Mobility Support using SIP

Elin Wedlund Ericsson elin.wedlund@netinsight.se
Henning Schulzrinne Columbia University hgs@cs.columbia.edu

ABSTRACT

Enabling mobility in IP networks is an important issue for making use of the many light-weight devices appearing at the market. The IP mobility support being standardized in the IETF uses tunnelling of IP packets from a Home Agent to a Foreign Agent to make the mobility transparent to the higher layer. There are a number of problems associated with Mobile IP, such as triangular routing, each host needing a home IP address, tunnelling management, etc. In this paper, we propose to use mobility support in the application layer protocol SIP where applicable, in order to support real-time communication in a more efficient way.

1 INTRODUCTION

The IETF has standardized IP mobility support [1] which provides for *transparent mobility*, in that it hides the change of IP address when the mobile host is moving between IP subnets. This is needed to keep TCP connections alive when a host moves from subnet to subnet. However, mobile IP is struggling with the problem of triangular routing, i.e., a packet to a mobile host travels via the home agent, whereas a packet from the mobile host is routed directly to the destination. The route optimization [2] solves this by sending binding updates to inform the sending host about the actual location of the mobile host. This solution has several problems, as will be discussed in Sec. 2. For real-time traffic such as voice or video over IP, it is more common to use the Real-Time Transport Protocol (RTP) [3] over UDP, and important issues are fast handoff, low latency, and – especially for wireless networks – high bandwidth utilization. Therefore, we see a need to introduce mobility awareness on a higher layer, where we can utilize knowledge about the traffic to make decisions on how to handle mobility in different situations. The application layer protocol Session Initiation Protocol (SIP) [4] already supports personal mobility[1], and the changes needed to support device mobility are minor. In this document, we will discuss how mobility support in SIP can improve the performance for realtime services in wireless networks, and propose an architecture for how this can be done. Also, application-layer mobility does not require any changes to the operating system of any of the participants and thus can be deployed widely much easier than mobile IP.

Throughout this document "mobile IP" refers to IP mobility support as defined in [1].

WoWMoM 99 Seattle WA USA

2 IP MOBILITY SUPPORT

Mobile IP has been proposed as a solution for mobility support in IP networks. A well-known problem with mobile IP is the triangular routing which adds delay to the traffic towards the mobile host, but not from the mobile host, see Fig. 1. Measurements [6] show that mobile IP increases the latency by 45% within a campus, which can be expected to increase in a wide area network, when the distance increases between the different entities. These numbers are also highly dependent on the mobile IP implementation, and the capacity of the home agent and foreign agent. For delay sensitive traffic, this is not acceptable, because we cannot afford a higher latency in the network than what is absolutely necessary. The fact that the packets are tunnelled also means that an overhead of typically 20 bytes (IP in IP encapsulation [7]) will be added to each packet. Compare this to the packet size for an audio packet, which is around 60 bytes including IP, UDP, and RTP headers, if the coder's bitrate is 6 kb/s.

Route optimization solves the triangular routing problem by using binding updates to inform the correspondent host about the current IP address. However, route optimization has several drawbacks:

- Route optimization requires changes in the IP stack of the correspondent host, since it must be able to encapsulate IP packets, and store care-of addresses of the foreign agent or mobile host.

[1] "Personal mobility is the ability of end users to originate and receive calls and access subscribed telecommunication services on any terminal in any location, and the ability of the network to identify end users as they move. Personal mobility is based on the use of a unique personal identity (i.e. 'personal number')." [5, p. 44].

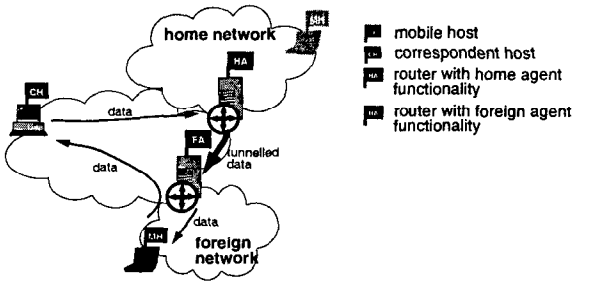

Figure 1: Mobile IP

- Only the home agent may send binding updates to correspondent hosts. This means that there will be an extra delay before the correspondent host finds out where to send the packets, during which the old foreign agent must forward the packets to the correct location.
- The mobile host needs to rely on the old foreign agent forwarding packets to its new foreign agent until the correspondent host has got the binding update. There is no requirement saying that the foreign agent must do so.
- The binding warnings and updates are not compulsory, and should be used sparingly, since it can be expected that many hosts will not support the binding update function.

Because of the requirements that are put on the correspondent hosts, it cannot be expected that route optimization will be widely employed in a near future. Moreover, the home and foreign agent can become bottlenecks since they must handle the tunnels for a possibly large number of mobile hosts. Another issue is that the mobile host needs a permanent home IP address, which can be a problem due to the address exhaustion in IP version 4. One issue still stands, though: Mobile IP provides transparent mobility which is needed to keep TCP connections alive as the user is moving. The solution suggested in this paper is to use mobile IP for long-lived TCP connections, e.g. telnet, ftp, irc, etc., but to use a more appropriate mobility support for real-time communication.

3 SIP MOBILITY SUPPORT

3.1 Introduction to SIP

The Session Initiation Protocol (SIP) [4] is an application-layer protocol used for establishing and tearing down multimedia sessions, both unicast and multicast. It has been standardized within the Internet Engineering Task Force for the invitation to multicast conferences and Internet telephone calls [8]. Entities in SIP are user agents, proxy servers and redirect servers. A user is addressed using an email-like address "user@hos", where "user" is a user name or phone number and "host" is a domain name or numerical address. SIP defines a number of methods, listed in Table 1. Responses to methods indicate success or failure, distinguished by status codes,

Message Name	Function
INVITE	Invite user(s) to a session. The session description is contained in the body of the message, e.g. using the Session Description Protocol (SDP) [9]. The session description contains the address where the host wants to receive media streams.
ACK	Acknowledgment of an INVITE request.
BYE	Sent when a call is to be released.
OPTIONS	Query server about capabilities.
CANCEL	Cancel a pending request.
REGISTER	Register with a SIP server.

Table 1: SIP Requests

1xx (100 to 199) for progress updates, 2xx for success, 3xx for redirection, and higher numbers for failure. Each new SIP transaction has a unique call identifier, which identifies the session. If the session needs to be modified, e.g. for adding another media, the same call identifier is used as in the initial request, in order to indicate that this is a modification of an existing session.

The SIP user agent has two basic functions: Listening for incoming SIP messages, and sending SIP messages upon user actions or incoming messages. The SIP user agent typically also starts appropriate applications according to the session that has been established. The SIP proxy server relays SIP messages, so that it is possible to use a domain name to find a user, rather than knowing the IP address or name of the host. A SIP proxy can thereby also be used to hide the location of the user. A redirect server returns the location of the host rather than relaying the SIP message. This makes it possible to build highly scalable servers, since it only has to send back a response with the correct location, instead of participating in the whole transaction which is the case for the SIP proxy. Both the redirect and proxy server accepts registrations from users, in which the current location of the user is given. The location can be stored either locally at the SIP server, or in a dedicated location server (more about the location server further below). Deployment of SIP servers enables personal mobility, since a user can register with the server independently of location, and thus be found even if the user is changing location or communication device. SIP requests and responses are generally sent using UDP, although TCP is also supported. A typical signalling case using a redirect server is shown in Fig. 2.

The INVITE message contains a session description expressed in SDP, and is received by a redirect server, which consults a location server to find out where to redirect the invitation. The function of the location server is not specified, but can be anything that can return a next hop address in the chain of finding the callee (which could be an address to another redirect server or a proxy). In many cases, the location server can simply be a table handled by the SIP server, containing the users' locations as they register with the SIP server. In Fig. 2, the location server returns the current address of the callee. A proxy server would, instead of redirecting the invitation, forward it to the callee. From now on, only redirect servers will be discussed, but this does not mean that a proxy server cannot be used instead. However, the load on a redirect server can be expected to be lower since it only needs to send an answer with the user's location, instead of participating in the whole signalling transaction. The SIP redirect server has properties resembling those of the home agent in mobile IP with route optimization, in that it

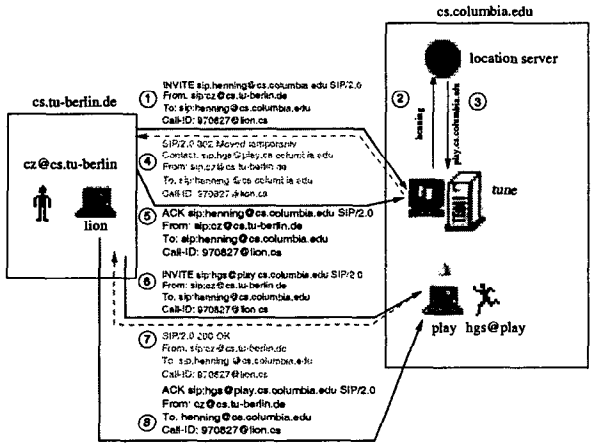

Figure 2: SIP transaction in redirect mode

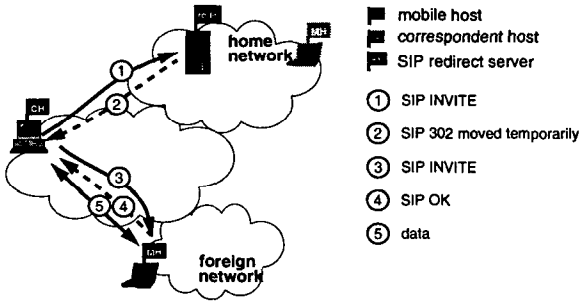

Figure 3: SIP mobility: setting up a call

tells the caller where to send the invitation. In addition, it can store preferences for the user regarding how to treat incoming requests depending on where the user is located, time of day, or the identity of the caller.

3.2 SIP Mobility Support

As was stated in the previous section, SIP supports personal mobility, i.e., a user can be found independent of location and network device (PC, laptop, IP phone, etc.). The step from personal mobility to IP mobility support is basically the roaming frequency, and that a user can change location (IP address) during a traffic flow. Therefore, in order to support IP mobility, we need to add the ability to move while a session is active. It is assumed that the mobile host belongs to a home network, on which there is a SIP server (in this example, a SIP redirect server), which receives registrations from the mobile host each time it changes location.This is similar to home agent registration in Mobile IP. Note that the mobile host does not need to have a statically allocated IP address on the home network. When the correspondent host sends an INVITE to the mobile host, the redirect server has current information about the mobile host's location and redirects the INVITE there (see Fig. 3)[2].

If the mobile host moves during a session, it must send a new INVITE to the correspondent host using the same call identifier as in the original call setup, see Fig. 4. It should put the new IP address in the Contact field of the SIP message, which tells the correspondent host where it wants to receive future SIP messages. To redirect the data traffic flow, it indicates the new address in the SDP field, where it specifies transport address.

The SIP INVITE (step 1 in Fig. 4) request could look as follows:

```
INVITE sip:alice@correspondent.com SIP/2.0
Via: SIP/2.0/UDP mh.current.location:5060
From: sip:betty@home.com
To: sip:alice@correspondent.com
Subject: a mobile session
Contact: betty@mh.current.location
```

[2]For conciseness, the ACK message, needed to confirm the receipt of the response, is left out in the figures throughout the rest of this paper.

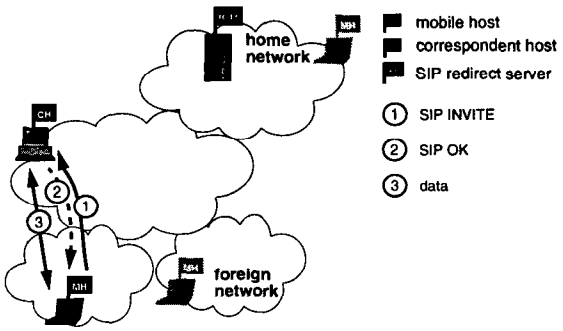

Figure 4: SIP mobility: mobile host moves

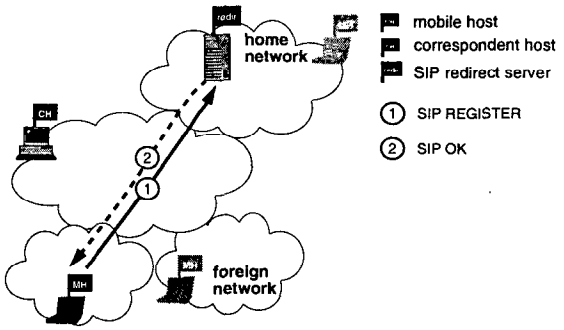

Figure 5: Mobile host registration

```
CSeq: 781769870 INVITE
Call-ID: <call-id of ongoing session>
Content-Length: ...
v=0
o=betty916340046 916340046 IN IP4 mh.current.location
t=2208988800 2208988800
c=IN IP4 mh.current.location
m=audio 50000 UDP 0
```

Betty owns the mobile host, and Alice is the user at the correspondent host. Betty's regular address (betty@home.com) is used in the From field, since that is used for identification, and can also be used as a fall back mechanism in case the communication is lost (more discussion on this in Section 3.2.2). The new address (mh.current.location) is put in the Contact field of the SIP message, and in the "c=" field of the session description (SDP) part of the message. Finally, the mobile host should update its registration at the home SIP server, so that new calls can be correctly redirected, see Fig. 5.

3.2.1 SIP Mobility Support with Mobile IP

If the mobile host is using mobile IP, it is not necessary, albeit useful, for the SIP server to have knowledge about the current location of the mobile host. It is however a waste of resources to keep duplicate information about the host's current address - both in the SIP server and in the home agent. One solution to avoid duplicate information is to co-locate the SIP redirect or proxy server and the home agent, or to allow the SIP server to query the home agent about the location of the mobile host. It would also be possible to actually send the invitation to the home address, let the home agent forward the invitation to the correct location, and let the mobile host provide information about its location in the response, using the Contact header.

3.2.2 Error Recovery

If the correspondent host for some reason has an outdated address of the mobile host, it must have a fall-back mechanism to break the error situation. One example of this is when we have two mobile hosts having a conversation; both lose contact for a while (e.g., by driving through a tunnel), and when they gain contact again, they have both got a new IP address, as illustrated in Fig. 6.

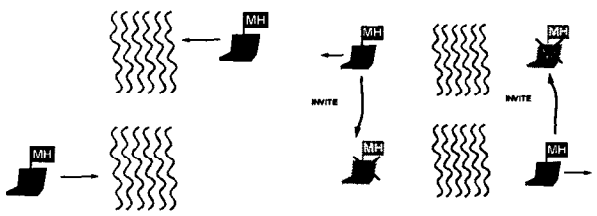

Figure 6: Stale address

In order to avoid situations like this, a host can send retransmissions of invitations also to the SIP server on the mobile host's home network. Since the SIP server has a fixed address, the mobile host can always send registrations to it. In this fashion, the correspondent host can re-locate a mobile host that has been lost. (If mobility rates are high and thus simultaneous mobility common, the MH could decrease the hand-off delay at the cost of additional packets, by sending an INVITE at the same time to the last known IP address of the MH and the home registrar.)

3.2.3 Security

In the SIP specification there is support for both authentication and encryption of SIP messages, using either challenge-response or private/public key cryptography.

4 PROPOSED ARCHITECTURE

By introducing SIP mobility support, we will avoid many of the problems with Mobile IP. However, SIP mobility cannot support TCP connections, so the best solution would be to use SIP mobility for realtime communication over UDP, and Mobile IP for TCP connections. This can be achieved if we allow the mobile host to choose when to use its home address or care-of address. When sending RTP streams it will use the care-of address, and when establishing TCP connections, it will use the home address and let the traffic be routed via the home agent. It may also use route optimization for the TCP connections. What we propose is a similar solution to the one presented in [6], which is to use mobile IP when necessary for long-lived TCP connections such as telnet, irc, ftp[3], etc.

However, for many of those applications that are likely to be used on mobile Internet terminals, even TCP applications may be quite usable without mobile IP underneath. These service include all those applications that are built on recoverable short transactions, such as web browsing, SMTP mail upload and POP or IMAP mail retrieval. In those cases, the TCP connections are usually short enough to make the cost of having to re-attempt an operation (e.g., an HTTP transaction that was aborted due to hand-off or the download of a single email) relatively small on average. The protocols mentioned also have application-layer recovery, since they are designed to operate in dial-up environments subject to sudden disconnects. (A smart application would know from receiving a TCP RST

[3]Even for ftp, application-layer recovery from address changes is supported in modern versions of ftp.

that it should retry the operation.) For real-time voice and video connections, our architecture supports mobility is supported by SIP.

The protocol stack is described in Fig. 7. As in [6], a mobile policy table is used for deciding what source address to use (home or care-of address), whether it should be tunnelled, or even use a bidirectional tunnel.

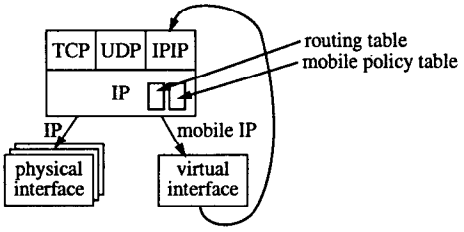

Figure 7: Protocol stack

An example of an MTP is shown in Table 2. In this table, we specify that telnet and ftp traffic should use mobile IP, and that all other traffic should not. For instance, web traffic will not use mobile IP, because we assume that the TCP connections are short, and will rarely be affected by a handoff.

5 IMPLEMENTATION

In order to provide SIP mobility in a network, a number of actions need to be taken. These actions are different depending on the type of device: mobile or non-mobile host, and SIP server.

5.1 Non-mobile Host

Each host that is to communicate with a mobile host, needs to support SIP. This is relatively simple, since SIP is an application level protocol, and can thus be downloaded and installed when it is needed. SIP clients complying with the SIP specification should react correctly to messages from a mobile SIP client.

5.2 Mobile Hosts

The SIP client on the mobile host needs to be integrated with the other mobility mechanisms, e.g. the driver for the wireless network card, and a DHCP [10] client (to obtain an IP address). When the host detects a new beacon from a base station, it should first check whether the new base station is on the same IP subnet or not. If the base station is on a new subnet, the host must acquire a new IP address, and initiate the SIP mobility.

Other events than beacons can trigger the SIP mobility mechanism, e.g. when a PCMCIA card is inserted to a laptop, or when a device is switched on (although this will probably not require a handoff). This erases the border between personal mobility and device mobility, since there is no difference whether the device is moving or the user is moving.

5.3 SIP Servers

There is no extra functionality needed in the SIP proxies or redirect servers. However, the load on the servers can be expected to be higher, since the mobile hosts will make a new registrations each time they change addresses. Some ideas on how to handle this scaling issue is discussed under future work in Sec. 8.2.

6 PERFORMANCE

It is not trivial to compare the performance of mobile IP vs. SIP mobility, because it very much depends on the distance between the mobile host, correspondent host, and the mobile host's home network.

6.1 End-to-End Delay

It is obvious that the end to end delay will be lower if packets are sent directly to the mobile host without being routed via the home network and/or being encapsulated. The extra latency introduced by mobile IP is basically proportional to the distance to the home network and the correspondent host. The delay introduced by the home agent and foreign agent are relatively small unless a congestion occurs and packets are buffered.

6.2 Handoff Delay

The handoff delay depends on the delays of several different operations:

- Both mobile IP and SIP-based mobility need to discover that they are in a new network. This depends on the wireless technology and the operating sytem interface of, say, a wireless LAN card.
- Then, a host needs to acquire an IP address via DHCP, which, depending on implementation [11], can be a major part of the overall handoff delay. A mobile-IP host needs to instead discover its new FA. The number of messages exchanged is roughly similar for either DHCP or FA discovery.
- A mobile IP host then has to register with the foreign and/or home agent (which in turn notifies the CH if route optimization is used), while a SIP-speaker needs to send an INVITE to the correspondent host, thus incurring misdirected packets for the one-way delay from MH to CH. Generally, the path from MH to HA to CH is going to be longer, possibly significantly so, than the direct path between CH and MH. This is a particular problem if the paths between MH and HA or HA and CH suffer from high packet loss, since that would significantly delay the binding update. The magnitude of this effect clearly depends on the relative location of the CH, MH and HA and thus can't be quantified with any generality. However, the typical one-way delay of about 20 to 50 ms[4] within the continental United States is roughly equivalent to the packetization interval for packet voice, so that no more than one or two packets will get lost due to handoff. This amount of packet loss can be compensated for by either forward-error correction [12, 13] or codec-level packet loss hiding [14], which are

[4] on links which have sufficient quality for real-time voice communications

dest. address	netmask	port	mobile IP	bidir.	Comment
0.0.0.0	0.0.0.0	23	yes	no	all telnet traffic should use mobile IP
0.0.0.0	0.0.0.0	21	yes	no	all ftp traffic should use mobile IP
0.0.0.0	0.0.0.0	0	no	no	all other traffic does not use mobile IP

Table 2: Example mobile policy table

likely to be needed in any event for "normal" Internet packet loss. Thus, as long as end-to-end delays of a round-trip time are acceptable, application-layer mobility as described here should be able to provide transparent mobility, even without lower-layer assistance (such as soft hand-offs).

In addition, for SIP mobility, the mobile host must register with the SIP server on the home network, although this does not factor into the handoff delay. In summary, we see that the difference in handoff delay is the same difference as mobile IP has for data packets between the current version and the use of route optimization.

Handling mobility at the application layer does introduce slight additional delays since an operating system context switch in the end host is required, but in modern OS, these are generally below one millisecond.

7 RELATED WORK

Much work has been done in the area of IP mobility support. The work that has had the most impact on this paper is the Mosquitonet project at Stanford [6]. In this project they are providing the Mobile Policy Table (MTP) which is used to define when mobile IP is to be used. However, other mobility support than mobile IP is not used in this project.

In [15], cellular IP is proposed to be used for mobility within IP subnets, in order to do faster handoffs. This could be used together with the SIP mobility and/or mobile IP. Foo and Chua [16] propose regional aware foreign agents for faster handoffs.

8 FUTURE WORK

8.1 Implementation

The SIP mobility support is currently being implemented. The mobile host has a mobile host daemon for mobile IP, which is the one implemented in the Mosquitonet project, and a SIP client based on tcl/tk. Moreover, the host is equipped with a wireless interface, and is using DHCP for getting an IP address. It would be desirable to also include a lower layer support, e.g., the cellular IP implementation [15] together with the SIP mobility and mobile IP in order to have a complete solution.

8.2 Hierarchical Registration Servers

When the mobile host is far away from its home network, sending a new registration to the home SIP server every time it moves can place an unnecessarily high load on the SIP server and network, especially if the home SIP server is serving many hosts. Instead, the mobile host can register with a closer SIP server, and the SIP server on the home network knows to which SIP server it should forward/redirect an incoming request. The basic approach is similar to both [15] and [16], but in this case, only the first SIP messages need to be routed via the servers in the hierarchical path.

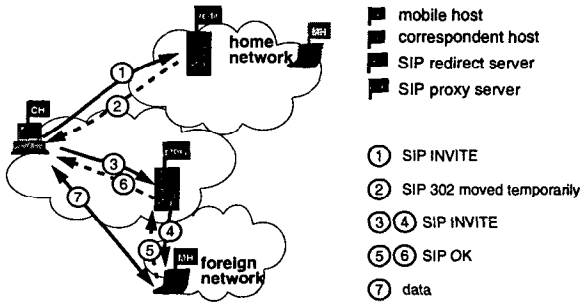

Figure 8: Hierarchical SIP servers

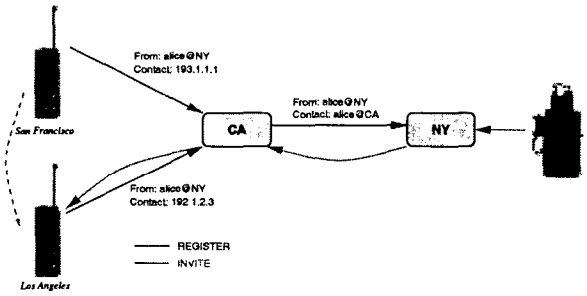

Figure 9: Registration with hierarchical SIP servers

Future SIP messages and the media stream can then be sent directly between the hosts. Fig. 8 shows a case with the home SIP server is a redirect server, and the "foreign" SIP server is a proxy server. This mechanism works without change to the SIP specification since registration requests can be proxied just as any other SIP request. A registration example is shown in Fig. 9, where the MH moves within California. It advertises the California server has its address to the home registrar, so that all INVITE requests from elsewhere go through the California server. The California server does not proxy registration changes within California to the home registrar, making such moves invisible to the home registrar.

9 CONCLUSIONS

This paper has proposed the use of mobility support in an application-layer signaling protocol, SIP, for real-time mobile communication. By moving the mobility handling to the application

layer, we eliminate the need for tunneling of the data stream. Moreover, the fact that SIP mobility is at the application layer, means that it can be installed easily, which allows the most common mobile application, voice, to enjoy mobility before mobile IP is widely deployed.

Another desirable effect of using SIP is that the distinction between personal and device mobility disappears — there is no difference whether I move my computer to a new network, or if I change communication device from my PC to my IP phone. SIP together with SDP has the support to both signal the changing transport addresses as well as the different capabilities of the devices.

Using SIP for mobility is possible without making any changes to the IP stack of the mobile host. If we want to support mobile IP as well, we use a mobile policy table for deciding when to use the home or care-of address in addition to the changes needed to support mobile IP itself. It is important to point out that unless route optimization should be supported, no changes to the kernel are needed for the correspondent host. What applies to both mobile IP and SIP mobility is that none of them is suitable for fast, or small scale mobility. The fast mobility should either be supported by lower layers or by some other, more suitable, scheme. One suggestion is Cellular IP, which can be used together with mobile IP or the SIP scheme.

REFERENCES

[1] C. Perkins, "IP mobility support," Request for Comments (Proposed Standard) 2002, Internet Engineering Task Force, Oct. 1996.

[2] C. Perkins and D. Johnson, "Route optimization in mobile IP," Internet Draft, Internet Engineering Task Force, Feb. 1999. Work in progress.

[3] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: a transport protocol for real-time applications," Request for Comments (Proposed Standard) 1889, Internet Engineering Task Force, Jan. 1996.

[4] M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg, "SIP: session initiation protocol," Request for Comments (Proposed Standard) 2543, Internet Engineering Task Force, Mar. 1999.

[5] R. Pandya, "Emerging mobile and personal communication systems," *IEEE Communications Magazine*, vol. 33, pp. 44–52, June 1995.

[6] X. Zhao, C. Castelluccia, and M. Baker, "Flexible network support for mobility," in *Proc. of Mobicom*, (Dallas, Texas), Oct. 1998.

[7] C. Perkins, "IP encapsulation within IP," Request for Comments (Proposed Standard) 2003, Internet Engineering Task Force, Oct. 1996.

[8] H. Schulzrinne and J. Rosenberg, "The session initiation protocol: Providing advanced telephony services across the internet," *Bell Labs Technical Journal*, vol. 3, pp. 144–160, October-December 1998.

[9] M. Handley and V. Jacobson, "SDP: session description protocol," Request for Comments (Proposed Standard) 2327, Internet Engineering Task Force, Apr. 1998.

[10] R. Droms, "Dynamic host configuration protocol," Request for Comments (Draft Standard) 2131, Internet Engineering Task Force, Mar. 1997.

[11] J.-O. Vatn and G. Q. M. Jr., "The effect of using co-located care-of addresses on macro handover latency," in *Proc. of 14th Nordic Tele-traffic Seminar*, (Technical University of Denmark, Lyngby, Denmark), Aug. 1998.

[12] J. Rosenberg and H. Schulzrinne, "An RTP payload format for generic forward error correction," Internet Draft, Internet Engineering Task Force, June 1999. Work in progress.

[13] C. Perkins, I. Kouvelas, O. Hodson, V. Hardman, M. Handley, J. C. Bolot, A. Vega-Garcia, and S. Fosse-Parisis, "RTP payload for redundant audio data," Request for Comments (Proposed Standard) 2198, Internet Engineering Task Force, Sept. 1997.

[14] H. Sanneck, A. Stenger, K. B. Younes, and B. Girod, "A new technique for audio packet loss concealment," in *Proceedings of Global Internet* (J. Crowcroft and H. Schulzrinne, eds.), (London, England), pp. 48–52, IEEE, Nov. 1996.

[15] A. G. Valk, "Cellular IP: A new approach to internet host mobility," *ACM Computer Communication Review*, vol. 29, pp. 50–65, Jan. 1999.

[16] S. Foo and K. Chua, "Regional aware foreign agent (RAFA) for fast local handoffs," Internet Draft, Internet Engineering Task Force, Nov. 1998. Work in progress.