

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: July 20, 2014

R.Sallantin
CNES/TAS/TESA
C.Baudoin
F.Arnal
Thales Alenia Space
E.Dubois
CNES
E.Chaput
A.Beylot
IRIT
January 16, 2014

Safe increase of the TCP's Initial Window
Using Initial Spreading
draft-sallantin-iccr-initial-spreading-00

Abstract

This document proposes a new fast start-up mechanism for TCP that can be used to speed the beginning of an Internet connection and then improved the short-lived TCP connections performance.

Initial Spreading allows to safely increase the Initial Window size in any cases, and notably in congested networks.

Merging the increase in the IW with the spacing of the segments belonging to the Initial Window (IW), Initial Spreading is a very simple mechanism that improves short-lived TCP flows performance and do not deteriorate long-lived TCP flows performance.

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), 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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
2	Terminology	4
3	Initial Spreading mechanism	4
4	Implementation considerations	4
4.1	Short Round Trip Time	5
4.2	Delayed Ack	5
4.3	TSO/GSO	5
4.4	RTT measure	6
5	Open discussions	6
5.1	Spacing Interval	6
5.2	Increasing the upper bound TCP's IW to more than 10 segments	7
5.3	Initial Spreading and LFN	7
6	Security Considerations	7
7	IANA Considerations	8
8	References	8
8.1	Normative References	8
8.2	Informative References	8
	Authors' Addresses	9

1 Introduction

The long Round Trip Time is probably the most detrimental constraint in Long Fat Networks (LFN), such as satellite networks, and notably for short-lived connections when the long delay significantly downgrades regular slow-start performance [FA11]. Several protocols and even new network architectures have been proposed to deal with this issue. The original idea of Initial Spreading [SB13] was to consider a long RTT as a resource to exploit, rather than as a constant to bypass. The long RTT can therefore be used as an opportunity to safely send a large amount of data during the first RTT after the connection has opened. Spacing the data along the whole RTT would in fact hopefully guarantee high independent probability that each segment is successfully received.

This approach resembles a combination of 2 TCP mechanisms: Pacing and Increase in the Initial Window. Both mechanisms have then been studied in depth to design Initial Spreading as an efficient fast start-up TCP mechanism, and notably avoid their respective flaws or weaknesses.

The original Pacing idea is to space the segments of a same window along an RTT to prevent generating bursts as far as possible. Hence, each segment arrives separately at the buffer and the impact on its queue is minimized. The bit rate can then reach its maximum. However, [AS00] has pointed out that this lack of bursts is responsible for poor performance. Pacing has a tendency to overload the network, and then cause a synchronization of the flows, that seriously damages both individual and global performance.

RFC 6928 [RFC6928] suggests to enlarge the IW size up to ten segments. Several articles and studies demonstrated that this would allow transmission of 90% of the connections in one RTT [DR10]. In most cases, and when the network is not congested in particular, this solution is probably the best one for dealing with short-lived TCP flows. However, in a congested environment, sending a large IW in one burst is likely to impact the buffers and then deteriorate the individual connection. Correlation between the segments of a same burst is responsible for major impairments when regarding the short-lived connections, and in particular for the connections that can be sent in one RTT (number of segments to be transmitted inferior to the upper bound value of the TCP's IW):

- o a decrease of the probability to successfully transmit the entire window.
- o an increase of the probability of successive segment losses.

o a significant reduction of the number of potential Duplicated Acknowledgements that are necessary to trigger fast loss recovery mechanisms and avoid to wait for an Retransmission Time Out.

In favor of a conservative approach, [RFC3390] recommended the use of an IW equal to 3.

Both mechanisms therefore suffer from a burst-related phenomenon, but in opposite ways.

Initial Spreading has been designed to tackle previous burst issues. Simulations and experimentations show that Initial Spreading is not only efficient in case of LFNS but also for other networks with small RTT.

2 Terminology

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 Initial Spreading mechanism

Initial Spreading [SB13] mechanism uses the permitted upper bound value of the TCP's IW (e.g; RFC 6928 [RFC6928] suggests to use 10 for this value). Initial Spreading spaces out a number of segments inferior or equal to this value across the first RTT before letting the TCP algorithm continue conventionally:

- (1) The RTT is measured during the SYN-SYN/ACK exchange.
- (2) The first RTT is split into n spaces with n the permitted upper bound value of the TCP's IW. Depending on the number of segments to be sent, until n segments are sent every RTT/n .
- (3) After the transmission of the IW, the regular TCP algorithm is used.

Thus, bursts do not downgrade the transmission of short-lived connections, but continue to prevent an overload of the network in the case of long-lived connections.

4 Implementation considerations

In this section, we discuss a number of aspects surrounding the

Initial Spreading implementations.

4.1 Short Round Trip Time

Whatever the different timer implementations, 2 segments can not be spaced of less than 1 Kernel timer (e.g: jiffy for Linux kernel).

In case where the time resulting of the division of 1 RTT by the upper bound value of the TCP's IW is inferior to 1 Kernel timer, Initial Spreading is not activated and TCP uses a regular slow start with a large IW.

4.2 Delayed Ack

The use of Delayed Ack (Del Ack) does not downgrade Initial Spreading efficiency.

Regarding long-lived connections and notably TCP's steady state, the effects of Del Ack are lessened by new TCP's flavors (such as TCP Cubic or Compound TCP [HR08][TS06]) which tend to adapt their congestion algorithm to take into account whether the receiver uses the Del Ack option or not. In doing so, they can prevent the connection from being too slow, and still continue to reduce acknowledgments traffic. In the event of short-lived connections, the use of Del Ack does not modify the transmission of the IW. There is then no change in the burst propagation.

4.3 TSO/GSO

TSO/GSO is used to reduce the CPU overhead of TCP/IP on fast networks. Instead of doing the segmentation in the kernel, large packets are sent to the Network Interface Card (NIC). The segmentation is then achieved by the NIC or just before the entry into the driver's xmit routine.

In its current design, Initial Spreading is not working when TSO or GSO are activated, but using Initial Spreading with an inactive TSO/GSO still enables better performance.

Two options can be foreseen for the joint use of Initial Spreading and TSO/GSO:

- (1) disable TSO/GSO for the first RTT, with no impact on performance since the throughput is limited by the IW.
- (2) implement Initial Spreading using the TCP Offload Engine (TOE) [RFC5522].

4.4 RTT measure

Initial spreading uses the SYN-SYN/ACK exchange to calculate the space between two segments. This measurement may not be perfectly accurate in congested networks when the RTT varies. Two different scenarios can then occur:

- o The measured RTT is superior to the RTT of the first segment of the IW, and an ACK arrives before that all the segments of the IW have been sent. Then, Initial Spreading **MUST** be stopped and only the segments transmission that is triggered by the received ACK is done.
- o The measured RTT is inferior to the RTT of the first segment of the IW. Consequences are negligible.

5 Open discussions

In this section, we introduce possible improvements for Initial Spreading and new perspectives.

5.1 Spacing Interval

It has been observed that most of the savings enabled by the Initial Spreading in congested environments comes from the independence of the segments sent during the first RTT. Indeed, experimentations have shown that preventing the bursts, Initial Spreading enables each segment of the IW to have an independent loss probability.

Currently, Initial Spreading waits RTT/n seconds before transmitting two segments of the IW, with n the permitted upper bound value of the TCP's IW.

This simple mechanism offers very good results but has two minor drawbacks:

- (1) An inaccuracy of the space measure (cf section 4.4).
- (2) In uncongested networks, Initial Spreading adds an extra delay that is equal to $(IW-1) * RTT/n$, with IW the number of segments ($\leq n$) that can be sent during the first RTT.

A solution could be to set the space not as a ratio of the measured RTT but as a minimal space that preserves the independence between the sent segments. Preliminary results show that smaller spacing interval may allow to maintain the independence of the segment loss probability. This may provide the same performances in congested

networks and improve the average delay in uncongested networks.

5.2 Increasing the upper bound TCP's IW to more than 10 segments

[DR10] have shown that an IW of 10 segments enables to send more than 90% of the web objects in one RTT. So the authors recommend to use Initial Spreading as a complement to [RFC6928].

If the average size of the web objects continues to evolve, Initial Spreading can be used to raise the IW size. Simulations and experiments showed even better results with an IW equal to 12.

Thus, Initial Spreading paves the way for larger IW. Further studies are needed to assess the impact on the networks, notably in terms of individual performance, fairness, friendliness and global performance.

5.3 Initial Spreading and LFN

The space community designed middleboxes to mitigate poor TCP performance for network with large RTT [FA11]. Proxy Enhancement Performance (PEP) are generally used in LFN and in particular in satellite communication systems [RFC3135] and offer very good TCP performance.

Nevertheless, some recent studies have emphasized major impairments occasioned by the use of satellite-specific transport solutions, and notably TCP-PEPs, in a global context. The break of the end-to-end TCP semantic, which is required to isolate the satellite segment, is notably responsible for an increased complexity in case of mobility scenarios or security context. This strongly mitigates PEPs benefits and reopens the debate on their relevance[DC10].

Many researchers have outlined that new TCP releases perform well for long-lived TCP connections, even in satellite environment [SC12], but continue to suffer from very poor performance in case of short-lived TCP connections.

Initial Spreading enables to reduce the RTT consequences for short-lived TCP connections and could be an end-to-end alternative to PEP.

6 Security Considerations

The security considerations found in [RFC5681] apply to this document. No additional security problems have been identified with Initial Spreading at this time.

7 IANA Considerations

This document contains no IANA considerations.

8 References

8.1 Normative References

- [RFC3390] A. Allman and S. Floyd, "Increasing tcp's initial window," RFC 3390, IETF, Proposed Standard, 2002.
- [RFC5532] T. Talpey, C. Juszczak, "Network File System (NFS) Remote Direct Memory Access (RDMA) Problem Statement," RFC 5532, IETF, Informational, May 2009.
- [RFC6928] J. Chu, N. Dukkipati, Y. Cheng, and M. Mathis, "Increasing tcp's initial window," RFC 6928, IETF, Experimental, Jan. 2013.
- [AH98] A. Allman, C. Hayes, and S. Ostermann, "An evaluation of TCP with Larger Initial Windows," ACM Computer Communication Review, 1998.
- [AS00] A. Aggarwal, S. Savage, and T. Anderson, "Understanding the performance of TCP pacing," in INFOCOM, vol. 3, mar 2000, pp. 1157-1165.
- [DR10] N. Dukkipati, T. Refice, Y. Cheng, J. Chu, T. Herbert, A. Agarwal, A. Jain, and N. Sutin, "An Argument for Increasing TCP's Initial Congestion Window," SIGCOMM Comput. Commun. Rev., vol. 40, no. 3, pp. 26-33, Jun. 2010.
- [SB13] R. Sallantin, C. Baudoin, E. Chaput, E. Dubois, F. Arnal, and A. Beylot, "Initial spreading: a fast start-up tcp mechanism," proceedings of LCN, 2013.

8.2 Informative References

- [RFC3135] J. Border, M. Kojo, J. Griner, G. Montenegro, Z. Shelby, "Performance Enhancing Proxies Intended to Mitigate Link-Related Degradations," RFC 3135, IETF, Informational, June 2001.
- [DF10] E. Dubois, J. Fasson, C. Donny, and E. Chaput, "Enhancing tcp based communications in mobile satellite scenarios: Tcp peps issues and solutions," in Proc. 5th Advanced

satellite multimedia systems conference (asma) and the 11th signal processing for space communications workshop (spsc), pages 476-483, 2010.

- [FA11] A. Fairhurst, G. Arjuna, H. Cruickshank, and C. Baudoin, "Transport challenges facing a next generation hybrid satellite internet," in International Journal of Satellite Communications and networking, 2011.
- [HR08] S. Ha, I. Rhee, and L. Xu, "CUBIC: A New TCP-Friendly High-Speed TCP Variant," SIGOPS Oper. Syst. Rev., vol. 42, no. 5, pp. 64-74, Jul. 2008.
- [LC09] R. Lacamera, D. Caini, C. Furrincieli, "Comparative performance evaluation of tcp variants on satellite environments," in ICC'09 Proceedings of the 2009 IEEE international conference on Communications, pages Pages 5161-5165, 2009.
- [SC12] R. Sallantin, E. Chaput, E. P. Dubois, C. Baudoin, F. Arnal, and A.-L. Beylot, "On the sustainability of PEPs for satellite Internet access," in ICSSC. AIAA, 2012.
- [TS06] K. Tan, J. Song, Q. Zhang, and M. Sridharan, "Compound TCP: A Scalable and TCP-friendly Congestion Control for High-speed Networks," in 4th International workshop on Protocols for Fast Long-Distance Networks (PFLDNet), 2006.

Authors' Addresses

Comments are solicited and should be addressed to the working group's mailing list at iccr@irtf.org and/or the authors:

Renaud Sallantin
CNES/TAS/TESA
IRIT/ENSEEIH 2, rue Charles Camichel BP 7122
31071 Toulouse Cedex 7
France
Phone: +33 6 48 07 86 44
Email: renaud.sallantin@gmail.com

Cedric Baudoin
Thales Alenia Space (TAS)
26 Avenue Jean Francois Champollion,
31100 Toulouse

France
Email: cedric.baudoin@thalesaleniaspace.com

Fabrice Arnal
Thales Alenia Space
Email: fabrice.arnal@thalesaleniaspace.com

Emmanuel Dubois
Centre National des Etudes Spatiales (CNES)
18 Avenue Edouard Belin
31400 Toulouse
France
Email: emmanuel.Dubois@cnes.Fr

Emmanuel Chaput
IRIT
IRIT / ENSEEIHT 2, rue Charles Camichel BP 7122
31071 Toulouse Cedex 7
France
Email: emmanuel.chaput@enseeiht.fr

Andre-Luc Beylot
IRIT
Email: andre-Luc.Beylot@enseeiht.fr