that include all the possible combinations of filtering levels $f_{s,P} \in \{2,5,6,7\}$, $f_{v,P} \in \{2,4,5,6\}$, $f_{d,P} \in \{3,4,5,6\}$, $f_{s,E} \in \{1,2,3,4\}$, $f_{v,E} \in \{1,2,3,4\}$, $f_{d,E} \in \{1,2,3,4\}$, with appropriate choices for the weights as follows:

$$\begin{cases} w_s = \dfrac{N_{s,P} \cdot R_{s,P} + N_{s,E} \cdot R_{s,E}}{R_{tot}} \\ w_v = \dfrac{N_{v,P} \cdot R_{v,P} + N_{v,E} \cdot R_{v,E}}{R_{tot}} \\ w_d = \dfrac{N_{d,P} \cdot R_{d,P} + N_{d,E} \cdot R_{d,E}}{R_{tot}} \end{cases} \tag{19}$$

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

By this initial population, we reduce, in case the number of flows $N_{\alpha,\beta}$ is high, the possibilities of obtaining no solutions (fitness function zero) when some solution exists.

V. SIMULATIONS RESULTS

In this Section, we explain the simulations experiments that have been carried out to evaluate the performance of the proposed algorithm. Different multimedia contents are downloaded to users via a 15 Mbps link. The transfer delay constraints that we consider for each traffic flow are provided in Table IV. We assume that the core network contains three nodes (routers), i.e., two hops. The transfer delay constraints for each traffic flow are obtained by dividing the maximum transfer delay in the core network (Table IV) by the number of hops. The filtering delays δ_1, δ_2 are set to 5 ms and 9 ms, respectively. We consider an average packet length of L=1000 bits. Each filter selects values for the coding rates, and formats, for the output traffic flows as indicated in Table III. To simplify the simulations, we assume that each of the incoming traffic flows is associated with the highest filtering level described in Table III. The filtering levels of the incoming traffic flows are fixed for our simulations, though in general these could be time-dependant. To test the algorithm, we compare a single node case without filtering (Case A), with another with filtering (Case B), see the architecture in Fig. 5.

V.1. Experiment 1

In this experiment, we evaluated the effectiveness of the proposed algorithm to enhance the capacity of the network in terms of the maximum number of users that could be admitted to the network without violating the pre-defined QoS requirements. To a service provider, this term signifies increasing revenues via utilizing the existing infrastructure more efficiently without offering new services that may require substantial new investments to upgrade the existing network facilities. Performance parameters of interest include the links' utilization, transfer delay of each traffic flow, and the average filtering degradation in the quality of the media content for the traffic flows downloaded by the Premium and Economy class users. Finally, we computed the increase in the revenues achieved by using the proposed algorithm and how this limits the number of admitted users.

In this experiment, each user requests a multimedia call containing all types of media: voice, video and data. First, we gradually load the link with only Premium class calls until the last user is blocked. We, then, repeat this experiment taking into consideration all possible combinations of Premium and Economy class users. Each simulation continues until the last user is blocked. The results are shown in Fig. 6. 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.

In Fig. 7, we compare 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. 8. This graph is a cross section of the 3D graphs shown in Fig. 7, 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 we overload the system, only the filtering case can support the users while maintaining high utilization, and without violating the delay constraints.

The values of the transfer delays of each traffic flow (voice, video, and data) are provided in Fig. 9. In the graph, the percentage of Premium and Economy class users is still maintained at 50% each, as in Fig. 8. For a node with filtering, we depict the transfer delays for both Premium and Economy class traffic flows. 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 we overload the system, 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. 9). 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. 8). 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.

In order to evaluate the effect of the filtering actions taken by the filter on the Premium and Economy class traffic flows, we define an average filtering reduction level for the Premium and Economy traffic flows. 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:

$$D_{\beta} = \frac{1}{3} \cdot \sum_{\alpha=s,v,d} D_{\alpha,\beta}. \quad (20)$$

In Fig. 10, we show the impact of the reduction in the filtering levels (or coding rates) on different traffic flows in three distinct scenarios. 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 we have mentioned in Section III, the filtering reduction is not directly proportional to the content reduction. In this case, if we replace the filtering levels by the coding rates in equations (13) and (20), we obtain an average reduction of 73% for the multimedia content. 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 play a more significant factor in the optimization function. In fact, as we have indicated in Section III, 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 Figure 10 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. 10 where 10% of Premium and 90% of Economy class users are admitted. Since this scenario can be considered as the most realistic case, we also present in Fig. 11 the coding rate values for all Premium and Economy traffic flows against the number of users.

As shown in Fig. 11, 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 with 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.

Fig. 12 elaborates the limitations on the gain achieved by increasing revenues via admitting more users. In computing the revenues, we used a simple revenue model [29] as follows:

$$\Phi = \sum_{\alpha=s,v,d} \sum_{\beta=P,E} N_{\alpha,\beta} \cdot T_{\alpha,\beta} \cdot R_{\alpha,\beta}. \quad (21)$$

The revenue measure Φ describes the total revenues generated due to the traffic downloaded by both Premium and Economy class users in a control period (Δt). In this model, $T_{\alpha,P}$ and $T_{\alpha,E}$ are the tariffs that Premium and Economy class users pay in a control period in order to download each type

of media stream. Without loss of generality, we assumed that the tariff is expressed in dollars /Mbps. The figure demonstrates two distinct scenarios.

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

$$T_{\alpha,P} = \begin{cases} 0.20 \ \$/Mbps, & f_{\alpha,P} > 4 \\ 0.15 \ \$/Mbps, & otherwise, \end{cases} \quad (22)$$

$$T_{\alpha,E} = 0.10 \ \$/Mbps, \ \forall f_{\alpha,E}. \quad (23)$$

In scenario 2, the costs 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:

$$T_{\alpha,P} = \begin{cases} 0.20 \ \$/Mbps, & f_{\alpha,P} > 4 \\ 0.10 \ \$/Mbps, & otherwise, \end{cases} \quad (24)$$

$$T_{\alpha,E} = \begin{cases} 0.10 \ \$/Mbps, & f_{s,E} \geq 2; f_{v,E} \geq 2; f_{d,E} > 2 \\ 0.05 \ \$/Mbps, & otherwise. \end{cases} \quad (25)$$

The increase in the revenue, achieved by the filtering, saturates at 100 users for scenarios 1 and 2, although, Fig. 12 indicates that approximately 220 users could be admitted. This means that there is no advantage for a service provider to admit any more users above this limit. In fact, admitting more users does not represent any additional revenue gain for the provider. This is an interesting result since it limits the operating region for the service provider despite the fact that the filtering algorithm would maintain the QoS parameters of each traffic flow even for a larger number of users.

Note that the revenues for the no filtering case is maximum at approximately 1.4 $ for 40 users since the network cannot admit more than 40 users. For the filtering case, however, the revenues reach 2.3 $ for 100 users. The revenue gain is approximately 50%, whereas the number of admitted users has more than doubled. Note that with the filtering case, each user is now paying less $/Mbps of usage. Hence, this is a more favorable operating region for both users and the operator alike.

In Fig. 13, we show the offered traffic-loads that trigger the activation of rate reduction via transcoding. Clearly, transcoding is activated only at high traffic-loads (100 users and above is the equivalent to more than 90% utilization).

We recall that the operating region of the provider has a maximum of 100 users due to saturation of revenues. Hence, transcoding would be rarely activated inside the core network; whereas, it would be more activated in the RNCs due to the fact 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.

V.2. Experiment 2

The second experiment was performed 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 III, are satisfied for the existing calls and 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.

In Fig. 14, we compare the call blocking rates between the case of no-filtering (Case A), and that with filtering (Case B). 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 we consider 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.

Finally, Fig. 15 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. 15 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. In this scenario, Premium class calls are blocked with higher percentages because some Economy class calls have been accepted in the previous control periods.

VI. Conclusions

In this paper, we presented a novel filtering strategy for controlling, adaptively, both the allocated bandwidth and the transfer delay of traffic flows downloaded from content servers to mobile users. A multi-objective optimization function has been proposed and solved using Genetic Algorithms in order to meet users' QoS demands while maintaining high links' throughput. Users are 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 strategy provides the network operator with the flexibility to control not just the coding rate of the content but also its type as well. The filters select the optimal coding rate and format for each traffic flow according to the dynamics of the links' traffic-loads and users' requests. Performance evaluation results show that the proposed filters provide significant gains in term of reducing the call blocking probability (an important measure 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 is unavailable. These gains have been achieved at a small price, which is a graceful degradation in the quality of the media contents that happens only when the network is overloaded. Finally, we note that the increase in the revenues saturates at a certain number of users beyond which admitting any more users will not be cost-effective for the operator.

References

[1] C. Aurrecoechea, A. Campbell, and L. Hauw, "A survey of QoS Architectures," *ACM/Springer Verlag Multimedia Systems Journal*, vol.6, no.3, pp. 138-151, May 1998.

[2] 3GPP, Technical Specification TS 23.107, *QoS Concept and Architecture*.

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

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

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

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

[7] S. Rappaport, *Wireless Communications*, Prentice Hall, New Jersey, 1996.

[8] A. Campora, "Wireless ATM: A perspective on issues and prospects," *IEEE Trans. Pers. Comm.*, vol. 3, no. 4, pp. 8-17, Aug. 1996.

[9] 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," *In Proceeding of ACM Multimedia '93*, Aug. 1993.

[10] 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.

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

[12] F. Garcia, D. Hutchinson, A. Mathue, and N. Yeadon "QoS support for distributed multimedia communication" *Proc. 1st Int. Conf. on Distributed Platforms*, Feb. 1996.

[13] N. Yeadon, F. Garcia, D. Hutchinson, and D. Shepherd, "Filters: QoS Support Mechanisms for Multi-peer Communications," *IEEE J. Select. Areas Commun.*, vol. 14, no. 7, pp. 1245-1262, Sept. 1996.

[14] M. Mirhakkak, N. Schult, and D. Thomson, "Dynamic Bandwidth Management and Adaptive Applications for a Variable Bandwidth Wireless Environment", *IEEE J. Select. Areas Commun.*, vol. 19, no. 10, pp. 1984-1997, Oct. 2001.

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

[16] 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.

[17] A. Vetro, H. Sun, and Y. Wao, "MPEG-4 Rate Control for Multiple Video Objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol.9, no. 1, pp. 186-199, Feb. 1999.

[18] A. Vetro, H. Sun, and Y. Wang, "Object-based Transcoding for Adaptable Video Content Delivery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 387-401, Mar. 2001.

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

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

[21] 3GPP, Technical Specification TS 26.234, *Transparent End-to-End Packet-switched Streaming Service (PSS), Protocols and codecs*.

[22] D. Goldberg, *Genetic algorithm in search, optimization and machine learning*, Addison Wesley, MA, 1989.

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

[24] M. Moustafa, I. Habib, and M. Naghshineh, "Efficient Radio Resource Control in Wireless Networks," *IEEE Trans. Wireless Commun.*, to be published.

[25] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, 1996.

[26] C. Houck, J. Joines, and M. Kay, "A genetic algorithm for function optimization," *NCSU-IE TR 95-09*, 1995.
[27] L. Davis, *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, NY, 1991.
[28] M. C. Chuah and O. Yue, "Engineering Wireless Multimedia Networks for QoS Differentiation," Technical Report, CUHK, Sept. 2001.
[29] 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.

Flavio De Angelis received a degree in telecommunication engineering cum laude from the University of Rome "Tor Vergata", Rome, Italy, in 1999. From 2000 to 2001, he worked as a network systems engineer in Comverse Technology where he was involved in configuring, testing and troubleshooting multimedia messaging systems for both wireless, and wireline networks.

He is currently pursuing the Ph.D. degree at the City University of New York, USA. His current research interests are in the areas of performance analysis, quality of service, and allocation of resources in wireless and IP networks.

Ibrahim Habib received the Ph.D. degree from the City University of New York, the M.Sc. degree from Polytechnic University of New York, USA, and the B.Sc. degree from Ain Shams University, Cairo, Egypt all in Electrical Engineering. From 1981 till 1983, and from 1984 till 1988, he was a computer networks engineer working on the planning, system engineering and installation of several IBM SNA networking projects in Egypt and Saudi Arabia. In 1991 he joined the Faculty of the City University of New York where is now a Full Professor. His research interests span different areas of traffic engineering in IP, wireless, and optical networking. He has published more than 80 technical papers and reports in these areas. From 1997 till 2000 he was with AT&T Labs, and from 2000 till 2001 with Telcordia Technologies, Applied Research Department working on the architecture design of IP over optical networks, optical control plane and Operations Support Systems (OSS). He is currently a Guest Editor of the IEEE Journal on Selected Areas in Communications (JSAC) on Metro Optical Networks, 2004. He was also a Guest Editor of the same Journal in 1997, and 2000. He served as a Guest Editor of the IEEE Communications Magazine twice in 1995 and 1997, and was an Editor of the same Magazine from 1994 till 1997. He is currently an Editor of the John Wiley Journal on Wireless Communications and Mobile Computing. He was the co-chairman of the Wireless Networking and Optical Networking Tracks at the IEEE GLOBECOM 2001, the High Speed Networks Symposium at ICC 2002, the chairman of both the Optical Network Symposium at GLOBECOM' 2003 and the Wireless Networking Symposium at ICC' 2004. He is listed in the Marquis's who is who in the World 2001, and 2003, and who is who in America 2002, and 2003, and 2004 editions.

Fabrizio A.M. Davide received a degree in electronic engineering cum laude from the University of Bologna, Italy, in 1988, and the Ph.D. in electronic engineering from the University of L'Aquila, Italy, in 1994.
He is currently an executive in the Telecom Italia Group. He is Director of Telecom Italia Learning Services Engineering where he leads a team of professionals working system integration and IT architectures, KM and e-learning application operation, and researchers working in the areas of technology enabled learning and cognition, hybrid biological-electronic systems, micro-sensors, virtual reality-based communications, non-linear dynamics theory and applications. He is also Adjunct Professor at Linköping University, Sweden, and Invited Professor at the University of Rome "Tor Vergata", Italy. He has published over 80 papers in refereed journals and conferences, as well as several book chapters and has a pending international patent application. He is editor of the series *Emerging Communication* (Amsterdam, Netherlands, IOS Press), and guest editor of *Sensors and Microsystems* (Singapore, World Scientific).

He has been a member of program committees and reviewer for several international conferences on sensors and system engineering (including *Biosensors and Bioelectronics* and *Eurosensors*), and has served on the Scientific Committees for eBeW (e-Business and e-work Conference), MEDICI (European Commission Framework of co-operation for multimedia access) and AISEM (the Italian Association for Sensors and Microsystems). He has coordinated a number of EU-funded R&D projects in the *Information Society Technologies Program* and has been a member of European Union panels for the design of research programs and the evaluation of research proposals.

Dr. Mahmoud Naghshineh is Director of emerging markets at IBM Technology Group. He is responsible for technology roadmap in the embedded space and defining software, services and solutions in support of Technology Group's offering. Prior to his current position, he was a Senior Manager at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, where he managed the Pervasive Computing Infrastructure Department and had world wide responsibility for IBM research's projects in Pervasive Computing and directs the work of more than 50 researchers at the TJ Watson Research Center. He manages projects in software and services infrastructure pervasive computing, mobile and wireless Internet, embedded software, secure platforms, network processor technologies, telecommunications services, and web-based payment infrastructure. He joined IBM in 1988. Prior to his current position, he has worked on communication and networking protocols, fast packet-switched/broadband IP networks, wireless and mobile networking optical networking, QoS provisioning, call admission, routing, and resource allocation, and network security. He has had several main technical contributions to IBM products in areas of networking technologies and software. He has contributed to IETF mobility protocols, IrDA, Bluetooth and IEEE 802.11 and .15 standards.

He received his doctoral degree form Columbia University, New York, in 1994. He is a Fellow member of the IEEE. Aside from IBM, he has been very active in external research and industrial forums. He is currently the Editor-in-Chief of IEEE Wireless Communications Magazine a leading magazine in mobile and wireless Internet. He has served as Program co-chair of MobiCom 2001 and as a technical editorial board member of many wireless and mobile networking/computing journals, as a member of technical program committee, session organizer and chairperson for many IEEE/ACM, NSF and Government conferences and workshops. Currently, he is an adjunct faculty member of the department of electrical engineering at Columbia University teaching a graduate course on wireless and mobile networking. He has published over 100 technical papers and holds a number of IBM awards and patents.

Figures captions:

Tables captions:

FIGURES

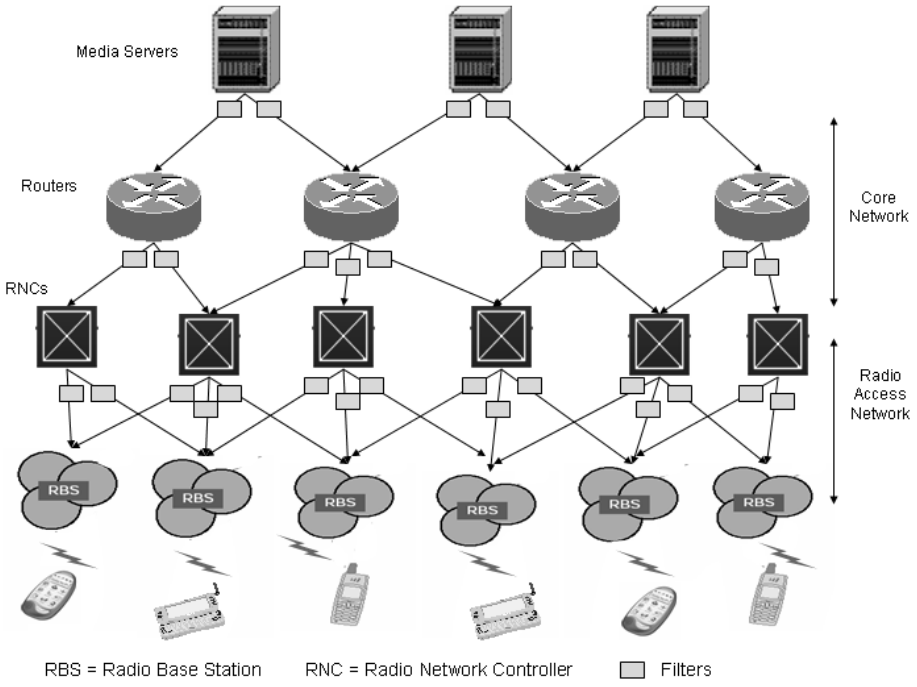

Fig. 1

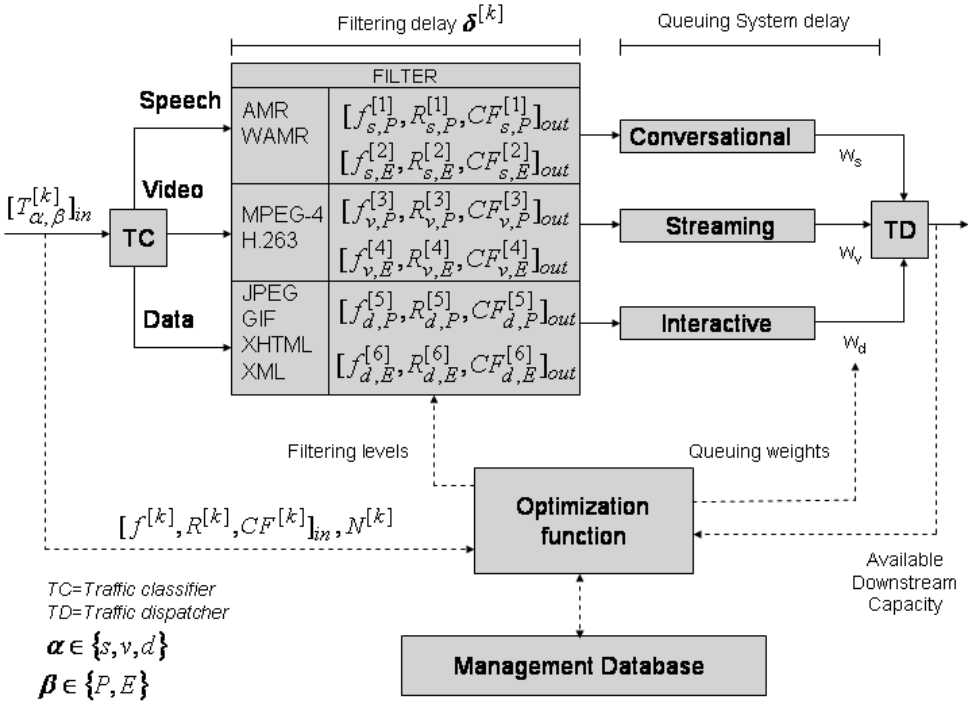

Fig.2

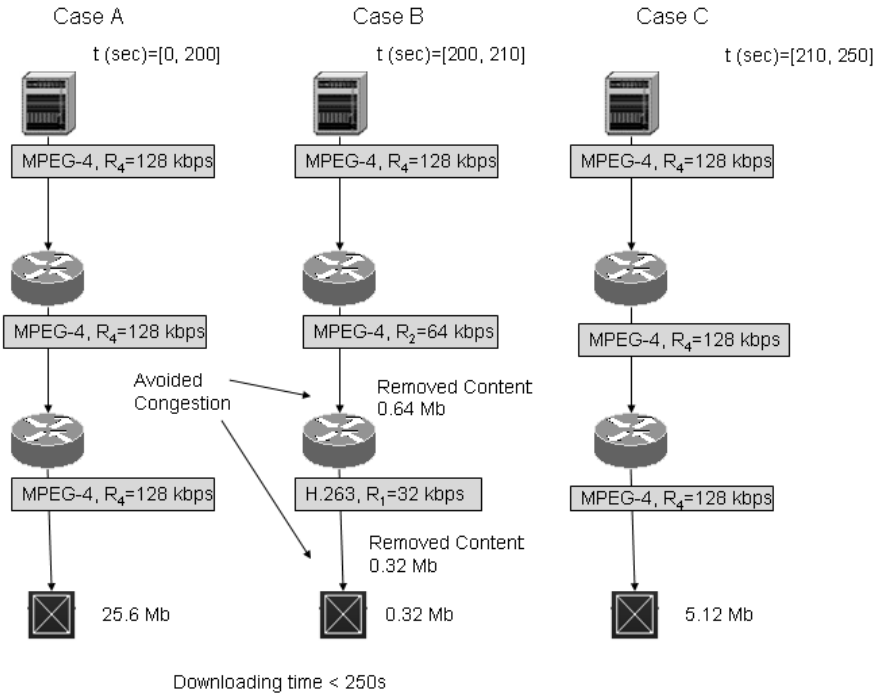
Case A
t (sec)=[0, 200]
MPEG-4, R4=128 kbps
MPEG-4, R4=128 kbps
MPEG-4, R4=128 kbps
25.6 Mb
Case B
t (sec)=[200, 210]
MPEG-4, R4=128 kbps
MPEG-4, R2=64 kbps
Avoided Congestion
Removed Content 0.64 Mb
H.263, R1=32 kbps
Removed Content 0.32 Mb
0.32 Mb
Case C
t (sec)=[210, 250]
MPEG-4, R4=128 kbps
MPEG-4, R4=128 kbps
MPEG-4, R4=128 kbps
5.12 Mb
Downloading time < 250s

Fig. 3

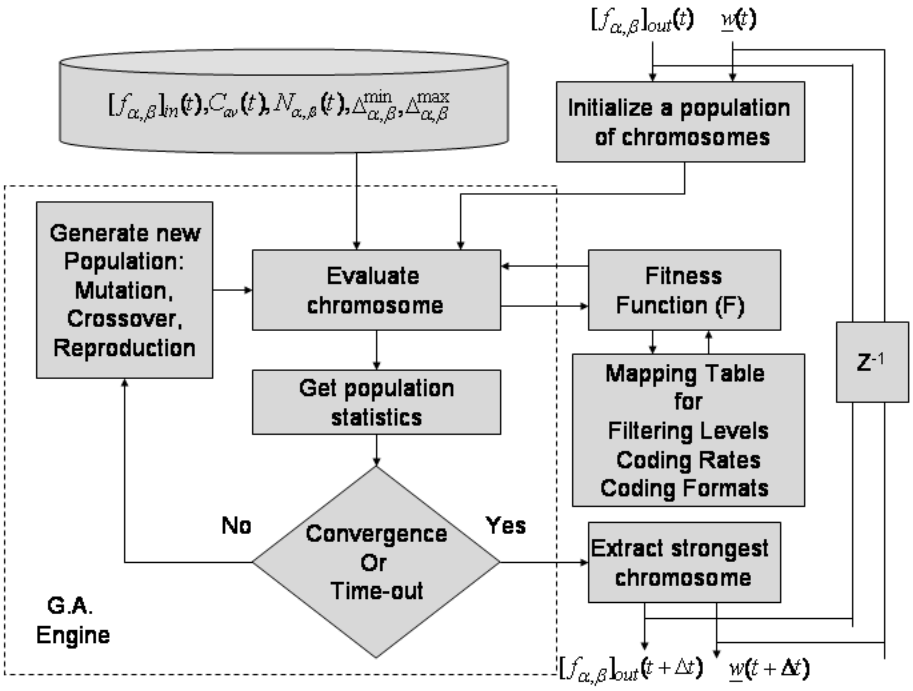
$[f_{\alpha,\beta}]_{out}(t)$
$\underline{w}(t)$
$[f_{\alpha,\beta}]_{in}(t), C_{\alpha}(t), N_{\alpha,\beta}(t), \Delta^{\min}_{\alpha,\beta}, \Delta^{\max}_{\alpha,\beta}$
Initialize a population of chromosomes
Generate new Population: Mutation, Crossover, Reproduction
Evaluate chromosome
Fitness Function (F)
Get population statistics
Mapping Table for Filtering Levels Coding Rates Coding Formats
Z-1
No
Convergence Or Time-out
Yes
Extract strongest chromosome
G.A. Engine
$[f_{\alpha,\beta}]_{out}(t+\Delta t)$
$\underline{w}(t+\Delta t)$

Fig. 4

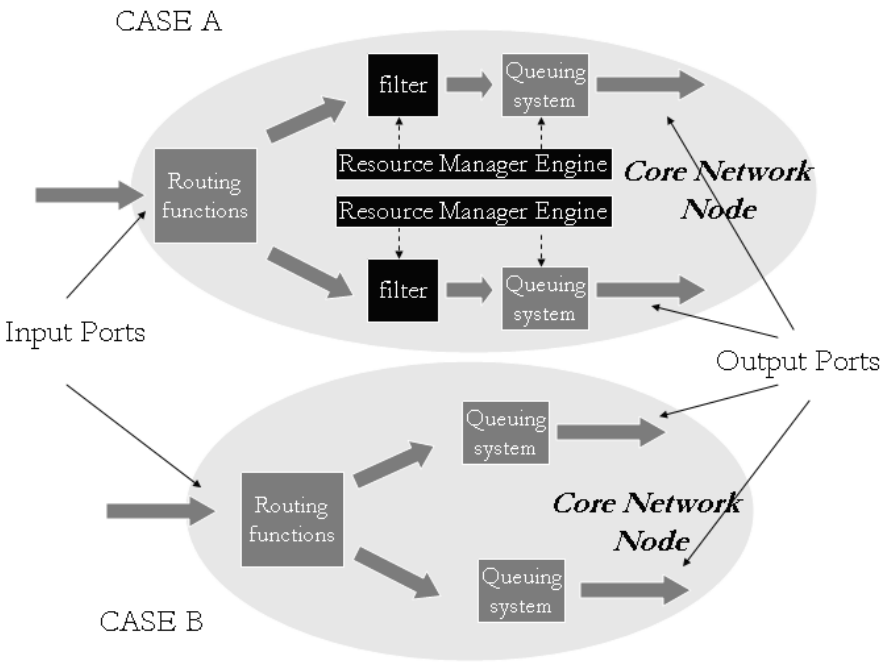
CASE A
filter
Queuing system
Resource Manager Engine
Core Network Node
Routing functions
Resource Manager Engine
filter
Queuing system
Input Ports
Output Ports
Queuing system
Routing functions
Core Network Node
Queuing system
CASE B

Fig. 5

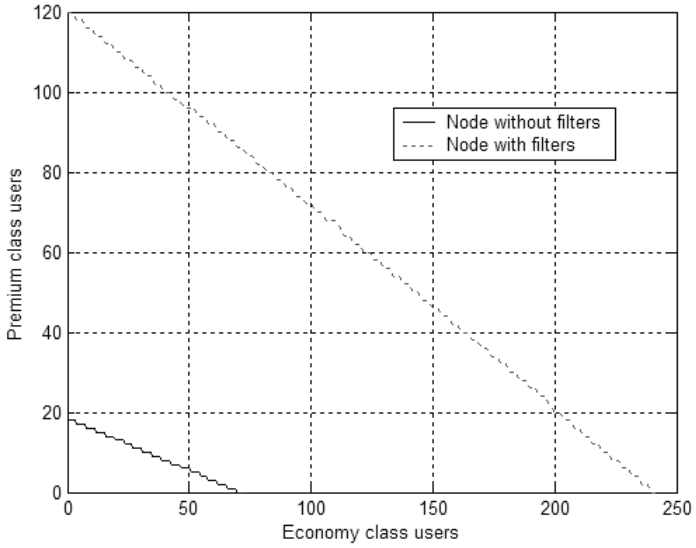
120
100
80
60
40
20
0
Premium class users
0
50
100
150
200
250
Economy class users
Node without filters
Node with filters

Fig. 6

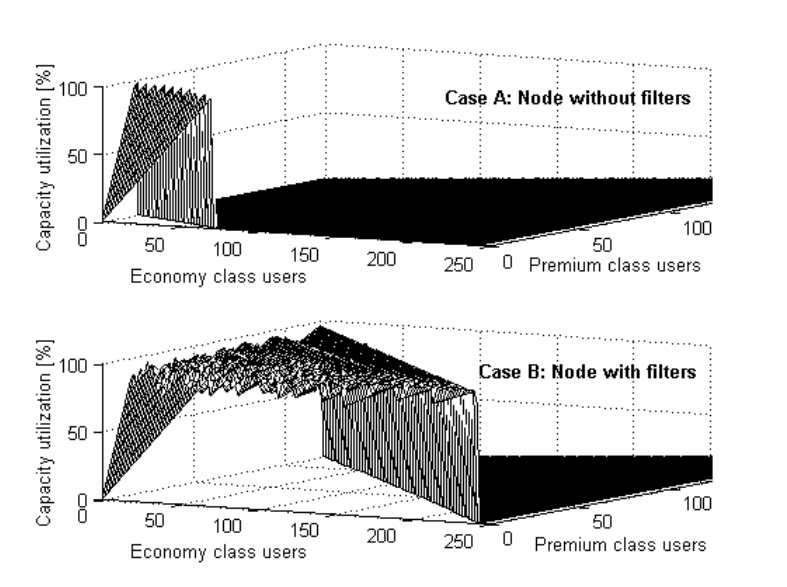
Case A: Node without filters
Capacity utilization [%]
Economy class users
Premium class users
Case B: Node with filters
Capacity utilization [%]
Economy class users
Premium class users

Fig. 7

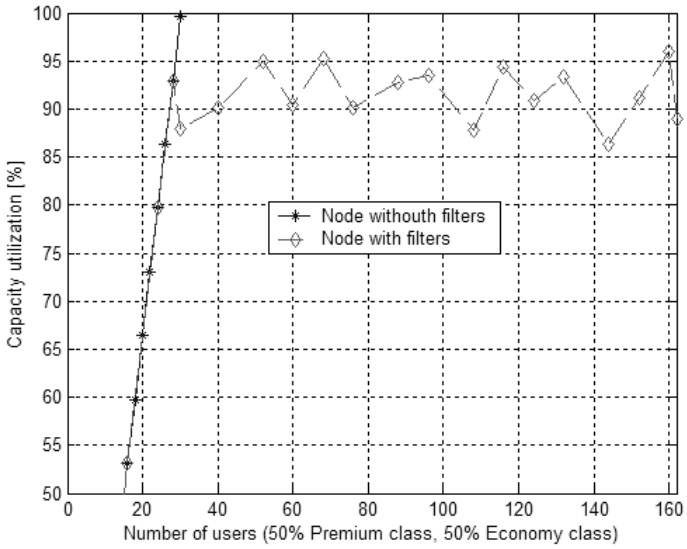
Node withouth filters
Node with filters
Capacity utilization [%]
Number of users (50% Premium class, 50% Economy class)

Fig. 8

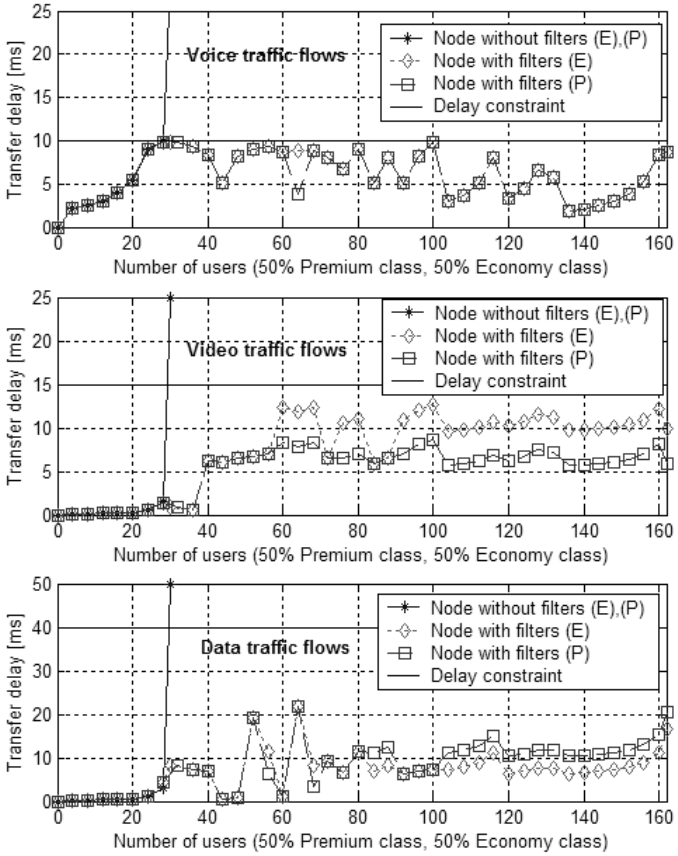
Voice traffic flows
Video traffic flows
Data traffic flows
Node without filters (E),(P)
Node with filters (E)
Node with filters (P)
Delay constraint
Transfer delay [ms]
Number of users (50% Premium class, 50% Economy class)

Fig. 9

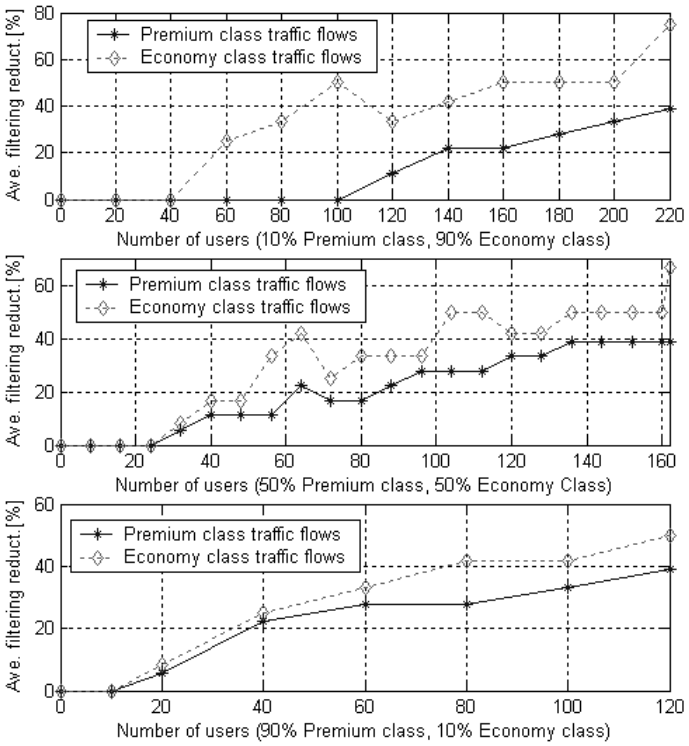
Premium class traffic flows
Economy class traffic flows
Ave. filtering reduct.[%]
Number of users (10% Premium class, 90% Economy class)
Premium class traffic flows
Economy class traffic flows
Ave. filtering reduct.[%]
Number of users (50% Premium class, 50% Economy Class)
Premium class traffic flows
Economy class traffic flows
Ave. filtering reduct.[%]
Number of users (90% Premium class, 10% Economy class)

Fig. 10

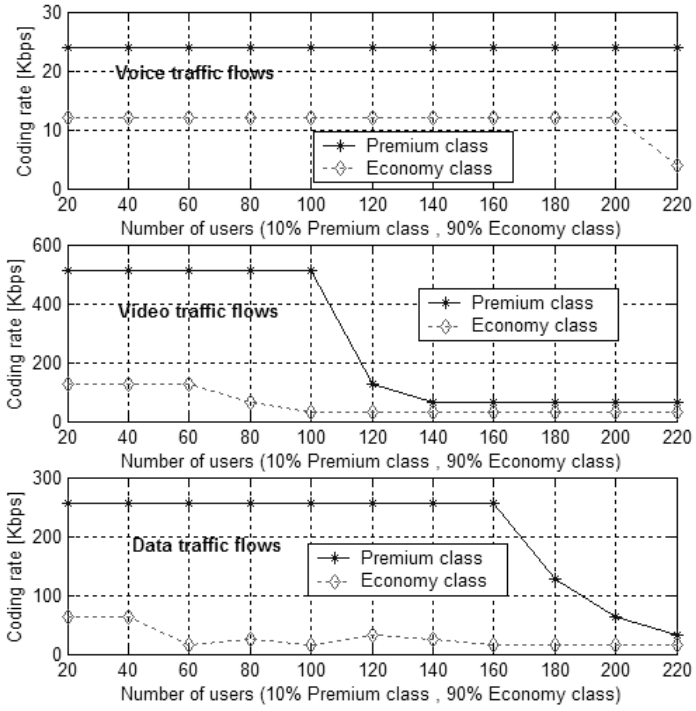
Voice traffic flows
Premium class
Economy class
Coding rate [Kbps]
Number of users (10% Premium class , 90% Economy class)
Video traffic flows
Premium class
Economy class
Coding rate [Kbps]
Number of users (10% Premium class , 90% Economy class)
Data traffic flows
Premium class
Economy class
Coding rate [Kbps]
Number of users (10% Premium class , 90% Economy class)

Fig. 11

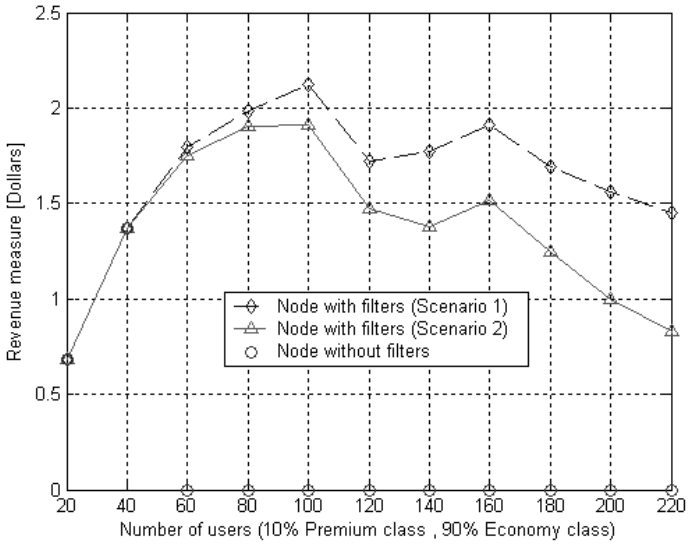
Revenue measure [Dollars]
Number of users (10% Premium class , 90% Economy class)
Node with filters (Scenario 1)
Node with filters (Scenario 2)
Node without filters

Fig. 12

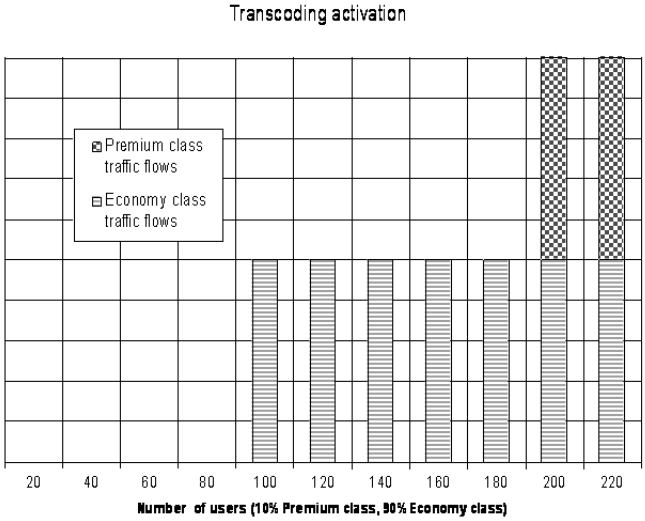
Transcoding activation
Premium class traffic flows
Economy class traffic flows
Number of users (10% Premium class, 90% Economy class)

Fig. 13

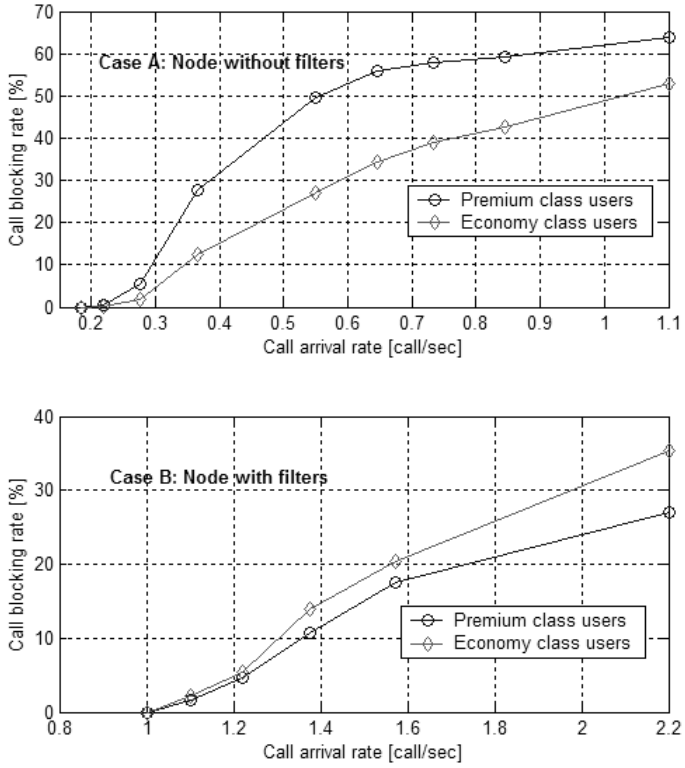
Case A: Node without filters
Call blocking rate [%]
Call arrival rate [call/sec]
Premium class users
Economy class users
Case B: Node with filters
Call blocking rate [%]
Call arrival rate [call/sec]
Premium class users
Economy class users

Fig. 14

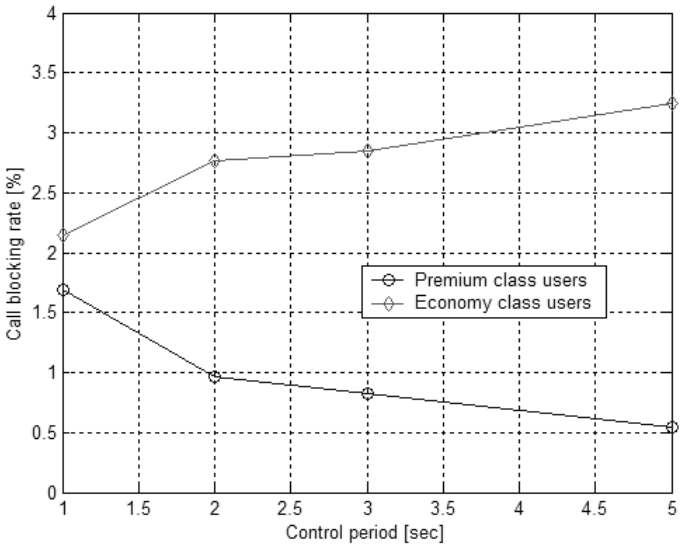
Call blocking rate [%]
Control period [sec]
Premium class users
Economy class users

Fig. 15

TABLES

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

Table I

Filtering levels $f^{[k]}$	Coding rates $R^{[k]}$	Coding formats $CF^{[k]}$	Grades of quality
1	R_1	CF_1	low
2	R_2	CF_1	…
…	…	…	medium
…	…	…	…
M	R_M	CF_2	high

Table II

Speech			Video			Data		
$f^{[k]}$	$R^{[k]}$ (kbps)	$CF^{[k]}$	$f^{[k]}$	$R^{[k]}$ (kbps)	$CF^{[k]}$	$f^{[k]}$	$R^{[k]}$ (kbps)	$CF^{[k]}$
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

Table III

	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 [2]	30 ms [2]	80 ms [28]
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

Table IV