# The Future of Instant Messaging - SIMPLE, autumn 2002

MIKKO MARTSOLA
Helsinki University of Technology, Espoo

SIMPLE - SIP for Instant Messaging and Presence Leveraging Extensions

This document describes shortly what SIP protocol is and how it is extended for instant message and presence leverage. Current implementations are not perfectly compatible with each other and this document describes some of these problems. The document contains an overview about current clients and server implementations. It also discusses about firewall and NAT compatibility and problems they cause.

## 1. INTRODUCTION

The fact is that people want to communicate with each other. This is the main reason why mobile phone industry has kept on growing so rapidly. But the voice communication is not always the best way for communication. In the past writing a letter and nowadays e-mail, have been good ways for communication. The SMS message boom has showed that there is a need for new text based instant messaging systems. Although most people nowadays have phones and they use SMS messages in their everyday life, there is still a great demand for better realtime instant messaging systems accessible for many different devices. One might say that you should use IRC (Internet Relay Chat) for instant messaging with your friends. That could be a good solution in some situations, but the truth is that IRC is not a very practical for the conversation that SMS's makes possible. As a result there are several competing instant message systems available in the Internet. The biggest players like Microsoft and AOL have their own instant message systems which are not compatible with each other. Some smaller companies have started to produce clients which are compatible with these message systems and allow same user to be signed on several IM systems with a single client software. The evitable question will be wheather it is possible that instant message systems could be compatible with each other?

## 2. WHAT IS SIP

Instant messaging clients need to create connection between talking parties. The Session Initiation Protocol (SIP) [5] is a protocol which enables internet endpoints (user agents) to discover one another and to agree on a characterization of a session they would like to share. Many Internet applications transfer real-time multimedia session data such as voice, video or text messages over the Internet and the SIP protocol works in concert with these protocols. The SIP enables the creation of an infrastructure of network hosts (called proxy servers) to which user agents can send registrations, invitations to sessions, and other requests. SIP is an agile, general purpose tool for creating, modifying and terminating sessions which works independently of underlying transport protocols and without dependency on the type of session that is being established.

[5]

## 3. WHAT IS SIMPLE

Session Initiation Protocol (SIP) provides a mechanism that can be used in presence leverage and in session oriented communication, but it does not provide mechanisms for instant messaging. This chapter will demonstrate two solutions for adding instant messaging to the SIP protocol. There is also a description how presence modes are transferred in SIP protocol.

[6]

### 3.1 Instant messaging

[6] defines instant messaging in the following way: *Instant messaging is defined as the exchange of content between a set of participants in near real time. Generally, the content is short text messages, although that need not be the case. Generally, the messages that are exchanged are not stored, but this also need not be the case. IM differs from email in common usage in that instant messages are usually grouped together into brief live conversations, consisting of numerous small messages sent back and forth.*

*Instant messaging as a service has been in existence within intranets and IP networks for quite some time. Early implementations include zephyr, the UNIX talk application, and IRC. More recently, IM has been used as a service coupled with presence and buddy lists; that is,*

Author's contact information: mikko.martsola@hut.fi

Published in *Internet Protocols for Mobile Computing - Seminar on Internetworking, Autumn 2002*, by Helsinki University of Technology, Telecommunications Software and Multimedia Laboratory. The complete publication can be found at http://www.tml.hut.fi/Studies/T-110.551/2002/papers/December/index.html. 
ISBN 951-22-6272-X - ISSN 1455-9749 - HUT - TML - TML-C9

*when a friend comes online, a user can be made aware of this and have the option of sending the friend an instant message. The protocols for accomplishing this are all proprietary, which has seriously hampered interoperability.*

The following subsection gives quite a technical description about how SIP protocol is to be extended to allow instant messaging. There are actually two different implementations. Both have their own advantages and disadvantages. The first one (paging mode) uses SIP signaling channel also for instant messaging and the second one (session mode) features separate signaling and pay load channels. The paging mode specification is quite ready and there are clients and servers available which uses it. The session mode is still under development and there are not any publicly available applications that uses it.

3.1.1 *Paging Mode.* Paging Mode uses new SIP MESSAGE method that is defined for carrying instant messages as payload in SIP protocol. This means that it uses the SIP signaling channel to transfer messages. In Paging Mode conversation between two users operates in the following manner: User generates new SIP request that contains MESSAGE method. The request URI (Uniform Resource Identifier) [8] of this request is normally in the following form: "im:user@domain", but it can also be a device address in situations where the client has current information about the recipients location. This can be implemented by making a user presence system to supply up-to-date user device address information for the client. SIP request body contains the instant message sent, which can be of any MIME types. The message may traverse trough a set of SIP proxies.The address resolution rules detailed in SIP [5] specifications are used to locate each hop. In normal case the final recipient replies with SIP 200 OK message which will be returned to the sender as with any other SIP request.
[6]

Althought the paging mode is the mode that is in use in currently available SIP IM clients, it still has some problems:

(1) Proxies can fork SIP IM's to multiple parties because the MESSAGE is sent as a regular SIP request. IM's should be delivered only to the user that it was originally ment for, not to anyone else.

(2) Messages can be redirected to somewhere else than they were originally sent and the other peer never gets the message.

(3) It can be possible that MESSAGE requests can get record-routed. This can cause the override of the original request route used.

(4) It is not clear how separate paths for signaling and payload could be used in this mode.

(5) Congestion control can not be implemented. SIP requests normally use UDP and there is no way to prevent MESSAGE request within a session from traversing a UDP hop.

Despite of these disadvantages, there are clear benefits using of the use of SIP MESSAGE method. Because SIP messages travel trough proxies, the proxies can work as a intermediaries which provide a solution for firewall and NAT traversal.
[7]

3.1.2 *Session mode.* The newer proposal by IETF SIMPLE working group [1] Is session mode instant messaging. Eventhough it uses different paths for signaling and payload transmission, it does not define a new protocol. It defines some new SDP attributes that enables the usage of the SIP MESSAGE request for session model.

How session mode works in practise? Typical conversation goes like this: Caller sends a SIP INVITE message to its proxy. Proxy adds relay information to the sip message and then sends it to callee. Callee replies with SIP 200 OK message to the proxy. Proxy adds relay information also to this message and forwards it to caller. After that proxy has no control to the conversation. Caller sends SIP ACK to callee and conversation may begin. All following SIP MESSAGE and SIP 200 OK messages have route headers which forces them to be delivered trough a relay server. Relay removes the route header from the send messages and puts the messages forward to their destination. When conversation ends, the SIP BYE and SIP 200 OK messages are sent directly from peer to peer.

Session mode proposal is designed to solve all the disadvantages defined in (Sec. 3.1.1)

(1) Forking is not possible. [5] says that the location service cannot be executed if the proxy does not have control over the domain in the requested address. Also forking can only occur when location service is executed. Proxy does not ever own the domain of the callee, which causes that forking never occur.

(2) Redirection can not happen for same reasons.

(3) Record-routes from the MESSAGE, if present, are always ignored. The route for the MESSAGE requests is learned from the Session Description Protocol.

(4) Signaling and payload travel trough their own paths. The path of signaling SIP requests is determined from the record-route headers. The path of MESSAGE messages of IM conversation is set by hop headers of Session Description protocol.

(5) The hop field, which tells the address of congestion controlled relay server, must be added to the messages by the proxies.

[7]

### 3.2 Presence Leverage

Presence means users ability or willingness to communicate with other users in the network. In many applications the presence information is limited to two modes: online and offline. This is also the situation with basic SIP protocol. SIMPLE working group is defining methods to provide more precise presence modes. Currently the definition is not quite strict and it has not stabilized, yet. This have caused that the MSN, HotSIP and Siemens SIP Client can not interchange presence information.

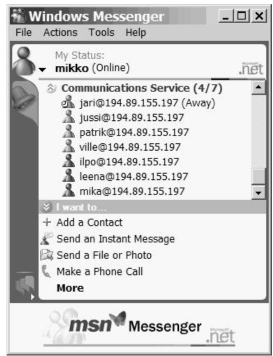


Fig. 1. MSN Messenger

3.2.1 *How IM client get information about other users presence.* Presence leverage works in this way. Client sends SIP SUBSCRIBE message to the proxy/presence server. This SUBSCRIBE contains information about other user. Server replies with SIP 200 OK or SIP 202 if the other user has not yet authorised the sender of the SUBSCRIBE message to see his presence mode. If so, the server sends SUBSCRIBE message to the subscribed party. If subscribed party is online he can reply with SIP 200 OK if he accepts the subscription. Server also immediately sends SIP NOTIFY message to the client. NOTIFY massage contains presence information on the requested user. This information can be faked to tell that the user is offline in the case there is not authorization yet. Client replies to the server with SIP 200 OK message. If the other user changes its presence mode, the server will send SIP NOTIFY message to all subscribed users. NOTIFY tells the new presence mode and clients reply to it with SIP 200 OK.

3.2.2 *How the user can change own presence mode.* SIP REGISTER message is used to inform SIP network about clients current communication address. Combined presence, proxy and registrar server is known as presence server. This server knows the user online / offline presence by using user registration information. When the user signs in, the instant message client sends SIP REGISTER message to the presence server and registers the user to the network. If the user changes his/hers state, a new REGISTER message is sent. When the user logs off the REGISTER message is also sent to the server.

3.2.3 *Using several clients.* User can have many devices for instant messaging. Any of these devices can be used for changing user's presence mode. When one of the user's devices changes presence mode, a notify message is sent to all devices.

3.2.4 *Error handling.* SIP REGISTER and SIP SUBSCRIBE messages have expire field that tells how long the information provided by the message is valid. Registrations and subscriptions must be renewed before expiration. If the server does not accept expiration time defined by client, it replies with message that sets a new expiration time. Expiration feature is useful in situations when IM client crashes or a network connection disappears.

## 4. DIFFERENCES THAT CAUSES PROBLEMS IN CURRENT IMPLEMENTATIONS

### 4.1 Instant messaging

Currently there are three working SIP instant messaging clients, that I know: HotSIP [12] and MSN Messenger [11] and Siemens SIP [13]. HotSIP works like IETF definition and MSN is more Microsoft proprietary. In this document MSN Messenger means messenger version 4.7 (4.7.0041). Version 5.0 (5.0.0537), which was recently released, does not support SIP at all. According to email conversation with Microsoft's agent, it seems probable that upcoming versions MSN Messenger 5.x will support SIP. All these three clients also work as a Voice over IP phone.

4.1.1 *SIMPLE spec defines that every IM has its unique call id.* Microsoft does not respect that every SIP instant message should have its own unique call id. Microsoft uses same call id for all messages sent from same messenger window. This is done to create some kind of session for each conversation, but I am not quite sure what are the advantages or if there any. This causes problems depending on how the server is implemented, because there is no specific definition. Server knows the order of the instant messages by the order received. Server must remember call id for every active sessions. If session is not used for a long time, server deletes the call id of the session from its memory. (Normally when the session is ended by closing Messenger client window. Closing the window causes a bye message sent by the Messanger.) If the client sends a message after that, the message has the same call id as the previous one, but server thinks that it is a first message for that session. If server sends an IM to the client it uses the call id it just received, but starts sequence numbering from the beginning (both ends, a server and a client have their own sequence numbering). When client receives the message, it checks the latest sequence number on the received message for that call id and detects that the sequence number is smaller than in previous message. This causes that the received message is dropped.

This problem can be bypassed by certain modification to the server. The same problem will occur if there is two

MSN Messengers configured to use each other as server and the other one crashes and is restarted. There is no solution for that case.

4.1.2 *SIMPLE group defines that there is only one transaction at the time.* SIMPLE group definition is that every message must have acknowledgement before a new message is sent. MSN messenger does not wait for the acknowledgements if user sends several messages rapidly. Therefore the might be several transactions open at the same time. If the server is implemented by the SIMPLE group definition, the server should drop messages which carries sequence number smaller than the last received sequence number. If several transactions are open at the same time there is a possibility that messages do not arrive in right order and some of the messages are therefore dropped.

### 4.2 presence leverage

4.2.1 *MSN implementation differs from the SIMPLE group.* Subscribe messages should be sent to presence server, but Messenger sends them through a proxy server to the other user. Presence information should be told to the presence server by sending a SIP REGISTER messages to the presence server. Then the presence server notifies the other clients by sending SIP NOTIFY messages to them. MSN Messenger sends these SIP NOTIFY messages directly to the other users. Messenger also sends the SIP notify message to subscribed users when the Messenger users logs off. The other clients send only SIP Register message to the presence server. This causes a problem. If someone tries to subscribe the user while the user is offline, then the subscriber must try subscribe until the user is online.

A better solution would be that subscriptions go through the proxy server and if user is offline, the server replies to the subscription and sends it to the user when he/she logs in next time.

Microsoft's implementation does not allow the user to have several clients open at same time.

## 5. CURRENT SIP INSTANT MESSAGING AND PRESENCE LEVERAGE SOLUTIONS

Several vendors are going to make their own SIP clients. At this moment the only publicly available clients are MSN Messenger version 4.7 (4.7.0041), HotSIP and Siemens SIP. MSN Messenger version 5.0 (5.0.0537) does not support SIP. [14; 15]

### 5.1 Clients

5.1.1 *MSN messenger.* MSN Messenger [11] Fig. 1 uses Microsoft proprietary protocol, but version 4.7 (4.7.0041) supports also SIMPLE. Version 5.0 (5.0.0537) does not support SIP. Messenger is free for all Microsoft Windows users.

5.1.2 *Hotsip.* HotSIP [12] HotSIP instant message client follows IETF SIMPLE specifications. The client is bundled with HotSIP server and the client is not authorised to be used with other servers.

5.1.3 *Siemens SIP.* Siemens SIP [13] Fig. 2 feels quite unstable at this moment. It is freely available. It is Voice Over IP and instant messaging capable.

### 5.2 Servers

Information in this subsection is directly taken from SIP Server manufacturer's homepage.

5.2.1 *Hotsip.* Features: SIP support/SDP - following latest drafts Proxy, redirect and registrar capabilities Advanced and extensive routing capabilities Open APIs designed for 3rd party service creation Net User Agent for web and WAP access Interoperable with SIP clients: Hotsip 's and Windows Messenger Contact list management Management, Charging (CDR), Provisioning, Configuration (XML), Logging and Statistics tools O&M tools using SNMP, RADIUS, and HTTP/HTTPS [12]

## 6. WORKING BEHIND FIREWALL

### 6.1 Which ports and protocols must pass firewall

Nowadays there are firewalls almost everywhere. Network Address Translation (NAT) is also in use in many places. These things can prevent instant messaging to work. Normally SIP protocol needs UDP port 5060 to be open. Instant messages travel through SIP protocol, therefore it is enough to allow UDP port 5060 packages pass the firewall. If instant message client can also be used for voice over IP or some other applications, then there is a lot more to be opened on firewall access list.

[9; 5]

### 6.2 NAT compatibility

6.2.1 *Instant messages in paging mode.* We assume that server has real IP address and client has NAT IP address. In paging mode the networks address translation does not prevent SIP protocol to work, if TCP is used as transport protocol. This is because the TCP stream is opened from the client. If UDP is used, then the network address translator does not know where it should send the incoming packages from the server to the client, because the server knows only the address of the NAT server, but not the address of the client.

6.2.2 *Instant messages in session mode.* In session mode SIP payload (instant messages) travels from client to client. This causes problems if one of the or both clients are behind the firewall. If the other client is behind NAT and the other is not, the session must be started from the client that is behind NAT. This works if TCP is used as a transport protocol. If UDP is used, then the client behind NAT can not receive messages until it has sent a message to the other client. If both clients are behind NAT, then there must be a special proxy that manipulates SIP packages and forces clients to communicate through it.

NAT problem is not yet solved with the SIP Voice over IP traffic. SIP working group is working on the problem and the same solution will be probably used in SIP instant messaging in the session mode.

SIP working group has this kind of a solution [10]: Client who is behind NAT uses TCP connection to send

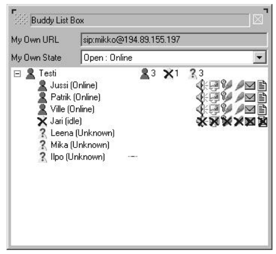


Fig. 2. Siemens SIP Client

the INVITE message, and to receive the response. Connection is left open permanently. This causes that there is no need to open new connections from the internet side of the NAT. If UDP is used, then the SIP proxy must be wise enough to check if the source address of SIP package and source address inside SIP packet differs. If addresses do not equal, then the proxy assumes that the client is behind NAT. In this case the proxy must send the response to the address and the port where the request originally came from.

## 7. CONCLUSIONS

Almost real time text based instant messaging is becoming a standard way of communication between people. New GPRS, WLAN, bluetooth and UMTS networks are enabling people to always be online. Currently the communication over GPRS network is quite costly and many operators use traffic based billing. SIP Instant messages require only small network bandwidth and are therefore efficient and quite inexpensive way of communication. J2ME (Java 2 Micro Edition) capable mobile phones are becoming more common and are enabling an easy way of implementing SIP instant message client to any mobile phone.

But there are some disadvantages. Do we have any privacy in the society where we are always connected to the Internet? Is there a possibility that someone actually can see where you are logged into the network and could use this feature for tracking down your location? I think these issues should be considered while designing new instant messaging services. If instant messaging is used for company's internal communication, it should be ensured that there is no way for people outside the company to read or manipulate messages. Most companies allow their employees to send emails to anyone during working hours. But is it okay to chat through an instant message system with your friends at work? To be able to use instant messaging efficiently, one should be able to use same client for communicating with co-workers, friends, relatives etc. But while at work it can get quite disturbing if your friends are trying to chat with you all the time. One solution for this problem could be the use of predefined profiles which make your presence information look different depending on who is requesting the information. These profiles could be manually changed by user and some features could be scheduled. For example your status could be shown as "busy" automatically during working hours to the friends and "online" to the co-workers.

I think that one important improvement to the SIP protocol will be the Session Mode instant messaging. Paging mode instant messaging works fine, but the feature richness of the SIP protocols causes some problems described earlier. The Session Mode instant messaging should be compatible with SIP protocol definition [5], so there should not be interoperability problems with current SIP proxies.

## REFERENCES

[1] IETF SIMPLE working group September 2002 URL: http://www.ietf.org/html.charters/simple-charter.html
[2] SIMPLE working group drafts September 2002 URL: http://www.softarmor.com/simple/drafts/
[3] SIMPLE working group mail archive September 2002 URL: http://mailman.dynamicsoft.com/pipermail/simple
[4] IETF SIP working group September 2002 URL: http://www.ietf.org/html.charters/sip-charter.html
[5] RFC3261, SIP: Session Initiation Protocol, J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler June 2002 URL: http://www.ietf.org/rfc/rfc3261.txt
[6] Session Initiation Protocol Extension for Instant Messaging, draft-ietf-sip-message-05 B. Campbell, J. Rosenberg,¡H. Schulzrinne, C. Huitema, D. Gurle June 26, 2002 URL: http://www.softarmor.com/simple/drafts/draft-ietf-sip-message-05.txt
[7] Using MESSAGE for IM Sessions J.Rosenberg May 2, 2002 URL: http://www.softarmor.com/simple/drafts/draft-rosenberg-simple-message-session-00.txt
[8] Uniform Resource Identifiers (URI): Generic Syntax URL: http://www.ietf.org/rfc/rfc2396.txt T. Berners-Lee, R. Fielding, U.C. Irvine, L. Masinter August 1998
[9] URL: http://www.vovida.org/pipermail/sip/2001-January/000324.html
[10] SIP Traversal through Residential and Enterprise NAT's and Firewalls J.Rosenberg,H.Schulzrinne March 2, 2001 URL: http://www.cs.columbia.edu/sip/drafts/draft-rosenberg-sip-entfw-01.txt
[11] Windows Messenger September 2002 URL: http://messenger.msn.com
[12] Hotsip September 2002 URL: http://www.hotsip.com
[13] Siemens SIP September 2002 URL: http://www.mysip.ch/download.html
[14] SIMPLE Chats New Standard to Make Instant Messaging Universal Paul Eng Aug. 24, 2002 URL:

http://abcnews.go.com/sections/scitech/CuttingEdge/cuttingedge010824.html
[15] AOL Time Warner delivers IM interoperability update Jennifer Jones , Cathleen Moore , and Matt Berger July 23, 2001 URL: http://iwsun4.infoworld.com/articles/hn/xml/01/07/23/010723hnaolmessanger.xml