

Communication Cost of SIP Signaling in Wireless Networks and Services

Elthea T. Lakay, Johnson I. Agbinya**

Private Bag X17, University of the Western Cape, Bellville, 7535, RSA; 2001084@uwc.ac.za

**Faculty of Engineering, University of Technology, Sydney, NSW 2007, Australia;
agbinya@eng.uts.edu.au

Abstract—Developers of SIP mostly concentrated on making SIP to do what no other protocol can do. In a way they failed to consider the weaknesses and dangers. From this realization we did a study on the weaknesses of SIP and SIP signaling in Wireless Networks and Services. From our study we identified Delay and Security as the most severe problems, which we discuss in this paper.

Index Terms—Delay, Quality, Network Security, QoS, SIP.

I. INTRODUCTION

SIP is currently receiving much attention and seems to be the most promising candidate as signaling protocol for the current and future IP telephony services, which is becoming a real competitor to plain old telephone service. For the realization of such a scenario, there is the obvious need to provide a certain level of quality and security, comparable to that provided by the traditional telephone system. We attempt to summarize the cost of using SIP signaling on wireless networks and devices. From the study, we identified security [3] and delay [1] as the most severe problems in SIP Signaling. Different types of security and delay, are discussed in Section II and III respectively. In section IV we discuss VoIP.

II. QUALITY AND SECURITY

Although security and privacy should be mandatory for an IP telephony architecture, most of the attention during the initial design of the IETF IP telephony architecture and its signaling protocol, SIP, has been focused on the possibility of providing new dynamic and powerful services, and simplicity. Less attention has been paid to security features.

If the new IP telephone architecture and SIP is to replace PSTN, it should provide the same basic telephony service with a comparable level of QoS (Quality of Service) and network security.

The following security characteristics should be guaranteed:

1. high service availability
2. stable and error-free operation
3. protection of the user-to-network and user-to-user network traffic

SIP messages may contain information a user or server wishes to keep private. The headers can reveal information about the communication patterns and content of individuals, or other confidential information. The SIP message body may also contain user information (media type, codec, address and ports, etc) that should not be revealed.

Securing SIP header and body information can be motivated by two different reasons:

1. Maintain private user and network information in order to guarantee a certain level of privacy
2. Avoiding SIP sessions being set up or charged by someone faking the identity of someone else.

The mechanisms that provide security in SIP can be classified as end-to-end or hop-by-hop protection. End-to-end mechanisms involve the caller and/or callee SIP user agents and are realized by features of the SIP protocol specifically designed for this purpose (e.g. SIP authentication and SIP message body encryption). Hop-by-hop mechanisms secure the communication between two successive SIP entities in the path of signaling messages. SIP does not provide specific features for hop-by-hop protection and relies on network-level or transport-level security. Hop-by-hop mechanisms are needed because intermediate elements may play an active role in SIP processing by reading and/or writing some parts of the SIP messages. End-to-end security cannot apply to these parts of messages that are read/written by intermediate SIP entities.

Two main security mechanisms are used with SIP: authentication and data encryption. Data authentication is used to authenticate the sender of the message, and to ensure that some critical message information was unmodified in transit. This is to prevent an attacker from modifying and/or replaying SIP requests and responses. Data encryption is used to ensure confidentiality of SIP communications, letting only the intended recipient decrypt and read the data. This is usually done using encryption algorithms such as Data Encryption Standard (DES) and Advanced Encryption Standard (AES).

SIP supports two forms of encryption:

1. end-to-end
2. hop-by-hop

End-to-end encryption provides confidentiality for all information. On the contrary, hop-by-hop encryption of whole

SIP messages can be used in order to protect the information that should be accessed by intermediate entities, such as *From*, *To*, and *Via* headers. Encryption of such information can prevent malicious users from determining who calls who, or accessing route information. Hop-by-hop encryption can be performed by security mechanisms external to SIP.

The security mechanisms must be combined properly to obtain a trusted network scenario. SIP communications are susceptible to several types of attacks:

1. The simplest attack in SIP is *snooping*, which permits an attacker to gain information on users identities, services, media, network topology and so on. This information can be used to perform other types of attacks.
2. Modification attacks occur when an attacker intercepts the signaling path and tries to modify SIP messages in order to change some service characteristics. This kind of attack depends on the kind of security used.
3. Snooping is used to impersonate the identity of a server or user to gain some information provided directly or indirectly by the attacked entity.
4. Finally, SIP is especially prone to denial of service attacks that can be performed in several ways, and can damage both servers and user agents.

The attack techniques may cause memory exhaustion, processor overload, and so on.

Although the security mechanisms provided with SIP can reduce the risk of attacks, there are some limitations in the scope of the mechanisms that must be considered.

III. DELAY

The SIP-T signaling system is a mechanism that uses SIP to facilitate the interconnection of PSTN with carrier class VoIP network. Based on IETF, the SIP-T signaling system not only promises scalability, flexibility, and interoperability with PSTN but also provides call control function of MGC (Media Gateway Controller) to set up, tear down and manage VoIP calls in carrier class VoIP network. The performance analysis of SIP-T signaling system plays an essential role in optimizing network QoS (Quality of Service). Queuing size, mean of queuing delay and the variance of queuing delay are the major performance of MGC in carrier class VoIP network. [1] assume a mathematical model of the M/G/1 queue with non-preemptive priority assignment to represent SIP-T signaling system. They also presented the formulas of queuing size, queuing delay and delay variation for the non-preemptive priority queue by queuing theory respectively.

Since a satellite component has been identified within Universal Mobile Telecommunication System (UMTS), there is a need to support SIP-based sessions over Satellite-UMTS(S-UMTS) as well as to achieve an end-to-end seamless IP-based terrestrial/satellite network integration. [4] aims at evaluating the performance of SIP-based session setup over the S-UMTS, taking into account the larger propagation delay over the satellite as well as the impact of the UMTS radio interface.

A. QUEUING DELAY

The queuing size was analyzed [1] using imbedded Markov chain and Semi-Markov process and the queuing delay and delay variation was analyzed using LST (Laplace-Stieltjes Transform) [9]. Queuing size is defined as the number of SIP-T messages in the system.

Mathematical				
SIP-T message (messages/s)	Mean (ms)	Std Dev (ms)	Buffer Size	
50	0.49	0.49	0.12	
100	0.62	0.68	0.23	
150	0.78	0.89	0.37	
200	1.00	1.14	0.52	
250	1.29	1.46	0.72	
300	1.71	1.90	0.97	
350	2.37	2.57	1.33	
400	3.57	3.75	1.92	
450	6.39	6.41	3.17	
500	25.00	18.57	8.20	

TABLE 1: MATHEMATICAL RESULTS [1]

Simulated				
SIP-T message (messages/s)	Mean and 95 th percent (ms)	Std. Dev (ms)	Buffer Size	Sample Size (SIP-T message)
59.2	0.49±0.01	0.47	0.12	4925
100.7	0.61±0.01	0.65	0.25	10074
150.4	0.79±0.01	0.88	0.39	15044
201.8	0.98±0.02	1.10	0.54	20178
251.4	1.27±0.02	1.40	0.76	25138
301.1	1.70±0.02	1.91	1.03	30115
348.9	2.36±0.03	2.57	1.40	34891
399.5	3.62±0.04	3.98	1.97	39955
450.7	6.49±0.06	6.67	3.49	45067
502.5	21.08±0.18	20.12	8.85	50251

TABLE 2: SIMULATION RESULTS [1]

Table1 and table2 shows the comparison data between mathematical and simulation results including buffer size, mean queuing delay and standard deviation of queuing delay for $P_u=0.004$. P_u denotes the error probability of SIP-T message.

The increase of the size of the queue, mean queuing delay and standard deviation of queuing delay vary slowly as arrival rate of SIP-T messages less that 450. However, the values increase dramatically with the arrival rate of SIP-T messages greater than 450. An intuitive and reasonable explanation for this phenomenon is that the SIP-T message arrival rate approaches the processing capability of the system for heavy traffic intensity. [1]

From the results in table1 and table2, we can see that the theoretical estimates are shown to be in excellent consistence with simulation results.

B. CALL SETUP DELAY

The SIP protocol enables a wide set of applications and Multimedia over IP (MoIP) is one of them. SIP-based video-telephony is one of the most challenging Multimedia over IP applications. A test in [2] has been carried out in a 3G network emulator, to measure **post-dialing delay, answer-signal delay and call-release delay**. These delays happen during the lifetime of a SIP call. This results were compared

to local, national, international and overseas Intranet LAN calls.

Post-Dialing Delay (PDD) is also called post-selection delay or dial-to-ring delay. This is the time elapsed between when the caller clicks the button of the terminal to call another caller and the time the caller hears his terminal ringing.

Answer-Signal Delay (ASD) is the time elapsed between when the callee picks-up the phone and the time the caller receives indication of this.

Call-Release Delay (CRD) is the time elapsed between when the releasing party (the caller) hangs-up the phone and the time the same party can initiate/receive a new call.

These delays are shown in the following table:

	Local SIP Call	National SIP Call	International SIP Call	Overseas SIP Calls	3G SIP Calls
PDD	24ms	38ms	153ms	24ms	62ms
ASD	23ms	31ms	147ms	237ms	45ms
CRD	11ms	30ms	138ms	230ms	50ms

TABLE 3: RESULTS OF PDD, ASD AND CRD FOR 3G SIP CALLS VS. INTRANET CALLS [2]

Tests incase of bandwidth limitations has been done [2]. The call success rate was always 100%, thus bandwidth would not stand in the way for a successful SIP call.

PDD for 2kbps bandwidth was less that 1 second. PDD for 5kbps bandwidth was approximately 420ms. Values of PDD decrease with increasing bandwidth. At the maximum bandwidth of 254kbps, the PDD was around 50ms. In realistic 3G network configurations, the bearer allocated for signaling could be a few kbps. [2]

ASD values were constantly 45ms for channels of at least 5kpbs, but ASD increased for very narrow channels (2kpbs to 166 ms). We think this due to the fact that loss is much higher for narrow bandwidth, whereby a 200/OK must be retransmitted. [2]

CRD could not be measured, because the media packets were queued up in the simulator, and they blocked the channel for a long time. [2]

Globally, these SIP signaling values are well in line with the Grade of Service (GoS) bounds proposed by the ETSI TIPHON QoS classes [10]. Air interface losses and narrow channels have a great impact on the overall SIP call setup time. This yields large call setup times. However, it is expected that UDP/IP header compression and SIP message compression algorithms would greatly reduce the SIP call setup delay over 3GPP networks.

C. MESSAGE TRANSFER DELAY

Tests done in [4] were for message transfer delay for the SIP INVITE at block error rates of 0%, 10% and 20%. Since the INVITE method is deemed to be the most important method in SIP, as it is the only method used to establish a session

between participants and normally contains the description of the session to be setup, it is interesting to ascertain its performance.

It is found that the message transfer delay increases as the message size increases. This is expected. It is found that the transfer delay of the INVITE request is substantially reduced compared to when no link-layer retransmission is employed, whereby the delay reduction is more as the message size and the block error rates increases. This is because without retransmission at the link-layer, a segment that is lost means that the whole message cannot be recovered at the receiver side and thus, the whole message needs to be retransmitted at the session-layer, according to the SIP reliability mechanism.

The tests also show that the delay is lower when having unsolicited and solicited STATUS report option set, compared to only having the solicited feedback. This is because by incorporating unsolicited feedback on top of solicited, the missing protocol data units (PDU) can be recovered faster since retransmissions of missing PDUs can be performed before polling; also the reduction in delay is more at a higher block error rate (BLER).

D. SESSION SETUP DELAY

From tests done [4], it was shown that, session setup delay and call blocking probability for a simple call setup sequence, whereby the call setup consists only of sending of the INVITE request and the 180 ringing response. Results were presented for the different STATUS report triggers settings with Tgood ranging between 0.5s and 10s, and Tbad equals to 0.5s, 2s and 4s. Comparing both schemes, it can be seen that when the channel is good, i.e. for a lower Tbad value a higher Tgood value, there is hardly any difference in performance, but as the channel gets worse, combining both unsolicited and solicited feedback options gives a lower delay and blocking probability.

Due to the inherent characteristics of SIP signaling being transactional-based and generous in size, the transport of these packets over the radio interface is not sufficient and when transversing over the error-prone wireless link plus a larger satellite propagation delay, the session establishment delay can be rather large. In our opinion SIP over Satellite is not advisable. From the study in [4], they have shown that with the presence of RLCAM, the session setup performance can be substantially improved. Also it is shown that the combination of unsolicited and solicited STATUS report trigger gives a better performance than just solicited alone in terms of delay and blocking probability in a more hostile environment.

E. HANDOFF DELAY

[6] define total handoff delay (D) as the time between detachment from old access medium and establishment of communication with correspondence node (CN). It consists of three (3) components:

1. time for switching lower layer medium to access network (D1).
2. time for detecting a new router and a new link (D2)

3. time for recovery of communication with a CN after detecting a new link (D3).

D3 is the one that [6] concentrated on. There are two main factors which contribute to delay D3; Duplicate Address Detection (DAD) and router selection. DAD imposes delay between receiving a Router Advertisement (RA) and sending packet out of the interface with auto configured IPv6 address. The purpose of DAD is to confirm the uniqueness of the IPv6 address on the link.

[6] measured the handoff delay of SIP terminal mobility in their IPv6 testbed. Two different scenarios have been considered:

- a) SIP mobility without kernel modification
- b) SIP mobility with kernel modification.

Measurement has been done for scenario a) and b). The following table shows the handoff delay, D3, which is related to signaling:

Handoff case	a)	b)
H12	38290.0 ms	171.4 ms
H23	3932.2 ms	161.6 ms
H31	1934.7 ms	161.1 ms

TABLE 4: HANDOFF DELAY OF SIGNALING [6]

The following table shows another handoff delay related to voice communication and the results is shown:

Handoff case	a)	b)
H12	38546.3 ms	420.8 ms
H23	4187.7 ms	418.6 ms
H31	1949.4 ms	408.4 ms

TABLE 5: HANDOFF DELAY OF MEDIA UDP PACKET [6]

We observed from tables 4 and 5, modified kernel has reduced handoff delay. We understand that even with the modified kernel, the delay figures are not acceptable for real-time multimedia communications. To complete the study in [6], they then integrated the MIPL MIPv6 in their testbed and performed the same experiment as done for SIP mobility. They observed that the handoff delay for signaling is about 2ms and the handoff delay for media UDP packet is less than 31ms. The results showed that MIPL MIPv6 with modified kernel outperforms the SIP mobility with modified kernel. However, they believe that application layer mobility, such as SIP mobility, is a potential candidate to support real-time applications.

IV. FAX OVER IP

For new Voice over IP (VoIP) networks to succeed, they will have to support legacy fax equipment installed at user premises worldwide. This fax requirement is attached to the PSTN, which is very robust and reliable. However, the reach and the popularity of the Internet combined with the fact that its use is almost free, is a major driver for VoIP networks and for the transfer of all applications that exist on the PSTN, to it. The most major of these applications is fax transmission. Fax transmission has special requirements, the first being that while the loss of a packet during a human conversation is not likely to affect a voice call a lot, it can easily affect a fax call. This is because fax transmission requires far more signaling and handshaking than a regular telephone call. This includes

negotiating details such as speed, paper size, and delivery confirmation. Apart from the signaling in a fax call, the sending and receiving of fax documents are mostly done and interpreted by automated fax machines. Therefore, any error in either the signaling or the actual transmission is likely to result in a lengthy recovery.

For three network models, [8] measured the link utilization of the inter-proxy-server link (IPSL), the link utilization of the auxiliary link (AUXL), the average end-to-end delay of SIP signaling packets, the average end-to-end delay of fax data packets, the average SIP call setup time, and the average fax call setup time. The three network models are discussed below:

Network Models:

1. calls are generated from the T.38 gateway, and all messages are sent to the SIP proxy server on its network. The path between the T.38 gateway and the SIP proxy server has an IP router on it. This is not necessary, but is done to maintain compatibility with the next two network models where IP routers play an important role. The SIP proxy server of the originator's network communicates with that of the recipient's network and initially sends SIP messages to set up a call. Once that is done, the T.38 gateway starts fax transmission. The originating T.38 gateway starts sending fax packets to the SIP proxy server. That is, the SIP proxy server routes and interprets SIP messages. But it also routes IFP packets without actually interpreting them. Such a scenario is likely if the SIP proxy server is implemented on a router within the network. In that case, it interprets SIP messages, possibly translating them, and maintaining state for them. But all other messages are not interpreted by it – they are just routed. This model is shown in figure 1.

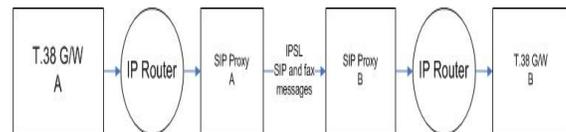


FIG. 1: SIP PROXIES ROUTING SIP AND FAX PACKETS[8]

2. has the same network elements but a different setup. It includes the components and links of the first model, but also has a direct link between the IP routers. That is, all the SIP signaling is carried out on the path that traverses the two SIP proxy servers. However, once the call is setup, all the fax data packet transmission is done through IP routers and links only. This is more likely to be a network where signaling travels on separate links, and data is sent across another set of links. This is possibly because data does not need to go through SIP proxy servers or other network entities. This frees up resources at such entities and segregates signaling from data transmission, much like PSTN networks where the Signaling System #7

(SS7) links are distinct from trunks. Figure 2 gives a view of this model.

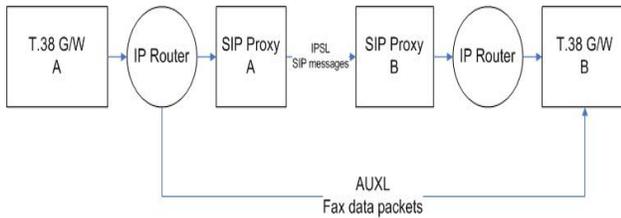


FIG. 2: SEPARATE PATHS FOR SIP MESSAGES AND FAX IMAGES [8]

- uses the same network as network model 2. However, not all SIP messages travel between the SIP proxy servers in this case. In general, all SIP terminals are configured to know their network's SIP proxy server. Hence, to set up a SIP call, they contact their SIP proxy server first and do not need to know the entire route to a receiver themselves. They only need to know how to route SIP calls they initiate to their designated SIP proxy server. The SIP proxy server then handles all the signaling for the rest of the call. For fax data packets, the AUXL link is used as is done in network model 2.

	Model 1	Model 2	Model 3
Average link utilization of IPSL (Mbps)	77.50	9.223	0.9546
Average link utilization of AUXL (Mbps)		69.03	75.96
Average end-to-end fax data packets delay (sec)	0.2023	0.2017	0.2017
Average end-to-end signaling delay (sec)	0.2014	0.2015	0.2009
Average SIP call setup times (sec)	3.4023	3.4042	3.4020
Average fax call setup time (sec)	4.2086	4.2106	4.2060

TABLE 5: EXPERIMENT RESULTS [8]

In terms of just link utilization values on IPSL, network model 3 is the most suitable. In terms of just link utilization values on AUXL, network model 2 is the most suitable, but it is only marginally better. In terms of end-to-end fax data packet delay, network models 2 and 3 are equally good. In terms of just packet end-to-end delay, network model 3 is the best. In terms of just average SIP call setup time and average fax call setup, network model 3 is the best.

V. CONCLUSION

In this paper the most severe problems, security and delay, of SIP and SIP signaling in Wireless Networks and Services are discussed. We found that though there are security mechanisms provided with SIP, that can reduce the risk of attacks, there are still some limitations in the scope. There are about five delay types we found, namely: Queuing Delay, Call Setup Delay, Message Transfer Delay, Session Setup Delay, and Handoff Delay, whereby Handoff Delay is the most severe delay. Each delay case where modified to

support real-time applications, but there is still room for improvement, especially for Handoff Delay.

The theoretical performance analysis of SIP-T signaling shows the robustness regardless of the effect of traffic intensity and error probability of SIP-T message for the non-preemptive priority queue. Thus we can determine how much the ratio of cost to performance can tolerate, how the planning and design is needed so as to meet the requirements of carrier class VoIP network.

This conclude that SIP and SIP signaling will soon be on even a better security and delay level then PSTN or any other used protocol.

ACKNOWLEDGMENT

The authors would like to thank Telkom COE (Centre of Excellence) and the Department of Computer Science for giving them this opportunity.

REFERENCES

- J.S. Wu and P.Y. Wang, "The Performance Analysis of SIP-T Signaling System in Carrier Class VoIP Network", 17th International Conference on Advanced Information Networking and Applications (AINA'03), IEEE March 2003, pp.39
- I.D.D. Curcio and M. Lundan, "SIP Call Setup Delay in 3G Networks", Seventh International Symposium on Computers and Communications (ISCC'02), IEEE July 2002, pp.835
- S. Salsano, L. Veltri and D. Papalilo, "SIP Security Issues: The SIP Authentication Procedure and its Processing Load", Network IEEE Nov/Dec 2002, vol. 16, Issue 6, pp.38-44
- V.Y.H. Kueh, R. Tafazolli and B. Evans, "Performance Evaluation of SIP-based Session Establishment over Satellite-UMTS", Vehicular Technology Conference 2003 (VTC 2003-Spring). The 57th IEEE Semiannual, IEEE 22-25 April 2003, vol. 2, pp.1381-1385
- J.W. Jung, R. Mudumbai, D. Montgomery and H.K. Kahng, "Performance Evaluation of Two Layered Mobility Management using Mobile IP and Session Initiation Protocol", Global Telecommunication Conference 2003 (GLOBECOM'03), IEEE 2003, Vol. 3, pp.1190-1194
- N. Nakajima, A. Dutta, S. Das and H. Schulzine, "Handoff Delay Analysis and Measurement for SIP based mobility in IPv6", International Conference on Communications (ICC'03), Vol.26, IEEE May 2003, pp.1085-1089
- N. Banejee, K. Basu and S.K. Das, "Handoff Delay Analysis in SIP-based Mobility Management in Wireless Networks", International Parallel and Distributed Processing Symposium (IPDPS'03), IEEE April 2003, pp.224b
- U. Choudhary, E. Perl and D. Sidhu, "Using T.38 and SIP for Real-Time Fax Transmission Over IP Networks", 26th Annual IEEE Conference on Local Computer Networks (LCN'01), IEEE November 2001, pp.74
- R.B. Cooper, "Queues Served in Cyclic Order: Waiting Times", The Bell System Technical Journal, Vol.49, No.3, March 1970, pp.399-413
- ETSI Telecommunications and Intranet Protocol Harmonization Over Networks (TIPHON), End to End Quality of Service in TIPHON System; Part2: Definition of Quality of Service (QoS) Classes, TS101 329-2 v.1.1.1.1, July 2000