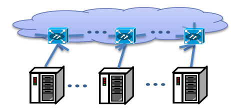
DATA CENTER NETWORK TOPOLOGIES USING OPTICAL PACKET SWITCHES

Yuichi Ohsita
Osaka University

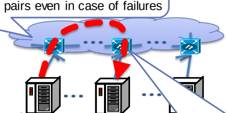
Data center

- Constructed of many servers and a network between servers.
  - Servers communicate with each other to handle large data.
  - Large data centers with hundreds of thousands of servers have been built.
- Network within the data center has large impacts on the performance of the data center.
  - Insufficient bandwidth prevents communication between servers.



Requirements for data center networks

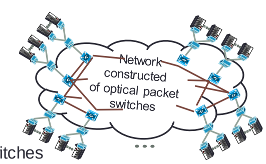
- Keep connectivity between all servers
  - To enable the communication between any server pairs even in case of failures



- Connect many servers
  - To be used in a large data center built in recent years
- Provide sufficient bandwidth between all servers
  - All servers cooperate with each other to handle large data

Data center network constructed of optical packet switches

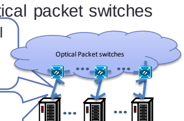
- Requirements for data center networks
  - Provide sufficient bandwidth between all servers
  - Consume small energy
  - Keep the sufficient bandwidth even in case of failures



- Advantages of optical packet switches
  - Provide large bandwidth between their ports with small energy consumption
- Optical packet switches can construct networks that provide sufficient bandwidth with small energy consumption.
- We construct data center network using optical packet switches that is robust to failures.

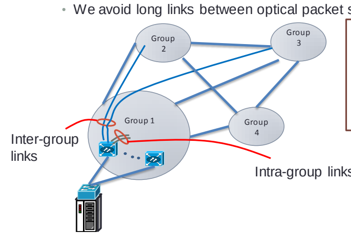
Goal of this work

- Goal
  - Construct a data center network using optical packet switches efficiently
    - Provide sufficient bandwidth between all servers by using optical packet switches.
    - Keep the connectivity between all servers even when some optical packet switches fail.
- Our data center networks using optical packet switches
  - Construct a core network using optical packet switches
    - To use the large bandwidth of optical packet switches efficiently
  - Connect one server rack to multiple optical packet switches.
    - To keep connectivity even when optical packet switches fail



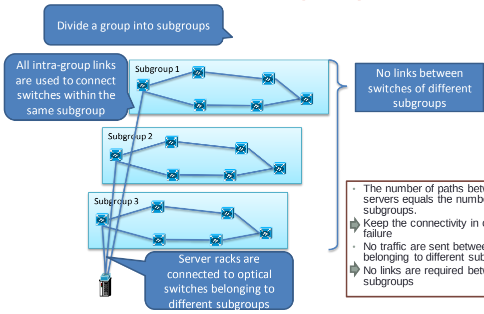
Our network structure

- We divide optical packet switches and server racks into multiple groups.
  - Each server rack is connected to the optical packet switches belonging to the same group.
    - We avoid long links between optical packet switches and server rack



Parameters
- Number of inter-group links per switch
- Number of intra-group links per switch
- Number of optical packet switches connected from one sever rack

Connection within a group



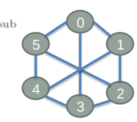
- The number of paths between servers equals the number of subgroups.
- ⇨ Keep the connectivity in case of failure
- No traffic are sent between switches belonging to different subgroups
- ⇨ No links are required between subgroups

Connection within a subgroup

- Connect optical packet switches according to the ID assigned to the switches.
  - Parameters
    - Number of switches in a subgroup
    - Number of links used to connect a switch to the switches belonging to the same subgroup
  - Steps
    1. Construct a ring topology by connecting switches of the nearest ID
    2. Add links between switches S_1 and S_2 if the following condition is satisfied.
       - Connect switches so that the intervals of switches connected to a certain switch are constant

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

$D^{\mathrm{opt}}(S_1)$	ID of switch S_1
P_{in}	Number of intra-group links per switch
N_{sub}	Number of switches within a subgroup
i	Integer variable



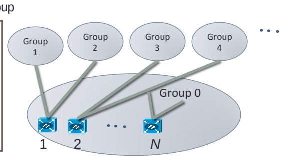
Connection between groups

- Connect switches according to the ID of group and the ID of switch
  - Connect switch S_1 and the switch belonging to $D^{\mathrm{gr}}(S_2)$ if the following condition is satisfied.

$$D^{\mathrm{in}}(S_1) = \begin{cases} \left\lfloor \frac{D^{\mathrm{gr}}(S_2)+K(G-1)}{P_{\mathrm{gr}}} \right\rfloor & (D^{\mathrm{gr}}(S_1) \geq D^{\mathrm{gr}}(S_2)) \\ \left\lfloor \frac{D^{\mathrm{gr}}(S_2)+K(G-1)-1}{P_{\mathrm{gr}}} \right\rfloor & (\mathrm{Otherwise}) \end{cases}$$

  - Connect so that the intervals of the IDs of switches connected to the same group become constant.
    - To avoid large number of hops to the groups
  - According to the destination group ID and the IDs of the switches within a group, we can identify the switch connected to the group

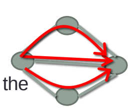
$D^{\mathrm{opt}}(S_1)$	ID of the switch S_1
$D^{\mathrm{gr}}(S_1)$	ID of the group the switch S_1 belongs to.
P_{gr}	Number of links between groups per switch
G	Number of groups
K	Positive Integer



Overview of parameter settings

- Parameters
  - Number of inter-group links per switch
  - Number of intra-group links per switch
  - Connection between sever racks and optical packet switches
- Input
  - Maximum traffic volume from a server rack
  - Number of server racks connected to one optical packet switch
  - Number of groups
  - Number of optical packet switches in a group
- Objective
  - Accommodate any traffic without limiting the bandwidth between servers
- Approach
  - Set parameters considering a load balancing method.

Valiant Load Balancing



- Avoid concentration of traffic by randomly selecting an intermediate switch regardless of the destination
  - Traffic from a server rack to an optical packet switch
    ＝Traffic volume from a server rack/Number of optical packet switches
  - Traffic from an optical packet switch to a server rack
    ＝Traffic volume to a server rack/Number of optical packet switches

⇩

We can calculate the maximum traffic volumes for all possible traffic

- We set parameters so as to accommodate traffic volume calculated considering the VLB

Setting parameters of inter-group connections

- Set the number of inter-group connections so as to satisfy the following condition.

Sum of bandwidths between a group pair
≧Total volume of traffic between a group pair

- Sum of bandwidths between a group pair
  =Number of links between a group pair × Bandwidth of a link
  - Number of links between a group pair
    =(Number of switches in a group × Number of intra-group links per switch) / Number of groups
- Total volume of traffic between a group pair
  ＝Number of servers in a group × Number of switches in a group
  × Traffic volume between a server rack and an optical packet switch
  - Traffic volume is calculated considering the VLB.

13

Setting parameters of intra-group connections

- Set the number of inter-group connections so as to satisfy the following condition.

Sum of bandwidths of intra-group links
≧Total volume of traffic passing links within a group

- Sum of bandwidths of intra-group links
＝Number of switches in a group × Number of intra-group links per switch
×Bandwidth of a link
- Total volume of traffic passing links within a group ＝
$\sum(\textit{Number of hops within a group} \times \textit{Traffic volume between a server rack and a switch})$

To accommodate more traffic, reduction of the number of hops is effective

14

Steps to set parameters of intra-group connections

1. Initialize the number of intra-group links as 2.
2. Construct the topology of the optical packet switches based on the current parameter
3. Connect servers to optical packet switches so that the number of hops between servers and switches become small
4. Check whether sufficient bandwidth can be provided in the current topology

If the bandwidth is insufficient → Add the number of intra-group links → (back to step 2)

15

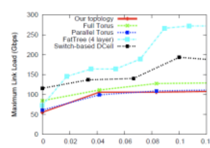
Evaluation

- Network parameters
  - Number of optical packet switches： 24
  - Number of groups： 6
  - Number of optical packet switches connected to the same server rack： 2
  - Number of servers connected under an optical packet switch: 200
- Topologies used in our comparison
  - We compare the topologies using the same number of optical packet switches with the same number of ports as our topology
    - FatTree、Dcell
    - Full Torus
      - Torus topology constructed of all optical packet switches
    - Parallel Torus
      - 2 torus topologies without links between the different torus topologies.
      - Each server rack is connected to both of torus topologies

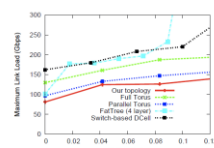
16

Maximum link load (Uniform random traffic)

- Metric
  - Maximum link load
    - Traffic：Randomly generated between all servers
    - Failures：Randomly selected optical packet switches fail
- Result
  - Even in case of failure, our topology has the smallest link load.



Minimum hop

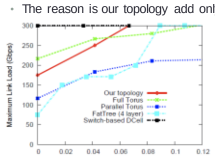


Valiant Load Balancing

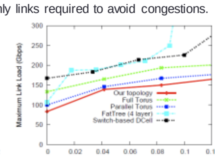
17

Maximum link load (Certain SW pair)

- Metrics
  - Maximum link load
    - Traffic：Each server rack generate traffic to only one selected server rack
    - Failure：Randomly selected optical packet switches fail
- Results
  - Our topology achieves the smallest link load by using the VLB even in case of failure
    - The reason is our topology add only links required to avoid congestions.



Minimum hop



Valiant Load Balancing

18

Conclusion

- Construct the data center network using the optical packet switches efficiently
  - Use the large bandwidth of optical packet switches efficiently.
  - Keep the connectivity between any servers even in case of failure.
- Propose a method to set parameters of our data center network suitable to the data center network using the optical packet switches
- Future work
  - Evaluation of topologies considering the properties of optical packet switches more.
  - Construct the topology considering the latency