Transport Area Working Group
Internet-Draft
Intended status: Informational
Expires: June 15, 2013

M. Suznjevic
University of Zagreb
J. Saldana
University of Zaragoza

December 12, 2012

Maximum Tolerable Delays when using Tunneling Compressed Multiplexed Traffic Flows

draft-suznjevic-tsvwg-mtd-tcmtf-00

Abstract

This document contains recommendations of maximum tolerable delays to be added by methods which improve bandwidth utilization through compression, multiplexing, and tunneling over a network path. Recommendations are presented only for real-time network services for which such bandwidth optimization techniques are applicable (i.e., services with low payload size to header size ratio).

Table of Contents

1. Introduction 3
2. Requirements Language 3
3. Considered services 3
4. Delay recommendations 4
4.1. VoIP 7
4.2. Online games 7
4.3. Remote desktop access 8
5. Acknowledgements 9
6. IANA Considerations 9
7. Security Considerations 9
8. References 9
8.1. Normative References 9
8.2. Informative References 10
Authors' Addresses 11

1. Introduction

This document extends the draft [TCMTF] with a set of recommendations of overall tolerable delays which can be added in the processes of compression, multiplexing, and tunneling. These recommendations are needed, since the techniques proposed in [TCMTF], while saving bandwidth, add additional network delay. Network delay is one of the main factors which can degrade the Quality of Experience (QoE) of real-time network services [TGPP_TR26.944]. In order to prevent QoE degradation of real-time services using TCMTF, a policy defining a multiplexing period can be employed. Values of maximum tolerable delays presented here form the base of such policy. The recommendations are presented for real-time network services in which TCTMF bandwidth optimization is applicable (i.e., services which have low payload to header size ratio which results in high protocol overhead).

The second set of recommendations focuses on different multiplexing policies and implementation issues which are service and link specific.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. Considered services

Under the term "real-time network services" we consider both conversational and streaming service classes of services as defined in [TGPP_TS]. Interactive and background services are considered non real-time. Fundamental requirements of real-time network services include conversational pattern (stringent and low delay) and preservation of the time relation (variation) between the information entities of the stream.

We are focused on real-time network services which have low payload to header size ratio and therefore are appropriate for bandwith optimizations presented in TCMTF. We identify the following services:

- o Voice over IP

- o Online games

- o Remote desktop services

While video services are considered real-time, they are not suitable for bandwidth optimization techniques proposed in [TCMTF], due to the high payload to header size ratio they present. Therefore, we neither take into account services using an approach in which all the calculations are deployed in the server, which sends a real-time video stream to the client. In these cases, TCMTF optimization is neither interesting nor applicable. On the other hand, TCMTF can be applied for web browsing in terms of payload to header size ratio, but since some studies have shown that web browsing delays of several seconds are acceptable to users, there is no need for policy limitations in TCMTF, as the multiplexing periods are shorter than that [ITU-T_G.1010].

4. Delay recommendations

The three normally considered network impairments in the studies related to subjective quality in real-time interactive games are:

- o delay - can be reported as one-way-delay (OWD) [RFC2679] and two-way-delay (Round Trip Time) [RFC2681]. In this document, under the term latency, one way end-to-end day is considered.

- o jitter - which is a statistical variance of the data packet inter-arrival time, in other words the variation of the delay.

- o packet loss - more important for certain applications, while other include very good algorithms for concealing it (e.g., some game genres).

In this document we give recommendations of overal tollerable delays for previously listed real-time network services. In an interactive service, the total delay is composed by the addition of delays as defined in 3GPP TR 26.944 [TGPP_TR26.944].

- o Transfer delay - from Host1 to Host2 at time T is defined by the statement: Host1 sent the first bit of a unit data to Host2 at wire-time T and that Host2 received the last bit of that packet at wire-time T+dT

- o Transaction delay - the sum of the time for a data packet to wait in queue and receive the service during the server transaction.

Figure 1 shows these delays. The labeled times (S and R) designate the times in which the packet is sent or received by the network card interface.

```
   +-----------+                          +-----------+
   |   Host1   |                          |   Host 2  |
   +-----------+                          +-----------+
          S-------                                   |   ^       ^
          |       -------                            |   |       |
          |              -------                     |transf.    |
          |                     -------              |   |       |
          |                            -------       |   v       |
          |                                   ------>R   ^       |
          |                                          |   |       |
          |                                          |transac. total RTT
          |                                          |   |       |
          |                                   -------S   v       |
          |                            -------       |   ^       |
          |                     -------              |   |       |
          |              -------                     |transf.    |
          |       -------                            |   |       |
          R<------                                   |   v       v
          |                                          |
                       S: Packet sent
                       R: Packet received
```

Figure 1

The use of TCMTF requires the addition of a multiplexer and a demultiplexer in the scenario. A number of flows are multiplexed together before being sent through the Internet. The packets are demultiplexed and rebuilt before being forwarded to the application server. An scheme of TCMTF is included in Figure 2:

```
        +------+
        |user 1|___
        +------+   \
                    \
                                 _  _
        +------+    +-----+     ( ‘   )_        +-------+   +------+
        |user 2|--->| mux |--> (    )     ‘) --->| demux |-->|server|
        +------+    +-----+   (_   (_ .  _) _)  +-------+   +------+
           .        /                Internet
           .       /
        +------+  /     <-----------tcmtf----------->
        |user n|_/
        +------+
```

Figure 2

This technique groups packets in order to build a multiplexed one. So "multiplexing period" has to be defined in the multiplexer. When it ends, all the arrived packets are sent together in the same bundle. Therefore, multiplexing delay caused by tcmtf optimization techniques can be seen as an increase of transfer delay. The delay in the multiplexer is caused by the time the packets are retained until the bundled packet is sent, plus processing time. However, in the demultiplexer packets are not retained, so only processing time is considered.

Figure 3 shows the total delay, when a multiplexer and a demultiplexer are added. It should be noticed that multiplexing can be deployed independently in the two directions, or only in one of them, as shown in Figure 3.

```
   +---------+     +--------+      +--------+    +---------+
   | Host 1  |     |  mux   |      |  demux |    | Host 2  |
   +---------+     +--------+      +--------+    +---------+
       S-------         |               |             |   ^       ^
       |       -------  |               |             |   |       |
       |              --R     ^         |             |   |       |
       |                |     |         |             |   |       |
       |                |    mux        |             |   |       |
       |                |     |         |             |   |       |
       |                |     |         |             | transf.   |
       |                S---- v         |             | (& mux)   |
       |                |    -------    |             |   |       |
       |                            ----R    ^        |   |       |
       |                                |  demux      |   |    total RTT
       |                                |    |        |   |       |
       |                                S--- v        |   v       |
       |                                |   --------->R   ^       |
       |                                              |   |       |
       |                                              |transac.   |
       |                                              |   |       |
       |                                      --------S   v       |
       |                              --------        |   ^       |
       |                      --------                |   |       |
       |              --------                        | transf.   |
       |       -------                                |   |       |
       R<------                                       |   v       v
       |                                              |
                       S: Packet sent
                       R: Packet received
```

Figure 3

If a policy defining a multiplexing period is used, then the average latency added to each packet will be half the multiplexing period. A tradeoff appears: the longer the multiplexing period, the higher the number of packets which can be grouped, thus obtaining better bandwidth savings. So in order to calculate the maximum multiplexing period, the rest of the delays have to be considered: if the sum of propagation, processing and transmission delays is under the maximum tolerable delay, then multiplexing will be possible without harming user’s experience. The calculation of the overal delay may be performed according to the ITU-T Y.1541 reccomendation [ITU-T_Y.1541] The difference will give the maximum recommended multiplexing period.

Next, we will report the maximum tolerable latencies for the previously listed real-time network services.

4.1. VoIP

For conversational audio, the International Telecommunication Union recommends in [ITU-T_G.114] less than 150 millisecond one-way end-to-end delay for high-quality real time traffic, but delays between 150ms and 400ms are acceptable. For a streaming audio, delay constraints are much looser, the delay should be less than 10s [ITU-T_G.1010].

4.2. Online games

Online games are a large area comprising many game genres which have different latency requirements. This draft focuses on real-time online games and endorses the general game categorization proposed in [Claypool_Latency] in which online games have been divided into:

- o Omnipresent, with the threshold of acceptable latency (i.e., latency in which performance is above 75% of the unimpaired performance) of 1000 ms. The most representative genre of omnipresent games are Real-Time Strategies.

- o Third Person Avatar, with the threshold of acceptable latency of 500ms. These games include include Role Playing Games (RPG) and Massively Multiplayer Online Role-Playing Games (MMORPG).

- o First Person Avatar, in which threshold of acceptable latency is 100ms. The most popular subgenre of them are First Person Shooters, such as "Call of Duty" or "Halo" series.

The study [Claypool_Latency] evaluated players’ performance of certain tasks while increasing latency, and reported latencies in which the performance dropped below 75% of the performance under unimpaired network conditions. While measuring objective performance metrics, this method highly underestimates the impact of delays on players’ QoE. Further studies accessing a particular game genre reported much lower latency thresholds for unimpaired gameplay.

A survey using a large number of First Person Shooter games was carried out in [Dick_Analysis]. As a result, they stated that latencies about 80ms could be considered as acceptable, since the games were rated as "unimpaired". Besides service QoE, it has been shown that delay has great impact on the user’s decision to join a game, but significantly lesser on the decision to leave the game [Henderson_QoS].

A study oriented to evaluation of Mean Opinion Score (MOS), based on variation of delay and jitter for MMORPGs, suggested that MOS drops below 4 for delays greater than 120 ms [Ries_QoEMMORPG]. The MOS score of 5 indicates excellent quality, while MOS score of 1 indicates bad quality. Another study focused on extracting the duration of play sessions for MMORPGs from the network traffic traces showed that the session durations start to decline sharply when latency is between 150ms and 200ms latencies[Chen_HowSensitive].

While original classification work [Claypool_Latency] states that latencies up to 1s are tolerated by omnipresent games, other studies argued that only latencies up to 200ms are tolerated by players of RTS games [Cajada_RTS].

4.3. Remote desktop access

For the services of remote computer access, the delays are dependent on the task performed through the remote desktop, which are categorized into audio, video and data (reading, web browsing, document creation). A QoE study indicates that for audio latencies below 225 ms and for data latencies below 200 ms are tolerated [Dusi_Thin].

We group all the results in the Table 1 indicating the maximum allowed of latencies and proposed multiplexing periods. Proposed multiplexing periods are guidelines, since the exact values are dependant of existing the delay in the network. It should be noted that multiplexing periods of about 1 second can be considered as enough for non real time services (e.g., web browsing and streaming audio).

Service	Tolerable latency (OWD)	Mux. period
Voice communication	< 400ms	< 80ms
Omnipresent games	< 300ms	< 60ms
First person avatar games	< 100ms	< 20ms
Third person avatar games	< 200ms	< 40ms
Remote desktop	< 200ms	< 40ms

Table 1: Final recommendations

5. Acknowledgements

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

All drafts are required to have a security considerations section. See RFC 3552 [RFC3552] for a guide.

8. References

8.1. Normative References

[ITU-T_G.1010]
International Telecommunication Union-Telecommunication, "End-user multimedia QoS categories", SERIES G: TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS; Quality of service and performance , 2001.

[ITU-T_G.114]
ITU-T, "ITU-T Recommendation G.114 One-way transmission time", ITU G.114, 2003.

[ITU-T_Y.1541]
International Telecommunication Union-Telecommunication, "; Network performance objectives for IP-based services", SERIES Y: GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS AND NEXT-GENERATION NETWORKS; Internet protocol aspects - Quality of service and network performance , 2011.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

[RFC2681] Almes, G., Kalidindi, S., and M. Zekauskas, "A Round-trip Delay Metric for IPPM", RFC 2681, September 1999.

8.2. Informative References

[Cajada_RTS]
Cajada, M., "VFC-RTS: Vector-Field Consistency para Real-Time-Strategy Multiplayer Games", Master of Science Disertation , 2012.

[Chen_HowSensitive]
Chen, K., Huang, P., and L. Chin-Luang, "How sensitive are online gamers to network quality?", Communications of the ACM 49, 2006.

[Claypool_Latency]
Claypool, M. and K. Claypool, "Latency and player actions in online games", Communications of the ACM 49, 2006.

[Dick_Analysis]
Dick, M., Wellnitz, O., and L. Wolf, "Analysis of factors affecting players' performance and perception in multiplayer games", Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games, pp. 1 - 7 , 2005.

[Dusi_Thin]
Dusi, M., Napolitano, S., Niccolini, S., and S. Longo, "A Closer Look at Thin-Client Connections: Statistical Application Identification for QoE Detection", IEEE Communications Magazine, pp. 195 - 202 , 2012.

[Henderson_QoS]
Henderson, T. and S. Bhatti, "Networked games: a QoS-sensitive application for QoS-insensitive users?", Proceedings of the ACM SIGCOMM workshop on Revisiting IP QoS: What have we learned, why do we care?, pp. 141-147 , 2003.

[RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, July 2003.

[Ries_QoEMMORPG]
Ries, M., Svoboda, P., and M. Rupp, "Empirical Study of Subjective Quality for Massive Multiplayer Games", Proceedings of the 15th International Conference on Systems, Signals and Image Processing, pp.181 - 184 , 2008.

[TCMTF] Saldana, J., Wing, D., Fernandez Navajas, J., Perumal, M., and F. Pascual Blanco, "Tunneling Compressed Multiplexed Traffic Flows (TCMTF)", Internet-Draft Jul, 2012.

[TGPP_TR26.944]
3rd Generation Partnership Project;, "Technical Specification Group Services and System Aspects; End-to-end multimedia services performance metrics", 3GPP TR 26.944 version 9.0.0 , 2012.

[TGPP_TS] 3rd Generation Partnership Project, European Telecommunications Standards Institute, "Quality of Service (QoS) concept and architecture", 3GPP TS 23.107 version 11.0.0 Release 11 , 2012.

Authors' Addresses

Mirko Suznjevic
University of Zagreb
Faculty of Electrical Engineering and Computing, Unska 3
Zagreb, 10000
Croatia

Phone: +385 1 6129 755
Email: mirko.suznjevic@fer.hr

Jose Saldana
University of Zaragoza
Dpt. IEC Ada Byron Building
Zaragoza, 50018
Spain

Phone: +34 976 762 698
Email: jsaldana@unizar.es