

What VOIP Requires From a Data Network

Introduction

Here is a very common story. A customer has a data network based on TCP/IP that is working well. He is able to transfer files and run networked database applications among several locations with adequate speed and efficiency. He is able to browse the web and download files from the Internet.

He installs VOIP devices at each of his sites and connects them to the data network. The voice technology works very poorly. He has intermittent problems with sketchy voice quality, dropouts in conversation, even dropped calls and calls that don't complete. His sturdy data network that works fine for all kinds of data applications barely works at all for voice applications.

The difficulty lies in the fact that voice applications require the network to provide some features that are not very important to data applications. File downloads and database programs require every byte to be delivered correctly, but they are flexible with regard to how long it takes to get the bytes from one location to another. Voice, on the other hand, requires the bytes to arrive in a very timely manner, although it is more flexible about losing a few bytes here and there.

In this document, we will address the issues that cause this situation and describe what is necessary to avoid it. These issues are network quality, available bandwidth, and packet competition. We will try to provide the reader with an understanding of what should be considered in planning a VOIP installation so that there will not be any unpleasant surprises.

Network Quality

The first thing that VOIP requires of a network has to do with its basic ability to deliver most packets on time. In general, TCP/IP networks do not guarantee that every packet sent will be delivered. Routers along the way have the option of dropping packets, if necessary, so some packets that are sent will never arrive at their destination. Since each packet can take a different route from the source device to the destination device, they can arrive out of order and take a different amount of time to travel the same distance. Finally, the total time it takes a packet to go from one location to the other may be very large; too large to allow voice conversations to progress normally. We will address each of these issues with regard to what is required for voice traffic.

Packet Loss

One aspect of network quality is packet loss. This is a quantity that can be measured by the network analysis tools that we will discuss later. The quantity is the percentage of packets that are sent from one end of the network connection that do not reach the other end. Networks with a packet loss of more than 3% are not good candidates for VOIP, as there will be dropouts in the audio.

Packet loss increases sharply at the point where the network is overloaded with traffic. For this reason, packet loss testing must be done in conjunction with bandwidth and QOS testing. On a lightly-loaded network, packet loss may be low, but it may become unacceptably high when the number of packets reaches the maximum that the network can accommodate.

Jitter

Another network quality issue to examine is jitter. Each packet of voice information takes a different amount of time to go from one end of the network to the other. This variation is called “jitter”. The VOIP equipment on the receiving end is responsible for putting the packets into a buffer so that they can be played out as an unbroken stream of audio.

The buffer that is used for the purpose is called a “jitter buffer” and it is a certain length in milliseconds. This length is called the “jitter buffer depth”. This depth should be about twice the size of the largest jitter value that actually occurs on the network.

In networks that display jitter values that are larger than 50 milliseconds, it is difficult or impossible to play the packets smoothly using a jitter buffer of reasonable depth. In these poorly-behaved networks, the receiving device will reset its jitter buffer frequently, leading to noticeable dropouts in audio.

It is important to set the jitter buffer depth in the VOIP devices to match the behavior of the network. If you set it too low, you will hear dropouts in the audio. If it is too high, there will be an unnecessary delay in the audio. Dropouts can also be caused by packet loss and QOS problems, so you cannot assume that the jitter buffer is too small just because there are dropouts.

Latency

The last aspect of network quality that we will examine is latency. This refers to the amount of time it takes a packet to get from one end of the network to the other. If this is long (more than 200 milliseconds), it can create problems for the VOIP equipment that lead to an echo in the audio. If it is very long (more than 400 milliseconds round-trip), then it will interfere with human conversations.

The echo problem only occurs in certain types of VOIP equipment, where audio is echoed back to the sender with some delay. When this delay is small (less than 200 milliseconds) there is an algorithm in the sending device that recognizes the echo and removes it. When the delay is very small (less than 28 ms), no echo handling is necessary, since a person will not notice the echo. When the delay is large (more than the size of the echo canceller’s buffer), the sending device will have forgotten that it sent that sound so it will not be able to recognize it as an echo.

When a connection is all-digital to a digital phone at the receiving end, then echo is not always an issue.

The problem in conversation, however, is always present when the round-trip latency (RTL) is more than 400 milliseconds or so. One person will start to say something, but before the sound of his words gets through the network, the other person will start to say

something. When they each hear the “interruption” of the other person speaking, they will pause to listen. When they hear that the other person has paused, they will try again to say something. This condition can be very frustrating, and will lead to half-duplex conversations in severe case. A half-duplex conversation is where each party will say “OVER” when they are done speaking.

Available Bandwidth

In order to use a network for VOIP, there must be enough bandwidth available to carry the voice packets. This available bandwidth is not the same thing as total (raw) bandwidth. It is a measurable amount of voice traffic that can be transported by the chain of routers and switches that make up the data network. This measurement is done by comparing the amount of voice data that we need to the amount of voice data that the network can carry. There are several subjects that you should understand to determine available bandwidth. These include raw bandwidth, bottlenecks, streaming UDP, and codecs.

Raw Bandwidth

Most customers who buy network services are familiar with the raw bandwidth of each of their connections. If they have a full T1 devoted to data networking, then they have about 1.5Mbits of raw bandwidth, half of a T1 is 768Kbits, etc. DSL is sometimes trickier, because raw bandwidth in one direction is often different than in the other. You might have 768Kbits of download capability with only 128Kbits of upload. Point-to-point radio, microwave, dial-up, and other “last-mile” technologies all have a specified raw bandwidth value that your provider can tell you.

What they cannot tell you is how much of the bandwidth will be available for your voice applications to use. Each type of connection loses some of its bandwidth due to the header information that is included for each packet. Some networks lose so many packets when you approach their maximum utilization that they become unusable for voice applications at that rate.

Bottlenecks

It may seem obvious that the connection between point A and point B that has the smallest available bandwidth will determine the available bandwidth for the connections between A and B. If one VOIP device is at the headquarters with a T1 connection to the network, but the other end of the conversation is at someone’s house with a 53Kbps dialup connection, then the total available bandwidth will be less than 53Kbps.

Streaming UDP

Some network technologies impose a very large per-packet price on network traffic. On a 768Kbps point-to-point radio connection, for instance, you may be able to download files at 600Kbps using large TCP packets. Voice, however, uses many very small UDP packets to transport its data. This has to be done to reduce the total delay from the time

one party says something and the time the other party hears it. Consequently, on the same 768Kbps radio connection you may only be able to carry 180Kbps of voice traffic. The actual bandwidth available for voice can only be determined accurately by measurement.

Codecs

Before you know if your network can carry the voice data, you will have to determine how much voice data you will have. You can get an estimate of this value by taking the number of calls that will use a network segment and multiplying it by the amount of traffic per call. This traffic-per-call value varies based on the codec that you choose. A codec is a method for encoding the audio as data. The G.711 codec will give toll-quality voice connections, but it uses about 80Kbps of available bandwidth per call. G.729 gives near-toll-quality voice, but it only uses 26Kbps per call.

The amount of available bandwidth per call also varies based on the number of milliseconds of audio that is included in each packet. Here is a table that shows some example bandwidth values:

| Codec | ms/packet | Bandwidth (Kbps) |
|-------|-----------|------------------|
| G.711 | 20 | 83 |
| G.711 | 30 | 76 |
| G.729 | 20 | 26.4 |
| G.729 | 40 | 17.2 |
| G.723 | 30 | 18.4 |
| G.723 | 60 | 12.3 |

Packet Competition

Once it has been determined that the network has low packet loss, low jitter, low latency, and sufficient available bandwidth, we arrive at the most complicated and troubling issue. We intend to use the same network to transport data and voice packets. When there are a lot of data packets to send (as in a file download), they will be put onto the network ahead of voice packets and they will interfere with voice conversations.

Congestion and FIFO Queuing

Routers are often connected between a fast network and a slow one. One of the router's jobs is to keep the slow side as busy as possible, to maximize the amount of data that travels through the network. It accomplishes this by queuing. Packets that arrive from

the fast network bound for the slow one are held in a queue and put on the slow network as fast as the slow network can take them.

This is like pouring syrup from a bucket into a bottle using a funnel. The bottle is the slow network, the funnel is the queue, and the bucket is the fast network. The funnel allows syrup to arrive from the bucket faster than the bottle will accept it.

TCP traffic works well in this routing system. A file download, for instance, flows out of the bucket into the funnel until it notices that the funnel is full and the syrup is running onto the floor (dropped packets). Then, the TCP source machine will automatically slow down the flow of packets to match the speed at which the network can deliver them to their final destination.

If there are two TCP streams going to two destinations, the analogy becomes a little forced. Imagine I have a funnel that feeds into a bean sorting/sacking machine. There are two people with buckets, one has a bucket of beans and one has a bucket of peas. They both pour from their buckets into the same funnel at the same time. With a First-in-First-out (FIFO) queue, the funnel drops bean, bean, pea, bean, pea, pea, bean, pea, etc. in the order that they arrive at the bottom of the funnel into the sorting machine. The machine then sorts beans and peas into their correct destination sacks.

When the funnel overflows, both men pour more slowly based on a complicated algorithm that involves watching how much of their product is getting through the machine and trying to match the speed.

Since this router only has a single queue (funnel), the 50 pound sack of beans will fill up faster, since each bean is bigger than a pea. This approach favors protocols that put more bytes in each packet.

The entire basis for TCP/IP routing has been developed around the principle that the men can pour beans at different rates and that the object is to fill the 50 pound sacks as quickly as possible.

Streaming UDP voice is not like that. The packets always flow at the same small rate. If they have to stop and wait for large packets that are ahead of them in a big queue, they will arrive too late to be played as audio. Voice packets that are late are useless. Throwing this trickle of very perishable packets into the funnel on top of the beans and peas causes the VOIP conversation to be broken up by dropouts, because they have to wait for all of the beans and peas to be sorted ahead of the voice packets before they can be delivered to the destination device.

Weighted Fair Queuing

To improve the queuing of packets in the router, a more advanced queuing scheme was developed. It is called “Weighted Fair Queuing” (WFQ), and it is an attempt to fix the problem where the beans are favored over the peas. In this system, the router tries to spot the different conversations that are going on and give each one its own queue. Then it takes packets from the queue in a way that will allow the number of bytes that come from each queue to be the same.

With this system in place, the man with the beans pours into the bean funnel and the man with the peas pours into a funnel that is just for peas. The machine takes more peas than

it does beans in a ratio that will fill 50 pound sacks of peas and beans at the same rate. In a router with two big file downloads and an interactive database session, each of the three conversations will be allowed to use about one third of the total bandwidth.

This queuing scheme works better for voice than the single FIFO queue, but it will still fail in many cases. Here is a simple example:

A branch office is connected to the corporate network by a DSL connection that provides 256Kbps of bandwidth in each direction.

There are two voice conversations using the G.711 codec in progress, using 83Kbps of bandwidth each.

There are also two file transfers in progress, due to web surfing.

Since there are four conversations, WFQ will give one fourth of the bandwidth to each, or about 64Kbps.

Without the necessary 83Kbps of throughput, both calls will suffer from dropouts, and the VOIP equipment cannot compensate.

Other Queuing Schemes

Sophisticated routers have several other schemes for queuing packets and delivering them from a fast network to a slower one. These can often be combined with QOS features of the router to achieve exactly what is needed to make the voice traffic work well in spite of simultaneous data traffic through the same link.

The important thing to remember is that routers usually default to a single FIFO queue or to the Weighted Fair Queuing strategy. This will usually result in poor voice quality due to packet competition. It is necessary to understand and implement QOS features in all of your network routers to correct this problem.

Quality of Service Configuration

Routers must be configured to treat voice packets in a special way, or the voice traffic will lose in the competition with the data packets. This special configuration is called “Quality of Service” and it will mean that VOIP conversations will work well even when there is a large amount of data traffic.

There are four broad categories of QOS configured networks: Best-Effort, Differentiated Service, Dedicated Service, and Guaranteed Service. These provide different levels of complexity and efficacy in the handling of voice traffic on the network when that voice traffic is in competition with data traffic.

Best-Effort QOS

This level of service is sometimes called “Lack of QOS”. Most network routers are configured in this manner by default. They use a queuing scheme that is appropriate for data traffic, and make no allowances for the special needs of voice traffic.

This is the network that most companies start out with. When there is no data traffic, or data traffic is very light, then the voice packets may get through on time and the voice

audio may sound fine. No one can depend on this network for toll-quality voice, because a large file download can start at any time and disrupt the voice.

Differentiated Service

The first step in solving the problem is to give the switches and routers in the network a way to differentiate between voice and data. Once the routers can classify the traffic, then they can be configured to route the packets using different schemes for different traffic types.

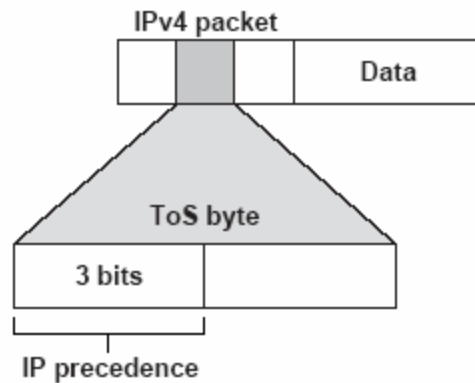
Traffic Classification

There is a spot in each packet that can be used to describe the type of data that the packet contains and how the packet should be handled by all of the switches and routers in the network.

The spot is called the TOS field. This is a byte that contains 8 bits. The first three bits are used for IP precedence, the next four are used for Type of Service, and the last bit is unused.

It is helpful to note that the four Type of Service bits are almost never used in real life. Only the three IP Precedence bits at the beginning of the Type of Service byte are commonly used. This leads to a great deal of confusion in terminology, because the Type of Service bits in the Type of Service byte are not used, but the IP Precedence Bits in the Type of Service byte are used. To make matters worse, the IP Precedence bits are sometimes called Class of Service, and Type of Service field is also called Diffserv Byte.

Figure 1 - IP Precedence Bits



Since the IP Precedence value is the first three bits of the TOS byte, some binary arithmetic is necessary to arrive at the correct TOS values to set each precedence value.

| TOS Byte | IP Precedence |
|----------|--------------------|
| 0xE0 | 7 Network Control |
| 0xC0 | 6 Internet Control |

| | |
|------|------------------|
| 0xA0 | 5 Critical / CP |
| 0x80 | 4 Flash Override |
| 0x60 | 3 Flash |
| 0x40 | 2 Immediate |
| 0x20 | 1 Priority |
| 0x00 | 0 Routine |

Packets are usually marked by the device that creates them. The place where the voice network connects to the data network will usually have a configuration item for TOS byte that can be set from the table above. 0x80 is a good choice for this setting. Values of IP Precedence that are greater than 4 are used by the routers and switches for special purposes, and they should be avoided.

Alternatively, the packets can be marked by one of the switches or routers in the edge network based on which physical port they came from, which VLAN they are from, what IP address they are bound for, etc. Routers and switches vary widely in their ability to do this type of marking.

Differential Treatment

Once the packets have been marked, routers should be configured to give them special handling. Low Latency Queuing (LLQ) is a good choice. This will move packets with a higher IP Precedence to the front of the queue for processing. Since voice traffic is a fixed rate, it should never overwhelm the capacity of a network.

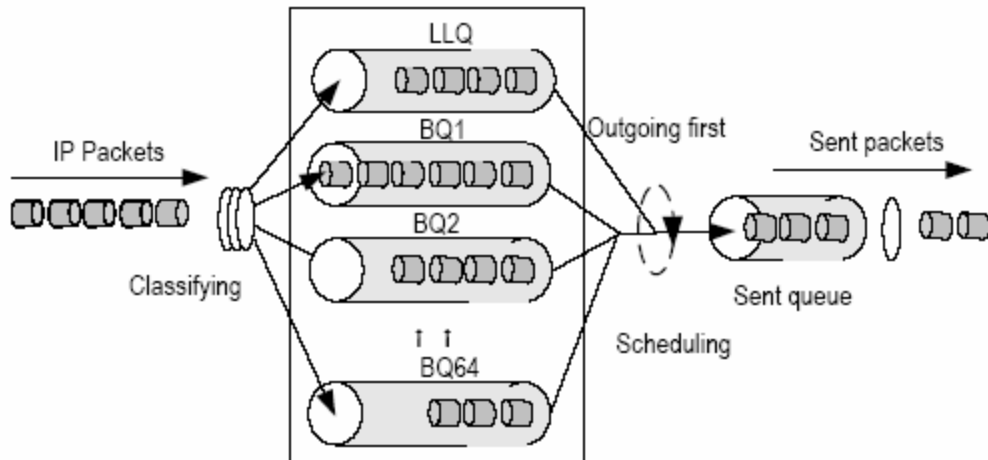
Capable routers offer lots of options for using the IP Precedence mark in their queuing algorithms. The details of these approaches are too complicated to go into here. It is sufficient to say that a competent network engineer should be able to configure each router in the network to deliver voice packets on time, regardless of data traffic.

Low Latency Queuing

The Low Latency Queuing (LLQ) scheme was designed for traffic like VOIP that must get to its destination on time. LLQ divides traffic into two categories, Normal and Low Latency. The Normal traffic is handled as data traffic, usually using WFQ. The Low Latency traffic is moved immediately to the head of the outgoing line so that it will experience the minimum possible delay.

At the same time, the Low Latency traffic is subject to an overall limit (sometimes called “policing”) which keeps it from becoming the only traffic on the link. Stifling the data traffic completely would be a bad thing that policing helps to avoid.

Figure 2 - Low Latency Queuing



IP RTP Priority Queuing

There is a special version of LLQ that is supported on Cisco routers called “IP RTP Priority”. It is more limited than LLQ, and is only useful for handling RTP voice traffic in a narrow range of network scenarios. It has the advantage of being easier to set up, but it has many disadvantages. Cisco has replaced it with the more general LLQ scheme in newer versions of its router software.

Dedicated Service

With dedicated service, routers are configured so that a certain bandwidth is permanently assigned to voice. This dedicated bandwidth is never used for data, so it is always available for voice packets. Sometimes this is done by marking voice packets for the router to see, but it is usually done by having a separate port on the router that is used for connection to the VOIP equipment.

When dedicated service is set up properly, it eliminates the problem of data traffic interfering with voice traffic on the same network. Like Differentiated Service, it is complicated to configure and has an effect on all of the switches and routers in the network.

The primary disadvantage of Dedicated Service is that a certain amount of your network bandwidth will sit idle when there is no voice traffic. This can seem wasteful to some people, who would rather be able to use that bandwidth for file downloads and database applications when there is no one on the phone. For these people, Differentiated Service with Priority Queuing might be a better solution.

However, the wasted bandwidth is well worth its cost if it provides you with the ability to depend on high-quality VOIP service at any time.

Guaranteed Service

Guaranteed service is the most complex approach to solving VOIP packet competition problems. It is sometimes called “Integrated Services Architecture” or “IntServ”. It requires the most expensive network equipment and the most talented network engineers to configure it.

Guaranteed service works by having each VOIP device make a “reservation” of the bandwidth needed for a call using the RSVP protocol. This reservation must be honored by each router that lies in the path between the VOIP source and destination devices.

Each router sets up a dedicated bandwidth for the call, but the dedication is temporary and only lasts until the call is finished. Then the bandwidth is again available for other calls or for data traffic.

Although Guaranteed Service is expensive in terms of equipment and trained personnel to configure it, it does not provide a significant advantage over Differential Service or Dedicated Service. If you have x amount of voice traffic that can use a particular link, then you need to provide x amount of available bandwidth for that traffic all the time. Guaranteed Service is a way to give back an “all trunks busy” indication when there is no more bandwidth available to make the requested voice call.