Research on High Performance Transmission Technology of DC Based on Network Awareness

Yanwei Wang
Architecture Research Department, Guangdong Inspur Intelligent Computing Technology Co., Ltd, and Inspur Electronic Information Industry Co., Ltd, Jinan Shandong China
wangyanwei@inspur.com

Cheng Huang
Architecture Research Department, Inspur Electronic Information Industry Co., Ltd, Jinan Shandong China
gaffer.c@foxmail.com

Jiaheng Fan
Architecture Research Department, Inspur Electronic Information Industry Co., Ltd, Jinan Shandong China
fanjiaheng@inspur.com

Le Yang
Architecture Research Department, Guangdong Inspur Intelligent Computing Technology Co., Ltd, Guangdong China
yangle01@inspur.com

Hongwei Kan
Architecture Research Department, Inspur Electronic Information Industry Co., Ltd, Jinan Shandong China
kanhongwei@inspur.com

Gaoming Cao
Digital Guangdong network construction Co., Ltd, Guangdong China
garincao@digitalgd.com.cn

ABSTRACT

In recent years, RDMA over Converged Ethernet (RoCE) provides high performance data transmission for Data center (DC). RoCE has excellent performance in lossless network, but when the network environment is unstable, the transmission performance of RoCE will decline rapidly. This paper proposes a DC network transmission technology based on network awareness. By monitoring the network status, the network status can be updated in real time. The transmission mechanism can be adjusted dynamically. In order to support the transmission in harsh environment, this paper constructs DC-TCP transmission based on Data Plane Development Kit (DPDK) TCP and DC-RoCE transmission based on RoCE, and designs a high-performance DC-TCP / DC-RoCE fusion technology framework. The simulation results show that the DC network transmission technology based on network awareness significantly improves the transmission capacity of the DC in complex environment.

CCS CONCEPTS

• **Networks**; • **Network architectures**; • **Network performance evaluation**; • **Network performance modeling**; • **Network services**; • **Network monitoring**;

KEYWORDS

DC, Network, RoCE, DPDK-TCP

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487153

ACM Reference Format:
Yanwei Wang, Cheng Huang, Jiaheng Fan, Le Yang, Hongwei Kan, and Gaoming Cao. 2021. Research on High Performance Transmission Technology of DC Based on Network Awareness. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487153

1 INTRODUCTION

With development of Internet, technologies such as big data, AI, and distributed computing have gradually matured. DC transmission has rapidly increased [1] [2] [3]. The main transmission technologies of the DC are DPDK-TCP and RoCE [4Edsall,] [5] [6]. TCP uses reactive congestion processing and cannot cope with burst traffic. DPDK-TCP is implement by high-performance software, with flexible mechanism and powerful functions, but high CPU usage [7] [8] [9]. RoCE is implemented through hardware, with simple functions, needs to be based on a lossless Ethernet network, and requires a good network environment. Once the lossless Ethernet is saturated, the performance will be greatly reduced, but the CPU occupancy rate is low [10] [11] [12]. No matter which method is adopted, it is unable to provide high-performance transmission in the DC under the complex network environment [13] [14] [15].

In response to the above problems, this paper proposes a DC network transmission technology based on network Awareness. This technology monitors network conditions, dynamically schedules DC communication resources, evolves the original communication protocol, integrates the advantages of multiple communication resources, and guarantees the communication quality of the DC in a complex network environment.

Table 1: DPDK-TCP VS RoCE Table

Features	DPDK-TCP	RoCE
Technical means	Software (flexible)	Hardware (fixed)
Communication processing delay	middle	Low
Congestion handling mode	Reaction	Lossless Ethernet
Burst traffic processing capability	Slower response Long delay	faster response Slow delay
CPU usage	High	Low
Flow control granularity	Stream	Port or priority
Performance when network overload	excellent	Poor

2 DC TRANSMISSION ARCHITECTURE BASED ON NETWORK AWARENESS

2.1 Technical Characteristics of DPDK-TCP and RoCE

The main work of this paper is to combine DPDK-TCP and RoCE, dynamically respond to complex network environment changes, and improve DC network environment adaptability. The main characteristics of the two include: technical means, communication processing delay, congestion handling mode, burst traffic processing capability, CPU usage, flow control granularity, and performance when network overload [16] [17] [18].

From the above table 1, both DPDK-TCP and RoCE have shortcomings. It is necessary to combine the advantages of them in order to release the cluster computing power of the DC.

2.2 Design of Data Center Transmission Architecture Based on Network Awareness

This paper designs data center transmission technology based on network state awareness. The technical framework model is as follows:

From the above figure 1, the technical architecture involves the network, data transmission layer, data processing layer, and application layer. The core of the research is the data transmission layer, and it is built on a hybrid network of lossless Ethernet and lossy Ethernet. The data transmission layer adopts data center network transmission technology based on network awareness, which mainly includes DC-TCP, DC-RoCE, network state awareness, network transmission scheduling, and network transmission migration.

As shown in figure 2, network status awareness is to aware changes in network status by monitoring network indicators. Network transmission scheduling schedules corresponding transmission methods to different streams. Through buffer migration, transmission parameters (such as: congestion window) inheritance and other functions to realize network transmission migration. In order to support the above functions, DC-TCP and DC-RoCE are evolving on DPDK-TCP and RoCE.

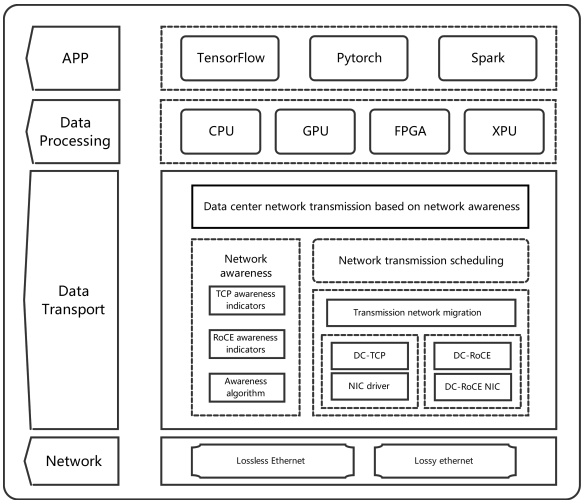

Figure 1: Figure of Technical Framework Model.

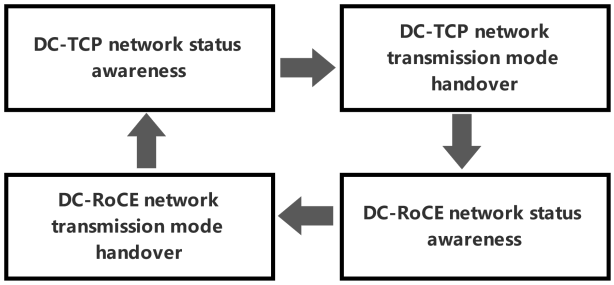

Figure 2: Figure of Mode Handover Model.

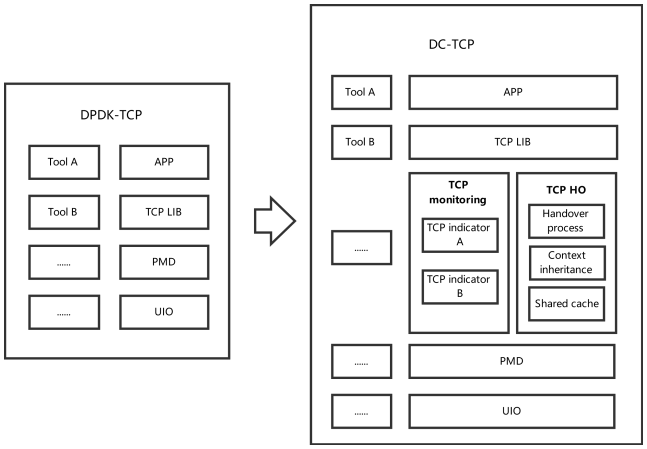

Figure 3: Figure of DPDK-TCP Evolution.

(1) DC DPDK-TCP evolution

As shown in figure 3, DC-TCP is a DPDK-TCP evolution technology. On the basis of DPDK-TCP, DC-TCP adds functions such as TCP network status monitoring, handover DC-RoCE, context inheritance, and shared cache.

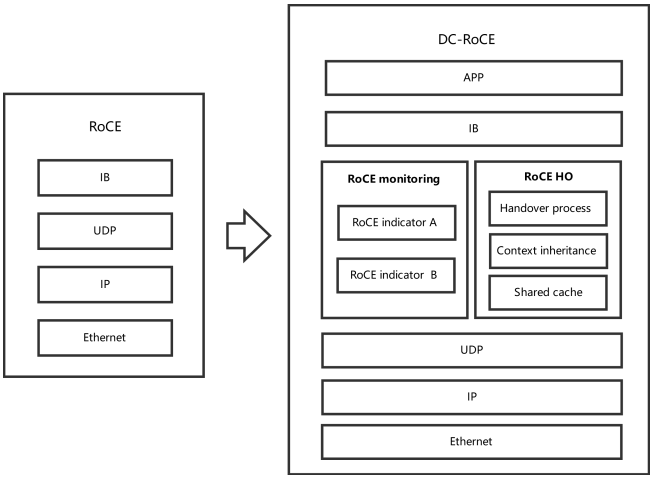

Figure 4: Figure of RoCE Evolution.

Table 2: Lossless VS Lossy Network Congestion Table

State	Lossless	Lossy
No congestion	No need for flow control	No need for flow control
congestion	Pause sending	Loss

(2) DC RoCE evolution

As shown in figure 4, DC-RoCE is a RoCE evolution technology. On the basis of RoCE, DC-RoCE adds functions such as TCP network status monitoring, RoCE Handover, context inheritance, and shared cache.

3 DC NETWORK TRANSMISSION STATUS AWARENESS

The network transmission status awareness determines the network transmission mode. The awareness result should give full advantages of different transmission modes. The high efficiency of RoCE in a non-congested network environment, and the flow control ability of TCP in a congested environment. Awareness indicators need to fully reflect the state of the network, and at the same time have both the efficiency and convenience of indicator collection.

3.1 Analysis of Complex Heterogeneous Ethernet Network Status

Ethernet is divided into lossy and lossless. Loss will cause packet loss when congested, and lossless will initiate back-pressure to suspend Ethernet port transmission. Both types of networks have congestion status, but the congestion handling is different. The comparison table is as follows in table 2

Both lossless and lossy Ethernet do not need flow control in a non-congested state. In this state, RoCE without slow start is more efficient. Congestion can be divided into two types: burst traffic congestion and long-term overload.

(1) Burst traffic congestion

Burst traffic congestion refers to short-term large traffic, but long-term overall traffic is not large. When a short burst of traffic is congested, the data center should use RoCE communication. RoCE has a high response speed and low CPU usage. Although performance will be significantly reduced during short-term congestion, as long as sudden congestion does not cause overall performance degradation, RoCE should continue to be used. The switch between RoCE and TCP has a certain cost, and frequent switch should be avoided.

(2) Long-term overload

When the overall traffic exceeds the network load for a long time, the DC should use TCP communication. At this time, due to the coarser granularity of RoCE flow control, there are problems such as deadlock, low efficiency of Back-to-N, and low communication efficiency. TCP is reactive congestion processing. When there is a long-term bottleneck in the intermediate node, it has high efficiency and should be switched to using TCP to transmit data.

(3) Judgment of burst traffic and long-term overload

Burst traffic and long-term overload are not a strictly distinguished state. The two states need to be judged based on the better performance of RoCE and TCP communication modes. Generally, after RoCE exceeds a certain rate, the performance will drop rapidly, but TCP can be dynamically adjusted to steadily use the network bandwidth to saturation. Therefore, the two states can be judged by comparing the network bandwidth when RoCE does not encounter back-pressure with the current network bandwidth.

3.2 TCP Network State Awareness

TCP meets the two necessary conditions of sufficient network and short-distance network to handover RoCE.

(1) Sufficient network

When the actual sending speed of TCP has not increased for a long time and no packet loss occurs, you can switch to RoCE. TCP continues to be unable to send to the network saturation state, use the following formula to judge:

$$sp < cwnd \tag{1}$$

(2) Short-distance network

The short distance of the network is a prerequisite for RoCE to have advantages. Too far network distance means that the network is unstable. RTT is an important indicator for identifying short-distance networks. Short RTT is an important prerequisite for short-distance networks. When the RTT is too large, it can be inferred that the network is at a remote end or the network status is very poor, and RoCE transmission should not be used. The short-distance network is judged by the following formula:

$$rtt < nt \tag{2}$$

Where, rtt is the return time of the sent data, and nt is the constant for short distance judgment. If the inequality is true, it is a short-distance network, otherwise it is a long-distance network.

3.3 RoCE Network State Awareness

When the network no longer adapts to RoCE, it should switch to TCP. The details are as follows:

- Packet loss occurs in RoCE operation

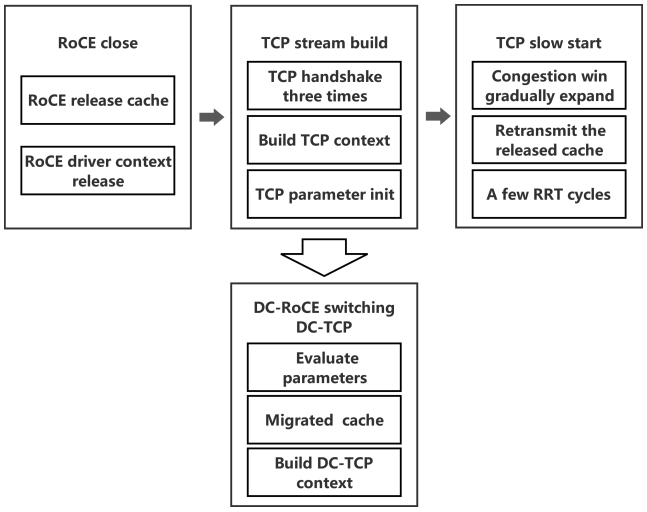

Figure 5: Figure of RoCE Switching TCP.

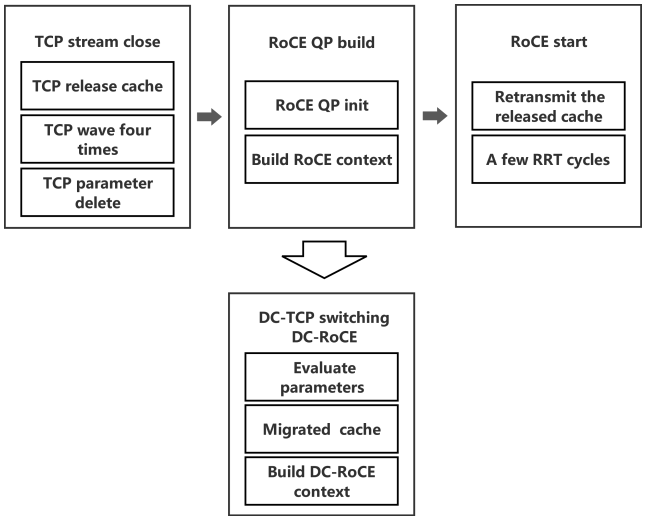

Figure 6: Figure of TCP Switching RoCE.

The packet loss of RoCE means that the overall performance of RoCE has been overloaded, and the transmission method needs to be adjusted. RoCE packet loss usually occurs when the network is severely congested.

- RTT is significantly prolonged

Since it may be a short burst traffic, it cannot be judged by instantaneous RTT, but should evaluate the long-term pressure of the overall network.

- Receive ECN

When the ECN is received, it indicates that the RoCE receiving end finds packet loss, which also means that network congestion occurs in the closer part of the receiving end.

- High-frequency back-pressure appears

Back pressure means that there is congestion close to the RoCE sender. Low-frequency back pressure may be caused by burst traffic. If high-frequency back-pressure occurs, you need to switch to TCP.

4 DC NETWORK TRANSMISSION MODE HANDOVER

It is a simple way to directly close the current connection and re-establish a new connection to switch the transmission mode. However, TCP slow start takes a long time, which seriously affects efficiency. The connection of reestablishing should not be closed. Instead, the current network status should be used to evaluate the parameters of the new connection method. At the same time, the existing cache is migrated to the new transmission mode cache.

As shown in the above figure 5, handover evolution of DC-RoCE to DC-TCP. The new handover method reduces unnecessary retransmissions and TCP slow start, and improves handover performance.

As shown in the above figure 6, handover evolution of DC-TCP to DC-RoCE. The new handover method reduces unnecessary retransmissions and RoCE initialization, and improves handover performance.

The DC-TCP context includes: congestion window, congestion parameter, RTT, RTT change rate, TCP buffer, local SN number, peer SN number, etc. The DC-RoCE context includes: RTT, ECN, RTT fluctuation, RoCE bandwidth, RoCE cache, local RoCE number, peer RoCE number, etc. During the switching process, these states need to be one-to-one correspondence with the context, and a suitable correspondence and conversion method must be found, so that the parameters after the switch are more in line with the current network state, and the performance consumption of restarting is avoided.

4.1 DC-RoCE Switching DC-TCP Process

To switch between DC-RoCE and DC-TCP, the status of DC-TCP must be evaluated through RoCE status, and the corresponding context should be transformed and migrated. The steps of the handover process are as follows:

- The DC-RoCE middleware monitors RoCE end-to-end flow control indicators.
- When the flow control index reaches the switching threshold, initiate the switching process from DC-RoCE to DC-TCP.
- Switch all streams with the same priority end-to-end to DC-TCP.
- DC-RoCE switches the corresponding context to DC-TCP.
- Convert and migrate the DC-RoCE context to DC-TCP.

4.2 DC-TCP Switching DC-RoCE Process

To switch between DC-TCP and DC-RoCE, the status of DC- RoCE must be evaluated DC-TCP through status, and the corresponding context should be transformed and migrated. The steps of the handover process are as follows:

- DC-TCP middleware monitors DC-TCP end-to-end flow control indicators.
- When the flow control index reaches the switching threshold, initiate the switching process from DC-TCP to DC-RoCE.

Table 3: Initialize the Connection Algorithm

NO.	Initialize the connection algorithm
1	The system allocates DC-TCP and DC-RoCE exclusive resources
2	Server test whether it supports DC-RoCE
3	**IF** support
4	Use DC-RoCE connection
5	**IF** RTT of DC-RoCE exceeds constant A
6	Disconnect, use DC-TCP connection
7	**ELSE**
8	Use DC-RoCE connection
9	**ELSE**
10	Use DC-TCP connection

Table 4: DC-RoCE State Switching Algorithm

NO.	DC-RoCE state switching algorithm
1	**IF** The receiving end receives ECN frequency higher than constant A
2	Switch to DC-TCP
3	**IF** Back-pressure frequency is higher than constant B
4	Switch to DC-TCP
5	**IF** The increase in RTT exceeds the constant C
6	Switch to DC-TCP
7	**IF** The overall load exceeds the capacity multiplied by the constant D
8	Switch to DC-TCP

- Switch all streams with the same priority from end to end to DC-RoCE.
- DC-TCP switches the corresponding context to DC-RoCE.
- Convert and migrate the DC-TCP context to DC-RoCE.

5 NETWORK STATE DRIVEN TRANSMISSION MODE SCHEDULING ALGORITHM

(1) Initialize the connection algorithm
The network transmission initialization algorithm is shown in the following table 3

(2) DC-RoCE state switching algorithm

DC-RoCE state switching algorithm is shown in the following table 4

(3) DC-TCP state switching algorithm

DC- TCP state switching algorithm is shown in the following table 5

6 SIMULATION EXPERIMENT AND ANALYSIS

The simulation experiment (Discrete event simulation) simulates and compares DC latency, bandwidth, and CPU usage in three network environments: good, medium, and bad. In the medium and bad network environment, there are both network congestion and non-congestion. The simulation experiment uses three communication methods: RoCE, data center transmission based on network awareness, and TCP. In order to simplify the expression, MIX is used to represent data center transmission based on network awareness.

Table 5: DC-TCP State Switching Algorithm

NO.	DC-TCP state switching algorithm
1	**IF** Did not receive duplicate ACK for a long time
2	Switch to DC-RoCE
3	**IF** Long time congestion sliding window not reaching limit
4	Switch to DC-RoCE

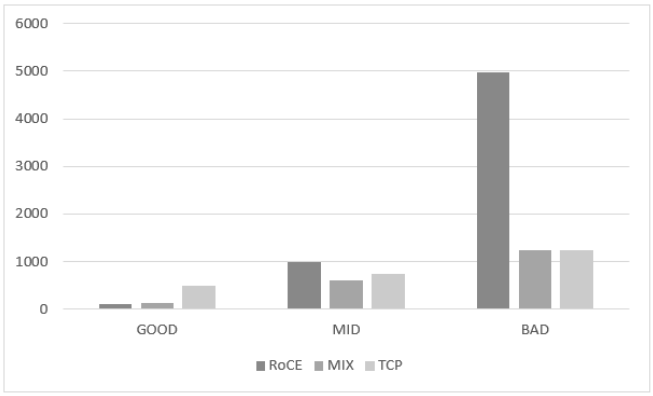

Figure 7: Figure of Total Delay (Unit: Milliseconds).

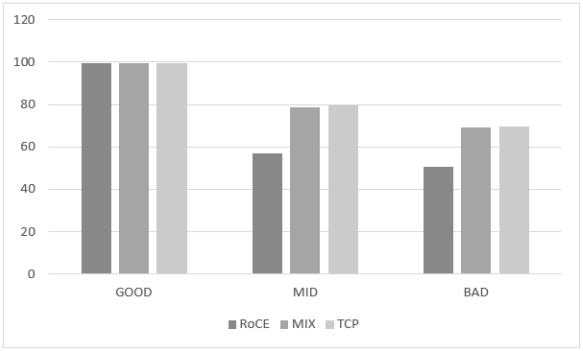

Figure 8: Figure of Average Bandwidth (Unit: Gb/s).

As shown in the above figure 7, when the network environment is good, the RoCE and MIX delays are similar and significantly shorter than TCP; when the network environment is medium, the MIX delays are the shortest; when the network environment is bad, the MIX and TCP delays are similar and significantly shorter In RoCE. Note that in the medium environment, MIX combines the advantages of the two transmissions and performs best.

As shown in the above figure 8, when the network environment is good, the three methods are all excellent; when the network environment is bad, the bandwidth of MIX and TCP is similar, and it is significantly larger than RoCE.

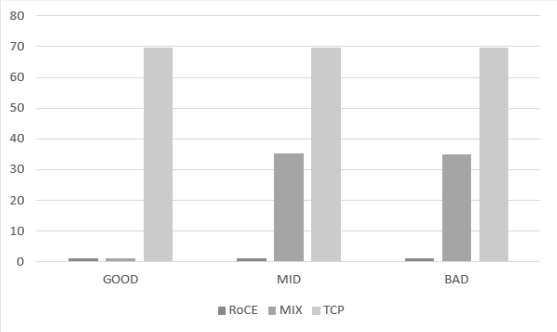

Figure 9: Figure of Average CPU Usage (Unit: Percent).

Table 6: ABBREVIATIONS TABLE

Abbreviations	Definitions
DC	Data center
RoCE	RDMA over Converged Ethernet
DPDK	Data Plane Development Kit

As shown in the above figure 9, when the network environment is good, the RoCE and MIX CPU usage are similar and significantly shorter than TCP; when the network environment is medium, the MIX CPU usage is medium; when the network environment is bad, the MIX CPU usage is medium. The CPU usage of MIX is significantly lower than that of TCP, the same as RoCE or higher than RoCE.

In summary, when the network environment is good, RoCE latency, bandwidth, and CPU usage are all excellent. However, when the network environment is poor, RoCE delay and bandwidth become the worst and cannot be used normally. MIX delay and bandwidth are better than RoCE, and the performance is similar to TCP, but it has a lower CPU usage than TCP. When the network environment is medium, RoCE delay and bandwidth is the worst and cannot be used normally. MIX combines the advantages of two transmissions in one with the lowest delay, and the CPU usage is much lower than TCP, and its performance is close to that of TCP, Higher than RoCE.

7 CONCLUSION

By studying the pain points of DC transmission, this paper proposes a DC high-performance network transmission technology based on network awareness. Simulation experiments have verified that the new technology has the characteristics of low latency, high bandwidth, and low CPU (compared to TCP) usage in a complex network environment. The next work will shift to the conversion of new technologies into engineering practices and the development of related commercial software and hardware.

ACKNOWLEDGMENTS

This work was supported in part by National Key R&D Program of China (2020YFB1805505), and 2019 Quancheng "5150" Talents Doubles Plan (Innovative Talents).

REFERENCES

[1] Putnam, A. , Caulfield, A. M. , Chung, E. S. , Chiou, D. , Constantinides, K. , & Demme, J. , *et al.* (2014). A reconfigurable fabric for accelerating large-scale datacenter services. Computer architecture news, 42(3), 13-24.

[2] Bari, M. F. , Boutaba, R. , Esteves, R. , Granville, L. Z. , Podlesny, M. , & Rabbani, M. G. , *et al.* (2013). Data center network virtualization: a survey. IEEE Communications Surveys & Tutorials, 15(2), 909-928.

[3] Mckeown, N. , Prabhakar, B. , & Shenker, S. . (2013). Pfabric: minimal near-optimal datacenter transport. Acm Sigcomm Computer Communication Review, 43(4), 435-446.

[4Edsall,] Tom, Fingerhut, Andy, Varghese, & George等. (2014). Conga: distributed congestion-aware load balancing for datacenters. Computer Communication Review A Quarterly Publication of the Special Interest Group on Data Communication.

[5] Wu, H. X. , Yang, X. L. , & Zhang, M. . (2013). Congestion control in data center networks: a survey and new perspectives. Applied Mechanics & Materials, 462-463, 1028-1035.

[6] Guo, B. , Shang, Y. , Zhang, Y. , Li, W. , Yin, S. , & Zhang, Y. , *et al.* (2019). Timeslot switching-based optical bypass in data center for intrarack elephant flow with an ultrafast dpdk-enabled timeslot allocator. Journal of Lightwave Technology, 37(10), 2253-2260.

[7] Lockwood, J. W. , & Monga, M. . (2015). Implementing Ultra Low Latency Data Center Services with Programmable Logic. IEEE Computer Society, 68-77.

[8] Tewari, M. . (2015). Enabling fine-grained network flow management in data center networks and servers. Dissertations & Theses - Gradworks.

[9] Kai, L. I. , Lin, Y. E. , Xiangzhan, Y. U. , & Yang, H. U. . (2017). Traffic dynamic load balancing method based on dpdk. Intelligent Computer and Applications.

[10] Tian, F. , Feng, W. , Zhang, Y. , & Zhang, Z. L. . (2020). A novel software-based multi-path rdma solutionfor data center networks.

[11] Xue, J. , Chaudhry, M. U. , Vamanan, B. , Vijaykumar, T. N. , & Thottethodi, M. . (2020). Dart: divide and specialize for fast response to congestion in rdma-based datacenter networks. IEEE/ACM Transactions on Networking, PP(99), 1-14.

[12] Barak, D. . Haifux club: infiniband, roce and rdma verbs - empowering supercomputing and data center interconnects.

[13] Hu, S. , Zhu, Y. , Peng, C. , Guo, C. , & Kai, C. . (2017). Tagger: Practical PFC Deadlock Prevention in Data Center Networks. International Conference.

[14] Zu, J. , Hu, G. , Yan, J. , & Tang, S. . (2021). A community detection based approach for service function chain online placement in data center network. Computer Communications, 169(1).

[15] Alaluna, M. , Vial, E. , Neves, N. , & Ramos, F. M. V. . (2019). Secure multi-cloud network virtualization. Computer Networks, 161.

[16] Li, W. , Liu, J. , Wang, S. , Zhang, T. , & Huang, J. . (2021). Survey on traffic management in data center network: from link layer to application layer. IEEE Access, PP(99), 1-1.

[17] Wang, Y., Kan, H., Su, D., Shen, Y., Liu, W., & Ou, M. (2020, November). Energy efficient computing offloading mechanism based on FPGA cluster for edge cloud. In Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering (pp. 1120-1125).

[18] Guoqiang, M., Rui, H., Jiangwei, W., Hongwei, K., & Rengang, L. (2020, October). A FPGA based intra-parallel architecture for PageRank graph processing. In 2020 IEEE International Conference on Edge Computing (EDGE) (pp. 31-38). IEEE.