A Quality of Service Negotiation-Based Admission Control Scheme for WCDMA Mobile Wireless Multiclass Services

Georgios S. Paschos, *Member, IEEE*, Ilias D. Politis, and Stavros A. Kotsopoulos

***Abstract*—This paper introduces a new call admission control scheme. Quality of service (QoS) parameters are negotiated during the admission procedure at heavy loaded cells and maximum capacity is achieved. Blocking and dropping are used to prevent the system from an outage situation but their probability of occurrence is greatly reduced. Reservation bandwidth is used thoroughly as a means of improving overall QoS. Furthermore, the greatest common divisor of all the application load factors is considered as the resource unit, the resulting resource pool is identified and the performance of the above algorithms is analyzed using a novel multistep Markov analysis. The analysis takes account of classes without time restrictions, which are served with a pre-emptive system. Although the increased channel utilization of our proposed scheme reduces the available bandwidth for calls with pre-emption, we compensate for this loss using an extra reservation bandwidth. Blocking probability, dropping probability, average time delay in the queue, and channel utilization factor are used as benchmarks for the proposed schemes.**

***Index Terms*—Call admission control, quality of service (QoS), wideband code division multiple access (WCDMA).**

I. Introduction

MULTIMEDIA applications with various bit rates and QoS requirements have steadily become to dominate the wireless communications environment. Wideband code division multiple access (WCDMA) technology has been established as the main air interface for third-generation (3G) mobile systems. The distinctive features of a WCDMA mobile system are the ability to support multimedia services and guarantee a requested grade of service (GoS). In an attempt to advance to 3G, Third-Generation Partnership Project (3GPP) was founded by five standards organizations ETSI (Europe), Committee T1 (USA), ARIB (Japan), TTC (Japan), and TTA (Korea), and the efforts for 3G standardization were united. In 1999, WCDMA was chosen by 3GPP as the access scheme for UMTS. In addition to UMTS the 3GSM expansion of the worldwide GSM system was announced to use WCDMA as the radio interface as well.

WCDMA is an interference-limited system, where each user is assigned an orthogonal code in order to gain access to the system. The orthogonal scrambling and channelization codes are used in the receiver for separation between the channels. However, along with the desired signal, every other WCDMA signal transmitted in the air interface is received and presented as wideband noise. As such, regardless of the number of codes used, the system can accept a limited number of users before the summed interference make every communication impossible. The situation, where the operation of the system is suspended due to the above or any other reason is called network outage. In this interference constraint system voice calls and data calls compete with each other. Due to mobility, calls are divided to new calls and handover calls, while due to quality of service (QoS) class differentiation, are divided into real-time (RT) calls and nonreal-time (NRT) calls [1]. The extensive categorization of calls is thoroughly described in a later section.

WCDMA technology introduces flexibility in radio resource management (RRM). RRM is the part of the network where the distribution of available resources is decided. It is responsible for the admission and management of call requests. The main part of RRM is the call admission control (CAC). The cell controller uses CAC algorithms in order to decide whether a new call should be accepted or not. A well-designed CAC algorithm should provide the user with the requested QoS as well as make an efficient use of the available capacity and prevent the system from an outage situation due to overloading. Such a QoS-driven admission control strategy protects the network from overloading and congestion, which will cause degradation of the requested QoS. In [2], more information about WCDMA and RRM can be found.

QoS in 3G networks has been studied in [3] and [4] and a simple CAC algorithm has been derived based on multiple access interference (MAI) measurements. In addition to that, a QoS-based CAC scheme for multiclass, multichip CDMA systems that embodies a dynamic nonpreemption area for better channel utilisation has been derived in [5]. CAC algorithms have been proposed that use threshold values for handover bandwidth reservation [6]. In [7], a CAC scheme is presented that includes bandwidth reservation and conditional access. High-priority calls are always accepted whereas low-priority calls are accepted under certain conditions. In [8], the author proposed a new RRM scheme for multichip rate DS/CDMA systems, which includes separate frequency bands for uplink and downlink. Soft handover and interfrequency handover were used to reduce call dropping. The authors in [9] and [10] studied guard channel admission control schemes, including new call bounding, cutoff priority and new call thinning schemes. These schemes were explained through a two-dimensional (2-D0 Markov analysis. Multidimensional Markov chains are used extensively in [11]

Manuscript received October 31, 2003; revised March 4, 2004, June 28, 2004, February 4, 2005. Paper approved by Associate Editor XXX.

The authors are with the Wireless Laboratory in the Telecommunications Department of University of Patras, Greece (e-mail: gpasxos@ee.upatras.gr; ipolitis@ee.upatras.gr; kotsop@ee.upatras.gr).

Digital Object Identifier 10.1109/TVT.2005.853455

and [12] as well. In the first study, a quality-based queuing model was developed for CDMA systems. The work was based on a 2-D Markov analysis. In the second, multidimensional Markov analysis is used to describe channel allocation in a multiclass environment.

CAC schemes for RRM in WCDMA systems have been studied extensively in recent years. In [13], the proposed CAC algorithm allows for higher throughput. A beam forming antenna array was used in this paper. In [14], the admission is granted if the load in the cell and in its neighboring cells is below a threshold value. A capacity based algorithm was proposed in [15], where a threshold value of the capacity is developed and a centralized demand algorithm is implemented. Two voice-based CAC schemes were presented in [16]. The first algorithm makes a decision upon a threshold value of SIR which depends on the propagation environment. The second scheme, denoted as Looking Around, despite its superiority to the first, requires knowledge of the neighboring cells load condition. In [17], the authors proposed an interference based CAC scheme applied to macro and micro layer. The algorithm employs several threshold values of acceptable interference level in order to include the concept of inter-layer handover. In [12], a CAC scheme was presented for integrated RT and NRT calls. The algorithm divides the available channels in the cell into three parts, one for RT calls, one for NRT calls and one for handover calls. The proposed algorithm allows for preemption of NRT data, which are queued in a buffer. Multidimensional Markov chains were used in this system as well.

In this paper, we present an admission scheme for WCDMA mobile systems that includes all the advantages of the already presented schemes and additionally we introduce the concept of QoS negotiation and renegotiation. Multiclass calls with different QoS requirements and assigned priorities are admitted in the system. The scheme achieves the minimization of blocking and dropping while at the same time the channel utilization is maximized. QoS negotiation and renegotiation in a CAC scheme allows the user to benefit from the available resources while the provider has full control of the system.

The rest of this paper is organized as follows. In Section II, several issues about capacity that are important for understanding the operation of CAC algorithm are analyzed. Section III includes a classification of QoS classes that will be used further in the paper. The model of the proposed system is presented in Section IV and is mathematically analyzed in Section V. Finally, the results are showcased in Section VI and the paper is concluded in Section VII.

II. Capacity Issues

The capacity of a WCDMA system is defined by means of interference. Each new admitted user provokes a noise rise. The maximum capacity is achieved when the cumulative interference becomes so great that the energy bit to noise density ratio E_b/N_0 requirements cannot be fulfilled. Further admission will cause network outage. For every application i, the bit energy requirements from [19] are

$$(E_b/N_o)_i = \frac{W}{\upsilon_i R_i} \cdot \frac{P_i}{I_{\text{total}} - P_i} \tag{1}$$

where W is the chip rate, υ_i is the application activity factor, R_i is the application bit rate, P_i the receiver signal power and I_{total} is the total interference received. If we solve for P_i, we see that for each set of requirements we have a specific power level for optimum functionality. More power causes unnecessary interference and less power produces inadequate E_b/N_0.

$$P_i = \frac{1}{1 + \frac{W}{(E_b/N_o)_i \cdot R_i \cdot \upsilon_i}} I_{\text{total}} \tag{2}$$

If we define the load factor of the application i as Lf_i, as in [19]

$$\text{Lf}_i \equiv \frac{P_i}{I_{\text{total}}} = \frac{1}{1 + \frac{W}{(E_b/N_o)_i \cdot R_i \cdot \upsilon_i}}. \tag{3}$$

Note that the load factor expresses the volume of interference incurred to the system by the user. Every QoS class has a specific load factor. Moreover, we have a system outage situation, when

$$\sum_i \text{Lf}_i \geq 1. \tag{4}$$

Following the above rationale, in a system with flexible capacity entities, the load factors can replace the concept of a channel as a resource measure in a fixed capacity system. In this paper, we use load factors as resource indicators.

When calculating the load factor of the cell we must take account other-cell interference as well. Other-cell interference is the interference caused by mobile users of neighbouring cells. In the case of uplink, the other-cell interference depends on the power generated by every mobile user active in the neighbouring cells and their respective positions. These signals constitute a wideband noise for the home cell node-b receiver. In the case of downlink load factor other-cell interference depends on the power transmitted by neighbouring base stations and the position of the user in the home cell. Other cell interference is denoted by I_{other} and the ratio of the other cell interference to P_i is denoted by i_F. In (5) the resulting formula is presented. In this study symmetric channel allocation in the uplink and downlink is assumed

$$\text{Lf}_i \equiv \frac{P_i + I_{\text{other}}}{I_{\text{total}}} = \frac{1 + i_F}{1 + \frac{W}{(E_b/N_o)_i \cdot R_i \cdot \upsilon_i}}. \tag{5}$$

III. QoS Classes Issues

QoS classification is a very important feature of the future mobile networks and UMTS as well. The gain of this characterization is the ability to manage the available resources while accounting for user-QoS. There can be many dimensions of categorization.

As described in [9], four distinct QoS classes are defined in WCDMA specifications; the conversational class, the streaming class, the interactive class and the background class. Additionally, QoS classes can be defined regarding the bit rate used. In

TABLE I
QoS CLASSES AND RELATIVE PARAMETERS

UMTS classes	Conversational	Streaming	Interactive	Background
Time requirements	Real time	Real time	Non- Real time	Non- Real time
Applications	Voice applications with AMR	Video & other streaming applications	Interaction applications & Computer gaming	Browsing & Downloading, Emails, MMS, SMS
Available Bit Rates	16Kbps (x1) 32Kbps (x2)	64Kbps (x4) 128Kbps (x8) 384Kbps (x24) 756Kbps (x48) 2Mbps (x128)	All rates supported	All rates supported
Activity factor	~ 0.67 (two-way call)	1	1	1
Static E_b/N_0	5 dB	1-1.5 dB	1 dB	1 dB
Calculated Lf_i for 16kbps demand	0.00875	0.005214 0.005848	0.005214	0.005214

Table I, the above classification is showcased and the respective characteristics of every class are listed.

As mentioned before, for every QoS class a different load factor can be calculated using the respective values of bit rate R_i, of activity factor u_i and the E_b/N_0 requirement. For reasons of simplicity in the calculations that will follow in a later section, we define $Lf_{\rm GCD}$ as the Greatest Common Divisor of the possible load factors Lf_i of each set of applications. In order to calculate $Lf_{\rm GCD}$, we simply turn all decimals to integers and find the greatest common divisor of the integers. Clearly, the load factor of each QoS class can be expressed as integer multiples i of $Lf_{\rm GCD}$

$$\frac{Lf_i}{Lf_{\rm GCD}} = i \tag{6}$$

where i is a positive integer.

The above definition allows a further QoS classification of the four basic QoS classes described in the table. The new classification is based on the sets of integer multiples i of $Lf_{\rm GCD}$ that are mapped on each QoS class and is used in the particular analysis. These values can be used by the RRM as the application resource requirements.

IV. SYSTEM MODEL

As it has been mentioned in the previous section, the system assumes three QoS classes, voice, RT data and NRT data. The call requests for voice and RT data applications are classified as new call request and handover call request, whereas for NRT data traffic such a classification is not needed. This type of application (background and interactive class) can be delivered in the background, thus the delay may even be minutes while the user does not expect the data within a certain time.

A. QoS Negotiation Scenario

The proposed QoS negotiation scenario is based on the range of allowable bit rates and respective load factors that a user application may have according to the billing policy that agrees on with the network provider. According to the above QoS class analysis, a user may be served the same application with several resource requirements, i.e., a voice call can be delivered with k and $k/2$ multiples of $Lf_{\rm GCD}$ (LFM). Hence, when a user makes a connection request, the system, which knows the range of LFMs that this user has agreed on for the particular application, allows the user to connect with the highest resource requirement so long as there are adequate spare resources in the cell. When the available resources are limited, the system negotiates the LFM with which it will allow the user to connect, within the boundaries of the range of bit rates (and respective LFMs) that the user has selected for the specific application. The QoS negotiation in this case allows a user to make a connection with the QoS requirements preserved within the acceptance range of the user and without causing overload to the system. The amount of gain in the system resources due to this procedure depends on the application and the user negotiation scenario. However, in any case, the resulting blocking probability and the utilization of the system resources is improved significantly as it can be seen in the analytic results later in the paper.

Nevertheless, the QoS negotiation with a user that requests a new or a handover connection, which allows the user to be connected with the lowest LFM accepted by the user, may still cause overload and degradation of the system. In this case, the system initiates a QoS renegotiation procedure with ongoing connections. QoS renegotiation strategy applies the QoS negotiation philosophy on ongoing calls. The ongoing call LFMs are reduced to the lowest accepted by their users, thus an additional resource allowance is gained. In case the re-negotiation attempt provides adequate capacity for the new or handover call request, the admission is granted, otherwise, the request is rejected and the system is fully loaded and negotiated.

B. Admission Control Strategy

A call admission control algorithm decides whether a radio access bearer request should be accepted or rejected. This decision is made upon the impact that this new connection will have on the system. The algorithm estimates the load increase that the new connection will cause, both for uplink and downlink, and admits the request only if the result is acceptable.

In this section, two QoS-negotiation-based CAC algorithms are proposed and presented, denoted by QoS CAC 1 and QoS CAC 2. The first algorithm includes a threshold for the system load factor, denoted by $\tau_H Lf_{\rm GCD}$, where τ_H is the multiples of $Lf_{\rm GCD}$ up to the threshold condition. When the current load factor is less than $\tau_H\ Lf_{\rm GCD}$, both new and handover call requests for voice and RT data traffic are served. On the other hand, when the load factor is greater than $\tau_H Lf_{\rm GCD}$ (the grey area in Fig. 1), only handover call requests are accepted. Therefore, the resource range (τ_H, M) constitutes the reserved handover bandwidth, where M is the total resource pool expressed in $Lf_{\rm GCD}$ multiples. The NRT class resource pool (denoted $M_{\rm NRT}$ here) is variable and defined by the resources that are left unused by the RT traffic. In the example of Fig. 1, where the system is in the condition j, the available bandwidth for RT handover calls is M-j, for RT new calls τ_H-j and for NRT calls M-j.

The NRT requests are queued in a buffer instead of being rejected. In case the system is fully loaded by a combination of

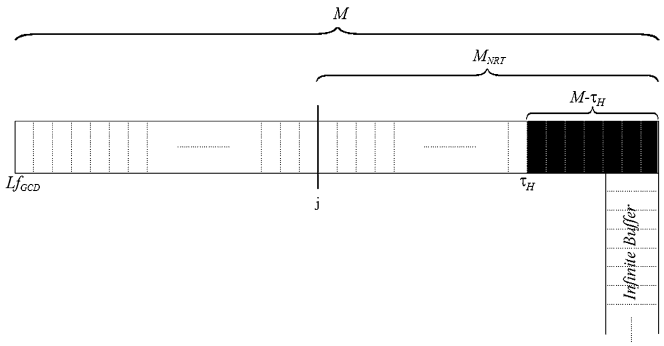


Fig. 1. Resource organization for QoS CAC 1.

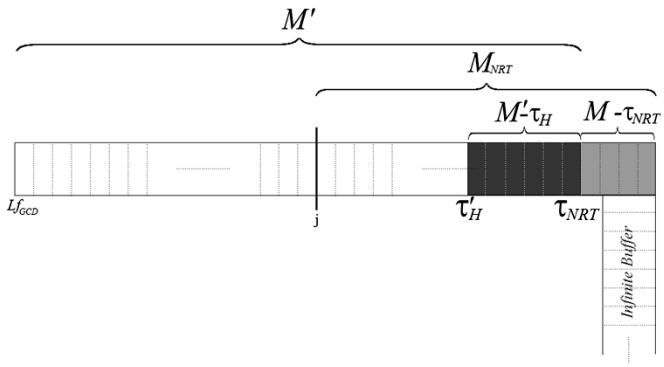


Fig. 2. Resource organization for QoS CAC 2.

applications from both RT and NRT classes, a new RT request preempts as many NRT applications as needed in order to be served. As such, with this algorithm, the total system resources are transparent to RT applications at all times. NRT applications can only use capacity that is not needed by the system.

In 3G mobile systems NRT data traffic is expected to be a popular application between users, since it allows the mobile terminal to become a potential portable personal computer. The most common applications, however, at least in the beginning, are expected to be the short messages (SMS), multimedia messages (MMS), and email [1]. The second proposed algorithm is more QoS oriented and efficient. The small part of resources that the algorithm reserves for NRT data in practice could only offer nothing but a limited increase in the QoS of RT traffic. On the other hand, this bandwidth reservation enables NRT users to have a limited access to the network. Hence, new and handover calls of voice and RT traffic can be served before Lf reaches $\tau_H \mathrm{Lf}_{\mathrm{GCD}}$, only handover call requests are allowed after τ_H and before τ_{NRT}, while NRT traffic is accommodated in whatever bandwidth is reserved after τ_{NRT}, as well as in the spare RT bandwidth. The necessity of this algorithm is imposed by the general concept of our QoS approach. Since the system initially accepts user calls with maximum rates, leaving negotiation for later stages, the available capacity is used long before the cell is fully loaded. Due to maximum channel utilization achieved by this concept, a situation of unfairness arises. Channel resources become unavailable for NRT applications long before the system is fully loaded. Thus the bandwidth reservation for NRT classes is crucial.

In Fig. 2, it is noted that the RT resource pool has become M' which is calculated as $M' = \tau_{\mathrm{NRT}}$. The light gray area is reserved for NRT calls only.

QoS CAC 1 algorithm consists of three strategies, each one corresponding to new, handover or NRT calls. Two of them are shown in Fig. 3.

Case 1) When a new call request of voice or RT data is issued, the system calculates the resulting load factor Lf_T. The call is admitted if the combined load factor for each application i is

$$Lf_T = Lf_P + Lf_N \leq \tau_H Lf_{\mathrm{GCD}} < 1 \tag{7}$$

In the above, Lf_P is the load factor of the system prior to the admittance of the new call. Lf_N is the load factor of the new application call. Thus, if the estimated load factor after

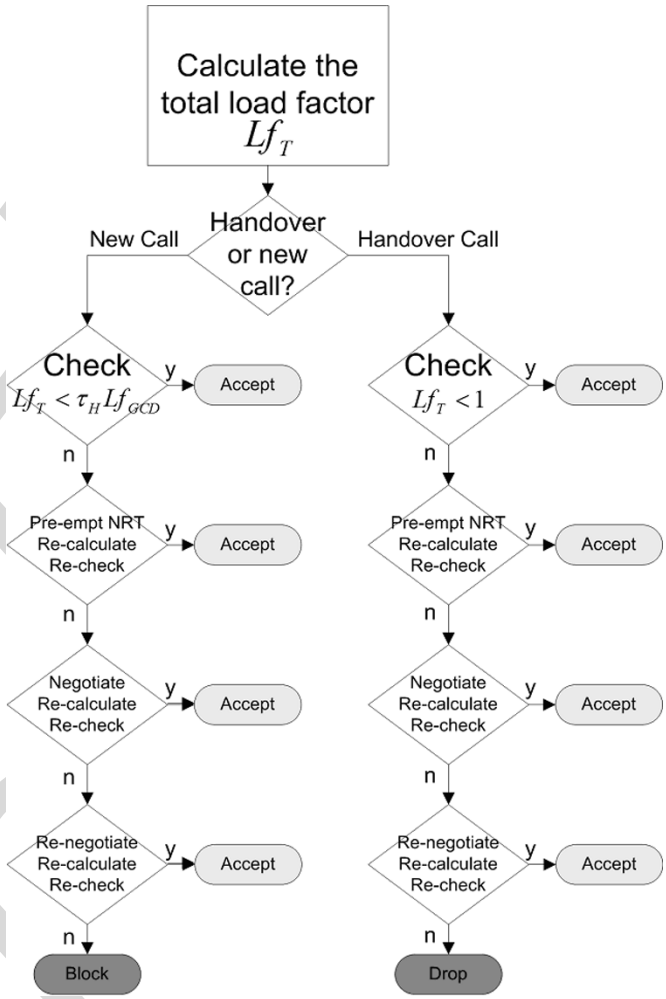


Fig. 3. CAC algorithm logic diagram for QoS CAC 1.

the admittance is larger than $\tau_H Lf_{\mathrm{GCD}}$, the system initiates the NRT preemption procedure.

If the preemption of NRT calls is not enough, the negotiation procedure over the LFM of the new call request is commenced. The new call will be admitted if the estimated load factor in (7) is within the boundaries of τ_H, with $R_{k,i}$ and respective LFM being the lowest in the range that the user has pre-selected for the specific application.

Otherwise, the system relies on the re-negotiation procedure over the LFMs of ongoing calls in the cell. There may be ongoing calls that had been admitted by the system when there were plenty of available resources, without negotiation. Therefore, the system renegotiates the bit rates $R_{k,i}$ of these calls and drops them to the lowest value accepted by the user. Hence, the system load factor decreases and new resources become available for the new call to be admitted. If the resource shortage is not amended, the call request will be blocked.

Case 2) When a handover call request of voice or RT data is issued, the same procedure is followed and the resulting load factor is estimated. In this case, the handover call is admitted if

$$Lf_T = Lf_P + Lf_H < 1 \tag{8}$$

where Lf_H is the load factor of the application that requested handover. Handover calls can be admitted until the system reaches the state of full capacity utilization (8). In this case, the system starts to preempt NRT calls. If the preemption does not solve the resource problem, the system negotiates the handover call over the LFMs as it has been described in Case 1). When, $\sum_i Lf_i \geq 1$, the renegotiation procedure is initiated over the bit rates of ongoing calls, until the load factor of the system after the admittance of handover call satisfies (8). The difference with the first case is that the system can accommodate handover calls in all the available bandwidth and after the load factor reaches τ_H then only handover calls are admitted.

Case 3) When a NRT data call request is issued it is assigned by the system as low-priority call and served only if RT spare resources exist

$$Lf_T = Lf_P + Lf_{\mathrm{NRT}} < 1 \tag{9}$$

where Lf_{NRT} is the load factor of the NRT application. In case no NRT available bandwidth exists or a RT call initiates a preemption procedure, the NRT calls are queued in a buffer. The length of the buffer is assumed infinite. It becomes clear that by using the above algorithm, although system outage never occurs, the delay of NRT calls in the buffer could be long and for extreme RT load situations the NRT call service would seem impossible even for requests of very small information size. Therefore, QoS CAC 2 provides for NRT applications a bandwidth reservation τ_{NRT} in order to reduce the waiting time in the buffer.

QoS CAC 2 also consists of three cases similar to QoS CAC 1.

Case 4) When a new call request of RT QoS class is made, it is only admitted if the cumulative load factor of each application accommodated in the system is below the threshold $\tau'_H Lf_{\mathrm{GCD}} = \tau_H Lf_{\mathrm{GCD}} \frac{M'}{M}$ like (8) in QoS CAC 1.

$$Lf_T = Lf_P + Lf_N \leq \tau'_H Lf_{\mathrm{GCD}} < 1 \tag{10}$$

Case 5) When a handover call request of a RT application is made, it is admitted only if the total load factor after the admittance is below the bandwidth reservation τ_{NRT}. Hence, in QoS CAC 2 the handover bandwidth is limited by τ_{NRT}. Similar to Case 2) of QoS CAC 1 we have:

$$Lf_T = Lf_P + Lf_H < \tau_{\mathrm{NRT}} Lf_{\mathrm{GCD}}. \tag{11}$$

Case 6) When a NRT data call is issued, it can be admitted in the system only if there are available resources in the NRT-bandwidth. As it is mentioned above, by reserving capacity for NRT applications, the algorithm QoS CAC 2 evolves to an even more QoS oriented scheme. The bandwidth reservation for NRT calls and handovers requests is

$$Lf_T = Lf_P + Lf_{\mathrm{NRT}} < 1 \tag{12}$$

The available bandwidth for NRT calls is the reservation bandwidth plus the spare RT bandwidth. The admittance is granted without negotiations over the bit rate of the application until all the available bandwidth is fully occupied. Therefore, all the excess requests of NRT applications are queued in the buffer. However, in this second algorithm, the delay time in the buffer is reduced significantly.

Finally, it must be noted that the above proposed algorithms can be characterized as NRT preemptive. In terms of negotiation, preemption is the total negotiation of connectivity that means that the proposed algorithms give high priority to RT classes. If equal priority is desirable, the NRT calls can use the same resource pool and follow certain negotiation rules the same way RT calls do. As such, the rest of the analysis would be the same with the exception of arrival rates where the sum of RT and NRT should be taken. However, NRT traffic negotiation can be enabled via NRT reservation bandwidth as explained in Section VI.

C. Important Issues

Soft handover is an important issue for WCDMA systems. It is founded on the idea of interference limited capacity. The terminal can maintain multiple connections with the base stations and divide the transfer of information using smaller load factors. Our proposed algorithm can be easily evolved to embody the soft handover feature using partial load factors for soft handover calls. However, it is appreciated that a multicell analysis is required for proper conclusion extraction. Therefore, soft handover is left as a future expansion of the present work.

Another point of interest is the connection delay. The proposed CAC algorithms contain a check and recheck procedure, a negotiation and renegotiation with the handsets and even the preemption of NRT calls. Since the WCDMA specifications include real time transmission and calculation of RF parameters, the information needed to calculate the load factor of every call and the total load factor are considered known to the system. As such, checking the load factors and threshold can be considered instantaneous. The negotiation procedure can be accelerated if the negotiation rules of every handset are known to the system since registration time. In this case, negotiations can last as short as one frame plus the time span which the handset need to switch its power characteristics. Renegotiation takes the same time as negotiation as long as the controller possesses the information of negotiation rules. Negotiation rules should better be modifiable by the user at all times. In any case, though, the information is instantaneously transferred to the system in order to be available at call request time. This knowledge enables the controller to perform all checks before it communicates with the handsets.

Finally, we consider preemption to be a fast procedure during which the resources are abruptly taken from NRT calls. The conclusion is that every admission procedure that takes place under this proposal is not expected to last more than two WCDMA frames.

V. Teletraffic Analysis

In this section of the paper, the proposed system is analysed and compared with the existing systems in view of improvement demonstration. The crucial parameters involved in a CAC scheme design are the blocking probability P_B, the outage probability P_O, the dropping probability P_D, the NRT class delay T_D and the resource utilization factor RU_F. We will calculate these features for our new schemes and compare them to the respective features of other systems in several load environments.

For the purposes of this analysis, markov chains will be used. In [9], [11], and [12], the class differentiation admission schemes were analysed with multidimensional chains. The classical teletraffic analysis of markov chains uses the channel as the basic resource unit. Every new request occupies an amount of bandwidth equal to the resource unit. There is no option for a new call with a requirement of one and a half channel and this works because multiclass systems were analysed with multidimensional chains until now.

In our analysis, we make the following assumptions:

The load factor of every new call is calculated and used as resource requirement.

1) We define the basic resource unit as the greatest common divisor (GCD) of the load factor of every possible application, Lf_{GCD}.
2) We assume exponential distribution for the remaining time until the next birth and next death with λ arrival rate and μ service rate.
3) We have a system with N different QoS classes that some of them issue negotiation between them.
4) We have a resource pool of $Lf_M = M \cdot Lf_{\mathrm{GCD}} = 1$ for RT applications, since $Lf_{\mathrm{total}} \leq 1$ is the necessary requirement so as the system will not suffer an outage situation.
5) Therefore we make the following steps in our analysis:
6) We use a linear chain with Lf_{GCD} steps and rates λ_i and μ_i for every application i.
7) We allow for multiple steps over the markov chain equal to Lf_i/Lf_{GCD}.
8) We use thresholds for handover reservation bandwidth and NRT class resource reservation.
9) We transform our model to accommodate QoS negotiation.

Fig

In Fig. 4, a markov chain for a variable-application system is showcased. Steps of one, two and three chains are depicted. Thus, the system can, i.e., be transported from the condition j directly to state $j+2$. This happens when an application with resource requirements $2 \cdot Lf_{\mathrm{GCD}}$ is admitted. As such, when the call is serviced, the system will travel back two steps. Similarly with the single-step case, every system state j has a probability of occurrence P_j. The total system conditions are $M+1$.

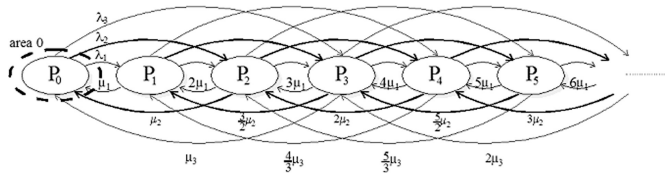


Fig. 4. Markov chain with multiple steps. Only three applications are described in this example for reasons of presentability.

We have assumed that λ_i is the call arrival rate and μ_i is the call service rate where i is defined in (6). Initially we assume that the system has a full set of applications having all the multiples of the basic resource unit Lf_{GCD}. For the real system analysis we can later set λ_i and μ_i zero for every multiple i that does not exist in our application variant.

The calculation of the service rate for every multi-step is based on the idea that the channel is assumed to be occupied by a random combination of users. A possible approximation could be to assume that $\lambda_i / \sum_{i=1}^{N} \lambda_i$ users are served with the application i. Then $\mu_i^j = \frac{j}{i}\mu_i \cdot \lambda_i / \sum_{i=1}^{N} \lambda_i$. The third multiplier represents the number of paralleled servers for the specific application. For the rest of the analysis we set $\mu_i' = \mu_i \cdot \lambda_i / \sum_{i=1}^{N} \lambda_i$. For $i=1$ and $i=2$, we obviously have $\mu_1', 2\,\mu_1', 3\,\mu_1', 4\,\mu_1', \ldots$ and $\mu_2', 3/2\,\mu_2', 2\,\mu_2', 5/2\,\mu_2', \ldots$ respectively. The concept of server and resource pool are directly relative to the bandwidth of the application in question.

Balance equations can be extracted by using the probability flux theorem [20] for the area surrounding every ring of the chain (as in Fig. 4). According to this, the probability flux through this area is zero. Any area can be used likewise in order to provide balance equations. The choice is made upon the complexity of the resulting equations:

$$\text{area } 0: \quad P_0 \sum_{i=1}^{N} \lambda_i = \sum_{i=1}^{N} \mu_i' P_i$$

$$\text{area } 1: \quad P_1 \left(\sum_{i=1}^{N} \lambda_i + \mu_1' \right) = \sum_{i=1}^{N} \frac{i+1}{i} \mu_i' P_{1+i} + \lambda_1 P_0$$

$$\text{area } 2: \quad P_2 \left(\sum_{i=1}^{N} \lambda_i + 2\mu_1' + \mu_2' \right) = \sum_{i=1}^{N} \frac{i+2}{i} \mu_i' P_{2+i} + \lambda_1 P_1 + \lambda_2 P_0$$

$$\vdots$$

$$\text{area } j: \quad P_j \left(\sum_{i=1}^{N} \lambda_i + \sum_{i=1}^{N} \frac{j}{i} \mu_i' \right) = \sum_{i=1}^{N} \frac{i+j}{i} \mu_i' P_{j+i} + \sum_{i=1}^{N} \lambda_i P_{j-i}$$

$$\vdots$$

$$\text{area } M-2: \quad P_{M-2}\left(\lambda_1+\lambda_2+\sum_{i=1}^{N}\frac{M-2}{i}\mu_i'\right) = (M-1)\mu_1' P_{M-1} + \frac{M}{2}\mu_2' P_M + \sum_{i=1}^{N}\lambda_i P_{M-2-i}$$

$$\text{area } M-1: \quad P_{M-1}\left(\lambda_1+\sum_{i=1}^{N}\frac{M-1}{i}\mu_i'\right) = M\mu_1' P_M + \sum_{i=1}^{N}\lambda_i P_{M-1-i} \tag{3}$$

The flux exiting the area on the left-hand side is equal to the flux entering the area on the right-hand side.

Moreover, we define the load of the system as ℓ

$$\ell = \sum_{i=1}^{N}\frac{\lambda_i}{\mu_i}. \tag{14}$$

Defining $\mathbf{P} = [P_1 \;\; P_2 \;\; \cdots \;\; P_M]^T$ we have the linear system of equations $\mathbf{\Lambda}\cdot\mathbf{P} = \mathbf{P_0}$ where (Please see the equation at the bottom of the page.)

The $\mathbf{\Lambda}$ matrix contains traffic information and is a $M \times M$ banded matrix. We showcased Λ for $M = N + 2$, without loss of generality, for reasons of apprehension. $\mathbf{P}$ matrix contains the system P_j state probabilities and $\mathbf{P}_0$ contains the unknown variable P_0. The system solution is expressed as a function of P_0 and then the final values are calculated with the help of the equation $\sum_{i=0}^{M} P_i = 1$. A very important issue when solving the above system of equation, i.e., when inverting the matrix $\mathbf{\Lambda}$, is to use the proper number of equations. Since the quantization of the resource pool is done with Lf_{GCD} steps, there is a chance that some of the last steps are never used by the system. That is, if there is no combination of different class users that can fully occupy the channel given by the equation $Lf_M \leq 1$. We found that the unnecessary steps R_S are given by

$$R_S = \mathrm{mod}\left(\frac{M}{i_{\min}}\right) \tag{15}$$

Where $i_{\min}$ is the minimum LFM used by any class in the given system. In case the application $i = 1$ is included, no unnecessary steps are observed.

A. Thresholds

In order to impose handover thresholds in this analysis, λ is split to $\lambda_{\mathrm{Hi}} + \lambda_{\mathrm{Ni}} = \lambda_i$. The function $\tau_H(i)$, denotes the threshold for handover reservation bandwidth as a function of application i. According to the above, the system ceases admitting any new calls from application i when it reaches the state $j = \tau_H(i)$. Despite having a fixed handover reservation bandwidth, $\tau_H(i)$ is not the same for every application. High-resource applications must be blocked several markov steps before the low-resource counterparts so as to preserve the crossing of the threshold from a new call. From the above, we have

$$\tau_H(i) = \tau_H - i \tag{16}$$

$$\tau_H(1) \equiv \tau_H. \tag{17}$$

Using $\tau_H(i)$ as input, we transform the equations by using λ_{Hi}

$$\sum_{i=1}^{\min(N,M-j)}\lambda_i = \sum_{i=1}^{\min(N,M-j)}\lambda_{H_i} + \sum_{i=1}^{\substack{\min(N,M-j)\\ \text{if } j<\tau_H(i)}}\lambda_{N_i}. \tag{18}$$

Using the handover ratio $\gamma = (\lambda_H)/(\lambda_H+\lambda_N)$ as in [18], the above is rewritten as

$$\sum_{i=1}^{\min(N,M-j)}\lambda_i = \sum_{i=1}^{\min(N,M-j)}\gamma\lambda_i + \sum_{i=1}^{\substack{\min(N,M-j)\\ \text{if } j<\tau_H(i)}}(1-\gamma)\lambda_i. \tag{19}$$

In QoS CAC 2, a provision for NRT class bandwidth reservation is made. For this algorithm we set $M' = Lf_{\tau\mathrm{NRT}}/Lf_{\mathrm{GCD}} < M$ the new resource pool for RT applications. The analysis for RT and NRT is clearly separate and no

$$\mathbf{\Lambda} = \begin{bmatrix} \mu_1' & \mu_2' & \mu_3' & \cdots & \mu_N' & 0 & 0 \\ -\left(\sum_{i=1}^{N}\lambda_i+\mu_1'\right) & 2\mu_1' & \frac{3}{2}\mu_2' & \cdots & \frac{N}{N-1}\mu_{N-1}' & \frac{N+1}{N}\mu_N' & 0 \\ \lambda_1 & -\left(\sum_{i=1}^{N}\lambda_i+\sum_{i=1}^{2}\frac{2}{i}\mu_i'\right) & 3\mu_1' & \cdots & \frac{N}{N-2}\mu_{N-2}' & \frac{N+1}{N-1}\mu_{N-1}' & \frac{N+2}{N}\mu_N' \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ \lambda_{N-2} & \lambda_{N-3} & \lambda_{N-4} & \cdots & (M-2)\mu_1' & \frac{M-1}{2}\mu_2' & \frac{M}{3}\mu_3' \\ \lambda_{N-1} & \lambda_{N-2} & \lambda_{N-3} & \cdots & -\left(\sum_{i=1}^{2}\lambda_i+\sum_{i=1}^{N}\frac{M-2}{i}\mu_i'\right) & (M-1)\mu_1' & \frac{M}{2}\mu_2' \\ \lambda_N & \lambda_{N-1} & \lambda_{N-2} & \cdots & \lambda_1 & -\left(\lambda_1+\sum_{i=1}^{N}\frac{M-1}{i}\mu_i'\right) & M\mu_1' \end{bmatrix}$$

$$\mathbf{P_0} = P_0\begin{bmatrix}\sum_{i=1}^{N}\lambda_i & -\lambda_1 & -\lambda_2 & \cdots & -\lambda_{N-1} & -\lambda_N & 0\end{bmatrix}^T.$$

thresholds for NRT classes are needed for the purposes of RT class analysis.

B. Negotiation and Renegotiation

Negotiation issues are examined with a new approach. We define n_{ij} the total percentage of class i users that accept as maximum negotiation a transition to class j (Obviously $j \leq i$). Consequently, we define a $N \times N$ matrix $\mathbf{Neg} = [n_{ij}]$. Following the class arrangement we considered previously, we easily conclude that $n_{ij} = 0$ for every $i \leq j$. Thus $\mathbf{Neg}$ is down triangular.

$$\mathbf{Neg} = \begin{bmatrix} n_{11} & 0 & \cdots & \cdots & 0 \\ n_{21} & n_{22} & \cdots & \cdots & 0 \\ \cdots & \cdots & & & \cdots \\ \cdots & \cdots & & & \cdots \\ n_{N1} & n_{N2} & \cdots & \cdots & n_{NN} \end{bmatrix} \tag{20}$$

In a real problem, several more percentages n_{ij} will be set to zero, as not every application can be negotiated to every other. This mapping is left to mobile system operator and this enables a powerful tool when negotiating QoS services at sales office time.

Using the above definition, we can substitute the rates λ_i with the equivalent rates

$$\lambda_i^{'} = \lambda_i \left(1 - \sum_{\substack{j=1 \\ j \neq i}}^{N} n_{ij} \right) + \sum_{\substack{j=1 \\ j \neq i}}^{N} n_{ji} \lambda_j \tag{21}$$

where λ_i' is the equivalent call arrival rate in the case of maximum negotiation (i.e., all calls are negotiated to the minimum user requirement and no further negotiations or re-negotiations can be performed). The second term on the right side expresses the reduction due to negotiation of the respective application to another down in the QoS chain. The third term expresses the increase in the rate due to negotiation from a better application. The high-resource class rates are decreased as the second term is greater than the third and the low-resource class rates are increased, as the opposite is the case.

If we analyze the above markov chain with λ_i' ratios, it is evident that the sum of condition probabilities $P_{\tau_H(i)} + P_{\tau_H(i)+1} + \cdots + P_M$ will express the blocking probability of class i with full negotiation (negotiation and re-negotiation) assumed. It should be noted that in practice when the system is loaded but the calls are not fully negotiated, the resources are fully used (the system is in τ_H condition) but we do not have a blocking or outage condition. Blocking will occur after all calls are fully negotiated and the algorithm does not permit outage. For systems without negotiation we will use the λ_i rates.

C. Probabilities Calculation

According to the above, blocking probability is

$$P_B(i) = \sum_{k=0}^{\min(M,M')-\tau_H(i)} P_{\tau_H(i)+k}. \tag{22}$$

Dropping probability is

$$P_D(i) = \sum_{k=0}^{i} P_{\min(M,M')-k}. \tag{23}$$

Outage is prevented. As such, outage probability is

$$P_O = 0. \tag{24}$$

D. NRT Class Analysis

For NRT applications we must estimate the average delay for a user that is queued in the buffer. According to the Kendall model [21], our case is a $M/M/M_{\mathrm{NRT}}/\infty/\infty$ case, which is solved by an Erlang C model [21], [24].

As we have mentioned before, the available resources for NRT applications is $M_{\mathrm{NRT}} = (M - M') + (M' - j)$, where $j \leq M'$. The first term describes the reservation bandwidth (for QoS CAC 1 we have $M = M'$) and the second term describes the unusable resources of the RT resource pool when the RT system is in the state j. Thus, the analysis for NRT classes is directly relative to the results of analysed RT system.

We need to perform a new markov analysis, this time for the NRT resource pool. M_{NRT} are the available resources, λ_{NRT} and $1/\mu_{\mathrm{NRT}}$ are the arrival and service rates using the application with the basic load factor (i.e., Lf_{GCD}). Here we will assume constant load factor traffic. The correctness of the analysis is left on the proper calculation of λ_{NRT} and $1/\mu_{\mathrm{NRT}}$ given the rates λ_i and $1/\mu_i$ for every application. In particular, we will have

$$\lambda_{\mathrm{NRT}} = \sum_{i=1}^{N_{\mathrm{NRT}}} i\lambda_i$$

$$\mu_{\mathrm{NRT}} = \mu. \tag{25}$$

The following analysis is based on the average available resources $\bar{M}_{\mathrm{NRT}}$ for NRT calls:

$$\bar{M}_{\mathrm{NRT}} = \sum_{j=0}^{M} P_j (M - j) \tag{26}$$

The possibility that a call will delay will be

$$P_D = \frac{\dfrac{\ell_{\mathrm{NRT}}^{\bar{M}_{\mathrm{NRT}}}}{\bar{M}_{\mathrm{NRT}}!} \dfrac{1}{1 - \ell_{\mathrm{NRT}}/\bar{M}_{\mathrm{NRT}}}}{\displaystyle\sum_{i=0}^{\bar{M}_{\mathrm{NRT}}-1} \frac{\ell_{\mathrm{NRT}}^{i}}{i!} + \frac{\ell_{\mathrm{NRT}}^{\bar{M}_{\mathrm{NRT}}}}{\bar{M}_{\mathrm{NRT}}!} \dfrac{1}{1 - \ell_{\mathrm{NRT}}/\bar{M}_{\mathrm{NRT}}}} \tag{27}$$

where, $\ell_{\mathrm{NRT}} = (\lambda_{\mathrm{NRT}}/\mu_{\mathrm{NRT}})$. The average number of calls waiting in the queue will be

$$E\{w\} = \frac{\ell_{\mathrm{NRT}}}{\bar{M}_{\mathrm{NRT}} - \ell_{\mathrm{NRT}}} P_D. \tag{28}$$

The average time waiting in the buffer is finally calculated

$$T_D = \frac{E\{w\}}{\lambda_{\mathrm{NRT}}}. \tag{29}$$

The above are found in [22] and [24].

E. Channel Utilization

The channel utilization is defined as the ratio of occupied resources to the total system resources. This measure is a binomially distributed variable, proportional to the state variable j of the RT system and j' of the NRT. As such, we use the expected value of channel utilization, namely resource utilization factor RU_F, as a function of load (in Erlangs) in order to compare the proposed schemes to the existing. Moreover, we are interested in comparing the systems over the RT resource usage since the NRT traffic in pre-emptive systems tends to fully utilize the system.

For the first scheme, we find

$$\mathrm{RU}_{F_1} \equiv E\{cu\} = 1 - \frac{1}{M}\sum_{j=0}^{M} jP_j. \tag{30}$$

For the second scheme, we find

$$\mathrm{RU}_{F_2} \equiv E\{cu\} = 1 - \frac{1}{M}\sum_{j=0}^{M'} jP_j \tag{31}$$

where j corresponds to the possible RT states.

VI. Numerical Results

In this section we analyze and compare our proposed schemes (QoS CAC 1 and QoS CAC 2) against the straightforward Erlang-B and Handover Reservation schemes. The selection of these two models is justified because of their popularity and widespread applicability. No comparison can be made between our schemes and other QoS schemes because the philosophy of negotiation we introduce is not used in the past papers. Note that, in view of compatibility, we have assumed preemptive systems in all four cases. System features are set equally in every case, as well.

In terms of the greatest common divisor of every application load factor, we make the approximation $Lf_{\mathrm{GCD}} = 0.01$. Consequently, the number of assumed resources is $M = 100$. Table II shows the chosen applications and the respective resource requirements as Lf_{GCD} multiples. These multiples correspond to the multiple markov steps used in the analysis. The rest of the assumptions are as follows.

1) Handover reservation is set to 10%.
2) NRT bandwidth reservation is set to 5%.
3) We have equal traffic for voice and data calls (without loss of generality).
4) The average call duration is 90 sec.
5) The voice calls are distributed equally to V1 and V2 class.
6) The data calls are distributed as shown in Table II.
7) Handover ratio is assumed to be $\gamma = 0.6$.

The negotiation matrix $\mathbf{Neg}$ is obviously chosen by the network operator. As such, we made a rational assumption and use the following matrix:

$$\mathbf{Neg} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 0.3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.9 & 0.1 & 0 & 0 \\ 0 & 0 & 0.9 & 0.05 & 0.05 & 0 \\ 0 & 0 & 0.7 & 0.13 & 0.12 & 0.05 \end{bmatrix}. \tag{32}$$

The mathematical analysis of the above case yields Figs. 6–14. In Figs. 6 and 7, the handover reservation phenomenon is clear. Fig Regarding blocking probability for RT calls, the Erlang-B model has the optimal performance. However, the other three models, using handover reservation bandwidth, make a provision for handover calls, thus maximizing QoS performance. In [23], it is made clear that overall QoS is much more sensitive to

TABLE II
Simulation Values for QoS Classes

QoS Class	Application	LFM Multiples	Load Percentage
1	V1	1	50%
2	V2	2	50%
3	D1	1	30%
4	D2	1	30%
5	D3	3	20%
6	D4	6	20%
	NRT	1	

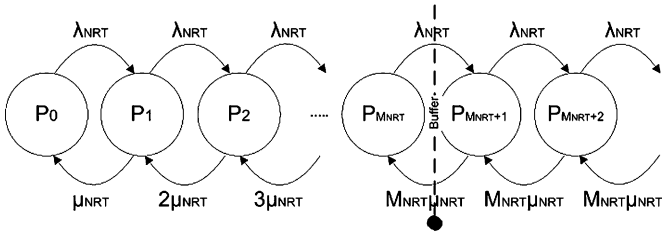


Fig. 5. Markov chain for NRT traffic with infinite buffer. The NRT resource is set to M_{NRT}.

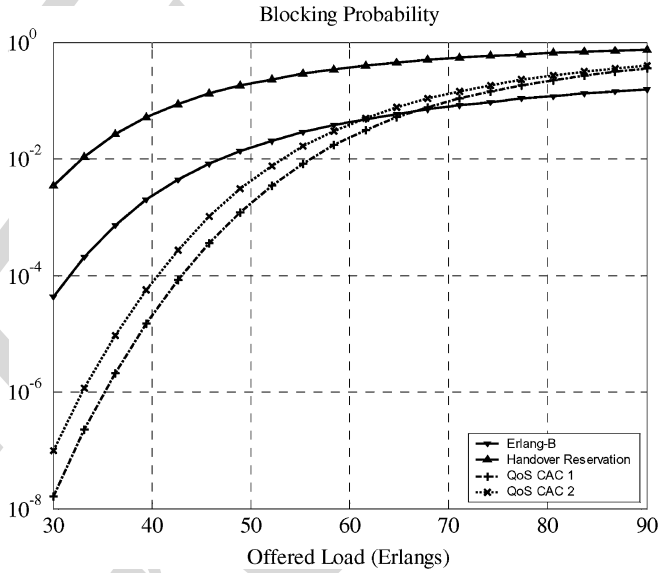


Fig. 6. Blocking probability for V1 calls versus overall offered load. Handover Reservation bandwidth is set to 10%, average duration is 90 s and handover ratio assumed 0.6.

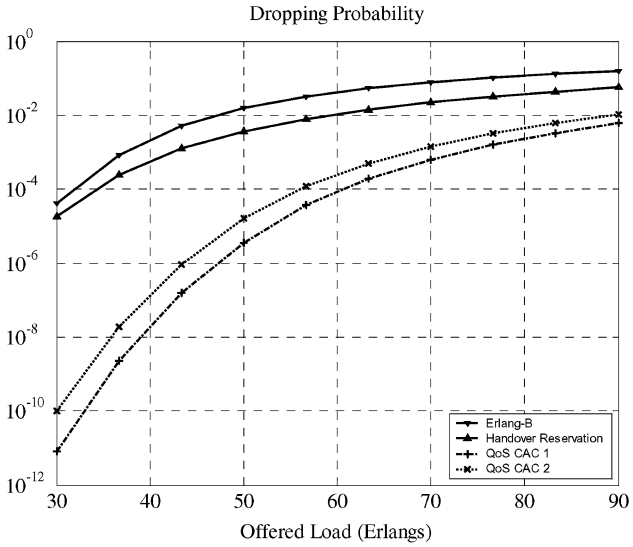


Fig. 7. Dropping probability for V1 calls versus overall offered load. Handover Reservation bandwidth is set to 10%, average duration is 90 sec and handover ratio assumed 0.6.

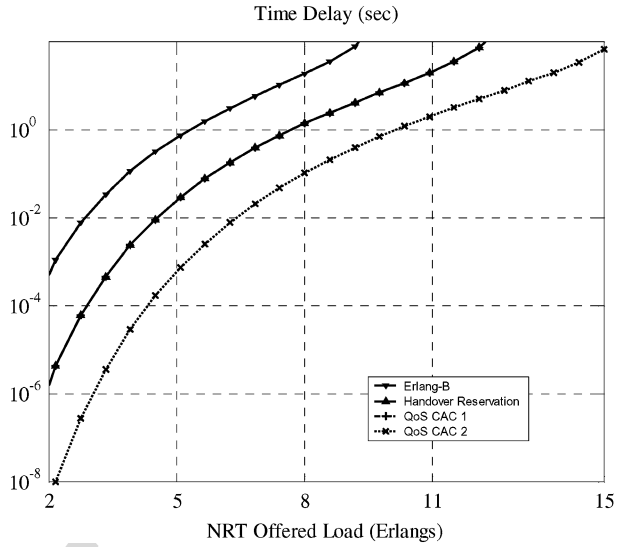


Fig. 9. Average time delay for 60 Erlangs RT traffic. Handover Reservation bandwidth is set to 10%, NRT reservation bandwidth is 5%, average duration is 90 s and handover ratio assumed 0.6.

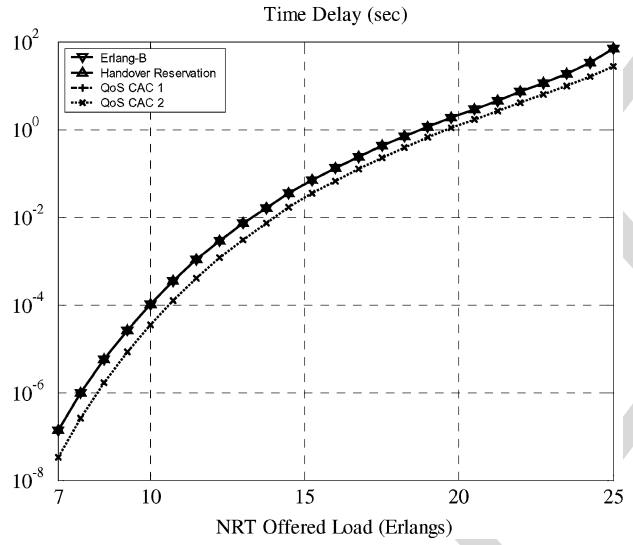


Fig. 8. Average time delay for 40 Erlangs RT traffic. Handover reservation bandwidth is set to 10%, NRT reservation bandwidth is 5%, average duration is 90 s and handover ratio assumed 0.6.

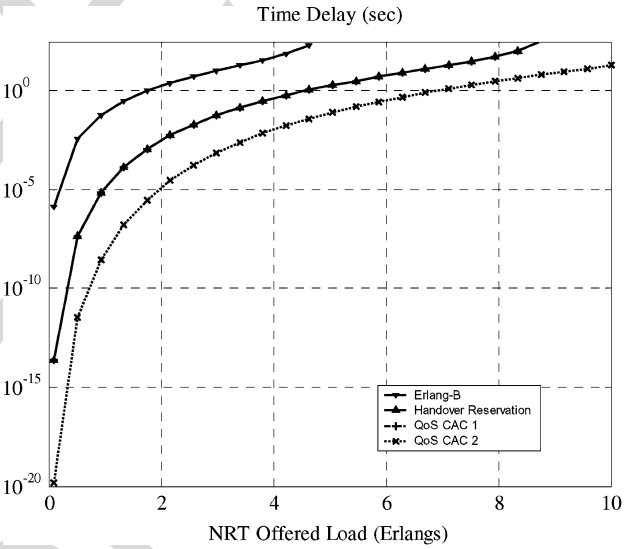


Fig. 10. Average time delay for 80 Erlangs RT traffic. Handover reservation bandwidth is set to 10%, NRT.

dropping percentage than blocking. Moreover, it is observed that the negotiation procedure has offered a great deal regarding blocking and dropping. In both cases, an advantage is gained towards the simple handover reservation model. It is shown that Erlang-B blocking performance can be reached.

Figs. 8–10 show the average delay that a NRT call will wait on the buffer when the RT traffic is, respectively, in Erlangs (E), 40, 60, and 80 E. In these figures, the gain from NRT reservation bandwidth is demonstrated. When the RT traffic is heavy (see Fig. 10), the resources are used up by RT traffic and the NRT calls are preempted in the buffer. QoS CAC 2 utilises a 5% NRT reservation bandwidth in order to continue to serve NRT calls in this case. From Fig. 8, is obvious that for light RT traffic, NRT behavior is the same for all schemes since most of the resource pool is available to NRT calls. In Figs. 9 and 10, it is shown that handover reservation bandwidth also reduces delay in the buffer and this is relative to resource utilization by RT traffic.

In , channel utilization factor is shown. This diagram shows Fig.11 that the Erlang-B is the most efficient in bandwidth utilization. In the Erlang-B scheme, the resources are equally available to all traffic demand. A scheme that uses bandwidth reservation is bound to have smaller utilization factor than Erlang-B. However, the negotiation proposed in our schemes enables the network operator to provide even greater quality to the user when the blocking probability remains under the Service Level Agreement value. This effect is demonstrated in . The utiliza- Fig.12 tion factor is calculated for the case in which the load of the proposed scheme is increased so as the blocking and dropping

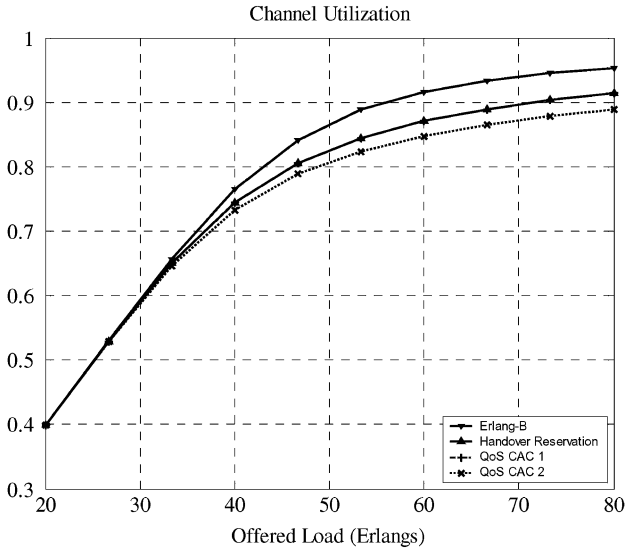


Fig. 11. Channel utilization factor. Handover reservation bandwidth is set to 10%, average duration is 90 sec and handover ratio assumed 0.6. NRT traffic is not accounted.

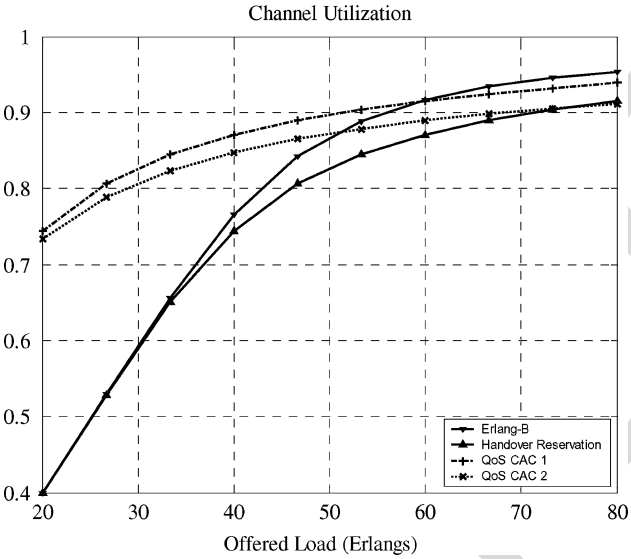


Fig. 12. Channel utilization factor. The traffic for the proposed schemes QoS CAC 1 and QoS CAC 2 is increased so as to equal the traffic of the other schemes when it is totally negotiated.

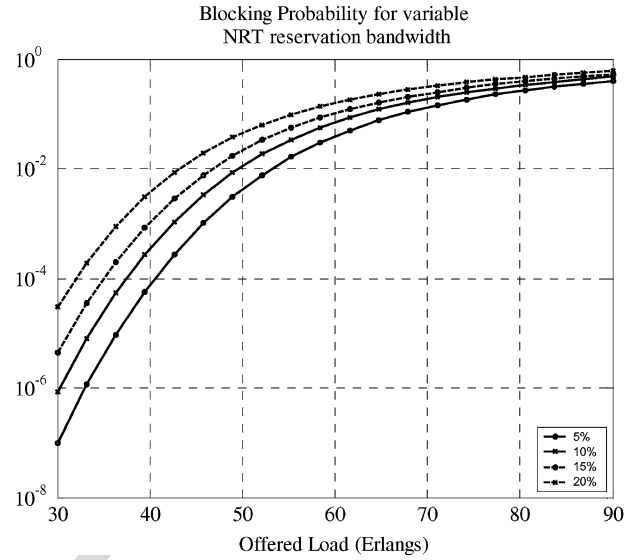


Fig. 13. Blocking probability of RT traffic for the QoS CAC 2 scheme. The NRT reservation bandwidth is set to 5%, 10%, 15%, and 20%.

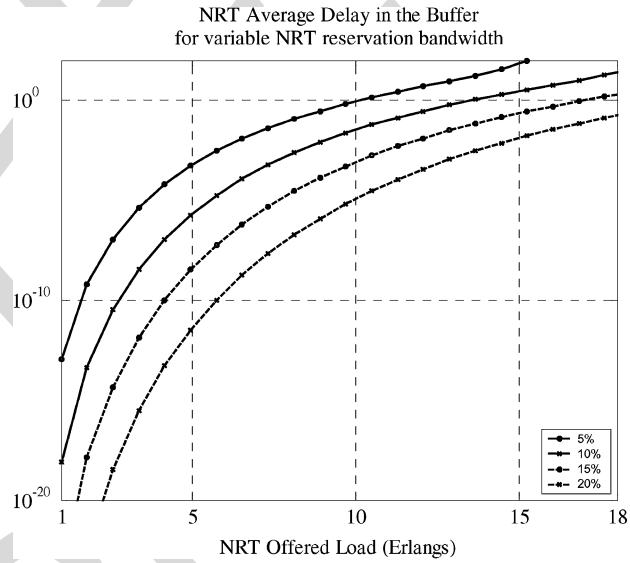


Fig. 14. NRT traffic average delay in the buffer for the QoS CAC 2 scheme. The NRT reservation bandwidth is set to 5%, 10%, 15%, and 20%. RT traffic is set to 60 E.

probabilities equal those of the Handover Reservation scheme. In this case, the negotiation procedure becomes evident. Due to negotiation, the channel utilization rises quickly and then it becomes steady close to 1.

In Figs. 13 and 14, we get different values for blocking probability and NRT delay of QoS CAC 2 scheme for several values of NRT reservation bandwidth. This negotiation procedure between RT and NRT traffic can take place in a real-time manner.

VII. Conclusion

Two new QoS-based CAC schemes are proposed; QoS CAC 1 and QoS CAC 2. Both schemes are designed to be used in the WCDMA radio interface which provides variable level quality. When the system is lightly loaded, calls are admitted and serviced with the maximum resource requirements of the user. When the outage condition is reached and no further calls can be accommodated, a negotiation procedure with the new and ongoing calls begins in order to free up some resources. Blocking occurs after every possible negotiation has been performed. This turbo channel utilization causes only one defect. Using preemption for NRT traffic, the delay for a NRT call can be enormous at RT load situations, due to this unfair channel occupation. We showed that this disadvantage can be solved by using a small proportion of the available resources as a NRT

reservation bandwidth. This new CAC technique can be evolved to be a powerful tool for network optimization at sales time.

REFERENCES

[1] R. Lloyd-Evans, *QoS in Integrated 3G Networks*. Boston, MA: Artech House, 2002.

[2] J. Laiho and A. Toskala, "Overview of WCDMA," *IEEE Veh. Technol. Soc. News*, vol. 50, no. 1, Feb. 2003.

[3] N. Dimitriou and R. Tafazolli, "Quality of service for multimedia CDMA," *IEEE Commun. Mag.*, Jul. 2000.

[4] W. S. Geon and D. G. Geong, "Call admission control for a CDMA mobile communications systems supporting multimedia services," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, Oct. 2002.

[5] M. S. Do, Y. Park, and J. Y. Lee, "Channel assignment with QoS guarantees for a multiclass multi-code CDMA system," *IEEE Trans. Veh. Technol.*, vol. 51, no. 5, Sep. 2002.

[6] J. Y. Lee, J. G. Choi, K. Park, and S. Bahk, "Realistic cell-oriented adaptive admission control for QoS support in wireless multimedia networks," *IEEE Trans. Veh. Technol.*, vol. 52, no. 3, May 2003.

[7] J. R. Moorman and J. W. Lockwood, "Wireless call admission control using threshold access sharing," *IEEE* , vol. 6, p. 2529, Nov. 2001.

[8] Y. W. Kim, D. K. Kim, J. H. Kim, S. M. Shin, and D. K. Sung, "Radio resource management in multi-chip-rate DS/CDMA systems supporting multiclass services," *IEEE Trans. Veh. Technol.*, vol. VT-50, May 2001.

[9] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks," *IEEE Trans. Veh. Technol.*, vol. 51, no. 2, Mar. 2002.

[10] H. Hlavacs, G. Haring, A. Kamra, and M. Bansal, "Modelling resource management for multi-class traffic in mobile cellular networks," in *Proc. IEEE 35th Annual Hawaii Int. Conf. on System Sciences*, Jan. 2002, pp. 1539–1548.

[11] C. N. Wu, Y. R. Tsai, and J. F. Chang, "A quality-based birth-and-death queuing model for evaluating the performance of an integrated voice/data CDMA cellular system," *IEEE Trans. Veh. Technol.*, vol. 48, no. 1, Jan. 1999.

[12] J. Wang, Q. A. Zeng, and D. P. Agrawal, "Performance analysis of a pre-emptive and priority reservation handoff scheme for integrated service-based wireless mobile networks," *IEEE Trans. Mob. Comp.*, vol. 2, Jan./Mar. 2003.

[13] K. I. Pedersen and P. E. Mogensen, "Directional power based admission control for WCDMA systems using beam forming antenna arrays system," *IEEE Trans. Veh. Technol.*, vol. 51, no. 6, Nov. 2002.

[14] F. Gunnarsson, E. Geijer-Lundin, G. Bark, and N. Wiberg, "Uplink admission control in WCDMA based on relative load estimates," in *Proc. IEEE International Conference on Communications*, vol. 5, 28 Apr./May 2002, pp. 3091–3095.

[15] H. Solana, A. V. Bardaji, and F. C. Palacio, "Capacity analysis and performance evaluation of call admission control for multimedia packet transmission in UMTS WCDMA system," *Proc. IEEE WCNC 2003*, vol. 3, pp. 1550–1555, Mar. 2003.

[16] Capone and S. Redana, "Call admission control techniques for UMTS," *Proc. IEEE VTC 2001 Fall*, vol. 2, pp. 925–929, Oct. 2001.

[17] W. Ying, Z. Jingmei, W. Weidong, and Z. Ping, "Call admission control in hierarchical cell structure," in *Proc. IEEE VTC Spring 2002*, vol. 4, May 2002, pp. 1955–1959.

[18] D. Hong and S. S. Rappaport, "Traffic model and performance analysis of cellular radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. VT-35, no. 3, Aug. 1986.

[19] H. Holma and A. Toskala, *WCDMA for UMTS*. New York, NY: Wiley, 2002.

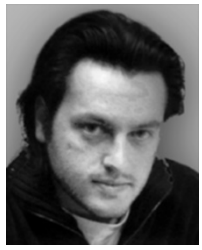
[20] W. J. Stewart, *Computations with Markov Chains*. New York, NY: Kluwer Academic, 1995.

[21] P. Bremaud, *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues*. New York, NY: Springer, 1999.

[22] F.-N. Pavlidou, *Digital Telephony*. Greece: Aristotle University of Thessaloniki, 1991.

[23] W. C. Hardy, *Measurement and Evaluation of Telecommunications Quality of Service*. New York, NY: Wiley, 2001.

[24] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed., New York, NY: McGraw-Hill, 2002.



Georgios Stavrou Paschos (M'99) was born in Athens, Greece, in 1978. He received the Diploma degree in electrical and computer engineering, Polytechnic School of Aristotle University of Thessaloniki in 2002. He is currently working towards the Ph.D. degree in telecommunications in the School of Electrical Engineering and Computer Science in the University of Patras, Greece.

His main interests are wireless networks, quality of service, and network management.

Ilias Politis received the B.Sc. degree in electronic engineering from Queen Mary College, London, U.K., in 2000 and the M.Sc. degree in mobile and personal communications from King's College, London, U.K., in 2001. He is currently working towards the Ph.D. degree in the Department of Electrical and Computer Engineering in University of Patras, Greece.

His research interests include wireless mobile ad hoc networks and hybrid wireless mobile networks, routing, and power management protocols.

Stavros Kotsopoulos was born in Argos-Argolidos, Greece, in 1952. He received the B.Sc. degree in physics in 1975 from the University of Thessaloniki, Greece, and received the Diploma degree in electrical and computer engineering from the University of Patras, Greece. He did his postgraduate studies in the University of Bradford, U.K. He also received the M.Phil. and Ph.D. degrees in 1978 and 1985, respectively.

Currently, he is Member of the Academic Staff of the Department of Electrical and Computer Engineering of the University of Patras and holds the position of Associate Professor. Since 2004, is the Director of the Wireless Telecommunications Laboratory and develops his professional life teaching and doing research in the scientific area of Telecommunications), with interest in mobile communications, interference, satellite communications, telematics applications, communication services and antennae design. Moreover he is the (co)author of the book titled "*mobile telephony.*" The research activity is documented by more than 160 publications in scientific journals and proceedings of International Conferences. He has been the leader of several international and many national research projects.

Dr. Kostopoulos is member of the Greek Physicists Society and member of the Technical Chamber of Greece.

Queries

Q1. Au: Please provide vol, pages, issue number-ED.
Q2. Au: Please provide title of publication end page number-ED.
Q3. Au: Please provide first initial(s)-ED.