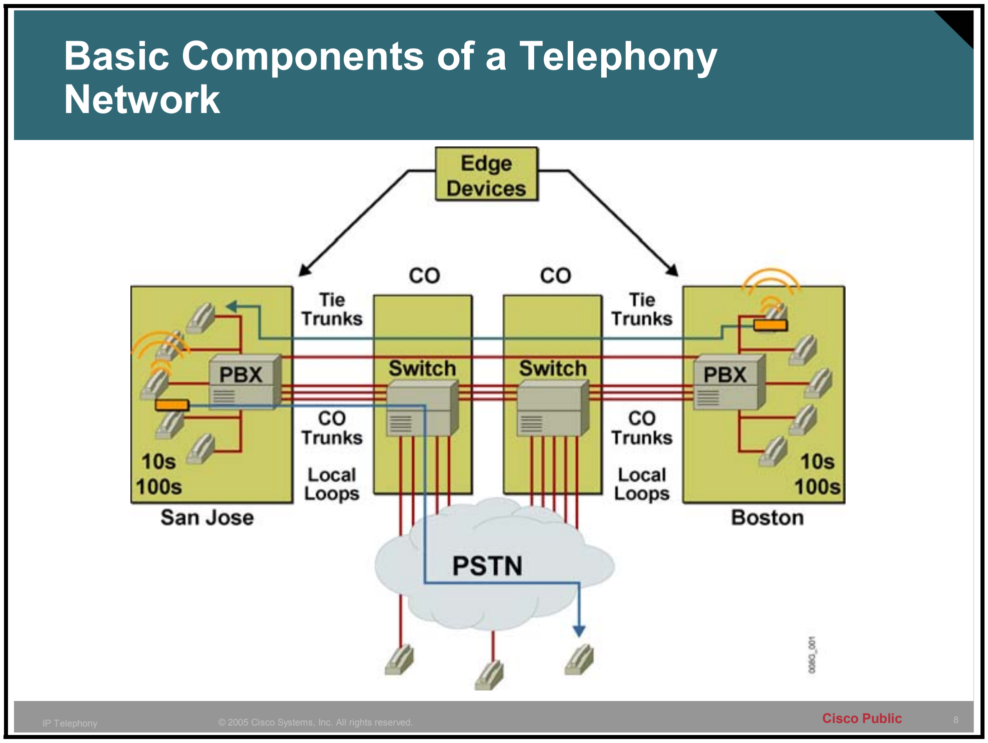
Differences between Traditional Telephony and VoIP

Traditional Telephony

This topic introduces the components of traditional telephony networks.



A number of components must be in place for an end-to-end call to succeed. These components are shown in the figure and include the following:

- Edge devices
- Local loops
- Private or central office (CO) switches
- Trunks

Edge Devices

The two types of edge devices that are used in a telephony network include:

- **Analog telephones:** Analog telephones are most common in home, small office/home office (SOHO), and small business environments. Direct connection to the public switched telephone network (PSTN) is usually made by using analog telephones. Proprietary analog telephones are occasionally used in conjunction with a PBX. These phones provide additional functions such as speakerphone, volume control, PBX message-waiting indicator, call on hold, and personalized ringing.

- **Digital telephones:** Digital telephones contain hardware to convert analog voice into a digitized stream. Larger corporate environments with PBXs generally use digital telephones. Digital telephones are typically proprietary, meaning that they work with the PBX or key system of that vendor only.

Local Loops

A local loop is the interface to the telephone company network. Typically, it is a single pair of wires that carry a single conversation. A home or small business may have multiple local loops.

Private or CO Switches

The CO switch terminates the local loop and handles signaling, digit collection, call routing, call setup, and call teardown.

A PBX switch is a privately owned switch located at the customer site. A PBX typically interfaces with other components to provide additional services; for example, voice mail.

Trunks

The primary function of a trunk is to provide the path between two switches. There are several common trunk types including:

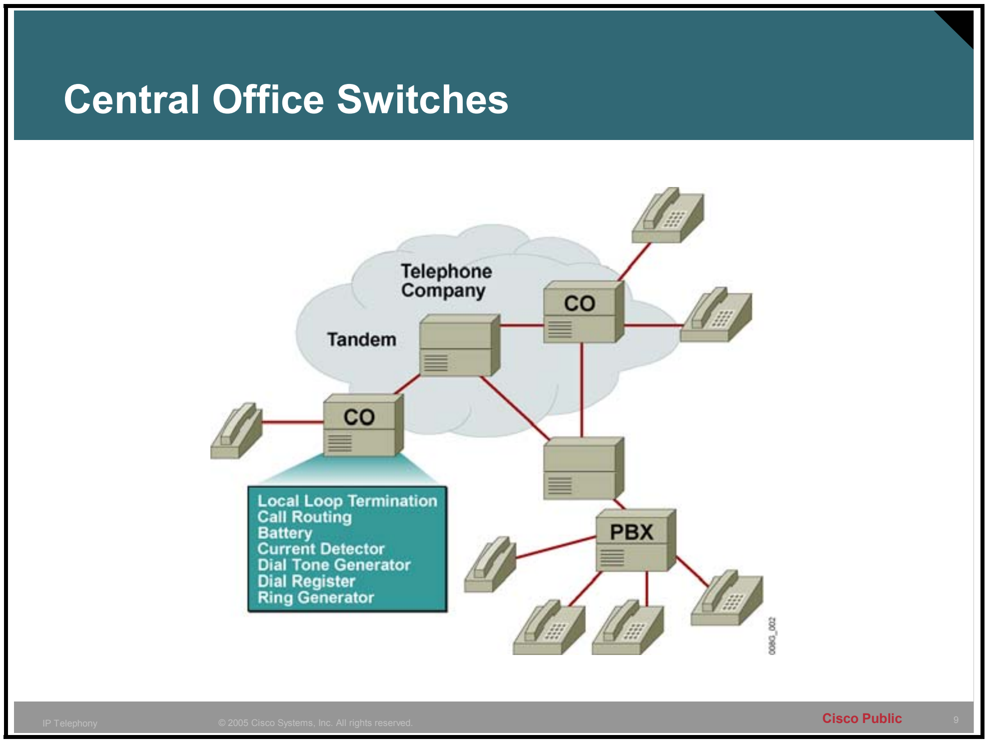
- **Tie trunk:** A dedicated circuit that connects PBXs directly

- **CO trunk:** A direct connection between a local CO and a PBX

- **Interoffice trunk:** A circuit that connects two local telephone company COs.

Traditional Telephony: Central Office Switches

This topic describes how CO switches function and make switching decisions.



The figure shows a typical CO switch environment. The CO switch terminates the local loop and makes the initial call-routing decision.

The call-routing function forwards the call to one of the following:

- Another end-user telephone if it is connected to the same CO
- Another CO switch
- A tandem switch

The CO switch makes the telephone work with the following components:

- **Battery:** The battery is the source of power to both the circuit and the telephone—it determines the status of the circuit. When the handset is lifted to let current flow, the telephone company provides the source that powers the circuit and the telephone. Because the telephone company powers the telephone from the CO, electrical power outages should not affect the basic telephone.

Note Some telephones on the market offer additional features that require a supplementary power source that the subscriber supplies; for example, cordless telephones. Some cordless telephones may lose function during a power outage.

- **Current detector:** The current detector monitors the status of a circuit by detecting whether it is open or closed. The table here describes current flow in a typical telephone.

Table 1: Current Flow in a Typical Telephone

Handset	Circuit	Current Flow
On cradle	On hook/open circuit	No
Off cradle	Off hook/closed circuit	Yes

- **Dial tone generator:** When the digit register is ready, the dial-tone generator produces a dial tone to acknowledge the request for service.

- **Digit register:** The digit register receives the dialed digits.

- **Ring generator:** When the switch detects a call for a specific subscriber, the ring generator alerts the called party by sending a ring signal to that subscriber.

You must configure a PBX connection to a CO switch that matches the signaling of the CO switch. This configuration ensures that the switch and the PBX can detect on hook, off hook, and dialed digits coming from either direction.

CO Switching Systems

Switching systems provide three primary functions:

- Call setup, routing, and teardown

- Call supervision

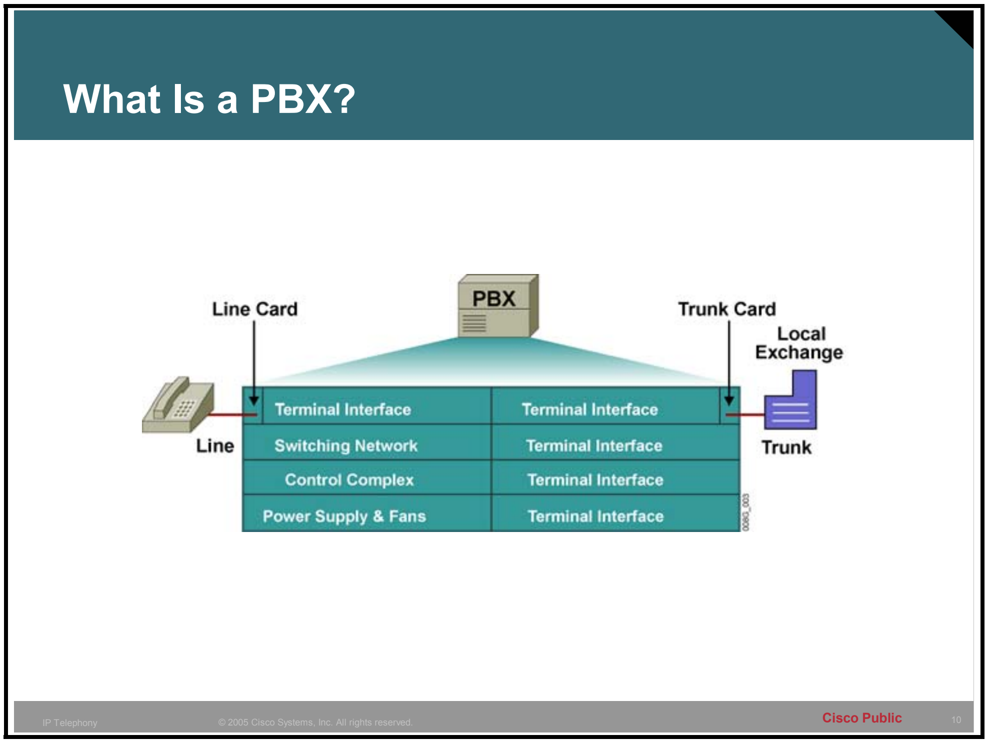
- Customer ID and telephone numbers

CO switches switch calls between locally terminated telephones. If a call recipient is not locally connected, the CO switch decides where to send the call based on its call-routing table. The call then travels over a trunk to another CO or to an intermediate switch that may belong to an inter-exchange carrier (IXC). Although intermediate switches do not provide dial tone, they act as hubs to connect other switches and provide interswitch call routing.

PSTN calls are traditionally circuit-switched, which guarantees end-to-end path and resources. Therefore, as the PSTN sends a call from one switch to another, the same resource is associated with the call until the call is terminated.

Traditional Telephony: PBX and Key Telephone System Functionality

In a corporate environment, where large numbers of staff need access to each other and the outside, individual telephone lines are not economically viable. This topic explores PBX and key telephone system functionality in environments today.



A PBX is a smaller, privately-owned version of the CO switches used by telephone companies.

Most businesses have a PBX telephone system, a key telephone system, or Centrex service. Large offices with more than 50 telephones or handsets choose a PBX to connect users, both in-house and to the PSTN.

PBXs come in a variety of sizes, typically from 20 to 20,000 stations. The selection of a PBX is important to most companies because a PBX has a typical life span of 7 to 10 years.

All PBXs offer a standard, basic set of calling features. Optional software provides additional capabilities.

The figure illustrates the internal components of a PBX: it connects to telephone handsets using line cards and to the local exchange using trunk cards.

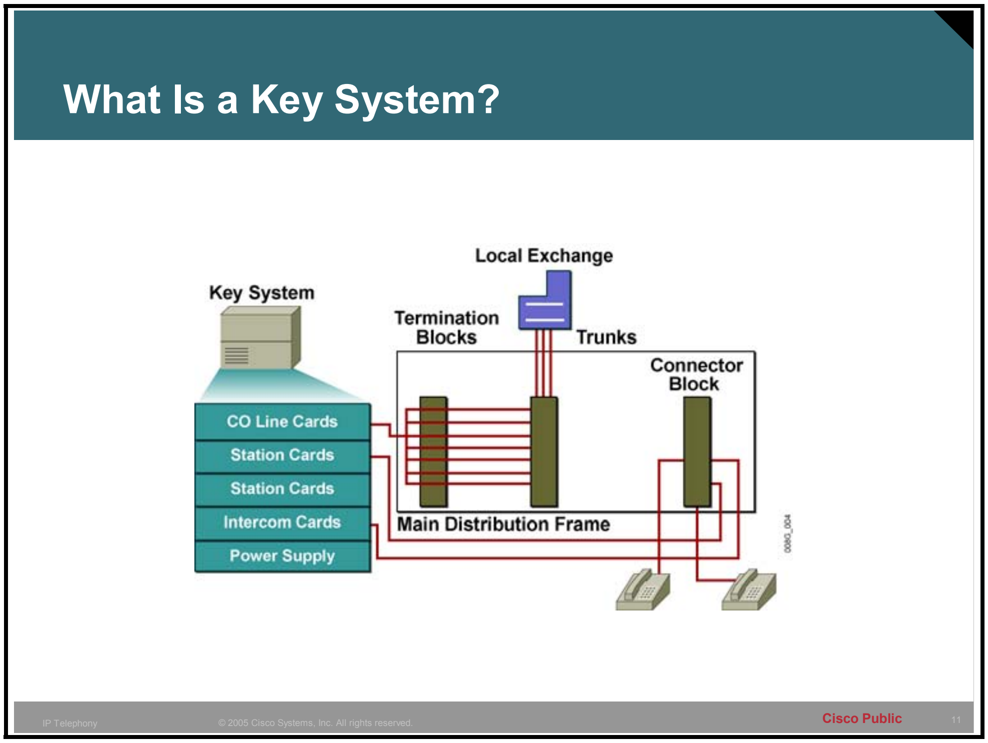
A PBX has three major components:

- **Terminal interface:** The terminal interface provides the connection between terminals and PBX features that reside in the control complex. Terminals can include telephone handsets, trunks, and lines. Common PBX features include dial tone and ringing.

- **Switching network:** The switching network provides the transmission path between two or more terminals in a conversation; for example, two telephones within an office communicate over the switching network.

- **Control complex:** The control complex provides the logic, memory, and processing for call setup, call supervision, and call disconnection.

Traditional Telephony: What Is a Key System



Small organizations and branch offices often use a key telephone system because a PBX offers functionality and extra features that they may not require. For example, a key system offers small businesses distributed answering from any telephone, unlike the central answering position required for a PBX.

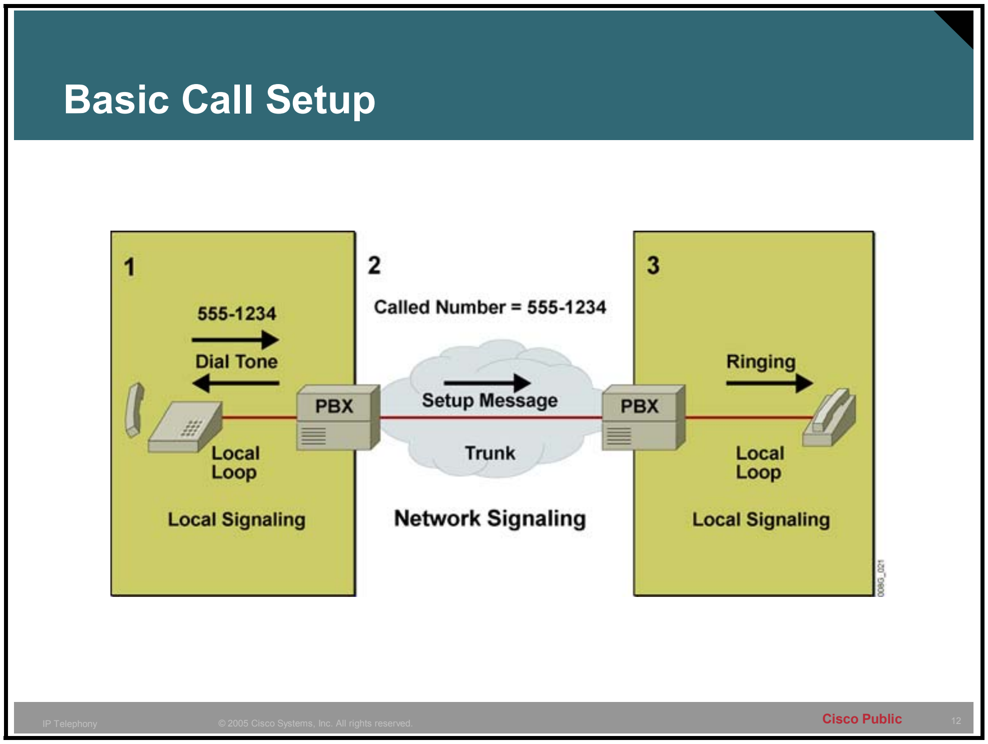
Today, key telephone systems are either analog or digital and are microprocessor-based. Key systems are typically used in offices with 30 to 40 users, but can be scaled to support over 100 users.

A key system has three major components:

- **Key service unit:** A key service unit (KSU) holds the system switching components, power, intercom, line and station cards, and the system logic.
- **System software:** System software provides the operating system and calling-feature software.
- **Telephones (instruments or handsets):** Telephones allow the user to choose a free line and dial out, usually by pressing a button on the telephone.

Traditional Telephony: Basic Call Setup

Call signaling, in its most basic form, is the capacity of a user to communicate a need for service to a network. The call-signaling process requires the ability to detect a request for and termination of service, send addressing information, and provide progress reports to the initiating party. This functionality corresponds to the three call-signaling types discussed in this topic: supervisory, address, and informational signaling.



The figure shows the three major steps in an end-to-end call. These steps include:

Step 1 Local signaling—originating side

The user signals the switch by going off hook and sending dialed digits through the local loop.

Step 2 Network signaling

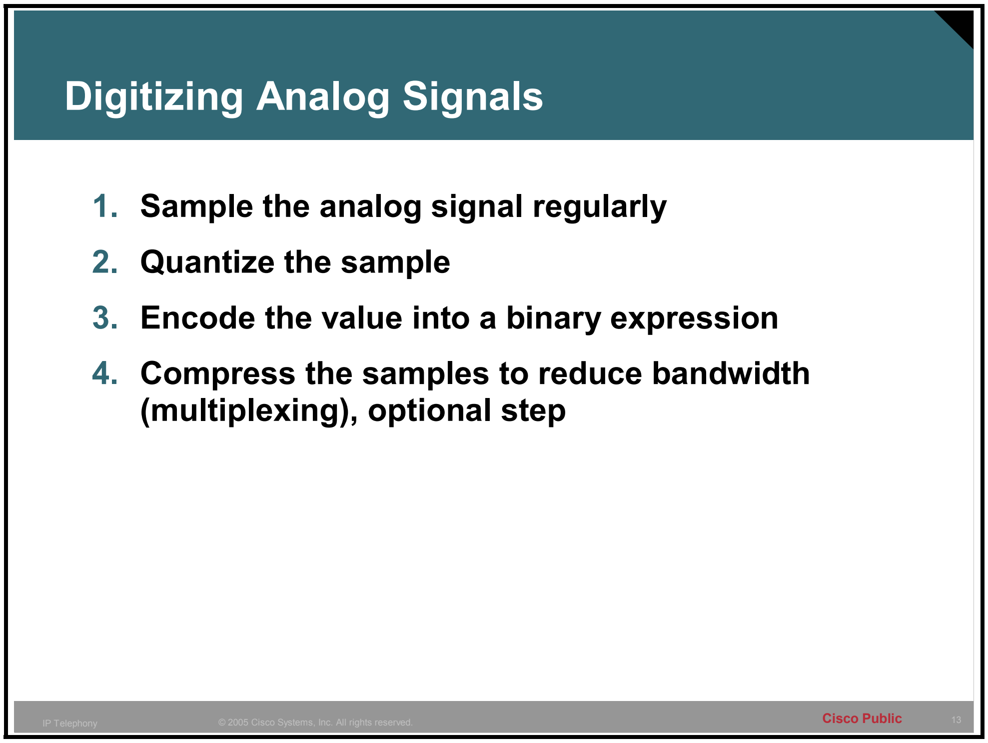
The switch makes a routing decision and signals the next, or terminating, switch through the use of setup messages sent across a trunk.

Step 3 Local signaling—terminating side

The terminating switch signals the call recipient by sending ringing voltage through the local loop to the recipient telephone.

PCM Theory

This topic describes the process of converting analog signals to digital signals.



Digitizing speech was a project first undertaken by the Bell System in the 1950s. The original purpose of digitizing speech was to deploy more voice circuits with a smaller number of wires. This evolved into the T1 and E1 transmission methods of today.

To convert an analog signal to a digital signal, you must perform these steps:

Note The last step is optional.

Table 1: Analog to Digital Signal Conversion

Step	Procedure	Description
1.	Sample the analog signal regularly.	The sampling rate must be two times the highest frequency to produce playback that appears neither choppy nor too smooth.
2.	Quantize the sample.	Quantization consists of a scale made up of 8 major divisions or chords. Each chord is subdivided into 16 equally spaced steps. The chords are not equally spaced but are actually finest near the origin. Steps are equal within the chords but different when they are compared between the chords. Finer graduations at the origin result in less distortion for low-level tones.
3.	Encode the value into 8-bit digital form.	PBX output is a continuous analog voice waveform. T1 digital voice is a snapshot of the wave encoded in ones and zeros.
4.	(Optional) Compress the samples to reduce bandwidth.	Although not essential to convert analog signals to digital, signal compression is widely used to reduce bandwidth.

Three components in the analog-to-digital conversion process include:

- **Sampling:** Sample the analog signal at periodic intervals. The output of sampling is a pulse amplitude modulation (PAM) signal.
- **Quantization:** Match the PAM signal to a segmented scale. This scale measures the amplitude (height) of the PAM signal and assigns an integer number to define that amplitude.
- **Encoding:** Convert the integer base-10 number to a binary number. The output of encoding is a binary expression in which each bit is either a 1 (pulse) or a 0 (no pulse).

This three-step process is repeated 8000 times per second for telephone voice channel service. Use the fourth optional step—compression—to save bandwidth. This optional step allows a single channel to carry more voice calls.

Note The most commonly used method of converting analog to digital is pulse code modulation (PCM).

Basic Voice Encoding: Converting Digital to Analog

This topic describes the process of converting digital signals back to analog signals.

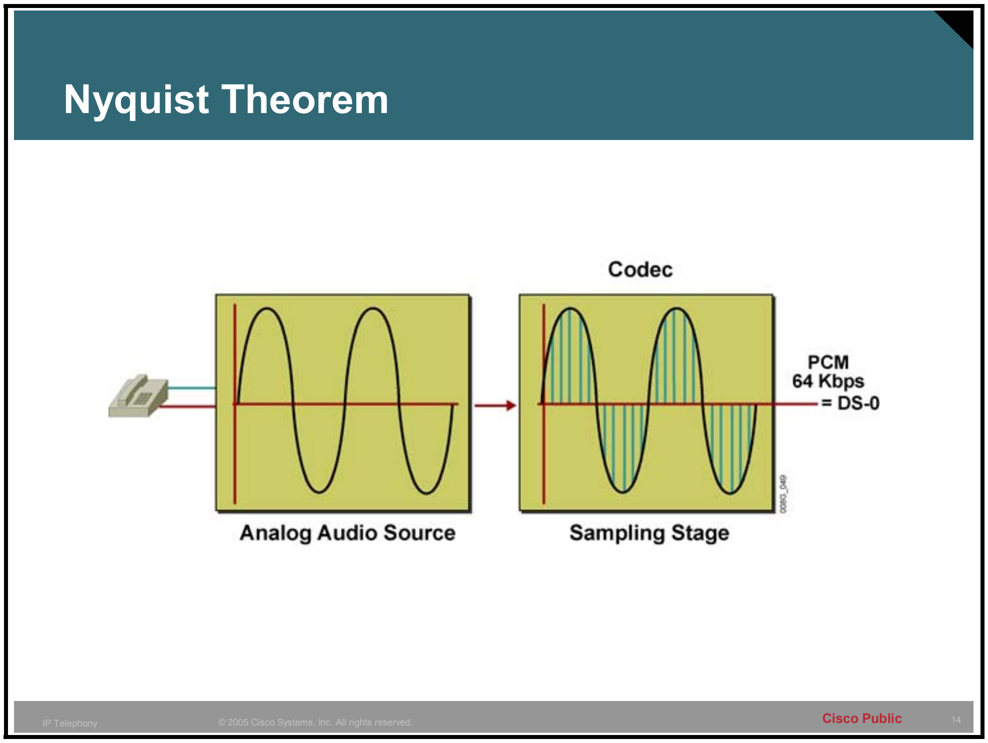
After the receiving terminal at the far end receives the digital PCM signal, it must convert the PCM signal back into an analog signal.

The process of converting digital signals back into analog signals includes the following two parts:

- **Decoding:** The received eight-bit word is decoded to recover the number that defines the amplitude of that sample. This information is used to rebuild a PAM signal of the original amplitude. This process is simply the reverse of the analog-to-digital conversion.
- **Filtering:** The PAM signal is passed through a properly designed filter that reconstructs the original analog wave form from its digitally coded counterpart.

PCM Theory

This topic describes the Nyquist Theorem that is the basis for digital signal technology.



Nyquist Theorem

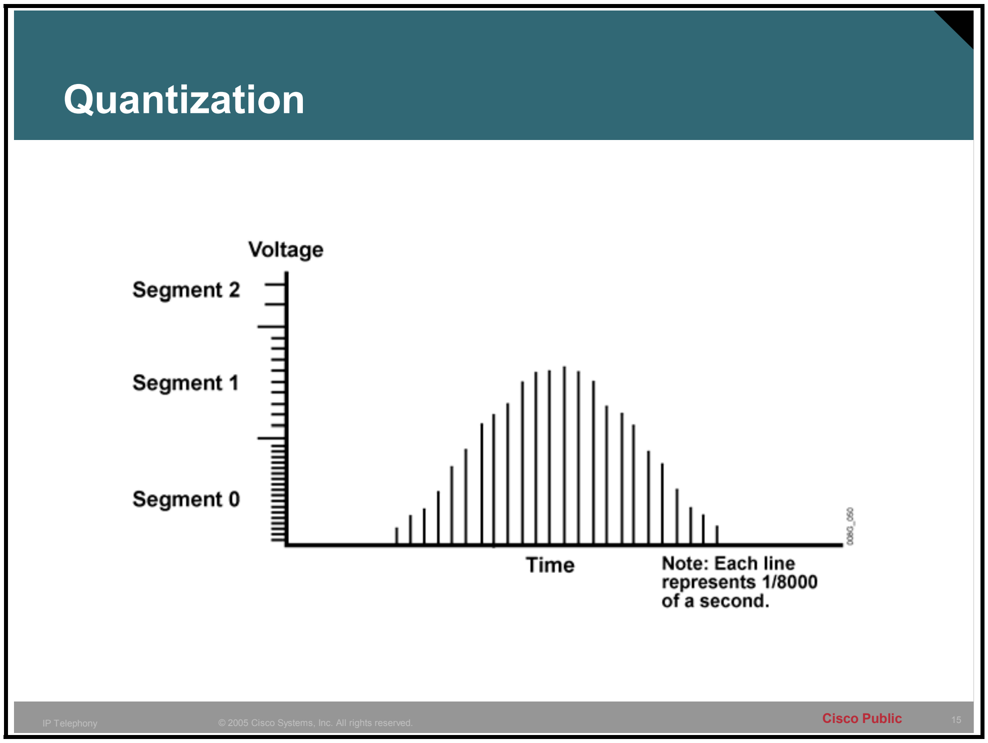
Digital signal technology is based on the premise stated in the Nyquist Theorem: when a signal is instantaneously sampled at the transmitter in regular intervals and has a rate of at least twice the highest channel frequency, then the samples will contain sufficient information to allow an accurate reconstruction of the signal at the receiver.

Example

While the human ear can sense sounds from 20 to 20,000 Hz, and speech encompasses sounds from about 200 to 9000 Hz, the telephone channel was designed to operate at about 300 to 3400 Hz. This economical range carries enough fidelity to allow callers to identify the party at the far end and sense their mood. Nyquist decided to extend the digitization to 4000 Hz, to capture higher-frequency sounds that the telephone channel may deliver. Therefore, the highest frequency for voice is 4000 Hz, or 8000 samples per second; that is, one sample every 125 microseconds.

PCM Theory: Quantization

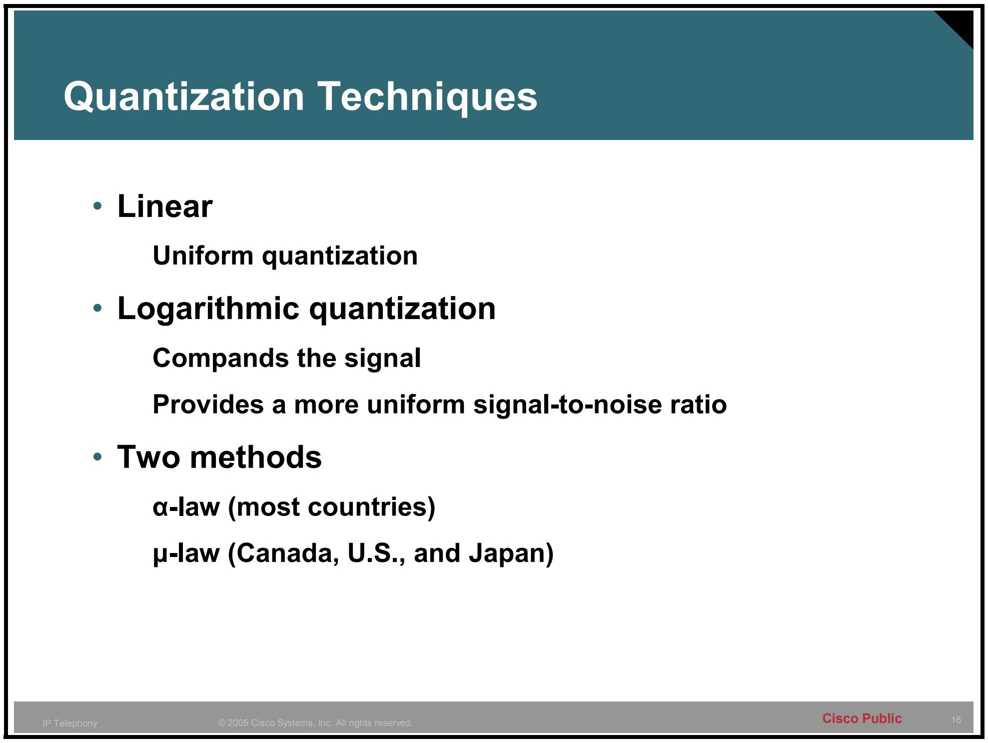
This topic explains quantization and its techniques.



Quantization involves dividing the range of amplitude values that are present in an analog signal sample into a set of discrete steps that are closest in value to the original analog signal. Each step is assigned a unique digital code word.

The figure here depicts quantization. In this example, the x-axis is time and the y-axis is the voltage value (PAM).

The voltage range is divided into 16 segments (0 to 7 positive, and 0 to 7 negative). Starting with segment 0, each segment has fewer steps than the previous segment, which reduces the noise-to-signal ratio and makes it uniform. This segmentation also corresponds closely to the logarithmic behavior of the human ear. If there is a noise-to-signal ratio problem, it is resolved by using a logarithmic scale to convert PAM to PCM.



Linear sampling of analog signals causes small-amplitude signals to have a higher noise-to-signal ratio, and therefore poorer quality than larger amplitude signals. The Bell System developed the *µ-law* method of quantization, which is widely used in North America. The International Telecommunication Union (ITU) modified the original µ-law method and created *α-law*, which is used in countries outside of North America.

By allowing smaller step functions at lower amplitudes—rather than higher amplitudes—µ-law and α-law provide a method of reducing this problem. Both µ-law and α-law compand the signal; for example, they both compress the signal for transmission and then expand the signal back to its original form at the other end.

The result of using µ-law and α-law is a more accurate value for smaller amplitude and uniform signal-to-noise quantization ratio (SQR) across the input range

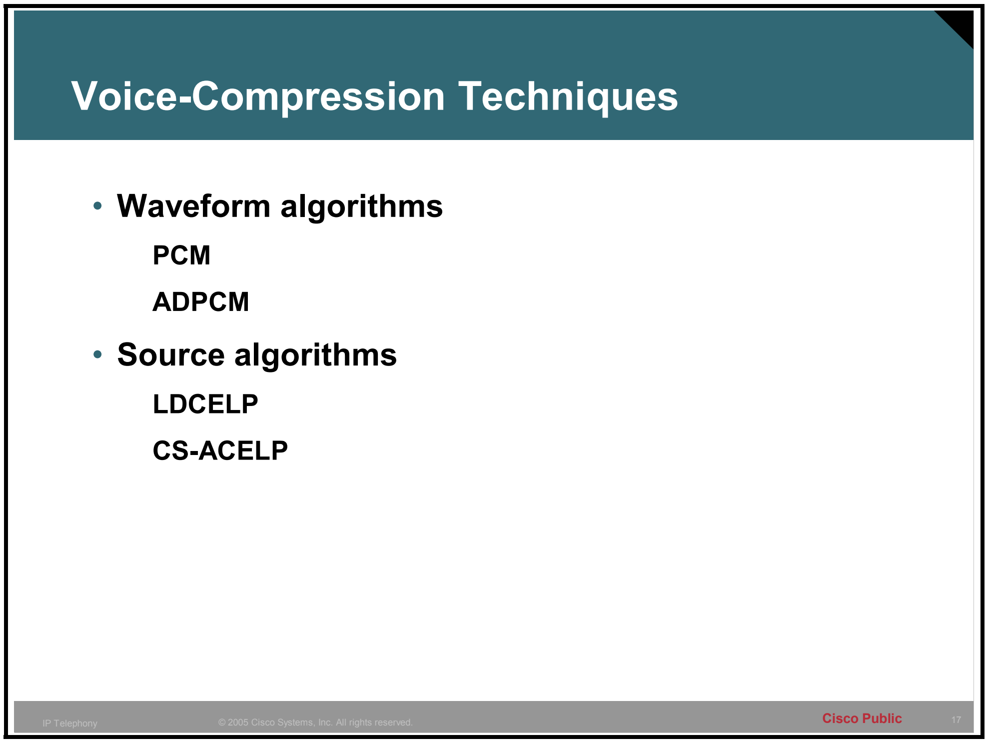
Both µ-law and α-law are linear approximations of a logarithmic input/output relationship. They both generate 64-kbps bit streams using 8-bit code words to segment and quantize levels within segments.

The difference between the original analog signal and the quantization level assigned is called quantization error, which is the source of distortion in digital transmission systems. Quantization error is any random disturbance or signal that interferes with the quality of the transmission or the signal itself.

Note	For communication between a µ-law country and an α-law country, the µ-law country must change its signaling to accommodate the α-law country.

Coder-Decoder

This topic describes two types of speech-coding schemes: waveform and source coding.

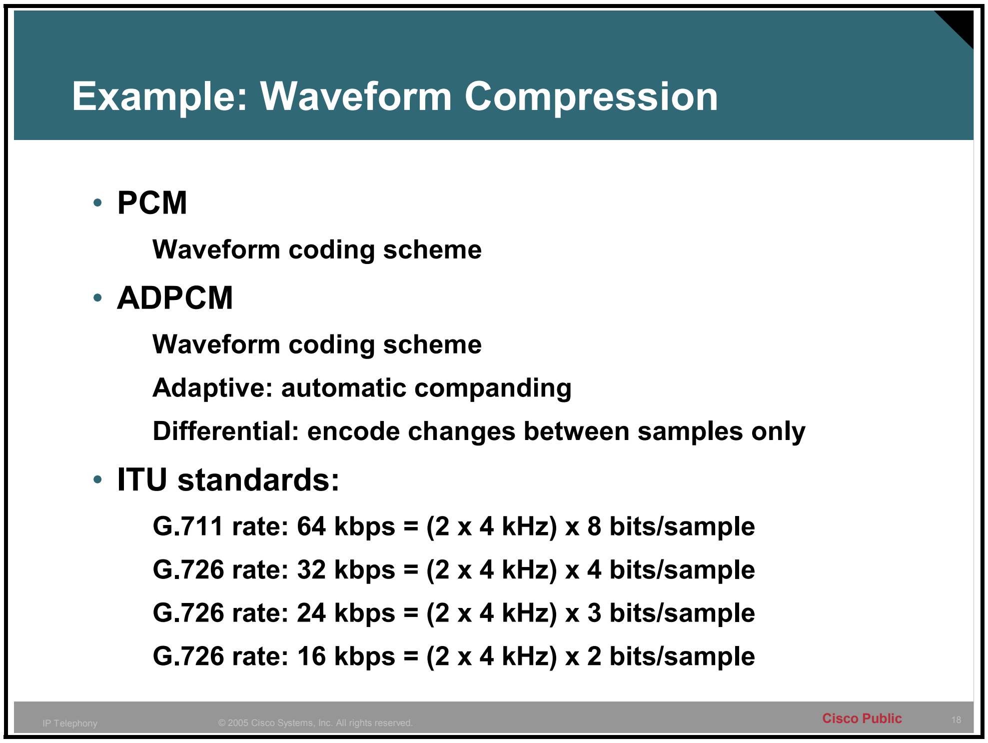


There are two voice compression techniques:

- Waveform algorithms (coders) function as follows:
  - Sample analog signals at 8000 times per second
  - Use predictive differential methods to reduce bandwidth
  - Bandwidth reduction highly impacts voice quality
  - Do not take advantage of speech characteristics
- Source algorithms function as follows:
  - Source algorithm coders are called vocoders. Vocoder is a term that describes 'Voice Coding', which is a device that converts analog speech into digital speech, using a specific compression scheme that is optimized for coding human speech.
  - Vocoders take advantage of speech characteristics.

— Codebooks store specific predictive waveshapes of human speech. They match the speech, encode the phrases, decode the waveshapes at the receiver by looking up the coded phrase, and match it to the stored waveshape in the receiver codebook.

Coder-Decoder: Waveform Compression



Standard PCM is known as ITU standard G.711.

Adaptive differential pulse code modulation (ADPCM) coders, like other waveform coders, encode analog voice signals into digital signals to adaptively predict future encodings by looking at the immediate past. The adaptive feature of ADCPM reduces the number of bits per second that the PCM method requires to encode voice signals.

ADCPM does this by taking 8000 samples per second of the analog voice signal and turning them into a linear PCM sample. ADCPM then calculates the predicted value of the next sample, based on the immediate past sample, and encodes the difference. The ADPCM process generates 4-bit words, therefore generating 16 specific bit patterns.

The ADPCM algorithm from the Consultative Committee for International Telegraph and Telephone (CCITT) transmits all 16 possible bit patterns. The ADPCM algorithm from the American National Standards Institute (ANSI) uses 15 of the 16 possible bit patterns. The ANSI ADPCM algorithm does not generate a 0000 pattern.

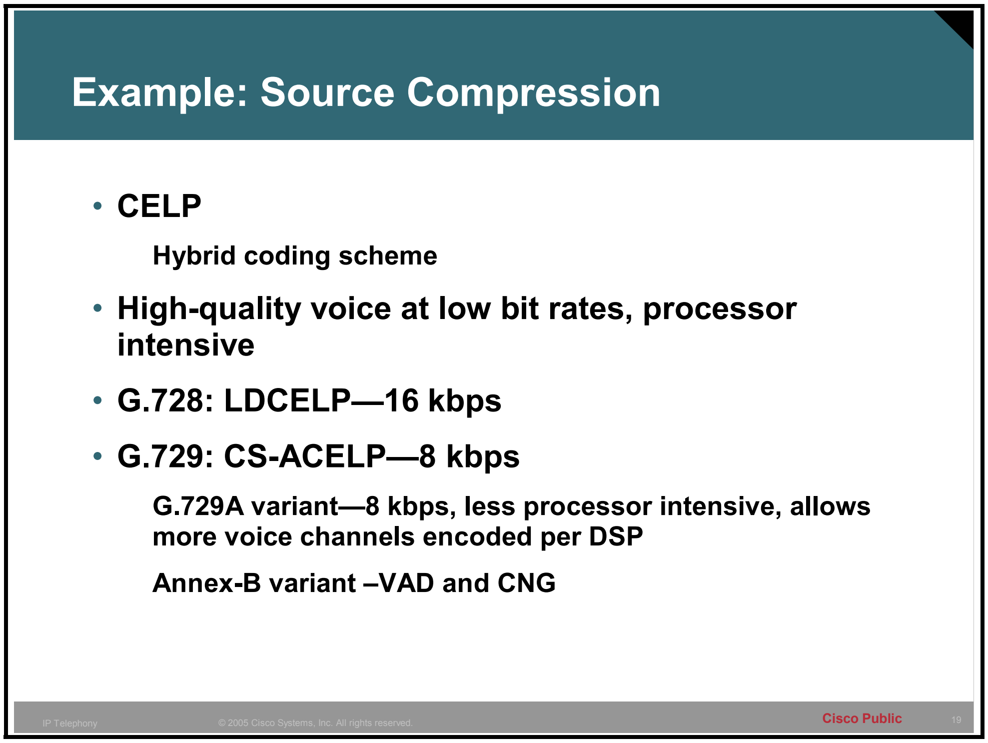
The ITU standards for compression are as follows:

- **G.711 rate:** 64 kbps = (2 x 4 kHz) x 8 bits/sample
- **G.726 rate:** 32 kbps = (2 x 4 kHz) x 4 bits/sample
- **G.726 rate:** 24 kbps = (2 x 4 kHz) x 3 bits/sample

- **G.726 rate:** 16 kbps = (2 x 4 kHz) x 2 bits/sample

Note CCITT is now called ITU-T.

Coder-Decoder: Source Compression



Code excited linear prediction (CELP) compression transforms analog voice signals as follows:

- The input to the coder is converted from an 8-bit PCM to a 16-bit linear PCM sample.
- A codebook uses feedback to continuously learn and predict the voice waveform.
- A white noise generator excites the coder.
- The mathematical result (recipe) is sent to the far-end decoder for synthesis and generation of the voice waveform.

Low-delay CELP (LDCELP) is similar to Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP), except:

- LDCELP uses a smaller codebook and operates at 16 kbps to minimize delay—or look-ahead—to 2 to 5 ms.
- The 10-bit codeword is produced from every five speech samples from the 8-kHz input.
- Four of these 10-bit codewords are called a subframe; they take approximately 2.5 ms to encode.

Two of these subframes are combined into a 5-ms block for transmission. CS-ACELP is a variation of CELP that performs these functions:

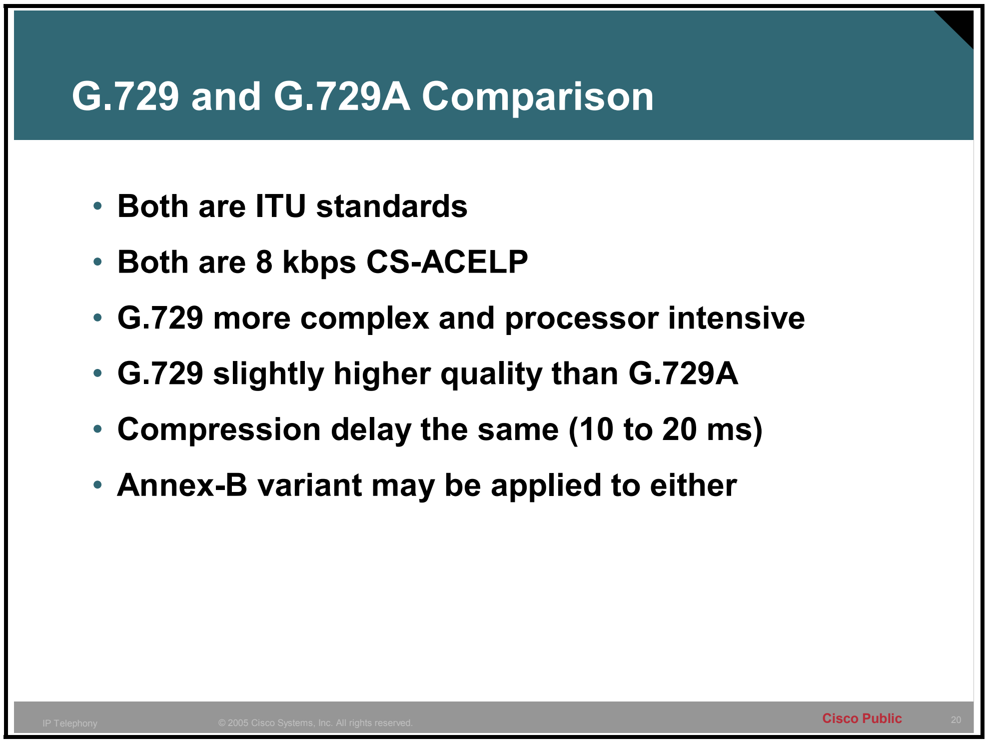
- Codes on 80-byte frames, which take approximately 10 ms to buffer and process
- Adds a look-ahead of 5 ms. A look-ahead is a coding mechanism that continuously analyzes, learns, and predicts the next waveshape.
- Adds noise reduction and pitch-synthesis filtering to processing requirements

Example

The Annex-B variant adds voice activity detection (VAD) in strict compliance with G.729B standards. When this coder-decoder (codec) variant is used, VAD is not tunable for music threshold. However, when Cisco VAD is configured, music threshold is tunable.

Coder-Decoder: G 729 and G 729A Compression

This topic compares G.729 and G.729A compression.



G.729, G.729 Annex-A (G.729A), G.729 Annex-B (G.729B), and G.729A Annex-B (G.729AB) are variations of CS-ACELP.

There is little difference between the ITU recommendations for G.729 and G.729A. All of the platforms that support G.729 also support G.729A.

G.729 is the compression algorithm that Cisco uses for high-quality 8-kbps voice. When properly implemented, G.729 sounds as good as the 32-kbps ADPCM. G.729 is a high-complexity, processor-intensive, compression algorithm that monopolizes processing resources.

Although G.729A is also an 8-kbps compression, it is not as processor-intensive as G.729. It is a medium-complexity variant of G.729 with slightly lower voice quality. The quality of G.729A is not as high as G.729 and is more susceptible to network irregularities such as delay, variation, and tandeming. Tandeming causes distortion that occurs when speech is coded, decoded, and then coded and decoded again, much like the distortion that occurs when a videotape is repeatedly copied.

Example

On Cisco IOS® gateways, you must use the variant (G.729 or G.729A) that is related to the codec complexity configuration on the voice card. This variant does not show up explicitly in the Cisco IOS command-line interface (CLI) codec choice. For example, the CLI does not display **g729r8** (alpha code) as a codec option. However, if the voice card is defined as medium-complexity, then the **g729r8** option is the G.729A codec.

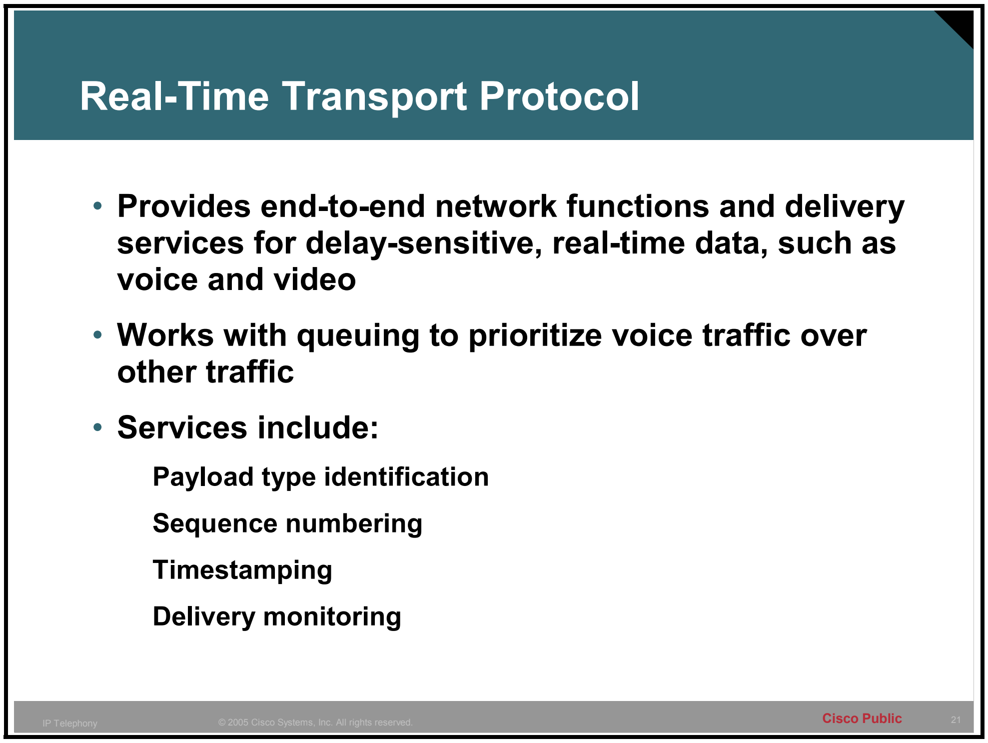
G.729B is a high-complexity algorithm and G.729AB is a medium-complexity variant of G.729B with slightly lower voice quality. The difference between the G.729 and G.729B codec is that the G.729B codec provides built-in Internet Engineering Task Force (IETF) VAD and comfort noise generation (CNG).

The following G.729 codec combinations interoperate:

- G.729 and G.729A
- G.729 and G.729
- G.729A and G.729A
- G.729B and G.729AB
- G.729B and G.729B
- G.729AB and G.729AB

Encapsulating Voice in IP Packets

This topic describes the functions of RTP and RTCP as they relate to the VoIP network.



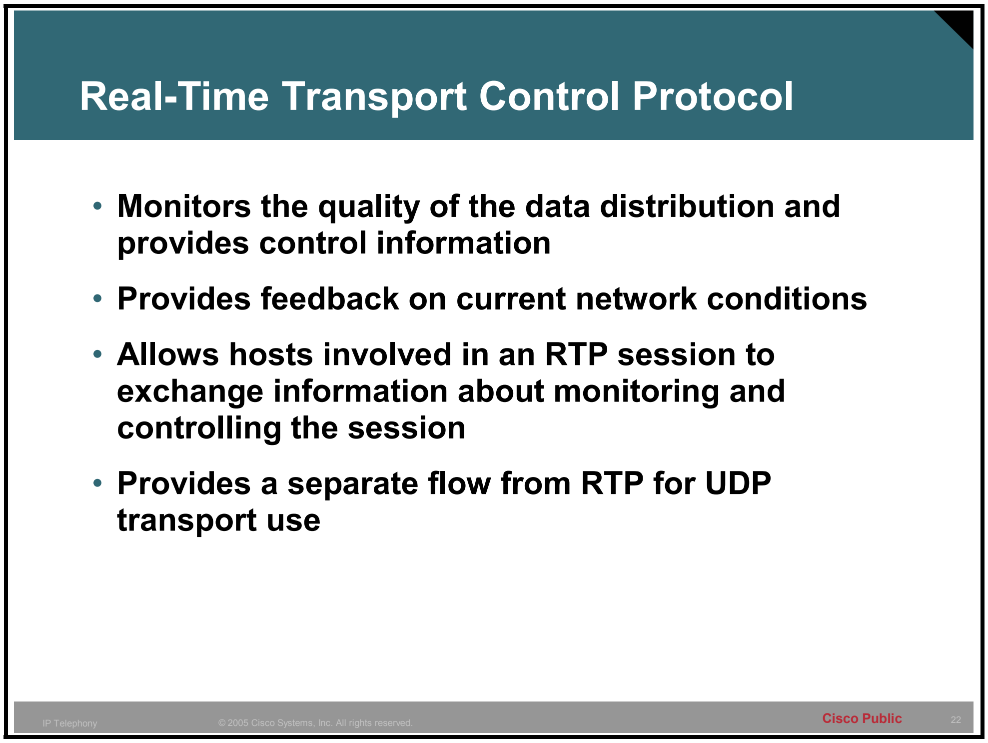
RTP provides end-to-end network transport functions intended for applications transmitting real-time requirements, such as audio and video. Those functions include payload type identification, sequence numbering, time stamping, and delivery monitoring.

RTP typically runs on top of UDP to utilize the multiplexing and checksum services of that protocol. Although RTP is often used for unicast sessions, it is primarily designed for multicast sessions. In addition to the roles of sender and receiver, RTP also defines the roles of translator and mixer to support the multicast requirements.

Example

RTP is a critical component of VoIP because it enables the destination device to reorder and retime the voice packets before they are played out to the user. An RTP header contains a time stamp and sequence number, which allows the receiving device to buffer and remove jitter and latency by synchronizing the packets to play back a continuous stream of sound. RTP uses sequence numbers to order the packets only. RTP does not request retransmission if a packet is lost.

For more information on RTP, refer to RFC 1889.



RTCP monitors the quality of the data distribution and provides control information. RTCP provides the following feedback on current network conditions:

- RTCP provides a mechanism for hosts involved in an RTP session to exchange information about monitoring and controlling the session. RTCP monitors the quality of elements such as packet count, packet loss, delay, and inter-arrival jitter. RTCP transmits packets as a percentage of session bandwidth, but at a specific rate of at least every 5 seconds.

- The RTP standard states that the Network Time Protocol (NTP) time stamp is based on synchronized clocks. The corresponding RTP time stamp is randomly generated and based on data-packet sampling. Both NTP and RTP are included in RTCP packets by the sender of the data.

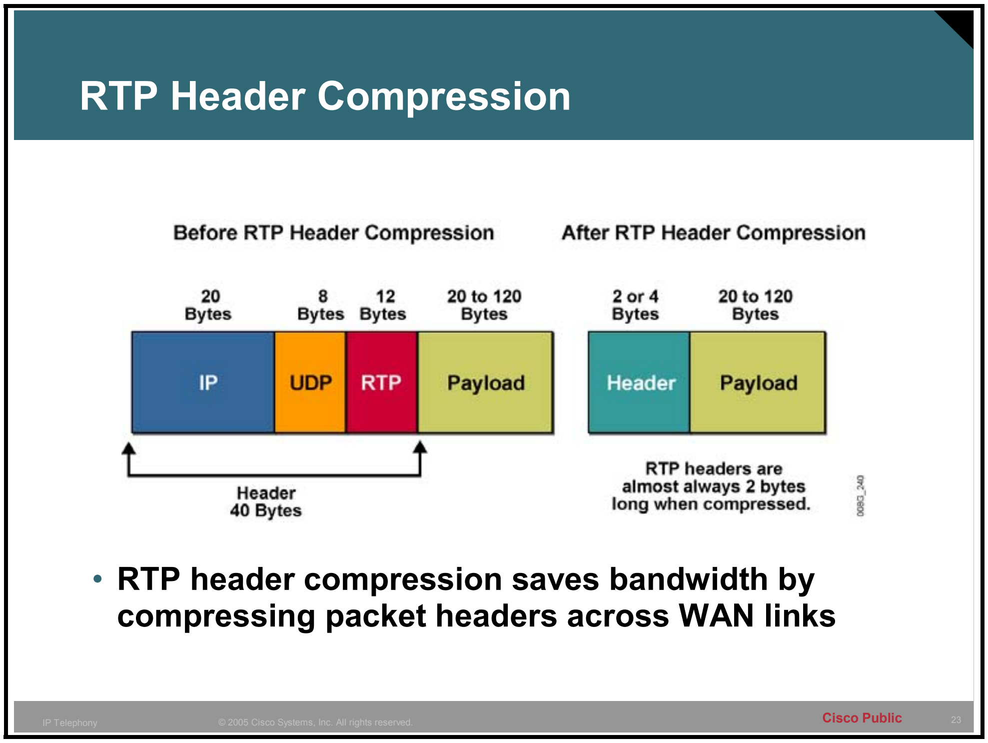
- RTCP provides a separate flow from RTP for transport use by UDP. When a voice stream is assigned UDP port numbers, RTP is typically assigned an even-numbered port and RTCP is assigned the next odd-numbered port. Each voice call has four ports assigned: RTP plus RTCP in the transmit directions and RTP plus RTCP in the receive direction.

Example

Throughout the duration of each RTP call, the RTCP report packets are generated at least every 5 seconds. In the event of poor network conditions, a call may be disconnected due to high packet loss. When viewing packets using a packet analyzer, a network administrator could check information in the RTCP header that includes packet count, octet count, number of packets lost, and jitter. The RTCP header information would shed light on why the calls were disconnected.

Encapsulating Voice in IP Packets: Compressed Real-Time Transport Protocol (CRTP)

This topic describes how IP voice headers are compressed using CRTP.



Given the number of multiple protocols that are necessary to transport voice over an IP network, the packet header can be large. You can use cRTP headers on a link-by-link basis to save bandwidth.

Using CRTP compresses the IP/UDP/RTP header from 40 bytes to 2 bytes without UDP checksums and from 40 bytes to 4 bytes with UDP checksums. RTP header compression is especially beneficial when the RTP payload size is small; for example, with compressed audio payloads are 20 and 50 bytes.

In addition, CRTP works on the premise that most of the fields in the IP/UDP/RTP header do not change, or that the change is predictable. Static fields include source and destination IP address, source and destination UDP port numbers, as well as many other fields in all three headers. For those fields where the change is predictable, the CRTP process is illustrated in the following table:

Table 1: CRTP

Stage	What Happens
The change is predictable.	The sending side tracks the predicted change.
The predicted change is tracked.	The sending side sends a hash of the header.
The receiving side predicts what the constant change is.	The receiving side substitutes the original stored header and calculates the changed fields.
There is an unexpected change.	The sending side sends the entire header without compression.

RTP Packet Components

In a packet voice environment using G.729 and when speech samples are framed every 20 ms, a payload of 20 bytes is generated. Without CRTP, the total packet size includes the following components:

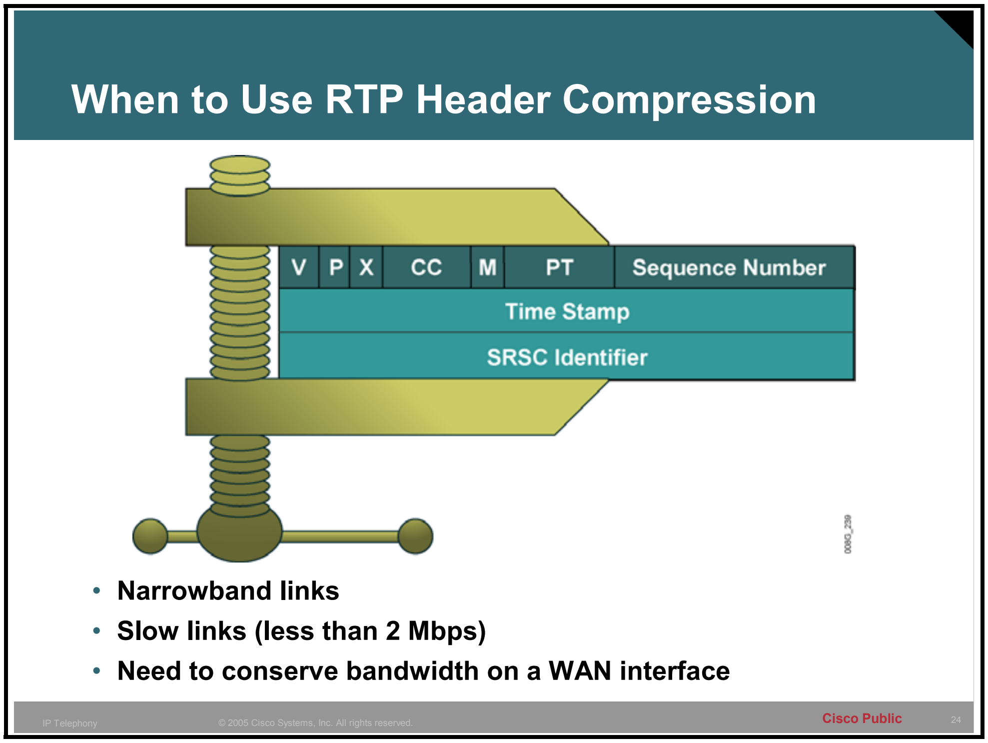
- IP header (20 bytes)
- UDP header (8 bytes)
- RTP header (12 bytes)
- Payload (20 bytes)

The header is twice the size of the payload; IP/UDP/RTP (20 + 8 + 12 = 40 bytes) vs. payload (20 bytes). When generating packets every 20 ms on a slow link, the header consumes a large portion of bandwidth.

In the figure, RTP header compression reduces the header to 2 bytes. The compressed header is 1/10th the payload size.

Encapsulating Voice in IP Packets: Using CRTP

This topic describes when to use CRTP.



You must configure CRTP on a specific serial interface or subinterface if you have any of these conditions:

- Congested WAN links
- Slow links (less than 2 Mbps)
- Need to conserve bandwidth on a WAN interface

Compression works on a link-by-link basis and must be enabled for each link that fits these requirements. You must enable compression on both sides of the link for proper results. Enabling compression on both ends of a low-bandwidth serial link can greatly reduce the network overhead if there is a significant volume of RTP traffic on that slow link.

Note	Compression adds to processing overhead. You must check resource availability on each device prior to turning on RTP header compression.

Example

If you want the router to compress RTP packets, use the **ip rtp header-compression** command. The **ip rtp header-compression** command defaults to active mode when it is configured. However, this command provides a passive mode setting in instances where you want the router to compress RTP packets *only* if it has received compressed RTP on that interface. When applying to a Frame Relay interface, use the **frame-relay ip rtp header-compression** command.

By default, the software supports a total of 16 RTP header compression connections on an interface. Depending on the traffic on the interface, you can change the number of header compression connections with the **ip rtp compression-connections** ***number*** command.

Note	Do not use CRTP if the link is faster than 2 Mbps.

