QoS Framework for SIP Signalling

Alexander A. Kist and Richard J. Harris
RMIT University Melbourne
BOX 2476V, Victoria 3001, Australia
Email: kist@ieee.org, richard@catt.rmit.edu.au

ABSTRACT
The Session Initiation Protocol (SIP) is widely accepted as the IETF alternative the ITU-T H.323 teleconferencing protocol to enable call and media session management and control. It is also used in carrier grade environments, such as the IP Multimedia Subsystem (IMS) of the 3rd Generation Partnership Project (3GPP) in emerging Universal Mobile Telecommunications System (UMTS) networks.

This paper proposes a framework that enables the Quality of Service (QoS) provisioning for signalling messages. The contributions are twofold: Firstly, it defines an overall framework to enable QoS provisioning for SIP signalling messages in carrier grade networks. Secondly, it outlines existing work that fits into the framework and identifies areas that require further investigation to implement the concepts in carrier grade networks.

KEY WORDS
Session Initiation Protocol (SIP), Quality of Service (QoS), Signalling, IP Multimedia Subsystem (IMS), Universal Mobile Telecommunications System (UMTS)

1 Introduction

Recent years have seen the IETF's *Session Initiation Protocol* (SIP) (RFC 3261 [1]) become the premier protocol choice for user location, session setup and session management tasks in IP environments. For example, the *3rd Generation Partnership Project* (3GPP), which is a global initiative to develop standards and specifications for next generation *Universal Mobile Telecommunications System* (UMTS) networks, uses SIP as the signalling protocol for its *IP Multimedia Subsystem (IMS)* in Release 5 (3GPP Technical Specification 23.228 (R5) [2], 24.228 (R5) [3], 24.229 (R5) [4]). Other examples include, but are not limited to, the use of SIP for air traffic control applications [5] and in IPv6 environments [6].

Where *Quality of Service* (QoS) issues for media transport in IP networks has been in the focus of the research community for many years, QoS provisioning specifically for signalling messages has received less attention. Signalling message delay and session initiation delay was addressed by Eyers [7] and Curcio [8]. Earlier work [9] outlined the need for QoS considerations on the SIP layer to provide equivalent telephony services, in particular, if SIP is to be used in 3GPP carrier-grade networks. This need is mainly based on two aspects: signalling network resources are shared with other services and guarantees to customers are only believable if service levels are defined.

This paper introduces an overall framework that enables QoS provisioning for SIP signalling traffic. It consists of several parts, including: *Virtual SIP Links* (VSLs), which allow the definition of virtual SIP networks, methodologies to calculate the size of signalling flows, dynamic resource allocation schemes and SIP routing methods.

Note that the IMS is used as an example network in this paper, but the concepts can be used in any environment that requires QoS SIP signalling. 3GPP uses it's own notation and introduces a number of SIP proxy servers called *Call Session Control Function* (CSCF). Commercial service providers require these servers to control session signalling message flows and enable authentication, billing, service provisioning, etc. *Proxy* CSCFs (P-CSCFs) are the network entry points for the *User Equipment* (UE). *Serving CSCFs* (S-CSCFs) hold a copy of the user profile, record session state information and provide higher level session handling functions. *Interrogating CSCFs* (I-CSCFs) are network entry points for terminating sessions and decide the message routing to S-CSCFs.

The contributions of this work are twofold: Firstly, it outlines the motivation and it defines an overall framework that enables QoS provisioning for SIP signalling messages in carrier grade networks. Secondly, it summarizes existing work that fits into the framework and identifies areas that require further investigation to use the concepts in carrier grade networks. Both tasks aid the ultimate goal of QoS provisioning for SIP service users in carrier grade networks.

The paper is organised as follows: It addresses the SIP signalling layer first; and then introduces the *Virtual SIP Overlay Network in Section* (VSON) 2. Section 3 summarises the framework and how the different concepts interact. Virtual SIP links are part of the VSON definition and are discussed in Section

4. VSLs can use Dynamic Resource Allocation (DRA) which is explained in Section 5. VSLs and DRA rely on traffic estimation and flow analysis which is discussed in Section 6. Messages on the VSON can take alternative paths; SIP message routing is discussed in Section 7. Section 8 outlines areas that require further attention.

2 SIP Layer Abstraction

In existing SIP signalling configurations, IP networks provide transport service for SIP messages. Any node that is connected to the IP network and has an IP address, is globally routable. In this situation, all nodes are logically fully meshed. The same is true for SIP signalling nodes that are connected to IP networks. An example of such a network is shown in Figure 1. It depicts the Transport/Network Layer and the SIP/Application Layer. The cloud symbolises the underlying transport network, and the nodes symbolise SIP servers.

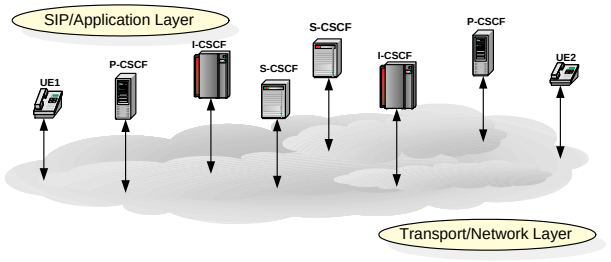

Figure 1. SIP and Network Layer

The IP resources of general-purpose transport networks are also used by services other than signalling, viz: This network configuration provides no dedicated signalling traffic resources. To define QoS levels solely for SIP, the SIP layer has to be separated from the transport layer. QoS measures on the SIP layer should be uncoupled from used transport technologies. On this SIP layer, service levels can be defined and an integrated QoS service concept can be developed. Figure 2 depicts such a transport independent virtual SIP overlay network.

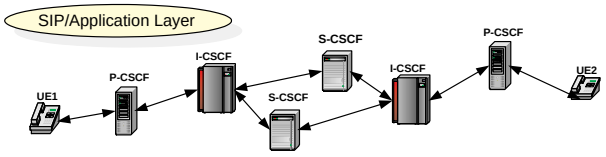

Figure 2. VSON Layer

The virtual network is reduced to well-defined links and nodes. Relevant issues of the underlying network have to be mapped onto this layer. This includes, but is not limited to, delays and bit errors. The VSON defines a signalling environment that enables the guarantee of service levels. Known methodologies can be applied and new strategies can be developed for this virtual overlay network.

The following sections outline issues that are relevant to the VSON. These include the definition of the connections between SIP nodes, possible resource allocation strategies for this link, signalling traffic management and SIP message routing.

3 Framework

The concept to provide QoS for SIP signalling consists of several models and methods that interact with each other. The overall goal is to define a virtual SIP overlay network with service guarantees as well as predictable and accountable behaviour. The relevant areas include: the SIP nodes, the connection between the nodes and the message routing in between the nodes. SIP nodes are server implementations that are similar to other existing Internet services; the connections and the routing are specific to the VSON.

Figure 3 depicts the different blocks of the framework and their interaction. The VSON (4) consists of SIP nodes and *Virtual SIP Links* (2). The VSL are dimensioned and set up by the *Dynamic Resource Allocation* DRA. The traffic model (1) is used by DRA and VSL to dimension these resources. Since the VSON defines it's own virtual network, routing strategies (5) are required on this level. To be able to use such

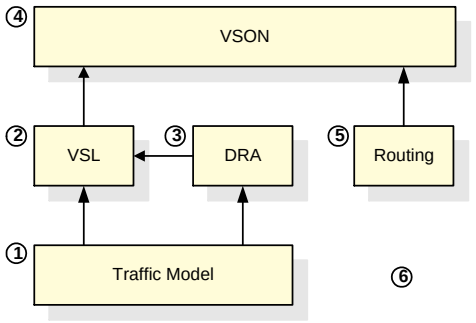

Figure 3. SIP QoS Framework

a scheme in an environment where signalling traffic and other services share network resources, it is important to guarantee QoS. To enable QoS on the application layer, the transport bearers have to provide QoS methodologies to protect traffic requirements for different services. The service level agreements, which define these needs, should be of a dynamic and adaptable nature. The detail of the *Service Level Agreement* (SLA) assignment process and implementation is not in the scope of this work.

4 Virtual SIP Links

Virtual SIP links are the abstraction of connections between SIP nodes. VSLs are logically located on the SIP layer and connect two SIP nodes. As outlined earlier, this is possible since the full connectivity provided by the IP network is not necessary to fulfil the functional requirements of VSONs.

The VSL definition is motivated by three main reasons: To define required resources, to enable traffic calculations and to enhance the performance of VSONs. To be able to introduce a quality aspect to VSONs, comparable and predictable parameters are required to classify the SIP signalling traffic. A well known method of classifying traffic is the *Leaky Bucket* (LB) methodology which is used to define VSLs. VSLs do not only define the flow size, but also limit the maximum queueing delay. Traffic models require the input of message loss probabilities. If resources are limited and classified by LBs, message loss boundaries can be estimated and traffic flows can be calculated [10].

VSLs have no physical equivalent. They are logical origin-destination associations with distinct traffic requirements. VSLs are defined by their *Traffic Specifications* (TSpec), i.e. the mean rate, the peak rate, the minimum policed message size, the burst size and the constraints of the message loss probability. These traffic specifications define a leaky bucket which enables traffic policing and shaping. Thus, VSLs implement a de facto admission control for signalling traffic. VSLs make connections comparable and accountable and they define clear interface points in the network.

VSLs are dimensioned for a maximum number of messages and a defined maximum message loss probability. Models that are addressed in Section 6 aid in the calculation of the LB parameters. Once the mean flow size is known, the required peak rate and buffer size can be calculated to comply with a minimum message loss probability. All defining VSL parameters are known. To deploy a VSL these parameters have to be accepted by the transport network. The transport network requires sufficient QoS technologies and has to be able to provide QoS resources.

VSLs require no additional hardware. Software modules in SIP servers can implement their functionality. In principle the use of VSLs for connections between SIP nodes is arbitrary. Connections with and without defined QoS conditions can coexist. If end-to-end QoS guarantees are required, all connections have to support VSLs. The mechanism of dimensioning and deploying can be executed on the fly by *Dynamic Resource Allocation* (DRA).

5 Dynamic Resource Allocation

The aim of dynamic resource allocation is twofold. Firstly, it is a methodology to enable the QoS provisioning for the virtual SIP signalling network. Secondly, it achieves the dimensioning automatically on the fly. DRA uses capabilities that mixed services IP transport networks provide. If the transport network supports dynamic service level agreements, nodes have the ability to reserve more/less bandwidth from the transport network to adapt to new requirements.

Dynamic resource allocation requires several supporting technologies. The underlying transport network has to support dynamic SLAs i.e. the ability to negotiate the resources with the network. It also requires that VSLs be used on the SIP layer. To keep the perceived QoS constant, the DRA process dimensions VSLs for SIP traffic. To be able to dimension VSLs appropriately, knowledge about network traffic flows is required. In such a context, the dynamic resource allocation methodology allows the automated configuration of resources and ensures QoS for signalling. DRA works as follows [11]:

DRA observes connections of SIP servers and adapt the resources as required. To achieve such functionality several sub functions are required: Data about the VSL state are collected as message arrivals per time unit and average message size. If these values cross upper or lower threshold limits, the scheme tries to increase or decrease the resource subscription. To ensure that actual trends are detected and not random fluctuation, the two thresholds need to have a minimum distance. The distance is estimated by known statistical methods.

The actual size of the resources that are subscribed can be calculated by using a flow model and methods for VSL dimensioning. The resource subscription is executed if thresholds are crossed. Additional resources are requested or resources are returned to the transport network. If the transport network cannot fulfil an increase-request, the resource subscription remains unchanged.

Both dynamic resource allocation and dynamic routing (Section 7) are processes that are executed on the fly, so it is important that they operate on different time scales. Frequent fluctuations in user/message numbers can be handled by dynamic routing schemes, long-term traffic shifts are handled by DRA.

DRA can yield considerable resource savings, since unused resources are returned to the transport network. This is in particular useful if the resources are billed on SLA bases.

6 SIP Flows and Traffic Estimation

To be able to use the methodology of VSLs and to get a fundamental understanding of traffic flows in SIP overlay networks, it is necessary to have models that describe SIP signalling message traffic. These models investigate the influence of message loss on the transmission media on SIP traffic and account for the fact

that SIP messages grow in size when they pass through the network.

A SIP fluid flow model does not take different SIP messages into account, but investigates how their traffic aggregates can be used to analyse flows. It allows also the calculation of the minimum bit error requirements for transport network connections. Such a flow model was introduced in [12]. A simplified version of this model relies on two parameters: the number of messages that are sent and the average size of these messages. Both parameters are locally available in nodes. Thus it is possible to estimate the mean flow size. The remainder of this section outlines a simplified model that allows flow size estimation between two SIP nodes.

Firstly the calculation of the message size is shown. The signalling traffic between two nodes depends on the number of messages that are sent between these nodes, including the number of messages that had to be resent. Figure 4 depicts an example-connection between two nodes a and b. The mean size of messages leaving node a is denoted by $\overline{m}_a$, the number of original messages sent on the connection under the assumption of zero loss is denoted by $n_{a,b}$. The connection has a message loss probability of $P_E(a,b)$. The number of messages that were dropped and resent on connection a,b and beyond are denoted by $d_a^{a,b}$ and the number of messages that were dropped beyond b and traversed a,b is denoted by $d_b^{a,b}$. Depending on

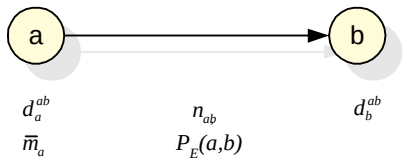

Figure 4. Example Connection

the local knowledge, the flow on link a,b can either be calculated for node a by Equation (1) or for node b by Equation (2)

$$FL_{a,b} = \overline{m_a} \cdot \left(n_{a,b} + d_a^{a,b}\right) \tag{1}$$

In the case of node a, $d_a^{a,b}$ is known and in the case of node b, $d_b^{a,b}$ is known.

$$FL_{a,b} = \overline{m_a} \cdot n_{a,b} \cdot \left(1 + \frac{P_E(a,b) + d_b^{a,b}}{1 - P_E(a,b)}\right) \tag{2}$$

The estimation of the average message size $\overline{m}_a$ is discussed next.

SIP messages consist of a message header and a message body. The message body is usually used for the session description which uses the *Session Description Protocol* (SDP). SIP messages appear with and without a message body, and therefore vary considerably in size.

State full servers that are traversed by requests are recorded in the "via" header field and in some cases in the "route-record" field of the message. This causes the SIP messages to grow while they pass through the network. Since the observations here are local the message size change has no impact in this context. The current message size estimate is of interest. A local node can calculate a recursive mean of all processed messages online. Equation (3) shows the calculation.

$$\overline{m}(t+1) = \overline{m}(t) + \frac{M(t+1) - \overline{m}(t)}{n} \tag{3}$$

The averaging count is denoted by n, i.e. the number of samples that are considered, the current message size is M and $\overline{m}$ is average. This value is simpler to calculate than the moving average. It does not require memory for the past n values. More recent measurements have a larger impact than older measurements.

This simple model allows the traffic estimation between two SIP nodes by dynamically measuring straightforward parameters, i.e. message size, message number and the number of dropped messages. The estimates help the dimensioning of signalling resources between these nodes as outlined in Section 5.

7 SIP Message Routing

Messages on the VSON layer can take alternative routes to their destination, and therefore message routing is required. Most SIP message routing decisions are made during the registration of users, since intermediate nodes record registration state and/or session state information. Once associations are formed they are persistent for subsequent requests.

In principle, message routing depends on two factors: Firstly the connectivity on the VSON layer, i.e. VSLs are defined between nodes and secondly, it depends on operator-specific policies and routing strategies. The first depends in general on functional specifications and the network structure, e.g. for 3GPP, the defined network entry point for UEs are P-CSCFs, the later can be specific to operators or single networks.

SIP message routing can be divided into two areas: Routing decisions that resemble load balancing schemes/server assignment problems; and routing problems that can be solved by shortest path like routing methodologies. For 3GPP, the first are relevant for routes from I-CSCFs to S-CSCFs and the later concern routing decisions that are required for routes from UEs/P-CSCFs/S-CSCFs to I-CSCFs. First load balancing and overflow routing is addressed.

The challenge of efficient message routing is similar to existing server/resource allocation problems which include server load balancing, hash routing and web caching schemes. In recent years, these areas have

been major research targets. The *SIP Message Overflow Routing Scheme* (SMORS) is introduced [13]. It targets efficient message routing for high volume SIP servers, in the 3GPP IMSs context and it specifically addresses message routing between I-CSCFs and S-CSCFs.

SMORS assigns users/messages to servers on the basis of generic routing information. If a threshold is reached, new users/messages are overflown to additional servers. The principal concept of overflow routing is not new and has been extensively studied in the context of *Public Switched Telephone Networks* (PSTNs). SMORS has several benefits: It is fast and requires little processing for subsequent message routing, it has load balancing, backup provisioning and overflow routing capabilities. One major advantage of this scheme is that the user-to-S-CSCF assignment is flexible. S-CSCFs can be assigned, by user priority, service subscription or alphabetically by user name. The assignments do not have to be calculated at runtime; these associations can be pre-calculated.

Shortest path-like routing protocols and methodologies in the IP routing context are well-understood and remain a continuing focus of the research community. Shortest path routing algorithms find paths between an *Origin-Destination Node Pair* (OD-pair) that satisfy a minimal optimisation metric. Traditional routing protocols such as *Open Shortest Path First* (OSPF) use scalar metrics to optimise the paths. Many possible alternative metrics are known, but the inverse of the capacity is most commonly used in IP networks. Factors that have an impact on the message routing on the VSL layer are the hierarchical and logical setup defined by the network structure, the availability of resources in nodes to handle additional requests, resource availability for additional signalling traffic on the transmission media, the close proximity of nodes and the quality on the connection. The routing process should consider these factors.

[14] proposed a multidimensional metric. During defined stages in the network operation one metric within the multidimensional metric will dominate the others. If, for example, two paths have the same number of available users and similar delays of acceptable length, the path with the maximum reliability should be selected. On the other hand, core network connections are likely to have similar reliabilities. In this case, the path selection is mainly based on delay which is a measure for distance in this context. If the number of available users on a certain path gets smaller, routing has to take this into account. It is important to note that in the combined metric, one metric dominates the others at a given time, the metric is not a simple weighted average calculation.

8 Further Work

This paper mentioned a number of models and methodologies that can work together to provide QoS for SIP signalling messages and improve the service experience of users. This section outlines areas for further research directions.

8.1 Transport Protocol

One issue that has a considerable impact on QoS issues is the transport protocol that is used. In principle, the SIP protocol is transport protocol independent. The drawback of the default transport protocol UDP is its unreliability and the disadvantage of TCP is the protocols' long connection set up time. UDPs unreliability can cause problems in SIP overlay networks with the high number of intermediate links and SIP's end-to-end reliability [9].

The *Stream Control Transmission Protocol* (SCTP) [15] could be an alternative, with considerable advantages for the session setup times and throughput. Internet Draft [16] discusses the SCTP protocol in conjunction with its use in SIP. It suggests that most of the benefits of SCTP occur under loss conditions. The situation in QoS signalling domains is different from the situation in general Internet environments, therefore further study is required to investigate the impact of transport protocols on SIP QoS.

8.2 QoS Network

VSLs assume that the transport network can provide QoS resources to signalling traffic. Further study should investigate the interaction between VSLs and the transport network, and possible implications. Furthermore, VSLs use dynamic SLAs. DRA assumes that these exist and can be used. Further study has to focus on the exact implementation of these dynamic SLAs, which may include SLA negotiation and further issues.

8.3 Generic Overlay Networks

Recent years have seen many new peer-to-peer services emerge, many of which define their own overlay network. Overlay networks are also a way to provide advanced services via the public Internet that are not provided by the network. In future IP networks, these services will be able to use QoS functionalities as well. Results such as the VSL principle, dynamic resource allocation and routing might be applicable in this context as well.

These efforts can also include dislocated corporate computer networks which use IP resources to connect various sites, as well as applications that define their own overlay network and intend to apply QoS

principles. Obviously, this outlines a wide research area with a number of interesting research problems to solve.

9 Conclusions

This paper introduced a framework that enables QoS provisioning for SIP signalling service in advanced carrier grade networks. It outlined a number of technologies and schemes that can aid the QoS effort. Future work has to investigate the practicality of these models in live networks. The major aim of this paper was to introduce some of these issues and possible solutions. These issues might gain importance, as more and more commercial services will use SIP as a session control and management. Providing QoS for SIP signalling is a core issue in providing QoS for consumers of future emerging (wireless) networks.

Acknowledgements

The authors would like to thank Ericsson AsiaPacificLab Australia and the Australian Telecommunications Cooperative Research Centre (ATcrc) for their financial support for this work.

References

[1] J.Rosenberg, H.Schulzrinne, G.Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. *SIP: Session Initiation Protocol*. IETF, June 2002. RFC 3261 (Obsoletes RFC 2543).

[2] 3rd Generation Partnership Project. *IP Multimedia (IM) Subsystem - Stage 2 (Release 5)*, January 2004. 3G TR 21.905 V5.11.0.

[3] 3rd Generation Partnership Project. *Signalling flows for the IP multimedia call control based on SIP and SDP - Stage 3 (Release 5)*, December 2003. 3GPP TR 24.228 V5.7.0.

[4] 3rd Generation Partnership Project. *Internet Protocol (IP) multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP) - Stage 3 (Release 5)*, January 2004. 3GPP TR 24.229 V5.7.0.

[5] K.Darilion, W.Kampichler, and K.Gschka. Event-based radio communication signalling using the session initiation protocol. *In The 11th IEEE International Conference on Networks (ICON 2003), Sydney, Australia*, September 2003.

[6] J.Fiedler D.Sisalem. SIP and IPv6: Why and how? *International Symposium on Applications and the Internet (SAINT2003), Orlando, USA*, Janury 2003.

[7] T.Eyers and H.Schulzrinne. Predicting Internet Telephony Call Setup Delay. In *In IPTel 2000 (First IP Telephony Workshop)*, April 2000.

[8] I.D.D. Curcio and M. Lundan. SIP call setup delay in 3G networks. *In The Seventh IEEE Symposium on Computers and Communications (ISCC'02), Taormina/Giardini Naxos, Italy*, July 2002.

[9] A.A. Kist and R.J. Harris. SIP Signalling Delay in 3GPP. *In Proceedings of Sixth International Symposium on Communications Interworking of IFIP - Interworking 2002, Perth Australia, October 13-16*, October 2002.

[10] A.A. Kist and R.J. Harris. Using virtual SIP links to enable QoS for signalling. *In The 11th IEEE International Conference on Networks (ICON 2003), Sydney, Australia*, September 2003.

[11] A.A. Kist and R.J. Harris. Dynamic resource allocation in 3GPP SIP overlay networks. *In Fourth International Conference on Information, Communications & Signal Processing and Fourth Pacific-Rim Conference on Multimedia (ICICS-PCM 2003), Singapore*, December 2003.

[12] A.A. Kist and R.J. Harris. A simple model for calculating SIP signalling flows in 3GPP networks. *In Second IFIP-TC6 Networking Conference 2002, Pisa, Italy May 19-24*, May 2002.

[13] A.A. Kist and R.J. Harris. SIP message overflow routing scheme (SMORS). *In 2003 Australian Telecommunications, Networks and Applications Conference (ATNAC), Melbourne, Australia*, December 2003.

[14] A.A. Kist and R.J. Harris. SIP routing methodologies in 3GPP. *In First International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs '03), Ilkley, West Yorkshire, U.K*, July 2003.

[15] R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, and V. Paxson. *Stream Control Transmission Protocol*. IETF, October 2000. RFC 2960.

[16] J. Rosenberg, H. Schulzrinne, and G. Camarillo. *The Stream Control Transmission Protocol as a Transport for the Session Initiation Protocol*. IETF, November 2003. Internet Draft <draft-ietf-sip-sctp-04.txt> (work in progress).