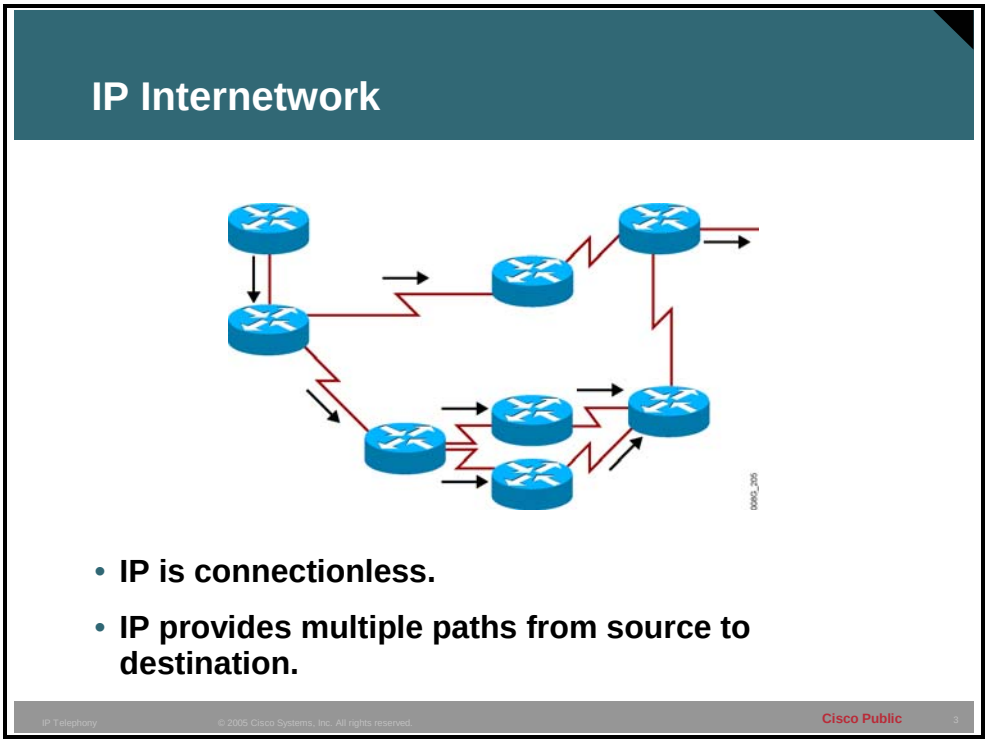
Requirements of Voice in an IP Internetwork

Real-Time Voice in a Best-Effort IP Internetwork

This topic lists problems associated with implementation of real-time voice traffic in a best-effort IP internetwork.



The traditional telephony network was originally designed to carry voice. The design of circuit-switched calls provides a guaranteed path and a delay threshold between source and destination. The IP network was originally designed to carry data. Data networks were not designed to carry voice traffic. Although data traffic is best-effort traffic and can withstand some amount of delay, jitter, and loss, voice traffic is real-time traffic that requires a certain quality of service (QoS). In the absence of any special QoS parameters, a voice packet is treated as just another data packet.

The user must have a well-engineered network, end to end, when running delay-sensitive applications such as VoIP. Fine-tuning the network to adequately support VoIP involves a series of protocols and features geared toward QoS.

Example: Real-Time Voice Delivery Issues

In the IP network shown in the figure, voice packets that enter the network at a constant rate can reach the intended destination by a number of routes. Because each of these routes may have different delay characteristics, the arrival rate of the packets may vary. This condition is called jitter.

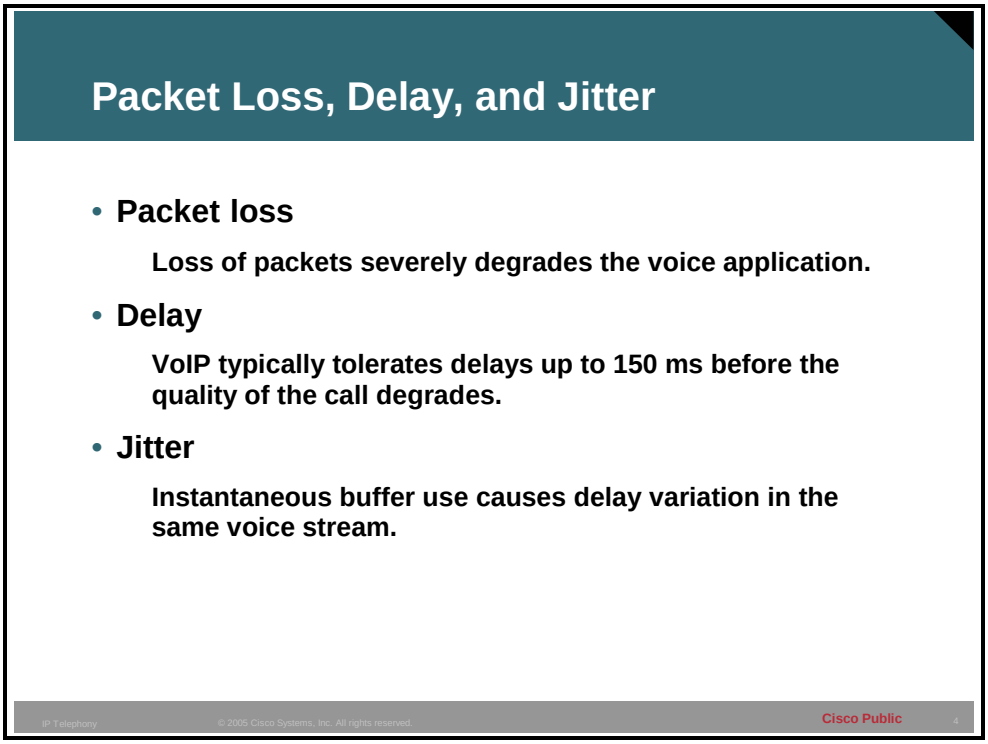
Another effect of multiple routes is that voice packets can arrive out of order. The far-end voice-enabled router or gateway has to re-sort the packets and adjust the interpacket interval for a proper-sounding voice playout.

Network transmission adds corruptive effects like noise, delay, echo, jitter, and packet loss to the speech signal. VoIP is susceptible to these network behaviors, which can degrade the voice application.

If a VoIP network is to provide the same quality that users have come to expect from traditional telephony services, then the network must ensure that the delay in transmitting a voice packet across the network, and the associated jitter, does not exceed specific thresholds.

Packet Loss, Delay, and Jitter

This topic discusses the causes of packet loss, end-to-end delay, and jitter delay in an IP internetwork.



In traditional telephony networks, voice has a guaranteed delay across the network by strict bandwidth association with each voice stream. Configuring voice in a data network environment requires network services with low delay, minimal jitter, and minimal packet loss. Over the long term, packet loss, delay, and jitter will all affect voice quality, as follows:

- **Packet loss:** The IP network may drop voice packets if the network quality is poor, if the network is congested, or if there is too much variable delay in the network. Codec algorithms can correct small amounts of loss, but too much loss can cause voice clipping and skips. The chief cause of packet loss is network congestion.
- **Delay:** End-to-end delay is the time that it takes the sending endpoint to send the packet to the receiving endpoint. End-to-end delay consists of the following two components:
  - **Fixed network delay:** You should examine fixed network delay during the initial design of the VoIP network. The International Telecommunication Union (ITU) standard G.114 states that a one-way delay budget of 150 ms is acceptable for high-quality voice. Research at Cisco Systems has shown that there is a negligible difference in voice quality scores using networks built with 200-ms delay budgets. Examples of fixed network delay include propagation delay of signals between the sending and receiving endpoints, voice encoding delay, and voice packetization time for various VoIP codecs.

- **Variable network delay:** Congested egress queues and serialization delays on network interfaces can cause variable packet delays. Serialization delay is a constant function of link speed and packet size. The larger the packet and the slower the link-clocking speed, the greater the serialization delay. Although this ratio is known, it can be considered variable because a larger data packet can enter the egress queue at any time before a voice packet. If the voice packet must wait for the data packet to serialize, the delay incurred by the voice packet is its own serialization delay, plus the serialization delay of the data packet in front of it.

- **Jitter:** Jitter is the variation between the expected arrival of a packet and when it is actually received. To compensate for these delay variations between voice packets in a conversation, VoIP endpoints use jitter buffers to turn the delay variations into a constant value so that voice can be played out smoothly. Buffers can fill instantaneously, however, because network congestion can be encountered at any time within a network. This instantaneous buffer use can lead to a difference in delay times between packets in the same voice stream.

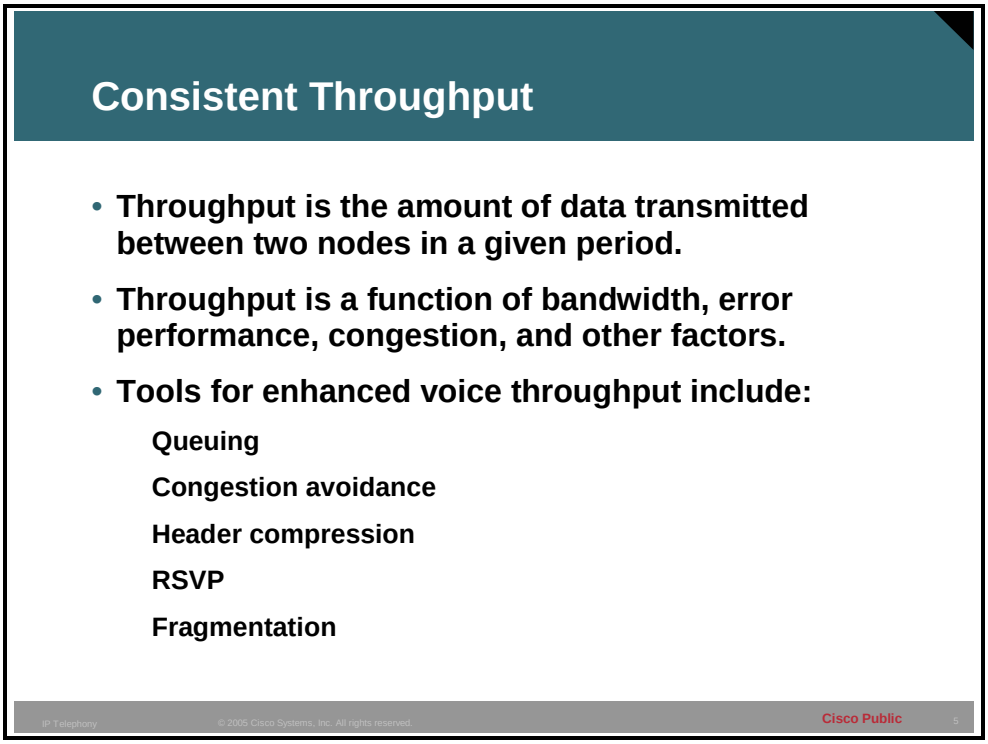
Example: Packet Loss, Delay, and Jitter Problems

The effect of end-to-end packet loss, delay, and jitter can be heard as follows:

- The calling party says, “Good morning, how are you?”
- With end-to-end delay, the called party hears, “……Good morning, how are you?”
- With jitter, the called party hears, “Good……morning, how……are you?”
- With packet loss, the called party hears, “Good m..ning, w are you?”

Consistent Throughput

This topic describes the methods that you can use to ensure consistent delivery and throughput of voice packets in an IP internetwork.



Throughput is the actual amount of useful data that is transmitted from a source to a destination. The amount of data that is placed in the pipe at the originating end is not necessarily the same amount of data that comes out at the destination. The data stream may be affected by error conditions in the network; for example, bits may be corrupted in transit, leaving the packet unusable. Packets may also be dropped during times of congestion, potentially forcing a retransmit, using twice the amount of bandwidth for that packet.

In the traditional telephony network, voice had guaranteed bandwidth associated with each voice stream. Cisco IOS software uses a number of techniques to reliably deliver real-time voice traffic across the modern data network. These techniques, which all work together to ensure consistent delivery and throughput of voice packets, include the following:

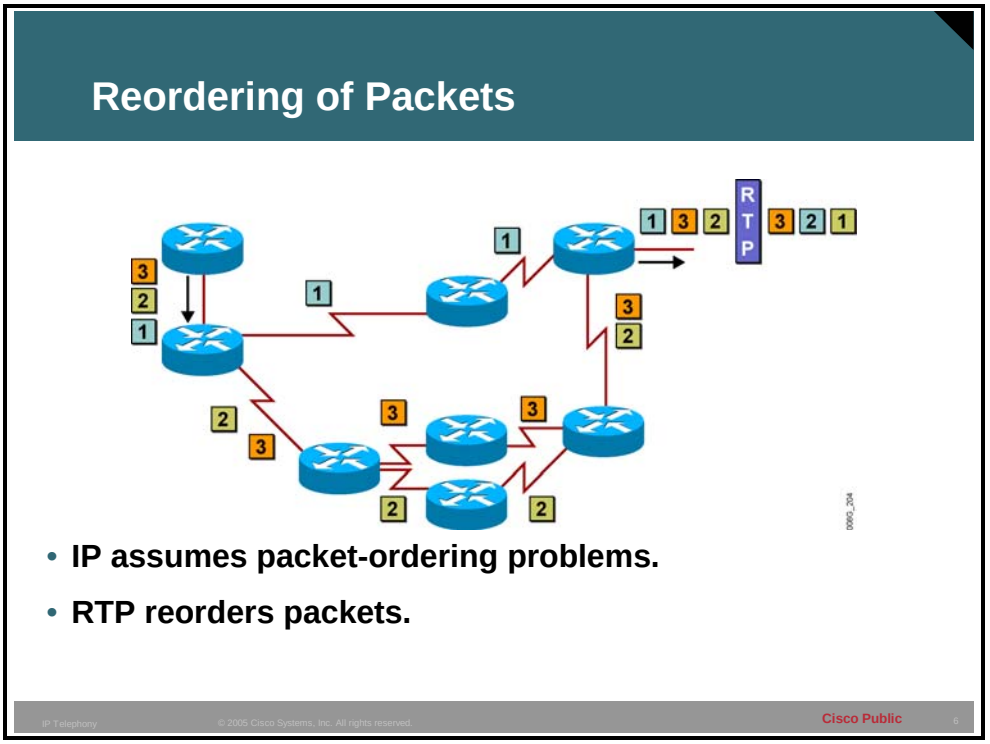
- **Queuing:** The act of holding packets so that they can be handled with a specific priority when leaving the router interface. Queuing enables routers and switches to handle bursts of traffic, measure network congestion, prioritize traffic, and allocate bandwidth. Cisco routers offer several different queuing mechanisms that can be implemented based on traffic requirements. Low Latency Queuing (LLQ) is one of the newest Cisco queuing mechanisms.
- **Congestion avoidance:** Congestion avoidance techniques monitor network traffic loads. The aim is to anticipate and avoid congestion at common network and internetwork bottlenecks before it becomes a problem. These techniques provide preferential treatment under congestion situations for premium (priority) class traffic, such as voice. At the same time, these techniques maximize network throughput and capacity use and minimize packet loss and delay. Weighted random early detection (WRED) is one of the QoS congestion avoidance mechanisms used in Cisco IOS software.

- **Header compression:** In the IP environment, voice is carried in Real-Time Transport Protocol (RTP), which is carried in User Datagram Protocol (UDP), which is then put inside an IP packet. This constitutes 40 bytes of RTP/UDP/IP header. This header size is large when compared to the typical voice payload of 20 bytes. Compressed RTP (CRTP) reduces the headers to 2 bytes in most cases, thus saving considerable bandwidth and providing for better throughput.
- **Resource Reservation Protocol:** Resource Reservation Protocol (RSVP) is a transport layer protocol that enables a network to provide differentiated levels of service to specific flows of data. Unlike routing protocols, RSVP is designed to manage flows of data rather than make decisions for each individual datagram. Data flows consist of discrete sessions between specific source and destination machines. Hosts use RSVP to request a QoS level from the network on behalf of an application data stream. Routers use RSVP to deliver QoS requests to other routers along the paths of the data stream. After an RSVP reservation is made, weighted fair queuing (WFQ) is the mechanism that actually delivers the queue space at each device. Voice calls in the IP environment can request RSVP service to provide guaranteed bandwidth for a voice call in a congested environment.
- **Fragmentation:** Fragmentation defines the maximum size for a data packet and is used in the voice environment to prevent excessive serialization delays. Serialization delay is the time that it takes to actually place the bits onto an interface; for example, a 1500-byte packet takes 187 ms to leave the router over a 64-kbps link. If a best-effort data packet of 1500 bytes is sent, real-time voice packets are queued until the large data packet is transmitted. This delay is unacceptable for voice traffic. However, if best-effort data packets are fragmented into smaller pieces, they can be interleaved with real-time (voice) packets. In this way, both voice and data packets can be carried together on low-speed links without causing excessive delay to the real-time voice traffic.

There are many QoS tools that can be used to ensure consistent throughput. When these mechanisms are employed, voice traffic on the network is assured priority and its delivery is more consistent.

Reordering of Voice Packets

This topic describes how RTP ensures consistent delivery order of voice packets in an IP internetwork.



In traditional telephony networks, voice samples are carried in an orderly manner through the use of time-division multiplexing (TDM). Because the path is circuit-switched, the path between the source and destination is reserved for the duration of the call. All of the voice samples stay in order as they are transmitted across the wire. Because IP provides connectionless transport with the possibility of multiple paths between sites, voice packets cannot arrive out of order at the destination. Because voice rides in UDP/IP packets, there is no automatic reordering of packets.

RTP provides end-to-end delivery services for data that require real-time support, such as interactive voice and video. According to RFC 1889, the services provided by RTP include payload-type identification, sequence numbering, time stamping, and delivery monitoring.

Example: Reordering Voice Packets

In the figure, RTP reorders the voice packets through the use of sequence numbers before playing them out to the user.

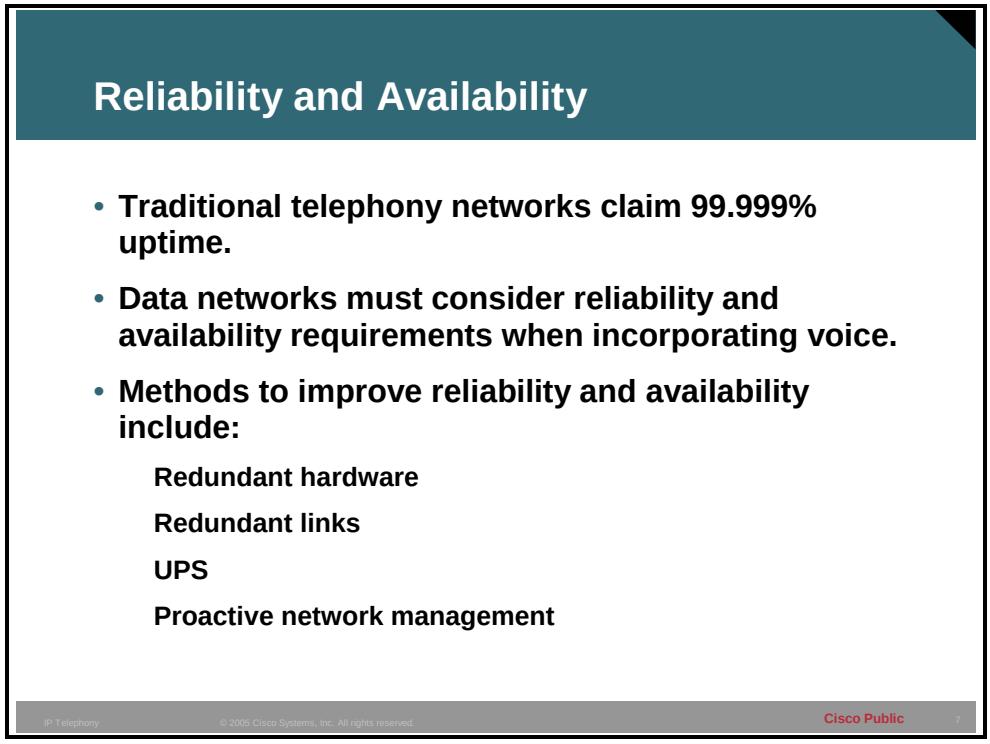
The table illustrates the various stages of packet reordering by RTP.

Sequencing of Packets by RTP

Stage	What Happens
Voice packets enter the network.	IP assumes packet-ordering problems.
RTP reorders the voice packets.	The voice packets are put in order through the use of sequence numbers.
RTP retimes the voice packets.	The voice packets are spaced according to the time stamp contained in each RTP header.
	The user hears the voice packets in order and with the same timing as when the voice stream left the source.
RTCP (Real-Time Transport Control Protocol) sends occasional report packet for delivery monitoring.	Both the sender and receiver send occasional report packets containing information, such as number of packets sent or received, the octet count, and the number of lost packets.

Reliability and Availability

The traditional telephony network strives to provide 99.999 percent uptime to the user. This corresponds to 5.25 minutes per year of downtime. Many data networks cannot make the same claim. This topic describes methods that you can use to improve reliability and availability in data networks.



To provide telephony users the same—or close to the same—level of service as they experience with traditional telephony, the reliability and availability of the data network takes on new importance.

Reliability is a measure of how resilient a network can be. Efforts to ensure reliability may include choosing hardware and software with a low mean time between failure, or installing redundant hardware and links. Availability is a measure of how accessible the network is to the users. When a user wants to make a call, for example, the network should be accessible to that user at any time a call is required. Efforts to ensure availability may include installing proactive network management to predict failures before they happen, and taking steps to correct problems in design of the network as it grows.

When the data network goes down, it may not come back up for minutes or even hours. This delay is unacceptable for telephony users. Local users with network equipment, such as voice-enabled routers, gateways, or switches for IP Phones, now find that their connectivity is terminated. Administrators must, therefore, provide an uninterruptible power supply (UPS) to these devices in addition to providing network availability. Previously, depending on the type of connection the user had, they received their power directly from the telephone company central office (CO) or through a UPS that was connected to their keyswitch or PBX in the event of a power outage. Now the network devices must have protected power to continue to function and provide power to the end devices.

Network reliability comes from incorporating redundancy into the network design. In traditional telephony, switches have multiple redundant connections to other switches. If either

a link or a switch becomes unavailable, the telephone company can route the call in different ways. This is why telephone companies can claim a high availability rate.

High availability encompasses many areas of the network. In a fully redundant network, the following components need to be duplicated:

- Servers and call managers
- Access layer devices, such as LAN switches
- Distribution layer devices, such as routers or multilayer switches
- Core layer devices, such as multilayer switches
- Interconnections, such as WAN links and public switched telephone network (PSTN) gateways, even through different providers
- Power supplies and UPSs

Example: Cisco Reliability and Availability

In some data networks, a high level of availability and reliability is not critical enough to warrant financing the hardware and links required to provide complete redundancy. If voice is layered onto the network, these requirements need to be revisited.

With Cisco Architecture for Voice, Video and Integrated Data (AVVID) technology, the use of Cisco CallManager clusters provides a way to design redundant hardware in the event of Cisco CallManager failure. When using gatekeepers, you can configure backup devices as secondary gatekeepers in case the primary gatekeeper fails. You must also revisit the network infrastructure. Redundant devices and Cisco IOS services, like Hot Standby Router Protocol (HSRP), can provide high availability. For proactive network monitoring and trouble reporting, a network management platform such as CiscoWorks2000 provides a high degree of responsiveness to network issues.

