A Call Admission Control protocol for Multimedia Cellular Networks

Ayman Elnaggar
Dept. of Electrical and Computer Engineering
Sultan Qaboos University
Email: ayman@squ.edu.om

Mokhtar Aboelaze Maan Musleh
Dept. of Computer Science and Engineering
York University
Email: {aboelaze, maan}@cse.yorku.ca

Abstract—**The performance of any cellular wireless network, as well as its revenue (number of customers using the network, and their degree of satisfaction) is determined to a great extent by its call admission control (CAC) protocol. As its name implies, the CAC determine if a new call request is granted, or rejected. In this paper, we propose a call admission control protocol for cellular multimedia wireless networks. Multimedia networks are characterized by a wide variety of bandwidth requests, priorities, and drop-off/rejection requirements by different customers. Our protocol depends on degrading the existing calls, according to their degradation priority, by reducing the bandwidth allocated to them in order to admit new calls according to their admission priority. Our protocol is too complicated for an analytical solution. However we present a Markov Model of a simplified version of our protocol for completeness, a Markov representation of the protocol is too complicated to be of any real value. Extensive simulation results show how our proposed protocol can improve the drop-off/rejection ratio for large bandwidth calls and at the same time maintain the quality of service requested by important calls.**

Index Terms—**call admission protocols; cellular networks; QoS; call degradation.**

I. Introduction

Wireless devices and connectivity through wireless networks are growing at an astonishing rate. Wireless networks are not only used for cellular phone applications, they are carrying different types of traffic, voice, video, and data [4]. Wireless networks range in coverage from desktop area networks to national cellular networks. Users of such networks expect a specific Quality of Service (QoS) depending on their application and service contracts. The network must guarantee the QoS requested by each user, and in the same time maximize its revenue by maximizing the number of users (calls) admitted to the network.

Admission control policies play a crucial role in the performance of any network. When deciding to admit a call to the network there are many factors to be taken into consideration most of them are contradictory (network utilization, revenue, QoS, fairness, …). The call admission control works in real-time; the algorithm used should be suitable for real-time implementation using the limited resources of the base station controller [16]. There has been a lot of research in call admission protocols that could be summarized as follows.

FCFS is probably the simplest of all call admission protocols, in this method if a call arrives and there is enough bandwidth to accommodate it, it is accepted, otherwise rejected. This method has shown to produce a good utilization of the bandwidth however it has been shown to be biased against calls with high bandwidth requirements. As a way of introducing some priority in FCFS, bandwidth is divided into segments and call requests are grouped into different categories such that a call request from category *i* can only be admitted if there is enough bandwidth in segment *i* otherwise rejected. The main problem with this technique is the waste of bandwidth since we could have enough bandwidth in one segments but a call that belongs to another segment is rejected.

In [6], the authors proposed a general framework for bandwidth degradation and call admission in a multi-class traffic wireless network. Their objective is to model the changes in the revenue of the network due to admitting new users; in the same time they estimated the cost of degrading an ongoing call and considered it to be negative revenue. The overall objective is to maximize the revenue. In their analysis they considered an exponential service time for the calls, and a Poisson arrival pattern. They did not differentiate between new calls and handoff calls; neither had they considered the effect of handoff at all.

In [18], the authors used the direction of the movement to predict the next cell and make an early reservation before the actual handoff occurs. They used both threshold distance and threshold time for early reservation. They assumed a real-time positioning technology in order to determine the position and the direction of movement of the user. They also used channel borrowing in the sense of migrating channels from cold cells (cells with light load) to hot cells (cells with high load), within constraints, not in the sense of borrowing bandwidth from existing calls as we use in this work.

El-Kadi et al in [7] proposed a new call admission scheme. Their protocol depends on dividing the connections into real-time, and data connections. The real time connections are guaranteed at least a minimum bandwidth. Where the data connections are assumed to tolerate a large delay and there is no guaranteed minimum bandwidth. In their protocol, a handoff call is admitted if there is enough B.W. in the cell, or if the minimum required bandwidth for this call could be achieved by borrowing from other connections in the cell. However, in their protocol they did not support any priority schemes either for the admitted or degraded calls.

In [11] the authors proposed a 2-level call admission scheme. In this protocol, the protocol is divided into 2 parts, Basic Call Admission Control (BCAC), and Advanced Call Admission Control (ACAC). The BCAC determine the admission based on the availability of B.W. The ACAC determine the call admission by utilizing the delay tolerance and priority queue algorithm, Depending on the type of the call blocked by the BCAC.

In [12] the authors proposed a novel but rather simple protocol in order to improve the fairness, they proposed the use of a single buffer in order to hold the call request if there is not enough bandwidth. If a call arrives and there is enough bandwidth to admit it, it is admitted right away. If there is not enough bandwidth, and the buffer is free, the call is admitted to the buffer and waits until there is enough bandwidth, then it is admitted and the buffer is cleared. If a call arrives and another call is waiting in the buffer, the new call is rejected. They proved that under a Poisson arrival and exponential service time, their protocol is optimally fair. However, one of the major drawbacks of their protocol is that it leads to a drop of utilization when there is a big difference in the bandwidth requirements of the different classes. Their protocol does not support different priorities levels for different users.

In [1], the author proposed a call admission protocol that can support differentiated fairness and maintain good resource utilization for calls with different and widely varied bandwidth requirements. However the author assumed a fixed bandwidth for every call and no call degradation is allowed in order to admit new calls.

In [3] the authors proposed a borrowing based CAC protocol. Their protocol depends on dividing the bandwidth into a reserved part for every call class, and a shared part. By controlling the reserved part of the bandwidth, they can control the call blocking and call-dropping ratio. They used attribute-measurement mechanism in order to dynamically adjust the reserved bandwidth. The main drawback of their technique is the overhead due to the measurement and update of the reserved part of the bandwidth.

[17] Proposes a dynamic auction-based scheme in order to allow the users to negotiate the required service level with the service provider. Their objective was fairness among users while maximizing the service provider revenue. In [10] the authors proposed an efficient resource allocation scheme for TCP services over IEEE802.16 networks. Their protocol estimates the uplink traffic of the subscriber station based on the downlink traffic to the subscriber station and accordingly allocates resources to the subscriber stations.

In [15] the authors proposed a scheme for crosslayer packet scheduling for video over downlink packet access network. Their scheme is suitable for IEEE802.16 wireless network. They used a utility function that dynamically controls the user data rate based on the channel quality.

In [2] the author proposed a scheme for borrowing bandwidth from existing calls in order to admit new calls. They showed that their scheme outperforms previous schemes in terms of call dropping probability and call blocking probability. However, they considered only two types of traffic and they did not include priority.

Call admission protocols that take into account the user velocity and the direction of movement in allocating bandwidth not only in the current cells but in the cell the user is moving into are studied in [8] and [9].

In this paper, we propose a borrowing based call admission control protocol for multimedia based wireless networks. Our protocol supports priority in both admission and borrowing. It does not perform any network measurement thus reducing the overhead experienced by the protocol and making it simple to implement in real time. Then we present a Markov model for our protocol. Our protocol is too complex to solve analytically, but a Markov model will be presented for our protocol under some simplifying conditions. We use extensive simulation in order to measure the performance of our protocol and compare it with similar protocols.

The organization of our paper as follows, Section II states the model we use in our simulation for both the network and the traffic. Section III explains our protocol. Section IV presents a Markov model for a simplified version of our proposed protocol. In section V we used extensive simulation to show the performance of our proposed protocol under different load and traffic assumptions; the paper ends with a conclusion and future work.

II. THE MODEL

In this section, we present the network model and the traffic model we used throughout the rest of the paper.

A. Network Model

We assume a cellular network architecture where every cell is served by a base station. Base stations are connected together by using a wireless or wireline network. Users are roaming in the coverage area and when moving from a cell to another cell, handoff occurs. The call admission control protocol in the base station is responsible for deciding whether to admit or to block a new or a handoff request. The call admission control protocol also determines if we borrow bandwidth from existing calls or not, and if yes, from what call(s).

Our protocol works equally well with wireless (non-cellular) networks where there is a central node that assign bandwidth to the different users based on requests. In our protocol there are no specific assumptions that are unique to cellular networks; we just assume that there is a "central" node that is responsible for admission and assigning bandwidth to the different users (including changing the bandwidth assigned to the different users).

When (in a cellular network) the network is congested and the CAC decides to turn down an admission request that is called a blocked call. However, if the admission request is coming from an active call in a neighboring cell that is moving into the cell's coverage area, that is called a dropped call. From the customer point of view, blocking a call is much more tolerable than dropping an active call. The admission control protocol must take this into consideration.

In this paper we did not assume any specific transmission technology such as TDM, FDM, or CDMA. The only assumption we made is that the user can receive a variable (discrete set of possible) bandwidth(s). Thus our protocol is completely independent of the underlying technology used for transmission.

In our simulation, we chose not to separately and explicitly consider the handoff calls. The reason is the rate of arrival of handoff requests to any cell depends on the cell size, and the speed of the mobile user. Although this is not difficult to simulate, however it will limit the usefulness of the results to that particular scenario. Instead, we chose to represent the handoff requests as a separate class of requests. The handoff class has the same bandwidth requirements and service rate as the corresponding calls originating in the cell, however, the handoff connection does have a higher priority than a similar type connection originating in the cell. Thus, we can control the call dropping and call blocking ratios.

We assume a Poisson arrival for the incoming calls. The service time for every call is exponentially distributed. Different classes of calls with different priorities, arrival rates, service times, and bandwidth requirements are considered. The call blocking ratio for each class and the resource utilization (network utilization) are the main parameters of interest. We also look at the distribution of the cell bandwidth among the different classes of calls. Since we use simulation to show the performance of our proposed protocol, changing the distribution of the arrival pattern and service time could be achieved very easily.

B. Traffic Model

We assume different types of traffic (audio, video and data). However for every type, we assume different classes or categories of users. For example consider a mobile terminal transmitting video; the quality of service offered to the terminal depends on the price paid for the service and can range over a wide range affecting the quality of the received video. The different types of traffic are characterized by different arrival rates, service rates, and bandwidth requirements.

Many applications especially audio and video transmission can support a variable bit rate. For example, MPEG-4 supports very low bit rate coding with bandwidth requirements of 5-64Kbps [4], while in Audio, using silence detection and sophisticated coding techniques [5] results in encoding that supports variable bit rate.

For data transmission such as file transfer, web browsing, and text messaging packet transmission is usually used (compared to circuit or virtual circuit for interactive audio/video transmission). That allows us to arbitrarily control the bandwidth according to network situation (by controlling the number of slots or packets assigned to any user in a specific time unit).

In this paper, we assume that every class has a maximum (requested) bandwidth, and a minimum bandwidth. The assumption is that, this class of traffic can supports degraded performance down to the minimum bandwidth. The network can, if needed, borrow some bandwidth from any user with the condition of not violating its minimum bandwidth requirements. Constant bit rate sources are a special case where the minimum bandwidth is the same as the maximum bandwidth

III. PROPOSED PROTOCOL

We assume *m* classes of users, such that users in class *i* require a maximum bandwidth of $B^i_{\max}$, and a minimum bandwidth of $B^i_{\min}$ for $1 \le i \le m$ the difference between these two values is the degradation that a user in class *i* can tolerate and is defined as the *degradable bandwidth* for class *i*. The degradable bandwidth ($B^i_{\max} - B^i_{\min}$) for a call in class *i* is divided into β_i segments; these segments may or may not be of equal values, where β_i determine the granularity of the bandwidth borrowed from a connection in class *i*. It also determines the priority of borrowing from class *i* connections, as we will see in the remaining part of this section. A call is said to be in level *s*, if there are *s* segments borrowed from it. The network borrows segments from ongoing calls in a way such that the difference between the states of any two calls that did not exhaust all their degradable bandwidth is at most 1. A network is said to be in state *s*, if each node that did not exhaust its degradable bandwidth is either level s or *s-1*, and all the calls that have already exhausted their degradable bandwidth do have a number of segment in their degradable bandwidth equal to $s_X \le$ s segments. Figure 1 shows the bandwidth requirements for 4 classes. Every class has its own $B^i_{\max}$, $B^i_{\min}$, and β_i.

Note that β_i also determines the priority of class *i* as far as borrowing bandwidth is concerned. Since the network borrows bandwidth from users in such a way that the difference between any 2 levels in the network, that did not exhaust their degradable bandwidth, is at most 1, which means that if the network borrowed from a call in class 1 a segment; it cannot borrow another segment from the same call unless it borrows a segment from all other calls in progress. That means the CAC protocol will

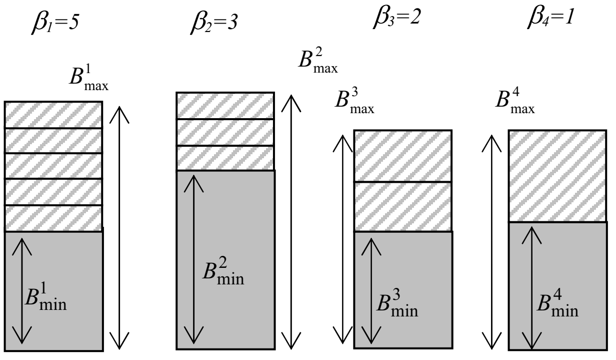


Figure 1 Bandwidth requirements for different types of traffic

borrow a segment from class 4, (The entire degradable bandwidth of class 4) before it borrows another segment from a call in class 1. Thus, giving class 1 a higher priority than class 4 (for bandwidth degradation purpose).

A. Prioruty

The proposed protocol also supports priority in call admission. Every traffic class has a priority tuple $< P_i, \rho_i >$. Where P_i is the *admission threshold*, and ρ_i is the *admission probability*

These two parameters determine the probability of admitting the call or rejecting it. If the network is in state $s \le P_i$, then the call is admitted (provided there is enough BW, or BW could be freed by degrading the existing calls and the state of the network stays less than or equal to P_i after the borrowing is completed). On the other hand, if $s > P_i$, the call is admitted with a probability ρ_i and is rejected with a probability $1-\rho_i$.

Our protocol has the advantage of having priority levels for different callers by treating them differently both in admitting them or not (based on the network state), and by deciding to borrow bandwidth from them. Low priority calls are not admitted if the network is beyond a specific state, also for high priority calls, we can increase the number of segments which leads to a smaller segment size, and allowing borrowing is smaller quantities from high priority callers. That leads to a situation that if the network is saturated beyond a specific threshold, we can be sure that we will borrow from high priority calls only to admit another high priority call.

B. Handoff calls

We opted not to deal with handoff calls any differently than new calls. In reality, handoff calls usually have higher priority than new calls. Customers consider getting a busy signal to be less of a nuisance than a call is terminated abruptly just because the customer moved to another cell. Thus handoff calls are usually given a higher priority compared to new calls. However, in this protocol we do have priority levels that are assigned to each category, in this case, handoff calls are considered to have a higher priority than new calls, thus giving them a preferential treatment when it comes to accepting the call or not. There is no reason for handoff calls to have a preferential treatment in bandwidth degradation or bandwidth borrowing compared to a new call originating in the cell, so they should have the same number of borrowable segments.

C. Protocol

In this section, we formally describe our proposed protocol using pseudo code.

```
L      The network level
$P_i$     Admission threshold for class I
$\rho_i$     Admission probability for class i
Case (event)  Arrival
IF $P_i$ < L reject with a probability (1-$\rho_i$)
ELSEIF There is enough BW  admit it
Else Find BW by degrading existing calls
       IF the new network level L'>$P_i$
       Reject with probability (1-$\rho_i$) and accept
       with probability $\rho_i$
       Else Reject /*could not free enough BW*/
CASE (event) departure
       If L ≠ 0 Redistribute the freed BW
```

IV Markov Model

In this section, we develop the Markov model for a simplified version of our proposed protocol. Our proposed protocol is extremely complex and is not suitable for a closed form solution using Markov chain. However, we present the Markov model of the protocol for completeness without solution. We assume the following:

- We assume there is no priority in admission, that is $\rho_i=1 \;\forall\, i$. That is to say if there is enough bandwidth, or enough bandwidth could be freed by borrowing, the call is admitted.
- The degradable bandwidth is divided into equal sized segments. That is a call in state *s* is assigned a bandwidth of $B^i_{\max} - s\delta_i$ where $\delta_i = \dfrac{B^i_{\max} - B^i_{\min}}{\beta_i}$
- $\forall i, j \; \min_i(B^i_{\min}) > \max_j(\delta_j)$ This assumption simply states that if the network is in state *s*>0, no call can be admitted (even with its minimum bandwidth without borrowing).

These assumptions are not necessary for the establishment of the model, but rather for the tractability of the equations governing the state transition of the model. The Markov model is *3m*-Dimensional Markov chain where *m* is the number of different call categories in the system.

States are represented by a *3m* tuple as follows: $N_1, N_2, \cdots, N_m, s_1, s_2, \cdots s_m, \alpha_1, \alpha_2, \cdots \alpha_m$, where

N_j= Number of calls of the j^{th} category in the system

s_j is the state of the j^{th} category(every call in category j is either in state s_j or state s_j*-1*)

α_j is the number of calls in category *j* in state s_j i.e. (N_j-α_j) calls in state s_j*-1.*

A call in state *s* in group *i* has an allocated BW of

$$bw(s_i) = \begin{cases} B^i_{\max} - s_i\delta_i & s_i \ge 0 \\ B^i_{\max} & s_i < 0 \end{cases} \qquad (1)$$

A condition on the set of legitimate states is:

$$\begin{array}{l} \left(\sum_{j=1}^{m} \left[\alpha_j bw(s_j) + (N_j - \alpha_j) bw(s_j - 1) \right] \le C \right) \wedge \\ \left(\begin{array}{l} if\ \exists s_j \ne 0 \Rightarrow \\ C - \sum_{j=1}^{m} \left[\alpha_j bw(s_j) + (N_j - \alpha_j) bw(s_j - 1) \right] < \delta_i \forall i \end{array} \right) \\ \wedge \left(if\ s_i = 0 \Rightarrow \alpha_i = N_i \right) \wedge \left(\alpha_i \le N_i \forall i \right) \end{array} \qquad (2)$$

Where *C* is the cell bandwidth.

The first equation simply states that the assigned bandwidth is less than the total cell capacity, while the second equation indicates that we will not borrow bandwidth and degrade connections without assigning the borrowed bandwidth to some call, and if a call is terminated, the freed bandwidth will be returned to the calls it was borrowed from if any. The third condition states that if there were no borrowing, then all the N_j calls in category *j* are in state 0. The fourth condition states that the number of calls in state s_i is less than or equal the total number of calls of category *i*.

The transition probability from state *S* to state *S'* where

$S = N_1, \cdots, .N_i, \cdots N_m, s_1, \cdots s_i, \cdots s_m, \alpha_1, \cdots \alpha_i, \cdots \alpha_m$

$S' = N'_1, \cdots, N'_i, \cdots, N'_m, s'_1, \cdots, s'_i, \cdots s'_m, \alpha'_1, \cdots, \alpha'_i, \cdots \alpha'_m$

is *P(S, S')*.

The transition probability depends on the event (arrival or departure) and the state of the network at the time. For arrivals, the call could be accepted without borrowing, accepted with borrowing, or rejected. For departure, the borrowed bandwidth (if any) should be returned back to the existing calls. We assume that λ_i is the arrival rate for category *i* and μ_i is the departure rate for category *i*

Before stating the transition probability between the states in The Markov chain, we will briefly describe conditions on the transition between the states in case of an arriving call request.

If the network is in state 0 and an arriving call from category *i* is accepted without borrowing, then.

$S = N_1, \cdots, N_i, \cdots N_m, 0, \cdots 0, \cdots 0, N_1, \cdots N_i, \cdots N_m$

$S' = N_1, \cdots, N_i + 1, \cdots N_m, 0, \cdots 0, \cdots 0, N_1, \cdots N_i + 1, \cdots N_m$

If an arriving call from category i is accepted with borrowing, then

$S = N_1, \cdots, N_i, \cdots N_m, s_1, \cdots s_i, \cdots s_m, \alpha_1, \cdots \alpha_i, \cdots \alpha_m$

$S' = N_1, \cdots, N_i + 1, \cdots N_m, s'_1, \cdots s'_i, \cdots s'_m, \alpha'_1, \cdots \alpha'_i, \cdots \alpha'_m$

and $s_j \le s'_j \wedge \alpha_j \le \alpha'_j\ \forall j$

We define the used bandwidth of the network in state *S* as

$$\Im(S) = \sum_{j=1}^{m} \left[\alpha_j bw(s_j) + (N_j - \alpha_j) bw(s_j - 1) \right]$$

The transition probability from state S to state S' where $S = N_1, \cdots, N_i, \cdots, N_m, s_1, \cdots, s_i, \cdots s_m, \alpha_1, \cdots, \alpha_i, \cdots \alpha_m$ $S' = N'_1, \cdots, N'_i, \cdots, N'_m, s'_1, \cdots, s'_i, \cdots s'_m, \alpha'_1, \cdots, \alpha'_i, \cdots \alpha'_m$ is $P(S, S')$ and can be described as

$$\begin{cases} \lambda_i & \begin{array}{l} \Big(\exists i \ni (N'_i = N_i + 1) \wedge (N'_j = N_j \forall j \ne i)\Big) \wedge \\ \Big(\big((s_j = s'_j = 0) \wedge (\alpha_j = N_j) \wedge (\alpha'_j = N'_j)\big) \forall j\Big) \end{array} \\ \\ \lambda_i / \Gamma & \begin{array}{l} \Big(\exists i \ni (N'_i = N_i + 1) \wedge (N'_j = N_j \forall j \ne i\Big) \wedge \\ \Big(C - \Im(S) < bw(s_i + 1)\Big) \wedge \Big(s_j \le s'_j, \alpha_j \le \alpha'_j \forall j\Big) \end{array} \\ \\ \lambda_i & \begin{array}{l} \Big(N_j = N'_j, s_j = s'_j, \alpha_j = \alpha'_j \forall j\Big) \wedge \\ \Big(C - \Im(S_{\min}) < B^i_{\min}\Big) \end{array} \\ \\ \mu_i & \begin{array}{l} \Big(\exists i \ni (N'_i = N_i - 1) \wedge (N'_j = N_j \forall j \ne i\Big) \wedge \\ \Big((s'_j = s_j = 0) \wedge (\alpha_j = N_j) \wedge (\alpha'_j = N'_j \forall j\Big) \end{array} \\ \\ \mu_i / \mathrm{H} & \begin{array}{l} \Big(\exists i \ni (N'_i = Ni - 1) \wedge (N'_j = N_j \forall j \ne i\Big) \wedge \\ \Big(s_i \ge s'_i \forall i\Big) \wedge \Big(\alpha_i \ge \alpha'_i \forall i\Big) \end{array} \end{cases}$$

Where s_{min} is the network state where every call is allocated its minimum required bandwidth. The state transition equation describes three cases for arrivals, and two cases for departure.

For arrival, the first case describes an arriving call of class *i* and the call is accepted without any borrowing. The second case describes an arriving call that will be accepted with borrowing. That implies that the arriving call could not find enough free bandwidth to be accepted without borrowing, but borrowing can accommodate this call. In this case, Γ represents the number of ways we can borrow bandwidth from the existing calls. Since our protocol did not specify the order of borrowing from existing calls, this is left to the implementation. The third case implies there is not enough bandwidth for the call to be accepted, and we could not borrow bandwidth even for the minimum bandwidth required for this call, in this case, the call is rejected and the network stays in the same state *S=S'*.

For departure, the first case where there was no borrowing, while the second case, we borrowed in order to accept the call, after the call is terminated, we will return the borrowed bandwidth to other calls. When returning the borrowed bandwidth, we can return it to the nodes in many different combinations. H represents the number of such combinations.

Example: Consider the following simple scenario where we have two types of calls, a regular phone call with bandwidth of 30Kbps, that call could be degraded to 25, and 20Kbps. The second type is an Email/ Fax call with a bandwidth of 20Kbps that could be degraded to 15Kbps or 10Kbps. The admission probability for both categories $\rho_1 = \rho_2 = 1$. The cell bandwidth is 1Mpbs. Figure 2. shows a part of the Markov chain for this example where.

We choose to include node <16, 26,0,0,0,0> because at this state, the utilization is 100% without any borrowing,

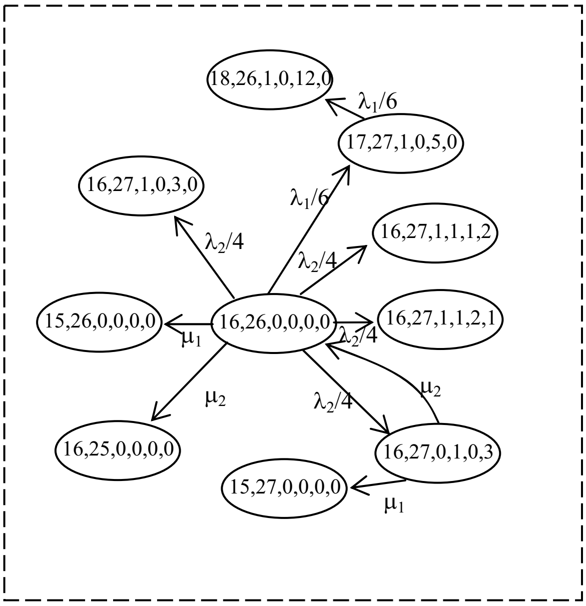


Figure 2: A part of the Markov chain for Example 1

any new arrivals will cause borrowing to start. For example if a call from category 2 arrives, we have to borrow 15Kbps in order to admit it at level 1.

The protocol does not specify the order of borrowing. The simulation assumes a round-robin borrowing from all the eligible nodes. In the Markov model we assume a random choice of the nodes. That implies that there are 4 different combinations for borrowing. We can borrow from zero node in category 1 (3 nodes in category 2), one node in category 1 (two nodes in category 2), two nodes in category 1 (1 node in category 2), or three nodes in category 1 (0 nodes in category 2) each with a probability 1/4. Thus there the transition probability from node <16,26,0,0,0,0> to any one of the four nodes < 16,27,0,1,0,3>, <16,27,1,1,1,2>, <16,27,1,1,2,1>, and <16,27,1,0,3,0>, is $\lambda_2/4$. If the arriving call request is from category 1, then we have 6 different possibilities (only one of them is shown in the Figure <17,26,1,0,5,0>)

Although the previous Example is a very simple case where we have only 2 types of traffic and 2 degradation levels for each. This chain has 34,419 states. If we increase the bandwidth to 2Mbps, the number of states is to 248,157 states. For a bandwidth of 3Mbps, that number jumps to 807,884 states. That means analytical solution for the model is not practically feasible for any real network. In the next section we use simulation and present some results about the performance of our proposed protocol.

V Simulation Results

We have simulated our protocol and we report our results in this section, the important factors that we considered are the call blocking ratios, the average bandwidth assigned to each call, and the utilization.

TABLE I. Traffic Characteristics

	B.W. Requirements max/min	Av. Duration	Type
1	30,30 Kbps	3 minutes	Voice
2	256,256 Kbps	5 minutes	Video conference
3	6,1 Mbps	10 minutes	Video on demand
4	10,10 Kbps	30 seconds	E-mail & FAX
5	512,64 Kbps	3 minutes	Data
6	10, 1 Mbps	2 minutes	File transfer

TABLE I. shows the workload we used in our simulation. We used the workload that was presented in [14], and [7]. The mix of the different types of traffic, the priority, and the number of segments will be stated for each experiment.

In our simulation, we used the maximum bandwidth requirements in [14], but not the average. The bandwidth requested at the time of admission is the maximum, then the bandwidth may fluctuate down to the minimum during the life of the call. We also used MATLAB [13] in our simulation.

One issue we had to consider is the mix of these different types of traffic. We could not find in the literature any model to predict the mix of traffic in wireless networks. We decided on using a mix that shares the cell bandwidth equally among the different types. The arrival rate for the different call categories is adjusted such that the load on the network from the different types of traffic is equal, thus the different types are equally sharing the network bandwidth. For example, the arrival rate for type 4 (require 20Kpbs for an average duration of 30 seconds) should be 9 times that of type 1 (require 30Kbps for a duration of 3 minutes) such that these two types have the same "load" on the network.

Our network model is as follows: 30Mbps total bandwidth per cell. We modeled handoff calls as a different category of calls with the same traffic characteristics as the new calls in the same group but with a higher priority when needed. When we modeled the call blocking probability (and call drop probability for handoff calls) we reported it vs. relative arrival rate. The reason of using the relative arrival rate instead of utilization is explained in the next paragraph.

As a measure of system load we use the relative arrival rate. The relative arrival rate is the arrival rate in calls per second normalized to the maximum possible arrival rate that achieves 100% utilization without borrowing. If the channel bandwidth is b_{max}, the duration of the call is T_c , and the required bandwidth is b_r , then the maximum possible arrival rate is $b_{\max}/(T_c \times b_r)$. We choose this measure instead of utilization since utilization will approach 99% just before the degradation starts and stays there while the calls degrade to allow more users at the same utilization level. On the other hand, relative arrival rate gives an indication of the network load compared to a network without call degradation (a relative arrival rate of 1 saturates a similar network without degradation).

In the following two sections, we present our simulation results for voice calls only, and for multimedia calls.

A. Vocie Calls Only

In this scenario, we consider only voice calls. We simulated a FCFS protocol with a bandwidth of 30Kbps per call, then we simulated our proposed protocol by using three bandwidth levels of 30, 25, 20 Kbps, and with an admission tuple of <2, 0.9> for new calls and admission tuple of <3, 1.0> for handoff calls (that means handoff calls are always admitted if we can free enough bandwidth, while new calls is admitted with a probability of 90% and rejected with a probability of 10% if the network in at level 3).

Figure 3. shows the relative arrival rate vs. the call block (drop) ratio for both new calls and handoff calls. Note that without degradation, arrival rate of 1 will result in a 100% utilization of the network and can lead to call drop. Note also that we assumed an equal load for new and handoff calls. That means the arrival rate shown in the figures is for both handoff and new calls. From the Figure, we see that at a relative arrival rate 0.9, the FCFS protocol starts to drop/block calls with equal probability for both new and handoff calls. For our proposed protocol, up to a relative arrival rate of 1.1, there are no dropped or blocked calls. At 1.1 (10% over load), the new calls starts to be dropped by the network, while the handoff calls has a 0 drop rate up to a relative bandwidth of 1.25 (at that point, the drop rate for new calls is 7%).

Figure 4. shows the relation between the call arrival rate and the number of simultaneous calls in the network (for the previous scenario). We can see that up to a relative arrival rate of 1, the number of calls grows linearly with the arrival rate. For FCFS protocol, the number stays at a maximum of approximately 500 calls for both handoff and calls that are originated in the cells (for a total of 1000 calls which is the network capacity). For a higher arrival rates, the extra calls are dropped and the number stays constant at 1000.

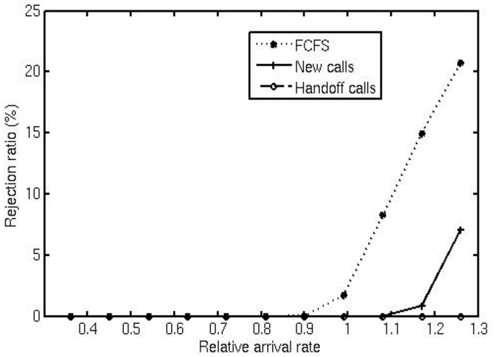


Figure 3. Arrival rate vs. rejection rate

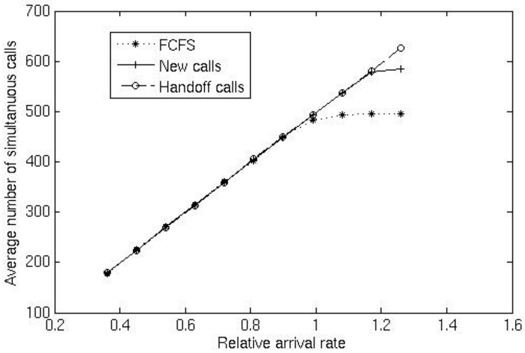


Figure 4. Arrival rate vs. the number of active calls

For our proposed protocol, the number of calls grows linearly up to a relative arrival rate of 1.2 That is of course on the expense of the bandwidth assigned to each call, that can potentially drops to 20Kbps. At rates higher than 1.2, the number of new calls stays the same, and the number of handoff calls increases. That clearly illustrates the higher priority given to hand off calls over new calls.

Figure 5. shows the relation between the arrival rate and the bandwidth assigned to each call. Again, up to an arrival rate 0.9, 30Kbps assigned to each call. After that, the bandwidth gradually drops to the min. of 20KHz per call for both new and handoff calls. (We show the results up to 24Kbps only since below that the drop rate for new calls are unacceptably high). Note that there is no difference in the bandwidth assigned to new and handoff calls, the only difference is in the rejection ratio.

For the previous experiments, we assumed the bandwidth is shared equally among new calls and handoff calls. Figure 6. shows the case when the percentage of the new calls are 30%, 50%, and 70% respectively.

From the figures we can see that disregarding the percentage of the bandwidth dedicated to new calls, the handoff calls up to a relative arrival rate of 1.4 (40% more than saturation) does not suffer any significant drop rate, while the new calls has a 10% drop rate. After 40% overload, then the handoff calls starts to suffer dropping at a much smaller rate than the new calls.

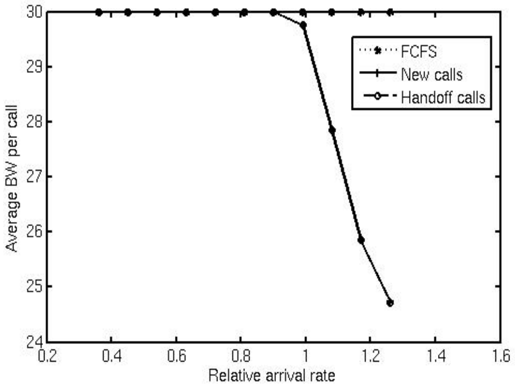
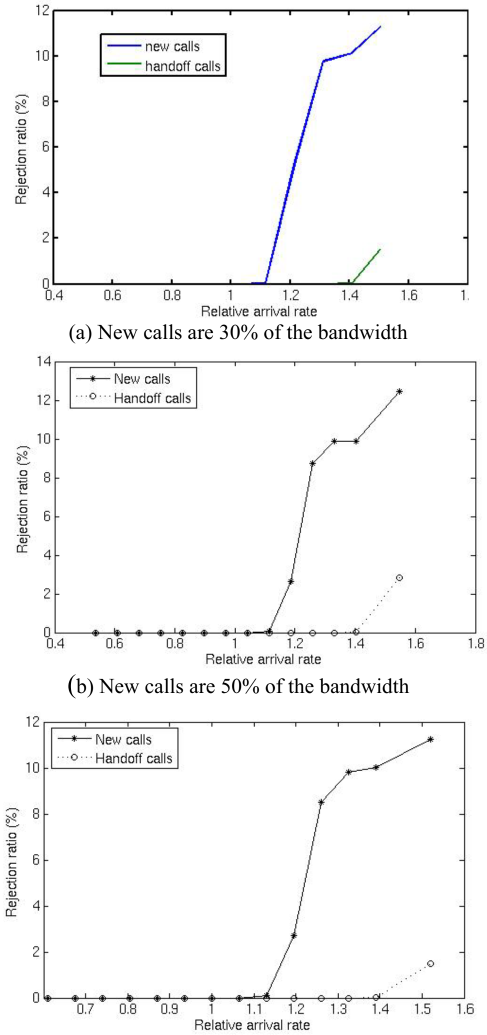


Figure 5. Average BW for cal vs. Relative arrival rate



(a) New calls are 30% of the bandwidth

(b) New calls are 50% of the bandwidth

(c) New calls are 70% of the bandwidth

Figure 6. The relative arrival rate vs. the drop rate for new and handoff calls under varying new call load.

B. Multimedia Calls

In this part, we consider multimedia calls competing for the bandwidth. We use a mix of the traffic in TABLE II. Since the bandwidth requirements and the average duration time of the different types are vastly different, we choose to adjust the arrival rates such that the bandwidth is divided equally among these 6 different types of traffic. For example, a relative arrival rate of 0.5 arrivals per second means that the system is running with an arrival rate of (0.5*5,000)/(30*180)=0.4429 arrivals per second for the first type (the 5000 Mbps is one sixth of the total cell bandwidth). At the same time, traffic type 4 have an arrival rate of (0.5*5,000)/(30*20)=4.1667 arrivals per second. That is a relative arrival rate of 1.0 means that every traffic type is having an arrival rate to saturate one sixth of the available bandwidth in the cell.

Figure 7. shows the rejection ratio vs. the relative arrival rate for a FCFS CAC protocol. From that Figure, we can see that the rejection ratio for traffic type 3 and 6 (max BW of 6Mbps and 10Mbps respectively) start to be considerably high from a relative arrival rate of 0.4 and increase significantly for larger relative arrival rate. While traffic type 1,2,4 and 5 with their relatively small data rate requirements sustain a reasonable rejection rate up to a relative arrival rate of 1.2. While at 1.2 relative arrival rate, virtually no type 6 calls are admitted at all.

TABLE II.   TABLE II. TRAFFIC MIX

Traffic type	P_i,ρ_i	BW requirement
1 Voice	<2 ,0.5>	{30 30 30} Kbps
2 Video conference	<2, 0.5>	{256 256 256}Kbps
3 Video on demand	<3, 1.0>	{6 3 1}Mbps
4 E-mail and Fax	<2, 0.5>	{20 10 5}Kbps
5 Data on demand	<2, 0.5>	{512 256 64}Kbps
6 File transfer	<3, 1.0>	{10 5 1}Mbps

The main reason for the 100% rejection of type 6 traffic is the large bandwidth required for such a call (10 Mbps). If the network is heavily loaded, there is no way to allocate that bandwidth to new calls, and new calls are always rejected. That will also help other traffic types by letting them utilize the unused bandwidth that was suppose to be used by type 6, thus allowing low bandwidth calls to have a very low rejection rate even at 120% arrival rate.

Figure 8. shows the effect of the FCFS protocol on the percentage of the bandwidth allocated to each type of traffic as a function of the relative arrival rate. While the arrival rate is adjusted such that the bandwidth is equally divided among all types. Ideally, the bandwidth grows linearly up to 16.6% (one sixth of the total cell bandwidth), then stays constant.

In the figure, we see that at low network utilization, all types equally share the bandwidth. As the load starts to increase, the percentage dedicated to types 3 and 6 starts to decrease, while the other low bandwidth types continue to increase. At an overload of 20%, type 6 traffic completely shuts down.

Figure 8. also shows that the drop in the bandwidth assigned to traffic types 6 and 3 is mainly transferred to traffic types 1,2, 4 and 5 since they require much less bandwidth compared to 3 and 6.

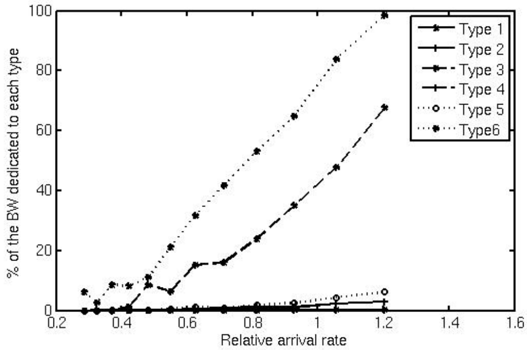


Figure 7. Rejection ratio vs. relative arrival rate for FCFS Admission Protocol

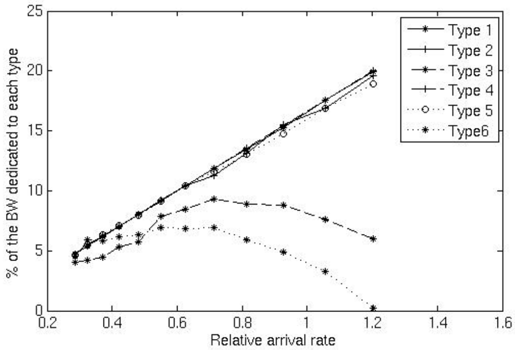


Figure 8. The percentage of BW dedicated to different types for FCFS CAC

Now, we simulate the same scenario using our protocol. We assume 3 levels with the bandwidth requirement as shown in TABLE II. In order to give some priority for traffic types 3 and 6, we assumed that traffic 6 and 3 have an admission threshold of 3 (highest priority) and an admission probability of 1. The rest of the traffic types have an admission threshold of 2, and an admission priority of 0.5.

Figure 9. shows the rejection ratio vs. the relative arrival rate for our proposed protocol. Note that at a relative arrival rate of 1, the rejection ratio for the large bandwidth traffic is less than 1%, while it is less than 0.5% for the low bandwidth traffic types.

Note also that at 100% over saturation (relative arrival rate =2) there is still a low rejection ratio especially for the low bandwidth traffic. That is mainly due to the fact that Type 6 traffic bandwidth requirements starts to go down from 10Mbps to 1Mbps thus freeing bandwidth for the other types.

At a relative arrival rate of 1.7, the rejection ratios for the high bandwidth traffic (types 3 and 6) starts to increase approaching 32% and 35% respectively at a load of 2.3, while the rejection ratios for the low bandwidth types remains below 3% even at a relative arrival rate of 2.3.

Figure 10. shows the relative bandwidth assigned to the different types of traffic using our protocol. As in the case of FCFS, at low load all types equally share the bandwidth. With increasing the load, each type share of the cell bandwidth starts to increase, although the high bandwidth types increases at a smaller rate compared to low bandwidth types.

At an arrival rate of 1, the high bandwidth types share of the total cell bandwidth starts to drop, also types 4 and 5 share of the bandwidth starts to drop since these two types are degradable, so the bandwidth assigned to each call starts to decrease in order to make room for new call requests. While, types 1 and 2 share of the bandwidth continue to increase, since these 2 types are not degradable at all.

Finally, Figure 11. shows the bandwidth assigned to the different types of traffic. We did not show types 1 and 2 since they require a constant bit rate, and the BW (bit rate) assigned to individual calls does not change with the increase in the relative arrival rate.

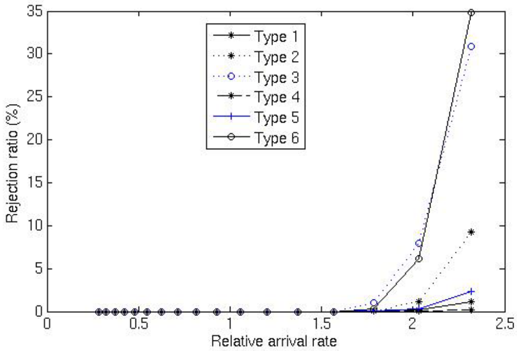


Figure 9. Rejection ratio vs. relative arrival rate

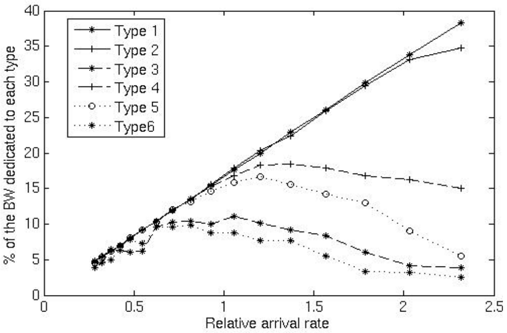


Figure 10. The percentage of BW dedicated to different types for our proposed protocol.

We can see that traffic types 3 and 6 starts at 6MHz and 10MHz respectively, then they go down to the minimum (1Mbps). Type 6 traffic (10 Mbps) starts to degrade at a relative arrival rate of 0.5. The main reason for that is that the bandwidth required for that type is one third of the total cell bandwidth that means even at low rates, the network can hardly afford two calls of type 6 before it starts degrading. At a relative arrival rate of 1.75, almost every call in the network is at its minimum bandwidth.

From the previous discussion we show that our protocol can efficiently utilize the cell bandwidth and share it among many traffic types according to their priorities by borrowing bandwidth from the existing calls in order to admit new ones.

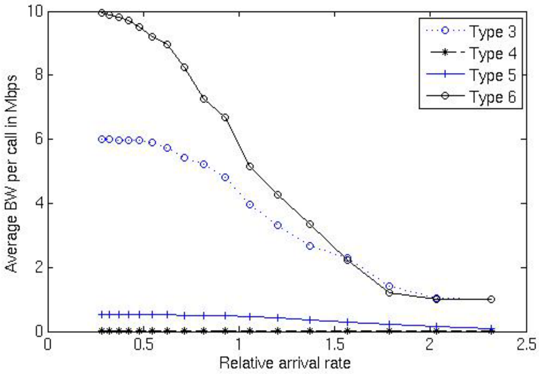


Figure 11. Arrival rate vs. average BW per call

CONCLUSION AND FUTURE WORK

In this paper we presented a call admission control protocol for cellular network. Our protocol depends on borrowing bandwidth from connections that can afford some performance degradation in order to admit new users to the network. Our protocol can assign priorities both in admitting traffic and in bandwidth degradation for different types of traffic. We presented a Markov chain representation for a simplified version of our protocol, and simulation for the full protocol. For future work, we would like to consider extending our protocol to guarantee QoS for different types of traffic by dynamically adjusting the admission tuples of each type to achieve the required QoS. Another issue to consider is the effect of bandwidth changes on the perceived QoS especially for real-time traffic.

REFERENCES

[1] M. Aboelaze "A call admission protocol for cellular networks that supports differentiated fairness" Proceeding of the *IEEE 59th Vehicular Technology Conference VTC2004 Spring*. May 2004.

[2] A. Al-Sharaeh "Dynamic rate-based borrowing scheme for QoS provisioning in high speed multimedia wireless cellular networks" *Journal of Applied Mathematics an Computations*. 179(2006) pp 714-724

[3] J.-Y Chang, and H.-L. Chen "A Borrowing-based call admission control policy for mobile multimedia wireless networks". *IEICE Trans on Communications.* Vol. E89-B, No. 10, Oct 2006. pp 2722-2732

[4] C. H. Chia, and M. S. Beg "Realizing MPEG-4 video transmission over wireless bluetooth link via HCI" IEEE Trans. on Consumer Electronic vol 49, No. 4 Nov. 2003 pp 1028-1034

[5] A. Crossman "A variable bit rate audio coder for videoconferencing" Proceedings of the IEEE Workshop on Speech Coding for telecommunication Oct. 13-15 1993 pp 29-30

[6] S. Das, S. K. Sen, K. Basu, and H. Lin "A Framework for bandwidth degradation and call admission control schemes for mobile multiclass traffic in next-generation wireless networks", IEEE Journal on Selected Areas in Communications, Vol. 21, No. 10, Dec. 2003 pp 1790-1802.

[7] M. El-Kadi, S. Olariu, and H. Abdel-Wahab "A rate based borrowing scheme for qos provisioning in multimedia wireless networks", IEEE Transactions on Para llel and Distributed Systems, Vol. 13, No. 2 pp 156-166, Feb. 2002

[8] M. M. Islam, M. Murshed, and L. S. Dooley "A direction based bandwidth reservation scheme for call admission control". Proceedings of the International Conference on Computer and Information technology. (ICCIT02), Dec. 2002. pp 345-349.

[9] W. S. Jeon, D. G. Jeong. "Call admission control for mobile multimedia communication with traffic asymmetry between uplink and downlink" IEEE Trans, on Vehicular technology, vol 50 No. 1 January 2001 pp 59-66

[10] E. Kim, J. Kim, and K. Kim "An efficient resource allocation for TCP services in IEEE802.16 wireless MANs". The 66th EEE Vehicular technology Conference VTC-2007. Sept. 2007 pp 1513-1517.

[11] M. I. Kim, and S. J. Kim "A 2-level call admission control scheme using priority queue for decreasing new call blocking & handoff call dropping" The 57th Semiannual Vehicular Technology Conference VTC2003-Spring Volume 1 pp 459-461 April 22-25, 2003

[12] Y. C. Lai, and Y. D. Lin "A fair admission control for large bandwidth multimedia applications" proceedings of The 22nd International Conference on Distributed Computing Systems Workshops, 2-5 July 2002. pp 317 - 322

[13] MATLAB www.mathworks.com July 2007

[14] C. Oliveira, J. Kim, and T. Suda "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks" IEEE journal on Selected Areas in Communications. Vol. 16, No. 6 August 1998, pp 858-874

[15] P. Pahalawatta, R. Berry, T. Pappas, and A. Katsaggeos "Content-aware resource allocation and packet scheduling for video transmission over wireless neworks". IEEE Journal on Selected Areas in Comuications. Vol. 25, No. 4 may 2007. pp 749-759

[16] Q. Ren, and G. Ramamurthy, "A real-time dynamic connection admission controller based on traffic modeling, measurement, and fuzzy logic control". IEEE Journal on Select. Areas Comm. Vol 18, Feb. 2000, pp 268-282.

[17] T. Taleb, and A. Nafaa "A fair and dynamic auction-based resource allocation scheme for wireless mobile networks". Proc. Of the IEEE International Conference on Communications ICC'08. May 208 pp 306-310

[18] M. Wu, E. Wong, and J. J. Li "Performance evaluation of predictive handoff scheme with channel borrowing" Proceedings of The 2003 IEEE International Performance, Computing, and Communications Conference pp 531-536, 2003.

Ayman Elnaggar was born in Cairo, Egypt. He has a B.Sc. from Cairo University in 1984, M. Sc. and Ph.D. from University of British Columbia in 1994 and 1997, respectively all in electrical and computer engineering.

He is an Associate Professor in the department of Electrical and Computer Engineering at Sultan Qaboos University, Muscat, Oman. His research interests are in DSP and networks.

Dr. Elnaggar is a member of the IEEE and a Professional member of the ACM.

Mokhtar A. Aboelaze was born in Cairo, Egypt. He has a B.Sc. from Cairo University in 1978, M. Sc. From University of South Carolina in 1984, and Ph.D. from Purdue University in 1988, all in electrical and computer engineering.

He is an Associate Professor in the department of Computer Science and Engineering at York University, Toronto, Ontario, Canada. His research areas are computer architecture and Networks.

Dr. Aboelaze is a senior member of the IEEE, and a Professional Engineer in the Province of Ontario.

Maan Musleh: was born in Minsk Russia and was raised between his homeland, Jerusalem and Dubai. He has his B.Sc. in computer engineering from the American University in Dubai, and currently he is working towards his M.Sc. at York University in Toronto, Canada. His research areas are DSP and networks.

He worked as a software developer in Dubai, and currently he is working as an eCommerce developer in Mississauga, Canada.