A New Call Admission Control Mechanism for Multimedia Traffic over Next-Generation Wireless Cellular Networks

Stylianos Chatziperis, Polychronis Koutsakis, Member, IEEE, and Michael Paterakis, Senior Member, IEEE

Abstract—The subject of Call Admission Control (CAC) for wireless networks has been studied extensively in the literature. Another subject on which many researchers have focused their attention is that of video traffic modeling. However, user mobility, combined with the rapidly growing number of “greedy” multimedia applications, in terms of bandwidth and Quality of Service (QoS) requirements, form a challenging and yet unresolved problem for third and fourth-generation wireless networks. In recent work, we have built a Discrete Autoregressive (DAR (1)) model to capture the behavior of multiplexed H.263 videoconference movies from variable bit rate (VBR) coders. Based on this model, we propose in this work a new efficient CAC scheme for wireless cellular networks, which differs from the existing proposals in the literature in that it uses precomputed traffic scenarios combined with online simulation for its decision making. Our scheme is shown, via an extensive simulation study comparison and a conceptual comparison with well-known existing approaches, to clearly excel in terms of QoS provisioning to users receiving videoconference and Web traffic. To the best of our knowledge, this is the first work in the relevant literature where such an approach has been proposed.

Index Terms—Call Admission Control, downlink wireless channel, traffic modeling, videoconferencing, H.263 video encoding, Web traffic.

1 INTRODUCTION

With the pervasive presence of mobile personal wireless computing devices, wireless communication technologies are rapidly evolving and influencing our daily way of life. The new wireless data and multimedia services, which are constantly added and supported on wireless networks, pose formidable challenges to these networks in their effort to satisfy strict quality-of-service (QoS) requirements. These challenges are further exacerbated by user mobility. To cope with the challenges related to supporting both the existing and the ever-increasing new services, next-generation wireless technologies will need to incorporate new sets of traffic control procedures.

Network congestion is difficult to resolve when real-time traffic, sensitive to both latency and packet loss, is present, without jeopardizing the QoS expected by the users of that traffic. Call Admission Control (CAC) 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 and existing calls. On the other hand, in order to provide enough bandwidth to accommodate broadband services to multiple mobile users, the size of the wireless cells is decreasing toward the picocell architecture [1]. In this environment, due to user mobility as mentioned above, the traffic conditions in the cells can change very quickly. Also, when mobile users change their point of attachment (handoff), the end-to-end path may be changed, whereas they still expect to receive the same QoS. An efficient CAC mechanism should be able to cope with this strict user requirement.

A substantial portion of the traffic carried by emerging wireless networks will be traffic from video services, especially videoconference traffic [52], [53], [54], [55]. Video packet delay requirements are strict because delays are annoying to a viewer: Whenever the delay experienced by a video packet exceeds the corresponding maximum delay, the packet is dropped and the video packet dropping requirements are equally strict. In [2], the authors state that there are three areas where single video source models are useful: studying what types of traffic descriptors are needed for parameter negotiation with the network at call setup, testing the rate control algorithms, and predicting the QoS degradation caused by congestion on an access link. In this work, we use a model for single and multiplexed videoconference H.263 sources recently developed by our group [3], [51] in order to propose and investigate the performance of a new CAC scheme for wireless networks, which differs from the existing proposals in the literature in that it uses precomputed traffic scenarios combined with online simulation for its decision making. This will be further explained later.

CAC schemes for wired and wireless networks are abundant in the literature, and they have been classified in two ways. We will briefly list the categories of each
classification in order to place our work within the relevant literature.

1.1 First Classification of CAC Schemes

The first type of classification regards the type of service offered by the network to the user and classifies services into three broad categories [4]:

1. The traditional service model is defined in [5] and [6] as guaranteed service. Admission control algorithms for the guaranteed service use the a priori characterizations of sources to calculate the worst case behavior of all the existing flows in addition to the incoming one.

2. The probabilistic service, described in [8], 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. The standard method of implementation for this service type is that each source is allotted an "equivalent bandwidth" (also called "effective bandwidth"), which is larger than its average rate but less than its peak rate. In most cases, the equivalent bandwidth is computed based on a statistical model [9] or a fluid flow approximation of traffic [10], [11].

3. In the case of applications that are tolerant of occasional delay bound violations, a third service type was proposed in [4], that is, the predictive service. The measurement-based admission control approach investigated in [4] uses the a priori source characterizations only for incoming flows and for flows very recently admitted. Flows already admitted for some time are characterized by measurements on their transmitted traffic. This approach can never provide the completely reliable delay bounds needed for a guaranteed or probabilistic service; therefore, measurement-based approaches can only be used in the context of relaxed service commitments of the network to the user.

The work presented in this paper generally falls within the second category, that is, of the probabilistic service, but does not adopt the "equivalent bandwidth approach." Instead, we propose the use of our video traffic model in order to be able to precompute a large number of traffic scenarios based on the traffic parameters (peak, mean, and standard deviation) which the video source will declare at admission. Flows already admitted for some time are characterized by measurements on their transmitted traffic. This approach can never provide the completely reliable delay bounds needed for a guaranteed or probabilistic service; therefore, measurement-based approaches can only be used in the context of relaxed service commitments of the network to the user.

1.2 Second Classification of CAC Schemes

The second type of CAC schemes' classification refers specifically to wireless networks and classifies wireless CAC schemes into two broad categories based on their handoff-priority policy [17]:

1. Guard channel (GC) schemes. In this type of scheme, some channels are reserved for handoff calls. Four different types of CAC schemes have appeared in the literature:
   - Cutoff priority scheme. A portion of the channel bandwidth is reserved for handoff calls, and when a channel is released, it is returned to the common pool of channels [18].
   - Fractional GC scheme. A new call is admitted with a certain probability, which depends on the number of busy channels [19].
   - Rigid division-based scheme [20]. All channels allocated to a cell are divided into two groups: one for the common use for all calls and the other solely for handoff calls.
   - New call bounding scheme [42]. A threshold is enforced on the number of new calls accepted into the cell.

2. Queuing priority (QP) schemes. In this type of scheme, calls are accepted whenever there are free channels. Depending on the approach, new calls are blocked and handoff calls are queued [21], vice versa [22], or all calls are queued and the queue is rearranged based on certain priorities [23].

The work presented in this paper falls conceptually within the second category, that is, of the QP schemes, as in our CAC scheme, both handoff calls and the new calls originating from within the picocell are accepted if enough free channel bandwidth exists to accommodate them, and no portion of the bandwidth is restricted for access of either type of call (therefore, our scheme is conceptually similar to [23]).

The rest of this work is organized as follows: Section 2 presents our system model. Section 3 discusses our videoconference traffic model and its accuracy, and Section 4 presents the Web traffic model that we use in our study (from [35] and [36]). Section 5 presents the adopted error model for the wireless channel, and Section 6 contains the description of our CAC scheme, as well as its conceptual comparison with quite a few CAC schemes of the literature. Section 7 includes the evaluation of the results comparison of our scheme with two well-known CAC approaches in the literature: in the case when video traffic is the only traffic present in the system and in the case when video traffic is integrated with Web traffic. Finally, Section 8 presents the conclusions of our study.

2 System Model

Within the picocell, spatially dispersed source terminals share a radio channel that connects them to a fixed base station (BS). The BS allocates channel resources, delivers feedback information, and serves as an interface to the mobile switching center (MSC). The MSC provides access to the fixed network infrastructure. Contrary to our study
in [16], which focused on the uplink channel, the study in this paper focuses on the downlink (BS to wireless terminals) channel, as we are interested in studying system behavior under high rates of Web traffic downloads. The downlink channel time is divided into time frames of equal length. Each frame has a duration of 12 ms (see [16] and [34]) and accommodates 256 information slots. The channel rate is 9.045 megabits per second (Mbps; from [34]). Each information slot accommodates exactly one fixed-length packet of asynchronous transfer mode (ATM) size that contains information (video and data information in our scheme) and a header. In the Multiple Access Control (MAC) schemes introduced in [40] and [41], it has been shown that the use in the uplink (and, respectively, in the downlink for acknowledgments) of a small portion of the channel bandwidth (less than 3 percent) for requests by terminals that wish to acquire slots to transmit is usually sufficient for high system performance. For this reason, we assume here that six of the 256 slots of the channel frame are used by the BS for transmitting acknowledgments and synchronization information to the source terminals; therefore, the available downlink channel bandwidth for information transmission is (250/256) + 9.045 Mbps = 8.833 Mbps.

We use computer simulations to study the performance of our mechanism. The simulations were conducted with the use of the C programming language in Unix SUN workstations. Each simulation point is the result of an average of 10 independent runs (Monte-Carlo simulation).

### 3 Videoconference Traffic Model

In this section, we present in brief our model for single and multiplexed videoconference H.263 sources, which was recently developed by our group [3], [51]. The model will be used in our proposal of the new CAC scheme.

#### 3.1 H.263 Streams

H.263 is a video standard that can be used for compressing the moving picture component of audiovisual services at low bit rates. It adopts the idea of PB frames, that is, two pictures being coded as a unit. Thus, a PB frame consists of one P-picture, which is predicted from the previous decoded P-picture, and one B-picture, which is predicted from both the previous decoded P-picture and the P-picture currently being decoded. The name B-picture was chosen because parts of B-pictures may be bidirectionally predicted from the past and future pictures. With this coding option, the picture rate can be increased considerably without increasing the bit rate much [24].

#### 3.2 The Distribution Fit for a Single Source

Our work in [3] and [51] regarding the development of a model for H.263 videoconference sources followed the steps of the well-known work by Heyman et al. [25], [26] for H.261 videoconference. In [25] and [26], Heyman et al. have shown that H.261 videoconference sequences generated by different hardware coders, using different coding algorithms, have gamma marginal distributions (this result has been extensively used in the relevant literature): They use this fact to build a Discrete Autoregressive (DAR) model of order 1, which works well when several sources are multiplexed.

In [3] and [51], we have used five different long sequences of H.263 encoded videos (from [27] and [28]) with low or moderate motion in order to derive a statistical model that fits well the real data. The length of the videos varies from 45 to 60 minutes, and the data for each trace consists of a sequence of the number of cells per video frame (VF; as already explained earlier, we use packets of ATM cell size throughout this work). We have investigated the possibility of modeling the five videos with quite a few well-known distributions: Our results have shown that the use of the gamma distribution is not a good choice, as the best fit among these distributions is achieved for all the studied traces with the use of the Pearson type-V distribution (it is not a perfect fit, but a very good one).

The five traces used were, respectively,

1. a video stream extracted and analyzed from a camera showing the events happening within an office ("Office Cam"),
2. a video stream extracted and analyzed from a camera showing a lecture ("Lecture Cam"),
3. a video stream extracted and analyzed from a parking security camera ("Parking Cam"),
4. a video stream extracted and analyzed from a talk show ("N3 Talk"),
5. a video stream extracted and analyzed from another talk show ("ARD Talk").

For each one of these movies, we have used the VBR coding version (in Quarter Common Intermediate Format (QCIF) resolution, which is a videoconferencing format providing a standard size for images produced by low-resolution digital cameras and video cameras; QCIF images are 176 pixels wide and 144 pixels tall). The mean, peak, and standard deviation of the VF sizes for each movie are given in Table 1, along with the respective parameters of the Pearson type-V distribution, which is shown from our results to provide the best fit for the traces among all the

<table>
<thead>
<tr>
<th>Movie</th>
<th>Duration (minutes)</th>
<th>Mean (bytes)</th>
<th>Peak (bytes)</th>
<th>Standard Deviation (bytes)</th>
<th>Pearson type V parameters (α, β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office</td>
<td>45</td>
<td>903</td>
<td>5191</td>
<td>327</td>
<td>(9.623, 7793.128)</td>
</tr>
<tr>
<td>Lecture</td>
<td>60</td>
<td>618</td>
<td>5760</td>
<td>370</td>
<td>(4.787, 2338.862)</td>
</tr>
<tr>
<td>Parking</td>
<td>60</td>
<td>2681</td>
<td>19680</td>
<td>502</td>
<td>(30.568, 79288.216)</td>
</tr>
<tr>
<td>N3 Talk</td>
<td>60</td>
<td>2545</td>
<td>13956</td>
<td>1454</td>
<td>(5.066, 10350.12)</td>
</tr>
<tr>
<td>ARD Talk</td>
<td>45</td>
<td>2374</td>
<td>13275</td>
<td>1296</td>
<td>(5.352, 10331.312)</td>
</tr>
</tbody>
</table>

### Table 1

Trace Statistics
The Probability Density Function (PDF) of a Pearson type-V distribution with parameters \((\alpha, \beta)\) is 
\[
f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}
\]
for all \(x > 0\); otherwise, it is zero. The mean and variance are given by
\[
\text{Mean} = \frac{\beta}{(\alpha - 1)}, \quad \text{Variance} = \frac{\beta^2}{((\alpha - 1)^2(\alpha - 2))}.
\]

The sets of parameters of these five traces comprise the “modes” adopted by videoconference users in our study (for example, the traffic generated by a videoconference user in “Office” mode has mean, peak, and standard deviation equal to the ones of the “Office Cam” trace).

To test statistically which distribution provides a good fit for the above traces, we have used Q-Q plots. The Q-Q plot is a powerful goodness-of-fit test [25], [29] which graphically compares two data sets in order to determine whether the data sets come from populations with a common distribution (if they do, the points of the plot should fall approximately along a 45-degree reference line). More specifically, a Q-Q plot is a plot of the quantiles of the data versus the quantiles of the fitted distribution (a \(z\)-quantile of \(X\) is any value \(x\) such that \(Pr(X \leq x) = z\)).

Since the focus of this work is on the new proposal of the CAC, we will present only indicative results of our videoconference model.

In Figs. 1 and 2, we have plotted the 0.03-, 0.06-, 0.09-, \ldots quantiles of the actual trace (\(x\)-axis) versus the respective quantiles of the various distribution fits (\(y\)-axis) for two of the five traces under study. The common characteristic of both figures is that the Pearson V distribution fit is the best in comparison to the gamma, log-normal, and exponential distributions, which are presented here (comparisons were also made with the negative binomial, log-logistic, and Pareto distributions, which were also worse fits than the Pearson V). However, a general comment, which stands for all traces of videoconference traffic, is that the autocorrelation coefficient is always very large; that is, traffic is highly correlated between successive frames (frames with small sizes are usually followed by similarly sized frames, and the same stands for frames with large sizes). This high autocorrelation can never be perfectly “captured” by a distribution independently generating frame sizes according to a declared mean and standard deviation, and therefore, none of the fitting attempts (including the Pearson V), as good as they might be, can achieve perfect accuracy. Still, very high accuracy can be achieved for multiplexed videoconference sources, similarly to the work in [25] and [26], and for this reason, we developed a DAR model of order 1.

### 3.3 The DAR(1) Model

A DAR model of order \(p\), denoted as DAR(\(p\)) [30], [31], 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 $\frac{1}{26}$ and $\frac{5}{26}$, 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 $X_n = V_n X_{n-1} + (1 - V_n) Y_n$, is a DAR(1) process.

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

$$P = \rho I + (1 - \rho) Q,$$

where $\rho$ is the autocorrelation coefficient, $I$ is the identity matrix, and $Q$ 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

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

where $\mu$ is the mean and $\sigma^2$ is the variance of the frame size for a specific video trace.

As in [25] and [26], where a DAR(1) model with a gamma distribution was used to model the number of cells per frame of VBR teleconferencing video, we built a model based on the Pearson V distribution. Figs. 3 and 4 present a comparison between our DAR(1) model and the actual traces of the Office Cam and ARD Talk sequence for a superposition of 15 and 20 traces, respectively. Figs. 5 and 6 present the respective autocorrelations versus the number of lags $W$ for various trace superpositions. It is clear from the figures that, even for a small number of lags (for example, larger than 10), the autocorrelation of the superposition of movies decreases dramatically for all the traces (for lags larger than 20, which are not shown in these figures, the autocorrelation remains very low, that is, lower than 0.05, for both the actual traces and the DAR(1) model of each trace).
Finally, although Figs. 3, 4, 5, and 6 suggest that the DAR(1) model very well captures the behavior of the multiplexed actual traces, they do not suffice as a result. Therefore, we proceeded with testing our model statistically in order to study whether it produces a good fit for the trace superposition. For this reason, we have again used Q-Q plots.

In Figs. 7 and 8, we have plotted the 0.025-, 0.05-, 0.075-, ..., quantiles (in cells/frame) of the actual Office Cam trace and ARD Talk trace versus the respective quantiles of their DAR(1) model for superpositions of 15 and 20 traces, respectively. To further verify our results for modeling videoconference traffic generated from a single trace, we used two DAR(1) models: the proposed one based on Pearson V and one based on the gamma distribution. As shown in the figures, the points of the Q-Q plot fall almost completely along the 45-degree reference line for our model, with the exception of the last 2.5 percent quantile (right-hand tail), for which our proposed DAR(1) model greatly overestimates the probability of frames with a very large number of cells (a similar result has been observed for the rest of the traces under study, with the difference that the last quantile shows an underestimation of the probability of frames with a very large number of cells). The very good fits show that the superposition of the actual traces can be modeled very well by a respective superposition of data produced by the DAR(1) model. On the contrary, in both figures, it is clear that the DAR(1) model,
which uses the gamma distribution, provides poor results in modeling the superposed traffic.

In Figs. 9 and 10, we again compare the quantiles of the DAR(1) models based on the Pearson V and the gamma distribution, with those of the actual superposed traffic from 50 and 60 videoconference sources, respectively. In the results presented in Fig. 9, we used 15 Office Cam traces, 12 ARD Talk traces, 10 Lecture Cam traces, eight Parking Cam traces, and five N3 Talk traces. In the results presented in Fig. 10, we used 20 Lecture Cam traces, 16 Office Cam traces, 10 N3 Talk traces, eight ARD Talk traces, and six Parking Cam traces. Again, the results clearly show the superiority of our proposed model in comparison to the one using the gamma distribution and, most importantly, the accuracy of our approach in modeling the actual superposed traffic.

In order to further verify the validity of our results, we also performed Kolmogorov-Smirnov tests (KS-tests) for all fitting attempts. The KS-test tries to determine if two data sets differ significantly. The KS-test has the advantage of making no assumption about the distribution of data; that is, it is nonparametric and distribution-free. The KS-test uses the maximum vertical deviation between the two curves as its statistic $D$. For more information on the KS-test, the interested reader is referred to [29]. The results of all our KS-tests confirmed our respective conclusions based on the Q-Q plots (that is, the Pearson V distribution is the best fit). Since this is not the main focus of this work, we present indicatively one such test in Fig. 11, in which, again, we compare the Pearson-V-based and gamma-based DAR(1) models, with the traffic originating from 15 superposed Office Cam traces. The model using the Pearson V distribution provides clearly accurate results, whereas the model using the gamma distribution is shown to fail to model the superposed traffic.

### 4 Data Traffic Model

We adopt the Web traffic model presented in [35] and [36], in which the distributions of the random variables concerning the composition of Web requests are the following:

- **Size of main object**: log-normal(1.31, 1.41) with mean $= 10$ Kbytes and standard deviation $= 25$ Kbytes.
- **Number of inline objects**: gamma(0.24, 23.42) with mean $= 5.55$ and standard deviation $= 11.4$.
- **Size of inline object**: log-normal(−0.75, 2.36) with mean $= 7.7$ Kbytes and standard deviation $= 126$ Kbytes.
- **Number of Web requests per World Wide Web session**: log-normal(1.8, 1.68) with mean $= 25$ pages and standard deviation $= 100$ pages.

The parameters above result in an average Web request size of about 50 Kbytes. The arrival process of Web sessions...
is assumed Poisson with rate $\lambda$ sessions per second. We do not simulate the Web request viewing time in our adopted data model, as it is done in [36]: We only compute the mean download time needed for a user to download all the Web pages in one Web session. This way, we actually consider a worst case scenario of a user asking for consecutive downloads and then viewing the requested material offline in order to check our system’s performance under very bursty data loads.

5 CHANNEL ERROR MODEL

It has been shown by several authors, for many different types of wireless channels, that the wireless channel can be modeled as a finite-state Discrete-Time Markov Chain (DTMC) [50]. The most widely adopted wireless channel error model in the literature is the Gilbert-Elliot [44], [45] model. The Gilbert-Elliot model is a two-state Markov model, where the channel switches between a “good state” (G; always error free) and a “bad state” (error prone). However, many recent studies have shown that the Gilbert-Elliot model fails to predict performance measures depending on longer term correlation of errors [46], minimizes channel capacity [47], and leads to a highly conservative allocation strategy [48].

A better choice for a more robust error model for wireless channels is the model presented in [49], which we adopt in our study. This model, with the use of the short and long error bursts, makes more accurate predictions on the long-term correlation of wireless channel errors than the Gilbert-Elliot model. The error model consists of a three-state DTMC, where one state is the G (error free), and the other two states are the “bad states”: the long bad (LB) and the short bad (SB) states, respectively (the Markov chain is shown in Fig. 12). A transmission is successful only if the channel is in the G; otherwise, it fails. The difference between the LB and SB states is the time correlation of errors: LB corresponds to long bursts of errors, SB to short ones.

The parameters of the error model are presented in Table 2. The average number of error bursts, in slots, experienced when the states LB and SB are entered are respectively given by $B_{LB} = 1/p_{bg,L}$ and $B_{SB} = 1/p_{bg,S}$, where $p_{bg,S}$ is the transition probability from state SB to G, and $p_{bg,L}$ is the transition probability from state LB to G. Similarly, the average number of consecutive error-free slots is given by $B_G = 1/p_{gb}$, where $p_{gb}$ is the probability to leave state G. The parameter $k$ is the probability that the Markov chain moves to state LB, given that it leaves state G. $k$ also represents the probability that an error burst is long (that is, the fraction of long bursts over the total number of error bursts).

We have chosen in our study the value of the probability $P_{bad}$, that is, the steady-state probability that the channel is in the bad state, to be equal to $8 \times 10^{-5}$. This value has been chosen in order to test an “almost-worst”-case scenario for our system, as the video packet dropping probability is set to $10^{-4}$, and by choosing a value of bad state probability larger than the upper bound on video packet dropping, the strict QoS requirement of video users would certainly be violated. The values for $p_{gb}$ and for the parameter $k$ were

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Error Model Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{bad} = 0.99992$</td>
<td></td>
</tr>
<tr>
<td>$B_{LB} = 1/p_{bg,L} = 65160$ slots</td>
<td></td>
</tr>
<tr>
<td>$B_{SB} = 1/p_{bg,S} = 2.38$ slots</td>
<td></td>
</tr>
<tr>
<td>$B_G = 1/p_{gb} = 59.53$ slots</td>
<td></td>
</tr>
<tr>
<td>$k = 0.05$</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 11. KS-test of DAR(1) Pearson V and gamma models versus the actual Office Cam trace for 15 superposed sources.

Fig. 12. Channel error model.
taken from [49], as well as the ratio between $p_{b\text{,eq}}$ and $p_{b\text{,L}}$. The value for $p_{b\text{,L}}$ is derived from the steady-state behavior of the Markov chain for the bad state probability chosen. A complete study on the wireless cellular channel model should also relate the transition probabilities to the velocity of each user; however, this parameter has minor importance for the specific work. The reason for this is our attempt to test the system “at its limits,” that is, we focus on studying network utilization in the “almost-worst-case” scenario, as explained earlier. Hence, the incorporation of user velocity into the model would only result in the user channel being, on the average, in a better condition than the one assumed in our study (if it was worse, then video traffic could not be serviced at all, as explained above).

6 THE CAC MECHANISM

The idea of our proposed CAC mechanism came as a result of the combination of

1. the quality of our video traffic model and
2. the fact that source bandwidth requirements for wireless videoconference can only be limited, by default, due to the type of the application. The difference, in terms of bandwidth requirements, between one person talking or gesturing, which is the case in videoconference traffic, and a bomb exploding near the protagonist in an action movie is quite clear. In the case of videoconference traffic, scene changes do not occur often, and this is the reason for the very high autocorrelation of lag-1, which was shown in Figs. 5 and 6.

Therefore, we were led to investigate the possibility of precomputing a large number of videoconference traffic scenarios (that is, to precompute the average bandwidth needed by videoconference sources in each traffic scenario) based on the five “modes” (sets of parameters) presented in Table 1. This precomputation will be the basis for admitting or rejecting a new videoconference call since the CAC will have positive knowledge of the network’s ability to accommodate a new source while satisfying the QoS requirements of the new call, as well as those of all the calls already admitted in the picocell.

6.1 Description of Our CAC Scheme and Its Implementations

6.1.1 First Implementation

By using the “mode” (peak, mean, or standard deviation) declared by the source in its QoS contract with the network at call setup, each incoming source can be “identified” and modeled with the use of our DAR(1) model (given that the number of “modes” is large enough and that videoconference traffic has limited variation in its bandwidth requirements, this should be an easy task). Then, the CAC simply checks in its precomputed scenarios’ list whether the admittance of a new videoconference call with the specific characteristics is possible or not (that is, it checks in the list to find the bandwidth needed for the new total number of calls, including the new call). All calls in this case are supposed to be generated within the picocell under study. This was the simplest framework within which our CAC mechanism was implemented.

6.1.2 Second Implementation

In this case, data (Web) traffic was integrated with videoconference traffic in the system. Our CAC policy for data is the following: A data user is admitted in the system by adding its declared mean rate requirements to the existing estimated bandwidth for videoconference and data sources and by checking whether the new estimated bandwidth is higher than the channel information rate. Also, the videoconference traffic is served by the BS with absolute priority (enforced via data preemption) as it has the strictest QoS requirements.

It should be noted that the bandwidth requirements of a Web user cannot be well estimated per session as there are many parameters in the Web model and the required bandwidth for each source deviates significantly from the estimated mean per session (the large standard deviation, in all the models used, is clearly shown by the model description in Section 4). This is logical, as quite often, Web browsing leads users from one Web page to another; however, this is not a critical factor in our study as the QoS contract agreement between the network and the user will be based on the data rate with which the network will provide the user for downloads. If the user attempts to exceed, through its Web requests, its agreed mean download rate, then the system can allow or prevent this behavior depending on whether the additional bandwidth that the user attempts to consume is free and on the algorithm on which the system’s Traffic Policing mechanism will be based. The simplest solution would be, in the case of high traffic load, to restrict the user to its specified rate through the use of a token bucket (leaky bucket) mechanism. The development of such a mechanism, however, is beyond the scope of this work, which is focused on CAC.

6.1.3 Third Implementation

This was the most complicated implementation of our CAC scheme, as videoconference users were divided based on their service class. Differentiated service contracts have already started being adopted by many wireless carriers based on the QoS demanded and paid for by the user. This fact was taken into account in our work in [16], where we considered cases in which a percentage of video users did not accept quality degradation (so that more calls can be accepted in the network and a higher channel throughput can be achieved). In [16], in the case of H.263 traffic, videoconference users were categorized in three service classes, each corresponding to just one type of video quality. In this work, we expand this approach by considering again three classes of users, corresponding this time to a total of five different types of video qualities. The three service classes are defined respectively as Premium-Quality (PQ), Standard-Quality (SQ), and Low-Quality (LQ) users. PQ users accept no quality degradation, SQ users accept quality degradation up to a certain level, and LQ users accept quality degradation up to the lowest level (still, if the channel load is low, SQ users and LQ users are entitled to the highest QoS agreed in their QoS contract). In the case when the channel load, with the admission of a new call, is computed to be higher than the channel information rate, users are gradually degraded up to the point where the new call can be admitted. Additionally, in this case, we have studied the situation when a portion of
the videoconference traffic in a cell originates from handoff calls. Handoff calls have absolute priority in obtaining an equal amount of channel bandwidth as the one that they were occupying in their previous picocell location; that is, handoff calls are not expected to endure any quality degradation as this would lead to user dissatisfaction. The rest of the bandwidth is free for allocation to any new videoconference calls generated within the picocell. As explained in Section 1, no portion of the bandwidth is exclusively dedicated to handoff calls: Our approach is a QP scheme with infinite queue length (as long as the QoS for video traffic is satisfied). Both handoff calls and new calls originating from within the picocell are accepted if enough free channel bandwidth exists to accommodate them, and no portion of the bandwidth is restricted for access of either type of call.

6.2 Discussion on the Proposed Scheme

A very important general comment that needs to be made on the proposed scheme is that, naturally, not all the possible video traffic scenarios can be precomputed. This is especially true since, in our scheme (and in next-generation wireless cellular networks), the wireless system will have a large number of degrees of freedom (in our scheme, this is expressed with the use of different modes and different service classes for video users). Hence, it would be impossible for the CAC module to precompute all possible traffic scenarios. On the contrary, it would be feasible to precompute only a small percentage of the total number of possible traffic scenarios. However, it should be taken into consideration that a wireless provider has total knowledge over the contracts of its customers (in terms of bandwidth and QoS requirements). Based on this knowledge, the wireless provider can precompute the traffic scenarios that are applicable to its clientele at a given period of time. These precomputed scenarios will, according to our proposal, be the initial precomputed traffic scenarios in the CAC module’s database. In the case that the network traffic entering the system at a given time does not correspond to a traffic scenario included in the database (this can easily take place as the number of clients of the wireless provider changes daily, as do their requirements), the CAC will use our video traffic model in order to compute online its estimation for the new traffic scenario. More specifically, the CAC will first identify the initial precomputed traffic scenario that is the closest (in terms of number of video users and their respective qualities). Then, it will estimate the total bandwidth needed by the sources already in the system plus the new source by adding to the precomputed estimation the estimation given for the new source by an “equivalent bandwidth” estimation method. This initial slightly conservative estimation of the total bandwidth (but still much less conservative compared to using the “equivalent bandwidth method” from scratch) will guarantee that, if the new user is admitted by the CAC, then this will not affect the QoS of existing users. Still, this overestimation will need to be used only for a short time window. During this time window, the CAC will proceed with using our DAR(1) modeling approach for estimating the new total traffic load. Due to the simplicity of the DAR(1) modeling approach, this is not a computationally heavy task for the CAC. After the end of this time window (which can be as low as 2 minutes, given the accuracy of the DAR model and the fact that 2 minute windows were shown to be a satisfactory choice in our earlier work in [16] for this type of traffic), the estimation of the total bandwidth will be made solely based on our modeling scheme. This online estimation can continue at the CAC during the whole duration of the new call so that, in given periods of time, corrective estimations can be made in the system; however, these corrections in the predicted total bandwidth needed by the sources will be small due to the accuracy of our modeling approach. After the completion of the call and the respective online estimation, the estimation will be added to the CAC module’s database; hence, in future cases of the same traffic scenario, even the small amount of time that was needed for the online estimation will not be necessary for the system to respond.

As shown from our results in Section 7, our scheme provides almost-optimal CAC both in the case of videoconference traffic being the sole type of traffic transmitted in the system and in the case of integrated videoconference traffic with data traffic. The very good performance of our scheme will also be shown in the case of handoff traffic arrivals.

6.3 Conceptual Comparison with Other CAC Approaches

In this section, we proceed to make a conceptual comparison between our scheme and the CAC approaches that have been used in the literature and were briefly explained in Section 1. Regarding the guaranteed service [5], [6], the comparison between the two mechanisms is fairly easy. Our mechanism provides an accurate estimation of the multiplexed sources’ behavior, whereas CAC algorithms for guaranteed service base their decisions on the worst case behavior of all the existing flows in addition to the incoming one. Network utilization under the guaranteed service approach is usually acceptable when flows are smooth. When flows are bursty, however, the guaranteed service inevitably results in very low utilization [7]. Since video traffic (even videoconference traffic, which is less bursty) is never smooth and does not often transmit at its peak rate for a significant amount of time, a large portion of the wireless channel bandwidth is left unused with the guaranteed service CACs. On the other hand, with the use of a guaranteed service CAC mechanism, sources never experience any packet loss, whereas, with our mechanism, packet loss always takes place; however, this loss is kept in our scheme well below the set maximum allowed video packet dropping rate of 0.01 percent [34].

As explained in Section 1, our scheme generally falls within the category of the probabilistic service [8]. However, the regular use in this type of scheme of the “equivalent bandwidth” method for estimating the aggregate bandwidth (for example, as in [10], [11], [12], [13], [14], [15], [16], [32], and [33]), which will be needed by a superposition of sources, is a quite conservative approach, as it has been shown in the literature. It only provides an approximate formula, which generally significantly overestimates the sources’ actual bandwidth requirements. This results in good QoS for all admitted users (customers), but as in the case of the guaranteed service, a significant portion of the wireless channel bandwidth is again left unused. More specifically, in [10], the equivalent bandwidth of a set of flows is defined as the bandwidth \( C(e) \), which is such that
the stationary bandwidth requirement of the set of flows exceeds this value, with a probability of at most $\epsilon$, where $\epsilon$ is the packet loss rate. However, if the packet loss rate is small enough, then it is expected that the algorithm of Section 5 will perform well. The authors point out in [4] that QoS contracts between the network provider and the customers will specify the level of violations over some macroscopic time scale (for example, days or weeks) rather than over a few hundred packet times. However, this only applies to specific tolerant applications in wired networks. The authors' result in [4], which shows that predictive service is a viable alternative to guaranteed service “for those applications willing to tolerate occasional delay violations, since it provides fairly reliable delay bounds,” is clearly not applicable in the case of applications such as videoconferencing over wireless networks, which have hard packet delay and packet dropping requirements. The CAC scheme presented in [38] is shown to generalize and outperform the scheme proposed in [4] by introducing a mechanism that exploits measured peak rate envelopes of the aggregate traffic flow in the network to allocate network resources. Since there is no assurance that the aggregate flow will continue to be bounded by its past behavior, the authors have developed a theory to quantify the confidence level of a schedulability condition, which attempts to incorporate the randomness of the aggregate envelope. However, although the proposed scheme in [38] is efficient, it suffers once again from the risk that is embodied in all predictive service CAC schemes. In their “conditional prediction of traffic envelope,” the authors point out that the mean rate can be optimally predicted only when the peak rate of the aggregate traffic flow is a correlated Gaussian process. This is not the case in our study for videoconference traffic, and additionally, the peak rate and mean rate are subject to quick and significant changes in a wireless network due to user mobility and the respective incoming/outgoing handoff calls [38] was designed for wired networks); therefore, the measurement window $T$ would have to change constantly over time to keep up with the changes in network load. This would require an increased computational complexity in the mechanism since the authors explain that a poorly set $T$ leads to an underutilization of network resources. Also, if the conditional prediction of the traffic envelope is not performed (for the above reasons), it is explained by the authors that their algorithm can lead to an admission of more traffic flows than the network can handle, given the users’ QoS requirements. A CAC scheme, which still falls within the category of predictive service but is the closest conceptually to our scheme, was proposed in [39]. The authors use measurements to determine admission control, but the admission decisions are precomputed. However, this precomputation is based on the assumption that a prior distribution is known for the offered load; therefore, this algorithm is not applicable to wireless networks, where an ongoing multimedia call in a given cell may often change its bandwidth due to a new call arrival, a call completion, or an incoming/outgoing handoff call [37].

Finally, regarding the second major classification of CAC schemes in the literature, our work excels conceptually in comparison to the GC scheme approach (for example, [1], [17], [18], [19], and [20]). The reasons for this are 1) our scheme’s great simplicity contrary to the complexity of the GC schemes in order to estimate the proper portion of channel bandwidth to be allocated solely to handoff calls and 2) our scheme’s significant accuracy due to the accuracy of our videoconference traffic model.

One key issue that needs to be addressed in respect to our CAC proposal is the applicability of the scheme as a more general proposal for wireless networks. Since the scheme is based on the accuracy of a videoconference traffic model for H.263 encoded video, it could be argued that the applicability of the scheme is restricted to the specific traffic scenario (for example, as correctly stated in [4], it is quite difficult, if not impossible, to provide accurate and tight statistical models for each individual flow in a network, especially as new wireless applications today continuously emerge at a high rate). However,

1. Videoconference applications gain in significance every day; hence, videoconference traffic is expected to soon become one of the key applications in wireless networks, and the most well-known and used video standards for this application today are H.263 and MPEG-4. For MPEG-4 videoconference traffic, our group has also recently completed the development of a traffic model based on DAR(1) modeling [56].
2. Videoconference applications are very greedy in terms of bandwidth requirements. Data traffic, as it is intuitively easy to understand but will also be shown from our results, is much less bandwidth consuming, regardless of the high burstiness of Web traffic. For this reason, it is much easier for the system to accommodate. The same stands for other types of traffic, such as voice and streaming video over wireless networks, which have the additional
advantage of limited or no burstiness in comparison to data traffic. Therefore, we believe that it is actually not necessary to adopt the same approach (that is, of basing the CAC scheme on traffic models) for all types of flows in the wireless network. If an accurate model exists for videoconference traffic (or, possibly, in the future, for an equally greedy type of flow), this would be enough for the provision of a CAC scheme that will be much less conservative than the guaranteed service and equivalent bandwidth probabilistic service approaches and much less risky than the predictive service approach.

Therefore, our work focuses on a specific type of traffic but a very important one. The remaining types of flows could be handled either with a measurement-based approach, since the network can have more relaxed service commitments to those users, or with the same approach used for Web traffic in our study.

7 Results and Discussion

As analyzed in Section 6.1, we have implemented our algorithm in three different cases, the results for which will be presented below. The QoS requirements are common in all these cases:

1. For videoconference traffic, the set maximum allowed video packet dropping rate is 0.01 percent [34], and the maximum transmission delay for the video packets of a VF is equal to the time before the expected arrival of the next VF with packets being dropped when the deadline is reached (the interframe period in H.263-encoded movies is not constant: It is an integer multiple of 40 ms).

2. For Web downloads, as explained in Section 4, we consider a worst case scenario of a user asking for consecutive downloads and then viewing the requested material offline in order to check our system's performance under very bursty data load. As shown from our results, regardless of the large number of 25 Web pages (requests) per data session, the very large standard deviation of 100 pages and the utilization of the largest portion of the channel bandwidth for videoconference downloads, the mean download delay does not exceed a few minutes in any of the studied scenarios.

The interarrival time both for handoff and new video calls is exponential.

7.1 First Algorithm Implementation

In this case, videoconference traffic is the only traffic type in the system, and all calls are generated from within the picocell. Our scheme of precomputed traffic scenarios is evaluated in 16 different scenarios versus the actual traffic generated by the real video traces. These scenarios are listed as follows:

Scenarios 1-7. Each of the five “modes” used in our study is used with a 20 percent probability (that is, a user who “wakes up” chooses one of the five “modes,” with a probability equal to 20 percent). The number of each scenario corresponds to the number of traces using each mode. For example, in scenario 4, there are 20 videoconference traces present in the system.

In each of the traffic scenarios (scenarios 8-16) presented in Table 3, we have considered various combinations of the cases where each of the five modes is selected by the users with probabilities of 10 percent, 15 percent, 20 percent, 25 percent, or 30 percent, and the total number of users present in the system is 31. The reason for this choice will be explained below.

One of the most well-known formulas for the estimation of the “equivalent bandwidth” of a set of flows was introduced in [10]. As explained in Section 6.3, the equivalent bandwidth of a set of flows is defined in [10] as the bandwidth \( C(\epsilon) \), which is such that the stationary bandwidth requirement of the set of flows exceeds this value with a probability of at most \( \epsilon \), where \( \epsilon \) is the packet loss rate (0.01 percent in our study). In order to investigate our mechanism’s performance, we compare, as shown in Table 4, the bandwidth utilization results from the use of the actual traces using the five modes with the respective results from our mechanism (using the models for each mode) and those from [10]. Each simulation point is the result of an average of 10 independent runs, each simulating 1 hour of network operation.

It is clear from the results presented in Table 4 that the equivalent bandwidth estimation with the use of the formula in [10] leads to an enormous overestimation of the actual bandwidth requirements of the superposition of videoconference sources (this overestimation ranges from 24.85 percent to 61.7 percent). On the contrary, the estimation provided by our mechanism yields a very small underestimation of the actual bandwidth requirements of the superposed sources (as an effect of the superposition of the DAR models), which ranges from a minimum of 0.6 percent to a maximum of 1.4 percent. The accuracy of our model leads to an equally accurate prediction in all
16 scenarios of the possibility of accommodating a superposition of videoconference sources, as not only is the required bandwidth very precisely estimated but also the percentage of dropped packets with the use of our model is indicative of whether a specific load can be supported by the system; that is, although the small underestimation with the use of our DAR models leads to a smaller packet dropping rate in our mechanism than the packet dropping when the real traces are used, there is no scenario among all the ones that we studied where the packet dropping of the real traces violated the upper bound of 0.01 percent without the same result also taking place for our scheme (for example, scenarios 7, 10, and 16 are shown to contain an overload of traffic with which the system cannot cope without excessive packet dropping both when using the real traces and our modeling approach).

This is not the case when the equivalent bandwidth estimation from [10] is used. The overestimation of the actual bandwidth requirements of the sources’ superposition is so large that four traffic scenarios (scenarios 6, 12, 13, and 14), which can be accommodated based on the actual sources' requirements (and are accommodated with the use of our scheme), are estimated as impossible to accommodate when the equivalent bandwidth estimation is used. This means, for example (scenario 6), that in the case of each mode being selected with an equal probability of 20 percent, the equivalent bandwidth estimation method predicts that up to 25 users can enter the system without a violation of the users' QoS requirements, whereas our scheme is accurate in computing that up to 30 users can enter the system without violating the users’ QoS requirements, whereas our scheme is accurate in computing that up to 30 users can enter the system without a violation of the users’ QoS requirements.

A final comment that needs to be made in order to make a completely fair comparison with the equivalent bandwidth-based CAC mechanisms is that the adoption of a single formula for the estimation of the equivalent bandwidth is not the best choice for the assignment of an “equivalent bandwidth.” As shown in [43], if the operation point of a queuing system can be determined, then an effective bandwidth assignment can be made, which will be quite close to the actual user needs. However, the solution of the operation point is often not an easy task, and this is especially true in the complex case of today’s and future wireless networks, where many different types of traffic and many different types of users of each type of traffic will be present.
7.2 Second Algorithm Implementation

In this case, Web traffic is integrated with videoconference traffic in the system, and again, all calls are supposed to generate from within the picocell. To avoid repetitive results, we do not present results with the use of the equivalent bandwidth estimation formula in this case, as it has been shown in Section 7.1 to provide clearly inferior performance in comparison to our scheme. Therefore, in this section, we only compare our scheme’s performance in estimating the users’ bandwidth requirements with the actual requirements generating from the integration of the real videoconference traces used in our study with Web traffic. Our CAC scheme for this case has been outlined in Section 6.1. In brief, we note that a data user is admitted in the system by adding its declared mean rate requirements to the existing estimated bandwidth for videoconference and data sources and checking whether the new estimated bandwidth is higher than the channel information rate. Videoconference traffic is served by the BS with absolute priority.

Table 5 presents the comparison of the bandwidth utilization results from the use of the actual traces with the respective results from our mechanism for the video/data integration case. More specifically, Table 5 presents the maximum Web session arrival rate that the system can sustain for each traffic scenario, the mean session delay, the percentage of dropped video packets, and the channel bandwidth consumption. No results are presented for scenarios 7, 10, and 16, as it has been shown from the results in Table 4 that the videoconference traffic generated in these scenarios cannot be accommodated by our system. Web users, in all the results in Table 5, adhere to their declared mean download rate.

It is clear from the above results that, once again, the accuracy of our videoconference traffic model leads to an equally accurate prediction, in all the scenarios under study, of the possibility of accommodating the superposition of videoconference and Web data sources. In this case, we observe in Table 5 that our mechanism provides in some cases an overestimation and in some other cases an underestimation of the actual bandwidth requirements of the superposed sources. The combination of the large standard deviation of the Web model used in our study with the burstiness of the videoconference traffic are responsible for this difference in comparison to the case of the first implementation of our CAC algorithm where our mechanism provided for all the studied scenarios a small underestimation of the real traces’ bandwidth requirements. This combination is also responsible for the “fluctuations” in the mean session delay, as shown in Table 5. Still, once again, the difference between our mechanism’s estimation and the actual video and data bandwidth requirements is very small, ranging from a minimum of 0.012 percent to a maximum of 6.47 percent and averaging at just 1.7 percent over all the studied scenarios. From the 13 scenarios for which results are presented in Table 5, in only three scenarios does the difference between our results and the results with the use of the actual traces exceed 1.5 percent (again, the large deviation in all the models used for the simulation of the Web downloads are responsible for this result), whereas, in eight of the 13 scenarios, the respective difference is well below 1 percent.

All the above results have been produced for an upper bound of 8 minutes in Web session download delay and for video packet dropping less than 0.01 percent. Given the average Web request size of about 50 Kbytes, the high mean number of pages per Web session (25), the very high standard deviation (100), and the presence of bursty video traffic in the system, we consider this to be a moderate upper bound.

As shown in Table 5, the maximum Web session arrival rate that the system can support so that both the above-mentioned QoS requirements are guaranteed to video and Web users is almost identical in our mechanism with the one actually supported by the system when real videoconference traces are used. Also, as shown in Table 5, in both the cases of real traces and DAR models, the maximum

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Maximum Web session arrival rate (sessions/sec)</th>
<th>Mean Session Delay (minutes)</th>
<th>Bandwidth (Mbps)</th>
<th>Maximum Web session arrival rate (sessions/sec)</th>
<th>Mean Session Delay (minutes)</th>
<th>Bandwidth (Mbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.044</td>
<td>6.24</td>
<td>7.057</td>
<td>0.045</td>
<td>4.24</td>
<td>7.276</td>
</tr>
<tr>
<td>2</td>
<td>0.041</td>
<td>7.44</td>
<td>7.694</td>
<td>0.038</td>
<td>3.68</td>
<td>7.307</td>
</tr>
<tr>
<td>3</td>
<td>0.031</td>
<td>3.62</td>
<td>7.361</td>
<td>0.031</td>
<td>3.38</td>
<td>7.254</td>
</tr>
<tr>
<td>4</td>
<td>0.026</td>
<td>2.71</td>
<td>7.587</td>
<td>0.026</td>
<td>3.32</td>
<td>7.546</td>
</tr>
<tr>
<td>5</td>
<td>0.022</td>
<td>4.21</td>
<td>8.063</td>
<td>0.018</td>
<td>3.21</td>
<td>7.541</td>
</tr>
<tr>
<td>6</td>
<td>0.013</td>
<td>2.51</td>
<td>7.835</td>
<td>0.013</td>
<td>2.52</td>
<td>7.811</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>0.018</td>
<td>3.58</td>
<td>7.782</td>
<td>0.019</td>
<td>2.38</td>
<td>7.829</td>
</tr>
<tr>
<td>9</td>
<td>0.021</td>
<td>3.96</td>
<td>8.176</td>
<td>0.018</td>
<td>3.79</td>
<td>7.837</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>0.016</td>
<td>2.94</td>
<td>7.909</td>
<td>0.016</td>
<td>4.81</td>
<td>7.901</td>
</tr>
<tr>
<td>12</td>
<td>0.013</td>
<td>2.31</td>
<td>7.942</td>
<td>0.013</td>
<td>2.61</td>
<td>7.872</td>
</tr>
<tr>
<td>13</td>
<td>0.012</td>
<td>2.24</td>
<td>8.251</td>
<td>0.012</td>
<td>5.35</td>
<td>8.226</td>
</tr>
<tr>
<td>14</td>
<td>0.012</td>
<td>2.49</td>
<td>8.049</td>
<td>0.012</td>
<td>6.21</td>
<td>8.079</td>
</tr>
<tr>
<td>15</td>
<td>0.015</td>
<td>4.01</td>
<td>7.984</td>
<td>0.015</td>
<td>4.21</td>
<td>7.985</td>
</tr>
<tr>
<td>16</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
system throughput is achieved again (as in Section 7.1) in scenario 13 and the respective throughputs are 91.22 percent and 90.95 percent.

7.3 Third Algorithm Implementation

As explained in Section 6.1, we have also implemented our CAC scheme on a more complex case than the ones presented in the previous sections. A portion of the total traffic originates from handoffed videoconference calls, which can be any of the five traces with equal probability and which accept no quality degradation and expect to be fully serviced by the new BS (hard handoff). On the contrary, videoconference calls originating from within the cell under study can accept quality degradation if they belong to SQ or LQ users. When the channel load, with the admission of a new call, is higher than the channel information rate, users are gradually degraded up to the point where the new call can be admitted.

Each of the five traces corresponds to a video quality. The highest video quality is that of the Parking trace, followed by the N3 Talk, ARD Talk, Office, and Lecture traces (these will be referred to as qualities 1-5 for the rest of the paper). This choice has been made after a careful observation of the traces’ behavior in terms of bandwidth requirements, which led us to conclude that the basis for our choice of video qualities should be the mean rate of each trace. The burstiness (peak/mean ratio) of the traces is of minor importance since all traces under study seldom transmit at their peak rate. PQ terminals use only quality 1 and accept no quality degradation, SQ terminals use qualities 2-4 and accept quality degradation up to quality 4, and LQ terminals use qualities 2-5 and accept quality degradation up to quality 5. The type of each user is determined with equal probability, and after the type determination, if the user is SQ or LQ, then the specific quality with which the user enters the system is again determined with equal probability.

In the results presented in Table 6, we have studied our scheme’s performance for scenarios 6, 8, 9, and 11-15. As in Section 7.2, no results are presented for scenarios 7, 10, and 16, as it has been shown from the results in Table 4 that the videoconference traffic generated in these scenarios cannot be accommodated by our system. Also, we do not present results for scenarios 1-5, as we focus on testing our system’s behavior under a heavy traffic load, and it has already been shown in Section 7.1 that the maximum number of traces that the system can accommodate in the “equal probability” case is 30. Therefore, we have used only the scenarios with 30 or 31 videoconference calls originating from within the cell.

Table 6 presents the comparison of the bandwidth utilization results from the use of the actual traces with the respective results from our mechanism when 5 percent, 10 percent, 15 percent, 20 percent, and 25 percent of the total videoconference traffic in the system is generated from handoffs (for example, in scenario 6, for the 25 percent handoff case, 30 videoconference calls are originating from within the cell under study, and 10 more calls originate from handoffs).

The results presented in Table 6 indicate once again that the difference between our mechanism’s estimation and the actual videoconference bandwidth requirements is very small: It ranges (either as an overestimation or an underestimation) from a minimum of 0.088 percent to a maximum of 1.28 percent and averages at a mere 0.59 percent over all the studied scenarios.

As shown in Table 6, once again, the maximum system throughput is achieved in scenario 13 for both the cases of real traces and DAR models and, more specifically, in the case of the 25 percent handoff traffic. This is expected, as, in this case, a larger number of video traces are accommodated by the system by the gradual degradation of SQ and LQ users to their lowest acceptable video quality. The respective throughputs for the real traces and our scheme are 76.15 percent and 76.04 percent.

All the above results have been produced for an upper bound of 0.01 percent for video packet dropping. Also, as in the results presented in Table 4, the video packet dropping rates for all the cases studied and presented in Table 6 were very similar when real traces and DAR models were used in the same traffic scenarios. The percentage of dropped video packets for users trying to establish a new call was, in all the studied cases, more than 96 percent of the total video packet dropping (the remaining percentage was caused by dropped video packets of handoff users). This result stems from our choice to offer absolute priority to handoff calls.

In order to avoid repetition, we do not add in this part of our work results for the case of integrated videoconference and Web traffic. Our simulations have once again shown, as it has been clearly presented in Section 7.2, that our scheme accurately predicts, in all the scenarios under study, the possibility of accommodating the superposition of videoconference and Web data sources, regardless of whether

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Real Traces - Bandwidth under various Handoff Loads (%)</th>
<th>DAR Model - Bandwidth under various Handoff Loads (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>5% 10% 15% 20% 25%</td>
<td>5% 10% 15% 20% 25%</td>
</tr>
<tr>
<td>9</td>
<td>5.534 5.649 5.943 6.175 6.296</td>
<td>5.501 5.607 5.897 6.147 6.239</td>
</tr>
</tbody>
</table>
Web downloads originate from handoffs or from within the cell under study.

However, there is always the case that users’ requirements may exceed the bandwidth requirements variability predicted by the wireless carrier. The number of wireless users is so large and growing so fast that certain users may not be satisfied by the “modes” offered by the carrier and demand/declare a different set of parameters for their videoconference call. The system, in this case, would have to search its pool of “modes” and find the “mode” that is most similar to the set of parameters declared by the user. We have simulated this case in our system, focusing on the most “difficult” of the implementations examined; that is, we used the scenarios for the third implementation of our algorithm (all users within the cell originate from known “modes”) with the alteration that the handoff traffic arriving at various time intervals in the system originates from “unknown modes” with which the system has to cope very quickly in order to incorporate the new calls. The two “unknown modes” used were those corresponding to two other videoconference-type traces from [27] and [28], the parameters of which are shown in Table 7.

We have used the Mean Square Error (MSE) measure in order to find which of the five “modes” in our pool is the most similar to those of the handoffed traffic. The “mode” selected in each case was the one with the smallest MSE from the handoffed trace when adding the MSEs for the mean, peak, and standard deviation and dividing by 3 (no “weights” are used, as the similarity of the mode to the mean, peak, and standard deviation, respectively, is considered of equal importance). The “Boulevard Bio” trace was found to be most similar to the “ARD Talk” mode and the “ARD News” trace is most similar to the “N3 Talk” mode.

All simulations were conducted for the case when 15 percent of the total videoconference traffic in the system is generated from handoffs.

The results presented in Table 8 show once again that the difference between our mechanism’s estimation and the actual videoconference bandwidth requirements is very small: It ranges (either as an overestimation or an underestimation) from a minimum of 0.076 percent to a maximum of 1.67 percent and averages at only 0.84 percent over all the studied scenarios.

However, in this case, there is a difference between the real traces’ video packet dropping results and our scheme’s video packet dropping results. The very high standard deviations (shown in Table 7) of the two traces used as “unknown modes” are not perfectly modeled by the modes selected from our mode pool; therefore, the video packet dropping is underestimated by our scheme by $3 \times 10^{-5}$ on the average. This means that, although our scheme very accurately predicts the bandwidth requirements of the unknown modes, it makes slightly less accurate predictions of their burstiness. Although the video packet dropping underestimation is quantitatively very small, it is significant in quality since the upper bound for video packet dropping is only $10^{-4}$. Therefore, we conclude that, in the case of “unknown modes,” the upper bound for video packet dropping when our mechanism is used should be even stricter (less than 0.01 percent) for our mechanism to be able to make “safe” decisions on the admittance of a new videoconference call.

Based on all the above results, we believe that the great precision in our scheme’s predictions can become even higher with the use of a slightly larger pool of videoconference “modes” from which traffic scenarios will be precomputed (for example, with the use of 10 modes, that is, a doubly sized pool of modes than the one used in this study). The use of a slightly larger number of modes will guarantee the existence of a 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. As explained in Sections 1 and 6, 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.

It should also be noted, as a general comment on all the presented results, that the choice of an ATM packet size further emphasizes the accuracy of our scheme, as with a larger packet size, even fewer video sources would need to be superposed for our DAR(1) model to work accurately (that is, for the superposition of movies to have low autocorrelation, which is needed for the model).

Finally, it should be emphasized that, in comparison to the measurement-based approach, the proposed scheme is the better choice for videoconference traffic due to its capability of predicting the bursty nature of the traffic

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Real Traces - Bandwidth (Mbps)</th>
<th>DAR Model - Bandwidth (Mbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6.612</td>
<td>6.617</td>
</tr>
<tr>
<td>8</td>
<td>6.181</td>
<td>6.078</td>
</tr>
<tr>
<td>9</td>
<td>6.204</td>
<td>6.145</td>
</tr>
<tr>
<td>11</td>
<td>6.417</td>
<td>6.402</td>
</tr>
<tr>
<td>12</td>
<td>6.614</td>
<td>6.691</td>
</tr>
<tr>
<td>13</td>
<td>6.712</td>
<td>6.779</td>
</tr>
<tr>
<td>14</td>
<td>6.708</td>
<td>6.802</td>
</tr>
<tr>
<td>15</td>
<td>6.703</td>
<td>6.686</td>
</tr>
</tbody>
</table>
(which is impossible for the measurement-based approach, in which the burstiness of the traffic would be “captured” and incorporated in the CAC after initially suffering from severe packet loss). This is of major importance for this type of traffic due to its hard packet delay and packet dropping requirements as opposed to the relaxed service commitments of the network to the user, which are assumed when the measurement-based approach is used. If, however, the two approaches were to be compared over longer time scales, the schemes would have comparable results.

8 Conclusions

Based on an accurate videoconference traffic model that has been developed by our group, we have proposed in this work a new efficient CAC scheme for multimedia traffic transmission over wireless cellular networks. The novelty of the scheme lies in the utilization of precomputed traffic scenarios, combined with online simulation, for decision making on the acceptance or rejection of a new videoconference call. The precomputation is based on the traffic parameters declared by the video source at call setup. 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. Our scheme is shown to excel both conceptually and in simulation results when compared to many well-known existing CAC approaches. As this work has focused on the evaluation of our proposed scheme over one cell (picocell) of the wireless network, our future work will include the extension of our scheme’s evaluation over a whole wireless network topology.

References


