Technical University of Crete

Dept. of Electronics & Computer Engineering

Telecommunications Division

Information & Computer Networks Laboratory

A New Call Admission Control Mechanism for Multimedia Traffic over Wireless Networks

Χατζηπέρης Στυλιανός

Undergraduate Thesis

Summer 2005

Στους γονείς μου, Νεκτάριο και Χαρίκλεια, και

τον αδερφό μου, Άγγελο.

Ευχαριστίες

Ένα μεγάλο ευχαριστώ, αρχικά, στον επιβλέποντα καθηγητή μου κ. Μιχάλη Πατεράκη για την εμπιστοσύνη που έδειξε στο πρόσωπο μου και τις δυνατότητες μου, για τις σημαντικές παρατηρήσεις και διορθώσεις πάνω στην διπλωματική μου εργασία και την πρόθυμη βοήθειά του σε οτιδήποτε χρειάστηκε. Επίσης, ένα μεγάλο ευχαριστώ στον συνεπιβλέποντα διδάσκοντα κ. Πολυχρόνη Κουτσάκη για την συνεργασία και την συνεχή και καθοριστική αρωγή του, χωρίς την οποία το αποτέλεσμα δεν θα ήταν, σε καμία περίπτωση, το ίδιο. Ένα ευχαριστώ στον καθηγητή κ. Νίκο Σιδηρόπουλο για την ανάγνωση της διπλωματικής μου και τα σχόλιά του. Ακόμα, ένα μεγάλο ευχαριστώ σε όλους τους καθηγητές μου, που θέτοντας ψηλά τον πήχη των απαιτήσεων τους με δίδαξαν πολλά και τέλος, ένα μεγάλο ευχαριστώ σε όλους τους φίλους μου τα τελευταία πέντε χρόνια, που με την φιλία τους έκαναν ευχάριστη την διαμονή μου στα Χανιά.

Σας ευχαριστώ όλους.

Χατζηπέρης Στυλιανός

Χανιά, Ιούλιος 2005

Εισαγωγή

Η παρουσία των ασύρματων επικοινωνιών στη ζωή μας γίνεται συνεχώς εντονότερη, καθώς όλο και περισσότερες υπηρεσίες δεδομένων και πολυμέσων παρέχονται πλέον από τα ασύρματα δίκτυα, δημιουργώντας ακόμα μεγαλύτερες προκλήσεις στην προσπάθεια των ασύρματων παροχέων να ικανοποιήσουν τις αυστηρές απαιτήσεις ποιότητας υπηρεσίας των χρηστών.

Για την παροχή της απαιτούμενης ποιότητας υπηρεσίας στους χρήστες ενός δικτύου, χρησιμοποιείται ένας μηχανισμός *ελέγχου αποδοχής κλήσης (Call Admission Control)* προκειμένου να ελέγξει τον αριθμό των συνδέσεων που εγκαθίστανται. Πολλοί μηχανισμοί ελέγχου αποδοχής κλήσης έχουν προταθεί στην βιβλιογραφία, τόσο για ενσύρματα όσο και για ασύρματα δίκτυα, και κατηγοριοποιούνται με βάση δύο κριτήρια:

1) Τον τύπο της παρεχόμενης υπηρεσίας από το δίκτυο στον χρήστη και

2) Τον τρόπο που χειρίζονται τις *αλλαγές σταθμού βάσης (handoffs)* και την πολιτική προτεραιοτήτων που χρησιμοποιούν.

Ο *μηχανισμός αποδοχής κλήσης* που προτείνουμε σε αυτήν την εργασία ανήκει στην κατηγορία των *probabilistic service μηχανισμών*, όσων αφορά το πρώτο κριτήριο, αφού χρησιμοποιεί κάποια *προϋπολογισμένα σενάρια κίνησης* λαμβάνοντας υπόψιν τις παραμέτρους κίνησης που δηλώνει ο χρήστης (μέσος ρυθμός, ανώτατος ρυθμός, τυπική απόκλιση). Ο μηχανισμός μας χρησιμοποιεί ένα μοντέλο που προσομοιώνει με ακρίβεια την κίνηση H.263 videoconference, ώστε να αναγνωρίσει και να «ταυτοποιήσει» κάθε καινούργια σύνδεση που ζητάει να εισέλθει στο σύστημα με κάποιο σύνολο παραμέτρων για το οποίο έχει προϋπολογίσει το σενάριο κίνησης. Βάσει του προϋπολογισμένου σεναρίου λαμβάνεται η απόφαση αποδοχής ή μη της νέας κλήσης.

Με βάση το δεύτερο κριτήριο ο μηχανισμός μας ανήκει στην κατηγορία *queuing priority,* αφού τόσο οι συνδέσεις με ενεργή κλήση που έρχονται από άλλη κυψέλη (handoff calls), όσο και οι νέες κλήσεις που προέρχονται από μέσα από την κυψέλη γίνονται αποδεκτές αν υπάρχει αρκετό διαθέσιμο εύρος ζωνης, ενώ κανένα μέρος του έυρους ζώνης δεν κρατείται για αποκλειστική πρόσβαση κανενός από τους δύο τύπους κλήσης.

Ο μηχανισμός μας συγκρίνεται με αρκετούς άλλους μηχανισμούς της βιβλιογραφίας και εμφανίζεται να υπερέχει, τόσο σε αποτελέσματα όσο και σε επίπεδο ιδεών.

ΠΕΡΙΕΧΟΜΕΝΑ

Chapters **Σελ.**

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 QoS requirements. These challenges are further exasperated 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 as well as 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 towards 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 while 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, and especially videoconference traffic. 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 rate control algorithms and predicting the quality of service 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], 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 for its decision-makin*g. This will be further explained below.

Call Admission Control 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.

A. 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, 6] as *guaranteed service*. Admission control algorithms for 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. Network utilization under this model is usually acceptable when flows are smooth; when flows are bursty, however, guaranteed service inevitably results in very low utilization [7].
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]. More specifically, in [10], the equivalent bandwidth of a set of flows is defined as the bandwidth C(ϵ) which is such that the stationary bandwidth requirement of the set of flows exceeds this value with probability at most ϵ, where ϵ is the packet loss rate. However, in the case of an absence of a very accurate statistical model for each individual flow the use of the effective bandwidth allotment leads to a significant overestimation of the sources' actual bandwidth requirements and therefore to a conservative CAC scheme, which fails to use efficiently all the available bandwidth, as shown in [11-15]. The approximate formula presented in [10] for effective bandwidth allotment to a source was used in a previous work from our group [16], which investigated the possibility of using an "effective bandwidth"-based probabilistic service CAC for videoconference traffic originating from latest technology encoders. Our results in [16] (in which no handoff traffic was considered, as in this work) have shown once again that the CAC which was based on the effective bandwidth idea was very conservative and provided much worse bandwidth utilization (throughput) results than a Traffic Policing mechanism which we also proposed in [16]. Since Traffic Policing is not enough on its own, as it can only amend problems caused by network overload and not prevent them, this was our motivation to adopt a completely different approach for Call Admission Control in the present work.

3. In the case of applications which are tolerant of occasional delay bound violations, a third service type was proposed in [4], 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 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 dissertation generally falls within the second category, 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, standard deviation) which the video source will declare at call setup. A logical assumption (adopted in our work) 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 use in our work each user's declared set of parameters in order to examine the respective precomputed traffic scenario which is based on our H.263 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 video traffic).

B. 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 schemes, some channels are reserved for handoff calls. Four different types of CAC schemes have appeared in the literature, namely:
    a. The *cutoff priority scheme*, in which 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].

   b. The *fractional guard channel scheme*, in which a new call is admitted with a certain probability, which depends on the number of busy channels [19].
   c. The *rigid division-based scheme* [20], in which 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.
   d. The *new call bounding scheme*, in which a threshold is enforced on the number of new calls accepted into the cell.
2. *Queuing Priority (QP) schemes*: In this type of schemes, calls are accepted whenever there are free channels. Depending on the approach, either new calls are blocked and handoff calls are queued [21], or vice versa [22], or all calls are queued and the queue is rearranged based on certain priorities [23].

The work presented in this dissertation falls conceptually within the second category, of the *queuing priority schemes*, as in our CAC scheme 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.

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 we use in our study (from [35, 36]). Section 5 contains the description of our CAC scheme, as well as its conceptual comparison with quite a few CAC schemes of the literature. Section 6 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 7 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 centre (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 dissertation 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 ([16, 34]) and accommodates 256 information slots. The channel rate is 9.045 Mb/s (from [34]). Each information slot accommodates exactly one fixed-length packet of ATM size[1] that contains information (video and data information, in our scheme) and a header. In the Multiple Access Control (MAC) schemes introduced in [41, 42] it has been shown that the use, in the uplink (and respectively, in the downlink for acknowledgements) of a small portion of the channel bandwidth (less than 3%) for requests by terminals which wish to acquire slots to transmit, is usually sufficient for high system performance. For this reason, we assume here that 6 of the 256 slots of the channel frame are used by the BS for transmitting acknowledgements 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.

Fading, in a wireless channel, is highly-varying with time, spatial dependencies and interference. We assume that the effect of fading can be mitigated by rich-function transmission/reception wireless sub-systems [37, 40].

[1] The choice of a different packet size would not alter the idea on which our mechanism is based, and therefore would have no qualitative effect in our results.

3. Videoconference Traffic Model

A. H.263 streams

H.263 is a video standard that can be used for compressing the moving picture component of audio-visual services at low bit rates. It adopts the idea of PB frames, i.e., 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].

B. The distribution fit for a single source

Our work in [3], 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, 26] the authors 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); the authors use this fact to build a Discrete Autoregressive (DAR) model of order one, which works well when several sources are multiplexed.

In [3], we have used five different long sequences of H.263 encoded videos (from [27, 28]) with low or moderate motion, in order to derive a statistical model which 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 (we use packets of ATM cell size throughout this work, but our mechanism can be used equally well with packets of other sizes). 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.
2. A video stream extracted and analyzed from a camera showing a lecture.
3. A video stream extracted and analyzed from a parking security camera.
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 which new video frames arrive every 80 msecs, in QCIF resolution. The mean, peak and standard deviation of the video frame 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 distributions investigated. 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. The mean and variance are given by the following expressions:

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

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

Table 1. Trace Statistics.

The sets of parameters of these five traces comprise the “modes” adopted by videoconference users in our study.

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 Figures 1-2, we have plotted the 0.03-, 0.06-, 0.09-,… quantiles of the actual trace versus the respective quantiles of the various distribution fits for two of the five traces under study (all values presented in the axes are in bytes).

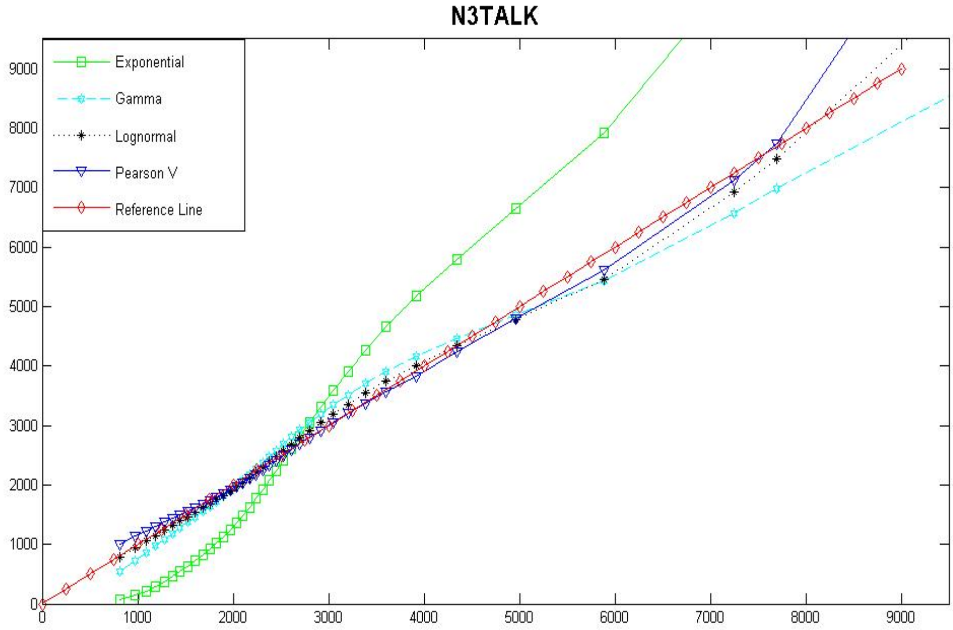

Figure 1. Q-Q plot for the N3 TALK trace.

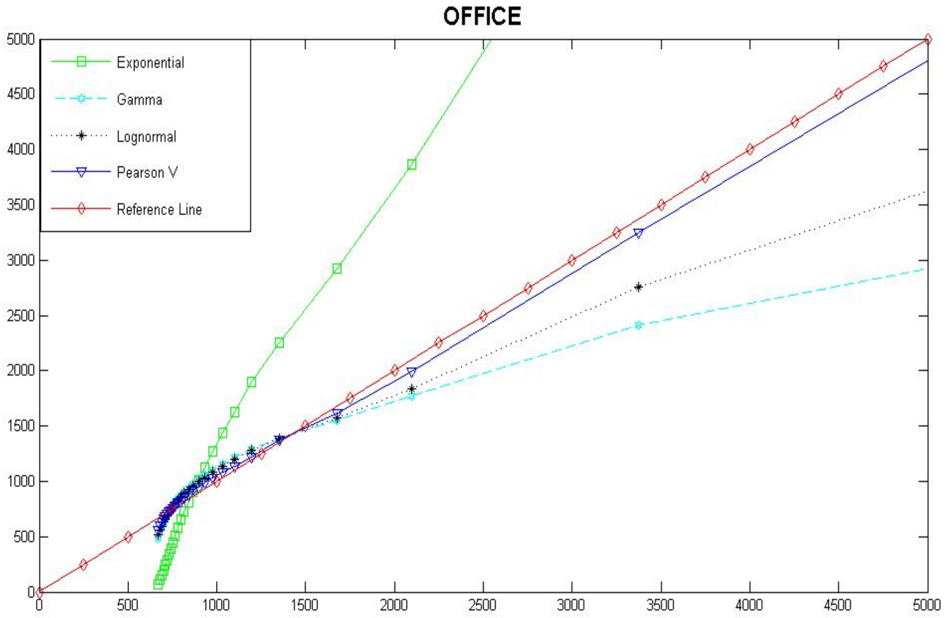

Figure 2. Q-Q plot for the Office Cam trace.

The common characteristic of both Figures is that the Pearson V distribution fit is the best in comparison to the gamma, lognormal and exponential distributions, which are presented here (comparisons were also made with the negative binomial 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, i.e., 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 generating independently 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, 26], and for this reason we developed a Discrete Autoregressive Model of order one.

C. The DAR(1) Model

A Discrete Autoregressive 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 1-ρ 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: $X_n = V_n X_{n-1} + (1- V_n) Y_n$, is a DAR(1) process.

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

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

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

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

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

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

As in [25, 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. Figures 3 and 4 present a comparison between our DAR(1) model and the actual traces of the office camera and ARD Talk sequence, for a superposition of traces. Figures 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, (e.g., 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 (lower than 0.05) for both the actual traces and the DAR(1) model of each trace).

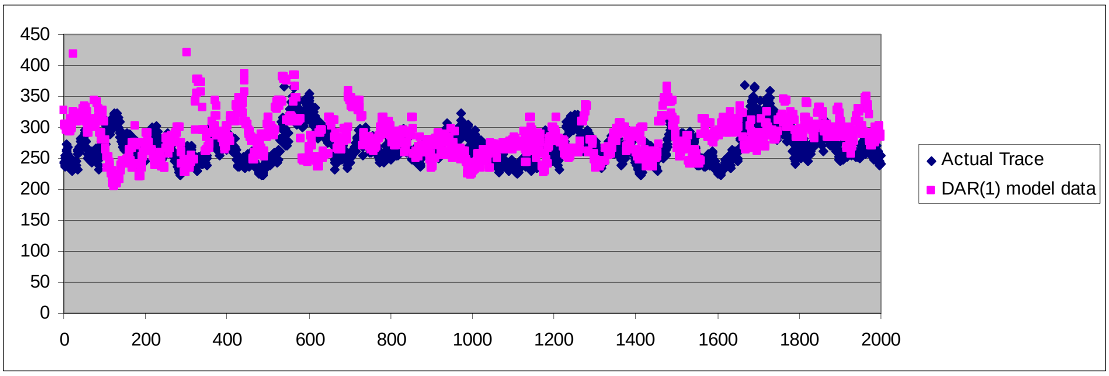

Figure 3. Comparison between the actual office camera trace and the DAR(1) model in number of cells/frame (Y-axis) for a 2000 frame trace for 15 superposed sources.

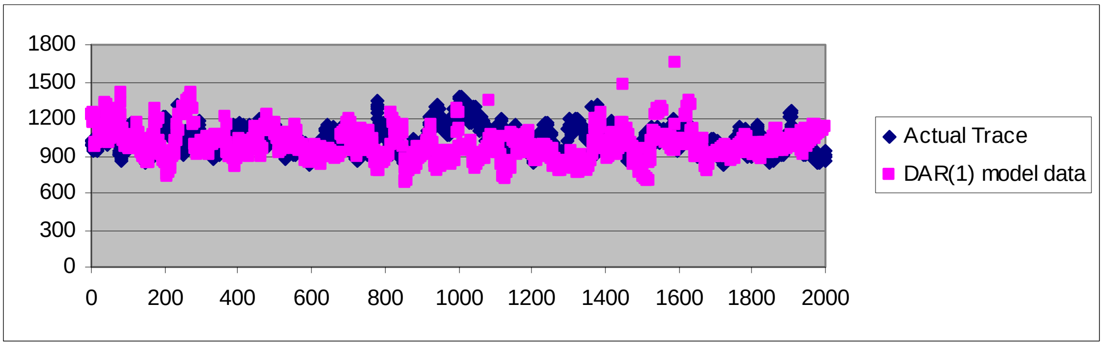

Figure 4. Comparison between the actual ARD Talk trace and the DAR(1) model in number of cells/frame (Y-axis) for a 2000 frame trace for 20 superposed sources.

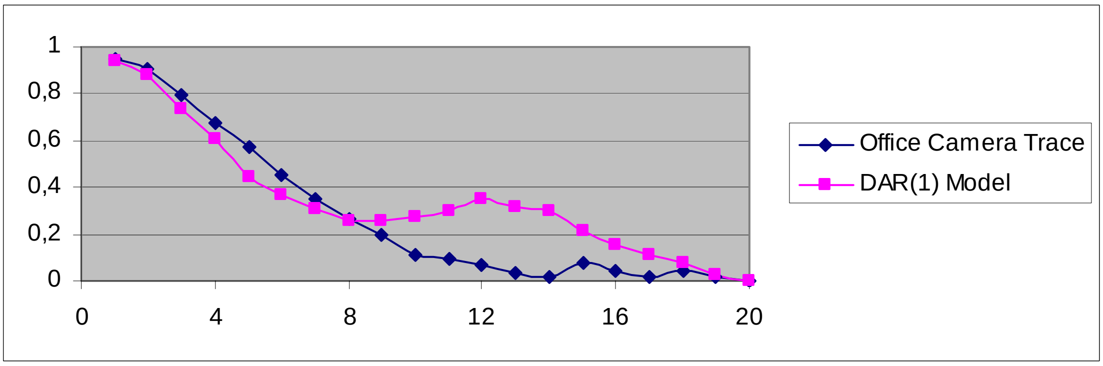

Figure 5. Autocorrelation vs. number of lags for the actual office camera trace and the DAR(1) model, for 15 superposed sources.

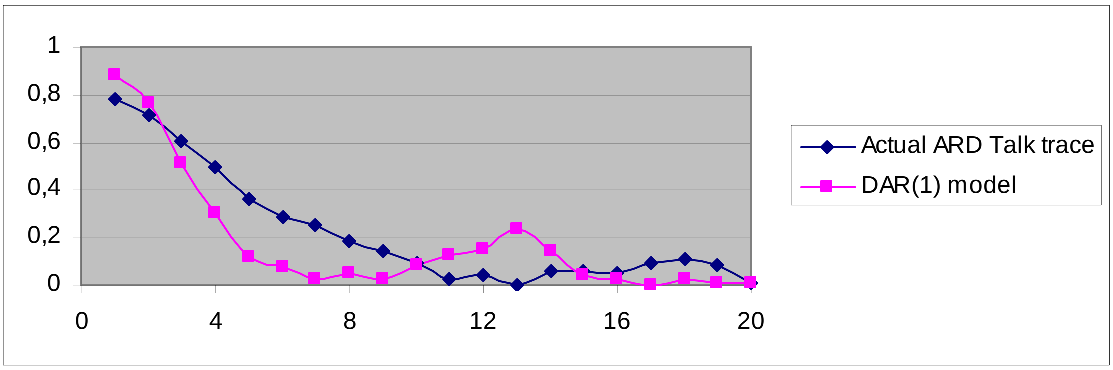

Figure 6. Autocorrelation vs. number of lags for the actual ARD Talk trace and the DAR(1) model, for 20 superposed sources.

Finally, although Figures 3-6 suggest that the DAR(1) model captures very well 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 used again Q-Q plots.

In Figures 7 and 8, we have plotted the 0.025-, 0.05-, 0.075-,… quantiles of the actual office camera trace and ARD Talk trace versus the respective quantiles of their DAR(1) model for superpositions of 15 and 20 traces, respectively. As shown in the Figures, the points of the Q-Q plot fall almost completely along the 45-degree reference line, with the exception of the last 2.5% quantile (right-hand tail), for which the 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.

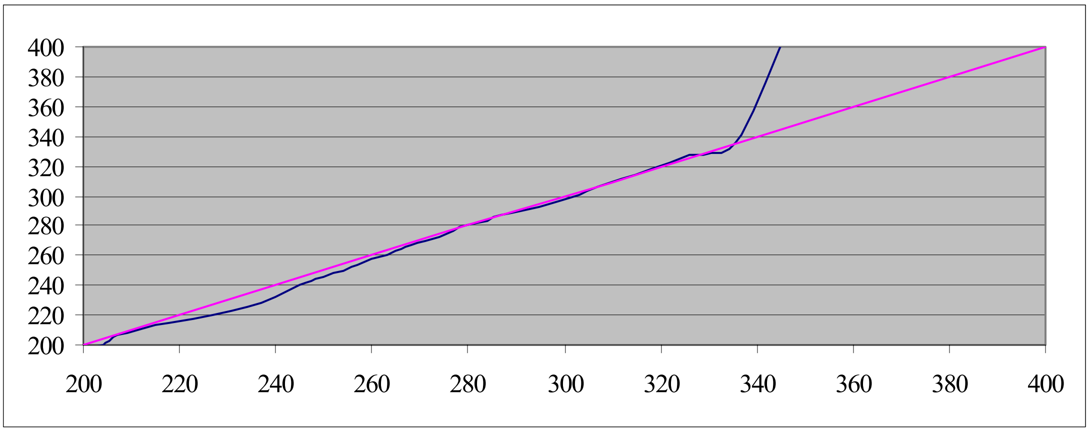

Figure 7. Q-Q plot of DAR(1) model versus the actual office camera trace for 15 superposed sources.

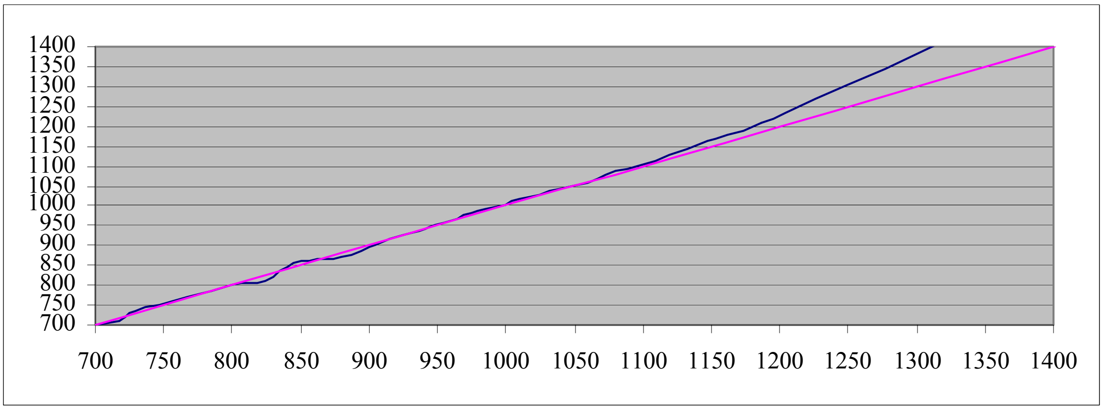

Figure 8. Q-Q plot of DAR(1) model versus the actual ARD Talk trace for 20 superposed sources.

4. Data Traffic Model

We adopt the web traffic model presented in [34, 35], in which the distributions of the random variables concerning the composition of web requests are the following:

- Size of main object: lognormal(1.31,1.41), with mean=10 KB and standard deviation=25KB.
- Number of inline objects: gamma(0.24,23.42), with mean=5.55 and standard deviation=11.4.
- Size of inline object: lognormal(-0.75,2.36), with mean=7.7 KB and standard deviation=126 KB.

- Number of web requests per www session: lognormal(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 50KB. The arrival process of web sessions is assumed Poisson with rate λ sessions per second. We do not simulate, in our adopted data model, the web request viewing time, 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 off-line, in order to check our system's performance under very bursty data load.

5. The Call Admission Control Mechanism

A. Description of our CAC Scheme and its Implementations

The idea of our proposed Call Admission Control mechanism came as a result of the combination of:

a. The quality of our video traffic model, and

b. 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, that is, scene changes do not occur often and this is the reason for the very high autocorrelation of lag-1, which was shown in Figures 5, 6.

Therefore, we were led to investigate the possibility of *precomputing* a large number of videoconference traffic scenarios (i.e., 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 Call Admission Controller 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. By using the "mode" (peak, mean, 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 Call Admission Controller simply checks in its precomputed scenarios' list, whether the admittance of a new videoconference call with the specific characteristics is possible or not. 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.

A second implementation case was the one in which data (web) traffic was integrated with videoconference traffic in the system. In this case our Call Admission Control 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 checking whether the new estimated bandwidth is higher than the channel information rate; also, 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, the system can allow or prevent this behavior, depending on whether the additional bandwidth 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 mechanism. The development of such a mechanism, however, is beyond the scope of this work, which is focused on Call Admission Control.

The third implementation case of our algorithm was the most complicated one, 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 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 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. 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 Queuing Priority scheme.

As it will be shown from our results, in Section 6, our scheme provides almost optimal Call Admission Control, 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.

B. Conceptual Comparison with other CAC Approaches

In this Section, we proceed to make a conceptual comparison between our scheme and the Call Admission Control approaches which 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; 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* Call Admission Controllers. 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% [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 schemes, of the "equivalent bandwidth" method for estimating the aggregate bandwidth (as, for example, in [10-16, 32, 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. An efficient CAC scheme which also generally falls in this category was proposed in [37]; however, this scheme is based on the assumption that for each source a *target bandwidth* is set, which the network aims at providing, and the way of estimating this target bandwidth is not specified, as it is assumed that all calls take varying bandwidth values from a set B, which contains integer multiples of a basic bandwidth unit. Therefore, for a source to choose its target bandwidth from B, an "equivalent bandwidth" estimation would have to be used again.

Regarding the *predictive* service and the measurement-based admission control approach with which it is mainly realized [4, 38], the basic problem of the method was stated in Section 1: this approach is very efficient only in the context of relaxed service commitments of the network to the user. 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 (e.g., 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 has 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 which 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 which is embodied in all predictive service Call Admission Control schemes; in their “conditional prediction of traffic envelope”, the authors point out that only when the peak rate of the aggregate traffic flow is a correlated Gaussian process can the mean rate be optimally predicted. 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 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 Guard Channel scheme approach (e.g., [1, 17-20]). The reasons for this are: a) 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 b) our scheme's almost optimal accuracy, due to the accuracy of our videoconference traffic model.

One key issue which 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 (e.g., 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:

a. 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 the latter our group is also currently completing the development a traffic model).

b. 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. Therefore, we believe that it is actually not necessary to adopt the same approach (i.e., 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 which will be much less conservative than the guaranteed service and equivalent bandwidth probabilistic service approaches, and much less risky than the predictive service approach. The remaining types of flows could be handled with the same approach used for web traffic in our study, or with any other of the many efficient approaches proposed in the literature.

6. Results and Discussion

As analyzed in Section 5, 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:

a. For videoconference traffic, the set maximum allowed video packet dropping rate is 0.01% [34] and the maximum 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 H.263-encoded movies is not constant; it is an integer multiple of 40 ms).

b. 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 off-line, in order to check our system's performance under very bursty data load. As it will be 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.

A. 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:

Scenarios 1-7. Each one of the five "modes" used in our study is used with a 20% probability (i.e., a user which "wakes up" chooses one of the five "modes" with a probability equal to 20%). The number of each scenario corresponds to the number of traces using each mode, e.g., in scenario no. 4 there are 20 videoconference traces present in the system.

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

Scenario 8. ARD Talk mode 20%, Lecture mode 30%, Office mode 25%, N3 Talk mode 20%, Parking mode 10%.

Scenario 9. ARD Talk mode 15%, Lecture mode 25%, Office mode 30%, N3 Talk mode 20%, Parking mode 10%.

Scenario 10. ARD Talk mode 20%, Lecture mode 10%, Office mode 15%, N3 Talk mode 25%, Parking mode 30%.

Scenario 11. ARD Talk mode 25%, Lecture mode 20%, Office mode 30%, N3 Talk mode 10%, Parking mode 15%.

Scenario 12. ARD Talk mode 30%, Lecture mode 15%, Office mode 25%, N3 Talk mode 10%, Parking mode 20%.

Scenario 13. ARD Talk mode 30%, Lecture mode 10%, Office mode 25%, N3 Talk mode 20%, Parking mode 15%.

Scenario 14. ARD Talk mode 25%, Lecture mode 10%, Office mode 30%, N3 Talk mode 20%, Parking mode 15%.

Scenario 15. ARD Talk mode 10%, Lecture mode 30%, Office mode 15%, N3 Talk mode 20%, Parking mode 25%.

Scenario 16. ARD Talk mode 25%, Lecture mode 10%, Office mode 15%, N3 Talk mode 30%, Parking mode 20%.

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 1, the equivalent bandwidth of a set of flows is defined in [10] as the bandwidth C(ϵ) which is such that the stationary bandwidth requirement of the set of flows exceeds this value with probability at most ϵ, where ϵ is the packet loss rate (0.01% in our study). In order to investigate our mechanism's performance, we compare, in Table 2, 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 one hour of network operation.

Scenario	Real Traces-Bandwidth (Mbps)	DAR Model-Bandwidth (Mbps)	Equivalent Bandwidth Estimation (Mbps)	Real Traces-Dropped Packets (%)	DAR Model-Dropped Packets (%)
1	1.01	1.004	1.261	0	0
2	2.016	2.002	2.68	0	0
3	3.028	3.005	4.201	0	0
4	4.045	4.009	5.803	0	0
5	5.061	5.011	7.478	0	0
6	6.071	6.006	9.215	0.000496	0.000132
7	7.093	7.009	11.01	0.052	0.0176
8	5.319	5.245	7.977	0	0
9	5.355	5.297	8.08	0	0
10	7.217	7.16	11.223	0.11	0.028
11	5.765	5.703	8.52	0.000005	0.000019
12	6.163	6.093	9.202	0.000021	0.0004
13	6.565	6.495	10.305	0.00708	0.00319
14	6.398	6.333	9.934	0.00286	0.00169
15	5.966	5.927	8.702	0.000193	0.000032
16	7.142	7.071	11.549	0.099	0.03

Table 2. First Algorithm Implementation-Results Comparison.

It is clear, from the results presented in Table 2, 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 25.6% to 63.3%); 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 0.65% to a maximum of 1.4%. 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 the required bandwidth is 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 actual packet dropping when the real traces are used, in none of the scenarios under study did the packet dropping of the real traces exceed the upped bound of 0.01% without the same result taking place also for our scheme (e.g., scenarios 7, 10, and 16, are shown to contain an overload of traffic, with which the system cannot cope without excessive packet dropping).

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 4 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%, the equivalent bandwidth estimation method predicts that up to 25 users can enter the system without a violation of users' QoS requirements, whereas our scheme is accurate in computing that up to 30 users can enter the system without QoS degradation. The overestimation of the equivalent bandwidth formula provides, of course, the gain of zero packet dropping, as a very significant amount of channel bandwidth is left unused; however, this is an insignificant gain in comparison to the bandwidth loss caused by this method, especially since the upper bound of 0.01% on video packet dropping, which our scheme adheres to, is already quite strict.

Also, it is easy to understand that the use of the equivalent bandwidth estimation is a bad choice even for the rest of the scenarios under study (the ones in which this estimation does not forbid the acceptance of all requesting users), as a Call Admission Controller based on this estimation would be misled into accepting a significantly smaller number of calls from other types of sources (e.g., data sources), as it would assume that a much larger portion of the available bandwidth is being used by videoconference users than the portion which actually would.

Based on the above, the proposed precomputation is shown to be a solid basis for admitting or rejecting a new videoconference call.

The reason for which we have chosen, in scenarios 8-16, the total number of videoconference users in the system to be 31 is that, as shown from Scenarios 6 and 7, the maximum number of traces that the system can accommodate in the "equal probability" case is 30; we decided to keep this number constant, and experiment with the varying probabilities used for each mode in scenarios 8-16. The reason for the additional 31st user is that the 15% and 25% probabilities, out of a total of 30, do not produce an integer number of traces as a result.

Finally, the results presented in Table 2 show that in both the cases of real traces and DAR models the maximum system throughput is achieved in Scenario 13, and the respective throughputs are 72.58% and 71.81%.

B. 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 in this case results with the use of the equivalent bandwidth estimation formula, as it has been shown in Section 6A 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 Call Admission Control scheme in this case has been outlined in Section 5. 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 3 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 3 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 2 that the videoconference traffic generated in these scenarios cannot be accommodated by our system. Web users, in all the results of Table 3, adhere to their declared mean download rate.

Scenario	**Real Traces**			**DAR Model**		
	Maximum web session arrival rate (sessions/sec)	**Mean Session Delay (minutes)**	**Bandwidth (Mbps)**	**Maximum web session arrival rate (sessions/sec)**	**Mean Session Delay (minutes)**	**Bandwidth (Mbps)**
1	0.044	6.24	7.057	0.045	4.24	7.276
2	0.041	7.44	7.694	0.038	3.68	7.307
3	0.031	3.62	7.361	0.031	3.38	7.254
4	0.026	2.71	7.587	0.026	3.32	7.546
5	0.022	4.21	8.063	0.018	3.21	7.541
6	0.013	2.51	7.835	0.013	2.52	7.811
7	-	-	-	-	-	-
8	0.018	3.58	7.782	0.019	2.38	7.829
9	0.021	3.96	8.176	0.018	3.79	7.837
10	-	-	-	-	-	-
11	0.016	2.94	7.909	0.016	4.81	7.901
12	0.013	2.31	7.942	0.013	2.61	7.872
13	0.012	2.24	8.251	0.012	5.35	8.226
14	0.012	2.49	8.049	0.012	6.21	8.079
15	0.015	4.01	7.984	0.015	4.21	7.985
16	-	-	-	-	-	-

Table 3. Second Algorithm Implementation-Results Comparison.

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 from Table 3 that our mechanism provides in some cases an overestimation and in some 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, when our mechanism provided for all studied scenarios a small underestimation of the real traces' bandwidth requirements; this combination is also responsible for the "fluctuations" in the mean session delay, shown in Table 3. 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% to a maximum of 6.47%, and averaging at just 1.7% over all the studied scenarios; from the 13 Scenarios for which results are presented in Table 3, in only 3 Scenarios the difference between our results and the results with the use of the actual traces exceeds 1.5% (again, the large deviation in all the models used for the simulation of the web downloads are responsible for this result), while in 8 of the 13 Scenarios the respective difference is well below 1%.

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%; given the average web request size of about 50KB, the high mean number of pages per web session (25), the very high standard deviation (100) and the presence of very bursty video traffic in the system, we consider this to be a moderate upper bound.

As shown in Table 3, 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 3, in both the cases of real traces and DAR models the maximum system throughput is achieved again (as in Section 6A) in Scenario 13, and the respective throughputs are 91.22% and 90.95%.

C. Third Algorithm Implementation

As explained in Section 5, 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 hand-offed 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 one 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 dissertation). This choice has been made after careful observation of the traces' behaviour 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, the specific Quality with which the user enters the system is again determined with equal probability.

In the results presented in Table 4, we have studied our scheme's performance for Scenarios 6, 8, 9, and 11-15. As in Section 6B, no results are presented for Scenarios 7, 10 and 16, as it has been shown from the results in Table 2 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 6A 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 4 presents the comparison of the bandwidth utilization results from the use of the actual traces with the respective results from our mechanism, when 5%, 10%, 15%, 20% and 25% of the total videoconference traffic in the system is generated from handoffs (e.g., in Scenario 6, for the 25% handoff case, 30 videoconference calls are originating from within the cell under study and 10 more calls originate from handoffs).

Scenario	**Real Traces - Bandwidth (Mbps) under various Handoff Loads (%)**					**DAR Model - Bandwidth (Mbps) under various Handoff Loads (%)**				
	5%	**10%**	**15%**	**20%**	**25%**	**5%**	**10%**	**15%**	**20%**	**25%**
6	6.329	6.282	6.574	6.654	6.641	6.288	6.346	6.512	6.646	6.603
8	5.534	5.649	5.943	6.175	6.296	5.501	5.607	5.897	6.147	6.239
9	5.589	5.706	5.876	6.057	6.295	5.549	5.633	5.836	6.013	6.245
11	5.919	6.107	6.231	6.535	6.527	5.881	6.072	6.185	6.481	6.477
12	6.309	6.526	6.572	6.787	6.836	6.274	6.484	6.532	6.781	6.778
13	6.686	6.829	6.801	6.851	6.888	6.638	6.801	6.813	6.832	6.878
14	6.578	6.705	6.749	6.829	6.859	6.554	6.674	6.758	6.841	6.836
15	6.154	6.387	6.459	6.708	6.663	6.112	6.343	6.431	6.669	6.625

Table 4. Third Algorithm Implementation-Results Comparison.

The results presented in Table 4 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% to a maximum of 1.28%, and averages at a mere 0.59% over all the studied scenarios.

As shown in Table 4, 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 25% handoff traffic. This is expected, as in this case a larger number of video traces is 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% and 76.04%.

All the above results have been produced for an upper bound of 0.01% for video packet dropping. Also, as in the results presented in Table 2, the video packet dropping rates for all the cases studied and presented in Table 4 were very similar when real traces and DAR models were used in the same traffic scenarios.

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 6B, 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 web downloads originate from handoffs or from within the cell under study.

However, there is always the case that users' requirements may exceed in variability the predicted bandwidth requirements variability 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" which 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, i.e., 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, 28], the parameters of which are shown in Table 5.

Movie	Mean (bytes)	Peak (bytes)	Standard Deviation (bytes)
Boulevard Bio	3001	11643	1616
ARD News	3442	15310	2649

Table 5. "Unknown Modes" Statistics.

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 hand-offed traffic. The "mode" selected in each case was the one with the smallest mean MSE from the hand-offed 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 most similar to the "N3 Talk" mode.

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

Scenario	Real Traces - Bandwidth (Mbps) under various Handoff Loads (%)	DAR Model - Bandwidth (Mbps) under various Handoff Loads (%)
6	**6.612**	**6.617**
8	**6.181**	**6.078**
9	**6.204**	**6.145**
11	**6.417**	**6.402**
12	**6.614**	**6.691**
13	**6.712**	**6.779**
14	**6.708**	**6.802**
15	**6.703**	**6.686**

Table 6. Third Algorithm Implementation-Results Comparison for Handoff Traffic from "unknown modes".

The results presented in Table 6 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% to a maximum of 1.67%, and averages at only 0.84% 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 5) 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 $3*10^{-5}$ on average (this result, combined with the very accurate prediction by our mechanism of the actual bandwidth requirements of the superposition of traces, means that our scheme predicts very accurately the bandwidth requirements of the unknown modes, but slightly less accurately their burstiness). Although the video packet dropping underestimation is very small in quantitative terms, it is significant in quality, since the upper bound for video packet dropping is 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%) 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 (e.g., with the use of 10 modes, i.e., 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 5, 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.

7. Conclusions

Based on an accurate videoconference traffic model which has been developed by our group, we have proposed in this work a new efficient Call Admission Control scheme for multimedia traffic transmission over wireless networks. The novelty of the scheme lies in the utilization of precomputed traffic scenarios 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 Call Admission Control approaches.

References

1. J. Misic, T. Y. Bun, “Adaptive admission control in wireless multimedia networks under nonuniform traffic conditions”, *IEEE Journal on Selected Areas in Communications,* Vol. 18, No. 11, 2000, pp. 2429-2442.
2. D. M. Lucantoni, M. F. Neuts, and A. R. Reibman, “Methods for performance evaluation of VBR video traffic models,” *IEEE/ACM Trans. Networking,* Vol. 2, No. 2, 1994, pp. 176-180.
3. P. Koutsakis, “On modeling VBR videoconferencing traffic from H.263 video coders”, submitted for publication.
4. S. Jamin, P. B. Danzig, S. J. Shenker and L. Zhang, “A measurement-based admission control algorithm for integrated service packet networks”, *IEEE/ACM Trans. Networking,* Vol. 5, No. 1, 1997, pp. 56-70.
5. D. D. Clark, S. J. Shenker and L. Zhang, “Supporting real-time applications in an integrated services packet network: architecture and mechanism”, *in Proc. of the ACM SIGCOMM*, 1992, Baltimore, USA, pp. 14-26.
6. D. Ferrari and D. C. Verma, “A scheme for real-time channel establishment in wide-area networks”, *IEEE Journal on Selected Areas in Communications,* Vol. 8, No. 3, 1990, pp. 368-379.
7. H. Zhang and D. Ferrari, “Improving utilization for deterministic service in multimedia communication”, *in Proc. of the IEEE International Conference on Multimedia Computing and Systems*, 1994, Boston, USA.

8. H. Zhang and E. W. Knightly, "Providing end-to-end statistical performance guarantee with bounding interval dependent stochastic models", *in Proc. of the ACM SIGMETRICS*, 1994, Nashville, USA.
9. H. Saito and K. Shiomoto, "Dynamic call admission control in ATM networks", *IEEE Journal on Selected Areas in Communications,* Vol. 9, No. 7, 1991, pp. 982-989.
10. W. Verbiest, L. Pinoo and B. Voeten, "The impact of the ATM concept on video coding", *IEEE Journal on Selected Areas in Communications,* Vol. SAC-6, 1988, pp. 1623-1632.
11. R. Guerin, H. Ahmadi and M. Naghsineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks", *IEEE Journal on Selected Areas in Communications,* Vol. 9, No. 7, 1991, pp. 968-981.
12. R. Guerin and L. Gun, "A Unified Approach to Bandwidth Allocation and Access Control in Fast Packet-Switched Networks", *in Proceedings of the IEEE Infocom' 92, pp.1-12, Florence, Italy,* 1992.
13. M. Naghshineh, and R. Guerin, "Fixed versus variable packet sizes in fast packet-switched networks", *in Proceedings of the IEEE Infocom' 93,* pp.217-226, San Francisco, USA, March 1993.
14. E. Gelenbe, X. Mang and R. Onvural, "Diffusion Based Call Admission Control in ATM", *Performance Evaluation,* Vol. 27&28, pp. 411-436, 1996.
15. J. Q.-J. Chak and W. Zhuang, "Capacity Analysis for Connection Admission Control in Indoor Multimedia CDMA Wireless Communications", *Wireless Personal Communications*, Vol. 12, 2000, pp. 269-282.
16. P. Koutsakis and M. Paterakis, "Call Admission Control and Traffic Policing Mechanisms for the Transmission of Videoconference Traffic from MPEG-4 and H.263 Video Coders in Wireless ATM Networks", *IEEE Transactions on Vehicular Technology*, Vol. 53, No.5, 2004, pp. 1525-1530.
17. Y. Fang and Y. Zhang, "Call Admission Control Schemes and Performance Analysis in Wireless Mobile Networks", *IEEE Transactions on Vehicular Technology*, Vol. 51, No.2, 2002, pp. 371-382.
18. D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures", *IEEE Transactions on Vehicular Technology*, Vol. VT-35, No.3, 1986, pp. 77-92.
19. R. Ramjee, D. Towsley and R. Nagarajan, "On optimal call admission control in cellular networks", *Wireless Networks*, Vol. 3, 1997, pp. 29-41.
20. M. D. Kulavaratharasah and A. H. Aghvami, "Teletraffic performance evaluation of microcellular personal communication networks (PCN's) with prioritized handoff procedures", *IEEE Transactions on Vehicular Technology*, Vol. 48, No.1, 1999, pp. 137-152.
21. R. A. Guerin, "Queuing-blocking system with two arrival streams and guard channels", *IEEE Transactions on Communications,* Vol. 36, No.2, 1988, pp. 153-163.

22. E. Del Re, R. Fantacci and G. Giambene, "Handover queuing strategies with dynamic and fixed channel allocation techniques in low earth orbit mobile satellite systems", *IEEE Transactions on Communications,* Vol. 47, No.1, 1999, pp. 89-102.
23. C. Chang, C. J. Chang and K. R. Lo, "Analysis of a hierarchical cellular system with reneging and dropping for waiting new calls and handoff calls", *IEEE Transactions on Vehicular Technology,* Vol. 48, No.4, 1999, pp. 1080-1091.
24. ITU-T Recommendation, H.263, 3/1996.
25. 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,* Vol. 2, No. 1, 1992, pp. 49-59.
26. D. P. Heyman, T. V. Lakshman, A. Tabatabai, H. Heeke, "Modeling teleconference traffic from VBR video coders", *in Proceedings of the IEEE International Conference on Communications (ICC) 1994,* New Orleans, USA, pp. 1744-1748.
27. [Online] http://www-tkn.ee.tu-berlin.de/research/trace/trace.html
28. F. H. P. Fitzek, M. Reisslein, "MPEG-4 and H.263 video traces for network performance evaluation", *IEEE Network,* Vol. 15, No. 6, 2001, pp. 40-54.
29. A. M. Law, W. D. Kelton, "Simulation modeling & analysis", 2nd Ed., McGraw Hill Inc., 1991.
30. A. Adas, "Traffic models in broadband networks", *IEEE Communications Magazine,* Vol. 35, No.7, 1997, pp. 82-89.
31. P. A. Jacobs and P. A. W. Lewis, "Time series generated by mixtures", *Journal of Time Series Analysis*, Vol. 4, No. 1, pp. 19-36, 1983.
32. Y. Iraqi and R. Boutaba, "A novel distributed call admission control for wireless mobile multimedia networks", *in Proceedings of the Third ACM International Workshop on Wireless Mobile Multimedia (WoWMoM) 2000*, Boston, USA, pp. 21-27.
33. F. Hu and N. K. Sharma, "Priority-determined multiclass handoff scheme with guaranteed mobile QoS in wireless multimedia networks", *IEEE Transactions on Vehicular Technology,* Vol. 53, No. 1, 2004, pp. 118-135.
34. D. A. Dyson and Z. J. Haas, "A Dynamic Packet Reservation Multiple Access Scheme for Wireless ATM", *Mobile Networks and Applications (MONET) Journal,* Vol. 4, No. 2, pp. 87-99, 1999.
35. H.-K. Choi, J. O. Limb, "A Behavioral Model of Web Traffic", *in Proceedings of the Seventh International Conference on Networking Protocols (ICNP)*, Toronto, Canada, 1999, pp. 327-334.
36. P. Tran-Gia, D. Staehle, K. Leibnitz, "Source traffic modeling of wireless applications", *International Journal of Electronics and Communications*, Vol. 55, No. 1, 2001, pp. 27-37.
37. T. Kwon, Y. Choi, C. Bisdikian and M. Naghsineh, "QoS provisioning in wireless/mobile multimedia networks using an adaptive framework", *Wireless Networks*, Vol. 9, 2003, pp. 51-59.

38. J. Qiu and E. W. Knightly, “Measurement-based admission control with aggregate traffic envelopes”, *IEEE/ACM Transactions on Networking*, Vol. 9, No. 2, 2001, pp. 199-210.
39. R. J. Gibbens, F. P. Kelly and P. B. Key, “A decision-based theoretic approach to call admission control in ATM networks”, *IEEE Journal on Selected Areas in Communications,* Vol. 13, No. 6, 1995, pp. 1101-1114.
40. J. Gomez, A. Campbell and H. Morikawa, “A systems approach to prediction, compensation and adaptation in wireless packet networks”, *in Proceedings of the ACM/IEEE International Workshop on Wireless Mobile Multimedia (WoWMoM) 1998*, Dallas, USA, pp. 92-100.
41. N. M. Mitrou, G. L. Lyberopoulos and A. D. Panagopoulou, "Voice and Data Integration in the Air-Interface of a Microcellular Mobile Communication System", *IEEE Transactions on Vehicular Technology*, Vol. 42, No. 1, 1993, pp. 1-13.
42. P. Koutsakis and M. Paterakis, "On Multiple Traffic Type Integration over Wireless TDMA Channels with Adjustable Request Bandwidth", *International Journal of Wireless Information Networks,* Vol. 7, No.2, pp.55-68, 2000.