A Multi-agent System for Resource Management in Wireless Mobile Multimedia Networks

Youssef Iraqi and Raouf Boutaba

University of Waterloo, DECE, Waterloo, Ont. N2L 3G1, Canada
{iraqi,rboutaba}@bbcr.uwaterloo.ca

Abstract. This paper introduces a multi-agent system for resource management developed for cellular mobile networks. The main feature of the proposed multi-agent system is a more efficient support for mobile multimedia users having dynamic bandwidth requirements. This is achieved by reducing the call dropping probability while maintaining a high network resource utilization. A call admission algorithm performed by the multi-agent system is proposed in this paper and involves not only the original agent (at the cell handling the new admission request) but also a cluster of neighboring agents. The neighboring agents provide significant information about their ability to support the new mobile user in the future. This distributed process allows the original agent to make a more clear-sighted admission decision for the new user. Simulations are provided to show the improvements obtained using our multi-agent system.

1 Introduction

Cellular mobile networks have to continue supporting their mobile users after they leave their original cells. This poses a new challenge to resource management algorithms. For instance a call admission process should not only take into consideration the available resources in the original cell but also in neighboring cells as well.

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

In this paper we propose a Multi-Agent system for call admission resource management designed for wireless mobile multimedia networks. The call admission process involves not only the cell that receives the call admission request but also a cluster of neighboring cells. The agents share important resource information so the new admitted user will not be dropped due to handoffs. Consequently, the network will provide a low call dropping probability while maintaining a high resource utilization.

The paper is organized as follows. In section 2, we describe the multi-agent

A. Ambler, S.B. Calo, and G. Kar (Eds.): DSOM 2000, LNCS 1960, pp. 218–229, 2000.

architecture proposed in this paper. Section 3 defines the dynamic mobile probabilities used by our multi-agent system. In section 4 we present the call admission process performed locally by agents in our system. Section 5 introduces the overall admission process involving a cluster of agents. Section 6 gives a description of agent's cooperation. Section 7 discusses the conducted simulation parameters and results. Finally, section 8 concludes this paper.

2 The Multi-agent Architecture

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

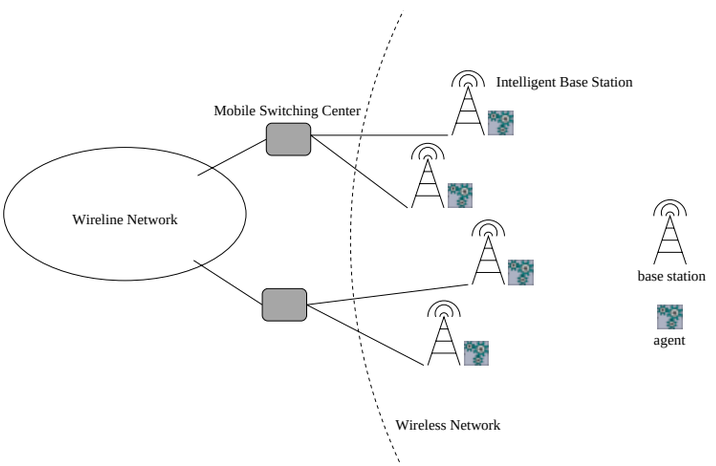

Fig. 1. A Wireless Network and the Multi-agent System

To reduce the call dropping probability, we propose a multi-agent system that allows neighboring cells to participate in the decision of a new user admission. Each cell or base station has an agent running on it. The agent keeps track of the cell's resources and shares information with neighboring agents to better support mobile users. Each involved agent in an admission request will give its local decision according to its available resources and information from other

agents and finally the agent at the cell where the request was issued will decide if the new request is accepted or not. By doing so, the new admitted connection will have more chances to survive after experiencing handoffs.

We use the notion of a cluster similar to the shadow cluster concept [5]. The idea is that every connection exerts an influence upon neighboring base stations. As the mobile terminal travels to other cells, the region of influence also moves. The set of cells influenced by a connection are said to constitute a cluster (see figure 2). Each user[1] in the network, with an active connection has a cluster associated to it. The agents in the cluster are chosen by the agent at the cell where the user resides. The number of agents of a user's cluster depend on factors such as user's current call holding time, user's QoS requirements, terminal trajectory and velocity.

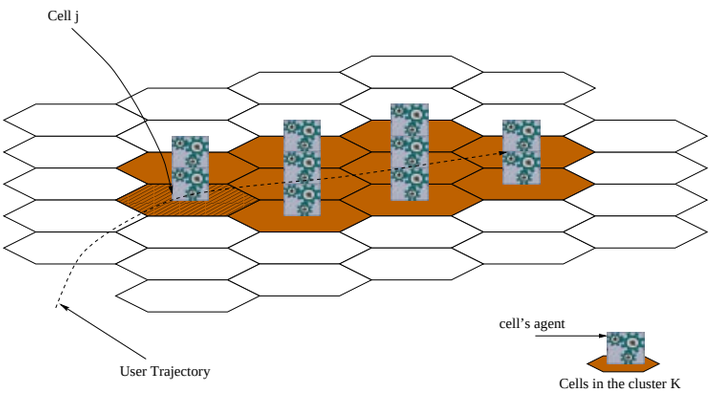

Fig. 2. Example of a User's Cluster

3 Dynamic Mobile Probabilities

We consider a wireless network where the time is divided in equal intervals at $t = t_1, t_2, ..., t_m$. Let j denote a base station (and the corresponding agent) in the network, and x a mobile terminal with an active wireless connection. Let $K(x)$ denote the set of agents that form the cluster for the active mobile terminal x. We denote $P_{x,j,k}(t) = [P_{x,j,k}(t_0), P_{x,j,k}(t_1), ..., P_{x,j,k}(t_{m_x})]$ the probability that mobile terminal x, currently in cell j, to be active in cell k, and therefore under the control of agent k, at times t_0, t_1, t_2, ..., t_{m_x}. $P_{x,j,k}(t)$ represents the projected probabilities that a mobile terminal will remain active in the future and at a particular location. It is referred to as the Dynamic Mobile Probability

[1] In the rest of the paper the term "user" and "connection" are used interchangeably

(DMP) in the following. The parameter m_x represents how far in the future the predicted probabilities are computed. It is not fixed for all users and can depend of the user QoS or the actual connection elapsed time.

Those probabilities may be function of several parameters such as: residence time of mobile x in cell j, handoff probability, the distribution of call length for a mobile terminal x when using a given service class, cell size and user mobility profile.

Of course, the more information we have, the more accurate are the probabilities, however the more complex is their computation.

For each user x in the network, the agent that is responsible for, decides the size of the cluster $K(x)$, those are the agents involved in the admission process, and sends the DMPs to all members in $K(x)$. The agent must specify if the user is a new one (in which case the agent is waiting for responses from the members of $K(x)$) or not.

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

DMPs can also be complex for example by including information about user mobility profiles. A method for computing dynamic mobile probabilities taking into consideration mobile terminal direction, velocity and statistical mobility data, is presented in [2]. Other schemes to compute these probabilities are presented in [3] [4]. To compute these probabilities, one can also use mobiles' path/direction information readily available from certain applications, such as the route guidance system of the Intelligent Transportation Systems with the Global Positioning System (GPS).

4 Local Call Admission Decision

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

In conjunction with the emergence of adaptive multimedia encoding [6] [7] [8], QoS adaptation schemes have been proposed to reduce handoff drops. In these schemes a connection's QoS can be downgraded if the available bandwidth in the new cell is not sufficient [9] [4]. Such schemes can be easily integrated in our system as part of the local call admission decision.

4.1 Computing Elementary Responses

At each time t_0 each agent, in a cluster $K(x)$ involved in our call admission (CA) process for user x, makes a local CA decision for different times in the future

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

User Types. An agent may be involved in the processing of different types of user. Possible user types at time t_0 are:

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

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

Local Call Admission Decision at Time t_0 for Time t_0. An agent can apply any local call admission algorithm to compute the elementary responses. In this work we assume that the agents use the Equivalent Bandwidth approach to compute these responses. Example of such a scheme is described in [1]. Other schemes can be downloaded to the agents from the management station.

The processing of local new users will be explained in section 5.

Local Call Admission Decision at Time t_0 for Time t_l ($t_l > t_0$). Each agent computes the equivalent bandwidth at different times in the future according to the DMPs of future users.

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

We denote $r_k(x, t)$ the elementary response of agent k for user x for time t. The agent sets in which order of users it will perform its call admission process. For instance, the agent can sort users in a decreasing order of their DMPs. If we assume that user x_i has higher priority than user x_j for all $i < j$, then to compute elementary responses for user x_j, we assume that all users x_i with $i < j$ that have a positive elementary response are accepted. As an example, if an agent wants to compute the elementary response r for user x_4, and we have already computed r for users $x_1 = 1$, $x_2 = 1$ and $x_3 = 0$, then to compute r for x_4 the agent assumes that user 1 and 2 are accepted in the system but not user x_3.

We propose also that the agent reserves some bandwidth in case of an erroneous prediction. This amount of reserved bandwidth is a parameter of our scheme and can be tuned to have the best performance. The choice of this parameter depends on the precision of the DMPs.

4.2 Computing the Final Responses and Sending the Results

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

We denote $C_k(x,t)$ the confidence that agent k has about its elementary response $r_k(x,t)$. The question arises on how the agent can compute (or simply choose) the confidence degrees $C_k(x,t)$, typically between 0% and 100%. One way to compute the confidence degrees is to use the percentage of available bandwidth when computing the elementary response as an indication of the confidence the agent may have in this elementary response.

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

$$R_k(x) = \frac{\sum_{t=t_0}^{t=t_{m_x}} r_k(x,t) \times P_{x,j,k}(t) \times C_k(x,t)}{\sum_{t=t_0}^{t=t_{m_x}} P_{x,j,k}(t)} \tag{1}$$

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

5 Taking the Final Decision

Here the decision takes into consideration the responses from all agents in the user's cluster. The admission process concerns only new users seeking admission to the network and not already accepted users.

We assume that agent j has already decided the cluster $K(x)$ and that agent j has already assigned to each agent k in the cluster $K(x)$ a weight $W_k(x)$. Each weight represents the importance of the contribution of the associated agent to the global decision process. Usually an agent that is involved more in supporting the user has a high weight value. Weights $W_k(x)$ depend on the DMPs and the time t.

We suggest to use the following formula to compute the weights $W_k(x)$:

$$W_k(x) = \frac{\sum_{t=t_0}^{t=t_{m_x}} P_{x,j,k}(t)}{\sum_{k' \in K} \sum_{t=t_0}^{t=t_{m_x}} P_{x,j,k'}(t)} \tag{2}$$

If we assume that each response $R_k(x)$, from agent k, is a percentage between 0% (can not be supported at all) and 100% (can be supported), then the agent computes the sum of $R_k(x) \times W_k(x)$ over k.

The final decision of the call admission process for user x is based on

$$D(x) = \sum_{k \in K} R_k(x) \times W_k(x) \tag{3}$$

If $D(x)$ is higher than a certain threshold then, user x is accepted; otherwise the user is rejected. The threshold can be specified by the user. The more higher is the threshold the more likely the user connection will survive in the event of a handoff.

Combining eq. 1 and eq. 2, eq. 3 can be written as:

$$D(x) = \frac{1}{\alpha} \sum_{k \in K} \sum_{t=t_0}^{t=t_{m_x}} r_k(x,t) \times P_{x,j,k}(t) \times C_k(x,t) \tag{4}$$

With $\alpha = \sum_{k' \in K} \sum_{t=t_0}^{t=t_{m_x}} P_{x,j,k'}(t)$. Only the value $\sum_{t=t_0}^{t=t_{m_x}} r_k(x,t) \times P_{x,j,k}(t) \times C_k(x,t)$ should be computed locally in each cell, and the final result is then, simply the sum of all responses from all the agents in the cluster K divided by α.

6 Agent's Cooperation

Each time t, an agent j should decide if it can support new users. It decides locally if it can support users of type 1 and 2 that have higher priority than other type of users (cf. user types in section 4.1). This is because, from a user point of view, receiving a busy signal is more bearable than having a forced termination. The agent also sends the DMPs to other agents and informs them about its users of type 3 (step 2 in figures 3, 4). Only those who can be supported locally are included, other users of type 3 that can not be accommodated locally are rejected. At the same time, the agent receives DMPs from other agents and is informed about users of type 4.

Using equation 1, the agent decides if it can support users of type 4 in the future and it sends the responses to the corresponding agents (step 3 in figures 3, 4). When it receives responses from the other agents concerning its users of type 3, it performs one of the two following steps (step 4 in figures 3, 4): If the agent can not accommodate the call, the call is rejected. If the agent can accommodate the call, then the call admission decision depends on equation 4.

Figure 3 shows the different steps of agent's cooperation when processing an admission request. Figure 4 depicts the admission process diagram at the agent receiving the admission request and at an agent belonging to the cluster. Because the admission request is time sensitive the agent waiting for responses from the agents in the cluster will wait until a predefined timer has expire then he will assume a negative response from all agents that could not respond in time.

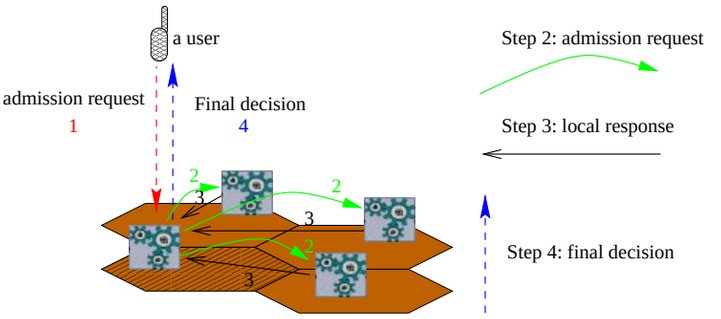

Fig. 3. Agent's Cooperation for the Admission of a User

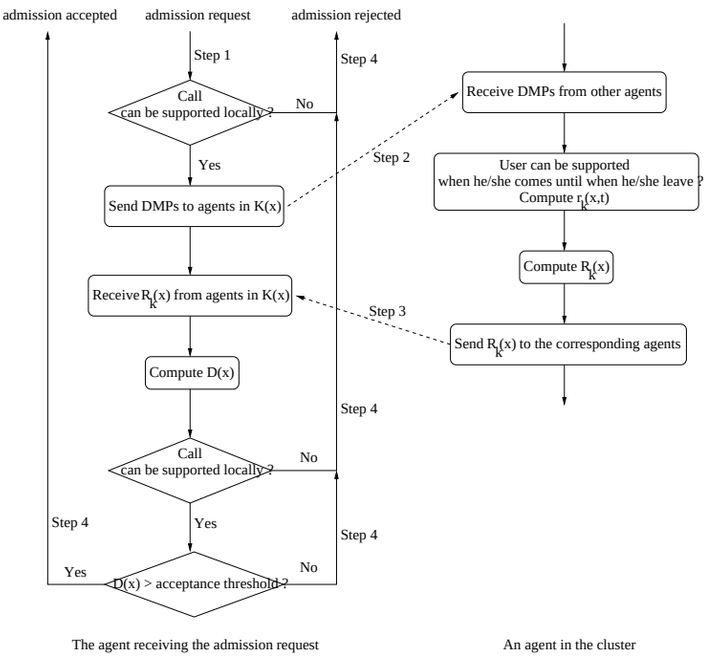

Fig. 4. Admission Process Diagram

7 Performance Evaluation

7.1 Simulation Parameters

For the sake of simplicity, we evaluate the performance of our Multi-Agent system for mobile terminals which are traveling along a highway. This is a simplest environment representing a one-dimensional cellular system. In our simulation study we have the following assumptions:

1. The time is quantized in intervals $T = 10s$
2. The whole cellular system is composed of 10 linearly-arranged cells, laid at 1-km intervals (see figure 5).

Fig. 5. A Highway Covered by 10 Cells

3. During each time interval, connection requests are generated in each cell according to Poisson process. A newly generated mobile terminal can appear anywhere in the cell with equal probability.
4. Mobile terminals can have speeds of: 70, 90, or 105 km/h. The probability of each speed is 1/3, and mobile terminals can travel in either of two directions with equal probability.
5. We consider three possible types of traffic: voice, data, or video. The probabilities of these types are 0.7, 0.2, 0.1 respectively. The number of bandwidth units (BUs) required by each connection type is: voice = 1, data = 5, video = 10. Note that fixed bandwidth amounts are allocated to users for the sake of simplicity.
6. Connection lifetimes are exponentially-distributed with mean value equal to 180 seconds.
7. Each cell has a fixed capacity of 40 bandwidth units.
8. m_x is fixed for all users and for the duration of the connection and is equal to 18. This means that the DMPs are computed for 18 steps in the future.
9. The size of the cluster $K(x)$ is fixed for all users and is equal to 5. This means that four cells in the direction of the user along with the cell where the user resides form the cluster.
10. We simulate a total of 4 hours of real-time highway traffic, with a constant cell load equal to 360 new calls/h/cell.
11. The DMPs are computed as in [2].
12. All users have the same threshold.
13. The confidence degree is computed as follows: $Confidence = e^{(1-p)} * p^3$ where p is a real number between 0 and 1 representing the percentage of available bandwidth at the time of computing the elementary response.

7.2 Simulation Results

In our simulations, a user x requesting a new connection is accepted into a cell only if the final decision $D(x)$ is above an acceptance threshold value. We varied this threshold value to observe its effect on the call dropping percentage and the average bandwidth utilization in the cells of the network.

By varying the value of the threshold in the simulations, we were able to decrease the percentage of dropped calls while maintaining a good average bandwidth utilization.

Figure 6 depicts the average bandwidth utilization of the cells in the network, and the corresponding percentage of dropped calls for different acceptance threshold values.

The top curve represents the average number of BU's that are used in all

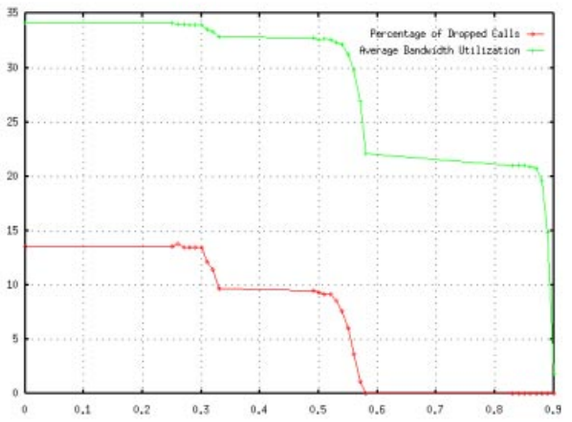

Fig. 6. Average Bandwidth Utilization and Percentage of Dropped Calls According to the Acceptance Threshold Value

cells in the network, considering the entire simulation time. When the threshold is equal to zero, the average bandwidth utilization is at its maximum value. In this case, the maximum bandwidth utilization is approximately equal to 34 BU's. The bottom curve depicts the percentage of dropped calls in the network. The highest percentage of dropped calls also occurs when the threshold is equal to zero; in this case, all connection requests are accepted regardless of the final decision $D(x)$, as long as there is available bandwidth in the cells where the connections are requested. For the simulated cell load, the maximum percentage of dropped calls is equal to 14%. By adjusting the threshold value, our Mutli-Agent system can control the percentage of calls that will be dropped. For example, with a threshold value of 57%, the percentage of dropped calls is reduced to the value of 1% while maintaining at the same time a high average bandwidth uti-

lization value of 27 BUs. The proposed scheme allow a tradeoff between average bandwidth utilization and the percentage of dropped calls. If the threshold value is 83% then no calls need to be dropped with a corresponding average bandwidth utilization of 21 BUs. Thus, the proposed scheme can reduce the percentage of dropped calls with an acceptable degradation in total bandwidth utilization.

8 Conclusion

In this paper, we have described a Multi-Agent system for resource management suitable for wireless multimedia networks. The proposed system operates in a distributed fashion by involving, in a call admission decision, not only the agent receiving the admission request, but also a determined number of neighboring agents. The goals underlying the design of our algorithm are: (1) to support mobile multimedia users with dynamic bandwidth requirements; (2) to reduce the call dropping probability while maintaining a high network resource utilization; and (3) to distribute call admission decision among clusters of neighboring agents to allow more clear-sighted decisions and hence a better user survivability in the network. More technically, our algorithm can integrate easily any method for computing Dynamic Mobile Probabilities (DMPs). It can also rely on different local call admission schemes including those designed for adaptive multimedia applications. Those schemes can be downloaded to the agents by the management system.

Simulations results have shown that by implementing the proposed multi-agent system, the wireless netwok is able to lower the call dropping probability while offering a high average bandwidth utilization. The wireless network is also able to maintain a high acceptance probability for new users. The signaling load induced by agent's communication is considered here acceptable as far as it only involves few messages exchanged between agents through the wired network which is assumed to be of high capacity. More simulations are envisaged in the future to evaluate our multi-agent system in more sophisticated situations, for example with users having dynamic bandwidth requirements, cell loads, and traffic distributions. Also envisaged is studying the influence of the number of agents involved in a call admission decision.

References

1. J. Evans and D. Everitt, 'Effective bandwidth based admission control for multi-service CDMA cellular networks,' IEEE Trans. Vehicular Tech., Vol 48, No 1, pp 36-46, January 1999.
2. D. A. Levine, I. F. Akyildz and M. Naghshineh, 'A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept,' IEEE/ACM Transactions on Networking, vol. 5, no. 1, Feb. 1997.
3. Sunghyun Choi and Kang G. Shin, 'Predictive and Adaptive Bandwidth Reservation for Hand-Offs in QoS-Sensitive Cellular Networks,' in Proc. ACM SIGCOMM'98, pp. 155-166, Vancouver, British Columbia, September 2-4, 1998.

4. Songwu Lu and Vaduvur Bharghavan, 'Adaptive Resource Management Algorithms for Indoor Mobile Computing Environments,' in Proc. ACM SIGCOMM'96, pp. 231-242, August 1996.
5. D. A. Levine, I. F. Akyildz and M. Naghshineh, 'The shadow cluster concept for resource allocation and call admission in ATM-based wireless networks,' in Proc. ACM Int. Conf. Mobile Comp. Networking MOBICOM'95, Berkeley, CA, pp. 142-150, Nov. 1995.
6. R. Rejaie, M. Handley, and D. Estrin, 'Quality Adaptation for Congestion Controlled Video Playback over the Internet,' in Proc. ACM SIGCOMM'99, September 1999.
7. S. McCanne, M. Vetterli, and V. Jacobson, 'Low-Complexity Video Coding for Receiver-Driven Layerd Multicast,' IEEE Journal on Selected Areas in Communications, Vol. 15, No. 6, pp. 983-1001, August 1997.
8. J. Hartung, A. Jacquin, J. Pawlyk, and K. Shipley, 'A Real-time Scalable Video Codec for Collaborative Applications over Packet Networks,' ACM Multimedia'98, pp. 419-426, Bristol, September 1998.
9. K. Lee, 'Supporting mobile multimedia in integrated service networks,' ACM Wireless Networks, vol. 2, pp. 205-217, 1996