EN2910A: Advanced Computer Architecture
Topic 06: Supercomputers & Data Centers

Prof. Sherief Reda
School of Engineering
Brown University

Material from:

- *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second Edition*
- *Computer Architecture: A Quantitative Approach* by Hennessy and Patterson

Warehouse-scale computers

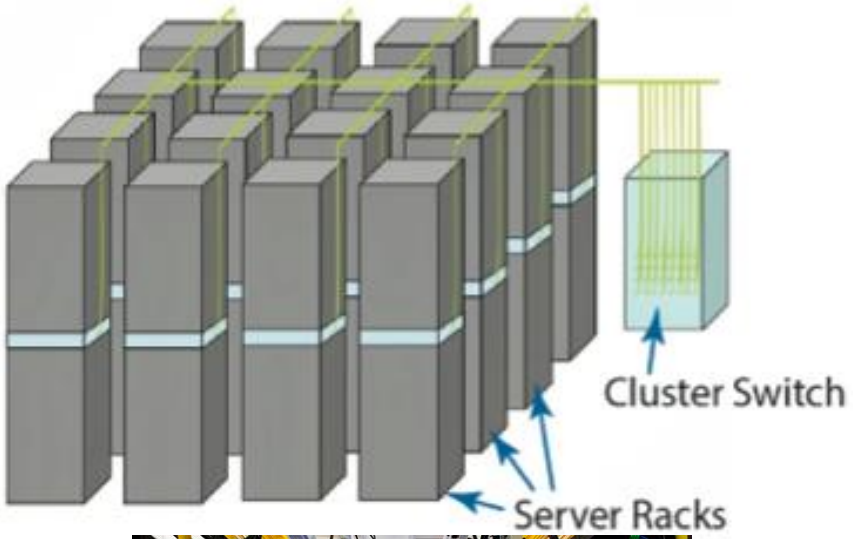
The data center as a computer. L. Barroso

Performance metrics for clusters

- **Supercomputers:**
  - Execution time
  - Threads communicate using message passing
  - FLOPS (FLOP/s): theoretical peak or using a standard benchmark (e.g., LINPACK is used for Top-500 supercomputer ranking)

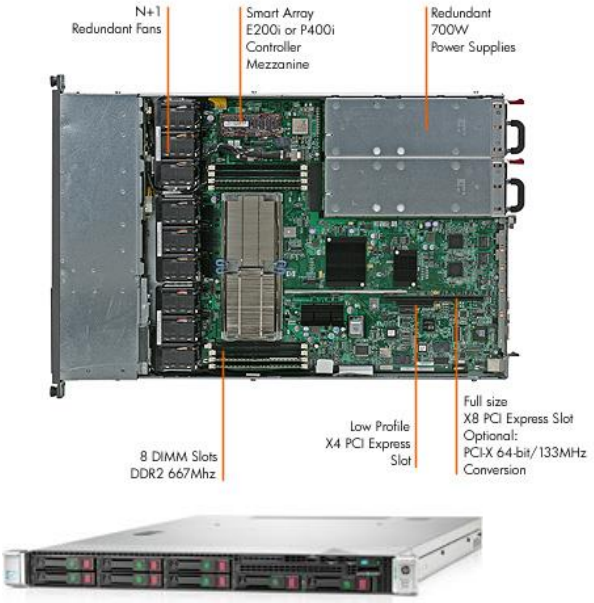
- **Data center scale:**
  - Abundant parallelism through request-level parallelism
  - Latency is important metric because it is seen by users
  - Bing study: users will use search less as response time increases
  - Service Level Objectives (SLOs)/Service Level Agreements (SLAs). E.g. 99% of requests be below 100 ms

Anatomy of data center

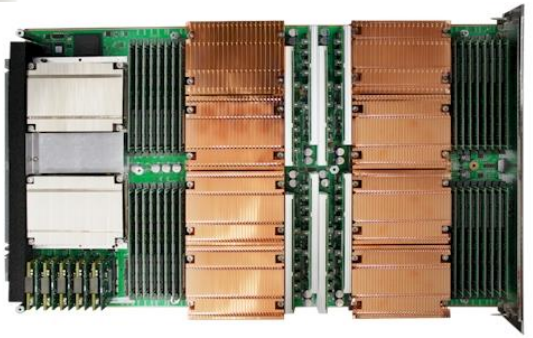


- 7-ft rack (42U) holds 40 1U servers and a rack-level Ethernet switch
- Rack switches have uplinks connecting to cluster-level switches

Servers

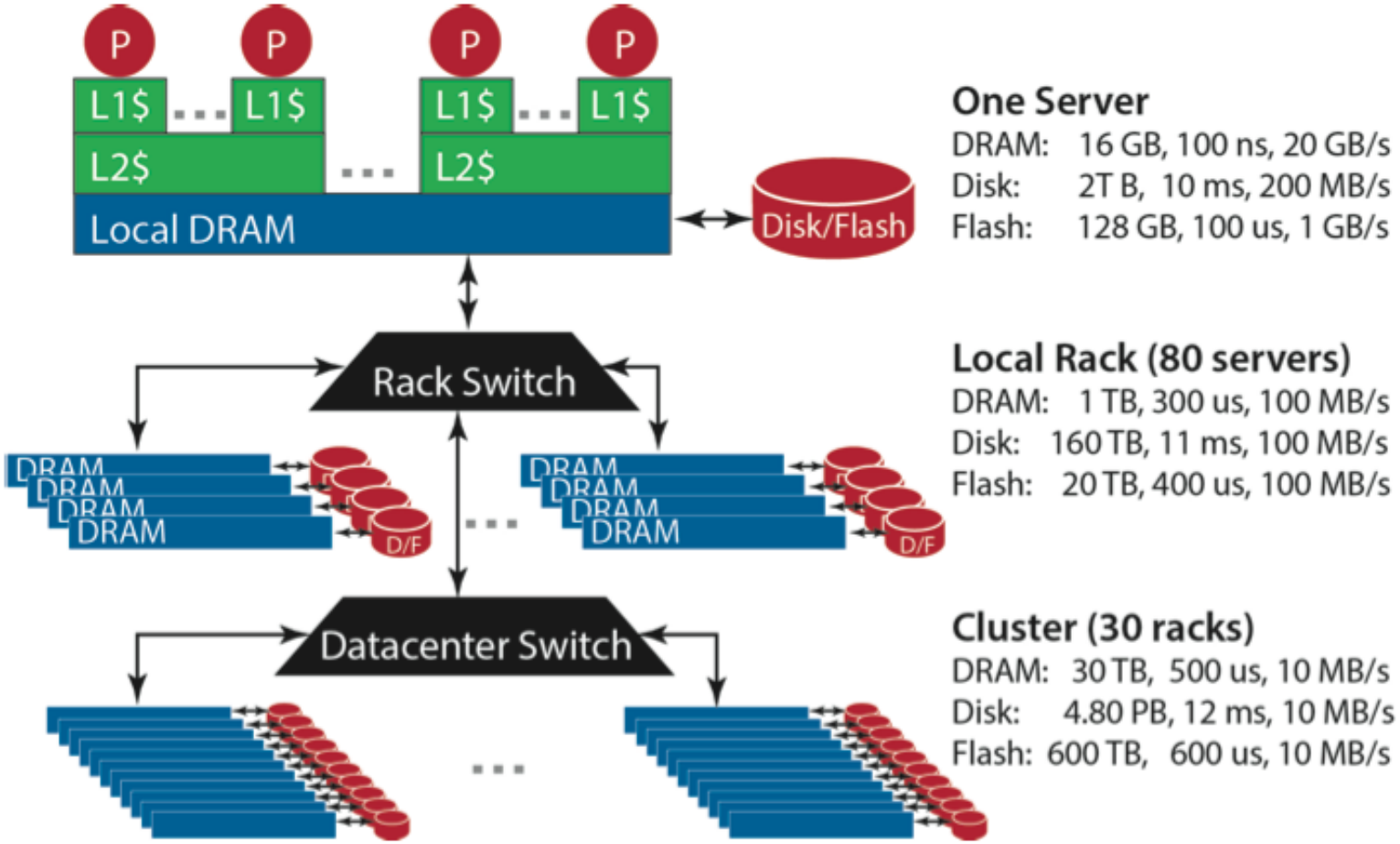


Blade chassis



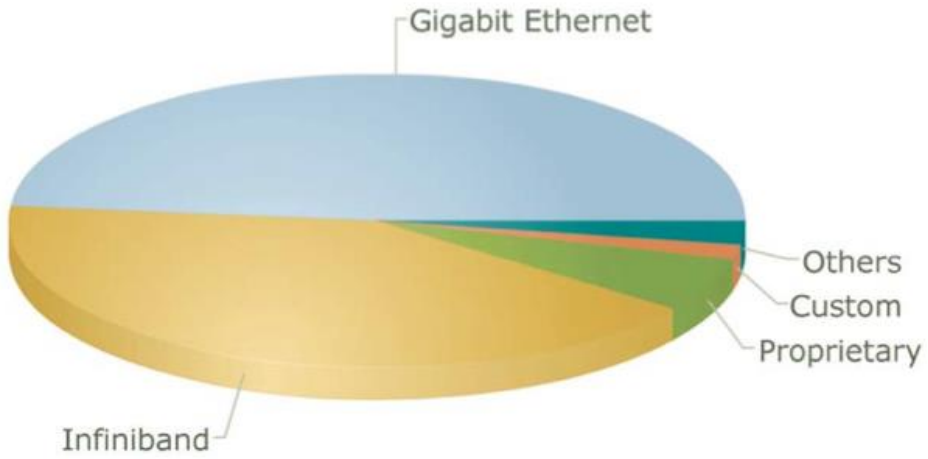
- Data centers:
  - 1U servers; each has two quad-core processors & 16 GB
  - Disk attached to each node, 10Gbe Ethernet
- Supercomputers
  - Use more expensive high-density blade servers and GPGPU
  - Blade chassis has number of blades (e.g., Cray XE6 blade with four 16-core AMD Opteron processors with 64 GB).
  - Chassis provides shared power supply and cooling

Storage hierarchy of a data center



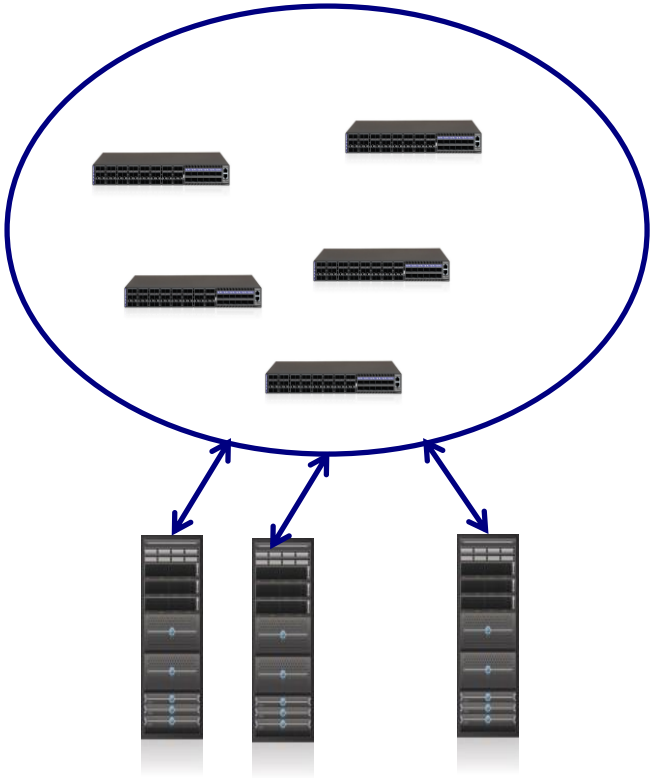
Data centers use global distributed storage (disks attached to each compute node), whereas supercomputers use Network Attached Storage (NAS) devices connected directly to the cluster network.

Supercomputer and data centers networks

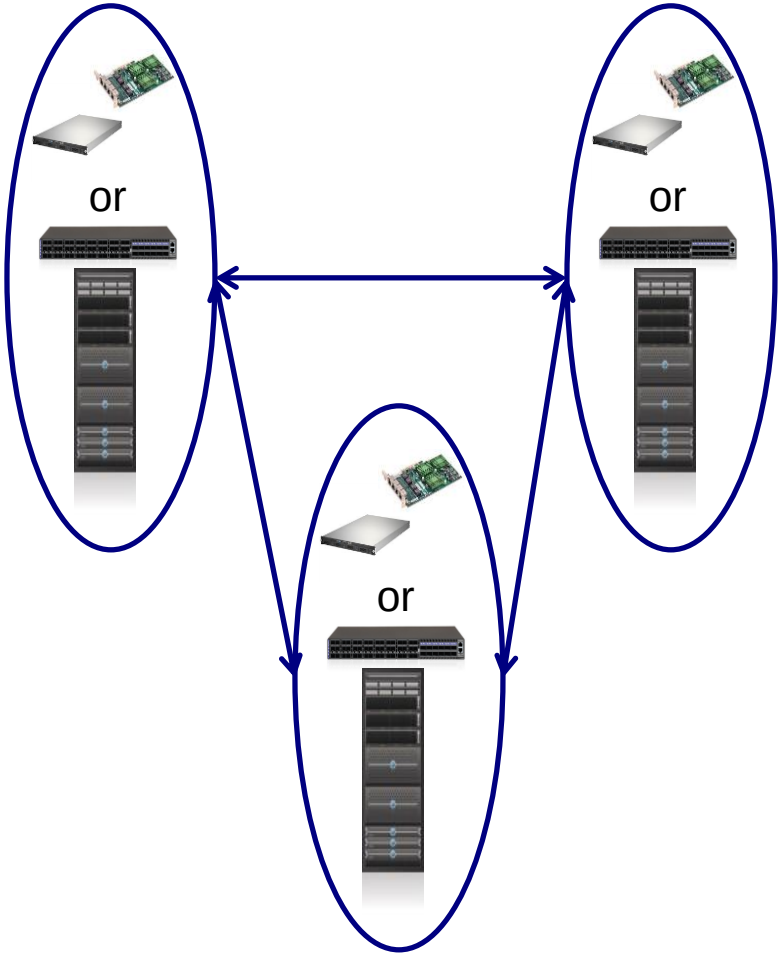


- Data centers use indirect network (e.g., Ethernet), which is good enough for request-level parallelism with limited inter-thread communication.
- Supercomputer use Infiniband and other proprietary networks which enable possible network topologies to facilitate quick communication and synchronization among threads on various servers.

Indirect and direct networks

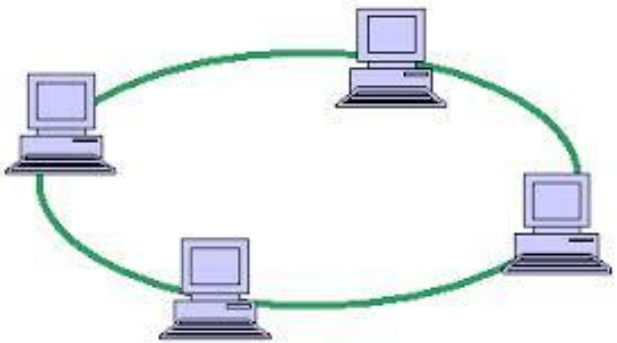


Indirect Network
Network as a “box” with no servers



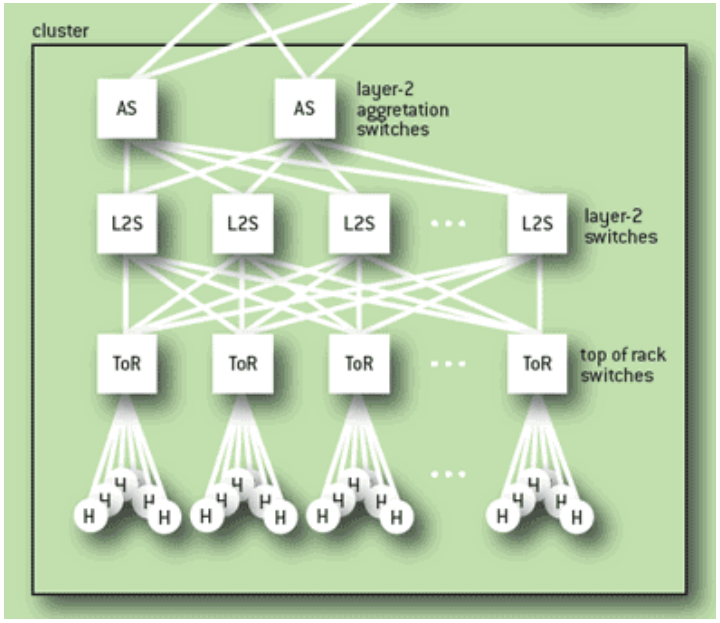
Direct Network
Node: computing + switch

Network metrics



- **Network size**: number of nodes
- **Node degree**: number of ports for each switch (switch complexity)
- **Network diameter**: Longest shortest path between any two nodes in the network
- **Network bandwidth**: best-case total bandwidth
- **Bisection bandwidth**: worst-case total bandwidth when half of nodes is communicating with other half

Datacenter network



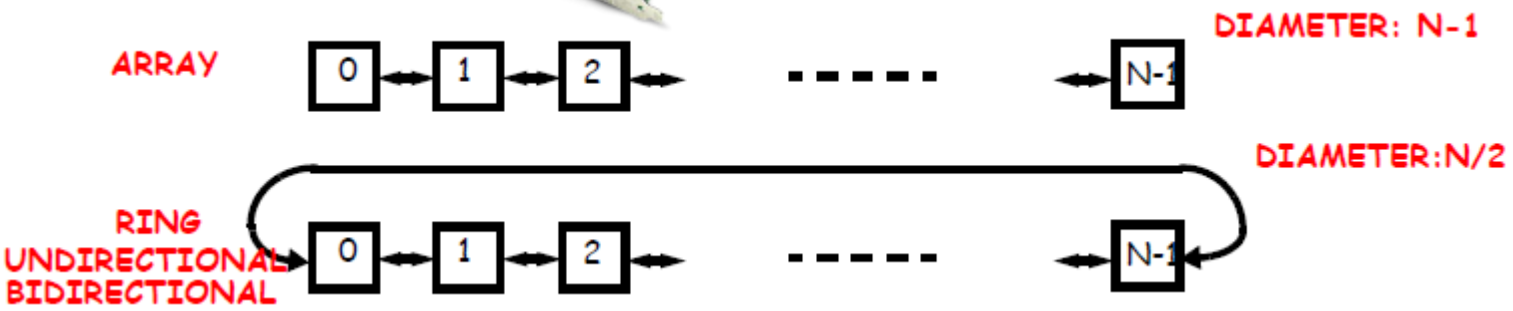
network card in each server

Rack switch example:
Mellanox SX 1024
48 ports of 10GbE
4 uplink ports of 40/56GbE.
BW: 704 Gb/s, 1.04 Bpackets/s
Latency: 250 ns

Aggregate switch example:
Cisco Nexus 7700

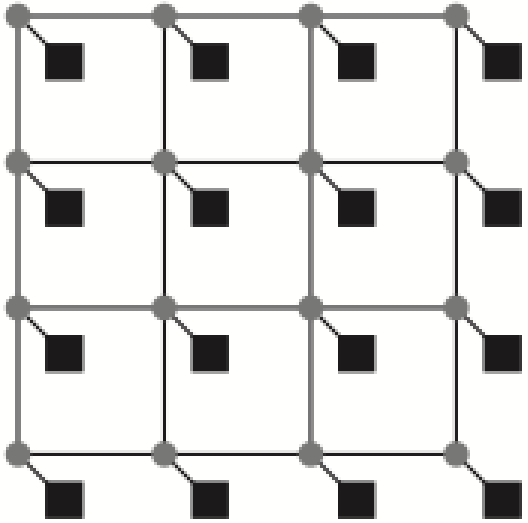
768 x 10 Gbps,
384 x 40 Gbps
192 x 100 Gbps
BW 83 Tbps

Direct networks: linear arrays and rings

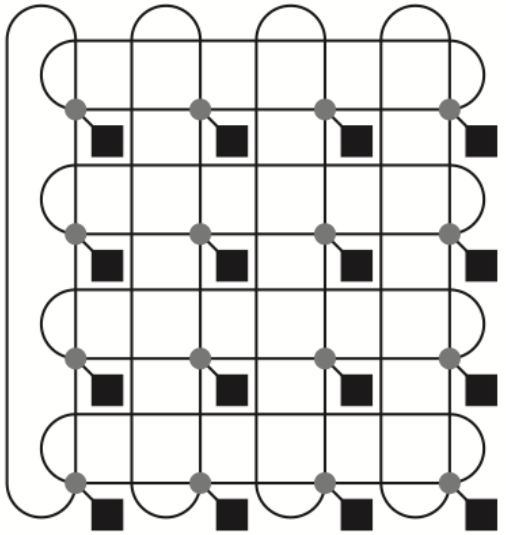


- Great aggregate BW → can exchange N-1 messages at a time
- Bisection BW = 1 for array and 2 for ring
- Layout of ring can be done to get short links

Direct networks: meshes and tori



mesh



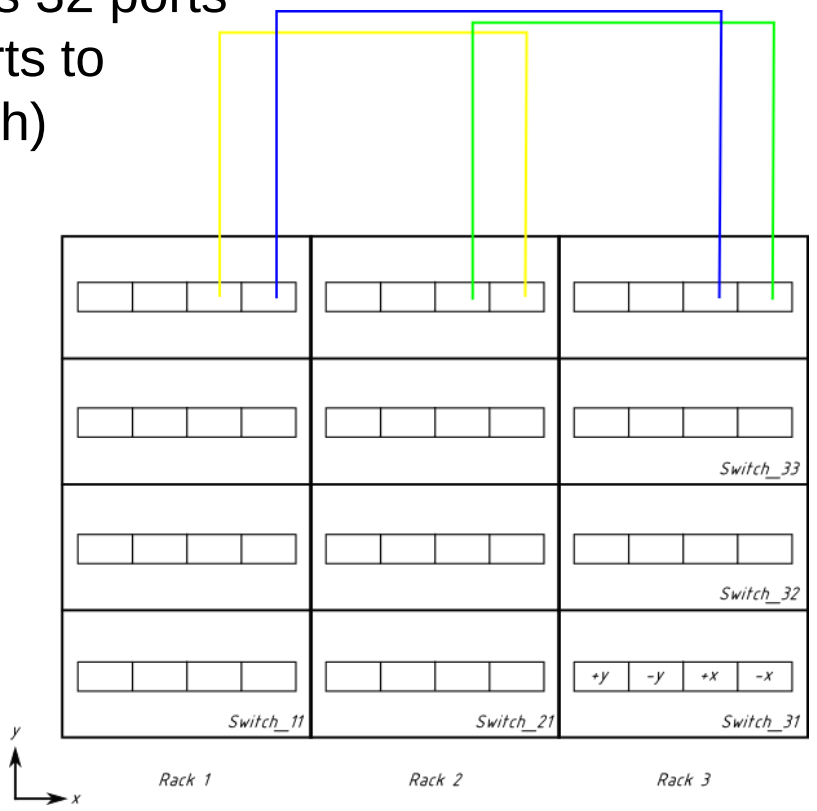
2D torus

- Improves bisection BW and diameter
- Increases number of links; requires more ports per switch (i.e., degree increases)
- What is the diameter?

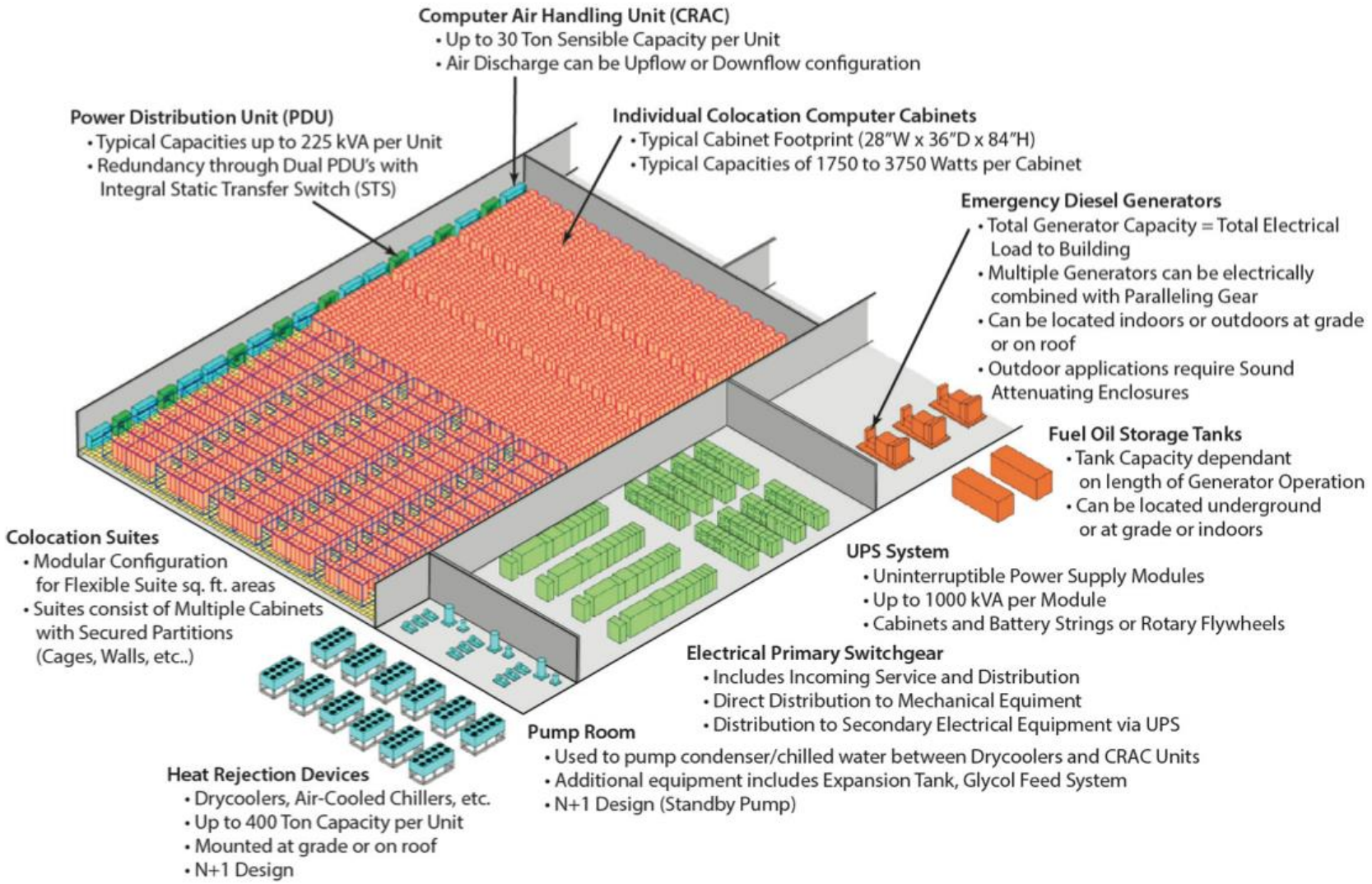
Designing a torus topology for HPC cluster

[example from clusterdesign.org]

- Cluster of 192 blades. Each server is 10U chassis with 16 blades and a switch with 32 ports
- Racks: 12 10U chassis → 4 chassis per rack and a total of 3 racks
- 2D torus: 4x3 torus. Each switch has 32 ports → 16 ports to the blades and 16 ports to each of the neighbors (4 cables each)



Support equipment



Summary

Supercomputers and data centers issues:

- Performance evaluation
- Server makeup
- Storage hierarchy
- Network topologies
- Support infrastructure for cooling and power delivery
- Power consumption concerns