ON THE ANALYSIS OF CELLULAR IP ACCESS NETWORKS

András G. Valkó*
Ericsson Research
andras.valko@lt.eth.ericsson.se

Javier Gomez, Sanghyo Kim, Andrew T. Campbell
Center for Telecommunications Research, Columbia University, New York
[javierg,shkim2,campbell]@comet.columbia.edu

Abstract

Mobile IP represents a simple and scalable global mobility solution but lacks support for fast handoff control and real-time location tracking found in cellular networks today. In contrast, 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. Future wireless networks should be capable of combining the strengths of both approaches without inheriting their weaknesses. In this paper we present analysis of Cellular IP, a new host mobility protocol which represents one such approach. Cellular IP incorporates a number of important cellular system features but remains firmly based on IP design principles. The protocol presented in this paper is implemented as extensions to the ns simulator.

1 Introduction

Recent initiatives to add mobility to the Internet mostly focus on the issue of address translation [2] through the introduction of location directories and address translation agents. In these protocols (e.g., Mobile IP [1]), packets addressed to a mobile host are delivered using regular IP routing to a temporary address

*Visiting Scientist, Center for Telecommunications Research, Columbia University, New York

assigned to the mobile host at its actual point of attachment. This approach results in simple and scalable schemes that offer global mobility support. It is not appropriate, however, for fast mobility and smooth handoff because after each migration a local address must be obtained and communicated to a possibly distant location directory or home agent (HA). Cellular mobile telephony systems are founded on radically different concepts. Instead of aiming at global mobility support, cellular systems are optimized to provide fast and smooth handoff in a restricted geographical area. In the area of coverage mobile users have wireless access to the mobility unaware global telephony network. A scalable forwarding protocol interconnects distinct cellular networks to support roaming between them.

Restricting the cellular coverage to a limited geographical area limits the potential number of connected users. This makes it feasible to maintain per mobile states which we believe is key to delivering fast handoff support to mobile hosts. Having per-mobile location information allows the cellular system to support location independent addressing avoiding the need to change addresses during each intra-network migration. Even in limited geographical areas, however, the number of users can grow to a point where using fast lookup techniques for per user data bases is no longer viable. In addition, mobility management requires mobile hosts to send registration information after migration. The resulting signaling overhead has significant impact on the performance of the wireless access network. To overcome this problem, cellular telephony systems require mobiles to register every migration only when they are engaged in "active" calls. In contrast, "idle" mobile hosts send registration messages less frequently and as a result can roam in large areas without loading the network and the mobility management system. The location of idle mobile hosts is only approximately known to the network at any one time. To establish a call to an idle mobile, the mobile host must be searched for in a limited set of cells. This feature called *passive connectivity* allows the cellular network to accommodate a very large number of users at any instance without overloading the network with large volumes of mobility management signaling information and messaging.

Cellular networks offer a number of desirable features which if applied correctly could enhance the performance of future wireless IP networks without loosing any of important flexibility, scalability and robustness properties that characterize IP networks. However, there are fundamental architectural differences between cellular and IP networks that make the application of cellular techniques to IP challenging. Cellular telephony systems rely on a restrictive "circuit" model that requires connection establishment prior to communication. In contrast, IP networks perform routing on a per packet basis. In addition, to-

day's cellular systems are strictly based on hierarchical networks and use costly mobile-aware nodes (e.g., MSC). We believe that a future Cellular Internet should be founded on IP, inheriting its simplicity, flexibility and robustness. A Cellular Internet should leverage mobility management and handoff techniques found in cellular networks. A single scalable host mobility protocol should be capable of flexibly supporting pico, campus and metropolitan area networks based on a set of simple and cheap network nodes that can be easily interconnected to form arbitrary topologies that operate without prior configuration.

In this paper, we present an analysis of *Cellular IP* [8] [9], a new mobile host protocol that is optimized to provide access to a Mobile IP enabled Internet in support of fast moving wireless hosts. Cellular IP incorporates a number of important cellular principles but remains firmly based on IP design principles. Because of its IP based design and the feature of passive connectivity, Cellular IP can scale from pico to metropolitan area installations. The Cellular IP distributed location management and routing algorithms lend themselves to a simple, efficient and low cost implementation for host mobility requiring no new packet formats, encapsulation or address space allocation beyond what is already present in IP. The paper is structured as follows. In Section 2, we present an overview of the Cellular IP protocol. Following this in Section 3 we analyze the protocol which is implemented as extensions to the ns simulator. In particular we discuss the handoff performance and cost of mobility management. We present some concluding remarks in Section 4.

2 Protocol Overview

2.1 Features

The universal component of a Cellular IP network is a *base station* which serves as a wireless access point but at the same time routes IP packets and integrates the mobility specific control functionality traditionally found in Mobile Switching Centers (MSC) and Base Station Controllers (BSC). Base stations are built using the regular IP forwarding engine, however, 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 their gateway as their Mobile IP care-of address. Figure 1 illustrates the path taken by packets addressed to a mobile host. Assuming Mobile IPv4 [1] and no route optimization [7], packets will be first routed to the host's home agent and then tunneled to a gateway. The gateway "detunnels" packets and forwards them toward the base stations. Inside the Cellular IP network, mobile hosts are identified by their home addresses and

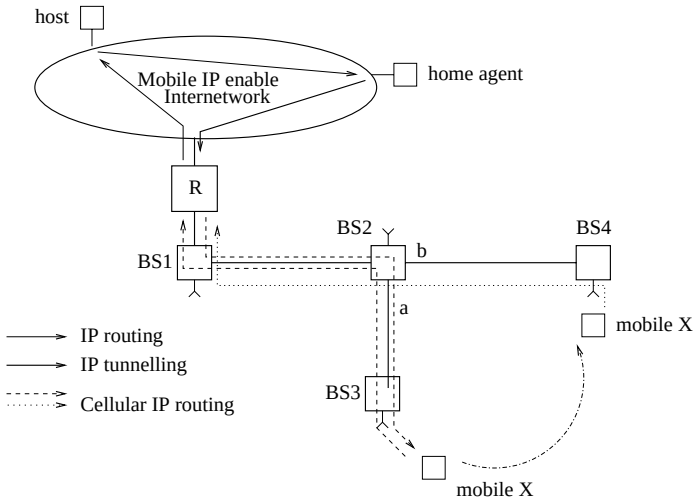

Figure 1: A Cellular IP Access Network Interconnected to a Mobile IP enabled Internet

data packets are routed without encapsulation, tunneling or address conversion. The Cellular IP routing protocol ensures that packets are delivered to the host's actual location; that is, the base station that serves as the mobile host's point of attachment to the Cellular IP access network. Packets transmitted by mobile hosts are first routed to a gateway and from there on to the global Mobile IP enabled 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 establish host location information. *Uplink* packets are routed from mobile hosts to the gateway on a hop-by-hop basis. The path taken by these packets is cached by 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 path routed. When the mobile host has no data to transmit it sends empty IP packets to the gateway to maintain its downlink "soft" routing state. Following the principle of passive connectivity mobile hosts that have not received packets for some period allow their downlink soft-state routes be cleared from the caches. In order to route packets to idle hosts a Cellular IP mechanism called *paging* is used. In what follows we provide a brief overview of the Cellular IP functions. For a full

discussion of Cellular IP see [8] and for a full specification of the protocol see [9].

2.2 Routing

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

Packets transmitted by a mobile host traverse the access network nodes destined for the gateway. As these packets pass each node on route to the gateway their route information is recorded as follows. Each base station maintains a "soft-state" *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 interface over which the packet entered the node. In the scenario illustrated in Figure 1 data packets are transmitted by a mobile host with IP address **X** and enter base station **BS2** through its interface **a**. In the routing cache of base station **BS2** this is indicated by a mapping **(X,a)**. This mapping remains valid for a system specific time called the *route-timeout* and its validity is renewed by each data packet that traverses the same interface coming from the same mobile host. As long as the mobile host is regularly sending data packets then base stations along the path (between the mobile's actual location and the gateway) maintain valid entries in their routing cache forming a soft-state route between the mobile host and gateway nodes. Packets addressed to the same mobile host are routed on a hop-by-hop basis using the 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 for this is when a mobile host is the receiver of a stream of UDP packets and has no data to transmit. To keep its routing cache mappings valid the mobile host transmits the *route-update packets* at regular intervals called *route-update time*. These packets are empty data packets addressed to the gateway. Route-update packets have the same effect on routing cache as normal data packets, however, they do not leave the Cellular IP network.

2.3 Handoff

The Cellular IP *hard handoff* is based on a simply approach that tolerates some potential packet loss in exchange for minimizing handoff messaging rather than

guaranteeing zero packet loss. Handoff is initiated by mobile hosts in a Cellular IP access network. Hosts listen to beacons transmitted by base stations and initiate handoff based on signal strength measurements. To perform a handoff a mobile host has to tune its radio to the new base station and send a route-update packet. This creates routing cache mappings on route to a gateway hence configuring the downlink route to the new base station. Handoff latency is the time that elapses between the handoff and the arrival of the first packet through the new route. For hard handoff, this time will be equal to the round-trip time between the mobile host and the cross-over point which is the gateway in the worst case. During this time downlink packets may be lost. Mappings associated with the old base station are not cleared explicity during handoff. Rather, they are cleared by a soft-state routing mechanism resident at each node in the Cellular IP access network on expiration of the route-timeout.

Before these mappings timeout a period exists when both the old and new downlink routes are valid and packets are delivered through both base stations. This feature is used in the *semisoft handoff* procedure that improves handoff performance but suits the lightweight nature of the base protocol by providing probabilistic guarantees instead of fully eliminating packet loss by, for example, retransmissions. Semisoft handoff adds a single temporary state to the soft state protocol in mobile hosts and base stations and scales well for a large number of mobile hosts and frequent handoffs.

The semisoft handoff procedure has two 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. When the mobile host initiates a handoff it first sends a *semisoft packet* to the new base station and immediately returns to listening to the old base station. While the host is still in connection with the old base station, the semisoft packet configures routing cache mappings associated with the new base station. After a *semisoft delay*, the host performs a regular handoff. The semisoft delay can be anything between the mobile-gateway round-trip time and the route-timeout. (In our ns simulation environment we use a conservative value of 100 ms). This delay ensures that by the time the host tunes its radio to the new base station its downlink packets are being delivered through both the old and new base stations.

2.4 Paging

Cellular IP defines an *idle mobile host* as one that has not received data packets for a system specific time called the *active-state-timeout*. Idle mobile hosts let their respective soft-state routing cache mappings timeout. These mobile

hosts transmit *paging-update packets* at regular intervals defined by the *paging-update-time*. The paging-update packet is an empty IP packet addressed to the gateway that is distinguished from a route-update packet by its IP type parameter. The mobile host sends its paging-update packets to the base station that has the best signal quality. Similar to data and route-update packets, paging-update packets are routed on a hop-by-hop basis toward the gateway. Base stations may optionally maintain *paging cache*. A paging cache has the same format and operation as a routing cache with two differences. First, paging cache mappings have a longer timeout period called the *paging-timeout*. Second, paging cache mappings are updated by any packet sent by mobile hosts including paging-update packets. In contrast, routing cache mappings are updated by data and route-update packets sent by mobile hosts. This results in idle mobile hosts having mappings in paging caches but not in routing caches. In addition, active mobile hosts will have mappings in both types of cache. 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 on all its interfaces except the one the packet came through. Paging cache is used to avoid broadcast search procedures found in cellular systems. Base stations that have paging cache will only forward the paging packet if the destination has a valid paging cache mapping for the mobile host and only to the mapped interface(s). Without any paging cache the first packet addressed to an idle mobile is broadcast in the access network. While the packet does not experience extra delay it does, however, load the access network. Using paging caches, the network operator can restrict the paging load in exchange for memory, processing and bandwidth cost.

Idle mobile hosts that receive a packet move from idle to active state and start their active-state-timer and immediately transmit a route-update packet. This ensures that routing cache mappings are established quickly potentially limiting any further flooding of messages to mobile hosts in Cellular IP access networks.

3 Analysis

In this section we analyze the handoff performance of Cellular IP access networks. We quantify the performance penalty associated with the Cellular IP handoff scheme which trades performance (e.g., packet loss and delay) for simplicity. Furthermore, we investigate the “cost” of mobility management

for routing and paging in Cellular IP access networks. Determining the mobility management cost is important because different cellular system installations (e.g., pico-cellular and macro-cellular access networks) will operate under different mobility conditions.

3.1 Simulation Environment

The Cellular IP protocol is implemented as extensions to the ns simulator [10] which is widely used by the networking community to analyze IP networks. The Cellular IP simulation environment used for the reported results is shown in Figure 1; note that the simulator supports Cellular IP access networks of arbitrary topology. The assumptions and limitations of the Cellular IP ns simulation environment are as follows. First, an "ideal wireless interface" is used; that is, packets transmitted over the wireless interface encounter no delay, bit error or loss and congestion over the air interface is not modeled. Next, the beacon messages transmitted by a Cellular IP gateway are not modeled. The network is configured when the simulation session is initiated and the topology remains constant for the duration of the simulation. Finally, wireless cells are assumed to overlap and mobile hosts move from one cell to another in zero time. This does not limit the ns simulator's ability to study packet loss during handoff because such packet loss is mainly a product of misrouted packets.

3.2 Handoff Performance

The design of a fast and efficient handoff algorithm is central to the performance of a cellular access network especially in the case of networks that are comprised of small wireless cells with fast moving mobile hosts. One of the design goals of Cellular IP is to operate efficiently at very high handoff frequencies. In accordance with this design goal, the Cellular IP handoff algorithm avoids explicit signaling messages (used for example in cellular telephony and Mobile IP systems) and buffering or forwarding of packets [5] [6]. As a result Cellular IP packets may be lost during handoffs. In such cases we assume that packet loss is dealt with by higher layer protocols (e.g., TCP). In this section we analyze the performance of Cellular IP handoff to determine the performance penalty we pay for our simple approach to host mobility.

3.2.1 Delay

The impact of handoff on ongoing sessions is commonly characterized by the *handoff delay*. Handoff delay is usually defined as the time taken to resume

normal traffic flow after a mobile host performs handoff. Though this does not fully determine the performance seen by applications, it is a good indication of the handoff performance. In [11] handoff delay is decomposed as *rendezvous* and *protocol time*. Rendezvous time refers to the time taken for a mobile host to attach to a new base station after it leaves the old base station. This time is related to wireless link characteristics, particularly to the inter-arrival time of beacons transmitted by base stations. Protocol time refers to the time taken to restore traffic flows/sessions once a mobile host has received a beacon from the new base station. In the following analysis we assume that the rendezvous time is small and handoff performance is determined by the protocol time. Rather than adopting the notations proposed in [11], we define the handoff delay as the time it takes a mobile host to receive the first packet through the new base station after it moved from the old to the new base station, which, as discussed earlier, we assume to take zero time.

In Cellular IP, handoff delay and packet loss are consequences of the time it takes for the distributed routing state to follow host mobility. As described in Section 2, immediately after handoff, mobile hosts transmit a route-update packet to reduce this time to a minimum. The route-update packet travels from the new base station to the gateway configuring the new soft-state downlink route toward the mobile host's new point of attachment. The old and new downlink routes both originate at the gateway but while the former routes packets to the old base station, the latter leads to the base station the host has just moved to.

A handoff scenario is illustrated in Figure 1. The node where the old and new routes join base station (**BS2**) in Figure 1 is referred to as the *cross-over node*. The new downlink route becomes operational when the first route-update packet transmitted through the new base station reaches the cross-over node. The time period during which time the mobile host is not receiving packets after initiating handoff represents the the time taken for the route-update packet to reach the cross-over node plus the time taken for the first downlink packet to travel from the cross-over node to the new base station. Handoff delay is equal to the round-trip time between the new base station and the cross-over node.

3.2.2 Packet Loss

In addition to handoff delay, application level service quality is also related to packet loss during handoff. To determine handoff packet loss, let us assume that a periodic stream of packets is being transmitted from the Internet to a mobile host. Before a handoff is initiated packets are routed along the old route. In the following calculation, we will assume that the cross-over node knows in

advance which of the stream's packets will be the last one to reach the mobile host at the old location. Let us assume that the cross-over node marks this packet. Upon receiving the marked packet, the mobile host performs a handoff and immediately transmits a route-update packet through the new base station. Downlink packets routed by the cross-over node after the marked packet but before the arrival of the route-update packet are routed to the old base station and are lost. This time interval is equal to the sum of the time taken for the marked packet to propagate from the cross-over node to the mobile host and the time taken for the route-update packet to reach the cross-over node. The loss of packets at handoff is related to the "handoff loop time" which is defined as the transmission time from the cross-over node to the mobile host's old location plus the transmission time from the mobile host's new location to the cross-over node. Specifically, the number of lost packets at handoff n_{loss} is equal to the number of packets arriving at the cross-over node during the handoff loop time T_L, that is

$$n_{loss} = wT_L \tag{1}$$

where w is the rate of downlink packets. Since the average handoff loop time is equal to the average handoff delay, the expected number of packets lost at handoff can equally be calculated using the handoff delay. In what follows we do not differentiate between these two values.

Handoff packet loss for a Constant Bit Rate (CBR) source using our simulation environment is plotted in Figure 2. The curve represents the average number of packets lost during handoff against down link packet inter-arrival time in seconds. The three curves correspond to T_L values of 0.002, 0.02 and 0.2 seconds, respectively (twice the link delay shown in Figure 2). The simulation results closely match the calculations presented above. These results are achieved with neither mobile hosts nor base stations having special states associated with handoff. In exchange for this simplicity, however, handoff performance is dependent upon the traffic conditions. In a highly loaded network the handoff delay and packet loss will be higher.

Real time Internet applications (e.g., voice over IP) are sensitive to packet delay and cannot typically tolerate the delay associated with the retransmission of lost packets. For these applications, the number of lost packets characterizes handoff performance. Other applications, however, use end-to-end flow control to respond to network and traffic conditions and retransmit packets and/or reduce transmission rate if errors occur. In what follows, we focus on TCP performance in the presence of handoff. TCP represents the most typical traffic type over today's Internet which carries World Wide Web, file transfer, re-

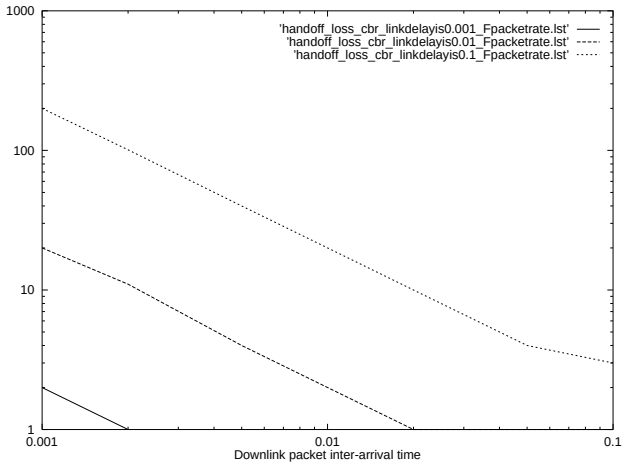

Figure 2: Packet Loss vs. CBR Packet Inter Arrival Time

mote login and other applications. Investigating TCP performance is important because its flow control has been shown to operate sub-optimally in wireless environments.

3.2.3 TCP Behavior

We will first use simulation to look at the behavior of a TCP session during handoff. The simulated configuration is identical to the Cellular IP simulation environment shown in Figure 1. In the first example TCP is used to download data to a mobile host. The TCP packet size is 1000 bytes and a mobile user has up to 5 Mbps downlink bandwidth, that is, the downlink packet rate w is 625 packets/sec. Packet transmission time between nodes in the simulated configuration is 2 ms, resulting in a handoff delay of 4 ms.

Figure 3 shows the sequence numbers of downlink data packets and uplink acknowledgments observed at the gateway during handoff; note that TCP Tahoe flow control is operational throughout. Handoff is initiated by the mobile host at 4 seconds into the simulation. In accordance with Equation 1 three consecutive packets get lost as indicated by the three consecutive missing acknowledgments. After the handoff delay packets continue to arrive at the mobile host. These packets are, however, out of sequence and cause the receiver to generate duplicate acknowledgments as indicated by the horizontal line of

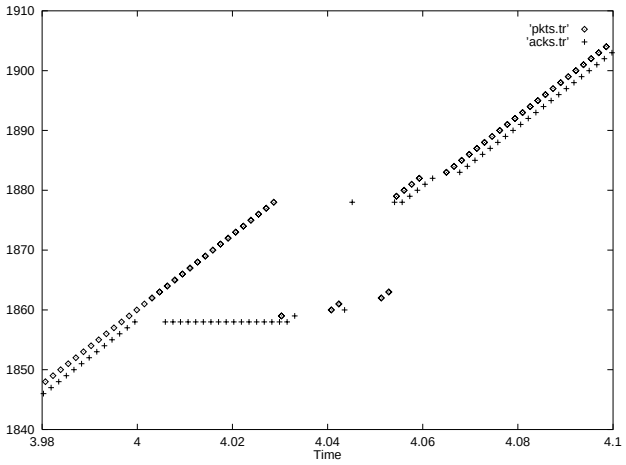

Figure 3: TCP Sequence Numbers at Handoff (Downlink Case)

acknowledgment sequence numbers. The duplicate acknowledgments inform the TCP transmitter about the losses and cause it to retransmit the lost packets. The first retransmitted packet arrives approximately 20 ms after the handoff (see Figure 3). Using Tahoe flow control, the transmitter remains silent until this packet is acknowledged and increases its transmission window size as further acknowledgments arrive. The full TCP rate is regained at 4.07 sec into the simulation as shown in Figure 3. The figure represents TCP sequence numbers at the client side transmitter for both packets and acknowledgements against time in seconds.

Cellular IP handoff is interpreted by a transmitter in the wired IP network as congestion which causes it to reduce its transmission rate. Using Tahoe flow control the handoff triggers slow-start which increases the performance impact of handoff packet loss. From the simulation results we observe that normal operation is resumed approximately 70 ms after handoff is initiated as shown in Figure 3.

In the next experiment TCP is used to carry data from the mobile host. In this case handoff packet loss affects acknowledgments instead of data packets. Figure 4 shows simulation results for a configuration that is identical to the previous one. Before handoff is initiated the TCP sender at the mobile host uses its maximum window size of 20 packets which is reflected in the difference between data packet and acknowledgment sequence numbers. At

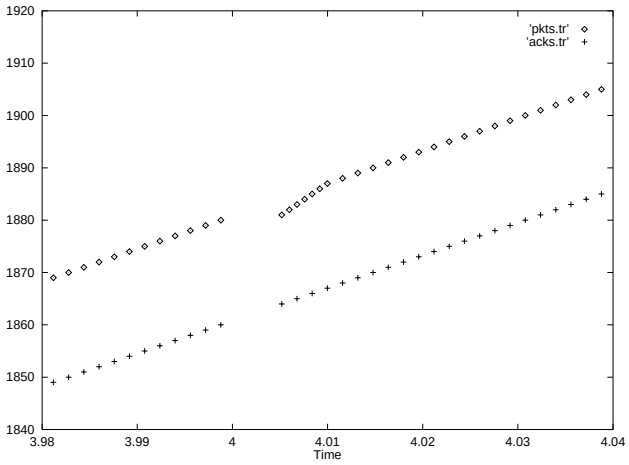

Figure 4: TCP Sequence Numbers during Handoff for the Uplink Case

4 sec (simulated time) the mobile host performs a handoff and stops receiving acknowledgments for a period of approximately 4 ms, which represents the handoff delay. During the handoff delay the sender does not transmit any packets since its window size is used up and it needs incoming acknowledgments to advance its transmission window.

In the next experiment (as shown in Figure 4) handoff is initiated when the TCP session is in a stabilized phase and acknowledgments keep arriving at the mobile host in a paced and continuous manner. After the handoff delay, acknowledgments are routed to the mobile host's new location. Due to the cumulative nature of TCP acknowledgments, the first acknowledgment that arrives at the mobile host after handoff informs the sender that all its transmitted packets have arrived at the receiver (up to the sequence number shown in the acknowledgment). This causes the transmitter to advance its transmission window and continue transmitting at the maximum available data rate. In the simulation example this rate is slightly higher than the rate dictated by TCP flow control which represents the long term average capacity. This results in a curve of data packet sequence numbers that is somewhat steeper during handoff. As observed in Figure 4, normal operation is resumed quickly with the result that handoff has little impact on the active data session.

We observe that the behavior is different if handoff occurs when a TCP session is in its initial slow start phase and acknowledgments are not regularly

arriving at the mobile host. In this case the new downlink route is established after the handoff delay but no acknowledgments arrive to the sender. If at this point the sender has used up all its transmission window and is waiting for acknowledgments then TCP can suffer a delay equal to the sender's retransmission timer. Mechanisms to avoid this problem are for further study.

3.3 Mobility Management Cost

3.3.1 Route Maintenance Overhead

The network operator will typically set the route-timeout to be a small multiple of the route-update time. This ensures that the mobile host's routing cache mappings remain valid even if a few route-update packets are lost. Let T_{ru} denote the route-update time and αT_{ru} the route-timeout where α is a small integer. To choose an optimal value for T_{ru}, the following trade-off should be observed. After an active host performs a handoff, its old routing cache mappings remain valid for a duration determined by route-timeout. During this time, packets addressed to this host continue to be delivered to the old base station increasing the network load and reducing network performance. A small value of T_{ru} should be used to minimize this condition. On the other hand, an active host that has no data to send must transmit route-update packets at a rate of $1/T_{ru}$. This load increases with decreasing T_{ru}. Let the cost of carrying a packet to or from the mobile host be defined as the size of the packet in bits. This model neglects differences in uplink and downlink cost due to different traffic conditions but is sufficient to characterize the T_{ru} trade-off. Consider a mobile host that is receiving data at a constant rate r bps (including headers) and let p denote the fraction of the time when it is not sending packets and is forced to transmit route-update packets instead. (We note that in some typical IP applications downlink traffic is considerably higher than uplink traffic. This, however, does not necessarily cause p to be high if acknowledgments are transmitted over the uplink.) The cost of transmitting route-update packets during time T is $R_{ru}pT/T_{ru}$ where R_{ru} is the size of a route-update packet in bits. During this time the mobile performs T/T_H handoffs where T_H (dwell time) is the mean time spent in a cell. After each handoff, the old route remains active for at most αT_{ru}, the exact value depending on when it was last updated before handoff. Hence the mean cost of sending packets along the old route after handoff is $rT_{ru}(\alpha - 1/2)$ and the total cost of misrouted packets during time T is $rTT_{ru}(\alpha - 1/2)/T_H$. The optimal route-update time $\hat{T}_{ru}$ is the one that minimizes the sum of these costs and is calculated as

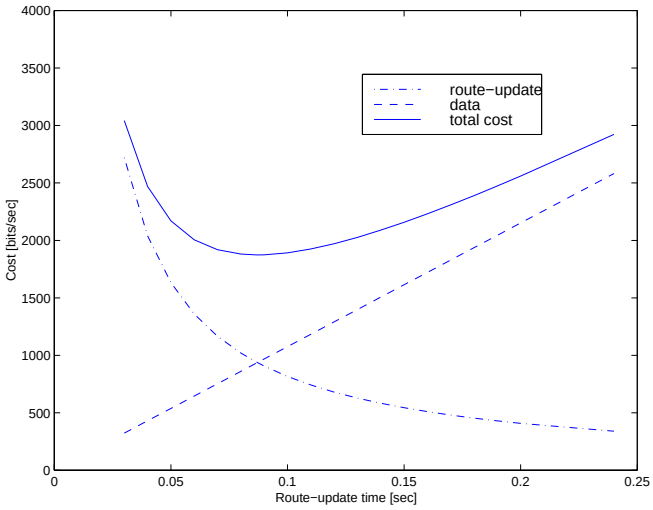

Figure 5: Location Management Cost vs. T_{ru} (Dotted Lines Represent Simulation Measurements)

$$\hat{T}_{ru} = \sqrt{\frac{pR_{ru}T_H}{r(\alpha - 1/2)}}$$

This theoretical result is plotted in Figure 5. The mobile host performs handoffs every 30 seconds while it is receiving data at a rate of 128 kbps. The size of route-update packets is 102 bytes, α is 3 and p is 0.1. The plot shows that the optimal route-update time in the described scenario is 87 ms. We also plotted the sum of these costs which can be interpreted as the total cost associated with the mobility of an active host and is calculated as

$$C_a = \frac{pR_{ru}}{\hat{T}_{ru}} + \frac{r\hat{T}_{ru}(\alpha - 1/2)}{T_H} = \sqrt{\frac{4pR_{ru}r(\alpha - 1/2)}{T_H}}$$

This cost is not proportional to the migration frequency but to its square root. In keeping with the original design goals, this shows Cellular IP's efficiency in supporting highly mobile hosts. Note that the mobility cost increases with increasing user data rate. This property applies to most mobility schemes (e.g., when data must be forwarded from one base station to another after handoff) but is more apparent in Cellular IP. This is related to the soft-state nature

of Cellular IP. Since there is no explicit signaling during handoff, which makes handoff transparent to the base stations, the base station is unaware that mobile hosts move into or out of its cell. Transmitting data to mobile hosts that have left the cell adds to the cost of mobility.

3.3.2 Paging Overhead

The paging-update time T_{pu} is subject to a similar trade-off as T_{ru}. A selected value that is too small will result in very frequent paging-update packets being sent by idle mobile hosts. On the other hand, considering that the paging-timeout is a small multiple of the paging-update time, increasing T_{pu} will result in an increase in the number of cells that an idle mobile host is paged in.

Paging is initiated when a new data session starts by a downlink packet, for instance a TCP connection is initiated *to* the mobile host. Let λ_P denote the arrival rate of such sessions and R_P the mean amount of traffic (bits) sent in paging packets. The paging packets are delivered to all the cells to which the mobile host has valid paging cache mappings. Let us first assume that all base stations have paging caches and that the probability of immediately revisiting a cell is negligible. Paging occurs in the 'primary' cell that the target mobile host resides in plus any other 'secondary' cells where the mobile host has valid paging cache mappings. Secondary cells represent cells that the mobile host has recently visited and that have valid paging cache for the target mobile host. Paging secondary cells is a waste of transmission resources and reflects the cost of our paging scheme. The mean number of secondary cells paged is $(\beta - 1/2)T_{pu}/T_H$, where β is the ratio between the paging-timeout and the paging-update time. The optimal paging-update value $\hat{T}_{pu}$ is the one that minimizes the sum of paging-update traffic and wasted paging traffic and is obtained as

$$\hat{T}_{pu} = \sqrt{\frac{R_{pu}T_H}{\lambda_P R_P(\beta - 1/2)}}$$

where R_{pu} is the size of paging-update packets in bits. Using this optimal paging-update time, the total cost C_i associated with the mobility of an idle host is

$$C_i = \sqrt{\frac{4R_{pu}\lambda_P R_P(\beta - 1/2)}{T_H}}$$

These results take a similar form to those obtained for the route-update time. However, the downlink data rate r (an important parameter in the route-update time trade-off) is now replaced by $\lambda_P R_P$ which is the rate at which data

arrives at the mobile host in paging packets. This rate depends largely on the application but will be in most cases orders of magnitude lower than r which justifies selecting a higher paging-update time than route-update time. This also accounts for the fact that the cost C_i associated with the mobility of idle hosts is significantly lower than the mobility cost of active users which is the basis of passive connectivity.

4 Conclusion

In this paper we have presented an analysis of the Cellular IP protocol. Cellular IP represents a new approach to IP host mobility that incorporates a number of important cellular system features but remains firmly based on IP design principles. A fundamental design objective of Cellular IP is implementational and functional simplicity. To reduce complexity, we omitted explicit location registrations and replaced them by implicit inband signaling. As a result, nodes in a Cellular IP access network need not be aware of the network topology or of the mobility of hosts in the service area. This design choice deliberately trades off performance for simplicity. As a result packets may be lost at handoff rather than explicitly buffered and redirected to mobile hosts as they move. Our analysis has focused on the performance of the Cellular IP hard handoff algorithm and on the network traffic overhead imposed by mobility management. We have found that a simple approach can offer fairly good service quality. We have presented an analytical and empirical study of protocol parameters that can be set to configure a Cellular IP access network to match local mobility and traffic characteristics. Future work is addressing new mechanisms to provide quality of service support while maintaining the same simple lightweight protocol approach to host mobility and wireless access to the Internet.

Acknowledgments

The authors wish to thank members of the IETF Mobile IP Working Group for their comments on the first version of this protocol. In addition, the COMET Group would like to thank Ericsson, IBM and Intel for for their on-going support of the Cellular IP Project (comet.columbia.edu/cellularip) at Columbia University.

References

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

[2] 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.

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

[4] Charles Perkins, “Mobile-IP Local Registration with Hierarchical Foreign Agents,” Internet Draft, draft-perkins-mobileip-hierfa-00.txt, Work in Progress, February 1996.

[5] H. Balakrishnan, S. Seshan, R. Katz, “Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks,” ACM Wireless Networks 1(4), December 1995.

[6] John Ioannidis, Dan Duchamp, Gerald Q. Maguire Jr., “IP-Based Protocols for Mobile Internetworking,” Proc. ACM Sigcomm’91, pp. 234-245, September 1991.

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

[8] András G. Valkó, “Cellular IP: A New Approach to Internet Host Mobility,” ACM Computer Communication Review, January 1999.

[9] A. Valkó, A. Campbell, J. Gomez, “Cellular IP,” Internet Draft, draft-valko-cellularip-00.txt, Work in Progress, November 1998.

[10] “Network Simulator - ns (version 2)”, ns home page, http://www-mash.cs.berkeley.edu/ns/ns.html.

[11] Ramon Cáceres, Venkata N. Padmanabhan, “Fast and Scalable Handoffs for Wireless Internetworks,” in *Proc. ACM Mobicom*, 1996.