Data Center Network Topologies

Raj Jain

Washington University in Saint Louis

Saint Louis, MO 63130

Jain@cse.wustl.edu

These slides and audio/video recordings of this class lecture are at:

http://www.cse.wustl.edu/~jain/cse570-21/

Overview

1. Data Center Physical Layout
2. Data Center Network Cabling
3. ToR vs. EoR
4. Clos and Fat-Tree topologies

Google's Data Center

Student Questions

❑ Do we measure the blade rack in U's as well? Or only the horizontal rack mountable?

Only height of the module is measured in U's. 1U=1.75 inch, Width of the rack is standard 19 inch. 10", 21", and 23" racks are also used. Total height of the rack is also standard 42 U or 45 U.

Ref: https://en.wikipedia.org/wiki/19-inch_rack

Washington University in St. Louis

http://www.cse.wustl.edu/~jain/cse570-21/

©2021 Raj Jain

Cooling Plant

Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/

Student Questions

Servers

Tower

Rack-mountable

Blade

Server Cabinets

Blade Chassis

1U

Mezzanine Card

Blade Server

Source: Santana 2014

1 Rack Unit = 1U=1.75 inch

Washington University in St. Louis

http://www.cse.wustl.edu/~jain/cse570-21/
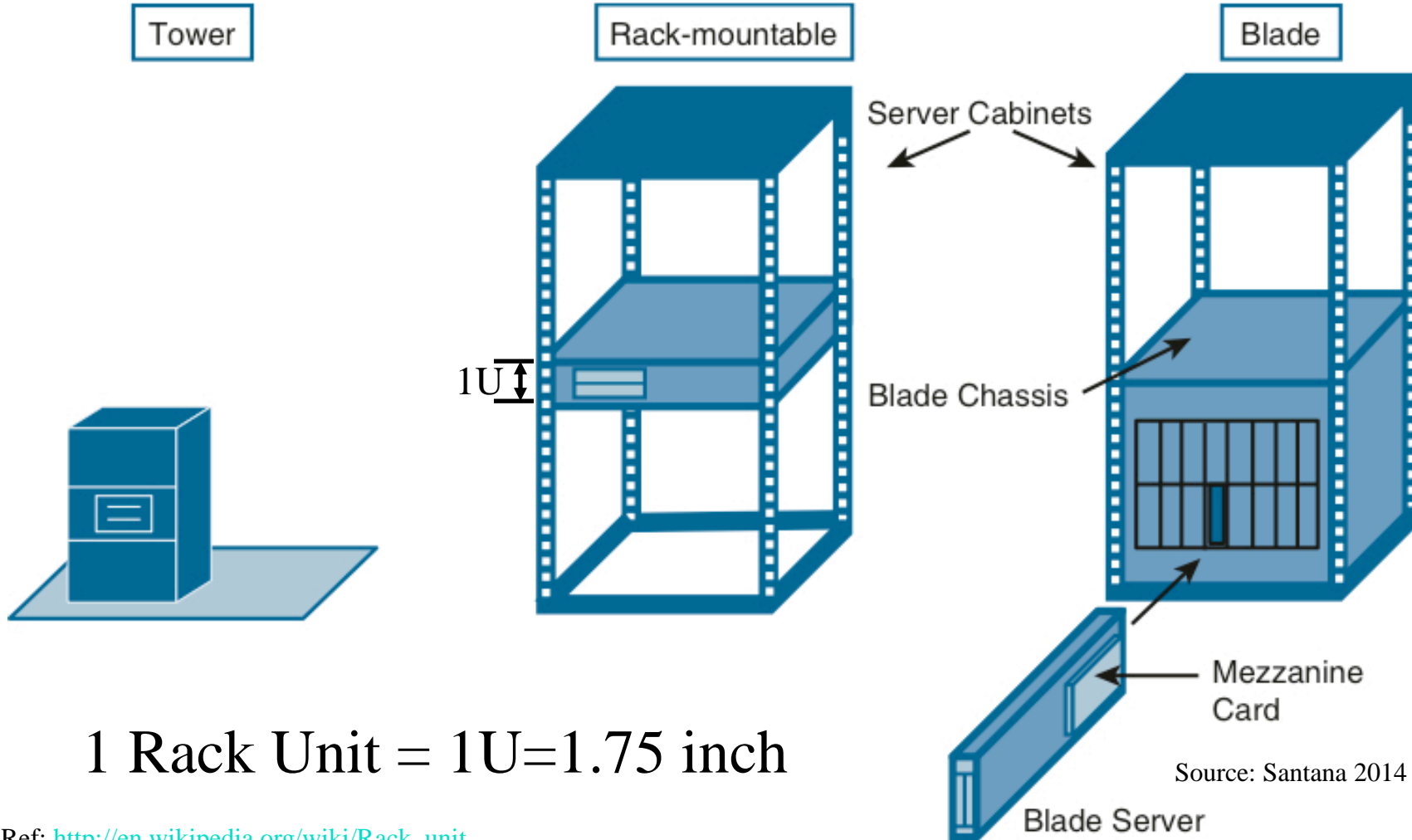
Student Questions

Modular Data Centers

- Small: < 1 MW, 4 racks per unit
- Medium: 1-4 MW, 10 racks per unit
- Large: > 4 MW, 20 racks per unit
- Built-in cooling, high PUE (power usage effectiveness) ≈1.02
  PUE = Power In/Power Used
- Rapid deployment

Ref: http://www.sgi.com/products/data_center/ice_cube_air/

©2021 Raj Jain

Student Questions

- Isn't the lower the PUE the better? Why the quiz answer said otherwise?

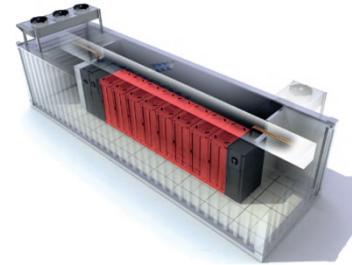*I made a mistake. The correct answer is "Yes, Lower PUE is better."*

- What is the mezzanine card used for in the data center?

Mezzanine=Expansion card, e.g., PCI card installed in parallel to the System board of the module/server

Ref:
https://www.techopedia.com/definition/21300/pci-mezzanine-card-pmc

Containerized Data Center

- ❑ Ready to Use. Connect to water and power supply and go.
- ❑ Built in cooling. Easy to scale.
  ⇒ Data Center trailer parks.
- ❑ Suitable for disaster recovery, e.g., flood, earthquake
- ❑ Offered by Cisco, IBM, SGI, Sun/ORACLE,…

Student Questions

- ❑ The companies listed offer the containerized data center, but do their data centers use this technique too? Or is this method meant for smaller companies? Basically, is this the most common approach now?

Containers may not be most cost-effective method. Aggregated cooling is cheaper than separated cooling. So large data centers may not use containers.

Ref: http://www.datacenterknowledge.com/archives/2010/05/31/iij-will-offer-commercial-container-facility/
http://www.cse.wustl.edu/~jain/cse570-21/

Washington University in St. Louis

©2021 Raj Jain

Unstructured Cabling

Student Questions

Washington University in St. Louis http://www.cse.wustl.edu/~jain/cse570-21/ ©2021 Raj Jain

Structured Cabling

Student Questions

Washington University in St. Louis   http://www.cse.wustl.edu/~jain/cse570-21/   ©2021 Raj Jain

Data Center Physical Layout

Power Backup Systems
Entrance Room
Telecommunications Room
Cooling System
Cabinets
Raised Floor
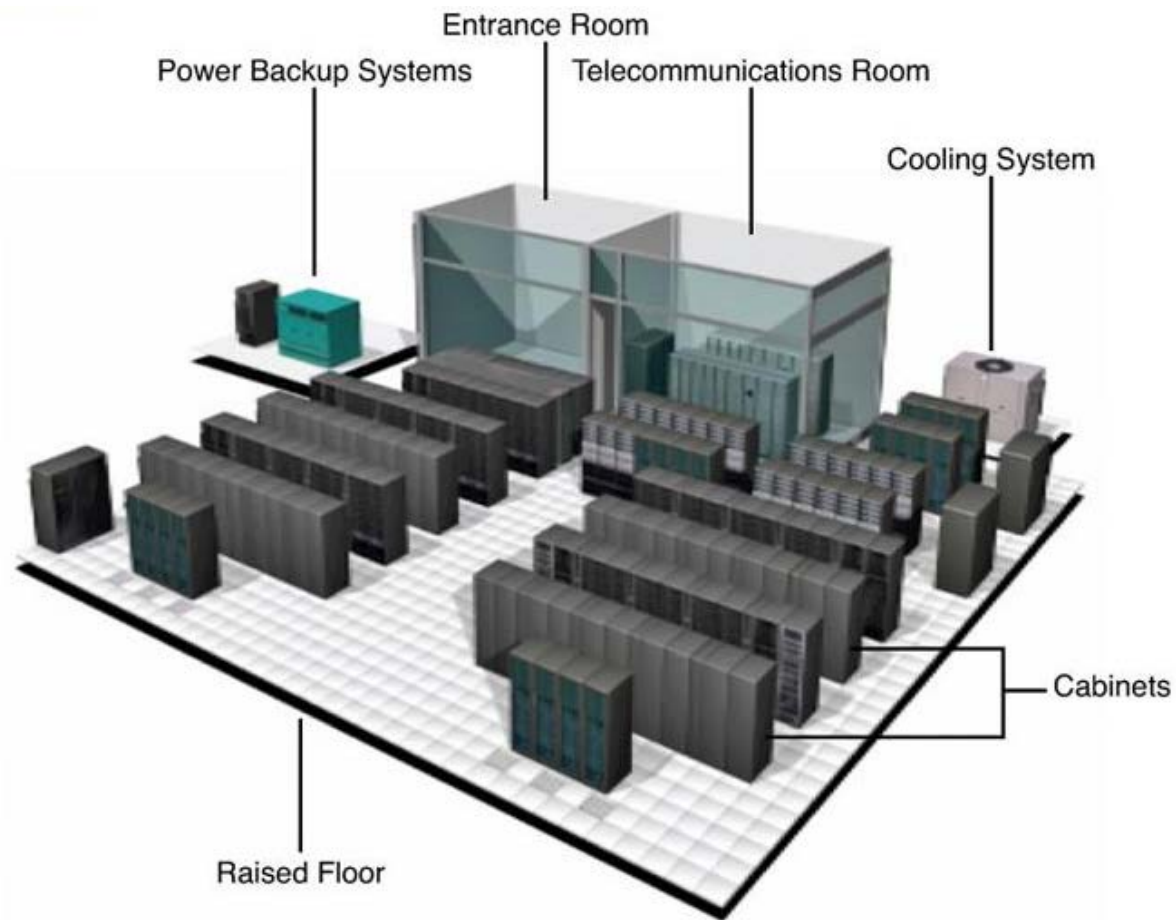
ANSI/TIA-942-2005 Standard

- Main Distribution Area (MDA)
- Horizontal Distribution Area (HDA)
- Equipment Distribution Area (EDA)
- Zone Distribution Area (ZDA)
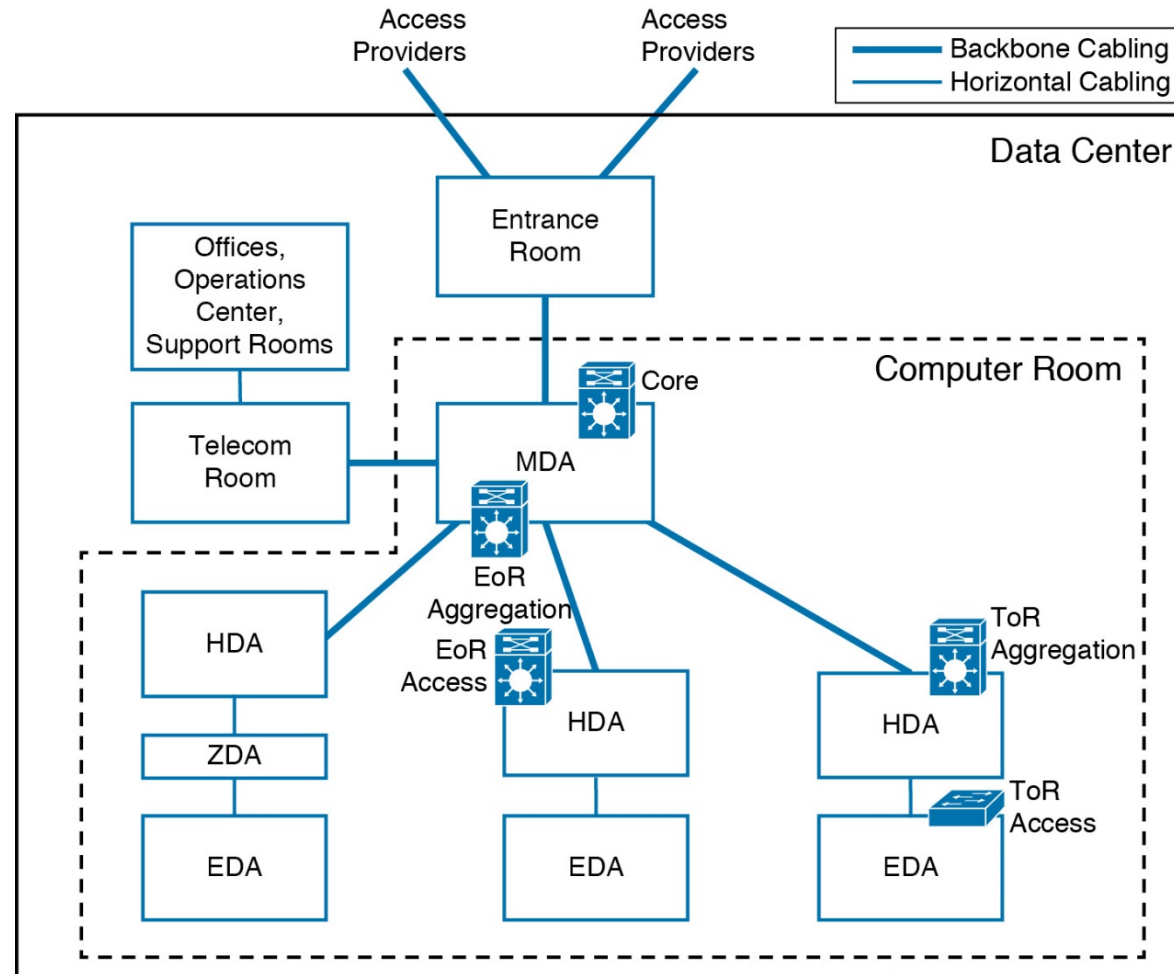
Student Questions

- What do EDA and ZDA do?

Basically these set up hierarchy. So that the networking speeds between the groups can set accordingly.
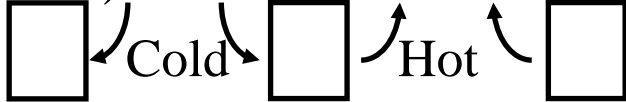*ZDA allows dividing large HDAs. It consists of only passive equipment.*
Passive=No power

Ref:
https://www.anixter.com/content/dam/Suppliers/CommScope/Documents/Data_Center_Topology_Guide.pdf

Source: Santana 2014
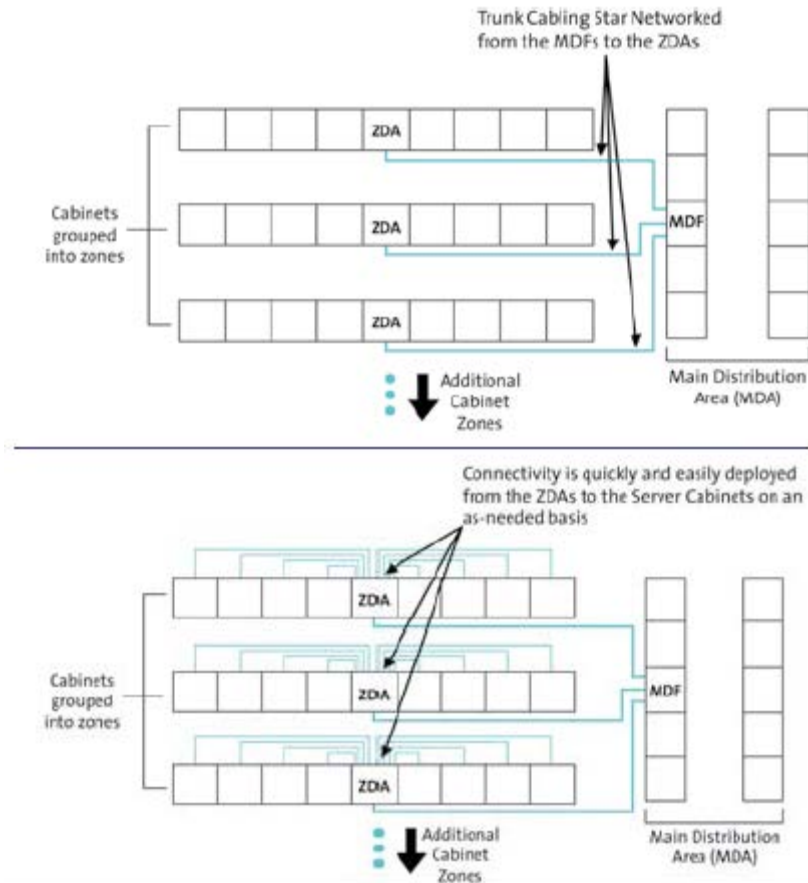
http://www.cse.wustl.edu/~jain/cse570-21/

©2021 Raj Jain

ANSI/TIA-942-2005 Standard

❑ Computer Room: Main servers

❑ Entrance Room: Data Center to external cabling

❑ Cross-Connect: Enables termination of cables

❑ Main Distribution Area (MDA): Main cross connect. Central Point of Structured Cabling. Core network devices

❑ Horizontal Distribution Area (HDA): Connections to active equipment.

❑ Equipment Distribution Area (EDA): Active Servers+Switches. Alternate hot and cold aisle. ▢ ↙Cold↘ ▢ ↗Hot↖ ▢

❑ Zone Distribution Area (ZDA): Optionally between HDA and EDA.

❑ Backbone Cabling: Connections between MDA, HDA, and Entrance room

Student Questions

Zone Distribution Area

Trunk Cabling Star Networked from the MDFs to the ZDAs

Connectivity is quickly and easily deployed from the ZDAs to the Server Cabinets on an as-needed basis

❑ High-fiber count cables connect ZDA to MDA or HDA. Low-fiber count cables connect ZDA to EDA as needed.

Ref: Jennifer Cline, "Zone Distribution in the data center,"
http://www.graybar.com/documents/zone-distribution-in-the-data-center.pdf

Data Center Network Topologies: 3-Tier

❑ Core, Aggregation, Access
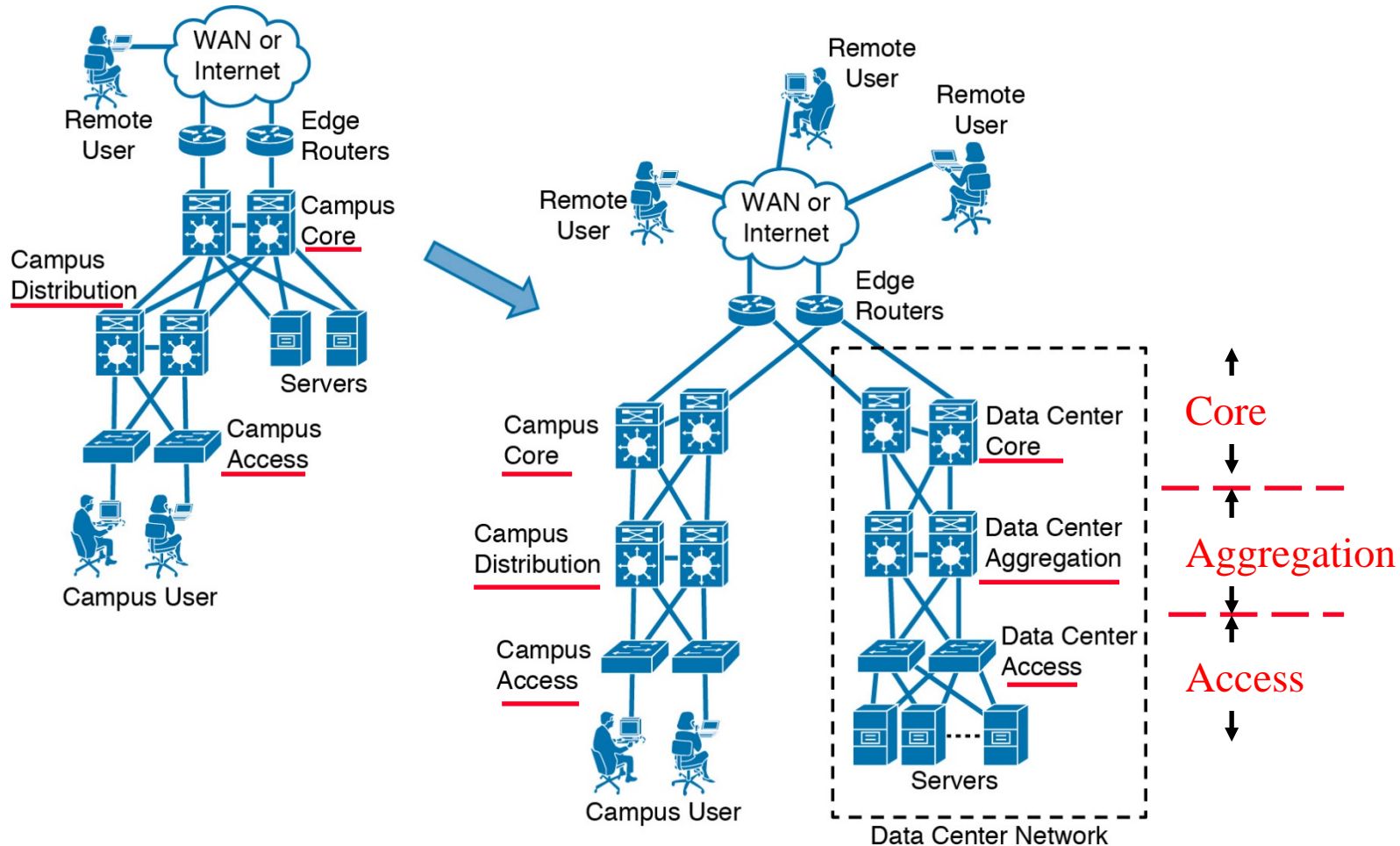
Student Questions

❑ Core, Aggregation, Access? Since the video doesn't show where you look at, please explain one more time.

See updates in red.
Access starts from users.

Source: Santana 2014
http://www.cse.wustl.edu/~jain/cse570-21/
©2021 Raj Jain

3-Tier Data Center Networks

- 20-40 servers per rack. Limited by power/cooling
- Each server connected to 2 access switches with 1 Gbps (10 Gbps becoming common)
- Access switches connect to 2 aggregation
- All switches below each pair of aggregation switches form a single layer-2 domain
- All traffic **north** of aggregation switches forwarded by L3 routing (South = Servers, North = Internet)
  $\Rightarrow$ Aggregation switches are L3 switches $\Rightarrow$ implement routing
- Aggregation switches connect to 2 core L3 switches
- Core L3 switches connect to edge routers
- Core layer forwards data center ingress and egress traffic

L3
L2

Student Questions

Ref: A. Greenberg, "VL2: A Scalable and Flexible Data Center Network," CACM, Vol. 54, NO. 3, March 2011, pp. 95-104, http://research.microsoft.com/pubs/80693/vl2-sigcomm09-final.pdf.

3-Tier Data Center Networks (Cont)

- Aggregation layer is also a place to put middleboxes, such as, firewalls, load balancers
- Access Layer provide high number of ports for connectivity.
- Low Latency: In high-frequency trading market, a few microseconds make a big difference.
  $\Rightarrow$ Cut-through switching and low-latency specifications.
- Each Layer 2 domain typically limited to a few hundred servers to limit broadcast
- Most traffic is internal to the data center.
- Most of the flows are small.
  Mode = 100 MB. DFS uses 100 MB chunks.
- Aggregation layer forwards server-to-server traffic in the data center $\Rightarrow$ Not ideal for East-West Traffic
- Network is the bottleneck.
  Uplinks utilization of 80% is common.

Student Questions
- For cut-through, how does the receiver know if the packet that shows up is bad if we sent it without checking? Or do we never tell them in the interest of speed?

CRC does not match and so the packet is thrown away by some one on the path or at the destination.
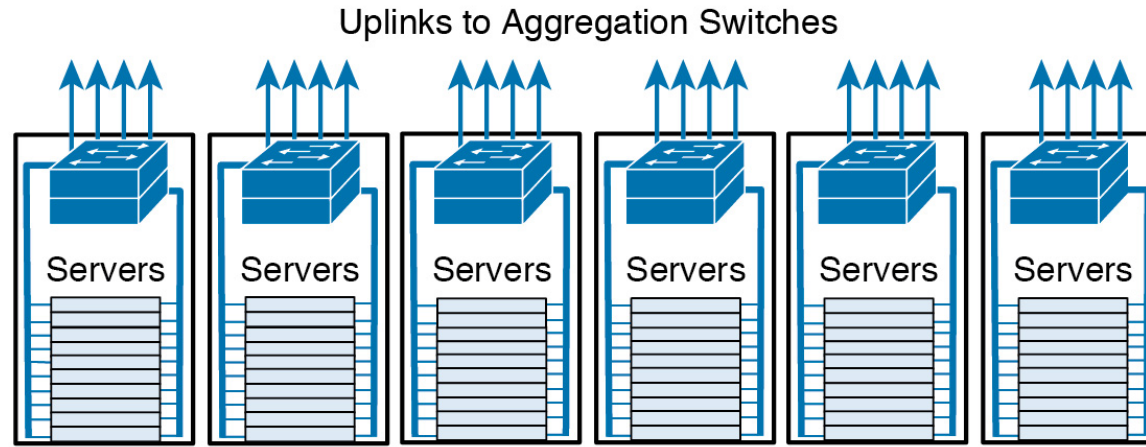
- Are there any other problems with 'data center internal traffic'?

This is still a topic of research. Remember "Network is the bottleneck."

http://www.cse.wustl.edu/~jain/cse570-21/
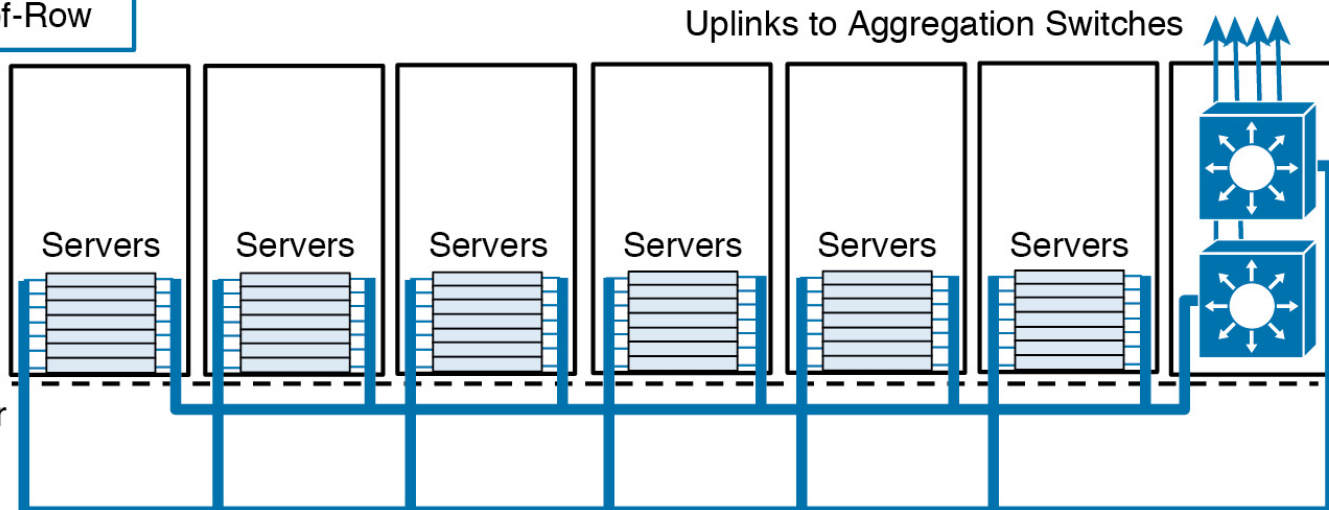©2021 Raj Jain

Switch Locations

Top-of-Rack

Uplinks to Aggregation Switches

Smaller cable between servers and switches
Network team has to manage switches on all racks

Servers Servers Servers Servers Servers Servers

Raised Floor

End-of-Row

Uplinks to Aggregation Switches

All network switches in one rack

Servers Servers Servers Servers Servers Servers

Raised Floor

Source: Santana 2014

Washington University in St. Louis

http://www.cse.wustl.edu/~jain/cse570-21/

©2021 Raj Jain

Student Questions

❑ The quiz said "all switches are put on the top of the rack." However, there is also end-of-row situation. Why "all switches are put on top of rack." is considered a true statement?

"All switches on ToR" is false.

❑ EOR is end of rack or end of row?

End of Row.

❑ Why are switches on the top of the rack if the cabling is along the floor?

Cabling can be under-floor or in ceilings.

3-17

ToR vs EoR

❑ ToR:

- ➢ + Easier cabling
- ➢ - If rack is not fully populated $\Rightarrow$ unused ToR ports
- ➢ - If rack traffic demand is high, difficult to add more ports
- ➢ - Upgrading (1G to 10G) requires complete Rack upgrade

❑ EoR:

- ➢ - Longer cables
- ➢ + Severs can be placed in any rack
- ➢ + Ports can easily added, upgraded

Student Questions

3-Tier Hierarchical Network Design

❑ All servers require application delivery services for security (VPN, Intrusion detection, firewall), performance (load balancer), networking (DNS, DHCP, NTP, FTP, RADIUS), Database services (SQL)

❑ ADCs are located between the aggregation and core routers and are shared by all servers

❑ Stateful devices (firewalls) on Aggregation layer

❑ Stateful = State of TCP connection

❑ Stateless, e.g., DNS
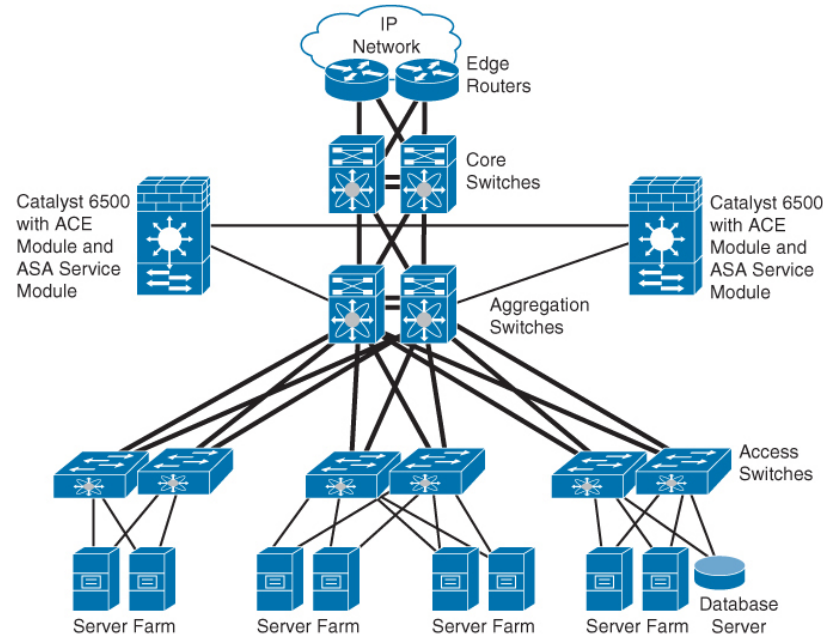
Source: Santana 2014

Student Questions

❑ Except longer cables, EoR seems to have more pros and why is ToR being used more?

Longer cables ⇒ Higher chances of failure.
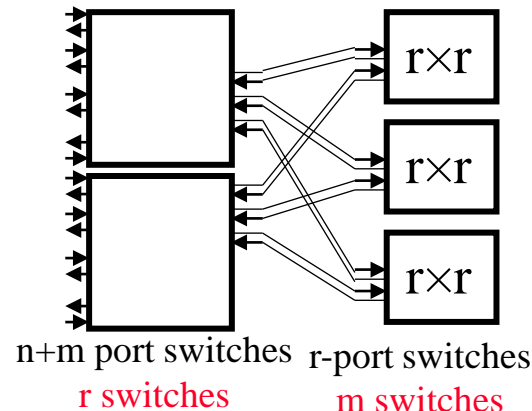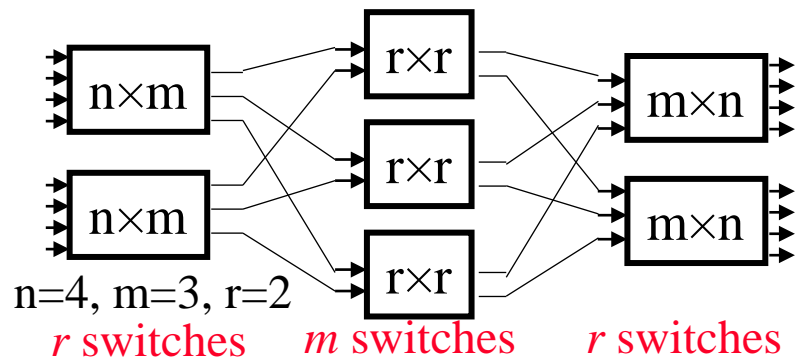
Problem with 3-Tier Topology

Student Questions

❑ Failure of a single link can reduce the available bandwidth by half

❑ With more than two aggregation switches, spanning tree becomes unpredictable in case of certain failures.

❑ Two aggregation switch => They are the bottleneck

❑ It is not possible for VLANs to span across multiple pairs of aggregation switches since the pairs are connected by L3

❑ VLAN provisioning becomes laborious

Washington University in St. Louis

http://www.cse.wustl.edu/~jain/cse570-21/

©2021 Raj Jain

Clos Networks

- Multi-stage circuit switching network proposed by Charles Clos in 1953 for telephone switching systems
- Allows forming a large switch from smaller switches
  The number of cross-points is reduced ⇒ Lower cost (then)
- 3-Stage Clos(n, m, r): ingress (rn×m), middle (mr×r), egress (rm×n)
- *Strict-sense non-blocking* if m ≥ 2n-1. Existing calls unaffected.
- *Rearrangeably non-blocking* if m ≥ n
- Can have any odd number of stages, e.g., 5
- **Folded**: Merge input and output in to one switch

n=4, m=3, r=2

r switches *m* switches *r* switches

n+m port switches r-port switches
r switches m switches

Student Questions

- Can you please provide an example on how to draw a 3-stage topology?
  Draw all possible connections
- Could you show how we got to the folded version of the Clos Network? I don't see how they are equivalent yet.
  Draw on a paper and fold in the center.
- How to define the number of input/output ports for the r switches? for the figure on the right? is it n×2?
  Yes. Each bidirectional line pair is one port.
- In the case where m>n what happens after folding? do we need to rearrange?
  You cannot have m>n. If there are n streams coming in, they will have to find a way to select n out of m output ports in side the switch and so the extra ports will be wasted.
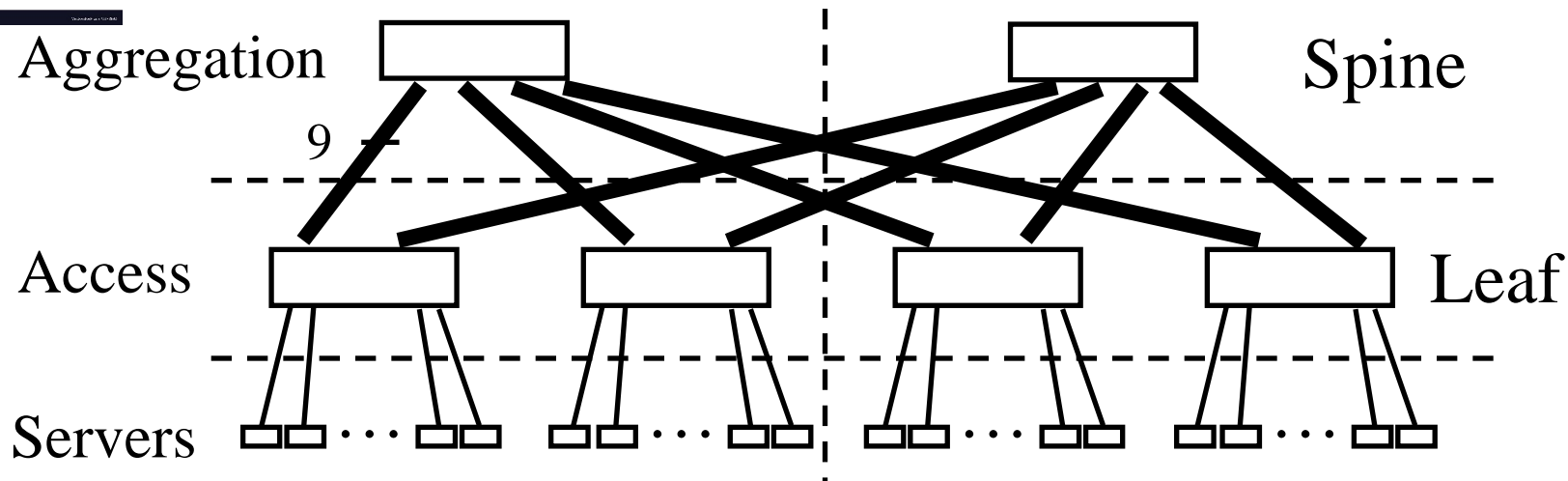
Homework 3A

❑ Draw a 3-stage clos(4, 5, 3) topology and its folded version. $n = 4$, $m = 5$, $r = 3$

Student Questions

http://www.cse.wustl.edu/~jain/cse570-21/
©2021 Raj Jain

Fat-Tree DCN Example

Aggregation   Spine

9

Access   Leaf

Servers

- ❑ 6 identical 36-port switches. All ports 1 Gbps. 72 Servers.
- ❑ Each access switch connects to 18 servers.
  9 Uplinks to first aggregation switch.
  Other 9 links to 2nd aggregation switch.
- ❑ Throughput between any two servers = 1 Gbps using ECMP
  Identical bandwidth (36 Gbps) at any bisection.
- ❑ Negative: Cabling complexity

Ref: Teach yourself Fat-Tree Design in 60 minutes, http://clusterdesign.org/fat-trees/
Washington University in St. Louis

http://www.cse.wustl.edu/~jain/cse570-21/

©2021 Raj Jain

Student Questions

- ❑ What do we need to take into consideration when we want to migrate a data center from 3-tier DCN to Fat-tree topology?

Don't' fix what is working.

- ❑ In 3-tier DCN topology, How does VLAN traffic route? I assume that the VLAN traffic needs to go to the aggregation level/tier, which increasing the aggregation switches' load.

VLANs have their own tree. All switches have to take care of VLANs going through them. Even the core switches. More to come during virtualization module.

- ❑ Does the VLAN traffic path differ on Fat-tree topology?

*Same path routing techniques (not yet discussed) work on all topologies.*

- ❑ What are the routing protocols used on 3-tier and Fat-Tree topology?

To be discussed in virtualization module. Depends on the virtualization level: L2 or L3.

Fat-Tree Topology (Cont)

❑ Half of leaf switch ports are towards servers and the other half towards spine

❑ With 36 port switches $\Rightarrow$ 18 ports to spine
$\Rightarrow$ 2, 3, 6, 9, 18 spine switches

❑ Maximum # of spine switches = ½ # of ports on leaf switches

Spine

6
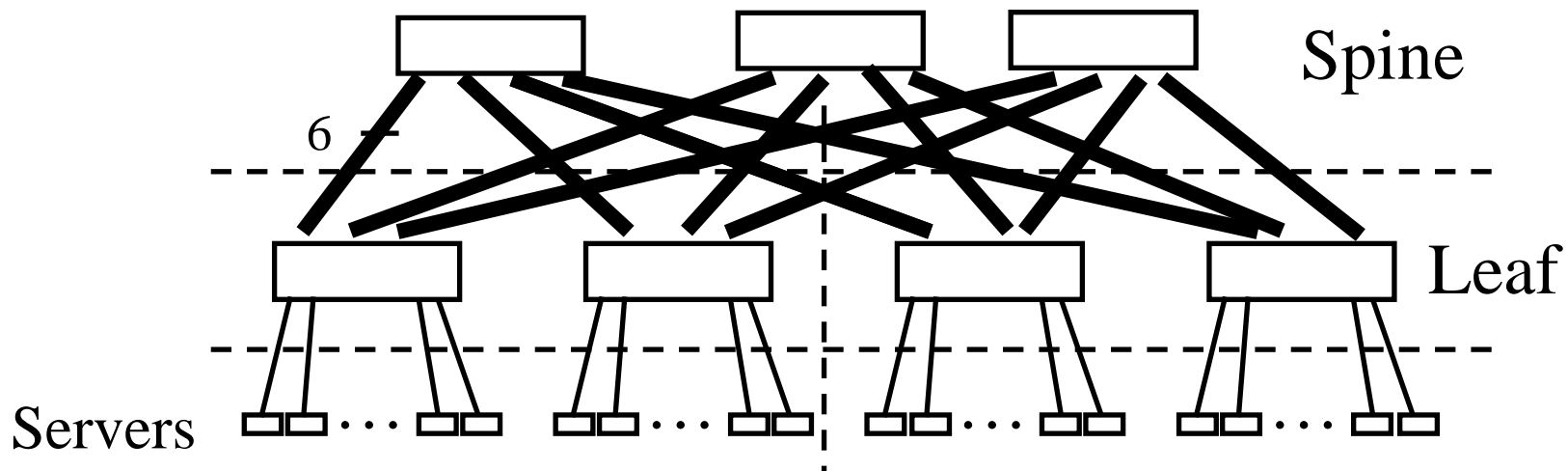
Leaf

Servers

❑ Largest configuration with n-port switches: $n^2/2$ servers can be connected using $n+n/2$ switches.
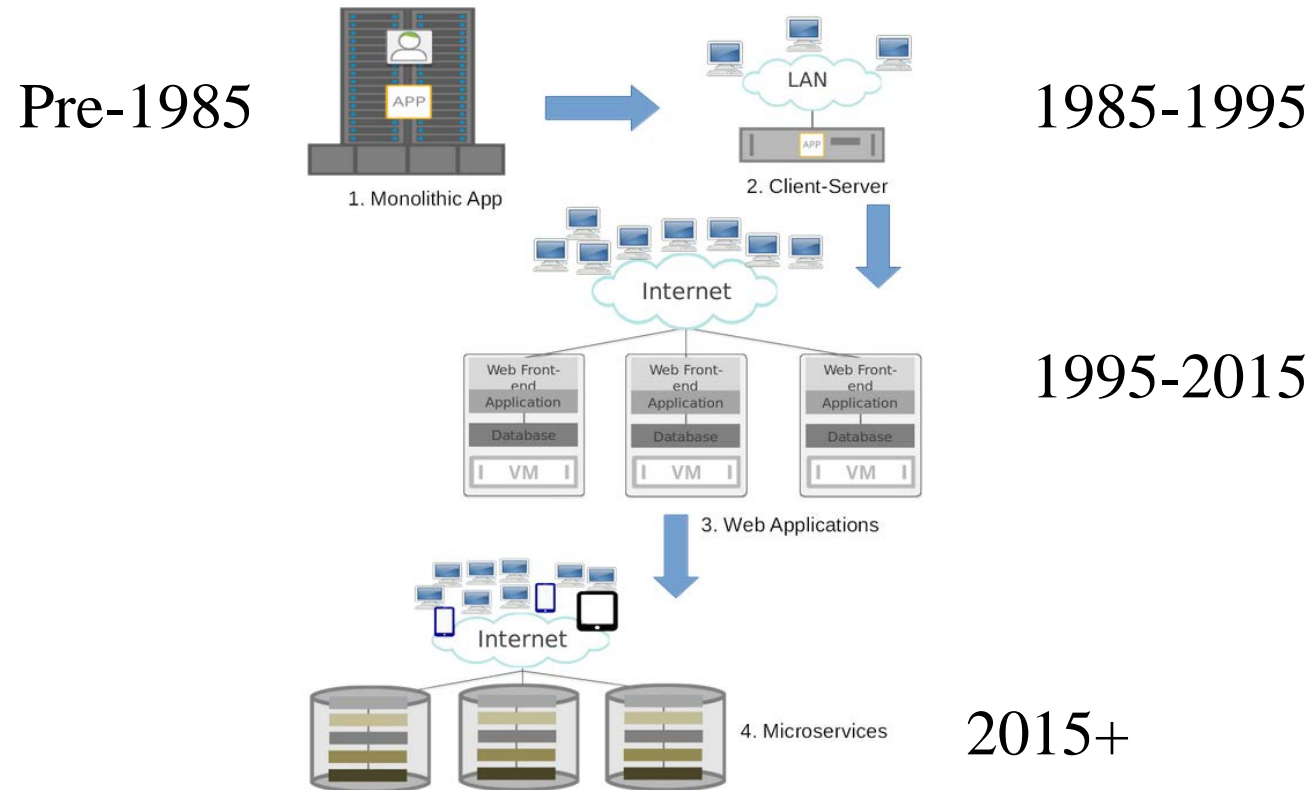
Homework 3B

1. Draw the largest Fat-tree topology using 4-port switches. Assume each server is connected to a single leaf switch while the leaf switches are multi-homed to spine switches. There is no core tier.

2. How many servers can be connected in the above configuration?

3. How many switches in all are required in the above configuration?

4. How many servers can be connected using 64-port switches.

5. How many switches are required to form the spine and the leaves using 64-port switches.

Student Questions
- Question three, "in all" means (spine + leaf)?

Yes

http://www.cse.wustl.edu/~jain/cse570-21/
©2021 Raj Jain

Evolution of Applications

Pre-1985

1985-1995

1995-2015

2015+

1. Monolithic App
2. Client-Server
3. Web Applications
4. Microservices

❑ **Larger Servers to Micro-Services ⇒ Increasing network demand**

Ref: Dinesh G. Dutt, "Cloud-Native Data Center Networking," O'Reilly Media, Inc., December 2019, ISBN: 9781492045595, Safari Book.
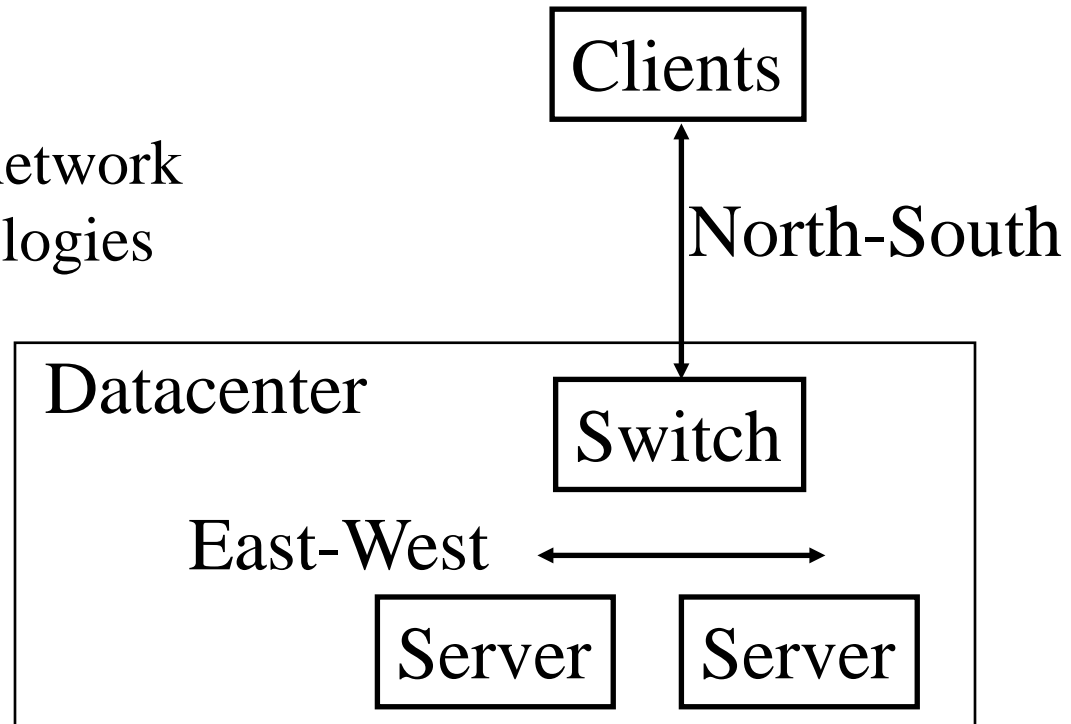

Student Questions

❑ Is the high-level motivation behind a flatter structure making leaf-leaf communication possible, easier, or faster? Does this change slow down the more traditional uses?

East-West (server-server) traffic is now more than north-south (server-client) traffic
⇒ Need leaf-leaf communication

North-South vs. East-West Traffic

❑ Previously, most of the traffic was north-south
⇒ Between servers in the data center and clients out-side

❑ Now the trend is towards traffic between servers for big data analysis
⇒ East-West traffic
⇒ Requires flatter network
⇒ Fat-tree like topologies

Advantages of 2-Tier Architecture

❑ Homogeneous Equipment: Spine and leaf switches both have the same number of ports with the same speed.
⇒ Maintenance and replacements is easier

❑ L2 forwarding is used only in each rack.
⇒ a new protocol (VXLAN) is used for routing between racks

❑ A leaf can reach any other leaf via any spine at the same cost
⇒ Equal cost multi-path (ECMP) simplifies routing

❑ All packets of a flow are sent using the same path to avoid out-of-order arrivals.

  ➢ Flow = {Source IP, Dest IP, L4 Protocol, Source Port, Dest Port)

  ➢ Flow hashing is used to select a spine switch

Washington University in St. Louis   http://www.cse.wustl.edu/~jain/cse570-21/   ©2021 Raj Jain

---

Student Questions

❑ Will some data packets be lost? How will it be fixed if it is lost?

L4 protocols take care of lost packets.

❑ Beside VXLAN, are companies using some proprietary protocols for servers communication?

There are no proprietary protocols.

❑ Can you please elaborate more on how hashing is used to select spine switch?

Hashing results in a pseudo-random number between 1 and n. That is the switch that is selected.

❑ In selecting the spine switch, how does this hash ensure equal distribution among the spine swithces?

By design, all numbers are equally likely.

❑ What is the difference between VXLAN and VLAN?

Virtual eXtended LAN extends over many different IP domains. It allows L2 over L3.

Variations

- Higher-speed Inter-Switch Links (ISLs) may be used:
  - 1 Gbps server/10 Gbps ISL, 10 Gbps Server/40 Gbps ISL
  - Reduces number of spine switches required
    (Smaller number of ECMP may result in some congestion.
    Also, loss of a spine may have a more severe impact)
- Two leaves per rack. Hosts are dual-ported.
- Three-tier Clos: $n^3/4$ servers using $n+n^2$ switches
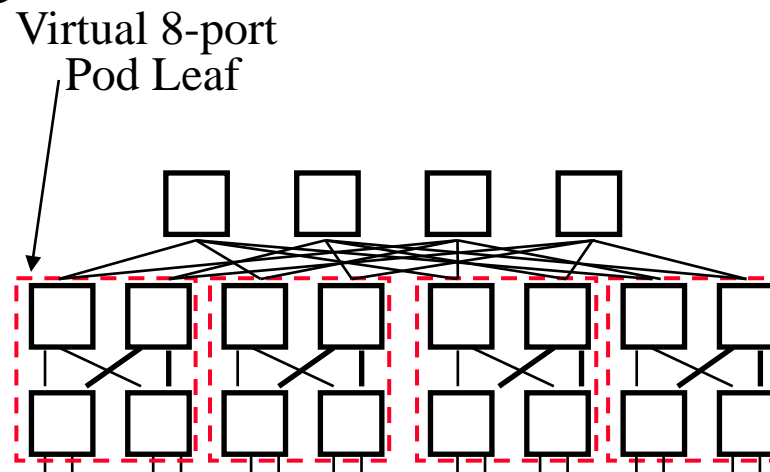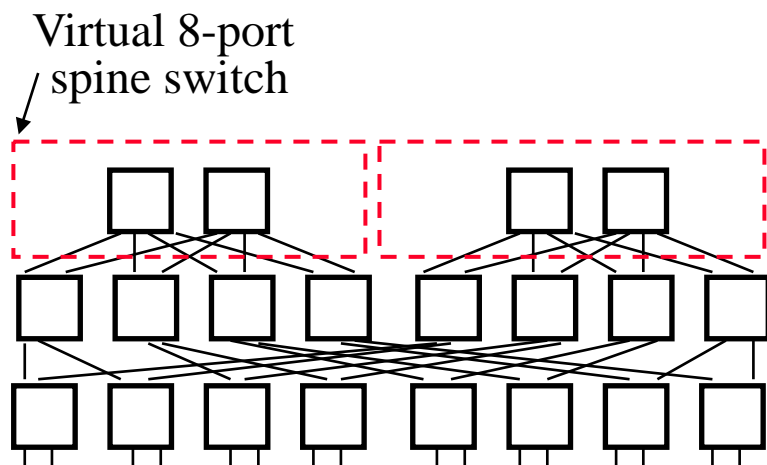
Virtual 8-port
spine switch

Virtual 8-port
Pod Leaf

Ref: Dinesh G. Dutt, "Cloud-Native Data Center Networking," O'Reilly Media, Inc., December 2019,
ISBN: 9781492045595, Safari Book.

http://www.cse.wustl.edu/~jain/cse570-21/

©2021 Raj Jain

Student Questions

- In a Clos network, increasing the number of ports on switches have huge gain (e.g. cubic return in a three-tier). Then how expensive is it to have switches with more ports (linear, quadratic, etc.)?

n-ports require n^2 internal connections. Each connection needs a queue/buffer area. At some point it becomes infeasible.

- Are there any intrinsic costs within the switches with more ports?

See above.

- In practice how many ports are there on the switches in Clos network?
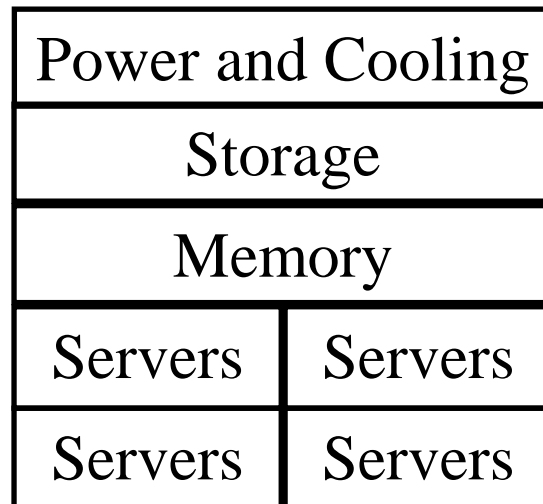
Could be several thousands.

- A two-tiered design has only the core and the edge tiers. It can support between 5K to 8K hosts. With 3-tiers, [1] targets 25,000 hosts similar to the picture at the bottom-right They call the layers Code, Edge and Aggregation.
[1] 10.1145/1402958.1402967.
Yes, each tier or switch could be a Clos network.

Rack-Scale Architecture

❑ Traditionally each server has its own cooling, storage, memory, and networking ⇒ Inefficient use of dedicated resources

❑ Shared resources ⇒ Rack-Scale Architecture (RSA)

❑ Memory, Storage, Cooling is shared by all servers on the rack Server "sleds" plug in to networking board on the back

❑ Buy complete racks rather than individual servers

❑ Being standardized by Open Compute Project (OCP)

Power and Cooling	
Storage	
Memory	
Servers	Servers
Servers	Servers

Washington University in St. Louis

http://www.cse.wustl.edu/~jain/cse570-21/

©2021 Raj Jain

Micro-Servers

❑ Micro-server = a small system on a chip (SOC) containing CPU, memory and multiple NICs

❑ Many micro-servers on a board (look like memory DIMMs)

❑ Micro-server sleds can replace server sleds in rack scale architecture

Student Questions

❑ So micro-servers are just SOCs such that we can fit many into a single server slot in a rack?

Yes.

❑ What differentiates micro-servers from micro-services?

Servers are hardware and cost money. Services are request and produce money.

Summary
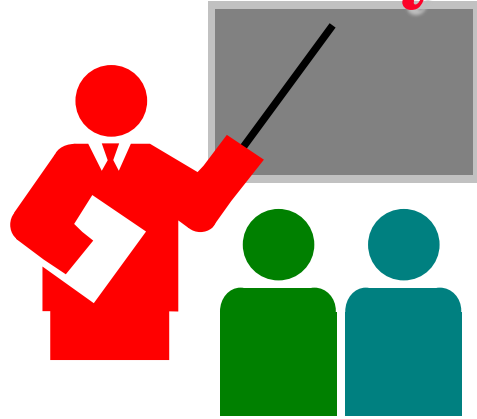
1. Modular data centers can be used for easy assembly and scaling
2. Three tiers:
   1. Access, Aggregation, Core
   2. Application delivery controllers between Aggregation and core.
   3. Need large L2 domains => Past
3. Clos-Based Fat-tree topology is being used to improve performance and reliability

Acronyms

ADC	Application Delivery Controller
ANSI	American National Standards Institute
BPE	Business Process Engineering
CSW	Core Switch
DCBX	Data Center Bridging eXtension
DCN	Data Center Network
DFS	Distributed File System
DHCP	Dynamic Host Control Protocol
DIMM	Dual Inline Memory Module
DNS	Domain Name System
ECMP	Equal Cost Multipath
EDA	Equipment Distribution Area
EoR	End of Row

Student Questions

❑

Acronyms (Cont)

ETS   Enhanced Transmission Selection

EVB   Edge Virtual Bridge

FC   Fibre Channel

FSW   Fabric switch

FTP   File Transfer Protocol

HDA   Horizontal Distribution Area

LACP   Link Aggregation Control Protocol

LAG   Link Aggregation

LLDP   Link Layer Discovery Protocol

MAC   Media Access Control

MDA   Main Distribution Area

MW   Mega-Watt

NIC   Network Interface Card

NTP   Network Time Protocol

NVGRE   Network Virtualization using Generic Routing Encapsulation

OCP   Open Compute Project

Student Questions

❑

Acronyms (Cont)

PFC   Priority Flow Control
PUE   Power Usage Effectiveness
RADIUS   Remote Authentication Dial-In User Service
RPC   Remote Procedure Call
RSA   Rack Scale Architecture
RSW   Rack switch
SOC   System on Chip
SQL   Structured Query Language
SSW   Spine Switches
STP   Spanning Tree Protocol
TIA   Telecommunications Industry Association
ToR   Top of Rack
TRILL   Transparent Interconnection of Lots of Link
VLAN   Virtual Local Area Network
VM   Virtual Machine
VPN   Virtual Private Network

Student Questions

❑

http://www.cse.wustl.edu/~jain/cse570-21/   ©2021 Raj Jain

Acronyms (Cont)

VRF   Virtual Routing and Forwarding
VXLAN   Virtual Extensible Local Area Network
ZDA   Zone Distribution Area

Student Questions

❑

Reading List

❑ Dinesh G. Dutt, "Cloud-Native Data Center Networking," O'Reilly Media, Inc., ecember 2019, ISBN: 9781492045595, Safari Book (Chapters 2 and 3)

❑ G. Santana, "Data Center Virtualization Fundamentals," Cisco Press, 2014, ISBN:1587143240 (Safari book) (Chapters 1 and 2)

Student Questions

❑

References

- A. Greenberg, "VL2: A Scalable and Flexible Data Center Network," CACM, Vol. 54, NO. 3, March 2011, pp. 95-104, http://research.microsoft.com/pubs/80693/*vl2*-sigcomm09-final.pdf

- http://en.wikipedia.org/wiki/Clos_network

- Teach yourself Fat-Tree Design in 60 minutes, http://clusterdesign.org/fat-trees/

- http://webodysseum.com/technologyscience/visit-the-googles-data-centers/

- http://www.sgi.com/products/data_center/ice_cube_air/

- Datacenter Infrastructure - mobile Data Center from Emerson Network Power, http://www.datacenterknowledge.com/archives/2010/05/31/iij-will-offer-commercial-container-facility/

- Jennifer Cline, "Zone Distribution in the data center," http://*www.graybar.com/documents/**zone*-*distribution*-in-the-data-center.pdf

Student Questions

-

Wikipedia Links

- http://en.wikipedia.org/wiki/Modular_data_center
- http://en.wikipedia.org/wiki/Data_center
- http://en.wikipedia.org/wiki/Structured_cabling
- http://en.wikipedia.org/wiki/Cable_management
- http://en.wikipedia.org/wiki/Raised_floor
- http://en.wikipedia.org/wiki/Data_center#environmental_control
- https://en.wikipedia.org/wiki/Hierarchical_internetworking_model
- http://en.wikipedia.org/wiki/Fat_tree
- http://en.wikipedia.org/wiki/Clos_network

Student Questions

-

Scan This to Download These Slides

http://www.cse.wustl.edu/~jain/cse570-21/m_03dct.htm

Raj Jain

http://rajjain.com

Student Questions

- To double confirm we jumped from slide 23 to 40 in the video min 1:11:16

Yes. All videos end with the QR code.

- Captioned words were not reviewed, they still appear as %%%.

My mistake.

- If possible, could you provide an overview of topology switching and your opinion on allowing applications to create custom topology?

Create routes based on application needs.
Ref: Kevin C. Webb, Alex C. Snoeren, and Kenneth Yocum, "Topology Switching for Data Center Networks," Hot-ICE 2011
https://www.usenix.org/conference/hot-ice11/topology-switching-data-center-networks

- We can only have an odd number of stages in Clos topology, is this true?

Only 3 stages.

Related Modules

CSE567M: Computer Systems Analysis (Spring 2013),

https://www.youtube.com/playlist?list=PLjGG94etKypJEKjNAa1n_1X0bWWNyZcof

CSE473S: Introduction to Computer Networks (Fall 2011),

https://www.youtube.com/playlist?list=PLjGG94etKypJWOSPMh8Azcgy5e_10TiDw

Wireless and Mobile Networking (Spring 2016),

https://www.youtube.com/playlist?list=PLjGG94etKypKeb0nzyN9tSs_HCd5c4wXF

CSE571S: Network Security (Fall 2011),

https://www.youtube.com/playlist?list=PLjGG94etKypKvzfVtutHcPFJXumyyg93u

Video Podcasts of Prof. Raj Jain's Lectures,

https://www.youtube.com/channel/UCN4-5wzNP9-ruOzQMs-8NUw

Student Questions