Realization and Application of Large-Scale Fast Optical Circuit Switch for Data Center Networking

Ken-ichi Sato, *Fellow, IEEE*

(Invited Paper)

***Abstract*—Applying optical switching to data center networks can greatly expand network bandwidth and reduce electrical power consumption, both of which are needed to meet the explosive traffic increase. The optical switch offers large bandwidth switching capability, and so eliminates the multistage switch network architecture needed with electrical switching. Furthermore, the single stage architecture of the optical switch greatly simplifies operating costs, which include cabling, while substantially reducing the number of transponders needed. Data center networks place very different demands on optical systems than communication networks. Grasping the right direction to proceed is of paramount importance. To realize the full potential of optical switches, such as scalability and cost effectiveness, we analyze the role of large-scale optical circuit switches and discuss the realization technologies that combine the two dimensions of space and wavelength. Our recent advances in large port-count optical switches are presented.**

***Index Terms*—Electrical and optical hybrid switching, intra datacenter network, low energy consumption, optical switch.**

I. Introduction

THE intra data center traffic is estimated to be about four times the global Internet traffic and the CAGR is about 30% [1]. In data centers, traffic flow is changing from North-South to East-West, that is between servers/storages within a data center rather than inbound/outbound traffic [2]. Optical technologies are currently being used merely in form of point-to-point optical links that interconnect top-of-rack switches (ToRs) and other electrical switches. The ever increasing traffic and the recent tangible demise of Moore's law in electronics [3] will trigger the bandwidth and power consumption crunch in data centers in the near future. The critical role of optics in defining the next generation data center is expected to lie in networking. To avoid the barriers imposed by large bandwidth electrical switches, it is essential to off-load large traffic flows from the electrical layer to the optical layer. Leveraging optical fast circuit switching alongside electrical packet switching is recognized to be one of the most attractive solutions [4]–[11]. With the hybrid network solution, the giant data flows between ToR switches or pods will be delivered by optical circuit switches. Optics is underlying most of the telecommunication networks, however, many remaining barriers are hindering its use in data centers, since the requirements are very different: lower cost and shorter reach. This paper discusses the realization and application of large-port count optical switches that will match the data center requirements. The paper is organized as follows. Chapter II briefly analyzes advances in electrical technologies taking the supercomputer as an example, since recent hyperscale data centers can be regarded as extremely large-scale computers. The recent saturation in performance that stems from the electrical power consumption bottleneck is discussed. Chapter III examines the characteristics of present intra data center networks and discusses how such large intra datacenter traffic and growth rates can be supported by relatively small throughput component Silicon switches whose projected throughput growth is relatively low. In Chapter IV, the tangible problem of the exponential increase in power consumption of the intra datacenter network is discussed. It is explained how large-scale fast optical switches can resolve the problem and the requirements for optical switches are analyzed. Chapter V explains our approach to developing the required large-port-count optical switches. In Chapter VI, we highlight our recent technical achievements towards this goal. Chapter VII concludes the paper.

II. Advances in Electrical Technologies

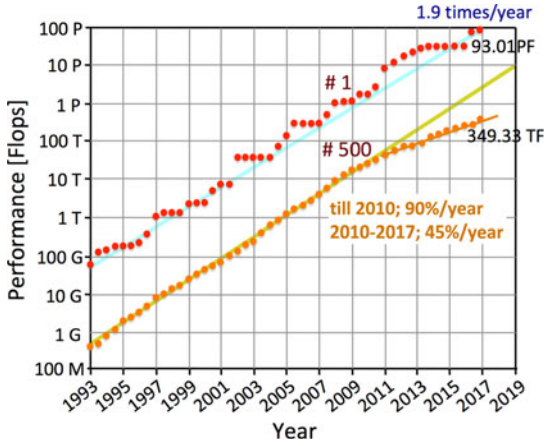
Fig. 1 depicts the performance (Flops, the LINPACK Benchmark) of the #1 and #500 supercomputers of the last 25 years [12]. They demonstrate a remarkable increase and the increase rate for the #1 supercomputer is 90% a year, however, that of #500 clearly shows a slow down after 2010, from 90% to 45%. Fig. 2 shows processor performance (SPEC CPU Benchmarks) and clock frequency growth relative to 1996 values [13]. The growth rates changed around 2010; they became rather smaller after 2010 as indicated by the rates in the figures. The clock frequency remained static after 2010, which mostly stems from the power consumption restriction of the Silicon chip. The processor performance advances are 1.38 and 1.14 before and after 2010, values that are smaller than the supercomputer performance advances (1.9 or 1.45). The gap is explained by the use of parallelism or the increase in the number of processor cores used. Fig. 3 shows the number of cores for #1 and the average

Manuscript received October 21, 2017; revised December 28, 2017 and January 29, 2018; accepted January 30, 2018. Date of publication February 1, 2018; date of current version March 1, 2018. This work was supported in part by NEDO and KAKENHI under Grant 26220905. *(Corresponding author: Ken-ichi Sato.)*

The author is with the Graduate School of Nagoya University, Nagoya 464-8603, Japan (e-mail: sato@nuee.nagoya-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JLT.2018.2801308



	Before ∼2010 (annual rate)	After ∼2010 (annual rate)
Performance (av. of Top 500)	1.9	1.45
Increase in # of cores (av. of Top 100): **A**	1.5	1.36
Processor performance advance: **B**	1.38	1.14
AxB	2.07	1.55

Fig. 1. Performance development in supercomputers and the major factors for the advances.

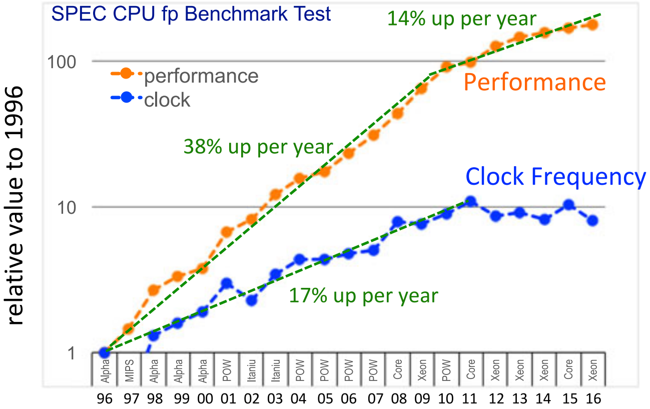


Fig. 2. Processor performance and clock frequency growth relative to 1996 values [11].

number of cores for the top 100 supercomputers [12]; the increase rate is also indicated. Considering this parallelism, the performance advance (CAGR) trend in supercomputers can be roughly explained as shown in the table in Fig. 1 (compare the product of A × B and the performance advances). Parallelism has almost saturated since the #1 supercomputer already uses more than 10 million cores and further expansion is becoming more and more difficult because of the total power limitation.

More fundamentally, CMOS performance improvements have slowed down. LSI (Large Scale Integration) power dissipation is proportional to the square of the driving voltage. Fig. 4 shows the reduction in CMOS driving voltage, together with predictions made by ITRS (the International Technology Roadmap for Semiconductor) in 1999, 2009, and 2013 [14], [15]. It is interesting to see that the predictions are repeatedly revised; the voltage reduction became more and more gentle. Scaling requires continual innovations in transistor structure and materials, however, the gate length of 5 nm, which corresponds to the near term target, is only ten or so times the Si atom lattice constant (0.54 nm). Thus power reduction in electronics is becoming more and more difficult to secure.

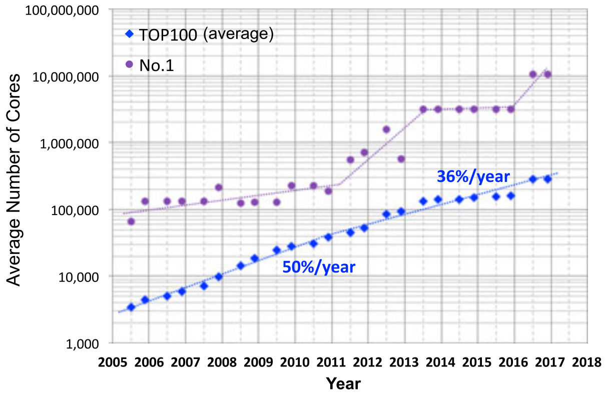


Fig. 3. Number of cores for #1 and the average number of cores for top 100 supercomputers.

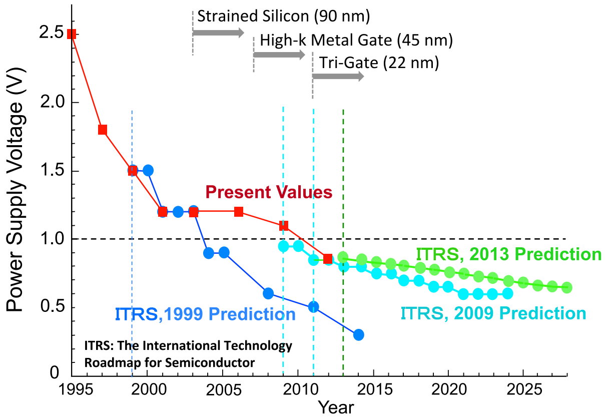


Fig. 4. CMOS LSI driving voltage.

The hyper-scale data center is sometimes recognized as a large-scale computer; some hundred thousands of servers/storages are interconnected through electrical switches. In Chapter III, we discuss present hyperscale data center networks that use electrical switches, and the current electrical switch bottleneck. Considering this, we then discuss how the bandwidth and power consumption crunch expected in the near future can be resolved by using optical switching technologies.

III. Intra Data Center Networks

A large-scale data center is created by the inter-connection of hundreds of thousands of server/storage systems. The switching network has to support huge bisection bandwidth, which can be 1 Pbps. The bisection bandwidth growth can be more than 100% a year in hyperscale datacenters [16], while the Silicon switch chip bandwidth advance is much smaller, around 50% [17] a

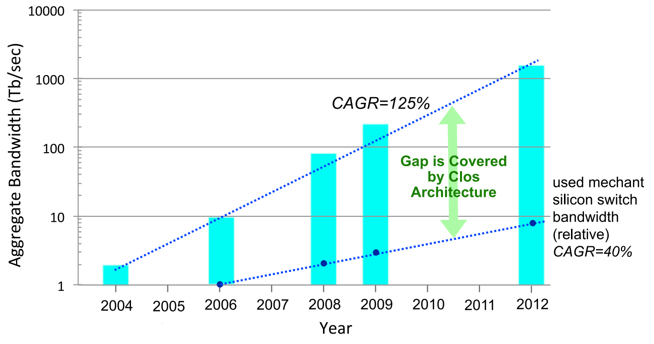


Fig. 5. Growth of Google datacenter aggregated bandwidth and merchant silicon switch bandwidth (relative to the year 2006).

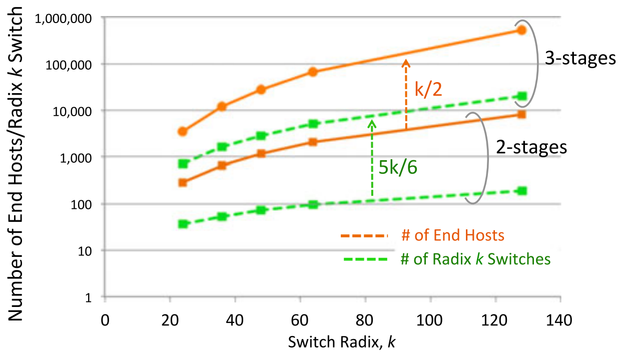


Fig. 6. Number of end hosts accommodated with multi-stage fat tree architecture with radix k switches.

year. Fig. 5 shows the growth in Google datacenter aggregated bandwidth and advances in the used merchant silicon switch chip bandwidth (relative to the year 2006) [16]. This switch chip bandwidth growth rate is expected to be reduced in the near future as discussed in Chapter II. Furthermore, the front surfaces of electrical switches tend to be fully covered with transceiver modules and the power consumption limits the module density [18]. To fill the gap between the needed bisection bandwidth growth and slow advances in Silicon switch chip bandwidth, a scalable networking technology that can effectively parallelize switch chip bandwidth is needed in electrical switch based intra datacenter networks.

Recent hyperscale data centers are based on the folded Clos (fat-tree) architecture using merchant Silicon switches or fabric switches, as this offers scalability [16], [19], [20]. Fig. 6 shows the number of end hosts ($k^n/2^{n-1}$) accommodated by the multi-stage (n-stage) fat-tree architecture with radix k switches, and the number of radix k switches needed ($k^{\text{n-1}}(1 + 1/2^{n-1})$). The number of hosts increases with the switch radix or number of stages, however, the number of radix k switches per host also increases as the number of stages increases for a fixed radix (Fig. 7). Thus you can increase the number of hosts accommodated by increasing the number of switching stages, at the cost of degrading switching efficiency or increasing number of switches per host. It is clear that higher radix switches accommodate more hosts cost effectively (less stages are needed), and hence advances in Silicon switch bandwidth (port speed × radix) are seen as key goals [17], [21]. Towards this, there are studies that use mid-board optics [21] as they are seen to be a useful approach to offset the stalled advances in electrical technologies (Moore's law) as discussed in Section II, by relaxing the size and power density limitation of present pluggable transceivers [22]. However, they do not resolve the Silicon switch chip bandwidth limitation, the fundamental problem of electrical switching technology limit that is now tangible. This paper focuses on another approach, introducing optical switches as discussed in Section IV.

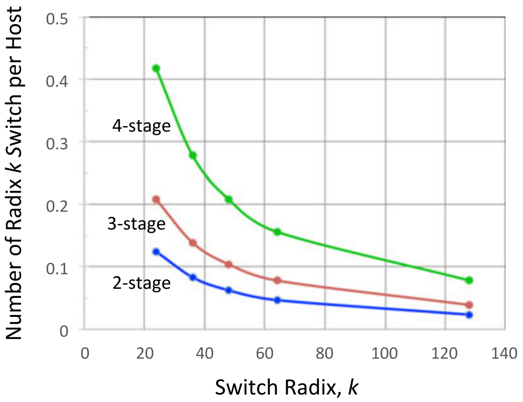


Fig. 7. Number of radix k switches per host in n-stage folded Clos network.

The traffic increase within data centers has also caused an explosion in the power consumption of electrical switches and routers. The degree of power reduction possible with CMOS technology advances is falling far behind the traffic growth as mentioned above. Indeed, the ratio of power consumption of networking and switching to the total datacenter power consumption is estimated to increase non-linearly, from 6.6% (2017) to 27% (2029) [23], if we rely on present electrical switching technologies. The application of optical switches can revolutionize data center networks through its inherent nature of transparency, which is discussed in the next chapter.

IV. Role of Optical Switches

Given that the traffic increase is over-running the advances in Silicon technology, optical technologies appear promising for switching large bandwidth traffic. Given that intra data center traffic is much larger than global IP traffic [1], optical switches will play a critical role in intra data center networks in the future. Please note that most optical switching schemes are transparent to the bitrates of the optical signals, which is completely different from electrical switching systems. In addition, the power consumption of optical switching, W/bit, is much smaller than that of electrical systems [24] and hence large bandwidth and low power consumption switch systems are possible [25], which is in contrast to the Silicon switch chip; the bandwidth is severely limited by its power consumption.

The optical switch offers large bandwidth switching capability, and so eliminates multi-stage switch network architecture needed with electrical switching [26]. Furthermore, the single stage architecture of the optical switch greatly simplifies operating costs, which include cabling, and substantially reduces the number of transponders needed [26]. Optical switching

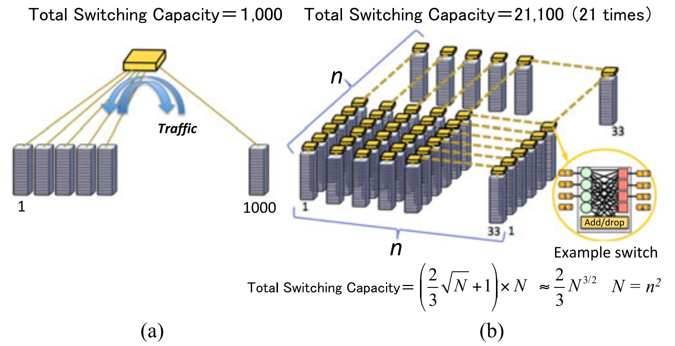
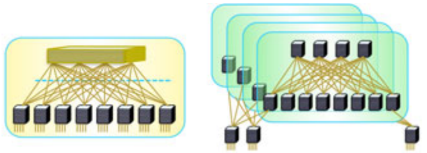

Total Switching Capacity＝1,000
Total Switching Capacity＝21,100（21 times）
Traffic
1
1000
n
n
1
33
33
1
Add/drop
Example switch
Total Switching Capacity＝$\left(\frac{2}{3}\sqrt{N}+1\right)\times N \approx \frac{2}{3}N^{3/2}$ $N=n^2$
(a)
(b)

Fig. 8. Comparison of required optical switch capacity. (a) Centralized switch and (b) Distributed switch configuration.



	Optical Switch		Electrical Switch
Switch Stages above ToR switches	single stage		2-stage (3-stage Fat Tree)
relative # of transceivers from one ToR to another ToR	link speed=k	link speed=1	link speed=1
	1/k	1	4

Fig. 9. Relative # of transceivers from one ToR to another ToR, where optical link speed is the same and single stage optical switches are assumed.

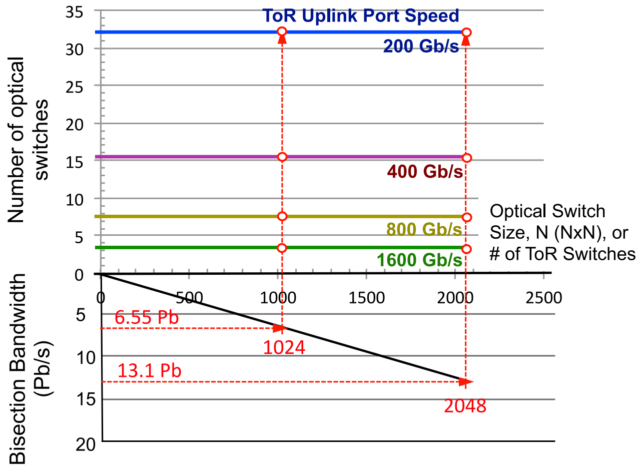


Fig. 10. Necessary number of optical switches versus optical switch size.

technologies have been widely deployed in the present core and metro networks as ROADMs/OXCs, however, one of the most stringent data centers requirements is cost effectiveness and scalability.

In data centers, traffic can be categorized according to flow size: mice flows and elephant flows. They differ in not only size, but also latency requirement. Elephant flows are associated with virtual machine migration, data backup, large file transfer including high-quality videos and so on, and most are not latency sensitive, while mice flows are very sensitive. Mice flows are dominant in terms of numbers, but elephant flows determine the total bandwidth [27]. Offloading elephant flows from electrical switches to optical switches (electrical and optical hybrid switch approach) can dramatically reduce the electrical switching bandwidth needed. For this purpose, fast optical circuit switching is necessary and is discussed hereafter.

In optical switch networks, large port count optical switches are effective, as discussed above, since employing small switches demands multi-stage configurations or the traversal of multiple switches between ToR switches. Fig. 8 compares two extreme cases where a single large port count optical switch or distributed small scale wavelength routing switches are used to connect N ToR switches. The latter approach demands $2/3N^{1/2}$ times more switching capacity (If $N = 1{,}000$, 21 times more capacity). Furthermore, traversing many switches complicates routing control (complex routing and wavelength assignment procedures are needed), increases connection set-up time, degrades optical transmission quality, and increases power consumption and total cost.

Reducing switching stages by using large port count switches can reduce the number of transceivers and interconnection fibers. Electrical switch and optical switch configurations are compared in Fig. 9. Here electrical ToR switches are connected through single stage optical switches or multi-stage electrical switches. For large data centers the number of switching stages is more than 2 (3 stages including ToR) [16], [19], [20] and hence optical switches reduce the number of transceivers by more than 75% (for same link speed), which substantially simplifies network configuration.

Fig. 10 shows the necessary number of optical switches and bisection bandwidth (no oversubscription) to interconnect ToR switches, each of which has throughput of 12.8 Tbps (128 × 128, 100 Gbps), versus the number of ToR switches. When the number of ToRs is 2048, the bisection bandwidth reaches 13 Pbps which is one order of magnitude higher than that of present large hyperscale data center networks. When ToR uplink speed is 200 Gbps, thirty two 2048 × 2048 optical switches are needed, while 3,072 12.8 Tbps electrical switches above ToRs are necessary to accommodate the same switching bandwidth. Please note that this parallel use of optical switches greatly reduces traffic collision at the destinations/sources, or the resultant sending delay (ToR switch is electrical and employ buffers). The performance (delay and possible data loss) can be analyzed using the multiple server model in queueing theory [28], where such parallelism greatly reduces information sending delay. The link speed and the necessary parallelism can be determined by considering the control policy of the datacenter accommodating different applications and link costs, but this exceeds the scope of this paper.

As discussed above, the introduction of optical switches substantially reduces the number of power hungry electrical switches and hence the electricity cost of intra datacenter networking can be greatly reduced. Furthermore, as shown in Fig. 9, the transparent optical switch system needs only 1/4 of the number of transceivers as the switching stage is reduced to one. The switch optical interface costs (including transceivers) can be a substantial part of network cost for present datacenter networks,

TABLE I
NUMBER OF SWITCHES AND OPTICAL LINKS NEEDED TO CREATE 0.74 PB BISECTION BANDWIDTH

	Electrical switch	Optical switch	
# of Leaf/Spine fabric switches	576	0	
# of NxN optical switches	0	1 (N=4,608)	3 (N=1,536)
# of 40G optical links	73,736	0	
# of 160 (or 200) G optical links	0	4,608	9,216

especially if a link demands speeds above 10 Gbps [29]. This indicates the potential network cost reduction possible with the introduction of optical switches.

Table I compares the number of component switches and optical links needed to create 0.74 Pb bisection bandwidth above ToRs with electric or optical switches; this assumes one example of the largest existing hypercale data centers that is constructed using 96 × 96 40 Gbps electrical switch fabric with four spine planes in the 3-stage Clos architecture [19]. Optical switches substantially simplify the overall switching system. Please note that, with optical circuit switching, only a limited number of electrical switches need to be used to create a hybrid switching system (not shown in Table I).

V. How to Create Large Port Count Optical Switches

The large bandwidth and low power consumption switching possible with optical systems stems from the inherent nature of bit rate agnostic optical switching. The power consumption of electrical switching is bit rate dependent and severely limits the switch chip bandwidth. While the requirements for optical circuit switches in data center applications or the breakeven points against electrical switching have yet to be identified, it is conceived that 100–300 ports will be inadequate and switching times of ∼100 μs will be the trigger to offloading traffic from electrical Ethernet switches [30]. Regarding the port count, 200 Gbps link speed and 300 ports demands a switch bandwidth (half Duplex) of 60 Tbps, which exceeds the capacity of the existing electrical switch fabric composed of many silicon switch chips. Of course we can create larger bandwidth electrical switching networks as discussed in Section IV by using multi-stage switch architectures, however, the required number of switches per host nonlinearly increases with the number of stages (see Fig. 6). Thus, optical switches become more effective as the port count is increased. In addition, the port cost of a switch should be low, the switch needs to offer cost-effective introduction even the early days (pay-as-you-grow capability), and it must be resilient to large scale failures (no single point of failure).

The 3D MEMS switch is not considered here because of its slow switching time (more than ms). Silicon photonics 2D planar matrix switches such as MEMS-actuated devices [31] and planar MZI devices [32] offer acceptable switching times, but available switch scale is rather small, less than 64. It will be difficult to expand the port count to more than 1,000. For example, a 1536 × 1536 switch needs ninety six 36 × 64, and sixty four 48 × 48 switches (3-stage Clos architecture), and traversing 3-stages may require an SOA/EDFA for compensating the three stage switch loss for each port. With matrix switches, the number of element switches (2 × 2 switch) increases nonlinearly (Fig. 11). While the hardware of wavelength routing switches increases linearly with the number of wavelengths, however, the available number of wavelengths is limited to around 100–200 (if channel spacing is 50 GHz, 88 channels in C-band). Our approach described in the next section effectively combines the different dimensions of space and wavelength routing. Silicon photonic technologies are adopted for creating space switches because of its switching speed and future envisaged cost effectiveness. Thus, the port count is given by the product of available wavelength number (∼100) and space switch port count; 96 wavelengths and 16 × 16 space switches yield 1536 × 1536.

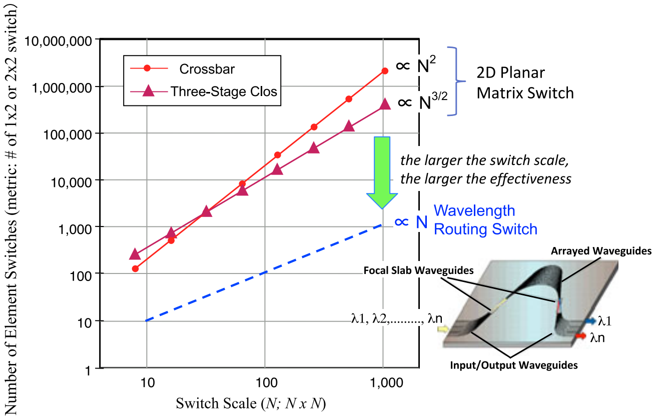


Fig. 11. Number of element optical switches versus optical switch size.

VI. Proposed Large Port Count Optical Switches

Fig. 12 depicts the architectures of our recently proposed optical switches. Table II summarizes necessary total and per port number of components for each architecture. Switching with wavelength routing can be implemented by tuning the optical source wavelengths or the tunable filters at optical receivers. Fig. 12(a) and (b) show the use of tunable lasers that will be implemented in ToR transceivers, while (c) shows using tunable filters at receiver side or the function is implemented with coherent detection. The architectures in Fig. 12(a) and (b) differ in wavelength routing switch arrangement. (a) uses wavelength routing subsystems that consist of cyclic AWGs (Arrayed Waveguide Gratings), while (b) uses combinations of an optical coupler and a conventional non-cyclic 1 × N AWG. Architectures (b) and (c) are almost symmetrical; the input and output directions are reversed and tunability at the input side is implemented in (b) by using wavelength tunable lasers, while that at the output side is done by using tunable filters in (c). Some details of each architecture are given below.

Please note that all proposed architectures yield large-port-count switches, and each uses disaggregated modules, which allows cost-effective introduction (pay-as-you-grow capability), and makes the systems resilient to large-scale failures (no single

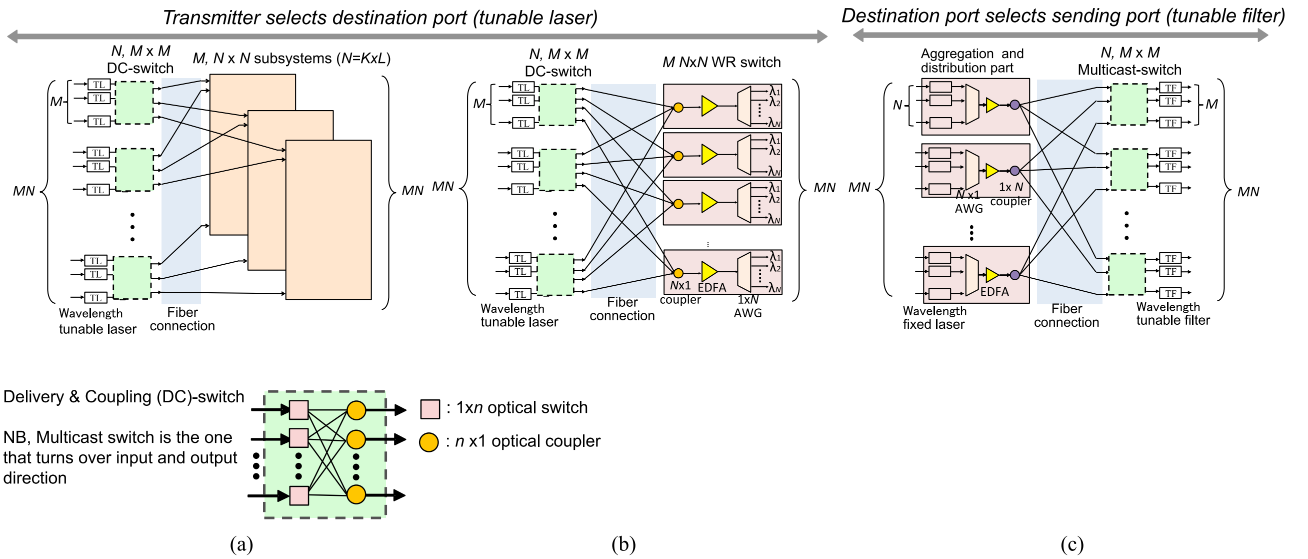


Fig. 12. Proposed *MNxMN* optical switch architectures (a) AWG based wavelength routing switch [33] (b) Tunable laser based wavelength routing switch [35] (c) Tunable filter based wavelength routing switch [36].

TABLE II
NUMBERS OF NECESSARY COMPONENTS FOR AN *MNxMN* OPTICAL SWITCH

Configuration		Tunable Laser	Fixed Laser	Tunable Filter	DC-switch	Nx1 OC	AWG	EDFA
(a)	Total #	MN	-	-	N	-	LM (KxK, cyclic) KM (LxL, cyclic)	-
	Per-port #	1	-	-	1/M	-	L/N (KxK, cyclic) K/N (LxL, cyclic)	-
(b)	Total #	MN	-	-	N	M	M (Nx1, non-cyclic)	M
	Per-port #	1	-	-	1/M	1/N	1/N	1/N
(c)	Total #	-	MN	MN	N	M	M (Nx1, non-cyclic)	M
	Per-port #	-	1	1	1/M	1/N	1/N	1/N

point of failure, another key weakness of large scale 3D MEMS switches).

A. Combination of Tunable Lasers and Cyclic AWGs

As for the WR (Wavelength Routing) switch, Fig. 12(a) uses combinations of cyclic Arrayed Waveguide Gratings (AWGs). The cross-talk of the cyclic AWG caused by optical signals from other input ports increases with the port count. To suppress the cross-talk, we developed a novel architecture that cascades small cyclic AWGs of different sizes (L $K \times K$ cyclic AWGs and K $L \times L$ cyclic AWGs) to create a $KL \times KL$ cyclic AWG [33], [34], where K and L ($\geq$2) must be mutually co-prime numbers [37]; then any input port can be connected to any output port with a specific wavelength. The architecture exhibits the useful features [34] of; 1) ITU-T fixed grid frequencies are utilized (a version of commercially available cost effective transceivers can be utilized), 2) the use of small cyclic AWGs offers small passband frequency deviation and low total cross-talk level, 3) high scalability, and 4) good modular growth capability. The total switch scale can be multiplied by using space switches that allow parallel use of the WR-switch part, which is attained with DC-switches (Delivery and Coupling switches) that consist of 1 × n tree or tap switches and n × 1 optical couplers [38], where the tree switch is composed of multi stages of MZIs (Mach-Zehnder interferometers). To verify the feasibility of the proposed architecture, we fabricated a prototype switch. The DC-switches were developed using silica PLC (Planar Lightwave Circuit) technologies; the switching time of the MZIs is ∼1 ms. The switching time can be substantially reduced by using Silicon photonic MZI switches as presented in Section VI-C. Fig. 13 shows a fabricated compact 270 × 270 optical switch (230 × 370 × 120 mm^3) with 3 × 3 DC-switches ($k = 3$), and 90 × 90 AWG subsystems each consisting of 10 9 × 9 and 9 10 × 10 AWGs. It can be easily expanded to 1440 × 1440 by setting k to 16. We can further expand the available port count by using the multi-stage switching architecture that uses AWGs and wavelength converters [39], however, the architecture prohibitively increases complexity, energy and cost.

B. Combination of Tunable Lasers and Non-Cyclic AWGs

The architecture in Fig. 12(b) uses optical couplers and AWGs in the wavelength routing part instead of cyclic AWGs as in Fig. 12(a), where the coupler loss can be compensated by EDFAs [40]. We developed the test system shown in Fig. 14; it interleaves the passbands of the paired 50 GHz grid AWGs with 25-GHz offset. The combination of an interleaver and AWGs makes the best use of their characteristics in a mutually complementary manner: the interleaver has few ports but a steep filter shape. Conversely, the AWG has gradual filter shape, but its port count can be large. With this proposal, we can construct a fine-resolution wavelength routing switch that enhances the spectral efficiency of the wavelength routing switch. Thanks to the high aggregation of wavelength signals and fine granular wavelength routing, we can achieve a large scale optical switch cost-effectively. The relatively expensive EDFA is shared by many wavelengths (for example 180 λs), which yields small per port cost (see Table II). We have tested the feasibility of the 1,440 × 1,440 switch by combining a fast-tunable laser,

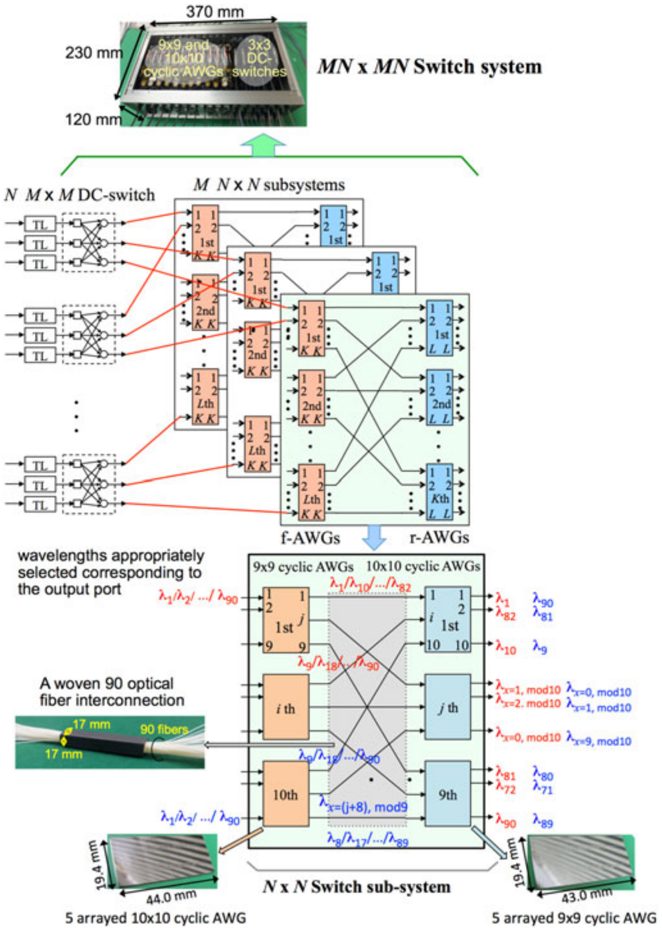


Fig. 13. Optical switch prototype that utilizes cascaded small cyclic AWGs.

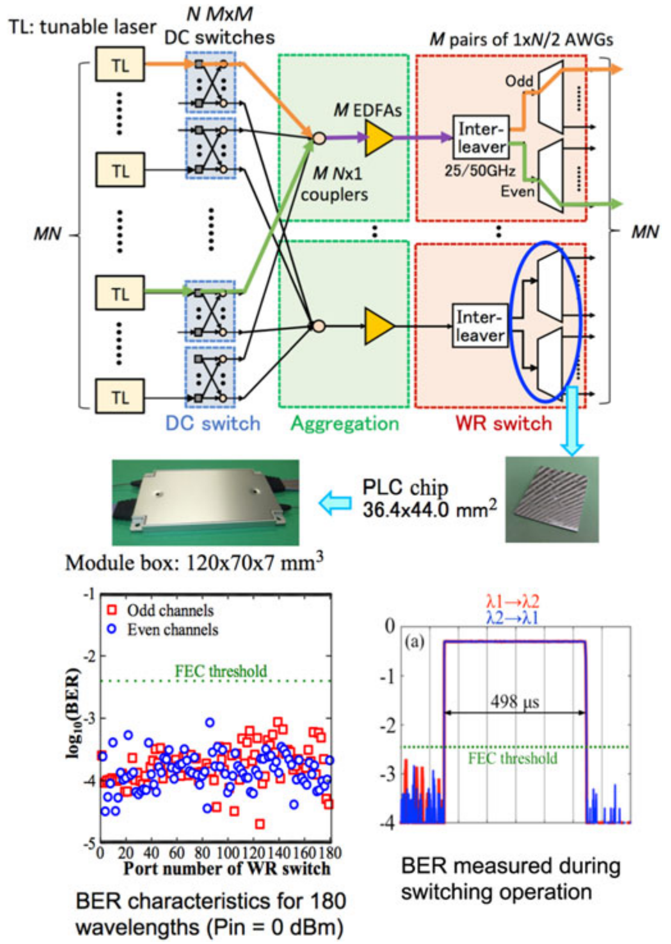


Fig. 14. Performance verification of optical switch that utilizes tunable lasers and non-cyclic AWGs.

8 × 8 DC switch, 180 × 1 coupler, EDFA, 1 × 2 interleaver, and pair of 1 × 90 AWGs monolithically integrated into a single PLC chip (36.4 mm × 44.0 mm) as shown in the inset of Fig. 14. Regarding the wavelength switching time, we measured 32,220 (180 × 179) combinations and the average and worst values were 348 μs and 436 μs, respectively, so shutter time was set to 498 μs with ~50 μs margin [40]. The laser wavelength switching time is expected to be reduced soon.

C. Combination of Fixed Wavelength Lasers and Tunable Filters

The architecture in Fig. 12(c) uses a combination of fixed wavelength lasers and tunable filters for wavelength routing [36]. EDFAs are also shared by many ports. The output port selects the desired input port. The fast wavelength tunability of lasers needed by the previous architectures may demand sophisticated laser technologies, however, a tunable filter can be created cost effectively by using Silicon photonic technology. The developed tunable filter uses thermo-optically controlled asymmetric Mach-Zehnder interferometers (MZI). It consists of 8 stages, and each stage MZI has a free-spectral range of $40/2^n$ nm ($n = 1 - 8$). By selecting one of the output of MZI stages, filter bandwidth can be selected to adjust to the transport signal bandwidth. The selectable 3-dB bandwidths were set at 17, 33, 66, or 256 GHz. A polarization insensitive multicast switch was also developed with Silicon photonic technology as shown in the Fig. 15 inset. Low loss (~15 dB; 9 dB 1 × 8 optical coupler intrinsic loss, less than 1 dB fiber coupling loss per facet, and less than 4 dB MZI and waveguide loss), low PDL (av. 0.6 dB), and low crosstalk (less than −35 dB) are attained [41]. The switching time of this switch is determined by that of the multicast switch or tunable filter. The two devices demonstrated switching times of 40 μs and 30 μs, respectively (see Fig. 15). The switching time can be reduced by applying a new MZI heater driving method [42]. Indeed, 6.1 μs turn-on and 4.3 μs turn-off switching times have been attained by a Si photonics MZI switch [42]. Proof of concept experiments for the proposed switch architecture were conducted using 10G(IM) and 56G(PAM-4) signals [36], and more technical details are given in [36].

D. Comparisons of Architectures

It is too early to identify the best architecture, which is likely to depend on the datacenter requirements such as scale and switching speed, both of which will be influenced by the applications. Some generic comparisons of the presented architectures are given below. One of the criteria in switch architecture

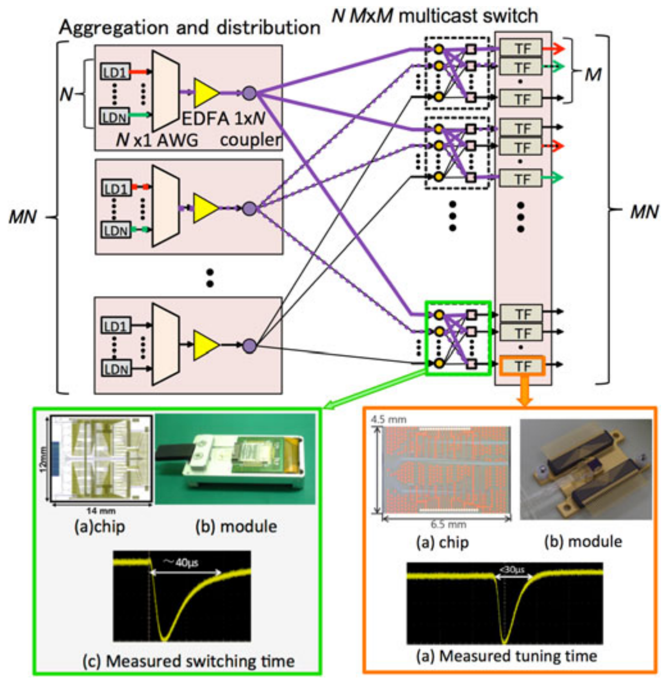


Fig. 15. Performance verification of optical switch that utilizes tunable filters.

selection is the availability of tunable lasers or tunable filter functions (in the future the function will be realized by coherent detection), where the cost is, of cause, a major criterion. Such devices are not yet mature enough and also their reliability needs to be assured (even for relatively shorter life times, say, ∼5 years, for datacenters), which can be a concern for active devices like tunable lasers.

Regarding the wavelength tuning speed of lasers, ∼1 μs is expected using the Distributed Bragg reflector (DBR) laser [43], and for tunable filters 4–6 μs has been achieved [42] with Silicon photonics technologies. Further reductions can be expected in the future. From the cost and the reliability aspects, the Silicon photonics tunable filter may be the first to be utilized. Another point to be noted is that the architecture of Fig. 12(a) has no channel aggregation points in the switch, which means that if the optical loss of each path, which is mostly determined by the space switch loss (optical coupler loss in the DC-switch part) and AWG loss, exceeds a certain limit, amplifiers will need to be added port by port since sharing by many ports is not possible. On the other hand, the architectures shown in Fig. 12(b) and (c) have signal aggregation points, so the relatively expensive optical amplifiers can be shared by many ports as depicted in Table II. The architectures depicted in Fig. 12 are basic and can be modified. In (a), the AWG stages can be extended to more than 3 to further expand the wavelength routing switch port number [44], which may need per port amplification (using EDFA or SOA) to compensate the increased losses. In (b), an N × 1 optical coupler located before the 1 × N non-cyclic AWG can be replaced by L (N/L) × 1 couplers and a L × N cyclic AWG. The L × N cyclic AWG functions as an L × 1 coupler and 1xN AWG, simultaneously. This would reduce the coupler loss by 10log L dB, while the AWG loss increase due to adopting the LxN cyclic AWG instead of non-cyclic 1xN AWG is smaller than the coupler loss reduction, and hence the total loss can be reduced and available switch size can be expanded. Using this scheme, a 153.6-Tbps throughput (1,536 × 1,536 at 10 Gbps) optical switch with Uniform-Loss and Cyclic-Frequency (ULCF) AWGs [45] was developed. Details will be presented soon [46]. Other modifications and optimizations to the necessary switch scale and channel speed are possible and work on this will be published elsewhere.

VII. Conclusion

Electrical switch power consumption in data center networks can be expected to increase non-linearly with the traffic [23]. In the face of the approaching demise of Moore's law, optical technologies will be needed to change the paradigm of intra datacenter networks. However, the requirements placed on optics when applied to data center networks are very different from those imposed by communication network use, including low cost, high expandability and short reach. Grasping the right direction to proceed and promoting the necessary technology developments are of supreme importance. One of the barriers hindering the use of optical networking technologies is the lack of large-scale cost-effective optical switches, which must be developed soon. Some of our recent developments towards this goal were presented. Developing associated technologies, network control technologies [10] and new protocols, are critical to making the best use of optical switch capabilities. The number of optical upper links of each ToR destined to optical switches or number of optical switches used in parallel determines the flow-sending delays, which also impacts total optical link cost. Optimization can be done by considering the different applications that each datacenter provides. These discussions are out of the scope of this paper.

Acknowledgment

Authors would like to thank Profs. H. Hasegawa, Y. Mori, and R. Shioya for their useful discussions. We are also indebted to our colleagues in AIST, NTT, and NEC.

References

[1] Cisco Global Cloud Index: Forecast and Methodology, 2011–2016.

[2] OIDA Workshop report, Future needs of scale-out data centers, 2013.

[3] R. Courtland, "Transistors could stop shrinking in 2021," *IEEE Spectrum*, vol. 53, no. 9, pp. 9–11, Sep. 2016.

[4] N. Farrington *et al.*, "Helios: A hybrid electrical/optical switch architecture for modular data centers," in *Proc. ACM SIGCOMM*, 2010, pp. 339–350.

[5] G. Wang *et al.*, "C-Through: Part-time optics in data centers," in *Proc. ACM SIGCOMM*, New Delhi, India, Aug. 2010, pp. 327–338.

[6] N. Farrington *et al.*, "A 10 μs hybrid optical-circuit/electrical-packet network for datacenters," in *Proc. 2013 Opt. Fiber Commun. Conf. Expo. Nat. Fiber Opt. Eng. Conf*, Anaheim, Mar. 2013, OW3H.3.

[7] Y.-K. Yeo, Q. Huang, L. Zhou, "Large port-count optical cross-connects for data centers," in *Proc. Photon. Switch.*, Corsica, France, Sep. 11–14, 2012

[8] P. N. Ji *et al.*, "Design and Evaluation of a Flexible-Bandwidth OFDM-Based Intra-Data Center Interconnect," *IEEE J. Sel. Topics Quant. Electron.*, vol. 19, no. 2, Mar./Apr. 2013, Art. no. 3700310.

[9] Z. Cao *et al.*, "Hi-LION: Hierarchical large-scale interconnection optical network with AWGRs [invited]," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 7, no. 1, pp. A97–A105, Jan. 2015.

[10] S. Spadaro, “Control plane architectures for photonic packet/circuit switching-based large scale data centres”, in *Proc. ECOC 2013 Symp.*, London, U.K., Sep. 24, 2013.

[11] J. Perelló *et al.*, “All-optical packet/circuit switching-based data center network for enhanced scalability, latency, and throughput,” *IEEE Netw.*, vol. 27,no. 6, pp. 14–22, Dec. 2013.

[12] [Online]. Available: https://www.top500.org

[13] R. Shioya, “Research trends of processor architecture,” The Institute of Electronics, Information and Communication Engineers (IEICE), General Conference, SS-7, Kusatus, Japan, Mar. 10–13, 2015.

[14] M. Takamiya, H. Shinohara, and T. Sakurai, “Low energy LSI with extremely low voltage operation,” *J. IEICE*, vol. 93, no. 11, p. 943, 2010 (in Japanese).

[15] International Technology Roadmap For Semiconductors 2013 Edition.

[16] A. Singh *et al.*, “Jupiter Rising: A decade of clos topologies and centralized control in google’s datacenter network,” in *Proc. SIGCOMM 15*, London, U.K., Aug. 17–21, 2015, pp. 183–197.

[17] M. Nowell, “50 Gb/s ethernet over a single lane and next generation 100 Gb/s & 200 Gb/s ethernet call for interest consensus presentation,” *IEEE 802.3*, Nov. 10, 2015.

[18] A. Ghiasi, “Large data centers interconnect bottlenecks,” *Opt. Exp.*, vol. 23, pp. 2085–2090, 2015.

[19] N. Farrington and A. Andreyev, “Facebook’s Data Center Network Architecture,” in *Proc. IEEE Opt. Interconnects Conf.*, 2013, pp. 49–50.

[20] D. A. Maltz, “Scaling bottlenecks in data center networks,” in *Proc. OFC 2014*, San Francisco, CA, USA, Mar. 9-13, 2014, Paper Tu2I.1.

[21] H. Dorren *et al.*, “Challenges for optically enabled high-radix switches for data center networks,” *IEEE J. Lightw. Technol.*, vol. 33, no. 5, pp. 1117–1125, Mar. 2015.

[22] R. Blum, “Scaling the compute and high speed networking needs of the data center with silicon photonics,” *presented at the ECOC 2017 Market Focus, Datacenters: Optics in Cloud Computing, session 1*, Gothenburg, Sweden, Sep. 19, 2017.

[23] International Technology Roadmap For Semiconductors 2.0, 2015 Edition.

[24] R. S. Tucker, “Green optical communications—Part II: Energy limitations in networks,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 17, no. 2, pp. 261–274, Mar./Apr. 2011.

[25] K. Sato and H. Hasegawa, “Optical networking technologies that will create future bandwidth-abundant networks [invited],” *J. Opt. Commun. Netw.*, vol. 1, no. 2, pp. A81–A93, Jul. 2009.

[26] K. Sato, “Realization and application of large-scale fast optical circuit switch for data center networking,” in *Proc. ECOC 2017*, Sep. 17–21, 2017, Paper Tu2F.1.

[27] A. Greenberg *et al.*, “VL2: A scalable and flexible data center network,” in *Proc. SIGCOMM*, Barcelona, Spain, 2009, pp. 51–62.

[28] L. Kleinrock, *Queueing Theory*. 1st ed. Hoboken, NJ, USA: Wiley, Jan. 2, 1975.

[29] L. A. Barroso *et al.*, *The Datacenter As a Computer, An Introduction to the Design of Warehouse-Scale Machines.* 2nd ed. San Rafael, CA, USA: Morgan & Claypool, 2013.

[30] A. Vahdat, “Delivering Scale Out Data Center Network- ing with Optics — Why and How,” in *Proc. Opt. Fiber Commun. Conf. Expo., 2012 Nat. Fiber Opt. Eng. Conf.*, San Diego, CA, USA, Mar. 2012, Paper OTu1B.1.

[31] M. C. Wu *et al.*, “MEMS-enabled scalable silicon photonic switches,” in *Proc. Front. Opt./Laser Sci.*, San Jose, CA, USA, 2015, Paper FW3B.2.

[32] K. Tanizawa *et al.*, “32 × 32 strictly non-blocking Si-wire optical switch on ultra-small die of 11 × 25 mm2,” in *Proc. Opt. Fiber Commun.*, Los Angeles, CA, USA, 2015, Paper M2B–5.

[33] T. Niwa *et al.*, “Large port count wavelength routing optical switch that consists of cascaded small-size cyclic arrayed waveguide gratings,” *IEEE Photon. Technol. Lett.*, vol. 24, no. 22, pp. 2027–2030, Nov. 2012.

[34] K. Sato *et al.*, “A large-scale wavelength routing optical switch for data center networks," *IEEE Commun. Mag.*, vol. 51, no. 9, pp. 46–52, Sep. 2013.

[35] K. Ueda *et al.*, “Fast optical circuit switch for intra-datacenter networking,” *IEICE Trans. Commun.*, vol. e100–b, no. 10, pp. 1740–1746, Oct. 2017.

[36] K. Ueda *et al.*, “Large-scale optical circuit switch for intra-datacenter networking using silicon-photonic multicast switch and tunable filter,” in *Proc. 42nd Eur. Conf. Opt. Commun.*, Dusseldorf, Germany, Sep. 2016, Paper W.2.F.2.

[37] K. Ishii, H. Hasegawa, and K. i. Sato, “Formulation of MUX/DEMUX functions for multiple input-output port cyclic AWG,” in *Proc. 2012 Asia Commun. Photon. Conf.*, Guangzhou, China, 2012, Paper AS4A.3.

[38] K. Sato, *Advances in Transport Network Technologies*. Norwood, MA, USA: Artech House, 1996.

[39] K. Xi, “A petabit bufferless optical switch for data center networks,” in *Optical Interconnects For Future Data Center Networks*, New York, NY, USA: Springer-Verlag, 2012, vol. 3, pp. 135–154.

[40] K. Ueda *et al.*, “Demonstration of 1,440 × 1,440 fast optical circuit switch for datacenter networking,” in *Proc. 21st OptoElectron. Commun. Conf./ 2016 Int. Conf. Photon. Switch.*, Niigata, Japan, July 3–7, 2016, Paper WF1-3.

[41] S. Nakamura, S. Yanagimachi, H. Takeshita, A. Tajima, T. Hino, and K. Fukuchi, “Optical Switches Based on Silicon Photonics for ROADM Application,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 22, no. 6, pp. 185–193, Nov./Dec. 2016, Art. no. 3600609.

[42] H. Matsuura *et al.*, “Accelerating switching speed of thermo-optic MZI silicon-photonic switches with “Turbo Pulse” in PWM Control,” in *Proc. Opt. Fiber Commun. Conf.*, Los Angeles, CA, USA, 2017, Paper W4E.3.

[43] S7500 Product Specification, CW Tunable Laser, Finisar

[44] K. Ueda *et al.*, “Large-scale optical switch utilizing multistage cyclic arrayed-waveguide gratings for intra-datacenter interconnection,” *IEEE Photon. J.*, vol. 9, no. 1, Feb. 2017, Art. no. 7800412.

[45] K. Okamoto, T. Hasegawa, O. Ishida, A. Himeno, and Y. Ohmori, “32 × 32 arrayed-waveguide grating multiplexer with uniform loss and cyclic frequency characteristics,” *Electron. Lett.*, vol. 33, no. 22, pp. 1865–1866, Oct. 1997.

[46] H. Nagai, Y. Mori, H. Hasegawa, and K. Sato, “ Demonstration of 153.6-Tbps throughput from 1,536 × 1,536 optical switch with uniform-loss and cyclic-frequency AWGs,” in *Proc. SPIE*, 2018, Paper 105600.

Ken-ichi Sato (M’87–SM’95–F’99) received the B.S., M.S., and Ph.D. degrees in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1976, 1978, and 1986, respectively.

He is currently a Professor with the Graduate School of Engineering, Nagoya University, and he is an NTT R&D Fellow. Before joining the university in April 2004, he was an Executive Manager with the Photonic Transport Network Laboratory at NTT. He has authored or coauthored more than 450 research publications in international journals and conferences. He holds 40 granted patents and more than 100 pending patents. His most significant achievements lie in two of the important transport network technology developments. One is ATM (Asynchronous Transfer Mode) network technology, which includes the invention of the Virtual Path concept. The other is photonic network technology, which includes the invention of the optical path concept and various networking and system technologies. His contributions extend to coediting the IEEE Journal on Selected Areas in Communications (four special issues) and the Journal of Lightwave Technology (three special issues); organizing several workshops and conference technical sessions; serving on numerous committees of international conferences including OFC 2016 General Chair and OFC 2014 Program Chair; authoring and coauthoring 14 books. He is a Fellow of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan. He served as the President of the IEICE during 2016–2017. He was the recipient of the Young Engineer Award in 1984, the Excellent Paper Award in 1991, the Achievement Award in 2000, and the Distinguished Achievement and Contributions Award in 2011 from the IEICE of Japan, the Best Paper Awards in 2007 and 2008 from the IEICE Communications Society, the Distinguished Achievement Award of the Ministry of Education, Science and Culture in 2002, and the Medal of Honor with Purple Ribbon from Japan’s Cabinet Office in 2014.