Cellular IP: A New Approach to Internet Host Mobility

András G. Valkó[1]
Ericsson Research
andras.valko@lt.eth.ericsson.se

Abstract

This paper describes a new approach to Internet host mobility. We argue that by separating local and wide area mobility, the performance of existing mobile host protocols (e.g. Mobile IP) can be significantly improved. We propose Cellular IP, a new lightweight and robust protocol that is optimized to support local mobility but efficiently interworks with Mobile IP to provide wide area mobility support. Cellular IP shows great benefit in comparison to existing host mobility proposals for environments where mobile hosts migrate frequently, which we argue, will be the rule rather than the exception as Internet wireless access becomes ubiquitous. Cellular IP maintains distributed cache for location management and routing purposes. Distributed paging cache coarsely maintains the position of 'idle' mobile hosts in a service area. Cellular IP uses this paging cache to quickly and efficiently pinpoint 'idle' mobile hosts that wish to engage in 'active' communications. This approach is beneficial because it can accommodate a large number of users attached to the network without overloading the location management system. Distributed routing cache maintains the position of active mobile hosts in the service area and dynamically refreshes the routing state in response to the handoff of active mobile hosts. These distributed location management and routing algorithms lend themselves to a simple and low cost implementation of Internet host mobility requiring no new packet formats, encapsulations or address space allocation beyond what is present in IP.

1 INTRODUCTION

As computers become smaller and global networking ubiquitous, the demand to provide network access to mobile users will grow rapidly. Recently, a number of IP frameworks have emerged to offer connectivity to mobile users [1] [2] [3] [4]. A basic difficulty that these protocols address is that the host address in IP has dual significance. First, as a unique identifier it should be kept constant regardless of host mobility. Second, in its role as a location pointer it should change as hosts change location [5]. These are competing requirements that mobile host protocols should efficiently resolve. A fundamental problem to solve is therefore the separation of these two goals while an up-to-date mapping of host identifiers to location information is made available. Overviews of existing protocol proposals are presented in [5] and [6].

In this paper we address host mobility in an environment where a wireless connection to the Internet is typical, rather than as it is today, an exception. We therefore assume an environment where highly mobile hosts often migrate during active data transfer and expect the network to manage these handoffs with minimum disturbance to ongoing data sessions. While people rarely read text or watch video while walking or driving, they may wish to, however, download files, browse the web, or talk on the Internet phone while on the move [4]. As small palmtop computers become more affordable, there is a need for cheap and ubiquitous wireless Internet access to make the vision of global mobile computing a reality. In such a world the popularity of cellular telephony would be out stripped by the use of versatile Internet-enabled palmtops that offered a variety of services and seamlessly migrated without any user intervention. It has been shown in [4] that in such an environment Mobile IP [7], optimized for macro-level mobility and relatively slow moving hosts, is no longer an optimal solution. Mobile IP requires that after each migration a location update mes-

[1] Visiting scientist at the Center for Telecommunications Research, Columbia University, New York.

sage be sent to a possibly distant home agent potentially increasing handoff latency and load on the global Internet. To overcome these limitations, and following on from the work by Caceres [4] on hierarchical mobility management approaches, we assume a mobile networking architecture where local *wireless access networks* handle local mobility while a Mobile IP capable Internet provides wide area mobility. In this case a mobile host's home agent is only informed when the host moves into a new access network and is unaware of the hosts mobility within an access network as illustrated in Figure 1. The main advantage of separating local and wide area mobility is that home agents need not be informed about local mobility within a wireless access network. We believe that this will become increasingly important as cells become smaller, host migration frequency faster and user population greater. By handling the majority of handoff control locally we argue that we can engineer faster handoffs and limit the impact of handoff on active data sessions while avoiding the exposure of local migration to distant home agents.

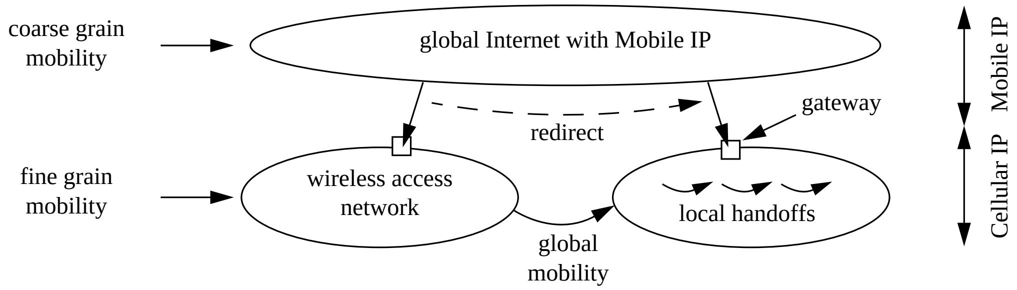

Figure 1 *Wireless access networks and Mobile IP*

A further benefit of a hierarchical architectural approach to mobility arises when many users wish to carry mobile-capable computers that need to be permanently 'connected' to the Internet as they move as is the norm today in mobile telecommunications. These users will generate frequent location update messages even when they are not actively transmitting data. Decreasing the cell size will place further demands on the system as the number of location updates increases relative to generated traffic. To allow for a large number of hosts to be connected, mobile host protocols need a simple location tracking scheme that imposes neither traffic nor processing load on the global network as long as the host is idle. We will refer to this important property as *cheap passive connectivity*.

In this paper we address these challenges by proposing *Cellular IP*, a new host mobility protocol that is optimized for wireless access networks and highly mobile hosts. The primary design goal of Cellular IP is *simplicity*: we envision that a Cellular IP wireless access point (base station) can be implemented as a small and cheap "commodity device". This makes the solution applicable to indoor systems where a base station is needed in each office or office floor. On the other hand, Cellular IP can also be implemented on top of regular IP routers to allow for easy migration from existing installations. Another key design consideration is *scalability*. The Cellular IP distributed location management makes it possible to use the same protocol and the same topology-unaware nodes in small indoor systems, large area networks or even heterogeneous systems. This allows mobile users to migrate freely and seamlessly between areas with different characteristics. They can obtain basic connectivity from the local cellular telephony operator's Cellular IP service while walking in streets but receive high bandwidth, using the same protocol, as soon as they enter an office building as envisioned in the case of wireless overlay networks [15]. "Performance transparency" or location independent service quality identified in [5] as a main requirement for mobile host protocols is not feasible in a wireless environment. Instead, we define *performance scalability*, that is the ability to use the same protocol in distinct environments, always obtaining the locally available level of service. Performance scalability also allows network operators to extend their networks when demand increases. This is important because in wireless systems the aggregate capacity is determined by the density of base stations and can be increased in exchange for increased equipment cost by installing new base stations.

Another benefit of Cellular IP is that it is fully compatible with IP. It requires neither new packet format or encapsulation, nor extra address space. Cellular IP systems use three types of control packets which can be implemented as a new IP option. This does not assume the updating of regular IP routers because these packets never leave the wireless access network. The paper is structured as follows. In Section 2, we discuss alternative host mobility approaches and compare our approach to the literature. A wireless access network model is presented in Section 3. This is followed in Section 4 by a description of the Cellular IP concepts. In Section 5 and 6 we discuss the design and implementation issues, respectively. Finally, in Section 7 we offer some concluding remarks.

2 RELATED WORK

A solution that requires mobile hosts to be restarted after migration supports portability and not mobility [5]. The primary design goal for mobile host protocols is therefore to allow for a host to change its point of access without being reconfigured by the user. This requirement is often referred to as operational transparency [5]. Users may sometimes wish to migrate during data transfer. A change of access point while connectivity is maintained is typically called a handoff. Another important design goal for mobile host protocols is to support handoffs without significant disturbance to ongoing data transmission.

It is shown in [6] that many of the proposed mobile host protocols, including the IETF Mobile IP solution [7] can be viewed as special cases of a “two tier addressing” architecture where a mobile host is logically associated with two IP addresses; that is, its home address that serves as an unchanged host-identifier and an address that reflects its current point of attachment to the Internet. This general architecture comprises three fundamental components. A Location Directory represents a data base, possibly distributed, that contains the most up-to-date mapping between the two address spaces. The translation of the host identifier to the actual destination address in each packet is performed by Address Translation Agents and involves the querying of either the Location Directory or a local cache. The final component of the generalized architecture is the Forwarding Agent that assures that packets arriving at the mobile host have its constant home address in the destination field.

While these solutions, in particular the IETF Mobile IP solution, meet the goals of operational transparency and handoff support, they are optimized for slowly moving hosts and become inefficient in the case of frequent migrations [4]. Mobile IP requires that the mobile host’s home agent be informed whenever the host moves to a new Foreign Agent. During the update messaging phase, packets will be forwarded to the old location and will not be delivered hence disturbing active data transmission. Similarly, during route optimization [8] [9], data transfer is disrupted while the correspondant host obtains a new binding. The effect of these delays grows with increasing handoff frequency. In addition, the update messages load both the Internet and the home agents even when the mobile host is idle while moving. This load is proportional to the number of mobile hosts and not to the generated traffic. This may be a problem as host mobility becomes more ubiquitous and cell sizes smaller.

To overcome the limitations of Mobile IP, a hierarchical mobility management approach is proposed in [4]. Three levels of mobility are defined, namely local mobility, mobility within an administrative domain and global mobility. A local mobility management protocol is proposed for the lowest level and Mobile IP is assumed in the highest global level. Cellular IP differs from this approach in two important aspects. First, instead of defining hierarchical levels of mobility, Cellular IP provides a framework that can scale operation from small office systems to large area networks. Building on the wireless overlay networks paradigm [15] we envision that small and large area Cellular IP networks will typically overlap. In an office building, one can be covered by the metropolitan area network, the campus network and an indoor system, each using the same Cellular IP protocol but with different settings of the location management devices (see Section 4). Second, Cellular IP includes an efficient location management and searching scheme that avoids tracking the mobility of idle users, thus reducing load on wireless access networks. This feature is needed to provide cheap passive connectivity that we identified as an important goal for future wireless IP systems.

Loosely tracking the location of idle mobile users and then “zooming in” on them when they are engaged in active communication (e.g., a conversation) is a technique familiar to cellular telephony. In the Global System for Mobile Communications (GSM), for instance, idle mobiles are located with a granularity of location areas and are searched for (i.e., *paged*) in the location area when an incoming call arrives [10]. By increasing the size of the location areas, paging traffic can be traded for location update traffic. The optimal operation point depends on the users’ call and mobility characteristics. Cellular IP builds on this concept and similar to cellular telephony systems avoids resource consuming location update procedures for idle users. Unlike voice systems, however, Cellular IP can not rely on a connection establishment phase to search for mobile hosts. In contrast, location management is based on a lightweight *soft-state* signalling system that distinguishes active and idle mobile hosts without introducing the notion of connections.

Such a vision also differentiates our approach from current efforts to provide data service over cellular telephony networks. The General Packet Radio Service (GPRS), for instance, requires that a “logical link context” be established between the mobile host and the network before data can be sent. As long as this attachment persists, this virtual “connection” migrates in the network with the mobile host [11]. In addition, GPRS is implemented on top of a GSM infrastructure which limits its applicability to small scale networks, especially indoor systems.

Third generation mobile systems, referred to as International Mobile Telecommunications 2000 (IMT-2000) are expected to offer world wide data services to mobile users [13]. These will be combined ATM/CDMA systems and will support services ranging from compressed voice to high quality multimedia [14]. Compared to this evolving standard, Cellular IP takes a simplistic approach by offering non-guaranteed datagram delivery and hence is more appropriate for best effort traffic. However, Cellular IP is designed around the IP paradigm and therefore can evolve to accomodate future IP quality of service schemes, e.g., differentiated services [16].

Recently, IP multicasting techniques have been proposed to solve location-independent addressing [12]. While Cellular IP is similar to those techniques in using a time-variant tree of paths to reach the mobile host, the multicasting-based solution requires that the location of even idle mobile hosts be continuously tracked by the network. This, however, does not meet the requirement for cheap passive connectivity. In addition, like most mobile host protocols [1] [2] [4], it requires that an enhanced IP router be integrated in each base station which is an expensive device in comparison to what we envision as a Cellular IP base station.

3 NETWORK MODEL

A wireless access network primarily consists of base stations interconnected by wired links as illustrated in Figure 2. Apart from the base stations, the network can contain nodes that have no radio device but serve as traffic concentrators or support mobility management functions. In Figure 2, all nodes with the exception of node **E** have radio devices.

Wireless access networks are connected to the global Internet by routers, called gateway routers. This router is also the best location for the home agent if the access network is the home network for some mobile hosts and it can serve as foreign agent for visiting hosts. We assume that in the global Internet, Mobile IP supports host mobility with a granularity of wireless access networks. Upon entering an access network (step 1 in Figure 2), the mobile host (**X**) registers with its home agent (step 2) which will forward packets addressed to the host to the access network (step 3). As long as the host is connected to the same access network, mobility is hidden from the home agent. Mobility between access networks occurs at a slower time scale, hence allowing Mobile IP to optimize for infrequent migrations.

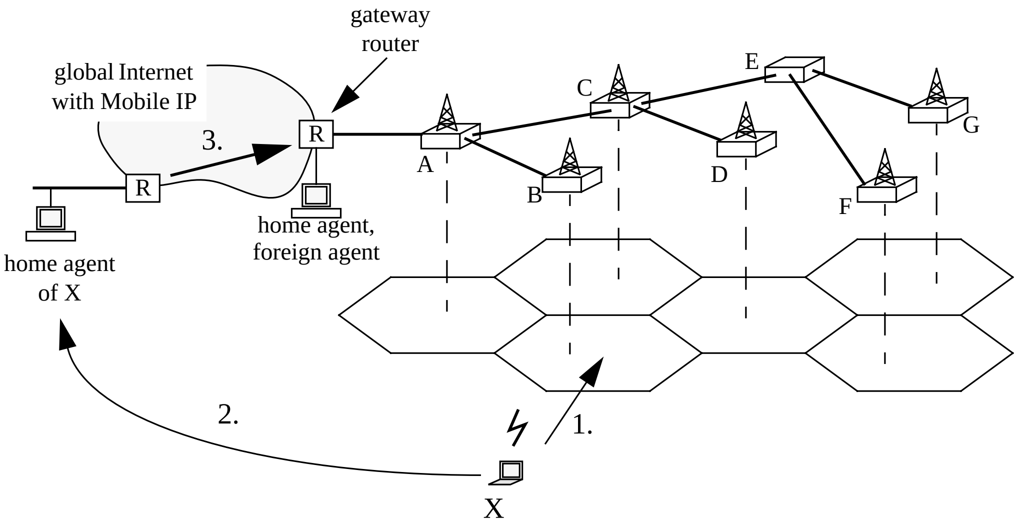

Figure 2 *Wireless access network model*

In what follows, we identify some key requirements for wireless access networks. Within the access network's service area, mobile hosts have no home location or dedicated point of attachment. Base stations periodically emit beacon signals to allow for hosts to identify an available base station. Visiting mobile hosts are treated the same as "home subscribers" as long as they are attached to the network. When a host connects or leaves the access network, it must inform its home agent as required by Mobile IP. In addition, the access network may require a registration and authentication procedure. To facilitate global migration, the access network specific registration must be simple and fast. Although for global reachability, the host must obtain a local care-of address, it is advantageous in terms of operational transparency if inside the access network, it is identified by its home IP address.

In an environment where users carry their wireless-enabled computers switched on most of the time, another important requirement is to allow for the maximum possible number of users in a given wireless access network. While traffic generated by mobile users will be limited by the network elements' capacity, idle users, by the mere fact of being reachable should impose little load. As mentioned in the Introduction, we refer to this requirement as "cheap passive connectivity".

Wireless cells may overlap facilitating seamless handoff support. In this case, the wireless access network protocol can support soft handoff, meaning that if the host changes cell without becoming temporarily unreachable, packets continue to be delivered with little disturbance. This is achieved by temporarily allowing simultaneous transmission to/from both base stations. However, if this is not possible because cells do not overlap, the protocol should still operate efficiently.

Because cells may cover small areas (e.g., rooms or sections of a highway), the wireless access network should perform well in the presence of very frequent migrations. Tracking mobile host movement with a granularity of cells requires that a control message be sent after every migration to a location data base and be processed there which becomes inefficient at high migration frequencies. However, letting the hosts roam in the service area untracked, and searching for them only when there is data to deliver is inefficient and unscalable. An efficient location management scheme is required that maintains location information of idle mobile hosts without overloading the network with location update messages. Location management should also incorporate a quick and efficient searching algorithm used when there is data to be routed to idle hosts. As the migration frequency may vary in networks, the location management scheme should be adaptable to local characteristics. In particular, in environments of low migration frequency, more accurate location information should be maintained for idle mobiles. In contrast, systems of high migration frequency may need to rely heavily on searching on-demand to limit the load imposed by location update messaging.

To allow for cheap end-user devices, a wireless access network should assume little complexity in mobile hosts. Ideally, a mobile host is memoryless in the sense that it keeps performing the same elementary actions to stay connected whenever it is within reach of a base station and remains idle otherwise. No special actions should be required at a handoff, or after a temporary radio channel black-out.

In summary, five key requirements of wireless access networks motivate the design of the Cellular IP protocol:

1. easy global migration;
2. cheap passive connectivity;
3. flexible handoff support;
4. efficient location management; and
5. simple memoryless mobile host behaviour.

4 CELLULAR IP

In addition to the requirements discussed in Section 3, the primary design objective of Cellular IP is to provide maximum scalability and robustness with minimal complexity. A Cellular IP network is fully distributed where

- nodes are unaware of the network topology;
- no centralized data bases or other single points of failure exist; and
- no element in the network must increase in complexity as the coverage area (and hence the potential number of connected hosts) increases.

4.1 Paging and Routing Mappings

For simplicity and scalability, in a Cellular IP network none of the nodes know the exact location of a mobile host. Packets addressed to a mobile host are routed to its current base station on a hop-by-hop basis where each node only needs to know on which of its outgoing ports to forward packets. This limited routing information is local to the host and does not assume that nodes have any knowledge of the wireless access network's topology. We refer to these information elements as *mappings* because they map mobile host identifiers (IP addresses) to node ports. Mappings are created by packets transmitted by mobile hosts. These packets travel in the access network toward the gateway router, routed on a hop-by-hop basis. By monitoring these packets and by mapping sender address to incoming port, nodes of the access network create a hop-by-hop reverse path for future packets addressed to the given host.

In order to minimize control messaging, mappings are not cleared in an explicit way after handoff. Rather, they are assigned timers to clear outdated mappings. This implies that to maintain its path of mappings, a mobile host must periodically transmit dummy packets when it has no real data to send. The combination of periodic transmitted packets and timed-out mappings ensure that as a mobile host roams in a service area, an up-to-date path of mappings will always exist between the gateway and the mobile host's base station. This scheme also results in easy migration between access networks because nodes need no advance information on a mobile host to create mappings and need not be informed when the host leaves the area.

Relying on timers, however, gives rise to the following trade-off. After a host performs a handoff, its path to the "old" base station will remain valid until the mappings are cleared. If in this period packets are routed to the host, they are delivered not only at its current base station but also to the old base station. This results in a waste of resources that can be minimized by selecting a small timeout interval. On the other hand, idle mobile hosts need to transmit dummy packets with a period comparable to the mapping timeout which may result in a significant load in the network. This can be particularly costly on scarce radio resources when the timeout interval is too small.

To overcome this problem, we first observe that the system has two characteristic time scales. To minimize resource waste due to unused but not-yet cleared mappings, the timeout should be in the order of the packet time scale. For a reasonable periodicity of dummy packets, on the other hand, it should operate at the host mobility time scale which may be orders of magnitude higher than the packet time scale.

Cellular IP solves this problem by using two parallel structures of mappings. Nodes maintain one set of mappings, called *Paging Caches (PC)*, for idle mobile hosts. These mappings have a timeout interval comparable to the migration frequency, possibly in the order of seconds or minutes. Independent of Paging Caches, nodes maintain another set of mappings called *Routing Caches (RC)*. These mappings are only maintained for mobile hosts currently receiving or expecting to receive data. For Routing Cache mappings the timeout can be in the packet time scale. Hence one can view PCs as a coarse location information data base for idle mobile hosts that zooms in on active hosts through the creation of RC mappings.

Figure 3 illustrates the relationship of PCs and RCs. While idle, the mobile host **X** keeps PCs up-to-date by transmitting dummy packets at a low frequency (step 1 in Figure 3). PCs have a relatively long timeout, hence they follow the roaming mobile host somewhat coarsely. When there are data packets to be routed to the mobile host, the PC mappings are used to find the host (step 2). As long as data packets keep arriving, the host maintains RC mappings, either by its outgoing data packets or through the transmission of dummy packets (step 3). Data packets addressed to the host are routed by RCs (step 4) that unlike PCs track the host closely.

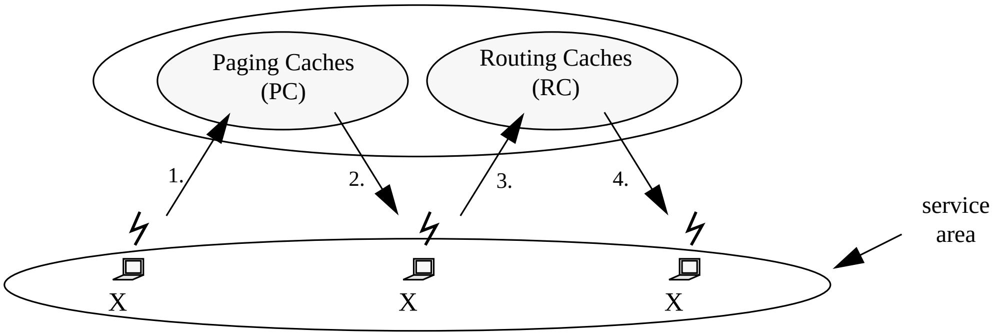

Figure 3 *Paging and Routing*

The separation of paging and routing has a further advantage. A wireless access network can have a large number of mobile hosts attached to it at a time, only a small percentage of which are receiving data packets. In this case, Paging Cache will contain a large number of hosts mappings at any time, making it a considerably larger data base than Routing Cache. As PCs are only used to search for mobile hosts, and not to route high bit rate data, the network operator can choose to place PCs in only a small number of well positioned nodes and let other nodes broadcast search messages. By creating more PCs, the location information can be made more precise thus reducing the size of the searched area. This design feature gives network operators the freedom to tune location management according to the network and mobility characteristics. We will discuss this issue in Section 6.

4.2 Paging

Idle mobile hosts periodically generate short control packets, called *paging-update packets* sending them to the nearest available base station. The paging-update packets travel in the access network toward the gateway router (GW), routed on a hop-by-hop basis, as illustrated in Figure 4. Nodes equipped with Paging Cache monitor passing paging-update packets and maintain the cache that maps mobile host identifiers (see

Section 5.1 on mobile addressing) to the port through which the paging-update packet arrived. The gateway router discards paging-update packets isolating Cellular IP specific operations from the Internet.

As illustrated in Figure 4, a mobile host **X** is currently in the cell of node **G**. Paging-update packets generated by the host travel toward the GW through nodes **G**, **E**, **C** and **A**. In this example network, nodes **A** and **E** contain PCs, but node **C** does not. Hence **C** simply forwards the paging-update packets toward the GW without registering location information about host **X**. Node **A** notes that the packets from **X** arrived via the port toward **C**, while **E** notes that they arrived via the port toward **G**.

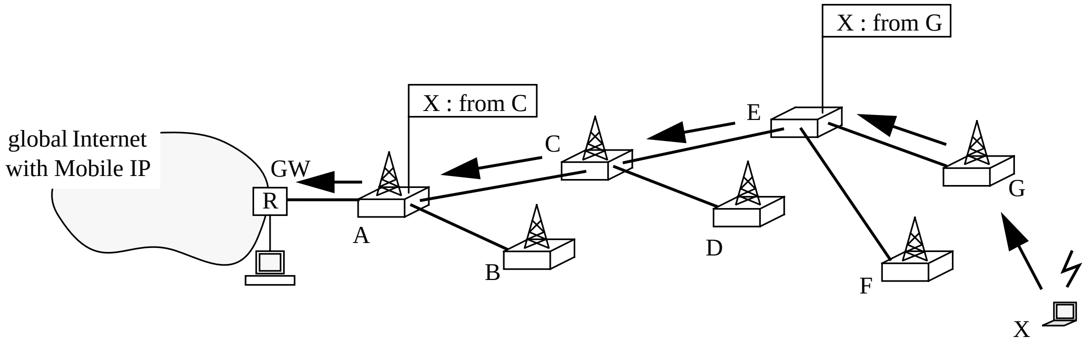

Figure 4 *Paging-update packets create mappings in PCs*

As the idle host moves, it keeps sending its paging-update packets to the nearest base station, forcing Paging Caches to have up-to-date mappings. Outdated mappings are cleared after a system-specific timeout. If, for instance, host **X** moves to cell **F**, its paging-update packets will now be sent to **F** as is illustrated in Figure 5. While node **A** will not notice a difference, in node **E** a new mapping for **X** will be created and after a while the old mapping will be timed out and cleared. For a short time the two mappings coexist guaranteeing that the host always remains reachable during migration.

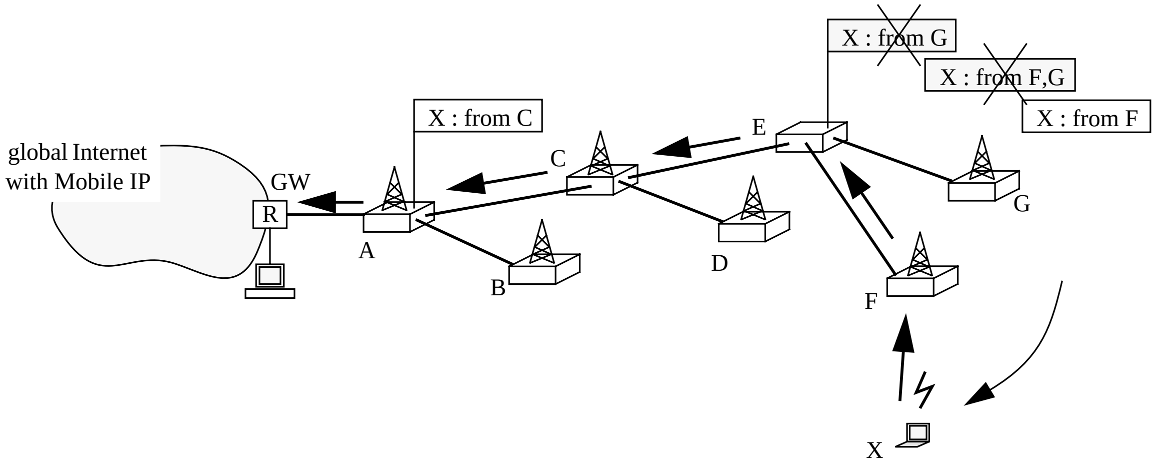

Figure 5 *PCs updated for a moving host*

When IP packets arrive at the gateway router, addressed to a mobile host for which no up-to-date routing information is available, the Paging Caches are used to find the host. The gateway queues the arrived IP packets and generates a control packet, called a *paging packet*, that contains the identifier of the mobile host being searched for. The paging packet is routed in the access network by Paging Caches that simply reverse the route taken by recent paging-update packets. If all nodes have Paging Caches, a full hop-by-hop route is available to the host's current location. If some nodes do not have PC, then they will forward the paging packet to all outgoing ports.

Continuing our example, in Figure 6 to route paging packets **A** checks its cache and finds that paging-update packets from **X** have recently arrived via the port toward node **C**. Hence **A** forwards the paging packet to **C** which in turn has no information about the host and forwards the packet to both possible directions. The paging packet sent to **D** is discarded because **D** knows that the host is not in its cell. The one sent to **E**, on the other hand, causes **E** to check its cache and find that **X** has been sending packets through **F**; therefore **E** forwards the paging packet to **F** and on to **X**.

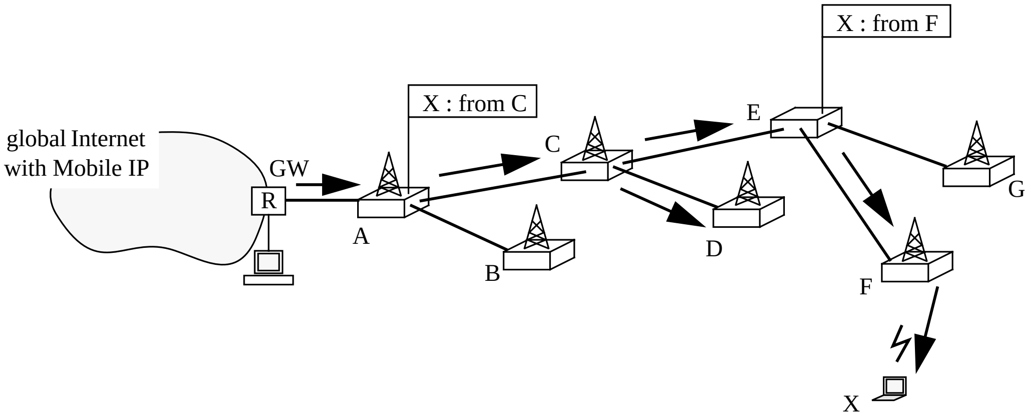

Figure 6 *Paging packets are routed to the mobile host by PCs*

Upon receiving the paging packet, the mobile host creates a control packet called a *route-update packet* and sends it to its base station (**F**). Similar to paging-update packets, route-update packets travel to the GW routed on a hop-by-hop basis, and create mappings for the sending mobile host in Routing Caches on the way. When the route-update packets reach the GW router, all RCs on the way are configured and data packet(s) queued in the GW can be delivered to the mobile host. This searching process delays the delivery of the first data packet(s) but once the path is established, following packets use it without repeated searching. However, if at any time during data transfer the host becomes temporarily unreachable and RCs time out, the next incoming data packets will automatically generate a new paging process. Hence such temporary radio channel black-outs need no interaction from the mobile host and result only in an extra paging delay in the communication.

4.3 Routing

Data packets transmitted by the mobile host are routed to the GW on a hop-by-hop basis. Nodes that contain Routing Cache monitor these passing data packets and use them to create a mapping of host identifiers to port numbers. Packets addressed to the mobile host are routed along the reverse path, hop-by-hop, by these Routing Caches and are broadcast where no routing information is available.

Table 1: Comparison of Paging and Routing

	Paging Cache (PC)	Routing Cache (RC)
driven by	all mobile-originated packets (data, route-update, paging-update)	mobile originated data and route-update packets
scope	both idle and active mobile hosts	active mobile hosts only
purpose	route paging packets	route mobile-addressed data packets
time scale	mobility	packet

The structure and basic operation of routing is much the same as that of paging. To clarify the duality between the two, we summarize the operation of PCs and RCs in Table 1. It may be worth noting once again that the two functions are separated because of the two intrinsic time scales characteristic to mobile systems. Routing deals with active hosts only (i.e. hosts receiving or transmitting data) and it is updated at a packet time scale. This allows Paging Caches to operate at a mobility time scale and hence avoid very frequent paging-updates by idle hosts.

The mobile host may keep receiving data packets without sending data for some time. To keep RCs configured and to avoid repeated paging, mobile hosts expecting data (when, for instance, a TCP connection is open) but having no packets to transmit must keep transmitting route-update packets periodically. Like data packets, route-update packets configure Routing Caches and ensure that the hop-by-hop route from the GW toward the mobile host remains up-to-date. We note that for reliability purposes, PCs do not stop tracking hosts while they are active. However, active hosts need not send paging-update packets because PCs are also configured by route-update and mobile-originated data packets.

4.4 Handoff

The mechanisms described above ensure that handoff, that is a migration during an ongoing data transfer, is handled automatically. Handoff in Cellular IP is always initiated by the mobile host. As the host approaches a new base station, it redirects its data packets from the old to the new base station. The first of these redirected packets will automatically configure a new path of RC mappings for the host, this time to the new base station. For a time equal to the timeout of RC mappings, packets addressed to the mobile host will be delivered at both the old and new base stations. This guarantees that if the host’s radio device is capable of listening to two logical channels, the handoff will be soft. If the host can not listen to both base stations at the same time then the performance of hard handoff will depend on the radio device. After a while, the path to the old base station will time out and clear, while packets will continue to be delivered to the host at its current location via the new base station.

Figure 7 illustrates a handoff scenario. The mobile host **X** is moving from cell **F** to **D** while it is sending and receiving data packets. Assuming that all nodes have Routing Caches, before the move, the RC mappings are as indicated in the flags. After the migration, the mobile-originated data packets, indicated by solid arrows, cause the cache in node **C** to create a new mapping. For some time, packets addressed to **X** are delivered both to **D** and to **F** (shown as dotted arrows). After the RC timeout, the old mapping to **E** is cleared, as is the mapping to **F** in **E**. From then on, only the up-to-date route is used. The cache in **A** is unchanged during the whole process.

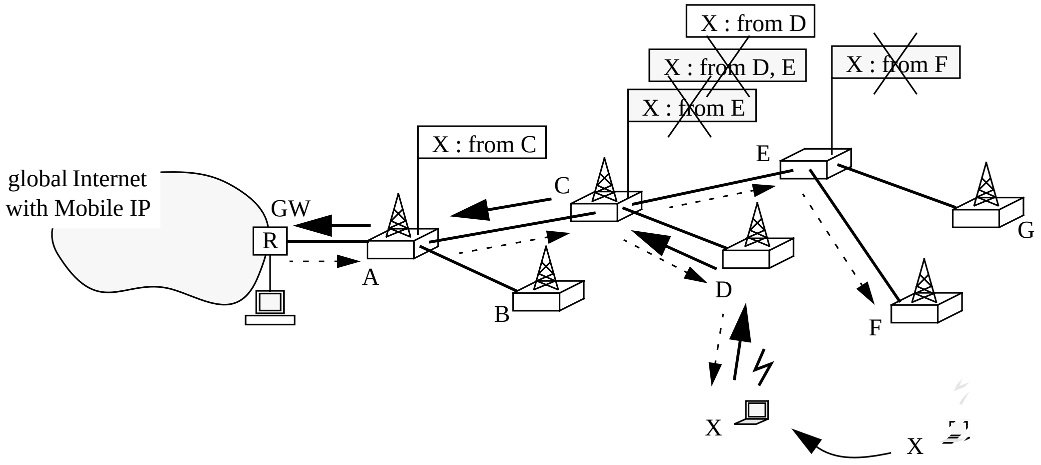

Figure 7 *Handoff*

This handoff process is simple, transparent and automatic. In nodes where the old and new paths take the same route, the old mappings are automatically reused rendering the search for an optimal cross-over point unnecessary in Cellular IP. In addition, if the old and new cells overlap, there is little interruption or disturbance in communications. If at handoff the mobile host remains temporarily out of radio contact while moving between two cells, the upper layers (e.g., TCP) may notice a delay and some packets maybe lost, but communication is resumed as soon as the host appears in the new cell. We note that this also applies to hosts becoming temporarily unreachable due to reasons other than handoff. If a host reappears before the RC timeout, service continues without any further delay. If RCs have timed out, they are reconfigured by the first packets transmitted by the host which does not even have to know about the disruption or notice whether it reappeared in the same or in another cell.

In the description of the handoff process, we implicitly assumed that the mobile host always has data packets to send. If this is not the case when the handoff occurs, the host would send route-update packets as soon as it arrives to the new cell to configure the new route. As route-update packets have the same effect as data packets, the handoff mechanism is the same in this case. Of course, if the host continues to have no data to send, it will keep sending route-update packets at its new location to maintain the route.

While this process would normally result in smooth handoff, in some cases (e.g. in indoor systems) handoffs can occur quickly or the mobile host, being on the border of two cells, can flip-flop between two base stations. To ensure continuous communication in these situations, the mobile host maintains a RC route to both base stations by sending its data packets to one and sending, in parallel, route-update packets to the other base station. In this case, the network is prepared for the handoff, and data transmission will be continuous if the host suddenly becomes unreachable by one of the base stations. We note that the same method can be used for idle mobile hosts to ensure reachability: instead of sending paging-update packets to just one base station, the host can send them to two or more base stations in parallel. Cellular IP can use these strategies to enhance reachability and handoff quality in exchange for network efficiency.

5 PROTOCOL DESIGN ISSUES

In the previous sections we described the Cellular IP concept and its functions of paging, routing and handoff. In the following, we discuss some algorithmic details of Cellular IP.

5.1 Addressing and Migration

As mobile host addresses have no location significance inside a Cellular IP network, any space of unique host identifiers can be used. The use of the home IP addresses is a simple solution and it has the advantage that IP packets can be used in the access network unchanged. Neither encapsulation, nor address conversion is needed in a Cellular IP network. In addition, the use of IP addresses as mobile host identifiers makes migration between access networks surprisingly simple. A mobile host entering a Cellular IP network simply has to communicate the local GW's address to its home agent as care-of-address.[1] The home agent will tunnel its packets to the GW, which will "detunnel" and forward them to the Cellular IP network. Paging and Routing Caches need no advance information to start creating mappings for the newly attached host, nor do they need to be informed when a mobile leaves the access network.

5.2 Control Packet Types

The three types of control packets (paging-update, route-update and paging) used by Cellular IP can be regular IP packets. All three types contain a single information element only: the mobile host's identifier. As this is indicated in the source or destination field of the packet (for paging-update/route-update and paging

[1] The GW's IP address can be included in the base station's beacon signal. It then also serves to identify the wireless access network when the host is covered by several overlapping access networks.

packets, respectively), the payload can remain empty. The control packets can be implemented by a new IP option that need not be understood by regular routers since these packets never leave the access network.

5.3 Node Configuration

In a Cellular IP network, none of the nodes needs a network-level view or topology information of the system, nor do they need to know their position in the network. Nodes do not have to be configured: Cellular IP is a plug-and-play solution which also allows for easy recovery after a node failure. However, a limited amount of routing information must be available at each node. A node must know which of its ports to use to route packets toward the GW. A simple way to maintain this information is to let the GW periodically broadcast a beacon message and let nodes record the port they last received the beacon through. In the description of the algorithms, this port will be referred to as the "uplink-port". In addition, the node must know which of its ports are connected to a uplink-port of another node.[1] These ports will be called "downlink-ports". It is important that the routing information stored in nodes be consistent in the sense that a route, free of loops, always exists from any node to the GW.

While frequent modifications of these routes can degrade network performance, rerouting after a failure or congestion does not jeopardize performance and can be used to guarantee fault tolerance when Cellular IP operates over a meshed network.

5.4 Node Algorithms

Nodes that have neither PCs, nor RCs simply forward all packets arrived through a downlink-port to their uplink-port and forward all packets arrived through the uplink-port to all their downlink-ports. A node with a PC must monitor all packets arriving from its downlink-ports. Apart from forwarding them to the uplink-port, it reads the source address and uses it to update the PC, as follows. If there is a mapping for the sending mobile host to the port through which the packet arrived, the timer for this mapping is reset. If this mapping does not exist, it is created and the timer is initiated. When a paging packet arrives through the uplink-port, the PC is checked. If valid mapping(s) exist(s) for the destination, the packet is forwarded to the mapped port(s). If there is no mapping, the packet is discarded.

The operation of a RC node differs from this in two aspects. First, as discussed in Section 4.3, paging-update packets do not update the RC. Second, instead of paging packets, the RC routes data packets arriving from the uplink-port.

A node can, of course, have a PC and an RC at the same time in which case it runs both algorithms. Nodes without PC forward all paging packets to all downlink-ports and nodes without RC forward all data packets arriving through the uplink-port to all downlink-ports. We outline the algorithms that run on nodes using pseudo-code in the Appendix.

5.5 Mobile Host Algorithm

Mobile hosts can be modelled as a simple two-state state machine as illustrated in Figure 8. It is important to note that though the names of the two states may suggest that they refer to whether or not the host is transmitting data, they are more related to incoming data. The host should be in "active" state when it expects incoming data and in "idle" state otherwise. In most cases data will be expected when there are also data or acknowledment packets to transmit which justifies the names of the two states.

In idle state, the host sends paging-update packets to the base station having the best signal quality with a system-specific constant inter-arrival time. Depending on the implementation, to ensure safe reachability, it can send paging-update packets to several base stations within reach. Regardless of the sending period, a paging-update packet should be sent to the new base station immediately after moving into a new cell.

[1] This information is needed to avoid loops when mobile-addressed packets are broadcast.

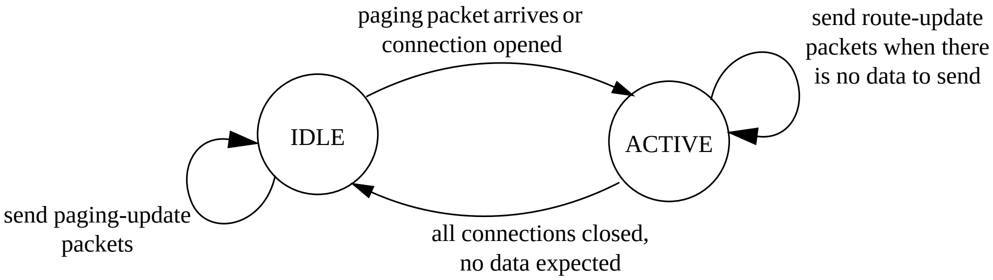

Figure 8 *Mobile host state machine*

The host moves to the "active" state if a paging packet arrives or a TCP connection is initiated or there is other reason to expect data. These reasons can be application specific and it is probably useful to let applications trigger the wireless interface to move to this state. However, this is not indispensable, and failure to move to the "active" state results in no other consequence, than an extra paging delay when the first data packets arrive addressed to the mobile host.

Once in "active" state, the mobile sends route-update packets periodically, with a system specific constant inter-arrival time that is likely to differ from the inter-arrival time of paging-update packets. (See Section 6.1 on timers.) Again, depending on the implementation, the host may be allowed to send route-update packets to more than one base stations in parallel, and it should in every case send one to the new base station immediately after a handoff. When the host has data packet to send, the sending of route-update packets is suspended.

The host can return to idle state if there are no open connections and there is no other reason to expect data. It is probably an efficient solution to assign a timer to the "active" state and return to idle state when the timer elapses without incoming data.

6 PROTOCOL IMPLEMENTATION ISSUES

Cellular IP defines a general framework that gives large freedom to a network operator to adjust the system to meet local characteristics. We are currently implementing the first experimental Cellular IP network that will serve as a testbed for studies on system performance and on its dependence on system settings.

6.1 Timers

Timers play an important role in Cellular IP efficiency. While the accuracy of timers is not crucial and hence synchronization can be omitted, the values of PC and RC timeouts, and the repetition rates of paging-update and route-update packets must be carefully selected for high performance. For both timeouts, a higher value results in less frequent control messages, but it extends the validity of unused paths.

If after migration the host always sends an extra paging-update packet to the new base station then the frequency of paging-update packets can be low without jeopardizing reachability. It is then ideal to set this frequency approximately equal to the migration frequency. To maintain the PC mapping even if some consecutive paging-update packets get lost, the PC timeout should be a few paging-update packets' time. With this setting, the load imposed by paging-update packets will be comparable to what explicit migration signalling would impose, while a mobile will usually be paged in just a few cells. The paging-update packet rate is hence the primary means of tuning Cellular IP mobility management according to a system's cell size and average user speed.

The performance is also sensitive to the RC timeout. This should be set so that a mobile in soft handoff receives a few consecutive packets from both base stations to ensure seamless handoff. A longer RC timeout should be avoided to avoid waste of resources. The route-update packet repetition rate should be such that a few route-update packets arrive during a RC timeout to maintain the mappings even if some route-update packets get lost. It is worth noting that in most communication sessions there will also be data packets or acknowledgments transmitted by the mobile host, hence route-update packets will not always be needed.

6.2 Caches

Another way to tune the network to local characteristics is its populating with Paging and Routing Caches. While the protocol supports nodes without RC, it is likely that for efficient utilization of the network's resources, each node will have a RC. PCs, which are considerably larger data bases, may be put in a few selected nodes only. Since nodes without PC simply broadcast paging packets, the paging process remains operational even with few PCs. More PCs result in less paging load in exchange for increased hardware cost.

The operator will typically place PCs in nodes with many leaf ports and in those that interconnect large disjoint areas of the network. Because the amount of routing is proportional to the number of data sessions and not to the aggregate traffic, adjusting the density of PCs is the way to tune the network to traffic characteristics. A larger number of PCs should be created if traffic mainly consists of frequent short bursts and less if long continuous data transmissions are dominant. The GW, however, should always contain a PC to avoid queueing packets for and paging hosts that are currently not attached to this access network.

Though in the description of the protocol this was not highlighted, it is possible to have nodes that have PCs for some of their downlink-ports but not for all. This may especially make sense in the case of base stations that should have a PC for their air interface ports, while their position in the network may not always justify configuring a PC. These nodes will behave as PC nodes for the ports they maintain a cache for, but appear as non-PC nodes for other leaf ports.

6.3 Paging

Depending on the network size, the time to page a mobile host may sometimes be long. Further studies are needed to see whether this delay is acceptable for most IP applications. If the paging delay is too high, it is possible to use the first data packet(s) as paging packets instead of queueing them in the GW and generating small paging packets. This solution also simplifies the GW implementations, though it results in a higher load in the network because data packets are larger than paging packets. More sophisticated implementations may choose between the two solutions depending on the size and payload type of packets.

7 CONCLUSIONS

In this paper we have discussed some limitations of existing mobile host protocols (e.g., Mobile IP), which become inefficient in wireless environments where hosts are highly mobile. We have presented a detailed discussion of a new protocol called Cellular IP which addresses these technical barriers. The Cellular IP architecture relies on the separation of local mobility from wide area mobility. We argued that while Mobile IP can efficiently support wide area mobility in the global Internet backbone, local mobility imposes special requirements not taken into account in the design and deployment of Mobile IP. We identified a set of key requirements, namely easy global migration, cheap passive connectivity, flexible handoff support, efficient location management and simple memoryless mobile hosts as motivating factors in our design. Cellular IP is optimized for wireless access networks and has been designed to satisfy these key requirements. Two further advantages of the protocol are its simplicity and robustness. A Cellular IP network scales well, using the same simple low-cost nodes for small indoor systems to metropolitan or large rural areas. The simplicity of a Cellular IP node and the capability of smooth interworking with Mobile IP eases the introduction of Cellular IP making it backward compatible, while the network can be easily extended in an incremental way. We are currently implementing the first experimental Cellular IP network that will serve as a testbed for performance studies. Results of these studies will be the subject of a future publication.

ACKNOWLEDGMENT

The author is grateful to Andrew T. Campbell for his invaluable help of reviewing and commenting earlier versions of this paper.

REFERENCES

[1] Fumio Teraoka, Yasuhiko Yokote, Mario Tokoro, “A Network Architecture Providing Host Migration Transparency, “ *Proc. ACM SIGCOMM’91*, pp. 209-220, September 1991.

[2] John Ioannidis, Dan Duchamp, Gerald Q. Maguire Jr., “IP-Based Protocols for Mobile Interworking,” *Proc ACM SIGCOMM’91*, pp. 235-245, September 1991.

[3] Charles Perkins, Andrew Myles, David B. Johnson, “IHMP: A Mobile Host Protocol for the Internet,” *Computer Networks and ISDN Systems*, 27(3), December 1994.

[4] Ramon Caceres, Venkata N. Padmanabhan, “Fast and Scalable Handoffs for Wireless Internetworks,” *Proc. ACM Conference on Mobile Computing and Networking* (Mobicom’96), pp. 56-66, 1996.

[5] Andrew Myles, David Skellern, “Comparing Four IP Based Mobile Host Protocols,” *Computer Networks and ISDN Systems*, pp. 349-356, November 1993.

[6] Pravin Bhagwat, Charles Perkins, Satish Tripathi, “Network Layer Mobility: an Architecture and Survey,” *IEEE Personal Communications Magazine*, Vol. 3, No. 3, pp. 54-64, June 1996.

[7] Charles Perkins, editor, “IP Mobility Support,” Internet RFC 2002, October 1996.

[8] David B. Johnson, Charles Perkins, “Route Optimization in Mobile IP,“ Internet Draft, draft-ietf-mobileip-optim-07.txt, November 1998, Work in Progress.

[9] David B. Johnson, Charles Perkins, “Mobility Support in IPv6,” Internet Draft, draft-ietf-mobileip-ipv6-07.txt, November 1998, Work in Progress.

[10] M. Mouly, M-B. Pautet, “The GSM System for Mobile Communications,” published by the authors, ISBN 2-9507190-0-7, 1992.

[11] Götz Brasche, Bernhard Walke, “Concepts, Services and Protocols of the New GSM Phase 2+ General Packet Radio Service,” *IEEE Communications Magazine*, pp. 94-104, August 1997.

[12] Jayanth Mysore, Vaduvur Bharghavan, “A New Multicasting-based Architecture for Internet Host Mobility,” *Proc. ACM Conference on Mobile Computing and Networking* (Mobicom’97), Budapest, October 1997.

[13] *IEEE Personal Communications* special issue on IMT-2000, Vol. 4, No. 4, August 1997.

[14] Göran Eneroth, Martin Johnsson, “ATM Transport in Cellular Networks,” *International Switching Symposium* (ISS’97), Toronto, September 1997.

[15] Mark Stemm, Randy H. Katz, “Vertical Handoffs in Wireless Overlay Networks,” *ACM Mobile Networking* (MONET), Special Issue on Mobile Networking in the Internet, Summer 1998.

[16] D. Clarc, J. Wroclawski, “An Approach to Service Allocation in the Internet,” Internet Draft, draft-clarc-diff-svc-alloc-00.txt, July 1997, Work in Progress.

APPENDIX

Packet in from uplink-port

```
mobile-id   <- identifier of destination mobile host

packet-type <- data/paging

IF (packet-type is data AND node has RC AND mobile-id has RC mapping) THEN
forward packet to mapped ports END IF
```

```
IF (packet-type is data AND node has RC AND mobile-id has no RC mapping) THEN
    IF (node is GW) THEN queue packet, create and send paging packet for mobile-id ELSE discard packet END IF
END IF
IF (packet-type is data AND node has no RC) THEN forward packet to all downlink-ports END IF
IF (packet-type is paging AND node has PC) THEN forward packet to mapped ports, discard if none END IF
IF (packet-type is paging AND node has no PC) THEN forward packet to all downlink-ports END IF
STOP
```

Packet in from downlink-port

```
mobile-id <- identifier of sending mobile host
port-id <- identifier of the downlink-port the packet came through
packet-type <- data/paging-update/route-update
IF (node has PC AND mobile-id has PC mapping to port-id) THEN reset timer of this mapping to PC-timeout END IF
IF (node has PC AND mobile-id has no PC mapping to port-id) THEN create mapping and set its timer to PC-timeout END IF
IF (packet-type is data or route-update) THEN
    IF (node has RC AND mobile-id has RC mapping to port-id) THEN reset timer of this mapping to RC-timeout END IF
    IF (node has RC AND mobile-id has no RC mapping to port-id) THEN create mapping and set its timer to RC-timeout END IF
        IF (node is GW AND there are data packets queued with destination=mobile-id) THEN dequeue these packets, forward them to port-id END IF
    END IF (node has RC)
END IF (packet-type is data or route-update)
IF (node is not GW OR packet-type is data) THEN forward packet to uplink-port END IF
STOP
```