

BANANA
Internet Draft
Intended Category: Informational

N. Leymann
C. Heidemann
Deutsche Telekom AG
L. Geng
China Mobile
J. Shen
China Telecom Co., Ltd
M. Zhang
L. Chen
Huawei
M. Cullen
Painless Security
February 6, 2018

Expires: August 10, 2018

BANdwidth Aggregation for interNet Access (BANANA)
Load Rebalance for Bonding Tunnels
draft-leymann-banana-load-rebalance-02.txt

Abstract

BANdwidth Aggregation for interNet Access (BANANA) makes use of a subscriber's multiple points of attachment to the Internet to provide the subscriber with higher bandwidth and reliability than what is provided by any single one of these attachments.

Various tunnel based methods have been developed to realize BANANA. This document specifies a throughput-increasing mechanism that can be commonly adopted by bonding tunnels methods. Basically, ingress node adaptively adjusts its load distribution function according to the quality of the bonding tunnels so as to make best use of the bonding bandwidth.

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

INTERNET-DRAFT Load Rebalancing for Bonding Tunnels February 6, 2018

<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) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](http://trustee.ietf.org/license-info) 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	2
2. Acronyms and Terminology	3
3. Problem: Bonding Reordering Buffer Bloating	3
4. Related Work	5
5. Load Rebalance	5
5.1. Adaptive Splitting Ratio	6
5.2. Adaptive Sequence Alignment	6
6. Protocol Extensions	7
7. Security Considerations	7
8. IANA Considerations	7
9. References	7
9.1. Normative References	7
9.2. Informative References	8
Author's Addresses	9

[1. Introduction](#)

BANdwidth Aggregation for interNet Access (BANANA) enables subscribers to make use of multiple access technologies to achieve

reliable and high bandwidth Internet access. Various bonding tunnel technologies have been proposed to realize BANANA [[GREbond](#)] [[GTPbond](#)] [[MIPbond](#)]. Since per packet traffic distribution is adopted by bonding tunnels, latency difference of the two tunnels may cause packet disorder to a single traffic flow that is being split across

INTERNET-DRAFT Load Rebalancing for Bonding Tunnels February 6, 2018

these two tunnels. Therefore, a reordering buffer for the bonding tunnels is used at the egress node to restore packet disorder. It is referred as "bonding reordering buffer" afterwards in this document.

The egress node places a limit (see `OUTOFORDER_TIMER` in [[RFC2890](#)]) on the time that a packet can wait in the bonding reordering buffer and places a limit on the number of packets in the bonding reordering buffer (`MAX_REORDER_BUFFER`, see `MAX_PERFLOW_BUFFER` in [[RFC2890](#)]). Any packet that would cause violation of either of the two limits MUST be forcibly delivered by the egress node. The bonding reordering buffer bloating issue may break these two limits, which lead to the mandatory packet delivery therefore causes mass loss of TCP packets. The throughput of the bonding tunnels may decrease dramatically. It is always important to minimize the usage of the bonding reordering buffer (or "Bonding Reordering Buffer Size") in order to reduce the possibility of breaking the above two limits.

BANANA may measure the Round Trip Time (RTT) and data rate of each tunnel and monitor the usage of the bonding reordering buffer. Based on the measurement, the ingress node may dynamically adjust the traffic distribution function in order to achieve a higher throughput of the bonding tunnels. For example, it may adaptively update the splitting ratio or adaptively arrange the packet sequence into the bonding tunnels.

[2.](#) Acronyms and Terminology

CIR: Committed Information Rate [[RFC2697](#)]

RTT: Round Trip Time

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.](#) Problem: Bonding Reordering Buffer Bloating

Latency difference of the two tunnels causes packet disorder to a traffic flow that is split across these two tunnels. The bonding reordering buffer based on the bonding sequence number at the egress is used to "absorb" this latency difference. Figure 3.1 illustrates the operation of the reordering.

INTERNET-DRAFT Load Rebalancing for Bonding Tunnels February 6, 2018

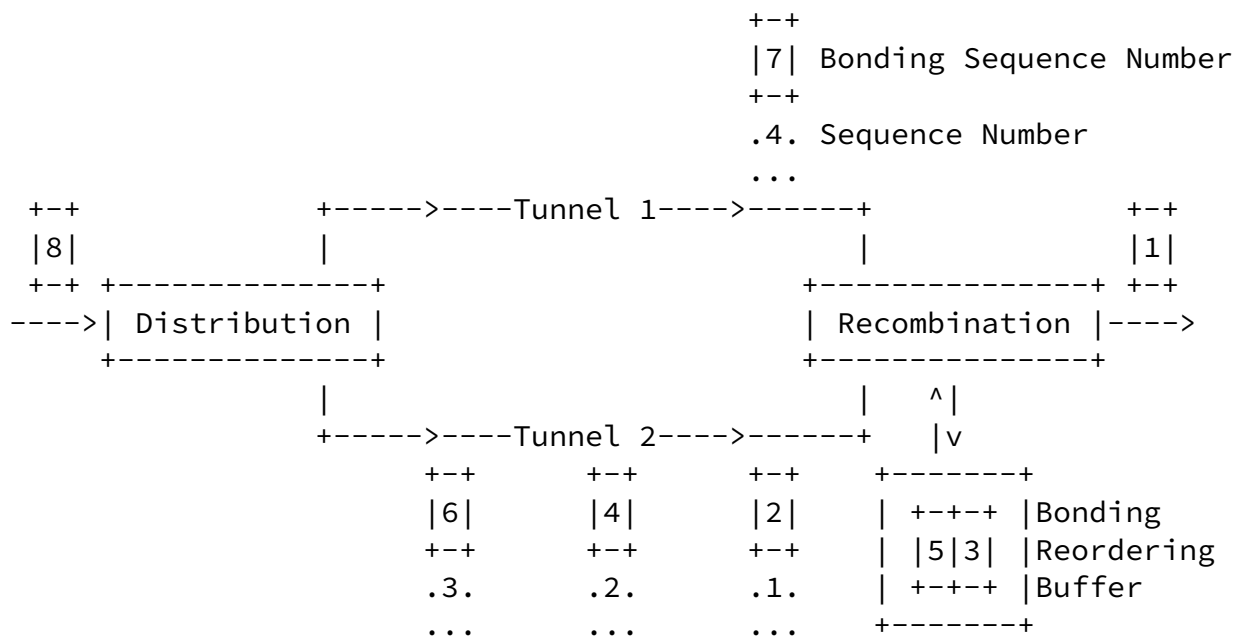


Figure 3.1: Bonding Tunnel Reordering Operation

[RFC2890] places two limits on the reordering buffer of a tunnel. One is the timer limit: `OUTOFORDER_TIMER` and the other is the size limit: `MAX_PERFLOW_BUFFER`. For bonding tunnels, the first limit is reused while the second parameter becomes the maximum bonding reordering buffer size of the entire bonding tunnel rather than a specific flow.

[RFC5681] defines Flight Size as the amount of data that has been sent but not yet cumulatively acknowledged. In this document, the Flight Size of a tunnel indicates the amount of data that has been

sent by the ingress node not to this tunnel but not yet pass through the reordering buffer (which is not shown in the figure) of this tunnel. The Flight Size of the entire bonding tunnel indicates the amount of data that has been sent by the ingress node by either tunnel but not yet pass through the bonding reordering buffer. From the sequence number of the last packet sent by the ingress node and the latest sequence number acknowledged by the egress node, the ingress node can monitor the Flight Size of a tunnel. For the entire bonding tunnel, the egress node might acknowledge the bonding sequence number via either of the two tunnels. The maximum bonding sequence number acknowledged by both tunnels is the latest acknowledged bonding sequence number.

As shown in Figure 3.1, the Flight Sizes of the tunnels can be used to estimate the load of the tunnels and the usage of the bonding reordering buffer. Suppose the Flight Size of tunnel_1 is F_1 , the Flight Size of tunnel_2 is F_2 , the Flight Size of the entire bonding tunnel is F_B while the Bonding Reordering Buffer Size is B . B can be calculated as

$$\begin{aligned} B &= F_B - F_1 - F_2 \\ &= 6 - 1 - 3 \\ &= 2 \end{aligned}$$

The bonding reordering buffer may bloat due to the large delay difference of the two tunnels. This bonding reordering buffer bloating issue might lead to the violation of the timer and/or the buffer size limit. The egress node has to deliver the violating packets, which will cause mass packet loss and retransmission of the carried TCP traffic. Throughput of the bonded tunnels will drop dramatically. Therefore, it is always important to minimize the size of the bonding reordering buffer.

[4.](#) Related Work

Several TCP congestion-avoidance algorithms are implemented for congestion control in the Internet. TCP New Reno, defined by [\[RFC6582\]](#), improves retransmission during the fast-recovery phase. In the absence of SACK [\[RFC2018\]](#), TCP New Reno responds to partial acknowledgments (ACKs that cover new data, but not all the data outstanding when loss was detected) and sends the next packet beyond the ACKed sequence number. The TCP [BIC] uses binary search to

iteratively find the proper congestion window size in each time interval of RTT. [[CUBIC](#)] is a less aggressive and more systematic derivative of BIC, in which the window is a cubic function of time so that RTT fairness is guaranteed.

However, traditional TCP congestion-avoidance algorithms are not applicable to bonding tunnels due to the following reasons. Bonding tunnels adopt per packet other than per flow load balancing. Bonding tunnels are established between a pair of network devices rather than host-to-host. The ingress node of bonding tunnels is not capable to alter the traffic sending rate. It does not keep sending buffers so it is not capable to retransmit lost packets either.

Explicit Congestion Notification (ECN [[RFC3168](#)]) notifies impending network congestion by setting a mark in the IP header instead of dropping packets. When the receiver echoes the congestion indication to the sender, the sender should reduce its transmission rate accordingly. The ECN mechanism could be applicable to tunnelling scenarios, but the mechanism itself must be specifically designed [[RFC6040](#)].

[5](#). Load Rebalance

Parameters such as the Round-Trip Time and the packet loss rate of each tunnel, the usage of the bonding reordering buffer and the data rate of the tunnels might be measured. The measurement could be done

in either an one-way or two-way manner. The ECN is a special case of such measurement. If the underlying network infrastructure of the bonding tunnels support ECN, the congestion indications of ECN could be used as measured information as well. The measured information might be carried either by data packets or control messages.

Based on the measured information, the ingress node can judge whether one tunnel is already congested so that the traffic proportion to be loaded on it should be decreased. The ingress node therefore can timely adjust the traffic distribution function to realize a "load rebalance". This load rebalance helps the BANANA system to make best use of the bandwidth of the two tunnels, and to reduce the queue length in the bonding reordering buffer before the congestion control of user's TCP traffic react.

[5.1.](#) Adaptive Splitting Ratio

Coloring mechanism is used to achieve per-packet traffic distribution across bonded tunnels [[GREbond](#)] [[GTPbond](#)]. Coloring mechanism is defined by [[RFC2697](#)] and [[RFC2698](#)]. The Committed Information Rate (CIR) determines the traffic rate distributed into a give tunnel. The CIR of the primary tunnel is fixed while the CIR of the secondary tunnel can be tuned dynamically. The ingress node may monitor the latency of the two tunnels via the measurement of RTT. If the latency difference of the two tunnels exceeds a pre-configured threshold (a value in the range from 0 to 100ms), the CIR for the secondary tunnel is decreased (e.g., by a half). Otherwise, its CIR is additively increased as high as to the maximum traffic rate of the secondary tunnel. As the ingress node tunes the CIR, the traffic splitting ratio will be adaptively changed as well.

[5.2.](#) Adaptive Sequence Alignment

The usage of the bonding reordering buffer is timely monitored and reported to the ingress node. A threshold for this usage is pre-configured according to bandwidth or calculated in real-time according to the traffic sending rate. Whenever this threshold is detected to be violated, the ingress node intentionally splits the next incoming packet parade to the lightly loaded (or faster) tunnel until the usage of the bonding reordering buffer drops below the threshold.

Alternatively, a RTT difference threshold could be used in the same way, i.e., the ingress node will temporarily stop sending packets to the heavily loaded (or slower) tunnel when the RTT difference of the two tunnels is detected to be larger than that threshold.

[6.](#) Protocol Extensions

TBD.

The specification about protocol extensions in this document is intended to be applicable to various bonding tunnel protocols.

[7.](#) Security Considerations

Security should be considered by specific bonding tunnel protocols.

8. IANA Considerations

This document does not require any allocations by the IANA and therefore does not have any new IANA considerations.

9. References

9.1. Normative References

- [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>>.
- [RFC2697] Heinanen, J. and R. Guerin, "A Single Rate Three Color Marker", [RFC 2697](#), DOI 10.17487/RFC2697, September 1999, <<http://www.rfc-editor.org/info/rfc2697>>.
- [RFC2698] Heinanen, J. and R. Guerin, "A Two Rate Three Color Marker", [RFC 2698](#), DOI 10.17487/RFC2698, September 1999, <<http://www.rfc-editor.org/info/rfc2698>>.
- [RFC2890] Dommety, G., "Key and Sequence Number Extensions to GRE", [RFC 2890](#), DOI 10.17487/RFC2890, September 2000, <<http://www.rfc-editor.org/info/rfc2890>>.
- [RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", [RFC 6040](#), DOI 10.17487/RFC6040, November 2010, <<http://www.rfc-editor.org/info/rfc3168>>.
- [RFC6582] T.Henderson, S.Floyd, A.Gurtov, Y.Nishida, "The NewReno Modification to TCP's Fast Recovery Algorithm", [RFC 6582](#), DOI 10.17487/RFC6582, April 2012, <<http://www.rfc-editor.org/info/rfc6582>>
- [CUBIC] I.Rhee, & L.Xu, "CUBIC: A New TCP-Friendly High-Speed TCP Variant", <<http://www4.ncsu.edu/~rhee/export/bitcp/cubic-paper.pdf>>

9.2. Informative References

- [RFC2018] M.Mathis, J.Mahdavi, S.Floyd, A.Romanow, "TCP Selective Acknowledgment Options", [RFC 2018](#), DOI 10.17487/RFC2018, October 1996, <<http://www.rfc-editor.org/info/rfc2018>>
- [RFC3168] K. Ramakrishnan, S. Floyd, D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", [RFC 3168](#), DOI 10.17487/RFC3168, September 2001, <<http://www.rfc-editor.org/info/rfc3168>>.
- [RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", [RFC 5681](#), DOI 10.17487/RFC5681, September 2009, <<http://www.rfc-editor.org/info/rfc5681>>.
- [GREbond] N. Leymann, C. Heidemann, M. Zhang, et al, "GRE Tunnel Bonding", [draft-zhang-gre-tunnel-bonding](#), work in progress.
- [GTPbond] P. Muley, W. Henderichx, G. Liang, H. Liu, "Network based Bonding solution for Hybrid Access", [draft-muley-network-based-bonding-hybrid-access](#), work in progress.
- [MIPbond] P. Seite, A. Yegin and S. Gundavelli, "Multihoming support for Residential Gateways", [draft-seite-dmm-rg-multihoming](#), work in progress.

Author's Addresses

Nicolai Leymann
Deutsche Telekom AG
Winterfeldtstrasse 21-27
Berlin 10781
Germany

Phone: +49-170-2275345
Email: n.leymann@telekom.de

Cornelius Heidemann
Deutsche Telekom AG
Heinrich-Hertz-Strasse 3-7
Darmstadt 64295
Germany

Phone: +4961515812721
Email: heidemannc@telekom.de

Liang Geng
China Mobile
32 Xuanwumen West Street,
Xicheng District, Beijing, 100053,
China

EMail: gengliang@chinamobile.com

Jun Shen
China Telecom Co., Ltd
109 West Zhongshan Ave, Tianhe District
Guangzhou 510630
P.R. China

EMail: shenjun@gsta.com

Mingui Zhang
Huawei Technologies
No.156 Beiqing Rd. Haidian District,
Beijing 100095 P.R. China

EMail: zhangmingui@huawei.com

INTERNET-DRAFT Load Rebalancing for Bonding Tunnels February 6, 2018

Lihao Chen
Huawei Technologies
No.156 Beiqing Rd. Haidian District,
Beijing 100095 P.R. China

EMail: lihao.chen@huawei.com

Margaret Cullen
Painless Security
14 Summer St. Suite 202
Malden, MA 02148 USA

EMail: margaret@painless-security.com

