Scale of Networks: Data Centers and Beyond
Laiyfeng Royston Lai, Angela Lin, Yiming Gu
Advisors: Alex Snoeren, George Porter

Background

Problem: maintaining high performance in an oversubscribed network. (connecting multiple devices to the same switch ports)

There are many design approaches to handle an oversubscribed network.

Choice 1: Topology - the way the constituent parts are arranged.
- Clos topology (Google)
- Dynamic topology (OSA)

For example, a clos topology is fully connected, while a dynamic topology is not fully connected, but can change the bandwidth allocation depending on the demands of the end hosts.

Choice 2: Type of Switching Network
- Circuit Switching Network
- Packet Switching Network

Two common types of switching networks today
- Optical Circuit Switching (OCS): sending data through a dedicated path using light particles.
- Electrical Packet Switching (EPS): transfer data through packets using any path in the network.

Architectures that employ EPS will have the advantage of more efficient data transfer rates than OCS, but EPS will not be able to support high transfer speeds at long distances due to the limits of copper cables.

We compare four network architectures that balance the tradeoffs between efficiency, power, and cost in various ways:
- REACToR is a Top-of-Rack switch (ToR) prototype that uses an overlaid OCS network on top of an existing EPS fabric. It seeks to increase the efficiency of a data center by improving the link speeds between server racks and pods.
- Helios is a predecessor to REACToR that also combines OCS and EPS, but it targets the link speeds from pods to the rest of the network.
- Jupiter is an EPS network design, capable of meeting the demands of Google’s data centers.
- OSA is a purely OCS design, targeting communications between ToRs in a data center.

Architectural Overviews

REACToR
- Handles inter-pod layer
- 9:1 bandwidth ratio (OPS: EPS)
- EPS always transferring data, even during OCS reconfiguration

Helios
- Handles pod-core switching layer
- 1:1 bisection bandwidth (OPS:EPS)

OSA
- Handles pod-core switching layer

Jupiter
- Handles all layers of data center

Citations

Acknowledgements
Christine Alvarado, Michael Barrow

This material is based upon work supported by the CSE Early Research Scholars Program founded by the National Science Foundation under Grant No. CNS-1033935.

Conclusion

All these architectures offer a solution to handling an over-subscribed network.

Jupiter provides a lot of bandwidth to each end-host, but it requires a lot of money to build such an architecture, not very scalable, and a lot of power is consumed by the switches.

OSA has a good data transfer rate, but the delays during reconfiguration are costly and heavily affects performance.

The hybrid architecture of REACToR offers a cost efficient means to get good performance, and being deployed at the ToR level, being very scalable. It reduces performance overhead from OCS delays by constant packet transfer through EPS.

Hence, if cost is not an issue, a clos topology network similar to Jupiter can be employed at all layers for high performance; otherwise, a cost efficient architecture possibility can be having REACToR connect the server racks, and a Jupiter layout connecting the pods to the core switches in the data centers.

Circuit Switching Network
Packet Switching Network
Type of Switching Network

Problem
Architectures Optimizations

Prototype
Network Optimizations
Advantages / Disadvantages

REACToR
Hybrid topology, EPS&OCS
- During OCS reconfiguration, Priority Flow Control (PFC) frame sent in advance to pause traffic to OCS at end hosts, and packets temporarily stored at end host buffer. Traffic to EPS undisturbed, which eases congestion during reconfiguration
- Easy to scale up
- Shorter reconfiguration times than Helios & OSA (+)
- Capable of running faster servers than Jupiter
- High throughput due to high bandwidth handled by OCS (+)
- Connecting multiple REACToRs together requires complex scheduling and synchronization algorithms (-)

Helios
Dynamic topology, EPS&OCS
- Helios has a topology manager that measures traffic, calculates the optimal topology, and then dynamically reconnects the topology to ameliorate the congestion
- Hardware independent (+)
- Generally higher power consumption (-)

OSA
Dynamic topology, EPS&OCS
ToRs can flexibly use more or less ports to change link capacity to beat congestion
- Achieves nearly 60% of non-blocking bandwidth in all-to-all communication (+)
- Not optimal for small flows due to large reconf delays (-)
- Hard to scale up(-)

Jupiter
Clos topology, EPS
- Connecting multiple REACToRs together
- Capable of running faster servers than Jupiter
- Shorter reconfiguration times than Helios & OSA (+)

Techniques to alleviate congestion:
- Discard lower priority traffic
- Employ link-level pause at ToRs to prevent servers from over-running oversubscribed uplinks.
- Utilize shared memory buffers of merchant silicon

Supports 1.3Pbps bisection bandwidth among servers in its largest configuration (+)
Can be used in very large data centers (+)
Fixed architecture once it is built (-)
Much of power spent on switching (-)
Very costly (-)

Techniques to alleviate congestion:
- Discard lower priority traffic
- Employ link-level pause at ToRs to prevent servers from over-running oversubscribed uplinks.
- Utilize shared memory buffers of merchant silicon

Supports 1.3Pbps bisection bandwidth among servers in its largest configuration (+)
Can be used in very large data centers (+)
Fixed architecture once it is built (-)
Much of power spent on switching (-)
Very costly (-)