Research of Optical Interconnection Technology in Datacenter Networks

Tianchun Hao
Beijing University of Posts and Telecommunications, Beijing, China

chuenhtc@163.com

Abstract. In recent years, with rapid development of cloud computing, the business volume of datacenter has increased unprecedentedly and the development of datacenter is facing great challenges. At this point, more efficient interconnection schemes are needed to maintain the increasing demand for communication bandwidth. With its great bandwidth advantage, optical fiber provides low-loss and low-cost solutions for telecom networks. At the same time, compared with the electric interconnection based on copper wire, optical switching technology greatly improves the transmission rate of the communication network. In this paper, the interconnection structure of the datacenter network is studied. Aiming at the interconnection for different top-of-racks (ToR), the a-Through and Helios structures based on photoelectric hybrid technology are reported. Moreover, some structures based on all-optical interconnection such as Proteus, OSMOSIS and IRIS are mentioned. For the interconnection in ToR, arrayed waveguide grating (AWG) based passive optical interconnection and the passive optical couplers-based interconnection schemes are described. Meantime, the main principles, advantages and disadvantages of various structures are analyzed. Finally, the future development of optical interconnection applications is stated.

1. Introduction
The 21st century is an era of information explosion, and more and more internet-related words become well known to a wider public with increasing frequency, such as big data, cloud computing and so on. Among them, big data is an important information asset, which is playing an important role in much more fields such as economy, education and medical treatment than before. As a computing method to deal with big data in a complicated and cumbersome way, cloud computing takes internet-centered as the core concept. It provides fast and safe cloud computing services and data storage on websites, so that everyone who is on the Internet can use the huge computing resources and datacenter on the network [1]. The rapid development of Internet technology and mobile access network technology has promoted people from a single individual to a node on the Internet. While enjoying the convenience brought by the information age, people also have a strong demand for the improvement of communication technology. For example, high capacity, high transmission speed storage space, low cost and efficient network information resources are expected by the public, while cloud storage services and efficient cloud computing services are the demands of enterprises. Datacenter provides a powerful supporting platform for cloud services, therefore, the core of improving service quality lies in the powerful upgrade of datacenter system.

One of the core values of cloud computing is the centralized processing of big data. With the increasing of datacenter traffic, the network bandwidth and performance are challenged. As the
number, scale and internal structure of traditional datacenter cannot meet the bandwidth requirements
of cloud computing and big data processing, datacenter network has become the bottleneck of modern
cloud computing. Nowadays, fiber optic communication is a mature and developing technology
worldwide. Compared with the original copper-based communication network, ordinary optical fiber
can not only improve the transmission rate, but also have the advantages of large bandwidth, small
size and low energy consumption. Based on the superiority of optical fiber communication for
network information communication, optical switching technology has been introduced into
commercial datacenter networks to replace high-energy electrical switches. Both optoelectronic hybrid
structure and all-optical interconnection structure are introduced. The choice of optical switching or
electrical packet switching is determined by the specific application. In this case, optical-circuit
switches handle large data streams, namely elephant streams, while an electrical packet switch is used
to exchange rat streams, namely, smaller data. In addition, datacenters need stronger internal networks.
Datacenter network structure is mainly based on three-tier tree topology network structure, in which
the layers are the core layer, the aggregation layer and the access layer [2]. In the access layer, also
known as the edge layer, a rack is usually composed of multiple servers, while the interconnection of
the two categories of datacenter interconnection takes place around the racks. On the one hand, the
servers inside the same rack are interconnected through the switch at the top-of-rack (ToR). On the
other hand, a ToR switch is connected to another ToR through switches in aggregation layer and core
layer. In order to provide high-speed forwarding services for packets accessing to the datacenter and
realize the fast data exchange between the access layer and core layer, the researchers suggest
structures such as fat-tree 59 and Quartz 60 [3] [4]. However, this kind of structure merely inserts
redundancy in the aggregation layer and core layer to improve the reliability of network inside the
datacenter, the data edge layer is usually out of effective protection because of the high cost and the
trust of self-healing ability. Reducing the usage of active devices and using passive optical
interconnect technology to make ToR connections at the edge layer of the datacenter will effectively
solve the problems mentioned above.

In this paper, the existing optical interconnection architecture of datacenter is summarized,
including the optical interconnection architectures between racks in datacenter and servers in the same
rack. Meantime, the main principles, advantages of various structures are analyzed. Finally, the future
development of optical interconnection applications is stated.

2. Structures of Datacenter Optical Interconnection

2.1. Optoelectronic Hybrid Interconnection Architecture

2.1.1. C-Through [5] C-Through is an internal datacenter network which is optoelectronic hybrid. As
shown in figure 1, the main part of its structure is dominated by traditional electrical switches. The
difference is ToR switches have communication interfaces for optical switching networks in addition
to the electrical transmission interfaces for aggregation level switches. Among them, the electrical
switch can deal with the problem of packet initiations with strong randomness, while the optical
switch can solve the problems occur when long data streams with high speed and stability need to be
transmitted. The optical exchange in the c-Through structure is realized by the optical switch array,
while only one optical transformation is allowed in ToR at the same time. As the demand changes, the
optical exchange needs to be reconfigured. In order to meet the needs of the large broadband
transmission between two racks, the c-Through structure sets up the flow monitoring of the server.
The propose is for real-time monitoring of the flow demand between different servers and different
racks. To support both electrical and optical switching networks, ToR takes virtual local area network
(Vlan) technology. When the ToR switch recognizes that the switch of packet's receiving and sending
has been connected via an optical link, it allocates the packets to the optically switched Vlan.
2.1.2. Helios [6] Helios is a kind of interconnection structure mainly oriented at the upper layer of the rack in the datacenter. The Helios architecture has many similarities to the C-Through architecture. They provide optical connections to ToR switches on the basis of electrical interconnection. There are two tiers of datacenter network in Helios, including pods and core layer, which is shown in Fig. 2. There are three different connection modes in this structure. Firstly, the servers inside the rack form an electrical connection to the ToR switch. Secondly, ToR switch regards multiple optical transceivers as the data interface for data exchange with the upper communication structure. Finally, ToR is connected to each core switch through a pair of optical transceivers, and the transceiver forms a super optical link with the core layer and the optical link switches through a wavelength division multiplexing (WDM) multiplexer.

2.2. All Optical Interconnection Architecture

2.2.1. Proteus [7] [8] The network structure of Proteus is shown in the Fig. 3. It is an all-optical interconnection structure of datacenter. Similarly, the optical switching matrix is still adopted as the core switching device, and WDM technology is used to realize the optical transmission link with large broadband and the data receiving and sending. The difference is that in Proteus structure, the complex optical switch and wavelength selective switch (WSS) are used for the connection between different wavelengths and the optical switch. In the transmitting link, WSS can group and reuse the multi-wavelength signals sent by ToR and allocate multi-channel optical signals to different switch array interfaces. In the receiving link, the wavelength demultiplexed signal will be accepted by the
corresponding receiver. According to the methods above, Proteus architecture realizes the multi-channel point-to-point connection between ToR.

Figure 3. The structure of Proteus [7]

2.2.2. OSMOSIS [9] [10] OSMOSIS architecture is a low-latency broadcast and selection structure, which is mainly based on WDM and space division multiplexing. In Fig.4, the network structure of OSMOSIS can be divided into broadcast layer and selection layer. The broadcast layer is a substructure for accounting eight ToR switches to access, while the photosynthetic combiner and the optical splitter are responsible for the confluence of optical signals and the space division multiplexing, respectively. After passing through the broadcast stage, each broadcast layer substructure bisected the power of all the multi-wavelength signals of the ToR switch and broadcast them to all the selective layer substructures. Selective layer is composed by two groups of semiconductor optical amplifier (SOA), divider and WDM, so as to realize the function of integrity choice of optical signal, including any broadcast substructure and signals in any wavelength. The whole structure is reconfigured through a programmable central arbitration unit, so the packet switching based on the optical network structure can be realized to some extent.

Figure 4. The structure of OSMOSIS [9]

2.2.3. IRIS [11] IRIS is an optical network structure inside a datacenter that adopts array waveguide grating routing (AWGR) and is realized by WDM technology. As Fig.4 shows, IRIS uses a three-level structure to realize a non-blocking optical switching network, including two space-domain switching parts and a time-domain switching structure. The spatial switching structure provides a long wavelength converter for each ToR switch. Through a series of wavelength converters, the synthesized multi-wavelength signal is finally sent into AWGR to realize the function of wavelength routing. The time domain switching structure adds the function of time caching to each input of the array waveguide grating (AWG). By demultiplexing the WDM signal, the multi-channel single wavelength signal passes through the fiber of different length in time buffer, realizing the signal delay, that is to say, realizing the function of time-domain cache.
2.2.4. Bidirectional photonic network [12] Bidirectional photonic network architecture focuses on the design of small switching structure, so as to realize the switching function of a large number of nodes. The structure is shown in Fig.6. It is composed of sub-switching parts of line selector switches based on multiple SOAs. At the same time, the butterfly topology structure is adopted to realize the interconnection of all nodes in a multi-level way. Within each substructure, the six SOAs correspond to all possible transport routes, significantly reducing the connection time between interfaces.

2.2.5. Interconnection based on multi-dimensional switching nodes [13] The circular datacenter interconnection structure based on multi-dimensional switching nodes was proposed by researchers from the Danish University of Science and Technology. It realized the multiplexing and demultiplexing of multi-core optical fiber by using photon integrated circuit technology. The model of the structure is shown in Fig.7. Servers connected to the ToR switch are interconnected through an optical circular network of multidimensional switching nodes. Each node supports the exchange of three dimensions, including space, wavelength and time. Optical circuit switches are the main components in different domains. There are fiber switching in the space domain, WSS in the wavelength domain and fast optical switches in the time domain. All connections to or from ToR should be fully arranged in wavelength and the time slots have to be disposed. To solve the problem, the same WSS or TDM switch provided by power splitting is strongly needed for multicasting and grooming of destining traffic. There may be several WSS and TDM switches per node in other
structures. However, in order to improve performance of multicasting and grooming among ToR ports, it can no longer be done in the same way as. For this reason, the fiber switch is used to enable these functionalities among ToR ports connected to different higher layer switches.

Figure 7. The structure of Interconnection based on multi-dimensional switching nodes [13]

2.3. Passive Optical Interconnection in Racks

2.3.1. Passive optical interconnection based on AWG [14] In the late 1980s, a professor from Delft University in the Netherlands came up with the concept of AWG. It is a group of gratings formed by an array of waveguides which have equal length difference, including an input waveguide, output waveguide, a focused planar waveguide, and an array waveguide. AWG can not only realize the function of multiplexing and demultiplexing, but also can be used for wavelength routing and interconnect among multiple ports. The structure consists of N servers and K links between ToR and the superstructure as shown in the Fig.8. The optical network interface (ONI) is used to receive and send optical signals from intra-rack and inter-rack. Thus, in order to realize all-to-all interconnection, this interconnection structure needs different wavelengths with the number of n+k. Due to the cyclic shift effect of AWG, the wavelength occupied inside and between racks needs to be planned in detail to avoid spectrum conflicts. Therefore, in order to achieve a conflict-free wavelength resource allocation, the communication between any two ports in the passive optical structure based on AWG can only be linked with fixed wavelength, and only through the channel [15].

The passive optical interconnection technology based on AWG has the advantages of high reliability. The composed network services are transparent and not limited by bandwidth, and it is easy to upgrade and expand capacity. Moreover, when the rack is expanded and the port is speeded up, the cost advantage of passive structure will become obvious due to its insensitivity to bandwidth changes compared with the traditional electric switch scheme. In addition, in the multi-server structure, the introduction of optical interconnection structure can greatly reduce energy consumption. However, the flexibility of the structure becomes limited because the communication between any two ports in the structure can only be connected with a fixed wavelength.
2.3.2. Interconnection based on passive optical coupler

An optical coupler is a passive device that can input any signal and broadcast it to all output ports. Its characteristics of broadcasting can meet the traffic demand of multicast in datacenter. In this structure, all the servers in the rack are connected through the optical network interface and the optical coupler, which has a pair of tunable lasers and filters for the transmitter and receiver respectively. For the traffic with the destination address inside the rack, the server port uses the broadcast method to send to all servers in the same rack. If the destination address is outside the rack, the wavelength direction is determined by the WSS. In Fig.9, the following three structures are examples of implementing an optical coupler-based interconnection structure on ToR. Scheme I is a N*2-based optical coupler interconnection structure, which is connected to the main body by a single fiber. Thus, it reduces the density of ToR port and improves the space utilization of the datacenter. Scheme II is a dual-fiber interconnection structure, which means that each server is connected to the coupler through a dual-port optical network unit. Therefore, at the cost of increasing the complexity of cabling, this structure reduces the insertion loss of communication. Scheme III shares the same basic principles as Scheme I, except that the structure is an interconnection based on a two-level cascade coupler.

Compared with AWG-based passive optical interconnection technology, the structure also has all the advantages of high reliability and large bandwidth. At the same time, in order to make full use of the spectral bandwidth advantage, the passive optical interconnection is considered to use dense wavelength division multiplexing (DWDM). The technology not only provides a huge amount of available bandwidth, but also offers interconnectivity for internal and external communications in the datacenter.

Nowadays, tunable transceivers are still quite expensive on the market and all three models mentioned above have different drawbacks. With the expansion of market demand and the progress of technology, it will be more competitive with the passive interconnection scheme of datacenter in the future.
3. Discussion

The c-Through architecture takes the optoelectronic network as a whole and leads the research in the field of optoelectronic hybrid interconnected datacenter network. This architecture combines the advantages of traditional electrical packet switching and optical circuit switching. It leads to the function that data between two switches can be propagated through optical circuits, while data not assigned to optical circuits can continue to be propagated through traditional network structures. However, only one micro-electro-mechanical system (MEMS) is used in the c-Through architecture, so there are extensibility problems caused by the limitation of the number of ports and delay issues caused by multiple ports. The MEMS also fails to achieve common many-to-one communication since optical circuit switching uses fixed single-wavelength routing. Helios is an optoelectronic hybrid datacenter network interconnection structure in which optical exchange exists as an auxiliary interconnection approach to high-speed transmission. Its electrical network uses multiple core switches connected to all ToR switches to achieve full interconnection. The structure adopts the WDM technology, so the Helios architecture can realize the many-to-one communication on the optical circuit switching. However, only one MEMS is used to connect to all the rack switches in the network, conducing Helios also has serious extensibility and delay problems. Proteus is suitable for data transmission with high data rate and low rate of change. It enlarges the number of connections between ToR switches and enhances link selectivity in optical switch arrays. With the coordination of WDM and WSS, the number of connections in ToR can be adjusted flexibly, thus the bandwidth elasticity will be enhanced. Finally, the large-capacity connections or one-to-many connections between switches can be realized. However, the disadvantages of this structure are the same as Helios. The switching rate of MEMS is low. Even if WSS is adopted, commercial WSS switching rate is close to that of MEMS, so the delay problem caused by reconfiguration of transmission link still exists. Thanks to the flexibility of the OSMOSIS architecture, the signals of a particular ToR switch can be accurately and independently selected for each switch. It achieves the switching function between any interface similar to that of an electrical
switch is realized. This structure has high throughput, especially when multiple light exchange unit structures are combined to achieve large-scale access capability. When the whole structure is treated as a switch, this access capability can be well applied in the scenario of an electrical switch. However, the adoption of a large number of SOAs results in a very bloated OSMOSIS structured system. It also brings the problem of increasing energy consumption and the great need for a well-behaved cooling system. The advantage of the IRIS structure is that it reduces the implementation difficulty of the control layer and eliminates the need to add random memory access to the system through a specific three-level structure. At the same time, the optical structure provides high network capacity and high-speed link switching. The obvious disadvantage of IRIS is that the system is bloated due to its complex three-stage structure and a large number of wavelength converters in the middle. The structure requires a large number of SOAs and active optical devices such as lasers, resulting in high power consumption and cooling requirements. The bidirectional photon network architecture reduces the amount of SOAs by changing the topology. It makes outstanding contributions which reduce the cost and efficiency of the entire architecture and create more flexible design and application scenarios. With the increase of the number of sub-switching structures, the capacity of the access node of the structure is greatly improved. The problem is that as the number of access points increases, the number of substructures is needed and a higher hierarchy is required to interconnect nodes. It means that the structure needs a precisely designed control layer and a well-arranged control procedure.

In the structure of interconnection based on multi-dimensional switching nodes, it provides excellent and uniform behavior along the full range of output ports and when increasing the group size for a single channel multicast, which could be a trend for future SDM-enabled optical data center networks. Asymmetric power ratio for 1 to 2 multicast and 2 to 1 grooming also results with relatively small penalty for low signal power. Moreover, bit error rate is extremely small for channels in different cores when simultaneous unicast and multicast switching as well as traffic grooming is performed with large throughput. Due to the lag of the research of space division multiplexing technology and the immaturity of the development of multi-core optical fiber, there is still some distance away from the commercial use of the structure.

4. Summary and Prospect
As the cornerstone of information society, datacenter has been playing a more crucial role in the development of today's society. Optical communication is born with the development of datacenter, which provides the upgrading of capacity and scale for the technical means of communication network. The proposed optical switching technology has developed into a very mature technical field, gradually completing the replacement of copper cable in the transmission network. The satisfaction of meeting the requirements of datacenter communication network in terms of low energy consumption, low cost, large capacity and low delay has been perfectly reflected. Several existing or proposed inter-optical network interconnection schemes of different datacenters has been introduced in this paper. Focusing on the optical interconnect technology between ToR, the e-Through and Helios architectures based on optoelectronic hybrid interconnection, as well as Proteus, OSMOSIS and IRIS architectures based on all-optical interconnect are discussed. In these structures, most optical interconnection schemes adopt WDM technology to increase the number of channels, network capacity and transmission rate. However, from the perspective of application, most of the schemes are still in the research stage and there will be some time before they are commercialized. The optical interconnection scheme in ToR includes passive optical interconnection based on AWG and passive optical coupler are stated as well. Compared with the former, the latter has more flexibility in the allocation of wavelength and spectrum resources, but the optical loss of the coupler makes the physical damage of the optical signal have a greater impact on the extensibility of the structure. Therefore, in the future, the datacenters still need a more efficient, agile and reliable structure, and the scalability of different optical interconnection schemes needs to be researched on the basis of different requirements.

References
[1] Luo Xiaohui. On the Development of Cloud Computing [J]. Electronic World, 2019, (8): 104.
[2] Bari M F, Boutaba R, Esteves R, et al. Data Center Network Virtualization: A Survey[J]. IEEE Communications Surveys & Tutorials, 2013, 15 (2): 909-928.
[3] R. N. Mysore, A. Pamboris, N. Farrington, et al. Portland: a scalable fault-tolerant layer 2 data center network fabric. ACM SIGCOMM Comput. Commun. Rev. 2009, 39(4): 39 – 50
[4] Y. Liu, P. Gao, B. Wong, et al. Quartz: a new design element for low latency DCNs. ACM SIGCOMM Comput. Commun. Rev., 2014, 44(4): 283 – 294
[5] G. Wang, D. G. Andersen, M. Kaminsky, et al. c-Through: part-time optics in data centers[J]. ACM SIGCOMM Computer Communication Review, 2010, 40(4): 327-338
[6] N. Farrington, G. Porter, S. Radhakrishnan et al. Helios: a hybrid electrical/optical switch architecture for modular data centers[C]. ACM SIGCOMM Comput. Commun. Rev. 2010. 40(41): 339-348
[7] Singla A, Singh A, Ramachandran K, et al. Proteus: a topology malleable data center network[C]. ACM Workshop on Hot Topics in Networks. HOTNETS 2010, Monterey, Ca, Usa - October, 2010: 1-6.
[8] Xu L. Feasibility Study on Topology Malleable Data Center Networks (DCN) Using Optical Switching Technologies [C]. Optical Fiber Communication Conference and Exposition, 2011: 1-3.
[9] Luijten R, Denzel W E, Grzybowski R R et al. Optical interconnection networks: The OSMOSIS, project[C]. Lasers and Electro-Optics Society, 2004. Leos 2004. The Meeting of the IEEE, 2004: 563-564 Vo 1. 2.
[10] Minkenberg C, Grzybowski R, Hemenway R, et al. Optical-packet-switched interconnect for supercomputer applications [Invited][J]. Journal of Optical Networking, 2004, 8(24): 900-913.
[11] Gripp J, Simsarian J E, Legrange J D, et al. Photonic Terabit Routers: The IRIS Project[C]. Optical Fiber Communication, 2010: 1 – 3
[12] Shacham A, Bergman K. An Experimental Validation of a Wavelength-Striped, Packet Switched. Optical Interconnection Network[J]. Journal of Lightwave Technology, 2009, 27 (7):841-850.
[13] V. Kamchevska, Y. Ding, K. Dalgaard, et al. On-chip SDM switching for unicast, multicast and traffic grooming in data centre networks. IEEE Photonics Technol. Letters, 2017, 29(2): 231-234
[14] M.K. Smit. New focusing and dispersive planar component based on an optical phased array. Electron. Lett., 1988, 24(7): 385-386
[15] Y. Cheng, M. Fiorani, L. Wosinska, et al. Reliable and cost efficient passive optical interconnects for data centers. IEEE Commun. Lett., 2015, 19(11): 1913 — 1916