This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

R-AQM: Reverse ACK Active Queue Management in Multitenant Data Centers

Xinle Du, Ke Xu, *Senior Member, IEEE, Member, ACM*, Lei Xu, Kai Zheng, *Senior Member, IEEE*, Meng Shen, *Member, IEEE*, Bo Wu, and Tong Li, *Member, IEEE*

***Abstract*— TCP incast has become a practical problem for high-bandwidth, low-latency transmissions, resulting in throughput degradation of up to 90% and delays of hundreds of milliseconds, severely impacting application performance. However, in virtualized multi-tenant data centers, host-based advancements in the TCP stack are hard to deploy from the operators' perspective. Operators only provide infrastructure in the form of virtual machines, in which only tenants can directly modify the end-host TCP stack. In this paper, we present R-AQM, a switch-powered reverse ACK active queue management (R-AQM) mechanism for enhancing ACK-clocking effects through assisting legacy TCP. Specifically, R-AQM proactively intercepts ACKs and paces the ACK-clocked in-flight data packets, preventing TCP from suffering incast collapse. We implement and evaluate R-AQM in NS-3 simulation and NetFPGA-based hardware switch. Both simulation and testbed results show that R-AQM greatly improves TCP performance under heavy incast workloads by significantly lowering packet loss rate, reducing retransmission timeouts, and supporting 16 times (i.e., 60 to 1000) more senders. Meanwhile, the forward queuing delays are also reduced by 4.6 times.**

***Index Terms*— Data center, multi-tenant, ACK, AQM.**

I. Introduction

DATA centers have evolved rapidly over the last few years, providing a wide variety of cloud services [5], [50] using TCP as the dominant transport layer protocol. However, the TCP incast problem causes drastic performance degradation when multiple senders synchronously send data to one receiver (i.e., many-to-one communication) with high-bandwidth and low-latency links [12], [62]. As the number of senders increases, bottleneck switches can quickly become overfilled. Inevitable packet drops would impose TCP retransmission timeout (RTO) for hundreds of milliseconds, resulting in goodput (the application-level throughput [39]) reduction of up to 90% [55], which affects the performance of applications.

Recently, a large number of improvements of TCP have been proposed [5], [18], [45], [55], [61], [63]. Some work identifies the cause of performance degradation and suggests adjusting existing congestion control (CC) parameters to match the data center network. For instance, Reducing-RTO [55] reduces the minimum retransmission timeout (RTO_{min}) value and reduces unnecessary waiting after packet drops. Others have suggested redesigning CC, using a new lossless RDMA (Remote Direct Memory Access) based network stack, or even designing entirely new data center transmission protocols. For example, DCTCP [5] accurately controls the total throughput through the explicit congestion notification (ECN) identifier provided by the switch to avoid overloading the switch buffer and packet loss. DCQCN [63] is a CC for the lossless network protocol RoCEv2 (RDMA over Converged Ethernet version 2) [8], which uses Priority-based Flow Control (PFC) [32] to avoid buffer overflow by forcing the immediate upstream entity to pause data transmission. NDP [45] redesigns the entire data center transport protocol, including routing and CC, to provide low latency and high throughput.

Although many of the above proposals have proven to be commercially available, they face a great challenge on real-world deployment in public and multi-tenant data centers [37]. This is because the common physical infrastructures such as switches and network interface cards (NIC) are shared by multiple tenants in the form of virtual machines (VMs). It is the tenants who are able to deploy applications in the VMs, select the corresponding transport layer protocols, and decide end-system protocol stack parameters such as ECN support and RTO_{min} value. Consequently, from the perspective of operators of multi-tenant data centers, when the tenants have already run the VMs, it requires more effort to modify the network protocol stack than to modify the common physical infrastructures (see Section II-B). In this case, simply changing the controllable physical infrastructure without any modifications to the legacy transport protocol stack in order to improve transmission performance transparently would be a contribution for data center operators.

Manuscript received 24 September 2021; revised 7 March 2022; accepted 14 July 2022; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor C. Peng. This work was supported in part by the China National Funds for Distinguished Young Scientists under Grant 61825204, in part by NSFC Project under Grant 61932016, and in part by the Beijing Outstanding Young Scientist Program under Grant BJJWZYJH01201910003011. *(Corresponding author: Tong Li.)*

Xinle Du and Lei Xu are with BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: dxl18@mails.tsinghua.edu.cn; thuxl07@gmail.com).

Ke Xu is with BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with PCL, Shenzhen 518066, China (e-mail: xuke@tsinghua.edu.cn).

Kai Zheng is with Huawei, Shenzhen 518129, China (e-mail: kai.zheng@huawei.com).

Meng Shen is with the Beijing Institute of Technology, Beijing 100081, China (e-mail: shenmeng@bit.edu.cn).

Bo Wu was with BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. He is now with Tencent Technologies, Shenzhen 518054, China (e-mail: wub14@tsinghua.org.cn).

Tong Li is with the Key Laboratory of Data Engineering and Knowledge Engineering, Information School, Renmin University of China, Beijing 100872, China (e-mail: tong.li@ruc.edu.cn).

Digital Object Identifier 10.1109/TNET.2022.3197973

1558-2566 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

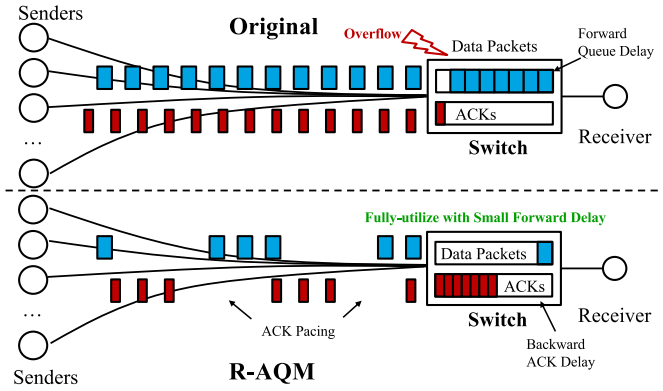


Fig. 1. The general idea of R-AQM – an illustrative example.

A basic idea of transparently enhancing the transport protocol stack is an intrusive modification to the headers of packets forwarded by switches. For example, HSCC [2] rewrites the value of the receive window (denoted by $rwnd$) to one MSS (Maximum Segment Size) in the ACK headers for all congested flows. These approaches, however, are limited by an artifact of the current window-based transport protocol design (e.g., NewReno [25], CUBIC [49], and DCTCP [5]), in which the window indicates the number of full-sized packets. In other words, $rwnd$ can not be rewritten to a proper fraction (i.e., $rwnd \notin (0,\ 1)$). This coarse granularity significantly limits the scale of concurrency.

In this paper, we present a new mechanism called R-AQM (Reverse Active Queue Management), which is transparent to end-systems and fine-grained. Figure 1 illustrates the general idea of R-AQM. The fundamental premise of R-AQM is ACK-clocking [34], i.e., ACKs not only acknowledge receipts of data packets but also trigger new packet sending. Unlike existing AQM schemes [9], [21] that intrusively modify the content of packets, R-AQM proactively intercepts ACKs to prevent the source from sending the next packet too fast, which also slows down the increase of the sending window. In this way, R-AQM is able to deploy active queue management for ACKs in the reverse path to adjust the in-flight traffic without overwhelming the switch in the case of incast congestion.

The rest of the paper is organized as follows. We introduce the background of the TCP incast problem, the deployment challenges for multi-tenant data centers, and the degradation of goodput by RTO in Section II. Section III illustrates the design rationale of our solution. The detailed design of R-AQM is demonstrated in Section IV. In Section V, we address the implementation of R-AQM on NetFPGA and P4. In Section VI, Section VII and Section VIII, we evaluate R-AQM in NS-3 and a small-scale testbed. Section IX surveys the related work. Finally, Section X concludes this paper.

II. Background and Motivation

A. TCP Incast Problem Hurts Performance

TCP incast is a catastrophic goodput collapse that occurs as the number of servers sending data to a client increases beyond the ability of an Ethernet switch to buffer packets. This scenario often happens intra data center communication when requesting data for file systems [52], during the shuffle phase of cloud computing systems [16], and in the partition/aggregate pattern of large-scale web applications [5]. The synchronous request workload causes packets to exceed the buffer on the bottleneck link, resulting in severe packet losses. Packet loss further causes costly timeout, which lasts for hundreds of milliseconds (varies in different scenarios). As a result, the goodput of the link drops due to wasting opportunities for sending data during the retransmission timeouts. We also give a quantitative analysis on the TCP incast problem below.

Assume that N incast flows with the same RTT are sharing a bottleneck link with the capacity of C. Each flow needs to send X bits, let n be the number of RTTs it takes to complete the transfer of one flow and m be the average RTO times. As illustrated in [2], the average goodput of the link is given by:

$$Goodput = \frac{X}{n \cdot RTT + m \cdot RTO + \frac{X \cdot N}{C}}$$

In the case that RTT $= 100\mu$s and RTO $=$ 100ms (the lower bound in the Linux implementations), since RTO is usually three orders of magnitude larger than RTT, it is easy to see that a single time of RTO can lead to a sharp drop in goodput.

B. Deployment Challenges for Multitenant Operators

To control the TCP incast, data center operators need to upgrade their hardware and (or) software. In private data centers, administrators can change the physical infrastructure such as switches and network interface cards (NIC), and also modify the transport protocol stack at end-systems. Therefore, improvements (e.g., Reducing-RTO [55], DCTCP [5], NDP [45]) are possible to be deployed by systematically upgrading infrastructures and systems.

However, in public and multi-tenant data centers, it requires more effort for operators to modify the network protocol stack than to modify the common physical infrastructures. In virtualized multi-tenant data centers [37], [59], the common physical infrastructures are shared by multiple tenants in the form of virtual machines (VMs). Generally, the data center operators deploy a default transport protocol stack in the system image of each VM. It is the tenants who are able to deploy applications or systems in the VMs, select the corresponding transport layer protocols, and decide end-system protocol stacks (e.g., Use BBR [11] between the user and the data center, use DCTCP [5] or NewReno [25] with ECN within the data center) and parameters (e.g., ECN support and RTO_{min} value).

Consequently, from the operators' perspective, it requires extra effort to modify the network protocol stack after tenants have already run the VMs. For example, enabling virtual CC in the hypervisor as specified in prior works such as AC/DC TCP [27] and vCC [15]. Both of which provide congestion agents in the hypervisor that transparently place efficient CCs for tenant VMs. However, these methods require full TCP state tracking and full TCP finite-state machines in

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

the hypervisor, which may overload the hypervisor and slow it down considerably. In addition, since incast usually happens on the last-hop switch, the end-to-end hypervisor-based way may still suffer from incast problems. In other words, simply applying the hypervisor-based solution only solves part of the problems [45].

Based on the above observations, we seek a solution that not only works on the incast problems but also transparent to the TCP stack at end hosts.

C. Fine Granularity Requirement of Window Control

HSCC is a switch-based congestion controller [2] that rewrites the value of receive window (denoted by $rwnd$) to one MSS (Maximum Segment Size) in the ACK headers for all congested flows without modifying TCP itself. However, HSCC cannot cooperate with legacy TCP CCs very well in data centers. Legacy TCP CCs in the Linux Kernel are almost window-based (NewReno [25], CUBIC [49], BBR [11], and DCTCP [5]). These window-based CCs have fundamental flaws in small RTT networks, because they cannot reduce the sending window infinitely (i.e., not less than 1 MSS). Since the bandwidth-delay product (BDP) in a data center network is usually small due to the small RTT, it is very easy for the in-flight packets to become larger than the BDP+buffer. In this case, the extra packets can only be dropped or be resent by retransmission. However, when incast occurs, flows may fail to build large enough in-flight packets to recover via fast retransmission (e.g., 3-duplicate ACKs). As a result, this coarse granularity significantly limits the scale of concurrency and we need fine-grained window control.

To motivate the requirement of fine-grained window control, we give a modeling analysis as below. Assume that the window size of flow i at time t is $w_i(t)$, the switch buffer and the link capacity are B and C, respectively. The queue size $q(t)$ in the switch at time t in the case of N incast flows is given by:

$$q(t) = \sum_{i=1}^{N} w_i(t) - C \cdot RTT$$

In the case of a large number of concurrent flows when $N \cdot MSS - C \cdot RTT \geq B$, a considerable proportion of flows may fall back to the stop-and-wait paradigm to avoid packet loss and RTO. That is, there must be some flows stopping sending data and setting the window to zero. This coarse granularity significantly limits the scale of concurrency. Particularly, the number of concurrent flows is limited to 40–60 in most modern switches [5], [10], [57]. However, this is not nearly enough to sustain real data center communications. For example, a cluster running data mining tasks have more than 80 concurrent flows per node [23], [60]; In Facebook's Memcached cluster [47], a single Web server may access over 100 Memcached servers. Worse, a production data center with 6000 servers supporting Web search applications has over 1000 concurrent traffic on work nodes [5]. It is obvious that even 1 MSS of sending window per flow is enough to overwhelm the switch buffer on a burst.

To better understand how does the granularity of window control impacts the scale of concurrency, we further conduct a simulation. Each sender sends a 320KB message to a fixed receiver. Three common CCs (NewReno with ECN, NewReno without ECN, DCTCP) and one particular CC that always sets the congestion window to 1 MSS (similar to HSCC [2]) are investigated. Figure 2 shows the goodput and RTO times with different scales of concurrency. Some insights are listed below:

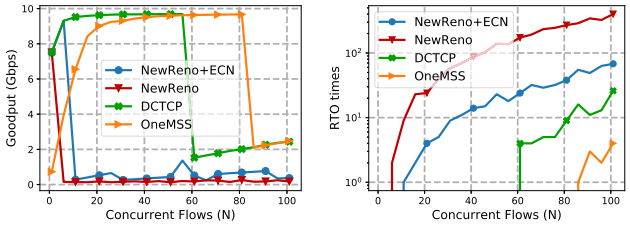
(1) Even one RTO occurs, the loss of goodput is enormous.

(2) NewReno does not work well. Even with ECN, the number of senders cannot exceed 10.

(3) DCTCP can alleviate the occurrence of incast collapse, but the concurrency can only be maintained around 60.

(4) Even if the sending window is always 1 MSS, only about 80 concurrent flows can be maintained.

In summary, a more fine-grained window control is needed to solve the incast problem. Meanwhile, switch-based mechanisms have the potential to overcome the deployment challenges for multi-tenant data centers. These greatly motivate the design ideas and principles of R-AQM.



(a) Goodput in incast (b) RTO times in incast

Fig. 2. The experiment to show the incast problem.

III. Design Rationale

Our goal is to design an incast control mechanism in the multi-tenant data center to handle as many concurrent connections as possible effectively. We came up with a new active ACK control approach called R-AQM. The critical factor that inspires the new ACK control approach is that if a sender does not receive an ACK, the sender cannot send the next data packet. If the ACK can be intentionally delayed, the senders' following sending action will also be delayed accordingly. The protocol that relies on the arrival of ACK packets to infer that the network can accept more packets is called the window-based ACK-clocking protocol [33], [34].

R-AQM enables the sources to alternate between two operation modes: one is the standard protocol mode (e.g., TCP additive-increase/multiplicative-decrease (AIMD)), where the source sends data according to its sending window in a legacy way. Another is the R-AQM mode, where the ACK delayed by the switch delays each source sending action. The alternation between the two modes happens in response to switch signals, based on the congestion level observed in the switch buffer. When the queue in the switch buffer builds up, the switch triggers the R-AQM mode. The switch can proactively intercept the ACK packet in the backward direction. When the queue recedes, the senders resume the sending window.

An implicit consequence of this scheme is that short-lived incast traffic is positively discriminated when it is most likely

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

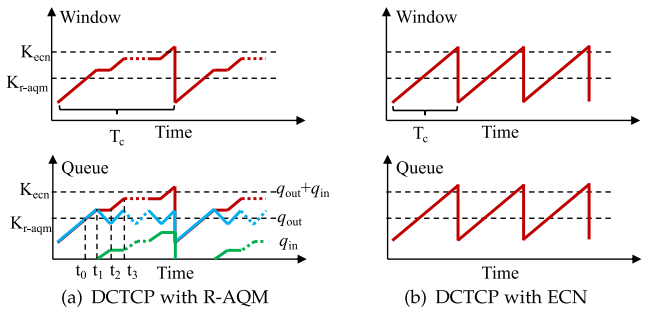


Fig. 3. Window size and queue size process.

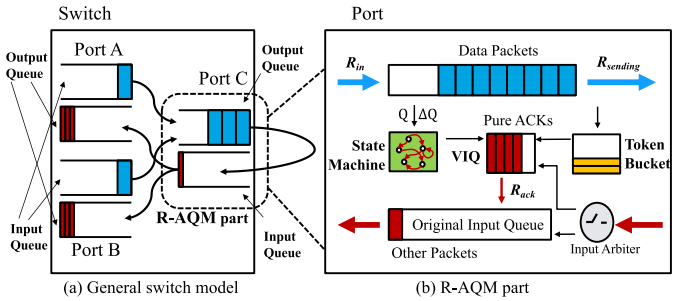


Fig. 4. R-AQM design.

to experience nonrecoverable losses immediately after the connection is set up. That is, when many synchronized flows surge, the buffer content builds up fast, and our scheme switches all ongoing flows to R-AQM mode. These flows react, typically 1/2 RTT later, by delaying their next sending action while the incast traffic flows are still within their three-way handshake (or sending their first few packets). The switch implicitly inhibits the sending window by proactively intercepting the ACK packet transparent to TCP in the end-systems.

Therefore, R-AQM naturally meets our requirements. When incast occurs, the switch can proactively intercept the ACK packet in the backward direction and send ACKs at a rate that does not make the ACK-triggered data packet overwhelming the switch. In this way, we can leverage active ACK control to adjust in-flight traffic without being constrained by the minimum window size shared by window-based solutions. We only need to adjust the ACK rhythm appropriately in the switch. Consequently, it is ideal for multi-tenant cloud data center networks. Moreover, because the bottleneck switch can capture the instantaneous queue length, it can sense incast more quickly and thus make decisions more quickly to prevent further congestion.

The rationale of R-AQM is to lower the nontrivial forward data queuing delay by introducing a trivial backward ACK queuing delay. In the incast traffic pattern applications, the forward packet is usually the service request packet (e.g., Reduce in MapReduce [16]), while the reverse packet is usually only the ACK without piggybacked data. Compared with the ACK's backward queuing delay, applications pay more attention to the forward queuing delay of the data packets. Forward delay refers to the time taken by a packet departing from the sender to the receiver, which is very important for the application's QoE. Backward delay in the reverse direction only delays the confirmation of a data packet [40], [41], which does not greatly impact QoE.

Figure 3 shows the window size and queue size with R-AQM and ECN. We now analyze the steady-state behavior of the R-AQM control loop in a simplified setting to understand how to convert the forward queuing delay to the backward queuing delay. We assume that the N flows are synchronized for convenience of understanding; i.e., their "sawtooth" window dynamics are in-phase. At time t_0, output queue length (Q_{out}) exceeds K_{r-aqm}, and the switch starts to buffer ACKs actively. From time t_0 to t_1, the senders need time to react to the ACK buffer action on the switch. From time t_1 to t_2, the senders do not receive ACKs, so the window remains unchanged. Meanwhile, the sending window does not change, the total number of packets in the network also does not change, so Q_{in}+Q_{out} remains unchanged. But the input queue length (Q_{in}) begins to grow, and the output queue length (Q_{out}) begins to decline. At time t_2, the source begins to receive ACKs again. R-AQM will repeat the same action to keep Q_{out} at a low level while the extra inflight ACK packets are stored in Q_{in}. Through the above steps, the data queue transforms into the ACK queue, and the forward queuing delay transforms into the backward queuing delay. R-AQM alleviates the excessive window growth rate of DCTCP (Figure 3(b)) without loss of throughput (Q_{out}>0).

With these benefits, the next questions are how to properly hold and send back ACKs in the switch, what problems active ACK interception can cause, how to fix it, and so on. In the next section, we introduce how we solve these problems by proposing R-AQM.

IV. R-AQM

R-AQM is an incast control mechanism that aims to mitigate buffer overflow problems by shaping ACKs in the switch through assisting legacy TCP. Figure 4(b) presents our design framework, which contains three main functional components: the Virtual Input Queue, the Token Bucket, and the State Machine. As shown in Figure 4, packets sent by the sender are queued on the bottleneck port as usual, and each packet a sender sends will be acknowledged by the receiver. (1) When the returned ACK enters the bottleneck port, the VIQ (virtual input queue) located in the switch input port recognizes ACKs, intercepts them and stores them; (2) The Token Bucket monitors the immediate egress sending packets and generates tokens to the bucket to trigger the VIQ dequeue action; (3) The State Machine parses queue length information, calculates the draining rate according to the congestion state. After the sender receives the ACK, the sender adjusts the sending rate and sends the next packet.

In this section, we propose our design by answering the following four questions:

- How to intercept and buffer ACKs?
- What is the ACK dequeue policy in the switch?
- How to determine the draining rate of ACKs?
- What are the side effects, and how to compensate?

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

A. How to Intercept and Buffer ACKs?

The first step is to distinguish between different ACKs, based on which the piggybacked ACKs (which can affect the application QoE) and ACKs with the FIN flag (which can not trigger a new data packet either) are excluded. In this paper, we define pure ACKs as ACKs without FIN and ACKs that are not piggybacked, which are determined by the combination of the packet size and the header tag. With the input arbiter, pure ACKs are queued in VIQ, and others are queueing in the original input queue. VIQ is located between the switch input port and the forward core. To better control the pure ACK draining rate, we need to separate the ACKs from other packets, setting up a virtual input queue for pure ACKs. To avoid reverse-path congestion causing packet loss or affecting RTT measurements, VIQ sets the highest priority of each ACK. Even if the ACK packet size is small, VIQ still needs some memory to store ACKs, so the design needs to consider how to drop packets. When the ACK is dropped, the sender will assume that the data was not received, which wastes forward throughput and might cause RTO.

There are two reasons why we choose to set up a VIQ instead of an ACK output queue. First, ACKs of congested flows should be proactively intercepted. As shown in Figure 4(a), port C's output queue is the congestion point. If the output queues of A and B are proactively intercepted, then the wrong ACKs from other ports may be buffered, affecting non-congested traffic. Second, deploying on an input queue is relatively easier. The input queue can directly obtain the output queue length within the same port, and the changes in operation logic are minimized, which does not affect the top design of the switch.

B. What Is the ACK Dequeue Policy in the Switch?

A proper switch implementation requires a hardware input queue, and its dequeue action needs to be controlled by a data plane. R-AQM uses the token bucket to trigger the dequeue action. The token bucket is an algorithm for traffic shaping in packet-switched networks. It can be used to check whether data transmission at the packet granularity conforms to the defined limits of bandwidth and burst, which measures the unevenness or variability of traffic. As shown in Figure 4(b), the token bucket controls the token input rate by monitoring the average value of $R_{sending}$. Each increment of a token triggers an enqueue action (not necessarily sending, see Section IV-C). Using the token bucket, on the one hand, we can regulate the sending rhythm of ACKs. On the other hand, the draining rate R_{ack} can be adjusted by controlling the proportion of the ACK consumption token.

C. How to Determine the Draining Rate of ACKs?

Having figured out how to buffer ACKs proactively, we need to figure out when to drain ACKs. We use the State Machine to judge the congestion state and adjust the ACK draining rate. A simple idea is mapping the output queue length directly to the ACK draining rate. The longer the queue, the more severe the congestion and the slower the ACK should be sent.

Algorithm 1 VIQ Send Algorithm. $\boldsymbol{state}$ Is the R-AQM State of One Port. α and n Are the Number of Tokens Consumed and the Number of ACKs Emitted at Each Time. α and n Control the ACK Draining Rate

```
function viq_send( )
    if state is NS and token ≥ α1 then
        token -= α1; VIQ.pop(n1) // Normal State
    else if state is DS and token ≥ α2 then
        token -= α2; VIQ.pop(n2) // Draining State
    else if state is CS and token ≥ α3 then
        token -= α3; VIQ.pop(n3) // Congest State
    end if
end function
```

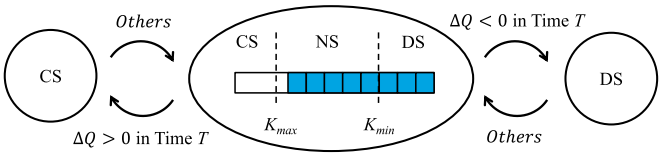


Fig. 5. R-AQM state transition diagram.

The shorter the queue, the less severe the congestion, and the faster the ACK should be sent. However, such a naive idea suffers from some issues.

First, mapping functions are costly for hardware [39]. The mapping of a linear function is difficult to implement in ASIC or FPGA hardware due to the requirement of the division operation. In general, the linear function is approximated by a step function [43], [48]. Second, the feedback latency causes oscillation. Since the network is a pipelined model, when R-AQM buffers ACKs, it does not immediately reduce congestion. Frequent changes in the ACK rate can cause oscillations. Third, queue length does not identify a burst. If a large amount of Incast traffic reaches the switch in a burst, the forwarding rate will be less than the queue input rate, and the queue length will continue to grow without any fluctuation [51], [53]. In such cases, R-AQM requires active buffering at the beginning of the burst, rather than waiting for the queue length to reach a specific threshold. This is because a low threshold is too sensitive to identify the burst, and a high threshold makes it difficult to identify a burst.

In R-AQM, we use queue length and its gradient to judge the congestion state comprehensively, and only three corresponding states are set. Figure 5 shows the state transition diagram, which is the most critical part of R-AQM. First, it calculates whether the queue length continues to grow or decline in time T. If the queue continues to grow, there will be a burst, so no matter how long the current queue is, it should trigger an active buffer to accommodate burst (left part of Figure 5). If the queue length continues to decrease, it means that the burst has ended. The ACK can be returned at this point, rather than waiting for the queue to decrease to a certain value (right part of Figure 5). When the network state is stable, just like the traditional AQM, it can be determined by the threshold (middle part of Figure 5).

Algorithm 1 illustrates the process of the ACK send action in the switch. Generating a token in the token bucket triggers

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

the procedure $viq\_send()$ at Line 1. There are three states to represent the different actions, namely Congest State (CS), Draining State (DS), and Normal State (NS). We use α to represent the number of tokens consumed and n to represent the number of ACKs emitted at each time. NS is the steady-state of the switch and requires only a uniform ACK response. In NS, $\frac{\alpha_1}{n_1} = 1$ (Line 2-3). DS indicates that the forward queue is about to empty, so we need to speed up emptying the reverse ACK queue. In DS, $\frac{\alpha_2}{n_2} < 1$ (Line 4-5). CS means extreme congestion. In CS, $\frac{\alpha_3}{n_3} \gg 1$ (Line 6-7). R-AQM needs to ensure that the ACK is sent at a low rate, but not stopped. First, it avoids RTO caused by senders that do not receive any ACKs for a long time. Second, it prevents some of the flows from starvation in the case of burst congestion.

D. What Are the Side Effects, and How to Compensate?

Interaction with TCP RTO:

One concern of R-AQM is its interaction with TCP RTO. R-AQM limits the rate of ACK in order to prevent RTO caused by packet loss, so it is inevitable to increase RTT. It is not sure whether this will cause RTT to be prolonged beyond RTO_{min}, leading to TCP timeouts and spurious retransmissions. For this reason, we specifically measure the RTTs in our experiments. We find that our ACK control does not adversely prolong the RTTs (for example, with 200 connections, the 99th percentile RTT is less than 0.3ms). And we do not observe any spurious retransmission.

Even though this phenomenon is rare, we still take into account the possibility and design counter-measures. As each pure ACK enters the switch, R-AQM records the time stamp in the auxiliary packet header. When each pure ACK exits the switch, R-AQM makes a judgment that if there are more than 5ms (which is recommended as the smallest RTO_{min} in [55]), it is considered as an old ACK (which may be caused by TCP timeout and retransmission), and will be dropped. The reason is that by dropping the out-of-order ACKs, R-AQM avoids disturbing the TCP at the sender for subsequent unnecessary retransmissions.

Interaction with TCP CC:

Another concern of R-AQM is its interaction with TCP CC. R-AQM takes effect before packet loss and ECN trigger. Therefore, when R-AQM senses congestion, it not only needs to delay the ACK transmission, but also needs to prevent the sender from increasing the sending window.

Two mechanisms are recommended to compensate. The first is a BECN-like mechanism [20] that directly marks the ECN-Echo in the TCP packet header of the ACK in the switch when there exists congestion. The senders therefore can use ECN to reduce the sending window, avoiding congestion quickly. Second, by considering packet loss as the congestion signal [38], it is recommended to adopt a mechanism similar to HSCC [2] that directly sets the ACK header's rwnd to 1 in the switch when incast congestion occurs. In R-AQM, when the output queue length + VIQ length exceeds the ECN threshold or HSCC threshold, the switch will be triggered to mark the ECN-echo flag on the ACK or set the rwnd of the ACK to 1.

TABLE I
VARIABLES OF FLUID MODEL

Variable	Description
w	Window size
N	Flow numbers
α	Estimated congestion degree of DCTCP
g	DCTCP's parameter
t	Time
q_{out}	Bottleneck output queue length
q_{in}	Bottleneck input queue length
q	$q_{in} + q_{out}$
R	Round-trip time (RTT)
C	Bottleneck link capacity
τ	the propagation delay

With these two mechanisms, R-AQM works well with existing CCs in the Linux kernel (NewReno [25], CUBIC [49], and DCTCP [5]). It can also coexist with different CCs and TCP settings, because R-AQM limits the sending window (rwnd by HSCC) to 1 MSS, therefore treating each sender fairly.

E. Stability Analysis

This section will use a flow model to demonstrate that R-AQM does not affect the original system's stability. We assume that the source uses DCTCP. The main symbols are summarized in Table I.

We now develop a fluid model by considering N long-lived flows traversing a single bottleneck link with capacity C. The following non-linear, delay-differential equations describe the dynamics of $w(t)$, $\alpha(t)$, and the queue size $q(t), q_{in}(t), q_{out}(t)$ in the switch:

$$\frac{dw}{dt} = \left(\frac{1}{R(t)} - \frac{w(t)\alpha(t)}{2R(t)}p(t-R^*)\right)(1 - r(t-R^*)) \tag{1}$$

$$\frac{d\alpha}{dt} = \frac{g}{R(t)}(p(t-R^*) - \alpha(t))(1 - r(t-R^*)) \tag{2}$$

$$\frac{dq}{dt} = (N\frac{w(t)}{R(t)} - C)(1 - r(t-R^*)) \tag{3}$$

$$\frac{dq_{in}}{dt} = C - C(1 - r(t-R^*)) \tag{4}$$

$$\frac{dq_{out}}{dt} = N\frac{w(t)}{R(t)}(1 - r(t-R^*)) - C \tag{5}$$

Here $p(t)$ indicates the ECN marking process at the switch and is given by:

$$p(t) = \mathbf{1}_{\{q(t) > K_{ecn}\}} \tag{6}$$

$r(t)$ indicates the R-AQM delaying process at the switch. To simplify the analysis model, we omit the Draining State and set the Congest State draining rate to 0, thus:

$$r(t) = \mathbf{1}_{\{q(t) > K_{r-aqm}\}} \tag{7}$$

and $R(t) = \tau + q(t)/C$ is the RTT, where τ is the propagation delay (assumed to be equal for all flows), and $q(t)/C$ is the queueing delay.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Equations (1) and (2) describe the DCTCP source, while (3)-(7) describe the queuing process at the switch and all AQM schemes. The source equations are coupled with the switch equations through the ECN marking process $p(t)$ and the R-AQM delaying process $r(t)$ which get feed back to the source with some delay. This feedback delay is approximately a fixed value $R^* = \tau + K_{ecn}/C$. The approximation aligns well with DCTCP's attempt to hold the queue size at around K_{ecn}.

Equation (1) models the window evolution and consists of the standard additive increase term, $1/R(t)$, and a multiplicative decrease term, $-w(t)\alpha(t)/2R(t)$. The latter term models the source's reduction of window size by a factor $\alpha(t)/2$ when packets are marked with ECN (i.e., $p(t-R^*)=1$). The term $(1-r(t-R^*))$ models the R-AQM process, which means that any behavior at the source side will be paused when R-AQM is in effect (i.e., $(1-r(t-R^*))=0$). Equation (2) is a continuous approximation of DCTCP's estimated congestion degree. Equation (3)-(5) models the queue evolution: $Nw(t)/R(t)$ is the network input rate and C is the service rate. Equation (4) models the input queue evolution: The first term C is the service rate and also the input queue's input rate. The last term $C(1-r(t-R^*)$ is the ACK draining rate controlled by R-AQM.

By setting the LHS (left hand side) of Equations (1)-(5) to zero, we see that the fluid model has a unique fixed point when $N \geq (C\tau + K_{ecn})/2$. Therefore, we have the following two operating regimes:

(i) $N \geq (C\tau + K_{ecn})/2$: In this regime, the model has a unique fixed point, namely:

$$(w, \alpha, q, q_{in}, q_{out}) = (2, 1, 2N - C\tau, 2N - C\tau - K_{r-aqm}, K_{r-aqm}) \quad (8)$$

This regime corresponds to the large N cases, where the system has a specific steady-state behavior: each source transmits two packets per RTT, of which $C\tau$ fills the link capacity, and the remaining $2N - C\tau$ build up a queue. All packets are marked as the queue constantly remains larger than K_{ecn}. There are $(2N - C\tau - K_{r-aqm})$ ACK packets in Q_{in} and K_{r-aqm} data packets in Q_{out}.

(ii) $N < (C\tau + K_{ecn})/2$: In this regime, the system does not have a fixed point. Instead, it has a periodic solution or limit cycle, characterized by a closed trajectory in state space. Figure 6 shows a sample phase diagram of the limit cycle projected onto the Window-Queue. As shown, all trajectories evolve towards the orbit of the limit cycle. The whole process is shown in Figure 2(a). This loop is similar to the pure DCTCP system loop [6], so we omit the proof for brevity.

DCTCP with R-AQM can guarantee the stability of the whole nonlinear system. Especially in the case of large N, R-AQM can store $(2N - C\tau - K_{r-aqm})$ ACK packets in Q_{in}, thus saving a large amount of storage space (ACK packet size: data packet = 64B:1460B). R-AQM ensures K_{r-aqm} data packets stored in the output queue, making the forward throughput not drop and the forward latency be small, so as to have a low completion time for mice flows. When N is small, the whole system is basically controlled by the DCTCP source.

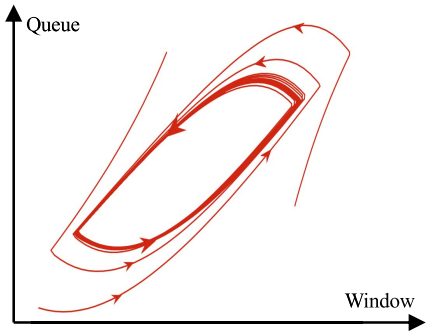


Fig. 6. Phase diagram showing occurrence of limit cycle.

R-AQM only ensures that the DCTCP window growth is not too aggressive, and at the same time, does not cause excessive throughput loss. In a word, R-AQM will not affect the stability of the original system. While easing the incast problem, it can also reduce the forward queue length and reduce the forward delay.

F. Discussions

Parameters Guidance: According to Algorithm 1, nine parameters of R-AQM need to be set. K_{max} and K_{min} represent the maximum and minimum values in the steady state, respectively, and T stands for burst duration. $\alpha_1, \alpha_2, \alpha_3$ and n_1, n_2, n_3 control the ACK draining rate in different state. The setting of these parameters is a trade-off. A small value of K indicates that R-AQM will be triggered when the forward queue is small, which will lead to low end-to-end latency but also low throughput. In order to keep high throughput, we need to make sure there are always packets in the forward queue. A small T means that R-AQM is very sensitive to burst and will slowdown active buffering, thus affecting queue length and packet loss rate. Therefore, K_{min} can be set to 0.5-1.0 times BDP, K_{max} can be set to 2 times BDP, and T can be set to 0.2-0.6 times RTT. As discussed in section IV-C, $\frac{\alpha_1}{n_1} = 1$, $\frac{\alpha_2}{n_2} < 1$ and $\frac{\alpha_3}{n_3} \gg 1$, this paper suggests $\frac{\alpha_2}{n_2} = 0.5$ and $\frac{\alpha_3}{n_3} = 10$. The recommended setting is also verified through simulations in Section VI-B.

Different ACK Mechanism: Piggybacked ACK and delayed ACK affect the use of R-AQM. First of all, typical inter-data center applications are MapReduce [16], GFS [22], etc., and piggybacked ACKs are rare. When the flows using pure ACKs coexist with flows using piggybacked ACKs, the flows using pure ACKs will respond to congestion more quickly and proactively. In this case, R-AQM can also support more senders, but it cannot guarantee the fairness of the two kinds of flows. Second, R-AQM assumes that the tenants are using per-packet ACK (one incoming packet triggers one ACK), which provides more precise control [14], [43]. Some tenants may need to modify their ACK mechanism, such as enabling Delayed ACK [42] so that an ACK can trigger more than one data packet at a time. Delayed ACK will not affect R-AQM's ability to resolve incast problem, only tenant performance. Hence data center operators should encourage tenants to use the default TCP stack to get better performance. In the worst case, this phenomenon can still be alleviated by adjusting K_{max} and K_{min}.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Symmetric Routing Dependency: R-AQM by design requires the ACKs to return on the same backward path as the data flow. This requirement can be easily met given the common deployment of ECMP routing in data centers [2], [14], [23]. At the same time, data center congestion often occurs on the last hop (Top of Rack (ToR) switches) [46], so we can also deploy R-AQM on last-hop switches to solve the Incast problems. Section VI-D and VII-D also show that R-AQM deployed on ToR switches only can also alleviate incast problems. However, using ECMP to ensure symmetric paths will limit the deployment of other flexible load balancing mechanisms [4], [17].

Non-ACK-clocking Cases: In a multi-tenant data center, the primary traffic is TCP/IP traffic [15], [27]. There is also RDMA traffic depending on business requirements [29]. Since R-AQM assumes ACK-clocking, it is hard work with non-TCP transport protocols. Currently, in real-world scenarios, RDMA traffic uses a separate priority. Therefore, R-AQM can be applied to priority queues using TCP, avoiding the influence of RDMA traffic and R-AQM. If the RDMA system uses the ACK-clocking version, R-AQM can also alleviate the Incast problem of RDMA. The experimental results of DCQCN+ win +R-AQM are also shown in Section VII. In addition, if the tenant uses rate-based or RTT-based CC, R-AQM will not make the congestion signal misjudged. Since ACK has the highest priority, R-AQM does not affect RTT measurement. Our experiment in Section VI-B also proved that R-AQM would also reduce RTT, which helps rate-based congestion control to judge congestion.

Elephant Flow and Mice Flow: R-AQM's motive is to address the TCP incast problem. Thus, like CC, R-AQM only controls elephant flows and has little control over mice flows, with only a few ACKs. However, R-AQM can reduce the FCT of mice flows. Most mice flows can end up in a single sending window, and R-AQM makes the forward queue length very small, so mice flows can pass quickly.

Maximum support senders and Memory usage: There is an upper limit to the number of R-AQM concurrency, which depends on the use of two pieces of memory (VIQ and the regular output queue). R-AQM controls the senders' sending action by shaping the ACKs, so it cannot reduce packet loss during the first control loop only by increasing the output queue memory. When traffic is stable, the number of senders depends on the VIQ size. Because the ACK is smaller than the data packet (60B v.s. 1460B), R-AQM can support more senders than traditional methods.

Malicious tenants: A malicious tenant can easily attack R-AQM in the following ways: a. Use an extremely large TCP initial window to cause packet drops; b. Send excessive ACK packets to impact other tenants. R-AQM does not focus on these security issues because these attacks can also impact data centers that are not using R-AQM. Generally, such problems can be alleviated by anomaly detection.

V. Implementation

R-AQM enabled Hardware Switch:

Ideally, the R-AQM's Token Bucket and the State Machine would be implemented in switch ASICs. We build a prototype of such a solution using the NetFPGA-SUME platform [64], a programmable hardware platform. It has four 10Gb/s Ethernet interfaces and a Xilinx Virtex-7 FPGA with QDRII+ and DDR3 memory resources.

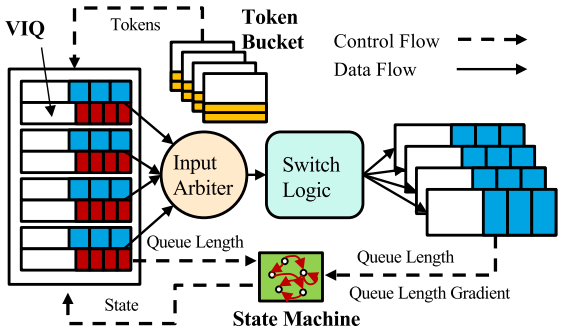


Fig. 7. R-AQM switch architecture on NetFPGA-SUME.

Figure 7 shows the top design of the R-AQM switch in NetPFGA. Packets enter one of the 10Gb/s interfaces and are stored in a regular input queue or VIQ. VIQ is allocated 36KB of SRAM, which separates from the regular input queue. ACKs are recognized while entering the input port. Pure ACKs enter VIQ, and others enter the regular input queue. The input arbiter takes packets from the input queues using a round-robin (RR) scheduling policy and feeds them to the L2 switching logic via a 256bit-wide 200MHz bus, which is fast enough to support more than 40Gb/s. The token bucket is used to trigger the VIQ sending action. The state machine determines the network congestion and adjusts the ACK draining rate through the queue length and its gradient. Pure ACK requires tokens and enters the input arbiter. Other packets can directly go to the input arbiter. After a conventional L2 forwarding decision is made, the packet reaches output queues.

Clearly, R-AQM implementation is quite simple, and the processing delay at the switch is very small. R-AQM does not operate normally for every packet, because it is only triggered to avoid packet dropping. Therefore, the additional processing delay at the switch is not introduced frequently. In addition, R-AQM introduces only a little resource consumption on switches. The R-AQM switch uses 58,610 LUTs (14% of the Virtex7's capacity), 29,370 FlipFlops (10%), and 1,470 blocks of RAM (45%). In comparison, the ECN-version FPGA switch uses 12%, 9%, and 40% respectively, so the complexity added by R-AQM is quite small.

We conclude that the implementation complexity, processing delay, and resource consumption of R-AQM are acceptable; thus, R-AQM can be built into commercial switches.

R-AQM Switch implementation in P4:

Figure 8 shows the implementation of R-AQM in P4. First, the token bucket is triggered based on events. Each packet triggers the accumulation calculation within the duration between the current packet and the previous successfully sent packet [28]. Second, the state decision machine needs to record the length of the queue and the change value of the length which is stored in two registers. We use an additional table (StateTokenRead table) and to read the state and the number of tokens from the register and save it as packet

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

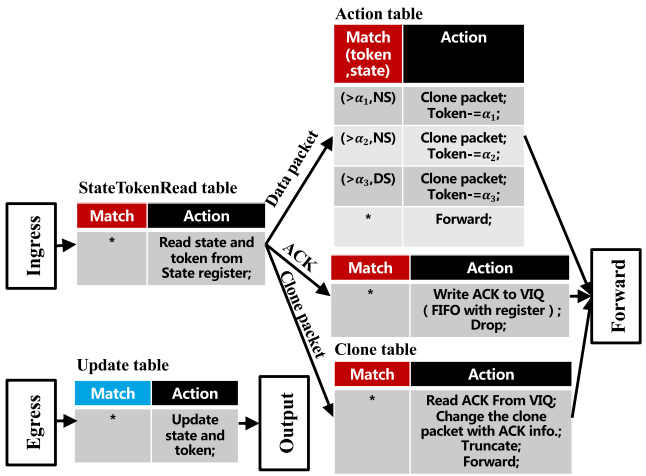


Fig. 8. R-AQM switch implementation on P4.

metadata [26]. Finally, VIQ is almost impossible to implement in P4, because the P4 operation primitive has nothing to do with buffer [28]. Here we use a slightly more complicated implementation. Each time an ACK enters the switch, we store the five-tuple of the ACK and the ACK Sequence Number into a FIFO which is made up of registers and drop the ACK packet (WriteVIQ table). If an ACK needs to be emitted, a data packet is cloned using the clone primitive (Action table). The cloned packet re-enters the switch via the recircle primitive (Clone table). The P4 switch will get an ACK information from VIQ and fill it into the cloned packet, using the truncate primitive to drop the payload, turns it into a real 'ACK'.

We have implemented R-AQM on BMv2,[1] however existing switches do not provide as much hardware resources and are not yet implemented on real P4 switches. In the future, we will be able to deploy R-AQM in a real environment when the relevant hardware resources are sufficient.

VI. Simulations of TCP

In this section, we conduct a simulation analysis of R-AQM performance with TCP using NS-3 [1]. Specifically, we evaluate three critical aspects of R-AQM as follows: (1) The flow scalability of R-AQM in terms of goodput, drop times, RTO times, latency and queuing. (2) The incast reaction details of R-AQM, concurrency, sensitivity and fairness analysis of R-AQM. (3) The effectiveness of R-AQM under all to all traffic.

A. Settings

The topology in the NS-3 simulations is a FatTree [3]. There are 16 Core switches, 20 Aggregation switches, 20 ToRs (Top-of-Rack switches) and 320 servers (16 in each rack), and each server has a single 10Gbps NIC connected to a single ToR. The capacity of each link between Core and Aggregation switches, Aggregation switches and ToRs are all 40Gbps. All links have a 1μs propagation delay, which gives a 12μs maximum base RTT. The switch per port's buffer is 300 packets (or about 400KB) derived from real device configurations.

[1]https://github.com/p4lang/behavioral-model

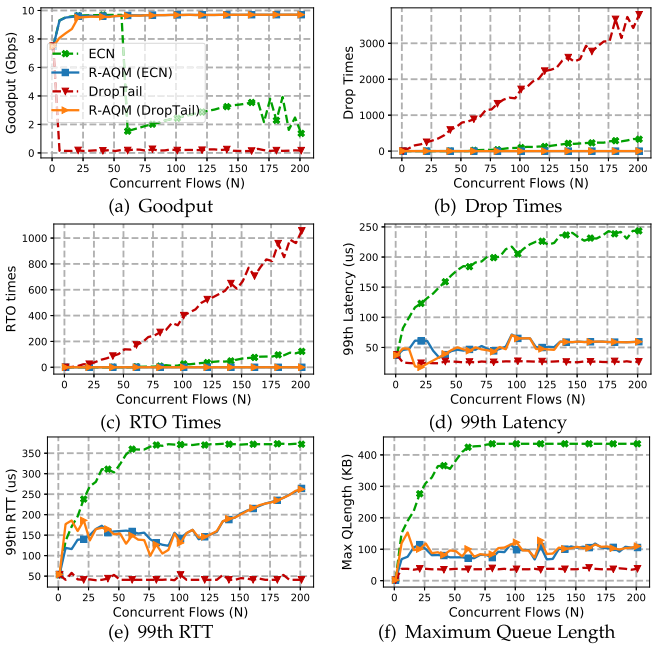


Fig. 9. Goodput, drop times, RTO times, latency, RTT and queue length with many concurrent flows.

We use two standard AQMs as baselines, ECN and droptail. We list the terminologies below:

- **ECN**: The corresponding senders' CC for ECN uses DCTCP.
- **R-AQM (ECN)**: The corresponding senders' CC for ECN uses DCTCP and the switch deploys R-AQM.
- **DropTail**: The corresponding senders' CC uses TCP NewReno.
- **R-AQM (DropTail/DropT)**: The corresponding senders' CC uses TCP NewReno and the switch deploys R-AQM.

The relevant parameters are set as follows: For R-AQM, we set K_{min} to 20 packets, K_{max} to 40 packets, T to 3μs, $\frac{\alpha_1}{n_1} = 1$, $\frac{\alpha_2}{n_2} = 0.5$ and $\frac{\alpha_3}{n_3} = 10$ in Algorithm 1. The switch VIQ is 500 packets, about 23KB. TCP is set to the default TCP RTO_{min} of 100 ms. For the ECN threshold, we scale the ECN threshold proportional to the link bandwidth. We set $ECNK_{min} = ECNK_{max} = 65$ packets according to [5].

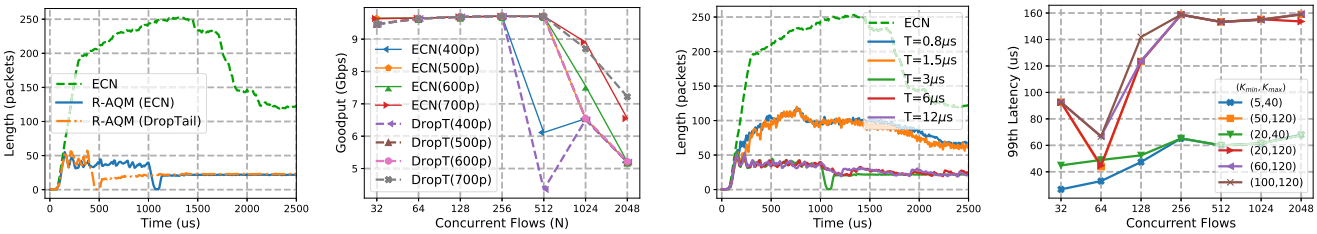
B. Incast

In incast, N hosts send a 3.2MB flow to a host. We vary the number of flows from 1 to 200. Figure 9 shows the results.

Goodput: First, we measured the goodput. In general, R-AQM can easily handle 200 concurrent connections without seeing any trend in performance degradation, while Droptail and ECN begin to downgrade when the numbers of connections exceed 5 and 60, respectively. When the number of connections is small, DCTCP with R-AQM shows little advantage over TCP with R-AQM in goodput (a few Gbps). For TCP, R-AQM only limits the sending window by limiting rwnd, so utilization decreases when the number of senders is small. However, R-AQM can continue to achieve near 9.8 Gbps goodput with an increasing number of connections.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a) Reaction Details Analysis of R-AQM (b) Sensitivity Analysis of Input Queue Length (c) Sensitivity Analysis of T (d) Sensitivity Analysis of K

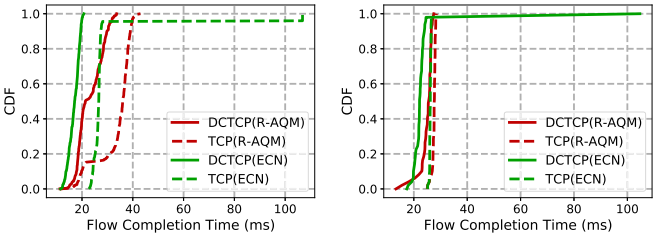
Fig. 10. Reaction details and sensitivity analysis of R-AQM.

Packet Drops and RTO: Second, we measured the drop times and RTO times. As shown in Figure 9(b) and 9(c), when the concurrency value is less than 200, R-AQM does not lose packets and trigger RTO. This also explains why R-AQM has no goodput loss. However, in the case of ultra-high concurrency, it still causes packet loss and RTO, which will be analyzed later.

Latency and RTT: Third, we measured forward one way latency and RTT. It can be seen that the latency can be significantly reduced with R-AQM. R-AQM achieves from 4.6× to 7.5× lower 99th latency compared to ECN. Droptail's latency is very low because its goodput is low, and the switch cannot be utilized effectively. As the number of concurrent connections increases, the 99th RTT increases but the 99th latency almost remains unchanged, indicating that R-AQM can keep the forward delay at a low value regardless of the number of concurrent flows. We also find that R-AQM delivers little impact on RTT. For example, when there are 200 concurrent connections, 99% of them are less than 0.3ms. Currently, many production data centers have reduced RTO_{min} to a low value (e.g., 10ms [5]). In Linux, the lowest possible RTO value is 5 jiffies (5ms) [55]. This suggests that R-AQM can work smoothly and will not result in issues like spurious timeouts and retransmissions in production datacenters with low RTO_{min}.

Queue Length: Fourth, we measure the buffer use of R-AQM in the last hop switch. Figure 9(f) shows that the buffer occupation of R-AQM is much lower than that of ECN. Moreover, as the number of senders increases, the buffer grows less significantly, while the ECN fills up when the sender is 70. In combination with Figure 9(d) and 9(e), with the sender number increasing, latency is unchanged while RTT is increasing, indicating that the reverse latency is increasing. This proves that the forward queue length is unchanged while the reverse VIQ is increasing all the time. The increase of ACK numbers makes little use of the buffer, so the utilization of the buffer can be reduced. For Droptail, a large number of packets are queued or lost in the first or middle hop, so the queue length is not the maximum one in the last hop.

Response Details of R-AQM: Fifth, we analyzed R-AQM response details and found why it could alleviate incast and reduce latency. Figure 10(a) shows the change in the queue length of the bottleneck switch overtime where N=36. R-AQM's queue starts to drop as it grows to 50 packets, while the ECN needs to grow to 250 before it can be adjusted. Fast adjustment of queue length can avoid more packet loss and reduce more RTO times. Moreover, R-AQM converges fast, which converges before 1000μs while ECN converges after 2000μs. Also, R-AQM negative feedback regulates queue length, maintaining a short queue length.



(a) 10-to-1 incast (b) 100-to-1 incast

Fig. 11. Fairness of different CCs.

Concurrency and Sensitivity Analysis of R-AQM: To explore the maximum number of connections that R-AQM can handle, we fix the total traffic volume and gradually increase the number of senders. We also repeat the simulation experiment using different parameters to assess the sensitivity of R-AQM to the setting of parameters. The results show that goodput is affected by the choice of the parameter VIQ length. Other parameters are not very sensitive to goodput. From Figure 10(b), we find that R-AQM can easily support more than 500 concurrent connections and sustain near 9Gbps goodput when facing 1000 senders. The goodput loss of 400p in R-AQM with 512 senders was due to VIQ dropping the ACK. After an ACK packet is lost, the sender cannot sense it and can only wait for RTO to resume sending. When there is more traffic, it will saturate more bandwidth, so the throughput will be higher. As shown in Figure 10(c) and 10(d), parameter T is sensitive to the reaction of the burst and K is sensitive to latency. We tested different T and K responses to burst, the results verify our analysis results in Section IV-F.

Fairness: Multi-tenant data centers might provide different CCs in use at the same time, so we also explore the fairness of hosts using different CCs. We observed the FCT distribution of different CCs' flow in the incast scenario. We discuss two scenarios, 10-to-1 incast and 100-to-1 incast, in which 50% of hosts use DCTCP, and the other 50% of hosts use TCP. In the 100-to-1 incast, the flow size is 3.2MB. In the 10-to-1 incast, to keep the time scale the same, we use 32MB flows. Figure 11 shows the results. In the 100-to-1 incast, R-AQM can reduce the gap between the two CCs. In the 10-to-1 incast, R-AQM can reduced the gap in 99th FCT. We believe this is because that R-AQM avoids RTO (abnormal state) effectively

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

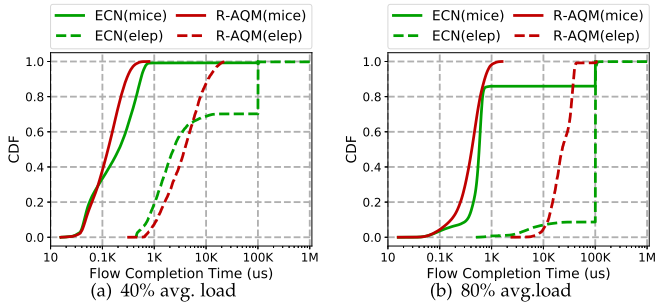


Fig. 12. Shuffle workload.

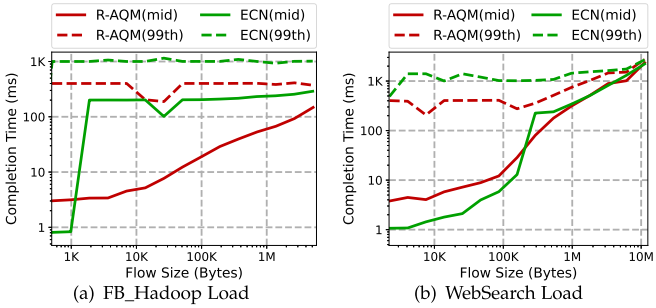


Fig. 13. Real workload.

and different CCs behave similarly in normal state. We also calculated Jane's Fairness Index (JFI) [35] for 10-to-1 incast and 100-to-1 incast, respectively. When the concurrency is 10, the index of R-AQM is 0.911, while that of ECN is 0.926. When the concurrency is 100, R-AQM is 0.981, and ECN is 0.982. From JFI value, R-AQM does not affect the fairness between TCP.

C. All to All

The all-to-all traffic patterns commonly happen in the shuffle step of MapReduce [16], which generates incast towards each host running a task. We simulate an all-to-all workload using NS-3. We select three machines on each rack, a total of 3×20.60 machines. Each machine sends a 500KB elephant flow and 50 5KB mice flows to the other 59 machines. So each machine sends and receives 3,540 (60×59) elephant flows and 177,000 ($60 \times 59 \times 50$) mice flows.

Figure 12 shows the CDF of the flow completion time under different loads. Because Droptail results are unsatisfactory, which causes many RTO, it is omitted here.

At the load of 40% scenarios, with R-AQM, the completion time of the mice flows is reduced. The 50th percentile FCTs are reduced from 208μs to 128μs and the 99th percentile FCTs are reduced from 768μs to 466μs in R-AQM. In ECN, 20% of the flow timeout, while in R-AQM there is no flow timeout. ECNs require queues long enough to trigger, so there is a lot of traffic resulting in RTOs due to packet loss, resulting in very long tail completion times. At 40% load, the R-AQM's mice-flow completion time was lower than that of the ECN, with no timeouts. This can prove that R-AQM can maintain a short forward queue, significantly improving the mice-flow application experience.

At the load of 80% scenarios, more than 15% of mice flows and 80% of elephant flows timeout in ECN, and only a small number of elephant flow timeouts in R-AQM. Even though the network load is very high, making ECN almost unusable, R-AQM guarantees that mice flows will not be RTO.

D. Real Workload

We use widely accepted and public available data center traffic traces, WebSearch [5] and FBHadoop [50] in real workload. Unlike traditional real workload scenario testing, we deployed only one ToR switch with R-AQM to validate the scenario of incremental deployment. To minimize the impact of congestion elsewhere on the measurement, we selected only one host per rack (out of a total of 20 racks) and sent data to the fixed nodes in a real flow size distribution. That is, in this experiment, we measured how much performance improvement we could achieve by simply replacing the most congested bottleneck switch.

Figure 13 shows the FCT under the two workloads. The 99th FCT of R-AQM was all the better than that of ECN. At 50th FCT, only ECN is good at the short flow part. The reason why ECN is good at short flows is that the ECN flows are RTO, so the link is idle and short flows can pass through quickly.

Through the above experiments, on the premise of avoiding RTO as much as possible, R-AQM can also provide a low latency for mice flows, which is enough to show that R-AQM can effectively alleviate the incast problem.

VII. Simulations of RDMA

In this section, we conduct a simulation analysis of R-AQM performance in RDMA with DCQCN [63] using NS-3 [1]. The purpose of R-AQM is to solve the TCP incast problem. In the RDMA network, the PFC mechanism can ensure that the network does not drop packets due to congestion, so it does not cause goodput loss. Therefore, we want to explore whether R-AQM can provide performance improvement in RDMA networks. R-AQM requires the sender CC to use the sending window, so our baseline scheme is DCQCN, DCQCN + win, and DCQCN + win + R-AQM.

A. Settings

The topology is the same as section VI. The switch is a shared memory switch. The dynamic threshold α is 1, and the memory is 2 MB.

The relevant parameters are set as follows: For R-AQM, we set K_{min} to 100 packets, K_{max} to 400 packets, T to 3μs, $\frac{\alpha_1}{n_1} = 1$, $\frac{\alpha_2}{n_2} = 0.5$ and $\frac{\alpha_3}{n_3} = 10$ in Algorithm 1. The switch VIQ is 500 packets, about 23KB. For the ECN threshold, we scale the ECN threshold proportional to the link bandwidth. We set $ECNK_{min}$ = 5 KB, $ECNK_{max}$ = 200 KB and $P_{max} = 1\%$ according to [63].

B. Incast

In incast, N hosts send a 3.2MB flow to a host. We vary the number of flows from 1 to 200. Figure 14 shows the results.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

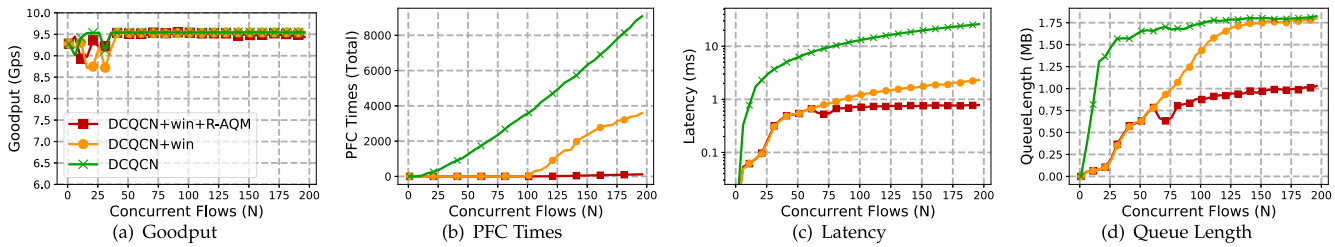


Fig. 14. Goodput, PFC times, RTO times, latency and queue length with many concurrent flows (DCQCN).

Goodput: First, we measured the goodput. Since RDMA does not drop packets due to congestion, there is no goodput loss, and both can easily handle 200 concurrent connections. Because the sending window will limit the rate of the sender, DCQCN + win will suffer a goodput loss when the sender is less than 50. R-AQM can mark ECN more accurately in the reverse path, so goodput loss is less than DCQCN + win.

PFC times: Second, we measured the PFC times. The PFC triggers only when the buffer of the switch exceeds the PFC threshold. PFC affects network performance, resulting in congestion spreading, PFC storms, and network deadlocks [24], [30], [43]. Therefore, avoiding PFC triggering is also an important performance indicator for RDMA networks. As shown in Figure 14(b), the number of times DCQCN triggers PFC is proportional to the number of concurrent flows. DCQCN + win can limit part of the PFC. However, when the number of senders exceeds 100, triggering the PFC is inevitable. R-AQM can always maintain a small number of PFC triggers.

Latency: Third, we measured forward one-way latency. It can be seen in Figure 14(c) that as the number of concurrent connections increases, the latency almost remains unchanged, indicating that R-AQM can keep the forward delay at a low value regardless of the number of concurrent flows.

Queue Length: Fourth, we measure the switch buffer use of R-AQM. Figure 14(d) shows that the buffer occupation of R-AQM is much lower than that of DCQCN and DCQCN + win. Moreover, as the number of senders increases, the buffer grows less significantly, while others fill up when the sender is 125.

Through the Incast experiment, we can prove that R-AQM can also take effect in the RDMA network. PFC times, latency, and queue length are all reduced with no goodput loss.

C. All to All

The all-to-all traffic patterns are also the same as Section VI. Figure 15 shows the CDF of the flow completion time under different loads.

At the load of 40% scenarios, with sending window, the completion time of the mice flows in DCQCN is reduced. The 50th percentile FCTs are reduced from 124 μs to 89 μs and the 99th percentile FCTs are reduced from 2030 μs to 900 μs with sending window. R-AQM can reduce the 99th FCT by a little bit compared to DCQCN + win. The 50th percentile FCT is the same as DCQCN + win, and the 99th FCT is 543 μs However, the R-AQM's elephant-flow completion time

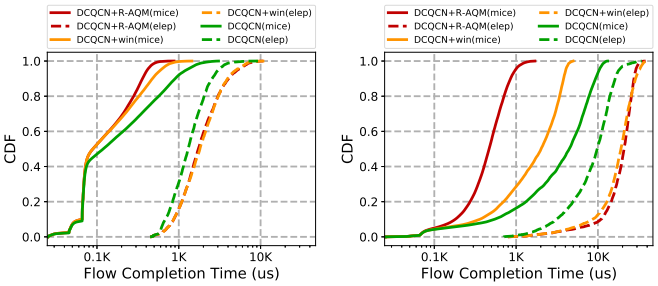


Fig. 15. Shuffle workload (DCQCN).

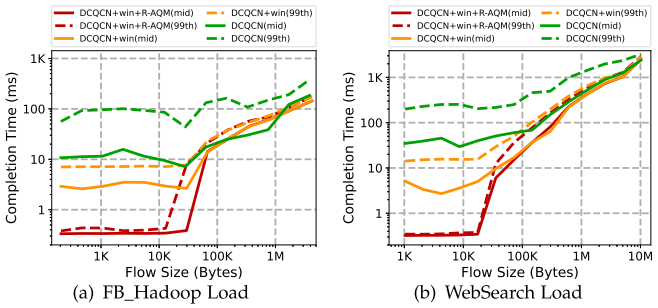


Fig. 16. Real workload (DCQCN).

is the same as DCQCN + win. This can prove that R-AQM can maintain a short forward queue, improving the mice-flow application experience. At the load of 80% scenarios, R-AQM can greatly reduce the FCT of mice flow. Through the above experiments, we find that R-AQM can also provide low latency for mice flows effectively in incast in RDMA networks.

D. Real Workload

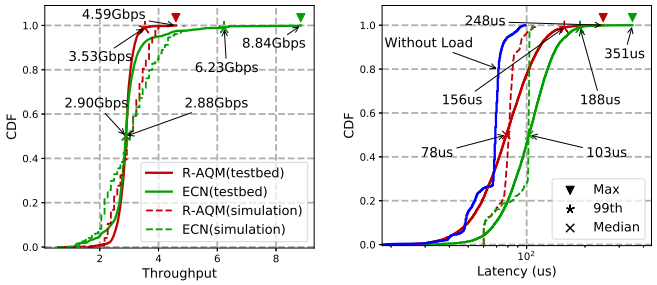
The settings are also the same as Section VI. Figure 16 shows the FCT of R-AQM with DCQCN under two workloads. The DCQCN+win+R-AQM is effective in reducing FCTs significantly for mice flows. Because the number of flows here is large, R-AQM can well control the queue length to give mice flows in a low-delay environment.

VIII. Testbed Experiments

In this section, we conduct a testbed experiment to validate the performance of R-AQM on the small-scale network. We verify R-AQM's ability to mitigate packet loss and low latency in a small-scale scenario through two typical workloads.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a) Throughput of CDF (b) Latecny of CDF

Fig. 17. Latency and throughput of CDF in the incast scenario.

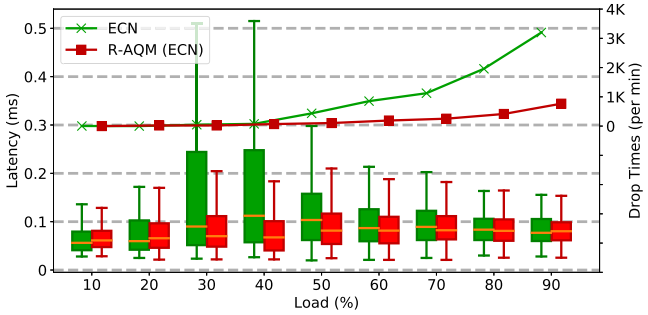


Fig. 18. Latency and drop times in the all to all scenario (testbed).

A. Settings

The topology of the testbed experiment mimics a small rack of the datacenter. The testbed includes one ToR NetFPGA switch and four servers connected via four 10Gbps links. Each server is equipped with a single Intel Xeno CPU and two dual-port Intel 82599 10G NICs. The CC algorithm at the hosts is DCTCP. TCP's results are similar and are therefore omitted later. On the NetFPGA switch, we also implemented the ECN version for comparison. The relevant parameters are set as follows: For R-AQM, we set K_{min} to 20 packets, K_{max} to 40 packets, T to 3μs, $\frac{\alpha_1}{n_1} = 1$, $\frac{\alpha_2}{n_2} = 0.5$ and $\frac{\alpha_3}{n_3} = 10$ in Algorithm 1. For the ECN threshold, R-AQM and pure ECN both set the 65 packets according to [5], and the switch output drop threshold is 300 packets.

B. Incast

We run a 3-to-1 incast: the frontend application on one host sends requests to another three servers. Upon receiving the request, each server replies with continuous elephant flow immediately. Meanwhile, a 1KB test mice flow is sent every second, which is used to measure the end to end latency. We calculate throughput and latency per second for each computer NIC port. At the same time, we apply the same scenario in the simulation to validate the testbed.

Figure 17 shows the throughput and latency of ECN and R-AQM in both simulation and testbed. We can see that the simulation experiment results are similar to the testbed experiment. For the 50th percentile throughput, R-AQM performs almost equally well compared to ECN, with the throughput of 2.88Gbps for ECN and 2.90Gbps for R-AQM. ECN has a long tail of 8.84Gpbs, which means an extremely unfair distribution of throughput could occur, with one port higher and the other two lower. The reason is that when a flow suffers from RTO caused by packet loss, other traffic takes up the bandwidth. In contrast, R-AQM shows a more even distribution of throughput and provides better fairness.

For latency, the 50th percentile and the worst case of R-AQM are 25μs and 103μs, respectively, both of which are lower than ECN. The 50th percentile of one-way delay for R-AQM is approximately 78μs, which is just slightly bigger (10μs) than the optimum transfer time in an idle network (67.7μs, labeled as "without load" in the figure). The improvement in latency by R-AQM is not outstanding due to the small scale of the testbed and the fact that the system kernel latency occupies a certain proportion of the overall latency. Despite that, it has been proved that R-AQM is effective in reducing latency.

C. All to All

To further demonstrate the functionality of R-AQM, we performed an all to all traffic pattern. Each host sends requests to the other three at the same time. The reply traffic's load varies from 10% to 90%. Meanwhile, a 1KB test mice flow is sent every second, which is used to measure the end to end latency. As shown in Figure 18, with the increase of load before 40%, the latency of ECN increases linearly, while R-AQM does not change. At 40%, R-AQM achieves 1.06× faster average latency than ECN, and the gap is more significant at the 99th percentile. This shows that R-AQM can effectively reduce end-to-end latency. Over 40%, the switch cannot handle more data, and packet loss begins, so the latency for both ECN and R-AQM starts to decrease. However, it is evident from the figure that R-AQM can keep the loss rate lower. Compared with ECN, R-AQM achieves up to 3.22 lower packet loss rates.

To conclude from our experiment, R-AQM effectively tames incast problems, decreases packet loss rate, reduces latency, and provides better fairness.

IX. Related Work

Many proposals address the incast congestion problem in data centers. Nevertheless, there has been relatively little effort to address incast in multi-tenant data centers. AC/DC TCP [27], vCC [15], DCTCP [5] and HSCC [2] can be used in multi-tenant data centers, but as discussed earlier, when incast arrived, even all senders' send windows are 1 MSS, the concurrency would not support much.

The most related works to R-AQM is PAC [10], which controls the sending rate of ACKs on the receiver to prevent incast congestion. We clarify the differences between R-AQM and PAC in two aspects. (1) PAC is an end-hosts mechanism, while R-AQM is a switch mechanism. (2) PAC cannot accurately predict the incast because the incast often occurs on the last-hop switch. R-AQM can directly obtain the queue length of the last-hop switch and take effect in advance.

In addition to the above works, a number of other data center transport designs have emerged, although their primary design space is not suitable for multi-tenant data centers.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Congestion control in private datacenter:

ICTCP [57] handles incast by adaptively adjusting the receiver side's receive window to throttle aggregate throughput. Tuning ECN [58] accelerates the delivery of congestion notifications using dequeue marking instead of traditional enqueue marking. D2TCP [54] adds deadline-awareness at the top of the DCTCP. It adjusts the congestion window to meet the deadline based on congestion conditions and deadline information. ExpressPass [14] uses credit packets for the preallocation of bandwidth to avoid congestion and to guarantee bounded queues.

DCQCN [63] and TIMELY [44] are proposed as the new end-to-end CC scheme designed for RDMA over Converged Ethernet v2 (RoCEv2) [8]. RoCEv2 enables lossless networks through Priority-based Flow Control (PFC), so there is no problem with large-scale RTO and goodput degradation. HPCC [43] also is a RoCEv2 CC that uses switch INT (in-network telemetry) to obtain the precise switch congestion state and calculates the remaining bandwidth. However, incast can also cause other congestion problems, such as PFC storm [43] and PFC deadlock [24], [30], resulting in high latency and unusable network.

All of the above approaches may face deployment issues in multi-tenant data centers and is out of the design scope of R-AQM. However, there is a theoretical possibility that R-AQM can be incrementally deployable with these approaches.

Switch-assisted mechanisms and CCs:

QCN [31] sends the quantized value of the congestion metric as feedback to senders, requiring fine parameters adjustment. PFC [32] allows switches to avoid buffer overflows by forcing the direct upstream switch or NIC to suspend data transfers. XCP [36] and RCP [19] use explicit feedback to measure the extent of congestion. D3 [56] achieves explicit rate control based on deadline information to guarantee deadlines. HULL [7] uses phantom queues to simulate a network at less than 100% utilization and relies on ECN to deliver congestion information. CP [13] and NDP [26] realizes fast packet loss notification by cutting packet payload in the switch and sending packet header back to the sender quickly.

These approaches share the same idea that switches cooperate with congestion control through congestion signals (packet loss, ECN, RTT, INT, etc.). However, most of them require an intrusive modification to the protocol stack. In contrast, R-AQM is an incremental and transparent design for end-systems, helping to alleviate the incast problem and gaining marginal benefit in terms of concurrency.

X. Conclusion

In this paper, we present R-AQM, a transparent reverse ACK active queue management design for multi-tenant data centers to tame the TCP incast problem through active ACK control. The basic principle of R-AQM is to reduce the nontrivial forward queue delay by introducing a trivial backward ACK delay. The critical design idea behind R-AQM is to proactively intercept the ACK in the switch and release it at a moderate rate to prevent too fast new packets from overwhelming the switch. R-AQM set up VIQs to buffer the ACK and use the Token Bucket to shape the flow. R-AQM also uses queue length and its gradient to judge burst and congestion. Our extensive simulations and experiments have shown that R-AQM can enhance existing CC solutions by supporting 16 times more senders and reducing forward queue delay by 4.6 times. One of the limitations of R-AQM is that it can only cooperate with window-based ACK-clocking protocols.

Acknowledgment

This work is not possible without the efforts of Xinping Chen and Shengjun Chen. The authors are grateful for conversations with and feedback from Kun Tan, Binzhang Fu, and Jincheng Bao. They also thank the reviewers for their valuable comments.

References

[1] (2019). *Network Simulator 3*. [Online]. Available: https://www.nsnam.org/

[2] A. M. Abdelmoniem and B. Bensaou, "Hysteresis-based active queue management for TCP traffic in data centers," in *Proc. IEEE*, Apr. 2019, pp. 1621–1629.

[3] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, 2008.

[4] M. Alizadeh *et al.*, "CONGA: Distributed congestion-aware load balancing for datacenters," in *Proc. ACM Conf. SIGCOMM*, Aug. 2014, pp. 503–514.

[5] M. Alizadeh *et al.*, "Data center TCP (DCTCP)," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 63–74, Oct. 2011.

[6] M. Alizadeh, A. Javanmard, and B. Prabhakar, "Analysis of DCTCP: Stability, convergence, and fairness," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 39, no. 1, pp. 73–84, 2011.

[7] M. Alizadeh *et al.*, "Less is more: Trading a little bandwidth for ultra-low latency in the data center," in *Proc. USENIX NSDI*, 2012, pp. 253–266.

[8] InfiniBand Trade Association. (Sep. 2014). *RoCE v2*. [Online]. Available: https://cw.infinibandta.org/document/dl/7781

[9] S. Athuraliya, S. H. Low, V. H. Li, and Q. Yin, "REM: Active queue management," *IEEE Netw.*, vol. 15, no. 3, pp. 48–53, May 2001.

[10] W. Bai, K. Chen, H. Wu, W. Lan, and Y. Zhao, "PAC: Taming TCP incast congestion using proactive ACK control," in *Proc. IEEE ICNP*, Oct. 2014, pp. 385–396.

[11] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson, "BBR: Congestion-based congestion control," *ACM Queue*, vol. 14, no. 5, pp. 20–53, 2016.

[12] W. Chen, F. Ren, J. Xie, C. Lin, K. Yin, and F. Baker, "Comprehensive understanding of TCP incast problem," in *Proc. IEEE INFOCOM*, Apr. 2015, pp. 1688–1696.

[13] P. Cheng, F. Ren, R. Shu, and C. Lin, "Catch the whole lot in an action: Rapid precise packet loss notification in data center," in *Proc. USENIX NSDI*, 2014, pp. 17–28.

[14] I. Cho, K. Jang, and D. Han, "Credit-scheduled delay-bounded congestion control for datacenters," in *Proc. ACM SIGCOMM*, Aug. 2017, pp. 239–252.

[15] B. Cronkite-Ratcliff *et al.*, "Virtualized congestion control," in *Proc. ACM SIGCOMM*, Aug. 2016, pp. 230–243.

[16] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[17] A. Dixit, P. Prakash, Y. C. Hu, and R. R. Kompella, "On the impact of packet spraying in data center networks," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 2130–2138.

[18] X. Du, K. Xu, T. Li, K. Zheng, S. Fu, and M. Shen, "Traffic control for data center network: State of the art and future research," (in Chinese) *Chin. J. Comput.*, vol. 43, no. 17, pp. 1–23, 2020.

[19] N. Dukkipati, *Rate Control Protocol (RCP): Congestion Control to Make Flows Complete Quickly*. Princeton, NJ, USA: Citeseer, 2008.

[20] S. Floyd, "TCP and explicit congestion notification," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 24, no. 5, pp. 8–23, Oct. 1994.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[21] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Trans. Netw.*, vol. 1, no. 4, pp. 397–413, Aug. 1993.

[22] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google file system," in *Proc. 19th ACM Symp. Operating Syst. Princ.*, 2003, pp. 29–43.

[23] A. Greenberg *et al.*, "VL2: A scalable and flexible data center network," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 51–62, 2009.

[24] C. Guo *et al.*, "RDMA over commodity Ethernet at scale," in *Proc. ACM SIGCOMM*, Aug. 2016, pp. 202–215.

[25] A. Gurtov, T. Henderson, S. Floyd, and Y. Nishida, *The NewReno Modification to TCP's Fast Recovery Algorithm*, document RFC 6582, Apr. 2012. [Online]. Available: https://rfc-editor.org/rfc/rfc6582.txt

[26] M. Handley *et al.*, "Re-architecting datacenter networks and stacks for low latency and high performance," in *Proc. ACM SIGCOMM*, Aug. 2017, pp. 29–42.

[27] K. He *et al.*, "AC/DC TCP: Virtual congestion control enforcement for datacenter networks," in *Proc. ACM SIGCOMM*, Aug. 2016, pp. 244–257.

[28] Y. He and W. Wu, "Fully functional rate limiter design on programmable hardware switches," in *Proc. ACM SIGCOMM Conf. Posters*, 2019, pp. 159–160.

[29] Z. He *et al.*, "MASQ: RDMA for virtual private cloud," in *Proc. Annu. Conf. ACM Special Interest Group Data Commun. Appl., Technol., Architectures, Protocols Comput. Commun.*, 2020, pp. 1–14.

[30] S. Hu *et al.*, "Tagger: Practical PFC deadlock prevention in data center networks," in *Proc. ACM CoNEXT*, 2017, pp. 451–463.

[31] *Congestion Notification*, Standard 802.11Qau, 2010.

[32] *Priority Based Flow Control*, Standard 802.11Qbb, 2011.

[33] V. Jacobson, "Congestion avoidance and control," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 18, no. 4, pp. 314–329, 1988.

[34] K. Jacobsson, L. L. H. Andrew, A. Tang, K. H. Johansson, H. Hjalmarsson, and S. H. Low, "ACK-clocking dynamics: Modelling the interaction between windows and the network," in *Proc. IEEE INFOCOM*, Apr. 2008, pp. 2146–2152.

[35] R. K. Jain *et al.*, "A quantitative measure of fairness and discrimination," Eastern Res. Lab., Digit. Equip. Corp., Hudson, MA, USA, Tech. Rep. DEC-TR-301, 1984, vol. 21.

[36] D. Katabi, M. Handley, and C. Rohrs, "Congestion control for high bandwidth-delay product networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 32, no. 4, pp. 89–102, Oct. 2002.

[37] T. Koponen *et al.*, "Network virtualization in multi-tenant datacenters," in *Proc. USENIX NSDI*, 2014, pp. 203–216.

[38] L. Li *et al.*, "A measurement study on multi-path TCP with multiple cellular carriers on high speed rails," in *Proc. ACM SIGCOMM*, Aug. 2018, pp. 161–175.

[39] T. Li, K. Wang, K. Xu, K. Yang, C. S. Magurawalage, and H. Wang, "Communication and computation cooperation in cloud radio access network with mobile edge computing," *CCF Trans. Netw.*, vol. 2, no. 1, pp. 43–56, Jun. 2019.

[40] T. Li, K. Zheng, and K. Xu, "Acknowledgment on demand for transport control," *IEEE Internet Comput.*, vol. 25, no. 2, pp. 109–115, Mar. 2021.

[41] T. Li *et al.*, "TACK: Improving wireless transport performance by taming acknowledgments," in *Proc. ACM SIGCOMM*, Jul. 2020, pp. 15–30.

[42] T. Li *et al.*, "Revisiting acknowledgment mechanism for transport control: Modeling, analysis, and implementation," *IEEE/ACM Trans. Netw.*, vol. 29, no. 6, pp. 2678–2692, Dec. 2021.

[43] Y. Li *et al.*, "HPCC: High precision congestion control," in *Proc. ACM SIGCOMM*, Aug. 2019, pp. 44–58.

[44] R. Mittal *et al.*, "TIMELY: RTT-based congestion control for the datacenter," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 537–550, 2015.

[45] R. Mittal *et al.*, "Revisiting network support for RDMA," in *Proc. ACM SIGCOMM*, Aug. 2018, pp. 313–326.

[46] B. Montazeri, Y. Li, M. Alizadeh, and J. Ousterhout, "Homa: A receiver-driven low-latency transport protocol using network priorities," in *Proc. ACM SIGCOMM*, vol. 2018, pp. 221–235.

[47] R. Nishtala *et al.*, "Scaling memcache at Facebook," in *Proc. USENIX NSDI*, 2013, pp. 385–398.

[48] K. Qian, W. Cheng, T. Zhang, and F. Ren, "Gentle flow control: Avoiding deadlock in lossless networks," in *Proc. ACM SIGCOMM*, Aug. 2019, pp. 75–89.

[49] I. Rhee, L. Xu, S. Ha, A. Zimmermann, L. Eggert, and R. Scheffenegger, *CUBIC for Fast Long-Distance Networks*, document RFC 8312, Feb. 2018. [Online]. Available: https://rfc-editor.org/rfc/rfc8312.txt

[50] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, "Inside the social network's (datacenter) network," in *Proc. ACM Conf. Special Interest Group Data Commun.*, Aug. 2015, pp. 123–137.

[51] D. Shan, W. Jiang, and F. Ren, "Analyzing and enhancing dynamic threshold policy of data center switches," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 9, pp. 2454–2470, Sep. 2017.

[52] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop distributed file system," in *Proc. IEEE MSST*, May 2010, pp. 1–10.

[53] S. Huang, M. Wang, and Y. Cui, "Traffic-aware buffer management in shared memory switches," in *Proc. IEEE INFOCOM*, May 2021, pp. 1–10.

[54] B. Vamanan, J. Hasan, and T. N. Vijaykumar, "Deadline-aware datacenter TCP (D2TCP)," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 115–126, Sep. 2012.

[55] V. Vasudevan *et al.*, "Safe and effective fine-grained TCP retransmissions for datacenter communication," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 303–314, 2009.

[56] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowtron, "Better never than late: Meeting deadlines in datacenter networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 50–61, Aug. 2011.

[57] H. Wu, Z. Feng, C. Guo, and Y. Zhang, "ICTCP: Incast congestion control for TCP in data-center networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, pp. 345–358, Apr. 2013.

[58] H. Wu, J. Ju, G. Lu, C. Guo, Y. Xiong, and Y. Zhang, "Tuning ECN for data center networks," in *Proc. ACM CoNEXT*, 2012, pp. 25–36.

[59] K. Xu *et al.*, "Modeling, analysis, and implementation of universal acceleration platform across online video sharing sites," *IEEE Trans. Serv. Comput.*, vol. 11, no. 3, pp. 534–548, May/Jun. 2018.

[60] K. Xu, L. Lv, T. Li, M. Shen, H. Wang, and K. Yang, "Minimizing tardiness for data-intensive applications in heterogeneous systems: A matching theory perspective," *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 1, pp. 144–158, Jan. 2020.

[61] L. Xu *et al.*, "ABQ: Active buffer queueing in datacenters," *IEEE Netw.*, vol. 34, no. 2, pp. 232–237, Mar. 2020.

[62] J. Zhang, F. Ren, and C. Lin, "Modeling and understanding TCP incast in data center networks," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 1377–1385.

[63] Y. Zhu *et al.*, "Congestion control for large-scale RDMA deployments," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 523–536, 2015.

[64] N. Zilberman, Y. Audzevich, G. A. Covington, and A. W. Moore, "NetFPGA SUME: Toward 100 Gbps as research commodity," *IEEE Micro*, vol. 34, no. 5, pp. 32–41, Sep./Oct. 2014.

Xinle Du received the B.E. degree from the Department of Computer Science and Technology, Xidian University, Xi'an, China, in 2014. He is currently pursuing the Ph.D. degree with Tsinghua University. His research interests include data-driven networks, data center network transport protocol, and AQM.

Ke Xu (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He is currently as a Full Professor with the Department of Computer Science and Technology, Tsinghua University. He has published more than 200 technical papers and holds 11 U.S. patents in the research areas of next-generation Internet, blockchain systems, the Internet of Things, and network security. He is a member of ACM. He has guest-edited several special issues in IEEE and Springer journals, and also served as the Steering Committee Chair for IEEE/ACM IWQoS. He is an Editor of IEEE INTERNET OF THINGS JOURNAL.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Lei Xu received the B.E. degree from the Department of Computer Science and Technology, Beijing Institute of Technology, China, in 2006, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, China, in 2018. He held a Visiting Scholar with the School of Computer Science and Electronic Engineering, University of Essex, U.K., in 2014. In 2018, he joined HiSilicon Technologies Company Ltd. focusing on DC networking, congestion control, and design of switch chip.

Kai Zheng (Senior Member, IEEE) is currently the Director of the Computer Network and Protocol Research Laboratory, Huawei Technologies. His research interests covered architectures and protocols for the next generation networks, such as 5G/IoT networks, cloud oriented data center networks, RDMA networks, and real-time multimedia networks.

Meng Shen (Member, IEEE) received the B.Eng. degree in computer science from Shandong University, Jinan, China, in 2009, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2014. He is currently a Professor at the Beijing Institute of Technology, Beijing. He has authored over 50 papers in top-level journals and conferences, such as ACM SIGCOMM, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC), and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (TIFS). His research interests include data privacy and security, blockchain applications, and encrypted traffic classification. He received the Best Paper Runner-Up Award from IEEE IPCCC 2014 and IEEE/ACM IWQoS 2020. He was selected by the Beijing Nova Program 2020 and the winner of the ACM SIGCOMM China Rising Star Award 2019. He has guest edited Special Issues on Emerging Technologies for Data Security and Privacy in IEEE NETWORK and IEEE INTERNET OF THINGS JOURNAL.

Bo Wu received the B.E. degree from the School of Software, Shandong University, China, in 2014, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, China, in 2019. He acted as a Visiting Scholar with the Department of Computer Science, Northwestern University, USA, from 2017 to 2018. He is currently as a Research Fellow at Tencent Technologies. His research interests include Internet architecture, network security, and machine learning.

Tong Li (Member, IEEE) received the B.E. degree from the School of Computer Science, Wuhan University, China, in 2012, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, China, in 2017. He was a Visiting Scholar with the School of Computer Science and Electronic Engineering, University of Essex, U.K., in 2014 and 2016. He worked as a Chief Engineer at Huawei before 2022, and currently he serves as an Associate Professor at the Renmin University of China. His research interests include networking, distributed systems, and big data.

Authorized licensed use limited to: Renmin University. Downloaded on August 24,2022 at 02:57:48 UTC from IEEE Xplore. Restrictions apply.