Priority-Determined Multiclass Handoff Scheme With Guaranteed Mobile QoS in Wireless Multimedia Networks

Fei Hu, *Member, IEEE,* and Neeraj K. Sharma, *Senior Member, IEEE*

***Abstract*—The provision of multiclass services is gaining wide acceptance and will be more ubiquitous in future wireless and mobile systems. The crucial issue is to provide the guaranteed mobile quality of service (QoS) for arriving multiclass calls. In multimedia cellular networks, we should not only minimize the dropping rate of handoff calls, but also control the blocking rate of new calls at an acceptable level. This paper proposes a novel multiclass call-admission-control mechanism that is based on a dynamic reservation pool for handoff requests. In this paper, we propose the concept of servicing multiclass connections based on priority determination through the combined analysis of mobile movement information and the desired QoS requirements of multimedia traffic. A practical framework is provided to determine the occurrence time of handoff-request reservations. In our simulation experiments, three kinds of timers are introduced for controlling the progress of discrete events. Our simulation results show that the individual QoS criteria of multiclass traffic such as the handoff call-dropping probability can be achieved within a targeted objective and the new-call-blocking probability is constrained to be below a given level. The proposed scheme is applicable to channel allocation of multiclass calls over high-speed wireless multimedia networks.**

***Index Terms*—Call-admission control (CAC), handoff, mobile networks, mobile quality-of-service (M-QoS), wireless networks.**

I. Introduction

In the next-generation mobile cellular-communication environment, the effective delivery of multimedia traffic will become an increasingly important issue as cell sizes shrink to accommodate continuously large demand for high capacity [1]. Mobile quality of service (M-QoS) is a set of performance parameters associated with wireless link, such as channel error rate and, with mobile units, such as handoff call-dropping probability (HDP) and new-call-blocking probability (NBP). It is a common practice to give a higher priority to the handoff calls as compared to new calls. On the other hand, giving too much priority to handoff calls will result in an excessive NBP. Denying too many new calls can bring an unacceptable ratio of carried-to-admitted traffic and a unsatisfactory revenue for network providers. How to allocate channels to meet the specified multimedia calls' bandwidth requirements is the main task of the connection-admission control (CAC) module that is carried out in the base station (BS).

Recently, limited work has been reported in the literature regarding CAC schemes in multiclass wireless networks [2], [3]. Most research works assume single-class traffic in the cells. The provision of multiclass services (also called multimedia communications) is gaining wide acceptance and will be more ubiquitous in the future wireless and mobile systems. Since the profiles of services are vastly different, the qualities of service (QoS) demanded by these services also differ greatly. It is a very challenging task to use the CAC for fairly allocating resources among the different mobile host (MH) users and to guarantee the required QoS of each class of calls. This paper proposes an effective framework for assigning wireless bandwidth to QoS-differentiated calls.

A multiclass CAC scheme based on adaptive bandwidth reservation has been proposed by Oliveira *et al.* in 1998 [4]. We refer it to as the Oliver98 scheme. One of the drawbacks of the Oliver98 strategy is that handoff prioritization, a crucial component of the CAC mechanism, is based on the concept of quality degradation (QD) [5]. QD should be used equally for all kinds of calls instead of only handoff calls. Another drawback of the Oliver98 strategy is that all of their simulations assume the interarrival times of handoff/new calls to follow a geometric distribution, which cannot reflect actual traffic conditions [2], [6]–[8]. The best assumption is general distribution. The potential resource-estimation scheme (PRES) is proposed by Ramanathan in [2]. This scheme is a multiclass extension of the adaptive resource-allocation scheme with GC estimation proposed in [9]. The obvious drawback of PRES is that it shows extremity for handoff prioritization. Handoff prioritization means that we should give handoff calls much higher priority over new calls. However, it does not imply that we should accept all of the handoff calls and consider only the admission control of each arriving new call. The one-step prediction scheme (OSPS) was suggested by Epstein in [10]–[12]. One of the drawbacks of OSPS is that it assumes the MH handoffs to all neighboring cells with equal probability when estimating one-step bandwidth. It overestimates the required bandwidth in those neighboring cells and unnecessarily denies many new calls, which makes the NBP unacceptably high when OSPS is applied to practical mobile-multimedia networks.

In this paper, we give a detailed and practical framework for handoff requests reservation. Our discussion assumes an accurate next-cell prediction scheme. With the successful application of the Kalman filter to the global position system (GPS) and other position-locating systems, a precise next-cell prediction

Manuscript received January 27, 2001; revised March 20, 2002, April 10, 2003, and July 23, 2003.

F. Hu is with Computer Engineering Department, Rochester Institute of Technology, Rochester, NY 14623-5603 USA (e-mail: fxheec@rit.edu; fei.hu@ieee.org).

N. K. Sharma was with Computer Engineering Department, Clarkson University, Potsdam, NY 13699 USA. He is now with Intel Corporation, MS C03-103, Beaverton, OR 97006-5771 USA (e-mail: neeraj.k.Sharma@intel.com).

Digital Object Identifier 10.1109/TVT.2003.819614

technology will become a reality in the next-generation mobile networks. It is unnecessary to assume that the MH will handoff to neighboring cells with undeterminable probability, such as in the Oliver98 strategy. It is also incorrect to regard the probabilities to all neighboring cells as the same value, such as in OSPS. The timing relationship is analyzed between handoff-request reservation and later handoff call admission. This is very meaningful for practical system implementation. The state transition map is given for our reservation-pool mechanism.

To guarantee the M-QoS of each class of handoff calls, we propose a new notion of reservation ordering (RO) of handoff requests. RO is about the assignment of admission priorities for multiclass calls. However, our admission-priority determination is made according to the MH's time-varying movement behaviors and the desired M-QoS requirements of the multiclass calls themselves. On the other hand, OSPS determines call priorities based on only calls' M-QoS profiles. For the computation of RO value, a weighted algorithm is proposed. Unlike LLCS and the Oliver98 strategy, we assume many traffic classes rather than just two (real-time and nonreal-time). The desired amount of bandwidth and delay requirements for these QoS profiles can vary greatly. Although PRES and OSPS also assume multiclass traffic, we analyze urgency details of different ATM AAL services instead of simply assuming K classes of mobile users. Such urgency details are used for computing the RO value. Our CAC approach is implemented in a distributed way. The algorithm needs only the signaling information between local BS and MH. This method can bring reduced computation load compared to mobile switch center (MSC)-centered control policy.

The rest of this paper is organized as follows. Section II describes the detailed procedure for forming a handoff-request reservation pool that is based on accurate next-cell prediction. This is followed by the presentation of the RO policy in Section III. The complete call-admission mechanism is presented in Section IV and Section V provides our simulation results and corresponding analysis. Section VI comments on the influences of practical factors on the efficiency of our scheme. Finally, we conclude in Section VII.

II. Forming of the Multiclass Reservation Pool

The profile-based CAC scheme is widely adopted in existing wireless academia to conduct bandwidth reservation and allocation of handoff/new calls. It assumes that we can get the mobility pattern of the MH using profile-based schemes. This assumption may not be valid in practical systems due to the following three reasons.

1) In many wireless systems, such as wireless ATM network environments, wireless components can be connected to wide-area networks (WANs), local-area networks (LANs), or even home, depending on what kind of ATM network is to be accessed. For such varied wired networks, it may not be possible to predict the arrival of MH to some cells, since the mobility patterns may not be available.
2) In profile-based schemes, the mobile system should deal with high computational overhead in terms of development, storage, and updating of the traffic patterns.
3) Varying traffic conditions suggest that such history-based schemes can never be fully reliable.

Therefore, we should use real-time position measurements to predict the future path of a moving MH. The greatest advantage of future position prediction is that we can determine the next cell that the MH will cross with high accuracy. Thus, we need to reserve wireless resources only in the next cell among all of the neighboring cells and eliminate the reservation of excessive bandwidth in those neighboring cells where the sum of arriving probabilities is less than some small value. Taking into consideration the limited radio resources compared to the wired part of the wireless network, such an advantage is valuable. We can use the following formula to express the above ideas:

$$\Re = \begin{cases} \sum_{i=1}^{M} \Phi_i, & \text{in Next cell where } \mathrm{P}_\Lambda \geq \text{Threshold} \\ 0, & \text{in other neighboring cells where} \\ & \quad \sum_{i=1}^{D-1} \mathrm{P}_i \leq (1 - \text{Threshold}) \end{cases} \quad (1)$$

where we denote the prediction accuracy for next cell as Threshold (in this paper, Threshold is assumed to be 90%, which is possible for existing position-location techniques [13]) and $\Re$ is the amount of wireless resource[1] reserved for arriving calls and M is the number of classes of multimedia connections with their individual resource reservation as $\Phi_i(1 \leq i \leq M)$. Furthermore, we assume that the computational-arriving probability for the next cell is P_Λ while others are $\mathbf{P}_i$ $(1 \leq i \leq D-1)$. D is the total number of neighboring cells, including the next cell.

Kalman filtering plays a crucial role in computing the arriving probabilities to neighboring cells and predicting the next cell, where handoff calls should be accepted. In this paper, we suggest elimination of hierarchical location prediction (HLP) for global intercell direction proposed in [14] and using only the local movement prediction.

The prerequisite of "accurate next-cell prediction" is also assumed in [15]. Our proposal is an extension to [15] in three aspects:

1) considering the multiclass CAC instead of single-class calls;
2) modifying their hybrid predictive-channel reservation (HPCR) to improve handoff priority (see Section IV for details);
3) analyzing the concrete time duration for handoff-request submission from the mobile hosts to the destination BS (see Section II.3).

The handoff-request reservations should take place before the actual acceptance of handoff calls. The acceptance of handoff calls means that the new BS should allocate channels for the calls. A MH will submit a handoff reservation request when:

1) received signal strength (RSS) in the current BS is below a threshold level;

[1] Typically, "resource" in cellular networks refers to the available bandwidth [such as time slots in the time-division multiple-access (TDMA) scheme, codes in the code-division multiple-access (CDMA) scheme, or frequency bands in the frequency-division multiple-access (FDMA) scheme], transmission power level, and the amount of buffer allocated for accommodating the incoming calls in the BS. For simplicity, here we assume only bandwidth.

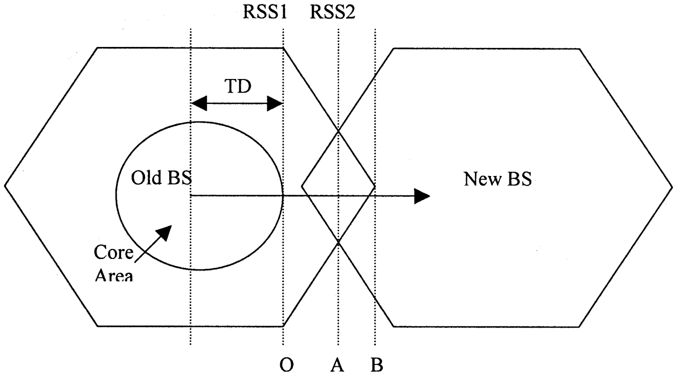


Fig. 1. Time for forming reservation pool (between O and A).

2) RSS in the next-cell BS is high enough to allow the receiving of signals from the MH.

In this paper, we give a practical way to determine the reservation deadline (RD) that is a time instance by that bandwidth assignment for the arriving handoff call should be completed. There are several handoff criteria for determining the value of RD [8]. A typical way to accept handoff calls starts as soon as the next-cell BS has the same strength of RSS as the current BS. This leads to too many unnecessary handoffs, since the RSS in the current BS is still adequate for communication.

To avoid blind selection of the start point of channel reservation for handoff requests, we define the concept of the core area (CA) with a radius of size threshold distance (TD) in the current cell, as shown in Fig. 1. In CA, there is a high probability for the MH to make a dramatic change in its direction and speed. The similar idea is proposed in [14] and [15]. However, if MH moves beyond the CA, the chances of a sudden change of direction are reduced.[2] Thus, we can improve the accuracy of next-cell prediction by using Kalman filter. The reasonable position to start making reservations can be chosen as O (see Fig. 1). From the point of view of RSS, position O corresponds to the value of RSS1 in the current cell. The relationship between the RSS and distance x from the transmitter in the BS is [16]

$$\mathrm{RSS}_{\mathrm{dB}} = -10\gamma \times Log(x) \tag{2}$$

where γ is the propagation path-loss coefficient.

To determine the value of RD, we consider the following two criteria.

1) The RSS level of the current BS drops below a threshold RSS2 so that it is somewhat difficult to keep the communication with MH. The position corresponding to RSS2 is shown as in position A.
2) The RSS level of next-cell BS is stronger than that of the current BS by a given hysteresis margin Δ. That is, we can only serve handoff calls in Position *B* (see Fig. 1).

As can be seen from Fig. 1, the RSS level-meeting condition (1) is on the right of line A, while for meeting condition (2) it is on the right of line B. Thus, to meet both conditions, we have to choose right of line B. Therefore, once a MH arrives at position B, we should stop the submitting of handoff request immediately. Then the reservation time duration Ω for an MH is from arriving time at position O to the arriving time at position B. Ω can be expressed as $\Omega = \mathrm{T}_{OB} = t_B - t_O$. If we consider predominantly walking and stationary users with an average speed of 2 m/s and a cell radius of 300 m, which is a common case in wireless ATM campus LAN, the typical value of Ω is about 5 s $\sim$15 s [16]. The value of Ω is important since all of the handoff reservation actions, such as RO and overflow-request queuing RO, which will be discussed later, should be finished during Ω. Also, the values of QDT and RET (discussed in Section V) are set up based on the value of Ω.

Let us first clarify the concept of wireless effective bandwidth (W-EB) before the discussion about the procedure of forming of multiclass reservation pool. Note the concept of effective bandwidth (EB) in [2] is adopted from a wired network instead of a wireless network, where EB means the minimum amount of bandwidth needed to provide a specific QoS given the traffic parameters of a connection and the buffer size at the multiplexer. Kim and Krunz adopted fluid-flow analysis in the mobile multimedia link layer to obtain an optimal bandwidth and code rate that can satisfy QoS parameters specified in terms of cell loss while maximizing the utilization of bandwidth [17]. They named this bandwidth with code rate as *W-EB*. In the following discussions, when we mention bandwidth requirements for each class of calls, it implies that the value of W-EB has been calculated using fast algorithms such as [17]. The number of channels required can be computed according to the value of W-EB.

Each handoff MH sends their W-EB requirements to the BS of next cell during their own Ω. These handoff-request reservations will form a varied-sized pool through marking unoccupied channels from free to reserved. As shown in Fig. 2, handoff calls of different classes can reserve highly varying sized channel blocks (CB). The term *CB* comes from the fact that, in a normal case, a handoff call belonging to some class will occupy a series of allocated time slots. The sizes of free and occupied bands are also varying, since at any time there are always occupied channels released due to calls completion or handoff to another cell.

In Fig. 2 the dark-shaded channel band is marked as *guard channel* (GC). In our scheme, we still maintain a small number of GCs for two reasons. First, because the computation of the cell-crossing probability using Kalman filtering is not perfectly accurate, it is possible that more handoff calls may arrive in the current cell than those we have reserved. Thus, in such scenarios, the GC can be used by the excessive arriving handoff calls. Second, in the case of congestion (the ratio of handoff calls to new calls increases beyond a certain threshold), the GC can give the handoff calls absolute priority as compared to new calls. However, the size of GC does not have to be as big as the fixed GC approach stated in [18], since we can use reservation pool to store reserved handoff requests with high accuracy. The exact number of GC could be slightly different in different cells, based on different traffic conditions such as the average number of handoff users.

To implement the above idea, we use OPNET [19] to simulate the state-transition map (STM) in Fig. 3. The transition from reserved to free state can be explained as follows.

[2] Generally, we choose the radius of the CA to be large enough so that once the mobile host moves beyond the CA, it is very close to the cell boundary. Thus, it does not have many chances to dramatically change its direction and velocity. Since the next-cell-predicting algorithm could be executed at a high speed within the small non-CA [14], the large size of CA should not be an adverse factor.

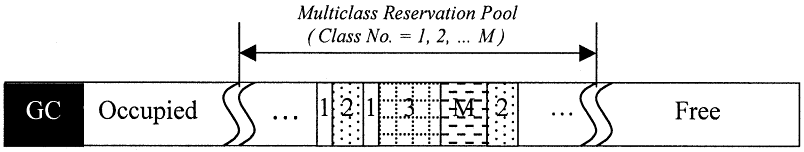


Fig. 2. A snapshot for the channels' status in the current cell.

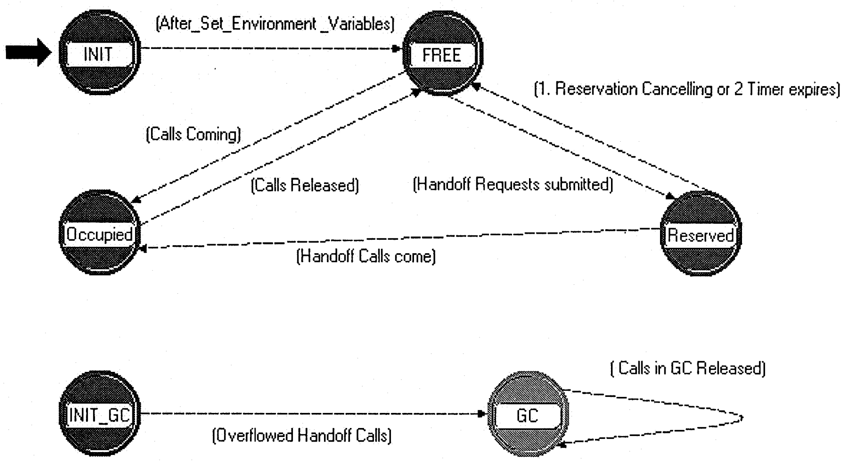

INIT
(After_Set_Environment _Variables)
FREE
(1. Reservation Cancelling or 2 Timer expires)
(Calls Coming)
(Calls Released)
(Handoff Requests submitted)
Occupied
Reserved
(Handoff Calls come)
INIT_GC
(Overflowed Handoff Calls)
GC
(Calls in GC Released)

Fig. 3. States-transition map (STM) for the channels with respect to time (OPNET simulation).

1) During Ω, the next-cell prediction algorithm can be executed periodically. The value of the period is determined based on the computation overhead of the system. Therefore, the resource reservations can happen several times. It was mentioned earlier that there is a very small probability for the MH to make a sudden change beyond the CA. However, due to accidents or other rare conditions, it is possible for the MH to make a sudden change of velocity. Thus, the current next-cell-prediction result will be different from last time. It necessitates that the MH should submit a reservation-cancellation message to the previous next cell and its BS should immediately change the corresponding mark from reserved to free.
2) The last time of the next-cell-prediction result can be incorrect in some rare cases due to the failure of prediction algorithm. It means that the MH does not go to the predicted next cell, although it had reserved a CB in that cell. As each CB can only be used by the MH that reserved it (the reason for doing that will be discussed in Section V), this CB will forever stay reserved if we do not recycle it into free state after a certain time.

Note in the STM that there is an independent state that is called GC. Since GCs are used only when handoff calls that failed to reserve CB cannot obtain free channels after the competition with new calls, it can only have two status: either unused or used by handoff calls. Thus, GC state cannot be part of the big cycle that happens in "non-GC" channels and consists of the other three states (free, occupied, and reserved). A released GC could not be called free in our STM since free in Fig. 3 only means free non-GC channels.

In Fig. 3, it can be noted that the condition "handoff calls arriving" could lead to two types of transformations.

1) From "reserved" to "occupied," as shown in Fig. 1. Once the mobile host moves beyond position B, immediate channel allocation should be executed for meeting its handoff call requirement. Since there is a one-to-one matching relationship between reserved channels and handoff calls, the system could allocate corresponding reserved channels to the coming handoff users.
2) From "free" to "occupied." Basically, most free channels are used for only new calls, since handoff calls already reserved their required channels (marked as "reserved"). However, there are few handoff hosts that could fail to reserve channels because of the rarely happening failure of the next-cell-prediction algorithm. For those handoff calls that could not find out their corresponding reserved channels, they will compete with new calls for the free channels.

Since handoff requests can be reserved with high accuracy, arriving handoff calls can be served with such a dynamically formed reservation pool that new calls cannot be used at any time. Traditional methods for handoff prioritization can be summarized as follows.

1) Keep a fixed number of reserved channels only for handoff calls *a priori*. This number should be much bigger than the GC in our approach. This method cannot adapt to varied traffic conditions since the number of reserved channels remains constant.
2) More people suggest using an adaptive resource-reservation scheme for call admission. This scheme can incur large errors, since their resource prediction is based on inaccurate mobile profiles.
3) Oliver98 uses QD to make the acceptance of handoff calls much easier than for new calls. The drawback has been explained in the introduction.

Our proposed approach uses accurate handoff request submission instead of simply keeping a fixed number of channels beforehand or dynamically estimating required channels based

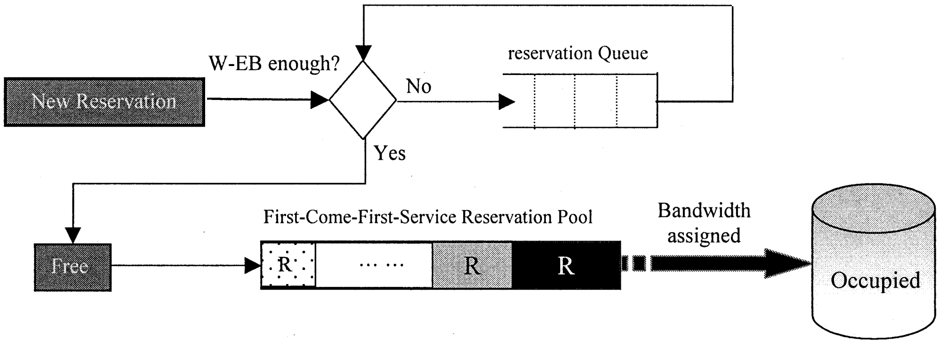


Fig. 4. Case without using RO.

on the possible arriving handoff traffic rate. New calls can only compete with unreserved handoff calls for the free channels. Because of the high submission accuracy, the number of unreserved handoff calls should be a very small percentage. Thus, we can give a much higher priority to handoff calls over new calls. This sort of handoff prioritization is also the future trend for the next-generation mobile and wireless systems.

Our multiclass bandwidth resource procedure could be sketchily summarized as follows.

1) Once the mobile host moves beyond the CA, it will send signaling message to the BSs that could receive its signal. The next-cell-prediction algorithm will run periodically in the BSs to determine the destination cell to which the mobile host is going.
2) Based on the next-cell-determination result, the mobile host will submit handoff requests periodically to the destination BS. The submission should be stopped once the mobile host arrives at Position *B* (see Fig. 1).
3) Each submission could not guarantee the success of channel reservation, since we will consider the reservation priority (to be discussed in the next section) for different classes of calls based on their QofS requirements.
4) Once the mobile host arrives at position *B*, the reserved channels for that mobile host should be immediately allocated. With time going, those allocated channels could be released when the mobile host terminates their connections or moves out of the current cell.
5) For new calls in the current cell, the mobile hosts do not run the next-cell-prediction algorithm and also do not have right to reserve channels. The networking system only checks whether there are unused channels available in the current cell and decides the admission of the new calls.

III. Priority Determination for Multiclass Handoff Calls: RO

For multiclass calls, we should assign each class of calls different priorities during resource allocation, unlike in the single-class case, where all calls are assumed to have the same priority. The role of RO is to make sure that the service order for each submitted handoff-request reservation is maintained.

Fig. 4 illustrates a case when a new handoff reservation comes at time t. Its desired W-EB can be expressed as $\Phi_{K,\chi}$, where χ is the identification number[3] of this MH, K $(1 \leq K \leq M)$ is the traffic class of this handoff request and the corresponding W-EB of the free channels at time t is $\Phi_{F,\Sigma}(t)$. Since the action of reservation is actually the marking of free channels to reserved, if we do not have enough free channels for marking as reserved, that is, if the following condition is met:

$$\Phi_{F,\Sigma}(t) < \Phi_{K,\chi} \tag{3}$$

this reservation cannot succeed at this time. For this over-flown reservation, we can use a reservation queue to buffer it. Once enough free channels are available, this queue can be served. However, the buffered request can be delayed by ΔT_{queue} before it is forwarded to the reservation pool. After experiencing ΔT_{queue} delay, there will be a further delay of ΔT_{pool} due to thte waiting period in the first-come–first-service (FCFS) pool, assuming that the total delay tolerance of this handoff call ΔT_{QoS} and if the condition

$$\Delta T_{\text{queue}} + \Delta T_{\text{pool}} > \Delta T_{\text{QoS}} \tag{4}$$

is met.

This call can be terminated because the application cannot run normally. Thus, we can see that by not using RO (see Fig. 4) can result in a serious consequence for handoff calls. If we could calculate the value of RO priority, we could discard the use of the FCFS pool in Fig. 4 and build a reservation pool with priority control (as shown in Fig. 5); that is to say, a handoff call that first submits handoff requests does not mean that it could successfully reserve channels with first priority, since its RO priority could be very low as compared to other handoff requests.

Note that there are dominate differences between reservation pool and reservation queue in Fig. 4, as follows.

1) Reservation pool is actually the marking of free channels to reserved channels with priority assignment. When the handoff user needs immediate connection from the new BS, the system will allocate its reserved channels to itself. The time instance of allocating channels is based on its RO priority.
2) Reservation queue is for temporarily storing handoff requests that fail to reserve channels. The system will periodically check whether there are occupied channels that are released to free channels. If there are free channels available at some time, the system will move out the

[3] In the actual system, χ often consists of two parts: one is the network adapter physical address and the other is the MH's home-agent (HA) address, which remains unchanged even if the MH undergoes a handoff to another MSC area.

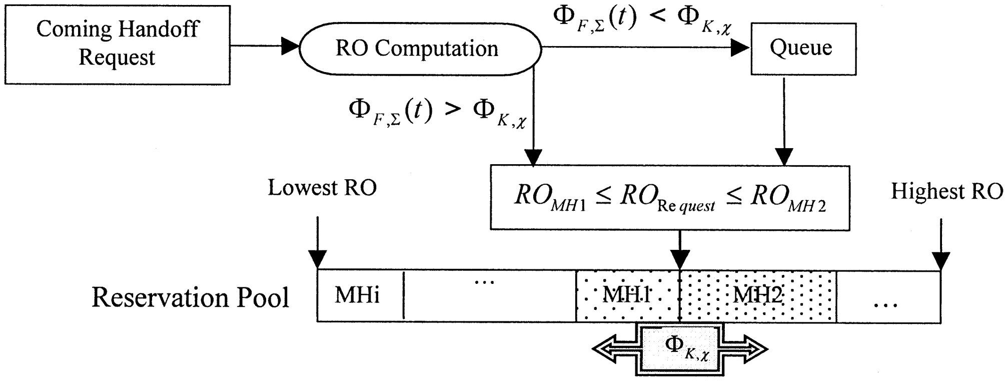

Coming Handoff Request
RO Computation
$\Phi_{F,\Sigma}(t) < \Phi_{K,\chi}$
Queue
$\Phi_{F,\Sigma}(t) > \Phi_{K,\chi}$
Lowest RO
$RO_{MH1} \le RO_{\mathrm{Re}quest} \le RO_{MH2}$
Highest RO
Reservation Pool
MHi
...
MH1
MH2
...
$\Phi_{K,\chi}$

Fig. 5. Dealing with each coming handoff request using RO.

TABLE I
TYPICAL MULTIMEDIA SERVICES IN FUTURE MOBILE MULTIMEDIA SYSTEMS

Class No.	Application Example	Transport Mode	Bandwidth Requirement	Mobility Degree	Average holding time	CU
1	Interactive Video	CBR	~5-15Mbps	Slow	10 min.	1.0
2	Videophone	CBR/VBR	384K-6Mbps	Normal/Slow	5 min.	0.8
3	Telephony /Voice	CBR	9.6K-256K	Fast/Normal /Slow	3 min.	0.6
4	WWW browsing	UBR	0.1-10M	Slow	10 min.	0.3
5	E-mail	UBR	~1M peak	Slow	120 sec.	0.0

TABLE II
POSSIBLE VELOCITY-NORMALIZATION RESULT

Average Velocity	< 20cm/s	1m/s	10m/s	20m/s	>30m/s
Practical example	Almost static	walking	Normal driving	Fast Car	Super Fast
Normalized ($\Delta RSS / \Delta t$)	0	0.2	0.4	0.7	1

handoff request with the highest priority in the reservation queue and finish the action of reservation. It should be noted that each handoff request in the reservation queue should be assigned a timer value. Once the timer timeouts, it should be erased from the reservation queue, since it makes no sense to keep it for a duration longer than reservation time duration Ω, as discussed in Section II.

For determining the RO priority for serving each handoff call, we define a term class urgency (CU) that represents the desired serving urgency degree. CU of the coming multimedia calls is determined by their M-QoS parameters, such as delay tolerance and HDP.

Table I [1] gives five typical multimedia services in future ATM systems and our assigned weights of CU that are between 0 and 1. These values are used in the simulations.

As shown in Table II, interactive video has the highest CU of 1. For these real-time services, we cannot use QD to degrade their qualities or use buffers to increase their delay [20]. For example, in a remote surgery guide (a type of telemedicine applications), if we degrade the quality of the video information sent out from the professional doctors, the remote place can have a dangerous surgery procedure.

However, CU cannot be used as the only factor for determining the value of RO. For example, when an MH is moving almost beyond the reservation area [from position O to position B in Fig. 1 (right)], we should possibly serve this handoff call immediately, even though its CU is low, since its RSS from the old BS is too weak to continue the communications. In other words, the RSS value can become another factor for determining the RO priority. Using RSS value as a priority factor has also been adopted in [7], where single-class traffic is assumed.

Varying speeds of MH can be a serious problem in a mobile multimedia environment in which very rapid fading is common due to its small cell size and low used power. To make the situation worse, the MH in the reservation area can wait in traffic jams, traffic lights, or at stop signs. For these cases, it is very improper to assign these MH to higher priorities just because their RSS is low. Since MH can travel at different speeds and directions, a faster MH will generally require an earlier handoff than a slower one. Thus, MH velocity can become another important factor for determining the RO priority. We can define the RO priority as a two-level weighted scheme

$$\mathrm{RO} = \left[W_1 \times \left(\frac{\Delta \mathrm{RSS}}{\Delta t}\right) + W_2 \times (\mathrm{RSS}) + W_3 \times (\text{class urgency})\right] \qquad (W_1 + W_2 + W_3 = 1) \tag{5}$$

where $\Delta\mathrm{RSS}/\Delta t$ reflects the value of MH velocity and RSS determines the distance of MH from its BS, as shown in (2).

TABLE III
BU Requirements for the Five Classes

Class No.	1 (Interactive Video)	2 (Videophone)	3 (Voice)	4 (WWW)	5 (E-mail)
Desired *BU*	30	10	1	10	5

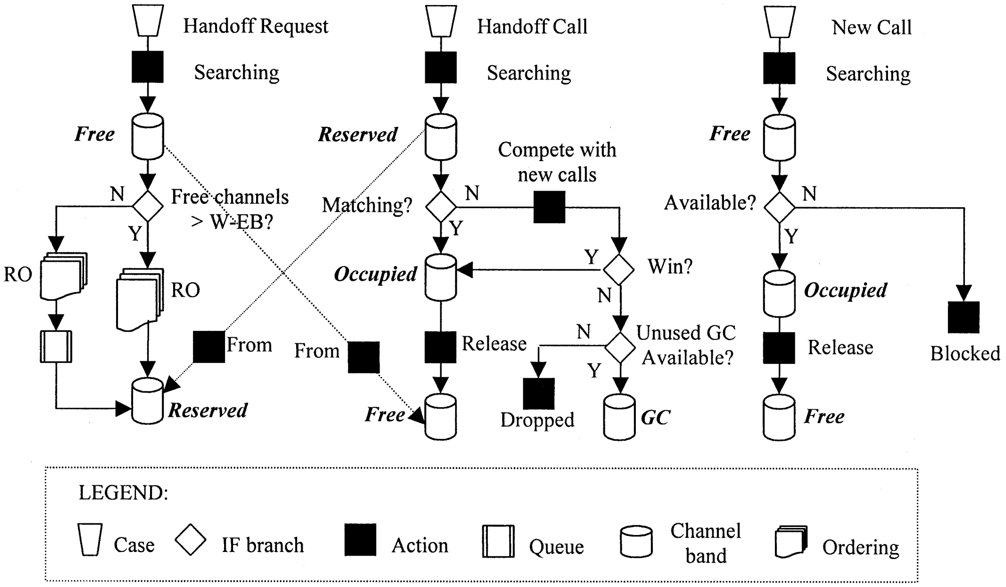


Fig. 6. Flow chat for our proposed multiclass CAC algorithm.

In a multiclass network, we can assign W_1, W_2, and W_3 based on the significance that the three above-mentioned factors may have on RO. A reasonable weight suite assignment is $W_1 = 0.1$, $W_2 = 0.4$, and $W_3 = 0.5$ since CU plays such an important role in multimedia networks.[4] Note that we should normalize the value of $\Delta\text{RSS}/\Delta t$ and RSS between 0 and 1. Table II shows a possible velocity normalization.

If a velocity v is between two neighboring values of Table III, for instance, v is between 1 and 10 m/s, its normalized value of $\Delta\text{RSS}/\Delta t$, represented as δ, could be calculated based on the solution of[5]

$$\frac{v - 1\, m/s}{10\, m/s - 1\, m/s} = \frac{\delta - 0.2}{0.4 - 0.2}.$$

Note that RO depends on two factors. One is the CU of handoff calls that is only determined by defined QoS class. The other is varying mobile behaviors of MH. We use velocity $(\Delta\text{RSS}/\Delta t)$ and position RSS to symbolize the latter factor. This scheme is different from OSPS, in which calls priorities are only determined by class QoS parameters.

There are already many good ways to measure MH velocity, such as in [14], [16], [21], and [22]. Thus, it is not difficult to obtain the value of $\Delta\text{RSS}/\Delta t$.

If the system has errors in estimating the MH velocity, (5) can produce incorrect RO for the corresponding calls. However, it should not be a big problem since in our scheme we assign the weight of velocity $(W1)$ a relatively small value 0.1 to reflect the fact that both the traffic classes and MHs positions play a more important role in determining the handoff priorities.

IV. Multiclass Call-Admission Algorithm

Our multiclass CAC algorithm is shown in Fig. 6. Two crucial details in Fig. 6 are worth mentioning.

1) *Handoff Call Dropping*: When an arriving handoff call is served according to its RO priority, it may not find the corresponding CB in the reservation pool (this can happen when an MH fails to reserve a CB due to the nonperfect next-cell-prediction algorithm).[6] Therefore, the handoff call has to compete with new calls for the free channels.

[4] When choosing the value of those three weights, we should consider the following facts: 1) because of the wide-spread multiclass applications, people pay much attention to the guaranteeing of QoS requirements such as the call-latency and handoff-dropping rates. Thus, it is reasonable that, in our simulation, we assign the value of the CU weight the highest value $(\text{W}_3 = 0.5)$; 2) comparing the two factors, i.e. velocity weight since once the mobile host moves beyond CA, it does not have a high probability to change velocity in a very narrow area; and 3) it could stop suddenly because of red lights or change velocity abruptly because of an approaching accident. However, it has a large possibility to continue to handoff to the next cell smoothly. Therefore, we set the value of $\text{W}_2 = 0.4$.

[5] In our simulation, we assign different class No's 1 ~ 5 to different handoff calls. Practically each coming handoff user will exchange a signaling message with the destination BS and let the system know what type of traffic it is carrying on. The actual traffic type could be constant bit rate (CBR) (such as voice), real-time variable bit rate (VBR) (such as interactive video), or unspecified bit rate (UBR) (such as e-mail data). In our simulations, the goal is to investigate the effectivity of our CAC procedure with priority control. Thus, we ignore the details of actual ATM traffic type and simply assume that the system will allocate a certain number of BUs to different classes of calls.

[6] For example, assume that there are three neighboring cells (cells 1, 2, and 3) and that a handoff user is in cell 1. It cooperates with all neighboring BSs and runs the next-cell-predicting algorithm. Assume that the result of the algorithm tells the system that the reservation should happen in cell 2. Then the system will reserve channels in cell 2. Unfortunately, the algorithm could make a mistake at a small probability and the user actually hands off to cell 3 rather than cell 2. Once the user enters cell 3, it could not find its corresponding reserved channels, because it actually reserved channels in cell 2.

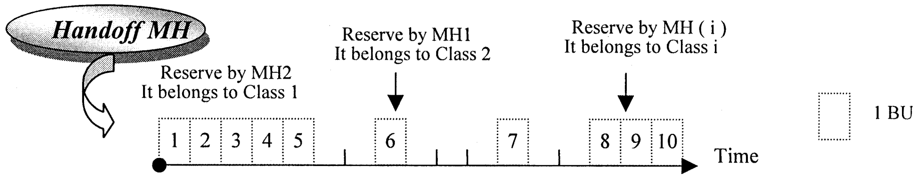
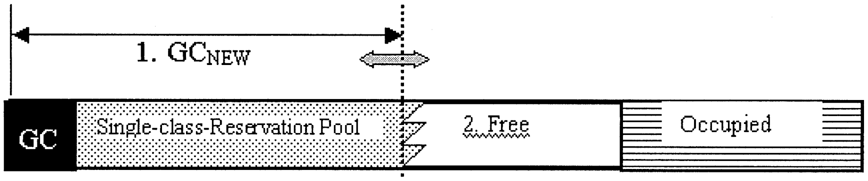
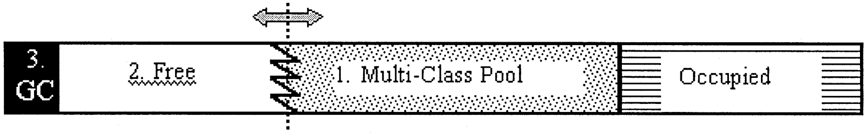


Fig. 7. Disadvantage of the shared-pool approach.



(a)



(b)

Fig. 8. Two channel-accessing schemes. (a) Scheme of [15] and (b) our proposed approach.

However, network congestion can occur due to too many arriving calls, which can result in the handoff call still being not served. Next, it will try to use GC. If all the GC are used by other handoff calls, this handoff call will be dropped.

2) *New-Call-Blocking*: If a new call cannot find free channels, it will be immediately blocked.

It should be noted that the new calls should be served based on their priorities. However, for determining their priorities, only class urgency is used, rather than RO, as in the case of handoff calls.

Different from the shadow cluster doncept [6], our algorithm does not need to involve the location prediction and resource reservation in a set of cluster cells. Thus, our algorithm could be more calculation effective.

Although we also use the idea of the reservation pool, as [15], our method is greatly different from it in the following aspects.

1) The approach in [15] only considered single-class calls and did not provide any priority considerations for the serving of handoff calls. Therefore, it assumed that a call uses only one channel instead of a CB, as in our scheme.
2) [15] suggested that a shared pool be used to reserve channels for each handoff call. "Shared-pool" means that a handoff MH can use any channel in the reservation pool, even though that channel was reserved by another MH. In our scheme, a one-to-one matching scheme is proposed. In other words, a CB can only be used by the MH that reserved it. The reason for doing so is exemplified below.

   Assuming that MH1 reserved a time slot in some carrier frequency, which is a common case in mobile multimedia, which use TDMA as the media-access protocol and MH2 reserved five-time slots (see Fig. 7). Assuming further that MH1 has a higher RO than MH2, we should serve MH1 before MH2. If we use a shared-pool approach, as in [15], there should be no differences among reserved time slots. It is possible that MH1 is assigned time slot 1 and that MH2 is assigned time slots 2–6. Thus, MH2 cannot obtain contiguous time slots such as 1 to 5. This is generally not accepted in a multirate system. It can also bring about difficulties for the implementation of actual signal-sampling hardware components.
3) The scheme in [15] suggests that the reservation pool and the fixed-sized GC can be combined into a new varying-sized GC, which we represented as $\mathrm{GC_{NEW}}$.[7] When handoff calls arrive, the system first checks the availability of $\mathrm{GC_{NEW}}$. If no $\mathrm{GC_{NEW}}$ exists, handoff calls compete with new calls in the free space. The relationship of channels is capacity $=$ $\mathrm{GC_{NEW}}$ $+$ free space $+$ Occupied space [refer to Fig. 8(a)].

   The drawback of this approach is that we can easily run out of GC and, thus, compromising our initial goal of assigning higher priorities to handoff calls.

   However, our approach adopts a different channel-assignment sequence for handoff calls. First, we check the reservation pool, since the handoff call has most likely reserved a channel in the pool. Otherwise, the call competes with new calls for a channel in the free space. If unsuccessful, the call uses the GC. Because we eliminate the definition of $\mathrm{GC_{NEW}}$ in [15] and keep GC as the last

[7] In [15], Section VI.B, the author showed by simulations that their HPCR scheme could get better results than the pure-GC scheme (no reservation). In their HPCR scheme, the total number of reserved channels, which we denoted as $\mathrm{GC_{NEW}}$, is the sum of the fixed number of GC plus the dynamically reserved PCR channels. In [15], Section III-A, based on their discussed HPCR scheme, handoff calls first use reserved channel, i.e., $\mathrm{GC_{NEW}}$ in Fig. 7(a) and then compete with new calls for free channels if no $\mathrm{GC_{NEW}}$ is available.
Note: Because both the HPCR of [15] and our proposal suggest the reservation of channels for handoff calls, the coming handoff users definitely will first check the reserved channels rather than free channels [see Fig. 7(a)].

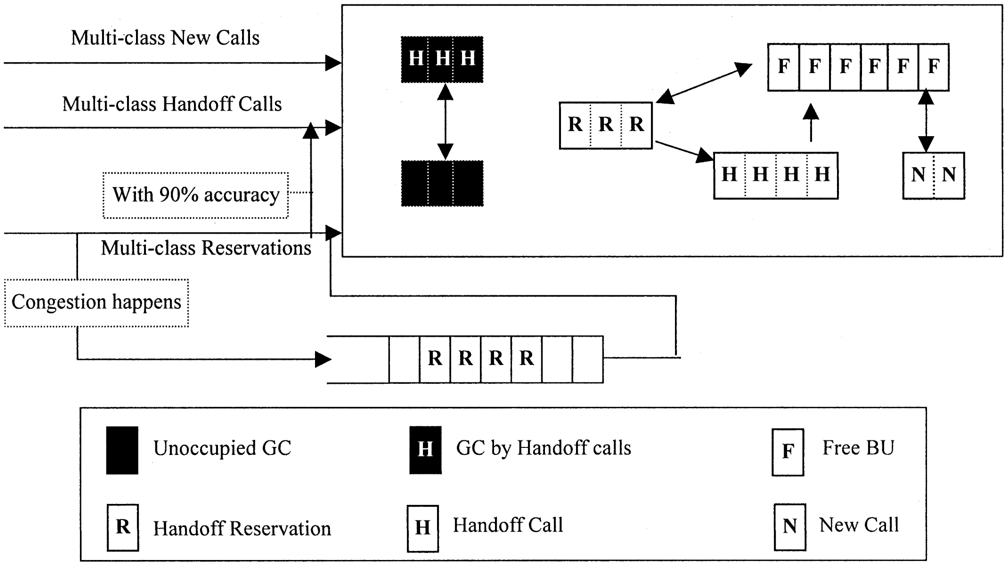


Fig. 9. Simulation model.

choice, we can further lower the handoff calls dropping rate. Fig. 8(b) shows our accessing sequence for handoff calls.

V. Simulation Experiments

A. Simulation Setup

Based on the proposed CAC algorithm, we built a discrete-event-based simulator, as shown in Fig. 9. In this simulation, we choose the total capacity of the current cell as 10 000 bandwidth units (BU).[8] The BU requirements for the five classes of calls are chosen as shown in Table III. In practical implementations, the BU could be matched to a certain number of time slots in a certain modulation frequency or other BUs.

The cell radius is assumed to be 500 m, which is a typical size for future mobile multimedia system. Three different velocities are assumed: 2 m/s (walking), 10 m/s (normal-speed car), and 20 m/s (high-speed vehicle). Furthermore, we assume that the five classes of calls have the same percentages of three velocities in order to emphasize the influence of class urgency on the computation of RO. A cluster of seven cells is assumed; each cell keeps contact with its six neighboring cells.

B. Simulation Stability

Our simulation model shows satisfactory convergence performance. The aggregation HDP and NBP can quickly converge to the stable values after a short simulation time. The larger capacity can lead to a slower convergence speed but a small fluctuation in the stable phase (see Fig. 10). This is because there are less occurrences of reservation-queue overflow when more bandwidth is available.

C. Definition of Three Timers

In this simulation, we define three kinds of timers that will be triggered as soon as they are reduced to zero during simulation cycles. The events triggered by the timers on timeout are determined by the STM shown in Fig. 3. The name of the timers and their initial value, purpose, and meaning are shown in Table IV.

The choice of 15 s for RET and 10 s for QDT is only for the convenience of simulation, although practically we should consider the different traffic densities, different cell sizes, and the mobile host's mobility status. In fact, the choice of RET value would not seriously influence the validity of our simulation results, since the reservation error will happen at a very small probability with an accurate next-cell-prediction algorithm. Also, because in practical cases we can reserve channels for handoff requests with a very high probability of success through empirical analysis on the channel assignment in different cells, the choice of QDT value will not bring dramatic impacts on our CAC scheme.

D. Simulation of Discrete Events With 90% Accuracy for Next-Cell Prediction

In this simulation, we simply assume that the next-cell-prediction algorithm described in [14] could be adopted and can predict the destination cell to which a mobile host will handoff with a 90% accuracy. For simulating the process of CAC based on next-cell prediction with 90% accuracy, we divide the simulation process into discrete time units (TUs). The handoff-request reservations in TU $(i-1)$ will be used by the handoff calls in TU (i), as shown in Fig. 11(a). The 10% inaccuracy comes from two factors that have equal probability of occurring [Fig. 11(b)].

1) There are 5% of reservations made by MH that actually did not arrive. These reservations will be assigned RET and will finally be recycled into free channels.
2) There are 5% of arriving handoff calls that did not reserve channels beforehand. These calls will have to compete with new calls for free channels.

[8] In this simulation, we choose this capacity value only to testify for the effect of our scheme. As a matter of fact, future mobile multimedia or even IMT-2000 should be expected to be able to provide an aggregate transmission capacity of 25 Mb/s when such systems are offered at frequency bands above 3 GHz [23].

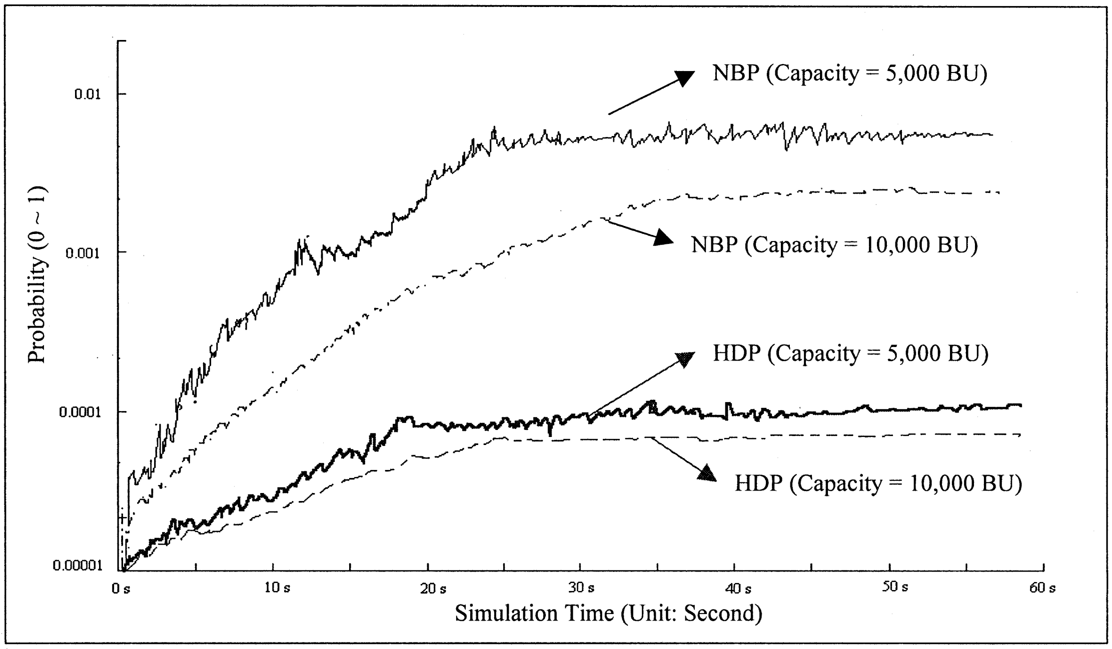


Fig. 10. Simulation stability.

TABLE IV
THREE TIMERS USED IN OUR SIMULATIONS

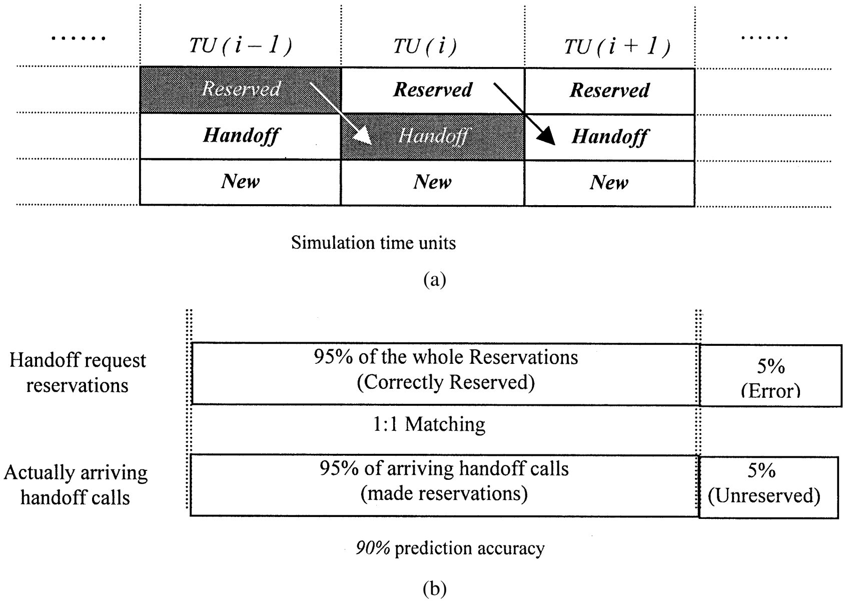
Timer name	Initial value	Purpose	Meaning
Call Holding Timer (CHT)	Randomly Generated (10 s~30 min)	Used for *occupied* channels	Only used in simulation. When it expires, *occupied* channels will be transferred to *free* ones. It is generated randomly based on the class PDF.
Reservation Expiration Timer (RET)	15 s	Used for *reserved* channels	If a *reserved* channel is not used by the MH due to reservation error, it can stay there forever unless we recycle it after a certain time. It's initial value can be set a little larger than the Reservation Duration Ω that is generally 5~15 s.
Queue Deadline Timer(QDT)	10 s	Used for reservation queue	If handoff request failed to reserve channels, it will be placed into the request queue. However, it cannot stay in the queue longer than Reservation Duration Ω.

E. Impact of Next-Cell-Prediction Accuracy Degradation on the Performance of the CAC Scheme

Our reservation-pool scheme assumed a high next-cell-prediction accuracy, as in [15]. To investigate the influence of the degradation of prediction accuracy on the performance of our CAC scheme, we modify the value of prediction accuracy in our simulation model (Section V-A) from 90% to 10% and draw our simulation results as inFigs. 12 and 13. Note that we do not discriminate different classes of calls and consider the performance of the aggregate calls. We investigate the varying trends of handoff-dropping probability and NBP under different HCD values.

Fig. 12 clearly demonstrates that the degradation of next-cell-prediction accuracy could lead to the dramatic deterioration of handoff call admission. Because handoff calls will compete with new calls for the free channels if they could not correctly reserve channels in the true cell that they will actually move into due to the large probability of next-cell-prediction failure, handoff calls could not get much priority over new calls. In Fig. 12, when the next-cell-prediction accuracy is below 20%, the handoff calls have almost the same granularity of denying probability as do new calls.

However, in Fig. 13 we could not see much improvement for NBP with the increase of the next-cell-prediction failure. This could be explained as follows. For a certain cell, there could be coming handoff calls that did not reserve channels in that cell because of the next-cell-prediction failure. At the same time, there could be some handoff users that did not actually come to this cell but reserved channels in this cell by mistake, due to the prediction failure. Typically to say, there are still large amount of reserved channels as compared to the case of perfect next-cell prediction. Thus, new calls still could not utilize more free channels. However, as shown in Fig. 13, there is still improvement for NBP when the next-cell-prediction accuracy is degraded. This is because each reserved channel is assigned a RET value (see Section V-C). When RET time-outs, the system will recycle those reserved channels to free channels. Thus, periodically the new calls could access more free channels.


......
TU(i − 1)
TU(i)
TU(i + 1)
......
Reserved
Reserved
Reserved
Handoff
Handoff
Handoff
New
New
New
Handoff request reservations
95% of the whole Reservations (Correctly Reserved)
5% (Error)
1:1 Matching
Actually arriving handoff calls
95% of arriving handoff calls (made reservations)
5% (Unreserved)

Simulation time units

(a)

90% prediction accuracy

(b)

Fig. 11. Discrete events with 90% accuracy. (a) Simulation time units and (b) 90% prediction accuracy.

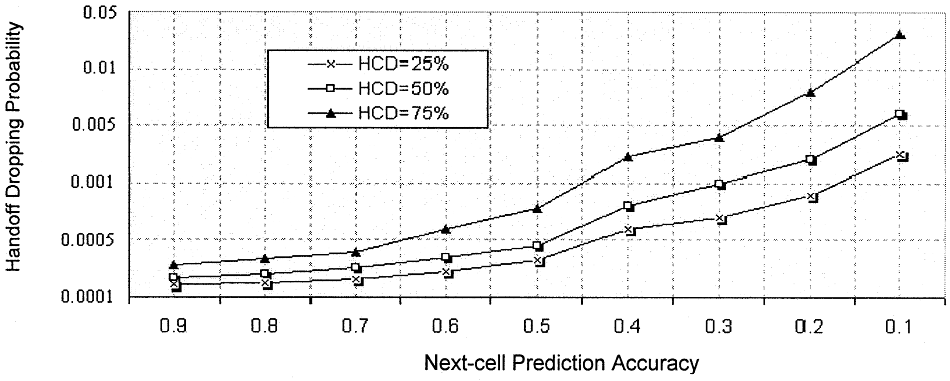


Fig. 12. Influence of next-cell prediction-accuracy degradation on handoff calls.

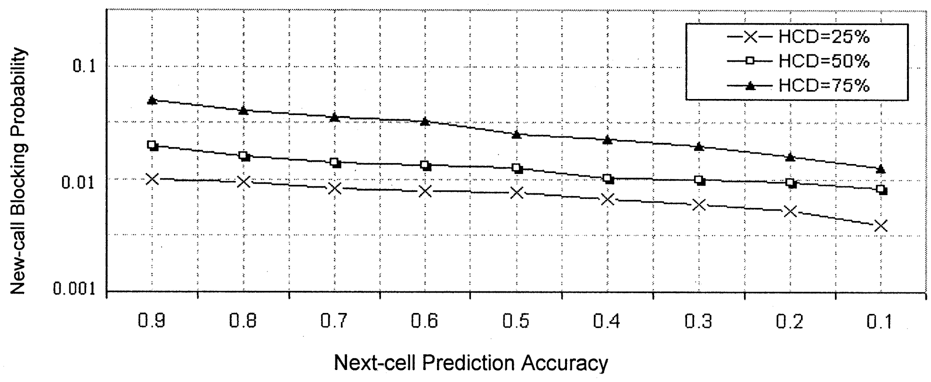


Fig. 13. Iinfluence of next-cell prediction-accuracy degradation on new calls.

F. Role of Queue

Our approach uses a queue for storing overflowing handoff reservations due to the lack of free channels. To investigate the effect of the queue, we assume the same numbers of five classes of handoff requests, that is, their percentage within the total handoff requests, is 20% individually. Because handoff congestion happens only when HCD is high, we let HCD = 80%, which makes the HDP almost 10 times larger than the HCD = 50% case.

The HDP results of five classes of handoff calls are shown in Fig. 14. Although each class of handoff calls experience a certain degree of improvement for their HDP due to the introduction of the reservation queue, the improvement values are different. It can be seen that class–5 calls have the most dominantly decreasing HDP while class-5 calls have the least improvement

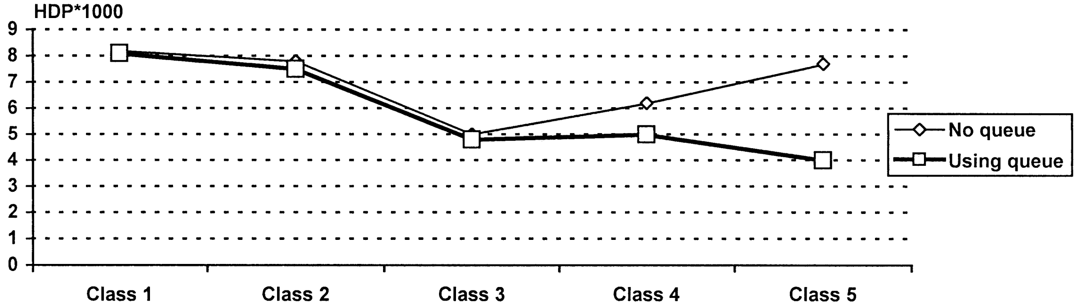


Fig. 14. Importance of the reservation queue.

as compared to the no-queue case. A possible explanation for this phenomenon is that class-5 calls have the lowest serving priority among the five classes of calls, since only class urgency is crucial for computing the value of RO after the elimination of other factors, such as mobile movements. Since the percentage of class-5 users is the same as other classes, class-5 calls will have the largest probability for being buffered into the reservation queue. Therefore, they benefit the most from the reservation queue.

G. Importance of Determining Multimedia Servicing Prioritization

If we assume that MH's position and velocity cannot influence much on the RO of each handoff call except for the CU of each class,[9] we can see the effect of RO on improving the HDP of each class of handoff calls.

We only consider two classes of calls: classes 1 and 5, since class-1 calls have the most crucial urgency requirements while class-5 calls have the least urgency requirements. Two important cases are considered: light handoff load $(\text{HCD} = 25\%)$ and heavy handoff load $(\text{HCD} = 75\%)$. The reason for choosing these two extreme cases is that we may see the effect of RO on HDP more clearly.

Fig. 15(a)–(d) are our simulation results. The X-axis represents the percentage of a given class of calls among all handoff calls. It varies from 20% to 100%. The Y-axis is the value of HDP multiplied by 10 000. It can be seen that HDP of class 1 calls decreases when RO is adopted. Although in a light-handoff-load case, the reduction is not very obvious [see Fig. 15(a)], in a heavy-handoff-load case, the effect of RO is very dominant [see Fig. 15(b)]. This is not a surprising result, since RO can assign class-1 calls the highest priority when only CU is considered.

Unfortunately, HDP increases for class-5 calls [see Figs. 15(c) and (d)], especially in a heavy-handoff-load case [Fig. 15(d)]. This is because class-5 calls get the lowest priority when their RO is compared to other classes. When the network is under congestion, the class-5 calls have the highest probability for being dropped among the five classes.

For dealing with this problem, we can use the crossover ATM switch to buffer those delay-insensitive class-5 ATM cells. When the handoff connection is rerouted from the old path to a new one, a crossover switch should be found by using a fast-searching algorithm [24]. Thus, the down-link data stream can be stored in the buffer of this switch.

H. As Compared to Other Multiclass CAC Schemes Proposed in the Literature

As stated in the Introduction, our reservation-pool scheme is different from other multiclass CAC schemes proposed in the literature in terms of guaranteeing handoff priority. On the one hand, we do give handoff calls higher priority than new calls, since handoff calls could use reserved channels exclusively. On the other hand, the forming of the reservation pool is through the one-to-one matching through accurate next-cell prediction, the size of the reservation pool is a reasonable value so that the new calls could access free channels with an acceptable probability.

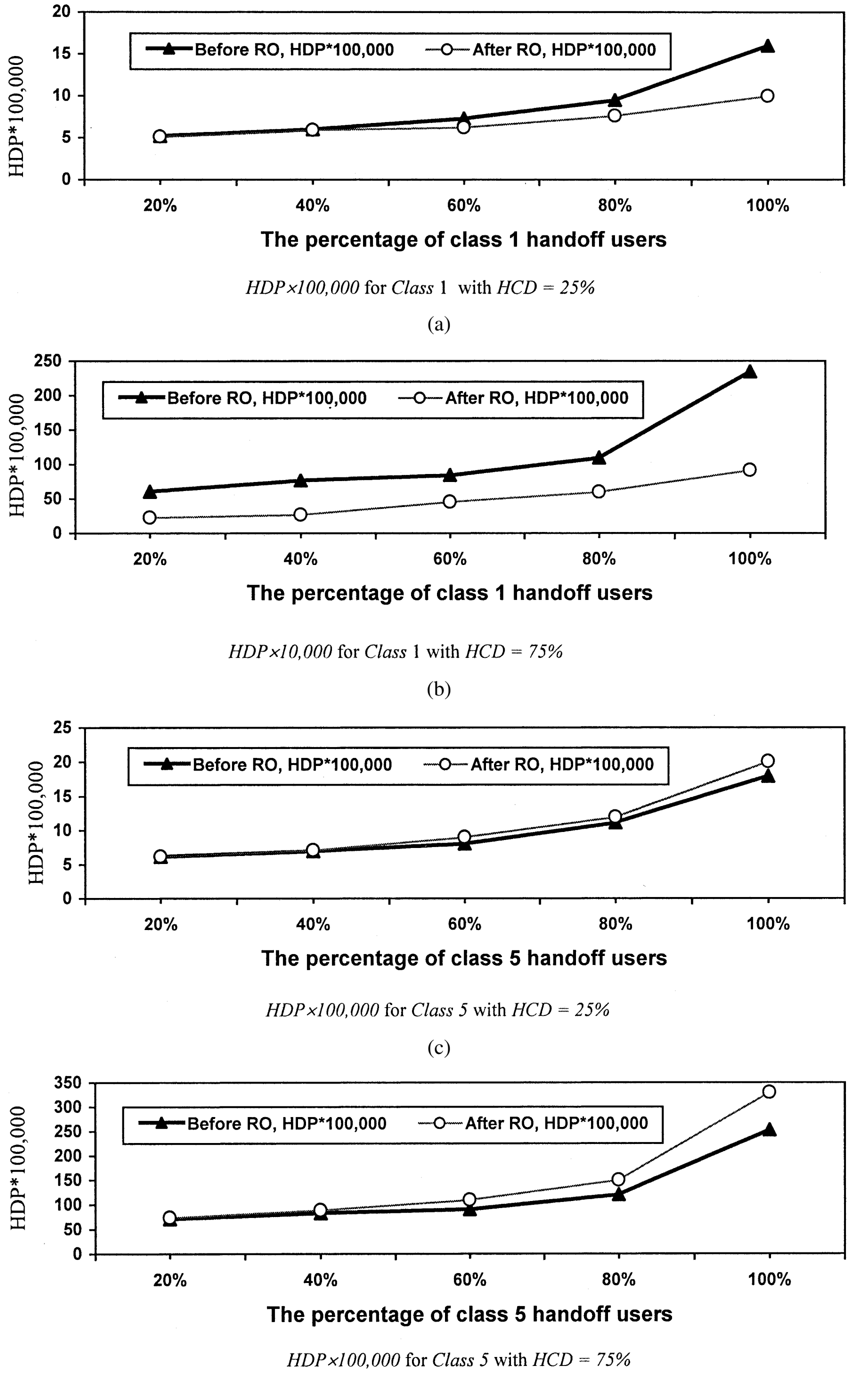
To see the advantage of our proposed approach, we compare our simulation results to that of the Oliver98 scheme proposed in [4]. Since in [4] there are only two classes of calls (real time and nonreal time), in this simulation we investigate only the performance of two classes of calls, class 1, such as interactive video (which can be considered as real-time calls), and class 5, such as e-mail (which can be considered as nonreal-time calls). We also adopt the same values of simulation parameters as in [4]:

1) interarrival times of handoff calls and new calls follow a geometric distribution;
2) number of cells in the mobile system is 100;
3) next-cell-prediction accuracy is set to 95%, which corresponds to 0.95 of probability of moving to a destination cell for the handoff user.

The comparison results are shown in Fig. 16. In this figure, we multiply the values of HDP with 10 000 times and the values of NBP with 10 times to observe their varying trends more clearly.

Fig. 16(a) shows a more serious deterioration for the handoff-dropping probability of class 1 calls in Oliver98, as compared to our proposed scheme. This could be explained as follows. Before the acceptance of class-1 (real-time) handoff calls, Oliver98 checks not only the availability of free bandwidth in the destination cell, but also makes sure that the system could reserve a certain amount of bandwidth in all neighboring cells. This approach could waste the limited wireless bandwidth and increase the threshold of acceptance since we actually could determine the next cell with high accuracy and need only reserve bandwidth in one of the neighboring cells. In addition, our scheme keeps a small amount of GCs as the last lifeboat for congested

[9] This can be achieved through assuming that each class of calls have the same percentage of all types of moving users, such as pedestrians and cars.



HDP×100,000 for *Class* 1 with *HCD = 25%*

(a)

HDP×10,000 for *Class* 1 with *HCD = 75%*

(b)

HDP×100,000 for *Class 5* with *HCD = 25%*

(c)

HDP×100,000 for *Class 5* with *HCD = 75%*

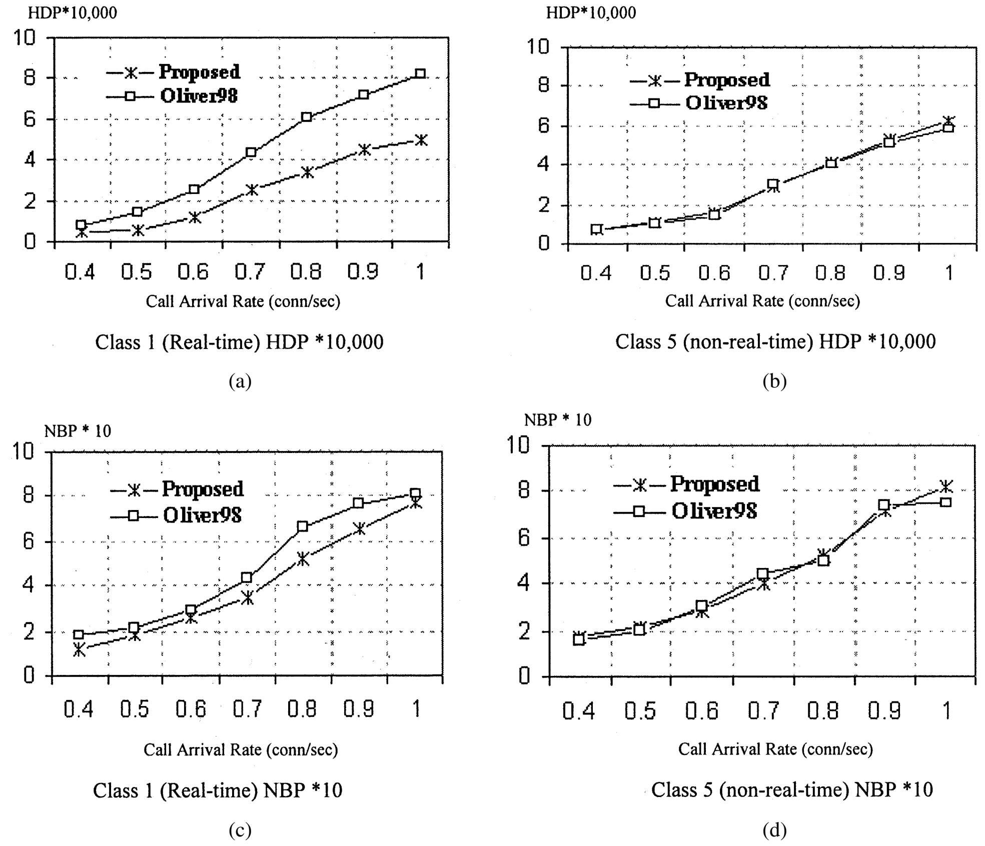
(d)

Fig. 15. Influence of RO on HDP. (a) HDP × 100 000 for class 1 with HCD = 25%, (b) HDP × 10 000 for class 1 with HCD = 75%, (c) HDP × 100 000 for class 5 with HCD = 25%, and (d) HDP × 100 000 for class 5 with HCD = 75%.

handoff calls, which could further improve the acceptance of handoff calls. Likewise, for class-1 new calls, Oliver98 also reserves bandwidth in all of the neighboring cells. If any of them could not successfully reserve bandwidth, this new call will be denied. This approach could deteriorate NBP more seriously than our proposed scheme [see Fig. 16(c)].

However, for class-5 (nonreal-time) calls, the performances of HDP and NBP are very similar between Oliver98 and our scheme [see Fig. 16(b) and (d)]. In our scheme, the class-5 handoff calls have the least class priority of reserving channels in the destination cell. In Oliver98, class-5 handoff calls could be accepted only when free channels are available. These two approaches could produce the close effects from the point of view of handoff-dropping probability. For class-5 new calls, our scheme and Oliver98 all simply check whether or not the free bandwidth is greater than or equal to the desired amount of bandwidth. If it does, this new call will be accepted.

It should be noted that although the acceptance probability of class-1 handoff calls could be increased through quality degradation, which is mentioned in the Oliver98 scheme, (an example



Class 1 (Real-time) HDP *10,000

(a)

Class 5 (non-real-time) HDP *10,000

(b)

Class 1 (Real-time) NBP *10

(c)

Class 5 (non-real-time) NBP *10

(d)

Fig. 16. Handoff-dropping probability and NBP for classes 1 and 5. (a) Class 1 (real-time) HDP * 10 000; (b) class 5 (nonreal-time) HDP * 10 000; (c) class 1 (real-time) NBP HDP * 10; (d) class 5 (nonreal-time) NBP HDP * 10.

of quality degradation is degrading the QoS requirement of the arriving traffic through adjusting its coding rate), for some critical class-1 multimedia applications such as interactive video, quality degradation may not be acceptable. Also, the mobile system could lead to a complex signaling between the handoff hosts and the BS.

I. Influence of GC

Based on the above descriptions, the GC can become the last lifeboat for handoff calls after finishing the following two processes:

1) handoff call cannot find its corresponding reserved channels;
2) handoff cannot obtain free channels after competing with new calls.

Therefore, GC can play an important role for decreasing HDP. This is testified to by our simulation results, as shown in Fig. 17(a)–(c). Before explaining the results in Fig. 15, we first define handoff calls density (HCD) as

$$\mathrm{HCD} = \frac{\sum_{\mathrm{Time}=0}^{\mathrm{Stable\_Status}} \mathrm{handoff}}{\sum_{\mathrm{Time}=0}^{\mathrm{Stable\_Status}} (\mathrm{handoff} + \mathrm{new})} \tag{6}$$

where Time $= 0$ means the initial simulation time and stable status is the time when the value of HDP has converged, which is determined by

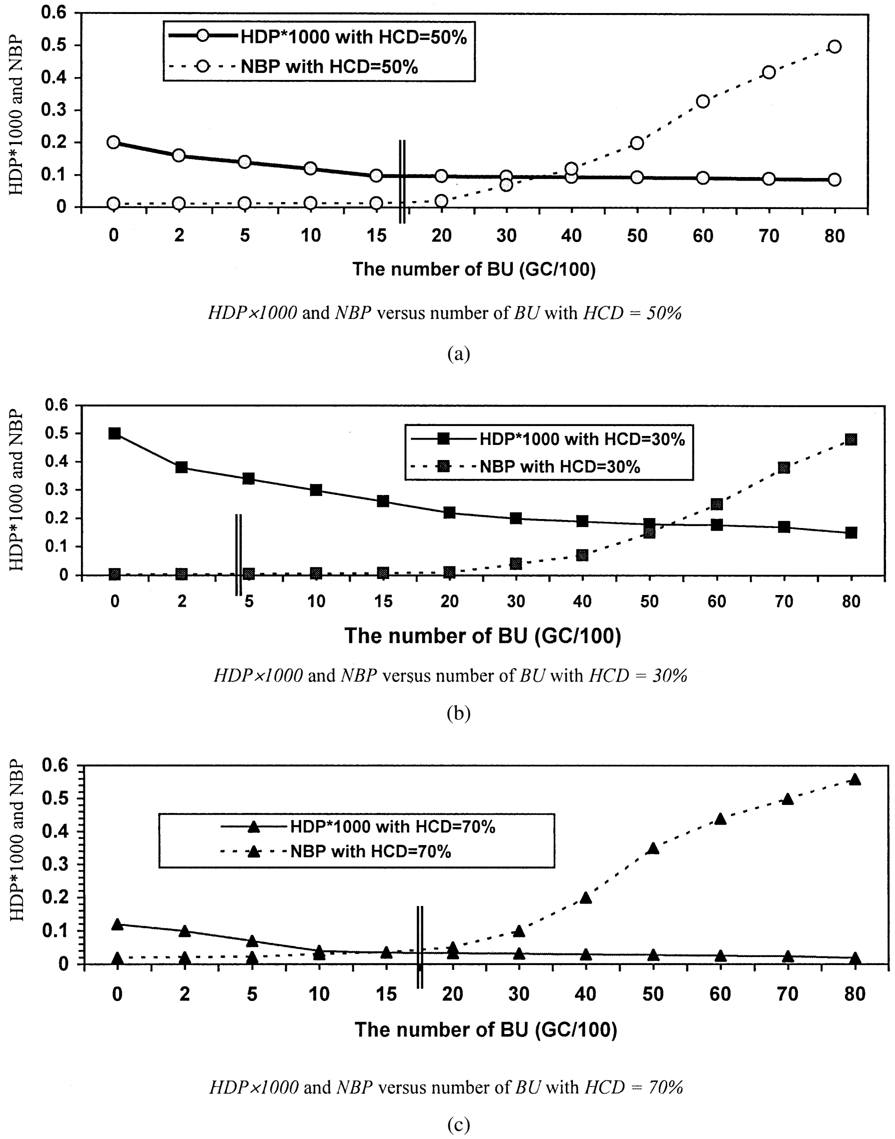
$$|\mathrm{HDP}_{\mathrm{current}} - \mathrm{HDP}_{\mathrm{previous}}| \leq \varepsilon. \tag{7}$$

In Fig. 17,[10] we show the aggregate HDP and NBP for all classes of handoff calls rather than an individual class of handoff calls. The average call-holding time and the average cell-residence time of each call could be generated based on a certain distribution, which will be discussed in Section V-J.

The common trend seen from Fig. 17 is that HDP will become smaller and NBP will become larger with the increasing of GC. We use HDP $\times$ 1000 instead of the original HDP value in order to compare HDP to NBP in the same graph.

We define the concept of warning line (WL) as the value of GC beyond WL. HDP does not have significant improvement while NBP can drastically increase. The importance of introducing WL is as follows. If the number of GCs is too high, our approach becomes similar to traditional schemes that reserve a large number of GCs for handoff calls based on an unreliable traffic profile. This could sacrifice the advantage of our reservation-pool approach, since there will be large number of handoff requests that could not successfully reserve channels within the small number of free channels. Although they can use the GCs after the failure of competing with new calls for free channels, this type of channel allocation is very aimless, since the number of GCs could never become proper enough as the size of reservation pool that is formed through one-to-one matching. On the other hand, too many GCs could largely decrease the number of free channels. Thus, new calls could be denied much more

[10] *Note*: In Fig. 17(a)–(d), we multiply HDP with 1000 times. If we just simply draw the HDP value without enlargement and the NBP value in the same graph, we could not clearly see the changing trend of HDP and NBP because the value of HDP is much smaller than NBP.



HDPx1000 and *NBP* versus number of *BU* with *HCD = 50%*

(a)

HDPx1000 and *NBP* versus number of *BU* with *HCD = 30%*

(b)

HDPx1000 and *NBP* versus number of *BU* with *HCD = 70%*

(c)

Fig. 17. Simulation results on GCs. (a) HDP × 1000 and NBP versus the number of BU with HCD = 50%. (b) HDP × 1000 and NBP versus number of BU with HCD = 30%. (c) HDP × 1000 and NBP versus number of BU with HCD = 70%.

frequently. The definition of WL could give us a threshold line beyond that we could not see a tradeoff result from the point of view of HDP and NBP. As discussed in Section II, the number of GC should be a small value just for overcoming the rarely happening errors of the next-cell-prediction algorithm.

We can see that WL can have different positions with varying HCD.

- Normal handoff traffic load [HCD $= 50\%$, Fig. 17(a)]: WL is located at GC $= 17$, which is actually 1700 BU. 1700 BU out of total 10 000 (17%) is close to our assumed next-cell-prediction inaccuracy, which is 10%.
- Light-handoff-traffic load [HCD $= 30\%$, Fig. 17(b)]: WL is located at GC $= 7$, which is a little smaller than the value of normal HCD case, since less handoff calls need less GC.
- Heavy handoff traffic load [HCD $= 70\%$, Fig. 17(c)]: WL is located at GC $= 22$, which is a little larger than the value of normal HCD case, as more handoff calls need more GC for decreasing HDP more effectively.

J. Determination of Call-Holding Time

An occupied CB will be released as soon as its CHT is reduced to zero. In a practical system, it means that a call leaves the current cell because it is either completed or incurs a handoff out of current cell. To obtain the probability density function (pdf) of the time spent by a class-K call in the current cell, we should know two pdfs.

1) The pdf of the duration of the class-K call in its whole lifetime, which we denote as $\mathrm{P}_K(t)$. This can represent the time distribution of the call from its origin until its termination by user, instead of handoff action.
2) The pdf of the unencumbered cell-residence time (i.e., cell-residence time if the connection is of an infinite duration), which we denote as $R_K(t)$.

Thus, we can express the pdf of the time duration only in the current cell $F_K(t)$ as

$$Y = F_K(t) = 1 - (1 - P_K(t))\,(1 - R_K(t)). \qquad (8)$$

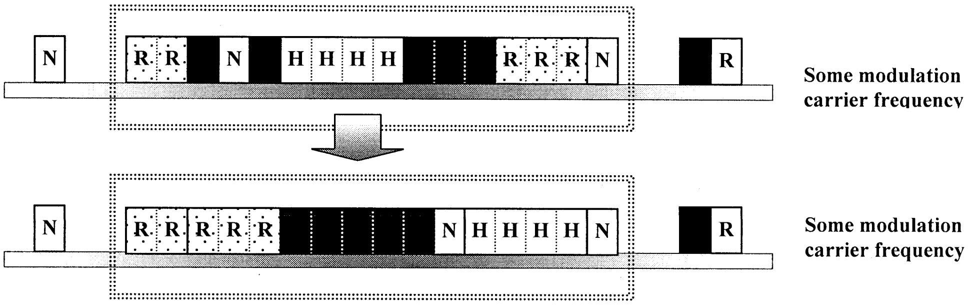

N
R R
N
H H H H
R R R N
R
Some modulation carrier frequency
N
R R R R R
N H H H H N
R
Some modulation carrier frequency

Fig. 18. Channel shuffling (CS) scheme.

If we assume that the call duration whose pdf is $\mathrm{P}_K(t)$ and the unencumbered cell residence time whose pdf is $R_K(t)$ are all exponentially distributed, the pdf of $F_K(t)$ is then also exponentially distributed based on (8). Thus, we have

$$\begin{cases} \mathrm{P}_K(t) = 1 - e^{-\varsigma t} \\ \mathrm{R}_K(t) = 1 - e^{-\xi t} \\ \mathrm{F}_K(t) = 1 - e^{-(\varsigma+\xi)t} \end{cases} \tag{9}$$

For simulations, we can use the inverse function of $\mathrm{F}_K(t)$ to generate a random value of call-holding time that will become the value of CHT for that CB

$$\mathrm{CHT}_K = t(Y) = F_K^{-1}\left(\mathrm{Random}(Y)\right) \tag{10}$$

where Y has an even distribution between zero and one.

Because $\mathrm{F}_K(t)$ can be any kind of pdf, we can assume general distribution rather than a single type of distribution. Based on the works of traffic modeling, it is very appropriate to use exponential distribution to describe tradition voice/audio services. For text data-transmission traffic, the ON-OFF model can be used. While for MEPG video or highly bursty data services, the traffic shows the feature of self-similarity and can be modeled with fractal brown motion (FBM) distribution.

K. Channel Shuffling

We extend the idea of channel shuffling (CS) in [3], to our multiclass reservation pool approach. To adapt to the multirate system requirement of future WATM, a CB is preferred to consist of a series of contiguous channels. Since each reserved CB will be occupied by the corresponding handoff call, we require the practical system to shuffle the communication carrier periodically to place all the reserved channels in one contiguous block. Our CS scheme is different from [3] since in [3] only free bandwidth is shuffled. Through the shuffling of the reserved channels, we in fact guarantee the contiguity of occupied channels and, finally, most of the free channels, since these different types of channels can be transferred into each other based on the STM in Fig. 3.

However, by shuffling only the reserved channels we cannot guarantee that all of the free channels are queued contiguously, since the released times of the occupied channels transferred from those reserved channels can be different. Thus, we should shuffle both free and reserved channels periodically at the same time. Our idea is shown in Fig. 18.[11]

To verify the efficiency of our CS scheme, we investigated the HDP of class-1 calls (real-time interactive video) and class-5 calls (nonreal-time, e-mail) (Fig. 19). We can see that the HDP of class-1 calls has a larger improvement as compared to class-5 calls. It can be explained as follows. Class-1 calls need much longer CB than class-5 calls, since video communication needs much larger bandwidth than common text transmission. Thus, if we use the CS scheme to produce more continuous CBs, we can accept more class-1 handoff requests. Text calls do not need long CB and, thus, do not have much improvement from the CS scheme.

L. On User Mobility

In the above simulation, we assume that users are moving in all directions randomly, at the same speed. In order to investigate the influence of user mobility, we make class-5 users move toward the reference cell with a higher probability than other three directions, i.e., set up a biased mobility mode. We then compare the HDP in two cases (even mobility and biased mobile modes) in Fig. 20.

Because the biased mobile mode causes more handoff requests in the reference cell, we can see that the HDP performance is worse. However, with the increase of user speed (from walking to city-driving speed), the HDP is lower (see Fig. 20). That result can be explained by (5): a faster handoff call has a higher RO and, thus, has a lower dropping rate.

VI. Discussion

1) The effect of fading, shadowing, and cochannel interference on the proposed handoff algorithm.

The fading, shadowing, and cochannel interference can make the RSS of the MH decrease to a certain threshold below which the connection with the old BS can not be maintained. It means that the new BS should reserve bandwidth for the incoming handoff calls in time. Thus, the determination of the radius of CA (see Fig. 1) is an important issue. In our handoff algorithm, we assumed a fixed range of CA. In a practical system, a flexible range of CA can be set up based on the degrees of fading, shadowing, and cochannel interference in different time instants for different mobile users.

2) The effect of nonequal traffic loading in different cells in the simulated cellular structure.

In our simulation, we assumed equal traffic loading in each cell. If different cells have nonequal traffic loading, some cells should reserve more bandwidth than others for accommodating

[11] Note that many wireless systems typically use a hybrid scheme (FDM and TDM) to allocate bandwidth to calls. In each modulation frequency, TDM is used to allocate time slots to each call.

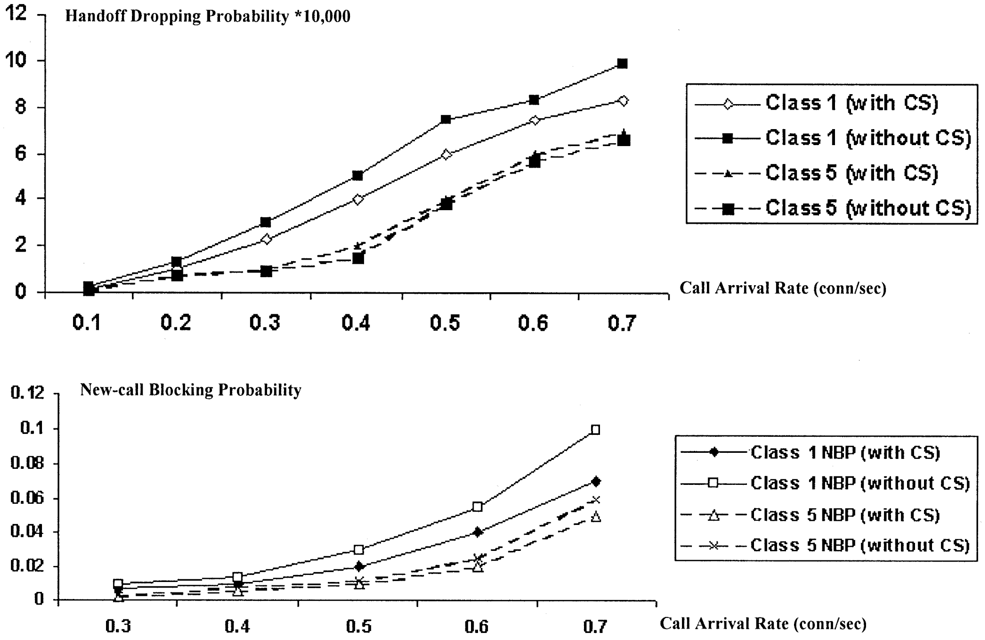


Fig. 19. Efficiency of our multiclass CS scheme.

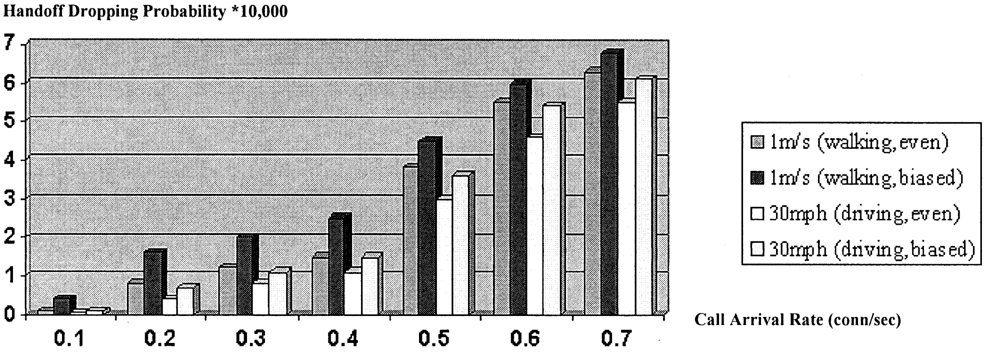


Fig. 20. Influence of user mobility mode.

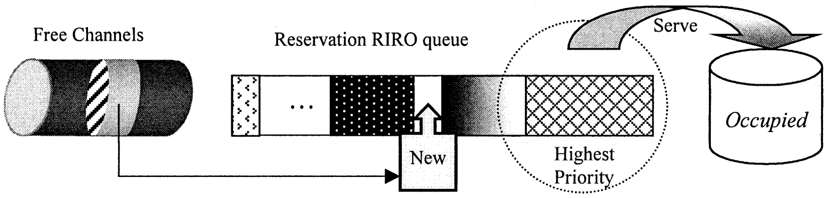


Fig. 21. Concept of the RIRO queue.

a heavier call-arriving rate. It is possible that some cells can run out of resource (reservation buffer and channel capacity). To solve that problem, we can adopt the concept of channel-borrowing mechanism (CBM) [25]. CBM states that the whole capacity of any cell is not a fixed value. Each cell only keeps a set of nominal channels and can borrow free channels from its neighboring cells to accommodate new calls. Through the using of CBM, the NBP can be further decreased. One of our future tasks is combining the CBM with our proposed approach to investigate the improvement of NBP.

3) The time order between handoff-request reservation and handoff-call admission.

In a practical mobile system, handoff-request reservation and handoff-call admission are carried out in parallel (Fig. 21). That is, the MH in neighboring cells keep submitting handoff requests to the current cell that uses the RO algorithm to order them. At the same time, the handoff request with the highest priority will be served and removed from the reservation pool. Then, the handoff request with the second priority is served, and so on. We can see that the reservation pool is just like a random-in–random-out (RIRO) queue with priority control.

VII. Conclusion

In this paper, we addressed the problem of providing M-QoS guarantee for multiclass calls in the mobile multimedia network. The network is assumed to be able to accurately predict next cells that the MH will cross. This assumption is reasonable for the developing mobile-position system such as GPS. A multiweighted algorithm for computing priorities of handoff requests was proposed in order to serve arriving multiclass calls with highly diverse QoS parameters. A feature of our approach is that we considered practical handoff-reservation duration for arriving multiclass handoff calls from the point of view of received signal strength. The future task is to derive analytical models to evaluate the performance of our CAC scheme. This paper provides a reservation-based call-admission strategy for guaranteeing the transport-layer QoS. Further work in this area will include translating the high-level resource allocations into scheduling at the low levels, such as the medium-access control (MAC) layer, so as to map the network QoS to MAC-oriented QoS.

References

[1] L. Hanzo, "Bandwidth-efficient wireless multimedia communications," *Proc. IEEE*, vol. 86, pp. 1342–1380, July 1998.

[2] P. Ramanathan *et al.*, "Dynamic resource allocation schemes during handoff for mobile multimedia wireless networks," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1270–1283, July 1999.

[3] S. K. Das *et al.*, "A call admission and control scheme for Quality-of-Service (QoS) provisioning in next generation wireless networks," *Wireless Networks 6*, pp. 17–30, 2000.

[4] C. Oliveira *et al.*, "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 858–873, Aug. 1998.

[5] K. Seal and S. Singh, "Loss profiles: a quality of service measure in mobile computing," *Wireless Networks*, vol. 2, no. 1, 1996.

[6] D. A. Levine *et al.*, "A resource estimation and cell admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Trans. Network.*, vol. 5, pp. 1–12, Feb. 1997.

[7] S. Tekinay and B. Jabbari, "A measurement-based prioritization scheme for handovers in mobile cellular networks," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 1343–1350, Oct. 1992.

[8] G. P. Pollini, "Handover rates in cellular systems: toward a closed form approximation," in *Proc. IEEE GLOBECOM '97*, 1997.

[9] O. T. W. Yu and V. C. M. Leung, "Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1208–1225, Sept. 1997.

[10] B. M. Epstein and M.Mischa Schwarz, "Predictive QoS-based admission control for multiclass traffic in cellular wireless networks," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 523–534, March 2000.

[11] ——, "QoS-based predictive admission control for multi-media traffic," in *Broadband Wireless Communications*, M. Luise and S. Pupolin, Eds. Berlin, Germany: Springer-Verlag, 1998, pp. 213–224.

[12] B. M. Epstein, "Resource allocation algorithms for multiclass wireless networks," Ph.D. dissertation, Columbia Univ., New York, 1999.

[13] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*, 3rd ed. New York: Wiley, 1997.

[14] T. Liu, P. Bahl, and I. Chlamtac, "Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 922–936, Aug. 1998.

[15] M.-H. Chiu and M. A. Bassiouni, "Predictive schemes for handoff prioritization in cellular networks based on mobile positioning," *IEEE J. Select. Areas Commun.*, vol. 18, March 2000.

[16] H. G. Ebersman and O. K.Ozan K. Tonguz, "handoff ordering using signal prediction priority queuing in personal communication systems," *IEEE Trans. Veh. Technol.*, vol. 48, pp. 20–35, Jan. 1999.

[17] J. G. Kim and M.Marwan Krunz, "Effective bandwidth in wireless ATM networks," in *Proc. MOBICOM*, Dallas, TX, 1998, pp. 233–241.

[18] S.-H. Oh and D.-W. Tcha, "Prioritized channel assignment in a cellular radio network," *IEEE Trans. Commun.*, vol. 40, July 1992.

[19] OPNET: An advanced networking simulation tool. *Available: http://www.opnet.com.* [Online]

[20] S. Sen *et al.*, "Quality-of-service degradation strategies for multimedia wireless networks," in *Proc. Vehicular Technology Conf.*, Ottawa, ON, Canada, May 1998, pp. 1884–1888.

[21] T. Liu, P. Bahl, and I. Chlamtac, "An optimal self-learning estimator for predicting inter-cell user trajectory in wireless radio networks," in *Proc. IEEE GLOBECOM*, 1997.

[22] M. Hellebrandt *et al.*, "Estimating position and velocity of mobiles in a cellular radio network," *IEEE Trans. Veh. Technol.*, vol. 46, Feb. 1997.

[23] M. Shafi *et al.*, "Wireless communications in the twenty-first century: a perspective," *Proc. IEEE*, vol. 85, pp. 1622–1637, Oct. 1997.

[24] C.-K. Toh, *Wireless ATM and Ad-Hoc Networks: Protocols and Architectures*. Norwell, MA: Kluwer , 1997, pp. 88–98.

[25] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunications systems: a comprehensive survey," *IEEE Pers. Commun.*, pp. 10–31, June 1996.

Fei Hu (M'99) received the Ph.D. degree in electrical and computer engineering from Clarkson University, Potsdam, NY, in 2002. He received the M.S. degree in telecommunication engineering from Shanghai Tiedao University, Shanghai, P. R. China, in 1996.

He is currently an Assistant Professor in the Department of Computer Engineering, Rochester Institute of Technology, Rochester, NY. His research interests include high-speed computer networks, wireless and mobile computing, Internet, and ATM.

Neeraj K. Sharma (M'90–SM'95) received the BSEE degree from the University of South Alabama, Mobile, AL, in 1987 and the M.S.E.E. and Ph.D. degrees from the University of Akron, Akron, OH, in 1989 and 1992, respectively, all in electrical engineering.

From 1993 to 1998, he was a Faculty Member with the Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Australia. From 1999 to 2000, he was an Associate Professor with the Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY. He is currently a System Engineer with Intel Corporation, Portland, OR. His research interests include fault-tolerant system design and performance and reliability analysis of computer systems and networks.