Performance Measurement and Analysis of H.323 Traffic *

Prasad Calyam[1], Mukundan Sridharan[2], Weiping Mandrawa[1], and Paul Schopis[1]

[1] OARnet, 1224 Kinnear Road,Columbus, Ohio 43212.
{pcalyam,wmandraw,pschopis}@oar.net
[2] Department of Computer and Information Science,
The Ohio State University, Columbus, OH 43210.
sridhara@cis.ohio-state.edu

Abstract. The popularity of H.323 applications has been demonstrated by the billions of minutes of audio and video traffic seen on the Internet every month. Our objective in this paper is to obtain Good, Acceptable and Poor performance bounds for network metrics such as delay, jitter and loss for H.323 applications based on objective and subjective quality assessment of various audio and video streams. To obtain the necessary data for our analysis we utilize the H.323 Beacon tool we have developed and a set of Videoconferencing tasks performed in a LAN and also with end-points located across multiple continents, connected via disparate network paths on the Internet.

1 Introduction

H.323 [1] is an umbrella standard that defines how real-time multimedia communications, such as audio and video-conferencing, can be exchanged on packet-switched networks (Internet). With the rapid increase in the number of individuals in industry and academia using H.323 audio and video-conferencing systems extensively, the expectation levels for better audio and video performance have risen significantly. This has led to the need to understand the behavior of audio and video traffic as it affects end user perceived quality of the H.323 applications over the Internet. Several studies have been conducted [2–4] and many approaches [5–7] have been proposed to determine the performance quality measures of H.323 applications. Many of the previous studies used pre-recorded audio and video streams and aimed at obtaining quality measures either based solely on network variations or on various audiovisual quality assessment methods.

Our primary focus in this paper is to understand how the various levels of network health, characterized by measuring delay, jitter and loss, can affect end user perception of audiovisual quality. By systematically emulating various network health scenarios and using a set of Videoconferencing 'Tasks' we determine performance bounds for delay, jitter and loss. The obtained performance bounds

* This work was supported in part by The Ohio Board of Regents and Internet2

are mapped to end-users perceptions of the overall audiovisual quality and are then categorized into Grades such as 'Good', 'Acceptable' and 'Poor'. We show that end-users are more sensitive to variations in jitter than variations in delay or loss. The results of this paper could provide ISPs and Videoconferencing Operators a better understanding of their end-user's experience of audiovisual quality for any given network health diagnostics.

To obtain the necessary data to support the various conclusions in this paper, we utilized the H.323 Beacon tool [8] we have developed and a set of Videoconferencing tasks. Over 500 one-on-one subjective quality assessments from Videoconferencing Users and the corresponding H.323 traffic traces were collected during our testing, which featured numerous network health scenarios in an isolated LAN environment and on the Internet. The collected traces provided objective quality assessments. The Internet testing involved 26 Videoconferencing end-points; each performing 12 Videoconferencing tasks and located across multiple continents, connected via disparate network paths that included research networks, commodity networks, cable modem connections, DSL modem connections and Satellite networks.

The rest of the paper is organized as follows: Section 2 provides a background pertaining to this paper, Section 3 describes our testing methodology, Section 4 discusses our analysis of the performance bounds for delay, jitter and loss and Section 5 concludes the paper.

2 Background

2.1 H.323 System Architecture

There are numerous factors that affect the performance assessments of H.323 applications. These factors can be subdivided into 3 categories: 1. Human factors 2. Device factors 3. Network factors. First, Human factors refer to the perception of quality of the audio and video streams and also the human error due to negligence or lack of training which results in performance bottlenecks which then affect the performance assessments. Secondly, essential devices such as H.323 terminals, Multipoint Control Units (MCUs), gatekeepers, firewalls and Network Address Translators (NATs) frequently contribute towards performance degradations in H.323 systems. Thirdly, the network dynamics caused by route changes, competing traffic and congestion cause performance bottlenecks that affect performance assessments. In this paper we are interested in studying aspects of the Human factors, which deal with end-user perception of audiovisual quality and the Network factors, which contribute to any network's health. The reader is referred to [9] for details related to Device factors.

2.2 Audiovisual Quality Assessment Metrics

There are two popular methods to assess audiovisual quality: Subjective quality assessment and Objective quality assessment. Subjective quality assessment involves playing a sample audiovisual clip to a number of participants. Their judgment of the quality of the clip is collected and used as a quality metric. Objective

Fig. 1. Voice Quality Classes

quality assessment does not rely on human judgment and involves automated procedures such as signal-to-noise ratio (SNR) measurements of original and reconstructed signals and other sophisticated algorithms such as Mean Square Error (MSE) distortion, Frequency weighted MSE, Segmented SNR, Perceptual Analysis Measurement System (PAMS) [10], Perceptual Evaluation of Speech Quality (PESQ) [11], and Emodel [5], to determine quality metrics. The problem with subjective quality assessment techniques is that human perception of quality is based on individual perception, which can vary significantly between a given set of individuals. The problem with objective quality assessment techniques is that they may not necessarily reflect the actual end-user experience. There have been studies [12] that show that when objective and subjective quality assessment are performed simultaneously, the results are comparable.

In our study, we employ both the subjective and objective quality assessment methods to determine end-user perception of audiovisual quality for various network health scenarios. To obtain subjective quality assessment scores from the participants, we extended the slider methodology presented in [7] and developed our own slider that was integrated into our H.323 Beacon tool. Participants ranked the audiovisual quality on a scale of 1 to 5 for various Videoconferencing tasks using what is basically the Mean Opinion Score (MOS) ranking technique. To obtain objective quality assessment scores we utilized the Telchemy VQMon tool [12] that implements the Emodel and uses traffic traces obtained for the various Videoconferencing tasks as an input for its analysis. The Emodel is a well established computational model that uses transmission parameters to predict the subjective quality. It uses a psycho-acoustic R-scale whose values range from 0 to 100 and can be mapped to MOS rankings and User Satisfaction as shown in Fig. 1. Though the Emodel fundamentally addresses objective quality assessment of voice, our collected data shows reasonable correlation of the subjective quality assessment scores for audiovisual quality provided by the participants and the objective quality assessment scores provided by VQMon. The reader is referred to [2, 5, 12] for more details relating to Emodel components.

2.3 Network Performance Metrics

The variables that affect the MOS rankings the most in H.323 system deployments are the dynamic network changes caused by route fluctuations, competing traffic and congestion. The network dynamics can be characterized by 3 network metrics viz. delay, jitter and loss as specified in [13].

Delay is defined as the amount of time that a packet takes to travel from the sender's application to the receiver's destination application. The components that contribute to the end-to-end delay include: (a) compression and transmission delay at the sender (b) propagation, processing and queuing delay in the network and (c) buffering and decompression delay at the receiver. The value of one-way delay needs to be stringent for H.323 audio and video traffic to sustain good interaction between the sender and receiver ends. It is recommended by [14] that delay bounds for the various grades of perceived performance in terms of human interaction can be defined as: Good (0ms-150ms), Acceptable (150ms-300ms), Poor ($> 300ms$).

Jitter is defined as the variation in the delay of the packets arriving at the receiving end. It is caused due to congestion at various points in the network, varying packet sizes that result in irregular processing times of packets, out of order packet delivery, and other such factors. Excessive jitter may cause packet discards or loss in the playback buffer at the receiving end. The playback buffer is used to deal with the variations in delay and facilitate smooth playback of the audio and video streams. There have been some indications in [15] about jitter bounds, which have been verified to be approximately correct in our earlier work [9] and are also supported by the studies conducted in this paper. However, there have not been well defined rules of thumb to suggest the accurate jitter bounds in terms of the various grades of H.323 application performance. Our studies suggest the following jitter values to be reasonably reliable estimates to determine the grade of perceived performance: Good (0ms-20ms), Acceptable (20ms-50ms), Poor ($> 50ms$).

Lastly, loss is defined as the percentage of transmitted packets that never reach the intended destination due to deliberately discarded packets (RED, TTL=0) or non-deliberately by intermediate links (layer-1), nodes (layer-3) and end-systems (discards due to late arrivals at the application). Though popular experience suggests loss levels greater than 1% can severely affect audio-visual quality, there have not been well defined loss bounds in terms of the various grades of H.323 application performance. Our studies in this paper suggest the following loss values to be reasonably reliable estimates to determine the grade of perceived performance: Good (0%-0.5%), Acceptable (0.5%-1.5%), Poor ($> 1.5\%$).

3 Test Methodology

3.1 Design of Experiments

Our approach in determining the performance bounds for delay, jitter and loss in terms of Good, Acceptable and Poor grades of performance, is to view the

network health as an outcome of the combined interaction of delay, jitter and loss. Our reasoning is that all three network parameters co-exist for every path in the Internet at any given point of time; regulating any one of these parameters affects the other parameters and ultimately the Quality of Service (QoS) perceived by the end-user in terms of H.323 application performance. A real-world example is illustrated in [16] where, resolving a loss problem in an Intercampus DS3 led to a decrease in the observed loss but unexpectedly led to an increase in the overall jitter levels in the network. This shows that network health needs to be sustained in a stable state where the network delay, jitter and loss are always within the Good performance bounds. In our design of experiments, we employ a full factorial design for the 3 factors, i.e. we emulate 27 scenarios that cover every possible permutation involving the various delay, jitter and loss levels.

We performed extensive LAN tests that covered all of the 27 possibilities and selected 9 scenarios shown in Table 1 of Section 4 for Internet measurements whose results in essence reflected the results of the 27 scenarios. The reason to use the 9 scenarios is that it is impractical and non-scalable to have each participant judging quality 27 times in a single subjective assessment test session.

For each of the 9 Internet test scenarios a Videoconferencing task was assigned. A Videoconferencing task could be any activity that takes place in a routine Videoconference. A casual conversation, an intense discussion, or a class lecture would qualify as a Videoconferencing task. There is significant literature recommending strategies for tasks that could be part of Subjective and Objective assessments of audiovisual quality [6, 7]. All of them recommend that in addition to passive viewing for assessments of audiovisual quality, the participants must be presented with realistic scenarios. Key guidelines proposed in the above literature were followed in task creation, participant training for scoring the audiovisual quality, task ordering and overall environment setup for the assessment. A subset of the Videoconferencing tasks performed by the test participants involved the audio and video loop back feature of the H.323 Beacon tool. The loopback feature enables local playback of remote H.323 Beacon client audio or video recorded at a remote H.323 Beacon server. The reader is referred to [8] for more details relating to the H.323 Beacon.

3.2 Test Setup

To obtain the performance bounds for delay, jitter and loss and to affirm our conclusions, we chose a two phase approach. In the first phase we performed extensive testing by emulating all the 27 scenarios in a LAN environment and obtained the performance bounds for delay, jitter and loss as stated in Section 2.3. In the second phase, we used the 9 scenarios described in Section 3.1 and performed the Internet tests. In both the phases of testing, we conducted one-on-one testing with participants at each of the LAN/Internet sites and collected traffic traces and objective and subjective quality assessments. Fig. 2 shows the overall test setup and Fig.3 shows the participating sites in the testing and their last mile network connections. NISTnet [17] network emulator was used to create the various network health scenarios by introducing delay, jitter and loss

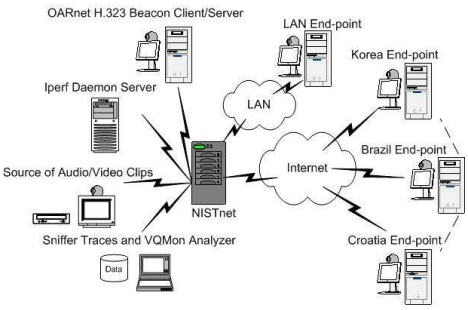

Fig. 2. Overall Test Setup

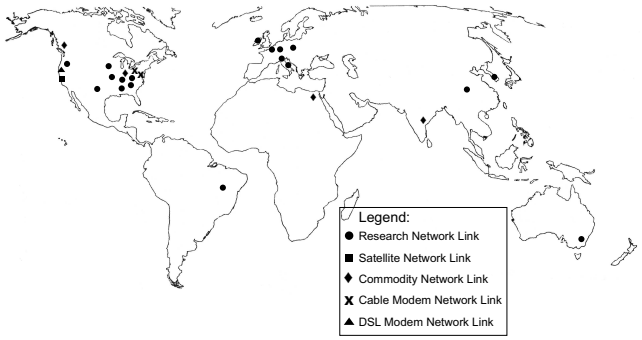

Fig. 3. World Map Showing the Test Sites Involved

in each path under test. Spirent SmartBits [18] technology was used to qualify whether NISTnet accurately introduced the delay, jitter and loss settings.

The difference between the LAN and Internet tests in terms of the NISTnet settings was that in the LAN environment the end-to-end delay, jitter and loss were completely controlled by the NISTnet settings; whereas, in the Internet the network paths already had inherent values of delay, jitter and loss. Therefore a network path characteristics pre-determination step was required for each test site before configuring additional end-to-end delay, jitter and loss on NISTnet. To accurately obtain the inherent network path characteristics, we used data from OARnet H.323 Beacon, NLANR Iperf and appareNet developed by Apparent Networks. The end-to-end delay, jitter and loss values configured on the NISTnet for the Internet tests were the values obtained by deducting the inherent path characteristics values from the LAN NISTnet settings.

4 Analysis of Performance Bounds

Sites with research network connections traversing Abilene and GEANT backbones had more consistent network path characteristics and overall results. This is in accord with the popular opinion about the superior performance levels of these backbones owing to the fact that there is limited cross traffic and low utilization levels on these networks. In contrast, sites that had commodity Internet connections, cable modem and DSL modem last mile connections and also academic sites such as Brazil and Korea had significant variations in network path characteristics and contributed the most to the variance in the overall results. Fig. 4 - 6 show the subjective and objective MOS rankings obtained during the testing for delay, jitter and loss values used in different scenarios. The variance observed in the results is contained well within the performance bounds values of delay, loss and jitter stated in Section 2.3. This clearly demonstrates that our LAN results are scalable to the Internet. Also the MOS values plotted in Fig. 4 - 6 include the values obtained for the tasks involving the H.323 Beacon. This proves the handy utility of the H.323 Beacon in determining user perceived audiovisual quality in H.323 production systems without remote end-user intervention.

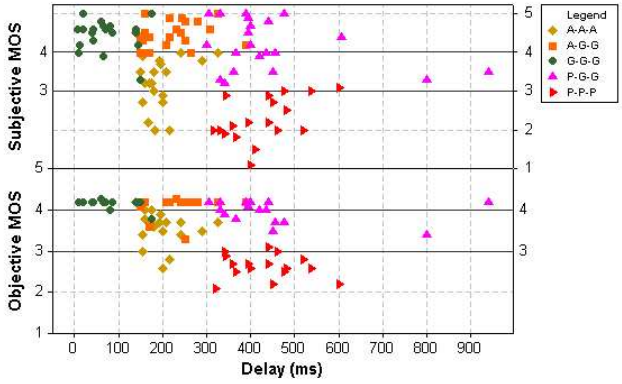

Fig. 4. Subjective and Objective MOS Vs Delay

The correlation between objective and subjective scores was observed to be in the above average to strong range. The Pearson correlation for delay, jitter and loss were 0.827, 0.737, and 0.712 respectively. The reason for all the correlations not being strong can be due to the following reasons: we used the Emodel objective scores that are modeled after only audio traffic streams; participants at the end-points involved in the testing were from very diverse backgrounds that included demographics of Videoconferencing Coordinators, Network Engineers, Graduate Students, Instructors and IT Managers; and various types of

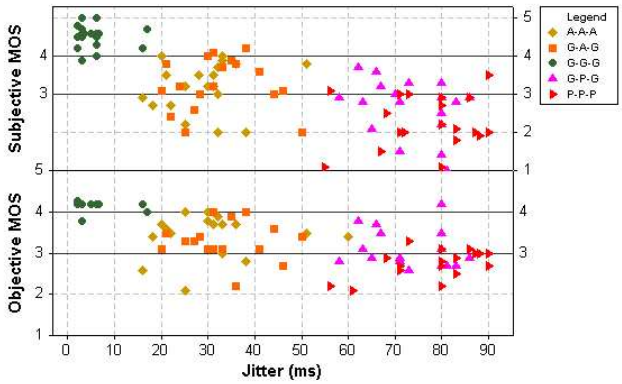

Fig. 5. Subiective and Obiective MOSS Vs Jitter

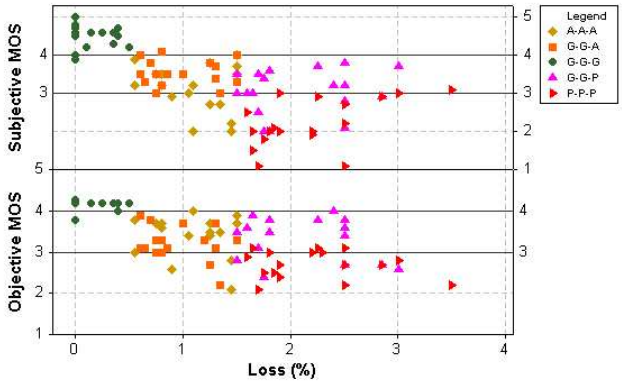

Fig. 6. Subjective and Objective MOS Vs Loss

hardware codec/software codec H.323 end-points were used in the testing. The various audio codecs observed at the endpoints were GSM, G.711 and G.722 and the various video codecs observed were H.261, H.262 and H.263.

Table 1 summarizes all the Subjective and Objective Quality Grade Assessments for the LAN/Internet tests. A '*' next to a paired-grade in the Result column indicates that an end-user will more often perceive that Grade of quality rather than the Grade without the '*' in that particular scenario. Values in the Result column that have more than one Grade and no '*' imply that there is an equal chance of an end-user perceiving either of the Grades of quality for that scenario.

In the pursuit to identify the most dominating factor amongst delay, jitter and loss that affects end-user perception of audiovisual quality, we normalized the scales of these factors and plotted them against both the Subjective and Objective MOS assessments as shown in Fig. 7 and 8. Each unit in the normalized scale corresponds to a delay of 150ms, jitter of 20ms and loss of 0.5%. We can observe that end-user perception of audiovisual quality is more sensitive

Table 1. Results of Quality Grade Assessments for LAN/Internet Tests

	S1	S2	S3	S4	S5	S6	S7	S8	S9
Delay	G	A	P	G	G	P	G	G	A
Jitter	G	A	P	G	P	G	G	A	G
Loss	G	A	P	P	G	G	A	G	G
Result	**G**	**A/P**	**P**	**A/P**	**A/P**	**G*/A**	**A**	**A*/P**	**G**

Legend:
$G \rightarrow Good$, $A \rightarrow Acceptable$, $P \rightarrow Poor$, $(S1 - S9) \rightarrow Scenarios$ 1-9

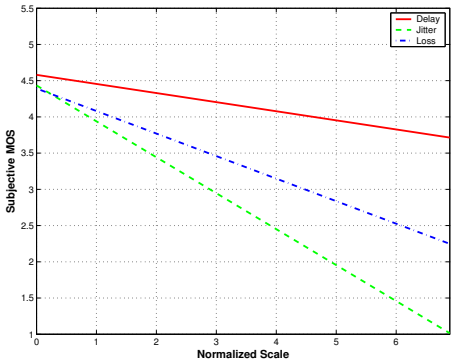

Fig. 7. Effects of Normalized Delay,Jitter and Loss variations on Subjecitve MOS

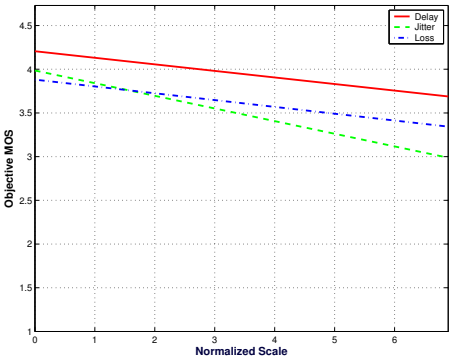

Fig. 8. Effects of Normalized Delay,Jitter and Loss variations on Objecitve MOS

to changes in jitter than to changes in delay and loss. In fact, the changes in delay have low impact on the end-user's perception of the audiovisual quality; although delay values more than 3 units on the normalized scale are deemed to be unsuitable for interactive communications [14].

5 Conclusion And Future Work

In this paper, we determined the performance bounds for network metrics such as delay, jitter and loss. We use these bounds to determine the impact of network health on end-user perception of audiovisual quality of H.323 applications. By emulating various network health scenarios both in the LAN and on the Internet and by using realistic Videoconferencing tasks, we show that end-user perception of audiovisual quality is more sensitive to the variations in end-to-end jitter than to variations in delay or loss. In the Internet tests, by considering almost every possible last-mile connection, we demonstrated that the results we obtained in the LAN tests scaled consistently to the Internet.

With the valuable network traces obtained from our one-on-one testing with various sites across the world we are currently studying the effects of jitter buffer sizes on packet discards under various network conditions using many popular audio and video codecs. We are also investigating effects of various packet sizes on the end-user perception of audiovisual quality of H.323 applications. The results of our studies could serve as trouble-shooting information during periods of suspected network trouble affecting H.323 audio and video-conferences. They can also foster broader understanding of the behavior of audio and video traffic over the Internet which can then lead to better designed networks in the future.

References

1. : ITU-T Recommendation H.323, Infrastructure of audiovisual services- Systems and terminal equipment for audiovisual services, Series H: Audiovisual and multimedia systems. (1999)
2. Markopoulou, A., Tobagi, F., Karam, M.: Assessment of VoIP quality over Internet backbones. In: Proceedings of IEEE Infocom. (2002)
3. Marsh, I., Li, F.: Wide area measurements of Voice Over IP Quality. SICS Technical Report T2003:08 (2003)
4. : Eurescom reports of Europe-wide experimentation of multimedia services (1999)
5. : ITU-T Recommendation G.107, The Emodel, a computational model for use in transmission planning. (1998)
6. : ITU-T Recommendation P.911, Subjective audiovisual quality assessment methods for multimedia applications. (1998)
7. Mullin, J., Smallwood, L., Watson, A., Wilson, G.: New techniques for assessing audio and video quality in real-time interactive communications. IHM-HCI Tutorial (2001)
8. : OARnet H.323 Beacon: a tool to troubleshoot end-to-end h.323 application performance problems. (http://www.itecohio.org/beacon)
9. Schopis, P., Calyam, P.: H.323 traffic characterization study. OARnet Technical Report submitted to American Distance Education Consortium (2001)
10. : ITU-T Recommendation P.800, Methods for subjective determination of transmission quality. (1996)
11. : ITU-T Recommendation P.862, An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. (2001)
12. Clark, A.: Modeling the effects of burst packet loss and recency on subjective voice quality. (2001)
13. Schulzrinne, H., Casner, S., Frederick, R., Jacobson, V.: RTP: A transport protocol for real-time applications. RFC 1889 (1996)
14. : ITU-T Recommendation G.114, One Way Transmission Time. (1996)
15. Kelly, B.: Quality of service in internet protocol (ip) networks (2002)
16. Indraji, A., Pearson, D.: The network: Internet2 commons site coordinator training (2003)
17. : NISTnet network emulation package. (http://snad.ncsl.nist.gov/itg/nistnet/)
18. : Spirent SmartBits network measurement suite. (http://www.spirentcom.com)