Enabling Real-Time All-IP Wireless Networks

Youssef Iraqi, Majid Ghaderi and Raouf Boutaba
University of Waterloo, Waterloo, Canada
{iraqi, ghaderi, rboutaba}@uwaterloo.ca

***Abstract*— We propose an all-IP wireless network architecture that does not require any change inside the network and can interwork with the existing wired network. This architecture is based on the operation of IntServ over DiffServ network. The standard RSVP protocol is used for signaling and reservation. The approach is based on probabilistic behavior of mobile users and does not require precise knowledge of user mobility specification. The architecture allows mobile users to specify both packet-level and connection-level Quality of Service (QoS) parameters. Simulation results show that the proposed architecture allows flexible network resource management while achieving high resource utilization.**

I. Introduction

Wireless IP networks consist of two components, wireless access networks and an IP backbone network. There are important issues that should be addressed in wireless IP networks in order to provide a seamless service on both fixed and mobile environments. Perhaps the most important issue is quality of service (QoS) provisioning. Current wired IP only offers the *best effort* service which treats all packets from all users equally. In wireless environments due to specific characteristics of wireless channel, QoS provisioning is even more challenging.

Lots of research has been done to address the QoS provisioning problem in wired IP networks and a number of QoS architectures have been proposed. To date, the IETF has adopted two architectures for providing end-to-end QoS in wired IP networks: Integrated Services (IntServ) and Differentiated Services (DiffServ).

Integrated Services have been extensively studied in wired networks. There have been many proposals for supporting real-time services in this framework. Among these proposals, the work by Jamin [1] has received a considerable attention.

In IntServ framework, flows are required to provide token bucket parameters at connection time. Each flow is described by an average rate r and bucket depth b. Jamin uses token bucket description of flows to analyze the effect of accepting a new flow on delay bounds of existing flows. If accepting a new flow will violate delay bounds of existing flows, then admission control will reject it.

Assume that there are N priority classes such that class i has higher priority than class j providing $1 \leq i < j \leq N$. Let b_i and r_i denote the sum of bucket depths and average rates for all flows in class i. And let μ denote the link capacity.

The worst-case class j delay, with FIFO discipline within the class and assuming infinite peak rates for the sources, is $D_j^* = \frac{\sum_{i=1}^{j} b_i}{\mu - \sum_{i=1}^{j-1} r_i}$ for each class j. Further, this delay is achieved for a strict priority service discipline under which class j has the least priority [2].

Talukdar *et al.* [3] extended the work of Jamin to wireless environments. Their scheme aims at accommodating real-time applications which cannot tolerate any QoS degradations due to mobility. They have assumed that the mobility of a user is predictable so that *mobility can be characterized precisely by mobility specification* which consists of the set of cells the mobile user is expected to visit during the life time of the flow. This is a very strong requirement and usually it is not feasible in practice.

According to Talukdar *et al.* [3], to implement this service model it is not enough to reserve resources along the path from the destination to the current location of the mobile host; *it is necessary to make reservations along all the paths to other locations the mobile host may visit*. However, it is not necessary to initiate the data flow along each of those paths.

In a series of papers Mahadevan *et al.* [4] have studied whether DiffServ, as defined for wired networks, is suitable for wireless networks. According to their study, several enhancements including signaling and mobility considerations should be made into DiffServ.

According to their study, due to user mobility, static provisioning of DiffServ is not sufficient in wireless environments. When a mobile executes a handoff, it is necessary to allocate resources dynamically. In addition to this, a signaling protocol is required as opposed to implicit admission control in DiffServ. This signaling protocol should consider the low bandwidth and mobility characteristics of the wireless network.

The previous wireless IntServ [3] and wireless DiffServ [4] approaches have drawbacks of their respective underlying architectures. Maintaining a per-flow state in every router is not scalable, and aggregation does not allow for quantitative services to be offered to flows. Furthermore, assuming that the exact mobility specification of a user is known beforehand is unrealistic. Also, reserving bandwidth in all the cells that the mobile user will visit is too conservative and will lead to poor network utilization. Both approaches require significant changes in the network infrastructure either by changing the behavior of routers inside the network or by introducing new signaling protocols.

In this paper, we propose an all-IP wireless network architecture that does not require any change inside the network and can interwork with the existing wired network. In this architecture, the standard RSVP protocol is used for signaling and reservation. The proposed approach allows flexible network resource management while achieving high resource

utilization. The scheme is based on probabilistic behavior of mobile users and does not require precise knowledge of user mobility specification. This architecture is flexible enough to support any target flow dropping probability. The admission control component of the architecture is based on our previous work in [5] which was proposed for admission control in wireless data networks.

The rest of this paper is organized as follows. In section II we present our all-IP wireless network architecture. Sections III, IV, and V are dedicated to the distributed resource management component of the architecture. We investigate the performance of the proposed scheme through simulation in section VI. Finally, Section VII concludes the paper.

II. All-IP wireless architecture

As we mentioned earlier, scalability concerns of the IntServ model in wired Internet forced the research community to design the DiffServ which is a simpler and more scalable framework. However, some disadvantages related to the static nature of the DiffServ model have been discovered. Therefore, cooperation of the two models have been proposed [6] to develop a dynamic and scalable architecture for wired Internet, which would be able to offer end-to-end quality of service.

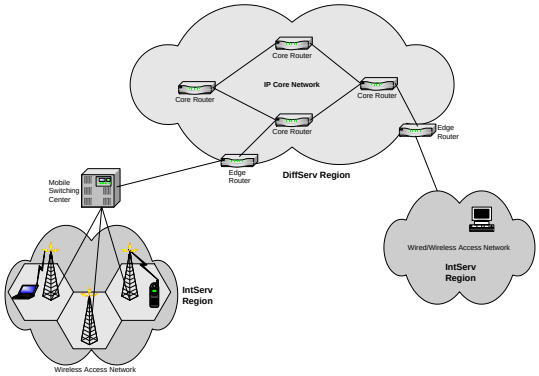

Fig. 1. All-IP wireless network architecture

In a wireless environment, resource utilization is a serious issue. The static nature of the DiffServ model will degrade the network performance. Therefore, we have designed our all-IP wireless network architecture based on the cooperation of IntServ and DiffServ as depicted in Figure 1. In this architecture IntServ operates at the wireless access network while core IP network is operating DiffServ. Since a wireless link can not accommodate a large number of flows as opposed to backbone links, scalability will not be the problem here. By using IntServ and RSVP in wireless access network, quantitative services can be offered to mobile users. When IntServ is used to provide access to DiffServ network, the most important node in the network is the edge router on the border of network regions. The router must implement two interfaces for IntServ/RSVP and DiffServ. Once a flow is admitted by the IntServ interface, its traffic is mapped to an appropriate PHB and packets will be marked accordingly. In the simplest case, guaranteed service is mapped to the EF PHB [7] and predictive service is mapped to an appropriate AF class [8].

A. Reservation Protocol

An important feature of our proposed architecture is that it operates with RSVP without any required change. This is because our admission control algorithm *explicitly* reserves resources in the local cell, where a connection request has been generated, but it *implicitly* reserves resources in other cells. Therefore, the normal operation of RSVP is enough to operate in this combined environment as addressed by Bernet *et al.* in [6].

RSVP [9] uses two types of messages to setup the reservation states in the routers, PATH message to setup the data flow path and RESV message to make the bandwidth reservation. The signaling process for end-to-end QoS starts when the sending host generates a PATH message. The PATH message is carried towards the receiving host. In the IntServ region the standard processing is applied and the PATH state is installed at the edge router and the message is sent towards the DiffServ region. In DiffServ network the PATH message is processed as normal IP packets until it reaches the IntServ region and the receiving host that will generate an RSVP RESV message. The RESV message is carried back towards the DiffServ region and the sending host. At the edge router the RESV message triggers admission control processing. If it is accepted then the message is forwarded through DiffServ region to the sending host until it reaches the local base station. When the local base station receives the RESV message it initiates a special admission control process which is described in section IV.

Hence, all the processing involved in this connection setup are the standard IntServ/DiffServ operations except the admission control process at the base stations.

B. Scheduling Algorithm

The unified scheduling algorithm of [10] is used in the IntServ region. In this scheme guaranteed service is provided by weighted fair queuing algorithm [2]. Weighted fair queuing assigns a portion of link capacity to each active flow. The scheduling discipline for predictive service is a priority queue. The flows within each predictive class are scheduled by FIFO algorithm. Best effort flows have the lowest priority in the scheduling. The satisfaction of end-to-end delay requirements is the responsibility of end systems. An end system could use any QoS routing mechanism to select a route that satisfies its end-to-end requirements.

III. System model

We consider a wireless network with a cellular infrastructure. Users can roam the network freely and experience a large number of handoffs during a typical connection. The wireless network must provide the requested level of service even if the user moves to an adjacent cell. A handoff could fail due to insufficient bandwidth in the new cell, and in such case, the connection is dropped.

To reduce the call-dropping probability, we make neighboring cells participate in the admission decision of a new user. Each cell will give its local decision and then the cell where the request was issued will decide if the new request is accepted or not. By doing so, the admitted connection will more likely survive handoffs.

As any distributed scheme, we use the notion of a cluster or group of cells (see Figure 2). Each user in the network with an active connection has a cluster associated to it.[1] The cells in the cluster are chosen by the cell where the user resides. These are the cells that are aware of the user. The shape and the number of cells of a user's cluster depend on factors such as the user's current call-holding time, QoS requirements, terminal trajectory and velocity.

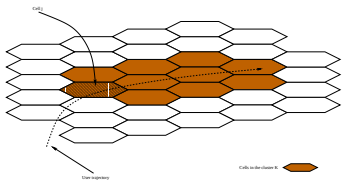

Fig. 2. Cell c_j and the cluster for a user

A. Dynamic mobile probabilities

We consider a wireless network where time is divided into equal intervals at $t_0, t_1, ..., t_m$ where $\forall i \geq 0\ t_{i+1} - t_i = \tau$. Let c_j denote a base station in the network,[2] and α a mobile terminal with an active wireless connection. Let $K(\alpha)$ denote the set of cells that form the cluster for user α. We write $[P_{\alpha,c_j,c_k}(t_0), P_{\alpha,c_j,c_k}(t_1), ..., P_{\alpha,c_j,c_k}(t_{m_\alpha})]$ for the probability that mobile terminal α, currently in cell c_j, will be active in cell c_k, and therefore under the control of base station c_k, at times t_0, t_1, t_2, ..., t_{m_α}. These probabilities are named differently by different researchers, but basically they represent the projected probabilities that mobile terminal α will remain active in the future and at a particular location. It is referred to as the Dynamic Mobile Probability (DMP) in the following. The parameter m_α represents how far in the future the predicted probabilities are computed.

DMPs may be functions of various parameters such as the handoff probability, the distribution of call duration for a mobile terminal α when using a given service class, the cell size, the user mobility profile, etc.

For each user α in the network, the cell responsible for this user determines the size of the cluster $K(\alpha)$. The cells in $K(\alpha)$ are those that will be involved in the admission process. The cell responsible for user α sends the DMPs to all members in $K(\alpha)$ specifying whether the user is a new one (in which case the cell is waiting for responses from the members of $K(\alpha)$).

In this paper, we assume that these probabilities are computed as in [5], however, the proposed admission control can use other methods to compute these probabilities as more precise and accurate methods become available. We believe that the DMPs approach is more realistic than assuming the user to have full knowledge about his/her mobility specifications.

[1]In this paper the terms "user," "connection" and "flow" are used interchangeably.

[2]We assume a one-to-one relationship between a base station and a network cell.

IV. The distributed admission control process

The distributed admission control component of the architecture, is based on our previous work in [5] which was proposed for admission control in wireless data networks. In this paper, we extend this scheme to include packet-level QoS parameters (e.g. delay), in addition to connection-level QoS parameters (e.g. call dropping probability).

Let us assume for now that each cell c_k in the cluster $K(\alpha)$ sends a response $R_{c_k}(\alpha)$ to tell the local cell c_j about its ability to support user α, and assume that $R_{c_k}(\alpha)$ is a real number between -1 (i.e. cannot accept user α), and +1 (i.e. can accept user α). Here, the admission decision takes into account the responses from all the cells in the user's cluster $K(\alpha)$. The cell has to combine the responses $R_{c_k}(\alpha)$ and take the final decision regarding the admission request. The cell has to decide the weight of each cell c_k in the user's cluster $K(\alpha)$. This will define the contribution of each cell to the final decision.

We have identified in [5] two factors for determining the weight of each cell in $K(\alpha)$: the *temporal relevance* and the *spatial relevance*.

If a cell c_{k1} in the user's cluster supports the user more than another cell c_{k2}, cell c_{k1} should have a higher impact on the admission of user α than cell c_{k2}. In general, the longer a cell is involved in supporting the user, the higher its impact. The temporal relevance $T_{c_k}(\alpha)$ represents this impact. As proposed in [5], we use the following formula for computing the temporal relevance $T_{c_k}(\alpha)$ of cell c_k:

$$T_{c_k}(\alpha) = \frac{\sum_{t=t_0}^{t=t_{m_\alpha}} P_{\alpha,c_j,c_k}(t)}{\sum_{c'_k \in K(\alpha)} \sum_{t=t_0}^{t=t_{m_\alpha}} P_{\alpha,c_j,c'_k}(t)} \qquad (1)$$

This parameter gives an indication of the percentage of time the user may spend in the considered cell c_k relative to the time the user is spending in the cluster. Equation 1 can be computed by the local cell c_j based only on the dynamic mobile probabilities.

To explain the idea of spatial relevance, we use the following example. Consider a linear highway covered by 10 square cells as in Figure 3. Assume that a new user, following the trajectory shown requests admission in cell number c_0 and that the CAC process involves five cells. Responses from cells numbered c_1, c_2, c_3 and c_4 are relevant only if cell c_0 can accommodate the user. Similarly, responses from cells c_2, c_3 and c_4 are relevant only if cell c_1 can accommodate the user when it hands off from cell c_0. This is because; a response from a cell is irrelevant if the user cannot be supported on the path to that cell. We note $S_{c_k}(\alpha)$ the spatial relevance of cell c_k for user α.

$S_{c_k}(\alpha)$ depends only on the topology of the cellular network and the responses from other cells in the cluster. In [11], we

Fig. 3. An example of a highway covered by 10 cells

proposed a method to compute the spatial relevance of a cell in one- and two-dimensional network cases. For the linear highway example of Figure 3, we use the following formula to compute the spatial relevance:

$$S_{c_0}(\alpha) = 1 \; and \; S_{c_k}(\alpha) = \prod_{l=c_1}^{c_k} f(R_{c_{l-1}}(\alpha)) \qquad (2)$$

Where $f(R) = (1+R)/2$.

This formula is chosen so that if one of the cells c_l before cell c_k has a negative response (i.e. $R_{c_l}(\alpha) = -1$), the spatial relevance of cell c_k is 0; and if all of the cells c_l before cell c_k have a positive response (i.e. $R_{c_l}(\alpha) = 1$), the spatial relevance of cell c_k is 1. Note that for each $c_k \in K(\alpha)$ we have $0 \leq S_{c_k}(\alpha) \leq 1$. Note also that in Equation 2, cell c_j (the cell receiving the admission request) has the index c_0 and that the other cells are indexed in an increasing order according to the user direction as in Figure 3.

In this distributed admission control algorithm, the cell receiving the admission request computes the sum of the product of $R_{c_k}(\alpha)$, $T_{c_k}(\alpha)$ and $S_{c_k}(\alpha)$ over c_k. The final decision of the call admission process for user α is based on:

$$D(\alpha) = \frac{\sum_{c_k \in K(\alpha)} R_{c_k}(\alpha) \times T_{c_k}(\alpha) \times S_{c_k}(\alpha)}{\sum_{c'_k \in K(\alpha)} T_{c'_k}(\alpha) \times S_{c'_k}(\alpha)} \qquad (3)$$

Note that $-1 \leq D(\alpha) \leq 1$ and that $\sum_{c'_k \in K(\alpha)} T_{c'_k}(\alpha) \times S_{c'_k}(\alpha)$ is never 0, since the spatial relevance, $S_{c_j}(\alpha)$, of cell c_j is always equal to 1, its temporal relevance $T_{c_j}(\alpha)$ is strictly positive, and all other $S_{c'_k}(\alpha)$ and $T_{c'_k}(\alpha)$ are positive or 0.

If $D(\alpha)$ is above a certain threshold, called acceptance threshold (T_{acc}), user α is accepted, otherwise, the user is rejected. The higher $D(\alpha)$, the more likely the user connection will survive in the event of a handoff.

V. Local admission control process

We show here how $R_{c_k}(\alpha)$ are computed. Without loss of generality, we assume that a user α is characterized by an average rate r_α and bucket depth b_α. The user can request any class of service (i.e. guaranteed, predictive or best effort).

A. Computing elementary responses

At each time t_0, each cell in a cluster $K(\alpha)$ involved in the admission control process for user α makes a local admission decision for different times in the future ($t_0, t_1, ..., t_{m_\alpha}$). Based on these decisions, which we call "elementary responses," the cell makes a final decision that represents its local response to the admission of user α to the network. Elementary responses are time-dependent. The computation of these responses varies according to the user location and type.

1) Local admission control at time t_0 for time t_0: Let μ denote the total capacity of the cell and N the number of predictive classes. For predictive class j, let $\nu_j = \sum_{\beta \in j} r_\beta$ and $b_j = \sum_{\beta \in j} b_\beta$ denote the aggregate rate and the aggregate bucket depth for all the flows belonging to the predictive class j. Let $\nu_G = \sum r_\beta$ denote the sum of all reserved rates for guaranteed service. Assume that flow α with token bucket parameters (r_α, b_α) has requested admission into the network.

1) *New Guaranteed Flow:* The flow α, is admitted to the guaranteed service class, if all the following conditions are satisfied at the base station:
   a) Sum of the requested flow rate r_α and the current rates of the flows in guaranteed and predictive classes should not exceed the cell capacity
   $$\mu > r_\alpha + \nu_G + \textstyle\sum_{i=1}^{N} \nu_i \qquad (4)$$
   b) The delay bounds of predictive classes should not be violated after the flow α is admitted
   $$D_j > \frac{\sum_{i=1}^{j} b_i}{\mu - \nu_G - \sum_{i=1}^{j-1} \nu_i - r_\alpha}, \; 1 \leq j \leq N. \qquad (5)$$
2) *New Predictive Flow:* The flow α, is admitted to the predictive service class l, if all the following conditions are satisfied at the base station:
   a) Sum of the requested flow rate r_α and the current rates of the flows in guaranteed and predictive classes should not exceed the cell capacity
   $$\mu > r_\alpha + \nu_G + \textstyle\sum_{i=1}^{N} \nu_i \qquad (6)$$
   b) The delay bound of the same priority class, D_l, should not be violated after the flow α is admitted
   $$D_l > \frac{\sum_{i=1}^{l} b_i + b_\alpha}{\mu - \nu_G - \sum_{i=1}^{l-1} \nu_i} \qquad (7)$$
   c) The delay bounds of the lower priority classes should not be violated after the flow α is admitted
   $$D_j > \frac{\sum_{i=1}^{j} b_i + b_\alpha}{\mu - \nu_G - \sum_{i=1}^{j-1} \nu_i - r_\alpha}, \; l < j \leq N. \qquad (8)$$

2) Local admission control at time t_0 for time $t_l (t_l > t_0)$: Each base station makes admission decision at different times in future according to the DMPs of future users.

Theorem 1: Let F be a flow described by the token bucket parameters $F = (r, b)$. And let F_1 and F_2 be two sub-flows such that $F_1 = p \times F$ and $F_2 = (1-p) \times F$, where $0 \leq p \leq 1$. If we set $F_1 = (pr, \; pb)$ and $F_2 = ((1-p)r, \; (1-p)b)$ then accepting flow F has the same effect on the delay bound experienced by all other classes as accepting both sub-flows F_1 and F_2.

Proof: Assume $F_1 = (r_1, b_1)$ and $F_2 = (r_2, b_2)$. It is clear that $r_1 = pr$ and $r_2 = (1-p)r$.

Let us now prove that $b_1 = pb$ and $b_2 = (1-p)b$. Assume that F belongs to class $1 \leq j \leq N$, where N denotes the number of predictive classes. Accepting flows from class j affects delay bound of classes at the same priority level or

at lower priority levels. Let D and D' denote the worst case delay after accepting F and $F_1 + F_2$, respectively. According to [2],

$$D' = \frac{\sum_{i=1}^{j} b_i + b_1 + b_2}{\mu - \sum_{i=1}^{j-1} r_i} = \frac{\sum_{i=1}^{j} b_i + b}{\mu - \sum_{i=1}^{j-1} r_i} = D$$

The same argument holds for delay bound of lower priority and guaranteed classes. ■

Assume user α, in cells c_j at time t_0, has a probability $P_{\alpha,c_j,c_k}(t_l)$ of being active in cell c_k at time t_l has token bucket parameters (r_α, b_α). Based on Theorem 1 cell c_k should consider a user α', for time t_l, with token bucket parameters $(P_{\alpha,c_j,c_k}(t_l) \times r_\alpha,\ P_{\alpha,c_j,c_k}(t_l) \times b_\alpha)$ and use it to perform its local admission control.

We write $E_{c_k}(\alpha, t)$ the elementary response of cell c_k for user α for time t. We assume that $E_{c_k}(\alpha, t)$ can take one of two values: -1 meaning that cell c_k cannot accommodate user α at time t; and $+1$ otherwise.

To determine the order in which a cell will perform its admission control it sorts the users in decreasing order of their DMPs.

B. Computing the final responses and sending the results

If, for user α, cell c_k has a response $E_{c_k}(\alpha, t)$ for each t from t_0 to t_{m_α} with a corresponding DMPs $P_{\alpha,c_j,c_k}(t_0)$ to $P_{\alpha,c_j,c_k}(t_{m_\alpha})$, then to compute the final response those elementary responses are weighted with the corresponding DMPs. The final response from cell c_k to cell c_j concerning user α is then :

$$R_{c_k}(\alpha) = \frac{\sum_{t=t_0}^{t=t_{m_\alpha}} E_{c_k}(\alpha, t) \times P_{\alpha,c_j,c_k}(t) \times C_{c_k}(\alpha, t)}{\sum_{t=t_0}^{t=t_{m_\alpha}} P_{\alpha,c_j,c_k}(t)} \quad (9)$$

where $C_{c_k}(\alpha, t)$ is the confidence that cell c_k has about the elementary response $E_{c_k}(\alpha, t)$. To normalize the final response, each elementary response is also divided by the sum of the DMPs in cell c_k over time t. Cell c_k, then, sends the response $R_{c_k}(\alpha)$ to the corresponding cell c_j. Note that $R_{c_k}(\alpha)$ is a real number between -1 and 1.

VI. Performance evaluation

A. Simulation model

All the evaluations are done for mobile terminals that are traveling along a highway as in Figure 3. In our simulation study we have the following simulation parameters and assumptions[3]:

1) The time is quantized in intervals of $\tau = 10s$.
2) The whole system is composed of 10 linearly-arranged cells, laid at 1-km intervals, numbered from 1 to 10.
3) Cells 1 and 10 are connected so that the whole cellular system forms a ring architecture as assumed in [3]. This avoids the uneven traffic load that would be experienced by these border cells otherwise.
4) Connection requests are generated in each cell according to a Poisson process with rate λ (connections/second). A newly generated mobile terminal can appear anywhere in the cell with equal probability.
5) Mobile terminals speeds are uniformly distributed between 80 and 120 km/h, and mobile terminals can travel in either of two directions with equal probability.
6) Each cell has a capacity of 2000 kb.
7) We consider two possible types of traffic: F1 and F2.
8) F1 = (64 kb/s, 1 kb), delay = 16 ms.
9) F2 = (128 kb/s, 10 kb), delay = 80 ms.
10) and 2 service classes: guaranteed (G) and predictive (P).
11) 50% ask for (G) service (90% F1 and 10% F2)
12) 50% ask for (P) service (80% F1 and 20% F2)
13) Connection lifetimes are exponentially distributed with a mean value of 180 seconds.

For the distributed admission control (DCAC) scheme we also assume that:

1) The DMPs are computed as in [5].
2) The weights are computed using Equation 1 and 2.
3) The confidence degrees are computed as in [5].
4) $m_\alpha = 18$. This means that the DMPs are computed for 18 steps in the future. And $K(\alpha) = 2$. This means that one cell in the direction of the user and the cell where the user resides form the cluster.

Five hours of traffic is simulated in each experiment that has been repeated several times to get results within the 95% confidence interval.

B. Simulated admission control algorithms

In addition to the proposed distributed admission control algorithm (DCAC), we have simulated the mobility independent admission control (MICAC) [3]. This scheme assumes that the mobility specification of the mobile is precisely known at connection setup time. In this scheme a flow is accepted only if all the cells that belong to the mobility specification have the requested bandwidth available for the lifetime of the flow. By reserving the requested bandwidth everywhere, MICAC achieves a zero call dropping probability.

We simulated a system that uses our distributed admission control scheme, and we computed important statistics like the Call Dropping Percentage (CDP), the Call Blocking Percentage (CBP) and the Average Bandwidth Utilization (ABU). Also we simulated a system that uses the MICAC scheme, and computed the same statistics.

The algorithms have been simulated subjected to loads of 1000, 2000 and 4000, which corresponds to normalized loads of 0.5, 1 and 2.

C. Simulation results

Even if our DCAC scheme can achieve any target CDP value, to compare with the MICAC scheme we have chosen the acceptance threshold so that DCAC achieves a zero CDP. The results are shown in Figures 4 and 5.

Figure 4, depicts the CBP achieved by the two schemes. The x axis represents the normalized load. According to the figure, MICAC has a higher CBP than DCAC irrespective of the offered load. This is because MICAC reserves the requested

[3]The simulation parameters used here are those used by most researchers

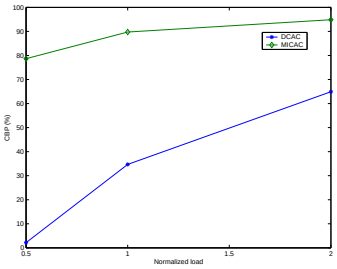

Fig. 4. CBP for the two schemes

bandwidth in all the cells that will be visited by the mobile for the lifetime of the flow. This reserved bandwidth prevents other cells from accepting new flows.

The MICAC behavior has a significant effect on the bandwidth utilization as shown in Figure 5.

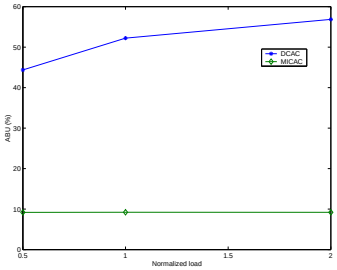

Fig. 5. ABU for the two schemes

Indeed, irrespective of the offered load, MICAC ABU is about 10%, while DCAC can reach more than 50% of average bandwidth utilization. The implicit reservation mechanism of DCAC allows for spatial statistical multiplexing among accepted flows. To guarantee the QoS, the only requirement is to make sure that the requested bandwidth is available when needed. If the bandwidth is reserved in a cell while the mobile is elsewhere, the bandwidth is wasted as it is the case in MICAC. DCAC performs implicit reservation only for times when the mobile is expected to be in a particular cell. By taking into consideration, both spatial and temporal mobile behavior, DCAC is able to better manage the network resources and accept more flows without sacrificing the QoS. Figure 5 also shows that DCAC ABU increases as the offered load increases. DCAC is able to take advantage of the spatial statistical multiplexing. MICAC, on the other hand, reserves bandwidth in all the cells that will be visited by the mobile, and is not able to accept any more flows irrespective of the offered load.

Furthermore, the acceptance threshold of DCAC allows the scheme to achieve any target CDP value. Indeed, we believe that 0% CDP is a very restrictive condition that will lead to poor network utilization. Most applications will not have such a strong requirement, rather, a target CDP of 5% to 10% seems more reasonable. MICAC does not allow such behavior. Table I, shows the performance of DCAC when the target CDP is 5% and 10% in case the normalized load is equal to 2. According to the table, DCAC achieves even higher bandwidth utilization if we allow for higher CDP. Note that any available bandwidth not used by guaranteed and predictive service classes can be used by the best effort service class.

CDP	CBP	ABU
5%	50%	76%
10%	45%	80%

TABLE I

DCAC PERFORMANCE FOR DIFFERENT TARGET CDP

VII. CONCLUSION

In this paper, we propose an all-IP wireless network architecture. It is based on the cooperation of IntServ and DiffServ models. The access network operates IntServ while the DiffServ is used in the core network. We take advantage of both worlds to develop a dynamic and scalable architecture for all-IP wireless networks. The standard RSVP protocol is used for signaling and reservation. We have also proposed a distributed admission control that accommodates both packet-level and connection-level QoS requirements. To our knowledge this is the first scheme to support these features in a wireless IP network. Simulation results show that our scheme achieves higher bandwidth utilization than another scheme designed to achieve a zero call dropping probability. Furthermore, the proposed scheme is flexible enough to support any target call dropping probability and achieves even higher bandwidth utilization.

REFERENCES

[1] S. Jamin, S. Shenker, and P. Danzig, "Comparison of measurement-based admission control algorithms for controlled-load service," in *Proc. INFOCOM'97*, vol. 3, (Kobe, Japan), pp. 973–980, April 1997.

[2] A. K. Parekh, *A generalized processor sharing approach to flow control in integrated services network*. Ph.D. Dissert., Dept. Electrical Engineering and Computer Science, MIT, Cambridge, USA, 1996.

[3] A. K. Talukdar, B. Badrinath, and A. Acharya, "Integrated services packet networks with mobile hosts: Architecture and performance," *ACM/Baltzer Wireless Networks*, vol. 5, pp. 111–124, 1999.

[4] I. Mahadevan and K. M. Sivalingam, "Architecture and experimental framework for supporting QoS in wireless networks using differentiated services," *ACM/Kluwer Mobile Networks and Applications*, vol. 6, pp. 385–395, 2001.

[5] Y. Iraqi and R. Boutaba, "When is it worth involving several cells in the call admission control process for multimedia cellular networks?," in *Proc. IEEE ICC*, pp. 336–340, 2001.

[6] Y. Bernet *et al.*, "A framework for integrated services operation over DiffServ networks." RFC 2998, IETF, November 2000.

[7] B. Budiardjo, B. Nazief, and D. Hartanto, "Integrated services to differentiated services packet forwarding: Guaranteed service to expedited forwarding PHB," in *Proc. IEEE LCN 2000*, (Tampa, USA), pp. 324–325, November 2000.

[8] T. Chahed *et al.*, "On mapping of QoS between integrated services and differentiated services," in *Proc. IWQoS 2000*, (Pittsburgh, USA), pp. 173–175, June 2000.

[9] B. Braden *et al.*, "Resource reservation protocol (RSVP): Version 1 functional specification." RFC 2205, IETF, September 1997.

[10] D. Clark, S. Shenker, and L. Zhang, "Supporting real-time applications in an integrated services packet network: Architecture and mechanism," in *Proc. SIGCOMM'92*, (Baltimore, USA), pp. 14–26, October 1992.

[11] Y. Iraqi and R. Boutaba, "Weight allocation in distributed admission control for wireless networks." submitted paper, 2003.