Special Section 1

Latest Trends for Data Center Networks

Ryo TAKAHASHI

Abstract

In order to realize a cyber-physical system, wherein the technology serves as the core of Society 5.0 [a vision about information-intensive society of the near future formulated by the Japanese government], it is necessary to dramatically improve the information processing capability of edge/cloud computing. For that, a transition to AI and high-performance computing (HPC) based systems is taking place at data centers as they begin to enter a period of major transformation greater than anything that was experienced before. This paper provides a bird's-eye view of the recent trends of research and development activities from the perspectives of data center networks, optical interconnection technology, and computing architecture.

Keywords : data center, edge computing, optical interconnection, computing architecture

1. Introduction

A cyber-physical system (CPS) that highly integrates cyberspace and physical space is essential for the realization of Society 5.0 (super smart society)[(1)] advocated by the Cabinet Office of Japan. (See Fig. 1). The realization of CPS requires, as something essential to it, a major improvement in the capability to process information in cyberspace along with the development of IoT technology in physical space and the development of massive data transfer technology (wireless and optical communication technology) for the connection between physical space and cyberspace. It has been customary until today to send data from physical space to large suburban data centers (DCs) for processing, but then the problem of propagation delay, which becomes greater with distance, produces difficulty in delivering real-time services like the support of the autonomous driving of automobiles.

Ryo TAKAHASHI Senior member (Network Research Institute, National Institute of Information and Communications Technology, Koganei-shi 184-8795 Japan)
E-mail : t.ryo@nict.go.jp
THE JOURNAL OF THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS Vol.106 No.2 pp.(1)-(10) February 2023

Considering that, progress has been made in recent years in the building of edge computing architecture by the distributed deployment of smaller DCs (termed μ-DCs, edge DCs, small DCs, etc.) in the periphery of a metro network.

Such DCs, playing a crucial role in cyberspace, have continued to evolve thanks to the advancement of such core technologies as conventional optical transceivers, switches, and servers. However, in recent years, explosive increases in traffic experienced by DCs began to produce growth gaps, namely, delays in the development of such core technologies, causing stress in different domains. For example, in the domain of networks, an increase in traffic experienced by DCs led to an increase in switch capacity, which in turn led to an increase in inter-switch link capacity, producing a situation in which the optical transceiver performance becomes binding upon the network performance. Likewise, in the domain of servers, an increase in AI parameters led to an improvement in the performance of processors and similar devices, which in turn led to an increase in the memory bandwidth requirement (memory wall problem), producing a situation in which the data transfer capacity is severely binding upon the improvement in information processing capabilities.

In recent years, the attempts to solve such problems

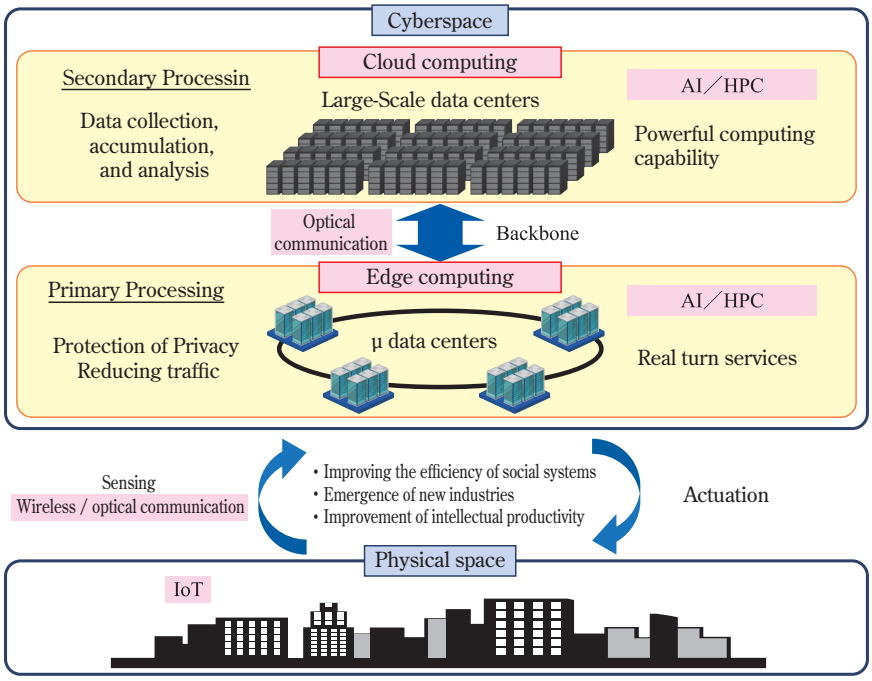


Fig. 1 Cyber-physical system

have begun to bring about drastic transformations in the various kinds of DCs. In the domain of networks, transformation is taking place around the network architecture and optical interconnection technology. In the domain of servers, a major transformation in the computing architecture is being conceived for the building of advanced systems centered around AI and high-performance computing (HPC). From the perspectives of networks and interconnections (hardware), this paper provides an overview of the recent trends in research and development addressing the imminent need to transform DCs in the various ways mentioned above.

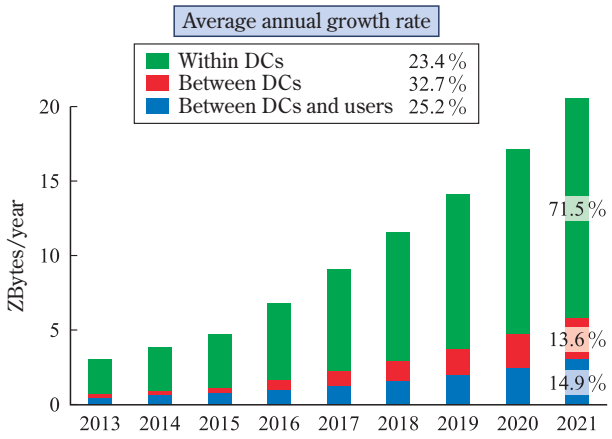


Fig. 2 Evolution of DC network traffic

2. Data Center Networks

DCs used to be located in such places as server rooms and large computer rooms of corporations until many dedicated operators began to run DCs as various internet services began to be commercialized from the beginning of the 1990s. In the 2000s, server virtualization technology (enabling a single server to run multiple virtual machines) and virtual machine migration technology (enabling a virtual machine to be moved to any server) appeared, which enabled the maximum use of the performance of multicore servers, and then the traffic through DC networks increased significantly. After that, in the 2010s, the advances in network virtualization and DC virtualization, combined with the coordinated operation of multiple DCs, resulted in the full-scale launch of cloud computing.

Fig. 2 reports the evolution of worldwide DC traffic [2]. Although the overall annual growth rate is 25%, DCs that are ahead of others in the deployment of AI (like Google DCs [3]) are experiencing an explosive growth in traffic at the annual rate of 70% or more. Another important point to note about DCs is that traffic of enormous volume, five times greater than traffic that takes place between DCs and users (north-south traffic), takes place within DCs between servers (east-west traffic). Because of the evolution of cloud

computing, inter-DC traffic is also increasing quickly.

2.1 Increase of Network Bandwidth

The history of the expansion of DC networks has been a struggle with the constant growth in service and traffic, and the networks continued to evolve supported by a variety of different technologies on the basis of the following principles :

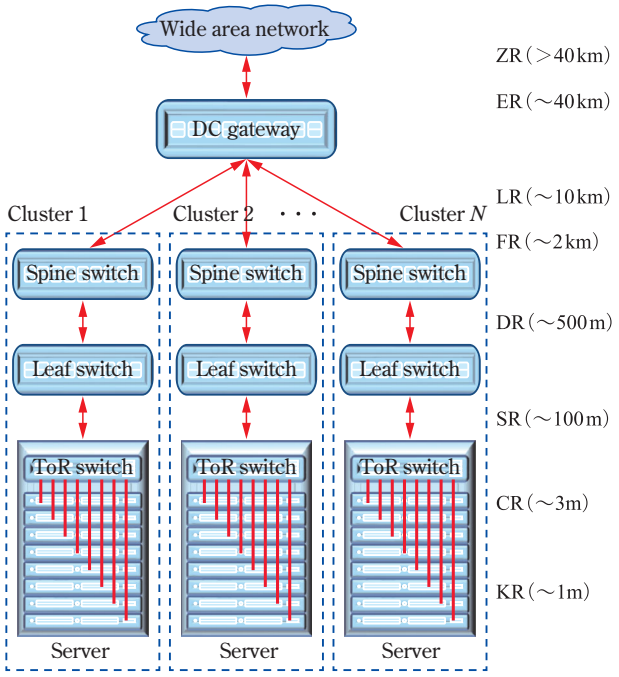
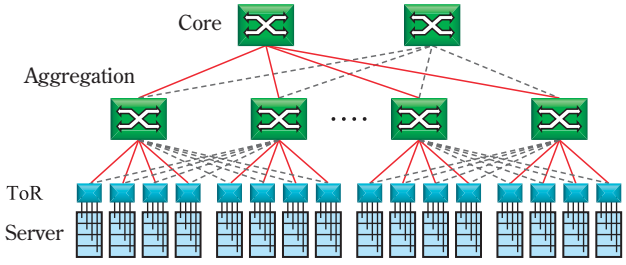


Fig. 3 DC network overview

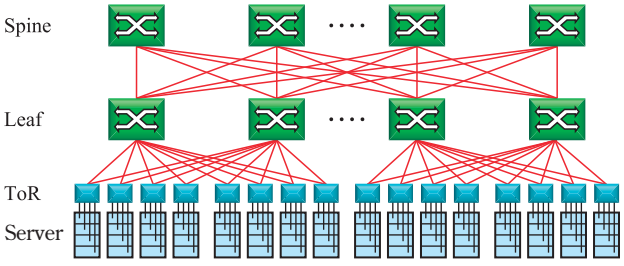
- Scale-up : increasing the link capacity of the networks
- Scale-out : increasing the number of servers and expanding the networks

As a result, massive DCs, with the number of servers in the order of one million, have been formed in recent years ; however, as shown in Fig. 3, servers are grouped into clusters of several tens of thousands of units to several hundreds of thousands of units, and they are connected to other DCs and to users via DC gateways and wide area networks (WANs). What you read on the right side of the diagram are the approximate distances between units and the name of the transceiver standard applicable to those distances (ranges). For use as SR/DR optical transceivers that are mainly used within DC networks (within clusters), optical transceivers of up to 800 Gbit/s have become available in recent years, and efforts are being made toward the development of units that would achieve 1.6 Tbit/s.

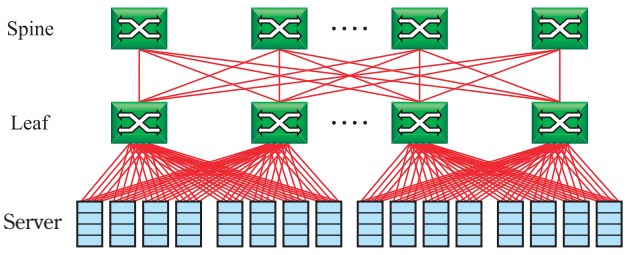
Fig. 4 shows examples of a network configuration within a cluster. Fig. 4 (a) shows a configuration that was used in the early period. It adopts a hierarchical fat-tree topology comprising top-of-rack (ToR) switches, aggregation switches, and core switches. It makes use of the Spanning Tree Protocol (STP), which is a layer 2 control protocol, and the links represented by dotted lines are blocked in order to prevent the interminable flow of data in the network loop. This network



(a) In the past: L2-based fat-tree network



(b) Today: L3-based Clos network



(c) Future: ToR-eliminated Clos network

Fig. 4 Evolution of DC networks

configuration causes the concentration of data traffic around the core switches, and therefore requires the use of switches of very large capacity. In the early period, when north-south traffic was dominant, traffic could be managed with this network configuration; however, later on as the east-west traffic increased with progress in server virtualization as mentioned earlier, this network configuration was found to be poor in scalability because it could not provide a sufficiently large network bandwidth. Therefore, from around the middle of the 2000s, there took place a transition to the network configuration that is widely used today, namely, the Clos network configuration shown in Fig. 4(b)[4]. (It is based on the interconnection of switches in three hierarchical layers, and it is also known as the folded Clos network configuration because of folding at the spine switches.) In this configuration, data transfer paths are greatly expanded with multiple redundant links interconnecting a large number of ToR, leaf, and spine switches. Moreover, the use of the layer 3 routing protocol (the Border Gateway Protocol [BGP] in most cases) made loop prevention unnecessary, allowing the active use of all links. As a result, this network configuration allows significant increases in the network bandwidth to cope with explosive increases in the east-west traffic.

2.2 Reduction of Power Consumption and Delay

Power consumption at DCs is expected to increase to around 8-10% of global power consumption in 2030, making the reducing of power consumption an extremely important target. As another issue, with increases in AI/HPC services that require advanced information processing capabilities, data transfer delays in networks have become a major problem. For the solution of these issues around power consumption and delay, several measures are being considered:

-Transition to next-generation optical transceivers

As the speed of optical links increases, the continued use of conventional pluggable optical transceivers becomes unadvisable due to complexities regarding signal compensation and error correction, which are likely to increase power consumption and delays. Toward the solution of this problem, the study of next-generation optical transceivers for placement near a switch ASIC (application specific integrated circuit) is conducted actively. This is discussed more in Section 3. (A switch ASIC is a large-scale integrated circuit specializing in the function of switch. See Fig. 6.)

-Transition from the store and forward method to the cut-through method

With the conventional store and forward method, the switch ASIC takes the entire input packet into memory before reading out the address from the header and forwarding the packet toward the destination. With the cut-through method, the switch ASIC initiates the address readout and forwarding procedure only after taking the header into memory, significantly reducing the transfer delay caused by the switch. For this, the input and output packets must be of the same format (without the modification of speed or the like), and this method is introduced to spine switches[4].

-Reducing the number of network layers; flattening

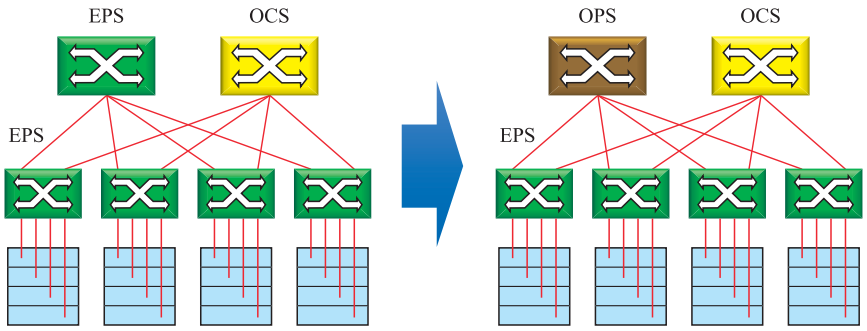
Currently, signals from servers are sent to ToR switches through electrical wirings, and then sent to leaf switches in the form of high-speed optical signals. There is an idea to do away with ToR switches in the future by having servers directly send out high-speed optical signals as shown in Fig. 4(c), reducing the number of network layers[5].

-Super high-radix switch utilizing a single switch ASIC

Within a single leaf or spine switch, a multiple port configuration (128 ports, for example) is realized by forming a Clos configuration with many switch ASICs[6]. As a result, a data transfer between servers requires passing through up to nine switch ASICs. However, in the future, when it becomes possible to realize a 400 Gbit/s×128 port switching device with a single large-capacity switch ASIC, delivering the capacity of 51.2 Tbit/s, for example, the number of ASICs may be reduced, and then the data transfer between servers will not require passing through of more than five ASICs. Yet, it is difficult to realize this with currently used pluggable optical transceivers, so the development of the earlier-mentioned next-generation ultra-compact optical transceivers is required[5].

-Transition to optical switch networks[7]

The ultimate approach to reducing power consumption and transfer delays is to replace electrical switches with optical switches by the deployment of optical switch networks. Fig. 5(a) shows, as the first step toward realizing such switch networks, the concept of a hybrid network composed of optical circuit switches (OCSs) and conventional electrical packet switches (EPSs).



(a) Hybrid network combining electrical packet switches (EPSs) and optical circuit switches (OCSs)

(b) Hybrid network combining optical packet switches (OPSs), EPSs, and OCSs

Fig. 5 Optical switch networks of the future

OCS is a circuit switching device with extremely slow switching speed in the order of milliseconds and does not have the feature of collision prevention. Therefore, it is not used for packet-by-packet switching; instead, it is used for the offloading of an extremely large mass of data like backup data. As a further step, the deployment of optical packet switches (OPSs), as shown in Fig. 5(b), is expected. An OPS, like an EPS, has the three elemental functions of header recognition, switching, and buffering and is capable of collision prevention, as well as packet-by-packet switching. However, optically realizing these functions is extremely difficult. The technical barrier is extremely high particularly for optical switches, which must play the crucial role, as they must satisfy requirements such as high speed, low loss, and polarization independence. Considering that, for the deployment of OPSs, the use of a network of a new topology (like torus and DCell) suitable for optical switches with a relatively small number of ports (something like 16×16) is proposed. For more about optical switch systems, the reader may be interested to read 4 "Trends for Optical Switch Networks" as another article on the latest trends for data center networks featured in this volume of the *Journal of IEICE*.

3. Optical Interconnections

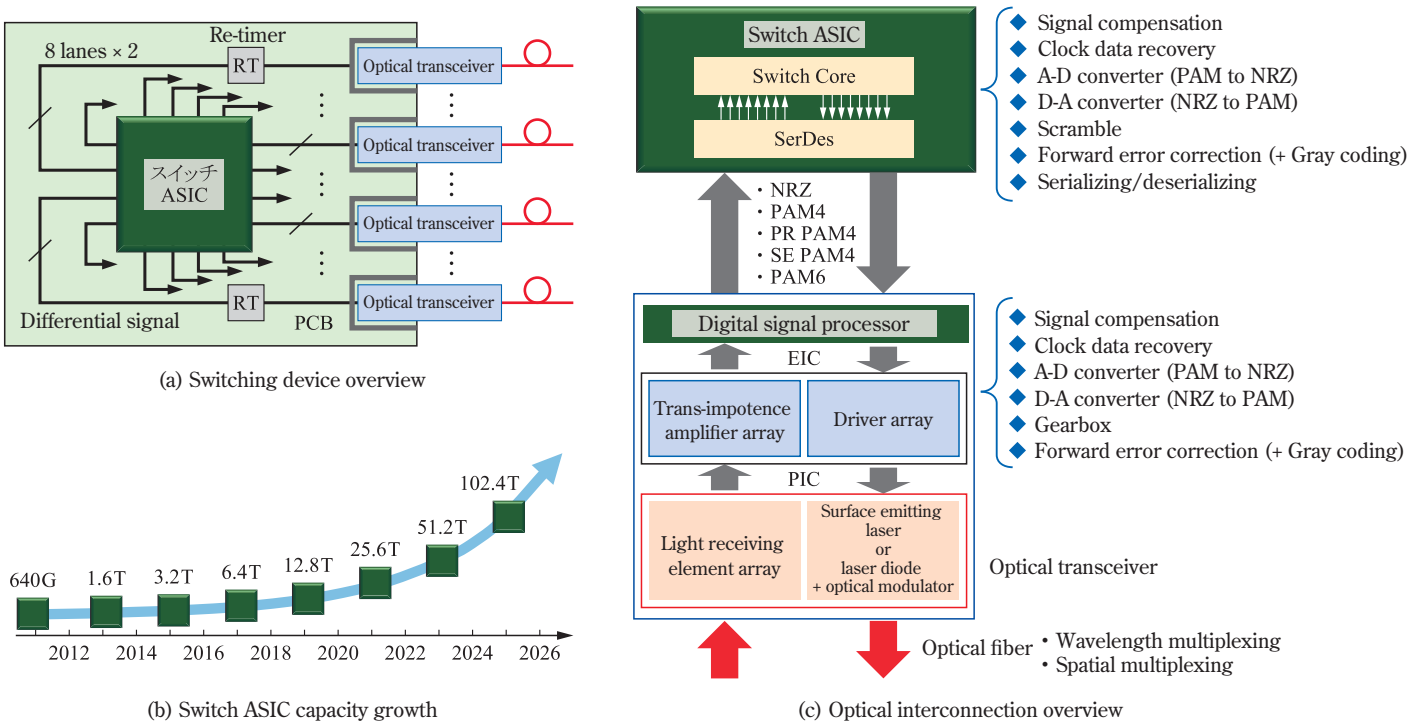
A currently used switching device is basically configured by placing a switch ASIC on the printed circuit board (PCB) and connecting it to each of the optical transceivers mounted onto the front panel using electric wiring of several tens of centimeters. (See Fig. 6 (a)) By interconnecting multiple pieces of such a device, a single leaf or spine switch is configured. The switch ASIC capacity is growing at the rate of becoming double every two years as shown in Fig. 6 (b). At present, high-capacity switch ASICs delivering the capacity of 25 Tbit/s are being developed. However, with such growth in the capacity of switch ASICs, switching devices are being faced with various problems described below:

-Restriction to the number of optical transceivers that may be mounted

The form factor (standard on physical dimensions) for optical transceivers evolved from CFP to QSFP toward the direction of becoming smaller, but after that, further downsizing was found to be difficult. Therefore, the maximum number of optical transceivers that may be mounted onto the front panel is limited to 32 pieces per 1 U.

-Issues around high-speed, highly dense electrical wiring

Since the number of input/output pins for connection between the PCB and the switch ASIC is restricted[8], each of the 32 pieces of the optical transceiver is connected to the switch ASIC by a differential signal path of eight lanes for input and output, respectively (through 32 pieces of electrical wiring in all). Currently, 400 Gbit/s optical transceivers make use of a pulse amplitude modulated PAM4 signal of 50 Gbit/s, but as the electrical signal becomes faster, the degradation of transfer characteristics (degradation of signal waveform) due to the dielectric loss of high-frequency components, crosstalk, and other causes becomes a major issue. In recent years, thanks to the development of flyover cables (fine coaxial cables) and Megtron-type low-loss boards enabled the realization of 800 Gbit/s optical transceivers, but many issues have yet to be



(a) Switching device overview

(b) Switch ASIC capacity growth

(c) Optical interconnection overview

Fig. 6 Switching device

overcome before the speed may be increased to 1.6 Tbit/s.

-Issues around signal processing and signal compensation

Since the signal waveform degrades significantly during transfer between the switch ASIC and the optical transceivers, signal processing and signal condensation of various kinds, as shown in Fig. 6 (c), becomes necessary. The I/O interface of the switch ASIC is equipped with a serializer/deserializer (SerDes) for the transformation of high-speed signals into internal clock regulated parallel low-speed signals. Besides being able to provide functions that are normally expected from SerDes, it also includes circuits for signal compensation (equalization), waveform shaping (CDR : clock and data recovery), and error correction (FEC : forward error correction). In addition, for the processing of the four-level PAM4 signal, analog-to-digital (A-D) and digital-to-analog (A-D) converters are required. The optical transceivers, on the other hand, are equipped with a digital signal processor (DSP) chip that similarly provides a variety of functions. Currently, as an increase in speed to 200 Gbit/s per lane is discussed [9], solutions, such as the inclusion of FEC into DSP, which used to be unnecessary in the past, and the use of a new modulation method (such as PR-PAM4 and PAM6/8), are being proposed. Thus, the continued use of conventional pluggable optical transceivers is coming close to the limit because, as electrical signals become faster, they increase power consumption and delays significantly due to complications regarding SerDes and DSP.

To solve this problem, it is important to make the distance between the ASIC and the optical transceivers as short as possible, minimizing the degradation of the electrical signal during transfer in these sections. For that, different mounting methods like the ones shown in Fig. 7 are proposed [5), (10), (11)], and in recent years, intense competitions are taking place worldwide particularly for the development of Near Package Optics (NPO) and Co-Packaged Optics (CPO) for compliant next-generation optical transceivers. XSR, VSR, and other acronyms in the figure stand for the names of Common Electrical I/O (CEI) Standards [12] for different distances (different amounts of transfer loss), and these standards specify functional requirements for SerDes and DSP. Since NPO/CPO devices do not make use of electrical wiring on the PCB, they have an increased number of parallel I/O wirings and allow the handling of NRZ signals. As a result, the demand for signal compensation by DSP becomes very light, and all other functions may be

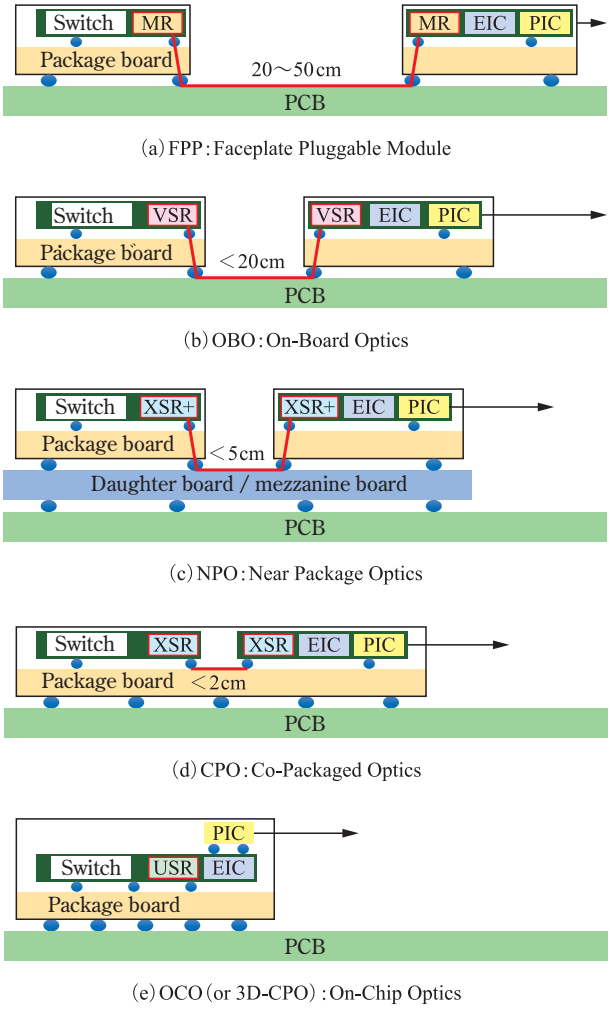


Fig. 7 Different approaches to optical transceiver mounting

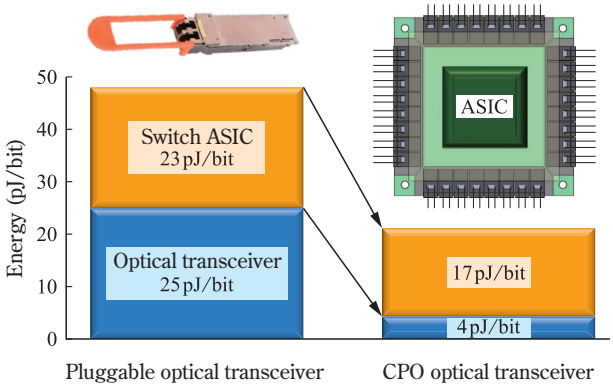


Fig. 8 Contribution to reduction in power consumption

removed, significantly reducing power consumption. Furthermore, by integrating the signal compensation and other similar circuits on the electrical integrated circuit (EIC) side, the DSP chip can be done away with, enabling major downsizing. Since this will reduce the functional demand on SerDes as well, the power consumed by the switch ASIC is also reduced. As a result, as shown in Fig. 8, energy spent for a single transfer via a switch will become less than half. Assuming the use of a 400G optical transceiver, the power consumed by SerDes is predicted to drop from 8 pJ/bit to 2 pJ/bit. For more about optical transceivers, the reader may be interested to read 2 "Data Center Optical Interconnection Technology" as another article on the latest trends for data center networks featured in this volume of the *Journal of IEICE*.

4. Computing Architecture

DCs of today that interconnect a large number of servers in a network configuration are approaching limit in allowing the improvement in the capability of computing systems centered around AI/HPC. Therefore, movements toward innovative transformation are taking place not only in the domain of networks but also in the domain of servers :

-Networks optimized for AI and machine learning

In recent years, the number of AI parameters has increased explosively, and AI processing requires the use of 100 or more accelerators. If such processing is executed on a conventional server system, the network transfer delay becomes a major problem. Fig. 9 (a) shows the concept of TPUv4 published by Google in 2021 [13]. Each tensor processing unit (TPU) board has four TPU chips mounted on it. Each TPU board is connected to a TPU host equipped with a CPU for controlling the TPU board. Each TPU board connects to ToR switches via the TPU host, and connects to external storage and servers via the DC network. Apart from the connection via the DC network, TPU boards are directly interconnected by optical links that form a three-dimensional torus network. (In the case of the earlier TPUv3, a two-dimensional network was formed instead.) The entire system has a TPU chip configuration of 4,096 chips, which enable the parallel processing of many instances of AI machine learning by partially forming neural networks.

-Resource-disaggregated computing

A conventional system, shown in Fig. 9 (b) (left), is composed of a large number (20 to 40) of rack-mounted

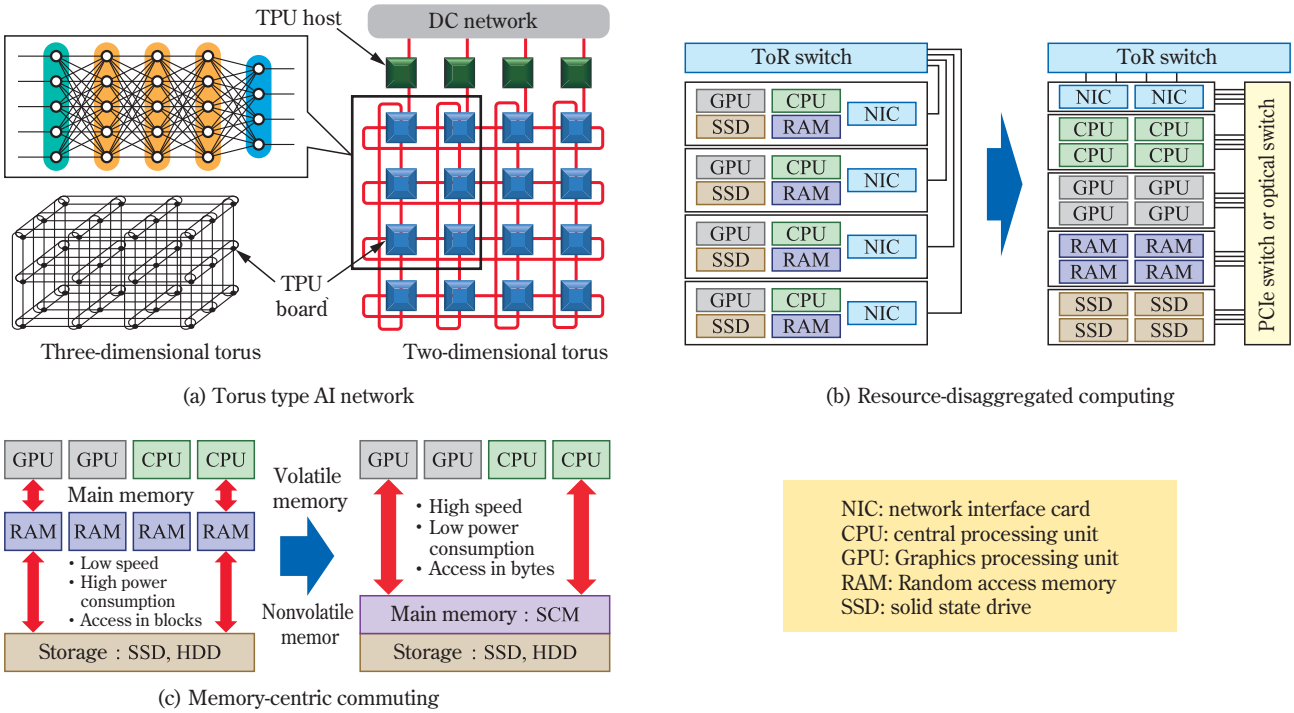


Fig. 9 New computing architectures

box-type servers, each bundled up into a set with resources, such as processor, accelerator, and memory, storage, and each such server connects to the DC network using a network interface card (NIC) to accomplish protocol conversion into Ether packets. Since there will be many resources that fall short of being fully used, the resource utilization rate is low, which results in poor energy efficiency.

Fig. 9 (b) (right) shows the concept of resource-disaggregated computing[14]. Resources are disaggregated and clustered into pools according to type, and they are connected via a PCIe switch (or an optical switch). Although this concept had been proposed since around the second half of the 2000s, the recent advances in AI/HPC and in direct memory access (DMA, direct access to memory made by other resources without intervention by CPU) accelerated the process of transition into such resource-disaggregated configuration. Such resource-disaggregated configuration has the following advantages:

- Enables application-by-application assignment of necessary resources
- Significantly improves the resource utilization ratio and energy efficiency
- Allows the scaling up, scaling down, and upgrading of resources by type

On the other hand, such resource-disaggregated configuration may have the following disadvantages:

- Requires enormously large number of links and an enormously broad bandwidth
- Increases delays in the transfer of data across resources of different types (particularly with memory)

As a solution to the above issues, expectations are placed on the development of large-capacity, low-loss, inter-chip optical interconnection technology using the earlier-mentioned CPO optical transceivers. Although the figure shows an example of rack-scale disaggregation, development into DC-scale disaggregation is expected in the future.

-Memory-centric computing

With the systems of today, data is copied from its storage (nonvolatile memory) into CPU/GPU-connected small-sized volatile main memory (DRAM) for processing, and immediately after the completion of processing, data has to be reentered (rewritten) into the storage. (See Fig. 9 (c) (left)) In this operation, a

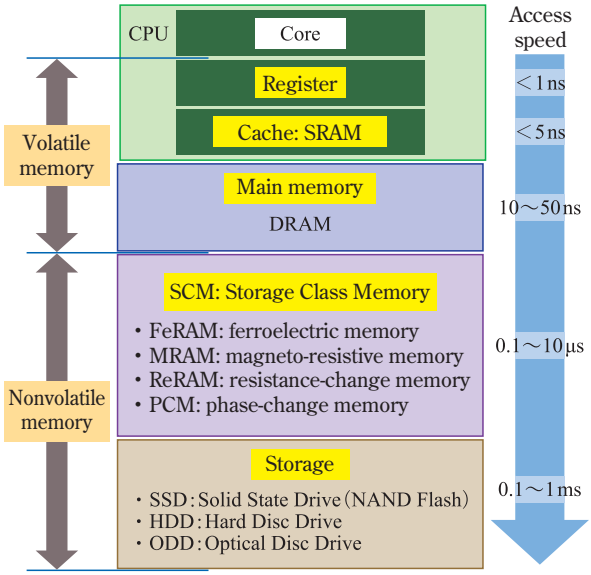


Fig. 10　New memory configurations

certain amount of time is spent in idleness due to a major difference in the access speed between the DRAM and the storage. Moreover, the energy consumed by the data transfer is greater than the energy consumed by the single instance of the computation process that takes place. Furthermore, since the transfer of data from the storage has to be in units of blocks, the main memory has to receive superfluous data as well. So, the issues are many.

The concept of memory-centric (or memory-driven) computing is basically dedicated to the minimizing of the data transfer that takes place frequently in conventional systems, and an example is shown in Fig. 9 (c) (right)[15]. The ultimate goal of such a concept is to place a nonvolatile large-capacity main memory (shared memory) at the center of the system for a direct connection to the processor. This will eliminate the need for transferring data from/to the storage at the time of processing (then the storage will serve as a buffer memory for storing data after processing). However, because of the above-mentioned handicap, conventional nonvolatile memory (SSD and HDD) can hardly be used as the main memory.

As a solution to this problem, expectations are placed on storage class memory (SCM) referred to as next-generation memory[16]. (See Fig. 10) SCM is suitable as the main memory because it is nonvolatile, yet nearly as fast as DRAM and, moreover, allow refined accesses in bytes. However, since some time will be needed before SCM of large capacity becomes available, the combined use of SCM and DRAM is considered in the meantime. For more about new computing architectures, the reader may be interested to read 3 "Computing and Network Technologies in Data Centers" as another article on latest trends for data center networks featured in this volume of the *Journal of IEICE*.

5. Conclusion

This paper provided an overview of the paradigm shifts of the different kinds experienced by data centers. When CPO ultra-compact optical transceivers, addressed by intense research and development in recent years, are realized, there will be a shift from the earlier concept of *inter-device optical connections* to the concept of *inter-chip optical connections*, strongly accelerating the transition to resource-disaggregated and memory-centric computing. Such optical interconnection technologies and computing architectures based on new concepts will greatly help data centers reduce power consumption, increase capacity, improve the resource utilization ratio, increase information processing capability, and save space. Then, edge/cloud computing, standing on the basis of such new technologies, is expected to help the emergence of new services of the kind that never existed before by supporting the evolution of the cyber-physical system and bringing about super smart society. For more about super smart society, the reader may be interested to read 5 "Roles of Edge Computing in B5G/6G and Development into a Super Smart Society" as another article on the latest trends for data center networks featured in this volume of the *Journal of IEICE*.

Reference

(1) 総務省, "Beyond 5G 推進戦略—6G へのロードマップ," 2020.
(2) Cisco Global Cloud Index : Forecast and Methodology, 2016-2021.
(3) X. Zhou, R. Urata, and H. Liu, "Beyond 1Tb/s datacenter interconnect technology : challenges and solutions," OFC2019, Tu2F.5, 2019.
(4) A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, H. Liu, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, and A. Vahdat, "Jupiter rising : A decade of clos topologies and centralized control in Google's datacenter network," Commun. ACM, vol. 59, no. 9, pp. 88-97, 2016.
(5) C. Minkenberg, N. Farrington, A. Zilkie, D. Nelson, C. Lai, D. Brunina, J. Byrd, B. Chowdhuri, N. Kucharewski, K. Muth, A. Nagra, G. Rodriguez, D. Rubi, T. Schrans, P. Srinivasan, Y. Wang, C. Yeh, and A. Rickman, "Reimagining datacenter topologies with integrated silicon photonics," J. Opt. Commun. and Netw., vol. 10, no. 7, pp. 126-139, 2018.
(6) A. Krishnamoorthy, H. Thacker, O. Torudbakken, S. Muller, A. Srinivasan, P. Decker, H. Opheim, J. Cunningham, I. Shubin, X. Zheng, M. Dignum, K. Raj, E. Rongved, and R. Penumatcha, "From

chip to cloud : Optical interconnects in engineered systems," J. Lightwave Technol., vol. 35, no. 15, pp. 3103-3115, 2017.
(7) F. Testa and L. Pavesi, Optical Switching in Next Generation Data Centers, Springer Verlag, New York, 2017.
(8) Y. Mori and K. Sato, "High-port-count optical circuit switches for intra-datacenter networks," J. Opt. Commun. and Netw., vol. 13, no. 8, 2021.
(9) OIF CEI-224G. https://www.oiforum.com/technical-work/hot-topics/common-electrical-i-o-cei-224g/
(10) A. Ghiasi, "Large datacenters interconnect bottlenecks," Opt. Express, vol. 23, no. 3, pp. 2085-2090, 2015.
(11) 高井厚志, "光トランシーバーの Form Factor の新動向(8)〜CPO/NPO と新しいデータセンター" EE Times, Jan. 2022.
(12) OIF CEI-112G. https://www.oiforum.com/technical-work/hot-topics/common-electrical-interface-cei-112g-2/
(13) Google Cloud. https://cloud.google.com/tpu/docs/system-architecture-tpu-vm
(14) K. Bergman, "Deeply disaggregated computing systems with embedded photonics," ARPA-E ENLITENED Program, 2021. https://arpa-e.energy.gov/sites/default/files/2021-01/DAY1_Bergman_ENLITENED_Phase2_Kickoff.pdf
(15) HPE Blog, "Memory-driven computing explained," 2017. https://www.hpe.com/us/en/newsroom/blog-post/2017/05/memory-driven-computing-explained.html
(16) G. Kranz, "Storge class memory (SCM)," TechTarget, Aug. 2021. https://www.techtarget.com/searchstorage/definition/storage-class-memory

(This paper was accepted on August 30, 2022, and finalized on September 16, 2022.)

Ryo TAKAHASHI (senior member)

In 1987, graduated from the Department of Electronics, Faculty of Engineering, University of Tokyo. In 1992, completed the doctoral program at the graduate school of the same university. Joined Nippon Telegraph and Telephone Corporation in the same year. Engages in research on ultra-high-speed optical devices, optical signal processing, optical networks, and in-vehicle optical communication systems. Currently serves as chief senior researcher at the Network Research Institute of the National Institute of Information and Communications Technology, Japan. A holder of doctoral degree in engineering.