

TechBrief

Extreme Networks

Combining Voice over IP with Policy-Based Quality of Service

Introduction

Businesses have traditionally maintained separate voice and data networks. A key reason for this is that legacy network technologies could not meet the diverse performance requirements of voice and data. Recent advances in networking technology, including fast Ethernet, wire-speed switching and Policy-Based Quality of Service management, have made it possible to build converged voice and data networks.

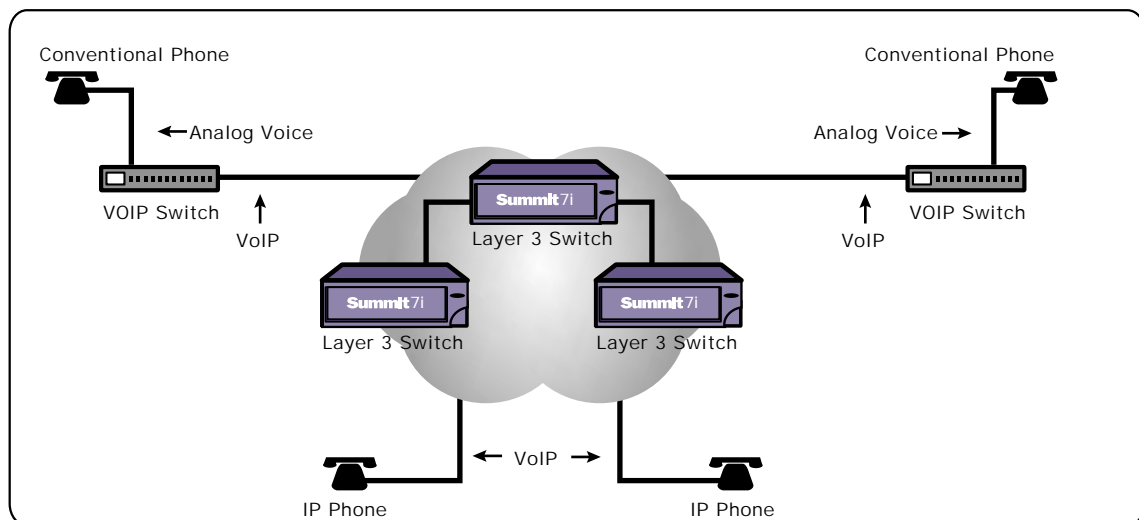
Converged networks reduce costs by eliminating redundant hardware, communications facilities and support staffs. Converged networks also enable a new generation of integrated voice/data applications. For example, users of web-based e-commerce applications can view product information while talking with customer service agents in a call center. With converged networks this can be done through a single network connection.

The focus of most converged network strategies is voice over IP (VoIP). VoIP refers to the transmission of telephone conversations over a packet-switched IP network. This IP network could be as small as a single subnet private LAN, or as large as the public Internet.

Voice over IP on the LAN

With VoIP on the LAN, phone conversations are converted to a stream of IP packets and sent over an Ethernet network. This network is usually restricted to a building or campus.

As VoIP technology matures, new conversion methods may emerge while existing ones become obsolete. Regardless of the method that is used to convert VoIP traffic for LANs, one fundamental process will remain the same: VoIP traffic will always traverse the LAN as a stream of IP packets.



Voice over IP on the LAN

Section II

Quality of Service and Infrastructure Requirements to Support Voice over IP

VoIP QoS Requirements

H.323, the global standard for packet-based multimedia communication like VoIP, provides telephone functionality that is comparable to the public switched telephone network. But the other key requirement for successful VoIP communications is quality of service (QoS). Voice communications requires networks with very low latency, low jitter, and minimal packet loss. Two factors drive these QoS requirements:

- Very high user expectations
- The technical requirements of real-time voice communications

Telephone users have very high expectations because they are accustomed to the QoS provided by the PSTN and private PBX-based networks. These connection-oriented, circuit-switched networks provide each user with dedicated bandwidth for the duration of each call. The result is extremely low latency and jitter, and minimal disruption due to "noise" on the connections. Low latency allows users to carry on natural conversations. Users differ in their delay tolerance, but a good rule of thumb is to limit one-way delay to about 150 milliseconds (ms). This delay budget includes the processing delays introduced by the end systems plus the latency of the network.

When coders/decoders (codecs) in VoIP terminals compress voice signals they introduce three types of delay:

- Processing, or algorithmic, delay – the time required for the codec to encode a single voice frame
- Look ahead delay – the time required for a codec to examine part of the next frame while encoding the current frame (most compression schemes require look ahead)
- Frame delay – the time required for the sending system to transmit one frame

The following are some commonly used ITU-T standard codecs and the amount of one-way delay that they introduce:

- G.711 uncompressed 64 Kbps speech adds negligible delay
- G.729 encodes speech at 8 Kbps and adds a one-way delay of about 25 ms.
- G.723.1 encodes speech at 6.4 Kbps or 5.3 Kbps and adds a one-way delay of about 67.5 ms.

In general, greater levels of compression introduce more delay and require lower network latency to maintain good voice quality. When using G.723.1 compression, for example, the network component of one-way delay should not exceed:
 $150\text{ms} - 67.5\text{ ms} = 82.5\text{ ms}$.

Building the Right Infrastructure to Support Voice over IP

One of the key challenges in implementing VoIP is to design and build an IP-based network that meets stringent QoS requirements and is comparable in performance to conventional circuit-switched telephone networks.

The high latency forwarding and best-effort delivery provided by traditional software based routers is generally not acceptable for streaming traffic like VoIP because it does not provide maximum latency guarantees or minimum bandwidth guarantees.

From the perspective of an IP-based Ethernet network, a VoIP packet containing part of a telephone conversation is no different than a data packet containing part of an e-mail. Both packets are received on an ingress port of an Ethernet switch and need to be forwarded out the egress port of an Ethernet switch. From the perspective of the end points, different types of traffic have very dissimilar requirements.

For example, e-mail traffic is typically handled using the store-and-forward process. It does not have to happen in real-time. An e-mail transmission does not have to be streamed from one end-point to another end-point to be successful.

Conversely, VoIP traffic is a real-time process. To complete a successful VoIP session, the network must be able to support the streaming of VoIP packets between the two end-points for the duration of the phone conversation. VoIP traffic requires a network to guarantee bandwidth and capacity for VoIP traffic.

To support VoIP traffic consistently and reliably, a network must be able to provide three things:

- Packet-forwarding latency that does not exceed the maximum tolerable level for a VoIP conversation
- Packet-forwarding jitter, which is the variation in latency over time, that does not exceed the maximum tolerable level to sustain a VoIP session
- Guaranteed network bandwidth and capacity for VoIP sessions during periods of network congestion

In other words, a network needs to provide performance – low latency and low jitter – and protection – quality of service.

Most VoIP sessions require one-way latency of not more than about 150 milliseconds. This delay budget is reduced by any delays introduced by codecs in the end systems. When round-trip delays exceed approximately 300 ms., natural human conversation becomes difficult. Anyone who has tried to carry on a conversation on a satellite link has experienced the affects of long delays.

Depending on the type of voice-compression method used, each one-way VoIP transmission requires between 32 Kbps to 64 Kbps of bandwidth. Some compression methods such as G.729 take the bandwidth required below 8 Kbps. As you can see, the bandwidth that is required for each VoIP session is relatively low. The challenge is to make that bandwidth available regardless of network utilization.

Section III

Optimizing Voice over IP with Policy-Based QoS and Wire-Speed Switching

Optimizing VoIP Performance with an Extreme Networks Infrastructure

Extreme Networks' Layer 3 switches, combined with Policy-Based QoS, are ideal for supporting VoIP traffic, and have been designed to accommodate VoIP traffic from the beginning. Our high-performance Summit stackable and BlackDiamond chassis switches deliver VoIP traffic at wire-speed with low latency and low jitter. At the same time, our ExtremeWare software suite delivers Policy-Based Quality of Service to protect the delivery and performance of mission-critical VoIP traffic, making sure it gets through – even during periods of network congestion. In other words, Extreme Networks' switches provide the performance and protection necessary for running VoIP traffic on a switched Ethernet network.

Extreme Networks delivers optimal performance for VoIP traffic by leveraging powerful switching and QoS capabilities that are common to all Summit and BlackDiamond switches:

- **Wire-speed Performance** – Wire-speed switching at Layer 2 and Layer 3 ensures that the packet forwarding speed, whether it is for a data packet containing part of an e-mail or a VoIP packet containing part of a telephone conversation, will never be an obstacle to deploying VoIP on the LAN.
- **Non-Blocking Architecture** – Every Extreme Networks switch has a non-blocking switch fabric, which means that the internal switch backplane will never be a source of congestion or packet loss for VoIP traffic.
- **Low Latency** – A byproduct of wire-speed switching is low forwarding latency. Extreme Networks switches introduce 8-12 microseconds (msec) of latency, or delay, when forwarding a 64-byte Ethernet packet at Layer 2 or Layer 3. This is an order of magnitude less than the 150-300 milliseconds (ms) of delay that VoIP packets can tolerate. Consequently, the packet-forwarding delay introduced by Extreme Networks switches is over 1,000 times less than the maximum that VoIP traffic will tolerate.
- **Low Jitter** – Jitter is a measure of the variation in latency over time. In an ideal situation with no jitter, every packet arrives at its destination with the exact same amount of latency over an infinite period of time. The reality with all networks is that some amount of jitter will exist. The challenge is to minimize that jitter. Extreme Networks switches exhibit an average jitter of 10 microseconds (msec). Jitter of less than 1 millisecond (ms) is excellent for VoIP traffic. Therefore, the Extreme jitter rate is a thousand times closer to 0 than what is required for a successful VoIP session.

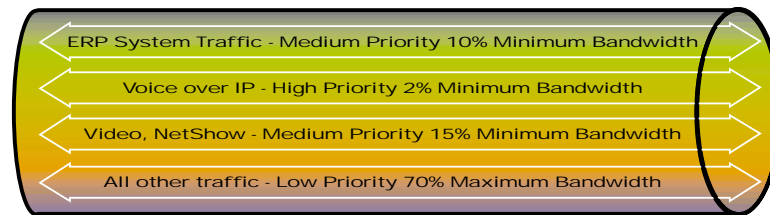
Protecting VoIP Services using Extreme's Policy-Based Quality of Service

Extreme Networks' Policy-Based Quality of Service capabilities provide a vital and necessary component to accommodating VoIP on the LAN – protection.

Policy-Based QoS allows network architects to create policies that guarantee network availability for VoIP traffic during times of network congestion. For example, real-time voice traffic can be guaranteed a specific amount of bandwidth to minimize latency and jitter while H.323 call signaling is simply assigned a high priority to ensure quick call set-up times.

Extreme Networks' performance capabilities ensure that the internal switch architecture will never be a point of congestion. However, network congestion, or over-subscription, is determined by traffic patterns.

If two wire-speed 100 Mbps traffic streams originating from different ingress ports are destined for the same 100 Mbps egress port, the egress port will be over-subscribed by a ratio of 2:1. In a non-QoS scenario, each 100 Mbps stream would be able to get half of its traffic through while the other half will be dropped. With Policy-Based QoS, you can create a policy that favors one stream over the other during periods of congestion or over-subscription.



A very effective QoS strategy is to first allocate bandwidth to broad classes of traffic and then prioritize traffic within each traffic class. For example, 10% of bandwidth could be allocated to ERP applications, 2% to Voice over IP, 15% to remote conferencing and streaming media, and the remainder shared by all other traffic.

Let's say that one of the streams represents VoIP traffic and the other stream represents e-mail traffic. Since e-mail traffic is not a real-time streaming application, it can easily tolerate a re-transmission of data, whereas VoIP traffic cannot. To ensure that the VoIP traffic gets all of its traffic through with low latency and low jitter during this over-subscription scenario, a QoS policy can be created that gives higher priority and more bandwidth to VoIP traffic. The higher relative priority and percentage of bandwidth assigned to VoIP traffic means that all VoIP traffic will be forwarded out of the egress port before any competing e-mail traffic is forwarded out of the egress port.

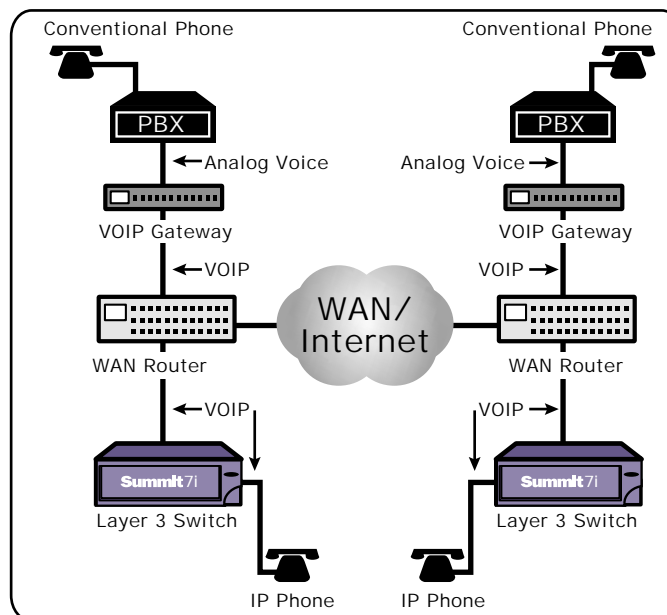
Voice over IP on the WAN

VoIP traffic can be sent over a variety of IP-based wide area networks (WANs). An IP-based WAN can be one of the following:

- A private enterprise WAN made up of leased lines, a frame relay service or an ATM service.
- A public IP carrier such as Qwest or Level 3
- The public Internet

For a business that is considering voice and data convergence, putting VoIP on the WAN is as important as VoIP on the LAN. The reasons for this are primarily economical. Cost savings can be immediate when long distance phone calls are diverted from the public switched telephone network (PSTN) and sent over an existing IP-based WAN. Running VoIP traffic on the WAN can be done in several ways:

- If the voice traffic is coming from a PBX, then a VoIP gateway will be required to convert voice traffic from the PBX into IP packets for transmission over the IP-based WAN. Similarly, a VoIP gateway will be required at the other end to convert VoIP traffic back into the format used by the PBX. The IP-based WAN can be a private data network, a public IP carrier or the public Internet.
- If the voice traffic has already been converted to VoIP traffic on the LAN, then the VoIP traffic will be transmitted over the IP-based WAN like any other IP data traffic.



Voice over IP on the WAN

Summary

Converged networks reduce costs by eliminating redundant hardware, communications facilities and support staffs. Converged networks also enable a new generation of integrated voice/data applications. Successful VoIP initiatives require the telephony functions defined by the H.323 standards plus IP networks that are capable of providing QoS comparable to that experienced by users of the PSTN and private PBX-based networks.

Policy-Based QoS allows network architects to create policies that guarantee network availability for VoIP traffic during times of network congestion. Policy-Based QoS allows network architects to create policies that guarantee network availability for VoIP traffic during times of network congestion. For example, real-time voice traffic can be guaranteed a specific amount of bandwidth to minimize latency and jitter while H.323 call signaling is simply assigned a high priority to ensure quick call set-up times.

Extreme Networks' Layer 3 switches, integrated with Policy-Based QoS, are ideal for VoIP traffic. Our high-performance Summit stackable and BlackDiamond chassis switches deliver VoIP traffic at wire-speed with low latency and low jitter. Extreme Networks' performance capabilities ensure that the internal switch architecture will never be a point of congestion. Extreme Networks' switches provide the performance and protection necessary for running VoIP traffic on a switched Ethernet network.

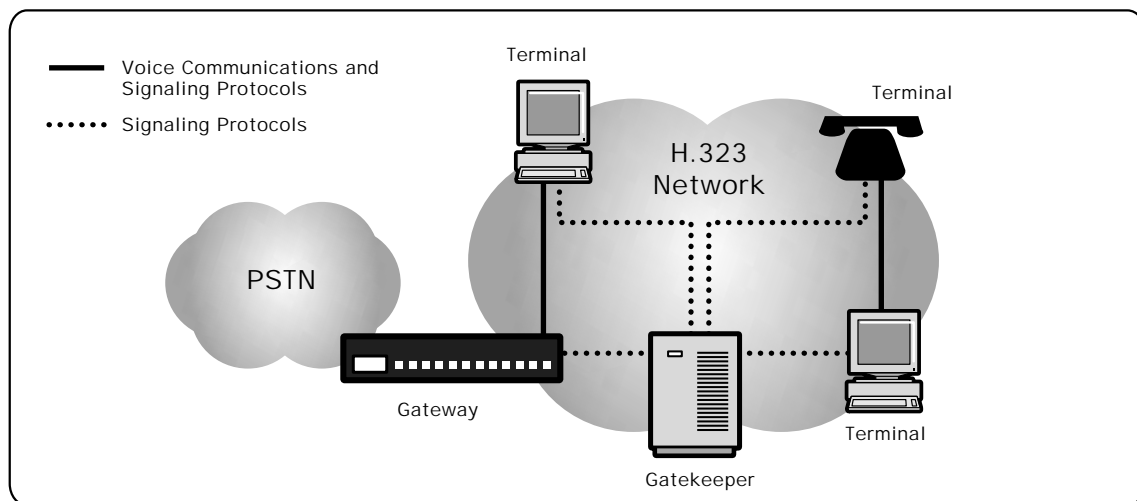
Section IV

An Overview of Voice-over-IP Standards and Protocol

An Overview of VoIP Standards

International Telecommunications Union (ITU-T) Recommendation H.323 is a global standard for packet-based multimedia communications, including VoIP. H.323 can be implemented on WANs or LANs.

H.323 is often referred to as an “umbrella” standard because it references a number of other standards that support multimedia communications. H.323 defines a set of network components and protocols that support real-time audio, video, and data communications. The following diagram shows the components that make up an H.323 VoIP network.



Components of an H.323 Voice-over-IP network

These components support real-time voice communications between end users and provide PBX-like network control functions.

- **Terminals** are the end-user devices that support two-way, real-time voice communications across H.323 networks. The most common terminal types are:
 - **IP phones** – This option uses a telephone with a built-in codec and embedded H.323 software, network interface card (NIC) and IP protocol stack. The IP phone is plugged directly into an Ethernet LAN just as a PC or other Ethernet station would be.
 - **PC phones** – This option uses PC software and hardware to enable a PC to become a VoIP phone on the network. This option is very similar to the IP phone option with one major difference: the PC phone typically has one NIC that will be used for both VoIP packets and data packets. In the case of an IP phone, the NIC is dedicated to only processing VoIP packets.
- **Gateways** enable communications between H.323 VoIP users and users of non-H.323 networks, most often the public switched telephone network (PSTN) or private PBX-based networks. Gateways allow users of conventional phones to communicate with VoIP users.
- **Multipoint Control Units (MCU)** are optional H.323 components that support multipoint conferencing, commonly called conference calls
- **Gatekeepers** provide control functions similar to those provided by PBXs and carrier switches in conventional voice networks. These devices control call setup and they can provide additional functions such as call forwarding, conference calling, and call waiting.

H.323 Protocols

H.323 defines the data stream formats and protocols that endpoints use to communicate with one another. It also defines the management and control protocols used between terminals, gatekeepers, gateways and MCUs. The following diagram shows the protocol stack implemented by H.323 endpoints (terminals and gateways) in VoIP networks.

Audio Application	Video Application	Terminal Control and Management				
Audio Codecs G.711 G.729 G.723.1	Video Codecs G.711 G.729 G.723.1	Real-Time Control Protocol	H.225.0 Registration, Admission, and Status	H.225.0 Call Signaling	H.245 Control Signaling	T-120 Data
Real-Time Protocol		User Datagram Protocol (UDP)			Transmission Control Protocol (TCP)	
Internet Protocol (IP)						

Protocol stack implemented by H.323 endpoints in a voice-over-IP network

Codecs digitally encode and decode audio or video signals for transmission across an H.323 network. In VoIP networks only audio codecs are used. Codecs differ in the encoding techniques that they use and the bit rate of their digital output streams. Codec compatibility is essential to VoIP interoperability. Endpoints cannot "speak" to one another unless they use compatible codecs.

H.323 end points must support at least one audio codec – the G.711 standard. This ensures basic interoperability between all H.323 terminals. The G.711 standard encodes audio at 64 Kbps and supports the pulse code modulation (PCM) that is widely used to encode voice on the PSTN. Other audio codecs may optionally be implemented. Some of these codecs and their transmission rates are:

- G.723.1 – 6.4 Kbps or 5.3 Kbps
- G.728 – 16 Kbps
- G.729 – 8 Kbps

Note that these standards use less bandwidth than G.711 because they compress audio signals. These compression algorithms take advantage the repetitive patterns found in human speech. Network designers should be aware that these low-bandwidth codecs output digital voice in very short frames – typically 10 to 30 bytes in length. In addition to compression, most VoIP terminals also implement silence suppression to eliminate network traffic during pauses in conversations.

Real-Time Protocols

Digitized audio and video streams are transported between endpoints by the real-time protocol (RTP). RTP is a connection-oriented, end-to-end protocol that is designed to transport delay-sensitive information. RTP identifies the encapsulated payload type and includes sequence numbers and time stamps that can be used to synchronize real-time information flows. RTP uses the connectionless, unreliable UDP transport protocols rather than TCP because retransmission delays disrupt real-time audio and video streams.

The real-time control protocol (RTCP) works with RTP to provide sending software with feedback on the quality of service being experienced by the receiver. RTP reports QoS parameters including packet loss and the amount of jitter being experienced. The sender can adjust transmission rates based on this feedback.

Signaling Protocols

Endpoints use the H.225 Registration, Admission, and Status (RAS) protocols to register with a gatekeeper. The H.225 RAS protocol runs over UDP.

The H.225 Call Signaling standard defines the protocols that endpoints use to set up and release connections. H.225 messages are transported on TCP connections.

After a call is set up, endpoints use H.245 Control Signaling to exchange information about their capabilities. For example, endpoints negotiate the use of audio codecs to ensure that both ends of the conversation are “speaking” the same language. H.245 messages are transported on TCP connections.