# Poster: The Congestion Path Multiplicity Problem in RDMA Multicast Congestion Control

Keita Aoki Miki Yamamoto
Faculty of Engineering Science, Kansai University
Suita, Osaka 564–8680 Japan
Email: {k971155, yama-m}@kansai-u.ac.jp

***Abstract*—In multicast RDMA, the sender overestimates congestion signals when multiple paths are congested. This congestion path multiplicity problem is hard to be resolved because RDMA sender cannot identify the sender of congestion signals due to RDMA protocol regulation. In this paper, we propose a new approach that the branching switch filters congestion signals from the worst congested interface. Our evaluation results show that our proposed method can adequately resolve the congestion path multiplicity problem.**

***Index Terms*—data center network, congestion control, multicast, RDMA**

## I. Introduction

In recent years, RDMA (Remote Direct Memory Access) [1]- [3] has been widely deployed in data center networks. RDMA can transmit data without heavy overhead of TCP/IP, and is suitable for delay sensitive services. Data replication is one of the most important services in a data center. RDMA is originally designed for unicast communications and data replication by unicast results in inefficient bandwidth usage. To address these issues, MC-RDMA [4] has been proposed.

MC-RDMA is well designed for reliable one-to-many communications. End points (hosts) can behave as unicast communications but networks (switches) operate some important tasks for reliable multicast communications. ACKs and NAKs are aggregated at the branching point of the multicast tree. However, it has no mechanism for aggregation of congestion signals, CNP (Congestion Notification Packet). This causes overly regulated throughput because of independent CNPs from multiple paths. We call this technical problem the *congestion path multiplicity problem*.

In this paper, we would like to propose Time window Max CNP Filtering (TMCF) method. In our proposed method, the switch at the branching point selects the worst congested interface and tentatively passes only CNPs from this selected interface during each time window. Our evaluation results show that TMCF can select the worst congested path and adjust transmission rate adequately.

## II. MC-RDMA

In MC-RDMA, multicast tree, i.e. multicast routing, is managed in the control plane. The sender transmits a packet whose destination address (destination Queue Pair number) is a virtual address which is similar to a multicast address in IP. At the branching point, when an interface has one destination, i.e. it is the final branching point on the path, this virtual address is exchanged to the real destination QP number. This operation is managed by the control plane and executed by P4 programmable switch. ACKs and NAKs are adequately merged. For example, only the ACK with the smallest ACK number is forwarded to the sender.

979-8-3503-5171-2/24/$31.00 ©2024 IEEE

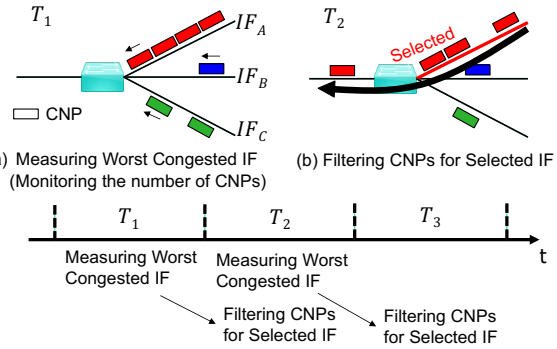


Fig. 1. Time Window Max CNP Filtering

In RDMA data centers, congestion control is one of the most important technical problems because RoCEv2 (RDMA over Converged Ethernet version 2) is designed to maintain lossless operations. For multicast traffic, it is also the case, but MC-RDMA fails to implement work-well CNP aggregation at the switch.

## III. Congestion Path Multiplicity Problem in Multicast RDMA

Loss path multiplicity problem (LPMP) is well-known technical problem for multicast congestion control [5]. Independently generated congestion signals (packet losses) on multiple paths in multicast tree throttle transmission rate, which causes over regulation and accordingly throughput degradation. One promising approach for LPMP is selection of the worst congested path at the sender.

In MC-RDMA, CNPs are not merged at switches and all CNPs arrive at the sender. However, RDMA protocol has no mechanism for identifying the source of a CNP (the receiver sending the corresponding CNP). This is because a CNP has only the destination QP number, i.e. QP number for the source node. This means the worst congested path cannot be identified at the source node in MC-RDMA.

When all CNPs are forwarded at each switch, congestion signals from all congested paths arrive at the source node. We call this new technical problem, similar to LPMP, congestion path multiplicity problem (CPMP).

## IV. Time Window Max CNP Filtering

In MC-RDMA, end points understand they behave as in unicast communications, and the sender cannot identify even

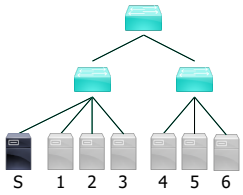


Fig. 2. Simulation Model

TABLE I
SIMULATION PARAMETERS

| | |
|---|---|
| $Bandwidth$ | 40[Gbps] |
| $RTT$ | 4[$\mu$s], 8[$\mu$s] |
| $K_{min}$ | 50[KB] |
| $K_{max}$ | 200[KB] |
| $P_{max}$ | 0.01 |
| $g$ | 1/256 |
| $Timeslot$ | 50[$\mu$s] |

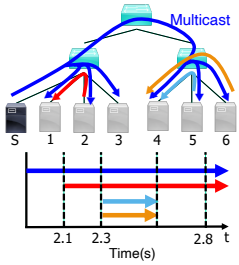


Fig. 5. Dynamic Bottleneck Model

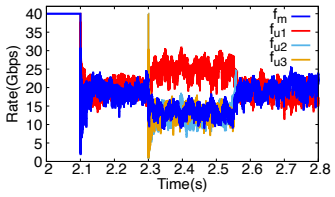


Fig. 6. Transmission Rate

though it receives all CNPs from all of the congested paths. We believe that the networks, i.e. the switches should operate adequately for multicast congestion control because end points cannot deal with CPMP.

To resolve CPMP in multicast RDMA, we propose a time window max CNP filtering method which is operating at each branching switch. The concept of TMCF is that a switch filters CNPs from the worst congested interface. When all switches on the worst congested path filter CNPs from their worst congested interface, the sender can receive CNPs from the worst congested path. TMCF is a time-slot based method. To identify the worst congested interface, a switch counts the number of arrived CNPs for each interface in each time slot. A switch selects the interface with the largest number of CNPs as the worst congested interface and forwards all CNPs from this selected interface in the next time slot. In Figure 1, CNPs from the selected interface, $IF_A$, in time slot $T_1$ are forwarded to the sender in time slot $T_2$. Even when a path includes multiple branching switches, TMCF can select the worst interface at each switch and accordingly only the CNPs from the worst congested path arrive at the source node. By this selection of the worst congested path, TMCF is expected to resolve CPMP.

## V. PERFORMANCE EVALUATION

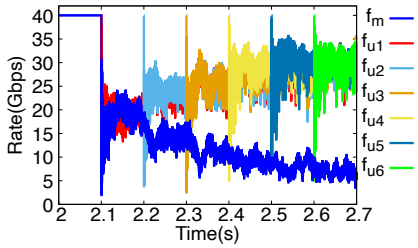


Fig. 3. Transmission Rate (DCQCN)

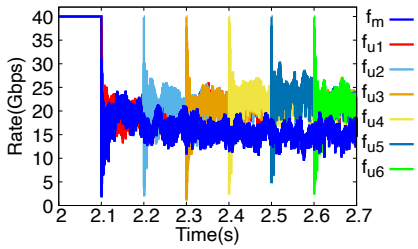


Fig. 4. Transmission Rate (DCQCN with TMCF)

We evaluate our proposed TMCF by network simulator. We implemented MC-RDMA and our proposed TMCF in RDMA package of ns-3 [6]. Simulation topology and parameter settings are shown in Fig. 2 and Table 1, respectively.

First, we evaluate CPMP. At the start time of simulation, only multicast traffic from the source node (S) to 6 receivers (node 1-6) starts its transmission. From 2.1s, one additional unicast flow is initiated every 0.1s. So, at 2.6s, there exists 6 unicast flows (1→2, 2→3, 3→1, 4→5, 5→6, 6→4), each of which shares a corresponding bottleneck link with the multicast flow. DCQCN [7] is used for RDMA congestion control algorithm for all flows. Figures 3 and 4 show transmission rate characteristics of DCQCN with and without TMCF, respectively. Without TMCF (Fig. 3), with increase of congestion paths (with increase of unicast flows), multicast transmission rate is degraded. The reason for this degradation is CPMP. With our proposed TMCF (Fig. 4) transmission rate of the multicast flow is not degraded with increase of unicast flows and link bandwidth of each bottleneck link (e.g. from 2.6s, there are 6 homogeneous bottleneck links) is fairly shared with corresponding unicast flow.

We also evaluate TMCF in the case that the worst congested path is dynamically changed. Multicast flow with 6 receivers starts first and a unicast flow from node 2 to 1 starts at 2.1s. In this case link connected to node 1 is the bottleneck link (Fig. 5). From 2.3s, two unicast flows (5→4 and 6→4) start and the bottleneck link is changed to the link connected to node 4. After these additional two unicast flows end their transmission[1], the bottleneck link goes back to the link connected to node 1. From evaluation results in Figure 6, in our proposed TMCF, transmission rate of multicast flow is adequately controlled to fairly sharing transmission rate with the corresponding unicast flow(s) on the bottleneck link.

## VI. CONCLUSIONS

In this paper, we proposed a time window max CNP filtering method to solve the congestion path multiplicity problem (CPMP) for multicast RDMA. In our proposed method, only the CNPs from the most congested interface at each branching switch are forwarded to the sender. This allows the sender to adequately adjust its transmission rate according to the most congested path in the multicast tree. Evaluation results showed that our method significantly improves CPMP.

**A**cknowledgement : This work is partly supported by JSPS KAKENHI Grant Number 23K21662 and 21H03434.

[1]In ns3 RDMA package, stop time of flows cannot be explicitly designated, so we just set flow length (file size of unicast flows). This is the reason why we do not explicitly depict end time of these two flows in Figure 5.

## REFERENCES


[1] Infiniband Trade Association, "InfiniBand architecture specification volume 1, release 1.3", https://cw.infinibandta.org/document/dl/7161.

[2] Infiniband Trade Association, "InfiniBand architecture specification volume 2, release 1.3", https://cw.infinibandta.org/document/dl/8125.

[3] Infiniband Trade Association, "Supplement to InfiniBand architecture specification volume 1 release 1.2.1 annex A17: RoCEv2 (IP routable RoCE)", https://cw.infinibandta.org/document/dl/7781.

[4] C. Huang, et al., "MC-RDMA: Improving Replication Performance of RDMA-based Distributed Systems with Reliable Multicast Support," in *Proc. IEEE ICNP'23*, Los Alamitos, CA, USA, Oct. 2023.

[5] S. Bhattacharyya, et al., "The loss path multiplicity problem in multicast congestion control," in *Proc. IEEE INFOCOM'99*, New York, NY, USA, Mar. 1999.

[6] https://github.com/alibaba-edu/High-Precision-Congestion-Control.

[7] Y. Zhu, et al., "Congestion Control for Large-Scale RDMA Deployments," in *Proc. ACM SIGCOMM'15*, London, United Kingdom, Aug. 2015.