

SIP in Mobile Environments - Applications and Possibilities

Marko Berg
Helsinki University of Technology
Marko.Berg@iki.fi

Abstract

With the recent emergence of a myriad of mobility-enabling technologies, and with many of them gaining foothold all over the globe, the issues of internetworking at the levels of communities and services have attracted a lot of attention. The Internet Protocol (IP) [1] has long ago proven to be able to serve most of the needs on the network layer, and recently all the more so with IPv6. However, the emerging global mobile community needs solutions in the application layer that allow mobile devices to access the same services and join their chosen communities over any technology available at a time and in a location. This has drawn a lot of attention to signaling in packet networks, and particularly to the Session Initiation Protocol (SIP) [2]. This paper gives an introduction to SIP and discusses its current position and near-future perspectives, particularly pertaining to mobile environments.

KEYWORDS: SIP, mobile communities, UMTS, service access, internetworking, mobility, packet network signaling, application layer mobility

1 Introduction

The need for session signaling stems from switched networks, where basically all types of communication require reserving the resources needed for the desired services beforehand. Setting up the communication path is what signaling is for, and it takes place prior to any actual service data transmissions. Also, the channel used for signaling data is often different from the one used for service communications.

Until the late 90's, strict demands for network resources were rarely encountered in packet network services. Most traffic was stateless and timewise non-critical, so that the open, unreliable and unpredictable best-effort nature of the Internet was a good fit. However, when both fixed and mobile networks became faster and ubiquitous, they provided a viable platform for services demanding a certain level of communications quality. The most famous examples of such services are voice and video over the Internet - the first in the form of Voice over IP (VoIP), and the latter as video conferencing.

SIP development started in good time - halfway through the 90's - in the sense that few services in packet networks needed that kind of signaling at the time. Thus SIP was in a good position when interest in packet network signaling started to increase. This paper examines the development of SIP, as well as existing and envisioned uses for it in mobile

networks, particularly in 3G.

Chapter 2 will introduce SIP in some detail. Chapter 3 discusses the challenges posed by mobility, and solutions provided by SIP. Chapter 4 highlights some famous applications of SIP.

2 SIP Basics

In this chapter the phases of SIP evolution until today are explained. Then an introduction to the services provided by SIP is given, and the central network elements are presented. Finally the protocol structure is described. This section is mostly based on [8] and [2].

2.1 Roots of SIP

Work on SIP started in the Internet Engineering Task Force (IETF) in the mid-nineties, and in 1999 the growing interest in the protocol and packet network signaling in general resulted in the creation of the dedicated SIP working group. Since then, in step with the growing popularity, other SIP-related working groups have been established to work on subjects such as application of the protocol, instant messaging and presence services, and internetworking between switched and packet networks.

SIP is heavily based on the Hyper Text Transfer Protocol (HTTP) [4] and Simple Mail Transfer Protocol (SMTP) [1]. The most notable features inherited from HTTP are the client-server communication model and the command model, along with the Multipurpose Internet Mail Extensions (MIME) -type [5] message structure and the addressing scheme using URLs and URIs. SMTP contributed the message encoding schemes and header formats. These features in combination with the ASCII-based messaging make the protocol easy to understand and interpret by network professionals. The high point of SIP propagation so far has been its adoption as the main signaling protocol of the IP Multimedia Subsystem of the third-generation network defined by the 3rd Generation Partnership Project (3GPP).

International Telecommunication Union (ITU) has also defined a signaling protocol alternative to the SIP called H.323. Interworking between the two has also been investigated [7], but H.323 will not be described in this paper.

2.2 SIP Services

The purpose of SIP is, simply put, session signaling in packet networks. The main features of SIP signaling are *End point*

location, Determining willingness to establish a session, Exchange of media information, Session modification and Session termination These are each presented in more detail below. The term "call" will from here on be used to denote the total of exchange between the parties, regardless of whether the communication procedure in question actually is a voice or video call. It could also consist of e.g. machine-to-machine data traffic or multimedia streaming.

END POINT LOCATION in SIP facilitates locating an end-user by a global address. An application can send a request for session creation to the network, and the user's current location will be solved by the network elements. The single recipient address may be directed to several locations or devices, transparently to the caller.

DETERMINING WILLINGNESS TO ESTABLISH A SESSION is useful e.g. in cases where the recipient of a session invitation has pre-configured a reply to be sent to a caller - either in the application or in a network node. The pre-configuration might include a redirection to another address or device, or rejection altogether. Any action taken may be chosen by arbitrary rules based on caller identity, time or date, recipient location etc. - just to name a few. The recipient may also choose to reply at the reception of an invitation over any rules, should the application support such functionality.

EXCHANGE OF MEDIA INFORMATION is an essential part of session setup, so that the call parties can query each other's capabilities and agree on session properties. Session properties could include e.g. security mechanisms, encodings used, media transfer protocols and so on, depending on the call type.

SESSION MODIFICATION is a SIP mechanism to renegotiate session properties at any time, so that the session can be e.g. adapted to changes in call party status, call state or other conditions. This can be necessary e.g. in the case of user or service mobility, discussed more thoroughly below in 3.1.

SESSION TERMINATION is an ostensibly trivial concept, but crucial in a situation where call parties hold resources reserved for the session. Graceful tear-down is essential to allow all session parties to free those resources and clean up the session state.

2.3 Network Elements

This subsection presents the network elements central to the SIP architecture and discusses their role in applications.

There are a few central components to the SIP architecture. These are *User Agents, Proxy Servers, Redirect Servers, Registrars* and *Location Servers*¹. Below the role and purpose of each of these is explained in turn. The elements do not necessarily represent physical units, but more than one of them may be collocated in a single network node. The logical separation is nevertheless significant.

²

A **User Agent (UA)** is the endpoint of a connection over which the session is created. It can be a device or application of the caller, or of the callee. The User Agents are the only

part of the network that maintains the session state. Both signaling (i.e. SIP traffic) and data traffic are transmitted between user agents, but usually along different paths. UAs function in either client or server mode in each session, depending on whether they originated or received the session initiation.

Proxy Servers function in both client and server roles. They receive signaling data from UAs or other servers, in which case they act in a server role themselves. They then open new connections in a client role and direct the signaling forward to other servers or to a UA. Proxy Servers are stateless, except possibly in the transaction context. A transaction is the exchange of signaling data consisting of a request and the responses and acknowledgments induced by it. The proxy server has the possibility of forking the request to more than one destinations, in which case the transaction state is must be stored for response routing.

Redirect Servers do not forward request, but only return a response denoting where the request should actually be sent. They provide the redirection data, but do not store it. They may e.g. consult a Location Server to resolve the redirection rules for an address. In practice the Redirect Server is often collocated in the same network node with a Registrar.

Registrars accept registrations from UAs, and can store e.g. location data and the connection statuses of users. Mobile users notify their change of location by contacting a Registrar and registering their current location there. The Registrar does not necessarily host the data itself, but might be backed up by e.g. a Location Server.

A **Location Server** can store the location, presence status and other data of users registered in the network or domain serviced by that server. It usually acts as the back-end system for Registrars and Redirect Servers.

2.4 Protocol Description

This chapter describes the structure and operation of the protocol. First a description of the communication model is given. Then the elements of the protocol are explained in more detail. Finally an example operation is presented. This subsection is mainly derived from [6].

SIP is a client-server protocol. A client - that is, the caller who wishes to establish a session with the callee that can be a person or a service - sends a request that always generates at least one response. The request contains a method that defines the operation being requested, in a quite similar way to HTTP. SIP operates in the application layer on virtually any transport layer protocol, and is only interested in session signaling. E.g. session state presentation and media transport are handled by different protocols, and SIP does not restrict their use in any way. Session Description Protocol (SDP) is required to be supported by all UAs, but any suitable protocol can be used.

The protocol can be functionally divided into two parts - the SIP core and the extensions. The core defines the basic parts of the protocol, i.e. most of what is discussed in this document, while the extensions are features added to the core protocol. Every SIP client has to support the features defined in the SIP core, and the core defines mechanisms to negotiate for the extensions that all call parties support.

¹Location server is not strictly speaking a part of the SIP architecture, but it is an essential part of most implementations, so it is described here too.

Consequently, the simplest possible SIP UA will always be able to communicate with the most advanced UA, while the use of advanced features can be negotiated between UAs that support them.

2.4.1 SIP Core

The core SIP specification defines these six types of requests [6]:

- INVITE
- ACK
- OPTIONS
- BYE
- CANCEL
- REGISTER

INVITE is used to request a person or a service to establish a session. INVITE uses a three-way handshake consisting of the request, a response, and an acknowledgment of the response. All the other request types - which are used mainly during an existing session - only get the response, mainly because an existing session necessitates fast transactions. As an exception, INVITE can also be used during a session, since a re-INVITE is the defined way to modify session parameters.

ACK is the acknowledgment of having received the response to an INVITE request. It exists mainly to make the three-way handshake possible, but is also useful e.g. in interworking with signaling protocols with differing message sequences, or when using third-party call control.

CANCEL is used to stop a pending INVITE transaction. It is useful particularly in situations where a Proxy Server forks an INVITE to several locations, and wishes to cancel the others once one of the destinations replies. CANCEL has no effect on existing sessions.

BYE is used to signal the termination of an existing session. In a two-party session this means that the session is torn down. In a session between multiple parties, e.g. conference calls, a BYE request signals that the transmitter is leaving the session, but session status remains unaffected. BYE requests are sent directly to the recipient, bypassing the signaling channel that in other situations goes through e.g. Proxy Servers - with the exception of servers that have specified in SIP header fields that they wish to remain on the signaling path.

REGISTER requests are used to send registrations to a Registrar. They update the users location information, and potentially other data, so that the information is available to other SIP servers on upcoming requests.

2.4.2 SIP Extensions

The SIP core has been extended with several features. Some of the most important of these are listed below, along with the way they are implemented:

- Session status propagation (using the PRACK method with reliable delivery)

- Mid-session transactions (INFO method)
- Verification of session conditions (COMET method)
- Asynchronous event notification (SUBSCRIBE and NOTIFY methods)
- Session transfer (REFER method)
- Machine commands (DO method)
- Caller preferences (header fields)
- Instant messaging (MESSAGE method)

Several of the extension are most usable in mobile environments, and some of them are discussed in more detail below.

3 SIP and Mobility

This chapter introduces different types of mobility, and discusses the level of support SIP provides in each case. Also possible extensions and alternative solutions are touched upon.

3.1 Types of Mobility

Types of mobility can be categorized in several ways. One of them is presented here, and the suitability of SIP to those cases is discussed. In cases where no perfect match currently exists, possible solutions are considered or alternatives presented to be used alongside SIP. Mobility in general can be divided e.g. into the following types [9]: *Terminal mobility*, *Personal mobility*, *Session mobility* and *Service mobility*. Each of these is explained below, and the support provided by SIP in each case is discussed. Most of the discussion is based on [9].

TERMINAL MOBILITY means roughly what is commonly understood by mobility in general - the ability to move a device to another location and have it communicate from the new location too. Regarding SIP, terminal mobility has three aspects that affect the protocol: *pre-call mobility*, *mid-call mobility* and *network partitions*.

Pre-call mobility is the simplest case to solve, as it is one of the main features of the core protocol. It means that the UA is able to leave its current location when not involved in a session, move to a new location, and be reachable there and able to create sessions from there. This is the case with the normal registration procedure of SIP.

Mid-call mobility is quite a bit trickier, since the session is not allowed to brake or communication be delayed when changing locations and networks. The hand-off procedure at the network border is handled with SIP so that the moving UA first registers to the new network, and then sends a new INVITE to the other party with the new connection information. Depending on the network structure, transport protocols, media properties, physical distances and so on, the hand-off delay may require support from other communication layers. Proposed solutions are e.g. Mobile IP in the network layer [12] and Predictive Address Reservation with SIP (PAR-SIP) in the link layer [11].

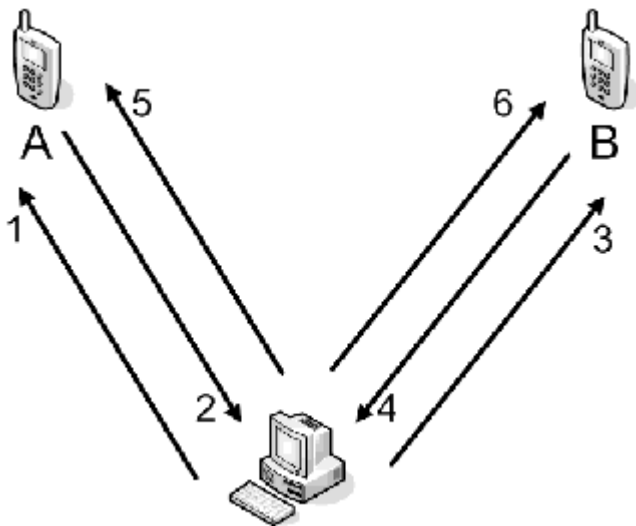


Figure 1: A description of 3rd-party initiated session between UAs

Network partition means the division of a single network due to e.g. router malfunction. Short breaks (up to 30 seconds) will not break sessions since SIP retransmits messages when no reply is received. On longer breaks, the session will need to be re-established, and in the case of moving users, even a new rendezvous may be needed.

PERSONAL MOBILITY is the ability of the user to change location transparently to other parties. The SIP registration mechanism supports this without modifications, since applications and devices can be programmed with arbitrary rules. A user may set her home UA to accept calls in the morning and in the afternoon, and redirect to a voice mail during the night. During the day the user may be registered on her UA at the office. Of course, manual registration can take place at any time and at any location.

SESSION MOBILITY means the ability to transfer an existing session to another UA. This could mean transferring a call started on a home PC to a mobile device when leaving the house, or even moving parts of a session, such as the voice stream of a video call to the audio system at home.

A simple way of transferring a session to another UA is to use the REFER method mentioned earlier. Using this method, either UA can send a reference to the other party. This will cause the party that received the REFER request to send an invitation to the referred location, and to continue the session with that end point.

A more general, and as such more widely applicable, solution is third-party call control that can be supported without changes to SIP. It basically means initiating a call between two parties by a third party. This is facilitated by the possibility to send an INVITE without a session description. The situation is depicted in Figure 1. In request 1 the session initiator, at a home computer, sends an INVITE with no session description to UA A. A replies in 2 with an OK response including a session description, which is then sent in an INVITE to UA B in request 3. B replies in 4 with an OK again containing a session description. This session description is sent to A in request 5 in an ACK message, and then an ACK is sent to B in 6. Now the session is ready and user at the

home computer can change the communication at either end at will.

SERVICE MOBILITY means being able to access the same services independent of location. The services in this case could mean personalized services such as a calendar or an address book that maintain data for the user. Such services can be accessible from anywhere only if they are either carried with the user, or if they can be accessed at a fixed address. The latter case would mean e.g. a server providing the service in the user's home network, but SIP provides no mechanism for this particular use.

4 Applications of SIP

This chapter briefly discusses current applications of SIP in fixed networks, and to a greater extent, in mobile third-generation networks.

4.1 SIP in Fixed Networks

Packet signaling is just as necessary in fixed packet networks as it is in mobile ones, even though it has gained a lot more popularity in the mobile world. This is a natural consequence of the facts that the telephone service is still the primary service, and it originated in fixed networks, so that in fixed locations it is often already available over a circuit-switched network. And with the common awkwardness of the still rather new packet services, most people have few reasons to adopt telephone service over fixed packet networks. However, the situation is likely to change with increasing speed in the near future, since at the same time as the quality of packet network communication services improves, new, interesting services unavailable in Public Switched Telephone Networks are developed at the same time.

At the moment there are no really wide-spread applications of SIP for fixed networks. While SIP would be a good fit also for VoIP services, the most popular implementations do not at the moment use it. SIP is gaining foothold all the time in smaller projects, but the SIP-based "killer application" for fixed networks has yet to emerge. The diffusion of SIP in fixed networks happens also largely as a byproduct of mobile services, since the services provided for packet networks are in many cases equally usable by mobile and fixed clients.

4.2 SIP as a Part of 3G

The pinnacle of SIP applications in mobile networks at the moment is the IP Multimedia Subsystem (IMS) of the third generation (3G) mobile networks. This section briefly introduces the IMS, and describes the use of SIP in IMS signaling. This section is mostly based on [6].

The wide publicity and hype around 3G networks is due to two aspects of the technology: first, it is expected to bring wide-band applications to mobile networks. Secondly, and probably more importantly, it is an apt example of the convergence of technologies taking place in several areas of communications, and as such is expected to "merge the Internet world and the cellular world" [6]. This merger also

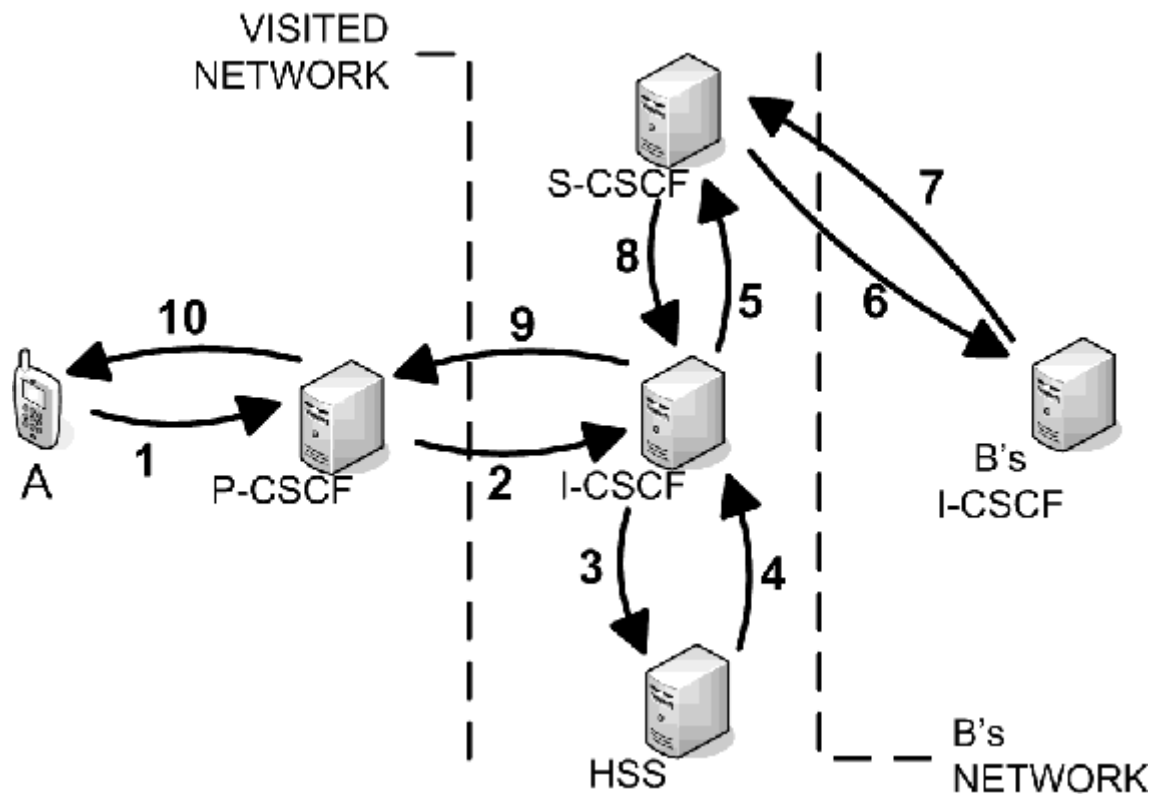


Figure 2: A description of session initiation by a roaming user in a 3G network

naturally brings packet network signaling to the focal point, for reasons discussed earlier.

The role of the IMS in 3G networks is to take care of transfers of voice and other media formats over the packet network. It provides e.g. security and quality of service on top of an inherently unreliable and open network, and thus signaling is one of its most important parts. The IMS relies on SIP in all service signaling with the user.

Call Service Control Functions (CSCFs) are the core of the IMS and operate in three different roles. A **Proxy CSCF** (P-CSCF) is the point of contact in the IMS for the mobile node, wherever it may roam. All requests to and from the mobile node pass through the P-CSCF that is currently assigned to serve the node. A **Serving CSCF** (S-CSCF) is the element that provides the actual network services to the mobile node, such as reachability and media services. Finally, an **Interrogating CSCF** (I-CSCF) exists to find the proper S-CSCF to serve the user. For this purpose it consults a **Home Subscriber Server** (HSS). From the SIP point of view, the HSS basically acts as a Location Server, introduced in 2.3, so that it provides information for the I-CSCF about which S-CSCF is to serve a particular user. Just as described earlier, the Location Server is a vital part of the network, but not a part of SIP communication per se.

SIP takes care of all end-to-end signaling that takes place in the IMS. The other end point can be another mobile node, a node in a fixed network, or a serving network node. Thus, whether a session is required for a video call between end-user nodes, for voicemail retrieval from a network service, or simply for registering to a new network - not that there's anything simple about that in the technical sense - SIP is the

work horse. The signaling is performed so that basically all the CSCFs function as SIP Proxy Servers that forward messages toward the destination. All mobile devices implement a SIP UA, as do the network nodes that provide services - within the IMS that means the S-CSCFs.

When a user enters a new network, i.e. crosses a network boundary or switches her mobile device on, a request with the REGISTER method is sent. During the registration, whether the user is in her home network or visiting a foreign one, the purpose of SIP signaling is to find the S-CSCF in the user's home network that from then on provides any provisioned services to the user. The first contacted network element in the mobile node's current location, a P-CSCF, is responsible for finding the user's home network and the I-CSCF located there that then takes the registration to the S-CSCF. Any incoming connections also go first to the S-CSCF in the home network that is serving the called user, and are then forwarded to the user's current location.

Figure 2 shows the creation of a session between users A and B. User A is roaming in a foreign network, and starts communication with user B by sending a session invitation to B. Only the first network element to contact on B's side of the network is shown, but otherwise that side is organized exactly the same way as A's side - whether B is at home or roaming elsewhere.

In step 1, A sends an invitation to the P-CSCF in the foreign network. The P-CSCF finds out where A's home network is, and in 2 sends the request to the I-CSCF in there. In step 3 the I-CSCF asks a HSS which S-CSCF should be used and gets a reply in 4 (this transaction does not use SIP, in line with the fact that Location Servers are outside of the core

protocol). The I-CSCF then forwards the invitation to the S-CSCF, which finds out the home networks of the target UA, in this case B. Messages inside B's network are not shown here, but in 5 the reply from B has passed through to the outermost network element, looking from B's side toward A. In 7 A's S-CSCF, again located in A's home network, receives the reply, which is in steps 8, 9 and 10 forwarded back to A through the I-CSCF and P-CSCF in that order. This is the signaling path between A and B in this configuration, but the media use different route. The only network node always on the communication path is the P-CSCF closest to each mobile node.

5 Conclusion

SIP is a mature protocol that provides good support for packet network signaling in the areas to which it is targeted. These areas are user mobility, message transfer and session negotiation [6]. User mobility has extended to terminal mobility and to some extent session mobility - with smart designs and no added complexity to the protocol - but that is the extent designed for SIP. The modular design of the protocol that separates the core and the extensions, and allows easy addition of features without compromising interoperability and compatibility, is an attractive property in several fields of communications. Together with the fact that SIP has definitely proven itself in the choice by 3GPP as the signaling protocol of the IMS, these properties are likely to lead to more wide-spread use of SIP. SIP is not a total mobility solution, though, but requires support from other layers in some situations [10] and [9] - nor should it be modified to be such, but to be used as a suitable tool for the tasks for which it was designed [6], p.163.

With the convergence of cellular networks and the Internet, and with circuit-switched networks losing their foothold in voice services, the global and ubiquitous packet network is coming ever closer. This creates a platform of an unprecedented scale for mobile communities, and session signaling is in a central role in a plethora of potential applications. Traditional voice and video calls, video mail services, instant messaging, location services and e.g. online games in a single network, combined with the levels of mobility provided by SIP, offer possibilities for quite extraordinary network applications.

References

- [1] Jon Postel(ed.). Internet Protocol. RFC 791, Information Sciences Institute, University of Southern California, September 1981.
- [2] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley and E. Schooler. SIP: Session Initiation Protocol. RFC 3261, IETF Network Working Group, June 2002.
- [3] Jonathan B. Postel. Simple Mail Transfer Protocol. RFC 821, Information Sciences Institute, University of Southern California, August 1982.
- [4] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. RFC 2068, IETF Network Working Group, January 1997.
- [5] N. Freed, N. Borenstein. Multipurpose Internet Mail Extensions, Part One: Format of Internet Message Bodies. RFC 2045, IETF Network Working Group, November 1996
- [6] Gonzalo Camarillo. SIP Demystified. McGraw-Hill Professional Book Group, 2001.
- [7] Kundan Singh and Henning Schulzrinne. Interworking Between SIP/SDP and H.323. Proceedings of the 1st IP-Telephony Workshop, Dept. of Computer Science, Columbia University, April 2000.
- [8] Alan B. Johnston. SIP: Understanding the Session Initiation Protocol. Artech House, second edition, 2004.
- [9] E. Wedlund and H. Schulzrinne. Mobility Support using SIP. In Proc. of Second ACM/IEEE International Conference on Wireless and Mobile Multimedia WoW-MoM99, Seattle Washington, USA, August 1999.
- [10] H. Schulzrinne and E. Wedlund. Application-layer Mobility using SIP. ACM SIGMOBILE Mobile Computing and Communications Review, Volume 4, Issue 3, July 2000.
- [11] W. Kim, M. Kim, K. Lee, C. Yu and B. Lee. Link Layer Assisted Mobility Support Using SIP for Real-time Multimedia Communications. ACM MobiWac, 2004.
- [12] M. Moh, G. Berquin and Y. Chen. Mobile IP Telephony: Mobility Support of SIP. In *Computer Communications and Networks, Proceedings*, 1999.