Computer Communications 32 (2009) 1781–1789

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Pricing-based call admission control for wireless MPEG-4 traffic

Aggelos Lazaris, Polychronis Koutsakis *

Department of Electronic and Computer Engineering, Technical University of Crete, Kounoupidiana Campus, 73100 Chania, Crete, Greece

ARTICLE INFO

Article history:
Available online 2 July 2009

Keywords:
Call admission control
MPEG-4 videoconference
Cellular networks
Traffic modeling
Provider revenue

ABSTRACT

Next generation wireless technologies will need to incorporate new sets of traffic control procedures in order to cope with the challenges related to supporting both the existing and the ever-increasing new multimedia services. The work presented in this paper focuses on traffic control for MPEG-4 videoconference over cellular networks, and is twofold. We first introduce a traffic model for medium quality MPEG-4 videoconference traffic. We then proceed to use this model, as well as a previously developed model for high quality MPEG-4 traffic, in the implementation of a new Call Admission Control (CAC) mechanism. Providers need to use efficient resource management mechanisms in order to keep existing clients satisfied and attract new customers, so that they can increase their revenue. Towards this goal, our CAC mechanism makes decisions on the acceptance/rejection of a new video call not only based on the predicted capacity that users will consume, but also on the possible revenue gained for the provider when degrading current users in order to accommodate new ones. Our mechanism is implemented with both a static and a dynamic pricing policy and is shown, via an extensive simulation study, to provide excellent Quality of Service (QoS) to wireless videoconference users.

1. Introduction

Network designers face a challenging problem when trying to control the traffic entered into a wireless cellular network. The reason is the rapidly growing number of multimedia applications, combined with user mobility which causes the traffic conditions in the cells to change very quickly. When mobile users change their point of attachment (handoff), the end-to-end path may be changed while they still expect to receive high QoS.

Emerging wireless networks will need to accommodate significant loads of traffic related to real-time video services, and especially videoconference traffic [1]. The QoS requirements of video users are particularly strict. The reason is that video packet transmission delays and the subsequent video packet dropping when the delay exceeds an upper bound result in the viewer's annoyance. For this reason, recent research efforts are focusing on improving end-to-end QoS provisioning in multimedia delivery [31]. Call admission control is a strategy used to limit the number of call connections into the network in order to reduce network congestion, therefore enabling the system to provide the desired QoS to newly incoming as well as existing calls.

In recent work [2] we have proposed an efficient CAC mechanism for high quality MPEG-4 videoconference traffic. Our mechanism adopted the idea of a *probabilistic* service, as described in [4]. This type of service does not provide for the worst-case scenario, but instead guarantees a bound on the rate of lost/delayed packets based on statistical characterization of the traffic. However, our mechanism did not adopt the standard method of implementation for this service type which is the use of an "equivalent bandwidth" estimation, larger than the average rate but less than the peak rate of the sources. Although widely used, "equivalent bandwidth"-based mechanisms are known to significantly overestimate the sources' actual bandwidth requirements and therefore to provide quite conservative CAC mechanisms, which fail to use efficiently all the available bandwidth [7,9]. Instead, we proposed the use of our modeling approach [16] for traffic originating from multiplexed high quality MPEG-4 videoconference traces, in order to propose a new CAC mechanism for videoconference traffic transmission over wireless cellular networks. Our CAC mechanism in [2] uses the traffic parameters (peak, mean, standard deviation) which the video source either declares at call setup or has agreed on in its contract with the wireless provider, in order to *precompute* or compute online a large number of traffic scenarios for its decision-making.

In the present work, we show that our modeling approach for high quality video traffic can also be applied in order to model videoconference traffic from multiplexed medium quality MPEG-4 videoconference traces. We then proceed to take advantage of the accuracy of our modeling approach both for medium quality and high quality MPEG-4 videoconference traffic, in order to propose a new CAC mechanism for wireless cellular networks. Our new CAC mechanism makes decisions not only based on the

* Corresponding author. Tel.: +30 28210 37235.
E-mail addresses: alazaris@telecom.tuc.gr (A. Lazaris), polk@telecom.tuc.gr (P. Koutsakis).

0140-3664/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2009.06.013

system's ability to accommodate newly arriving users in terms of capacity (as in [2]), but also on the profit that can be made by the provider if existing users are degraded in order for new video calls to be accepted. When evaluating customer profitability, marketers are often reminded of the 80/20 rule (80% of the profits are produced by top 20% of profitable customers and 80% of the costs are produced by top 20% of unprofitable customers) [5]. This is especially true when discussing the case of bursty video users who may be willing to accept high costs but expect to transmit at high rates with excellent QoS, or, respectively, may not be willing to pay more than a minimum fee but are still attempting to send packet bursts over the network. In [32] the authors examine a scenario in which CAC is strongly needed, and will be even more needed in the future; this is the case of organized leisure areas, where visitors can enjoy themselves experiencing different types of entertainment and which are continuously multiplying in large cities. In these areas handoff video calls between WLAN and UMTS need to be serviced, and the strict QoS requirements of video traffic makes the need for an accurate CAC mechanism imperative.

To the best of our knowledge, this is one of the first works in the relevant literature focusing on revenue-based CAC for videoconference traffic over wireless networks (the study in [12] addresses the pricing problem, but for pre-encoded video, not live videoconference and without taking CAC into account). A much earlier version of the algorithm was presented in [17]. The differences between [17] and the present work are:

(a) [17] was implemented on one cell of the network. On the contrary, our present work considers the problem of mobility and proposes a new bandwidth reservation mechanism in adjacent cells which implements the concept that bandwidth reservation in adjacent cells is desirable, but not obligatory for a call to be accepted.
(b) in [17] we used ad hoc revenue weights. On the contrary, in the present work we consider both a static and a dynamic pricing mechanism in order to compute the possible revenue from the acceptance of a new user into the network.

The paper is organized as follows. Section 2 presents our model for multiplexed traffic from medium quality MPEG-4 sources. In Section 3 we present our system model. Section 4 discusses our CAC algorithm and its relation to the relevant literature on revenue-based CAC. Section 5 presents our results, showing the efficiency of the mechanism. Finally, Section 6 contains the conclusions of our study.

2. Medium quality MPEG-4 videoconference traffic model

In this work, we study the medium quality version of three different long sequences of MPEG-4-encoded videos from [11]. The difference between the medium and the high quality encoding of the movies lies in the quantization parameters used in each case, in [11]. The three traces ("Office Cam", "Lecture Room Cam", "Boulevard Bio") are movies with low motion.

We have investigated the possibility of modeling the traces with a number of well-known distributions (Gamma, Lognormal, Log-logistic, Exponential, Geometric, Weibull, Pearson V). Our work follows the steps of the work presented in [27], where Heyman et al. analyzed three videoconference sequences coded with a modified version of the H.261 video coding standard and two other coding schemes, similar to the H.261. The authors in [27] found that the marginal distributions for all the sequences could be described by a Gamma (or equivalently Negative Binomial) distribution and used this result to build a Discrete Autoregressive (DAR) model of order one, which works well when several sources are multiplexed. In [30] the authors confirmed the results of [27], as they studied H.261-encoded realistic videoconference sessions transmitted over an IP-network videoconference, and found that the video frame sizes can be satisfactorily approximated by the PDF of a Gamma distribution; they also found that a method exploiting the convolution property of Gamma densities (instead of the usual moments matching approach) can be more appropriate with respect to queuing. In [28,29] different modeling approaches are proposed for MPEG-1 traffic, based on the Lognormal and Gamma distributions.

Our results (derived with the use of Q–Q plots [10], Kolmogorov–Smirnov (KS) tests [10] and Kullback–Leibler (KL) tests [8]) have shown that, similarly to our work in [16] on modeling high quality MPEG-4 videoconference traffic, the best fit among these distributions for modeling a single movie is achieved for all traces examined with the use of the Pearson type V distribution (also known as the inverted gamma distribution). The data for each trace consists of a sequence of the number of cells per video frame. The length of the videos varies from 45 to 60 min. Table 1 presents the trace statistics for each trace, in the Medium Quality (MQ) columns. Table 1 also contains the trace statistics for the high quality version of the same movies, as both versions will be needed in our work on an efficient CAC mechanism. *The sets of parameters of the traces comprise the "modes" adopted by videoconference users in our study. This will be further explained in Section 4.*

The Probability Density Function (PDF) of a Pearson type V distribution with parameters (α, β) is

$f(x) = [x^{-(\alpha+1)}e^{-\beta/x}]/[\beta^{-\alpha}\Gamma(\alpha)]$, for all $x > 0$, and zero otherwise ($\Gamma(n)$) denotes the gamma function, which is an extension of the factorial to complex and real number arguments. It is related to the factorial by $\Gamma(n) = (n-1)!$, if n is a positive integer; for a complex number z with a positive real part the Gamma function is defined by $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$.

The mean and variance are given by the equations:

$$\text{Mean} = \beta/(\alpha-1), \quad \text{Variance} = \beta^2/[(\alpha-1)^2(\alpha-2)].$$

However, although the Pearson V was shown to be the better fit among all distributions, the degree of goodness-of-fit for the Pearson V varied significantly, and even in the cases of a quite good fit, the fit was not perfectly accurate. This was expected, as the gross differences in the number of bits required to represent I, P and B frames impose a degree of periodicity on MPEG-encoded streams, based on the cyclic GOP formats. Any model which purports to reflect the frame-by-frame correlations of an MPEG-encoded video stream must account for GOP cyclicity, otherwise the model could produce biased estimates of cell loss rate for a network with some given traffic policing mechanism [3]. Hence, we proceeded to study

Table 1
Statistics for the high and medium quality versions of the video traces.

Movie	Mean (Mbps)		Peak (Mbps)		Standard deviation (Mbps)		Revenue weight		
	HQ	MQ	HQ	MQ	HQ	MQ	HQ($\mu + U * \sigma$)	HQ (formula (1)), q = 25%	MQ
Office	0.4	0.11	2	1	0.434	0.253	5.95	4.86	2.23
Lecture	0.21	0.058	1.5	0.69	0.182	0.094	2.92	2.18	1
Boulevard Bio	0.65	0.19	2.6	1.3	0.368	0.197	8.19	6.07	2.79

the frame size distribution for each of the three different video frame types (I, P, B), in the same way we studied the frame size distribution for the whole trace. The Pearson V distribution once again provided the best fitting results for all types of video frames' sequences, and the modeling results were much improved in comparison with those of modeling the trace as a whole. We present, indicatively, the results from our KS-tests for the I, P and B frames of the lecture trace in Figs. 1–3. The KS-test tries to determine if two data sets differ significantly and has the advantage that it makes no assumption about the distribution of data, i.e., it is non-parametric and distribution-free. It uses the maximum vertical deviation between the two curves as its statistic D. The use of KS-tests is a good statistical tool; however it has the drawback that KS-tests give the same weight to the difference between the actual data and the fitted distribution for all values of data, whereas many compared distributions differ primarily in their tails; for this reason we confirmed our modeling results with Q–Q plots and KL-tests. The results of our KS-tests, in Figs. 1–3 show that the Pearson V distribution is the best fit, as it has the smallest maximum vertical deviation from all the distributions. Similar results were deducted by all our statistical tests.

The goal of our work in this section is to build a model which, based on the good but not perfect fit of the Pearson V distribution for modeling a single movie, will accurately capture the behavior of *multiplexed medium quality MPEG-4 videoconference movies* from VBR coders.

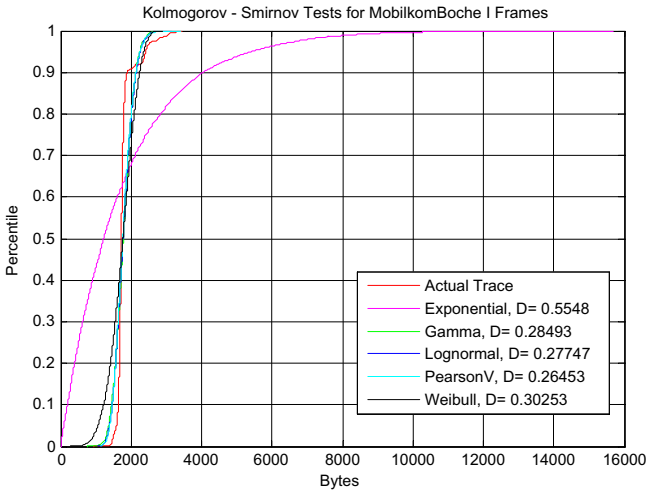

Fig. 1. KS test for the lecture movie I frames.

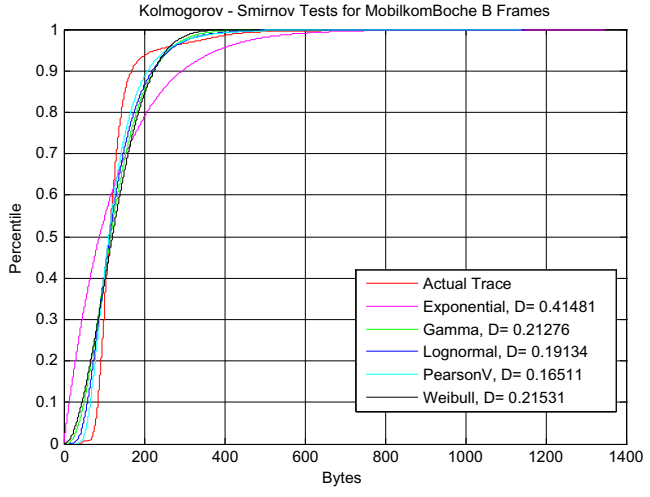

Fig. 2. KS test for the lecture movie B frames.

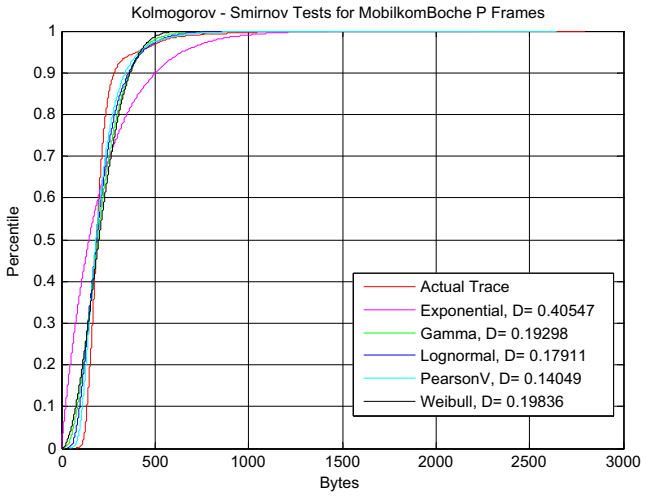

Fig. 3. KS test for the lecture movie P frames.

A discrete autoregressive model of order p, denoted as DAR(p) [14], generates a stationary sequence of discrete random variables with an arbitrary probability distribution and with an autocorrelation structure similar to that of an Autoregressive model. DAR(1) is a special case of a DAR(p) process and it is defined as follows: let $\{V_n\}$ and $\{Y_n\}$ be two sequences of independent random variables. The random variable V_n can take two values, 0 and 1, with probabilities $1-\rho$ and ρ, respectively. The random variable Y_n has a discrete state space S and $P\{Y_n = i\} = \pi(i)$. The sequence of random variables $\{X_n\}$ which is formed according to the linear model: $X_n = V_n X_{n-1} + (1 - V_n)Yn$, is a DAR(1) process.

A DAR(1) process is a Markov chain with discrete state space S and a transition matrix:

$$\mathbf{P} = \rho\mathbf{I} + (1-\rho)\mathbf{Q}, \tag{1}$$

where ρ is the autocorrelation coefficient, $\mathbf{I}$ is the identity matrix and $\mathbf{Q}$ is a matrix with $Q_{ij} = \pi(j)$ for $i, j \in S$.

Autocorrelations are usually plotted for a range W of lags. The autocorrelation can be calculated by the formula:

$$\rho(W) = E[(X_i - \mu)(X_{i+w} - \mu)]/\sigma^2, \tag{2}$$

where μ is the mean and σ^2 the variance of the frame size for a specific video trace.

We build *for each video frame type* a model based only on the above parameters which are either known at call set-up time or can be measured without introducing much complexity in the network. In short, DAR(1) provides an easy and practical method to compute the transition matrix and gives us a model based only on four physically meaningful parameters, i.e., the mean, peak, variance and the lag-1 autocorrelation coefficient ρ of the offered traffic (which is typically very high for videoconference sources). We proceeded with testing our models statistically (with the same methods used for single traces) in order to study whether it produces a good fit for the trace superposition. The accurate fits in our results have shown that the superposition of the actual traces can be modeled well by a respective superposition of data produced by our modeling approach.

In Fig. 4, we have plotted the 0.01-, 0.02-, 0.03-,... quantiles of the actual office camera trace versus the respective quantiles of the DAR(1) model for the superposition of 30 traces' P frames. As shown in the figure, the points of the Q–Q plot fall either very close

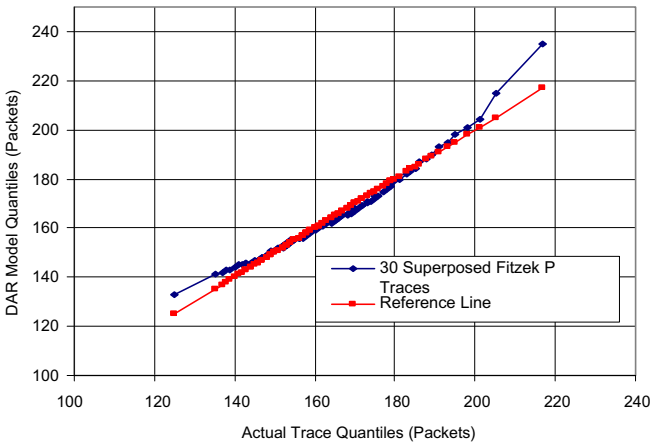

Fig. 4. Q–Q plot of DAR(1) model versus the actual office camera trace for the *P* frames of 30 superposed sources.

or completely along the 45-degree reference line (which would correspond to a perfect match of the actual trace quantiles), with the exception of the first and last 3% quantile (left- and right-hand tail), for which the DAR(1) model greatly overestimates the probability of frames with a very large number of cells. The very good fit (shown from all our results, which are similar in nature to those in [16] and omitted here to avoid repetition) shows that the superposition of the actual traces can be accurately modeled by a respective superposition of data produced by the DAR(1) model.

The accuracy of our modeling approach will be used in our proposal for a revenue-based CAC mechanism for next generation wireless cellular networks. By using the DAR(1) model, the CAC mechanism is able to precompute the bandwidth that will be needed by the superposition of video users and then combine this precomputation with estimates on the provider revenue, as it will be explained in Section 4.

3. System model

Our system model is taken from [18,19]. We use the notion of a cluster similar to the shadow cluster concept [19]. The idea is that every user exerts an influence upon neighboring base stations. As the mobile terminal travels to other cells, the region of influence also moves. The set of cells influenced by a user are said to constitute a cluster. Each user in the network with an active connection has a cluster associated with it. The cells in the cluster are chosen by the cell where the user resides. The shape and the number of cells of a user's cluster depend on factors such as the user's current call holding time, the user's QoS, terminal trajectory and velocity. We consider a wireless network where the time is divided into equal intervals at $t = t_1, t_2, \ldots, t_m$. Let j denote a base station in the network, and x a mobile terminal with an active wireless connection. Let $K(x)$ denote the set of cells that form the cluster for the active mobile terminal x. We denote $P_{x,j,k}(t) = [P_{x,j,k}(t_0), P_{x,j,k}(t_1), \ldots P_{x,j,k}(t_{mx})]$ the probability that mobile terminal x, currently in cell j, will be active in cell k, and therefore under the control of base station k, at times $t_0, t_1, t_2, \ldots, t_{mx}$. $P_{x,j,k}(t)$, known as Dynamic Mobile Probability (DMP), represents the projected probabilities that a mobile terminal will remain active in the future and at a particular location. The parameter m_x represents how far in the future the predicted probabilities are computed; it is not fixed for all users and can depend on the user QoS or the elapsed time since the start of the connection. The DMP can be based on simple or complex information. The latter case is considered in [19] and in the present work, as we compute DMP in the same way as in [19], taking into consideration the mobile terminal direction, velocity and statistical mobility data (we assume for the sake of simplicity that mobile terminals are traveling along a highway, and that each mobile terminal has been observed in the past, and the probabilistic information on the times this mobile terminal typically spends in the cells along the highway has been collected and is available). This is the simplest environment, representing a one-dimensional cellular system. In our simulation study we have made the same assumptions as those in [18]. More specifically:

(a) time is quantized in intervals of $T = 10$ s,
(b) the whole cellular system is composed of 10 linearly arranged cells, laid at 1 km intervals,
(c) during each time interval connection requests are generated in each cell according to a Poisson process and a newly generated mobile terminal can appear anywhere in the cell with equal probability,
(d) mobile terminals can have speeds of: 70, 90, or 105 km/h (with equal probability) and can travel in either of two directions with equal probability,
(e) connection lifetimes are exponentially distributed with mean value equal to 180 s,
(f) m_x is fixed for all users and for the duration of the connection and is equal to 18. This means that the DMPs are computed for 18 steps in the future. Several of the calculated active mobile probabilities $P_{x,j,k}(t)$ will be zero, since the probability that a mobile terminal will be in those cells at a given time may be zero,
(g) the size of the cluster $K(x)$ is fixed for all users and is equal to 5. This means that four cells in the direction that the user is moving towards, along with the cell where the user resides, form the cluster.

Finally, we assume that the effect of fading in the wireless channel, which is highly varying with time and spatial dependencies and interference, can be mitigated by rich-function transmission/reception wireless subsystems [20].

4. Revenue-based call admission control

The use of the accurate DAR(1) MPEG-4 videoconference traffic model from [16], in order to precompute or compute online various traffic scenarios based on the traffic parameters declared by the video sources at call setup, led our mechanism to outperform the "equivalent bandwidth" approach in [2]. These parameters are used for the "identification" of the source as a user adopting a specific "mode". In order to explain what a "mode" is, we first note that a logical assumption for next generation wireless networks is that videoconference users will be allowed to adopt one of a few specific "modes", each corresponding to a set of traffic parameters. Therefore, we used in our work each user's declared set of parameters in order to examine the respective precomputed traffic scenario, based on our MPEG-4 model for a source with such a set of parameters. This approach is especially plausible for wireless videoconference traffic, as the number of variations between source bandwidth requirements is naturally restricted by the type of application (a much larger pool of modes would have to be used in the case of regular video traffic). In [15] we have shown that our mechanism works equally well for H.263 videoconference traffic and excels, again, in comparison to the equivalent bandwidth approach. The accuracy of our video traffic models, for both MPEG-4 and H.263 traffic, helps alleviate the disadvantage that can be found in a number of efficient CAC mechanisms in the literature (e.g., [20,22,25]), which is the frequent use of bandwidth adapta-

tion; as pointed out in [24], frequent bandwidth switching among different levels can consume a significant amount of resources, due to increased network signaling overhead. Regarding [25], which introduced a revenue-based CAC, it should also be pointed out that the linear change which the authors proposed in the number of guard channels (depending on the change in revenue) is not an efficient choice for video traffic, due to its bursty nature.

Similarly to our work in [2] we denote in the present work as "modes" for the MPEG-4 videoconference users the sets of traffic parameters presented in Table 1. Hence, we use six modes for MPEG-4 videoconference traffic, one high and one medium quality mode for each one of the three traces. High-paying users adopt the high quality (HQ) modes, due to the increased bandwidth these modes offer.

The CAC mechanism proposed in [2] focused only on the system's ability to accommodate a newly arriving user in terms of the total channel capacity which is needed for all terminals after the inclusion of the new user. In the case when the channel load, with the admission of a new call, was precomputed (or computed online) to be higher than the channel information rate, users which had agreed in their contracts on the possibility of degradation were gradually degraded up to the point where the new call could be admitted. One parameter not included in our study in [2] was that in a real-life scenario, the decision of admitting or rejecting a new call in the network will be made by the provider not only based on the capacity needed to accommodate the call, but also on the revenue that the admission of the new call will provide. That is, if the admission of a new call (and the subsequent increase in bandwidth utilization) can only be made with the degradation of a higher-paying customer who enjoys higher QoS, the CAC module should compute whether this is a profitable decision.

Hence, in our new CAC mechanism we not only adopt the idea of [2] for precomputation and online computation of various traffic scenarios and implement it for MPEG-4 traffic, but we also assign "revenue weights" to each one of the six MPEG-4 modes, thereby differentiating them into different service classes. Some recent relevant studies (e.g., [21,22,26]) have proposed dynamic pricing strategies in mobile networks, as opposed to static pricing which we propose in this work and has also been proposed in other works in the literature (e.g., [23], which however does not consider video traffic). Dynamic pricing gives negative incentives to users, based on current network conditions (e.g., congestion), in an attempt to shape the aggregate traffic in the network. However, dynamic pricing does not take into account the fact that frequent changes in the price of a call are certain to cause user dissatisfaction, as the user will have to choose constantly whether or not they will make or continue a call after receiving information from the network on how much the call will cost.

We use two approaches to assign the revenue weights, and we will evaluate our CAC mechanism with both. The first approach uses static pricing, while the second uses dynamic pricing. Based on our reasoning above, we believe that the first approach is preferable, i.e., the use of a CAC mechanism combining accurate traffic prediction with static pricing is the better choice. However, we also include in our study the use of the second approach, in order to show the efficiency of our CAC mechanism regardless of the pricing policy. More specifically:

(a) With the first approach (static pricing), revenue weights are assigned based on the traffic parameters of the traces, i.e., based on the mean, standard deviation and peak (or the burstiness, defined as peak/mean ratio). We need to choose one parameter as the "revenue weight unit"; some good parameter choices (the choice depends on the provider's resources, short-term and long-term financial goals) can be:

1. The smallest mean, among all the MQ modes of the traces examined (shown in Table 1). The logic behind this choice is that the "weight unit" corresponds to the mean of the customer with the lowest mean bandwidth requirements. Alternatively, the median among all HQ and MQ mean values shown in Table 1 can also be used (the logic being that the "weight unit" corresponds to the mean of the customer who has average bandwidth requirements, and the provider is mainly interested in customers asking at least for this median bandwidth allocation). Both of these choices, however, do not take the burstiness of the video sources into account and for this reason are not used in our work.
2. The smallest sum $(\mu + F(\sigma))$, where μ is the mean of the video trace, σ is the standard deviation and F is a function. We have chosen in our work to use as a "revenue weight unit" the smallest sum $(\mu + U * \sigma)$, where U is the smallest constant equal to μ/σ among all modes (the values of U vary from 0.435 to 1.766, as the mean is for some modes larger and for some modes smaller than the standard deviation). Therefore, what we are actually controlling in this way is that $U * \sigma$, never becomes larger than μ for any of the modes, so that it does not become the ruling factor (more important than the mean) in the definition of the "revenue weight unit". The smallest value of U is 0.435 for the MQ mode of the "office" trace. This is the value of U used in our simulations, and by using it the smallest sum $(\mu + U * \sigma)$ among all modes is related to the MQ mode of the "lecture" trace. This is used as our "revenue weight unit", and by computing the ratio of each one of the rest of the sums $(\mu + U * \sigma)$ versus the sum for the MQ "lecture" trace, we get the fifth column of Table 1, containing the revenue weights for all modes with the use of this approach.

(b) In the second approach, we use dynamic pricing and we utilize the formula for the demand function from [26], since it is implemented for different priority users, which fits our system's assumptions. From [26], we derive:

$$p_h = p_o + p_o * \sqrt{-4\ln(q)}/2, \quad p_h \geqslant p_o, \tag{3}$$

where p_o is the price for a medium quality user, p_h is the price charged to high quality users and q is the percentage of high quality users who accept dynamic pricing (i.e., they do not accept degradation). The probabilities p_h and p_o have three different sets of values, one for each mode of Table 1. Without loss of generality, we have kept the values of p_o equal to those we found with the use of the previous approach, and we dynamically calculate the values of p_h for each mode, based on the value of q in every time interval of $T = 10$ s (at the start of our simulations, q is set to 25% and the starting revenue weights are shown in Table 1, but the percentage of active HQ users who do not accept degradation varies at any given time).

Users choose one of the six modes with a probability when they enter the system (in Section 5 we will also discuss our results when altering this probability). Users adopting the high quality "Boulevard Bio" MPEG-4 videoconference mode are the ones demanding the highest QoS and paying, respectively, for it, followed by users adopting the high quality "office" mode and the high quality "lecture" mode; users adopting the MQ versions of the traces are the low-paying users.

Our CAC mechanism uses the traffic models presented in Section 2 for medium quality and in [16] for high quality MPEG-4 traffic, in order to *precompute* a number of traffic scenarios. Naturally, not all traffic scenarios can be precomputed, due to the very large number of all possible traffic loads; still, with the use of an adequate number of precomputed scenarios and our accurate video

model, when a non-precomputed traffic load occurs in the system an online simulation can be conducted relatively quickly by our system in order to compute the "deviation" between the bandwidth needed currently and the "closest" (in terms of the synthesis of modes) precomputed traffic scenario. This new traffic scenario will then be added into the CAC mechanism's database of precomputed scenarios.

As already explained, our CAC does not make its decision based only on the maximization of system capacity, but also on the maximization of provider revenue. The current revenue R is computed as

$$R = \sum_i N_i * W_i$$

where N_i is the total number of video users of mode i, and W_i is the revenue from each user of mode i (shown in the last columns of Table 1). Then our proposed CAC algorithm proceeds with the following steps, at the arrival of a new user request in a cell. At the end of the algorithm, we explain what we mean by the phrase "reserve the proper bandwidth in the cluster cells".

Our CAC mechanism is mainly focused on the local cell (the cell in which a call is initiated or to which a call is hand-offed). The logic of the CAC algorithm is that, when a new video user arrives (either from within the picocell or from handoff), the system first checks whether it can be accommodated in terms of the total bandwidth which will be needed when the user is multiplexed with the existing users in the system. If this is not possible, the algorithm attempts to degrade the user, if the user accepts degradation. The rationale behind this decision is that the arrival of a new user should cause the minimum possible number of degradations, and hence irritation, to users who are already in the system, therefore it is preferable that the new user is accepted with degradation. One point which needs to be stressed here is that in most of the relevant works in the literature (including our previous work in [2,15]), it is commonly accepted that handoff calls have absolute priority in obtaining an equal amount of channel bandwidth as the one they were occupying in their previous picocell location, i.e., handoff calls are not expected to endure any quality degradation, as this would lead to user dissatisfaction. We take a different approach in this work. It is indeed crucial for a handoff user not to experience call dropping when moving from one picocell to the next, as this would lead to significant user irritation (call dropping is much more irritating than the blocking of the call of a new user who attempts to transmit). However, if the mobile user experiences, during handoff, a degradation for which he has agreed in his contract, this should not be a cause for user irritation and therefore is allowed in our algorithm. If after degradation (of either a new or a handoff videoconference call) the acceptance of the call is still not possible, the CAC mechanism checks all possibilities of degrading users of the same or lesser priority of the new call in order to accommodate it. If such a possibility exists and the call comes from handoff, it is accepted. If, however, it is a new call originating from within the picocell, it will be accommodated *only if* its acceptance will lead to higher revenue; otherwise, even if the total bandwidth that will be used with the acceptance of the new call is

1. Identify the User Service Class ("Mode")
2. Check precomputed scenarios or compute online the Total Bandwidth (TB) needed for all video calls in the current cell, with the inclusion of the new call
3. ***if*** TB $\leqslant$ Channel Rate ***then***
4. Admit Call, Compute new revenue R^*, Reserve the proper bandwidth in the cluster cells
5. ***else***
6. ***if*** new user accepts degradation ***then***
7. Degrade New User to its next lower mode
8. Check precomputed scenarios or compute online the TB needed for all video calls, with the inclusion of the new degraded call
9. ***if*** TB $\leqslant$ Channel Rate ***then***
10. Admit New Degraded Call, Compute new revenue R^*, Reserve the proper bandwidth in the cluster cells.
11. ***else***
12. Check precomputed scenarios or compute online to find the smallest number Y of users of lesser priority "modes" or equal priority "modes" with the new user, who would need to be degraded for TB $\leqslant$ Channel Rate. Procedure starts with users of lesser priority mode.
13. ***if*** $Y > 0$ ***then***
14. ***if*** new user is handoff ***then***
15. Degrade All Y users
16. Admit New Degraded Call, Reserve the proper bandwidth in the cluster cells
17. ***else***
18. Compute the revenue R^{**}, which can be acquired for this Y and with the acceptance of the new user.
19. ***if*** $R^{**} > R$ ***then***
20. Degrade All Y users
21. Admit New Degraded Call, Reserve the proper bandwidth in the cluster cells.
22. ***else***
23. Reject New Call
24. ***end if***
25. ***end if***
26. ***else*** ($Y = 0$, which means that step 12 cannot be satisfied)
27. Reject New Call
28. ***end if***
29. ***end if***
30. ***else*** (new user does not accept degradation)
31. Repeat steps (12–28)
32. ***end if***
33. ***end if***

larger than the bandwidth previous used, there is no reason to degrade a significant number of users (and cause them even a slight irritation) if the provider will receive no extra revenue. If the new call does not accept any degradation, the attempt is still made to degrade lesser or equal priority users who are already in the system, and a new call from within the picocell is again accepted only if it leads to higher revenue.

As mentioned above, our CAC mechanism focuses mainly on the local cell. The reason is that, as good as the mobility prediction might be, it is impossible to ever predict with perfect accuracy to which cell a user might move, and additionally, video calls can be degraded, therefore even if the necessary bandwidth to accommodate a handoff user may not be available at a given time in its entirety, the additional needed bandwidth can be "squeezed" out of existing calls. Therefore, regarding the bandwidth reservation in the cluster cells, to which we refer in the algorithm, we use the probabilities $P_{x,j,k}(t)$ in order to make partial reservations in neighboring cells. More specifically, we use as an estimate the median among all probabilities that a mobile x might move into cell k at a future time and we reserve, in that cell, bandwidth equal to $P_{med,x,j,k}(t) * B$, where B is the bandwidth currently utilized by the mobile in the cell where it resides. This partial reservation helps alleviate the problem of accommodating a new bursty video user who enters a cell, since part of the needed bandwidth has already been reserved and the rest will most likely be acquired through degradation of users who accept it. At the same time, since it is impossible to be certain whether a mobile will actually move to a next cell, it is impractical to reserve the whole of B in the cells comprising the cluster; this would hinder new calls originating from within cell k to acquire their needed bandwidth since a large portion of the cell's bandwidth would remain unused. The bandwidth reserved for possible handoffs, in our mechanism, is not dedicated to the user for whom it was initially reserved (since this user may never enter the cell, finally) but is entered into a *reserved bandwidth pool*, from which handoff users entering the cell can be serviced (if the user moves to another cluster or terminates his call, the bandwidth reserved for this user is removed from the reserved bandwidth pool of the adjacent cells of the previous cluster). For this reason, even if the $P_{med,x,j,k}(t) * B$ bandwidth is impossible to be reserved in an adjacent cell, the call is still accepted in the present cell, i.e., *bandwidth reservation in adjacent cells is desirable, but not obligatory for a call to be accepted*; its acceptance relies only on the local cell's ability to accommodate the call. The reasons that the call is accepted even if the respective bandwidth for it cannot be reserved in an adjacent cell are: (a) that it is impossible to predict if the user will actually move to that cell, and (b) that the call may still be serviced from the reserved bandwidth pool of that cell.

Regarding the applicability of our mechanism as a more general proposal for wireless cellular networks, it needs to be stressed that it is actually not necessary to adopt the same approach (i.e., of basing the CAC on traffic models) for all types of flows in the wireless network. If an accurate model exists for video traffic, which is the most bursty type of traffic in the network, this would be enough for the provision of a revenue-based CAC mechanism which will be much less conservative than the equivalent bandwidth probabilistic service approach, which was shown in [2,15] to be inferior to the prediction-based approach. The remaining types of flows (e.g., voice and data flows), which are much less greedy in terms of their bandwidth requirements, could be admitted based simply on their declared mean rate, or with any other of the many efficient approaches proposed in the literature.

5. Results and discussion

Fourth generation mobile data transmission rates are planned to be up to 20 Mbps, therefore in this work we study a channel of this rate. The maximum allowed transmission delay for the video packets of a Video Frame (VF) is equal to the time before the arrival of the next VF, with packets being dropped when the deadline is reached (the interframe period in the MPEG-4-encoded movies we studied is 40 ms, since the frame rate is 25 frames/s). The maximum acceptable video packet dropping probability is set to 0.01% [13].

Our mechanism is evaluated over the traffic generated by the real video traces. We studied three scenarios:

(a) The six modes in our study are used with a 16.67% probability (i.e., a user which "wakes up" chooses one of the six modes with equal probability). These probabilities refer to the initial mode with which a user enters the network; this mode may change due to degradation, based on the user's contract. When studying this scenario, we found that the maximum number of users that the system could accommodate, without violating the strict QoS requirement of 0.01% maximum video packet dropping, was 51. As shown in Table 2, this corresponds to less than 75% utilization of the channel capacity. The burstiness of video traffic is responsible for the system's inability to accommodate more sources without violating their QoS requirements.

(b) In this scenario, all MQ modes are selected with 23.33% probability, while all HQ modes are selected with 10% probability. The rationale behind this choice of scenarios was that low-paying users will be the majority among all customers of a wireless provider. The maximum number of users that the system could accommodate, depending on the traffic mix, was found to be 63 in this scenario.

(c) In this scenario, the MQ mode with the lowest bandwidth requirements ("lecture" MQ mode) is selected with 30% probability, the other two MQ modes are selected with 20% probability, the "lecture" HQ mode is selected with 15% probability, the "office" HQ mode with 10% probability and the "Boulevard Bio" HQ mode with 5% probability. Again, the rationale here is that the highest quality modes will be selected by the minority of customers. The maximum number of users that the system could accommodate was found to be 72 in this scenario.

Each simulation point is the result of an average of 10 independent runs, each simulating three hours of network operation. Table 2 presents in its first column the bandwidth which is actually needed by the video traces, in its second column the estimated bandwidth that the traces will need based on our DAR(1) modeling approach, and in its third and fourth columns the bandwidth that is

Table 2
Estimations of the required bandwidth and bandwidth utilization with the CAC mechanism.

Real traces-bandwidth (Mbps) for the 3 scenarios			DAR model-bandwidth (Mbps) for the 3 scenarios			Actual bandwidth used with the CAC scheme (Mbps) for the 3 scenarios-static pricing			Actual bandwidth used with the CAC scheme (Mbps) for the 3 scenarios-dynamic pricing		
S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
14.56	15.35	15.62	15.38	16.07	16.49	14.01	14.75	15.28	13.47	14.43	14.99

utilized with the use of our CAC mechanism, both in the case of static and of dynamic pricing. Two significant conclusions can be drawn from the Table. The first is that the estimation provided by our DAR(1) video traffic modeling approach yields an overestimation of the actual bandwidth requirements of the superposed sources (the reasons for this have to do with an overestimation of the *I* frames size and are explained in our modeling work in [16]); still, this overestimation is small and ranges in all simulated scenarios from a minimum of 4.69% to a maximum of 5.63%. The average overestimation provided by our mechanism is 5.3% over all the studied scenarios, which is acceptable, especially given that a small overestimation of the actual bandwidth requirements of video traffic is usually preferable, in order for the system to cope with the bursty nature of video users. The second conclusion has to do with the efficiency of our CAC mechanism. By comparing the four respective columns in Table 2 it is clear that the actual bandwidth utilized when our revenue-based CAC mechanism is enforced (regardless of static or dynamic pricing) is smaller than the bandwidth needed by the real traces, and hence also smaller than the estimated bandwidth with the use of the DAR(1) model. The difference between the bandwidth utilized with our CAC and the actual bandwidth needed by the traces, ranges for static pricing in all simulated scenarios from a minimum of 2.34% to a maximum of 4.12%, with an average of 3.41 %, and for dynamic pricing ranges from a minimum of 4.03% to a maximum of 7.48%, with an average of 5.84%. The reason for these differences between the CAC scheme and the bandwidth needed by the traces is that some of the MQ videoconference calls are rejected from the system in order to achieve higher revenue; hence, with the use of the DAR(1) modeling approach we reserve slightly more bandwidth than actually needed, and then, with the use of the CAC mechanism we hinder a number of MQ users from accessing the system in order to keep HQ users continuously content with the service they are receiving (i.e., they seldom need to be degraded).

The reason for the larger gap between the dynamic pricing CAC and the bandwidth needed by the traces (in comparison to the static pricing CAC) is that for all values of q >5%, the revenue weights from dynamic pricing for HQ users are lower than those for static pricing, based on the respective formulae used for each policy. Our CAC algorithm defines that in order to accept a new user there needs to be a degradation of a number of HQ users (denoted as Y) *and* the combination of the degradation with the acceptance of the new user needs to be more profitable for the provider than rejecting the new user and letting the Y HQ users keep their high quality mode. The degradation, in the case of the static pricing, causes a larger revenue loss. The reason is that in static pricing HQ users bring in higher revenue; therefore, often the degradation is not profitable, and the new user is not accepted, whereas he may be accepted in the case of dynamic pricing, where the revenue from HQ users is lower. By keeping more HQ users in the system in the case of static pricing, we get more bandwidth utilization than with dynamic pricing. This is also confirmed in our results in Fig. 5, which follows. Therefore, not only is the static pricing CAC preferable conceptually as we discussed in Section 4, but it also excels in bandwidth utilization, when combined with accurate traffic modeling. Of course, it could be argued that, by using a different set of traces and hence a different set of modes, and by choosing different formulae for the static and dynamic pricing, the system could be set up in such a way that the revenue weights from dynamic pricing are higher than those of static pricing. However, even if that is the case, as we explained earlier static pricing guarantees the absence of user irritation (which can be caused by dynamic pricing and users needing to decide on whether they will continue their call with current or degraded QoS); for this reason, static pricing is not only preferable in our view, but a provider using it is also justified to set higher prices, since users know a priori what quality of service they will be receiving and how much it will be costing them.

The high precision in our mechanism's predictions can show even better results with the use of a slightly larger pool of videoconference modes from which traffic scenarios will be precomputed. The use of a slightly larger number of modes will guarantee the existence of a larger variety of parameter sets, so that an incoming call's traffic parameters will always be well-matched with those of one of the modes in the pool.

We further investigate our mechanism's performance in the results presented in Fig. 5, where we present both the estimation provided by the DAR model and the actual bandwidth usage from our revenue-based CAC mechanism (for both the static and the dynamic pricing cases); we also indicatively present the estimation provided by the use of the equivalent bandwidth approach from [6]. All the above are presented versus the normalized real system utilization; this indicates the actual traffic load generated by the traces, normalized to the channel capacity, e.g., a traffic load equal to 40% represents 40% of the 20 Mbps uplink capacity, i.e., 8 Mbps system throughput (these loads have been created for each of the three scenarios under study, and the results presented are the average over all conducted simulations). As shown in the figure, the equivalent bandwidth estimation significantly overestimates the actual traffic load in all cases (this confirms our results in [2,15]). Additionally, the points of the plot of the DAR model estimation fall very close to the 45-degree reference line, showing that our estimation is in all cases very accurate. Most importantly, the figure shows once again that the use of our CAC mechanism leads to a slight underallocation to the videoconference users, in comparison to their offered load. The reason is that some of the MQ videoconference calls are rejected from the system in order to achieve higher revenue for the provider. As the offered load increases, this underallocation increases as well, in order not to allow low-paying users to fill the channel capacity at the cost of degrading high-paying ones. As also shown and explained in the discussion of the results of Table 2, the use of the CAC mechanism with static pricing is shown in the figure to achieve higher bandwidth utilization in comparison to the CAC with dynamic pricing. Finally, it should be noted that for a normalized offered load larger than 78.2%, the QoS requirements of video users are violated (video packet dropping exceeds 0.01%), for both the static and the dynamic pricing implementations of our CAC, due to the burstiness

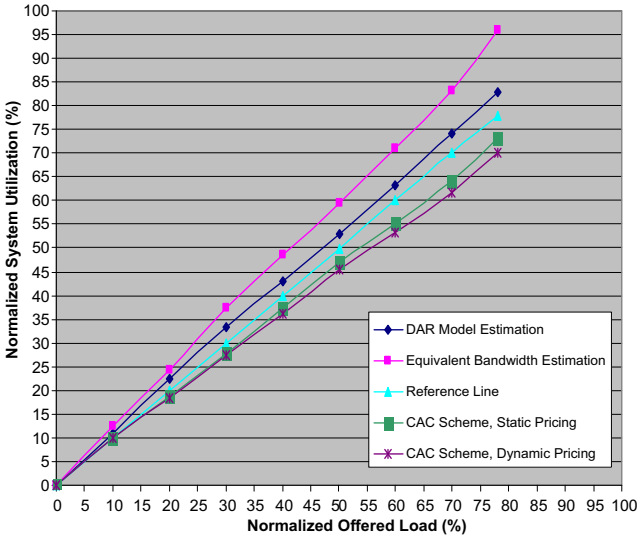

Fig. 5. Capacity utilization with the CAC mechanism.

of video traffic. This is the reason that no results are included in the figure for higher normalized loads.

6. Conclusions

In this work, we have proposed a new, highly efficient Call Admission Control mechanism for MPEG-4 videoconference traffic transmission over next generation cellular networks. Our mechanism, which is evaluated over a cellular network with 5 cells per cluster, uses precomputed traffic scenarios for decision-making on the acceptance or rejection of a new MPEG-4 videoconference call. The precomputation (or online simulation) is based on the traffic parameters declared by the video source at call setup (both high and medium quality sources are used in our study); these parameters are used for the "identification" of the source as a user adopting a specific mode from the pool of modes which have provided the basis for the precomputation of our traffic scenarios. Then, this precomputation is used in combination with an algorithm which maximizes the provider revenue, in order to make decisions on the acceptance or rejection of a new videoconference call (both for static and dynamic pricing policies). Our results, derived from an extensive simulation study, show that our mechanism is able to provide excellent QoS to wireless videoconference users and outperforms, in terms of bandwidth utilization, a CAC mechanism based on equivalent bandwidth estimation.

References

[1] M. Etoh, T. Yoshimura, Advances in wireless video delivery, in: Proceedings of the IEEE, vol. 93, No. 1, 2005, pp. 111–122.
[2] P. Koutsakis, Efficient call admission control for MPEG-4 wireless videoconference traffic, in: Proceedings of the IEEE ICCCN, Honolulu, USA, 2007.
[3] M. Frey, S. Ngyuyen-Quang, A gamma-based framework for modeling variable-rate video sources: the GOP GBAR Model, IEEE/ACM Transactions on Networking 8 (6) (2000) 710–719.
[4] H. Zhang, E.W. Knightly, Providing end-to-end statistical performance guarantee with bounding interval dependent stochastic models, in: Proceedings of the ACM SIGMETRICS, Nashville, USA, 1994.
[5] S.-Y. Kim, T.-S. Jung, E.-H. Suh, H.-S. Hwang, Customer segmentation and strategy development based on customer lifetime value: a case study, Expert Systems with Applications 31 (1) (2006) 101–107.
[6] W. Verbiest, L. Pinoo, B. Voeten, The impact of the ATM concept on video coding, IEEE Journal on Selected Areas in Communications SAC-6 (1988) 1623–1632.
[7] R. Guerin, H. Ahmadi, M. Naghsineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, IEEE Journal on Selected Areas in Communications 9 (7) (1991) 968–981.
[8] K.P. Burnham, D.R. Anderson, Model Selection and Multi-Model Inference, Springer Verlag, New York, 2002.
[9] J.Q.-J. Chak, W. Zhuang, Capacity analysis for connection admission control in indoor multimedia CDMA wireless communications, Wireless Personal Communications 12 (2000) 269–282.
[10] A.M. Law, W.D. Kelton, Simulation Modeling & Analysis, second ed., McGraw Hill Inc., 1991.
[11] F.H.P. Fitzek, M. Reisslein, MPEG-4 and H.263 video traces for network performance evaluation, IEEE Network 15 (6) (2001) 40–54.
[12] P. Seeling, M. Reisslein, Video pricing for wireless networks, in: Proceedings of the IEEE CCNC, Las Vegas, USA, 2006.
[13] D.A. Dyson, Z.J. Haas, A dynamic packet reservation multiple access mechanism for wireless ATM, Mobile Networks and Applications (MONET) Journal 4 (2) (1999) 87–99.
[14] P.A. Jacobs, P.A.W. Lewis, Time series generated by mixtures, Journal of Time Series Analysis 4 (1) (1983) 19–36.
[15] S. Chatziperis, P. Koutsakis, M. Paterakis, A new call admission control mechanism for multimedia traffic over next generation wireless cellular networks, IEEE Transactions on Mobile Computing 7 (1) (2008) 95–112.
[16] A. Lazaris, P. Koutsakis, M. Paterakis, A new model for video traffic originating from multiplexed MPEG-4 videoconference streams, Performance Evaluation 65 (1) (2008) 51–70.
[17] A. Lazaris, P. Koutsakis, Revenue-based call admission control for MPEG-4 wireless videoconference traffic, in: Proceedings of the 9th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Newport Beach, USA, 2008.
[18] Y. Iraqi, R. Boutaba, A novel distributed call admission control for wireless mobile multimedia networks, in: Proceedings of the ACM WOWMOM, Boston, USA, 2000.
[19] D.A. Levine, I.F. Akyildiz, M. Naghsineh, A resource estimation and call admission control algorithm for wireless multimedia networks using the shadow cluster concept, IEEE/ACM Transactions on Networking 5 (2) (1997) 87–99.
[20] T. Kwon, Y. Choi, C. Bisdikian, M. Naghshineh, QoS provisioning in wireless/mobile multimedia networks using an adaptive framework, Wireless Networks 9 (2003) 51–59.
[21] J. Hou, J. Yang, S. Papavassiliou, Integration of pricing with call admission control to meet QoS requirements in cellular networks, IEEE Transactions on Parallel and Distributed Systems 13 (9) (2002) 898–910.
[22] S. Mandal, D. Saha, A. Mahanti, A technique to support dynamic pricing strategy for differentiated cellular mobile services, in: Proceedings of the IEEE GLOBECOM, St. Louis, USA, 2005.
[23] M. Mandjes, Pricing strategies under heterogeneous service requirements, in: Proceedings of the IEEE INFOCOM, San Francisco, USA, 2003.
[24] C. Chou, K.G. Shin, Analysis of combined adaptive bandwidth allocation and admission control in wireless networks, in: Proceedings of the IEEE INFOCOM, New York, USA, 2002.
[25] S. Nelakuditi, R.R. Harinath, S. Rayadurgam, Z.-L. Zhang, Revenue-based call admission control for wireless cellular networks, in: Proceedings of the IEEE International Conference on Personal Wireless Communications, Jaipur, India, 1999.
[26] S.Yaipairoj, F.Harmantzis, Dynamic pricing with alternatives for mobile networks, in: Proceedings of the IEEE Wireless Communications Networking Conference (WCNC), Atlanta, USA, 2004.
[27] D.P. Heyman, A. Tabatabai, T.V. Lakshman, Statistical analysis and simulation study of video teleconference traffic in ATM networks, IEEE Transactions on Circuits and Systems for Video Technology 2 (1) (1992) 49–59.
[28] M. Krunz, S.K. Tripathi, On the characterization of VBR MPEG streams, in: Proceedings of ACM SIGMETRICS, vol. 25, June 1997.
[29] U.K. Sarkar, S. Ramakrishnan, D. Sarkar, Modeling full-length video using Markov-modulated gamma-based framework, IEEE/ACM Transactions on Networking 11 (4) (2003) 638–649.
[30] C. Skianis, K. Kontovasilis, A. Drigas, M. Moatsos, Measurement and statistical analysis of asymmetric multipoint videoconference traffic in IP networks, Telecommunication Systems 23 (1–2) (2003) 95–122.
[31] T. Pliakas, G. Kormentzas, C. Skianis, Scalable video streaming traffic delivery in IP/UMTS networking environments, Journal of Multimedia 2 (2) (2007) 37–46.
[32] P. Makris, C. Skianis, Multi-scenario based call admission control for coexisting heterogeneous wireless technologies, in: Proceedings of the IEEE GLOBECOM, New Orleans, USA, 2008.