

Internet Engineering Task Force
INTERNET DRAFT
November 16, 1998
Expires May 16, 1999
<[draft-seth-sigtran-req-00.txt](#)>

Taruni Seth
Albert Broscius
Christian Huitema
Huai-An P. Lin
Bellcore

Performance Requirements for Signaling in Internet Telephony

T. Seth, A. Broscius, C. Huitema, H. P. Lin
Bellcore

Status of this document

This document is an Internet-Draft. Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

To view the entire list of current Internet-Drafts, please check the "1id-abstracts.txt" listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa), ftp.nordu.net (Northern Europe), ftp.nis.garr.it (Southern Europe), munnari.oz.au (Pacific Rim), ftp.ietf.org (US East Coast), or ftp.isi.edu (US West Coast).

Abstract

To allow interoperability between the existing telephone network and Internet Telephony (IT) it is necessary for the signaling performance to be comparable to that of the current standards to avoid introducing degradation in the service. In this Internet Draft, we highlight the problem of providing high-quality signaling across an IP network that is built on a SONET infrastructure. We show that there are cases where the current PSTN standards are not satisfiable by a naive mapping of the IT signaling directly to the UDP or TCP transport protocols, even neglecting packet loss in router queues.

Table of Contents	Page
1. Introduction	2
2. Context	3
3. Mandatory requirements	5
3.1. General performance target for SS7 networks	5
3.2. Sequencing requirements	6
3.3. SS7 performance expectations for Delay	6
3.4. Performance targets for ISUP	8
3.5. Q.931 performance requirements	8
3.6. TCAP performance requirements	8
4. Call Setup	9
4.1. Initial response to a call request	9
4.2. Continuity test	11
5. Expected performances of the underlying IP network	12
5.1. Basic Internet quality	13
5.2. Internet Telephony IP quality	13
5.3. Underlying SONET characteristics	14
6. Adequacy of TCP	16
6.1. TCP version	16
6.2. Delay distribution	16
6.3. Effect of timer granularity	18
6.4. Effect of the Nagle algorithm	18
6.5. Can TCP meet our hard requirements?	18
7. Adequacy of UDP	19
8. Summary and Conclusion	19
9. References	20
10. Authors' addresses	22

[1.](#) Introduction

A public switched telephone network (PSTN) based telephone call involves the delivery of voice over a dedicated circuit-switched network (CSN) and the delivery of call processing signaling messages over a separate packet switched network called the Common Channel Signaling (CCS) network. Segmenting signaling data and voice on the network allows the performance guarantees of the different traffic components to be set independently, since they differ in their loss and delay tolerance requirements.

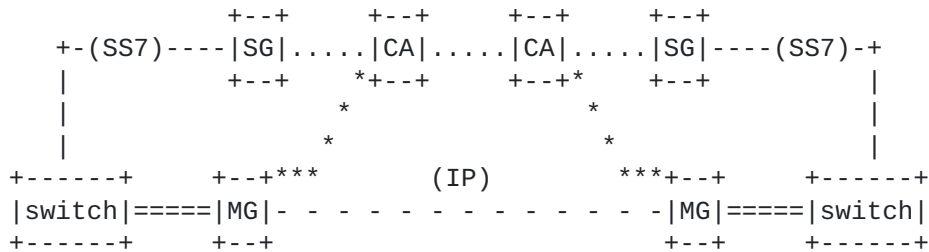
PSTN call processing involves two types of signaling messages: the ISUP (ISDN User Part) [[1](#)] messages which are responsible for the basic setup, management and teardown of a telephone call and the TCAP (Transaction Capabilities User Part) [[2](#)] messages, which are used for advanced call setup features, and require access to network databases, such as the database of valid calling card PIN numbers. Both of these protocols have specific performance requirements. To inter work with PSTN, Internet Telephony must process these ISUP and TCAP messages. Furthermore,

it may involve a variety of other PSTN signaling messages, such as Q.931, SCCP and possibly even the MTP-3 protocols. Each of these protocols has different requirements and will require to be encapsulated in specific ways.

This Internet Draft focuses mainly on some of the issues related to the message loss and delay requirement of the ISUP protocol in particular, which is required for every call processing.

2. Context

A commonly envisioned Internet Telephony system architecture includes an IP network as the core communication infrastructure. Both reliable and unreliable data is transported over the IP infrastructure through the use of a variety of upper layer protocols. For transporting voice information, the Real-Time Protocol (RTP) has been proposed as specified in RFC 1889 [3]. This protocol provides audio framing information in addition to the payload. Though voice quality issues are important, in this paper, we focus on the signalling performance, such as for example the time elapsed between the end of dialing and the beginning of the ringing/busy tone.



```

- - - RTP/UDP/IP
==== Voice Trunk
---- SS7
***** SGCP or MGCP or IPDC /IP
  
```

-- Figure 1: Reference architecture: transit service --

We will study the signalling quality problems in the context of "transit networks", where an Internet Protocol network is used to relay calls between classic telephony switches. The reference architecture for these services is shown in figure 1, where

switch is a telephony switch,

SS7 is a common carrier signalling network carrying signalling messages between the switch and the SG,

SG is a signalling gateway,

MG is a media gateway, connected to the switch by a set of trunks,

CA or MGC

is a "call agent", or "media gateway controller." The call agent receives signalling from the switch through the SG, and may send signalling over the IP network to other call agents. It controls the media gateway to set up RTP sessions or data calls as a function of this signalling.

Signalling quality requirements could be expressed by simply stating that call set-up time, and generally signalling delays, should be similar to those observed in classic telephony networks. However, one could certainly envisage different trade off between network performance, network cost, and user services. When analyzing the requirements, we will distinguish between absolute requirements, which are mandated for proper interaction with the classic telephone network, and quality objectives, which are mostly desirable goals.

We also study whether classic transport protocols, such as TCP and UDP, can meet these requirements, and we will attempt to provide the general requirements of a signalling transport protocol.

3. Mandatory requirements

The mandatory requirements of telephony systems are specified in several "General Requirement" documents published by Bellcore and ITU-T. We need to derive loss and delay bounds from the existing telephony standards for inter working of Internet Telephony and the PSTN. Both loss and latency affect the perceived user quality when establishing telephone calls across the Internet Telephony infrastructure. Excessive delay may cause call setup failure through end-switch time-outs, requiring the user to re-dial. ISUP loss may also cause call setup failures through timeouts that may leave resources in the network held in an active state after a call teardown message is lost.

Transmission times for connections with digital segments includes delay due to equipment processing as well as propagation delay. This is a critical parameter for any application whose overall performance is dependant on user or terminal interactivity. Delay allocation rules, in most standards, apply to processing time only, as the propagation time portion is determined by the distance and speed of the signal in the

transmission facility. Moreover, performance related metrics for delay are not the sole criteria and metrics such as traffic volume and economics sometimes dictate routing choices and network layouts.

3.1. General performance target for SS7 networks

The performance objectives of a SS7 network are dealt with in two ITU-T recommendations, Q.706 [11] and Q.709[12]. The design of the SS7 provides for error detection, correction and sequential transfer of signal units. The main objectives to be met are:

- 1) To limit delay in signalling connections in the network.
- 2) To achieve a high degree of availability of signalling connections.

The availability and dependability objectives for the transport of signalling messages by the MTP in these networks are:

- * No more than one in $10E+7$ (1 in 10,000,000) messages should be lost.
- * No more than one in $10E+10$ messages should be delivered out of sequence or duplicated.
- * No more than one in $10E+9$ message errors should remain undetected.
- * The signalling route between an origination and destination SP should be available 99.9998% of the times or better. This implies a maximum permissible downtime or unavailability of 10 minutes per year per route.
- * Though there are no specific end-to-end delay objectives for SS7, they are specified for specific services or uses of the SS7 protocol. Further there are delay objectives for some network components, and others can be calculated. Thus an estimate can be made for any given network configuration.

3.2. Sequencing requirements

Since the underlying SS7 network is connectionless, a stringent requirement for mis-sequenced messages has been set, as it is often easier to recover from the loss of a message by a timeout than from one delivered out-of-sequence. The MTP is able to maintain a high probability of message sequencing. This is ensured by the MTP user, which generates a value for a Signaling Link Selection (SLS) field as a parameter for each message. As the message is routed through the network, wherever there is a choice to be made between alternate routes, the link selection is made based on the SLS value in the message.

When in-sequence delivery of several messages is required, e.g. in the event a message is fragmented, such as an IAM which contains Source, Destination and Circuit Identification Code (CIC) information, then the user must specify the same SLS parameter for all the fragmented messages. They will then follow the same path through the network and be delivered in order.

3.3. SS7 performance expectations for Delay

Performance parameters for specifying signalling delays are given in terms of a hypothetical signalling reference connection (HSRC), for international working, and are defined in Q.709. Further the performance requirements are generally given for the basic building blocks of the SS7 network, namely the signalling points (SPs), Signalling Transfer Points (STPs), and the signalling links.

In ITU-T terms, signaling delays refer to the time taken for the transfer of SS7 messages from the originating SP to the destination SP in the HSRC. The maximum signalling delay is a function of several parameters, such as the propagation time on the signalling links (which is variable of distance), number of SPs and STPs involved in each connection as well as the processing time, emission and queuing delays within each of these network elements.

The maximum signalling delays in International and National Components of an HSRC, estimated for link-by-link processing over the entire connection, for 95% of signalling connections are [[16](#)]:

- * National Component (large): 520 (800) ms
- * National Component (average): 390 (600) ms
- * International Component (large to large): 410 (620) ms
- * International Component (large to average): 540 (820) ms
- * International Component (average to average): 690 (1040) ms

The values are for simple message type and for complex message types in parentheses. The definition for an average-sized country is one where maximum distance of a subscriber from an ISC is within 1,000 km or where the number of subscribers is fewer than $n \times 10$ million (n is not yet specified).

Moreover the maximum number of nodes (includes all SPs, STPs, SEPs etc), in an HSRC, for 95% of signalling connections are also fixed.

- * National Component (large): 8

- * National Component (average): 6
- * International Component (large to large): 7
- * International Component (large to average): 9
- * International Component (average to average): 12

The delays within each of the network elements such as the STP are made up of the Processor Handling time and the Message Transfer time. The values vary with the length of the message being transmitted. For a range of message lengths of (23 - 279) bytes, mean and 95% values for these times are:

STP Processor Handling Time:

Normal Processor Load	19 - 55 ms	35 - 75 ms
+ 30% Load	60 - 160 ms	120 - 320 ms

STP Outgoing Link Delay with Basic Error Correction and no Disturbances:

Link Load 0.2 erlang	4.0 - 39.6 ms	14 - 61.5 ms
Link Load 0.4 erlang	5.2 - 46.9 ms	18.6 - 87.1 ms

Thus the allowed signalling link delays can be computed for any given network. For an average-sized country with 6 nodes, a simple message of **50 bytes, with normal processor load and 0.2 erlang link load will** require about 290 ms of signalling delays within its various nodes. This implies we have approximately 100ms at most for signalling link delay.

3.4. Performance targets for ISUP

The performance objectives for the ISUP as specified in ITU-T recommendation Q.766 [13] include:

- * Unsuccessful calls due to signalling malfunctions should not exceed 2 calls in 10E+5 calls.
- * No more than 1 out of 10E+4 messages should be delayed by more than 300 ms due to error correction by retransmission.

3.5. Q.931 performance requirements

The Q.931 requirements [17] are a lot less stringent than the ISUP requirements, hence any specifications that satisfy ISUP should suffice.

3.6. TCAP performance requirements

The TCAP messages are more complex and each application has its own set

of performance requirements and timers. Moreover, at present it is unclear if we will require encapsulation of SCCP and some or all of the MTP-3 message in the transmission of TCAP over IP.

Other noted end-to-end setup times are in the range of about 1s, for the AT&T CCS-SS7 networks [18].

4. Call Setup

4.1. Initial response to a call request

The signalling of ISUP messages involved in call setup, is performed on a link-by-link basis in the circuit switched connection between the exchanges. All information necessary to establish a call is sent by the calling party to the originating exchange. Part of this information, relevant to call control, is mapped to the Initial Address Message (IAM) of the ISUP and is sent to the next exchange. This is always the first ISUP message and occurs for all call setups.

We will analyze here the two most salient delay requirements for call setup. This involves the timer values between the initiation of a call by the caller and the hearing of a ringing/busy tone:

- * Initial phase of call set-up, between the initial address message (IAM) and the return of an address complete message (ACM) coinciding with the tone. The timer "T-IAM", for when sending an IAM and awaiting an ACM, ANM, or REL message is 20-30sec [7].
- * Response to a continuity check request in an IAM. The timer "T-COT" awaiting return of a suitable tone, after the continuity test is 2 sec [7].

Messages from a telephony SS7 system can be transported using the Transmission Control Protocol (TCP) [4] over an IP based network. In the IP network, there exist the CA, which are the counterparts to the switches in the PSTN network. These CAs contain the Call Controller, which provides signaling functionality for call setup, management and teardown. Communication across the IP network can use TCP for ISUP signaling between the Signaling Gateways (SG) and the CA and other signaling, e.g. SGCP or MGCP over UDP, between the CA and the MG (Figure 2).

In this minimal scenario of a call setup (Figure 2), the incoming signaling message to the ingress signaling gateway (SG1) is processed and sent to the CA which will then exchange messages with the voice gateways to check and reserve resources for the voice traffic. In a distributed environment more than one CA will be involved. The message is then sent to the egress signaling gateway (SG2) to complete call setup over the PSTN. The actual call setup process also involves a set of message

exchanges between the CA and the MG.

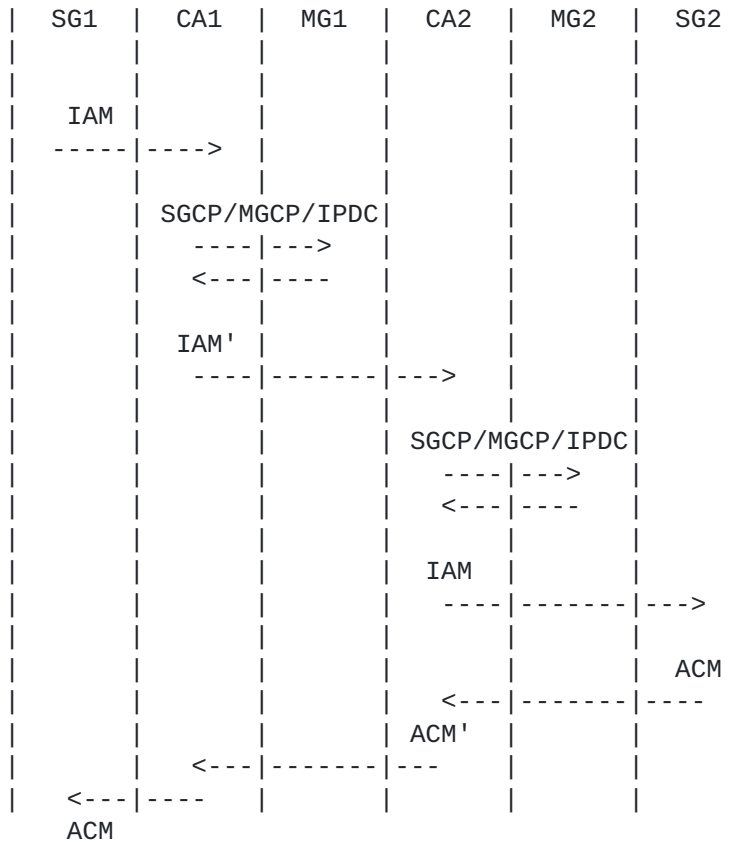


Figure 2: simplified scenario of a call setup

The sequence of messages that must complete successfully before the calling party hears the ringing tone is as follows:

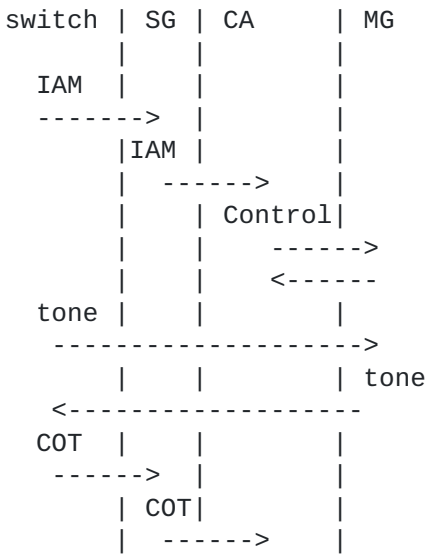
- * IAM message processing by the ingress SS7 Gateway (SG1)
- * IAM message transport to the ingress call agent (CA1),
- * IAM message processing by CA1,
- * CA1 exchange with ingress media gateway (MG1),
- * Call control message (IAM' is encapsulated IAM over IP) transport to the egress call agent (CA2),
- * CA2 exchange with egress media gateway (MG2),

- * IAM message transport to Egress SS7 Gateway (SG2),
- * Transmission of the IAM message to the egress switch, processing, and return of an ACM message,
- * Processing of ACM message by SG2,
- * Transmission of ACM message between SG2 and CA2,
- * Processing of ACM message by CA2,
- * Transmission of an ACM-like message between CA2 and CA1,
- * Processing of ACM message by CA1,
- * Transmission of ACM message between CA1 and SG1,
- * Processing of ACM message by SG1,
- * Transmission to the incoming switch.

We assume the lower value of 20 seconds for the timer "T-IAM", to prevent any switch timeout and to ensure wide compatibility. Further assuming that the transit network should not use more than half of the total delay, we see that we should not spend more than 10 seconds for the transmission and, possibly, retransmission, of 11 messages. This would imply that we have less than 1 sec to send the message over the network.

4.2. Continuity test

Continuity tests are required only on analog trunks, whereas the messages and requirements of 4.1. always hold. The most stringent timing specified in [6] is that of continuity tests. According to the specification, the test will be conducted as specified in figure 3. The tone here is not a message but is an analog tone and an electrical loopback.



-- figure 3: continuity test --

The timers specified in [7] are the following:

- * Between the sending of the IAM and the sending of the first tone by the switch (T.COT,d): 50 to 500 milliseconds.
- * Waiting for the return tone after sending the IAM: 2 seconds.

If we allocate from these 2 seconds a delay of 200 ms for processing in the CA, another delay of 100 ms for transmission over SS7, and a delay of 200 ms for processing the tones inside the MG, analogous to processing time requirements of a switch [10], we are left with at most 1.5 second for the transmission of the IAM between the SG and the CA, plus transmission of one control message between the CA and the MG (the action takes place even if the acknowledgement is lost.)

5. Expected performances of the underlying IP network

The quality of service delivered by IP transport mechanisms depends on the quality of the underlying IP network service. We have studied this quality under three assumptions:

- * Basic Internet quality, as derived from the observation of today's networks.
- * Internet telephony quality, supposing that the IP network has been engineered to provide a quality of service compatible with the transmission of voice over PSTN.

- * Best possible quality, assuming that the signalling runs at a high level of priority, that there are no congestion losses, and that we are only limited by the underlying loss rate of the transmission network, which we will assume here to be a SONET based network.

5.1. Basic Internet quality

Several measurement efforts going on today have reported figures of transmission quality that vary heavily with the network path:

- * Statistical measurements and analysis by Guy Almes [[14](#)] and also Sanghi et. al. [[15](#)], show that the losses in the Internet today are in the range of 2-10%. Losses have a direct correlation with delay and we will discuss some of these issues in the section that covers TCP and UDP.
- * Experiments conducted at Bellcore show that the busy hour average packet loss rate on random cross-Internet connections can be as high as 16%.
- * The busy hour transmission delays round trip delay between well connected sites varies between 100 and 300 ms.

A general conclusion of these observation is that the basic Internet quality, today, would not really allow the transmission of toll quality voice, except on some "lucky" subsets.

5.2. Internet Telephony IP quality

We may expect that Internet Telephony will often be transported over dedicated IP networks, and that prioritization and access control will be used to guarantee a level of service that is compatible with quality expectation of telephony users.

Telephony applications can be described as relatively tolerant to a small amount of packet losses (e.g. 1% or 2%), but very dependent on a small network delay. In fact, the key characteristic of the quality of service is the end to end voice delay [[9](#)]:

- * An end to end delay lower than 100 or maybe 150 ms is generally deemed compatible with "toll quality."
- * An end to end delay between 150 and 350 ms is generally considered mediocre but can be accepted in some circumstances. System using geo-stationary satellites incur this kind of delay.
- * An end to end delay larger than 350ms is generally not considered acceptable, except under exceptional circumstances such as, for

example, space exploration. The end to end speech delay is the sum of several components:

- * Coding and packetization delays,
- * Network transmission delays,
- * Jitter compensation delays at the receiver,
- * Decoding and play out delays.

In consequence, we may expect cross network transmission delays to not exceed 50 to 100 ms, while the packet loss rate could reach value of 1% or 2%.

5.3. Underlying SONET characteristics

Certain factors such as the speed of light and fiber path distances are lower bounds to the minimal delay in transmission of a signal through any medium. We compute theoretical values for loss bounds for a SONET-based network.

The following calculations assume that loss is due only to bit errors within the SONET links composing an IP path used for transport of ISUP messages. This assumption relies on prioritizing ISUP signaling traffic to minimize loss via QoS-based differentiated services mechanisms within the IP network.

Based on the SONET requirements [6], we compute the ISUP message loss rate in paths made of SONET links assuming no queuing loss within the routers composing that path. Three different scenarios are considered: a metropolitan area, the continental US, and an international path.

A Metropolitan area:

- * Mileage across a Metropolitan area is ~ 30 miles ~ 50 kms.
- * Each SONET span ~ 40 kms.
- * Therefore, there are 2 spans across the metropolitan area. However, there are likely to be greater than two spans in a metropolitan ring. We assume 10 in this example.
- * The worst case SONET bit-error rate [6] = $1 \times 10E-10$ / span.
- * The Bit Error Rate for SONET the metropolitan path ~ $10 \times 10E-10$.
- * Assuming an ISUP message size (20 to 200) bytes => (160 to 1,600)

bits.

- * Message Loss Rate for ISUP $\sim (160 \text{ to } 1,600) \times 10E-9 = 2 \times 10E-7 \text{ to } 2 \times 10E-6$.

The Transcontinental US:

- * Mileage across continental US is $\sim 3,000$ miles $\sim 5,000$ kms.
- * Each SONET span ~ 40 kms.
- * Therefore, there are 125 spans across the continental US.
- * The worst case SONET bit-error rate [6] = $1 \times 10E-10$ / span.
- * The Bit Error Rate for SONET transcontinental path $\sim 125 \times 10E-10 = 1.25 \times 10E-8$.
- * Assuming an ISUP message size (20 to 200) bytes \Rightarrow (160 to 1600) bits.
- * Message Loss Rate for ISUP $\sim (160 \text{ to } 1,600) \times 1.25 \times 10E-8 = 2 \times 10E-6 \text{ to } 2 \times 10E-5$.

An International Path:

- * Mileage across an international network is $\sim 15,000$ miles $\sim 25,000$ kms.
- * Each SONET span ~ 40 kms.
- * Therefore, there are 625 spans across an international network.
- * The worst case SONET bit-error rate [6] = $1 \times 10E-10$ / span.
- * The Bit Error Rate for SONET an international path $\sim 625 \times 10E-10 = 6.25 \times 10E-8$.
- * Assuming an ISUP message size (20 to 200) bytes \Rightarrow (160 to 1,600) bits.
- * Message Loss Rate for ISUP $\sim (160 \text{ to } 1,600) \times 6 \times 10E-8 = 1 \times 10E-5 \text{ to } 1 \times 10E-4$

It should be noted that Guy Almes in [] reports that the packet loss rate on the links that he observes never drops below $10E-4$. This observation should tell us that there are sources of packet losses other than congestion errors and SONET transmission errors. Possible culprits

include interconnection busses inside routers, transmission errors on Ethernet segments between a router and a server, and possible local losses in the network interfaces of work stations. Since $10E-4$ is also the worse guaranteed lost rate that we can expect from a SONET compliant international connection, it is safe to not assume that congestion control and other prioritization schemes could guarantee a packet loss rate better than $10E-4$.

6. Adequacy of TCP

The requirement of losing less than $10E-7$ packets could, on the surface, be met by simply sending packets over a TCP-IP connection. TCP meets the reliability strategy by retransmitting lost packets either after a transmission delay (duplicate ack detection) or after the expiration of a timer. However, the delay introduced by these retransmissions affect the remaining in-sequence packets and may lead to expiration of the signaling (e.g. ISUP) timers. Thus, guaranteed delivery by TCP may in fact be its pitfall.

In order to evaluate the adequacy of TCP, we performed measurements of a specific implementation, as well as analysis of the results for an emulated IP network.

6.1. TCP version

Multiple implementations of TCP are available, with slightly different retransmission algorithms, timer management, etc. In order to assess the performance of TCP, we had to pick one version.

The transmission strategy applied to our testing incorporates standard algorithms such as, Congestion Avoidance [8], Slow Start and Fast-Retransmit to ensure low-latency and a reliable operation. We will show that the loss of a packet correlates to delay being spread over subsequent packets due to the in-order delivery requirements of TCP. We discuss the mechanism of fast retransmit, following loss of a packet and its affect on the transmission. We assume there is negligible delay in transmission of packets from the application to the sender TCP and similarly between the receiving TCP and application.

6.2. Delay distribution

% of Packets Received	time (ms)
91	55
1	70
1	90
1	110
1	130
1	150
1	170
1	190
1	210
1	230

-- Figure 4: Distribution of TCP delays after fast-retransmit (one-way delay set to 50ms) --

It is observed that the delay of the number of consecutive packets following a fast-retransmit depends on the round trip time (RTT) and the Inter Packet Interval (IPI). Figure 4 gives a the delay distribution of packets over time and shows that the loss of a single packet and delay of subsequent packets at the reciever is separated by the IPI. In our example, we consider a RTT of about 100ms (based on a 50 ms one way time - delay (OWT) assumed for propagation in optical fiber across the transcontinental US) and an IPI of 20ms, which shows that loss of 1-packet results in delay of 9 packets. This implies that for fast retransmit we have:

1% loss results in 9% of packets being delayed > 1 OWT.

Packets following a loss event must wait to be delivered to the application until the missing packet is resent and received correctly at the destination host. This effect of the in-order delivery requirement of TCP results in each loss event depicted resembling a comb: several spikes with the same amplitude are equally spaced by the packet interarrival interval. The first three spikes represent the number of duplicate acknowledgments needed at the source to trigger a fast retransmit. The remaining spikes represent the number of additional packets in

flight that are delayed at the receiver from being delivered to the application by the absence of the lost data packet.

6.3. Effect of timer granularity

The emulations that we performed showed the behavior of TCP under steady load. However, when the load is less than steady, we will find situation where the "last" packet of a batch is lost. In this case, the retransmission will have to be triggered by a timer.

A similar problem arises when the loss affects a retransmitted packet. Such packets will also have to be retransmitted through a timer mechanisms.

TCP implementations try to compute the timer value through a timer estimation algorithm. The algorithm tries to estimate a value that is large enough to prevent undue retransmission, yet small enough to not cause long delays. The computation often use a coarse clock, with a 500 ms granularity, resulting in values that are never less than 500 milliseconds, and often larger than 1 second.

6.4. Effect of the Nagle algorithm

In the 1980's, TCP was often used for remote terminal applications that would send "one character per packet". It was observed that TCP's flow control, which limits the number of bytes in transit, was very inefficient under these circumstances, because it would allow stations to transmit a very large number of small packets before reaching the flow control window. For this reason, TCP implementations incorporate a rate limiting algorithm. A station that transmits a "small" packet should wait for its acknowledgement before transmitting the next packet.

The SS7 packets are short enough that if sent one at a time, they may well trigger the rate limitation algorithm, which will have two effects:

- * The next packet (or packets) will be queued for a full round trip time before transmission, which will affect performance.
- * Because the short packet will be the last of a batch, losses of that packet will have to be corrected by timer-based retransmission.

6.5. Can TCP meet our hard requirements?

We can summarize the hard requirements of ISUP by saying that transmission delays should not be larger than 1 second in more than in one case in 10,000,000. The main problem with TCP in these conditions is the timer based retransmission: a typical timer value of 1 second, combined

with three transmission delays, exceeds the 1 second limit.

The worse performances will be obtained in the case of isolated packets, due to either a low level of traffic or the triggering of a rate limitation algorithm. In this case, the only way to guarantee that the performance will be met is to make sure that the packet loss rate is lower than $10E-7$, which is a very low number. It is much lower than the average packet loss rate values observed.

Hence the calculations for theoretical performance levels of ISUP messages over a SONET network using TCP/IP may get affected by another order of magnitude. If the delay introduced by these losses exceeds the delay bounds set by the ISUP timeouts, then those messages are considered to be lost. Hence for the network comprising the transcontinental US the packet loss rates for ISUP messages are,

* Delay due to loss of packets, increases net message loss rate by factor of 10 => $2 \times 10E-5$ to $2 \times 10E-4$.

7. Adequacy of UDP

The above analysis shows the pitfalls of using TCP for the signalling in Internet Telephony. The complex connection-oriented protocol state machines in TCP add overhead for a simple request/response exchange between two hosts. Moreover, the retransmission mechanisms in TCP get triggered by any unacknowledged byte, adding an unnecessary delay to a number of subsequent signalling messages.

UDP on the other hand does not provide any loss protection to the messages transported. ARQ enhancements allow for retransmission of lost or corrupted packets, but these require at least an additional 1.5 round trip times. Recently, versions of UDP are being discussed which are supposed to provide a guarantee against loss of data. One such proposal is Reliable UDP or RUDP which supports different levels of services based on the reliability negotiated between the two endpoints. RUDP extends the datagram service of UDP to include reliable and ordered delivery, based on timer values which trigger retransmission. However, whether these will allow us to overcome all the problems of loss and delay and provide a protocol to meet the ISUP requirements is not yet established. These areas need a lot more study and analysis before any conclusion can be reached.

8. Summary and Conclusion

In this draft we have summarized the mandatory and desirable performance requirements for an Internet Telephony infrastructure that can interoperate with the existing PSTN services.

Using ISUP directly over UDP, these computations show that we cannot get adequate loss performance to meet the LSSGR ISUP loss requirements. It may be possible to use UDP with enhancements to the protocol to ensure loss and sequencing guarantees. Use of a simple TCP's retransmission mechanisms to protect against loss of ISUP messages is feasible, but at the cost of introducing latency. If the time required for correction of these losses exceeds the delay bounds set by the LSSGR, then delay of control messages within the Internet Telephony system could cause signaling failures.

Note that each ISUP message received at an ingress signaling gateway requires more than one transmission across the IP network. Additionally, the ISUP message receipt triggers other signaling protocols to be exchanged across the IP network (e.g., MGCP). While the success of these messages does not strictly affect ISUP message loss rate, their loss may induce a timeout and a subsequent loss of the associated ISUP message. This means that the effective ISUP message loss rate may be higher than that computed.

This study also shows a theoretical analysis of Internet Telephony signaling performance without queuing losses. The message loss rate for **200 byte ISUP messages can vary from 10E-7 to 10E-4 for SONET-based metropolitan to international scale Internet Telephony networks.**

The question then arises are we able to give an adequate performance level to ISUP using an underlying IP layer with some modification either to the network or to the protocols being implemented.

9. References

- [1] American National Standard Institute (ANSI), "Signaling System No. 7 (SS7) - Integrated Services Digital (ISDN) User Part," ANSI standard T1.113, January 3, 1995.
- [2] Bellcore, "AIN Switch - Service Control Point(SCP)/ Adjunct Interface Generic Requirements", GR-1299-CORE, Issue 2, Dec. 1994, Section 2-TCAP
- [3] H. Schulzrinne, S.Casner, R. Frederick, Van Jacobson, "RTP: Transport Protocol for Real-Time Applications," [RFC 1889](#).
- [4] J. B. Postel, "Transport Control Protocol (TCP)," [RFC 793](#).
- [5] A. R. Modarressi, R. A. Skoog, "Signaling System Number 7: A Tutorial", IEEE Communications, pp. 19-43, Vol. 28, No. 7, 1990
- [6] "Synchronous Generic Criteria", Bellcore document GR-253-CORE, Issue 1, December 1994.

- [7] Bellcore, "LSSGR:Switching System Generic Requirements for Call Control Using the Integrated Services Digital Network User Part (ISDNUP)", GR-317-CORE, Issue 2, Dec. 1997.
- [8] Van Jacobson, "Congestion Avoidance and Control," In Proceedings of SIGCOMM '88, Stanford, CA, ACM.
- [9] One-Way Transmission Time. Transmission Systems and Media. ITU-T recommendation G.114, rev. 1. International Telecommunication Union, 1993.
- [10] Bellcore, "Common Channel Signaling for Network Interface Specification (CCSNIS) Supporting Network Interconnection, Message Transfer Part(MTP) and Integrated Services Digital Network User Part(ISDNUP)", GR-905-CORE, [Appendix B](#), Issue 2, Dec. 1996.
- [11] Specifications of Signaling System No. 7. Message Transfer Part Signaling Performance. ITU-T recommendation Q.706. International Telecommunication Union, 1996.
- [12] Specifications of Signaling System No. 7. Message Transfer Part. Hypothetical Signaling Reference Connection. ITU-T recommendation Q.709, rev. 1. International Telecommunication Union, 1993.
- [13] Specifications of Signaling System No. 7. ISDN User Part. Performance Objectives in the Integrated Services Digital Network Application. ITU-T recommendation Q.766, rev. 1. International Telecommunication Union, 1993.
- [14] Guy Almes, "Loss and Delay Measurement Plots", <http://ippm-db.advanced.org/plots>, Advanced Network & Services, Inc.
- [15] D. Sanghi et.al., "Experimental Assessment of End-to-End Behavior on Internet", Proc. IEEE INFOCOM '93, March 1993, pp 867-874.
- [16] P. K. Bhatnagar, "Engineering Networks for Synchronization, CCS 7 and ISDN", IEEE Telecommunications Handbook Series.
- [17] Digital Subscriber Signalling System No. 1 (DSS 1) - ISDN User-Network Interface Layer 3 Specification For Basic Call Control. ITU-T recommendation Q.931, International Telecommunication Union, 1993.
- [18] AT&T Webpage, <http://www.att.com/attlabs/people/fellows/lawser.html>

10. Authors' addresses

Taruni Seth
Bellcore
445 South Street, MCC-1G209R
Morristown, NJ 07960-6438

Phone: 973 829-4046
Email: tseth@notes.cc.bellcore.com

Albert Broscius
Bellcore
445 South Street, MCC-1A264B
Morristown, NJ 07960-6438

Phone: 973 829-4781
Email: broscius@bellcore.com

Christian Huitema
Bellcore
445 South Street, MCC-1J244B
Morristown, NJ 07960-6438

Phone: 973 829-4266
Email: huitema@bellcore.com

Huai-An P. Lin
Bellcore
445 South Street, MCC-1A216R
Morristown, NJ 07960-6438

Phone: 973 829-2412
Email: hlin@bellcore.com