Predicting Internet Telephony Call Setup Delay

Tony Eyers
TITR
University of Wollongong
Australia
tony@elec.uow.edu.au

Henning Schulzrinne
Department of Computer Science
Columbia University
New York, NY 10027
hgs@cs.columbia.edu

September 23, 1999

Abstract

Internet telephony has been the focus of much recent effort by ITU and IETF standards bodies, with initial, albeit small-scale deployment in progress. While Internet telephony voice quality has been studied, call setup delay has received little attention. This paper outlines a simulation study of Internet Telephony Call Setup delay, based on UDP delay/loss traces. The focus is signaling transport delay, and the variations arising from packet loss and associated retransmissions. Of particular interest are the differences arising from H.323 signaling, which uses TCP, and SIP, which can use UDP with additional error recovery. Results show that during high error periods, H.323 call setup delay significantly exceeds that of SIP. We also consider PSTN/Internet telephony interworking, and show that high blocking rates are likely if either H.323 or SIP are used across the public Internet.

1 Introduction

Internet telephony is experiencing significant growth, prompted initially by low-price long distance calls [1]. Longer term growth will be motivated by the greater service flexibility offered by IP networks, compared to the Public Switched Telephone Network (PSTN) [2]. This flexibility arises from the increased signaling capability of IP end systems compared to current handsets, as well as the ability to support multiple media types. To be widely accepted however, Internet telephony QOS must match or exceed that of the PSTN. From a user's perspective, the quality of service consists of the reliability and amount of time of setting up the call and then the audio and video quality of the actual call. While the latter aspect has received considerable attention, experience has shown that Internet telephony call setup times can be much longer than the essentially instantaneous call setups that have become routine for the PSTN. This paper attempts to predict call setup delays over the public Internet.

Over the 120 year history of telephony, signaling performance has improved markedly. For example, Gherardi and Jewett [3] note that call setup time dropped from 4 minutes (!) in 1923 to 1.2 minutes in 1928. Duffy and Mercer [4] reports that in 1978, the average time between end of dialing and ringback was about 10.9 s. In 1998, AT&T [5] claimed a call setup time of less than two seconds for toll calls, and 2.5 seconds for calls requiring database lookups. Since Internet telephony signaling uses the same high-speed backbone links as for data while most SS7 systems are still connected by 64 kb/s links, call setup delay could be significantly less for Internet telephony. We will explore this in detail below.

Call setup delay (also known as post-dialing delay or post-selection delay [6]) is defined as the interval between entering the last dialed digit and receiving ringback. Another, related, measure is the time between entering the last dialed digit and when the callee's phone starts to ring. We will refer to this delay as the *dial-to-ring delay*, as there does not seem to be a standard designation. In a traditional phone system, there is no acoustic feedback between dialing and ringback, so that an excessive delay until ringback may lead the caller to believe that "something is wrong" and abandon the call. Internet telephony has the advantage that it can provide additional feedback during call setup, before ringback. For example, SIP servers can send any number of *provisional responses* that indicate the progress of address translations or other network actions, as discussed in Section 3. E.721 [6] recommends an average delay of no more than 3.0, 5.0 or 8.0 s, for local, toll and international calls, respectively. The 95th percentiles are set at 6.0, 8.0 and 11.0 s, respectively.

The importance of the dial-to-ring delay depends on the type of call. For example, if a fax machine is "blast-faxing" to a number of receivers, the call setup time becomes an important component of the achievable throughput. It is similarly important for short data calls like checking email, although the connection setup delay of modem calls appears to be dominated by the modem training time and PPP delays that often take ten seconds or more. For completed calls, that is, about 70% of all calls [4], it takes the callee about on average 8.5 s to pick up the phone, so that reducing the call setup time much below a second probably yields limited improvement for human-to-human calls. Traditional benchmarks for signaling performance cannot distinguish between these different uses of the telephone system, but the distinction may be important if Internet telephony is primarily used for human-to-human contact, with data and fax using other mechanisms.

Another important signaling delay is post-pickup delay (or, more formally, answer-signal delay [6]), which roughly measures the delay between the time the callee picks up the receiver and the time the caller receives indication of this. The actual definition in E.721 only considers the message transfer delay, not any delays incurred in the end systems. If the speech path from callee to caller only gets cut through when this message reaches the caller, the first "hello" of the callee may get lost, leading to confusion. While this paper does not provide measurements for this particular delay, we will describe how it relates to post-dial delay. E.721 [6] recommends average delays of 0.75 s for local, 1.5 s for toll and 2.0 s for international connections, with 1.5 s, 3.0 s, and 5.0 s as 95% values.

The design and performance modeling of circuit switched networks has been an active research area for most of the 20th century. In particular, the advent of common channel signaling (SS7) has prompted many performance studies [7]. The result has been a robust PSTN with tightly engineered QOS, particularly with regard to call setup delay. Associated with this is a series of ITU recommendations which specify performance targets for signaling transport [8] and call processing [9]. These underpin the signaling network engineering (e.g., [10]) and switch design needed to ensure call setup delay performance.

The recent arrival of Internet Telephony (ca. 1996) and the much more varied network infrastructure have precluded the same depth of performance study. Some Internet Telephony call setup delay targets have been proposed [11], based on ITU recommendations for the PSTN. However, delay targets for Internet telephony call setup components encompassing both IP signaling transport and server delay are not yet in place. Indeed, it appears unlikely that an Internet standard for signaling delay will emerge, except possibly in connection with ensuring delay targets for SS7 networks. Instead, customers may make signaling delay part of their service level agreement (SLA) with a carrier.

Internet Telephony uses new end-to-end signaling protocols, such as H.323 [12] and SIP [13], with IP networks providing signaling message transport. We present a simulation study of Internet Telephony call setup delay, based on these protocols and Internet delay traces. The purpose is threefold: to determine call setup delays arising from signaling transport within the public Internet, to compare the relative performance of SIP and H.323, and to investigate blocking probabilities arising from PSTN/Internet telephony interworking. A key finding is that TCP error control, used in H.323, can significantly increase call setup delay compared to the UDP-based approach commonly used in SIP.

Section 2 outlines previous Internet telephony performance studies, then reviews ITU recommendations for call setup delay, and their applicability to Internet telephony. Section 3 presents H.323 and SIP call setup procedures, highlighting the different packet loss recovery techniques. Section 4 describes the Internet delay traces used for this study. Section 5 presents the simulation methodology. Results in section 6 compare

H.323 and SIP over a variety of paths. Discussion and conclusions are in sections 7 and 8, respectively.

2 Prior Work

A major thrust of Internet telephony research has been protocol development. H.323 [12] and SIP [13, 14, 15] have emerged as the key peer-to-peer call setup protocols, with G.729 and G.723.1 the leading low-bit-rate audio codecs. Current protocol issues include billing, address resolution [2] and resource allocation [16]. QOS arising from the established and developing protocols is remains a key issue.

Initial Internet telephony QOS research has considered voice quality arising from deployment of the new codecs over the public Internet. Kostas *et al.* [17] generate UDP trace records over six months, and, from the mean delay and standard deviation, conclude that acceptable voice quality would generally be available over the intra-USA paths considered. Maxemchuk and Lo [18] extend this work by incorporating compensation within the codecs for lost and delayed voice packets. They conclude that acceptable performance is usually available within the USA, but that voice quality on international calls is often poor.

Call setup delay is a key and easily discernable QOS parameter. To date however, there appears to have been no Internet telephony call setup studies incorporating signaling transfer delays. Elwalid *et al.* [19] apply a queueing analysis to an H.323-based switch, to determine the intra-server delay, that is, the call setup delay distribution within the switch. The 99th percentile of this delay is, seemingly arbitrarily, bounded at 1.5 seconds, with this bound used to determine the maximum server load. Signaling message transfer delay between switches is not considered. In this paper we take the opposite approach, by modeling the signaling transfer delay for SIP and H.323 messages. The total call setup delay also includes server call processing and the translation between domain names and IP addresses via DNS. The server call processing delay can vary widely, depending on whether the server, for example, makes calls to networked databases or processes per-call scripts [20]. Given the variability of both components, we do not attempt to characterize them here. From our experience, a basic SIP redirection operation takes between 10 and 100 ms, depending on whether an external process in invoked or not.

Lin *et al.* [11] propose Internet telephony call setup delay targets based on the ITU Q.725 targets for the PSTN. The ITU figures for signaling transfer point (STP)[1] message transfer delay, maximum number of signaling hops and ISUP message transfer delay are combined to provide mean and 95th percentile Internet telephony call setup delay targets. The delay targets in [11] and [19], while both pertaining to post-dial delay, are different. In particular, the 99th percentile is specified in [19], even though no source is listed for this figure. While the delay targets in [11] are based on ITU sources, they do not appear to incorporate queueing delay for signaling messages (ITU figures for this are in E.733 [21]). In addition, [11] does not mention ITU recommendation E.721 [6], which provides mean call setup delay targets. However, the figures in E.721 and [11] are similar.

While ITU delay targets provide a starting point for Internet telephony, there are clear differences. Directly mapping SS7 signaling transport delay and ISDN cross-switch delay to Internet telephony implies that the ratio of these two components is the same for the PSTN and the Internet. Given the difference between SS7 and IP message transport, it seems more appropriate to define specific targets for each. There is another constraint on Internet call setup delay, which applies when interconnecting with ISDN switches. These switches may abandon a call if a reply from a setup attempt (i.e., an IAM signaling message) is not received within two seconds [22]. Hence, for PSTN/Internet interworking, an additional Internet call setup delay target is required, which keeps this loss rate within acceptable bounds.

[1] An SS7 "router"

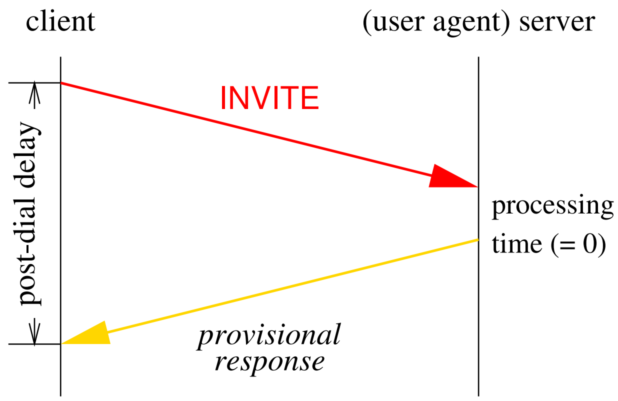

Figure 1: SIP call setup (initial portion)

3 H.323 and SIP Call Setup

Both H.323 and SIP are *peer-to-peer* signaling protocols, used by Internet telephony end systems to establish multimedia sessions. H.323 and SIP allow a variety of call setup mechanisms, ranging from a single message exchange between caller and callee (e.g., H.323v2 Fast Connect), to more complex calls which traverse a number of servers before reaching their destination. Fundamental to all these call types are provisions for reliable message transfer in the face of packet losses, which, in turn, determine the upper bounds for call setup delay. We outline the H.323 and SIP call setup procedures, focusing on the error recovery techniques.

3.1 SIP

Figure 1 shows part of a basic SIP call setup. A Client sends an INVITE call setup message to a User Agent Server (callee). Usually, the UAS returns one or more provisional response messages indicating receipt of the INVITE request and call progress. This is roughly equivalent to the ISDN IAM/ACM message exchange, with the delay representing the post-dial delay. Figure 2 shows a more complex SIP call, where the client first queries a *redirect server*, whose response contains either the address of the final destination or that of another redirect server. The rest of the call continues as in Figure 1. More complex SIP call types are outlined in [15].

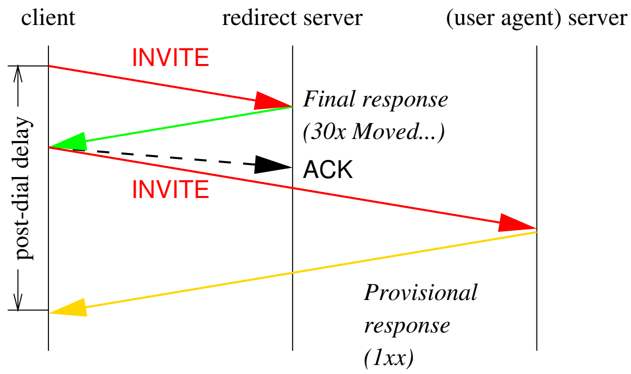

Figure 2: SIP call setup with redirect server

The signaling messages in Figures 1 and 2 are generally sent via UDP, although SIP also supports TCP [13]. An application-level timeout and retransmission scheme recovers from UDP errors. An INVITE message is retransmitted until the first provisional response is receiver, first after 500 ms, then again after one additional second, then two seconds and finally every four seconds[2]. INVITE message transmissions cease after seven attempts. The server simply transmits a provisional response for each INVITE received, without any timers. When the call is answered, redirected or fails for some reason, the UAS transmits a *final response*, as shown in the top part of Fig. 2. The final response is retransmitted with the same spacing as the INVITE, until the caller sends an ACK message. The post-pickup delay is determined by how long it takes for the first final response to reach the caller.

3.2 H.323

Figure 3 shows the simplest H.323 call setup, the Fast Connect option available in H.323v2. A wide variety of other H.323 scenarios are possible [12], some using a gatekeeper for address resolution, connection admission control and call signaling. A general comparison of H.323 and SIP appears in [23]. For this study, we consider the Fast Connect option only, the aim being to highlight the delay differences between this call type and the corresponding SIP one. A significantly more involved call setup mechanism requiring several TCP connection is specified in the 1998 version of [12].

As these differences arise principally from the use of TCP in H.323, we review briefly TCP connection setup and error control. TCP connection setup requires an exchange of SYN messages, followed by an ACK to complete the three way handshake, as shown in Fig. 3. Data transfer then begins, which, in this case, comprises the H.323 SETUP/CONNECT message exchange. The SYN and SETUP/CONNECT messages time out if not acknowledged. The timeout value increases exponentially, usually by a factor of two, each time a given message is retransmitted. TCP timeout values are generally determined by the measured round trip delay and delay variance. However, the default TCP timeout values apply for new connections, such as those shown here.

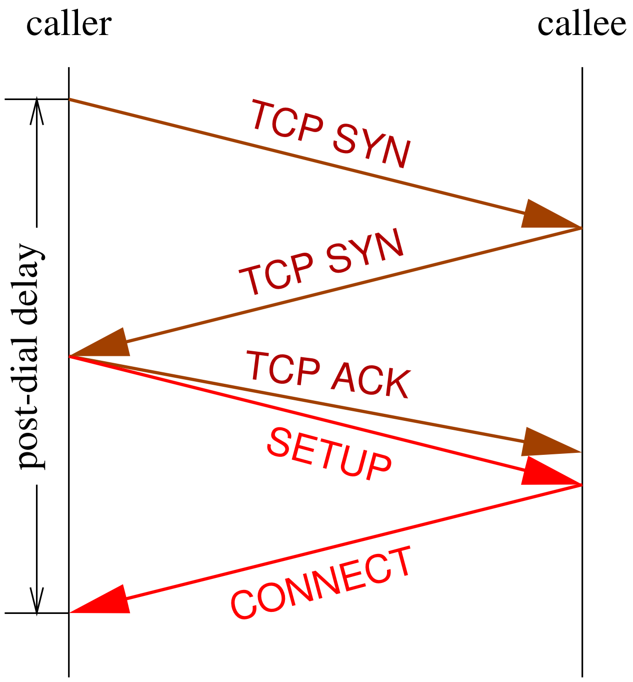

Figure 3: H.323 Fast Connect call setup

[2] The value of 500 ms was chosen since it represents a reasonable upper bound for interactive voice communications.

RFC 1122 [24] specifies an initial timeout of three seconds[3], however some implementations start at six seconds [25]. Either value is too high for Internet telephony.

Concerns about the suitability of TCP error control for signaling are raised in [26], which highlights delays arising from TCP timer granularities (generally set to 500 ms). The TCP delayed acknowledgment mechanism, designed to reduce traffic loads, adds more delay [27]. Clearly TCP needs tuning for Internet telephony signaling, with lower initial timeout values, such as the SIP ones, and immediate acknowledgments. Some systems, such as Solaris using `ndd`, allow the system-wide tuning of these parameters, while others, such as Winsock, do not. To avoid biasing the results by operating system settings, we assume these lowered initial timeout values.

In many cases, H.323 uses *gatekeeper-routed* signaling, where signaling is propagated along a chain of TCP connection from client through one or more gatekeepers to the destination. Since the number of such gatekeepers are hard to predict, we ignore them in our model. Also, if these gatekeepers exchange signaling messages regularly, they may already have an existing TCP connection, so that they can avoid the TCP connection setup overhead. Our model would then apply.[4]

4 Internet Delay Traces

Packet traces have received much attention in recent years. The major focus has been on interarrival distributions, which have displayed long range dependencies which are at odds with traditional traffic models [28]. The scope of trace results has been extended by the IETF IP Performance Measurement Working Group (IPPM), which has developed metrics and techniques for one way delay and loss measurements [29]. These techniques underpin the Surveyor project [30], run by Advanced Networks and Services, which provides the Internet delay and loss statistics used in this paper.

The Surveyor project, which began in 1997, provides continuous monitoring of UDP delay and loss between selected sites. There are currently 38 of these, mostly in the USA, with some in Europe and the Pacific region. UDP packets of 40-byte length are sent at exponentially distributed intervals with a mean of 500 ms [31]. Using GPS receivers for synchronization, the receiver measures the one way delay with 50 μs resolution, while the packet headers allow loss detection. Results are collated at Advanced Networks, with delay and loss histograms available at the Web site. We have used the individual trace results for our simulations, which essentially provide a delay sample or loss indication every 500 ms.

5 Simulation Methodology

The simulation estimates the call setup delay distribution experienced by SIP and H.323 system operating over the public Internet. The SIP and H.323 call setup message exchanges outlined previously are modelled, using the Internet delay and loss figures gathered by the Surveyor project. Unfortunately, we cannot simply map signaling requests to a corresponding Surveyor sample, since the spacing of the Surveyor samples is not uniform, with additional gaps due to packet losses. We approximate the network delay behavior by assuming that the instantaneous UDP delay is the one experienced by the most recent Surveyor sample corresponding to the simulated transmission time of the SIP or H.323 request or response. If this sample was lost, then the most recent delay sample before that is used.

[3]The Solaris operating system, for example, uses a three-second time out.

[4]The authors are not sure whether H.323 connection re-use is possible or implemented in practice. The only constraint on the Call Signaling Channel is that it "shall be a reliable channel" [12, p. 38]). H.323 does imply that a separate connection will be used per call, by saying that "For the Gatekeeper Routed method, the Gatekeeper may choose to close the Call Signaling Channel, or it may keep it open for the duration of the call to support supplementary services. Only the Gatekeeper shall close the Call Signaling Channel and it should not be closed when a Gateway is involved in the call." (p. 38)

The simulation aims to capture the effect of UDP burst errors. A two-state error model is used, which operates as follows: The number of UDP packet losses in the last 200 samples (E200) and the last 20 samples (E20) is recorded. This corresponds to the previous 100 seconds and 10 seconds, respectively. If the number of errors in the last 20 samples is zero or one, then a "good" error state is assumed. The UDP error probability used in the simulation is E200/200, the mean error rate over the previous 100 seconds. Otherwise, a "bad" error state is assumed, where the error probability is E20/20, the mean error rate over the last 10 seconds.

While two-state error models are commonly used, the heuristic presented here and the values chosen are arbitrary. The results which follow test the sensitivity of this error model.

The simulation uses these extrapolated UDP delay and loss statistics to determine the respective H.323 and SIP call setup delay distributions. We assume that the UDP statistics can be applied to TCP packets. The H.323 results consider the Fast Connect call setup shown in Fig. 3. The SIP results encompass a simple call setup (as in Fig. 1), and a redirect server interaction, followed by a simple call setup, as in Fig. 2). For the SIP calls, the path between the client and user agent server (UAS) may traverse multiple stateless proxies.

Delay distributions are generated as if calls were made over one hour, on a specified day, according to the Surveyor data. A one hour delay distribution is based on results from thirty two independent one hour simulations, each with 180,000 simulated calls.

6 Results

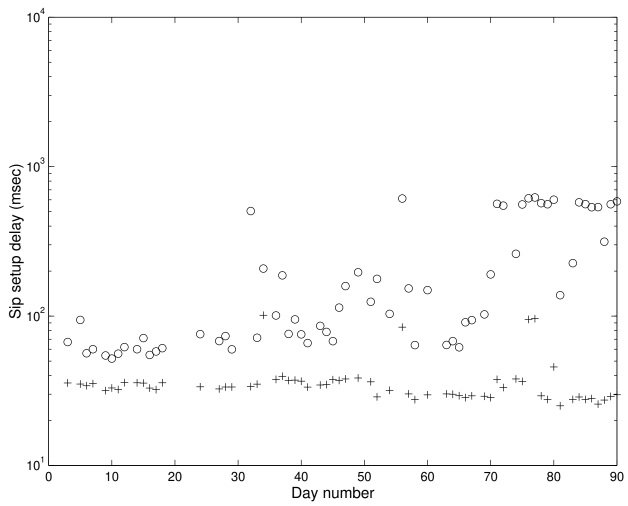

Figure 4: Minimum and 95th percentile SIP setup delay, New York → Boston

The scope of the Surveyor database allows the gathering of results over extended periods, providing insight into average and worst case performance of H.323 and SIP. In particular, we highlight TCP delays in H.323, and investigate delay increases arising from more complex SIP calls which incorporate redirect servers and intermediate proxies.

Similar to circuit switched network engineering, we identify an Internet telephony "busy hour". While a detailed analysis of Internet busy hours is beyond the scope of this paper, the Surveyor trace histograms show 16:00 hours (Eastern Time) to be a reasonable busy hour choice. The results here cover the first 90 business days of 1999, and consider one hour each day, starting at 16:00 hours.

We investigate three paths within the USA. All originate at the Advanced Networks headquarters in New York, and extend to either Boston (Harvard University), Chicago (University of Chicago) or the West Coast

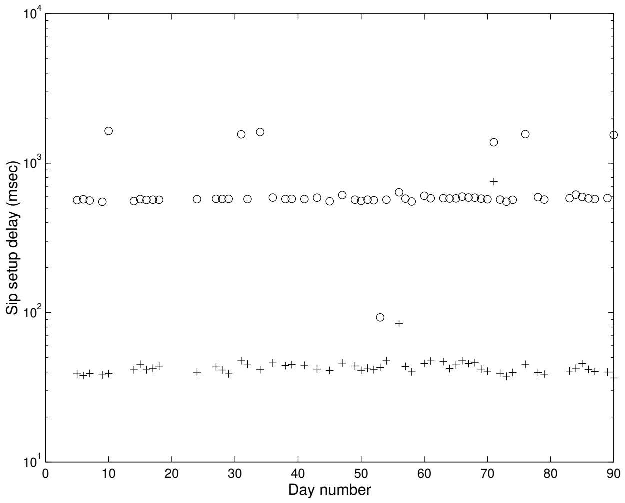

Figure 5: Minimum and 95th percentile SIP setup delay, New York → Chicago

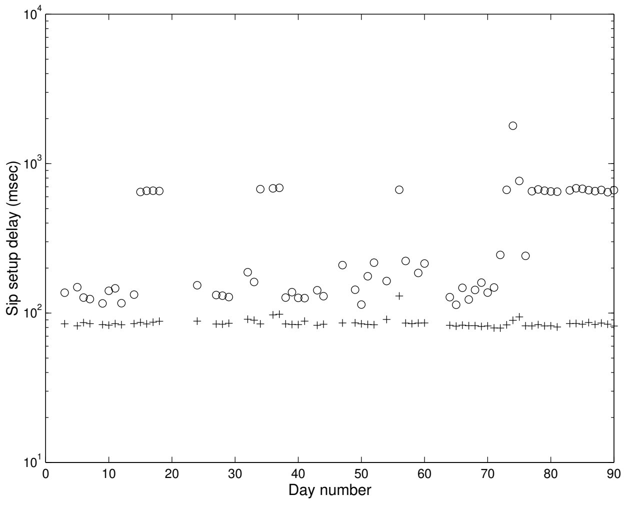

Figure 6: Minimum and 95th percentile SIP setup delay, New York → West Coast

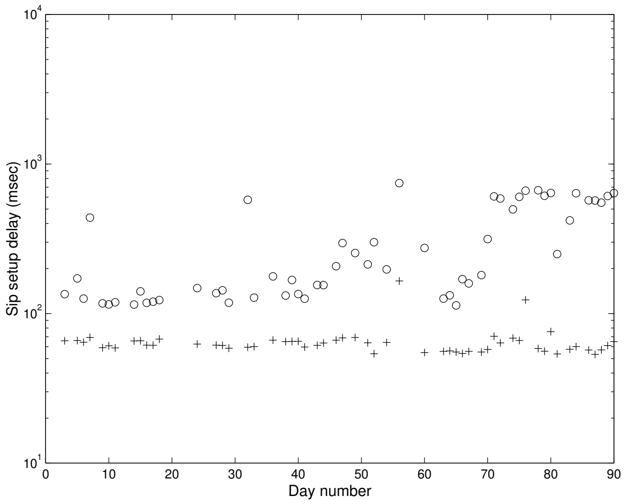

Figure 7: Minimum and 95th percentile SIP setup delay, New York → Boston, redirected in Washington, D.C.

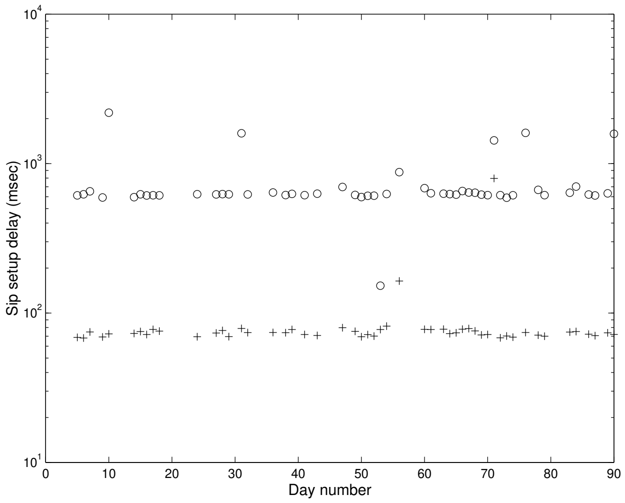

Figure 8: Minimum and 95th percentile SIP setup delay, New York → Chicago, redirected in Washington, D.C.

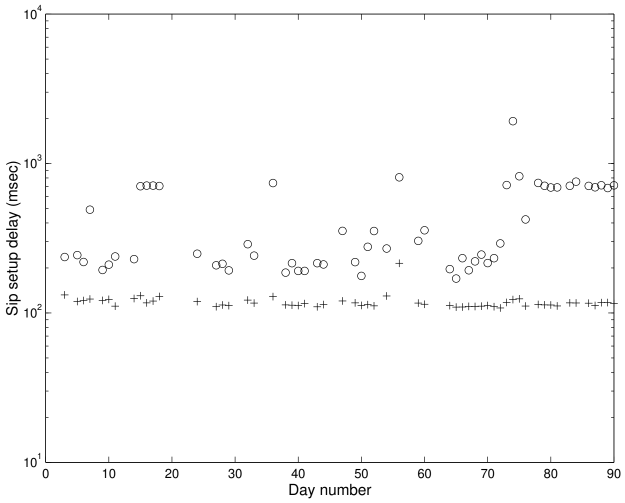

Figure 9: Minimum and 95th percentile SIP setup delay, New York → West Coast, redirected in Washington, D.C.

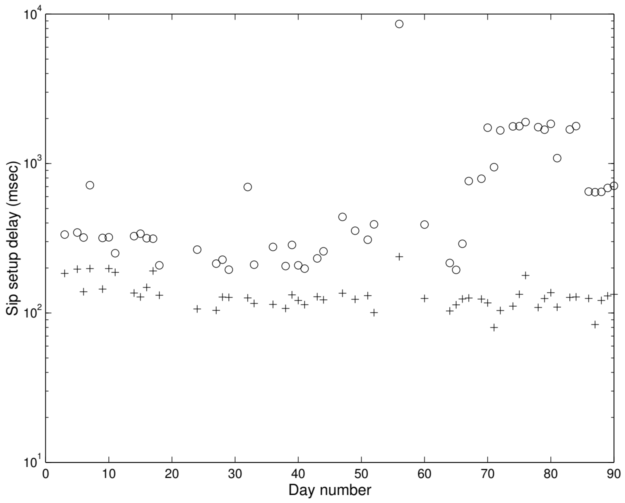

Figure 10: Minimum and 95th percentile SIP setup delay, New York → Chicago

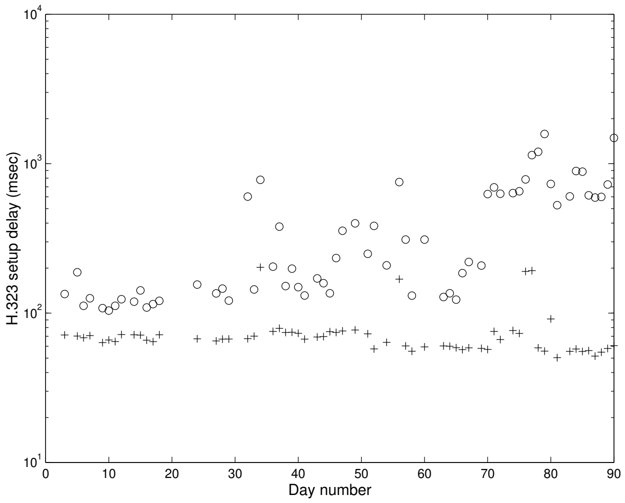

Figure 11: Minimum and 95th percentile H.323 setup delay, New York → Boston

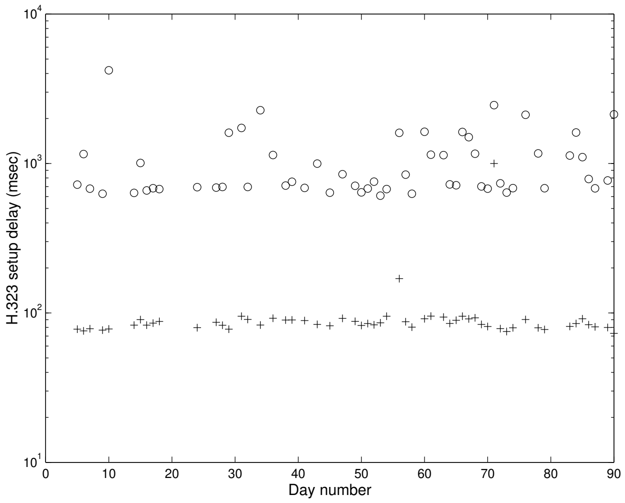

Figure 12: Minimum and 95th percentile H.323 setup delay, New York → Chicago

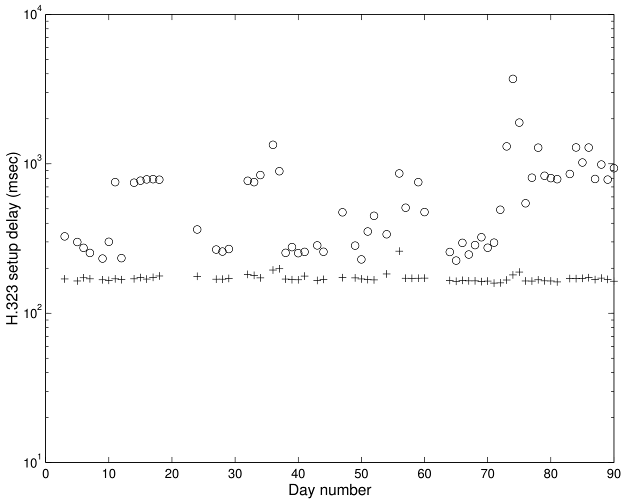

Figure 13: Minimum and 95th percentile H.323 setup delay, New York → West Coast

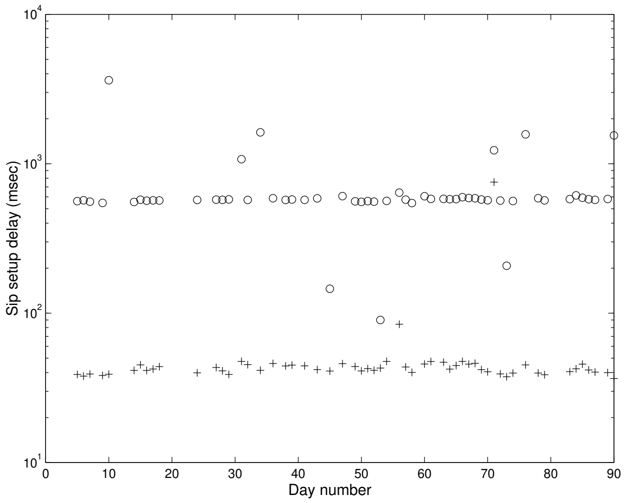

Figure 14: Minimum and 95th percentile SIP call setup delay, New York → Chicago, with error window of 4 samples

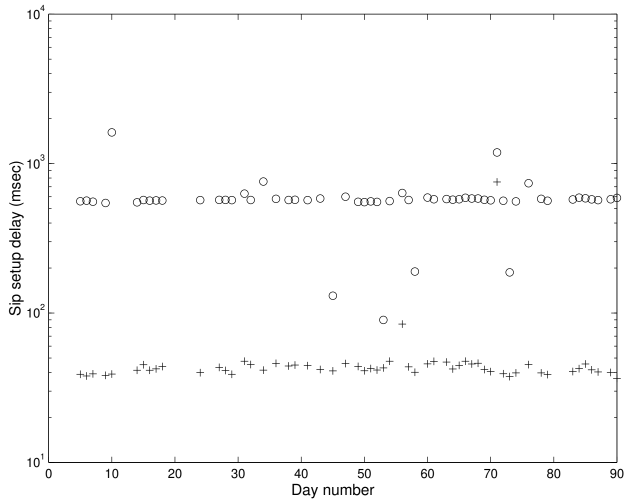

Figure 15: Minimum and 95th percentile SIP call setup delay, New York → Chicago, with average error

(NASA AMES near Sunnyvale, California). They are 306, 1158 and 4128 km away from New York, with one-way propagation delays of 1.5, 5.8 and 20.6 ms, respectively.

We consider the following scenarios:

- A one-hop SIP call setup over each of the three paths, that is, an INVITE/provisional response exchange between the source and destination (Fig. 4 to 6).
- A SIP call over each path, which first queries a redirect server at George Washington University, in Washington, D.C. (328 km from New York), then exchanges an INVITE/provisional response, as before (Fig. 7 to 9).
- Fig. 10 extends the SIP call in Figure 9, by passing the INVITE/provisional response messages through a stateless proxy in Boston.
- Fig. 11 to Fig. 12 consider a one-hop H.323 call setup ("fast connect"), and represent the H.323 equivalent of the SIP calls in figures 4 to 6. The difference is the H.323 TCP connection establishment, shown in Figure 3. The TCP timeout values used in the simulation are the same as the SIP ones and thus do not represent the much longer values likely to be encountered by "stock" operating systems.

The plots show, for each day, the minimum call setup delay (a "+") and the 95th delay percentile (an "o").

A key aim of this study has been to investigate the effect of UDP burst errors, as captured by the Surveyor traces. As indicated, the simulation maintains two error windows, which indicate "good" and "bad" error states. As mentioned earlier, the default window sizes are 200 samples and 20 samples respectively. The sensitivity to these parameters is tested in figures 14 and 15, which repeat the New York/Chicago SIP call from Fig. 5. Figure 14 keeps the "good" window at 200 samples, but reduces the "bad" window size to four samples ($\equiv$ 2 seconds). Fig. 15 measures the mean UDP error rate over the entire hour, and uses that value for all calls, ignoring error correlation.

Section 2 indicated a hard call setup delay limit of 2 seconds for PSTN/Internet telephony interworking since ISDN switches abandon call attempts which exceed this limit. If we require that no more than 1% of such calls fail during the busy hour, we need to ensure that the 99th delay percentile is below 2 seconds. As the simulation results represent signalling delays only, we assume a constant delay of 500 ms for other call processing tasks such as DNS lookups, resulting in a delay budget of 1.5 s. Table 1 shows the percentage of days for which this limit is not achieved, for one-hop SIP and H.323 calls between the destinations shown. (The H.323 figures are shown below the corresponding SIP figure.)

7 Discussion

Due to the high speed links used in these paths (T3 and above), the minimum delay results essentially show the round trip message propagation delay [31]. Figures 4 to 6 show this increasing from around 30 ms for New York – Boston to 40 ms for New York – Chicago, and around 80 ms for New York – West Coast.

The 95th percentile results in Figures 4 to 6 show two causes of increasing delays, namely increased queueing delay and retransmissions. The queueing delay is essentially the difference between the mininum delay, i.e., the propagation delay, and the associated 95th percentile. Figures 4 to 6 show this difference to be between 100 and 200 ms.

In Figures 4 to 6, the total delays are mostly around 600 ms, due to the 500 ms SIP initial timeout, propagation and queueing delays. In Figures 5 and 6, some 95th percentile results are around 1.6 s. Here, two timeouts occur in the same message exchange, a 500 ms timeout, then a 1 s timeout. The Chicago route experiences the most errors, with almost all days showing 95th percentiles of around 600 ms.

Figures 7 to 9 add the SIP redirect server interaction, increasing the number of one-way paths from two to four. This produces a 30 ms increase in the minimum delay. The 95th percentile results due to queueing delay only (i.e., those less than 500 ms) roughly double. This is essentially the convolution of the queueing delays on each path. However, the 95th percentile results due to retransmissions (those above 500 ms) are largely unchanged from those in Figures 4 to 6.

The retransmission timeout delay clearly dominates the 95th percentile results. Figure 10 investigates this delay further, by routing the New York-West Coast messages via Boston. Comparing these results with Figure 9 (i.e., without the Boston leg), we see increases in the minimum delay of around 30 to 50 ms. However the 95th percentile results are worse, with many delays in the range of 1.6 s. This more complex SIP call, while not greatly increasing the minimum delay, worsens the 95th delay percentile, and hence the perceived call setup delay QOS.

Clearly the number of paths traversed during a call setup determines the message loss probability, and hence the delay due to retransmissions. In this context we consider the H.323 results in Figures 11 to 12. These essentially repeat the SIP results in Figures 4 to 6, with an additional message exchange for the TCP connection setup. Hence we see a doubling of the minimum delay. Figures 11 and 12 show the 95th percentile results arising from retransmission timeouts (i.e., those around 600 ms) to be largely the same as the SIP ones in Figures 4 and 6. Fig. 12, the New York–Chicago route, however shows substantially worse 95th percentile results than the corresponding SIP ones in Fig. 5. In particular, the number of days with a 95th percentile of one second or above (indicating multiple timeouts in the message exchanges) increases by a factor of four.

Delay targets for Internet telephony signalling transport have yet to be established, as indicated. However a 95th percentile of less than a second appears reasonable (and lies well within the ranges outlined in [11]). This target is achieved for almost all the simple SIP call setups (Figures4 to 6). The H.323 calls also achieve this limit on some paths. However the TCP delays shown here are "best case", with timeout values less than the default ones, and without the additional delays arising from timer granularity. As Fig. 10 shows, call setup delay targets are less likely to be achieved over the public Internet for more complex SIP and H.323 call types.

[TABLE 1: HOW DID COLORADO GET IN THERE? IT'S NOT MENTIONED ELSEWHERE...]

While these results indicate delay trends, they do not measure availability. Gaps in the results, particularly in Fig. 10, are due, in part, to a lack of Surveyor results for those days. The simulation also dropped

	Bos.	Chi.	West	Wash.	Col.
New York	20.3	77.2	32.3	9.1	15.4
	28.2	94.7	40.0	20.0	18.5
Boston		1.6	31.5	0.0	5.4
		1.6	31.5	0.0	10.8
Chicago			34.3	5.2	28.6
			34.3	6.9	61.4
West Coast				33.3	45.3
				36.7	57.3
Washington State					6.6
					6.6

Table 1: Percentage of days where PSTN/Internet telephony blocking probability exceeds 1%, for SIP (top row) and H.323 (bottom row)

days which exhibited

- a gap of more than 5 seconds between trace files, or
- a gap of more than 60 seconds between trace records.

The aim was to avoid biasing results by including gaps which may have arisen from faults in the measuring equipment, rather than actual path unavailability. Hence the results here apply to "good" days, with continuous path availability. It is possible that the delay results from some of the missing days are far worse than the ones shown here. The next stage of this project will investigate Internet telephony availability.

Figures 14 and 15 test the sensitivity of the error model outlined in Section 5. Moving the "bad state" error window from 20 samples (Fig. 5) to four samples (Fig. 14) produces almost no change in the results. Fig. 15 ignores error bursts, by using the mean error rate over the entire hour for all calls. While the Figure 15 results are similar to the Fig. 5 ones, the "bad" error days (i.e., with a 95th percentile around 1.6 seconds) are not detected. Hence, while the two-state error model shows bursty error effects, as desired, the results appear to be insensitive to the window size chosen.

Table 1 shows that, for PSTN/Internet Telephony interworking, reasonable blocking targets (here, 1%) are not likely to be achieved. The table increases the number of samples by including paths to Colorado. The majority of the source/destination pairs show many days (more than 20%) when the 1% blocking probability is not reached. The H.323 results are generally worse than the SIP ones, due to the TCP connection setup delays. However, neither SIP nor H.323 provide satisfactory performance. Hence, from the perspective of blocking probability, the best-effort Internet appears not well suited for PSTN interworking. For this application, dedicated IP signalling capacity is more appropriate.

8 Conclusions

This paper has considered Internet call setup delays, focusing on the delay component arising from signalling transport. While initial Internet telephony call setup delay targets have been proposed elsewhere, individual targets for the signalling component are still needed.

Drawing on delay and loss traces from the public Internet, our simulation study has shown that, for the paths considered, acceptable SIP call setup delay is available for simple call types. More complex SIP calls, which traverse multiple paths, display variable delay performance. Our results show that the TCP connection setup associated with H.323 calls substantially increases call setup delay over errored paths, even after tuning TCP implementations for more rapid retransmission.

If large Internet telephony gateways dominate Internet telephony, the number of signaling paths for each such gateway will likely be small. In those cases, substantially better signaling performance can be achieved if retransmission timers are tuned, based on previous calls, for the round-trip delays to the destination or, if the request interarrival rate is less than a third of the SIP time out value, by using TCPs fast retransmit.

While acceptable call setup delay performance is at times available over the public Internet, our results show that unacceptable blocking rates are likely when interconnecting with the PSTN.

Acknowledgements

The authors would like to thank the staff at Advanced Networks and Services, and in particular Sunil Kalindindi, for providing access to the data used in this project.

References

[1] D. Clark, "A taxonomy of internet telephony applications," in *Proc. of 25th Telecommunications Policy Research Conference*, (Washington, DC), Sept. 1997.

[2] C. A. Polyzois, K. H. Purdy, P.-F. Yang, D. Shrader, H. Sinnreich, F. Mnard, and H. Schulzrinne, "From POTS to PANS – a commentary on the evolution to internet telephony," *IEEE Network*, Vol. 13, pp. 58–64, May/June 1999.

[3] B. Gherardi and F. B. Jewett, "Telephone communication system of the united states," *Bell System Technical Journal*, Vol. 9, pp. 1–100, Jan. 1930.

[4] F. P. Duffy and R. A. Mercer, "A study of network performance and customer behavior during-direct-distance-dialing call attempts in the USA," *Bell System Technical Journal*, Vol. 57, no. 1, pp. 1–33, 1978.

[5] AT&T, "AT&T sets the industry standard for network reliability," Mar. 1998. http://www.att.com/network/standrd.html.

[6] International Telecommunication Union, "Network grade of service parameters and target values for circuit-switched services in the evolving isdn," Recommendation E.721, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, May 1999.

[7] V. A. Bolotin, P. J. Kuhn, C. D. Pack, and R. A. Skoog, "Common channel signaling networks: Performance, engineering, protocols and capacity management," *IEEE Journal on Selected Areas in Communications*, Vol. 12, pp. 377–544, Apr. 1994. Special issue.

[8] International Telecommunication Union, "Message transfer part signalling performance," Recommendation Q.706, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, Mar. 1993.

[9] International Telecommunication Union, "Signalling performance in the telephone application," Recommendation Q.725, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, Mar. 1993.

[10] R. A. Skoog, "Engineering common channel signaling networks for ISDN," in *Twelfth International Teletraffic Congress*, Vol. 2, (Torino), pp. 1–7 (2.4A,), June 1988.

[11] H. Lin, T. Seth, A. Broscius, and C. Huitema, "VoIP signaling performance requirements and expectations," Internet Draft, Internet Engineering Task Force, June 1999. Work in progress.

[12] International Telecommunication Union, "Visual telephone systems and equipment for local area networks which provide a non-guaranteed quality of service," Recommendation H.323, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, May 1996.

[13] M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg, "SIP: session initiation protocol," Request for Comments (Proposed Standard) 2543, Internet Engineering Task Force, Mar. 1999.

[14] H. Schulzrinne and J. Rosenberg, "Internet telephony: Architecture and protocols – an IETF perspective," *Computer Networks and ISDN Systems*, Vol. 31, pp. 237–255, Feb. 1999.

[15] H. Schulzrinne and J. Rosenberg, "The session initiation protocol: Providing advanced telephony services across the internet," *Bell Labs Technical Journal*, Vol. 3, pp. 144–160, October-December 1998.

[16] P. Goyal, A. Greenberg, C. Kalmanek, B. Marshall, P. Mishra, D. Nortz, and K. K. Ramakrishnan, "Integration of call signaling and resource management for ip telephony," *IEEE Network*, Vol. 13, pp. 24–33, May/June 1999.

[17] T. J. Kostas, M. S. Borella, I. Sidhu, G. M. Schuster, J. Grabiec, and J. Mahler, "Real-time voice over packet-switched networks," *IEEE Network*, Vol. 12, pp. 18–27, Jan. 1998.

[18] N. F. Maxemchuk and S. Lo, "Measurement and interpretation of voice traffic on the internet," in *Conference Record of the International Conference on Communications (ICC)*, (Montreal, Canada), June 1997.

[19] A. I. Elwalid, G. G. Freundlich, P. M. Gerhardt, H. Hagirahim, K. G. Ramakrishnan, and D. Tse, "An overview of the multimedia communications exchange (mmcx) and its performance characterization," *Bell Labs Technical Journal*, Vol. 2, Spring 1997.

[20] J. Rosenberg, J. Lennox, and H. Schulzrinne, "Programming internet telephony services," Technical Report CUCS-010-99, Columbia University, New York, New York, Mar. 1999.

[21] International Telecommunication Union, "Methods for dimensioning resources in signalling system no. 7 networks," Recommendation E.733, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, 1988 Nov.

[22] Bellcore, "Lssgr: Switching system generic requirements for call control using the integrated services digital network user part (isdnup)," Tech. Rep. GR-317-CORE, Bellcore, Morristown, New Jersey, Dec. 1997. Issue 2.

[23] H. Schulzrinne and J. Rosenberg, "A comparison of SIP and H.323 for internet telephony," in *Proc. International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, (Cambridge, England), July 1998.

[24] R. T. Braden, "Requirements for internet hosts - communication layers," Request for Comments (Standard) 1122, Internet Engineering Task Force, Oct. 1989.

[25] W. R. Stevens, *TCP/IP illustrated: the implementation*, Vol. 2. Reading, Massachusetts: Addison-Wesley, 1994.

[26] T. Seth, A. Broscius, C. Huitema, and H. Lin, "Performance requirements for signaling in internet telephony," Internet Draft, Internet Engineering Task Force, Nov. 1998. Work in progress.

[27] M. Allman, "On the generation and use of TCP acknowledgments," *ACM Computer Communication Review*, Vol. 28, pp. 4–21, Oct. 1998.

[28] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, "The changing nature of network traffic: Scaling phenomena," *ACM Computer Communication Review*, Vol. 28, pp. 5–29, Apr. 1998.

[29] V. Paxson, G. Almes, J. Mahdavi, and M. Mathis, "Framework for IP performance metrics," Request for Comments (Informational) 2330, Internet Engineering Task Force, May 1998.

[30] Advanced Networks and Services, "The Surveyor Project home page," 1999. http://www.advanced.org/csg-ipmm.

[31] S. Kalidindi and M. J. Zekauskas, "Surveyor: An infrastructure for internet performance measurements," in *Proc. of INET*, (San Jose, California), June 1999.