INTERNET-DRAFT
Intended Status: Informational
Expires: Feb 09, 2019

F. Chen
W. Sun
Huawei Technologies

Aug 10, 2018

Problem Statement of RoCEv2 Congestion Management
draft-chen-nfsv4-rocev2-cm-problem-statement-00

Abstract

On IP-routed datacenter networks, RDMA is deployed using RoCEv2 protocol. RoCEv2 specification does not define the congestion management and load balancing methods. RoCEv2 relies on the existing Link-Layer Flow-Control IEEE 802.1Qbb(Priority-based Flow Control, PFC)to provide a lossless network. RoCEv2 Congestion Management(RCM) use ECN(Explicit Congestion Notification, defined in RFC3168) to signal the congestion to the destination and use the congestion notification to reduce the rate of injection and increase the injection rate when the extent of congestion decreases. More and more practice of congestion management for RoCEv2 appear in the industry, such as DCQCN(Data Center Quantized Congestion Notification). There is a demanding for the new RoCE protocol(temporary alias RoCEv3) to provide stronger congestion management and load balancing mechanisms for RDMA deployment in modern datacenter.

Table of Contents

1 Introduction

With the emerging Distributed Storage, AI/HPC, Machine Learning, etc., modern datacenter applications demand high throughput(40Gbps and above) with ultra-low latency of < 10 microsecond per hop from the network, with low CPU overhead. Remote Direct Memory Access (RDMA) can meet these needs on Ethernet.

On IP-routed datacenter networks, RDMA is deployed using RoCEv2 protocol. RoCEv2 is a straightforward extension of the RoCE protocol that involves a simple modification of the RoCE packet format. RoCEv2 packets carry an IP header which allows traversal of IP L3 Routers and a UDP header that serves as a stateless encapsulation layer for the RDMA Transport Protocol Packets over IP[1].

RoCEv2 Congestion Management (RCM) provides the capability to avoid congestion hot spots and optimize the throughput of the fabric. RCM relies on the existing Link-Layer Flow-Control IEEE 802.1Qbb(PFC) to provide a drop free network. RoCEv2 Congestion Management(RCM) also use ECN(RFC3168) to signal the congestion to the destination and use the congestion notification to reduce the rate of injection and increase the injection rate when the extent of congestion decreases.

More and more practice of congestion management for RoCEv2 appear in the industry, such as DCQCN, etc. Shall we consider to develop next Generation RoCE protocol(alias RoCEv3) with stronger congestion management and load balancing mechanisms for RDMA deployment in modern datacenter?

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 Abbreviations

RCM - RoCEv2 Congestion Management

PFC - Priority-based Flow Control

ECN - Explicit Congestion Notification

DCQCN - Data Center Quantized Congestion Notification

AI/HPC - Artificial Intelligence/High-Performance computing

ECMP - Equal-Cost Multipath

4 Problem statement & requirements

Network congestion happens in the network switches when the incoming traffic is larger than the bandwidth of the outgoing link on which it has to be transmitted. Congestion is the primary source of loss and in the network, congestion leads to dramatic performance degradation.

Generally, RoCEv2 relies on Link-Layer Flow-Control IEEE 802.1Qbb(PFC) to provide a lossless underlying networks. Lossless networks implement mechanism of flow control, which pauses the traffic in the incoming link before the buffer overfills, and by that prevents case of dropping packets[2]. However, PFC can lead to poor application performance due to problems like head-of-line blocking and unfairness[3]. In order to avoid the problems involved by PFC, there is another faction research on the congestion control mechanisms over the lossy network.

We need a kind of protocol with stronger capability of congestion management to achieve the high throughput and low latency in the large-scale datacenter network with more flexible requirement to the underlay network. The interoperability is also required among the industry practice.

5 Current Congestion Management for RoCEv2

5.1 PFC

RDMA is deployed using the RoCEv2 protocol, which relies on IEEE 802.1Qbb Priority-based Flow Control (PFC) to enable a drop-free network.

PFC is a link level protocol that allows a receiver to assert flow control telling the transmitter to pause sending traffic for a specified priority. However, because PFC will stop all traffic in a particular traffic class at the ingress port, the flows destined to other ports will also be blocked.

The known problems of PFC are head-of-line blocking, unfairness, deadlock[4].

5.2 ECN

Explicit congestion notification (ECN) enables end-to-end congestion notification between two endpoints on TCP/IP based networks. ECN notifies networks about congestion with the goal of reducing packet loss and delay by making the sending device decrease the transmission rate until the congestion clears, without dropping packets. RFC 3168, The Addition of Explicit Congestion Notification (ECN) to IP, defines ECN.

6. Congestion Management Practice

6.1 Packet Retransmission

NICs were not designed to deal with losses efficiently. Receiver discards out-of-order packets. Sender does go-back-N on detecting packet loss. RoCEv2 adopt Go-back-N loss recovery and needs lossless layer 2 (by using PFC) for good performance[5].

If new RDMA protocol does not rely on the lossless layer 2 network, an efficient method of Packet Retransmission is necessary.

6.2 Congestion Control Mechanisms

6.2.1 RTT-based Congestion Control

The typical practice of RTT based Congestion Control is TIMELY[6]. It introduces the simple packet delay, measured as round-trip times at hosts, is an effective congestion signal without the need for switch feedback. TIMELY measures RTT with microsecond accuracy, and that these RTTs are sufficient to estimate switch queueing. TIMELY can adjust transmission rates using RTT gradients to keep packet latency low while delivering high bandwidth. TIMELY is a delay-based congestion control protocol for use in the datacenter.

Because the RDMA transport is in the NIC and sensitive to packet drops, so PFC is necessary because drops hurt performance badly. That is to say TIMELY needs PFC to provide lossless underlay network.

6.2.2 Credit-based Congestion Control

ExpressPass[7] is an end-to-end credit-scheduled, delay-bounded congestion control for datacenters. ExpressPass uses credit packets to control congestion even before sending data packets, which enables to achieve bounded delay and fast convergence. It uses end-to-end credit transfer for bandwidth allocation and fine-grained packet scheduling.

6.2.3 ECN-based congestion control

Data Center Quantized Congestion Notification (DCQCN)[3] is an end-to-end congestion control scheme for RoCEv2. DCQCN is a combination of ECN and PFC to support end-to-end lossless Ethernet. The idea behind DCQCN is to allow ECN to do flow control by decreasing the transmission rate at the sender when congestion starts, thereby minimizing the time PFC is triggered.

Although RoCEv2 standard[1] does not list DCQCN as the RCM mechanism, but it is widely used in the industry practice.

6.3 Re-ordering

When the packets arrive at the destination out-of-order, the destination should store the packets to restore the order. Destination should assign special buffer resource to perform re-ordering. There are many methods to implement the re-ordering either on switch or on NIC side. Here will not go into the details.

6.4 Load Balancing

6.4.1 ECMP

RoCEv2 packets use an opaque flow identifier in their UDP Source Port field for ECMP method to implement path selection mechanisms for load balancing and improve utilization of the fabric topology.

Traditional ECMP can not balance loads well in the data center network because it splits loads at the granularity of flow.

The finer the granularity of load balancing, the more effective the load balancing is and the higher the utilization of network bandwidth can be achieved.

6.4.2 Flowlet

The typical Flowlet-based load balancing is CONGA[8]. CONGA is a network-based distributed congestion-aware load balancing mechanism for datacenters. It splits TCP flows into flowlets, estimates real-time congestion on fabric paths, and allocates flowlets to paths based on feedback from remote switches.

Flowlets are bursts of packets from a flow. The idle interval between two bursts of packets is larger than the maximum difference in latency among the paths. So the second burst can be sent along a different path than the first without reordering packets.

6.4.3 Per-packet

The effect of packet-based load balancing is the best because the corresponding granularity is the smallest. The consequence is that packets belonging to the same flow will be allocated to different paths. When the forwarding delays of paths are different, it is possible that packets may arrive at the receiver out-of-order.

7 Summary

The new emerging RoCE based applications urge the practice of different congestion management mechanisms to be practiced in kinds of modern large-scale datacenter network. In this problem statement, not all the mainstream mechanisms are introduced. It is still needed to extend when considering the future RoCE protocol(temporary alias

RoCEv3) with robot congestion management capability and more flexible requirement on layer 2 network which might be the next direction.

8 Security Considerations

This document does not introduce any additional security constraints.

9 IANA Considerations TBD

10 References

[1] Infiniband Trade Association. Supplement to InfiniBand architecture specification volume 1 release 1.2.2 annex A17: RoCEv2 (IP routable RoCE), 2014.

[2] Understanding RoCEv2 Congestion Management, https://community.mellanox.com/docs/DOC-2321

[3] Zhu, Yibo, et al. "Congestion Control for Large-Scale RDMA Deployments." Acm Sigcomm Computer Communication Review 45.5(2015):523-536.

[4] Hu, Shuihai, et al. "Deadlocks in Datacenter Networks: Why Do They Form, and How to Avoid Them." The, ACM Workshop ACM, 2016:92-98.

[5] Mittal, Radhika, et al. "Revisiting Network Support for RDMA." (2018).

[6] Mittal, Radhika, et al. "TIMELY: RTT-based Congestion Control for the Datacenter." ACM Conference on Special Interest Group on Data Communication ACM, 2015:537-550.

[7] Cho, Inho, D. Han, and K. Jang. "ExpressPass: End-to-End Credit-based Congestion Control for Datacenters." (2016).

[8] Alizadeh, Mohammad, et al. "CONGA: distributed congestion-aware load balancing for datacenters." ACM Conference on SIGCOMM ACM, 2014:503-514.

Authors' Addresses

Fei Chen
Huawei Technologies Co., Ltd.
Email: chenfei57@huawei.com

Wenhao Sun
Huawei Technologies Co., Ltd.
Email: sam.sunwenhao@huawei.com