“Digitalizing” Optical Layer for The Green Computing Continuum As The Future Digital Infrastructure

Shu Namiki and Kiyo Ishii
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan,
{shu.namiki, kiyo-ishii}@aist.go.jp

Abstract: This talk will review, from the perspective of photonics, the technological trends of the digital infrastructure migrating toward the “computing continuum,” where the optical network and computing infrastructure are converged. The functional block-based disaggregation (FBD) model will be introduced as a key to incorporate the optical layer switching into the future digital infrastructure. © 2022 The Authors

1. Introduction

The disaggregation of servers in data centers, where various, heterogeneous compute resources are pooled and optimally reconfigured for different purposes on a common virtualized platform, is attracting increasing attention as a compelling countermeasure against the demise of Moore’s law because it can save the hardware resources, and hence the overall energy consumption of the data center. Indeed, the concept of disaggregated computing is extended not only from intra-rack to multi-rack scale, but also from cloud to edge computing indefinitely, thus ultimately forming the so-called “computing continuum [1]”. However, it is rarely pointed out that this seemingly compelling trend contains an inherent paradox: the more the disaggregated computing scales, the ever-higher performances the interconnect/network will require. For example, one big switch inter-connecting all disaggregated compute nodes is demanded for next generation hyperscale data centers [2]. The pertinent performances of the network here are the bandwidth, latency, energy efficiency, security, and dependability. These performances depend on transceivers and switches, while the latter consume energy predominantly, and are totally subject to the slowdown of Moore’s law, which leads thus regressively to the energy crunch of the entire digital infrastructure [3].

Optical layer switching is uniquely attractive because it can offer simultaneously a physically guaranteed high bandwidth, speed-of-light low latency, extremely high energy efficiency, and physics based high security. Among various types of optical switches [4,5], silicon photonic switches comprising Mach-Zehnder Interferometers with thermo-optic phase shifters [6] are one of the most robust options with decent yet unique advantages, such as compatibility with CMOS processes, fast (microsecond) switching speed, thermal and mechanical stability, reliable and established packaging, and high throughput of calibration and testing. Recently, authors’ group demonstrated 16-ch.×32-Gbaud QPSK WDM transmission by nine-time cyclic propagations through a fully loaded 32×32 silicon photonics switch, corresponding to a total bi-section bandwidth of 125 Pb/s for a nine stage Clos network configuration (131,072 ports × 0.952 Tb/s) [7,8]. The 32 x 32 switch used in this experiment had a wall-plug power consumption of 23.6 W. Considering a bi-section bandwidth of 30.5 Tbps in this experiment, this switch has an energy efficiency of approximately 0.8 pJ/bit. This value is remarkably superior to a value 68 pJ/bit that is the current wall-plug efficiency of a 25.6-Tbps spine/leaf switch blade used in the hyperscale data centers [9]. This tremendous difference (by almost two digits) in energy efficiency motivates us to ponder how to exploit optical layer switching with such huge potential for the real systems.

However, a vital shortcoming of optical layer switching is that it can only offer fast circuit switching and not full packet switching functionality. Therefore, a real challenge lies in how to overcome such a shortcoming and fully enjoy the merits of optical layer switching. At least, this challenge may not be able to resolve until a cross-layer holistic systems approach is taken such that the computing layer help overcome and/or circumvent such formidability of optical layer switching. One straightforward means would be the hybrid use of electrical packet switching and fast optical circuit switching [10]. For example, if the 30 % of the traffic could be successfully off-loaded from ASIC based switches to optical layer switching (say 100 more energy efficient), then 29.7% of energy could be saved in theory. However, the real traffic patterns are not always favorable for such hybrid switching, presumably influenced by not only algorithms but also the extent of how data are physically fragmented over the racks. In this regard, we must seek for new compute-network converged architectures that e.g. control the traffic pattern always favorable for the hybrid switching, and/or that suit in the context of emerging disaggregated computing architectures [11]. This talk will address the digitalization of the optical layer by means of the disaggregation of hardware along with its machine-readable model that automatically links the optical layer to upper layers up to the computing layer.

2. Disaggregation of servers by optical layer switching

The concept of disaggregated computing is substantiated by a mechanism that allows to reconfigure disaggregated heterogeneous compute nodes, such as CPUs, accelerators, xPUs, memories, and network interface cards. Ideally, all of these compute nodes are connected to one common big switch with a flat topology. However, as discussed in Introduction, this generic approach is leading to the energy crunch of the network. There have been attempts to utilize a hybrid use of optical circuit switches and electrical switches within disaggregated servers [12, 13], which showed that the communication latency mostly set by the physical propagation delay severely limited the computation performances. This poses a more vital drawback of using such one big switch with a large footprint as it results in high latency on average. As a result, the core elements of a disaggregated server have to be located within a certain physical distance permitted by the latency requirement. This reversely limits the optimum number of disaggregated nodes in a disaggregated server. Here let us consider a case that there are N disaggregated nodes with an I/O bandwidth of C bit/s each in a data center, and k disaggregated servers are configured such that each disaggregated server contains N/k nodes on average. If we adopt the one big switch approach, then the switch must have a bi-section bandwidth of NxC. On the other hand, if we build k disaggregated servers individually, each server carries a switch with a throughput of N/k x C on average. Assuming that the energy consumption of switches scales quadratically with throughput, the ratio of the total energy consumption of the latter to the former case would be ∝ 1/k. In a hyperscale data center, k can be more than tens of thousands. This simple estimation favors the use of k small switches, rather than one big switch. Then, optical layer switching will play a role of reconfiguring disaggregated nodes to each of small switches. Figure 1 illustrate an example of such a system. The above-mentioned silicon photonics switches are suitable for this purpose. The number of ports required for this switch is at least mk where m denotes the number of disaggregated servers per rack. For instance, mk = 8x16 (128). Because silicon photonics switches scale well [7], the optical switches in Fig. 1 offer inter-rack interconnects for multi-rack scaling as far as required latency allows.

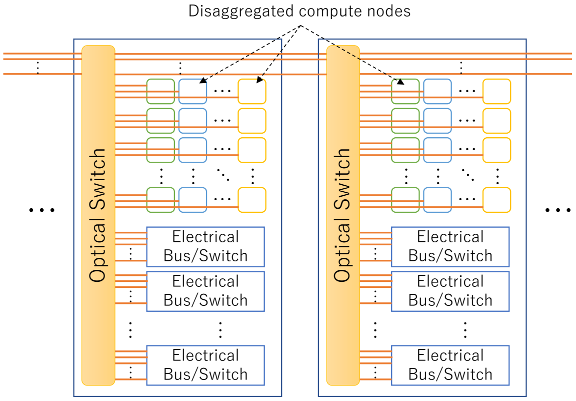

Fig. 1 Disaggregated server racks reconfigured by optical layer switching

3. Disaggregation of the optical layer

The wide use of optical layer switching in disaggregated computing will naturally extend to seamless connections outside data centers by virtue of optical fiber communications. For example, remotely located data centers could be seamlessly connected between each other and regarded as one large, disaggregated data center. This process will also be converged with mobile/multi-access edge computing, then migrate toward the so-called computing continuum where the compute and network will no longer be managed separately on nationwide scale. However, today's optical networks outside data centers are rigidly segmented by various topologies such as ROADM ring and star topologies. By the era of 6G, the current optical layer infrastructure has to be substantially upgraded [11,14]. In general, the optical layer can take any topologies by combining various kinds of optical switches such as optical cross-connects, wavelength selective switches, multicast and select switches, splitters/couplers. The disaggregation of the optical layer is a concept that by regarding each of these devices as optical functional block, any optical network topologies can be constructed even by unskilled entities. The concept is depicted in Fig. 2. The idea is that each of functional blocks operates as a pluggable module, and therefore, automatic plug-and-play architecture to "digitalize" the optical layer is the key.

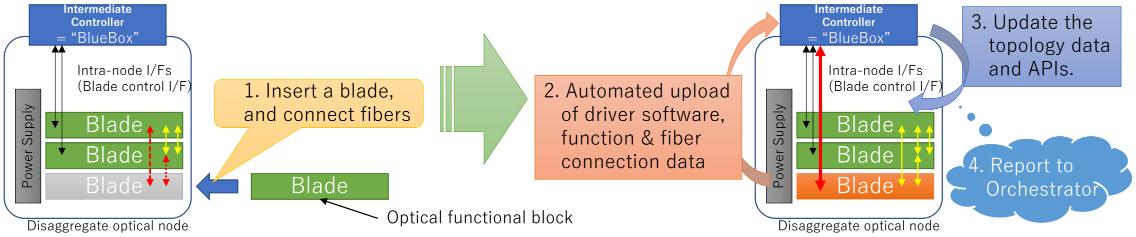
Intermediate Controller = "BlueBox"
Intra-node I/Fs (Blade control I/F)
Power Supply
Blade
Blade
Blade
Disaggregate optical node
1. Insert a blade, and connect fibers
Blade
Optical functional block
2. Automated upload of driver software, function & fiber connection data
Intermediate Controller = "BlueBox"
Intra-node I/Fs (Blade control I/F)
Power Supply
Blade
Blade
Blade
Disaggregate optical node
3. Update the topology data and APIs.
4. Report to Orchestrator

Fig. 2 Conceptual diagram of the disaggregation of the optical layer. Each disaggregated optical functional block is enclosed in a blade.

4. Ontology, then digitalization of the optical layer

For the disaggregated optical layer to converge with upper layers up to the computing layer, the detailed status of the optical layer must be at least accountable in the digital domain. In other words, the creation and maintenance of a digital twin, or so-called dynamic map of the optical layer is required. The functional block-based disaggregation (FBD) model is a component-level model that provides a one-to-one correspondence between the model and the actual hardware [15], which means any optical layer with any topology can be precisely described *as it exists*. Likewise, FBD model offers a platform for the digitalization of the optical layer that includes the automatic acquisition and generation of the topology information, and the control of the signal integrity over every optical path, as well as the means to exchange pertinent information with the upper layers through appropriate translation and/or abstraction.

The hierarchical relationships among the models are shown in Fig. 3. Because the FBD model serves as database with a complete set of information that can reproduce the switching functionalities and node configurations, it can generate mappings between the real hardware composition and any other abstracted models. For example, transformation algorithms bridging the OpenROADM device model and the FBD model have been demonstrated in [16]. Automated network operations, including node structure updates for failure recovery as well as optical path establishment/removal, were successfully demonstrated on a field testbed. The node structure update was completed within 5 minutes, and the multidomain cooperative optical path recovery triggered by the update was swiftly accomplished without manual configuration. Another example is about TAPI: the transformation algorithms and the path computation service based on the FBD model have been demonstrated in [17], where externalization of the path computation function was achieved simply owing to the component-level model described in a machine-readable manner. Moreover, optical path provisioning within 30 s has been successfully demonstrated on a real hardware testbed where an SDN controller (cf. a WDM orchestrator and an open line system controller (OLS-C)) was operated at a “node” level abstraction, whereas the path computation function was operated at the “optical component” level, which enabled an accurate path computation service for any node structure without increasing the complexity of the network OS. These transformation algorithm developments (i.e., automatic nodal structure analysis) and the simple path computation server implementation are enabled by the generic ILP-based path computation mechanism explained in [15].

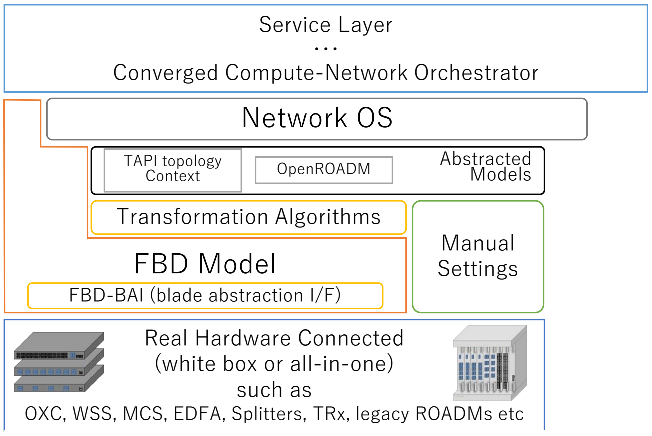
Service Layer
…
Converged Compute-Network Orchestrator
Network OS
TAPI topology Context
OpenROADM
Abstracted Models
Transformation Algorithms
Manual Settings
FBD Model
FBD-BAI (blade abstraction I/F)
Real Hardware Connected
(white box or all-in-one)
such as
OXC, WSS, MCS, EDFA, Splitters, TRx, legacy ROADMs etc

Fig. 3. Hierarchical architecture of the models.

5. Conclusion

The optical layer switching with high and guaranteed bandwidth, low latency, high energy efficiency, and high security is expected to substantially empower the future digital infrastructure in the post-Moore’s law era, in which compute and optical network will be converged to form the so-called computing continuum. It was addressed, however, that a real challenge lied in how to “digitalize” the optical layer in order to converge with the computing layer. To handle various topologies of the optical layer generally and automatically, the functional block-based disaggregation (FBD) model was discussed as the key. The FBD model substantiates a component-level automated management mechanism, and thereby will serve as an important platform of fully disaggregated computing and/or optical networks, underpinning the future computing continuum.

Acknowledgement: This work was partly supported by JSPS KAKENHI JP19H02164.

[1] See e.g., A. J. Ferrer et al., arXiv:2103.06026.
[2] K. Shi, et al., OFC2021, Tu4A.1.
[3] S. Namiki, TP1: Plenary, ISUPT2019; Plenary, APC2017, JM1A.2.; OFC2018 WS S1B, Team A Presentation.
[4] N. Parsons, et al., ECOC 2016, W.2.F.1.
[5] T. J. Seok, et al., Optica 6, 490 (2019).
[6] K. Suzuki, et al., JLT 37, 116 (2019).
[7] R. Matsumoto, et al., OFC2021, Tu6A.2.
[8] K. Suzuki et al., Invited talk in this OFC (OFC2021).
[9] R. Stone, et al., ECOC 2020, Mo2C-1.
[10] K. Sato, JLT 36, 1411 (2018), and references therein.
[11] See e.g., https://iowngf.org/
[12] G. Zervas, et al., OFC2017, W3D.4.
[13] V. Mishra, et al., JOCN 13, 126 (2021).
[14] J. Kani, et al., JOCN 12, D48 (2020).
[15] K. Ishii, S. Namiki, OFC 2021, F1C.1, and references therein.
[16] K. Ishii, et al., JLT, 39, 821 (2021).
[17] K. Ishii, et al., ECOC2020, We1K-2