15-744: Computer Networking

L-24 Data Center Networking

Overview

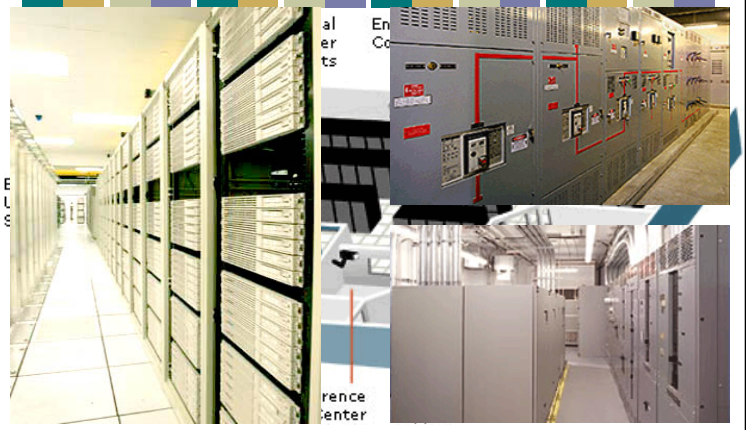
- Data Center Overview
- Routing in the DC
- Transport in the DC

Datacenter Arms Race

- Amazon, Google, Microsoft, Yahoo!, … race to build next-gen mega-datacenters
  - Industrial-scale Information Technology
  - 100,000+ servers
  - Located where land, water, fiber-optic connectivity, and cheap power are available
- E.g., Microsoft Quincy
  - 43600 sq. ft. (10 football fields), sized for 48 MW
  - Also Chicago, San Antonio, Dublin @$500M each
- E.g., Google:
  - The Dalles OR, Pryor OK, Council Bluffs, IW, Lenoir NC, Goose Creek , SC

Google Oregon Datacenter

Computers + Net + Storage + *Power* + *Cooling*



5

Energy Proportional Computing

"The Case for Energy-Proportional Computing," Luiz André Barroso, Urs Hölzle, *IEEE Computer* December 2007

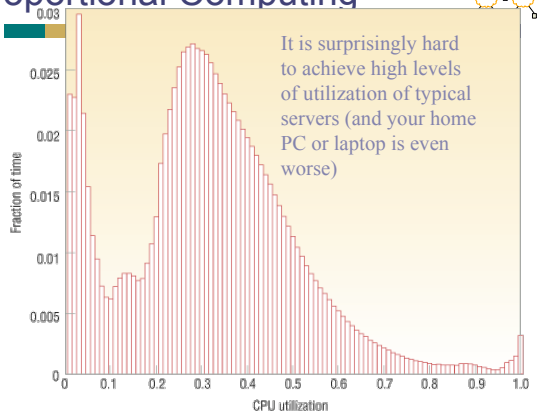


Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum

6

Energy Proportional Computing

"The Case for Energy-Proportional Computing," Luiz André Barroso, Urs Hölzle, *IEEE Computer* December 2007

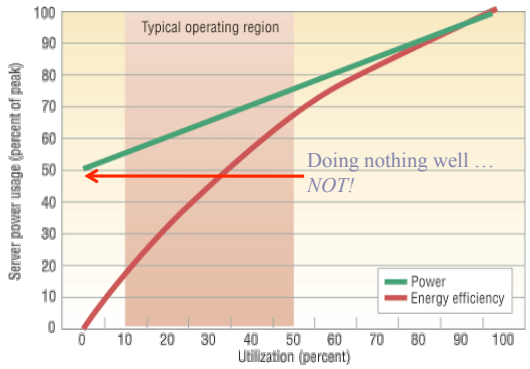


Figure 2. Server power usage and energy efficiency at varying utilization levels, from idle to peak performance. Even an energy-efficient server still consumes about half its full power when doing virtually no work.

7

Energy Proportional Computing

"The Case for Energy-Proportional Computing," Luiz André Barroso, Urs Hölzle, *IEEE Computer* December 2007

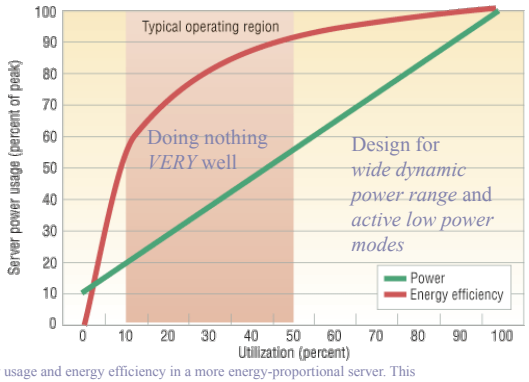
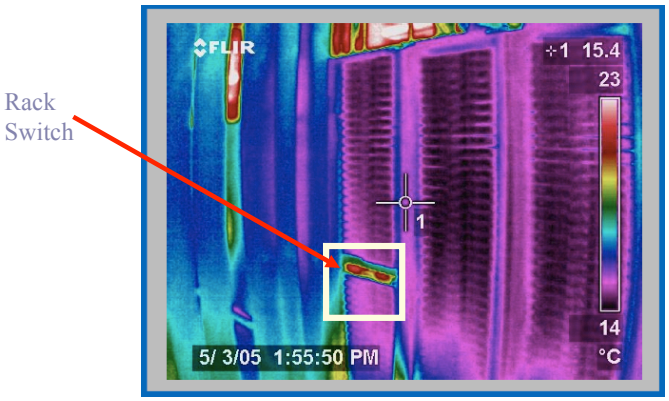


Figure 4. Power usage and energy efficiency in a more energy-proportional server. This server has a power efficiency of more than 80 percent of its peak value for utilizations of 30 percent and above, with efficiency remaining above 50 percent for utilization levels as low as 10 percent.

8

Thermal Image of Typical Cluster



M. K. Patterson, A. Pratt, P. Kumar,
"From UPS to Silicon: an end-to-end evaluation of datacenter efficiency", Intel Corporation

9

DC Networking and Power

- Within DC racks, network equipment often the "hottest" components in the hot spot
- Network opportunities for power reduction
  - Transition to higher speed interconnects (10 Gbs) at DC scales and densities
  - High function/high power assists embedded in network element (e.g., TCAMs)

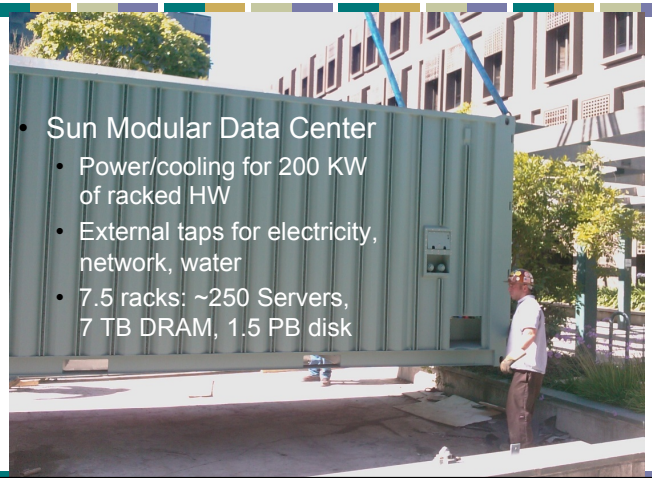
10

DC Networking and Power

- 96 x 1 Gbit port Cisco datacenter switch consumes around 15 kW -- approximately 100x a typical dual processor Google server @ 145 W
- High port density drives network element design, but such high power density makes it difficult to tightly pack them with servers
- Alternative distributed processing/communications topology under investigation by various research groups

11

Containerized Datacenters



- Sun Modular Data Center
  - Power/cooling for 200 KW of racked HW
  - External taps for electricity, network, water
  - 7.5 racks: ~250 Servers, 7 TB DRAM, 1.5 PB disk

13

Containerized Datacenters

Summary

- Energy Consumption in IT Equipment
  - Energy Proportional Computing
  - Inherent inefficiencies in electrical energy distribution
- Energy Consumption in Internet Datacenters
  - Backend to billions of network capable devices
  - Enormous processing, storage, and bandwidth supporting applications for huge user communities
  - Resource Management: Processor, Memory, I/O, Network to maximize performance subject to power constraints: “Do Nothing Well”
  - New packaging opportunities for better optimization of computing + communicating + power + mechanical

15

Overview

- Data Center Overview
- Routing in the DC
- Transport in the DC

16

Layer 2 vs. Layer 3 for Data Centers

Technique	Plug and play	Scalability	Small Switch State	Seamless VM Migration
Layer 2: Flat MAC Addresses	+	-	-	+
Layer 3: IP Addresses	-	+	+	-

17

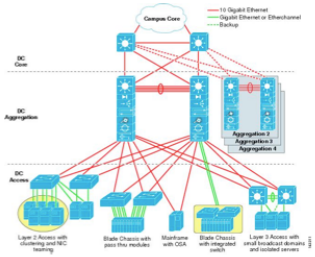
Flat vs. Location Based Addresses

- Commodity switches today have ~640 KB of low latency, power hungry, expensive on chip memory
  - Stores 32 – 64 K flow entries
- Assume 10 million virtual endpoints in 500,000 servers in datacenter
- Flat addresses ➔ 10 million address mappings ➔ ~100 MB on chip memory ➔ ~150 times the memory size that can be put on chip today
- Location based addresses ➔ 100 – 1000 address mappings ➔ ~10 KB of memory ➔ easily accommodated in switches today

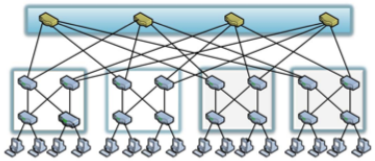
18

PortLand: Main Assumption

- Hierarchical structure of data center networks:
  - They are multi-level, multi-rooted trees



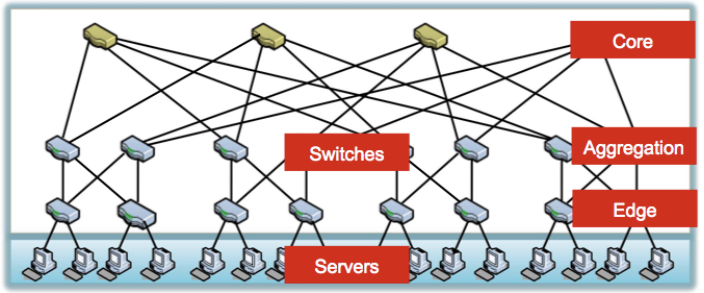
Cisco Recommended Configuration



Fat Tree

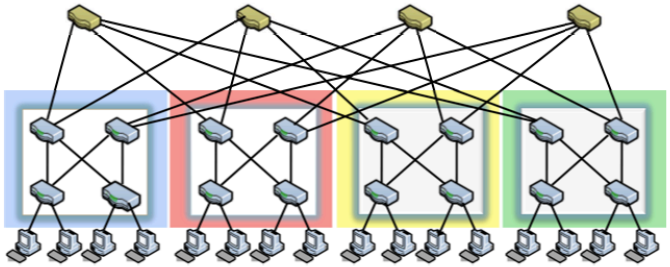
19

Data Center Network



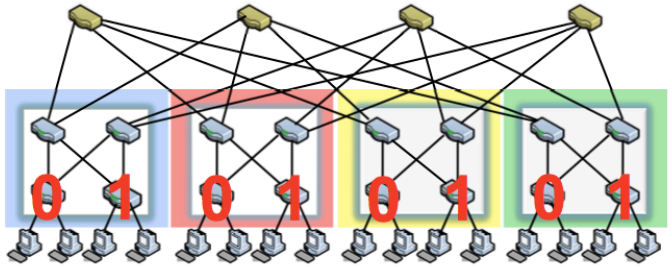
20

Hierarchical Addresses



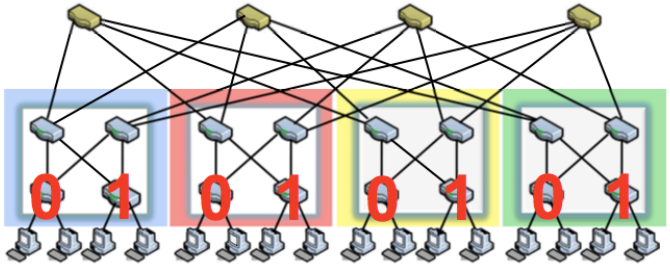
21

Hierarchical Addresses



22

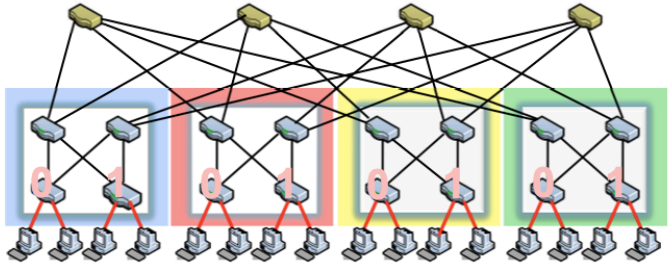
Hierarchical Addresses



PMAC: pod.position.port.vmid

23

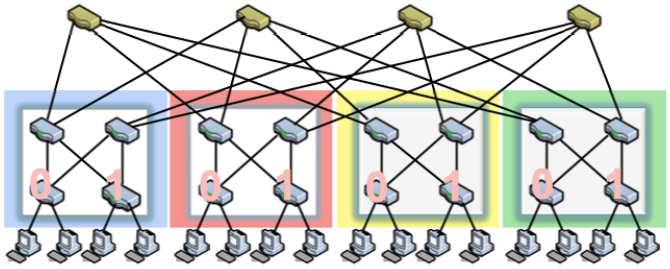
Hierarchical Addresses



PMAC: pod.position.port.vmid

24

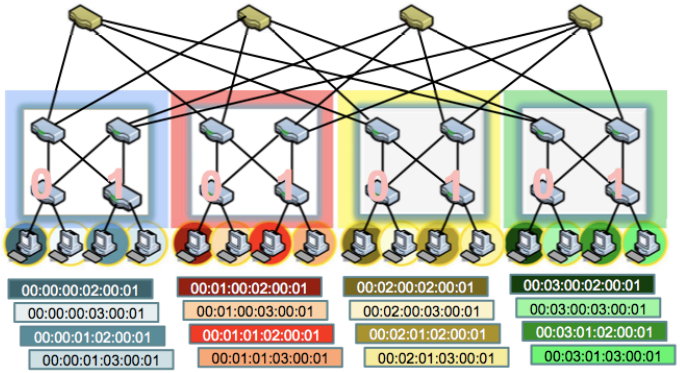
Hierarchical Addresses



PMAC: pod.position.port.vmid

25

Hierarchical Addresses



26

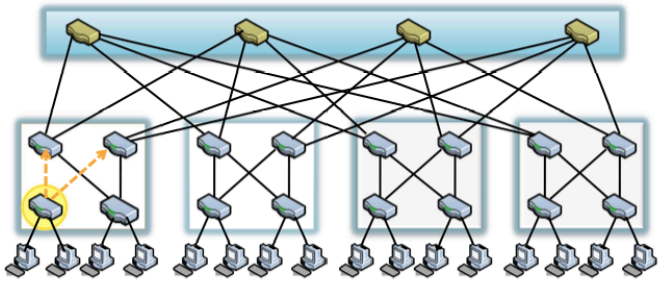
PortLand: Location Discovery Protocol

- Location Discovery Messages (LDMs) exchanged between neighboring switches
- Switches self-discover location on boot up

Location characteristic	Technique
1) Tree level / Role	Based on neighbor identity
2) Pod number	Aggregation and edge switches agree on pod number
3) Position number	Aggregation switches help edge switches choose unique position number

27

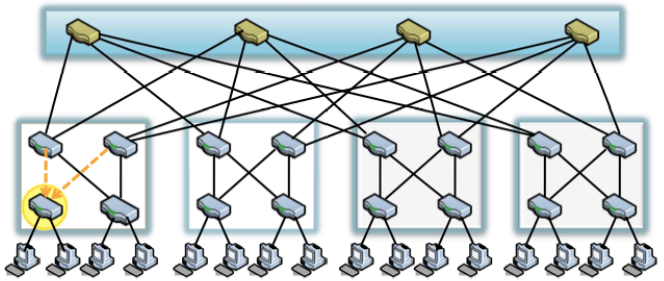
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
A0:B1:FD:56:32:01	??	??	??

28

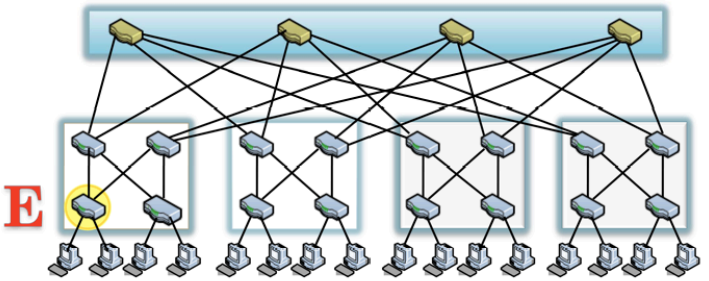
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
A0:B1:FD:56:32:01	??	??	??

29

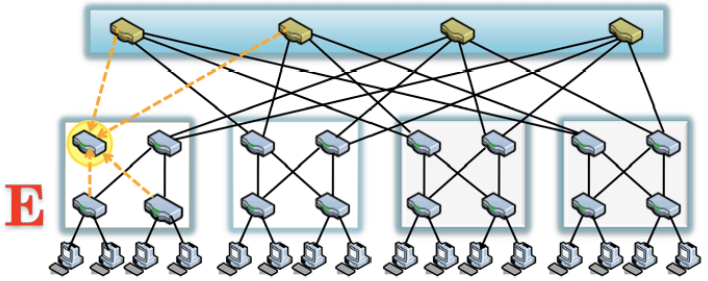
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
A0:B1:FD:56:32:01	??	??	0

30

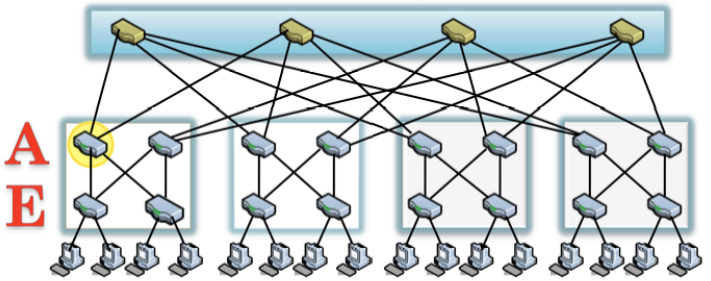
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
B0:A1:FD:57:32:01	??	??	??

31

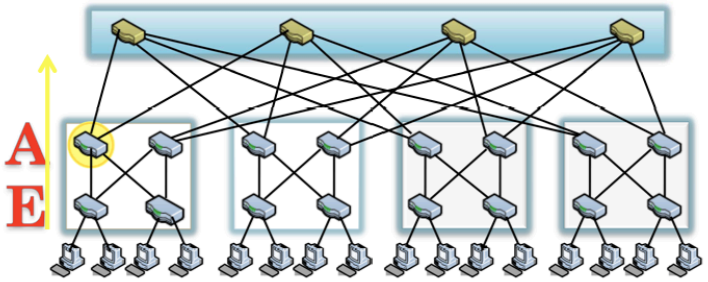
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
B0:A1:FD:57:32:01	??	??	1

32

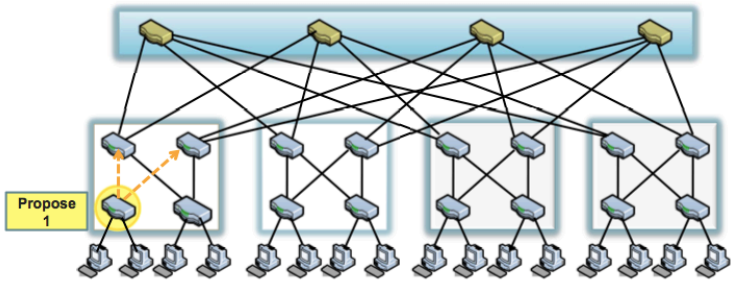
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
B0:A1:FD:57:32:01	??	??	1

33

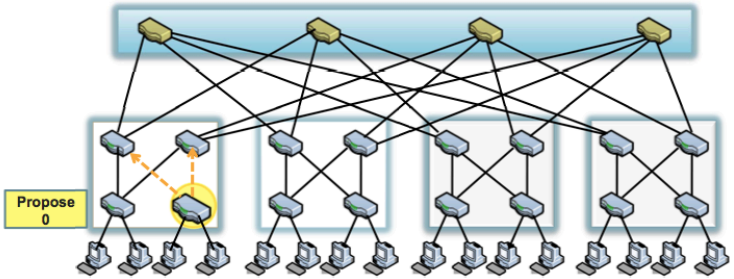
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
A0:B1:FD:56:32:01	??	??	0

34

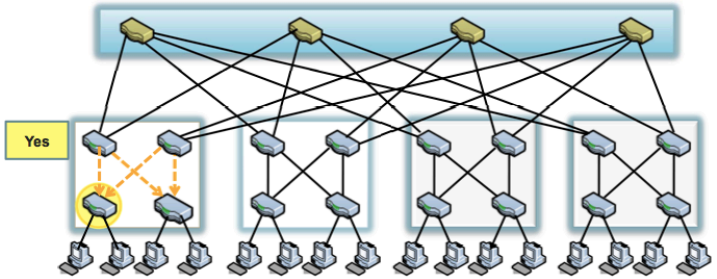
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
D0:B1:AD:56:32:01	??	??	0

35

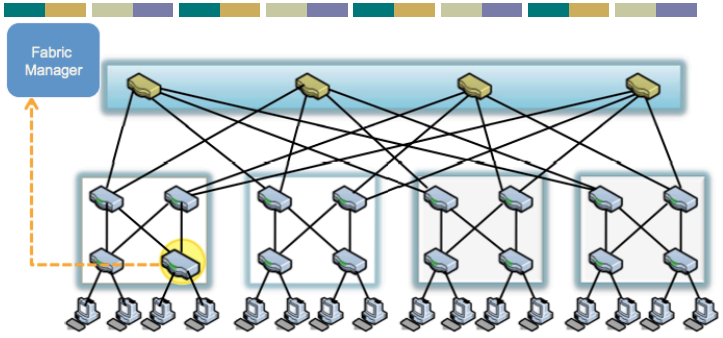
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
A0:B1:FD:56:32:01	??	1	0

36

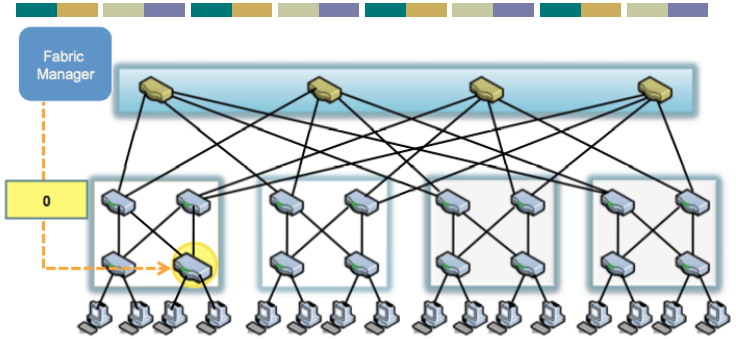
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
D0:B1:AD:56:32:01	??	0	0

37

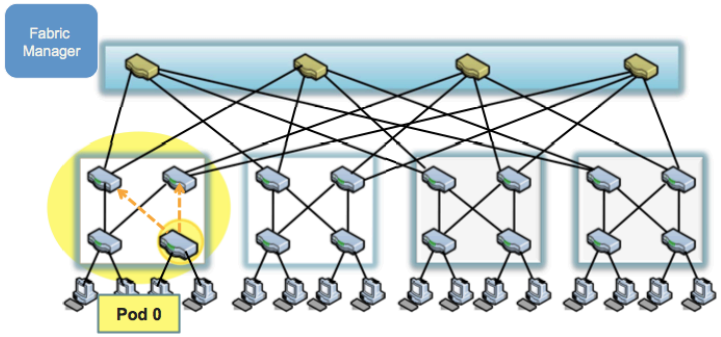
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
D0:B1:AD:56:32:01	??	0	0

38

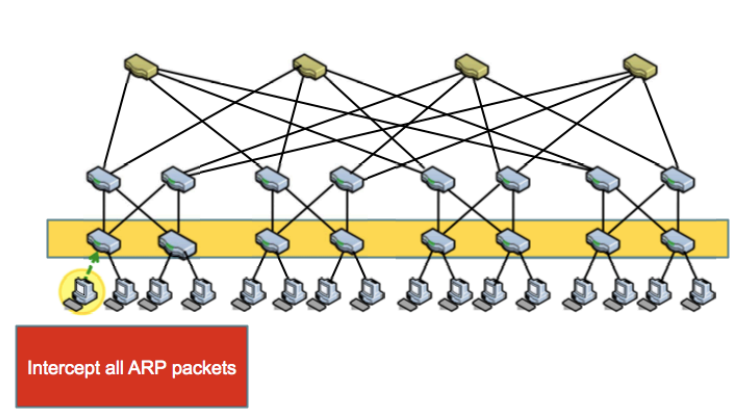
Location Discovery Protocol



Switch Identifier	Pod Number	Position	Tree Level
D0:B1:AD:56:32:01	0	0	0

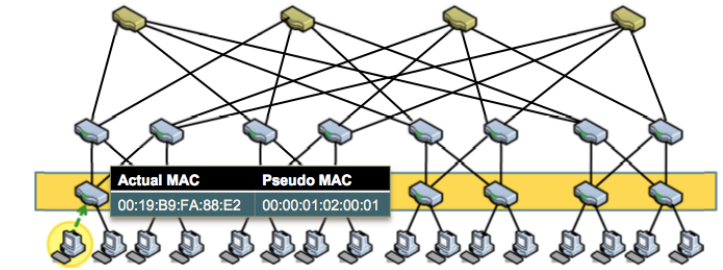
39

Name Resolution



40

Name Resolution

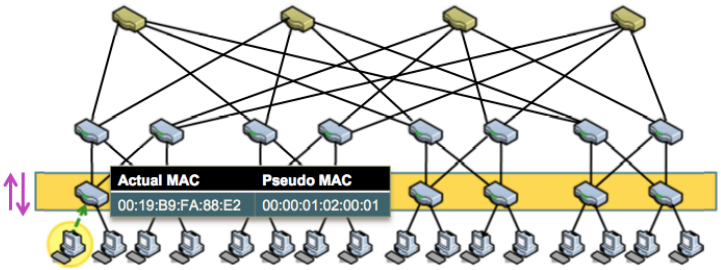


Intercept all ARP packets

Assign new end hosts with PMACs

41

Name Resolution



Intercept all ARP packets

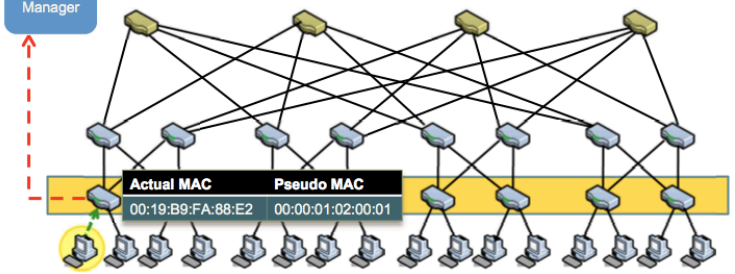
Assign new end hosts with PMACs

Rewrite MAC for packets entering and exiting network

42

Name Resolution

Fabric Manager

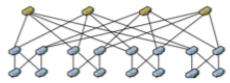


43

Fabric Manager

Fabric Manager

IP	Pseudo MAC
10.5.1.2	00:00:01:02:00:01
10.2.4.5	00:02:00:02:00:01



ARP mappings

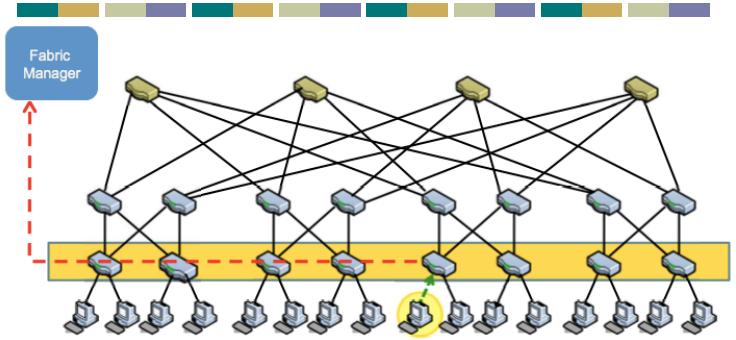
Network map

Soft state

Administrator configuration

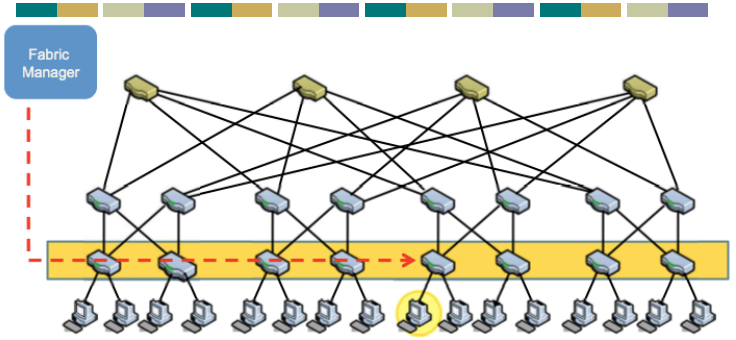
44

Name Resolution



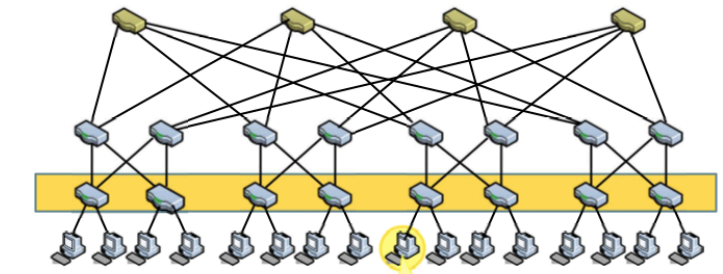
45

Name Resolution



46

Name Resolution



47

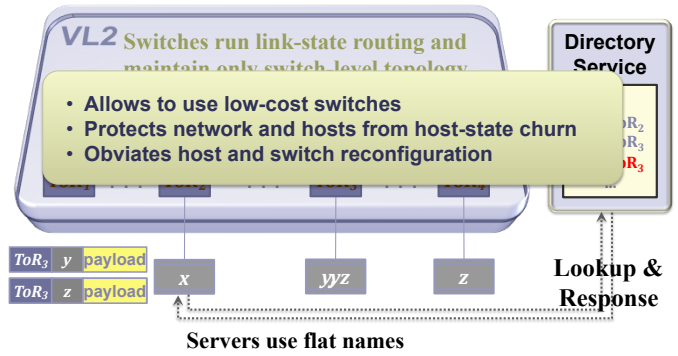
Other Schemes

- SEATTLE [SIGCOMM '08]:
  - Layer 2 network fabric that works at enterprise scale
  - Eliminates ARP broadcast, proposes one-hop DHT
  - Eliminates flooding, uses broadcast based LSR
  - Scalability limited by
    - Broadcast based routing protocol
    - Large switch state
- VL2 [SIGCOMM '09]
  - Network architecture that scales to support huge data centers
  - Layer 3 routing fabric used to implement a virtual layer 2
  - Scale Layer 2 via end host modifications
    - Unmodified switch hardware and software
    - End hosts modified to perform enhanced resolution to assist routing and forwarding

48

VL2: Name-Location Separation

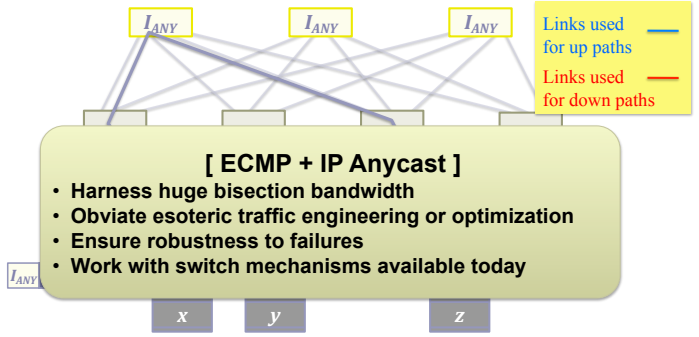
Cope with host churns with very little overhead



49

VL2: Random Indirection

Cope with arbitrary TMs with very little overhead



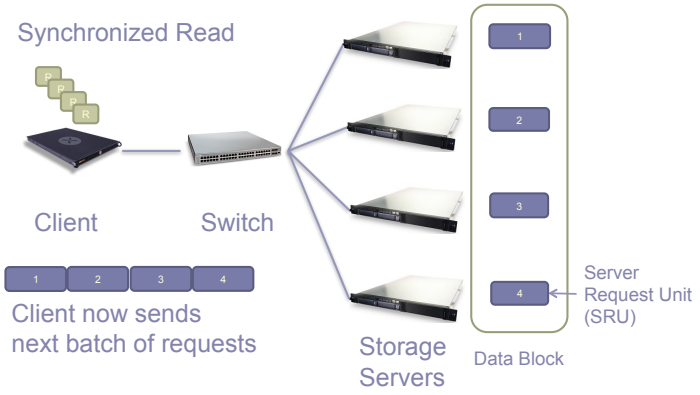
50

Overview

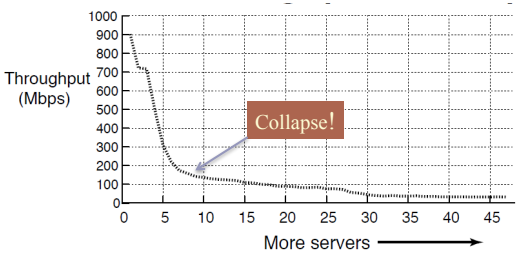
- Data Center Overview
- Routing in the DC
- Transport in the DC

51

Cluster-based Storage Systems



TCP Throughput Collapse

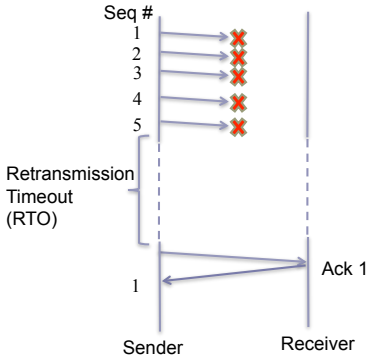


Cluster Setup
1Gbps Ethernet
Unmodified TCP
S50 Switch
1MB Block Size

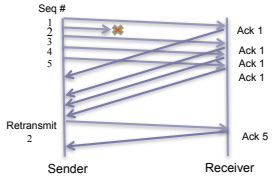
- TCP *Incast*
  - Cause of throughput collapse: coarse-grained TCP timeouts

TCP: Loss recovery comparison

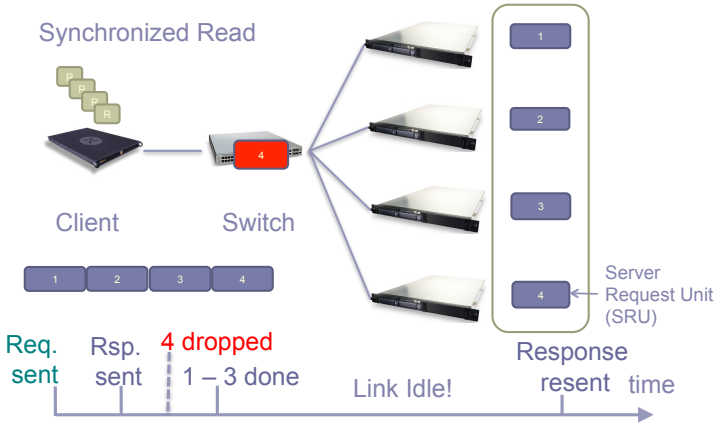
Timeout driven recovery is slow (**ms**)



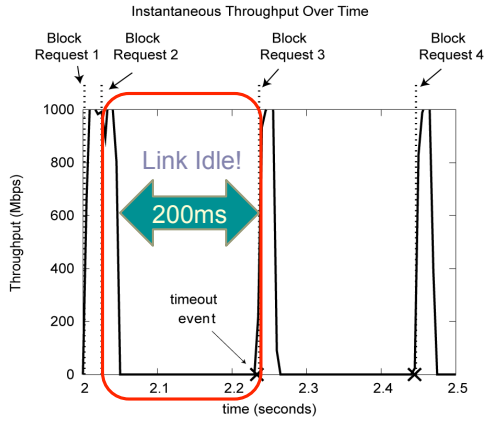
Data-driven recovery is super fast (**µs**) in datacenters



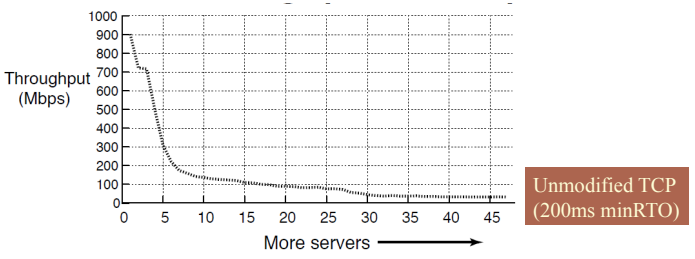
Link Idle Time Due To Timeouts



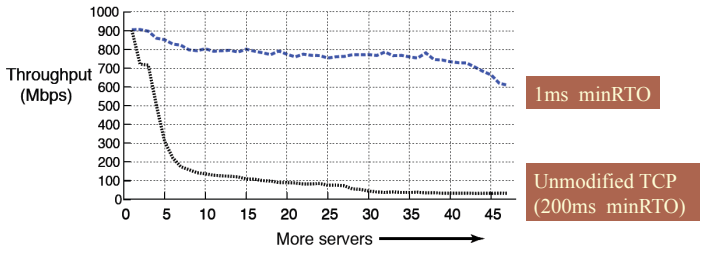
Client Link Utilization



Default minRTO: Throughput Collapse

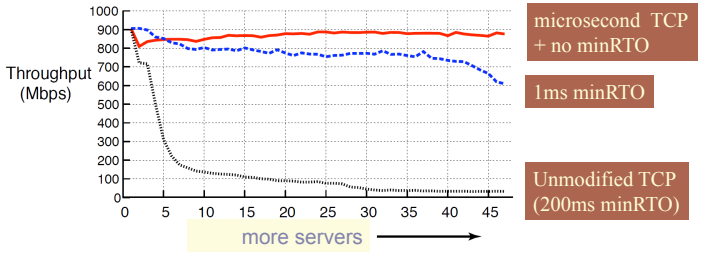


Lowering minRTO to 1ms helps



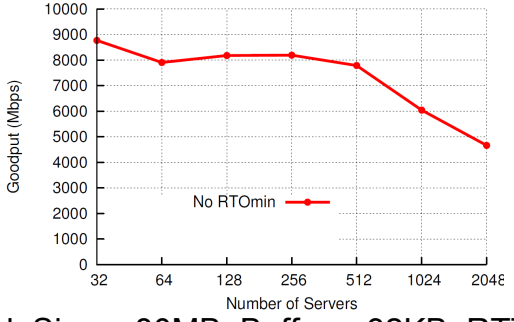
Millisecond retransmissions are not enough

Solution: **µsecond TCP + no minRTO**



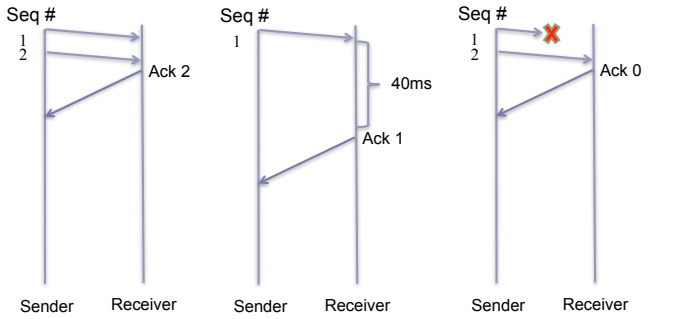
✓ High throughput for up to 47 servers

Simulation: Scaling to thousands



Block Size = 80MB, Buffer = 32KB, RTT = 20us

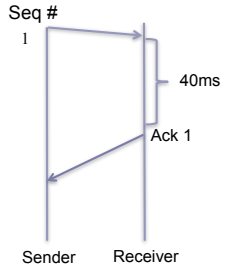
Delayed-ACK (for RTO > 40ms)



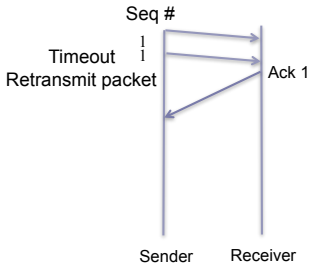
Delayed-Ack: Optimization to reduce #ACKs sent

µsecond RTO and Delayed-ACK

RTO > 40ms



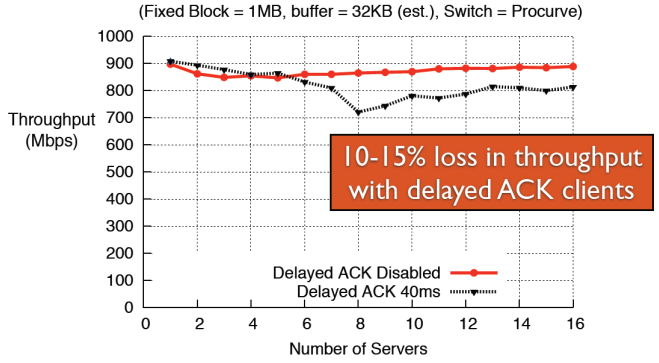
RTO < 40ms



Premature Timeout

RTO on sender triggers before Delayed-ACK on receiver

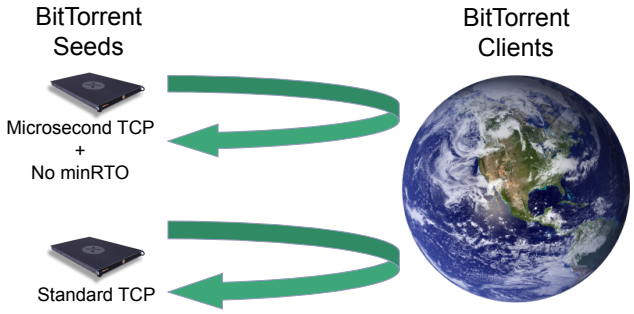
Impact of Delayed-ACK



Is it safe for the wide-area?

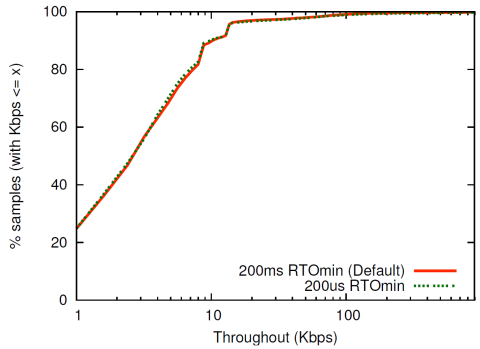
- Stability: Could we cause congestion collapse?
  - No: Wide-area RTOs are in 10s, 100s of ms
  - No: Timeouts result in rediscovering link capacity (slow down the rate of transfer)
- Performance: Do we timeout unnecessarily?
  - [Allman99] Reducing minRTO increases the chance of premature timeouts
    - Premature timeouts slow transfer rate
  - Today: detect and recover from premature timeouts
  - Wide-area experiments to determine performance impact

Wide-area Experiment



- Do microsecond timeouts harm wide-area throughput?

Wide-area Experiment: Results



No noticeable difference in throughput

Other Efforts

- Topology
  - Using extra links to meet traffic matrix
    - 60Ghz links → MSR paper in HotNets09
    - Reconfigurable optical interconnects → CMU and UCSD in Sigcomm2010
- Transport
  - Data Center TCP → data-center only protocol that uses RED-like techniques in routers

67

Next Lecture

- Topology
- Required reading
  - On Power-Law Relationships of the Internet Topology
  - A First-Principles Approach to Understanding the Internet's Router-level Topology
- Optional reading
  - Measuring ISP Topologies with Rocketfuel

68

Aside: Disk Power

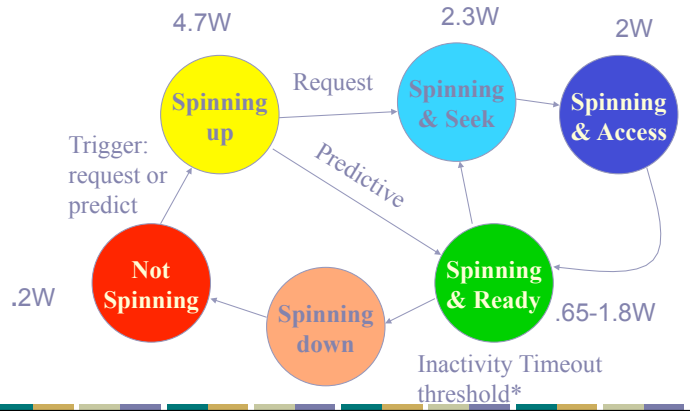
IBM Microdrive (1inch)

- writing 300mA (3.3V) 1W
- standby 65mA (3.3V) .2W

IBM TravelStar (2.5inch)

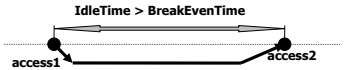
- read/write 2W
- spinning 1.8W
- low power idle .65W
- standby .25W
- sleep .1W
- startup 4.7 W
- seek 2.3W

Spin-down Disk Model

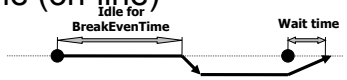


Disk Spindown

- Disk Power Management – Oracle (off-line)



- Disk Power Management – Practical scheme (on-line)



71

Spin-Down Policies

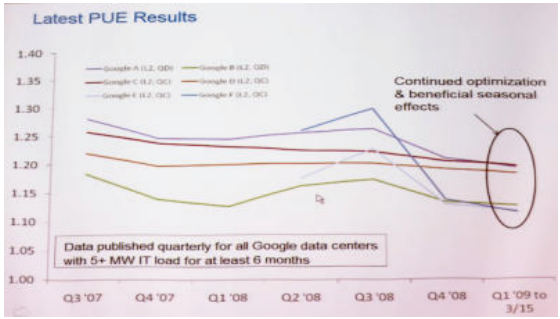
- Fixed Thresholds
  - T_{out} = spin-down cost s.t. $2*E_{transition} = P_{spin}*T_{out}$
- Adaptive Thresholds: T_{out} = f (recent accesses)
  - Exploit burstiness in T_{idle}
- Minimizing Bumps (user annoyance/latency)
  - Predictive spin-ups
- Changing access patterns (making burstiness)
  - Caching
  - Prefetching

Google

- Since 2005, its data centers have been composed of standard shipping containers--each with 1,160 servers and a power consumption that can reach 250 kilowatts
- Google server was 3.5 inches thick--2U, or 2 rack units, in data center parlance. It had two processors, two hard drives, and eight memory slots mounted on a motherboard built by Gigabyte

73

Google's PUE



- In the third quarter of 2008, Google's PUE was 1.21, but it dropped to 1.20 for the fourth quarter and to 1.19 for the first quarter of 2009 through March 15
- Newest facilities have 1.12

74