Abstract

Wireless access to Internet services will become typical, rather than the exception as it is today. Such a vision presents great demands on mobile networks. Mobile IP represents a simple and scalable global mobility solution but lacks the support for fast handoff control and paging found in cellular telephony networks. In contrast, second- and third-generation cellular systems offer seamless mobility support but are built on complex and costly connection-oriented networking infrastructure that lacks the inherent flexibility, robustness, and scalability found in IP networks. In this article we present Cellular IP, a micro-mobility protocol that provides seamless mobility support in limited geographical areas. Cellular IP, which incorporates a number of important cellular system design principles such as paging in support of passive connectivity, is built on a foundation of IP forwarding, minimal signaling, and soft-state location management. We discuss the design, implementation, and evaluation of a Cellular IP testbed developed at Columbia University over the past several years. Built on a simple, low-cost, plug-and-play systems paradigm, Cellular IP software enables the construction of arbitrary-sized access networks scaling from picocellular to metropolitan area networks. The source code for Cellular IP is freely available from the Web (comet.columbia.edu/cellularip).

Design, Implementation, and Evaluation of Cellular IP

ANDREW T. CAMPBELL, JAVIER GOMEZ, SANGHYO KIM, ANDRÁS G. VALKÓ, AND CHIEH-YIH WAN, COLUMBIA UNIVERSITY, NEW YORK

ZOLTÁN R. TURÁNYI, TECHNICAL UNIVERSITY OF BUDAPEST

The development of commodity-based palmtop devices with built-in high-speed packet radio access to the Internet will have a major impact on the way we communicate. Large numbers of mobile users equipped with wireless IP-enabled communicators will have access to a wide array of Web-based mobile multimedia services. Future wireless network infrastructure will have to support a wide variety of users, applications, and access needs. While access rates for traditional services such as mobile telephony are well understood, requiring small and fixed-size bandwidth allocation, wireless data rates may vary by a number of orders of magnitude. Regulatory and environmental factors influence access speed, service quality, and pricing policy. For example, outdoor wireless access will remain at lower speeds and premium rates compared to unlicensed indoor communications.

High-speed access can be achieved by using smaller and smaller cell sizes, resulting in coverage areas with a larger number of base stations. One can imagine a scenario where each person's office has its own access point offering tens to hundreds of megabits per second of wireless access. These types of picocellular environments call for simple, low-cost wireless infrastructure that ultimately must compete with wireline LAN service quality, costs, security, and plug-and-play scalability. Mobile users will expect the same level of service quality as wireline users. That will translate to high-speed access with *seamless mobility*, which we define as the ability of the network to support fast handoff between base stations with low delay and minimum or zero packet loss. As base station density increases, however, so will handoff rates. This places significant demands on future mobile network architecture, protocols, and services to support seamless mobility.

Recent initiatives to add mobility to the Internet have mostly focused on the issue of address translation [1] through the introduction of location directories and address translation agents. The problem of address translation is fundamental to global Internet mobility and comes from the hierarchical nature of IP addressing. In Mobile IP [2] packets addressed to a mobile host are delivered using regular IP routing to a temporary address assigned to a mobile host at its actual point of attachment. This approach results in a simple and scalable scheme that offers global mobility. Mobile IP is not appropriate, however, for seamless mobility because after each migration a local address must be obtained and communicated to a possibly distant location directory or home agent (HA). We believe that support for seamless mobility will be needed in order to provide good service quality to mobile users, particularly in picocellular environments where the rate of handoff and associated signaling load grows rapidly.

Network support for seamless mobility was not a primary design consideration when Mobile IP was first defined in the early '90s. More recently the Mobile IP Working Group has been addressing this issue. With frequent handoff *micro-mobility protocols* have been proposed [3–6] to handle local movement of mobile hosts without interaction with the Mobile-IP-enabled Internet. This has the benefit of reducing delay and packet loss during handoff, and eliminating registration between mobile hosts and distant HAs when mobile hosts remain inside their local coverage areas. Eliminating registration in this manner reduces the signaling load experienced by the core network in support of mobility. Reducing signaling in this manner is necessary for the wireless Internet to scale to support very large volumes of wireless subscribers.

We envision a wireless Internet with many hundreds of millions of wireless subscribers. As in the case of the cellular phone, we imagine that wireless IP communicators will be turned on around the clock, ready to receive or initiate services. In fact, the vast majority of subscribers will not be actively communicating most of the time. Rather, wireless IP communicators will be switched on, ready for service, constantly reachable by the wireless Internet. In essence, mobile hosts will be in an idle state but passively connected to the network infrastructure. As

1070-9916/00/$10.00 © 2000 IEEE

in the case of the mobile telephony network, it will be sufficient for the wireless Internet only to know the approximate location of its population of idle users. The exact location of idle mobile hosts only becomes important when data needs to be forwarded to them, in which case the network needs to be able to efficiently search and find these users in a scalable and timely manner. In cellular telephony systems this process is called *paging*.

As the number of mobile subscribers grows, so does the need to provide efficient location tracking in support of idle users and paging in support of active communications. In order to achieve scalable location management, the wireless Internet needs to handle the location tracking of active and idle mobile hosts independently. Support for *passive connectivity* balances a number of important design considerations. For example, only keeping the approximate location information of idle users requires significantly less signaling, and thus reduces the load over the air interface and core network. Reducing signaling over the air interfaces also has the benefit of preserving the power reserves of mobile hosts.

Currently, Mobile IP does not support the notion of seamless mobility, passive connectivity, or paging. We argue that the future wireless Internet will need to support these requirements in order to deliver service quality, minimize signaling, and scale to support hundreds of millions of subscribers. In this article we present an evaluation of Cellular IP [4, 7], a micro-mobility protocol that is optimized to provide local access to a Mobile-IP-enabled Internet in support of fast-moving wireless hosts. Cellular IP incorporates a number of important design features present in cellular networks but remains firmly based on IP design principles. The protocol is specifically designed to support seamless mobility, passive connectivity, and paging. Cellular IP access networks can be constructed in a plug-and-play manner scaling from picocellular to metropolitan area networks. Distributed location management, routing, and handoff algorithms lend themselves to a simple, efficient, and low-cost software implementation for host mobility requiring no new packet formats, encapsulation, or address space allocation beyond that present in IP.

This article is structured as follows. We discuss the related work in the field. We provide an overview of Cellular IP and discuss the design choices made. The performance of the protocol is evaluated using measurements from an experimental testbed developed at Columbia University. Finally, we present some concluding remarks.

Related Work

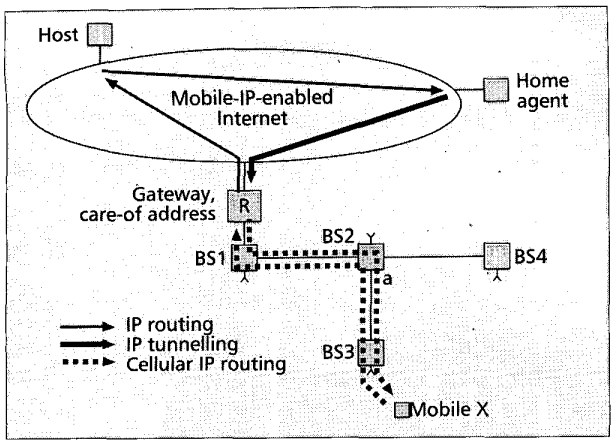
A number of micro-mobility solutions have been discussed in the literature. In [3] a hierarchical mobility model is described where independent wireless access networks interwork with a global mobility protocol. Address translation and security are functions of the global mobility solution. In contrast, wireless access networks provide mechanisms for the support of local location management and mobility. In [6, 8] Mobile IP is extended by arranging foreign agents in a hierarchy. The top of the hierarchy is rooted at the edge of the access network and is defined by the care-of address registered with HAs. Upon reception of a packet, the foreign agent at the top of the hierarchy interacts with a local database to determine to which lower-level foreign agent located in the access network to forward the packet. This procedure may be repeated, depending on the depth of the routing hierarchy. Similar ideas are adopted in the case of campus and domain foreign agents [3] and local registration schemes [9].

We observe that Cellular IP and the protocols discussed above employ per-mobile-host states and hop-by-hop routing to achieve fast handoff control. These hierarchical mobility proposals do not, however, support the notion of passive connectivity with its separation of active and idle users, as does Cellular IP. In these proposals, a foreign agent maintains database entries for each mobile host in its region, having to search a potentially large database in order to route each packet. In contrast, Cellular IP routing cache only contains entries for mobile hosts that have recently transmitted packets. This reduces the search time and increases protocol scalability. Other differences exist. Hierarchical foreign agent schemes operate on top of IP, whereas Cellular IP is itself a layer three routing protocol; that is, Cellular IP replaces IP routing in the wireless access network but without modifying the IP packet format and forwarding mechanism. To increase efficiency, location management is integrated with routing in Cellular IP access networks. The per-host location information stored in Cellular IP nodes is not a network address. Rather, per-host location state represents the next hop route to forward packets to for a given mobile host. Such an integrated approach simplifies both routing and location management in wireless access networks.

In [10] off-the-shelf Ethernet switches and wireless LAN cards are used to build wireless access networks. The learning feature of Ethernet switches is used for location management. Data frames transmitted by mobile hosts are used to establish and refresh location information inside the access network. Although this approach of using Ethernet switches for location management results in simple, cheap, and efficient access networks, the concept is hard to extend with desirable features, such as smooth and secure handoff or paging. Cellular IP uses data packets to refresh location management state and can operate at layer two or three. However, mobility support is built into Cellular IP nodes.

Support for seamless mobility, passive connectivity, and paging is fundamental to improving scalability, minimizing power consumption, and delivering suitable service quality to mobile hosts. Few solutions, however, support these features [11]. One protocol that supports seamless mobility, passive connectivity, and paging is Hawaii [5]. In contrast to Cellular IP nodes, which preserve the simplicity of the Ethernet switch solution discussed above, Hawaii nodes are IP routers. It is interesting to note that low-cost layer two switches can be used to build Cellular IP access networks supporting tens of thousands of mobile hosts [12]. We believe that this approach becomes increasingly important when constructing low-cost picocellular infrastructure. The use of an all-IP-based router solution for picocellular networks may become prohibitively expensive. This motivates the need to have a layer two and three solution to micro-mobility. Hawaii assumes that an intradomain routing protocol is operating in the access network, allowing each node to have routes to other nodes. This routing information is used to exchange explicit signaling messages, and forward packets between old and new access points during handoff. The use of explicit signaling messages is limited in Cellular IP, which uses the IP data packets to convey location and paging information.

Different proposals have different scaling properties. The base stations associated with the original Columbia protocol [13] represent radio-enabled routers operating in campus area networks. Base stations broadcast search messages among each other in order to determine the location of mobile hosts. By tunneling packets between base stations, the Columbia scheme effectively creates a mobile overlay network on top of the wired campus network. This protocol works well for small numbers of mobile hosts, but encounters scalability problems due to the nature of the broadcast search algorithm used. The local mobility protocol proposed by [3] uses workstations as base stations, and hence is more appropriate in networks with small cells. However, this protocol is similar to commercially available solutions [14] in the respect that it only provides mobility within the area covered by a local area network. A key design requirement



■ **Figure 1.** *Cellular IP access network.*

of Cellular IP is its capability to scale from local to metropolitan area networks. Cellular IP can be deployed across a number of different installations, including office, campus, and wireless Internet service provider (ISP) coverage areas [12].

Many of the proposals discussed above are capable of minimizing service disruption during handoff. In [15] an IP multicasting technique is used to support fast handoff. Here mobile hosts are identified by multicast IP addresses. Base stations are capable of joining a mobile host's multicast group. This includes the base station the mobile host is currently connected to as well as others which it may move to after handoff. In the latter case, packets are delivered to the new base station even before the host has migrated. In Hawaii, seamless handoff is achieved by exchanging a series of signaling messages between the old and new base stations. This facilitates the forwarding of packets from the old base station to the new one during handoff. Both of these approaches require nodes to either be multicast capable routers or process signaling messages. Cellular IP handoff aims at simplicity, eliminating the reliance on multicast and minimizing explicit signaling.

Protocol Overview

As the name suggests, Cellular IP inherits cellular principles for mobility management such as passive connectivity, paging, and fast handoff control, but implements them around the IP paradigm. Cellular IP access networks require minimal configuration (e.g., similar to switched Ethernet LANs), thereby easing the deployment and management of wireless access networks. An important concept in Cellular IP design is simplicity and minimal use of explicit signaling, enabling low-cost implementation of the protocol. In what follows, we present an overview of Cellular IP access networks and discuss support for routing, handoff, paging, and security in these networks. For a full discussion of the protocol and its specification see [4, 7], respectively.

The Network Model

The universal component of Cellular IP access networks is the *base station* which serves as a wireless access point and router of IP packets while performing all mobility-related functions. Base stations are built on a regular IP forwarding engine with the exception that IP routing is replaced by Cellular IP routing and location management. Cellular IP access networks are connected to the Internet via *gateway* routers. Mobile hosts attached to an access network use the IP address of the gateway as their Mobile IP care-of address. Figure 1 illustrates the path taken by packets addressed to a mobile host. Assuming Mobile IPv4 [2] and no route optimization [16], packets are first routed to the host's HA and then tunneled to the gateway. The gateway detunnels packets and forwards them toward a base station. Inside a Cellular IP network, mobile hosts are identified by their home address, and data packets are routed without tunneling or address conversion. The Cellular IP routing protocol ensures that packets are delivered to the host's actual location. Packets transmitted by mobile hosts are first routed toward the gateway and from there on to the Internet.

In Cellular IP, location management and handoff support are integrated with routing. To minimize control messaging, regular data packets transmitted by mobile hosts are used to refresh host location information. *Uplink* packets are routed from a mobile host to the gateway on a hop-by-hop basis. The path taken by these packets is cached by all intermediate base stations. To route *downlink* packets addressed to a mobile host, the path used by recently transmitted packets from the mobile host is reversed. When the mobile host has no data to transmit, it sends small, special IP packets toward the gateway to maintain its downlink routing state. Following the principle of passive connectivity, mobile hosts that have not received packets for some period of time allow their downlink routes to be cleared from the cache as dictated by a soft state timer. *Paging* is used to route packets to idle mobile hosts in a Cellular IP access network.

Routing

The Cellular IP gateway periodically broadcasts a beacon packet that is flooded in the access network. Base stations record the neighbor they last received this beacon from and use it to route packets toward the gateway. All packets transmitted by mobile hosts, regardless of their destination address, are routed toward the gateway using these routes.

As these packets pass each node en route to the gateway, their route information is recorded as follows. Each base station maintains a *routing cache*. When a data packet originated by a mobile host enters a base station, the local routing cache stores the IP address of the source mobile host and the neighbor from which the packet entered the node. In the scenario illustrated in Fig. 1, data packets are transmitted by a mobile host with source IP address **X**. In the routing cache of BS2 this is indicated by a *mapping* (X, BS3). This soft-state mapping remains valid for a system-specific time called *route-timeout*. Data packets are used to maintain and refresh mappings. As long as mobile host X regularly sends data packets, base stations along the path between the mobile's actual point of attachment and the gateway will maintain valid routing cache mappings, forming a soft-state path between the mobile host and gateway node. Packets addressed to mobile host X are routed on a hop-by-hop basis using this established routing cache.

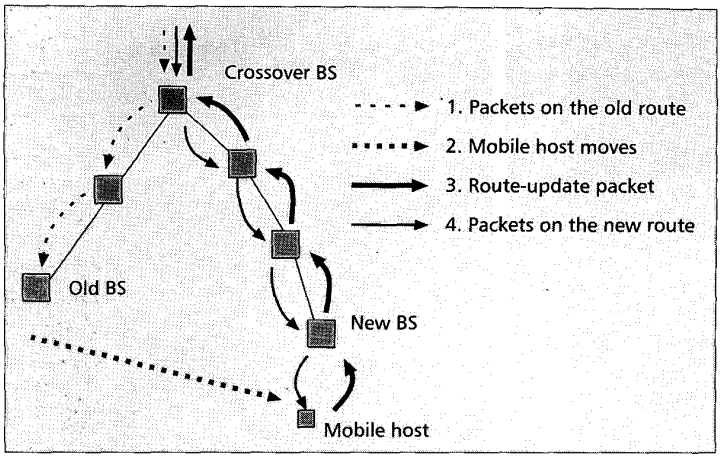
A mobile host may sometimes wish to maintain its routing cache mappings even though it is not regularly transmitting data packets. A typical example of this is when a mobile host receives a UDP stream of packets on the downlink but has no data to transmit on the uplink. To keep its routing cache mappings valid, the mobile host transmits *route-update packets* on the uplink at regular intervals called *route-update time*. These packets are special ICMP packets addressed to the gateway. Route-update packets update routing cache mappings as is the case with normal data packets. However, route-update messages do not leave the Cellular IP access network.

Handoff

Cellular IP supports two types of handoff scheme. Cellular IP *hard handoff* is based on a simple approach that trades off some packet loss for minimizing handoff signaling rather than trying to guarantee zero packet loss. Cellular IP *semisoft handoff* exploits the notion that some mobile hosts can simultaneously receive packets from the new and old base stations

during handoff. Semisoft handoff minimizes packet loss, providing improved TCP and UDP performance over hard handoff.

Hard Handoff — Mobile hosts listen to beacons transmitted by base stations and initiate handoff based on signal strength measurements. To perform a handoff, a mobile host tunes its radio to a new base station and sends a route-update packet. The route-update message creates routing cache mappings en route to the gateway configuring the downlink route cache to point toward the new base station. Handoff latency is the time that elapses between handoff initiation and the arrival of the first packet along the new route. In the case of hard handoff this is equal to the round-trip time between the mobile host and the *crossover base station* as illustrated in Fig. 2. We define the crossover base station as the common branch node between the old and new base stations, an example of which is illustrated in the figure. In the worst case the crossover point is the gateway. During this interval, downlink packets may be lost. Mappings associated with the old base station are not cleared when handoff is initiated. Rather, mappings between the crossover node and the old base station timeout and are removed. No packets are transmitted along the old path once the route-update message has created a new mapping at the crossover base station that points toward the new base station.



■ **Figure 2.** *Cellular IP handoff.*

Although packets may get lost during a hard handoff, the time taken to redirect packets to the new point of attachment is shorter than that in Mobile IP. This is due to the fact that only a local node has to be notified rather than a possibly distant HA in the case of Mobile IP.

There are several ways to reduce packet loss during handoff. One approach relies on interaction between the old and new base stations [5] during handoff. In this case the new base station notifies the old base station of the pending handoff. Packets that arrive at the old base station after notification of handoff are forwarded to the new base station and onto the mobile host. In contrast, packets that arrive at the old base station before notification is complete will be lost. If the notification time (i.e., the round-trip time between the new and old base stations) is not smaller than handoff duration (i.e., the round-trip time between the new and cross-over base stations), this approach does not significantly improve handoff. An additional cost of these schemes is that communications, signaling, and information state exchange are required between base stations for this approach to work. To preserve the simplicity of hard handoff, Cellular IP employs a different approach to counter the problem of packet loss.

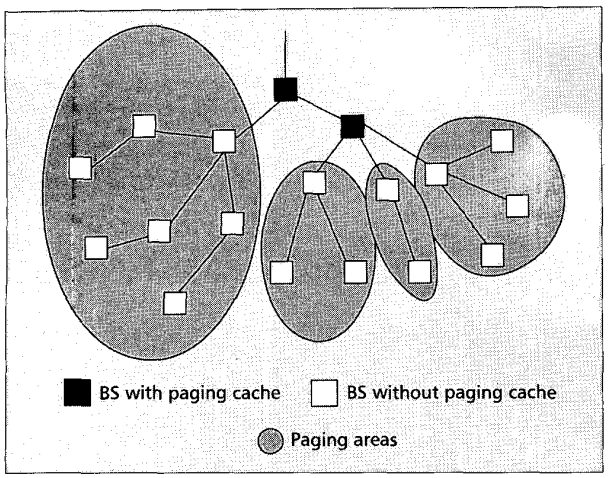
Semisoft Handoff — After hard handoff, the path to the old base station remains in place until the soft-state cache mappings time out. We leverage this feature to support a new handoff service called semisoft handoff that improves handoff performance while maintaining the lightweight nature of the base Cellular IP protocol. Semisoft handoff calls for one temporary state variable to be added to the protocol running in the mobile hosts and base stations.

Semi-soft handoff scales well for large numbers of mobile hosts and frequent handoff, and comprises two architectural components. First, in order to reduce handoff latency, the routing cache mappings associated with the new base station must be created before the actual handoff takes place. Before a mobile host hands off to a new access point, it sends a *semisoft packet* to the new base station and immediately returns to listening to the old base station.

The purpose of the semisoft packet is to establish new routing cache mappings between the crossover and new base stations. During this route establishment phase the mobile host is still connected to the old base station. After a *semisoft delay*, the mobile host performs a regular handoff. The semisoft delay can be an arbitrary value that is proportional to the mobile-to-gateway round-trip delay. This delay ensures that by the time the mobile host finally tunes its radio to the new base station, its downlink packets are being delivered through both the old and new base stations. We observe that downlink packets consume twice the amount of resources during this period. However, this period represents a short duration when one considers the complete semisoft handoff process.

While the semisoft packet ensures that mobile hosts continue to receive packets immediately after handoff, it does not, however, ensure smooth handoff between base stations. Depending on the network topology and traffic conditions, the time to transmit packets from the crossover point to the old and new base stations may differ, and the packet streams transmitted through the two base stations are typically unsynchronized. If the new base station is behind the old one, the mobile host will receive duplicate packets, which does not disrupt many applications. For example, TCP will not be forced into slow start due to the arrival of duplicate acknowledgments. If the new base station is ahead, packets will be missing from the stream received at the mobile host.

The second architectural component of semisoft handoff resolves this issue of the new base station getting ahead. The solution to this problem is based on the observation that perfect synchronization of packet streams is unnecessary. This condition can be eliminated by temporarily introducing a constant delay along the new path between the crossover and new base stations using a simple delay device mechanism. The device needs to provide sufficient delay to compensate, with high probability, for the time difference between the two streams traveling on the old and new paths. Optimally, the device delay should be located at the crossover base station. The crossover base station is aware that a semisoft handoff is in progress from the fact that a semisoft packet arrives from a mobile host that has mapping to another interface. Mappings created at crossover points by the reception of semisoft packets include a flag to indicate that downlink packets must pass through a delay device before being forwarded for transmission along the new path. After handoff is complete, the mobile host sends a data or route-update packet along the new path. These packets have the impact of clearing the flag causing all packets in the delay device to be forwarded to the mobile host. Base stations only need a small pool of delay buffers to resolve this issue. Packets that cannot sustain additional delay can be forwarded



■ **Figure 3.** *Paging areas.*

without passing through the delay device. This differentiation can be made on a per-packet basis, using, say, differentiated service or transport (e.g., TCP, UDP, or RTP) codepoints.

Paging

Typically, fixed hosts connected to the Internet (e.g., desktop computers) remain online for extended periods of time, even though most of the time they do not communicate. Being always connected in this manner results in being reachable around the clock with instant access to Internet resources. Mobile subscribers connected to the wireless Internet will expect similar service. However, in the case of mobile hosts maintaining location information to support being continuously reachable would require frequent location updates which would consume precious bandwidth and battery power.

Cellular systems employ the notion of passive connectivity to reduce the power consumption of idle mobile hosts. Base stations are geographically grouped into *paging areas*, as illustrated. When there is no call ongoing, mobile hosts only need to report their position to the network if they move between paging areas. This makes location update and handoff support for idle hosts unnecessary. When an incoming call is detected at the gateway, a paging message is transmitted to the mobile host's current paging area to establish the call. The mobile node informs the infrastructure of its location as a result of the paging process and transition to active mode to take the call.

While the definition of an idle mobile device is well understood in the context of cellular systems, which are connection-oriented in nature, its meaning in IP-based mobile networks is unclear. Cellular IP defines an idle mobile host as one that has not transmitted packets for a system-specific *active-state-timeout*. Due to lack of updates, the soft-state routing cache mappings of idle mobile hosts will time out in a fully distributed manner. In order to remain reachable, mobile hosts transmit *paging-update* packets at regular intervals defined by a *paging-update-time*. A paging-update packet is an ICMP packet, which is addressed to the gateway and distinguished from route-update packets by its type parameter value. Mobile hosts send paging-update packets to base stations that have better signal quality. As in the case of data and route-update packets, paging-update packets are routed toward the gateway on a hop-by-hop basis. Base stations may optionally maintain *paging cache*. Paging cache has the same format and operation as routing cache with the following exceptions. Paging cache mappings have a longer timeout period called *paging-timeout*; hence, a longer interval exists between consecutive paging-update packets. In addition, any packet sent by mobile hosts, including route-update packets, can update paging cache. However, paging-update packets cannot update routing cache. This results in idle mobile hosts having mappings in the paging cache but not in the routing cache. In contrast, active mobile hosts will have mappings in both routing and paging caches.

Packets addressed to a mobile host are normally routed by routing cache mappings. Paging occurs when a packet is addressed to an idle mobile host, and the gateway or base stations find no valid routing cache mapping for the destination. If the base station has no paging cache, it will forward the packet to all of its interfaces except the one the packet came through. Cellular IP has no explicit paging control message. Rather, the first data packet that arrives at the gateway forms an implicit paging message which is forwarded in the access network. Paging cache is used to avoid broadcast search procedures. Base stations that have paging cache will only forward a paging packet if the destination has a valid paging cache mapping. In this case the paging message is only forwarded to the mapped interface. If there is no paging cache in an access network, the first packet addressed to an idle mobile will be broadcast, increasing the load on the access network.

The network operator can limit paging load in exchange for memory and processing cost by using paging cache in the access network. By placing paging cache in base stations, paging areas can be defined as needed. An operator can construct paging areas and determine which nodes in the access network should support paging cache and which should not. For example, paging cache could be located at the gateway only or at the majority of the base stations in the access network. The construction of paging areas (i.e., the number of base stations that make up a paging area) and the distribution of paging cache within a paging area (i.e., which nodes do and do not have paging cache) is a configuration issue, some examples of which are illustrated in Fig. 3.

In the case of Cellular IP, a paging area identifier is broadcast as part of beacon messages. Idle mobile hosts will only transmit paging-update packets when they move between paging areas. An idle mobile host that receives a paging packet transitions from idle to active state and immediately transmits a route-update packet toward the gateway. This ensures that routing cache mappings are quickly established, limiting any further paging in the location area.

Security

Cellular IP has been designed to support seamless and secure handoff. Mobile systems are open to a number of security problems that do not exist in their stationary counterparts. In a fixed network, the prefix of a subnet is usually configured manually, and the location of the prefix is communicated between routers that either have some form of inherent trust model or use a secure protocol. This makes it hard to impersonate someone. Mobile hosts, on the other hand, must update their location while moving. These location messages make impersonation possible unless properly secured. Wireless access networks compound these security problems because packets can be snooped over the air interface. Cellular IP faces impersonation and snooping attacks because it is wireless and mobile.

Cellular IP addresses these security issues. First, only authenticated packets can establish or change cache mappings in a Cellular IP access network. By authenticating paging and routing update control messages, malicious users are prevented from capturing traffic destined for mobile hosts. In Cellular IP access networks, only control packets are authenticated. In this case data packets are not authenticated, which would be costly in terms of transport performance. Control messages establish and change existing mappings. In contrast, data packets can only refresh existing mappings. Active mobile hosts transmit route-update packets during handoff to create a new chain of soft-state cache mappings pointing to the new point of attachment.

In Cellular IP seamless handoff is of primary importance. Therefore, session keys used by mobile hosts to perform authentication must be promptly available at the new base station during handoff. Timeliness of the authentication process is critical in the case of micro-mobility due to the requirement of fast handoff control. In contrast, global mobility solutions may have broader requirements such as user identification, bilateral billing, and service provisioning agreements. These broader requirements outweigh the need to support fast handoff control where the scalability of the global authentication, authorization, and accounting (AAA) [14] system is of more importance than seamless handoff. One can envision, however, micro-mobility protocols that build on global AAA preferences by offering enhanced services (e.g., fast session key management) to aid seamless handoff.

During handoff, the new base station could hypothetically acquire a session key by contacting the old base station, the cross-over base station or some central key management server. In Cellular IP fast session key management operates as follows. Rather than defining new signaling, a special session key is used in Cellular IP access networks. Base stations can independently calculate session keys. This eliminates the need for signaling in support of session key management, which would inevitably add additional delay to the handoff process. The session key is a secure hash, which combines:

- The IP address of a mobile host (IP_{MH})
- A random value (R_{MH}) assigned to a mobile host when it first registers with an access network
- A network secret ($K_{network}$) known by all base stations within an access network

The session key is calculated using an MD5 hash function: ($K_{session}$ = MD5 (IP_{MH}, R_{MH}, $K_{network}$)

A session key is first calculated and transmitted to a mobile host when it first contacts the Cellular IP network during global mobility authentication and authorization. The random value R_{MH} is assigned to the mobile host at this point.

Control packets carry this random value(R_{MH}) together with their authentication information. A timestamp is used for replay protection. The session key is used to perform authentication. Base stations can quickly calculate the session key by combining the IP address and the random value found in the control packet with the network secret. Base stations can validate the authentication easily with the session key. The base stations perform the validation process without any further communication or pre-distributed subscription databases. This results in fast and secure handoff. To enhance security, the network key could be periodically replaced thereby triggering session key changes making brute force attacks more difficult.

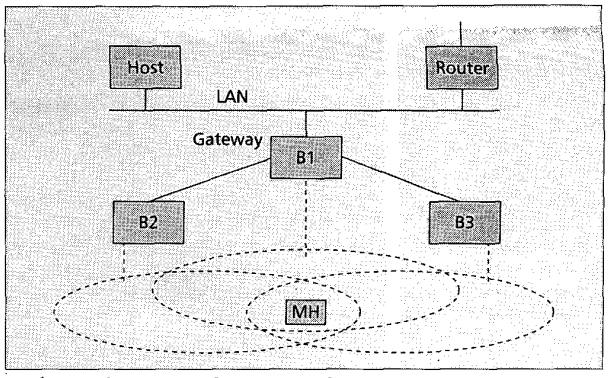
Evaluation

To evaluate Cellular IP performance we have built a testbed and designed a set of experiments to analyze the protocol. In what follows we describe our Cellular IP testbed and experimental results.

Testbed

The goal of the experiments is to evaluate the performance and scalability of the protocol. Cellular IP has been implemented and evaluated on FreeBSD 2.2.6 software platform. Note that other operating systems are supported including Windows and Linux. In this article we report and evaluate the FreeBSD version of the protocol. The Cellular IP base station and mobile software modules execute in user space and use the Berkeley Packet Filter's Packet Capture library (PCAP) [18] for processing and forwarding of IP packets.

The experimental results reported in this article are based on measurements taken from the Cellular IP testbed illustrated in Fig. 4. The access network consists of three base stations based on multihomed 300 MHz Pentium PCs hardware. One of the base stations also serves as a gateway router to the Mobile IP enabled Internet. Base stations are interconnected using 100 Mb/s full duplex links. Mobile hosts are 300 MHz Pentium PC notebook. Mobile hosts and base stations are equipped with 2 Mb/s WaveLAN 2.4 GHz radio interfaces. Note that the current software release of the protocol supports device drivers for a number of 11 Mb/s radios including IEEE 802.11 WaveLAN and Aironet radios. The 2.4 GHz WaveLAN radios can operate at eight different frequencies to avoid interference between adjacent cells. In the testbed the base stations are statically assigned frequencies while mobile hosts can dynamically change frequency to perform handoff. Throughout the experiments the mobile host shown in Fig. 4 was in an overlapping region of cells. For experimentation purposes a utility tool located on the mobile host was capable of periodically triggering handoff regardless of the measured signal quality. Handoff initiated by the utility tool is, however, identical to the Cellular IP mobile initiated handoff.



■ **Figure 4.** *The Cellular IP testbed.*

Handoff

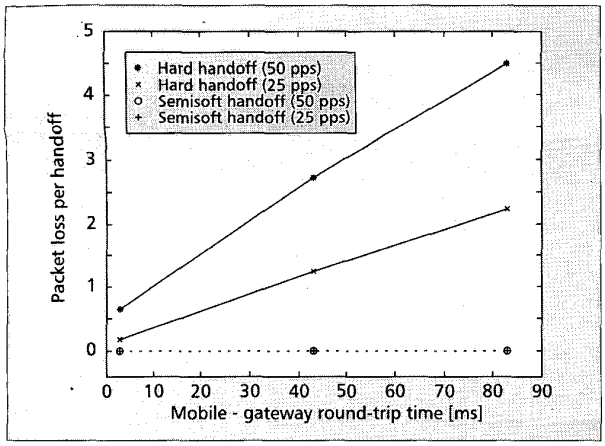
An important objective of this experiment is to analyze the performance of hard and semisoft handoff and investigate the impact of handoff on UDP and TCP performance. Here we measure the packet loss for hard and semisoft handoff, respectively.

UDP Performance — During this experiment the mobile host shown in Fig. 4 receives 100 byte UDP packets at rates of 25 and 50 packets/s while making periodic handoffs (driven by the utility tool) between base stations **B2** and **B3** every 5 s.

The measurement results are plotted in Fig. 5. Each point on the graph was obtained by averaging loss measurements over 50 consecutive handoffs. The solid lines in Fig. 5 show that hard handoff causes packet losses proportional to the round-trip time and to the downlink packet rate. Under these experimental conditions hard handoff results in at least 1 packet loss for small mobile to gateway round-trip delays and up to four packet losses for delays of 80 ms.

The dashed line in Fig. 5 represents the packet loss results from Cellular IP semisoft handoff.

The experimental conditions for semisoft and hard handoff are identical. In this experiment, a delay device buffers packets before they are forwarded along the new downlink path. Each downlink packet is inserted into the delay device at the cross-over base station **B1** until the arrival of the next downlink packet at which point the first packet is dequeued and forwarded toward the new base station. When the semisoft handoff is complete, the last packet is cleared from the buffer and is sent to the mobile host. Figure 5 illustrates that semisoft handoff eliminates packet loss. Note that buffering a



■ **Figure 5.** *UDP packet loss with handoff (pps: packets/s).*

single packet in the delay device is sufficient to eliminate loss even in the case of a large round-trip time where hard handoff results in the loss of up to four packets.

TCP Performance —In the next experiment, we study the impact of handoff performance on TCP Reno throughput. The mobile host performs handoff between **B2** and **B3** at fixed time intervals. We measure TCP throughput using `ttcp` by downloading 16 MBytes of data from a correspondent host to a mobile host. Each data point is an average of 6 independent measurements.

The TCP throughput to the mobile host performing hard handoff is shown by the solid curve in Fig. 6. The throughput measured at zero handoff frequency (i.e., no handoffs) is marginally lower than the 1.6 Mb/s achieved using standard IP routing in the same configuration. The difference between IP and Cellular IP forwarding is attributed to the fact that IP is implemented in the kernel and Cellular IP in user space. In addition, Cellular IP uses PCAP to forward packets which is not optimized for IP forwarding. We observe that the performance of TCP degrades as the hard handoff frequency increases due to packet loss. As the handoff rate increases TCP has less time to recover from loss. This force TCP to operate below its optimal operational point resulting in a significant drop in transport performance as the handoff rate approaches one per second. Note that a mobile handing off every second is an aggressive handoff rate.

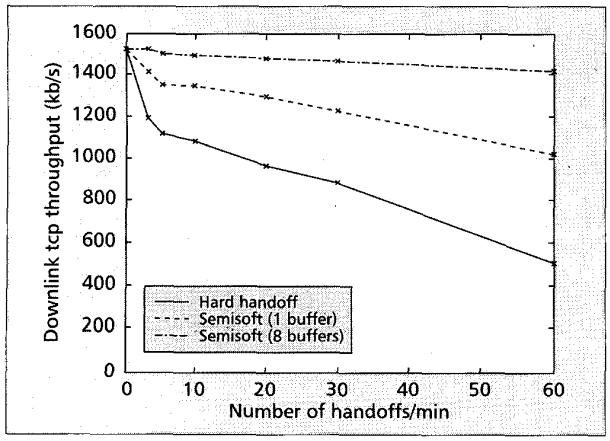
The next experiment investigates TCP improvement gains using semisoft handoff. The experimental conditions for the semisoft and hard handoff experiments are identical. The dashed curve for a 1-packet delay device in Fig. 6 shows the TCP throughput achieved by mobile host that performs semisoft handoff at an increasing rate. From the figure we can observe that semisoft handoff reduced packet loss and significantly improved the transport throughput in relation to the hard handoff scheme.

Unlike the semisoft handoff experiment for UDP traffic, packet loss is not entirely eliminated with TCP. This can be observed in the decline in the measured throughput as the handoff frequency increases. We attribute the lack of synchronization and subsequent loss to the single buffer delay device used. Buffering packets is tied to the packet inter-arrival time, which is both shorter and more irregular in TCP streams than in the case of the UDP experiment. To introduce sufficient delay, we configure the semisoft delay device to support an 8-packet circular buffer. In Fig. 6 the dash curve for the 8-packet delay device shows performance results associated with using the larger buffer. We observe from the graph that packet loss is eliminated at the higher handoff rates. A slight disturbance remains at handoff rates approaching one handoff per second due to the transmission delay variations encountered during handoff. The semisoft handoff results look promising. Even at the highest handoff rate TCP throughput is almost identical to that of a stationary host as shown in Fig. 6.

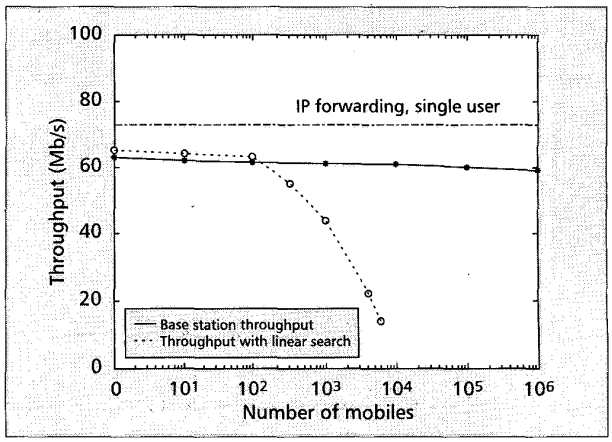
Scalability

The use of per mobile host routes in Cellular IP access networks naturally raises concerns about the ability of the protocol to scale to support high throughput with very large numbers of mobile hosts. As the number of active mobile hosts grows, so will the routing tables in the access network. Routing cache needs to be efficiently searched for each data packet forwarded by a base station. In the case of routing cache misses, the paging cache will be searched for the delivery of downlink packets. The routing cache will maintain mappings for packets that have been recently forwarded. The paging cache is therefore rarely accessed for these packets. Per-host route lookup time in Cellular IP does not limit the number of users connected to the Cellular IP network. Rather, the number of active users is limited. In this case Cellular IP networks can support an order of magnitude more users than other micro-mobility protocols that do not implement passive connectivity and paging.

To estimate the impact of different routing cache sizes on our user space Cellular IP implementation, we create random cache mappings and place them permanently into the routing cache. The solid line in Fig. 7 shows the base station throughput measured for a multihomed 300 MHz Pentium PC base station using `ttcp` and 1500-byte packets for different routing



■ **Figure 6.** *TCP throughput with handoff.*



■ **Figure 7.** *Base station throughput.*

cache sizes. In this experiment we substitute a 100 Mb/s Ethernet connection for a radio interface. The fact that the throughput curve hardly decreases with increasing routing cache size suggests that the performance bottleneck is not the cache lookup time. As shown in Fig. 7, the Cellular IP base station throughput is somewhat lower than the standard IP throughput. This is due to the additional packet processing involved with PCAP, and additional packet copies that take place across kernel and user space domains. We note that the operation of routing cache is very similar to the self-learning operation of Ethernet switches, which can maintain tables of tens of thousands of entries at gigabit speeds. Our results indicate that Cellular IP software base stations are capable of supporting large numbers of mobile hosts and high aggregate throughput. We observe that per-host routes can be supported without diminishing the performance of base station implementation.

Conclusion

In this article we present the design, implementation, and evaluation of the Cellular IP protocol in an experimental testbed setting. Cellular IP represents a new approach to IP host mobility that incorporates a number of important cellular system features such as passive connectivity, paging, and seamless handoff. The Cellular IP routing, handoff, paging, and security algorithms are simple and scalable, resulting in the development of highly scalable software base stations using off-the-shelf PC hardware, operating systems, and radios. We report on the ability of Cellular IP to offer seamless mobility for TCP and UDP applications operating in highly mobile environments. In [12] we evaluate the ability of the protocol to operate under a number of diverse networking scenarios, including picocellular, campus wireless IP, and wireless IP telephony networks. In addition we analyze the mobility management cost of our routing and paging schemes.

The experimental testbed reported in this article has shown that stronger control and management features can be built into commodity IP-based mobile networks without the need for costly and complex circuits. However, a number of challenges remain. Further work is required to extend the protocol with suitable quality of service provisioning to support mobile multimedia. Here we believe that Cellular IP has the fundamental hooks to deliver wireless differentiated services [19] to mobile hosts. More work is required to analyze the protocol response to link and node failure. Issues of particular interest are the consistency of routes after failure and the time to reestablish route after failure.

Further research is required to support multiple gateways in Cellular IP networks. In multiple gateway access networks, a Cellular IP mobile host will use the IP address of one of the gateways as its care-of address and should be capable of changing gateways during normal operations if need be. Finally, the Cellular IP protocol specification, source code, and ns simulation environment are available from the Web (comet.columbia.edu/cellularip).

Acknowledgments

This work is supported in part by Broadcom Research, Ericsson, Fujitsu, IBM, Intel, and Nortel Networks. The authors would like to thank Bill Paul for his invaluable help in implementing Cellular IP. We also thank members of the IETF Mobile IP working group for their helpful comments on an early version of the protocol specification.

References

[1] P. Bhagwat, C. Perkins, and S. Tripathi, "Network Layer Mobility: an Architecture and Survey," *IEEE Pers. Commun.*, vol. 3, no. 3, June 1996, pp. 54–64.
[2] C. Perkins, Ed., "IP Mobility Support," IETF RFC 2002, Oct. 1996.
[3] R. Caceres and V. N. Padmanabhan, "Fast and Scalable Handoffs for Wireless Internetworks," *Proc. ACM Mobicom*, 1996.
[4] A. G. Valko, "Cellular IP: A New Approach to Internet Host Mobility," *ACM Comp. Commun. Rev.*, Jan. 1999.
[5] R. Ramjee *et al.*, "HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area Wireless Networks," *Proc. IEEE Int'l. Conf. Network Protocols*, 1999.
[6] E. Gustafsson, A. Jonsson, and C. Perkins, "Mobile IP Regional Tunnel Management," Internet draft, draft-ietf-mobileip-reg-tunnel-01.txt, Internet draft, Aug. 1999; work in progress.
[7] A. Campbell *et al.*, "Cellular IP," Internet draft, draft-ietf-mobileip-cellularip-00.txt, Dec. 1999; work in progress.
[8] S. F. Foo and K. C. Chua, "Regional Aware Foreign Agent (RAFA) for Fast Local Handoffs," Internet draft, draft-chuafoo-mobileip-rafa-00.txt, Nov. 1998; work in progress.
[9] M. C. Chuah and Y. Li, "Distributed Registration Extension to Mobile IP," Internet draft, draftchuahli-mobileip-dremip-00.txt, Oct. 1997; work in progress.
[10] J. W. Lockwood, "Implementation of Campus-wide Wireless Network Services using ATM, Virtual LANs and Wireless Basestations," *Wireless Communications and Networking Conf. (WCNC'99)*, New Orleans, LA, Sept. 1999.
[11] M. Mouly and M.-B. Pautet, "The GSM System for Mobile Communications," 1992.
[12] A. Campbell *et al.*, "Peformance of Cellular IP Access Networks," Tech. rep., CTR, Columbia Univ., Jan. 1999.
[13] J Ioannidis, D. Duchamp, and G. Q. Maguire Jr., "IP-Based Protocols for Mobile Internetworking," *Proc. ACM Sigcomm '91*, Sept. 1991, pp. 234–45.
[14] "WaveLAN Air Interface," Data Manual, AT&T Corporation, Doc. no. 407-0024785 rev. 2 (draft), July 11, 1995.
[15] H. Balakrishnan, S. Seshan, and R. Katz, "Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks," *ACM Wireless Networks*, vol. 1, no. 4, Dec. 1995.
[16] D. B. Johnson and C. Perkins, "Route Optimization in Mobile IP," Internet draft, draft-ietf-mobileip-optim-07.txt, Nov. 1998; work in progress.
[17] "Mobile IP Authentication, Authorization, and Accounting Requirements, Internet draft, draftietf-mobileip-aaa-reqs-03.txt, Mar. 2000; work in progress.
[18] S. McCanne and V. Jacobson, "The BSD Packet Filter: A New Architecture for User-Level Packet Capture," *USENIX '93*, San Diego, CA.
[19] S. Blake *et al.*, "An Architecture for Differentiated Services," IETF RFC 2475, Dec. 1998.

Biographies

ANDREW T. CAMPBELL (campbell@ctr.columbia.edu) is an assistant professor in the Department of Electrical Engineering and member of the COMET Group at the Center for Telecommunications Research (CTR), Columbia University, New York. His research interests include open programmable networks, mobile networking, and quality of service. He is active in the IETF and the international working group on Open Signaling (OPENSIG). He is a past co-chair of the 5th IFIP/IEEE International Workshop on Quality of Service (IWQOS '97) and 6th IEEE International Workshop on Mobile Multimedia Communications (MOMUC '99). He received his Ph.D. in computer science in 1996.

JAVIER GOMEZ CASTELLANOS [StM] (javierg@comet.columbia.edu) obtained a B.S. degree with honors in electrical engineering in 1993 from National Autonomous University of Mexico (UNAM), and an M.S. degree in electrical engineering in 1996 from Columbia University, New York. Since 1996 he has been a Ph.D. student in the COMET Group at CTR, Columbia University, New York. His research interests cover routing, QoS, and power-aware design for cellular and ad-hoc networks.

SANGHYO KIM's biography was not available when this issue went to press.

ANDRÁS VALKO received his M.Sc. (EE) and Ph.D. (CS) degrees in 1994 and 1999, respectively, both from the Technical University of Budapest (TUB), Hungary. In 1994 he joined the High Speed Networks Laboratory at TUB where he worked mainly on ATM performance issues. Since 1996 he has been with Ericsson's Traffic Analysis and Network Performance Laboratory. His fields of research include Internet mobility, wireless/mobile Internet access, third-generation cellular mobile networks, mobile ad hoc networks, and mobile system performance analysis. The work presented in this article was carried out while he was a visiting researcher at CTR, Columbia University, New York.

CHIEH-YIH WAN received his M.S. degree in electrical engineering from Columbia University in 1999. Since 2000 he is with Comet Group in Columbia University as a Ph.D. student. He is interested in the areas of wireless networking, IP micro-mobility support, and sensor networking/smart space.

ZOLTÁN RICHÁRD TURÁNYI started his university studies at the Technical University of Budapest, and graduated in informatics (M.Sc.) and started his Ph.D. studies in 1996. Currently he works in Ericsson TrafficLab on his thesis. His research interests include mobile and ad hoc packet networks, QoS, and network simulation.