Challenges for Mobile Voice-Over-IP

Prathima Agrawal[1], Jyh-Cheng Chen[1], and Cormac J. Sreenan[2]

[1] Applied Research, Telcordia Technologies, Morristown, NJ 07960, USA
{pagrawal,jcchen}@research.telcordia.com
[2] Department of Computer Science, University College Cork, Cork, Ireland
cjs@cs.ucc.ie

Abstract. The use of IP to transport voice is receiving significant attention in the telecommunications and data networking industries. Commonly known as Voice-over-IP (VoIP), this technology has the potential to allow user communication involving multiple media types, and between terminals that offer improved capabilities and user interfaces. Meanwhile there has been explosive growth in the demand for wireless connectivity, raising the question of how VoIP can be extended to situations where users and terminals can be mobile. In this paper we survey the technologies necessary to provide mobile VoIP, and identify relevant technical challenges.

1 Introduction

IP telephony is expected to see continued growth in supporting corporate and commercial services. This growth points to a large business and market value for Voice-over-IP (VoIP) in the near future. Currently IP telephony provides voice service to end terminals that are attached to wired networks. With the proliferation of mobile and wireless services and as more users disconnect from their fixed access points and become mobile, there is an increasing need to adapt VoIP to the mobile and wireless domain [10].

However, the mobile environment is much more dynamic and subject to change than the traditional wireline environment. The uncertainties of the wireless and mobile environments call for an increased level of adaptability. For more robust real-time communications, a signaling protocol is vital in providing highly reliable and robust connectivity in such communications environment. In addition to establishing and releasing a call, a signaling protocol may also need to monitor and maintain connectivity when the end terminal is moving and/or the transmission capabilities are varying. Dealing with issues of user and terminal mobility in an internetwork takes on special significance when voice transport is involved, because of latency and general performance constraints that are not as important for non-real time communication. Similarly, the service quality necessary for voice communication demands packet scheduling and resource reservation techniques which can operate effectively in the presence of mobility.

This paper is organized as follows. Section 2 provides a survey of signaling for VoIP, and describes a signaling architecture for use with mobile VoIP terminals.

Section 3 explains how user and terminal mobility are handled in the telephone system and in the Internet. Performance issues for supporting voice to mobile IP terminals are discussed. Section 4 discusses how to provide support for quality of service (QoS) for packet voice communication over wireless links and where terminals are non-stationary. Finally, Section 5 concludes the paper.

2 Signaling

The first challenge of IP telephony is to initiate interactive communication sessions between users. Two major standards have recently emerged for the signaling and control of VoIP. The first is H.323 [8] supported by the ITU-T, and the second is the Session Initiation Protocol (SIP) [6] proposed by the IETF. These two protocols provide similar signaling functionality, however H.323 is widely deployed in commercial products. Microsoft released for example NetMeeting, a H.323-compliant product, in 1996. On the other hand, SIP is a lightweight protocol which provides a simpler implementation and a greater degree of efficiency than H.323. These two protocols, however, currently do not provide support for resource reservation and QoS guarantees. DOSA (Distributed Open Signaling Architecture) [5] and TOPS (Telephony Over Packet networkS) [9], proposed by AT&T, provide a framework to integrate call signaling and resource management for IP telephony, and to enable creative new services for packet telephony. ITSUMO [7], a Telcordia-Toshiba joint research project, aims on providing IP signaling along with terminal mobility and QoS support.

2.1 H.323

Originally H.323 was developed for visual telephone terminal conferencing over non guaranteed QoS LANs and is an umbrella standard covering audio and video coders, call signaling, connection control, data and conference control, media transport, etc. In H.323, the signaling functionality is migrated to end terminals which are intelligent end points instead of the dumb end points used in the PSTN (Public Switched Telephone Network). H.323 is also not tied to a single transport mechanism, and can run over a variety of networks, including ATM and ISDN.

Fig. 1 shows a typical architecture and components of an H.323 LAN which is called a H.323 zone. The terminal in H.323 usually is a multimedia PC. The gateway (GW) is an endpoint on the network which provides for real-time, two-way communications between H.323 terminals on the packet-based network, other ITU terminals on a switched circuit network, and other H.323 gateways. The Multipoint Control Unit (MCU) is an endpoint which provides the capability for three or more terminals and gateways to participate in a multipoint conference. The gatekeeper (GK) is an entity that provides address translation and controls access to the network for H.323 terminals, gateways, and MCUs. The gatekeeper may also provide other services to the terminals, gateways and MCUs such as bandwidth management and locating other gateways.

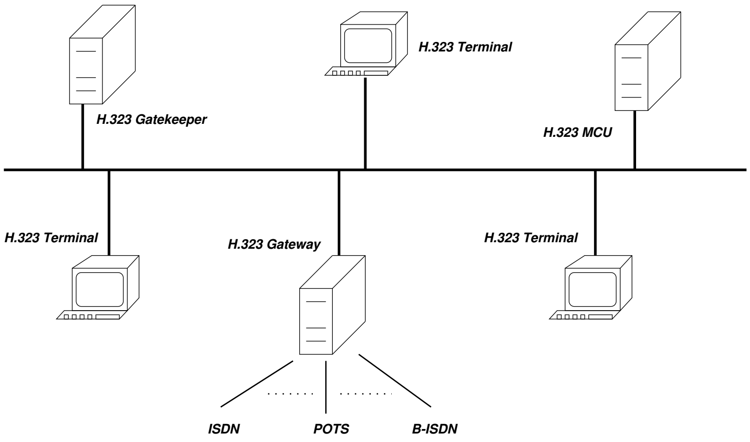

Fig. 1. H.323 zone

H.323 uses H.225.0 as the connection establishment protocol and H.245 as the control protocol between H.323 clients to establish a call, negotiate terminal capability and open logical channels. In the call that involves gatekeepers, a terminal that wants to set up a call first sends a request to its GK. After the GK exchanges message with the other GK, the GK responds to the call originator (caller) with the address of the desired destination (callee). The caller then sends the request directly to the callee. The callee then asks its GK for permission after sending a call proceeding message to the caller. Once the permission is granted by its GK, the callee responds to the caller. After the negotiation between the end points, the media channels are opened and the terminals then exchange media data using RTP/RTCP.

To provide terminal mobility for H.323-based IP telephony, [11] proposes leveraging the dynamic join and departure of multipoint conference with location update. It supports mobility for H.323 terminals without adding new entities, and with minimal modifications to the standard.

2.2 SIP

SIP (Session Initiation Protocol) is a signaling protocol for IP telephony developed by the IETF. SIP originated out of a lightweight Internet approach and reuses many of the header fields, encoding rules, error codes and authentication mechanisms of HTTP. As with HTTP, SIP operates independently of the packet layer and uses text to encode its messages, which facilitates parsing and helps to simplify debugging.

To place a call in SIP, the call originator (caller) locates an appropriate SIP server, typically by DNS. The caller then sends an INVITE request to the SIP server. The handling of INVITE depends on the nature of the server. In the case of a proxy server, the server queries a location server to find the address of the desired destination (callee), and forwards the INVITE message to the callee. The callee sends an ACK to the proxy server which forwards it back to the caller. All subsequent signaling between caller and callee is then handled through the proxy server.

In the case of a redirect server, the server queries the location server to find the callee, then supplies the location of the callee to the caller. The caller then issues an INVITE message to the callee. All subsequent signaling is direct. SIP can also be used simply from a client to another client. SIP deals with user mobility by relying on location servers, but was not designed to address the issues of terminal mobility. Section 3 discusses a solution that uses SIP to provide terminal mobility.

2.3 DOSA and TOPS

Traditional circuit telephony service ensures that end-to-end resources are reserved before the called party's phone is made to ring. However, both H.323 and SIP are signaling protocols without any consideration of resource reservation and QoS guarantees. DOSA (Distributed Open Signaling Architecture), which is a framework for call signaling and resource management for IP telephony proposed by AT&T, provides similar service to conventional telephony. In DOSA, telephony clients, such as PCs or multimedia terminal adaptors, participate in end-to-end capability negotiation, call signaling, and resource reservation. DOSA incorporates an explicit coordination between the call signaling protocol and the resource management protocol such that users are authenticated and authorized before receiving access to the enhanced QoS associated with the telephony service.

TOPS (Telephony Over Packet networkS) also proposed by AT&T, on the other hand, allows users to use mobile terminals or to move between terminals but still can be reached by the same user distinguishing name. Users can control how calls are routed to them. Terminals in TOPS have a range of capabilities to support different media. The major elements of TOPS include a directory service, an application layer signaling (ALS) protocol, a logical channel (LC) abstraction, and mechanisms to support a variety of conferencing modes. The directory service provides flexible mapping of a user's name to a set of terminals where the user can be reached. The ALS negotiates capabilities, establishes and maintains call state, and support advanced features. The LC abstraction isolates the applications from the underlying networks. The TOPS system has been implemented and is being used experimentally within AT&T Labs - Research.

2.4 ITSUMO Signaling

ITSUMO (Internet Technologies Supporting Universal Mobile Operations), a Telcordia and Toshiba joint research project, envisions an end-to-end wireless/ wireline IP platform for supporting future real-time and non-real-time multimedia services. The goal is to use IP and third (and above) generation wireless packet technologies to design a wireless platform that allows mobile users to

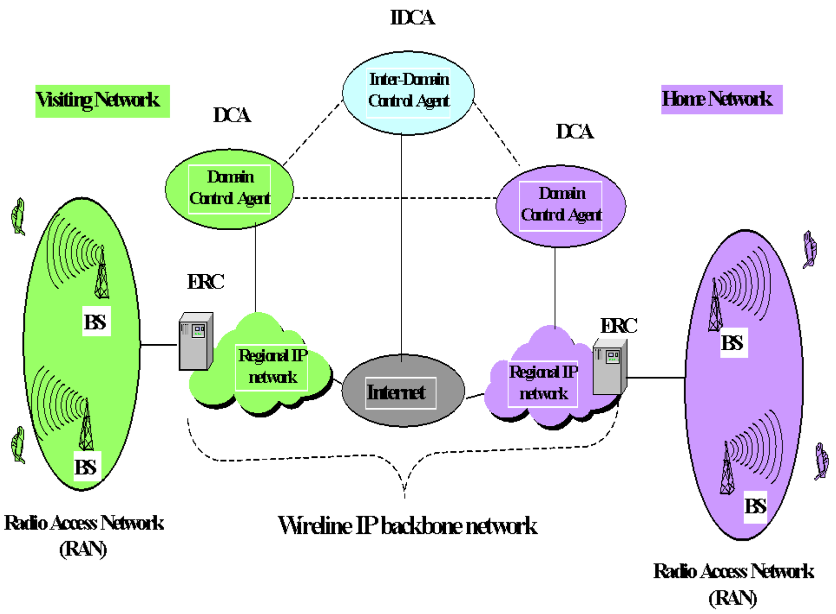

Fig. 2. ITSUMO's long term network architecture

access all services on a next generation Internet. Fig. 2 depicts the end-to-end packet platform of ITSUMO's all IP wireless/wireline network, which comprises all IP wireless access networks and an IP backbone network. The backbone network is an end-to-end wireline IP infrastructure that will comprise regional providers' wireline networks that are connected through the Internet. Besides mobile stations/terminals, a wireless access network also comprises a radio access network (RAN), and an edge router and controller (ERC). In order to support the needs of its users, a wireless access network interacts with the network control entities that are shown as Domain Control Agents (DCA) in Fig. 2.

The long-term control architecture of ITSUMO is shown in Fig. 3. It uses SIP for its end-to-end control (e.g., DCA, IDCA, etc.) architecture. All mobile stations (MSs) and fixed hosts have SIP user agents that interact with the SIP servers (i.e., proxy servers, redirect servers, and registrar) of the DCA/IDCA within the network. In Fig. 3, the SIP server entity of a regional IP network DCA represents the set of SIP proxy and SIP redirect servers within the regional network that perform the network control and signaling functions. Similarly, the registrar represents the server and database (or set of servers and databases) that accepts (accept) SIP REGISTER requests and provides (provide) location services that are similar to those of the home/visiting location registries (HLR/VLR) in today's cellular telephony. As Fig. 3 shows the MAAAQ (mobility, authentication, authorization, accounting, and QoS) entity is built on top of

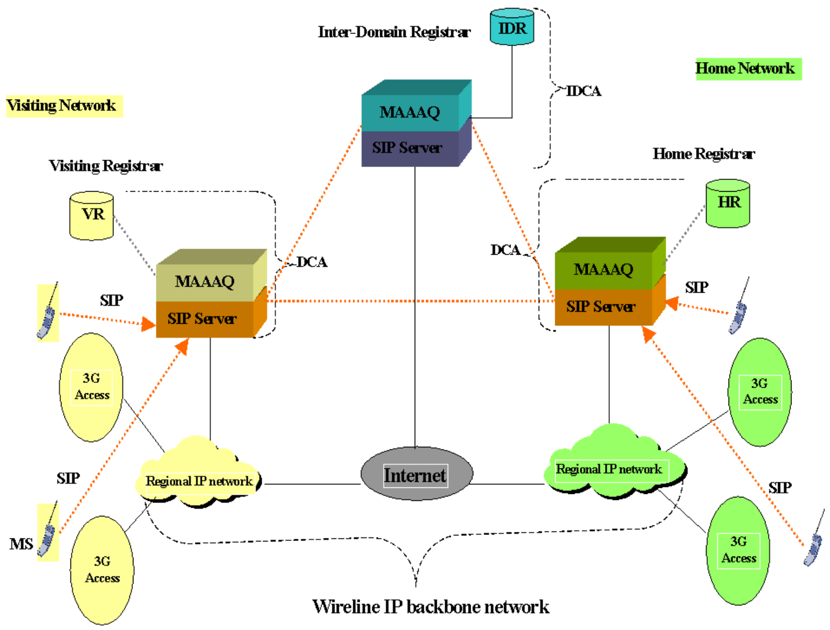

Fig. 3. ITSUMO's long term signaling architecture

a SIP-based DCA (and/or IDCA) system, and uses the location and signaling services of SIP to support roaming users.

3 Mobility

The issue of mobility can be discussed by first considering its implementation in the telephone system, and then user and terminal mobility in the Internet.

3.1 Mobility in the Telephone System

In a telephony system, a fundamental requirement is to be able to locate a user in order to establish real-time communication. Conventional telephony does not provide a means of locating a user per-se, but rather assumes a relatively static association of a fixed terminal with each user, and relies on the caller to consult a directory to map user names to terminal numbers (addresses). The user being called can temporarily circumvent this scheme by enabling the call forwarding feature, allowing all calls addressed to a given terminal to be redirected to an alternate terminal. In most telephone systems it is also possible to get a so-called personal number - a telephone number associated with an individual user rather than a given terminal. The system maintains a dynamic mapping between the personal telephone number and the number of a specific telephone. This mapping is updated explicitly by the user to whom the number is assigned.

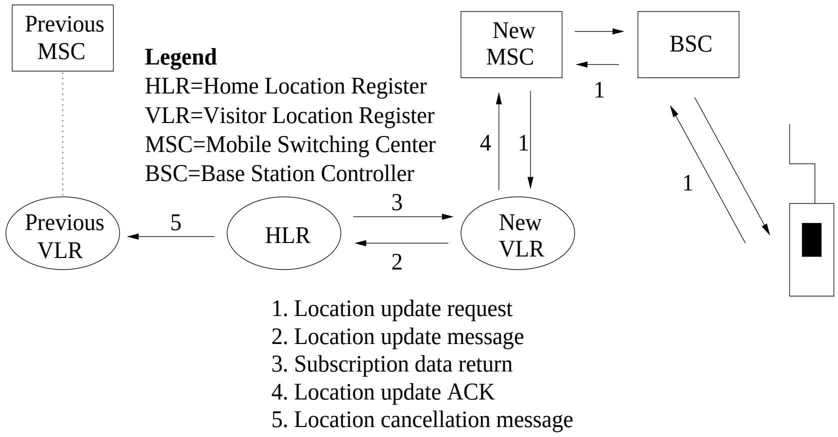

Fig. 4. Mobility in the Cellular Telephone System.

Cellular telephony uses the same basic model as conventional telephony except of course that terminals are mobile and allowed to roam between coverage areas (cells). To establish a call, the system must first locate the current location of the target terminal. This is achieved by accessing a database at the "home" location of the terminal, as determined by the terminal number. This database, usually called the Home Location Register (HLR), maintains an entry identifying the terminal's last recorded location. Each coverage area has its own database that maintains entries for all visiting terminals - usually known as the Visitor Location Register (VLR). Fig. 4 shows the steps involved in this terminal location process. As a terminal moves from one coverage area to another in a process known as handoff, an entry is created in the new VLR, the entry in the previous VLR is removed, and the HLR is updated to point at the new coverage area. Should the terminal be involved in a call when the move between coverage areas occurs, the handoff mechanism arranges for the voice connection to be transferred across to the new cell.

3.2 User Mobility in the Internet

In the Internet, the association between a user and a terminal number is not assumed to be static. Users expect to be able to access the Internet from any computer, and it is commonly the case that the terminal number (IP address) assigned to a given terminal is generated dynamically using the Dynamic Host Configuration Protocol (DHCP). Thus, to locate the address of the user's current terminal requires that a dynamic binding be maintained and consulted whenever one wishes to locate that user. Early VoIP products relied on the use of a special rendezvous service, such as an Internet Relay Chat (IRC) server, to allow users to browse the list of users online, and search for the current IP address associated with a user. Each time the VoIP client software executed it took care of the registration procedure on behalf of the user.

More recent approaches include the H.323 Gatekeeper, which provides a mapping between a user alias such as an email-address and the corresponding transport address. Use of an email address to identify a user provides a naming space which is convenient, scalable and can provide unique identification. H.323 client software can take care of registering the user with a Gatekeeper. The SIP Location Server provides the same function, but in addition to explicit registration, the server can also be configured to use other techniques to try to locate a user, including for example the user of the finger utility to probe a set of likely computers. The TOPS directory service allows user registration, but in addition supports the association of more than one terminal with a given user, and call processing logic to select amongst different terminals based on attributes such as caller identification, time-of-day and media types desired.

3.3 Internet Terminal Mobility

Host mobility refers to the situation where a terminal can be moved from one location to another. There are two cases to consider. Nomadic computing is when a terminal is used in many different locations, but in each case obtains a local IP address via DHCP. The salient point here is that any entries in user directory servers must be updated with the new address. The other model is mobile computing, where a terminal is allowed to retain its IP address no matter in which location it happens to be operating. This is the problem that the Mobile-IP protocol [12] solves.

Mobile-IP introduces the concept of a terminal having two IP addresses: a home IP address and a "care-of" IP address. The home IP address refers to the address the terminal is assigned in its home subnet. The care-of address is a temporary IP address assigned for use by the terminal while it is visiting another subnet, so it changes each time the terminal moves from one subnet to another. The basic idea is as follows. When a terminal moves outside its home subnet to another subnet it first acquires a local care-of address, and registers that address with a Home Agent back on its home subnet. Packets addressed to the terminal continue to arrive on the home subnet, are picked up by the home agent and forwarded to the terminal at its current care-of address (possibly via a Foreign Agent). This is shown in Fig. 5. When a terminal moves between subnets there is a handoff procedure that takes care of updating registration information at the Home Agent.

Several performance issues arise when considering the use of Mobile-IP for supporting VoIP. First there is the issue of end-to-end latency, which for voice telephony must be tightly constrained. In the basic proposal for Mobile-IP, all packets destined for a mobile host travel via the host's home subnet - resulting in a situation known as triangular routing. Clearly this can result in significant increases in the end-to-end delay for a VoIP call. A solution to triangular routing exists as a route optimization protocol, which allows so-called corresponding hosts to learn the current care-of address of the target host. This protocol is designed as an optimization and can result in latency which is highly variable.

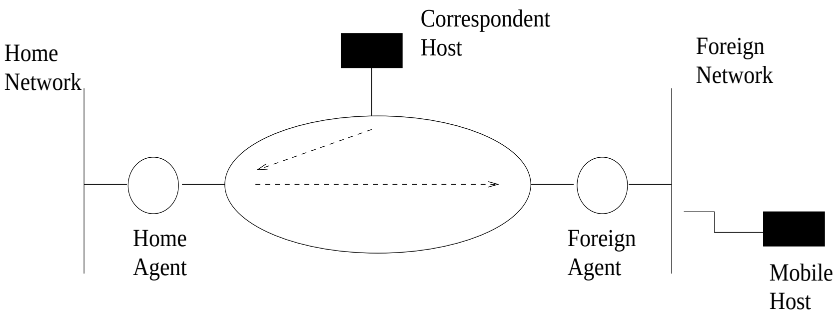

Fig. 5. Mobile-IP operation.

Another performance issue is the latency of handoff in Mobile-IP. Because of the need to interact with the Home Agent in order to perform a handoff, the latency involved can be significant. There have been several proposals designed to improve the performance of handoff such as [2]. This proposal introduces a hierarchy which allows several coverage areas to be managed as a group, within which host movement is invisible to the Home Agent and correspondent hosts. Related to that is the Cellular IP work [3] which suggests the use of location management techniques similar to those used in the cellular telephone system for providing a scalable mobility management solution with Mobile-IP. Finally, [13] proposes using SIP to provide low-latency mobility for terminals involved in multimedia sessions. In that solution SIP signaling messages are re-routed as usual based on the operation of Mobile-IP, but the media streams are re-routed at the application layer by having SIP renegotiate the various streams to use the care-of address.

4 Service Quality

In addition to mobility-related performance issues as discussed above, VoIP communication involving mobile terminals requires two key service quality issues to be addressed: the ability to schedule voice and non-real time data on a wireless access medium, and protocols to provide resource reservations in situations where one or more terminals can be mobile.

4.1 Medium Access

The issue of providing scheduled medium access has of course received considerable interest in the research literature, both for wireline and wireless networks. Popular wireless local area networks have followed a contention-based access model comparable to that of an Ethernet network, making them generally unsuitable for multimedia traffic. For example, the IEEE 802.11 standard for wireless LANs implements carrier sense multiple access (CSMA) as its default mode of operation, known as the distributed coordination function (DCF). Just as with Ethernet a consequence of this approach is an inability to predict access latencies, since they are dependent on the current offered load. The 802.11 standard also provides an optional mode of operation called the point coordination function (PCF) which introduces an access controller for each coverage area. The access controller implements a contention-free access period in addition to the contention-based period of the default mode. In the contention-free period the access controller polls those hosts which have indicated a desire to have predictable access, giving them a regular opportunity to access the medium.

A different solution is exemplified by the Wireless Integrated Services Platform (WISP) [4]. WISP supports a mix of constant bit rate voice and best-effort data by employing a medium access protocol based on token passing. In essence, a controller in each coverage area runs an algorithm which, on request from a terminal, grants access to the medium by passing a token to the terminal. For voice traffic, the terminal exchanges messages with the controller to specify its bandwidth requirements and hence the rate at which it will require tokens, allowing the controller to perform admission control.

4.2 Resource Reservation

Multiple service classes in an IP network can be provided by the resource reservation protocol (RSVP). Based on the notion of a data flow, RSVP allows a sender to describe the traffic it plans to transmit, and receivers the class of service they desire. This information is used to perform admission control and establish flow state at each of the intermediate routers. In RSVP, this information is soft-state, i.e. it times out and is discarded if not regularly refreshed. A key problem that arises is how RSVP works in the presence of terminals that are mobile. In particular, having established a reservation, if either or both sender and receiver move to a different network then a new reservation is required. This has performance implications relating to the delay of establishing a new reservation. It also raises the possibility of a terminal moving into a coverage area in which the desired resources are not available.

An extension to RSVP has been proposed to deal with these issues. Known as Mobile-RSVP [1], the idea is to define a reservation as being either mobility independent or mobility dependent. For mobility independent service it is necessary to make spatial reservations, i.e. at each of the locations that a mobile terminal might visit during the lifetime of the flow. This set of locations must be specified by the subscriber as part of a mobility specification. The scheme distinguishes between two types of reservation: an active reservation to the current location of the terminal, and passive reservations to all others cells in the mobility specification. A passive reservation becomes an active reservation when the terminal enters the corresponding coverage area. Subscribers to the mobility dependent service make a reservation from the sender only to their current cell. To avoid low utilization of the network, the resources associated with passive reservations are allowed to be used by other flows as if they were unreserved.

5 Conclusions

The use of IP to transport voice is receiving significant attention in the telecommunications and data networking industries. Commonly known as Voice-over-IP (VoIP), this technology has the potential to allow user communication involving multiple media types, and between terminals that offer improved capabilities and user interfaces. Meanwhile there has been an explosive growth in the demand for wireless connectivity, raising the question of how VoIP can be extended to situations where users and terminals can be mobile. In this paper we have surveyed the technologies necessary to provide mobile VoIP. It is clear that mobility has serious implications for locating users in order to establish a call, and maintaining connectivity with appropriate QoS throughout a call, even in the presence of mobility. Mobility also has an impact on call signaling, especially in relation to having to adapt call characteristics and media destinations as a terminal moves. We have discussed these key issues and given direction towards possible solutions.

Designing a network to provide wireline VoIP alongside other services is in itself a very difficult task, and introduces a host of technical and operational problems. In a mobile VoIP system there are additional issues of tracking users that can roam between different autonomous systems, and providing seamless handoff between wireless access infrastructures that are owned and operated by different organizations. These remain complex issues that need to be resolved within the overall goal of providing multi-service networks where voice telephony is just one of a set of services all based on IP delivery to the end-terminal.

References

1. Badrinath, B., and Talukdar, A. IPv6 + MOBILE-IP + MRSVP = Internet Cellular Phone? In *Proc. of IFIP Int. Workshop on Quality of Service* (May 1997), pp. 49–52.
2. Caceres, R., and Padmanabhan, V. N. Fast and scalable wireless handoffs in supports of mobile Internet audio. *ACM/Baltzer Mobile Networks and Applications 3*, 4 (1998), 351–363.
3. Campbell, A., Gomez, J., Kim, S., Wan, C., and Valko, A. A cellular IP testbed demonstrator. In *Proc. of the 6th IEEE International Workshop on Mobile Multimedia Communications (MOMUC)* (1999), pp. 145–148.
4. Goel, S., Mishra, P., Saran, H., and Sreenan, C. Design and evaluation of a platform for mobile packet telephony. In *Proc. of IEEE International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)* (July 1998), pp. 71–81.
5. Goyal, P., Greenberg, A., Kalmanek, C. R., Marshall, W. T., Mishra, P., Nortz, D., and Ramakrishnan, K. K. Integration of call signaling and resource management for IP telephony. *IEEE Network 13*, 3 (May/June 1999), 24–32.
6. Handley, M., Schulzrinne, H., Schooler, E., and Rosenberg, J. SIP: session initiation protocol. IETF RFC 2543, Mar. 1999.
7. ITSUMO Group. Benchmarking of ITSUMO's all IP wireless architecture. Mobile Wireless Internet Forum (http://www.mwif.org/) <mwif2000.028.0>, Jan. 2000.

8. ITU-T REC. H.323. Packet-based multimedia communications systems, Oct. 1997.
9. KALMANEK, C., KAPLAN, A., MARSHALL, W., MISHRA, P., ONUFRYK, P., RAMAKRISHNAN, K., AND SREENAN, C. TOPS: an architecture for telephony over packet networks. *IEEE Journal on Selected Areas in Communications 17*, 1 (Jan. 1999), 91–108.
10. KANTER, T., OLROG, C., AND JR, G. Q. M. VoIP for wireless and mobile multimedia applications. In *Proc. of the 1999 Personal Computing and Communications Workshop* (Nov. 1999), pp. 141–144.
11. LIAO, W. Mobile Internet telephony: mobile extensions to H.323. In *Proc. of IEEE INFOCOM* (New York, NY, Mar. 1999), pp. 12–19.
12. PERKINS, C. IP mobility support. IETF RFC 2002, Oct. 1996.
13. WEDLUND, E., AND SCHULZRINNE, H. Mobility support using SIP. In *Proc. of 2nd ACM International Workshop on Wireless Mobile Multimedia* (1999).