THEMIS: Efficiently Mitigating Congestion-Induced Fairness Disparities in Long-Haul RDMA Networks

Rixin Liu[1], Menghao Zhang[1], Zihan Niu[1], Zili Meng[2], Xiaohe Hu[3]
[1]Beihang University [2]HKUST [3]Infrawaves

1 INTRODUCTION

In recent years, cross-DC (Data Center) services have emerged to meet the growing demand for big data processing, scientific research, and artificial intelligence. Due to its kernel-bypass and protocol-offload feature, RDMA networks naturally provide high throughput and low latency. Therefore, to obtain high network performance, it is a natural design choice for cloud service providers to apply RDMA to cross-DC scenarios (e.g., Microsoft Azure [5]).

However, extending RDMA to long-haul scenarios is non-trivial. Since a sender has to wait for information from receiver before responding to network congestion, inter-DC traffic takes hundreds of times longer to react than intra-DC traffic. Therefore, existing congestion control mechanisms based on feedback from receiver (e.g., DCQCN [8], HPCC [7]) ultimately results in intra-DC traffic being suppressed unfairly. To illustrate this, we conduct an experiment in a three-tire topology without oversubscription, as shown in Figure 1. The latency of inter-DC link is 500us (i.e., corresponding to 100km distance), and the hop-to-hop latency within DC is 1 us. We use the WebSearch workload [7], which is characterized by small requests and large responses, with the ratio of intra-DC traffic to inter-DC traffic around 5:1 [2]. We use the same metric as HPCC, FCT Slow Down (a flow's actual FCT normalized by its ideal FCT when the network only has this flow), to indicate the increase in flow latency. Table 1 illustrates the unfairness issue is severe with common congestion control algorithms. Note that although HPCC usually performs better than DCQCN in intra-DC scenarios, it adversely deteriorates the unfairness problem in cross-DC circumstances, since it exacerbates the suppression of intra-DC flows due to its focus on minimizing queue buildup.

While there are several works (e.g., Bifrost [4], Swing [6]) attempting to optimize the PFC buffer requirement in long-haul RDMA networks, the congestion-induced fairness disparities are largely ignored. Annulus [2], a representative work suitable for cross-DC RDMA networks, uses a dual congestion control loop which relies on L3 routed QCN to reduce feedback time from near-source congestion point. Since Annulus requires an L2 table to implement the L3 routed QCN mechanism, this mechanism is only deployed on ToR switches. Hence, Annulus does not eliminate the long feedback loop of congestion signals issued from remote bottlenecks, and cannot address near-destination unfairness. In addition, conventional wisdom has already realized the RTT-fairness issue in TCP networks. Prague algorithm [3]

Table 1: Average FCT Slow Down in commonly used RDMA congestion control algorithms.

Average FCT Slow Down	DCQCN	HPCC
Intra-DC	33.179	56.946
Inter-DC	1.7787	2.3561
Unfairness	18.654	24.170

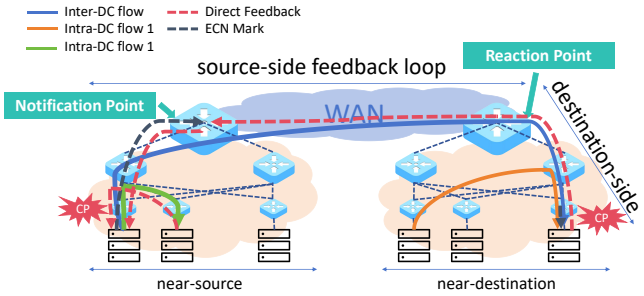

Figure 1: THEMIS Overview and Workflow.

sets a minimum value of 25ms for *rtt_virt* of all flows to reduce the gap between the RTTs of long and short flows. And it uses the value of $\frac{rtt\_real}{rtt\_virt}$ to instruct changes in the congestion window (cwnd), which essentially reduces the cwnd growth/decrease rate of short-RTT flows. However, this approach can not be applied to RDMA networks directly. On one hand, most commodity RDMA NICs [10] do not support window-based congestion control. On the other hand, unlike TCP that receives ACKs per packet, RDMA receives ACKs per operation, which thus has much larger ACK intervals than TCP and cannot get RTT feedback timely.

Long feedback loop of inter-DC traffic is the crux of unfairness in long-haul networks. During the period from when congestion starts until the feedback signal arrives at the remote inter-DC traffic sender, only intra-DC traffic is suppressed. Meanwhile, inter-DC traffic continues at its original speed, causing queue buildup at the congestion point. This results in a significant increase in intra-DC FCT, while inter-DC traffic FCT is relatively less affected. To mitigate this issue, we present THEMIS, a fairness maintenance patch for widely deployed RDMA congestion control algorithms, via preemptive notification points and reaction points at external switch. It is built on the following key principles: (1) **Equal feedback loop for intra/inter-DC traffic.** We expect that both types of traffic experience an equal feedback loop. Therefore, we consider to use the External Switch (SW) to achieve proactive response and feedback. (2) **Easy for deployment.** We aim to achieve minimum modifications to existing data center infrastructure, and the required functionalities should be commonly available in commercial

devices. Therefore, we use only programmable External SW to implement the entire design. Our preliminary evaluations demonstrate Themis is highly effective in mitigating this unfairness issue.

2 THEMIS DESIGN

Since DCQCN is the default and most widely used congestion control algorithm in large-scale RDMA networks, we are currently focusing only on the adaptation to DCQCN. Note that Themis can be easily applied to switch-assisted algorithms like HPCC. Figure 1 shows the workflow of Themis. First, to address near-source unfairness, the External SW generates CNPs in response to ECN marks, advancing the notification point, and it will selectively forward CNPs from the receiver. Second, to address near-destination unfairness, External SW responds to feedback signals from its local data center, advancing the reaction point.

2.1 Notification Point on External SW

Unlike traditional approaches that use the receiver as the notification point, Themis leverages the external switch to shorten the feedback path. Although Annulus uses the QCN mechanism to treat the ToR switch as a near-source notification point, its mechanism violates the extensibility requirement. Even if it extends the L3 routed QCN mechanism to higher-level switches, it still faces the issue of over-suppression in the presence of multiple bottlenecks. A naive solution is to mark the packets of a flow after a switch sends feedback, preventing downstream switches from issuing multiple feedback for the same flow. However, this approach requires that all the switches in the network are capable of generating CNP packets, which does not align with the easy deployment we previously mentioned.

Fortunately, we found that External SW is a sweet spot to be a preemptive notification point, which can easily achieve both fairness and to avoid over-suppression. First, Themis will not modify any mechanisms or devices within the data center, and the ECN mechanism in DCQCN will function normally. Second, besides sending a CNP when receiving a near-source ECN mark, Themis also clears the ECN mark to prevent the External SW on the other side of WAN from reacting, thus avoiding over-suppression. Third, to align with DCQCN CNP generation mechanism, Themis creates a hash mapping based on the five-tuple and timestamp to record the time t_{last} it last sent a CNP for a specific flow. When it receives a CNP from the WAN or encounters another ECN mark, it will only forward or generate the CNP if the difference between the current time t_{cur} and t_{last} is larger than $t_{interval}$ (e.g., 50us). However, when near-destination congestion occurred, the time difference in receiving feedback between the senders of inter-DC flows and intra-DC flows remains significant. Therefore, this module alone cannot fully resolve the unfairness issue, and we need the next module.

2.2 Reaction Point on External SW

Instead of waiting for the CNP from the receiver, delayed by the wide-area network, the external SW in Themis responds directly to CNPs from its local data center. Moreover, to avoid overreacting to the target flow, the reaction point marks the CNP it receives with a single bit, so the the external SW near the sender will not react to this CNP. In this way, Themis segments the entire feedback loop into two sections, one from the sender to the external switch near the receiver, the other is the from the external switch to the receiver, as shown in Figure 1. Therefore, the root cause of near-destination unfairness can be mitigated.

However, proactively controlling the sending rate of a specific flow is challenging to achieve at switches. We observe that `resubmit`, an operation for employing multiple packet processing procedure on the same packet [9], has the potential to fulfill our goal. It has the following two low-overhead characteristics: first, it occurs in the ingress pipeline, avoiding waste of the switch's egress pipeline processing capacity; second, it only moves some metadata without relocating the packets themselves, reducing the impact on switch memory. Leveraging this advantage, we can achieve the switch-level reaction point at minimum cost. Further, similar to the reaction point algorithm in the DCQCN algorithm, Themis keeps track of the last time a CNP was received, and dynamically adjusts the blocking time of the target flow based on the frequency of received CNPs, denoted as T_{block}. After determining the blocking time, and since the queue depth of the switch is dynamically changing, we need to calculate the minimum number of resubmits n in real time, according to $\sum_{i=1}^{n} \frac{qlen_i}{rate_{drain}} = T_{block}$. In the equation, n represents the number of resubmit operations and $qlen_i$ represents the queue length at the beginning of the i-th resubmit. And since the External SW has only one port connected to the WAN, the draining rate of the ingress queue is usually equal to the line rate of the link. To avoid out-of-order packet delivery, the number of resubmits required for each packet is no less than the number of resubmits for its preceding packets.

3 EVALUATION AND FUTURE WORK

We implement an open-source Themis prototype [11] based on the HPCC NS3 project [1]. As Table 2 shows, Themis significantly mitigates the unfairness issues at both near the source and near the destination. In the future, we plan to implement a full Themis prototype with programmable switches (e.g., Tofino), support more congestion control algorithms (e.g., RTT-based, and conduct extensive evaluations in real testbeds.

Table 2: Themis significantly improves fairness.

Average FCT Slow Down	DCQCN	DCQCN w/ Themis
Intra-DC	33.179	12.389
Inter-DC	1.7787	1.9116
Unfairness	18.654	6.4810

REFERENCES

[1] Alibaba. 2020. High Precision Congestion Control. https://github.com/alibaba-edu/High-Precision-Congestion-Control.

[2] Ahmed Saeed et al. 2020. Annulus: A Dual Congestion Control Loop for Datacenter and WAN Traffic Aggregates. In *ACM SIGCOMM 2020*.

[3] K. D. Schepper et al. 2024. Prague Congestion Control. https://www.ietf.org/archive/id/draft-briscoe-iccrg-prague-congestion-control-04.html.

[4] P. Yu et al. 2023. Bifrost: Extending RoCE for Long Distance Inter-DC Links. In *ICNP 2023*.

[5] Wei Bai et al. 2023. Empowering Azure Storage with RDMA. In *USENIX NSDI 2023*.

[6] Y. Chen et al. 2023. Swing: Providing Long-Range Lossless RDMA via PFC-Relay. *TPDS 2023* (2023).

[7] Yuliang Li et al. 2019. HPCC: high precision congestion control. In *ACM SIGCOMM 2019*.

[8] Yibo Zhu et al. 2015. Congestion control for large-scale RDMA deployments. In *ACM SIGCOMM 2015*.

[9] The P4.org Architecture Working Group. 2021. PSA Specification. https://p4.org/p4-spec/docs/PSA.html#sec-clone.

[10] NVIDIA Mellanox. 2024. ConnectX-6. https://www.nvidia.com/en-sg/networking/ethernet/connectx-6/.

[11] Themis. 2024. Themis-NS3. https://github.com/Eternal579/Themis.