Latency and QoS for Voice over IP

Karie Gonia

Karie Gonia
Version 2.4b, Option 1
Latency and QoS for Voice over IP

Introduction

This paper will familiarize the reader with the fundamentals of a VoIP (Voice over IP) implementation and its effects. VoIP has been a front runner for companies looking to take advantage of IP. There are cost benefits, however with VoIP there are challenges. VoIP is latency driven therefore your Infrastructure must be able to support it. Latency is the delay that occurs when a packet crosses a network connection, from sender to receiver. Quality of Service (QoS) is yet another challenge. QoS challenges derive from different traffic types requiring different service levels from the network. QoS in a network must be adept of; prioritizing traffic types; interpreting traffic types (applications running over IP); and then conveying them over the network so that QoS requirements can be met. We will discuss the technical issues involved with assuring QoS, as well as solutions.

Basics of VoIP

With technology advances on the rise VoIP's (Voice over IP) popularity is ever increasing. VoIP or IP Telephony technology allows the delivery of voice information using the Internet Protocol (IP). Voice information is sent via a packet switched network instead of the traditional voice (circuit-switched) network protocols of the PSTN (Public Switched Telephone Network). For VoIP companies can build integrated networks for voice and data. To eliminate dependency on proprietorship adopting open standards provide muiltivendor interoperability. With their own anomalous characteristics, independent standards arose. We will be discussing the similarities and differences between the different protocols for VoIP.

IP Telephony or Voice over IP is the process of transmitting telephone calls over IP data networks instead of using "traditional" telephone networks. Telephone networks use a basic concept called Circuit Switching. Circuit Switching is the means of connecting communications between two points, in two directions – between the sender and receiver – this connection is called a circuit. Circuit Switching networks are the underpinning of the Public Switched Telephone Network (PSTN). VoIP does not utilize circuit switching to transmit data, but rather uses data networks to transmit the data. This technique is called Packet Switching. Packet switching does not keep the connection open like circuit switching. Packet switching allows the transmission of data in small blocks, called packets. The connection is only opened for the duration of the packet's transmission. Packet switching is an effective way to minimize the connection that is maintained between systems, resulting in not loading down the network. In packet switching technology users have the ability to share the same data path in a network. This type of communication method between sender and receiver is often referred to as connectionless. IP traffic over the internet mostly utilizes packet switching technology. The internet is considered basically to be a connectionless network. Another advantage of using packet switching

technology for VoIP is that packet switching allows the transmission of multiple calls to inhabit the same space used by only one call in a circuit switched network environment. To utilize the benefits of VoIP your company will need to install telephones and a digital private branch exchange (PBX). PBX is a switch used to connect phones/extensions to outside phones lines, as well as internally to each other. Gateways are also used. Gateways connect devices on different networks so they can communicate. Gateways translate circuit-switches signals into digital data that can be transmitted over a packet-switched network using the Internet protocol (IP). The PBX can also be an example of a Gateway.[1] IP is the language used by most data networks. Let's look at a brief description of how a classic telephone call works and compare it to a call using VoIP over a packet-switched network:

1. A User picks up the receiver of a telephone and waits for a dial tone.
2. Once a dial tone is established you have a connection to the local telephone carrier.
3. User dials the number of the person/place they would like to talk to.
4. The call is then routed through the switch at the local telephone carrier to the person/place you are calling.
5. A connection is established, opening the circuit between your telephone and the person/place's telephone line.
5. A conversation is established. Once finished, the User hangs up the telephone receiver.
6. The circuit is closed once the User hangs up the telephone. The line is now closed.

VoIP over a packet-switched network is established by the following:

1. User picks up the receiver sending a signal to the PBX
2. The PBX sends a dial tone once it picks up the signal. This allows the user to identify that they have established a connection with the PBX.
3. User dials the number of the person/place they want to talk to. The PBX will then temporarily store the number.
4. The PBX will verify that the number entered is valid.
5. The PBX establishes who to map the number to. The number is adjoined to the IP address of the IP Host. The IP host is normally another PBX that is connected to the phone system of the number you dialed. In most cases the IP host is the system you are trying to connect with.
6. The PBX establishes a connection between your PBX and the other party's IP host. As part of the session establishment each system needs to use the same communication protocol. Each system will be expecting packets of data from the other systems. The systems will have bi-directional channels open as part of the session.

[1] http://www.computer.howstuffworks.com/ip-telephony.htm/printable

7. Users establish communication and talk for a length of time. During communication you're PBX and the IP host of the party you are talking to transmit packets when there is data to be sent. Your PBX allows the circuit to stay open between itself and your phone while sending packets to and from the IP host.
8. Users complete the call and hang up the receiver.
9. Once you hang up the receiver, the circuit closes the connection between the PBX and your phone.
10. The PBX will then terminate the session of the other party by contacting the IP host of the party you called.
11. The PBX eliminates the number-to-IP-host mapping from memory once the session has been terminated.

If you currently have a phone and/or a computer you can talk to someone using VoIP without having to purchase additional equipment. The following bullets lists the methods used to talk to someone using VoIP:

1. **Computer-to-Computer –** This is the easiest and most effective way to use VoIP. Computer-to-computer connections allow endless benefits to include cost. Users won't have to pay for long-distance calls. There is software available that can be used for this method of VoIP. All that is needed is: an Internet connection, sound card, microphone, speakers, and software. Most of the software available is free or low-cost. An example of the software is MSN Explorer.

2. **Telephone-to-telephone –** Using gateways Users can connect with any telephone in the world. One of the benefits of using telephone-to-telephone technology is that there are discounted services offered by multiple companies. First the User must call into one of the companies' gateways. Once the user has called into the gateway/s they must enter the number of the party they wish to call. A connection will then be established through their IP-based network. Use of this method derives cost benefits due to the rates are much lower than standard long distance. However, some find a disadvantage in having to call a number first.

3. **Telephone-to-computer -** Another option companies are providing are calling cards or special numbers that permit users who are using a standard telephone to call a user with a computer. Again, there are cost benefits. The call cost is cheaper than conventional long-distance calls. The downfall is that the computer user must have installed, and have the vendors software currently running.

4. **Computer-to-telephone -** This method proves to be popular because it allows a user to call anyone from his/her computer. The only requirement is that the person they are calling has to have a phone. Again, software is required, yet typically free. Net2Phone offers competitive rates, as well as, free calls to anyone in the United States for the first five minutes. If the user's call runs over

the 'free' five minutes Net2Phone charges a rate of 3.9 cents per minute. Their international rates fluctuate depending on where you call, 3.9 cents to $7.52 per minute. Another advantage is Net2Phone is easy to administer and use.[2]

Understanding VoIP Standards and Protocols

VoIP encircles many standards and protocols. Basic fundamentals must be understood in order to comprehend the applications and usage of VoIP. The following definitions provide assistance in gaining a useful starting ground (the protocols are listed in alphabetical order):

- *H.248* is an ITU Recommendation that defines "Gateway Control Protocol." H.248 is the result of a joint collaboration between the ITU and the Internet Engineering Task Force (IETF). It is also referred to as IETF RFC 2885 (Megaco), which defines a centralized architecture for creating multimedia applications, including VoIP. In many ways, H.248 builds on and extends MGCP.

- *H.323* is an ITU Recommendation that defines "packet-based multimedia communications systems." In other words, H.323 defines a distributed architecture for creating multimedia applications, including VoIP.

- *IETF* refers to the Internet Engineering Task Force (http://www.ietf.org/), a community of engineers that seeks to determine how the Internet and Internet protocols work, as well as to define the prominent standards.

- *ITU* is the International Telecommunication Union (http://www.itu.int/home/index.html), an international organization within the United Nations System (http://www.unsystem.org/) where governments and the private sector coordinate global telecom networks and services.

- *Megaco*, also known as IETF RFC 2885 and ITU Recommendation H.248, defines a centralized architecture for creating multimedia applications, including VoIP.

- *Media Gateway Control Protocol (MGCP)*, also known as IETF RFC 2705, defines a centralized architecture for creating multimedia applications, including VoIP.

[2] http://www.computer.howstuffworks.com/ip-telephony.htm/printable

- *Real-Time Transport Protocol (RTP)*, also known as IETF RFC 1889, defines a transport protocol for real-time applications. Specifically, RTP provides the transport to carry the audio/media portion of VoIP communication. RTP is used by all the VoIP signaling protocols.

- *Session Initiation Protocol (SIP)*, also known as IETF RFC 2543, defines a distributed architecture for creating multimedia applications, including VoIP. [3]

For the purpose of this paper we will be looking at only four of the above mentioned standards. Since this paper's focus is Latency and QoS, we will be examining the H.323, Megaco/MGCP, and SIP standards and how they apply in making a VoIP infrastructure successful.

H.323

H.323 is considered an "umbrella" specification, covering multiple sub protocols related to call setup and signaling. The H.323 standard grew in the International Telecommunications Union (ITU) from videoconferencing applications. It has since evolved to meet the growing needs of VoIP networks. H.323 is currently the most commonly used VoIP signaling and call-control protocol. In its infancy, Version1, June 1996, the H.323 standard addressed communications over IP-based local area networks (LANS). Version 2, January 1998; outspread the protocol for wide area use and for general purpose IP networks. H.323 delineates all aspects of call transmission, call establishment and network resource availability. Several sub-protocols are defined by H.323; such as Registration, Admission, and Status (RAS) protocols for call routing, H.245 the "call control channel", and H.225 protocols for call set up. [4] Two protocols used, which define the essential requirements for transporting real-time data over a packet network are the Real-Time Control Protocol (RTCP) and/or the Real-Time Transport Protocol (RPT).

H.323 applies four major elements for a network-based communications system:

1. Gatekeepers
2. Mulitipoint Control Units (MCU)
3. Terminals
4. Gateways

[3] Cisco, p.2.
[4] Elachi, p.1.

Gatekeepers and Gateways help adjudicate the PSTN interconnection, while Multipoint Control Units (MCU) permits multiparty audio and videoconferences. Considerable issues in IP telephony (addressing problems) are handled by Gateways. Gateways enable the use of standard telephones to converse over the Internet instead of multimedia computers. In order to call another PC user, you must have their Internet Protocol (IP) address. Using a gateway product, you only need to dial their phone number. [5]

Gatekeepers provide address translation and bandwidth management by mapping IP addresses and telephone numbers. In order to have a successful IP network that will command all call traffic commencing and concluding at regular telephones, gatekeeper services are essential.

H.323 requires that TCP connections be used to transport the messages, resulting in an additional roundtrip exchange. The protocol uses multiple roundtrip messages to delegate signaling and control for calls between two terminals. There is intricacy and time involved in setting up a call. Recently version 3 has improved and incorporates a "Fast Connect" procedure that compounds the Q.931 messages exchanged between terminals and a tunneling procedure that allows H.245 to share a single TCP connection with Q.931. [6]

MGCP/Megaco

The Media Gateway Control Protocol (MGCP) and H.248/Megaco were designed to provide an architecture where specific communications elements and telephony gateways could be centrally added a VoIP network. In this instance, an architecture using these protocols closely coincides with the existing PSTN services. In its current form, MGCP is an aggregate of two earlier protocols, SGCP and IPDC. MGCP places call signaling control and processing intellect in call agents or media gateway controllers. Media gateways are telephony gateways that assist multi-service packet networks, translating data packets and audio signals. Just as a gatekeeper in H.323, the MGCP call agent performs the same call routing functions, but with constricted control. The gateways are counted on to execute commands forwarded by the call agents.

Megaco, known also as H.248, its ITU epithet, took the first major step towards accrediting the proposed Megaco/H.248 protocol standard in August 2000. Megaco RFC 3015 was developed by the IETF Megaco Working Group in alliance with ITU-T Study Group 16. [7] Megaco delivers the affiliation between the Media Gateway (MG) and the Media Gateway Controller (MGC) by conjoining the Call Control existence and the Media Processing (MG) entity. The Media Gateway (MG) adapts media provided by one type of network into the framework of another type of network. All the while MGC controls the call states that relate to connection control for media channels in MG. Like H.323, MGCP is a complementary protocol. The newer Megaco and MGCP are equipped as internal

[5] Elachi, P.2.
[6] Elachi, p.2.
[7] Elachi, p.3.

protocols for traffic between MGCs and MGs for corroded gateway architectures. The MGC utilizes the signaling layers of H.323 and acquaints itself as an H.323 gatekeeper or as an H.323 endpoint. Focusing on the audio signal translation, MGs converse the audio signals transported on telephone circuits and data packets funneled over the Internet or packet switched networks. The protocols MGCP and Megaco compliment both H.323 and SIP by aiding multimedia calls, multipoint, at the media level.

SIP

Session Initiation Protocol (SIP) is the Internet Engineering Task Force's (IETF) standard for multimedia conferencing over IP. [8] SIP began in late 1996 and was not until 1998 when the protocol was approved as an RFC by the IETF.

SIP then began to establish acceptance as an IP telephony protocol. SIP is an application-layer control protocol that can be used to create, prolong and abort calls between multiple endpoints. As defined in RFC 2543, SIP addresses the activity of session management within a packet telephony network, as well as, signaling. [9] Signaling provides the call information to be transported across network borders. Session administration assists the ability to execute the distribution of an end-to-end call. SIP utilizes the model; ask for and reply. The same organic model is used by Hypertext Transfer Protocol (HTTP), for instigating communication sessions between users. Unlike MGCP a SIP call can commence and conclude strictly between two clients, without the use of a call agent.

To instigate a session the caller sends a request to a callee's address, by sending a simple text command. The callee then responds with consent or refusal of the invitation. The call can be act as a go-between by a redirect server for routing purposes or by a proxy server. A number of capabilities are provided and supported by SIP, such as; address resolution, call redirection, name mapping. SIP can establish a session between the initial and destination endpoint – if the call is able to be completed, SIP initiates the session. SIP is multifunctional in that it can handle the addition of another endpoint joining the conference or the changing of media. Transfer and terminations of calls can be handled by SIP as well. SIP supports the transfer of calls from endpoint to endpoint with ease. The transferring party must specify the new endpoint so that SIP may initiate the session. SIP will then terminate the session between the transferring party and the transferee. Once the call is complete SIP will terminate the session between all call parties. [10] To accommodate telephony services a multitude of standards and protocols must join together. Depending on what standards a particular device supports results in interoperability. Most of the VoIP technologies put into practice, to date, are single vendor trials. The success of competing standards far out- way the devices which fail to support the standards themselves.

[8] Halpern, p.48.
[9] Halpern, p.48.
[10] Elachi, p.3.

Latency and its Effects

Unlike the data on IP networks that can tolerate delay-sensitivity well VoIP is vastly different. VoIP is latency-sensitivity driven. The definition of Latency is as follows:

- The time it takes for a packet to cross a network connection, from sender to receiver.
- The period of time that a frame is held by a network device before it is forwarded.[11]

There are ample problems within IP networks and the Internet specifically. Packets can get lost resulting in latency and poor service levels. While other "low-bandwidth" applications, such as email, may not suffer small delays equates to a major nuisance in VoIP.

250 milliseconds are considered by many to be the maximum acceptable latency allowable in a VoIP network. [12] In two-way phone conversations latency can be very undesirable. To achieve high quality voice the maximum desired one-way latency is 150ms if round trip delays exceed 250ms voice users will notice delays, and callers will start to talk over each other. Anything beyond 500ms deems the call impractical. With a delay over 500ms it would be like asking a question and not hearing the answer until hours later. Due to network congestion packet loss tends to be a key factor in lost voice signals, as well as, unreasonable delay, and jitter.

VoIP Latency Challenges

There are many challenges with implementing a VoIP solution instead of relying on the traditional PSTN system. The challenges include:

1. Degraded voice quality due to packet delay
2. Packet Loss
3. Poor echo cancellation
4. Distortion
5. Latency (Latency is the most bothersome because the entire network is involved from end to end.)

There are many components of latency, one of which is the jitter buffer. The jitter buffer is used to collect the packets at the gateway to the PSTN or from the receiving VoIP phone for the outbound voice signal. An advantage of utilizing the jitter buffer is that it may

[11] http://www.dictionary.com

[12] Lackey, p.2.

reduce packet loss by simply increasing the size of the buffer. However, it leads to extensive latency.

As previously discussed, VoIP technology can employ many methods of transmission. Voice calls can be made from a VoIP phone or a traditional phone on the PSTN.

Latency factors, including voice signals, that come in to play from the VoIP phone to the PSTN phone contain delays in many networks elements; VoIP phone, IP network routers or switches, wires, delays in the PSTN environment and the IP to PSTN gateway.

The signal, at the phone, must be sampled, encoded, and packaged as Real Tim Protocol (RTP) packets. Routers that come into contact with the IP network will contain input and output buffers.[13] The packets will come across additional buffers including the jitter buffer at the gateway. There will also be delays related with decoding and reassembly of the signal. Transmission delays will occur as well, due primarily to passage through wiring, and would result in a call between Los Angeles and New York to be roughly 20ms. Another latency factor to consider is transmission through the PSTN environment. This could entail breaking up the signal into frame relay or ATM packets and reconstructing them after transmission through an optical fiber. For this example, input and output buffers come across each other again. If the purpose is a VoIP phone on another network, the phone call must be converted back into RTP packets, including encoding and decoding, and buffer delays.

In regards to the PBX interface or PSTN there is very little delay at the gateway. Incoming analog signals are digitized to 64Kpbs Pulse Code Modulation (PCM), T1 or E1's are already digital signals. They are then passed on to the compression subsystem. If the compression is not executed on a committed DSP processor that is currently servicing the telephony interface impediments may occur in the buffering between the telephony subsystem and the compression subsystem. [14] DSP's that handle incoming PCM samples have the ability to also perform voice compression tasks without presenting buffering and data transfer latency.

Latencies and the IP network seem to go hand and hand. Both can be extremely volatile when dealing with IP telephony. IP has become a virtually worldwide protocol used for transmitting data. The expansion of the Internet has created endless configurations and routing potential. Thus, making it difficult to predict delays between gateways connected to one another over the Internet. Voice packets that transmit over the Internet have very little control in regards to the route taken by the packets between two gateways. This may induce long delays, high packet loss and erratic packet jitter (out of sequence packets). Let's discuss the ramifications of packet loss and packet delay.

Packet Loss and Packet Delay

[13] Skoog, P., Arnold, D., p.1.
[14] Mockingbird, p.3.

Packet Delay can wreak havoc in latency-sensitive applications such as VoIP. Packets transmitted over IP networks are passed through routers. Delays may occur in the data network dependent upon configuration, capacity, performance and load. It is known that each router introduces roughly 10ms of latency. There are many components that can alter this number. High volumes of data packets that arrive at the router concurrent with IP voice traffic. To minimize router latency for latency-sensitive applications, such as voice, networks may prioritize IP telephony ports over data ports. Later in this paper we will take a look at QoS and prioritizing latency-sensitive applications over the network.

Like packet delay through a router, packet loss may occur within a router. Packets have the likelihood of actually being lost in a router. Routers send out all of the incoming packets to the right outgoing port. Errors can, and will, occur if a router is overloaded. The router will then choose to abandon IP packets. In IP telephony packet loss is unacceptable. The performance of an IP call will suffer greatly if packet loss occurs. The quality of the conversation will lag if packet loss reaches more than 5%. Most gateways utilize vocoding algorithms to assist with packet loss. Let's take a look at the Vocoder algorithms out there today.

Vocoders

A vocoder is the device that translates analog and digital voice signals. Vocoders take analog voice signals convert them into digital signals and then translate them into speech sounds. As described by the ITU-T's G-series recommendations below are the most popular vocoders currently used for IP telephony:

- **G.711,** The G.711 algorithm encodes non-compressed speech streams running at 64 Kbps. (mockinbird) G.711 encoded voice is already in the correct format for digital voice delivery in the public phone network or through private branch exchanges (PBXs).

- **G.726,** G.726 describes ADPCM coding at 40,32,24 and 16kbps – ADPCM voice may also be interchanged between packet voice and public phone or PBX networks, provided that the latter has ADPCM capability.

- **G.728,** G.728 describes a 16-kbps low-delay variation of CELP voice compression – CELP voice coding must be transcoded to a public telephony format for delivery to or through telephone networks.

- **G.729,** G.729 describes CELP compression that enables voice to be coded into 8-kbps streams – Two variations of this standard (G.729 and G729 Annex A) differ largely in

computational complexity, and both generally provide speech quality as good as that of 32-kbps ADPCM.

- **G.723.1,** G.723.1 describes a compression technique that can be used for compressing speech or other audio signal components of multimedia service at a very low bit rate, as part of the overall H.324 family of standards – This coder has two bit rates associated with it – 5.3 and 6.3 kbps; the higher bit rate is based on MP-MLQ technology and has greater quality; the lower bit rate is based on CELP, gives good quality, and provides system designers with additional flexibility.[15]

Uncompressed speech data that is transmitted over the network expends a lot of bandwidth. Vocoders are used to compress speech before it is transmitted. A vocoder will then decompress the speech when it reaches its destination. In order to complete the process of compressing speech and decompressing speech when it reaches its destination the vocoder must buffer the data to assess speech fragments. The vocoder can impart a small delay while it "looks ahead" in the course of its mathematical computations.

This is a minimal delay, often 15 to 45 ms for typical vocoders that look ahead. Vocoder delay, as well as, buffering delay is frequently called algorithmic delay. During computations to compress speech for transport the vocoder can present further delay. The computer processor running the vocoder is where the compuations take place. Depending on the processor, additional delays can occur. The system delay plus the actual time required to calculate the speech is called compression delay. Gateways are multifaceted to serve an array of purposes, one of which is to manage compression delay. Vocoders estimate the analog waveform of speech. When the speech is reconstructed from its compressed format there may be a decline in the quality of the speech.

There is a method to measure the quality of the vocoders that are wildly used today. It is called the Mean Opinion Score or MOS. The MOS functions on a scale type measuring system; from low to high (0 being low and 5 being the highest). The MOS measures the bit rate and sample size for the VoIP vocoders. There are VoIP algorithms that operate a much lower bit rate than others, therefore causing a greater delay. This being said, it takes longer to for an algorithm with a low bit rate to encode speech than the algorithms that run at a much higher bit rate. For optimum performance it is feasible for the application and network to strike a balance between latency, voice quality and bit rate.

QOS Introduction

[15] Halpern, p.50.

Quality of Service (QoS) refers to the capability of a network to provide managed bandwidth and better service to preferential network traffic.
QoS can provide priority included, but not limited too, controlling jitter; latency and drop precedence. QoS also provides dedicated bandwidth. QoS technology is widely used for VoIP as it enables you to allocate priority service to voice. Benefits included in QoS are ensuring delay sensitive/mission critical applications are not compromised by other applications. QOS technology for voice also provides reliability and predictability. QOS ensures voice packet delivery without packet loss. Another feature of QOS technology eliminates poor quality voice transmission, together with missing syllables and crackles that relinquish the call unsatisfactory.

Without QOS latency-sensitive applications, such as VoIP, are unable to be supported by IP. Where there are benefits of IP-based QOS there are also deployment challenges. Due to voice traffic being time sensitive it requires higher QOS guarantees than data traffic. Email typically requires low bandwidth and is not considered mission critical. However, voice traffic is latency-sensitivity driven and requires special care. In the following pages we will be discussing the technical challenges and solutions involved with a QOS implementation.

Technical Challenges of QoS Implementation

There are multiple issues that may well affect the QOS distribution for network based integrated services. These issues result in different traffic types requiring different services levels. Voice relies on UDP/RPT as its transport and UDP is not reliable.

Delays in voice traffic create gaps in the transmission that may be heard by the receiver, resulting in unhappy customers. QOS technology features concrete priority service to voice traffic to establish predictable delivery. Usually small in size, transmission of voice packets range from 80 to 256 bytes. Unless QOS techniques are used such voice packets can be delayed between larger data packets. QOS techniques used are packet fragmentation and interleaving. One of the crucial technical issues with QOS is that in order to be effective it must be supported end-to end. For VoIP to be of functional quality the network should essentially have a bare minimum data rate and bounded delay variation.

If QoS mechanisms are supported on only portions of the network there are no guarantees that the traffic will get the handling end-to-end that is necessary to achieve success. The concept of an end-to-end QoS implementation can get convoluted in the case of “any-to-any” communications (not known in advance, any user to any user). [16] In the data communications of “any-to-any”, many service providers can get involved. In some cases it may become necessary to have wide-area carriers, voice switching providers, application service providers (ASPs), storage are network providers (SANs), and others. QoS

[16] http://www.riverstonenet.com/pdf/qos.pdf

implementation depends largely on the coordinated efforts of multiple parties. In order for end-to-end QoS to be functional the information needs to be communicated to each provider so that the QoS along the transmission route can be honored in an unfailing way. There have been a number of IP based QoS standards defined to ensure QoS interoperability. This provides flexibility in achieving specific sets of QoS parameters intertwined to accomplish the particular handling of different types of traffic. The down-fall of these standards is that they are inclined to be used within an Intradomain only (single network domain). , resulting in non standardized service classes for the whole network. The QoS technologies offered today are a means to manage jitter and delay for voice.

QoS Mechanisms: IntServ, DiffServ and RSVP

Integrated Services/IntServ

Integrated Services or IntServ is protocol architecture designed to provide QoS over the internet. It was developed by the IETF, as documented in RFC1633, RFC2212, and RFC2215. [17] The IETF's efforts to identify IP-based Qos mechanisms began in the early 1990's. IntServ was designed to provide unambiguous, end-to-end QoS for a sequence of packets that have the same source and destination. There is a growing need for QoS in each network environment, as well as the need for applications to specify their Qos requirements. IntServ works by allowing hosts to request per-flow, resources end-to-end along a data path. [18] The network will then provide feedback to the hosts notifying them whether or not the network can assist with such requests. During the development of IntServ the focus primarily was on three control mechanism functions:

1. Packet classification
2. Scheduling
3. Admission control

These control mechanisms define many of the essential principles on which IP-based QoS is assembled. IntServ permits flexibility in regards to how packets are classified. However, packets get the same treatment once classified. This is done by the packet scheduler. The packet scheduler also administers packet forwarding using queues, timers and other mechanisms. Another control mechanism is admission control. Admission control is necessary to make certain that new flows would be established QoS handling without impacting the existing traffic flows on the network. Admission control is the decision about resource availability. To insure that the commitments to certain classes of packets are met, appropriate resources needs to be requested. This needs to take place

[17] URL:ftp://ftp.rfc-editor.org/in-notes/rfc2215.txt, URL:ftp://ftp.rfc-editor.org/in-notes/rfc2212.txt, URL:ftp://ftp.rfc-editor.org/in-notes/rfc2215.txt

[18] Riverstone, p.5.

so that if the resources are not available the request can be refused. Under IntServ, applications utilize a signaling protocol to communicate their QoS requirements to the network.

The signaling protocol is also accountable for installing and maintaining a QoS control state in the network. This provides the ability the network needs in order to communicate management information back to the application.
There are multiple signaling protocols available; one such signaling protocol the IETF defined Resource Reservation Protocol (RSVP). We will be discussing RSVP in greater detail later in this paper. IntServ developers developed control and signaling mechanisms for QoS, as well as QoS services: controlled load and guaranteed services. Controlled load and guaranteed services are the only QoS services defined by the IETF.

Their provisions indicate the mechanisms that networks must support in order to provide QoS. Let's take a look at the differences between controlled load and guaranteed services:

- **Controlled load service -** Controlled load service is intended for applications that require a large quantity of their packets to be delivered with minimal delay. Examples of these applications are: non-real time audio and video. Controlled load service was designed to offer "best-effort" service on a network that is unloaded or lightly loaded.

- **Guaranteed service -** Guaranteed service controls the maximum queuing delay to ensure that packets arrive within a specified delivery time. Examples of these applications are video and interactive voice, remote control, financial transactions, and other delay-sensitive applications.

The IntServ model has scalability concerns but has recently been revitalized. RSVP has been revitalized to use with MLPS. We will now take a look at the RSVP protocol.

RSVP

RSVP is a signaling protocol used to identify the communications between senders and receivers. The RSVP protocol intervenes with the network fundamentals for QoS flows. Users who wish to communicate with each other communicate via RSPV messages. Each message contains a header and a set of objects.

The RSVP protocol can be used with IntServ, thus rejuvenating the IntServ protocol that had loss its popularity in recent years. If the RSVP protocol is used in conjunction with IntServ each RSVP messages identify resources that have been requested to the network using IntServ constraints. RSVP may also be used with MLPS therefore changing the scope from its use with the IntServ protocol. If used with MLPS, RSVP transports label binding and precise routing information. [19] The QoS mechanism protocol RSVP is a

[19] URL:http://www.riverstonenet.com/pdf/qos.pdf

receiver-based protocol. Receiver-based protocols are services that are requested by the receiver, replying to messages received from the sender. To initiate a RSVP session the sender sends a PATH message. A PATH's message function is to locate a path through the network for a specific data flow allowing the route to be bound for use by RSVP.

In regards to QoS, the host that is receiving sends a reservation-request (RESV) message indicating what service class it is able to support. The reservation messages travel through the network in a reverse path until the sending host receives them. The RSVP protocol makes an effort to make a resource reservation at each network component from which the application flow will pass through. To be aware of QoS requirements particular in each RSVP request the RSVP-enabled routers, as well as, other network components need to maintain "state" on the individual RSVP flows. RSVP institutes a temporary, or soft, state on all RSVP devices between sending and receiving nodes. The protocol will also send intermittent messages to refresh state. State will be deleted if there are no refresh messages received within an allotted timeframe.

There are pro's and con's of using and relying on soft state. The benefits of relying on soft state include:

1. The ability for RSVP to acclimatize to network changes.
2. RSVP can modify the path amid receivers and senders in reply to routing changes, by simply checking and refreshing state.

The disadvantages of RSVP include:

1. High overhead involved with RSVP's per-flow direction.
2. Limited scalability.
3. Per flow reservation computation.
4. Memory resources in each router.
5. Extensive volume of message exchange.

To be able to support traffic flows and use for tunneling the IETF has set forth to decrease RSVP's shortcomings. As previously stated, RSVP is used in the MPLS environment to transport precise routing and label binding information. RSVP within the MPLS environment transports routing and label binding information instead of QoS information. Network operators have options when using RSVP to set up precise routed Label Switched Paths (LSP's), or tunnels. They have the capability to allot resources along the data path by means of RSVP reservations and IntServe service classes.[20] RSVP allows tunneling support and has been improved to provide a single reservation the ability to combine other RSVP reservations. Combined reservations may include packets from a large number of flows that require related treatment and/or the same traffic class. There is a "new kid in town" in regards to VoIP QoS protocols. The IETF have been looking towards using Differentiated Services, also known as DiffServ, to specify which traffic can be gathered together.

[20] URL:http://www.riverstonenet.com/pdf/qos.pdf

DiffServ has the ability to determine which traffic flow with the same DiffServ marking could be managed by an aggregated reservation.

Differentiated Services/DiffServ

Differentiated Services or DiffServ is a QoS mechanism that differs greatly to the IntServ and RSVP QoS mechanisms. IntServ and RSVP are alike in the fact that they both have an ambitious approach to QoS, using an individual flow-based reservation model.

DiffServ has a much simpler approach towards QoS, making it more scaleable than other QoS mechanisms. DiffServ permits network traffic to be broken up into small flows for appropriate marking. The flows are marked so that the devices on the network amongst the traffic path will be able to identify and provide the appropriate treatment. There are different functions within the DiffServ architecture. They are broken out into two structures; edge functions and network core.

The functions that are more complex are handled by hosts or routers, also known has edge functions. An example of a complex function would be traffic conditioning. Functions that are less complex are handled by devices in the interior network. An example of a less complex function would be packet queuing. DiffServ calls for network managers to determine the amount bandwidth or latency needed for a class, or classes of networks. They will need to configure there routers to accommodate the request.

There is a field created by the DiffServ working group to assist with packet marking. The defined field is the DS field. The information contained in the DS field is the six most important bits in the IPv4 Type of Service (TOS) octet and the IPv6 Traffic Class octet. DiffServ currently does not utilize the two least significant bits. [21] DiffServ Code Points or DSCP's are the values found in the DS field. The DSCP's determine how packets are taken care of as they are transmitted across the network. The hosts or routers at the edge of a network have the responsibility of determining how packet classification and DSCP marking is handled. As previously mentioned hosts or routers can also be called edge nodes or edge devices. The edge devices have multiple functions. They are responsible for collecting individual flows and combining them into macro-flows and shaping and dropping traffic so it corresponds with the traffic profile.

The edge node or device is also tasked with determining once the traffic coincides. Once that has been established the edge node will mark it properly and disclose it to the DiffServ portion of the network. The packets are marked by queuing the packets properly. This service is handled by the routers at each hop alongside the path toward the destination providing access to reserved bandwidth and executing QoS functions. Per Hop Behavior or PHB is the particular treatment a router grants a packet with a specific DiffServ Code Point.

[21] URL:http://www.riverstonenet.com/pdf/qos.pdf

PHB identifies how the packet should be treated at each hop. PHBs. There have been multiple standardized PHB's created to date. There are currently three PHBs. They are:

1. **Default PHB:** Best effort forwarding

2. **Expedited Forwarding:** Expedited forwarding may also be known as EF. Expedited forwarding guarantees that each DiffServ node presents low-delay, low-jitter and low-loss. Expedited forwarding also allows a certain amount of traffic to enter the network. Once it has entered the network it will see to it that the traffic flows obtain minimal queuing. It will then drop off the additional traffic at the network ingress.

3. **Assured Forwarding:** Assured forwarding may also be known as AF. Assured forwarding grants lower-level functions compared to EF. AF provides resources based on four classes of traffic.

Depending on the resources, each specific traffic class will acquire a certain type of handling that is allotted to it and drop precedence. Examples of resources may be a buffer space and bandwidth. Per-Domain Behavior or PDB was specified by the DiffServ working group to assist end-to-end QoS in a given network. A PDB identifies a specific assortment of packets obtains edge to edge as they traverse the DiffServ domain. For each PDB there are measurable characteristics that originate from a precise forwarding treatment (PHB), traffic classification, traffic loading and a domain's topology. [22] DiffServ and PDBs' components were intended to assist network operators in characterizing differentiated services. One of the pros of QoS based services is that it allows flexibility for network managers to combine tools and/or created their own. There are many advantages of using DiffServ, such as scalability and its capability to handle collective traffic. DiffServ is much more practical to implement than IntServ.

Conclusion

There are many factors to consider when converging voice and data infrastructures. As we have discussed Latency and QoS are only a few technical challenges of a VoIP implementation. Like any other new-found technology VoIP will require a learning curve. Taking proper and careful consideration of how one implements the technology will be necessary. Assurance of QoS, low latency, low jitter and high availability is essential in the success of an integrated voice and data network. The capability to have an adaptable mechanism that will enable users to customize the QoS policy to their precise needs is an essential component of a successful, overall integrated network. The goal of any corporation implementing VoIP should be to utilize a variety of techniques to optimize scalability, availability, and manageability. Optimization provides your critical business with the high bandwidth, low delay, and controlled jitter required to be successful.

[22] URL:http://www.riverstonenet.com/pdf/qos.pdf

Bibliography

1. Braden, R., Clark, D., Shenker, S. "Integrated Services in the Internet Architecture: an Overview." June 1994. URL:ftp://ftp.rfc-editor.org/in-notes/rfc1633.txt (14 March 2003).

2. Elachi, Joanna. "Standards Snapshot: The State Of The Big 3 in VoIP Signaling, Protocols" 27 November 2002. URL:http://www.commweb.com/article/COM20001127S0008 (13 Feb 2003).

3. "End-to-End QoS: Challenges and Practical Solutions." URL:http://www.riverstonenet.com/pdf/qos.pdf

4. Halpern, Jason. "SAFE: IP Telephony Security in Depth." 30 July 2002. URL:http://www.cisco.com/warp/public/cc/so/cuso/epso/sqfr/safip_wp.pdf (12 Feb 2003).

5. Handley, M., Schulzrinne, H., Schooler, E., Rosenberg, J.. "SIP: Session Initiation Protocol." March 1999. URL:ftp://ftp.rfc-editor.org/in-notes/rfc2543.txt (13 Feb 2002).

6. Held, Gilbert. "Emerging Technology: Reducing Voice over IP Latency." 10 July 2000. URL:http://www.networkmagazine.com/article/NMG20000710S0012 (21 Feb 2003).

7. Kuhn, Dr. Robert M.. "Understanding Voice Over IP." URL:http://www.compassconsulting.com/articles/voipintro.html (6 Feb 2003).

8. Lackey, Jason. "VoIP – Building a sound Foundation for Voice over IP." 21 November 2002. URL:http://www.riverstonenet.com/solutions/voip.shtml (13 Feb 2002).

9. Shenker, S., Partridge, C., Guerin, R. "Specification of Guaranteed Quality of Service." September 1997. URL:ftp://ftp.rfc-editor.org/in-notes/rfc2212.txt (14 March 2003).

10. Shenker, S., Wroclawski, J.. "General Characterization Parameters for Integrated Service Network Elements." September 1997. URL:ftp://ftp.rfc-editor.org/in-notes/rfc2215.txt (14 March 2003).

11. Skoog, Paul., Arnold, Doug. "Synchronization Essentials of VoIP." URL:http://www.truetime.com/DOCSn/Sync_VoIP.pdf (13 March 2003).

12. Stringfellow, Brian. "Secure Voice Over IP." 15 August 2001.
URL:http://www.sans.org//rr/voip/sec voice.php (13 Feb 2003).

13. "Transporting Voice over IP." V3.1.
URL:http://www.mbird.com/pdfs/wp transportvoice.pdf (13 March 2003).

14. Tyson, Jeff. "How IP Telephony Works."
URL:http://computer.howstuffworks.com/ip-telephony.htm (19 March 2003)

15. "Using QoS to Optimize Voice Quality in VoIP Networks." 14 November 2002.
URL:http://www.cisco.com/warp/public/779/smbiz/community/qos voip.html (14 March 2003).

16. Weiss, Eric. "Security concerns with VoIP." 20 August 2001.
URL:http://www.sans.org/rr/voip/sec concerns.php (19 Feb 2003).

Last Updated: November 27th, 2020

Upcoming SANS Training

Click here to view a list of all SANS Courses

SANS Essentials Australia 2021	**Melbourne, AU**	**Feb 15, 2021 - Feb 20, 2021**	**Live Event**
SANS OnDemand	**OnlineUS**	**Anytime**	**Self Paced**
SANS SelfStudy	**Books & MP3s OnlyUS**	**Anytime**	**Self Paced**