Transport Area Working Group
Internet-Draft
Intended status: Informational
Expires: December 15, 2016

M. Suznjevic
University of Zagreb
J. Saldana
University of Zaragoza
June 13, 2016

Delay Limits for Real-Time Services
draft-suznjevic-tsvwg-delay-limits-00

Abstract

Network delay is one of the main factors which can degrade the Quality of Experience (QoE) of the users of network services. This document surveys a set of recommendations about the maximum latency tolerated by the users of delay-constrained services. Some recommendations already exist for VoIP, but emerging services as e.g. online gaming, have different requirements. Different papers in the literature reporting these constraints are surveyed, and a summary of the latency limits for each service is finally provided.

Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on December 15, 2016.

Copyright Notice

carefully, as they describe your rights and restrictions with respect to this document.

Table of Contents

1. Introduction 2
1.1. Requirements Language 3
2. Considered services 3
2.1. Real-time services 3
2.2. Non real-time services 3
3. Definitions 3
4. Delay recommendations 5
4.1. VoIP 5
4.2. Online games 5
4.3. Remote desktop access 7
4.4. Non real-time service 7
4.5. Summary 7
5. Acknowledgements 8
6. IANA Considerations 8
7. Security Considerations 8
8. References 8
8.1. Normative References 8
8.2. Informative References 9
Authors' Addresses 11

1. Introduction

The "Workshop on Reducing Internet Latency" [Workshop], sponsored by the Internet Society and some research projects in 2013, discussed different ways for reducing Internet latency, stating that "For Internet applications, reducing the latency impact of sharing the communications medium with other users and applications is key."

Network delay is one of the main factors which can degrade the Quality of Experience (QoE) of network services [RFC6390] [TGPP_TR26.944]. In order to prevent the degradation of the perceived quality of the services with delay constraints, a maximum limit can be defined. This "latency budget" has to be taken into account when considering the possibility of adding new network functions (e.g. through middleboxes), since every optimization adds some delay as a counterpart. These new functions not only exist at upper layers, but they can also be found in Layer 2. For example, in [IEEE.802-11N.2009], a number of Protocol Data Units can be grouped and transmitted together, but this will add a new delay required to gather a number of frames together.

This document surveys a set of recommendations about the maximum latency tolerated by the users of services with delay constraints.

Some recommendations already exist for e.g. VoIP [ITU-T_G.114], but emerging services as e.g. online gaming, have different requirements, which may also vary with the game genre.

1.1. 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].

2. Considered services

2.1. Real-time services

Under the term "real-time network services" we consider both conversational and streaming service classes 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 identify the following real-time network services, as those with the most stringent real-time constraints:

- Voice over IP
- Online games
- Remote desktop services

2.2. Non real-time services

Non real-time services such as streaming audio or video, and instant messaging also have delay limits, but different studies have shown that acceptable delays for these services are up to several seconds [ITU-T_G.1010].

Some types of machine to machine (M2M) traffic (e.g., metering messages from various sensors) for these services can be go up to an hour [Liu_M2M].

3. Definitions

The three network impairments normally considered in the studies related to subjective quality in delay-constrained services are:

- 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 delay is considered.

- delay variation - which is a statistical variance of the data packet inter-arrival time, in other words the variation of the delay as defined in [RFC3393].

- packet loss - more important for certain services, while other include very good algorithms for concealing it (e.g. some game genres receive accumulative updates, so packet loss is not important).

In this document we give recommendations for overall tolerable delays to be taken into account when adding new middleboxes or functionalities in the network. In an interactive service, the total delay is composed by the addition of delays as defined in 3GPP TR 26.944 [TGPP_TR26.944]. The overall delay may be calculated according to the ITU-T Y.1541 recommendation [ITU-T_Y.1541].

- 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." Thus, it includes the transmission delay (the amount of time Host1 requires to push all of the packet’s bits into the wire) and the propagation delay in the network (the amount of time it takes for the head of the packet to travel from Host1 to Host2).

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

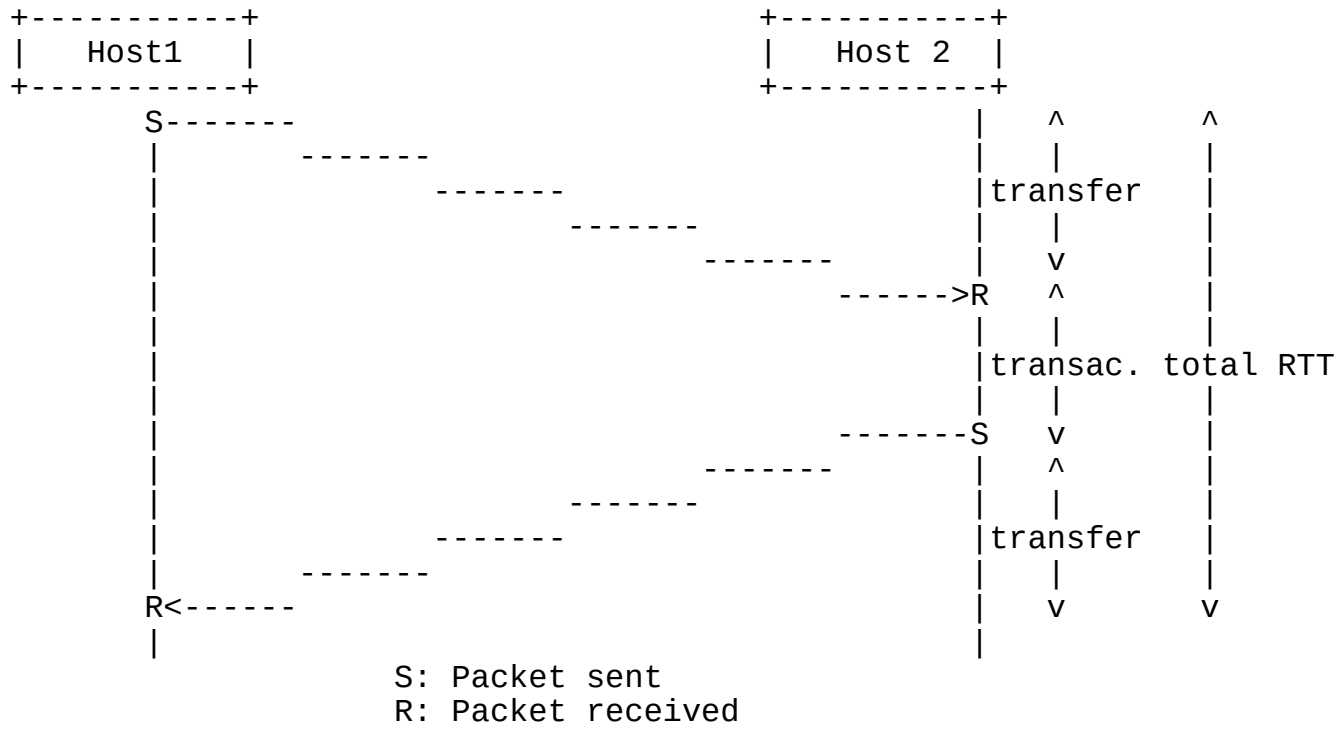

Figure 1

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

4. Delay recommendations

4.1. VoIP

For conversational audio, the International Telecommunication Union recommends [ITU-T_G.114] less than 150 millisecond one-way end-to-end delay for high-quality real time traffic, but delays between 150 ms and 400 ms are still acceptable. When considering conversational audio, it should be noted that this delay limits include jitter buffers and codec processing. For streaming audio, delay constraints are much looser, so the delay should be less than 10 s [ITU-T_G.1010].

4.2. Online games

Online games comprise game genres which have different latency requirements. This document focuses on real-time online games and endorses the general game categorization proposed in [Claypool_Latency] in which online games have been divided into:

- 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.

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

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

As remarked in [Bernier_Latency] and [Oliveira_online], different methods can be employed to combat delay in online games. The so-called "client-side prediction" has been largely used in First Person Shooters. It can be divided into "input prediction" and "dead reckoning," where input prediction hides the latency for the client-controlled actions while dead reckoning hides the latency of other participating players.

The study [Claypool_Latency] evaluated players' performance in certain tasks, while increasing latency, and reported values at 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.

Other approach some studies have taken is to perform "objective measurements" [Kaiser_objective] a number of identical "bots", i.e. virtual avatars controlled by Artificial Intelligence, are placed in the same virtual scenario and a number of parties between them are performed. If the number of parties if high enough, then the score will be the same for all the bots. Then, different network impairments (latency, jitter, packet loss) are added to one of the bots, and another set of tests is performed. The performance degradation of the network-impaired bot can then be statistically characterized.

A survey using a large number of First Person Shooter games has been carried out in [Dick_Analysis]. They state that latency about 80 ms could be considered as acceptable, since the games have been 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 less on the decision to leave the game [Henderson_QoS].

A study on Mean Opinion Score (MOS) evaluation, 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 round trip time is between 150 ms and 200 ms [Chen_HowSensitive].

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

4.3. Remote desktop access

For the remote computer access services, the delays are dependent on the task performed through the remote desktop. Tasks may include operations with audio, video and data (e.g., reading, web browsing, document creation). A QoE study indicates that for audio latency below 225 ms and for data latency below 200 ms is tolerated [Dusi_Thin].

4.4. Non real-time service

Under this category we include services for M2M metering information, streaming audio, and instant messaging. M2M metering services present a one way communication (i.e., most information travels from sensors to the central server) [Liu_M2M]. The signalling information related to M2M can also be optimized. Internet of Things application layer protocols such as CoAP [RFC7252], used in Constrained RESTful Environments (CoRE)[RFC6690]. The ACK_TIMEOUT period in CoAP is set to 2 seconds. Instant messaging (despite "instant" in its name) has been categorized as data service by the ITU-T, and it has been designated with acceptable delays of up to a few seconds [ITU-T_G.1010].

4.5. Summary

We group all the results in Table 1 indicating the maximum allowed latency and proposed multiplexing periods. Proposed multiplexing periods are guidelines, since the exact values are dependant of the existing delay in the network. It should be noted that reported tolerable latency is based on values of preferred delays, and delays in which QoE estimation is not significantly degraded. Multiplexing periods of about 1 second can be considered as sufficient for non real-time services (e.g., streaming audio).

Service	Tolerable latency (OWD)	Mux. period
Voice communication	< 150ms	< 30ms
Omnipresent games	< 200ms	< 40ms
First person avatar games	< 80ms	< 15ms
Third person avatar games	< 120ms	< 25ms
Remote desktop	< 200ms	< 40ms
Instant messaging	< 5s	< 1s
M2M (metering)	< 1hour	< 1s

Table 1: Final recommendations

5. Acknowledgements

Jose Saldana was funded by the EU H2020 Wi-5 project (Grant Agreement no: 644262).

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

No relevant security considerations have been identified

8. References

8.1. Normative References

[IEEE.802-11N.2009]
"Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications - Amendment 5: Enhancements for higher throughput", IEEE Standard 802.11n, Oct 2009, <http://standards.ieee.org/getieee802/download/802.11n-2009.pdf>.

[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, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[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.

[RFC3393] Demichelis, C., Chimento, S., and P. Zekauskas, "IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)", RFC 3393, November 2002.

[RFC6390] Clark, A. and B. Claise, "Guidelines for Considering New Performance Metric Development", RFC 6390, October 2011.

[RFC6690] Shelby, Z., "Constrained RESTful Environments (CoRE) Link Format", RFC 6690, August 2012.

[RFC7252] Shelby, Z., Hartke, K., and C. Bormann, "The Constrained Application Protocol (CoAP)", RFC 7252, DOI 10.17487/RFC7252, June 2014, <http://www.rfc-editor.org/info/rfc7252>.

8.2. Informative References

[Bernier_Latency]
Bernier, Y., "Latency Compensating Methods in Client/ Server In-Game Protocol Design and Optimization", Proc. Game Developers Conference, San Jose Vol. 98033. No. 425., 2001.

[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.

[Kaiser_objective]
Kaiser, A., Maggiorini, D., Boussetta , K., and N. Achir, "On the Objective Evaluation of Real-Time Networked Games", Proc. IEEE Global Telecommunications Conference (GLOBECOM 2009) , 2009.

[Liu_M2M] Liu, R., Wu, W., Zao, H., and D. Yang, "M2M-Oriented QoS Categorization in Cellular Network", Master of Science Disertation , 2012.

[Oliveira_online]
Oliveira, M. and T. Henderson, "What online gamers really think of the Internet?", Proceedings of the 2nd workshop on Network and system support for games (NetGames '03). ACM, New York, NY, USA pp. 185-193, 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.

[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.

[Workshop]
Ford, M., "Workshop report: reducing internet latency", SIGCOMM Comput. Commun. Rev. 44, 2 (April 2014), 80-86. , 2013.

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