TSVWG R. Even
Internet-Draft Huawei
Intended status: Informational March 10, 2019
Expires: September 11, 2019

Fast Congestion Response
draft-even-fast-congestion-response-00

Abstract

The high link speed (100Gb/s) in Data Centers (DC) are making network transfers complete faster and in fewer RTTs. The short data bursts requires low latency while longer data transfer require high throughput. This document describes the current state of flow control and congestion handling in the DC using RoCEv2 and suggests new directions for faster congestion control.

Table of Contents

1. Introduction

The high link speed (100Gb/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. [RFC8257] titled Data Center TCP (DCTCP): TCP Congestion Control for Data Centers is an Informational RFC that extends the Explicit Congestion Notification (ECN) [RFC3168] processing to estimate the fraction of bytes that encounter congestion, DCTCP then scales the TCP congestion window based on this estimate. DCTCP does not change the ECN reporting in TCP. Other ECN notification mechanisms are specified for RTP in [RFC6679] and for QUIC [I-D.ietf-quic-transport]. The ECN notification are reported from the end receiver to the sender and the notification includes only the occurrence of ECN in the TCP case and the number of ECN marked packet for RTP and QUIC. What is common for TCP, RTP and QUIC is that the switches in the middle just monitor and report while the analysis and the rate control are done by the data sender.

In Data Centers the InfiniBand Architecture (IBA) offers a rich set of I/O services based on an RDMA access method and message passing semantics. RDMA over Converged Ethernet (RoCEv2) [RoCEv2] is using UDP as the transport for RDMA. RoCEv2 Congestion Management (RCM) provides the capability to avoid congestion hot spots and optimize the throughput of the fabric. RCM relies on the Link-Layer Flow-Control IEEE 802.1Qbb(PFC) to provide a lossless network. RoCEv2 Congestion Management(RCM) use ECN [RFC3168] to signal the congestion to the destination. The ECN notification is sent back from the receiver to the data sender using RoCEv2 Congestion Notification Packet (CNP) that notifies the sender about ECN marked packets. The rate reduction by the sender as well as the increase in data injection is left to the implementation.

2. Conventions

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 [RFC2119] [RFC8174]

3. Problem statement

The congestion control using ECN in the DC is done between the receiver and the sender. The network measures the traffic and informs the receiver about problems by the ECN bit. The Receiver will send to the Sender in the RoCEv2 case, a CNP message and the sender adapts by reducing the rate. The sender reduces the rate based on pre-defined policy. The sender has also a policy about when to start sending at a higher rate and by how much to increase the traffic. In the DC network when latency and high transfer rate is important there is a need to define a congestion response mechanism that will be optimized for the DC network. The behavior of the sender on congestion is not specified by RoCEV2.

This type of congestion management is re-active. The high link speed in the DC (100Gb/s) are making network transfers complete faster and in fewer RTTs; allocating flows their proper rates as quickly as possible becomes a priority. The convergence time must become a primary metric for congestion control in high speed networks.

A pro-active direction will provide more information to the sender about the congestion that can be used to optimize the congestion response allowing the network to adapt faster to the changes in the traffic conditions. This information should be available to the sender to allow fast response (RTT or lower).

The entity that measures the congestion is the switch in the network. Currently it just notifies about congestion to the receiver (ECN), may drop packets (the receiver may use IEEE 802.1Qbb to provide a lossless network). The receiver NIC informs the sender about the ECN; the sender will analyze, control and execute an action to address the congestion based on some predefined policy.

The requirement is to allow the network to control the traffic instead of the end points. The proposal is to allow the network to analyze the congestion and inform the sender (QPSource in terms of ROCEv2)) how to handle the congestion when in the transport layer (directly to the data sender). In the case of RoCEV2 as the transport protocol can be a new Congestion Notification Message. This requires a new message from the network to the sender (backward notification). The proposed solution for the DC should only be deployed in an intra-data-center environment where both endpoints and the switching fabric are under a single administrative domain.

4. Security Considerations

TBD

5. IANA Considerations

No IANA action

6. References

6.1. 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>.

[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>.

6.2. Informative References

[I-D.ietf-quic-transport] Iyengar, J. and M. Thomson, "QUIC: A UDP-Based Multiplexed and Secure Transport", draft-ietf-quic-transport-18 (work in progress), January 2019.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <https://www.rfc-editor.org/info/rfc3168>.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, DOI 10.17487/RFC3550, July 2003, <https://www.rfc-editor.org/info/rfc3550>.

[RFC6679] Westerlund, M., Johansson, I., Perkins, C., O'Hanlon, P., and K. Carlberg, "Explicit Congestion Notification (ECN) for RTP over UDP", RFC 6679, DOI 10.17487/RFC6679, August 2012, <https://www.rfc-editor.org/info/rfc6679>.

[RFC8257] Bensley, S., Thaler, D., Balasubramanian, P., Eggert, L., and G. Judd, "Data Center TCP (DCTCP): TCP Congestion Control for Data Centers", RFC 8257, DOI 10.17487/RFC8257, October 2017, <https://www.rfc-editor.org/info/rfc8257>.

Author's Address

Roni Even
Huawei

Email: roni.even@huawei.com