A Thesis for the Degree of Ph.D. in Engineering

Optical Access/Intra Data Center Network with High Energy Efficiency and Reliability

July 2017

Graduate School of Science and Technology
Keio University

Masahiro HAYASHITANI

Contents

Lists of Figures

Lists of Tables

Summary

Traffic to and within data centers have been increasing rapidly because of cloud service adoption. In the near future, traffic to the data center will dominate the traffic in access networks, and flows will be largest within the data center network. Therefore, it is assumed that the power consumed by data centers will increase and the repercussions of failure will worsen as the traffic increases. The network connection to the data center (defined as access data center network hereinafter) and the network within the data center (defined as intra data center network hereinafter), need network architectures that offer power savings and survivability for mission critical services in order to well handle the increasing traffic.

The present access data center network basically consists of PON (Passive Optical Network). However, PON is weak in terms of scalability and resource utilization because splitters, which simply broadcast packets, are used between OLT (Optical Line Terminal) and ONU (Optical Network Unit). Therefore, a new network architecture of the aggregation type is needed as traffic to the data center and the number of users will increase rapidly. This dissertation focuses first on the access data center network that offers high energy efficiency with significant scalability.

Current intra data center networks generally use electrical switches and routers. Unfortunately, the power consumption of these devices increases rapidly when the traffic on the data center network increases. Therefore, the intra data center network needs optical technologies that can realize power savings and larger network scale are needed.

This dissertation also focuses on the intra data center network that offers power savings with high data center performance and high reliability. To achieve these goals, this dissertation proposes optical access/intra data center networks with high energy efficiency and reliability. For the optical access data center network, the proposed scheme realizes high energy efficiency and scalability by minimizing the number of active OLT to suit user traffic and maximizing unnecessary switching times and ONU sleep by accelerated slot assignment. For the optical intra data center network, the proposed scheme realizes power savings by minimizing working buffers and servers by assigning VMs (Virtual Machine) based on VM groups thresholding for high data center performance. Furthermore, the proposed scheme realizes high reliability of medium-priority services by rapidly suspending low-priority services upon the bidirectional failure notification issued by the failure detecting node.

This dissertation is organized as follows. Chapter 1 describes the background of the dissertation and clarifies the challenges in data center networks and the position of the dissertation. Chapter 2 illustrates technologies of access/intra data center networks and schemes for power saving and survivability. Chapter 3 examines the goal of high energy efficiency in the optical access data center network. The scheme proposed therein offers a 47% reduction in power consumption compared to PON. Chapter 4 proposes a high energy efficiency scheme for the optical intra data center network. The proposed scheme can reduce network power consumption by 40% and server power consumption by 59% compared to the conventional optical data center. Chapter 5 proposes a high reliability scheme for the optical intra data center network. The proposed scheme can realize constant recovery time regardless of the low-priority traffic condition. Chapter 6 draws this dissertation to its conclusion with a useful summary of the advances raised herein.

Chapter 1

Introduction

Chapter 1 describes the background of the dissertation, and clarifies the requirements placed on access/intra data center networks and the position of the dissertation.

1.1 Background

Traffic to data centers and that within data centers have been increasing rapidly because of cloud service development. In the near future, most access network traffic will that generated by the data center, and 70% of flows which are generated in the data center go to servers within the data center [1-1]. Figure 1.1 shows IP traffic growth triggered by the data center. Both the access data center network and the intra data center network will need network architectures that offer greatly improved power efficiency and survivability if the mission critical services are to be maintained in the face of the relentless increase in traffic.

1.2 Access Data Center

The present access data center network is mainly implemented as a PON (Passive Optical Network). However, PON has a problem with both scalability and resource utilization because splitters, which simply broadcast packets, are used between OLT (Optical Line Terminal) and ONU (Optical Network Unit). Therefore, an aggregation type network architecture is needed as traffic to the data center and the number of users are increasing

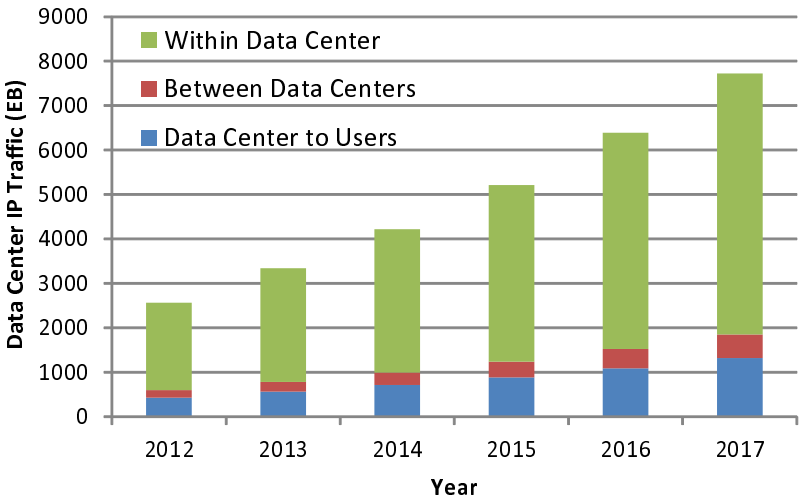

Figure 1.1. IP traffic growth triggered by data center.

rapidly. In the near future, traffic to the data center will be most traffic in the access networks. This dissertation focuses enhancing the energy efficiency and scalability of the access data center network.

In PON, an OLT (Optical Line Terminal) is deployed in a central office, while one ONU (Optical Network Unit) is deployed in each subscriber. Between the OLT and the ONUs, there are splitters to connect all ONUs with the OLT. If the central office must accommodate more subscribers, more OLTs are needed due to splitter loss. AON (Active Optical Network) was proposed for accommodating more subscribers and providing flexible resource assignment [1-2, 3]. AON replaces the splitters with optical switches, and optical wavelength resources are divided into slots that lie on a regular grid. AON transfers content by assigning slots, and a network user can, in some slots, access large bandwidth. However, AON uses many optical switches, and the switches are active in every slot. Thus, the power consumption of a large scalable AON is a problem.

Therefore, an energy-efficient optical slot switching architecture is proposed for the access data center network. The energy-efficient switching network minimizes working

OLTs to suit user traffic and maximizes unnecessary switching times and ONU sleep by accelerated slot assignment.

1.3 Intra Data Center

Current intra data center networks use electrical switches and routers. Unfortunately, the power consumption of these devices increases rapidly as the traffic within the data center network increases. Therefore, optical technologies which can realize power savings and larger scale networks are needed if the data center is to handle the increase in traffic. This dissertation focuses on raising the data center performance and reliability of the intra data center network while realizing power savings multiple services including mission critical services.

Annual global data center IP traffic is expected to reach 7.7 zettabytes by the end of 2017 given the compound annual growth rate of 31% from 2012 to 2017. Current data center networks suffer from excessive power consumption due to the sheer number of electrical switches in their core parts. The power consumption of Cisco Nexus 7000, commonly used in data center networks, is 2.5 W/Gbps [1-4]. Electrical routers are the major power consumers in intra data centers, and the consumption rises with the size of the intra data center network. Low latency is crucial for intra data center networks because big data analyses, based upon real-time CEP (Complex Event Processing), demand the exchange of huge quantities of data.

The power consumption of an optical switch is a just a few dozen mW/Gbps. Therefore, deploying more optical switches will lower the power consumption of the network [1-5]. Some of the architectures proposed for the data center network use optical switch technologies [1-6, 7, 8, 9, 10, 11]. Of interest is the HOPR-based data center network [1-12]. HOPR uses CMOS-based electrical buffers in order to avoid packet contention and to counter the optical power attenuation created by long hops as the buffers regenerate the

packets.

In addition, VM (Virtual Machine) migration schemes have been proposed for power saving in the intra data center network [1-13, 14]. The scheme turns off unused links and switches by selecting which switches must be turned on to achieve the desired data center network performance and fault tolerance [1-13]. However, the scheme turns off links and switches by monitoring just the traffic, and does not reduce power consumption of the data center network by considering VM situation in the servers. Another scheme turns off switches by considering VM situation in servers [1-14]. However, it assumes that only one VM in a server is active. A real data center will have multiple VMs in a server for more efficient server utilization. These schemes use the fat-tree topology which needs switches with many ports, and thus are suitable for networks based on HOPR which have fewer ports.

Therefore, an energy efficient intra data center network is proposed here. The proposed network minimizes working buffers and servers to VM assignment based on VM group thresholding for high data center performance.

In the intra data center network, multiple-class traffic is considered where each class may have different traffic demands. In general, 1+1 protection is best for high-priority traffic; it achieves high-speed protection as backup paths are always available for data transmission, and destination nodes simply switch to the backup paths in the case of a failure. Mission critical and broadcasting services, for example, are taken to be high-priority class. In general, 1:1 protection [1-15, 16, 17, 18, 19] is best for middle priority traffic. In 1:1 protection, the idle backup paths are used to carry low-priority traffic [1-20]. The 1:1 protection scheme can thus achieve a good balance between high-speed transmission and efficient path utilization. Enterprise services, for example, are classed as middle-priority traffic. Low-priority traffic is not protected and is dropped in favor of backup path traffic in case of failure. The Internet is classed as low-priority traffic. If a

network fault occurs, the low-priority traffic is suspended, and the middle-priority traffic is switched to the backup paths. A scheme suspending low-priority traffic enables the network to carry low-priority traffic between nodes [1-21, 22], but the process for requesting acknowledgment and suspension makes it difficult to provide high-speed protection.

Therefore, a reliable intra data center network is proposed. The proposed intra data center network suspends low-priority service rapidly upon failure notification by a failure detecting node.

1.4 Target of Dissertation

Figure 1.2 shows the conventional access/intra data center networks. It is assumed that the power consumption of data center networks increases and effects of failure are expanded as the traffic increases. Both the access data center network and the intra data center network need a network architecture that offers power savings and survivability of mission critical services. Figure 1.3 shows proposed access/intra data center networks. In the optical access data center network, the proposed scheme realizes high energy efficiency with large scalability by minimizing the number of active OLT while supporting user traffic and maximizing switching latency and ONU sleep by accelerated slot assignment. In the optical intra data center network, the proposed scheme realizes power savings by minimizing working buffers and servers by assigning VMs based on VM group thresholding while achieving high data center performance. Furthermore, the proposed scheme realizes high reliability of middle-priority services by rapidly suspending low-priority services upon failure notification. Table 1.1 shows an outline of the proposals in this dissertation and the corresponding chapters. In this dissertation, the scale of intra data center network is assumed to be the enterprise-level data center network and so has several thousand servers. The access data center network supports not just for the data center, but most traffic of access data center network will go to the data center.

Table 1.1. Outline of proposal in this dissertation.

Chapter 3	Purpose	High energy efficiency in access data center network.
	Problem	More energy efficiency with larger scalability.
	Proposal	Minimize working OLT according to user traffic. Maximize unnecessary switching times and ONU sleep.
	Achievement	Reduce network power consumption by 47% as compared to PON.
Chapter 4	Purpose	High energy efficiency in intra data center network.
	Problem	More energy efficiency with higher data center performance.
	Proposal	Minimize working buffers and servers by controlling VM assignment based under high DC performance.
	Achievement	Reduce network power consumption by 40% and server power consumption 59% with high data center performance as compared to conventional HOPR based data center.
Chapter 5	Purpose	High reliability in intra data center network.
	Problem	High reliability of middle-priority service.
	Proposal	Suspend low-priority traffic rapidly by sending failure notification on bidirectional sides.
	Achievement	Constant total failure-recovery time of middle priority service even if low priority service is offered.

Figure 1.4 shows a future image of data center migration. In inter data center networks, data center migration will be needed for further energy and resource efficiencies. This dissertation focuses on the energy efficiency and reliability of the access and intra data center networks.

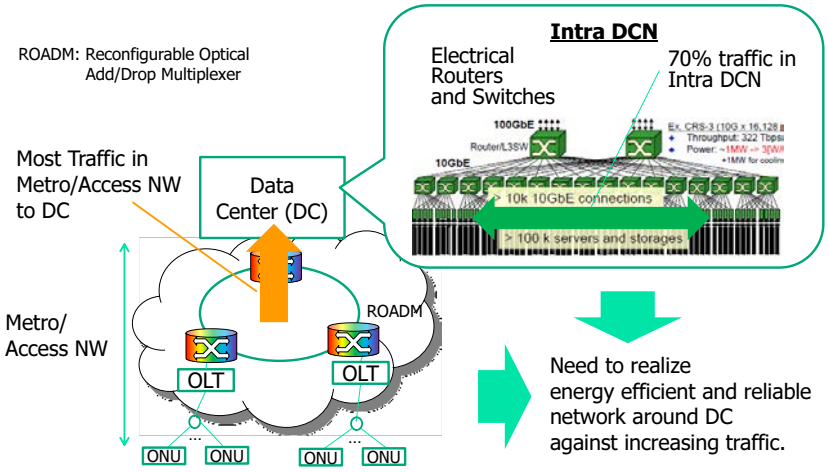

Figure 1.2. Conventional access/intra data center network.

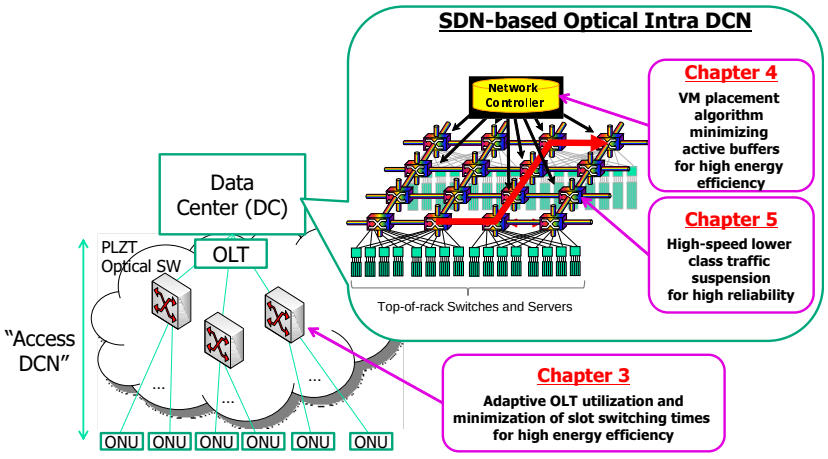

Figure 1.3. Proposed access/intra data center network.

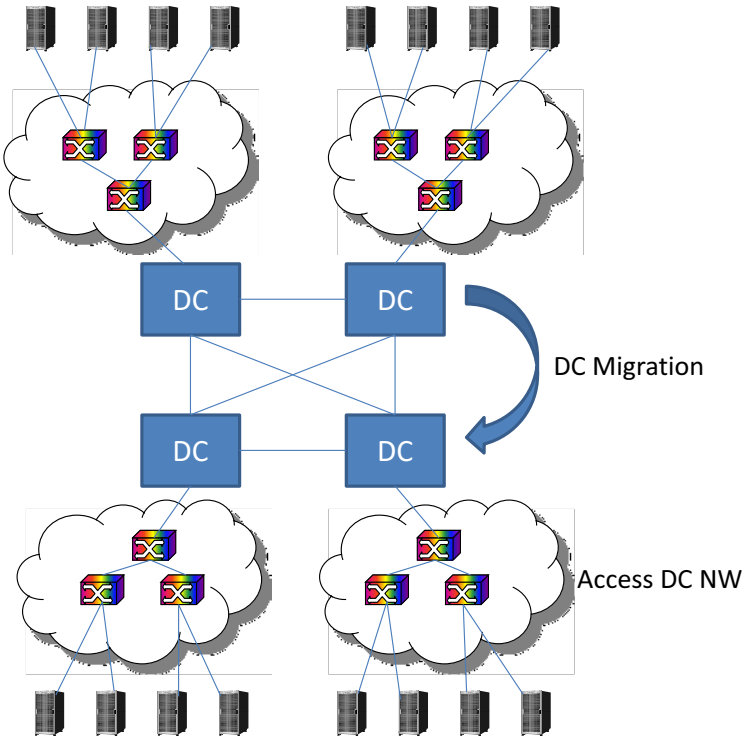

Figure 1.4. Future image of data center migration.

References

[1-1] D. Kilper, K. Bergman, V. WS Chan, I. Monga, G. Porter, and K. Rauschenbach, “Optical networks come of age,” Optics and Photonics News, vol.25, no.9, pp.50-57, Sep. 2014.

[1-2] T. Nomura, H. Ueda, T. Tsuboi, and H. Kasai, “Development of New Optical Access Network System Based on Optical Packet Switches,” in Proc. ECOC (European Conference Optical Communication), no.4.4.3, Sept. 2007.

[1-3] T. Nomura, H. Ueda, T. Tsuboi, H. Kurokawa, and H. Kasai, “Novel Optical Packet Switched Access Network Architecture,” Proc. OFC (Optical Fiber Communication), no.OTuJ6, Anaheim, USA, Mar. 2006. .

[1-4] Cisco Nexus 7000 Series,
http://www.cisco.com/c/en/us/td/docs/switches/datacenter/hw/nexus7000/installation/guide/n7k_hig_book.html

[1-5] P. K. Pepeljugoski, J. A. Kash, F. Doany, D. M. Kuchta, L. Schares, C. Schow, M. Taubenblatt, B. J. Offrein, and A. Benner, “Low Power and High Density Optical Interconnects for Future Supercomputers,” in Proc. OFC, no.OThX2, San Diego, CA, USA, 2010.

[1-6] C. Kachris, K. Kanonakis, and I. Tomkos, “Optical interconnection networks in data centers: Recent trends and future challenges,” IEEE Commun. Mag., vol.51, no.9, pp.39-45, Sept. 2013.

[1-7] D. T. Neilson, “Photonics for switching and routing,” IEEE J. Sel. Topics Quantum Electron., vol.12, no.4, pp.669-678, July/Aug. 2006.

[1-8] W. Zhang, H. Wang, and K. Bergman, “Next-generation optically-interconnected high-performance data centers,” IEEE/OSA J. Lightw. Technol., vol.30, no.24, pp.3836-3844, Dec. 2012.

[1-9] H. Mehrvar, H. Ma, X. Yang, Y. Wang, S. Li, A. Graves, D. Wang, H. Y. Fu, D. Geng, D. Goodwill, and E. Bernier, “Photonic switching of native ethernet frames for data centers,” in Proc. Photonics Switching, no.JT5C.2, San Diego, CA, USA, 2014.

[1-10] Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella, and S. J. B. Yoo, “LIONS: An AWGR-based low latency optical switch for high-performance computing and data centers,” IEEE J. Sel. Topics Quantum Electron., vol.19, no.2, article 3600409, Mar./Apr. 2013.

[1-11] J. Gripp, J. E. Simsarian. J. D. LeGrange, P. Bernasconi, and D. T. Neilson, “Photonics terabit routers: The IRIS project,” in Proc. OFC, no.OThP3, San Diego, CA. USA, 2010.

[1-12] K. Kitayama, Y. Huang, Y. Yoshida, R. Takahashi, T. Segawa, S. Ibrahim, T. Nakahara, Y. Suzaki, M. Hayashitani, Y. Hasegawa, Y. Mizukoshi, and A. Hiramatsu, “Torus-Topology Data Center Network Based on Optical Packet/Agile Circuit Switching with Intelligent Flow Management,” IEEE/OSA J. Lightw. Technol., vol.33, no.5, pp.1063-1071, Mar. 2015.

[1-13] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, “ElasticTree: Saving Energy in Data Center Networks,” in USENIX NSDI, April 2010.

[1-14] V. Mann, P. Dutta, S. Kalyanaraman, and A. Kumar, “VMFlow: Leveraging VM Mobility to Reduce Network Power Costs in Data Centers,” NETWORKING 2011. Springer Berlin Heidelberg, 2011.

[1-15] T. Shiragaki, et al., “Network Resource Advantages of Bidirectional Wavelength-path Switched Ring,” IEEE Photon. Technol. Lett., vol.11, no.10, pp.1325-1327, Oct. 1999.

[1-16] D. Forbes, et al., “Optical Shared Protection Ring Performance,” in Proc. ECOC, vol.2, pp.52-53, Nice, France, Sep. 1999.

[1-17] D. S. Levy, et al., “Optical Layer Shared Protection Using an IP-based Optical Control Network,” in Proc. OFC, no.TuO8, Anaheim, CA, Mar. 2001.

[1-18] M. J. Li, et al., “Transparent Optical Protection Ring Architecture and Applications,” IEEE/OSA J. Lightwave Technol., vol.23, no.10, Oct. 2005.

[1-19] S. Kim, et al., “Rapid and Efficient Protection for All-optical WDM Mesh Networks,” IEEE J. Sel. Areas Commun., vol.25, no.9, Dec. 2007.

[1-20] M. J. Li, et al., “Design and Experiment of Transparent Four-fiber Optical Channel Shared Protection Ring,” in Proc. NFOEC (National Fiber Optic Engineers Conference), pp.2018-2025, Dallas, TX, 2002.

[1-21] S. Seno, et al., “Optical Path Protection with Fast Extra Path Preemption,” IEICE Trans. Commun., vol.E89-B, no.11, pp.3032-3039, Nov. 2006.

[1-22] S. Seno, T. Fujii, M. Tanabe, E. Horiuchi, Y. Baba, and T. Ideguchi, “A Proposal and Evaluation of Multi-class Optical Path Protection Scheme for Reliable Computing,” High Performance Computing and Communications, vol.3726, pp.723-732, Sept. 2005.

Chapter 2

Access/Intra Data Center Network Technologies

Chapter 2 introduces technologies related to research in this dissertation. First, access data center network technologies that target power saving and survivability are illustrated. Second, intra data center network technologies for power saving and survivability are illustrated. Finally, the specific position of this dissertation is described.

2.1 Access Data Center Network

This section shows access data center network technologies and schemes for power saving and survivability.

2.1.1 Energy Efficiency

We start with raising the energy efficiency of access data center networks. Power saving approaches for PON, AON, and Optical Metro/Access Integrated Network are discussed.

Power Saving Approaches in PON

The architecture and terminologies of PON systems are shown in Figure 2.1. The key elements are OLT (Optical Line Terminal), usually located in a central office, ONU (Optical Network Unit), typically placed at the subscriber's premises, and the ODN (Optical

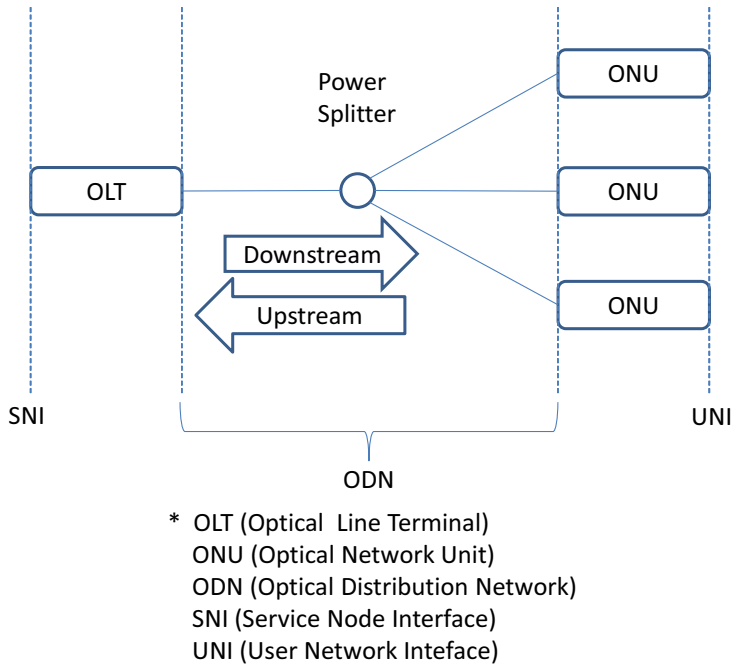

Figure 2.1. PON architecture and terminologies.

Distribution Network) made up of fiber and optical power splitters that forms the outside plant [2-1]. There two main types of power saving approaches for PON. One is ONU control [2-2], the other is OLT control [2-3].

ONU Control In the current PON, OLT physically broadcasts downstream traffic to all ONUs which remain active even if they are not the destination of any data. Figure 2.2 shows the potential energy savings if the ONUs stay active only in the receiving time slots (shaded) and switch to sleep mode in empty time slots (white). By exploiting the statistical multiplexing property of the downstream traffic, a significant reduction in energy consumption can be achieved. In order to enable sleep mode in the current PON, an ONU must wake up from the sleep mode and resynchronization with the network. Figure 2.3 illustrates the necessary phases for the ONU to wake up; the ONU must recover the OLT clock ($T_{recovery}$) and regain synchronization with the network (T_{sync}.). This incurs an overhead, $T_{overhead} = T_{recovery} + T_{sync}$., after which the ONU can listen for bandwidth

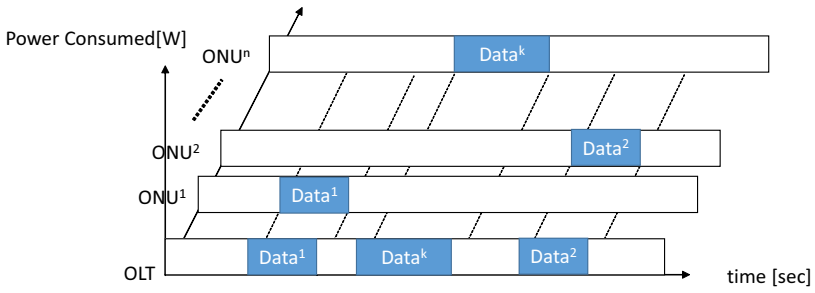

Figure 2.2. Potential energy savings.

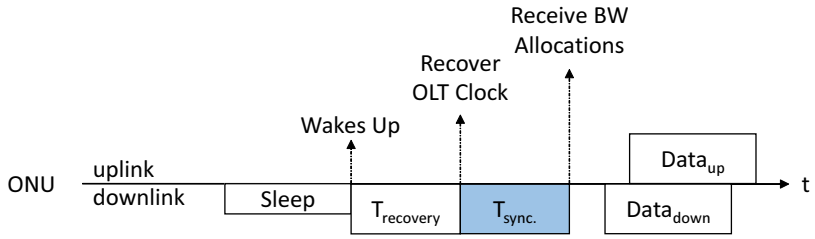

Figure 2.3. Necessary phase for ONU to wake up.

assignment and send or receive data. The synchronization time, T_{sync}., depends on the framing and the MAC protocol utilized. The following paragraphs assume the clock has been correctly recovered and focuses on how the ONU can regain synchronization with the network from sleep mode.

Figure 2.4 shows GPON (Gigabit PON) frames. In Fig. 2.4, GTC (GPON Transmission Convergence) specifies a GTC frame that lasts 125 μs. The beginning of each GTC frame holds PCB (Physical Control Block), which has 4 bytes for physical synchronization (Psync) and variable bytes for upstream bandwidth allocation (US BW map). An ONU waking from sleep mode would use Psync to gain synchronization so as to capture the first bytes of the GTC frame. If the ONU wakes up in the middle of an active GTC frame, it would receive a Psync header no later than 125 μs plus the 4 bytes (32 ns) and then commence synchronization to the GTC frame. Thus, the maximum synchronization delay is 125 μs. After synchronizing to the GTC frame, the ONU can resume sending or

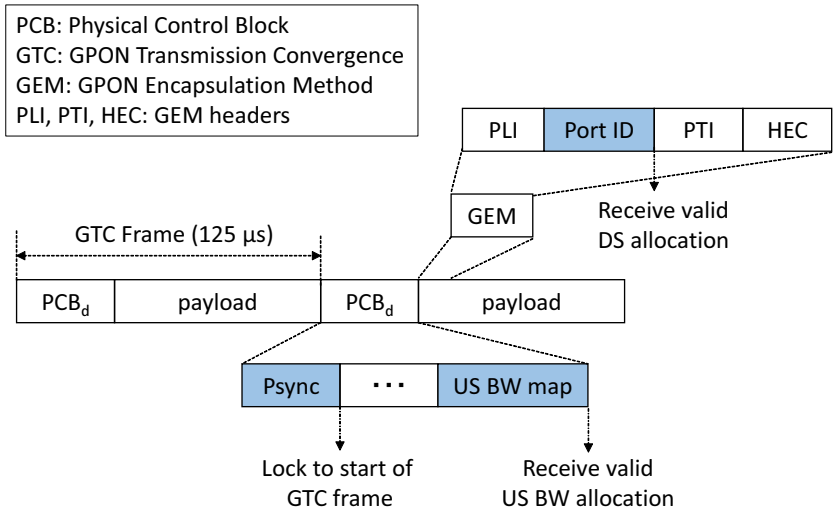

Figure 2.4. GPON frames.

receiving data if it receives an upstream or downstream bandwidth allocation. In GPON, this is done using either the US BW map field embedded at the end of the PCB or the Port ID field from GEM (GPON Encapsulation Method) frame in the ensuing GTC payload section.

Figure 2.5 shows the receiver architecture of current GPON ONUs. In this architecture, the received optical signal is first converted into an electrical signal by APD (Avalanche PhotoDiode). Two stage electrical amplifiers, including TIA (TransImpedance Amplifier) and LA (Limiting Amplifier), amplify the electrical signal before sending it to the CM-CDR (Continuous Mode-Clock and Data Recovery) circuit. CM-CDR forwards the recovered data and clock to the de-serializer (DMUX). DMUX then outputs parallelized data to a lower speed, back-end digital circuit for further processing. The back-end digital circuit consists primarily of memory and various digital data processing blocks. In this architecture, when the ONU enters sleep mode, the entire front-end analog circuit and part of the back-end digital circuit are turned off. Some parts of the back-end digital circuit, such as the clock and volatile memory, cannot be turned off and thus the ONU still consumes power during the sleep mode. The ONU uses CM-CDR to recover the OLT clock

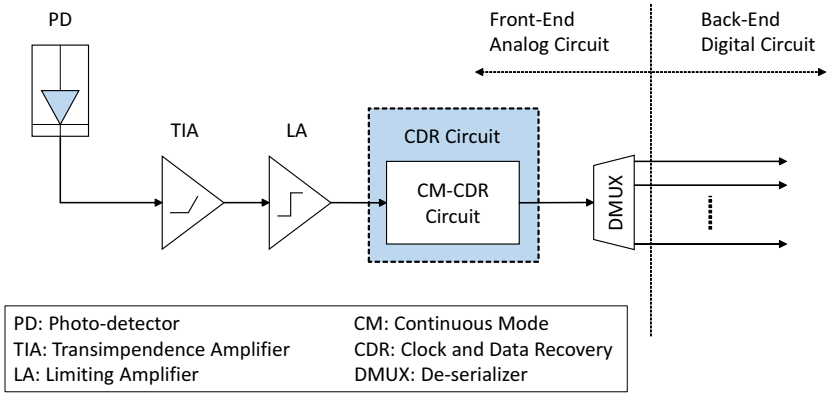

Figure 2.5. Receiver architecture of current GPON ONUs.

when it wakes from sleep mode. CM-CDR takes a significant amount of time, i.e. 2-5 ms, to recover the OLT clock. The long recovery time ($T_{recovery}$) significantly increases the wake-up overhead, which in turn degrades the energy savings possible.

Figure 2.6 shows a modified ONU architecture [2-2]. In this architecture, the entire frontend analog circuit is the same as in the ONU architecture of Fig. 2.5, except for the additional sleep mode control circuit placed before DMUX. Unlike the ONU architecture in Fig. 2.5, CM-CDR (as well as the APD photodiode, plus TIA and LA circuit) circuit is left on to maintain OLT clock accuracy whereas DMUX is switched off in sleep mode. The sleep control circuit selects the counter data path when the ONU enters sleep mode. The counter uses the recovered clock to time sleep mode. When the counter expires, it sends a WAKEUP signal to the controller and the controller reconnects the DMUX to the recovered clock and data to resume receiving mode.

OLT Control Existing proposals take advantage of the bursty nature of the traffic at the user side to better control the sleep mode and provided an adaptive line rate to efficiently reduce ONU power consumption. It is, however, challenging to apply “sleep” mode to the OLT to reduce its energy consumption for the following reasons. In PON, the OLT serves as the central access node and so controls the network resource access of the ONUs.

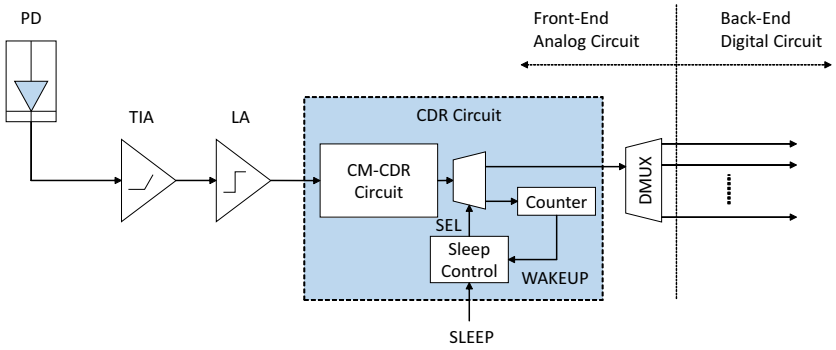

Figure 2.6. Modified ONU architecture.

Putting the OLT to sleep can easily result in service disruption of the ONUs communicating with the OLT. Thus, a more sophisticated scheme is needed to reduce OLT energy consumption without degrading services of end users. An energy-efficient OLT structure has been proposed [2-3]. The structure can activate its OLT line cards to suit the actual traffic arriving. To avoid service degradation during the process of powering on/off OLT line cards, proper devices are added to the legacy OLT chassis to ensure that all ONUs can communicate with active line cards.

In the central office, one OLT chassis typically comprises multiple OLT line cards, each of which communicates with a number of ONUs. In currently deployed GPON systems, one OLT line card usually communicates with either 16 or 32 ONUs. To avoid service disruptions of ONUs connected to the central office, all these OLT line cards are usually kept active all the time. To reduce OLT energy consumption of OLT, the main idea is to activate only as many OLT line cards in the OLT chassis as are needed to support the real-time incoming traffic.

C denotes the provisioned data rate of one OLT line card, L the total number of OLT line cards, N the number of ONUs connected to the OLT chassis, and $r_i(t)$ the arrival traffic rate of ONU i at time t. By activating all OLT line cards, the overall data rate accommodated by the OLT chassis equals $C \cdot L$. $C \cdot L$ is likely to be greater than the

real-time incoming traffic, i.e.,$\sum_{i=1}^{N} r_i(t)$. l denotes the smallest number of OLT line cards that can provision $\sum_{i=1}^{N} r_i(t)$ data rate. Then,

$$l = \lceil \sum_{i=1}^{N} r_i(t)/C \rceil.$$

The ultimate objective is to activate only l OLT line cards to serve all N ONUs at any given time t instead of activating all L line cards. However, deactivating an OLT line card may disrupt the service of ONUs using the card. To avoid service disruption, any OLT line card should be able to provision bandwidth to any ONUs connected to the OLT chassis. To address this issue, several modifications to the legacy OLT chassis have been proposed.

In order to dynamically configure the communications between OLT line cards and ONUs, one scheme places an optical switch in front of each OLT line card as shown in Figure 2.7 [2-3]. The function of the optical switch is to dynamically configure the connections between the OLT line cards and ONUs. When the network is heavily loaded, the switches are configured such that each PON branch communicates with one OLT line card. When the network is lightly loaded, the switches are configured that multiple PON branches communicate with one line card. The idle OLT line cards are powered off, thus reducing energy consumption.

It is assumed that the energy consumption of the optical switch is negligible. Compared to the scheme of keeping all L line cards active, the partial activation of $\lceil \sum_{i=1}^{N} r_i(t)/C \rceil$ line cards can achieve energy savings (relative) as large as

$$1 - \frac{\lceil \sum_{i=1}^{N} r_i(t)/C \rceil}{L}.$$

The average energy saving over time span T equals to

$$\frac{\int_{t=0}^{T} 1 - \lceil \sum_{i=1}^{N} r_i(t)/C \rceil / L}{T}.$$

Figure 2.8 (a)-(d) illustrates the configuration of switches for the case that one OLT chassis contains four OLT line cards. We define traffic load as $\sum_{i=1}^{N} r_i(t)/L \cdot C$, where $\sum_{i=1}^{N} r_i(t)$

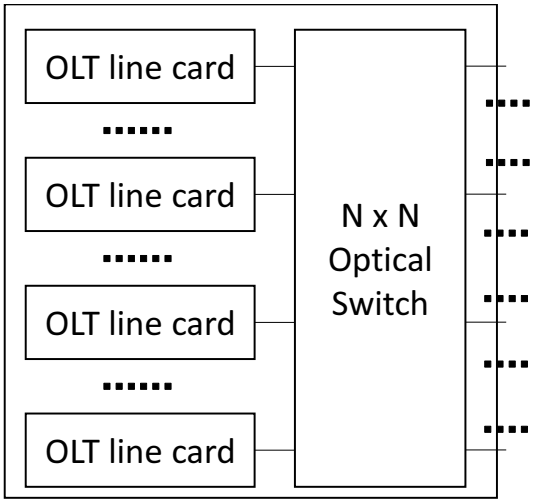

Figure 2.7. OLT with optical switch.

is the total arrival traffic rates of all ONUs and $L \cdot C$ is the capacity provisioned by all OLT line cards. By dynamically configuring switches, the number of active OLT line cards is reduced from four to x when the traffic load falls between $x/4$ and $(x + 1)/4$. Thus, a significant amount of power can be saved.

The power saving approaches in PON can reduce power consumption in the access data center network. However, PON has problems in scalability and resource utilization due to splitters.

Power Saving Approach in AON

Figure 2.9 shows the general architecture of AON. In AON, optical switches are used as relay points.

The GE-OSAN (Gigabit Ethernet-Optical Switched Access Network) architecture realizes higher security and longer spans than the conventional PON [2-4]. Figure 2.10 shows the basic configuration of GE-OSAN. It has a tree topology similar to PON. OSM (Optical Switch Module) switches optical signals frame by frame between OLT and ONUs so higher security than PON can be achieved. In addition, it offers longer transmission spans

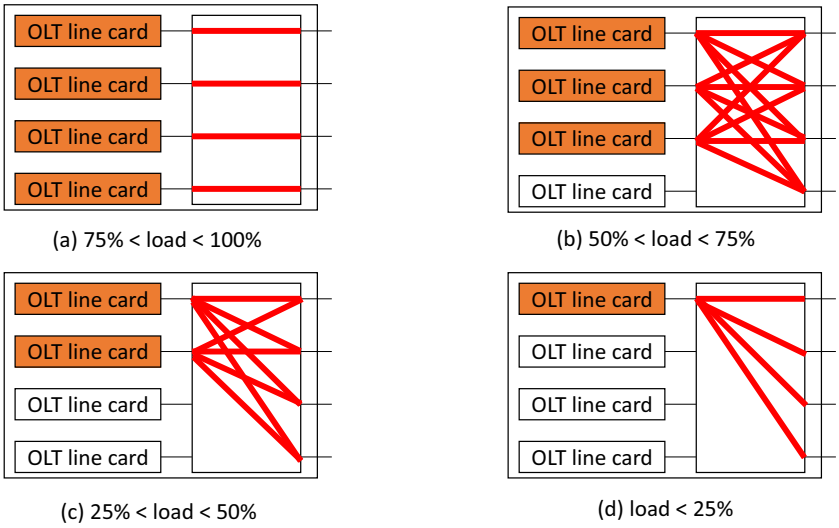

Figure 2.8. OLTs with configuration of optical switches.

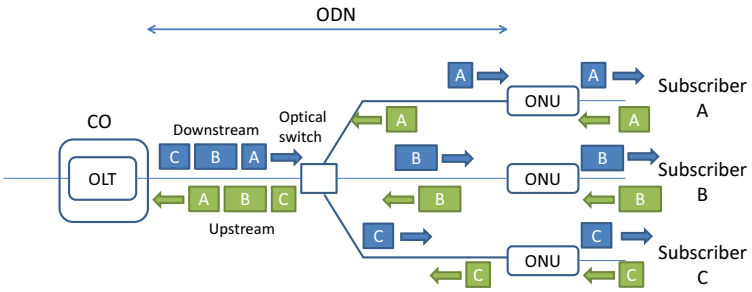

Figure 2.9. General architecture of active optical network.

since the optical switches have lower insertion loss than the optical splitters used in PON.

ActiON (Active Optical access Network) adopts PLZT high-speed optical switches [2-5, 6]. PLZT is a new waveguide material and consists of Pb, La, Zr, and Ti. Figure 2.11 shows the structure of the PLZT waveguide switch. The thickness and width of the PLZT channel waveguide are about 5 μm and 3 μm, respectively. The electrode structure is a simple capacitor type with electrode length of 3.5 mm. The switching time of a 1x2 PLZT switch element, less than 3 ns, is limited by the RC time constant of the capacitor type electrode geometry. PLZT is very attractive due to its dense integration, device

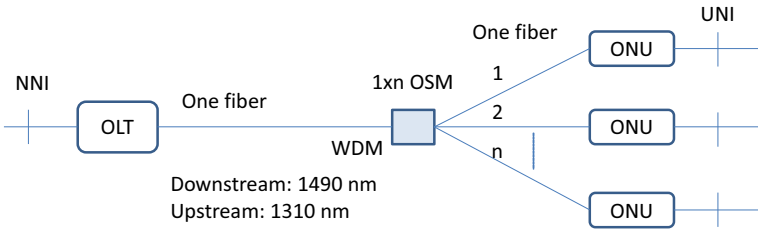

Figure 2.10. Basic configuration of GE-OSAN.

miniaturization, high-speed control, and low power consumption. Its efficient voltage-induced index change, that is, EO (Electro-Optic) effect, can enable miniaturization of electrodes, and low power consumption [2-7, 8]. There are several optical switches that offer high-speed switching including the LN (LiNbO3) switch based on EO materials, InGaAs (Indium Gallium Arsenide) switch, and SOA (Semiconductor Optical Amplifier). Major limitations of the LN are driving voltage versus device length, polarization dependence, and DC drift. InGaAs has disadvantages of insertion loss, polarization dependence and power consumption. The SOA has difficulty in terms of monolithic scalability and power consumption. On the other hand, PLZT has the advantages of polarization independence, switch size, and power consumption. The switching time of the PLZT optical switch is better than 10 ns.

In AON, the power saving approaches mainly target ONU or OLT. Recent studies do not consider the power consumption of the optical switches themselves.

Power Saving Approach in Optical Metro/Access Integrated Network

An optical metro/access integrated network should cover existing metropolitan and access networks by employing optical access technologies. OLTs are deployed on the boundaries of existing core and metropolitan area networks. The optical metro/access integrated network accommodates many more ONUs than the existing PON system. Activities (SARDANA and EλAN) towards the optical metro/access integrated network are

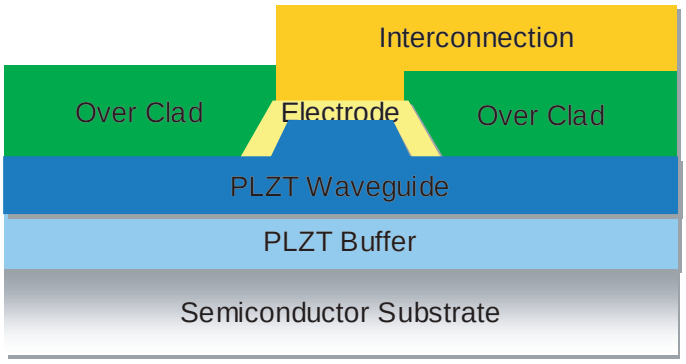

Figure 2.11. Structure of PLZT waveguide switch.

introduced. Different long-reach optical access networks have been studied with a view to their power efficiency. The SARDANA project [2-9] introduced the hybrid TDM/WDM (Wavelength Division Multiplexing) PON, in which remote amplification is utilized for reach extension [2-10].

Figure 2.12 shows the SARDANA architecture. The SARDANA network is an LR-PON (Long Reach-PON) that transparently merges TDM single-fiber passive tree sections with a main WDM double-fiber ring by means of passive RNs (Remote Nodes) [2-9]. The 100- km WDM ring transports 32 wavelengths for $1,000$ users, with a splitting ratio of 1:32 for a TDM tree and only 1 wavelength per TDM tree. Network protection and traffic balancing properties are provided by the ring configuration and the resilient design of the RNs, guaranteeing a continuous connection of each RN to the CO (Central Office) even in the case of fiber cut. The WDM ring is implemented by a double-fiber to avoid main Rayleigh backscattering impairments. Bidirectional propagation takes place in the single-fiber TDM trees.

The wavelength transparency of the ring and the ONUs, as well as the wavelength add/drop feature of the RNs enables sharing of the same network infrastructure by several operators, up to the RNs or even the ONU. It also allows users to select an operator by switching easily exchangeable filters at the ONU. In this way, SARDANA offers a

possible solution to implementing multi-operability in the physical layer; a set of downstream/upstream wavelength channels is simply allocated to each operator.

The TDM/WDM overlay in the SARDANA network eases the migration process from legacy PON solutions such as standard Gigabit Ethernet PONs to next-generation 10G versions. At the same time, SARDANA aims to reduce the complexity of the network infrastructure by implementing a fully passive (no power supply) plant from the CO to the end user premises with several RNs interconnecting the WDM ring to the TDM trees. This means a zero power consuming plant that focus on the hot spots of the CO and at the end user premises where low cost devices are utilized.

By ensuring intelligent OLT location with both capillary fiber optic diffusion and the implementation of a regional/sub-regional SARDANA infrastructure would, in principle, make it possible to reduce the number of COs from 820 to 11, thus attaining a tremendous reduction in the carbon footprint, as well as operator maintenance cost. The SARDANA external fiber plant can be fully passive thanks to the design of the RNs, in the sense that electrical power feeding is not needed, further reducing the carbon footprint and the OPEX (OPerating EXpenditure) of this PON.

Figure 2.13 shows the architecture of EλAN [2-11, 12]. EλAN integrates today's metropolitan area and the access networks of different network services. The transmission distance between P-OLTs and P-ONUs is extended compared to current access network systems. P-OLTs are located on the boundaries of today's core and metropolitan area networks.

The functions of each component are briefly described in the following paragraphs. The virtual layer-2 network transfers data frames of each service to/from appropriate P-OLTs.

P-OLTs are deployed at a number of aggregated COs. Each P-OLT configures L (Logical)-OLTs and provides the MAC and PHY functions required for each service such as frame processing, time synchronization, MPCP (Multi-Point Control Protocol), DBA

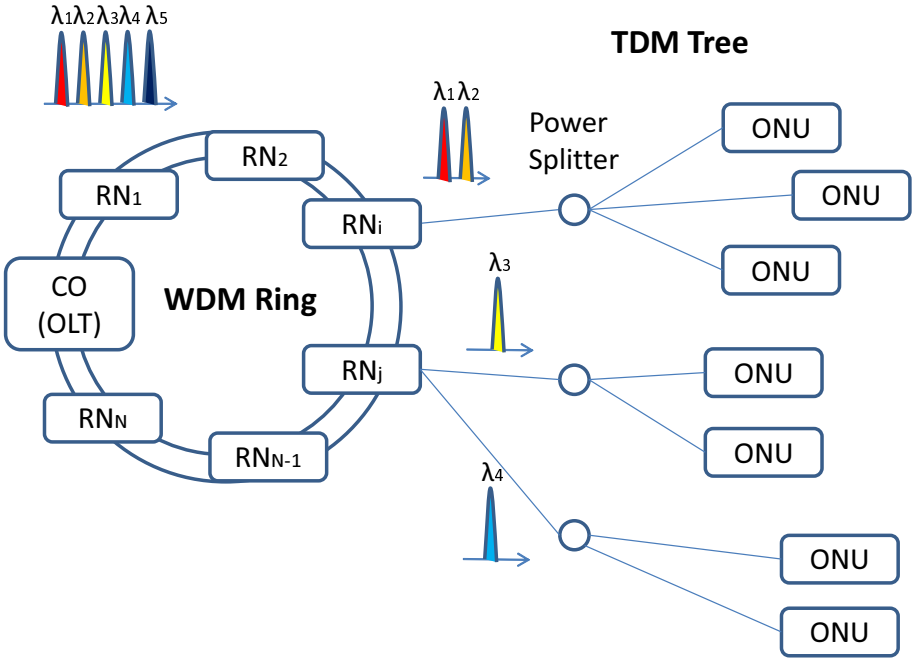

Figure 2.12. SARDANA architecture.

(Dynamic Bandwidth Allocation), rate adaptation, and FEC (Forward Error Correction) coding. Specific P-OLT implementation methods are still under study, but there is a high possibility that P-OLTs will use programmable logic devices to process data frames at high speed. Figure 2.14 shows one possible P-OLT implementation that uses FPGAs (Field Programmable Gate Arrays). A single P-OLT hosts a number of configuration boards to realize several L-OLTs. Each configuration board has two FPGAs for MAC and PHY, and the FPGAs can be reconfigured to provide the aforementioned functions. In the same way, P-ONUs are deployed at subscribers' premises and support network services by generating L-ONUs. It is assumed that P-ONU has a single transponder.

The NMS (Network Management System) performs coordinated control of the virtual layer-2 network, P-OLTs, and ODN. The NMS provisions a L-OLT configured on a P-OLT and an access path to satisfy a new connection request from a subscriber.

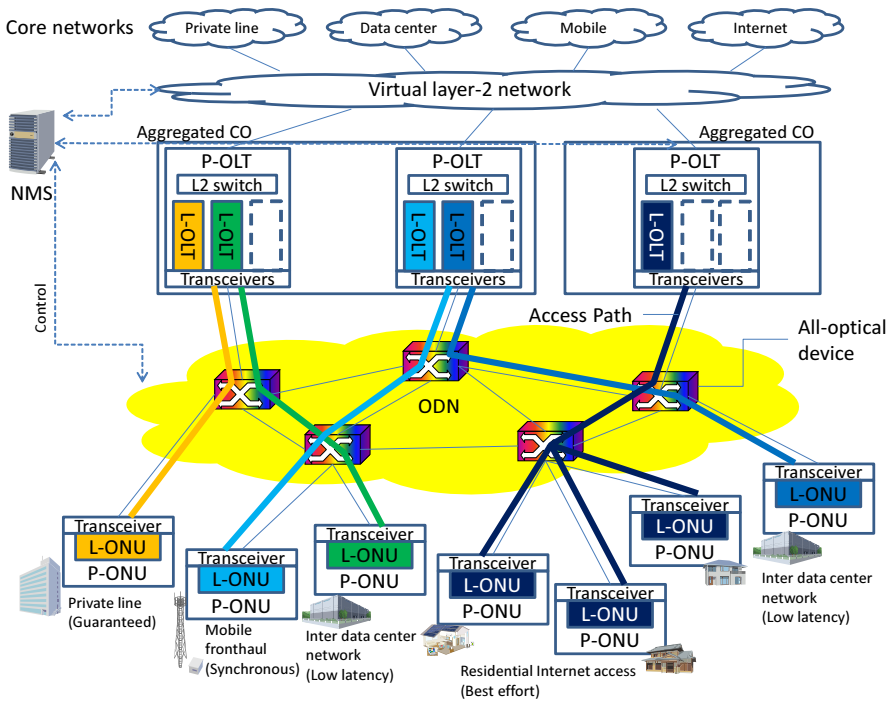

Figure 2.13. EλAN architecture.

Power Saving in EλAN EλAN introduced L-OLT migration to achieve more flexible utilization of network resources [2-11, 13]. Figure 2.15 shows an image of L-OLT migration. In the migration procedure, a L-OLT is replicated from one P-OLT to another P-OLT as directed by the NMS. The NMS also reconfigures the virtual layer-2 network and ODN at the same time so that data frames are transferred via the newly replicated L-OLT. An access path on the ODN is rerouted according to the replication of the L-OLT. Frequency slots that are allocated to the access path may be changed. As a result, subscribers continue to receive the same network service via the other P-OLT. The now quiescent L-OLT in the original P-OLT is released after L-OLT migration.

There are several advantages to implementing L-OLT migration in EλAN. The most notable merit is that the energy savings possible in the COs strengthen with the traffic fluctuation. For example, L-OLTs are aggregated to a limited number of P-OLTs when the total traffic amount is low. By putting unused P-OLTs to sleep, the network operator

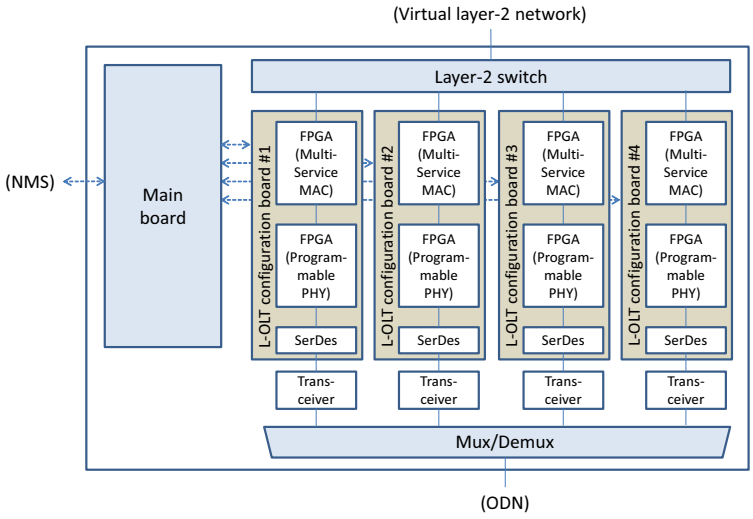

Figure 2.14. P-OLT implementation using FPGAs.

can reduce the total power consumption, which reduces OPEX. In [2-13], the energy reduction effect of L-OLT placement optimization was evaluated on the assumption that the arrival rate of connection requests fluctuates sinusoidally on a daily basis. The number of P-OLTs in sleep mode increased 16.7% on average and 35.8% at maximum, compared to the situation without L-OLT placement optimization. Another merit is the improvement in fault tolerance. When a failure occurs in a P-OLT or optical devices on an access path, P-ONUs under the P-OLT become unable to receive network services. If the control line between the NMS and P-OLTs is still active, the NMS searches for another P-OLT that can support an access path to these P-ONUs. The NMS then switches the L-OLTs from the faulty P-OLT to the other P-OLT, which recovers the connections.

In the Optical Metro/Access Integrated Network, the main power saving targets are the ONU and/or OLT. Optical switches have not been targeted for reducing the power consumption in recent research.

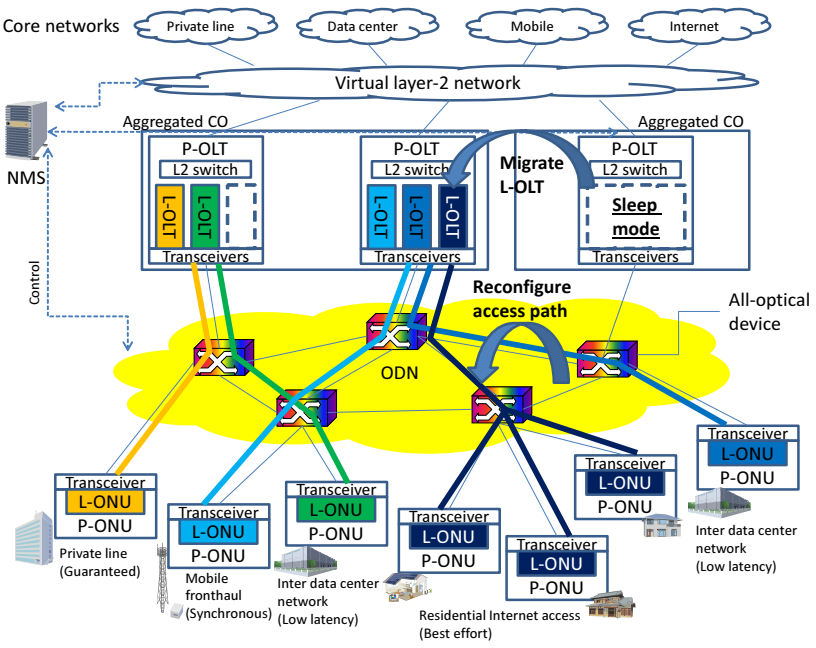

Figure 2.15. Image of L-OLT migration.

Comparison of Energy Efficient Method in Access Data Center Network

Table 2.1 shows a comparison of energy-efficient methods for the access data center network. PON [2-2, 3], AON [2-4, 5, 6], and Optical Metro/Access Integrated Network [2-11, 13] are also assessed. PON has problems in scalability and resource utilization. AON has problems with power consumption and switch control. Optical Metro/Access Integrated Network has a slight problem with power consumption, but its complex architecture is costly. The scheme proposed in Chapter 3 [2-37] achieves low power consumption while attaining excellent scalability and resource utilization by minimizing working OLTs and maximizing unnecessary switching times and ONU sleep.

2.1.2 Reliability

Access data center networks that offer high reliability are discussed. PON protection, AON protection, and Optical Metro/Access Integrated Network protection are explained.

Table 2.1. Comparison of energy-efficient methods in access data center network.

Scheme	Power Consumption	Pros	Cons
PON [2-2, 3]	Good	Cost	Scalability, Resource Utilization
AON [2-4, 5, 6]	Poor	Scalability, Resource Utilization	Switch Control
Optical Metro/Access Integrated Network [2-11, 13]	Fair	Scalability, Resource Utilization	Complex Architecture, Cost
Proposed in Chap. 3 [2-37]	Good	Scalability, Resource Utilization	-

Reliable Architecture in PON

A PON protection scheme is needed to ensure the system reliability required for business use, because PON is based on sharing a single fiber via an ODN [2-14]. This means that if a failure occurs in the feeder fiber, splitter, or OLT, all the users on that ODN branch experience connection problems. A variety of different protection architectures and schemes have been proposed for PON systems. Three redundant PON topologies for constructing practical systems are defined in ITU-T G.984.1. Types A, B, C, and D are shown in Figures 2.16 and 2.17. Type A protects only the feeder fiber and is useful for accidents or disasters that damage the fiber. Type B protects the OLT equipment with separate OLT blades, but no equipment redundancy is provided in the ONUs or feeder fibers. Type B offers automated switching and puts an additional PON port on the OLT. Type C provides two fully redundant links but is not cost-effective because it requires one backup ONU for every working ONU. Type D protection is similar to type C, but offers response flexibility for failures. For example, when the feeder fiber of the primary path is cut, type D allows subscribers to receive data from the spare OLT without switching ONUs, while subscribers have to switch to spare ONUs with type C.

Type B protection is a highly effective solution, because the failure rate of fiber is very low compared to that of OLTs. In Type B protection, the switching time is important. However, the connection between the OLT and ONUs is broken during protection switching, so the ONUs cannot synchronize their local clocks to the OLT due to downstream signal loss. As a result, the switching time is increased by the need to reestablish the ONU link.

Figure 2.18 shows a protection scheme named type H [2-15]. Type H uses the $1:N$ scheme to protects several PON trees, that is, a spare OLT is shared by N working OLTs. The spare OLT is connected to power splitters of the PON trees by a $1 \times N$ optical switch. Type H deals with a single point failure in OLTs or feeder fibers more economically than

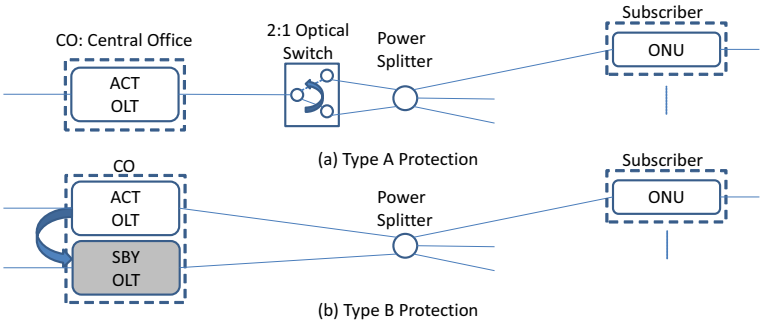

Figure 2.16. PON protection scheme (Type A and B).

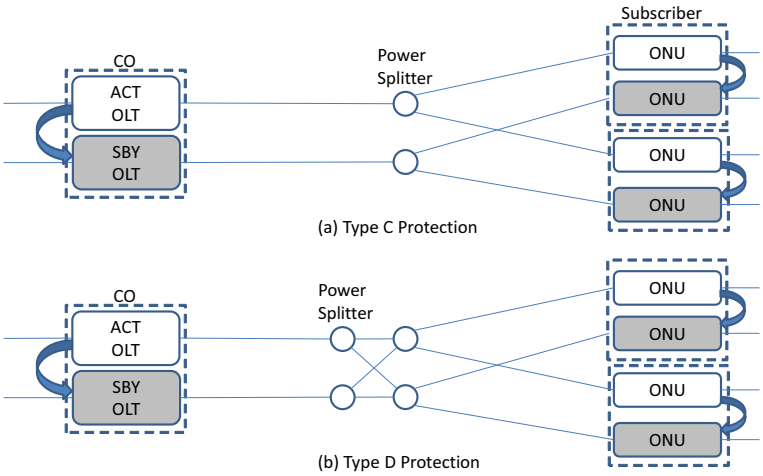

Figure 2.17. PON protection scheme (Type C and D).

type B.

The protection schemes in PON can realize high-speed protection in the simple architecture. However, PON has problems in resource utilization due to splitters, and cost due to additional devices.

Reliable Architecture in AON

A highly-energy-efficient network using an optical aggregation network and a service cloud has been proposed to greatly reduce the power consumption from today's Inter-

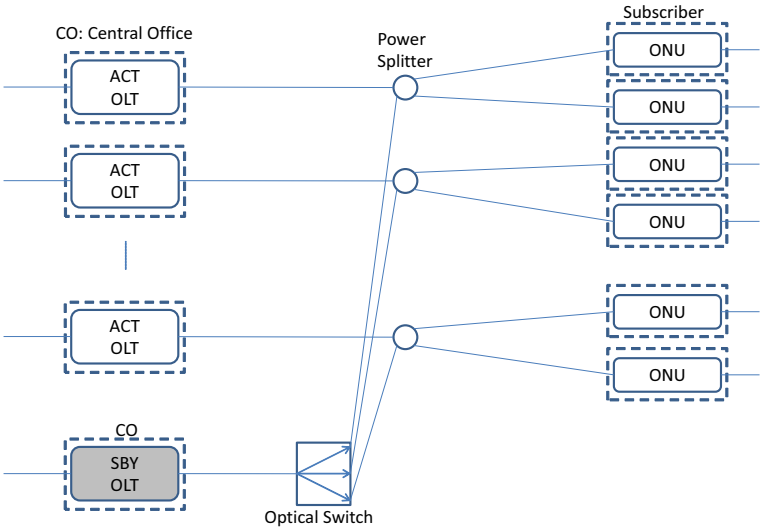

Figure 2.18. PON protection scheme (Type H).

net [2-16]. The optical aggregation network connects a solitary giant router to users by a logical tree topology. Figure 2.19 shows the architecture for the optical aggregation network. In this architecture, small-sized rings are connected in a tree-like topology by using 2x2 optical switches. Fig. 2.19 shows the proposed architecture, where the number of ring stages, *M*, is 2. With the introduction of the tree structure, the proposed architecture reduces the optical loss between the giant router and users, and thus minimizing the unavailability of the ring topology. In addition, two different routes that do not share any optical switch are set to every user. Therefore, all users can continue to communicate with the giant router when any of the 2x2 optical switches fails. In Fig. 2.19, the two different routes for user 2 are shown as dotted arrows. X-marked ports are unused because it is impossible to set two different routes to a user who connects to these ports. As all ports of the 2x2 optical switch are used (except X-marked ports), this architecture reduces *S* compared to the duplex tree topology.

The protection scheme in AON has a slight problem with cost because of optical switches.

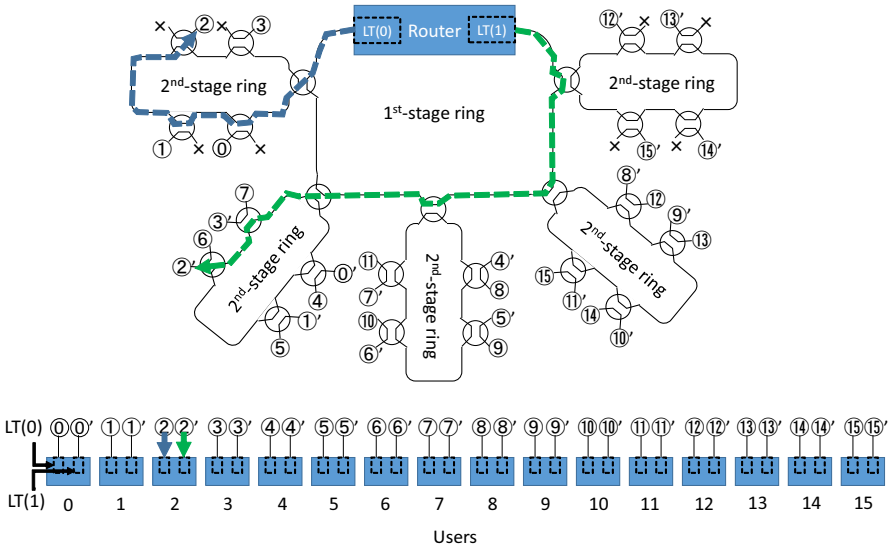

Figure 2.19. Architecture for optical aggregation network.

Reliable Architecture in Opitcal Metro/Access Integrated Network

SARDANA Apporach First tests of the SARDANA network have been performed using different configurations [2-17]. Figure 2.20 shows the protection scheme and results for a 105 km ring between Rome and Pomezia cities, 10G down- and 2.5G up-stream, with 2 RNs and 3 channels; the pump power is below 1.2 W at 1480nm. Sensitivities are -33 dBm and -36 dBm respectively. Protection against fiber cut was validated in down- and up-stream directions, with less than 1 dB penalty at rerouting. With the burst mode upstream operation, any gain transients at CO or RNs EDFAs can be mitigated by means of predistortion of data packets at the ONU, allowing a strong reduction, to 30%, of packet overshoot. On the other hand, because of the strong variation in optical traffic on the ring, it is useful to develop an automatic method based on a genetic algorithm to assess and minimize the impact of nonlinear crosstalk in the WDM ring channels; by optimizing channel frequencies and powers, the budget can be improved by 3-5 dB.

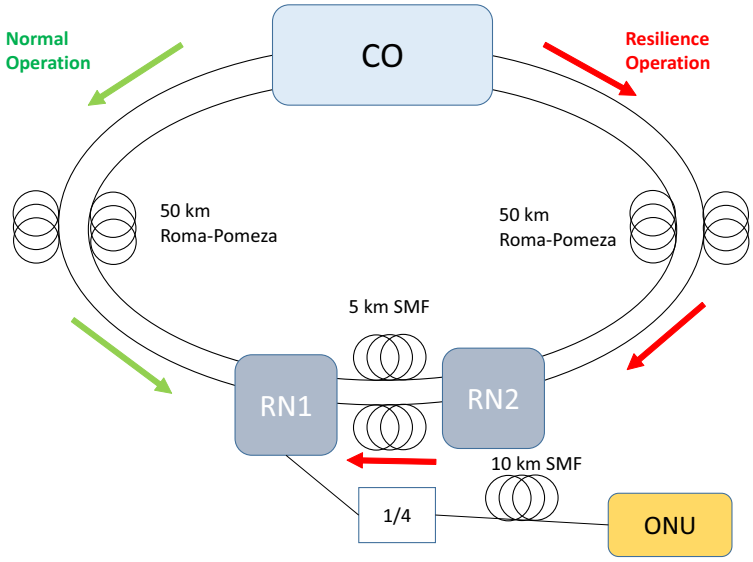

Figure 2.20. Protection scheme and test results.

EλAN Approach Figure 2.21 shows a conceptual model of the TDMA based ONU group recovery method [2-18]. It assumes that CO A is collapsed and its P-ONUs are accommodated by the P-OLT in CO B. First, L-ONUs are divided into ONU groups. One example of how to make ONU groups is as follows: select up to 256 geographically-close L-ONUs and group them. How to make efficient ONU groups is beyond the scope of this paper. A single L-OLT will support 2 or 4 ONU groups using virtualized 2 or 4 L-OLTs in TDMA manner. Fig. 2.21 describes ONU group #1 and #2.

The TDMA process in L-OLT is as follows. First, an L-OLT activates virtual L-OLT#1 for ONU group #1. Ranging and registration are done in this period and parameters for ONU group #1 are stored in virtual L-OLT#1. Second, the L-OLT activates virtual L-OLT#2 for ONU group #2. Ranging and registration are done in this period and parameters for ONU group #2 are stored in virtual L-OLT#2. The L-OLT periodically exchanges virtual L-OLT#1 and L-OLT#2 by data transfer using MPCP. L-OLT sends HOLDOVER message to accommodated L-ONUs before L-OLT switches ONU group. HOLDOVER message maintains the logical link even if OLT detects optical signal loss.

L-ONUs that receive HOLDOVER(start) message stop communication. L-OLT send HOLDOVER(end) message to L-ONUs when the L-ONUs are to resume communication. It is assumed that switching between the virtual L-OLTs takes less than 50 ms, and transfer operation time of a few seconds is assigned to each group. This means that in the first transfer operation time, virtual L-OLT#1 accommodates ONU group #1 without registration and transfers the data from buffer to ONU group #1 and receives data from ONU group #1. In the last MPCP operation, 0 byte is assigned to stop data transmission from ONU group #1. During this period, downstream data to ONU group #2 may be lost or buffered. After the 50 ms period for virtual L-OLT switching, L-OLT#2 becomes active.

The TDMA based L-OLT sharing method should take into consideration the overhead time of switching among ONU groups. Reducing the number of switching ONU groups can increase the communication efficiency, but to cover more subscribers, the number of groups should be as large as possible. Therefore, it is necessary to increase the service time to accommodate ONU groups. However, this will affect the service quality especially of voice communication, since long service times increase the communication interval.

The protection schemes in Optical Metro/Access Integrated Network offer better resource utilization and lower cost due to scalable architecture.

Comparison of Reliable Network Method in Access Data Center Network

Table 2.2 shows a comparison of three reliable network methods for the access data center network: PON protection [2-14], AON protection [2-16], and Optical Metro/Access Integrated Network Protection [2-17, 18]. PON protection has problems in resource utilization and cost. AON has a slight problem with cost. Optical Metro/Access Integrated Network offers better resource utilization, reliability, and lower cost.

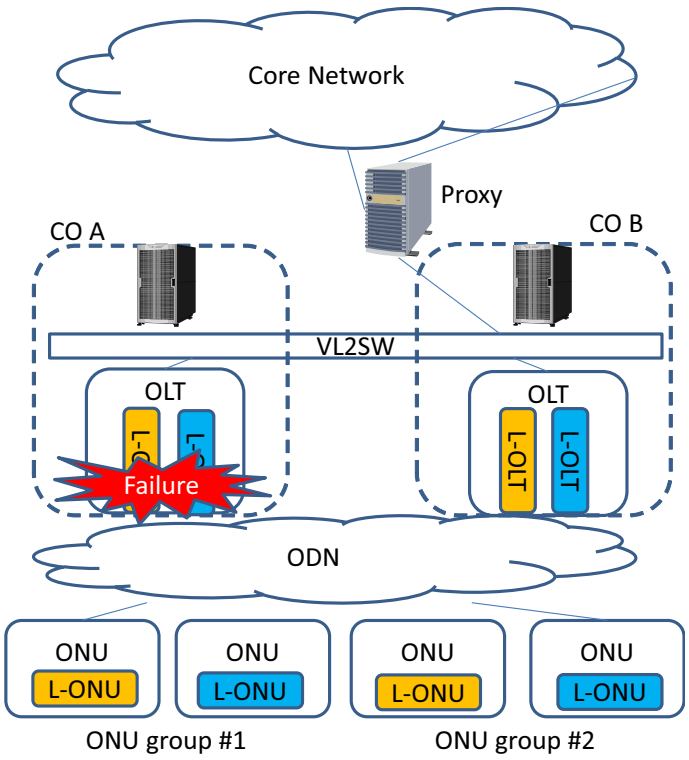

Figure 2.21. Conceptual model of TDMA based ONU group recovery method.

2.2 Intra Data Center Network

This section shows the intra data center network technologies and schemes for power saving and survivability.

2.2.1 Energy Efficiency

Intra data center networks realizing high energy efficiency are shown. The optical based data center network, VM migration, and SDN (Software-Defined Networking) based data center network for power saving are explained.

Table 2.2. Comparison of reliable network methods in access data center network.

Scheme	Reliability (Rapid Recovery)	Pros	Cons
PON Protection [2-14]	Good	Simple Architecture	Resource Utilization
AON Protection [2-16]	Good	Resource Utilization	Switch Cost
Optical Metro/Access Integrated Network Protection [2-17, 18]	Good	Resource Utilization	Complex Architecture

Optical based Data Center Network for Power Saving

Some groups have proposed architectures with optical switch technologies for reducing the power consumption of the data center network [2-19, 20, 21]. The data center network proposals replace some switches with OCS [2-19] or with OPS [2-20, 21].

Mehrvar et al. [2-19] replace some switches in the data center network with OCS, and some traffic is offloaded to OCS. The offloaded traffic can be guaranteed by the bandwidth assigned to OCS. Furthermore, the architecture reduces the power consumption of the data center network by using OCS, as some power-hungry switches are eliminated. However, OCS fails to support frequent changes in traffic when the traffic exceeds the reserved bandwidth. Therefore, no data center network is completely based on OCS, and the effectiveness of the power saving efficiency is limited.

Some groups proposed a data center network with OPS [2-20, 21]. In this data center network, all switches except ToR (Top of Rack) switches are replaced by OPS. OPS can well handle frequent changes in traffic and realize green data center networks because

OPS does not need to reserve bandwidth for packet transfer. In addition, OPS can attain packet switching times of nano-second order. However, current OPS devices are limited to just 8 input ports in [2-22]. Therefore, many switching stages are needed to manage the sheer number of servers in the data center. It is difficult to manage many servers in data center because the number of switch stages is limited due to optical power attenuation.

The HOPR-based data center network is now explained [2-23]. HOPR has both OPS and OCS, and uses them differently to serve traffic demands. In addition, HOPR has an electrical buffer based on CMOS in order to avoid packet contention; its regeneration function offsets the optical power attenuation created by the many stages.

Figure 2.22 shows a data center network based on HOPR with torus topology, similar to current supercomputers such as CRAY, Blue Gene, K Computer [2-24]. The data center network uses 100 Gbps links to connect each HOPR to neighboring HOPRs. Each HOPR is also connected to a group of ToR switches through 10 Gbps links. Compared to the conventional fat-tree data center network, the torus network offers better scalable even if no switch has many ports and high capacity. In addition, the torus network is robust because it can provide several alternative routes to the destination.

Figure 2.23 shows the HOPR architecture. From Fig. 2.23, Each ToR switch is connected to a shared electrical buffer. The buffer is used to achieve the following.

- Forward packets from ToR to HOPR
- Avoid packet contention in HOPR
- Avoid optical power attenuation

Packets can be transferred without the buffer if none of the above functions are needed. However, the power consumption in the buffer makes up a large fraction of that in HOPR. Therefore, it is necessary that the working buffers is reduced as much as possible in order to reduce power consumption in the intra data center network.

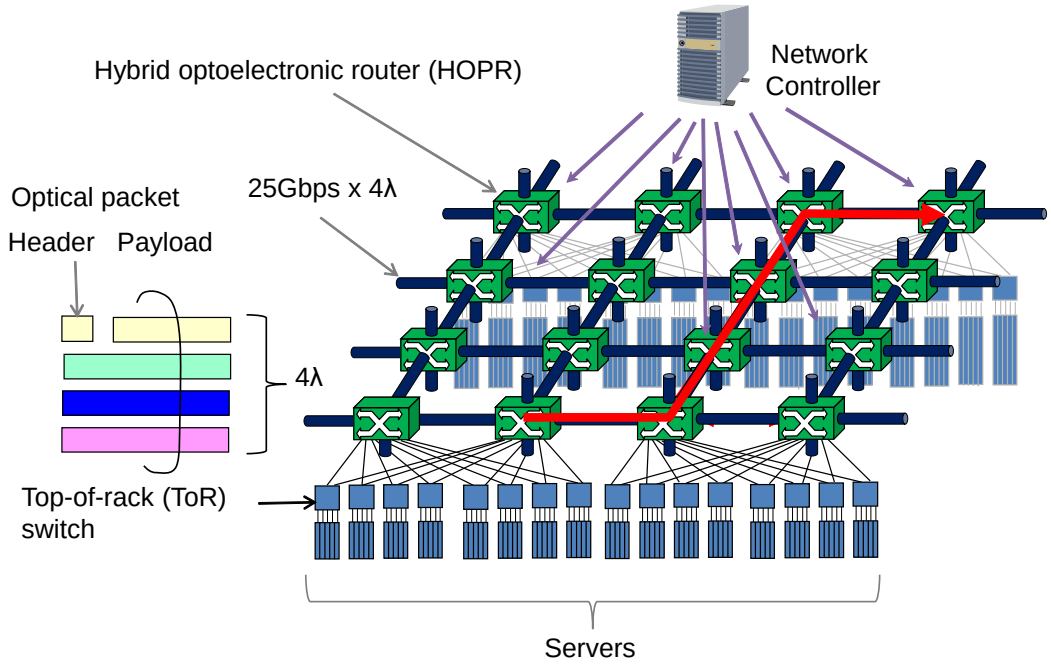

Figure 2.22. Data center network based on HOPR.

VM Migration for Power Saving

ElasticTree [2-25] and VMFlow [2-26] was proposed as power saving schemes in the intra data center network.

ElasticTree ElasticTree is a system that can dynamically adjust the energy consumption of a data center network. ElasticTree consists of three logical modules (optimizer, routing, and power control) as shown in Figure 2.24. The optimizer's role is to find the minimum power network subset that satisfies the current traffic demands. Its inputs are the topology, traffic matrix, a power model for each switch, and the desired fault tolerance properties (spare switches and spare capacity). The optimizer outputs a set of active components to both the power control and routing modules. Power control toggles the power states of ports, line cards, and entire switches, while routing chooses paths for all flows, then pushes routes into the network.

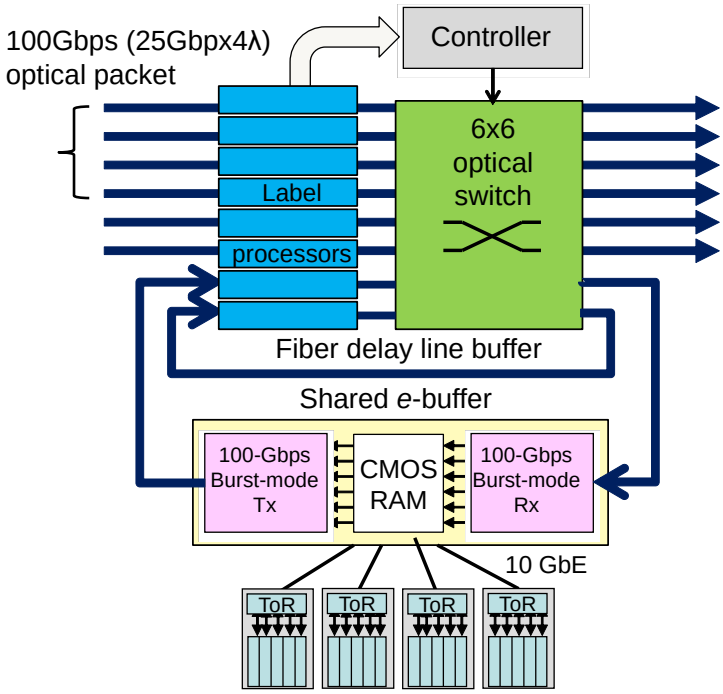

Figure 2.23. HOPR architecture.

ElasticTree turns off unused links and switches by selecting only those switches that must be turned on to satisfy the performance and fault tolerance demands of the data center network. Therefore, ElasticTree realizes the green data center network. However, ElasticTree turns off links and switches by monitoring only traffic, and does not consider the VM situation in servers.

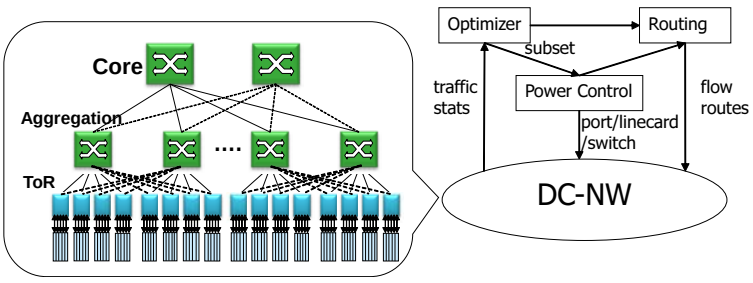

Figure 2.24. ElasticTree modules.

VMFlow ElasticTree also proposed a topology-aware approach that yields shorter computation times than greedy bin packing. However, the solution does not change the network power consumption. The VMFlow framework fundamentally differs from ElasticTree because it exploits the flexibility of VM placement available in the current data centers.

VMFlow realizes the green data center network by considering VM situation in servers, when deactivating switches. However, VMFlow assumes that each server has only one active VM. Furthermore, VMFlow mainly considers initial VM placement. In real data centers, each server will host multiple VMs for greater efficiency. In addition, VM migration is repeatedly performed for aggregating VMs on one side in in-service data center. Therefore, VMFlow does not consider operation in real data centers, and does not consider VM situation when realizing the green data center network by.

These schemes use the fat-tree topology which needs switches with many ports in the performance evaluations to date; they do not consider networks based on HOPR which does not have many ports.

SDN based Data Center Network for Power Saving

In the context of virtualized data centers, network programmability provides a modular interface that separates physical topologies from virtual topologies; each can be managed and evolved independently. Many architectural proposals rely on network programming technologies such as OpenFlow [2-27, 28]. SDN makes it possible to conceive a fundamentally different parallel and distributed computing engine that is deeply embedded in the intra data center network, realizing application-aware service provisioning. Also, SDN-enabled networks in the intra data center network can be adaptively provisioned to boost data throughput for specific applications in reserved time periods.

The ECDC (Energy efficient Data Center) approach is a smart mechanism for finding a

good balance between the various customer application requirements and the efficient use of the available resources in the data center [2-29]. It leverages monitoring information from machines as well as network devices and environmental data to create a coherent view of the current situation in the data center. This enables a single network operator to react to system-wide changes as soon as they happen. However, the monitoring data is also used by a smart control application to react in certain situations by redistributing VMs, traffic flows and VLANs as well as powering devices up and down. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Figure 2.25 shows the ECDC architecture in a simple data center scenario. The type A customer is a private home user who wants to use an entertainment service, e.g. video streaming, hosted by a service provider as a virtual infrastructure in rack B and C in the data center. The type B customer is a business user who uses a business application set up by his company in the data center, e.g. a virtual desktop infrastructure. As both types of customers have different demands and requirements, their traffic is kept in separate VLANs in the data center network so that they can be managed independently. The network itself is OpenFlow-enabled by OFC (OpenFlow Controller). The control connection for each network element is established via a physically isolated management network. The management station queries monitoring information on CPU-, network-, and memory-load, as well as power consumption from the connected devices. Armed with this information, the management station generates the appropriate network policy for the OFC, distributes virtual machines across the servers, and powers down unused devices in order to ensure the efficient utilization of all resources while maintaining a good

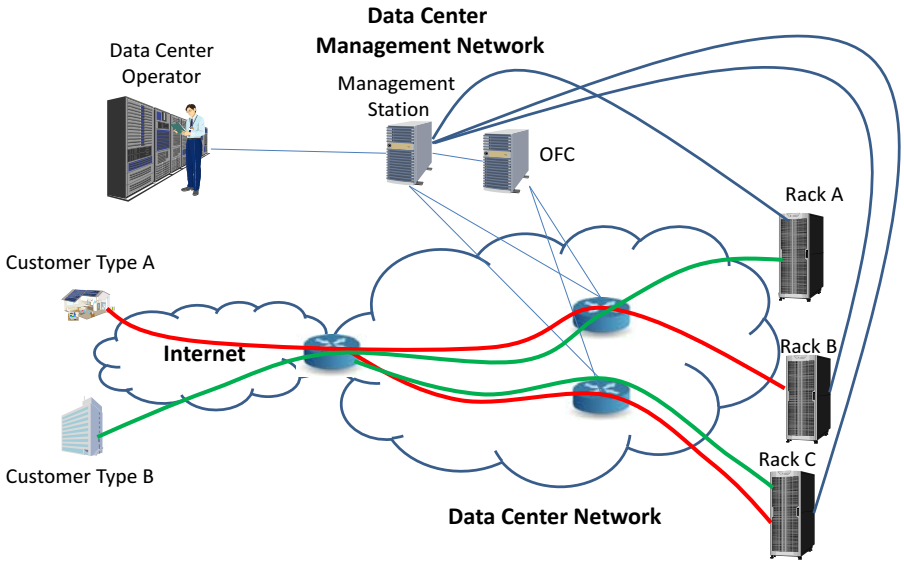

Figure 2.25. ECDC architecture.

service quality for the customer at all times. The management station achieves this by observing a number of configured thresholds and timeouts for each service class.

SDN-based data center network for power saving can control servers and networks in order to shut off idle devices. However, the network does not consider power saving in optical-based data center such as HOPR-based.

Comparison of Energy Efficient Method in Intra Data Center Network

Table 2.3 shows a comparison of three energy-efficient methods for the intra data center network: HOPR-based data center [2-23], data center with VM migration [2-25, 26], and SDN based data center [2-28, 29]. The HOPR-based data center has problems in router and server control. Data center with VM migration and SDN base data center are based on electrical routers. The scheme proposed in Chapter 4 [2-38, 39] realizes router and server control for optical based routers by minimizing working buffers and servers by controlling VM assignment for high data center performance.

Table 2.3. Comparison of energy-efficient methods in intra data center network.

Scheme	Power Consumption	Pros	Cons
HOPR based Data Center [2-23]	Good	Optical-based	Router Control, Server Control
Data Center with VM Migration [2-25, 26]	Good	Router Control, Server Control	Electrical-based
SDN based Data Center [2-28, 29]	Good	Router Control, Server Control	Electrical-based
Proposed in Chap. 4 [2-38, 39]	Very Good	Optical-based, Router Control, Server Control	-

2.2.2 Reliability

Intra data center networks realizing high reliability are shown. Basic resilient system, multiple-class protection, and SDN considering protection are explained.

Basic Resilient System

In the context of data centers, fault-tolerance covers failure handling of components in the data plane (e.g., switches and links) and control plane (e.g., lookup systems) [2-27]. Most of the architectures are robust against failures of the data plane components. For instance, SecondNet [2-30] uses a spanning tree signaling channel to detect failures, and an allocation algorithm to handle them. VL2 [2-31] relies on the routing protocols such as OSPF (Open Shortest Path First).

Basic resilient system such as 1+1 protection in optical network domains can realize high-speed protection with simple operation. However, the system has problems in resource utilization and power consumption due to wasting backup resource.

Multiple-class Protection

Figure 2.26 shows a network topology and an example wavelength-path configuration for multiple-priority-class traffic. The topology is a data center network based on 2D torus. Each node has optical add/drop ports and achieves cut-through transmission. The wavelength for the primary path differs from that for the backup path. The wavelength for lower-priority traffic is the same as that for the backup path, and all the nodes are eligible to transmit and receive lower-priority traffic. The control channel is assigned a dedicated wavelength.

Multiple-priority class traffic in accordance with the demand for each class of traffic in the network is considered. Table 2.4 shows the protection schemes appropriate for

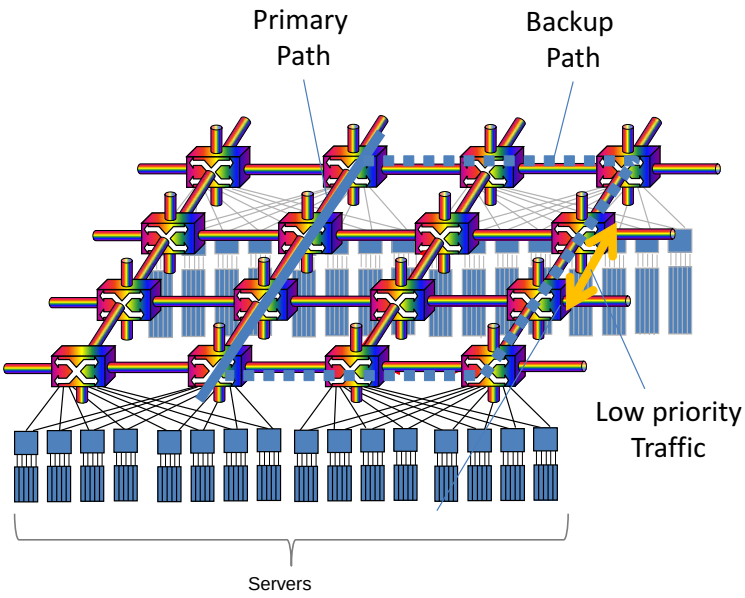

Figure 2.26. Network topology and example of wavelength-path configuration under multiple-priority class traffic.

different types of traffic with different priorities. In general, 1+1 protection is best for high-priority traffic because it achieves high-speed protection as backup paths are always available for data transmission, and destination nodes simply switch to backup paths in the case of failure. Mission critical and broadcasting services, for example, are classed as high-priority class. In general, 1:1 protection with low-priority traffic is best for middle-priority traffic. In 1:1 protection, backup paths are unoccupied, so low-priority traffic can be transmitted on idle backup paths. The 1:1 protection scheme can thus achieve a good balance between high-speed transmission and efficient path utilization. Enterprise services, for example, are classed as middle-priority traffic. Low-priority traffic is not protected and is replaced by backup path traffic in case of network failure. The Internet is classed as low-priority traffic.

The multiple class protection scheme used for suspending low-priority traffic [2-32, 33] in WDM ring networks is explained. It is assumed that control information such as

Table 2.4. Parameters in multiple priority class.

Traffic Type	Traffic Priority	Protection Scheme
Mission Critical, Broadcasting	High	1+1
Enterprise services (VPN etc.)	Middle	1:1 with low-priority traffic
Internet (with upper layer protection)	Low	None

failure notifications, requests to suspend low-priority traffic, and requests to switch paths is transmitted on a control channel.

First, the multiple class protection scheme is overviewed, in which the destination node of a primary path detects a path failure when it loses the signal on a data channel. The node looks up the path information table and identifies the nodes sending low-priority traffic. The node sends a request to suspend low-priority traffic to all nodes transmitting this traffic. After receiving acknowledgment of the suspensions, the node sends a request to switch paths to the source node of the primary path in both directions (clockwise and counterclockwise). If the node cannot receive acknowledgments, the node does not move to next action. When the source node receives this request, it switches the primary path to the corresponding backup path. The destination node receives data from the backup path, and protection is completed. Figure 2.27 shows the actions taken by a node when it detects a failure in the multiple class protection scheme.

Next, specific operations in the multiple class protection scheme are explained. Figure 2.28 shows an example of path protection in the multiple class protection scheme. A link between Nodes B and C has failed. Here, a link failure means a single fiber cut (i.e., unidirectional path failure). The solid arrows from each node indicate the flow of the signal on the control channel, and the two-dots and chain (– -) arrow from Node A to Node E indicates the flow of the signal on the data channel. The heavy line on each node indicates the processing time for each operation. Nodes A and E have path-information

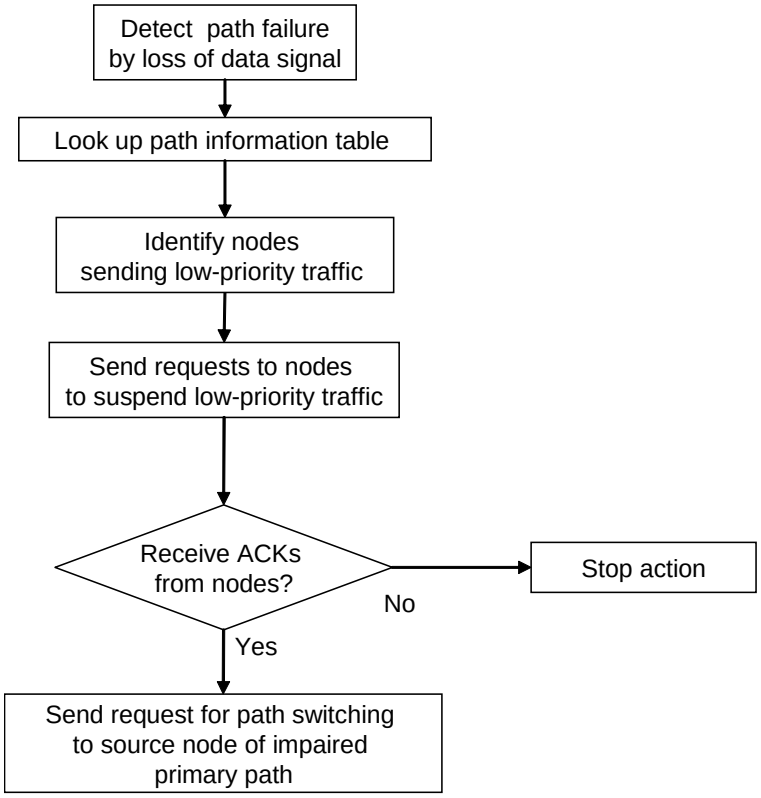

Figure 2.27. Actions taken by node when it detects a failure in multiple class protection scheme.

tables showing the primary paths set by each node, the corresponding backup paths, and the nodes transmitting low-priority traffic on the wavelengths of the backup path. Nodes B, D, F, and H are transmitting low-priority traffic on the wavelength ($\lambda 2$) of the backup path. Therefore, Nodes A and E have information about the low-priority traffic transmitted by these nodes.

Node E detects a path failure between Nodes A and E. It checks the path-information table (as partially shown in Fig. 2.28) to identify which nodes are transmitting low-priority traffic. It then sends requests to those nodes to suspend that traffic. These nodes suspend the traffic and return acknowledgments of the suspension to Node E. When Node E receives the acknowledgments, it sends a request for path switching to the source node of the impaired primary path, i.e., Node A, in both directions. When Node A receives the

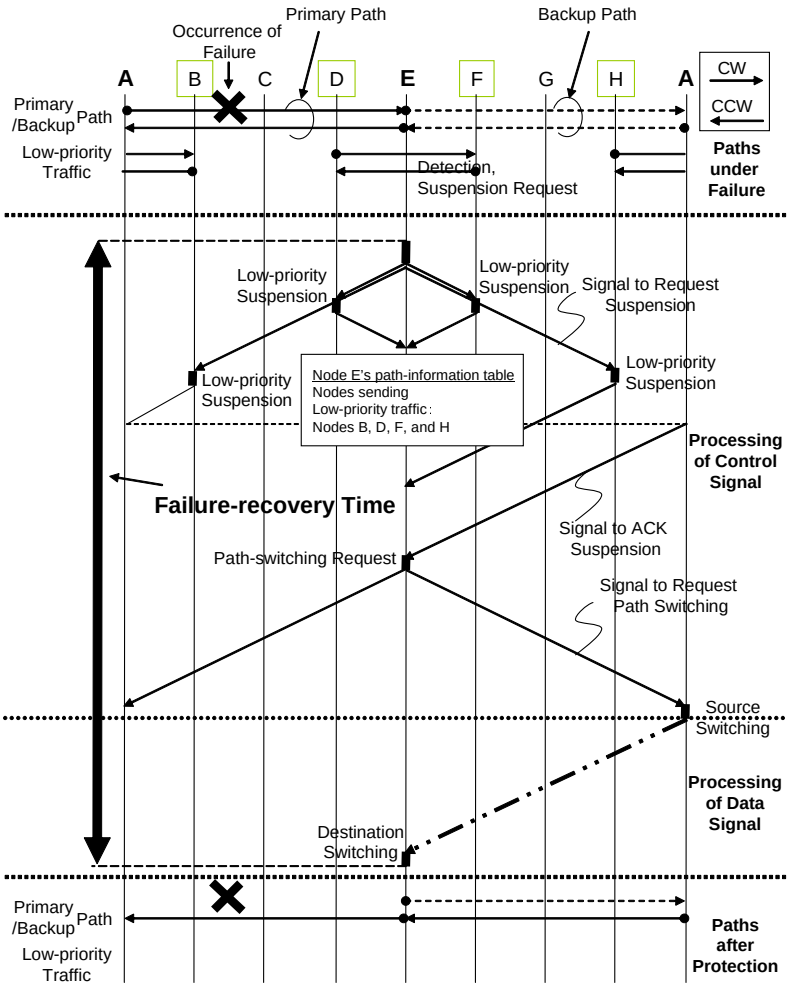

Figure 2.28. Example of path protection in multiple class protection scheme (link failure).

request, it switches the primary path to the corresponding backup path. Node E receives the data from the backup path, and protection is completed.

Figure 2.29 shows another example of path protection in the multiple class protection scheme. A span failure has occurred between Nodes B and C. Here, a span failure means two fiber cuts (i.e., bidirectional path failure). Nodes A and E each detect both path failures; here, we focus on the actions taken by Node E. Because of the span failure, Node B cannot receive the request for suspension in the counterclockwise direction directly and must therefore indirectly receive it in the clockwise direction; it also cannot directly send notification of the suspension in the clockwise direction and must indirectly send it in

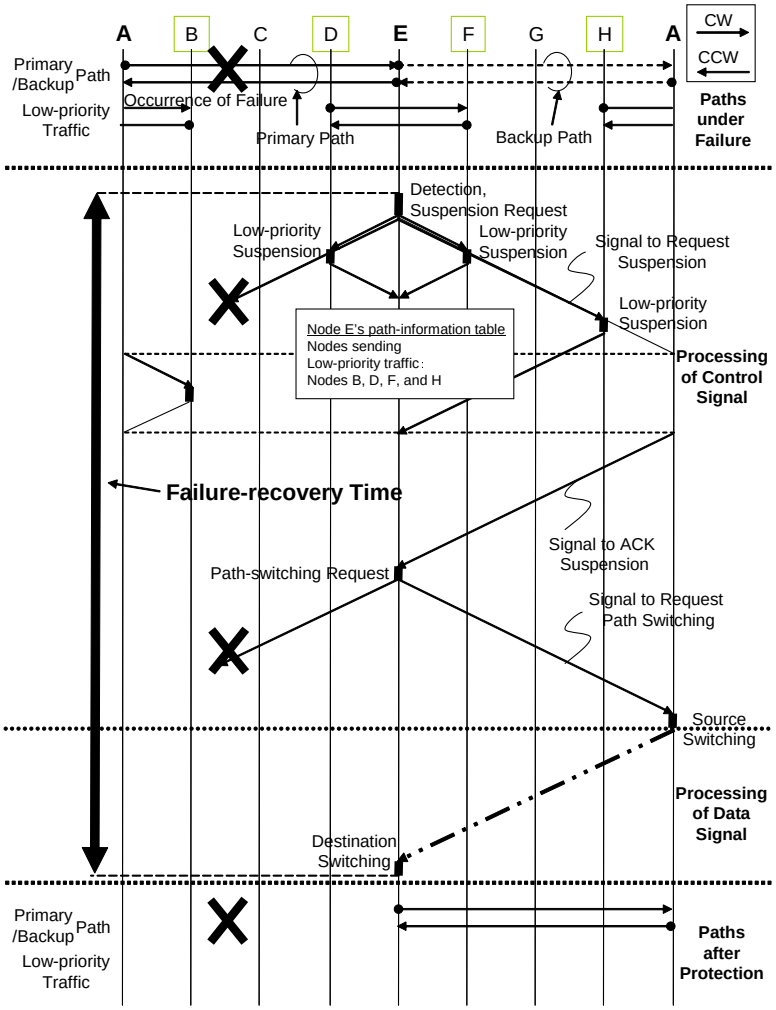

Figure 2.29. Example of path protection in multiple class protection scheme (span failure).

the counterclockwise direction. When Node E receives the notification, it sends a request for path switching to the source node of the impaired primary path, i.e., Node A, in both directions. Node A switches the primary path to the appropriate backup path after receiving this request. In the case of span failure, requests for low-priority traffic suspension and path switching cannot be directly sent; therefore, in the conventional scheme, it takes more time to complete protection after span failure than it does after link failure.

In the multiple class protection scheme, a destination node that detects a path failure

sends requests to suspend low-priority traffic to all nodes that are handling this traffic on the wavelength of the backup path, and it also receives acknowledgments of the suspensions from the nodes. These operations by the destination node are needed before it can send a request for path switching to the source node. The greater the number of nodes transmitting low-priority traffic, the longer it takes to start path-switching operations. It takes even more time in the case of span failure. Therefore, the multiple class protection approach poses difficulties in providing high-speed protection.

SDN with Protection

SlickFlow SlickFlow is an OpenFlow-based fault tolerant routing design that allows switches to recover from failures by specifying alternative paths in the packet headers [2-34]. The approach relies on defining, at the source, the primary route and alternative paths. This enables packets to skip from failures without the intervention of centralized controllers. The key idea of SlickFlow is to represent a path as a sequence of segments that will be used by each switch to perform the forwarding operation. A segment carries the information of the next node of the primary path, and an alternative path related to the current node. If the primary path is not available, then the current switch rebuilds the header and forwards the packet to the alternative path.

Route computation is performed by the controller which has a centralized view of the network, and is able to compute all the routes among endpoints. When a new flow arrives at the network edge, the controller installs rules at source switches (ingress) to embed the SlickFlow header into the packet and at destination switches (egress) to remove the header in order to deliver the packet to the destination endpoint.

The protection mechanism allows switches to reroute packets directly in the data plane as a faster alternative to controller-based path restoration. The source routing approach simplifies the forwarding task in every switch, since it only requires local knowledge of

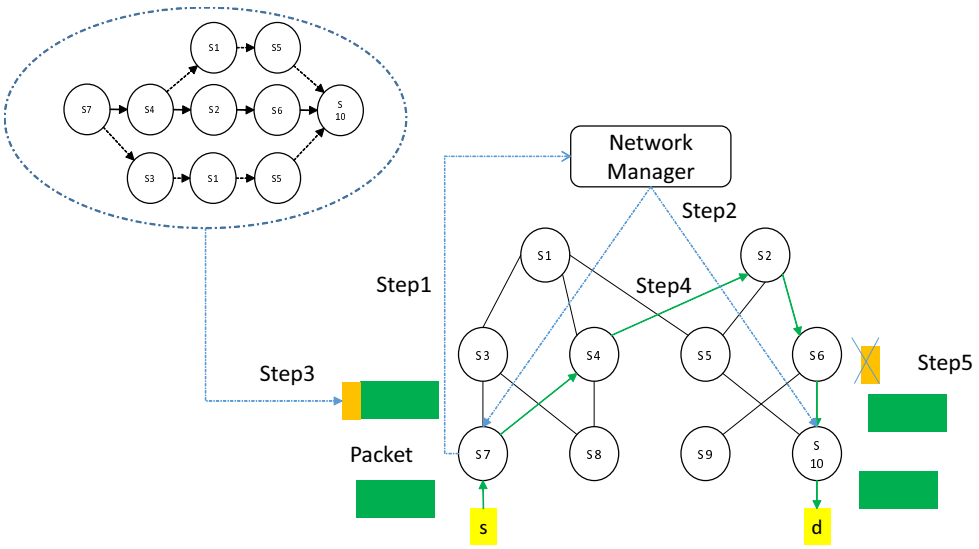

Figure 2.30. Example to illustrate design.

its neighbors, rather than a routing entry for each potential destination.

Figure 2.30 shows an example that illustrates this approach. Suppose that host *s* wishes to communicate with destination *d*. As the flow table of source ToR switch $S7$ is empty, the packet matches no rule and is forwarded to the NM (Network Manager) (Step I). From a list of pre-computed existing paths, NM selects the primary path ($S7$–$S4$–$S2$–$S6$–$S10$) and alternative paths and installs one rule (OpenFlow entry) at the source and destination ToRs (Step II). The rule at source ToR $S7$ instructs the switch to (*i*) embed the selected paths into the packet header fields (Step III), and (*ii*) forward the packet via the output port to $S4$. At $S4$, $S2$, and $S6$, all the forwarding decisions are based on the contents of the SlickFlow header carried by the packet (Step IV). Finally, the rule in the destination ToR $S10$ removes the header and delivers the packet to the destination *d* (Step V). In case of a failure event, the current switch looks for alternatives paths in the SlickFlow header. If found, the packet is rerouted to the backup path without intervention of the NM.

It is worth to note that, SlickFlow forwarding does not use IP addresses for packet flow matching within the data center network. Unlike the traditional hierarchical assignment

of IP addresses, there are no restrictions on how addresses are allocated. In essence, host location is separated from host identifier so routing is transparent to end hosts and compatible with existing commodity switches.

Recovery Scheme Supporting Multiple Failures A failure recovery scheme that supports multiple failures in the data center network is introduced. Figure 2.31 shows SDN path configuration [2-35]. The data plane is MPLS-TP (Multi-Protocol Label Switching-Transport Profile). MPLS-TP provides high reliability and manageability to packet switching networks, and offers strong OAM (Operational Administration and Maintenance) and protection functions equivalent to SONET/SDH [2-36]. The control plane is OpenFlow, and the control plane can setup normal paths, protection paths, and monitoring signals automatically. Therefore, SDN realizes centralized control of the overall transport network by using high reliable MPLS-TP paths. An OpenFlow controller performs topology/path management, path computation/setup, and OAM monitoring/failure recovery. In path setup, the OpenFlow controller connects to MPLS-TP switches by the OpenFlow secure channel, and sends Flow-mod message in OpenFlow to set flow entries in the MPLS-TP switches. The OpenFlow controller considers protection operation in path setup, packet loss/delay in OAM monitoring, and restoration as well as protection in failure recovery. Protection in path setup uses Group-mod message in OpenFlow, and the MPLS-TP switches perform protection by referring to Group table in OpenFlow without contacting the controller. Therefore, high-speed protection is realized. Figure 2.32 shows protection operation in SDN.

The protection schemes using SDN technology can perform multiple recoveries according to network information on centralized control. However, the schemes has small problem in high-speed protection due to centralized control.

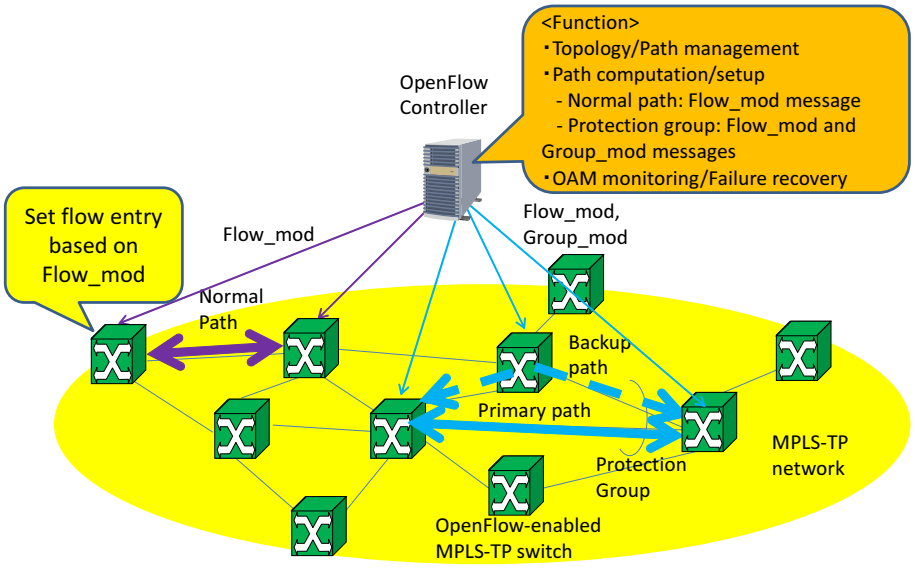

Figure 2.31. SDN path configuration.

Comparison of Reliable Network Method in Intra Data Center Network

Table 2.5 shows a comparison of thee reliable network methods for the intra data center network: 1+1 protection [2-27, 30], multiple class protection [2-32, 33], and SDN with protection [2-34, 35]. 1+1 protection has problems in resource utilization and power consumption. Multiple class protection has small problem in reliability. SDN with protection has slight difficulty with reliability. The scheme proposed in Chapter 5 [2-40, 41] enhances resource utilization, power saving, and reliability by suspending low-priority service rapidly upon bidirectional failure notification from a failure detecting node.

2.3 Conclusion

Based on the related technologies and work described in Sections 2.1 and 2.2, the position of research in this dissertation has been summarized. The ultimate goal is to realize access/intra data center networks that offer high energy efficiency and reliability.

Figure 2.33 shows the position of the dissertation with regard to the access data center

Table 2.5. Comparison of reliable network methods in intra data center network.

Scheme	Reliability (Rapid Recovery)	Pros	Cons
1+1 Protection [2-27, 30]	Good	Simple Operation	Resouce Utilization, Power Consumption
Multiple Class Protection [2-32, 33]	Fair	Resouce Utilization	-
SDN with Protection [2-34, 35]	Fair	Resource Utilization, Power Consumption	-
Proposed in Chap. 5 [2-40, 41]	Good	Resource Utilization, Power Consumption	-

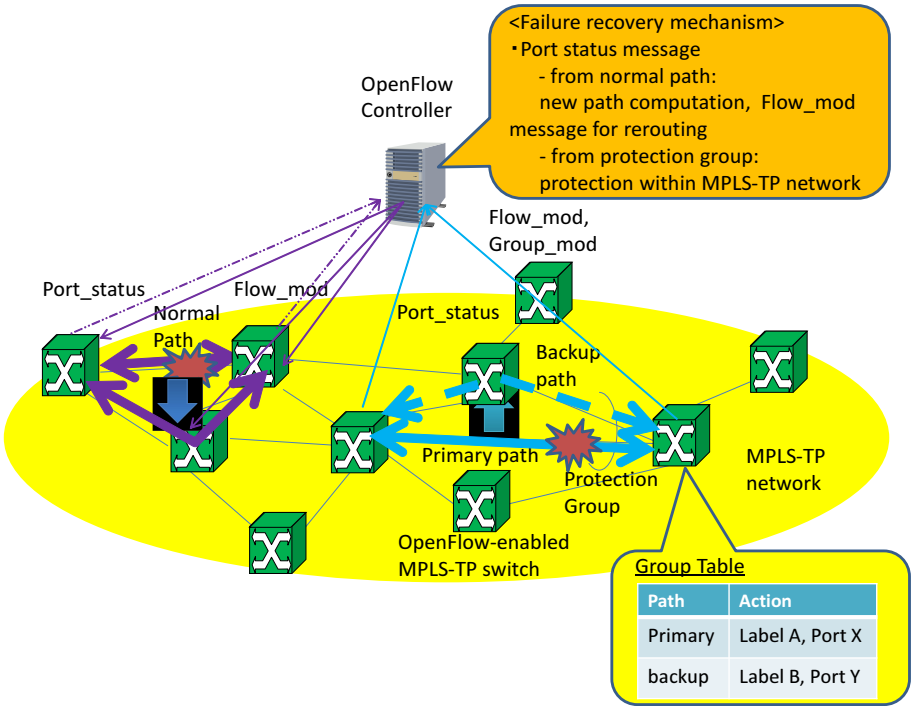

Figure 2.32. Protection operation in SDN.

network. Approaches for energy efficiency in the access data center network are PON, AON, optical metro/access integrated network. Approaches for reliability in the access data center network are PON protection, AON protection, and optical metro/access integrated network protection. The proposal in Chapter 3 for high energy efficiency is based on AON and PON; it minimizes working OLTs according to user traffic and maximizes unnecessary switching times and ONU sleep by accelerated slot assignment [2-37].

Figure 2.34 shows the position of dissertation with regard to the intra data center network. Approaches for energy efficiency in the intra data center network are optical based data center, VM migration for power saving, and SDN based data center. Approaches for reliability in the intra data center network are basic resilient system, multiple-class protection, and SDN with protection. The proposal in Chapter 4 for high energy efficiency is based on optical data center, VM migration, and SDN data center; it minimizes working buffers and servers according to VM assignment based on VM group thresholding for

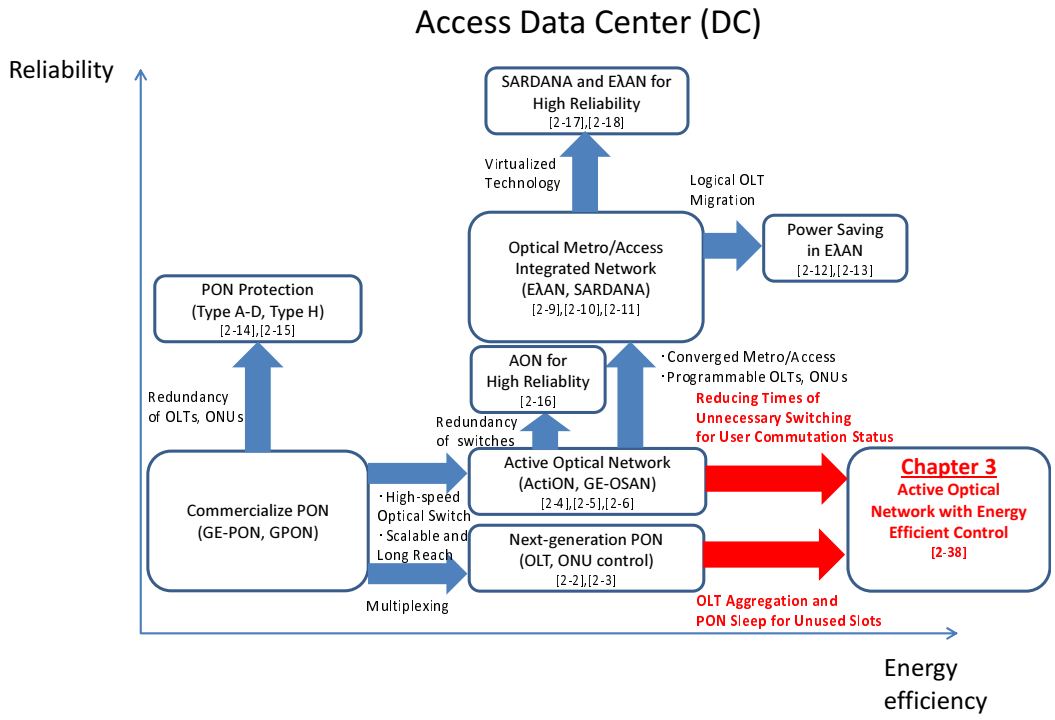

Figure 2.33. Position of dissertation in access data center network.

high data center performance [2-38, 39]. The proposal in Chapter 5 for high reliability is based on SDN based optical data center and multiple-class protection; it suspends low-priority service rapidly on bidirectional failure notification from a failure detecting node [2-40, 41].

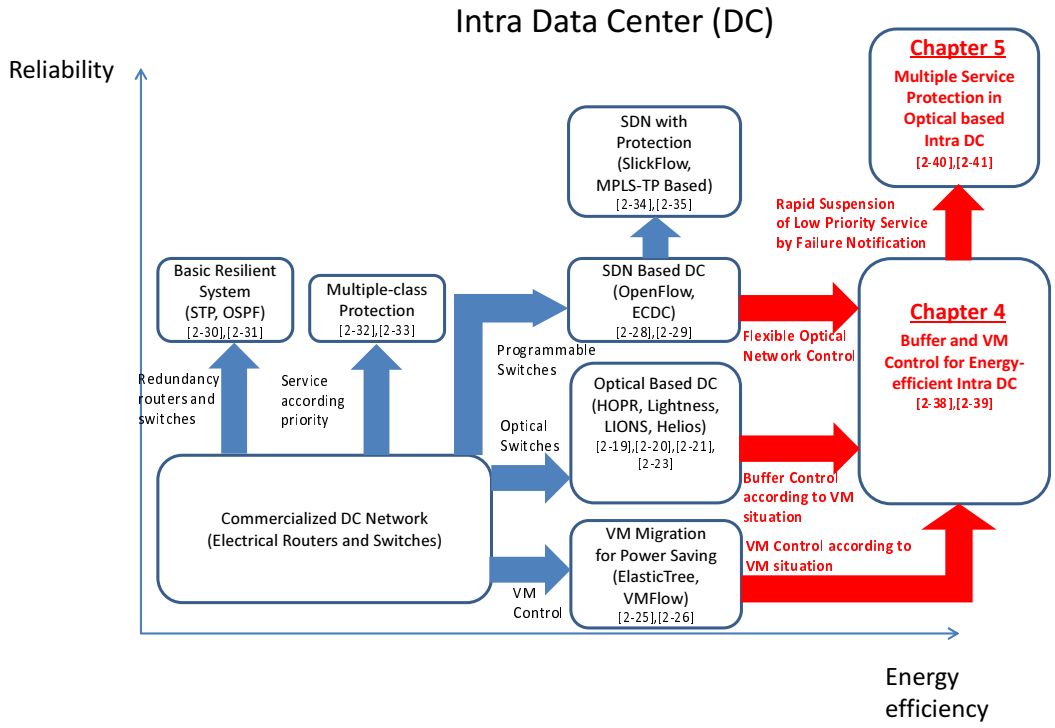

Figure 2.34. Position of dissertation in intra data center network.

References

[2-1] D. Nesset, “NG-PON2 Technology and Standards,” IEEE/OSA J. Lightw. Technol., vol.33, no.5, pp.1136-1143, Mar. 2015.

[2-2] Shing-Wa Wong, L. Valcarenghi, She-Hwa Yen, D.R. Campelo, S. Yamashita, L. Kazovsky, “Sleep mode for energy saving PONs: Advantages and drawbacks,” GreenComm2, IEEE GLOBECOM Workshops, 2009.

[2-3] J. Zhang, T. Wang, N. Ansari, “Designing energy-efficient optical line terminal for TDM passive optical networks,” Sarnoff Symposium,2011 34th IEEE, May 2011.

[2-4] T. Nomura, H. Ueda, T. Tsuboi, and H. Kasai, “Development of New Optical Access Network System Based on Optical Packet Switches,” in Proc. ECOC, no.4.4.3, Sept. 2007.

[2-5] M. Hayashitani, T. Kasahara, D. Ishii, Y. Arakawa, S. Okamoto, N. Yamanaka, N Takezawa, and K. Nashimoto, “Design and Implementation of GMPLS-Based Optical Slot Switching Access-Distribution Network Using PLZT Ultra-High Speed Optical Switch,” in Proc. OFC, no. OWC4, Ahaheim, CA, Mar. 2007.

[2-6] M. Hayashitani, T. Kasahara, D. Ishii, Y. Arakawa, S. Okamoto, N. Yamanaka, N Takezawa, and K. Nashimoto, “10 ns High-speed PLZT optical content distribution system having slot-switch and GMPLS controller,” IEICE Electronics Express, vol.5, no.6, pp.181-186, Mar. 2008.

[2-7] K. Nashimoto, N. Tanaka, M. LaBuda, D. Ritums, J. Dawley, M. Raj, D. Kudzuma, and T. Vo, “High-Speed PLZT Optical Switches for Burst and Packet Switching,” in Proc. Broadband Networks, pp.195-200, Boston, USA, Oct. 2005.

[2-8] G. H. Haertling and C. E. Land, “Hot-Pressed (Pb,La)(Zr,Ti)O_3 Ferroelectric Ceramics for Electrooptic Applications”, Journal of American Ceramic Society, vol.54, no.1, pp.1-11, Jan. 1971.

[2-9] European 7th Framework Programme Project SARDANA. Available: www.ict-sardana.eu.

[2-10] A. Lovric, S. Aleksic, J. A. Lazaro, G. M. T. Beleffi, J. Prat, and V. Polo, “Power efficiency of SARDANA and other long-reach optical access networks,” in Proc. ONDM, Bologna, Italy, Feb. 2011.

[2-11] T. Sato, K. Tokuhashi, H. Takeshita, S. Okamoto, and N. Yamanaka, “A study on network control method in elastic lambda aggregation network (EλAN),” in Proc IEEE HPSR, pp.29-34, July 2013.

[2-12] T. Sato, K. Ashizawa, H. Takeshita, S. Okamoto, N. Yamanaka, and E. Oki, “Logical Optical Line Terminal Placement Optimization in the Elastic Lambda Aggregation Network With Optical Distribution Network Constraints,” J. Opt. Commun. Netw. vol.7, pp 928-941, Sept. 2015.

[2-13] T. Sato, Y. Higuchi, S. Okamoto, N. Yamanaka, and E. Oki, “Logical OLT migration in elastic lambda aggregation network,” IEICE Trans. Commun., vol.J97-B, no.7, pp.474-485, July 2014.

[2-14] T. Nishitani, J. Mizuguchi, and H. Mukai, “Experimental Study of Type B Protection for TWDM-PON System,” IEEE/OSA J. Opt. Commun. Netw. vol.7, pp.A414-A420, March 2015.

[2-15] D. J. Xu, W. Yen, and E. Ho, “Proposal of a new protection mechanism for ATM PON interface,” in Proc. ICC, vol.7, pp.2160-2165, June 2001.

[2-16] T. Sato, K. Ashizawa, K. Tokuhashi, D. Ishii, S. Okamoto, E. Oki, N. Yamanaka, “A Design Method of High-availability and Low-optical-loss Optical Aggregation Network Architecture,” in Proc. IEEE-ISAS 2011, no.GS1-A-2, pp.7-12, June 2011.

[2-17] J. Prat, J. Lazaro, P. Chanclou, R. Soila, A. M. Gallardo, A. Teixeira, G. M. Tosi-Beleffi, and I. Tomkos, “Results from EU Project SARDANA on 10G Extended Reach WDM PONs,” in Proc. OFC, No. OThG5, San Diego, CA, Mar. 2010.

[2-18] A. Kotsugai, T. Sato, H. Takeshita, S. Okamoto, and N. Yamanaka, “TDMA-based OLT sharing method to improve disaster tolerance in Elastic Lambda Aggregation Network,” in Proc. ECOC, P.7.6, Canne, France, Sept. 2014.

[2-19] H. Mehrvar, H. Ma, X. Yang, Y. Wang, S. Li, A. Graves, D. Wang, H. Y. Fu, D. Geng, D. Goodwill, and E. Bernier, “Photonic switching of native ethernet frames for data centers,” In Proc. Photonics Switching, no.JT5C.2, San Diego, CA, USA, 2014.

[2-20] Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella, and S. J. B. Yoo, “LIONS: An AWGR-based low latency optical switch for high-performance computing and data centers,” IEEE J. Sel. Topics Quantum Electron., vol.19, no.2, article 3600409, Mar./Apr. 2013.

[2-21] J. Gripp, J. E. Simsarian. J. D. LeGrange, P. Bernasconi, and D. T. Neilson, “Photonics terabit routers: The IRIS project,” In Proc. OFC, no.OThP3, San Diego, CA. USA, 2010.

[2-22] EpiPhotonics,
http://epiphotonics.com/products.html

[2-23] K. Kitayama, Y. Huang, Y. Yoshida, R. Takahashi, T. Segawa, S. Ibrahim, T. Nakahara, Y. Suzaki, M. Hayashitani, Y. Hasegawa, Y. Mizukoshi, and A. Hiramatsu, "Torus-Topology Data Center Network Based on Optical Packet/Agile Circuit Switching with Intelligent Flow Management," IEEE/OSA J. Lightw. Technol., vol.33, no.5, pp.1063-1071, Mar. 2015.

[2-24] M. Yokokawa, F. Shoji, A. Uno, M. Kurokawa, and T. Watanabe, "The K computer: Japanese next-generation supercomputer development project," in Proc. IEEE/ACM Int. Symp. Low-power Electron. Design, pp.371-372, 2011.

[2-25] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "ElasticTree: Saving Energy in Data Center Networks," in USENIX NSDI, April 2010.

[2-26] V. Mann, P. Dutta, S. Kalyanaraman, and A. Kumar, "VMFlow: Leveraging VM Mobility to Reduce Network Power Costs in Data Centers," NETWORKING 2011. Springer Berlin Heidelberg, 2011.

[2-27] M. F. Bari, R. Boutaba, R. Esteves, L. Z. Granville, M. Podlesny, M. G. Rabbani, Q. Zhang, and M. F. Zhari, "Data Center Network Virtualization: A Survey," IEEE Commun. Surveys & Tutorial, vol.15, no 2, pp.909-928, Second Quarter 2013.

[2-28] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, " Openflow: enabling innovation in campus networks," ACM SIGCOMM Comput. Commun. Review, vol.38, no.2, pp. 69-74, Apr. 2008.

[2-29] M. Jarschel, R. Pries, "An OpenFlow-Based Energy-Efficient Data Center Approach," In Proc. SIGCOMM, pp.87-88, Aug. 2012.

[2-30] C. Guo, G. Lu, H. Wang, S. Yang, C. Kong, P. Sun, W. Wu, and Y. Zhang, "SecondNet: A Data Center Network Virtualization Architecture with Bandwidth Guarantees," in Proc. ACM CoNEXT, Dec. 2010.

[2-31] A. Greenberg, J. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. Maltz, P. Patel, and S. Sengupta, "VL2: A Scalable and Flexible Data Center Network," in Proc. SIGCOMM, pp. 51-62, Aug. 2009.

[2-32] S. Seno, T. Fujii, M. Tanabe, E. Horiuchi, Y. Baba, and T. Ideguchi, "Optical Path Protection with Fast Extra Path Preemption," IEICE Trans. Commun., vol.E89-B, no.11, pp.3032-3039, Nov. 2006.

[2-33] S. Seno, T. Fujii, M. Tanabe, E. Horiuchi, Y. Baba, and T. Ideguchi, "A Proposal and Evaluation of Multi-class Optical Path Protection Scheme for Reliable Computing," High Performance Computing and Communications, vol.3726, pp.723-732, Sept. 2005.

[2-34] R. M. Ramos, M. Martinello, and C. E. Rothenberg, "SlickFlow: Resilient Source Routing in Data Center Networks Unlocked by OpenFlow, " In Proc. IEEE LCN, pp.606-613, Sydney, Australia, Oct. 2013.

[2-35] M. Hayashitani, Y. Hasegawa, K. Suzuki, and Y. Mizukoshi, "Flexible and automated operational control in SDN transport-base virtual router," in Proc. OFC, no.W1E.1, San Francisco, CA, Mar. 2014.

[2-36] N. Jenkins, et al., Requirement of an MPLS Transport Profile, RFC 5654, Sep. 2009.

[2-37] M. Hayashitani, T. Kasahara, D. Ishii, Y. Arakawa, S. Okamoto, N. Yamanaka, N Takezawa, and K. Nashimoto, "GMPLS-based optical slot switching access-

distribution network with a 10 ns high-speed PLZT optical switch," Journal of Optical Networking, vol.7, no.8. pp.744-758, Aug. 2008.

[2-38] M. Hayashitani, K. Suzuki, and Y. Mizukoshi, "A Study on Relationship between Data Center Performance and Network Power Consumption with Buffer Management in HOPR-based Data Center," in Proc. iPOP (IP + Optical Network), no.T4-1, Okinawa, Japan, Apr. 2015.

[2-39] M. Hayashitani, K. Suzuki, and N. Yamanaka, "Control Scheme of HOPR-Based Data Center Network by Considering VM Situation," IEICE Trans. Commun. (Japanese Edition), vol.J99-B, no.4, pp.334-344, Apr. 2016.

[2-40] M. Hayashitani, M. Sakauchi, and K. Fukuchi, "A high-speed protection scheme for multiple-priority-class traffic in WDM ring networks," in Proc. APCC (Asia-Pacific Conference on Communications), pp.1-5, Tokyo, Japan, Oct. 2008.

[2-41] M. Hayashitani, M. Sakauchi, and K. Fukuchi, "A High-speed Protection Scheme for Multiple-Priority-Class Traffic in WDM Ring Networks," IEICE Trans. Commun., vol.E93-B, no.5, pp.1172-1179, May 2010.

Chapter 3

Active Optical Network with Energy-efficient Control

Chapter 2 introduced technologies related to researchers in this dissertation, and access data center network technologies about power saving were illustrated. The conventional approaches for power saving in access data center have problems in scalability and resource utilization, and even if the approaches satisfy the scalability and resource utilization, they have problems in power-saving effect and complex architecture. Chapter 3 proposes active optical network with energy-efficient control in the access data center network [3-1]. The proposal in the access data center network minimizes working OLT according to user traffic and maximizes unnecessary switching times and ONU sleep by accelerated slot assignment. The accelerated slot assignment realized continuous slot assignment for user of most requested slot to reduce switching time and, the assignment created vacant slots in back part for ONU sleep. The proposed scheme can reduce power consumption by 47% as compared to PON.

3.1 Chapter Introduction

Traffic to the data center has been increasing rabidly because of cloud service development. In the near future, traffic to the data center will be most of traffic in access networks as shown in Chapter 1. In the access data center network, network architecture for power saving is needed in order to treat increasing traffic.

In the present access data center network, PON is mainly used. However, PON has a problem about scalability and resource utilization because splitters, which only broadcast packets, are used between OLT and ONU. Therefore, network architecture with aggregation type is needed when traffic to the data center and the number of users increase rapidly. This dissertation focuses on the access data center network realizing higher energy efficiency with larger scalability.

In PON, an OLT is deployed in a central office, each ONU is deployed in each subscriber. Between the OLT and the ONUs, there are splitters to connect the all ONU with the OLT. PON needs more OLTs due to loss of splitters when the central office accommodates more subscribers. AON is proposed for accommodating more subscribers and flexible resource assignment [3-2, 3]. In AON, the splitters are replaced by optical switches, and an optical wavelength is divided into slots that occur cyclically. In AON, the network transfers a content by using slots, and a network user can, in some slots, access large bandwidth. However, AON uses many optical switches, and the switches switch in every slot. Thus, the power consumption of AON with larger scalability is a problem.

Therefore, an energy-efficient optical slot switching network is proposed in the access data center network. The energy-efficient switching network minimizes working OLT according to user traffic and maximizes unnecessary switching times and ONU sleep by accelerated slot assignment.

The rest of this chapter is organized as follows. The energy-efficient access data center network is presented in Section **3.2**. The experiments and performance evaluation of the access data center network are shown in Section **3.3** and Section **3.4**. Finally, this chapter is concluded in Section **3.5**.

3.2 Proposed Active Optical Network

Figure 3.1 shows a proposed access distribution network. This network well supports content distribution in the access network. The proposed network has tree topology from a service provider to clients through optical switches. The slot switching network consists of a control plane and a data plane, and it synchronizes the devices in the network. The control plane employs the GMPLS (Generalized Multi-Protocol Label Switching) [3-4, 5] extension protocol and reserves optical slots based on [3-6]. GMPLS is a set of network control protocols to envision a next generation high performance transport networks. Unlike TDM (Time Division Multiplexing), GMPLS enables the slot switching network to reserve and release slots dynamically, and it realizes the distributed control of optical switches in the network.The data plane is an all optical network based on the PLZT optical switches manufactured by Nozomi Photonics (present EpiPhotonics) [3-7]. This product is the result of a collaboration between Keio University and Nozomi Photonics. The switch device specifications of this product were designed through the collaboration of Keio University and Nozomi Photonics. Nozomi fabricated the switch device, and the switch subsystem was developed by Keio, who also made the controller of the switch subsystem, and the protocol controller for multiple switch systems. Frames are used in the data plane, and each frame has several slots. The data is transferred by using slots. In this section, the optical slot reservation by GMPLS extension protocol and the slot switching by PLZT high-speed optical switch are explained.

3.2.1 Accelerated and Tentative Reservation

The TDM-LSP (TDM-Label Switched Path) scheme in RSVP-TE (Resource reSerVation Protocol-Traffic Engineering) [3-8] is extended, and optical slot reservation is realized. RSVP-TE is standardized as the GMPLS signaling protocol.

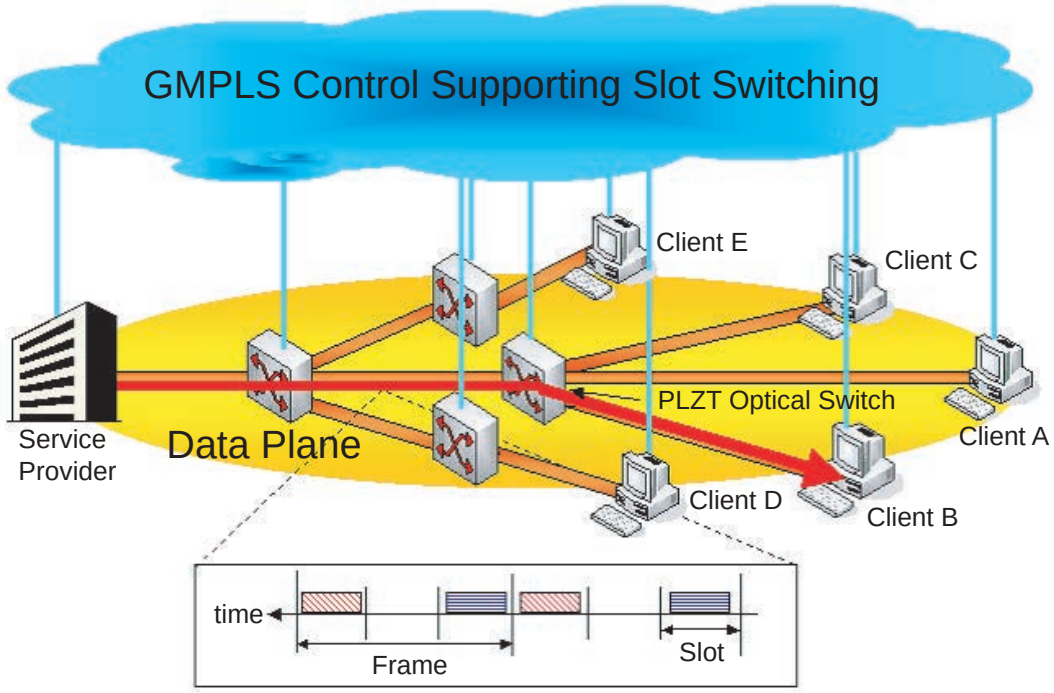

Figure 3.1. Proposed access distribution network.

Figure 3.2 shows an example of slot reservation in the slot switching network. Each vertical line is a time line, and each time line is divided into frames consisting of several slots. In Fig. 3.2, the number of slots in a frame is 3. It is assumed that Service Provider receives a content request from Client A. Service Provider sends a PATH message for Client A after receiving the request. Intermediate Nodes receiving the PATH message confirm whether there are vacant slots. If there are vacant slots, Intermediate Nodes store the information about vacant slots and send the PATH message to the next node. Upon receiving the PATH message, Client A selects Slot 2, the earliest slot among all available slots on all the links of the route, and sends a RESV message for Service Provider. Intermediate Nodes receiving the RESV message reserve Slot 2 and send the RESV message to the next node. Each Intermediate Node maintains information about the reserved slots and the corresponding output port in the PLZT switch. Upon receiving the RESV message, Service Provider transfers the content by using the reserved slots in

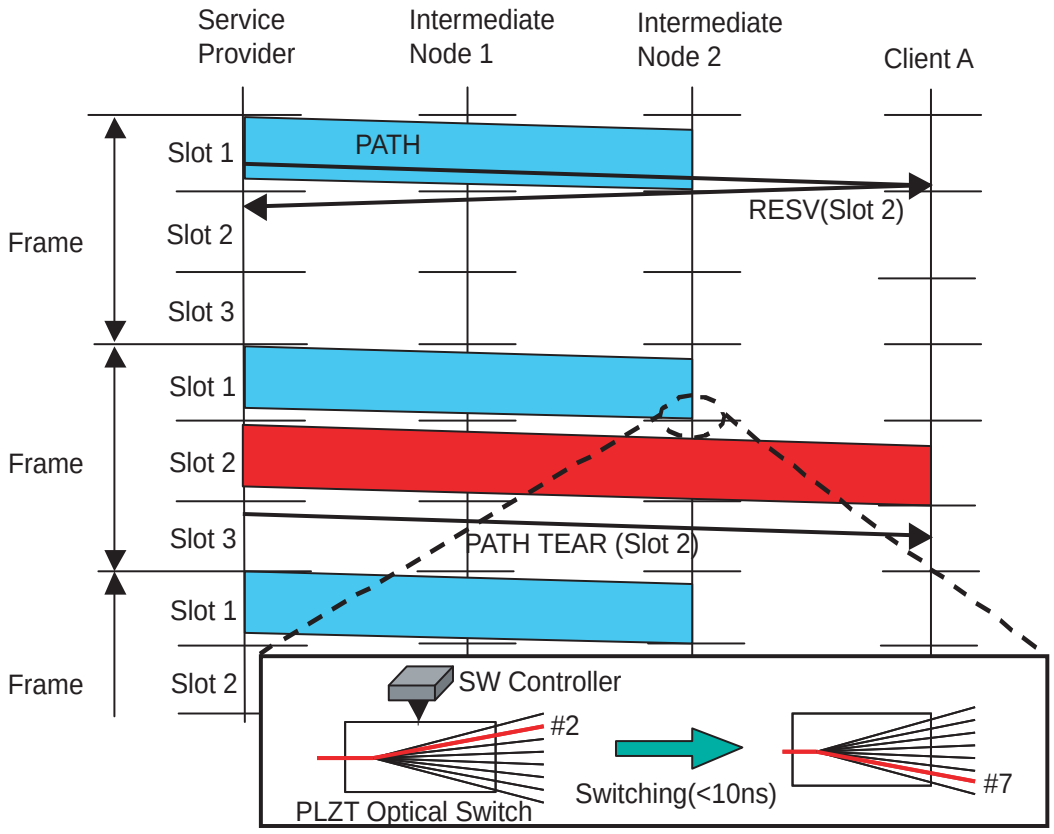

Figure 3.2. Example of slot reservation in the slot switching network.

each frame. Service Provider sends a Path Tear Down message for Client A and releases the slots after finishing content transfer. Each Intermediate Node sends a switch control signal to the PLZT switch based on the information about the reserved slots and the corresponding output port. Each PLZT switch selects an output port based on the switch control signal. In Fig. 3.2, Intermediate Node 2 receives a RSVP signal in the first frame and reserves Slot 2 in the next frame. In the second frame, The switch controller send the switch control signal to the PLZT optical switch in the guard time between Slot 1 and 2 based on the information about the reserved slot, and the PLZT switch switches from Port 2 to Port 7. The switch in Intermediate Node 2 switches from Port 2 to Port 7. Therefore, Client A can receive the content in Slot 2.

The proposed scheme applies the accelerated and tentative slot reservation technique if

by

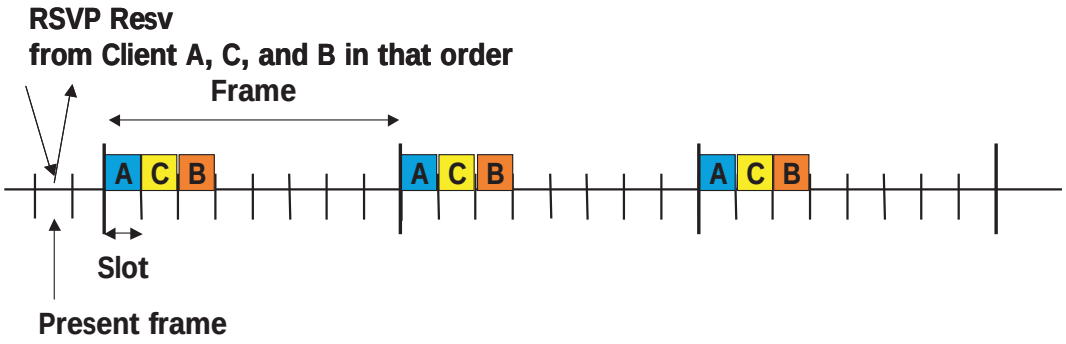

Figure 3.3. Basic slot reservation in the slot switching network.

there are vacant slots in the next frames. First of all, the basic slot reservation technique is explained in the intermediate node of the slot switching network. Figure 3.3 shows the basic slot reservation method. The number of clients is 8, and the number of slots in a frame is 8. In the basic slot reservation method, the proposed scheme receives the RSVP signal in a frame and reserves slots in the next frame. In Fig. 3.3, the proposed scheme receives the RSVP signals from Client A, C, and B, in that order and reserves slots in the next few frames on a first-come-first-served basis. The scheme allocates a slot in a frame for clients in order to guarantee the minimum bandwidth. If the network bandwidth is 10 Gbps, the proposed scheme can guarantee about 1.25 Gbps for all clients in Fig. 3.3.

Next, the accelerated and tentative slot reservation method is explained in the intermediate node. Figure 3.4 shows the flowchart of the accelerated and tentative reservation method, and Figure 3.5 shows the flowchart of data transfer. In Fig. 3.4, vacant slots include the slots tentatively reserved to cover the case of accelerated reservation in order to increase the success probability of accelerated reservation. Accelerated reservation is performed in the next frame, and tentative reservation is performed in the following frames. Therefore, the reservation range of accelerated reservation is smaller than that of tentative reservation. Each switch system has a related table to release the reserve slots when the tentative reserved slots are used for data transfer, as shown in Fig. 3.5.

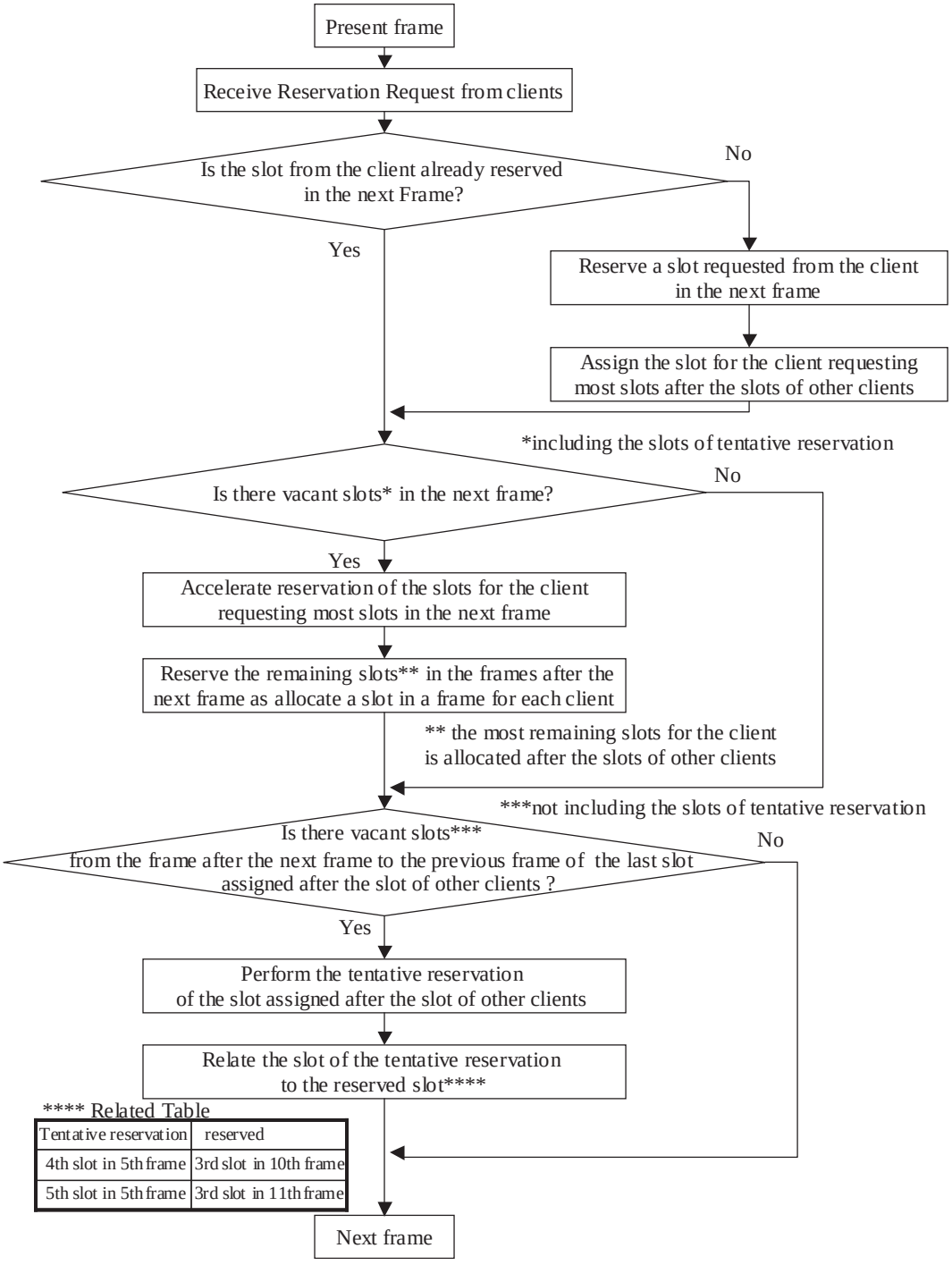

Figure 3.4. Flowchart of the accelerated and tentative reservation.

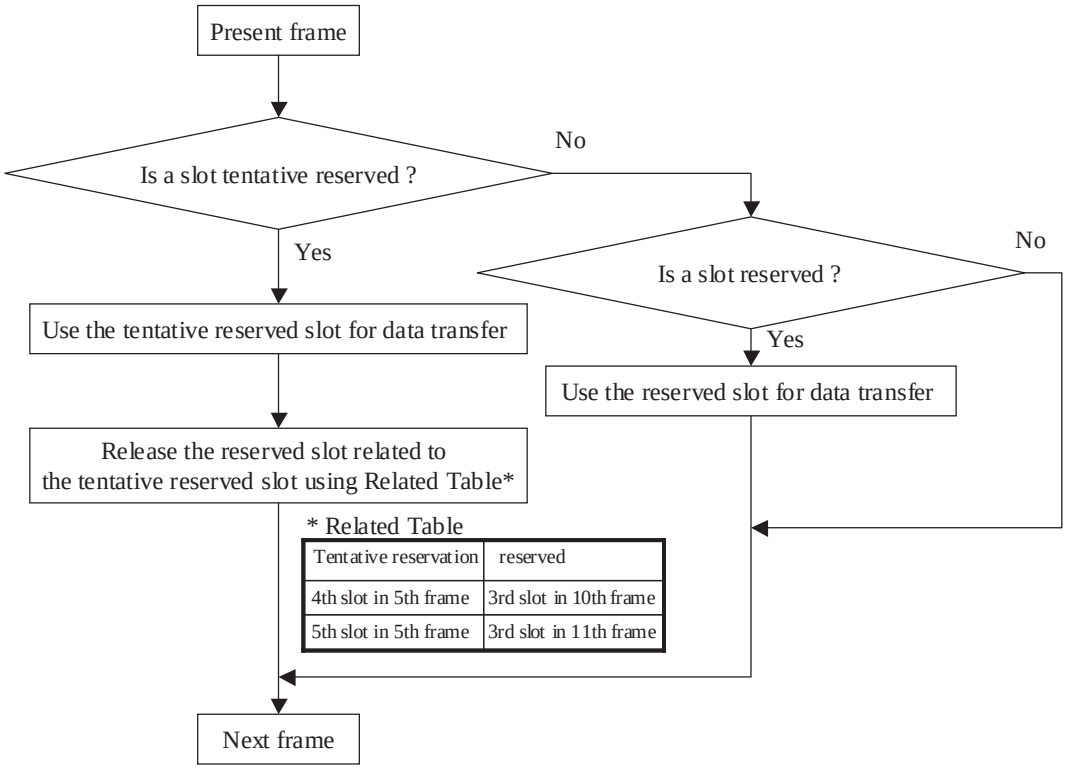

Figure 3.5. Flowchart of data transfer.

Figure 3.6 shows an example of accelerated reservation and tentative reservation. First, the accelerated reservation is explained. As in basic slot reservation, the number of clients is 8, and the number of slots per frame is 8. In Fig. 3.6, the proposed scheme receives RSVP signals from Client A, B, and C. Client A requests 10 slots, Client B requests 6 slots, and Client C requests 3 slots. After receiving the RSVP signals, the scheme reserves slots in the next frame. The order of slot reservation follows the number of slots requested by clients. In Fig. 3.6, Client A requests the most slots, and so its slot is assigned after those of the other clients. This is because the scheme reduces unnecessary switching between slots by reserving continuous sequences of slots for the same client. The scheme checks for vacant slots in the next frame. If there are vacant slots, the slots for the client that requested the most slots are accelerated. In this figure, the slots for Client A are accelerated. Accelerated reservation shortens the delay.

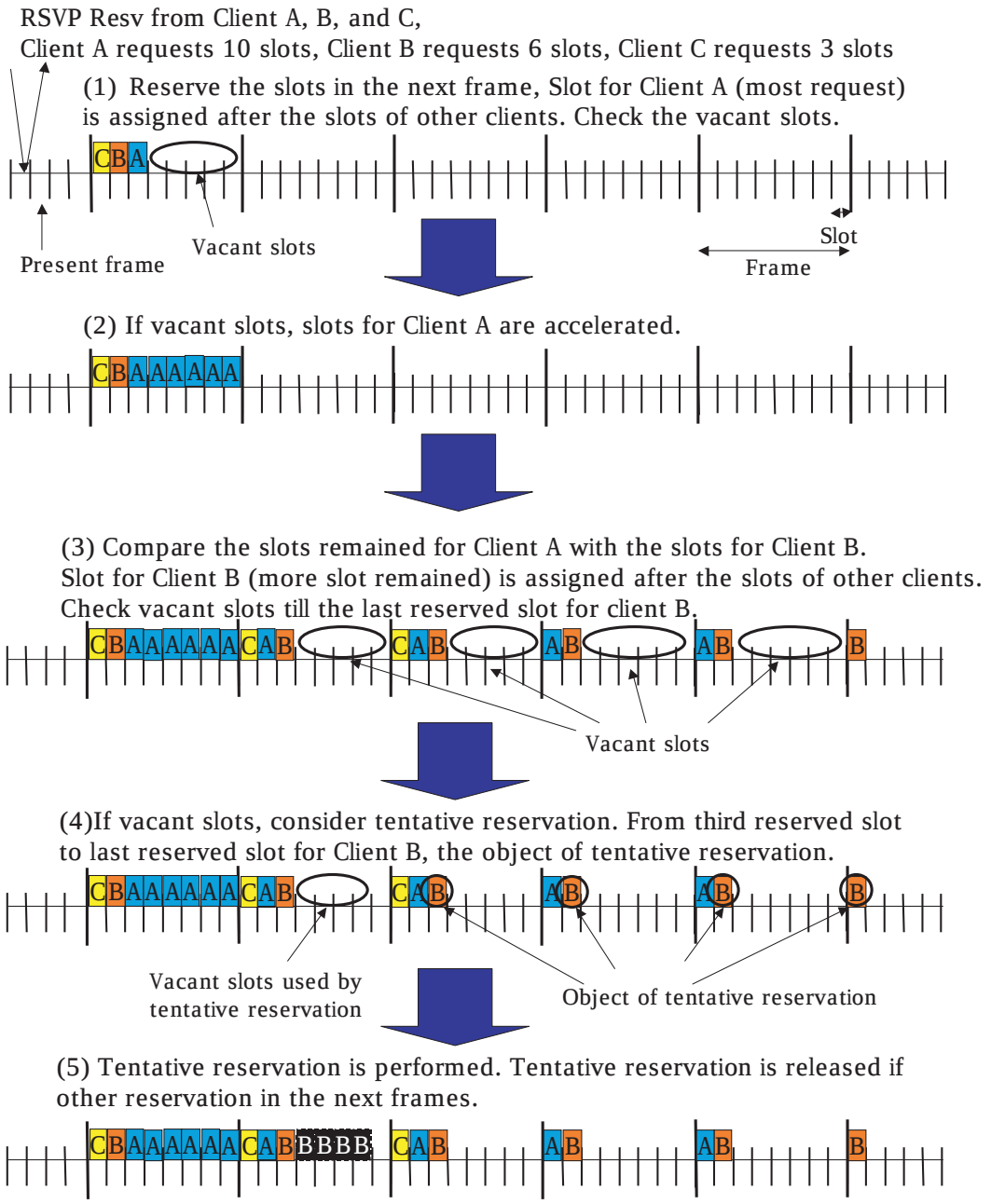

Figure 3.6. Accelerated and tentative slot reservation.

Next, the tentative reservation is explained. The proposed scheme compares the number of slots remaining for the accelerated client with the number of slots for the client that requested the second most slots. The slot for the client that has more remaining slots is assigned after those of the other clients in the next frames for tentative reservation. In this case, Client A has 4 remaining slots, and Client B has 5 remaining slots, so the slot for Client B is assigned last. The scheme repeats the vacant slot checking process up to the last reserved slot for the client that has more remaining slots. If there are vacant slots, the proposed scheme performs tentative reservation. The reserved slots after the vacant slots assigned to tentative reservation are the targets of tentative reservation. In Fig. 3.6, 4 slots for Client B after the vacant slots are the targets of tentative reservation. Next, tentative reservation is performed. In tentative reservation, slots of only one client are targets of tentative reservation. This is because the implementation of the method is made simple as much as possible. If there is other reservations in the next frames, the tentative reserved slots are released. If the tentative reserved slot are used for data transfer, the reserved slot corresponding to the tentative reserved slot is released. Tentative reservation can reduce the delay as can accelerated reservation. Here, the accelerated and tentative reservation method will require signaling overhead. However, the method reserves slots not in the present frame, but in the next frames. Therefore, the signaling overhead in the method does not have a major impact on delay in the proposed system.

Therefore, the proposed scheme can support client reservation flexibly, reserve slots efficiently, and shorten the delay. The proposed scheme applies the rule that the client requesting the most slots is assigned acceleration slots. This seems to be unfair to the other clients, however, all clients can use at least one slot in a frame in order to guarantee the minimum bandwidth. The proposed scheme can satisfy fairness in terms of the minimum bandwidth. In addition, the proposed scheme can reduce unnecessary switching between slots by reserving continuous sequences of slots for the same client.

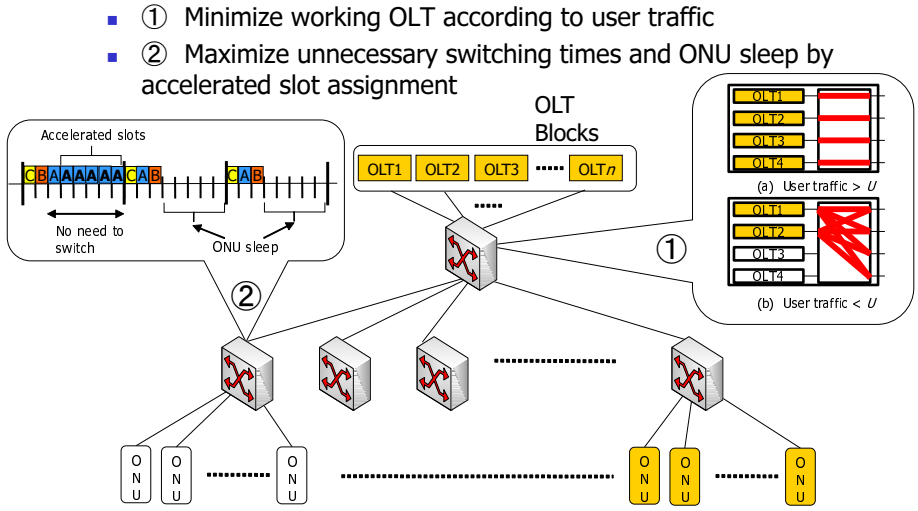

Figure 3.7. Energy-efficient control in access data center network.

3.2.2 Energy-efficient Control in Access Data Center Network

The proposal for high energy efficiency in access data center network controls OLT, optical switches and ONUs. The energy-efficient network aggregates OLT and performs accelerated and tentative slot assignment under optical slot switching condition in order to reduce unnecessary switching and working ONUs. OLT aggregation saves power about OLT. Accelerated and tentative slot assignment saves power about switches and ONUs.

Figure 3.7 shows energy-efficient control in access data center network. The proposed scheme performs power control in every frame. First, each ONU requests slots. Next, the scheme controls the number of L-OLT (Logical OLT) according to the request. The scheme assigns slots by accelerated tentative reservation. Finally, the scheme sleeps ONUs according to the slot assignment.

Therefore, the proposed scheme realizes energy-efficient access data center network.

3.2.3 Implementation of Slot Switching

In the slot switching network, optical switches are used as relay points. In order to realize the slot switching network, it is very important which switch is used. First of all, we consider the MEMS switches are applied for the slot switching network. The MEMS switch is widely used as optical switch and can have a large switch size. However, the switching time of the MEMS switch is several hundred msec. The slot switching network largely depends on the switching time of optical switch. In the proposed network, the slot size is supposed to be μs order if the network user transfers several kbyte data in a slot through 10 Gbps network. If the MEMS switches are used in the slot switching network, the large guard time needs to be put between slots by considering the switching time as shown in Figure 3.8 (a). The guard time is msec order in case of the MEMS switch. Therefore, the guard time almost equals the slot size and causes the overhead in the slot switching network. Thus, the MEMS switch is not suitable for the slot switching network.

The PLZT high-speed optical switch is applied for the slot switching network. PLZT has the advantages of polarization independence, the switch size, and power consumption. The switching time of the PLZT optical switch is less than 10 ns. If the PLZT switch is used in the slot switching network, the very small guard time has only to be put as shown in Figure 3.8 (b). The guard time is nsec order in case of the PLZT switch. Therefore, the network user can use bandwidth efficiently because the PLZT switch can reduce the guard time as compared to the MEMS switch. Thus, the PLZT switch is suitable for the slot switching network.

Figure 3.9 shows the bandwidth efficiency in the slot switching network. In Fig. 3.9, the guard time between slots is twice the switching time. The guard time of MEMS switch is 200 ms, and The guard time of PLZT and SOA is 20 ns in Fig. 3.9. The bandwidth efficiency is defined as

$$\frac{\text{slot size}}{\text{slot size + guard time}}. \tag{3.1}$$

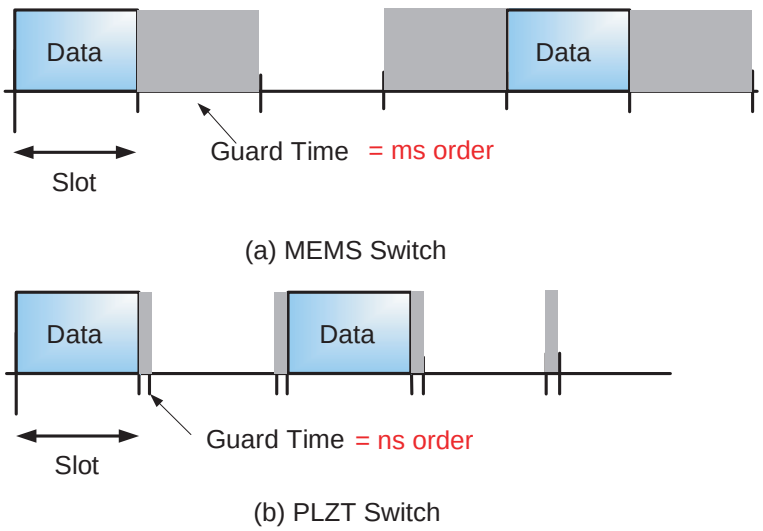

Figure 3.8. Effect of the guard time between slots.

Table 3.1. Slot size for different combinations of bitrate and guard time.

bitrate / slot size	10 μs	100 ms
100 Mbps	125 B	1.25 MB
10 Gbps	12.5 kB	125 MB

Fig. 3.9 shows that the MEMS switch reduces the bandwidth efficiency at short slot sizes because of the overhead of the guard time. The PLZT and SOA switches achieve good bandwidth efficiency even with the short slot sizes because they allow the guard time to be greatly reduced. Table 3.1 shows the slot size as a function of bit rate and guard time. From Table 3.1, slot sizes of μsec order are desirable for a 10 Gbps network given the number of bytes in a slot. This confirms that the PLZT switch should be applied to the slot switching system for a high-speed optical network rather the MEMS switch.

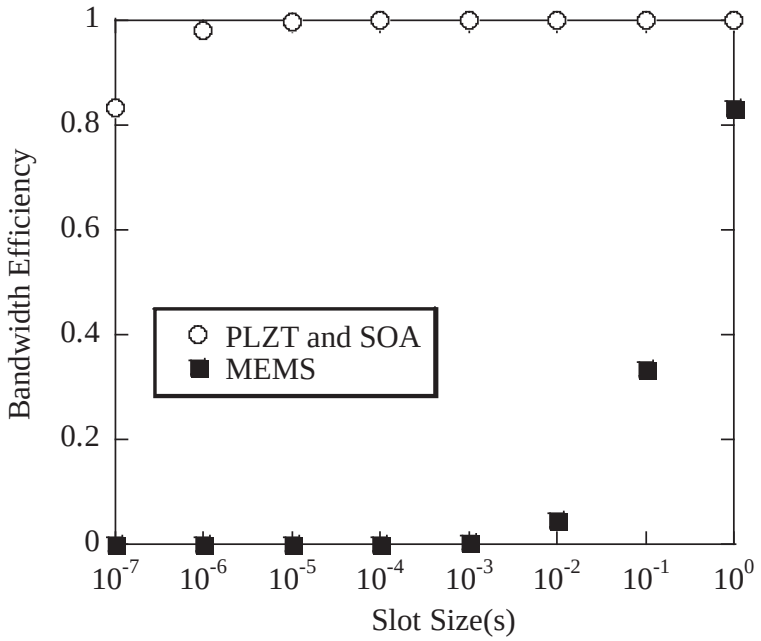

Figure 3.9. Bandwidth efficiency in the slot switching network.

3.3 Experiments

In this section, the experiments of the slot switching network are explained. In this study, to realize the slot switching network, a PLZT optical switch system with a GMPLS-based controller is developed. A pulse pattern generator is used to activate the PLZT optical switch. So, the developed switch system has enabled us to control the PLZT optical switch from GMPLS-based PC. And it is assumed that we the proposed network is applied for the access network, and the distribution network is implemented. It is assumed 8^N users, and N is the number of the stage of switch system. Besides, the proposed network is compared with PON. And the performance evaluation of the accelerated and tentative slot reservation is discussed according to the computer simulation.

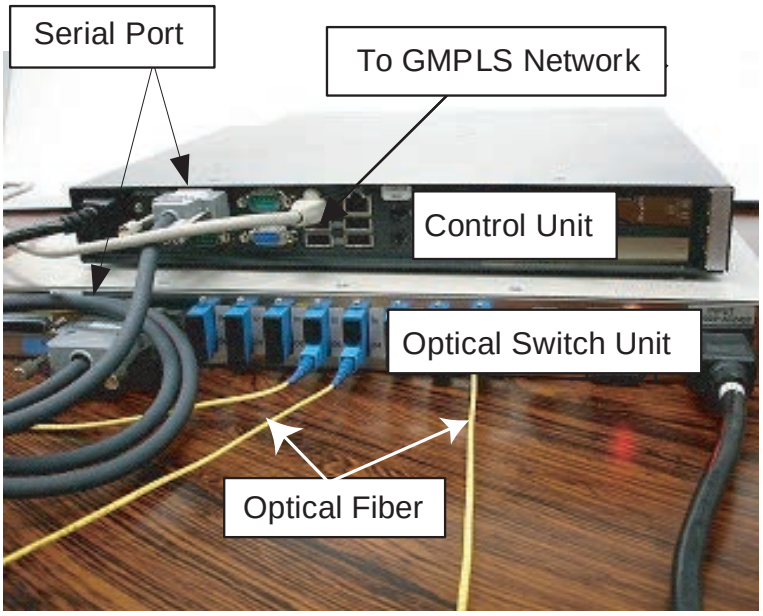

Figure 3.10. PLZT optical switch system with the GMPLS-based controller.

3.3.1 PLZT Optical Switch System

In order to realize the proposed slot switching network, The PLZT optical switch system with the GMPLS-based controller is developed. The developed switch system is the result of a collaboration between Keio University and Nozomi Photonics. Figure 3.10 shows the switch system with the controller. The system consists of a control unit and an optical switch unit. The control unit is a Linux-based PC with GMPLS software and is connected to the optical switch unit via a serial link. Figure 3.11 shows a block diagram of the system. The optical switch unit consists of an ultra-high speed driver board, a fast driver, and an optical switch body. The driver board includes an FPGA that has a pair of 4000 pattern memory banks. It reads and writes the banks based on signals from the controller and sends the appropriate switching pattern signal to the fast driver. The fast driver sends switch signals to the switch body upon receiving signals from the driver board.

The operation of this system is explained when the system receives an RSVP signal. The system gets the absolute time information or frame information from the service

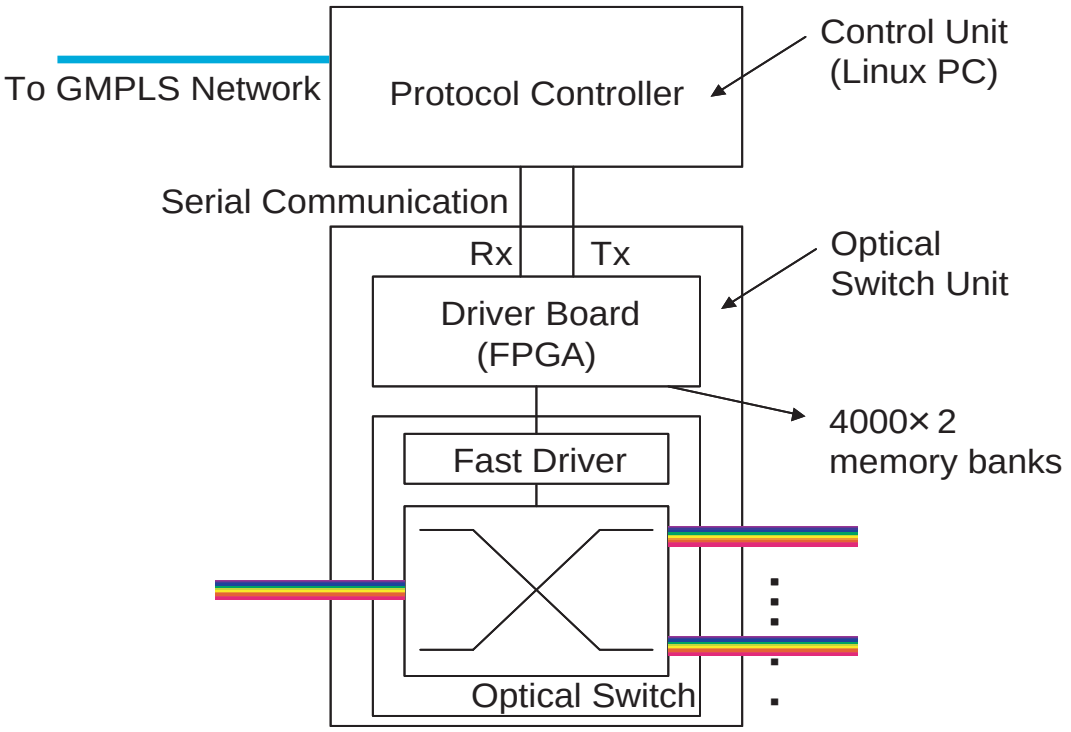

Figure 3.11. Block diagram of the switch system.

provider that is the root of the access network. However, in this prototype switch system, the information is distributed from the master switch system, the root of the switch system. In addition, each switch is synchronized by the frame edge trigger signal from the master switch system in the guard time between slots. This guard time is determined by the switching time and propagation delay among switches. The protocol controller receiving the RSVP signal from the GMPLS network converts the signal into a signal that the optical switch unit can receive. The controller sends the converted signal to the driver board in the optical switch unit via the serial link. The driver board receiving the converted signal stores the reservation information in the memory banks. When the time comes that the slot is assigned, the optical switch unit reads the reservation information from the memory banks and sends the information to the optical switch. The optical switch knows which port the system switches to from the information, and the switch system can switch to the desired port according to the information, and the switch system can switch to the desired

Table 3.2. Specifications of the switch system.

item	numerical value
Switch Size	1x8
Switching Speed	about 10 ns
Switch Loss	about 12 dB
Crosstalk	about -30 dB
Extinction	about 15 - 20 dB
Bit Error Rate	$< 10^{-9}$ (Power > -30 dBm)
Slot Number	4000 Configuration
Slot Period	From 1 Hz to 100 MHz
Memory Bank	2 banks (Switch in 1 clock)

port in the guard time according to the information. This is the operation of this switch system. Therefore, the proposed switch system can activate the switch by slot according to an extended RESV message from the GMPLS control plane.

Table 3.2 shows the specifications of the serial communication and the switch system. The typical insertion loss of 1x8 switch is 12 dB including 3 dB waveguide loss, 6 dB electrode loss, and 3 dB coupling loss. With a modified device structure to minimize the electrode loss, the loss will be reduced to 6 dB in the near future. The extinction ratio of the 1x8 switch is about 15-20 dB. The bit error rate is lower than 10^{-9} when the output power exceeds -30 dBm. In Table 3.2, 1 configuration means a 24-bit switch control signal and 1 memory bank can store 4000 configurations. In addition, the table shows that the system has two memory banks. This enables the system to write a switch control signal in one bank while reading the other bank. When the controller board receives a new slot reservation signal, the two memory banks enable the system to write the reservation signal without having to stop reading the bank. Therefore, the system can realize slot switching without a break.

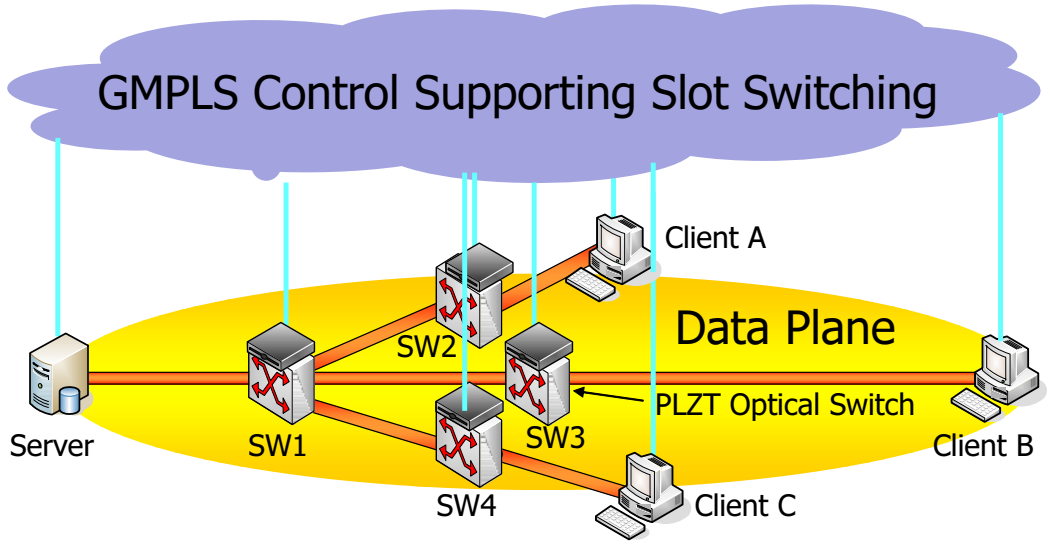

Figure 3.12. Experimental network of the slot switching network.

3.3.2 Experimental Network

Figure 3.12 shows the setup used to implement the slot switching network. In the experimental network, there are a server and three clients. The experimental network has tree topology from the server to the clients through the switch systems. Each PC is a Linux-based PC with GMPLS software. The server and clients all use a newly developed 1.5 μm optical NIC (Network Interface Card). Each Control PC is a control unit of the switch system. Each optical SW, the optical switch unit of the switch system, is a 1x8 switch. The network synchronizes the switch systems by connecting the systems via a synchronization cable. The synchronization cable is a serial cable that transfers the information needed for to synchronize the switch systems.

Figure 3.13 shows an experimental result of the slot switching network. Initially, the slots are reserved for Clients B, A, and C, in that order. Note that the time line is the reverse of the time line shown in Fig. 3.1. Pictures in Fig. 3.13 show the optical waveform measured by an oscilloscope. The second slot for Client A is assigned, and you can see that its optical signal is received in the second slot by Client A. You can also see that

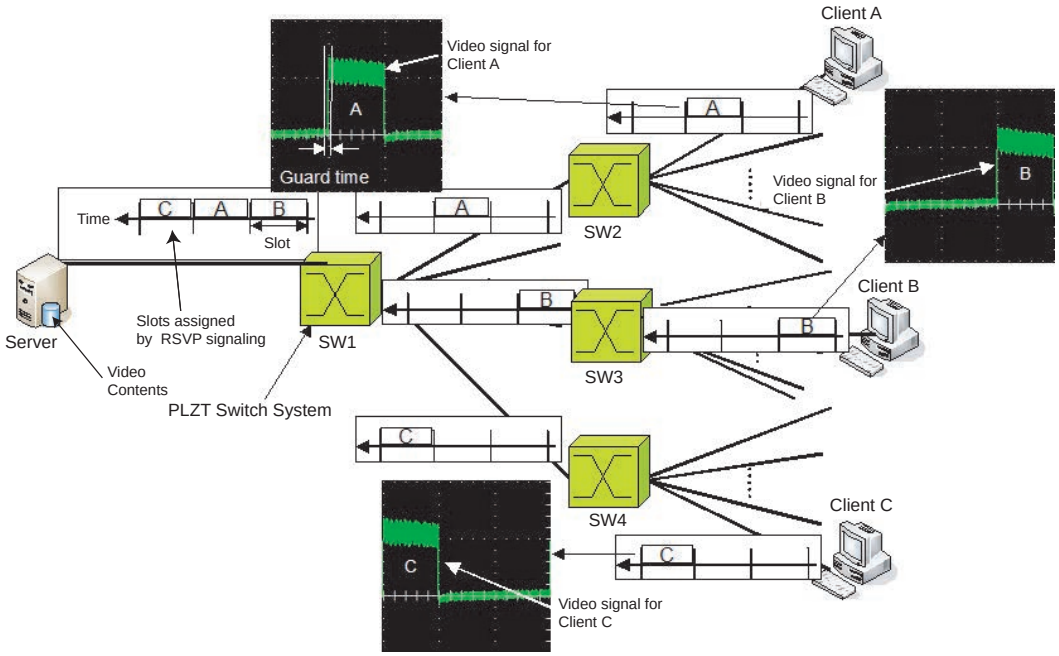

Figure 3.13. Experimental result in the slot switching network.

Client B receives its optical signal in the first slot and Client C receives its optical signal in the third slot. Figure 3.14 shows the switching waveform during the guard time. Fig. 3.14 shows that the switching time of the experimental system is 12.5 ns. This value includes measurement system factors such as the response time of the oscilloscope. The O/E conversion needed to display the waveform on the oscilloscope would affect the rise-up time. The best rise-up time of the same switch device is less than 10 ns as reported in [3-9]. The experimental result confirms that the optical slot switching network can be realized.

3.4 Performance Evaluation

In this section, the scalability and power saving are evaluated in the proposed scheme.

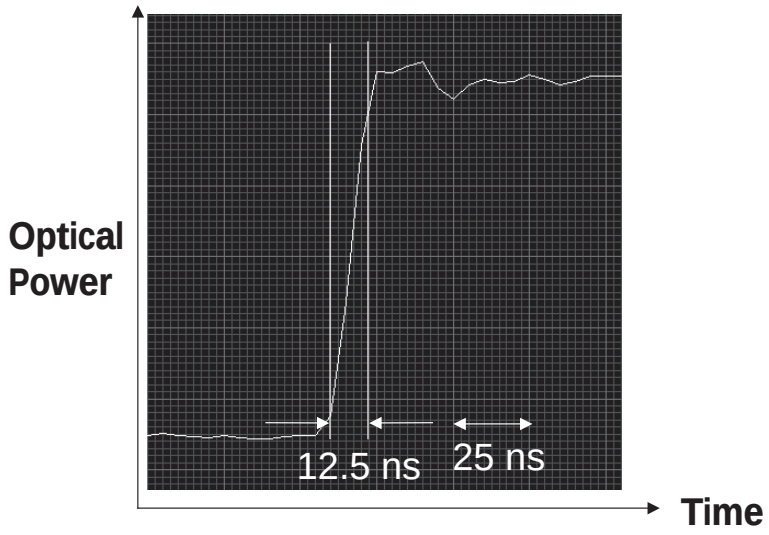

Figure 3.14. Switching waveform during the guard time.

3.4.1 Comparison with PON in Scalability

As described above, the proposed slot switching network well supports content distribution in the access network. This network is called SDSN (Switched Distribution Slot Network). SDSN is compared with PON in terms of the scalability. PON uses splitters in place of optical switches. The loss in the splitter is big because the splitter broadcasts the optical signal. For example, the loss is at least 9 dB if the splitter has 8 branches. SDSN uses optical switches. The loss in the PLZT optical switch is 6 dB and is smaller than that in the splitter because the optical switch switches only to desired port [3-7]. Figure 3.15 shows how many subscribers a service provider covers in terms of loss. It is assumed a 1x8 switch in SDSN. This figure shows that SDSN can cover many more subscribers than PON. If the permissible loss of switches or splitters is 18 dB, SDSN can handle over 500 subscribers while PON can handle only 64 subscribers. Moreover, SDSN can realize a transparent and secure network because the service provider in SDSN sends data to just the client requesting it.

SDSN has the advantages of larger coverage area and stronger security. The former

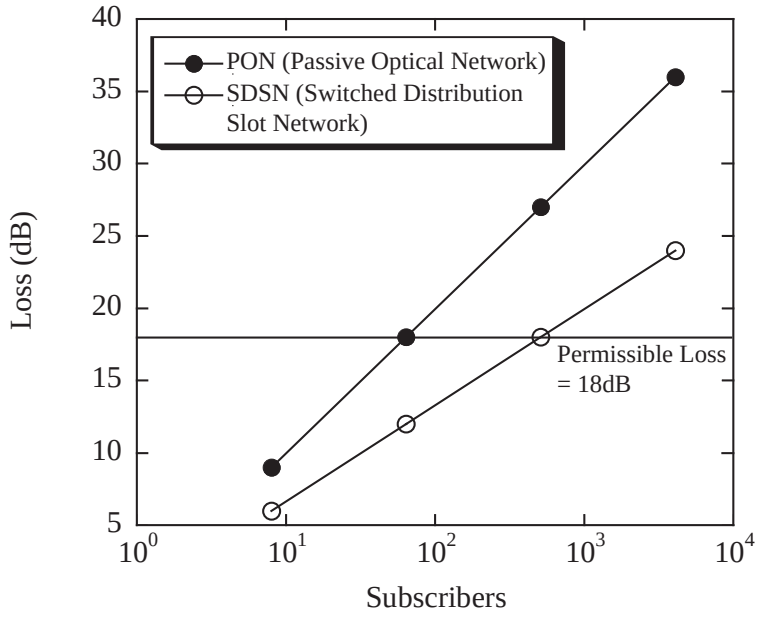

Figure 3.15. Scalability of PON and SDSN.

allows the number of service providers needed to be reduced. Moreover, rather than splitting and broadcasting the data, SDSN sends the data to just the client requesting it. Therefore, SDSN is suitable for content distribution in the access network. Meanwhile, PON has the advantages of low cost and broadcast service. It is important to use these two schemes as the situation demands.

3.4.2 Evaluation of Accelerated and Tentative Slot Reservation

The accelerated and tentative slot reservation scheme is compared with the basic slot reservation scheme according to the computer simulation. The number of slots in a frame is the number of clients. Slot size is 10 μs. The data size is uniformly distributed from 1 to 5 slots. Figure 3.16 shows the delay versus the load in slot reservation. In Fig. 3.16, the number of clients is 8. The delay is defined as the time from when slots for a client is reserved until data transfer is terminated, and the load is defined as the generation probability of calls in one slot. From Fig. 3.16, the accelerated and tentative reservation

scheme reduces the delay compared to the basic reservation scheme under low load. This is because the accelerated and tentative reservation scheme accelerates reservation and performs tentative reservation if there are vacant slots in the next frames. The figure also shows that the difference between the accelerated and tentative reservation scheme and the basic reservation scheme becomes small as the load increases. This is because the number of vacant slots becomes small, which weakens the advantage of the accelerated and tentative reservation scheme. Therefore, the proposed scheme can shorten the delay.

The data size and the number of clients are discussed. Figure 3.17 shows the delay versus the data size. The slot size is 10 μs and 12.5 kByte in 10 Gbps networks, as shown in Table 3.1. The data size corresponds to the number of slots. The data size 1 means 12.5 kByte in 10 Gbps networks. In Fig. 3.17, the load is 0.25 and the number of slots in a frame is 8. Fig. 3.17 shows the difference in delay between the two schemes increases with the data size. This is because large data is more likely to be the target of accelerated and tentative reservation. Figure 3.18 shows the delay versus the number of clients. In Fig. 3.18, the load is 0.25. Fig. 3.18 shows that the delay increases with the number of clients. This is because the number of slots in a frame increases with the number of clients; the two slots assigned to the same client in two sequential frames are offset by a larger amount of time. However, tentative reservation has more vacant slots to utilize which makes the proposed scheme more effective.

Figure 3.19 shows the frequency of switching versus the load in slot reservation. The number of slots in a frame is 8. The frequency of switching is defined as the probability of performing slot switching between slots. Fig. 3.19 shows that the accelerated and tentative reservation scheme reduces the frequency of switching compared to the basic reservation scheme under low load. This is because the accelerated and tentative reservation scheme reserves continuous sequences of slots for the same client. At and above medium loads, the two schemes have basically the same frequency of switching. This is

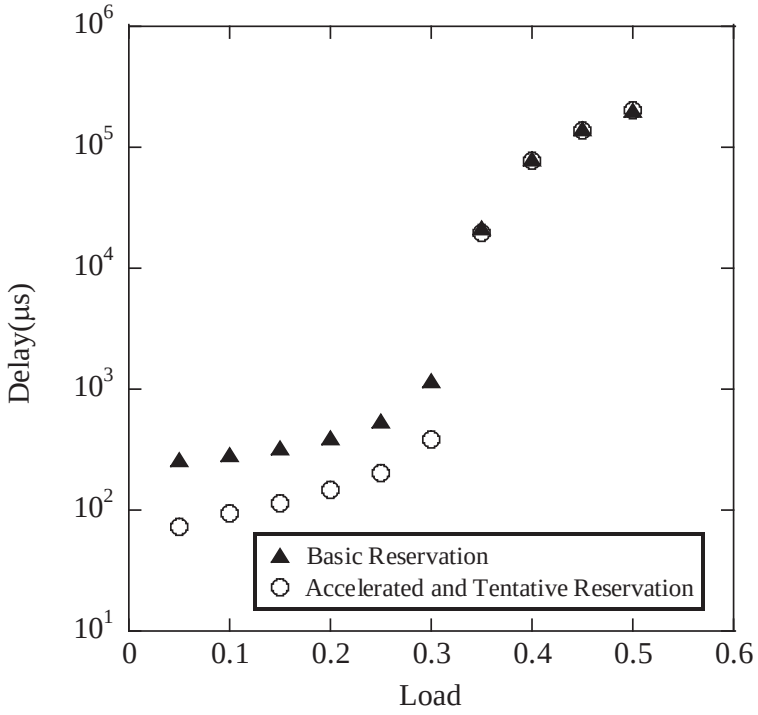

Figure 3.16. Delay versus the load in the slot reservation.

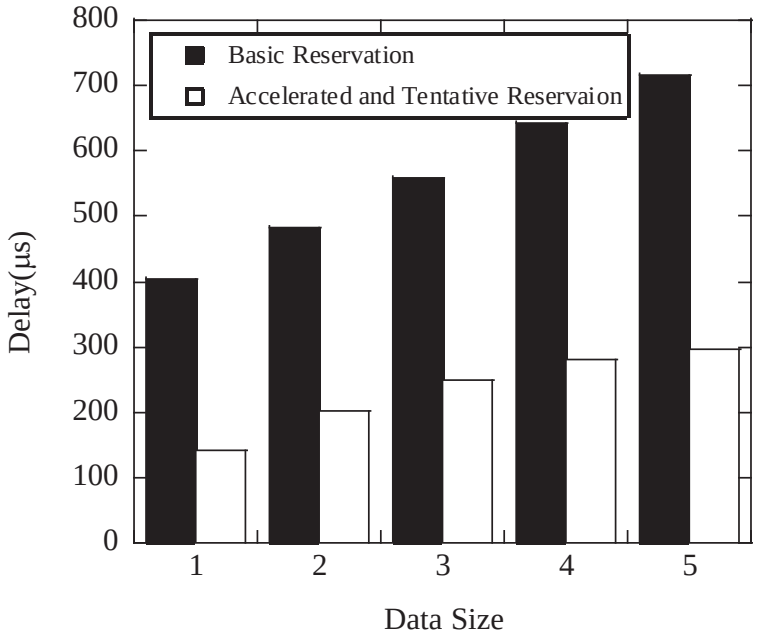

Figure 3.17. Delay versus the data size (Load:0.25).

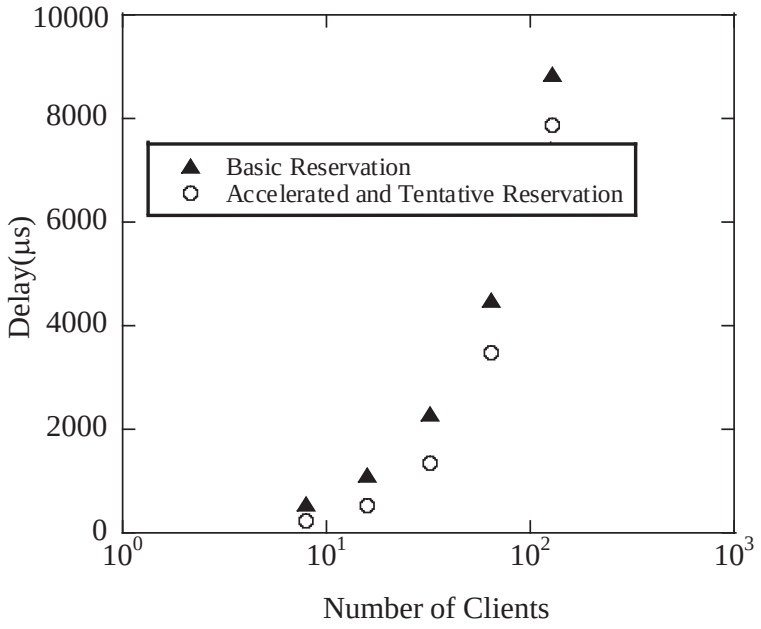

Figure 3.18. Delay versus the number of clients (Load:0.25).

because there are few vacant slots available for performing tentative reservation. Overall, the proposed scheme can reduce the frequency of unnecessary switching between slots and utilize the bandwidth more efficiently.

3.4.3 Evaluation of Power Consumption in Access Data Center Networks

This subsection compares power consumption in access data center networks. Figure 3.20 shows comparison of power consumption in access data center networks. Table 3.3 shows power consumption of devices. Power consumption of switch with less switching means that based on the result from Fig. 3.19. From Fig. 3.20, the proposed scheme can reduce power consumption by 47% as compared to PON. This is because the proposed scheme minimizes working OLT according to user traffic and maximizes unnecessary switching times and ONU sleep by the accelerated slot assignment.

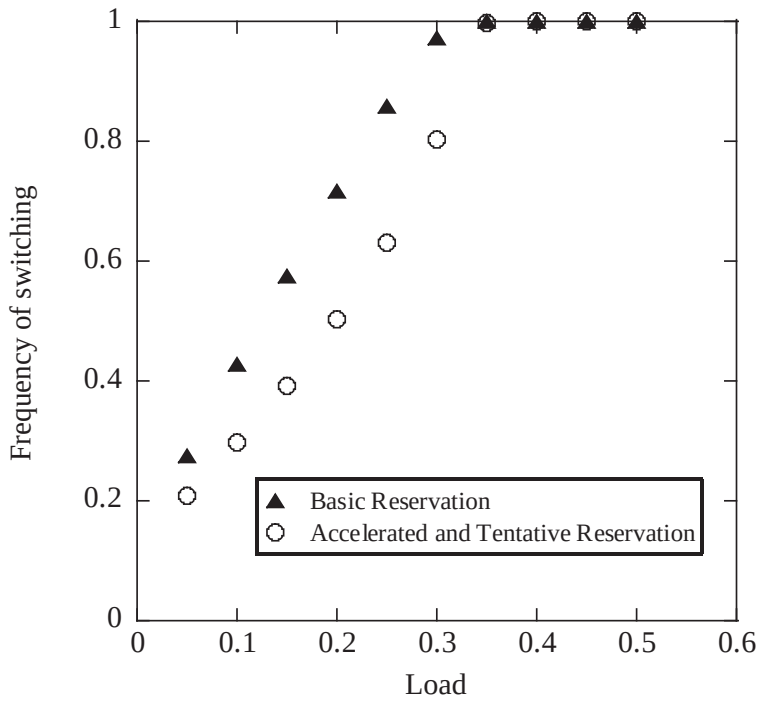

Figure 3.19. Frequency of switching versus the load in the slot reservation.

Table 3.3. Power consumption of devices (user: 512).

Device	W
OLT	400
ONU	5
Switch (1x8)	25.0
Switch (1x8) with less switching	18.8

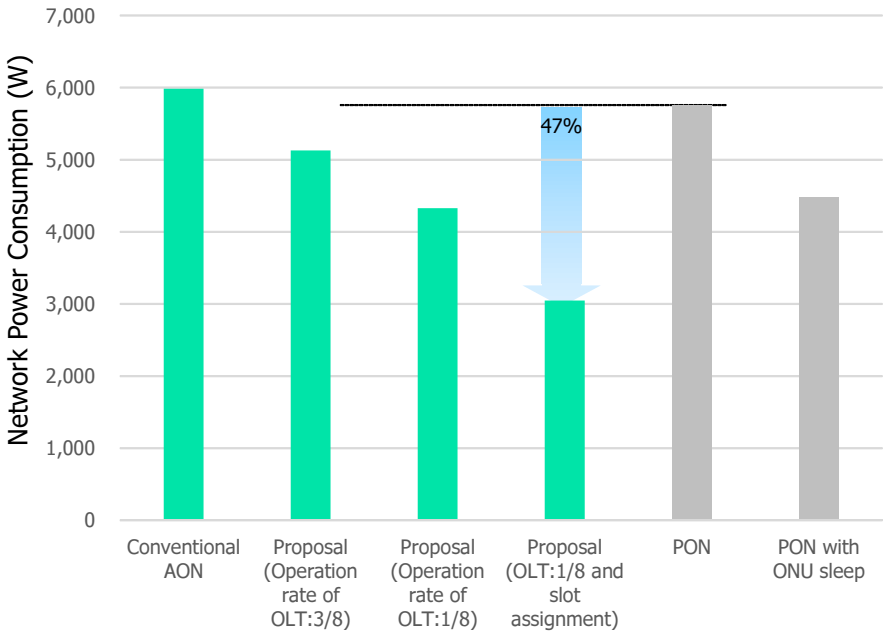

Figure 3.20. Comparison of power consumption in access data center networks.

3.5 Conclusion

Chapter 3 proposed active optical metro/access network with energy-efficient control. The proposal for high energy efficiency in the access data center network minimized working OLT according to user traffic and maximized unnecessary switching times and ONU sleep by accelerated slot assignment. The accelerated slot assignment realized continuous slot assignment for user of most requested slot to reduce switching time and, the assignment created vacant slots in back part for ONU sleep.The proposed scheme reduced power consumption by 47% as compared to PON.

References

[3-1] M. Hayashitani, T. Kasahara, D. Ishii, Y. Arakawa, S. Okamoto, N. Yamanaka, N Takezawa, and K. Nashimoto, “GMPLS-based optical slot switching access-distribution network with a 10 ns high-speed PLZT optical switch,” Journal of Optical Networking, vol.7, no.8. pp744-758, Aug. 2008.

[3-2] T. Nomura, H. Ueda, T. Tsuboi, and H. Kasai, “Development of New Optical Access Network System Based on Optical Packet Switches,” in Proc. ECOC, no.4.4.3, Sept. 2007.

[3-3] M. Hayashitani, T. Kasahara, D. Ishii, Y. Arakawa, S. Okamoto, N. Yamanaka, N Takezawa, and K. Nashimoto, “Design and Implementation of GMPLS-Based Optical Slot Switching Access-Distribution Network Using PLZT Ultra-High Speed Optical Switch,” in Proc. OFC, no. OWC4, Ahaheim, CA, Mar. 2007.

[3-4] E. Mannie. ed., “Generalized Multi-Protocol Label Switching (GMPLS) Architecture,” IETF RFC 3945, Oct. 2004.

[3-5] T. S. El-Bawab, “Generalized Multi-Protocol Label Switching (GMPLS): In Quest of Data and Transport Integration,” Proc. Globecom 2004 Workshops, pp.343-344, Dallas, USA, Dec. 2004.

[3-6] T. Kasahara, M. Hayashitani, D. Ishii, and N. Yamanaka, “Optical Slot Switching Architecture based on dynamic path setup using ultra-high speed PLZT optical

switch," in Proc. International Symposium on Contemporary Photonics Technology (CPT) 2007, no.G-1, pp.67-68, Tokyo, Japan, Jan. 2007.

[3-7] EpiPhotonics,
http://epiphotonics.com/products.html

[3-8] L. Berger, ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extentions," IETF RFC 3473, Jan. 2003.

[3-9] K. Nashimoto, N. Tanaka, M. LaBuda, D. Ritums, J. Dawley, M. Raj, D. Kudzuma, and T. Vo, "High-Speed PLZT Optical Switches for Burst and Packet Switching," in Proc. Broadband Networks, pp.195-200, Boston, USA, Oct. 2005.

Chapter 4

Buffer and VM Control for Energy-efficient Intra Data Center Network

Chapter 2 introduced intra data center network technologies about power saving were illustrated as well as the access data center network technologies. The conventional approaches for power saving in intra data center have problems in realizing technologies under optical-based routers. Chapter 4 proposes optical energy-efficient intra data center with buffer and VM control [4-1, 2]. The proposal for high energy efficiency based on the optical data center realizes minimizes working buffers and servers according to VM assignment based on threshold in VM groups under high data center performance. The VM assignment had VM aggregation for power saving and distribution for high data center performance. The proposed scheme can reduce network power consumption by 40% and server power consumption by 59% as compared to the conventional HOPR-based data center.

4.1 Chapter Introduction

Traffic in the data center has been increasing rabidly because of cloud service development. 70% of traffic in the data center flows within the data center network as shown in Chapter 1. In the intra data center network, network architecture for power saving is needed in order to treat increasing traffic.

In the present intra data center network, electrical switches and routers are usually

used. However, in the data center based on electrical technologies, power consumption increases rapidly when the traffic within the data center network increases. Therefore, optical technologies which can realize power saving and larger scale network for the data center are needed in order to treat increasing traffic in the intra data center network. This dissertation also focuses on the intra data center network realizing power saving with high data center performance.

Current intra data center network is facing technical challenges of high power consumption due to large volume electrical switches in core part of data center network. The power consumption of Cisco Nexus 7000 commonly used in the data center network is 2.5 W/Gbps [4-3]. Electrical routers are the major contributors to power consumption in intra data centers, where the power scales up as the size of the intra data center network increases. The low latency is crucial for intra data center networks because analysis of big data, based upon real-time CEP, will be performed by exchanging huge quantities of data on the intra data center network.

The power consumption of optical switch is a few dozen mW/Gbps. Therefore, it leads lower power consumption to deploy more optical switches in networks [4-4]. Some groups proposed architecture of the data center network with optical switch technologies [4-5, 6, 7, 8, 9, 10]. In the technologies, HOPR-based data center network is proposed [4-11]. HOPR has an electrical buffer based on CMOS in order to avoid packet contention and treat long hops due to optical power attenuation.

In addition, VM migration schemes are proposed as power saving schemes in the intra data center network [4-12, 13]. A scheme turns off unused links and switches by selecting switches which must be turned on according to performance of data center network and fault tolerance. However, the scheme turns off links and switches by monitoring only traffic, and does not save power in the data center network by considering VM situation in servers. Other scheme turns off switches by considering VM situation in servers. How-

ever, it is assumed that only one VM in a server works. In real data center, multiple VMs in a server works for efficient use of servers. These schemes use fat-tree topology which needs switches with many ports in performance evaluation, and do not consider the network based on HOPR which does not have many ports.

Therefore, an energy efficient intra data center network is proposed. The proposed energy efficient intra data center network minimizes working buffers and servers according to VM assignment based on threshold in VM groups under high data center performance.

The rest of this chapter is organized as follows. The proposed scheme is shown in Section **4.2**, and performance of the proposed scheme is explained in Section **4.3**. This chapter is concluded in Section **4.4**.

4.2 SDN Based Data Center with Buffer Control

In this section, a control scheme of HOPR-based green data center network by considering VM situation is proposed. The proposed scheme performs VM aggregation according to threshold by network controller, and turns off the unused buffers when servers under the HOPRs do not work. The scheme also performs VM distribution according to threshold in order to prevent from reducing performance of data center due to exceeding VM aggregation. In addition, the scheme sets VM groups and limits the number of hops in VM migration by considering overhead of VM migration and connection between VMs providing the same service. The network controller controls (turns on or off) the buffers after it performs VM aggregation and distribution according to group constraint.

The propose scheme realizes green data center network with high performance of data center. Power control of HOPR buffer is explained. Next, VM aggregation and distribution considering VM groups are explained. Finally, VM aggregation, distribution, and buffer control are shown.

4.2.1 Power Control of HOPR Buffer

Figure 4.1 shows buffer control in the proposed scheme. The network controller confirms that all servers under a HOPR do not work, and turns off the HOPR buffer. The power consumption in buffer is two-thirds of that in HOPR. Therefore, it has great impact on power consumption to turn off the buffer. HOPR can forward packets even when the buffer is turned off. Conventional electrical switches in data center cannot forward packets when the buffer is turned off. Therefore, the network controller in HOPR-based data center can perform VM migration flexibly for aggregation of VM with low power consumption. This is because packets between VMs are transferred flexibly for VM migration in HOPR-based data center network. Figure 4.2 shows an example of VM migration. In the proposed scheme, the network controller configure Express Path between HOPRs for VM migration. Express Path means optical path in HOPR-based network, and occupies bandwidth between HOPRs. The network controller performs VM migration after configuring Express Path. The controller confirms that VMs under the HOPR which is the source of VM migration do not work, and turns off the HOPR buffer.

4.2.2 VM Aggregation and Distribution Considering VM Groups

Lower power consumption in data center network is expected by turning off buffers in the proposed scheme. However, performance of data center is degraded because higher load concentrates on servers under HOPR which buffer is turned on. Therefore, the propose scheme performs VM distribution as well as VM aggregation in order to prevent from avoiding exceeding VM aggregation. Figure 4.3 shows a policy of VM aggregation and distribution. The proposed scheme sets each threshold for VM aggregation and distribution, and makes a decision about VM aggregation and distribution. Th_a is the threshold for VM aggregation, and Th_d is the threshold for VM distribution. VM groups are defined

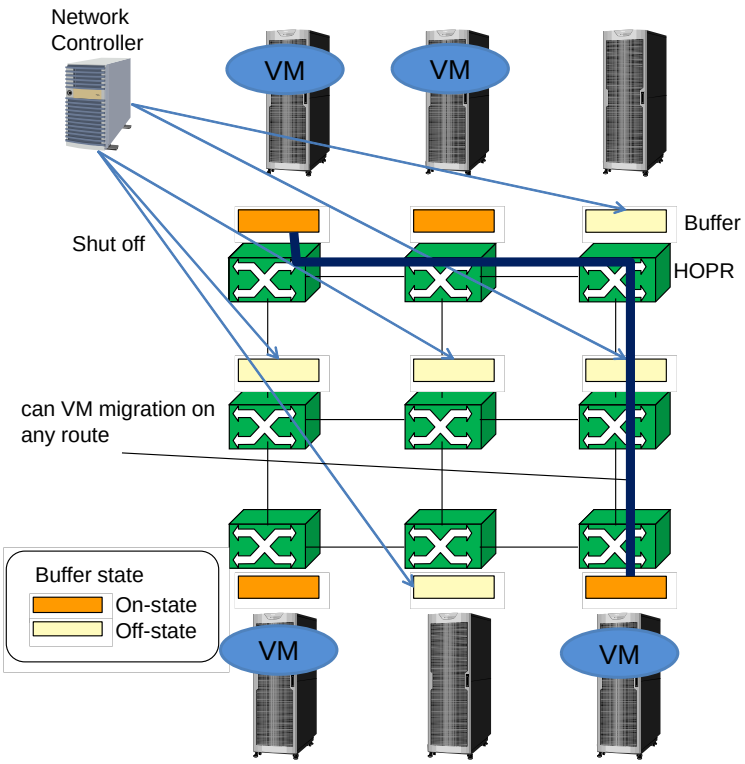

Figure 4.1. Buffer control in proposed scheme.

in order to set range of VM migration. The number of hops between VMs is limited by applying VM groups. The VM groups are explained as below.

The purpose of VM aggregation is power saving of data center network. From Fig. 4.3 (a), the proposed scheme performs VM migration for aggregation from PM_i (PM: Physical Machine), source of VM, to PM_j, destination of VM, when $O_{PM}(i)$, the operation rate of PM_i, is less than Th_a. The operation rate of PM_i is defined as

$$O_{PM}(i) = \sum_{j=1}^{N} O_{VM}(i, j)U(i, j).$$

N is the number of working VM in PM_i, $O_{VM}(i, j)$ is the operation rate of VM_j in PM_i, and $U(i, j)$ is the CPU usage of VM_j in PM_i. Table 4.1 shows parameters in the proposed scheme.

The purpose of VM distribution is performance retention of data center. From Fig. 4.3 (b), the proposed scheme performs VM migration for distribution from PM_i to PM_j when

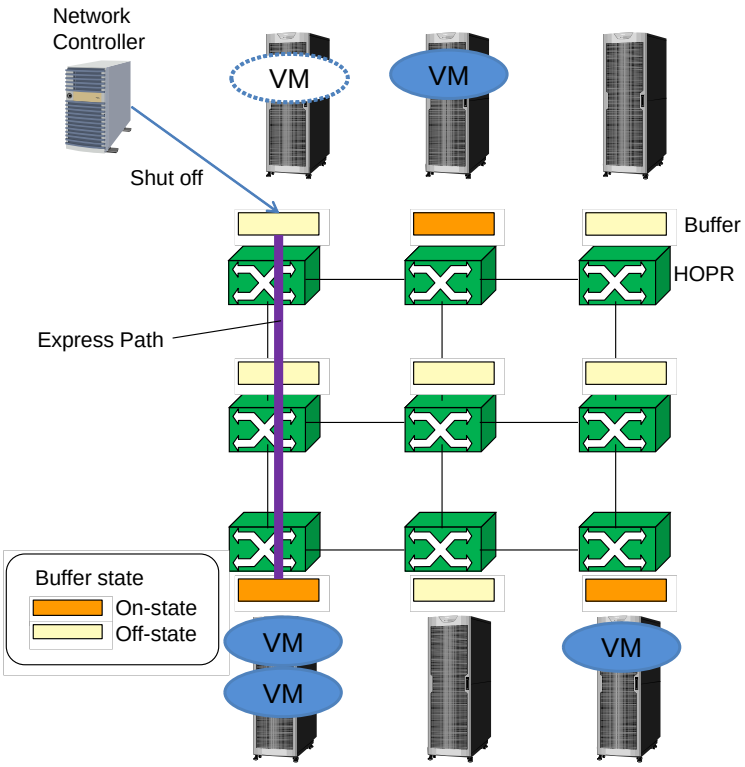

Figure 4.2. Example of VM migration.

$O_{PM}(i)$ is more than Th_d.

The proposed scheme defines VM groups in order to set range of VM migration. A groups of VMs realizes one service for a user. For example, VMs of Web server, Application server, and DB servers are defined as a VM group in case of the Web. The network controller limits the number of hops for VM migration by setting VM groups in order to

Table 4.1. Parameters in proposed scheme.

Th_a	Threshold for VM aggregation
Th_d	Threshold for VM distribution
$O_{PM}(i)$	Operation rate of PM_i
$O_{VM}(i, j)$	Operation rate of VM_j in PM_i
$U(i, j)$	CPU usage of VM_j in PM_i

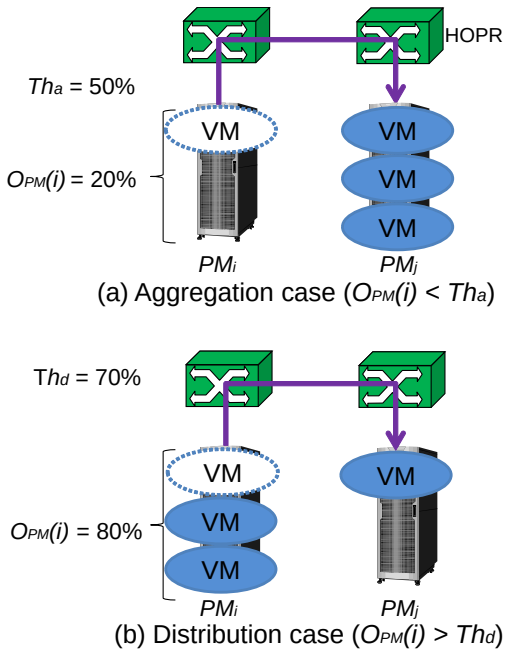

Figure 4.3. Policy of VM aggregation and distribution.

put VMs in near PMs. Figure 4.4 shows an example of VM groups. The number of hops in VM Group 1 is limited to 2, and that in VM Group 2 is limited to 3 for VM migration in Fig. 4.4. The proposed scheme can reduce link resource occupied by Express Path and delay for communications between VMs by limiting the number of hops.

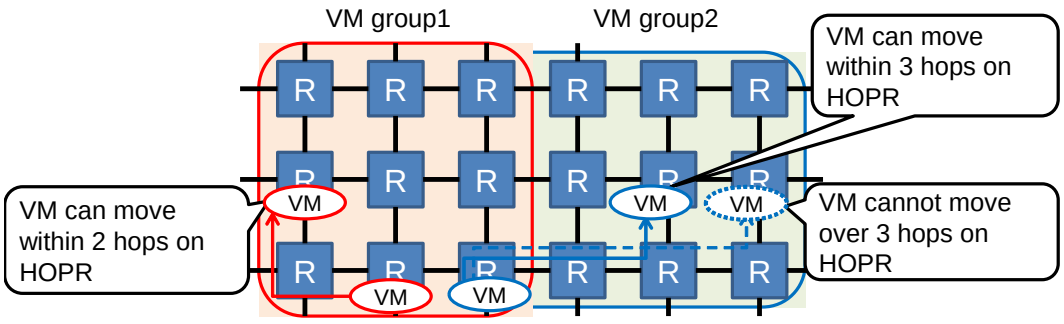

Figure 4.4. Example of VM groups.

4.2.3 VM Aggregation, Distribution, and Buffer Control

Figures 4.5 and 4.6 show flowcharts of the proposed scheme in VM aggregation and distribution. Here, the relationship between Th_a and Th_d is assumed to $Th_a < Th_d$. Each PM in the data center has a unique number, and Each VM in a PM has a unique number for management.

First, the operation of VM aggregation and buffer control is explained.

1. Determine if $O_{PM}(i)$ is less than Th_a. Search PM_j treating one or more working VMs in a VM group if $O_{PM}(i)$ is less than Th_a. Perform VM distribution if $O_{PM}(i)$ is more than Th_d (See Fig. 4.6).

2. Determine if $O_{PM}(j)$ is less than Th_d. Move all VMs in PM_i to PM_j if $O_{PM}(j)$ is less than Th_d. Here, $O_{PM}(j)$ shows the operation rate after VM aggregation. Search PM_j again in the VM group if $O_{PM}(j)$ is more than Th_d.

3. Check if there is working PM under HOPR where PM_i exists. Turn off a HOPR buffer if there is no working PM.

Next, the operation of VM distribution and buffer control is explained.

1. Determine if $O_{PM}(i)$ is more than Th_d. Search PM_j treating one or more working VM in a VM group if $O_{PM}(i)$ is more than Th_d. Process is finished if $O_{PM}(i)$ is less than Th_d.

2. Determine if $O_{PM}(j)$ is less than Th_d. Move VMs until $O_{PM}(i)$ is less than Th_d if $O_{PM}(j)$ is less than Th_d. Here, $O_{PM}(j)$ shows the operation rate after VM distribution.

3. Check if all PM in the VM group are applied for VM distribution if $O_{PM}(j)$ is more than Th_d. Return to Step 1 if there are PM not applied for VM distribution. Search PM_j again.

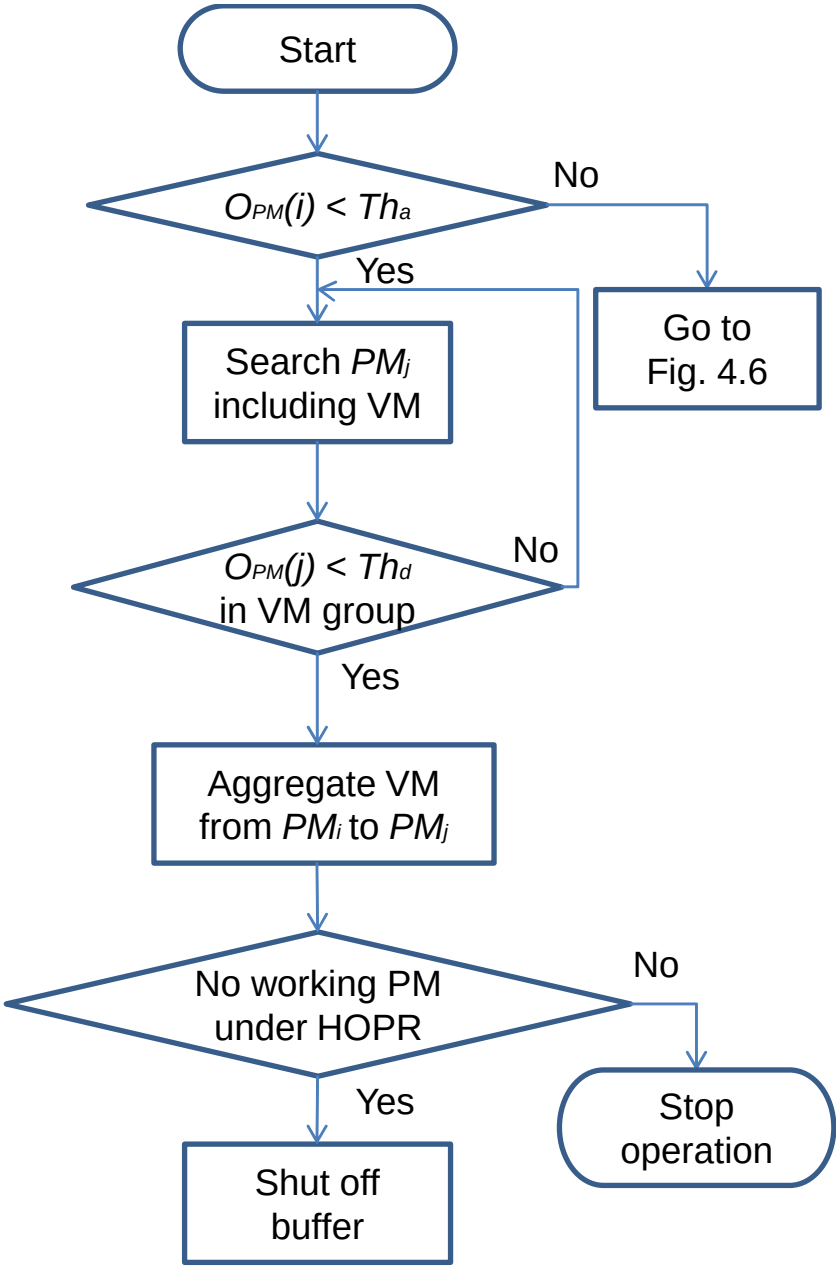

Figure 4.5. Flowchart of proposed scheme in VM aggregation.

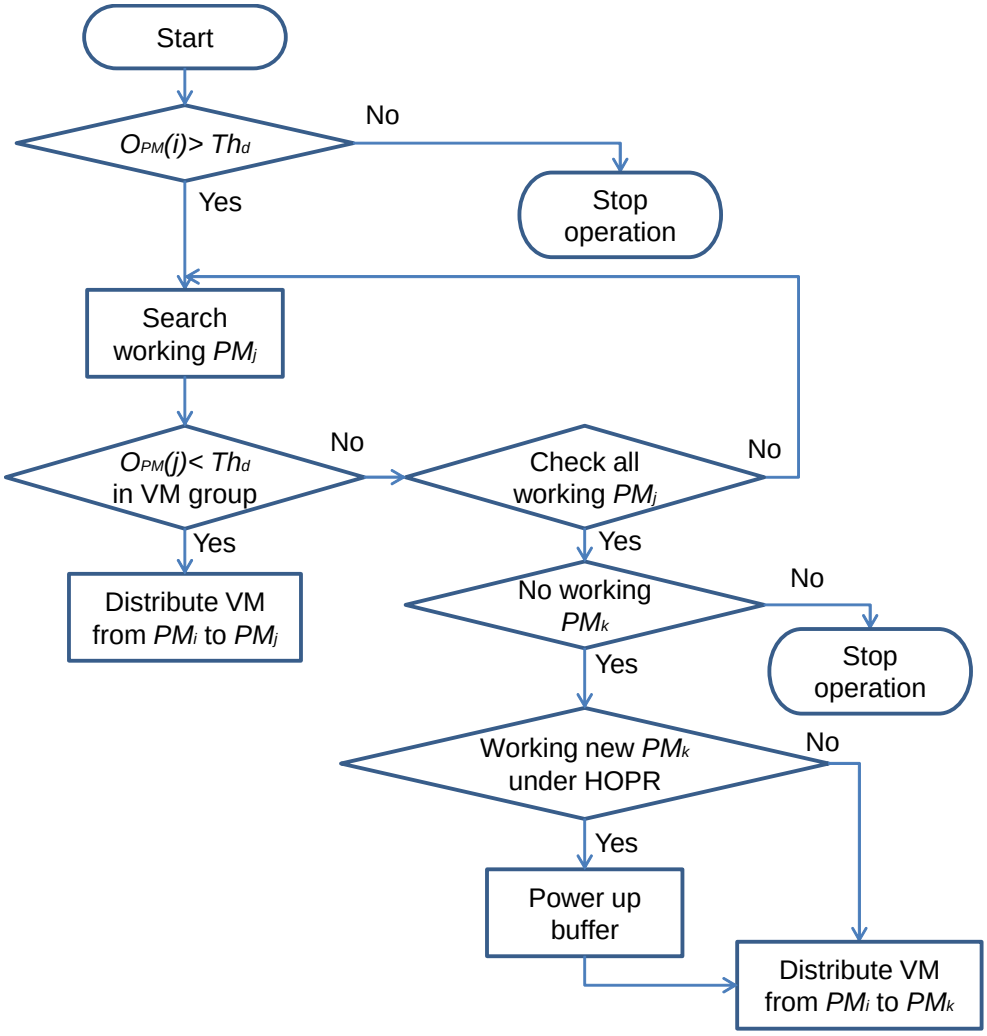

Figure 4.6. Flowchart of proposed scheme in VM distribution.

4. Check if there are non working PM, PM_k, in the VM group when all PM are applied for VM distribution.

5. Check status of PM which is under the HOPR with PM_k if PM_k exists. Turn on the HOPR buffer if there are no working PM under the HOPR.

6. Move VMs in PM_i to PM_k until $O_{PM}(i)$ is less than Th_d regardless the result of previous step.

The network controller searches PM in ascending order in the first step for easy imple-

mentation. The controller also searches VM in a PM in ascending order.

The proposed scheme maintains data center performance under green data center network by performing VM aggregation and distribution according to VM status.

4.3 Performance Evaluation

In this section, the power consumption in data center network and data center performance are evaluated in order to confirm that the proposed scheme realize green data center network under high performance of data center.

4.3.1 Evaluation Condition

Table 4.2 shows simulation parameters. Here, a 2D torus topology is used for the evaluation. The 2D torus topology has four input and output ports. The effect of green data center network is evaluated as the number of working buffers which occupy large part of power in HOPR. The number of responses per second in a PM is defined as performance of PM, and the total of each PM performance is defined as data center performance [4-14]. The performance of PM_i is defined as

$$P_{PM}(i) = \min(\sum_{j=1}^{N} P_{VM}(i, j), M).$$

Here, $P_{VM}(i, j)$ is the performance of VM_j in PM_i. N is the number of VMs in PM_i. M is the upper bound in performance of PM_i. Figure 4.7 shows the performance of VM versus operation rate of VM. The operation rate of VM is assumed to be the CPU usage of VM. the VM performance is defined as the number of responses per second as shown in Fig. 4.7. The VM performance are proportional to the operation rate of VM. However, the VM performance has upper bound like the PM performance.

Table 4.2. Simulation parameters.

Parameter	Value
HOPR (except Core Router)	100
ToR switch per HOPR	10
PM per ToR	10
Max VM per PM	4
Max performance in PM	2,500 response/sec
Max performance in VM	800 response/sec
VM group	50
Max hop in VM group	7
Max (Min) VM in VM group	50 (25)
Topology	2D torus

4.3.2 Evaluation about Power Saving

Figures 4.8 and 4.9 show the number of working buffer versus average operation rate of VM. Fig. 4.8 shows the number of working buffers when Th_d is 75%, and Fig 4.9 shows the number of working buffers when Th_d is 90%. Here, the proposed scheme is compared with a conventional scheme which does not turn off HOPR buffers regardless of VM status. From Fig. 4.8, the number of working buffers is reduced when Th_a increases. This is because HOPRs where no servers are working increases through the influence of VM aggregation when Th_a increases. The proposed scheme can reduce power consumption of HOPR-based data center network by up to 40% when the operation rate of VM is 50%, and the power consumption in buffer is two-thirds of that in HOPR. In this case, HOPRs whose buffer is working is 40% of the total HOPRs, and minimum value as shown in Fig. 4.8. Thus, the proposed scheme can turn off HOPR buffers by up to 60% in case of Fig. 4.8. Therefore, the scheme can reduce the power consumption of HOPR-based data center network by $60\% \times \frac{2}{3} = 40\%$. From Fig. 4.9, the number of working buffers is reduced when Th_a increases. In addition, the number of working buffers does not

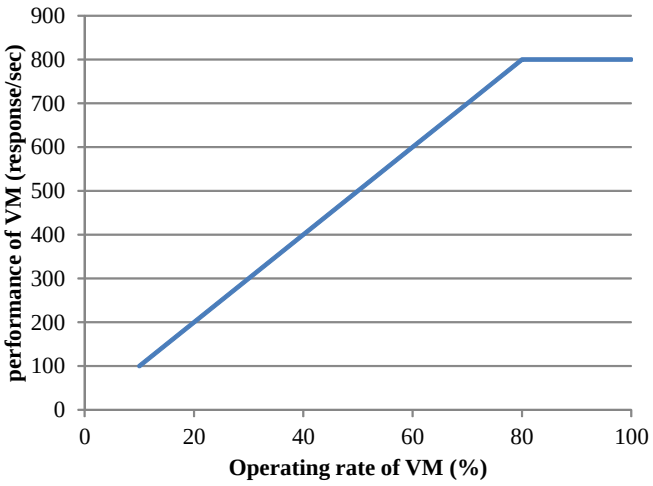

Figure 4.7. Performance of VM versus oparation rate of VM.

change even when Th_d changes as shown in Fig. 4.8 and Fig. 4.9. This is because the network controller performs VM distribution by searching HOPRs whose buffer is working preferentially.

4.3.3 Evaluation about Data Center Performance

Figures 4.10 and 4.11 show the data center performance versus average operation rate of VM. Fig. 4.10 shows the performance when Th_a is 25%, and Fig. 4.11 shows the performance when Th_a is 50%. From Fig. 4.10, the data center performance of the proposed scheme is going on that of the conventional scheme when Th_d decreases. Especially the data center performance of the proposed scheme is the same as that of the conventional scheme when Th_d is 60%. This is because the proposed scheme can perform VM distribution before the total VM performance per PM reaches the upper bound of the PM performance. From Fig. 4.11, the data center performance of the proposed scheme is going on that of the conventional scheme when Th_d decreases. In addition, the data center performance does not change even when Th_a changes as shown in Fig. 4.10 and Fig.

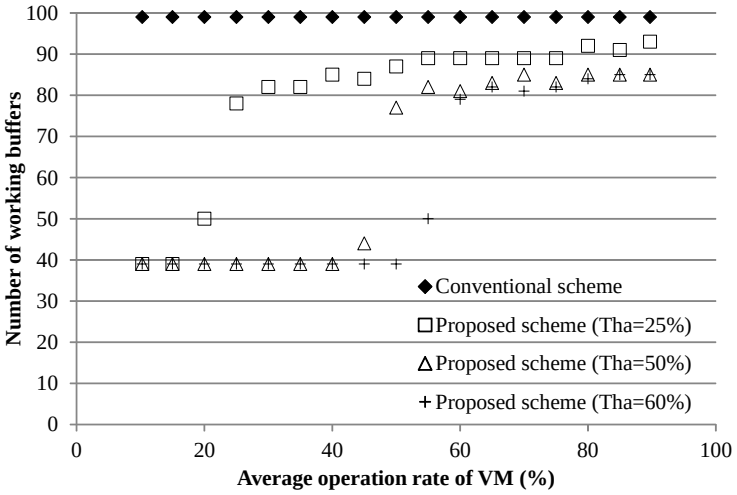

Figure 4.8. Number of working buffer versus average operation rate of VM (Th_d = 75%).

4.11. This is because the total VM performance per PM does not reach the upper bound of the PM performance even if the proposed scheme performs VM aggregation.

The proposed scheme can realize green data center network with high performance of data center by increasing the threshold for VM aggregation in case of lower operation rate of VM and decreasing the threshold for VM distribution in case of higher operation rate of VM.

4.3.4 Effect of Traffic by VM Migration

The effect of traffic by VM migration is evaluated in order to confirm effect of the proposed scheme.

The proposed scheme occupies bandwidth between HOPRs by using Express Path during VM migration. Therefore, the effect of traffic by VM migration is evaluated. Table 4.3 shows parameters about VM for evaluation of the effect. Figure 4.12 shows the occupation rate of link between HOPRs versus average operation rate of VM. The occupation

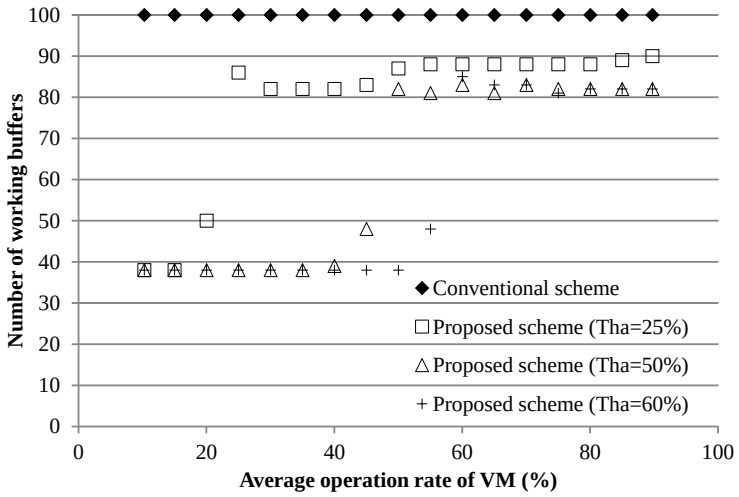

Figure 4.9. Number of working buffer versus average operation rate of VM (Th_d = 90%).

rate of link between HOPRs is define as

$$C = \frac{\sum_{i=0}^{O} h(i)t_{mig}(i)}{\sum_{i=0}^{O} t_{mig}(i)L} \times 100.$$

Here, O is the number of VM migration, $h(i)$ is the number of hops between HOPRs in ith migration, $t_{mig}(i)$ is time of ith migration, and L is the total number of links between HOPRs. The numerator is the total time of link occupation, and the denominator is total elapsed time of all links in VM migration. Each VM migration is performed one by one, more than two VM migrations are not performed simultaneously.

From Fig. 4.12, the occupation rate in the proposed scheme is less than 1.6%, and very low rate. If the permitted occupation rate is 16% for VM migration, the proposed scheme can perform up to ten VM migrations. The occupation rate for VM migration needs to be set according to status of data center network.

The proposed scheme does not have a effect on data center traffic when each VM migration is performed one by one. The number of VM migrations needs to be controlled according to status of data center network, the control scheme is a future work.

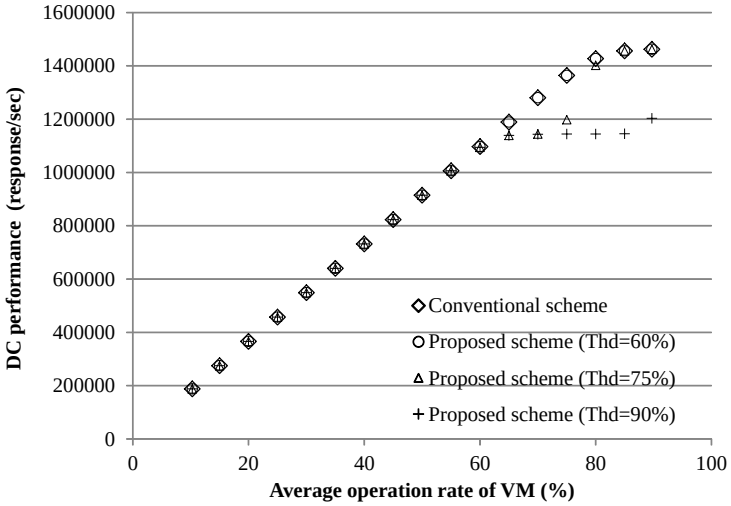

Figure 4.10. Data center performance versus average operation rate of VM (Th_a = 25%).

Table 4.3. Parameters about VM.

Parameter	Value
VM size (MB)	128,256,512,1024,2048,4096
VM ratio	1/15,2/15,3/15,4/15,3/15,2/15

4.3.5 Effect by VM Group

In this subsection, the effect of the number of VM and the maximum number of hops in VM migration are evaluated.

The proposed scheme defines VM groups in order to set the range of VM migration. We confirm how the number of VM groups affects power saving and link occupation between HOPRs. Here, Each VM in a VM group is put on PM randomly. Table 4.4 shows the parameters about VM group, the number of VMs is 2500.

Figure 4.13 shows the number of working buffer and occupation rate of link between HOPRs versus the number of VM groups. From Fig. 4.13, the proposed scheme should set the number of VM groups one when the proposed scheme only reduces the number

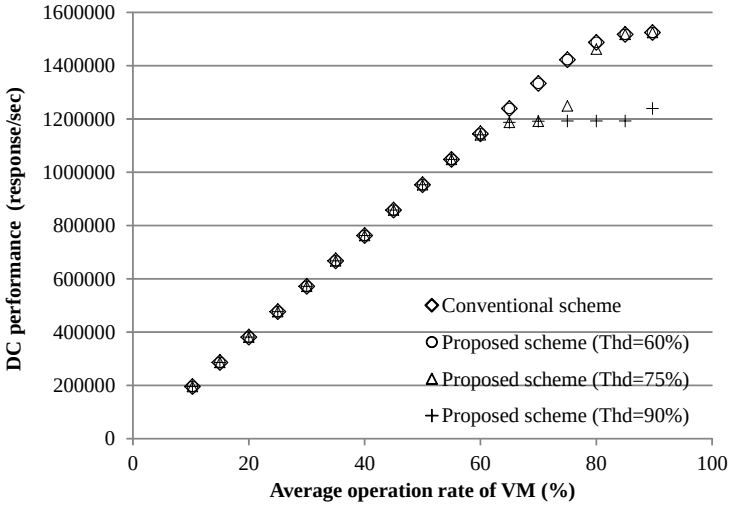

Figure 4.11. Data center performance versus average operation rate of VM (Th_a = 50%).

of working buffers. This is because it is likely to perform VM aggregation with larger area for VM migration. However, the occupation rate increases when the number of VM groups is one. Therefore, the number of VM groups needs to be set in order to keep a balance between the number of working buffers and the occupation rate. The data center performance in the proposed scheme does not differ in the number of VM groups, and the performance is the same as that in the conventional scheme. This is because the total VM performance per PM does not reach the upper bound of the PM performance even if the proposed scheme performs VM migration.

Next, we confirm the relationship between power saving and link occupation between HOPRs by changing the number of hops for VM migration. Here, the number of VM groups is 50, and the number of VMs in a VM group is 50.

Figure 4.14 shows the number of working buffer and occupation rate of link between HOPRs versus the number of maximum hops in VM migration. From Fig. 4.14, the proposed scheme should set the number of maximum hops higher when the proposed scheme only reduces the number of working buffers. This is because it is likely to perform VM aggregation with larger area for VM migration. However, the occupation rate increases

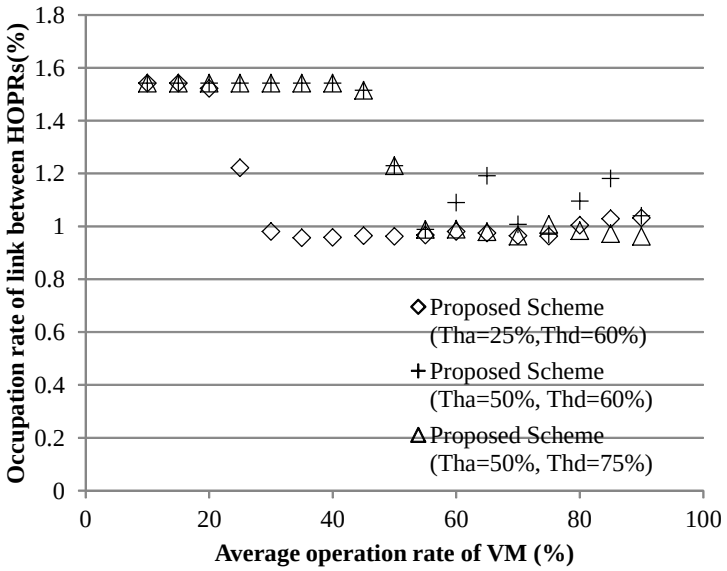

Figure 4.12. Occupation rate of link between HOPRs versus average operation rate of VM.

when the number of maximum hops is higher. Therefore, the number of maximum hops needs to be set in order to keep a balance between the number of working buffers and the occupation rate. The data center performance in the proposed scheme does not differ in the number of maximum hops, and the performance is the same as that in the conventional scheme. This is because the total VM performance per PM does not reach the upper bound of the PM performance even if the proposed scheme performs VM migration.

The proposed scheme sets the number of VM groups lower and the number of maximum hops higher when the proposed scheme only consider power saving. However, VM migration needs to be performed according to the status of traffic in data center network. Therefore, the number of VM groups and maximum hops need to be set in order to keep a balance between power saving and link occupation.

Table 4.4. Parameters about VM group.

Parameter	Value
Th_a	50%
Th_d	60%
VM size (MB)	128,256,512,1024,2048,4096
VM ratio	1/15,2/15,3/15,4/15,3/15,2/15
Number of VMs	2500
Average operation rate of VM	40%

Table 4.5. Power consumption of devices (servers: 2500).

Device	W
Router	960
HOPR	120
HOPR w/o buffer	40
Server	500
Server w/ sleep mode	10

4.3.6 Evaluation of Power Consumption in Intra Data Center Network

This subsection compares power consumption in intra data center networks. Figure 4.15 shows comparison of network power consumption in the intra data center networks. Figure 4.16 shows comparison of server power consumption in the intra data center networks. Table 4.5 shows power consumption of devices. From Fig. 4.15, proposed HOPR can reduce network power consumption by 40% as compared to conventional HOPR. From Fig. 4.16, proposed HOPR can reduce server power consumption by 59% as compared to conventional HOPR. This is because proposed HOPR realizes energy-efficient intra data center network by controlling routers and VMs under electrical buffer situation.

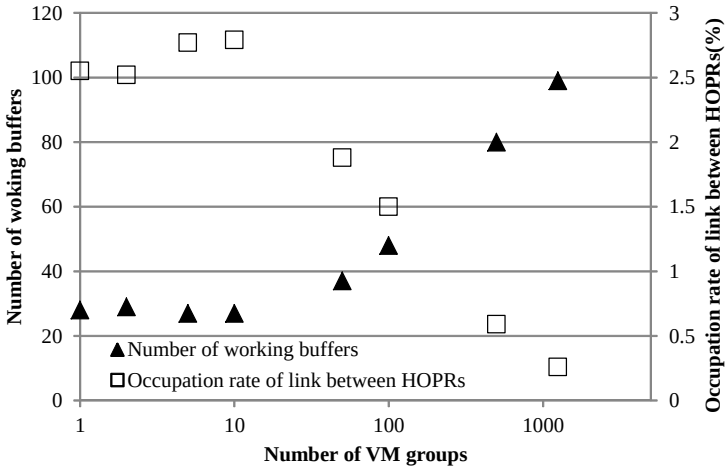

Figure 4.13. Number of working buffer and occupation rate of link between HOPRs versus number of VM groups.

4.4 Conclusion

Chapter 4 proposed optical energy-efficient intra data center with buffer and VM control. The proposal for high energy efficiency based on the optical data center minimizes working buffers and servers according to VM assignment based on threshold in VM groups under high data center performance. The VM assignment had VM aggregation for power saving and distribution for high data center performance. The proposed scheme reduced network power consumption by 40% and server power consumption by 59% as compared to the conventional HOPR-based data center.

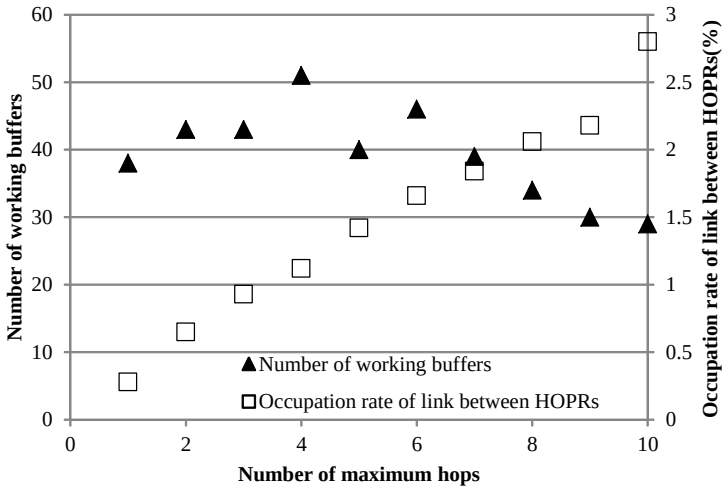

Figure 4.14. Number of working buffer and occupation rate of link between HOPRs versus number of maximum hops in VM migration.

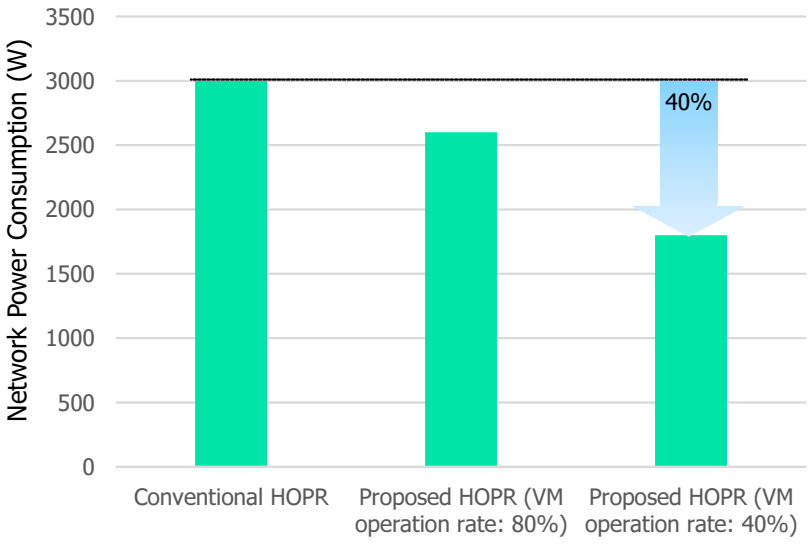

Figure 4.15. Comparison of network power consumption in intra data center networks.

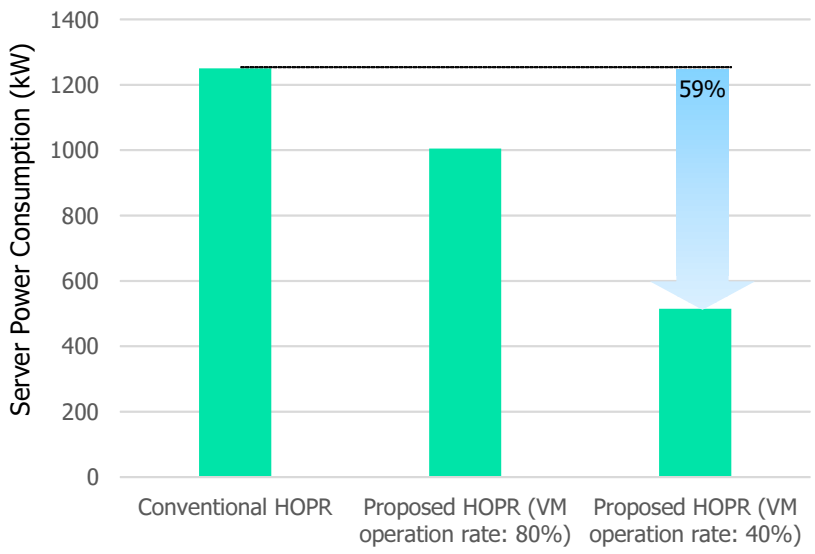

Figure 4.16. Comparison of server power consumption in intra data center networks.

References

[4-1] M. Hayashitani, K. Suzuki, and Y. Mizukoshi, “A Study on Relationship between Data Center Performance and Network Power Consumption with Buffer Management in HOPR-based Data Center,” in Proc. iPOP (IP + Optical Network), no.T4-1, Okinawa, Japan, Apr. 2015.

[4-2] M. Hayashitani, K. Suzuki, and N. Yamanaka, “Control Scheme of HOPR-Based Data Center Network by Considering VM Situation,” IEICE Trans. Commun, (Japanese Edition), vol.J99-B, no.4, pp. 334-344, Apr. 2016.

[4-3] Cisco Nexus 7000 Series,
http://www.cisco.com/c/en/us/td/docs/switches/datacenter/hw/nexus7000/installation/guide/n7k_hig_book.html

[4-4] P. K. Pepeljugoski, J. A. Kash, F. Doany, D. M. Kuchta, L. Schares, C. Schow, M. Taubenblatt, B. J. Offrein, and A. Benner, “Low Power and High Density Optical Interconnects for Future Supercomputers,” In Proc. OFC, no.OThX2, San Diego, CA, USA, 2010.

[4-5] C. Kachris, K. Kanonakis, and I. Tomkos, “Optical interconnection networks in data centers: Recent trends and future challenges,” IEEE Commun. Mag., vol.51, no.9, pp.39-45, Sept. 2013.

[4-6] D. T. Neilson, “Photonics for switching and routing,” IEEE J. Sel. Topics Quantum Electron., vol.12, no.4, pp.669-678, July/Aug. 2006.

[4-7] W. Zhang, H. Wang, and K. Bergman, “Next-generation optically-interconnected high-performance data centers,” IEEE/OSA J. Lightw. Technol., vol.30, no.24, pp.3836-3844, Dec. 2012.

[4-8] H. Mehrvar, H. Ma, X. Yang, Y. Wang, S. Li, A. Graves, D. Wang, H. Y. Fu, D. Geng, D. Goodwill, and E. Bernier, “Photonic switching of native ethernet frames for data centers,” in Proc. Photonics Switching, no.JT5C.2, San Diego, CA, USA, 2014.

[4-9] Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella, and S. J. B. Yoo, “LIONS: An AWGR-based low latency optical switch for high-performance computing and data centers,” IEEE J. Sel. Topics Quantum Electron., vol.19, no.2, article 3600409, Mar./Apr. 2013.

[4-10] J. Gripp, J. E. Simsarian. J. D. LeGrange, P. Bernasconi, and D. T. Neilson, “Photonics terabit routers: The IRIS project,” in Proc. OFC, no.OThP3, San Diego, CA. USA, 2010.

[4-11] K. Kitayama, Y. Huang, Y. Yoshida, R. Takahashi, T. Segawa, S. Ibrahim, T. Nakahara, Y. Suzaki, M. Hayashitani, Y. Hasegawa, Y. Mizukoshi, and A. Hiramatsu, “Torus-Topology Data Center Network Based on Optical Packet/Agile Circuit Switching with Intelligent Flow Management,” IEEE/OSA J. Lightw. Technol., vol.33, no.5, pp.1063-1071, Mar. 2015.

[4-12] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, “ElasticTree: Saving Energy in Data Center Networks,” in USENIX NSDI, April 2010.

[4-13] V. Mann, P. Dutta, S. Kalyanaraman, and A. Kumar, “VMFlow: Leveraging VM Mobility to Reduce Network Power Costs in Data Centers,” NETWORKING 2011. Springer Berlin Heidelberg, 2011.

[4-14] IBM Knowledge Center, http://www-01.ibm.com/support/knowledgecenter/ssw_i5_54/rzamy/prftuneqtip.html

Chapter 5

Multiple Service Protection in Optical Based Intra Data Center Network

Chapter 2 introduced intra data center network technologies about reliability were illustrated as well as the intra data center network about power saving. The conventional approaches for reliability in intra data center have problems in resource utilization and power consumption, and even if the approaches satisfy the resource utilization, they have problem in reliability effect. Chapter 5 proposes multiple service protection in the optical base intra data center network [5-1, 2]. The proposal for high reliability based on the optical based intra data center suspended low-priority service rapidly on the failure notification of bidirectional sides from the failure detecting node. The source node can switch to backup path when the node receives the failure notifications from the bidirectional sides because all the node suspends low-priority services. The proposed scheme can suppress total recovery time of middle-priority service to constant level regardless of low-priority traffic condition.

5.1 Chapter Introduction

In the present intra data center network, electrical switches and routers are usually used. Unfortunately, the power consumption of these devices increases rapidly as the traffic within the data center network increases. Therefore, optical technologies which can realize power savings and larger scale networks are needed if the data center is to handle

the increase in traffic. This dissertation focuses on raising the data center performance and reliability of the intra data center network while realizing power savings multiple services including mission critical services.

In the intra data center network, multiple-class traffic is considered in accordance with the demand of each class of traffic. In general, 1+1 protection is best for high-priority traffic because it achieves high-speed protection by always using backup paths for data transmission, and destination nodes only switch to backup paths in the case of a failure. Mission critical and broadcasting services, for example, are classed as high-priority class. In general, 1:1 protection with low-priority traffic is best for middle-priority traffic [5-3, 4, 5, 6, 7]. In 1:1 protection, backup paths are unoccupied, so low-priority traffic can be transmitted on the backup paths. The 1:1 protection scheme can thus achieve a good balance between high-speed transmission and efficient path utilization. Enterprise services, for example, are classed as middle-priority traffic. Low-priority traffic is not protected and is replaced with backup path traffic in case of the failure. The Internet is classed as low-priority traffic [5-8]. The low-priority traffic is removed and replaced with backup path traffic in case of network failure and has no backup path. If a network fault occurs, the low-priority traffic is suspended, and the middle-priority traffic is switched to the backup path. A scheme suspending low-priority traffic enables the network to carry low-priority traffic between nodes[5-9, 10], but the process for requesting acknowledgments and suspension makes it difficult to provide high-speed protection.

Therefore, a reliable intra data center network is proposed. The proposed intra data center network suspends low-priority service rapidly on a failure notification of bidirectional sides from a failure detecting node.

The rest of this chapter is organized as follows. The proposed protection scheme is described in Section **5.2**, and performances are discussed in Section **5.3**. The reliability with high energy efficiency is discussed in Section **5.4**. This chapter is concluded in

Section **5.5**.

5.2 Proposed Multi-service Protection Scheme

In response to the problems associated with the conventional scheme [5-9, 10], a protection scheme that rapidly suspends low-priority traffic is developed. Each node has a path-information table that includes the primary paths transmitted by the node or passing through it, the corresponding backup paths, and low-priority traffic transmitted by the node. If a node adjacent to a failure point detects a link failure through the optical loss of the control signal in the physical layer, it checks its path-information table to see if any impaired primary paths are transmitted by the detecting node or passing through it. If there is no impaired primary path in the path-information table, the node stops operation. If there are any impaired paths in the path-information table, the node sees if it is transmitting low-priority traffic on the corresponding backup path. If it is transmitting, it suspends transmission of that traffic. It also sends a failure notification to the source node of the faulty primary path through intermediate nodes in both directions (clockwise and counterclockwise). The control signal for this notification includes the location of the failed link. Figure 5.1 shows the action taken by the failure-detecting node in the proposed scheme.

Each node that receives the notification confirms the location information of a failed link. It checks the path information table and confirms the impaired primary path. And, it sees if it is transmitting low-priority traffic on the corresponding backup path. If it is, it suspends transmission of that traffic. When the node is the source node of impaired primary path, the node sees if it receives the failure notification in both directions. If it receives in both directions, it switches the primary path to the corresponding backup path after verifying that the nodes transmitting low-priority traffic have suspended transmission. If the source node does not receive in both directions, it waits to receive the failure

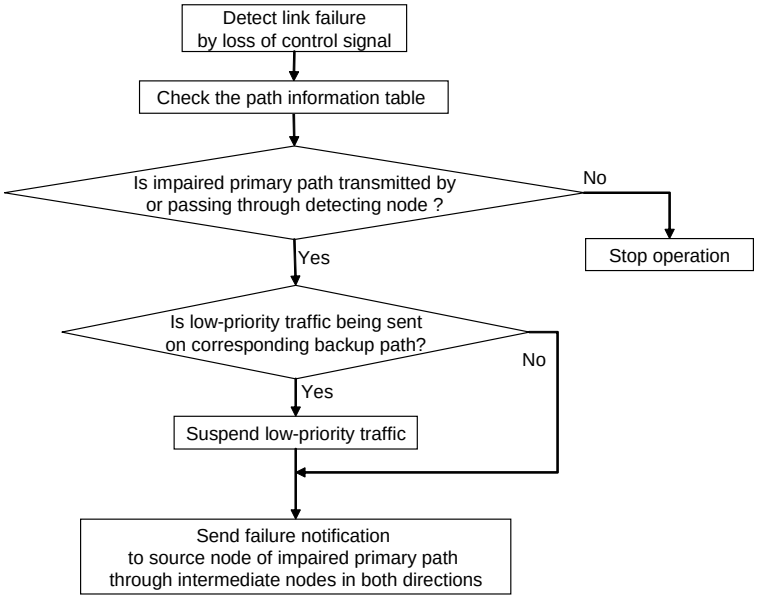

Figure 5.1. Actions taken by failure-detecting node in proposed scheme.

notification in both directions. The destination node receives data from the backup path, and protection is completed. Figure 5.2 shows the action taken by the nodes that received the failure notification.

Next, specific operations are discussed in the proposed scheme. Figure 5.3 shows an example of path protection in the proposed scheme. A link failure has occurred between Nodes B and C. The solid arrows from each node indicate the flow of the signal on the control channel, the two-dots chain (– -) arrow from Node A to Node E indicates the flow of the signal on the data channel, and the heavy line on each node indicates the processing time for each operation. Here, Nodes A, B, C, D, and E have information about the primary path between Nodes A and E and the corresponding backup path in each node's own path-information table, and Nodes B, D, F, and H have information about low-priority traffic transmitted by each node. The path-information table for Node D is shown in Fig. 5.3 as an example.

Node C detects a link failure and then confirms that the node is not transmitting low-priority traffic on the backup path by checking the path-information table. It then sends a

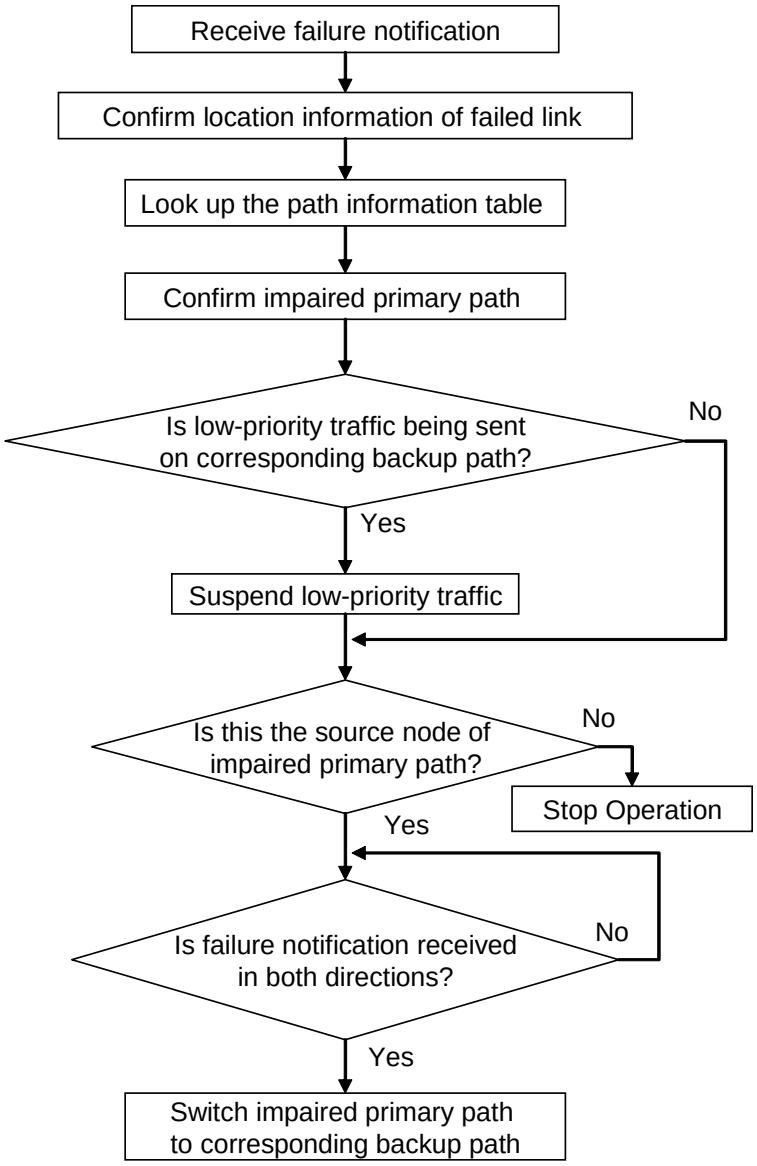

Figure 5.2. Actions taken by nodes receiving failure notification in proposed scheme.

failure notification to Node A, i.e., the source node of the primary path in the clockwise direction, through intermediate nodes in both directions. Nodes B, D, F, and H receive the failure notification and then determine whether or not they are transmitting low-priority traffic on the backup path by checking both their path-information table and the location information on the failed link included in the failure notification. Any nodes transmitting low-priority traffic immediately suspend transmission. For example, Node D confirms that the primary path between Nodes A and E is impaired by referring to the location of a failed link, i.e., the link between Nodes B and C in the clockwise direction. It also checks if the node is transmitting low-priority traffic on the corresponding backup path. The low-priority traffic from Nodes D to F is being transmitted on the backup path between Nodes E and A in the clockwise direction. Therefore, Node D suspends this low-priority traffic. Nodes E and G receive the failure notification and they then confirm that no nodes are transmitting low-priority traffic on the backup path by referring to the failure notification and their own path-information table. Node A receives the first failure notification in the counterclockwise direction and then confirms that it is not transmitting low-priority traffic by referring to the failure notification and the path-information table. And the node receives the second failure notification in the clockwise direction and confirms that the nodes transmitting low-priority traffic have suspended transmission. The node then switches the primary path to a corresponding backup path. Node E, i.e., the destination node of the primary path in the clockwise direction, receives the data from the backup path, and protection is completed.

Figure 5.4 shows another example of path protection in the proposed scheme. A span failure has occurred between Nodes B and C. Each node immediately detects each link failures through the loss of the control signal. Node B sends a failure notification to Node E, the source node of the primary path in the counterclockwise direction, and Node C sends a failure notification to Node A, i.e., the source node of the primary path in the

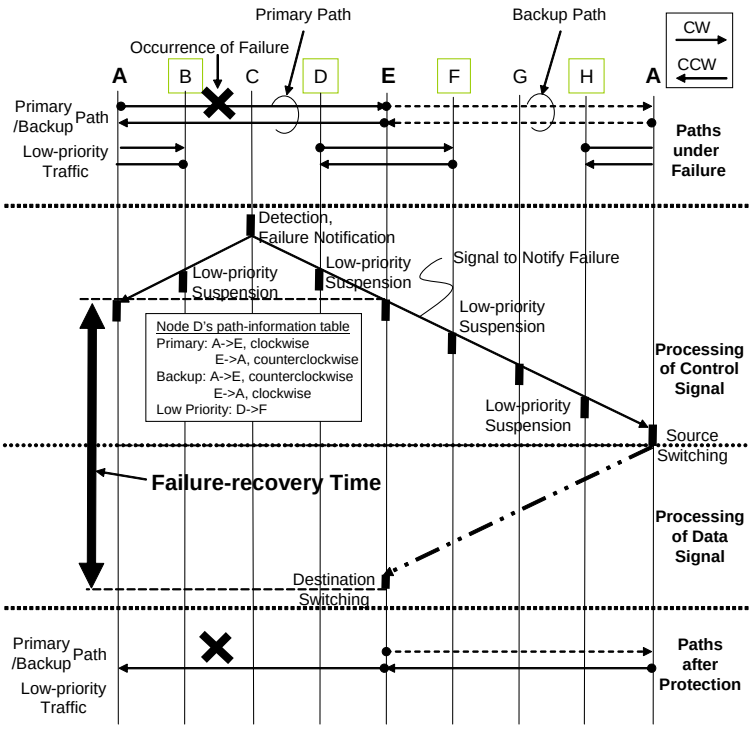

Figure 5.3. Example of path protection in proposed scheme (link failure).

clockwise direction. In Fig. 5.3, Node A receives two failure notifications from Node C in both directions, while in Fig. 5.4, Node A receives the failure notification from Node B in the counterclockwise direction, and Node A receives the other failure notification from Node C in the clockwise direction. In the case of span failure, the source node of the impaired primary path receives two failure notifications from two separate nodes; however, the source node of the impaired primary path receives the second notification with the same timing after a failure regardless of whether it is a link or span failure. For example, Node A receives the second notification from Node C with the same timing, as seen in Figs. 5.3 and 5.4. Therefore, the proposed scheme achieves an equal failure recovery time regardless of whether it is a link or span failure.

Every node in the proposed scheme suspends low-priority traffic by autonomously referring to the failure notifications and to their own path-information tables. A control signal for failure notification is sequentially and immediately sent from the node that de-

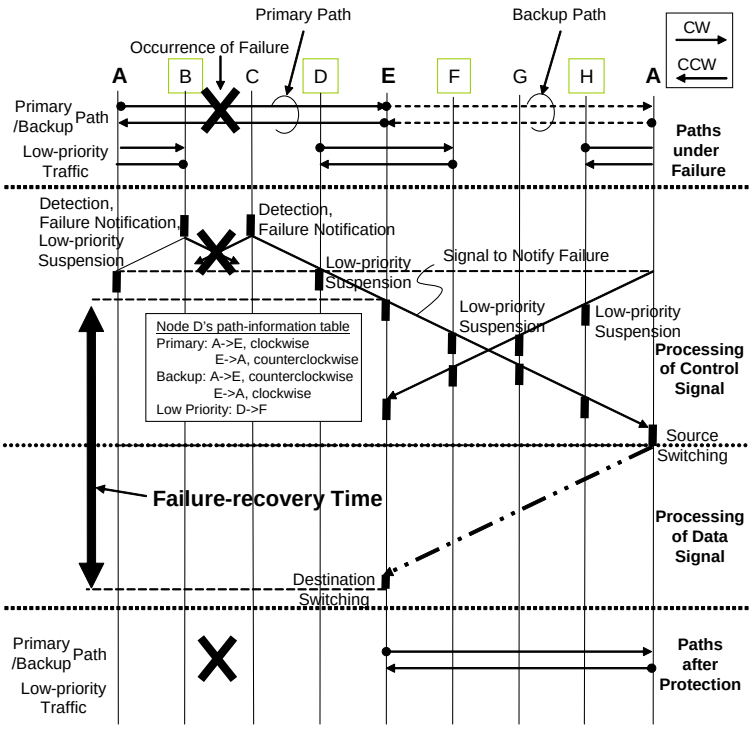

Figure 5.4. Example of path protection in proposed scheme (span failure).

tects the failure to the source node of the primary path through intermediate nodes, and each node rapidly suspends low-priority traffic. As a result, the proposed scheme provides high-speed protection. In addition, the amount of control signal traffic is lower than with the conventional scheme because only a control signal relating to failure notification is needed. The conventional and proposed schemes are compared in Table 5.1.

5.3 Performance Evaluation

Computer simulation is used in order to evaluate the failure-recovery time with both schemes. The failure-recovery time is defined as the period of time in which a destination node cannot receive data when failure occurs. The network topology was a two-fiber WDM ring with 159 data channels and 1 control channel. The data signal was transmitted on the basis of an optical transport network (OTN) frame defined in ITU standard G.709.

Table 5.1. Comparison of conventional and proposed schemes.

	Conventional Scheme [5-9, 10]	Proposed Scheme
Failure-detecting Node	Destination Node	Node adjacent to Failure Point
Failure Detection	Loss of Data Signal	Loss of Control Signal
Traffic-suspension Trigger	Receive Request to Suspend	Receive Failure Notification and Check Path-information Table
Operation after Suspension	Send ACK of Suspension	None
Path-switching Trigger	Receive Request for Path Switching	Receive Failure Notification
Control-signal Type	Request to Suspend, ACK of Suspension and Request for Path Switching	Failure Notification only

The data-channel bandwidth per wavelength was 10 Gbps, and the control-channel bandwidth was 155 Mbps. A control signal was transmitted on the control channel in the same way packets are forwarded. The packet in the control signal included the source address, destination address, type of packet, and location of the failed link, and the packet size is 10 bytes. The packet types were “Request for suspension,” “ACK of suspension,” “Request for path switching,” or “Failure notification.” Location information on the failed link was used only in the proposed scheme. The process time in each node is set as below. It is assumed that the process time in each node is much shorter than the propagation delay in each link. If a link distance is more than 20 km, the propagation delay in each link is over 100 μs. The table lookup time of one path information is 10 ns. The low-priority traffic suspension time is 100 ns, and the control packet creation time is 100 ns. The control

packet is terminated in each node, and O/E/O conversion is needed. The control packet delay time in node is 1 μs. It is assumed that there would be two types of failures, i.e., link and span failures.

A full-mesh path configuration is considered because it is expected to be dominant in the near future due to the rapid increase in P2P traffic. Full-mesh primary paths were set up and assigned odd wavelengths (λ1, λ3,$\cdots$). Corresponding backup paths were set up in the counter direction of the primary paths, and assigned even wavelengths (λ2, λ4,$\cdots$). For example, the primary path from Node A to Node E was set up in the clockwise direction, and the corresponding backup path was set up in the counterclockwise direction. All low-priority traffic was one-hop traffic. Hub-and-spoke path configurations are typical in current WDM ring networks, so in addition to the full-mesh path configuration, a hub-and-spoke configuration is investigated.

5.3.1 Full-mesh Path Configuration

First, the failure-recovery time versus the proportion of low-priority traffic is investigated. The proportion of low-priority traffic is defined as the proportion of wavelength links used for low-priority traffic to all links with wavelengths assigned for backup paths. Let's consider the case in which there are eight nodes, the number of wavelength links assigned for backup paths is 10, and four links are used for low-priority traffic for each wavelength. The number of links with the wavelength assigned for backup paths is 80 (10x8), and the number of wavelength links used for low-priority traffic is 40 (10x4). Therefore, the proportion of low-priority traffic is 0.5 (40/80).

Figure 5.5 plots the failure-recovery time versus the proportion of low-priority traffic in eight nodes. The link distance is 20 km. from Fig. 5.5, the proposed scheme provides high-speed protection compared with the conventional approach when the proportion of low-priority traffic exceeds 0.125. As this proportion increases, nodes far from the desti-

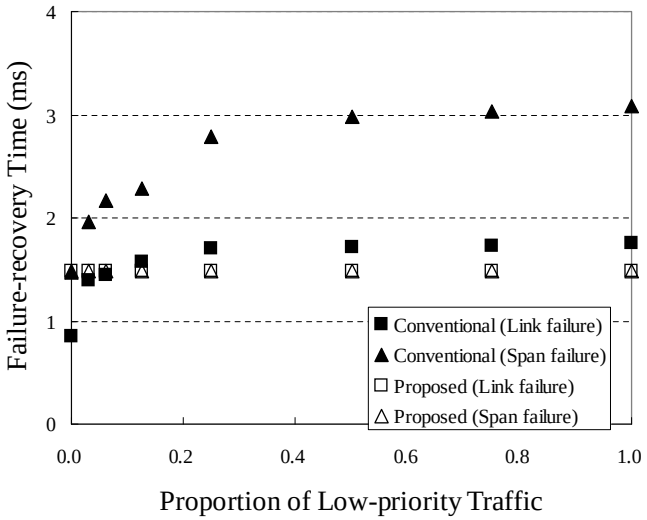

Figure 5.5. Failure-recovery time versus proportion of low-priority traffic in eight nodes.

nation node are more likely to transmit low-priority traffic. Therefore, it takes a substantial amount of time to start path-switching operations with the conventional scheme. In addition, with the conventional scheme, the failure-recovery time following span failure is longer than that following link failure. This is because requests to suspend low-priority traffic and switch path cannot be sent directly due to span failure. Therefore, the requests are indirectly sent in span failure, and the failure-recovery time of span failure is longer than that of link failure. We also found the failure-recovery time increased as the proportion increased in the conventional scheme. As the proportion increased, the number of nodes transmitting low-priority traffic increased. Therefore, the number of processes for suspending low-priority traffic and receiving acknowledgments increased. Consequently, the delay caused by the processes increased the failure-recovery time.

We found that the proposed scheme achieved the same performance even if the proportion of low-priority traffic changed due to link or span failure. In this scheme, if a node detects a failure, it sends a failure notification to the source nodes of the primary paths through intermediate nodes in both directions, and the operation time for this is the same

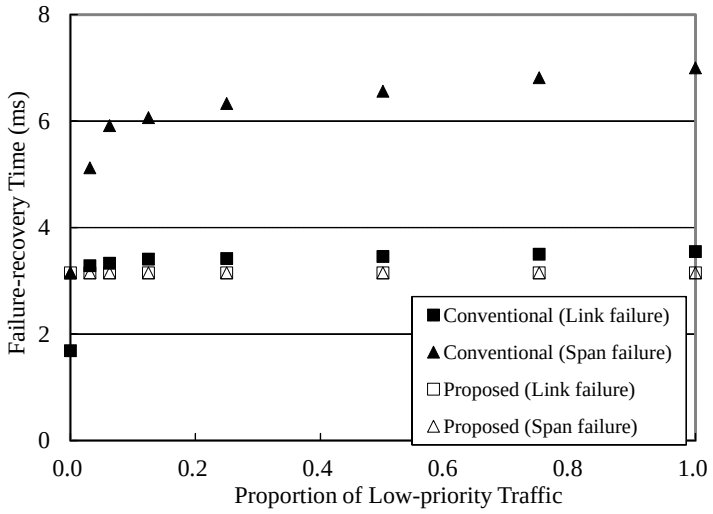

Figure 5.6. Failure-recovery time versus proportion of low-priority traffic in 16 nodes.

regardless of the failure type or the proportion.

Figures 5.6 and 5.7 plot the failure-recovery time versus the proportion of low-priority traffic in 16 nodes and 24 nodes, respectively. Figures 5.5 to 5.7 show that when the number of nodes was large, the failure-recovery time with the conventional scheme was more likely to increase as the proportion of low-priority traffic increased. In this case, the proportion of low-priority traffic was more than 0.125. As the number of nodes increased, the number of wavelengths used for low-priority traffic drastically increased, thereby increasing the delays in both the suspension of low-priority traffic and the receiving of acknowledgments.

Figure 5.8 plots the failure-recovery time versus the link distance. There were eight nodes, and the proportion of low-priority traffic was 100%. The differences in the failure-recovery times between the schemes increased with the link distance because the transmission delay for suspension requests and acknowledgments increased more for the conventional scheme than the proposed one.

Figure 5.9 plots the failure-recovery time versus the number of nodes. The link distance was 20 km, and the proportion of low-priority traffic was 100%. The failure-recovery time

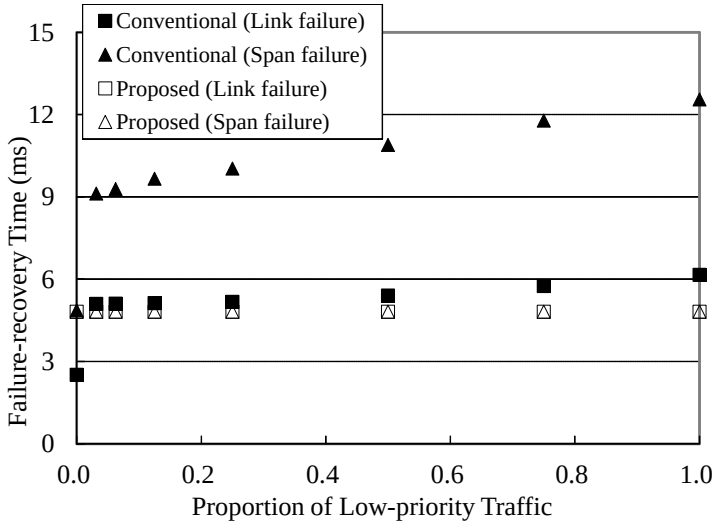

Figure 5.7. Failure-recovery time versus proportion of low-priority traffic in 24 nodes.

with the conventional scheme increased exponentially with the number of nodes, while with the proposed scheme it increased linearly. With both schemes, the transmission delays for the control and data signals increased linearly with the number of nodes. Since the amount of low-priority traffic drastically increased with the number of nodes, the process delay in suspending low-priority traffic and receiving acknowledgment exponentially increased with the number of nodes with the conventional scheme. With the proposed scheme, when there were 24 nodes and a span failure occurred, the failure recovery time was about 60% lower compared with the conventional scheme.

5.3.2 Hub-and-spoke Path Configuration

Next, we confirm a hub-and-spoke path configuration in which the number of impaired primary paths varies depending on the location of the failed link. Figure 5.10 shows hub-and-spoke primary paths used in the clockwise direction. There are eight nodes, and the hub is Node A. Corresponding backup paths were set up in the direction opposite the primary paths. All low-priority traffic was one-hop traffic, the link distance was 20 km,

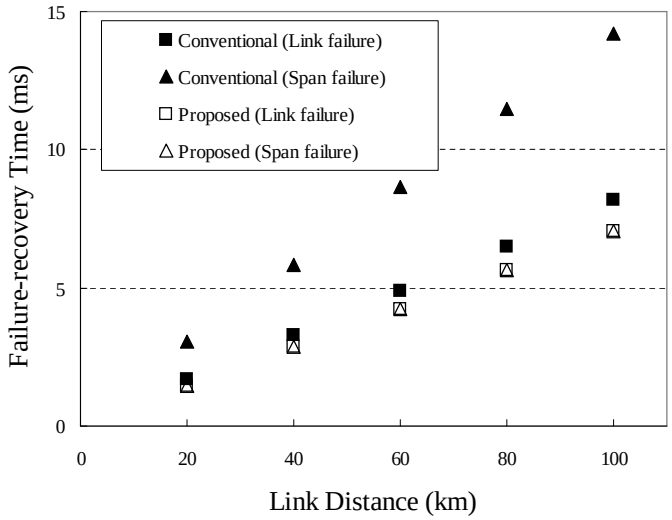

Figure 5.8. Failure-recovery time versus link distance.

and the proportion of low-priority traffic was 100%.

Figure 5.11 shows the failure-recovery time versus the location of the failed link(s). It shows that the failure-recovery time depended on the location of the failed link. This is because the number of impaired primary paths depended on the failure location. For example, the primary paths with one to four hops were impaired when the link between Nodes A and B failed, and only a primary path with four hops was impaired when the link between Nodes D and E failed. In addition, we found that the failure-recovery time was the same regardless of the location of the failed link in the conventional scheme (link failure). The total transmission delay for the control signal and data signal was equivalent to the time it took a signal to make two circuits of the ring, regardless of the location of the failed link, after a destination node had detected a failure. As a result, the delay was always constant.

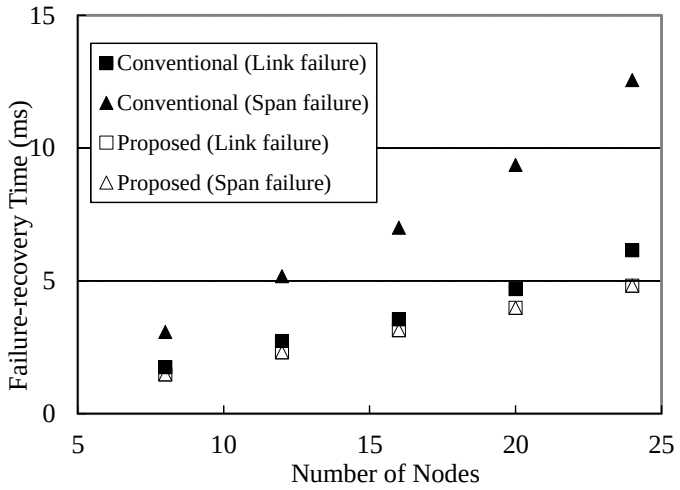

Figure 5.9. Failure-recovery time versus proportion of low-priority traffic in 24 nodes.

5.3.3 Evaluation of Total Recovery Time

This subsection compares total recovery time in intra data center network. The total recovery time is defined as failure-recovery time of all middle class path applied 1:1 protection. Figure 5.12 shows total recovery time versus proportion of low-priority traffic in 24 nodes. From Fig. 5.12, the proposed scheme can realize constant recovery time regardless of low-priority condition. This is because the scheme realizes rapid suspension of low-priority service by sending failure notification of both directions and suspending the service according to the notification.

5.4 Reliability on Energy Efficient Intra Data Center Network

Chapter 4 proposed the energy efficient intra data center network [5-11, 12]. We confirm reliability by applying the technology proposed in Chapter 4. Figure 5.13 shows path configuration when a part of HOPR buffers is turned off. HOPR can forward pack-

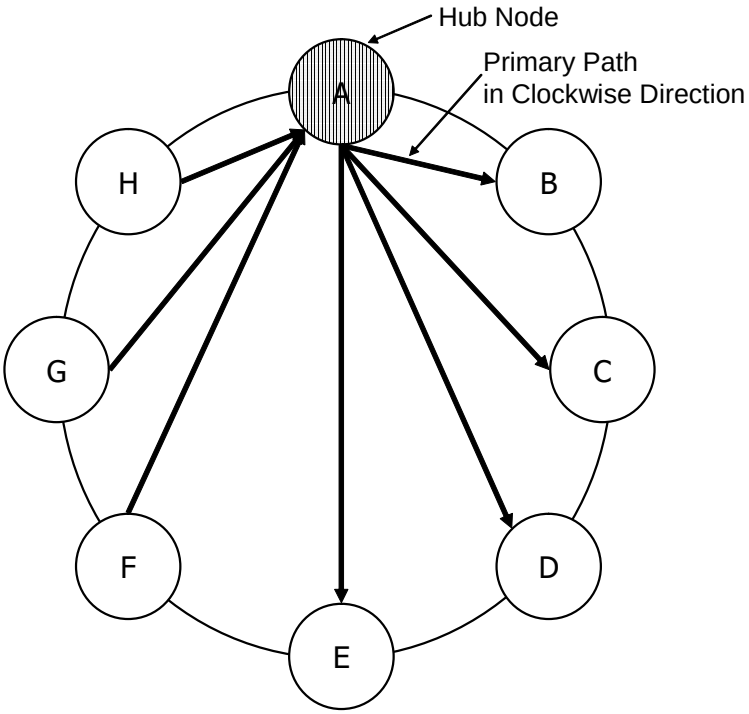

Figure 5.10. Hub-and-spoke-primary-paths used in clockwise direction.

ets even when the buffer is turned off. However, HOPR cannot transfer the packets from the transponders. Thus, the range of low priority traffic is limited. The action of failure recovery in the proposed scheme does not change regardless of the range.

Figure 5.14 shows rate of transponder sending data versus rate of buffers turned off. Figure 5.15 shows failure-recovery time versus of buffers turned off. In this case, the number of nodes is 16. When the rate of buffers turned off increases, the rate of transponder sending data reduces. This is because HOPR cannot transfer the packet from the transponder without buffer. When the rate of buffers increases, the failure-recovery time does not change. This is because failure-detecting HOPR sends a failure notification to source HOPR of the primary paths through intermediate HOPRs in both directions, and the operation time for this is the same regardless of the range.

Therefore, the proposed scheme can realize high reliability with high energy efficiency.

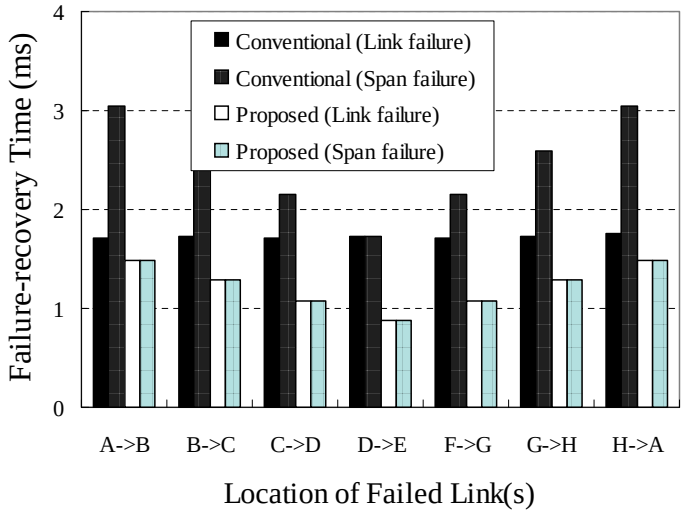

Figure 5.11. Failure-recovery time versus location of failed link.

5.5 Conclusion

Chapter 5 proposed multiple service protection in the optical base intra data center network. The proposal for high reliability based on the optical based intra data center suspends low-priority service rapidly on a failure notification of bidirectional sides from a failure detecting node. The source node can switch to backup path when the node receives the failure notifications from the bidirectional sides because all the node suspends low-priority services. The proposed scheme suppressed total recovery time of middle-priority service to constant level regardless of low-priority traffic condition.

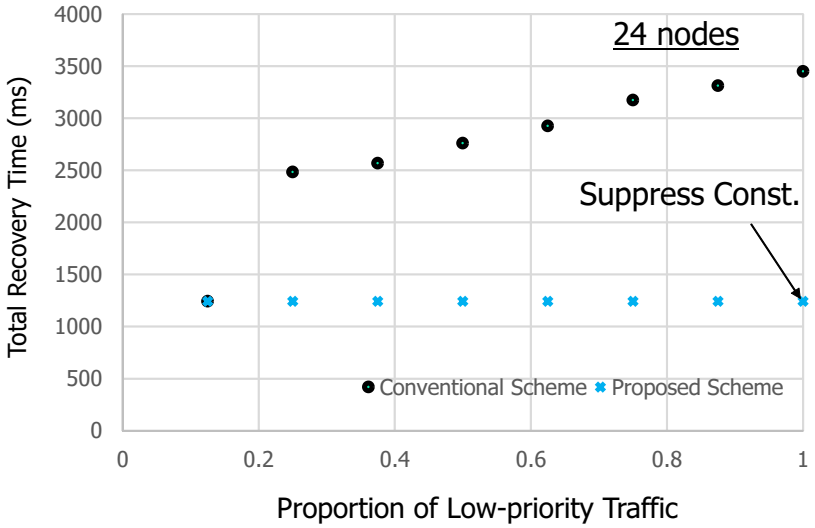

Figure 5.12. Total recovery time versus proportion of low-priority traffic in 24 nodes.

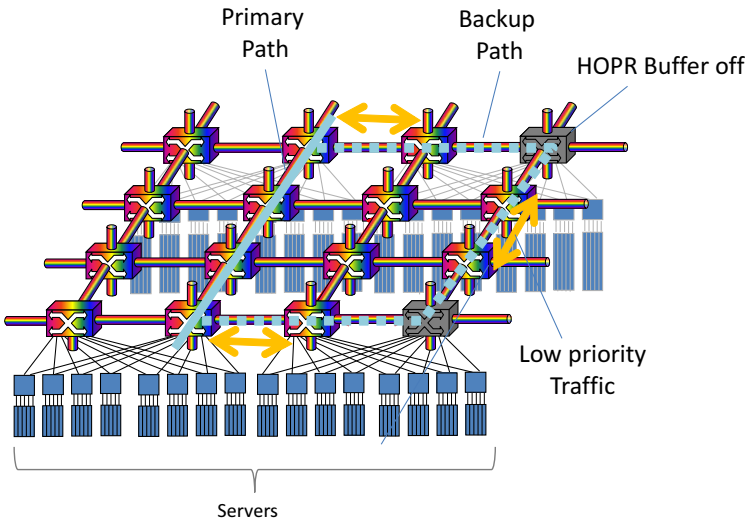

Figure 5.13. Path configuration when a part of HOPR buffer is turned off.

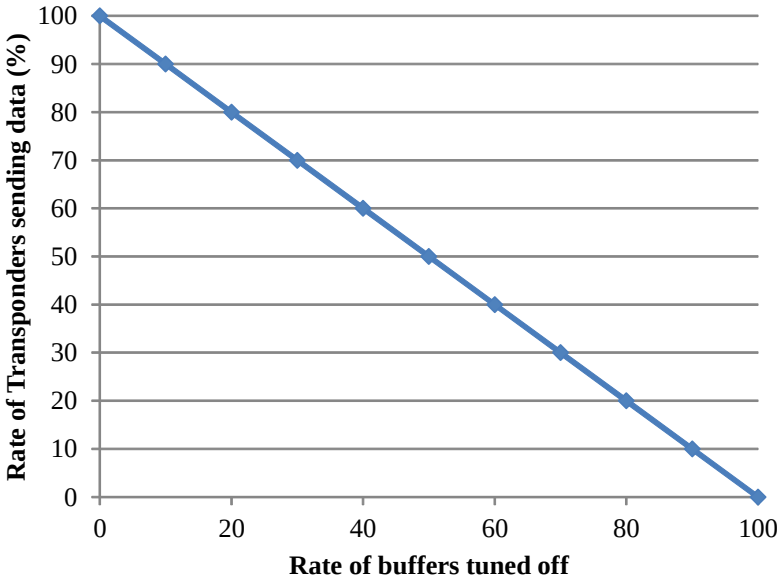

Figure 5.14. Rate of transponder sending data versus rate of buffers turned off.

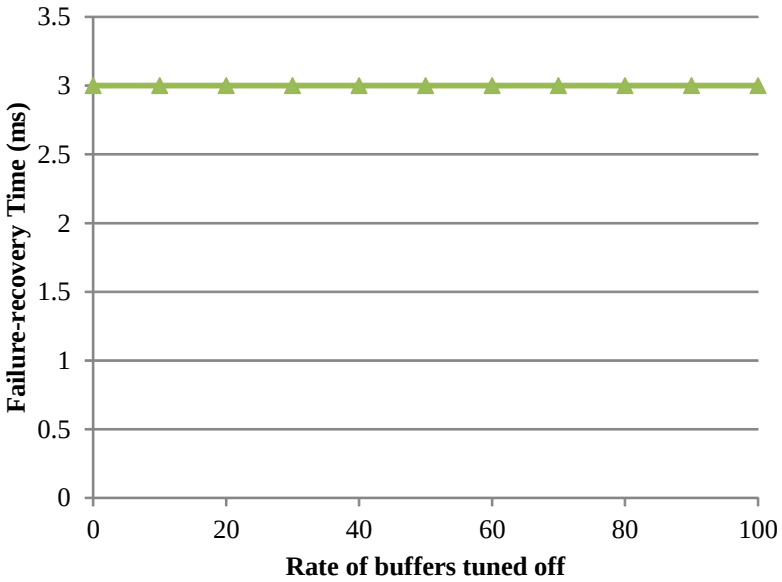

Figure 5.15. Failure-recovery time versus rate of buffers turned off.

References

[5-1] M. Hayashitani, M. Sakauchi, and K. Fukuchi, “A high-speed protection scheme for multiple-priority-class traffic in WDM ring networks,” in Proc. APCC (Asia-Pacific Conference on Communications), pp.1-5, Tokyo, Japan, Oct. 2008.

[5-2] M. Hayashitani, M. Sakauchi, and K. Fukuchi, “A High-speed Protection Scheme for Multiple-Priority-Class Traffic in WDM Ring Networks,” IEICE Trans. Commun., vol.E93-B, no.5, pp.1172-1179, May 2010.

[5-3] T. Shiragaki, et al., “Network Resource Advantages of Bidirectional Wavelength-path Switched Ring,” IEEE Photon. Technol. Lett., vol.11, no.10, pp.1325-1327, Oct. 1999.

[5-4] D. Forbes, et al., “Optical Shared Protection Ring Performance,” in Proc. ECOC, vol.2, pp.52-53, Nice, France, Sep. 1999.

[5-5] D. S. Levy, et al., “Optical Layer Shared Protection Using an IP-based Optical Control Network,” in Proc. OFC, no.TuO8, Anaheim, CA, Mar. 2001.

[5-6] M. J. Li, et al., “Transparent Optical Protection Ring Architecture and Applications,” IEEE/OSA J. Lightwave Technol., vol.23, no.10, Oct. 2005.

[5-7] S. Kim, et al., “Rapid and Efficient Protection for All-optical WDM Mesh Networks,” IEEE J. Sel. Areas Commun., vol.25, no.9, Dec. 2007.

[5-8] M. J. Li, et al., “Design and Experiment of Transparent Four-fiber Optical Channel Shared Protection Ring,” in Proc. NFOEC, pp.2018-2025, Dallas, TX, 2002.

[5-9] S. Seno, et al, “Optical Path Protection with Fast Extra Path Preemption,” IEICE Trans. Commun., vol.E89-B, no.11, pp.3032-3039, Nov. 2006.

[5-10] S. Seno, T. Fujii, M. Tanabe, E. Horiuchi, Y. Baba, and T. Ideguchi, “A Proposal and Evaluation of Multi-class Optical Path Protection Scheme for Reliable Computing,” High Performance Computing and Communications, vol.3726, pp.723-732, Sept. 2005.

[5-11] M. Hayashitani, K. Suzuki, and Y. Mizukoshi, “A Study on Relationship between Data Center Performance and Network Power Consumption with Buffer Management in HOPR-based Data Center,” in Proc. iPOP, no.T4-1, Okinawa, Japan, Apr. 2015.

[5-12] M. Hayashitani, K. Suzuki, and N. Yamanaka, “Control Scheme of HOPR-Based Data Center Network by Considering VM Situation,” IEICE Trans. Commun. (Japanese Edition), vol.J99-B, no.4, pp.334-344, Apr. 2016.

Chapter 6

Overall Conclusion

This dissertation proposed optical access/intra data center networks that offer high energy efficiency and reliability.

Traffic to the data center and that in the data center have been increasing rapidly because of cloud service development. In the near future, traffic within the data center will dominate, so raising the efficiency of the data center network is essential. It logical assumption is that the power consumption of data center networks will increase and impact failure rates as the traffic increases.

The architectures of access and intra data center networks must be designed for power saving and to achieve the survivability demanded by mission critical services. Most existing access data center networks adopt the PON architecture. However, PON has problems in terms of poor scalability and resource utilization because splitters, which simply broadcast packets, are used between OLT and ONU. Therefore, network architectures of the aggregation type are needed when traffic to the data center and the number of users increase rapidly. This dissertation focuses on realizing access data center networks that have higher energy efficiency and far greater scalability.

Present intra data center networks employ electrical switches and routers, so their power consumption increases rapidly when the traffic within the data center network increases. Therefore, optical technologies which can realize dramatic power savings and larger network scales for the data center must be adopted in order to support the increasing traffic loads of the intra data center network. This dissertation focuses on intra data center net-

works with excellent power efficiency, high data center performance, and high reliability.

To solve the above problems, this dissertation proposed advanced optical access/intra data center networks. In the optical access data center network, the proposed scheme realizes high energy efficiency with large scalability by minimizing the number of working OLT to suit user traffic and maximizing unnecessary switching times and ONU sleep by accelerated slot assignment. In the optical intra data center network, the proposed scheme realizes power savings by minimizing working buffers and servers according to VM assignment based on VM group thresholding with high data center performance. Furthermore, the proposed scheme enhances the reliability of middle-priority service by suspending low-priority service rapidly after bidirectional failure notification from the failure detecting node.

Chapter 1 described the background of the dissertation and clarified the challenges in data center networks and the position of the dissertation. Chapter 2 introduced advanced technologies for access/intra data center networks and schemes for enhancing power savings and survivability.

Chapter 3 tackled the energy efficiency of the optical access data center network. The proposal minimizes the number of active OLT to suit user traffic and maximized unnecessary switching times and ONU sleep by accelerated slot assignment. Accelerated slot assignment offers disruption-free slot assignment to the user of the most requested slot to reduce switching time; the assignment creates idle ONUs that can put to sleep. The proposed scheme reduces the power consumption by 47% compared to PON.

Chapter 4 addressed the energy efficiency of the optical intra data center network. The proposal minimizes the number of working buffers and servers according to VM assignment based on VM group thresholding with high data center performance. VM assignment targets VM aggregation for power saving and load distribution for high data center performance. The proposed scheme reduces the network power consumption by 40%

and server power consumption by 59% compared to the conventional HOPR-based data center.

Chapter 5 showed the feasibility of higher reliability in the optical intra data center network. The proposal suspends low-priority service rapidly upon bidirectional failure notification from the failure detecting node. The source node can switch to backup path when the node receives a bidirectional failure notification because all associated nodes suspend low-priority services. The proposed scheme holds the total recovery time of middle-priority service constant regardless of the low-priority traffic condition.

Future work will consider inter data center networks. These networks must support data center migration for greater energy and resource efficiency. As network scale will continue to increase, scalability and energy-efficiency must be considered together with reliability.

The overall conclusion is that this dissertation has contributed to the realization of high energy-efficient and reliable optical access/intra data center networks.

List of the Related Papers

Journal papers

Papers Related to this Ph.D. Dissertation

(1) **Masahiro Hayashitani**, Kazuya Suzuki, and Naoaki Yamanaka, “Control Scheme of HOPR-Based Data Center Network by Considering VM Situation,” IEICE Transactions on Communications, vol.J99-B, no.4, pp.334-344, Apr. 2016 (in Japanese).

(2) **Masahiro Hayashitani**, Masahiro Sakauchi, and Kiyoshi Fukuchi, “A High-speed Protection Scheme for Multiple-Priority-Class Traffic in WDM Ring Networks,” IEICE Transactions on Communications, vol.E93-B, no.5, pp.1172-1179, May 2010.

(3) **Masahiro Hayashitani**, Teruo Kasahara, Daisuke Ishii, Yutaka Arakawa, Satoru Okamoto, Naoaki Yamanaka, Naganori Takezawa, and Keiichi Nashimoto, “GMPLS-based optical slot switching access-distribution network with a 10 ns high-speed PLZT optical switch,” Journal of Optical Networking, vol.7, no.8. pp.744-758, Aug. 2008.

(4) **Masahiro Hayashitani**, Teruo Kasahara, Daisuke Ishii, Yutaka Arakawa, Satoru Okamoto, Naoaki Yamanaka, Naganori Takezawa, and Keiichi Nashimoto, “10 ns High-speed PLZT optical content distribution system having slot-switch and GMPLS controller,” IEICE Electronics Express, vol.5, no.6, pp.181-186, Mar. 2008.

Other Papers

(1) Ken-ichi Kitayama, Yue-cai Huang, Yuki Yoshida, Ryo Takahashi, Toru Segawa, Salah Ibrahim, Tatsushi Nakahara, Yasumasa Suzaki, **Masahiro Hayashitani**, Yohei Hasegawa, Yasuhiro Mizukoshi, and Atsushi Hiramatsu, "Torus-Topology Data Center Network Based on Optical Packet/Agile Circuit Switching with Intelligent Flow Management," IEEE/OSA Journal of Lightwave Technology, vol.33, no.5, pp.1063-1071, Mar. 2015.

(2) Hiroyuki Miyagi, **Masahiro Hayashitani**, Daisuke Ishii, Yutaka Arakawa, and Naoaki Yamanaka, "Advanced Wavelength Reservation Method Based on Deadline-Aware Scheduling for Lambda Grid Networks," IEEE/OSA Journal of Lightwave Technology, vol.25, no.10, pp.2904-2910, Oct. 2007.

International conference papers

(1) **Masahiro Hayashitani**, Satoru Okamoto, Eiji Oki, and Naoaki Yamanaka, "Source-based Wavelength-path Protection Scheme with Tree-shaped Backup-path Configuration in WDM Networks," in Proc. IEEE HPSR (High Performance Switching and Routing), pp.55-60, Yokohama, Japan, June 2016. (Presented by Masahiro Hayashitani)

(2) **Masahiro Hayashitani**, Kazuya Suzuki, and Yasuhiro Mizukoshi, "A study on Relationship between Data Center Performance and Network Power Consumption with Buffer Management in HOPR-based Data Center," in Proc. iPOP (IP + Optical Network), no.T4-1, Okinawa, Japan, Apr. 2015. (Presented by Masahiro Hayashitani)

(3) **Masahiro Hayashitani**, Yohei Hasegawa, Kazuya Suzuki, and Yasuhiro Mizukoshi,

"Flexible and automated operational control in SDN transport-base virtual router," in Proc. OFC (Optical Fiber Conference), no.W1E.1, San Francisco, CA, Mar. 2014. (Presented by Masahiro Hayashitani)

(4) Ken-ichi Kitayama, Yue-Cai Huang, Yuki Yoshida, Ryo Takahashi, and **Masahiro Hayashitani**, "Optical packet and path switching intra center network: Enabling technologies and network performance with intelligent flow control," in Proc. ECOC (European Conference on Optical Communication), pp.1-3, Cannes, France, Sept. 2014. (Presented by Ken-ichi Kitayama).

(5) Ken-ichi Kitayama, Soumitra Debnath, Yuki Yoshida, Ryo Takahashi, Atsushi Hiramatsu, Yuichi Oshita, Masayuki Murata, and **Masahiro Hayashitani**, "Green, smart optical packet switching network with flow control for data centers," in Proc. IEEE Photonics Society Summer Topical Meeting Series, pp.252-253, Waikoloa, HI, July 2013. (Presented by Ken-ichi Kitayama)

(6) **Masahiro Hayashitani**, Yohei Hasegawa, Yoshihiko Kanaumi, Shuichi Saito, Yasuhiro Mizukoshi, and Shuji Ishii, "Demonstration of packet transport node for software-defined multi-layer networking in wide area network testbed," in Proc. iPOP (IP + Optical Network), no.T1-2, Tokyo, Japan, May 2013. (Presented by Masahiro Hayashitani)

(7) **Masahiro Hayashitani**, Yohei Hasegawa, and Yasuhiro Mizukoshi, "A selection method of failure-recovery scheme in OpenFlow-enabled MPLS-TP networks," in Proc. iPOP (IP + Optical Network), no.P-3, Tokyo, Japan, May 2013. (Presented by Masahiro Hayashitani)

(8) **Masahiro Hayashitani**, Itaru Nishioka, Ippei Akiyoshi, and Yasuhiro Mizukoshi, "Design and Implementation of OpenFlow-enabled MPLS-TP Switch Prototype,"

in Proc. iPOP (IP + Optical Network), no.T3-4, Yokohama, Japan, May 2012. (Presented by Masahiro Hayashitani)

(9) **Masahiro Hayashitani**, and Masahiro Sakauchi, "Demonstration of High-Speed Signaling for Protection under Multiple-Priority-Class Traffic in WDM Ring Networks," in Proc. OFC (Optical Fiber Conference), no.JThA43, San Diego, CA, Mar. 2010. (Presented by Masahiro Hayashitani)

(10) **Masahiro Hayashitani**, Masahiro Sakauchi, and Kiyoshi Fukuchi, "A high-speed protection scheme for multiple-priority-class traffic in WDM ring networks," in Proc. APCC (Asia-Pacific Conference on Communications), pp.1-5, Tokyo, Japan, Oct. 2008. (Presented by Masahiro Hayashitani)

(11) Hiroyuki Miyagi, **Masahiro Hayashitani**, Daisuke Ishii, Yutaka Arakawa, and Naoaki Yamanaka, "A Deadline-Aware Scheduling Scheme for Wavelength Assignment in λ Grid Networks," in Proc. ICC (IEEE International Conference on Communications), pp.2383-2387, Glasgow, UK, June 2007. (Presented by Hiroyuki Miyagi)

(12) Teruo Kasahara, **Masahiro Hayashitani**, Yutaka Arakawa, Satoru Okamoto, and Naoaki Yamanaka, "Design and Implementation of GMPLS-basd Optical Slot Switching Network with PLZT High-speed Optical Switch," in Proc. HPSR (High Performance Switching and Routing), pp.1-6, Brooklyn, NY, May 2007. (Presented by Teruo Kasahara)

(13) **Masahiro Hayashitani**, Teruo Kasahara, Daisuke Ishii, Yutaka Arakawa, Satoru Okamoto, Naoaki Yamanaka, Naganori Takezawa, and Keiichi Nashimoto, "Design and Implementation of GMPLS-Based Optical Slot Switching Access-Distribution Network Using PLZT Ultra-High Speed Optical Switch," in Proc. OFC (Optical Fiber Conference), no.OWC4, Ahaheim, CA, Mar. 2007. (Presented by Masahiro Hayashitani)

(14) **Masahiro Hayashitani**, Teruo Kasahara, Daisuke Ishii, Yutaka Arakawa, Satoru Okamoto, Naoaki Yamanaka, Naganori Takezawa, and Keiichi Nashimoto, "GMPLS-based Optical Slot Switching Access-distribution Network with 10ns High-speed PLZT Optical Switch,", in Proc. CPT (International Symposium on Contemporary Photonics Technology), no.B-4, pp.23-24, Tokyo, Japan, Jan. 2007. (Presented by Masahiro Hayashitani)

(15) **Masahiro Hayashitani**, Kohei Okazaki, and Naoaki Yamanaka, "A new burst assembly technique supporting fair QoS about the number of hops in OCBS multi-hop networks," in Proc. COIN-NGNCON (Joint International Conference on Optical Internet and Next Generation Network), pp.40-42, Jeju, Korea, July 2006.

Technical reports

(1) **Masahiro Hayashitani**, Yohei Hasegawa, Kazuya Suzuki, Yasuhiro Mizukoshi "A study on SDN transport technology for gradual migration from existing network," Technical Report of IEICE, vol.114, no.28, NS2014-36, pp.69-73, May 2014. (in Japanese) (Presented by Kazuya Suzuki)

(2) **Masahiro Hayashitani**, Yohei Hasegawa, Kazuya Suzuki, Yasuhiro Mizukoshi "SDN transport technology for realizing reliable virtual router," Technical Report of IEICE, vol.113, no.244, NS2013-93, pp.19-23, Oct 2013. (in Japanese) (Presented by Masahiro Hayashitani)

(3) **Masahiro Hayashitani**, and Masahiro, Sakauchi "High-speed Optical Path Protection Scheme with Tree-shaped Backup Path Configuration in WDM Mesh Networks," Technical Report of IEICE, vol.108, no.417, PN2008-61 pp.7-12, Jan. 2009. (in Japanese) (Presented by Masahiro Hayashitani)

(4) **Masahiro Hayashitani**, Masahiro, Sakauchi, and Kiyoshi Fukuchi, “Protection Scheme Suspending Low Priority Traffic Rapidly in WDM Ring Networks,” Technical Report of IEICE, vol.107, no.544, PN2007-94, pp.117-122, Mar. 2008. (in Japanese) (Presented by Masahiro Hayashitani)

(5) **Masahiro Hayashitani**, Teruo Kasahara, Daisuke Ishii, Yutaka Arakawa, Satoru Okamoto, and Naoaki Yamanaka, “ Design and Implement of GMPLS-based Optical Slot Switching Network Using PLZT Ultra-high Speed Optical Switch,” Technical Report of IEICE, vol.106, no.208, PN2006-15, pp.29-34, Aug. 2006. (in Japanese) (Presented by Masahiro Hayashitani)

(6) Teruo Kasahara, Jun-ichiro Homma, **Masahiro Hayashitani**, Daisuke Ishii, Yutaka Arakawa, and Naoaki Yamanaka, “Optical Slot Switch Architecture based on dynamic path setup using ultra-high speed PLZT optical switch,” Technical Report of IEICE, vol.105, no.667, PN2005-103, pp.27-32, Mar. 2006. (in Japanese) (Presented by Teruo Kasahara)

(7) Hiroyuki Miyagi, **Masahiro Hayashitani**, Daisuke Ishii, Yutaka Arakawa, and Naoaki Yamanaka, “ A Deadline-Scheduling Scheme for Wavelength Assignment in Lambda Grid Network,” Technical Report of IEICE, vol.105, no.667, PN2005-109, pp.57-62, Mar. 2006. (in Japanese) (Presented by Hiroyuki Miyagi)

(8) **Masahiro Hayashitani**, Daisuke Ishii, Kohei Okazaki, and Naoaki Yamanaka, “A New Burst Assembly Technique for QoS Support Regardless of the Number of Hops in OCBS Multi-hop Networks,” Technical Report of IEICE, vol.104, no.600, PN2004-83, pp.39-44, Jan. 2005. (Presented by Masahiro Hayashitani).

Acknowledgments

This dissertation has been written under the direction and guidance of Prof. Naoaki Yamanaka in Department of Information and Computer Science, Keio University, Japan.

My sincere gratitude and deepest appreciation should be first given to my supervisor Prof. Naoaki Yamanaka for his valuable suggestions, guidance and continuous encouragement throughout my research work. With the guidance of Prof. Yamanaka, I was able to conclude my studies and gain splendid experiences in the Ph.D. course.

I owe a great deal of thanks to the members of dissertation committee, Prof. Iwao Sasase in Department of Information and Computer Science, Keio University, Japan, Prof. Hideharu Amano in Department of Information and Computer Science, Keio University, Japan, Prof. Hiroyuki Tsuda in Department of Electronics and Electrical Engineering, Keio University, Japan, and Prof. Bijan Jabbari in Electrical and Computer Engineering Department, George Mason University, USA, for their comments, suggestions, and careful and critical reading of this dissertation.

I am deeply grateful to Prof. Satoru Okamoto of Yamanaka Lab. in Department of Information and Computer Science, Keio University, Japan. He gave insightful comments and suggestions, especially about access networks.

I am very grateful to Prof. Eiji Oki in Graduate School of Informatics, Kyoto University, Japan, for his valuable support and comments on my studies and dissertation.

I would like to thank to Dr. Kazuya Suzuki, NEC Corporation, who provided a lot of technical advice about network systems, especially about data center networks.

I wish to express my sincere gratitude to Dr. Keiichi Nashimoto who is the founder, president and CEO of EpiPhotonics Corp. High speed optical switches such as the PLZT optical switch developed by EpiPhotonics Corp. are essential for access networks.

I am deeply grateful to Dr. Kohei Okazaki, NEC Corporation for providing a lot of technical advice about optical networks and positive comments when he was a Ph.D. candidate. We had good discussions not only on my research topics but also on wide range topics.

I would like to thank to Dr. Yutaka Arakawa, Associate Prof. in Graduate School of Information Science, Nara Institute of Science and Technology, Japan, for his constructive comments and warm encouragement when he was an assistant in Yamanaka Lab.

I am very grateful to Dr. Daisuke Ishii, Hitachi Ltd. Central Research Laboratory, for providing a lot of technical advice on optical networks when he was a Ph.D. candidate.

I am really thankful to Dr. Takehiro Sato, Assistant Professor of Yamakaka Lab. in Department of Information and Computer Science, Keio University, Japan. He gave me a lot of support, comments, and advice about the dissertation.

I must thank other current and prior members of Sasase Lab. and Yamanaka Lab., Dr. Hidetoshi Takeshita, Dr. Sho Shimizu, Dr. Kunitaka Ashizawa, Dr. Kazumasa Tokuhashi, Dr. Ko Kikuta, Dr. Shan Gao, Mr. Junichiro Honma, Mr. Kohki Ohba, Mr. Takamasa Isohara, Mr. Motoki Shirasu, and Mr. Kazuhiko Hasegawa, Mr. Hiroyuki Ishikawa, Mr. Teruo Kasahara, Mr. Tomohiro Tsuji, Mr. Hiroyuki Miyagi, Mr. Masahiro Tateno, Mr. Ryota Usui, Ms. Fumiko Uehara, Mr. Mikio Kanako, Mr. Taku Kihara, Mr. Shimpei Koda, Mr. Masahiro Nishida, Mr. Kazuki Irie, Ms. Motomi Akagi, Mr. Yusuke Okazaki, Ms. Midori Terasawa, Mr. Hirofumi Yamashita, Mr. Jun Matsumoto, Ms. Haruka Yonezu, Mr. Soushi Yamamoto, Mr. Shanming Zhang, Mr. Takuro Okano, Mr. Harunaga Onda, Mr. Yuki Higuchi, and Ms. Akiko Hirao. All have inspired me in my research life.

I appreciate the efforts of the secretaries of Yamanaka Lab., Ms. Ayumi Sato, Ms. Kaori Kozakai, Ms. Atsuko Morimoto, and Ms. Yasuyo Komai.

Finally, I would like to express my deepest gratitude to my wife, daughter, and son for their moral support and warm encouragement. Finally, I must thank to my father, mother, and sisters for warm support. This dissertation could not be completed without their support and encouragement.

School of Science for Open and Environmental Systems

Graduate School of Science and Technology

Keio University

Masahiro Hayashitani

July 31st, 2017