Moving to a Multi-Service Network: Pitfalls and Solutions

By John Bartlett and Peter Sevcik
April 2002

The buzz about convergence has been with us for years, and yet it never quite seems to happen. Why does this buzz persist, when the vision has yet to materialize? There are two reasons to integrate real-time traffic with data traffic: to save money and to increase productivity through powerful new features.

The term "convergence" embodies the cost savings rationale. Integrating data and telephone networks can save money. The idea is that we will have only one converged network to design, install, support and manage. Because resources can be shared between voice and data, the greater capacity data network will carry the voice at low incremental cost. Furthermore, the flat rate billing structure of a data network provides an attractive cost savings for voice users accustomed to per-minute billing.

So why hasn't it happened? There are two major strikes against it. The first is that the technical issues are complex, and good solutions are only just starting to appear. Standardization, integration across carriers, reliability and quality of service (QoS) are all issues that are unresolved. The second impediment is that we have a perfectly serviceable system now, the PSTN, which establishes a high standard of quality and convenience. If it works, don't fix it!

So why continue to struggle against the status quo? We continue to beat the drum for a converged network because the opportunity is so attractive. A converged network by another name is a multi-service network. It does not just combine real-time and data traffic. It leverages the opportunity created by having two endpoints involved in a voice or video conversation and combines this with the ability to tap into an entire cyber-world of data resources.

This opportunity will give us not just another way to do the same things we do today, but will make things much easier and cheaper, and will enable completely new applications that make the workplace more productive. Today's most elaborate PBX features will be managed far more easily and inexpensively in the multiservice network. Adding or moving offices or people will be greatly simplified. Having your office context follow you into a conference room or another facility will be as simple as a login. Instantly putting the right information into your conference will be a few clicks away.

So What is the Problem?

Why haven't we created this multi-service network? Why is mixing real-time traffic and data traffic so difficult? Because real-time and data traffic have very different characteristics and very different needs, which interfere with each other in a network to the detriment of both.

Data Traffic is Bursty

Computers want to utilize the entire available bandwidth for short periods of time to move blocks of data across the network. Transfer a file, send email, download a web page, backup a file. This creates the traffic profile we are all familiar with, having very sharp peaks, and a large peak to average ratio. Figure 1 shows typical data bandwidth, with a peak to average ratio of 10. Computers can share a network because these peaks can be interleaved. Waiting a short period of time to use the network does not cause the computer any problems.

NetForecast Report 5059

Data traffic, using the TCP protocol, is able to recover from packet loss through retransmission. Hence if the peaks of utilization come together in the network and momentarily overwhelm a router queue, the protocol sends the information again to insure it gets through. Of primary importance to data traffic is that it get through 100% correct. If a short delay is required to insure reliable delivery, so be it.

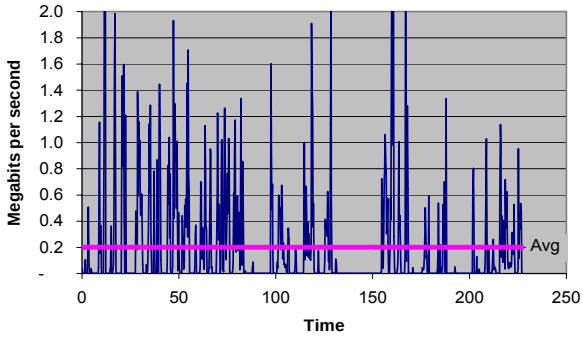

Figure 1 - Typical Data Traffic Profile

Lastly, data traffic is self-regulating. If a network connection is busy, packet loss causes each TCP connection to throttle back the rate at which it is sending data until the packet loss is lowered to an acceptable amount. The completion of the task (e.g., a file transfer) is stretched out, but the task is eventually completed with 100% of the data delivered.

Real-time Traffic is Constant

Real-time traffic is not bursty, it is constant and ongoing. Real-time traffic begins when the connection is established (e.g., a VoIP call), and continues for the duration of that call. The bandwidth never exceeds the nominal bandwidth of the connection, and the peak to average ratio is very close to one, as illustrated in Figure 2. This stream of data represents a real world continuous function, such as a video image, someone's voice, or music. This information has an important time component; it must be delivered in a very regular fashion so that the voice or image can be recreated at the other end of the connection, using the same timing with which it was created.

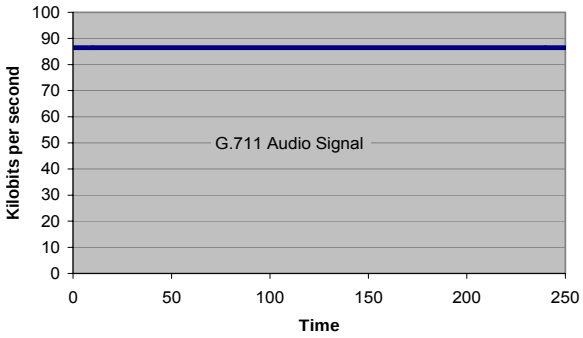

Figure 2 - Audio Bandwidth Profile

Real-time traffic does not recover from packet loss, because it has no time to do so. Once the receiving end has discovered a packet is missing, there is no time left to request that it be sent again from the source. The receiver has to make-do, using various algorithms to cover up the missing data as best it can. Thus packet loss has a direct impact on the quality of the voice and video reproduction.

Real-time traffic does not regulate itself as the network becomes more or less congested. Some devices have the ability to pick an encoding algorithm based on current network parameters, choosing a higher or lower bandwidth codec. However, once the choice is made, the stream emanates from the source at a constant bandwidth, independent of the state of the network. Congestion in the net that causes packet loss will cause the quality of the connection to fluctuate as the traffic load varies.

Lastly, real-time traffic is sensitive to latency and jitter. Latency, the delay from source to destination, affects the interactiveness of a real-time conversation. Streaming traffic, those situations where there is one source and one or more listeners (e.g., a video clip on CNN) can sustain extended latency because there is little interaction between the listener

and the source. However a voice conversation or a videoconference require latency to be below 150 ms for natural interactions to occur [1].

Jitter is the variation in latency across the network. If jitter is relatively small, (i.e., latency is constant,) packets arrive in a timely manner and can be played promptly. If jitter is large, then a timing buffer is required on the receiving end to properly align packets in time before they are played. This jitter buffer adds overall latency to the connection, and will drop packets if they fall outside of its timing window.

Latency and Jitter

So where are latency and jitter introduced to the network? The primary contributors are path length (not much can be done about that) and queues, the latter the real culprit. Each router and switch has output queues through which packets must pass on their way to the output port, as shown in Figure 3. When utilization of a link is low, queues are nearly empty, which means packets pass through quickly. When traffic is heavy, or when traffic is bursty (momentarily heavy), queues suddenly have to store packets waiting for the output port. Waiting behind other packets causes incoming packets to be slowed down. Latency, the time from the source to the destination, increases. Jitter is the variation in latency. Jitter occurs when short bursts fill queues, and then let them empty. Some packets will experience deep queues, while others in the same flow will see nearly empty queues.

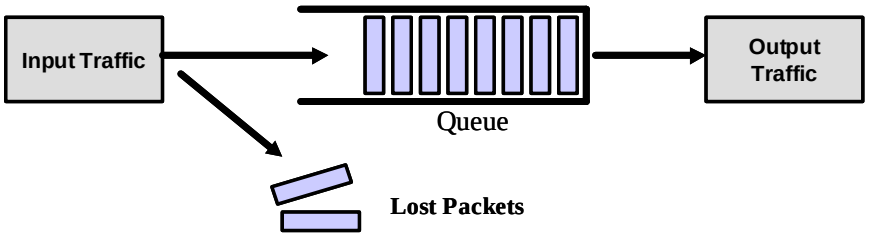

Figure 3 - Queuing Causes Delay and Loss of Packets

Well-behaved traffic, meaning packet streams that are of a consistent packet-length and are spaced consistently in time, will not have the effect on queues described above. Queues carrying this traffic will not back up until the line utilization is very high. Voice and video signals are examples of well-behaved traffic. However, when mixed with data traffic, queuing occurs at much lower levels of utilization and can have adverse affects on the video and/or voice quality.

How Does a Voice and Data Mix Behave?

NetForecast built a model to simulate voice and web traffic running on a common link with equal priority. Two different mixes of voice and web traffic were simulated; one using 20% voice traffic, and 80% web traffic. The second uses a 40% voice and 60% web mix.

The model uses as a source a traffic profile captured from real voice over IP (VoIP) and n web users. A set of users is modeled using each application (voice and web) over a 1-hour period. Web users have randomized start times, while voice users run continuously for the duration of the simulation. Queue depth is modeled on 100ms intervals by finding the difference between the offered load, and the load able to be carried by the line. Packet loss is counted whenever the queues overflow. Latency is measured as the time necessary to empty the queue.

Bursty Traffic Causes Queuing

The model clearly shows the effect of burstiness on queue depth. Web traffic is quite bursty, whereas voice is very well controlled. When voice traffic is mixed with web traffic, the composite traffic becomes more managed as more voice is added to the mix.

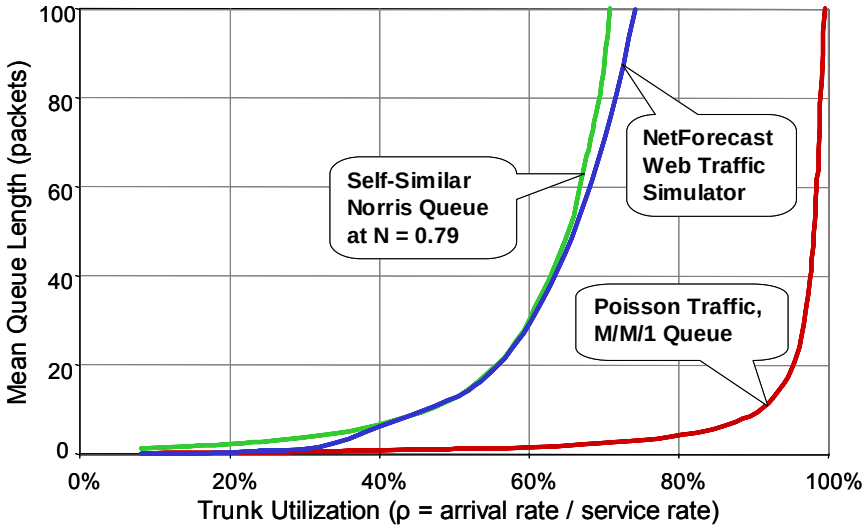

Figure 4 - Queuing Depth for Different Amounts of Burstiness

Figure 4 shows queue depth vs. line utilization for typical data traffic, as well as the standard MM1 queuing curve for reference [2]. Also shown is the Norris curve[2], which is an analytical model, intended to better represent the effect of bursty traffic. Note how the simulated traffic follows the Norris curve. As the voice mix increases, this simulation curve (blue) slides towards the right, approaching the M/M/1 curve.

Results of Modeling

Using the model described above, three networks were simulated. The characteristics of these networks are shown in Table 1. Each network was modeled with a T1 link at the edges, 2 10-Megabit/sec hops in the access ISP, an OC12 to the backbone and OC48s in the backbone itself. The edge of each network is the same; the longer networks having more core (OC48) hops than the shorter networks.

The model uses a captured traffic profile to generate a traffic load. The voice traffic in this simulation is using a G.711 codec. The remaining traffic is HTTP web traffic. The load is scaled up by adding instances (e.g., application users.) The accumulated traffic load is then fed into a queue model simulating the output queues of the routers on the network path. The resulting packet loss and latency of the network are determined.

Table 1 - Modeled Network Characteristics

Network	Length	Hops
Metro	25 mi.	10
National	3,000 mi.	15
Global	12,000 mi.	25

Packet loss and latency are converted to a Mean Opinion Score (MOS) using the ITU e-model [3] that estimates the effect of packet loss and latency on voice transmission quality. A score of 5 represents excellent quality, whereas a score of 1 is very poor quality. A score of 4 or better is considered 'toll' quality, equivalent to the performance of today's PSTN.

MOS Results

NetForecast ran the model with varying loads to determine the network behavior across a wide range of utilizations. Figure 5 shows the calculated Mean Opinion Score (MOS) for each network. The two groups of curves represent a 20% voice mix and a 40% voice mix respectively.

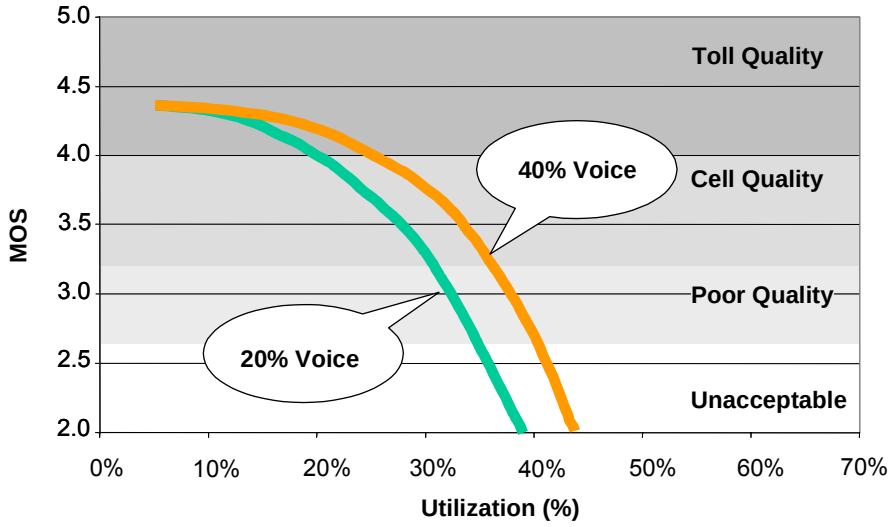

Figure 5 - MOS Modeling Results

In this figure, we see that a rapid drop in MOS occurs as soon as the network begins to lose packets. This packet loss comes from two sources. The first is the loss of packets due to queue overflows. Any packet dropped at a router along the network path, causes degradation in the voice quality of the receiving party. Packet loss also comes from an accumulation of jitter through the network. Each router queue adds jitter to the packets as they traverse the network. If the accumulated jitter of a packet is larger than the jitter buffer on the receiving system, then the packet is dropped. This adds to the percentage of packets dropped, and decreases quality further.

Note that the 40% mix curve is further to the right. This indicates the network can be run at a higher level of utilization before crossing a quality threshold. This occurs because the voice traffic is very consistent in nature, never bursty. Much of the need for queuing comes from the bursty nature of data traffic. Hence as the mix moves more heavily towards voice traffic, the network behaves better at any given utilization level. The first VOIP call on the network is the one that is hardest to guarantee.

The widths of the curves in Figure 5 represent the difference among the metro, national and global networks. Note that the e-Model used to derive the MOS values is not very sensitive to latency. Thus the size of the network (Metro, National or Global) does not show a

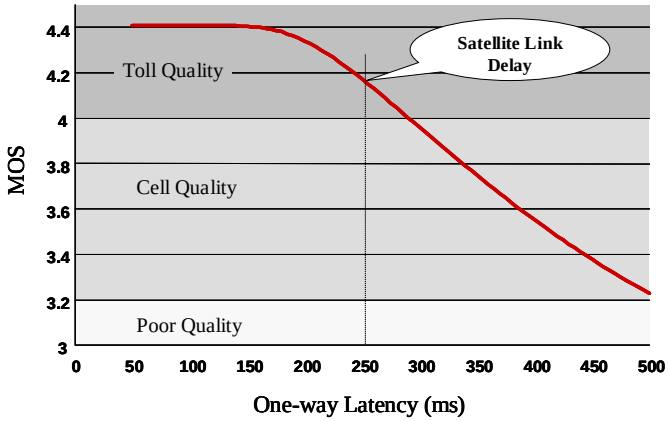

Figure 6 - MOS as a Function of Latency

significant different in these results. Figure 6 shows the effect of latency on MOS using the ITU e-model. A one-way latency of 250 ms, which represents a satellite link phone call from Boston to Los Angeles, is shown as still being a toll quality call. However, anyone who has experienced a satellite call knows that the long round trip delay time causes the interactivity of the call to be very poor.

Cost and Utilization vs. Voice Quality

It is clear that voice quality is best when network utilization is low. But low utilization also means that the resources of the network are underutilized, resulting in a high cost solution. A productive business or profitable carrier network cannot afford this luxury.

Solutions are required that allow high quality voice in the heavily utilized network, so that both the voice quality goal and the low network cost goal can be simultaneously realized. Quality of Service techniques have promised this utopia, but have been slow to deliver, and remain problematic today.

Voice Quality Bands

Four voice quality bands are shown on the MOS charts in this report. The values for these bands were derived from Table B.1 in the ITU-T G.107 specification, and from calculating the MOS score of various cellular telephone coding algorithms, as given in the ITU-T G.113 specification (table I.1/G.113.) The cellular scores are degraded only due to the codec algorithm; they assume a clear radio channel. Table B.1/G.107 indicates that a MOS score of 4.3 is required before users are 'Very Satisfied'.

Quality of Service

Quality of service (QoS) comes in many forms. Some forms of QoS move special traffic onto its own separate network, where it can be better managed. This is an effective, but expensive solution. All forms of QoS that maintain differentiated traffic types on the same network use some form of prioritization to insure time-sensitive traffic is delivered quickly and consistently.

Figure 7 shows the interference of spikes of data traffic into the constant bandwidth consumption of a real-time traffic flow. The average bandwidth of the data traffic is low, however its instantaneous peaks cause interference with the real-time traffic. Quality of Service techniques solve this problem by giving priority to the real-time stream. This in effect holds down those momentary peaks so they do not interfere with the real-time stream. The impact on the data stream is minimal, since sufficient bandwidth exists to complete the data transfer quickly.

This simplified example becomes complex when there are many high-priority streams requiring different levels of priority, and when overall utilization is high. When multiple real-time streams are mixed at the same level of priority, the even queuing behavior we see with voice streams, no longer occurs because queues are carrying packets of varying lengths. These additional streams may be bursty, which creates additional queuing to absorb those bursts when they exceed the output line rate.

Furthermore, when there are multiple levels of priority simultaneously using the same link (e.g., different classes of QoS), each level has an effect on the other. The highest priority class gets the link bandwidth and timing it requires. The traffic in this priority level has some effect on the next class below it, because it has preempted some network bandwidth. This next level down, in conjunction with the traffic in the highest priority level, has an effect on the next lower traffic level, and so on down the line.

The queuing behavior of these combinations of traffic quickly move back towards the best effort queuing behavior described above, again giving us poor performance as the network utilization increases. It is especially difficult to provide different types of QoS

service in this converged environment, such as a low latency connection *and* a guaranteed bandwidth connection.

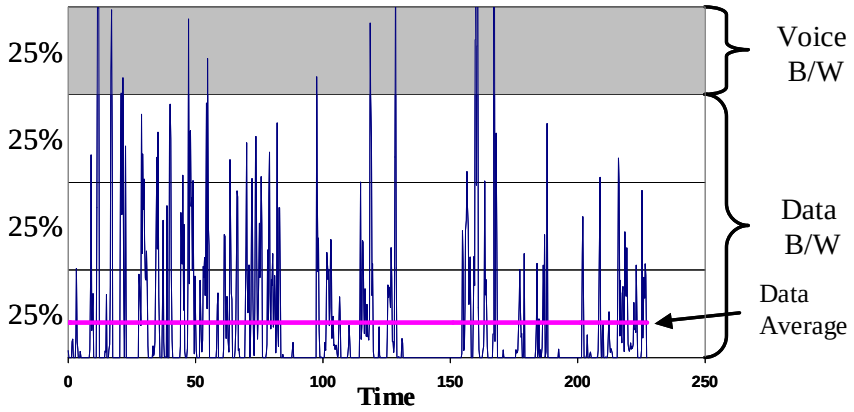

Figure 7 - Mixing Data and Voice

A New Solution - Sequenced Traffic

A new approach to solving this problem is being built by Cetacean Networks, in Portsmouth, New Hampshire (www.cetacean.com). To untie the Gordian knot of QoS, Cetacean has re-introduced a lost dimension, time.

Lets draw an analogy between data networks and the airline business. Best effort traffic always flies standby. If there is a seat available at the last minute, the packet goes. If not, the packet waits. High priority traffic flies standby in first class. First class is then extended as far back in the plane as necessary to accommodate the various levels of priority passengers. But if a seat is not available, the packet is still waiting.

Sequenced traffic introduces the idea of an itinerary to IP networks. A specific seat, in the form of a time slot appointment, is assigned to the sequenced traffic flow on each leg of the flow's path. Each appointment will always give the sequenced traveler precedence over any other. However, if the sequenced traveler is not there, the seat is given to a best-effort standby traveler instead.

How Sequencing Works

To make sequencing possible, switches must have a synchronized view of time. Packets that pass between switches indicating the beginning of a synchronization window accomplish switch synchronization. Appointments for each packet flow are assigned at each switch, and the collection of appointments for a flow is its itinerary. These appointments are assigned in such a way that a packet flowing through the network will always arrive just in time for its appointment at the next switch. Thus each appointment has accounted for the flight time from the previous switch on this path.

When a new traffic flow is about to begin, such as at the beginning of a voice or video conversation, the network determines its path and itinerary for this flow. A sequencing agent, like a travel agent, is responsible for finding the best path and itinerary for this flow. This sequencing agent exists for each administrative span of control in a sequenced network.

Each switch is aware of active flows, and the appointments assigned to them. As the time for an appointment comes near, the switch holds off other traffic, making the output port available for the sequenced packet imminently arriving. The switch only stops unsequenced traffic to the output port if a packet's length is such that it would still be

consuming the output port when an appointment starts. The sequenced packet arrives, and is passed directly to the waiting output port without any queuing delay.

Sequencing Increases Network Utilization Potential

The key differentiator between sequenced traffic and QoS (priority based) traffic is that each sequenced flow is completely independent of the other flows using the same links because they are separated in time. Thus the highest priority voice connection does not limit the bandwidth or add jitter to a lower priority voice stream. Nor does the combination of the voice and video streams affect the interactive traffic that is also sequenced, to insure user productivity remains high. Because this technique isolates each stream from the others, very high link utilizations can be accomplished with no degradation in the behavior of sequenced flows.

Figure 8 shows the same MOS curves seen earlier, with a new line indicating the performance of sequenced traffic. Because the sequenced traffic is isolated in time from the other traffic flowing on the link, there is no degradation as the link utilization approaches 100%. When this traffic was modeled, NetForecast found a very slight degradation in the global model due to the distance halfway around the globe, but no degradation due to interference from other traffic.

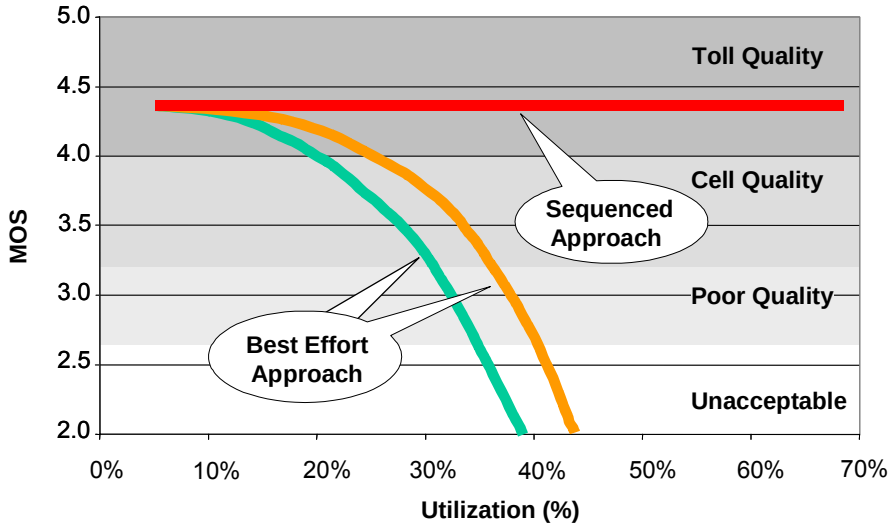

Figure 8 - Sequenced vs. Best Effort Traffic

Sequencing vs. TDM

So how is this different than the channelized architecture of the PSTN? Why not just allocate bandwidth in a SONET Ring for voice instead of integrating it with IP?

Sequencing has the same advantages as a TDM network, in that the bandwidth is always allocated and available, the latency is predictable and short, and jitter is minimal. Its advantages over TDM are two fold. Firstly, although a sequenced flow always has the bandwidth available to it, it does not waste that bandwidth if there is temporarily no use for it. A voice call that uses silence suppression will have time slots in which no data is sent. In the sequenced network, these slots are used by best effort traffic. Once a switch determines that a sequenced packet is not arriving, it will de-queue other traffic and utilize the available bandwidth.

Secondly, integration with the data network brings all the advantages of being tightly connected to the rich data environment it provides, and reduces overall costs by eliminating redundant networks.

Sequencing Eliminates Jitter

We noted above that a sequenced flow has very low jitter because there is no queuing of the traffic as it passes through the network. Because jitter is virtually eliminated, no jitter buffer is required on the receiving endpoint. This simplifies the logic of the receiver, and more importantly, reduces the end-to-end latency of the connection, important for interactive voice or video communications.

Conclusions

The bursty nature of data traffic causes queuing, which in turn causes delay, variations in delay (jitter) and packet loss. All three of these problems increase exponentially as utilization increases. Our modeling shows that packet loss starts to occur at about 30% utilization, which correlates well with industry experience.

Quality of Service techniques that rely on prioritization have a difficult time meeting all their latency, jitter and packet loss goals due to the interactions between multiple levels of priority. Single point solutions (e.g., two levels of priority, voice and everything else) will work, but more complex combinations create a management nightmare.

Cetacean's new sequenced traffic approach, adds a welcome dimension to the network, which allows real-time traffic flows to meet their latency, jitter and packet loss goals without degrading other traffic flows on the link. This approach will allow enterprises and carriers to run their networks at much higher link utilizations without sacrificing connection quality.

The increased economic pressure on enterprise, ISP and carrier networks will push network operators to higher traffic utilization. A highly utilized network is a cost-effective network, assuming it satisfies business goals. Sequenced traffic promises to allow business goals to be met while running a cost-effective network. ¤

References

[1] "General Characteristics of International Telephone Connections and International Telephone Circuits: One-Way Transmission Time", ITU-T Recommendation G.114, February 1996, http://www.itu.org, or http://www.itu.int/rec/recommendation.asp?type=folders&lang=e&parent=T-REC-G.114

[2] "High-Speed Networks, TCP/IP and ATM Design Principles" by William Stallings, Prentice Hall, (Chapter 8.3 Performance Implications of Self-Similarity)

[3] "The E-model, a computational model for use in transmission planning", ITU-T G.107 E-Model, http://www.itu.org, or http://www.itu.int/rec/recommendation.asp?type=folders&lang=e&parent=T-REC-G.107

Authors

John Bartlett is Vice President of NetForecast, and has 24 years of experience designing at the chip, board, system and network levels, with a focus on performance. John led the team that built the first VLAN implementation, one of the first ATM switches, and he is a leading authority on the behavior of real-time traffic on the Internet. He can be reached at john@netforecast.com.

Peter Sevcik is President of NetForecast and a leading authority on network technology. He specializes in performance analysis and holds the patent on application response time prediction. Peter has contributed to the design of more than 100 networks including the Internet. He can be reached at peter@netforecast.com.