

Voice over IP (VoIP) and Network Management

The Internet Protocol (IP) network was primarily designed for carrying best-effort (no assurance for delivery without any impairment) data traffic. Traditionally, voice communication was primarily carried over the public switched telephone network (PSTN), which is a global network. Because of the success of IP in becoming a worldwide standard and the ubiquity of IP networks, interconnectivity of IP networks provides global connectivity to its users. The global nature of IP networks provides alternative solutions to carrying media, including voice, as long as it can be carried as a payload in an IP packet.

The change from circuit-switched technology to packet-switched technology leverages some common technologies but also requires some unique methods for transporting voice packets. In the PSTN, the analog voice signal (human speech is analog voice) needs to be digitized using pulse code modulation (PCM), and the signaling happens transparently to the end user to provide dial tone, call routing, and other call features. The IP networks digitize voice signals at the telephone set by employing digital signal processors (DSP), and they put it in an IP packet. The IP packet containing the digitized voice traverses the IP network, leveraging its routing protocols using various available paths to reach the end destination. During the course of its journey, the voice packet shares links/paths with other packets carrying other data traffic, whereas in the PSTN, the circuit or path is established during the initial call setup and is dedicated for that voice call for its entire duration.

The PSTN has a dedicated call-signaling plane called Signaling System 7 (SS7). SS7 behaves like an IP routing and call-signaling protocol; however, in IP networks, call signaling shares common network resources with the data traffic.

The packet-switching mechanism in IP networks not only includes packetization of analog voice with the help of coders and decoders called CoDec but also encapsulation of this digitized voice data and voice signaling in packets as it is guided through the IP network based on routing protocols on the routers. This transport method is referred as

Voice over Internet Protocol (VoIP). VoIP introduces several new issues on the packet-switched network that a circuit-switched network does not have:

- Serialization delay
- Propagation delay with variance
- Packet loss during transmission
- Service degradation as the voice packets share a link on the LAN and Internet with data traffic

This chapter discusses some of these issues briefly.

VoIP Technology

VoIP technology was initially targeted for achieving toll savings called toll bypass because it uses the available bandwidth at the client's local-area network (LAN), wide-area network (WAN), and the Internet. VoIP technology turns analog voice into digital data packets that can be stored, searched, manipulated, copied, combined with other data, and distributed to virtually any device that connects to the IP network.

VoIP can interact seamlessly with other IP-based data systems because of the common standards defined by the International Telecommunications Union (ITU-T), the Internet Engineering Task Force (IETF), and the Institute of Electrical and Electronics Engineers (IEEE) that cover the VoIP call-signaling protocols, encryption, compression, encoding, and decoding techniques. This interconnectivity not only has made VoIP almost as ubiquitous as the time-division multiplexing (TDM)-based voice technology, but it now also facilitates enhanced collaboration experience by integrating with business process applications in manufacturing, health care, finance, government, education, and many other industry verticals. VoIP has provided opportunities to the service providers to offer new communication, enabling services to consumers and enterprises that contribute significantly to their revenue. The growing strategic importance of VoIP makes the management of the network extremely critical.

Network management's purpose is to maintain the quality and optimum delivery of services provided by the network. This is critical for service providers whose business is network-centric and enterprise or commercial segments because of business dependency on information technology. Network management has evolved to keep up with the technology advancements and convergence such as VoIP. This evolution is not limited to the technology involved in network management, including collection methods, configuration management, SLA management, trouble ticketing software, and so on. The advancements in network management include process improvement and best practices formulation into standards such as Telecommunications Management Network (TMN), Fault Configuration Accounting Performance Security (FCAPS), enhanced Telecom Operations Map (eTOM) framework, and Information Technology Infrastructure Library (ITIL). At a detailed level, we have seen continuous improvements and advancements in collection methods using technologies such as Simple Network Management Protocol versions

(SNMPv1, SNMPv2, SNMPv3), Remote network MONitoring (RMON), and eXtensible Markup Language (XML) that provide a central view of overall performance of the network by integrating the various network management tools and correlating data to provide intelligent and actionable reports.

This chapter lays the foundation for this book by explaining the fundamental workflow and process management concepts in network management and by providing an overview of TMN, FCAPS, and ITIL. We cover how these concepts apply to converged networks and then highlight the specific aspects that are fundamental to the topic of this book, that is, VoIP performance management and optimization based on analysis of key performance indicators or metrics. In the next section, we look at VoIP technology basics, its underlying protocols, common network problems in VoIP networks, and some voice quality-related problems in IP networks. We then cover the VoIP basics and network management concepts and discuss how they apply specifically to VoIP networks. We also discuss the strategic value of VoIP to service providers and enterprises and the importance of managing VoIP networks.

VoIP Overview

The PSTN is considered a *connection-oriented network*, in which a path from the source to the destination is established using TDM trunks between the intermediate central offices before the audio path is cut through, so end users can start their conversations over their phones. TDM is a *multiplexing scheme* in which two or more bit streams of digitized voice signal are transferred apparently and simultaneously as subchannels by taking turns on the common physical channel. The time domain is divided into several recurrent time slots of fixed length, one for each subchannel. This subchannel is referred to as Digital Signal 0 (DS0). DS0 has a bandwidth of 64,000 bits per second (bps), which is determined by the Nyquist theorem, which states that the minimum sampling rate of twice the frequency of the signal to be sampled will result in an accurate representation of the original signal. Because the human voice is limited to 4000 hertz, a sampling rate of 8000 samples per second, or every 125 microseconds, is used. The conversion process begins by analyzing each voice sample and converting it into an 8-bit word, also called an *octet*. If there are 8 bits per sample and 8000 samples per second, the product is 64,000 bits per second. Hence, the bandwidth of DS0 is 64 kbps.

The communication starts with the end user taking the telephone off the hook, which notifies the network that the service is requested. The network then returns a dial tone, and the end user dials the destination number. The call routing information is relayed by an overlay signaling network known as Signaling System 7 (SS7), which is a global network. All the central offices, including international gateways, connect to it. When the destination party answers, the end-to-end connection is confirmed through the various central offices along the path. When the conversation is complete, the two parties hang up, and their network resources can be reallocated for someone else's conversation. Also note that the voice signal can also come from data devices such as modems rather than just a phone.

Because of the connection-oriented nature of the PSTN, which holds the call path or “circuit” as constant for the duration of the call, the characteristics of that path, including serialization delay at different connection points, propagation delay, and information sequencing, remain constant for the duration of the call. Because these constants add to the reliability of the system, the term *reliable network* is often used to describe a connection-oriented environment. However, this reliability comes at a cost of a dedicated network meant for only one purpose, that is, to switch calls.

In contrast, traditional data networks, including intranets and the Internet, are considered *connectionless networks*, in which the full source and destination address are attached to a packet of information, and then that packet is passed through the network for delivery to the ultimate destination. As this packet traverses through the network routers and switches, they route or switch the packet based on the header information in the packet and the routing configuration and routing protocols on the Layer 3 devices in the IP network. The dynamic nature of IP routing might cause these packets for the same call to take different routes during the entire duration of the call. This increases the potential for packets arriving out of sequence and with varying delay with increased probability of drop. The network devices might have different processing power, capacity, and connections of varying speed and bandwidth, which make the serialization and propagation delay variable and potentially large.

There are some inherent problems with digitization of analog voice signals, such as noise and echo, which are introduced by impedance mismatch in the hybrid (2-wire to 4-wire conversion) coder/decoder (codec) circuits. Because of the additional delay in IP networks, these problems get exacerbated and are more noticeable. For these reasons, the terms *best effort* and *unreliable* are often used to describe a connectionless environment. These problems can be mitigated through proper planning and employing special techniques to guarantee quality of service (QoS) for media traffic and proactive network monitoring to continue to optimize the network characteristics for better end-user experience.

Note VoIP is also referred to as IP telephony, or IPT. Both terms are used to refer to sending voice packets across an IP network. The distinction is based on the endpoints used in the communication. In a VoIP network, the TDM-based PSTN network interconnects with the IP network, usually through a voice gateway—enabling communication between an IP endpoint and another traditional endpoint in the PSTN. In an IP telephony environment, the communication typically takes place between two IP endpoints.

Unified Communications (UC) is a framework of hardware and software products that facilitate multiple communication means such as voice video, presence, instant messaging, and other collaboration technologies over an IP network that might integrate with other external networks. UC also encompasses management software for monitoring and configuring the devices in the network. The term *UC* is more commonly used in the enterprise context.

This book uses VoIP and IP telephony interchangeably. UC will be used in an enterprise context, where services other than just voice, such as voicemail and presence, are involved.

Before we elaborate on some of the common problems (and their root causes) encountered by service providers and network architects when deploying voice over an IP infrastructure, we briefly cover some protocol basics.

Media Transport Protocol for VoIP—RTP

Using Transmission Control Protocol (TCP) for transport on top of IP provides enhanced reliability (albeit with additional protocol overhead) as compared with User Datagram Protocol (UDP), although it is still not a true equivalent of a connection-oriented service. But a connection-oriented transport infrastructure such as PSTN is not absolutely necessary to support interactive communication. This is why Real Time Protocol (RTP) was introduced to carry real-time, interactive media traffic with greater efficiency and reliability on a fundamentally unreliable protocol (IP).

RTP runs on top of UDP to avoid the overhead associated with the otherwise more reliable TCP. RTP is currently the cornerstone for carrying real-time traffic across IP networks. To date, all VoIP signaling protocols utilize RTP/UDP/IP as their transport mechanism for voice traffic. Often, RTP packet flows are known as *RTP streams* or *media streams*. Therefore, you can use IP in conjunction with UDP and RTP to replace a traditional voice circuit.

Figure 1-1 illustrates the different fields within the RTP header.

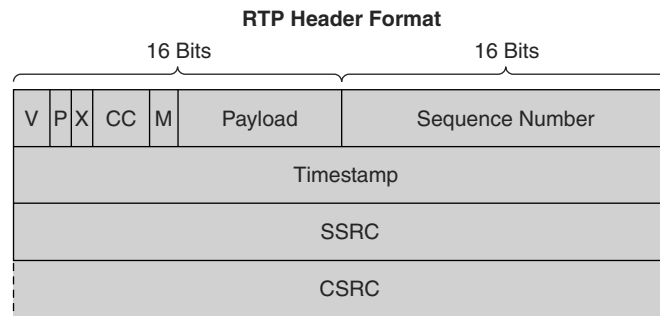


Figure 1-1 Real-time Transport Protocol Header Fields

The RTP header fields are described in the following list:

- **Version (V):** This field is 2 bits long and identifies the version of RTP. Most of the current applications use version 2. RFC 3550 covering RTP also defines version 2.
- **Padding (P):** This field is 1 bit long. If the padding bit is set, the packet contains one or more additional padding octets at the end that are not part of the payload.
- **Extension (X):** This field is 1 bit long. If the extension bit is set, the fixed header *must* be followed by exactly one header extension.

- **CSRC Count (CC):** This field is 4 bits long. The Contributing SouRCe (CSRC) count contains the number of CSRC identifiers that follow the fixed header.
- **Marker (M):** This field is 1 bit long. The interpretation of the marker is defined by a profile.
- **Payload Type (PT):** This field is 7 bits long and identifies the format of the RTP payload and determines its interpretation by the application.
- **Sequence Number:** This field is 16 bits long. The sequence number increments by one for each RTP data packet sent and can be used by the receiver to detect packet loss and to restore packet sequence. The initial value of the sequence number should be random (unpredictable).
- **Timestamp:** This field is 32 bits long. The timestamp reflects the sampling instant of the first octet in the RTP data packet. The initial value of the timestamp should be random. Several consecutive RTP packets have equal timestamps if they are (logically) generated at once, for example, if they belong to the same video frame. Consecutive RTP packets can contain timestamps that are not monotonic if the data is not transmitted in the order it was sampled, as in the case of MPEG-interpolated video frames.
- **SSRC:** This field is 32 bits long. The Synchronization SouRCe (SSRC) identifier field identifies the synchronization source. This identifier is chosen randomly, with the intent that no two synchronization sources within the same RTP session have the same SSRC identifier.
- **CSRC:** This field is 32 bits long and contains 0 to 15 items, 32 bits each. The CSRC list identifies the contributing sources for the payload contained in this packet. The number of identifiers is given by the CC field. If there are more than 15 contributing sources, only 15 can be identified. CSRC identifiers are inserted by mixers using the SSRC identifiers of contributing sources.

The IP packet containing the voice payload has an IP packet header that is 40 bytes (IP = 20 bytes, UDP = 8 bytes, and RTP = 12 bytes). The IP packet containing the voice payload varies in size based on the codec type, its bit rate, and the codec packetization period, which is also known as the *codec sample interval*. For example, a G.711 codec with a bit rate of 64 kbps and a sample interval of 20 milliseconds (ms) would yield a voice payload of 160 bytes. This is calculated as follows:

$$\begin{aligned}
 \text{Voice payload size} &= \text{Codec bit rate} * \text{Codec sample interval} \\
 &= (64 \text{ kbps for G.711}) * (20 \text{ ms}) \\
 &= (8000 \text{ bytes per second}) * (.02 \text{ seconds}) \\
 &= 160 \text{ bytes}
 \end{aligned}$$

The total length of the voice packet is 200 bytes (40 bytes IP header + 160 bytes voice payload), and this does not include the Layer 2 header overhead. Figure 1-2 illustrates the RTP encapsulation format.

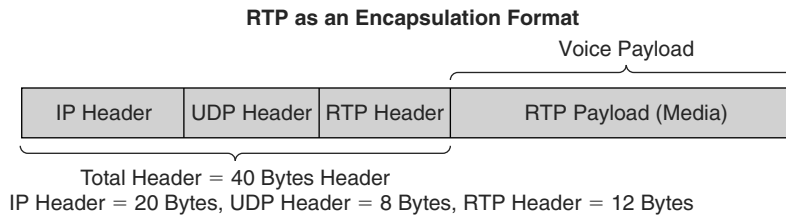


Figure 1-2 *Real-time Transport Protocol Encapsulation Format*

Similarly, for a G.729 codec with a bit rate of 8 kbps and a sample interval of 20 ms, the voice payload size comes out to 20 bytes. The total length of the voice packet in this case comes out to 60 bytes (40 bytes header + 20 bytes voice payload). Because voice packets are relatively small, a loss of one to two packets does not severely impact the quality of the voice conversation, and in most cases, the user might not experience a noticeable difference in the quality of the voice call.

In the case of packet loss, the receiving station waits for a period of time (per its jitter buffer) and then runs a *concealment strategy*. This concealment strategy replays the last packet received, so the listener does not hear gaps of silence. Because the lost speech is only typically 10 or 20 ms, the listener most likely does not hear the difference. You can accomplish this concealment strategy only if one packet is lost. If multiple consecutive packets are lost, the concealment strategy is run only once until another packet is received. In IP networks, it is common and normal for packet loss to occur. In fact, TCP/IP was built to utilize packet loss as a means of controlling the flow of packets. In TCP/IP, if a packet is lost, it is retransmitted. In most real-time applications, retransmission of a packet is worse than not receiving a packet because of the time-sensitive nature of the information. That is why real-time applications do not use TCP and instead use UDP.

The ITU-T recommends a one-way delay of no more than 150 ms. In a Cisco VoIP network, the unidirectional delay might be 120 ms (currently, 65 to 85 ms of that 120-ms delay is derived from two Cisco VoIP gateways when using G.729). If the receiving station must request that a packet be retransmitted, the delay is too large and large gaps and breaks in the conversation occur.

RTP has a timestamp field that records the exact time the packet was sent (in relation to the entire RTP stream). This information is used by the device terminating/receiving the audio flow. The receiving device uses the RTP timestamps to determine when a packet was expected, whether the packet was in order, and whether it was received when expected. All this information helps the receiving station determine how to tune its own settings to mask any potential issues such as delay, jitter, and packet loss, which are the result of the inherent nature of an IP network. These network-related problems are discussed in more detail later in this chapter.

VoIP Signaling Protocols

Some of the commonly used VoIP signaling protocols discussed in this book include H.323, the Media Gateway Control Protocol (MGCP), and the Session Initiation Protocol (SIP). Although other VoIP signaling protocols exist, such as H.248/Megaco, they are not covered in this book because the deployment models discussed are primarily based on H.323, MGCP, and SIP:

- H.323 is an ITU-T specification for transmitting audio, video, and data across an IP network, including the Internet. When compliant with H.323, vendors' products and applications can communicate and interoperate with each other. The H.323 standard addresses call signaling and control, multimedia transport and control, and bandwidth control for point-to-point and multipoint conferences. The H series of recommendations also specifies H.225 for connection establishment and termination between endpoints, H.245 for multimedia communications, H.320 for Integrated Services Digital Network (ISDN), and H.324 for plain old telephone service (POTS) as transport mechanisms. H.323-based deployment models are covered in detail in Chapter 5, "VoIP Deployment Models in Enterprise Networks."
- Media Gateway Control Protocol (MGCP) is defined in RFC 3435. MGCP is a protocol used by media gateway controllers (MGC), also known as call agents, to control media gateways (MG). MGCP is based on a master/slave relationship in which MGC is the master that issues commands to the MG (slave). The MG acknowledges the command, executes it, and notifies the MGC of the outcome (successful or not). In this architecture, the MG handles the media functions, such as conversion of TDM/analog signals into Real-time Transport Protocol (RTP)/Real-time Transport Control Protocol (RTCP) streams. MGC handles the call-signaling functions. MGCP-based call control is also referred to as a centralized switching deployment model. This is discussed in more detail in Chapter 3, "VoIP Deployment Models in Service Provider Networks."
- Session Initiation Protocol (SIP) is defined in RFC 3261. SIP is a signaling protocol that controls the initiation, modification, and termination of interactive multimedia sessions. The multimedia sessions can be as diverse as audio or video calls among two or more parties, chat sessions, or game sessions. SIP extensions have also been defined for instant messaging, presence, and event notifications. SIP is a text-based protocol that is similar to HTTP, Simple Mail Transfer Protocol (SMTP), and SDP.

SIP is a peer-to-peer protocol, which means that network capabilities such as call routing and session management functions are distributed across all the nodes (including endpoints and network servers) within the SIP network. This is in contrast to the traditional telephony model, where the phones or end-user devices are completely dependent on centralized switches in the network for call session establishment and services. SIP-based deployment models are covered in detail in Chapter 4, "Internet Telephony."

Common Network Problems in VoIP Networks

Deploying a VoIP infrastructure introduces a new set of challenges that do not exist in circuit-switched networks like the PSTN. Some of the common network problems encountered by providers deploying VoIP infrastructure include the following:

- Delay/latency
- Jitter
- Packet loss
- Voice Activity Detection (VAD)
- Other issues

These issues can affect the quality of voice service and result in a poor user experience when voice packets are transported over an IP infrastructure. Figure 1-3 categorizes the type of voice quality issues experienced by the users and shows their associated root causes.

Voice Quality Issues and Their Associated Root Causes

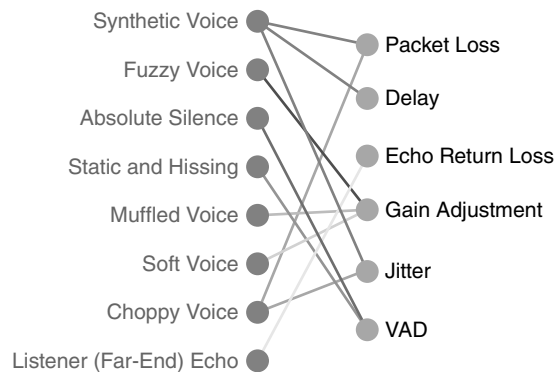


Figure 1-3 *Voice Quality Issues and Their Associated Root Causes*

Before we continue, it is important to understand the meaning and correlation of the previously mentioned issues and their impact on voice quality.

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. Three types of delay are inherent in today's packet-based voice networks:

- Propagation delay is caused by the length a signal must travel through light in fiber or electrical impulse in copper-based networks.
- Processing delay defines many different causes of delay (actual packetization, compression, and packet switching) and is caused by devices that forward the frame through the network.
- Serialization delay, also called queuing delay, is the amount of time it takes to actually place the voice packet onto an interface.

Propagation Delay

Light travels through a vacuum at a speed of 186,000 miles per second, and electrons travel through copper or fiber at approximately 125,000 miles per second. A fiber network stretching halfway around the world (13,000 miles) induces a one-way delay of about 70 ms. Although this delay is almost imperceptible to the human ear, propagation delays in conjunction with processing delays can cause noticeable speech degradation.

Processing Delay

As mentioned previously, devices that forward the frame through the network cause a processing delay. Processing delays can impact traditional phone networks and packet-based networks. This section discusses the different processing delays and describes how they affect voice quality.

The time taken by the DSP to compress a block of PCM samples is called *compression* or *coder delay*. The compression time for a Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP) process ranges from 2.5 ms to 10 ms based on the loading of the DSP processor. If the DSP is fully loaded with four voice channels, the coder delay is 10 ms. If the DSP is loaded with only one voice channel, the coder delay is 2.5 ms.

Packetization delay is the time taken to fill a packet payload with encoded/compressed speech. This delay is a function of the sample block size required by the coder and the number of blocks placed in a single frame. Packetization delay can also be called *accumulation delay*, as the voice samples accumulate in a buffer before they are released.

In the Cisco IOS VoIP product, the digital signal processor (DSP) can generate a speech sample every 10 ms when using G.729. Two of these speech samples (both with 10-ms speech data which causes a delay) are then placed within one packet. The packet delay is, therefore, 20 ms. An initial look-ahead of 5 ms occurs when using G.729, which gives an initial delay of 25 ms for the first speech frame. The compression algorithm relies on known voice characteristics to correctly process sample block N. The algorithm must have some knowledge of what is in block N+1 to accurately reproduce sample block N. This look-ahead is also known as the *algorithmic delay*.

Vendors can decide how many speech samples they want to send in one packet. In the previous example, G.729 uses 10-ms speech samples; each increase in samples per frame raises the delay by 10 ms. In fact, Cisco IOS enables users to choose how many samples to put into each frame.

Cisco gave DSP much of the responsibility for framing and forming packets to keep router/gateway overhead low. The RTP header, for example, is placed on the frame in the DSP instead of giving the router that task. So the primary task of the router is to forward the voice packets as fast as possible to reduce the switching or forwarding delay.

Serialization/Queuing Delay

A packet-based network experiences delay for several reasons. Two of these are the time necessary to move the actual packet to the output queue and queuing delay.

When packets are held in a queue because of congestion on an outbound interface, the result is queuing delay. Queuing delay occurs when more packets are sent out than the interface can handle at a given interval.

As mentioned previously, the ITU-T G.114 recommendation specifies that for good voice quality, no more than 150 ms of one-way, end-to-end delay should occur. With today's Cisco VoIP implementation, two routers with minimal network delay (back to back) introduce or require only about 60 ms of end-to-end delay. This leaves up to 90 ms of network delay to move the IP packet from source to destination.

Jitter

Jitter is the variation of packet interarrival time. Jitter is one issue that exists only in packet-based networks. While in a packet voice environment, the sender is expected to reliably transmit voice packets at a regular interval (for example, send one frame every 20 ms). These voice packets can be delayed throughout the packet network and not arrive at that same regular interval at the receiving station (for example, they might not be received every 20 ms). The difference between when the packet is expected and when it is actually received is jitter.

To compensate for jitter and conceal interarrival packet delay variation, VoIP endpoints are equipped with jitter buffers. Voice packets in IP networks have highly variable packet interarrival intervals. Recommended practice is to count the number of packets that arrive late and create a ratio of these packets to the number of packets that are successfully processed. You can then use this ratio to adjust the jitter buffer to target a predetermined, allowable late-packet ratio. This adaptation of jitter buffer sizing is effective in compensating for delays.

Note that jitter and total delay are not the same thing, although having plenty of jitter in a packet network can increase the amount of total delay in the network. This is because the more jitter you have, the larger your jitter buffer needs to be to compensate for the unpredictable nature of the packet network. Because more packets are accumulated in the jitter (or dejitter) buffer before play-out, it contributes to the overall delay budget because of larger accumulation time.

Most DSPs do not have infinite jitter buffers to handle excessive network delays. When network delay is predictable and has fewer variations, it is better to just drop packets or

have fixed-length buffers instead of creating unwanted delays in the jitter buffers. As long as the packet drop because of jitter buffer overflow is minimal, it might not significantly affect the voice quality compared to excessive delay. If your data network is engineered well and you take the proper precautions, jitter is usually not a major problem, and the jitter buffer does not significantly contribute to the total end-to-end delay. Although in situations where network delay is not constant, especially where network path traverses the Internet, adaptive or variable-length jitter buffers will be beneficial.

RTP timestamps are used within Cisco IOS Software to determine what level of jitter, if any, exists within the network.

The jitter buffer found within Cisco IOS Software is considered a dynamic queue. As voice frames arrive too quickly, the queue fills up. Similarly, when voice frames arrive too slowly, the queue empties out. This way the voice frame play-out rate is constant.

Although many vendors choose to use static jitter buffers, Cisco found that a well-engineered dynamic jitter buffer is the best mechanism to use for packet-based voice networks. Static jitter buffers force the jitter buffer to be either too large or too small, thereby causing the audio quality to suffer, because of either lost packets or excessive delay. The Cisco jitter buffer dynamically increases or decreases based upon the interarrival delay variation of the last few packets. More details on configuring jitter buffer can be found in *Troubleshooting Cisco IP Telephony*, by Paul Giral, Addis Hallmark, and Anne Smith.

Packet Loss

Packet loss in data networks is both common and expected. Many data protocols use packet loss to detect the condition of the network and can reduce the number of packets they are sending. Data can tolerate a packet being retransmitted. But retransmitting a packet containing voice traffic (RTP) is not an option because the packets might arrive at the destination out of order and with excessive delay. Voice traffic can tolerate a small amount of packet loss as long as the gap in voice is not perceivable to human ear, although excessive packet drop will cause a noticeable gap in voice, making it sound choppy. Therefore, when putting voice on data networks, it is important to use a mechanism like quality of service (QoS) to make the voice traffic somewhat resistant to periodic packet loss by prioritizing it over data traffic, although QoS might not completely prevent packet loss.

Cisco has developed many QoS tools that enable administrators to classify and manage traffic through a data network. If a data network is well engineered, you can keep packet loss to a minimum.

The Cisco VoIP implementation enables the voice-configured router to respond to periodic packet loss. If a voice packet is not received when expected (the expected time is variable), it is assumed to be lost and the last packet received is replayed. If the packet lost is only 20 ms of speech, the average listener does not notice the difference in voice quality.

Voice Activity Detection (VAD)

In normal voice conversations, someone speaks and someone else listens. Today's toll networks contain a bidirectional, 64,000-bps (bits per second) channel, regardless of whether anyone is speaking. This means that in a normal conversation, at least 50 percent of the total bandwidth is wasted. The amount of wasted bandwidth can actually be much higher if you take a statistical sampling of the breaks and pauses in a person's normal speech patterns.

When using VoIP, you can utilize this “wasted” bandwidth for other purposes when VAD is enabled. VAD works by detecting the magnitude of speech in decibels (dB) and deciding when to cut off the voice from being transmitted.

Typically, when the VAD detects a drop-off of speech amplitude, it waits a fixed amount of time before it stops putting speech frames in packets. This fixed amount of time is known as *hangover* and is typically 200 ms.

With any technology, trade-offs are made. VAD experiences certain inherent problems in determining when speech ends and begins and in distinguishing speech from background noise. This means that if you are in a noisy room, VAD is unable to distinguish between speech and background noise. This is also known as the *signal-to-noise threshold*. In these scenarios, VAD disables itself at the beginning of the call.

Another inherent problem with VAD is detecting when speech begins. Typically, the beginning of a sentence is cut off or clipped. This phenomenon is known as *front-end speech clipping*. Usually, the person listening to the speech does not notice front-end speech clipping.

Other Issues

Besides the common problems mentioned previously, other issues can impact voice quality. These issues can include the following:

- **Physical layer impairments:** Noise, interference in the line, loose connectors, badly terminated punch-down block, and so on
- **Last-mile connection bandwidth:** Low-speed connections, oversubscription of circuits resulting in congestion, and so on
- **Network resource overutilization:** High CPU and memory utilization on network devices, oversubscription of IP links resulting in congestion, high number of input/output drops under interfaces, lack of QoS for voice, and so on
- **VoIP application issues:** Poor software implementation on PC-based soft clients, lack of QoS and prioritization of resources for voice, and so on

The impact of these issues on voice quality and their fixes are discussed in detail in Chapter 6, “Managing VoIP Networks,” and Chapter 7, “Performance Analysis and Fault Isolation.”

However, we now look at some of the common voice quality problems that occur as a result of these issues.

Common Voice Quality Problems in IP Networks

The network problems listed in the previous section can seriously impact the quality of voice in an IP network. Some of the common voice quality issues experienced by providers include the following:

- **Noise:** This is typically any noise on the line introduced by an analog source in addition to the voice signal. Noise will typically leave the conversation intelligible but still far from excellent. Static, hum, crosstalk, and intermittent popping tones are examples where the calling and called parties can understand each other, but with some effort. Some noises are so severe that the voice becomes unintelligible.
- **Voice distortion:** This is typically any problem that affects the voice (RTP/media stream) itself. This category is further divided as follows:
 - **Echoed voice:** Echo voice is where the voice signal is repeated on the line. It can be heard at either end of the call, in varying degrees and with many combinations of delay and loss within the echoed signal.
 - **Garbled voice:** A garbled voice signal is one where the actual character of the voice is altered to a significant degree and often has a fluctuating quality. On some occasions, the voice becomes unintelligible.
 - **Volume distortion:** Volume distortion problems are associated with incorrect volume levels, whether constant or in flux.

Table 1-1 summarizes the different types of voice quality issues, their impact, and root causes.

In a nutshell, the IP network should be architected and configured in a way to transmit voice traffic the fastest way possible as a steady, smooth stream. Delay should be kept under 150 ms in one direction, and average jitter should be below 30 ms. Delay sources should be reduced, and dejitter buffers should be used carefully because compensating for jitter itself can create additional delay. Any out-of-order packets should be dropped because voice does not tolerate delay associated with packet retransmitting.

Quality of service best practices, traffic engineering for capacity planning, and proper network readiness assessment that takes into consideration delay budget planning in the predeployment phases for VoIP address these requirements. Echo- and noise-related issues can be compensated for by proper network wide-gain adjustment and use of echo cancellers. However, this must be done proactively in the design or pilot stages because changing the gain levels at the network boundary, where the IP and TDM network interface, can potentially affect the install base that did not originally have issues. Network transmission loss plan (NTLP) is a key step in VoIP implementation project. The concepts of network readiness assessment, traffic engineering for voice, delay budget planning, and NTLP are discussed in detail in Chapter 6.

Table 1-1 *Voice Quality Issues and Their Causes*

Voice Quality Issue	Symptom	Root Cause
<i>Noise</i>		
Absolute silence	This type of silence between speech can be understood if you have ever had the experience of not knowing whether the other person is still there because there is no sound on the line.	A common cause for this problem is VAD without comfort noise. To experience this symptom, the background noise is usually loud enough for the silence insertion to be noticeable but soft enough so that VAD will be engaged.
Clicking	Clicking is an external sound similar to a knock that is usually inserted at intervals.	A common cause is clock slips or other digital errors on the line.
Crackling	Crackling is an irregular form of very light static, similar to the sound a fire makes.	A common cause is poor electrical connections, in particular poor cable connections. Other causes are electrical interference and a defective power supply on the phone.
Crosstalk	Crosstalk is a familiar concept where you can hear someone else's conversation on the line. Commonly the other parties cannot hear you. There are also forms of crosstalk where all parties can hear each other.	Wires in close proximity, where the signal of one is induced into the other, is a common cause of this problem.
Hissing	Hissing is more driven and constant than static. White noise is a term often associated with strong hissing. Pink noise is a less constant hissing noise, and brown noise is even less constant.	A common cause of hissing is VAD. When VAD kicks in, comfort noise packets are introduced into the audio stream. The hissing sound is caused by the introduction of comfort noise into the conversation.
Static	Severe static is an example of static that in addition to creating background noise, affects the dial and ring tones and the voice itself. Another name for this symptom might be scratchy or gravel voice.	A common cause is A-law/Mu-law codec mismatch. A-law is a codec companding scheme used outside of the United States, whereas Mu-law is a U.S.-specific codec companding scheme. This is typically involved in international calls originating or terminating in the United States.

continues

Table 1-1 *Voice Quality Issues and Their Causes (continued)*

Voice Quality Issue	Symptom	Root Cause
<i>Echoed Voice</i>		
Listener echo	Listener and talker echo sound similar, although the signal strength of listener echo might be lower. The essential difference between them is who hears the echo and where it is produced. Listener echo is the component of the talker echo that leaks through the near-end hybrid and returns again to the listener, causing a delayed softer echo. The listener hears the talker twice.	Common causes are Insufficient loss of the echo signal. The reduction in the echo level produced by the tail circuit without the use of an echo canceler is referred to as Echo Return Loss (ERL). So if a speech signal enters the tail circuit from the network at a level of X dB, the echo coming back from the tail circuit into the terminal of the echo canceller is (X – ERL). Long echo tail. Echo cancellers in the gateway adjacent to the near-end hybrid circuit not activating.
Talker echo	Talker echo is the signal that leaks in the far-end hybrid and returns to the sender (talker). The talker hears an echo of his own voice.	Common causes are Insufficient loss of the echo signal. Echo cancellers in the gateway adjacent to the far-end hybrid not activating. Acoustic echo caused by the listener's phone.
Tunnel voice	Tunnel voice sounds similar to talking in a tunnel or on a poor-quality mobile phone car kit.	A common cause is tight echo with some loss. For example, 10-ms delay and 50 percent loss on the echo signal.
<i>Garbled Voice</i>		
Choppy voice	Choppy voice describes the sound when there are gaps in the voice. Syllables appear to be dropped or badly delayed in a start-and-stop fashion. Note: Other terms used to describe this sound are <i>clipped voice</i> and <i>broken voice</i> .	Common causes are consecutive packets being lost or excessively delayed such that DSP predictive insertion cannot be used and silence is inserted instead, for example, delay inserted into a call through contention caused by a large data packet.

Table 1-1 *Voice Quality Issues and Their Causes*

Voice Quality Issue	Symptom	Root Cause
Synthetic (robotic) voice	The term <i>synthetic</i> means that the sound of the voice is artificial and with a quiver. Predictive insertion causes this synthetic sound by replacing the sound lost when a packet is dropped with a best guess from a previous sample. Synthetic voice and choppy voice commonly occur together.	A common cause is single packet loss or delay beyond the bounds of the dejitter buffer playout period. DSP predictive insertion causes the synthetic quality of the voice, for example, when a call was provided insufficient bandwidth (such as a G711 codec across 64 kbps).
Underwater voice	Unintelligible underwater voice describes a distortion that makes it impossible to understand the voice. Descriptions of this sound include the sound of a cassette tape being fast forwarded, a gulping sound, and a wishy-washy sound.	A common cause of this problem is a G729 IETF and pre-IETF codec mismatch.
<i>Volume Distortion</i>		
Fuzzy voice	Fuzzy voice sounds similar to the radio being turned up too loud and the voice is shaky. This can only occur at certain signal levels within the sentence depending on the level of gain applied.	This is often caused by too much gain on the signal, possibly introduced at one of a number of points in the network. It can also apply to IP phones when used in noisy environments or when the volume is set to the high end.
Muffled voice	Muffled voice sounds similar to speaking with your hand over your mouth.	A common cause is an overdriven signal or some other cause that eliminates or reduces the signal level at frequencies inside the key range for voice (between 440 and 3500 Hz).
Soft voice	Soft voice is like a low voice that is hard to hear.	Soft voice is usually caused by too much attenuation on the signal, possibly introduced at one of a number of points in the network such as a voice gateway when trying to reduce echo.
Tinny voice	Tinny voice is similar to listening to an old-fashioned wireless broadcast.	A common cause is an overdriven signal or some other cause that eliminates or reduces the signal level at frequencies outside the key range for voice (less than 440 Hz and greater than 3500 Hz) but important to the richness of the voice.

Before we get into the discussion of network management concepts and how they apply specifically to VoIP, we discuss the strategic value of VoIP in providing business-critical services to customers and hence the importance of managing VoIP networks.

Strategic Importance of VoIP and Management

Businesses and consumers have been able to find applications for the communication network and turn them into critical services. This phenomenon has been in effect since the invention of the telegraph in the mid-1800s, when stock and commodity traders used it for obtaining stock prices, and it became the de facto stock ticker for Wall Street. Telephone systems have been around for over 100 years now. You might lose power but the phone still works. Users have gotten to the point where they expect the phone system to always be working. The public telephony system has supported life support services and is the backbone of how businesses conduct day-to-day activities. Therefore, any interruption in services provided by a communications network can have dire consequences.

VoIP technology is no different when it comes to the business criticality of the services it provides. Initially, cost cutting might be the deciding factor for service providers in making the initial investment by leveraging their data networks. The cost savings aspect is rooted in this fundamental technology difference: PSTN networks use the SS7 protocol that runs on a dedicated signaling network TDM to set up a call path or circuit. This circuit requires 64 kbps of network bandwidth for a single voice channel throughout the network. Packet telephony or VoIP uses the network bandwidth more efficiently by using statistical analysis to multiplex voice traffic, including both call signaling and the voice stream, alongside data traffic and sharing it across multiple logical connections. This reduces the overall bandwidth requirement, especially if it is combined with compression techniques that are possible only with packet switching. In addition to optimized bandwidth utilization, service providers can relay the toll savings to their consumers by making their services more competitive by leveraging the Internet instead of paying for interconnection charges.

Similarly, enterprises are also interested in leveraging their intranet, which spans all their corporate footprint, especially if it has excess capacity, to save on long-distance toll charges. As smartphones penetrate the cellular phone industry with technologies such as Edge, 3G, and PDSN to access the Internet, the consumers of these devices have begun experimenting with VoIP by leveraging clients such as Skype regardless of bandwidth constraints and the best effort for packet delivery nature of Internet. This has made mobile service providers seriously consider offering VoIP as they make the transition to 4G networks with more efficient IP transport to prevent (or compensate for) revenue erosion because of the diminishing adoption of traditional voice over the public land and mobile network (PLMN).

VoIP technology in general and Unified Communications in particular allow businesses to push additional applications to their employees, partners, and customers in their ecosystem to increase productivity by coupling them with their business processes to provide a comprehensive business-centric collaboration platform. This includes using

unified messaging that makes the delivery of voice and voicemail through any IP-capable medium, including email clients, web browsers, or smartphones. Other advanced features include coupling IP phones with web services, tracking users' present status and location, integrated information systems, the ability to initiate a call on demand from a website, mobility over IP network, Single Number Reach (SNR) for greater access, and location awareness.

A few of these examples include a collaboration-enabled radiology system in a hospital that allows a radiologist to contact the referring physician from his or her workstation used to read the radiology images without the need to look up the contact information using the preferred way of communications. The radiologist can share images and other diagnostic details related to a particular study. This saves time, which increases productivity and additionally provides compliance with health-care standards because the entire transaction can be coordinated and recorded over an IP infrastructure for auditing purposes. Similarly, a nurse in a hospital setup can scan a list of available doctors when confronted with an urgent medical need, and with the click of a mouse, speak instantly with the most appropriate specialist, wherever he or she is located. If a VoIP call control system is integrated with hospital databases and monitoring systems, it can provide a real-time view of the patient's history and vital statistics during the call, expedite test orders, and keep track of prescriptions by converting system messages into alerts and notifications and immediately delivering them to the medical staff. The VoIP system can also tie into external systems such as medical insurance providers to make the entire care provider system more efficient.

Some collaboration-enabled business processes are applicable in other market segments, such as finance, manufacturing, education, and government, when a VoIP system is integrated with other productivity applications and business systems process management systems, including SAP, Oracle, PeopleSoft, Salesforce.com, and other Business Process Management (BPM) platforms.

Because IP packets carrying voice or any other payload can be rerouted, copied, stored, originated from different IP-capable interfaces such as a telephone, instant messaging, email, and web presence, it is changing the way customer relationship management (CRM) systems, including contact centers, are modeled. VoIP provides agility when designing CRM systems by providing contact center agents with consolidated and up-to-date contextual customer records along with a customer's preferred medium of communication. Contact center software typically uses either time or skill-level selection. Time selection is based upon agent free time. Skill-level selection is what is mentioned here. In the past, time selection was mainly used. Skill selection requires more intelligent software. Also, the contact center agents can be located anywhere in the world, which allows an organization to have access to different pools of expertise anytime from anywhere. This flexibility is cheaper over IP networks and allows easier expansion as business needs grow over the longer period and can bring agents online on demand on short notice.

Because VoIP provides more than a mechanism for voice transport and allows organizations to develop their critical business processes, its strategic value is significant. We have to keep in mind that VoIP networks are now required to provide the same level of service as the PSTN, including emergency services, and comply with the same regulations that

were once meant for the PSTN only. This makes the management of a VoIP network a business-critical function. The next section discusses various management methodologies that can be used to devise a comprehensive plan to manage the entire life cycle of a VoIP network.

Network Management Methodologies

Because of the critical nature of a communications network, as discussed earlier, the Consultative Committee for International Telegraph and Telephone (CCIT), dating back to 1865, has provided guidance and structure for network management. Its focus was primarily telecommunications network management for stability and facilitating interoperability between national networks to enable global communication. The International Telecommunications Union (ITU-T), which was created in March 1993 and replacing CCIT, developed one of the first modern network management methodologies, including the Telecommunications Management Network (TMN), which was followed by a more comprehensive methodology that defines key management areas covering Fault, Configuration, Accounting, Performance, and Security (FCAPS). The United Kingdom's Central Computer and Telecommunications Agency (CCTA) created the Information Technology Infrastructure Library (ITIL), focusing on the service delivery and support methodologies. The TeleManagement Forum (TMF) defines a framework called the enhanced Telecommunications Operations Map (eTOM) to help its members reduce the costs and manage risks associated with creating and delivering services profitably.

Telecommunications Management Network

The International Telecommunications Union (ITU-T) introduced recommendation M.3010 in May 1996. This delivered the concept of the Telecommunications Management Network (TMN). Recommendation M.3010 provided a framework for service providers to manage their service delivery network. This framework defined four management architectures at different levels of abstraction: functional, physical, informational, and logical layers. The TMN-prescribed framework also provided a common methodology and logic that was applicable to the management of private corporate-owned IT networks. Figure 1-4 includes the four layers of abstraction defined in M.3010.

The Business Management Layer (BML) and the Service Management Layer (SML) provide a relationship between the IT and the business of the corporation. The Network Management Layer (NML) deals with fault and performance data for the network. The Element Management Layer deals with configuration management, fault, and performance at the device level.

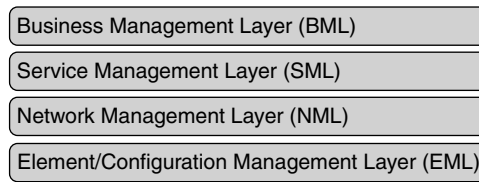


Figure 1-4 *TMN Logical Layers*

FCAPS Model

The ITU-T introduced another recommendation, M.34000, in April 1997, that further defined the general management functionality. Recommendation M.3400 breaks the management system into five functional key areas of FCAPS. The FCAPS model does not have a particular focus on the business-related role of a management system within the telecommunications network. However, this functional model does provide informational elements that can help in the business aspects of the telecommunications network, such as analyzing accounting and performance data to extract meaningful information to make business decisions about the network's capacity for new service offerings. This concept is discussed in detail in Chapter 8, "Trend Analysis and Optimization." The following sections describe the five functional areas in the FCAPS model.

Fault Management

Fault management is about recognizing the problem through continuously monitoring the entire network, correlating the fault data, and isolating the problem to the source. Fault management involves the entire life cycle, including fault detection, handling of alarms, fault isolation by the filtration and correlation process, fault correction for network recovery, tracking the incident by error logging, and managing the entire workflow through a trouble ticketing system.

Configuration Management

Configuration management deals with managing the configuration change control process, including commissioning and decommissioning of network devices, backing up and restoring the methodology for configurations, and overall workflow management for the administrators performing the configuration changes.

Accounting Management

Accounting management covers methods to track usage statistics and costs associated with time and services provided with devices and other network resources. Some aspects of accounting management overlap with fault management to provide comprehensive data to validate service-level agreement compliance.

Performance Management

This area covers the network management system's (NMS) capability to track system statistics to help identify network trends. In a sense, this provides feedback to the fault management layer to establish thresholds for determining the network and device-level faults proactively. It also provides data for capacity planning for network growth and new service offerings (by determining network readiness).

Security Management

Security management addresses access rights that include authentication and authorization, data privacy, and auditing security violations.

Figure 1-5 includes a list of the FCAPS functional elements.

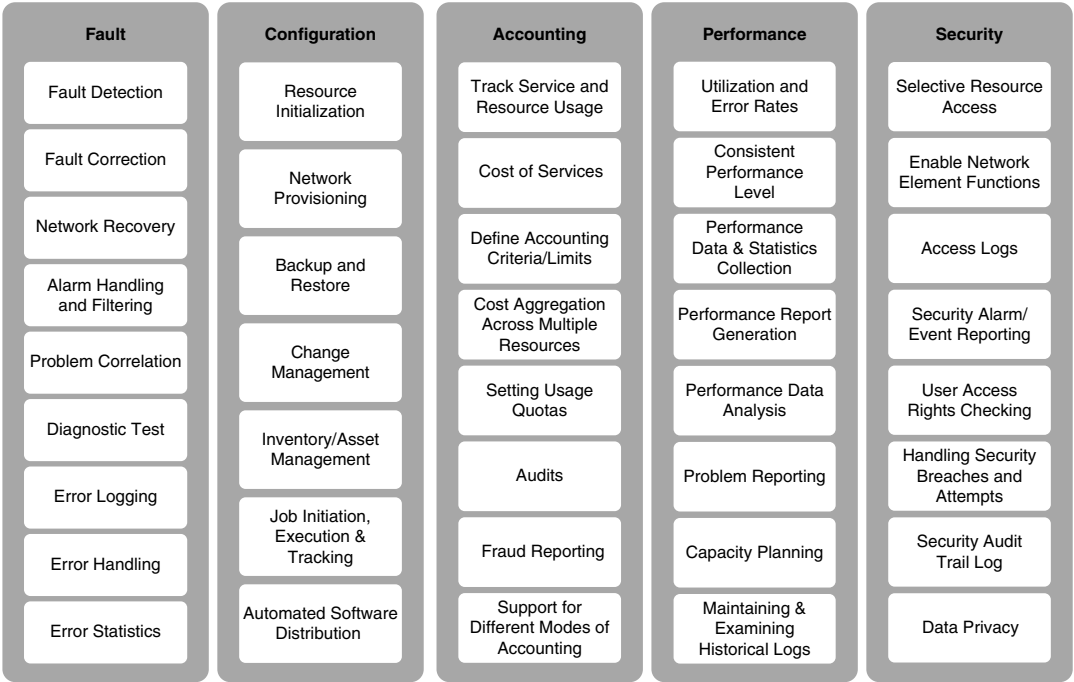


Figure 1-5 FCAPS Functional Elements

Information Technology Infrastructure Library (ITIL)

ITIL is a set of best practices for IT service management. It has become a worldwide de facto standard in service management since the late 1990s. ITIL provides network administrators and CIOs a customizable framework for best practice to achieve quality service and overcome difficulties associated with the growth of networks and the services it offers to its subscribers. Because of its customizable nature, ITIL standards are being adopted by organizations both big and small to improve their business processes related to IT.

These are the five fundamental processes defined by ITIL version 3 that cover the entire life cycle of an IT project, from starting its architectural planning, spanning through the design and implementation phases, and covering the operational phases to a continued loop with the services optimization.

Service Strategy

Service strategy provides guidance to all IT service providers about the following:

- What services should be offered
- Who the services should be offered to
- How the customer and stakeholders perceive and measure the value, and how this value is created
- How to evaluate and leverage partners for partial or complete sourcing of the services
- How visibility and control over value creation are achieved through financial management
- How the allocation of available resources is tuned to optimal effect across the portfolio of services
- How service performance is measured

The financial management aspect of service strategy covers the function and processes responsible for managing a provider's budgeting, accounting, and charging requirements. It provides IT with the quantification in financial terms about the value of network services and the infrastructure upon which they are delivered and the qualification of operational forecasting.

Service portfolio management is a continuous and proactive process that deals with defining services for the planning/concept phase, design, and transition pipeline, and maintaining them through these phases.

Demand management is targeted to understanding and influencing customer demand for services and the provision of capacity to meet these demands.

Service Design

Service design starts with a set of business requirements and ends with the development of a service solution designed to meet documented business requirements and outcomes for handover into service transition. It covers the following management aspects:

- Service Catalogue Management provides a single consistent source of information on all the agreed services and ensures that it is available to all the authorized users.
- The Service Level Management (SLM) process is intended to ensure that all operations services and their performances are measured in a consistent, professional manner throughout the IT organization, and that the services and the reports produced meet the needs of the business and customers. This includes service-level agreements (SLA), operational-level agreements (OLA), and the production of the Service Improvement Plan and the Service Quality Plan.
- Capacity Management includes business, service, and capacity management across the service life cycle. It is therefore important to consider capacity management at the onset of the design stage.
- Availability Management deals with availability-related issues pertaining to services, components, and resources to ensure that the availability targets in all areas are measured and achieved in a cost-effective manner.
- IT Service Continuity Management is targeted toward maintaining the appropriate ongoing recovery capability within IT services to match the agreed needs, requirements, and time scales of the business. It also ensures that all the activities are aligned with business continuity plans and business priorities.
- Information Security Management is part of the overall corporate governance framework. Its purpose is to ensure that complete information is available when required to the authorized personnel with focus on its authenticity and nonrepudiation.
- The Supplier Management process ensures that suppliers and the services they provide are managed to support IT service targets and business expectations while conforming to all the terms and conditions of their contracts and agreements.

Figure 1-6 illustrates the ITIL Services and Support delivery components. Service Design, Service Transition, and Service Operation are cyclic processes based on Service Strategy. Each of these phases provides feedback to the next step for continuous improvement, as shown by the curved arrows in the diagram.

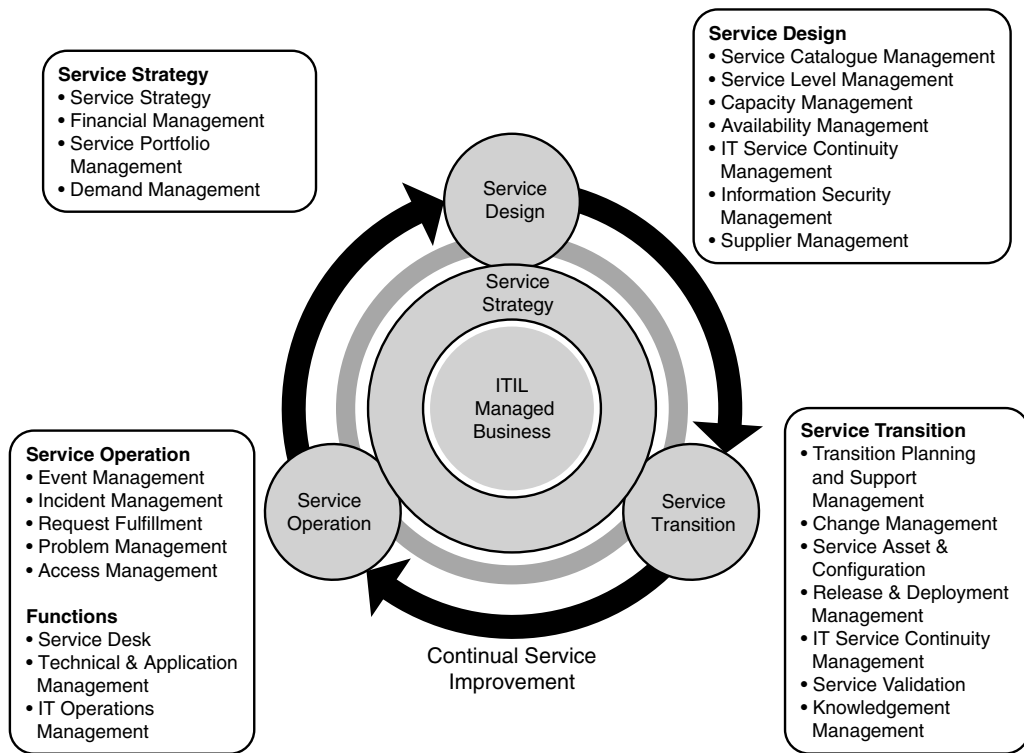


Figure 1-6 *ITIL Service and Support Delivery Components*

Service Transition

The main role of Service Transition is to deliver services that are required by the business into operational use. This is achieved through the following processes and activities:

- Transition Planning and Support is geared toward planning and coordination of resources to ensure that the requirements of Service Strategy as encoded in Service Design are effectively realized in Service Operations. Transition Planning is done by identifying, managing, and controlling the risks of failure and disruption across transition activities.
- Change Management ensures that changes are planned, prioritized, evaluated, tested, authorized, and implemented while being recorded and documented in a controlled manner.
- Service Asset and Configuration Management lay a framework for providing accurate information and control across all assets and relationships that go up an organization's infrastructure. Its scope can also be extended to non-IT assets and to internal and external service providers where shared assets need to be controlled. Cloud computing, hosted services, and software as a service (SaaS) are some of the examples for external or hybrid resources.

- Release and Deployment management covers the entire assembly and implementation of new or changed services for operational use from release planning through to early life support. It ensures effective release, and deployment delivers significant business value by delivering changes at optimized speed, risk, and cost, and offering a consistent, appropriate, and auditable implementation of useful business services.
- Knowledge Management ensures that the right person has the right knowledge at the right time to deliver and support the services required by the business.

Service Operation

Service Operation's main role is to deliver agreed-upon levels of service to the users and customers and to manage the applications, technology, and infrastructure that support the delivery of the services. It includes the following management processes:

- Event Management generates, detects notifications, and proactively monitors the status of network elements. It also includes managing the responses to these notifications received either reactively or proactively by prioritizing and correlating them for further analysis.
- Incident Management's main purpose is to restore normal service as quickly as possible and to minimize the adverse impact on business operations. The process includes prioritizing incidents according to urgency and business impact and categorizing them so that the appropriate skill set can be applied for remediation. The loop is closed only when the resolution to the problem has been confirmed by the end user as communicated through a Service Desk. This management process is often aided by network management tools.
- The Request Fulfillment process enables users to request and receive standard services, to source and deliver these services, to provide information to users and customers about services and procedures for using them, and to assist with general information, complaints, and comments.
- The Problem Management process is geared toward preventing problems and resulting incidents from happening, to eliminate recurring incidents, and to minimize the impact of incidents that are inevitable. This is done by diagnosing causes of incidents, determining the resolution, and ensuring that the resolution is implemented and verified.
- The Access Management process manages the rights and permissions for users to be able to access a service or group of services. It also tracks access and role changes for auditing purposes. In that sense, access management ensures the confidentiality, availability, and integrity of data and intellectual property.

Service Operations relies on key functions, including Service Desk, which is a central point of contact for users of services; Technical and Application Management, including the people's technical expertise and management of the network infrastructure and the applications running on it; and IT Operations Management, which is staffed by operators carrying routine operational tasks.

Continual Service Improvement

Continual Service Improvement is considered a feedback loop that provides a way for an organization to identify and manage appropriate improvements by contrasting its current position and the value it is providing to the business with its long-term goals and objectives, and identifying any gaps that exist. This is a continuous process to address changes in business requirements, to refresh technology cycles, and to ensure that high quality is maintained. Continual Service is shown in Figure 1-6 by arrows representing the feedback between service phases transition.

Enhanced Telecom Operations Map (eTOM)

Enhanced Telecom Operations Map, or eTOM, defines a business process framework that is part of TeleManagement Forum's Next Generation Operations Systems and Software (NGOSS). It is the most commonly adopted framework among telecommunications service providers. The eTOM framework defines three major process areas that are fundamentally technology and industry segment agnostic. Each area has four levels of processes defined, with each level covering specific process details. These three process areas include the following:

- **Strategy, Infrastructure and Product:** This area covers services offering a strategy and commit process and the entire infrastructure and product life cycle management. The four hierarchical levels in order include marketing and offer management, service development and management, resource development and management, and supply chain development and management.
- **Operations:** This area addresses operations support and readiness and fulfillment, assurance, and billing (FAB). The corresponding four levels include customer relationship management, service management and operations, resource management, and partner relationship management.
- **Enterprise Management:** This is an overlying process area that covers strategic planning, financial assets management, brand management and associated marketing activities, human resource management, research and development, stakeholder and external relations management, disaster recovery, security and fraud management, and quality management.

Note Because service providers offer products and services to enterprises relevant to VoIP and Unified Communications, including infrastructure such as Metro-Ethernet, MPLS, and other more comprehensive services such as UC as a Service (UCaaS), there was a growing need to map eTOM to ITIL. ITIL is widely adopted by enterprises. This mapping was targeted to better align SP's IT service management framework with its enterprise customers. In 2004, the TM forum formulated a team to focus on mapping eTOM to ITIL. The TM Forum's technical report is referred as "An Interpreter's Guide for eTOM and ITIL Practitioners." It is available at http://www.tmforum.org/community/groups/the_business_process_framework.

Comprehensive Network Management Methodology

The FCAPS model provides essential foresight and knowledge to optimize network performance through fault, performance, and configuration management by focusing on technology management aspects. It also addresses security, which is a major concern to network operators with respect to theft of service and data protection. ITIL provides the methodology and discipline to execute essentially all aspects of service management encompassing the network used for delivery of these services. The management framework and functions, as described earlier in the “Service Operation” section of the ITIL v3 foundation discussion in this chapter, overlap with specific steps laid out in the FCAPS framework. However, the ITIL methodology addresses certain tasks such as configuration under its Service Transition process definitions. TMN also possesses similarities with ITIL and FCAPS, where its definition of Business Management and Service Management layers match ITIL’s Service Strategy and Service Design definitions. The Fault and Performance and Element/Configuration Management layer covers the operational aspect of the service delivery network in the same manner as the Service Transition and Service Operation definition in ITIL and all the aspects of FCAPS. However, ITIL covers the entire life cycle of the service delivery network starting from planning, designing, implementation/commissioning, operation, and optimization phases.

There has been an increase in services in the telecom world because of competition and technological maturity that has brought us services such as VoIP and Collaboration services, including web conferencing, video, presence, and single-number reach, to name a few. This situation becomes more complex with new service models such as cloud computing and hosted or remotely managed services. IT has to continuously evolve with the introduction of new service businesses by being more agile to support and grow them. At the same time, typically there is less focus on business-to-IT alignment. Sometimes there is a tendency to measure productivity, profit, and loss for individual departments in an organization. This is because of a lack of a service-centric approach, which often ignores the fact that these services are being delivered on the converged network. An absence of services-centric views increases the time to adapt to new business models, market needs and makes it difficult to efficiently manage them. This fragmented approach also affects budgeting decisions that are made in silos and might not optimally realize the potential of the converged network. The management processes defined under the ITIL v3 Foundation provide a comprehensive framework to bridge the gap between business processes and IT operations.

The ITIL methodology is flexible enough to build a corporate-specific framework to correlate discrete data collection points illustrated by FCAPS and present them in service-centric dashboards that are relevant to the core business of the corporation. For example, it can provide IT decision makers with a view of the collaboration application on the network, concentrating on their availability, adoption, and capacity to grow in a business context that shows the savings as a result of business transaction efficiencies achieved by using these tools and applications. Chapter 8 introduces the concept of dashboards that provide an overall network view by focusing on key performance indicators in the context of the services offered by the network.

Any comprehensive solution must begin at the planning and design stage. When a problem is identified after introducing a new technology by following ITIL management methodology and FCAPS-based metrics, it might not be easy to remediate it. A proactive planning and design methodology executed with foresight ensures that the network is fundamentally ready for the introduction of new technologies such as voice and video on converged networks. This entails performing network readiness assessment, traffic engineering (that might be covered through capacity management as prescribed in ITIL), network transmission loss plan establishment, a policy for quality of service classes, and corporate operational readiness assessment. We cover these proactive aspects related to overall network performance and specific to VoIP in Chapter 6.

Because of the comprehensive nature of ITIL, the customization of its various processes make the ITIL compliance a rather long project, especially for large service providers and corporations. Of course, the ITIL methodology adoption is easier for greenfield deployments, but the existing service delivery networks should employ the fundamental operations strategy as laid out in FCAPS.

This concept, showing a relationship among FCAPS, ITIL, TMN, Planning, and Design methodology, is shown in Figure 1-7.

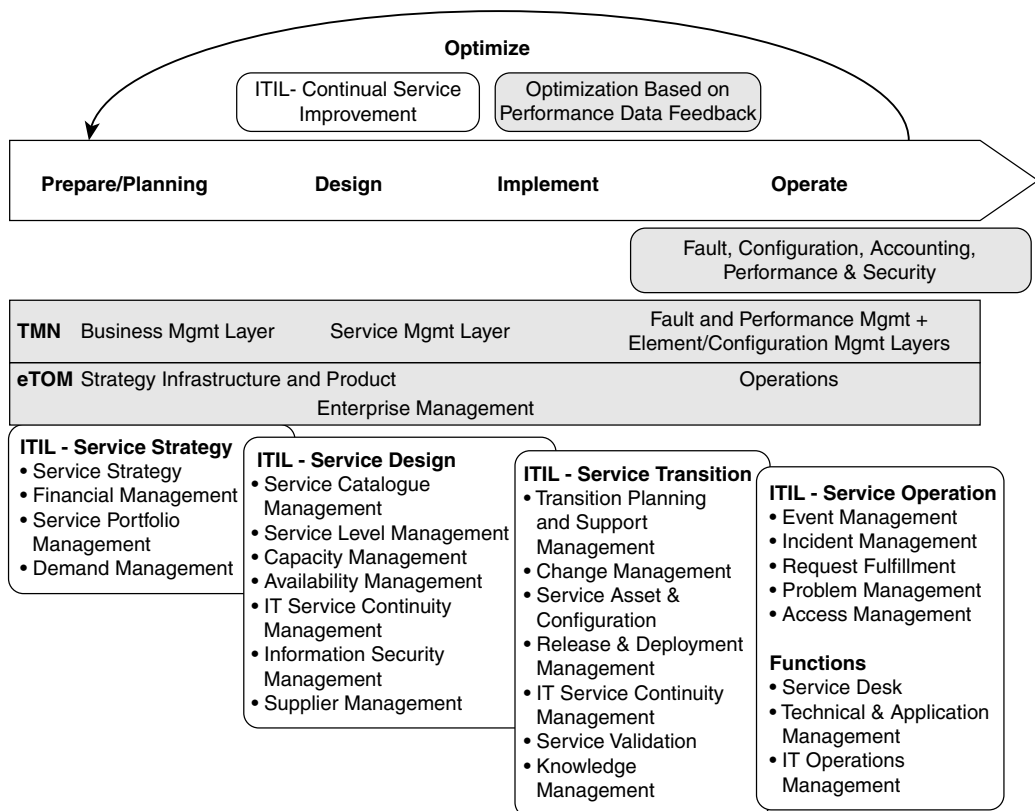


Figure 1-7 Comprehensive Network Management Methodology

Focusing on Performance Metrics

Since the advent of the service delivery networks, network operators have focused on managing availability. This includes proactive checks, including SNMP polling and reactive notifications through alarms and traps; for example, if the router is up, the link is down or congested, or the server needs to be restarted. However, these days, the service provider and enterprise networks have been able to achieve 99.9 percent uptime. The reasons include technology improvements, regulatory compliance with stricter penalties, and considerably heavier reliance on the networks where most of the enterprise applications are networks that motivate network managers to put more redundancy and other resiliency measures in the network.

A Yankee Group report titled “Performance is the New Mandate for Network Management” refers to a study on enterprise application management. This study found that enterprises report an average productivity drop of 14 percent when experiencing performance degradations in Oracle, SAP, PeopleSoft, Siebel, and custom .NET and J2EE applications. Ashton, Metzler & Associates conducted a survey of 176 IT professionals that indicated that over a quarter of network operations centers (NOC) do not meet their organizations’ current needs. To fulfill current and emerging requirements, NOCs are being driven to do a better job of managing application performance, to implement more effective IT processes, and to be able to troubleshoot performance problems faster.

To adopt a more proactive approach as opposed to solely relying on alarms, traps, and other outage notifications, network operators must be able to do the following:

- Monitor the health of the network at the device level by looking at key performance indicators (KPI) to determine availability, processor and memory utilization, and interface utilization, including packet errors and discards.
- Maintain the visibility of the network utilization, including traffic patterns on different layers, segments, classes of service, and various links including delay, jitter and packet loss, application performance, and when users are the most active in relationship with their interaction with the network services and application. A comprehensive network visibility is especially critical for voice and video service because the user experience is subjective to large extent and cannot be directly measured with individual KPIs. Having contextual data also helps troubleshoot the problems faster and reduces the Mean Time To Resolve (MTTR) problems.
- Establish a baseline for normal performance. It involves real-time monitoring and analysis of historical records such as call detail and diagnostic records in the case of VoIP applications and performance logs from servers running vital applications for the business. After the baseline is established over a period of time through performance monitoring, the baseline represents a normal network activity level. This characterization helps the administrator define a threshold that is approximately equivalent to the baseline with some additional tolerance built into it. This tolerance level might depend on the network link and/or component specifications for safe operation. Accurate establishment of network performance thresholds can also aid in determining deviation as they start to occur. In that sense, this practice provides an opportunity to mitigate them before they cause service interruption. Network baseline also

provides guidance in determining when and where there is inappropriate and wasteful use of network resources. Also, it helps network engineers understand where there is the potential for cost savings and efficiencies on the network. These concepts are discussed in more detail in Chapters 7 and 8.

To accomplish these three key tasks, the network operator needs a framework and a methodology to determine what is practical to monitor, where to monitor, how often to monitor (to avoid adverse effects on the monitored devices), and what information needs to be correlated to provide meaningful information. There is a famous network management idiom: “You cannot manage what you don’t measure, and you cannot measure what you don’t collect.” Performance metrics are core to an FCAPS-based management scheme. The Fault and Performance Management layer in TMN, Service Operations and its functions in ITIL, and Accounting and Performance in FCAPS revolve around KPIs or key performance metrics (KPM). In that sense, performance metrics have a pivotal role in overall network management in the operational stage as well as in network optimization and growth planning.

The main focus of this book is to define and illustrate an approach of collecting, analyzing, and trending data using performance counters. This approach is based on defining and measuring KPIs at different layers (the physical layer, data link layer, IP layer, and application layer) from various devices in the VoIP network and using this data to gauge the health of the VoIP network using a consolidated dashboard view. Figure 1-8 shows elements of network management methodology that contribute to KPIs for performance measurement and proactive fault detention.

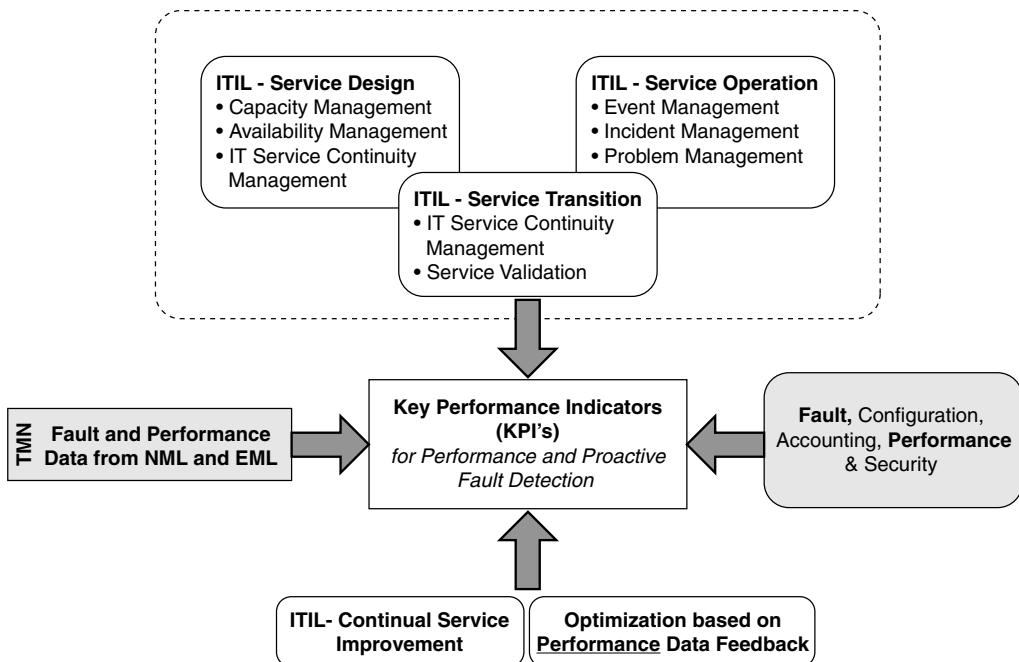


Figure 1-8 Role of KPIs in Network Management Methodology

Summary

Some of the voice quality problems such as noise, echoed voice, and volume distortion are typically related to planning, design, and implementation problems in the VoIP network. The root cause of these issues can vary from physical impairments to codec issues to problems with echo cancellation. The VoIP provider needs to take proper steps in planning and designing the network to avoid these issues. Other issues such as garbled voice and related problems are primarily caused by network problems such as delay, jitter, and packet loss. We discussed the key network management standards, including TMN, FCAPS, ITIL, and eTOM, that define a framework for comprehensive network management. The strategic nature of a communications network in general and VoIP in particular demands a comprehensive network management strategy that covers the application and the underlying transport network.

The next chapter discusses a comprehensive network management strategy for VoIP networks and describes how this approach helps service providers/enterprises in performance management and optimization of the VoIP network. The performance management and optimization strategies are discussed in detail in Chapters 7 and 8.

Reference

1. Biswas, Suparno. *From FCAPS to ITIL: An Optimized Migration*. February 21, 2005.
2. International Telecommunications Union website for information about ITU-T. <http://www.itu.int/ITU-T>.
3. Parker, Jeff. FCAPS, TMN & ITIL—Three Key Ingredients to Effective IT Management. May 6, 2005.
4. Davidson, Jonathan, James Peters, Manoj Bhatia, Satish Kalidindi, and Sudipto Mukherjee. *Voice over IP Fundamentals*, Second Edition. Indianapolis, IN. Cisco Press, 2006.
5. Siddiqui, Talal. “Managing Voice Quality in Converged IP Networks,” BRKVVT-2301 presentation. Cisco Systems Inc., July 2007. The TeleManagement Forum. <http://www.tmforum.org>.
6. NetQoS. “Performance First.” <http://www.netqos.com>, 2009.
7. Werbach, Kevin. “Using VoIP to Compete.” *Harvard Business Review*, Vol. 83, No. 9, September 2005.