Data Center Network Topologies: A Guided Tour through Data Center Networking

Hakim Weatherspoon

Assistant Professor, Dept of Computer Science

CS 5413: High Performance Systems and Networking

September 17, 2014

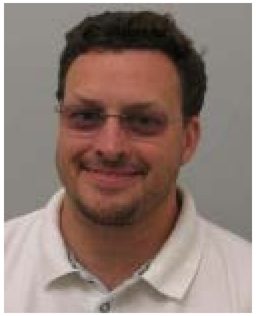
Goals for Today

- A Guided Tour Through Datacenter Networking
  - D. Abts and B. Felderman. Communications of the ACM (CACM), Volume 55, Issue 6 (June 2012), pages 44-51.

Authors

- Bob Felderman
  - Princeton and UCLA
  - Currently a Principle Engineer at Google
  - Founded Myricom
    - Myricom pioneered some kernel bypass approaches
    - Used in cluster computing due to low latency and high performance
  - Also, founded Precision IO

- Dennis Abts
  - PhD from U. of Minnesota
  - Currently a member of Technical Staff at Google
    - System architecture and next-gen large scale clusters
    - research interests include scalable coherence protocols, memory consistency models, interconnection networks, fault tolerant computing and robust system design
  - Sr. Principal Engineer and System Architect for Cray Inc



Goals for Today

- A Guided Tour Through Datacenter Networking
  - D. Abts and B. Felderman. Communications of the ACM (CACM), Volume 55, Issue 6 (June 2012), pages 44-51.
- Background: Principles and central ideas of data center networks
- Data Center Traffic
- Data Center Network Architecture
- Network Performance
  - Flow Control
  - Network Stack
- Scalable, Manageable, and Flexible
- Reliable and Available

Background

- High Performance Computing (HPC)
  - Expensive and highly tuned
  - High bandwidth
  - Low latency
  - Fine-grained
  - E.g. HPC Application like scientific computing and financial enterprise systems

Background

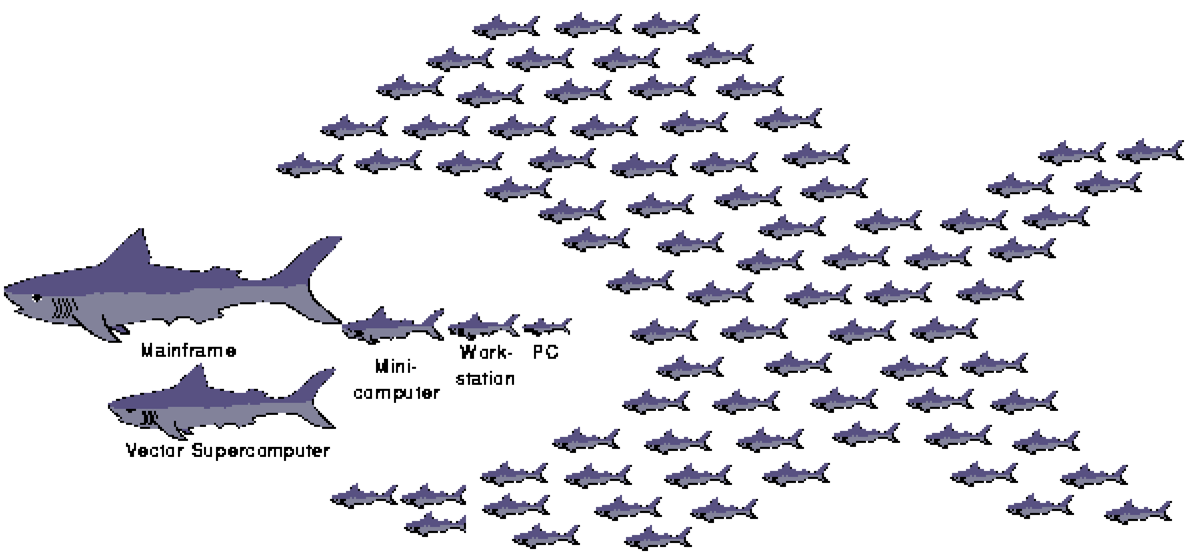
- Ethernet networks
  - Cheap and general (COTS; commodity off the shelf)
  - Increasing bandwidth (1GbE, 10GbE, 40GbE, 100GbE)
  - E.g. 42% of Top500 use Ethernet in 2012 (2% in 2002)
  - E.g. Web and cloud applications

Background

- Modern Data Center
  - 10s to 100s of thousands of hosts
  - Each host many processing cores, memory, network interface, and local storage (HDD and/or SDD)
- Clusters
  - 10s of racks with 10s of servers in each rack
  - Homogeneous
  - Individual request may contact many clusters
    - Performance based on slowest response
    - Performance of Remote memory vs local disk
    - Network and resp variance congestion can reduce performance
    - Overprovisioning may be too expensive
    - QoS (quality of service): Implemented via NIC with flow classification and priorities

Background

- Modern Data Center
  - 10s to 100s of thousands of hosts
  - Each host many processing cores, memory, network interface, and local storage (HDD and/or SDD)
- Clusters
  - 10s of racks with 10s of servers in each rack




cooling
towers
warehouse-scale
computer
power substation

Goals for Today

- A Guided Tour Through Datacenter Networking
  - D. Abts and B. Felderman. Communications of the ACM (CACM), Volume 55, Issue 6 (June 2012), pages 44-51.
- Background: Principles and central ideas of data center networks
- Data Center Traffic
- Data Center Network Architecture
- Network Performance
  - Flow Control
  - Network Stack
- Scalable, Manageable, and Flexible
- Reliable and Available

Data Center Traffic

Bimodal: Elephant and Mice

- Average the same by variance is significant
- Mice
  - Short lived
  - Most flows
- Elephant
  - Long lived and bursty
  - Less than 1% of flows
  - Performance impact is significant
    - Lead to temporary congestion on a shared bottleneck link
    - Oversubscription: Hierarchical datacenter topology
    - Inter-rack communication less orchestrated than intra-rack

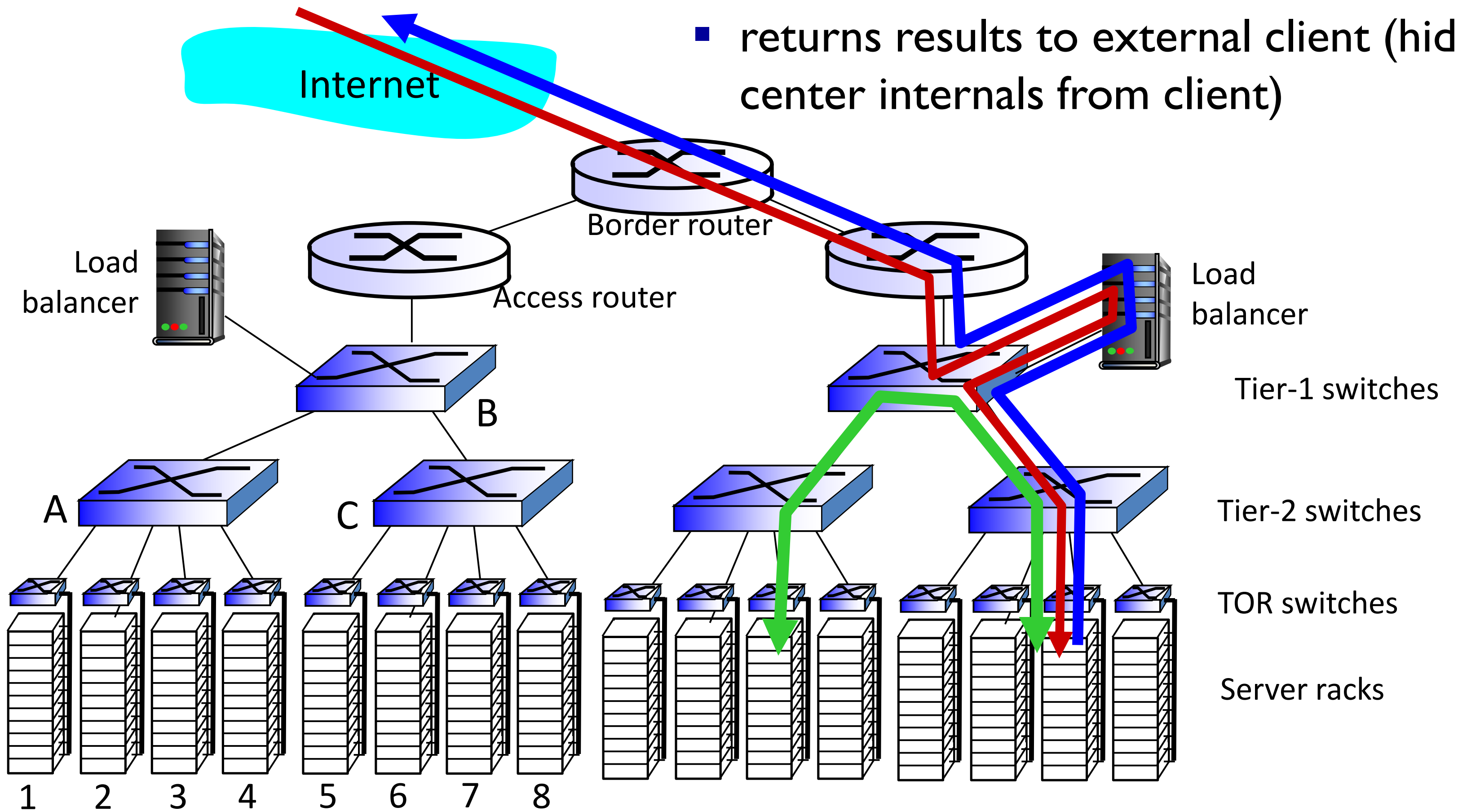
Data Center Network Architecture

Inside a 40-ft Microsoft container, Chicago data center

Data Center Network Architecture

load balancer: application-layer routing

- receives external client requests
- directs workload within data center
- returns results to external client (hiding data center internals from client)



Data Center Network Architecture

- How to identify hosts
  - Endpoint identifiers (Local Area IP address)
  - Statically assigned identifiers or DHCP

- Limitations of Layer 2 and 3 routing
  - ARP (broadcasts)
    - Switches participate in spanning tree protocols (STP) or transparent interconnect of lots of links (TRILL)
  - 64k entries: limitation of packet-forwarding tables

Data Center Network Architecture

Limitations

- *Topology:*
  - 2 layers: 5K to 8K hosts
  - 3 layer: >25K hosts
  - Switches:
    - Leaves: have N GigE ports (48-288) + N 10 GigE uplinks to one or more layers of network elements
    - Higher levels: N 10 GigE ports (32-128)
- *Multi-path Routing:*
  - Ex. ECMP
    - without it, the largest cluster = 1,280 nodes
    - Performs static load splitting among flows
    - Lead to oversubscription for simple comm. patterns
    - Routing table entries grows multiplicatively with number of paths, cost ++, lookup latency ++

Data Center Network Architecture

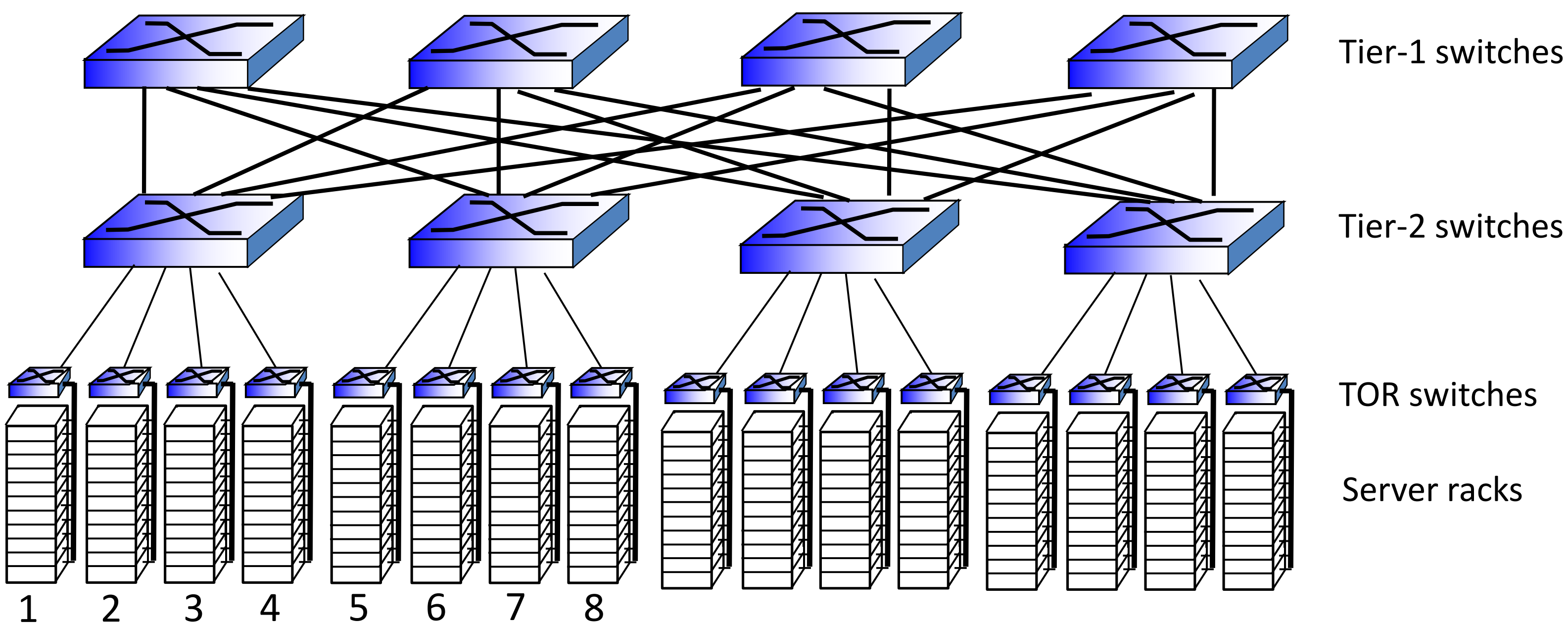
Issues with Traditional Data Center Topology

- *Oversubscription:*
  - Ratio of the worst-case achievable aggregate bandwidth among the end hosts to the total bisection bandwidth of a particular communication topology
  - Lower the total cost of the design
  - Typical designs: factor of 2:5:1 (400 Mbps)to 8:1(125 Mbps)
- *Cost:*
  - Edge: $7,000 for each 48-port GigE switch
  - Aggregation and core: $700,000 for 128-port 10GigE switches
  - Cabling costs are not considered!

Data Center Network Architecture

FatTree overcomes limitations

- rich interconnection among switches, racks:
  - increased throughput between racks (multiple routing paths possible)
  - increased reliability via redundancy



Goals for Today

- A Guided Tour Through Datacenter Networking
  - D. Abts and B. Felderman. Communications of the ACM (CACM), Volume 55, Issue 6 (June 2012), pages 44-51.
- Background: Principles and central ideas of data center networks
- Data Center Traffic
- Data Center Network Architecture
- Network Performance
  - Flow Control
  - Network Stack
- Scalable, Manageable, and Flexible
- Reliable and Available

Network Performance

- Flow control
  - L1: Propagation delay
  - L2/3: Buffering
    - Stable vs unstable networks
  - L4: end-to-end flow control—TCP

- End-host Network Stack performance
  - Kernel (OS) bypass
  - Zero-copy
  - Limitations: Interrupt Coalescing

  - What about virtualization?
    - Multi-queue NICs

Goals for Today

- A Guided Tour Through Datacenter Networking
  - D. Abts and B. Felderman. Communications of the ACM (CACM), Volume 55, Issue 6 (June 2012), pages 44-51.
- Background: Principles and central ideas of data center networks
- Data Center Traffic
- Data Center Network Architecture
- Network Performance
  - Flow Control
  - Network Stack
- Scalable, Manageable, and Flexible
- Reliable and Available

Perspective

- Data Center Networks have unique requirements

- However, network stack remains intact, but innovation at individual layers: (L1 – optical, L2/L3 – topologies, L4 – TCP (DCTCP), L5 – sockets)

Before Next time

- Project Proposal
  - **due this Friday, Sept 19**
  - Meet with groups, TA, and professor
- Lab2
  - Multi threaded TCP proxy
  - **CHANGE: Due this Friday, Sept 22**

- ***Required review and reading***
  - "A Scalable, Commodity Data Center Network Architecture," M. Al-Fares, A. Loukissas, A. Vahdat . *ACM SIGCOMM Computer Communication Review*, Volume 38, Issue 4 (October 2008), pages 63-74.
  - http://dl.acm.org/citation.cfm?id=1402967

- Check piazza: http://piazza.com/cornell/fall2014/cs5413
- Check website for updated schedule