Proactive Resource Management for the Mitigation of Service Discontinuation in Mobile Networks

George Alyfantis, Stathes Hadjiefthymiades, and Lazaros Merakos
Communication Networks Laboratory
Department of Informatics and Telecommunications, University of Athens
Athens 15784, Greece
{alyf, shadj, merakos}@di.uoa.gr

***Abstract*— A scheme for the proactive allocation of network resources to mobile users, based on a pricing framework, is proposed. The objective is the reduction of service discontinuation events attributed to handovers in the cellular infrastructure. The future base station, where network resources have to be reserved proactively, is determined by means of a path prediction algorithm. The network receives a fee for providing the advance reservation service to the user. The exact price is determined after a sequential bargaining procedure, modeled as a two-person non-cooperative game between the mobile user and the network. The efficiency of the proposed scheme is evaluated through simulations.**

Keywords*— *Proactive resource management, bargaining, game theory, pricing

I. INTRODUCTION

The occurrence of handovers in cellular mobile networks is a very important issue and the main research drive for the design of resource management algorithms. A session (call) may have to be terminated when the mobile terminal (MT) is handed over to a new base station (BS), which does not have adequate resources to support the quality of service (QoS) requirements of the particular session. This event is referred to as *handover blocking*, and is very annoying for the user [2].

The handover blocking probability may be reduced through the use of proactive resource reservation in the neighborhood of the present cell of a MT [6]. After the occurrence of the handover, the MT does not compete for a share of the finite resources but enjoys a pre-arranged configuration. Network resources that could be managed through such schemes include (but are not limited to) bandwidth, MAC frames/slots, files, and packets [1].

The scheme introduced in this paper is based on a pricing model for the proactive resource reservation. A user agent residing in the MT negotiates with the network, in order to pre-reserve an amount of resources in the most likely to be visited cell. The network is paid for this premium service, so as to compensate for keeping these resources unused. The price that the user pays is determined after a bargaining process that is modeled as a two-person non-cooperative game.

The rest of the paper is organized as follows. We discuss related prior work, in Section II. In Section III, we discuss how the considered commodity (i.e., bandwidth) is valuated by the mobile user and the network. Section IV is devoted to the algorithms used for the management of network resources. Section V provides simulation results, while Section VI our conclusions.

II. PRIOR RELATED WORK

In [3], a simple adaptive call admission control (CAC) algorithm, which takes advantage of guard channels, has been introduced. Other adaptive reservation and admission control schemes based on guard resources are studied in [4], and [5]. In [6], time is partitioned into equal intervals, and the probability of each MT being in any cell within the shadow cluster (set of BSs to which a MT will probably attach in the near future) for future time intervals is estimated. The BSs exchange information on the predicted bandwidth demands for future time intervals in order to determine the feasibility of admitting new call requests. In [9], an advance reservation is kept valid for a limited period of time starting at the *Earliest Arrival Time* and ending at the *Latest Arrival Time* for a specific cell-user combination. The two times are calculated geometrically over the cellular network topology. There are also simpler schemes assuming a fixed amount of guard bandwidth as in [7], which, however, are inefficient under non-stationary traffic conditions. From the above, it is evident that the management of guard resources has attracted significant research efforts. However, to the best of our knowledge, this problem has not been studied yet using concepts from game theory and pricing.

III. MT AND BS RESOURCE VALUATION

In this paper, we propose a proactive resource management scheme for the reduction of handover blocking events in mobile networks. The resources of the BS are used for both active sessions, and sessions anticipated from neighboring cells. The BS receives a fee for providing the advance reservation service to the MT, in order to compensate for keeping those resources unused. The exact price is determined by means of a bargaining process (described in Section IV), which takes into account the resource valuations of both the BS and the MT.

In this section, we discuss how the MTs and the BSs valuate the network resources for the considered advance reservations. Let p_i ($/(Kb/s)) be the unit price agreed between the MT and the BS for session i (for details on the derivation of p_i see Section IV). The utility function of the MT and the BS is given in (1) and (2), respectively.

$$U_{MT} = Benefit_{MT,i} - p_i \quad (1)$$

$$U_{BS} = p_i - Cost_{BS} \quad (2)$$

In (1), $Benefit_{MT,i}$ denotes the valuation of the MT for one unit of bandwidth. Such valuation depends upon two factors (see (3)): a user-specific weight, $f \in [0,1]$, denoting the criticality of the application session (which can derive from a user profile), and the observed duration of the session until present time, d_i. We assume that a session that has lasted for long will persist in the future, in which case the corresponding bandwidth is considered as of great importance. This assumption is usually adopted in the engineering of WWW caches [11].

$$Benefit_{MT,i} = f \cdot \frac{d_i}{d_{\max}} \quad (3)$$

d_{max} denotes the largest session duration that the MT has observed so far, and is used for normalization purposes. For valuations in the $[b_{min}, b_{max}]$ interval, $Benefit_{MT}$ can be linearly transformed to assume values in the desired range, i.e., $Benefit'_{MT} = b_{min} + (b_{max} - b_{min}) \cdot Benefit_{MT}$.

In (2), $Cost_{BS}$ represents the cost valuation of the BS for making a reservation. Such valuation is proportional to the local resource utilization, as shown in (4).

$$Cost_{BS} = \frac{C_{Total} - C_{Free}}{C_{Total}} \quad (4)$$

C_{Free} denotes the current amount of free resources, while C_{Total} the total amount of resources in the BS. For valuations in the [s_{min}, s_{max}] interval, $Cost_{BS}$ can be linearly transformed to assume values in the desired range, i.e., $Cost'_{BS} = s_{min} + (s_{max} - s_{min}) \cdot Cost_{BS}$.

IV. ARCHITECTURE ANALYSIS

In this section, we describe the proposed architecture, and the algorithms used for the management of network resources. The proposed architecture relies on three basic components (Fig. 1): 1) a *path prediction algorithm* (PPA), 2) a bargaining mechanism, and 3) a resource management framework. The BS, by keeping statistics on the offered traffic load, updates the lowest acceptable price per unit of bandwidth that an MT has to pay for an advance reservation, $Cost_{BS}$, as discussed in Section III. The MT agent uses the output of the PPA[1], in order to determine the most likely BS for the next handover. After valuating the resource (i.e., calculate $Benefit_{MT,i}$ for every active session i), the MT commences the bargain with the target BS.

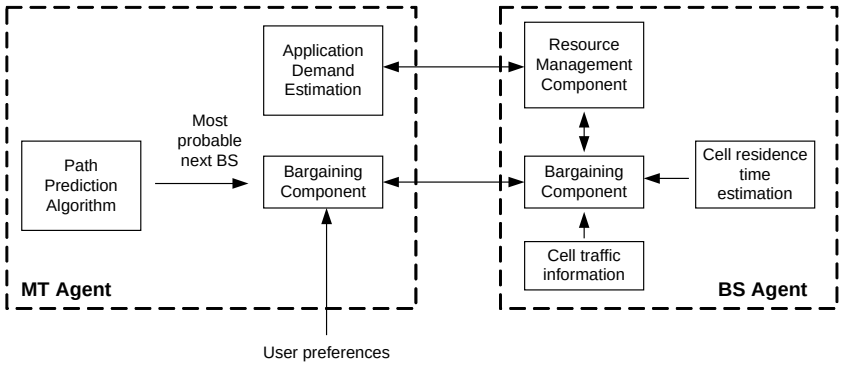

Figure 1. MT Agent and BS Agent overview

A. The Bargaining Mechanism

The bargaining mechanism introduced in this paper is based on the model studied in [12], i.e., an infinite horizon bargaining model with two-sided uncertainty [13]. The BS is the player that possesses the good (bandwidth), while the MT is the buyer. The players bargain over the price of a bandwidth unit (e.g., 1 Kb/s). Specifically, the MT starts a bargain with the target BS (i.e., the BS that is most probable to be visited in the future), after spending an arbitrary time interval, t_0, in the current cell[2]. The BS starts by making expensive offers, but gradually makes his offers more attractive, until the MT accepts to buy, at time t_b. Offers are issued every τ seconds. Fig. 2 illustrates the discussed process. At time t_h, the MT executes the anticipated handover. Note that $t_h > t_b$; in the opposite case, the bargain would be forced to terminate, and the MT would make the handover with no reserved resources in the

[1] The analysis and evaluation of PPAs is outside the scope of this work. We assume the use of the PPA reported in [10]. Further discussions on the path prediction issue can be found in [8], [9].
[2] The MT can only negotiate with one BS in the neighborhood.

new BS. This can be observed in the *no-gap case* [13], where the trade may not provide gains to the involved players[3], or if the handover takes place sooner than expected. Notice that, as the importance between different application types may vary for a given user, the MT makes a separate bargain for every different application type (e.g., FTP, voice over IP, streaming video).

In the considered model, both players incur costs when delaying the bargaining conclusion. Such costs express the stress that time places on the players, as the MT may fear that the handover will occur before the agreement, and the BS may worry that the MT will possibly prefer another, more inexpensive, reachable BS that belongs to another network [15]. This situation is modeled by discounting the payoffs of the players in the subsequent rounds according to the factor δ_b and δ_s ($0 < \delta_b, \delta_s < 1$) for the buyer (MT) and the seller (BS), respectively.

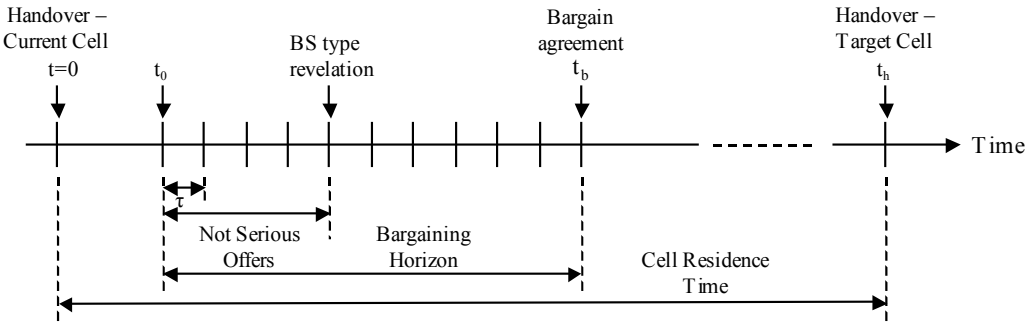

Figure 2. Bargaining Process

In the sequential equilibrium of the discussed bargaining game, information (i.e., player valuation) is gradually revealed over time and the rate of revelation depends on the costs of delay. Bargainers expecting larger gains (who therefore are more impatient) reach agreement before those that expect smaller gains [12].

B. Resource Management Scheme

The successful completion of a bargain implies that the MT and the BS have mutually agreed upon the bandwidth unit price. However, the exact amount of resources that the BS will have to reserve is not a result of the bargain. The MT is free to update the amount of resources that wants to be reserved, before the time of the handover. Specifically, the MT checks continuously, for every application type (for which the bargain has successfully terminated), if the corresponding reservation at the future BS can cover the needs of the active sessions. The *resource management component* (RMC), which resides on the BS (see Fig. 1), gathers all such MT reservation requests, and allocates them resources so as to limit the application session discontinuation probability. Moreover, it tries to maximize the monetary benefit of the BS, subject to the adopted pricing policy, which can be summarized as follows:

- The tariff charged for the basic connection service, for S bandwidth units, and t time units, is $c(t,S) = p \cdot S \cdot t$, where p ($/(Kb/s)·s) is the standard price charged by the network per bandwidth unit.
- The tariff charged for reservation request i, for s_i bandwidth units, is $c_i = p_i \cdot s_i$, where p_i ($/(Kb/s)) is the unit price (accepted by both parties after the bargain). Note that the network is paid only if the requested resources are successfully reserved, at the time of the handover.

[3] The no-gap case refers to the situation where the valuation of the buyer may be lower than the valuation of the seller, in which case an agreement cannot be reached.

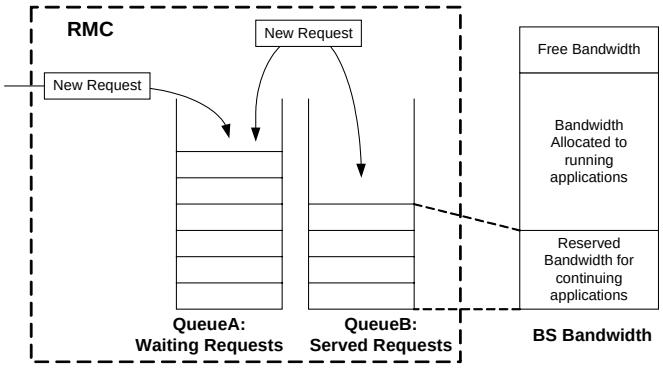

Figure 3. *QueueA* stores new MT requests. *QueueB* keeps served requests

When a MT issues a new reservation request, this request is inserted into *QueueA* (see Fig. 3). When an amount of bandwidth is released (e.g., a session terminates, or a MT makes a handover to an adjacent cell), the most "important" requests of *QueueA* are inserted into *QueueB*, and considered served. The "importance" of a request is determined by two factors: 1) the mutually agreed unit price p_i, for the particular session request i, and 2) the "age" of the request, i.e., the time since the MT entered the current cell (t_i), since the older a requests is the sooner the handover is anticipated to take place. Therefore, we define (importance) metric Z_A, by which requests in *QueueA* are sorted, as follows:

$$Z_A = p_i \cdot t_i$$

Note that the transfer of a request to *QueueB* does not imply that it will be kept there forever; it may be evicted (and inserted back into *QueueA*) at the occurrence of specific events, e.g., initiation of a new application, or handover of another MT, as will be discussed below.

1) New Application Arrival Event

Consider the situation illustrated in Fig. 4. At time t_n, MT_y (found at cell l) wants to start a new session, requiring S bandwidth units. At the same time, MT_x, which has successfully made a reservation to cell l, is anticipated to arrive from cell k. According to the assumed pricing policy, if the BS accepts the new session, it will charge MT_y with $p \cdot S$ \$ per unit time. On the other hand, when MT_x arrives, the BS will immediately receive $p_x \cdot s_x$ \$.

Given an estimation of the *cell residence time* (CRT)[4], and assuming that the new application will not be of very short duration, the network profit, G, for the time interval $[t_n, t_h]$, can be calculated as follows:

$$G(t_n, t_h) = \begin{cases} p_x \cdot s_x, & \text{if new appl. is blocked} \\ p \cdot S \cdot (t_h - t_n), & \text{otherwise} \end{cases} \qquad (5)$$

[4] The CRT can be estimated in many ways. One of them is to use a low pass filter of the form: $CRT = a \cdot CRT_{new} + (1-a) \cdot CRT$, where $0 \le a \le 1$

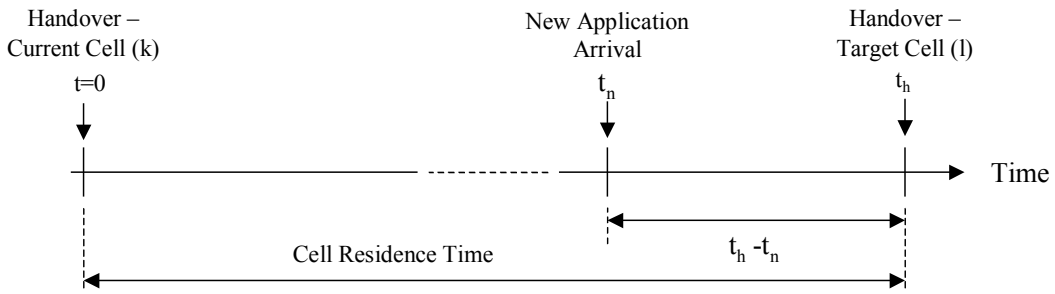

Figure 4. New session arrival

```
Event: A new application demands S units of bandwidth
/* It is assumed that there is not sufficient free bandwidth */

gatheredResources = 0
gatheredValue = 0
criticalElement = -1
for i = QueueB.length:(-1):1
   request = QueueB.getElem(i)
   gatheredResources = gatheredResources + request.wantedResources
   reqValue =
(request.wantedResources*request.price)/timeLeftForHandover()
   gatheredValue = gatheredValue + reqValue
   if gatheredResources >= S then
      if gatheredValue < p*S then
         criticalElement = i
         break
      endif
   endif
endfor

if criticalElement <> -1 then
   for i = criticalElement :QueueB.length
      request = QueueB.getElem(i)
      QueueA.insertElem(request)
      QueueB.remove(i)
   endfor
   return OK
else
   return FAILED
endif
```

Listing 1. New application arrival event

The BS follows the action that yields the higher payoff. If the BS decides to grant access to the new session, it has to expunge requests from *QueueB*, until sufficient resources are available. However, it is very important to determine which particular requests to evict. The cumulative value of these requests must be minimal. Moreover, the total amount of released resources must be greater than or equal to the resources required by the new session. Lastly, the monetary value of the set must be less than the value of the new session. This is a typical constrained optimization problem, which is formulated as follows:

$$\min_{y} \sum_{j=1}^{n} p_j s_j y_j \text{ , s.t. } \sum_{j=1}^{n} s_j y_j \geq S, \sum_{j=1}^{n} \frac{p_j s_j}{t_j} y_j < p \cdot S$$
$$y_j \in \{0,1\}, \; j \in N = \{1,2,...n\}$$

where N is the set of elements in *QueueB*, t_j is the expected remaining CRT, while y_j are binary variables indicating whether the solution contains the corresponding element, j, or not. S denotes the amount of resources that has to be released. This is an integer linear programming problem, which resembles the 0-1 knapsack problem. We adopted an approximate algorithm (linear programming relaxation) with $O(n)$ complexity (see Listing 1). We assume that elements in *QueueB* are sorted by the price to needed capacity ratio

$$Z_B = p_i / s_i.$$

2) Handover Event

Here, we discuss the actions taken, in the event of a handover, if the RMC has not yet served all the reservation requests of the handed over MT – i.e., when all, or a portion of the requests remain in *QueueA*. In such a case, it is known to the RMC that requests currently stored in *QueueB* (i.e., served requests) are inactive, as the corresponding MTs have not arrived yet. It is also known that granting to the MT the requested resources by expunging some requests from *QueueB* will result in the agreed payment. The expunged requests will, possibly, have the opportunity to be served once again before the arrival of their MT. The RMC selects the set of requests to be evicted by solving the following optimization problem, similarly to the new session arrival event (see previous paragraph). The solution of the problem is achieved by means of the algorithm presented in Listing 2.

```
Event: A MT was handed over and a request demanding S units of bandwidth
is expunged from QueueB
/* It is assumed that the RMC has not included the request in QueueB
*/

gatheredResources = 0
criticalElement = -1
for i = QueueB.length:(-1):1
   request = QueueB.getElem(i)
   gatheredResources = gatheredResources +
request.wantedResources
   if gatheredResources >= S then
      criticalElement = i
      break
   endif
endfor

if criticalElement <> -1 then
   for i = criticalElement :QueueB.length
      request = QueueB.getElem(i)
      QueueA.insertElem(request)
      QueueB.remove(i)
   endfor
   return OK
else
   return FAILED
endif
```

Listing 2. Handover event

$$\min_{\boldsymbol{y}} \sum_{j=1}^{n} p_j s_j y_j, \text{s.t.} \sum_{j=1}^{n} s_j y_j \geq S$$
$$y_j \in \{0,1\},\ j \in N = \{1,2,...n\}$$

Other events, e.g., handover to a wrongly predicted BS (due to a PPA failure), can be handled similarly, and, thus, are not described, in this paper, for the sake of brevity.

V. SIMULATIONS

A. Simulation Setup

In order to assess the performance of the proposed scheme, we performed a series of simulations. Specifically, we assumed a floor layout covered by 10 BSs, as shown in Fig. 5. The capacity of each BS (C_{Total}) was 20,000 units (Kb/s). A line connecting two BSs denotes that the corresponding cells are neighboring. In the simulations, a population of 2,000 MTs was roaming stochastically within the network, which resulted in a heavily loaded and congested network. MT cell residence times follow the generalized Gamma distribution, as in [14]. We also assumed that the PPA manages to correctly identify the next cell with probability 0.8 [10].

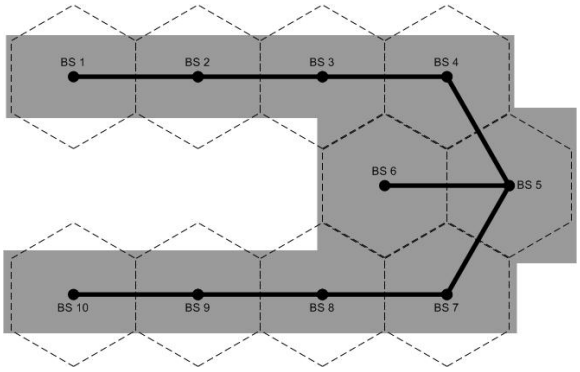

Figure 5. The floor plan of the simulation area

With respect to the user traffic characteristics, we assume a simple traffic model with four kinds of applications, namely FTP, HTTP, VoIP, and video. The characteristics of these applications are summarized in Table I. Furthermore, we assume that the maximum number of application sessions running concurrently at a particular MT are limited. Specifically, a user can run in parallel up to three FTP, three HTTP, one VoIP, and one video session. The duration of an application session was modeled as a random variable that follows the exponential distribution with mean value as indicated in Table I. Application session arrivals follow the Poisson distribution. With regards to the bargaining parameters, we assumed that $Benefit_{MT}$ and $Cost_{BS}$ are uniformly distributed random variables in the (0.5,1.5) and (0,0.5) interval, respectively, and that $\delta_s = \delta_b = 0.75$.

TABLE I. APPLICATION CHARACTERISTICS

Application Type	Requested Bandwidth (Kb/s)	Mean Session Duration (s)	Mean Session Interarrival Time (s)
FTP	100	300	1800
HTTP	50	3	60
VoIP	64	120	1200
Video	512	300	1800

B. Simulation Results

For the assessment of the proposed scheme, we considered the probability of blocking of a new or a continuing application session, P_n and P_h respectively. P_n is defined as the number of session initiation failures over the number of session initiation attempts, while P_h is the number of discontinued sessions as a result of handovers over the number of application sessions that were subject to handover.

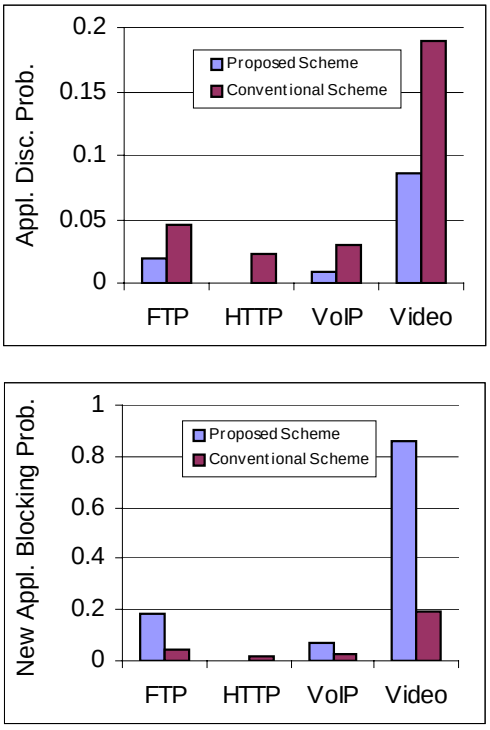

Figure 6. (a) Application discontinuation probability (b) New application blocking Probability

For standard price p = 0.1 \$/((Kb/s)·s), Fig. 6a and Fig. 6b depict our simulation results. Note that the proposed scheme improves the application discontinuation probability, P_h, for all types of applications, whereas, the new application blocking probability, P_n, is higher compared to the conventional scheme (i.e., without proactive reservation). This is anticipated, as a portion of the network resources is used for the handed over sessions, thus, reducing the amount of available resources for new sessions.

We now study the effect of the standard bandwidth unit price per second p on P_n and P_h. The value of p is taken into account by the admission control algorithm described in Section IV.B.2. When p is large, new sessions take priority over the handed over sessions, which means that it is difficult for a handed over session to obtain the required resources. This implies that in the case of failure, the session will try to be admitted, considered as a new session (i.e., it will switch role). This has as a result that it will have to compete with other sessions anticipated to be handed over (over which it has priority). Hence, an increase in p may have an indirect positive impact on P_h. This is depicted in Fig. 7a. Observe that by increasing the value of p there is a decrease of the application discontinuation probability. However, by further increasing the value of p, the application discontinuation probability increases.

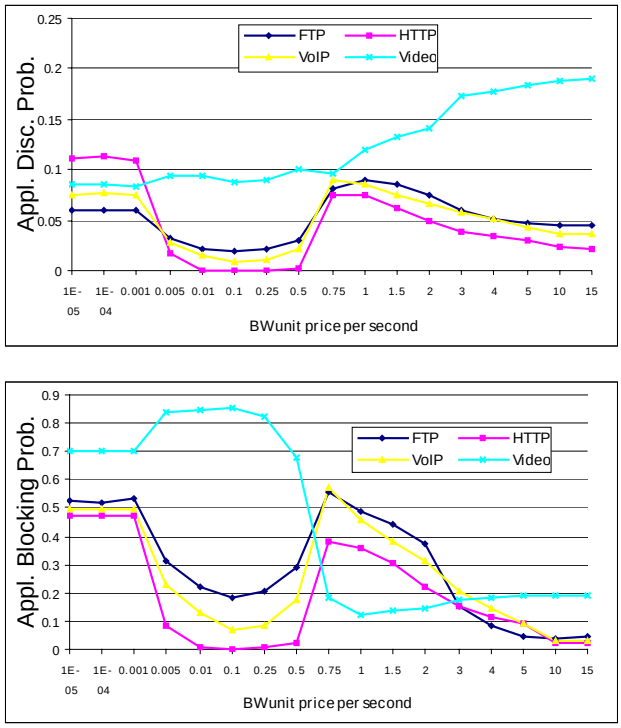

Figure 7. Effect of the unit price on (a) the application discontinuation probability (b) the new application blocking probability

From Fig. 7a and Fig. 7b, we may make the following three observations: 1) the effect of p on probabilities P_h and P_n is not monotonic, 2) there is a range of values for which both P_h and P_n reach a minimum for the FTP, VoIP, and HTTP applications, and 3) as p increases, P_h and P_n converge to the corresponding blocking probabilities of the conventional scheme (see Fig. 6a and Fig. 6b). Hence, there is a tradeoff regarding the value of p. Moreover, we observe that the proposed mechanism favors low-rate applications. According to the discussion in Section IV.B.1 and Section IV.B.2, this is reasonable, as, in the case of congestion, a low-rate request is more likely to find the necessary resources than a high-rate request.

VI. CONCLUSIONS

In this paper, we have proposed a proactive resource management scheme for the reservation of network resources, prior to the handover of the MT. A sequential bargaining procedure, modeled as a two-person non-cooperative game, between the MT and the target BS, concludes with a mutually agreed bandwidth unit price. Following this procedure, the MT can request advance resource

reservations from the BS. The BS accumulates such reservation requests from different MTs, and on the occurrence of specific events (e.g., session initiation, handover, session termination), decides which requests will obtain the requested resources.

Simulation results show that the proposed mechanism is capable of reducing the application discontinuation probability compared to the conventional scheme. However, as anticipated, the new application session blocking probability is increased. Moreover, it was shown that the price charged to active sessions affected the performance of the system. For certain values of this price, the session discontinuation probability was minimized. It was also observed that the mechanism favors application sessions with relatively low bandwidth requirements. In the future, we plan to further study the behavior of the proposed scheme, by measuring the effects of uncertainty that exists in such random and unstable environments (e.g., sensitivity analysis on the next BS prediction probability, or the uncertainty regarding the time of the handover). Moreover, we would like to focus our study on how such mechanisms could be applied to other types of network resources besides bandwidth, and provide a unified framework for the proactive management of resources in such mobile and highly volatile environments.

ACKNOWLEDGEMENTS

The first author would like to thank the Alexander S. Onassis Public Benefit Foundation for its financial support.

REFERENCES

[1] S. Hadjiefthymiades, S. Papayiannis and L. Merakos, "Using Path Prediction to Improve TCP Performance in Wireless/Mobile Communications", IEEE Communications Magazine, Vol.40 No.8, 2002

[2] M. Sidi and D. Starobinski, "New Call Blocking versus Handoff Blocking in Cellular Networks", Kluwer Wireless Networks, Vol. 3, Issue 1, 1997.

[3] Y. Zhang and D. Liu, "An Adaptive Algorithm for Call Admission Control in Wireless Networks", Proc. IEEE Global Communications Conference, San Antonio, TX, pp.3628-3632, 2001

[4] O. Yu and V. Leung, "Adaptive Resource Allocation for Prioritized Call Admission over an ATM-based Wireless PCN", IEEE JSAC, Vol. 15, no. 7, pp. 1208–25, 1997

[5] C. Oliveira, J. Kim and T. Suda, "An Adaptive Bandwidth Reservation Scheme for High-speed Multimedia Wireless Networks", IEEE JSAC, vol. 16, no. 6, pp. 858–74, 1998.

[6] D. Levine, I. Akyildiz and M. Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept", IEEE/ACM Trans. Networking, vol. 5, no.1, pp. 1–12, 1997.

[7] R. Guerin, "Queuing-blocking system with two arrival streams and guard channels," IEEE Trans. Commun., vol. 36, pp. 153−63, 1988.

[8] G. Liu and G. Maguire Jr, "A Class of Mobile Motion Prediction Algorithms for Wireless Mobile Computing and Communications", MONET, Vol.1, pp.113-121, 1996.

[9] A. Aljadhai and T. Znati, "Predictive Mobility Support for QoS Provisioning in Mobile Wireless Environments", IEEE JSAC, Vol.19, No.10, 2001.

[10] M. Kyriakakos, S. Hadjiefthymiades, N. Fragkiadakis and L. Merakos, "Enhanced Path Prediction for Network Resource Management in Wireless LANs", in IEEE Wireless Communications Magazine, Vol.10 No.6, 2003.

[11] M. Rabinovich and O. Spatscheck, "Web Caching and Replication", Addison Wesley, 2001

[12] P. Cramton, "Bargaining with Incomplete Information: An Infinite-Horizon Model with Two-Sided Uncertainty", Review of Economic Studies, 51, pp. 579–593, 1984.

[13] D. Fudenberg and J. Tirole, "Game Theory", MIT Press, Cambridge (MA), 1991.

[14] M. Zonoozi and P. Dassanayake, "User Mobility Modeling and Characterization of Mobility Patterns", IEEE JSAC, Vol.15, No.7, 1997.

[15] H. Lin, M. Chatterjee, Sajal K. Das and K. Basu, "ARC: an integrated admission and rate control framework for CDMA data networks based on non-cooperative games", Proc. MOBICOM 2003, pp. 326-338