SWING: Providing Long-Range Lossless RDMA via PFC-Relay

Yanqing Chen, Chen Tian, Jiaqing Dong, Song Feng, Xu Zhang, Chang Liu, Peiwen Yu, Nai Xia, Wanchun Dou, and Guihai Chen

Abstract—Remote Direct Memory Access (RDMA) has been widely deployed in datacenters for its high performance. Large-scale high performance cloud services built on geographically distributed datacenters require long-range RDMA for performance requirements. However, existing RDMA solutions can hardly satisfy the stringent requirements of the emerging large-scale high-performance cloud services built on geo-distributed datacenters in terms of throughput and delay. On the one hand, lossless RDMA suffers from a deep buffer and potential suboptimal throughput for inter-datacenter traffic due to delayed response to Priority Flow Control (PFC) messages. On the other hand, lossy RDMA with selective retransmissions suffers from poor performance when multiple flows with different round-trip times (RTTs) coexist in cross-datacenter scenarios. This article proposes SWING , which expands the high-performance lossless RDMA to long-distance links through PFC-Relay. SWING ensures the throughput of long-distance links while minimizing the buffer requirement for long-range RDMA. It enables long-range RDMA without making any modifications to existing in-datacenter networks. The evaluation shows that SWING can reduce the average flow completion time (FCT) by 14%-66% in a variety of traffic scenarios.

Index Terms—Inter datacenter communication, Datacenter networks, Flow control, PFC, RDMA

1 INTRODUCTION

RDMA has been widely adopted by high performance computing (HPC) systems [1], [2]. It provides applications with ultra-low latency, high throughput, and low CPU resource consumption. It has become a trend for datacenters to deploy RDMA to enhance the performance of the accommodated cloud services. Solutions for building RDMA networks in datacenters generally fall into two categories: lossless RDMA and lossy RDMA. RDMA over Converged Ethernet v2 (RoCEv2) [3] is a typical lossless solution. It is compatible with IP/Ethernet and requires the PFC mechanism to ensure a lossless network. In terms of lossy solutions, the improved RoCE NIC (IRN) [4] is proposed as a selective retransmission scheme for building RDMA over lossy networks [4], [5], [6].

Large-scale high performance cloud services built on geographically distributed datacenters require long-range RDMA. First, more and more services provided by datacenter operators like Amazon [7], Microsoft [8], Google [9], and Facebook [10], are deployed across multiple regionally connected small-scale datacenters. These datacenters are connected by dedicated optical cables directly, which are different from wide area networks (WAN). This gives the services the ability to establish RDMA connections over datacenter interconnection (DCI). Second, these services usually adopt RDMA to communicate inside the datacenter for stringent performance requirements. When these services communicate across these datacenters, it is better to continue using existing RDMA technology to achieve single-connection high throughput over long-distance links. In addition, using RDMA can maintain API consistency and reduce the complexity of application deployment across datacenters. There are several requirements for a solution that provides long-range RDMA. First, the solution needs to be compatible with the existing RDMA technologies. It should not interfere with existing network protocols. Second, the solution should keep the existing network equipment inside the datacenter unchanged. Modifications to existing infrastructures should be avoided. Third, it is expected to fully utilize the long-distance links and provide the upper-layer applications with high performance. Last but not least, it should use as less buffer as possible to avoid long queuing delays. However, existing RDMA solutions cannot meet these requirements. Typically, most datacenters usually adopt deep buffer switches for inter-datacenter connections. These switches

- *Yanqing Chen, Chen Tian, Chang Liu, Peiwen Yu, Nai Xia, Wanchun Dou, and Guihai Chen are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China. E-mail: dz1933003@smail.nju.edu.cn, {tianchen, xianai, douwc, gchen}@nju.edu.cn, liuchang_1307@163.com, pavin0702@foxmail.com.*
- *Jiaqing Dong is with the State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China. E-mail: jiaqing.dong@outlook.com.*
- *Song Feng is with Network Information Center, Xiangya Hospital, Central South University, Changsha 410017, China. E-mail: fs205@sina.com.*
- *Xu Zhang is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210093, China. E-mail: xzhang17@nju.edu.cn.*

Manuscript received 1 April 2022; revised 12 October 2022; accepted 12 October 2022. Date of publication 19 October 2022; date of current version 16 November 2022.
This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B0101390001, in part by the National Natural Science Foundation of China under Grants 92067206, 62072228 and 61972222, in part by the Fundamental Research Funds for the Central Universities, in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization, and in part by Jiangsu Innovation and Entrepreneurship (Shuangchuang) Program.
(Corresponding authors: Chen Tian and Song Feng.)
Recommended for acceptance by K. Gopalan.
Digital Object Identifier no. 10.1109/TPDS.2022.3215517

are more complex and expensive than shallow-buffer switches. In addition, the deep buffer will introduce long queuing delays for RDMA connections. For shallow-buffered switches, the PFC mechanism can cause throughput loss. Even though connected by dedicated optical fibers, a long distance between two datacenters introduces a long propagation delay with a large Bandwidth Delay Product (BDP). The long propagation delay causes PFC RESUME unable to take effect on the upstream switch in time. The large BDP also makes the shallow buffer of the downstream switch drain empty before the resumed traffic from the upstream switch arrives, thus resulting in throughput loss. In addition, lossy RDMA does not fit the long-range scenario either. IRN [4] is a typical lossy RDMA solution. It maintains several sets of registers as bitmaps to track the status of each packet. The optimal size of the bitmap is determined according to the RTT of the flow. Therefore, the fixed bitmap design of IRN makes it impossible to guarantee the performance of flows with different RTTs at the same time. A dynamic bitmap design of IRN could be a good solution, but it requires upgrading all the related network interface cards (NICs). And for large BDPs, the resource occupied by the bitmap on the NIC is also a considerable overhead.

This paper proposes SWING , which provides long-range lossless RDMA via the PFC-relay mechanism. SWING plugs a "relay" device at each end of the long-distance link close to the external switch. Frequent cyclic PFC signals generated by the receiving side can enforce the local relay device to send data at the draining rate of the remote switch. The remote relay device will also be enforced by the delayed cyclic PFC signals to send data at the draining rate of the local switch. In this way, the SWING transparently "relays" these cyclic PFC signals over the long-distance link for both sides. Furthermore, the relay device in SWING only requires half of the buffer on each port compared to the DCI switch. The proposed solution satisfies all the aforementioned four requirements of long-range RDMA. First, cross-datacenter applications still use RDMA communications while no conversion is required in the middle. Second, the relay is only deployed on the long-distance link and does not affect the network inside the datacenter. Third, SWING can keep the long-distance links fully utilized. Finally, as analyzed in Section 5.2 , the relay only needs half of the buffer required by deep buffer switch solutions. It is worth noting that SWING only modifies the flow control mechanism of the long-distance links. The RDMA congestion control algorithm remains unchanged. The contributions of this paper can be summarized as follows. We theoretically analyze the native PFC and prove its throughput loss in long-range RDMA. We propose the PFC-relay mechanism for inter-datacenter RDMA connection without making modifications to the existing network inside datacenters. Based on the PFC-relay mechanism, we propose our solution, SWING , which supports long-range RDMA and eliminates the throughput loss with half the required buffer size compared to native PFC. SWING enables existing inter-datacenter connections to support high-performance long-range RDMA without making any modifications to in-datacenter networks. The evaluation results demonstrate that SWING can reduce the average FCT for inter-datacenter traffic by 14% - 66%.

2 BACKGROUND

2.1 RDMA in Datacenter

RDMA is a networking technology that provides high-performance data transmission. Compared with traditional network transmission, RDMA can directly transmit data to the memory of the remote node by bypassing memory replications and interruptions in the operating system on both sides. It can easily achieve low latency and high throughput without huge resource consumption such as CPUs.

InfiniBand [11], one kind of RDMA technology, is originated in the HPC community first, where applications are homogenous, highly parallel, and require both high bandwidth and low latency communication between nodes. Then with the increasing network demand for high-performance services in the datacenter, operators are looking for a solution to deploy RDMA into datacenters. However, due to the closed architecture and high costs of InfiniBand, it is hard to deploy with the existing networking infrastructure. In addition, there will be two separate networks if the operator deploys both Ethernet and InfiniBand in one datacenter [12]. So another solution, RoCEv2, is based on IP/Ethernet which can be deployed in a three-layer network. As IP/Ethernet is still the dominant communication network in the datacenters, datacenters deploy RoCEv2 at scale [5], [13]. iWarp [14] supports RDMA with the TCP (Transmission Control Protocol) stack built inside the network card. It can support RDMA over a more general network due to the widely adopted TCP. However, due to the complexity of TCP, iWarp is more expensive than the alternative RoCEv2 to achieve the same performance [15].

Lossless RDMA. In datacenters that have supported RDMA for years, they need a lossless network to ensure low latency and high throughput. Both InfiniBand and RoCEv2 leverage hop-by-hop flow control mechanisms to guarantee zero packet loss [16], [17]. Specifically, InfiniBand clusters use credit-based flow control to make the network lossless [18], while RoCEv2 relies on the PFC mechanism to control the behavior of the sender to avoid packet loss [19]. The PFC mechanism is proposed by Data Center Bridging (DCB) task group [16] to ensure lossless Ethernet for RoCEv2 [20]. PFC sets two thresholds in the ingress buffer of the switch, namely `XOFF` and `XON`. When the ingress queue length exceeds the threshold `XOFF`, the receiver can send a `PAUSE` message to pause the sender's transmission, and when the ingress queue length drops below the threshold `XON`, the receiver can resume transmission by sending a `RESUME` message. The difference between the total ingress buffer and the `XOFF` threshold is the headroom buffer, which is used to absorb in-flight packets before the `PAUSE` message takes effect. Consequently, if the headroom is greater than the BDP, the switch buffer will not overflow.

Since the PFC pause mechanism causes packets to stay in the switch, it incurs the following problems. First, the pause mechanism damages some innocent flows that do not cause congestion, which leads to the head-of-line (HOL) blocking and the PFC storm [13]. Furthermore, the pause mechanism causes queue lengths to accumulate and propagate to other switches, triggering the pause mechanism again. Second, cyclic buffer dependency (CBD) can lead to deadlocks [13], [21], [22]. In this case, on this cycle

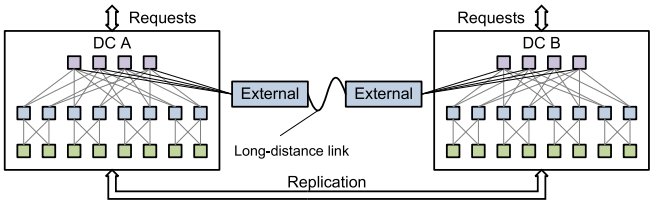


Fig. 1. Cross-datacenter application communication activities in a typical active-active architecture.

the ports are paused due to PFC, waiting for upstream PFC messages to resume transmission.

Lossy RDMA. Solutions have been proposed for lossy RDMA. Mellanox supports RoCE Selective Repeat to recover from packet loss [23]. Similarly, IRN [4] has been proposed to build lossy RDMA. Because the implementation of RoCE Selective Repeat is closed source, we mainly discuss IRN for lossy RDMA. There are two reasons why IRN can achieve better performance than the lossless solution. First, IRN uses the selective retransmission mechanism. The receiver sends a SACK message after receiving an out-of-order packet to notify the sender to retransmit the lost packets. As complementation, a timeout mechanism is used to cover edge cases. For the short flow (such as remote procedure call requests) whose length is less than three MTU, the timeout RTO_{low} of each packet is set to one RTT to achieve fast retransmission. For other flows, the timeout RTO_{high} of each packet is set to the RTT including full queuing time to detect packet loss and avoid over-transmission. The second is to use a static bitmap that bounds the number of in-flight packets to just fill a BDP. This mechanism can reduce congestion in the network.

2.2 Cross-Datacenter Applications

Large-scale high performance cloud services are being deployed across multiple datacenters. When the number of users reaches a certain scale, service providers will face the challenge of high concurrency and massive data. However, for a single datacenter, there is an upper limit on the computing power and storage capacity. The cross-datacenter deployment becomes an important solution. First, the regional datacenter is able to provide better performance services for local customers. Second, this solution can keep the continuity of services and enhance the ability to resist risks. Under the influence of natural disasters such as typhoons, earthquakes, and floods, datacenters may face power outages and network interruptions. Human factors such as software errors and configuration mistakes may also cause the services in a single datacenter to be unavailable [24], which in turn leads to overall service paralysis. For cross-datacenter services, the distance between the two connected datacenters is restricted by the regional service level agreement (SLAs). The length of the optical fiber connecting the datacenters is typically limited to 120 km [8]. The propagation delay of the optical fiber is about 5 μs per kilometer, so the one-way delay can be up to 600 μs. These fibers are dedicated to the inter-datacenter traffic, while not responsible for the traffic of WAN. There are several solutions proposed for connecting multiple datacenters [25]. Regarding the equipment for establishing the datacenter interconnection, the traditional solution is using a router or a DCI switch. Take the two-datacenter scenario in Fig. 1 as an example. Each datacenter has an external switch that connects to its spine switches. Then a transceiver plugged into the external switch converts electrical signals into optical signals, and then connects to the other side through Dense Wavelength Division Multiplexing (DWDM) [26].

2.3 Long-Range RDMA Requirements

It has become a trend for high performance services in datacenters to adopt RDMA to satisfy the stringent performance requirements in terms of low latency and high bandwidth [1], [2], [6]. When these services are deployed across multiple datacenters, long-range RDMA support is required.

For network performance, there have been some works comparing RDMA and TCP over long-distance links. The results in [27] show that a link packet loss rate of 0.001% will cause TCP throughput to drop sharply in a congested scenario. In contrast, RDMA can maintain high throughput under lossless networks. Another choice is iWarp, which has not been widely deployed in datacenters. Because there are some well-known complexities and problems of TCP in the datacenter environment, iWarp cannot perform as well as RoCEv2 in the same generation [28].

For applications, there are three reasons that RDMA is required compared to the traditional TCP. First, applications deployed across datacenters require real-time synchronization, such as databases in the transaction system. [29] shows that RDMA consumes much fewer CPU resources than TCP for each connection. The saved CPU resources can be used to process more transactions to improve the response time of the application. Second, applications like remote data access, large file transmission, and video streaming require the network to transmit large amounts of data. The single-flow throughput of TCP is lower than that of RDMA [29]. The high throughput provided by RDMA can reduce the data transmission time. Third, maintaining the usage of RDMA for cross-datacenter communications can help applications or datacenter operators avoid implementing two sets of communication interfaces. In addition, iWarp [14] involves the modification of the datacenter infrastructure.

For deployment, long-distance RDMA is mature in HPC to interconnect high-performance computing clusters. ESnet [30] connects cross-state and cross-border sites through RDMA over Infiniband [31]. To support the use of Infiniband on long-distance links, Obsidian [32], Vcinity [33], and Mellanox [34] have launched corresponding products. Although these technologies are based on Infiniband and cannot be used directly in the datacenter, they provide a reference for deploying RDMA between datacenters. On long-distance links, the benefits like high-throughput and low-CPU resource consumption that RDMA can provide are still attractive for computing-intensive and IO-intensive applications.

3 Related Work

RDMA Over WAN. There has been a lot of work on the measurement of RDMA on WAN. Researchers in [27] aim at implementing RDMA on GridFTP [35]. To evaluate the performance, they test the performance of RDMA and TCP

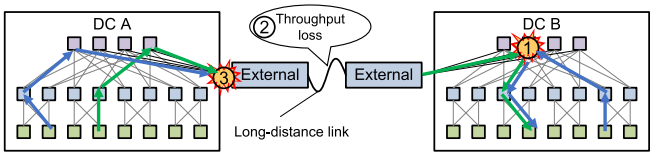


Fig. 2. Problems using PFC and IRN in long-range RDMA.

under different loss rates and WAN delays under a 10 Gbps network. In the scenario of 120 ms delay and no packet loss, the single-stream throughput of RDMA is 20% higher than that of TCP. When the packet loss rate rises to 0.001%, this advantage expands to 300%. Another work [29] is dedicated to measuring the performance of RDMA over WAN. Its main background is the data transmission between Science DMZs and the research on the factors that affect transmission performance. They test the throughput of various protocols like TCP, TCP-splice, RoCE, and applications like xfer_test, GridFTP, and netperf [36] directly in ESnet [30], where xfer_test is a benchmarking tool developed by the researchers in this work. The results show that RDMA is better than TCP in terms of throughput and CPU utilization in all applications except that GripFTP supports RDMA imperfectly. There is also a work [37] to test the performance of HPC middleware under WAN. This work evaluates the performance in terms of RTT, RDMA connection methods (such as Reliable Connection and Unreliable Datagram), and RDMA message size. They point out the impact of message size on throughput under long-haul links but did not associate throughput with the switch buffer size and link length. These works laid the groundwork for us to deploy RDMA over long-distance links.

Flow Controls for RDMA. Many works have observed that PFC has many side effects under congestion conditions. Researchers in [38] utilize the ingress and egress queue statistics to improve PFC. The idea mainly comes from Quantized Congestion Notification (QCN), which detects the length of the egress queue and determines the flow that contributes most to the congestion. They alleviated the HOL blocking problem and reduced the flow completion time. A work [39] uses the rate of change of the queue length to send PFC messages instead of using a fixed threshold. This predictive mechanism enables it to effectively reduce the switch queue length and protect innocent flows to a certain extent. The result of reducing queue length also leads to improved tail latency. Another work [22] solves the deadlock in lossless networks, which lets the PFC message directly control the link rate by mapping to the queue length. However, these solutions need to modify the switch PFC triggering algorithm in the datacenter and do not consider the performance on long-distance links.

4 Analysis

This section analyzes potential solutions for deploying long-range RDMA between datacenters and figures out their limitations respectively. The first one is the PFC mechanism, which is widely adopted by RoCEv2 for building a lossless network. The second one is lossy RDMA, of which we take IRN [4] as the representative.

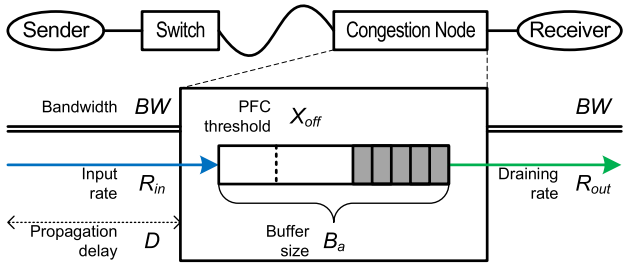


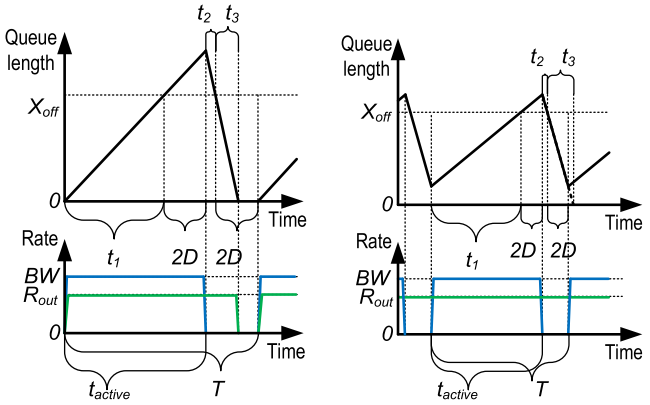
Fig. 3. The theoretical model of the PFC mechanism.

4.1 PFC Needs Deep Buffer

Cross-datacenter applications generate two types of traffic. Inter-datacenter traffic and intra-datacenter traffic will compete with each other, causing congestion and backpressure on long-distance links. Without the deep buffer provided by a high-end switch, the PFC mechanism can result in reduced throughput, which is shown in Figs. 2 ① and 2 ②. To find the reason, two questions need to be answered. The first question is: *What is the relationship between the link length and the switch buffer size to guarantee zero packet loss?*

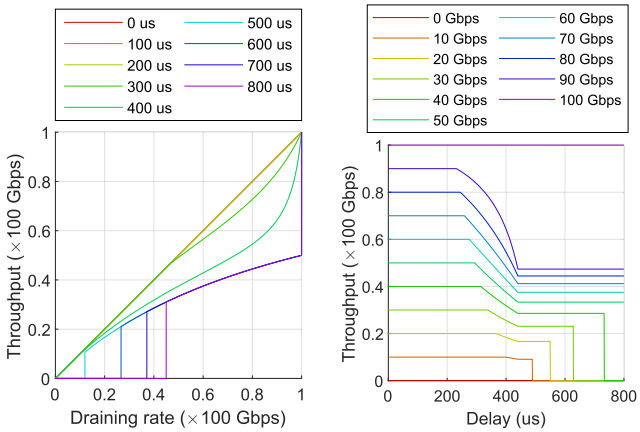
Although PFC can guarantee zero packet loss in the datacenter, there is a precondition, that is, the headroom needs to be set to at least one BDP to absorb the in-flight packets to avoid packet loss. In detail, the propagation delay introduced by a long-distance link decides how long the PFC signal takes effect. After the ingress queue length of the downstream switch exceeds the threshold `XOFF`, the switch sends `PAUSE` to the upstream switch. It will take a propagation delay to arrive at the upstream switch. After `PAUSE` reaches the upstream switch port, it stops sending data. The last packet sent from the upstream switch takes another propagation delay to arrive at the downstream switch. However, we find that the buffer of one BDP is not enough to guarantee the throughput. Here comes the second question: *What is the relationship between the link length and the switch buffer size to guarantee the throughput?*

As shown in Fig. 3, the transmission of a pair of nodes passes through a bottleneck node. The flow control works on the switch ingress port. For convenience, only one priority queue is modeled here. The buffer allocated to this queue is B_a. The link bandwidth is BW. The long-distance link propagation delay is D. The input rate of this queue is R_{in} and its draining rate is R_{out}. The PFC threshold `XOFF` is set to X_{off}. In practice, `XON` is lower than `XOFF` to reduce the queue length. To ensure maximum throughput, we set `XON` and `XOFF` to be the same. Here are two auxiliary conditions that do not affect the conclusion of this question: (i) The draining rate R_{out} is a fixed value, and the ratio to the bandwidth is α, that is, $R_{out} = \alpha \cdot BW$. For traffic on short-distance links, the frequency of triggering PFC is much higher than that on long-distance links. Therefore, we regard R_{out} as the average value over a while. Otherwise, we cannot track changes in queue length. (ii) The input rate R_{in} is alternated between 0 and BW by PFC. This is based on the on-off traffic pattern on the long-distance link caused by the PFC mechanism. Our purpose is to use the least buffer to support the longest link under any congestion without throughput loss. B_a and BW are properties of the



(a) Throughput loss. (b) No throughput loss.

Fig. 4. Changes in queue length and throughput of the congested node with PFC enabled.



(a) Throughput under different draining rates. (b) Throughput under different delays.

Fig. 5. With a fixed total buffer of 11 MB, the best performance of PFC under different congestion and link propagation delays.

switch. X_{off} has its optimal setting scheme. Therefore the model variables are D and α.

With these conditions, we can easily draw schematic diagrams of the queue length and long-distance link throughput, as shown in Fig. 4. The blue curve is the input rate and the green curve refers to the draining rate. The figures show two cases. In the first case, the throughput of the long-distance link is dropped, where the draining rate is zero for a period of time. In the second case, there is no throughput loss. We will now calculate how much throughput has been wasted. First, we need to define the time to empty the queue from the PFC threshold

$$t_3 = \frac{X_{off}}{\alpha \cdot BW}, \tag{1}$$

which can be used to separate these two situations in Fig. 4.

Next, we define the time it takes for the queue length to grow from the lowest value to reach the PFC threshold XOFF in each cycle. When $t_3 < 2D$, as shown in Fig. 4a, the queue length grows from 0. When $t_3 \geq 2D$, as shown in Fig. 4b, the queue length increases after a $2D$ decrease from X_{off}. The minimum queue length is $X_{off} - 2D \cdot \alpha \cdot BW$. After simplification, the following formula can be obtained

$$t_1 = \begin{cases} \frac{X_{off}}{(1-\alpha)BW} & \text{if } t_3 < 2D, \\ \frac{2D \cdot \alpha}{(1-\alpha)} & \text{if } t_3 \geq 2D. \end{cases} \tag{2}$$

Then the receiver will send PAUSE back to the upstream, which will take effect after one RTT, consuming a headroom of $2D(1-\alpha)BW$. Therefore, the time it takes for the queue length to fall back to the PFC threshold is

$$t_2 = \frac{2D(1-\alpha)}{\alpha}. \tag{3}$$

Finally, we have the transmission time on the long-distance link $t_{active} = t_1 + 2D$ and the time of one cycle $T = t_1 + 2D + t_2 + 2D$. In order to maximize the throughput and buffer utilization, the PFC threshold is set to $X_{off} = B_a - 2D \cdot BW$. In addition, to ensure zero packet loss, the maximum link propagation delay is $B_a/(2BW)$. In other words, all buffers are used to store in-flight packets. Under these two constraints, the utilization ratio of the long-distance link is given by the following formula

$$\phi(D,\alpha) = \frac{t_{active}}{T} = \begin{cases} \frac{\alpha(2BW \cdot D \cdot \alpha - B_a)}{2BW \cdot D(-1+\alpha+\alpha^2) - \alpha \cdot B_a} & \text{if } t_3 < 2D, \\ \alpha & \text{if } t_3 \geq 2D.. \end{cases} \tag{4}$$

According to this formula, it can be verified that $\phi(D,\alpha) < \alpha$ when $t_3 < 2D$. Therefore, the boundary condition to ensure that the throughput is not compromised is $t_3 \geq 2D$, that is, $B_a \geq 4D \cdot BW$, which equals to $2 \times BDP$. In other words, the PFC mechanism requires a minimum buffer of two BDPs to ensure the performance.

For a more intuitive understanding, we assume that the buffer that can be allocated to the long-distance link port is 11 MB, and the port bandwidth is 100 Gbps. The long-distance link throughput varies with the link propagation delay and the draining rate as shown in Fig. 5. Note that we set the PFC threshold close to zero when the headroom is smaller than one BDP. In addition, we set the throughput to zero when there is packet loss. It can be seen that when the link propagation delay is greater than 200 μs, the throughput is less than the draining rate.

Now we can answer the second question. When the PFC mechanism is enabled, to ensure that the link throughput does not decrease, each port needs to allocate a buffer of *at least* twice the BDP. The headroom size is at least one BDP to ensure zero packet loss. The remaining buffer is used to compensate for the link transmission idling caused by the RESUME message transmission delay and the minimum size for this part is one BDP. For cross-datacenter services, two datacenters have a distance of tens of kilometers or even hundreds of kilometers. For each 100 Gbps link, each port requires at least 125 KB of buffer per kilometer. To establish a lossless link using PFC between the two datacenters and ensure that the throughput does not decrease, we need deep buffer switches, which inevitably introduce high queuing delays.

4.2 IRN Hurts Performance

Unlike the PFC mechanism, IRN will not pause the traffic on the long-distance link when there is competition between the long-distance traffic and the traffic in the datacenter. Instead, it will drop packets at the congestion node. This mechanism can potentially solve the head-of-line blocking,

"pause spreading" and other problems mentioned in many papers in recent years caused by PFC [5], [13], [21].

To achieve good performance, IRN needs a difficult-to-congest topology and an appropriate bitmap size. However, these conditions are difficult to be satisfied in cross-datacenter scenarios. First, the inter-datacenter bandwidth is expensive and thus usually smaller than the intra-datacenter bandwidth. If one datacenter sends burst traffic to another one, the external switch will be congested and drop packets, as shown in Fig. 2 ③. In addition, the traffic that has reached the remote datacenter will compete with the internal traffic of that datacenter, as shown in Fig. 2 ①. Both of these situations show that congestion is easy to occur in the cross-datacenter scenario.

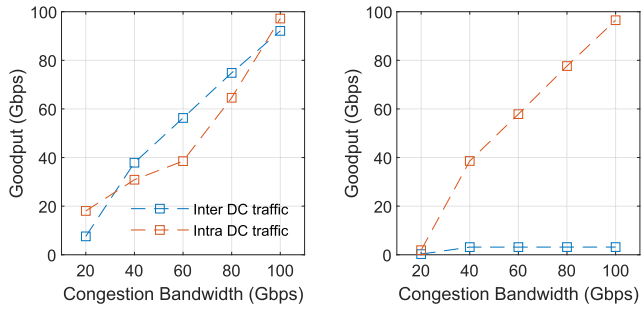
Second, there are two kinds of traffic in cross-datacenter services. One is intra-datacenter traffic, and the other one is inter-datacenter traffic. The bitmap in IRN is used to track in-flight packets, which is static. And the optimized size of the bitmap is just one BDP of the network. As the BDPs of the inter-datacenter link and intra-datacenter link differ greatly, it is difficult to have a configuration that satisfies all kinds of traffic.

We continue to use the model in Fig. 3 to analyze the performance of IRN. A sender is connected to a receiver behind the congestion node through a local switch. Because IRN does not provide a formal analysis and its mechanism is complex, we use simulation to evaluate its performance under long-distance links. The length of the long-distance link in the model is the variable. The longer one is used to simulate a long-range RDMA link with a large BDP. The shorter one is used to simulate a local datacenter link with a small BDP. For each case, an optimal IRN configuration is given as recommended in [4]. We then swap the IRN configurations of the two networks to show how much throughput is lost when the configuration is not optimized for its traffic.

For the long-range RDMA simulation, the bandwidth is set to 100 Gbps, the propagation delay is set to 400 μs, and the ingress buffer is set to 20 MB. The size of the IRN bitmap is set to 6,686 packets. RTO_{high} and RTO_{low} are set to 2,480 μs and 802 μs, respectively. For the local datacenter simulation, we adjusted the propagation delay to 12 μs and the ingress buffer to 3 MB (the longest 6-hop path in a typical CLOS topology). The size of the IRN bitmap is set to 219 packets. RTO_{high} and RTO_{low} are set to 110 μs and 26 μs, respectively. According to the recommendation of IRN [4], we disable PFC on all switches and hosts. The test scenarios are respectively no congestion and congestion throughput from 20 Gbps to 80 Gbps. One large flow is sent from the sender to the receiver without any congestion controls to eliminate their interferences.

The throughput of transferring useful data, i.e. goodput, is shown in Fig. 6. As expected, for the intra-datacenter traffic, a larger bitmap optimized for a long-distance link causes too much transmission, resulting in congestion and retransmission, as shown in Fig. 6a. The smaller bitmap optimized for the intra-datacenter link leads to under-throughput of inter-datacenter traffic because in-flight packets are smaller than one BDP, as shown in Fig. 6b. Therefore, IRN with a fixed bitmap size cannot guarantee the throughput of long-range RDMA.

Furthermore, even if there was an IRN mechanism with the bitmap size adjustable according to the congestion level,



(a) Goodputs of two types of traffic with configuration optimized for the long-range RDMA.

(b) Goodputs of two types of traffic with configuration optimized for the local datacenter.

Fig. 6. Goodput with different configurations and different traffic with IRN.

the resource occupied by the bitmap is too large for the long-range RDMA scenario. Calculated based on the MTU of 1 KB, the fiber bandwidth of 100 Gbps, and the propagation delay of 5 μs per kilometer, a 125-byte bitmap is required per kilometer per queue pair [4]. On an RDMA network interface card that supports two thousand queue pairs and a maximum link length of 120 km, the bitmap occupies at least 30 MB in the chip, which will increase the cost of the network interface card greatly.

4.3 Summary and Goals

The analysis above illustrates that the existing lossless and lossy networks are not suitable for the long-range RDMA. In the lossless network with the PFC mechanism, a part of the BDP-sized buffer is wasted to compensate for the delay of the RESUME message. For the lossy network solution, the fixed-size bitmap of the IRN cannot guarantee the performance of traffic with different RTTs in a long-range RDMA scenario. And the deployment of IRN needs to upgrade all NICs in the datacenter. Considering the pros and cons of these solutions, we propose to support long-distance lossless RDMA with minimal buffer without changing the PFC mechanism of the external switches at both ends.

5 Design

In this section, we present the design of Swing. Swing extends PFC with a modified "Relay" mechanism to support long-range lossless RDMA. It introduces a relay device to enable minimal modification to the existing infrastructures, reduce required buffer size, and provide extendability at the same time.

5.1 Long-Distance Link Deployment

In a typical long-distance link deployment scenario, a datacenter is connected to another datacenter through the datacenter interconnection (DCI) fiber with an external switch and DWDM, as shown in Fig. 7 ①. A straightforward way to deploy the modified flow control mechanism is to replace the external switch. However, the buffer of the external switch is limited and cannot be expanded, which will limit the number of fiber connections with other datacenters. For example, for a 100 Gbps port that supports a propagation delay of 400 μs, each port requires a 10 MB buffer. A 32 MB

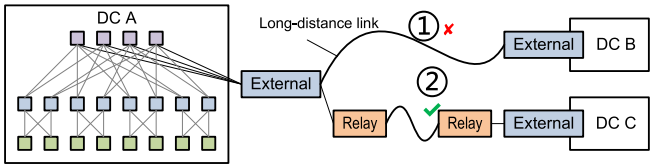


Fig. 7. Long-distance link deployment methods.

buffer switch can support up to three long-distance links like this. If more DCI links need to be deployed, the switch has to be replaced. Additionally, DCIs and data center networks (DCNs) are generally managed and operated separately. The operator should only add additional devices to support the new DCI to ensure minimal changes to existing devices.

The basic idea of SWING is to plug a relay device into both ends of the long-distance link, as shown in Fig. 7 ②. We call this device a "relay" because it can take over the PFC mechanism on the long-distance link, which allows us to modify the PFC mechanism without replacing the external switch. Moreover, it can buffer the in-flight packets to release the buffer usage of the external switch. In this way, the external switch does not need to have a large buffer, nor does it need to reserve an unused buffer to deal with future inter-datacenter links. As a complementary device, it is more flexible and less expensive than the switch upgrade solution like PFC.

5.2 PFC-Relay

The previous section has proved that PFC can cause throughput loss on long-distance links. The root cause of the throughput loss on the long-distance link is that PFC RESUME cannot take effect in time. The latency of RESUME leads to the need for additional buffers. Therefore, our goal is to minimize the latency of RESUME. Our intuition is to forward the downstream PFC signal directly to the upstream without triggering through the XON and XOFF thresholds on its queue. In this way, the PFC signal is no longer delayed, thus avoiding an additional buffer of at least one BDP to ensure throughput.

Fig. 8 illustrates a pair of ports on the PFC-relay device. One is called Local (L) Port, which is connected to the local external switch. The other is called Remote (R) Port, which is connected to the remote external switch through the long-distance link. The PFC mechanism of the external switches does not need any modification. The PFC message responding behavior of the relay device is consistent with the standard PFC mechanism. More specifically, any port stops sending when it receives `PAUSE` and resume sending when PFC RESUME arrives. The only modification to the PFC mechanism in the relay device is that, the *local port* which is connected to the local switch requires to relay the received PFC messages to the remote side.

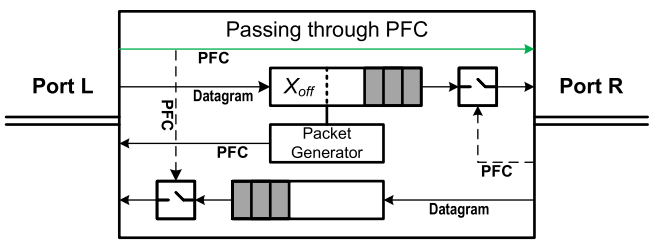


Fig. 8. The mechanism of PFC-Relay.

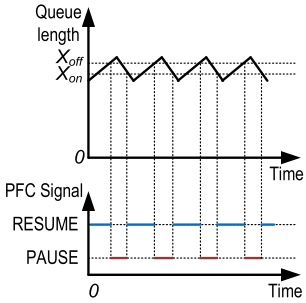


Fig. 9. Frequent cyclic PFC signals generated by the external switch.

Let's take the case in Fig. 7 as an example. For DC A, as the relay device is closely connected to the local external switch, the PFC messages generated by the switch will get feedback from the local relay instantly. Consequently, as illustrated in Fig. 9, the switch will generate frequent cyclic PFC messages which maintain the switch queue length fluctuating near the `XOFF/XON` threshold, just like a 'swing'. In this way, these frequent cyclic PFC messages enforce the *local port* of the relay device to send data at the draining rate of the switch. Furthermore, as the local relay device forwards the PFC messages to the remote side, the remote relay device will receive the same PFC controlling signal sequences after a propagation delay, as illustrated in Fig. 10. As the line rate of all relay ports is the same, the remote relay device will be forced by the delayed cyclic PFC messages to send at the same rate as the local switch's draining rate.

Analysis. First, we use the relay on the left in Fig. 10 as the local relay to analyze the upper bound of the ingress buffer that is connected to the long-distance link. The relay on the right in Fig. 10 is the remote relay, which sends data to the local relay. Suppose the RTT of the long-distance link equals $2D$. Denote the average sending rate from the local relay device to the local external switch at time t as $\overline{R}_{out}(t)$. The average sending rate of the remote relay to the local relay at time t will equal $\overline{R}_{out}(t-D)$ because the PFC messages will take a delay of D to arrive at the remote relay. And denote the average input rate of the local relay device at time t as $\overline{R}_{in}(t)$. As a packet from the remote relay will take another D to arrive at the local relay, we have

$$\overline{R}_{in}(t) = \overline{R}_{out}(t-2D), \tag{5}$$

which describes a delayed rate control effect.

Fig. 11 illustrates this delayed control effect. It's worth noting that the shape of the curve is arbitrary and is just intended to show the delayed rate control effect and the required buffer size. This figure can also tell how much the relay device's buffer will be used under varying throughput. In the figure, the green curve indicates the draining rate $\overline{R}_{out}$ while the blue curve indicates the input rate $\overline{R}_{in}$, which equals the draining rate exactly one RTT ($2D$) ago. The required buffer size equals the maximum value of

$$\int_{t_0}^{t_0+2D} (R_{in}(t) - R_{out}(t)) \cdot dt,$$

for any t_0. In the figure, the blue area marked with '+' minus the green area marked with '-' corresponds to the value of this equation. Similar to the calculation of the area of a parallelogram, the maximum queue length is exactly one BDP

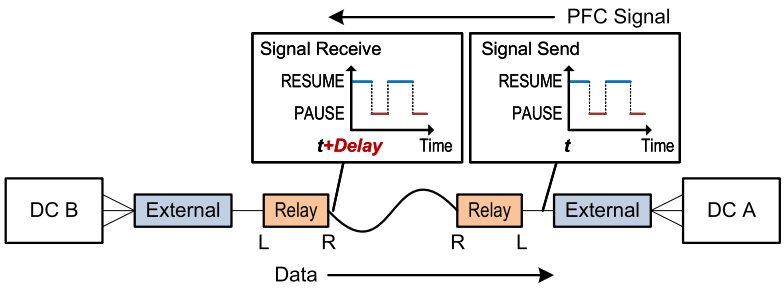


Fig. 10. Workflow of the PFC-Relay mechanism.

$(2D \cdot BW)$. Therefore, the minimum buffer required for the PFC-Relay queue is one BDP. In contrast, the deep-buffer switch solution requires a minimum buffer of two BDPs, which is twice as much as the proposed solution.

Second, we analyze the throughput of PFC-Relay using the same model and definition as the analysis of the PFC mechanism in Section 4. Let R_{out} be the relay device's draining rate, which is also the sending rate from the relay to the local external switch. Note that $R_{out} = \alpha \cdot BW$. We can easily get its utilization ratio of the long-distance link

$$\phi(D, \alpha) = \alpha, \tag{6}$$

which shows that the throughput of PFC-Relay is only related to the draining rate, and is independent of the link propagation delay and the buffer size.

For a more intuitive understanding, we make the same assumption with the analysis of PFC, that is, a 100 Gbps port with an 11 MB ingress buffer. The long-distance link throughput varies with the link propagation delay and the draining rate is shown in Fig. 12. The throughput is zero when the packet loss happens. These two figures both illustrate that, in the case of sufficient buffer, the throughput of PFC-Relay is only related to the draining rate and is independent of the link delay.

Parameter Settings. According to the previous analysis, there is no difference between the PFC configuration of Port L and the switch configuration inside the datacenter. The PFC-Relay configuration of Port R has only one parameter, which is the size of the ingress buffer. Taking into account the buffer fluctuation caused by PFC and the delay fluctuation caused by hardware, the buffer is required to be a little larger than one BDP of the long-distance link.

5.3 Comparison

In this section, we compare SWING with the state of the art briefly in several aspects.

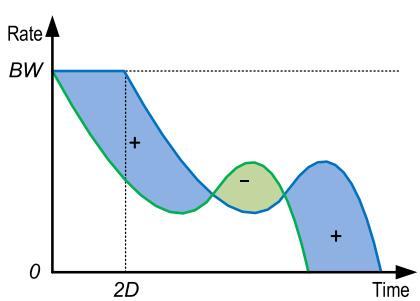


Fig. 11. The theoretical buffer lower bound of PFC-Relay.

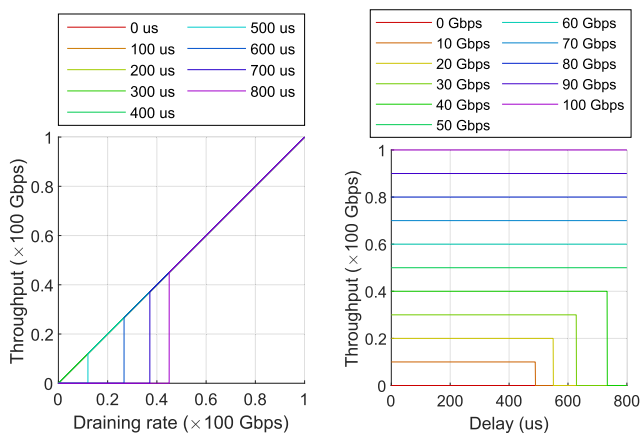
Network Card Requirements. Both the deep-buffered PFC and SWING solutions do not need to replace the network card. For IRN, the bitmap of the network card is fixed and only supports the intra-datacenter RDMA. So modifying IRN requires replacing the network cards in the entire datacenter, which is unacceptable.

External Switch Requirements. The external switch used by the PFC solution requires a deep buffer with a Tbps-level throughput chip to support long-range RDMA. SWING only needs half of the buffer used by the PFC switch to support the same link length. Moreover, it does not need to implement complex functions such as switching and traffic management and only needs to support hundred Gbps-level throughput for a pair of ports, which makes the chip of relay much more economical.

Scalability. The maximum link length and link number that the PFC solution can support is limited by the buffer size of the external switch. In contrast, for SWING, the maximum link number is as same as the relay device number, which can be easily expanded. Regarding the link length, SWING can support twice the link length of the PFC solution with the same buffer resources.

6 EVALUATION

In this section, we evaluate the performance of SWING and compare it with the native PFC mechanism and IRN. The congestion control algorithm used in the experiments is DCQCN [5]. These three flow control mechanisms, including



(a) Throughput under different draining rates.

(b) Throughput under different delays.

Fig. 12. With a fixed total buffer of 11 MB, the best performance of PFC-Relay under different congestion and link propagation delays.

SWING , native PFC, and IRN, are all implemented in the NS3 simulator [40].

6.1 Evaluation Settings

Network Topologies. We use the same topology as in Fig. 1 when evaluating native PFC and IRN. Both DC A and DC B are built with the FatTree [41] topology. Each Top-of-Rack (ToR) switch is connected to two servers. The bandwidth of the long-distance link is 400 Gbps, and the bandwidth of other links is 100 Gbps. When evaluating SWING , we use the topology shown in Fig. 7. Compared with the previous topology, the only difference is that there are two PFC-relay devices plugged near the external switches. Except for long-distance links, the propagation delay of all links is 1 μs. The propagation delay of the long-distance link is 400 μs, corresponding to a distance of 80 km.

Native PFC Settings. For ports that are connected to the long-distance link, we set the ingress buffer to 41 MB. The PFC threshold at these ports is set to 1 MB. According to the analysis in Section 4.1, this will result in a decrease in the throughput of the long-distance link. The buffer of other ports is twice the BDP of one datacenter, where the headroom is equal to the BDP of the link it is connected to. In our topology, the headroom of these ports is 30 KB, and the PFC threshold is 288 KB. The switches inside the datacenter have a 10 MB shared buffer with a dynamic PFC threshold so that the PFC mechanism will be triggered when an ingress queue consumes more than 25% of the available shared buffer.

SWING*Settings.* Except for external switches and relay devices, other switch settings are consistent with the native PFC settings. For the external switch, the headroom of the port connected to the local datacenter is set to 30 KB, and the PFC threshold is set to 288 KB. For the port on the link connecting the relay and external switch, as it supports higher bandwidth without being connected to a long-distance link, the headroom of the port is set to 120 KB, while the PFC threshold is set to 198 KB. The sum of the two is consistent with the other ports. The port of the relay that is connected to the long-distance link does not need to set the PFC threshold, and the port buffer is 41 MB.

IRN Settings. The buffer size of all ports is the same as the configuration in the lossless network, but PFC is not enabled. As the IRN configuration is related to BDP and there are two different BDPs in the long-range RDMA scenario, we set two IRN configurations in the evaluation. One is the optimal configuration based on inter-datacenter flows (IRN-inter). The bitmap size of IRN is set to 6,809, while RTO_{high} and RTO_{low} are 1875 μs and 817 μs, respectively. Another one is the optimal configuration based on intra-datacenter flows (IRN-intra). The bitmap size, RTO_{high} and RTO_{low} are 106, 140 μs and 13 μs, respectively. The above configurations of IRN are based on the recommended settings in [4].

Congestion Control. DCQCN [5] is a congestion control algorithm of layer 3 which is widely deployed in datacenters. PFC is a layer 2 flow control mechanism and is usually enabled together with DCQCN in datacenters. Therefore, the two are orthogonal. To compare the performance of PFC, SWING, and IRN and exclude the influence of other factors like congestion control algorithms, we disable DCQCN. In this way, we can exclude the influence of the congestion control algorithm on the results of these flow control algorithms. In the final comprehensive evaluation, we enable DCQCN to simulate a real datacenter environment. The DCQCN configuration in the evaluation is based on Mellanox's recommendations [42]. The configuration of ECN is based on HPCC [43]. In addition, ECN is disabled on switches connected to long-distance links, which has little effect on our results.

Traffic Loads. The evaluations adopt commonly used datacenter traffic traces, i.e., WebSearch [44], and FB_Hadoop [45]. The WebSearch workload is characterized by small requests and large responses. Among them are mainly long flows, 95% of the flows exceed 1 MB [46]. 70% of the flows in the FB_Hadoop workload are smaller than 10 KB, but 90% of the traffic is contributed by flows larger than 100 KB. In our evaluation, the average load of the long-distance link is varied from 30% to 70%, and the number of receivers is adjusted from 16 to 4 respectively. In the mixed traffic test, we set a 30% load of background traffic inside both datacenters.

Metrics. We have four performance metrics. (i) Average flow completion time (FCT); (ii) Tail FCT; (iii) Long-distance link throughput; (iv) PFC pause duration and IRN retransmission time.

6.2 The Performance of Inter-DC Flows

We first evaluate the performance of the four settings, PFC, PFC-Relay, IRN-intra, and IRN-inter, when there is only long-distance link traffic. We set the inter-datacenter traffic to be unidirectionally sent from the 16 servers of DC A to the 4 servers of DC B. The average load of long-distance links is 70%.

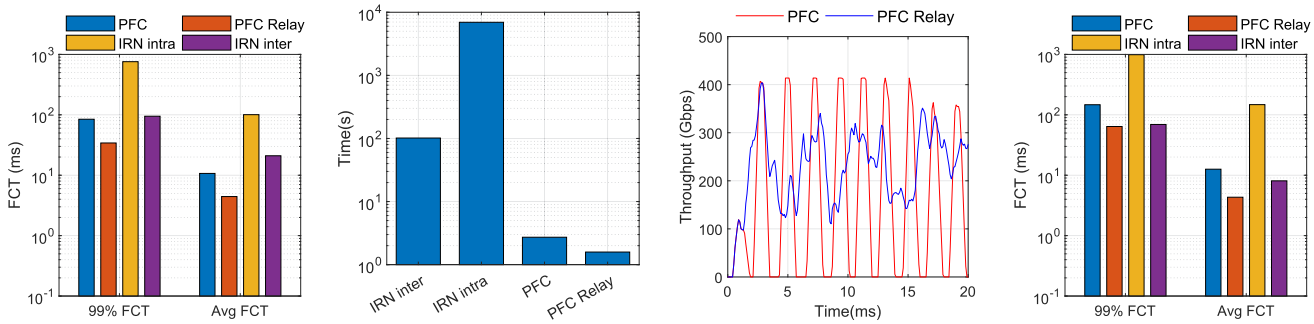
Better Performance Without CC. We first exclude the influence of congestion control on PFC-Relay, and the result is shown in Fig. 13a. Compared with PFC, PFC-Relay can reduce the 99% percentile latency by 60%, and the average FCT by 59%. Due to the long feedback delay, both settings of IRN do not perform well.

Lower Wasted Transmission Time. As shown in Fig. 13b, PFC-Relay also helps to reduce the PFC pause time, saving 42% of the transmission time compared to PFC. IRN retransmission time is calculated by dividing the number of retransmission bytes by the bandwidth of the server NIC, which represents the retransmission time wasted by the server. IRN-intra causes retransmissions due to timeout, therefore reducing performance. IRN-inter also causes retransmissions due to the long feedback delay and the large bitmap.

Higher Long-Distance Link Utilization. We also test the rate of the long-distance link, as shown in Fig. 13c. When the average load of long-distance links is 70%, PFC-Relay can quickly feedback the RESUME message of the downstream switch, so the link is always active. For PFC, there is periodic non-throughput on long-distance links, exacerbating head-of-line blocking and flow latency.

Maintaining Performance Under DCQCN. Due to the delay of explicit congestion packets (CNP), the performance improved by DCQCN on long-distance links is limited, as shown in Fig. 13d. Compared with PFC, the tail latency and average FCT of PFC-Relay are still reduced by 56% and 66%, respectively.

Other Scenarios. To evaluate the performance of PFC-Relay in various scenarios, we adjust the average load of the long-distance link and the number of receivers respectively.



(a) 70% avg. load, 16 to 4, No CC

(b) PFC pause time and IRN retransmission time

(c) The rate of long-distance link

(d) 70% avg. load, 16 to 4, DCQCN

Fig. 13. The performance of each settings when there is only inter-datacenter traffic.

Fig. 14a shows that when the link load is 30%, the performance improved by PFC-Relay is not significant, however, the tail latency and average FCT are still reduced by 47% and 32% respectively compared to PFC. IRN-inter has much less packet loss in this scenario, and its performance is very close to PFC-Relay. IRN-intra is still dominated by timeout retransmission, and its performance is not ideal.

When we adjust the number of receivers to 16, we find that even if the congestion of each server drops, the ports of the external switch can still trigger PFC. In this case, PFC-Relay can still ensure the performance of inter-datacenter flows as much as possible, as shown in Fig. 14b.

6.3 The Impact on Intra-DC Traffic

We evaluate whether the performance improvement of inter-datacenter flows will affect the intra-datacenter traffic. We keep the flows from the 16 servers of DC A to the 4 servers of DC B without congestion control, and the average load of the long-distance link is 70%. Then we add 30% background traffic load to DC B and DC A respectively.

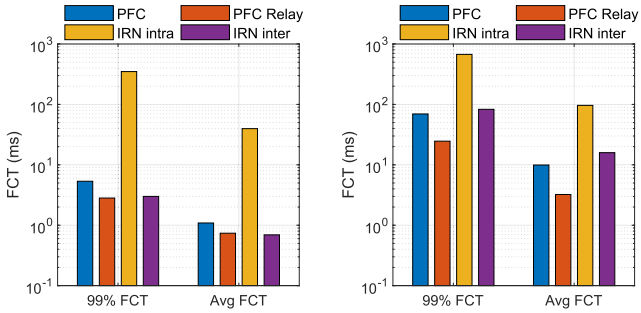
Small Impact on the Remote DC Traffic. First, we evaluate the scenario with DC B background traffic. This means that inter-datacenter flows from DC A will compete with the flows in DC B. The result is shown in Fig. 15. Due to the congestion of DC B, PFC-Relay does not significantly improve the performance of inter-datacenter flows. Compared with PFC, the average FCT and tail latency of inter-datacenter flows are reduced by 16% and 17%, respectively. Because PFC-Relay reduces the throughput of the remote datacenter flows to repay the throughput of the inter-datacenter flows. The performance improvement of inter-datacenter flows results in a decrease in the performance of DC B traffic. But the overall performance of PFC-Relay is still better. Compared with PFC, the average FCT and tail latency of all flows are reduced by 14% and 18%, respectively.

Improving the Performance of the Local DC Traffic. Then we evaluate the scenario with DC A background traffic. In this scenario, the server sends both inter-datacenter and intra-datacenter flows, thereby influencing each other. As shown in Fig. 16, for PFC-Relay, because the increase in long-distance link throughput reduces the congestion in DC A, the server can send more data, which improves the performance of intra-datacenter flows as well. Compared to PFC, PFC-Relay reduces the average FCT and tail latency of inter-datacenter flows by 52% and 59%, of intra-datacenter flows by 63% and 55%.

6.4 Combine All Traffic Together

Last, we evaluate the overall performance of PFC-Relay. The inter-datacenter flows are still sent from the 16 servers of DC A to the 4 servers of DC B. The average load of long-distance links is 70%. DC A and DC B both have a 30% background traffic load. The WebSearch and FB_Hadoop traffic traces are used to generate flows. DCQCN is enabled on all server NICs. The results are shown in Fig. 17.

Because PFC-Relay is a flow control mechanism, it is not sensitive to the composition of the flow. The tail latency and average FCT of PFC-Relay are both better than the other three settings, with a minimum reduction of 44% and 53%. Compared with PFC, PFC-Relay combines the two advantages of high inter-datacenter flow throughput and lower congestion of DC A. For IRN-intra, due to the too short timeout, retransmission takes up a lot of transmission time. For IRN-inter, a large bitmap causes over transmission, and also cannot limit excessive retransmissions.



(a) 30% avg. load, 16 to 4, No CC

(b) 70% avg. load, 16 to 16, No CC

Fig. 14. The performance of inter-datacenter flows in different scenarios.

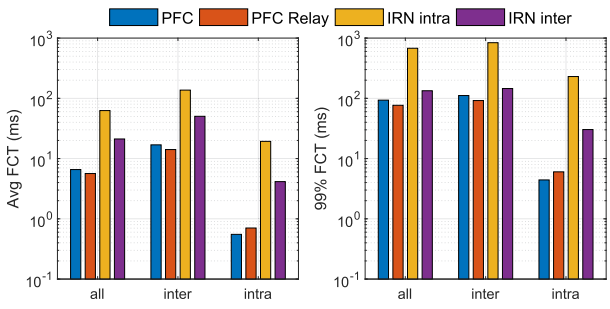


Fig. 15. The impact of inter-datacenter flows on the remote datacenter.

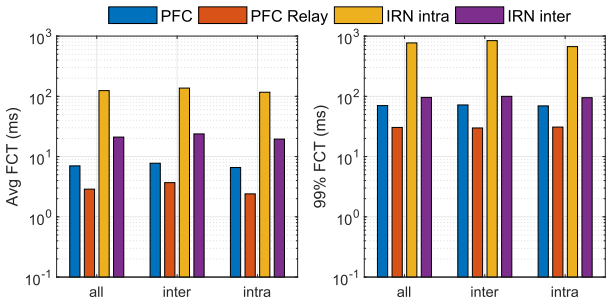
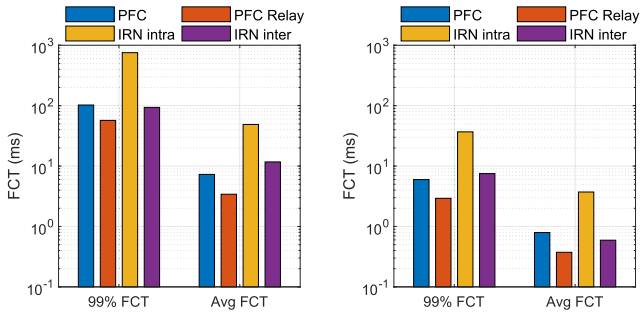


Fig. 16. The impact of inter-datacenter flows on the local datacenter.



(a) WebSearch (b) FB_Hadoop

Fig. 17. Overall performance.

7 Discussion

The Impact of Existing Work on Optimizing PFC on Swing. There are some works to address the problems of the PFC storm and deadlock. The method used to resolve the PFC storm in [13] is to detect the network status and disable the PFC mechanism on the NIC and top-of-rack (ToR) switch. In addition, several works [47], [48], [49] on datacenter congestion control are devoted to avoiding triggering PFC. These works are compatible with Swing because they do not involve modification of switches and have no special requirements for topology. The work of solving deadlock [22] needs to modify the flow control mechanism. Fortunately, Swing only works on long-distance links, and there is no CBD, so Swing is compatible with it.

Cost and Implementation of Swing. To add a long-distance link that supports lossless RDMA, the datacenter operator can choose to add a deep-buffer switch of the traditional PFC solution or a relay device of Swing. The advantage of the relay device is that it only needs half the buffer of the switch. Moreover, it does not need to support packet switching, and the maximum throughput it supports is less than one-tenth of that of the switch. Therefore, the cost of ASIC in the relay device is less than that of the switch. In addition, the deployment cost of relay devices and switches is not much different. Both of these solutions need to be connected and tuned with DCI specifically.

Support PFC-Relay on DCI Switches. The PFC-relay mechanism can be added to the DCI switch, which requires the switch to generate high-frequency PFC packets based on throughput and congestion to replace the "swing" behavior. However, as described in Section 5.1, we hope to decouple the buffer from the DCI switch and let the relay undertake the flow control of the long-distance link and manage it independently from the DCN.

Need Better Lossy RDMA for Flows With Different RTTs. In our evaluation, the fixed bitmap of the IRN does not guarantee the performance of flows with different RTTs. For long-distance links, a too large bitmap size will cause over-transmission and retransmission, and a too short retransmission timeout will also cause a large number of retransmissions, both of which have a great impact on performance. But we still believe that lossy RDMA can and is suitable for deployment on long-distance links. In the future, we will consider adapting a similar mechanism to long-distance links and optimize for traffic with different RTTs and congestion.

Need Better CC for Inter-DC flows. In the evaluation, we found that DCQCN does not significantly improve the performance of inter-datacenter flows, and in some cases, the performance even drops. When the external switch has a deep buffer, how to improve the performance of inter-datacenter flows is a difficult problem. Some works like [50], [51], [52] are aiming to improve the congestion control mechanism for the hybrid network of DCN and WAN. The main consideration of these works is to use different congestion control algorithms for traffic on different networks. Our future work will also consider the congestion control on Swing, and explore the use of relay devices to proxy congestion control for these two types of traffic.

8 Conclusion

This paper proposes Swing to enable long-range lossless RDMA via PFC-relay. Swing is fully compatible with existing network protocols and requires no modifications to existing infrastructures inside datacenters. It plugs a "relay" device close to the external switch, which will generate periodic frequent cyclic PFC messages to enforce the relay device to send data at the switch's draining rate. Swing guarantees no throughput loss for long-distance links with half the buffer size required by the native PFC mechanism. We evaluate Swing against native PFC and IRN with inter-datacenter traffic and intra-datacenter traffic in various scenarios. The results demonstrate that Swing achieves better inter-datacenter flow performance and overall performance than native PFC and IRN.

Acknowledgments

The authors would like to thank anonymous reviewers for their valuable comments.

References

[1] P. MacArthur and R. D. Russell, "A performance study to guide RDMA programming decisions," in *Proc. IEEE 14th Int. Conf. High Perform. Comput. Commun. IEEE 9th Int. Conf. Embedded Softw. Syst.*, 2012, pp. 778–785.

[2] T. S. Woodall, G. M. Shipman, G. Bosilca, R. L. Graham, and A. B. Maccabe, "High performance RDMA protocols in HPC," in *Proc. Eur. Parallel Virt. Mach./Message Passing Interface Users' Group Meeting*, 2006, pp. 76–85.

[3] I. T. Association, "Infiniband architecture specification volume 1 release 1.2.1 annex A17: RoCEv2," 2014. [Online]. Available: https://cw.infinibandta.org/document/dl/7781

[4] R. Mittal et al., "Revisiting network support for RDMA," in *Proc. Conf. ACM Special Int. Group Data Commun.*, 2018, pp. 313–326. [Online]. Available: https://doi.org/10.1145/3230543.3230557

[5] Y. Zhu et al., "Congestion control for large-scale RDMA deployments," in *Proc. ACM Conf. Special Int. Group Data Commun.*, 2015, pp. 523–536. [Online]. Available: https://doi.org/10.1145/2785956.2787484

[6] S.-Y. Tsai and Y. Zhang, "Lite kernel RDMA support for datacenter applications," in *Proc. 26th Symp. Operating Syst. Princ.*, 2017, pp. 306–324. [Online]. Available: https://doi.org/10.1145/3132747.3132762
[7] S. Alam, P. Agnihotri, and G. Dumont, "AWS re:invent. enterprise fundamentals: Design your account and VPC architecture for enterprise operating models," 2016. [Online]. Available: https://www.slideshare.net/AmazonWebServices
[8] M. Filer, J. Gaudette, Y. Yin, D. Billor, Z. Bakhtiari, and J. L. Cox, "Low-margin optical networking at cloud scale [invited]," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 11, no. 10, pp. C94–C108, Oct. 2019.
[9] X. Zhou and H. Liu, "Pluggable DWDM: Considerations for campus and metro DCI applications," Stresemannallee 15 60596, Frankfurt/Main Germany, 2016.
[10] Y. Sverdlik, "Facebook rethinks in-region data center interconnection," 2018. [Online]. Available: https://www.datacenterknowledge.com/networks/facebook-rethinks-region-data-center-interconnection
[11] I. T. Association, "Infiniband architecture specification Volume 2 Release 1.3.1. 2016," 2014. [Online]. Available: https://cw.infinibandta.org/document/dl/8125
[12] W. Cheng, K. Qian, W. Jiang, T. Zhang, and F. Ren, "Re-architecting congestion management in lossless ethernet," in *Proc. 17th USENIX Symp. Netw. Syst. Des. Implementation*, 2020, pp. 19–36. [Online]. Available: https://www.usenix.org/conference/nsdi20/presentation/cheng
[13] C. Guo et al., "RDMA over commodity ethernet at scale," in *Proc. ACM SIGCOMM Conf.*, 2016, pp. 202–215. [Online]. Available: https://doi.org/10.1145/2934872.2934908
[14] R. Recio, B. Metzler, P. Culley, J. Hilland, and D. Garcia, "A remote direct memory access protocol specification," Tech. Rep. RFC 5040, Oct., 2007.
[15] Mellanox, "RoCE vs. iWARP competitive analysis," 2021. [Online]. Available: https://www.mellanox.com/related-docs/whitepapers/WP_RoCE_vs_iWARP.pdf
[16] IEEE, "Data center bridging task group," 2013. [Online]. Available: https://www.ieee802.org/1/pages/dcbridges.html
[17] Mellanox, "InfiniBand white paper," 2020. [Online]. Available: https://www.mellanox.com/resources-library/white-papers
[18] Mellanox, "InfiniBand architecture released specification," 2020. [Online]. Available: https://www.infinibandta.org/ibta-specifications-download/
[19] NVIDIA, "RDMA over converged ethernet (RoCE)," 2022. [Online]. Available: https://docs.nvidia.com/networking/pages/viewpage.action?pageId=25134510
[20] IEEE, "802.1Qbb–Priority-based flow control," 2010. [Online]. Available: https://1.ieee802.org/dcb/802–1qbb/
[21] S. Hu et al., "Deadlocks in datacenter networks: Why do they form, and how to avoid them," in *Proc. 15th ACM Workshop Hot Topics Netw.*, 2016, pp. 92–98. [Online]. Available: https://doi.org/10.1145/3005745.3005760
[22] K. Qian, W. Cheng, T. Zhang, and F. Ren, "Gentle flow control: Avoiding deadlock in lossless networks," in *Proc. ACM Special Int. Group Data Commun.*, 2019, pp. 75–89. [Online]. Available: https://doi.org/10.1145/3341302.3342065
[23] Mellanox, "ConnectX6 firmware," 2020. [Online]. Available: https://docs.mellanox.com/display/ConnectX6DxFirmwarev22271016/Changes_and_New_Features
[24] K. Veeraraghavan et al., "Maelstrom: Mitigating datacenter-level disasters by draining interdependent traffic safely and efficiently," in *Proc. 13th USENIX Symp. Operating Syst. Des. Implementation*, 2018, pp. 373–389.
[25] A. Singh et al., "Jupiter rising: A decade of CLOS topologies and centralized control in Google's datacenter network," in *Proc. ACM Conf. Special Int. Group Data Commun.*, 2015, pp. 183–197. [Online]. Available: https://doi.org/10.1145/2785956.2787508
[26] V. Dukic et al., "Beyond the mega-data center: Networking multi-data center regions," in *Proc. Annu. Conf. ACM Special Int. Group Data Commun. Appl. Technol. Architectures Protoc. Comput. Commun.*, 2020, pp. 765–781. [Online]. Available: https://doi.org/10.1145/3387514.3406220
[27] E. Kissel and M. Swany, "Evaluating high performance data transfer with RDMA-based protocols in wide-area networks," in *Proc. IEEE 14th Int. Conf. High Perform. Comput. Commun. IEEE 9th Int. Conf. Embedded Softw. Syst.*, 2012, pp. 802–811.
[28] V. Vasudevan et al., "Safe and effective fine-grained TCP retransmissions for datacenter communication," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 303–314, 2009.
[29] E. Kissel, M. Swany, B. Tierney, and E. Pouyoul, "Efficient wide area data transfer protocols for 100 gbps networks and beyond," in *Proc. 3rd Int. Workshop Netw.-Aware Data Manage.*, 2013, pp. 1–10. [Online]. Available: https://doi.org/10.1145/2534695.2534699
[30] ESnet, "The network," 2021. [Online]. Available: https://www.es.net/engineering-services/the-network/
[31] L. Rotman, "ESnet, Orange silicon valley, and bay microsystems demonstrate the world's first long distance 40Gbps RDMA data transfer," 2011. [Online]. Available: https://www.es.net/news-and-publications/esnet-news
[32] O. S. Inc., "Longbow devices eliminate the range limitations imposed by the InfiniBand specification," 2021. [Online]. Available: https://obsidianstrategics.com/products/longbow/index.html
[33] Vcinity, "Vcinity radical," 2021. [Online]. Available: https://vcinity.io/radical
[34] NVIDIA, "NVIDIA mellanox MetroX-2 long haul systems," 2021. [Online]. Available: https://www.nvidia.com/en-us/networking/infiniband/metrox-2/
[35] W. Allcock, J. Bresnahan, R. Kettimuthu, and M. Link, "The globus striped GridFTP framework and server," in *Proc. ACM/IEEE Conf. Supercomputing*, 2005, pp. 54–54.
[36] Netperf, "The netperf homepage," 2022. [Online]. Available: https://hewlettpackard.github.io/netperf/
[37] S. Narravula, H. Subramoni, P. Lai, R. Noronha, and D. K. Panda, "Performance of HPC middleware over infiniband WAN," in *Proc. IEEE 37th Int. Conf. Parallel Process.*, 2008, pp. 304–311.
[38] S. N. Avci, Z. Li, and F. Liu, "Congestion aware priority flow control in data center networks," in *Proc. IFIP Netw. Conf. Workshops*, 2016, pp. 126–134.
[39] C. Tian et al., "P-PFC: Reducing tail latency with predictive PFC in lossless data center networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 6, pp. 1447–1459, Jun. 2020.
[40] ns 3, "ns-3 homepage," 2020. [Online]. Available: https://www.nsnam.org/
[41] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proc. ACM SIGCOMM Conf. Data Commun.*, 2008, pp. 63–74. [Online]. Available: https://doi.org/10.1145/1402958.1402967
[42] Mellanox, "DCQCN parameters," 2020. [Online]. Available: https://community.mellanox.com/s/article/dcqcn-parameters
[43] Y. Li et al., "HPCC: High precision congestion control," in *Proc. ACM Special Int. Group Data Commun.*, 2019, pp. 44–58. [Online]. Available: https://doi.org/10.1145/3341302.3342085
[44] M. Alizadeh et al., "Data center TCP (DCTCP)," in *Proc. ACM SIGCOMM Conf.*, 2010, pp. 63–74. [Online]. Available: https://doi.org/10.1145/1851182.1851192
[45] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, "Inside the social network's (datacenter) network," in *Proc. ACM Conf. Special Int. Group Data Commun.*, 2015, pp. 123–137. [Online]. Available: https://doi.org/10.1145/2785956.2787472
[46] B. Montazeri, Y. Li, M. Alizadeh, and J. Ousterhout, "Homa: A receiver-driven low-latency transport protocol using network priorities," in *Proc. Conf. ACM Special Int. Group Data Commun.*, 2018, pp. 221–235.
[47] Y. Zhu et al., "Congestion control for large-scale RDMA deployments," *SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 523–536, Aug. 2015. [Online]. Available: https://doi.org/10.1145/2829988.2787484
[48] B. Montazeri, Y. Li, M. Alizadeh, and J. Ousterhout, "Homa: A receiver-driven low-latency transport protocol using network priorities," in *Proc. Conf. ACM Special Int. Group Data Commun.*, 2018, pp. 221–235. [Online]. Available: https://doi.org/10.1145/3230543.3230564
[49] M. Handley et al., "Re-architecting datacenter networks and stacks for low latency and high performance," in *Proc. Conf. ACM Special Int. Group Data Commun.*, 2017, pp. 29–42. [Online]. Available: https://doi.org/10.1145/3098822.3098825
[50] A. Saeed et al., "Annulus: A dual congestion control loop for datacenter and wan traffic aggregates," in *Proc. Annu. Conf. ACM Special Int. Group Data Commun. Appl. Technol. Architectures Protoco. Comput. Commun.*, 2020, pp. 735–749. [Online]. Available: https://doi.org/10.1145/3387514.3405899
[51] S. Zou, J. Huang, J. Liu, T. Zhang, N. Jiang, and J. Wang, "GTCP: Hybrid congestion control for cross-datacenter networks," in *Proc. IEEE 41st Int. Conf. Distrib. Comput. Syst.*, 2021, pp. 932–942.
[52] G. Zeng et al., "Congestion control for cross-datacenter networks," *IEEE/ACM Trans. Netw.*, vol. 30, no. 5, pp. 2074–2089, Oct. 2022.

Yanqing Chen received the BS degree from the Department of Computer Science and Engineering, Southeast University, China, in 2019. He is currently working toward the PhD degree with the Department of Computer Science and Technology, Nanjing University, China. His research interests include programmable switches and datacenter networks.

Chen Tian received the BS, MS, and PhD degrees from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, China, in 2000, 2003, and 2008, respectively. He is a professor with the State Key Laboratory for Novel Software Technology, Nanjing University, China. He was previously an associate professor with the School of Electronics Information and Communications, Huazhong University of Science and Technology, China. From 2012 to 2013, he was a postdoctoral researcher with the Department of Computer Science, Yale University. His research interests include data center networks, network function virtualization, distributed systems, Internet streaming and urban computing.

Jiaqing Dong received the BS degree in computer science and technology from Peking University, in July 2013, and the PhD degree in computer science and technology from Tsinghua University, in July 2020. He is currently an assistant researcher with the State Key Laboratory of Media Convergence and Communication, Communication University of China. From December 2020 to July 2021, he worked as a visiting scholar with the Department of Computer Science and Technology, Nanjing University, China. His research interests include data center networks and distributed systems.

Song Feng received the bachelor's degree from East China Normal University, in 1994, and the master's degree in engineering from Central South University, in 2007. He is now a senior engineer with the Network Information Center of Xiangya Hospital, Central South University. His research interests include medical informatization, Big Data, data governance and analysis.

Xu Zhang received the BS degree in communication engineering from the Beijing University of Posts and Telecommunications, China, in 2012, and the PhD degree in computer science from the Department of Computer Science and Technology, Tsinghua University, China, in 2017. He is with the School of Electronic Science and Engineering, Nanjing University, China. His research interests include artificial intelligence, multimedia communication, cloud/edge computing, and network measurement. He was the co-recipient of 2019 IEEE Broadcast Technology Society Best Paper Award.

Chang Liu received the BS degree from the School of Computer Science and Engineering, Northeastern University, China, in 2021. She is currently working toward the master's degree with the Department of Computer Science and Technology, Nanjing University, China. Her research interests include programmable switches and datacenter networks.

Peiwen Yu received the BE degree from the School of the Environment, Nanjing University, China, in 2020. He is currently working toward the 1st-year ME degrees with the Department of Computer Science and Technology, Nanjing University. His research interests include datacenter networks and network architecture.

Nai Xia received the BS and PhD degrees from the Department of Computer Science, Nanjing University China, in 2001 and 2007, respectively. He is currently an assistant professor with the Department of Computer Science and Technology, Nanjing University. His research interests include computer networking, operating systems and software security.

Wanchun Dou received the PhD degree in mechanical and electronic engineering from the Nanjing University of Science and Technology, China, in 2001. He is currently a full professor at the State Key Laboratory for Novel Software Technology, Nanjing University. From April 2005 to June 2005 and from November 2008 to February 2009, he respectively visited the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, as a visiting scholar. Up to now, he has chaired three National Natural Science Foundation of China projects and published more than 60 research papers in international journals and international conferences. His research interests include workflow, cloud computing, and service computing.

Guihai Chen received the BS degree in computer software from Nanjing University, in 1984, the ME degree in computer applications from Southeast University in 1987, and the PhD degree in computer science from the University of Hong Kong, in 1997. He is a distinguished professor of Nanjing University. He had been invited as a visiting professor by Kyushu Institute of Technology, Japan, University of Queensland, Australia and Wayne State University. He has a wide range of research interests with focus on parallel computing, wireless networks, data centers, peer-to-peer computing, high-performance computer architecture and data engineering. He has published more than 350 peer-reviewed papers, and more than 200 of them are in well-archived international journals such as *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Transactions on Computers*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE/ACM Transactions on Networking* and *ACM Transactions on Sensor Networks*, and also in well-known conference proceedings such as HPCA, MOBIHOC, INFOCOM, ICNP, ICDCS, CoNext and AAAI. He has won 9 paper awards including ICNP 2015 best paper award and DASFAA 2017 best paper award.