15-744: Computer Networking

L-24 Data Center Networking

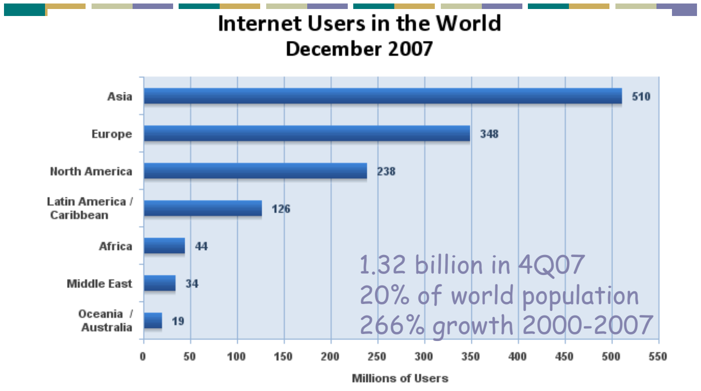
Overview

- Data Center Overview
- Networking in the DC

“The Big Switch,” Redux

“A hundred years ago, companies stopped generating their own power with steam engines and dynamos and plugged into the newly built electric grid. The cheap power pumped out by electric utilities didn’t just change how businesses operate. It set off a chain reaction of economic and social transformations that brought the modern world into existence. Today, a similar revolution is under way. Hooked up to the Internet’s global computing grid, massive information-processing plants have begun pumping data and software code into our homes and businesses. This time, it’s computing that’s turning into a utility.”

Growth of the Internet Continues …



1.32 billion in 4Q07
20% of world population
266% growth 2000-2007

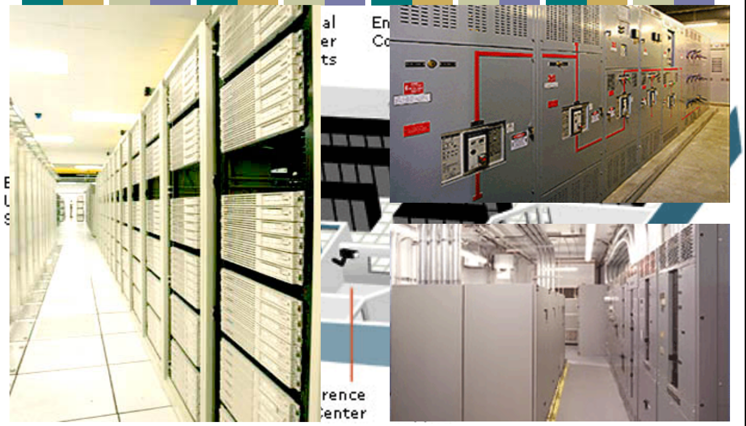
Note: Total World Internet Users estimate is 1,319,872,109 for year-end 2007
Copyright © 2008, Miniwatts Marketing Group - www.internetworldstats.com

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*



Energy Proportional Computing

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

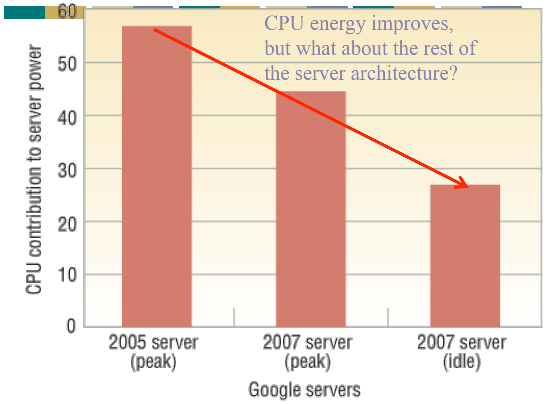
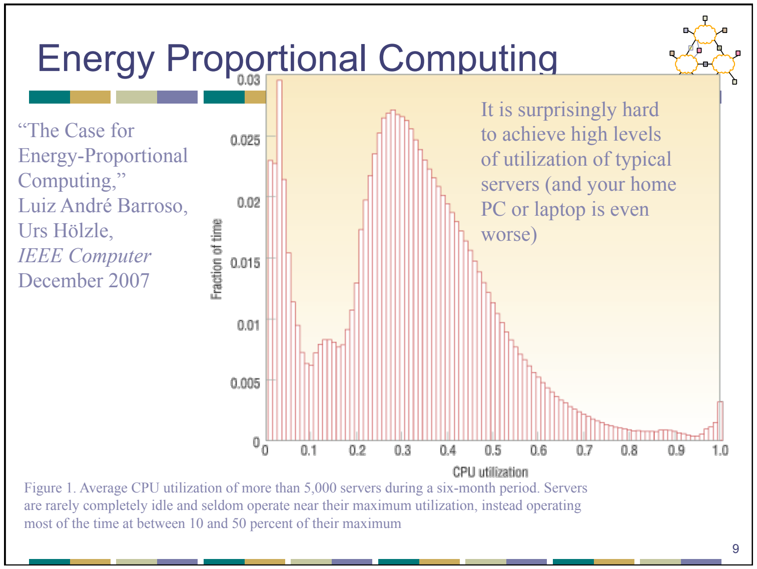


Figure 3. CPU contribution to total server power for two generations of Google servers at peak performance (the first two bars) and for the later generation at idle (the rightmost bar).

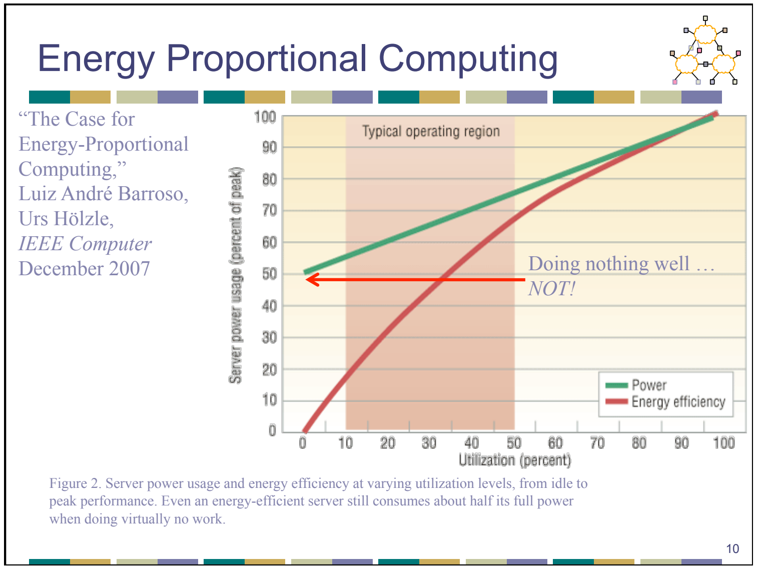


Energy Proportional Computing

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

It is surprisingly hard to achieve high levels of utilization of typical servers (and your home PC or laptop is even worse)

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

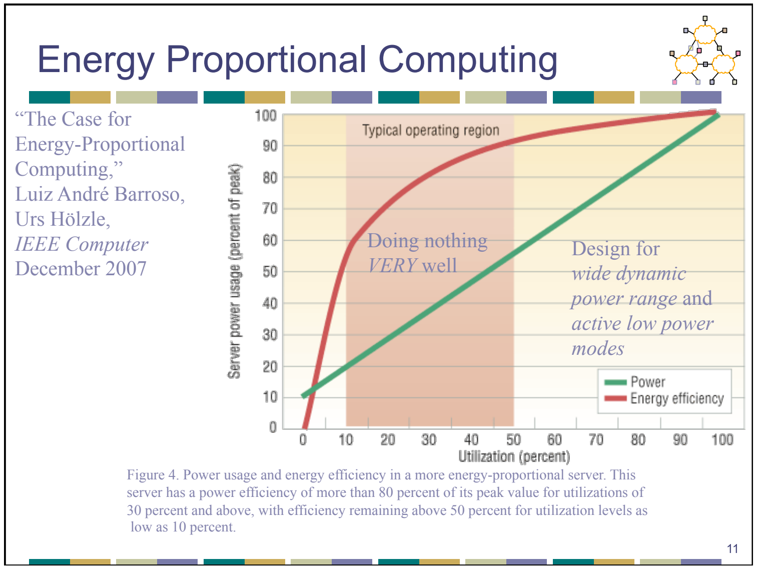


Energy Proportional Computing

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

Doing nothing well … *NOT!*

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.



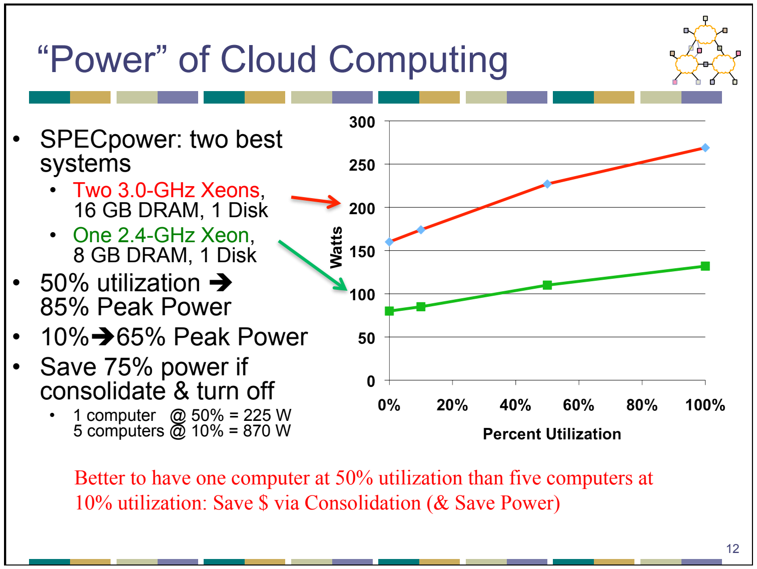
Energy Proportional Computing

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

Doing nothing *VERY* well

Design for *wide dynamic power range* and *active low power modes*

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.



"Power" of Cloud Computing

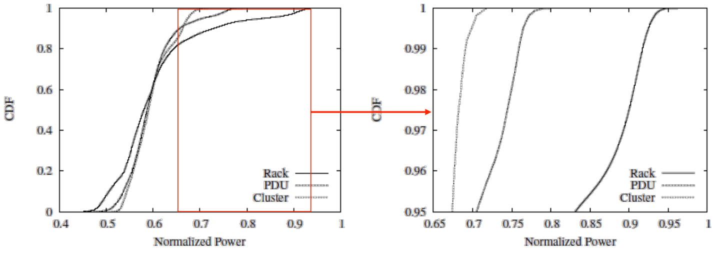
- SPECpower: two best systems
  - Two 3.0-GHz Xeons, 16 GB DRAM, 1 Disk
  - One 2.4-GHz Xeon, 8 GB DRAM, 1 Disk
- 50% utilization ➔ 85% Peak Power
- 10%➔65% Peak Power
- Save 75% power if consolidate & turn off
  - 1 computer @ 50% = 225 W
    5 computers @ 10% = 870 W

Better to have one computer at 50% utilization than five computers at 10% utilization: Save $ via Consolidation (& Save Power)

Bringing Resources On-/Off-line

- Save power by taking DC "slices" off-line
  - Resource footprint of applications hard to model
  - Dynamic environment, complex cost functions require measurement-driven decisions -- opportunity for statistical machine learning
  - Must maintain Service Level Agreements, no negative impacts on hardware reliability
  - Pervasive use of virtualization (VMs, VLANs, VStor) makes feasible rapid shutdown/migration/restart
- Recent results suggest that conserving energy may actually improve reliability
  - MTTF: stress of on/off cycle vs. benefits of off-hours

Typical Datacenter Power



Power-aware allocation of resources can achieve higher levels of utilization – harder to drive a cluster to high levels of utilization than an individual rack

X. Fan, W-D Weber, L. Barroso, "Power Provisioning for a Warehouse-sized Computer," ISCA'07, San Diego, (June 2007).

Aside: Disk Power

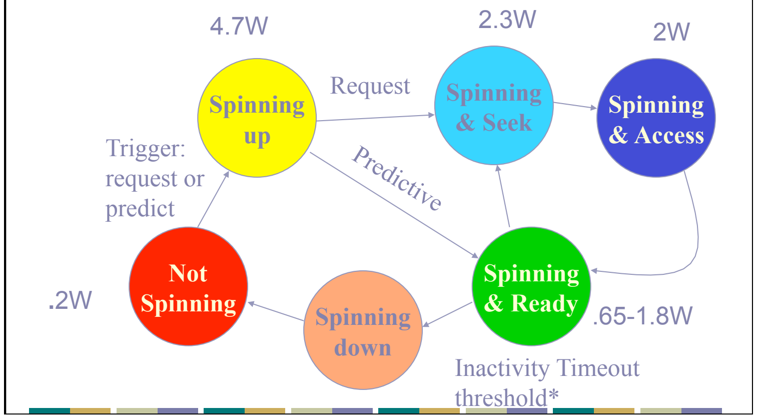
IBM Microdrive (1inch)

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

IBM TravelStar (2.5inch)

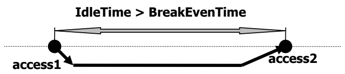
- read/write 2W
- spinning 1.8W
- low power idle .65W
- standby .25W
- sleep .1W
- startup 4.7 W
- seek 2.3W

Spin-down Disk Model

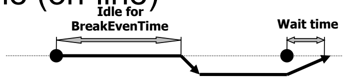


Disk Spindown

- Disk Power Management – Oracle (off-line)



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

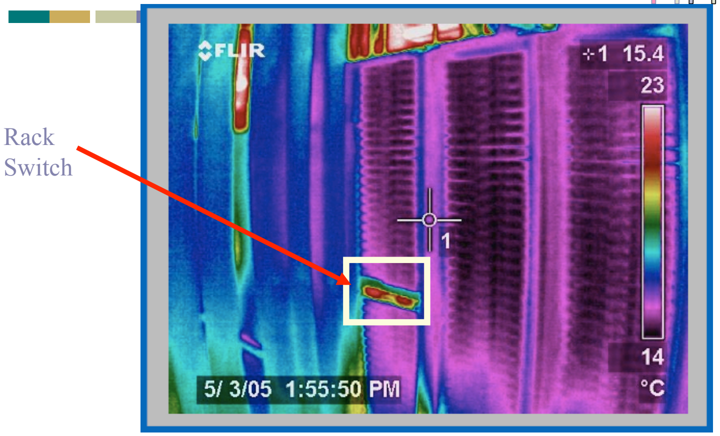


Source: from the presentation slides of the authors

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

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

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)

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

EEKLABS

Keep on trucking

AMERICAN POWER CONVERSION CORP.'S InfraStruxure Express mobile data center can deliver power and Internet connectivity when there are no other options.

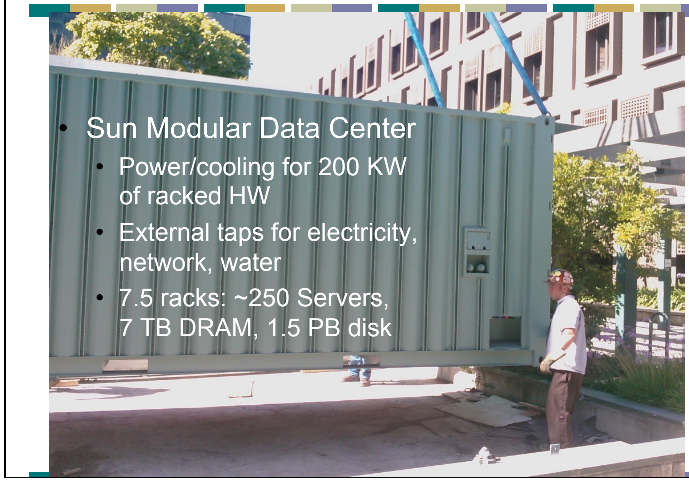
InfraStruxure Express is a fully opera-

officials said that the cost of a lease depends on financing options but that companies could expect to pay about $20,000 per month. They added that InfraStruxure Express can be delivered anywhere in the continental United

provide as much as 400 kilowatts of power, and it has external feeds that can be used to deliver temporary power to buildings.

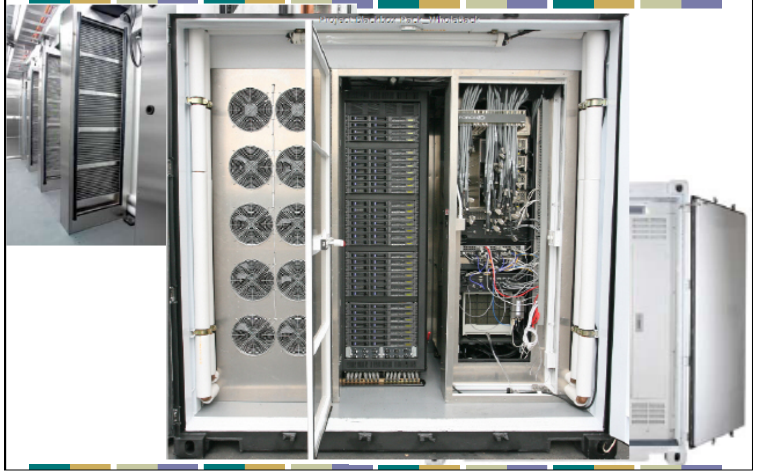
The on-board cooling is adequate for data center environments, and the trailer is

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

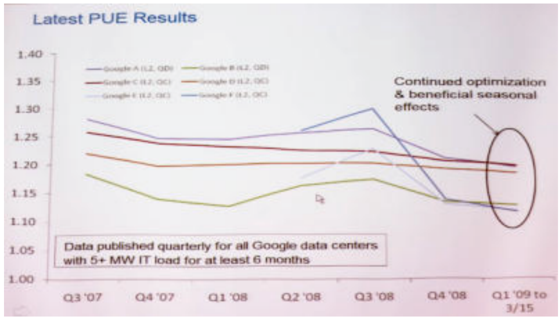
Containerized Datacenters



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

Google's PUE


Latest PUE Results

Continued optimization & beneficial seasonal effects

Data published quarterly for all Google data centers with 5+ MW IT load for at least 6 months

- 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

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

Overview

- Data Center Overview
- Networking in the DC

Layer 2 vs. Layer 3 for Data Centers

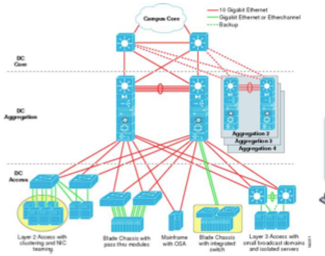
Technique	Plug and play	Scalability	Small Switch State	Seamless VM Migration
Layer 2: Flat MAC Addresses	+	−	−	+
Layer 3: IP Addresses	−	+	+	−

Flat vs. Location Based Addresses

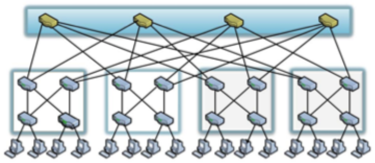
- Commodity switches today have ~640 KB of low latency, power hungry, expensive on chip memory
  - Stores 32 – 64 K flow entries
- Assume 10 million virtual endpoints in 500,000 servers in datacenter
- Flat addresses ➔ 10 million address mappings ➔ ~100 MB on chip memory ➔ ~150 times the memory size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

PortLand: Main Assumption

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

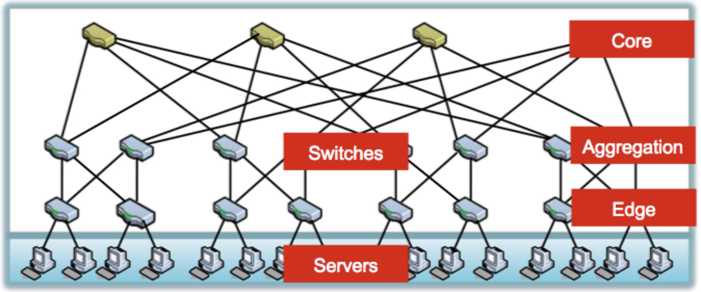

Campus Core
10 Gigabit Ethernet
Gigabit Ethernet or Etherchannel
Backup
DC Core
DC Aggregation
Aggregation 2
Aggregation 3
Aggregation 4
DC Access
Layer 2 Access with clustering and NIC teaming
Blade Chassis with pass thru modules
Mainframe with OSA
Blade Chassis with integrated switch
Layer 3 Access with small broadcast domains and isolated servers

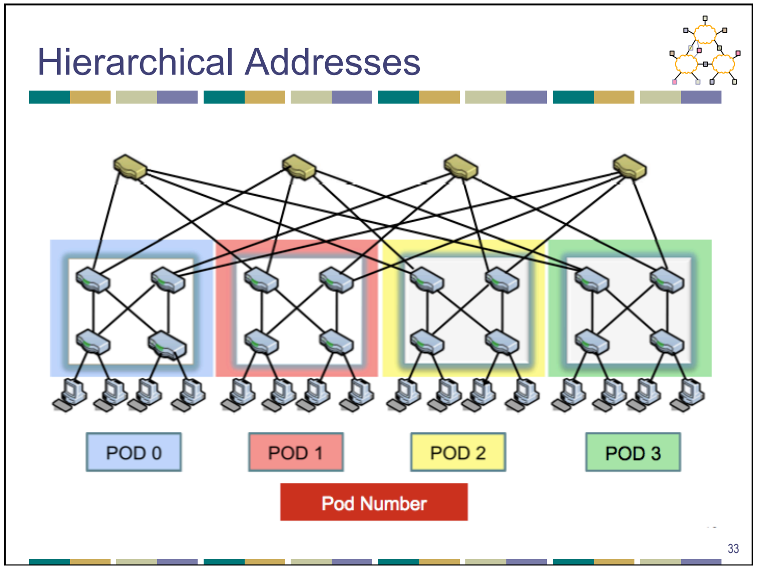
Cisco Recommended Configuration



Fat Tree

Data Center Network

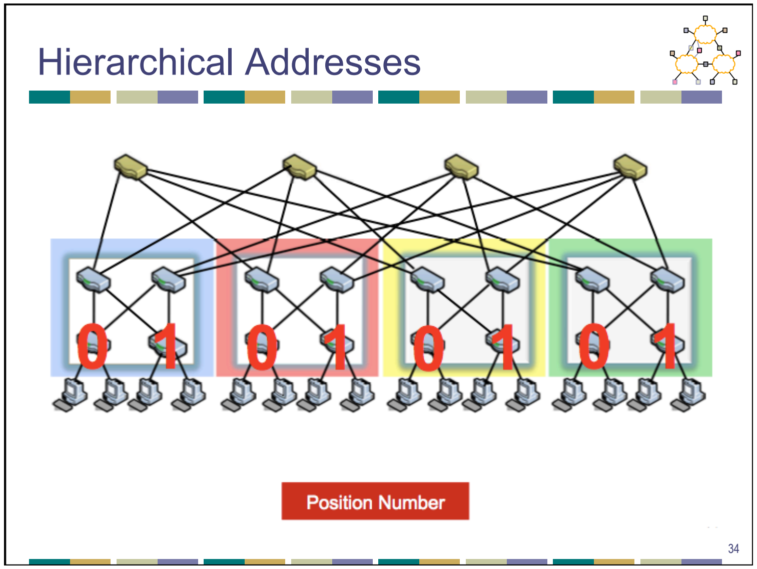




Hierarchical Addresses

Pod Number

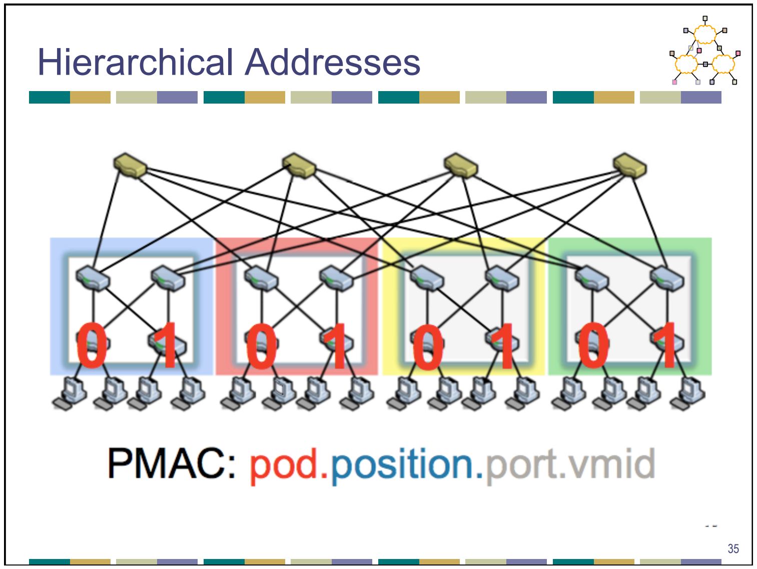
33



Hierarchical Addresses

Position Number

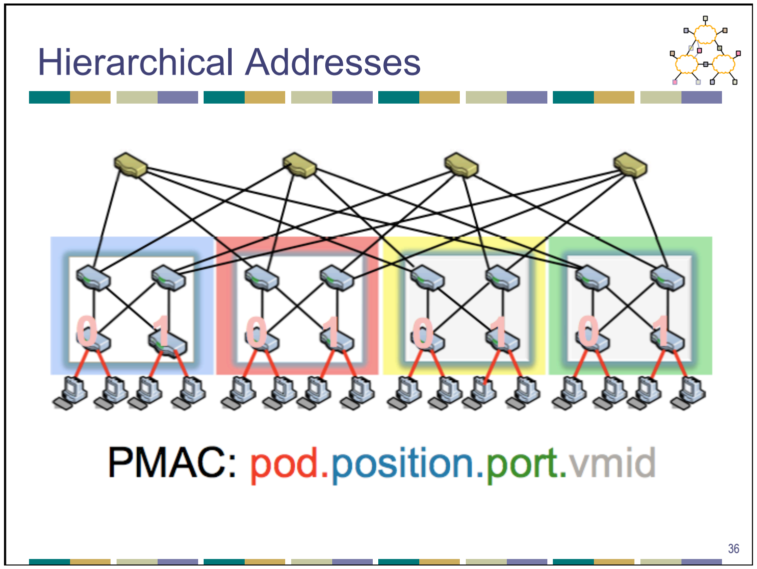
34



Hierarchical Addresses

PMAC: pod.position.port.vmid

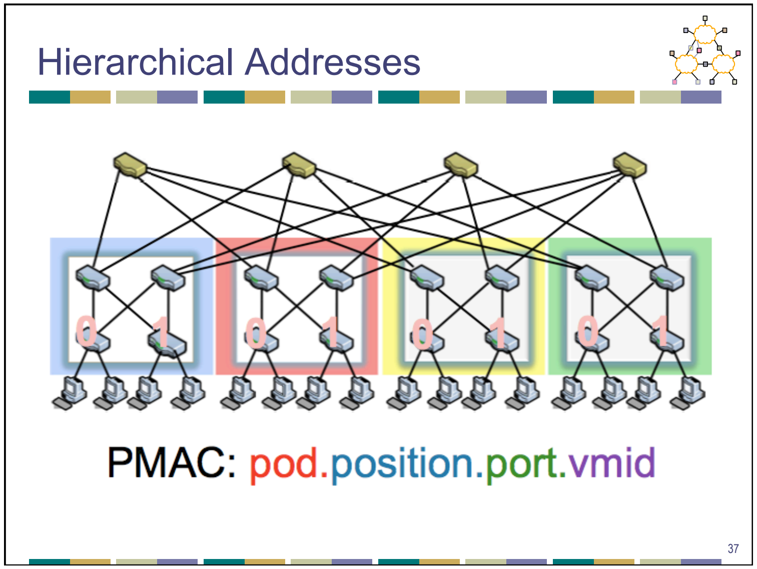
35



Hierarchical Addresses

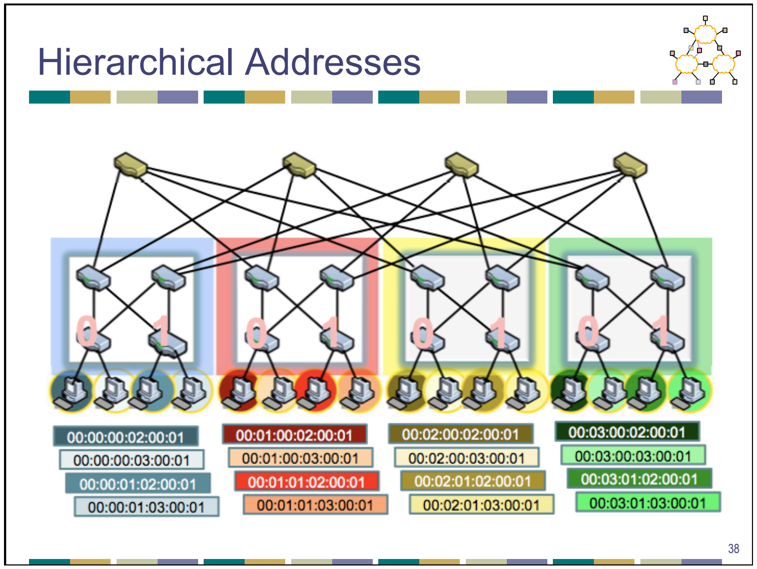
PMAC: pod.position.port.vmid

36

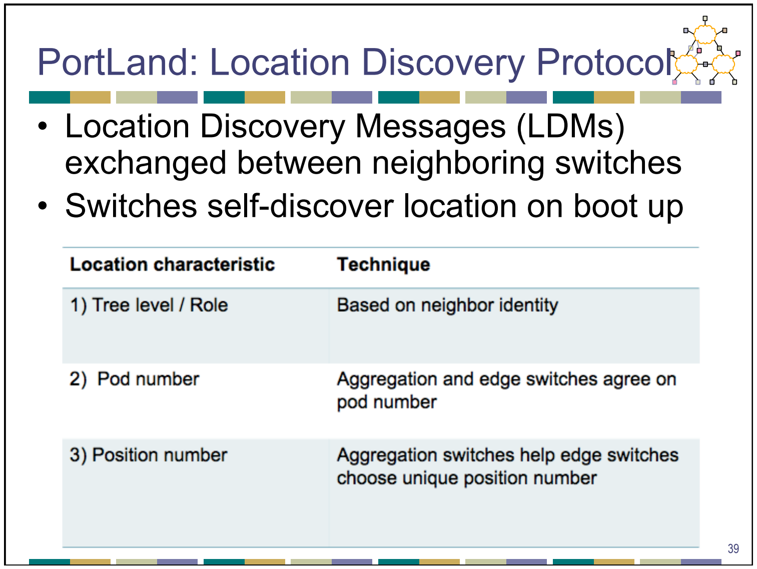


Hierarchical Addresses

PMAC: pod.position.port.vmid



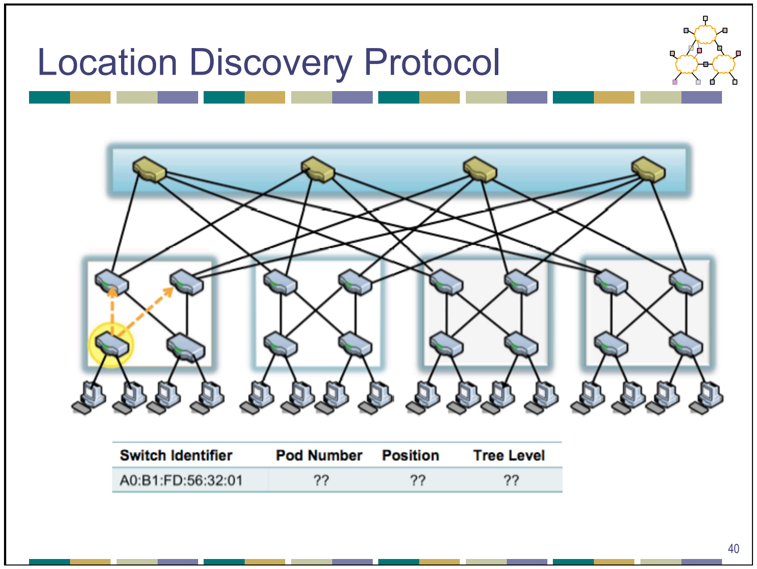
Hierarchical Addresses



PortLand: Location Discovery Protocol

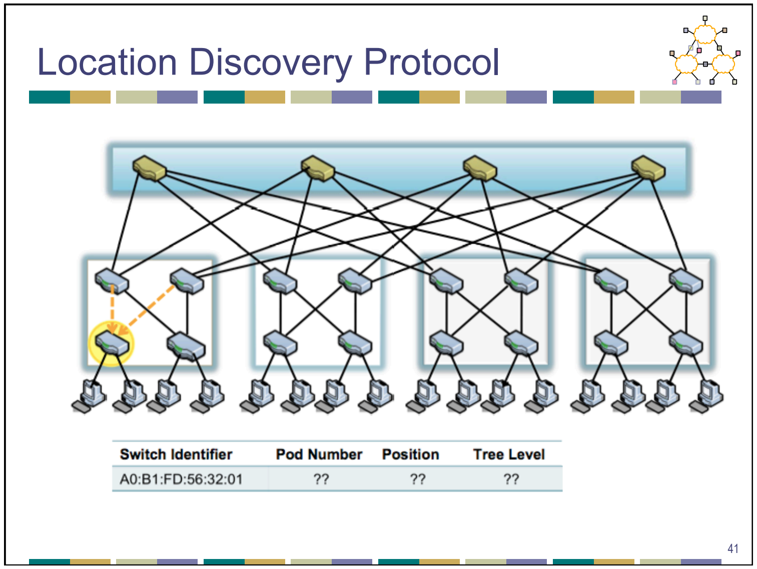
- Location Discovery Messages (LDMs) exchanged between neighboring switches
- Switches self-discover location on boot up

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



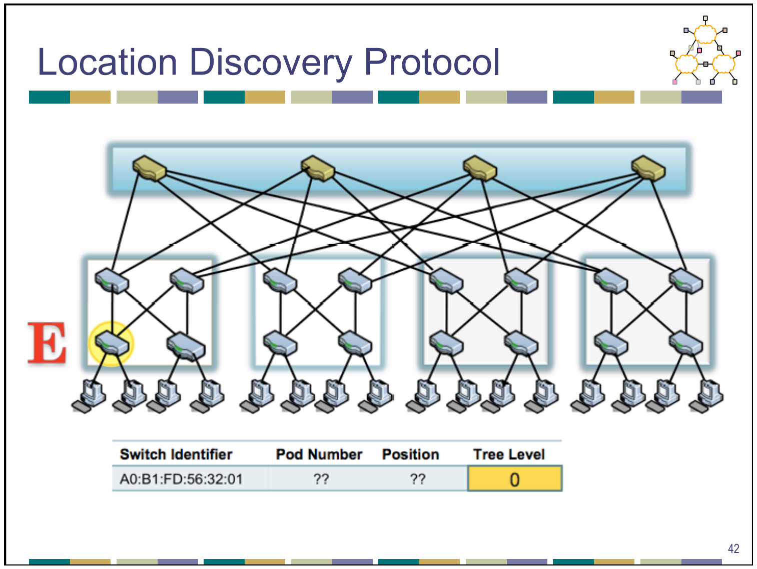
Location Discovery Protocol

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



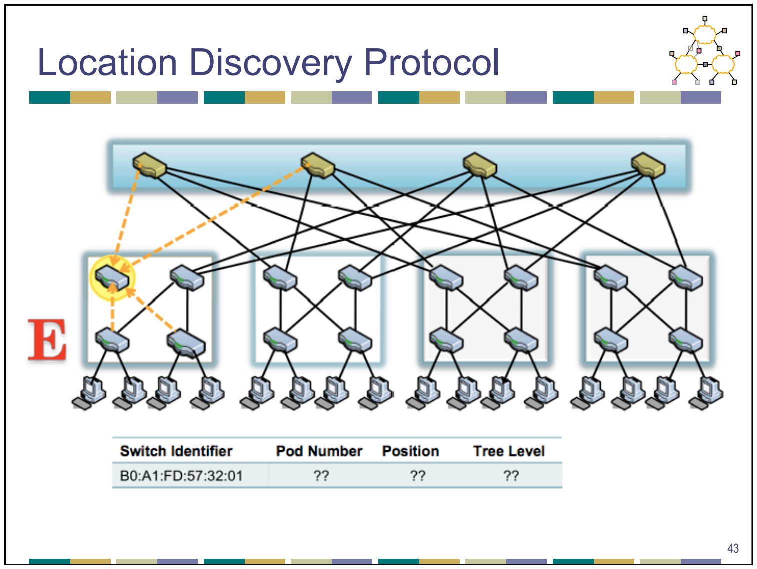
Location Discovery Protocol

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



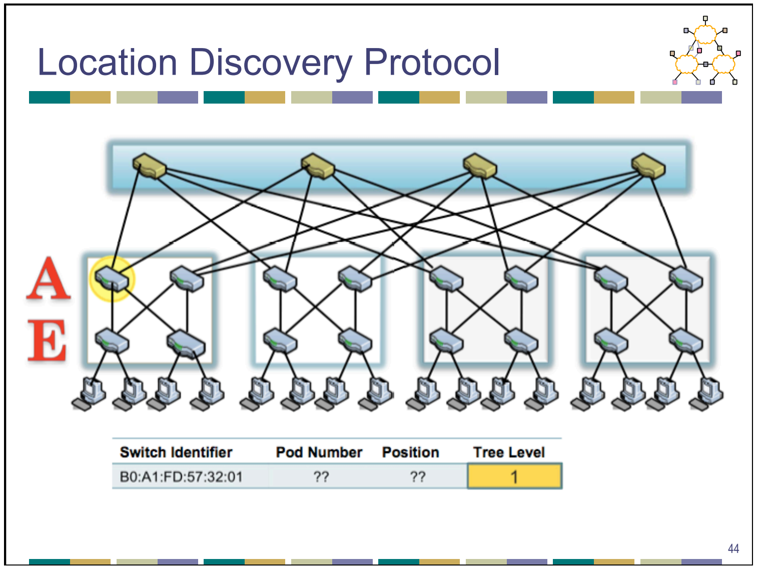
Location Discovery Protocol

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



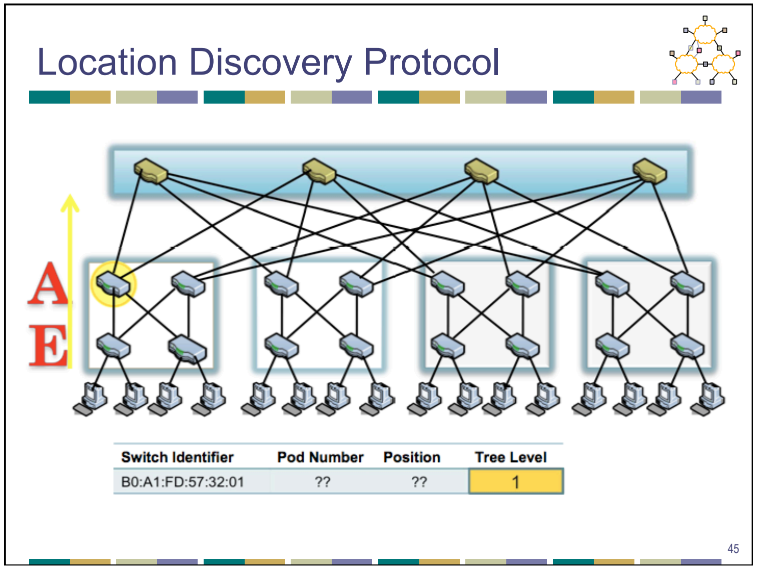
Location Discovery Protocol

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



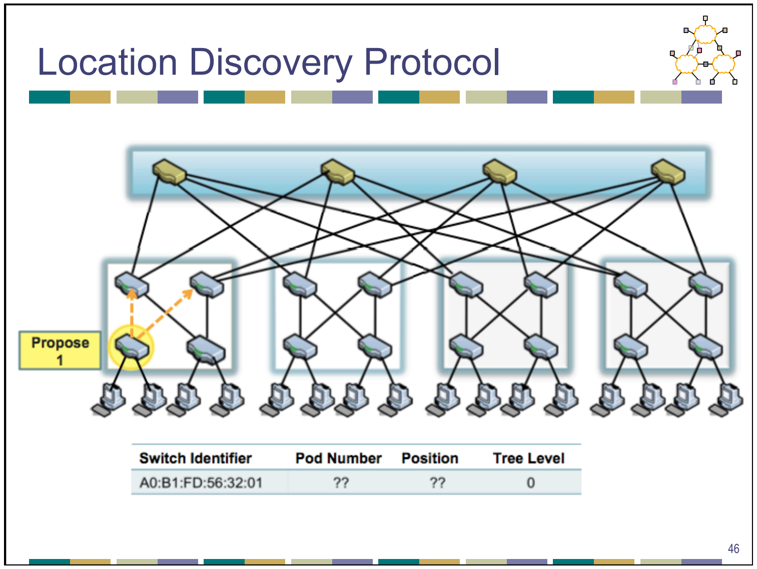
Location Discovery Protocol

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



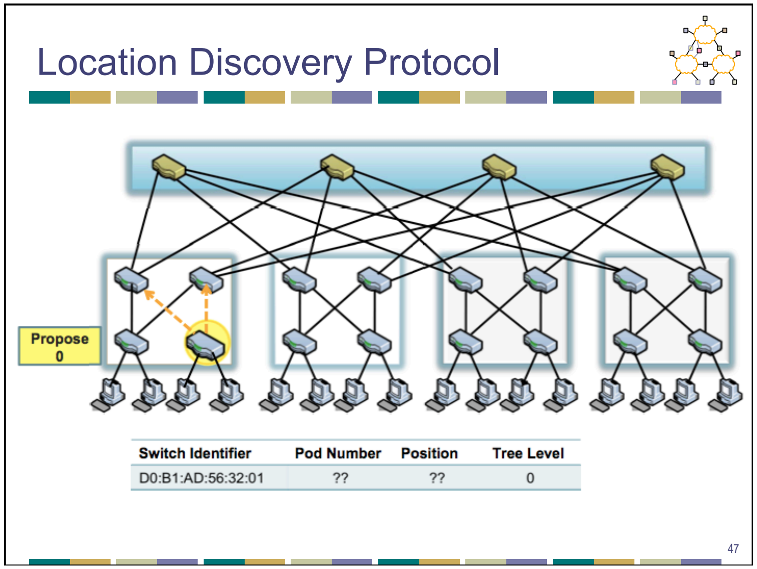
Location Discovery Protocol

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



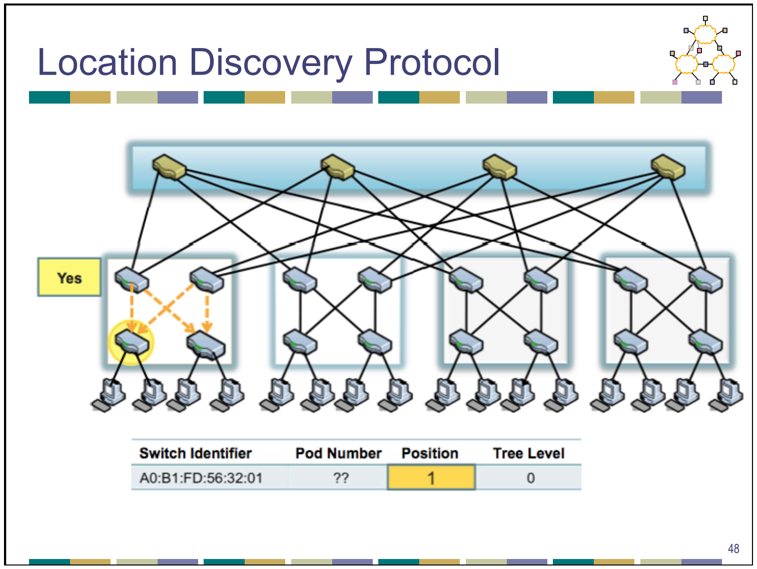
Location Discovery Protocol

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



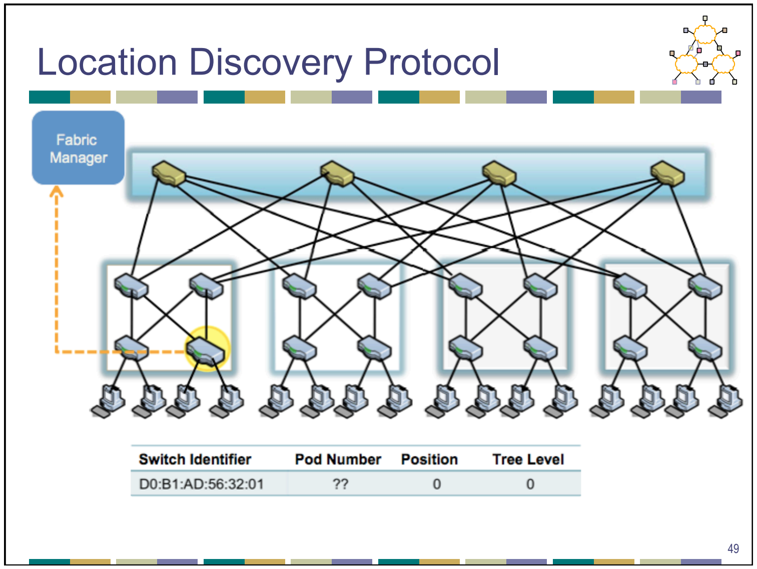
Location Discovery Protocol

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



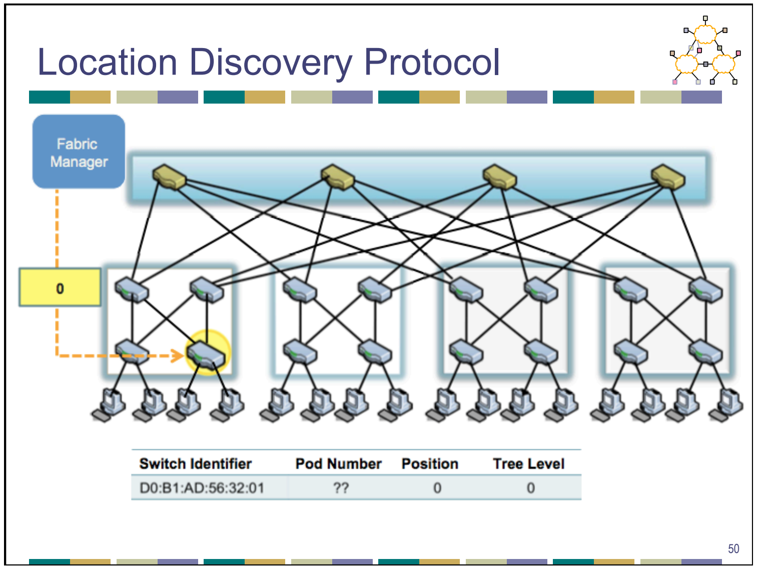
Location Discovery Protocol

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



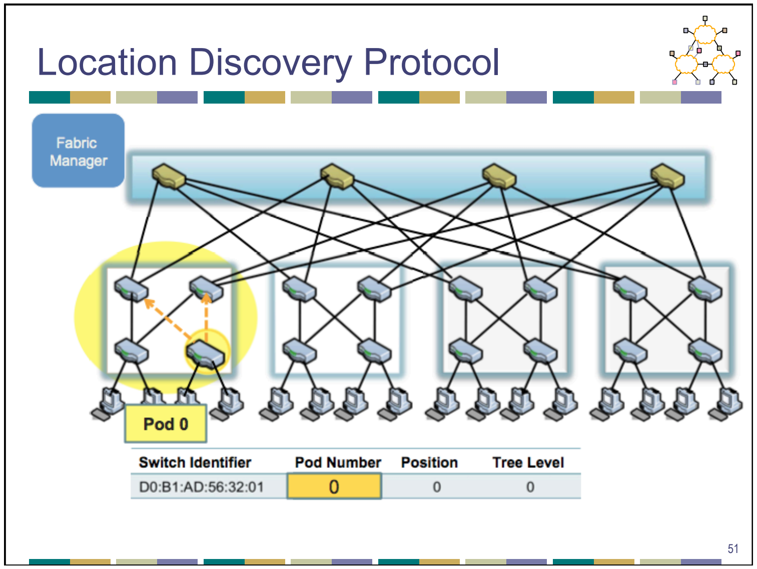
Location Discovery Protocol

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



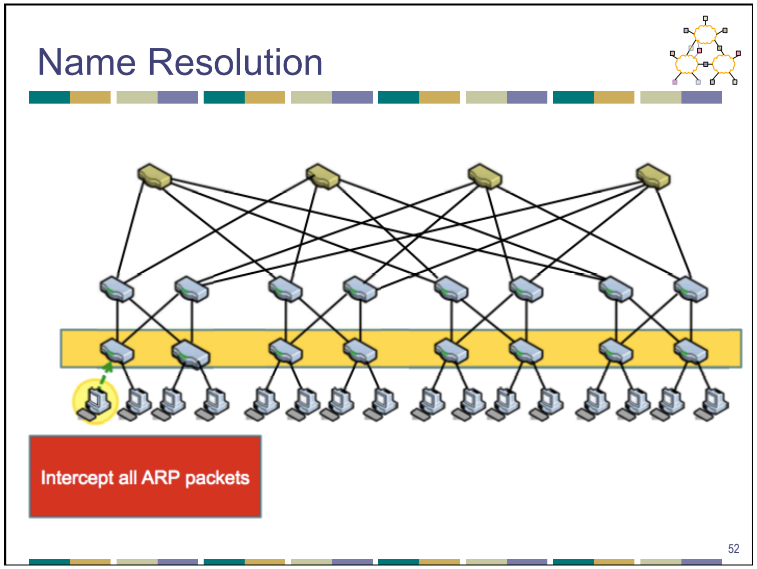
Location Discovery Protocol

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



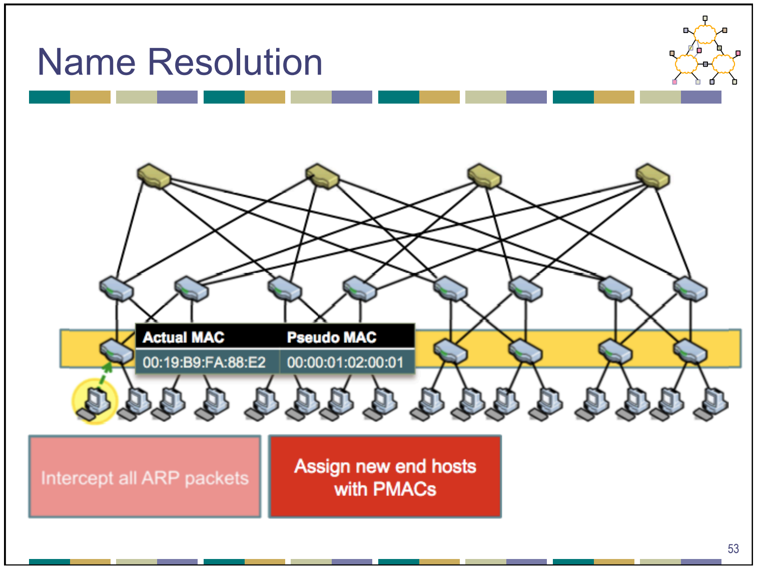
Location Discovery Protocol

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



Name Resolution

Intercept all ARP packets



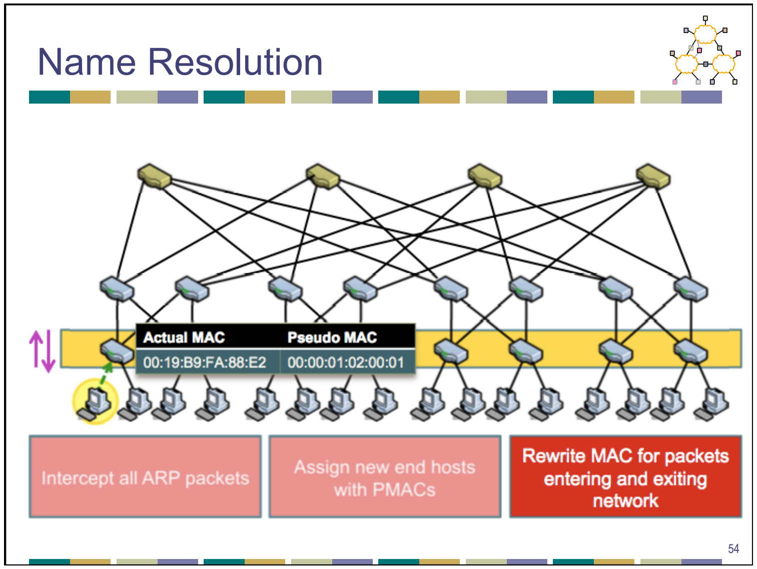
Name Resolution

Actual MAC	Pseudo MAC
00:19:B9:FA:88:E2	00:00:01:02:00:01

Intercept all ARP packets

Assign new end hosts with PMACs

53



Name Resolution

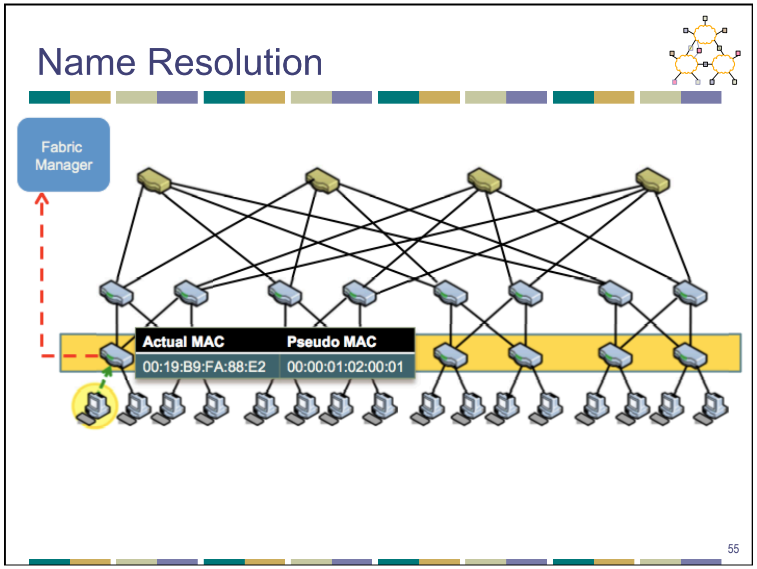
Actual MAC	Pseudo MAC
00:19:B9:FA:88:E2	00:00:01:02:00:01

Intercept all ARP packets

Assign new end hosts with PMACs

Rewrite MAC for packets entering and exiting network

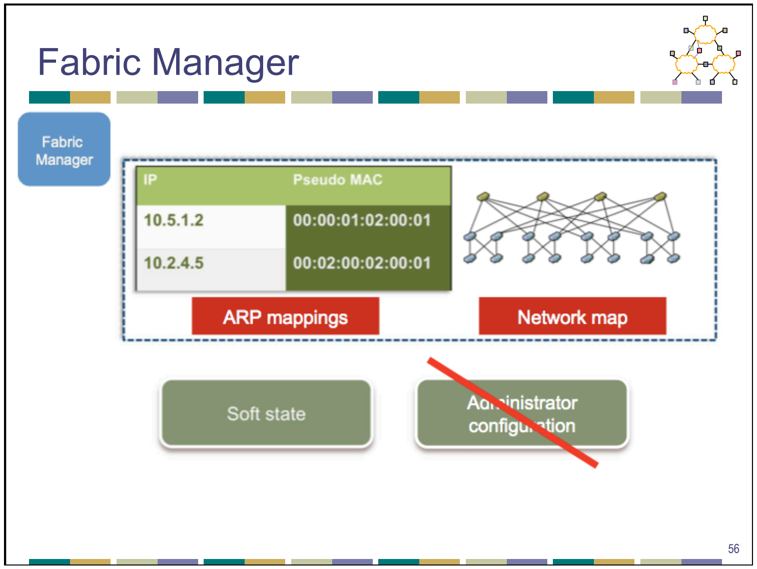
54



Name Resolution

Actual MAC	Pseudo MAC
00:19:B9:FA:88:E2	00:00:01:02:00:01

55



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

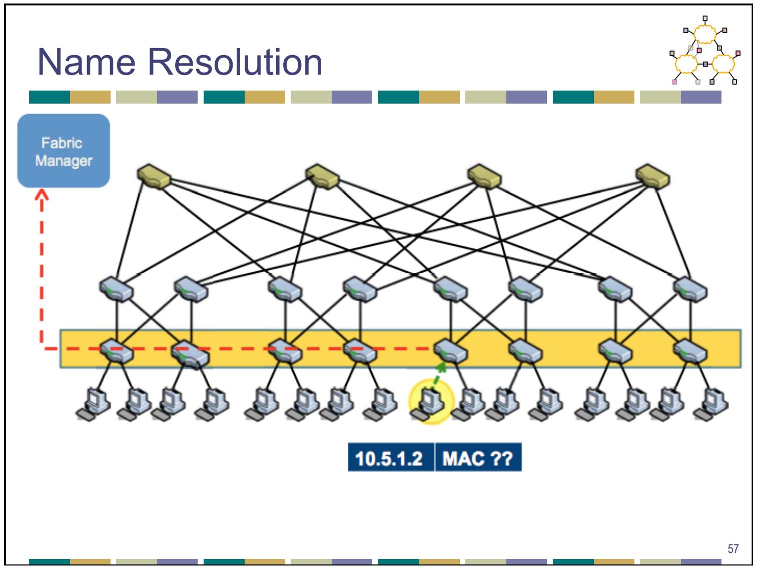
ARP mappings

Network map

Soft state

~~Administrator configuration~~

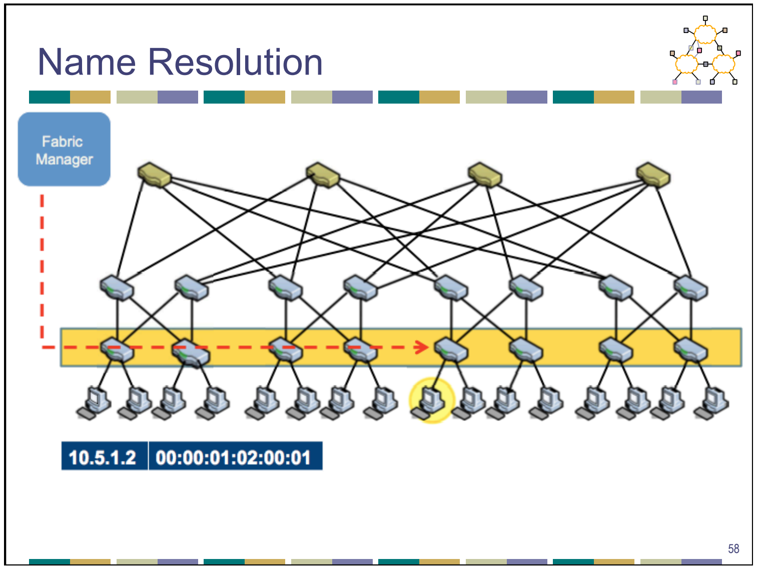
56



Name Resolution

Fabric Manager

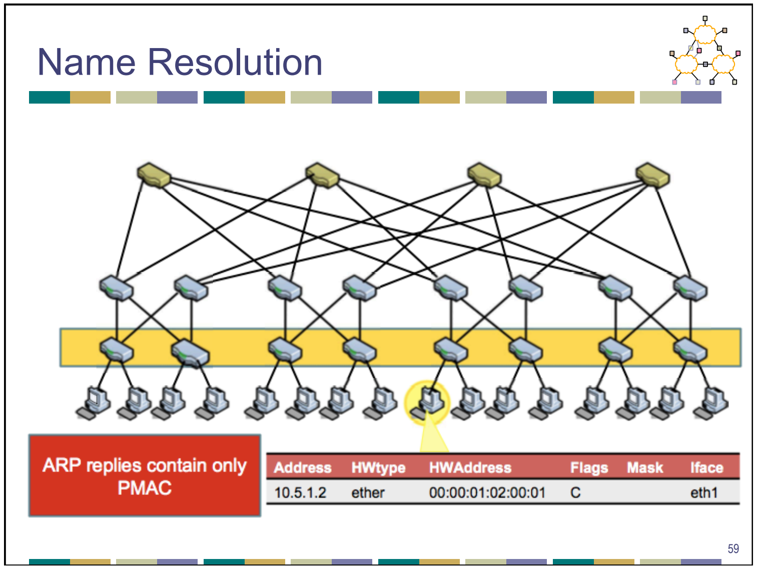
10.5.1.2	MAC ??



Name Resolution

Fabric Manager

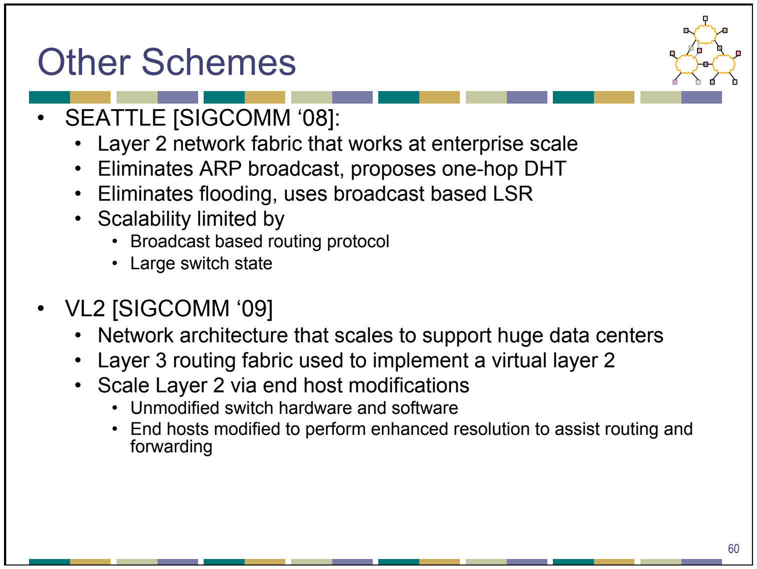
10.5.1.2	00:00:01:02:00:01



Name Resolution

ARP replies contain only PMAC

Address	HWtype	HWAddress	Flags	Mask	Iface
10.5.1.2	ether	00:00:01:02:00:01	C		eth1

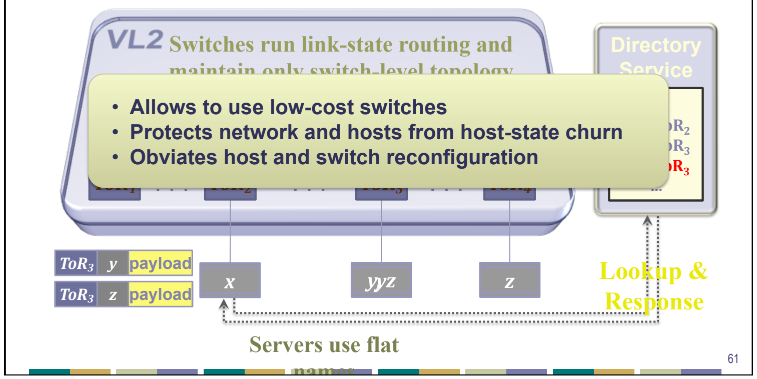


Other Schemes

- 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

VL2: Name-Location Separation

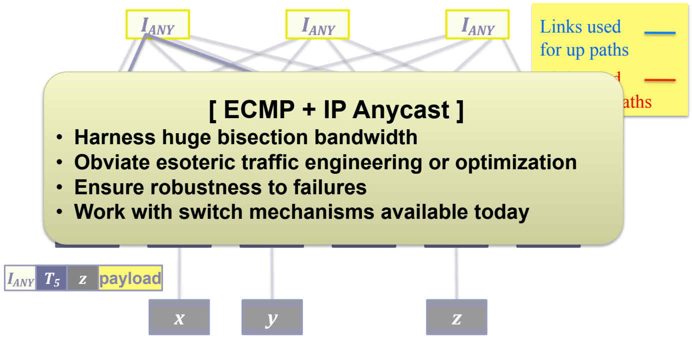
Cope with host churns with very little overhead



- Allows to use low-cost switches
- Protects network and hosts from host-state churn
- Obviates host and switch reconfiguration

VL2: Random Indirection

Cope with arbitrary TMs with very little overhead



[ECMP + IP Anycast]

- Harness huge bisection bandwidth
- Obviate esoteric traffic engineering or optimization
- Ensure robustness to failures
- Work with switch mechanisms available today