A Scalable, Commodity Data Center Network Architecture

Paper Note
Junzhi Gong

September 22, 2018

1 Backgrounds

- Clusters in datacenters consisting of tens of thousands of PCs are common in large institutions. Important applications classes include scientific computing, financial analysis, data analysis and warehousing, and large-scale network services.
- Today, the principle bottleneck in large-scale clusters is often inter-node communication bandwidth.
- There are two high-level choices for building the communication fabric for large-scale clusters. One option leverages specialized hardware and communication protocols, and the other option leverages commodity Ethernet switches and routers to interconnect cluster machines.
- Communication bandwidth in large clusters may become oversubscribed by a significant factor depending on the communication patterns.

2 Target

Design a data center communication architecture that meets the following goals:

1. Scalable interconnection bandwidth: it should be possible for an arbitrary host in the data center to communicate with any other host in the network at the full bandwidth of its local network interface.
2. Economies of scale: just as commodity personal computers became the basis for large-scale computing environments, we hope to leverage the same economies of scale to make cheap off-the-shelf Ethernet switches the basis for large-scale data center networks.
3. Backward compatibility: the entire system should be back- ward compatible with hosts running Ethernet and IP. That is, existing data centers, which almost universally leverage com- modity Ethernet and run IP, should be able to take advantage of the new interconnect architecture with no modifications.

3 Current Topologies

- **Topology.** Typical topology used in datacenters consists of two levels or three levels. Those levels include root switches, aggregate switches (optional), and leaf switches.
- **Oversubscription.** Many data center designs introduce oversubscription as a means to lower the total cost of the design. The oversubscription is defined as the ratio of the worst-case achievable aggregate bandwidth among the end hosts to the total bisection bandwidth of a particular communication topology.
- **Multi-path routing.** Delivering full bandwidth between arbitrary hosts in larger clusters requires multi-path routing, like ECMP.
- **Cost.** The cost for building a network interconnect for a large cluster greatly affects design decisions. Overall, we find that existing techniques for delivering high levels of bandwidth in large clusters incur significant cost and that fat-tree based cluster interconnects hold significant promise for delivering scalable bandwidth at moderate cost.
- **Clos networks and the fat-tree.** Clos network topology delivers high levels of bandwidth for many end devices by appropriately interconnecting smaller commodity switches. Fat-tree is a special instance of Clos network topology. Advantages are:
  - All switching elements are identical, enabling us to leverage cheap commodity parts for all of the switches in the communication architecture.
  - Fat-trees are rearrangeably non-blocking, meaning that for arbitrary communication patterns, there is some set of paths that will saturate all the bandwidth available to the end hosts in the topology.