Application Notes

▷ **Title**   **Impact of Delay in Voice over IP Services**

▪ **Series**   VoIP Performance Management

▪ **Date**   January 2006

▷ ### Overview

This application note describes the sources of delay in Voice over IP services, the impact of this delay on conversational quality and, how to define acceptable levels of delay in service level definitions or agreements.

▷ ### Contents

Introduction

Delay has been a source of problems with telephone networks for many years. Previously, the major source of delay was satellite connections, typically 700 milliseconds per satellite hop. However, significant delay can now occur, even on local connections, due to the use of VoIP and cellular technology.

Sources of delay

In a voice conversation, a call participant is aware of the time taken for the remote user to react. This time includes the round trip delay experienced by the voice signal and the reaction time of the remote user. The user is not aware of any asymmetry that may be present in the time taken by the outgoing versus incoming voice signal.

Various delays can be measured or specified. *Round trip delay* refers to the total delay for the sending and receiving direction combined. One-way delay refers to the delay in either the sending or receiving direction. Symmetric one-way delay refers to the delay in the sending or receiving direction with the assumption that they are equal (i.e. half the round trip delay).

Delay within a Voice over IP service has three key components:

- IP Network Delay

- End System Delay

- External Delay

IP Network Delay

The IP Network delay is generally the round trip delay for an IP packet within the IP network. This may be measured using tools such as "ping" or protocols such as RTCP. It is important to note that some IP networks treat the RTP packets used to transport voice with higher priority than other packet types and hence the delay measured by ping or RTCP may be different to that actually experienced by voice packets.

IP network delay comprises the sum of transmission delays and queuing delays experienced by a packet traveling through the collection of routers, switches and other hardware that comprise the network. There is a certain minimum level of delay that will be experienced due to the time it takes to transmit a packet serially through a link. Onto this is added a more variable level of delay due to network congestion. IP network delays can range from just a few milliseconds to several hundred milliseconds.

If an IP phone or gateway is connected via DSL, a cable modem, or a low speed digital connection, then congestion related delay often could be significant. It takes approximately 100 milliseconds to send a typical maximum-size IP packet through a 128-kilobit per second link. This means that if voice and data traffic are both present on the link, then the voice traffic can be significantly delayed by data traffic. This can also cause a high level of jitter, which can lead to an increase in jitter buffer size (see below).

End System Delay

The End System Delay within a VoIP service is the sum of the encoding, decoding, jitter buffer, and other data handling delays that occur within a typical IP phone or gateway.

The encoding delay is the time taken to accumulate and encode a voice packet, and will always be at least as large as the packet size, typically 10 to 30 milliseconds.

The decoding delay is the time taken to decode and start to play out a received voice packet. This delay can be quite short.

IP phones and gateways use jitter buffers to remove short-term packet delay variations from the received packet stream by inserting some additional delay that allows the system to re-insert late-arriving packets in the proper order before they are decoded and played out. Jitter buffers are often adaptive and can grow in size to several hundred milliseconds.

The End System Delay can therefore typically range from 10-20 milliseconds up to several hundred milliseconds and can potentially form a major component of overall perceived delay.

Impact of delay on perceptual quality

Delay can have a considerable impact on conversational quality. High levels of delay can lead to conversational interaction problems. However, even low levels of delay can make any echo that may be present on a call much more annoying.

Conversational difficulty

Normal conversation is an interactive process, and if delay is introduced, it can lead to call participants interrupting each other (doubletalk) or to excessive pauses in speech. Frequent interruptions can obviously be quite annoying and excessive silence periods can be misinterpreted as delays in response, which can alter the apparent emotional content of speech.
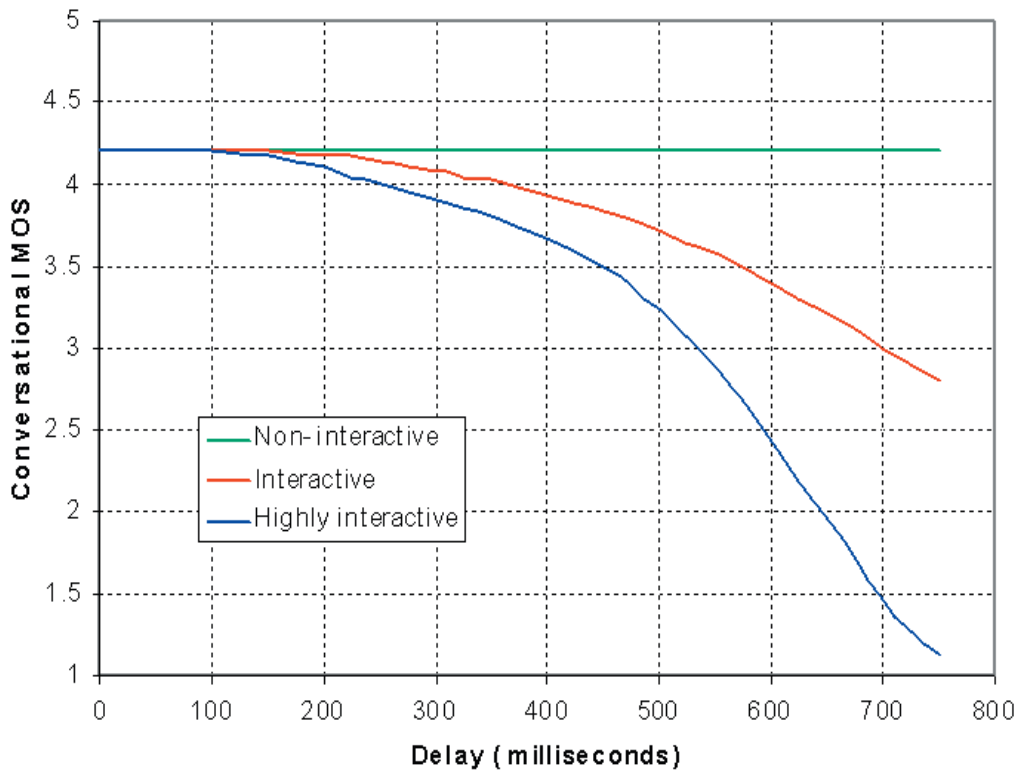
*Figure 1. Example of the relationship between one-way delay and
conversational quality for three conversational models*

Delay effects are also task dependant. If the conversational flow is largely in one direction, then delay may have little effect, whereas if the conversation is highly interactive, then even low levels of delay can be problematic.

Figure 1 above illustrates the task dependency of delay for three situations. If the task is non-interactive (i.e. a one way flow in the conversation) then delay has no effect. For a moderately interactive conversation there is little effect at 300 mS, however for a highly interactive conversation there is some noticeable effect at 100 mS and a rapid degradation as delay increases.

Impact of delay on echo

Echo is a common problem in telephony systems. The effects of echo are very dependant on delay - if the delay is very low then a high level of echo can be tolerated. This is illustrated in Figure 2 on the following page.

Figure 2 shows that with 10 mS of delay, a high echo level of 25 dB can be tolerated whereas with 100 mS of delay the tolerable echo level is 46 dB, which is quite low.
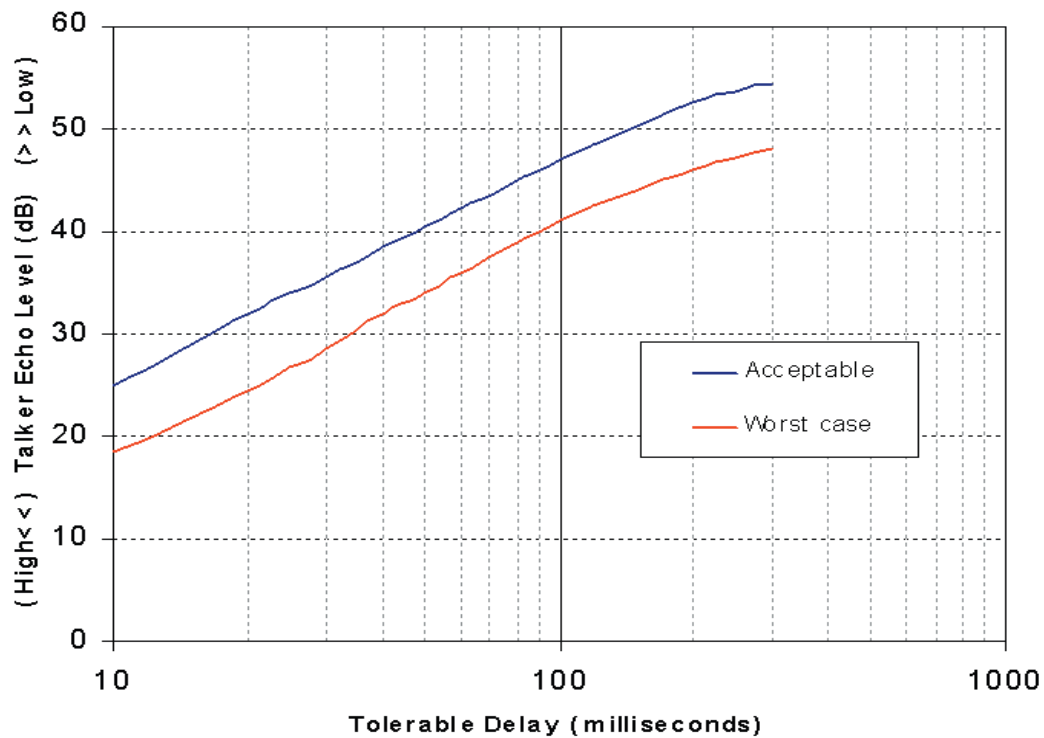
Figure 2. Example of the relationship between tolerable one-way delay and echo level (TELR)

Echo can be caused by acoustic feedback from the earpiece to mouthpiece of a telephone handset or headset, or by coupling at a 4-wire to 2-wire analog connection point. Echo cancellers are used to reduce or remove echo, however they are not always effective.

Diagnosing delay problems

The general steps in diagnosing delay problems are:

- Use either RTCP measured delay or a tool such as ping to determine if the source of delay appears to be the network or the end system (but note that this is not an accurate means of measuring delay).

- If the problem appears to be in the end system then check:

    (a) Jitter level and jitter buffer size. If the problem is caused by excessive jitter then investigate the cause of local network congestion.

    (b) Examine the packet size. Some systems are configured to use large voice packets or to pack multiple voice packets per IP packet. If this is the problem, use smaller packet sizes.

- If the problem appears to be in the network then check:

    (a) Access link (or internal network link) speeds, which should ideally be at least 500 kilobits per second

    (b) Network congestion, which can increase both delay and jitter.

RTCP XR (RFC3611) can be of great help in diagnosing delay problems as it (the VoIP metrics report block) reports network delay, end system delay and jitter buffer configuration.

Specifying acceptable targets for delay

When determining acceptable delay targets it is important to remember the following key points:

- Delay effects are task dependant.

- Delay experienced by users incorporates both end system delay (jitter dependent) and network delay.

- Tighter delay targets tend to imply higher bandwidth and lower congestion or utilization, which potentially increases service costs.

The table below provides an illustration of the approach that can be used for delay specification:

In this case, separate limits have been established for less interactive (typical office) and highly interactive (stock trader) applications. Approximate budgets are established for the endpoint-related parameters (related to packet size and jitter buffer size), and for network delay.

This approach allows expected and maximum delays to be estimated and compared with a worst-case limit for the type of application.

	Less interactive application	Highly interactive application
Delay related to packet size, approx voice packet duration	20 mS packets → 20 mS	10 mS packets → 10 mS
Jitter buffer size, approx 2 x expected typical jitter level	20 mS jitter → 40 mS jitter buffer	20 mS jitter → 20 mS jitter buffer
Worst case jitter buffer size approx 2 x max jitter level	50 mS jitter → 100 mS jitter buffer	20 mS jitter → 40 mS jitter buffer
Expected typical IP network delay (one way)	50 mS typical	50 mS typical
Maximum IP network delay	200 mS maximum	100 mS maximum
Expected total one way delay	*20+40+50 = 110 mS*	*10+20+50 = 80 mS*
Max total one way delay	*20+100+200 = 320 mS*	*10+40+100 = 150 mS*
Worst case limit	*400 mS*	*250 mS*

Summary

Delay in VoIP systems is a constant source of concern for the network operator because increasing amounts of delay contribute to steadily decreasing customer satisfaction with the service. Delay can come from a variety of sources that include the end system (e.g. IP phone), the local LAN, the access to the transport network and the transport network itself.

Excessive delay can cause difficulties in normal conversation, and its effects increase with the level of interaction in the conversation and the amount of delay.

Diagnoses and identification of delay problems can be done in real-time using simple tools such as RTCP and ping, but using more sophisticated tools such as VQmon will enable measurements that are more accurate. Once sources of delay are identified, then steps can be taken to reduce the delay to acceptable levels for the required application.

Notes

Acronyms

dB	Decibels
DSL	Digital Subscriber Line
IETF	Internet Engineering Task Force
IP	Internet Protocol
ITU	International Telecommunications Union
LAN	Local Area Network
MOS	Mean Opinion Score
ms	Milliseconds

RFC	Request For Comments
RTP	Real Time Protocol
RTCP	Real Time Control Protocol
RTCP XR	RTP Control Protocol Extended Report
TELR	Talker Echo Loudness Ratings
VoIP	Voice Over Internet Protocol
VQmon/EP	VQmon End Point
VQmon/SA	VQmon Stream Analysis
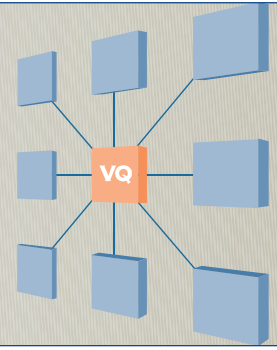
References

[1] IETF RFC3611 RTP Control Protocol Extended Reports (RTCP XR)

[2] ITU-T Recommendation G.131 Control of Talker Echo

[3] Pure Delay Effects on Speech Quality in Telecommunications, N. Kitawaki, K. Itoh, IEEE JSAC, vol. 9, no. 4, pp. 586-593, May 1991

About Telchemy, Incorporated

Telchemy, Incorporated is the global leader in VoIP Performance Management with its **VQmon**® and **SQmon**™ families of call quality monitoring and analysis software. Telchemy is the world's first company to provide voice quality management technology that considers the effects of time-varying network impairments and the perceptual effects of time-varying call quality. Founded in 1999, the company has products deployed and in use worldwide and markets its technology through leading networking, test and management product companies. Visit www.telchemy.com.

Telchemy Application Notes

- **Series** Understanding VoIP Performance
  - **Title** Voice Quality Measurement
  - **Title** Jitter Buffer Configuration
  - **Title** RTCP XR User Guide

- **Series** **VoIP Performance Management**
  - **Title** **Impact of Delay in Voice over IP Services**
  - **Title** Echo in Voice over IP Systems
  - **Title** Managing IP Centrex Performance
  - **Title** Managing Cable Telephony Performance
  - **Title** Managing Enterprise IP Telephony
  - **Title** Managing IP-based Cellular Services
  - **Title** Managing Carrier Core VoIP Services
  - **Title** SLA Monitoring Architecture

- **Series** VQmon Application Concepts

- **Series** SQmon Application Concepts

Information

- **Email**
  sales@telchemy.com
  info@telchemy.com
- **Phone**
  +1-678-387-3000

- **Telchemy, Incorporated**
  105 Nobel Ct
  Suite 100
  Alpharetta, GA 30005
  USA