TSVWG R. Even
Internet-Draft M. Liu
Intended status: Informational Y. Zhang
Expires: August 7, 2020 Huawei

February 4, 2020

Data Center Fast Congestion Management
draft-even-tsvwg-datacenter-fast-congestion-00

Abstract

A good congestion control for data centers (DC) should provide low latency, fast convergence and high link utilization. Since multiple applications with different requirements may run on the DC network it is important to provide fairness between different applications that may use different congestion algorithms. An important issue from the user perspective is to achieve short Flow Completion Time (FCT). This document proposes data center congestion control direction aiming to achieve high performance while proving fairness.

Status of This Memo

This Internet-Draft is submitted 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 https://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 August 7, 2020.

Copyright Notice

Copyright (c) 2020 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 (https://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. Conventions 3
3. Congestion Handling Cases 3
3.1. Congestion only in leaf switch connected to receiver . . 3
3.2. Congestion in the Spine switch 4
3.2.1. ECN case 4
3.2.2. Spine and leaf switches share information 4
3.2.3. FCR from spine and leaf switches 4
3.3. Congestion in leaf switch connected to data sender . . . 4
4. Summary 4
5. Rate Information 5
6. Requirements 5
7. Implementation Options 6
8. Tests results 6
8.1. Many senders to one receiver 6
9. Security Considerations 8
10. IANA Considerations 8
11. References 9
11.1. Normative References 9
11.2. Informative References 9
Authors' Addresses 10

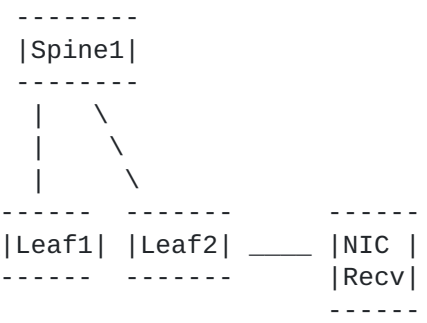
1. Introduction

The major use case that we are looking at is congestion control for Data Centers, a controlled environment as specified in RFC8085[RFC8085]. With the emerging Distributed Storage, AI/HPC (High Performance Computing), Machine Learning, etc., modern datacenter applications demand high throughput (40Gbps and above) with ultra-low latency of less than 10 microsecond per hop from the network, with low CPU overhead. The end to end latency should be less than 50usec, this value is based on DCQCN [DCQCN]. The high link speed (>40Gb/s) in Data Centers (DC) are making network transfers complete faster and in fewer RTTs. Network traffic in a data center is often a mix of short and long flows, where the short flows require low latencies and the long flows require high throughputs.

A good congestion control for data centers (DC) should provide low latency, fast convergence and high link utilization. Since multiple applications with different requirements may run on the DC network it

is important to provide fairness between different applications that may use different congestion algorithms. An important issue from the user perspective is to achieve short Flow Completion Time (FCT).

A typical DC architecture is composed of a spine-leaf topology where there are three hop switches at most for a flow. If we look from the flow view then we can assume that for the first hop switch there is low probability for congestion. The congestion will happen in higher probability at the spine or the last hop. The figure bellow shows a simple spine-leaf topology; in a typical DC there will be multiple Spines and Leaves.



2. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Congestion Handling Cases

3.1. Congestion only in leaf switch connected to receiver

The leaf switch is congested and does not receive any ECN CE marking on incoming streams. The leaf switch sends FCR (Fast Congestion Response) message to all sending NICs. The general case requires that the leaf switch will know who are the senders and if they support FCR. There is also a requirement to define how the congested leaf connects and send the FCR message to the senders. If not all senders whose streams are congesting the same egress port support FCR the congested leaf switch will drop back to use ECN CE marking to the

receiver. Another option is to send FCR to the senders that support it and use ECN CE marking on the flows from senders that do not support FCR, in this case the switch should wait for at least one RTT

before sending a second FCR to allow all senders to drop their sending rate.

3.2. Congestion in the Spine switch

There are a couple of options for supporting this case. The Spine and leaf switch will need to be aware of which option is in use.

3.2.1. ECN case

The leaf switch receives ECN CE marks from the spine. The leaf switch does not know what rate information it can send regardless if it is congested or not. The leaf switch will convey ECN marking to the receiver.

3.2.2. Spine and leaf switches share information

The Spine switch provides rate/congestion information to the downstream leaf switch. The leaf switch may be congested or not but will be responsible to send the FCR message to the sending NICs. The information from the spine may provide rate information using an FCR like message.

3.2.3. FCR from spine and leaf switches

The Spine switch will send FCR to the sending NICs and will not send ECN marking to the downstream leaf switch. In this option if there is also congestion on the downstream leaf a second FCR message will be sent from the leaf to the sending NIC who will have to use the lower recommended rate information.

3.3. Congestion in leaf switch connected to data sender

This case has lower probability but in case of congestion the leaf switch will send FCR message to all the contributing NICs of the flow causing the congestion on the congested egress port. If FCR is not supported by all the congesting NICs, the switch will CE mark these flows, this will cause the FCR supporting NICs to respond faster and the switch should allow the other streams to respond (wait little over an RTT time) before sending another FCR.

4. Summary

If all NICs currently sending data to the leaf switch support FCR messages it is safe to use FCR and if the congestion is in the Spine switch the action will be according to the options in section Section 3.2.

If the leaf switch knows that not all NICs sending data through the switch support FCR, the leaf switch may fall back to ECN marking. Another option is to use mixed mode by sending FCR to supporting NICs and ECN marks towards the receiver, the senders that support FCR should use the received FCR and ignore the ECN message from the receiver.

In the case where there are multiple congestion points, the NIC should use the lowest rate information from all received FCRs.

5. Rate Information

The leaf switch needs to supply rate information using the FCR message. The same rate information will be sent to all data senders to the congested port regardless of the rate they needed. This may cause underutilization of the available bandwidth if some of them have no need for all the recommended rate; this will be addressed by the leaf switch sending updated rate information based on the current usage after a number of RTTs. The leaf switch may also send updated FCR message when more bandwidth is available, for example when senders stop sending. Note that sending such information may cause congestion on upstream switches; another option is to use the sender congestion control to raise the sending rate according to its CC algorithm.

In the tests that were done so far, the solution was that all senders received the same rate information. We need to specify what we would like to send as the content of the rate information in the FCR message (bits/sec, number of bytes to send similar to wnd in TCP).

6. Requirements

To support FCR based on the above use cases requires:

1. The congested leaf should be able to know which data sources support FCR.

2. The congested leaf should be able to send the FCR message in-path for example by using TCP/UDP options or in the UDP applications back channel. Another option is to establish a connection to the data senders and send FCR messages to them.

3. Sender should be able to start sending at maximum rate if the new stream is the only stream sent by the sender.

7. Implementation Options

The FCR message from the network to the data sender MUST only be deployed in a controlled environment [RFC8085] such as Data Centers. The FCR message should provide an identification of the stream for example by providing the source and destination IP and Port number of the flow.

FCR should only be deployed in an intra-data-center environment where both endpoints and the switching fabric are under a single administrative domain. FCR MUST NOT be deployed over the public Internet

1. The tests are based on ROCEv2 [RoCEv2] using a revised CNP message and assume all senders support FCR. To use this option for ROCEv2 the data sender should mark support for the revised CNP message, this will allow the leaf switch to know if it can send back the revised CNP. This implementation mode is for testing only, we do not propose this mode for a solution.

2. For the proposed solution for the general case there may be a couple of options. The preference is to use a generic message at the transport level (TCP/UDP) otherwise will need a different message per application. The suggested proposal is to use new TCP option [RFC0793] and [RFC2460] for IPv6 that can be piggy backed on the ack message. There should be an FCR support option sent by the data sender. For UDP where a back channel is usually in the application layer we can use UDP options [I-D.ietf-tsvwg-udp-options] for announcing FCR support and using the application back channel in an application extension or in a UDP option to send FCR (In the testing we used a revised CNP message for ROCEv2). Another option is to use IOAM like mechanism (the general IOAM specification is [I-D.ietf-ippm-ioam-data], the loopback option is in [I-D.ietf-ippm-ioam-flags], sending message from the leaf switch can be based on https://tools.ietf.org/id/draft-ioamteam-ippm-ioam-direct-export-00.txt)

8. Tests results

Note: this can be an appendix later if relevant

8.1. Many senders to one receiver

In this test scenario we had six senders and one receiver on a single

switch on a 25 Gbit/sec connection. Five senders were sending long flows to create congestion and the sixth sender sent continuous 8 Bytes packets to test latency.

Network Average Load	NIC CC	Network CC	Improvement percentage
30%	1.61	1.61	0.00%
50%	2.68	2.68	0.00%
80%	4.23	4.24	0.24%
100%	4.36	4.51	3.44%

Sender NIC BW(Gbps)

The bandwidth of NIC CC and Network CC are almost the same in the long flows case

Network Average Load	NIC CC	Network CC	Improvement percentage
30%	5.89	5.79	1.70%
50%	6.04	6.04	0.00%
80%	7.33	6.67	9.00%
100%	7.45	6.78	8.99%

Latency flow result(us) - Average

Network Average Load	NIC CC	Network CC	Improvement percentage

30%	8.89	7.65	13.95%
50%	9.14	8.46	7.44%
80%	12.94	8.74	32.46%
100%	11.6	8.74	24.66%

Latency flow result(us) - 99%

Network Average Load	NIC CC	Network CC	Improvement percentage
30%	21.77	8.79	59.62%
50%	24.89	11.8	52.59%
80%	23.45	9.36	60.09%
100%	22.91	9.19	59.89%

Latency flow result(us) - 99.9%

We can see that the average latency is reduced by maximum 9% and the 99.9% latency which indicates the maximum queue size is reduced by maximum 60%.

The results show in that for the long flow many-to-one situation the Network CC achieves the same bandwidth as the NIC CC and better latency for mice flow.

9. Security Considerations

The FCR message is hard to secure, sending an FCR message from the network to the source has security risks since it can be easily used for DOS attack. This solution must only be used in a managed network [RFC8085]. The FCR message must be terminated in the managed network and should not cross the network domain.

Since this message is sent in a closed managed network it does not have the same security concerns as ICMP source quench message [RFC5927] defined on the general Internet.

An attacker can send an FCR message with lower or higher rate information. This may cause an underutilization of the network or congestion. The network entity closest to the receiver should provide an alert if an unexpected rate is being used which may hint that such an attack is taking place. A sender may also try to identify if the FCR message has rate information in the expected range.

10. IANA Considerations

TBD

11. References

11.1. 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, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

11.2. Informative References

[DCQCN] Zhu, Y., Eran, H., Firestone, D., Guo, C., Lipshteyn, M., Liron, Y., Padhye, J., Raindel, S., Yahia, M. H., and M. Zhang, "Congestion control for large-scale RDMA deployments. In ACM SIGCOMM Computer Communication Review, Vol. 45. ACM, 523-536.", 8 2015, <https://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p523.pdf>.

[I-D.ietf-ippm-ioam-data] Brockners, F., Bhandari, S., Pignataro, C., Gredler, H., Leddy, J., Youell, S., Mizrahi, T., Mozes, D., Lapukhov, P., Chang, R., daniel.bernier@bell.ca, d., and J. Lemon, "Data Fields for In-situ OAM", draft-ietf-ippm-ioam-data-07 (work in progress), September 2019.

[I-D.ietf-ippm-ioam-flags] Mizrahi, T., Brockners, F., Bhandari, S., Sivakolundu, R., Pignataro, C., Kfir, A., Gafni, B., Spiegel, M., and J. Lemon, "In-situ OAM Flags", draft-ietf-ippm-ioam-flags-00 (work in progress), October 2019.

[I-D.ietf-tsvwg-udp-options] Touch, J., "Transport Options for UDP", draft-ietf-tsvwg-udp-options-08 (work in progress), September 2019.

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <https://www.rfc-editor.org/info/rfc793>.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, DOI 10.17487/RFC2460, December 1998, <https://www.rfc-editor.org/info/rfc2460>.

[RFC5927] Gont, F., "ICMP Attacks against TCP", RFC 5927, DOI 10.17487/RFC5927, July 2010, <https://www.rfc-editor.org/info/rfc5927>.

[RFC8085] Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <https://www.rfc-editor.org/info/rfc8085>.

[RoCEv2] "Infiniband Trade Association. Supplement to InfiniBand architecture specification volume 1 release 1.2.2 annex A17: RoCEv2 (IP routable RoCE).", <https://cw.infinibandta.org/document/dl/7781>.

Authors' Addresses

Roni Even
Huawei

Email: roni.even@huawei.com

Mengzhu Liu
Huawei

Email: liumengzhu@huawei.com

Yali Zhang
Huawei

Email: zhangyali369@huawei.com