

Is Your WAN Ready For Voice?

John Bartlett

I returned from VoiceCon in San Francisco a few weeks ago, where we heard about how enterprises are continuing to roll out voice, and in some cases video communications across their IP networks. Many enterprises started their voice over IP (VOIP) deployments small to test the technology and learn how to manage it. These small deployments were often isolated to headquarters or to a new branch office. Now that enterprises are gaining confidence, VOIP is being extended across the wide area network (WAN).

I have been moderating a session at VoiceCon for 2 years now discussing how to design a voice-ready network, and this year in San Francisco we did a second session focused on the WAN. A number of key issues have consistently been identified as critical to ensuring proper support of voice and video in the WAN as follows:

- Quality of service (QoS)
- Bandwidth management
 - Dedicated versus converged WAN links
 - Determining call volume and needed bandwidth
 - Managing bandwidth—Call admission control (CAC)
- Network resiliency
- Voice security
- Remember the mobile worker
- Monitoring and quality of experience (QoE)

Quality Of Service (QoS)

Different types of applications require different treatment by the network. This is not because some applications are more important to the organization than others, though undoubtedly this is true. It is because different types of applications

are more susceptible to packet loss and jitter than others, and need better treatment in order to work well. This is like needing high-octane gasoline in a high-performance engine; the engine doesn't work well on regular gas.

There are important considerations for those applications that are, in fact, more important to the organization, but those will be dealt with in the next section, bandwidth management.

The wide area network (WAN) is the first place where QoS should be deployed. QoS provides priority to some packet streams over others, when there is queuing or congestion in the network. This congestion occurs most frequently where traffic is flowing from high-speed connections into lower-speed ones, like at the boundary between the LAN and the WAN access link. Just like the slowdown on the highway when one lane is closed, packet traffic backs up in a queue at the WAN boundary, because the WAN access link, due to cost considerations, is usually the lowest-bandwidth link in the network path.

Voice over IP (VOIP) is one of the most sensitive applications to packet loss and jitter. Voice quality degrades quickly as the network starts dropping packets. We have high expectations for the quality of our voice connections, so users notice the quality drop, and complain.

Voice traffic should be prioritized as the highest-priority application on the WAN link, followed by interactive videoconferencing, interactive data applications, and finally bulk data applications. The IETF has recently completed an RFC with guidelines for classifying applications into DiffServ categories, RFC-4594. This document gives descriptions for each application category to help users determine how to assign DiffServ priorities to each enterprise application.

A common misconception about QoS is that it is only needed if the link is heavily utilized. Thus, goes the argument, if your links are only 50 percent utilized, QoS never comes into play, and the added complexity of configuring QoS is not worthwhile.

The problem here is one of defining the appropriate time scales for measuring link utilization. Congestion often occurs only for very short periods of time. If link

utilization is being measured over a day, an hour or even five minutes, momentary congestion might be averaged out and the network technician may see average utilization numbers that are quite low. However, during the short congestion period, the link was 100 percent utilized, and traffic was backed up in the output queues.

During these brief events, voice packets may be delayed (which causes jitter), or dropped due to an overfull queue (packet loss), causing voice degradation. If QoS is deployed, it should solve these momentary congestion loss problems and maintain the voice quality throughout the voice call.

Bandwidth Management: Dedicated vs. Converged Links

A number of aspects to bandwidth management are important to the WAN. The first is the mix of real-time traffic (voice and video) with data traffic on the link. Data traffic is bursty in nature, often creating peaks of demand that are 10 times higher than its average utilization rate.

Traditional LAN metrics have assumed that a link is "full" when its average utilization is about 35 percent. The large bursts of utilization we see in data traffic are required to achieve good application performance. It is during those peaks of data movement that our users are waiting for their screens to update. Further loading of the link begins to slow down the data applications, again causing users to complain.

When we mix voice or video (i.e., real-time traffic) with data traffic, we use a priority mechanism (QoS) as described above to ensure the voice or video are transported without loss or jitter. But every time a real-time packet is prioritized over a data packet, the data application is slowed down a little bit. If the real time traffic starts to exceed 35 percent of the link capacity, the priority mechanism does not work as well, and the data traffic begins to suffer significant degradation.

Thus, the rule of thumb for a converged network is to not let the real-time traffic exceed 35 percent of the total link capacity. So a T1 link should be limited to about 500 kbps of voice or video, and a 10-Mbps metro Ethernet link should be limited to 3.5 Mbps voice or video.

John Bartlett is a consultant and VP with NetForecast (www.netforecast.com), specializing in data and real-time application performance on enterprise networks and the Internet. He can be reached at john@netforecast.com

One option chosen by enterprises that need significant amounts of real-time traffic support in their WAN, is to use an overlay, or a separate WAN link or service that is dedicated to real-time traffic. Links that carry only real-time traffic can be loaded to much higher utilization levels without adverse affects.

The reason this works is because the real-time traffic is well behaved, meaning that it does not have the same bursty nature as data traffic. Real-time traffic arrives at router queues with a much more predictable Poisson distribution which queues can handle without packet loss, and with minimal jitter. It is possible to utilize dedicated links at 90 percent capacity without problems.

Bandwidth Management—Bandwidth Demand

The second bandwidth problem to understand and manage is the bandwidth demand. Each simultaneous voice call added to an individual WAN link will increase the real-time traffic flow in a linear manner. The network needs to be designed to handle the expected demand, so the first task is to accurately predict the demand so the network can be properly configured.

The bandwidth required to support voice calls is determined by the number of simultaneous calls and by the codec used.

The number of simultaneous voice calls can be estimated using Erlangs (www.erlang.com). This calculation makes assumptions about the average call length, and allows input on what percentage of call attempts it is OK to block (present with a busy signal), and then generates the number of “lines,” or simultaneous calls that have to be supported.

Choosing the codec also has a significant effect on the bandwidth required. The tradeoff is between call quality and the amount of bandwidth used. The G.711 codec has the best call quality, and it degrades slowly when packet loss is introduced. However it has the highest bandwidth consumption, using about 80 kbps per voice stream. G.729, the other common choice, is used most often across WAN links. G.729 provides lower overall voice quality, but uses considerably less bandwidth, consuming only 24 kbps per voice stream.

Videoconferencing streams consume

much higher amounts of bandwidth, and often are in a call for much longer periods of time. Videoconferencing endpoints can be configured to use 384 kbps, 512 kbps or 768 kbps for a call. Newer high definition (HD) video conferencing units use 1 Mbps, 2 Mbps or even 4 Mbps. All of these bandwidth numbers have to then be increased by about 20 percent to account for the IP overhead, so a 384-kbps video call actually consumes close to 460 kbps, and a 4-Mbps call consumes almost 5 Mbps of network bandwidth.

The demand numbers generated by determining how many calls will simultaneously take place must be mapped to the network topology, to determine which links will be required to carry the traffic, and how much traffic will be on each link. This information is then used by the network team to properly size those links, and to program the bandwidth limits for the QOS deployment.

Bandwidth Management—Call Admission Control

The third bandwidth issue is call admission control. When the network is configured for QOS, the configuration will include a bandwidth limit for the high-priority voice or video class. This bandwidth limit will be different for each major WAN link, and will be related both to the demand and to the total link capacity, as described above.

The bandwidth limit is needed by the network to protect itself from overuse by the high priority classes. If a large amount of high-priority traffic is suddenly introduced to the network, lower-priority traffic could be choked off, causing important data applications to fail. Limiting high-priority traffic to a predetermined bandwidth limit prevents this problem.

What this means to the high priority traffic is that total demand must remain within the capacity defined for each link. If traffic exceeds a link’s capacity, all packets that use the high-priority class will start to be dropped. In other words, the quality of the voice or video will begin to degrade for all users on that link.

Voice and video systems provide call admission control to help manage this issue. The call manager tracks how many concurrent calls are in progress on each link of the network. By knowing the

bandwidth requirement for a call and the number of calls active at any time, the call manager can determine if a new call will exceed the network capacity for high priority traffic on that link. If the limit is exceeded, the call can be denied (busy signal), rerouted via the Public Switched Telephone Network (PSTN), or remarked as lower-priority traffic and sent best-effort.

Resiliency

The very high reliability of the PSTN has created a high expectation for the reliability of all voice communications networks (at least landline networks), and users are disappointed if these expectations are not met. Data networks have not traditionally met the high availability standards of the PSTN, so as we move our voice onto the “data” network, we need to address availability as well.

A simple way to provide resiliency for voice is to provide for a local PSTN connection at each remote office location. If the WAN link to headquarters fails, the remote office can still make phone calls via the PSTN. Not all the features of the IP-based voice system will be available when using the PSTN, and there may be limits on how many calls can be placed because of the size of the PSTN connection, but at least voice connectivity will be maintained.

In the past, WAN resiliency was obtained by using multiple dedicated or frame relay links from each branch office to other sites. If a mesh or ring configuration is built, then the failure of an individual link does not necessarily isolate an office, since the second or third path is still available.

Today many enterprises are moving toward using a single service provider for office connectivity, through an MPLS cloud. The service provider now worries about the resiliency of their cloud, but a single access link to each office can become the point of failure. This can be overcome by provisioning two access links, preferably to separate service provider points of presence (PoPs), and preferably via separate physical paths out of the branch office. Physical redundancy is intended to prevent simultaneous failure due to physical interference, such as a cable being dug up by a backhoe or a car hitting a telephone pole.

Some enterprises go a step further and contract with two separate service

providers, each of whom provides an access link to each enterprise office. Using this approach, a failure within a service provider's MPLS cloud can still be managed by diverting traffic to the second provider until service with the first is restored.

Security

As IP-based voice and video systems become more prevalent, they become targets for attacks. Thus, security is becoming an important aspect of deployment. Three aspects of security should be considered:

- Security of the voice/video equipment
- Security of the conversations/meetings
- Maintaining network security as voice/video calls cross the firewall

Equipment Security: Administrative rights for the call manager, endpoint configurations and other infrastructure components should be limited to trusted administrators. By limiting access for configuration or software upgrades, you limit the opportunities for malicious or accidental changes that can affect voice quality or availability.

Conversation Security: Encryption of voice streams is the best way to achieve conversation security. IP-based voice provides many more opportunities for a third party to capture packets and determine call patterns, or even reassemble voice calls and hear the conversation. Encryption across the WAN can provide security against this threat.

Network Security: Traditional firewalls open ports for data traffic that is initiated on the trusted side of the firewall, but bar traffic initiated on the un-trusted side from traversing the firewall boundary. Voice and video systems must establish traffic streams in each direction, so if there is a firewall in the path, one of those

streams will not be able to cross. Either a session border controller (SBC) or firewalls that explicitly understand the SIP or H.323 protocol must be used to allow appropriate voice and video traffic to cross the firewall without reducing network security.

Remember The Mobile Worker

Mobility is one of the key features provided by IP-based telephony. Employees who often work at home, travel to other offices or connect from hotel rooms or WiFi hotspots can benefit greatly from the flexibility that IP-telephony provides.

To support these wandering workers, we should include them in the network plan. Security can be addressed through a corporate VPN and/or an SBC or voice-aware firewall. Bandwidth can be addressed by ensuring that sufficient Internet access bandwidth is available for the expected call volume from traveling workers.

Quality of service is often not possible for the remote worker, because the Internet does not yet provide QoS capabilities. Quality can be improved by giving priority to outgoing voice packets, and by providing more than enough bandwidth. These steps will help avoid packet loss caused by the Internet access link. But if there are problems deeper in the Internet, users have to just try again or wait for a less congested time to make their calls.

Network And Experience Monitoring

Last but not least is measurement and monitoring. When we deploy voice and/or video traffic on our data networks, we are introducing a very new type of traffic. Voice and video use UDP streams instead of TCP streams, and they require all the special treatment described above.

Because our networks have not carried this type of traffic in the past, most network teams do not have the right kind of monitoring equipment in place to ensure that the traffic is being carried correctly.

Test tools that measure the ability of the network to carry packets with low loss and low jitter across the network from end-to-end are required. Without tools that test across the network, rather than looking at the packet loss at a specific router or switch, the network engineer cannot determine if the network is the cause of a user complaint, or if the voice equipment itself is failing.

Newer tools are now becoming available that take this testing further up the application stack, and test not only the ability of the network to support low loss and jitter, but also test the quality of the voice or video being delivered. This testing, often referred to as quality of experience, or QOE, focuses more closely (as the name suggests) on the user experience. Testing QOE allows the network engineer to track voice quality over time, notice trends towards degradation in parts of the network and resolve issues before they become user complaints.

Conclusions

Supporting voice and video on the WAN requires some careful attention to the new traffic type and its demands on the network. Some of these changes require a new understanding within the IT department about how the network affects application performance, and how to balance limited resources to provide the best functionality and user experience possible. As IT teams learn how to manage this new traffic, it will become an integral part of the network and the daily work process, and will soon not be the challenge it initially presents □