Why Voice over IP?

• Traditional TDM (Time-division multiplexing)

– High recurring maintenance costs

– Monolithic switch design with proprietary interfaces

– Uses dedicated, voice-only bandwidth in HFC network

• IP

– Many services, one network

– Leverages existing data infrastructure

– Enhanced services

– Open standards

VoIP Transport

• **Packet Encapsulation**

– RTP – Real TimeTransport Protocol

• Provides timestamp and packet sequence numbering so destination applications can reassemble and playback speech

– UDP – User Datagram Protocol

• Provides port number addressing, so that the correct destination application can receive the packet

• Provides data integrity via CRC (Cyclic Redundancy Check)

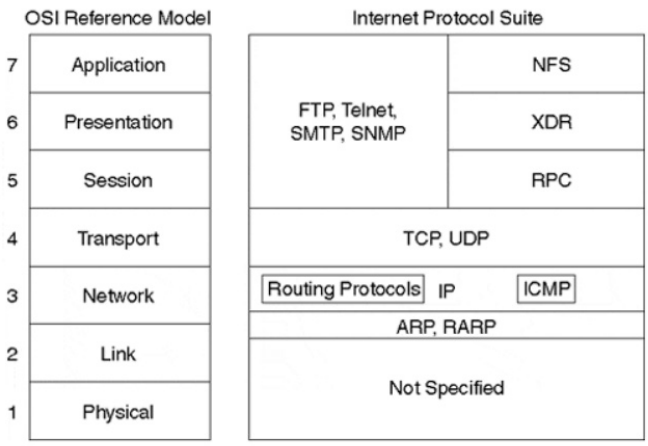
– IP – Internet Protocol

• Provides routing info for packets to get to appropriate destination

• Provides packet prioritization

Many of the benefits of Voice over IP (VoIP) are derived from the use of Internet Protocol (IP) as the transport mechanism. To truly understand these benefits, you must first under-stand what IP actually means. What are the behavioral characteristics of IP, and what does an IP packet look like?

Before you can understand what IP can do for you and ways you can run applications through IP, you must first become familiar with the Open Systems Interconnection (OSI) reference model and how it applies to IP.



The Application Layer

Most users are familiar withthe application layer. Some well-known applicationsinclude the following:

- E-mail
- Webbrowsing
- Word processing

The Presentation Layer

The presentation layer ensures that information sent by the application layer of one system is readable by the application layer of another system. If necessary, the presentation layer translates between multiple data formats by using a common data representation format.

The Session Layer

As its name implies, the session layer establishes, manages, and terminates sessions between applications.

Sessions consist of dialogue between two or more presentation entities (recall that the session layer provides its services to the presentation layer).

The Transport Layer

The transport layer is responsible for ensuring reliable data transport on an internetwork. This is accomplished through flow control, error checking (checksum), end-to-end acknowledgments, retransmissions, and data sequencing.

The Network Layer

The network layer provides for the logical addressing which enables two disparate systems on different logical networks to determine a possible path to communicate. The network layer is the layer in which routing protocols reside.

On the Internet today, IP addressing is by far the most common addressing scheme in use.

The Data Link Layer

The data link layer provides reliable transport across a physical link. The link layer has its own addressing scheme. This addressing scheme is concerned with physical connectivity and can transport frames based

upon the datalink layer address.

The Physical Layer

The physical layer is concerned with creating 1s and 0s on the physical medium with electrical impulses/voltage changes.

IP Transport Mechanisms

TCP and UserDatagram Protocol (UDP) have different characteristics that various applications can use. If reliability is more important than delay, for instance, you can use TCP/IP to guarantee packet delivery. UDP/IP does not utilize packet re-transmissions, however.

This can lower reliability, but in some cases a late retransmissionis of no use.

TCP

- TCP provides full-duplex, acknowledged and flow-controlled service to upper-layer protocols. It moves data in a continuous, unstructured byte stream where bytes are identified by sequence numbers.

- TCP can support numerous simultaneous upper-layer conversations. The port numbers in a TCP header identify an upper-layer conversation. Many well-known TCP ports are reserved for File Transfer Protocol (FTP), World Wide Web (WWW), Telnet, and so on.

- Within the signaling portion of VoIP, TCP is used to ensure the reliability of the setup of a call. Due to the methods by which TCP operates, it is not feasible to use TCP as the mechanism to carry the actual voice in a VoIP call. With VoIP, packet loss is less important than latency.

UDP

- UDP is a much simpler protocol than TCP and is useful in situations where the reliability mechanisms of TCP are unnecessary.

- UDP also is connectionless and has a smaller header, which translates to minimal overhead.

- The UDP header has only four fields: source port, destination port, length, and UDP checksum (optional).

- The source and destination port fields serve the same functions as they do in the TCP header. The length field specifies the length of the UDP header and data, and the checksum field enables packet integrity checking.

- UDP is used in VoIP to carry the actual voice traffic (the bearer channels). TCP is not used because flow control and retransmission of voice audio packets are unnecessary.

- Because UDP is used to carry the audio stream, it continues to transmit, regardless of whether you are experiencing 5 percent packet loss or 50 percent packet loss.

- If TCP were utilized for VoIP, the latency incurred waiting for acknowledgments and retransmissions would render voice quality unacceptable. With VoIP and other real-time applications, controlling latency is more

important than ensuring the reliable delivery of each packet.

- To create a proper network design, it is important to know all the caveats and inner workings of networking technology.

- Below we will explain many of the issues facing Voice over IP:

Delay/Latency

VoIP delay or latency is characterized as the amount of time it takes for speech to exit the speaker's mouth and reach the listener's ear.

Jitter

Jitter is defined as a variation in the delay of received packets.

Pulse Code Modulation

- Although analog communication is ideal for human communication, analog transmission is neither robust nor efficient at recovering from line noise. In the early telephony network, when analog transmission was passed through amplifiers to boost the signal,

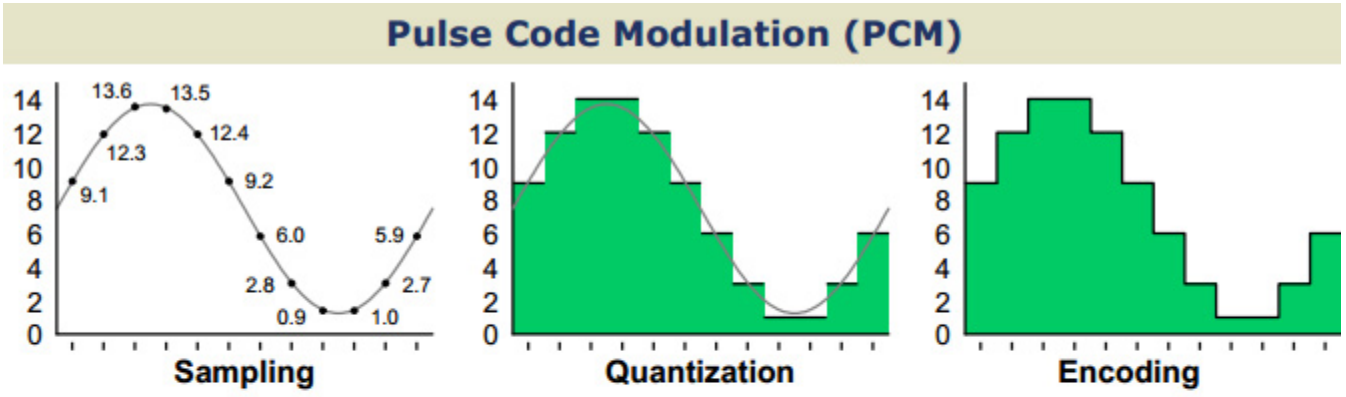
not only was the voice boosted but the line noise was amplified, as well.

- This line noise resulted in an often-unusable connection.

- It is much easier for digital samples, which are comprised of 1 and 0 bits, to be separated from line noise. Therefore, when analog signals are regenerated as digital samples, a clean sound is maintained. When the benefits of this digital representation became evident, the telephony network migrated to pulse code modulation (PCM).

- PCM convertsanalog sound into digitalform by sampling the analog sound 8000 times per second and converting each sample into a numeric code. The Nyquist theorem states that if you sample an analog signal at twice the rate of the highest frequency of interest, you can accurately reconstruct

that signal back into its analog form. Because most speech content is below 4000Hz (4 kHz), a sampling rate of 8000 times per second (125 ms between samples) is required.



Sampling
8000 discrete signal measurements are taken at equal intervals every second

Quantization
The level of each sample is rounded to the nearest expressible value

Encoding
Digital values are encoded as binary numbers for encapsulation

Compression (Optional)
The digital signal is compressed in real time to consume less bandwidth

Voice Coding Standards

The ITU-T standardizes CELP, MP-MLQ PCM, and ADPCM coding schemes inits G-series recommendations.

The most popular voice coding standards for telephony and packet voice include:

• G.711—Describes the 64 Kbps PCM voice coding technique outlined earlier; G.711-encoded voice is already in the correct format for digital voice delivery in the public phone network or through Private Branch eXchanges (PBXs).

• G.726—Describes ADPCM coding at 40, 32, 24, and 16Kbps; you also can interchange ADPCM voice between packet voice and public phone or PBX networks, provided that the latter has ADPCM capability.

• G.728—Describes a 16 Kbps low-delay variation of CELP voice compression.

• G.729—Describes CELP compression that enables voice to be coded into 8 Kbps streams; two variations of this standard (G.729 and G.729 Annex A) differ largely in computational complexity, and both generally provide speech quality as good as that of 32 Kbps ADPCM.

• G.723.1—Describes a compression technique that you can use to compress speech or othe raudio signal components of multimedia service at a low bit rate, as part of the overall H.324 family of standards. Two bit rates are associated with this coder: 5.3 and 6.3 Kbps. The higher bit rateis based

on MP-MLQ technology and provides greater quality. The lower bit rate is based on CELP, provides good quality, and affords system designers with additional flexibility.

Voice Codecs				
	MOS	**Bandwidth**	**Complexity**	**Free**
G.722 SB-ADPCM	4.13	48-64 kbps	Medium	Yes
G.711 PCM	4.1	64 kbps	Low	Yes
iLBC	4.1	15.2 kbps	High	Yes
G.729 CS-ACELP	3.92	8 kbps	High	No
G.726 ADPCM	3.85	32 kbps	Medium	Yes
G.729a CS-ACELP	3.7	8 kbps	Medium	No
G.728 LD-CELP	3.61	16 kbps	High	No

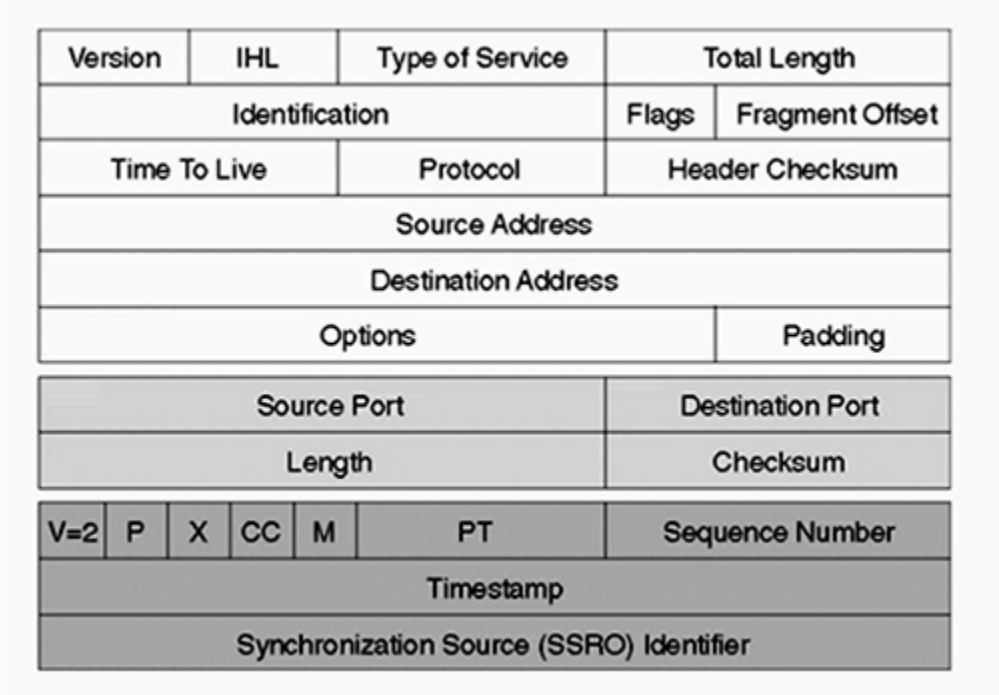
Transport Protocols

- 2 main types of traffic ride upon Internet Protocol (IP): User Datagram Protocol (UDP) and Transmission Control Protocol (TCP). In general, you use TCP when you need a reliable connection and UDP when you need simplicity and reliability is not your chief concern.

- Due to the time-sensitive nature of voice traffic, UDP/IP was the logical choice tocarry voice. More information was needed on a packet-by-packet basis than UDP offered,however. So, for real-time or delay-sensitive traffic, the Internet Engineering Task Force (IETF) adopted the RTP. VoIP rides on top of RTP, which rides on top of UDP. Therefore, VoIP is carried with an RTP/UDP/IP packet header.

RTP

- RTP is the standard for transmitting delay-sensitive traffic across packet-based networks. RTP rides on top of UDP and IP. RTP gives receiving stations information that is not in the connectionless UDP/IP streams. As shown in the below figure, two important bits of information are sequence information and timestamping. RTP uses the sequence information to determine whether the packets are arriving in order and it uses the time-stamping information to determine the interarrival packet time (jitter).



- You can use RTP for media on demand, as well as for interactive services such as Internet telephony. RTP consists of a data part and a control part, the latter called RTP Control Protocol (RTCP).

- The data part of RTP is a thin protocol that provides support for applications with real-time properties, such as continuous media (for example, audio and video), including timing reconstruction, loss detection, and content identification.

- RTCP provides support for real-time conferencing of groups of any size within an Internet. This support includes source identification and support for gateways, such as audio and videobridges as well as multicast-to-unicast translators. It also offers QoS feedback from receivers to the multicast group, as well as support for the synchronization of different media streams.

IP Signaling Protocols: Session Initiation Protocol (SIP)

- The Session Initiation Protocol (SIP) is an application-layer signaling-control protocol used to establish, maintain, and terminate multimedia sessions. Multimedia sessions include Internet telephony, conferences and other similar applications involving such media as audio, video and data.

- You can use SIP invitations to establish sessions and carry session descriptions.SIP supports unicast and multicast sessions as well as point-to-point and multipoint calls. You can establish and terminate communications using the following five SIP facets: user location, user capability, user availability, call setup, and call handling. SIP, on which Request For Comments (RFC) 2543 is based, is a text-based protocol that is part of the overall Internet Engineering Task Force (IETF) multimedia architecture.

- Internet Protocol (IP) telephony is still being developed and will require additional signaling capabilities in the future. The extensibility of SIP enables such development of incremental functionality. SIP message headers are versatile, and you can registeradditional features with the Internet Assigned

Numbers Authority (IANA). SIP messageflexibility also enables elements to construct advanced telephony services, including mobility type services.

User Agents

User agents are client end-system applications that contain both a user-agent client (UAC) and a user-agent server(UAS), otherwise known as client and server, respectively.

- Client—Initiates SIP requests and acts as the user's calling agent.

- Server—Receives requests and returns responses on behalf of the user; acts as the user-called agent.

Addressing

- SIP addresses, also called SIP Universal Resource Locators (URLs), exist in the form of users @ hosts.

- Similar to e-mail addresses, a SIP URL is identified by user@host. The user portion of the address can be a username or telephone number, and the host portion can be a domain name or network address. You can identify a user's SIP URL by his or her e-mail address. The following example depicts two possible SIP URLs:

sip:support@mizu-voip.com

sip:test1@191.191.191.19

Locating a Server

A client can send a SIP request either directly, to a locally configured proxy server, or to the IP address and port of the corresponding SIP URL. Sending a SIP request directly is relatively easy, as the end-system

application knows the proxy server. Sending a SIP request in the second manner is somewhat more complicated, for the following reasons:

- The client must determine the IP address and port number of the server for which the request is destined.

- If the port number isnot listed in the requested SIP URL, the default port is 5060.

- If the protocoltype is not listed in the request SIP URL, the client must first attempt to connect using User DatagramProtocol (UDP) and then TransmissionControl Protocol (TCP).

- The client queries the Domain Name System (DNS) serverfor thehost IP address. If it finds no address records, the client is unable to locate the server and cannotcontinuewiththe request.

SIP Transactions

- After addressing is resolved, the client sends one or more SIP requests and receives one or more responses from the specified server. All the requests and responses associated with this activity are considered part of a

SIP transaction. For simplicity and consistency, the header fields in all request messages match the header fields in all response messages.

You can transmit SIP transactions in either UDP or TCP.

SIP Messages

- Two kinds of SIP messages exist: requests initiated by clients, and responses returned from servers. Every message contains a header that describes the details of communication. SIP is a text-based protocol with message syntax and header fields identical to Hypertext Transfer Protocol (HTTP). SIP messages are sent over TCP or UDP with multiple messages carried in a single TCP connection or UDP datagram.

Message headers

- You use message headers to specify the calling party, called party, route, and message type of a call. The four groups of message headers are:

- General headers—Apply to requests and responses.
- Entity headers—Define information about the message body type and length.
- Request headers – Enable the client to include additional request information
- Response headers—Enable the server to include additional response information.

General Headers	Entity Headers	Request Headers	Response
Accept	Content-Encoding	Authorization	Allow
Accept-Encoding	Content-Length	Contact	Proxy-Authentic
Accept-Language	Content-Type	Hide	Retry-After
Call-ID		Max-Forwards	Server
Contact		Organization	Unsupported
CSeq		Priority	Warning
Date		Proxy-Authorization	WWW-Authenti
Encryption		Proxy-Require	
Expires		Route	
From		Require	
Record-Route		Response-Key	
Timestamp		Subject	
To		User-Agent	
Via			

To: identifies the recipient of the request

From: indicates the initiator of the request

Subject: describes the nature of the call

Via: indicates the path taken by the request

Call-ID: uniquely identifies a specific invitation or all registrations of a specific client

Content-length: identifies the size of message body in octets

Content-Type: indicates the media type of message body

Expires: indicates when the message content expires

Message Requests

- SIP communication features six kinds of message requests. These requests, also referred to as methods, enable user agents and network servers to locate, invite, and manage calls. The six SIP requests are as

follows:

- **INVITE -** This method indicates that the user or service is invited to participate in a session. It includes a session description and for two-way calls, the callingparty indicates the media type. A successful

response to a two-party INVITE (200 OK response) includes the called party's receive mediatype. With this simple method, users can recognize the capabilities of the other end and open a

conversation session with a limited number of messages and round trips.

- **ACK** - These requests correspond to an INVITE request. They represent the final confirmation from the end system and conclude the transaction initiated by the INVITE command. If the calling party

Includes a session description in the ACK request, no additional parameters are used in the session. If a session description is absent, the session parameters in the INVITE request are used as the default.

- **OPTIONS** - This method enables you to query and collect user agents and network server capabilities. This requestis not used to establish sessions, however.

- **BYE -** This method is used by calling and called parties to release a call. Before actually releasing the call, the user agent sends this request to the server indicating the desire to release the session.

- **CANCEL -** This request enables user agents and network servers to cancel anyin-progress request. This does not affect completed requests in which final responses were already received.

- **REGISTER** - This method is used by clients to register location information with SIP servers.

Message Responses

- SIP message responses are based upon the receipt and interpretation of a corresponding request. They are sent in response to requests and indicate call success or failure, including the status of the server. The six classes of responses, their status codes,

and explanations of what they do are provided in the below table:

Class of Response	Status Code	Explanation
Informational	100	Trying
	180	Ringing
	181	Call is being forwarded
	182	Queued
Success	200	OK
	300	Multiple choices
	301	Moved permanently
	302	Moved temporarily
	303	See other
	305	Use proxy
	380	Alternative service
Client-Error	400	Bad request
	401	Unauthorized
	402	Payment required
	403	Forbidden
	404	Not found
	405	Method not allowed
	406	Not acceptable
	407	Proxy authentication required
Client-Error	408	Request timeout
	409	Conflict
	410	Gone
	411	Length required
	413	Request entity too large
	414	Requested URL too large
	415	Unsupported media type
	420	Bad extension
	480	Temporarily not available
	481	Call leg or transaction doesn't exist
	482	Loop detected
	483	Too many hops
	484	Address incomplete
	485	Ambiguous
	486	Busy here
Server-Error	500	Internal server error
	501	Not implemented
	502	Bad gateway
	503	Service unavailable
	504	Gateway timeout
	505	SIP version not supported
Global Failure	600	Busy everywhere
	603	Decline
	604	Does not exist anywhere
	606	Not acceptable