ADAPTIVE DIGITAL TECHNOLOGIES

WHITE PAPER

Voice Latency

for VoIP Applications

Adaptive Digital Technologies, Inc.

March 22, 2013

File: VoiceLatency.doc

1. Introduction

This document describes the time (or latency) that it takes for a voice signal picked up at the microphone of one telephone to appear at the speaker of a second telephone when the two telephones are connected via a G.PAK-based VoIP system.

This document will analyze the progression of an analog impulse as it makes its way the the system. Each process that affects the latency of a Voip signal as it traverses through the various processes of a VOIP system will be discussed. For convenience, the processing will be divided into three major phases of operation:

1. Encoding,
2. Network traversal, and
3. Decoding.

The latency for each phase can be calcaluted summing the times that it takes to perform each of the individual processes within the phase, including any scheduling and buffering delays that may be inherent between processes.

The following table lists the latency as a function of frame size assuming that the channel is running G.711 and G.168 echo cancellation. This table does not account for delays through the IP network.

Frame Size	End to End Latency (msec)
1	9.5 + Jitter Buffer Size
5	25.5 + Jitter Buffer Size
10	45.5 + Jitter Buffer Size
30	125.5 + Jitter Buffer Size

The remainder of the document describes the contributing factors toward the overall latency.

2. Latency Formulae

2.1 Notations

Latency values in blue indicate a fixed amount of latency. Values in green indicate that latencies of up to that amount MAY occur dependent upon system loading and timing.

2.2 Encoding Latency (EL)

The encoding latency is calculated by summing the values E1 through E6. See Section 3 for details.

EL = ¼ ms + 1 ms + frame time + (frame time + algorithm delays) + 1 ms

2.3 Network Latency (NL)

The latency, NL, due to network transversal is highly dependent upon the number of nodes (switches, hubs, routers, gateways, proxies, etc.) that the packet must pass through to reach its destination and the available data bandwidith between those nodes.

2.4 Decoding Latency (DL)

The decoding latency is calculated by summing the values D1 through D6. See Section 4 for details.

DL = 1 ms + 1 ms + (min delay + max jitter depth + frame time) + (frame time + algorithm delays) + 1 ms +¼ ms

3. Encoding (Microphone to Network).

This section describes the processing and latency that occurs from the time that an analog impulse is detected at the microphone of the telephone until it is placed onto the network as an IP packet.

3.1 A/D Sampling (E1 = ¼ msec)

This step is performed by hardware and converts the analog input into digital samples and places the samples on a TDM bus. Samples are stored to and removed from the TDM bus at the rate of 8000 kHz. This mechanism adds latency E1 = up to 1/4th of a millisecond of latency between the time when an 1/8th of a millisecond long analog signal is processed by the A/D hardware and the time that it is subsequently read from the TDM bus.

3.2 TDM Demultiplexing (E2 = 1 msec)

This step demultipexes the data on a TDM bus into individual streams of data for channel processing. During every one millisecond interval, one millisecond of TDM samples are stored into a (ping) DMA buffer while the previous millisecond of samples are demuxed from (pong) DMA buffer into their corresponding framing buffers. This step occurs periodically and adds latency E2 = 1 millisecond.

3.3 Framing (E3 = frame time)

Voice algorithms and encoders require that one of more samples be processed together as a 'frame' of data. The amount of time required to collect the number of samples that comprise this frame of data is called the frame time. A complete frame's worth of samples must be collected prior to invoking the frame processing. The framing step delays processing of voice samples until a complete frame is available. This step occurs periodically and adds a latency E3 = frame time.

3.4 Processing + Encoding (E4 = frame time + algorithm delays)

This step performs algorithm processing such as: echo cancellation, noise suppression, tone detection, and sampling rate conversions. After the algorithm processing is complete the signal is then encoded for delivery to the network. This operation occurs periodically and may add up to a frame time of latency. Algorithms such as speech compression, echo cancellation, noise reduction, filtering, and sampling rate conversions may also add latency.

3.5 Delivery of Encoded data to IP stack (E5 = 1 msec)

This step places the encoded data into a queue for processing by the IP stack. This queue is serviced at a minimum of once every millisecond for a latency of E5 = up to 1 millisecond.

3.6 Formatting of IP Packet (E6 = 0)

This step formats the IP packet and places the packet into a queue for delivery by the hardware to the ethernet. This operation is negligible; E6 ~ 0 ms.

4. Decoding (Network to Speaker)

This section describes the processing and latency that occurs from the time that a voice packet is received from the network until it appears at the speaker of the telephone.

4.1 IP Stack Reception Delay (D1 = 1 msec)

This step receives the IP packet from the ethernet and places the packet into a queue for delivery to an application socket. This operation is driven by a semaphore with a latency D1 of up to 1 millisecond.

4.2 Jitter Buffer Insertion (D2 = 1 msec)

This step removes IP packets from a queue and places them into a jitter buffer. This process is called periodically and has a latency D2 of up to 1 millisecond.

4.3 Jitter Delay (D3 = min delay + max jitter depth + frame time)

This step delays the playout of packets to adjust the playback time of voice packets to account for network jitter. This process adds a fixed latency of min delay. A second latency of up to max jitter depth is also added to allow dynamic jitter buffers to adjust to variable latency. A third latency of up to frame time is also added to account for the playout time of the packet relative to the start of the frame.

4.4 Decoding + Processing (D4 = frame time + algorithm delays)

This step decodes the packet into individual samples for playout. After decoding is is complete algorithm processing such as: echo cancellation, sampling rate conversions, and automatic gain control is performed. This operation occurs periodically and adds a frame time of latency. Algorithms such as speech decompression, echo cancellation, sampling rate conversions and filtering may also add latency.

4.5 TDM Mux (D5 = 1 msec)

This step multipexes the data from individual channel streams onto the TDM bus. During every one millisecond interval, one millisecond of samples are muxed into a (ping) DMA buffer while the previous millisecond of muxed samples are transferred to the TDM bus from a second (pong) DMA buffer. This step occurs periodically and adds latency D5 = 1 millisecond.

4.6 D/A (D6 = ¼ msec)

This step is performed by hardware and converts the digital samples from the TMD bus to an analog signal for output to the speaker. Samples are placed on and removed from the TDM bus at the rate of 8000 kHz. This mechanism adds latency D6 = up to 1/4th of a millisecond of latency between the time when a sample is place on the TDM bus and the subsequent 1/8th of a millisecond long analog signal is produced by the D/A hardware.