SIP Mobility Modes: Application Layer and Data Link Layer

Abdullah Azfar† **Md. Sakhawat Hossen**† **Razib Hayat Khan** ††

†School of Information and Communication Technology,
The Royal Institute of Technology (KTH), Stockholm, Sweden

††Department of Telematics,
Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Summary
The Session Initiation Protocol (SIP) is one of the most widely used protocols for Voice over Internet Protocol (VoIP). Mobility is a very sophisticated service in VoIP. VoIP mobility performance depends on the handoff delay in the Data Link layer. As VoIP is application layer application, application layer mobility support has a vital role in supporting VoIP mobility. This paper presents an overview of different modes of VoIP mobility at both the application layer and data link layer. Different aspects of these modes are explored.

Key words:
Session Initiation Protocol (SIP), Mobility, Handoff delay, Mobile Host (MH), Corresponding Host (CH).

1. Introduction

SIP is an application layer protocol standardized by IETF [1]. SIP can establish, modify and terminate multimedia sessions. SIP is a signaling protocol which does not provide directly services, but provides primitives to implement different services. For establishing multimedia sessions, SIP supports five functions: determining a user’s location, user’s availability, user capabilities, session setup, and session management.

The SIP message flow is illustrated in figure 1. A SIP User Agent Client sends an invitation to the Outbound Proxy Server [2]. The Outbound Proxy Server consults a Domain Name System (DNS) server to find the Inbound Proxy Sever for the Destination User Agent based on the Uniform Resource Identifier (URI) sent by the User Agent Client. The Inbound Proxy Server then looks for the location of the Destination User Agent via the Location Server and forwards the invitation to its final destination. The resulting media sessions occur between the User Agents through Real Time Protocol (RTP) after the session is established. This message flow structure is referred to as the SIP Trapezoid.

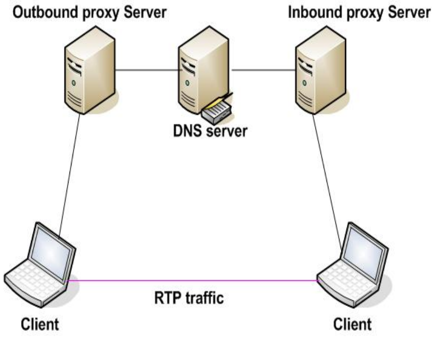

Fig. 1 SIP Trapezoid.

On the other hand, a SIP session is established following a three way handshake (see figure 2) [3]. The caller sends an INVITE request. The called party replies with an OK message. The caller originating the INVITE request sends an ACK message after receiving the response. For session termination, a BYE message is sent by either of the parties.

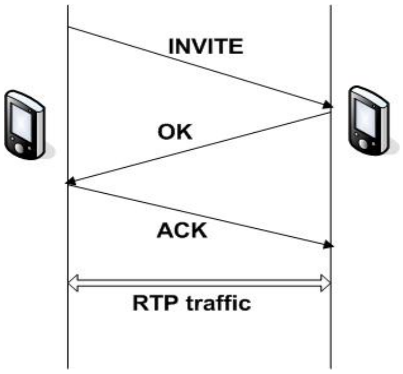

Fig. 2 SIP three way handshaking.

Manuscript received December 5, 2009
Manuscript revised December 20, 2009

Many research papers have been written SIP mobility. SIP is mainly concerned with the application layer, therefore it supports application layer mobility. However, the performance of the application will depend on the total handoff delay. Handoff is performed at the data link layer, the network layer, and via SIP at the application layer. This paper tries to review different modes of mobility as supported by SIP with respect to providing application layer mobility with a focus on handoff issues.

The rest of the paper is organized as follows: different modes of SIP mobility are discussed in section 2. Application layer handoff techniques and data link layer handoff techniques for SIP are discussed in section 3 and 4. A general discussion about the mobility issues is presented in section 5. Finally, some conclusions are drawn in section 6.

2. SIP Mobility Modes

There are four modes of mobility supported by SIP as discussed in [4]. Brief overviews of the four modes of mobility are given below.

2.1 Terminal Mobility

In terminal mobility the mobile device moves across one subnet to another. Terminal mobility ensures uninterrupted communication with the Mobile Host (MH) despite its movement from one subnet to another. Mobile IP plays a vital role in supporting terminal mobility. Whenever a MH moves from one subnet to another a new IP address has to be assigned to it. This can be done through a DHCP server.

Whenever a MH changes its location, the change is detected by the Home Agent (SIP proxy). The corresponding host (CH) gets the new location of the MH from SIP proxy and sends an INVITE message to the MH. This can be done without any loss of information if the MH and CH are not in an ongoing conversation. This is known as pre-call mobility [4]. But for an ongoing conversation known as mid-call mobility this becomes a real challenge. The MH sends an INVITE request directly to the CH to update its IP address. The real time traffic is carried over RTP/UDP. So while assigning a new IP, special care should be taken to avoid triangular routing and any kind of encapsulation [5]. C. H. Yeh, Q. Wu and Y. B. Lin developed a new architecture design on the protocol stack of SIP [6] to get a better performance of delay in terminal mobility. The detail of this architecture is beyond the scope of this paper.

2.2 Session Mobility

Session mobility allows a user to change devices or terminals within an ongoing session. Session mobility provides ubiquitous service to a user. It allows changing terminal seamlessly, continuing the media session without any disruption or re-establishment of the session. For example, a user can view a video stream on his laptop and want to continue viewing the stream via his mobile device after going outside. Session mobility provides this facility. Depending on the user’s requirements, session mobility can provide different facilities. Whenever a session has been transferred to other device, it should be possible to bring back to the session to the previous device. On the other hand, if multiple media types are involved in a session, then user should be able to choose whether to transfer the whole session or to transfer a portion of the session. For example, in a multimedia session the user should be able to choose to continue with video in one device and audio via another device.

There are four approaches to provide session mobility. They are- Media Gateway Control Protocol (MGCP), REFER method, Third Party Call Control (3PCC), and Split a SIP session over multiple devices (SSIP).

2.2.1 Media Gateway Control Protocol (MGCP)

MGCP is a call control architecture based on gateways and call control agents [7]. The call control agents will synchronize with each other to communicate with the gateways. The gateways will execute the commands sent by the control agents. A gateway supports two or three connections per endpoint. This enables the gateway to provide three way calling service. The gateways can establish a call between two parties operated by a single call agent, or two different call agents. After establishing the call, the call agents can exchange information. A call agent can change the IP address or port number by sending a “modify connection’’ command to the gateway for an established call.

2.2.2 REFER Method

REFER method [8] is a method provided by SIP to support session mobility. The sender sends information about a new terminal to the recipient. The recipient is supposed to contact this UA and establish a session. The REFER request is simply a request to the recipient to contact a new UA. For example, Bob is using his fixed phone to contact Alice. Now, Bob leaves his house and he wants to continue the session. He could send a REFER request to Alice containing his mobile device identity in the ‘’REFER TO’’ header field. This causes Alice to send and INVITE message to Bob’s mobile device in order to initiate communication. On the other hand, Alice’s communication with Bob’s fixed phone comes to an end.

2.2.3 Third Party Call Control (3PCC)

In 3PCC, one entity controls the call between two parties. The third party acts like an operator. For example, a user can click a link on a web page to dial the operator, then the operator will establish a call. There are four different call flows of 3PCC [9]. Each of them has benefits and drawbacks. The simplest of the call flows is as follows. The operator first sends an INVITE message without any Session Description Protocol (SDP) to the caller. The caller responds with a 200 OK message which contains an OFFER for the callee. The operator then sends an INVITE message with the OFFER to the callee. The Callee sends a 200 OK message with an ANSWER. The operator acknowledges this by an ACK message to the callee and ACK ANSWER message to the caller. Now the media session between the caller and the callee is established. 3PCC can provide session mobility using this method. Whenever a user wants to change the device the operator can inform the other party about the new device as described above.

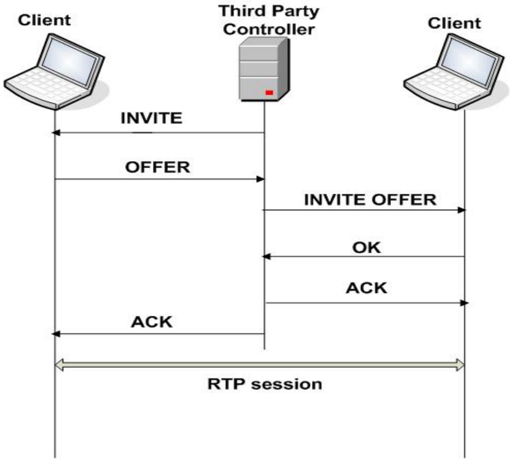

Fig. 3 Simple call flow in 3PCC.

2.2.4 Split a SIP Session Over Multiple Devices (SSIP)

SSIP provides the service of splitting an ongoing session over multiple devices. Some improvements over the user agent make it possible to split the session. An extension header named “Mobility” is added to the REFER method to make it transparent to the callee [10]. Using SSIP, a session can be completely split, partially split or a new session can be added to the ongoing session. Initially SSIP did not have the ability to retrieve a split session back into one session. Later the retrieval mechanism was implemented by using a Nested REFER method [11] [12].

2.3 Personal Mobility

Personal mobility allows a user to initiate and receive calls from any location and any device [13]. A user can use more than one device to send and receive calls. For example, one can use a PC, a mobile phone and a PDA. Personal mobility provides a way to communicate with the user through the same user name via a logical address mapping. The reverse process where the user can be reached by numerous addresses at a single device is also provided by personal mobility. SIP forking proxies can be used to reach a person on multiple devices.

The amount of traffic sent to numerous devices to reach a person is a central issue in personal mobility. INVITE messages can be sent to the devices to find the active device by forking, but this is a time consuming process. T. P. Wang and H. Y. Lee have proposed sending the INVITE messages to the devices by dividing them into ‘active’ and ‘standby’ groups [14]. Whenever a user registers a new address, the address will be stored with a new qvalue [1]. The qvalue will indicate the precedence of the new address. Higher priority will be assigned to address with higher qvalues. Hence, the least used addresses should have a lower qvalue. If the value of qvalue exceeds the maximum range then the lowest valued address will be placed into a standby group. Now, any INVITE request will be forked to the active group. If no response is received, it will be forked to the standby group. If the address is found in the standby group, then this address will be returned to active group replacing the least recently used address in the active group. The qvalue of an address will always be increased whenever it sends an INVITE message. In this way, the most active address can stay in the top of the list.

2.4 Service Mobility

There are a number of services that users will want to use whenever they are changing devices. The services that a user will want to use when changing devices or networks include personal data, contact book, call logs, speed dial lists, and so on. Service mobility gives the user the ability to continue to use these services. The user need not to re-enter personal data, contact number, call logs, etc. each time they change their devices or network. The service mobility capability of SIP makes it possible for the user to have up to date information updated in any device.

Initially it looks simple to provide service mobility. A user could have an USB flash memory device to carry personal data, contact numbers and plug it to the device from which he wants to use his SIP account [4]. The device can read the data from the USB flash memory. Or the user can have a GSM SIM card to carry all these data. However, these techniques have a limitation. The devices which will be used to login to the SIP account will have to have USB ports or GSM SIM card slots.

In practice, it is feasible to update the user data from the user agent to the home server frequently. Whenever the user logs in, the necessary data will be retrieved from the home server. For example, a process could update the local device with any new information once an hour. The user agent sends a timestamp to the home register while sending user data. This helps the home server to decide whether to update its own database or to update the user agent.

An architecture has been proposed by R. Shacham, H. Schulzrinne, W. Kellerer, and S. Thakolsri for location based service mobility [15]. The location sources can be divided into two groups: stationary and mobile. Stationary location sources are fixed and they identify a user entering a room and publish the information. Mobile location sources are the handheld devices carried by the users. The mobile location sources receive information through Bluetooth, DHCP, or GPS and send it to the home server. This architecture also defines a “Room state” which defines the current users in a room. The user profiles are represented in XML format.

3. Application Layer Handoff Techniques

This section discusses different handoff schemes used by SIP at the application layer. Application layer handoff delay does not take into account the movement detection, IP address discovery, or configuration. As discussed in terminal mobility, the MH sends an INVITE message to the CH after getting a new IP address. If the MH and CH are far away from each other, it takes a long time to redirect data from the CH to MH. This may cause loss of traffic. This can be solved by multicasting data during the change over period. Several ways of achieving fast handoff in SIP are discussed by A. Dutta, S. Madhani, W. Chen, O. Altintas, and H. Schulzrinne in [16].

3.1 SIP Registrar and RTP Translator

The MH sends an INVITE message to the CH of the visited network after obtaining a new IP address. However, a delay may occur to register with the new subnet. If every subnet has an RTP translator it can avoid data loss in this period. Whenever the MH moves to a new subnet, the SIP proxy server in the older subnet sends the IP address of the moved MH to the RTP translator. Any packets received in the former subnet will now be processed by the RTP translator. The RTP translator redirects the received packets to the new IP address. The RTP translator assumes that the MH has successfully established a connection with CH when no packet has been received for a while. At this point, the RTP translator removes the IP table entry for the MH.

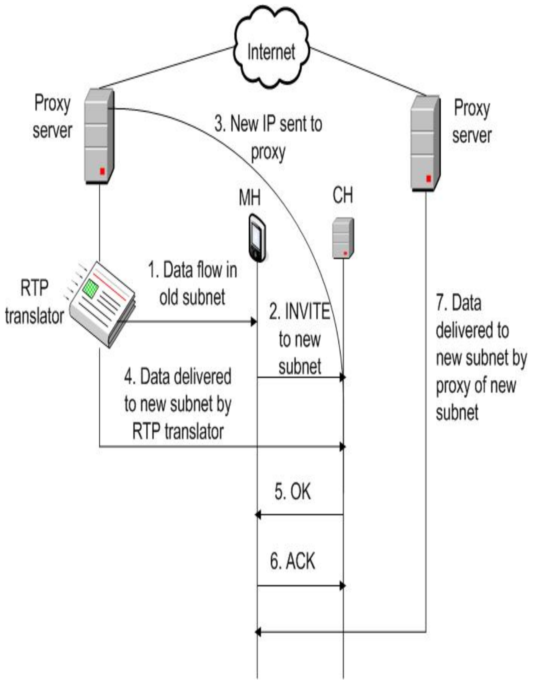
Internet
Proxy server
3. New IP sent to proxy
Proxy server
MH
CH
RTP translator
1. Data flow in old subnet
2. INVITE to new subnet
7. Data delivered to new subnet by proxy of new subnet
4. Data delivered to new subnet by RTP translator
5. OK
6. ACK

Fig. 4 RTP Translator.

3.2 SIP Outbound Proxy

Whenever a MH visits a new network, it sends an INVITE request to the outbound proxy of the SIP registrar in the previously visited network. The outbound proxy remembers this INVITE message. Whenever the MH moves to a new subnet, the outbound proxy issues a re-INVITE message to the CH based on the INVITE message. The outbound proxy can use the data sent by the MH to configure the RTP translator.

3.3 Back-to-Back User Agent

A Back to Back User Agent (B2BUA) [1] [17] is a logical entity. The B2BUA sits in between the MH and CH. When receives a request from the MH it acts as user agent. The other user agent issues a new request and forwards this request to the CH. The second user agent sends its own address as the destination. Whenever a MH moves from one subnet to another, the MH sends the INVITE request to the B2BUA. The B2BUA then sends a re-INVITE message to the CH. Until the B2BUA gets a connection established with CH, media streams are delivered to the previous address.

4. Data Link Layer Handoff Techniques

The application layer is not concerned with address allocation to the MH. Address allocation is taken care of by the data link layer. A fair amount of handoff delay is associated with the address allocation. Address allocation time may be more than a second in case of Dynamic Host Configuration Protocol (DHCP). But in case of *adhoc* networks, DHCP has some shortcomings. The server has to be preconfigured and the clients can request an address from the server only.

A protocol named Dynamic Registration and Configuration Protocol (DRCP) has been proposed by R. Vaidyanathan, L. Kant, A. McAuley, and M. Bereschinsky in [18]. DRCP uses a Client-Proxy-Server model. The client configures itself automatically by using information from the proxy or server. A proxy can be any client in the network which advertises the server information. The proxies can be regarded as virtual clients. The proxies propagate data between clients and servers. The servers own addresses and lease them to clients and proxies. Servers also authenticate clients (and proxies) and control access to the network. The DRCP sever or proxy broadcasts DRCP_ADVERTISE message. Any client wishing to join the network sends a DRCP_REQUEST message to the server or proxy. The DRCP server or proxy then sends a DRCP_REPLY to establish a session.

Another mechanism of Predictive Address Reservation with SIP (PAR-SIP) has been proposed by W. Kim, M. Kim, K. Lee, C. Yu, and B. Lee in [19]. PAR-SIP performs address reservation before link layer handoff. Whenever a MH encounters a new subnet and the signal to noise (SNR) ratio exceeds a threshold, Then the MH selects a new Access Point (AP). The Base Station (BS) plays a vital role here. The BS maintains a table containing the Media Access Control (MAC) addresses and network addresses of the neighboring BSs. The information in this table is used to perform an IP address reservation and to create an AP list for the MH. Whenever the MH predicts a new AP, the BS confirms the network address of the AP. The BS then reserves a new IP address in the new subnet via a DHCP server. This IP address will be used as the IP address of the MH after hand off. The MH sends a re-INVITE message to the CH with the new reserved address. Afterwards, the CH sends an OK message and handoff is completed.

5. Discussion

Different types of mobility modes at the application layer have different issues regarding handoff delay. As an application layer module, the handoff delay for SIP terminal mobility depends on the underlying address detection procedure and the time require to resume the media transmissions. Issues such as detachment from the old access point and attachment to new access point are link layer issues. There will always be some handoff delay associated with terminal mobility. There will be a subsequent delay if all the packets are tunneled through the home agent. The address encapsulation adds further delay while assigning new IP [4].

In session mobility, a disruption can occur when transferring sessions between devices. A long disruption can cause loss of communication. SSIP establishes communication with the new device before tearing down the previous communication. For this reason SSIP does not suffer from loss of communication problem.

Personal mobility suffers from delay issues too. Sending a fork message to all devices can be time consuming as well as bandwidth consuming. The location based service mobility scheme allows many devices to be used by a user. It reduces the need to carry a GSM SIM or USB flash memory.

The RTP translator method used in application layer handoff reduces the packet loss during handoff. Experiments performed in [16] show that it can achieve 80% improvement over packet loss compared with normal handoff process. But there is a possibility of duplicate packets being received. This is because the RTP translator can send some packets even after the handoff is completed. The B2BUA approach is also an efficient approach but, it needs additional user agents. The SIP outbound proxy must be able to access to the Session Description Protocol (SDP) in order to configure the Network Address Translator (NAT).

The performance analysis of data link layer handoff shows that the average configuration time for DRCP increases linearly with the increase of DRCP advertisement frequency. On the other hand, the DRCP message traffic decreases in proportion to the advertisement frequency. But in both cases the linearity is maintained up to a threshold. This is because if advertisement frequency increases too high then a cluster of clients wait in a queue. With each DRCP_ADVERTISE the proxy server has to serve more clients. The PAR-SIP method provides better performance by reducing handoff delay.

6. Conclusions

This paper has delineated a brief overview of SIP mobility issues at both application layer and data lank layer. The overview reveals that the handoff delay is the most important criteria of concern. No matter what the source of delay, a handoff delay for a significant amount of time can cause loss of data. The present solutions for reducing the handoff delay are very useful, but new solutions can improve the quality of service a lot.

References

[1] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, “SIP: Session Initiation Protocol”, IETF RFC 3261, *IETF Network Working Group*, June 2002.

[2] A.B. Johnston, “SIP: Understanding the Session Initiation Protocol”, *2nd Edition, Artech House Telecommunications Library,* 2004, ISBN 1-58053-655-7.

[3] G. Camarillo, “SIP Demystified”, *1st Edition, Mcgraw-Hill,* 2002, ISBN 0-07-137340-3.

[4] H. Schulzrinne, E. Wedlund, “Application-Layer Mobility Using SIP”, *ACM SIGMOBILE Mobile Computing and Communications Review,* Vol. 4, Issue 3, pp. 47 – 57, July 2000, ISSN: 1559-1662.

[5] A. Dutta, F. Vakil, J. C. Chen, M. Tauil, “Application Layer Mobility Management Scheme for Wireless Internet”, *Proceedings of IEEE International Conference on Third Generation Wireless and Beyond (3Gwireless ’01),* San Francisco, pp. 379-385, May 2001.

[6] C. H. Yeh, Q. Wu, Y. B. Lin, “SIP Terminal Mobility for both IPv4 and IPv6”, *Proceedings of the 26th IEEE International Conference Workshops on Distributed Computing Systems,* pp. 53-58, July 2006, ISBN :1545-0678 , ISSN: 0-7695-2541-5.

[7] M. Arango, A. Dugan, I. Elliott, C. Huitema, S. Pickett, “Media Gateway Control Protocol (MGCP)”, IETF RFC 2705, *IETF Network Working Group*, October 1999.

[8] R. Sparks, “The Session Initiation Protocol (SIP) Refer Method”, IETF RFC 3515, *IETF Network Working Group*, April 2003.

[9] J. Rosenberg, J. Peterson, H. Schulzrinne, G. Camarillo, “Best Current Practices for Third Party Call Control (3pcc) in the Session Initiation Protocol (SIP)”, IETF RFC 3725, *IETF Network Working Group,* April 2004.

[10] M. X. Chen, C. J. Peng, R. H. Hwang, "SSIP: Split a SIP session over multiple devices", *Computer Standards and Interfaces*, Vol. 29, No. 5 pp. 531-545, July 2007.

[11] M. X. Chen, F. J. Wang, “Session Mobility of SIP over Multiple Devices”, *4th International Conference on Testbeds Research Infrastructures for the DEvelopment of NeTworks COMmunities (TRIDENTCOM 2008),* Innsbruck, Austria , March 2008.

[12] R. Sparks, “The Session Initiation Protocol (SIP) Referred-By Mechanism”, IETF RFC 3892, *IETF Network Working Group,* September 2004.

[13] R. Pandya, "Emerging Mobile and Personal Communication Systems", *Communications Magazine, IEEE*, Vol. 33, Issue 6, pp. 44-52, June 1995.

[14] T. P. Wang, H. Y. Lee, “User Location Management for Personal Mobility in SIP-based VoIP Services”, Third International Conference on Communications and Networking in China, Hangzhou, China, pp. 910-914, August 2008, ISBN: 978-1-4244-2374-3.

[15] R. Shacham, H. Schulzrinne, W. Kellerer, S. Thakolsri, “An Architecture for Location-based Service Mobility Using the SIP Event Model”, *MobiSys 2004 Workshop on Context Awareness*, Boston, June 2004.

[16] A. Dutta, S. Madhani, W. Chen, O. Altintas, H. Schulzrinne, “Fast-Handoff Schemes for Application Layer Mobility Management”, *15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications,* Vol. 3, pp. 1527- 1532, September 2004, ISBN: 0-7803-8523-3.

[17] N. Banerjee, S. K. Das, A. Acharya, “SIP-based Mobility Architecture for Next Generation Wireless Networks”, *Third IEEE International Conference on Pervasive Computing and Communications (PerCom'05),* Kauai Island, Hawaii, pp. 181-190, March 2005, ISBN: 0-7695-2299-8.

[18] R. Vaidyanathan, L. Kant, A. McAuley, M. Bereschinsky, “Performance Modeling and Simulation of Dynamic and Rapid Auto-Configuration Protocols for Ad-hoc Wireless Networks”, *36th Annual Simulation Symposium,* Orlando, Florida, pp. 57- 64, 30 March-2 April 2003, ISBN: 0-7695-1911-3.

[19] W. Kim, M. Kim, K. Lee, C. Yu, B. Lee, “Link Layer Assisted Mobility Support Using SIP for Real-time Multimedia Communications”, *Proceedings of the Second International Workshop on Mobility Management & Wireless Access Protocols,* Philadelphia, pp. 127-129, October 2004, ISBN: 1-58113-920-9.

Abdullah Azfar is doing his MS in Erasmus Mundus NordSecMob program specialized in Security and Mobile Computing in Norwegian University of Science and Technology (NTNU), Norway and Royal Institute of Technology (KTH), Sweden. He received his BSc degree in Computer Science and Information Technology from Islamic University of Technology (IUT), Gazipur, Bangladesh in 2005. He served as a lecturer in the Islamic University of Technology during the period March 2006 – July 2008. He also served as a lecturer in Prime University, Dhaka, Bangladesh during the period October 2005 – February 2006. He received the Erasmus Mundus scholarship from the European Union for his MS studies and OIC (Organization of the Islamic Conference) scholarship for three years during his BSc studies. His research interest is mainly focused on Information Systems Security. At present he is working with security issues in VoIP.

Md. Sakhawat Hossen is doing his MS in Internetworking in the Royal Institute of Technology (KTH), Sweden. He received his BSc degree in Computer Science and Information Technology from Islamic University of Technology (IUT), Gazipur, Bangladesh in 2004. Formerly he was a lecturer in the department of Computer Science and Engineering (CSE) of Stamford University Bangladesh. His research interest is mainly focused on Evolutionary optimization, Internet security, wireless sensor network (WSN), IP network and VoIP. Currently he is doing his Master thesis on secure session initiation protocol (SIP) user agent with key escrow capability to facilitate lawful interception (LI).

Razib Hayat Khan is doing his PhD at Department of Telematics, Norwegian University of Science and Technology (NTNU), Norway. He completed his M.Sc. in Information & Communication Systems Security specialized in Security in Open Distributed System from Royal Institute of Technology (KTH), Sweden in 2008. He worked under VRIEND project (http://vriend.ewi.utwente.nl) as part of his M.Sc. thesis which was sponsored by Sentinels, a joint initiative of the Dutch Ministry of Economic Affairs, the Netherlands organization for Scientific Research Governing Board and the Technology Foundation STW and the industrial partners were Philips Electronics, AkzoNobel, Corus, and DSM. He also worked as research engineer, Multimedia technologies at Ericsson AB, Sweden. He received his B.Sc. degree in Computer Science and Information Technology from Islamic University of Technology (IUT), Gazipur, Bangladesh in 2004. He served as a lecturer in Stamford University, Dhaka, Bangladesh during the period November 2004 – August 2006. He received the OIC (Organization of the Islamic Conference) scholarship for three years during his BSc studies. His research interest is mainly focused on Network performance modeling, Information Systems Security. At present he is working with performance and security issues in Communication system.