RCC: Enabling Receiver-Driven RDMA Congestion Control With Congestion Divide-and-Conquer in Datacenter Networks

Jiao Zhang, *Senior Member, IEEE*, Xiaolong Zhong, *Graduate Student Member, IEEE*, Zirui Wan, Yu Tian, Tian Pan, *Senior Member, IEEE*, and Tao Huang, *Senior Member, IEEE*

***Abstract*— The development of datacenter applications leads to the need for end-to-end communication with microsecond latency. As a result, RDMA is becoming prevalent in datacenter networks to mitigate the latency caused by the slow processing speed of the traditional software network stack. However, existing RDMA congestion control mechanisms are either far from optimal in simultaneously achieving high throughput and low latency or in need of additional in-network function support. In this paper, by leveraging the observation that most congestion occurs at the last hop in datacenter networks, we propose RCC, a receiver-driven rapid congestion control mechanism for RDMA networks that combines *explicit assignment* and *iterative window adjustment*. Firstly, we propose a network congestion distinguish method to classify congestions into two types, last-hop congestion and in-network congestion. Then, an Explicit Window Assignment mechanism is proposed to solve the last-hop congestion, which enables senders to converge to a proper sending rate in one-RTT. For in-network congestion, a PID-based iterative delay-based window adjustment scheme is proposed to achieve fast convergence and near-zero queuing latency. RCC does not need additional in-network support and is friendly to hardware implementation. In our evaluation, the overall average FCT (Flow Completion Time) of RCC is 4∼79% better than Homa, ExpressPass, DCQCN, TIMELY, and HPCC.**

***Index Terms*— Datacenter, RDMA, congestion control, receiver-driven, PI controller.**

Manuscript received 22 December 2021; revised 15 May 2022; accepted 18 June 2022; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor K. Chen. Date of publication 8 July 2022; date of current version 16 February 2023. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61872401 and Grant 62132022 and in part by the Fok Ying Tung Education Foundation under Grant 171059. This paper is an extended version of the work that first appeared with title "Receiver-Driven RDMA Congestion Control by Differentiating Congestion Types in Datacenter Networks" at IEEE ICNP 2021 [DOI: 10.1109/ICNP52444.2021.9651938]. *(Corresponding author: Tao Huang.)*

Jiao Zhang, Tian Pan, and Tao Huang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the Purple Mountain Laboratories, Nanjing 211111, China (e-mail: jiaozhang@bupt.edu.cn; pan@bupt.edu.cn; htao@bupt.edu.cn).

Xiaolong Zhong and Zirui Wan are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: xlzhong@bupt.edu.cn; wanzr@bupt.edu.cn).

Yu Tian is with the School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: tianyu2992@bupt.edu.cn).

Digital Object Identifier 10.1109/TNET.2022.3185105

I. INTRODUCTION

DATACENTERS are increasingly dominating the market for different types of high-end computing and distributed data storage services [2], [3]. These workloads put enormous pressure on datacenter networks to deliver ever faster throughput and extremely low latency at a low cost. More specifically, with the tendency of deploying high I/O speed storage media in datacenters, such as NVMe (Non-Volatile Memory express), the storage speed and access latency are significantly improved [3]. Therefore, datacenters become a good fit for applications with great demand for computation and storage capacity. However, to take full advantage of the distributed and high-speed computation and storage resources in datacenters, the networking stack requires to guarantee high throughput and microsecond latency communications among distributed nodes. Otherwise, the communication latency will become the bottleneck of the applications [4], [5].

Unfortunately, the traditional TCP/IP network stack incurs a lot of overhead [6]. CPU spends much time managing data transfers for write-intensive workloads, reducing the overall performance of these tasks. To solve this issue, the RDMA (Remote Direct Memory Access) technique is becoming widely used in datacenter networks [4]–[6]. The direct connection of RDMA NICs reduces the involvement of the CPU during data transmission. Meanwhile, combined with fast storage like NVMe, the RDMA can cut end-to-end communication latencies down from milliseconds to microseconds.

However, deploying RDMA in datacenters poses great challenges on datacenter networking. Limited by the hardware resources in NICs, current RDMA congestion control relies on a simple go-back-N method to recover lost packets. Once the loss rate becomes higher, the performance of RDMA connections will dramatically deteriorate. Thus, PFC (Priority Flow Control) is used to guarantee in-network losslessness. However, PFC potentially brings fatal problems like PFC deadlock and PFC pause frame storm [6], [7]. Therefore, much attempt has been conducted to design RDMA-dedicated congestion control mechanisms to avoid packet dropping.

The goal of congestion control mechanisms is to allocate the bandwidth of congested links efficiently. The key challenge lies in that end-hosts can not obtain accurate information on network conditions easily. Most of the existing RDMA

congestion control mechanisms use various metrics, such as ECN mark and RTT, to detect network conditions at the sender side [5], [8], [9]. Then iterative window adjustment schemes are proposed to solve network congestion. HPCC [4] suggests using INT (In-band Network Telemetry) to obtain accurate information on network conditions and then precisely controls the congestion window at the sender side to achieve faster convergence and lower latency. It can not be deployed if INT support is absent.

In this paper, we instead ask, *is it essential to accurately measure in-network information for congestion control mechanisms in all cases*? Most congestion happens at the last hop due to the many-to-one traffic pattern in datacenter networks, even in over-subscribed datacenter networks [10]–[13]. We call this kind of congestion **last-hop congestion**. The other congestion, which happens at other places, is referred to as **in-network congestion**. A study of Google's production datacenters reveals that *the predominant source of congestion, accounting for **62.8%**, comes from the last hop in datacenter networks* [13]. Fortunately, receivers can easily obtain the last-hop congestion information. Therefore, *there is potential for designing a simple and efficient mechanism to solve the major last-hop congestion without obtaining in-network congestion information, while the remaining small part of in-network congestion can be further addressed by another more complicated scheme.*

Enlightened by the above investigation, we propose a novel RDMA congestion control mechanism, RCC, that *combines explicit assignment and iterative window adjustment* at the receiver side. Firstly, we propose a network congestion differentiation method to detect whether the last-hop congestion happens. Inspired by the fast recovery mechanism in TCP, we use n consecutive measured RTT values to infer whether network congestion occurs or not. Then the last-hop average throughput is used to further distinguish whether last-hop congestion happens. For *last-hop congestion*, we propose an Explicit Window Assignment scheme to adjust the sending rate according to the connection number perceived at the receiver side and piggyback the sending window through ACK packets to senders. Besides, we combine per-ACK window adjustment and packet pacing to avoid instantaneous large queuing caused by Incast flows. In this way, the last-hop congestion that takes the majority of network congestion in datacenters can be solved in one RTT. For *in-network congestion*, we design a new iterative window adjustment scheme based on the PID (Proportional Integral Derivative) control theory. By combining the proportional and derivative terms, RCC can converge to a unique fixed point and achieve high utilization with near-zero queuing latency. Besides, RCC sets the upper bound of sending rate through the Explicit Window Assignment mechanism for each flow. In this way, RCC can avoid overlarge PID-based iterative window adjustment results.

The main advantages of RCC include: *1) it can achieve high throughput, near-zero queuing latency, fast convergence, and fairness simultaneously, 2) it does not need additional in-network features and thus can be readily deployed with traditional commodity switches, 3) it requires only a small amount of extra memory for each RDMA connection, which makes it friendly to hardware implementation.*

We analyze the stability and convergence of RCC based on a mathematical model and the PID control theory. Then we use a simulation-based phase margin and loop bandwidth analysis to show how to configure parameters in RCC to ensure stability and convergence.

Furthermore, we evaluate the performance of RCC both in testbed and NS3 simulator [14] by conducting micro-benchmark experiments as well as large-scale simulations using realistic workloads from Google and Facebook datacenters. We show RCC outperforms Homa, ExpressPass, DCQCN, TIMELY, and HPCC in terms of mean and tail flow completion time, convergence rate, fairness, and queuing latency. Large-scale simulations show that RCC achieves 55% lower average FCT and 79% lower 99^{th} percentile FCT than TIMELY and DCQCN for typical datacenter topology and workload settings. Compared with HPCC, RCC is fairer and achieves better performance in multiple scenarios.

In summary, our key contributions are:

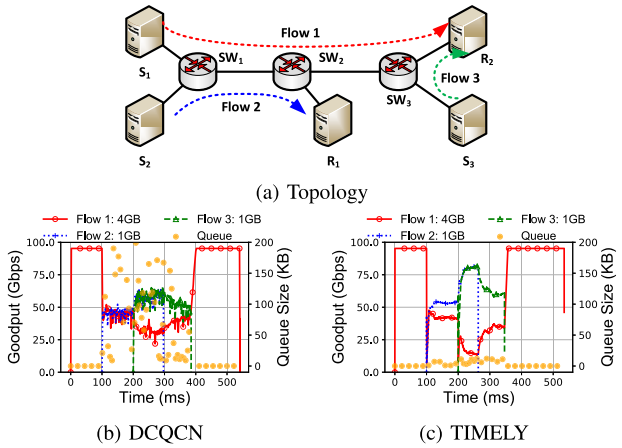
- We leverage the characteristic that most congestion happens at the last hop to design RCC, a high-performance transport for RDMA in datacenter networks;
- We propose an Explicit Window Adjustment mechanism to fairly assign the last-hop bandwidth to senders in one-RTT for last-hop congestion. And we design a PID-based window adjustment mechanism to simultaneously achieve fairness and a guaranteed steady-state latency for in-network congestion. Besides, per-ACK window adjustment and packet pacing are combined to mitigate instantaneous large queuing latency;
- We theoretically analyze RCC on its stability and convergence and show how to tune parameters of RCC under various network conditions;
- We evaluate RCC with both DPDK implementation and large-scale simulations in comparison with state-of-the-art RDMA and receiver-driven schemes. Results show that RCC achieves $4 \sim 79\%$ lower overall average FCT than Homa, ExpressPass, DCQCN, TIMELY, and HPCC.

II. Background and Motivation

A. RDMA in Ethernet

Traditional TCP suffers from high CPU overhead and large latency [15]. By offloading the transport layer function to the hardware chip, RDMA is able to access (i.e., read from or write to) memory on a remote machine without interrupting the processing of the CPU(s) on that system. RDMA was previously used in lossless InfiniBand networks. To use RDMA in Ethernet and IP networks, RoCE [16] is proposed. RoCE follows the original design of RDMA for lossless networks, using PFC [17] to avoid packet loss in Ethernet and using a go-back retransmission mechanism to recover lost packets.

PFC is a hop-by-hop flow control mechanism to prevent buffer overflow on Ethernet switches and end-host NICs. It works in the queue granularity and sends PAUSE/RESUME frames from downstream devices to notify upstream devices to pause/resume sending packets. Because of a coarse-grained



(a) Topology

(b) DCQCN (c) TIMELY

Fig. 1. Performance of DCQCN and TIMELY in a multi-bottleneck topology. Flow 1, 2, 3 starts at 0, 100, 200 ms, respectively.

queue-level operation, PFC possibly leads to poor performance for individual flows, such as unfairness, flow transmission and head-of-line blocking. Even worse, unexpected interaction between PFC and Ethernet packets flooding possibly breaks the up-down routing and could lead to occasional deadlocks.

Go-Back Retransmission. The original design of RDMA in Ethernet employs go-back-0 retransmission to handle occasional packet drops, which suffers from the live-lock problem. To address this issue, a modified go-back-N retransmission is employed. Go-back-N scheme solves the live-lock problem but still wastes time and bandwidth for sending redundant packets, potentially increasing the probability of congestion.

B. Defects of State-of-the-Art Congestion Control Solutions

In order to reduce the side effects of the above go-back mechanisms, flow-based RDMA congestion control solutions like DCQCN, TIMELY and HPCC, have been proposed.

DCQCN is an end-to-end rate-based congestion control mechanism [5] proposed for RoCEv2. It achieves high link utilization by fast rate increase similar to QCN and fairness through DCTCP-like fine-grained rate control [18].

TIMELY is an RTT-based and rate-based RDMA congestion control mechanism [9], which iteratively adjusts the sending rate based on the RTT gradients.

HPCC [4] leverages precise link load information carried by INT probes supported by switches to calculate the appropriate sending window for connections.

However, limited by the intrinsic long end-to-end control loop, though proven to be effective or even widely deployed, the algorithms above easily fall short under specific scenarios. We conduct a simulation to illustrate defects of the state-of-the-art with three competitive flows traversing a typical multi-bottleneck topology, in which congestions exist both at the network edges and inside. As shown in Fig. 1(b) and 1(c), DCQCN fails to give a fast and precise response to congestions, thus leading to drastic queue oscillation and under-utilization; TIMELY cannot simultaneously achieve fairness and guaranteed queuing delay.[1]

[1]As the best of these three schemes, we elaborate the performance degradation of HPCC under specific scenarios in § V-B.

C. Most Congestion Happens at the Network Edge

Topology. Datacenter network topology plays a vital role in determining the communication bandwidth and latency between each pair of nodes. The tree-based hierarchical topology with two or three tiers according to the network scale is widely used in practice [19]–[22]. These topologies generally have sufficient cross-sectional bandwidth and the core network will not become a bottleneck [10].

Communication Pattern. Datacenters employ the scale-out method to support large-scale applications. Generally, applications are hosted in tens of hundreds of servers [26], and there are frequent communications between different pairs of nodes to support various tasks from users. In fact, due to the sufficient cross-sectional bandwidth and the widely existed many-to-one/many-to-many communication patterns [11], *most congestion in full-bisectional/over-subscribed datacenters happens at the network edge (or more specifically, the last hop)* [10], [23], [24]. For example, some popular applications (key-value stores [25], data mining [26], parameter servers [27] used in distributed machine learning frameworks, etc.) often generate a number of scatter-gather [11], [28] and batch computing tasks [29], causing the many-to-one communication pattern. A study of Google's production datacenters reveals that *the predominant source of congestion, accounting for **62.8%**, comes from the last hop in datacenter networks* [13].

D. Brief Summary

Existing state-of-the-art RDMA congestion control solutions follow a similar rationale: senders iteratively adjust sending rates of flows according to the network congestion signals. There is still room for improvement from the perspective of fairness, convergence rate, and end-to-end latency. It is potential to design a more concise and efficient RDMA congestion control algorithm by leveraging the special characteristic that congestion often happens at the last hop in datacenters.

III. Design

A. Basic Idea and Challenges

The key idea of RCC is to leverage the observation that most of the congestion happens at the receiver edge and allocate bandwidth to connections according to different kinds of congestion types. We classify network congestion into two types: **C1** happens at the last hop; **C2** happens at other places. The first type of congestion takes the majority of network congestion in datacenter networks [30] and can be easily solved at the receiver in one RTT. For the other in-network congestion, RCC adjusts the bandwidth allocated to each connection based on the network delay feedback.

To realize the basic idea of RCC, there are three main challenges to be solved.

1. How to differentiate different types of network congestion at receivers. To obtain the network congestion type, we first need to detect whether and where network congestion happens accurately and responsively. Generally, network congestion can be detected based on widely-used advanced

signals, such as RTT, ECN, loss rate. However, instantaneous value directly using the instantaneous value of them will possibly lead to overreaction. On the other hand, using the average value of them will possibly be unresponsive to network congestion.

2. How to obtain accurate bandwidth sharing when addressing congestion C1. In order to assign the last-hop bandwidth accurately and fairly to RDMA connections, a straightforward method is that receivers count the precise number of active flows and then explicitly assign the average bandwidth to each connection. However, consider an Incast scenario, flows usually start one by one with a quite short interval. In this case, at first, the receiver may assign a higher congestion window to senders, causing the aggregate sending rate of all the flows higher than the last-hop bandwidth.

3. How to achieve fast convergence and near-zero queuing latency when addressing congestion C2. Due to the limited information provided by ECN marks, it is hard to utilize ECN to achieve fast convergence as well as near-zero queuing latency. Existing delay-based congestion control protocols such as TCP vegas [31], FAST TCP [32], and Compound TCP [33] have inherent limitations to achieve both fast convergence and low latency in current high-speed datacenter networks. They only react after queue build-up. Although TIMELY mitigates the problem by using the delay gradient, it fails to converge to a fixed point.

B. Framework

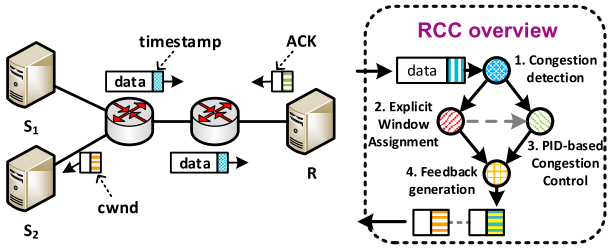
Before proceeding to describe the framework of RCC, we first explain why RCC is window-based and delay-based.

Window-based or rate-based. In rate-based congestion control schemes, packets are continuously sent before receiving feedback, which may further aggravate congestion when feedback is delayed due to congestion. Window-based solutions can avoid this problem by limiting the number of inflight packets even if the feedback is delayed. In this way, congestion will not be magnified, making the network more stabilized.

Delay-based or ECN-based. ECN is per-hop feedback, which can prevent packet loss efficiently. However, ECN-based schemes fail to effectively control the end-to-end queuing length as the number of hops increases, while RTT is end-to-end feedback information, which can be used to control the end-to-end queuing length more effectively.

Fig. 2 shows the framework of RCC. It includes three main functions: Differentiating Congestion Types (§ III-C), Explicit Window Assignment (§ III-D) and PID-based Iterative Adjustment (§ III-E). Each flow starts at line rate like other RDMA congestion control mechanisms [4], [5], [9]. Each data packet has a timestamp field to indicate the packet's sending time. As shown in Alg. 1, the receiver calculates the one-way delay by subtracting the timestamp value from the current time when the packet arrives (Line 3). Besides, if the packet belongs to a new flow or is the last packet for an existing flow, the receiver will update the number of active flows and calculate the new fair share (Line 4).

If the flow has been in PID-based congestion control procedure, it will stay in this state until the end (Line 6-9). This



(a) Workflow (b) RDMA NIC components

Fig. 2. The overview of RCC framework.

Algorithm 1 RCC Algorithm at Receiver Side

1: **INPUT:** data packet pkt
2: **OUTPUT:** sending window $cwnd$
3: $rtt \leftarrow$ CALCULATERTT(pkt)
4: $num \leftarrow$ UPDATEFLOWNUMBER(pkt)
5: $fair\_share \leftarrow$ EXPLICITWINDOWASSIGNMENT(num)
6: **if** flow already in PID-based congestion control **then**
7: $cwnd \leftarrow$ PIDCONTROL$(rtt, fair_share)$
8: **return**
9: **end if**
10: **if** $RX\ rate >= NIC\ speed * \eta$ **then**
11: $cwnd \leftarrow fair\_share$
12: **else**
13: $in\_network \leftarrow$ CONGESTIONDETECTION(rtt)
14: **if** $in\_network == true$ **then**
15: $cwnd \leftarrow$ PIDCONTROL$(rtt, fair_share)$
16: **else**
17: $cwnd \leftarrow fair\_share$
18: **end if**
19: **end if**

is because most flows are quite short in high-speed datacenter networks and switching between Explicit Window Assignment and PID-based congestion control may cause in-network queue oscillation. And the PID-based congestion control results will be limited by Explicit Window Assignment. Thus, last-hop congestion will not happen again.

Otherwise, the receiver uses one-way delay and other information to determine if in-network congestion occurs (Line 13). If in-network congestion does not happen, the receiver explicitly assigns the sending window to the fair share (Line 11 and 17). For in-network congestion, the receiver adjusts the sending window using the PID-based congestion control mechanism and the upper bound of the sending window is set to be the fair share (Line 15). After the adjustment of sending window, the receiver piggybacks this information by ACK packets to senders. The sender adjusts its sending window after receiving each ACK packet.

C. Congestion Differentiation

Detecting Network Congestion. In RCC, each packet carries the sending timestamp in its header. Upon receiving a packet, a receiver can obtain the real-time one-way delay of the

corresponding connection. Note that here we assume that the clock at senders and receivers are synchronized [34].

Let RTT_i^{base} and $RTT_i(t)$ represent the base and measured one-way delay of connection i, respectively. We can use the difference between $RTT_i(t)$ and RTT_i^{base} to infer whether network congestion happens or not. If the difference between $RTT_i(t)$ and RTT_i^{base} exceeds a threshold, then RCC will decrease the congestion window of connections.

However, many flows in datacenter networks are extremely short [35], maybe containing only several packets. Besides, each connection starts at line rate. Thus, these extremely short flows possibly incur ephemerally high $RTT_i(t)$. If we directly use $RTT_i(t)$ to detect network congestion and decrease the congestion window of all connections once the instantaneous $RTT_i(t)$ is larger than RTT_i^{base}, network bandwidth will possibly suffer from under-utilization.

*Enlightened by the **Fast Recovery** mechanism in TCP, we use n consecutive $RTT_i(t)$ values to infer whether network congestion happens or not.* Specifically, if n consecutive $RTT_i(t)$ values satisfy the following inequation:

$$RTT_i(t) > RTT_i^{base} \times (1+\delta), \quad 0 < \delta < 1, \tag{1}$$

then we can infer that network congestion happens. δ represents the allowed congestion level caused by queuing.

Determining Congestion Type. First, we calculate the received bytes of all connections, B_R, in the last round.[2] The sample duration is set to RTT_i^{base}. Let $\eta \in (0,1)$ represent the expected link utilization of the last hop. If $B_R > C \times \min_i(RTT_i^{base}) \times \eta$, then we can infer that the bandwidth of the last hop has been fully utilized, where c represents the bandwidth of the last hop. Thus, network congestion happens at the last hop. Otherwise, congestion happens at other places.

D. Explicit Window Assignment

Counting the Number of Messages, N. Unlike the stream-oriented protocol TCP, RDMA is a message-oriented one. Thus, we can easily count the number of transmitted messages based on the begin/end mark in `IB BTH` (InfiniBand Basic Transport Header). For example, in an `RDMA Write` operation, the first packet's `opcode` field in `BTH` header is set to `RDMA Write First`; the final packet of the message has an `opcode` as either `RDMA Write Last` or `RDMA Write Last With Immediate`. Thus, we can track the number of messages accurately in RDMA NICs by checking the `opcode` field of each packet. Similarly, we can also count the number of messages generated in other RDMA operations.

Computing Accurate Bandwidth Sharing. A receiver computes the congestion window for each connection i, $W(i) = \frac{C}{N}$. This computed value will be delivered to senders by ACKs. We mitigate the impacts of the second challenge by combining per-ACK window adjustment and packet pacing. Receivers piggyback the assigned window in each ACK packet based on the current active flow number. Therefore, the improper larger window sent to the senders will only last a very short time, that is, the time between two consecutive ACK packets. Moreover, NICs at senders use the packet pacing to add space between consecutive packets of all flows. Through these mechanisms, each ACK packet that carries larger window information can only trigger a small number of extra data packets. Correspondingly, the overall sending rate of Incast flows will not cause too large instantaneous queuing.

[2]The term 'round' here refers to RTT. That is, B_R is the sum of the bytes received by all connections in the previous RTT.

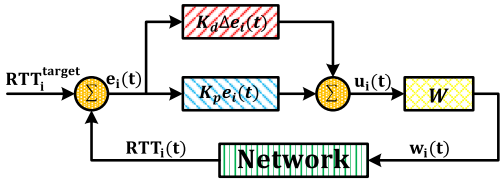


Fig. 3. Structure of PID-based congestion control.

E. PID-Based Congestion Control

For in-network congestion, RCC uses a PID controller to govern the dynamics of the sending window. In general, PID control is an adaptive optimization method widely used in various closed-loop feedback systems [36]. By applying error-based proportional (P), integral (I) and differential (D) regulation terms to the controlled object, the system can quickly respond to external disturbances and converge to the steady-state. Compared with the parameter-fixed AIMD[3] adjustment scheme adopted in traditional TCP, PID control can flexibly select the combination of regulation terms according to the characteristics of the actual system. In addition, the calibration based on real-time error sampling enables the controlled system to adaptively realize dynamic adjustment according to the real-time state of the system, avoiding continuous oscillation caused by coarse-grained control.

Specifically, in RCC, the controller continuously adapts the window to the estimated delay in order to match RTT_i^{target}. RTT_i^{target} controls the tradeoff between the bandwidth utilization and the steady-state queue length. It should be a little larger than RTT_i^{base} and smaller than the RTT value in congestion scenarios. Thus, we let

$$RTT_i^{target} = RTT_i^{base} \times (1 + \frac{\delta}{2}) \tag{2}$$

Computing the Control Factor. Fig. 3 illustrates the structure of PID-based congestion control by which RCC handles in-network congestion. It continuously calculates the error value $E_i(t)$ of connection i as the difference between the actual measured value $RTT_i(t)$ and RTT_i^{target}, that is,

$$E_i(t) = RTT_i(t) - RTT_i^{target} \tag{3}$$

To compute the control factor $U_i(t)$, which is used to adjust the sending window, the controller applies a correction based on proportional, integral, and derivative terms. In RCC, the parameter of integral term K_i is set to 0. Equation (4) expresses the overall control function. The proportional term

[3]Additive Increase in Congestion Avoidance Phase and Multiplicative Decrease in Fast Recovery Phase.

gives an instantaneous response to the error value $E_i(t)$, while the derivative term is an estimation of the future trend of $E_i(t)$.

$$U_i(t) = U_i(t-1) + K_p \times E_i(t) + K_d \times [E_i(t) - E_i(t-1)] \tag{4}$$

The proportional term can ensure that the PID-based congestion control mechanism converges to a fixed point, while the derivative term is used to achieve rapid convergence speed. Finally, with proper settings of these two parameters, RCC can maintain near-zero steady-state queues in the network without compromising other performance metrics.

Computing the Sending Window. RCC uses the control factor $U_i(t)$ to adjust the sending window of flows. If $U_i(t) > 0$, which indicates that the inflight packets is larger than the network capacity, RCC will perform a multiplicative window decrement. Otherwise, a multiplicative window increment will be conducted since the network has available bandwidth. For ease of deployment, we use $tanh(\cdot)$ (a function ranges in $(-1, 1)$) to scale the window size as follows:

$$W_i(t) = W_i(t-1) \times [1 - tanh(U_i(t))] \tag{5}$$

Due to the derivative term in (4), the increment of window size will gradually decrease as the window becomes larger, eliminating the unfairness caused by pure MIMD algorithms.

It is worth noting that formally, equation (4) adopts proportional and differential terms in standard PID control on the error value of RTT, but for the control factor $U(t)$ which finally determines the dynamics of sending window, (4) actually exerts feedback gain for its difference term (i.e., $U_i(t) - U_i(t-1)$). After performing cumulative summation on both sides of (4), it can be seen that RCC essentially adopts PI regulation for the system.[4]

F. Parameters and Overhead of RCC

Parameters of RCC. The congestion differentiation module of RCC has three parameters: n, δ, η. n and δ control the tradeoff between throughput and transient queue length. To maximize throughput, transient queues are inevitable. We set $n = 3$ and $\delta = 0.2$ for high throughput and near-zero queuing. And the expected utilization of the last hop, η is set to 0.95.

The PID-based scheme has two additional parameters: K_p and K_d. They control the speed of convergence to fairness and steady-state. The larger the K_d is, the faster the convergence speed will be. However, larger K_d will cause oscillation. We will discuss this in detail in § IV.

Overhead of RCC. Table I summarizes all state variables maintained by each RCC connection. Collectively, RCC uses 52 bytes in the sender and receiver for each RDMA connection. This memory footprint is comparable to other state-of-the-art RDMA congestion control protocols.[5] Besides, the computation overhead is negligible with a background thread.

TABLE I
STATE VARIABLES OF RCC

	State Variable	Description	Size (Byte)
Sender	cwnd	Congestion Window	4
	snd_una	Send Unacknowledged	4
	snd_nxt	Send Next	4
Receiver	flow_num	Current Flow Number	4
	rcv_nxt	Receive Next	4
	last_rtt	The Last Measured RTT	8
	last_rtt_diff	The Last RTT Diff	8
	cwnd	Congestion Window	4
	update_timer	RTT Update Timer	4
	base_rtt	Base RTT	8
Total			52

G. Deployment Choices

Clock Synchronization. The deployment of RCC relies on high precision clock synchronization throughout the datacenter network. Some recent research efforts can reduce the upper bound of clock synchronization within a datacenter to a few hundred nanoseconds, which is sufficient for our work [37], [38]. And the recent work, On-Ramp, is also trying to use the one-way delay to solve datacenter network congestion [39].

Even without high precision clock synchronization, RCC can still be practically deployed by moving the delay calculation and PID-based congestion control to the sender side; the receiver side only calculates and feeds back the flow number N and the throughput B_R. This solves the problem of not being able to obtain the one-way delay.

IV. THEORETICAL ANALYSIS

In this section, we explore the stability, convergence and fairness of RCC based on the fluid model and the control system theory [36].

A. Model Formalization

Considering N long-lived flows traversing a single bottleneck link with capacity C, taking account of the relationship between Q and RTT, i.e., $R_i(t) = \frac{q(t)}{C} + d$, the non-linear, delay-differential equations below describe the dynamics of $W_i(t)$, $Q(t)$ and $R_i(t)$:

$$\frac{dW_i(t)}{dt} = -\frac{W_i(t)tanh(U_i(t))}{R_i(t)} \tag{6}$$

$$\frac{dQ(t)}{dt} = \sum_{i}^{N} \frac{W_i(t)}{R_i(t)} - C \tag{7}$$

$$\frac{dR_i(t)}{dt} = \frac{1}{C}\frac{dQ(t)}{dt} \tag{8}$$

where d is the shared propagation delay, $R_i(t)$ and $\frac{Q(t)}{C}$ denote the RTT and shared queuing delay, respectively.

As for the control factor $U_i(t)$ in (4), we take *bilinear transformation*[6] [40] to convert it into a continuous one:

$$\frac{dU_i(t)}{dt} = \frac{K_p}{R_i(t)}[R_i(t) - R_{ref}] + (K_d + \frac{K_p}{2})\frac{dR_i(t)}{dt} \tag{9}$$

where $R_{ref} = RTT^{target}$ is the expected value in equilibrium.

[4] In fact, this indirect PI control is also used in [40]–[43]. For the simplicity and consistency throughout this paper, we use the notation "PID" and symbols in equation (4) to represent the closed-loop feedback control of RCC.

[5] For instance, $\sim$ 60 bytes in DCQCN [5].

[6] A tool for converting continuous-time (s-transform) and discrete-time (z-transform) system functions without affecting the stability of the system, also known as Tustin transform. The conversion formula is $z = \frac{2+sT}{2-sT}$, where T is the system sampling interval.

Equation (6) describes the variation of the sending window along with the difference between the real value and the reference value of RTT measured at receiver side, while equation (7) indicates the queuing process at the switch. Equations (8) and (9) capture the evolution of direct and indirect control signals, respectively. By letting the LHS of (6)-(9) equal 0, with the assumption that all flows are synchronized and peak simultaneously (which is obvious), it is easily verified that W, Q and R do not reach the steady-state until they satisfy:

$$tanh(U_i)* = 0 \quad and \quad R_i* = R_{ref} \tag{10}$$

$$\frac{W_1*}{R_1*} = \frac{W_2*}{R_2*} = \cdots = \frac{W_N*}{R_N*} \tag{11}$$

where symbol $*$ indicates the value at the fixed points, which also represents the fairness of all flows in equilibrium.

B. Derivation of the System Transfer Function

Referring to the linearization and Laplace transformation method used in [41]–[43], we give the derivation of the system function for RCC as follows:

For simplicity, we denoting W_i* and R_{ref} as W_0 and R_0, respectively. To linearize the fluid model around the fixed points represented in equations (10) and (11), we firstly redefine the RHS of equations (6), (7) and (9) by:

$$\begin{aligned} f(W,R,U) &\doteq -\frac{W_i(t)tanh(U_i(t))}{R_i(t)} \\ g(W,R) &\doteq N\frac{W_i(t)}{R_i(t)} - C \\ h(R) &\doteq \frac{K_p}{R_0}[R_i(t) - R_0] + (K_d + \frac{K_p}{2})\frac{dR_i(t)}{dt} \end{aligned} \tag{12}$$

According to the relationship between R_i and Q in equation (8) and evaluating partials at the fixed point $(W_0, R_0, U*)$ of (12) gives:

$$\begin{aligned} \frac{\partial f}{\partial W} &= -\frac{tanh(U_i)*}{R_0} = 0 \\ \frac{\partial f}{\partial R} &= \frac{W_0 tanh(U_i)*}{R_0^2} = 0 \\ \frac{\partial f}{\partial U} &= -\frac{W_0[1 - tanh^2(U_i)*]}{R_0} = -\frac{C}{N} \end{aligned} \tag{13}$$

$$\begin{aligned} \frac{\partial g}{\partial W} &= -\frac{N}{R_0} \\ \frac{\partial g}{\partial R} &= -\frac{NW_0}{R_0^2} = -\frac{C}{R_0} \end{aligned} \tag{14}$$

$$\begin{aligned} \frac{\partial h}{\partial Q} &= \frac{K_p}{CR_0} \\ \frac{\partial h}{\partial \dot{Q}} &= \frac{1}{C}(K_d + \frac{K_p}{2}) \end{aligned} \tag{15}$$

where $\dot{Q}$ is the differential of Q.

With the denotions $\delta W \doteq W - W_0$, $\delta Q \doteq Q - Q*$, $\delta U \doteq U - U*$, we make linearization around the operation point and obtain:

$$\delta\dot{W}(t) = -\frac{C}{N}\delta U(t)$$

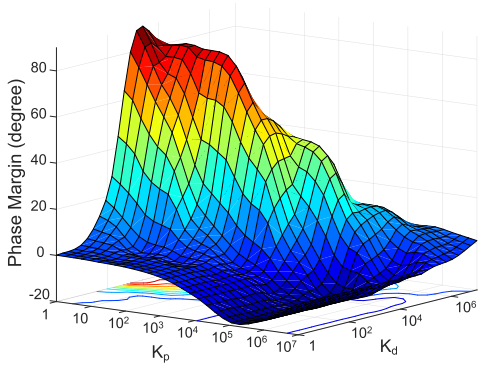


Fig. 4. Phase margin as a function of parameter K_p and K_d.

$$\begin{aligned} \delta\dot{Q}(t) &= \frac{N}{R_0}\delta W(t) - \frac{C}{R_0}\delta R(t) \\ \delta\dot{U}(t) &= \frac{K_p}{CR_0}\delta Q(t) + \frac{1}{C}(K_d + \frac{K_p}{2})\delta\dot{Q}(t) \end{aligned} \tag{16}$$

Performing the Laplace transform on differential equations in (16) we get:

$$\begin{aligned} W(s) &= -\frac{C}{Ns}U(s) \\ Q(s) &= \frac{N}{sR_0 + 1}W(s) \\ U(s) &= -\frac{\frac{K_p}{R_0} + (K_d + \frac{K_p}{2})s}{Cs} \end{aligned} \tag{17}$$

Combining functions in (17), we finally get the open-loop transfer function of the whole system as follows:

$$G(s) = K\frac{1 + \frac{s}{z}}{s^2(s + \frac{1}{R_0})} \tag{18}$$

where $K = \frac{K_p}{R_0^2}$ and $z = \frac{K_p}{(K_d + K_p/2)R_0}$.

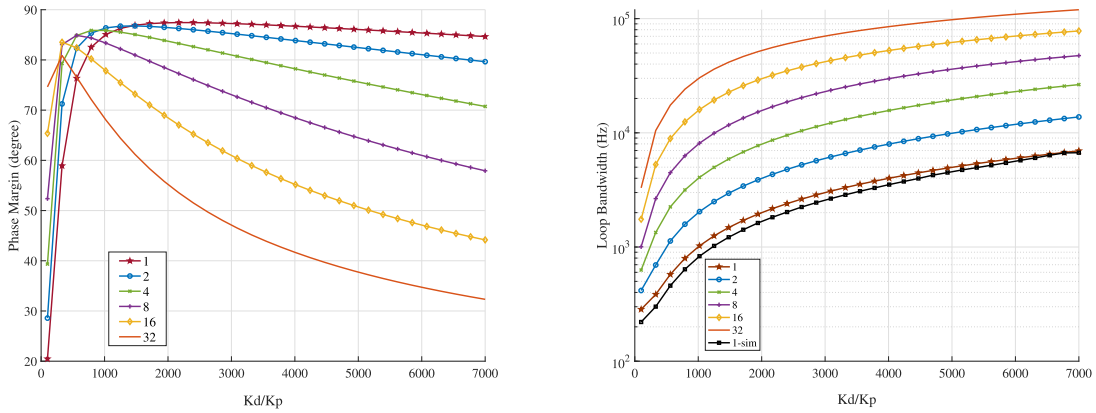
C. Stability Analysis and Control Parameters

According to the Bode Stability Criteria, a system is stable only if the phase margin in the Bode diagram of the transfer function $G(s)$ is above 0. And the higher the phase margin, the more stable the system. Taking the setting mentioned above for parameter δ with 0.2, i.e., $R_{ref} = 13.2\mu s$, Fig. 4 shows the variation of the phase margin relative to different (K_p, K_d) pairs. As shown in Fig. 4, setting $K_p \in [1, 10^4]$ and $K_d \in [10^3, 10^5]$ ensures an acceptable phase margin above 30 degrees. Besides, to make the system more stable, K_p should be one or two orders of magnitude smaller than K_d. In fact, the response speed (time taken to reach equilibrium) and the optimal operation range (with guaranteed open-loop gain) of a system are often opposite, which will be discussed below.

D. Convergence Analysis

Proof As concluded in [44], PID controllers (like the one used in RCC) that can be characterized by a SISO ARMAX[7] model shown in equation (19) and be optimized

[7]SISO ARMAX stands for *Single-Input-Single-Output AutoRegressive Moving Average with eXtra input*, which describes a controlled system with external interference.



(a) Phase margin trend with different pairs of (K_p, K_d) values

(b) Loop frequency trend with different pairs of (K_p, K_d) values

Fig. 5. The trend of RCC stability and convergence under different (Kp,Kd) pairs.

by a generalized minimum variance objective function like equation (20) have global convergence[8] in equation (21).

$$A(q^{-1})y(t) = q^{-d}B(q^{-1})u(t) + C(q^{-1})\omega(t) \tag{19}$$

where $y(t)$ and $u(t)$ are the output and input sequences of the system, respectively. d is the system delay, $\omega(t)$ is a random sequence defined in the probability space (Ω, F, P), $A(q^{-1})$, $B(q^{-1})$ and $C(q^{-1})$ are model operator polynomials.

$$J = E\{[\mathring{P}y(t+d) - \mathring{R}y_t(t)]^2 + [\mathring{Q}'u(t)]^2\} \tag{20}$$

where $\mathring{P}$, $\mathring{R}$ and $\mathring{Q}'$ are known weighted polynomials about q^{-1}, and $y_t(t)$ is the target reference output.

$$\lim_{N\to\infty} \sum_{t=1}^{N} \frac{[e(t) - \mathring{E}\omega(t)]^2}{r(t-1)} \tag{21}$$

where $e(t) = y(t) - y_t(t)$, $r(t)$ is an observable random sequence and $\mathring{E}$ is a polynomial solution satisfying the Diophantine equation.[9]

As demonstrated in [44], all operators and noises (i.e., random sequences $\omega(t)$ and $r(t)$) in equations (19)-(21) have nothing to do with the design details of RCC, that is, for an arbitrarily set reference value y_t (R_{ref} in RCC) of the controlled signal, a specific parameter combination can always be found to meet the convergence condition shown in equation (21). Intuitively, under the optimization of (20), the error signal $e(t)$ (e.g., $E_i(t)$ in RCC) of the system will eventually approach a fixed value; that is, the system has global convergence under the adjustment of the controlled parameters. See [44] for complete proof. □

To further explore the trade-off between system convergence and stability, we use numerical simulation to solve the phase margin (PM) and loop bandwidth (LB) under a series of (K_p, K_d) pairs, where the absolute values of these two terms are positively correlated with the degree of stability and the convergence speed, respectively. We start K_p from 1 with five doubling rounds, and construct a wide variation range for K_d by offering 31 sets of scale factors (i.e., K_d/K_p).

The trends of PM and LB are shown in Fig. 5. We find that PM has a maximum with the increase of K_d for any given K_p, and a smaller K_p owns a wider stable range. Besides, under the same scale factor, PM sharply decreases as K_p doubles, which indicates that the proportional term, K_p, as the primary control item of PID control, needs to be adjusted prudently. However, for any given K_p, LB increases monotonically with respect to the scale factor, and under a fixed scale factor, the increase of K_p will cause a synchronous growth of LB, which will bring a faster system response. In a nutshell, K_p has a mutually-exclusive impact on stability and convergence.

Besides, as shown in Fig. 5(b), RCC achieves near-ideal convergence speed, the difference between numerical (brown) and NS3 (black) simulations reduces (22.56% → 0.54%) as scale factor grows. We take the $(1, 1000)$ pair as the reference, which gives a $\frac{1}{1000Hz} = 1ms$ ideal convergence time. Considering stability and convergence speed comprehensively with the actual RTT magnitude, we limit $K_p \in [1, 16]$ and $K_d \in [10^3, 7 \times 10^3]$ in actual use.

V. Simulation Evaluation

In this section, we evaluate the performance of RCC by conducting both micro-benchmarks and large-scale simulations.

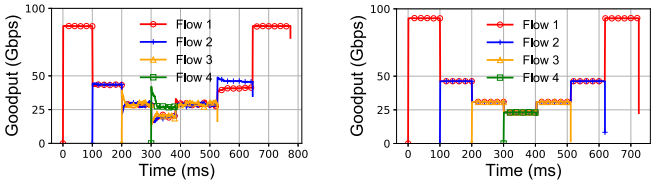
A. Evaluation Setup

Network Topology. A fat-tree topology is used in large-scale simulations. It consists of 320 hosts, 20 ToR switches, 20 aggregation switches, and 16 core switches. Hosts and ToRs are connected with 100 Gbps links, while all switches are connected via 400 Gbps links. The delay of each link is set to $1\mu s$, which gives a $12\mu s$ base RTT. The switch buffer size is set to $32MB$ according to real device configurations [4].

Schemes Compared. We compare RCC with DCQCN, TIMELY, HPCC, Homa [45] and ExpressPass [46]. Homa and ExpressPass are typical receiver-driven transport protocols, although they are not designed for RDMA networks. We use the open-source code of DCQCN [47], Homa [48] and HPCC

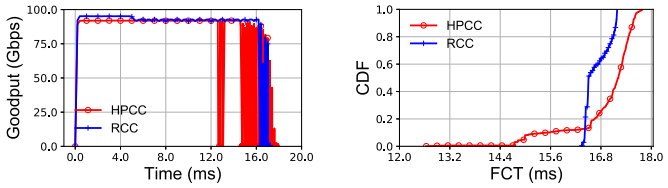
[8]Global convergence is a kind of strong attribute, where the system will always eventually stabilize around a unique point under any initial conditions.

[9]The Diophantine equation describes the relationship between $\mathring{E}$ and operators A, B and C, which is independent of the specific expressions relevant operators.



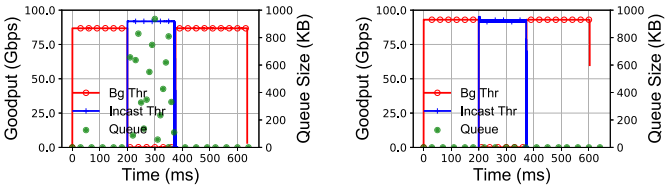
(a) HPCC (b) RCC

Fig. 6. Fast convergence and fairness.



(a) Goodput (b) CDF of FCT

Fig. 7. Goodput and FCT in the Incast scenario.



(a) HPCC (b) RCC

Fig. 8. Queuing length under Incast scenario.

in our evaluation and implement ExpressPass and TIMELY based on their algorithms.

Parameter Settings. For RCC, we set $\delta = 0.2$, which leads to a $13.2\mu s$ RTT_i^{target}. Then, we set $K_p = 1 \times 10^4$ and $K_d = 1 \times 10^5$ based on the analysis in § IV. For other schemes, we use the parameters suggested in the corresponding papers. We also scale the ECN marking threshold proportional to the link bandwidth suggested in [4] for DCQCN.

Benchmark Workloads. We use four kinds of realistic datacenter workloads (web search [18], data mining [20], web server [12] and cache follower [12]) that are widely used in prior literature for large-scale simulations.

Performance Metrics. The performance of RCC is evaluated using the following metrics: (1) goodput, (2) convergence speed/fairness, (3) average FCT, (4) tail FCT, and (5) in-network queuing length.

B. Micro-Benchmarks

We compare RCC with HPCC in several micro-benchmarks since HPCC is regarded as the best solution for the time being.

1) Fast Convergence and Fairness. We show that RCC flows can quickly react to changes in available bandwidth and rapidly converge to appropriate flow rates. Besides, we evaluate the fairness of RCC.

Setup: We use a dumbbell topology, where four senders and one receiver are connected with the same 100 Gbps bottleneck link. Four senders transmit one flow to the receiver in turn with an interval of $100ms$. The lengths of the four flows are 4.4, 2.2, 1.1 and 0.27 GB, respectively.

Result: Fig. 6 shows the goodput of the four flows in RCC and HPCC. RCC quickly throttles the rate of existing flows upon a new flow starts, and recovers the rate of other flows after a flow ends. This is because the receiver has precise information about the flow number being transmitted. When a flow finishes, RCC only needs one RTT to re-assign the sending window for existing flows to fully utilize the newly available bandwidth. As a result, the overall goodput of RCC is more stable. However, HPCC needs several RTTs to converge to a stable rate after a flow arrives/finishes.

Fig. 6 also illustrates the fairness of RCC and HPCC. RCC provides better fairness even in a short time scale. All flows evenly share the bottleneck bandwidth and grab their fair share quickly. Specifically, when N varies from 1 to 4, each flow's goodput quickly converges to $\sim \frac{100 \times 0.95}{N} Gbps$, giving a Jain fairness index within $0.998 \sim 0.999$ (1 is optimal). However, flows in HPCC can not get the fair share under a various number of concurrency and the goodput suffers oscillation.

2) Incast. Next, we evaluate the performance of RCC under the Incast scenario.

Setup: One receiver initiates connections with 1000 senders and requests $200KB$ data from each sender simultaneously. This workload is similar to the workload used in [49]. Besides, there is also a long-lived background flow to the receiver. We evaluate the overall FCT and the goodput of flows.

Result: Fig. 7 illustrates how RCC and HPCC react to congestion caused by *Incast*. The aggregate goodput of RCC and HPCC is similar. The total goodput of the 1000 flows remains stable as time goes on, at around $94.98Gbps$. As for the average FCT of Incast flows, RCC performs much better than HPCC. RCC improves the median of FCT by about 5.2% and improves the 99^{th} percentile FCT by about 4.1%.

3) Low Queuing Latency. Incast traffic likely causes instantaneous large queuing latency due to its burstiness. We next show the queuing latency of RCC in this scenario.

Setup: We use the same Incast traffic as above and measure the queue length at the bottleneck switch.

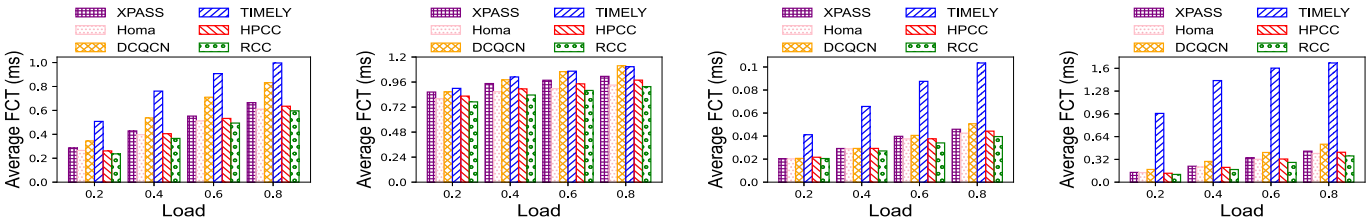
Result: Fig. 8 shows the queue length of HPCC and RCC under Incast. RCC keeps near-zero queue length since it explicitly assigns window size and combines per-ACK feedback and packet pacing. In HPCC, the queue builds up to $900KB$ due to its slow responsiveness to the large flow concurrency.

C. Large-Scale Simulations

Now, we evaluate the performance of RCC in large-scale scenarios. We use the fat-tree topology mentioned in § V-A and generate traffic according to the four realistic workloads.

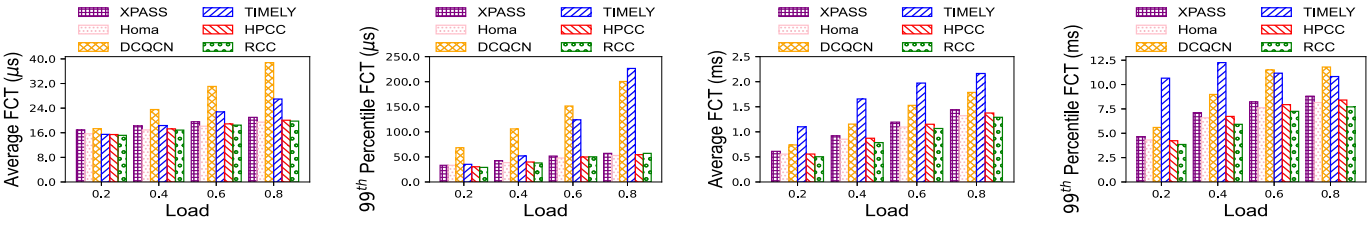
1) Overall Performance.

Result: Fig. 9 shows the average FCT achieved by each scheme under four workloads and different link loads. The performance of RCC is better than Homa, ExpressPass, DCQCN,



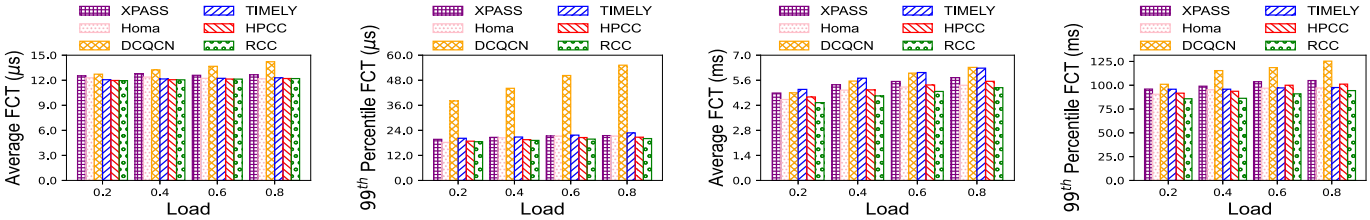
(a) Web Search (b) Data Mining (c) Web Server (d) Cache Follower

Fig. 9. Overall performance under four realistic workloads.



(a) 0-100KB average FCT (b) 0-100KB tail FCT (c) >100KB average FCT (d) >100KB tail FCT

Fig. 10. Overall performance in web search workload.



(a) 0-100KB average FCT (b) 0-100KB tail FCT (c) >100KB average FCT (d) >100KB tail FCT

Fig. 11. Overall performance in data mining workload.

TIMELY and HPCC. In the web search workload, the overall average FCT with RCC is up to 9%, 11%, 11%, 30% and 45% lower compared with HPCC, Homa, ExpressPass, DCQCN and TIMELY, respectively. And the results change to 7%, 4%, 15%, 18% and 19% under the data mining workload. Besides, RCC delivers the best performance under the web server and cache follower workloads. The improvement of RCC over HPCC in the overall average FCT is also obvious: $5 \sim 11\%$ for the web server workload and $\sim 14\%$ for the cache follower workload.

2) FCT Breakdown Based on Flow Size.

Result: We break down the FCT across flow sizes as shown in Fig. 10 and Fig. 11. We omit the results of web server and cache follower workloads since the performance is consistent with the other workloads.

For short flows, though under which Homa is proved to be highly effective, RCC performs better than it and HPCC, and greatly outperforms ExpressPass, DCQCN and TIMELY. This is because RCC keeps near-zero queue length and handles congestion rapidly by combining the Explicit Window Assignment mechanism with the PID-based congestion control, while the core packet spraying scheme used in Homa may encounter various issues. Fig. 11(a) and 11(b) show that RCC has a slightly better performance than HPCC both in average FCT and 99^{th} percentile FCT under the data mining workload. Besides, RCC improves the average/99^{th} percentile FCT by $10 \sim 25\%$ and $21 \sim 50\%$ compared with DCQCN, respectively. In the web search workload, RCC still achieves better performance than HPCC and gains similar FCT reduction over the other four schemes as under data mining workload.

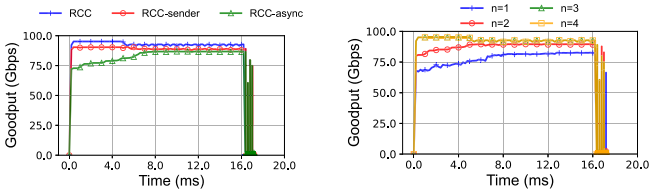
As for long flows, RCC performs better than all the other schemes. In the data mining workload, RCC performs better than HPCC in terms of the average and 99^{th} percentile FCT. RCC also achieves great performance for long flows in the web search workload. Compared with DCQCN, RCC reduces by $25 \sim 31\%$ for the average FCT and up to 37% for the 99^{th} percentile FCT. This is because RCC can rapidly reach the fair share by the Explicit Window Assignment algorithm.

D. RCC Deep Dive

In this section, we dig deeper into RCC's design by conducting a series of targeted simulations.

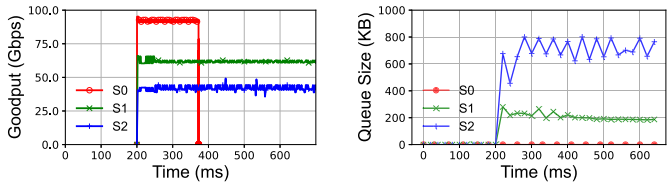
1) Impact of Clock Synchronization.

Setup: To explore the impact of clock synchronization mechanism on the performance of RCC, we construct two control groups, in which RCC-async simulates the interference of asynchronous clock by introducing random time jitter, while



(a) Impact of clock synchronization (b) Impact of parameter n

Fig. 12. The impacts of clock synchronization and parameter n on RCC.



(a) Goodput (b) Queue legnth

Fig. 13. The impact of different scale functions.

RCC-sender moves the PID control logic to the sender side (as elaborated in § III-G). We run RCC, RCC-async and RCC-sender using the same settings in § V-B-2).

Result: Fig. 12(a) illustrates that time jitter caused by clock synchronization does lead to a deterioration on the goodput, but not pretty severe as the result in RCC-sender achieves similar steady and high throughput compared with RCC, which again verifies the feasibility of deployment choice in § III-G. In fact, the impact caused by the congestion signal transition from one-way delay to RTT will be alleviated by the order-of-magnitude difference between end-to-end delay and the per-ACK window adjustment manner, thus there will be no avalanche decline in performance.

2) Settings of Parameter n in Congestion Detection.

Setup: As a congestion control mechanism, whether congestion can be detected efficiently and accurately is the key to achieve optimal performance, which is dominated by parameter n. To determine the least boundary conditions of n, we take RCC (i.e., $n = 3$) in 1) as the baseline and measure the goodput of bottleneck link with the same simulation settings.

Result: The results in Fig. 12(b) reveal that instantaneous delay rag ($n = 1$) would fuzz up congestion detection thus degrading performance of RCC compared with recommended parameter settings ($n = 3$). In addition, the goodput of the bottleneck link keeps approaching the optimal steady-state performance as the RTT sampling value (i.e., n) increases. Actually, RCC gives a similar performance when $n > 2$, thus we choose $n = 3$ as the least boundary for congestion detection for simplicity and effectiveness.

3) Impact of the Scale Function in PID Controller.

Setup: To explain in depth the reason for introducing $tanh(\cdot)$ to shape the congestion signal $U_i(t)$, we denote the scaling term $1 - tanh(\cdot)$ in equation (5) as S_0, and then use $sigmoid(\cdot)$[10] to construct two new scaling terms $S_1 = 1 - sigmoid(\cdot)$, $S_2 = 2 - 2sigmoid(\cdot)$. S_1 is similar to S_0 in form but has a smaller range, while S_2 has the same range with S_0 but cannot derive the delay-differential equation like (6). We modified RCC with different scale functions and rerun the simulations in § V-B-3).

Result: We can observe from Fig. 13 that the alter of scale function would lead to an apparent reduction on goodput and a persistent oscillation in bottleneck queue length. This further shows the robustness of recommended $tanh(\cdot)$ reshaper. In fact, as a centrosymmetric function about reference point (0,1), $tanh(\cdot)$ performs homomorphic window adjustment on both sides of the ideal equilibrium point, thus ensuring the fairness of the whole system. It also avoids excessive interference of noise signal through non-uniform scaling, which has a vital impact on the stability, fairness and convergence of PID controller. And as a window-based congestion control algorithm, the application of $tanh(\cdot)$ in RCC constrains the upper/lower limits of window adjustment, which is consistent with the maximum halving boundary setting in classical algorithms like DCTCP and DCQCN. In comparison, although S_1 and S_2 share either the same differentiability or same value range with S_0, the lack of the other property results in the sluggish response with S_1 and the instability with S_2.

[10] $Sigmoid(\cdot)$ is another widely used normalization function with a range of (0,1). Its expression is $S(x) = \frac{1}{1+e^{-x}}$.

TABLE II
PAUSE TIME (%) UNDER DIFFERENT INCAST SEVERITY

	# of concurrent flows					
	16	**32**	**64**	**128**	**192**	**256**
DCQCN	0	0	0	0	27.6%	42.1%
TIMELY	0	0	3.1%	7.2%	33.9%	54.9%
RCC	0	0	0	0	0	0.3%

4) Robustness to PFC Activation.

Setup: To quantify the superior performance of RCC in preventing PFC activation, we conduct a series of N-to-1 Incast scenarios with each sender sends 200KB data.

Result: As shown in Table II, benefiting from the receiver-driven congestion divide-and-conquer, PFC is rarely triggered under the control of RCC (only 0.3% under 256 concurrency). By contrast, lots of PFC pause frames are generated in DCQCN and TIMELY with concurrent flows larger than 128; this can easily impair the performance of the entire network.

5) The Effects of Explicit Window Assignment and PID-Based Congestion Control.

Setup: To evaluate the effect of the schemes in § III-D and § III-E, we compare the complete RCC with RCC-PID (RCC without Explicit Window Assignment) and RCC-EWA (RCC without PID-based Congestion Control) in the fat-tree topology with the web search workload.

Result: Fig. 14 shows the average FCT of RCC, RCC-PID and RCC-EWA. We observe that there is always a gap in the FCT results with RCC in the last two schemes, which indicates that both EWA and PID are vital for better performance in RCC. Specifically, RCC outperforms RCC-PID and RCC-EWA by up to 28% and 19% on average, and the result of tail FCT breakdown is similar. The gap would be remedied

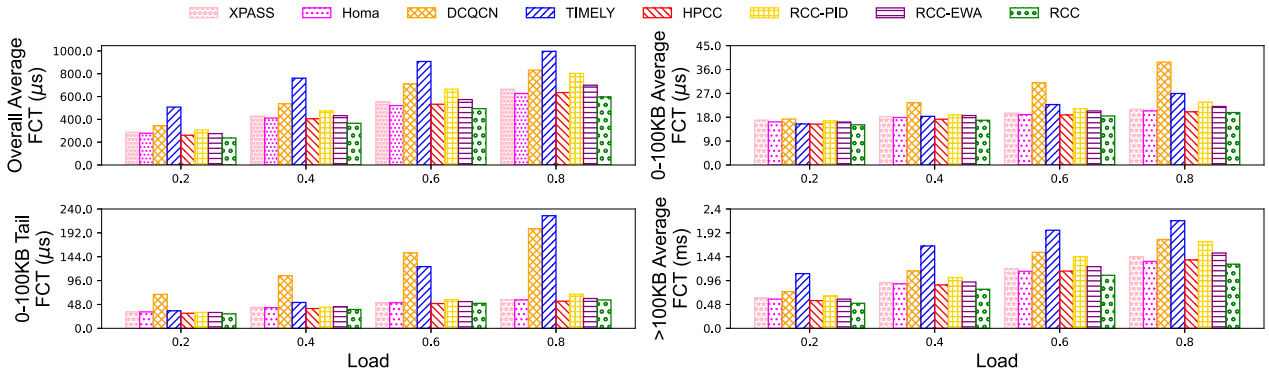
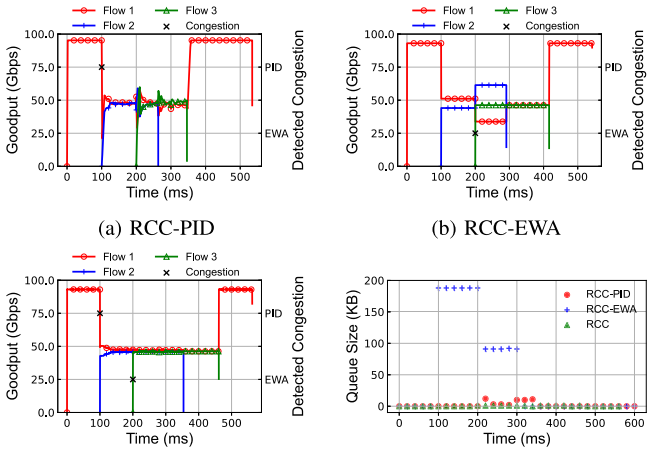


Fig. 14. RCC deep dive in web search workload.



(a) RCC-PID (b) RCC-EWA

(c) RCC (d) Queuing length

Fig. 15. RCC deep dive with multi-bottleneck topology.

by equipment with either the EWA or PID part, as detailed in the next simulation.

6) The Effects of EWA and PID in More Detail.

Setup: We run RCC-PID, RCC-EWA and RCC using the multi-bottleneck topology in Fig. 1(a) with three flows to quantify the benefits of combining these two schemes. Flow 1 begins to transmit data at line rate, then we start flow 2, 3 at $100ms$ and $200ms$ to construct both in-network and last-hop congestion with flow 1. Besides, we add flags in EWA and PID units to check the validity and accuracy of congestion detection and differentiation algorithms. All links are $100Gbps$.

Result: Fig. 15 shows the goodput of three flows in each scheme. We observe from Fig. 15(a) that RCC-PID requires several iterations to reach the proper sending rate every time the number of flows changes. Fig. 15(b) and 15(d) show that flow 1 and 2 can not share the bottleneck bandwidth fairly and build a large queue in the network. However, Fig. 15(c) illustrates that RCC handles different kinds of congestion perfectly by combining EWA and PID parts and achieves near-zero queuing length. Moreover, the consistency of detected congestion type with constructed congestions (denoted as $\times$) in Fig. 15(a)-(c) indicates that the congestion detection and differentiation in RCC are effective and accurate.

VI. Testbed Evaluation

In this section, we implement RCC with DPDK [50] and conduct testbed experiments to validate its performance compared with simulation results.

A. Implementation and Settings

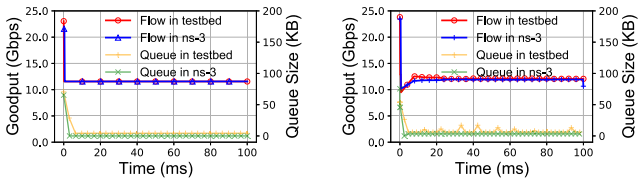
In order to emulate the kernel bypass property of RDMA, we implement RCC based on the user-level mTCP stack [51]. Specifically, function *ProcessACK()* in *tcp_in.c* covers the main algorithm of RCC and corresponding state variables are added to basic packet headers. We plug in a Mellanox ConnectX5-EN NIC on each of the two Dell PowerEdge R730 servers to act as sender and receiver, respectively. Each NIC has two $25Gbps$ Ethernet ports, so the server can work as two senders or receivers. We build a two-tier network topology with three switches and then connected the servers to the two leaf switches. The rate of all links is $25Gbps$. Each server in this topology is equipped with two Intel Xeon E5-2609 v4 CPUs (8 cores, 1.7 GHz), 128 GB 2133 MT/s DDR4 RAM.

B. Cross Validation of EWA and PID Components

The highlight of RCC lies in its differentiated treatment of different types of network congestion. To verify the basic performance of the algorithm design, we run two different experiments on both the testbed and NS3 simulator and measure throughput and switch egress queue length. In both testbed and simulator, the RTT is about $12\mu s$. So the parameters are set to be the same in hardware experiments and NS3 simulations.

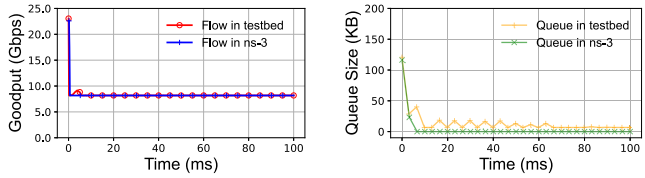
In the first experiment, each sender generates one flow to the same receiver, which results in the last-hop congestion. Fig. 16(a) shows the sending rate stabilizes at $12Gbps$ with a near-zero queue length for both the testbed and simulation.

In the second experiment, each sender generates one flow to different receivers, which results in the in-network congestion. Fig. 16(b) shows that the sending rate stabilizes at $12Gbps$ for both the testbed and simulation. In the steady-state, the throughput of two senders oscillates with low amplitude in both testbed and simulation and the queue length is limited to a very low level. Results verify that both EWA and



(a) Last-hop congestion scenario. (b) In-network congestion scenario.

Fig. 16. Cross validation with simulations.



(a) Goodput (b) Queue legnth

Fig. 17. Performance of RCC testbed under 3-to-1 Incast.

PID components in RCC would behave as expected (that in simulations) under real datacenter network environments.

C. Performance Under Incast Scenario

To cross-validate the ability of RCC to deal with abnormal traffic patterns in real scenarios, we reuse the environment settings in § VI-B, but connect these two servers to one switch, which gives a 3-to-1 Incast scenario. The frontend application on one end-host (receiver) sends requests to the other three end-hosts (senders). Upon receiving the request, each sender replies with continuous elephant flow immediately. We apply the same settings in NS3 simulator and record the goodput and switch queue length. Fig. 17 demonstrates that despite the presence of oscillation and jitter, the results of the simulation and the experiment can still maintain an absolute degree of consistency, which is acceptable and valuable.

VII. Related Work

Congestion control (CC) is an enduring topic in data center networks (DCN), and a plethora of novel proposals have emerged over the last decade. Here we briefly introduce some closely related work.

A. Sender-Based Congestion Control

DCTCP [18] is the first ECN-based CC algorithm used in DCN. It adjusts each flow's sending rate according to the ECN-mark fraction in each RTT. With the prosperity of RDMA in DCN, sender-based RDMA-specific CC mechanisms are proposed, which passively adjust the transmission behavior at senders according to various congestion signals. DCQCN [5] reacts to ECN marks, TIMELY [9] and Swift [52] use RTT variation, while HPCC [4] relies on precise link load information. Due to the long end-to-end control loop, these solutions require several RTTs to iteratively converge to the fair share, while RCC needs only one RTT for handling the dominant last-hop congestion.

B. Receiver-Driven Congestion Control

To avoid possible performance degradation under Incast scenario, some proposals suggest shifting the control entity to the receiver side. pHost [53] and ExpressPass [46] prevents congestion by explicitly sending well-controlled tokens/credits to senders. However, RDMA NICs are hard to maintain different timers for flow-level packet pacing implement. Homa [45] uses priority queues to schedule packets dynamically. However, since the core packet spraying [54] scheme likely incurs packet reordering, it is not well-supported in RDMA networks. Aeolus [55] assists receiver-driven CCs by prioritizing scheduled packets over unscheduled along with effective selective dropping and recovery schemes, but issues like core congestion are pendent. In essence, the effectiveness of receiver-driven solutions coping with the last-hop congestion highly depends on the assumption that most congestion happens at the ToR downlink [13]. RCC inherits this good feature but pushes further for designing a dedicated scheme for the in-network congestion.

C. Switch-Driven Solutions

XCP [56] and RCP [57] are two proactive rate control schemes designed primarily for TCP. With the participation of the switch, the former adjusts the window size information in the packet header, while the latter calculates each link's fair rate. TFC [58] proposes a token-based bandwidth allocation scheme based on the active flow number in each time interval, while BFC [59] tracks active flows to achieve accurate per-hop per-flow flow control and is compatible with end-to-end solutions. RoCC [40] proactively computes the fair flow rate and piggybacks it to the sender based on PI controller at the switch, RCC differs from it in indirect control and continuous control factor scaling.

VIII. Conclusion

This paper presents RCC, a receiver-driven transport for RDMA in datacenters. It can efficiently utilize the network bandwidth while keeping near-zero in-network queue length. By differentiating network congestion types, RCC employs novel *Explicit Window Assignment* and *PID-based congestion control* to address the last-hop and in-network congestion, respectively. The results of the testbed experiments and large-scale simulations validate that RCC achieves low queuing latency, high bandwidth utilization, and fairness simultaneously.

References

[1] J. Zhang *et al.*, "Receiver-driven RDMA congestion control by differentiating congestion types in datacenter networks," in *Proc. IEEE 29th Int. Conf. Netw. Protocols (ICNP)*, Nov. 2021, pp. 1–12.
[2] K. Hazelwood *et al.*, "Applied machine learning at facebook: A datacenter infrastructure perspective," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2018, pp. 620–629.
[3] A. Klimovic, C. Kozyrakis, E. Thereska, B. John, and S. Kumar, "Flash storage disaggregation," in *Proc. 11th Eur. Conf. Comput. Syst.*, Apr. 2016, pp. 1–15.
[4] Y. Li, R. Miao, H. H. Liu, Y. Zhuang, F. Feng, and Z. Cao, "HPCC: High precision congestion control," in *Proc. ACM SIGCOMM*, 2019, pp. 44–58.

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

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

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

[8] Y. Gao, Y. Yang, T. Chen, J. Zheng, B. Mao, and G. Chen, "DCQCN+: Taming large-scale incast congestion in RDMA over Ethernet networks," in *Proc. IEEE 26th Int. Conf. Netw. Protocols (ICNP)*, Sep. 2018, pp. 110–120.

[9] R. Mittal, V. T. Lam, N. Dukkipati, E. Blem, H. Wassel, and M. Ghobadi, "TIMELY: RTT-based congestion control for the datacenter," in *Proc. ACM SIGCOMM*, 2015, pp. 537–550.

[10] M. Handley *et al.*, "Re-architecting datacenter networks and stacks for low latency and high performance," in *Proc. Conf. ACM Special Interest Group Data Commun.*, Aug. 2017, pp. 29–42.

[11] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: Measurements & analysis," in *Proc. ACM SIGCOMM*, 2009, pp. 202–208.

[12] 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.

[13] A. Singh *et al.*, "Jupiter rising: A decade of clos topologies and centralized control in Google's datacenter network," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 183–197, 2015.

[14] *The NS3 Simulator*. Accessed: Apr. 10, 2021. [Online]. Available: https://www.nsnam.org/

[15] I. Marinos, R. N. M. Watson, and M. Handley, "Network stack specialization for performance," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 175–186, Feb. 2015.

[16] *Supplement to InfiniBand Architecture Specification Volume 1 Release 1.2.2 Annex A17: RoCEv2 (IP Routable RoCE)*, Infiniband Trade Association, Armonk, NY, USA, 2014.

[17] *802.1Qbb—Priority-Based Flow Control*. Accessed: Apr. 10, 2021. [Online]. Available: http://www.ieee802.org/1/pages/802.1bb.html

[18] M. Alizadeh *et al.*, "Data center TCP (DCTCP)," in *Proc. ACM SIGCOMM Conf. (SIGCOMM)*, 2010, pp. 63–74.

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

[20] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, and P. Lahiri, "VL2: A scalable and flexible data center network," in *Proc. ACM SIGCOMM*, 2009, pp. 51–62.

[21] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, and Y. Shi, "BCube: A high performance, server-centric network architecture for modular data centers," in *Proc. ACM SIGCOMM*, 2009, pp. 63–74.

[22] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "DCell: A scalable and fault-tolerant network structure for data centers," in *Proc. ACM SIGCOMM*, 2008, pp. 75–86.

[23] V. Jeyakumar, M. Alizadeh, D. Mazières, B. Prabhakar, A. Greenberg, and C. Kim, "EyeQ: Practical network performance isolation at the edge," in *Proc. USENIX NSDI*, 2013, pp. 297–311.

[24] Q. Zhang, V. Liu, H. Zeng, and A. Krishnamurthy, "High-resolution measurement of data center microbursts," in *Proc. Internet Meas. Conf.*, Nov. 2017, pp. 78–85.

[25] A. Kalia, M. Kaminsky, and D. G. Andersen, "Using RDMA efficiently for key-value services," in *Proc. ACM Conf. SIGCOMM*, Aug. 2014, pp. 295–306.

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

[27] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, and V. Josifovski, "Scaling distributed machine learning with the parameter server," in *Proc. USENIX OSDI*, 2014, pp. 583–598.

[28] D. Zats, T. Das, P. Mohan, D. Borthakur, and R. Katz, "DeTail: Reducing the flow completion time tail in datacenter networks," in *Proc. ACM SIGCOMM*, 2012, pp. 139–150.

[29] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: Distributed data-parallel programs from sequential building blocks," in *Proc. ACM EuroSys*, 2007, pp. 59–72.

[30] V. Vasudevan, A. Phanishayee, H. Shah, E. Krevat, D. G. Andersen, and G. R. Ganger, "Safe and effective fine-grained TCP retransmissions for datacenter communication," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 303–314, 2009.

[31] L. S. Brakmo and L. L. Peterson, "TCP Vegas: End to end congestion avoidance on a global internet," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 8, pp. 1465–1480, Oct. 1995.

[32] C. Jin, D. X. Wei, and S. H. Low, "FAST TCP: Motivation, architecture, algorithms, performance," in *Proc. IEEE INFOCOM*, 2004, pp. 2490–2501.

[33] K. Tan, J. Song, Q. Zhang, and M. Sridharan, "A compound TCP approach for high-speed and long distance networks," in *Proc. 25th IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2006, pp. 1–12.

[34] *IEEE Approved Draft Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems*, IEEE Standard 1588-2019, 2019.

[35] A. Kalia, M. Kaminsky, and D. Andersen, "Datacenter RPCs can be general and fast," in *Proc. USENIX NSDI*, 2019, pp. 1–16.

[36] G. F. Franklin, J. D. Powell, A. Emami-Naeini, and J. D. Powell, *Feedback Control of Dynamic Systems*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[37] Y. Geng, S. Liu, Z. Yin, A. Naik, B. Prabhakar, and M. Rosenblum, "Exploiting a natural network effect for scalable, fine-grained clock synchronization," in *Proc. USENIX NSDI*, 2018, pp. 81–94.

[38] Y. Li, G. Kumar, H. Hariharan, H. Wassel, P. Hochschild, and D. Platt, "Sundial: Fault-tolerant clock synchronization for datacenters," in *Proc. USENIX OSDI*, 2020, pp. 1171–1186.

[39] S. Liu, A. Ghalayini, M. Alizadeh, B. Prabhakar, M. Rosenblum, and A. Sivaraman, "Breaking the transience-equilibrium Nexus: A new approach to datacenter packet transport," in *Proc. USENIX NSDI*, 2021, pp. 47–63.

[40] P. Taheri *et al.*, "RoCC: Robust congestion control for RDMA," in *Proc. ACM CoNEXT*, 2020, pp. 17–30.

[41] M. Alizadeh, A. Kabbani, B. Atikoglu, and B. Prabhakar, "Stability analysis of QCN: The averaging principle," in *Proc. ACM SIGMETRICS*, 2011, pp. 49–60.

[42] C. V. Hollot, V. Misra, D. Towsley, and W.-B. Gong, "A control theoretic analysis of RED," in *Proc. IEEE Conf. Comput. Commun. 20th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, Apr. 2001, pp. 1510–1519.

[43] R. Pan *et al.*, "PIE: A lightweight control scheme to address the bufferbloat problem," in *Proc. IEEE 14th Int. Conf. High Perform. Switching Routing (HPSR)*, Jul. 2013, pp. 148–155.

[44] R. P. Borase, D. K. Maghade, S. Y. Sondkar, and S. Y. Pawar, "A review of PID control, tuning methods and applications," *Int. J. Dyn. Control*, vol. 9, pp. 818–827, Jul. 2020.

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

[46] I. Cho, K. Jang, and D. Han, "Credit-scheduled delay-bounded congestion control for datacenters," in *Proc. Conf. ACM Special Interest Group Data Commun.*, Aug. 2017, pp. 239–252.

[47] Y. Zhu, M. Ghobadi, V. Misra, and J. Padhye, "ECN or delay: Lessons learnt from analysis of DCQCN and TIMELY," in *Proc. ACM CoNEXT*, 2016, pp. 313–327.

[48] *Homa Simulation*. Accessed: Apr. 10, 2021. [Online]. Available: https://github.com/PlatformLab/HomaSimulation/

[49] Y. Chen, R. Griffith, J. Liu, R. H. Katz, and A. D. Joseph, "Understanding TCP incast throughput collapse in datacenter networks," in *Proc. 1st ACM Workshop Res. Enterprise Netw. (WREN)*, 2009, pp. 73–82.

[50] *DPDK*. Accessed: Apr. 10, 2021. [Online]. Available: https://www.dpdk.org/

[51] E. Jeong *et al.*, "mTCP: A highly scalable user-level TCP stack for multicore systems," in *Proc. USENIX NSDI*, 2014, pp. 489–502.

[52] G. Kumar, N. Dukkipati, K. Jang, H. M. Wassel, X. Wu, and B. Montazeri, "Swift: Delay is simple and effective for congestion control in the datacenter," in *Proc. ACM SIGCOMM*, 2020, pp. 514–528.

[53] P. X. Gao, A. Narayan, G. Kumar, R. Agarwal, S. Ratnasamy, and S. Shenker, "pHost: Distributed near-optimal datacenter transport over commodity network fabric," in *Proc. ACM CoNEXT*, 2015, pp. 1–12.

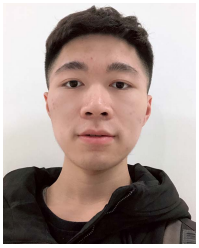
[54] A. Dixit, P. Prakash, Y. C. Hu, and R. R. Kompella, "On the impact of packet spraying in data center networks," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 2130–2138.

[55] S. Hu *et al.*, "Aeolus: A building block for proactive transport in datacenters," in *Proc. ACM SIGCOMM*, 2020, pp. 422–434.

[56] D. Katabi, M. Handley, and C. Rohrs, "Congestion control for high bandwidth-delay product networks," in *Proc. Conf. Appl., Technol., Architectures, Protocols Comput. Commun. (SIGCOMM)*, 2002, pp. 89–102.

[57] N. Dukkipati, "RCP: Congestion control to make flows complete quickly," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2006.
[58] J. Zhang, F. Ren, R. Shu, and P. Cheng, "TFC: Token flow control in data center networks," in *Proc. ACM EuroSys*, 2016, pp. 1–14.
[59] P. Goyal, P. Shah, K. Zhao, G. Nikolaidis, M. Alizadeh, and T. E. Anderson, "Backpressure flow control," in *Proc. USENIX NSDI*, 2022, pp. 779–805.

Jiao Zhang (Senior Member, IEEE) received the bachelor's degree from the School of Computer Science and Technology, Beijing University of Posts and Telecommunications (BUPT), China, in July 2008, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, China, in July 2014. From August 2012 to August 2013, she was a Visiting Student with the Networking Group of ICSI, UC Berkeley. She is currently an Associate Professor with the School of Information and Communication Engineering, BUPT. Her research interests mainly include data center networking, network function virtualization, and future internet architecture.



Xiaolong Zhong (Graduate Student Member, IEEE) received the bachelor's degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China, where he is currently pursuing the master's degree with the State Key Laboratory of Networking and Switching Technology. His current research interests include data center networking and software-defined networking.

Zirui Wan received the bachelor's degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China, where he is currently pursuing the Ph.D. degree with the State Key Laboratory of Networking and Switching Technology. His current research interests include data center networking and software-defined networking.

Yu Tian is currently an Associate Professor with the School of Science, Beijing University of Posts and Telecommunications. She is major in the theory of dynamics system and differential equations, geometric singular perturbation theory, and variational approach.

Tian Pan (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2015. He was a Post-Doctoral Researcher with the Beijing University of Posts and Telecommunications (BUPT) from 2015 to 2017, where he is currently an Associate Professor with the School of Information and Communication Engineering. His research interests include router architecture, software-defined networking, programmable data plane, satellite networks, and machine learning for network applications.

Tao Huang (Senior Member, IEEE) received the B.S. degree in communication engineering from Nankai University, Tianjin, China, in 2002, and the M.S. and Ph.D. degrees in communication and information system from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004 and 2007, respectively. He is currently a Professor with BUPT. His current research interests include network architecture, routing and forwarding, and network virtualization.