A Representative Based Architecture for Handling Mobility in Connection Oriented Radio Networks

Subir Kumar Biswas[†*]
[†]Computer Laboratory
University of Cambridge, UK.

Andy Hopper[‡†]
[‡]Olivetti Research Limited
Cambridge, UK.

Abstract

This paper presents a connection management architecture for handling mobility in a radio network environment. A new concept of per roaming entity software agent, known as mobile representative, *is used for insulating fixed network entities from the effects of user mobility. It is shown that mobile representatives can be used for distributing mobility management load within fixed backbone network, depending on the instantaneous physical locations of mobile terminals. The paper describes a representative management architecture followed by a novel connection caching strategy which is used for achieving low-latency and scalable connection handovers.*

1 Introduction

Multimedia networking is an emerging technological trend of the 1990s and there exists an increasing demand for supporting multimedia traffic in wide and local area wireless network environment [1]. By now it has began to be understood that in order to guarantee the strict performance requirements of multimedia traffic, connection oriented approaches [2] are more efficient compared to conventional datagram based networking. However, designing connection oriented wireless networks requires to overcome some serious restrictions imposed by mobility of the roaming network users. In addition to a complex routing policy, during relocation of the end users of a connection the end-to-end connectivity needs to be maintained by performing a signalling operation, known as *handover* [2, 3].

This paper presents a mobility management architecture for supporting connection oriented multimedia streams in a radio network environment. In addition to tackling the mentioned problems, the present architecture attempts to achieve the following design goals. First, it has been decided that user mobility should be isolated from fixed network entities as much as possible so that the existing fixed network protocols and applications can co-exist with its wireless counterparts without any major modifications. Second, no assumptions should be made about backbone network topology so that wireless network services can be smoothly deployed around the existing high speed wired networks. Finally, mobility processing load should be uniformly distributed over the network and, if possible, should be dynamically re-distributed following the movement of roaming users. A concept of per mobile software entity, known as *mobile representative*, is introduced in order to achieve the mentioned design goals. A representative acts as an intermediate agent between its client mobile unit and other remote peer entities. In addition to isolating mobility from non-mobile network entities, representatives enable a network to distribute mobility processing load by dynamically migrating themselves, following their client mobile users. A connection handover is dealt with by a novel caching scheme where a bank of connection *segments* are dynamically created and destroyed, following the mobility pattern of an end user. During a handover, instead of creating a new end-to-end connection, those segments are appropriately switched in and out in order to maintain the desired connectivity.

The rest of the paper is organized as follows. The next section outlines network model and a brief description of the major wireless services, necessary for supporting mobility. The concept of mobile representative is introduced in Section 3. Sections 4 and 5 present brief outlines of the proposed location and representative management procedures. Mobile connection related issues, including handover, are discussed in Sections 6 and 7. Finally, Section 8 concludes the paper with a summary of this work.

2 Network Model and Services

The present wireless mobile network is composed of a wired high speed *backbone* and a wireless (radio) interconnect. The wireless part is partitioned into few *transmission cells*, with a control point, known as a *base station* (BS), in each of the cells. Base stations, which are attached to the wired backbone, provide communication gateways between the radio and backbone networks. Any relocatable entity, known as a *mobile unit* (MU), communicates with any other network entity through the base stations and the backbone network.

The network consists of large number of very small transmission cells, often referred to as pico-cells. The spatial as well as temporal profiles of the pico-cells within a building are highly irregular and unpredicatable because of complicated terrain and attenuation characteristics [4] of the building materials. To partially mitigate this unreliability, caused by unpredictable pico-cell characteristics, a cell overlapping strategy is introduced where an MU is likely to be seen simultaneously by multiple base stations. *Pico-cell overlapping* is also used as a key component while designing low-latency handovers, described in Section 7. In a single carrier frequency scenario [5], however, the cost paid for this is a reduction of overall radio frequency reuse efficiency.

Along with the non-mobile network entities, each MU is assigned with a globally unique network layer address which is independent of the instantaneous geographical location of the MU. This static addressing policy enables the conventional network layer addressing and naming schemes (e.g. IP, ATM) to be applicable in the current network model. *Names* will be used in the rest of this paper, for identifying the mobile units.

A layered representation of the major wireless network functionalities is pictured in Figure 1. Mobility handling services are split into two independent func-

*Present contact address: NEC USA Inc., 4 Independence Way, Princeton, NJ 08540. E-mail: skb@ccrl.nj.nec.com

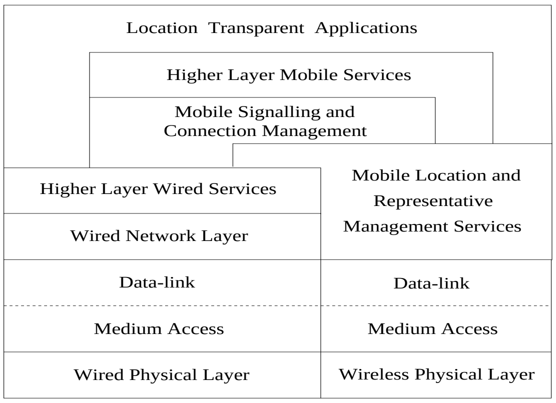

Figure 1: Layered representation of network services

tional modules, namely, a connection manager and a location/representative manager. Because of the connection oriented nature of the network, most of the location service functions are required to be kept isolated from data and connection services. Note that the present mobility management architecture does not assume any specific medium access or data-link layer protocol and thus it could be implemented on top of a wide range of lower layer protocols (as in [5]) which can accommodate the mentioned *cell overlapping* feature.

3 Mobile Representative

While setting up a mobile connection[1], the source and intermediate network entities need to know about the location of the end mobile unit in order to route the connection. Any post-connection mobility information also needs to be disseminated, so that the connection and reliable data transfer through it can be maintained. This requires all involved network entities to be aware of changing location information regarding the relevant MU. The present approach is to provide a control and data interface between a mobile entity and the rest of the network in order to avoid global distribution of mobile location information. Such a software interface, known as *mobile representative* (MR), is employed for partially isolating the effects of user mobility from the rest of the network. Any application, interested in talking to an MU, contacts its representative and the rest is handled by the corresponding MR. This requires a representative to keep up-to-date location information about its client MU. Figure 2 illustrates the way a connection to an MU always passes through its representative. When the mobile unit relocates itself from BS_1 to BS_2, only parts of the connection from the mobile unit to its representative are affected. The rest of it remains unaltered, as long as the representative process remains stationary. Since a representative process is provided for each MU, by choosing appropriate locations of the MR processes the mobility management functionalities can easily be distributed according to the physical locations of the roaming entities.

Placement and mobility of the representatives are two crucial design decisions. One extreme approach is to nail down a representative process on a fixed network entity which is expected to remain close to its client MU, most of the time. Although it makes MU mobility completely transparent to the network, the scheme suffers from lack

[1] Any connection, with an MU at one end, is termed as a mobile connection.

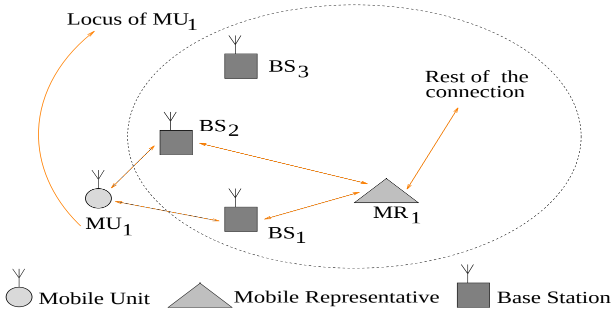

Figure 2: Use of *representative* for mobility insulation

of reliability and non-optimal utilization of network resources, caused by excessive location update traffic. At other extreme, an MR closely follows its client MU and usually resides within its geographically nearest possible stationary entity. While yielding a better location update performance, this approach fails to ensure the *mobility hiding* feature of the representatives. An intermediate measure is taken in the current implementation [6], where an MR is arranged to be much less mobile than its client MU and a representative migration takes place only when its client MU relocates to a new geographical or administrative domain. Note that in addition to mobility management load balancing, representative migration allows the design not to assume any specific fixed network topology. This is done by introducing a dynamic representative host selection scheme, based on network topology and user mobility.

4 Location Service

A distributed location server is designed where each network domain, known as a *cluster*, maintains a location management entity, called as a *Cluster Location Server* (CLS). Only the connection management related operations of the location server are highlighted in this section. For details about the location server design and its implementation, the reader is referred to [6].

In addition to maintaining distributed location information, the location server also aids the MRs to keep up-to-date location information about their respective client mobile units. A CLS needs to supply the following information to a location query client. <u>MU location:</u> Represented by the addresses of the base stations, currently seeing the mobile unit, and their cluster identifiers. <u>MR location:</u> Indicated by the address of the host-entity on which the relevant representative process is running currently.

A client sends its location queries to the CLS of the local cluster. Upon reception of a query, the CLS first attempts to resolve it by searching the local databases. If the local search fails, it consults the relevant remote CLSs to acquire the necessary information for the querying client. More about the distributed query searching and location update strategies are presented in [6].

5 Representative Management

Creation and migration are identified as the primary management operations associated with a representative process.

<u>Creation:</u> Both during initial creation and migration, new host for a representative is chosen by the appropriate cluster location servers. The cluster, in which a represen-

tative MR_j would reside, is selected by the home CLS[2] of MU_j and the particular representative host-entity is selected by the CLS of chosen cluster. When MU_j is simultaneously seen by multiple clusters, home CLS of MU_j uses any one or a combination of the following heuristics for choosing an appropriate cluster in order to create or relocate MR_j. *i*)Choose the cluster where MU_j was seen most recently. This is useful when not much information is known about MU_j's *mobility profile*[3]. *ii*)Select the cluster where maximum number of base stations are currently being seen by MU_j. This strategy is useful for minimizing average distance between MR_j and all the base stations, seen by MU_j.

Once a cluster is selected, CLS of the chosen cluster uses the following criteria to pickup a host-entity for MR_j process. First, it is necessary for a host to be located in a fairly central position with respect to the cluster base stations. A variation of this policy is to give more priority to the BSs which are currently seeing the client mobile unit MU_j. In addition, the chosen machine should have sufficient computing and communication resources left for hosting a representative process.

Migration: As explained in Section 3, the semi-mobile MR processes migrate only when their client MUs move across cluster boundaries. On relocation of MU_j to a new cluster, its home CLS_h may decide to change the host-entity of representative MR_j and, if decided, a set of heuristics are used by cluster location servers to choose a new host for MR_j. Since a mobile connection needs to pass through the involving MRs (see Section 3), an online representative migration would require the MRs to move with all the passing connections. While maintaining transparency, this requires additional complexity for performing *representative handovers* in addition to the standard *base station handovers* of mobile connections.

Instead, a *transient representative duplication* mechanism is introduced where a new representative MR_{new} is created soon after a migration decision is taken by the relevant home CLS. While all old connections continue to operate through the old representative MR_{old}, new connections are setup through MR_{new}. From this point onwards the entire network, including location servers, considers MR_{new} as the new representative process and MR_{old} becomes defunct except for the old connections. Once all the old connections are closed, MR_{old} terminates itself automatically.

6 Mobile Connections

Structure of a mobile connection, designed for supporting low-latency handovers, is pictured in Figure 3. A mobile connection can be seen as a set of wired and wireless *segments*, put together to achieve end-to-end connectivity. Although all wireless segments are single-hop, number of hops in a wired segment depends on relative physical locations of base stations and host-entities of the end mobile units' representative processes. Any combination of segments, providing end-to-end connectivity, constitutes a *path*. At any time, there exists only one *active path* which carries traffic through a connection. For example, MU_j-BS_p-MR_j-MR_k-BS_t-MU_k in Figure 3 indicates an *active path* for connection MU_j-MU_k.

Multiple paths are used for the handover control scheme, explained in Section 7. It is to be noted that the multiplicity of edge segments (MU-BS and BS-MR) and the paths in a mobile connection depends on the number of base stations, simultaneously seen by end MUs. During a handover, those edge segments are switched in and out of a connection by MUs and their representatives. These switching operations are coordinated according to the patterns of inter base MU relocation.

6.1 Connection Setup

While creating a new mobile connection, only the segments necessary to construct an *initial path* are setup and data transfer starts through it immediately. If necessary, other feasible segments of the connection are created afterwards. An example connection and segment setup procedure is pictured in Figure 3 and the corresponding signalling sequence is depicted in Figure 4.

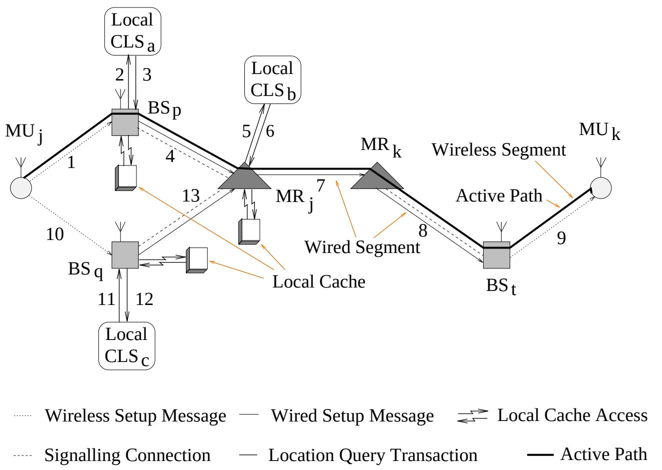

Figure 3: Connection segments and their setup management

Each connection setup cycle involves at least one location query to locate the relevant representative process. Two queries are needed when both the end entities are mobile. Messages '2', '3', '5' and '6' represent such location query transactions during the example setup procedure.

A combination of *source* and *hop-by-hop* approaches is used for routing a mobile connection. Source routing is implemented on a per segment basis whereas within a segment, a connection is dynamically routed in a hop-by-hop fashion. This requires base station BS_p, for instance, to put the network layer address of MR_j host in the connection setup message (labelled '4'), for short term source routing. Segment routing from BS_p to MR_j, however, is hop-by-hop and is dynamically performed by using any standard wired connection routing policy. Note that the one-hop wireless setup ('1' and '9') and reply messages do not need any explicit routing to be performed.

6.2 Segment Management

While data shipping continues through the *initial path*, additional segments are created and destroyed, following the end MU mobility patterns. The segment, connecting two mobile representatives, remains unaltered during the lifetime of a connection; only the ones, with an MU or a BS as one of the end entities, are created and terminated dynamically. An off-line segment setup procedure,

[2]This CLS belongs to the cluster where the mobile unit MU_j comes from.

[3]Mobility profile is specified [7] as temporal and spatial locality information about the movements of an MU.

triggered by MU_j's relocation into BS_q's pico-cell, is presented in Figure 4; corresponding messages are labelled from '10' through '13'. Other than when a mobile connection is terminated, a segment through base station BS_p is deleted when MU_j goes out of BS_P's pico-cell. Note that, all these segment management tasks are described on a per connection basis.

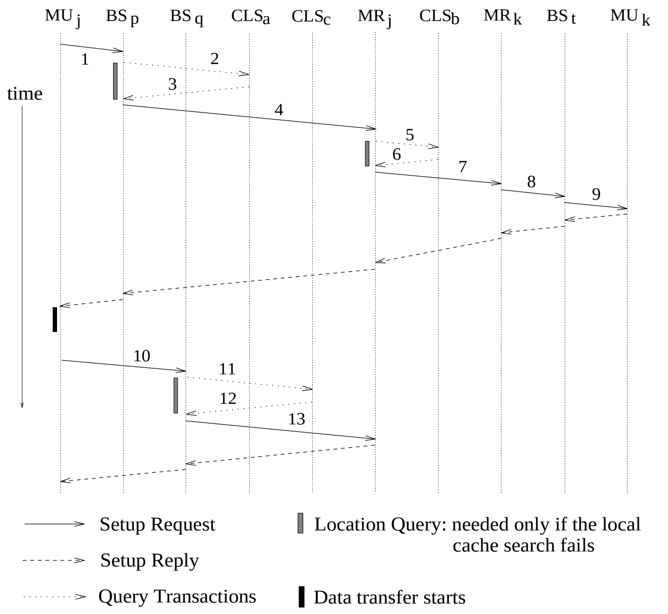

Figure 4: Signalling for connection and segment setup

A mobile connection may close-down either because of an explicit termination by an end entity or due to a fault developed during an MU relocation. In either case, all associated segments need to be cleared while closing the connection.

7 Handovers

In order to reduce handover latency, a segment caching scheme is introduced where necessary segments of a connection are created before an actual handover takes place. Once the handover is initiated, relevant segments are switched in and out of the connection to maintain an end-to-end *path* at any point of time. Although this network layer handover protocol attempts its best to prevent packet loss and out-of-ordering, currently no mechanism is provided for recovery from those error situations. Higher layer protocols will have to recover from these, depending on specific performance requirements of transported traffic. A carefully designed signalling sequence is used for reducing packet loss and out-of-ordering probability during handovers.

In order to expedite handover operations, signalling between base and representative is carried out through pre-established signalling connections. This requires a BS to create and maintain a handover signalling connection with an MR, if there exists at least one mobile connection segment between them.

Sudden failure of a radio link, constituting a part of the *active path* of a mobile connection, triggers a *without hint* handover [2]. On the other hand, when quality of the radio link between an MU and a BS degrades gradually, the MU may like to handover few or all of its connections to another BS with better radio link with it. These are referred to as *with hint* handovers.

7.1 With Hint

Segment switching and its corresponding signalling sequence during a *with hint* handover, initiated by MU_j, is pictured in Figure 5. It is assumed that current *active path* is passing through BS_p and segments j and l, involving BS_q, are already setup during off-line segment management procedure. Consistent improvement of radio link quality of segment j, compared to that of i, triggers MU_j to initiate the handover.

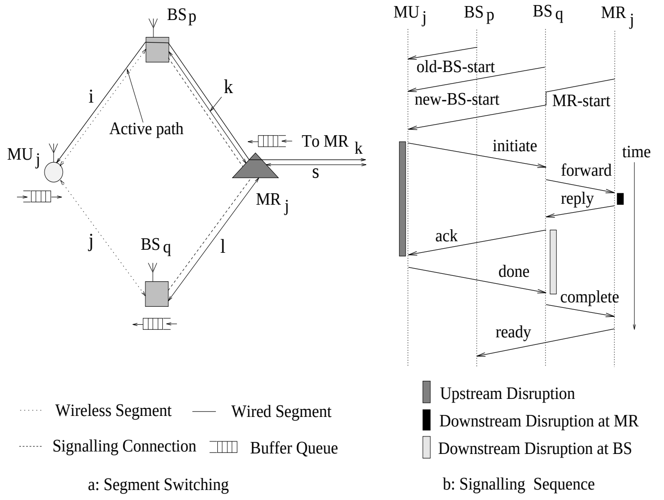

Figure 5: Signalling for a *with hint* handover

Soon after MU_j decides to initiate a handover, it stops sending traffic through segment i and puts the user data in a temporary buffer; however, it continues receiving through segment i. MU_j also sends an *initiate* message to the newly chosen base BS_q. When BS_q receives this message, it puts a buffer in between l and j. Until an explicit message (*done*) is received, data stored in this buffer is not sent through segment j. Once the buffer is set, BS_q sends a *forward* message to representative MR_j through a previously established handover signalling connection. Upon reception of *forward* message, MR_j breaks the existing path by disconnecting segments s and k and constructs a new one by connecting segments l and s. Soon after this segment switching is over, MR_j replies to BS_q with a *reply* signalling message. BS_q simply forwards this to MU_j as an *ack* message. BS_q is currently buffering all traffic received through segment l but forwarding the data (if there is any) received through j. On reception of the *ack* message, MU_j stops receiving through segment i; then it clears the buffered data by sending it through segment j and starts transmitting and receiving regularly through this segment. MU_j declares the completion of handover operation by sending a *done* message to BS_q. When BS_q receives this *done* message, it clears the queued traffic through segment j and starts regular data forwarding from segment l to j. BS_q also sends a *complete* message to representative MR_j, informing the latter about a successful completion of the handover procedure. Finally, MR_j sends a *ready* message to old base station BS_p so that the latter, depending on quality of its radio link with

MU_j, can either delete segments i and k or it can prepare for future handovers.

In situations where a *with hint* handover has to be initiated by a BS or an MR, messages *old-BS-start*, *new-BS-start* and *MR-start* are used by the respective entities in order to instruct an end mobile unit to perform a handover, as described above.

7.2 Without Hint

When a failure of radio link with MU_j is detected by BS_p, it sends a *failure inform* message to MR_j (see Figure 6). Additionally, the base station initiates deletion operations for segments i and k. On reception of that message, MR_j stops forwarding traffic through segment k and corresponding data is buffered within a local queue. Data received through segment k, however, continues to be forwarded through segment s. From this point onwards MR_j acts very much like as in *with hint* situation. The only difference is that when MR_j receives a *forward* message from BS_q, along with other usual segment switching activities, it flushes all buffered data and forwards it through segment l.

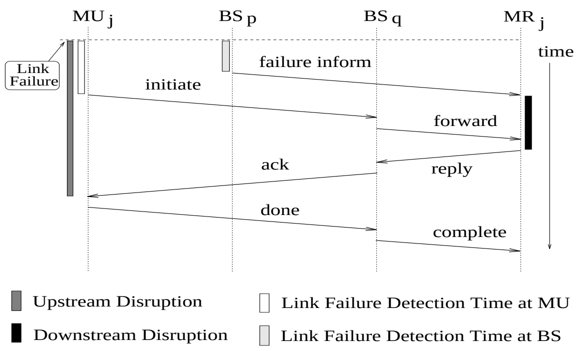

Figure 6: Signalling for a *without hint* handover

On the other side, upon detection of the link failure, MU_j starts handover by sending an *initiate* message to the newly chosen base station BS_q. Unlike for a *with hint* case, MU_j stops both transmission and reception through segment i and immediately starts reception over segment j. BS_q handles the *initiate*, *forward*, *reply* and *ack* messages in the same way as described in *with hint* situation. The only difference to be noted is that BS_q does not have to provide any buffering for MR_j-to-MU_j traffic. This is because of the immediate reception switching of MU_j from segment i to segment j, mentioned in the last step.

During a *without hint* handover, although the newly chosen BS does not contribute towards *downstream*[4] disruption, the MR counterpart of it is quite high because of link failure detection delay and *failure inform* message transit time. Upstream latency is also affected by link failure detection latency.

7.3 Performance Issues

Isolation of handover signalling at both the ends of a mobile connection makes this protocol scalable. Since representative migration policy (see Section 5) ensures that an MR is not terminated as long as at least one mobile connection is passing through it, handover signalling never gets affected by a non-stationary representative process.

Note that MU_j, in Figures 5 and 6, exchanges all handover signalling messages only with new base station BS_q. Better statistical radio link quality of segment j compared to that of segment i (possibly, this is why the handover was initiated), improves reliability of this scheme.

It is worth mentioning that the present protocol does not distinguish between *inter* and *intra* cluster handovers; the only difference, seen by applications, is increased latency because of larger geographical separations between the signalling entities in an inter-cluster situation. However, inter-cluster location and representative migration management operations are much more complicated compared to the intra-cluster ones. More details regarding a prototype implementation of the described architecture and its performance can be found in [6].

8 Concluding Remarks

This paper presented a design of a connection management architecture for supporting mobility in a radio network. A concept of mobile representative, for making mobility transparent to non-mobile network entities, has been introduced. It has been shown that these representatives can efficiently migrate over the fixed backbone network in order to distribute the mobility processing load. Detail management policies for representative creation and migration were presented. A segment caching scheme was also proposed for designing scalable and low-latency connection handovers. While the present scheme does not address Quality of Service (QoS) issues, there exists scope for tailoring the architecture for supporting stream QoS guarantee.

References

[1] T. Imielinski and B. R. Badrinath, "Mobile Wireless Computing: Solutions and Challenges in Data Management," *Technical report*, DataMan project, Rutgers University, April 1990.

[2] K. Keeton, B. A. Mah, S. Seshan, R. H. Katz and D. Ferrari, "Providing Connection-Oriented Network Services to Mobile Hosts," *Proceedings of the USENIX symposium on Mobile and Location-Independent Computing,* Cambridge, Massachusetts, August 1993, pp. 83-102.

[3] K. S. Meier-Hellstern, G. P. Pollini and D. J. Goodman, "Network Protocols for the Cellular Packet Switch," *To be appeared in the IEEE Transactions on Communications.*

[4] H. Hashemi, "The Indoor Radio Propagation Channel," *Proceedings of IEEE,* vol. 81, no. 7, July 1993, pp. 943-968.

[5] J. D. Porter and A. Hopper, "An ATM Based Protocol for Wireless LANs," *Olivetti Research Ltd. Technical Report,* no. 94-2, April 1994.

[6] S. K. Biswas, "Handling Realtime Traffic in Mobile Networks," *Ph.D. Thesis, Cambridge University Computer Laboratory,* August 1994.

[7] S. Rajagopalan "An Adaptive Location Management Scheme for Mobile Internetworking," *WINLAB Technical Report,* no. TR-93, December 1994.

[4]Represents traffic from backbone to MU. The reverse direction is referred to as *upstream*.