INTERNET DRAFT Taruni Seth

Internet Engineering Task Force Albert Broscius

February 26, 1999 Christian Huitema

Expires August 26, 1999 Huai-An P. Lin

<draft-ietf-sigtran-tcap-perf-req-00.txt> Bellcore

Performance Requirements for TCAP Signaling in Internet Telephony

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

Status of this document

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 discuss the

performance requirements for TCAP signaling across an IP network. We also highlight the dependency on the SCP database location and thus problems related in providing high-quality service for TCAP based applications.

Table of Contents

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. PSTN call processing involves two types of signaling messages: the ISUP (ISDN User Part) [2] messages which are responsible for the basic setup, management and teardown of a telephone call and the Transaction Capabilities User Part (TCAP) [3] messages, which are used for non-circuit related messages used in advanced call setup features, and those requiring access to network databases, such as the database of valid calling card PIN numbers. To interwork with PSTN, Internet Telephony must process these ISUP and TCAP messages. Both of these protocols have specific performance requirements. Requirements for ISUP messages were discussed in [1]. This Internet Draft focuses mainly on some of the

issues related to the performance requirement of the TCAP messages.

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. In the IP network, generally, all the traffic competes against each other in a single flow. Segmenting signaling, data and voice traffic flows 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.

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. 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 based on user expectations.

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. TCAP 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.

Earlier work specified the requirements for ISUP Signaling over IP based on the PSTN specifications[1]. ISUP messages are used for basic call control and setup, whereas TCAP messages are particular to each specific application and do not have generic performance requirements. However, there are some specifications for the network elements involved in a TCAP transaction, which allow us to estimate some other measures. Here we will try and establish a similar set of parameters for TCAP messaging, based on signalling performance metrics such as the query-response delay.

In this Internet draft, we focus on the performance requirements for TCAP applications. We discuss mandatory requirements as established by the timers in PSTN network elements, which determine the tolerance of these applications to delays and losses. We will also discuss the user expectations and values of these requirements as available from current

PSTN implementations. We summarize the implications of these in a VoIP framework and discuss merits of existing technologies to enhance these performance requirements in the IP networks.

3. Overview of TCAP

3.1. TCAP in a Nutshell

TCAP messages are designed for accessing databases or other switches to retrieve information or invoke features. TCAP enables the deployment of advanced intelligent network services by supporting non-circuit related information exchange between signaling points using the signaling connection control part (SCCP) connectionless service for message transport. This is a fundamental difference between ISUP and TCAP protocols. ISUP messages follow a particular path used to establish a circuit connection and use the Message Transfer Part (MTP) to route its messages, whereas TCAP information is not related to any one circuit and must be transferred through the network using end-to-end signaling, which is achieved by the SCCP protocol above MTP. The term "Transaction Capabilities" refers to the application layer protocol, called TCAP, plus the supporting Presentation, Session, and Transport layers, called the Application Services Part (ASP). A common example of TCAP usage is in dialing a 800, 888, or 900 number. An SSP uses TCAP to query an SCP to determine the routing number(s) associated with the dialed digits. The SCP uses TCAP to return a response containing the routing number(s) (or an error) back to the SSP.

TCAP messages consist of two portions: a Transaction Portion composed entirely of protocol control information, and a Component Portion which contains protocol-related information as well as data concerning the application process.

The transaction portion of the TCAP message identifies whether the transaction between two nodes is expected to consist of a single message (i.e. one way communication) or multiple messages (i.e. interactive communication). There are five types of TCAP messages, called Package Type: Query, Response, Conversation, Unidirectional, and Abort. The transaction portion provides the information necessary for the signaling point to route the component information to its destination. It contains the Transaction ID and the package type identifier. The Transaction ID is a reference to correlate messages within the same transaction and associate the TCAP transaction with a specific application at the originating and destination signaling points.

In a single transaction, one or more operations may occur. For each operation, one or more components may be involved. Components provide the information that request an action, invoke an operation or provide the reaction to a previous request. Component types include Invoke,

Return Result, Return Error, and Reject. Components include parameters which contain application-specific data carried unexamined by TCAP.

3.2. Signaling Connection Control Part

The SCCP provides connectionless and connection-oriented network services above MTP Level 3. SCCP provides two major functions that are lacking in the MTP. The first of these is the capability to address applications within a signaling point. The MTP can only receive and deliver messages from a node "as a whole"; it does not deal with software applications within a node. While MTP Level 3 provides point codes to allow messages to be addressed to specific signaling points, SCCP provides subsystem numbers to allow messages to be addressed to specific applications (called subsystems) at these signaling points. SCCP is used as the transport layer for TCAP-based services such as toll free (800) phone, calling card, and wireless services. The SCCP controls provide efficient routing of TCAP like messages that are independent of voice network connections. Their routing information may contain the Link Selection information besides the Originating and Destination Point Codes, if necessary.

3.3. Global Title Translation

SCCP also provides the means by which a Signal Transfer Point (STP) can perform global title translation (GTT), a procedure by which the destination signaling point and subsystem number (SSN) is determined from digits (i.e., the global title) present in the signaling message. The global title digits may be the dialed 800/888 number, calling card number, or mobile subscriber identification number depending on the service requested. Since an STP provides global title translation, originating signaling points do not need to know the destination point code or subsystem number of every potential destination to which they might have to route a message for the associated service. Only the STPs need to maintain a database of destination point codes and subsystem numbers associated with specific services and possible destinations. The STP examines the message, and determines where the message should be routed.

Switches can generate queries addressed to their local STPs, which, using global title translation, select the correct destination to which the message should be routed. STPs must maintain a database that enables them to determine to where a query should be routed. Global title translation effectively centralizes the problem and places it in a node (the STP) that has been designed to perform this function. Further, an STP can perform "intermediate global title translation," in which it uses its tables to find another STP further along the route to the destination. Intermediate global title translation minimizes the need for STPs to maintain extensive information about nodes which are far removed from them. Global Title Translation is also used at the STP to share load

among mated Service Control Point (SCP) databases in both normal and failure scenarios. It can select an SCP on either a priority basis (referred to as primary -- backup) or to allow load sharing across all available SCPs.

3.4. Service Control Point (SCP)

An SCP site refers to the hardware and node software needed to support applications. An SCP is a network system that supports the execution of service logic in response to queries from switching systems. These services are implemented as features of switching systems and network databases. The SCP node serves as a network host to each application and provides common functions as distributing incoming messages to the appropriate application, assigning external signaling links, SCP maintenance and controlling SCP overload procedures. The SCP applications consist of application-specific software and data that can be accessed by other nodes on the signaling network, such as a switching system or an Operator Services System (OSS). Each application may provide many individual services, and these are transparent to the SCP node. In return, this application formulates the parameters of a response message, which the node formats and routes to the appropriate network entity.

3.5. TCAP Transaction Flow Diagram

In the VoIP framework, TCAP messages must be supported not only in order to interoperate with the PSTN, but to allow development of new services. In the IP network, there exist the Media Gateway Controllers (MGC), which are the counterparts to the switches in the PSTN network. These MGCs contain the Call Controller, which provides signaling functionality for call setup. The TCAP signaling may be envisioned within IP networks between the MGC and/or a SG and an IP-based SCP (IP-SCP) (Figure 1). Further, as shown in the architecture document [4], the TCAP signaling may also be used for cross-access between entities in the SS7 domain and the IP domain, such as: - access from an SS7 network to an IP-SCP - access from an SS7 network to an MGC - access from an MGC to an SS7 network element (SCP) - access from an IP-SCP to an SS7 network element

In this minimal scenario of a TCAP call query (Figure 1), the signal may be an incoming TCAP query to the ingress signaling gateway (SG) or may originate at the MGC. It is accordingly processed and sent to the IP-SCP which will then process it and return an appropriate response.

Call Related TCAP Transaction/Message Flow Diagram [5]

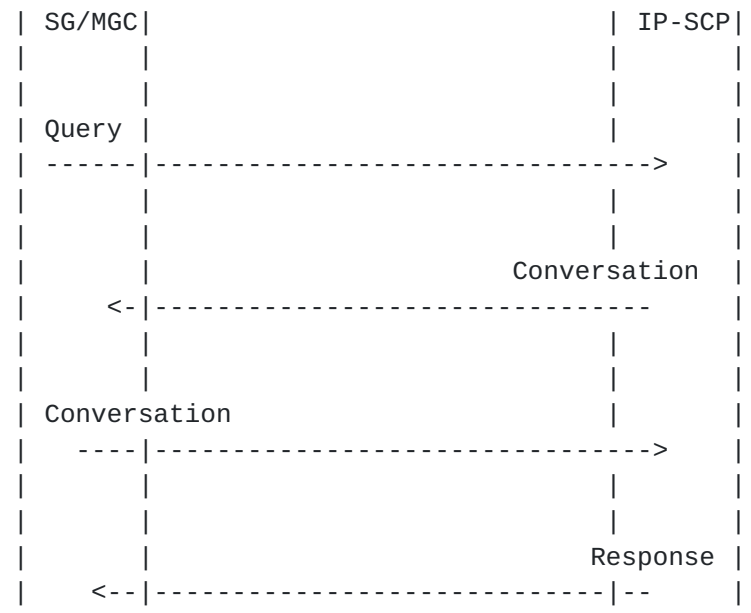

Figure 1: simplified scenario of a TCAP query and response

The sequence of messages that must complete successfully before the TCAP transaction query is satisfied is as follows:

* Query message processing by the SG or MGC

* Query message transport to the IP database (IP-SCP),

* Query message processing by the IP-SCP,

* Conversation message processing by the IP-SCP,

* Conversation message transport from the IP-SCP to the SG or MGC,

* Conversation message processing by the SG/MGC,

* Conversation message transport from the SG/MGC to the IP-SCP,

* Conversation message processing by the IP-SCP,

* Response message processing by the IP-SCP,

* Response message transport from the IP-SCP to the SG/MGC,

* Response message processing by the SG/MGC,

4. Performance Requirements

The end-to-end AIN performance is a function of the performance of each AIN network element (e.g., an SSP, STP) and each network system (e.g., a SCP) and their interfaces. Thus to ensure overall network performance, the performance objectives and requirements must be defined for each component. We described the requirements for STPs in our earlier document [1], and here we discuss them for SSPs and SCPs. The SSP generally originates a query to the SCP and awaits its response. The performance of the TCAP applications relies on the Query Response Time. The SCP performance depends not only on its several components and interfaces but also on the application process(es) involved. Thus its conformance testing depends on the use of standard application processes called benchmark transactions, that emulate the potential AIN service on a SCP.

4.1. Query Response Time

Query Response Time is defined as the time it takes to send a query to a database host and for the database to process the query and return data to the querying entity (e.g. SSP, STP). The Query response time for the Network User Identifier (NUI) database host and the NUI database to process a Public Packet Switched Network (PPSN) query and return the data to the querying entity is given as [6]:

* Mean value 0.25 to 0.5 seconds

The common user expectation for most simple TCAP query-response applications appears to be on the order of 0.5sec. However, the actual PSTN working data shows that these are in the range of 250-350ms.

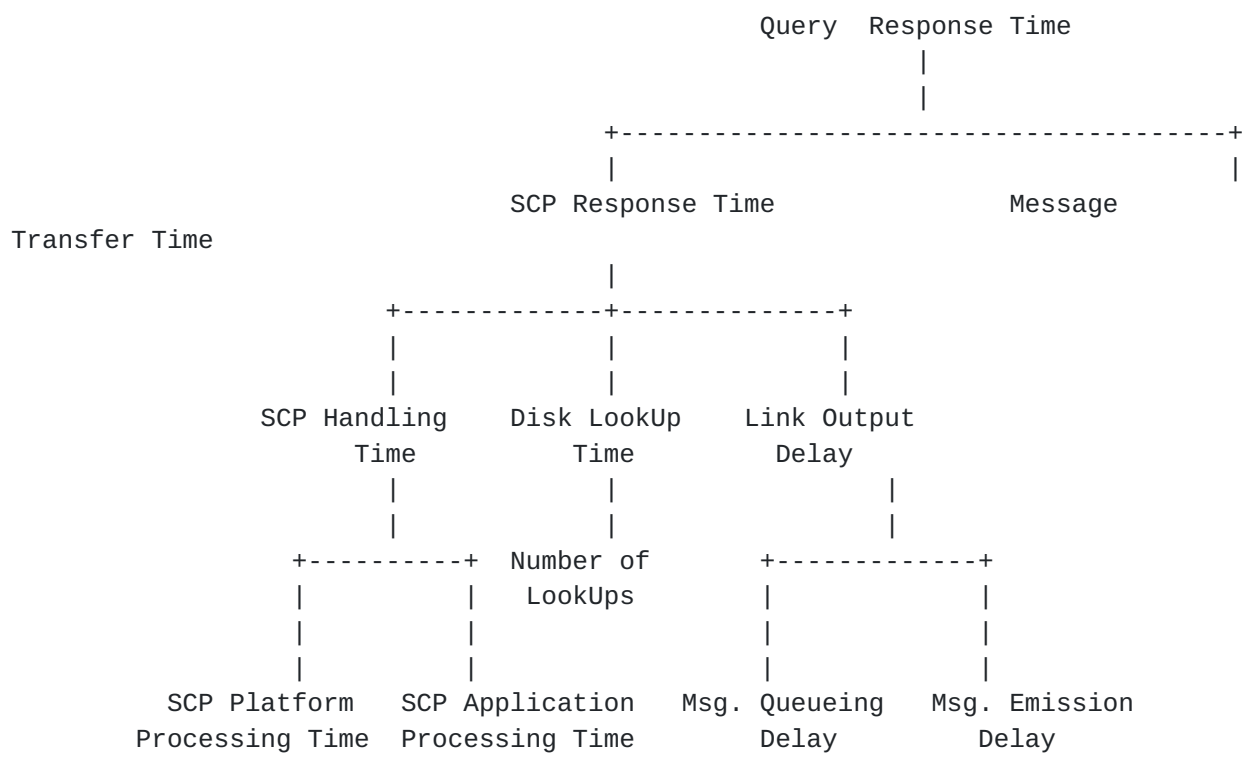

Figure 2: Time components in a TCAP query

4.2. SSP Response Time

The SSP uses a timer T1 to establish the time for a message from the SCP in response to a message sent by the SSP to the SCP. Each instance of this timer T1 is associated with a particular TCAP transaction, and only one timer is set for a given transaction. The allowable range for timer T1, is from 0.1 to 30.0 seconds, with 0.1 second increments, and a default value is 3.0 seconds [7].

The SSP starts this T1 timer after it sends a Query (or Conversation Package) to the SCP and then awaits an SCP call-related response (or Conversation Package) for the particular TCAP transaction. The SSP cancels the timer on receipt of this response, closes the particular TCAP transaction, and continues call processing.

4.3. SCP Response Time

The SCP response time is calculated as a sum of the SCP Handling Time, Disk Lookup Time and Link Output Delay. It is defined as the interval that begins when the last bit of a Call Related message enters the SCP, and ends when the last bit of a Call Related message leaves the SCP.

4.3.1. SCP Handling Time

It is defined as the interval that begins when the last bit of a Call Related message enters the SCP, and ends when the last bit of a Call Related message is placed at the outgoing signaling link buffer, excluding time taken for a disk lookup. It is further subdivided into the SCP Platform Processing Time and the SCP Application Processing Time.

SCP Platform Processing Time: This has three time contributions. The first begins when the last bit of a Call Related message enters the SCP, and ends when the TCAP message is made available to the application process. Second involves execution time of any application support processing functions needed by the SCP application, and the third begins when the platform receives the outgoing message from the application process and ends when the last bit of a Call Related message is placed at the outgoing signaling link buffer. The mean and 95th percentile SCP Platform Processing Time in processing a Call Related message which does not involve a disk look up [7], (but does involve processing of critical operations, administration, and maintenance functions) is

* Mean value <= 100 ms

* 0.95 prob. of not exceeding <= 120 ms

SCP Application Processing Time: It is the difference between the SCP Handling Time and the SCP Platform Processing Time. It is a function of desired overall service delay, network architecture, deployment etc. and is negotiated and tailored to each application's need.

4.3.2. Disk Lookup Time

Certain applications require access to data on the disk, and some messages may even require multiple disk lookups. The Disk Lookup Time required in determining a SCP response message, must be

* <= 30 ms for each Lookup.

4.3.3. Link Output Delay

It is the interval that begins when last bit of a SCP Response message is placed at the outgoing signaling link buffer, and ends when the last bit of the Call Related message leaves the SCP on the outgoing signaling link. The two components of the Link Output Delay are the message queuing delay and the message emission delay. Queuing delay is a function of link occupancy and the message length distribution. Emission delay is a function of the signaling link speed and the message length distribution. Under normal conditions, messages are expected to be shorter than **279 octets. The Link Output Delay [7] calculated using the M/G/1 queuing**

formula for messages of length 279 octets and at a Link load of 0.4 erlang should be

	56 Kb/s	64Kb/s
Mean value	<= 55 ms	<= 47 ms
0.95 prob. of not exceeding	<=102 ms	<= 89 ms

4.3.4. Total SCP Response Time

The industry requirement for SCP response times for a simple TCAP transaction (one query, one response, for data retrieval from a LIDB for an operator system) can be summarized as follows [6]:

	Daily Peak (aka Reference Load A)	Yearly Peak (aka Reference Load B)
Mean value	<= 250 ms	<= 400 ms
0.95 prob. not exceed	300 ms	600 ms

4.4. Message Transfer Time

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 SSPs, STPs, (or SG/MGC) and SCPs involved in each connection as well as the processing time, emission and queuing delays within each of these network elements. 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. However it is possible to estimate the upper bound for the time available for message transfers from the maximum query response time and a sum of the times required by the various components of the network involved in the query.

Consider a mean time of 350ms for a simple TCAP query-response application, which uses a single disk lookup, a 25ms average SCP Application Processing Time and a 64Kbps link. Then, the transmission time available between the querying agent and the database is of the order of 150 ms for a query and its response.

Mean Values

SCP Platform Processing Time	100 ms

* Disk Lookup (Assuming one)	30 ms
* SCP Application Processing Time (Assumed)	25 ms
* Link Output Delay (64 kbps link)	47 ms
---	---
* SCP Response Time	202 ms
* Average Query Response	350 ms
* Implied Message Transfer Time	148 ms

Thus, if the SCP Response Time is independent of the network technology (other than the Link Output Delay), we have roughly a 150 ms budget for TCAP message transfer--round trip, or 75 ms one way, to achieve a response time comparable to the PSTN.

4.5. Other General Parameters

Certain parameters such as maxResponseTime and maxPendingTime are also defined as general performance measures over the CCS network.

4.5.1. maxResponseTime

It is the maximum time the SCP allows itself to respond to a message received over the CCS network. It takes into account the time it will take for the SCP reply to reach the sending node i.e. the SSP. It is an administrable value with a default set to 2.0 seconds [6].

4.5.2. maxPendingTime

It is the maximum time the SCP will wait for a reply to a Conversation or Query Package Type. This value is meant to be a "catch all" in case of an error. Applications that need to monitor the time or depend on the timing of the reply to the message sent set separate timers. It is an administrable value with a default set to 15.0 minutes [6].

5. Implications to VoIP

Telephony applications can be described as relatively intolerant to packet losses and network delay. The quality of service delivered by an IP transport mechanisms depends on the quality of the underlying IP network service. Statistical measurements and analysis by Guy Almes [8] and also Sanghi et. al. [9], show that the losses in the Internet today are in the range of 2-10%. Losses have a direct correlation with delay. A tentative conclusion from this is that the basic Internet quality, today, would not really allow the transmission of toll quality voice,

except on some "lucky" subsets.

We may expect that Internet Telephony traffic will often be transported over dedicated IP networks, and that prioritization and access control will be used to minimize loss and delay (to signaling traffic), via QoS-based differentiated services mechanisms. This will guarantee a level of service that is compatible with quality expectation of PSTN users. In IP routing, the use of differentiated services via traffic schedulers such as Weighted Fair Queuing (WFQ), and Priority Queuing, allows traffic flows to be distinguished and prioritized. This enables allocation of QoS parameters to different flows. However, preliminary laboratory tests of commercial routers supporting differentiated services indicate that there are some overhead delays associated with implementing these mechanisms and may lead to some unexpected complications. These delays are probably processing delays of the WFQ scheduling algorithm and some additional queuing delays. The average shift in the delay comprises of:

1. **A one-time overhead of 10ms delay associated with tagging the incoming signaling packets with the required precedence bits at the** ingress router.

2. **A 10ms per-node (router) delay arising from applying the WFQ algorithm to the different traffic flows.**

The delays associated with the implementation of WFQ will probably increase linearly with the number of routers in the transmission path. The network designer would have to very carefully decide if they can justify the use of QoS to guarantee reliability, if use of differentiated services imposes unacceptable delays on the transmission of TCAP signaling messages. For example, if one-way transmission time is about 75ms and there are 3 routers in the path between the querying entity and the IP-SCP, then applying ToS based WFQ to the signaling packets would reduce available transmission time to roughly 35ms (10ms for tagging and 30ms for WFQ at the 3 routers). It may be argued that these delays are vendor specific and can be improved over time and that packets may be tagged at source. However, tagging at ingress routers across different network domains as a security measure may still be an issue.

In the IT architecture the IP-SCP is still an ambiguous network element. It would allow the SG or the MGC to directly access the database to satisfy the TCAP queries. Since the IP network is not as robust as the existing PSTN, the high loss probability of messages will impose strict restrictions on the location of the IP-SCP within the IP network. Moreover, if the TCAP query originates in the IP network and needs to access an PSTN based SCP, it would involve complex message processing and transmission. The estimate of the message transfer time in Section 4.4,

would imply that there is little scope for retransmission in the event of network losses. At this stage, we can only conclude that these delays need to be investigated more thoroughly before deciding the effectiveness of using Differentiated Services to prioritize signaling traffic.

6. References

[1] T. Seth, et al, "Performance Requirements for Signaling in Internet Telephony" <draft-seth-sigtran-req-00.txt>, Nov. 1998.

[2] American National Standard Institute (ANSI), "Signaling System No. 7 (SS7) - Integrated Services Digital (ISDN) User Part," ANSI standard T1.113, January 3, 1995.

[3] Bellcore, "AIN Switch - Service Control Point(SCP)/ Adjunct Interface Generic Requirements", GR-1299-CORE, Issue 2, Dec. 1994, Section 2-TCAP.
.IP [4] L. Ong, " Architectural Framework for Signaling Transport" <draft-sigtran-framework-arch-01.txt>, Feb. 1999.

[5] Bellcore, "Advanced Intelligent Network Generic Requirements (AINGR): Switch - Service Control Point(SCP)/ Adjunct Interface, ", GR-1299-CORE, Issue 4, Sept. 1997.

[6] Bellcore, "Service Control Point Node, Generic Requirements", TR-NWT-000029, Issue 1, Sept. 1990.

[7] Bellcore, "Advanced Intelligent Network (AIN) Service Control Point(SCP), Generic Requirements", GR-1280-CORE, Issue 1, Aug. 1993.

[8] Guy Almes, "Loss and Delay Measurement Plots", http://ippm-db.advanced.org/plots, Advanced Network & Services, Inc.

[9] D. Sanghi et.al., "Experimental Assessment of End-to-End Behavior on Internet", Proc. IEEE INFOCOM '93, March 1993, pp 867-874.

7. Authors' Addresses

Taruni U Seth
Bellcore
445 South Street, MCC-1G209R
Morristown, NJ 07960-6438
Phone: 973 829-4046

Email: tseth@notes.cc.bellcore.com

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

Full Copyright Statement

HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.