

Network Working Group  
Internet-Draft  
Expires: April 2, 2006

Y(J) Stein  
RAD Data Communications  
Sept 29, 2005

Great Real-Time Problem Statement  
draft-stein-great-00.txt

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with [Section 6 of BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on April 2, 2006.

Copyright Notice

Copyright (C) The Internet Society (2005).

Abstract

VoIP is commonly perceived to be a low quality, but low cost, alternative to standard telephony. This poor perception is often well deserved, being fueled by implementations designed without regard to characteristics of IP networks. This problem statement attempts to catalog the shortcomings of current implementations, in order to explore the IETF community's interest in working to improve this situation.

Internet-Draft

great

Sept 2005

## 1. Introduction

Consider the placing of a phone call over the PSTN. The end-user terminal is extremely simplistic and inexpensive, the scalability of the PSTN being based on 'dumb' terminals at end-points, with all the intelligence concentrated in the core. From the moment the user requests service by off-hooking, an imperceptible amount of time passes before the network indicates that is ready to receive signaling by delivering audible dial-tone. Since the service availability is 'five nines' (i.e. 99.999 percent) the user will probably not remember an event where dial-tone was not heard immediately after off-hooking. The user then enters the required part of a hierarchical destination address, and will then receive feedback as to the usage status of the destination terminal, in the form of ringback or busy tone, usually within seconds. Assuming that the destination terminal is not in use and the called party is present and decides to accept the call, a session is established an imperceptible time after off-hooking of the destination terminal.

For the duration of the conversation the voice is guaranteed to be 'toll quality' defined to be at least 4 on a Mean Opinion Scale (MOS) scale from 1 to 5. This quality is admittedly imperfect, due to the audio spectrum being truncated at 4 kHz (thus making differentiation of various unvoiced fricatives impossible, and distorting music) but preserves speaker identity and does not impede understandability for native speakers of the language spoken. There will, in general, be no unusual noises or audible artifacts (unless due to sources radiating close to the end-user terminals), and no gaps or discontinuities in the received information. Furthermore, the one-way propagation delay is usually close to the physically minimum possible (i.e. the time taken for light to travel between the two points) and no perceivable echo is introduced due to the telephone electronics. With extremely high probability the session will only be terminated when either the originator or called party decide to terminate.

Now, for comparison, let us consider a typical VoIP call over the public Internet. The end-user terminal may either be a personal computer (PC) or IP-phone, the former being a multifunctional computational device and the latter smaller and less computationally able, but relatively expensive terminal. Assuming a PC as terminal, the user initiates a call by typing an identifier if IP address, or by choosing the desired destination from a list. Thereafter follows

a rather prolonged period during which the user has no call progress feedback; the duration is usually longer for peer-to-peer systems, but is often considerable even for systems with centralized registries. Afterwards a simulation of ringback or busy-tone is commonly played, and assuming the destination terminal is powered and

the called party is willing to take the call, a bi-directional session is setup.

Once the session commences the voice quality will usually not be as good as that experienced in the PSTN (see [Section 3](#) for a discussion). In fact, the quality may be variable ranging from telephone-like to incomprehensible. Depending on network characteristics there will often be gaps when the sound completely disappears, or becomes metallic, or sounds like Martians are speaking. At times artifacts such as beeps may be heard. When using the public Internet the round-trip delay will often be so high (over one half second) that free conversation is impossible, and the parties to the conversation may repeatedly speak at the same time, or may purposely leave long pauses (for the other to interrupt or to aver that the connection is still operational), or say 'over' as in push-to-talk systems. The session may also terminate unexpectedly, and then may or may not be restored by reconnecting.

The sum total of the user's perception of the audio quality, delay, reliability and other factors is sometimes called the 'user experience'. Why would anyone use a VoIP systems if the user experience is significantly inferior to that of the standard telephone system? The situation is analagous to that of cellular phones, which also have noticeably lower audio quality and may unexpectedly disconnect, but the mediocre user experience is tolerated due to a new feature, namely mobility. Here some enthusiasts have suggested that the attraction of VoIP is due to the additional functionality that is, or will be, available (e.g. instant messaging, video). However, in most cases it is probably either the economics (free calls) or the ready accessibility for people already seated at a PC (along with presence indications) that induces most people to tolerate the poor quality. In fact, many times the latter type of user will start a conversation on VoIP in order to ask whether they can call over the PSTN. The feeling of most users is that the quality is good enough for casual, hobby type conversations (reminding some of us of our ham radio origins), and thus such users

are willing to use it to speak with remote acquaintances, mothers-in-law, etc. They might not, however, choose to use VoIP to call their bank branch, an important client, or their boss.

Of course, much of what was said above is specific to the present state of the public Internet, while well engineered, highly overprovisioned, networks suffer much less from these troubles. However, this does not mean that the public Internet is inherently unsuitable for quality transport of voice traffic, nor that it is imperative to make major changes in order for it to become suitable (although such changes may help). Many of the above problems can be amended, although not completely solved, by taking the

characteristics of the Internet into account at all stages of the VoIP implementation. We call such an implementation and its components, 'PSN-aware'.

The above discussion focused on VoIP, but similar statements could be made concerning other forms of real-time traffic transported over the Internet, such as videoconferencing. On the other hand not all real-time traffic is as problematic. For example, streaming audio that can be delivered after a certain delay may be able to exploit retransmission mechanisms, and thus be immunized to many of the above hindrances. The essential ingredients are real-time constraints and delay insensitivity, characteristics present in interactive real-time applications.

## [2.](#) Characteristics of PSNs

The design philosophy of the Public Switched Telephone Network (PSTN) presumes that routing is expensive but bandwidth plentiful, while that of Packet Switched Networks (PSNs), such as the Internet, presupposes bandwidth to be dear while routing affordable. The former tenets lead to a circuit switched network that naturally supports reliable and high quality interactive audio sessions, while the resource sharing required by the latter postulates makes providing such services a challenge.

The very fact that PSN users share bandwidth means that no user traffic receives treatment identical to that of a PSTN circuit. The major sources of performance degradation for real-time delay-sensitive PSN traffic can be identified as follows:

- \* packet creation time
- \* network propagation delay
- \* packet delay variation
- \* packet loss and mis-ordering
- \* congestion events
- \* lack of inherent timing transport
- \* bandwidth conservation algorithms
- \* emulation mechanisms

Unlike PSTN traffic, PSN traffic is sent in packets. The first byte of data placed in the packet experiences latency corresponding to the time required to fill the packet at the source. Although the last byte placed in the packet experiences only minimal delay, it is the last to be played out, and thus all data experiences latency equal to the packet creation time (PCT). In VoIP systems this may be less than 1 millisecond (for example When using the G.728 LD-CELP encoder), it is typically tens of milliseconds (for example 10 millisecond for G.729, 60 millisecond for a two-frame superpacket of G.723.1). PCT is a frame-size related latency introduced by the

source, but additional delay is usually added at the destination. Most speech decoders require 'lookahead', and (as will be discussed below) jitter buffer based systems require storing of packets. These additional delays may greatly increase the overall one-way delay.

While TDM switches typically add 1/8000 of a second latency per switch, Queuing delay in IP routers may be orders of magnitude higher.

This aforementioned latency is not constant from packet to packet, and successive packets do not even necessarily follow the same route. For these reasons packets injected into the PSN at a constant rate exit it at stochastic intervals. As we wish to play out audio at a constant rate, this packet delay variation (PDV) must be compensated. There are two ways this may be accomplished. In jitter buffer based systems Incoming packets are not directly played out, but rather placed in a 'jitter buffer' and later played out at a constant rate. The jitter buffer is usually configured to be able to absorb the maximum expected PDV, and thus introduces a significant amount of delay. In 'shock absorber' based systems packets are played out as they arrive, and when a packet is not yet available, a signal processing algorithm is employed to extrapolate based on previous

packets, until such time as a packet arrives. These systems introduce only minimal additional latency, but require considerably more computational power.

IP networks are intrinsically best-effort, and thus there is no guarantee that a packet injected into the PSN is actually received. In fact, all PSNs introduce some percentage of packet loss (PL), due to packets rejected due to detectable errors, packets dropped due to congested resources, and packets dropped due to policy decisions. Packet loss due to random errors will be independently distributed, but other types may cause bursts of lost packets. In addition, when parallel paths exist, packets may be received out-of-order, and must be either reordered (may be possible in jitter buffer based systems) or treated as lost. When a packet has not been received a decision must be made as to what to play out. One possibility is silence, but this will lead to reduced perceived audio quality. Depending on the expected percentage of packet loss, packet loss concealment (PLC) mechanisms may need to be employed.

Another consequence of the bandwidth sharing of PSNs is the possibility of congestion events, statistically infrequent peaks of activity during which there is insufficient bandwidth or processing power to transport all packets. For non-real-time traffic there are self-regulating rate control mechanisms, but for real-time traffic it is not clear that such mechanisms can be useful.

The PSTN is based on TDM networks that inherently transport timing information in the physical layer along with the data. PSNs do not include such a physical layer clock, and when such a clock is required, an appropriate mechanism must be supplied. This mechanism may rely on a clock source external to the PSN (e.g. GPS satellites), or may involve clock recovery over the PSN itself (e.g. NTP).

By bandwidth conservation algorithms we mean all source codings employed for reduction of data rate to closer to the Shannon rate. These range from lossless data compression, through speech encoding, fax image encoding, to video encoding. Except for lossless compression, all such mechanisms introduce some quality reduction, and all (including lossless compression) reduce robustness to errors and packet loss.

The final source of degradation is emulation mechanisms internal to gateways that enable access to the PSN. These mechanisms may try to simulate behavior of a PSTN system, to terminate or relay PSTN-specific signaling, or to optimize operation of interactive real-time traffic over the PSN. These mechanisms are typically required to detect various characteristics of the incoming real-time signals, and need to do so rapidly, with high probability of detection, and with low false alarm rate. When such a mechanism fails, the gateway may enter a state from which it may take time to exit, creating a severe anomaly in user perceived performance.

### 3. Bandwidth and Audio Quality Problems

Even assuming a perfect PSN, i.e. one with no packet loss (PL) nor packet mis-ordering and only minimal packet delay variation (PDV), the perceived voice quality of VoIP calls is highly dependent on bandwidth reduction mechanisms. First, in order to minimize bandwidth consumption speech encoding algorithms are employed that reduce the MOS to somewhere between 3.5 and 3.8. Second, voice activity detection (VAD) is typically employed to mute (or replace with locally generated 'comfort noise') one direction of the conversation; this VAD is never perfect and may clip the start of voice spurts. Due to the speech compressions not passing various tones (e.g. DTMF), are passed using special relay functions; false alarms in such detection produce annoying beeps known as 'talk-offs'.

When the present generation of speech encoders was developed, the only design criteria were compression ratio, speech quality (MOS), and to a certain degree delay (although G.723.1 was supposedly designed with VoIP in mind, its round-trip combined delay of 75 milliseconds is not conducive to use over the public Internet). At about the same time speech encoders were developed for satellite

applications that were built to be robust to individual bit errors; but no encoders were built to be robust to loss of entire packets. Indeed, even the common event of the loss of a single packet may cause a disruption to the decoded audio that may last for a long time. Later the iLBC speech coder (described in [RFC 3952](#)) was designed to eliminate this problem (and today other encoder techniques are known that are inherently insensitive to missing data). When the packet loss problem was better understood, PLC

mechanisms were added to speech encoders used over PSNs, but these PLCs helped mainly for loss of isolated packets. Typical PL patterns of IP networks (e.g. loss bursts) were not taken into account.

As the development of speech encoding algorithms has in general proceeded without detailed knowledge of PSN characteristics, required functionality, such as PLC, has been added on a posteriori. Higher efficiency and performance may be gained by a priori design of PSN-aware speech and other audio (and later video) encoders and PSN-aware PLC mechanisms.

In addition, when the end-user terminals are no longer POTS phones, one may ask why we are still limiting ourselves to 4 kHz bandwidth. Wideband telephony (8 kHz bandwidth) speech is noticeably superior, and may go far to convincing users that VoIP quality may actually exceed that of the PSTN. Design of standardized PSN-aware wideband encoders is a worthwhile task waiting to be tackled.

Most speech encoders used today take in a constant number of bytes of uncompressed audio, and produce a constant number of compressed bytes. Some speech coders are called adaptive multirate, in that they may be configured to produce a specified number of compressed bytes. Truly variable rate compression techniques vary in output rate according to the character of the input sounds. While the use of constant rate transport infrastructures dictates constant rate encoders, PSN packets may vary in size from packet to packet, and thus variable rate encoders may be used. It is an open question as to how to match these encoder parameters to PSN characteristics.

#### [4.](#) Delay and Delay Variation Problems

Standard PSTN practice places tight constraints on the tolerable end-to-end and round-trip delays. Although the more modern approach is to consider the effect of delay along with other degradations, one-way transmission times of up to 150 milliseconds are considered universally acceptable, assuming adequate echo control is provided. Echo cancellation is required when the delay exceeds about 20 milliseconds.

The one-way delay in PSNs is greater than that of the PSTN, due at

very least to PCT and lookahead, and often to queuing delays and



jitter buffer latency. Indeed, network propagation times alone may be in the 100 millisecond range, and thus incompatible with the minimum delay introduced by G.723.1. Thus a sensible approach would be to start with a specification of the network delay, and to derive allowable buffering and processing budgets. This would probably require smaller frame sizes and minimization of lookahead, and innovative designs would be needed to keep bit rates reasonable.

More attention should be drawn to the perfection of shock absorber based systems. These may need to be more fully integrated into the encoder, perhaps more specifically into the PLC mechanism.

## 5. Congestion Problems

When congestion is detected, either by explicit notification or via detection of packet loss, even real-time systems should heed the network's warning of imminent trouble. In addition to PLC on any missing packet, in the other direction rate cutback needs to be attempted, e.g. by lowering VAD thresholds, via adaptation of the rate of adaptive multirate encoders or the average output rate parameter of variable rate encoders, and in extreme cases by deliberate dropping of packets that are likely to be more effectively concealed by the PLC. Although all these activities reduce the user's perception of voice quality, they do so less drastically than complete loss of all audio.

Adaptive multirate encoders can generally change rate on a packet by packet basis in 'hitless' fashion, but it is unknown how to do this when changing encoder. There has not been sufficient study of how to identify packets that may be less harmful to discard.

## 6. Emulation Problems

The lack of precise clock synchronization between source and destination (play out) clocks is usually considered unimportant for voice. This is because even a missed or extra speech sample every few minutes is undetectable to the ear. The situation is different when the system is used to transport non-speech data, such as fax and data modem transfer without appropriate relays. In such cases it is necessary to match the destination clock to that of the source in order to eliminate sample slips.

Accurate (line or acoustic) echo cancellation is essential for high ratings of user experience. At present echo cancellation is typically performed where its computational cost is minimized, i.e. close to the place where the echo is generated, rather than where it would be heard. It would be useful to be able to employ an echo

cancellation server anywhere in the network, but there are problems that need to be solved before this can be accomplished. For example, the relative timing of the signals flowing in opposite directions needs to be determined (including clock synchronization), and the fact that neither signal may be echo-free.

Real-time monitoring of voice quality has been previously considered. Such measures may be based on acoustic models or on measurement of network degradations and use of previously determined calibrations. Timely feedback of such end-to-end information quality may be useful in improving the audio quality, but the precise mechanisms need to be worked out.

Another problem that may be addressed concern multi-user conferencing. Many present-day systems choose a single dominant speaker, squelching others desiring to talk. This introduces various perceived quality degradations, in addition to giving a bad impression to the user wanting to 'break in'. Complete summing of audio from all users is problematic for several reasons. It requires decompression and recompression of user audio, and rescaling to avoid excessive signal levels. Advances would be welcome here.

Reduction of the connection setup delay, and the related delays for entering/exiting fax-relay and modem-relay modes is an important signalling problem to be solved.

Integration of real-time delay-sensitive traffic along a time line with other applications may be interesting. The most important application here is lip syncing, but syncing text for Karaoke, whiteboard motions to spoken words, etc. may need to be addressed.

## 7. Security Considerations

Although not directly related to the real-time character of the traffic authentication, encryption, and methods for lawful interception (CALEA) need to be integrated in a standard way into VoIP systems.

## 8. IANA Considerations

This Internet Draft does not propose a protocol, nor a change to any existing protocol, and thus no IANA considerations are raised.

Internet-Draft

great

Sept 2005

Author's Address

Yaakov (J) Stein  
RAD Data Communications  
24 Raoul Wallenberg St., Bldg C  
Tel Aviv 69719  
ISRAEL

Phone: +972 3 645-5389  
Email: yaakov\_s@rad.com

---

Internet-Draft

great

Sept 2005

### Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in [BCP 78](#) and [BCP 79](#).

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at [ietf-ipr@ietf.org](mailto:ietf-ipr@ietf.org).

### Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE

INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED  
WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

#### Copyright Statement

Copyright (C) The Internet Society (2005). This document is subject to the rights, licenses and restrictions contained in [BCP 78](#), and except as set forth therein, the authors retain all their rights.

#### Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.