Gen2 Ventures

White Paper

Clearing the Way for VoIP

An Alternative to Expensive WAN Upgrades

Executive Overview

Enterprises have traditionally maintained separate networks for their voice and data traffic. Their circuit-switched voice networks provide the controlled environment needed for high-quality voice conversations while packet-switched IP networks deliver the flexibility and low-cost bandwidth needed to support ever-changing data requirements.

But the days of separate voice and data networks are numbered. Driven by new voice over IP (VoIP) technologies and the need to reduce network costs, many enterprises are designing new converged networks capable of handling both voice and data. The key challenge is to reconcile the performance requirements of voice with the unpredictable nature of data on a single network.

This white paper analyzes VoIP performance and bandwidth requirements and shows how Expand Networks ACCELERATORs help to deliver the required performance while reducing WAN costs in converged networks.

The VoIP Performance Challenge

The motivation for running voice over IP networks is to eliminate the expense of maintaining separate voice and data networks. It sounds easy enough to run voice over IP network – just encapsulate digitized voice in IP packets and go. Digitizing and packetizing voice is fairly straightforward, but there's one other key issue that is much tougher to deal with.

The key challenge in building converged networks is performance. Voice communications has much more stringent performance requirements than data communications. The best way to understand voice performance requirements is to analyze the traditional voice communications network – the public switched telephone network (PSTN).

First, with the exception of most "last mile" copper loops, the PSTN is a digital network. Since human voice is analog, voice traffic must be digitized before it enters the network and then converted back to analog on the receiving end. Pairs of codecs (coder/decoders) at the endpoints perform the conversions between analog and digital signals.

To provide high quality voice the codecs use a technique called Pulse Code Modulation (PCM) that samples analog voice every 125 microseconds (1/8,000 of a second) and digitally encodes each sample as an 8-bit code. Since 8,000 of these 8-bit samples must be transmitted every second, PCM requires 64 kbps of bandwidth for each call.

To ensure the quality of each call, the PSTN uses multiplexing and circuit-switching technology to allocate a fixed 64 kbps channel for the duration of each call. Since the required bandwidth is always available, there is very little end-to-end latency, no jitter (variation in latency), and virtually no data loss. The net result is a consistently high level of voice quality, called toll quality.

Gen2 Ventures
5339 Prospect Road #321
San Jose, CA 95129

(408) 446-2277
www.gen2ventures.com

Toll quality is the standard of comparison for VoIP because we all take the performance of the PSTN for granted and expect similar quality from any new technology that attempts to take its place.

Voice over IP – Mixing Oil and Water?

To run voice traffic over IP networks it is first digitized and packetized. Digitized audio streams are transported between endpoints by the real-time protocol (RTP). RTP is a connection-oriented end-to-end protocol that is designed to transport delay-sensitive information. RTP identifies the encapsulated payload type and includes sequence numbers and time stamps that are use to synchronize real-time information flows. RTP uses the connectionless, unreliable user datagram protocol (UDP) transport protocols because retransmission of lost or corrupted data disrupts real-time audio streams.

Delivering high quality voice communications over IP networks is a challenge because these networks have none of the characteristics that enable the PSTN to provide toll quality voice service. Unlike the PSTN, IP networks use packet switching rather than circuit switching technology. Packet switching works well for data because it maximizes bandwidth utilization by allowing all users to dynamically share network bandwidth.

The downside of IP's dynamic resource sharing is that it provides only a best-effort delivery service which does not guarantee the performance levels of specific traffic flows such as voice conversations. To overcome these IP performance limitations enterprises are beginning to employ bandwidth management techniques such as prioritization to ensure that critical applications get the performance they need.

But bandwidth management alone simply allocates bandwidth to critical applications at the expense of other applications, many of which are also important to the enterprise. Similarly, just adding more bandwidth is usually ineffective because any additional bandwidth will be consumed by the most aggressive applications, not the most important ones.

What is needed, particularly on WANs where bandwidth is scarce and expensive, is a combination of adequate bandwidth and the ability to manage that bandwidth. Since VoIP consumes predictable amounts of bandwidth for each call in progress one of the first tasks in planning for VoIP is to determine the amount of bandwidth needed for the number of active calls to be supported by the network.

VoIP Bandwidth Requirements

VoIP creates two types of network traffic – the call control messages used to setup and manage connections between users, and the digitally encoded voice conversations. The call setup and management protocols involve simple messaging between IP phones and an IP PBX. These protocols use very little bandwidth and they do not have stringent latency requirements. A delay of a few seconds in setting up a call is usually acceptable.

The real challenge is to satisfy the bandwidth demands of the digitized voice streams between users. Each call consumes a nearly constant amount of bandwidth for the duration of the call. How much bandwidth is needed for each call? That depends primarily on the voice encoding technique used as well as a couple of other variables.

Two voice encoding standards are widely supported by VoIP products. The first is the G.711 standard that uses the same PCM encoding used on the PSTN at a bit rate of 64 kbps. In contrast to the PSTN approach of sending 8-bit PCM voice samples at 125 microseconds

intervals, G.711 packs multiple samples into each IP packet sent. Packing multiple PCM voice samples into a single IP packet reduces packet header overhead. Each VoIP packet is made up of IP/UDP/RTP headers in addition to the voice sample payload. Because these headers total 40 bytes per packet it is important to minimize the total number of packets sent. The maximum payload sizes are limited by the encoding latency as payload size is increased. G.711 payloads are usually limited to 160 bytes (20 ms. of voice) or 240 bytes (30 ms. of voice) because larger payloads would increase the encoding latency beyond acceptable limits and cause perceptible delays in conversations.

G.729 is another widely supported voice encoding standard. G.729 encodes voice at a bit rate of 8 kbps by compressing as well as digitizing the voice signals. This compression is lossy and can degrade voice quality compared to G.711 encoding. The payloads of G.729 packets are typically 20 or 40 bytes.

Although G.711 and G.720 encode voice at bit rates of 64 kbps and 8 kbps respectively, the actually link bandwidth consumed is greater because of the IP/UDP/RTP packet header overhead. The actual link bandwidth requirements for G.711 and G729 are:

G.711 with 160 byte payloads	83 kbps
G.711 with 240 byte payloads	76 kbps
G.729 with 20 byte payloads	26.4 kbps
G.729 with 40 byte payloads	17.2 kbps

Link bandwidth requirements can be reduced for all encoding schemes by using a technique called RTP Header Compression (cRTP). cRTP operates hop-by-hop and compresses the 40 byte IP/UDP/RTP headers to 2 or 4 bytes. Link bandwidth requirements when using cRTP are:

G.711 with 160 byte payloads	68 kbps
G.711 with 240 byte payloads	66 kbps
G.729 with 20 byte payloads	11.2 kbps
G.729 with 40 byte payloads	9.6 kbps

Another technique, called Voice Activity Detection (VAD) can further reduce link bandwidth requirements by detecting periods of silence in conversations and preventing packets of silence from being sent. VAD works with all encoding standards and can typically reduce the per call traffic volume by about one third, but its statistical nature means that actual link bandwidth requirements are reduced only in situations where a large number of VoIP calls share a link.

VoIP Performance Requirements

VoIP has three specific performance requirements that have to be met in order to provide toll quality voice conversations. The first is end-to-end latency. Anyone who has ever tried to carry on a conversation over a satellite link knows how excessive latency impacts quality. Long delays make it difficult for callers to determine when the person at the other end has finished talking. This results in very unnatural speech patterns.

How much latency is too much? A rule of thumb is that one-way latency should not exceed 150 milliseconds. 150 millisecond delays are noticeable, but when latency exceeds 250 milliseconds it becomes difficult to carry on a conversation. Latency is a non-issue on the PSTN, but delays on IP networks can easily cause latency to exceed 150 milliseconds.

End-to-end latency is the sum of encoding/decoding latency and transmission latency. The level of compression provided by the codec is proportional to the encoding/decoding latency it introduces. For example, G.711 performs no compression and adds negligible latency while G.729 codecs compress voice to 8 kbps but add a one-way delay of about 25 ms.

More significant delays can occur when voice packets are transmitted across a network, particularly when low speed WAN links are involved. The following chart shows the latency that results when voice packets get “stuck” behind data packets of different sizes being sent over WAN links.

	64 kbps	**128 kbps**	**1.5 Mbps**	**2 Mbps**
1,500 Bytes	188 ms	93 ms	8 ms	6 ms
1,000 Bytes	125 ms	62 ms	5 ms	4 ms
500 Bytes	63 ms	31 ms	3 ms	2 ms
250 Bytes	31 ms	16 ms	2 ms	1 ms

On T1/E1 and faster links this latency is only a small fraction of the total one-way latency budget of 150 ms., but on low-speed links the situation is very different. A single 1,500 byte packet on a 64 kbps link will push the latency beyond the 150 ms mark and even on a 128 kbps link, nearly two thirds of the total delay budget is consumed by just the transmission delay.

This problem is compounded by the fact that compressed voice formats such as G.729 are more likely to be used over low-speed WAN links, and these algorithms contribute their own latency to the total end-to-end delay.

Even when voice packets are not blocked by data packets they are subject to their own serialization delay – the amount of time that is takes to clock the bits onto a serial link. Again, this delay is determined by packet size and link speed. Reductions in packet size result in less serialization delay and therefore, lower end-to-end latency.

Another key performance metric is jitter. Jitter is the amount of variation in latency that is experienced over time. IP phones have some ability to buffer incoming audio streams to compensate for jitter, but excessive jitter can disrupt conversations. Again, the PSTN has virtually no latency and therefore no jitter, but enterprise IP networks are subject to jitter caused by congestion on LANs and WANs and by packet buffering in routers and other network devices.

The third important performance metric is packet loss. Since VoIP is a real-time audio service that uses UDP transport protocols, there is no way to recover lost packets. Packet loss can result in a metallic sound or dropouts in conversations that can be very frustrating to users. The PSTN experiences virtually no loss of digitized voice, but IP networks routinely experience packet loss due primarily to congestion.

The key to meeting all of the VoIP performance requirements is adequate bandwidth, and the simplest solution is to throw bandwidth at the problem. This approach is being used successfully on enterprise LANs that have been upgraded to switched 100 Mbps and gigabit Ethernet. The real challenge is the wide area network.

Private WAN facilities such as frame relay and private lines are very expensive and as a result most enterprise still have very limited bandwidth between their headquarters and their remote offices. Half of all WAN links between corporate headquarters and remote offices are 56kbps/64kbps or lower. Most other remote offices operate at speeds of 128kbps to 512kbps and fewer than 10% are T1/E1 or greater.

How Expand's ACCELERATORs Meet VoIP Performance Requirements

Expand's ACCELERATOR product line can help enterprises address the performance requirements of all enterprise applications including VoIP. First, ACCELERATORs change the economics of wide area networking by squeezing an average of 100% - 400% more bandwidth with peaks of 1000% depending on traffic mix. This frees up link bandwidth to support high quality VoIP services – and it does it without expensive WAN upgrades. It is also important to note that ACCELERATORs do not use lossy compression schemes that might degrade voice quality and they add less that one millisecond of latency.

In fact, the ACCELERATOR's compression actually reduces end-to-end latency by reducing serialization delays on WAN links. For example, it takes 125 ms to serialize a 1,000 byte packet on a 64 kbps link, but if an ACCELERATOR increases the effective bandwidth by 4X to 256 kbps, the serialization delay is reduced by a factor of four to 31 ms. The following formula can be used to calculate the serialization delay for any combination of packet size and link speed:

Packet Size (in bytes) x 8 / Link Speed (in kbps) = Serialization Delay (in ms)

In addition to freeing up bandwidth normally consumed by data applications, ACCELERATORs are able to reduce WAN bandwidth requirements for different VoIP codecs. In tests that are described later in this paper, ACCELERATORs reduced G.711 bandwidth requirements by 20% and G.729 by 70%. As a result, WAN links can carry more simultaneous voice calls and the performance of other applications may also be improved.

ACCELERATORs solve increased jitter and latency caused by large data packets over slow WAN links by fragmenting large data packets and injecting VoIP packets at regular intervals. This feature allows VoIP and data to co-exist even on branch office WAN links. For example, normally, a VoIP packet "stuck" behind a 1,500 byte packet on a 64kbps lin will be delayed by 188ms (see table on page 3). Using the ACCELERATOR's packet fragmentation will result in the data packet being reduced in size (accelerated - say from 1500 bytes to 500b bytes) and then fragmented into smaller data packets (say - 2 packets of 250 bytes each). In this case, the latency for the VoIP packet will go down from 188 ms to 31ms!

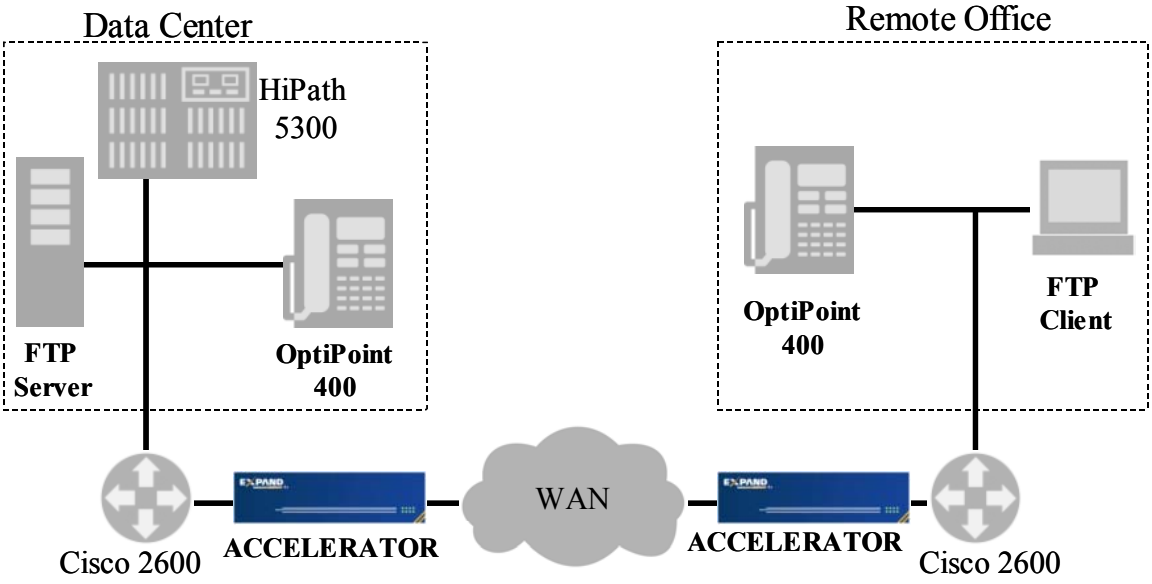
In addition to increasing WAN capacity for both data and VoIP while reducing latency and jitter, ACCELERATORs also manage WAN bandwidth to ensure that critical applications like VoIP get the bandwidth they need. Expand's ACCELERATORs include an Instant QoS feature that prioritizes application access to WAN bandwidth. Without such prioritization, the additional effective bandwidth provided by ACCELERATORs could be consumed by aggressive, non-critical applications such as file sharing.

ACCELERATOR's AppView feature provides graphical visibility for all application traffic sharing a link. AppView can be used to monitor WAN utilization and to plan future capacity requirements.

And finally, ACCELERATORs have a set of data integrity features that are designed to stop the packet loss that can degrade voice quality. A flow control mechanism reduces packet loss caused by link congestion and a packet recovery feature ensures that any lost packets are transparently recovered at the link level before they can cause voice quality problems.

Validating ACCELERATOR Performance in a VoIP Environment

To test the effectiveness of ACCELERATORs in VoIP environments, Expand Networks partnered with Siemens ICN, a leading supplier of IP telephony solutions. They created a test environment based on the Siemens HiPath 5300 VoIP Server and OptiPoint 400 IP Phones. The test configuration is shown in the following diagram.



Various combinations of voice (G.711 & G.729) and FTP data were sent over a simulated WAN running at 128kbps. The objectives of the test were to:

1. Measure the additional bandwidth created by ACCELERATORs
2. Verify that ACCELERATORs add negligible latency and do not have a negative impact on voice quality
3. Verify the effectiveness of the ACCELERATOR queuing mechanisms in prioritizing voice traffic over non real-time protocols

The following table summarizes some of the test results.

Traffic Mix	End-to-End Latency	Acceleration %	Voice Quality
Data only	< 20 msec	356%	N/A
G.711 voice only	< 20 msec	20%	Excellent
G.729 voice only	< 20 msec	70%	Excellent
G.711 voice & data	< 20 msec	139%	Very Good
G.729 voice & data	< 20 msec	308%	Excellent
G.711 voice & data with voice prioritization	< 20 msec	224%	Excellent

Test notes:

1. Acceleration of voice traffic enables WAN links to carry more calls while maintaining voice quality. For example, a link that could previously support four concurrent G.711 calls could

support five calls, and a link that previously supported four concurrent G.729 calls could support seven calls

2. G.729 acceleration percentages are significantly higher than G.711 due to the fact that smaller G.729 packets contain a greater percentage of highly compressible header information.
3. When the ACCELERATOR is configured to prioritize voice over data, the voice quality improved from Very Good to Excellent.
4. Siemens lab personnel based the voice quality ratings on subjective observations.

These tests demonstrate the effectiveness of ACCELERATORs in environments where voice and data traffic coexist on low-speed WAN links. They enable WAN links to handle more voice and data traffic while maintaining service quality for all users.

About Gen2 Ventures

Gen2 Ventures, led by industry veteran Donald Czubek, is a leading analyst firm specializing in emerging technologies that accelerate and manage the performance of networked applications. Focus areas include network acceleration and QoS management, Web server acceleration, and enterprise CDNs. Gen2 Ventures provides research reports, consulting services, and training to vendors, service providers, and enterprise IT clients.