

VOICE OVER THE INTERNET: A TUTORIAL DISCUSSING PROBLEMS AND SOLUTIONS ASSOCIATED WITH ALTERNATIVE TRANSPORT

GEORGE SCHEETS, OKLAHOMA STATE UNIVERSITY

MARIOS PARPERIS, WILTEL COMMUNICATIONS

RITU SINGH, NDRS TECHNOLOGIES

ABSTRACT

This article provides a tutorial overview of voice over the Internet, examining the effects of moving voice traffic over the packet switched Internet and comparing this with the effects of moving voice over the more traditional circuit-switched telephone system. The emphasis of this document is on areas of concern to a backbone service provider implementing Voice over IP (VoIP). We begin by providing overviews of the Plain Old Telephone Service (POTS) and VoIP. We then discuss techniques service providers can use to help preserve service quality on their VoIP networks.

Next, we briefly discuss Voice over ATM (VoATM) as an alternative to VoIP.

Finally, we offer some conclusions.

This article provides a tutorial overview of one of the Internet traffic types likely to see significant future growth, Voice over the Internet (a.k.a. Voice over IP or VoIP). The effects of moving voice traffic over a packet-switched statistically multiplexed network such as the Internet, which was never really designed to handle this type of source, are examined and compared with the effects of moving voice over the more traditional circuit-switched time division multiplexed telephone system, more affectionately known as POTS. The emphasis of this article is on areas of concern to a backbone service provider implementing VoIP.

This article is organized as follows. We first provide an overview of POTS. We then examine a non-traditional, packet-based, voice transport, VoIP. We next survey techniques service providers can use to help preserve quality on their VoIP networks, while the following section considers the effects that design choices and failings in the network have on this quality. We discuss the trade-offs involved in terms of maximizing the number of calls supported while simultaneously preserving end-to-end delivery bounds and meeting minimum quality standards. Comments regarding another alternative voice transport system, ATM, are offered. Our conclusions are then summarized.

POTS

POTS today consists of a mix of the very old and the very new. Figure 1 presents a simplified view of POTS connectivity. End-user telephones typically use twisted pair copper cabling to deliver analog voice signals to the central office (CO). The vast majority of inter-office communications are now carried over fiber optic trunks.

Today, most long-haul voice traffic is digitized. At the CO switch, the analog signals are passed through a narrow-bandwidth bandpass filter, sampled at a rate of 8,000 samples/second and quantized (rounded off) to one of 256 unequally spaced voltage values. A technique known as Pulse Code Modulation (PCM) is then used to assign an 8-bit code word to each voltage, resulting in 64 kb requiring transport every second. This 8-bit code word is the basic transmission entity of POTS. POTS backbones use Time Division Multiplexing (TDM) and circuit switching to efficiently transmit this entity. Circuit switching is used to dedicate sufficient bandwidth to support this bit rate for the duration of the phone call. An end-to-end path is set up, and resources are dedicated, prior to the initiation of the actual voice conversation. As PCM outputs 8 bits every 1/8000th of a second in a predictable, deter-

ministic manner, for every trunk traversed, regardless of the trunk speed used, TDM will set aside 8 bits every 1/8000th of a second for each simplex call. The resulting traffic is moved to the destination central office switch, and converted back to an analog signal for transmission over the local loop [1].

Transmission protocols originally developed for the voice network, including state of the art Synchronous Optical Network (SONET) and the now mostly obsolete T Carrier systems, are all built around frames, of which there are 8,000 each second. Frames typically consist of some control bits or bytes, and time slots into which 8-bit voice code words can be inserted.

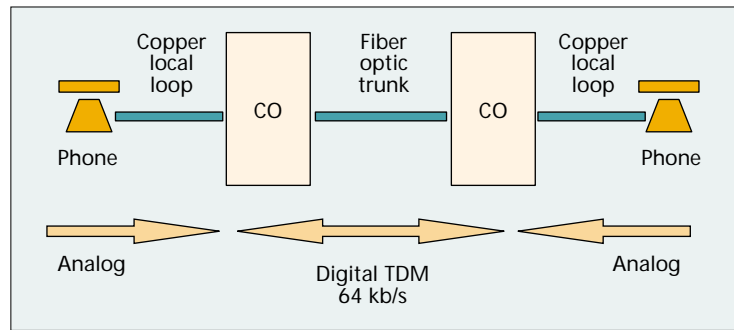
Few central office switches are directly connected as shown in Fig. 1. More typically one or more digital TDM telephone switches are in the end-to-end path a POTS call traverses. To maximize network use, intermediate voice switches (known as tandem switches) may move voice code words to an output line time slot that differs from the input line's time slot. This may be necessary, for example, when voice traffic occupying the same time slot is received on two different input lines, but both need to be placed on the same output line. This time slot interchange (TSI) may be accomplished by writing voice code words into the tandem switch's memory in the order they are received, and then reading them back out of memory at the appropriate time but in a different order [1]. Time slot interchange may cause a worst-case delay of 1/8000th of a second, which can occur if a voice code word occupying time slot k of a frame must be delayed to time slot $k-1$ of the following frame.

Voice delays through a POTS network are smaller than those seen through a VoIP network. Essentially, the end-to-end delivery delay on a POTS network consists of the propagation delay of the electromagnetic energy over the physical cabling, the PCM encoding delay at the source central office (ideally 1/8000th of a second), TSI delays on intermediate tandem switches ($\leq 1/8000$ th of a second per switch), and the PCM decoding delay at the destination central office that is also ideally 1/8000th of a second. Practically speaking, coding and decoding delays may be approximately 10-20 times larger, depending on the demands placed on the source and destination digital signal processors. Figure 2 shows a simplified diagram of these delays.

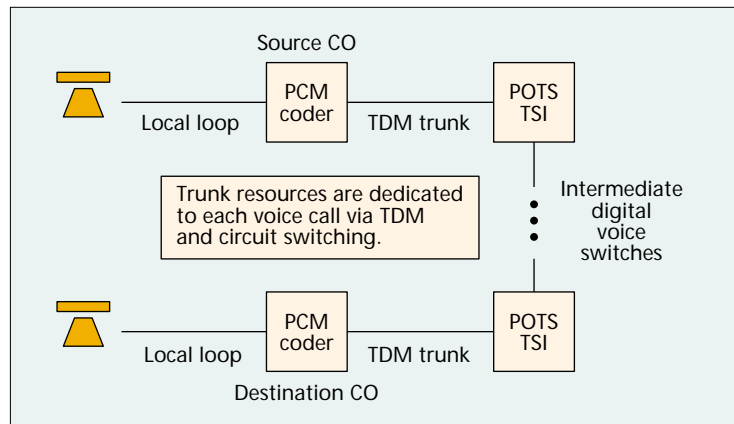
NON-TRADITIONAL TELEPHONY: VOIP

Probably the most significant difference with VoIP, as compared to POTS, is that backbone trunking resources are not assigned in a dedicated methodical manner to support a voice call. Instead, trunk bandwidth is assigned by routers on an seemingly random, as needed basis via statistical multiplexing and packet switching. The result is that while statistics can be generated for the average packet delay and variation, the delivery times for each individual packet are not predictable. With POTS, voice traffic arrives at the destination PCM decoder in a constant 8-bits-every-1/8000th-of-a-second manner. With VoIP, packetized voice traffic arrives at the destination decoder at unpredictable intervals and ordering. This randomness necessitates a more complex environment in order to smooth the flow of voice bits to the destination voice decoder, which typically requires a steady bit arrival rate.

Figure 3 shows a simplified diagram of the delays associated with this type of network. Below we highlight the key functions of each block.



■ FIGURE 1. Typical POTS connectivity.



■ FIGURE 2. Sources of POTS delay.

VOICE CODER

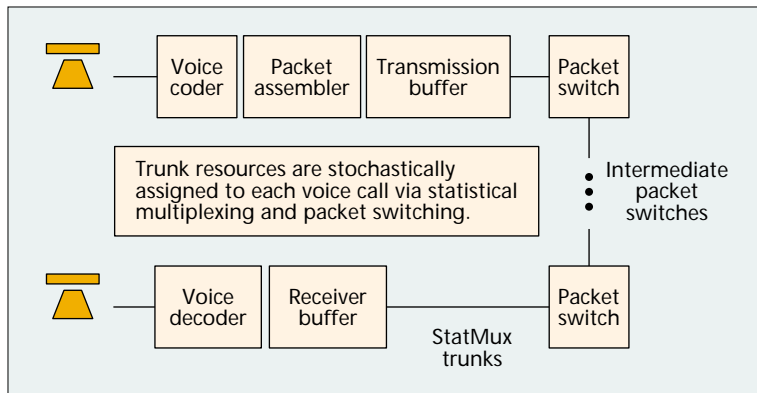
POTS networks almost universally use the ITU G.711 64 kb/s PCM standard. As previously noted, this type of coder outputs 8 bits every 1/8000th of a second, that is, the frame of this coder is 1/8000th of a second, and the frame size is 8 bits.

Other coders exist that can reduce the generated bit rate, but usually with a slightly reduced perceived quality. Table 1 compares selected ITU coders, including their associated bit rates [2, 3].

Of these codes, G.729 has evoked considerable interest for VoIP providers as it has comparable quality to G.711, but at a greatly reduced bit rate. G.729 has a frame length of 10 msec, and its compression algorithm outputs 80 bits every 1/100th of a second, yielding an 8 kb/s bit rate.

POTS TDM backbones are based around fixed-rate coders. For example, a G.711 coder outputs 64 kb/s at all times, regardless of whether or not the voice source is talking or listening. The statistical as-needed allocation of backbone bandwidth on VoIP networks offers the opportunity to deploy variable-rate coders, which output traffic at one bit rate when the voice source is talking, and a lower bit rate (possibly zero) when the voice source is quiet. For example, G.729B with silence suppression examines each 10 msec voice frame and makes a voice/no voice decision. If the coder detects voice energy, it will output the standard 80 bits of compressed, digitized voice for that frame. If the coder decides this frame does not contain voice energy, the coder will output a reduced block of bits containing comfort noise, information that the receiver will use to generate background noise so that the user does not think the connection has been lost [3]. Alternatively, the coder could output nothing at all and the receiver could generate comfort noise.

Experiments have shown that in a typical two-way interactive voice conversation, voice sources are only active 40 percent of the time [4]. The 60 percent idle time includes pauses



■ FIGURE 3. Sources of VoIP delay.

while listening to the other party talking, as well as pauses between sentences, and even pauses between some words. A G.729B coder with silence suppression will output 8 kb/s during talk spurts (40 percent of the time), and nothing or a reduced bit rate during intervening silence intervals (60 percent of the time). The use of silence suppression potentially allows a G.729 coder to reduce its average output from 8 kb/s to 3.2 kb/s (alternating 8 kb/s and 0 kb/s bursts).

PACKET ASSEMBLER

One decision faced by every operator of a VoIP telephony voice switch is how many frames from the voice coder to include in each transmitted packet. A typical VoIP packet requires about 47 bytes of overhead: 8 bytes for the User Datagram Protocol (UDP), 12 bytes for Real Time Transport Protocol (RTP), 20 bytes for the Internet Protocol (IP), and 7 bytes for the Point-to-Point Protocol. Header compression of the UDP and IP headers can reduce the amount of overhead on low-speed links, but is not a standardized option on high-speed carrier backbones. It would not be cost effective to take the one byte frame output of a G.711 coder, packetize it, and immediately transport this 48 byte packet, as 98 percent (47 out of 48) of the bytes transmitted would be overhead. In this case, it is better to place the output of multiple frames into one packet. The disadvantage of this is that time is lost at the transmitter site waiting for the desired number of voice frames to be generated.

The choice of the number of frames in a packet can seriously impact the number of phone calls a VoIP network can support. Too few frames/packet results in a considerable percentage of bandwidth being allocated to the movement of overhead bytes. Too many frames in a packet results in an excessive amount of time being lost assembling a packet at the transmitter, reducing the time remaining to meet the end-to-end delivery time bound. In the latter case, trunk loads may have to be decreased in order to maintain switch queuing delays at tolerable levels.

Once assembled, the transmission buffer provides a place for the packet to be stored while waiting for network bandwidth to become available.

RECEIVER BUFFER

Unlike the POTS network where voice traffic arrives at the destination in byte sized pieces at regular time intervals, in a VoIP network traffic arrives irregularly in packet sized chunks. Voice decoders are typically expecting a steady feed rate. For example, a G.711 decoder expects to receive 8 bits every 1/8000th of a second. A G.729 decoder expects 80 bits every 1/100th of a second. The receiver buffer's primary function is to remove the jitter associated

with the irregular packet arrivals and provide a smooth traffic flow to the voice decoder. As a result, the receiver buffer is also known as a de-jitter buffer. Generally, this buffer is allowed to partially fill before any play-out to the voice decoder commences.

An appropriately sized receiver buffer is critical to the proper operation of VoIP systems. By sizing, we are referring to two parameters: the fill delay mentioned immediately above, and the buffer storage capacity. If the fill delay is chosen too small and a group of packets arrives later than expected, it is possible that the buffer will empty, which will result in a loss of voice output. If the fill delay is chosen too large, then excessive time is lost in the

receiver buffer waiting to be played out, which reduces the time allotted to other devices in order that end-to-end delivery goals be met. If the buffer storage capacity is too small then a burst of packets received may result in inadequate storage space being available and dropped packets.

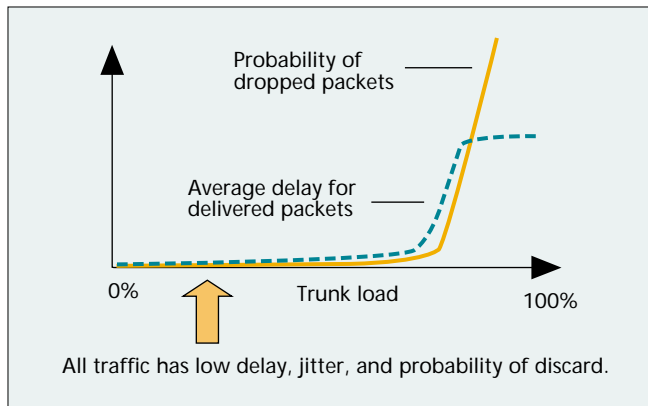
With the use of time lines, and assuming no packets are dropped by intermediate packet switches, it is possible to show that to be 100 percent certain that the receiver buffer does not empty, packet playback should commence no earlier than $WCDeliv$ seconds after assembly, where $WCDeliv$ is equal to the worst case packet end-to-end delivery time. A more complicated statistical analysis would be necessary to determine tolerable receiver buffer delay if packets are dropped by the network, or if a possibility of the receiver buffer emptying is acceptable. If VoIP traffic is routed over an Internet backbone augmented with voice gateways handling connection admission control to limit the number of voice calls, and these gateways are knowledgeable of the end-to-end voice paths through the network and worst-case delays at transited routers, the above technique is, at least in theory, possible.

However, if the traffic is routed over the commodity Internet, where the exact path through the network and the worst-case delay through routers is not known, determination of $WCDeliv$ may be difficult, if not impossible, to obtain. In this case the packet playback might be delayed a value large enough such that it is (hopefully) rarely exceeded by the actual packet end-to-end transit time. An alternative option, and one that tends to be favored today, would be to adaptively adjust the de-jitter buffer delay to account for the time varying changes of the voice packet delivery delays.

To summarize, in comparing Fig. 2 with Fig. 3, it should be noted that a VoIP system is more complex than POTS. VoIP has more sources of delay, and the delay through these ele-

Codec	G.711	G.723.1	G.726-32	G.729
Coding rate (kb/s)	64	5.3-6.3	32	8
Frame size (ms)	0.125	30	0.250	10
Algorithm	PCM	MP-MLQ	ADPCM	CS-ACELP
Processing delay (ms)	0.125	30	0.250	10
Look ahead delay (ms)	0	7.5	0	5
Payload (Bytes)	1	20-24	1	10
Quality	Good	Good/fair	Good	Good
Complexity	Lowest	Highest	Low	High

■ Table 1. Selected ITU voice coders and key characteristics.



■ FIGURE 4. Light trunk loading.

ments is generally considerably larger than that experienced crossing corresponding POTS elements. Note also that individual packets will see different transit times across the network. VoIP designers have a more difficult job insuring that end-to-end delivery delays and overall quality remain at tolerable values.

Examining Fig. 3, note that to meet the required end-to-end delivery delay, the sum of the voice coding delay ($VCDelay$), packet assembly delay ($PADelay$), service time and queuing delays at the voice source and all intermediate packet switches ($QuDelays$), receiver de-jitter buffer delay ($JitDelay$), end-to-end propagation delay ($PropDelay$), and voice decoding delay ($VDDelay$), all must be less than or equal to the target end-to-end voice delivery delay ($Target$). Mathematically,

$$VCDelay + PADelay + QuDelays + JitDelay + PropDelay + VDDelay \leq Target. \quad (1)$$

PRESERVING QUALITY ON VOIP NETWORKS

Preserving the quality on VoIP networks requires careful engineering by the system engineer [5]. In this section we discuss available design options [6–14].

CONTROLLING THE NUMBER OF INTERMEDIATE PACKET SWITCHES

It is not unusual for traffic traversing the commodity Internet to traverse 10–20 packet switches (routers), even on short-distance trips. As examples, a Traceroute launched from the lead author's house in Stillwater, Oklahoma to Cisco System's Web site in California traversed routers in Oklahoma City, Dallas, Fort Worth, Anaheim, and San Jose, 13 in all. A Traceroute launched from the same location to Oklahoma State University's Web server located two miles away was first shipped to Oklahoma City, then back to Stillwater, traversing a total of 10 routers.

Most of the variability in the inter-arrival time of VoIP packets at the destination is due to varying queue times at the intermediate packet switches. A large amount of resulting jitter will require a large receiver de-jitter buffer delay to smooth things out. Keeping the number of intervening packet switches low can make a significant difference here. This may be difficult to accomplish on the commodity Internet.

CONTROLLING THE DELAYS AT PACKET SWITCHES

Queuing delays in packet switches can be controlled by either controlling the load on interconnecting trunk links, prioritizing the voice traffic, or enabling some combination of the two.

If the trunks are kept lightly loaded, best-effort first in,

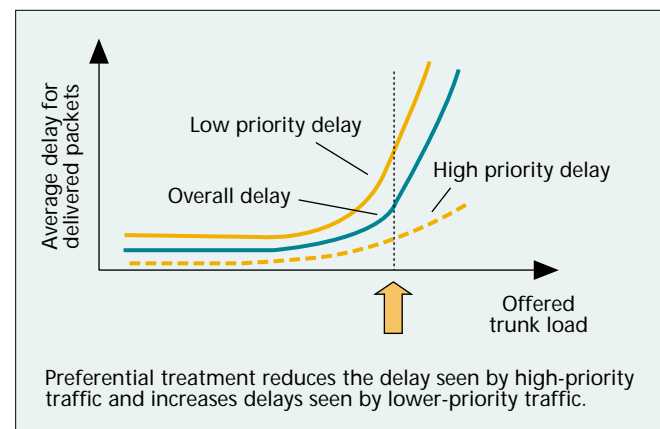
first out routers can be deployed on the backbone. Figure 4 shows a classic delay versus load plot for a statistically multiplexed switch, modified to account for finite-length buffers. As the trunk load increases, a point will eventually be reached where the switch buffers fill up, the probability of a packet being dropped becomes unacceptably high, and the delay seen traversing a switch by packets that are not dropped is limited by the buffer size. At low loads, the average delays through a router will be low, the probability of dropped packets will be small, and quality will be acceptable for all traffic. Note, however, that this option may not be a cost effective approach as it does not scale well if VoIP minutes increase significantly.

At heavier loads Quality of Service (QoS)-enabled routers should be deployed, and priorities used to give preferential treatment to time-sensitive traffic such as interactive VoIP. Figure 5 shows an example where the prioritized voice traffic sees similar performance from a QoS-enabled router (dashed orange line) that the unprioritized voice would see in a more lightly loaded best effort router (orange line). The widespread deployment of Differential Services (DiffServ) or Internet Protocol Version 6 will be essential in order to offer standardized priorities.

CONTROLLING PACKET LOSS AT PACKET SWITCHES

Related to the problem of controlling delay at intervening switches is the problem of controlling packet loss. Both are a function of the amount of buffer space in the switch. In the delay-limited case, there is plenty of buffer space such that the probability of packet discards can be ignored. The allowable delay through the switch sets the load that the output trunks can support. In the buffer-limited case, the finite amount of buffer memory in a switch is the limiting factor that sets the tolerable load that the switch can handle. This finite storage for queued traffic results in packet discard probabilities being the limiting factor, not the tolerable delay.

Choosing an upper bound on the allowable packet loss on a VoIP network is not a trivial task. The subjective voice degradation resulting at the receiver is a complex function of the amount of coder compression (generally, the greater the amount of compression the greater the impact of a loss of information), the number of coder frames placed in each packet (the loss of a packet carrying multiple frames is more serious than the loss of a packet carrying one voice frame), and the decoder's ability to mask any lost traffic [15]. This issue is discussed in more detail in the next section.



■ FIGURE 5. Heavy trunk loading

SEGREGATE OR INTEGRATE

Another choice faced by the VoIP system designer is whether or not to segregate the time-sensitive voice traffic from the data traffic, where the latter refers to man-computer or computer-computer communications. Potentially, carriers will gain the most economy by integrating all types of traffic over a common core. However, given the current state of the commercial Internet, which by and large remains a best-effort network, and traffic increases that strain a carrier's ability to deploy bandwidth fast enough to remain ahead of the growth curve, integrating time-sensitive voice traffic with other traffic on Internet backbones is not necessarily a good idea if the carrier's goal is to offer a VoIP service with quality approaching that of POTS.

Given the state of the art, some carriers have chosen to segregate their VoIP traffic and move it via a dedicated VoIP network separate from the network moving data.

MAINTAINING THE NETWORK: TRAFFIC ENGINEERING

One problem with the Internet today is that, by and large, the system still transmits traffic as datagrams wherein switching devices forward each packet independently of all other packets. As a result there is no guarantee that traffic that is part of the same information transfer will follow the same path. As routers typically update their routing tables several times an hour it is possible that the end-to-end path taken by the voice traffic may shift, possibly to a path with more hops, a larger delay, or more delay variation. Should this happen, the carefully engineered VoIP system could be thrown into disarray and fail to meet specifications.

One advantage of Multi-Protocol Label Switching (MPLS) is that it enables virtual circuits, which allow the path through the system to be designated in advance and fixed in place [9]. Designating the path in advance allows the option of setting aside and reserving switch resources, such as buffer space and bandwidth, for specific traffic flows [10]. Having predictable paths through the system makes traffic engineering simpler and more feasible, increasing the probability that the system can be configured to perform reliably.

Installing voice gateways at access points to the VoIP backbone can allow the implementation of connection admission control (CAC), such that the number of allowed voice calls can be monitored and calls blocked if network resources are insufficient to support a new request.

DEALING WITH DEGRADATION: THE IMPACT OF PACKET LOSS, CODER DISTORTION, AND END-TO-END DELAY ON VOICE QUALITY

The previous section discussed techniques that carriers can use to help insure that a backbone carrying VoIP traffic is able to provide reliable and timely delivery of voice packets. However, despite careful engineering, end-to-end delivery delay is going to be greater than that on POTS, and packets will be lost. In this section, we discuss how these factors, coupled with distortion in the voice coding algorithms, can effect the perceived quality of the reconstructed voice at the receiver. We also note one important tool that can be used to estimate the overall voice quality.

CODER DISTORTION

Coder distortion is frequently assessed via the mean opinion score (MOS), which is a measure that specifies the "perceptual similarity between an uncoded source signal and its coded version" [16]. This score is subjective in nature, and is assigned by a panel of trained listeners awarding values ranging from 1 to 5, with a "5" being no degradation noted and a "4" being perceptible degradation noted, but not annoying. An MOS score of "1" is awarded if the degradation is very annoying. Of the coders listed in Table 1, the POTS mainstay, G.711, has the highest MOS score of 4.3, G.723.1 has the lowest with a 3.9, and the other two lie in between [1]. While the caliber of these coders is very similar, a VoIP system using G.729 does begin its journey from the source coder with a slight quality disadvantage compared to POTS.

PACKET LOSS

Packet loss can severely affect playback quality. One packet can contain one or more speech frames depending on the implementation. While the actual impact of a lost packet will depend on the amount of compression delivered by the voice coder and the number of voice frames per packet, a typical requirement is that the loss rate be one percent or less [15, 17].

Error masking can be used to help overcome the deterioration of quality of the decoded speech due to packet loss. Information in the last correct frame can be used by the receiver decoder to mask missing information. Unfortunately, packet losses encountered in digital transmission over the commodity Internet generally appear in bursts. This tends to reduce the effectiveness of voice decoder error concealment, as the decoder may need to mask the loss of several consecutive packets, each carrying multiple frames [18]. Good CAC procedures at gateways to a carrier VoIP backbone can help reduce the probability of lost packets.

END-TO-END DELIVERY DELAY

Tests have shown that the perceived quality of an interactive voice conversation depends heavily on the time that elapses between voice energy hitting a microphone and playing out on the destination speaker. As this time increases, the quality steadily degrades in that users become more likely to accidentally talk over each other. This time need not be very large before problems commence. The lead author has personally experienced this difficulty during satellite-based phone calls between Korea and the United States back in the early 1980s. The one-way delivery time of approximately 0.3 to 0.4 seconds, coupled with brain processing time at the far end while a reply was formulated, and then the return trip of the audio, was perceived as an unnaturally long time interval that required some practice to adjust to. There are other effects as well. For example, a pause due to a lengthy round-trip time might be mistaken as hesitation on the part of the party with whom you are talking. International Telecommunications Union (ITU) standard G.114 recommends 150 msec as the maximum allowable end-to-end delivery value for VoIP systems, and that will be considered the worst case target value for this article.

ECHO CONTROL [19]

The POTS local loop typically uses two wires which carry both inbound and outbound analog voice traffic, whereas the long distance telephone network uses what are called four wire systems with the inbound audio (ideally) separated from the out-

bound audio and each carried on its own path. Connections between two wire and four wire systems are known as hybrids. Imperfections in these interconnections allow some of the inbound long distance audio to couple into the outbound audio path and return to the sender. Echo in the network results from this coupling between the transmit and the receive path.

While an end-to-end VoIP network will have no hybrids, and hence would have no internal echo paths, it may still be subject to echo from other sources. Coupling between a speaker and a microphone, either acoustically or via mechanical vibrations, is one potential return path back to the original source. Electrical pick up between circuits or wires (known as cross-talk) is another. These echo sources are usually weaker than hybrid echo.

The severity of an echo depends on both the strength of the echoed signal and the time it takes to return to the talker. Given a constant amplitude echo, the longer the round trip time, the louder the echo will sound. As mentioned previously, VoIP networks will have a longer end-to-end delay than POTS. As a result, echo that is inaudible on POTS may become noticeable in a packet based VoIP system.

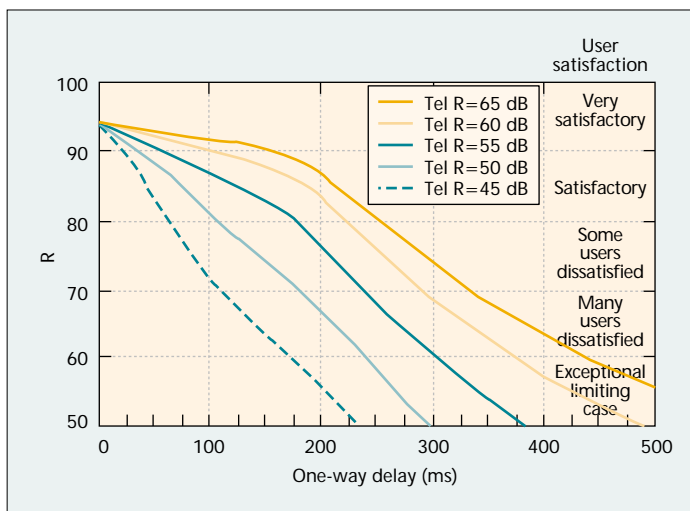
Two key techniques used to control echo in the Public Switched Telephone Network (PSTN) and on VoIP networks are echo cancellers and loss planning. Depending on the strength of the expected echo and its associated time delay, these methods could be used alone or may be combined.

An echo canceller is a device that detects and eliminates echo. Adaptive filters are usually used here. These filters search for delayed and attenuated replicas of the inbound audio in the outbound return path. The adaptive filter will then generate an estimate of this echo signal and subtract that estimate from the return path, reducing the strength of the echo by up to 25–30 dB. Any residual echo can then be removed by a processor that eliminates all weak signals below a certain threshold.

Echo control on calls with short end-to-end delays can be managed effectively by deliberately introducing loss in the path, reducing the echoes to a point below the typical individual's audio threshold. Loss planning (also known as padding) in the Public Switched Telephone Network is largely intended to keep all sources of echo inaudible. Loss planning can only remove echo where the mouth-to-ear delay is up to 20–25 ms. If the delays become greater than this, echo begins to again become audible due to the previously mentioned phenomena of echo subjected to longer round trip times sounding louder than the same echo signal with a shorter round trip delay. Even with short delays, it is generally necessary to control signal levels to ensure that echo control devices function properly. Any echo must be low enough that an echo canceller will not mistake it for a direct signal. Level control becomes of great importance in VoIP networks because of the increase of the end-to-end delay. This requires the deployment of technologies such as automatic level control and Noise Reduction (as in ITU Recommendation G.169) to insure natural, comfortable speech.

THE E-MODEL

ITU standard G.107, known as the E-Model, attempts to predict how a typical user would rate the overall quality of a phone call. Based on a large set of subjective experiments, the model can generate a numerical score that accounts for factors such as coder distortion, packet loss, end-to-end delay, and echo, as well as other parameters such as noise. This resulting score is known as the transmission rating factor R ,



■ FIGURE 6. The effect of echo loudness and end-to-end delay on the perceived quality of a voice call [20]. (Reproduced under written permission from Telecommunications Industry Association.)

and varies between 0 and 100. Given the R rating, the quality is somewhat akin to the standard school grade scale where ≥ 90 is an “A,” ≥ 80 is a “B,” etc.

The R rating is defined as

$$R = R_o - I_s - I_d - I_e + A, \quad (2)$$

where R_o accounts for the effects of noise, I_s the effects of speech transmission impairments such as too much side tone or a too loud signal, I_d the impairments associated with delay (including echo), I_e equipment imperfections, and A is an advantage factor that accounts for the willingness of users to accept substandard performance on a new service.

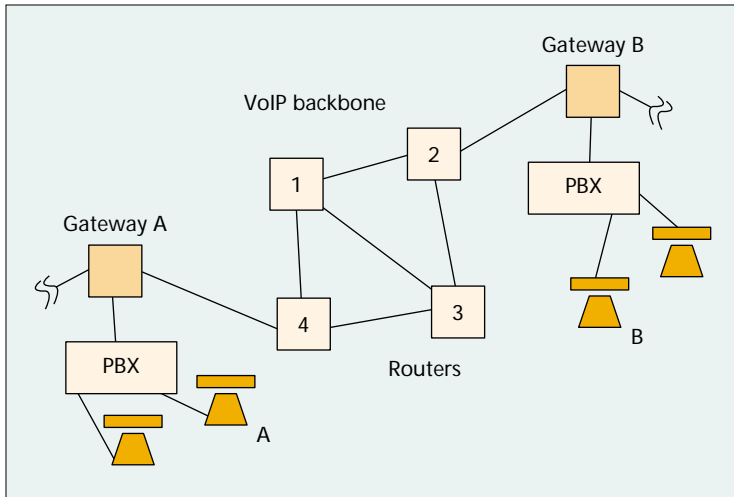
For example, the talker echo loudness rating (TELR) is the loudness loss between the talker's mouth and the ear via the echo path. It is the sum of the losses around the loop, from one set's transmitter back to the receiver on the same set. The family of curves in Fig. 6 shows the effect of echo on the quality of the voice signal, as predicted by the E-Model. Note that as TELR is reduced and an echo is louder, the amount of end-to-end delay available for the connection for a given performance quality objective on the R scale is also reduced.

Given some target voice quality, the system designer can use the E-Model to gain an understanding as to how backbone and access network impairments might affect the caliber of the transmitted voice signal over the VoIP network, and insight regarding which steps might be taken to bring a substandard result up to the quality level desired. We shall use this model shortly when we estimate the expected quality of the signals in the test network described in the following section.

For a more detailed discussion of the E-Model, the interested reader is referred to [15] or [20].

ENGINEERING ON-TIME DELIVERY AND MAXIMIZING THE NUMBER OF SUPPORTED VOICE CALLS

To maximize resulting revenues, the system engineer will want to maximize the number of calls the network will support, while meeting or exceeding the quality level indicated above. To better understand the impact certain design choices have on the ability of a VoIP system to support calls, it is constructive to look in some detail at a specific example. Fig. 7 shows the network used in this example, where phones are connect-



■ FIGURE 7. Example carrier VoIP network.

ed to PBX telephone switches which, in turn, direct external traffic to a VoIP gateway and a VoIP network. This configuration is similar to that used by carriers that have segregated their interactive IP voice traffic from the data. For the numerical example to follow, we make the following assumptions:

- The voice signal is analog until it hits the PBX, where 64 kb/s G.711 coding is used. The coding delay here is 1/8000th of a second. POTS technology then transports this digital signal to the gateway.
- At the gateway, an analog signal is reconstructed and fed to a G.729 voice coder that outputs one frame of 10 octets every 10 msec (8 kb/s). To help prevent audible glitches from frame-to-frame, G.729 includes 5 msec of look-ahead information which overlaps the next frame. Hence 15 msec are required to acquire the necessary voice information to code a frame.
- If silence suppression is being used and the voice source is silent, 0 kb/s is assumed output with any necessary comfort noise generated at the receiver.
- At the gateway, N G.729 Frames are acquired for placement in a packet for transmission. This yields a $47 + N \cdot 10$ -byte packet for transmission (7 bytes Point-to-Point Protocol, 20 bytes IP, 8 bytes UDP, 12 bytes RTP, and $10 \cdot N$ bytes of traffic).
- It requires $10 \cdot N$ msec + 5 msec to acquire sufficient voice to generate N G.729 frames for emplacement in a packet. The voice coder will then have up to 10 msec to complete compression before the next following frame is collected and ready for processing. Considering this, we use a voice coding and packet assembly delay of 15 msec + $10 \cdot N$ msec below.
- The PBXs are connected to a gateway with a T-1 line. Not shown are other PBXs or CO switches connected to the gateways.
- The gateways are connected to the VoIP backbone by OC-3s. Packet switching and statistical multiplexing are used on these connections. Not shown are other gateways that are also attached to the backbone routers.
- The VoIP routers are connected by OC-12s. These connections are also packet switched and statistically multiplexed.
- Propagation delays between the PBX and the gateways are inconsequential and are ignored. The propagation delays between the routers are assumed to be 10 msec.
- The network is engineered such that packet discards due to queue overflows or bit errors are negligible and can be ignored. Hence CAC procedures at gateways are in effect, and fiber optic systems are widely deployed.

- End-to-end packet flows are connection oriented, and therefore packets will arrive at the destination in order of transmission.
- Using the information provided in RTP, the receiver de-jitter buffer delays playback until WC_{Deliv} seconds after packet assembly in order to insure the receiver buffer never empties. All follow-on packets will therefore also commence playing WC_{Deliv} seconds after assembly, spending a total of WC_{Deliv} seconds in transit, stored in the queues of intermediate switches, or stored in the receiver buffer. Note that in this situation, $QuDelays + JitDelay + PropDelay = WC_{Deliv}$ for all packets.

As mentioned earlier, current deployments of VoIP are likely to use fixed or adaptive de-jitter buffer delays. A large fixed de-jitter buffer delay would have the same effect as above. An adaptive de-jitter buffer would likely result in a reduced de-jitter buffer delay (since the worst case delay may not be that common), but also occasional glitches due to an emptied playback buffer, adversely affecting the R rating.

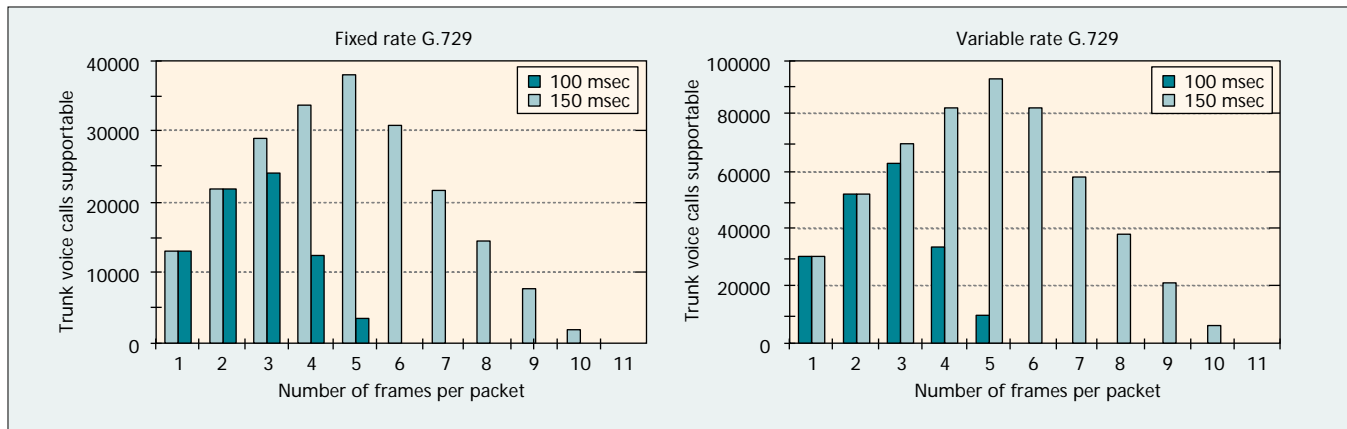
Considering this scenario and Eq. 1, we have

$$\begin{aligned}
 & VCDelay + PADelay + QuDelays + JitDelay \\
 & + PropDelay + VDDelay = \\
 & .015 + .010 \cdot N + QuDelays + JitDelay \\
 & + PropDelay + .010 = \quad (3) \\
 & .015 + .010 \cdot N + WC_{Deliv} + .010 \leq Target, \text{ or} \\
 & .010 \cdot N + WC_{Deliv} \leq Target - .025
 \end{aligned}$$

To maximize profits, a VoIP carrier desires to move as many voice calls as feasible over their VoIP backbone. Equation 3 illustrates two of the key trade-offs that can affect this: the number of frames in a packet and the tolerable delays at intervening switches. As one increases, the other must decrease. As the number of frames in a packet decreases, more time can be allocated to queuing delays, and trunk loads can be increased in order to carry more traffic, but a larger percentage of that traffic is packet overhead, not more phone calls. Conversely, as the number of frames in a packet increases, less time is available for packets to spend in intervening switches. Trunk loads must be decreased, reducing the number of phone calls that can be conveyed. Using a technique outlined in [21], Figs. 8a and 8b show plots of the number of voice calls supportable over a backbone OC-12 trunk for both the fixed-rate (Fig. 8a) and variable-rate (Fig. 8b) G.729 coders, for 100 msec and 150 msec target end-to-end delay bounds, using the network of Fig. 7. A target delivery time of 50 msec is seen to be impractical for this example network, even with one frame per packet, as from Equation 3 it is seen that WC_{Deliv} must be less than 15 msec, an infeasible value when the propagation delay alone is 20 msec. A similar situation occurs at 11 frames per packet with a target delivery time of 150 msec, and at six frames per packet and a target delivery time of 100 msec.

Figure 8 indicates that multiple frames per packet are necessary to maximize the number of voice calls the system can support, and that given some network configuration the choice of the number of voice frames to carry in a packet can significantly impact the number of calls the system is able to support. As a comparison, note that an OC-12 can carry 8,192 POTS phone calls. The combination of compression and silence suppression allows the VoIP system to potentially service a significantly larger number of customers.

The choice of receiver de-jitter buffer delay also has a con-



■ FIGURE 8. a) Fixed rate G.729 VoIP calls supportable over OC-12 trunks for 100 and 150 msec delivery delays; b) Variable rate G.729 VoIP calls supportable over OC-12 trunks for 100 and 150 msec delivery delays.

siderable impact on the number of calls the network can support. Equation 3 is based on the conservative choice of delaying the initial packet of a call or talk spurt such that it plays back WC_{Deliv} seconds after construction at the transmitter. This choice insures the buffer will not empty, but it delays play back an amount of time that is only necessary in the worst case, an event that may be unlikely. If the de-jitter buffer delay is decreased, time can be freed up and transferred to another entity of Equation 1, such as the packet assembly delay (allowing a reduction in the percentage of overhead) or delays spent at packet switches (allowing an increased trunk load). The interested reader is referred to [22] and [23] for further information on these issues.

What kind of quality does the E-Model predict for the network of Fig. 7? We use the following values in this example [20]:

- A base R_0 value of 94, corresponding to the value associated with a typical PSTN voice connection.
- A delay impairment value I_d of 2, 4, and 6 for echo delays of 50, 100, and 150 msec, respectively. This corresponds to a system using echo cancelers sufficient to generate a TELR of 60 dB in Fig. 6. Perfect echo cancellation requires a TELR of 65 dB.
- An equipment impairment value I_e of 10 for a G.729 fixed-rate coder, and 11 for a variable-rate coder. These values reflect the voice quality degradation of these reduced bit rate coders as compared to the standard G.711 coder.
- A and I_s are both set equal to zero, indicating the user expects similar quality on a VoIP phone as on a regular phone, and that there are no speech impairments other than those associated with the voice coder, as reflected in I_e .

Substituting these into Eq. 2 yields R values ranging from 78 to 82 for the fixed-rate coder, and 77 to 81 for the variable-rate coder (with the lower values associated with the higher end-to-end delivery delays). Note that these R values would be reduced further were packet losses of any significance.

If the minimum acceptable R rating is a 70, these all pass muster in terms of quality. But the values calculated do correspond to a perceived voice quality on the borderline between where most users are satisfied and where some of the more critical users will be unsatisfied.

ALTERNATE TRAFFIC SOURCES FOR A VOIP BACKBONE

Two other types of networks are likely to be connected to a carrier VoIP backbone and are mentioned briefly below.

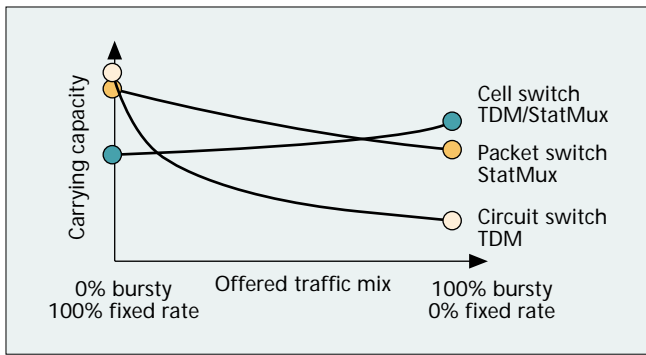
Instead of using standard PSTN techniques to interconnect mobile telephone cell sites to the rest of the world, VoIP

backbones can be used. An analysis of this type of network would differ from the network of Fig. 7 in several significant ways. First, the mobile phone voice coder would take longer than a G.711 coder to generate voice bits for movement over the mobile phone radio frequency (RF) link. This would reduce the time available for the VoIP backbone to move the traffic to the destination. If this destination were another mobile phone, the voice decoder would also require more decoding time than POTS' G.711. Second, the probability of voice frames being lost over the mobile phone's RF link will generally be higher than that of a wired link, adversely affecting the R rating. Third, were trans-coding necessary to convert the mobile phone-compressed voice to a different protocol being used on the VoIP backbone, some additional voice degradation would result compared to the example above, further impacting the R rating. In short, if a mobile telephone is the source or destination of a call being carried by a VoIP backbone, additional engineering attention must be focused on maintaining quality. A reduction in the allowed end-to-end delay (to reduce echo delay impairments) may be necessary to offset quality degradation caused by frame losses over the RF link and trans-coding degradation. A reduction in the number of supportable calls over the backbone would result.

Another type of system likely to be connected to a carrier VoIP backbone has one (or both) of the POTS PBX systems of Fig. 7 replaced with LAN-based VoIP systems. Traffic from the LAN VoIP phone will be mixed in with data traffic on the corporate LAN prior to being segregated and routed to the VoIP network gateway. It will be much more difficult for a carrier to offer and engineer high QoS when a significant part of the end-to-end network is on a LAN, especially if in the LAN the VoIP traffic is not prioritized. This situation will have some of the same characteristics as that of mobile phone voice sources and sinks, specifically higher delays at the LAN voice coders and decoders, possible trans-coding degradation at the gateways, and higher packet losses on the LAN. Engineering steps to offset these would be similar to that mentioned in the paragraph above, but would also include an increase required in the receiver de-jitter buffer delay to offset the jitter inherent with LAN traffic. Again, a reduced number of supportable calls over the backbone would likely be required in order to meet QoS goals.

WHAT ABOUT ATM?

Designed in the late 1980s for a mixed traffic environment, ATM already has many of the QoS protocols tested and in place that IP networks are now seeking to deploy.

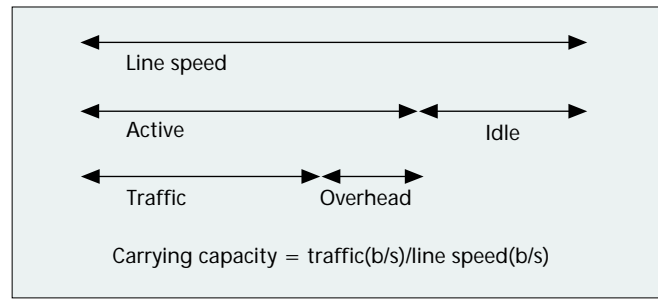


■ FIGURE 9. Switched network carrying capacities for high-speed trunks.

Figure 9 shows a plot of the carrying capacity for the three main digital technologies that have been deployed by telecommunications carriers over the last 40 years: circuit switching and TDM; cell switching and a combination of TDM and statistical multiplexing; and packet switching and statistical multiplexing. Carrying capacity is defined here as the ratio of application traffic moved, divided by the line speed required to carry it. From Fig. 10 it can be seen that this parameter accounts for both the inability of certain multiplexing techniques to fully load a line and maintain acceptable performance, and the overhead associated with moving the traffic.

Referring back to Fig. 9, the carrying capacity is plotted against offered traffic mixes ranging from 100 percent fixed-rate and no bursty traffic on the left hand side of the plot, to 100 percent bursty and no fixed-rate traffic on the right side. In the middle of the horizontal axis, the offered load is therefore half fixed-rate traffic and half bursty traffic such as data. Endpoints on this graph were plotted based on the following logic. If fixed-rate traffic is fed to a circuit-switched TDM network, trunk lines can be fully loaded. Likewise, an ATM network could map the fixed-rate traffic onto constant bit rate virtual circuits whose cells could be moved by TDM, also fully loading the trunk, but in this case slightly more than 10 percent of the bandwidth would be lost to ATM overhead. Fixed-rate traffic fed to a packet network would have to be packetized and statistically multiplexed onto the trunk bandwidth. Due to a mix of variable-length packets, traffic sources randomly entering and leaving, and the need to try to hold jitter down to negligible levels in order to mimic “fixed-rate” quality, we assumed for this plot that approximately 30 to 40 percent of the bandwidth will be lost due to overhead and an inability to fully load the trunk. Endpoints on the right hand side of the graph for the packet-switched and cell-switched points were assigned based on average packet sizes of 300 bytes requiring movement (including overhead) [24], and 70 to 80 percent statistically multiplexed trunk loads for the packet and cell networks. The cell network maps the packets into ATM cells. The circuit-switched TDM network’s right hand end point is assigned under the assumption that circuits (leased lines) do not have their bursty traffic aggregated. It is somewhat arbitrarily specified at 25 percent based on average traffic statistics observed on corporate Frame Relay access lines. Points in between the endpoints are weighted averages.

A more rigorous analysis would clearly be required to pinpoint the exact traffic mixes where the different technologies cross over in terms of their efficacy in moving the offered load. In [25] it is discussed in detail how this can be accomplished. Nevertheless, Fig. 9 does offer some insight into why different technologies have come to the forefront over the years. In the 1970s and 1980s the traffic mix offered to the typical telecommunications carrier was almost wholly fixed-rate voice. Figure 9 shows that particular mix is most effectively carried by a circuit-switched TDM network. ATM was



■ FIGURE 10. Carrying capacity. Accounting for overhead and idle bandwidth.

developed in the late 1980s and early 1990s, and was extensively adopted by providers as they realized that a cell-switched architecture allowing a combination of TDM and statistical multiplexing would be more cost effective in an environment with a mixture of fixed-rate and bursty traffic. As the traffic mix continues to become even more heavily data oriented at the beginning of the 21st century, Internet technologies are now justifiably receiving considerable attention.

In terms of its ability to support voice calls, Voice over ATM (VoATM) is superior to any other technology previously discussed in this article. If the routers of Fig. 7 are replaced by ATM switches, and fixed-rate 8 kb/s G.729 voice coders supported by CBR virtual circuits are used over the backbone, the OC-12s would be able to support approximately 55,000 phone calls. If, instead, variable-rate G.729B coders using silence suppression are used, the ATM system using variable bit rate/real-time virtual circuits would again be able to support a greater number of phone calls than the variable-rate VoIP system, at a comparable quality level.

However, the viability of ATM likely lies not in its ability to move voice, but in its ability to adapt to the changes in the offered traffic mixes faced by carriers today, specifically the increasing percentage of bursty data. The cost effectiveness of ATM is not so strong in this type of an environment.

CONCLUSIONS

This article has provided a tutorial overview of some of the problems and possible solutions identified with packet-switched voice transport. The focus was on carrier issues associated with preserving the quality of VoIP, but traditional POTS technology as well as non-traditional VoATM were also discussed. For papers that focus on other aspects of Internet telephony, especially the signaling process, the reader is referred to [26] and [27].

VoIP technology is clearly viable, but deploying a high-quality system requires careful system engineering, which is not a trivial task. Quality comparable to POTS can be obtained. Couple this with VoIP’s potential ability to support a greater number of phone calls than a POTS network of comparable bandwidth, and the recent shift to a traffic mix dominated by bursty computer-to-computer data that is best carried by packet switching, and the conclusion is that voice over the Internet is a technology with a promising future ... at least for the immediate future.

However, the communications industry does not stand still. As MPLS penetration increases and operators become more familiar with its use, voice over MPLS may see increased usage as, like VoIP, this technique can easily take advantage of voice compression, and it has considerably less overhead than that associated with a typical VoIP packet [28].

Additionally, there may be at least one more major paradigm shift in the coming years, that being the emergence of high-quality, high-bit-rate video that will require reliable

and timely delivery. Internet technologies may not be the best choice when video dominates the traffic mix offered to carriers [29]. If either of these proves to be the case, the transport techniques used to move voice will have to evolve once again.

ACKNOWLEDGMENTS

The authors wish to acknowledge the benevolent prodding of the editor and reviewers, which resulted in a much improved article than that originally submitted.

REFERENCES

- [1] J. Bellamy, *Digital Telephony*, 3rd Ed., John Wiley & Sons, 2000.
- [2] R. Cox, "Three New Speech Coders from the ITU Cover a Range of Applications," *IEEE Commun. Mag.*, Sept. 1997.
- [3] A. Benyassine *et al.*, "ITU-T Recommendation G.729 Annex B: A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications," *IEEE Commun. Mag.*, Sept. 1997.
- [4] P. Brady, "A Model for On-Off Speech Patterns in Two-Way Conversation," *Bell System Tech. Journal*, Sept. 1969.
- [5] P. P. Mishra and H. Saran, "Capacity Management and Routing Policies for Voice over IP Traffic," *IEEE Network*, Mar./Apr. 2000.
- [6] S. Shenker, "Fundamental Design Issues for the Future Internet," *IEEE JSAC*, Sept. 1995.
- [7] C. Metz, "IP QoS: Traveling in First Class on the Internet," *IEEE Internet Computing*, Apr. 1999, pp. 84–88.
- [8] X. Xiao and L. Ni, "Internet QoS: A Big Picture," *IEEE Network*, Mar./Apr. 1999.
- [9] T. Li, "MPLS and the Evolving Internet Architecture," *IEEE Commun. Mag.*, Dec. 1999.
- [10] C. Metz, "RSVP: General-Purpose Signaling for IP," *IEEE Internet Comp.*, June 1999.
- [11] M. Perkins *et al.*, "Speech Transmission Performance Planning in Hybrid IP/SCN Networks," *IEEE Commun. Mag.*, July 1999.
- [12] F. Schneider *et al.*, "Building Trustworthy Systems: Lessons from the PTN and Internet," *IEEE Internet Comp.*, Dec. 1999.
- [13] B. Li *et al.*, "QoS-Enabled Voice Support in the Next-Generation Internet: Issues, Existing Approaches, and Challenges," *IEEE Commun. Mag.*, Apr. 2000.
- [14] L. Mathy *et al.*, "The Internet: A Global Telecommunications Solution?," *IEEE Network*, July/Aug. 2000.
- [15] J. Janssen *et al.*, "Assessing Voice Quality in Packet-Based Telephony," *IEEE Internet Computing*, June 2002.
- [16] P. Noll, "Audio Coding," chapter in *The Communications Handbook*, 1997, CRC & IEEE Press.
- [17] P. Goyal *et al.*, "Integration of Call Signaling and Resource Management for IP Telephony," *IEEE Network*, May/June 1999.
- [18] M. Borella *et al.*, "Internet Packet Loss: Measurement and Implications for End-to-End QoS," *Proc. 1998 ICPP Wksp.*, Aug. 1998.
- [19] Nortel Networks White Paper, "Voice over Packet: An Assessment of Voice Performance on Packet Networks", <http://www.nortelnetworks.com/products/library/collateral/74007.25-09-01.pdf>
- [20] TSB116, "Voice Quality Recommendations for IP Telephony," *Telecommun. Industry Association*, Mar. 2001.
- [21] G. Scheets, R. Singh, and M. Parperis, "Analyzing End-to-End Delivery Delay in Pure VoIP Networks," *45th IEEE MWSCAS*, Aug. 2002.
- [22] D. Vleeschauwer *et al.*, "An Accurate Closed-Form Formula to Calculate the Dejittering Delay in Packetized Voice Transport," *Proc. IFIP-TC6/European Commission Int'l. Conf. Net.*, May 2000.
- [23] D. Vleeschauwer, J. Janssen, and G. Petit, "Delay Bounds for Low-Bit-Rate Voice Transport over IP Networks," *Proc. SPIE Conf. Perf. and Control of Network Systems III*, vol. 3841, Sept. 1999, pp. 40–48.
- [24] K. Thompson *et al.*, "Wide Area Internet Traffic Patterns and Characteristics," *IEEE Network*, Nov./Dec. 1997, p. 10–22.
- [25] G. Scheets and M. Allen, "Switched Network Carrying Capacities," chapter in *CRC Handbook of Communications Technologies: The Next Decade*, 2000, CRC Press.
- [26] B. Goode, "Voice Over Internet Protocol (VoIP)," *IEEE Proc.*, Sept. 2002.
- [27] W. Jiang *et al.*, "Integrating Internet Telephony Service," *IEEE Internet Computing*, June 2002.
- [28] D. Wright, "Voice over MPLS Compared to Voice over Other Packet Transport Technologies," *IEEE Commun. Mag.*, Nov. 2002.
- [29] J. Lubacz, "The IP Syndrome," *IEEE Commun. Mag.*, Feb. 2000.

BIOGRAPHIES

GEORGE SCHEETS (scheets@okstate.edu) received a B.S. in general engineering from the United States Military Academy at West Point in 1975, and following five years in the U.S. Army as a Signal Corps officer, he received M.S. and Ph.D. degrees in electrical engineering from Kansas State University in 1984 and 1987, respectively. He then joined the faculty at Oklahoma State University. His research interests include telecommunications network analysis and design, signal processing for intercept receivers, and teaching in a virtual environment.

MARIOS PARPERIS (marios.parperis@wgc.com) received the Higher National Diploma in electrical engineering from the Higher Technical Institute in Nicosia, Cyprus in 1991, and the B.Eng. and M.Eng. degrees in electrical engineering from Carleton University in Ottawa, Canada in 1997 and 1999, respectively. In 1999 he joined Newbridge Networks/Alcatel as a strategic network engineer. He is currently a network architect with WilTel Communications. His research interests are in voice quality issues in VoIP networks. He has been an IEE/IEEE member since 1990.

RITU SINGH (rt_singh@hotmail.com) received the B. A. degree with honors in English literature from Kurukshetra University, India, in 1980, B.S. (magna cum laude) and M.S. degrees in electrical engineering from the University of Tulsa, Tulsa, OK, in 1990 and 1996, respectively, and the Ph.D. degree in electrical engineering from Oklahoma State University in 2002. Currently she is president and CEO of NDRS Technologies and engaged in the development of software for medical and electromagnetic applications.