EXTENSIONS OF SESSION INITIATION PROTOCOL FOR NEXT GENERATION SERVICES

Ida Čelam
Research & Development Center
Ericsson Nikola Tesla
HR-10002 Zagreb, Croatia
Ida.Celam@etk.ericsson.se

Zoran Kokolj
Research & Development Center
Ericsson Nikola Tesla
HR-10002 Zagreb, Croatia
Zoran.Kokolj@etk.ericsson.se

Abstract — **The IETF's Session Initiation Protocol (SIP) is a flexible, yet simple tool for establishing interactive connections across the Internet. It was originally developed for initiation of multimedia sessions. SIP, internet multimedia, voice over IP and IP telephony have become quite popular, both within IETF and other standard groups, and the applications of SIP have grown. Result of this popularity has been a continual flood of suggestions for SIP modifications and extensions. One of SIP's most important features is flexibility, i.e. ease of its extension (with new methods, new header fields, new body types, and new parameters). There have been countless proposals to do just that. The intention of this paper is to give a short overview of some of the most interesting extensions of SIP protocol.**

Keywords — SIP; extensions; multimedia; services; intreworking;

I. INTRODUCTION

Popular for its most important strength – flexibility, IETF's Session Initiation Protocol (SIP) [1], originally designed for establishment of multimedia sessions, has become target of new ideas for its extension. There are many suggestions, including new methods, header fields, body types and parameters. Their intension is to broaden the current functionality of SIP protocol and make it more efficient.

The IETF management task has been to keep the SIP protocol development focused on its core strengths and applications it does best. Therefore, an IETF process has been put into place to define how extensions are to be made to the protocol [2]. The process is designed to ensure that proposed extensions are appropriate for SIP (as opposed to being done in some other protocol), that they fit within the protocol model and framework, are consistent with its operation, and that they solve problems generically rather than for a specific use case.

There are many extensions currently suggested for SIP and some of them are already accepted and being put into standard. The intention of this paper is to give a short overview of some of the most interesting extensions like SIP for telephony, reliable provisional responses, resource reservation, session timer, refer method, presence and instant messaging and requirements from Third Generation Partnership Project (3GPP).

II. SIP IN NEXT GENERATION NETWORKS

Next generation networks are multiservice networks based on packet switching technology. A dominant approach to their implementation is based on IP protocol, the only available technology that provides end to end connectivity.

Because packet networks are ill equipped to handle real time multimedia traffic, inherently sensitive to delay and jitter, a wide range of protocols was introduced to handle newly arisen requirements. The most important elements of multimedia network protocol architecture are depicted by Figure 1.

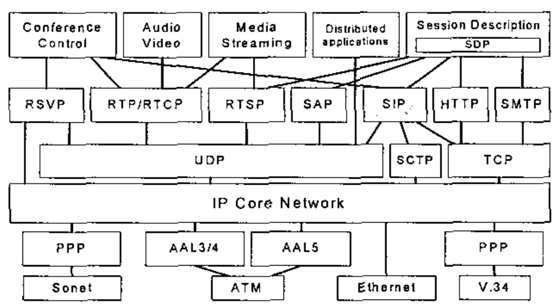

Figure 1. Multimedia network protocol architecture

Session Initiation Protocol is a signalling protocol whose primary purpose is to control multimedia sessions. There is another contender for this title, ITU-T supported H.323 protocol. To avoid going into too many details, we will just state that main advantages of SIP are its simplicity and broader applicability than just session handling.

SIP is an application level protocol. Its basic functionality includes establishment, maintenance and termination of multimedia session between two or more users. It also provides support for a wide range of services, such as multimedia conferencing, internet telephony and user mobility.

7th International Conference on Telecommunications - ConTEL 2003
ISBN: 953-184-052-0, June 11-13, 2003, Zagreb, Croatia

Role of SIP is somewhat similar to that of ISUP or Q.931 protocols in PSTN networks, with an important difference that it does not perform resource reservation and is practically independent of the underlying transport protocol. SIP signalling flow is distinct from media flow, usually carried by RTP protocol. Protocol structure is based on HTTP, which relates to textual message encoding, message syntax, transactions and client-server communication model. Table I lists basic SIP methods (requests) and response classes.

TABLE I. BASIC SIP REQUESTS AND RESPONSE CLASSES

REQUESTS	RESPONSE CLASSES
INVITE	100-199 Provisional informational
ACK	200-299 Success
BYE	300-399 Redirection
CANCEL	400-499 Client error
REGISTER	500-599 Server error
OPTIONS	600-699 Global failure
INFO	

Messages contain headers and payload. Payload is usually described with SDP protocol but, as it is transparently carried through SIP network, a number of formats are allowed. Support of MIME payload is particularly important for interaction with other network services, such as e-mail and WWW. SIP network architecture includes four network elements which are: *user agents* (transaction initiators), *proxy servers* (SIP traffic routers on application network level, with an option to modify outgoing messages), *redirect servers* (perform functionality similar to that of the proxy server, but produce reply with the destination address instead of routing requests by themselves) and *registrars* (user agent's interfaces towards abstract location network service).

Growing acceptance of SIP revealed several shortcomings in session handling procedures, but also an opportunity to use it in different applications with minimal changes to the original protocol. Subsequent chapters will present several extensions of the base protocol specification.

III. SIP FOR TELEPHONY

Internet telephony is one of the key services of a multiservice network. It allows transfer of voice data over IP based networks, with or without quality of service provisions. Currently SIP architecture does not include network gateway node. Applications of this node can be summarised in two broad categories: voice trunking over IP based network and voice communication between users in different networks. The main requirements put on those applications is total transparency of the packet network for PSTN signalling protocols. Semantics of the basic SIP protocol can not express all constructs required to interwork with PSTN network protocols. For that reason an extension, popularly known as SIP-T, was introduced [3]. Primary purpose of this extension is to allow seamless conversion between SIP and ISUP. Long-term objective is to enhance the specification until it is able to cope with all relevant PSTN protocols, including Q.931.

SIP-T handles the requirement for transparent transport of ISUP signalling with two mechanisms. ISUP message is stored in the SIP message as MIME payload. To enable message routing through SIP network relevant information is extracted from ISUP messages and transformed into SIP message header data. The most important parts of information are called party number and media stream description. This extension also provides support for in-band signalling, supported in ISUP by INF/INR signals, by introducing the INFO method. Unfortunately, not all problems are solved, as overlap receiving and DTMF signal transfer is still an open issue. Figure 2 introduces the network architecture of distributed network gateway and related nodes, capable of providing PSTN and SIP internetworking.

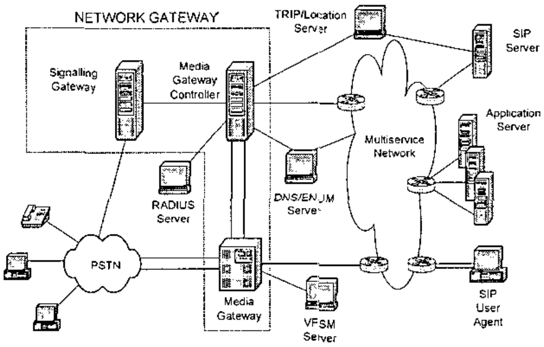

Figure 2. Distributed PSTN- SIP network gateway

IV. RELIABLE PROVISIONAL RESPONSES

Internetworking between SIP and PSTN networks, made possible by network gateways, requires higher transaction reliability than provided by the basic SIP protocol. The main problem is unreliable delivery of provisional responses (like *183 Session Progress Message*) to the call originator, especially when UDP is used as a transport protocol. These messages are particularly important when internetworking with ISUP, as some in-band signalling needs to be exchanged before a final response is issued.

Solution to this problem is presented in [4]. It proposes a new PRACK method, which acknowledges reception of provisional responses in a similar way as ACK acknowledges INVITEs. Important difference is that PRACK initiates a new transaction, terminated by a final response. This behaviour is essential to allow seamless protocol extension.

V. RESOURCE RESERVATION

SIP was intentionally made independent of the quality of service and resource reservation mechanisms. In the typical case Resource Reservation Protocol (RSVP) is used for that purpose. Independence causes serious problems in some applications, including internet telephony. One problematic scenario can be summarised as follows. When a SIP user agent receives an INVITE request it has to indicate acceptable session parameters in one of the subsequent responses, either provisional or final. Compatibility of communicating sides and

ConTEL 2003, ISBN: 953-184-052-0

availability of network resources can only be determined when this response is received by the calling side. However, called user agent indicates the existence of incoming call to the user immediately after INVITE is received. This causes the phone to start ringing although there is no guarantee that call establishment is possible. QoS provision problem is closely related. Best effort service approach is not applicable in a number of situations. Example is again internet telephony. Amount of available network resources can be sufficient for call establishment, but too scarce to ensure acceptable voice quality. This is of outmost importance when charging functionality exists in the network, as users demand appropriate service for their money.

The latest proposal on how to handle such situations is given in [5]. Novelties are explicit expression of preconditions and an UPDATE method [6]. In some previous proposals a COMET method was used for similar purposes, but it is now depreciated. As preconditions listed in the INVITE request are usually not fulfilled a priory, the called user agent does not alert its user. Only after successful negotiation and resource reservation will the called user be notified. If that is not the case, call is silently discarded. Introduction of UPDATE method also provides support for early media, a one-way media flow used to inform the calling user about call progress or play various announcements. This flow is not subject to charging.

VI. SESSION TIMER

The SIP does not define a keepalive mechanism for the sessions it establishes. Although the user agents may be able to determine if the session has timed out using session specific mechanisms, call stateful proxies will not be able to do so. For instance, when a user agent fails to send a BYE message at the end of a session or when BYE message is lost due to network problems, a call stateful proxy will not know when the session has ended. In this situation the call stateful proxy will retain state for the call and has no deterministic method of finding out when the call state information no longer applies. Similar problem occurs during construction of SIP Network Address Translator (NAT) Application Level Gateway (ALG). The ALG embedded in a NAT needs to maintain state for whole duration of call. This state must eventually be removed. Relying on a BYE to trigger the removal of state, besides being unreliable, provides opportunity for denial of service attack.

To resolve these problems a new extension called *session timer* [7] is proposed. It allows periodic refreshes of SIP sessions through a re-INVITE or UPDATE request. User Agent sends periodic re-INVITE or UPDATE requests (referred to as session refresh requests) to keep the session alive. The interval for the session refresh requests is determined through a negotiation mechanism. If a session refresh request is not received before the interval expires, the session is considered terminated. Both UAs are supposed to send a BYE, so that call stateful proxies can remove the state of the call.

The extension is backwards compatible with SIP as it works as long as either of the two participants in the dialog understands it. Two new header fields, Session-Expires (conveys the duration of the session) and Min-SE (conveys the minimum allowed value for the session expiration), and a new response code, 422 *Session Interval Too Small*, are defined.

VII. REFER METHOD

REFER method is a new SIP extension introduced in [8] which demands that the recipient REFERs to a resource provided in the request. It provides a mechanism allowing the party sending the REFER (referrer) to be notified of the outcome of the referenced request. This can be used to enable many applications, including call transfer and conference invitation. The REFER method indicates that the recipient (identified by the Request-URI) should contact a third party using the contact information provided in the request. A REFER request implicitly establishes a subscription to the refer event. Besides defining a new SIP request called REFER, the extension also introduces a new request header field called Refer-To. It should be emphasized that implementation of this method requires changes in User agents, while SIP proxies, registrars and redirect servers can support it without modifications.

REFER method can be used for implementing useful applications such as call transfer fully described in [9]. The participants in a basic transfer should indicate support for the REFER and NOTIFY methods in Allow header fields in INVITE, 200 OK to INVITE and OPTIONS messages.

VIII. INSTANT MESSAGING AND CALLER PRESENCE

SIP network allows transparent redirection of call setup messages based on information stored in the location servers. SIP user agents provide this information with REGISTER method, issued on user login/logoff or explicit request for location change. However, as the same user can simultaneously register on several terminals or be temporarily absent without logging off, redirection facilities are not enough to guarantee that he can be reached at some point in time. For that reason a new service, called presence, should be introduced. Full potential of this service becomes apparent in conjunction with other network services, such as instant messaging. Instant messaging allows instant delivery of notifications between network endpoints. Originator of the message can be a person or a network server. This service is somewhat similar to *Simple Messaging Service* of GSM network, but with an important difference that messages can also be exchanged during session setup and execution.

Presence service is implemented using SIP event notification framework [10]. SIP user agent that wants to be informed about specific events needs to register to the server side, which is either another user agent or a network server. Two new methods are introduced for this purpose. SUBSCRIBE is used by the originating user agent to register for indication of a particular event, while NOTIFY is issued when the event occurs. The solution also requires a new network server, called presence server, which holds the presence information of the registered users. Presence server is expected to be co-located with the registrar server.

Instant messaging service profits from presence service because the probability that the message will reach the intended user is much higher. For that SIP is extended with the

MESSAGE method [11]. Because user information is added as a MIME type in MESSAGE method, its contents is not limited to textual format. This allows multimedia applications similar to *Multimedia Messaging Service* of the mobile networks.

It is expected that instant messaging and presence, along with SIP event notification framework, will be used to implement a wide range of new services. Examples are third party call control, networked control of electronic devices and user agents with graphical user interfaces that list currently reachable users.

IX. SIP IN 3GPP

As in other parts of the IP based world, there were two candidates for the role of session control protocol in Internet Multimedia Subsystem (IMS) of the 3G core network: SIP and H.323. SIP prevailed for a number of reasons, but mainly because of its simplicity and extensibility. Because of those characteristics SIP is constantly expanding its application area, and a typical example is the instant messaging and presence functionality presented in the previous chapter. On the other hand, applicability of H.323 seems to be bound to session control. Adoption of SIP as the primary signalling protocol in IMS [12] imposed additional requirements on the protocol characteristics. It was decided that 3GPP consortium will not undertake the endeavour of introducing the necessary extensions by itself, but rather leave it in the hands of IETF organisation. As a consequence the expansion set is still undefined, but there are guidelines and proposals **Error! Reference source not found.** on how to handle a number of identified issues.

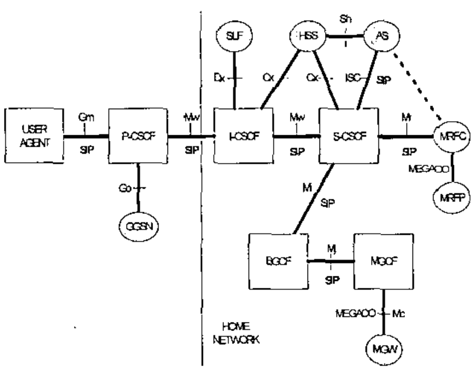

Figure 3. Internet Multimedia Subsystem

Figure 3 presents the structure of the Internet Multimedia subsystem and lists network interfaces that use SIP as a standardised protocol solution. Note the use of the Media Gateway Control Function, a bridge to legacy PSTN networks.

X. CONCLUSIONS

SIP has been proposed as a solution for numerous problems, including mobility, configuration and management, QoS control, call control, device control, third party call control. Clearly, not every problem can be solved by a SIP extension. More importantly, some problems that could be solved by a SIP extension probably shouldn't be. There are two broad criteria for determining whether a SIP extension is an appropriate solution to some problem. First, the problem should fit into the general purvey of SIP's solution space. Secondly, the solution must conform to the general SIP architectural model.

While the first criteria might seem obvious, numerous extensions to SIP have been proposed just because some function is needed in a device which also speaks SIP. The argument that is generally given is – "I'd rather implement one protocol than many". As an example, user agents, like all other IP hosts, need some way to obtain their IP address. This is generally done through DHCP. SIP's multicast registration mechanisms might supply an alternate way to obtain an IP address. This would eliminate the need for DHCP in clients. One must be careful when introducing such extensions, since protocols should be defined to provide specific functions, rather than being defined based on all communications requirements between a pair of devices. The first approach to protocol design yields modular protocols with broad application. It also facilitates extensibility and growth; single protocols can be removed or changed without affecting the entire system. This approach to protocol engineering mirrors object oriented software engineering.

Second criteria, that the extension must conform to the general SIP architectural model, ensures that the protocol remains manageable and broadly applicable.

As SIP is gaining increasing acceptance in numerous applications of next generation networks, its future seems to be bright. We expect its position will continue to strengthen as it becomes one of the pillars of the new age in communications.

REFERENCES

[1] IETF RFC 3261: SIP: Session Initiation Protocol, June 2002

[2] IETF RFC 3427: Change process for the Session Initiation Protocol (SIP), December 2002

[3] IETF RFC 3372: Session Initiation Protocol for Telephones (SIP-T): Context and Architectures, September 2002

[4] IETF RFC 3262: Reliability of Provisional Responses in the Session Initiation Protocol, June 2002

[5] IETF RFC 3312: Integration of Resource Management and Session Initiation Protocol, October 2002

[6] IETF RFC 3311: The Session Initiation Protocol (SIP) UPDATE method, October 2002

[7] IETF draft: Session Timers in the Session Initiation Protocol (SIP), November 2002

[8] IETF draft: The SIP Refer Method, November 2002

[9] IETF draft: Session Initiation Protocol Call Control – Transfer, October 2002

[10] IETF RFC 3265: Session Initiation Protocol (SIP)-Specific Event Notification, June 2002

[11] IETF RFC 3428: Session initiation protocol extension for instant messaging, September 2002

[12] 3GPP Technical Specification TS 23.228: IP Multimedia Subsystem, Stage 2, Release 6, January 2003

[13] IETF draft: 3GPP Release 5 Requirements on the Session Initiation Protocol, October 2002