Data Center Network Topologies using Optical Packet Switches

Yuichi Ohsita†, Masayuki Murata‡

† Graduate School of Economics, Osaka University
‡ Graduate School of Information Science and Technology, Osaka University

***Abstract*— The large data center network constructed of only the electronic packet switches consumes large power to provide enough bandwidth for all server pairs. One approach to construct the data center network that provides enough bandwidth with small energy consumption is to use the optical packet switches. In the data center network using the optical packet switches, however, the failures of the optical packet switches may have large impacts on the communication between servers. In this paper, we propose the data center network topology using the optical packet switches that can provide enough bandwidth even when some optical packet switches fail. We evaluate our topology and clarify that our topology can provide enough bandwidth even when some optical packet switches fail.**

***Index Terms*— Data Center Network, Topology, Optical Packet Switch**

I. Introduction

In recent years, large data centers with tens of thousands of servers have been built to handle very large amounts of data generated by various online applications. In a data center, servers handle such very large amounts of data by communicating with each other via the network within the data center. Thus, the network within the data center has large impacts on the performance of the data center.

However, the traditional data center network, which is constructed as the tree topology, is not suitable to large data centers. The traditional data center network cannot provide enough bandwidth between all servers in a large data center, because the links connected to the root switch may become the bottlenecks. The traditional data center network consumes large energy because a large number of switches or switches with a large number of ports, whose energy consumption is large, are required to connect a large number of servers.

There are many researches to construct networks for large data centers [1–9]. To provide sufficient bandwidth between servers, many methods [1–3] use the topology called FatTree. The FatTree is the tree topology with multiple root switches. In this topology, each switch uses a half of its ports to connect it to the switches of the upper layer, and the other half of its ports to connect it to the switches of the lower layer. This topology, however, requires many switches to provide enough bandwidths between a large number of servers, which leads to large energy consumption.

Methods to connect a large number of servers with small energy consumption have also been proposed [5–9]. In these methods, the network is constructed with a small number of switches and servers having multiple ports by directly connecting server ports. The network constructed by these methods, however, may not provide enough bandwidths between all servers, because servers cannot use all the bandwidths of their ports since the servers also relay the traffic between other server pairs.

One approach to provide enough bandwidth with small energy consumption is to use optical packet switches [10–12]. Optical packet switches can provide large bandwidth between their ports with small energy consumption. In recent years, the optical packet switch architecture with a large number of ports for data centers, which is constructed by using multiple optical switches, has been introduced [11], [12]. However, the network using the optical packet switches with a large number of ports is vulnerable to the failure of the optical packet switch, because most of the traffic between servers traverses the optical packet switch.

In this paper, we discuss the network topology using the optical packet switches with a small number of ports that can provide enough bandwidths between all server pairs even when failures occur. In our topology, we use the optical packet switches to construct the core network of the data center. Similar to the traditional data center, we deploy the electronic packet switch called the top-of-rack (ToR) switch in each server rack. All servers in a server rack are connected to the ToR switch in the same server rack. The ToR switches are connected to the core network by connecting them to optical packet switches. By connecting each optical packet switch to multiple ToR switches and aggregating traffic from them, we use the large bandwidth between optical packet switches efficiently. Moreover, by connecting each ToR switch to multiple optical packet switches, we keep the connectivity between all servers even when optical packet switches fail.

The rest of this paper is organized as follows. In Section II, we explain the overview of the data center network using the optical packet switches. In Section III, we propose a topology of the core network constructed of the optical packet switches, and the method to connect the ToR switches to the optical packet switches. Then, we evaluate our topology and clarify that our topology can provide sufficient bandwidth even when optical packet switches fail in Section IV. Finally, Section V provides a conclusion.

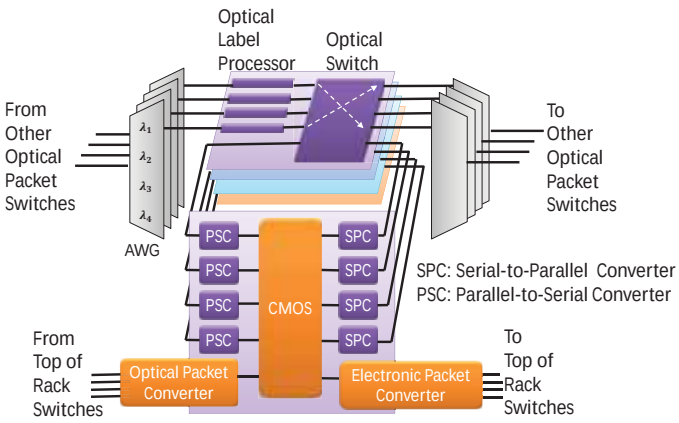


Fig. 1. Optical Packet Switch Achitecture

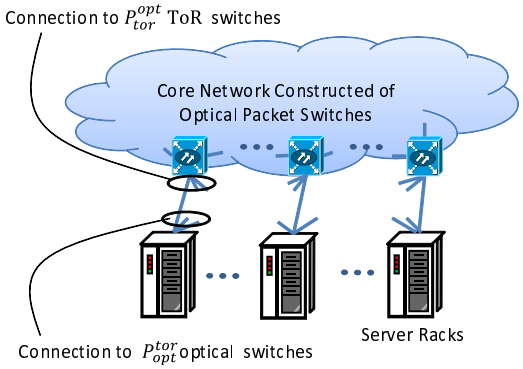


Fig. 2. Data Center Network Using Optical Packet Swtiches

II. Data Center Networks Using Optical Packet Switches

A. Opticak packet switches

In this paper, we modify the optical packet switch proposed by Urata et al. [10] to suit data center networks. Figure 1 shows the optical packet switch architecture used in this paper. In this architecture, optical packets, constructed of multiple wavelengths, are relayed between optical packet switches. The optical packets from other optical packet switches are demultiplexed into optical signals of each wavelength. Then, after label processing, the optical signals are relayed to the destination port and multiplexed into optical packets. In case of collision, the optical packets are stored in the shared buffer constructed with CMOS after serial-to-parallel conversion. Then, we try to relay the packets again after parallel-to-serial conversion.

Packets from ToR switches are aggregated to the optical packets and stored in the shared buffer. Then, the packets are relayed after parallel-to-serial conversion. Optical packets whose destination is the ToR switches connected to the optical packet switch are also stored in the shared buffer. Then the packets are sent to the destination ToR switches after demultiplexing the optical packet into the packets to each ToR switch.

Optical packet switch can provide large bandwidth with small energy consumption. Thus, we can construct the network that can provide enough bandwidth with small energy consumption by using the optical packet switches.

In this paper, each optical packet switch has $P_{\mathrm{opt}}^{\mathrm{opt}}$ ports to the other optical packet switches and $P_{\mathrm{tor}}^{\mathrm{opt}}$ ports to the ToR switches. The bandwidths of the links between optical packet switches are B^{opt} Gbps and the bandwidths of the links between optical packet switches and ToR switches are 10 Gbps.

B. Data Center Networks using Optical Packet Switches

Figure 2 shows the data center network using optical packet switches. In this network, the optical packet switches are used to construct the core network of the data center.

Similar to the traditional data center, we deploy the ToR switch in each server rack. $P_{\mathrm{tor}}^{\mathrm{svr}}$ servers in a server rack are connected to one ToR switch with a 1 Gbps links. The ToR switches are connected to the core network by connecting them to optical packet switches. As mentioned in Section II-A, each optical packet switch is connected to $P_{\mathrm{tor}}^{\mathrm{opt}}$ ToR switches, and aggregates traffic from them to efficiently use the large bandwidth between optical packet switches. Each ToR switch is also connected to $P_{\mathrm{opt}}^{\mathrm{tor}}$ optical packet switches to keep the connectivity even when optical packet switches fail.

III. Topologies suitable to data center networks using optical packet switches

In this section, we propose a topology satisfying the following points; (1) we efficiently use the links between optical packet switches, whose bandwidths are much larger than those of ports of ToR switches, by aggregating traffic from multiple ToR switches, and (2) we keep the connectivity between all servers even when optical packet switches fail by connecting each ToR switch to multiple optical packet switches.

In Section III-A, we propose a topology for data center network using the optical packet switches. In Section III-B, we explain a method to set parameters of our topology so as to provide enough bandwidth between all servers.

A. Topology of Data Center Network Using Optical Packet Switches

In our topology, we divide the data center network into multiple groups. By connecting each ToR switch to optical packet switches belonging to the same group, we avoid long links between optical packet switches and ToR switches. We denote the number of ToR switches in each group, the number of optical packet switches in each group, and number of groups as $N_{\mathrm{in}}^{\mathrm{tor}}$, $N_{\mathrm{in}}^{\mathrm{opt}}$, and G respectively. Each optical packet uses P_{in} ports to connect optical packet switches belonging to the same group, and P_{gr} ports to connect optical packet switches belonging to other groups.

We also divide each group into $P_{\mathrm{opt}}^{\mathrm{tor}}$ subgroups. Each ToR switch is connected to optical packet switches belonging to different subgroups. All of P_{in} ports of each optical packet switch are used to connect optical packet switches belonging to the same subgroup. No links are constructed between optical

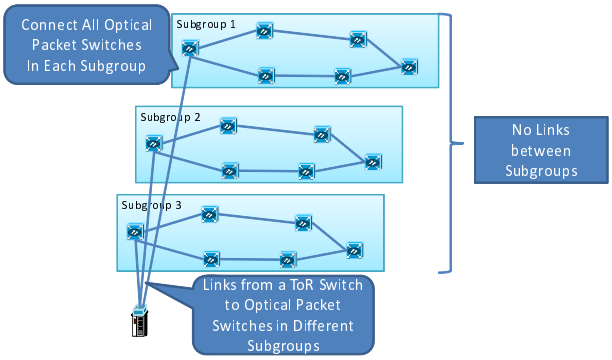


Fig. 3. Connection withing a Group

packet switches belonging to different subgroups as shown in Figure 3.

In this topology, we have $P_{\rm opt}^{\rm tor}$ distinct paths between all ToR switch pairs. Thus, we can keep the connectivity between all ToR switch pairs even when optical packet switches fail.

In addition, this topology effectively uses the ports of optical packet switches. The set of ToR switches connected to each subgroup is the same. Thus, the links between optical packet switches belonging to different subgroups are not required. By using all of $P_{\rm in}$ ports of each switch to connect optical packet switches of the same subgroup, we make the number of hops between ToR switches and optical packet switches small.

We assign the unique ID to the groups, the subgroups in each group and the optical packet switches in each subgroup. We denote the group ID, the subgroup ID and the optical packet switch ID of optical packet switch s as $D^{\rm gr}(s)$, $D^{\rm sub}(s)$ and $D^{\rm opt}(s)$ respectively.

1) Connection within a Group: We first connect the optical packet switches belonging to the same subgroups. Then, we connect each ToR switch to $P_{\rm opt}^{\rm tor}$ optical packet switches belonging to different subgroups.

The optical packet switches belonging to the same subgroup are connected by the following steps. First, we construct a ring topology by connecting the optical packet switches of the nearest optical packet switch IDs. Then, we add links between optical packet switches S_1 and S_2 if the following constraint is satisfied;

$$D^{\rm opt}(S_2) = \lfloor D^{\rm opt}(S_1) + iN_{\rm sub}/(P_{\rm in}-1)\rfloor \bmod N_{\rm sub}, \quad (1)$$

where $N_{\rm sub}$ is the number of optical packet switches belonging to each subgroup and i is a positive integer. If the optical packet switch S_2 satisfying Eq. (1) does not have enough ports used for the connection within a group, we connect S_1 to the optical packet switch that has enough ports and has the optical packet switch ID close to S_2.

2) Connection between Groups: We connect groups by adding links between optical packet switches belonging to different groups. The number of links used to connect a group to other groups is $N_{\rm in}^{\rm opt}P_{\rm gr}$. If $N_{\rm in}^{\rm opt}P_{\rm gr} \geq G-1$, we can add links between all group pairs. In this paper, we assume that we can add links between all group pairs.

To connect groups, we select the optical packet switches on the both ends of links between the groups. We select the optical packet switch S_1 as the optical packet switch to be connected to the Kth link between groups $D^{\rm gr}(S_1)$ and $D^{\rm gr}(S_2)$ if the following constraint is satisfied;

$$D^{\rm in}(S_1) = \begin{cases} \lfloor \frac{D^{\rm gr}(S_2)+K(G-1)}{P_{\rm gr}} \rfloor & (D^{\rm gr}(S_1) \geq D^{\rm gr}(S_2)) \\ \lfloor \frac{D^{\rm gr}(S_2)+K(G-1)-1}{P_{\rm gr}} \rfloor & (\text{Otherwise}) \end{cases}, \quad (2)$$

where $D^{\rm in}$ is the number defined by

$$D^{\rm in}(S_1) = D^{\rm sub}(S_1)\frac{N_{\rm in}^{\rm opt}}{P_{\rm opt}^{\rm tor}} + D^{\rm opt}(S_1).$$

3) Routing for Our Topology: In our topology, we can calculate routes from ToR switches to optical packet switches and from optical packet switches to ToR switches by using the ID assigned to optical packet switches, $[D^{\rm gr}(s), D^{\rm sub}(s), D^{\rm opt}(s)]$, without exchanging any route information.

a) Routes from ToR Switches to Optical Packet Switches: The routes from ToR switch in the group $D^{\rm gr}(s)$ to the optical packet switch d are calculated by the following steps.

If the destination optical packet switch d belongs to $D^{\rm gr}(s)$, the source ToR switch first sends the packet to the optical packet switch that is directly connected to the source ToR switch and belongs to the same subgroup as the destination optical packet switch d, (i.e., the subgroup $D^{\rm sub}(d)$). The intermediate optical packet switch selects the next hop by calculating $H(d,a)$, defined by Eq. (3), for all neighbor optical packet switches a.

$$H(d,a) = |D^{\rm opt}(d) - D^{\rm opt}(a)| \quad (3)$$

The optical packet switch a having the smallest $H(d,a)$ is close to the destination optical packet switch d. Thus, we select the optical packet switch having the smallest $H(d,a)$ as the next-hop optical packet switch. If there are multiple optical packet switches having the smallest $H(d,a)$, we regard all optical packet switches having the smallest $H(d,a)$ as the candidates of the next hop, and balance the load by selecting the next-hop optical packet switch randomly from the candidates.

If the destination optical packet switch does not belong to $D^{\rm gr}(s)$, we first select the intermediate optical packet switch in the group $D^{\rm gr}(s)$ having a link to an optical packet switch belonging to the subgroup $D^{\rm sub}(d)$ in the group $D^{\rm gr}(d)$. The intermediate optical packet switch is selected by the following steps. First we calculate the range of $\tilde{k}$ in Eq. (2) where $\tilde{k}$th link between the groups $D^{\rm gr}(s)$ and $D^{\rm gr}(d)$ are connected to optical packet switches belonging to the subgroup $D^{\rm sub}(d)$ in the group $D^{\rm gr}(d)$ by solving the following inequation;

$$D^{\rm sub}(d)\frac{N_{\rm in}^{\rm opt}}{P_{\rm opt}^{\rm tor}} \leq \tilde{D}^{\rm in}(d) < (D^{\rm sub}(d)+1)\frac{N_{\rm in}^{\rm opt}}{P_{\rm opt}^{\rm tor}}, \quad (4)$$

where

$$\tilde{D}^{\rm in}(d) = \begin{cases} \lfloor \frac{D^{\rm gr}(s)+\tilde{k}(G-1)}{P_{\rm gr}} \rfloor & (D^{\rm gr}(d) \geq D^{\rm gr}(s)) \\ \lfloor \frac{D^{\rm gr}(s)+\tilde{k}(G-1)-1}{P_{\rm gr}} \rfloor & (\text{Otherwise}) \end{cases}.$$

Then, we identify the optical packet switch s' connected to an optical packet switch belonging to the subgroup $D^{\rm sub}(d)$ in

the group $D^{\mathrm{gr}}(d)$, by substituting $\tilde{k}$ to K, s' to S_1, and d to S_2 in Eq. (2).

After selecting the intermediate optical packet switch, we calculate the routes from the source ToR switch to the intermediate optical packet switch and from the intermediate optical packet switch to the destination optical packet switch by the same steps as the case that the destination optical packet switch d belongs to $D^{\mathrm{gr}}(s)$.

b) Routes from optical packet switches to ToR switches: If the destination ToR switch belongs to the same group as the source optical packet switch, we first select the intermediate optical packet switch d^{opt} that belongs to the same subgroup as the source optical packet switches and is directly connected to the destination ToR switch. In this paper, we assume that each optical packet switch knows the connections between all optical packet switches and all ToR switches within its group. Thus, each optical packet switch can calculate d^{opt}. Then, we calculate the routes from the source optical packet switch to the intermediate optical packet switch d^{opt} by using $H(d^{\mathrm{opt}}, a)$ in the same manner as the case of routes from the ToR switch to the optical packet switch.

If the destination ToR switch does not belong to the same group as the source optical packet switch, we select the intermediate optical packet switch having a link to the group of the destination ToR switch. The intermediate optical packet switches having a link to the group of the destination ToR switch are obtained by Eq. (2). Then, we calculate the routes from the source optical packet switch to the intermediate optical packet switch, and from the intermediate optical packet switch to the destination ToR switch, by the same steps as the case that the destination ToR switch belongs to the same group as the source optical packet switch.

c) Routes between ToR Switches: We can calculate routes between ToR switches by selecting an intermediate optical packet switch and calculating the routes from the source ToR switch to the intermediate optical packet switch and from the intermediate optical packet switch to the destination ToR switch. By selecting the intermediate optical packet switch at an end of a link between the group of the source ToR switch and the group of the destination ToR switch based on Eq. (2), we can avoid large hop counts between ToR switches.

d) Handling Failures: If the optical packet switch S_1 cannot find no suitable next-hop optical packet switch for the destination d because of failures, it returns the packet to the previous-hop optical packet switch S_2. By receiving the returned packet, S_2 identifies that S_1 has no suitable path to the destination d. Thus, S_2 removes S_1 from the candidates of the next-hop optical packet switches to d, and relays the packet to one of the other candidates. If S_2 cannot also find no suitable next-hop optical packet switch after removing S_1 from the candidates, S_2 also returns the packet to the previous hop of S_2. By continuing the above steps, all optical packet switches can remove the switches having no suitable routes to d from their candidates of the next-hop switch to d.

B. Parameter Settings

Our topology has three kinds of parameters, P_{gr}, P_{in} and connection between the ToR switches and the optical packet switches. In this subsection, we set these parameters so that our topology can accommodate any traffic without limiting the bandwidth between servers.

When setting parameters, we assume that traffic is balanced by Valiant Load Balancing (VLB) [13]. In the VLB, we select the intermediate nodes randomly regardless of the destination to avoid the concentration of traffic on certain links even when traffic volume of certain node pairs is large.

Applying the VLB to our topology, we select an intermediate optical packet switch randomly with the probability of $\frac{1}{N_{\mathrm{in}}^{\mathrm{opt}}G}$. Then, traffic is sent via the selected intermediate optical packet switch. Applying the VLB, the traffic volume from a ToR switch to an optical packet switch, $T_{\mathrm{tor,opt}}$ and the traffic volume from an optical packet switch to a ToR switch, $T_{\mathrm{opt,tor}}$ satisfy the following conditions;

$$T_{\mathrm{tor,opt}} \leq \frac{P_{\mathrm{tor}}^{\mathrm{svr}}}{N_{\mathrm{in}}^{\mathrm{opt}}G}, \tag{5}$$

$$T_{\mathrm{opt,tor}} \leq \frac{P_{\mathrm{tor}}^{\mathrm{svr}}}{N_{\mathrm{in}}^{\mathrm{opt}}G}. \tag{6}$$

Thus, we set the parameters of our topology so as to accommodate traffic of $T_{\mathrm{tor,opt}}^{\max} = T_{\mathrm{opt,tor}}^{\max} = \frac{P_{\mathrm{tor}}^{\mathrm{svr}}}{N_{\mathrm{in}}^{\mathrm{opt}}G}$ between all ToR switch and optical packet switch pairs.

1) Parameter of Connection Between Groups: By applying the VLB, the sum of traffic sent between a certain group pair T^{gr} is constrained by

$$T^{\mathrm{gr}} \leq (T_{\mathrm{tor,opt}}^{\max} + T_{\mathrm{opt,tor}}^{\max}) N_{\mathrm{in}}^{\mathrm{opt}} N_{\mathrm{in}}^{\mathrm{tor}}.$$

We have $\frac{P_{\mathrm{gr}}N_{\mathrm{in}}^{\mathrm{opt}}}{G-1}$ bidirectional links between each group pair whose bandwidths are B^{opt} Gbps. Thus, to avoid congestion on the links between groups, we set P_{gr} so as to satisfy the following condition;

$$\frac{2B^{\mathrm{opt}}P_{\mathrm{gr}}N_{\mathrm{in}}^{\mathrm{opt}}}{G-1} \geq (T_{\mathrm{tor,opt}}^{\max} + T_{\mathrm{opt,tor}}^{\max}) N_{\mathrm{in}}^{\mathrm{opt}} N_{\mathrm{in}}^{\mathrm{tor}}. \tag{7}$$

2) Parameters of Connection within a Group: We denote the traffic amount on link l as X_l and the set of links between optical packet switches within a group as L. We also denote the set of traffic from a ToR switch to an optical packet switch as $F_{\mathrm{opt}}^{\mathrm{tor}}$, and the set of traffic from an optical packet switch to a ToR switch as $F_{\mathrm{tor}}^{\mathrm{opt}}$.

The sum of the traffic amounts traversing the links within a certain group $\sum_{l\in L} X_l$ satisfies the following condition;

$$\sum_{l\in L} X_l \leq \sum_{i\in F_{\mathrm{opt}}^{\mathrm{tor}}} M_i T_{\mathrm{tor,opt}}^{\max} + \sum_{i\in F_{\mathrm{tor}}^{\mathrm{opt}}} M_i T_{\mathrm{opt,tor}}^{\max},$$

where M_i is the number of links within the group passed by traffic i.

We have $\frac{P_{\mathrm{in}}N_{\mathrm{in}}^{\mathrm{opt}}}{2}$ bidirectional links between optical packet switches within a group. Thus the sum of the bandwidth of the links within a group is $B^{\mathrm{opt}}P_{\mathrm{in}}N_{\mathrm{in}}^{\mathrm{opt}}$. Therefore, Eq. (8) should be satisfied to provide enough bandwidth between all ToR switches.

$$B^{\mathrm{opt}}P_{\mathrm{in}}N_{\mathrm{in}}^{\mathrm{opt}} \geq \sum_{i\in F_{\mathrm{opt}}^{\mathrm{tor}}} M_i T_{\mathrm{tor,opt}}^{\max} + \sum_{i\in F_{\mathrm{tor}}^{\mathrm{opt}}} M_i T_{\mathrm{opt,tor}}^{\max} \tag{8}$$

TABLE I
TOPOLOGIES USED IN OUR EVALUATION

	# of Servers	# of Optical Packet SW	# of Links between Optical SW
Our Topology	2400	24	48
Full Torus	2400	24	48
Parallel Torus	2400	24	48
FatTree (3 layer)	2400	20	32
FatTree (4 layer)	2400	56	140
Switch-based DCell	2400	30	60

Eq. (8) indicates that one approach to provide enough bandwidth between ToR switches is to reduce the average number of hops between ToR switches and optical packet switches. Thus, we connect ToR switches to optical packet switches so as to minimize the average number of hops between the ToR switches and the optical packet switches. Then, we check whether the condition of Eq. (8) is satisfied. If the condition of Eq. (8) is not satisfied, we add more links between optical packet switches within the group.

We set the parameter $P_{\rm in}$ and connection between the ToR switches and the optical packet switches by the following steps.

Step 1 Initialize $P_{\rm in}$ to 2.
Step 2 Construct the topology between optical packet switches including both intra- and inter-group connection based on the current parameter, $P_{\rm in}$.
Step 3 Connect ToR switches to optical packet switches so that the average number of hops between ToR switches and optical packet switches is minimized.
Step 4 Check whether Eq. (8) is satisfied for all groups. If Eq. (8) is satisfied, go to Step 5. Otherwise, go back to Step 2 after incrementing $P_{\rm in}$ by 1.
Step 5 End.

At Step 2 mentioned above, it is required to minimize the average number of hops between ToR switches and optical packet switches. However, it is difficult to obtain the optimal connection between ToR switches and optical packet switches among all possible solutions. In this paper, we select one optical packet switch to be connected to a certain ToR switch so as to minimize the average number of hops from the ToR switch to all optical packet switches at each step, instead of finding the optimal solution among all possible solutions. By continuing this step, we connect all ToR switches to optical packet switches.

IV. EVALUATION

A. Topologies

In this section, we evaluate our topology by comparing it with the topologies shown in Table I.

a) Our Topology: In our evaluation, we set the number of optical packet switches connected to one ToR switch, $P_{\rm opt}^{\rm tor}$ to 2, and the number of ToR switches connected to one optical packet switch, $P_{\rm tor}^{\rm opt}$ to 10. Each ToR switch is connected to 20 servers within a rack. We set the number of optical packet switch within a group $N_{\rm in}^{\rm opt}$ to 6, and the number of groups G to 4. Thus, the number of optical packet switches in our topology is 24. We set the parameters $P_{\rm group}$ and $P_{\rm in}$ by the steps described in Section III-B, setting $B^{\rm opt}$ to 100 Gbps. As a result, $P_{\rm group}$ and $P_{\rm in}$ are set to two.

b) Full Torus: We construct the torus topology using the same number of optical packet switches and the same number of links as our topology. In this evaluation, each optical packet switch of our topology has four ports. Thus, in the full torus topology, we also use the optical packet switches with four ports, and we connect optical packet switches as the 4×6 torus. Similar to our topology, we connect each ToR switch to two optical packet switches and each optical packet switch to ten ToR switches.

c) Parallel Torus: We construct $P_{\rm opt}^{\rm tor}$ torus topologies without links between the different torus topologies. We connect each ToR switch to optical packet switches in the different torus topologies. We use the same number of optical packet switches and the same number of links as our topology. That is, in this evaluation, we use 24 optical packet switches with four ports, and construct two 3×4 torus topologies. Similar to our topology, we connect each ToR switch to two optical packet switches and each optical packet switch to ten ToR switches.

d) FatTree: We construct the FatTree topology using optical packet switches with four ports by method proposed by Al-Fares et al. [1]. This topology is the tree topology with multiple roots, where the half of the ports of an optical packet switch are used to connect it to nodes of the upper layer and the other half of the ports of an optical packet switch are used to connect it to nodes of the lower layer.

Though the method proposed by Al-Fares et al. [1] constructs the 3-layer FatTree, which is constructed of root switches and the pods containing two layers of switches, we can construct higher-layer FatTree topologies. The k-layer FatTree constructed of optical packet switches with four ports includes $(2k-1)2^{k-1}$ optical packet switches.

For our evaluation, we construct two kinds of the FatTree topologies; the 3-layer FatTree topology and the 4-layer FatTree topology using optical packet switches with four ports. We connect ToR switches to the optical packet switches at the lowest layer only. We connect the same number of ToR switches as our topology to both topologies. We set the number of optical packet switches connected to each ToR switch to 2. The number of ToR switches connected to each optical packet switch is 30 and 15 for the 3-layer and 4-layer FatTree topologies, respectively.

e) Switch-based DCell: DCell is the topology for data center networks proposed by Guo et al. [5]. Since the original DCell is constructed by connecting server ports directly, we modify the DCell so as to be used for the connection between optical packet switches. We call the modified version of the DCell *switch-based DCell*.

In the switch-based DCell, a high-layer DCell is constructed from low-layer DCells. We denote the number of optical packet switches in one layer-k DCell as $N_k^{\rm DCell}$. The switch-based DCell is constructed by the following steps. First, layer-0 DCells are constructed by adding links between all pairs of $N_0^{\rm DCell}$ optical packet switches. Then, layer-k DCells are constructed from $N_{k-1}^{\rm DCell}+1$ layer-$k-1$ DCells so that each

layer-$k-1$ DCell is connected to all other layer-$k-1$ DCells with one link.

In our evaluation, we construct the layer-1 switch-based DCell with $N_0^{\mathrm{DCell}} = 5$. Thus, the number of optical packet switches is 30 and the number of ports per optical packet switch is 5, which are larger than our topology. We connect the same number of ToR switches as our topology and set the number of optical packet switches connected to each ToR switch to 2. Thus, the number of ToR switches connected to one optical packet switch is 8. Comparing our topology with this topology, we clarify that our topology can accommodate more traffic than the switch-based DCell even though the switch-based DCell has more links.

B. Properties of Topologies

We compare the topologies by the following metrics.

Edge Betweenness — The edge betweenness of the link l, C_l is defined by

$$C_l = \sum_{s,d \in V, l \in L} \frac{|F_{s,l,d}|}{|F_{s,d}|},$$

where V is the set of nodes which are the source or destination nodes of traffic, L is the set of links, $F_{s,l,d}$ is the set of the shortest paths from nodes s to d passing the link l, and $F_{s,d}$ is the set of the shortest paths from nodes s to d. The edge betweenness indicates the expected number of traffic passing the link. Thus, the topology having the large edge betweenness is easy to be congested. In our evaluation, we calculate the maximum edge betweenness for the traffic between ToR switches.

Minimum Cut — The minimum cut indicates the smallest number of link failures to make the source node unable to reach the destination node. In our evaluation, we calculate the minimum cut for all ToR switch pairs. In all topologies used in our evaluation, each ToR switch is connected to two optical packet switches. Thus, the minimal cut is at most 2.

Table II shows the results. From this table, the minimum cuts of all topologies are 2. That is, all server pairs can communicate with each other even when one link fails in all topologies.

The FatTree topologies have large edge betweenness regardless of the number of layers. Especially, even though the 4-layer FatTree uses more than double optical packet switches and links between optical packet switches compared with other topologies, its edge betweenness is larger than our topology and the parallel torus, and is similar to the full torus. This is caused by the large average number of hops between ToR switches. In the FatTree topologies, a large amount of traffic passes the root optical packet switches, which causes the large average number of hops. The large average number of hops leads to the large expected number of traffic passing a link.

TABLE II
PROPERTIES OF TOPOLOGIES

	Edge Betweenness	Minimum Cut
Our Topology	1000	2
Full Torus	1600	2
Parallel Torus	1200	2
FatTree (3 layer)	2700	2
FatTree (4 layer)	1575	2
Switch-based DCell	2065	2

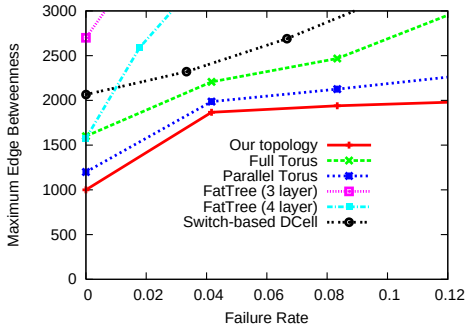


Fig. 4. Edge Betweenness in Case of Failure

The switch-based DCell also has large edge betweenness, even though the switch-based DCell used in our evaluation has more links than our topology and torus topologies. This is because the switch-based DCell has only one link between each layer-0 DCell pair. In the switch-based DCell, we connect many layer-0 DCell pairs by limiting the number of links between each layer-0 DCell pair to one. This makes the number of hops between optical packet switches small. One link between each layer-0 DCell pair, however, cannot provide enough bandwidth.

Compared with the full torus, the parallel torus has smaller edge betweenness. This is caused by close connection between optical packet switches connected to different ToR switches. The parallel torus has more links between optical packet switches connected to different ToR switches instead of connecting optical packet switches connected to the same ToR switch, while the full torus has links between optical packet switches connected to the same ToR switch. This close connection between optical packet switches connected to different ToR switches makes the number of links passed by the traffic between ToR switches small, and reduces the number of traffic between ToR switches passing each link.

Among the topology used in our evaluation, our topology has the smallest edge betweenness. Similar to the parallel torus, our topology uses more links between optical packet switches connected to different ToR switches instead of connecting optical packet switches connected to the same ToR switches. In addition, the parameters in our topology are set by the steps described in Section III-B, which aims to avoid concentration of traffic on certain links. As a result, the parameters of our topology are set so as to make the maximum edge betweenness small.

We also compare the maximum edge betweenness when the randomly selected optical packet switches fail. In this comparison, we generate 100 patterns of random failures, and calculate the average of the maximum edge betweenness for the cases that all servers can communicate with each other. By using this metric, we compare the possibility that congestion occurs when some optical packet switches fail. Figure 4 shows the results. In Fig. 4, the horizontal axis indicates the failure rate of the optical packet switches, and the vertical axis indicates the maximum edge betweenness.

As shown in Fig. 4, the maximum edge betweennesses of the FatTree topologies increase faster than other topologies as the failure rate increases. In the FatTree topologies, because the number of shortest paths between ToR switches passing each link is large, the failure of each optical packet switch affects many ToR switch pairs. Moreover, the paths between the ToR switch pairs affected by the failure also pass many links. As a result, the failure of each optical packet switch has large impacts on the edge betweennesses of many links.

Fig. 4 also indicates that our topology has the smallest edge betweenness even when some optical packet switches fail. As discussed above, our topology has the smallest edge betweenness in the case of no failures. In addition, unlike the FatTree topologies, because the number of shortest paths between ToR switches passing each link and the average number of hops between ToR switches are small, the failure of each optical packet switch affects only few paths between ToR switch pairs, and few links. As a result, the edge betweenness of our method remains the smallest even when some optical packet switches fail.

We have also confirmed that the edge betweenness of our topology does not become large even when the failure rate becomes more than 0.12. However, the probability that ToR switch pairs unable to communicate with each other exist becomes large as the failure rate increases in our topology. In our topology, any optical packet switch has important links that connects different groups. Thus, as the failure rate increases, the number of redundant paths between groups decreases. Finally, when the number of paths between groups becomes 0 due to the failures, the ToR switches belonging to the different groups become unable to communicate with each other. However, as shown in Table II, considering the worst case of failure, no topologies are more robust to failures than our topology. In addition, by setting $P_{\mathrm{opt}}^{\mathrm{tor}}$ to a large value, we can make our topology more robust to failures.

C. Maximum Link Load

In this subsection, we define the link load as the sum of traffic volume passing the link, and we compare the maximum link load without limiting the sum of traffic volume passing each link. In this evaluation, we generate the following two kinds of traffic.

- Uniform Random: Traffic is generated between all server pairs. We add the traffic, whose volume is randomly generated, between the randomly selected server pairs until the NICs of all servers have no remaining bandwidth.
- Certain SW Pair: All of servers connected to the same ToR switch communicate with the servers connected to a certain ToR switch.

For each type of traffic, we randomly generate 20 patterns of traffic and calculate the maximum link load.

In our evaluation, routes of traffic between ToR switches are calculated by the following policies.

- ECMP: Traffic between ToR switch is equally divided among all shortest paths.
- VLB: One intermediate optical packet switch is selected randomly regardless of the destination. Then the traffic is sent from the source ToR switch to the selected intermediate optical packet switch, and from the intermediate optical packet switch to the destination ToR switch.

In this evaluation, similar to Fig. 4, we generate the random failure of optical packet switches and investigate the maximum link utilization in the case that all server can communicate with each other. Figure 5 shows the results. In Fig. 5, the horizontal axis indicates the failure rate of the optical packet switches, and the vertical axis indicates the maximum link loads.

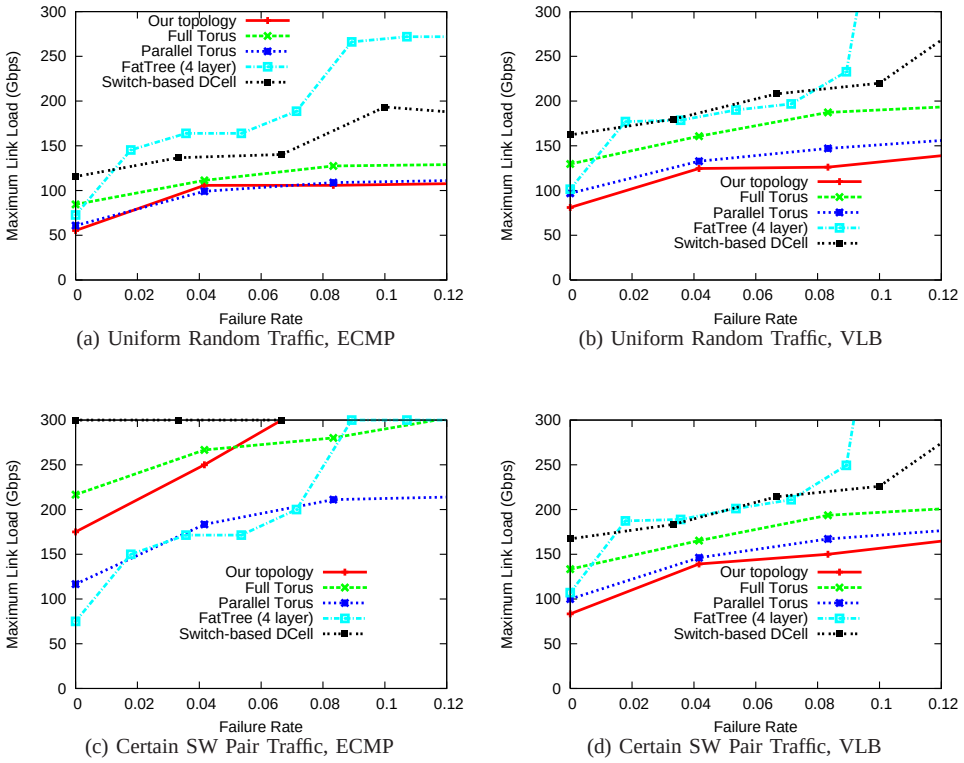
Figs. 5(a) and 5(b) indicate that our topology has the smallest link loads in the case of the uniform random traffic regardless of the routing. In the case of the uniform random traffic, link loads are proportional to the edge betweennesses. Thus, our topology, having the smallest edge betweenness as shown in Fig. 4, has the smallest link loads.

In the case of the certain switch pair traffic, our topology using the ECMP has much larger link loads than the parallel torus. This is caused by the number of distinct shortest paths. While the torus has many distinct shortest paths, the number of distinct shortest paths in our topology is small, which causes concentration of traffic on certain links.

By calculating routes with VLB, however, our topology achieves the smallest link loads even in the case of the certain switch pair traffic. This is because the parameters of our topology are set so as to avoid concentration of traffic on certain links when the routes are calculated by VLB. As shown in Fig. 5, among all pairs of the topologies and routing methods used in our evaluation, only the 4-layer FatTree topology using the ECMP achieves slightly smaller link loads than our topology in the case of no failures. The 4-layer FatTree, however, uses more than double optical packet switches and links between optical packet switches of our topology. In addition, similar to the edge betweenness shown in Fig. 4, the link loads of the 4-layer FatTree increase fast as the failure rate increases. Therefore, our topology is the most suitable topology for accommodating traffic between ToR switches when some optical packet switches fail.

V. Conclusion

The large data center network constructed of only the electronic packet switches consumes large power to provide the enough bandwidth for all server pairs. One approach to construct the data center network that provides sufficient bandwidth with small energy consumption is to use the optical



(a) Uniform Random Traffic, ECMP

(b) Uniform Random Traffic, VLB

(c) Certain SW Pair Traffic, ECMP

(d) Certain SW Pair Traffic, VLB

Fig. 5. Maximum Link Load

packet switches. In the data center network using the optical packet switches, however, the failures of the optical packet switches may have large impacts on the communication between servers.

In this paper, we proposed the data center network topology using the optical packet switches that can provide enough bandwidth even when some optical packet switches fail. We evaluated our topology and clarified that our topology can provide enough bandwidth even when some optical packet switches fail.

One of our future research topics is to compare our topology using optical packet switches with the existing data center networks using only electronic packet switches.

Acknowledgement

This research was supported by the National Institute of Information and Communications Technology (NICT).

References

[1] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proceedings of ACM SIGCOMM*, vol. 38, pp. 63–74, Aug. 2008.

[2] 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," *ACM SIGCOMM Computer Communication Review*, vol. 39, pp. 51–62, Aug. 2009.

[3] R. N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "PortLand: A scalable fault-tolerant layer 2 data center network fabric," *ACM SIGCOMM Computer Communication Review*, vol. 39, pp. 39–50, Aug. 2009.

[4] J. Kim, W. Dally, and D. Abts, "Flattened butterfly: a cost-efficient topology for highradix networks," in *Proceedings of the 34th annual international symposium on Computer architecture*, vol. 35, pp. 126–137, June 2007.

[5] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "DCell: A scalable and fault-tolerant network structure for data centers," *ACM SIGCOMM Computer Communication Review*, vol. 38, pp. 75–86, Aug. 2008.

[6] D. Li, C. Guo, H. Wu, K. Tan, Y. Zhang, S. Lu, and J. Wu, "Scalable and cost-effective interconnection of data-center servers using dual server ports," *IEEE/ACM Transactions on Networking*, vol. 19, pp. 102–114, Feb. 2011.

[7] D. Guo, T. Chen, D. Li, Y. Liu, X. Liu, and G. Chen, "Bcn: expansible network structures for data centers using hierarchical compound graphs," in *Proceedings of INFOCOM*, pp. 61–65, Apr. 2011.

[8] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: A high performance, server-centric network architecture for modular data centers," *ACM SIGCOMM Computer Communication Review*, vol. 39, pp. 63–74, Aug. 2009.

[9] Y. Liao, D. Yin, and L. Gao, "Dpillar: Scalable dual-port server interconnection for data center networks," in *Proceedings of ICCCN*, pp. 1–6, Aug. 2010.

[10] R. Urata, T. Nakahara, H. Takenouchi, T. Segawa, H. Ishikawa, A. Ohki, H. Sugiyama, S. Nishihara, and R. Takahashi, "4x4 optical packet switching of asynchronous burst optical packets with a prototype, 4x4 label processing and switching sub-system," *Optics Express*, vol. 18, pp. 15283–15288, July 2010.

[11] H. J. Chao and K. Xi, "Bufferless optical clos switches for data centers," in *Proceedings of OFC*, Mar. 2011.

[12] K. Xi, Y. H. Kao, M. Yang, and H. J. Chao, "Petabit optical switch for data center networks." Technical Report, Polytechnic Institute of New York University, `http://eeweb.poly.edu/~chao/publications/petasw.pdf`.

[13] M. Kodialam, T. V. Lakshman, and S. Sengupta, "Efficient and robust routing of highly variable traffic," in *Proceedings of HotNets*, Nov. 2004.