A Call Admission Control for QoS-Sensitive Wireless Mobile Networks

Youssef Iraqi and Raouf Boutaba

University of Waterloo, Canada

E-mail: {iraqi,rboutaba}@bbcr.uwaterloo.ca

In this paper we propose a novel Distributed Call Admission Control (DCAC) framework designed for wireless mobile multimedia networks. The call admission process involves not only the cell that receives the call admission request but also a cluster of neighboring cells so the user will not be dropped due to handoffs. We evaluate the performance of the DCAC scheme in terms of average bandwidth utilization, call dropping probability and call blocking probability. Simulations demonstrate that for multimedia support, explicit involvement of cells gives better performance. We also investigate the impact of the number of involved cells in the DCAC process on the achieved performance.

1. Introduction

As the mobile network is often simply an extension of the fixed network infrastructure from the user perspective, mobile wireless users will demand wireline-similar services. Such demand will continue to increase with the growth of multimedia computing and collaborative networking applications. If any QoS is to be provided in a wireless environment, admission control is required. Admission control is already a difficult problem in traditional wired networks and it becomes more complex when considering wireless networks. In a wireless network mobiles are free to move arbitrarily; thus, the network access point may change rapidly at unpredictable times. This has a tremendous effect on QoS.

Cellular mobile networks have to continue supporting their mobile users after they leave their original cells. This rises a new challenge to Call Admission Control (CAC) algorithms. A call admission process should not only take into consideration the available resources in the original cell but also in neighboring cells as well.

Mobile users are in a growing demand for multimedia applications, and the next generation wireless networks are designed to support such bandwidth greedy applications. The (wireless) bandwidth allocated to a user will not be fixed for the lifetime of the connection as in traditional cellular networks, rather the base station will allocate bandwidth dynamically to users. Many evolving standards for Wireless Broadband Systems, UMTS and IMT2000 have proposed solutions to support such capability [1] [2] [3].

Call admission schemes that involve other cells in the decision process can be divided into two categories. Implicit involvement schemes and explicit involvement schemes. In the implicit involvement schemes, cells are involved implicitly in the CAC process. An example of such scheme is the well known Guard Channel scheme [4] where each cell reserves a number of channels (or bandwidth units) for exclusive use by handoff users. In this scheme cells are implicitly involved in the CAC process. In the explicit category, cells are explicitly involved in the CAC process by exchanging important information about users and about their capabilities to support these users.

Several call admission algorithms have been proposed for wireless networks to support multimedia users with dynamic bandwidth requirements (e.g. [5]). These algorithms take only local information in the admission decision process, which result in a high call dropping probability. Call Dropping Probability (CDP) is an important connection level QoS parameter in wireless mobile networks. To reduce the call dropping probability, few other CAC algorithms which take into consideration information from neighboring cells have been proposed [6–8]. However, those algorithms only support users with fixed bandwidth requirements.

To our knowledge, no call admission control algorithm has been proposed for supporting multimedia users with dynamic bandwidth requirements while offering a low call dropping probability.

To achieve this goal, we propose a novel Distributed Call Admission Control framework designed for wireless mobile multimedia networks. The call admission process involves not only the cell that receives the call admission request but also a cluster of neighboring cells so the user will not be dropped due to handoffs which will provide a low call dropping probability.

In this paper, we present a performance evaluation of the CAC scheme in terms of Average Bandwidth Utilization (ABU) and Call Dropping Probability (CDP). Other parameters such as Call Blocking Probability (CBP) are also investigated. We compare our scheme with the Guard Channel scheme. We demonstrate that for multimedia support, explicit involvement of cells gives better performance in terms of CDP, CBP and ABU.

We also investigate the impact of the number of involved cells in the CAC process on the achieved performance, in terms of CDP, CBP and ABU.

The paper is organized as follows. In section 2, we describe the model of the system considered in this paper. Section 3 defines the dynamic mobile probabilities used by our distributed call admission control algorithm. In section 4 we present the call admission control performed locally by the cells in our system. Section 5 introduces the overall call admission control scheme involving a cluster of neighboring cells and gives the detailed steps of the distributed call admission control algorithm. Section 6 describes the algorithm used to dynamically achieve a target call dropping probability. Sections 7, 8 and 9 discuss the conducted simulations and present a detailed analysis of the obtained results. Finally, section 10 concludes the paper.

2. System Model

We consider a wireless/mobile network with a cellular infrastructure that can support mobile terminals running applications that demand a wide range of resources. Users can freely roam the network and experience a large number of handoffs during a typical connection. We assume that users have a dynamic bandwidth requirement. The wireless network must provide the requested level of service even if the user moves to an adjacent cell. A handoff could fail due to insufficient bandwidth in the new cell[1], and in such case, the connection is dropped.

To reduce the call dropping probability, we allow neighboring cells to participate in the admission decision of a new user. Each involved cell will give its local decision and then the cell where the request

[1] or in a neighboring cell if a mechanism like the directed retry is used

was issued will finally decide if the new request is accepted or not. By doing so, the new admitted connection will have more chance to survive after experiencing handoffs.

We use the notion of a cluster similar to the shadow cluster concept [9]. The idea is that every connection exerts an influence upon neighboring base stations. As the mobile terminal travels to other cells, the region of influence also moves. The cells influenced by a connection are said to constitute a cluster (see figure 1). Each user in the network, with an active connection has a cluster associated to it[2]. The cells in the cluster are chosen by the cell where the user resides. The shape and the number of cells of a user's cluster depend on factors such as user's current call holding time, user's QoS, terminal trajectory and velocity.

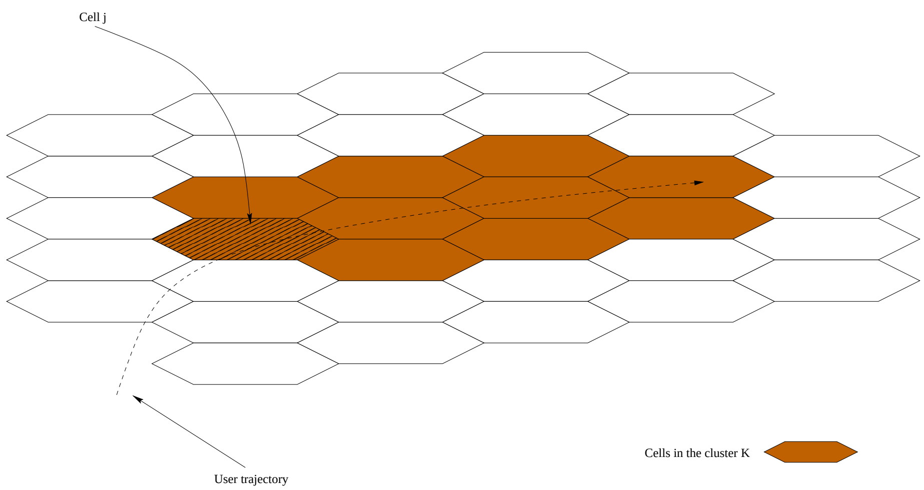

Figure 1. Cell j and the cluster for a user

3. Dynamic Mobile Probabilities

We consider a wireless network where the time is divided in equal intervals at $t = t_0, t_1, ..., t_m$ where $\forall i \geq 0 \; t_{i+1} - t_i = \tau$. Let j denote a base station in the network[3], and x a mobile terminal with an active wireless connection. Let $K(x)$ denote the set of cells that form the cluster for user x. We write $P_{x,j,k}(t) = [P_{x,j,k}(t_0), P_{x,j,k}(t_1), ..., P_{x,j,k}(t_{m_x})]$ the probability that mobile terminal x, currently in cell

[2] in the rest of the paper the term "user" and "connection" are used interchangeably

[3] we assume a one to one relationship between a base station and a network cell

j, to be active in cell k, and therefore under the control of base station k, at times t_0, t_1, t_2, ..., t_{m_x}. $P_{x,j,k}(t)$ represents the projected probabilities that mobile terminal x will remain active in the future and at a particular location. It is referred to as the Dynamic Mobile Probability (DMP) in the following. The parameter m_x represents how far in the future the predicted probabilities are computed. It is not fixed for all users and can depend of the user's QoS or the actual connection elapsed time.

Those probabilities may be function of several parameters such as: the handoff probability, the cell size, the user mobility profile, etc.

For each user x in the network, the cell responsible for this user decides the size of the cluster $K(x)$. The cells in $K(x)$ are those involved in the CAC process. The cell responsible for user x sends the DMPs to all members in $K(x)$ specifying whether the user is a new one (in which case the cell is waiting for responses from the members of $K(x)$) or not.

DMPs could range from simple probabilities to complex ones. Simple probabilities can be obtained by assuming, for example, that call length is exponentially distributed, call arrival process follows a Poisson distribution, handoff probabilities are equal in any direction and so on.

DMPs can also be complex, for example by including information about user mobility profiles. A method for computing dynamic mobile probabilities taking into consideration mobile terminal direction, velocity and statistical mobility data, is presented in [6]. Other schemes to compute these probabilities are presented in [7,8]. To compute these probabilities, one can also use mobiles' path/direction information readily available from certain applications, such as the route guidance system of the Intelligent Transportation Systems with the Global Positioning System (GPS).

4. Local Call Admission Control

User's traffic can be either voice, data or video. Voice users are usually characterized by a fixed bandwidth demand. Data and video users have a dynamic bandwidth requirement due to the burstiness of the carried traffic. Without loss of generality, we assume that all users are characterized by a bandwidth demand distribution $f_x(E_x(c), \sigma_c)$. Where $E_x(c)$ and σ_c are the mean and the standard deviation of the distribution f_x respectively, and c is the type of traffic for user x. Note that $E_x(c)$ depends of the traffic type c (voice, data or video).

In conjunction with the emergence of adaptive multimedia encoding [10–12], QoS adaptation schemes have been proposed to reduce handoff drops. In these schemes a connection's QoS can be downgraded if the available bandwidth in the new cell is not sufficient [7]. Such schemes can be easily integrated in our framework as part of the local call admission control.

4.1. Computing Elementary Responses

At each time t_0, each cell in a cluster $K(x)$ involved in our CAC process for user x, makes a local CAC decision for different times in the future $(t_0, t_1, ..., t_{m_x})$. Based on these CAC decisions, we call "Elementary Responses", the cell makes a final decision which represents its local response to the admission of user x in the network. Elementary responses are time dependent. The computation of these responses is different according to the user location and type: The user can be either a local new user or a new user that has a non null probability to be in this cell in the near future.

4.1.1. User types

A cell may be involved in the processing of different types of user. Possible user types at time t_0 are:

1. Old users local to the cell
2. Old users coming from another cell (executing a handoff)
3. New users (at time t_0) from within the cell
4. New users (at time t_0) from other cells

New users are defined as all users seeking admission at time t_0. Users of type 1 have the highest priority. Priority between other users is subject to some ordering policy. The network tries to support old users if possible and uses the DMPs to check if a cell can accommodate a new user who will possibly come to the cell in the future.

4.1.2. Local CAC at time t_0 for time t_0

The cell can apply any local call admission algorithm to compute the elementary responses. In this work we assume that the cells use the Equivalent Bandwidth approach to compute these responses. Example of such a scheme is described in [5]. The processing of local new users will be explained in section 5.

4.1.3. Local CAC at time t_0 for time t_l ($t_l > t_0$)

Each base station (BS) computes the equivalent bandwidth at different times in the future according to the DMPs of future users.

If user x, in cell j at time t_0, has a probability $P_{x,j,k}(t_l)$ to be active in cell k at time t_l and has a bandwidth demand distribution function $f_x(E_x(c), \sigma_c)$, then cell k should consider a user x', for time t_l, with a bandwidth demand distribution function $f'_{x'}(E_x(c) \times P_{x,j,k}(t_l), \sigma_c)$ and use it to perform its local admission control.

We write $r_k(x,t)$ the elementary response of cell k for user x for time t. We assume that $r_k(x,t)$ can take one of two values: -1 meaning that cell k can not accommodate user x at time t; and $+1$ otherwise.

The cell sets in which order of users it will perform its admission control. For instance, the cell can sort users in a decreasing order of their DMPs.

If we assume that user x_i has higher priority than user x_j for all $i < j$, then to compute elementary responses for user x_j, we assume that all users x_i with $i < j$ that have a positive elementary response are accepted. As an example, if a cell wants to compute the elementary response r for user x_4, and we have already computed r for users $x_1 = 1$, $x_2 = 1$ and $x_3 = -1$, then to compute r for x_4 the cell assumes that user 1 and 2 are accepted in the system but not user x_3.

4.2. Computing the Final Responses and Sending the Results

Since the elementary responses for future foreign users are computed according to local information about the future, they should not be assigned the same confidence degree. Indeed, responses corresponding to the near future are more likely to be more accurate than those of the far future.

We denote $C_k(x,t)$ the confidence that has cell k about its elementary response $r_k(x,t)$. The question arises on how the cell can compute (or simply choose) the confidence degrees $C_k(x,t)$, typically between 0% and 100%. The confidence degrees depend of many parameters. It is clear that the time in the future for which the response is computed has great impact on the confidence of that response. The available bandwidth when computing the elementary response also affects the confidence. In fact to give a complete picture the confidence degree also depends on the efficiency of the call admission routine. However, this later is difficult to determine.

To compute the confidence degrees we use a formula that uses the percentage of available bandwidth when computing the elementary response as an indication of the confidence the cell may have in this elementary response. The confidence degrees are computed using eq. 1:

$$C_k(x,t) = e^{(p-1)} * p^n \tag{1}$$

where p is a real number between 0 and 1 representing the percentage of available bandwidth at the time of computing the elementary response. And $n \geq 1$ is a parameter that is chosen experimentally to obtain the best efficiency of the call admission routine.

If for user x, cell k has a response $r_k(x,t)$ for each t from t_0 to t_{m_x} with a corresponding DMPs $P_{x,j,k}(t_0)$ to $P_{x,j,k}(t_{m_x})$, then to compute the final response those elementary responses are weighted with the corresponding DMPs. The final response from cell k to cell j concerning user x is then :

$$R_k(x) = \frac{\sum_{t=t_0}^{t=t_{m_x}} r_k(x,t) \times P_{x,j,k}(t) \times C_k(x,t)}{\sum_{t=t_0}^{t=t_{m_x}} P_{x,j,k}(t)} \tag{2}$$

where $C_k(x,t)$ is the confidence that has cell k about the elementary response $r_k(x,t)$. To normalize the final response each elementary response is also divided by the sum over time t of the DMPs in cell k. Of course, the sum $\sum_{t=t_0}^{t=t_{m_x}} P_{x,j,k}(t)$ should not be null (which otherwise means that all the DMPs for cell k are null!). Cell k, then, sends the response $R_k(x)$ to the corresponding cell j. Note that $R_k(x)$ is a real number between -1 and 1.

5. Distributed Call Admission Control

Here the decision takes into consideration the responses from all the cells in the user's cluster. The admission process concerns only new users seeking admission to the network, not already accepted ones. We assume that cell j has already decided the cluster $K(x)$ and that cell j has already assigned to each cell k in the cluster $K(x)$ a weight $W_k(x)$. Each weight represents the importance of the contribution of the associated cell to the global decision process. Usually, the more a cell is involved in supporting the user, the higher is its weight value. Weights $W_k(x)$ depend on the DMPs. We propose the following formula to compute the weights $W_k(x)$:

$$W_k(x) = \frac{\sum_{t=t_0}^{t=t_{m_x}} P_{x,j,k}(t)}{\sum_{k' \in K(x)} \sum_{t=t_0}^{t=t_{m_x}} P_{x,j,k'}(t)} \quad (3)$$

5.1. Relevance

Here, we introduce a new parameter that we call spatial relevance or simply relevance of a cell. To explain the idea of relevance, let's take the following example: consider a linear highway covered by 10 square cells as in figure 2. Assume that a new user, following the trajectory shown in figure 2, is requesting admission in cell number 0 and that the CAC process involves 5 cells. Responses from cells number 1, 2, 3 and 4 are relevant only if cell number 0 can accommodate the user. Similarly, responses from cells 2, 3 and 4 are relevant only if cell 1 can accommodate the new user when it hands off from cell 0. And the same principle applies to the other cells. This is because, in our scheme, a response from a cell is irrelevant if the user can not be supported until that cell. We write $\Phi_k(x)$ the relevance of cell k for user x.

The relevance depends only on the topology of the considered cellular network. For the linear highway

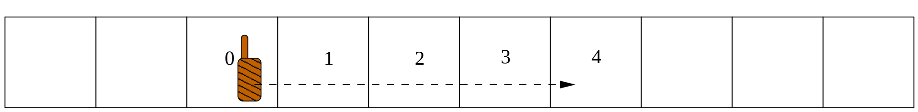

Figure 2. An example of a highway covered by 10 cells

example of figure 2, we propose the following relevance formula:

$$\Phi_0(x) = 1 \ and \ \Phi_k(x) = \prod_{l=1}^{k} \frac{(1 + R_{l-1}(x))}{2} \tag{4}$$

Note that for each $k \in K(x)$, $\Phi_k(x)$ is between 0 and 1. Note also that in eq. 4, cell j (the cell receiving the admission request) has the index 0 and that the other cells are indexed in an increasing order according to the user direction as in figure 2.

The cell, receiving the admission request, computes the sum of the product of $R_k(x)$, $W_k(x)$ and $\Phi_k(x)$ over k. The final decision of the call admission process for user x is based on:

$$D(x) = \frac{\sum_{k \in K(x)} R_k(x) \times W_k(x) \times \Phi_k(x)}{\sum_{k' \in K(x)} W_{k'}(x) \times \Phi_{k'}(x)} \tag{5}$$

Note that $-1 \leq D(x) \leq 1$ and that the denominator is never null, since the relevance, $\Phi_0(x)$, of cell j is always equal to 1, its weight $W_j(x)$ is strictly superior to 0, and all other $\Phi_{k'}(x)$ and $W_{k'}(x)$ are positive or null.

If $D(x)$ is higher than a certain threshold, we call acceptance threshold (T_{acc}), user x is accepted; the user is rejected otherwise. The higher is $D(x)$ the more likely the user connection will survive in the event of a handoff.

5.2. The Algorithm

Each time t, a cell j should decide if it can support new users. It decides locally if it can support users of type 1 and 2 that have higher priority than other types of user (cf. user types in section 4.1.1). This is because, from a user point of view, receiving a busy signal is more bearable than having a forced termination. The cell also sends the DMPs to other cells and informs them about its users of type 3. Only those who can be supported locally are included, other users of type 3 that can not be accommodated locally are rejected. At the same time, the cell receives DMPs from other cells and is informed about users of type 4.

Using eq. 2, the cell decides if it can support users of type 4 in the future and sends the responses to the corresponding cells. When it receives responses from the other cells concerning its users of type 3, it performs one of the two following steps: If the cell can not accommodate the call, the call is rejected. If the cell can accommodate the call, then the CAC decision depends on the value of $D(x)$.

6. Maintaining a Target Call Dropping Probability

In this section we explain how our algorithm varies the value of the acceptance threshold to maintain a target CDP value. We assume that each Mobile Switch Center (MSC), controlling a set of cells in the network, modifies the acceptance threshold of the cells it controls in order to maintain a target CDP. The following is the pseudo-code of the algorithm for adjusting the acceptance threshold, we will refer to as algorithm 1 in the remaining of the paper.

Algorithm 1 Pseudo-code for adjusting the acceptance threshold T_{acc}.

$w = \lceil \frac{1}{target\ CDP} \rceil$; $wobs = w$; $nA = 0$; $nD = 0$

if a user is accepted

{ $nA++$;

if ($nA \geq wobs$)

{if ($nD == wobs/w$) {$wobs = w$; $nA = 0$; $nD = 0$;}

else {$wobs+ = w$; if ($T_{acc} > -1.0$) $T_{acc}- = 0.01$;}}}

if a user is dropped

$nD++$;

if ($nD > wobs/w$) {$wobs+ = w$; if ($T_{acc} < 0.95$) $T_{acc}+ = 0.01$;}

The MSC begins by selecting a reference observation window w according to the target CDP as follows: $w = \lceil \frac{1}{target\ CDP} \rceil$. Note that we do not include the case where the target CDP is equal to zero, since this one is almost impossible to achieve and not practical from the provider point of view.

The variable representing the observation window $wobs$ is set to w, and the number of accepted users nA as well as the number of dropped users nD are set to zero.

If a new user is accepted in the system then nA is incremented by one. If we have observed at least

$wobs$ accepted users ($nA \geq wobs$) then, if the number of users dropped is equal to the maximum allowed dropping value, we set $wobs$ to w and set nA and nD to zero and restart from the beginning. If the number of users dropped is less than the maximum allowed, then we increase $wobs$ and decrease the acceptance threshold. This means that we will allow more users to be admitted in the system.

In case a user is dropped then nD is incremented by one. If the number of dropped users exceeds the maximum allowed value, then we increase $wobs$ and increase the acceptance threshold. This means that we increase our observation window and will allow less users to be admitted in the system.

Note that the proposed algorithm aims to achieve exactly the target CDP. This can easily be modified to let the actual CDP lay between a maximum and a minimum allowed values. Note also that the maximum allowed acceptance threshold is set to 0.95 in algorithm 1 in order for the network to accept a minimum number of users even if a congestion occurs.

7. Performance Evaluation

For simplicity, all the evaluations are done for mobile terminals that are traveling along a highway as in figure 2. This is a simple environment representing a one-dimensional cellular system. In our simulation study we have the following simulation parameters and assumptions:

1. The time is quantized in intervals $\tau = 10s$

2. The whole cellular system is composed of 10 linearly-arranged cells, laid at 1-km intervals. Cells are numbered from 1 to 10.

3. Cell 1 and 10 are connected so that the whole cellular system forms a ring architecture as assumed in [13]. This allows to avoid the uneven traffic load that will be experienced by cell 1 and 10 otherwise.

4. Connection requests are generated in each cell according to a Poisson process with rate λ (connections/second). A newly generated mobile terminal can appear anywhere in the cell with equal probability.

5. Mobile terminals can have speeds of: 70, 90, or 105 km/h. The probability of each speed is 1/3, and mobile terminals can travel in either of two directions with equal probability.

6. We consider three possible types of traffic: voice, data, and video. The probabilities associated with these types are R_{voice}, R_{data} and R_{video} respectively. The number of bandwidth units (BUs) required by each connection type is: $B_{voice} = 1$, $B_{data} = 5$, $B_{video} = 10$. Note that fixed bandwidth amounts are allocated to users for the sake of simplicity.

7. Connection lifetimes are exponentially-distributed with a mean value equal to 180 seconds.

For the Distributed Call Admission Control scheme we assume also that:

1. The DMPs are computed as in [6].

2. The relevance is computed using eq. 4.

3. The confidence degrees are computed using eq. 1 with $n = 3$.

Note: The offered load per cell, L, is defined as connection generation rate $\times$ connections' bandwidth $\times$ average connection lifetime, i.e.,

$L = (R_{voice}.B_{voice} + R_{data}.B_{data} + R_{video}.B_{video}).\lambda.180.$

In the following sections we will present the simulations we have done to evaluate the performance of our scheme. Two main sets of simulations were conducted. The first set of simulations, presented in section 8, evaluates the performance of our distributed call admission control scheme by comparing it to the Guard Channel scheme. The results prove that it is definitely advantageous to take neighboring cells' resource information into consideration in the call admission control process. The second set of simulations, presented in section 9, determines when it is worth involving several cells in the call admission control process.

8. Performance Evaluation: First set

We compare our algorithm (DCAC) with the guard channel (GC) scheme where each cell reserves a number of channels for exclusive use by handoff users [4].

We simulate a system that uses the GC scheme. We change the number of reserved channels (from 0% to 100% step 2%) for each simulation and we compute several important parameters like the Call Dropping Percentage, the Call Blocking Percentage and the Average Bandwidth Utilization. Also we simulate a system that uses our Distributed Call Admission Control scheme, we change the value of the acceptance threshold (from 0.4 to 0.7 step 0.01) for each simulation and we compute the same parameters. In this first set of simulations, since the acceptance threshold is fixed for every simulation, algorithm 1 is not used.

8.1. Simulation parameters

In this simulation set we have the following simulation parameters and assumptions:

1. Each cell has a fixed capacity of 50 bandwidth units (or channels).

For the distributed call admission control scheme we assume also that:

1. $m_x = 18$ for all users. This means that the DMPs are computed for 18 steps in the future.
2. The size of the cluster $K(x)$ is fixed for all users and is equal to 2. This means that one cell in the direction of the user along with the cell where the user resides form the cluster.
3. All users have the same acceptance threshold.

Knowing the average connection lifetime and bandwidth, we choose the connection generation rate to have a cell load of 100% and we simulate a total of 24 hours of real-time highway traffic for each simulation.

8.2. Simulation results

8.2.1. Data users only

In these first simulations we assume that only one class of service is used by all users. We compare the performance of DCAC and GC with only data users, which means that the probabilities associated with the three types of traffic are $R_{voice} = 0$, $R_{data} = 1$ and $R_{video} = 0$ respectively.

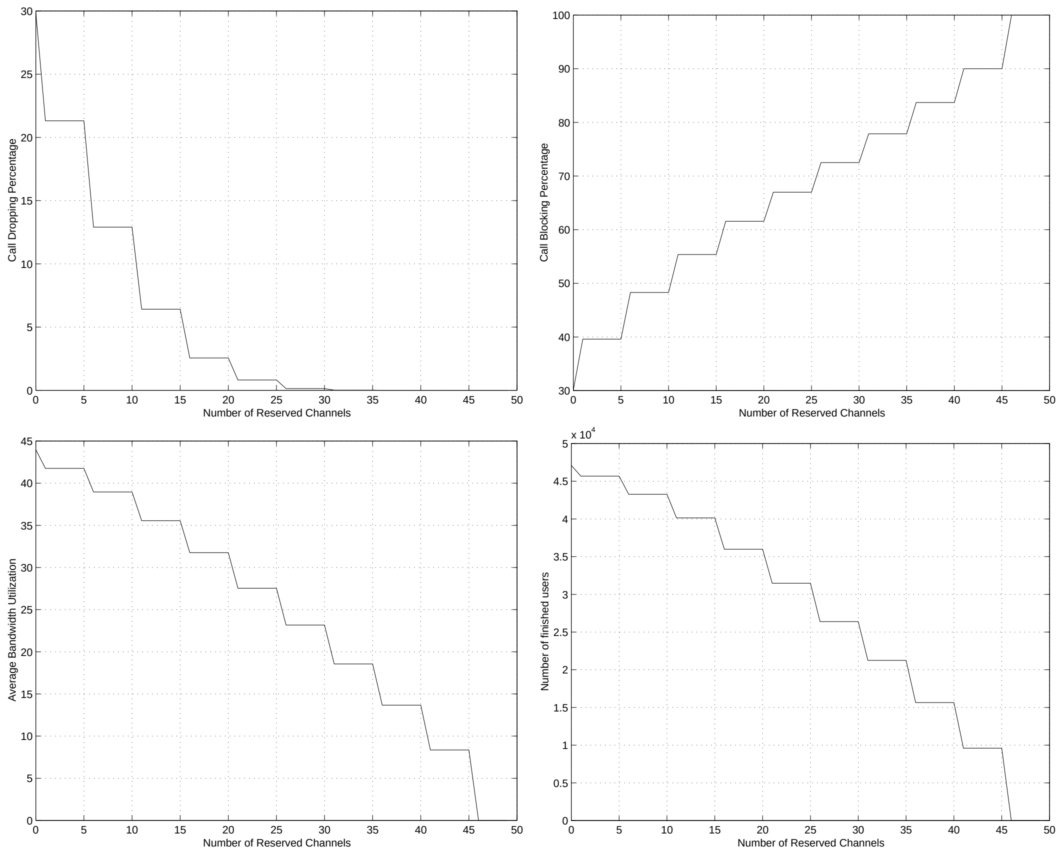

Figure 3. (3.a, 3.b, 3.c and 3.d) Simulation results for GC according to the number of reserved channels

Figure 3 depicts the performance results of the GC scheme. Figure 3.a presents the Call Dropping Percentage experienced by data users according to the number of reserved channels. As expected, when the number of reserved channels increases the CDP decreases. We can notice that since the number of bandwidth units used by data users is 5, the guard channel gives stepped curves with a step of 5. As indicated by figure 3.a some CDP values can not be obtained using the GC scheme. For example, using this scheme, and under the considered cell load, we can not achieve a CDP of 10%.

Figure 3.b depicts the Call Blocking Percentage experienced by data users according to the number of reserved channels. Of course, when the number of reserved channels increases the CBP increases also which affects the average number of resources used as depicted by figure 3.c.

Figure 3.d presents the number of finished data users which is the number of data users that were able to access the network and to finish their calls without being dropped. The number of finished data users decreases as the number of reserved channels increases.

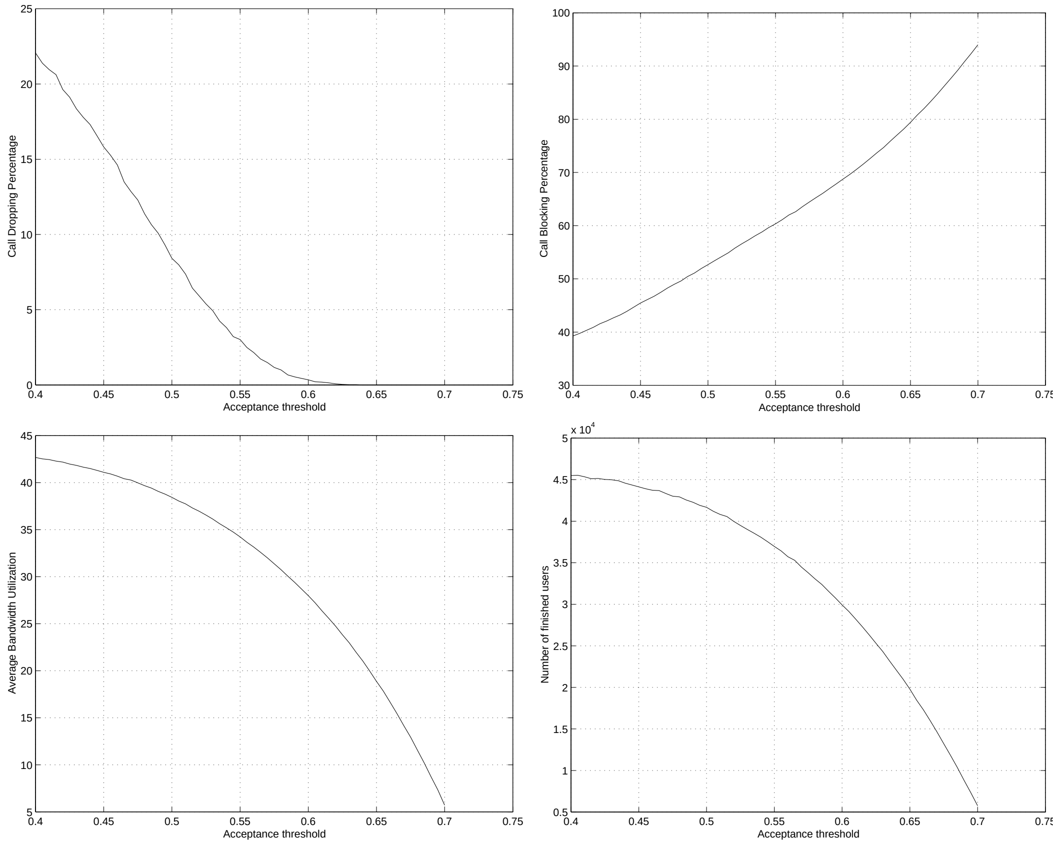

Figure 4. (4.a, 4.b, 4.c and 4.d) Simulation results for DCAC according to the acceptance threshold

Figure 4 depicts the performance results of the DCAC scheme. In this scheme only users that have a decision value (see eq. 5) above the acceptance threshold (T_{acc}) are admitted to the network.

Figure 4.a presents the Call Dropping Percentage experienced by data users according to the value of T_{acc}. As expected, when T_{acc} increases, a smaller number of users are admitted to the network and hence users incur lower CDP. In contrast with the GC scheme, the DCAC allows for any value of the CDP. Indeed, by choosing the right T_{acc} value, we can obtain any desired (and possible) CDP value.

Figure 4.b depicts the Call Blocking Percentage experienced by data users according to the value of T_{acc}. Of course, when the value of T_{acc} increases the CBP increases also. This is because we require a higher value of the decision results (i.e. $D(x)$). Figure 4.c shows as expected the decrease in the average bandwidth utilization incurred when the value of the acceptance threshold increases.

Figure 4.d presents the number of finished data users.

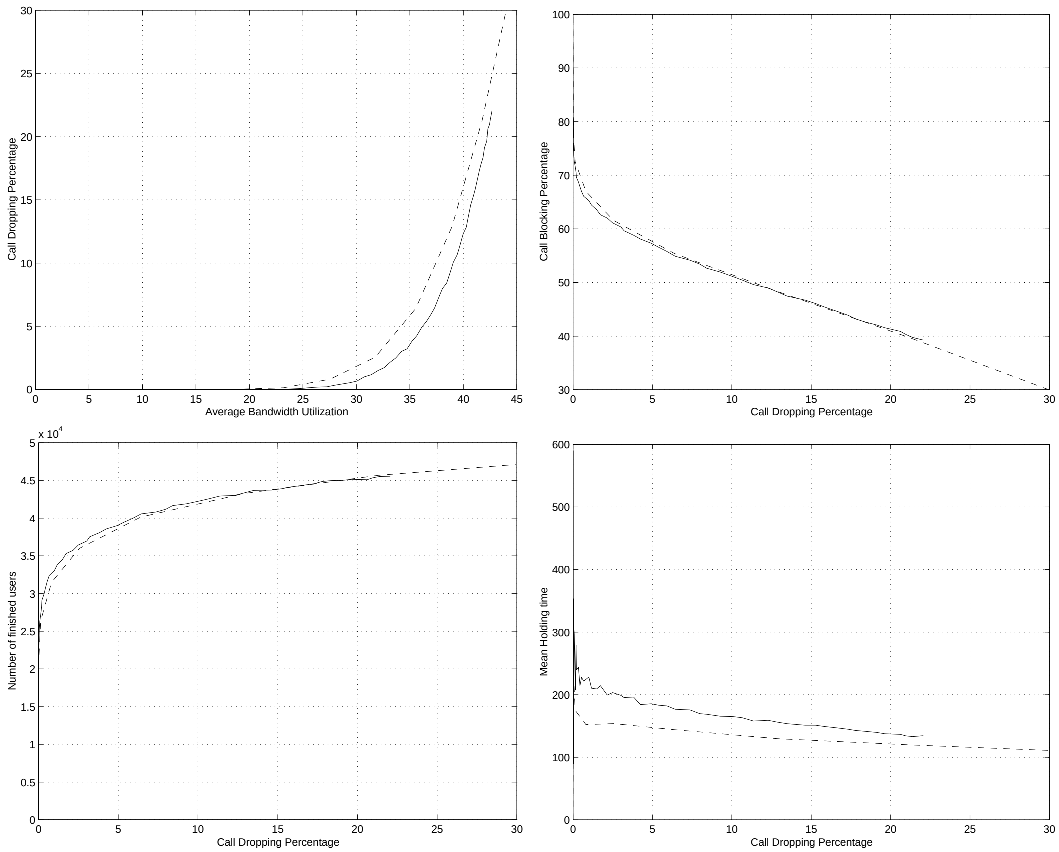

Figure 5. (5.a, 5.b, 5.c and 5.d) Comparing the two schemes (data traffic only)

Figure 5 presents the comparison results of the two schemes. Solid lines refer to the DCAC scheme and dotted lines refer to the GC scheme.

Figure 5.a depicts the Call Dropping Percentage experienced by data users according to the average bandwidth utilization achieved by the two schemes. As presented by figure 5.a, for the same achieved average bandwidth utilization the DCAC scheme allows for a lower CDP. In other words, the DCAC is able to achieve the same CDP while allowing a higher resource utilization. It is important to notice that there is always a tradeoff between the CDP, CBP and average resource utilization. So, we can not have the lower CDP and at the same time the higher resource utilization and a lower CBP. A good scheme is the one that optimizes one parameter without sacrificing the other parameters.

Figure 5.b presents the Call Blocking Percentage as a function of the Call Dropping Percentage. According to this figure, the two schemes achieve the same CBP for the same CDP, with an advantage

of DCAC over GC when the CDP is bellow 10%. It is worth noting that typical interesting CDP values are bellow 10%. Figure 5.c depicts the number of finished users as a function of the CDP. Here again, the DCAC scheme allows more users to finish their calls when the CDP is bellow 10%.

Figure 5.d presents the mean holding time of dropped users as a function of the CDP. The mean holding time presented in this figure is computed for only dropped users and reflects the mean holding time for users who were not able to finish their calls. As presented in this figure the DCAC is able to protect users in the network from being dropped. Not only the scheme is able to achieve better performance in terms of CDP and resource utilization but also to increase the mean holding time before being dropped. This means that all users with shorter holding time are saved from being dropped.

The obtained results of this comparison can be explained by the fact that the DCAC scheme is able to differentiate between those data users who are most likely to be dropped and those who are not. By denying access to those data users who have responses bellow the acceptance threshold even if the resources are available locally, the DCAC scheme protects the other users from being dropped. However, the GC scheme is unable to make this differentiation and data users with high dropping probability get access to the network. Those users consume resources which may prevent other users from surviving a handoff.

The improvement of using the DCAC when only data users are present is not very big however, results show that even with one traffic type the DCAC achieves better performance. In the next subsection, we compare the two schemes when the three types of traffic are present.

8.2.2. All classes of service

In these simulations we simulate a system where users can have one of three types of traffic; voice, data or video. The probabilities associated with these types are now $R_{voice} = 0.3$, $R_{data} = 0.4$ and $R_{video} = 0.3$.

Figure 6 presents the comparison results of the two schemes when the three types of traffic are present. Solid lines refer to the DCAC scheme and dotted lines refer to the GC scheme.

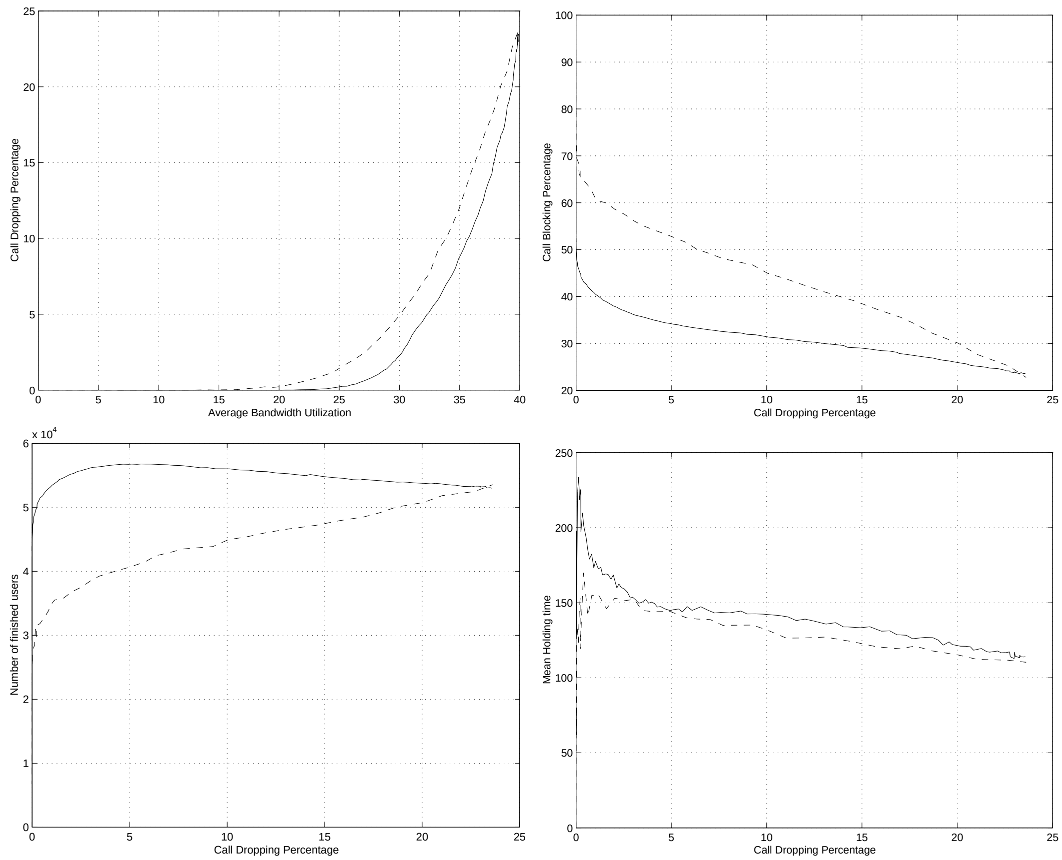

Figure 6. (6.a, 6.b, 6.c and 6.d) Comparing the two schemes (all classes of service)

Figure 6.a depicts the Call Dropping Percentage experienced by all users according to the average bandwidth utilization achieved by the two schemes. As presented by figure 6.a, for the same average bandwidth utilization the DCAC scheme allows for a lower CDP. The DCAC is able to achieve the same CDP while allowing a higher resource utilization. We can notice that the improvement in this case is higher than in the case of data traffic only (in figure 5.a).

Figure 6.b presents the Call Blocking Percentage as a function of the Call Dropping Percentage. According to this figure, the DCAC scheme achieves a very low CBP for the same CDP value. This is a very important result, since it shows the importance of differentiating between users in an admission control scheme. Being able to differentiate between users allows the DCAC to accept more users without sacrificing the CDP. In the same case, to achieve the same CDP the GC scheme has no choice but to block more users. This is better shown in figure 6.c which depicts the number of finished users as a function of the CDP for the two schemes. According to this figure, it is clear that the DCAC scheme

achieves better performance as a very high number of users are able to finish their calls without being dropped as opposed to the GC case.

Figure 6.d presents the mean holding time of dropped users as a function of the CDP. Here again, the DCAC scheme achieves better performance.

To better understand the behavior of the two schemes, let us have a look at the performance of the

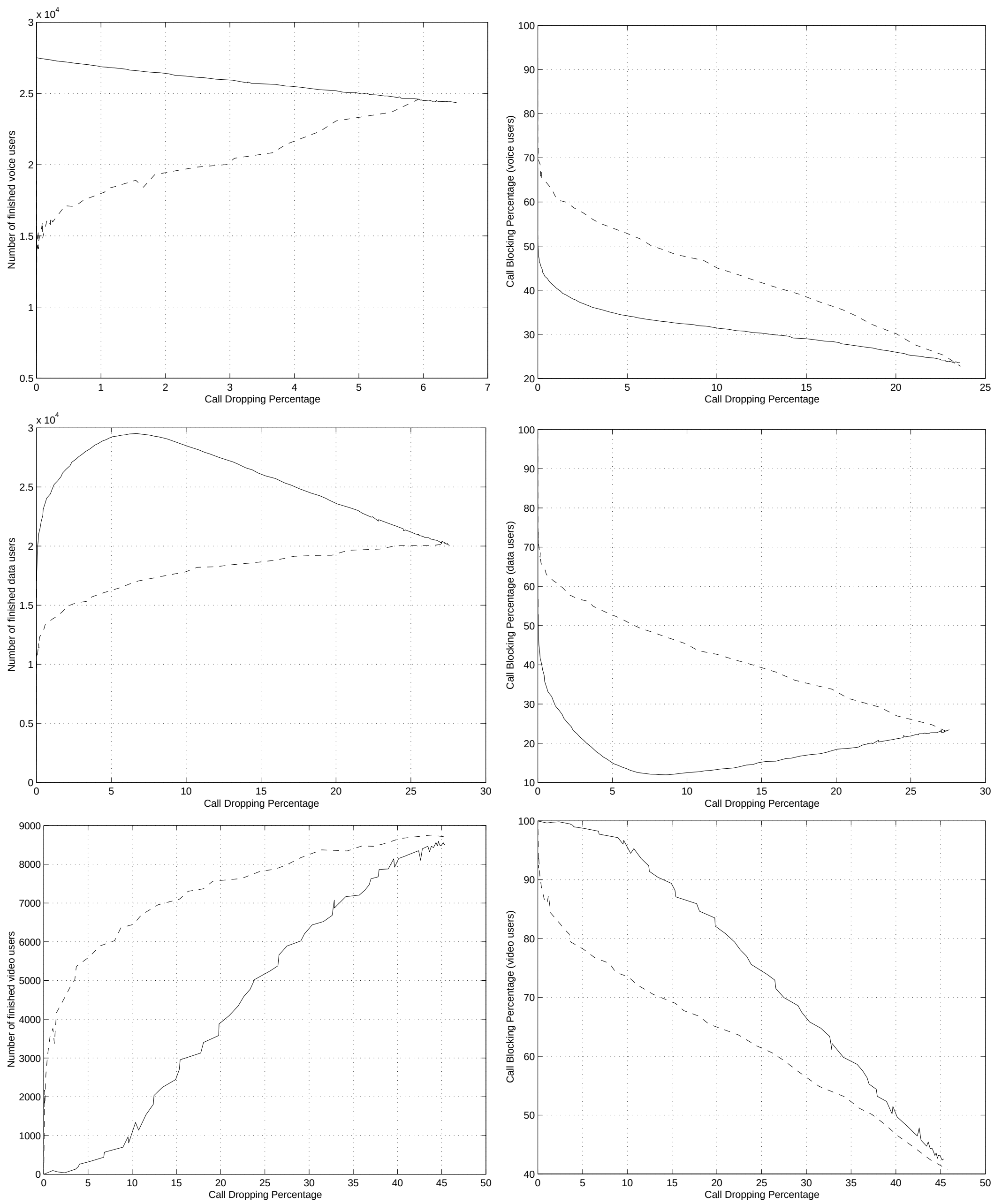

Figure 7. (7.a, 7.b, 7.c, 7.d, 7.e and 7.f) Comparing the two schemes (all classes of service)

two schemes with regards to the three type of traffic separately.

Figure 7 depicts the performance of the two schemes for each type of traffic. Figure 7.a presents the number of finished voice users according to the CDP. As the GC scheme requires a higher number of reserved channels to get a lower CDP, the number of finished voice users decreases while the CDP decreases. However, for the DCAC scheme, the number of finished voice users increases while the CDP decreases. At first sight, this result may seam abnormal since the number of finished voice users is expected to decrease as the CDP decreases (i.e. as the T_{acc} increases). However, as the T_{acc} increases, less data and video users are accepted to the network which gives voice users more chance to get accepted and to finish their calls without being dropped.

Figure 7.b depicts the Call Blocking Percentage experienced by voice users according to the CDP. As shown by this figure, voice users experience a lower CBP with DCAC than with GC for the same CDP.

Figure 7.c presents the number of finished data users according to the CDP. As expected, for the GC scheme, the number of finished data users decreases as the CDP decreases. However, the DCAC scheme presents an interesting behavior. As the CDP decreases, the number of finished data users increases and then decreases. The increase in the number of finished data users can be explained by the fact that as CDP decreases (i.e. T_{acc} increases), less video users are accepted to the network. This gives data users more chance to get accepted to the network and finish their calls. However, as the CDP gets smaller and smaller, the DCAC can no longer protect data users and will then block more data users from accessing the network which explains why the number of finished data users decreases when the CDP value gets bellow 7%. Figure 7.d depicts the Call Blocking Percentage experienced by data users according to the CDP. Here again, for the GC scheme, the CBP experienced by data users increases while CDP decreases. But for the DCAC scheme, the same behavior as presented in figure 7.c happens. The CBP for data users decreases as the CDP decreases. However, when the CDP reaches a certain value, no more data users can be accepted to the network without sacrificing voice users, so the CBP for data users start increasing while the CDP decreases.

Figure 7.e and 7.f presents the number of finished video users and the CBP for video users according

to the CDP respectively. These two figures show that the DCAC scheme is able to detect the congestion and start blocking video users that will most likely be dropped due to their high bandwidth demand (10 BUs). The GC scheme, on the other hand, will still allow more video users to access the network at the expense of a higher CDP value and at the expense of other voice and data users. It is worth noting that the simulations conducted assume a cell load of 100% which means that the network is congested.

Similar comparative simulations were conducted with a cell load of 50% and similar results were obtained.

9. Performance Evaluation: Second set

We evaluate here the DCAC scheme with different numbers of cells involved in the CAC process. We compare the performance of the scheme in the two following scenarios:

1. two cells are involved in the CAC process. This scheme will be referred to as **SC1**.
2. five cells are involved in the CAC process. This scheme will be referred to as **SC2**.

9.1. Simulation parameters

In this simulation set we use the following parameters and assumptions:

1. Each cell has a fixed capacity of 100 bandwidth units except cells 3, 4 and 5 which have 50, 30 and 50 bandwidth units respectively. This is to create a local congestion that will remain for a relatively long period. An example of such case is a temporary increase in the interference level which prevents the cells from using all their capacity.
2. $R_{voice} = 0.3$, $R_{data} = 0.4$ and $R_{video} = 0.3$.
3. We simulate a total of 10 hours of real-time highway traffic, with a constant cell load equal to 720 new calls/h/cell.
4. All users with a specific type of service have the same acceptance threshold. Algorithm 1 is used to adjust the acceptance threshold T_{acc} of all 10 cells and the target CDP is 10%. We assume that all

10 cells are under the control of one Mobile Switching Center. The accepted thresholds for voice, data and video users are set to $1.7 * T_{acc}$, $1.2 * T_{acc}$ and T_{acc} respectively. This is to achieve fairness between voice, data and video users. Indeed, if we use the same acceptance threshold for all users irrespective to their class of service, very few video users will be admitted to the network. This is because video users require more capacity than the other users, and hence it is more difficult to obtain high responses ($D(x)$).

The following additional simulation parameters are used for the **SC1** scheme:

- $m_x = 18$ for all users. This means that the DMPs are computed for 18 steps in the future.
- $K(x) = 2$ for all users. This means that one cell in the direction of the user and the cell where the user resides form the cluster.

For **SC2** scheme, the following additional simulation parameters are assumed:

- $m_x = 25$ for all users. This means that the DMPs are computed for 25 steps in the future.
- $K(x) = 5$ for all users. This means that four cells in the direction of the user and the cell where the user resides form the cluster.

9.2. Simulation results

In our simulations, a user x requesting a new connection is accepted into a cell only if the final decision $D(x)$ is above the acceptance threshold corresponding to the user class of service (voice, data or video). Figure 8 depicts the call dropping percentage achieved when using scheme **SC2**. The call dropping percentage represents the ratio of dropped users to the number of admitted users in the system. This is the aggregate call dropping percentage including all types of service. We can notice that algorithm 1 allows the actual CDP to approach the target CDP by varying the value of the acceptance threshold T_{acc}.

In figure 9.a, we compare the percentage of refused calls, given the offered load, when using scheme **SC1** and **SC2**. We can notice that **SC2** refuses less users than **SC1**. Indeed, **SC2** accepts about 8% more users than **SC1**. At a first sight, this result may seem abnormal. Indeed, scheme **SC2** involves

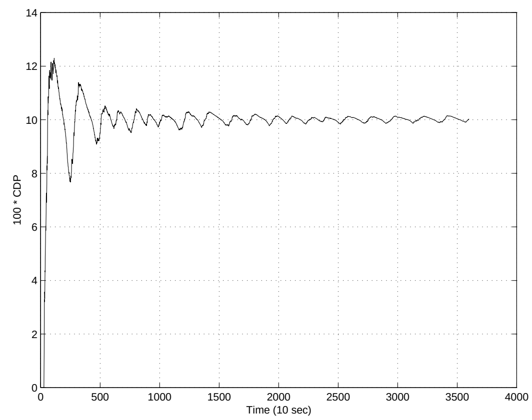

Figure 8. Call dropping percentage

five cells in the CAC decision process (3 more cells than **SC1**), and thus it is more difficult for a new user to be admitted by **SC2** than **SC1**. However, as we will see later in this section, **SC2** has the ability to avoid admitting those users who are most likely to be dropped and can use the saved bandwidth to accept more users who can most likely be supported.

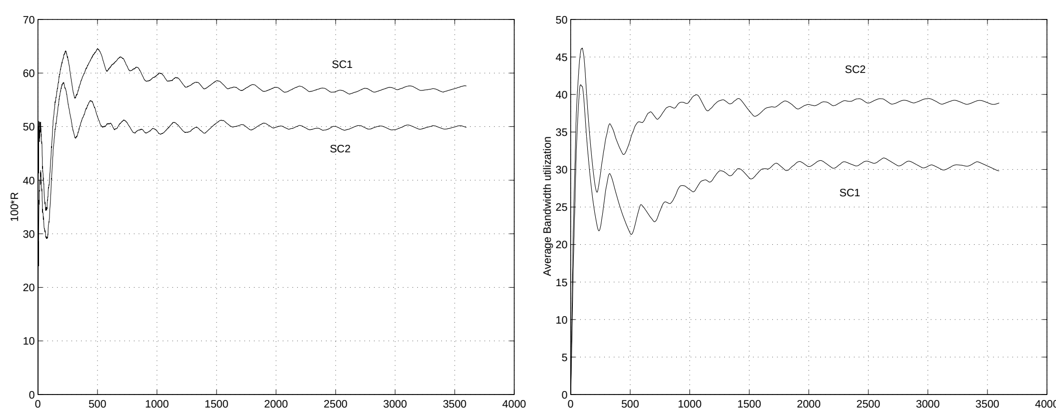

Figure 9. (a) Percentage of refused calls and (b) average bandwidth utilization

Figure 9.b shows that **SC2** not only accepts more users than **SC1** but also allows for a better resource utilization. In fact, **SC2** uses almost 10 bandwidth units more than **SC1**. It is worth noting that the low resource utilization experienced by the two schemes is due to the number of video users in the system and to the assumption that the whole system is controlled by one MSC. The latter assumption means that when a part of the network experience a congestion, the whole network is affected by refusing more

users (since the MSC increases the acceptance threshold for all the cells in the network). Although the simulated one MSC configuration is not likely to happen in the real-world, simulation results show the potential benefit of using scheme **SC2** compared to scheme **SC1**.

To further compare the two considered schemes, we compute the individual dropping percentage among the three considered classes of service, namely voice, data and video. The simulation results are shown in figure 10. In this figure, we can observe that the two schemes, **SC1** and **SC2**, achieve almost

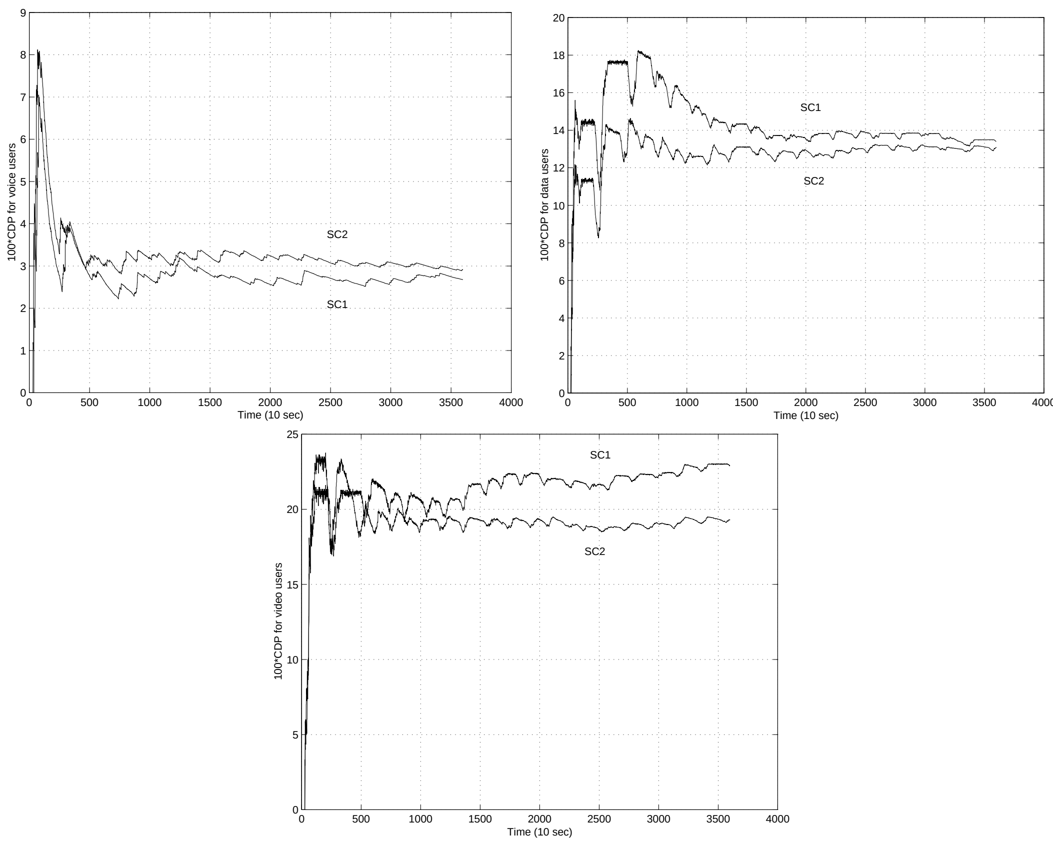

Figure 10. Percentage of dropped voice, data and video users

the same dropping percentage for voice and data users respectively, with a slightly better performance of **SC2** in case of data users. However, **SC2** drops almost 4% video users less than **SC1**.

As the percentage of dropped users depicted in figure 10 is computed according to the number of accepted users in each class of service, the comparison will not be fair if we do not observe the number of admitted users within each class of service for the two schemes. Figure 11 shows the percentage of refused calls within each class of service, and figure 12 plotted the number of accepted users within each class of service when using the two schemes.

According to figure 11, **SC2** refuses less users than **SC1** irrespective of users classes of service. This means that **SC2** accepts more users while achieving the same CDP in case of voice and data users, and that it allows more video users to be admitted to the network while achieving a lower CDP compared to **SC1**. According to figure 12, **SC2** accepts about 1500 video users more than **SC1** for the 10 real-time hours considered.

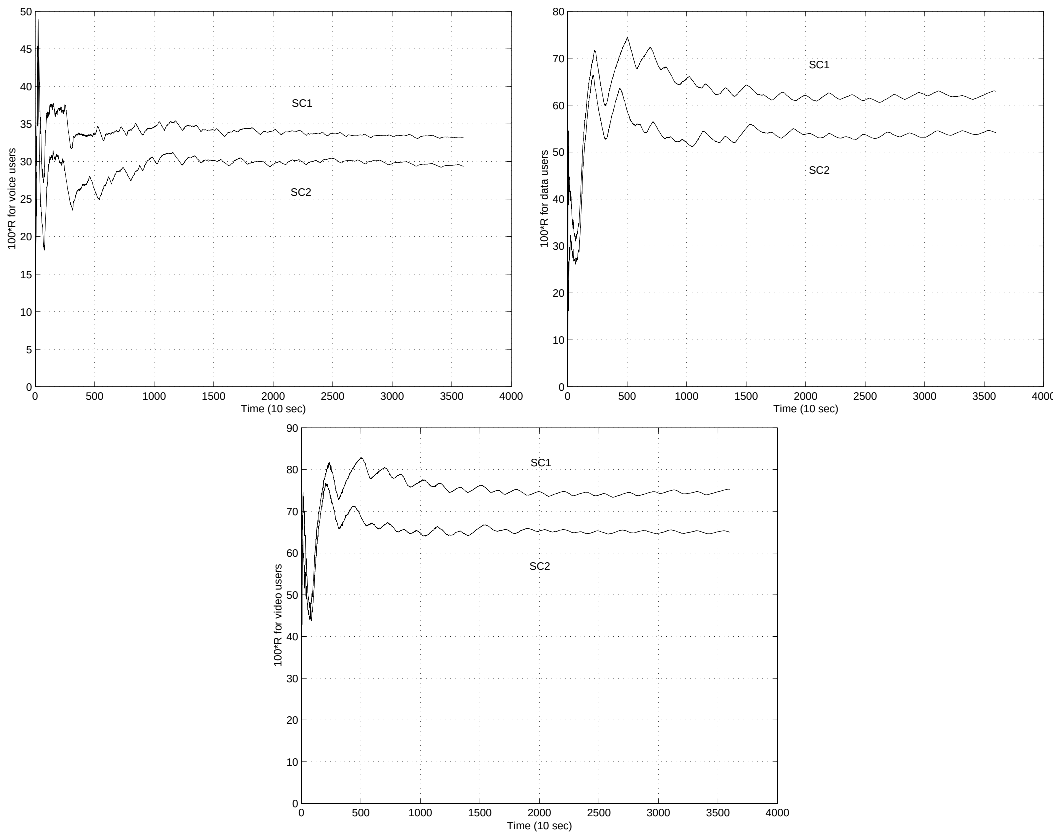

Figure 11. Percentage of refused voice, data and video calls

The bad performance achieved by **SC1** is explained by the fact that this scheme can not differentiate between those users who can be supported and those who can not. Its short sight prevents the scheme from being informed about a far congestion. Thus, the only way for **SC1** to reduce the CDP to the target value is to accept less users in the network, which results in a poor resource utilization.

On the other hand, since **SC2** involves more cells in the CAC process than **SC1**, the scheme is able to distinguish between those users who can be supported and those who are most likely to be dropped due

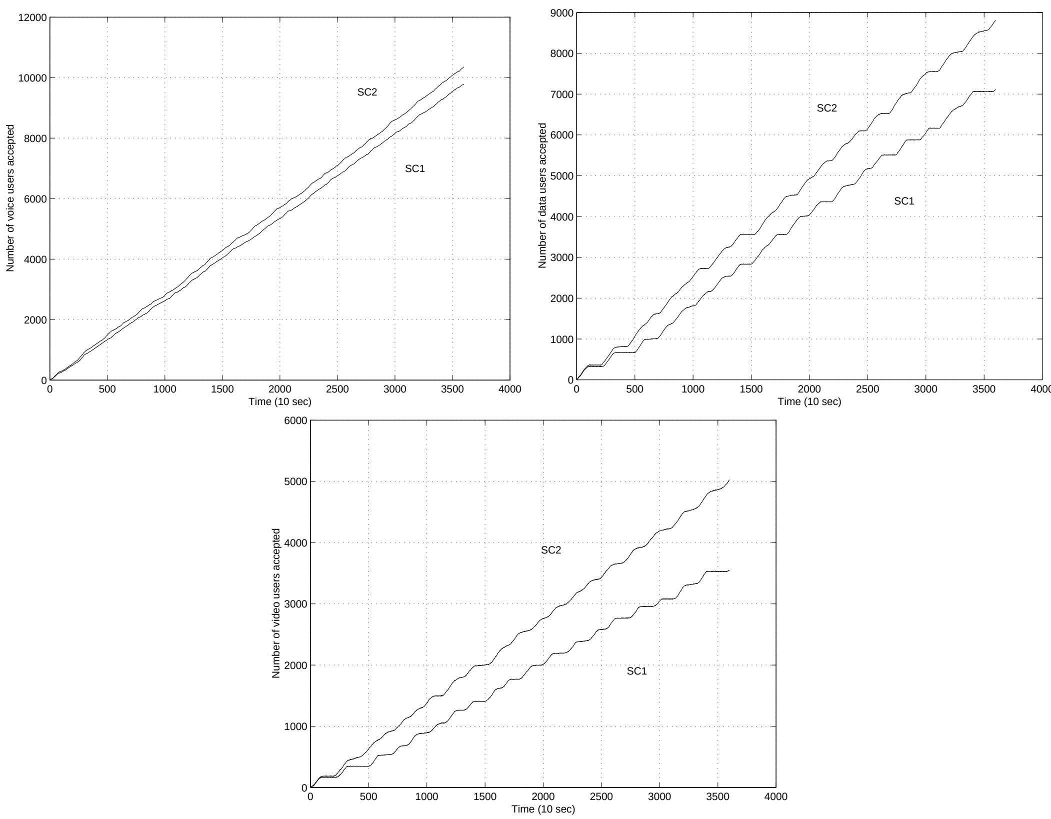

Figure 12. Number of accepted voice, data and video calls

to some congestion. This has the two following benefits: (1) the scheme can accept more users without sacrificing the CDP; (2) the bandwidth saved from not allowing some "bad" users to be admitted in the network, can be used to admit more "good" users.

We have conducted several other simulations with different offered loads and different simulation parameters (such as different mean holding time). The main observation worth highlighting here is that the two schemes **SC1** and **SC2** achieve almost the same performance in case of no congestion or in case of a uniformly distributed congestion. The latter case is less important since it can be solved off-line by increasing the network capacity. We have observed in the simulations presented in this paper, **SC2** achieves a better performance in case of a local congestion. The fact that the two schemes achieve the same results in case of a non congested network or in case of a uniformly distributed congestion is intuitively predictable. This is mainly because the responses from the three additional cells in **SC2** (cells 2, 3 and 4 in figure 2) only confirm what the two involved cells in **SC1** (cells 0, 1 in figure 2) have decided.

Of course, **SC2** does not have only advantages. As **SC2** involves more cells in the CAC decision process, it induces more communications between base stations and also requires more processing power than **SC1**. These resources are less critical compared to the wireless network bandwidth. A good compromise is to use **SC1** when the network is not congested and use **SC2** when a congestion is detected. The process of selecting the good scheme is out of the scope of this paper and is subject to future work.

10. Conclusion

We have described a call admission control scheme suitable for wireless multimedia networks. The proposed scheme operates in a distributed fashion by involving, in a call admission decision, not only the cell where the call originated, but also a determined number of neighboring cells. This distributed process allows the original cell to make more clear-sighted admission decisions.

The goals underlying the design of our algorithm are: (1) to support mobile multimedia users with dynamic bandwidth requirements; (2) to reduce the call dropping probability while maintaining a high network resource utilization; and (3) to distribute call admission control among clusters of neighboring cells to allow more clear-sighted decisions and hence a better user survivability in the network. Previous works have addressed these issues separately which makes this work, in pursuing these goals simultaneously, of a pioneering nature. More technically, our algorithm can integrate easily any method for computing Dynamic Mobile Probabilities (DMPs). It can also rely on different local call admission control schemes including those designed for adaptive multimedia applications. It is worth noting that the signaling load induced by the algorithm operation is considered here acceptable as far as it only involves few messages exchanged between base stations through the wired network which is assumed to be of high capacity.

Simulations results have shown a significant improvement of the distributed CAC over the Guard Channel scheme. By implementing the distributed call admission control scheme, the system is able to lower the call dropping probability while offering a high resource utilization by selecting only those users that are more likely to complete their call without being dropped.

We also presented an analysis of the comparison between two call admission control schemes involving different number of cells in the decision process. We have observed that it is worth involving more cells in the CAC decision in case of local congestion. This allows the scheme to take a more clear-sighted admission decision for new users, hence, achieving better resource management and quality of service. We have compared the performance of our Distributed Call Admission Control scheme when involving two and five cells in the admission decision respectively. However, our scheme can involve any number of cells. We have demonstrated that in some cases it is worth involving several cells in the admission process. The choice of the number of cells to involve and when this should happen is an important issue that will be addressed in the future.

References

[1] Universal Mobile Telecommunications System (UMTS), http://www.etsi.org/umts/

[2] Broadband Radio Access Networks (BRAN), http://www.etsi.org/bran/

[3] International Mobile Telecommunications (IMT2000), http://www.itu.int/imt/

[4] D. Hong and S. Rappaport, 'Traffic Modeling and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures,' *IEEE Trans. Vehicular Tech.*, vol. 35, pp. 77-92, 1986.

[5] J. Evans and D. Everitt. Effective bandwidth based admission control for multiservice cdma cellular networks. *IEEE Trans. Vehicular Tech.*, 48(1):36–46, 1999.

[6] D. A. Levine, I. F. Akyildz, and M. Naghshineh. A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept. *IEEE/ACM Trans. on Networking*, 5(1), 1997.

[7] S. Lu and V. Bharghavan. Adaptive resource management algorithms for indoor mobile computing environments. In *ACM SIGCOMM'96 Proceedings*, pages 231–242, 1996.

[8] S. Choi and K. G. Shin. Predictive and adaptive bandwidth reservation for hand-offs in qos-sensitive cellular networks. In *ACM SIGCOMM'98 Proceedings*, pages 155–166, 1998.

[9] D. A. Levine, I. F. Akyildz, and M. Naghshineh. The shadow cluster concept for resource allocation and call admission in atm-based wireless networks. In *ACM MOBICOM'95 Proceedings*, pages 142–150, 1995.

[10] R. Rejaie, M. Handley, and D. Estrin. Quality adaptation for congestion controlled video playback over the internet. In *ACM SIGCOMM'99 Proceedings*, 1999.

[11] S. McCanne, M. Vetterli, and V. Jacobson. Low-complexity video coding for receiver-driven layerd multicast. *IEEE JSAC*, 15(6):983–1001, 1997.

[12] J. Hartung, A. Jacquin, J. Pawlyk, and K. Shipley. A real-time scalable video codec for collaborative applications over packet networks. In *ACM Multimedia'98 Proceedings*, pages 419–426, 1998.

[13] M. Naghshineh and M. Schwartz, 'Distributed Call Admission Control in Mobile/Wireless Networks.' *IEEE JSAC*, Vol. 14, No. 4, pp. 711-717, 1996.