Transport Service for Session Initiation Protocol in SIP-T Scenarios

Luigi Alcuri, Michele L. Fasciana
DIE - Dipartimento di Ingegneria Elettrica
Università degli studi di Palermo
Palermo, Italia
E-mail: {luigi.alcuri; michele.fasciana}@tti.unipa.it

Abstract – **The traditional telephone service has undergone deep changes in the years concerning the techniques of commutation and signalling. Today Public Switching Telecom Networks (PSTNs) and Private Integrated Services Networks (PINSs) are strongly directed to adopt IP-based backbone. If it's possible to transfer the voice on packet switching network, in such reliable way, using Voice over IP (VoIP) techniques, there is not still a valid alternative to the signalling systems, reliable and unfailing, like Signalling System n. 7 (SS7). Session Initiation Protocol (SIP) seems to be the best candidate. SIP is an application layer signalling protocol and the problem remains still opened for what it concerns the transport service to use. This paper discusses about pros and cons of the possible candidates. To evaluate performances we have achieved some simulations, using a new module for Network Simulator 2 [1].**

Keywords: SIP,SIP-T, Transport Services, NS2

I. Introduction

The Session Initiation Protocol (SIP) [2] is an application layer protocol for creating, modifying and terminating sessions. SIP is designed in a modular way so that does not depend on the session type established and on the lower-layer used transport protocol. Its modularity is one of the most important strengths of SIP protocol. It makes SIP flexible and easy to extend with new features.

The SIGTRAN working group developed the Stream Control Transmission Protocol (SCTP) [3], that was first intended to transport telephony signalling over an unreliable network such as an IP network.

The SIP specification describes how the protocol operates over TCP and over UDP, but several papers [4] have broadly discussed the advantages of SCTP in comparison to TCP as reliable transport service in call signalling applications.

However, considered that SIP provides some mechanisms of reliability [5], the choice of a reliable or less transport service is still an open issue.

This paper discusses advantages from the choice of SCTP for SIP in comparison to UDP in the different scenarios contemplated in a set of documents known as SIP for Telephones (SIP-T) [6]. To appraise the performances we have effected some simulation using a new module for NS2 [1], called MGC-module, able to reproduce the behaviour of a decomposed gateway (SG/MGC/MG system).

The remainder of this paper is organized as follows. Section 2 describes SIP-T basic operations, then pros and cons of each transport protocol are analyzed. Section 3 describes shortly how MGC-module for NS2 works and what hypothesis we have adopted in the simulations. Section 4 shows the achieved results. Finally section 5 outlines some conclusions.

II. Background

A. Signalling Protocols

Multimedia applications use SIP like application-layer control protocol to establish, to modify and to terminate their sessions or calls. These applications include Internet telephony and similar. SIP is one of the key protocols used to implement Voice over IP (VoIP).

ISUP is a level 4 protocol used in SS7 networks. This protocol is used for controlling telephone calls, maintenance of the network (blocking and resetting circuits etc.) and providing Intelligent Network services (IN).

QSIG is a signalling protocol that operates between Private Integrated Services eXchanges (PINX) within a Private Integrated Services Network (PISN).

Nowadays many PSTN and PINS backbones are IP-based and use standardized techniques for forwarding user payload (voice samples), but there is not a reliable standard for signalling transfer[1]. In this paper we refer to Next Generation Public Telephone Network (NG-PTN) and Next Generation Virtual Private Network (NG-VPN) relating to respectively PSTN and PISN .

B. SIP for Telephones (SIP-T)

SIP-T [6] is a set of mechanisms for interfacing traditional telephone signalling with SIP. The purpose of SIP-T is to provide protocol translation and feature transparency across points of ISUP-SIP or QSIG-SIP interconnection.

The SIP-T effort provides a framework for the integration of legacy telephone signalling into SIP messages.

To reach the goal, two techniques, known as *encapsulation* and *translation*, are used.

---

[1] In this case PINS are called *Virtual Private Network (VPN)*.

Encapsulation

SIP protocol, by *encapsulation*, can carry any type of information (like an e-mail application), in every message between MGCs, during the setup of Real Time application. Gateways can take advantage of the full range of services afforded by the existing telephone network when placing calls from PSTN to PSTN across a SIP network. The procedure by which SIP messages are used to transport ISUP (or QSIG) payloads from gateway to gateway is called SIG-*Encapsulation*. In the next chapter, we show two simulated scenarios with ISUP-Encapsulation (NG-PTN) and QSIG-Encapsulation (in NG-VPN).

Translation

Translation encompasses all aspects of signalling protocol conversion between SIP and ISUP [8] or SIP and QSIG [9]. There are two components to the problem of translation:

- *ISUP/SIP (QSIG/SIP) message mapping*: this describes a mapping between ISUP (QSIG) and SIP at the message level. In SIP-T deployments gateways are entrusted with the task of generating a specific ISUP message for each SIP message received and vice versa.
- *ISUP/SIP (QSIG/SIP) header mapping*: a SIP request which is used to set up a telephone call should contain information that enables it to be appropriately routed to its destination by proxy servers in the SIP network - for example, the telephone number dialed by the originating user.

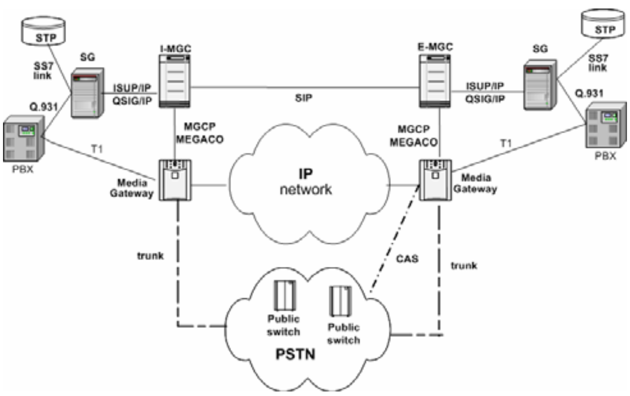

Figure 1. SIP-T framework

Figure 1 shows a network configuration using SIP-T. In this example Media Gateways (MGs) are connected to the Switched Circuit Network (SCN) via SS7 trunks, ISDN trunks, and CAS trunks. Media Gateway Controllers (MGSs) manage MG by MGCP or MeGaCo Protocol. The Ingress Media Gateway Controller (I-MGC) may receive a signalling call from Signalling Gateway connected to SS7. The signalling information from these trunks must be processed by the I-MGC to establish the originating call part, and to determine the identity of the Egress Media Gateway Controller (E-MGC) required to complete the call. The I-MGC uses SIP to communicate the necessary information to the E-MGC to complete the call. The E-MGC is able to establish the terminating call part on any of the supported trunk types.

At the edge of the depicted network, an MGC converts the ISUP signals to SIP requests, and sends them to other MGCs. Although this figure describes only two MGCs, VoIP deployments have many such points of interconnection with the PSTN and VPN (usually to diversify among PSTN rate centers).

Figure 2 shows one of several call flows related SIP-T with voice calls originating and terminating in the PSTN (via gateways).

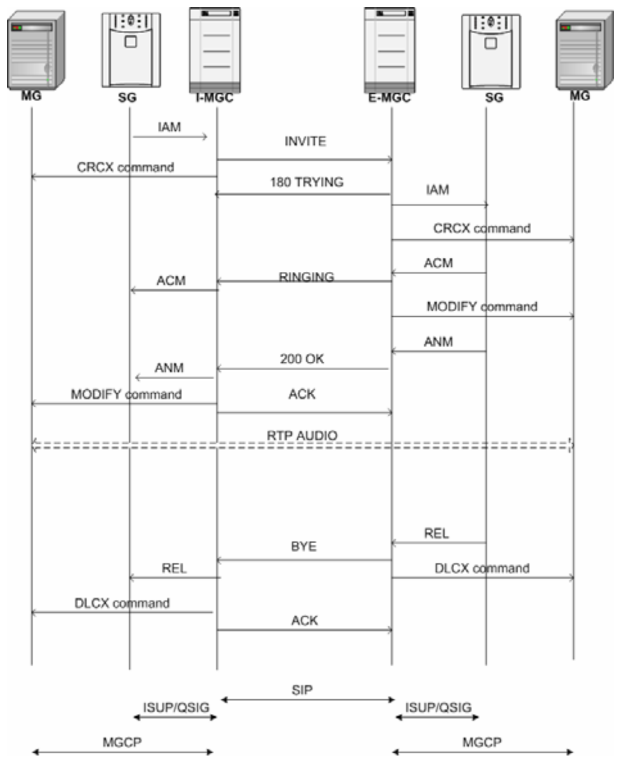

Figure 2. SIP-T call flow

C. *SIP and reliability*

SIP provides some mechanisms to give reliability to its transactions [5]. In particular, the INVITE Timer manages the retransmission of the INVITE messages. This occurs when no related TRYING or RINGING messages are received from the sender after a proper time (0.5 sec) from the instant of the forwarding INVITE.

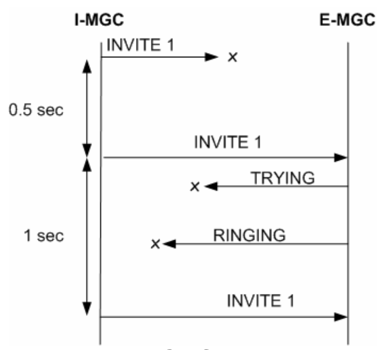

Figure 3. SIP timer

D. Transport Service for SIP

SIP is transport-independent, and can run over any reliable or unreliable message or stream transport. In recent times many discussions have been made about possibility to use the transport service provided by SCTP (Stream Control Transmission Protocol), deployed by SIGTRAN-WG [4].

SCTP has been designed as a new transport protocol for the Internet (or intranets), at the same layer as TCP and UDP. It has been designed with the transport of legacy SS7 signalling messages in mind. Many of the features designed to support transport of such signalling are also useful for the transport of SIP (the Session Initiation Protocol).

All the advantages that SCTP has over UDP regarding SIP transport are also shared by TCP. Below there is a list of the general advantages that a connection-oriented transport protocol such as TCP or SCTP has over a connection-less transport protocol such as UDP.

- **Fast Retransmit:** SCTP can quickly determine the loss of a packet, as a result of its usage of SACK and a mechanism which sends SACK messages faster than normal when losses are detected. The result is that losses of SIP messages can be detected much faster than when SIP is run over UDP (detection will take at least 500ms, if not more). Also TCP provides SACK and fast retransmit options. Over an existing connection, this results in faster call setup times under conditions of packet loss. This is probably the most significant advantage respecting SCTP for SIP transport.

- **Congestion Control:** SCTP maintains congestion control over the entire association. For SIP, this means that the aggregate rate of messages between two entities can be controlled. When SIP is run over TCP, the same advantages are afforded. However, when run over UDP, SIP provides less effective congestion control. This is due to the congestion state (measured in terms of the UDP retransmit interval) is computed on a transaction by transaction basis, rather than across all transactions. Congestion control performance is thus similar to opening N parallel TCP connections, as opposed to sending N messages over 1 TCP connection.

- **Transport layer fragmentation:** SCTP and TCP provide transport layer fragmentation. If a SIP message is larger than the MTU size, it is fragmented at the transport layer. When UDP is used, fragmentation occurs at the IP layer. IP fragmentation increases the likelihood of having packet losses and make it difficult (when not impossible) NAT and firewall crossing. This feature will become important if the size of SIP messages grows dramatically.

In comparison to TCP, SCTP offer some advantage relating to call signalling transmission.

- **Head of the Line**: SCTP is message based as opposed to TCP that is stream based. This allows SCTP to separate different signalling messages at the transport layer. TCP just understands bytes. Assembling received bytes to form signalling messages is performed at the application layer. Therefore, TCP always delivers an ordered stream of bytes to the application. On the other hand, SCTP can deliver signalling messages to the application as soon as they arrive (when using the unordered service). The loss of a signalling message does not affect the delivery of the remaining the messages. This avoids the head of line blocking problem in TCP, which occurs when multiple higher layer connections are multiplexed within a single TCP connection. A SIP transaction can be considered an application layer connection. Between proxies there are multiple transactions running. The loss of a message in one transaction should not adversely affect the ability of a different transaction to send a message. Thus, if SIP is run between entities with many transactions occurring in parallel, SCTP can provide better performance than SIP over TCP (but not SIP over UDP; however SIP over UDP is not ideal from a congestion control standpoint, see above).

- **Easier Parsing**: Another advantage of message based protocols such as SCTP and UDP over stream based protocols such as TCP is that they allow easier parsing of messages at the application layer. There is not need of establishing boundaries (typically using Content-Length headers) between different messages. However, this advantage is almost negligible.

- **Multihoming**: An SCTP connection can be associated with multiple IP addresses on the same host. Data is always sent over one of the addresses, but if it should become unreachable, data sent to one can migrate to a different address. This improves fault tolerance; network failures making one interface of the server unavailable do not prevent the service from continuing to operate. SIP servers are likely to have substantial fault tolerance requirements. It is worth nothng that because SIP is message-oriented, and not stream oriented, the existing SRV procedures defined in [1] can accomplish the same goal, even when SIP is run over TCP. In fact, SRV records allow the "connection" to fail over a separate host. Since SIP proxies can run statelessly, failover can be accomplished without data synchronization between the primary and its backups. Thus, the multihoming capabilities of SCTP provide marginal benefits.

TABLE I. TRANSPORT SERVICES: RELIABILITY

Reliability mechanism	UDP	TCP	SCTP
Fast Retransmit	N	Y	Y
Congestion Control	N	Y	Y
Transp. layer frag.	N	Y	Y
Head of the Line	N	N	Y
Easier Parsing	N	N	Y
Multihoming	N	N	Y

III. A TOOL FOR PERFORMANCE EVALUATION : MGC-MODULE

NS2 [1] is a network simulator, free and widely used in research ambit. Its core engine is written in C++ language and use Tcl (Tool Command Language) to describe several simulation scenarios. We have projected a new tool for NS2, able to simulate MGC behaviour in environments like that described in Figure 4.

MGC module provides to simulate a bidirectional dialogue between two SIP agents that send INVITE requests. Only when call setup is correctly performed, a new bidirectional voice session starts between two nodes. Then developed voice traffic depends on the number of call setup correctly performed.

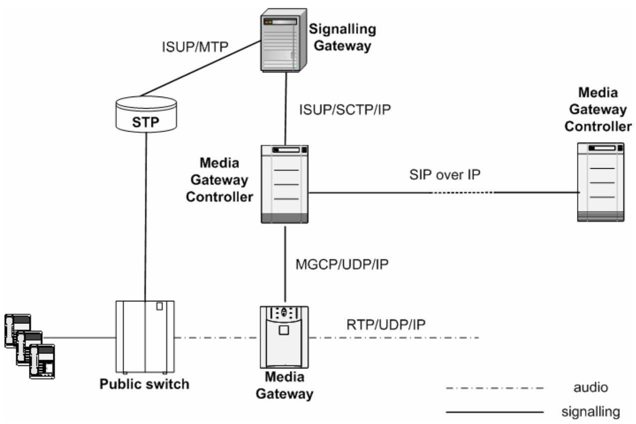

Figure 4. MGC module: reference scenario

A. *MGC-MGC dialogue : SIP*

The inter-arrival process between two consecutive sent INVITE messages is poissonian. According to call flow described in Figure 2, the TRYING message is sent instantly as soon as an INVITE message is received; timers manage the forwarding of SIP messages properly: RINGING timer (provides uniform delay) and 200OK timer (provides exponential delay). Finally BYE timer controls the stop of the single call (the time is stochastic too).

B. *MGC-MG dialogue :MGCP*

Between a SIP message and the other, MGC sends a set of commands to MG using MGCP protocol. Among the problems that the simulation of the scenery SIP-T introduces, there is that to synchronize the departure of the vocal calls (in the reality managed through MGCP) with the instant in which MGC agents receives the message 200OK; it is only in this moment, in fact, that all the fit procedures to the installation of the call (the information exchange through SDP, opening of the doors RTP and UDP in the Media Gateway, choice of the codecs, etc.) are dispatched.

NS2 provides some particular procedures able to modify the carrying out of the simulation (execution of the tcl script) directly from the rising C++ code. Several Tcl procedures (Insproc) are been inserted for managing voice sources during the simulations. In particular:

- Instproc *CONFIGURESESSION*:
  it refers to CRCX MGCP command
- Instproc *STARTCALL*:
  it refers to MDCX MGCP command
- Instproc *STOPCALL*:
  it refers to DLCX MGCP command

By using this procedures in the simulation script MGC module can manage autonomously set up, start and stop of the voice sources.

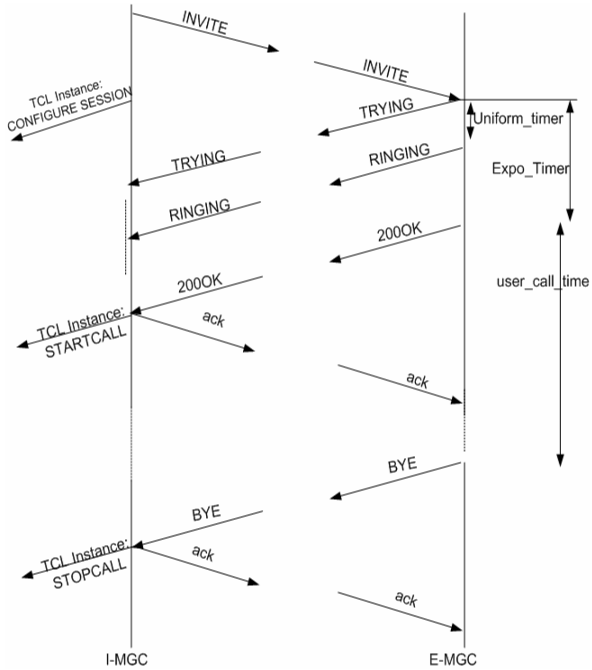

Figure 5. MGC Module: call flow

C. *MGC modules and Transport Services*

In real implementation, SIP protocol can enjoy of a transport both reliable that unreliable, then we have given the possibility to choose between UDP (connectionless service) and SCTP (reliable service).

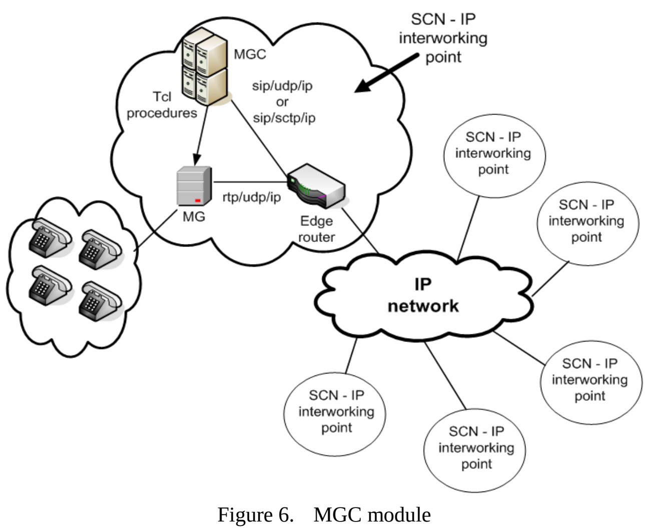

Figure 6. MGC module

IV. SIMULATING SIP-T

Using the MGC modules, we have conducted a set of suitable simulations to appraise the efficiency of the mechanisms of reliability foreseen by SIP protocol and the achieved performances using two different transport services.

A. Service Models: Best Effort and Expedited

The simulations have been conducted also foreseeing two possible behaviors as it regards the service offered by the network: *Best Effort* (BE) and Expedited *Forwarding – DiffServ* (EF-DS) for SIP messages. The first one is the default service in IP networks, the second guarantees best performances concerned to loss packet rate and latency. We think that *EF-DiffServ* can provide that essential quality of service for call signalling transfer in critical scenarios as NG-PTN. Figure 7 shows how EF-DS aware router works.

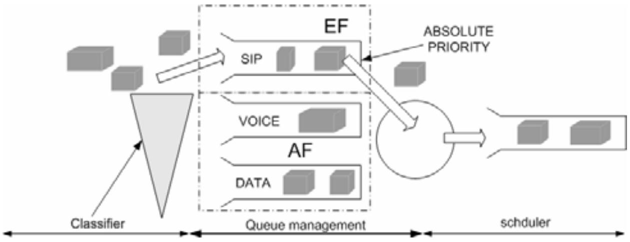

Figure 7. EF-DS: Reference model

B. General setting

- Voice sources
  The reference model for the voice sources it is that of Brady: exponential ON-OFF sources with average idle time of 1.3 sec and average burst time equal to 1 sec. For the coding we have hypothesized the use of the simple G.711 PCM.
- Http traffic
  To simulate WWW traffic we have used a simple script that provides exponential flow arrival and flow lengths drawn from an empirical distribution. A random number of packets is associated to each new flow, according to an empirical distribution. Here 4 different lengths have been considered: 50% of flows have 1 packets, 20% have 6 packets, 20% have 18 packets, the remaining 10% have 190 packets.
- FTP sources
  FTP sources are been inserted too for producing background traffic.

C. Measurements

As it regards the effected measurements, it has not been possible to individualize a reference standard for the definition of the call setup in SIP-T. For this we have chosen those parameters that are seemed more meaningful in this context. In particular:

- *Average Call setup*:
  Mean temporal interval between the dispatch of an INVITE and the receipt of the related TRYING or RINGING message.
- *INVITE mean delay*:
  Mean temporal interval between the dispatch from MGC and the receipt at other MGC of an INVITE message.
- *INVITE to TRYING mean delay*:
  Mean temporal interval between the dispatch of an INVITE and the receipt of the related TRYING message.

Case of EF-DS service (no packet loss occours), *average call setup* and *INVITE to TRYING mean delay* provide identical values. Other measurements have concerned the following loss percentages:

- *Loss call percentage*;
- *Loss INVITE percentage*;
- *Loss TRYING percentage* (related to the number of sent INVITE messages);
- *Loss RINGING percentage* (related to the number of sent INVITE messages).

D. NG-PTN scenario

In the simulation, described in figure 8, we have hypothesized a scenario where two MGCs constitute the access to the backbone IP of an NG-PTN. 11 call requests to the second come (load offered equal to the 82% of the central link with mean time of the single call equal to 100 sec). Using UDP protocol, we have considered two cases: activated SIP timer (UDP+T), disabled SIP timer. With SCTP the timer has been disabled.

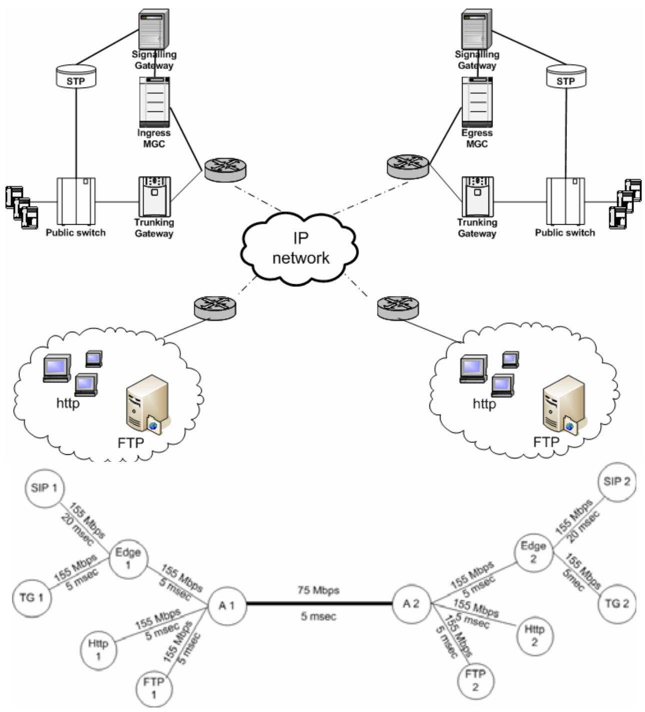

Figure 8. NG-PTN

TABLE II. NG-PTN (BE SERVICE)

	Theor.	UDP	UDP + TIMER	SCTP
call setup [msec]	110	110.974 +-0.03743	114.499 +- 2.048	158.975 +- 6.84
INV delay [msec]	55	55.3706 +- 0.0085	58.5155 +- 1.0512	55.1164 +- 0.0315
INV toTRY delay [msec]	110	110.611 +- 0.01415	113.635 +- 1.01191	158.975 +- 6.84
loss call (%)	-	0.5641% +- 0.0709%	-	-
loss INVITE (%)	-	0.5517% +- 0.0694%	-	-
loss TRY (%)	-	1.52271% +- 0.1%	1.31923% +- 0.7439%	-
loss RING (%)	-	1.52511% +- 0.0 789%	1.35158% +- 0.712134	-

TABLE III. NG-PTN (EF-DS SERVICE)

	Theor.	UDP	UDP + TIMER	SCTP
call setup [msec]	110	110.259 +- 0.0027	110.259 +- 0.00233	110.298 +- 0.0043
INV delay [msec]	55	55.1933 +- 0.001299	55.1929 +- 0.001522	55.2126 +- 0.0027
loss call (%)	-	-	-	-
loss INVITE (%)	-	-	-	-
loss TRY (%)	-	-	-	-
loss RING (%)	-	-	-	-

Analyzing the data shown in the two tables is possible to make some considerations. The use of a reliable transport service, as that offered by SCTP, would seem forced in the case in which the network doesn't provide some DIFFSERV mechanism. In fact, using simply the SIP-timer, it’s possible to guarantee the delivery of the message of INVITE and the departure of the set up, the same it cannot tell itself for the following messages, that could be lost. In effect, in case of UDP+T, we have obtained zero loss call: this is due to the fact that we have considered valid also the case in which
TRYING message is goes lost but related RINGING is received.

Opposite results we have achieved with the second hypothesis of service (EF-DS), whose clearly resulted as the typology of adopted transport is totally indifferent.

E. *NG-VPN scenario*

In the simulation described in figure 9 the couple of MGC constitutes the access to the backbone IP of two branches of a company whose network constitutes a NG-VPN, a virtual private network in which the telephone signalling travels through the protocol SIP. In this case we have decided to conduct a set of simulations modifying of time in time the offered load.

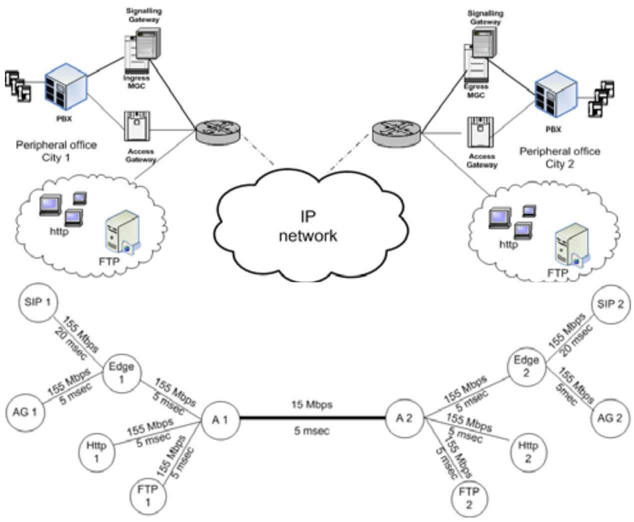

Figure 9. NG-VPN

Best Effort Service

Also in this scenario it is observed as the SIP timer results inefficient, in absence of further mechanisms of reliability. But, differently from NG-PTN scenario, the graphs related to the call setup show how, with percentages of data traffic (http and ftp) comparable with voice traffic, the SCTP protocol foresees problems, especially with elevated utilization factors of the bottleneck: beginning to work the mechanisms of reliability in presence of aggressive traffic as that http the transmission of the SIP messages comes notably, raising in such way the times of call setup. Using SIP timer INVITE mean delay increases. Also in this case with SIP timer we have not observed loss calls or loss INVITE, but TRYING messages have not undergone losses only with SCTP. Simulation results are shown in the following graphs.

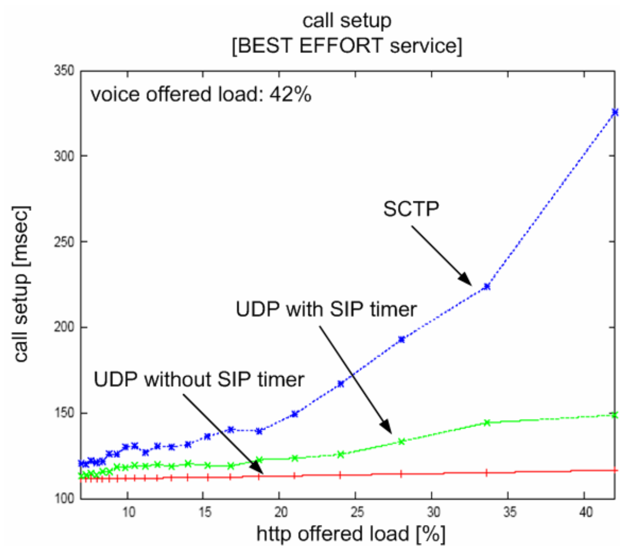

Figure 10a . call setup vs http load

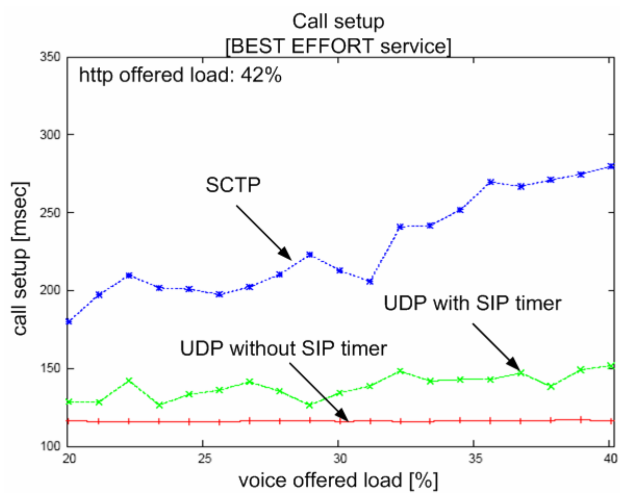

Figure 10b. call setup vs voice load

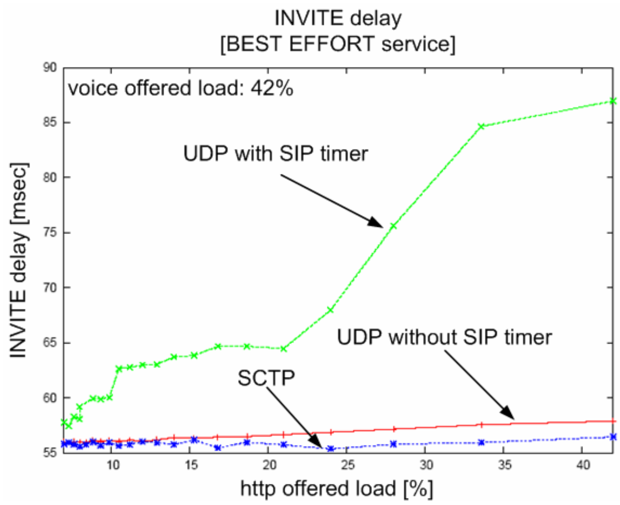

Figure 11a. INVITE mean delay vs http load

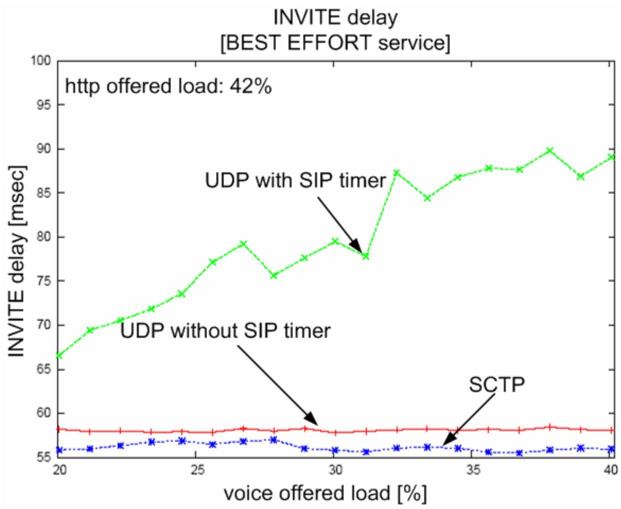

Figure 11b. INVITE mean delay vs voice load

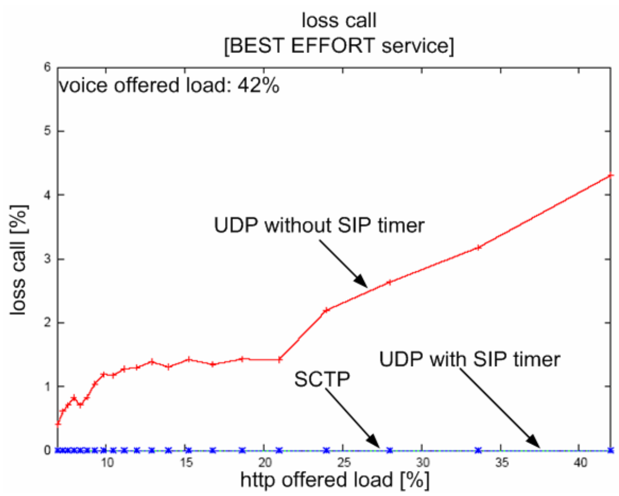

Figure 12a. loss call vs voice load

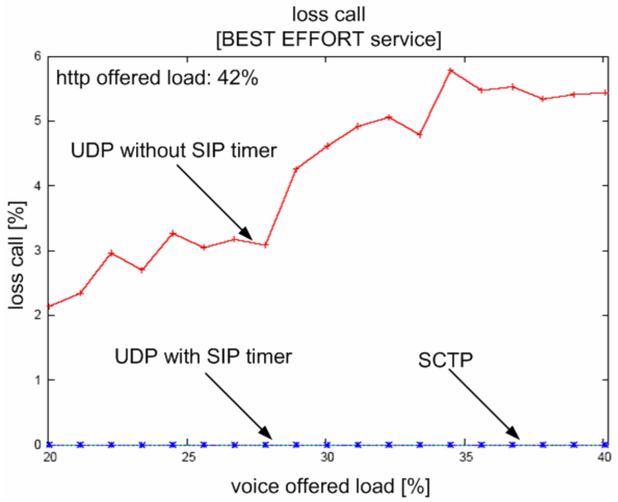

Figure 12b. loss call vs voice load

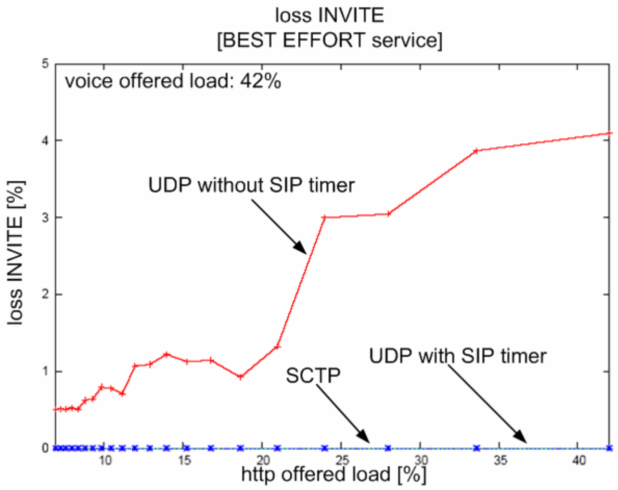

Figure 13a. loss INVITE vs http load

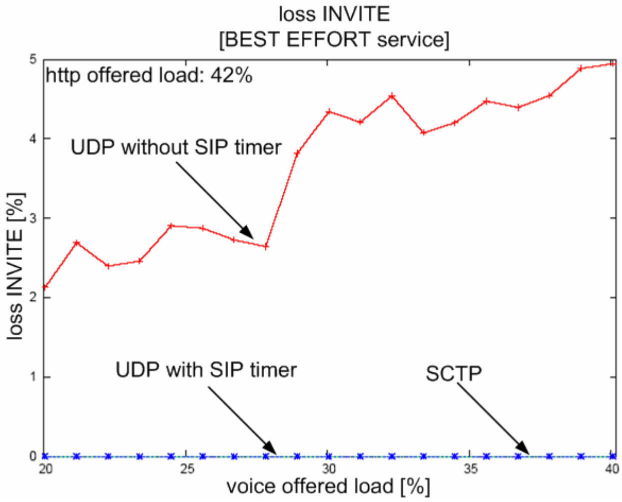

Figure 13b. loss INVITE vs voice load

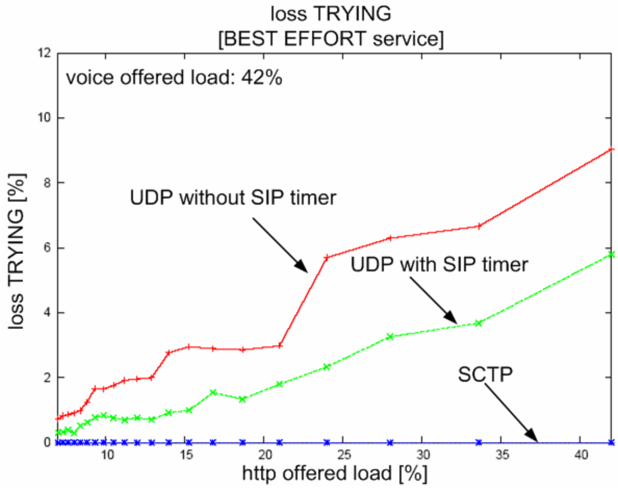

Figure 14a. loss TRYING vs voice load

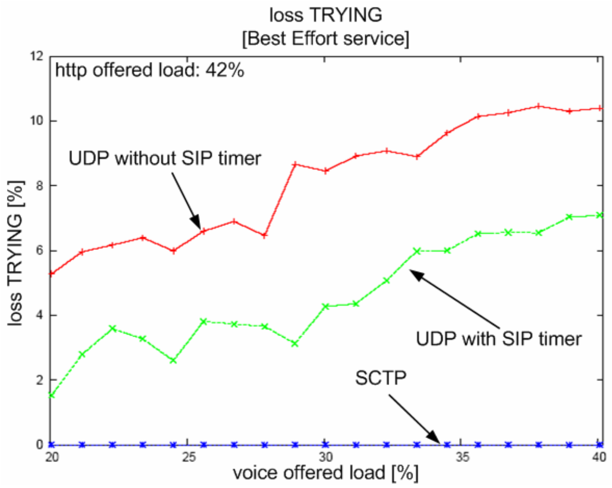

Figure 14b. loss TRYING vs voice load

A further result not shown in the graphs has been the following: using SCTP and activating at the same time the INVITE timer, the transmission of the SIP messages results practically jammed with a percentage of sent INVITE messages equal to less than a quarter of those expectations.

Expedited Forwarding - DiffServ

Also in this scenario, if the network is able to provide Expedited Forwarding service for SIP call signalling, the best performance are achieved. We have not find remarkable differences (see the graphs) among the use of UDP that SCTP: any loss in relief and least differences on the delays to debit themselves, probably, to the same simulator. Simulation results are shown in the following graphs.

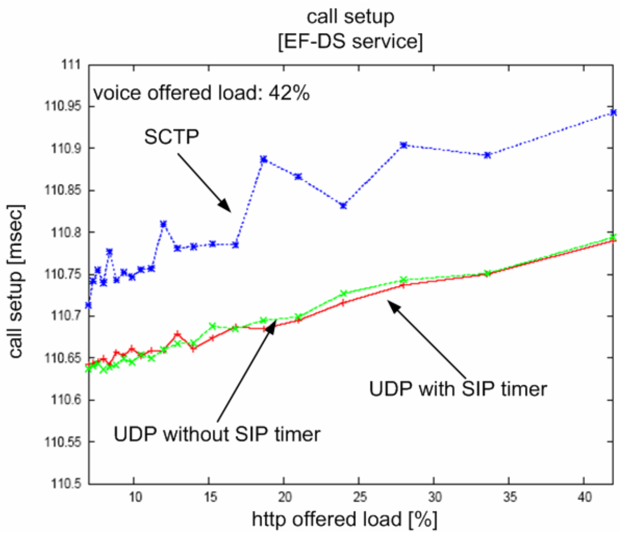

Figure 15a. call setup vs http load

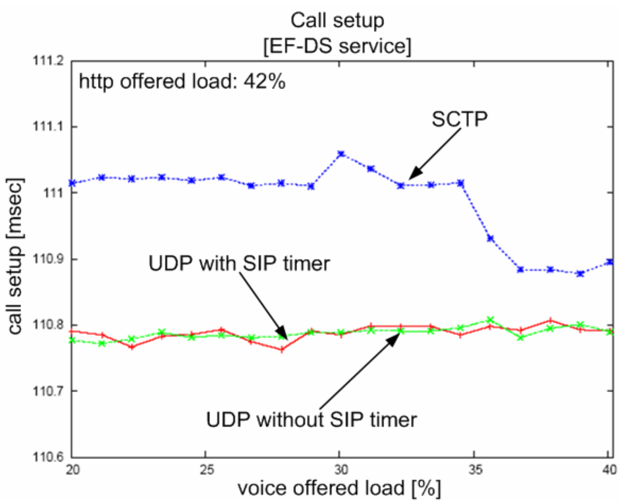

Figure 15b. call setup vs voice load

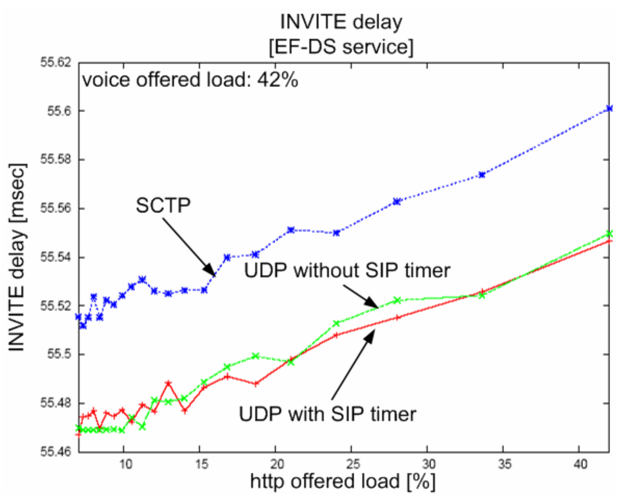

Figure 16a. INVITE mean delay vs http load

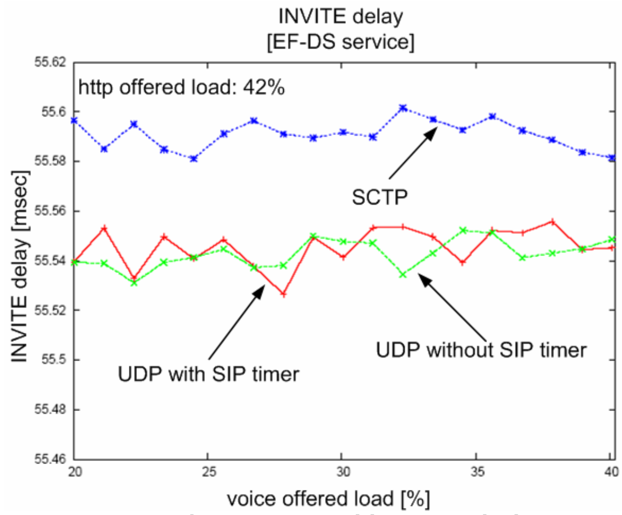

Figure 16b. INVITE mean delay vs voice load

V. CONCLUSION

In this paper we have discussed about the performance between two different transport services for Session Initiation Protocol in telephone applications. We have observed how SIP timer doesn't guarantee the correct execution of call setup using UDP protocol over Best Effort networks. Nevertheless if the network provides Quality of Service and can assure an Expedited Forwarding treatment to the call signalling, we have observed as results totally indifferent using a transport service respect to other. For all these reasons we think that the use of an unreliable service, as that provided by UDP, can be a choice by not to forsake, above all of the public telephone networks which, to keep on guaranteeing an high level of reliability and quality, have to orient themselves towards diffserv-aware frameworks.

REFERENCES

[1] The Network Simulator -2", http://www.isi.edu/nsnam/ns/

[2] [Handley M., Schulzrinne H., Schooler E., Rosenberg J., "SIP: Session Initiation Protocol", RFC 2543. IETF. March 1999.

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

[4] Rosenberg J, Schulzrinne H., "SCTP as a Transport for SIP", draft-rosenberg-sip-sctp-00.txt. IETF. Jone 2000. Work in progress.

[5] Rosenberg J., Schulzrinne H., "Reliability of Provisional Responses in SIP", draft-ietf-sip-100rel- 03.txt. IETF. March 2001. Work in progress.

[6] A. Vemuri, J. Peterson, "SIP for Telephones (SIP-T): Context and Architectures", draft-ietf-sipping-sipt-04.IETF. June 2002. Best Current Pactice

[7] Zimmerer E., Lemuri A., Nadkarni Z., Camarillo G., "SIP Best Current Practice for Telephony Interworking", draft-zimmerer-sip-bcp-t-00.txt . IETF. April 2000. Informational

[8] G. Camarillo, A. Roach, J. Peterson, L. Ong, "ISUP to SIP Mapping", draft-ietf-sipping-isup-06.IETF.April 2002. Internet-Draft

[9] J. Elwell, F. Derks, P. Mourot, O. Rousseau, "Interworking between SIP and QSIG", draft-elwell-sipping-qsig2sip-03.txt.IETF. October 2002. Internet-Draft

[10] A.Johnston, S. Donovan, R. Sparks, C. Cunningham, K. Summers , “Session Initiation Protocol PSTN Call Flows”, draft-ietf-sipping-pstn-call-flows-02.txt.IETF. April 2003. Internet-Draft

[11] M. Arango, A. Dugan, I. Elliott, C. Huitema, S. Pickett, “Media Gateway Control Protocol (MGCP Request for Comments: 2705.IETF. October 1999.

[12] E. Zimmerer, J. Peterson, A. Vemuri, L. Ong, F. Audet, “MIME media types for ISUP and QSIG Objects”, Request for Comments: 3204. IETF. December 2001.Standards Track

[13] Xie Q., Stewart R., Sharp C., Rytina I., “SCTP Unreliable Data Mode Extension”, draft-ietf sigtranusctp- 01.txt. IETF. February 2001. Work in progress.