SIP MARKET OVERVIEW

An analysis of SIP technology and the state of the SIP market

September 2003

Jonathan Cumming
Director of Marketing, Protocol Software, Data Connection
Jonathan.Cumming@dataconnection.com

Data Connection Limited
100 Church Street
Enfield
EN2 6BQ
United Kingdom
http://www.dataconnection.com



Copyright © 2003-2004 Data Connection Ltd. All Rights Reserved

Executive Summary

Session Initiation Protocol (SIP) is continuing to develop rapidly and it is difficult to keep up with all of its innovations and uses. This white paper is aimed at people who want to understand the concepts and drivers behind SIP adoption, and how it is evolving to face new challenges.

This paper summarizes where SIP has come from, how it works, and what makes it such a useful protocol. It then describes how SIP is used in applications including telephony, conferencing and messaging, and how it is being extended to provide innovative services and accommodate the requirements of real-world deployment, where NATs, service level agreements and regulators exist.

In covering this broad range of SIP-related topics, it provides a summary of the state of this increasingly important protocol.

About the Author

Jonathan Cumming is Director of Marketing, Protocol Software at Data Connection. Previously, he was development manager for DC-SIP, Data Connection's SIP User Agent and Proxy Server Toolkit, and retains product management responsibility for the product.

Jonathan has over 15 years' experience in the communications software industry. He holds an MBA from INSEAD and an Engineering degree from Cambridge University.

Table of Contents

1 Introduction 1
1.1 SIP concepts 1
1.2 Definition of terms 2
1.3 Where is SIP discussed? 5
2 History 7
2.1 The origins of SIP 7
2.2 How SIP developed 7
2.3 The return to reality 8
3 SIP applications 9
3.1 Telephony 9
3.2 Instant Messaging (IM) 10
3.3 Presence 11
4 SIP deployments 13
4.1 Existing SIP services 13
4.2 Interoperating with other protocols 14
5 Issues complicating SIP deployment 16
5.1 Reliability 16
5.2 Security 18
5.3 Quality of Service (QoS) and Resource Reservation 20
5.4 Scalability 22
5.5 Accounting 23
5.6 Privacy 24
5.7 NAT and Firewall traversal 25
5.7.1 Types of NAT 25
5.7.2 Using SIP through NATs 27
5.7.3 Application Level Gateways (ALGs) 28
5.7.4 Devices behind the same NAT 29
5.8 Device configuration 30
5.9 IPv6 30
6 SIP and the PSTN 31
6.1 Interoperability 31
6.1.1 Overlap signaling 31
6.1.2 Early media 32
6.1.3 Application Control with a traditional phone keypad 33
6.2 Regulatory requirements 34
6.2.1 Wire-tapping 35

Copyright © 2003-2004 Data Connection Limited. All Rights Reserved.

http://www.dataconnection.com

6.2.2 Emergency calls 36
7 Enhanced applications for SIP 38
7.1 Mobile (3G) 38
7.2 Caller preferences 39
7.3 Third party Call control 39
7.4 Conferencing 41
7.5 Click-to-call or click-to-dial 43
7.6 ENUM 44
8 The future 45
9 Further information 47
9.1 Web-sites 47
9.2 IETF RFCs and drafts 47
9.2.1 Application Control with traditional keypad 47
9.2.2 Early media 47
9.2.3 Overlap dialing 47
9.2.4 3G Mobile 48
9.2.5 AAA and security 48
9.2.6 Caller Preferences 48
9.2.7 Conferencing 48
9.2.8 NAT and firewall traversal 48
9.2.9 Device configuration 49
9.2.10 Presence and Instant Messaging 49
9.2.11 QoS 49
9.2.12 Other documents 49
10 About Data Connection Limited (DCL) 50

Copyright © 2003-2004 Data Connection Limited. All Rights Reserved.

1 Introduction

Session Initiation Protocol (SIP) is a signaling protocol for controlling multi-media sessions. In other words, it provides a way to establish voice, video and messaging communication between devices. From its initial use in Internet Telephony, SIP is spreading into many new areas, including advanced telephony applications, conferencing and instant messaging, and its functionality is expanding to meet the new requirements from its increased scope.

This paper provides an overview of the current state of SIP, and explains both the technology and the business requirements that are driving development in order to give a context in which to understand the issues involved.

This document is not a SIP primer, although it does explain the main concepts and terms that SIP uses, and is aimed at people who are

- working with SIP and wanting to increase their understanding of other ways that it is used
- looking at developing or deploying SIP-capable devices
- just interested in understanding SIP a bit better.

As with any fast-moving field, any document that describes the current state of the market is always out of date, so this paper provides a snapshot from September 2003. Nevertheless, the concepts on which SIP is based and the problems that it addresses do not change, so the majority of this information will remain valid even if the details have altered. The further information section should provide useful pointers for anyone who wishes to investigate particular areas in more detail.

1.1 SIP concepts

SIP's view of the network matches that used in the Internet: intelligent devices communicate directly with each other over a simple transport infrastructure. This contrasts with the traditional telephone network, where transport between dumb endpoints is provided through an intelligent network core that is an active party in any conversation.

This difference allows the network to become a commodity and allows any device attached to the network to provide a service to any other. This increases competition, which drives down prices, and helps innovation, because the investment required to set up a new service is very small. With the traditional intelligent telephony network, only the telephone company can provide new services, and this requires the network core to be upgraded, which is an expensive and slow process.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

While the above explains why IP telephony is helping to drive down the general cost of telephony, and why there is a high level of SIP innovation, the following SIP features show why it is such a powerful framework.

- Mobility: SIP allows a client to register dynamically with a fixed location, so that calls can be routed to it using a well-known address, similar to an email address.
- Flexible message structure: SIP's message structure makes it much easier to extend for new applications than equivalent existing protocols, such as H.323 which uses the ITU's opaque ASN.1 encoding standard instead of text, and it is seen as being much simpler and more flexible.
- Distribution of function between devices: SIP enables requests to be dynamically routed through different devices, enabling functionality to be distributed and requests routed through the relevant devices.
- Negotiation of supported features: This makes SIP very adaptable, as the media and protocol extensions to be used for a particular call are negotiated between the clients on that call. As a result, SIP can be used to set up any type of media conversation, including voice, video and messaging.
- Separation of signaling and media: In SIP, the paths of the signaling and the media are totally independent. The signaling and media may traverse different routes through independent sets of devices on different physical networks.
- Forking: This allows multiple devices to be associated with a single address, so that all or a selection of these devices can be contacted simultaneously or sequentially, according to local policy.

These features are equally applicable to many areas, including telephony and messaging, and have been the drivers for SIP's adoption by the major players in these fields.

1.2 Definition of terms

SIP communication is made up of messages that are sent between the devices using UDP, TCP, or another transport protocol. These messages are either requests or responses and contain a set of headers, which are the parameters of the message, and one or more message bodies, as required by the application.

A single SIP request and all its responses form a SIP transaction. Different types of transaction are used for different protocol functions. For example, an INVITE starts a telephone call, and a MESSAGE sends an instant message.

A SIP dialog is a persistent link between two devices that is used to associate transactions and to provide ordering between them. SIP transactions can exist within or outside a SIP dialog, and transactions are used to establish and terminate dialogs. For example, in telephony, the initial INVITE that starts the call also establishes a dialog between the participants. To end the call, one participant sends a BYE within the context of this dialog. This BYE transaction terminates both the call and the associated dialog.

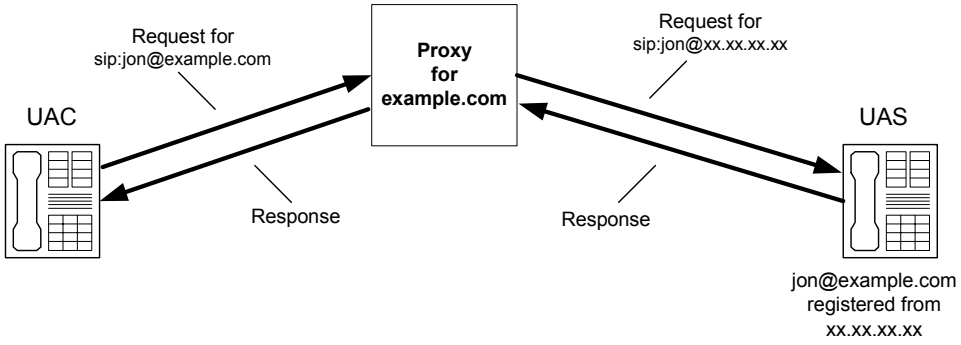
Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

The high-level concept of a call does not simply map to a SIP dialog, because a single telephone call may include conversations with several people and devices, for example receptionists and voicemail systems. These individual connections need separate SIP dialogs, so the call can contain multiple dialogs. SIP messages contain a call identifier field (Call-ID) that is sometimes used to link the dialogs and transaction into an application-level concept of a call, although this use is strictly outside the standard.

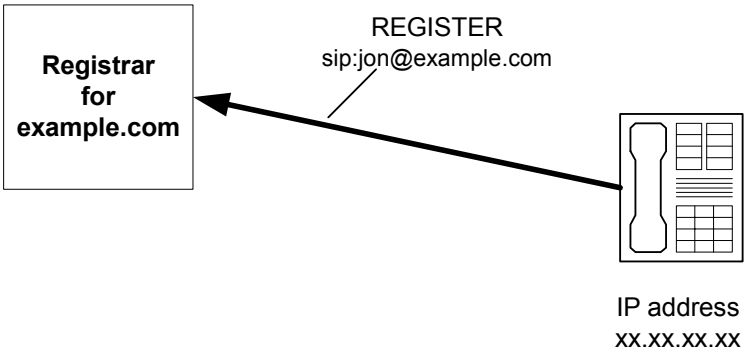
The following terms are used to describe SIP devices.

- **User Agents (UA)** are endpoint devices that terminate the SIP signaling. They can be clients (UAC) that initiate requests, servers (UAS) that respond to requests, or more normally a combination of the two.

- **Proxies** are devices in the signaling path between User Agents that route requests on towards their destination. They may add parameters to the requests and may reject requests, but they may not initiate requests or respond positively to any request that they receive. Proxies pass unrecognized messages through unchanged; this means that many new features can be deployed in a network by upgrading only the User Agents and leaving the proxies to continue with their default behavior.



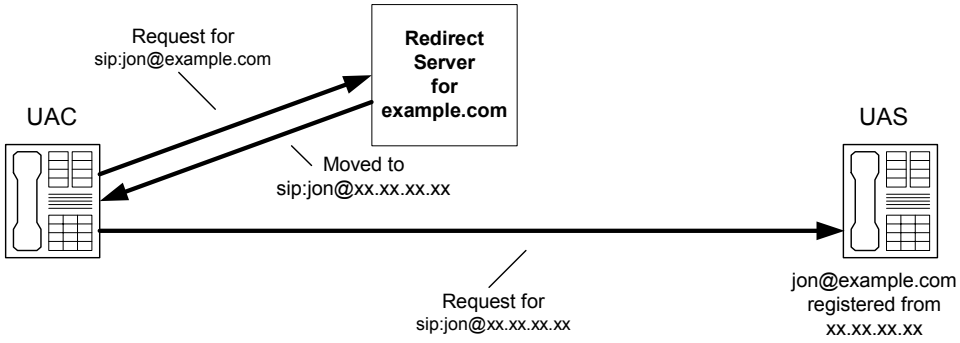
- **Registrars** are specialized User Agent Servers that handle REGISTER requests. SIP devices use REGISTER requests to dynamically register their current location, and this enables them to be contacted when mobile.



The registrar now knows the current IP address at which jon is reachable.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

- **Redirect Servers** are specialized User Agent Servers that respond to requests by redirecting them to another device.



  The redirect server responds to the request containing the address to which the request should be redirected.

Many real devices contain several of the above elements. For example, a Registrar will normally be linked with a proxy or redirect server, so that the proxy or redirect server can use the location information that it receives to send requests on to the registered devices.

However, the action that a device takes on receipt of a SIP request is not determined purely by the protocol; it is also determined by the application. An application may decide to forward the request on to another server for further processing, such as authentication, instead of forwarding it directly to its destination. The generic term for such a device is an **application server**. From a SIP view, an application server may behave as a User Agent, a Proxy or a combination of the two, depending on the situation.

A common configuration is what is known as a Back-to-Back User Agent (B2BUA) where the device is similar to a proxy in its behavior, but actually terminates the SIP signaling on both sides, so that it can initiate requests to control the dialogs passing through it. This requires that the B2BUA is a trusted party in the communications, which prevents end-to-end encryption and authentication of the messages.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

1.3 Where is SIP discussed?

The main forum of SIP standardization is in the Internet Engineering Task Force (IETF), which is the primary standards body for Internet protocols. The IETF has set up the following three working groups to work on the protocol and its application.

- The SIP working group covers enhancements to the core protocol.
- The SIPPING working group covers applications of SIP.
- The SIMPLE working group covers Instant Messaging and Presence applications of SIP.

The distinction between these groups is that the SIPPING and SIMPLE working groups discuss applications of SIP and decide how SIP should be used in each of them. If they determine that the requirements of a particular application cannot be handled by the core protocol, then these requirements are passed to the SIP working group for a solution. This enables the SIP working group to maintain control over extensions to the protocol, while limiting the scope of its discussions.

Other IETF working groups whose areas touch on SIP include the following.

- IPTEL (Internet routing of telephone calls)
- MMUSIC (responsible for Session Descriptor Protocol (SDP), which SIP uses to describe its media sessions)
- MIDCOM (Middlebox communication – firewall and NAT traversal)
- SPIRITS (PSTN – Internet telephony interoperation)
- ENUM (Internet use of traditional PSTN phone numbers)

Several industry groups are also discussing how to standardize the use of SIP in their environment. These include

- Packetcable (www.packetcable.com), who are using SIP for telephony over cable
- 3GPP (www.3gpp.org), who have adopted SIP for 3G mobile
- Multi-service Switching Forum (MSF) (www.msforum.org), which has defined SIP-T conformance levels and is now working to ensure that SIP can be deployed in large scale PSTN networks.
- ETSI TIPHON (Telecommunications and Internet Protocol Harmonization Over Networks) (www.etsi.org), who are working to ensure that SIP is suitable for deployable telephony applications.

There is a continual conflict between the requirements of the traditional telephone providers, who need to provide an end-to-end billable solution that meets their regulatory requirements, and the less controlled environment of the Internet. This is resulting in concern over the interoperability of the different flavors of SIP, including 3GPP SIP, PacketCable SIP, and IETF SIP, and discussions are ongoing to ensure that they all work together.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

There is a separate initiative to standardize the programming interfaces to SIP and other telephony protocols. This work covers the following interfaces.

- JAIN (java.sun.com/products/jain) – Java APIs to SIP and other Next Generation telecom protocols.
- Parlay (www.parlay.org) – High-level, protocol independent APIs that allow the development of telecommunications applications that are independent of the underlying network.
- Call Processing Language (CPL) - XML-based language that can be used to describe and control Internet telephony services (draft-ietf-iptel-cpl-08).
- Common Gateway Interface (CGI) - HTTP CGI compatible extensions to providing SIP services on a SIP server (RFC 3050)

These standardized interfaces help the development of SIP applications that are not tied to a specific implementation of the protocol. This makes the resulting application more portable and reduces the developer's dependence on one supplier, but they can add a processing overhead that may reduce the overall efficiency of the system. The protocol independent interfaces also limit the ability to exploit the advantages of a particular protocol.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

2 History

2.1 The origins of SIP

SIP was originally developed around 1996 in an academic project to control multicast media distribution. Its message structure was based on SMTP (email), with the simple, text-based, extensible form that had helped to make email so successful. When interest in Internet Telephony increased, this initial work was used as the basis of the new protocol, and it was standardized by the IETF in March 1999 as RFC 2543.

SIP has since been extended for use in instant messaging and presence, and continues to find new applications in the establishment of sessions between devices whose location and capabilities may change.

2.2 How SIP developed

The initial work on SIP received strong backing from the venture capital community, with a number of well-funded companies set up to develop SIP-based products. This, together with adoption by MCI WorldCom, Cisco and ETSI TIPHON, led to an explosion of interest in the protocol.

Early standardization work concentrated on the use of SIP for Telephony (SIP-T), and it became clear that RFC 2543 would have to be extended in many ways to handle all the new requirements. The huge number of extensions that were proposed overwhelmed the SIP working group and led to long delays in their standardization. As a result, the standards lagged behind the requirements, and many new features were added through proprietary mechanisms. Although many of these extensions have now been either adopted as standards or replaced by standard mechanisms, this divergence has led to interoperability problems in function beyond that defined in the core specifications.

After three years of rapid development and extension to SIP's function, RFC 2543 was finally replaced in 2002 by a new set of SIP standards based on RFC 3261. These new standards clarified and extended the original protocol, and improved its scalability and security. Products supporting RFC 3261 are now appearing on the market, although support of some aspects of the protocol, for example transport level security (TLS), is still limited.

In around 2000, 3GPP (Third generation mobile) also selected SIP as the basis for its communications infrastructure, and, as a result, there has recently been a major drive to standardize the extensions required for mobile telephony.

Current work is focusing on areas including NAT traversal, conferencing and security. These and other areas are discussed in more detail in later chapters.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

2.3 The return to reality

The initial enthusiasm for SIP coincided with the Internet bubble, as SIP offered a way to replace the existing expensive telephone system. The combination of venture capital backing, which expected short-term returns, and over-optimistic claims from the protocol's exponents placed unrealistic demands on the protocol and the products being developed. This resulted in a drop in the quality of both the standards definition and the products that came to market, as competitors raced to support too many features. In addition, the impression that SIP was a simple protocol resulted in the development of many SIP implementations, written in different programming languages to different versions of the standard, and providing very different levels of quality and completeness. This caused real interoperability problems and raised concerns over SIP's fitness for any commercial purpose.

This "bad press" could have killed the protocol, but with influential backers, including Cisco, Microsoft and Nokia, and its fundamental strengths, SIP continued to develop and mature. Today, there are over 20 SIP-related RFCs and over 100 SIP-related drafts being discussed in the working groups, and almost every major telephony equipment manufacturer is developing SIP-capable products.

Interoperability is improving as the standards and the implementations mature. Traditionally, SIP interoperability has been determined at the closed-door SIPit events that are coordinated by the SIP Forum http://www.sipforum.org. However, although these events are invaluable for ensuring good interoperability, the results are confidential and cannot be used by a potential customer to determine whether particular devices are compatible.

SIP device resellers are therefore assembling product combinations that they have tested to offer complete solutions, but pressure from customers for a better measure of interoperability is encouraging the establishment of independent conformance tests for SIP devices. The first stage in this process is the definition of suitable sets of functionality that should be supported by particular devices. Once these are agreed, it will be possible to establish independent testing of any claims.

Various industry consortia, including the MSF and PacketCable, have developed conformance levels for their applications, and others, including the SIP Forum, are developing a more generic framework for SIP conformance. Many bodies are claiming to produce conformance test tools and programs, but until the standards and conformance levels have stabilized, these will only be able to validate basic functionality.

SIP products are also now being designed to handle real-world requirements of reliability, security and manageability, but SIP is still an immature protocol that has not been proven in large-scale deployments and it is still evolving to support more advanced applications. In normal operation, the protocol is fairly stable and robust, but some serious issues with the design of the protocol remain to be resolved. For example, there is continuing work to improve the handling of error conditions and the behavior under heavy load. These, and other major issues that must be considered when using SIP in a real environment, are discussed in more detail in Chapter 5.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

3 SIP applications

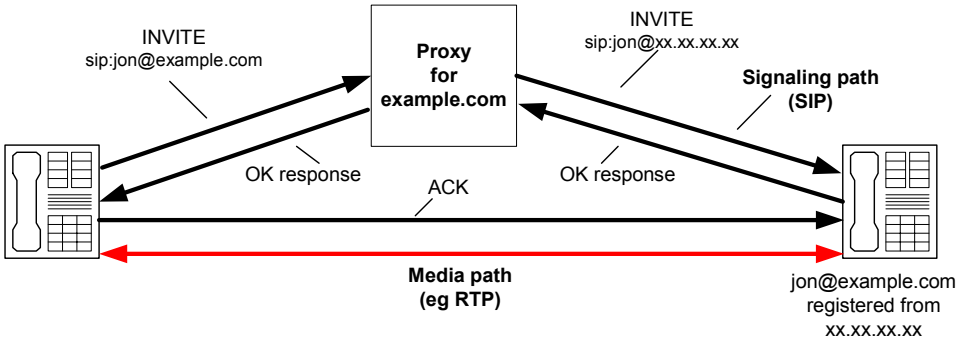
Current SIP use falls into three main categories: telephony (including conferencing), instant messaging and presence. The following sections describe how SIP works in each of these areas.

3.1 Telephony

Protocols for audio and video telephony are, in principle, straightforward in an IP environment, because the underlying network provides a routable infrastructure over which to send the media. However, a usable telephone requires additional features, including the ability to find the subscriber and to negotiate a compatible media type for the conversation.

To make a SIP telephone call, a SIP UA sends an INVITE request. In the message body of this request, it puts the SDP description of its available media channels. This request is forwarded by proxies across the network until it reaches its destination, or until it is rejected with an error response.

When the called UA receives the INVITE request, it checks whether it is capable of accepting the call, and then starts the phone ringing. In the meantime, it sends a provisional response back to the caller to tell it that the phone is now ringing. When the phone is answered, the called UA sends a final positive response with the SDP description of its media channels back to the caller. On receipt of this response, both parties now have the SDP descriptions of the other's media, and can establish the media channels agreed. The caller UA also acknowledges the successful receipt of the response by sending an ACK, which is a special type of request, back to the called UA.



Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

If, during the call, either party wants to change the media, for example to open a video channel, then it can send a re-INVITE (an INVITE within the established dialog) with an SDP body describing the new media. If acceptable, the recipient responds positively with its SDP. Otherwise, it rejects the request and the session continues unchanged. When either party wants to hang-up the call, it sends a BYE request.

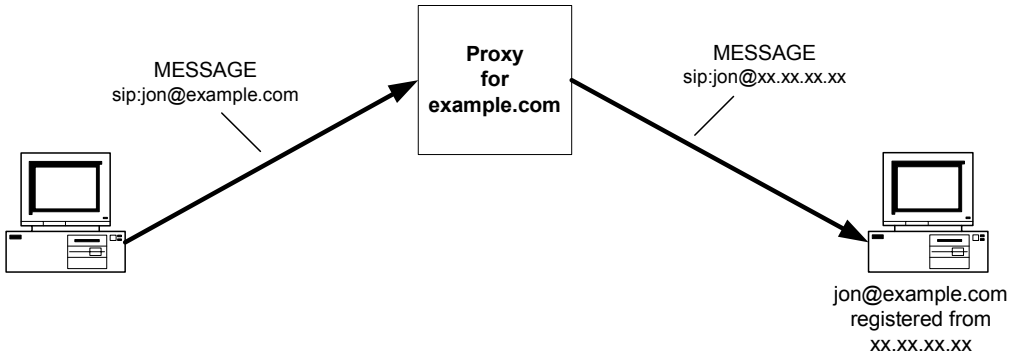
This set of primitives allows the establishment of a telephone service, but there are many complications and variations to this scenario; some of these are covered in Chapters 5 and 6.

3.2 Instant Messaging (IM)

IM provides the ability to send messages to other individuals. The underlying requirement is very similar to email, but the user experience is very different. Instant messages are analogous to the sentences in a conversation; they are normally short, informal, and expect a quick response. Email, on the other hand, is an electronic letter; it has a more formal structure and delivery process.

Many see IM as the next killer application. Existing IM services, as provided by Yahoo!, AOL and Microsoft, have been extremely successful, as has the analogous short message service for mobile phones (particularly in the UK).

SIMPLE (SIP Instant Messaging and Presence Leveraging Extensions) defines how SIP can be used for IM. It uses SIP registration to enable users to be contacted using their URLs, for example sip:jon@myserver.com, at a changing IP address. Messages addressed to the users are then redirected or proxied by their home server onto their current location.



SIMPLE defines the following two modes of operation.

- In page mode, every message is independent of every other. No persistent protocol-level connection is established between the User Agents, and each message is routed independently to its destination. This is directly analogous to the operation of email.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

- In session mode, a persistent connection is established between the two User Agents, and a separate media channel carries the message contents. This operates in the same way as in telephony, except that the media session that is established uses Message Session Relay Protocol (MSRP), as defined in draft-ietf-simple-message-session-01, rather than RTP.

The limitation of page mode is that there is no protocol-level link between messages. As a result, although the protocol provides a reliable transport, it lacks flow control and message ordering, and is therefore unsuitable for carrying large amounts of data or high message flow rates.

Page mode also sends all the data through the signaling channel and any routing proxies. This limits the scalability of the solution, because all the messages traverse the central routing proxies. This puts an unnecessary load on what may be a bottleneck, and restricts messages to types that are understood by all the devices in the signaling path.

In session mode, flow control and ordering of the data is provided by MSRP. The data is sent directly between User Agents or through specified message relays. This is normally a quicker route than sending through all the proxies in the signaling path, and it reduces the load on the proxies. For small numbers of messages in a conversation, session mode has a higher overhead because more SIP messages are required and the media channel has to be established and closed. For longer conversations or large amounts of data, session mode is more efficient because the media messages do not need to include the routing and authentication information that would be required in every page mode message.

In some environments, for example financial institutions, additional security or message monitoring is needed, which requires access to all the message contents at some intermediate monitoring device. In page mode, this can be provided in any of the proxies along the signaling path. In session mode, the message relays in the media path can be used instead.

IM is growing very fast, and the use of SIMPLE is growing at an even faster pace, due to the drive towards open standards and the benefits of compatibility between IM and telephony. Current implementations are based on page mode, but the use of session mode will increase, because its improved scalability is required for larger installations.

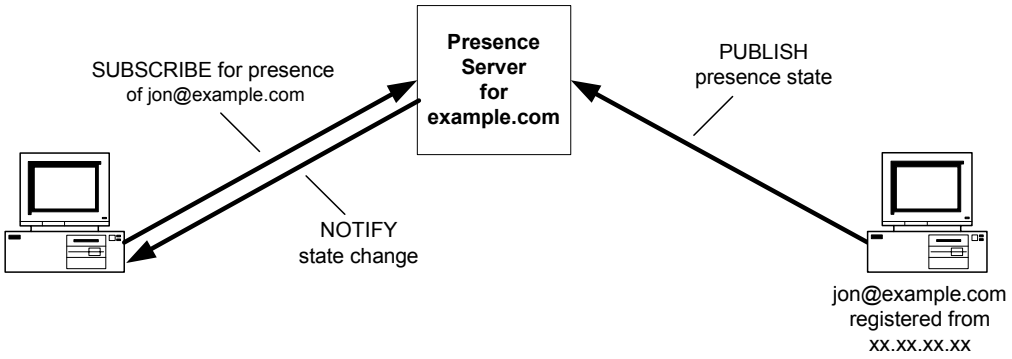
3.3 Presence

Presence is the ability to publish your state, for example whether you are at your desk, and to subscribe to other people's state and be notified when it changes. For example, this can be used to tell your colleagues whether you are available to take their calls.

Presence is handled in SIP using a generic event monitoring and notification mechanism, which is defined in RFC 3265 – SIP Specific Event Notification. This allows a device to subscribe to an event package that is supported by another device and to receive change notifications from it. Event packages define a set of state information for a specific context; for example, draft-ietf-simple-presence-10 defines the package for presence. Event packages are being defined for a wide range of applications.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

Presence also defines the concept of a presence server. A presence server collects the presence state from a set of devices, and enables a client to subscribe to it in order to receive notifications whenever the state of these devices changes. The advantage of a presence server is that an individual device only has to publish its state to a single server, rather than to each interested party, which aids scalability.



Presence is normally used with telephony or IM, and it is this combination that is so powerful. For example, an intelligent proxy can automatically route calls directly to your mobile phone when you are out of the office, or a conference server can start a conference and invite a pre-arranged set of participants as soon as all the key people signal their availability. However, the most common use of presence today is between friends and colleagues.

The use of presence in an informal environment works well, but there are privacy concerns when it is used more widely. In particular, who should be told what information about you, what is a suitable level of detail, and what are they allowed to do with this information? There are some very subtle effects of this; for example, will you appear rude or inefficient if you ignore a phone call after having published your availability? As a result of the increased information that is available about us, we are going to have to be much clearer about what information we want to give to whom, and how it might be used to monitor us. This issue is not completely solved, and it is discussed in detail later in section 5.6 on Privacy.

Finally, it is not clear how the increased information that presence provides will affect productivity; given that interruptions generally lower one's efficiency, and the existence of presence information is likely to increase someone's likelihood of contacting you, will the increased number of interruptions lower productivity, or will the time saved from only calling people when they are available and the increased responsiveness raise it?

Presence is an extremely powerful feature, as the earlier examples demonstrate, but it will be the societal issues that will limit its acceptance of presence, rather than any technical ones.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

4 SIP deployments

SIP can be used throughout a network: as a peer-to-peer protocol between endpoints, between the endpoints and the devices in the core, and between devices within the core. However, SIP can also be used only in parts of the network. The reduced scope of this sort of limited application makes it suitable for early adoption of the protocol, because it requires only a subset of function and interoperability with a limited range of devices. Today, SIP is being used in a range of situations: as an end-to-end protocol by early adopters, and as part of the telephone network to back-haul traffic over IP links between switches. It is therefore forming an ever-larger part of the network as the protocol matures.

The initial driver for SIP adoption in telephony was cost, but as the monopoly of telephony service providers has been reduced, prices have dropped in many markets to a level where cost is no longer a significant factor. For example in Japan, Yahoo!BB has been so successful at attracting customers to its SIP-based telephone services that NTT, the incumbent supplier, has been forced to respond with similar pricing plans.

In the future, SIP adoption will not be driven primarily by cost, but by the new services that it can provide and the convenience of converged voice and data networks.

4.1 Existing SIP services

Current SIP use falls into the following categories.

- Internet-only services. These consumer-orientated services provide a central SIP registrar and enable free calls across the Internet to other SIP phones. There may also be some interconnectivity with the PSTN, but only to freephone numbers and with limited ability to receive calls, because in both cases the party on the PSTN side pays for the call. No charges are levied and therefore minimal security and administrative overheads are required. The Internet provides the bandwidth for the SIP signaling and media.

  Free operators, including Free World Dialup, are offering this type of service as a loss-leader, in order to establish a strong market presence that they hope to be able to exploit in the future. There are strong precedents for this business approach on the Internet in the form of Google and Hotmail.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

- PSTN and Internet service. In addition to calls between SIP phones across the Internet, the service provider supplies PSTN gateways to allow calls to be made to PSTN numbers, and a phone number that allows calls to be made directly to the SIP phone from the PSTN. This requires a commercial arrangement between the user and the service provider, and Vonage, Deltathree and MCI (WorldCom) all provide this type of service. The overhead of maintaining this commercial relationship makes this commercially viable only for high volume users. However, where there is existing commercial relationship, for example with a DSL service provider, telephony offers a very easy add-on; this is the model being used so successfully by Yahoo!BB.

- Enterprise use. In this case, the service is provided within an organization for inter-office calls, and through gateways controlled by the enterprise into the PSTN. There is only a single commercial relationship between the enterprise and the telephone company, so this offers an efficient way to make a large cost saving.

- Specialized Use. SIP can also be used to back-haul traffic between particular switches, or to communicate between components within a single system. In these situations, SIP is only being used internally, so the business case is purely based on its effectiveness for the purpose against any competing technologies.

As Internet telephony becomes more popular, these models are likely to evolve into a structure that offers end-to-end SIP between what are currently islands of SIP, with the increased flexibility and functionality that this offers.

Practical deployment issues and governments regulations, including QoS, wire-tapping and access to emergency services, may restrict this spread, and these issues are discussed in the following chapters. In addition, the incumbent telephony service providers will attempt to restrict the growth of SIP telephony through regulatory pressures and predatory pricing.

4.2 Interoperating with other protocols

SIP is only one of many protocols being used to provide telephony and messaging services. There is therefore demand from customers to provide services between these protocols, and a number of manufacturers are developing gateways to do this conversion. Interoperability of basic function is normally straightforward, and the complexities arise when mapping more subtle concepts between the systems: for example state levels when the scales do not match, or permissions when the same group concepts do not exist.

There has been a great deal of work on interoperability in telephony in the various standards bodies to produce standard mappings. These include RFC 3398, which defines the mapping between ISUP and SIP messages to provide ISDN/SIP interoperation, and draft-ietf-sipping-qsig2sip-02, which proposes a mapping between Q.SIG and SIP.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

In IM, although the major providers have agreed to standardize on SIP, and many of their proprietary protocols are being phased out, the IETF is standardizing two IM protocols: SIMPLE, which is SIP-based, and Jabber, which is an XML-based standard from the open-source community. Both protocols provide similar functionality and will have to co-exist, and there are proposals to use SIP to establish Jabber sessions.

Standardization of protocol conversion is incomplete and some aspects will always remain proprietary, but significant work has been done to ensure interoperability across a heterogeneous network. This work will continue, driven by the need for SIP to be installed into existing environments and to interoperate with a huge range of existing devices.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

5 Issues complicating SIP deployment

Chapter 3 described how SIP can be used to provide simple telephony, IM and presence services. However, commercially deployable technologies require a far richer feature set, and the following sections cover some of the issues that need to be addressed in real products.

Although SIP standard solutions now exist for many of these areas, the required features are still missing from the current generation of SIP devices because this functionality has only recently been standardized. Therefore, the solutions described may not yet be deployable.

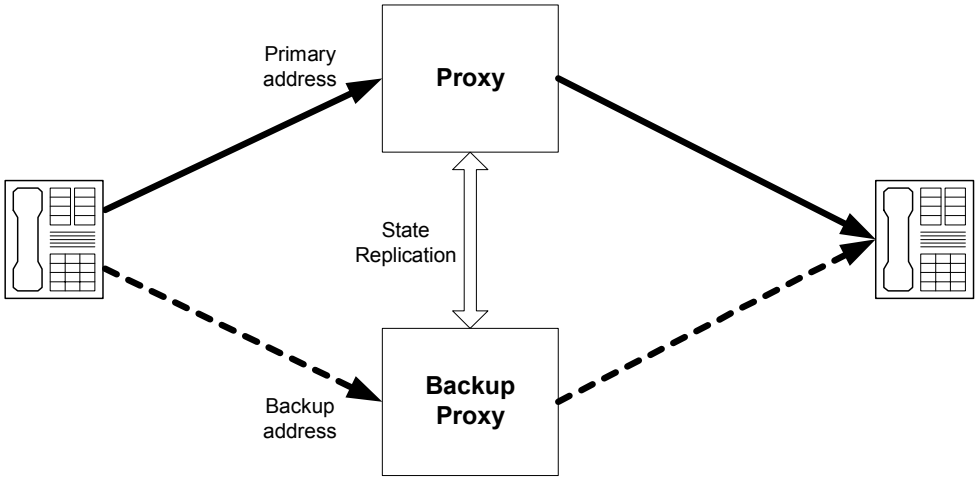
5.1 Reliability

Telephone services are expected to provide a very high level of reliability. This is often referred to as "5 9s" and indicates that the service should be available 99.999% of the time, or less than 5 minutes' downtime in any year, including system maintenance and upgrades. Mobile telephony and IM have traditionally had a lower level of reliability, but expectations even in these areas are rising as the technologies mature. Traditional PSTN equipment provides this level of reliability using expensive fault-tolerant hardware, but SIP attempts to provide it using Domain Name Service (DNS) to reroute the messages around failures.

DNS provides the mapping between services, domain names and IP addresses, and it allows multiple alternate domain names to be configured for a single service and multiple IP addresses to be configured for a single domain name. Using DNS, a SIP device can retrieve the list of alternate addresses and, if its request to the first one fails, it can automatically reroute the request to an alternate backup address.

Using DNS, it is possible to remove any single point of failure from the system, but this does require state replication between any stateful devices in the system. These will normally include any User Agents clients and any proxies that are controlling the allocation of resources.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.



However, for a SIP device to reroute a message requires it to detect that the initial request has failed, before attempting to use an alternative address. When using SIP, this detection mechanism may be very slow, particularly over UDP. In addition, each new request should also be routed using the same algorithm, so it too will be routed first to the failed server and will exhibit the same poor recovery characteristics. The issues raised by this are discussed in more detail in <draft-sparks-sip-noninvite-00>.

The use of a reliable transport protocol such as TCP or TLS, instead of UDP, greatly improves the speed of failure detection, but this relies on the failure to establish a reliable connection, which also takes time to detect. Proprietary mechanisms that continually monitor the status of partners are required for more responsive recovery. The use of such mechanisms is pushing the architecture towards that used in the traditional telephony network, where the transport layer continuously monitors the state of the links between a defined set of connected switches.

Alternative solutions use redundant hardware to provide failover within a single box, or in a cluster. These techniques enable the remote party to be reasonably unaware of a failover: any TCP or TLS connection and any outstanding transactions may fail, but existing SIP dialogs should continue unchanged. One complexity in these solutions is that the IP addresses must remain unchanged during any failover; this can be achieved using a load-balancing front end, a redundant LAN routing protocol, or by the backup taking over the real IP address of the failed machine.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

5.2 Security

The requirement of any security framework is to enable the identification of participants, and to ensure the integrity and confidentiality of any conversations. SIP was not originally designed to be secure, as it was developed to operate within reasonably trusted environments. This makes the protocol more efficient when used within a trusted world, but, as a result, it is vulnerable to attacks from

- external devices
- devices in the signaling path (man-in-the-middle attacks)
- endpoints.

Example of attacks include

- espionage, including eavesdropping and monitoring to obtain private information
- fraud, to gain unauthorized access to resources or to avoid payment
- denial of service (DoS) attacks
- use of incorrectly formed messages to exploit flaws in specific devices.

These security issues, which are described in detail in RFC 3261, are being addressed by extensions to the protocol, including the following.

- The sips: prefix, defined in RFC 3261, which is analogous to https: and mandates the use of a secure transport protocol, such as TLS, between trusted entities. This limits the ability for external devices to launch successful attacks.

- S/MIME (RFC 1847) support for end-to-end message authentication and validation, and encryption of message bodies. These protect from man-in-the-middle attacks, as they prevent intermediaries from accessing or modifying messages.

- Enhancements for Authenticated Identity Management in SIP <draft-ietf-sip-identity-01>, which proposes a mechanism for validating that the author of a message is reachable using the return address given.

- SIP Authenticated Body (AIB) Format <draft-ietf-sip-authid-body-02>, which provides a portable message signature to verify the author of a message.

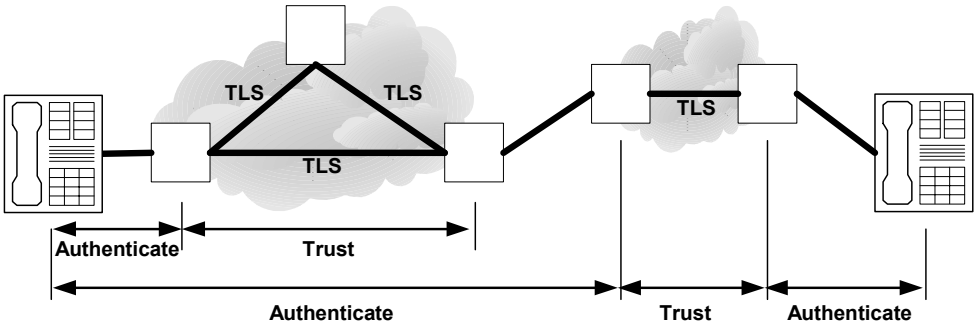
However until these extensions are widely deployed, SIP networks will remain vulnerable.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

These mechanisms provide the ability to authenticate the participants and secure the SIP communications, but it is unlikely that the entire network will use a single point of authentication. As a result, the security architecture is likely to include

- a shared-secret based authentication to identify endpoints to a local server using a username and password
- established trust relationships between servers with key-based authentication and secure transport
- identity authentication provided by the local server, on behalf of the endpoint, for other endpoints that need to validate the endpoint's identity.

This model can be extended with separate secured network segments, with trusted relationships internally and authentication at the borders.



In the above example, authentication at the border of each domain is made directly with the calling UA. However, an individual user may not want to negotiate separate agreements with every network provider, so agreements will often be made between providers to allow seamless transition over their combined network.

SIP provides an extensible authentication architecture that enables it to use a variety of authentication algorithms. SIP extensions for each algorithm define how SIP carries the particular fields required by that algorithm. The draft <sip-ietf-sipping-aaa-req-03> describes the Authentication, Authorization and Accounting requirements for SIP in more detail. In many systems, the authentication itself may be delegated to a separate authentication server that holds the authentication policies and keys. This can use a protocol such as RADIUS.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

5.3 Quality of Service (QoS) and Resource Reservation

When making a telephone call, it is expected (and regulated) that

- the delay before it is possible to speak after the call is connected will be short
- the sound will be reasonable (low jitter and packet loss)
- the delay across the network (latency) will be acceptable
- the call will not be charged for unless it succeeds.

This requires mechanisms to

- guarantee media availability when a call connects and before billing
- control the bandwidth and latency of the media.

The base SIP standard contains no mechanisms for controlling network bandwidth and latency availability, and most current IP networks do not provide this either. However, with the rise of MPLS-based networks, and the use of SIP to control media flows over ATM and other QoS networks, guaranteed quality can be provided.

The use of SIP over non-IP media networks is supported through extensions to SDP to set up the non-IP media channels. For example, RFC 3108, Conventions for the use of the SDP for ATM Bearer Connections, defines how to use SDP to negotiate ATM channels. QoS is provided by the underlying network and negotiated end-to-end using these parameters.

On an IP network, there are two main ways in which a service provider can provide guaranteed QoS across its network. These can be characterized as follows.

- Integrated Services (IntServ) networks use a protocol like RSVP (RFC 2210) to set up a separate bandwidth reservation across the network for each requested media stream. This process reserves resources on every link and at every node that the media path traverses. The problem with IntServ is that it does not scale, because every media reservation requires explicit bandwidth allocations at multiple devices. This generates a huge volume of traffic, especially for VoIP, where a large number of calls are either very short or never get answered. IntServ is therefore not suitable for large VoIP installations.

- Differentiated Services (DiffServ) networks classify all traffic into a series of predefined classes, and then prioritize this traffic throughout the network based on its class. This requires DiffServ-capable routers throughout the network to understand the prioritization and to modify their behavior accordingly, but does not require separate reservations for each media stream, so this mechanism does scale.

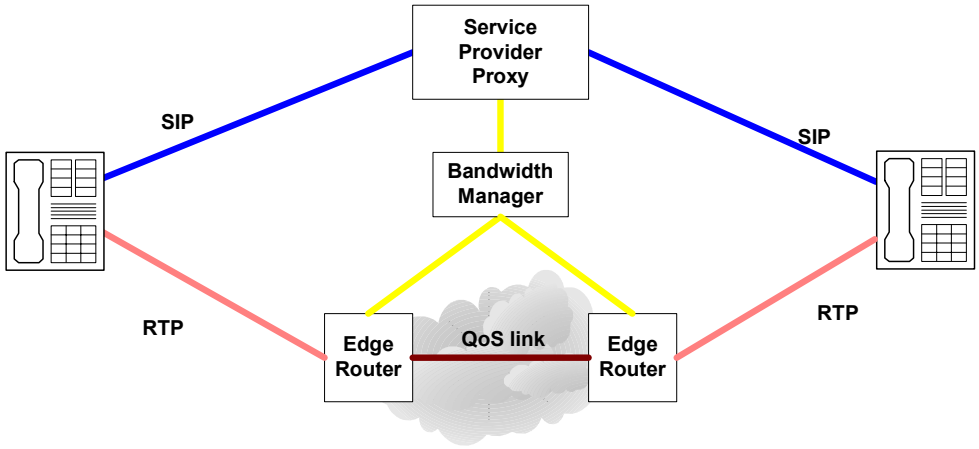
DiffServ networks also require routers at the boundaries of the network to assign priority to packets received from the outside, and to monitor the traffic to ensure that the network is not overloaded.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

Whichever mechanism is used, the service provider needs to control access to the network so as to ensure that adequate resources are available to meet the agreed QoS levels and to prevent degradation of the network by unauthorized traffic. This control will normally be provided by a device at the edge of the network, an Edge Router.

To use SIP across such a QoS network requires a SIP proxy in the signaling path to understand any media requests and open the necessary pinholes in the Edge Router firewalls. This works as follows.

- When the SIP request reaches the proxy, the UA and proxy negotiate the parameters required for the media path. The proxy instructs a Bandwidth Manager to set up the media channel.
- The Bandwidth Manager is responsible for authorizing media channel requests made by through the Service Provider's SIP Proxy. It monitors the loading on the network and controls the Edge Routers' policy to ensure that QoS is maintained within the network. It will open and close pinholes in the Edge Routers to let specific media channels through the network in response to requests from the proxy.
- When the Edge Router receives the media, the necessary pinholes have already been opened, so the media can pass through the network with a known QoS.



It would be possible to use a SIP B2BUA at the boundary of the provider network, and to hide the reservation process from the UA. However this would limit the new services that the UA could develop, because the B2BUAs in the network would have to understand any extensions in order to be able to allocation the right resources. Involving the UA in the reservation minimizes the intelligence that must be implemented in the network core.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

RFC 3312, Integration of Resource Management and SIP, defines an extension to SIP that enables media reservation before the phone rings. This ensures that, when the phone is picked up, the media channel is already in place. RFC 3313, Private SIP Extensions for Media Authorization, defines how this can be used to negotiate and reserve the quality of the media channel, and to refuse the call if a suitable channel is unavailable. This feature is not yet widely available, but is increasingly being mandated for equipment in the core of the network.

Currently, QoS is not normally provided out into a customer's LAN, but when voice and data start sharing the same LAN, QoS becomes important. This is because voice requires a fairly low bandwidth to be available on demand with consistent latency to provide good sound quality, but it is fairly tolerant to transmission errors. On the other hand, data can use high bandwidth and can handle high and variable latency, but is intolerant to errors. If these two very different types of traffic are mixed on the same network without differentiation, the overall performance will degrade rapidly as the loading increases. When QoS does become available in LANs, support for RFC 3312 and RFC 3313 will also be required in SIP phones to provide end-to-end QoS.

5.4 Scalability

The existing PSTN network supports billions of telephone subscribers; this is a huge number of addresses to track and for which to maintain routing information. The network also has to handle large numbers of calls, particularly at peak times, with consistent reliability. This presents two separate scalability issues: the first is the ability to route quickly to the required destination during call set-up, and the second is the ability for devices in the core of the network to handle the traffic associated with all the active calls.

For call setup, SIP uses the proven, scalable DNS framework as described above. DNS can handle the required number of addresses and is able to control local caching, which allows consistent information to be distributed throughout the network and minimizes the load on the master database. SIP proxies spread through the network can then provide distributed SIP routing and authentication. Once a call has been established, SIP provides direct communication between the devices over the IP backbone, without any centralized point of control that might become a bottleneck.

Within an individual server, the SIP protocol also scales well, because it includes identification fields for rapid matching of messages to dialogs and transactions, and suitable implementations can load balance across clusters of machines using DNS.

However, because of security, audit and network incompatibilities, both the signaling (SIP) and the media (RTP or another protocol) are often routed through intermediate devices that do more processing than just IP forwarding. One example is recording for billing or an audit trail, where a company, service provider, or government diverts all traffic through a specific device to record the required information. In such situations, these intermediate devices are in the path of all the communications and may become bottlenecks for the machines that they serve.

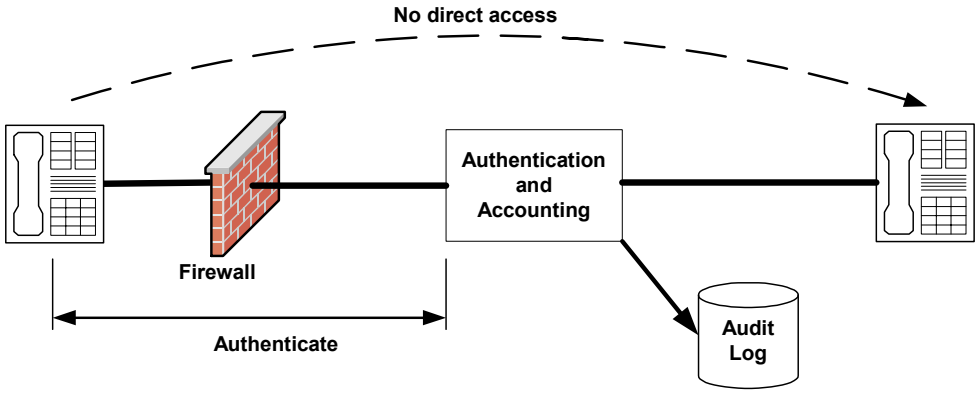
Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

QoS also imposes a heavy load, as it requires the monitoring of bandwidth usage and availability through the network. IntServ does not scale well, because it requires a separate bandwidth reservation across the network for every call. DiffServ or MPLS-TE based solutions scale better, because the bandwidth allocation is performed at a higher, aggregate level, but these require a second level of control and monitoring to ensure that the allocated resources are not themselves overloaded.

5.5 Accounting

In order to charge for something, it must be possible to monitor and control access to it. This requires the ability to identify users (authentication), check that they are eligible to use the resource and then permit use (authorization), and record the usage (accounting). Authentication and checking of eligibility using SIP is covered in section 5.2, Security, above, and control of access and monitoring are covered in this section.

For services that are accessed through the signaling, for example status requests using SIP Events, proxies on the signaling path can control and monitor usage, although it must not be possible to bypass the accounting proxies to access the resource directly. This can be achieved using TLS or a firewall to limit direct access to the resource.



For media-based services, like telephony, the network must be able to restrict access to the media; this is only possible in a network that limits direct communication between endpoints. Billing records can then be linked to the resource reservations, as described in Quality of Service (QoS) and Resource Reservation above.

A non-QoS network can use firewalls to control media access in a similar way. These can be managed using the same techniques, with the firewall opening media pinholes and tracking usage. However, an extremely strict firewall policy is needed to prevent a customer bypassing the firewall, and such control limits the general usability of the network, although this solution may be suitable for dedicated telephone networks.

More generally, SIP signaling is not designed for the time-based billing used in traditional telephone networks. The separation between signaling and media that SIP provides makes it difficult to time calls accurately, as is required by telephony regulations. This is an area of ongoing study and debate.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

A prepaid service, where the network retains the ability to disconnect a call after it has started, imposes further constraints. In the earlier examples, the intermediate gateways authenticate the user and participate in the media negotiation, but otherwise stay in the signaling path only to handle media changes and to clean up at the end of the call. In SIP terms, the gateways are proxies, because they cannot initiate requests. However, the prepayment application server retains control of the call signaling, and is therefore a B2BUA rather than a proxy. This distinction is important when deciding how to design a SIP server to provide a chosen set of services.

5.6 Privacy

Privacy is the control of information, including

- who receives what information
- the level of detail that is provided
- what the recipient is allowed to do with any information received.

This is a complex area to define and even more difficult to enforce. For this reason, government regulations exist to control the behavior of some recipients of private information.

Using SIP, private information may be distributed through the following two mechanisms.

- Implicit distribution. Some information is required for the protocol to work. This includes headers to tell the recipient who has sent the message and how to reply, as well as lower level information, such as the IP address to which the media must be sent. SIP UAs can avoid much of this information by obscuring the return addresses and many other identifiable fields, but they are unable to remove all indications of the message source. In order to provide a fully anonymous service, a separate anonymizing server (implemented as a B2BUA) is required in the signaling and media paths to hide all identifiable fields.

- Explicit distribution. The UA may choose to provide information to trusted third parties, however it may want this to be hidden from others. For example, a network may require user identification for authentication purposes, which should not be passed to the destination. In these cases, the recipient of the information must remove it from any messages that are passed on and must restrict its own use of the information.

  For presence information, this situation is even more sensitive, as a much richer set of private information is being made available to third parties. This requires the ability to specify which groups of users can access each part of its state information.

RFC 3323 describes the requirements for maintaining privacy in more detail, and how privacy servers within the core of the network can provide this. Work on maintaining privacy of presence information is ongoing.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

5.7 NAT and Firewall traversal

NATs (Network Address Translators) exist to overcome the limit on the number of available IPv4 addresses, and to provide privacy and security for devices within a private LAN. All NATs set up bindings between external IP address/port combinations and internal IP addresses and ports, to allow packets to be routed back from the external network to devices within the LAN that do not have a globally routable IP address. These bindings may be statically configured to allow access to services within the LAN for external users, for example a website, or dynamically configured to allow packets to be routed back to an internal machine for a particular communication session.

Firewalls implement an organization's security policy and may be configured to allow or disallow particular protocols, including SIP. They work by restricting the flow of packets through them based on configurable criteria, which may include the packet's source or destination address or port, or the protocol being used. It is the responsibility of the organization to configure its firewall to allow or disallow SIP traffic according to its own policies.

NATs and firewalls are often co-resident, because the management of NAT bindings is readily integrated with additional security. However, they are logically separate in their function, and it is only the NAT function that presents a technical challenge for SIP to overcome; it is not the intent of SIP to bypass firewall policy, though SIP should be firewall-friendly.

5.7.1 Types of NAT

There are different types of NAT, distinguished by the characteristics of their bindings. The following lists the major types of NAT.

- Basic NATs do not change the port number. The bindings link an internal IP address to an external IP address for selected ports, but the port numbers are unchanged across the NAT.
- Full-cone NATs set up a single binding between an external IP address and port, and an internal IP address and port. Once this binding is established, any packet that is received from the external network to this address and port will be forwarded to the internal address and port.
- Restricted cone NATs (and Port restricted cone NATs) operate as above, but only accept packets that are received from the same IP address (IP address and port) as the destination of the outgoing packet that established the mapping.
- In each of the above cases, a particular internal IP address and port always maps to the same external IP address and port. However, Symmetric NATs set up a different binding each time, so the same internal IP address and port may appear as different IP addresses and ports to different destinations, and several devices can share the same external address and port when communicating to different remote hosts.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

These NAT characteristics result in the following effects.

- The party inside the NAT must initiate communication to each remote address and port to create the new dynamic binding, or a separate protocol must be used to create new bindings. If no external mechanism is used to create the bindings, then a device behind a NAT may be able to make SIP calls but not be able to receive them. Even in this situation, symmetric RTP must be used to allow media to flow in both directions through a single RTP connection initiated from inside the NAT.

- To maintain a dynamic binding, packets must be sent between the parties at regular intervals (the required frequency of these retransmissions is not defined and can be under a minute), or the communication must use a session-based transport, such as TCP. For this reason, the use of a session-based transport protocol is strongly recommended. If UDP is used, then the device behind the NAT must continually resend registration or other messages to maintain the bindings, which is a waste of resources.

- Two ports on the same internal address may be mapped to different external IP addresses, and the external ports may bear no relation to the internal ports - as a result, the value of addresses and ports cannot be inferred from the other addresses or ports. This breaks some of the existing standards that assume a numerical relationship between port numbers. Several extensions have been developed to address this issue, including RFC 3581 for symmetric response routing in SIP, and draft-ietf-mmusic-sdp4nat-05, which extends SDP to specify additional port numbers for RTP.

- An internal device has to use a separate protocol to determine the address at which it will appear to external devices. In SIP, this requirement is minimized because the recipient of a message sets the return address to be the address from which the message is received, rather than address that the sender believes is correct. However, additional protocols are required to determine valid addresses for the media.

These issues are common to all VoIP protocols, not only SIP, so the IETF has established the MIDCOM working group to discuss general solutions to NAT traversal by VoIP. Their solutions fall into the following categories.

- NAT detection protocols that allow a device inside the NAT to determine the NAT's behavior and bindings indirectly, and to modify the protocol messages appropriately. STUN, as defined in RFC 3489, describes such a protocol.

- NAT control protocols that allow a device inside the NAT to control the NAT to set up dynamic NAT bindings and to determine the external address that will be presented. uPnP provides one mechanism, which is supported by Microsoft and is being discussed by the uPnP forum, rather than the IETF.

- Application Level Gateways (ALGs), which modify the signaling messages and may provide a media relay. ALGs can work around limitations in the protocol and provide a short-term solution. These are discussed in more detail later.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

- Relays in the external network with globally routable addresses to relay the messages. TURN provides this functionality.

NATs are not required in IPv6 networks, so it is hoped that they will eventually disappear, but they will exist for many years, and SIP must work through them.

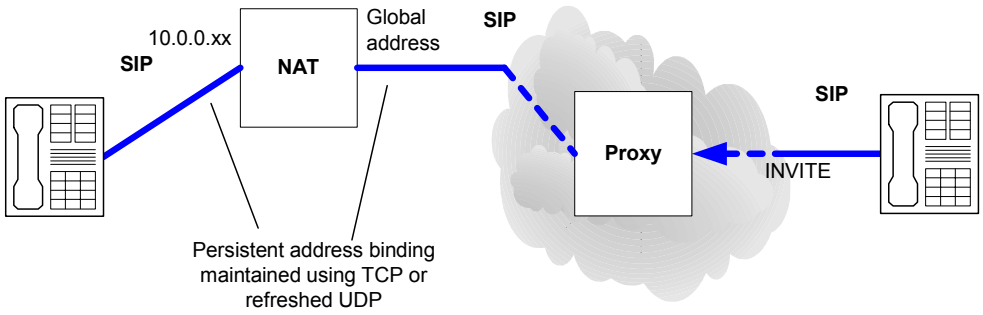
This functionality is likely to change as the standards for NAT and firewall control improve, and the best solution will be a combination of the above, dependent on the precise scenario.

5.7.2 Using SIP through NATs

As mentioned earlier, SIP contains several features to help its operation through NATs, but the following issues still remain.

- How do you send a SIP message to a device that is behind a NAT?
- How do you establish a media session with it?

Once a device has received a SIP message from another device that is behind a NAT, it can respond to the address and port from which the message was received, and these addresses remain valid as long as the NAT binding is kept alive. However, if the first SIP message is to the device behind the NAT, another mechanism is required. This first SIP message can be sent through the proxy with which the device registered its location, as long as the device maintains its NAT binding with the proxy. As discussed earlier, this can be achieved by using a TCP connection or by refreshing its registration at regular intervals. By Record-Routing all requests, the proxy can also ensure that it remains in the path of all future requests, and that external devices do not try to contact the device behind the NAT directly. As result, this mechanism works for even the most restrictive NATs.



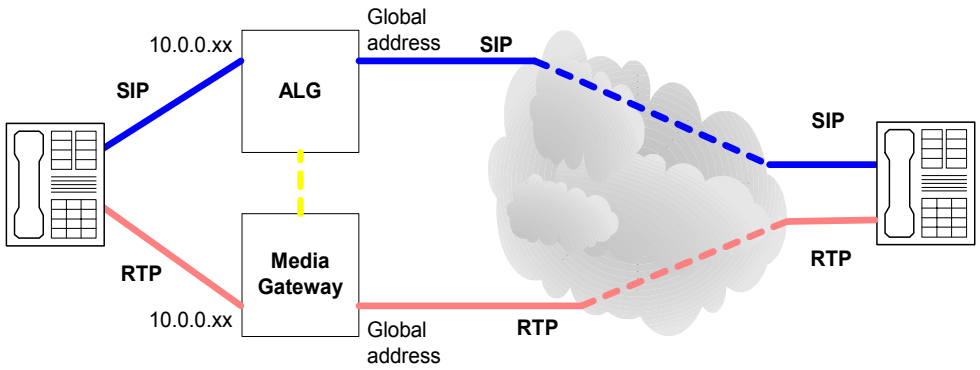
When only one device is behind a NAT, the device behind the NAT can successfully start the media session and, by using symmetric RTP, this session can be used to send media in both directions. However, when both devices are behind NATs, the situation is more difficult because neither has a valid address with which to establish the media session. If another protocol is not available to determine a globally routable address to which to direct the media, then a media relay may also be required.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

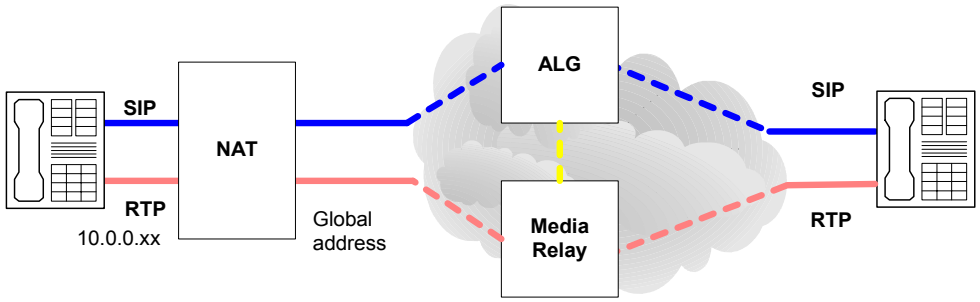
5.7.3 Application Level Gateways (ALGs)

ALGs are devices that understand higher-level protocols, and may dynamically open additional pinholes through the firewall to let data through according to each protocol's requirements. For example, a SIP ALG may open pinholes in the firewall to allow the media to flow.

ALGs can also be integrated with NATs and be used to modify the messages as they pass through to convert any internal IP addresses to their external equivalents. This provides a method for NAT traversal that does not require any changes to the endpoints.



An ALG can be positioned outside the NAT to enable SIP communication with devices behind a NAT. In this situation, it incorporates a media relay and modifies the SIP messages to direct all media through its relay. Because the ALG presents globally routable addresses, it can successfully set up connections with endpoints that are behind a NAT, and can therefore be used as an intermediary in calls between endpoints even if both are behind NATs.



Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

Because of their ability to work through NATs with the current generation of SIP, ALGs are a fundamental part of today's SIP offerings and form the basis of specialized products such as Session Border Controllers. However, SIP ALGs are implemented as B2BUAs, not proxies, because they modify the SIP signaling messages beyond that allowed by a proxy. As a result, there are various problems with ALGs, including the following.

- The SIP messages cannot be encrypted end-to-end, because the ALG needs to be able to interpret it. This limits security and privacy and makes the ALG a trusted party in all communication.
- The protocol cannot be extended without upgrading the ALG. Again, the ALG needs to understand the protocol to control the firewall or media relay appropriately.

For these reasons, ALGs are not able to support new protocol extensions and service innovation by end users, and cannot be recommended as a long-term solution.

5.7.4 Devices behind the same NAT

If both of the endpoints are behind the same NAT, it is more efficient for them to use the internal IP addresses instead of globally routable addresses, because the messages can then remain within the LAN. For both SIP and SDP signaling, this can be achieved by using a fully qualified domain name rather than an IP address to advertise the server ports, and by providing a local DNS server that returns the internal address rather than the globally routable IP address. However, if a globally routable DNS address for the endpoint does not exist, this solution is not possible. Also, not all endpoints may support domain names within SDP, which limits the applicability of this solution in some environments.

A fuller explanation of the scenarios and a mechanism that handles many of these scenarios is presented in <draft-rosenberg-sipping-ice-01>.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

5.8 Device configuration

SIP devices do not require a lot of configuration information, but the way that this information is entered varies significantly between devices. This makes support of SIP devices more complex than it should be.

The following configuration information is normally required.

- Local (outbound) proxy, to handle local policy and NAT/firewall traversal
- Registrar (one or more)
- Username and password (one or more)

Rather than agreeing a single standard mechanism for automated configuration under centralized control, several alternative mechanisms are being recommended. These include the use of RFC 3361 – DHCP Option for SIP and well-known DNS and multicast addresses. To coordinate these separate mechanisms, draft-ietf-sipping-config-framework-00 defines a single configuration process that tries each in turn until one succeeds.

In some environments, it is unclear who should control the endpoint configuration. For example, users may need different outbound proxies depending on the service required and may not want their network service provider to control this choice, whereas the network service provider may have a financial incentive to route all of the SIP traffic through its servers. In others environments, for example enterprises, centralized management can be used to minimize end-user support and enforce corporate policy. Different solutions may therefore suit each situation.

Endpoint configuration is also possible using SNMP, or another MIB-based management protocol. The standard MIBs for the configuration and monitoring of SIP devices are well advanced <draft-ietf-sip-mib-07>, although arguments remain over the level of detail that should be available through the MIB. MIBs are particularly suitable for the management of larger SIP devices, such as servers, where they provide a high level of configuration detail and status information and can be easily integrated into a larger system management suite.

5.9 IPv6

The introduction of IPv6 is being driven by the lack of IPv4 addresses, particularly in the Far East, by the standardization on IPv6 for 3G mobile, and by government initiatives in countries including the UK and US.

SIP and SDP are fully compatible with IPv6, so are ideally suited to this environment.

The only IPv6 specific standard for SIP is an updated DHCP option to configure the SIP outbound proxy <RFC 3319>. This is required because DHCP has changed slightly between IPv4 and IPv6.

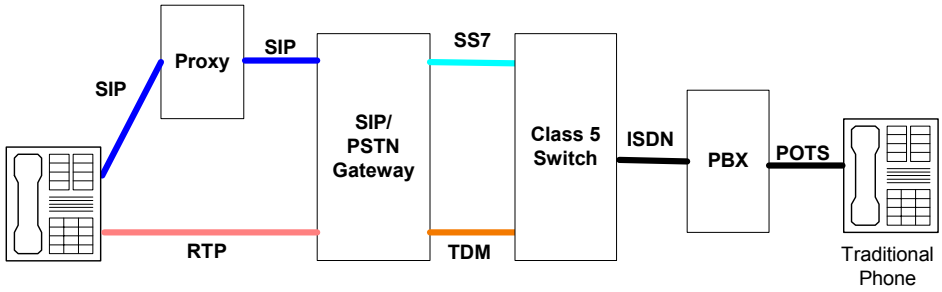
Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

6 SIP and the PSTN

Telephony is the most developed SIP application, and the PSTN adds a range of specific requirements. These requirements fall into the categories of interoperability and regulatory, and the following sections describe the issues to be addressed in each area.

6.1 Interoperability

Full PSTN interoperability implies that a SIP phone, operating through a SIP to PSTN gateway, is a fully functional replacement to a traditional phone. In other words, a subscriber can access all existing services with a SIP phone even when some of those services are provided by a third party, for example corporate voicemail. This level of interoperability does not prevent a SIP phone from providing new services that cannot be provided on a traditional phone.



SIP was designed for Internet telephony, and not designed to replicate the PSTN, and this means that it cannot readily handle all PSTN features. The following is a list of the some of the more important areas of work to use SIP in the PSTN.

6.1.1 Overlap signaling

Overlap signaling is required when it is not possible to determine whether a particular sequence of digits represents a valid phone number without attempting to place the call. This situation exists in various networks, including several European countries. In these networks, it is not possible to wait until the entire number has been entered before dialing, because the only way to detect this would be a pause in the entered digits. To allow a user to dial slowly, for example when referring to a telephone directory, a large delay between individual digits must be allowed (>10 seconds). If the exchange were to wait this length of time after the last digit to determine that the number was complete before placing the call, the delay in call-setup would be unacceptable. Therefore, when the user starts dialing, the telephone exchange waits for a minimum number of digits and a pause of a few seconds before using the digits collected to route the call onto the next hop. If the next hop has sufficient information, it continues to route the call onto its destination; otherwise it waits for further digits from the user before continuing.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

This mechanism does not map easily onto SIP, because one subset of a number may not be routed the same way as another. As a result, when additional digits are received, a completely new SIP call must be made incorporating this new information, to enable the call to be routed independently. For all but one of the calls, the number will not represent a valid destination, and the call will fail with an "Address Incomplete" type of response, so only a single call remains.

It should be noted that a native SIP phone should not generate overlap dialing, because the user can be forced to enter the complete number before attempting to dial, as with mobile phones. However, when interoperating with traditional phones through SIP adaptor, or through a SIP gateway to the PSTN, overlap dialing cannot be avoided.

If an overlap-dialed call has to be routed from the SIP network into the PSTN, then all the calls placed as a result of overlap signaling must reach the same gateway and be correlated together. Otherwise, the gateway will not be able to generate overlap signaling in the PSTN, and will instead place multiple independent calls, which uses more resources.

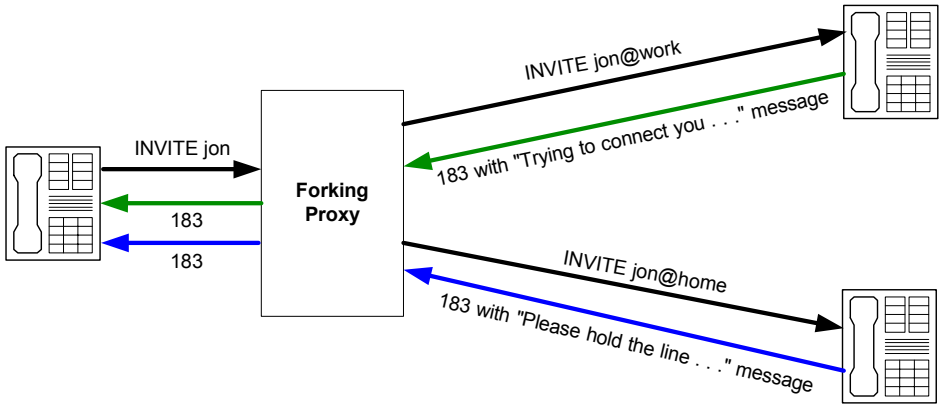
RFC 3578 describes this mechanism in more detail, although the standard is not yet widely implemented.

6.1.2 Early media

Early media describes media sessions that are started before the call setup completes. This is used in the PSTN for announcements during connection, such as "Trying to connect you …", and to minimize the delay before the establishment of the media session once the call has connected.

Early media sessions require a mechanism to negotiate media channels before the call setup completes. This requires an ongoing exchange of messages between the caller and called party during call setup to agree these channels and any changes. The SIP protocol extension for reliable provisional responses (RFC 3262) provides such a mechanism, is the basis of early media in SIP, and is well supported.

Although early media works successfully, it does not work well with forking because a single forked call may establish multiple media streams. This is shown in the following diagram.



Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

Here the call is forked and causes two phones to ring simultaneously. When both send back early media, what does the caller hear? Should one of the streams take priority over the other? The handling of this situation is under the control of the client application, but this complicates its design and there is no simple solution. For example, if the client application chooses one of the streams and then the other completes first, the caller may hear a very confused media stream.

This remains an area of ongoing concern, although it is not currently presenting a serious problem because forking is not widely used.

6.1.3 Application Control with a traditional phone keypad

The telephone keypad is often used to control telephony applications. These applications include

- information services, such as share prices and timetables
- calling-card services
- voicemail and unified messaging services.

Traditionally, key presses are encoded as DTMF and transmitted over the line with the voice. Using SIP, there are two methods to transport key presses: one is in the signaling channel, and the other is in the media channel. Both methods are needed to handle all of the above applications.

Calling-card applications need to monitor all key presses to control call setup, but once the call is put through, they are only interested in specific key sequences to regain control of the call to allow placement of a follow-on call; other key presses must be sent to the true destination of the call to control any application there. Monitoring the entire media stream by the calling-card application to detect these key sequences would be inefficient and would tie up media resources, so this type of application needs a mechanism to receive information on the key presses through the signaling channel.

Voicemail applications, on the other hand, may allow the user to record messages and announcements, controlled using key presses. In this situation, it is important that the keystrokes and media are synchronized in time, so that the recording starts and ends at the right time. This requires that the keystrokes be sent over the media channel, because this correlation cannot be provided over the signaling channel.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

After much discussion and use of non-standard mechanisms, the following solutions have been proposed to handle each requirement.

- For the transport of key presses in the media stream, RFC 2833 provides suitable functionality, and this standard is now widely supported in SIP phones and application servers. It encodes the key presses into packets in the RTP media stream.

- There is still no final agreement on how to carry key presses in the signaling channel, but current proposals allow a device to ask the UA to send it each keystroke in a new SIP message. Further proposals include the ability to download a digit map to the client to allow it to monitor particular key sequences. The advantage is that this can decrease the number of messages required, but it also increases the complexity of the UA, especially if multiple devices want to monitor simultaneously.

Several existing SIP implementations use the INFO message to carry all key presses to devices in the signaling path. This method is inefficient because it sends all of the keystrokes through the signaling path, even when not required. It also raises scalability concerns, because there is no flow control mechanism to control the large number of messages that may be generated. As a result, the use of INFO messages is strongly discouraged.

There are also concerns over how multiple servers that are monitoring a single call should interact. For example, it is possible for several of them to place meaning onto the same key sequence; this is known as feature collision. One proposal to solve this uses Service Brokers, which act as a central point for other feature servers to interact with the call and resolve any conflicts.

6.2 Regulatory requirements

Telephony is heavily regulated because of its importance to the economy; it is fundamental to most businesses, provides access to the emergency services, and is monitored by the security services. To provide a full PSTN replacement service, the SIP network has to meet the same regulatory standards for features, quality and reliability.

Regulation of Internet telephony is already happening in many countries, although it is unclear how successful this process can be, given the ability to make phone calls over the Internet without any central point of control. The incumbent Telcos are also working to increase regulation of the Internet telephony service providers, in order to limit its growth and to raise the barriers to entry into the industry.

Compliance with these regulations also brings benefits in the form of government subsidies in many countries. There is therefore an incentive for Internet telephony service providers to comply with government telecommunications regulations whenever possible.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

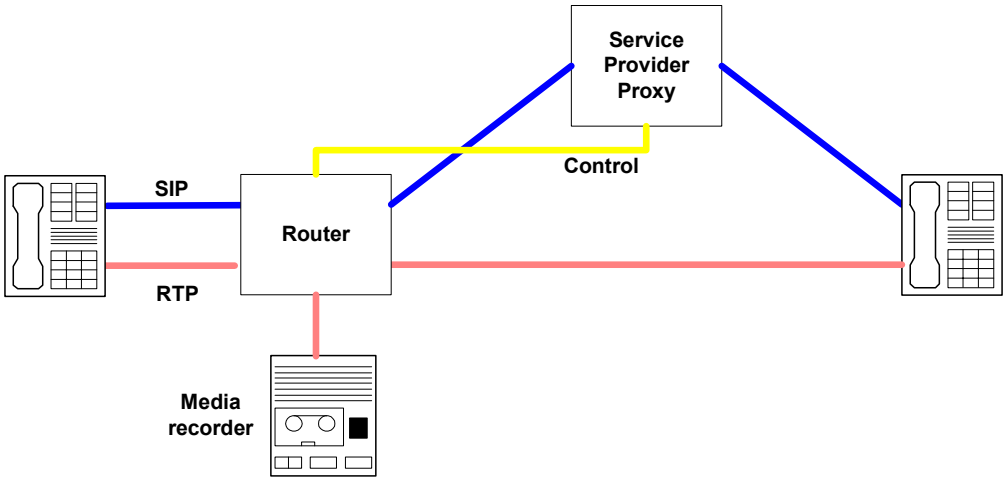
For a SIP-based telephone network to satisfy all its regulations, it will need to look a lot like the PSTN, with redundancy, media reservation, local feature servers and wire tapping capabilities. Reliability, scalability and QoS are more general requirements, and these were covered in Chapter 5. The following sections describe the other PSTN-specific requirements in more detail.

6.2.1 Wire-tapping

In most countries, the government is able to monitor selected telephone calls to or from individuals, without the knowledge of that individual. For traditional telephone companies, this is provided through the local exchange, which handles both the signaling and the media for every call.

In the IP world of SIP, there may not be a telephone service provider, and there is no longer a simple central point at which to monitor the calls. However, assuming that a SIP telephone provider is being used, then their proxy may be used to monitor the signaling and to record information including the source, destination and duration of any calls.

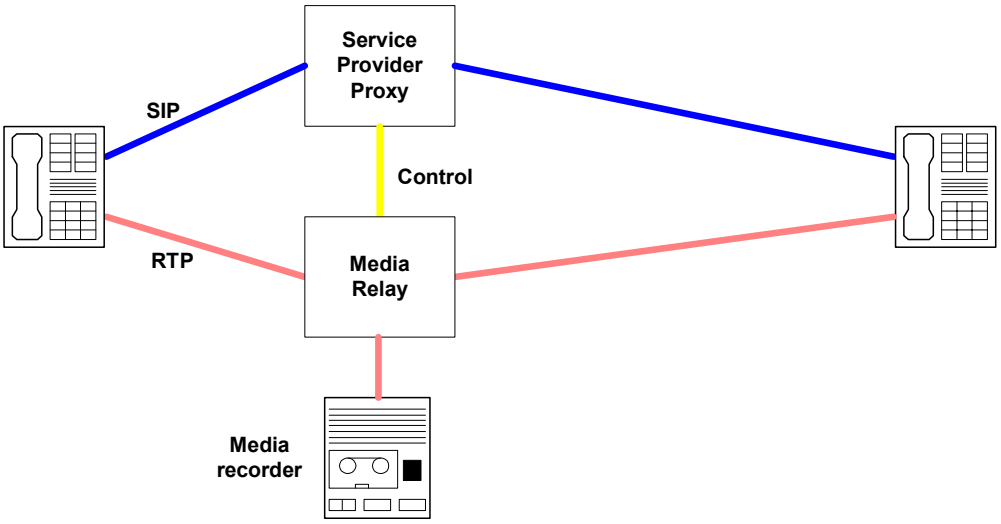
Monitoring of the media is more complicated, as media is normally sent directly between user agents over a separate route. The only way to monitor such traffic is by packet sniffing at the router at the boundary of the customer's connection to the Internet, as shown below. This is a very processor-intensive process, and is further complicated if the customer has multiple links.



An alternative solution to direct only the monitored traffic through a media relay, where the call could be recorded, is also not possible, because one of the requirements is that it is not possible to detect that you are being monitored. The very act of redirecting the media identifies that the call may be being monitored.

It may also be possible to direct all traffic through a media relay, and some equipment manufacturers are using this solution, but this puts a heavy load onto equipment in the core of the network. In addition, there is no way to enforce this over the Internet without installing very restrictive firewalls to prevent direct media communication.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.



This issue remains unresolved, although the CALEA requirements in the US and similar proposals in other countries are addressing this issue. Current indications are that the regulations will impose the ability to monitor all traffic at the network edge, including telephony. However there is significant lobbying to limit the resulting intrusion of privacy and its enormous implementation costs.

6.2.2 Emergency calls

Due to their importance, calls to the emergency services are regulated separately from other calls. These regulations include the following.

Location determination

- The call should be handled by the local emergency service, so that the local police or ambulance service is always called. This requires knowledge of the location of the user, which is not available through SIP, because the IP address cannot always identify the location.
- Caller identification is required to allow the emergency service to know the location of the call, to allow them to dispatch help to the correct location, even when the caller cannot convey their location. This private information must also be withheld from other users.

Both of these can be provided by local configuration and the inclusion of caller location fields in emergency call requests. However, the use of local configuration risks the information being out of date. The IETF GEOPRIV working group is discussing the management of location information using DHCP, which would enable a device to determine its location from a central server. This potentially provides a long-term solution to this problem, but also places a requirement on service providers to manage this additional information.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

Special handling of emergency calls

- Emergency calls should be given higher priority by the network. draft-ietf-sip-resource-priority-01 defines additional SIP headers that categorize the priority of a request. These new headers do not affect the operation of any IP routers in the network, but may be used by the SIP-enabled devices to prioritize their processing of the messages and to allocate higher priority to the IP packets to enable faster routing through the network.

- Calls to the emergency services are allowed even if the user is not an authenticated user of the network, for example with roaming mobile phones. There is no standardized method to allow this, and in particular it is not clear how the phone would know where to call without being authenticated and receiving local configuration information.

Work on all these issues is ongoing in the standards bodies, together with close liaison with the regulators to ensure that any solution is acceptable to them.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

7 Enhanced applications for SIP

This chapter discusses some of the areas for which SIP is being developed, which will enhance the range of facilities that are currently available.

7.1 Mobile (3G)

SIP was mandated for call signaling for revision 5 of the 3GPP proposals for mobile networks. In revision 6, SIP's use is being extended to include presence. Revision 6 is scheduled to be frozen in March 2004.

The mobile environment presents a very different environment from a traditional SIP network, and this has required several extensions to the protocol. Its main characteristics and their effects include the following.

- Bandwidth is expensive in any radio-based environment. SIP is a text-based protocol that was designed for high-bandwidth environments, and can be compressed to significantly reduce the bandwidth required. SigComp <RFC 3320> provides a generic compression framework that is suitable for SIP. It is optimized for a particular protocol through the use of a standard dictionary of commonly used terms within that protocol. The standard dictionary for SIP and SDP is defined in RFC 3485.

- IPv6 has been mandated by 3GPP for use throughout the network. The use of SIP with IPv6 was covered earlier, and presents no problems.

- Mobile users move between radio cells, but as they move they maintain the same IP address. As a result, this movement is invisible to SIP, and the signaling is unaffected.

- Extended registration is required to allow mobile users to roam (use their phones with foreign network providers), while maintaining their relationship with their home provider so that they receive a single invoice, and to access their personal settings such as voicemail. This is achieved by routing communications through a local proxy (to impose local rules and access to local resources) and through a home proxy (to provide consistent global services and access private settings).

  Some minor SIP extensions have been defined that force messages to travel through several proxies, and to obtain the necessary configuration information from the different domains. These extensions include RFC 3327 (SIP Extension Header Field for Registering Non-Adjacent Contacts) and <draft-ietf-sip-scvrtdisco-04> (SIP Extension Header Field for Service Route Discovery During Registration).

With these extensions, SIP provides a flexible signaling framework for mobile telephony onto which new services, including presence and messaging, can be built.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

7.2 Caller preferences

SIP has the ability to set up different types of communications session, including voice, video and instant messaging, and an individual user may have several SIP devices: for example at home, at the office and for mobile use.

When a call is received, the called party may, using pre-defined rules in a proxy or through an interactive choice, direct the call to any specific device. This choice may, for example, depend on the time of day, the identity of the caller, or the type of media requested. However, the caller may also have a preference over the device that is used to answer the call. For example, the caller may only want to talk if the called party is available at work, and does not want to be put through to voicemail.

Caller preferences allow the caller to request that the call only completes if certain conditions are met. Proxies and the recipient then use this information to decide how to route the call. The final destination of the call will therefore depend on both the caller's preferences and the called party's policy for handling incoming calls.

The success of this functionality relies on standard definitions for the types of device that are available to answer the call, and the willingness of the user to provide this information to a third party. <draft-ietf-sip-callee-caps-00> defines a way to describe the capabilities of a SIP device, and <draft-ietf-sip-callerprefs-09> defines how a caller can request to connect only to devices meeting selected criteria. These drafts, which are now getting close to standardization, provide the basis for this powerful feature.

7.3 Third party Call control

Third party call control refers to the ability for a device that is not one of the ends of the SIP signaling to affect a SIP dialog. It is required to provide PBX style services, such as call transfer and call screening, when there is no central PBX. There is no way to achieve this within the core SIP protocol, because the protocol is secured from end to end within a dialog, so several SIP extensions have been defined to enable this functionality.

The following example shows a call screening service, which requires third party call control when there is no central PBX.

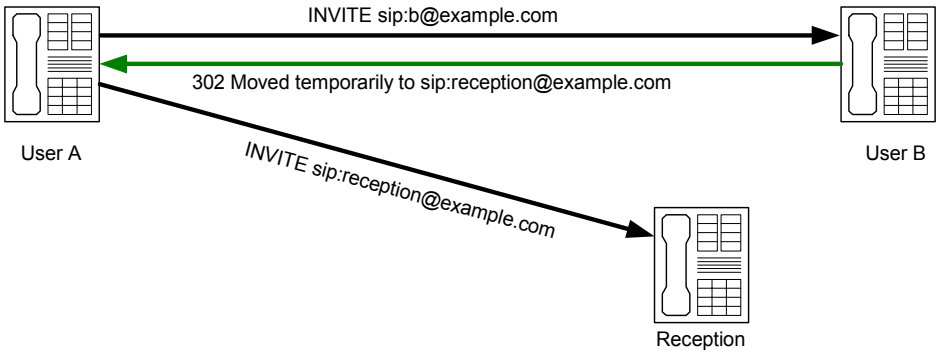
- A calls B, who has his calls forwarded to a receptionist.
- The receptionist checks with B whether he wants to take the call.
- The receptionist puts the caller through to B.

Third party call control is required for the receptionist to put the caller through to B, and to be removed from further involvement in the call.

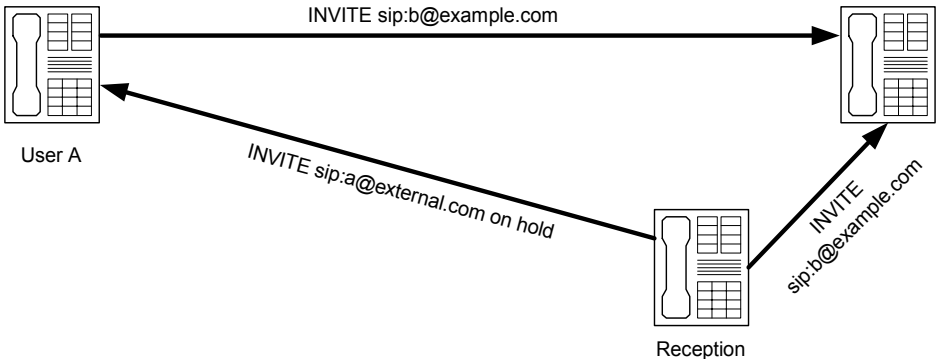
Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

Using SIP, this can work in the following way.

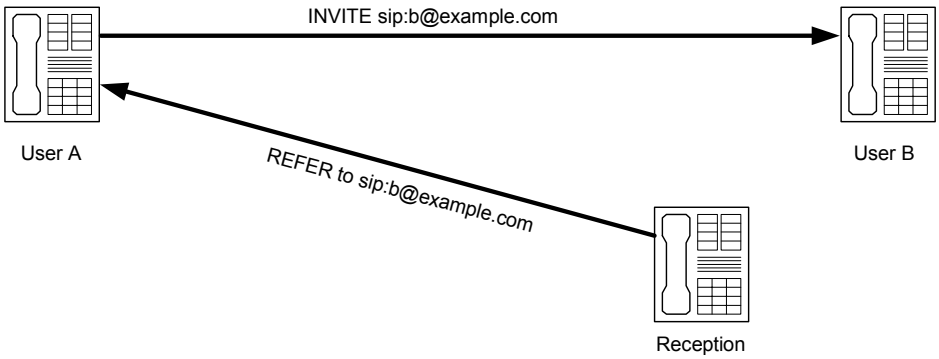
- B redirects all his calls to the receptionist using either his phone or a proxy, so that the initial call is established with the receptionist.



- After answering the call from A, the receptionist puts it on hold and calls B.



- The final stage requires the receptionist to set up a call between A and B to replace the two existing calls and to take the receptionist out of the loop.



Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

The SIP REFER method (RFC 3515) allows a third party to request a SIP device to perform a defined action. In the call-screening example, REFER is used by the receptionist to cause A to call B directly.

There are several issues with this mechanism, in particular in relation to security. For example, if a person transfers their caller to a premium rate number, who pays for this call? Also, how is the second call put straight through, whereas the first call is diverted to the receptionist? Furthermore, it must not be possible for A to reuse any of this information in order to make a call directly to B at a later time and bypass his call forwarding.

REFER incorporates a security mechanism using a token that is passed by the REFERer to the REFERee to enable the REFERee to validate the authority of the REFERer. This provides the security required above, but it also requires an existing trust relationship between the referrer and referee to interpret the token. Although the meaning of the token is dependent on the particular environment, the current lack of standardization will cause interoperability problems between different vendor solutions.

7.4 Conferencing

VoIP conferencing today primarily uses H.323 as the signaling protocol. H.323 is well established in the market and has been extended to include conference control features. The use of SIP in conferencing applications is an area of intense interest and standardized mechanisms are being defined to add conference control.

Conferences fall into the following two categories.

- Tightly-coupled conferences have a central point of control. This is the traditional conferencing system, where a single server controls the conference and the media mixing.
- Loosely-coupled conferences have no central point of control; the users communicate directly with each other, and the control and media mixing may be distributed through the network.

The distinction between these conference types, and the requirements that they have, are discussed in detail in draft-ietf-sipping-cc-framework-01.

The use of SIP for tightly-coupled conferencing is well advanced, because this can be achieved using only standard telephony function, although more advanced features are being planned. Loosely-coupled conferences present a much more difficult problem, because of the difficulty of maintaining consistent state across conference participants as participants enter and leave. The control of loosely-coupled conferences is still an area of academic study, so the rest of this section is devoted to tightly-coupled conferences.

Conferencing imposes a wide range of high and low-level requirements, including the following.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

Session control

- Conversion of a two-party call into a conference with three or more participants
- Conversion of a conference back into a two-party call when the other participants leave
- Invitation to new participant (dial-out)
- Acceptance of a new participant (dial-in)

Conference floor control

- view information on the other conference participants
- control who may join and speak in the conference.

Application-level conference control

- prearrange conferences
- create conferences on demand

With a SIP conference server, the session control requirements are covered by standard telephony and third party call control mechanisms. For example, REFER can be used to redirect a call to the right conference bridge.

Conference floor control requires that the participants have additional information about the other participants and can control their behavior. This could be provided by a conference-aware SIP phone, which might, for example, present a list of all the participants, and allow the conference chair to choose the next speaker or disconnect a participant.

The SIP Events mechanism <RFC 3265> provides the ability for one party to request additional status from another. This mechanism can be used by a conference server to request the status of the participants, and also by the participants to request the status of the other participants from the server. The information that would be provided is not yet standardized, but the conference state package <draft-ietf-sipping-conference-package-00> defines what this might include.

Conference control using SIP is still not well standardized, but work is continuing to bring consensus to this area. Solutions based on proprietary extensions are being developed, but until the standards mature, there will be limited interoperability of the higher-level features.

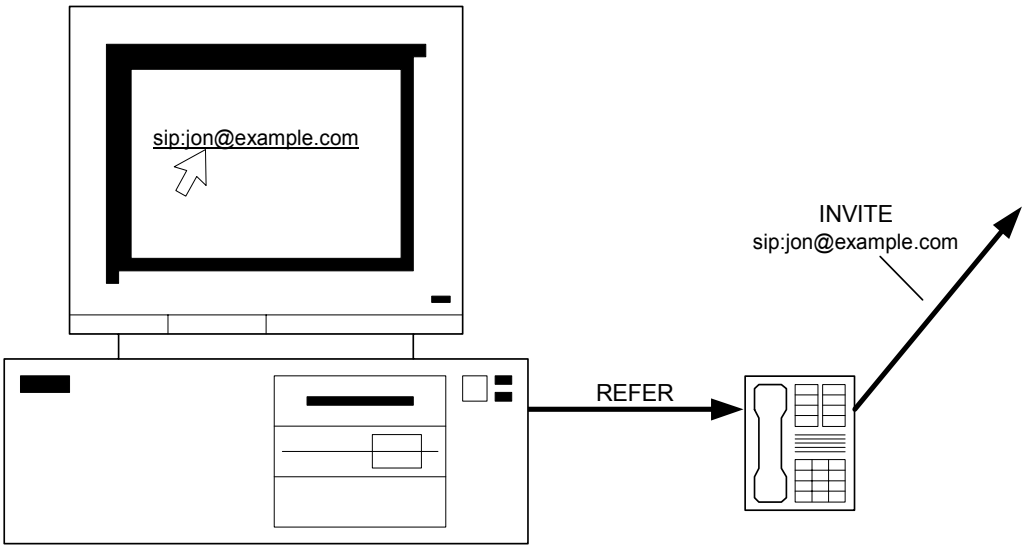
Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

7.5 Click-to-call or click-to-dial

Click-to-call describes the ability for a hyperlink on a web page to initiate a telephone conversation to the referenced destination. This would be extremely useful in a web-based directory service or as a marketing tool. For example, it could be used to link directly from a company's website to a sales representative. However, in order for a click on a browser to initiate a phone call, the browser must be able to control the phone, and this functionality is not currently widely available.

Most people use a PC as their web browser, so the phone must either be a soft-phone on the same PC and integrated with the browser, or a hardware phone that is somehow under the PC's control. Soft-phones do not generally offer a great user experience, because PCs are not designed as telephones, and there is little integration between web browsers and proprietary phone systems to allow control of separate phones.

This integration of web and telephony was one of the early promises of SIP, but this is not yet widely used. A pure SIP solution to this problem requires a click in a browser to issue a REFER request to a designated phone to cause it to make the call. This is equivalent to the third party call control scenario described earlier.



Now that third party call control using SIP is being standardized, and suitable security enhancements are being defined to ensure that the system is secure, standardized browser extensions are possible to provide this functionality with any SIP phone.

The prevalence of this integration will increase rapidly as SIP replaces proprietary protocols in enterprise telephony systems, and as standard browser add-ins become available to control enterprise phones directly, using SIP, and indirectly through control of the PBX.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

7.6 ENUM

ENUM aims to leverage the familiarity of existing telephone numbers on to Internet addresses. It defines a unique mapping between international phone numbers and host names in a way that enables DNS to be used to resolve the host name to an IP address, and the responsibility for maintenance of the DNS records to be delegated to the relevant country and regional authorities. The mapping is defined in RFC 2916, and, for example, +44 20 8366 1177 maps to 7.7.1.1.6.6.3.8.0.2.4.4.e164.arpa.

For SIP, ENUM provides a standard mapping between traditional phone numbers and Internet addresses, which could simplify the creation of an integrated PSTN and IP telephony system. However, there is currently very limited adoption of the standard and it is not gaining rapid traction. It would be straightforward to provide a default service provider as a gateway from the Internet into the PSTN, but it is not clear how this would be configured if multiple service providers are providing equivalent gateways.

draft-ietf-sipping-e164-04 describes the details of how to use ENUM to map between telephone numbers and SIP uris.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

8 The future

Over the past 5 years, SIP has evolved from a flexible but limited protocol suitable for use in NAT-less IP networks, to a protocol in use across the Internet and at the core of the next generation of commercial telephony networks with their hybrid IP/TDM networks. Much work has been done to enhance SIP to support the QoS and other regulatory requirements, and it appears that most are now close to resolution.

With large-scale deployments such as Vonage and Yahoo!BB, and SIP phones now being mass-produced and available for under $100, the residential and SOHO markets are beginning to take off. At the same time, the increasing availability of SIP-enabled PBX solutions is driving enterprise adoption, and SIP deployments by the major carriers to replace the PSTN will start once all the regulatory issues are resolved. Every indication is that this combination will continue an exponential growth in SIP usage over the coming years.

The potential demand for VoIP is huge, but it is worth remembering that its users care about the services offered and the cost, rather than the underlying technology. Now that SIP equipment is becoming easy to install and use, broadband Internet providers can provide a basic telephony service at a very low cost, and increasingly they will offer such a service. Countering this, there will be increased charging for broadband connections based on bandwidth use, due to the spread of bandwidth-hungry applications, but this is unlikely to be at a level to impose a significant cost on audio services.

As margins are squeezed due to this increased competition, the network will increasingly becomes a commodity, and additional services, such as higher quality of service (QoS), interconnection to the PSTN, unified messaging systems, and mobile coverage will be the products that can command a premium.

The main risks to this picture are that

- the standards diverge as a result of the competing demands of its different uses, and that SIP loses the simplicity, interoperability and flexibility on which it was based
- the regulators limit the use of SIP telephony, or incumbent telephony suppliers maintain their monopoly grip and limit competition in the network provision
- SIP is dropped for use in next generation networks because its advantages are overwhelmed by commercial and regulatory requirements.

However, SIP has the opportunity to provide a flexible framework for true telephony interoperability between fixed, wireless, free and commercial services, and to provide seamless enhanced services across multiple networks.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

Many powerful organizations are backing the use of SIP. Not all of them will be winners as a result of its success. Which ones are depends crucially on how the protocol develops and is deployed. If the protocol remains open and interoperable, then the user will benefit from increased competition and enhanced services. However, if the protocol becomes non-interoperable islands, then this promise will be delayed, although it seems unlikely that this progress will be stopped completely.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

9 Further information

9.1 Web-sites

SIP working group
- Official site: http://www.ietf.org/html.charters/sip-charter.html
- Supplemental site: www.softarmor.com/sipwg

SIPPING working group
- Official site: http://www.ietf.org/html.charters/sipping-charter.html
- Supplemental site: www.softarmor.com/sipping

SIMPLE working group
- Official site: http://www.ietf.org/html.charters/simple-charter.html
- Supplemental site: www.softarmor.com/simple

Others
- SIP Centre: www.sipcenter.com
- SIP Forum: www.sipforum.org
- Henning's SIP pages: www.cs.columbia.edu/sip
- PacketCable: www.packetcable.com
- Multiservice Switching Forum (MSF): www.msforum.org
- International Packet Communications Consortium (IPCC): www.packetcomm.org

9.2 IETF RFCs and drafts

9.2.1 Application Control with traditional keypad

RFC 2833	RTP Payload for DTMF Digits, Telephony Tones and Telephony Signals

9.2.2 Early media

RFC 3311	SIP UPDATE message
RFC 3262	Reliability of Provisional Responses
draft-camarillo-sipping-early-media-02.txt	Early Media and Ringing Tone Generation in SIP

9.2.3 Overlap dialing

RFC 3578	Mapping of Integrated Services Digital Network (ISDN) User Part (ISUP) Overlap Signalling to SIP

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

9.2.4 3G Mobile

RFC 3320	Signaling Compression (SigComp)
RFC 3327	SIP Extension for Registering Non-Adjacent Contacts
RFC 3485	The SIP and SDP Static Dictionary for Signaling Compression (SigComp)
draft-ietf-sipping-3gpp-r5-requirements-00.txt	3rd-Generation Partnership Project (3GPP) Release 5 requirements on SIP

9.2.5 AAA and security

draft-ietf-sip-authid-body-02	SIP Authenticated Identity Body (AIB) Format
draft-ietf-sip-identity-01	Enhancements for Authenticated Identity Management in SIP
draft-ietf-sip-smime-aes-01	S/MIME AES Requirement for SIP
draft-ietf-sipping-aaa-req-03.txt	Authentication, Authorization and Accounting Requirements for SIP
draft-mahy-sipping-smime-vs-digest-01.txt	Discussion of suitability: S/MIME instead of Digest Authentication in SIP
draft-jennings-sipping-certs-01	Certificate Discover for SIP

9.2.6 Caller Preferences

draft-ietf-sip-callerprefs-09	Caller Preferences for SIP
draft-ietf-sip-callee-caps-00	Indicating User Agent Capabilities in the Session Initiation Protocol (SIP)

9.2.7 Conferencing

draft-ietf-sipping-3pcc-04	Best Current Practices for Third Party Call Control in SIP
draft-rosenberg-sipping-conferencing-framework-01.txt	A Framework for Conferencing with SIP

9.2.8 NAT and firewall traversal

RFC 3489	STUN - Simple Traversal of UDP Through NATs
RFC 3581	An extension to SIP for Symmetric Response Routing
draft-ietf-mmusic-sdp4nat-05	RTCP attribute in SDP
draft-ietf-mmusic-rtsp-nat-01	How to Enable Real-Time Streaming Protocol (RTSP) traverse Network Address Translators (NAT) and interact with Firewalls.
draft-rosenberg-midcom-turn-01	Traversal Using Relay NAT (TURN)
draft-rosenberg-sipping-ice-01	Interactive Connectivity Establishment (ICE): A Methodology for NAT Traversal for SIP

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

9.2.9 Device configuration

RFC 3361	DHCP Option for SIP
draft-ietf-sipping-config-framework-00	A Framework for SIP User Agent Configuration
draft-ietf-sip-mib-07	Management Information Base for Session Initiation Protocol (SIP)

9.2.10 Presence and Instant Messaging

RFC 3265	SIP Specific Event Notification
RFC 3428	Session Initiation Protocol Extension for Instant Messaging
draft-ietf-simple-message-sessions-01	Instant Message Sessions in SIMPLE
draft-ietf-simple-presence-10	A Presence Event Package for the Session Initiation Protocol (SIP)
draft-houri-simple-arch-01	SIP/SIMPLE Based Presence and IM Architecture

9.2.11 QoS

RFC 3312	Integration of Resource Management and SIP
RFC 3313	Private SIP Extensions for Media Authorization
draft-ietf-sip-resource-priority-01.txt	Communications Resource Priority for SIP

9.2.12 Other documents

MSF Technical Report MSF-TR-QoS-001-FINAL	Quality of Service for Next Generation Voice over IP Networks
PacketCable Specification PKT-SP-DQOS-I03_021116	PacketCable Dynamic Quality of Service Specification

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.

10 About Data Connection Limited (DCL)

Data Connection Limited (DCL) is the leading independent developer and supplier of portable protocol software suites for VoIP (SIP, MGCP, Megaco), VPN (RFC 2547 MPLS/BGP, Martini, VPLS), IP Routing (OSPF, IS-IS, BGP, CSPF), MPLS (GMPLS, UNI, NNI), ATM (PNNI, SPVC, UNI) and SNA, and Conferencing, Messaging, and Directory solutions. Customers include Alcatel, Cabletron, Cisco, Fujitsu, Hewlett-Packard, Hitachi, IBM Corp., Microsoft, Mitel, NEC, Nortel, Siemens, SGI and Sun.

DCL is headquartered in London UK, with US offices in Reston, VA and Alameda, CA. It was founded in 1981 and is privately held. During each of the past 21 years its profits have exceeded 20% of revenue. Last year, sales exceeded $40 million, of which over 90% were outside the UK, mostly in the US. Even through the current severe downturn, Data Connection's financial position remains secure, as does its employee base: its 200+ software engineers have an average length of service of 8 years, with turnover of <3% annually.

DC-SIP provides a complete SIP User Agent and Proxy toolkit for building high-performance SIP devices. DC-SIP supports the latest RFCs, including RFC 3261 and many extensions, and is used by customers around the world to build scalable and robust SIP devices. DC-SIP is supplied pre-integrated with Windows, Solaris, Linux, VxWorks, OSE and LynxOS, and is readily ported to other environments.

All of the Data Connection protocol implementations are designed for scalability, distribution across multiple processors, and fault tolerance. We have extremely consistent development processes that result in on-time delivery of highly robust and efficient software. This is backed up by an exceptionally responsive and expert support service, staffed by engineers with direct experience in developing the protocol solutions.

DCL also supplies integrated solutions incorporating SIP and its other technologies in web-conferencing and unified messaging solutions, and as a complete class 5 replacement switch through its Metaswitch division.

Data Connection is a trademark of Data Connection Limited and Data Connection Corporation. All other trademarks and registered trademarks are the property of their respective owners.

Copyright © 2002-2004 Data Connection Limited. All Rights Reserved.