RDMA Transports in Datacenter Networks: Survey

Jinbin Hu, Houqiang Shen, Xuchong Liu

***Abstract*—Remote Direct Memory Access (RDMA) has become an important building block of modern datacenter network (DCN) infrastructure given the merits of kernel bypass, zero memory copy techniques and low CPU overhead. However, the increasingly stringent performance requirements of both ultra-low latency and high throughput from the booming datacenter applications raise challenges to large-scale RDMA deployments in DCNs. In this article, we conclude the existing problems of RDMA in Ethernet-based DCNs, and revisit the corresponding efforts of improving RDMA performance in the past decade. We also point out the insights behind these RDMA transport schemes and discuss the future research opportunities with the hope to inspire new interests and shed light on the following research in this field.**

***Index Terms*—Data center, RDMA, transport control, survey.**

I. Introduction

As the datacenter network (DCN) bandwidth continuous to increase, from 10Gbps to 100Gbps or even higher, network transmission becomes the bottleneck for datacenter applications [2, 3]. Specifically, traditional kernel TCP cannot meet the demand of latency-sensitive services like online search or bandwidth-intensive applications like distributed machine learning and cloud storage [6, 11, 15]. In recent years, Remote Direct Memory Access (RDMA) network has become an attractive trend in data centers to meet the above rising stringent demands by providing ultra-low latency, high throughput and low CPU overhead with the kernel bypass and zero memory copy techniques. Nowadays, RDMA networks have been deployed using RDMA over Converged Ethernet Version 2 (RoCEv2) protocol in production data centers such as Microsoft, Google and Alibaba data centers [1, 2, 6, 15].

However, RDMA networks encounter several fundamental challenges during the process of actual deployment and operation [4]. If deploying RDMA directly in Ethernet-based DCNs, RDMA throughput degrades dramatically, because only the simple go-back-N retransmission mechanism is implemented in the RDMA NICs (RNICs) due to limited on-chip resources [6]. Consequently, hop-by-hop Priority-based Flow Control (PFC) is used to enable lossless RDMA networks by avoiding buffer overflow at the switches. However, PFC brings new issues including head-of-line blocking, congestion spreading and deadlock, decreasing RDMA's performance and hurting RDMA's scalability [2, 11, 12]. Recently, efficient and scalable lossy RDMA has attracted extensive attention in academia and industry while redesigning RDMA protocol becomes the new trend [15].

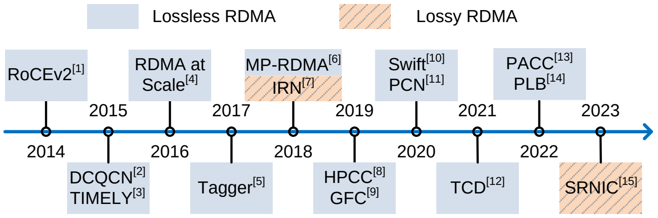

Fig. 1. Timeline of typical RDMA transport solutions in the past decade.

In the light of the above problems in lossless and lossy RDMA networks, many research studies have been working on how to improve RDMA transmission performance in DCNs. Figure 1 shows the timeline of typical RDMA transport solutions in the past decade. Since RoCEv2 [1] was proposed in 2014, these works can be divided into three categories from the technical perspective, as shown in Figure 2. First, DCQCN [2], TIMELY [3], IRN [7], HPCC [8], Swift [10], PCN [11] and PACC [13] focus on designing single-path congestion control mechanism to improve application performance. Second, MP-RDMA [6] and PLB [14] are proposed to make full use of multiple equal-cost paths and alleviate congestion simultaneously. Third, Tagger [5], GFC [9] and TCD [12] are helpful to address specific PFC issues such as head-of-line blocking, PFC storms and deadlock revealed in [4]. The latest work SRNIC [15] has taken a step towards improving RDMA scalability by designing scalable RDMA NIC.

In this article, we conduct a comprehensive survey of RDMA transport techniques from the perspectives of single-path congestion control, multi-path transport, and miscellaneous issues shown in Figure 2. We first present the background and motivations for each research topic, and then discuss their research challenges. Next, we introduce the state-of-the-art RDMA transport solutions and discuss the future research directions for RDMA transports in Ethernet-based DCNs. Finally, we conclude the article.

II. Single-path Congestion Control

A. Background and Motivations

With the link speed growing rapidly from 10Gbps to 100Gbps or higher in DCNs, new flows starting at line rate will aggressively seize the available bandwidth, leading to deep queueing buildup [15]. Moreover, with the datacenter traffic bursty strength increasing continuously, especially in the Partition/Aggregate communication patterns, a bulk of in-flight packets transmitted simultaneously from a large number of distributed source end-hosts to a small number of destinations are easy to cause serious transient congestion.

Manuscript created January, 2023; This work was developed by the IEEE Publication Technology Department. This work is distributed under the LaTeX Project Public License (LPPL) (http://www.latex-project.org/) version 1.3. A copy of the LPPL, version 1.3, is included in the base LaTeX documentation of all distributions of LaTeX released 2003/12/01 or later. The opinions expressed here are entirely that of the author. No warranty is expressed or implied. User assumes all risk.

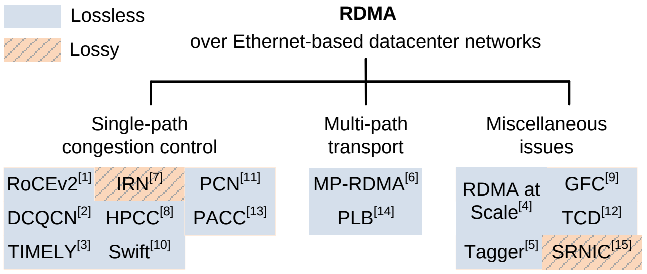

Fig. 2. Categories of typical RDMA transport solutions in the past decade.

Given that RoCEv2 protocol offloaded on the hardware-based RDMA NIC lacks congestion control mechanism, how to detect congestion and adjust sending rate in coordination with RoCEv2 is a crucial research topic. In addition, limited by the hardware resources of the RDMA NIC, the packet loss recovery mechanism in RoCEv2 implements the simple go-back-N strategy, which retransmits all the packets that have not received acknowledge (ACK) packets after the lost packet. The inefficient retransmission causes RDMA to be very sensitive to packet loss. Research shows that RDMA throughput degrades dramatically to nearly zero when the packet loss rate exceeds 0.1% [6], which also makes congestion control particularly important.

Without timely and effective congestion control, the application performances in both lossless and lossy RDMA networks will be seriously damaged due to large tail latency caused by queueing. One line of research concerns the PFC-enabled lossless DCNs [2, 3, 8, 10, 11], once the ingress queue length exceeds a specified threshold, PFC mechanism is triggered to pause the upstream port or a priority class to prevent buffer overflow. This coarse-grained flow control brings a series of negative problems such as the well-known head-of-line blocking, congestion spreading and even extremely serious PFC storms, deadlock, etc. Another line of research focuses on the lossy RDMA networks [7, 15], a lot of lost packets due to congestion will trigger many unnecessary retransmission or even timeouts, leading to large tail latency. Therefore, congestion control is critical to the system performance in data centers.

B. Challenges

Congestion control has always been a crucial way to improve the datacenter application performance and user experience. However, effective congestion control is not easy to achieve, especially in modern DCNs with increasing link bandwidth. The key challenges are the timeliness and accuracy of congestion control.

The first challenge is timely control congestion. The measurement studies have shown that more than 60% of the flows in realistic datacenter workloads can be finished just in one RTT or a few RTTs [14]. It is difficult to control these tiny flows in an end-to-end manner. Although the above existing single-path congestion control protocols effectively alleviate persistent congestion, they are not able to handle transient congestion well. The reason is that these end-to-end solutions require at least one RTT control loop from congestion sensing to sending rate adjustment based on the congestion feedback. Therefore, how to improve tolerance for bursty traffic, mitigate the instantaneous congestion in time, shorten the period from congestion detection to congestion control taking effects, and eliminate PFC triggering in lossless networks or reduce congestion losses in lossy networks are extremely challenging research topics.

The second challenge is accurately control congestion. First, it is not trivial to choose appropriate signals to convey congestion correctly. At present, congestion signals such as ECN, RTT and INT have their own pros and cons. Specifically, ECN is binary signal that is marked as long as the queue length of one hop exceeds a preset threshold. Thus, ECN can only indicate that there is congestion on the path, but cannot reflect the congestion degree at each local switch. RTT variation also makes it difficult to distinguish whether the delay is caused by in-network congestion or processing delay at the end-hosts. INT technology introduces non-negligible additional traffic overhead. Therefore, accurate congestion detection requires to take multiple signals indicating congestion into consideration. Second, distinguishing the congested flows that really responsible for the congestion is quite challenging due to multiple congestion bottlenecks. Third, performing accurate rate adjustment for the congested flows is also quite challenging. Existing solutions generally require multiple RTTs for rate evolution to converge to the target sending rate accurately.

In brief, it is quite challenging to detect and control congestion timely and accurately in RDMA DCNs. Especially under the bursty congestion scenarios, the rate adjustment of end-to-end congestion control is difficult to match with the fast operations of hop-by-hop PFC in the current mainstream lossless networks.

C. Existing Solutions

To enable RDMA networks operate effectively, many research efforts have been spent on designing efficient congestion control to reduce PFC triggering or packet dropping in lossless or lossy RDMA networks in the past few years.

For Lossless RDMA Networks. As the earliest representative RDMA congestion control protocol, DCQCN [2] leverages Explicit Congestion Notification (ECN) marked packets to indicate queue length at the switches and uses Congestion Notification Packets (CNP) generated at the receiver to inform senders to adjust sending rate. DCQCN has been implemented in Mellanox ConnectX series NIC and is being deployed in Microsoft's data centers. TIMELY [3] and its evolution protocol Swift [10] are deployed in Google data centers, they use the Round Trip Time (RTT) variation measured in the hardware NIC with microsecond accuracy to predict congestion and correspondingly control the sending rate at the end-hosts. HPCC [8] is presented as a high precision congestion control, it employs in-network telemetry (INT) technology supported by new switching Application Specific Integrated Circuit (ASIC) to obtain precise link utilization and compute accurate flow rate without gradual evolution and convergence process.

However, the above solutions do not recognize the congested flows that really responsible for congestion, they treat all flows passing through the congestion point equally, resulting in unreasonable rate adjustment. To further control traffic precisely, PCN [11] is the first congestion control protocol that identify real congested flows and make discriminative rate adjustment for congested and uncongested flows. Moreover, PCN adopts a receiver-driven rate adjustment scheme to alleviate congestion quickly within one RTT.

For Lossy RDMA Networks. Unlike the above congestion control mechanisms, which assume working in a lossless networking fabric, the recently proposed IRN [7] operates in the lossy RDMA networks without enabling PFC and its negative impacts. IRN redesigns a more efficient packet loss recovery mechanism and bounds the number of in-flight packets for each flow by the bandwidth-delay product (BDP) of the network on the RoCE NIC with low overhead. The improved packet selective retransmission and packet level flow control method make RDMA more resilient to the large-scale lossy DCNs. However, the performance in terms of convergence and fairness needs further exploration and improvement.

D. Future Opportunities

First, based on the respective advantages of various signals that can piggyback more different information, synthesizing multiple signals to make them complementary and jointly indicate congestion is a feasible congestion detection mechanism. In addition, supported by the advanced NIC and in-network programmable switching ASICs, combining multiple congestion feedback points with collaborative notification to distinguish real congested flows and locations where congestion actually occurs in a fine-grained manner is a possible method worth trying.

Second, the existing end-to-end congestion control protocols have not yet solved the challenges of controlling bursty congestion. However, in practice, a large number of small flows send all the data before the congestion control schemes take effects. Therefore, designing new resilient congestion control protocols collaboration with in-network devices and end-hosts to alleviate transient congestion due to bursty tiny flows is an important direction. Thanks to the vigorous development of machine learning methods and strong support of computing resources, adaptive adjusting the congestion control method or tuning congestion parameters for time-varying network environments and specific traffic characteristics generated by various applications is also a possible direction.

Third, based on the hardware offloading technologies and sufficient on-chip computing and storage resources of the hardware such as Field Programmable Gate Array (FPGA), it is possible to design better hardware-based transport protocols to control congestion with low CPU overhead.

III. Multi-path Transport

A. Background and Motivations

Modern datacenter networks such as Clos topologies provide massive equal-cost parallel paths between end-hosts pairs [6]. The existing single-path transports fall short to take advantage of rich network capacities. In order to make full use of the multiple paths, researchers have proposed a series of multi-path transports and load balancing mechanisms for DCNs [6, 14]. However, these solutions designed for traditional DCNs have poor performance if they are directly employed in RDMA networks.

The specific reasons are as follows: First, aligned with the recent trend of hardware offloading, the entire protocol stack processing sinks to the RDMA NIC. Unfortunately, the hardware on-chip resources of the NIC is limited. If the traditional multi-path transmission protocols such as MPTCP are directly deployed in RDMA networks, they will consume more memory than all the available Static Random Access Memory (SRAM) on the NIC [6]. Because these protocols maintain excessive states for flows and sub-flows, leading to high latency and CPU overhead caused by the additional memory footprint between the host memory and the on-NIC SRAM. Second, the signals sensing path status in traditional load balancing schemes cannot perceive PFC PAUSE/RESUME correctly and timely in the PFC-enabled lossless DCNs. Moreover, the congestion feedback signals are potentially blocked due to PFC pausing. For these reasons, it is likely that the sending rate adjustment and load balancing decisions are incorrect or inaccurate. Third, RDMA networks are more sensitive to lost and out-of-order packets. While the existing multi-path transports cannot preserve order transmission or control disorder degree well while utilizing parallel paths, resulting in larger flow completion time due to serious retransmission. In short, the above issues motivate researchers to redesign multi-path transport for RDMA networks.

B. Challenges

The current trend is to offload RDMA protocols to the hardware NICs. It is quite challenging to design and implement a practical multi-path transports on the NIC to operate efficiently in production DCNs. First, to make congestion-aware load balancing decisions, the multiple transports require to track the congestion states for each path and maintain the information for all the active flows on the hardware with limited computing and storage resources. These states consume considerable hardware resources overhead, and they increase linearly with the number of parallel paths and active flows. Second, due to the diversity of multiple paths in terms of latency, the packets of the same flow allocated on different parallel paths potentially arrive at the receiver out of the order. Especially in high-speed RDMA networks, the simple hardware-based retransmission caused by out-of-order packets will significantly downgrades throughput. Third, in addition to the challenges encountered by the single-path congestion control in RDMA networks, solving the problems of congestion isolation and congestion mismatch among multiple paths for multi-path transmission is also quite challenging.

C. Existing Solutions

The existing solutions for multiple transports fall into two main categories, including multi-path congestion control and load balancing mechanisms. As the first representative work of RDMA multi-path protocols, MP-RDMA [6] focuses on utilizing multiple paths with minimal memory footprint between host memory and RDMA NIC memory to achieve high throughput and low latency. MP-RDMA proposes three new techniques to solve the challenges of limited RDMA NIC memory. First, to avoid per-path status, MP-RDMA employs a multi-path ACK-clocking mechanism to balance load among parallel paths in a congestion-aware manner. Second, to control the degree of disorder, MP-RDMA utilizes an out-of-order aware path selection scheme to shield the congested path with large delay and maintains a bitmap with low storage overhead to track the reordering packets. Third, MP-RDMA designs a synchronize mechanism to guarantee in-order memory updating. However, MP-RDMA is also an end-to-end transport, and it is difficult to deal with the transient congestion due to bursty traffic in a timely manner. In addition, although MP-RDMA discards the concept of sub-flow to reduce memory consumption caused by states maintenance, it leads to poor congestion isolation among multiple paths, leading to congestion mismatch and longer convergence process.

As a complementary mechanism to congestion control mechanism, PLB [14] focuses on balancing load effectively at the switches in PFC-enabled datacenter networks. Through extensively experiments, PLB has pointed out that the existing load balancing solutions for lossy DCNs cannot work well in lossless PFC-enabled RDMA networks, because the congestion signals used to guide rerouting, such as separate ECN, RTT or local queue length, could not accurately and timely feed back PFC pausing and resuming. Thus, PLB proposed a PFC-aware load balancing mechanism by using RTT-level and sub-RTT level signals to reflect path status simultaneously, requiring switches modifications.

D. *Future Opportunities*

Modern datacenter networks enable abundant multiple paths to provide bisection bandwidth. It is helpful to improve throughput and reduce flow completion time by making full use of parallel paths. However, to the best of our knowledge, there are no multiple transports including congestion control protocols and load balancing mechanisms for lossy RDMA networks, and few solutions for lossless DCNs. Therefore, designing efficient and reliable multi-path transports for both lossless and lossy RDMA networks to improve application performance is a future direction.

One of the possible direction is multi-path congestion control focusing on adjusting sending rate to the target one and allocating packets to the appropriate paths simultaneously. Current multi-path transports employ one congestion window for each flow for all parallel paths. Thus, the sending rate is affected by the congestion states of multiple paths, inevitably causing the current rate cannot match the status of all paths. To avoid poor congestion isolation among multiple paths and congestion mismatch between the flow rate and path status, we can design fine-grained congestion control schemes to reasonably adjust the sending rate according to the congestion degree of each path.

The other possible direction is load balancing focusing on rerouting traffic to alleviate congestion without rate adjustment. By leveraging the programmable switches, we can flexibly modify the packets forwarding logic at the data plane for selecting appropriate egress port. Fully considering the path diversities, we can research how to avoid out-of-order packets to guarantee in-order transmission and how to solve the cache miss problem with low memory overhead. Moreover, we can explore the interaction between load balancing and congestion control, and further study which flows need to be adjusted rate to alleviate congestion by injecting fewer in-flight packets, which flows only need to be switched paths to prevent queue buildup by distributing traffic, and which flows need to be rerouted and adjusted rate concurrently.

IV. Miscellaneous Issues

A. *Background and Motivations*

Nowadays, RDMA over Converged Ethernet with hop-by-hop PFC mechanism has been deployed in DCNs at scale by public cloud providers such as Microsoft, Alibaba and Google [2–4, 8, 10]. Although PFC effectively prevents packet loss due to buffer overflow, some serious phenomenons like deadlock and PFC storm occur concomitantly in the practical scenarios. Deadlock is caused by Cyclic Buffer Dependency (CBD), while all switches in this circle wait for each other to resume packet transmission. Once deadlock happens, it cannot break the interdependence loop spontaneously due to all ports in the CBD cease forever. Worse still, a global deadlock or PFC pause frame storm caused by congestion spreading will damage and shutdown the whole network. Due to the above serious problems occurred in the limited-scale networks, the enterprise datacenter operators cannot running RoCEv2 at large-scale DCNs.

To avoid fatal PFC's side-effects and enable large-scale RDMA networks, lossy RDMA DCNs have attracted significant attention recently [7, 15]. Without the guarantee of no packet loss, it is necessary to solve the problem of how to effectively retransmit caused by buffer overflow. Otherwise, a large number of inefficient retransmissions caused by inherent bursty traffic cannot enhance the scalability of RoCE. In addition, the connection scalability is also a crucial problem that the goodput collapses with the increase number of connections built on the RDMA NIC. The root cause of the throughput performance degradation is the cache miss between the limited on-chip SRAM on the NIC and the host memory for context switching.

B. *Challenges*

With the scale of datacenter network growing rapidly and the traffic increasing drastically, the datacenter operators cannot manipulate network configurations well manually to avoid deadlock and congestion spreading issues. Thus, these issues have attracted considerable attentions. However, PFC triggering involves many factors, including the switch buffer size, link bandwidth, number of priority queues, number of congested flows, flow sending rate, number of related ingress ports and PFC triggering threshold. It is very challenging to take all these factors into consideration simultaneously to predict and prevent PFC triggering in distributed networks, and it is more difficult to avoid negative impacts of various PFC problems after PFC triggering.

Traditional RoCEv2 transport performs inefficient go-back-N retransmission and only supports hundreds of connections. However, scalable RDMA transport in lossy RDMA network requires to be quite resilient to packet loss and massive connections. Based on the limited hardware resources on the RDMA NIC, the improved RoCE transport meets a series of new challenges, including redesigning reliable transmission algorithm, optimizing data structure and memory footprint, dealing with cache missing, etc. Moreover, following the trend of programmable transports, the advanced RoCE NIC requires to be programmable for each module in transport protocols to support flexible modification and reconfiguration of the data delivery algorithms such as congestion control and packet loss recovery mechanisms. Another key challenge is that the implementation of transport logic including packet generation and flow states query needs to satisfy the hardware timing constraints and consider the heterogeneity of the hardware infrastructures.

C. Existing Solutions

For Lossless RDMA Networks. To avoid the system traps in deadlock, GFC [9] and Tagger [5] are proposed to solve the cyclic buffer dependency problem by breaking the necessary conditions of deadlock and eliminating CBDs, respectively. Specifically, GFC [9] explores a new solution to break the hold and wait condition by manipulating the sending rate at a fine granularity. The flow control scheme ensures that the input rate at the ingress port matches the draining rate at the egress port and maintains the queue length in a steady state, so that all flows can continue to pass through the network even CBD exists. Tagger [5] reserves the expected lossless paths dynamically and allows the packets that not transmitted on lossless paths to be dropped under extreme congestion scenario, resulting in no CDB formation. In TCD [12], the authors develop a new congestion detection scheme to identify congested ports and flows accurately. TCD defines ternary states of switch ports, including congestion, non-congestion and undetermined states, and captures these states transition based on the queue length evolution.

For Lossy RDMA Networks. To improve network scalability and connection scalability, SRNIC [15] devotes great efforts to designing a scalable RDMA NIC architecture. Guided by the insight that the on-chip data structures and their cache requirements in the RDMA NIC can be minimized by simultaneously optimizing protocol and architecture, SRNIC implements cache-free queue pairs scheduler and memory-free selective repeat to remove involved memory as many as possible. The experiments results show that SRNIC supports 10K performant connections on chip with high throughput and low latency. However, when the number of lost packets increases, the CPU overhead caused by software retransmission also increases.

D. Future Opportunities

First, since the shutdown of the whole network is the most fatal damage for the distributed large-scale DCNs, one possible direction is to timely isolate the negative impacts of the hop-by-hop PFC mechanism in a limited scope of controllable local area networks. Although a series of serious problems such as deadlock and pausing storm caused by PFC triggering under the heavy congestion scenario, PFC can play a good role in a small range in terms of the advantages of simple deployment and effective preventing buffer overflow. Therefore, we can design a hybrid RDMA network by co-designing lossless and lossy networks divided from the perspective of datacenter topologies or switch priority queues. Moreover, it is a good opportunity to design machine learning based algorithms to dynamically tune the parameters in the partition mechanism to adapt to the network congestion changes. In this way, we can avoid the whole network paralyzed and continue to utilize the advantages of PFC to deliver reliable transmission.

Second, the existing solutions for eliminating packet loss impose many restrictions on RDMA network configurations and harmful side-effects on performance. The other direction is to adapt RoCE to the lossy network to improve the scalability of RDMA network. With the help of relatively rich computing and storage resources in the advanced hardware such as FPGA, data processing unit (DPU) and Application Specific Integrated Circuit (ASIC), how to reasonably divide software and hardware functions for RoCE transports, including virtualization, tunneling, connection and buffer management, to combine the slow path and fast path for reliable data transmission directly affect the delay performance of RDMA. Complying with the trend of hardware offloading technique, how to implement a simple and efficient loss recovery mechanism to make RDMA network scalability and how to abstract the common logic modules to provide flexible RDMA transports programming on the high-speed RoCE NIC without hardware resources constraints are the important directions for further exploration and research.

Third, in addition to the efforts on general RDMA transports, domain specific RDMA transports for diverse applications such as storage, distributed machine learning and high-performance computing are also the essential research directions. According to the traffic generation modes and traffic distribution characteristics of special applications, it is desirable to eliminate the transmission bottlenecks caused by the connection between heterogeneous hardware devices. Nowadays, there is a huge disparity in the development pace between network speed, computing capacity and cache size, so how to coordinate hardware resources and how to support direct network connectivity is a quite important issue in DCNs. However, only a few efforts have been made to allow the physical equipments such as FPGA accelerators and GPUs to directly control the RDMA NICs. Therefore, there is still a broad space for exploration and research.

V. Conclution

In this article, we made a survey on the current state-of-the-art of RDMA related work in Ethernet-based datacenter networks. We categorized the existing solutions into three categories, including single-path congestion control mechanisms, multi-path transports and miscellaneous issues, to show a general overview of this active research field common concerned by both academia and industry. We described the background and motivations, discussed the design challenges, reviewed the existing schemes and also pointed out the future research directions hoping to inspire new ways of thinking and make new contributions in this area.

VI. Acknowledgment

This work is supported by the National Natural Science Foundation of China (62102046, 62072056), the Natural Science Foundation of Hunan Province (2022JJ30618, 2023JJ50331), the Scientific Research Fund of Hunan Provincial Education Department (22B0300) and by Hunan Provincial Key Research and Development Program (2022GK2019).

References

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

[2] Y. Zhu, H. Eran, D. Firestone, et al. Congestion Control for Large-Scale RDMA Deployments. In Proc. ACM SIGCOMM, 2015.

[3] R. Mittal, V. T. Lam, N. Dukkipati, et al. Timely: RTT-based Congestion Control for the Datacenter. In Proc. ACM SIGCOMM, 2015.

[4] C. Guo, H. Wu, Z. Deng, et al. RDMA over Commodity Ethernet at Scale. In Proc. ACM SIGCOMM, 2016.

[5] S. Hu, Y. Zhu, P. Cheng, et al. Tagger: Practical PFC Deadlock Prevention in Data Center Networks. In Proc. ACM CoNEXT, 2017.

[6] Y. Lu, G. Chen, B. Li, et al. Multi-Path Transport for RDMA in Datacenters. In Proc. USENIX NSDI, 2018.

[7] R. Mittal, A. Shpiner, A. Panda, et al. Revisiting Network Support for RDMA. In Proc. ACM SIGCOMM, 2018.

[8] Y. Li, R. Miao, H. H. Liu, et al. HPCC: High Precision Congestion Control. In Proc. ACM SIGCOMM, 2019.

[9] K. Qian, W. Cheng, T. Zhang, F. Ren. Gentle Flow Control: Avoiding Deadlock in Lossless Networks. In Proc. ACM SIGCOMM, 2019.

[10] G. Kumar, N. Dukkipati, K. Jang, et al. Swift: Delay is Simple and Effective for Congestion Control in the Datacenter. In Proc. ACM SIGCOMM, 2020.

[11] W. Cheng, K. Qian, W. Jiang, T. Zhang, F. Ren. Re-architecting Congestion Management in Lossless Ethernet. In Proc. USENIX NSDI, 2020.

[12] Y. Zhang, Y. Liu, Q. Meng, F. Ren. Congestion Detection in Lossless Networks. In Proc. ACM SIGCOMM, 2021.

[13] X. Zhong, J. Zhang, Y. Zhang, Z. Guan, Z. Wan. PACC: Proactive and Accurate Congestion Feedback for RDMA Congestion Control. In Proc. IEEE INFOCOM, 2022.

[14] J. Hu, C. Zeng, Z. Wang, H. Xu, J. Huang, K. Chen. Load Balancing in PFC-Enabled Datacenter Networks. In Proc. ACM APNet, 2022.

[15] Z. Wang, L. Luo, Q. Ning, et al. SRNIC: A Scalable Architecture for RDMA NICs. In Proc. USENIX NSDI, 2023.

Jinbin Hu (jinbinhu@csust.edu.cn) received the B.E. and M.E. degrees from Beijing Jiao Tong University, China, in 2008 and 2011, respectively, and the PhD degree in computer science from Central South University, China, in 2020. She is currently a Post-Doc in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, and working in the School of Computer and Communication Engineering, Changsha University of Science and Technology, China. Her current research interests are in the area of datacenter networks and distributed systems.

Houqiang Shen (shq@stu.csust.edu.cn) is currently pursuing the M.S. degree in the School of Computer Science and Engineering at Changsha University of Science and Technology, China. His research interests are in the areas of datacenter networks.

Xuchong Liu (14117874@qq.com) is a professor at the Department of Information Technology of Hunan Police Academy. He is currently the director of the Information Technology Department of Hunan Police Academy, the director of the Hunan Provincial Key Laboratory of Network Investigation Technology, the director of the Hunan Provincial Key Laboratory of Network Crime Investigation, etc. His current research interests include Network Crime Investigation, Electronic Evidence Collection, Data Center Networking and Machine Learning Systems.