

Interested in Voice over IP?

How to proceed

Is the time right to make the move to Voice-over-IP (VoIP)?

There's an easy way to tell. If your monthly recurring telecom costs exceed \$10,000, get to work. The success of cellular phone service proved that consumers would accept unreliable and less-than-perfect voice services, and with IP traffic entering nearly half the businesses and households in the U.S., the opportunity is just too large to ignore.

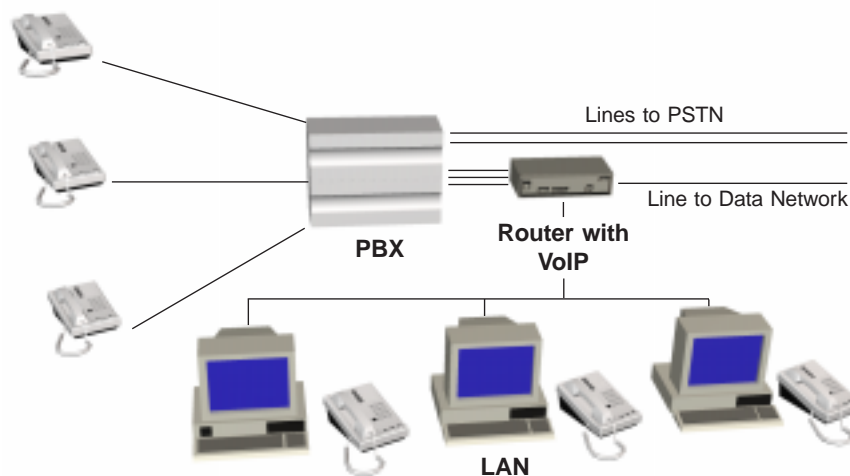
However, there's no "one-size-fits-all" solution, and there's still uncertainty about VoIP equipment. But the ITU is moving fast, even reusing existing IETF standards to get the job done and issuing pre-announcements of coming standards, and products have been rolling out at a furious pace. Vendors are adding voice to existing devices, acquiring the technology from someone else, or building it from scratch.

There are the usual crowd—such as Cisco, Lucent, and Nortel—but the excitement has stirred many others to action. A few have been at this game for a while, selling voice over other transports, such as frame relay. This may or may not give them a leg up on the competition, as the difficult part of the game changes radically with an unreliable transport.

Curiously, it's transport, the ubiquitous IP, that is driving the excitement. So how do you create a short list? Start with the equipment that you already have in your network, because you may be able to upgrade existing gear to add voice-processing capability. Even if an upgrade doesn't pan out, you still have the advantage of a familiar partner when venturing into these uncharted waters. Watch out for potholes in the upgrade route though; it's going to require more than just digitizing voice to provide acceptable voice quality. You're going to need more processing power and also more memory.

The other candidates for your short list are likely to become obvious when you look at the markets that the vendors are targeting. While the major vendors have several different solutions, each aimed at a particular sized organization, most of the newer second-tier vendors have chosen a model where they build for a particular size organization.

Factor in standards compliance by becoming familiar with the H.323 family. While this is an evolving standard, it's evolving very fast, and it's clearly the target. If a vendor can't tell you how they are going to work with these standards, be sure to factor in early retirement for the equipment they provide. This isn't to say that you should be overly concerned with standards, since you should be looking at a payoff before the standards have even reached full ratification.



The easy and obvious place to put this equipment is between your PBX and your router (or integrated inside the router), and reap the monthly savings on the separate leased lines or switched calls. But, it's worth considering bringing voice IP packets right to the desktop. While it may not be an obvious business case unless you're about to change your voice switch, it will put you in a great position to take advantage of either data sharing or video, when the time comes. There are suppliers out there now—NBX Corp. (acquired by 3Com), Selsius (acquired by Cisco), Shoreline Teleworks, and Touchwave Inc., for example—ready to let you share your existing LAN wire plant between data and voice.

Where to Start?

Start by characterizing the existing data traffic on your network—how much is there, and how bursty is it—in other words, establish a good baseline. You also need to get a good reading on how much voice traffic currently exists in the parts of your organization that you are considering moving to VoIP.

If your datacom and telecom services aren't currently managed by the same people, this could be a revealing exercise. The data folks may be surprised to find that the voice folks have tools and techniques to characterize their traffic well. If they are doing almost any management of the voice network, they will have all the important baseline numbers, things like:

- Average number of calls per hour.
- When busy times occur in a day, month, and year.
- The peak number of calls in the busy hour.
- Average duration of the calls.
- Acceptable ratio of blocked to completed calls.

This is pretty much the same information as for the data network, with the glaring difference that for telecom you know exactly how much bandwidth each call will use. This won't change with VoIP, except that the fixed amount should be reduced considerably.

Ferguson and Huston, in their book *Quality of Service*, classify network applications into three useful categories: elastic, intolerant real-time, and tolerant real-time. Elastic applications—like HTTP, FTP, and most business processes—were written to run as fast as the network allows. The usual brake pedal for these applications is the network itself. As the network becomes congested, acknowledgments are delayed, the sender waits and the flow slows down. Tolerant real-time applications are typically streaming video or audio which are seen or heard after some delay, because the receiver buffers the incoming data to damp network variations.

Voice traffic falls into the intolerant real-time category. Humans are intolerant of speech delays of more than about 200 milliseconds (ms). The dilemma is that while elastic applications can tolerate a fair amount of delay, they usually try to consume every bit of network capacity they can. In contrast, voice traffic only needs small amounts of the network, but wants that small amount to be available immediately. By the time a network is congested, it's too late to be adding intolerant real-time applications to the mix.

Most of the attention has been focused on implementing VoIP as a wide-area solution. This makes sense considering the amount of money spent there, but we think a LAN solution is worthwhile just to eliminate wire clutter. If you're about to spend money on telephone gear, or setting up or retrofitting an office infrastructure, you have a golden opportunity to do the right thing. Although the quality-of-service (QoS) problem is the same for both WAN and LAN, the solutions are quite different.

WAN

If you're only considering placing voice traffic on wide-area links, IP QoS or other bandwidth management solutions may help in throttling back the peaks of the existing data traffic to allow room for the voice.

Today's standards-based approach to QoS is the Resource Reservation Protocol (RSVP), shipped in most recent routers. A proxy RSVP client in your router can carve out an adequate amount of bandwidth for the voice traffic and provide the low delay forwarding that voice requires. While this only applies directly to routers that you control, RSVP does provide for the ability to pass through other routers. This could be effective in combination with a guaranteed performance service from an ISP. For example, if you have an Internet service that can consistently deliver 100ms latency for some class of traffic, you still have about 100ms available to work with at your ingress and egress points, and RSVP should ensure that you hold to your half of the performance arrangement.

You may hear that RSVP doesn't scale well enough to support VoIP. While this is probably true for networks owned and operated by phone companies, IXCs, ISPs, and so on. However, if properly implemented, RSVP should work quite well for large companies and it certainly should handle the needs of small- to medium-sized organizations.

Also, remember that the H.323 family of standards was designed for LANs with no QoS so probably won't apply to your current WAN configurations.

LAN

Here what's needed is a way to control the data flow from each device attached to the network. Ideally, each application would control its own rate of sending into the network. The problems with this are that there are too many groups developing applications (most with only a vague notion of the problem), and they would need an awareness of how much network bandwidth is available at any point in time. In other words, this isn't going to happen soon.

Another solution would be for the operating system on each host to control the rate of traffic being injected into the network. This, at least, has possibilities; in fact Microsoft is talking about this capability as part of their ADSI (Active Directory Service Interface), so it may be real someday.

In the here and now, the closest you will get to a standards-based solution is IEEE 802.1q. Combined with 802.1p, this could even be an end-to-end solution, but the requirement that all your network gear conform to the standard makes it appropriate only when all the NICs and switches are being changed at once. There are other approaches; for example, layer 4 switches with implicit QoS capability. With switching to every device, this should provide the bandwidth control, but you will be giving up the standards battle.

The answers are out there and the equipment manufacturers are eager to give them to you. Take the time to look at several and test them thoroughly. You need to simulate, at the very least, the heaviest traffic conditions that you see in your production network, with a similar mix of disparate applications and protocols.

Begin Planning-Now

Once you have a handle on the traffic in your network (both voice and data) and you have decided whom to bring to the dance, it's time to get into the details. There are many technical considerations, and they will be different for each installation. As a sampling, these seem to come up regularly (and rarely occur to folks in advance).

MTU sizes and low latency

The usual maximum frame size on a data network is 1,500 bytes. This is reasonable for elastic applications, but at 56 kbps, 1,500 bytes takes longer than our total latency goal of 200ms just to put on the wire. You will have to set up a smaller maximum transmission unit (MTU) size to be forwarded through the lower speed links. While this will cost a small amount of data efficiency, you shouldn't need to change anything other than a line of router configuration. Current Microsoft Windows TCP/IP stacks routinely do "black hole discovery" and adjust their MTU accordingly.

Bandwidth usage during silence

In most phone calls, one person is talking while the other is listening. It would be nice to recover the bandwidth that is being chewed up in one direction while you patiently wait your turn to talk. This is called "silence suppression" or VAD (voice activation detection), and turns out to be one of the tougher pieces of the puzzle. You may want to try it in a solution that you're considering, but for planning purposes, assume that you will use the bandwidth required full time.

Recovery speed

Reconnect times during a circuit-switch failure are measured in milliseconds, and may not even be perceptible to the user. In contrast, route failures in an IP network may take many seconds to re-converge, if there is even an alternate route available. Don't assume you'll be able to configure QoS capabilities on an ISDN or dial backup line. It may be easier, and a lot less risky, to provide a reduced "plain old telephone service" to back up the new VoIP service.

Modeling and Testing

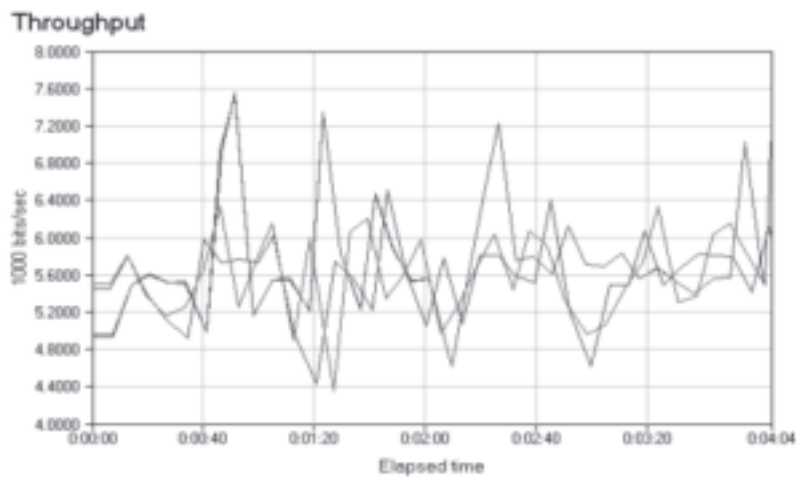
If you have a network-modeling tool to predict the effect of changes in the network, now is a good time to crank it up. The numbers that need to be input to the new model include the reduction in available bandwidth for data applications, with the amount varying by the number of simultaneous calls; that number, in turn, will vary over time. Watch out for seasonal peaks in your business.

Use a reduced MTU size for the slower-speed links. Add in the control traffic associated with the voice IP packets-both the RSVP flows and the direct voice measurement flows, such as the real-time control protocol (RTCP) flows. It may be difficult to predict the performance impact on your routers of setting up and managing the QoS functions.

Note however, if you're not already familiar with the modeling tools, this is probably not a good time to decide to get into it. They can take a substantial amount of time to get comfortable with, and a lot of their utility is derived from experience: using the tool over time and feeding results back into each subsequent use.

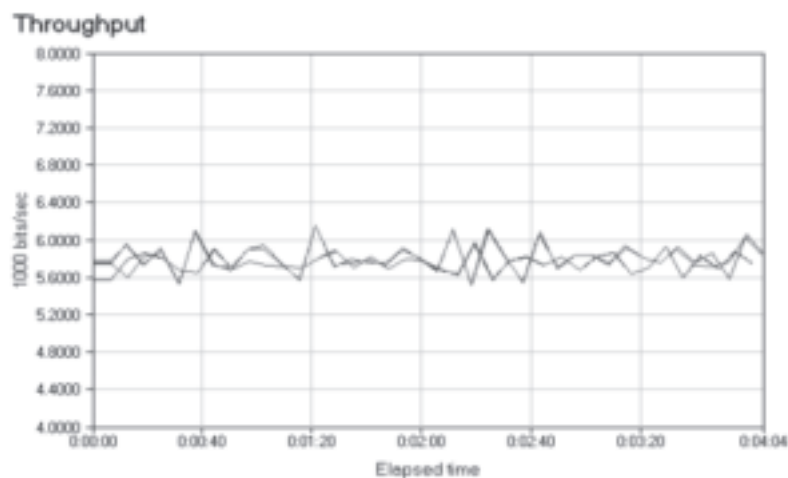
Whether you plan to use your existing equipment with add-on functionality or do a major upgrade, be sure to test the performance of whatever solutions you consider. You need to find out how well voice traffic behaves, and the effect on response times of your critical applications.

We used Chariot, by Ganymede Software, to test whether two G.723 compressed-voice conversations could maintain a constant enough rate across a 64Kbps leased line, in the presence of other traffic. We added a persistent TCP file transfer in both directions through the same pair of connected computers. Other traffic was only allowed during window acknowledgments. There were a pair of routers on the path between the two end computers.



Our expectation was that the file-transfer traffic would seriously disrupt the flow of the simulated voice traffic. As the following graph shows, we were not disappointed. The graph has four lines, showing the throughput for each of the four UDP connections between two computers—one in each direction for the two telephone conversations. (The background TCP file-transfer traffic is not shown.) While striving for an average throughput of 5.7kbps, the conversations experience nearly 50% throughput variation during the 4-minute measurement period.

Without Weighted Fair Queuing, the Chariot test of four 5.7 kbps UDP connections showed substantial variation in the network throughput over four minutes.



We then enabled Weighted Fair Queuing on the router interfaces, and re-ran the same test with no other changes. We expected this to be an intermediate step on the way to using RSVP, but were surprised that we achieved these perfectly adequate results (that is, low variation) simply with Weighted Fair Queuing. The throughput variation was less than 10% during the second run of the 4-minute test.

With Weighted Fair Queuing enabled in the intermediate routers, the throughput variation was much lower for the four connections. A minor configuration change gave a rather dramatic improvement in throughput variation and, therefore, perceived voice quality.

Conclusion

The technology is here to save a bundle on wide-area telecom costs, and to do it in a standards-based way. Industry experts are saying that the underlying costs of moving voice over IP packets is anywhere from 30-300 percent less expensive than traditional circuit switching. Even at the low end, that could be a lot of savings.

Going all the way to the desktop could be tough right now, but it's going to get easier. If you really have the commitment to managing your LAN resources-let's say switched Ethernet to the desktop for starters-you can do it. If you do, you're going to be nicely prepared for the next level of convergence.

Being able to treat voice as another, albeit critical, application on the data network is the really big benefit. At least some of the savings should go into creating a healthier data network infrastructure. The key to making it work is testing-test your current network before you start, test each change as you put it in place, and test after you've put it all together. Knowing what your network performance will look like after you converge will put a smile on your face.

About the authors

Daniel J. McCullough has held a variety of positions in Information Technology, both as a supplier and a consumer, over the past 22 years. He is currently a Product Marketing Engineer for Cisco Systems Inc. in RTP, NC.

John Q. Walker is the vice president of product development and a founder of Ganymede Software Inc. in RTP, NC. He holds a Ph.D. in computer science from the University of North Carolina.

An initial version of this paper appeared in Voice 2000, a supplement to Business Communications Review (BCR), April 1999, pages 16-22.

About Ganymede Software

Ganymede Software Inc. is the first to introduce a solution for enterprise performance management, which will enable IT departments to deliver reliable, predictable application performance. With over 700 customers, Ganymede is the recognized leader in end-to-end network performance management. Ganymede was recently named by *PC Week* as one of 15 top companies to watch in 1999. Ganymede's Chariot network testing system is used by all major independent test labs and has won numerous awards, including *Data Communications'* Hot Products award and *Network World's* World Class award. Pegasus™ monitors end-to-end network performance on an ongoing basis and won a New Product Achievement award at ComNet. Ganymede Software is based in Morrisville, North Carolina and can be reached at 919-469-0997. On the Internet, visit <http://www.Ganymede.com>