CE693: Adv. Computer Networking

L-12 Data Center Networking

Acknowledgments: Lecture slides are from the graduate level Computer Networks course thought by Srinivasan Seshan at CMU. When slides are obtained from other sources, a a reference will be noted on the bottom of that slide. A full list of references is provided on the last slide.

Overview

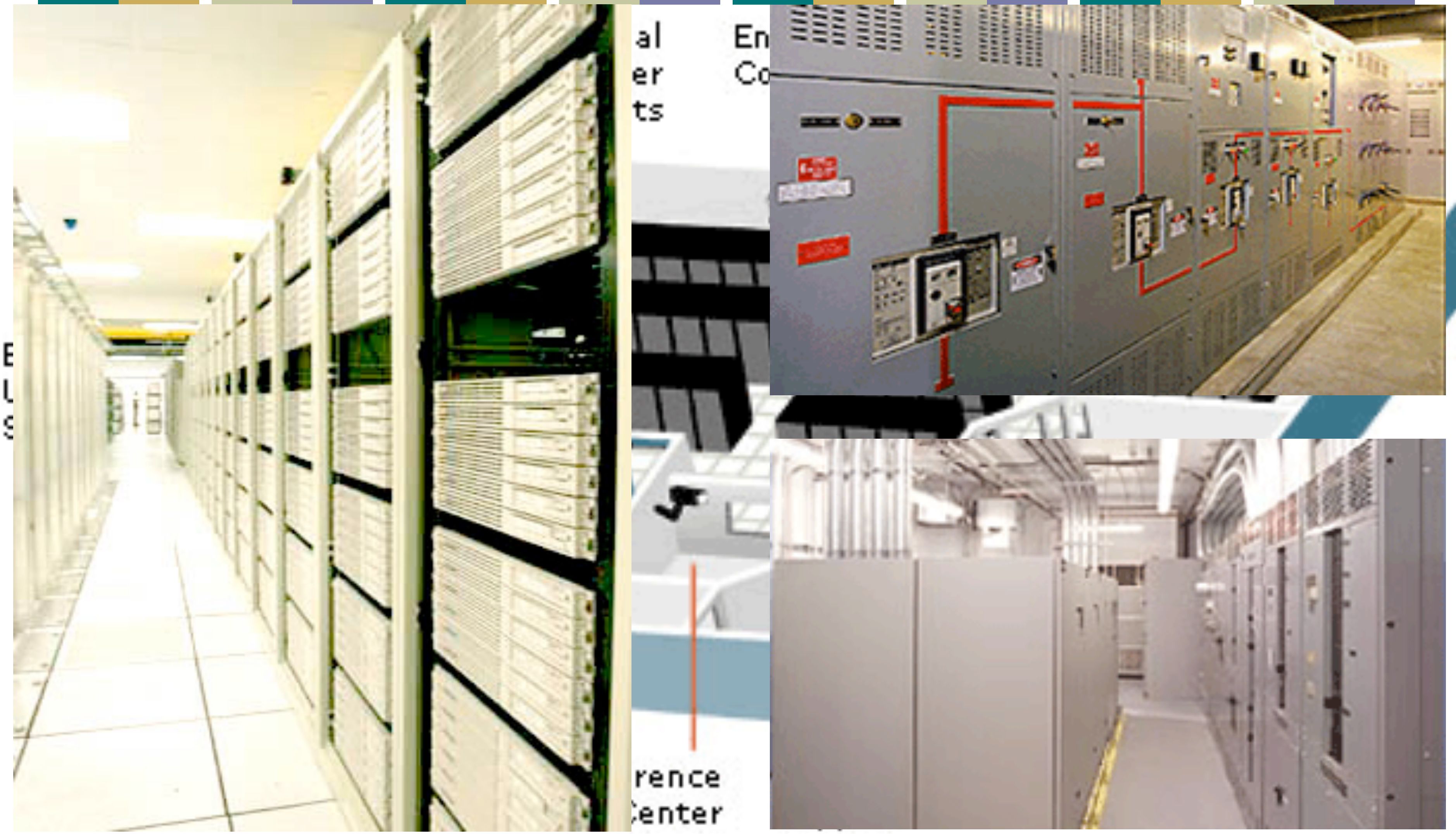
- Data Center Overview
- Routing in the DC
- Transport in the DC

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*



Energy Proportional Computing

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

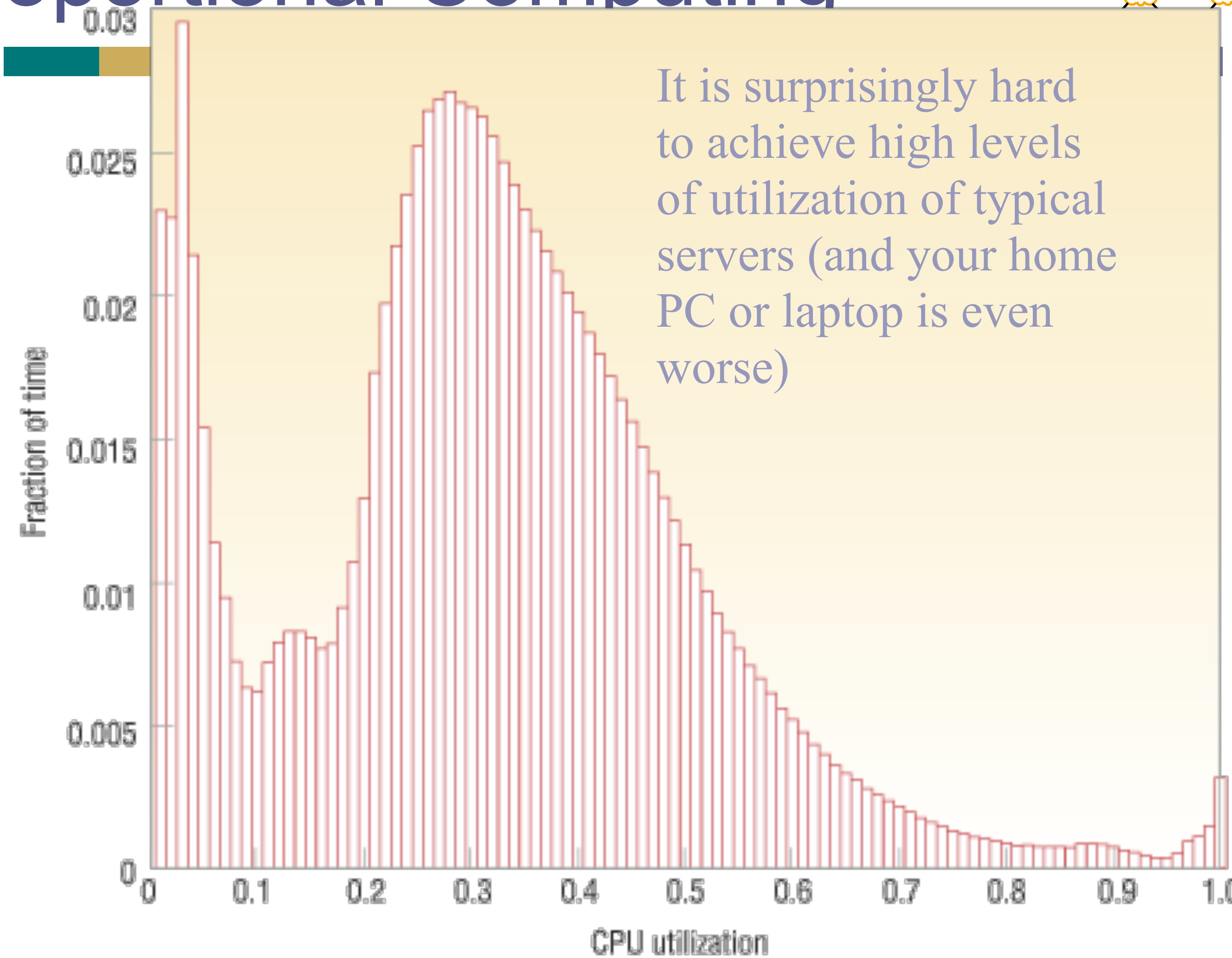


Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum

Energy Proportional Computing

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

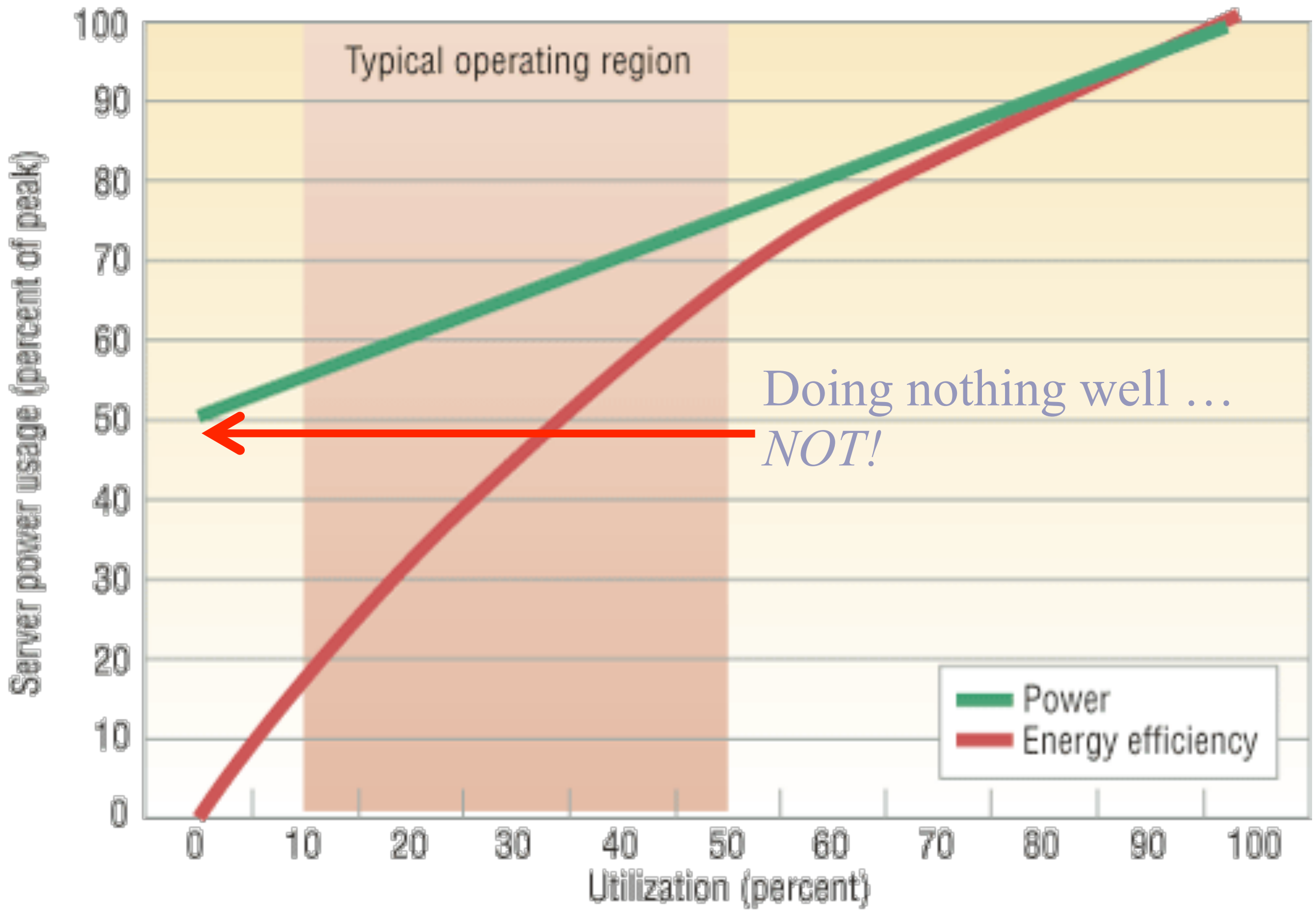


Figure 2. Server power usage and energy efficiency at varying utilization levels, from idle to peak performance. Even an energy-efficient server still consumes about half its full power when doing virtually no work.

Energy Proportional Computing

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

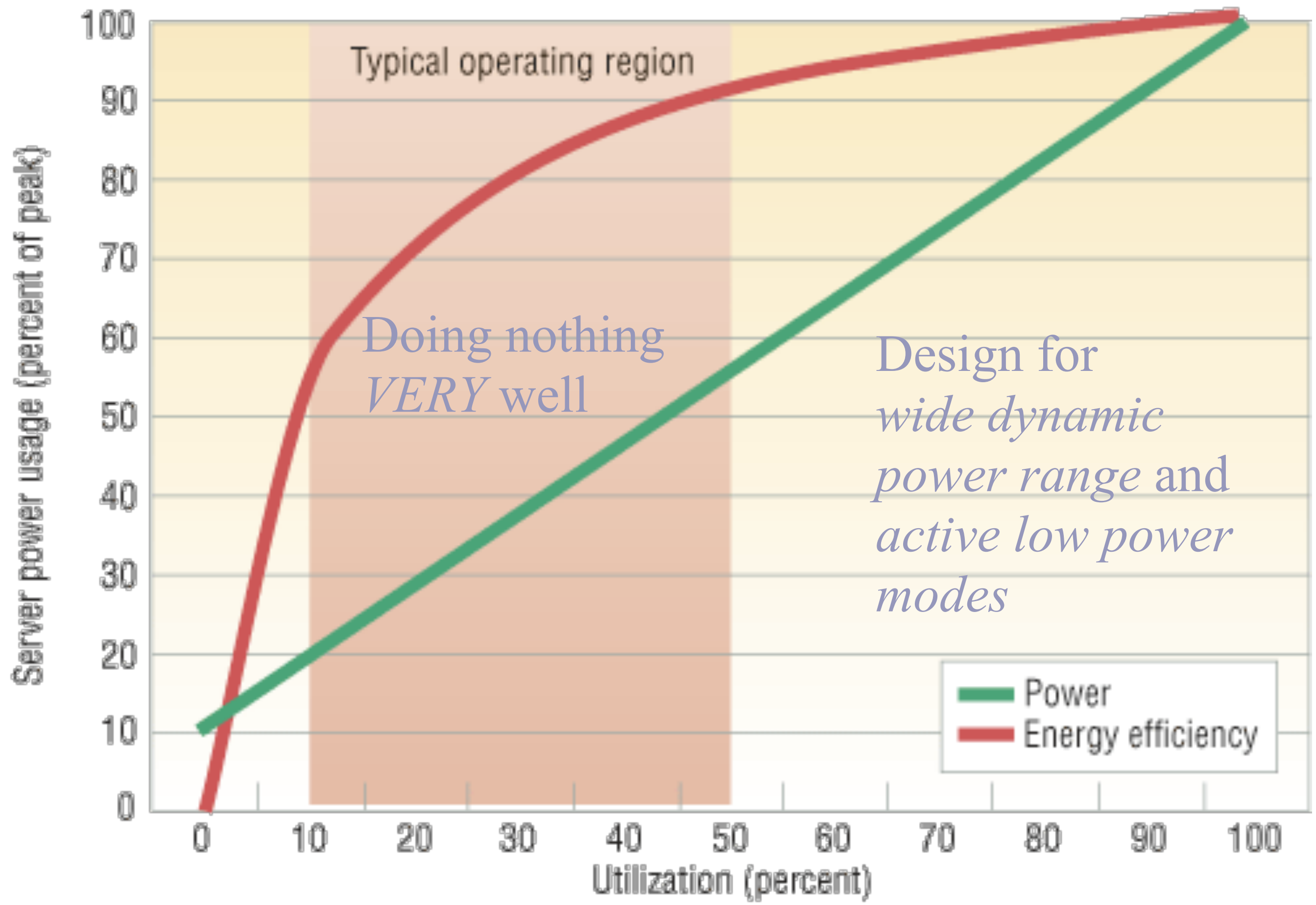
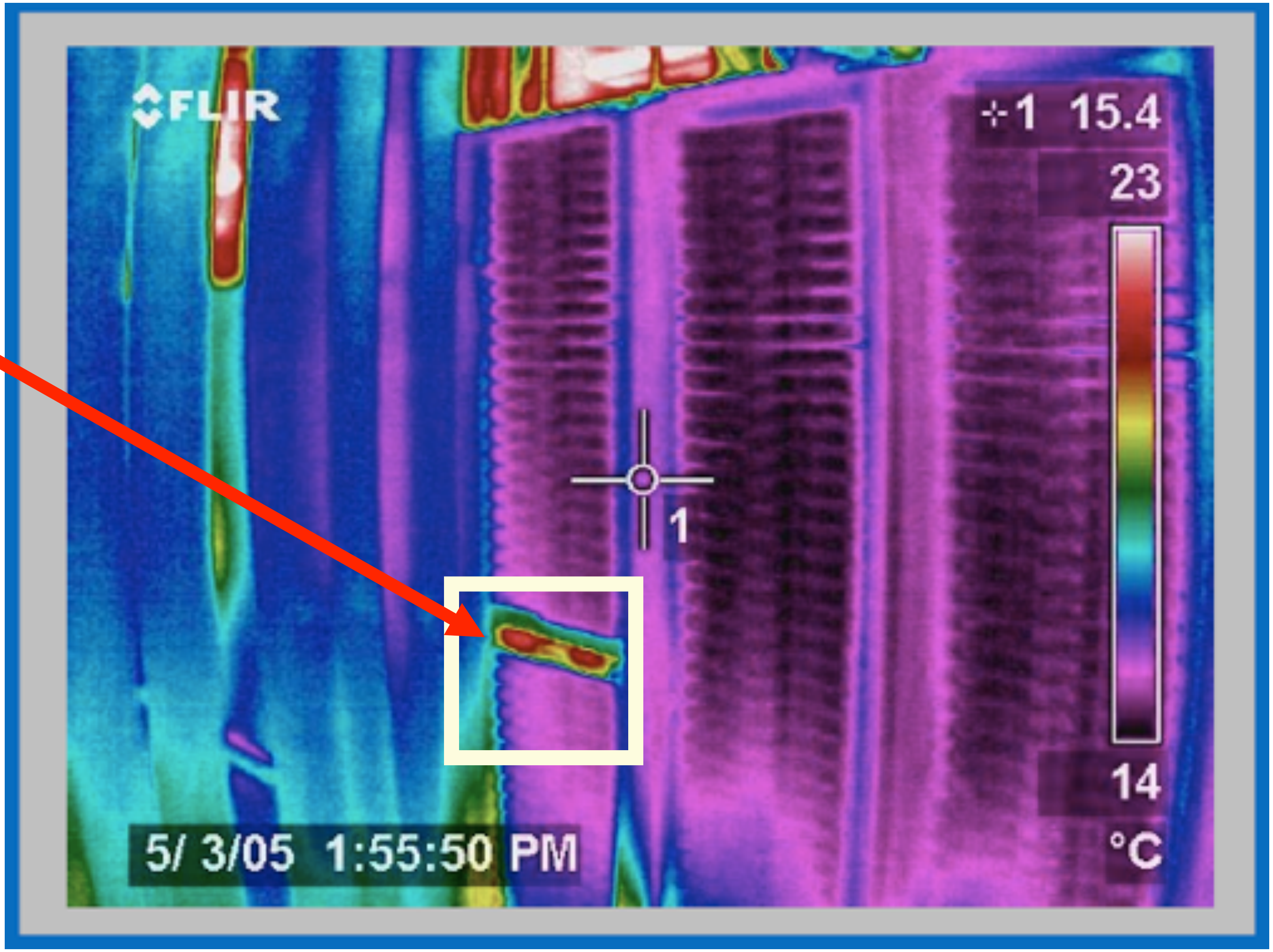


Figure 4. Power usage and energy efficiency in a more energy-proportional server. This server has a power efficiency of more than 80 percent of its peak value for utilizations of 30 percent and above, with efficiency remaining above 50 percent for utilization levels as low as 10 percent.

Thermal Image of Typical Cluster

Rack Switch



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

- 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

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.

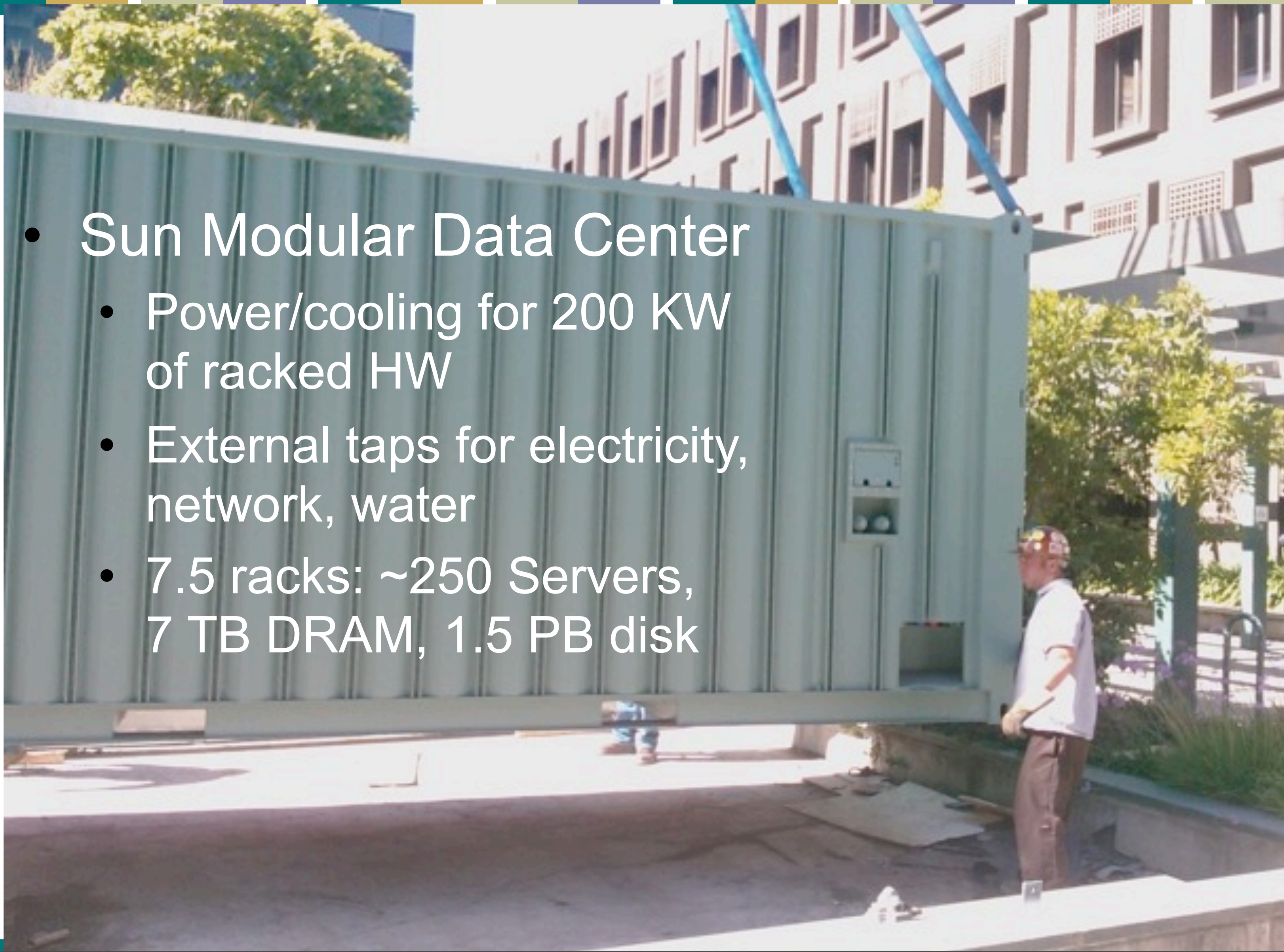
InfraStruxure Express is a fully operational data center on wheels (see photo)

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 States within three and a half days.

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

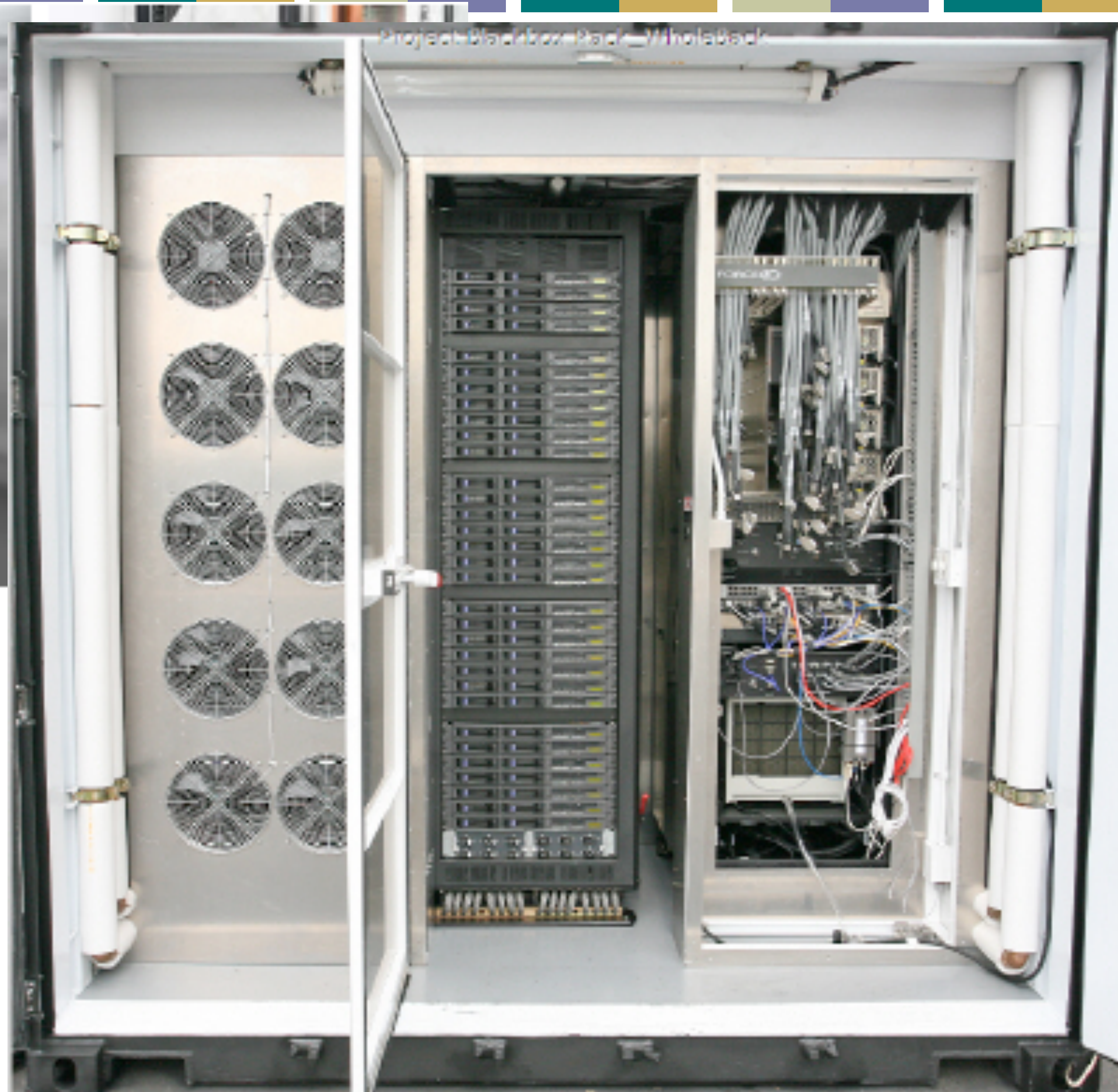
The on-board cooling is adequate for data center environments, and the trailer is configured with hot and cold aisles for

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



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
- Routing in the DC
- Transport in the DC

Layer 2 vs. Layer 3 for Data Centers

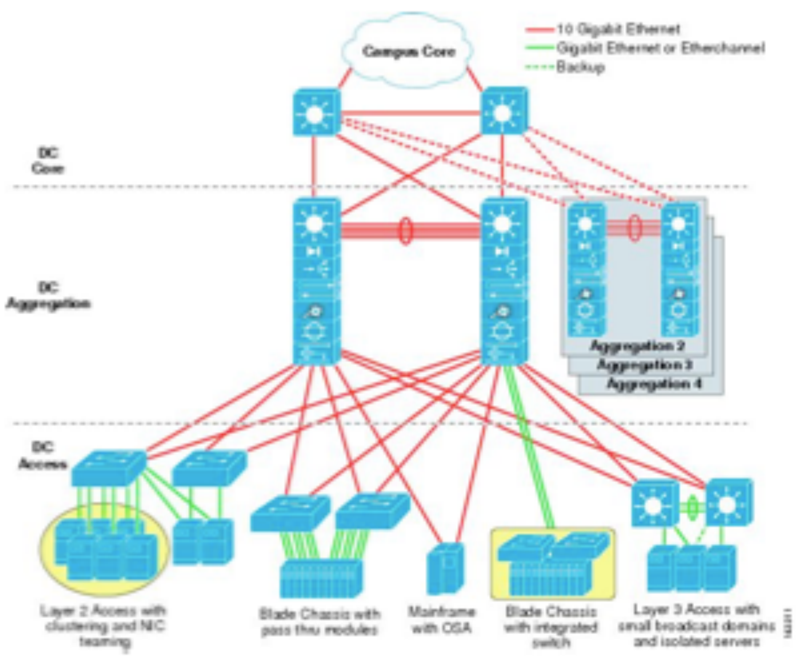
Technique	Plug and play	Scalability	Small Switch State	Seamless VM Migration
Layer 2: Flat MAC Addresses	+	–	–	+
Layer 3: IP Addresses	–	+	+	–

Flat vs. Location Based Addresses

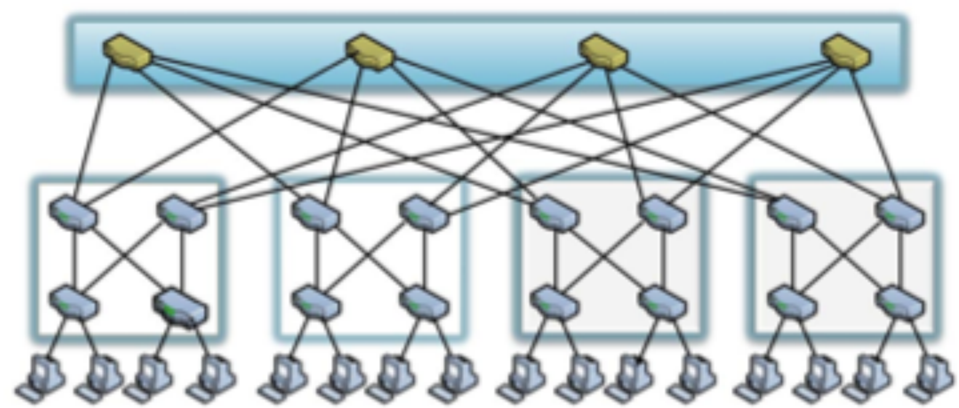
- Commodity switches today have ~640 KB of low latency, power hungry, expensive on chip memory
  - Stores 32 – 64 K flow entries
- Assume 10 million virtual endpoints in 500,000 servers in datacenter
- Flat addresses ➔ 10 million address mappings ➔ ~100 MB on chip memory ➔ ~150 times the memory size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

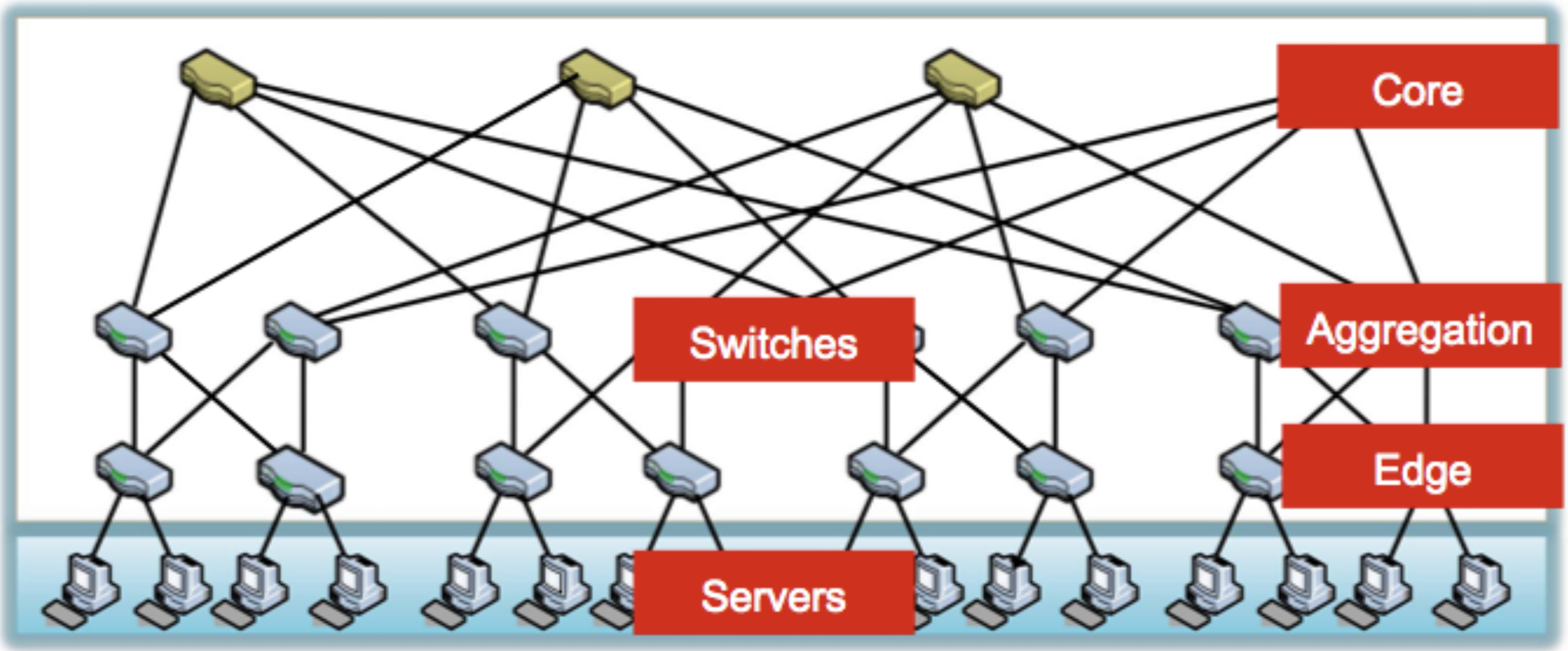


Cisco Recommended Configuration

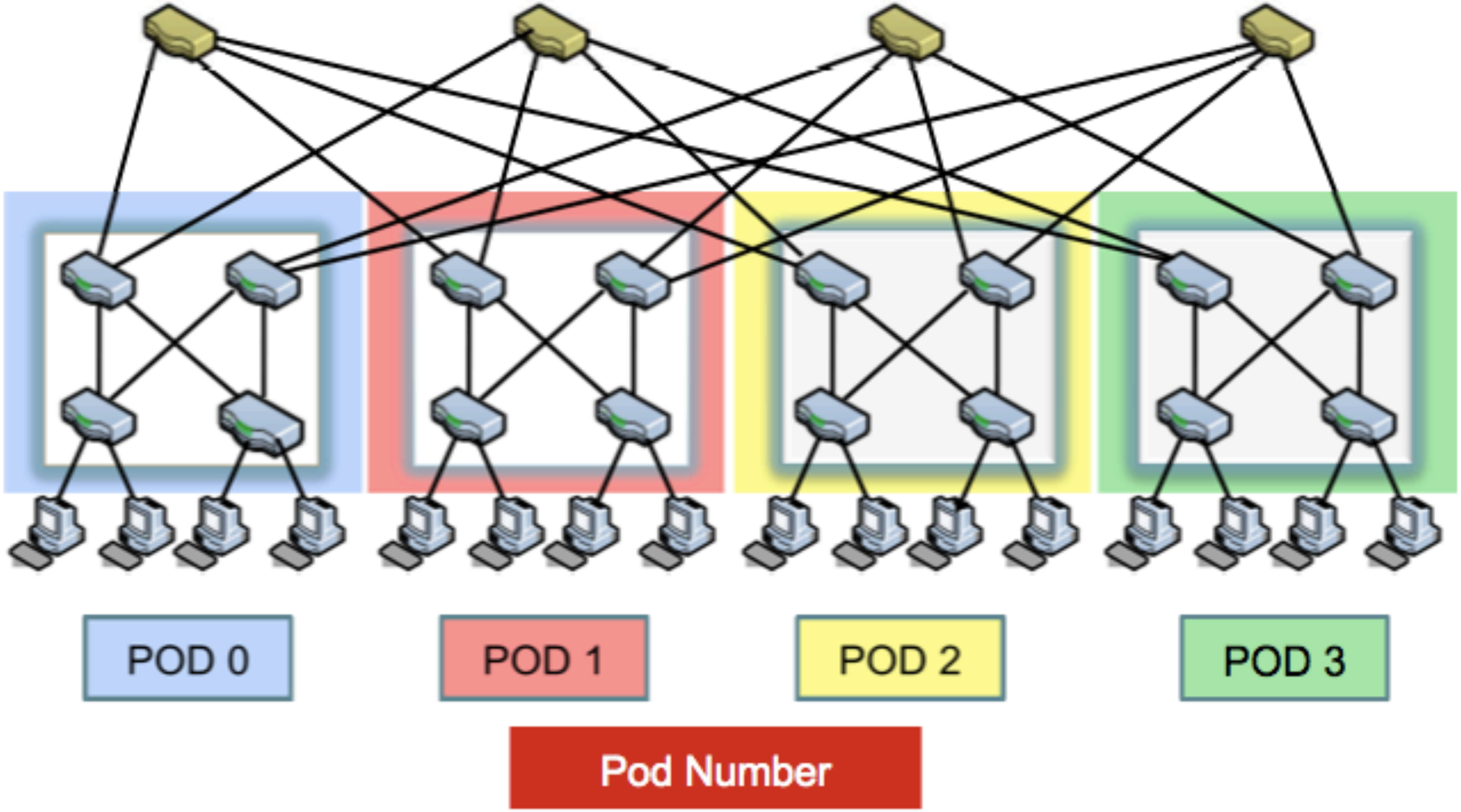


Fat Tree

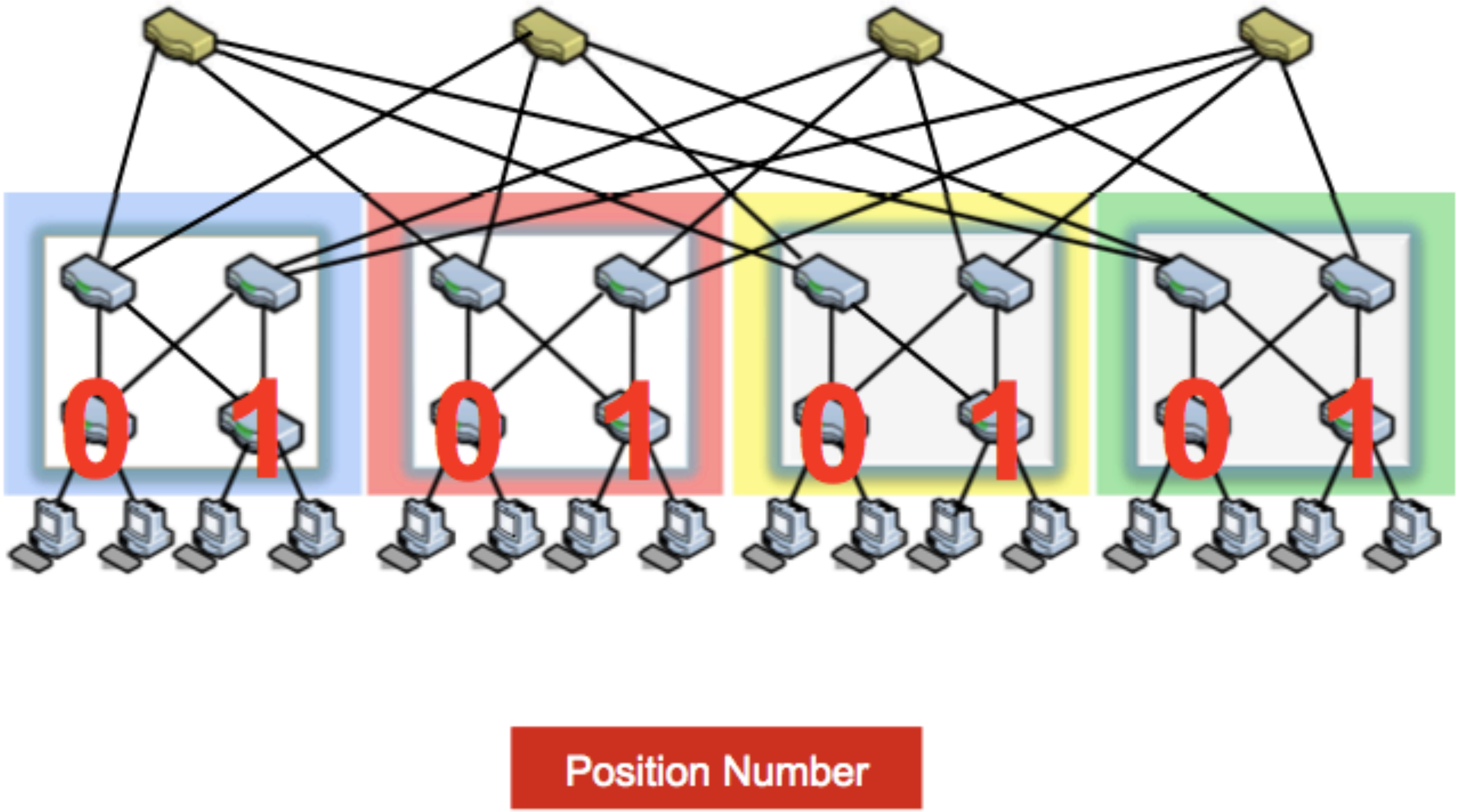
Data Center Network



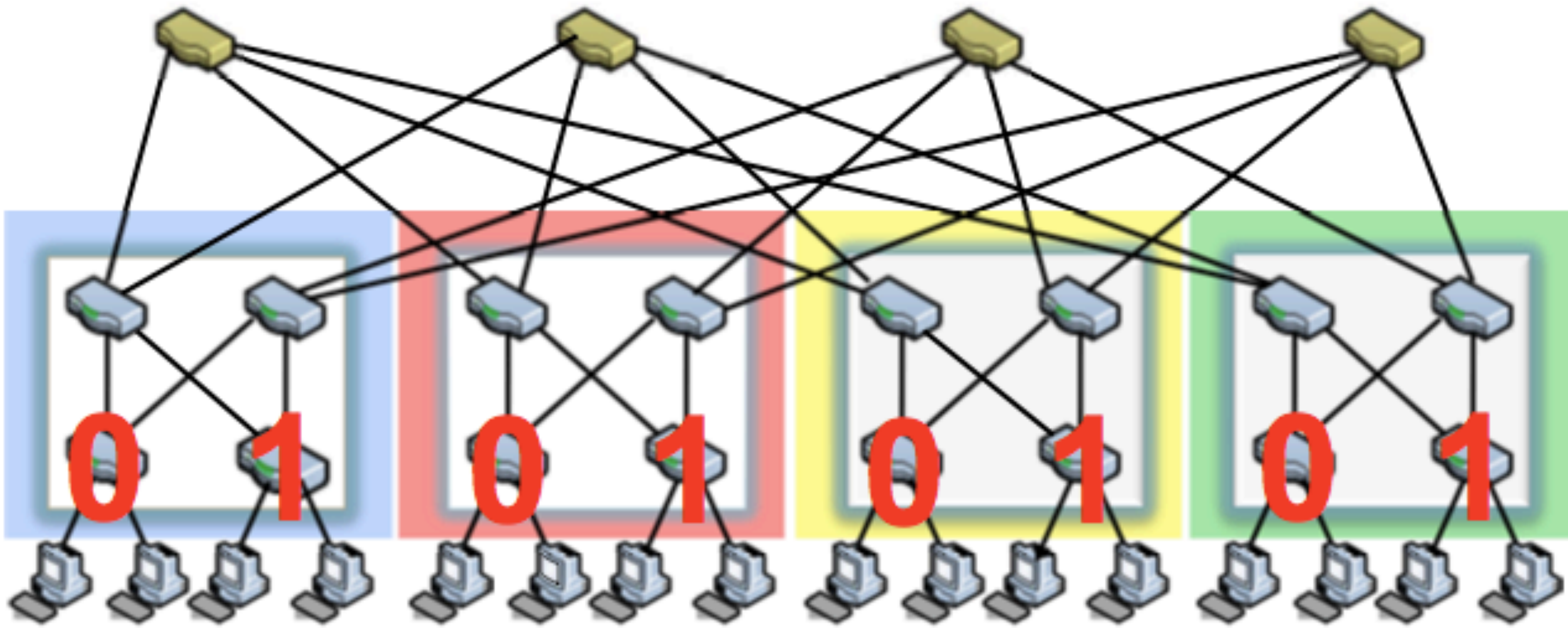
Hierarchical Addresses



Hierarchical Addresses

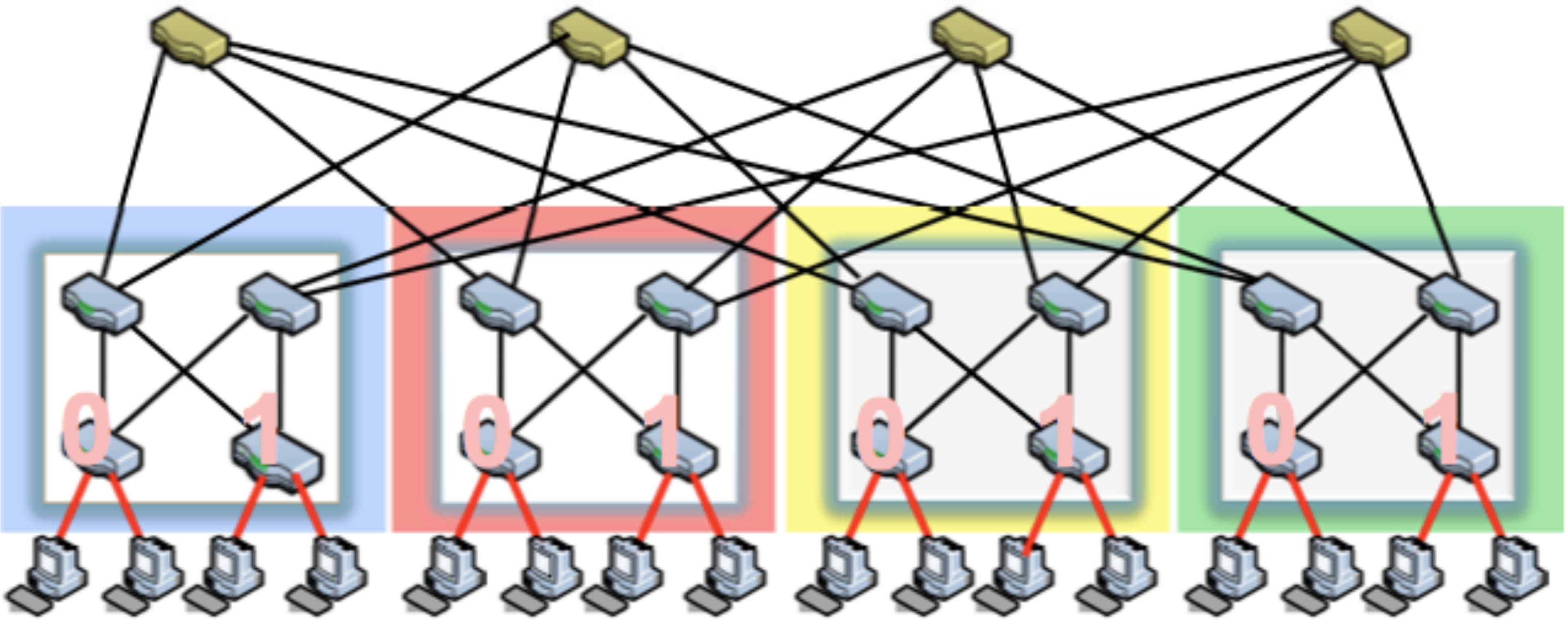


Hierarchical Addresses



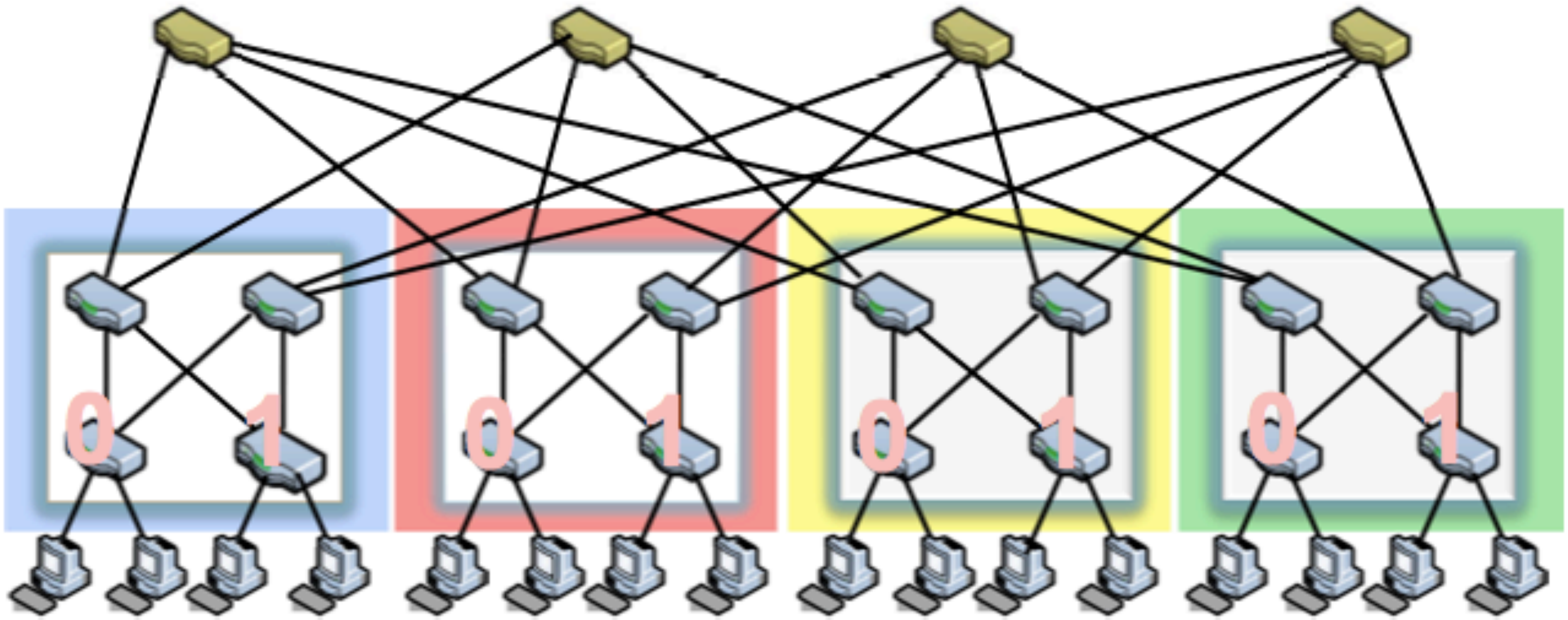
PMAC: pod.position.port.vmid

Hierarchical Addresses



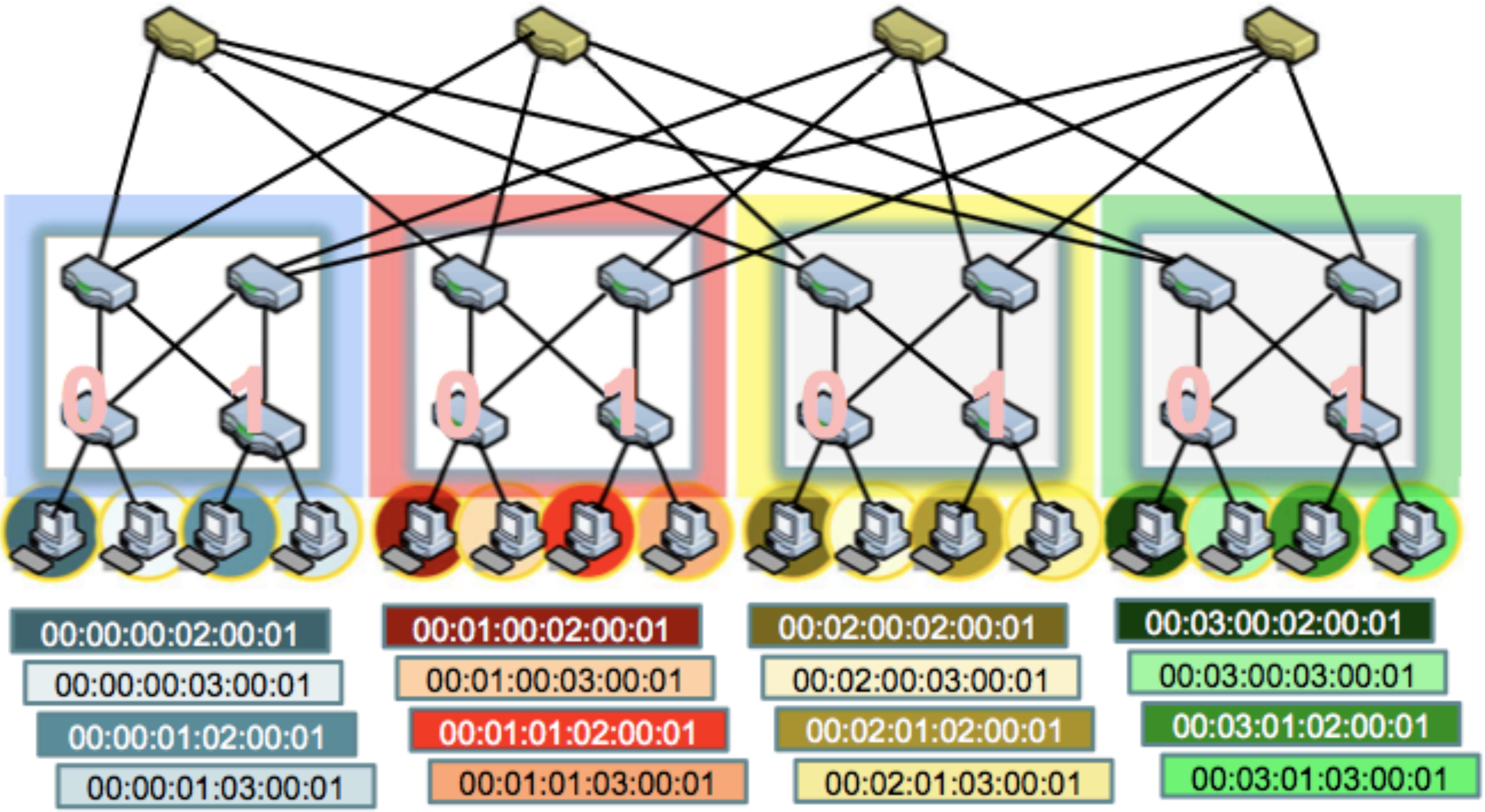
PMAC: pod.position.port.vmid

Hierarchical Addresses



PMAC: pod.position.port.vmid

Hierarchical Addresses

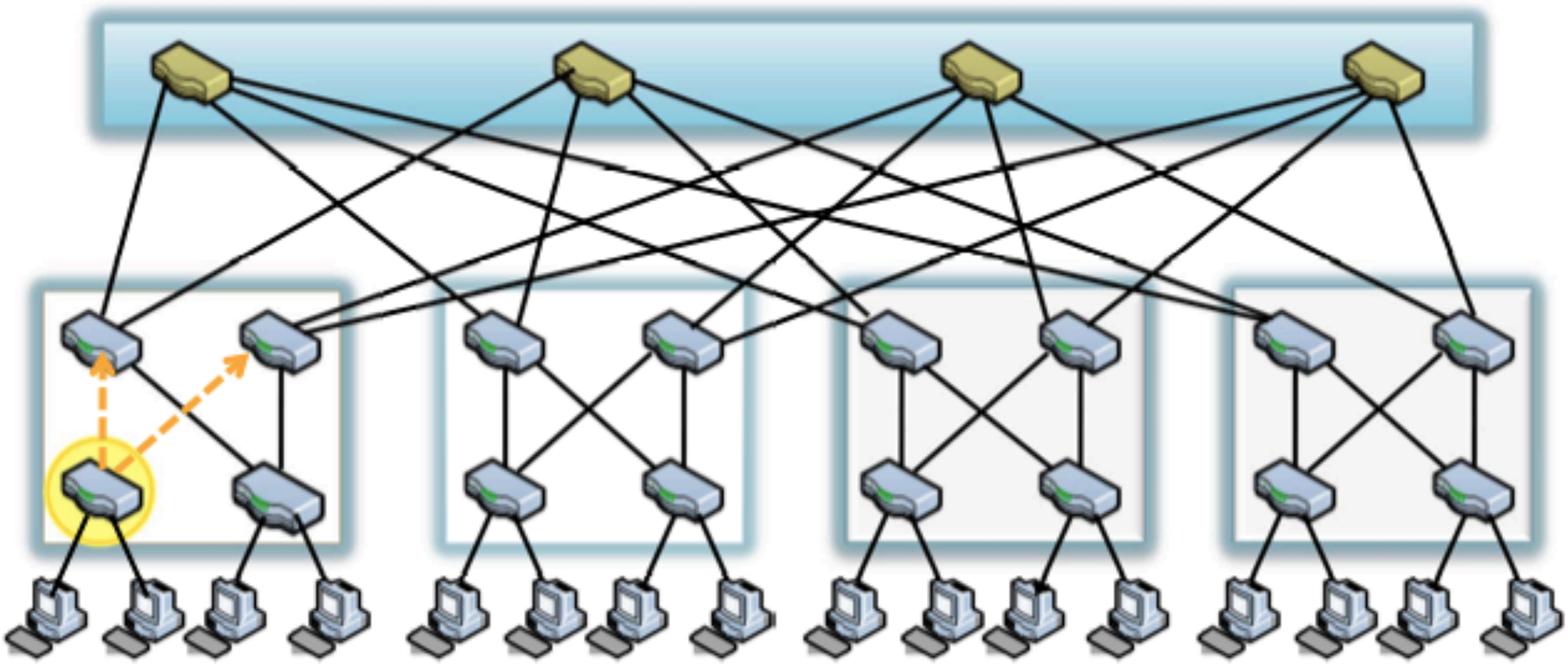


PortLand: Location Discovery Protocol

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

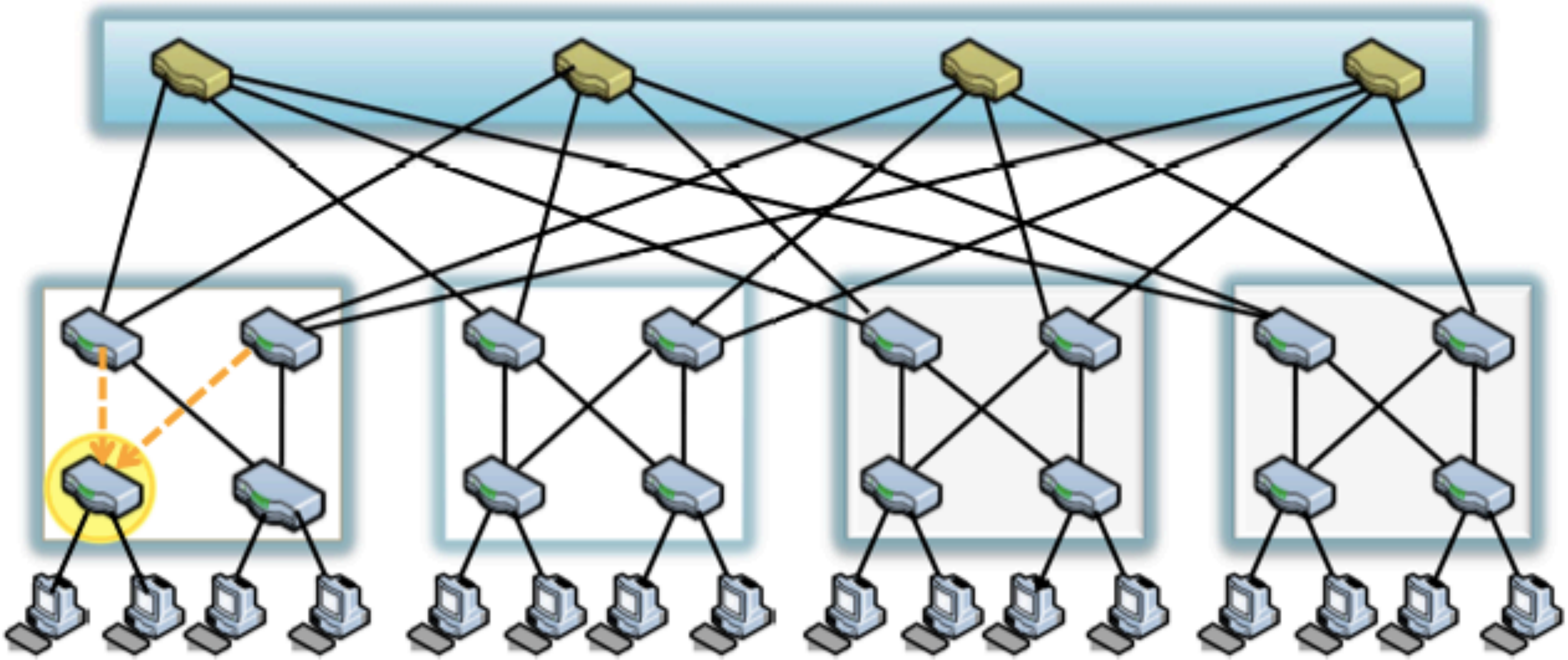
Location characteristic	Technique
1) Tree level / Role	Based on neighbor identity
2) Pod number	Aggregation and edge switches agree on pod number
3) Position number	Aggregation switches help edge switches choose unique position number

Location Discovery Protocol



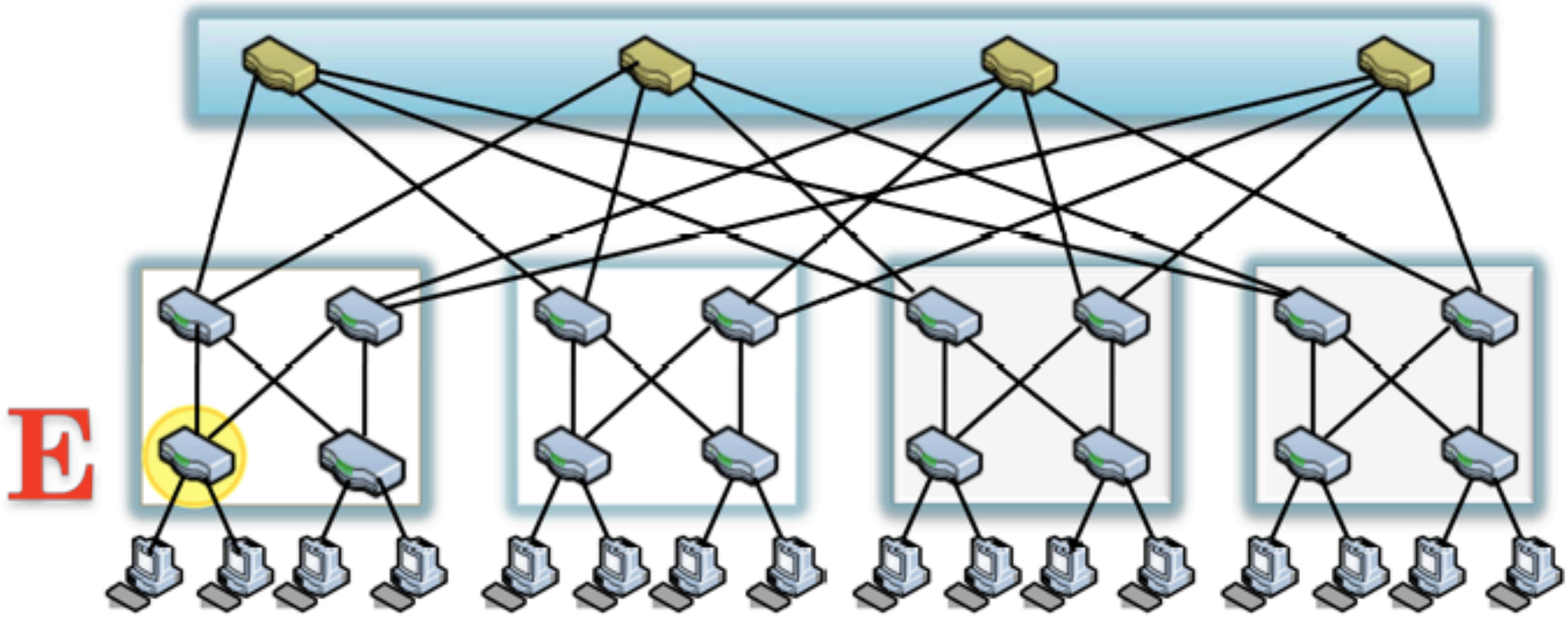
Switch Identifier	Pod Number	Position	Tree Level
A0:B1:FD:56:32:01	??	??	??

Location Discovery Protocol



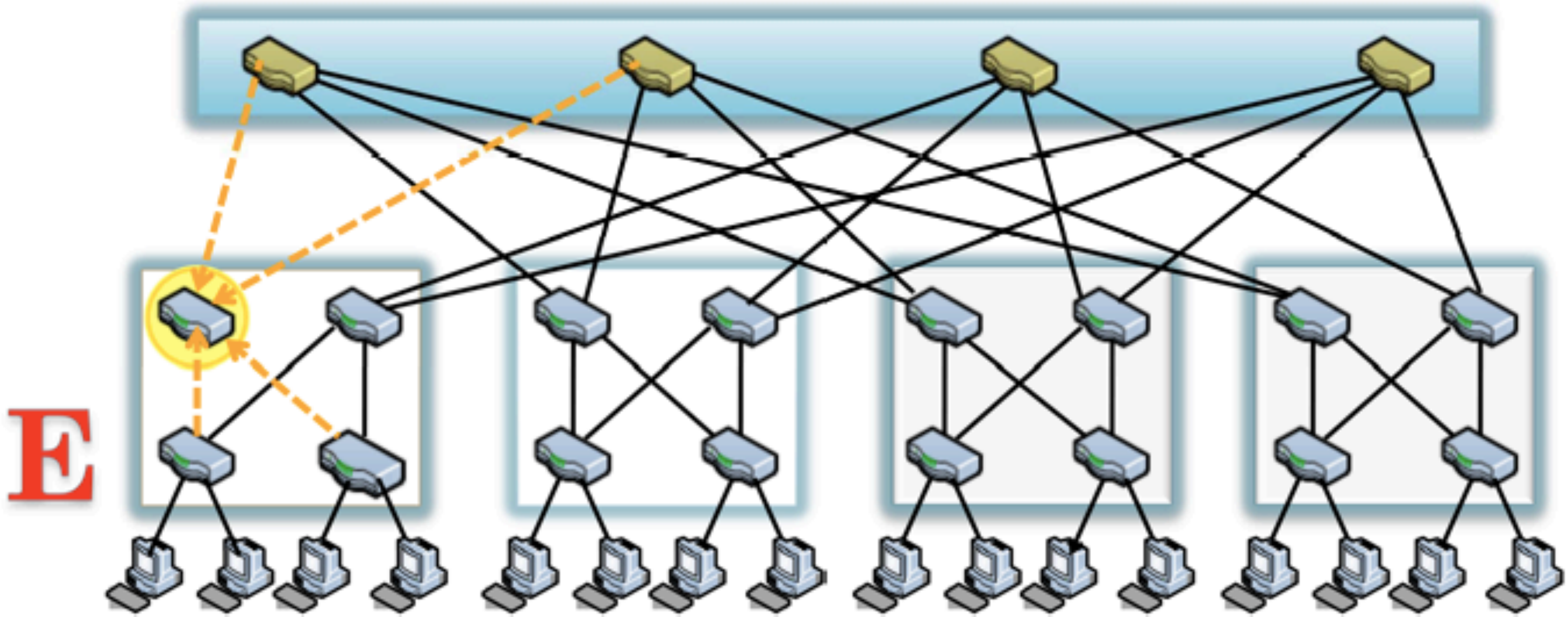
Switch Identifier	Pod Number	Position	Tree Level
A0:B1:FD:56:32:01	??	??	??

Location Discovery Protocol



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

Location Discovery Protocol

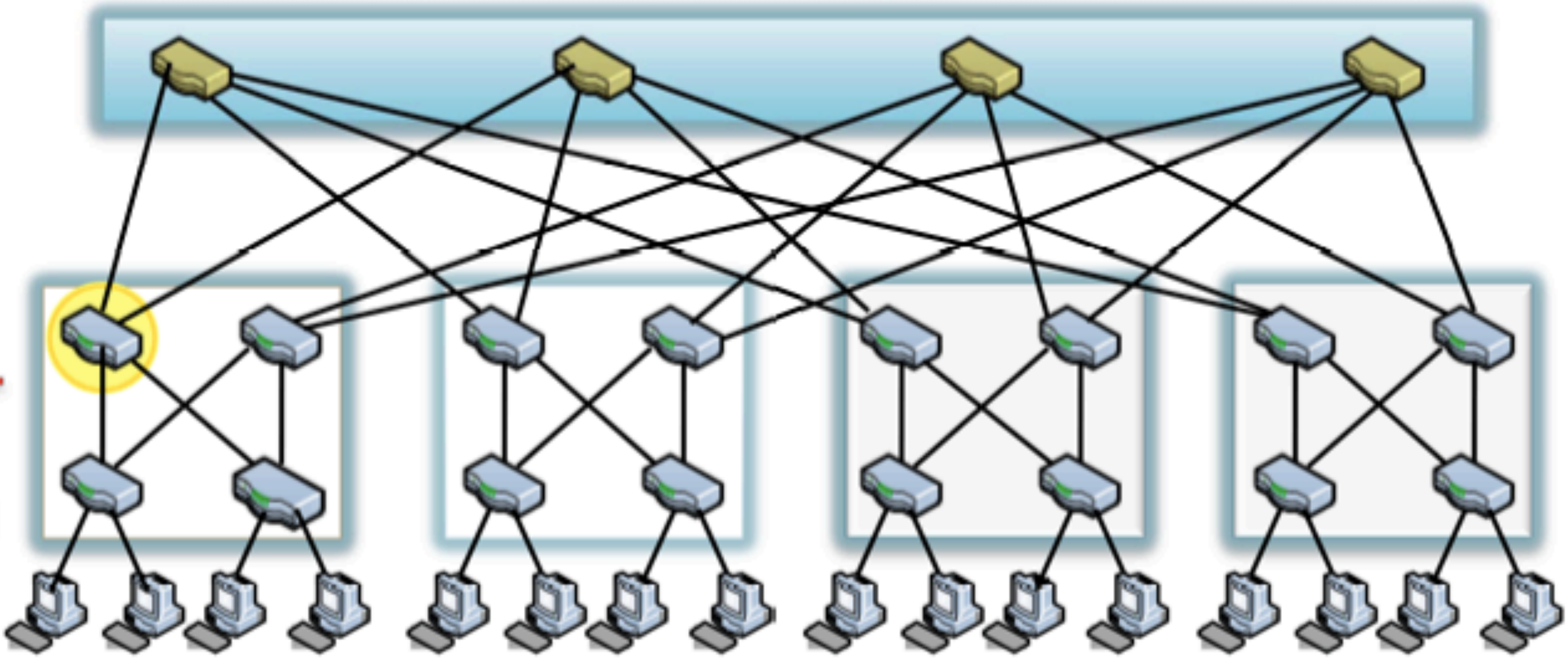


Switch Identifier	Pod Number	Position	Tree Level
B0:A1:FD:57:32:01	??	??	??

Location Discovery Protocol

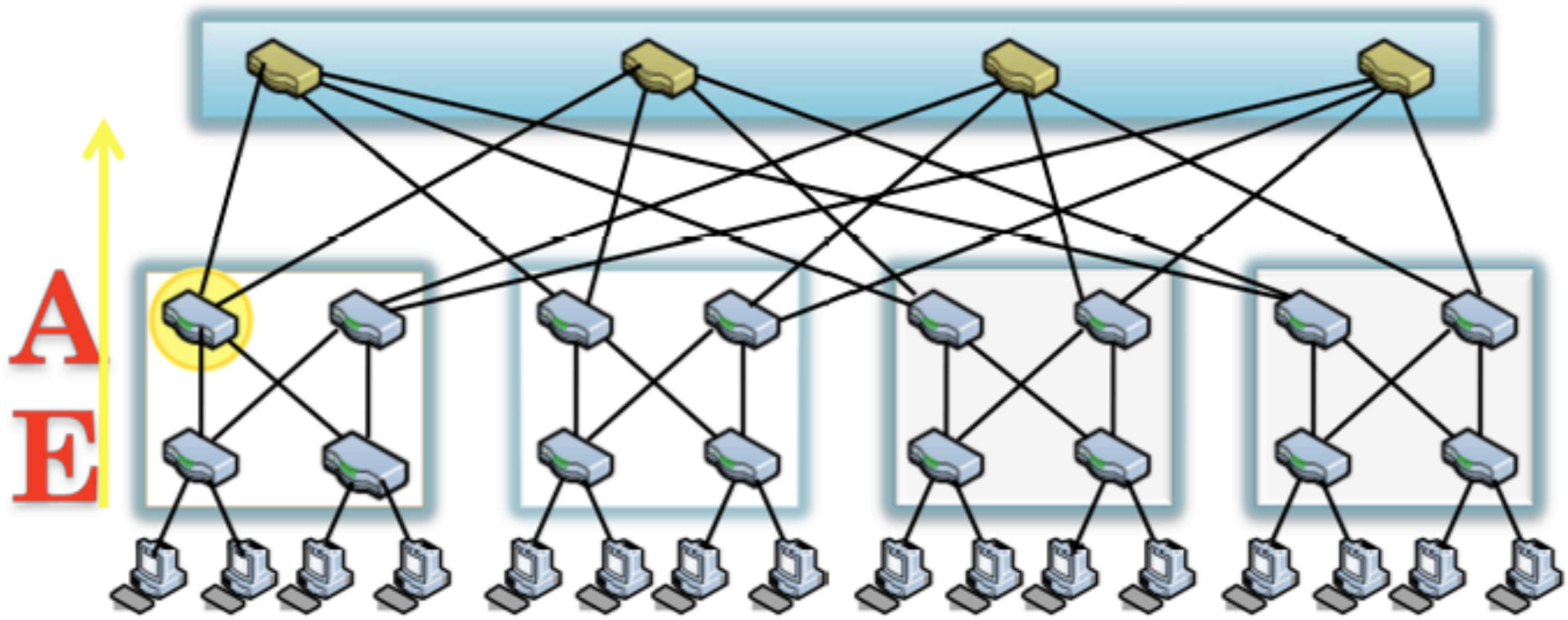
A

E



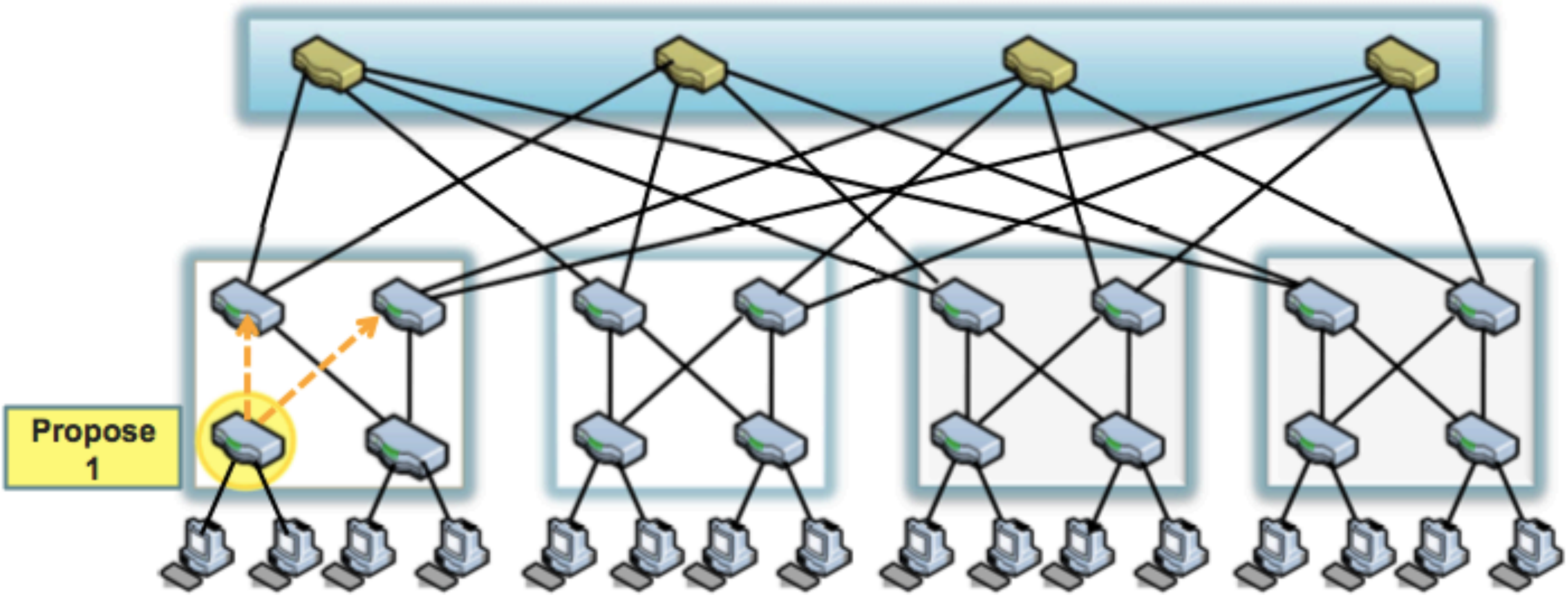
Switch Identifier	Pod Number	Position	Tree Level
B0:A1:FD:57:32:01	??	??	1

Location Discovery Protocol



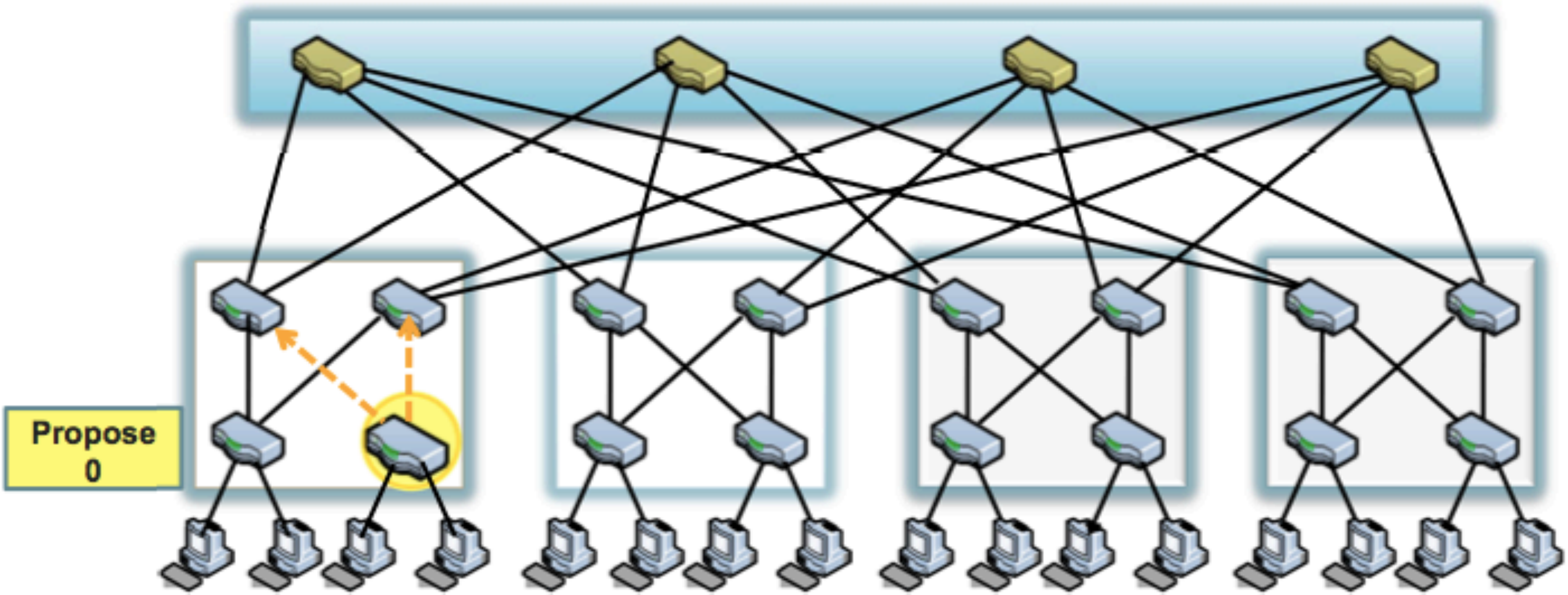
Switch Identifier	Pod Number	Position	Tree Level
B0:A1:FD:57:32:01	??	??	1

Location Discovery Protocol



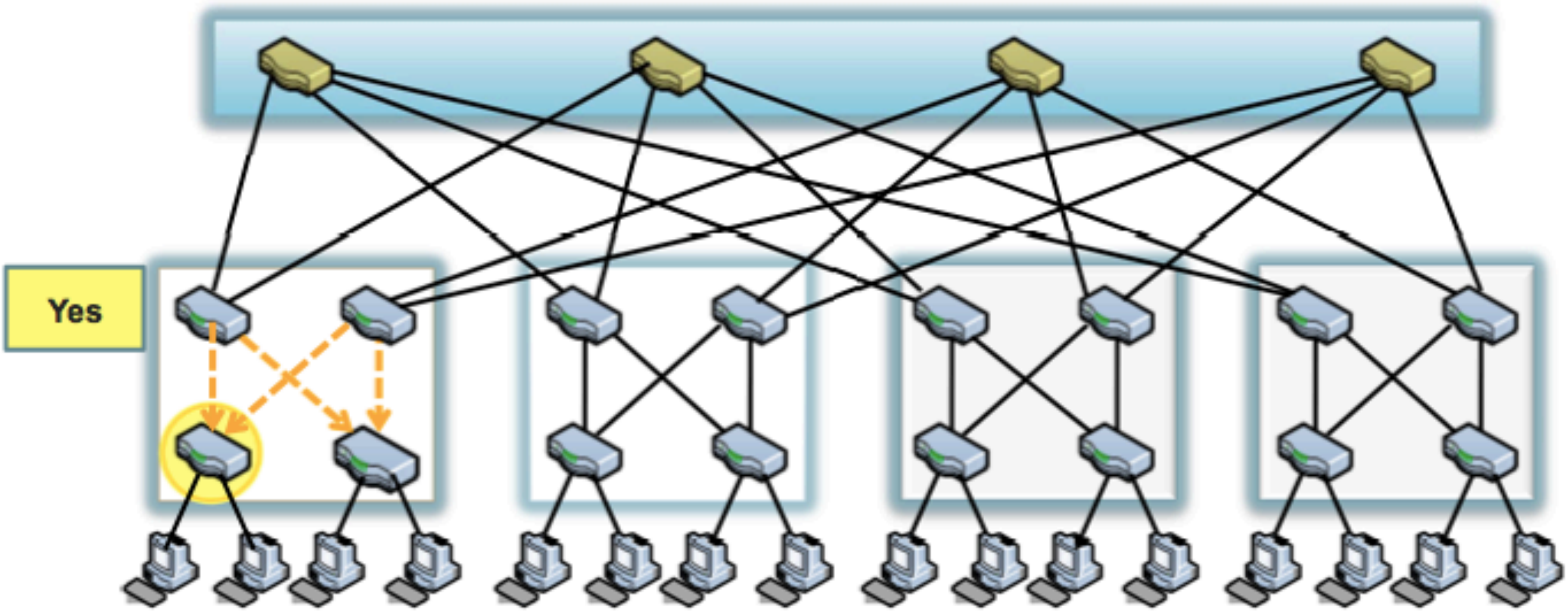
Switch Identifier	Pod Number	Position	Tree Level
A0:B1:FD:56:32:01	??	??	0

Location Discovery Protocol



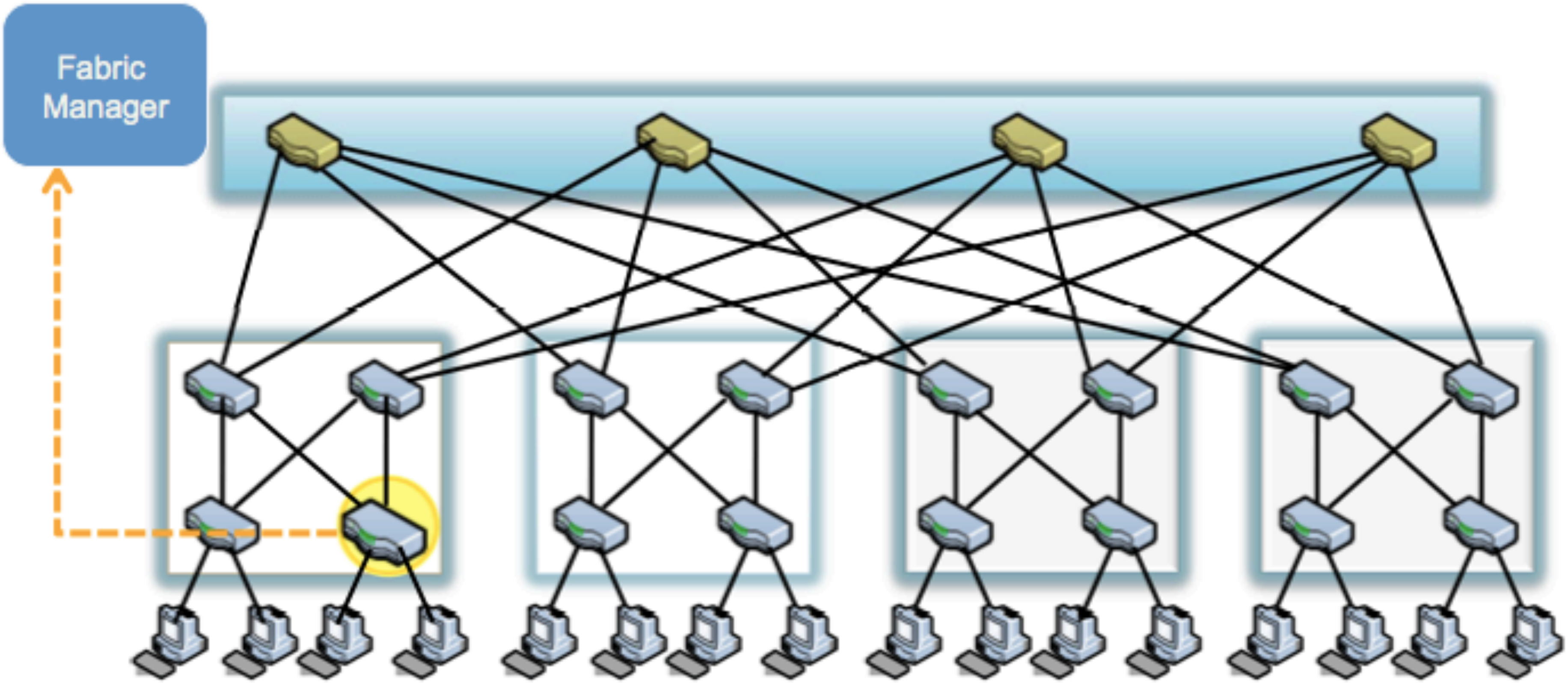
Switch Identifier	Pod Number	Position	Tree Level
D0:B1:AD:56:32:01	??	??	0

Location Discovery Protocol



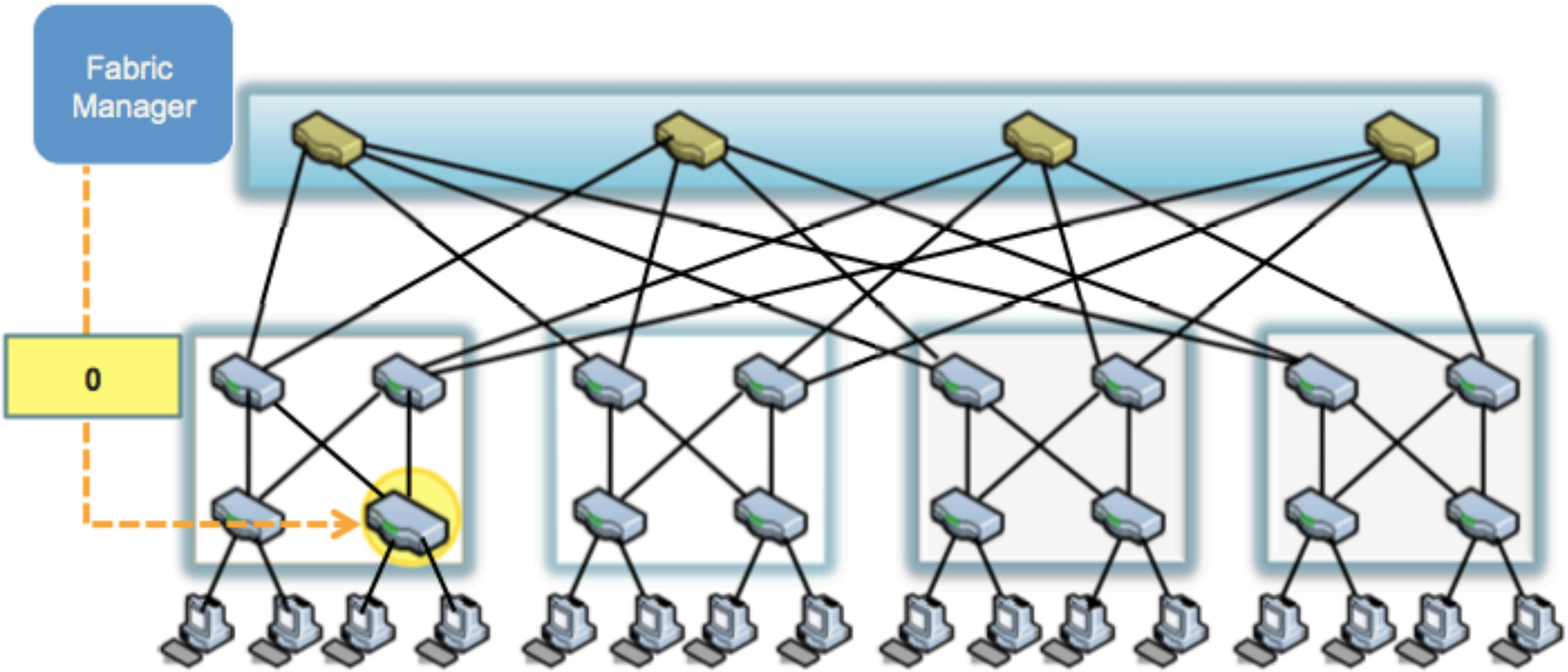
Switch Identifier	Pod Number	Position	Tree Level
A0:B1:FD:56:32:01	??	1	0

Location Discovery Protocol



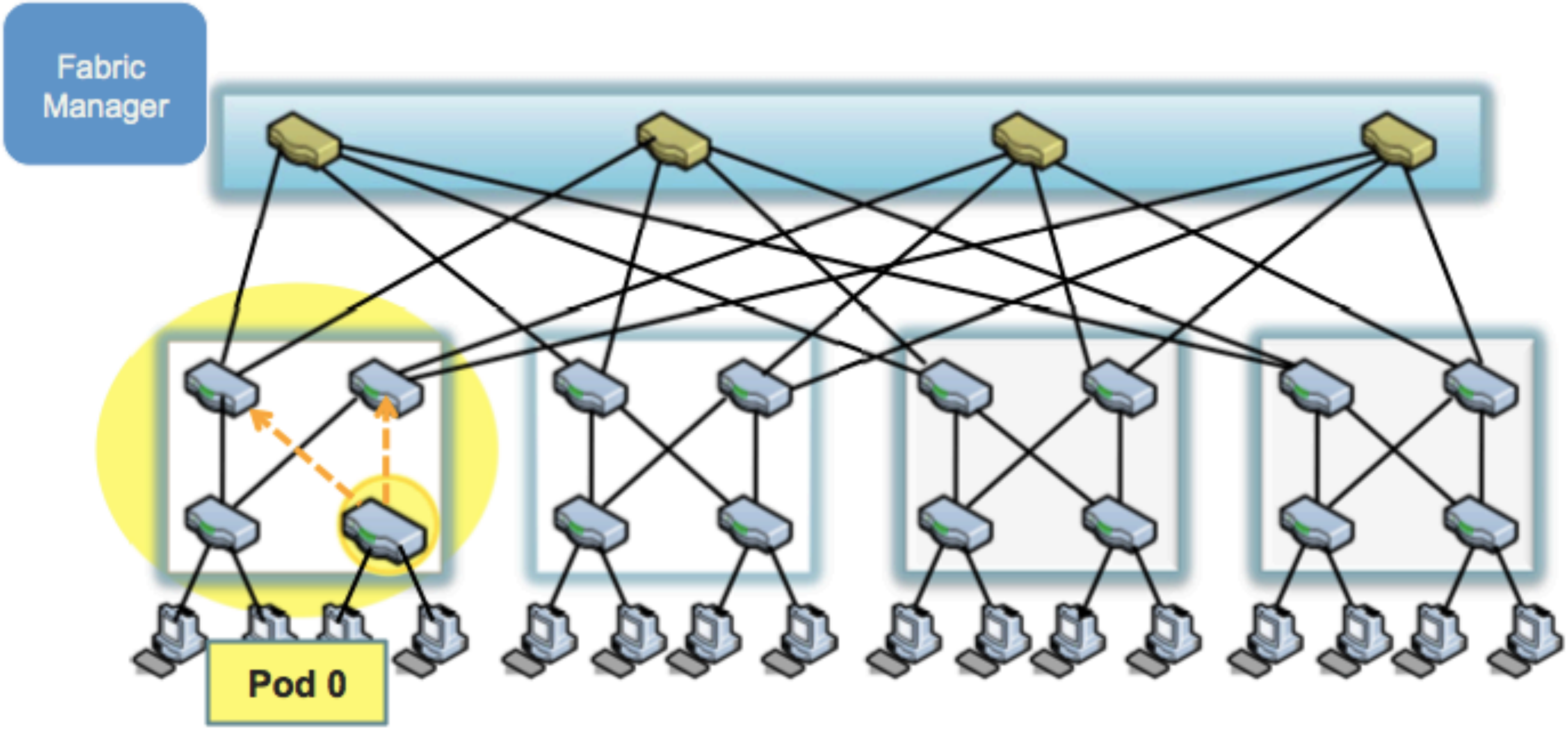
Switch Identifier	Pod Number	Position	Tree Level
D0:B1:AD:56:32:01	??	0	0

Location Discovery Protocol



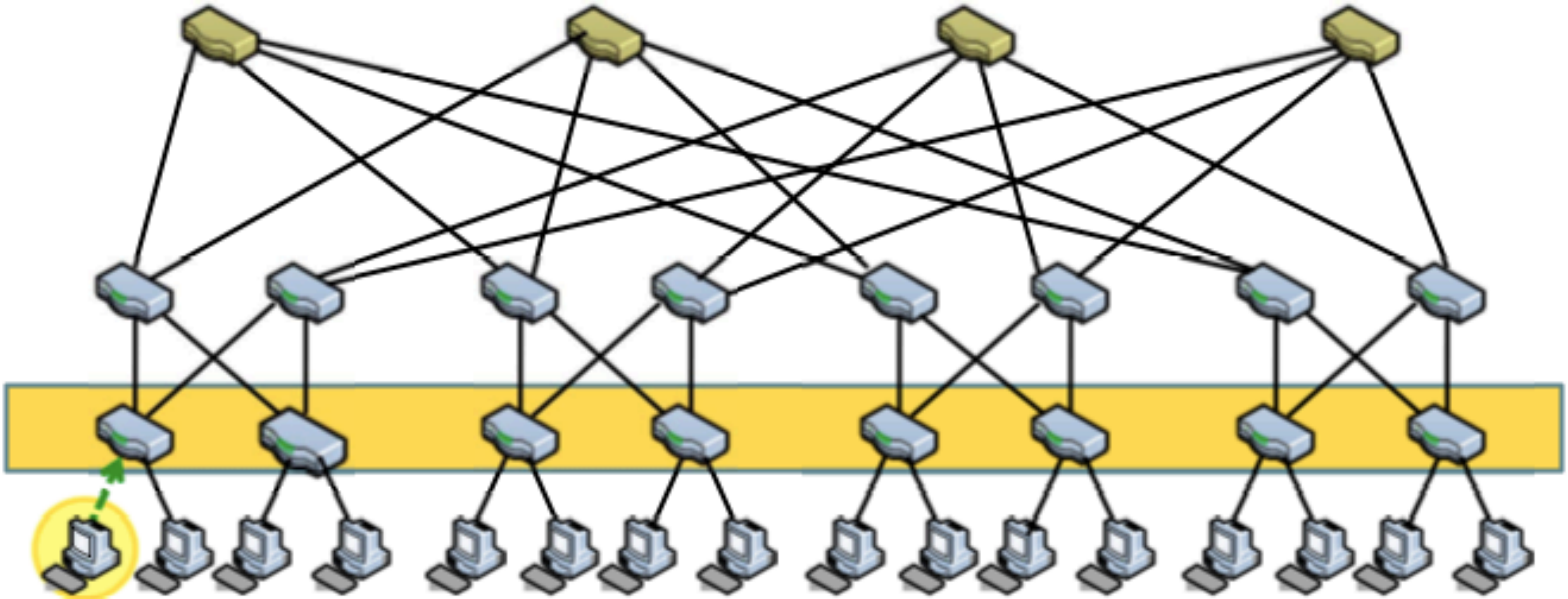
Switch Identifier	Pod Number	Position	Tree Level
D0:B1:AD:56:32:01	??	0	0

Location Discovery Protocol



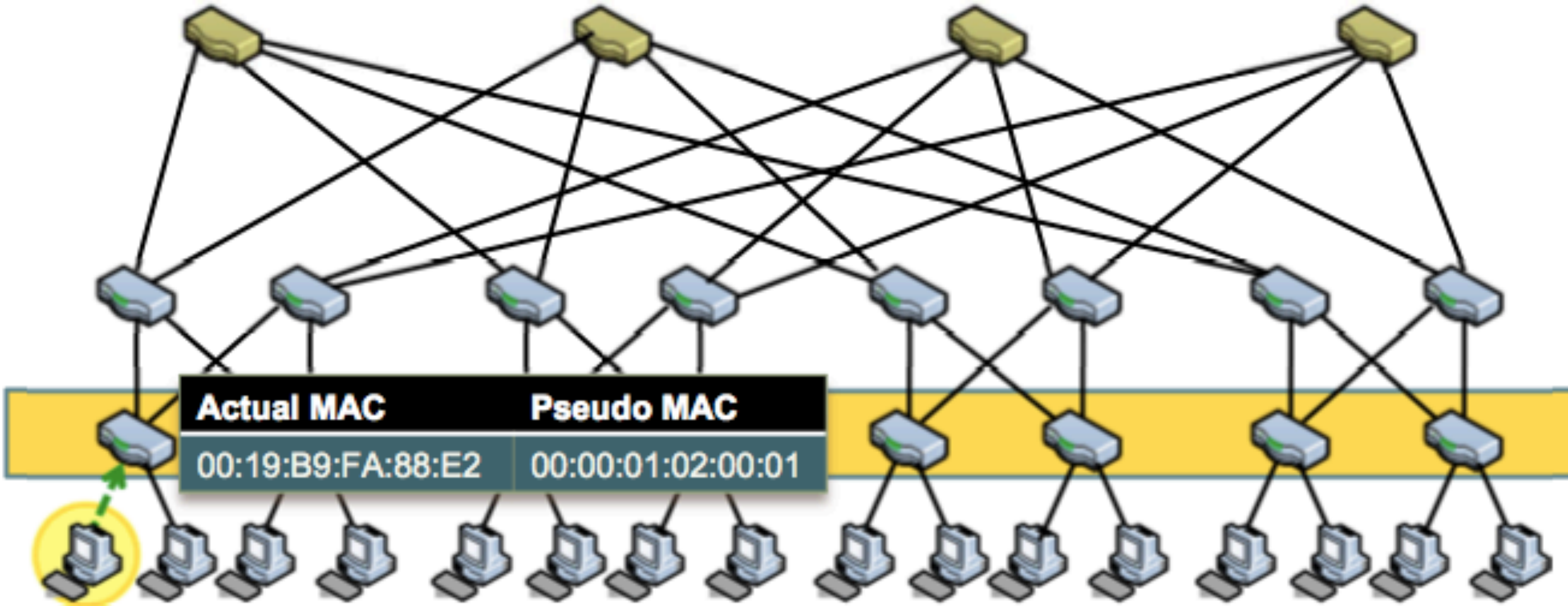
Switch Identifier	Pod Number	Position	Tree Level
D0:B1:AD:56:32:01	0	0	0

Name Resolution



Intercept all ARP packets

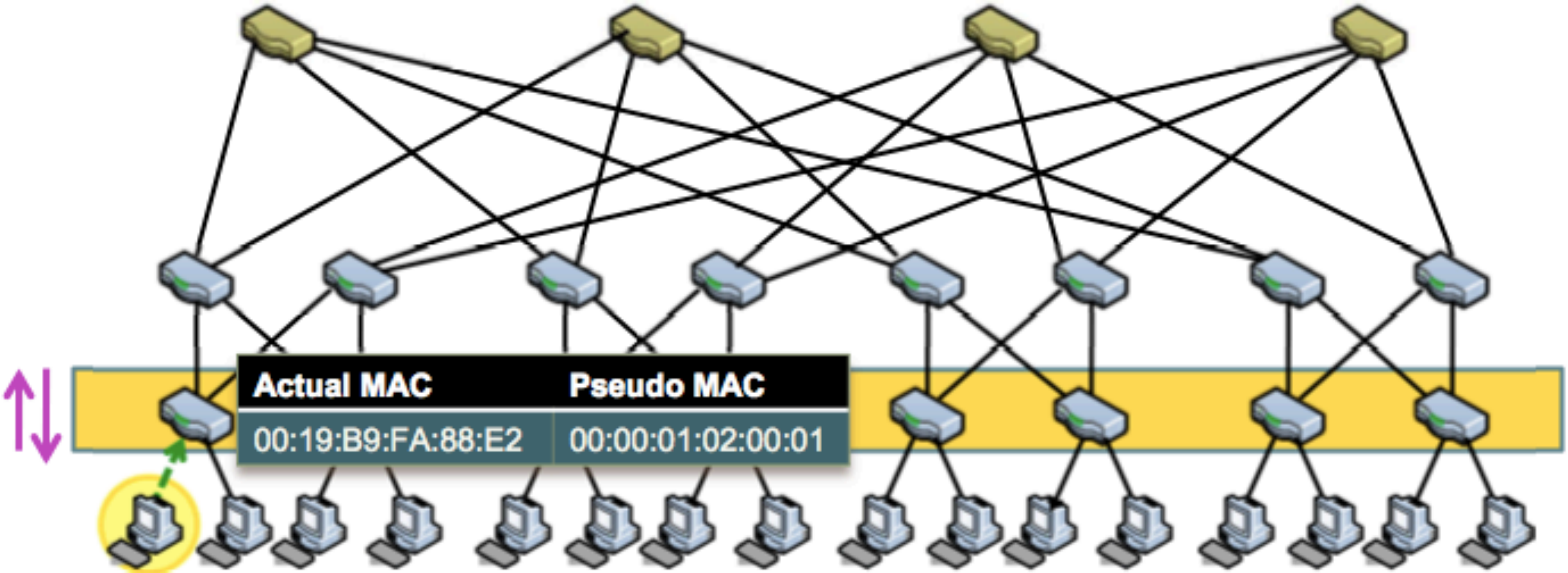
Name Resolution



Intercept all ARP packets

Assign new end hosts with PMACs

Name Resolution

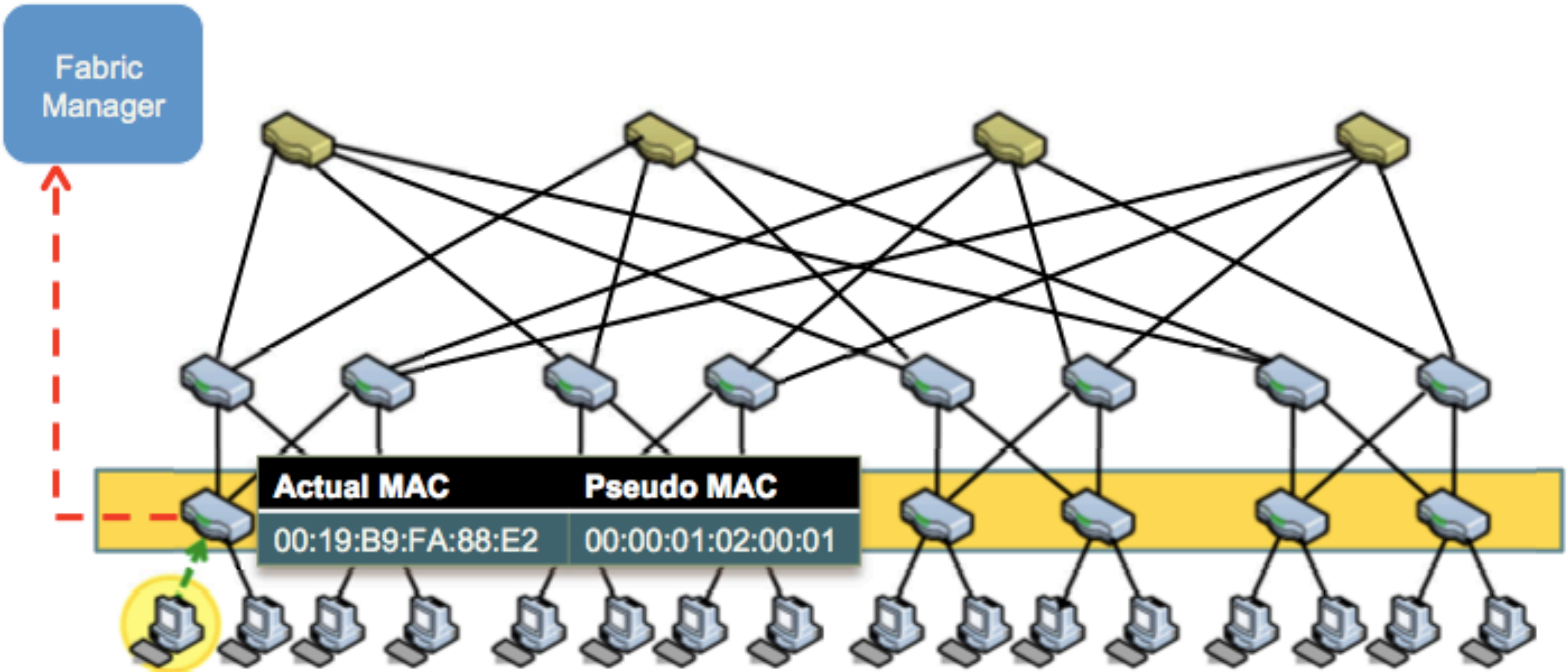


Intercept all ARP packets

Assign new end hosts with PMACs

Rewrite MAC for packets entering and exiting network

Name Resolution

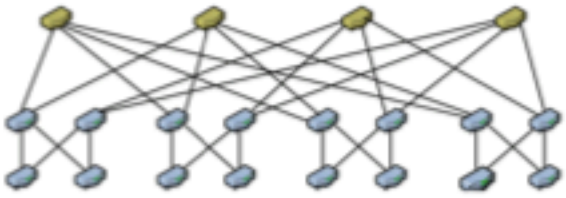


Fabric Manager

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

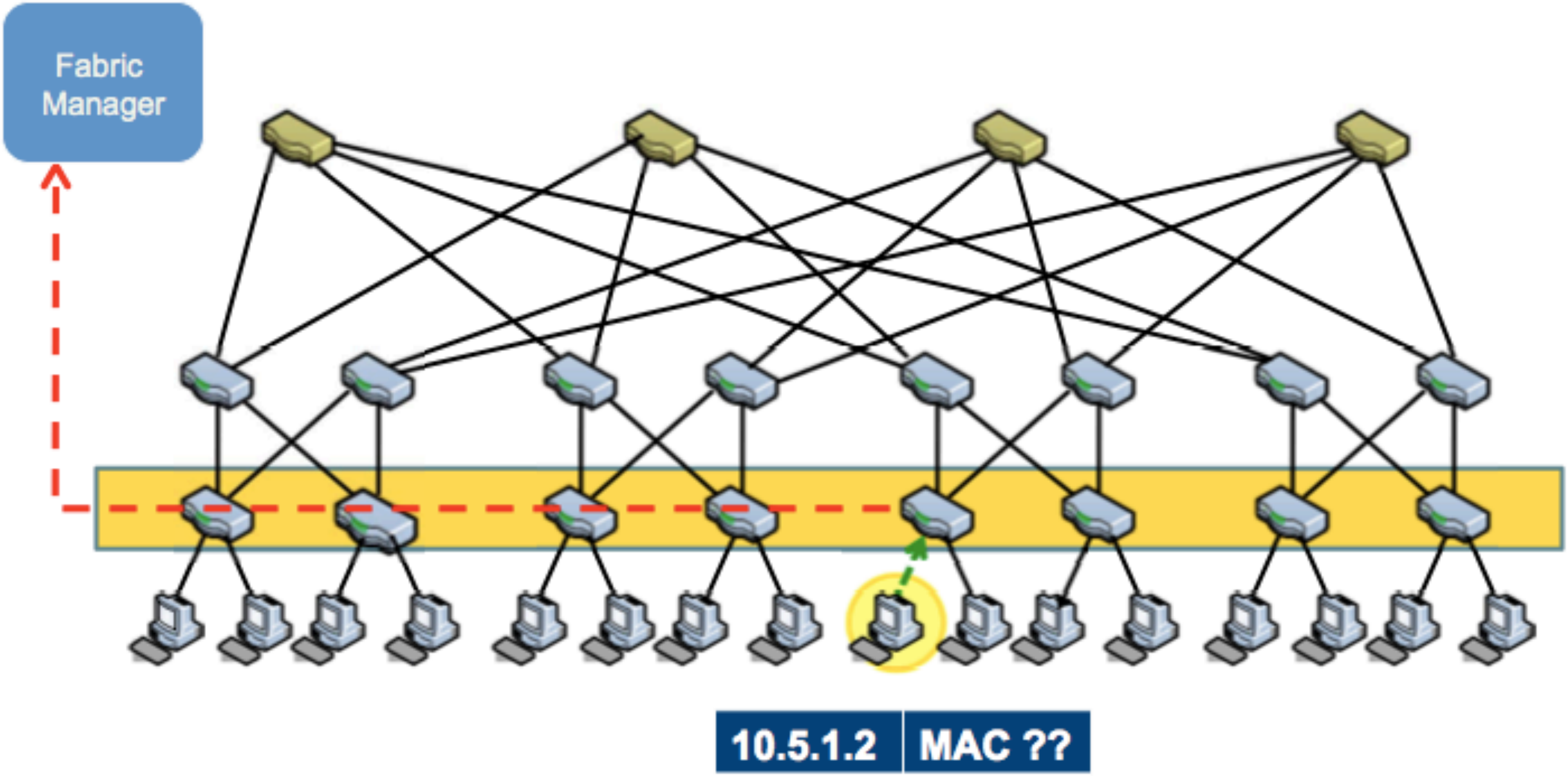


Network map

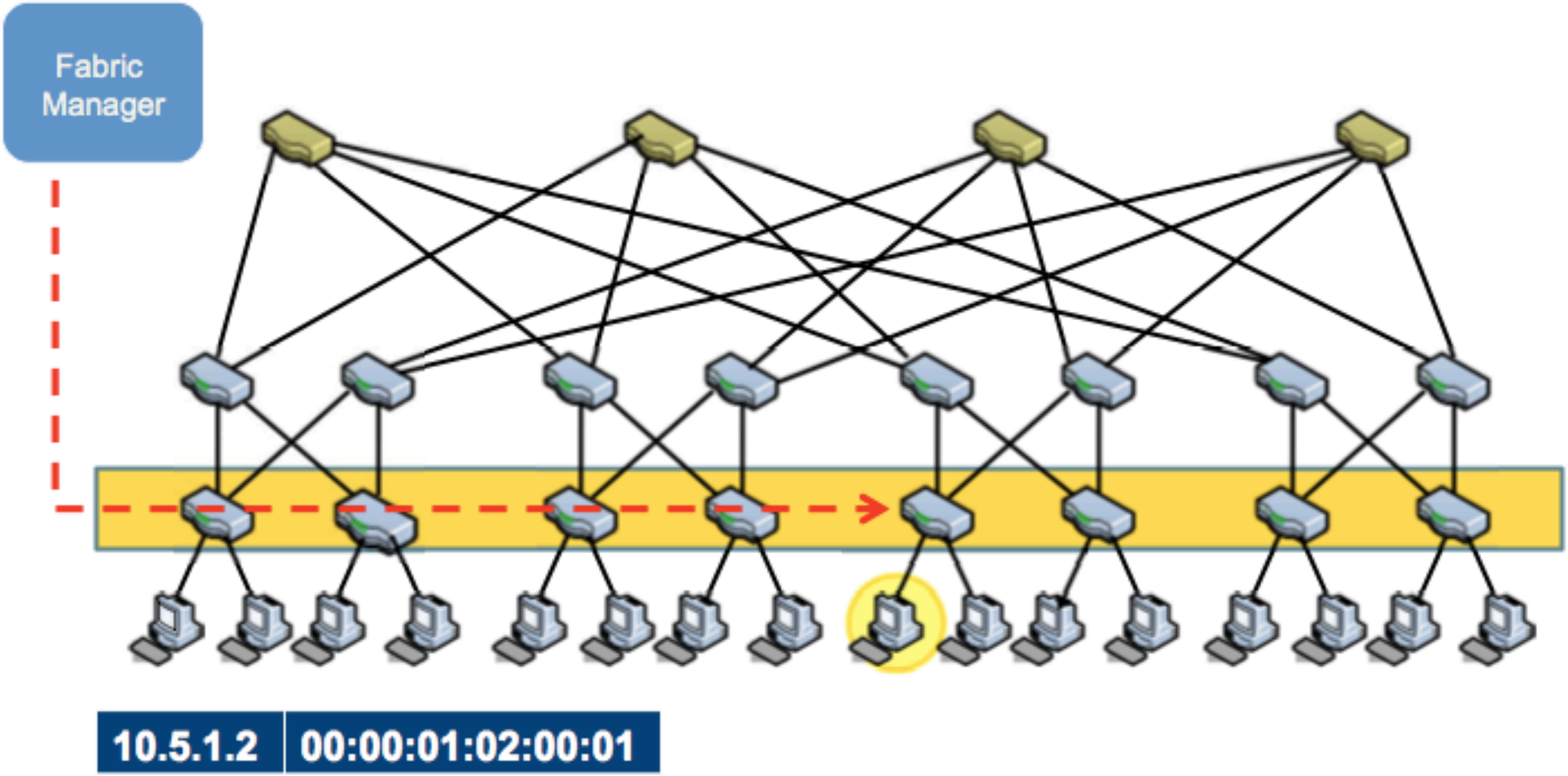
Soft state

~~Administrator configuration~~

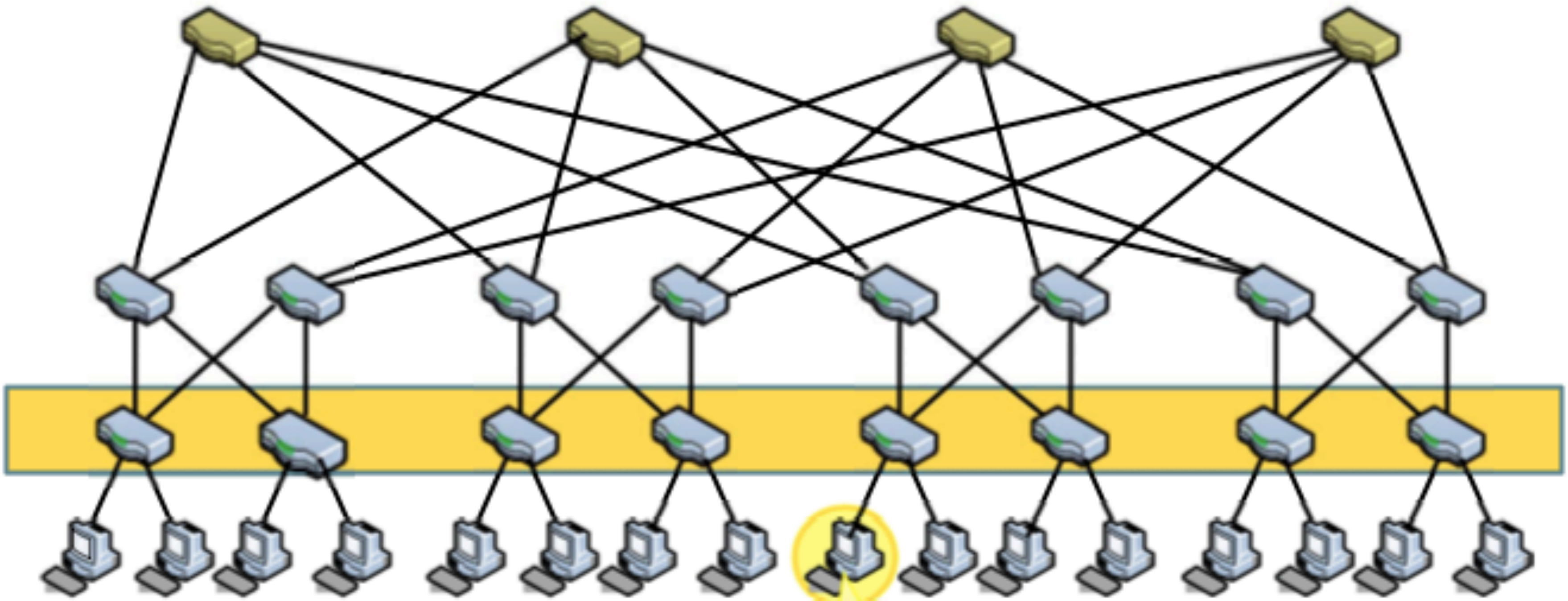
Name Resolution



Name Resolution



Name Resolution



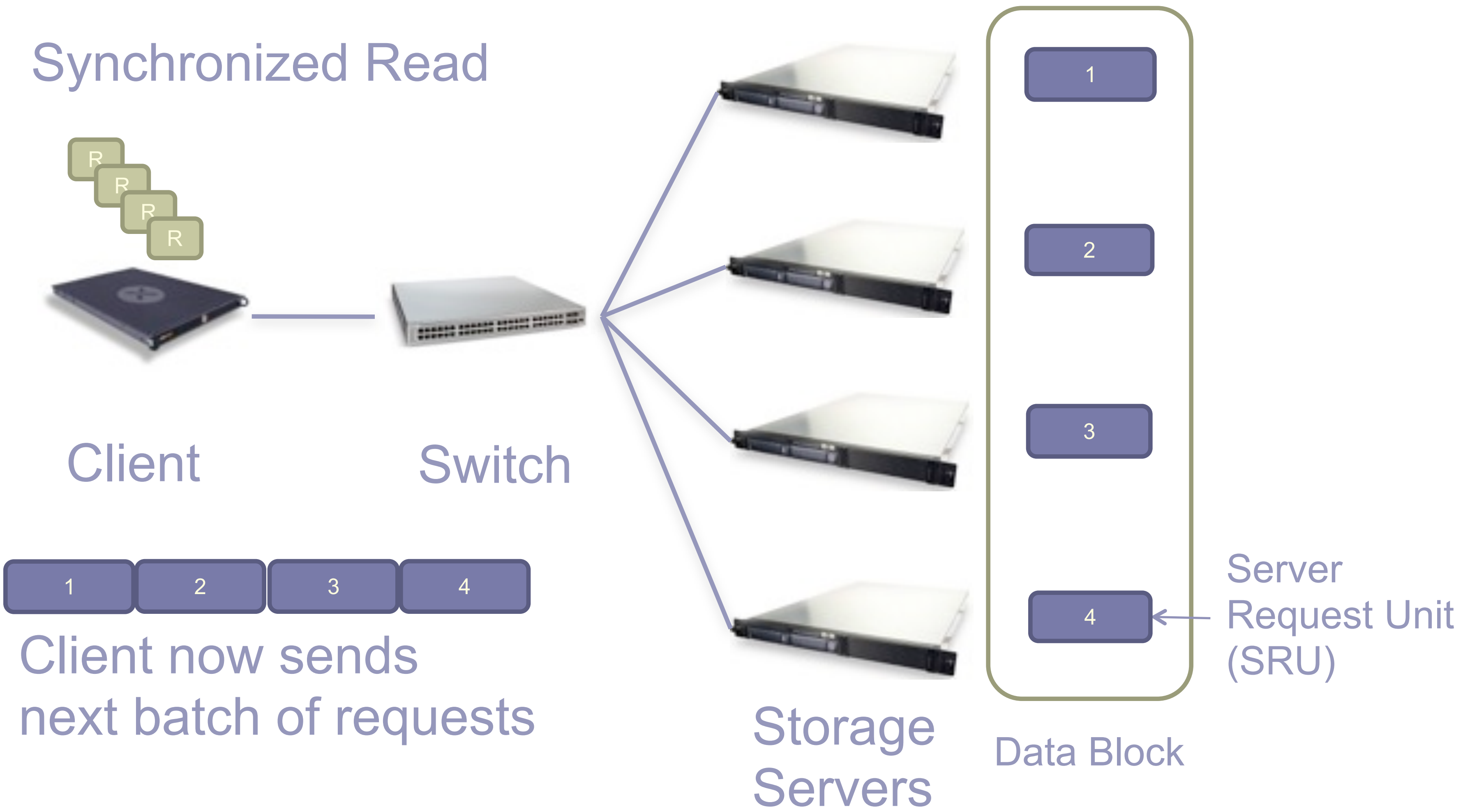
ARP replies contain only PMAC

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

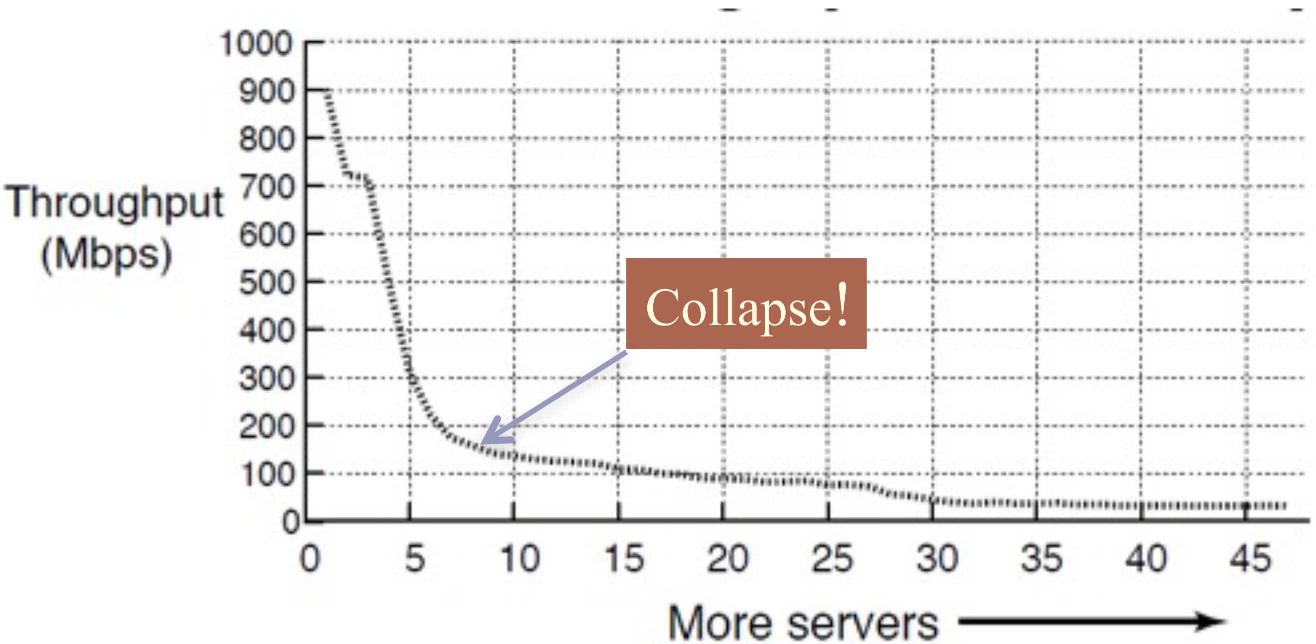
Overview

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

Cluster-based Storage Systems



TCP Throughput Collapse



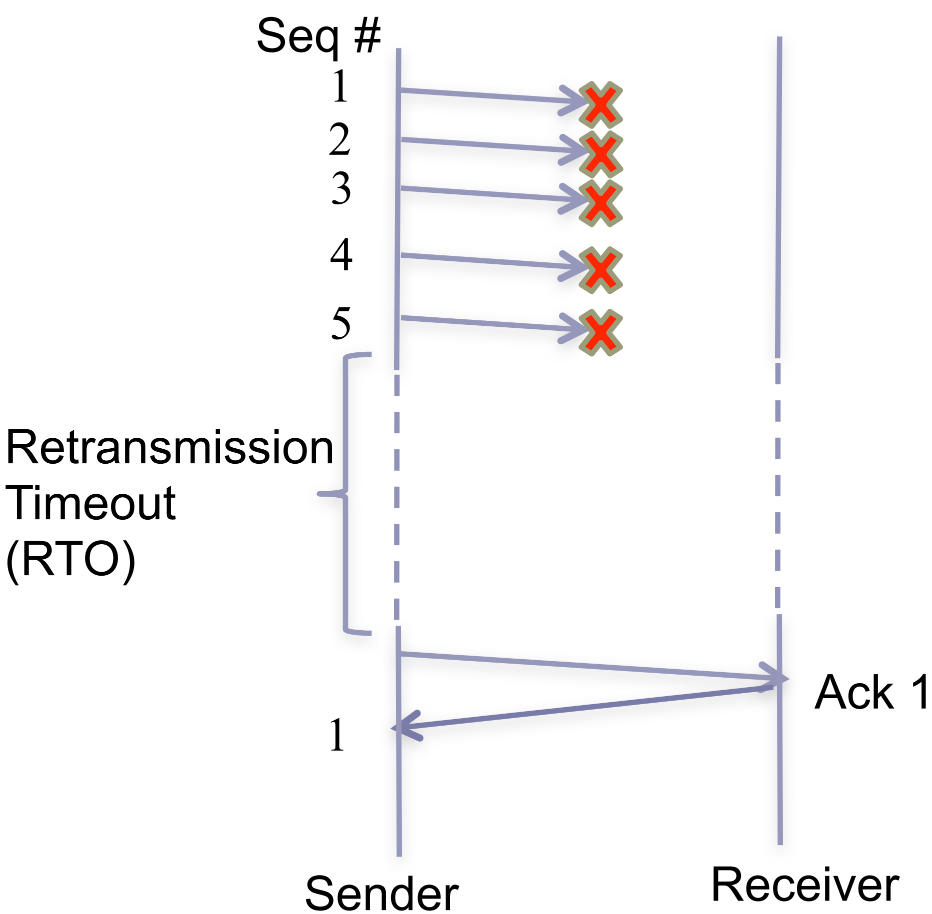
Cluster Setup
1Gbps Ethernet
Unmodified TCP
S50 Switch
1MB Block Size

- TCP *Incast*
  - Cause of throughput collapse:

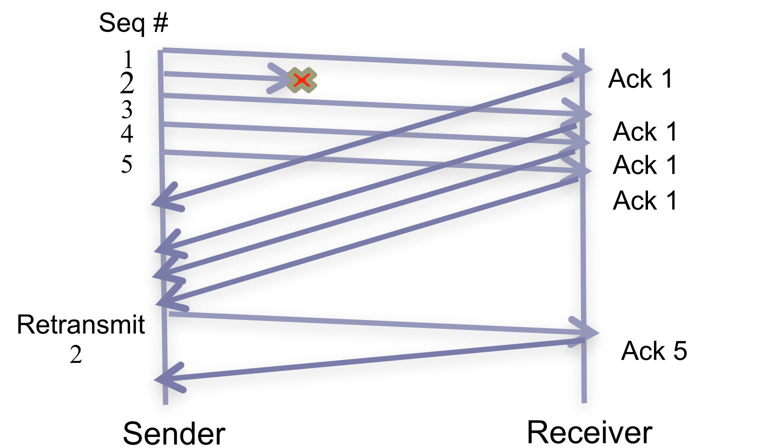
    coarse-grained TCP timeouts

TCP: Loss recovery comparison

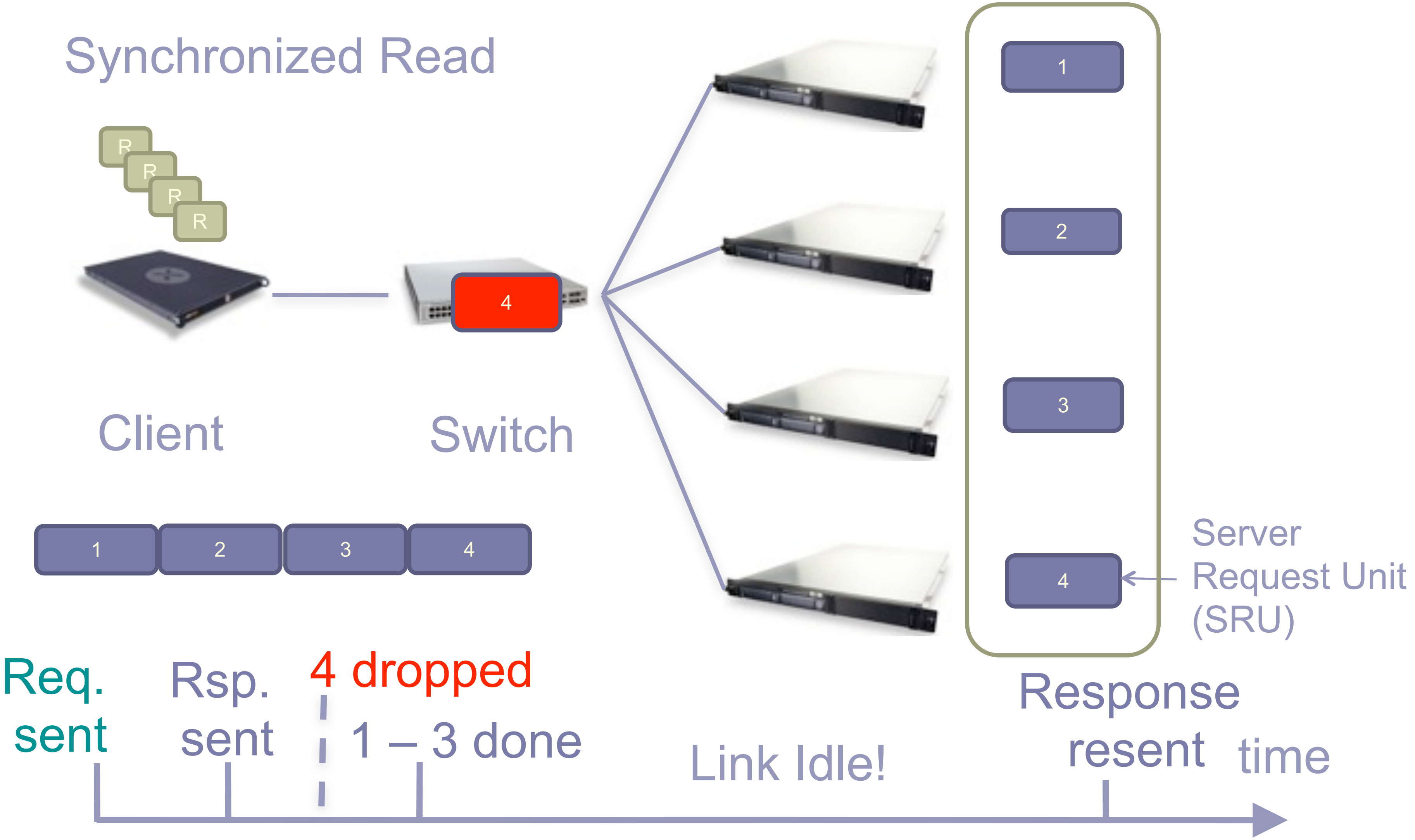
Timeout driven recovery is slow (**ms**)



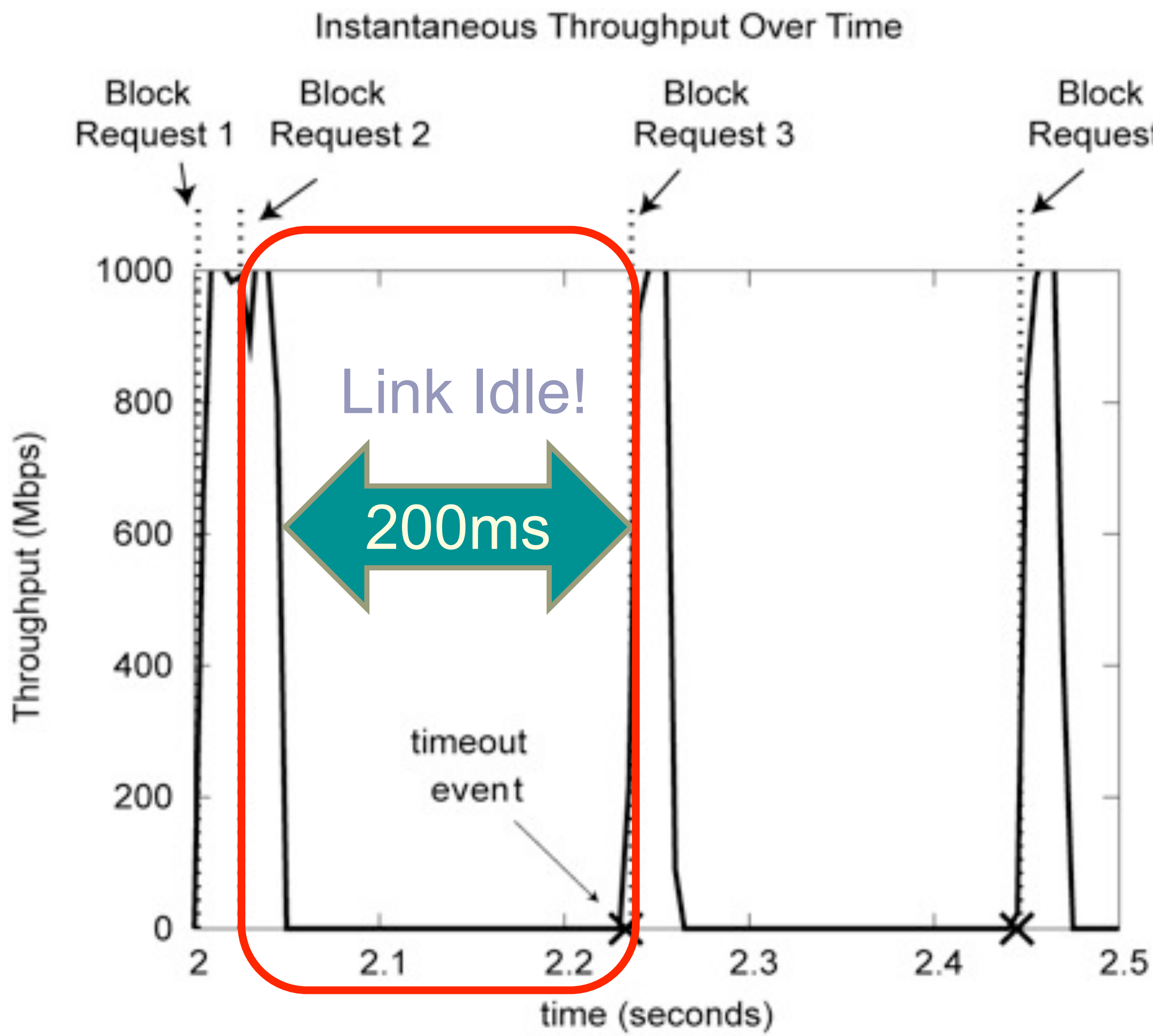
Data-driven recovery is super fast (**µs**) in datacenters



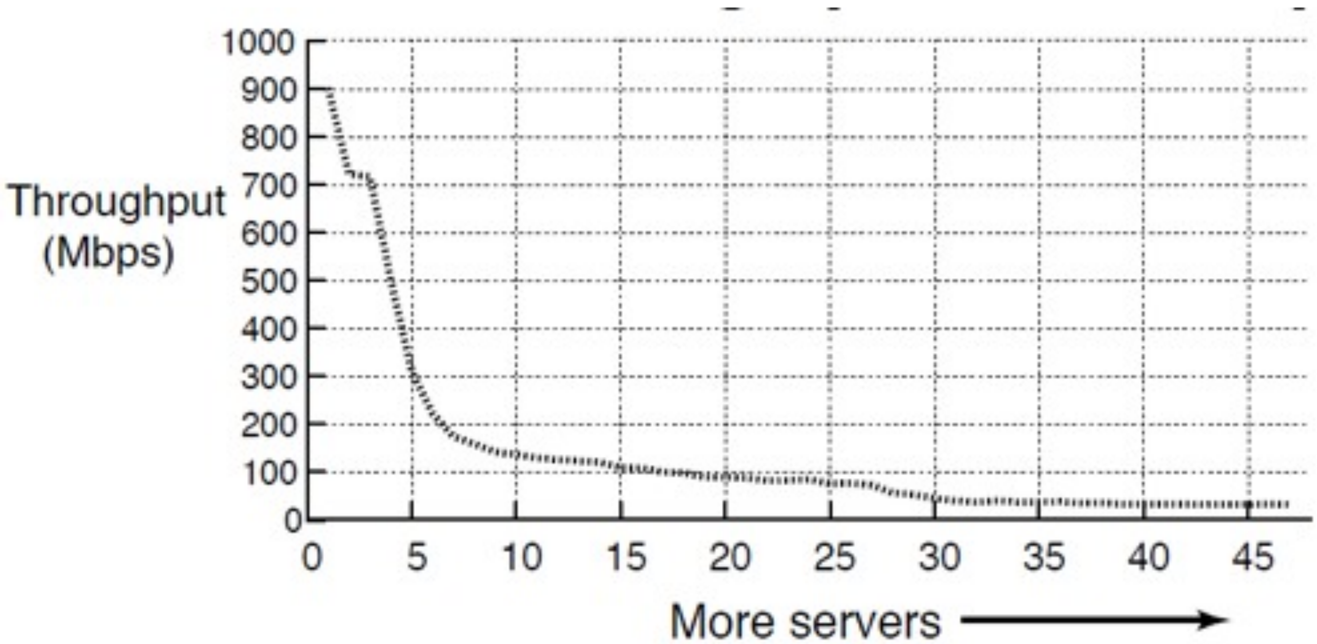
Link Idle Time Due To Timeouts



Client Link Utilization

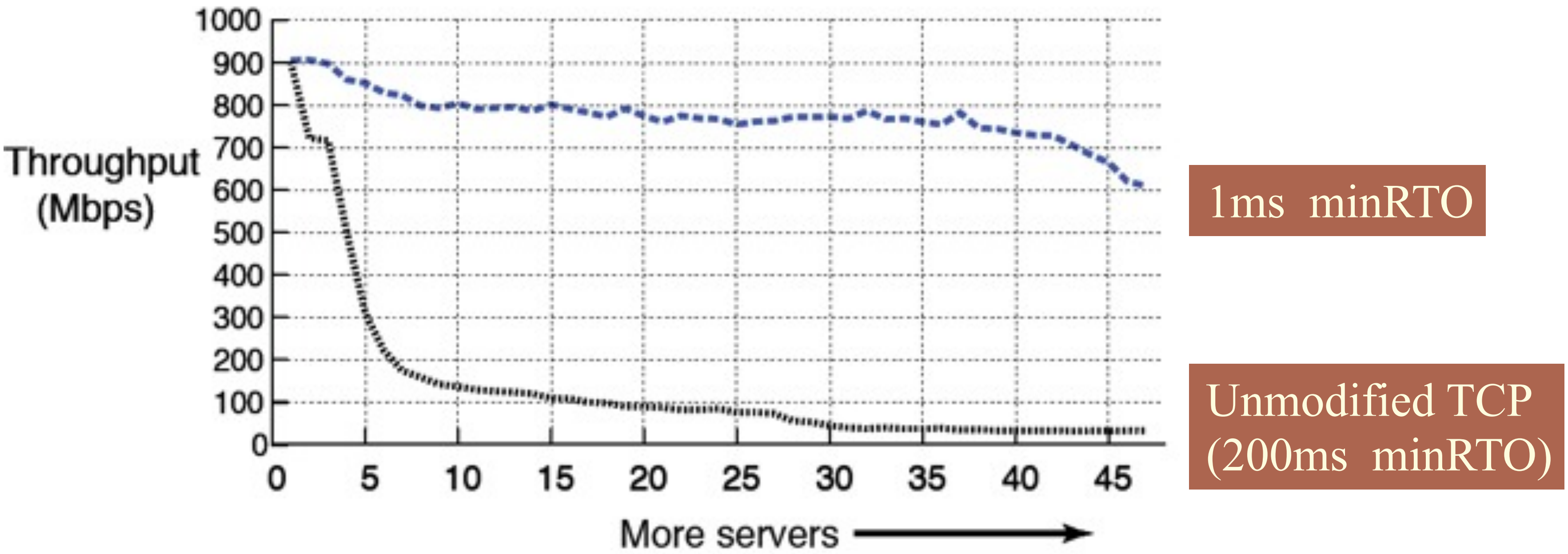


Default minRTO: Throughput Collapse



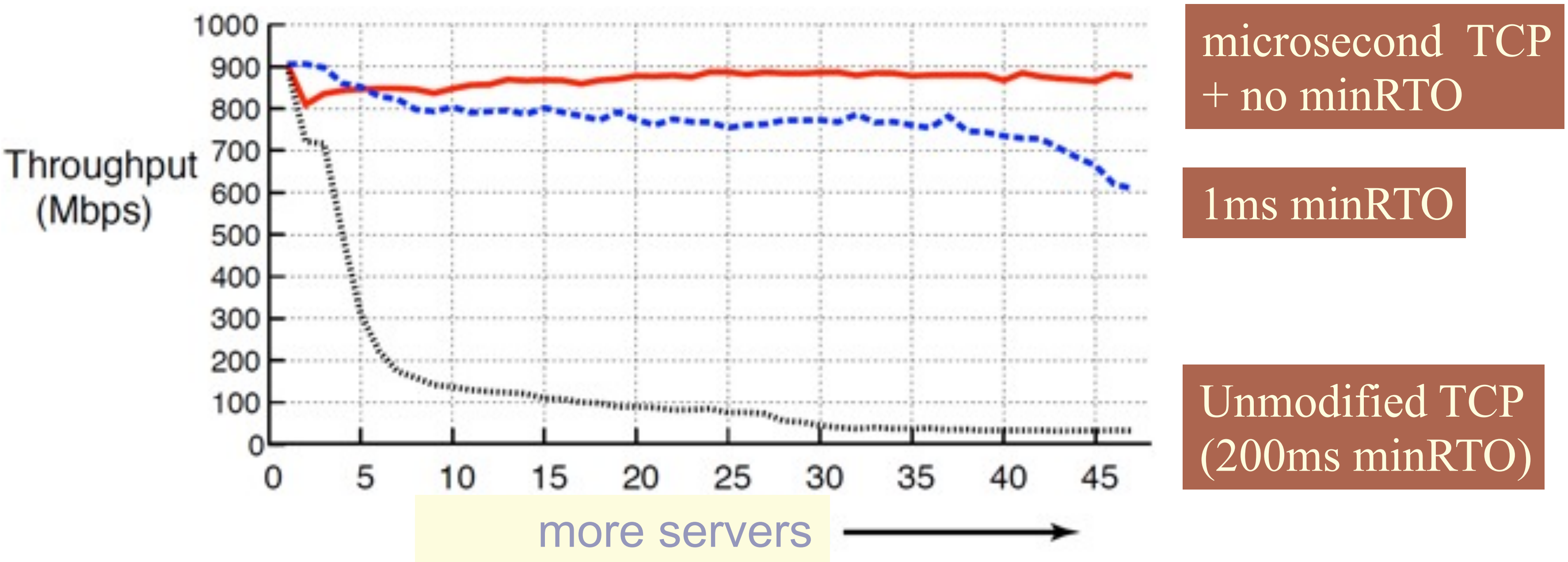
Unmodified TCP
(200ms minRTO)

Lowering minRTO to 1ms helps



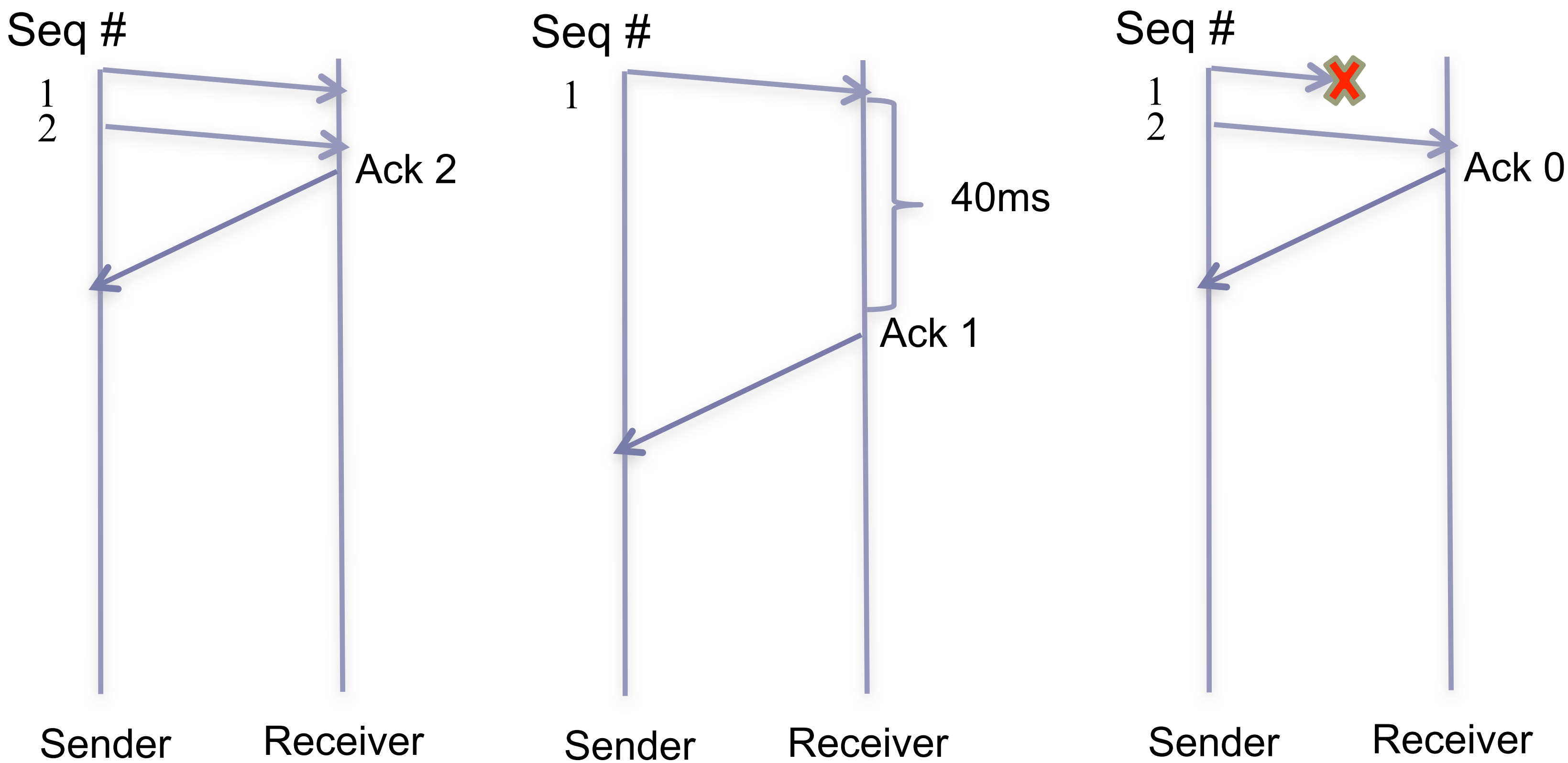
Millisecond retransmissions are not enough

Solution: **μsecond TCP + no minRTO**



✓ High throughput for up to 47 servers

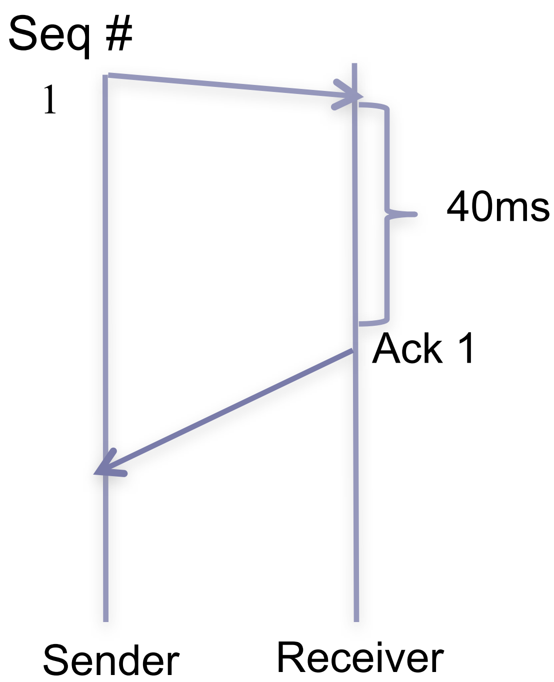
Delayed-ACK (for RTO > 40ms)



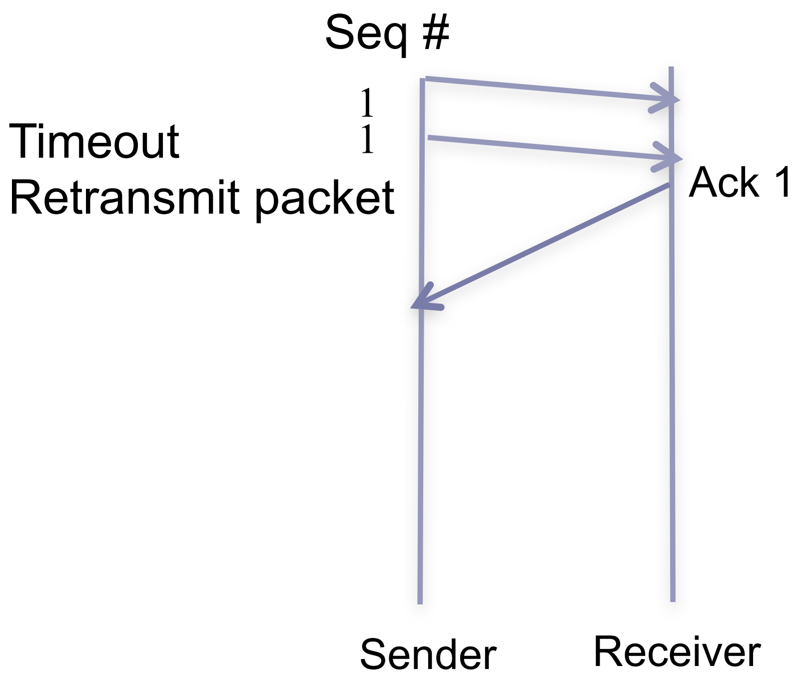
Delayed-Ack: Optimization to reduce #ACKs sent

µsecond RTO and Delayed-ACK

RTO > 40ms



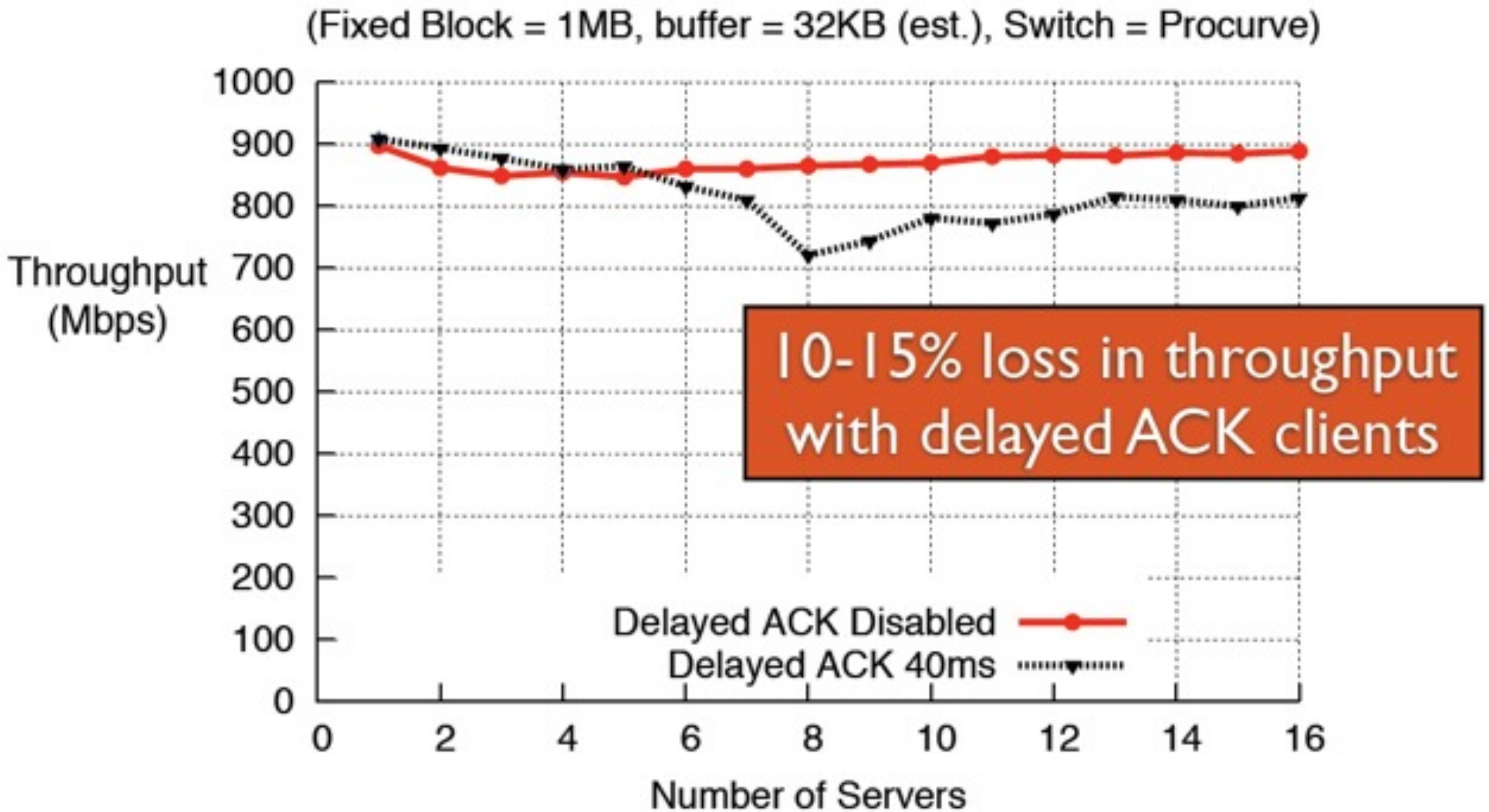
RTO < 40ms



Premature Timeout

RTO on sender triggers before Delayed-ACK on receiver

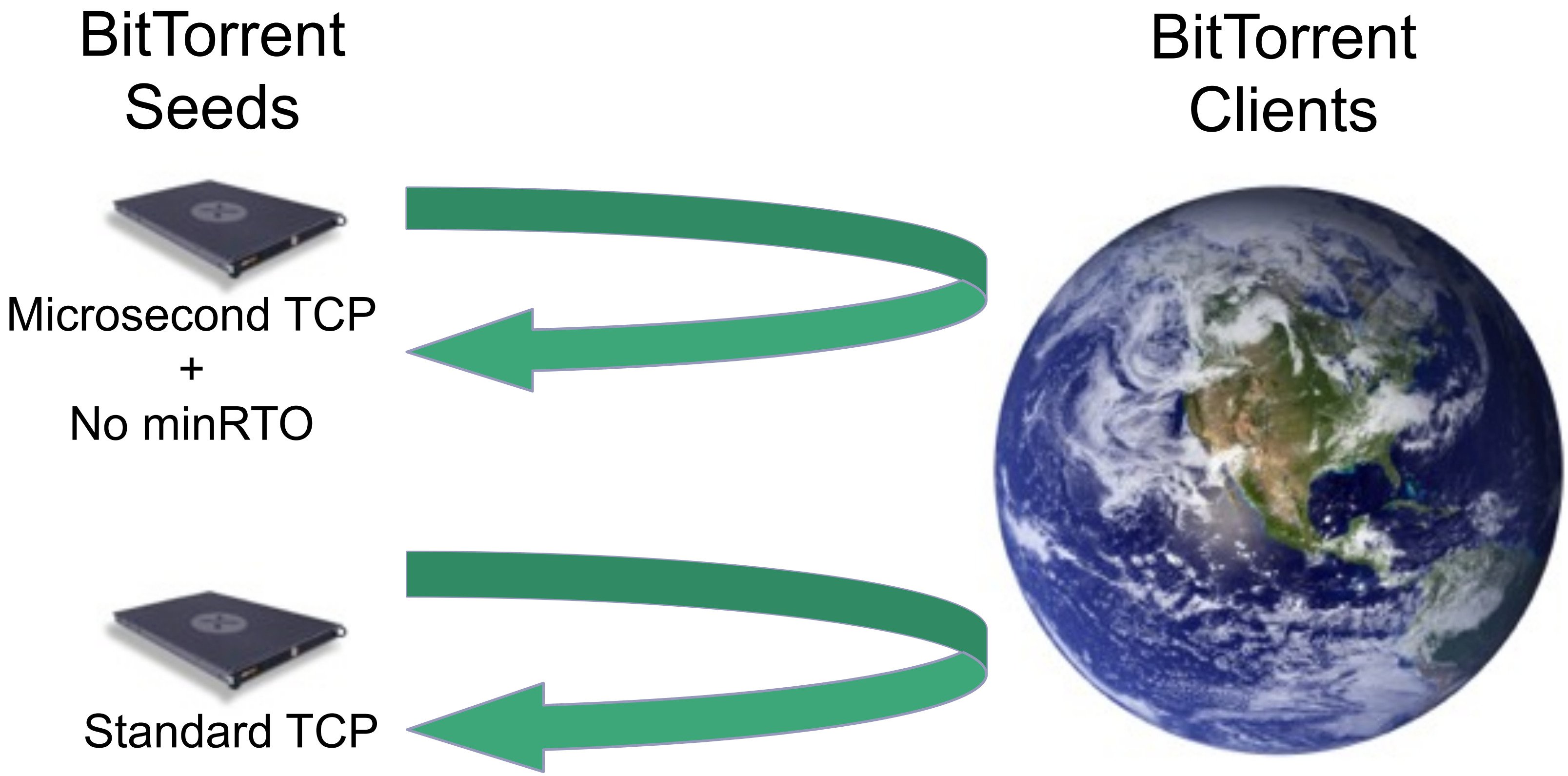
Impact of Delayed-ACK



Is it safe for the wide-area?

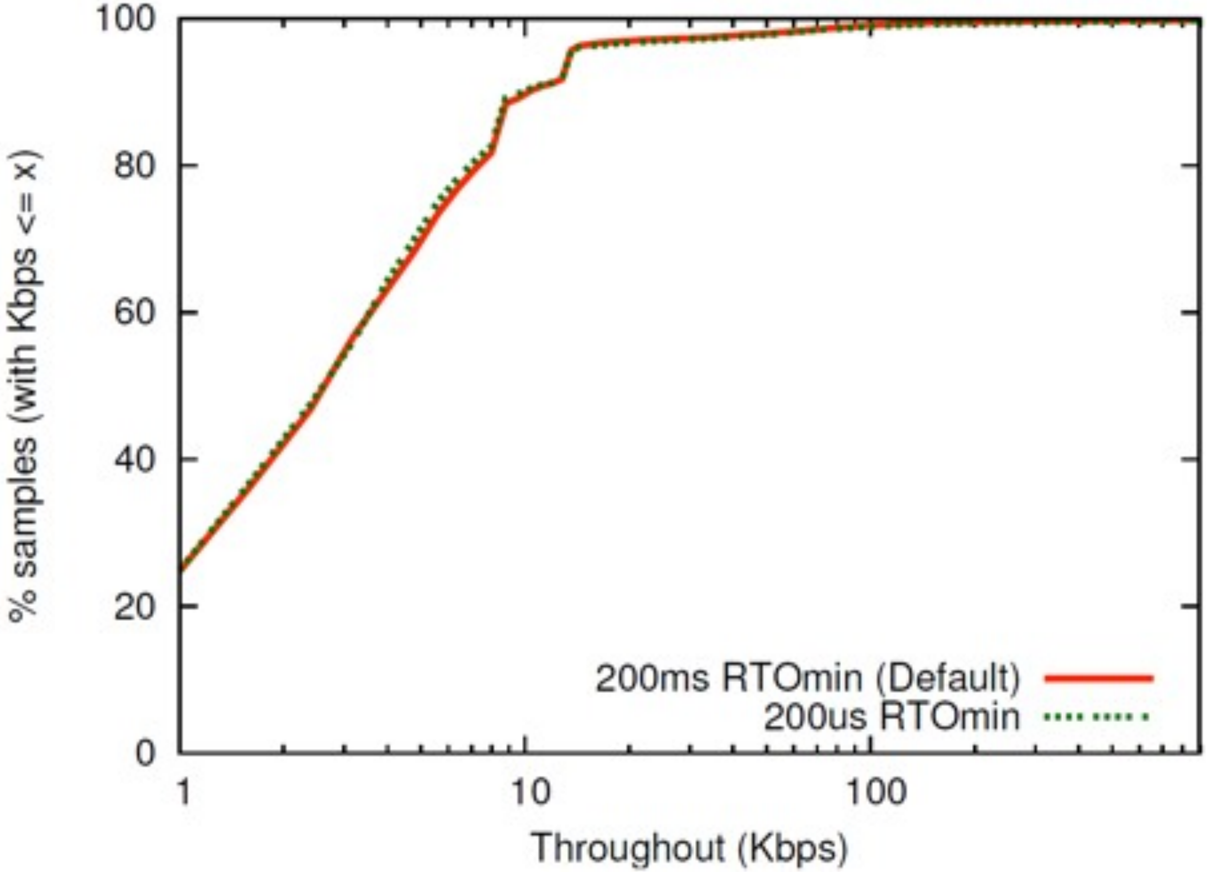
- Stability: Could we cause congestion collapse?
  - No: Wide-area RTOs are in 10s, 100s of ms
  - No: Timeouts result in rediscovering link capacity (slow down the rate of transfer)
- Performance: Do we timeout unnecessarily?
  - [Allman99] Reducing minRTO increases the chance of premature timeouts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

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