Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

Technology trends in large-scale high-efficiency network computing*

Jinshu SU†‡[1,2], Baokang ZHAO†‡[1], Yi DAI[1], Jijun CAO[1], Ziling WEI[1], Na ZHAO[1], Congxi SONG[1], Yujing LIU[1], Yusheng XIA[2]

[1]*School of Computer, National University of Defense Technology, Changsha 410073, China*

[2]*Academy of Military Science, Beijing 100091, China*

†E-mail: sjs@nudt.edu.cn; bkzhao@nudt.edu.cn

Received May 18, 2022; Revision accepted Oct. 11, 2022; Crosschecked Nov. 15, 2022

Abstract: Network technology is the basis for large-scale high-efficiency network computing, such as supercomputing, cloud computing, big data processing, and artificial intelligence computing. The network technologies of network computing systems in different fields not only learn from each other but also have targeted design and optimization. Considering it comprehensively, three development trends, i.e., integration, differentiation, and optimization, are summarized in this paper for network technologies in different fields. Integration reflects that there are no clear boundaries for network technologies in different fields, differentiation reflects that there are some unique solutions in different application fields or innovative solutions under new application requirements, and optimization reflects that there are some optimizations for specific scenarios. This paper can help academic researchers consider what should be done in the future and industry personnel consider how to build efficient practical network systems.

Key words: Supercomputing; Cloud computing; Network technology; Development trends
https://doi.org/10.1631/FITEE.2200217 **CLC number:** TP393.0

1 Introduction

Cloud computing, big data, and artificial intelligence (AI) have endless demands for massive information processing and transmission capabilities. To meet the resource demands of the development of applications, it is urgent to connect massive processing nodes into a larger system by various network technologies. The above large computing system is called large-scale high-efficiency network computing.

From the perspective of the network, we investigate the network technologies used in various large-scale high-efficiency network computing, hoping to deeply analyze the typical network characteristics and provide a design basis for the development of network technologies to support high-efficiency network computing.

With the large increase in the number of Internet users, the Internet has become a complex giant system. At present, there are 4.57 billion Internet users, accounting for 58% of the world's population. The Internet has three important characteristics.

First, networking technologies are changing rapidly. Ethernet has developed into a rich family, from 100 Mbps, 1 Gbps, 2.5 Gbps, 10 Gbps, 40 Gbps, 100 Gbps, to 400 Gbps. It makes a variety of technologies coexist, while some technologies (e.g., Token Ring, FDDI) disappear. In April 2022, the number of core protocol specifications of the Internet (i.e., RFC) was larger than 9000 (https://rfc.ietf.org). There are

‡ Corresponding authors

* Project supported by the National Natural Science Foundation of China (Nos. 61972412, 62202486, and 12102468)

ORCID: Jinshu SU, https://orcid.org/0000-0001-9273-616X; Baokang ZHAO, https://orcid.org/0000-0001-9200-9018

more than 70 000 border gateway protocol (BGP) routing domains on the Internet. There are more than 100 million routers at the carrier and enterprise levels, not including home routers.

Second, there is a rapid development of data centers. In 2021, there were 326 ultralarge data centers in the world. The data transmission rate of data centers reaches 1.7 ZB/month, that is, 52.47 Pbps. In addition, the global data centers have 524 691 100-Gbps links, with an average of 830 100-Gbps exports per data center.

Third, the computing power of the Internet is developing at an unimaginable speed. Taking the computing power of the global TOP500 as an example, both the computing power of the fastest supercomputer and the total computing power rise almost linearly, following the 10-fold law. Fig. 1 shows the computing power of the TOP500 in different years. The green line indicates the total computing power of the TOP500, the brown line indicates the computing power of the fastest supercomputer, and the blue line indicates the computing power of the supercomputer ranking last in each year. As shown in Fig. 1, the TOP500 was first proposed in 1993, with a total computing power of less than 1 TFlops. However, in 2021, the computing power reached exascale. At the end of 2021, among the TOP500, China ranked first in the total number of units, and the United States ranked first in the total computing power (https://www.top500.org).

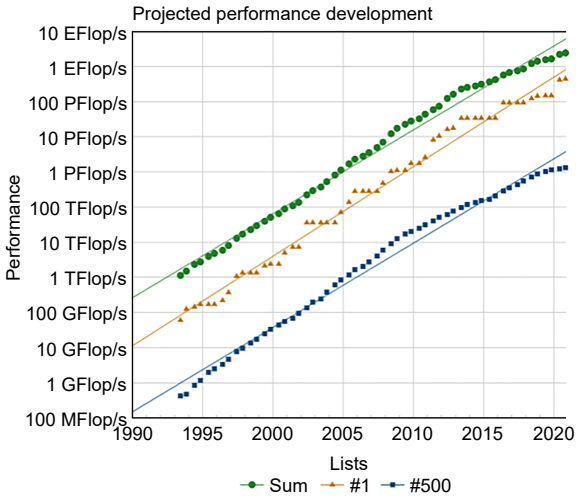

Fig. 1 Diagram of supercomputing TOP500 performance from 1993 to 2021

The graph is generated from the www. top500. org website. References to color refer to the online version of this figure

Under the premise of constant computing power, the network performance and network bandwidth become the decisive factors for data processing. The data provided by Mellanox indicate that using 100 Gigabit Ethernet (GbE) is 6.5 times faster than using 10 GbE under constant computing power. Thus, network performance has become a widely studied capability (Guo et al., 2016).

At the same time, the ratio of network bandwidth to CPU computing power is changing dramatically. Before 2010, the annual growth rate of network bandwidth was approximately 30%. Then, it increased slightly to 35% in 2015 and reached 45% in 2020. However, the growth rate of CPU computing power dropped from 23% before 2010 to 12% before 2015 and then to an average annual growth rate of 3.5% in recent years. The above trend brings benefits and opportunities to heterogeneous network engines.

To this end, the high speed and diversification of the network have become the development trend, and it is necessary to explore new models and new paths for the development of network technologies. Prof. Jiangxing WU proposed that the existing network technology development paradigm of self-evolution can no longer meet new development needs. It must be completely reformed from the perspective of thinking, methodology, and practice norms. Wu (2022) and Ji et al. (2022) proposed a polymorphic intelligent network environment (PINE, multimodal network for short). It provides a new paradigm and new ideas for the research and development of network technology.

To well investigate the current state of the network technology involved in high-efficiency network computing, we divide the network technologies into three fields: supercomputing, data centers, and the Internet (a narrow sense). The corresponding technical characteristics are summarized in Table 1. As shown in Table 1, the requirements for the network computing techniques are different in different application fields.

Overall, the development trend of network technology is integration, differentiation, and optimization. From a macro point of view, the performance demands are similar under different scenarios. For example, high throughput and low latency are the main goals for both the data center and supercomputing. Thus, techniques that can support common goals can achieve rapid development. There are no clear boundaries in

Table 1 Technical characteristics of large-scale high-efficiency network computing

Performance indicator	Supercomputing (intranet)	Supercomputing (extranet and auxiliary network)	Data center	Internet
Performance	Extremely high	High	High	Middle
Scalability	Middle	Middle	Extremely high (smooth upgrading)	Extremely high (smooth upgrading)
Compatibility	Middle	High	High	Extremely high
Security	Middle	Middle	Extremely high	High
Operation and maintenance	High	High	Extremely high	High

network technologies in different fields. In this case, the techniques of different systems show a trend towards integration. However, from a microscopic point of view, within computer networks, many new technologies have evolved according to the goals pursued by different applications. Due to the fertile soil of the Internet, nutrition is improving, and the thin waist of the Internet is gradually transitioning to a thick waist. In other words, it shows the trend of differentiation. To satisfy the high performance requirement, optimization is a permanent topic.

On one hand, many fields of high-efficiency network computing, such as Internet point of presence (POP), scientific computing, cloud computing, large-scale data processing, and AI computing, show the trend of network technology integration, complementarity, or even unity. On the other hand, to pursue specific performance demands, there is a trend of differentiation. Generally, network technologies in the field of high-efficiency network computing show a trend of coexistence of integration, differentiation, and optimization. For example, for the data center and supercomputing, 100/400 Gbps Ethernet is widely adopted as it is highly cost-effective, which shows the trend of integration. However, for different special demands, we need to take some targeted designs. It shows the trend of differentiation. For example, under the case of low latency, the InfiniBand (IB) technique is introduced. In the case of security transmission, the quick user datagram protocol (UDP) Internet connections (QUIC) mechanism is introduced. To satisfy the high performance requirement, continuing efforts are made to optimize the different kinds of techniques. For example, in-network computing is proposed to decrease the service latency and increase the throughput. It shows the trend of optimization. Due to space limitations, this paper focuses mainly on the point-to-point and end-to-end issues in the network, and other networking issues such as software-defined network (SDN) will be discussed in another paper.

2 Integration trend

The integration trend is manifested mainly in the network layer, link layer, and physical layer.

2.1 Heterogeneous network complementarity

Taking high-performance computing (HPC) as an example, the inherent networking requirements are multifaceted, and heterogeneous network integration is a trend. The interconnection between computing nodes pursues high bandwidth, low latency, and zero packet loss. However, the connection between storage systems and visualization systems requires high bandwidth and compatibility. For the connection of the control, debugging, and diagnosis systems, generality instead of high bandwidth is needed. The origin of the heterogeneous network complementation is implemented in IBM's Blue Gene's L-series HPC system. In the development process, Blue Gene's Q series, Tianhe computer system, etc., are also milestones for heterogeneous networks in HPCs.

There are five types of networks in the Bule Gene L series (Coteus et al., 2005). The first one, named 3D-Torus, connects all computing nodes, providing high-bandwidth and low-latency connections for large-scale computing. The global collective network and the global interrupt network are designed together with 3D-Torus, occupying 2–4 bits, and they perform global aggregation operations. The fourth is the I/O network, which uses GbE technology to access the external storage network. The fifth is the service network, which uses fast Ethernet to connect all joint test action group (JTAG) information. JTAG is used to diagnose hardware at the signal level.

Fig. 2a is a diagram of four RACKs. Each RACK is divided into the upper part and the lower part. The red part is the midplane, connecting 16 nodes. Each node includes four communication nodes and 32 computing nodes. Fig. 2b is a top-level view of the Blue Gene L control network, connecting multiple Ethernet switches. Fig. 2c is the configuration diagram of half RACK. The midplane includes a service card, a built-in Ethernet switch connecting 16 nodes, and four link control CFPGAs.

The Q series developed by IBM in 2012 is a milestone in the complementarity of heterogeneous networks. In the Q series, in addition to maintaining five types of networks, IBM has increased the interconnection of computing nodes from 3D to 5D, thereby reducing the radius of the system and the latency between nodes. In the service node, the Q series adopts 10 GbE or IB as the interconnection of the file server nodes, which improves the compatibility, versatility, and performance.

In the Tianhe supercomputer system, the storage network and the control network are also well designed. The former is used for connecting the storage systems, and the latter is used to collect the running status of each node.

2.2 Generalization of link technology

It is shown that the generalized interconnection technology occupies the majority of the massively parallel processing (MPP) link technology. From the changes in the TOP500 interconnection technology, it can be observed that the trend is to improve the link speed and shorten the point-to-point delay. The main method to shorten the delay is to reduce the connection radius of the MPP system and improve the difference between the local memory access time and the remote node access time, such as Thinking Machine's 2D mesh, Cray's 3D-Torus, FatTree structure, and the hybrid network structure of 5D and other multiple networks in IBM Blue Gene.

As shown in Fig. 3, the proportion of general technology has gradually increased. For example, 415 units, accounting for 83% of the list in November 2021, use IB or Ethernet technology. At the same time, among the top 10 supercomputer systems in the TOP500, systems with customized and proprietary interconnection networks account for up to 40%. With the rapid development of Ethernet and IB network technologies, Myrinet (Boden et al., 1995), Quadrics (Petrini et al., 2002), SP Switch, NUMAlink, and other interconnection networks used in early HPC systems have been gradually withdrawn from the historical stage.

2.3 Link transmission unification

At present, the technique of the link layer shows unprecedented unity, and SerDes technology is widely used. SerDes technology was originally used in optical fiber communication. With increasing speed, high-speed serial interfaces have become the mainstream. The serial interface uses mainly differential signal transmission technology, which has the characteristics of low power consumption, strong anti-interference, and high speed. SerDes technology can be divided into

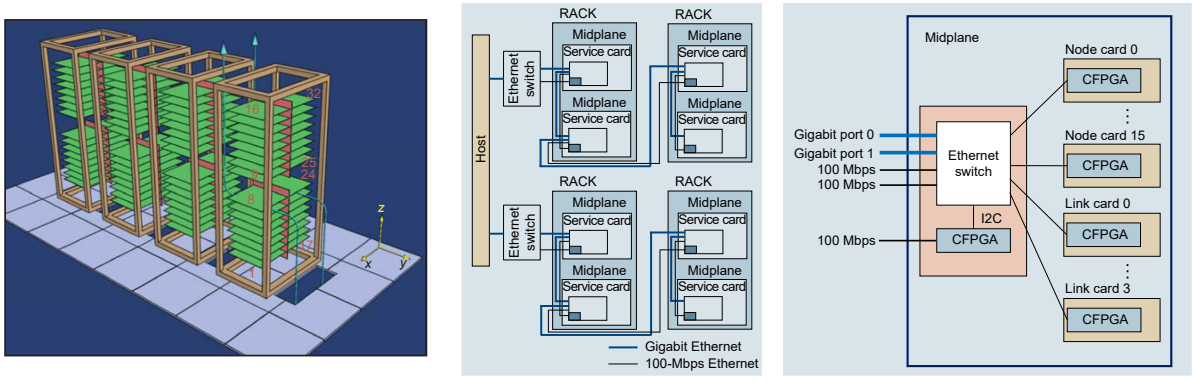

Fig. 2 Diagram of IBM Bule Gene Q series heterogeneous networks: (a) diagram of four RACKs; (b) top-level view of the Blue Gene L control network; (c) configuration diagram of half RACK (Coteus et al., 2005)

References to color refer to the online version of this figure

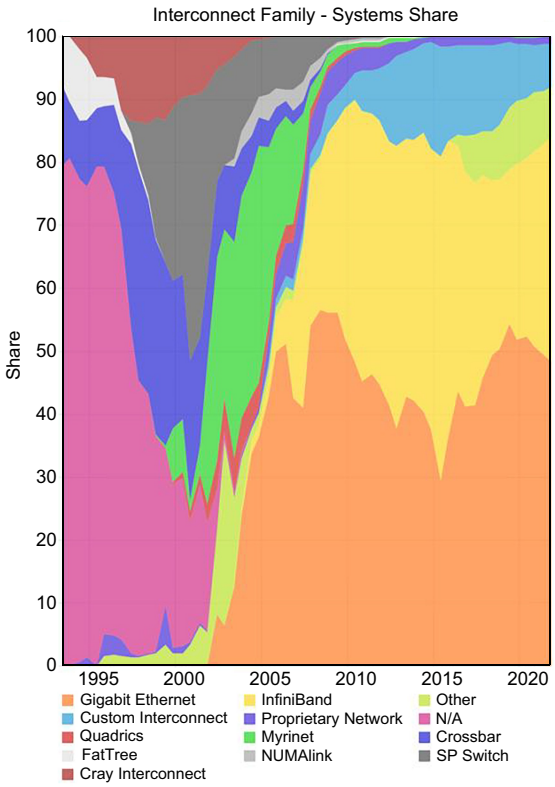

Fig. 3 Diagram of the network technology trend in TOP500 supercomputing

The graph is generated from the www.top500.org website. References to color refer to the online version of this figure

four categories: (1) parallel clock SerDes, which serializes the parallel wide bus into a plurality of differential signal pairs and transmits the clock in parallel with the data; (2) 8B/10B encoding SerDes, which maps each data byte to a 10-bit code, and then serializes it into a single signal pair; (3) embedded clock SerDes, which serializes the data bus and clock into a serial signal pair; (4) bit-interleaved SerDes, which aggregates bits from multiple input serial streams into faster serial signal pairs. The disadvantage of SerDes technology is that it requires super precise and ultralow jitter components to provide the reference clocks needed to control high data rate serial signals.

In summary, SerDes technology is widely used because of its high bandwidth, low signal number, and many other benefits (such as reduced routing conflicts, reduced switching noise, lower power consumption, and low packaging costs). Applications include PCI-e Gen3/4/5, 4/8/16 Lane, Ethernet 40/100/400/800 Gbps, IB, SATA, etc.

3 Differentiation trend

The transport layer, network layer, and link layer have targeted designs for special needs, for example, the advanced and customized interconnection system developed for high-end supercomputing, the integrated design of internal routers and network cards for data centers, and the implementation HTTP/3 based on QUIC and transport layer security (TLS) for optimizing data security transmission efficiency.

3.1 High-end system customization

Among the top 10 systems of the TOP500, to meet the urgent need for communication performance brought by the exponential improvement of HPC performance, advanced enabling technologies are widely adopted in high-end systems and customized interconnection networks. In addition, it can be specifically optimized for the communication characteristics of HPC applications. The above approaches are adopted by top HPC systems, even at a high cost. For example, among the 57th HPC TOP500 list of supercomputing systems released in June 2021, 7.4% and 1.2% of the listed systems use customized interconnection and proprietary interconnection respectively, but their performance share ratios are as high as 11.59% and 17.66% respectively. Among the top 10 systems, customized and proprietary interconnection networks account for as much as 40%. Typical high-end customized and proprietary supercomputing interconnection technologies include Cray XC30 and Slingshot interconnects (de Sensi et al., 2020), TH Express (Liao et al., 2015), Tofu Interconnect 2 (Ajima et al., 2014), and Bull BXI Interconnect (Derradji et al., 2015).

The high-index routing chip of the Tianhe custom network adopts the tile-based multiport binding scalable switching architecture MBTR (Dai et al., 2019). By integrating multiple physical port-related buffering and arbitration logic resources in a single tile, the on-chip storage overhead can be reduced by 50%–75% and can achieve 100% throughput. In addition, with the hierarchical single-cycle high-order arbitration mechanism, the critical path transmission delay on network on chip (NoC) is only 30 ns. On the other hand, with the single-lane bandwidth of the high-speed serial transmission varying from 56 to 112 Gbps, or even 224 Gbps, the link error rate increases with the

increase of link bandwidth. The low latency and high reliability become important challenges for the design of high-index routing chips. To achieve a low-latency and high-bandwidth hardware transmission protocol stack, the Tianhe interconnect network adopts the lane-adaptive multilink dynamic binding physical code sublayer (PCS), low-latency forward error correction (FEC) code, efficient retransmission at the link layer based on the sliding windows mechanism, and full connection/semiconnection end-to-end reliable data transmission.

The development trend is of high order and high redundancy. Customized and proprietary HPC interconnection networks typically use a tiled distributed switching fabric, where the number of physical ports per tile is referred to as the order of the higher-order switch chip. In HPC systems, the use of high-level switching chips to build an interconnection network can reduce the radix of the system interconnection and the average number of hops in communication between nodes, thereby reducing the communication delay between nodes. High redundancy can not only provide flexibility for interconnecting network topology design but also improve reliability through redundant links. As predicted by Kim et al. (2005), with the continuous progress of integrated circuit technology, switching chips are constantly developing in a higher-order direction. However, switching chips also face many challenges, such as the design of high-throughput scalable switching structures and scalable network topology design.

3.2 Integration design for the network and interface

In pursuit of high throughput, low latency, low cost, and easy management, data center networks generally use a single network technology, which is different from high-end HPC. Google proposed a tightly coupled data center network (Gibson et al., 2022), which changed the hierarchical design approach. It designs the network card and the top of the cabinet switch together and compresses the transmission delay. In this case, it shows the following features: predictable, high bandwidth, and low latency.

Google takes the TiN ASIC chip instead of the CPU chip as the core of the computing system. Each TiN provides three kinds of networking methods: (1) Two 16*PCIe Gen3.0s are connected to two servers. (2) One 100-Gbps Ethernet is connected to the classic backbone switch of the data center, thus incorporating the server pool of the new structure into the computing, networking, and storage capabilities of the overall data center. (3) Among 32 dedicated links of 25 Gbps, eight are used for interconnection between points of delivery (PODs), and 24 are used for interconnection within PODs.

The basic module of POD contains two TiN ASIC chips and four servers. Six basic modules constitute a POD. The POD has excellent networking and computing capabilities. The POD has 12 100-Gbps Ethernet interfaces to connect with other computer systems in the data center. Each chip has 600 Gbps interconnection capability, a total of 7.2 Tbps intra-POD connection capability, and 9.6 Tbps inter-POD connection capability.

The innovation of this structure is manifested in three aspects. First, the traditional interconnect routing node and the network interface card are combined into one chip while supporting three types of protocols. Second, the three stages of delay from the bus to the network interface card and then to the routing chip are compressed to one stage so that the delay is greatly shortened. Third, a dedicated low-latency data center interconnect link is proposed and designed.

Google established a prototype system consisting of 576 TiNs and 500 servers. It is tested under the delay-based congestion control scenario. When the network card reaches wire-speed traffic, the delay of the switching fabric can be maintained at 40 μs. If the load is under 70%, the latency is less than 20 μs. In addition, it is proven that we can effectively isolate low-latency traffic data from high-latency traffic data by adopting two load modes.

3.3 Transmission protocol customization

Transmission control protocol (TCP) is a common transmission protocol for the Internet, data centers, and HPC systems. TCP has always been considered unshakable in terms of reliable transmission. However, the problem of restricting the application development has always been hoped to be overcome for TCP. For example, since the TCP is usually implemented in kernel mode, its update and expansion will result in the update of the entire kernel, resulting in a lower speed of iteration of the TCP.

"Many-to-one" data transfer is common in data centers, so it is easy to produce the TCP incast problem (a catastrophic TCP throughput collapse). In this case, the data center TCP (DCTCP) and the incast congestion control for TCP (ICTCP) were born. As the scale of data centers continues to expand, more efficient transmission protocols need to be proposed to meet the characteristics of data center scenarios (such as multipath transmission and differentiated application requirements).

Hypertext transfer protocol (HTTP), hypertext connection, and uniform resource descriptor provide the technical basis for the popularity of web on the Internet. As security concerns grow, hypertext transfer protocol secure (HTTPS) over TLS has become the standard. However, TCP plus TLS results in a large transmission delay. At the same time, a TCP session can be transmitted only in sequence, not concurrently. Therefore, there is a tendency for differentiation at the transport layer. The HPC developed the efficient transport layer protocol of RDMA several years ago. In the following subsections, we describe mainly the secure transmission technologies such as QUIC and MPQUIC in recent years. QUIC and MPQUIC show rapid development. QUIC has advantages over TCP, such as protocol entrenchment, implementation entrenchment, handshake delay, and head-of-line blocking delay. However, most existing applications are designed to focus on TCP. In addition, QUIC shows some deficiencies in some scenarios. For example, the firewall cannot decrypt the QUIC traffic for packet inspection, and thus malicious traffic easily enters the network. Thus, some researchers still believe that TCP cannot be replaced by QUIC in the short term. In the long term, QUIC instead of TCP may be the mainstream. In addition, this paper focuses on high-efficiency computing networks. In this case, in the following subsections, QUIC/MPQUIC is still discussed under scenarios of network transport protocols in high-efficiency computing domains.

3.3.1 QUIC mechanism

Although TCP is the basic protocol of the Internet, breaking through the limitations of TCP has attracted much attention due to the need for secure transmission. In particular, since 1 to 3 round trip time (RTT) delays are required in the TCP plus TLS solution, the approach is widely criticized.

In response to this demand, Google proposed the QUIC protocol, which is a UDP-based transport layer protocol. It was deployed in 2012. After nearly 10 years of running-in, the IETF finally launched the standard RFC9000 in May 2021 (Iyengar and Thomson, 2021). An RFC lasting 12 years is extremely rare. At present, the industry has designed a new QUIC protocol stack based on the flexibility of the user mode implementation of the QUIC protocol to support new application requirements of various businesses, such as streaming media transmission, web page loading, and real-time virtual reality/augmented reality (VR/AR) interaction.

The QUIC protocol is implemented in user mode, providing the possibility to flexibly customize protocol functions according to user needs. In addition, the QUIC protocol can provide reliable and secure encrypted transmission based on 0-1 RTT low-latency connection establishment. The QUIC protocol currently has 24 kinds of protocol implementations. Chrome enables the use of the QUIC protocol by default. Microsoft Edge, Firefox, and Safari also provide QUIC protocol support. In 2018, the HTTP working group and QUIC working group of the IETF jointly released the QUIC-based protocol stack, i.e., HTTP over QUIC, later renamed HTTP/3 (Bishop, 2021). As of 2019, 4.6% of websites (approximately 9.1% of traffic) used QUIC, and 42.1% of Google's traffic was transmitted through the QUIC protocol. As of 2021, 75% of Facebook's traffic was transmitted through QUIC (https://www.ietf.org/blog/quic-industry/).

The QUIC protocol stack includes the transport layer, TLS, and part of the application layer. The header is unencrypted text, and the content part is encrypted. At the same time, QUIC supports multistream multiplexing on a single connection by logically opening multiple streams concurrently between applications, thereby alleviating the head-of-line blocking problem. The QUIC protocol encrypts the packet content to avoid tampering with the packet content by the middleware. The sequence number of QUIC messages is increased, and the sequence number of retransmission is different from the previous message sequence number to avoid ambiguity caused by retransmissions. At the same time, the ACK frame

contains the round-trip delay for more accurate RTT measurement. QUIC connections are identified by connection-ID instead of traditional IP/port five-tuple, which can achieve more convenient connection migration (Langley et al., 2017).

3.3.2 MPQUIC mechanism

The MPQUIC protocol extends multipath on the basis of the QUIC protocol to provide high bandwidth and reliability for transmission. As we all know, fixed network, Wi-Fi, mobile communication, satellite communication, and some other networking approaches are available at the same time in many occasions. To this end, it supports the establishment of multiple paths in one transmission connection at the same time, realizes concurrent transmission, and provides alternate paths, having the advantages of improving network link utilization, enhancing network robustness and multipath bandwidth aggregation, supporting mobility, and improving transmission reliability. Although the IETF proposed the multipath transmission protocol MPTCP in 2013 (Ford et al., 2020), it cannot solve the problems of concurrent transmission and slow connection.

de Coninck and Bonaventure (2017) proposed a multipath transmission MPQUIC based on the QUIC protocol and a draft of the MPQUIC technology, which was updated to version 01 (Liu et al., 2022) in February 2022. The QUIC protocol has been widely adopted due to its flexibility and low latency, although the application of MPQUIC technology is still in its infancy.

de Coninck and Bonaventure (2021) extended multipath transmission based on the plug-in protocol implementation of MPQUIC and then proposed MFQUIC. MFQUIC regards bidirectional paths as two unidirectional flows, and each unidirectional flow is indexed by UCID. This design can effectively improve the transmission efficiency of networks with large differences in uplink and downlink (such as ADSL and satellite network). Based on the self-developed XQUIC protocol stack, the Alibaba Tao Department technical team proposed a multipath transmission XLINK architecture for the short video application (Zheng et al., 2021). The architecture includes priority-oriented multigranularity reinjection scheduling algorithms, packet scheduling algorithms based on user experience feedback, and path control algorithms. By designing the extension fields of MPQUIC packets and adopting a fast path for responses, it can reduce the first video frame latency and the playback delay of short videos, and guarantee the balance between transmission performance and overhead. XLINK has passed more than three million short video transmission experiments. The test shows that compared with single-path QUIC, XLINK reduces the request completion time by 19%–50%, the first frame delay by 32%, and the buffer rate by 23%–67%, while introducing only 2.1% redundant traffic. The team submitted XLINK as a draft, updated to version 04 (Liu et al., 2020) as of October 2021.

In terms of the MPQUIC packet scheduling mechanism, the default scheduling algorithms include the minRTT algorithm that preferentially selects the path with the lowest latency and the round robin algorithm that selects paths in a round robin manner. In both algorithms, the latency and fairness are taken into account when scheduling packets. However, in a heterogeneous network environment, when packets are transmitted to the receiving end by paths of different qualities, a large number of out-of-order packets will be generated, causing serious problems of head-of-line blocking and buffer expansion. To better adapt to the transmission of heterogeneous links, Ferlin et al. (2016) and Lim et al. (2017) designed the BLEST and ECF packet scheduling algorithms, respectively. The idea of the algorithms is to estimate the congestion status of the path or the download completion time and then to decide to wait or send packets on the slower path.

In addition, due to the combination of multistream concurrency and multipath selection in the MPQUIC protocol, the traditional multipath packet scheduling strategy adds a dimension of scheduling requirements, that is, stream granularity scheduling. Shi et al. (2020) made improvements on the basis of the original work and proposed PStream. At the same time, they found that streams with large amounts of data are more suitable for transmission on the basis of high bandwidth, and that streams with small amounts of data are suitable for transmission on the basis of low latency. In the same year, the work SRPT (Jonglez et al., 2020) adopted the transmission strategy that small streams have high priority in flow granularity scheduling. By combining the ECF path selection strategy, the completion time was reduced

compared with other algorithms in the experiment of loading Wikipedia pages.

4 Optimization trend

With the independent development of computing and networks, some computing operations with low computational loads, such as sum operations and max operations, can be executed in network devices by integrating some computing components into network devices (i.e., switches and routers). In this case, the network devices can well support the demands of high-efficiency network computing. For example, a 32-bit computing operation is integrated into the switch of Mellanox. Thus, in-networking computing or computing-in-networking is proposed. The field of data centers is generally referred to as in-network computing. In the field of supercomputing, it is generally called collective computing. In-network computing in data centers and supercomputers' collective computing use mainly programmable network devices (programmable switch ASICs, network processors, FPGAs, programmable network cards, etc.) to add repetitive and relatively simple computing functions to the data plane of the network, which may be on a chip or device. Protocol offloading aims at relatively complex computing functions, which can further integrate computing and networking capabilities.

4.1 In-network computing in data centers

The innovation of software and hardware has promoted the rapid development of in-network computing (https://conferences.sigcomm.org/sigcomm/2018/workshop-netcompute.html). On the hardware side, many hardware vendors have released programmable products with performance guarantees, such as Barefoot Tofino, Intel FlexPipe, Cavium XPliant, and Netronome Agilio. On the software side, in addition to new network features such as in-network telemetry and layer-4 load balancing, application-level features such as key-value store (KVS) (Jin et al., 2017; Li BJ et al., 2017) and consensus protocols (Dang et al., 2016, 2020) have been proposed. The overhead of in-network computing is usually very small and it does not require additional space, cost, or power.

The service latency in a cloud environment is an important performance metric, and thus reducing latency is important. In-network computing means that a transaction is terminated in its path without reaching the end host for reprocessing, which results in lower latency optimizations.

Another advantage of in-network computing is throughput. The switch's ASIC processes up to 10 billion packets per second at wire speed, thus potentially supporting billions of operations per second. Since in-network computing is an additional function of the network equipment, it is processed in the flow of data, which eliminates the delay of data receiving, data buffering, and data sending caused by additional equipment and reduces energy consumption. For example, on a switch, the energy "cost" of one million KVS queries is less than one watt.

Jin et al. (2017) proposed a key-value storage architecture called NetCache, which exploits the power and flexibility of a new generation of programmable switches to handle queries for hot items and balance the load among storage nodes. It can provide high aggregate throughput and low latency. The authors implemented a NetCache prototype on Barefoot Tofino switches and commodity servers. A single switch can process more than two billion queries per second while consuming only a fraction of hardware resources. The query object is 6.4×10^4 items, the key is 16 bytes, and the value is 128 bytes. For high-performance in-memory KVSs, NetCache increases throughput by 3 to 10 times and reduces latency by up to 50% for 40% of queries.

Li BJ et al. (2017) proposed a high-performance KVS method, KV-Direct, which leverages a programmable NIC to extend the RDMA primitives so that the remote direct key-value can access the host memory directly. A single NIC KV-Direct can achieve up to 1.8×10^8 key-value operations per second, equivalent to the throughput of dozens of CPU cores. Compared to the CPU-based KVS implementation, KV-Direct is three times more power efficient while keeping tail latency below 10 μs. In addition, KV-Direct can achieve approximately linear scalability through multiple network cards. By installing 10 programmable NICs on commodity servers, KV-Direct achieves 1.22 billion KV operations per second.

Li YJ et al. (2019) proposed a reinforcement learning training acceleration solution iSwitch in switches, which transfers gradient aggregation operations from server nodes to network switches. In this case, it can reduce the number of network hops for gradient aggregation. This not only reduces the end-to-end network latency for synchronous training but also improves the convergence speed with fast weight updates for asynchronous training. iSwitch redesigns the distributed reinforcement learning training algorithm and proposes a hierarchical aggregation mechanism to further improve the parallelism and scalability of rack-scale distributed reinforcement learning training.

4.2 Collective computing for supercomputing

As the scale of computing units in supercomputing systems continues to grow, there is an urgent need to open up and develop high-level parallelism. The system architecture must be re-examined from the perspective of massively parallel computing and communication to fully exploit the computing and communication parallelism. In the field of HPC, a design idea of processing data at a suitable position in the system is proposed to reduce the amount of data communication between nodes. From the perspective of architecture, new system components are introduced to reasonably divide the distributed processing of data instead of processing all the data in the local or remote CPU. This collaborative architecture design spans various computing components, networks, and storage infrastructure, ensuring that each component can be regarded as a system accelerator. Thus, it can help improve system efficiency and optimize system performance. Collective computing makes the traditional CPU-centric processing mode evolve into a data-centric processing mode based on network offloading, which accelerates communication and computing.

Massage passing interface (MPI) application statistics made by the HPC Advisory Council (HPCAC) show that the collective communication time of a large number of scientific computing and engineering applications accounts for up to 80% of the total MPI communication time and 60% of the total execution time. Deep learning applications are also sensitive to collective communication performance. Therefore, optimizing the performance of collective communications is critical to the overall performance of HPC and deep learning applications. The collective communication hardware offload technology effectively reduces the multiple transmissions of data between various endpoints by offloading collective communication operations from the CPU to the interconnecting network chip, and improves the execution efficiency of the application. This innovative method of performing computation in the network reduces the amount of data transmitted over the network, greatly reduces the amount of data traffic on the network, and frees up valuable CPU resources for computation. Mellanox proposed the idea of a collective communication coprocessor, which processes data in the process of data transmission. It develops the scalable hierarchical aggregation reduction protocol (SHARP^{TM}) by offloading collective communication operations to Mellanox network chips, using a communication tree to receive and reduce data from source node groups, and distributing the reduction results within the group. Network switching chips or NICs can be used as aggregation nodes of the logical SHARP tree. An aggregation node can join multiple communication trees at the same time, but the communication tree can perform only one operation at a time. The SHARP protocol effectively reduces the MPI collective communication time by offloading the collective communication network, and the NIC data throughput is increased by more than twice. In addition, it releases many CPU resources, realizes a high overlap of computing and communication, and effectively accelerates machine learning applications. For example, after the MPI_ AllReduce operation is offloaded to the network, the communication delay is reduced by 75%. As the node size increases, the communication delay maintains scalable and stable growth. MPI_AllReduce is frequently called in deep learning applications. Therefore, the performance of the benchmark program ResNet, which uses the TensorFlow distributed deep learning framework Horovod, is improved by 16% after SHARP acceleration (Song, 2019).

4.3 Protocol offloading

Different from in-network computing or collective computing, protocol offloading generally achieves

acceleration using relatively complex protocols as individual components after adding physical or logical processing components. For example, the TCP offload engine TOE (Wang et al., 2008), or remote direct memory access (RDMA), is completely offloaded on the network card. By bypassing CPU intervention, low-latency and high-bandwidth direct communication of memory data between nodes is achieved.

In the data center, the commercial Ethernet protocol uses RDMA to support highly reliable and delay-sensitive services. With the rapid development of data center applications such as cloud storage, the standard TCP/IP protocol can no longer meet the data center's requirements, such as high network bandwidth, low latency, and low CPU overhead. In HPC, the RDMA communication mechanism is used to achieve efficient internode communication. However, deploying RDMA in a data center is difficult. Commercial RDMA is deployed using mainly IB technology or dedicated customized network protocols, while most data centers are constructed using IP and Ethernet technologies, which are incompatible with IB protocol stacks. Managers are also reluctant to deploy and manage two separate networks in the data center. In this case, the industry has formulated the RDMA over converged Ethernet (RoCE) standard (InfiniBand Trade Association, 2010) and its upgraded version RoCEv2 (InfiniBand Trade Association, 2014). RoCEv2 retains only the transport layer of IB and uses IP and UDP encapsulation to replace the IB network layer (L3). The second layer (L2) is replaced by Ethernet. The reason for the network performance improvement brought by RoCE is that RDMA offloads the entire transport layer logic to the NIC. Compared with traditional software transport protocols, RDMA can bypass the CPU to achieve direct access to remote memory, thus providing low CPU overhead and zero-copy low-latency communication. Microsoft further proposed a priority-based flow control (PFC) under the IP layer based on differentiated services code point (DSCP), which extends RDMA from Ethernet to the IP layer (L3). It realizes the large-scale deployment of RDMA and the coexistence between RDMA and TCP in data centers. RDMA is used for data center internal communication, and TCP communication is still used between data centers (Guo et al., 2016). With the large-scale deployment of RDMA in data centers, Microsoft data centers (Zhu et al., 2015; Guo et al., 2016), Google data centers, Microsoft Azure cloud computing, Baidu Machine Learning, and Tencent have all leveraged RDMA to meet the strict requirements for network latency, throughput, and CPU computing energy performance of online services, large-scale data centers, and cloud computing. In some time periods, it formed a situation of "Everything is over RDMA."

The early implementation of offloading generally used FPGA or dedicated network cards (also called smart network cards). By accelerating some common operations, it releases the computing pressure of the server and improves the performance-price ratio of the overall solution. The extensiveness of offloading requirements and the diversity of acceleration functions have created a living space for the proposal and development of data processing units (DPUs).

DPU is considered to be the third largest processing unit after CPU and GPU. The CPU performs general computing, the GPU achieves accelerated computing, and the DPU is responsive to the data movement and data processing within the data center. DPU generally contains three elements. The first element is the high-performance network interface, which can realize syntax analysis and data transmission at wire speed. The second one is a high-performance, multicore, programmable CPU with industry standards, which is precisely coupled with other components. The last element is the flexible and programmable acceleration engine that can accelerate machine learning, security, communications, storage, and so on.

The DPU works well for the following situations: (1) syntax analysis of data packets, which is beneficial to the implementation of Open VSwitch (OVS); (2) TCP acceleration, including receive side scaling (RSS), large receive offload (LRO), and checksum; (3) RDMA data transmission acceleration; (4) GPU direct accelerator, which directly provides network data for the GPU; (5) VXLAN network virtualization and VTEP offloading; (6) a traffic shaping accelerator, which enables multimedia streaming, content distribution, and transmission of 4K and 8K video over IP networks; (7) single root I/O virtualization (SR-IOV) and VirtIO; (8) online acceleration of IPSEC and TLS, which can be used for security

isolation, trust root, security root, secure firmware upgrade, and authorized container.

In recent years, Smart NIC has further enhanced the offload processing of network security and storage-related loads by integrating DPU components (https://www.mellanox.com/products/BlueField-SmartNIC-Ethernet). In the field of HPC, Ohio State University's Network Based Computing Laboratory (NBCL) designed a DPU-accelerated BluesMPI, which offloads the nonblocking converged communication MPI_Alltoall to the ARM core of the BlueField™ smart network card (Bayatpour et al., 2021). Recently, the NBCL has used DPU-integrated ARM cores for the first time to accelerate different stages of deep learning training, including training data loading, data augmentation, and training model validation. These offloading operations can reduce the overall deep learning training time by 15% (Jain et al., 2021).

The DPU can be an independent component. It is expected to become a key component of the next-generation server.

4.4 Cross-layer optimization

The development trend of multinetwork integration in high-performance networks reflects the following three aspects. First, with the rise of lossless Ethernet in recent years, RoCE technology has continuously demonstrated "affinity" in supporting HPC communication. For example, Amazon's Cloud HPC adopts scalable reliable datagram protocol (SRDP) technology similar to RoCE to support HPC applications with millions of cores. Second, the HPC network supports the Ethernet protocol stack and its upper-layer applications through network interface virtualization technology. For example, while supporting HPC programming models such as MPI, SHMEM, and PGAS, the open fabrics network standard interface uses the address active registration/address unicast query mechanism to realize the address resolution protocol (ARP) function. In addition, it implements the virtual layered network communication mechanism IP over Express (IPoE) based on the kernel encapsulation interface, thus providing support for TCP/IP high-bandwidth communication. Third, the network hardware infrastructure directly provides multimode configurable functions and supports the mutual conversion and interoperability between multiple network protocols, to realize the interconnection and interoperability of different types of networks and provide communication support for different types of applications. At present, network products with converged features have appeared, such as the ConnectX-6 VPI developed by Mellanox (de Sensi et al., 2020), which realizes 200 Gbps IB and Ethernet converged interconnect chips. It has high performance, such as low latency and high bandwidth, which can greatly improve the performance of the HPC system and data center. ConnectX-6 VPI also integrates network virtualization offload technology to create efficient hyperscale cloud and SDN/NFV data centers.

In addition, the Slingshot Switch produced by Cray has strong Ethernet compatibility and availability. It can support supercomputing and data centers at the same time, allowing the Cray system to build a large-scale interconnected network on 250 000 computing terminals with a diameter of three network hops. Slingshot Switch can also connect directly to third-party Ethernet storage devices and Ethernet networks. In summary, with the continuous development of multi-network fusion technology, the boundaries between HPC environments and data centers are increasingly blurred. Using the same set of infrastructure to support HPC, big data processing, and AI computing will be an important trend in the development of HPC intranets.

Note that Cray's Slingshot E-class HPC interconnection architecture has completed the interoperability test with the Mellanox Ethernet interface card ConnectX-5 and implemented the data center RoCE protocol on the HPC interconnection network (de Sensi et al., 2020). HPC networks and data center network facilities are gradually moving towards integration.

5 Summary

Driven by the development of the Internet, supercomputing, cloud computing, big data, AI, and AR/VR of the Metaverse, increasing demands on network technology have been proposed. From the perspective of the overall development of network technology, this paper summarizes three development trends of network technology, i.e., integration, differentiation, and optimization. It aims to provide useful guidance

for related system designers and key technical points for researchers.

Contributors

Jinshu SU initiated the work. Jinshu SU, Baokang ZHAO, Yi DAI, Jijun CAO, Ziling WEI, Na ZHAO, Congxi SONG, Yujing LIU, and Yusheng XIA drafted the paper. Jinshu SU, Ziling WEI, and Congxi SONG revised and finalized the paper.

Compliance with ethics guidelines

Jinshu SU, Baokang ZHAO, Yi DAI, Jijun CAO, Ziling WEI, Na ZHAO, Congxi SONG, Yujing LIU, and Yusheng XIA declare that they have no conflict of interest.

References

Ajima Y, Inoue T, Hiramoto S, et al., 2014. Tofu Interconnect 2: system-on-chip integration of high-performance interconnect. Proc 29th Int Conf on Supercomputing, p.498-507. https://doi.org/10.1007/978-3-319-07518-1_35

Bayatpour M, Sarkauskas N, Subramoni H, et al., 2021. BluesMPI: efficient MPI non-blocking alltoall offloading designs on modern BlueField smart NICs. Proc 36th Int Conf on High Performance Computing, p.18-37. https://doi.org/10.1007/978-3-030-78713-4_2

Bishop M, 2021. Hypertext Transfer Protocol Version 3 (HTTP/3). Internet-Draft draft-ietf-quic-http-34. Internet Engineering Task Force.

Boden NJ, Cohen D, Felderman RE, et al., 1995. Myrinet: a gigabit-per-second local area network. *IEEE Micro*, 15(1): 29-36. https://doi.org/10.1109/40.342015

Coteus P, Bickford HR, Cipolla TM, et al., 2005. Packaging the Blue Gene/L supercomputer. *IBM J Res Dev*, 49(2-3): 213-248. https://doi.org/10.1147/rd.492.0213

Dai Y, Lu K, Xiao LQ, et al., 2019. A cost-efficient router architecture for HPC inter-connection networks: design and implementation. *IEEE Trans Parall Distrib Syst*, 30(4): 738-753. https://doi.org/10.1109/TPDS.2018.2873337

Dang HT, Canini M, Pedone F, et al., 2016. Paxos made switch-y. *ACM SIGCOMM Comput Commun Rev*, 46(2): 18-24. https://doi.org/10.1145/2935634.2935638

Dang HT, Bressana P, Wang H, et al., 2020. P4xos: consensus as a network service. *IEEE/ACM Trans Netw*, 28(4):1726-1738. https://doi.org/10.1109/TNET.2020.2992106

de Coninck Q, Bonaventure O, 2017. Multipath QUIC: design and evaluation. Proc 13th Int Conf on Emerging Networking Experiments and Technologies, p.160-166. https://doi.org/10.1145/3143361.3143370

de Coninck Q, Bonaventure O, 2021. Multiflow QUIC: a generic multipath transport protocol. *IEEE Commun Mag*, 59(5):108-113. https://doi.org/10.1109/MCOM.001.2000892

Derradji S, Palfer-Sollier T, Panziera JP, et al., 2015. The BXI interconnect architecture. Proc IEEE 23rd Annual Symp on High-Performance Interconnects, p.18-25. https://doi.org/10.1109/HOTI.2015.15

de Sensi D, di Girolamo S, McMahon KH, et al., 2020. An in-depth analysis of the slingshot interconnect. Proc Int Conf for High Performance Computing, Networking, Storage and Analysis, p.1-14. https://doi.org/10.1109/SC41405.2020.00039

Ferlin S, Alay Ö, Mehani O, et al., 2016. BLEST: blocking estimation-based MPTCP scheduler for heterogeneous networks. Proc IFIP Networking Conf and Workshops, p.431-439. https://doi.org/10.1109/IFIPNetworking.2016.7497206

Ford A, Raiciu C, Handley M, et al., 2020. TCP Extensions for Multipath Operation with Multiple Addresses. RFC 8684. Internet Engineering Task Force.

Gibson D, Hariharan H, Lance E, et al., 2022. Aquila: a unified, low-latency fabric for datacenter networks. Proc 19th USENIX Symp on Networked Systems Design and Implementation, p.1249-1266.

Guo CX, Wu HT, Deng Z, et al., 2016. RDMA over commodity Ethernet at scale. Proc ACM SIGCOMM Conf, p.202-215. https://doi.org/10.1145/2934872.2934908

InfiniBand Trade Association, 2010. Supplement to InfiniBand Architecture Specification Volume 1 Release 1.2.2 annex A16: RDMA over Converged Ethernet (RoCE).

InfiniBand Trade Association, 2014. Supplement to InfiniBand Architecture Specification Volume 1 Release 1.2.2 annex A17: RoCEv2 (IP Routable RoCE).

Iyengar J, Thomson M, 2021. QUIC: a UDP-Based Multiplexed and Secure Transport. RFC 9000. Internet Engineering Task Force.

Jain A, Alnaasan N, Shafi A, et al., 2021. Accelerating CPU-based distributed DNN training on modern HPC clusters using BlueField-2 DPUs. Proc IEEE Symp on High-Performance Interconnects, p.17-24. https://doi.org/10.1109/HOTI52880.2021.00017

Ji XS, Wu JX, Jin L, et al., 2022. Discussion on a new paradigm of endogenous security towards 6G networks. *Front Inform Technol Electron Eng*, 23(10):1421-1450. https://doi.org/10.1631/FITEE.2200060

Jin X, Li XZ, Zhang HY, et al., 2017. NetCache: balancing key-value stores with fast in-network caching. Proc 26th Symp on Operating Systems Principles, p.121-136. https://doi.org/10.1145/3132747.3132764

Jonglez B, Heusse M, Gaujal B, et al., 2020. SRPT-ECF: challenging Round-Robin for stream-aware multipath scheduling. Proc IFIP Networking Conf, p.719-724.

Kim J, Dally WJ, Towles B, et al., 2005. Microarchitecture of a high radix router. Proc 32nd Int Symp on Computer Architecture, p.420-431. https://doi.org/10.1109/ISCA.2005.35

Langley A, Riddoch A, Wilk A, et al., 2017. The QUIC transport protocol: design and Internet-scale deployment. Proc Conf of the ACM Special Interest Group on Data Communication, p.183-196. https://doi.org/10.1145/3098822.3098842

Li BJ, Ruan ZY, Xiao WC, et al., 2017. KV-Direct: high-performance in-memory key-value store with programmable NIC. Proc 26th Symp on Operating Systems Principles,

p.137-152. https://doi.org/10.1145/3132747.3132756

Li YJ, Liu IJ, Yuan YF, et al., 2019. Accelerating distributed reinforcement learning with in-switch computing. Proc ACM/IEEE 46th Annual Int Symp on Computer Architecture, p.279-291.

Liao XK, Pang ZB, Wang KF, et al., 2015. High performance interconnect network for Tianhe system. *J Comput Sci Technol*, 30(2):259-272. https://doi.org/10.1007/s11390-015-1520-7

Lim YS, Nahum EM, Towsley D, et al., 2017. ECF: an MPTCP path scheduler to manage heterogeneous paths. Proc 13th Int Conf on Emerging Networking Experiments and Technologies, p.147-159. https://doi.org/10.1145/3143361.3143376

Liu Y, Ma Y, Huitema C, et al., 2020. Multipath Extension for QUIC. Internet-Draft: draft-liu-multipath-quic-04. Internet Engineering Task Force.

Liu Y, Ma Y, de Coninck Q, et al., 2022. Multipath Extension for QUIC. Internet-Draft: draft-ietf-quic-multipath-01. Internet Engineering Task Force.

Petrini F, Feng WC, Hoisie A, et al., 2002. The Quadrics network: high-performance clustering technology. *IEEE Micro*, 22(1):46-57. https://doi.org/10.1109/40.988689

Shi X, Wang L, Zhang F, et al., 2020. PStream: priority-based stream scheduling for heterogeneous paths in multipath-QUIC. Proc 29th Int Conf on Computer Communications and Networks, p.1-8. https://doi.org/10.1109/ICCCN49398.2020.9209682

Song QC, 2019. Mellanox In-Network Computing for AI and the Development with NVIDIA (SHARP-NCCL). Mellanox.

Wang XF, Shi XQ, Su JS, 2008. A TOE-based approach to zero-copy data transmission. *Comput Eng Sci*, 30(2):135-138 (in Chinese).

Wu JX, 2022. Revolution of the development paradigm of network technology system—network of networks. *Telecommun Sci*, 38(6):3-12 (in Chinese). https://doi.org/10.11959/j.issn.1000-0801.2022140

Zheng ZL, Ma YF, Liu YM, et al., 2021. XLINK: QoE-driven multi-path QUIC transport in large-scale video services. Proc ACM SIGCOMM Conf, p.418-432. https://doi.org/10.1145/3452296.3472893

Zhu YB, Eran H, Firestone D, et al., 2015. Congestion control for large-scale RDMA deployments. *ACM SIGCOMM Comput Commun Rev*, 45(4):523-536. https://doi.org/10.1145/2829988.2787484