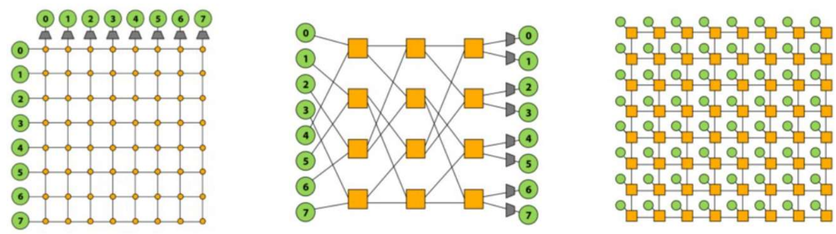
Addendum to Lecture # 5 (Interconnects)

We got confused about O(N) cost for the Mesh. Here are few suggestions:

(1) We need to be clear what is N. Are they nodes ("things" that can potentially read or write to the interconnect)? Or is N the number of links in the topology? At times, we can define number of links in terms of number of nodes.
(2) We need to be clear cost measures what? Is it number of switches? Is it number of links?

Review: network topologies



Topology	Crossbar	Multi-stage log.	Mesh
Direct/Indirect	Indirect	Indirect	Direct
Blocking/ Non-blocking	Non-blocking	Blocking (one discussed in class is, others are not)	Blocking
Cost	$O(N^2)$	O(N lg N)	O(N)
Latency	O(1)	O(lg N)	O(sqrt(N)) (average)

CMU 15-418/618, Spring 2017

Ungraded Homework:

Following is from Dr. Rana Asif's Slides. Here instead of N, we are using P. Try to prove the numbers. You might like to use **Induction** for the proof or might like to use the **specific geometry** of the topology.

➨ ***Arc-connectivity:*** *The minimum number of arcs or links that must be removed from the network, to break the network into two disconnected networks*

Evaluating Static Interconnections

Network	Diameter	Bisection Width	Arc Connectivity	Cost (No. of links)
Completely-connected	1	$p^2/4$	$p-1$	$p(p-1)/2$
Star	2	1	1	$p-1$
Complete binary tree	$2\log((p+1)/2)$	1	1	$p-1$
Linear array	$p-1$	1	1	$p-1$
2-D mesh, no wraparound	$2(\sqrt{p}-1)$	$\sqrt{p}$	2	$2(p-\sqrt{p})$
2-D wraparound mesh	$2\lfloor\sqrt{p}/2\rfloor$	$2\sqrt{p}$	4	$2p$
Hypercube	$\log p$	$p/2$	$\log p$	$(p\log p)/2$

CS3006 - Spring 2022

References for going deeper:

(1) Prof. Onur Mutlu's video lectures on interconnects:
   a. http://www.youtube.com/watch?v=jnJpbZUKrJ4
   b. http://www.youtube.com/watch?v=rksmfG_5fXo

(2) See freely available appendix F on interconnects. There is a section on historical perspectives as well.
https://elsevier.widen.net/content/nwcwklhics/original/CompanionAsset_9780128119051_Hennessy_References_Appendices.zip?u=ebnrhc&download=true

(3) Book: Principles and Practices of Interconnection Networks https://a.co/d/eJSouVC

Datacenter networking:

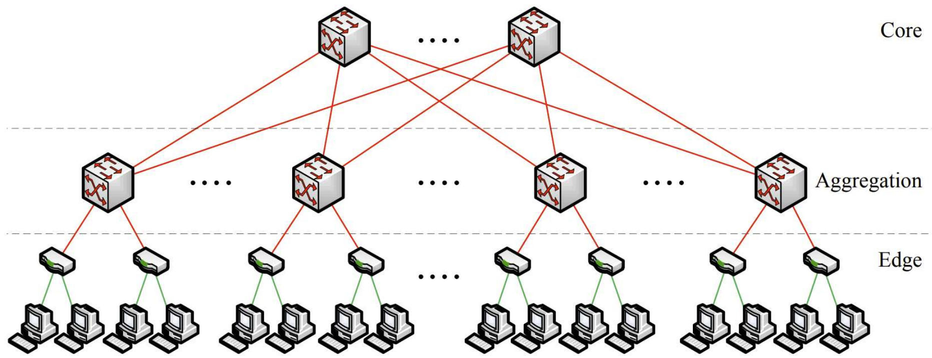


Figure 1: Common data center interconnect topology. Host to switch links are GigE and links between switches are 10 GigE.

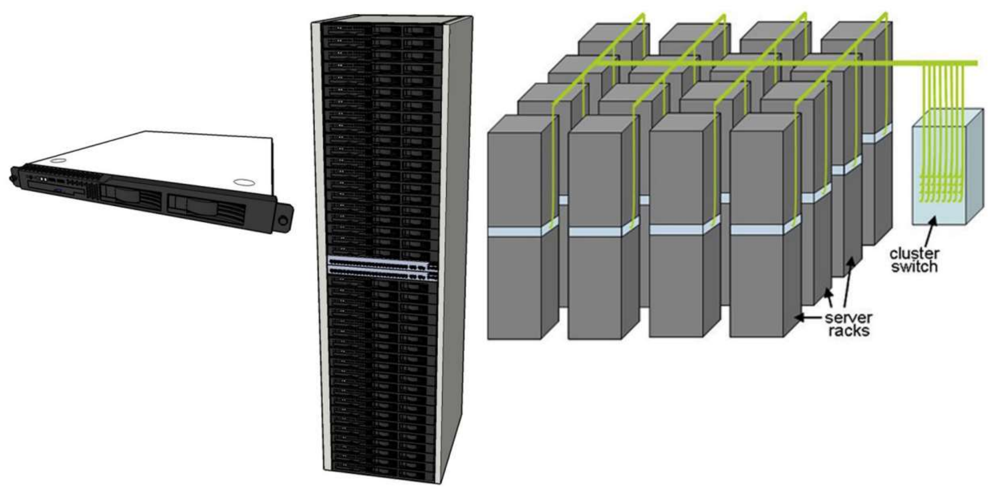
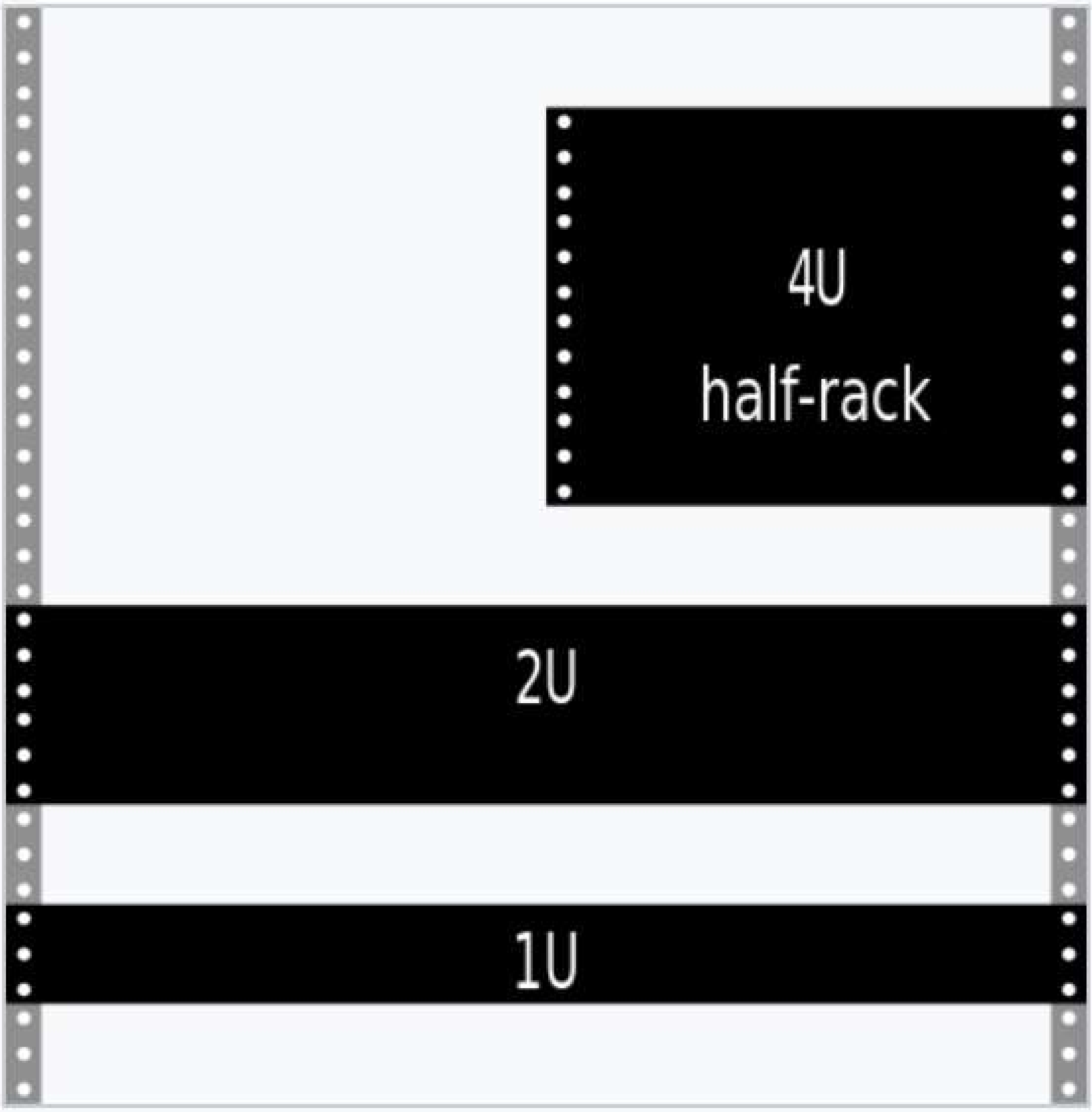


FIGURE 1.1: Typical elements in warehouse-scale systems: 1U server (left), 7´ rack with Ethernet switch (middle), and diagram of a small cluster with a cluster-level Ethernet switch/router (right).



Rack with sample component sizes including an A/V half-rack unit

“Ordinary servers are usually 3U high, meaning a rack theoretically can hold 14 servers. However, by reducing server height to 2U or 1U, a rack can hold 21 or **42 servers**—increasing the processing power by 50% to 100% in the same floor space.”

“One rack **unit (1U) is 1.75”** (44.45 mm) of vertical space, or typically the equivalent of three rack hole spaces tall. One of the first criteria to consider when purchasing a rack is how many RUs your equipment requires.”

“Oversubscription

Many data center designs introduce oversubscription as a means to lower the total cost of the design. **We define the term oversubscription to be the ratio of the worst-case achievable aggregate bandwidth among the end hosts to the total bisection bandwidth of a particular communication topology**. An oversubscription of 1:1 indicates that all hosts may potentially communicate with arbitrary other hosts at the full bandwidth of their network interface (e.g., 1 Gb/s for commodity Ethernet designs). An oversubscription value of **5:1 means that only 20% of available host bandwidth is available for some communication patterns**. Typical designs are oversubscribed by a factor of 2.5:1 (400 Mbps) to 8:1 (125 Mbps) [1]. Although data centers with oversubscription of 1:1 are possible for 1 Gb/s Ethernet, as we discuss in Section 2.1.4, the cost for such designs is typically prohibitive, even for modest-size data centers. Achieving full bisection bandwidth for 10 Gb/s Ethernet is not currently possible when moving beyond a single switch.”

[Credit: https://www.cs.yale.edu/homes/yu-minlan/teach/csci599-fall12/papers/fattree.pdf]

Following figures are Circs 2008:

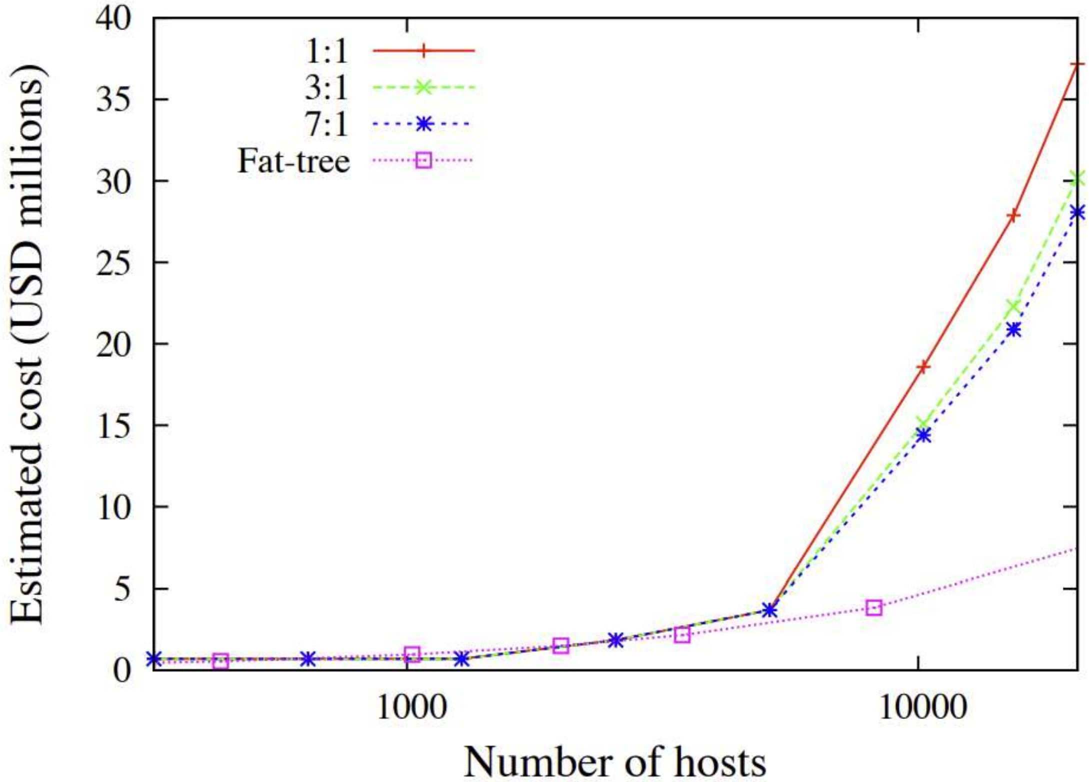


Figure 2: Current cost estimate vs. maximum possible number of hosts for different oversubscription ratios.

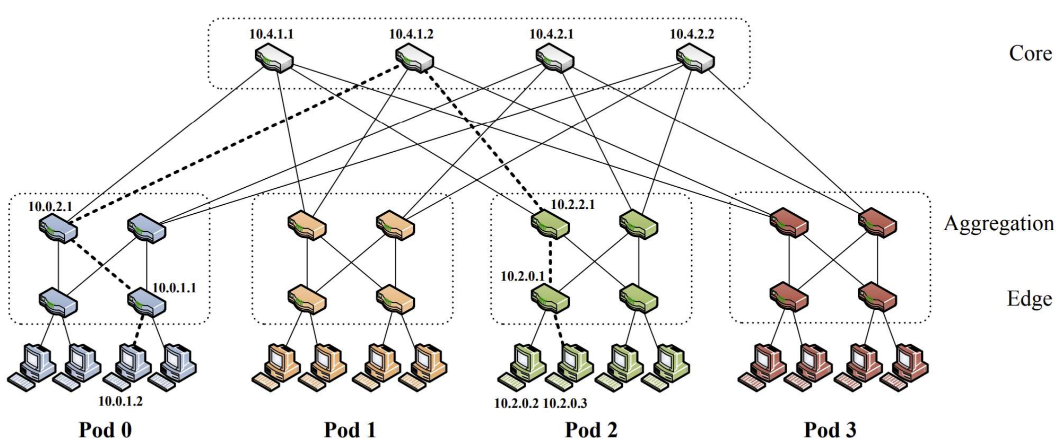
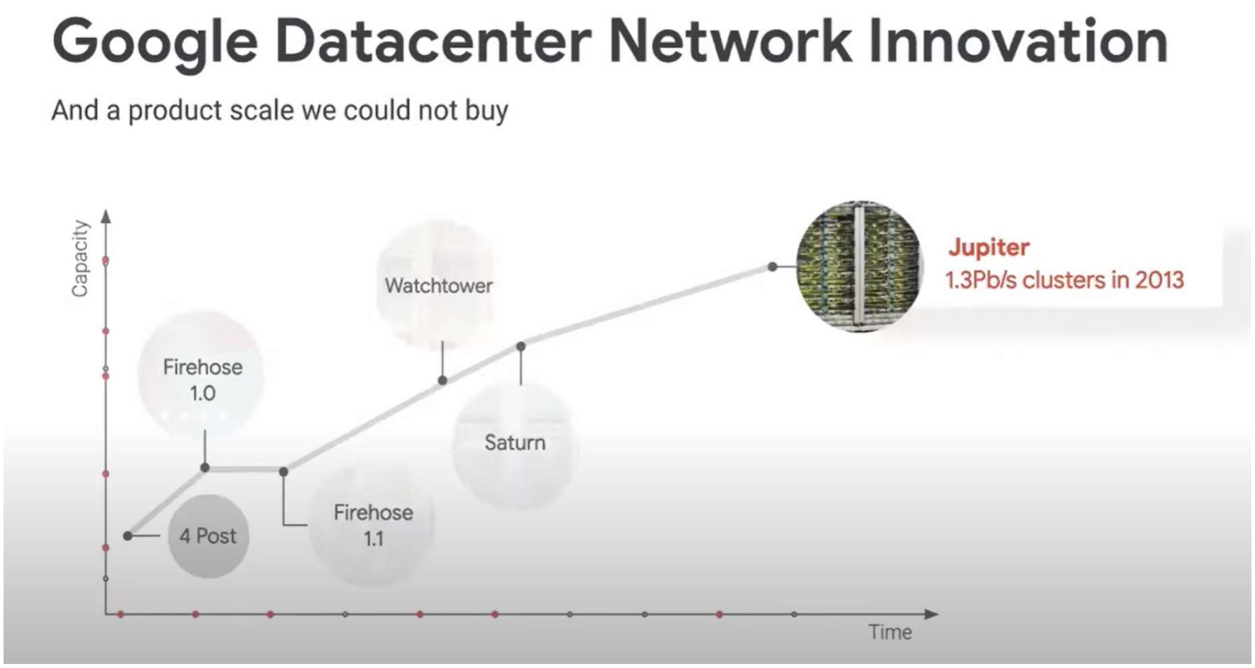
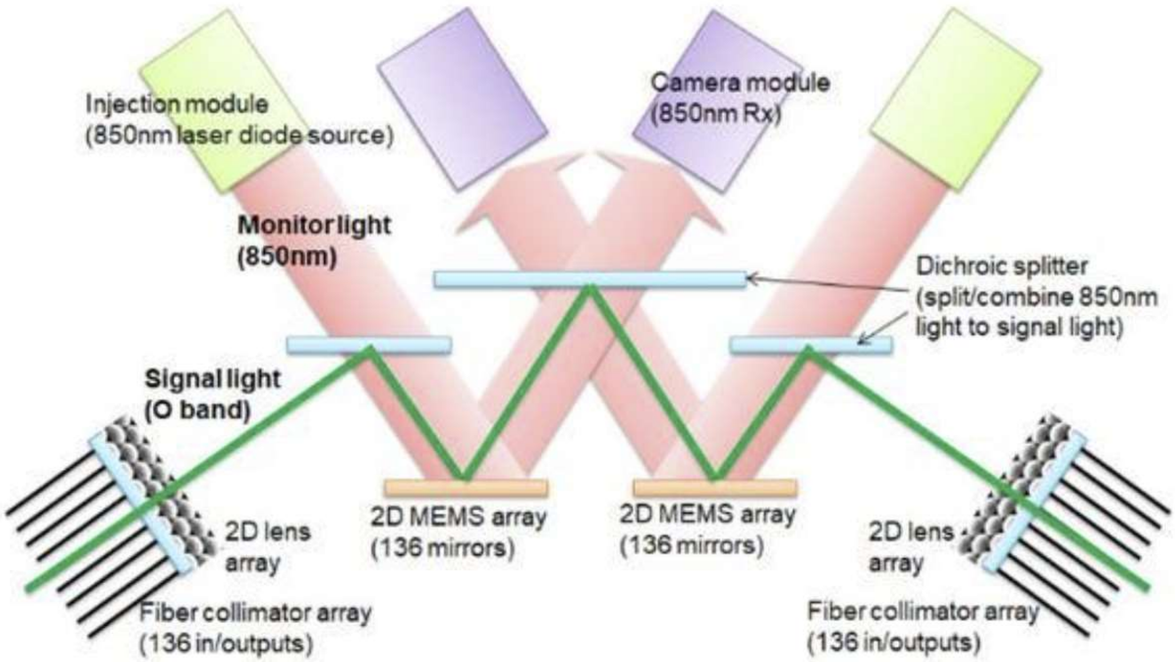


Figure 3: Simple fat-tree topology. Using the two-level routing tables described in Section 3.3, packets from source 10.0.1.2 **to destination** 10.2.0.3 **would take the dashed path.**



[Credit: Screengrab from: https://www.youtube.com/watch?v=Am_itCzkaE0&t=2990s]

Latest innovations: Brining **optical circuit switching** to our Jupiter data center networking and in support of large-scale ML training



Google Apollo: The >$3 Billion Game-Changer in Datacenter Networking

[Credit: https://www.linkedin.com/posts/vahdat_google-apollo-the-3-billion-game-changer-activity-7043400337700880384-yrEt/]

There are tons of details that we cant go into in this class. That is a topic for another class such as graduate level course on Datacenter Networking.