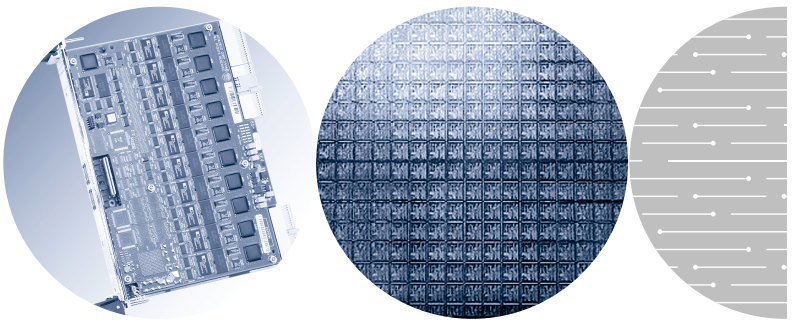
Whitepaper

Overcoming Barriers to High-Quality Voice over IP Deployments

Intel in Communications



Contents

Executive Summary

Quality of Service (QoS) issues are critical to the successful deployment of Voice over IP (VoIP) systems. Such impairment factors as latency, jitter, packet loss, and echo must be considered along with the ways in which they interact. The ability of a network to support high-quality VoIP calls and the configuration of various hardware and software parameters are important in ensuring caller satisfaction.

Introduction

Deployment of VoIP is growing, partly because QoS issues have been successfully addressed. In a properly configured VoIP installation today, conversations in VoIP calls can sound as good, and sometimes better, than in circuit-switched calls.

In a VoIP system, QoS depends on how a call is set up. More than telephone equipment and server technology must be considered. The network itself must be "VoIP-ready" and endpoint devices need to be tuned for QoS.

Creating a high-quality VoIP system depends on understanding basic QoS concepts such as latency, jitter, packet loss, and echo. Specific strategies can then be put in place to identify potential deployment issues at a specific site and resolve them.

Understanding QoS

Measuring QoS

Measures of quality tend to be very subjective in communications systems, and over the years several attempts have been made to predict a caller's feelings about QoS using objective criteria.

The traditional measure of a user's perception of quality is the Mean Opinion Score (MOS) defined in *Methods for Subjective Determination of Voice Quality* (ITU-T P.800). In P-800, an expert panel of listeners rated pre-selected voice samples of voice encoding and compression algorithms under controlled conditions. An MOS score can range from 1 (bad) to 5 (excellent), and a MOS of 4 is considered toll quality. The Pulse Code Modulation (PCM) algorithm (ITU-T G.711) has a MOS score of 4.4.

ITU-T G.107 presents a mathematical model, known as the E-Model, which attempts to predict QoS scores using more objective impairment factors. TIA/EIA TSB116 provides a comparison of E-Model Rating Values (R) and MOS scores. See Table 1 for details. An R-Value of 94 is equal to a MOS of 4.4[1]

R-Value	Characterization	MOS
90-100	Very satisfied	4.3+
80-90	Satisfied	4.0-4.3
70-80	Some Users Dissatisfied	3.6-4.0
60-70	Many Users Dissatisfied	3.1-3.6
50-60	Nearly All Users Dissatisfied	2.6-3.1
0-60	Not Recommended	1.0-2.6

Table 1. Comparison of R-Values and MOS Scores

[1] Table 1 draws data from Figure 1 in TIA/EIA TSB116, p. 5.

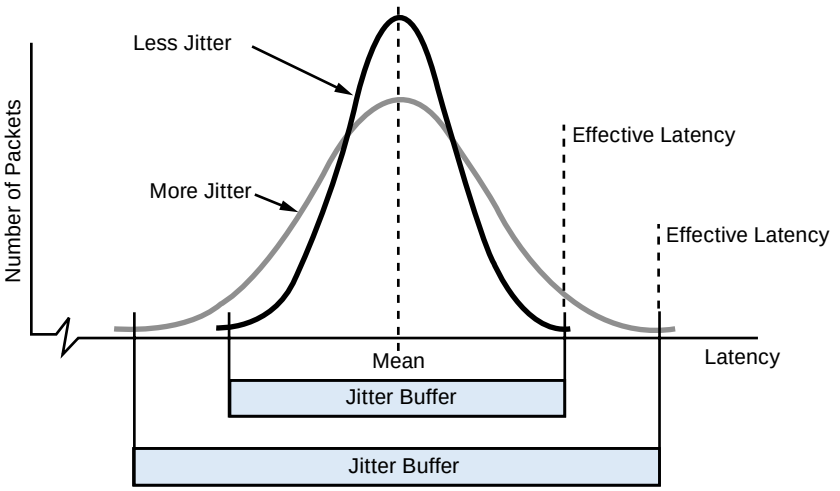


Figure 1. Jitter Increases Latency

Factors Affecting QoS

Impairment factors important to VoIP that are used as input parameters to the E-Model are:

- Delay
- Packet Loss
- Speech Compression
- Echo

These and other impairment factors must be controlled and mitigated in the network and at endpoints to ensure high QoS ratings and customer satisfaction.

Basic Delay: Latency

The transmission of voice data packets is not instantaneous, and latency is the term used to describe the time needed for all of the following:

- A packet of voice data to move across the network to an endpoint
- Encoding and packetization at the transmitting end
- De-jittering and decoding at the receiving end

Total latency is also called end-to-end latency or mouth-to-ear latency.

Conversations generally involve “turn-taking” on 200 ms breaks. When network latency approaches this value, the flow of a conversation becomes distorted. The two parties may start to talk at the same time or interrupt each other.

At over 170 ms delay, even a perfect signal degrades rapidly in acceptability.[2]

Variable Delay: Jitter

When discussing networks, latency is often given as an average value. However, VoIP is time-sensitive, and such an average can be misleading.

When a packet stream travels over an IP network, there is no guarantee that each packet in the network will travel over the same path, as in a circuit-switched network. Because they do not take the same path, intervals between packet arrival times vary since one packet may take more “hops” than the others, delaying its arrival considerably and causing it to have a much higher latency.

Parts of an IP network can also be momentarily overwhelmed with other processing duties in, for example, a router. Both of these circumstances cause latency to become irregular, and this irregular delay is called jitter.

To lessen the effects of jitter, packets are gathered in a jitter buffer at the receive end.

[2] For an important discussion of delay, including the expected delay for IP codecs, see ITU-T G.114.

Figure 1 shows that with the same average latency, increased jitter requires a larger jitter buffer, which consumes additional memory and yields greater latency.[3]

The jitter buffer must be sized to capture an optimal proportion of the data packets while keeping the effective latency as low as possible. In Figure 1, packets on the left and right ends of the bell curve fall outside of the jitter buffer and are, in effect, lost.

Packet Loss

Because IP networks treat voice as if it were data, voice packets will be dropped just as data packets are dropped under severe traffic loads and congestion.

Lost data packets can be re-transmitted, but, of course, this is not an acceptable solution for voice packets, which can contain up to 40 or as many as 80 ms of speech information.

Packet loss, then, can significantly reduce QoS. Even a 1% loss can significantly degrade the user experience with the ITU-T G.711 voice coder (vocoder), which is considered the standard for toll quality. Other coders degrade even more severely because they compress the data more rigorously.

Lost packets are ignored in the calculation of jitter since they are, in effect, packets with infinite delay and would skew the calculations.

Bandwidth

Bandwidth is the raw data transmission capacity of a network, and inadequate bandwidth causes both delay and packet loss. Since IP network traffic is irregular, packets will often be delayed without some kind of prioritization.

Several techniques can be used to prioritize packets. These include CoS, ToS, DiffServ, and IntServ.

Class of Service (CoS)

The IEEE extension 802.1P describes Class of Service (CoS) values that can be used to assign a priority. Network devices that recognize three-bit CoS values will deliver high-priority packets in a predictable manner. When congestion occurs, low-priority traffic is dropped in favor of high-priority traffic.

Type of Service (ToS)

RFC 1349 describes Type of Service (ToS), an in-band signaling of precedence for IP, which allows the Layer 3 IPV4 header to contain eight precedence values (0-7). These values are examined by routers and can also be used by Level 3 switches. DiffServ (described below) has superseded ToS.

Differentiated Services (DiffServ)

RFC 2474 redefines the ToS field in the Layer 3 IPV4 header as Differentiated Services (DiffServ). The redefinition is backwardly compatible. DiffServ defines a small set of Per-Hop Behaviors (PHBs) to define packet treatment. Examples are expedited forwarding, which simulate a virtual leased line; assured forwarding, which provides a high probability of delivery; and best effort. PHB is applied to each packet at each node, and the technique is highly scalable, performing classification at the edge.

Internet Services (IntServ RSVP)

RFC 2205 describes RSVP, an out-of-band QoS signaling protocol for reserving resources, such as bandwidth, for a “flow” or network path. Each flow is unidirectional, and two flows must be set up for each call. RSVP is not easily scalable, and DiffServ is expected to eclipse it.

Echo

Telephone handsets are designed to provide sidetone, which is necessary for a satisfactory user experience during a call. Microphone input is injected back into the earpiece, which allows callers to hear their own words and feel that they are being transmitted correctly. Echo, however, is always a negative factor and should always be minimized. In circuit-switched networks, latency is normally so low that echo is perceived in the same way as sidetone and is not a significant impairment.

[3] The jitter buffers in Figure 1 would handle 99% of all packet latencies.

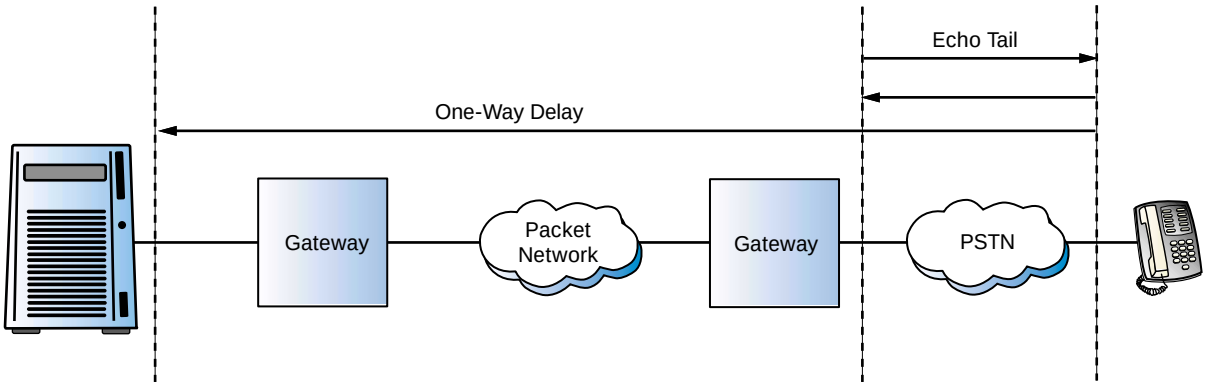


Figure 2. Echo and Delay in a VoIP Configuration

On an analog phone, echo is generated at the 2- to 4-wire hybrid. On a digital or IP phone, echo is generated in the analog section of the phone (at the handset cord capacitive coupling, acoustic coupling, etc.). TSB116 (p.10) provides detailed diagrams of these types of connections.

In a VoIP configuration, the tail (that is, the time of a round trip from the gateway to the hybrid and back) on a gateway echo canceller only needs to handle delay on the circuit-switched leg of the call as illustrated in Figure 2. An echo tail of 16 ms is usually adequate with 32 ms required in France. Intel's IP boards provide a tail of 16 ms, but the Intel® NetStructure™ IPT6720C provides a tail of 64 ms.

Echo loudness must also be controlled. ITU-T G.168 recommends an echo loudness rating (ELR) of >=55 dB of echo path loss for echo cancellers in gateways. Echo cancellation is never perfect, and the more echo that is eliminated, the higher the computational load. The ITU recommendation is quite stringent and difficult to achieve.

Echo interacting with delay can compound a negative user experience. TSB116 (p. 10) provides various talker echo loudness ratings (TELRs) mapped to delay, which shows how different loudness levels track to user satisfaction ratings.

Optimizing QoS

Some of the areas covered in this section correspond to the recommendations for IP telephony voice quality in TSB116.[4]

Reduce Latency and Jitter

Controlling delay is key to optimizing QoS, and should be kept well below 170 ms. Because jitter increases effective latency, it is important to take steps to control both latency and jitter.

Because the size of a jitter buffer directly affects the latency perceived by the caller, networks must be provisioned for both low latency and low jitter both at the endpoint and from end-to-end.

Reducing Delay at an Endpoint

Several techniques can be used to reduce delay at an endpoint:

- Optimize jitter buffering – Intel's packet loss recovery algorithm provides an adaptive jitter buffer for its IP boards and Intel NetStructure Host Media Processing (HMP) software.
- Optimize packet size – A packet size of 10 ms is optimal for reduced packetization latency. However, a larger number of smaller packets with relatively greater bandwidth overhead may be worse overall than a 20 ms packet size in some situations.
- Avoid asynchronous transcoding

[4] A summary of recommendations appears on p. 1 of TSB116.

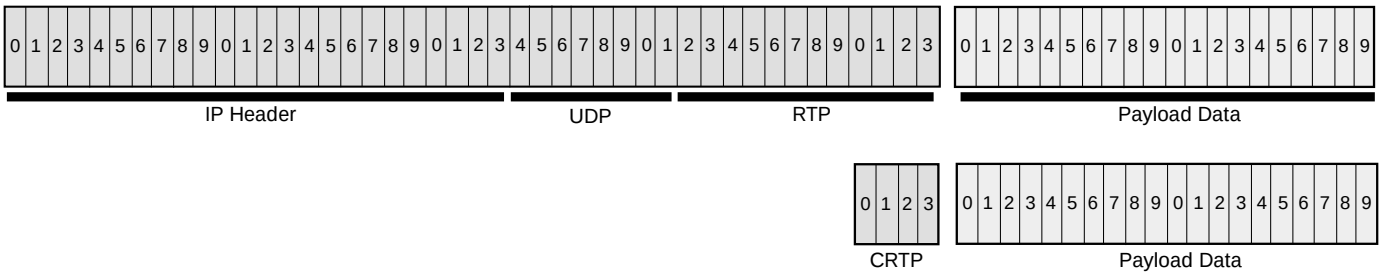


Figure 3. Header Compression

- Use a stable packet size
- Use a low compression codec such as G.711
- Ensure that network protocol stacks are efficient and correctly prioritized for VoIP traffic.

Reducing End-to-End Delay by Prioritization

To reduce end-to-end delay, VoIP packets can be given a higher priority at Layer 2 and Layer 3 by using the following:

- Class of Service (COS) – implemented for Ethernet
- Type of Service (TOS) – field in the IP header can be provisioned using the ipm_SetParm() on Intel IP boards.
- DiffServ – implemented at the router by static provisioning based on TOS bits.
- RSVP for bandwidth reservation – implemented at the router by static provisioning based on the transmitting port.
- Policy-based network management
- Multi-Protocol Label Switching (MPLS) [RFC 3031], which uses RFC 3212 CR-LDP for QoS

In order for prioritization to be effective, network stacks at endpoints must also prioritize the VoIP packets.

Dealing with Slow Links

On very slow links, a single large data packet could take up all available bandwidth for an unacceptable length of time. For example, a one-kilobyte data packet (equivalent to eight kilobits) traversing a 128 kilobits per second link blocks the link for 62.5 ms, which will delay six 10 ms packets of voice, greatly increasing jitter. Fragmenting large data packets and interleaving them with small VoIP packets, which are time-sensitive, can remedy this situation. Fragmentation and interleaving can be transparently provisioned between WAN routers with low-speed links.

Another technique that should be considered for reducing bandwidth demands on slow links is header compression, and IETF's RFC 2508 describes a very effective method for such compression. Normally a header consists of nested headers for real-time transport protocol (RTP), user datagram protocol (UDP), and Internet Protocol (IP) with a total size of 44 bytes. The payload data according to G.729A at 20 ms in such a packet is only 20 bytes. However, by using a compressed real-time protocol (CRTP) header, the header is reduced to 2 to 4 bytes, which means the packet uses only one-third of the bandwidth (24 bytes instead of 64 bytes). See Figure 3 for an illustration of this technique.

Compensate for Packet Loss

Packet Loss Concealment (PLC) or Packet Loss Recovery (PLR) algorithms at the endpoint can compensate for packet loss. Even a 5% rate of packet loss can be acceptable when the G.711 PLC algorithm is used.[5] Many speech coders based on Codebook Excited Linear Prediction (CELP), such as G.723.1, G.728, and G.729, have PLC built-in.

[5] See TSB116, pp. 37-8. Figure I.1/G.711 provides an excellent illustration of the dramatic effect of the frame erasure concealment algorithm for the G.711 vocoder (ITU Recommendation T-G.711 Appendix I).

IP boards and HMP software from Intel use PLC to repeat the previous packet when a packet is lost, and then inject silence for the rest of the burst. IPT boards from Intel use a similar strategy.

Intel's packet loss recovery algorithm is also valuable because it optimizes buffer size by adapting to current conditions. This algorithm is available with IP boards and HMP software from Intel.

Payload redundancy (RFC 2198) can also be used to prevent packet loss, but it increases bandwidth requirements. It is supported with the ipm_SetRemoteMedia() parameter in Intel's IP boards .

Ensure Enough Bandwidth Is Available

Speech compression should be considered because it reduces bandwidth requirements end-to-end.

Speech compression algorithms are described in ITU-T G.723, G.729, and other standards. This kind of compression does reduce bandwidth requirements, but it also reduces perceived sound quality. In addition, packet loss has much more serious consequences when high compression codecs are used because more data is lost per packet. Even with small delays, compression codecs are barely acceptable.[6] For this reason, G.711 is the preferred voice coder, even though it has high bandwidth requirements.

G.711 also provides other advantages. As previously mentioned, G.711 Appendix I supplies a powerful packet loss concealment algorithm. G.711 Appendix II provides two powerful tools for bandwidth conservation: a voice activity detector (VAD) and a comfort noise generator (CNG). The VAD senses when no voice is present, and sends sparse control packets rather than full packets of silence. The CNG plays background noise instead of no sound at all, which users find preferable to silence.

The bandwidth conservation techniques can provide about a 50% bandwidth savings simply by suppressing the normal silence in voice calls because the connection is full duplex with 64 Kbps in each direction. Since only one person talks at a time, the bandwidth consumed by the other person is always empty and can be coded into silence packets.

Start with a Good E-Model "R" Factor

Use the best codec possible since the network degrades codec performance. G.711 is a good choice, although higher bandwidth codecs such as G.722 can deliver superior sound quality if enough bandwidth is available. Consider low bandwidth codecs such as G.723 or G.729 only if bandwidth is constrained. Use CRTP to conserve bandwidth on slow links.

For more recommendations, see Appendix A.

Conclusion

User perception of VoIP quality can radically improve when network and telephone equipment are correctly set up. Provisioning the network for QoS is paramount, but alone it is not enough. Impairment factors such as latency, jitter, packet loss, and echo, and the way these factors interact, are also critical. When careful attention is paid to controlling these factors and sufficient bandwidth is available, callers can be as satisfied with VoIP calls as they are with calls carried over a circuit-switched network.

[6] See TSB116, p.12.

Appendix A. Configuring Hardware and Software Optimally

Before making specific changes to your hardware and software configuration, you should complete a general VoIP readiness survey to understand your VoIP requirements and current traffic patterns. You should also remedy any network limitations that would inhibit effective VoIP deployment. You will then be ready to look at the specific configuration changes listed in the rest of this appendix.

1. Set Phone Defaults

- Enable VLAN tag, which contains the CoS field
- Set CoS and the DiffServ Code Point (DSCP) for voice traffic
- Set codec to G.711
- Set packet to smallest (10 ms)
- Enable PLC
- Enable redundancy

2. Set NIC Defaults

Enable classification and tagging for VoIP with CoS and DSCP

3. Set Parameters on Intel® Products

To optimize QoS, it is important that specific parameters on IP boards and HMP software from Intel are set.

If Intel signaling stacks for SIP and H.323 are in use, do the following:

- For setting codes and frame size for calls, use gc_MakeCall()
- For setting TOS and RFC 2833 redundancy level, use gc_SetParm()

If other signaling stacks under split call control are in use, do the following

- For setting codes and frame size for calls, use ipm_SetMediaInfo()
- For setting TOS and RFC 2833 redundancy level, use ipm_SetParms()

4. Make Other Software Changes

Set the following parameters on a call-by-call basis.

- Codec selection
- PLC
- Redundancy
- Jitter buffer

Record Real-Time Control Protocol (RTCP)[7] information with call details. This allows parameters to be tuned, and applications can query for RTCP information at any time with an ipm function.

Use call quality alarms. This allows thresholds to be set, and administrators to be alerted when quality falls below those thresholds.

5. Consider Call Control Signaling and Media Streaming

Call control signals should be given a higher priority than the media stream because a network segment congested with media streaming can cause an interruption in the transmission of call control signals. If you find this to be a problem, you might consider using a separate NIC for Session Initiation Protocol (SIP), H.323, or other control signals. Isolate them through the use of a separate Ethernet hub or use an Ethernet switch with significant headroom capacity.

[7] IETF RFC 768 defines the User Datagram Protocol (UDP), an unreliable but fast protocol for IP. IETF RFC 1889 uses UDP and specifies a media streaming protocol called the Real Time Protocol (RTP). RTCP is a subset of the RTP specification.

Appendix B. Key to Board References

Two board categories are referenced in this paper. This appendix lists the boards in each category.

IP Boards from Intel

Intel NetStructure DM/IP241-1T1-PCI-100BT

Intel NetStructure DM/IP301-1ET1-PCI-100BT

Intel NetStructure DM/IP481-2T1-CPCI-100BT

Intel NetStructure DM/IP601-2E1-CPCI-100BT

Intel NetStructure DM/IP601-CPCI-100BT

IPT Boards from Intel

Intel NetStructure IPT1200C

Intel NetStructure IPT2400C

Intel NetStructure IPT4800C

Intel NetStructure IPT6720C

To learn more, visit our site on the World Wide Web at **http://www.intel.com**.

1515 Route Ten
Parsippany, NJ 07054
Phone: 1-973-993-3000

Printed in the USA Printed on recycled paper. 03/03 00-8539-001