The future of communication using SIP

Autor(en): **[s. n.]**

Objekttyp: **Article**

Zeitschrift: **Comtec : Informations- und Telekommunikationstechnologie = information and telecommunication technology**

Band (Jahr): **80 (2002)**

Heft 11

PDF erstellt am: **28.06.2024**

Persistenter Link: https://doi.org/10.5169/seals-877252

Session Initiation Protocol (SIP)

The Future of Communication Using SIP

The Session Initiation Protocol (SIP) has been called the ISUP of next-generation networks. Certainly, the two protocols share many characteristics, but whereas ISUP embodies the evolution of common-channel signalling over the span of many years in the history of telephony, SIP has the potential to be a groundbreaking, disruptive force of change.

We foresee that the adoption of SIP will usher in a new era of multimedia communications that leverage the strengths of Internet technologies to provide opportunities for innovation.
This article describes the session initiation protocol (SIP), which is the new standard for establishing multiparty, multimedia communication in IP-based networks, and some of the implications that its adoption carries for the future of communications. It also gives a brief overview of the business drivers, which led to the specification and development of the SIP-based IP Multimedia System (IMS).
Ericsson is poised to capitalise on the revolution that SIP heralds, which offers an exciting promise for the communication networks of the future.

How SIP works

SIP is an application-level control protocol for establishing, modifying and terminating multimedia sessions between one or more participants; it supports multimedia conferencing, Internet telephone calls, registration and redirection services, and is easily extended. It traces its roots to a number of multiparty conferencing initiatives in the history of the IETF as well as the Web and Internet email. As a result, SIP embodies not only a distinct protocol, with syntax and semantics for the messages to be exchanged, but also a philosophy of end-to-end control over session establishment with support from servers in the network.

The place of SIP in the Internet protocol space

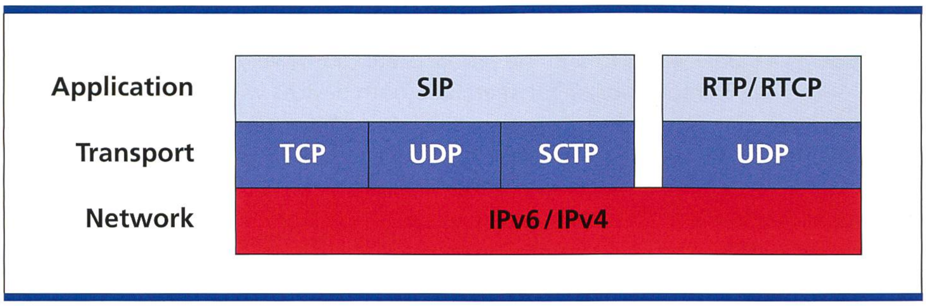
Figure 1 shows the relationship of SIP to various other protocols in use in the Internet.

Basic session establishment

In the simplest case, two users wish to communicate with one another using a variety of media types (such as audio, video and text messages) over an IP network. The software application that enables each party's communication is known as a User Agent (UA). The UA could be running as a "softclient" on a PC, as the operating software in a mobile device, or in the firmware of a desktop SIP phone.
Party A, wishing to communicate with party B, sends an INVITE message to B, who is listening on the official SIP port (5060). The INVITE body contains information encoded with the Session Description Protocol (SDP) indicating the types of media he is willing and able to use. B's response similarly indicates his preferred media types. Once A sends back an ACK (acknowledge) message, both parties are aware of each other's IP addresses and port numbers on which they want to receive the media streams. Both SIP parties know which types and bandwidth of media the other is able to receive. Upon sending and receiving of the ACK, both ends may begin transmitting data to the corresponding receiver ports via a separate media connection using the Real Time Protocol (RTP), or some other appropriate transport protocol. Throughout the duration of the session, either party can make updates (indicating a new set of media types, addition of new parties to the session, or other changes), by sending additional SIP messages. At the end of communication, either party may send a BYE message, which indicates termination. When the other party sends a response the session is terminated.
For performance reasons, UDP (User Datagram Protocol) is a required transport protocol. TCP and others are optional. Because of the unreliable nature of the datagram service in UDP, SIP contains its own retransmission mechanisms, including the three-way exchange described above between nodes for establishment of sessions.

Syntax and Addressing

Destinations in SIP may use Uniform Resource Identifiers (URI) with the same format as email addresses. A valid SIP



Relationship of various protocols used on the Internet.

Terminals

address might be something like: sip:Timon.Springer@ericsson.com. This implies the use of DNS to map host and domain names to IP addresses, which is a key aspect of the integration of SIP with web and email enabled technologies, which are already familiar with the concepts of URIs and their interpretations.
The close connection between SIP and DNS facilitates interoperability with telephony systems as well as their addressing mechanisms. Support for E.164 numbers in DNS (ENUM) allows SIP servers and clients to send and receive telephone numbers in place of SIP URIs, and to route them in a sensible way.
SIP is a text-based protocol, and it reuses the structure of HTTP and SMTP messages with numerous informational headers followed by a (possibly multipart) body. As a result, scripting languages such as Perl and Python are well suited to automating many session-processing tasks in a SIP server.

Indirection

A key concept in the way SIP works is indirection. With indirection the current location of a user is hidden behind a more permanent user identity or URL (uniform resource locator). A network-based SIP server then binds this mobile identity to the more permanent URL, much in the same way a home agent acts in a mobile IP network. The two principal mechanisms in SIP to support this are redirection and proxying.
In the event that party A does not know party B's address directly, it is possible to first send the INVITE to what is known as a Redirect Server. The redirect server sends back a response indicating where party B can be found (usually in the form of a SIP URI), which party A can use to send a new INVITE.
The use of a Proxy Server allows users to have a node in the network that performs some intermediary function before routing SIP messages to their destination on behalf of the UA. In the event that such a node exists, SIP messages that it receives are forwarded to the appropriate destination. Responses are forwarded in the reverse direction such that the proxy looks to each endpoint as if it were the other (from a signalling point of view).
In the case of both a redirect and a proxy server, a Location Server may be consulted for information about the current SIP address of the indicated destination. The interface between the proxy or redirect server and the location server is not defined in the SIP RFC, but can be some appropriate querying interface such as LDAP, HTTP, or DIAMETER. SIP supports real-time updates to the location server's database via a REGISTER message that indicates the user's current location via a SIP URI. The SIP Server that receives these REGISTER messages and updates the location database is known as a SIP Registrar. Through information obtained from the registrar, the redirect or proxy server is able to re-route a SIP request to the destination where the user wants to be reached.
However, indirection is not limited to the user, but may also be applied to the SIP servers themselves, using DNS. A number of DNS mechanisms exists to map a symbolic name for a SIP server into a concrete IP address where that server may be reached. Particularly interesting are SRV records, which allow the definition of one or more SIP servers that are the first point of contact for a given domain. For example, four separate SIP servers that share the load for that domain may service the ericsson.com domain. The use of DNS makes it very easy to establish these network topologies. Therefore, the process by which a user is contacted using SIP involves both determination of the server(s) for that user as well as determination of the location of that server. Indirection may be used in both of these steps allowing for a very flexible and fluid communications network.

Forking

A particularly useful aspect of the behaviour of SIP proxies is their ability to forward a SIP message to multiple destinations simultaneously, in a procedure known as forking. This allows a user to have multiple destinations registered simultaneously (say, a mobile device as well as a desk phone) and have both destinations alerted when a new session request arrives. The proxy server handles correlation of the responses received from various branches and ensures that only a single upstream response is sent to the client. This enables a number of interesting and useful features.

Using SIP in various environments

In order to bring useful SIP implementations to the market, it is necessary to overcome some hurdles that exist in particular environments. Knowing these hurdles and their solutions helps explain why Ericsson made various architectural choices for the solutions under development.

SIP on the public Internet

SIP on the public Internet is the default case, given the high bandwidth and low latency between network elements (making QoS mechanisms less important) and the lack of differentiated domains that would introduce such elements as firewalls, NATs (Network Address Translator) or other gateways.

SIP in protected domains

For many users, the default scenario described above is incomplete. Most notably, the network is often not "transparent" end-to-end, at or above the IP layer. Provisions must be made to ensure unchanged behaviour from an end-user perspective, and the solutions are not always simple.

For security purposes, many IP networks are protected from external traffic by the installation of a firewall, which only allows packets to flow through designated secure points. For both security purposes and address conservation, many ISPs and enterprises allocate private IP addresses to their users that are then mapped to public IP addresses by a NAT. For "pure" client-server applications such as the Web, file transfer, or email, in which the precise real-time network address of the end-user is irrelevant, application-level proxies can provide unimpeded access to the services needed, and firewalls and NATs are not show-stoppers.

However, although SIP differentiates the signalling from the media in establishing a real-time communication session, the two are nevertheless highly interdependent, and firewalls and NATs can be problematic.

A SIP proxy that is used to pass messages through a firewall, for example, must not only provide the indirection support for locating the users on either side, but also have some means of instructing the firewall to allow the media to pass through as well (since the media may use an entirely different combination of IP addresses and port numbers).

Similarly, if a NAT transforms the private IP address used by a SIP UA into a public address, then the private media address described by the SDP in the SIP message will be unusable by the remote UA. Therefore, an Application Layer Gateway (ALG) must be coupled to the NAT to ensure that the SDP is re-written to reflect the public address that will be used to transit the media across the domain boundary.

SIP in a mobile environment

Limited bandwidth over the air interface means that the amount of signalling must be kept to a minimum. The size of SIP messages may become quite large because of the number of steps taken in routing between endpoints and the fact that SIP is a text protocol rather than binary. This necessitates the establishment of a node that can perform signalling customisations for efficiency without affecting end-to-end transparency. Such a node (a SIP proxy) can also serve as a useful default inbound and outbound proxy for such things as address translation services, invocation of locally significant services such as taxi locators, weather, or news, and guaranteeing QoS for network resources.

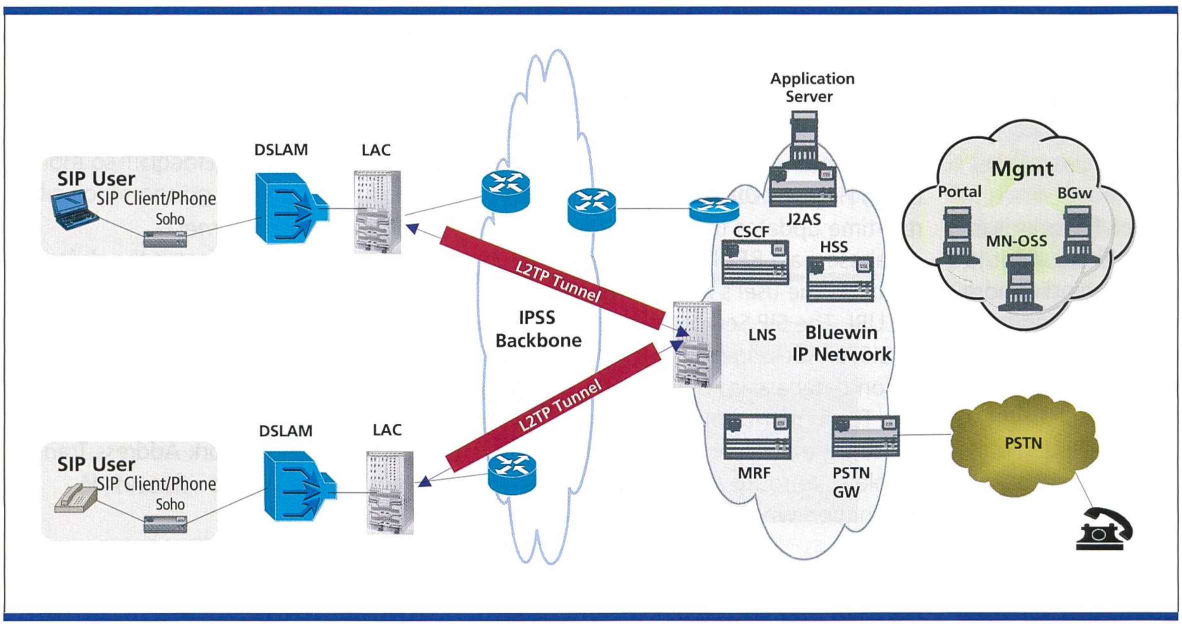
Consequences

Because of the particular requirements imposed on SIP-based networks in these environments, there is great benefit to be derived from an architecture that supports some or all of them in a flexible manner. As we shall see in a later section, a number of specialized nodes have been identified by, among others, 3GPP to accommodate these requirements and provide operators with a powerful framework for delivering added value to the end-user experience.

Capabilities Supported by SIP

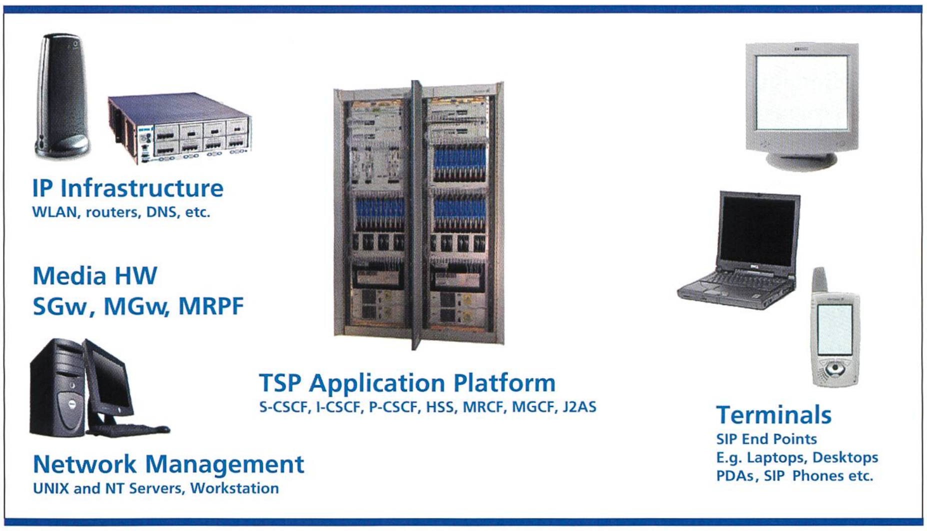
The following section lists some examples of services and service building blocks that are made possible by the basic mechanisms of SIP.

- Add/Drop Media: SIP supports the ability to add new media types (or remove unwanted media) in the middle of a session. Additionally, disparate media types can be supported at the various endpoints without degrading anyone's experience (i.e. lowest-common-denominator communications are not necessary).
- Find Me/Follow Me: Register (with a SIP Registrar) in multiple locations simultaneously according to your daily patterns. Different SIP-enabled devices like fixed phone at home and at work, mobile phone, PC at work and/or at home, and the phone and/or PC of a secretary. Incoming calls may ring simultaneously in all locations or sequential in a user-defined order, and when answering the call on any of the SIP devices the other devices stop ringing. The person calling is never aware of anything but you answering. Note that none of what we have described so far requires specialised PBX equipment!
- Presence and Instant Messaging: Sending and receiving instant messages via SIP. SIP will be one of the leading



EMM in Swisscom Network

choices for transport of presence and instant messaging information between users, as it is extremely well suited to this purpose.
- Conferencing and Distance Working.
- The Abolition of Class 5 Services: Traditional voice services that previously provided substantial revenue to incumbent local operators are made trivial by SIP. Services such as caller ID, call waiting and call hold are handled by basic SIP mechanisms within the UA and require no input or control by the network.
- Multiparty Gaming: SIP can be used to transport a wide range of real-time information, including gaming events for multiplayer games. Additionally, SIP (whether supported by the game client itself or via a separate communicator application) can be used during the game for audio, video and chat between players.
- VPNs Made Simple: The use of DNS, ENUM, and the indirection capabilities of SIP make VPNs simple to develop and manage. A single SIP proxy can provide address mapping and forwarding services for a remote location, making it appear to the user that they are located within the same corporate domain as their colleagues across the country.



Equipment Overview

Business Motivation for SIP-based Systems

We are currently witnessing a dramatic change in the way we communicate. The monolithic incumbent telcos, like Swisscom, who only a few years ago provided all or most of our communications needs, are now under constant pressure.

Within the next decade operators will need a new generation of networks, primarily based on IP technology. These will open up new revenue opportunities and lead to customer-driven services being the key to profitability. At the same time, these new networks do not have the same value chain as traditional telephony, and the business models are not yet established. Operators need to find ways to take advantage of these new networks and move their businesses into a totally new space. Providing the same set of services using new technology will not enhance the end-user value. In fact, as we have seen, a re-implementation of traditional services will only drive margins down and eventually lead to operator losses. The deployment of SIP-based networks can be a key part of the solution to this need.

Elements of the SIP technology and business value proposition, as compared to existing systems, include but are not limited to:
- Presence: Presence means that a group of individuals can share information about their current availability status. SIP provides new creative ways of developing services based on presence information. The value is not so much presence as an individual service, but in combining presence with other multimedia capabilities such as "combinational" service offerings.
- Combinational services: It will be easy to combine conversational multimedia services with other categories of services such as directory information, web browsing, positioning and presence. For example, a location based service can be developed where a conversational communication session is combined with positioning information and maps to provide information related to the geographical location of the parties involved.
- Access independence: SIP itself, being an application-layer protocol in the IP-based suite, is access independent in nature and offers seamless service capabilities between fixed and mobile networks. This is a key element in making the promise of fixed-to-mobile convergence a reality.
- New charging models: The operator will have the flexibility to define new charging models such as charging based on actual media usage. For example, if the communication between two parties starts with a real-time voice session and video is added later on, it's possible to charge for the sessions individually and for the actual media usage during the sessions.
- Quality of Service (QoS): Given the importance of conversational multimedia services, where real-time sensitive sessions involving voice and video streams are shared with less time sensitive streams, it is crucial that sufficient QoS mechanisms are in place, guaranteeing a rich experience for the end-user. New product offerings will provide a robust and flexible architecture supporting the required QoS and security requirements in mission critical applications. SIP-based products sit on top of this IP network and take advantage of the capabilities of the underlying network to provide QoS.

Conclusion

SIP is a central part of the value proposition of future multimedia networks. This value stems from several key aspects of the protocol, including:
- The flexibility of addressing, routing and modifying messages using the protocol.
- The support for a wide range of media types, simultaneously invoked or selectively added as the need arises.
- The wealth of information that can be communicated to all nodes in the network, fostering an end-to-end view of services and applications.
- The ready integration with web-enabling technologies that springs from SIP's origins in the IETF.

Ericsson brings several new products and solutions into this new technology space. These offerings will enable operators to position themselves in new markets and realise new revenue streams while protecting their investment in legacy systems. 6

Ericsson AG
Telekommunikation Zentrale
Ruchstuckstrasse 21
CH-8306 Brüttisellen
Tel. 01 807 22 22
Homepage: www.ericsson.ch

References

Mark Handley et al.:
The Session Initiation Protocol, draft-ietf-sip-rfc2543bis-05.txt, http://www.ietf.org, October 2001
Gonzalo Camarillo: SIP Demystified, McGraw-Hill, 2002

J. Rosenberg et al.: SIP Extensions for Presence, draft-ietf-simple-presence-04.txt, http://www.ietf.org, Novermber 2001

J. Rosenberg et al.: SIP Extensions for Instant Messaging, Draft-ietf-simple-im-01.txt, http://www.ietf.org, July, 2001

Summary

SIP oder die Zukunft der Kommunikation

Das Session Initiation Protocol (SIP) wird auch das ISUP der nächsten Netzgeneration genannt. Tatsächlich haben die beiden Protokolle vieles gemeinsam, aber während das ISUP in der Geschichte der Telefonie die Jahre dauernde Entwicklung der Zentralkanalsignalisierung verkörpert, hat das SIP die Voraussetzungen, die Dinge von einem Tag auf den anderen von Grund auf zu verändern. Das SIP leistet einen entscheidenden Beitrag zur Wertbeständigkeit künftiger Multimedia-Netze. Ericsson bedient diese Technik mit zahlreichen neuen Produkten und Lösungen. Diese eröffnen den Anbietern die Chance, sich neue Märkte und neue Umsatzquellen zu erschliessen, ohne ihre Investitionen in traditionelle Systeme zu gefährden.

FIRMEN UND PRODUKTE

Netzwerk-Systemarchitektur

Sun präsentiert ihre Roadmap für N1, die Vision für die nächste Systemgeneration. N1 bündelt die weit verteilten Rechenressourcen (Server, Speicher, Software und Netzwerke) zu einer leistungsfähigen, zentralen Einheit. Durch die Automatisierung der komplexen Verwaltungstechnologie können Business- und IT-Manager den Auslastungsgrad, die Effizienz und die Flexibilität ihrer Datenzentren erheblich verbessern. N1 baut auf 20 Jahren Erfahrung von Sun in der Forschung und Entwicklung von Systemarchitekturen und bildet den Eckstein der Sun-Systemstrategie der nächsten Generation. Dank netzwerkzentrierter Computing-Innovationen wie NFS (Network File System), Dynamic System Domains, Java™, Solaris™ und die Software Sun ONE Grid Engine stehen den Kunden von Sun alle Vorteile der neuen N1-Architektur zur Verfügung. N1 wandelt herkömmliche IT-Infrastrukturen in ein effizienteres Computing-Modell um. Die N1-Architektur ermöglicht echtes Computing nach Bedarf: Die Ausgaben der Kunden richten sich nach ihrem Wachstum, und die Verwaltung der Technologie in Unternehmen erreicht ein weitaus höheres Niveau. Dies ist eine langfristig angelegte Strategie. Sun wird deshalb die N1-Produkte ab diesem Kalenderjahr phasenweise auf dem Markt einführen. Die Roadmap stellt sich wie folgt dar:

Phase 1: Virtualisierung von Ressourcen (ab 2002)
In dieser ersten Phase wird die grundlegende Infrastruktur für N1 bereitgestellt. Die Kunden wandeln allmählich individuelle Computer, Netzwerkelemente und Speichersysteme in einen zentralen Ressourcenpool um. Das System weist die Verwendung dieser Ressourcen zu, überwacht und misst sie.

Phase 2: Bereitstellung von Services (ab 2003)
In Phase 2 definieren die Administratoren die gewünschten Services, wie zum Beispiel E-Banking, und N1 übernimmt die Bereitstellung der erforderlichen Ressourcen aus dem in Phase 1 aufgebauten virtuellen Computer.

Phase 3: Automatisierung von Richtlinien (ab 2004)
In Phase 3 verwaltet N1 schliesslich die Vorgaben für die Applikationsdienstebenen automatisch. Mit Richtlinien, welche die geschäftlichen Anforderungen und Prioritäten widerspiegeln, werden die Applikationen und die erforderlichen netzwerkweiten Ressourcen gesteuert. Der in Phase 2 eingerichtete E-Banking-Service lässt sich beispielsweise so konfigurieren, dass VIP-Kunden vorrangigen Zugriff haben.

Sun Services unterstützt die Kunden beim Aufbau der für N1 notwendigen Systemarchitektur und übernimmt die Anleitung beim Übergang zu einer N1-Umgebung – angefangen bei der Konzeption bis hin zur Umsetzung und Verwaltung, unterstützt durch wissensbasierte, intelligente Technologien.

Sun Microsystems (Schweiz) AG
Rolf Widmer
Javastrasse 2, Hegnau, 8604 Volketswil
Tel. 01 908 90 00
E-Mail: rolf.widmer@sun.com
Homepage: www.sun.ch