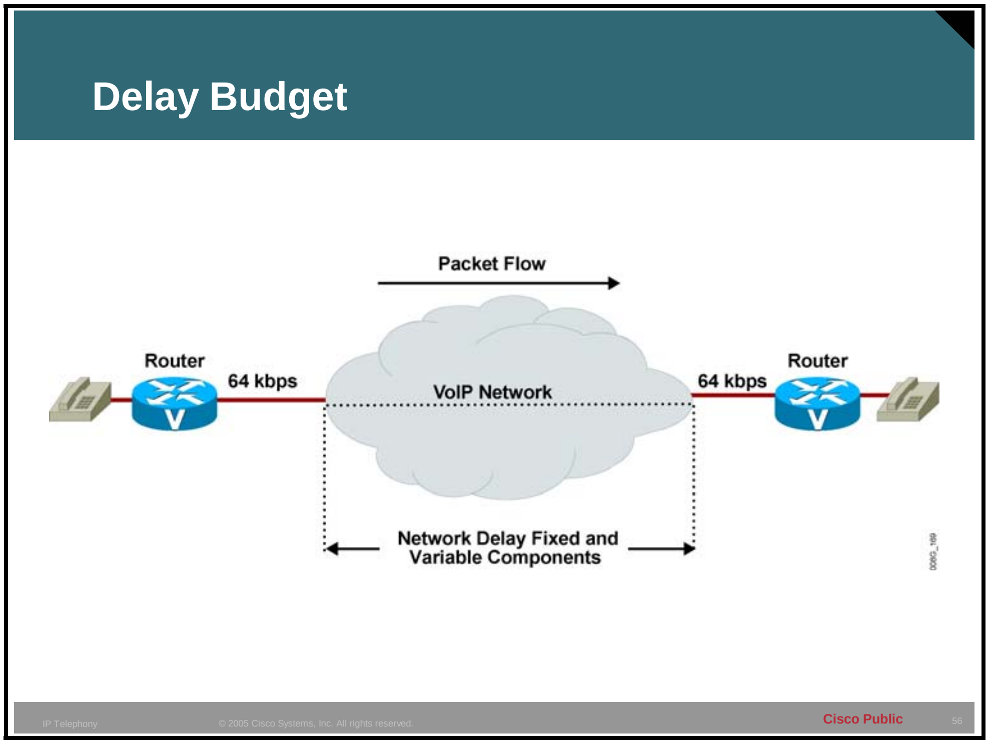
Delay

Need for a Delay Budget

The end-to-end delay in a VoIP network is known as the delay budget. Network administrators must design a network to operate within an acceptable delay budget. This topic explains the concept of delay budget and how to measure it.



Delay Budget

IP Telephony © 2005 Cisco Systems, Inc. All rights reserved. Cisco Public 56

Delay is the accumulated latency of end-to-end voice traffic in a VoIP network. The purpose of a delay budget is to ensure that the voice network does not exceed accepted limits of delay for voice telephony conversation. The delay budget is the sum of all the delays, fixed and variable, that are found in the network along the audio path. You can measure the delay budget by adding up all of the individual contributing components, as shown in the figure. The delay budget is measured in each direction individually, not round-trip.

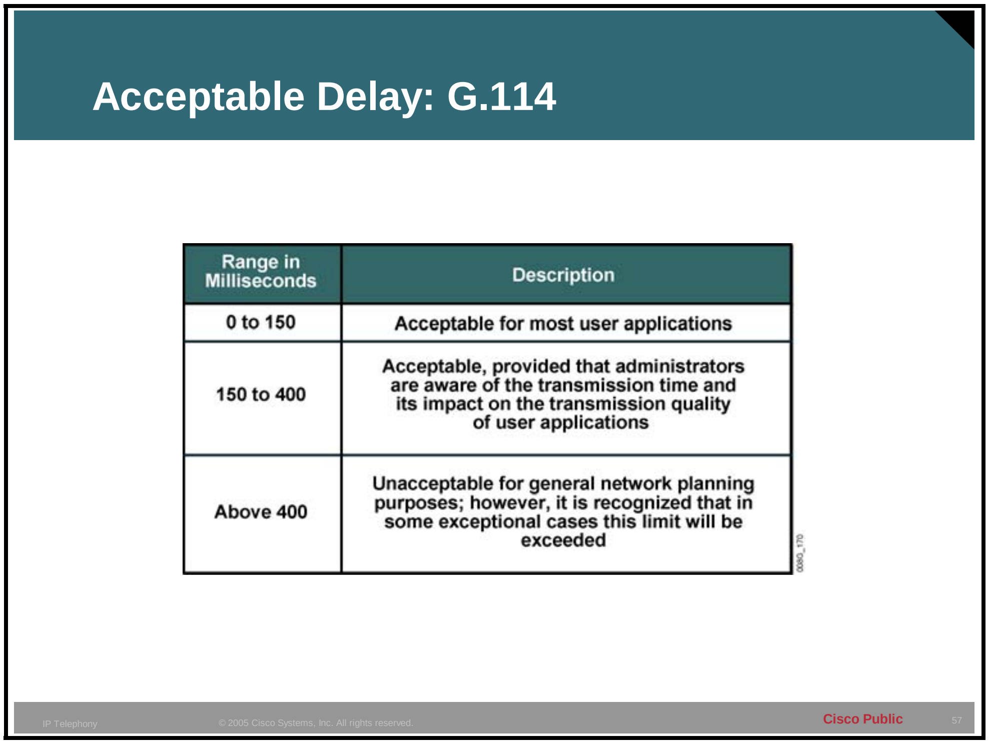
Network administrators must be aware that delay exists, and then design their network to bring end-to-end delay within acceptable limits.

Example: Need For a Delay Budget

As delay increases, talkers and listeners become unsynchronized and often find themselves speaking at the same time or both waiting for the other to speak. This condition is commonly called *talker overlap*. While the overall voice quality may be acceptable, users may find the stilted nature of the conversation unacceptably annoying. Talker overlap may be observed on international telephone calls that travel over satellite connections. Satellite delay is about 500 ms: 250 ms up and 250 ms down.

Guidelines for Acceptable Delay

International telephony communications must adhere to a delay standard. This topic defines the standard and its limits.



Acceptable Delay: G.114

Range in Milliseconds	Description
0 to 150	Acceptable for most user applications
150 to 400	Acceptable, provided that administrators are aware of the transmission time and its impact on the transmission quality of user applications
Above 400	Unacceptable for general network planning purposes; however, it is recognized that in some exceptional cases this limit will be exceeded

IP Telephony © 2005 Cisco Systems, Inc. All rights reserved. Cisco Public 57

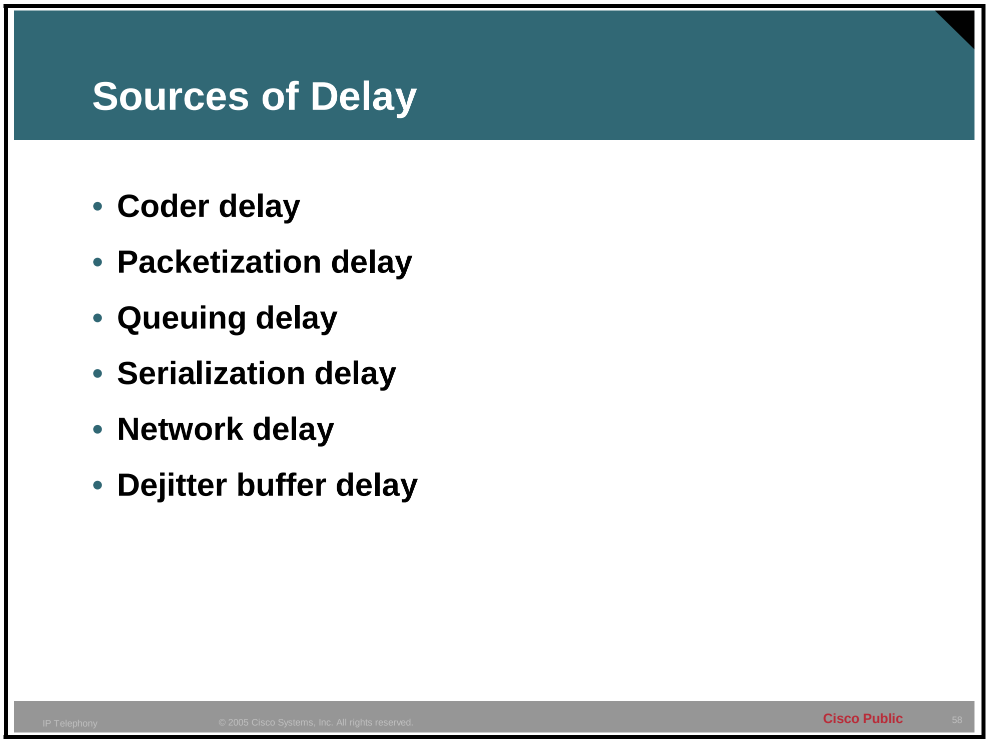
The ITU addresses network delay for voice applications in Recommendation G.114. This recommendation is oriented to national telecommunications administrations and is more stringent than what is normally applied in private voice networks. When the location and business needs of end users are well known to the network designer, more delay may prove acceptable.

Example: Acceptable Delay

As shown in the figure, acceptable delay time is from 0 to 400 ms for private networks. Delay times that exceed 400 ms are unacceptable for general network planning. However, all networks should be engineered to recognize and minimize the voice-connection delay. For example, suppose an enterprise is doing the planning for a new IP telephony roll-out. They plan and design to meet the 150-ms one-way delay, but have some locations in which 250-ms one-way delay is the best they can do. Since most of these calls will be on-net calls, the organization has to decide if the lower quality of the calls is acceptable for these locations.

Sources of Delay

Many factors add to overall delay. This topic lists six factors that are sources of delay.



Sources of Delay

- Coder delay
- Packetization delay
- Queuing delay
- Serialization delay
- Network delay
- Dejitter buffer delay

IP Telephony © 2005 Cisco Systems, Inc. All rights reserved. Cisco Public 58

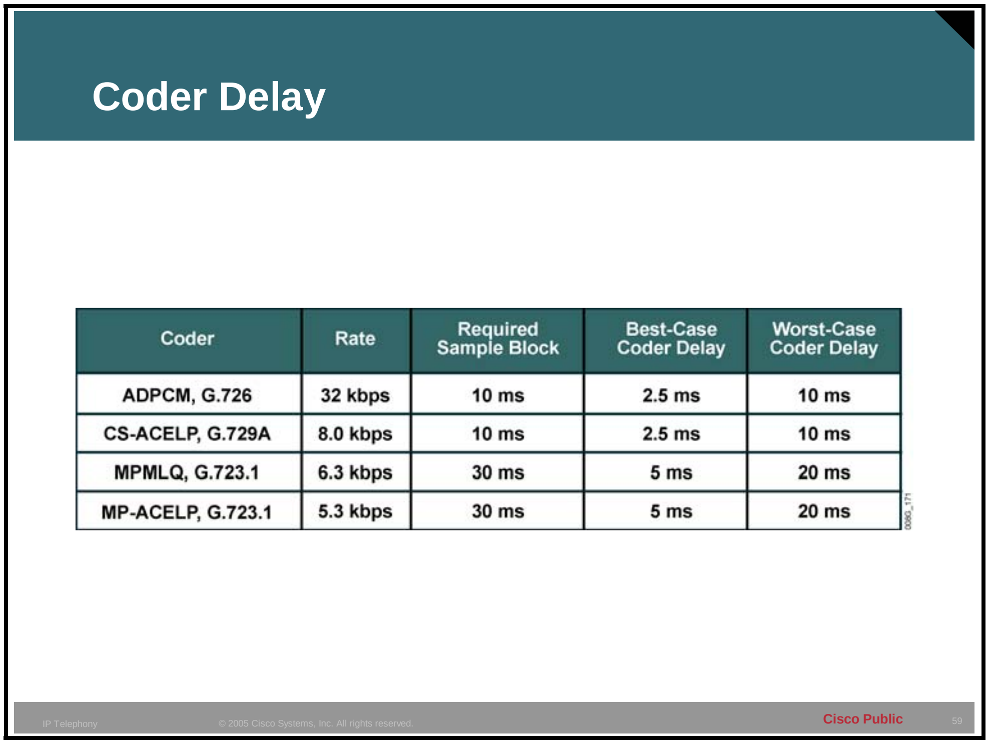
Following is an explanation of six major factors that contribute to overall fixed and variable delay:

- **Coder delay:** Also called processing delay, coder delay is the time taken by the DSP to compress a block of pulse code modulation (PCM) samples. Because different coders work in different ways, this delay varies with the voice coder that is used and the processor speed.
- **Packetization delay:** Packetization delay is the time it takes to fill a packet payload with encoded or compressed speech. This delay is a function of the sample block size that is required by the vocoder and the number of blocks placed in a single frame. Packetization delay is also called accumulation delay because the voice samples accumulate in a buffer before being released. With typical payload sizes used on Cisco routers, packetization delay for G.711, G.726, and G.729 does not exceed 30 ms.
- **Queuing delay:** After the network builds a compressed voice payload, it adds a header and queues for transmission on the network connection. Because voice should have absolute priority in the router or gateway, a voice frame must wait only for a data frame already playing out or for other voice frames ahead of it. Essentially, the voice frame waits for the serialization delay of any preceding frames that are in the output queue. Queuing delay is a variable delay and is dependent on the trunk speed and the state of the queue.
- **Serialization delay:** Serialization delay is the fixed delay that is required to clock a voice or data frame onto the network interface; it is directly related to the clock rate on the trunk.

- **Network delay:** The public Frame Relay or ATM network that interconnects the endpoint locations is the source of the longest voice-connection delays. These delays are also the most difficult to quantify. If a private enterprise builds its own internal Frame Relay network for the purpose of wide-area connectivity, it is possible to identify the individual components of delay. In general, the fixed components are from propagation delays on the trunks within the network; variable delays are the result of queuing delays that clock frames into and out of intermediate switches. To estimate propagation delay, a popular estimate of 10 microseconds/mile or 6 microseconds/km (G.114) is widely used, although intermediate multiplexing equipment, backhauling, microwave links, and other features of carrier networks create many exceptions. Typical carrier delays for U.S. Frame Relay connections are 40 ms fixed, and 25 ms variable, for a total worst-case delay of 65 ms.
- **Dejitter buffer delay:** Because speech is a constant bit-rate service, the jitter from all the variable delays must be removed before the signal leaves the network. In Cisco routers and gateways, this is accomplished with a dejitter buffer at the far-end (receiving) router or gateway. The dejitter buffer transforms the variable delay into a fixed delay by holding the first sample that is received for a period of time *before* playing it out. This holding period is known as the initial playout delay. The actual contribution of the dejitter buffer to delay is the initial playout delay of the dejitter buffer *plus* the actual amount of delay of the first packet that was buffered in the network. The worst case would be twice the dejitter buffer initial delay (assuming the first packet through the network experienced only minimal buffering delay).

Effects of Coders and Voice Sampling on Delay

The process of encoding an analog voice sample into a compressed digitized bit stream contributes to delay. This topic describes the effect of coder delay. Best-case and worst-case coder delays are shown in the figure.



Coder Delay

Coder	Rate	Required Sample Block	Best-Case Coder Delay	Worst-Case Coder Delay
ADPCM, G.726	32 kbps	10 ms	2.5 ms	10 ms
CS-ACELP, G.729A	8.0 kbps	10 ms	2.5 ms	10 ms
MPMLQ, G.723.1	6.3 kbps	30 ms	5 ms	20 ms
MP-ACELP, G.723.1	5.3 kbps	30 ms	5 ms	20 ms

IP Telephony © 2005 Cisco Systems, Inc. All rights reserved. Cisco Public 59

The compression time for a Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP) process ranges from 2.5 to 10 ms, depending on the loading of the DSP. If the DSP is fully loaded with four voice channels, the coder delay will be 10 ms. If the DSP is loaded with one voice channel only, the coder delay will be 2.5 ms. For design purposes, use the worst-case time of 10 ms.

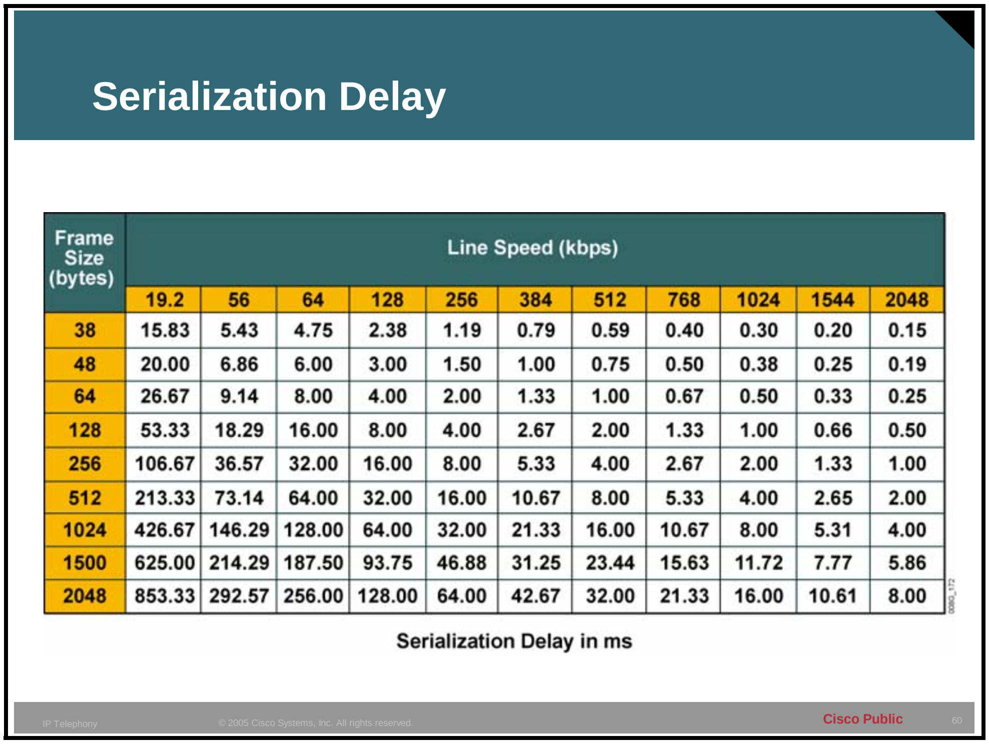
Decompression time is roughly 10 percent of the compression time for each block. However, because there may be multiple samples in each frame, the decompression time is proportional to the number of samples per frame. Consequently, the worst-case decompression time for a frame with three samples is 3 x 1 ms, or 3 ms. Generally, two or three blocks of compressed G.729 output are put in one frame, while only one sample of compressed G.723.1 output is sent in each frame.

Example: Coder Delay

The figure shows examples of some coder delays. CS-ACELP, for example, lists a best-case delay of 2.5 ms with a worst-case delay of 10 ms. In all calculations, the worst-case number should be used.

Managing Serialization Delay

An important part of delay is the serialization delay on the interface. This topic describes serialization delay and its management.



Serialization Delay

Frame Size (bytes)	Line Speed (kbps)										
	19.2	56	64	128	256	384	512	768	1024	1544	2048
38	15.83	5.43	4.75	2.38	1.19	0.79	0.59	0.40	0.30	0.20	0.15
48	20.00	6.86	6.00	3.00	1.50	1.00	0.75	0.50	0.38	0.25	0.19
64	26.67	9.14	8.00	4.00	2.00	1.33	1.00	0.67	0.50	0.33	0.25
128	53.33	18.29	16.00	8.00	4.00	2.67	2.00	1.33	1.00	0.66	0.50
256	106.67	36.57	32.00	16.00	8.00	5.33	4.00	2.67	2.00	1.33	1.00
512	213.33	73.14	64.00	32.00	16.00	10.67	8.00	5.33	4.00	2.65	2.00
1024	426.67	146.29	128.00	64.00	32.00	21.33	16.00	10.67	8.00	5.31	4.00
1500	625.00	214.29	187.50	93.75	46.88	31.25	23.44	15.63	11.72	7.77	5.86
2048	853.33	292.57	256.00	128.00	64.00	42.67	32.00	21.33	16.00	10.61	8.00

Serialization Delay in ms

IP Telephony © 2005 Cisco Systems, Inc. All rights reserved. Cisco Public 60

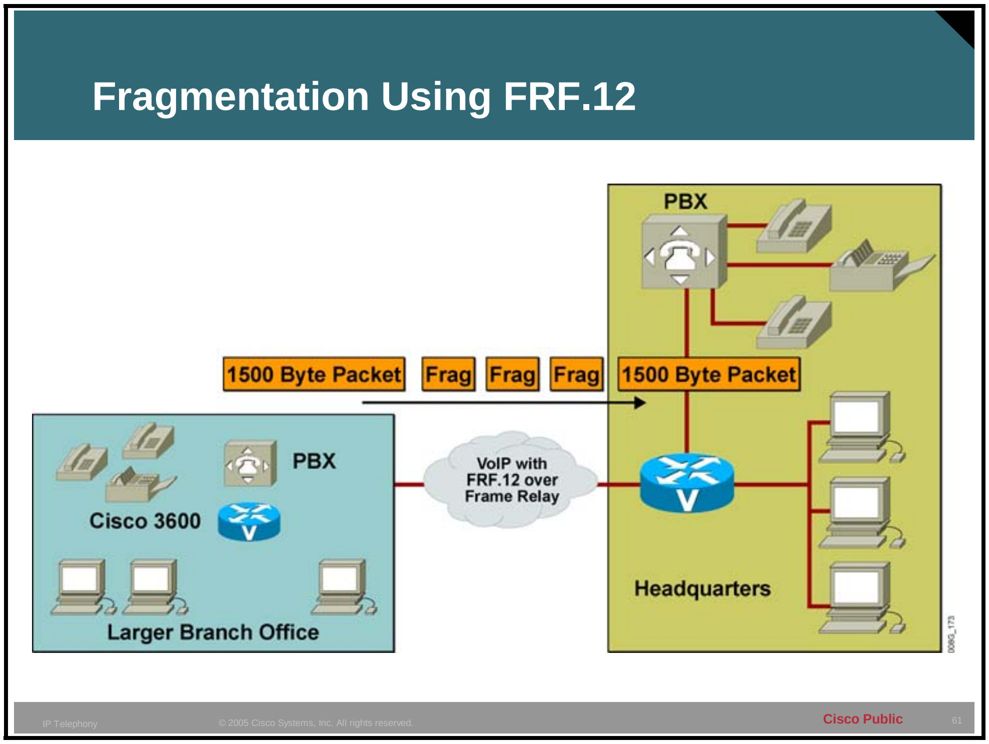
The figure shows the serialization delay required for different frame sizes at various line speeds. This table uses total frame size, not payload size, for computation. As an example, reading from the graphic, on a 64-kbps line, a CS-ACELP voice frame with a length of 38 bytes (37 bytes + 1-byte flag) has a serialization delay of 4.75 ms. If the line speed is increased to 1.544 Mbps, the serialization delay goes down to 0.2 ms. Cisco recommends a 10-ms serialization delay, not to exceed 20 ms.

You can effectively change the serialization delay of a voice packet by performing the following:

- **Increasing the link speed:** Solves the problem, but is expensive
- **Decreasing the packet size:** May not be possible for all codec types, and also increases bandwidth overhead

Managing Queuing Delay

Queuing delay contributes to overall delay. This topic explains queuing delay and offers a standard solution.



Fragmentation Using FRF.12

IP Telephony © 2005 Cisco Systems, Inc. All rights reserved. Cisco Public 61

Queuing delay occurs when other elements in the outbound queue (voice or data packets) cause voice packets to be delayed. For example, the serialization delay for a 1500-byte packet is 214 ms to leave the router over a 56-kbps link. If a 1500-byte data packet that is not real-time is being sent, real-time (voice) packets are queued until the large data packet is transmitted. This delay is unacceptable for voice traffic. If data packets that are not real-time packets are fragmented into smaller frames, they are interleaved with real-time (voice) frames. In this way, both voice and data frames can be carried together on low-speed links without causing excessive delay to the real-time voice traffic. One way to implement this fragmentation is to use FRF.12 on VoIP over Frame Relay networks. FRF.12 serves to fragment Frame Relay frames into smaller frames, even from different permanent virtual circuits (PVCs).

Example: Fragment Size Configuration

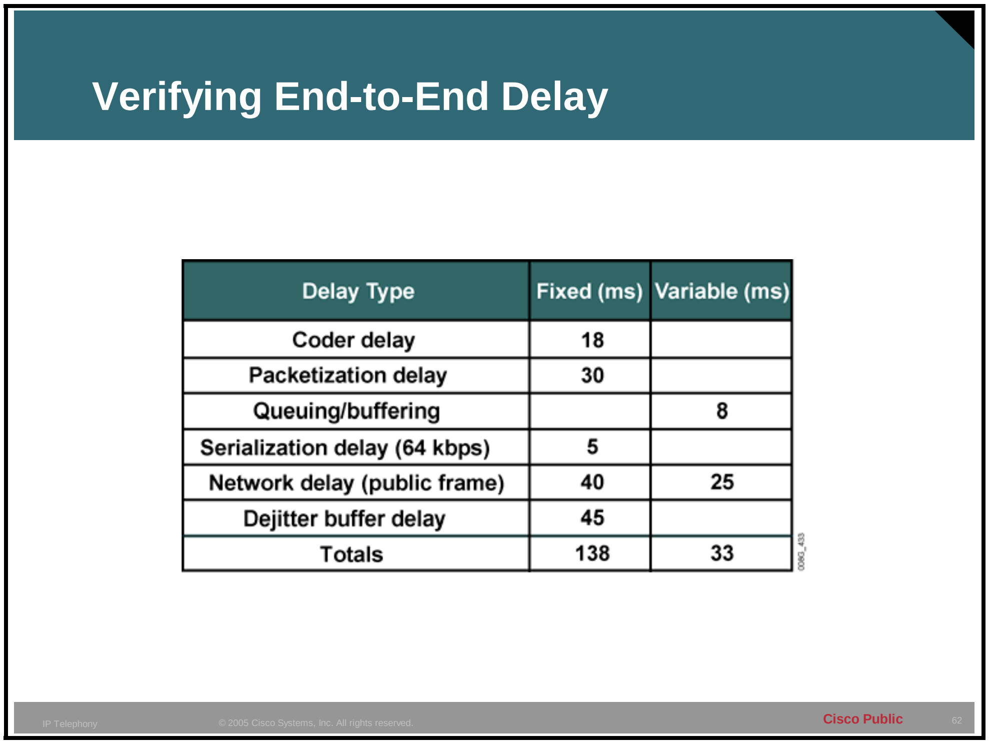
You can configure fragment size using the **frame-relay fragment** *fragment_size* command in a Frame Relay map class. The *fragment_size* argument defines the payload size of a fragment and excludes the Frame Relay headers and any Frame Relay fragmentation header. The valid range is from 16 bytes to 1600 bytes; the default is 53 bytes.

The *fragment_size* argument should be set so that the serialization delay is close to 10 ms; for example, if using a 384-kbps link, the fragmentation size should be set at 512 kbps.

Set the fragmentation size so that the largest data packet is not larger than the voice packets.

Verifying End-to-End Delay

End-to-end delay is calculated and compared to the G.114 recommendation. This topic illustrates the process.



Verifying End-to-End Delay

Delay Type	Fixed (ms)	Variable (ms)
Coder delay	18	
Packetization delay	30	
Queuing/buffering		8
Serialization delay (64 kbps)	5	
Network delay (public frame)	40	25
Dejitter buffer delay	45	
Totals	138	33

006G_433

IP Telephony © 2005 Cisco Systems, Inc. All rights reserved. Cisco Public 62

A typical one-hop connection over a public Frame Relay connection may have the delay budget that is shown in the figure. To calculate one-way delay, simply add all of the contributing components together. The goal is to allow a one-way delay as recommended by G.114.

Example: Verifying End-To-End Delay

The figure shows an acceptable one-way delay of 138 ms, plus 33 ms, for a total of 171 ms.