                    PW Congestion Considerations
                    draft-ietf-pwe3-congcons-01

Abstract

   Pseudowires (PWs) have become a common mechanism for tunneling
   traffic, and may be found competing for network resources both with
   other PWs and with non-PW traffic, such as TCP/IP flows.  It is thus
   worthwhile specifying under what conditions such competition is safe,
   i.e., the PW traffic does not significantly harm other traffic or
   contribute more than it should to congestion.  We conclude that PWs
   transporting responsive traffic behave as desired without the need
   for additional mechanisms.  For inelastic PWs (such as TDM PWs) we
   derive a bound under which such PWs consume no more network capacity
   than a TCP flow.

Status of this Memo

Copyright Notice

Table of Contents

1.  Introduction

   A pseudowire (PW) is a construct for tunneling a native service over
   a Packet Switched Network (PSN)(see [RFC3985]), such as IPv4, IPv6,
   or MPLS.  The PW packet encapsulates a unit of native service
   information by prepending the headers required for transport in the
   particular PSN (which must include a demultiplexer field to
   distinguish the different PWs) and preferably the 4 byte PWE3 control
   word.  PWs have no bandwidth reservation mechanism, meaning that when
   multiple PWs are transported in parallel there is no defined means
   for guaranteeing network resources for any particular PW.  This
   competition for resources may translate to a particular PW not being
   able to deliver the QoS required to emulate the native service.  For
   example, MPLS-TE enables achieving a particular desired allocation of
   resources between multiple LSPs; however, when multiple Ethernet PWs
   are placed in a single MPLS tunnel, there is no way to similarly
   divide resources amongst them (although DiffServ QoS prioritization
   may be available for PWs).  The use of PWs in service provider MPLS
   networks is well understood and will not be discussed further here.

   While PWs are most often placed in MPLS tunnels, there are several
   mechanisms that enable transporting PWs over an IP infrastructure.
   These include:
      TDM PWs ([RFC4553][RFC5086][RFC5087]) that define UDP/IP
      encapsulations,
      L2TPv3 PWs,
      MPLS PWs directly over IP according to RFC 4023 [RFC4023],
      MPLS PWs over GRE over IP according to RFC 4023 [RFC4023].
   Whenever PWs are transported over IP, they may compete with
   congestion-responsive flows (e.g., TCP flows).  Hence in order to
   prevent congestion collapse the PWs MUST behave in a fashion that
   does not cause undue damage to the throughput of such congestion-
   responsive flows [RFC2914].

   At first glance one may think that this would require a PW
   transported over IP to be considered as a single flow, on a par with
   a single TCP flow.  Were we to accept this tenet, we would require a
   PW to back off under congestion to consume no more bandwidth than a
   single TCP flow under such conditions (see [RFC5348]).  However,
   since PWs may carry traffic from many users, it makes more sense to
   consider each PW to be equivalent to multiple TCP flows.  We will
   discuss whether PWs consisting of elastic flows need a back-off
   strategy in Section 2.

   TDM PWs ([RFC4553][RFC5086][RFC5087]) represent inelastic constant
   bit-rate (CBR) flows that may require lower or higher throughput than
   that consumed by an otherwise-unconstrained TCP flow would under the
   same network conditions.  In any case a TDM PW is not able to respond

to congestion in a TCP-like manner; on the other hand, the total
bandwidth they consume remains constant and does not increase to
consume additional bandwidth as TCP rates back off.  If the bandwidth
consumed by a TDM PW is considered detrimental, the only available
remedy is to completely shut down the PW.  Such a shutdown would
impact multiple users, and the service restoration time would in
general be lengthy.  We will discuss when the shutdown of inelastic
PWs can be avoided in Section 3.


2.  PWs Comprising Elastic Flows

   In this section we consider Ethernet PWs that primarily carry
   congestion-responsive traffic.  We will show that we automatically
   obtain the desired congestion avoidance behavior, and that additional
   mechanisms are not needed.

   Let us assume that an Ethernet PW aggregating several TCP flows is
   flowing alongside several TCP/IP flows.  Each Ethernet PW packet
   carries a single Ethernet frame that carries a single IP packet that
   carries a single TCP segment.  Thus, if congestion is signaled by an
   intermediate router dropping a packet, a single end-user TCP/IP
   packet is dropped, whether or not that packet is encapsulated in the
   PW.

   The result is that the individual TCP flows inside the PW experience
   the same drop probability as the non-PW TCP flows.  Thus the behavior
   of a TCP sender (retransmitting the packet and appropriately reducing
   its sending rate) is the same for flows directly over IP and for
   flows inside the PW.  In other words, individual TCP flows are
   neither rewarded nor penalized for being carried over the PW.  On the
   other hand, the PW does not behave as a single TCP flow; it will
   consume the aggregated bandwidth of its component flows, and backs
   off much less sharply than a single flow would.

   We claim that this is precisely the desired behavior.  Any fairness
   considerations should be applied to the individual TCP flows, and not
   to the aggregate.  Were individual TCP flows rewarded for being
   carried over a PW, this would create an incentive to create PWs for
   no operational reason.  Were individual flows penalized, there would
   be a deterrence that could impede pseudowire deployment.

   There have been proposals to add additional TCP-friendly mechanisms
   to PWs, for example by carrying PWs over DCCP.  In light of the above
   arguments, it is clear that this would force the PW to behave as a
   single flow, rather than N flows, and penalize the constituent TCP
   flows.  In addition, the individual TCP flows would still back off
   due to their end points being oblivious to the fact that they are

carried over a PW.  This will further degrade the flow's throughput
as compared to a non-PW-encapsulated flow.  Thus, such additional
mechanisms contradict the behavior previously described as desirable.


3.  PWs Comprising Inelastic Flows

   TDM PWs ([RFC4553][RFC5086][RFC5087]) are more problematic than the
   elastic PWs of the previous section.  Being constant bit-rate (CBR),
   they can not be made responsive to congestion.  On the other hand,
   being CBR, they also do not attempt to capture additional bandwidth
   when TCP flows back off.

   Since a TDM PW continuously consumes a constant amount of bandwidth,
   if the bandwidth occupied by a TDM PW endangers the network as a
   whole, the only recourse is to shut it down, denying service to all
   customers of the TDM native service.  We should mention in passing
   that under certain conditions it may be possible to reduce the
   bandwidth consumption of a TDM PW.  A prevalent case is that of a TDM
   native service that carries voice channels that may not all be
   active.  Using the AAL2 mode of [RFC5087] (perhaps along with
   connection admission control) can enable bandwidth adaptation, at the
   expense of more sophisticated native service processing (NSP).

   In the following we will show that for many cases of interest a TDM
   PW, treated as a single flow, will behave in a reasonable manner
   without any additional mechanisms.  We will focus on structure-
   agnostic TDM PWs [RFC4553] although our analysis can be readily
   applied to structure-aware PWs (see Appendix A).

   There are two network parameters relevant to our discussion, namely
   the one-way delay D and the loss probability p.  The one-way delay of
   a native TDM service consists of the physical time-of-flight plus 125
   microseconds for each TDM switch traversed.  This is very small as
   compared to PSN network-crossing latencies.  Many protocols and
   applications running over TDM circuits thus require low delay, and we
   need thus only consider delays of up to about 32 milliseconds.

   The TDM PW RFCs specify the egress behavior upon experiencing packet
   loss.  Structure-agnostic transport has no alternative to outputting
   an "all-ones" AIS pattern towards the TDM circuit, which if long
   enough in duration is recognized by the receiving TDM device as a
   fault indication (see Appendix A).  International standards place
   stringent limits on the number of such faults tolerated.
   Calculations presented in the appendix show that only loss
   probabilities in the realm of fractions of a percent are relevant for
   structure-agnostic transport (see Appendix A).

Structure-aware transport regenerates frame alignment signals thus
hiding AIS indications resulting from infrequent packet loss.
Furthermore, for TDM circuits carrying voice channels the use of
packet loss concealment algorithms is possible (such algorithms have
been previously described for TDM PWs).  However, even structure-
aware transport ceases to provide a useful service at about 2 percent
loss probability.

RFC 5348 on TCP Friendly Rate Control (TFRC) [RFC5348] provides the
following simplified formula for throughput that is used as the basis
for TFRC's sending rate control.

$$X\_Bps = \frac{S}{R \; ( \sqrt{2p/3} + 12 \sqrt{3p/8} \; p \; (1+32p^2) )}$$

where
    X_Bps is average sending rate in Bytes per second,
    S is the segment (packet payload) size in Bytes,
    R is the round-trip time in seconds,
    p is the loss probability.

We can use this formula to determine when a TDM PW consumes no more
bandwidth than a TCP flow between the same endpoints would consume
under the same conditions.  Replacing the round-trip delay with twice
the one-way delay D, setting the bandwidth to that of the TDM service
BW, and the segment size to be the TDM fragment TDM plus 4 Bytes to
account for the PWE3 control word, we obtain the following condition
for a TDM PW.

$$D < \frac{(TDM + 4)}{BW \; f(p) \; / \; 4}$$

where
    D is the one-way delay,
    TDM is the TDM segment size in Bytes,
    BW is TDM service bandwidth in bits per second,
    $f(p) = \sqrt{2p/3} + 12 \sqrt{3p/8} \; p \; (1+32p^2)$.

One may view this condition as defining a safe operating envelope for
a TDM PW, as a TDM PW that consumes no more bandwidth than a TCP flow
would not affect congestion more than were it to be TCP traffic.
Under this condition it should hence be safe to mix the TDM PW with
congestion-responsive traffic such as TCP, without causing
significant additional congestion problems.  Were the TDM PW to
consume significantly more bandwidth a TCP flow, it could contribute
disproportionately to congestion, and its mixture with congestion-

responsive traffic may be inappropriate.

We derived the condition assuming steady-state conditions, and thus two caveats are in order.  First, the condition does not specify how to treat a TDM PW that initially satisfies the condition, but is then faced with a deteriorating network environment.  In such cases one additionally needs to analyze the reaction times of the responsive flows to congestion events.  Second, the derivation assumed that the TDM PW was competing with long-lived TDM flows, because under this assumption it was straightforward to obtain a quantitative comparison with something widely considered to offer a safe response to congestion.  Short-lived TCP flows may find themselves disadvantaged as compared to a long-lived TDM PW satisfying the condition.  These dynamic cases will be considered in future versions of this draft.

The results are displayed in the accompanying figures (available only in the PDF version of this document).  TCP compatible behavior is obtained for the area under curves appropriate for each TDM fragment size.

We see in Figure 1 that a TDM PW carrying an E1 native service (2.048 Mbps) satisfies the condition for all parameters of interest if each packet carries at least S=512 Bytes of TDM data.  For the SAToP default of 256 Bytes, as long as the one-way delay is less than 10 milliseconds, the loss probability can exceed 0.3 percent.  For packets containing 128 or 64 Bytes the constraints are more troublesome, but there are still parameter ranges where the TDM PW consumes less than a TCP flow under similar conditions.  Similarly, Figure 2 demonstrates that an E3 native service (34.368 Mbps) with the SAToP default of 1024 Bytes of TDM per packet satisfies the condition for delays up to about 5 milliseconds.

Note that violating the condition for a short amount of time is not sufficient justification for shutting down the TDM PW.  While TCP flows react within a round trip time, PW commissioning and decommissioning are time consuming processes that should only be undertaken when it becomes clear that the congestion is not transient.  Future versions of this draft will provide guidance as to when a TDM PW should be terminated.

Figure 1 TCP Compatibility areas for T1 SAToP

Figure 2 TCP Compatibility areas for E1 SAToP
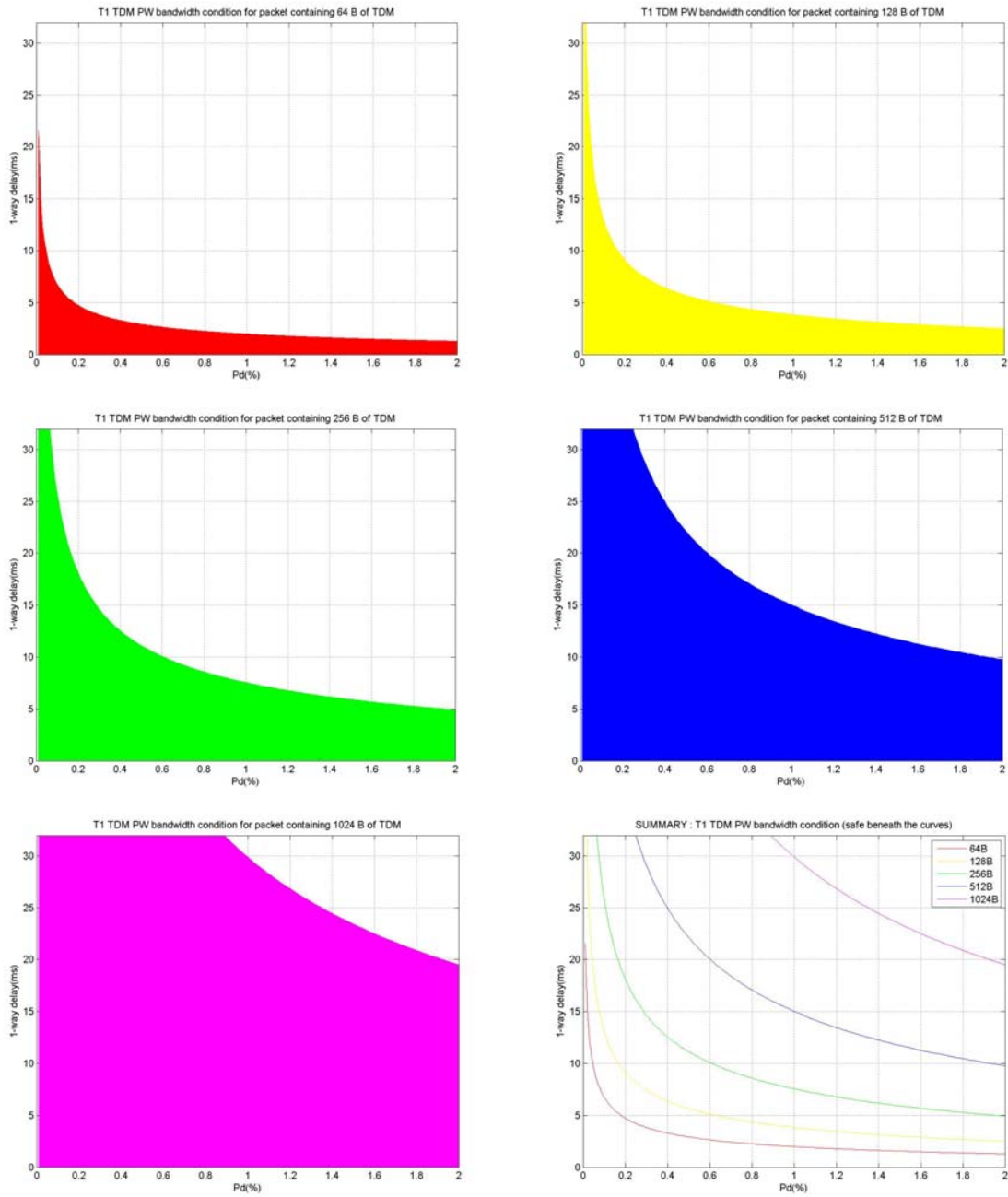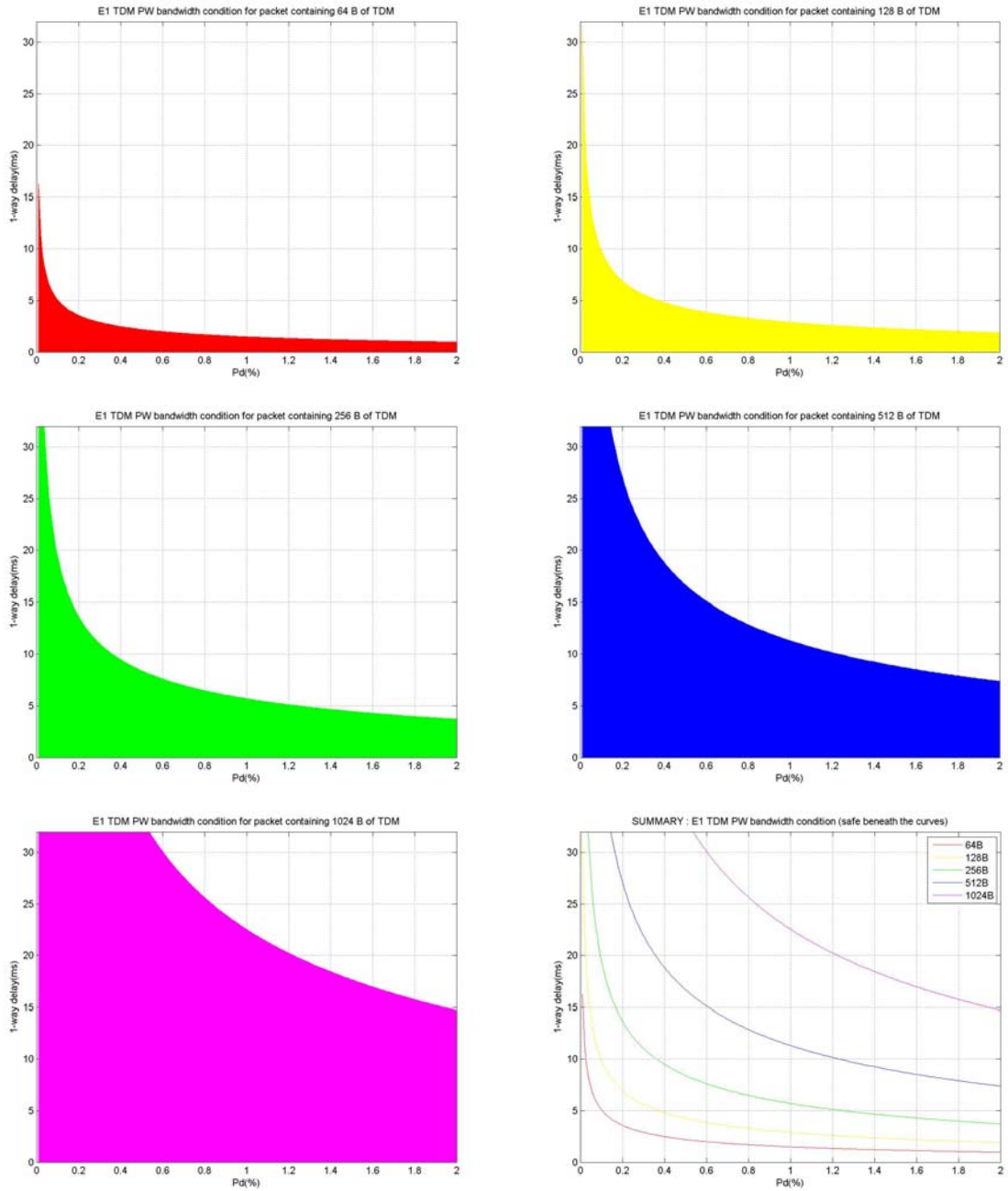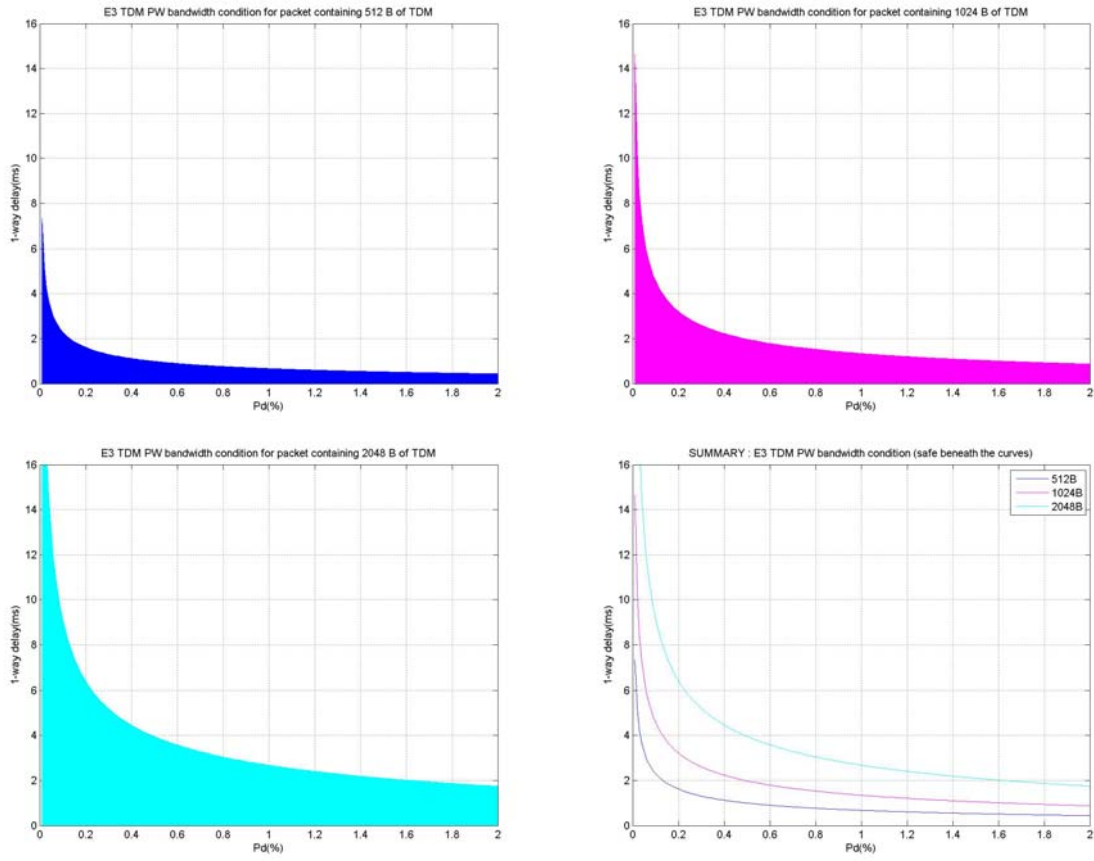
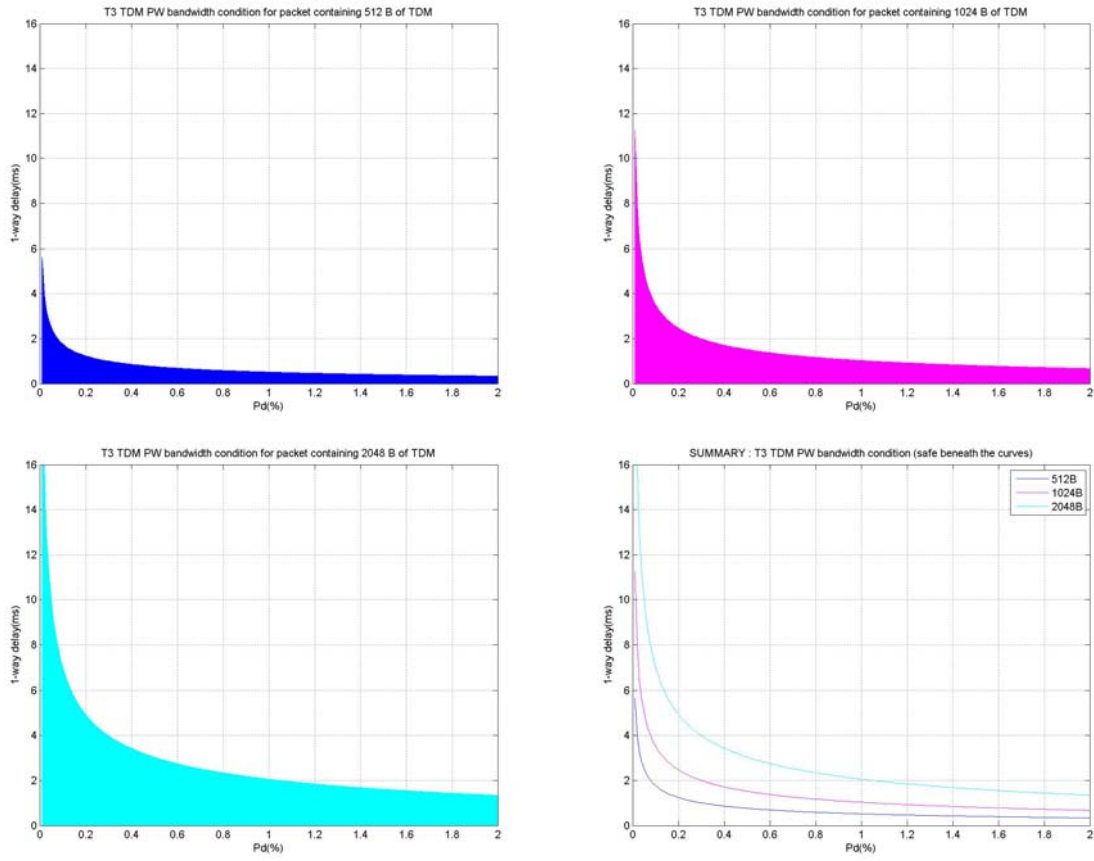Figure 3 TCP Compatibility areas for E3 SAToP

Figure 4 TCP Compatibility areas for T3 SAToP

4.  Security Considerations

   This document does not introduce any new congestion-specific
   mechanisms and thus does not introduce any new security
   considerations above those present for PWs in general.


5.  IANA Considerations

   This document requires no IANA actions.

6.  Informative References

   [RFC2914]   Floyd, S., "Congestion Control Principles", BCP 41,
               RFC 2914, September 2000.

   [RFC3985]   Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-
               Edge (PWE3) Architecture", RFC 3985, March 2005.

   [RFC4023]   Worster, T., Rekhter, Y., and E. Rosen, "Encapsulating
               MPLS in IP or Generic Routing Encapsulation (GRE)",
               RFC 4023, March 2005.

   [RFC4553]   Vainshtein, A. and YJ. Stein, "Structure-Agnostic Time
               Division Multiplexing (TDM) over Packet (SAToP)",
               RFC 4553, June 2006.

   [RFC5086]   Vainshtein, A., Sasson, I., Metz, E., Frost, T., and P.
               Pate, "Structure-Aware Time Division Multiplexed (TDM)
               Circuit Emulation Service over Packet Switched Network
               (CESoPSN)", RFC 5086, December 2007.

   [RFC5087]   Stein, Y(J)., Shashoua, R., Insler, R., and M. Anavi,
               "Time Division Multiplexing over IP (TDMoIP)", RFC 5087,
               December 2007.

   [RFC5348]   Floyd, S., Handley, M., Padhye, J., and J. Widmer, "TCP
               Friendly Rate Control (TFRC): Protocol Specification",
               RFC 5348, September 2008.

   [G775]      International Telecommunications Union, "Loss of Signal
               (LOS), Alarm Indication Signal (AIS) and Remote Defect
               Indication (RDI) defect detection and clearance criteria
               for PDH signals", ITU Recommendation G.775, October 1998.

   [G826]      International Telecommunications Union, "Error Performance
               Parameters and Objectives for International Constant Bit
               Rate Digital Paths at or above Primary Rate",
               ITU Recommendation G.826, December 2002.

   [P862]      International Telecommunications Union, "Perceptual
               evaluation of speech quality (PESQ): An objective method
               for end-to-end speech quality assessment of narrow-band
               telephone networks and speech codecs", ITU Recommendation
               G.826, February 2001.

   [I-D.stein-pwe3-tdm-packetloss]
               Stein, Y(J). and I. Druker, "The Effect of Packet Loss on
               Voice Quality for TDM over Pseudowires", October 2003.

Appendix A.  Loss Probabilities for TDM PWs

   ITU-T Recommendation G.826 [G826] specifies limits on the Errored
   Second Ratio (ESR) and the Severely Errored Second Ratio (SESR).  For
   our purposes, we will simplify the definitions and understand an
   Errored Second (ES) to be a second of time during which a TDM bit
   error occurred or a defect indication was detected.  A Severely
   Errored Second (SES) is an ES second during which the Bit Error Rate
   (BER) exceeded one in one thousand ($10^{-3}$).  Note that if the error
   condition AIS was detected according to the criteria of ITU-T
   Recommendation G.775 [G826] a SES was considered to have occurred.
   The respective ratios are the fraction of ES or SES to the total
   number of seconds in the measurement interval.

   For both E1 and T1 TDM circuits, G.826 allows ESR of 4% (0.04), and
   SESR of 1/5% (0.002).  For E3 and T3 the ESR must be no more than
   7.5% (0.075), while the SESR is unchanged.

   Focusing on E1 circuits, the ESR of 4% translates, assuming the worst
   case of isolated exactly periodic packet loss, to a packet loss event
   no more than every 25 seconds.  However, once a packet is lost,
   another packet lost in the same second doesn't change the ESR,
   although it may contribute to the ES becoming a SES.  Assuming an
   integer number of TDM frames per PW packet, the number of packets per
   second is given by packets per second = 8000 / (frames per packet),
   where prevalent cases are 1, 2, 4 and 8 frames per packet.  Since for
   these cases there will be 8000, 4000, 2000, and 1000 packets per
   second, respectively, the maximum allowed packet loss probability is
   0.0005%, 0.001%, 0.002%, and 0.004% respectively.

   These extremely low allowed packet loss probabilities are only for
   the worst case scenario.  In reality, when packet loss is above
   0.001%, it is likely that loss bursts will occur.  If the lost
   packets are sufficiently close together (we ignore the precise
   details here) then the permitted packet loss rate increases by the
   appropriate factor, without G.826 being cognizant of any change.
   Hence the worst-case analysis is expected to be extremely pessimistic
   for real networks.  Next we will go to the opposite extreme and
   assume that all packet loss events are in periodic loss bursts.  In
   order to minimize the ESR we will assume that the burst lasts no more
   than one second, and so we can afford to lose no more than packet per
   second packets in each burst.  As long as such one-second bursts do
   not exceed four percent of the time, we still maintain the allowable
   ESR.  Hence the maximum permissible packet loss rate is 4%.  Of
   course, this estimate is extremely optimistic, and furthermore does
   not take into consideration the SESR criteria.

   As previously explained, a SES is declared whenever AIS is detected.

There is a major difference between structure-aware and structure-agnostic transport in this regards.  When a packet is lost SAToP outputs an "all-ones" pattern to the TDM circuit, which is interpreted as AIS according to G.775 [G775].  For E1 circuits, G.775 specifies for AIS to be detected when four consecutive TDM frames have no more than 2 alternations.  This means that if a PW packet or consecutive packets containing at least four frames are lost, and four or more frames of "all-ones" output to the TDM circuit, a SES will be declared.  Thus burst packet loss, or packets containing a large number of TDM frames, lead SAToP to cause high SESR, which is 20 times more restricted than ESR.  On the other hand, since structure-aware transport regenerates the correct frame alignment pattern, even when the corresponding packet has been lost, packet loss will not cause declaration of SES.  This is the main reason that SAToP is much more vulnerable to packet loss than the structure-aware methods.

For realistic networks, the maximum allowed packet loss for SAToP will be intermediate between the extremely pessimistic estimates and the extremely optimistic ones.  In order to numerically gauge the situation, we have modeled the network as a four-state Markov model, (corresponding to a successfully received packet, a packet received within a loss burst, a packet lost within a burst, and a packet lost when not within a burst).  This model is an extension of the widely used Gilbert model.  We set the transition probabilities in order to roughly correspond to anecdotal evidence, namely low background isolated packet loss, and infrequent bursts wherein most packets are lost.  Such simulation shows that up to 0.5% average packet loss may occur and the recovered TDM still conform to the G.826 ESR and SESR criteria.

Appendix B.  Effect of Packet Loss on Voice Quality for TDM PWs

   Packet loss in voice traffic can cause in gaps or artifacts that
   result in choppy, annoying or even unintelligible speech.  The
   precise effect of packet loss on voice quality has been the subject
   of detailed study in the VoIP community, but VoIP results are not
   directly applicable to TDM PWs.  This is because VoIP packets
   typically contain over 10 milliseconds of the speech signal, while
   multichannel TDM packets may contain only a single sample, or perhaps
   a very small number of samples.

   The effect of packet loss on TDM PWs has been previously reported
   [I-D.stein-pwe3-tdm-packetloss].  In that study it was assumed that
   each packet carried a single sample of each TDM timeslot (although
   the extension to multiple samples is relatively straightforward and
   does not drastically change the results).  Four sample replacement
   algorithms were compared, differing in the value used to replace the
   lost sample:
   1.  replacing every lost sample by a preselected constant (e.g., zero
       or "AIS" insertion),
   2.  replacing a lost sample by the previous sample,
   3.  replacing a lost sample by linear interpolation between the
       previous and following samples,
   4.  replacing the lost sample by STatistically Enhanced INterpolation
       (STEIN).
   Only the first method is applicable to SAToP transport, as structure
   awareness is required in order to identify the individual voice
   channels.  For structure aware transport, the loss of a packet is
   typically identified by the receipt of the following packet, and thus
   the following sample is usually available.  The last algorithm posits
   the LPC speech generation model and derives lost samples based on
   available samples both before and after each lost sample.

   The four algorithms were compared in a controlled experiment in which
   speech data was selected from English and American English subsets of
   the ITU-T P.50 Appendix 1 corpus [P.50App1] and consisted of 16
   speakers, eight male and eight female.  Each speaker spoke either
   three or four sentences, for a total of between seven and 15 seconds.
   The selected files were filtered to telephony quality using modified
   IRS filtering and downsampled to 8 KHz.  Packet loss of 0, 0.25, 0.5,
   0.75, 1, 2, 3, 4 and 5 percent were simulated using a uniform random
   number generator (bursty packet loss was also simulated but is not
   reported here).  For each file the four methods of lost sample
   replacement were applied and the Mean Opinion Score (MOS) was
   estimated using PESQ [P862].  Figure 5 depicts the PESQ-derived MOS
   for each of the four replacement methods for packet drop
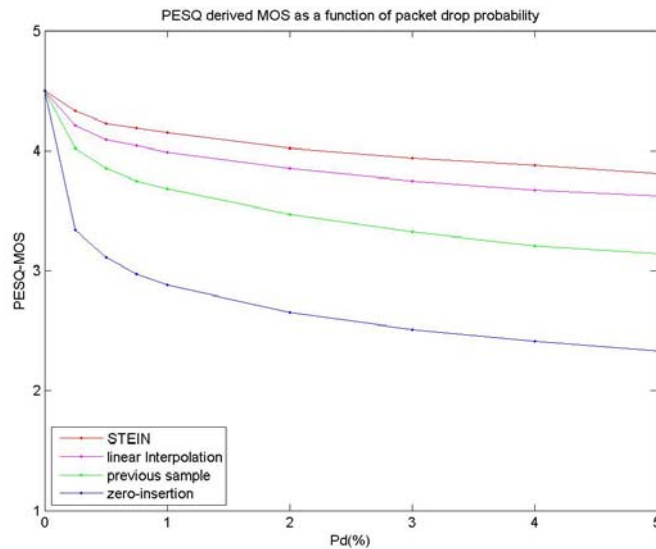   probabilities up to 5%.

Figure 5 PESQ derived MOS as a function of packet drop probability


For all cases the MOS resulting from the use of zero insertion is
less than that obtained by replacing with the previous sample, which
in turn is less than that of linear interpolation, which is slightly
less than that obtained by statistical interpolation.

Unlike the artifacts speech compression methods may produce when
subject to buffer loss, packet loss here effectively produces
additive white impulse noise.  The subjective impression is that of
static noise on AM radio stations or crackling on old phonograph
records.  For a given PESQ-derived MOS, this type of degradation is
more acceptable to listeners than choppiness or tones common in VoIP.

If MOS>4 (full toll quality) is required, then the following packet
drop probabilities are allowable:
    zero insertion - 0.05 %
    previous sample - 0.25 %
    linear interpolation - 0.75 %
    STEIN - 2 %

   If MOS>3.75 (barely perceptible quality degradation) is acceptable,
   then the following packet drop probabilities are allowable:
      zero insertion - 0.1 %
      previous sample - 0.75 %
      linear interpolation - 3 %
      STEIN - 6.5 %

   If MOS>3.5 (cell-phone quality) is tolerable, then the following
   packet drop probabilities are allowable:
      zero insertion - 0.4 %
      previous sample - 2 %
      linear interpolation - 8 %
      STEIN - 14 %

Authors' Addresses

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

   Phone: +972 (0)3 645-5389
   Email: yaakov_s@rad.com


   David L. Black
   EMC Corporation
   176 South St.
   Hopkinton, MA  69719
   USA

   Phone: +1 (508) 293-7953
   Email: david.black@emc.com


   Bob Briscoe
   BT
   B54/77, Adastral Park
   Martlesham Heath
   Ipswich  IP5 3RE
   UK

   Phone: +44 1473 645196
   Email: bob.briscoe@bt.com
   URI:   http://bobbriscoe.net/