NetQoS
Performance Experts

VoIP: Do You See What I'm Saying?

Managing VoIP Quality of Experience on Your Network

NetQoS, Inc.

Chapter 3 - VoIP Call Quality Performance

Once a call is connected successfully, you can begin an interactive conversation. Just as the speed of call setup affected your opinion of VoIP system performance, the audio quality of the conversation plays a key role in your perception of the overall user experience for the call. In the past, some phone companies advertised the fact that the quality of their calls was superior to that of other companies. But the differences among them were actually quite minor; in fact, we've become accustomed to the very high level of quality achieved through decades of innovation in the PSTN.

The PSTN took advantage of the fact that once a call was connected, the resources needed for that call were reserved for the duration of the conversation. In the world of VoIP, as calls are added to an IP-based network, the necessary resources are most often shared by all network users. Quality cannot be guaranteed and must be carefully managed from the network performance perspective to provide a high level of network quality of service (QoS).

The rise of the cellular or mobile phone has somewhat prepared us for sub-PSTN call quality, but the reduced quality associated with cell phones comes with the significant advantage of mobility. For business communications, reduced call quality can have a direct impact on the bottom line. If you can't talk to your business partners and customers, revenue will likely suffer. As more workers become mobile and need the capability of high-quality business communications wherever they may be, a new premium is being placed on VoIP call quality.

The quality of experience that you and your users have with the phone system is closely related to the perceived quality of each call. Call quality is not a completely objective measurement, and thus it is important to understand some of the different standards for voice quality and the methods used to evaluate it.

Call Quality Standards

Call quality has always been a somewhat subjective measurement. You can ask a group of volunteers to listen to the quality of the voice signal during a call and rate it based on the same set of criteria, but everyone's opinion is slightly different. That being said, standards are in place to provide highly accurate predictions of user opinions of call quality. The industry standard for subjective measurement of voice quality is the mean opinion score, or MOS. The MOS is defined in the International Telecommunications Union (ITU) recommendation P.800.

MOS

The mean opinion score is just what the name implies: it's a score derived from users' opinions. A sentence is read aloud over the telephone to a number of listeners. (For example, a commonly used sentence for this type of testing is "You will have to be very quiet.") After hearing the sentence, the listeners score the conversation based on their opinion of how it sounded. The scores are averaged to come up with the mean opinion score. After years of testing, the ITU used data from listener opinions to codify a scoring standard.

The MOS can range on a scale from 5 to 1, with 5 being the best and 1 being the worst rating. Table 3-1 shows the MOS values and associated quality rating.

MOS	Quality Rating	Listening Effort
5	Excellent	No effort required
4	Good	No appreciable effort required
3	Fair	Moderate effort required
2	Poor	Considerable effort required
1	Bad	No meaning understood with any feasible effort

Table 3-1 – Mean Opinion Score Scale (from ITU P.800 specification).

MOS is the de facto standard for call quality and provides a good, high-level comparison for the quality among different phone calls. A MOS value of 4 or higher is generally considered toll quality, or equivalent to typical PSTN quality. As you can imagine, this approach is pretty difficult to sustain for an ongoing VoIP monitoring and management system; you can't add a team of call-quality evaluators to your payroll and allow them to listen in on phone conversations to provide quality ratings. However, a large amount of research has been done to correlate how people rate calls in the presence of common network impairments like latency and packet loss. Using this information, other standards have evolved that can take impairment information and map it to a MOS.

Different types of MOS are currently being used to rate call quality, depending on the factors that are included in the measurement. The most commonly referenced type of MOS is listening quality (MOS-LQ). Another type of MOS that you may encounter is conversational quality (MOS-CQ). The main difference is that MOS-LQ does not account for factors that can affect the conversational (or two-way) nature of the call, such as latency. MOS-LQ focuses on the quality from the perspective of the listener and only takes into account the loss impairments, like network packet loss and jitter buffer loss.

Let's take a look at some of the other standards that can provide call quality mapped to MOS.

PESQ

The Perceptual Evaluation of Speech Quality (PESQ) is a method that involves active testing to determine call quality. Defined in the ITU P.862 standard, PESQ is the successor to some older quality standards known as PSQM and PAMS. Using PESQ, a reference signal is played through the system, and the received version is compared to the original reference version. Any degradation that occurred is measured, and a quality score is computed. The resulting score is often mapped to a MOS-LQ value.

Because PESQ is an active, or intrusive, monitoring approach, it may not be ideal for networks that are already near capacity.

E-Model

The E-Model is defined in ITU recommendation G.107. The E-Model is useful for VoIP call quality measurement and monitoring because it takes into account a whole range of specific data-network impairments, like packet loss and latency. The E-Model was developed by performing many subjective MOS tests with different codecs and varying degrees of network impairments. The E-Model takes the information about codec, packet loss, and latency and derives an R-value. The R-value can be mapped to a MOS.

The E-Model begins the calculation with a theoretical reference R-value that is free from any impairment. The associated impairments are then subtracted from the reference value to derive the perceived quality value. Table 3-2 shows how the E-Model provides a mapping of R-values and MOS to user satisfaction.

User Satisfaction	R-value	MOS
Very satisfied	90	4.34
Satisfied	80	4.03
Some users dissatisfied	70	3.60
Many users dissatisfied	60	3.10
Nearly all users dissatisfied	50	2.58

Table 3-2 – Relationship between User Satisfaction, R-value, and MOS (from ITU G.107).

The E-Model was originally designed for network planning exercises and provides a good way to estimate call quality as it is affected by underlying network impairments.

P.VTQ

The ITU P.VTQ standard defines an endpoint MOS algorithm based on PESQ. A passive type of monitoring approach, P.VTQ looks at the call quality of real phone calls as they progress, and is usually implemented in the IP phone or gateway. The P.VTQ standard defines a quality value called the K-factor. Like the R-value, the K-factor is mapped to the MOS. More precisely, the K-factor is a MOS-LQ value because it includes impairments such as packet loss and jitter discards, but it does not include delay-related impairments. The K-factor is usually calculated over 8-second samples of the audio for a given call.

IP phones and gateways from Cisco Systems implement the P.VTQ standard as a means of calculating a MOS for VoIP phone calls. At the end of a call, the phones and gateways can provide the quality information for that call.

Our discussion of the standards used to calculate call quality mentioned several network-related factors that can impair that quality and lower the quality score, however it is calculated. Now let's look at some of the key call quality metrics in greater detail and discuss how they can impact call quality.

Key Call Quality Metrics

Just as with any network application, the underlying network performance will have a direct effect on user perceptions of call quality. Metrics like transaction time and network round-trip time are often used to measure TCP application performance and can be helpful when you need to tune network components. However, these metrics are not directly related to call quality.

A more VoIP-specific focus is required when tuning the network to carry high-quality phone calls. As a result, IP telephony experts instead rely on several metrics that relate directly to call quality to help them measure the likely user experience when interacting with the VoIP system. Let's discuss these key metrics individually.

Latency

The time it takes for a VoIP packet to travel from the speaker to the listener is called the delay, or latency, of the system. VoIP traffic is extremely sensitive to latency. The ITU G.114 standard includes research showing that 150 ms of one-way latency is the point where call quality begins to degrade. Delays greater than 150 ms can make conversing difficult and result in a degraded user experience. Latency is something that needs to be accounted for end-to-end—from the speaker's mouth to the listener's ear. Often, the round-trip delay in the network is used and divided by 2 to get a one-way latency value.

As a VoIP packet traverses the network, latency can be introduced in a number of different ways:

- Packetization – The time it takes for a codec to encode and decode a packet for transmission or reception.
- Queuing – The time spent in router queues along the path through the network. The more congestion, the greater the queuing delay.
- Serialization – The time it takes to put a packet on a network link interface. This value increases as the packet size increases and as the link speed decreases.
- Propagation – The time it takes to travel from one point to another in the network. This time is a fixed value that's directly related to geographical distance. A good rule of thumb is 10 microseconds per mile.
- Jitter buffer – Each VoIP phone and each voice gateway provides a jitter buffer to smooth out the effects of network jitter, or variations in delay among packets in the same stream. This buffer adds delay as the packet is held for playout.

Most of the components introducing delay for a given VoIP call are not related to bandwidth. The one exception is the queuing delay, which could be alleviated by extra bandwidth. So if you have a call performance problem due to delay, additional bandwidth probably won't fix the problem. We'll discuss what types of quality issues bandwidth can help and what types of issues that bandwidth does not help, later on in this chapter.

Figure 3-1 shows the components and areas of the network where latency can be introduced.

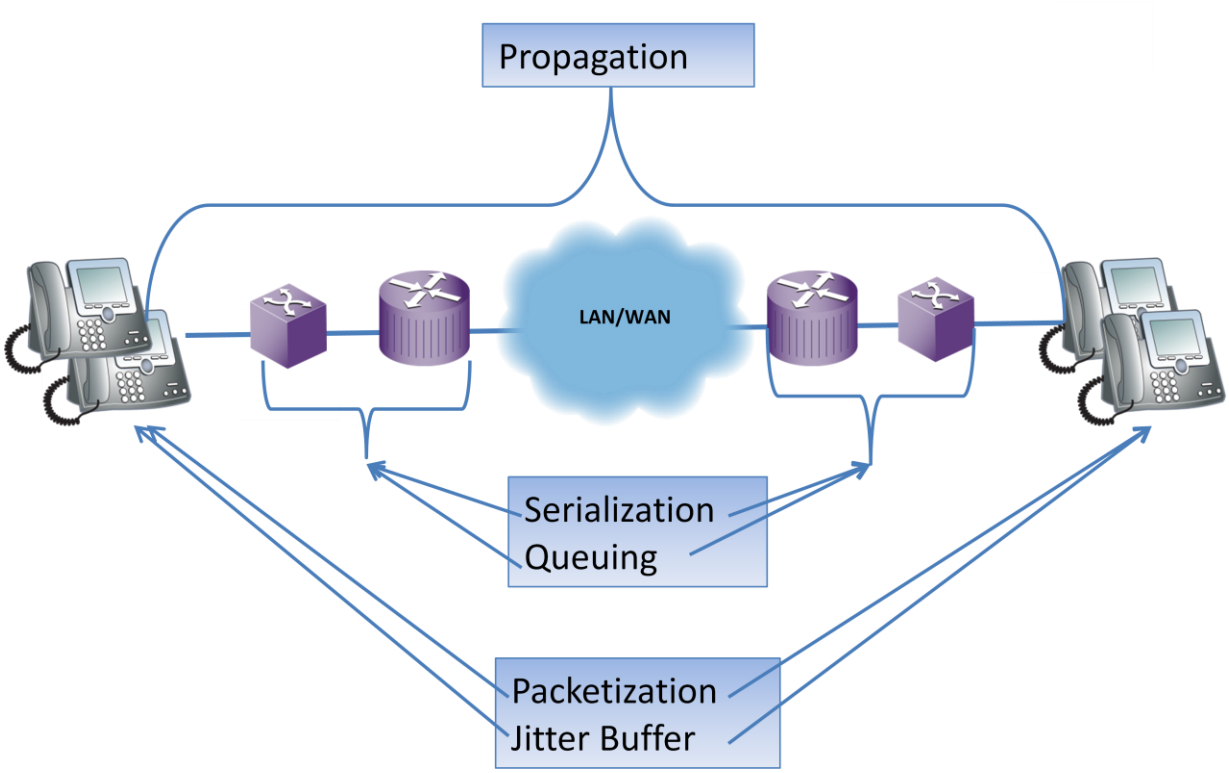

Figure 3-1 – Latency can be introduced at a number of different locations in the network

Latency is a key component of conversational quality because the higher the latency, the more difficulty you have determining whose turn it is to speak. If both parties in a phone call begin to speak at the same time, not only will you not be able to hear everything the other person is saying, but you also may start to feel that you are talking on walkie-talkies and need some protocol to allow for a party to start and stop talking ("Over and Out").

As the latency increases, the likelihood of encountering other quality impairments like echo increases as well. Echo is often present in PSTN phone calls, but due to low latency, it is usually inaudible. In a VoIP network, higher latency values can reveal this normally dormant impairment. We'll discuss it again below, when we introduce the ACOM metric.

Packet Loss

VoIP packets are sent using RTP, so they don't get retransmitted if they are lost. The impact of lost packets varies depending on the number of packets lost and the manner in which they are lost. Packet loss is another key network performance metric that affects the call quality.

There are two types of packet loss:

1) **Network packet loss**: The receiver did not receive the packet because it was discarded somewhere between sender and receiver. Network packet loss has numerous causes, but the two most common are:
   - Network congestion: Anytime there is too much traffic on a link or queue, a router must make the decision to discard some of the packets. Packets are dropped by the router according to a variety of different QoS mechanisms.
   - Network errors: Link errors caused by physical media can lead to packet loss. If a packet is corrupted and can't be reconstructed after transmission, it will be discarded

and thus lost. Duplex errors between network cards and switches is a leading cause of packet loss.

2) **Jitter buffer loss** – The receiver received the packet, but it was too early or too late to be accommodated by the jitter buffer and was therefore discarded.

Concealment issues are also created when packets are lost, either through the network or by the jitter buffer. When a packet is either not received by or is discarded by the phone, many codecs use concealment algorithms to try to preserve as much of the call quality as they can. These concealment algorithms range from simple – just play the last known audio sample – to complex – try to interpolate what the next audio sample might be. The goal is to "conceal" the lost packet. The resulting audio doesn't always sound better, however.

Packet loss occurs in different profiles. Random loss occurs when a packet is occasionally dropped, either by the network or jitter buffer. This loss pattern can impair the call quality, but in most cases, it is not very noticeable. By contrast, a more detrimental type of packet loss is "bursty" packet loss. This loss pattern describes what happens when a number of consecutive packets are lost in bursts. Because the packets are carrying audio information, bursty loss can cause entire syllables or words to be missing from the conversation.

The impact of packet loss on the call quality is also determined to some extent by the codec. The codec defines the packet size that is transmitted and fills the packet with audio samples. As we mentioned in a previous chapter, packet sizes range from 20 bytes for the G.729 codec up to 240 bytes for the G.711 codec. Larger packets typically contain more audio information; thus, their loss has a greater impact on quality. In addition, codecs may compress the audio information to fit more data into each packet. If a high compression setting is used, losing a packet can result in the loss of more audio information.

Codecs are making steady improvements in handling packet loss. Newer codecs, such as Microsoft's RTAudio, can operate with good quality in Internet-type conditions where packet loss may be 10% or higher.

Jitter

Jitter, often referred to as delay variation, is a measure of the differences in arrival times among all VoIP packets sent during a call. When a phone sends a VoIP packet, it sets the timestamp value in the RTP header. When the packet is received, the receiver generates a timestamp and compares it with the sender's timestamp. The timestamps are used to calculate the packet's delay variation or jitter value, which can then be reported by the IP phone or by other monitoring equipment. Too much jitter can cause packets to arrive too early or too late to be processed by the jitter buffer. When this happens, the phone discards the packets – a form of packet loss.

IP phones send data packets at constant rate. The time between packets depends on the codec. For example, G.711 typically sends a data packet every 20 milliseconds. You can think of this process as similar to loading a truck. The truck waits 20 ms for the packet to fill up with audio information and then departs for the destination. The codec outputs the packets to the network at constant intervals, with no variation. Figure 3-2 shows how the network can affect packet arrival times and create jitter.

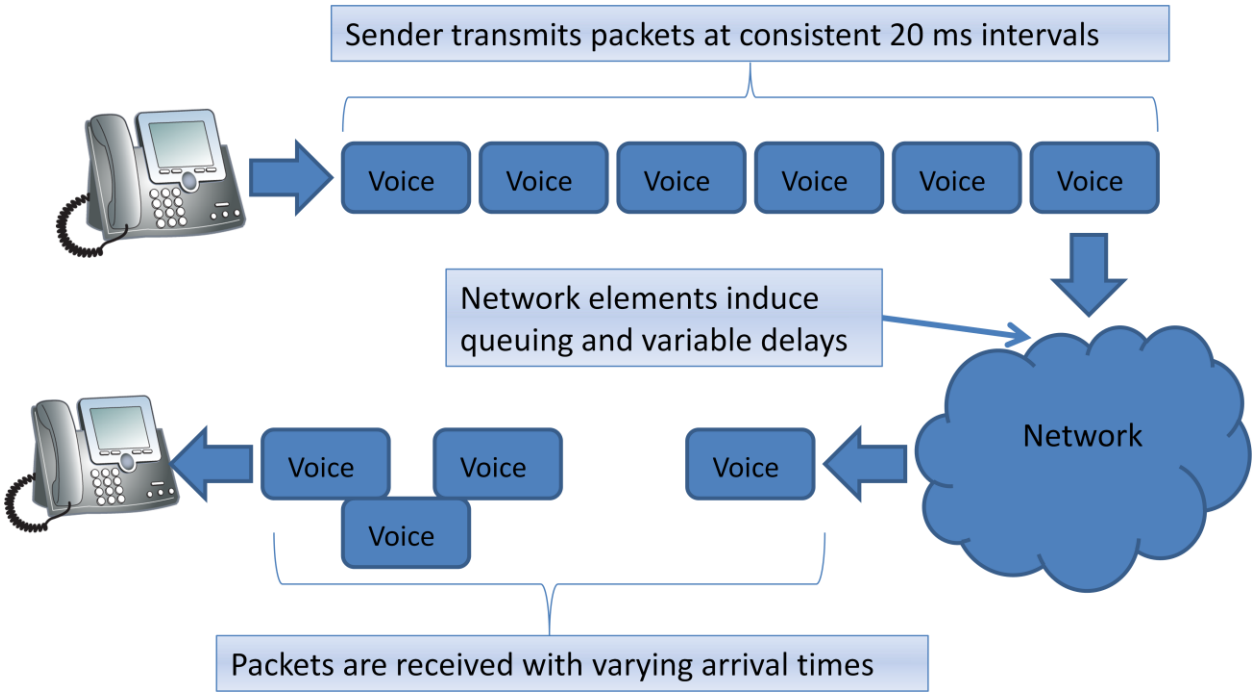

Figure 3-2 – Variable delays in the network creates jitter

To reduce the impact of jitter, the receiver employs a jitter buffer to receive the incoming packets. You may have seen audio or video applications that report that they are "buffering" data for playback. They are actually buffering data in a jitter buffer. Most jitter buffers in IP phones and voice gateways hold at least two VoIP packets. They are adaptive and can adjust, becoming larger or smaller depending on network conditions. As packets are received, they are placed in the jitter buffer, which holds them in order to supply a constant stream of packets to the receiving codec. The jitter buffer may have to account for early- or late-arriving packets and may have to re-order any packets that arrive out of order. Packets that don't fit into the jitter buffer are discarded, resulting in packet loss.

So, why not just make the jitter buffer large enough to handle all possible variations? The reason is that every millisecond increase to the jitter buffer results in a millisecond of additional latency for the VoIP packet. Jitter buffer delay must be factored into the overall latency budget for VoIP.

Combined Echo Return Loss (ACOM)

An impairment to call quality that almost everyone has encountered at some point is echo. It's very distracting to hear your words, or those of the person you are talking with, repeated. Echo is a function of delay. When the delay is less than 25 ms, echo is not perceptible to the human ear. This is usually the case in the PSTN, with its dedicated circuits. However, in VoIP environments, delay is often greater: up to 150 ms, which is a good limit in a well-designed system. With additional delays and PSTN connections through voice gateways, the likelihood of hearing echo on VoIP calls increases. For this reason, voice gateways include echo cancellation components to help reduce, or "cancel," the echo on VoIP calls to the PSTN. Echo cancellation works by comparing the signal pattern entering the PSTN tail circuit with the

signal pattern that returns. Variations in echo strength and the time period during which the echo canceller looks for the echo can affect the efficiency of the cancellation process.

There's not a specific metric that tells you how much echo is present in a call. However, some VoIP monitoring equipment can report a metric that gives you a sense for how well echo cancellation is working: the standard known as Combined Echo Return Loss, or ACOM. ACOM is defined in the ITU G.168 standard and measured in decibels (dB). It represents the combined echo return loss through the system – the attenuation of echo from all possible means (Echo Return Loss (ERL) + Echo Return Loss Enhancement (ERLE)). ACOM is measured on the IP side of the echo cancellation device and includes all sources of echo loss, providing a good gauge of the remaining echo strength. Figure 3-3 shows the measurement of ACOM in relation to the voice gateway, PSTN, and echo canceller.

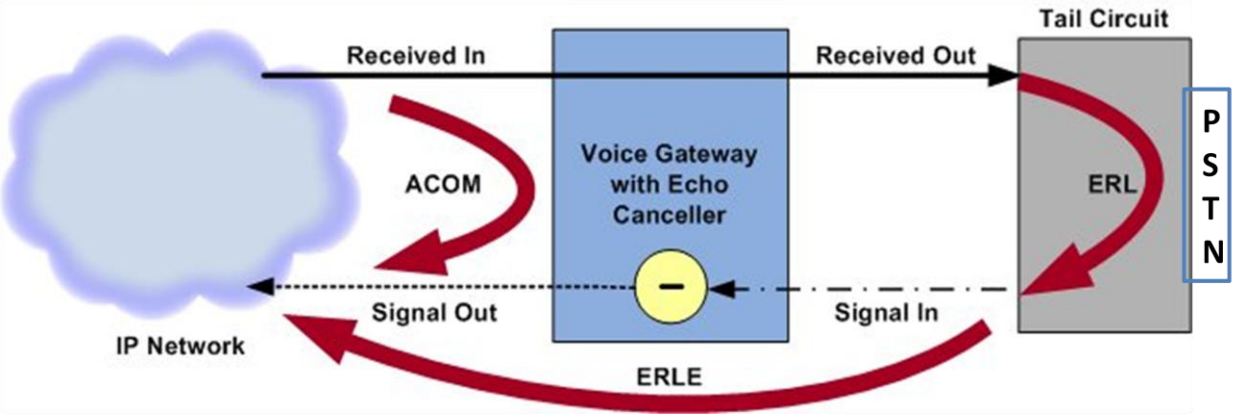

Figure 3-3 – Echo measurement and echo cancellation

Since ACOM measures the loss volume between the original signal and the echo, you want the ACOM values for voice gateway calls to be high. In other words, the greater the ACOM value, the lower the level of echo. For example, an ACOM value of 45 dB means that the echo heard is 45 dB lower than the original conversation sounds. ACOM values of 15 dB or higher are usually considered good. If the ACOM drops below 6 dB, the echo canceller usually cannot suppress the echo.

ACOM is measured within the voice gateway and does not apply to phone calls directly between two IP phones.

In the previous sections, we've introduced some of the key call quality metrics:

- MOS – The de facto standard for call quality.
- Latency – The delay of packets between talker and listener.
- Packet Loss – Packets that are lost in the network.
- Jitter Buffer Loss – Packets that are received by the listener, but that arrive too early or late for the jitter buffer to handle, and are therefore discarded.
- ACOM – The combined echo return loss, which gives an indication of the effectiveness of echo cancellation.

Now let's consider an approach to troubleshooting call quality performance issues that uses these metrics in the process.

Troubleshooting Call Quality Performance

Just as with any network or application performance issue, identifying the problem is the first step in tackling a call quality issue. Looking in the right place is a good second step. Now that you understand the key VoIP call-quality metrics, you can begin to identify the problem by finding out which call quality performance metric is impacted.

You have a couple of options for gaining the required visibility into call quality metrics. One method is to simulate phone calls and measure the results. This is the general approach taken by the Cisco IP SLA function that is a part of almost all Cisco routing and switching devices. This approach actively generates RTP packets and sends them through the network. The receiving router or switch calculates call quality metrics for the simulated call. Another way to identify potential call quality problems is by deploying a hardware or software monitoring tool to passively monitor the call quality metrics reported by your phones as users make real calls, and receiving notifications when the call-quality metrics signal trouble. Both approaches provide value to a proactive management solution.

The rest of our discussion of troubleshooting is going to assume that you are using a performance management and monitoring tool. A good hardware- or software-based monitoring system is essential if you want to provide high-quality VoIP to your network users. And then you need to develop a smart approach to configuring and using your monitoring tool.

VoIP call quality issues are most likely to occur in two areas of the network:

- Calls that are traversing a WAN link or links
- Calls to the PSTN through a voice gateway

Paying close attention to these areas can help you troubleshoot call quality issues as they arise. When you are monitoring call-quality metrics, you need to be able to quickly see which users are affected and which metrics are showing the impact of the underlying problem.

A good troubleshooting process must therefore include a step-by-step approach to isolating the problem. Figure 3-4 outlines a troubleshooting process for call-quality performance issues.

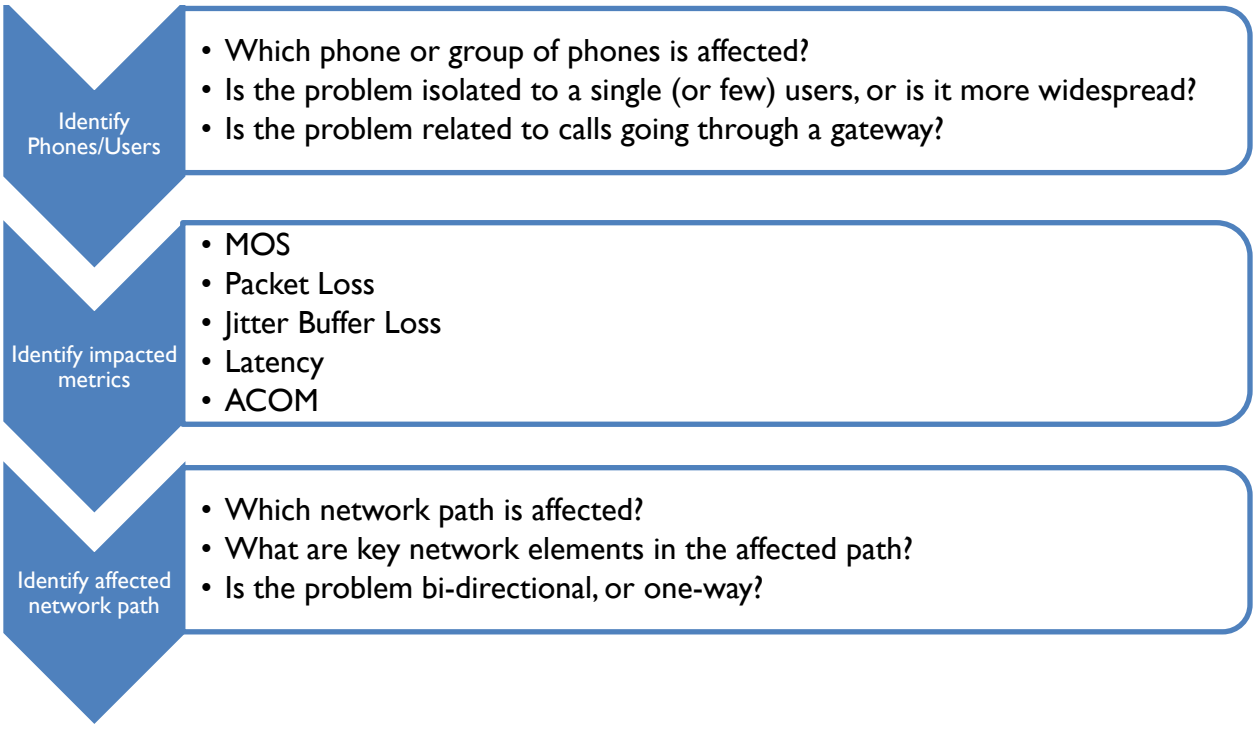

Figure 3-4 – Step-by-step call quality problem identification

Once you know the affected users or phones, performance metric, and the network path involved, you can take a look at the devices and network segments to determine whether a given problem is device- or network-related. Knowing which metrics are affected can help you look in the right place from the outset. In the sections to follow, we'll outline an approach to identifying call quality issues that will put you well on your way toward resolving them.

Performance Incidents

Applying thresholds to the call quality metrics we've discussed in the previous sections is a good way to take a proactive approach to managing call quality performance. Multi-tier thresholds provide a way to alert the network manager when performance is deteriorating, before the problem becomes acute. By setting up good thresholds, you can be alerted to performance problems before you get complaints from irate users. But what are good threshold settings? Are they industry-standard, or network-specific?

Most VoIP systems would benefit from a combination of standard threshold values and values specifically tailored to unique network conditions. On a busy network, where potentially thousands of calls could be running simultaneously, your monitoring tool needs a mechanism for avoiding duplicate notifications in response to degraded call performance. And you need the flexibility of granular thresholds that alert you to changes in all the key metrics that affect call quality.

Since the MOS is the widely accepted standard for call quality measurement, you should consider setting thresholds to alert you to degraded MOS values. In addition, the underlying performance metrics like latency, packet loss, jitter buffer loss, and ACOM have either standards-based thresholds or generally accepted best-practices thresholds.

Because we are dealing with the likely user experience when we attempt to manage the performance of a VoIP system, the subject of thresholds will necessarily be a little subjective. However, industry guidelines are in place and offer a starting point. For example, we discussed earlier the ITU G.107 mapping of user satisfaction to MOS. From this standard, good thresholds can be derived that help alert you when user satisfaction with the call quality is most likely declining. Figure 3-5 lists default thresholds derived from industry standards and best practices.

Metric	Degraded Threshold	Excessive Threshold	Minimum Call Minutes
MOS	MOS 4.03	MOS 3.6	15
Packet Loss	Percentage 1	Percentage 5	15
Jitter Buffer Loss	Percentage 1	Percentage 5	15
Latency	Milliseconds 150	Milliseconds 400	15
ACOM	dB 15	dB 6	15

Figure 3-5 – Guidelines for call quality performance thresholds

For any type of call monitoring, look at a set of calls for a given time period—15 minutes, for example. During this time period, a number of calls are completed. Monitoring systems are notorious for generating tons of alerts, so you may want to filter some of the noise by choosing a reasonable value for a minimum number of call minutes that must be completed by phones in the monitored system before any alerts can be raised (see the "Minimum Call Minutes" column in the above image). In the threshold guidelines shown above, we selected 15 call minutes as the minimum threshold.

For a proactive monitoring solution, you'd like to know if the metrics for the calls were rated Normal, Degraded, or Excessive. These can be color coded Normal=Green, Degraded=Yellow, and Excessive=Orange, to help easily spot performance degradations. A performance incident or alert should include information to help you quickly identify the location of the problem. Information such as the phone location (network subnet) and the call server or gateway that handled the call setup should be included.

Network performance is variable, and the same factors that affect network latency, packet loss, and jitter will affect VoIP call quality performance. As you configure thresholds and alerting for performance issues, consider the possibility that certain network locations may consistently have worse call quality performance. For example, a group of phones in a remote location at the other end of a slow-speed WAN link is likely to receive below-average call quality performance due to the higher latency in the WAN. For these phones, apply a set of threshold values that allow for routinely higher latency metrics at the specific remote location. Otherwise, you'll keep seeing the same alerts each time someone uses those phones to place a call.

Once you have good metric thresholds defined, then you can begin the process of analyzing call quality metrics and take a proactive approach to troubleshooting.

Identify Phones/Users

The first step in troubleshooting a VoIP call-quality issue is to identify phones and users that are experiencing degraded call quality performance. This identification process is made easier by mapping the network and organizing subnets into groups and using those groups to configure reporting options in your VoIP monitoring tool. Grouping phones that are expected to achieve similar performance into distinct *locations* is a good first step.

A VoIP location definition might contain a subnet or group of subnets. Phones within those subnets would be grouped for reporting purposes based on the VoIP equipment and links that they all access. They might all access the same call server, for example. Looking at the measured quality of calls to and from the locations in your organization will allow you to spot potential quality problems and assess their impact. For example, you need to know whether the problem is widespread, across multiple network locations, or whether it is confined to one or two locations.

Figure 3-6 shows a report from our VoIP monitoring system. It contains a listing of locations where calls were made, with three locations that are experiencing quality issues during the last day (indicated by orange timeslots, where the quality was rated as "Excessive").

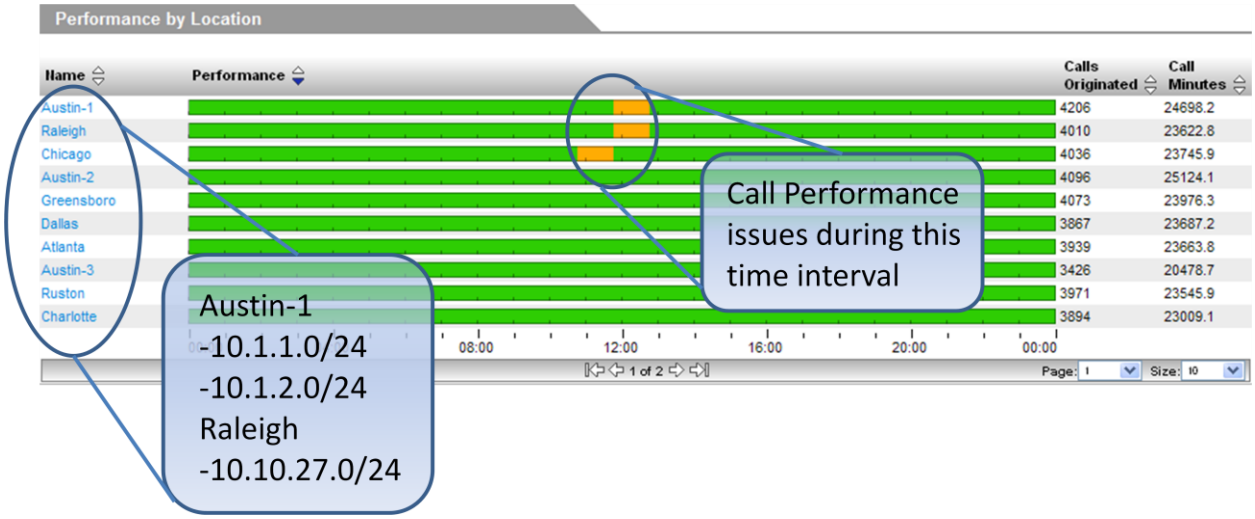

Figure 3-6 – View of call performance for each location in the enterprise

From this type of report, organized by network location, you can quickly spot the places that need further investigation. To determine whether the problem within a location is widespread or isolated to particular phones or users, you also need information about the calls that are occurring within that location. Reporting quality information on a per-call basis, can help you spot problems that may be isolated to individual users or phones.

The next troubleshooting step to take is to identify the call-quality metrics that are showing signs of performance degradation.

Identify the Affected Metrics

Above, we discussed the key metrics that directly impact VoIP call quality. After determining the network locations that are experiencing quality issues, we need to identify the specific metrics that are affected—the ones that contributed to the "Excessive" call quality ratings. In

the above example, the locations Austin-1, Raleigh, and Chicago were reported to be experiencing some call quality issues. As a logical next step, we would like to drill down to the Austin-1 location and see the call quality metrics for calls to/from this location. Figure 3-7 shows an example of the performance ratings for calls to and from the Austin location.

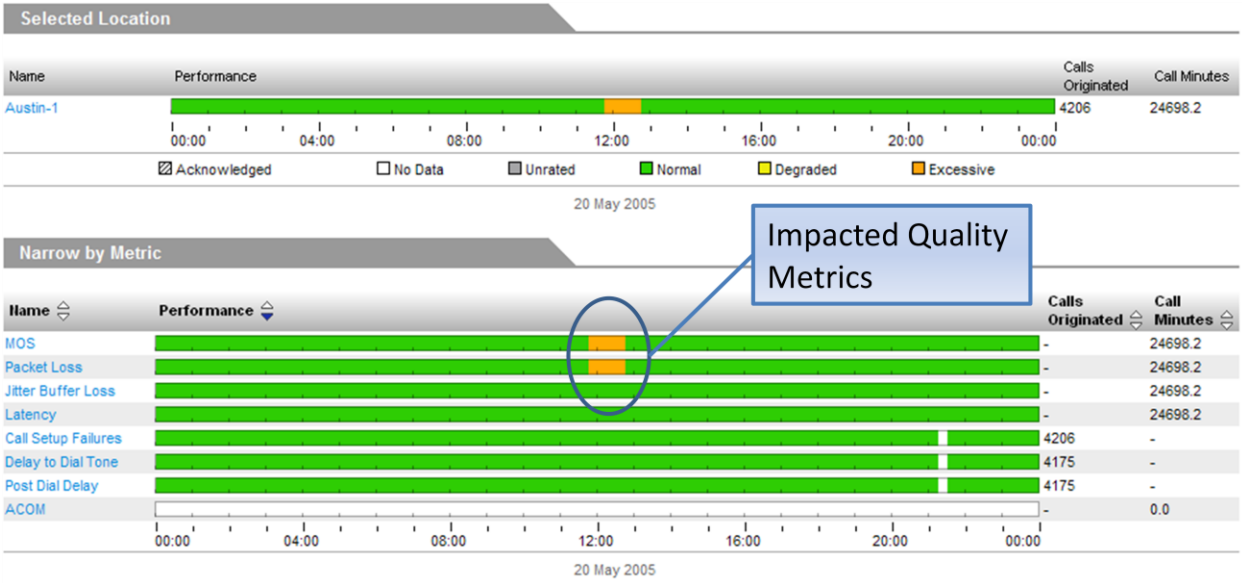

Figure 3-7 – Identify the call quality metrics that are contributing to degraded user experience

The MOS metric is rated as "Excessive," which is indicated by the orange timeslot. Looking at the other metrics, we see that the Packet Loss performance metric is also rated as "Excessive" and is likely the contributing factor to the low MOS. The other metrics, Jitter Buffer Loss and Latency, are normal, as indicated by the green timeslots.

To further aid your troubleshooting, it's important to take a look at the actual metric values for the time period when the call quality issue occurred. Looking at the metric values can help you understand the specifics of a given time period: Is the packet loss 1% or 5%? Is the latency 10 ms or 100 ms? In addition, the metric value details can point out times where the MOS declined due to other contributing performance metrics.

So far, the approach we've taken for troubleshooting call quality issues has allowed us to determine that we have a call quality issue in a particular network location. We know the metric or metrics that were degraded, the actual values of the metrics during the time period in question, and any correlation between the metric values, such as which performance metric lowered the MOS value. The next step is to identify the network path or paths that were carrying the calls affected by the underlying problem that caused the quality issue.

Identify Affected Network Path(s)

As we pointed out in a previous chapter, each VoIP call consists of two media streams that do not have to traverse the network over the same path. In the example above, users in the Austin-1 location were experiencing degraded MOS values caused by packet loss. If we look at the RTP media streams that were received by phones in the Austin-1 location, we can see which location has been sending the data that was lost. Figure 3-8 shows an example of the Austin-1 location and performance ratings for all other locations that were parties to calls with phones

at Austin-1. In the following chart, we can see that the Raleigh location is the only other location that was impacted.

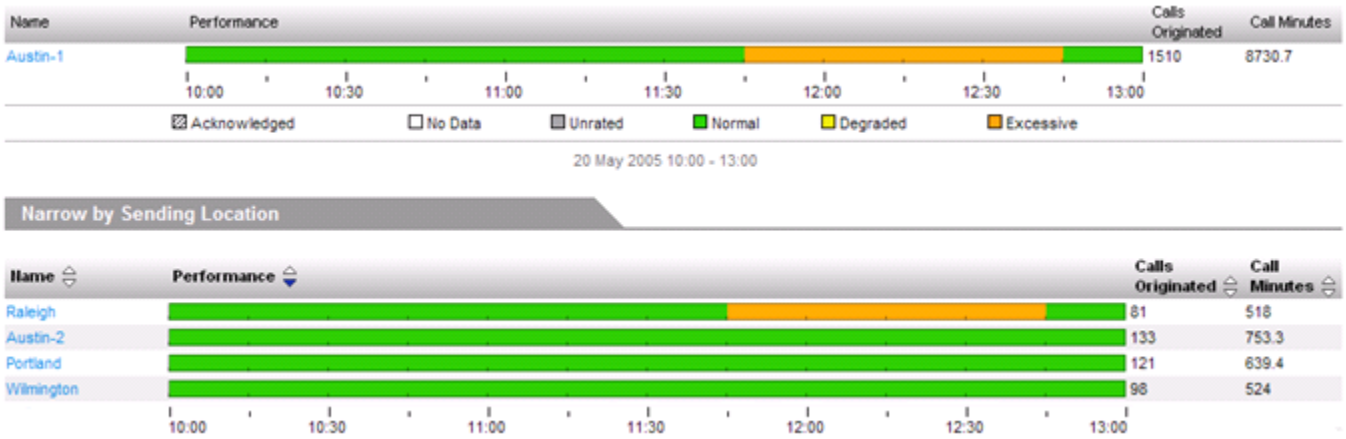

Figure 3-8 – Identify the affected network path(s)

We have now determined that the affected network path is Raleigh → Austin. Calls between our Raleigh and Austin-1 locations are experiencing significant packet loss somewhere along this path.

The network path provides useful information in troubleshooting call quality issues because a device or link in that network path is probably causing the problem. Using other network performance data sources, such as NetFlow, SNMP, and IP SLA, can help further pinpoint the root cause of the packet loss in this scenario. A consolidated view of network performance information is a useful part of a proactive call quality management strategy.

However, there are times when the overall network performance looks good, and the call quality problem is isolated to a specific call or phone. In these cases, there is another layer of detailed information that can be helpful for troubleshooting.

Troubleshooting Information for Specific Calls

A call quality problem with a single phone or user may be more difficult to troubleshoot than an issue that is affecting an entire network location. For any VoIP call, data will be sent in two RTP streams between the IP phones or between an IP phone and a gateway. In these cases, you need as much information about the two sides involved (the phone or the gateway) and more granular quality information – quality metrics while the call is in progress. Getting this level of quality information for every single call in your enterprise will likely result in data overload – you'll have so much data that you won't be able to make sense of it. But for specific cases, this type of visibility is invaluable to getting to the core of the problem.

IP Phone and Gateway Information

IP phone and voice gateway information (both configuration and runtime) can be very useful when debugging call quality problems for specific calls. Table 3-3 lists the information for IP phones that you should collect.

Phone Information	Comment
Name	The unique name for the phone. Often this name may include the phone hardware or MAC address.
Phone Number	The directory number or phone number for this phone.
IP Address	The IP address for this phone.
Port	The UDP port number used for the RTP stream. Useful information for firewall traversal issues.
Location	The network location for this phone. May be a name given to a subnet or group of subnets.
Model	The model or type of phone. All phone models are not created equal. Some have more features, some have less features.
Call Server	The call server the phone is registered with.
Codec	The codec used for the call. Useful for troubleshooting issues where different codecs are used for the same call.
Firmware Version	The version of firmware running on the phone. Newer versions of firmware may have bug fixes that you need.
Serial Number	The serial number for the phone. Useful for inventory tracking purposes.
Switch IP Address	The management IP address of the network switch that the phone is connected to. Knowing the switch that the phone is connected to is useful for troubleshooting issues like duplex mismatch, power-over-Ethernet problems, and VLAN configuration.
Switch Name	The DNS name for the switch where the phone is connected.
Switch Port	The switch port where the phone is connected. Useful for troubleshooting port specific issues.

Table 3-3 – Useful phone information for troubleshooting.

Likewise, for calls that go through a voice gateway (coming from the PSTN to the IP network, or the reverse), information for the gateway and the individual voice interface on that gateway is very useful for troubleshooting purposes. Table 3-4 lists the gateway information that you should collect.

Voice Gateway Information	Comment
Name	The DNS name for the gateway.
Phone Number	The calling or called PSTN phone number.
IP Address	The IP address for this gateway.
Call Server	The call server which the gateway is communicating with for call setup purposes.
Codec	The codec used for the call. Useful for troubleshooting issues where different codecs are used for the same call.
Voice Interface	The interface that the call is using to connect to the PSTN. Often this will be the PRI, with information about the slot, subunit, and port used for the call. Useful for troubleshooting problems with

	connections to a particular carrier.
Voice Channel	The channel the call is using to connect to the PSTN. Useful for troubleshooting problems with a specific interface channel.

Table 3-4 – Useful gateway information for troubleshooting.

The gateway provides a number of voice interfaces and channels for connection to the PSTN. Different interfaces may be connected to different carriers or providers. The interfaces may be plugged into different hardware slots and subunits within the gateway. Understanding which call used which interface can be useful for debugging, even if your only role in this part of the process is calling the equipment vendor or service provider and outlining the issues.

Transient Quality Issues

Some call quality issues can be very transient and difficult to track down. A faulty switch port that doesn't negotiate parameters correctly with the phone network interface card can lead to sporadic packet loss. For these types of quality issues, monitoring the call quality metric values as the call is in progress is important. If you are monitoring the metrics while the end user is still on the phone, you can see variations over time as network conditions change. Correlating this performance information with metrics from other network performance data sources can provide insight into the root cause of a call-quality problem.

An important aspect of any call quality troubleshooting process is taking a look at the familiar network factors, like bandwidth, protocol usage, QoS, and WAN optimization, all of which affect the performance of all your network applications along with VoIP. Let's discuss a VoIP-specific approach to these ever-present threats to optimal performance.

Network Considerations

When addressing call quality performance problems, you certainly can't rule out general network issues. The network plays a key role in call quality performance, so bandwidth usage and QoS consistency are important parts of the network management equation that must be considered when troubleshooting VoIP call quality. Understanding these two items requires network traffic analysis, with visibility into the composition and volume of network data flow. This type of analysis, best performed after maximum visibility into that data flow is achieved, is a critical component for a smooth and successful VoIP deployment and will help you far into the future as you plan for upgrades and the inevitable expansion of the VoIP system.

NetFlow is an example of a data source that provides information needed to effectively manage call quality performance. NetFlow is built into most Cisco network routers and switches. Statistics are kept about the protocol bandwidth usage and QoS markings for voice call traffic (and all other) flows. You can then use a third-party monitoring package to collect, parse, and report on the NetFlow data. The NetFlow reporting package will nicely supplement VoIP-specific monitoring and management tools.

Bandwidth Usage

Once you have the required tools in place to help you analyze and understand the traffic that's flowing over your network links, you are better poised to avoid capacity and utilization issues that could potentially affect the VoIP system. We discussed the bandwidth usage associated with the most popular VoIP codecs in the first chapter. Do you have any idea how much

bandwidth is consumed by VoIP calls on key WAN links? Do you understand the percentage of usage traceable to VoIP when compared to the mix of other application traffic?

These are good questions that can be answered by looking more closely at bandwidth consumption on your network. Figure 3-9 shows a breakdown of protocol bandwidth usage for a single router interface. This report was taken from a NetFlow reporting tool.

Figure 3-9 – Voice traffic bandwidth usage on a particular interface

Bandwidth consumption for VoIP traffic can be deceptive. Protocol headers add overhead, and two bi-directional streams are required per call. On the example network being monitored for the above report, voice traffic was 89.03% of the protocol mix for a particular router interface. If call traffic had grown heavier, the link might have become overburdened, resulting in packet discards.

In a Cisco environment, you can use the visibility provided by NetFlow to determine the typical mix of protocols and their bandwidth usage on particular links in your network. You may find that VoIP call traffic is using more bandwidth than you had allocated for it in your planning, or that it is going over an interface that you didn't expect.

QoS Mismatch

QoS consistency for call packets is another item that can be determined via NetFlow analysis. Good VoIP call quality requires QoS mechanisms to prioritize the voice traffic and provide low-latency queuing. IP phones and gateways will mark each packet sent with the desired QoS setting. The TOS byte in the IP header is commonly used to contain the priority information. As the packet traverses the network, routers along the way may alter the marking of the packets. This is particularly true of MPLS networks, where a packet entering a carrier network with one type of QoS marking may leave the network with a different QoS marking. QoS is only as good as its weakest link, and the configuration of any router can impact the prioritization that the packet receives along the network path. VoIP packets are relatively small, making them likely candidates for queuing behind larger application packets if they don't have the correct prioritization bit settings.

NetFlow information can be used to show the breakdown of QoS packet markings for packets passing through a router interface. The TOS byte in the IP header contains the QoS marking. This byte is also known as the DiffServ Code Point (DSCP). The bits within the DSCP byte

represent different QoS levels, and the value of the byte is usually included in common DiffServ naming conventions. For example, a value of 46 in the DSCP portion (first 6 bits) of the TOS byte field would be known as DSCP46. Just as the field has multiple names (TOS and DiffServ), the values have different names as well. For voice traffic, you are likely to see a value of Expedited Flow (EF) which is also known by its DiffServ name as DSCP46. Using the TOS byte values from NetFlow, a breakdown of the different QoS values can be reported. Figure 3-10 illustrates this concept with a report from a NetFlow reporting tool.

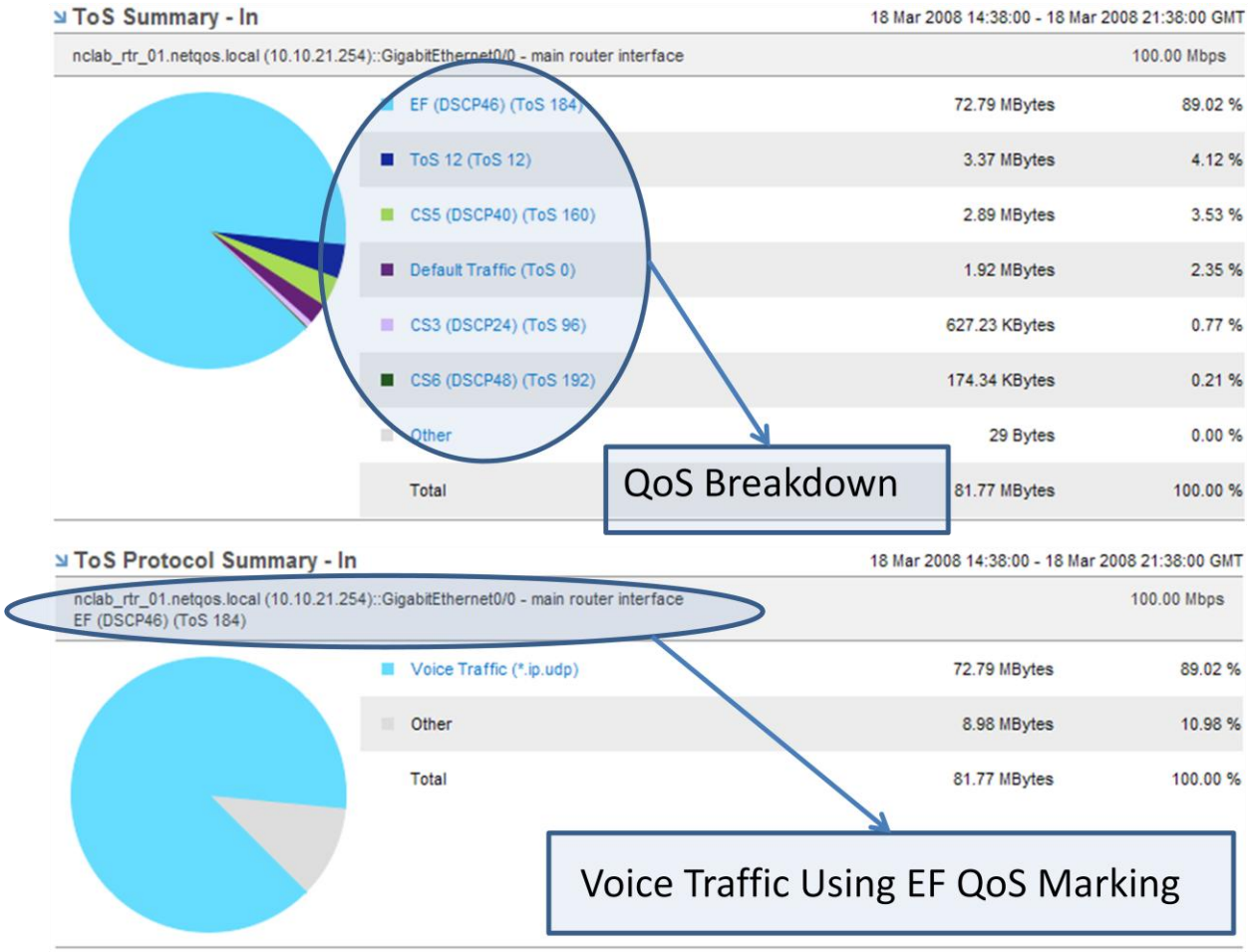

Figure 3-10 – Use QoS marking information to look for mismatches.

In addition to bandwidth allocation and QoS configuration, the manner in which you deploy phones, gateways, and call servers can also greatly impact call quality. Let's discuss some of the deployment models and considerations.

Network Deployment Considerations

A number of different network deployment models are valid for a VoIP system and can deliver excellent performance. But whenever possible, it's best to plan carefully and select optimal placement for critical components. For example, the locations of the IP phones with respect to other IP phones and voice gateways can easily have an impact on call quality performance.

A common enterprise deployment model is multiple sites with centralized call processing. Call servers are centralized, and phones may be local or remote. For the case where phones are located in remote or branch offices, the calls between sites may have to traverse a WAN. For these reasons, many enterprises are moving to MPLS networks, which can provide QoS mechanisms for the VoIP traffic.

Class-based QoS is commonly used to provide separate queues for VoIP call traffic. With class-based QoS, you typically define a policy that controls the amount of traffic allowed for each traffic class. Each traffic class is provided with a queue for its traffic. A separate queue is important to minimize latency and queuing delays for the VoIP traffic.

As you add network hops between the phones or between phones and voice gateway, you inevitably introduce additional latency. Processing at each device along the way adds up, and don't forget the propagation delay, which is related directly to the physical distance that the packets must travel: they can't go faster than the speed of light. Some network scenarios that have the highest potential for creating or exacerbating call-quality performance issues:

- Phones or gateways located at the end of slow-speed links, or links with high latency.
- A wireless phone accessing the network via a high-latency wireless network.
- Phones or gateways using a WAN link (or links) to connect to other phones.
- Call traffic that must traverse a carrier network (that you don't own).
- Gateways located far from the IP phones that are making calls through them.

Good network design principles are critical for ensuring optimal call quality performance. Include QoS early, or you will have to add it at a later date when it might not be as easy.

WAN Optimization

Many enterprises are centralizing data-center resources. As this occurs, a relatively new technique known as WAN optimization is being used to improve the performance of critical customer applications and the LANs at remote locations. WAN optimization is an umbrella term for many different techniques, including TCP optimization, application protocol optimization, and data caching. The goal of optimization is to reduce network flows in an effort to decrease application latency and make more bandwidth available.

Should you consider using WAN optimization techniques with VoIP in an attempt to improve call performance? Let's discuss the key points:

- WAN optimization does not help UDP-based protocols. Most WAN Optimization technology is focused on TCP applications. The VoIP call traffic uses RTP over UDP.
- WAN optimization techniques do not work well with protocols that have small packet sizes. VoIP media packets range from 20 bytes to 160 bytes, depending on the codec.
- WAN optimization data-caching techniques would not help VoIP call traffic. Data caching is used for cases where large amounts of the same data are frequently requested by client applications on the WAN side. A VoIP call consists of encoded audio data, which is changing constantly over time.

It is likely that WAN optimization techniques would have a negative, direct impact on call performance—but only if an attempt was made to optimize the RTP protocol. This doesn't mean that WAN optimization could not offer indirect help. By applying WAN optimization techniques, other application traffic bandwidth consumption could be reduced. This would

effectively provide more bandwidth and less resource contention for the VoIP call traffic—thus reducing the likelihood of queuing delays and improving the overall performance conditions for the VoIP calls.

We've discussed several network conditions that can affect call quality performance. Now we'll broaden the scope of our discussion and take a look at some other factors that can reduce VoIP call quality.

Other Considerations

In addition to the network considerations, other components in your VoIP system can affect the call quality. Some areas to watch for are usage of voice activity detection, voice gateway translations, and faulty end stations which can affect call quality.

Voice Activity Detection (VAD)

Voice Activity Detection, sometimes referred to as "silence suppression," is a feature available in most VoIP systems. It is primarily used for bandwidth savings. The idea is that during many portions of a conversation, the parties are not talking at the same time. During most conversations, at any given moment, one party is talking and the other listening. VAD is an algorithm in the IP phone or gateway that detects when one party is not talking and instructs the phone or gateway codec to stop sending data, thus reducing the bandwidth required for the call.

VAD can create quality issues such as clipping and audio loss because the silence detection algorithms are not perfect. For example, many telephone conversations are very interactive, with both parties interjecting during the conversation. VAD may clip the beginning or ending portions of the conversation. The tradeoff you need to consider with VAD is whether to exchange quality for bandwidth savings.

Voice Gateway Translation

Before convergence, all network managers had to worry about was typically the IP network. But the PSTN is not going away anytime soon. Anytime you have a place in the network where translation between protocols occurs, such as the voice gateway, where analog calls are translated to digital, there is always the potential for call quality performance issues.

The voice gateway acts as the interface between the IP network and the PSTN. When you make a call that goes through the gateway, the bidirectional RTP streams travel between IP phone and gateway. The gateway terminates the RTP stream and translates the packets to information that can be transferred to the PSTN. In some cases, the gateway may need to translate between different codecs. For example, you may use a G.729 codec with reduced bandwidth requirements for calls over certain WAN links. The gateway may have to translate between phones using G.711 on one side and G.729 on the other side. The voice gateway contains digital signal processors (DSPs) that provide the codec translation in the hardware. The DSPs take the incoming PSTN audio and generate packets to transmit to the IP phone. The quality and speed of this process can have a direct impact on the quality of the call.

Endpoint Characteristics

Other factors that can play a role in call quality reside within the IP endpoint, be it a softphone or a regular, "hard" phone. If the endpoint is a softphone, consider the kind of microphone and speakers that are being used. Computers running a softphone quite often do not have very

sophisticated microphones or speakers. What about volume levels? Does the user have the volume set on the phone correctly? Are other user errors in play here?

Environmental factors can play a role as well. Is a speaker phone being used for the call? If so, is a large amount of background noise interfering with the reception?

If a particular phone is continually reporting low MOS values, begin by taking a look at the underlying network metrics, as we suggested above. But if the network metrics look good, the problem may be attributable to the specific phone; or to the surrounding environment where the call is being placed or received.

Increase the Bandwidth to Improve the Quality?

When call quality performance problems occur, what steps can you take to solve the problem? Can you solve VoIP performance issues by throwing more bandwidth at them? Some in the industry seem to think so. Yet engineering a data network to provide the same level of VoIP call quality performance that we are accustomed to in the PSTN is a complex undertaking with many elements to consider besides bandwidth.

Bandwidth Helps With Additional Calls

The amount of bandwidth required by a VoIP call can be deceptive. The codecs send data at relatively slow data rates. G.711 typically consumes the most bandwidth, operating at 64 kbps. Although this is the nominal data rate, there is more to the story. For every VoIP packet that is sent, an RTP, UDP, and IP header are added. These headers add 12, 8, and 20 bytes respectively. So while G.711 sends data at 64 kbps, the actual amount of bandwidth required is more like 87 kbps due to the additional headers.

When deploying VoIP, it's a good idea to understand the call volume in your network. For example, what's the busy hour call traffic between two office locations? Once you determine your typical usage levels, you can calculate the amount of bandwidth required to run VoIP traffic over your network links. If specific links are trying to support too many calls, congestion will occur, which will lead to call performance issues. So adding bandwidth to a specific network link can enable those links to support more VoIP calls with higher call quality.

Bandwidth Helps to Alleviate Congestion

Congestion is one of the leading causes of lost packets and jitter, two network impairments that can lead to poor application performance and can be deadly to the quality of a VoIP call. Congestion is a fairly simple concept: there are too many users for a given resource. Rush hour on busy freeway provides a good illustration. In a network, congestion might occur on a given link because too much traffic is flowing over that link; VoIP is often combined with traditional data-networked applications like email, ERP, and Web. TCP has congestion control built into the protocol, but for UDP-based applications, like VoIP, there's nothing in the protocol to help. In fact, when faced with congestion, the TCP applications will actually back off and slow down or stop sending data (creating potential performance issues for the TCP application users), while the VoIP phones will just keep on sending it.

Adding bandwidth to a link can allow for more calls and for more application traffic to traverse the link without that link's becoming congested. Adding more bandwidth than will be utilized is typically called oversubscribing a link. Oversubscription can help prevent congestion, but it's usually expensive.

Bandwidth Helps Enable YouTube Browsing

Adding bandwidth might actually enable applications like YouTube video viewers to run on your network. Without the additional bandwidth, video viewing might not have been possible – or the quality might have been so bad that no one would try it. With additional bandwidth, your users may start to enjoy watching videos from their desk. This is an interesting twist on VoIP call quality problems. You add the bandwidth to fix the quality issues, but in doing so, you actually enable new bandwidth-hungry applications, which consume all the additional bandwidth; leaving you with the same call quality problems, which still must be addressed!

Bandwidth Doesn't Help Reduce Delay or Latency

Because VoIP traffic is extremely sensitive to delay, slowing down a VoIP packet by more than 150 ms can cause serious quality problems. Earlier in "Key Call Quality Metrics", we discussed the different components of the network that can create delays for a VoIP packet.

Most of the components introducing delay for a given VoIP call, such as jitter buffers, are not affected by additional bandwidth. The one exception is the queuing delay, which could be reduced by the extra bandwidth. So if you have a call quality problem due to delay, additional bandwidth may not fix the problem.

Bandwidth Doesn't Help Avoid Echo or PSTN-to-IP Conversion Issues

Whenever a call crosses from the IP network to the PSTN, there is always a chance for quality issues to be introduced. For example, echo is one of the most annoying quality issues that can occur on VoIP-to-PSTN calls. The IP portion of the call is not the direct cause of the echo, but it can be indirectly related. In many cases, the echo may already be located in the analog portion of the call, but it is imperceptible until the packets hit the data network. Additional delays added when traversing the IP network and any gateways doing analog-to-digital translation can cause the echo to be audible by the phone users.

As we've stated earlier, additional bandwidth is usually not a good solution for solving VoIP quality issues caused by delay.

Bandwidth Doesn't Help Improper QoS Mechanisms

VoIP is one of the applications that can take advantage of the move to MPLS-based networks. Many companies are migrating to MPLS-based networks, and at the same time, or shortly thereafter, considering rolling out VoIP. QoS mechanisms are critically important for VoIP traffic, but the QoS policies must be applied end-to-end to be effective. One link in the path with improper QoS is equivalent to no QoS for the VoIP call. As VoIP traffic traverses an MPLS network, it may enter and exit the "Enterprise" and "Carrier" networks at different points. If the QoS marking is different at these ingress and egress points, the quality of the call is likely to suffer. This is typically a configuration problem that would not be solved by additional bandwidth.

Bandwidth Doesn't Help Network Architecture: A Weak, Slow Link is Still a Weak, Slow Link

If you decided to add bandwidth to improve VoIP call quality, where would you add it? VoIP calls will be traveling from location to location, through gateways, potentially over VPNs, and even through wireless networks. You might upgrade some of the WAN links, but if there is still even one weak, slow link in the path, you could still have quality problems. In addition, the upgraded links could produce new bottlenecks within the network; any place where higher-speed links converge into an area of the network serviced by a slower-speed link is a potential

traffic jam. So while additional bandwidth might help with calls traversing a certain path, the question is: Does it also introduce new problem areas, where slower links are still traversed for part of the VoIP call path?

Bandwidth Doesn't Help Your IT Budget

Additional bandwidth can be a costly solution to a VoIP call quality issue, especially if it doesn't solve the problem. It's important to proactively manage VoIP call performance, understand the causes of any quality issues, and know the composition of the traffic that's flowing over your network. With better visibility and knowledge, you'll know when additional bandwidth might be an appropriate solution – and when it won't be!

Chapter Summary

In this chapter, we've discussed the concepts that are important for ensuring optimal call quality. Why should you be concerned about call quality performance? The reason is that call quality performance can have a dramatic impact on your overall business. If you can't talk to customers and partners, then it will affect your business. The overall user experience of the phone system is tied closely to the quality of the calls. In order to understand this aspect of user experience, you need to monitor certain call quality metrics very closely:

- MOS
- Latency
- Packet Loss
- Jitter Buffer Loss
- ACOM

Managing call quality performance can go a long way in providing for a successful VoIP deployment. Many enterprises are considering VoIP as the stepping stone for other Unified Communications applications like video, presence, and instant messaging. Getting VoIP right is a crucial first step down the road to true unified communications.

In the next chapter of this ebook, we'll examine the path to Unified Communications and consider:

- What is Unified Communications?
- What are the new Unified Communications applications?
- How will Unified Communications affect network performance?
- How can you manage Unified Communications applications?

Unified Communications offers the vision of great productivity gains as integrated multi-modal communications are built into our most commonly used business applications. But any time new applications are added to the network, it's always good to take a step back and analyze the potential impact on network performance.