Next Generation Telephony: A look at Session Initiation Protocol

White Paper

Author:	Thomas Doumas
Version:	A.00.03
Last modified:	04/05/99 9:02 AM
File:	SIP_White_Paper.doc

Contents

1. INTRODUCTION 3

1.1 HISTORICAL PERSPECTIVE 3
1.2 IP TELEPHONY BENEFICIARIES 5
1.3 THE IP TELEPHONY CHALLENGES 5
1.4 IP TELEPHONY SIGNALING 6
1.5 VOIP SIGNALING PROTOCOLS 7

2. INTRODUCING SIP 7

3. SIP OPERATION 8

3.1 PROXY SERVER OPERATION 8
3.2 REDIRECT SERVER OPERATION 10
3.3 SIP ADDRESSING 11

4. SIP MESSAGES 12

4.1 THE INVITE MESSAGE 12
4.2 RESPONSE MESSAGES 13

5. SDP 14

5.1 MEDIA INFORMATION 14

6. SIP VS. H.323 16

7. SIP STACK 15

8. TESTING SIP 18

9. VOIP FUTURE 19

10. FURTHER INFORMATION 20

11. BIOGRAPHY 21

1. INTRODUCTION

Next generation telephony and the role of Session Initiation Protocol is more easily understood with a little background on the telephone industry and the Internet. The following sections present a brief overview of the current telephone system and the Internet.

1.1 Historical perspective

The telephone system dates back to 1876 when the Bell Company was formed. This was the original public switched telephone network (PSTN). Over the years the telephone system evolved to a complex network that provides advanced services in addition to voice calling. These services include direct dialing, billing options (credit card, calling card, collect, prepaid minutes, etc), privacy options (caller id, caller id block, etc), convenience options, directory information, etc. The fundamental circuit switched architecture remains, while most everything else has changed. Until recently, the telephone system in the United States was a monopoly. As is common with monopolies, the innovation is introduced slowly and costs are high. Along with the monopoly came strict federal regulation of the quality and availability of service.

The telephone system of the United States had great influence on the telephone systems of other countries. Although differences in protocols and physical interfaces exist, the principles are the same. Many telephone systems outside of the United States are still monopolies under governmental control.

The telephone system architecture today makes use of two distinct functional layers, the transport layer and the control layer. The transport layer consists of a mesh of circuit switches. The switches on the edge, providing connection to the end-user, are called Class 5, end office switches (from the AT&T 5ESS). The switches in the core, providing connection among end office switches, are called Class 4 tandem switches (from the AT&T 4ESS). The control network consists of a mesh of computers and databases that control the behavior of the circuit switches of the transport layer. The separation between the control layer and the transport layer provides “common channel signaling” with which the transport resources are freed from signaling traffic. All signaling traffic travels on the Signaling System 7, or SS7 control network. Figure 1 presents a diagram of the telephone system that highlights the architecture.

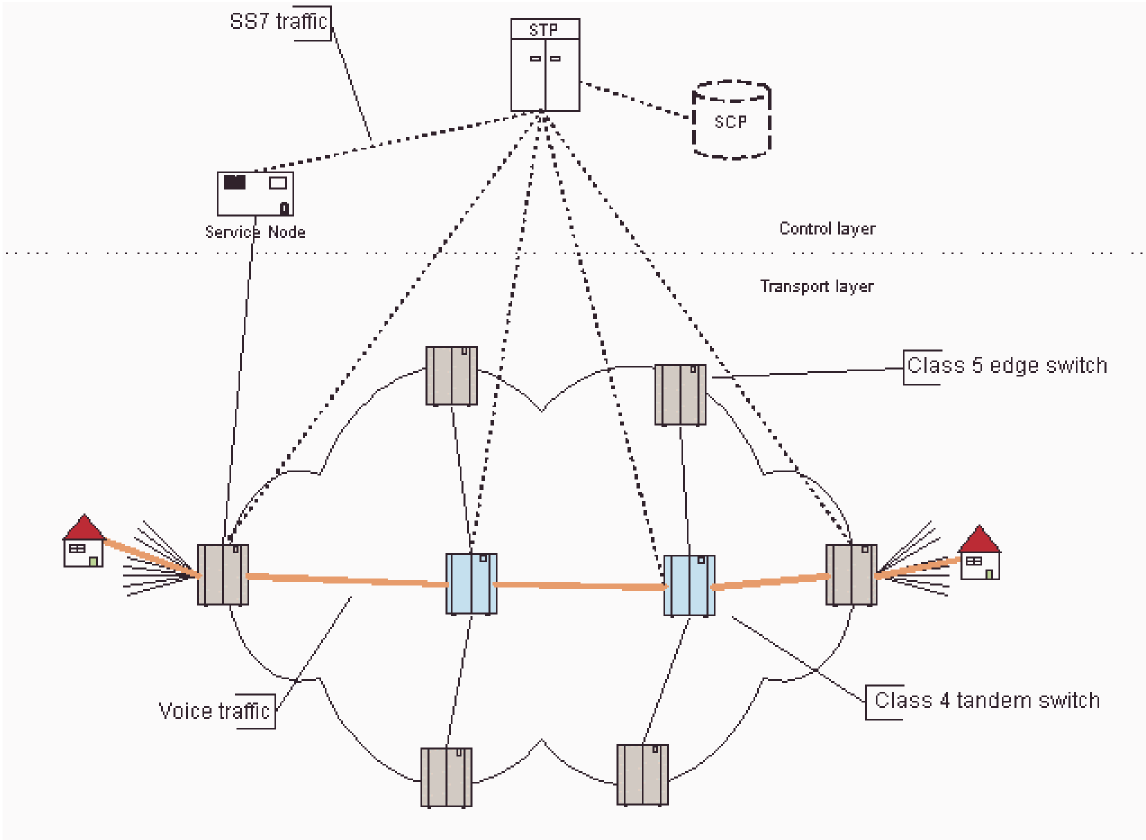

Figure 1 PSTN architecture

On the other hand, the Internet grew out of United States government and university research projects. The intent of these projects was to develop a communications network that could survive catastrophes. The Internet does not have nor requires a separate control layer. The reason for this is that unlike the telephone system, the Internet employs a packet switched transport layer. There is no signaling because circuits do not need to be established in advance of data transfer as in a circuit switch transport. Each packet contains sufficient information so that the packet switches (routers) can forward it to its ultimate destination.

The basis of the Internet is a suite of open standards (IP, TCP, etc). Over the years this open-ness has allowed fierce competition among providers of network equipment used to build the physical Internet. The competition has produced network equipment that is relatively inexpensive when compared to the proprietary equipment that makes up the telephone network. Another difference with respect to the telephone system concerns regulations. There is fundamentally no governmental regulations putting a burden on the Internet. A key consequence of this is that the Internet lacks the high level of service quality and availability of the telephone system.

The open standards of the Internet have resulted in a large number of independent software developers producing innovative applications. One of these applications is telephony service. A great amount of research effort focusing on real-time applications on the Internet has now yielded an exciting opportunity for a next generation telephone system. This next generation telephone system will be based on the core Internet technologies of IP, TCP and UDP. There is an important distinction between building the next generation telephone system on the Internet and building it on Internet technologies. There are too many technical and management problems to overcome to use the actual Internet. Instead next generation telephone companies will build private data networks based on Internet or IP technology.

The monopolies of long distance telephony are gone, the Internet technologies have matured, and it is time for the next generation telephone system to be built as a **Voice over IP** (VoIP) system.

1.2 IP telephony beneficiaries

The preceding sections provide some insight into why there is benefit to a new, IP telephony system. This section presents the primary beneficiaries. The driving factor focuses on value. The next generation telephone system promises to better telephony services at a lower cost and lower cost for basic services to consumers. The next generation telephone system promises to provide profits and growth the companies that provide the networks and equipment for the networks. We expect to see new carriers, new equipment manufacturers, new opportunities for existing carries and new opportunities for existing equipment opportunities. Almost over night new companies such as Qwest, and Level3 have formed to provide telephony services over their own managed IP networks. The advantages that these companies seek to exploit are inexpensive packet switching equipment, rapidly increasing performance and lower cost software development.

Furthermore, there will be new opportunities for software development companies to provide innovative telephony applications that take advantage of the World Wide Web and computer/telephony integration. For example product support call centers are looking towards “click to call”. This “click to call” also has application in electronic commerce.

1.3 The IP telephony challenges

There are three main challenges in the path from the Internet technologies of today to a next generation telephone system. The first is that some call signaling capability needs to be brought to packet switching. The second is that quality of service must be controlled. The third challenge is building a converged PSTN/VoIP network. The transition from PSTN to VoIP will be a gradual process. Significant technical and business issues will need to be solved. Since VoIP and PSTN networks will co-exist for many years a converged network will need to be built in order to bridge the VoIP and PSTN networks. The converged network will allow calls to originate on a VoIP network and terminate on a PSTN network (and vice versa). The converged network will make use of VoIP gateway devices to bridge the two networks. Figure 2 presents a conceptual diagram of a converged network.

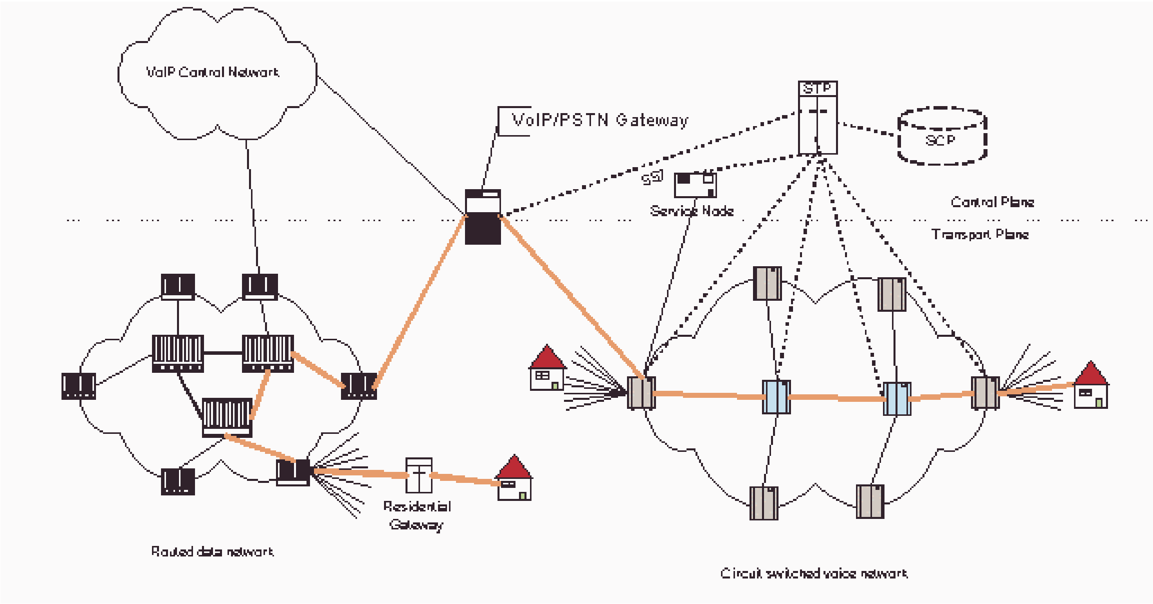

Figure 2 Converged architecture

There are efforts underway all over the world focused on these problems. This paper explores the topic of call signaling in the next generation telephone network.

1.4 IP telephony signaling

Previous sections have described IP, packet switched networks as not needing call signaling. Telephony applications introduce the requirement for signaling into IP networks. The objective of signaling protocols is to establish operating parameters to be used for a call prior to the data transfer. For example, the called and calling parties must establish

- the encoding mechanism for the audio, or video data,
- the transport addresses to be used to transfer the voice data (IP address, transport protocol and port number),
- the bandwidth requirements,
- proper authorization for initiating and accepting a call,
- call transfer, call diversion,
- the location of the called party, etc.

In addition, the call signaling must provide for an interface between the existing telephone system and the IP telephony system. The following sections explore the current status of VoIP signaling protocols.

1.5 VoIP signaling protocols

The ITU-T H.323 protocol suite is the dominant VoIP protocol suite as measured by the number of commercially available products. The H.323 protocol suite is also the dominant VoIP protocol suite as measured by the size and complexity of the specifications. This leads to a steep learning curve for everyone involved. The end result is high cost of implementation, high connection set-up latency, and difficulty in achieving interoperability in heterogeneous networks.

Although the ITU-T H.323 protocol suite currently dominates voice over IP (VoIP) protocols, there exists a lightweight contender for call signaling that avoids all the complexity, high connection set-up latency and implementation difficulties of H.323. The Session Initiation Protocol or SIP brings simplicity, familiarity and clarity of purpose to IP telephony that Internet savvy network professionals will appreciate.

Whether you are a next generation telephony service provider, a network or IT manager, or an established telephony carrier breaking into VoIP, SIP is a protocol that you are likely to encounter. Network professional should welcome SIP as a text-based, call signaling protocol that leverages the power of the Internet by borrowing such common elements as the format of HTTP, Domain Name System (DNS) and email style addressing.

2. INTRODUCING SIP

SIP is a text-based protocol that leverages the power of the Internet by borrowing such common elements as the format of HTTP, Domain Name System (DNS) and email style addressing. Further, SIP makes use of the Session Description Protocol or SDP for specification of the session parameters. SIP provides the necessary protocol elements to provide services such as call forwarding, call diversion, personal mobility, calling and called party authentication, terminal capabilities negotiation and multicast conferencing.

Like H.323, SIP is independent of the underlying network and transport protocols. SIP may be run over IPX, frame relay, ATM AAL5 or X.25. The most common and typical use of SIP is over UDP. All subsequent examples will address only SIP running over IP networks. This paper presents the SIP architecture, examines the operation of SIP, and compares SIP to H.323.

The most fundamental SIP operating model is that the calling client sends a call invitation to the called client. The called client may accept the invitation, acknowledging it with a response of “OK”. Finally, the calling client will “close the loop” with the called client by send an acknowledgement back to the called client. Figure 3 depicts this call setup process. Audio and/or video data exchange follows the call setup process.

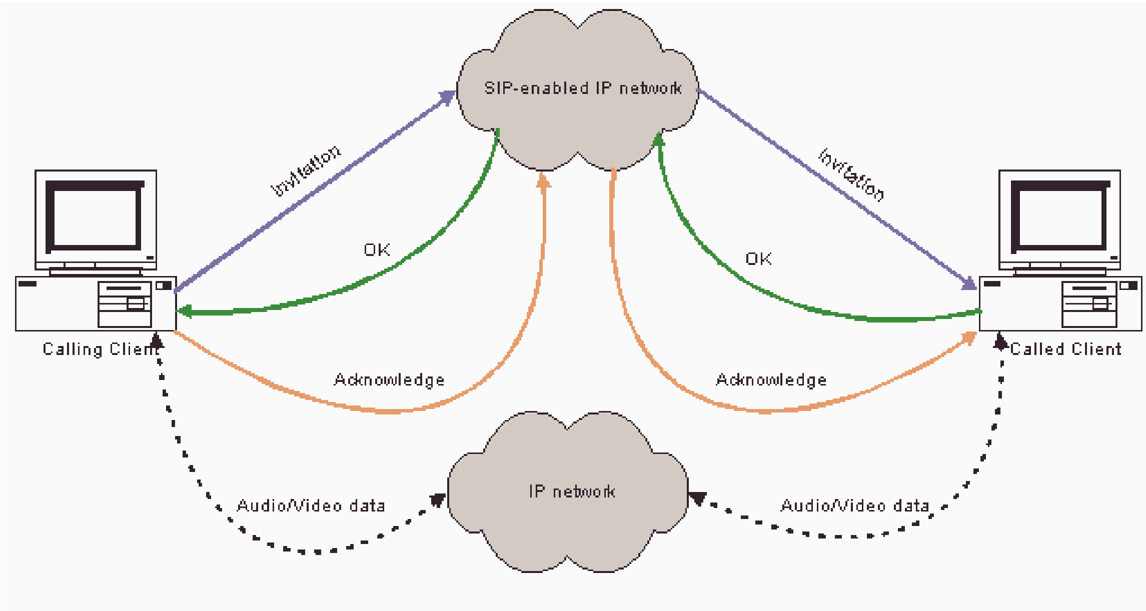

Figure 3 SIP operation

The SIP-enabled network is the next topic of discussion. A SIP-enable network is an IP network that includes the usual components and services such as routers, and DNS servers, as well SIP servers. The function of the SIP server is to support SIP-based telephony by providing a single access point for locating clients, mapping friendly names to addresses, and re-directing requests. There are two types of SIP servers, the proxy server and the redirect server. SIP operation is dependent upon which type of server is used. In the proxy model, the SIP proxy server is the only point of contact that the clients have for signaling messages. In the redirect model, the SIP redirect server lets the calling client know the location of the called client and then gets out of the way of subsequent signaling messages. The following sections will describe in greater detail the operation of SIP using the proxy and redirect models.

3. SIP OPERATION

In their simplest form, SIP calls involve two SIP clients and a SIP server. There are two distinct SIP call models, the proxy model and the redirect call model. The calling client sends an invitation to the called client directly or through the SIP server. The SIP clients locate the SIP server by a configuration parameter similar to the proxy-server parameter of Internet browsers.

3.1 Proxy server operation

The proxy call model makes use of a SIP proxy server. The SIP proxy server plays a roll similar to that of the HTTP proxy server in an HTTP system. The proxy server routes signaling messages between the called and calling parties. The RTP audio or video packets are sent directly between the clients after the call has been established. Figure 4 presents a typical call signaling procedure using a SIP proxy server.

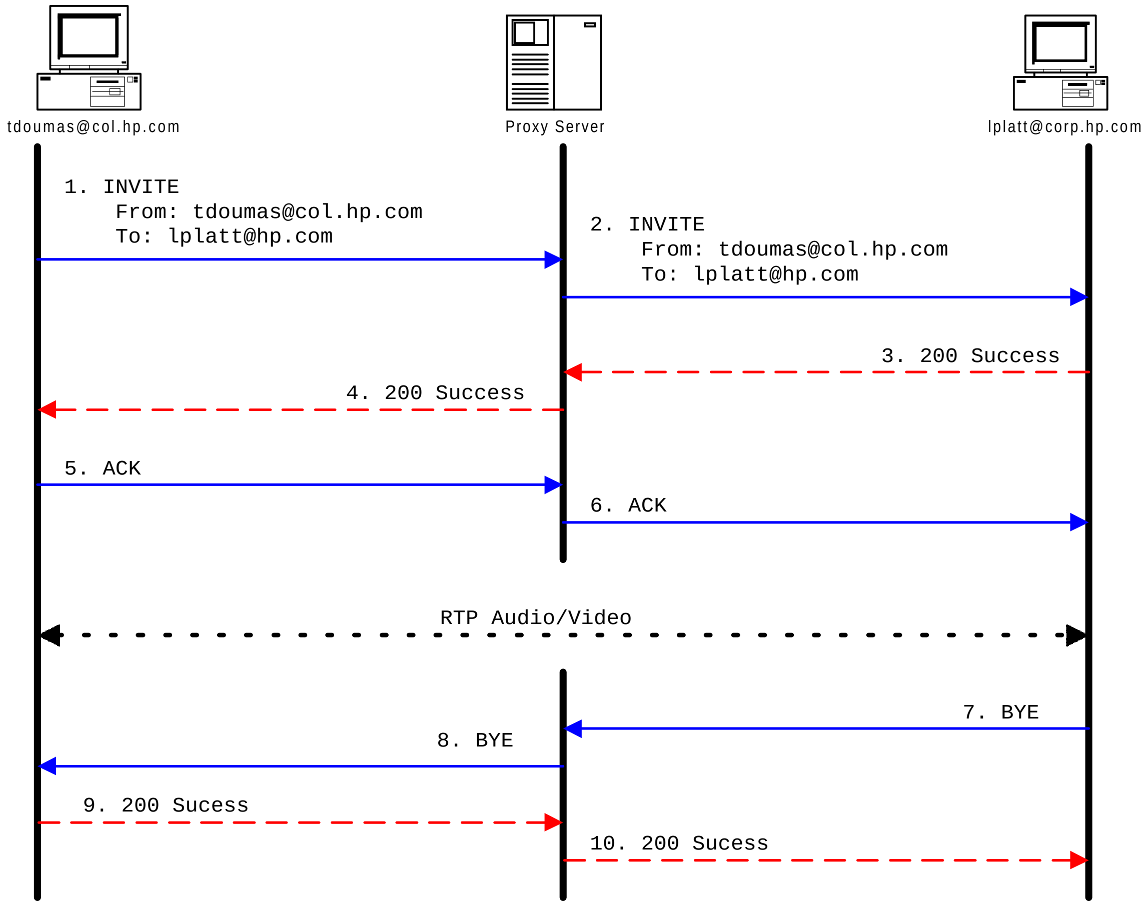

Figure 4 Proxy server operation.

3.2 Redirect server operation

The redirect server model makes use of a SIP redirect server. The SIP redirect server informs the calling client of the SIP URL for the called client. The calling client then proceeds to set up the call directly with the called client. Figure 4 presents a typical call.

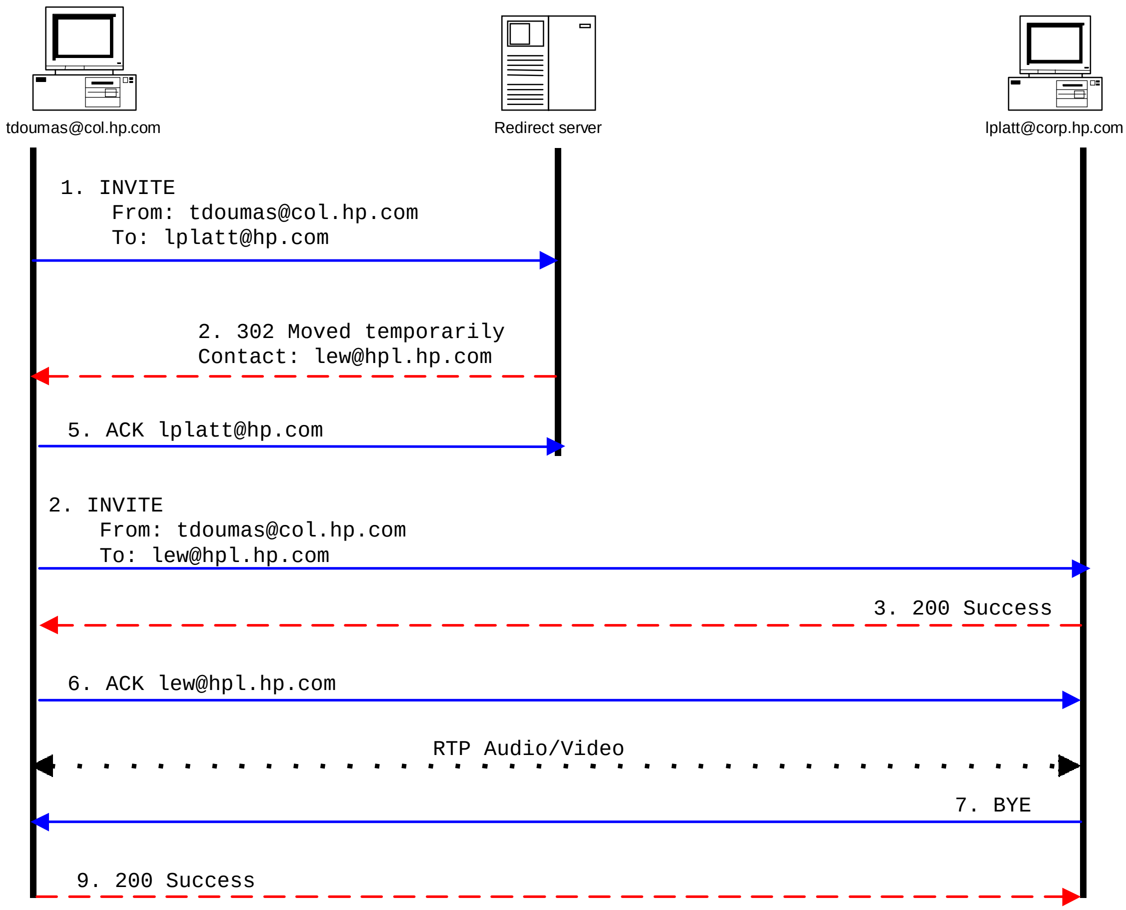

Figure 5 SIP redirect operation

3.3 SIP addressing

The preceding figures presented familiar looking addresses for the SIP clients. SIP clients are identified by a SIP URL. The SIP URL follows the `“user@host”` form of email addresses. The user part may be a user name or a telephone number. The host part may be a domain name, a CNAME or a numeric network address. Some typical SIP URLs are:

- `sip:elvis@graceland.com` (host independent)
- `sip:bill@oval.office.gov` (host specific)
- `sip:+1-800-555-1212@information.att.net`

A common user need might by to call a person for whom the SIP URL is not known. It is easy to imagine a process of using WWW search engines and directory services to resolve a name. For example, let’s say that a user wants to call the governor of Minnesota. A WWW search engine might turn up the common name “Jesse Ventura”. A directory service would resolve the name to a host-independent URL such as “jesse.ventura@minnesota.gov”. A SIP server would resolve the host-independent URL to a host-specific URL such as “sip:jventura@governorsmansion.minnesota.gov”. Ultimately, a DNS server would resolve the address to a specific numeric IP address.

SIP URLs may also specify the specific port on the host, the transport protocol (e.g. UDP) and a multicast address.

4. SIP MESSAGES

SIP defines two basic message types, the request and the response. The request messages are used to initiate, confirm, modify and terminate calls. The response messages are used to convey either provisional information such as “ringing”, or “moved temporarily” response or final information such as “busy” or “does not exist”. Table 1 presents the SIP request methods.

Request method	**Purpose**
INVITE	Initiate a call.
ACK	Confirm the final response to an INVITE.
BYE	Terminate a call.
CANCEL	Cancel searches and “ringing’
OPTIONS	Communicate features supported
REGISTERED	Register a client with a location service.

Table 1 SIP Request methods

4.1 The INVITE message

The `INVITE` message is sent by the calling client to initiate a call with another client. There are 5 mandatory parameters for the `INVITE` message. Table 2 presents the mandatory parameters of the `INVITE` message.

Parameter	**Description**
Call-ID	Uniquely identifies a particular invitation.
CSeq	A monotonically increasing sequence number used to identify the sequence of requests associated with a given Call-ID.
From	A SIP URL that identifies the initiator of the request. May include a “friendly name”.
To	A SIP URL that identifies the recipient of the request. May include a “friendly name”.
Via	Indicates the path taken by the request so far. The Via parameter is used to prevent looping of requests, assures that replies take the same route as

	requests and assists in unusual routing situations.

Table 2 `INVITE` parameters

4.2 Response messages

SIP Response messages indicated either call progress information or final status information. The response messages contain a `Status-Code` and a `Reason-Phrase`. The `Status-Code` de is a three digit integer that indicates the outcome of the request. The `Reason-Phrase` provides a textual description intended for humans. *Table 3* presents a summary of SIP response message categories and their use.

Status-Code	**Category**	**Example information**
1xx	Informational	trying, ringing, call is being forwarded, queued
2xx	Success	OK
3xx	Redirection	Moved permanently, moved temporarily, etc
4xx	Client error	Bad request, unauthorized, not found, busy, etc
5xx	Server error	Server error, not implemented, bad gateway, etc.
6xx	Global failure	Busy everywhere, does not exist anywhere, etc.

Table 3 Response codes

5. SDP

SIP makes use of the Session Description Protocol (SDP) to describe the attributes of SIP sessions. The SDP headers are encapsulated as the message body of a SIP request. In the H.323 world, SDP plays a similar role as that of H.245. Like SIP, SDP headers are encoded with ASCII text. The SDP headers are of the simple form `<type>=<value>.` The `<type>` is always a single character and `<value>` is a text string whose format is dependent on `<type>.`

SDP is not really a protocol as much as it is a format for describing multimedia sessions. SDP headers specify:

- Session name and purpose
- Time(s) the session is active
- The media comprising the session
- Transport address and media format of the session
- Bandwidth to be used by the session
- Contact information for the person responsible for the session

5.1 Media information

A key component of SDP is the description of the media of the session. SDP media descriptions include:

- Media type (audio, video)
- Transport protocol (UDP, TCP, RTP, etc)
- Media format (H.261, MPEG, etc)
- Multicast address for IP multicast sessions
- Transport port for IP multicast sessions
- Remote address for IP unicast sessions
- Transport port for IP unicast sessions
- Session start and stop times

6. SIP STACK

SIP is transport layer indepedent. It can run over any datagram or stream protocol such as UDP, TCP, ATM, etc). A common SIP implementation is on a TCP/IP infrastructure. The TCP/IP infrastructure provides inexpensive widespread connectivity, directory services, naming services and a widely known development environment. Possible implementation scenarios are an enterprise Intranet, a next-generation telephony service provider's private IP-based packet network or the Internet. The audio or video data streams are transported using RTP over UDP. The SIP call signaling messages can be carried over UDP or TCP with UDP being the preferred method. One important consideration when using SIP over UDP is that the entire message should fit within a single packet. If a SIP message is fragmented into multiple datagrams, the probability of losing the entire message increases with the number of fragments.

Figure 6 presents an example SIP running in a TCP/IP environment.

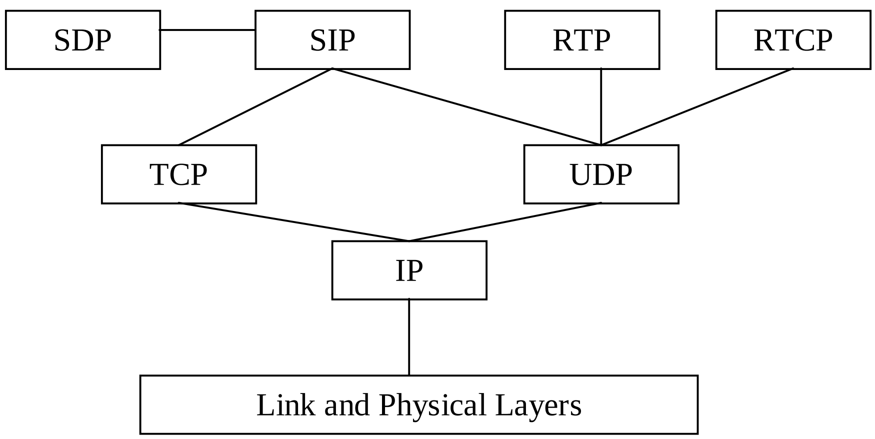

Figure 6 SIP protocol stack.

7. SIP VS. H.323

SIP together with SDP provide the same basic call signaling, call control services and supplementary services as H.323. Table 4 presents the various telephony services and their associated protocols for H.323 and SIP.

It is clear from the table that H.323 makes use of more protocols and there must define the interactions between multiple protocols. The net result is that more effort is required to implement an H.323 device. In addition, the computing resources for an H.323 device are greater than that of a SIP device. In fact a fully functional SIP client can be implemented with only two months of engineering effort [1].

Service/Element	H.323 protocol	SIP protocol
Friendly name mapping	RAS	Existing directory service
Server discovery	RAS	Domain Name System
Authentication, security	RAS, H.235	SIP/SDP, web infrastructure
Call signaling	Q.931	SIP
Terminal capabilities exchange	H.245	SIP/SDP
Supplementary services	H.450.1, 2, and 3	SIP/SDP
Audio/Video transport	RTP	RTP
codecs	ITU-T only	Any IANA registered.
Session descriptions	H.245, ASN.1 PER	SDP

Table 4 SIP/H.323 protocols

A comparison of specification quantity demonstrates one measure of the difference between SIP and H.323 complexity and difficulty of implementation. Table 5 presents a comparison of the size of specifications of SIP and H.323.

H.323		SIP	
Specification	**Length (pages)**	**Specification**	**Length (pages)**
H.323 Version 2	117	SIP	153
H.225.0 Version 2	171	SDP	42
H.245 Version 3	354		
X.691 (ASN.1 PER)	70		
H.450.1	22		
H.450.2	47		
H.450.3	65		
TOTALS	**846**		**195**

Table 5 SIP/SDP Complexity

SIP, because it is text-based simplifies the implementation task. Scripts may be implemented in Perl, Tcl or other text based scripting languages. On the other hand implementation of an H.323 based device requires an ASN.1 PER encoder/decoder. This is a formidable undertaking requiring not only the development of the encoding and decoding software but also the external specification of the semantics. The code space requirements will restrict the use of H.323 in inexpensive client devices. In addition, simple scripts cannot easily manipulate binary ASN.1 PER encoded data.

8. TESTING SIP

This section presents some testing issues related to SIP systems. The focus is on the problems that might occur during the call setup procedure.

8.1 SIP proxy address

In either of the SIP operating models, the SIP clients must be properly configured with the name or address of the appropriate SIP server. SIP relies on the Internet Control Message Protocol (ICMP) to return error information to the client in the event of an unreachable server. In analyzing a SIP sequence, one might see the INVITE message transmitted and an ICMP "Host unreachable", "Network unreachable" message returned. This might indicate that the IP address configured for the SIP server is incorrect.

Troubleshooting such a problem can easily be accomplished with a protocol analyzer that has the capability to search for ICMP messages.

8.2 Called party URL

If the IP address of the SIP URL for the called party is incorrect, then the INVITE message that is forwarded from the SIP proxy server would result in an ICMP message similar to that in the preceding example of an incorrect SIP proxy address.

The SIP URL may include a UDP port number for the called party. If the UDP port number is not included, then the assumption is that the default UDP port of 5060 is to be used. In the event that the called party is not using the port number expected by the calling party, then an ICMP message will result. In this case, the ICMP message would indicate "Port unreachable".

8.3 Proxy specific problems

There are specific rules for the behavior of SIP proxies. One key rule is that proxies must not reorder or modify fields in the SIP header (except Via). This includes the restriction that proxies must not change how fields are split across multiple lines.

A simple comparison of the fields transmitted to the proxy with the fields that the proxy forwards to the called client would illuminate such a problem with the implementation of a proxy. This comparison could be easily accomplished with a protocol analyzer that can decode and present the fields of SIP in a user-friendly manner.

8.4 Message size problems

When running over UDP, SIP messages will generally be small enough to fit into a single packet. If they do not, then fragmentation at the IP layer will break the message up into multiple datagrams. Although SIP supports fragmentation.The probability of a lost datagram increases with the number of datagrams. Since the entire message must be re-transmitted if a single datagram is lost, fragmentation would negatively impact the performance and reliability. Packet sizes that are at the limit of Ethernet would be indicators of fragmentation and possible SIP problems. Further analysis of the IP header using a protocol analyzer would isolate if problems were being caused by fragmentation when using SIP over UDP.

9. HEWLETT-PACKARD SOLUTIONS

Hewlett-Packard's Internet Advisor provides support for SIP troubleshooting. Protocol decodes provide a detail view of the fields of each protocol. **Error! Reference source not found.** presents a typical SIP message sequence. The summary view shows a high level view of the sequence, while the detailed view shows each SIP message header.

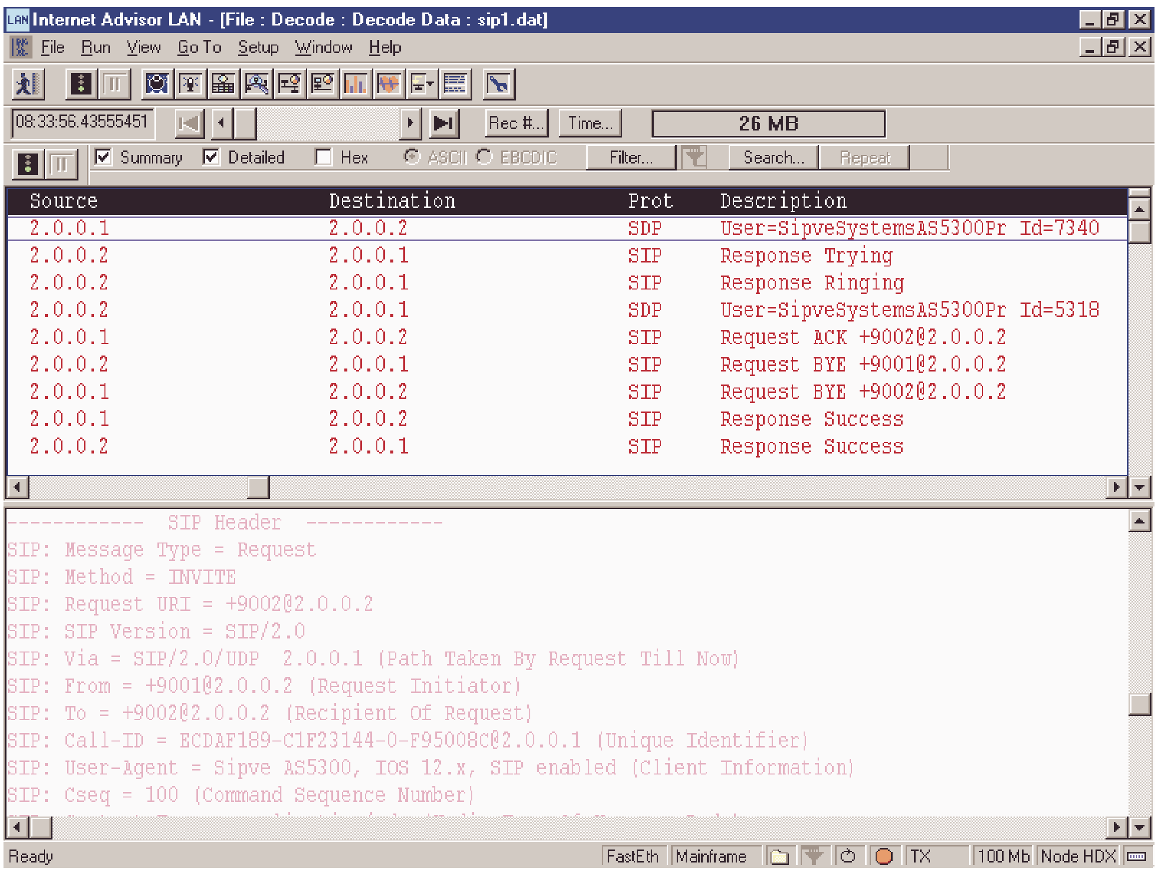

Figure 7 SIP message sequence

Figure 8 presents the details of the SDP headers.

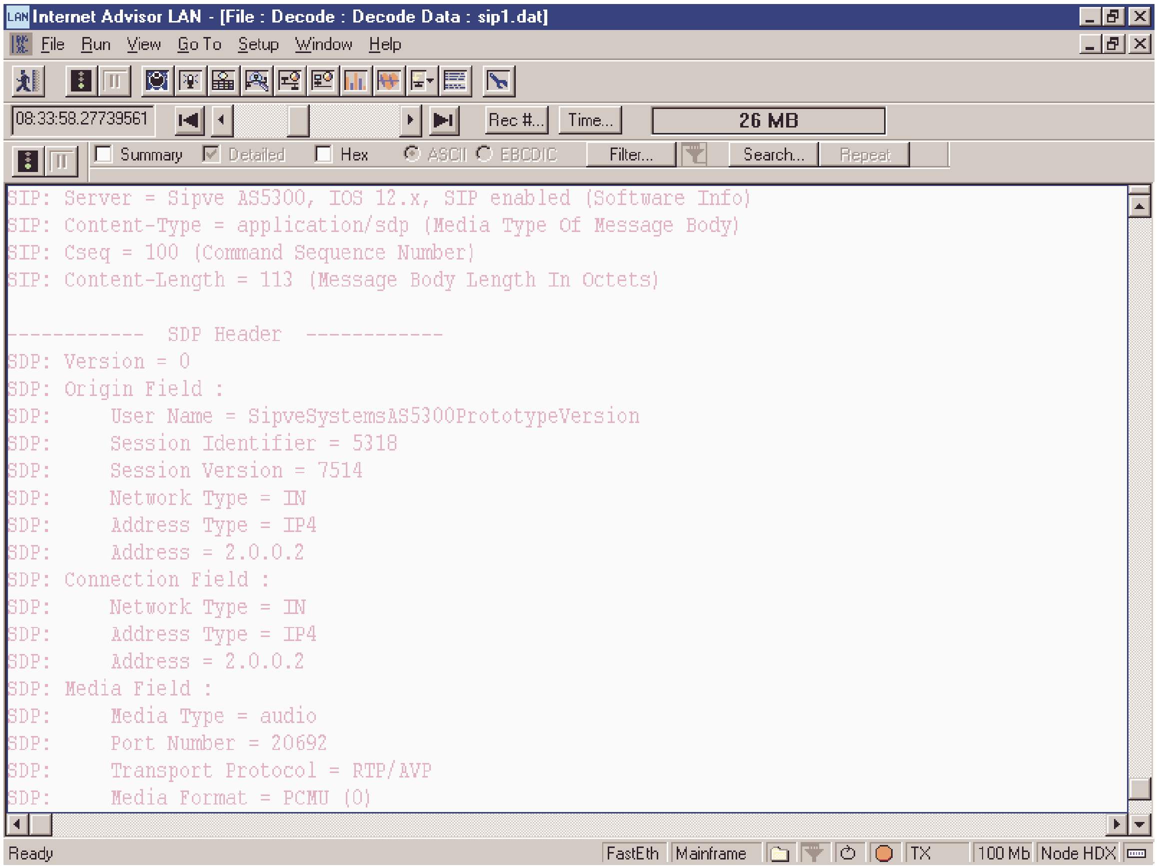

Figure 8 SDP details

In addition, the Hewlett-Packard Internet Advisor provides filtering and searching to allow the network troubleshooter the ability to display the desired frames. The statistical analysis may be used to monitor the health of the network and to “drill down” to the connections or protocols that are consuming bandwidth.

10. VOIP FUTURE

There is one. Details at 11:00.

11. REFERENCES

[1] H. Schulzrinne, J. Rosenberg, “A Comparison of SIP and H.323 for Internet Telephony”, Technical Report, Columbia University, New York, New York, 1998

12. FURTHER INFORMATION

Further information about SIP may be obtained from the IETF drafts and RFCs:

M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg, "SIP: Session Initiation Protocol", Request for Comments (Proposed Standard) 2543 , Internet Engineering Task Force, April 1999.

RFC 2327 Session Description Protocol.

13. BIOGRAPHY

Thomas Doumas is a member of the technical staff at the Hewlett-Packard Network Systems Test Division (NSTD) since 1984. Thomas has been involved in the management and software development of network test equipment for WAN, ATM and LAN networks. Thomas holds a BSEE and MSEE from the University of Wisconsin-Madison. Thomas' area of study focused on microprocessor based instrumentation.