A Network Architecture for Mobile Computing*

Kevin Brown
Dept. of Computer Science
Univ. of South Carolina
Columbia, SC 29208
kbrown@cs.sc.edu

Suresh Singh
Dept. of Computer Science
Univ. of South Carolina
Columbia, SC 29208
singh@cs.sc.edu

September 7, 1995

Submitted for publication

Abstract

In this paper we report on an ongoing project to design and build the network and transport layers for mobile networking. The network architecture used is unique in that it separates the mobile network(s) from fixed networks and provides connectivity between the two via special gateways. These gateways provide QOS guarantees to mobile users for all their open connections. We provide summaries of all the protocols we are implementing (or have implemented) and discuss possible improvements.

```
Technical area : Wireless Networks and
                 Protocols

Corresponding author: Suresh Singh

Telephone: (803)777-2596
      Fax: (803)777-3767
```

1 Introduction

Mobile computing is an emerging new computing environment incorporating both wireless and high-speed networking technologies. Users equipped with *personal digital assistants* or PDAs

*Funding for this work was provided by the NSF under grant number NCR-9410357.

(palm-top computers) will have access to a wide variety of services that will be made available over national and international communication networks. Mobile users will be able to access their data and other services such as electronic mail, electronic news, yellow pages, map services, electronic banking and videotelephony services while on the move. To receive these services, mobile users will be connected to fixed networks via wireless networks (or mobile networks).

The **goal** of this paper is to present a comprehensive solution to the problem of wireless networking for mobile computing. We propose a mobile network architecture, network layer design and transport protocols that, we believe, will make it possible to offer all of the above services in an integrated manner. Such a system is currently being built at University of S. Carolina and, in this paper, we discuss the design in detail.

1.1 Challenges in Mobile Networking

Providing the type of services discussed above to mobile users requires high data rates on the wireless link and several authors (see for example Goodman[7, 8] Joseph[11]) have proposed an average data rate of between 1–2Mbps per mobile user. The third generation cellular system being developed in Europe (UMTS – Universal Mobile Communication System), for instance, also propose bandwidth in the same range (see DaSilva[4]). In order to support such high data rates, a *microcellular* network architecture has been proposed, see Goodman[7] and Duchamp[5]. Here, a geographical region such as a campus is divided into *microcells* with a diameter of the order of hundreds of meters. All mobile users within a microcell communicate with a central host machine within that cell who serves as a gateway to the wired networks; this machine is called a *mobile support station* (MSS).

What are some of the *networking issues* we need to address in order to provide the different types of service discussed above? Two broad issues that need to be considered are the following:

- Design of an efficient network architecture to support mobility and corresponding network layer protocols. The problems here include,
  - Tracking mobile users as they roam,
  - Routing messages and forwarding them to the current location of the mobile user,
  - Providing flow-control and buffering for open connections.
- Developing transport layer protocols that mesh easily with protocols that will be made available over high-speed networks. Some requirements here include,
  - Developing mobile network analogues of commonly used protocols such as TCP and UDP,
  - Developing protocols that support real-time applications such as voice and video or on-line data services,

– Maintaining quality of service guarantees for applications even in the presence of mobility.

Some of the problems mentioned above have been addressed by other researchers as we discuss below.

Routing in Mobile Networks: In mobile networks, since the hosts are mobile, routing is a problem. Ioanidis[9] proposes a solution called the **IPIP** ("IP-within-IP") protocol. Here every MH has a unique IP address called its 'home address'. To deliver a packet to a remote MH, the source MSS first broadcasts an ARP to all other MSS nodes to locate the MH. Eventually some MSS responds. The source MSS then encapsulates each packet from the source MH within another packet containing the IP address of the MSS in whose cell the destination MH is located. Upon receiving this packet the destination MSS extracts the original packet and attempts to deliver it to the destination MH. If the MH has moved away, the destination MSS attempts to locate it by broadcasting an ARP request. As discussed in Teraoka[15] this method is not easily scalable.

Teraoka[15] proposes a more flexible solution to the problem called – the Virtual Internet Protocol or **VIP**. Here every host has a *virtual network address* (VIP address) that is unchanging. In addition, hosts have associated *physical network addresses* (traditional IP addresses) that may change as the host moves around. At the transport layer, the target node is always specified by its VIP address only. The address resolution from the VIP address to the IP address takes place at either the network layer of the same machine or at a gateway. Both, the host machines and gateways, maintain a cache of VIP to IP mappings with associated time stamps. This information is in the form of a table and is called AMT (or *address mapping table*). Routing is achieved by referring to these AMT tables.

Transport Protocols for Mobile Networks: Since mobile hosts will expect the same services that are offered to fixed hosts, it is necessary to implement transport services in the mobile domain that are similar to those offered in the fixed networks. TCP is one such protocol. If we use TCP without any modification in mobile networks we have a serious problem of efficiency. This is because in TCP the sender begins retransmission of packets if they are not acknowledged within a short amount of time (hundreds of milliseconds). In a mobile environment, as a user moves between cells, there is a brief blackout period while the mobile unit performs a 'handshake' with the new MSS. These blackout periods may also be caused by physical obstacles in the environment that interfere with radio signals. These periods can be of the order of 1 second thus delaying the transmission of acknowledgements for packets received. This results in the TCP sender timing out and retransmitting the unacknowledged packets thus greatly reducing the efficiency of the connection. A solution to this problem is the I-TCP protocol (Indirect-TCP), implemented at Rutgers University as part of the DataMan project (see Barke[1]), that provides efficient reliable communication for the wireless environment. A benefit of the above implementation is that it allows mobile hosts to be connected over the Internet. We examine the I-TCP protocol in more detail in section 3.2.1 where we compare it against our own proposal.

1.2 Summary of Paper

We are currently building a mobile network from the ground up and this paper discusses our design and initial experience. Specifically, we propose a design for the network layer and transport layers for 3rd generation wireless systems and provide arguments in support of this design. The implementation of the protocol stack is been done under Unix (we use NetBSD) running on Pentium PCs.

- In section 2 we discuss our architecture for the mobile subnetwork that, we feel, best addresses the various issues raised in section 1.1.
- Our network layer design is presented in section 3.1. The sketch of our transport layer is presented in section 3.2. We also address management and control questions specific to the mobile environment (e.g., feedback of dynamic bandwidth changes to applications, etc.). Special protocols for notification applications (e.g., pager service) and continuous media are also incorporated.

2 Overview of our Proposed Architecture

The solutions discussed in section 1.1 use a microcellular architecture where the base station for each cell is a node on the internet. The solutions proposed for routing and TCP implementations thus assume that the underlying subnetwork is a datagram network. We believe that this assumption has only one justification – compatibility with existing technology – which, in our view, is insufficient. Some of the problems with this approach are,

- The base stations (or MSSs) are responsible for tracking mobile users and forwarding packets to their new locations. This adds to the *cost* and *complexity* of the base stations and since we expect cell sizes to be small (100m), in order to accommodate high-bandwidth applications, the total cost of a mobile network will be very high.
- As a user roams between cells, the bandwidth available in each cell may also vary. If the user has open connections, it will have to renegotiate QOS parameters frequently. This is clearly an undesirable situation.
- Cell latency times (staying time in a cell) are typically small (several seconds). This exacerbates both of the above problems.

Our architecture is discussed in detail in Singh[14]. We summarize the main points in this section. Our architecture may be viewed as a three-level hierarchy (see Figure 1). At the lowest level are the mobile hosts (MH) who communicate with MSS nodes in each cell. Several MSSs are controlled by a machine called the Supervisor Host (SH). The SH is connected to the wired network and it handles most of the routing and other protocol details for the mobile users. In addition it maintains connections for mobile users, handles flow-control and is responsible for maintaining the negotiated quality of service. A single SH may thus control all MSS nodes within a small building. Our architecture separates the mobile network from the high-speed wired network and provides connectivity between the two via supervisor hosts (SHs) who serve the function of a **gateway**.

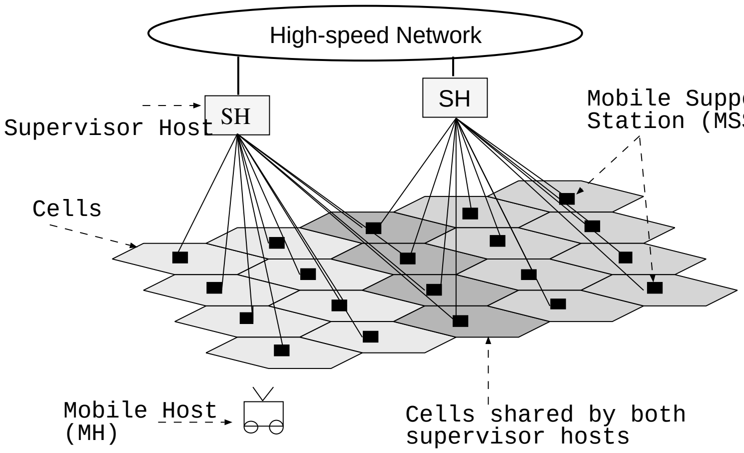

Figure 1: Proposed Architecture.

A mobile user may set up connections where the other end-point is either another mobile user or a fixed host (e.g., a service-provider) in the fixed network. In either case the connection is managed by the current SH of the mobile host(s) (see Figure 2). The reason for splitting the connection between the MH and the service-provider is to shield fixed nodes from the idiosyncrasies of the mobile environment. Thus, the service-provider sets up a connection with the SH assuming the SH is the other end-point of the connection. The SH sets up another connection to the MH. Thus for every **MH – service-provider** connection the QOS parameters are defined *separately* for the **MH – SH** part and for the **SH – service-provider** part of the connection. Connections between two MHs are broken in three (if the two MHs are controlled by different SHs) – **MH1 – SH1**, **SH1 – SH2** and **SH2 – MH2**. Note that the SH–SH part of any connection is established over the fixed network. The benefits of our architecture, thus, are:

- The MSSs are simple and cheap devices because they merely serve as a point of attachment for MHs.
- Since several MSSs are controlled by a single SH, the roaming MH remains within the domain of the same SH for long time periods[1]. This makes it easier to guarantee QOS

[1]e.g., a MH may roam frequently between rooms in an office building but remain for many hours within the building – each room is a cell and all cells in the building are controlled by one SH.

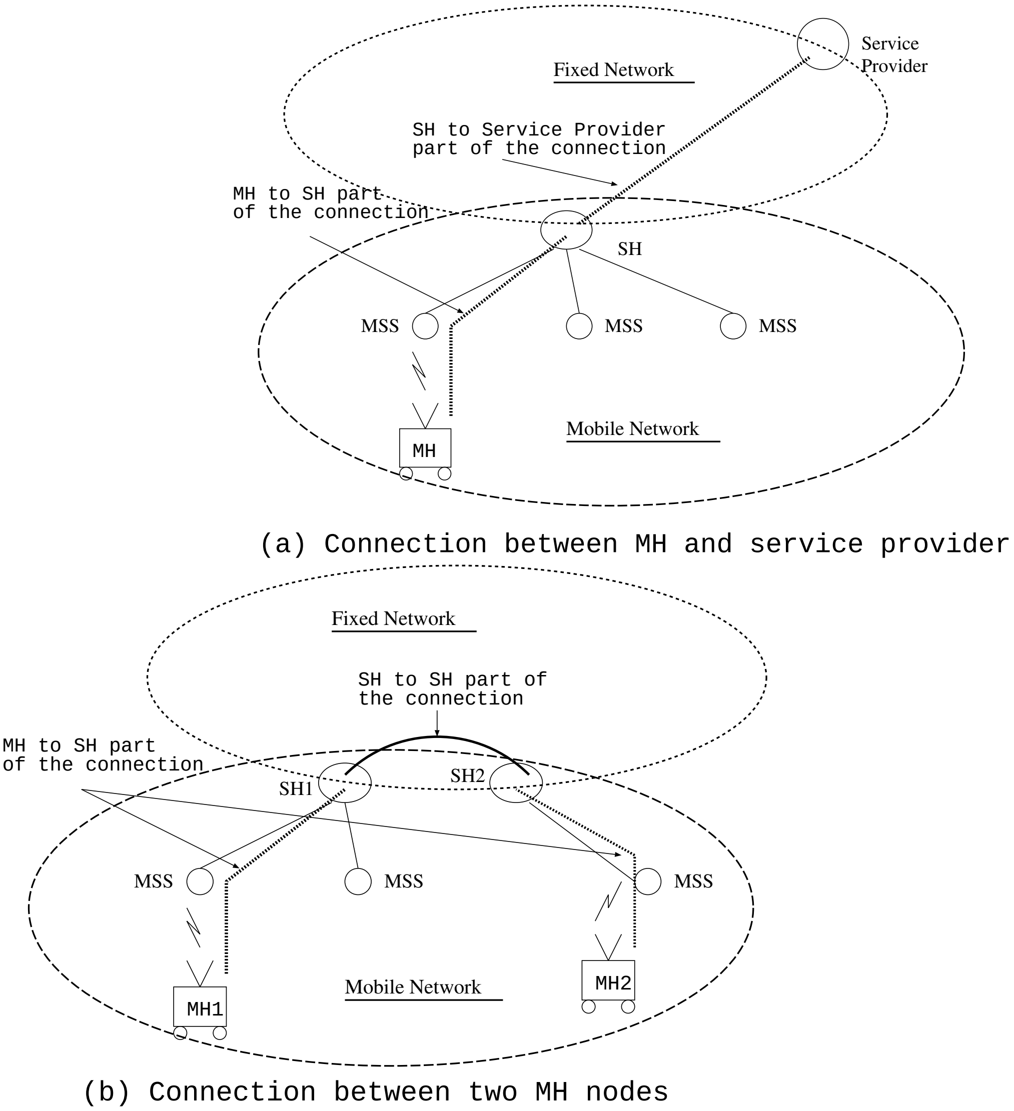

Figure 2: Connections for MHs are managed by SHs.

parameters for MH connections.

3 Network and Transport Layer Protocols

Our view of the mobile network in relation to fixed wired networks is illustrated in Figure 3. The mobile network is actually composed of many sub-networks each of which is connected to the fixed networks via a SH node. SH nodes communicate with one another over the fixed network. Each SH controls several MSS nodes. Physical communication between a SH node and its MSS nodes is accomplished either over a dedicated network consisting of dedicated wiring (perhaps several MSSs are connected via twisted-pair to a hub and several hubs are connected directly to the SH) or over the fixed network itself – i.e., the MSS nodes are regular hosts on the fixed network and can be addressed with IP addresses. The latter is a cheaper solution and provides a *migration* path to having dedicated mobile networks.

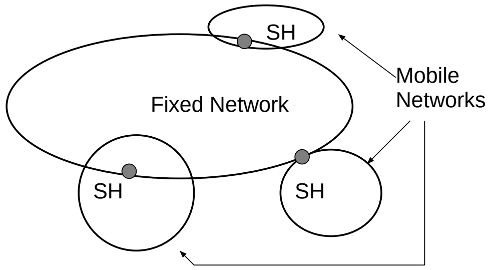

Figure 3: Relationship of mobile networks to fixed networks.

3.1 Network Layer Design

3.1.1 Routing and Tracking

Recall that all connections set up by a MH pass through its SH. For instance, a connection between MH M_1 and M_2 located within the same mobile network (i.e., same SH S) is set up as M_1–S–M_2. If the two MHs are in different mobile networks S_1 and S_2 the connection is M_1–S_1–S_2–M_2. The S_1–S_2 portion of the connection passes over the fixed network. A connection from M_1 to a fixed host F is set up as, M_1–S–F [2]. In all of these cases, notice that routing consists of two components – routing within a given mobile network from the SH to the MH and routing over the fixed network between SHs or between SHs and fixed hosts. Let us consider each of these two components separately.

In our design we implement *virtual circuits* at the network layer *in each mobile sub-network*. This means that the network layer will deliver all packets *in order* to the current MSS of the MH. Thus even if the MH moves between cells, so long as the MH remains within the same mobile sub-network, all packets will be delivered to its MSS in order. It is important to observe that the network layer *does not* guarantee delivery of packets *to the* MH; it only guarantees delivery to the MSS. This is because the wireless link is very unreliable and error recovery is best left to the transport layer which is responsible for implementing service guarantees. We discuss this point in greater detail in section 3.2. Moreover, if the MH moves into the domain of another SH, there are no guarantees made regarding the delivery of packets in transit.

To route within one mobile sub-network (i.e., within the domain of one SH), the network layer at the SH maintains a location table consisting of entries for each MH currently within its domain and the location of the MH (i.e., the identity of its MSS). This table is updated via control messages passed between the MSS and SH every time a MH moves. If the MSS nodes are

[2]The reason for routing all connections through a SH is that the SH can provide network level buffering and flow-control. If we set up a connection directly between two MHs, for instance, flow-control and retransmission of lost packets will require an exchange of control messages between the MHs consuming scarce wireless bandwidth.

connected via dedicated links to the SH, there is no need for them to have IP numbers. The SH simply transmits on the appropriate port. If the MSS nodes are nodes on the fixed network, then the SH needs to route messages to the MSS nodes on the fixed network. In our implementation all MH nodes have unique IP addresses (with some 'home' network as part of the address). Here the SH nodes route messages to the MH by using the IP *loose source routing option* (following the work of Johnson[10]). The destination address in the header is set to the IP address of the MSS and the MH IP address is contained as the first IP address in the option part of the header. The MSS examines the datagram and delivers it to the MH (if it is present in its cell). If it has moved away, the MSS discards the datagram.

To implement reliable delivery (in the sense discussed earlier), every datagram is given a sequence number. A MSS sends an ACK for each datagram transmitted to the MH (note that the datagram may not be received by the MH because of fading or other interference). Until a datagram is ACKed, it is buffered at the network layer of the SH. If a MH moves away from its current MSS to a new MSS, the old MSS discards all messages but simultaneously informs the SH of the sequence numbers of these discarded messages. These messages are then retransmitted to the new location of the MH. A detailed protocol is presented in Gahi[6] (though our current implementation contains many changes).

For routing over the fixed network (e.g., between SHs or between an SH and a service provider), the existing routing protocol provided over the fixed network is used. Thus the network layer shown in Figure 4 is local to the mobile sub-networks only. The network layer shown in Figure 4 consists of two sub-layers – a tracking and VC maintenance sub-layer sitting on top of IP. The tracking sub-layer is responsible for maintaining location information for each MH currently in that mobile sub-network. VC maintenance refers to the task of guaranteeing reliable delivery of datagrams. It is noteworthy that we currently use IP for routing purposes. This choice is dictated more by budgetary constraints than scientific ones.

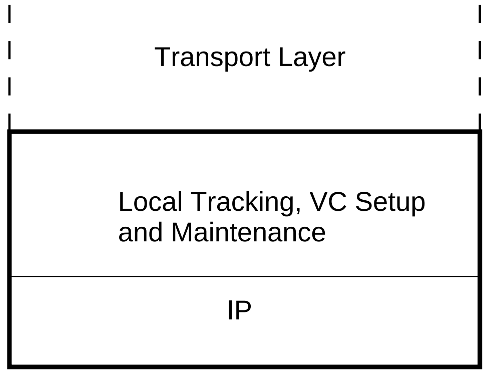

Figure 4: Network Layer.

3.2 Transport Layer Design

When developing transport layer protocols for the mobile environment, it is important to keep the following constraints in mind:

- The wireless link is very fragile and error-prone. This means that any reliable protocol must perform a significant amount of error-recovery.
- The bandwidth of the wireless link will always be a limited resource. Thus, all protocols need to be 'lean'. For instance, in the case of TCP-like protocols, it is not a good idea to have end-to-end flow control (where one end is mobile) because of the high number of control messages that will be sent.
- Available bandwidth within a cell may change dynamically (because it is impossible to control the number of users per cell). This leads to all kinds of problems in guaranteeing QOS parameters such as delay bounds and bandwidth guarantees.
- Mobile hosts frequently encounter extended periods of disconnection (caused due to handshake or due to physical interference with the signal) and this will result in significant losses of data for UDP-like protocols. We need to redefine best-effort service for such cases.
- Mobile hosts may move between SHs after opening transport connections. Should the old SH continue to be responsible for these connections or should control be transferred to the new SH?

All of the above issues can be summarized in the form of two questions. The answer to these questions will determine the transport layer design.

(1) Should the transport layer be aware of the mobility of MHs?

(2) Should the transport layer be aware of bandwidth fluctuations at the wireless link?

If we were to strictly follow the layering idea of the OSI hierarchy, mobility and bandwidth fluctuations will have to be concealed from the transport layer. For mobile networks, however, even though we can adhere to this philosophy, it will result in degraded performance and increased message overhead. To see why this is the case, let us consider two scenarios that will be commonplace in a mobile environment. In the first, a MH with open data connection(s) moves into a cell where there are many other MHs. It is likely that the negotiated QOS for the connections of this MH can no longer be satisfied and will have to be renegotiated. Since renegotiation of QOS parameters involves the transport layer (and the network layer), it is not possible to conceal bandwidth fluctuations from the transport layer. We discuss this problem further in section 3.2.4.

For the second scenario, consider what happens if a MH with an open TCP connection moves from its current SH (where the TCP connection was established) to another SH. In this case it is not necessary to re-establish the TCP connection (because the network layer can forward datagrams from the old SH to the new SH). However, the efficiency of the TCP connection will be degraded for reasons discussed in section 1.1 (recall that I-TCP attempts to alleviate this efficiency problem by breaking the TCP connection in two). In our architecture, as we discuss in section 3.2.1 and section 4, since we perform a great deal of bandwidth management within each mobile sub-network (controlled by a single SH), it becomes necessary to re-establish connections when a MH moves from the sub-network of one SH into the sub-network of another. Thus, we believe, the transport layer will need to be aware of the mobility of MHs and bandwidth fluctuations for each open connection.

Our view for the transport layer of the mobile sub-network is shown in Figure 5. The prefix M stands for 'mobile'. Thus we have a version of mobile-TCP and mobile-UDP. M-CM refers to the mobile-continuous media protocol and is useful for implementing real-time services such as voice or video. It is loosely based on the continuous media protocol of Moran[12]. The mobility management module deals with the problems of mobility (i.e., re-establishing transport connections when a MH moves between sub-networks and renegotiating QOS parameters).

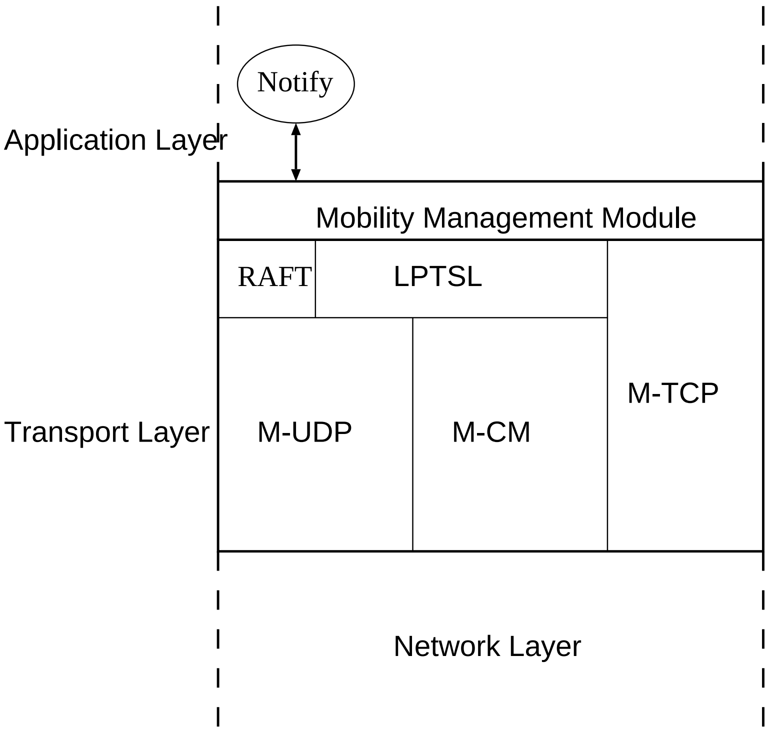

Figure 5: Transport Layer.

3.2.1 M-TCP

In section 1.1 we touched upon the problems associated with providing TCP-like service over mobile networks. Specifically, if the receiver is a MH, the TCP sender times out frequently because of the 'blackout' periods in mobile networks. The solution of Barke[1] breaks the connection in two (similar to our idea with the difference that they consider mobile networks to be part of the internet) – from the fixed host to the MSS and from the MSS to the MH. The MSS effectively serves as the TCP connection end-point from the point of view of the fixed host. The MSS is then responsible for forwarding all data reliably to the MH. Note that the *semantics* of this implementation differ from TCP semantics – it is possible for the sender to think that all data has been correctly delivered to the MH (since the MSS has sent ACKs) even if this is not the case.

In our design, the TCP connection is broken in two – fixed host to SH and SH to MH. The fixed host to SH part of the connection uses regular TCP while M-TCP is used for the SH to MH part of the connection. Since M-TCP is layered on top of a reliable virtual circuit connection (reliable within one mobile sub-network), its design is relatively simplified. Even though this design looks similar to I-TCP, unlike Brake[1], we implement almost TCP-like semantics that make it easier for the sending application to recover from the type of error described above. The TCP client on the SH always ACKs *all but the last byte* of data received from the sender. The last byte is ACKed only after it has been successfully sent to the MH by M-TCP at the SH. In this scenario, if the MH disconnects before receiving all data from the SH, the sender will never receive an ACK for the last byte. Thus, the sending application will know that the connection has failed and can take remedial action. In addition to the above change, we ensure that the buffers at the SH are not exhausted (which can happen, for instance, if the MH has been temporarily disconnected) by linking buffer availability (i.e., receiver window size) to the *available bandwidth on the wireless link*. This causes the TCP client to automatically choke the sender when the MH is in a crowded cell! This linkage between expected wireless bandwidth and TCP buffers is another unique feature of our design.

It is important to note, however, that M-TCP semantics are still slightly different from TCP semantics. This is best illustrated by considering a *talk* application. The application displays data on the user screen as and when it is received over the TCP connection. If the SH crashes, it is possible for the sender to think that almost all the lines it has typed thus far have appeared on the receiver's screen (because ACKs have been received for all but the last byte) even though this may not be the case (i.e., the SH crashed before that data was sent to the MH). This situation is not completely hopeless, however, because the sender will eventually realize that the SH has crashed and it can then take remedial action.

What happens when a MH moves from the sub-network of one SH into the sub-network of another? In our implementation this is handled in a manner similar to I-TCP. The old SH transfers TCP state to the new SH after it has been informed by the mobility management module about the move (recall that whenever a MH enters a new cell, it performs a handshake procedure with the new MSS. The information exchanged during this handshake includes the identity of the previous SH and information about all open connections, see Ghai[6]). Meanwhile all datagrams arriving at the old SH are sent on to the new SH via IP loose source routing.

In our implementation, we require the MH to maintain the identity of the original SH (who first set up the TCP connection). If the MH moves from $SH_{original}$ to SH_1 to SH_2, after SH_1 has transferred TCP-state to SH_2, $SH_{original}$ is given the IP address of SH_2 so that it can route datagrams directly to SH_2 without going through SH_1 first. This method keeps the cost of forwarding datagrams small. In the future, if IP is replaced by a protocol like VIP, we will not have to worry about this particular problem. This is because the intermediate routers in the fixed network will automatically associate the IP address of SH_2 with the VIP address of the MH (see Singh[14] for more details) and route datagrams accordingly.

3.2.2 M-UDP

If we were to implement UDP in the mobile network without any changes, the performance seen by a MH would be very poor. This is because packets transmitted while the MH is moving between cells or is blocked by some physical obstruction are lost (note that in TCP these packets would be retransmitted). Only a small percentage of loss is due to lack of bandwidth. A high-level view of our M-UDP protocol as implemented is the following (a detailed protocol may be found in Brown[2]).

- Every UDP packet is buffered at the SH.
- The SH discards a packet if it has run out of buffer space or if it has been transmitted n times to the MSS.

The semantics of M-UDP are almost identical to UDP in the sense that delivery is not guaranteed. However, M-UDP attempts a 'best effort' service that is constrained only by buffer space availability at the SH and by bandwidth availability on the wireless link. In our experiments we observed a 2 to 3 fold improvement in the number of packets delivered by M-UDP in comparison to UDP (for mobile hosts).

3.2.3 M-CM

A large percentage of future applications will require transmission of data at regular intervals. This kind of data is referred to as continuous media and some examples include voice communication, video communication, etc. Continuous media applications have severe time deadlines and bandwidth requirements and thus cannot be implemented on top of message-based transport protocols such as M-TCP or M-UDP. Following Moran[12] we propose a separate transport protocol suite called M-CM (mobile-continuous media) that will provide the functionality required by such applications. Unfortunately, however, the protocols proposed in Moran[12] and other solutions proposed for the high-speed network domain cannot be adapted to the mobile networks because of two reasons:

- bandwidth availability varies in an unpredictable manner as hosts roam,
- fading and handoff cause periods of disconnection.

All proposals for CM-type protocols require the network layer to provide strict guarantees regarding bandwidth availability and delay bounds for each connection. Because of the above reasons, however, this is not possible in the mobile domain.

Our approach is to provide a "best-possible" service to M-CM connections. Intuitively, this means that the SH will arbitrate between various MH connections to determine how much bandwidth must be allocated to each open connection. Thus, if a MH has an open ftp connection (via M-TCP) and an open video connection (via M-CM), a *scheduler* process (CS process, see section 4) will starve the ftp connection in favor of the video connection if the available bandwidth, within the current cell of the MH, gets reduced. The scheduler also interfaces with LPTSL in case some fraction of data along the M-CM connection(s) need to be discarded. This M-CM protocol has not yet been fully specified.

3.2.4 LPTSL (*Loss Profile Transport Sub-Layer*)

Future applications to be provided to mobile users will include audio (e.g., telephone, audio conferencing, etc.) and video applications (e.g., map information, viewing movies, etc.). These applications have real-time constraints and therefore need to be implemented over M-CM. However, we have a unique problem of *dynamic bandwidth changes during the lifetime of a connection* caused due to the **unpredictable mobility** of mobile hosts.

To illustrate a consequence of this unpredictable mobility, consider a situation where several mobile users have opened high-bandwidth connections. When these connections are set up, the network ensures that the users receive some guaranteed bandwidth. Since these users are all mobile, it is possible that many of them could move into the same cell. In such a situation, it is very likely that the requested bandwidth of the cell will exceed available bandwidth resulting in the original QOS (quality of service) parameters being violated. This situation does not arise in high-speed networks because users are not mobile during the life-time of a connection.

To deal with this situation, we propose that most open connections, *than can tolerate losses*, within the choked cell be penalized (either equally or differentially, based on some need-based policy). Thus, a penalized connection will see a $x\%$ reduction in available bandwidth. The question now is – does the MH renegotiate the QOS parameters to force the sender to reduce the connection bandwidth (e.g., use a higher compression ratio for a video connection) or is data for that connection discarded by the SH? The first option sounds attractive but is not necessarily the right choice for the following reasons:

- The bandwidth crunch is probably temporary and will be alleviated as soon as a MH roams out of the cell. When this happens, the QOS parameters will have to be renegotiated.

- The cell latency of a MH is of the order of several seconds. The end-to-end renegotiation process is time consuming and thus the bandwidth available may change even before this renegotiation is completed.

- While the renegotiation is going on, data will continue to be sent at the old rate. Since the wireless bandwidth is small, the buffers at the SH will possibly overflow.

We propose that the SH *judiciously* discard data for each penalized connection. Since the SH operates at the transport layer (it is a gateway), it can do this. Note, however, that indiscriminate discarding of data may result in garbage at the MH (e.g., if data is randomly thrown out of a compressed video stream, no video can be reconstructed). To solve this problem, we have proposed a new sub-layer called the Loss Profile Transport Sub-Layer. The sending application puts *flags* in the data stream by making calls to LPTSL. All data between a pair of flags represents a *logical segment* (e.g., one logical segment may be a single compressed frame in JPEG-video). The LPTSL at the SH discards entire logical segments in the event of a bandwidth crunch to ensure a $x\%$ reduction in bandwidth (note that the application at the MH can be informed of the location of the discarded segments in case that information is required – as in MPEG-2 video). A detailed protocol is presented in Seal[13]. See Figure 6 for an explanation of the operation of this layer. Here the data stream at the sender is broken into data segments (logical segments) separated by special flags. These flags are inserted by the LPTSL layer at the sender. The LPTSL layer at the SH is informed of the available bandwidth in the current cell of the MH and determines if any data needs to be discarded. If so, it discards entire logical segments (all data between consecutive flags). The LPTSL at the MH passes up the arriving data to its receiving application *and indicates the location and size of the discarded segments.*

A reason for discarding data at the LPTSL is because LPTSL provides different *discarding functions.* When a connection is set up, a QOS parameter negotiated is the discarding function(s) to be used in the event of a bandwidth crunch. For instance, if the connection is an audio connection, the user may prefer uniform random loss as opposed to bursty loss. On the other hand, if the data is compressed video, random loss will prove to be disastrous. In this case the user may opt for bursty loss (i.e., discard entire frames rather than random bytes from several frames). These discarding functions are provided in the form of a indexed library of sub-routines at the LPTSL layer of the SH.

3.2.5 RAFT (*Repetitive Almost-reliable Fast Transport*)

In addition to applications discussed in section 2, PDAs will be used as devices whereby critical data can be transmitted to the user quickly – much like today's pagers or beepers but with a

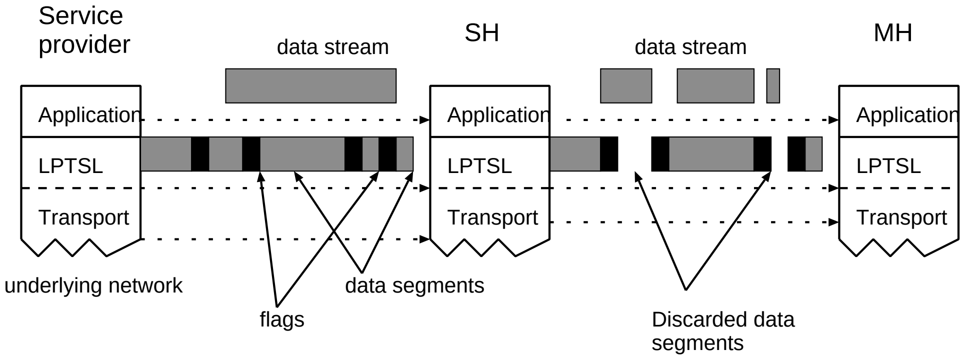

Figure 6: Loss Profile Transport Sub-Layer.

greater degree of sophistication. These type of *notification* applications need to be able to transmit data *quickly* and *reliably* to the mobile user. To facilitate the development of such applications we propose a transport sub-layer called RAFT that is built on top of M-UDP. The approach we use is the following – RAFT at the SH sends data several times. When the MH eventually receives all data correctly it sends a *shut_off* message to RAFT at the SH. RAFT data takes precedence over all other connections.

4 Management and Control

In Figure 7, we show the transport and network layer control and management functions at the SH. The SH needs to ensure that QOS parameters for open connections are maintained. This implies appropriate bandwidth management within each cell and buffer allocation for each connection at the SH. In Figure 7 the management and control functions of the protocol stack for the mobile sub-network is shown in detail. The different arrows indicate control paths, management paths, data paths and QOS negotiation paths.

At the network layer we have a process (CBM) that monitors the bandwidth utilization within each cell controlled by the SH and passes this information up to a transport layer process (CS) via a management path. The CS process arbitrates bandwidth within each cell between all open connections. Thus, if there is a real-time connection (e.g., audio) and a data connection into a cell, the CS may choose to starve the data connection in order to ensure the delay bounds for the real-time connection are met. Another control process at the network layer (NL_QOS) monitors the QOS being delivered to each VC and sends this information to the CS as well. This management path is required to ensure that QOS parameters (such as bandwidth usage or delays) for M-CM connections are met. If some M-CM connection has not been receiving its negotiated QOS, the CS process allocates more bandwidth to that connection.

CS is the process which is responsible for maintaining QOS for all M-CM connections and ensuring that data connections (M-UDP or M-TCP) do not get starved in the process. The CS receives information for each M-CM connection's QOS contract. Note that this contract may be renegotiated during the lifetime of the connection and thus CS needs to be informed of this change via the management path. M-UDP connections are subject to data discarding (via LPTSL) in the event of a bandwidth crunch. Thus, we assume that some QOS negotiation also takes place for UDP connections and this information is passed on to the CS process as well. CS *periodically* informs M-CM, M-UDP and M-TCP of the available bandwidth for each connection via control paths. It is up to these protocols to ensure that they control each open connection (either choke the sender as in M-TCP, or discard data via LPTSL as in M-CM) adequately.

QOS negotiations take place between M-CM's QOS process, NL_QOS and possibly the QOS process on the fixed network (in case the connection is to a fixed host) via QOS negotiation paths. QOS control information is also exchanged between LPTSL and the QOS processes of M-UDP and M-CM. Finally, the buffer manager processes at both, the network layer and the transport layer, are responsible for buffer allocation to different connections. The data paths followed for some typical connections are illustrated in the figure as well.

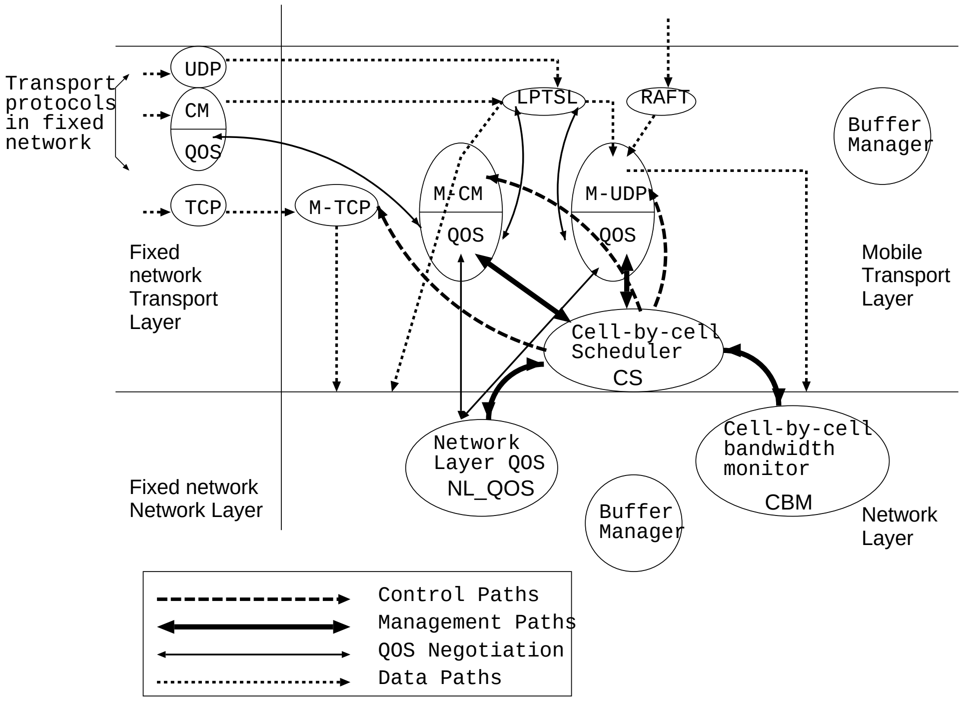

Figure 7: Management and Data flow at the transport and network layers of the SH.

5 Conclusions

In this paper we have proposed a complete design of the network layer and transport layer in a manner that best addresses the problems of the mobile environment. Our design is a radical departure from other researchers in that we propose that mobile networks be viewed as separate from wired networks with connectivity being provided by special gateway nodes. These nodes provide a variety of transport level services that best meet the constraints of the mobile environment. A complete implementation of this architecture is underway and Figure 8 indicates the current status of this implementation.

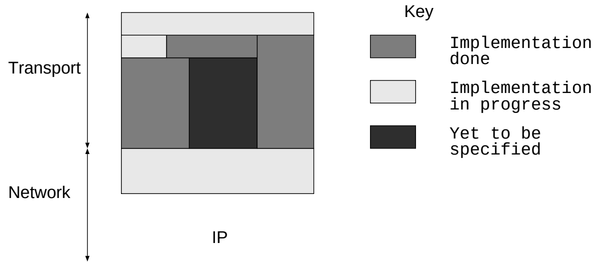

Figure 8: Current status of implementation.

References

[1] A. Bakre and B. R. Badrinath,"I-TCP: Indirect TCP for Mobile Hosts", *Technical Report* DCS-TR-314, Rutgers University, Piscataway, NJ 08855.

[2] K. Brown and S. Singh,"M-UDP: Mobile UDP", Manuscript.

[3] I. Arieh Cimet, "How to Assign Service Areas in a Cellular Mobile Telephone System", *IEEE ICC'94*, pp. 197-200, May 1994.

[4] J. S. DaSilva and B. E. Fernandes,"The European Research Program for Advanced Mobile Systems", *IEEE Personal Communications Magazine*, February 1995, pp. 14-19.

[5] D. Duchamp, Steven K. Feiner and G. Q. Maguire, "Software technology for wireless Mobile computing" *IEEE Network Mag.*, pp 12-18, November 1991.

[6] R. Ghai and S. Singh,"An Architecture and Communication Protocol for Picocellular Networks",*IEEE Personal Communications Magazine*, Vol. 1(3), 1994, pp. 36-46.

[7] David J. Goodman, "Cellular Packet Communications", *IEEE Trans. on Comm.*, vol. 38, no. 8, pp 1272-1280, August 1990.

[8] David J. Goodman, "Trends in Cellular and Cordless Communications",*IEEE Communications Magazine*, pp 31-40, June 1991.

[9] J. Ioanidis, D. Duchamp and G. Q. Maguire, "IP-based protocols for mobile internetworking" *Proc. of ACM SIGCOMM'91*, pp 235-245, September 1991.

[10] D. B. Johnson,"Mobile Host Internetworking Using IP Loose Source Routing", Technical Report CMU-CS-93-128, Carnegie Mellon University, Pittsburgh, PA 15213, 1993.

[11] C.S. Joseph, *et al*, "Propagation Measurement to Support Third Generation Mobile Radio Network Planning", *43rd IEEE Vehicular Tech. Conf.*, May 1993, pp. 61-64.

[12] M. Moran and B. Wolfinger,"Design of a Continuous Media Data Transport Service and Protocol", Technical Report TR-92-019, Computer Science Division, University of California Berkeley, April 1992.

[13] K. Seal and S. Singh,"Loss Profiles: A Quality of Service Measure in Mobile Computing", *J. Wireless Networks*, (submitted).

[14] S. Singh,"Quality of Service Guarantees in Mobile Computing", *Journal Computer Communications*, (to appear).

[15] F. Teraoka and M. Tokoro, "Host Migration Transparency in IP Networks: The VIP Approach", *SIGCOMM*, Vol. 23, No. 1, Jan 1993, pp. 45-65.