Synopsis of Basic VoIP Concepts

The Catalyst 4224 Access Gateway Switch (Catalyst 4224) provides Voice over IP (VoIP) gateway applications for a *micro branch* office. This chapter introduces some basic VoIP concepts.

This chapter contains these sections:

VoIP Overview

The VoIP application allows a Catalyst 4224 to convert analog voice signals such as telephone calls and faxes into digital IP packets and distribute these packets across a WAN. In VoIP technology, a digital signal processor (DSP) segments the voice signals into frames and stores them in packets. These packets are transported using IP in compliance with the International Telecommunication Union-Telecommunication Standardization Sector's (ITU-T's) specification H.323, the specification for transmitting multimedia (voice, video, and data) across a network.

Because VoIP is a delay-sensitive application, you need to fine-tune your network from end to end before implementing VoIP. Fine-tuning your network to support VoIP incorporates a series of protocols and features that improve quality of service (QoS). Furthermore, you must take traffic shaping into account to ensure the reliability of VoIP.

A Voice Primer

This section describes some basic telephony concepts that might help you understand VoIP:

How VoIP Processes a Typical Telephone Call

The general flow of a two-party call follows this process:

1. The caller picks up the handset. This signals an off-hook condition to the VoIP signaling application in the Catalyst 4224.
2. The session application issues a dial tone and waits for the caller to dial a telephone number.
3. The caller dials the telephone number. The session application stores the dialed digits.
4. After enough digits are accumulated to match a configured destination pattern, the telephone number is mapped to an IP host via the dial plan mapper. The IP host has a direct connection to either the destination telephone number or a private branch exchange (PBX) that is responsible for completing the call to the configured destination pattern.
5. The session application runs the H.323 session protocol to establish transmission and reception channels for each direction over the IP network. If the call is being handled by a PBX, the PBX forwards the call to the destination telephone. If Resource Reservation Protocol (RSVP) has been configured, RSVP reservations are put into effect to achieve the desired quality of service (QoS) over the IP network.

6. The coder-decoder compression schemes (codecs) are enabled for both ends of the connection using Real-Time Transport Protocol/User Datagram Protocol/Internet Protocol (RTP/UDP/IP) as the protocol stack.
7. Any call-progress indications (or other signals that can be carried inband) are cut through the voice path as soon as an end-to-end audio channel is established. Signaling that can be detected by the voice ports is also trapped by the session application at each end of the connection. Signaling carried over the IP network is encapsulated in Real-Time Transport Control Protocol (RTCP) using the RTCP application-defined (APP) extension mechanism.
8. When either person hangs up the phone, RSVP reservations are torn down (if RSVP is used) and the session ends. Each end becomes idle, waiting for the next off-hook condition to trigger another call setup.

Numbering Scheme

The standard Public Switched Telephone Network (PSTN) is a large, circuit-switched network. It uses a specific numbering scheme, which complies with the ITU-T's international public telecommunications numbering plan (E.164) recommendations. For example, in North America, the North American Numbering Plan (NANP) is used. NANP consists of an area code, an office code, and a station code:

- Area codes are assigned geographically.
- Office codes are assigned to specific switches.
- Station codes identify a specific port on that switch.

The format in North America is 1Nxx-Nxx-xxxx, where N = digits 2 through 9, and x = digits 0 through 9. Internationally, each country is assigned a one- to three-digit country code, and the country's dialing plan is dictated by the country code.

Analog Versus Digital

Analog transmission is not robust or efficient at recovering from line noise. Because analog signals degrade over distance, they need to be amplified periodically. This amplification boosts both the voice signal and the ambient line noise, resulting in degradation of the quality of the transmitted sound.

In response to the limitations of analog transmission, the telephony industry migrated to digital transmission using pulse code modulation (PCM) or adaptive differential PCM (ADPCM). In both cases, analog sound is converted into digital form by sampling the analog sound 8000 times per second and converting each sample into a numeric code.

codecs

PCM and ADPCM are examples of *waveform* codec and are compression techniques that exploit the redundant characteristics of the waveform itself. In addition to waveform codecs, there are source codecs that compress speech by sending only simplified parametric information about voice transmission. Thus, these codecs require less bandwidth. Source codecs include linear predictive coding (LPC), code-excited linear prediction (CELP) and multipulse-multilevel quantization (MP-MLQ).

Coding techniques for telephony and voice packet are standardized by the ITU-T in its G-series recommendations. The Catalyst 4224 uses the following coding standards:

- G.711—Describes the 64-kbps PCM voice coding technique. In G.711, encoded voice is already in the correct format for digital voice delivery in the PSTN or through PBXs.
- G.729—Describes CELP compression where voice is coded into 8-kbps streams. There are two variations of this standard (G.729 and G.729 Annex A) that differ mainly in computational complexity, but both provide speech quality similar to 32-kbps ADPCM.

Mean Opinion Score

Each codec provides a certain quality of speech. The quality of transmitted speech is a subjective response of the listener. A common benchmark used to determine the quality of sound produced by specific codecs is the mean opinion score (MOS). With MOS, a wide range of listeners judge the quality of a voice sample (corresponding to a particular codec) on a scale of 1 (bad) to 5 (excellent). The scores are averaged to provide the MOS for that sample. Table B-1 shows the relationship between codecs and MOS scores.

Table B-1 Compression Methods and MOS Scores

Compression Method	Bit Rate (kbps)	Framing Size (ms)	MOS Score
G.711 PCM	64	0.125	4.1
G.729 CS-ACELP[1]	8	10	3.92
G.729 x 2 encodings[2]	8	10	3.27
G.729 x 3 encodings	8	10	2.68
G.729a[3] CS-ACELP	8	10	3.7

1. Conjugate structure-algebraic code-excited linear prediction.
2. A G.729 voice signal is tandem-encoded two times.
3. G.729 Annex A.

Although it might seem logical from a financial standpoint to convert all calls to low bit-rate codecs to save on infrastructure costs, you should be aware of the drawbacks of designing voice networks with low bit-rate compression. One of the main drawbacks is signal distortion due to multiple encodings (called tandem encodings). For example, when a G.729 voice signal is tandem-encoded three times, the MOS score drops from 3.92 (very good) to 2.68 (unacceptable). Another drawback of low bit-rate codecs is codec-induced delay.

Delay

One of the most important design considerations in implementing voice is minimizing one-way, end-to-end delay. Voice traffic is real-time traffic; if there is too long a delay in voice packet delivery, speech becomes unrecognizable. Delay is inherent in voice networking and is caused by a number of different factors. An acceptable delay is less than 200 milliseconds.

There are two kinds of delays inherent in today's telephony networks: propagation delay and handling delay.

Propagation delay is caused by the characteristics of the speed of light traveling via a fiber-optic-based or copper-based medium.

Handling delay (sometimes called serialization delay) is caused by the devices that handle voice information. Handling delays significantly degrade voice quality in a packet network.

Delays caused by codecs are considered handling delays. Table B-2 shows the delay introduced by different codecs.

Table B-2 codec-Induced Delays

codec	Bit Rate (kbps)	Framing size (ms)	Compression Delay (ms)
G.711 PCM	64	0.125	5
G.729 CS-ACELP	8	10	15
G.729a CS-ACELP	8	10	15

Another handling delay is the time it takes to generate a voice packet. In VoIP, the DSP generates a frame every 10 milliseconds. Two of these frames are then placed within one voice packet, so the packet delay is 20 milliseconds.

Another source of handling delay is the time it takes to move the packet to the output queue. Cisco IOS software expedites the process of determining packet destination and getting the packet to the output queue. The actual delay at the output queue is another source of handling delay and should be kept under 10 milliseconds whenever possible by using queuing methods that are optimal for your network.

In Voice over Frame Relay, you need to make sure that voice traffic is not crowded out by data traffic.

Jitter

Jitter is another factor that affects delay. Jitter is the variation between the time a voice packet is expected to be received and when it actually is received, causing discontinuity in the real-time voice stream. Voice devices such as the Cisco 3600, Cisco MC3810, and the Catalyst 4224 compensate for jitter by setting up a playout buffer to play back voice in a smooth fashion.

Playout control is handled through RTP encapsulation, either by selecting adaptive or non-adaptive playout-delay mode. In either mode, the default value for nominal delay is sufficient.

End-to-End Delay

Figuring out the end-to-end delay is not difficult if you know the end-to-end signal paths/data paths, the codec, and the payload size of the packets. Adding the delays from the endpoints to the codecs at both ends, the encoder delay (which is 5 milliseconds for the G.711 and G.726 codecs and 10 milliseconds for the G.729 codec), the packet delay, and the fixed portion of the network delay yields the end-to-end delay for the connection.

Echo

Echo is hearing your own voice in the telephone receiver while you are talking. When timed properly, echo is reassuring to the speaker. But if the echo exceeds approximately 25 milliseconds, it can be distracting and cause breaks in the conversation.

In a traditional telephony network, echo is normally caused by a mismatch in impedance from the four-wire network switch conversion to the two-wire local loop and is controlled by echo cancellers. In voice-packet-based networks, echo cancellers are built into the low bit-rate codecs and are operated on each DSP. Echo cancellers are, by design, limited by the total amount of time they will wait for the reflected speech to be received. This amount of time is called an *echo trail*. The echo trail is normally 32 milliseconds. VoIP has configurable echo trails of 8, 16, 24, and 32 milliseconds.

Signaling

Although there are various types of signaling used in telecommunications today, this document describes only those with direct applicability to Cisco voice implementations. One signaling type involves access signaling, which determines

when a line has gone off-hook or on-hook. Foreign Exchange Station (FXS) and Foreign Exchange Office (FXO) are types of access signaling. There are two common methods of providing this basic signal:

- Loop start is the most common technique for access signaling in a standard PSTN end-loop network. When a handset is picked up (goes off-hook), this action closes the circuit that draws current from the telephone company's central office (CO), indicating a change in status. This change in status signals the CO to provide a dial tone. An incoming call is signalled from the CO to the handset by a standard on/off pattern signal, causing the telephone to ring.
- Ground start is another access signaling method used to indicate on-hook/off-hook status to the CO, but this signaling method is primarily used on trunk lines or tie-lines between PBXs. Ground-start signaling works through ground and current detectors, allowing the network to indicate off-hook or seizure of an incoming call independent of the ringing signal.

Another signaling technique used mainly between PBXs or other network-to-network telephony switches is known as Ear and Mouth (E&M). There are five types of E&M signaling, as well as two different wiring methods.