

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/cse570-23/>

Student Questions



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

Student Questions

Google's Data Center



Student Questions

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

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

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

Source: <http://webodyssey.com/technologyscience/visit-the-googles-data-centers/>

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

Cooling Plant



Student Questions

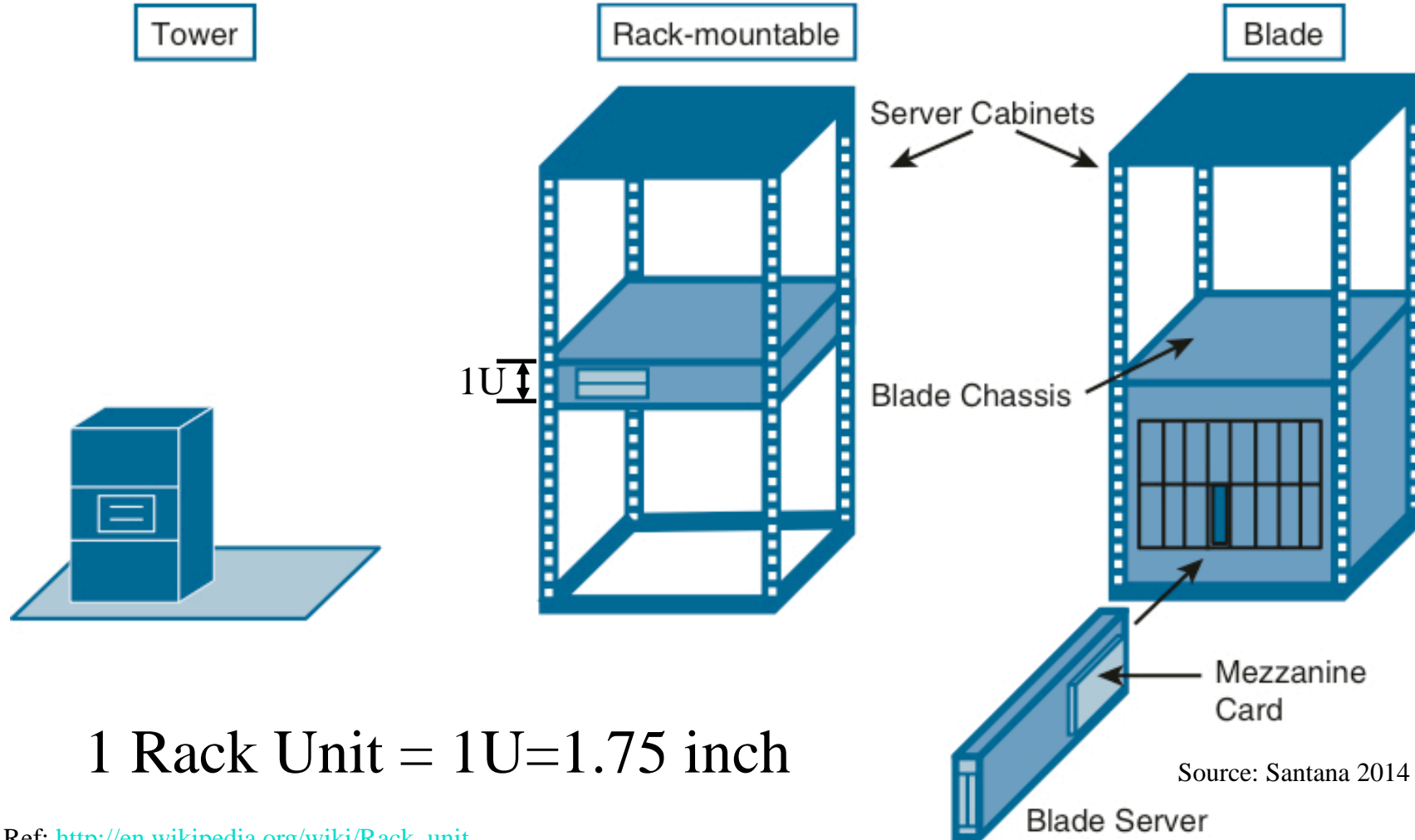
Source: <http://webodyssey.com/technologyscience/visit-the-googles-data-centers/>

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

Servers



1 Rack Unit = 1U=1.75 inch

Ref: http://en.wikipedia.org/wiki/Rack_unit

Ref: G. Santana, "Data Center Virtualization Fundamentals," Cisco Press, 2014, ISBN:1587143240

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

Source: Santana 2014

©2023 Raj Jain

Student Questions

- ❑ Is the NAS server a kind of rack or blade server?

NAS = Network attached storage => Remote storage. Depending upon the size, it could be a rack or blade server. A few terabytes can be easily put on a blade. For Petabytes, you may need a rack.

- ❑ Can a blade server operate without mezzanine cards?

Yes. Mezzanine cards generally provide some extra functions that may not be necessary.

- ❖ When do we use Rack, and when will we use Blade?

A Blade is a small component such as a single computer. A Rack is usually multiple computers.

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/

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

Student Questions

- ❑ Isn't the lower the PUE, the better? Why did the quiz answer say 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>

- ❑ Is PUE always measured in Megawatts?

PUE is a ratio. It has no unit.

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/

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

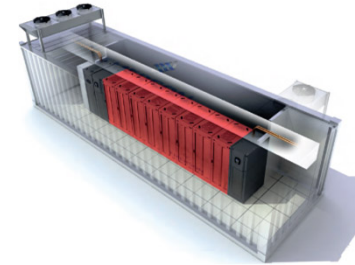
©2023 Raj Jain

Student Questions

- ❑ Does most power consumption come from running the cooling/plants or the servers?

*Cooling requires more power.
Computer systems are becoming energy efficient, faster, smaller, and better. Cooling systems have yet to progress at that rate.*

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

- ❑ Containerized data centers are convenient, but are there project-specific reasons customers might want to create their own data center over choosing convenience?

Of course!

Unstructured Cabling



Student Questions

Source: <http://webodyseum.com/technologyscience/visit-the-googles-data-centers/>

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

Structured Cabling



Student Questions

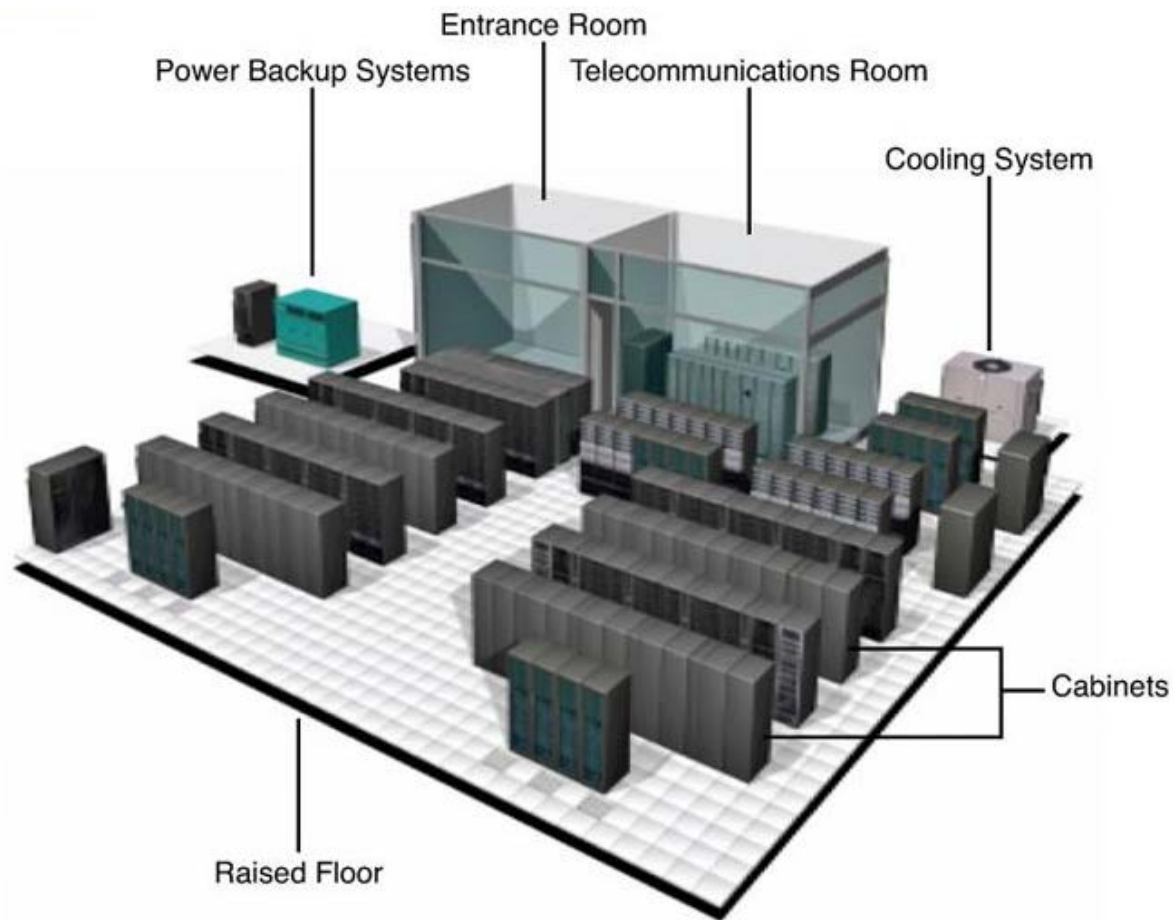
Source: <http://webodyssey.com/technologyscience/visit-the-googles-data-centers/>

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

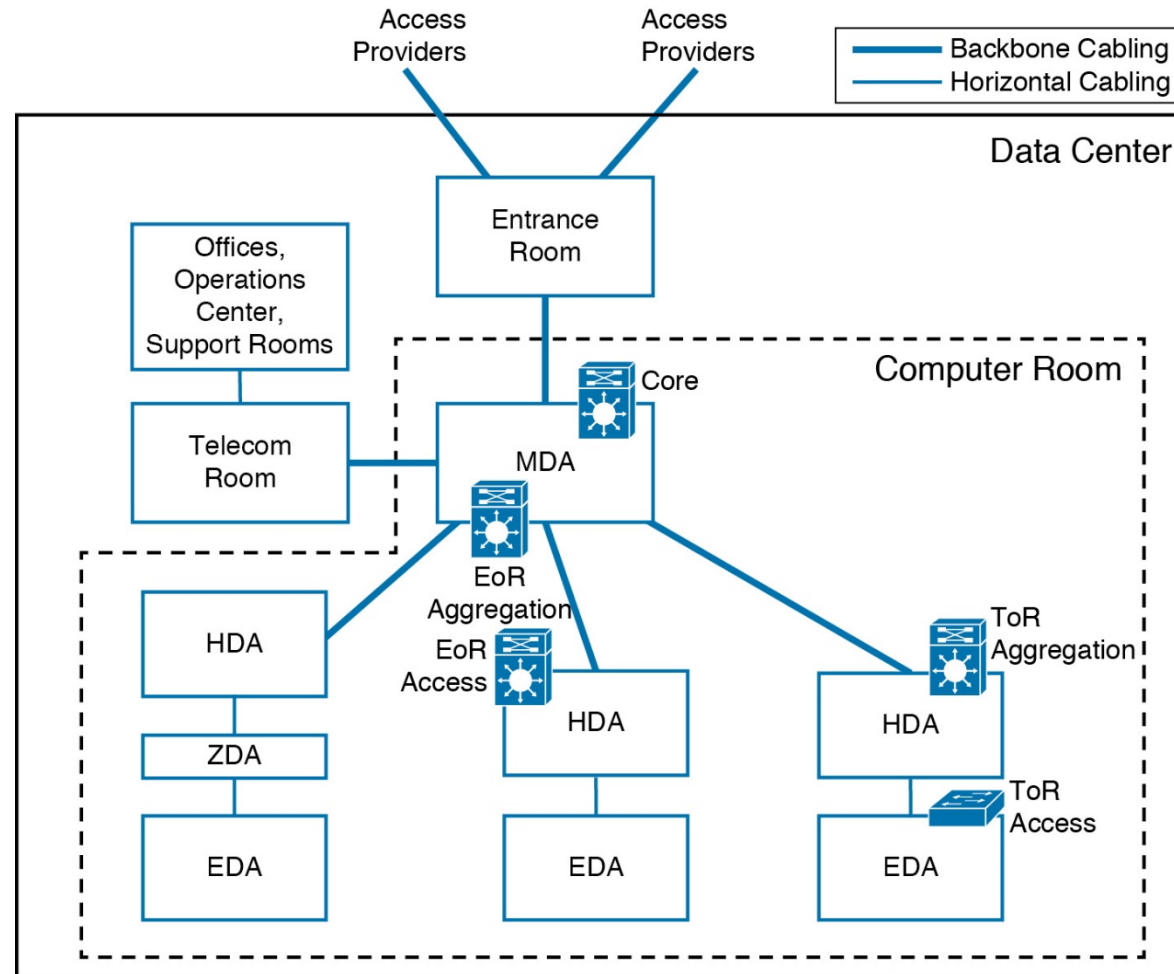
Data Center Physical Layout



Student Questions

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 a hierarchy. So that the networking speeds between the groups can be 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

- ❑ Do access providers own their data centers? Or share with others that all buy space at a private data center?

The trend is towards sharing and using clouds.

Source: Santana 2014

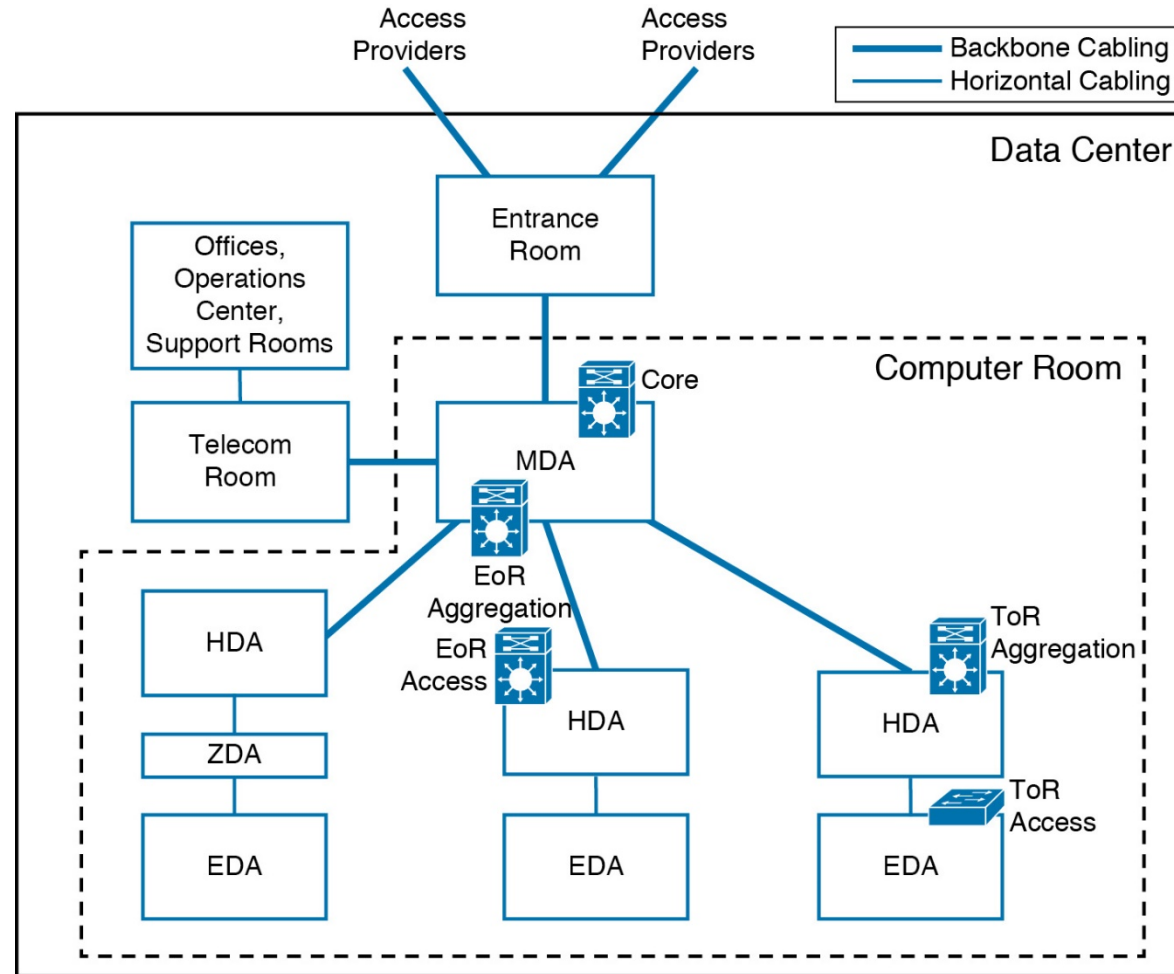
Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

ANSI/TIA-942-2005 Standard

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



Student Questions

- ❑ Do the cables within this standard layout use optic fiber for any part (like the backbone cables)?
Yes. Many cables, even in the racks, could be optical fibers.

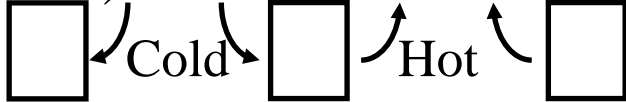
Source: Santana 2014

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 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 aisles. 
- ❑ Zone Distribution Area (ZDA): Optionally between HDA and EDA.
- ❑ Backbone Cabling: Connections between MDA, HDA, and

Entrance room

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

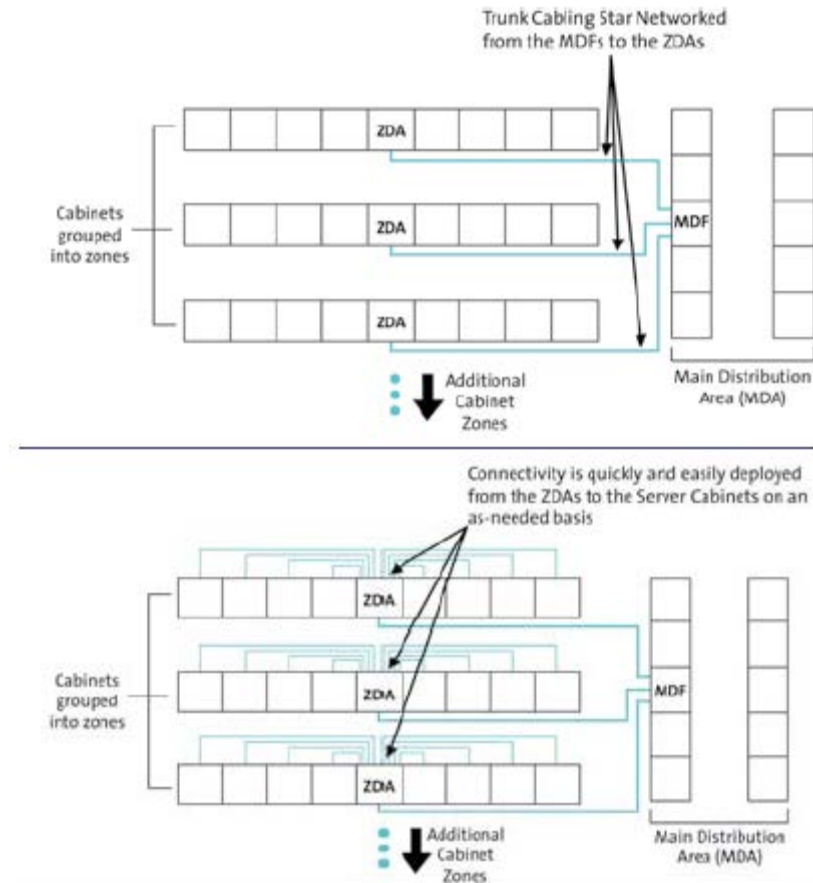
Student Questions

- ❑ What are the purposes of?
The question needs to be completed.
- ❖ Can you explain the difference between HDA's and EDA's? Are they not the same active equipment?

EDAs are the end-user servers.

HDAs connect or provide services to EDAs.

Zone Distribution Area



- ❑ 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>

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

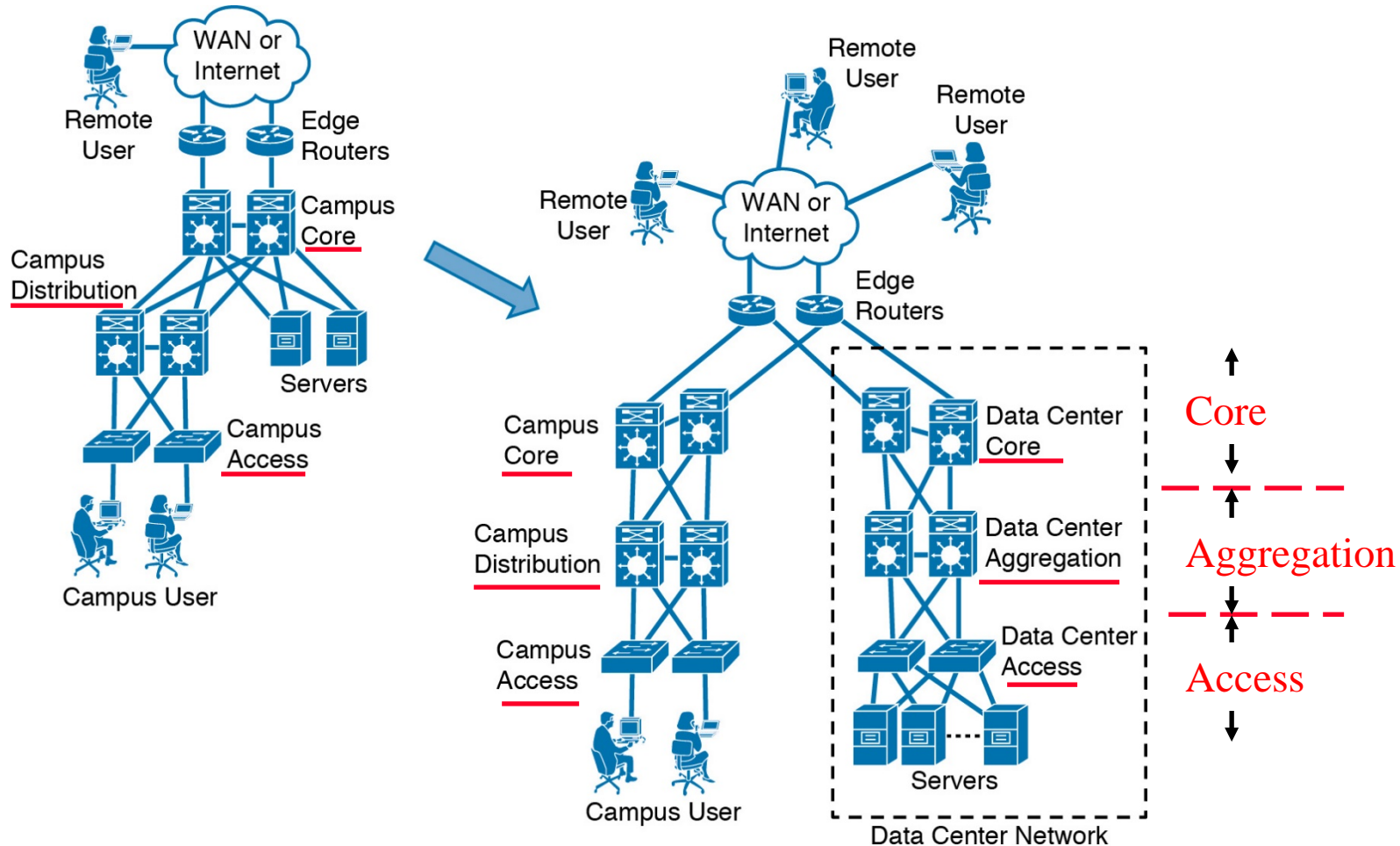
Student Questions

- ❑ Does ZDA here serve for VLAN?

VLANs can be used with or without ZDA. There is no relationship.

Data Center Network Topologies: 3-Tier

Core, Aggregation, Access



Student Questions

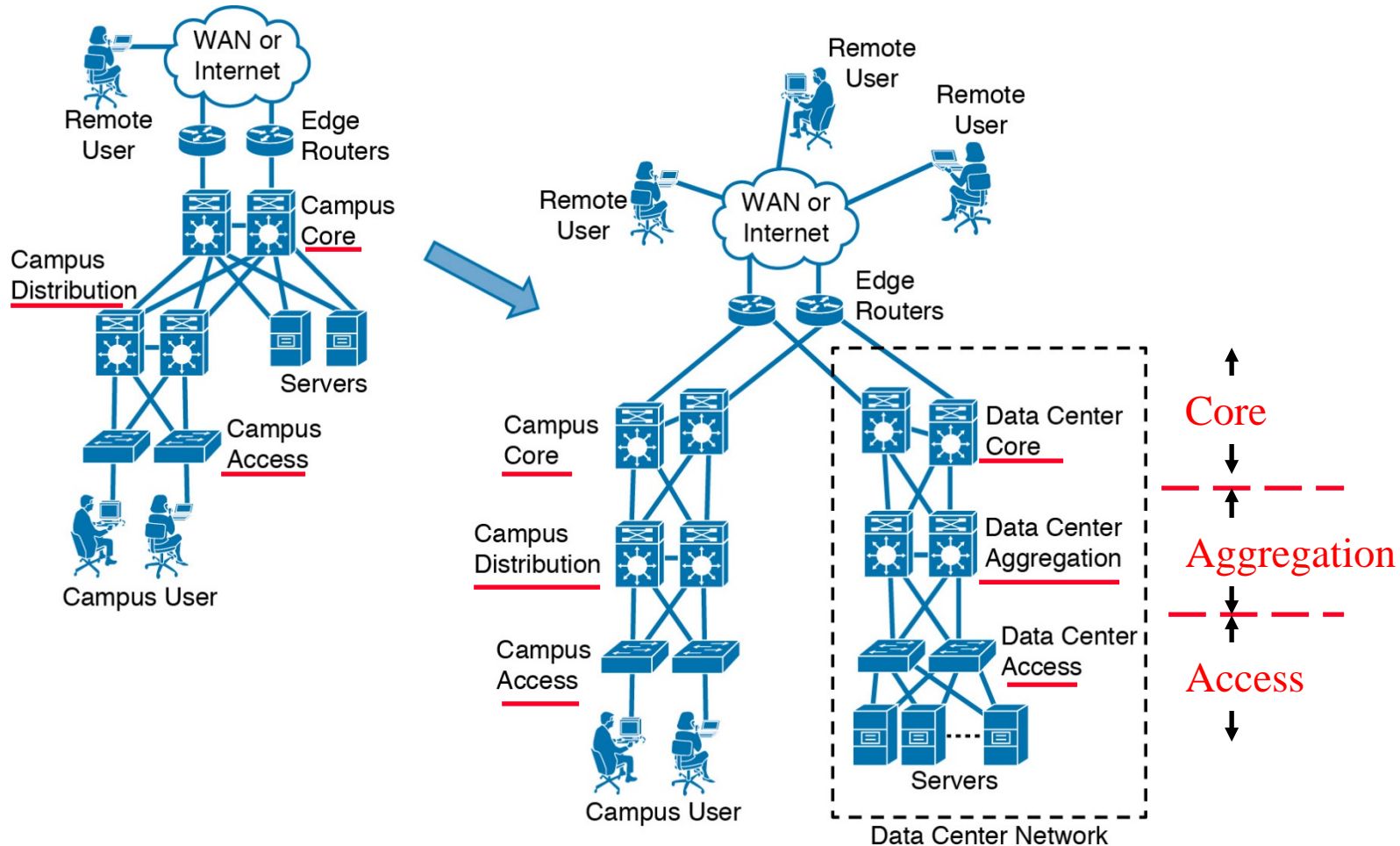
- ❑ Core, Aggregation, Access? Since the video doesn't show where you look at, please explain one more time.

See updates in red. Access starts from users.

- ❑ Are there any new trends in network topologies with the increase of better and faster computing systems?
Fat trees and two-layer architectures are discussed towards the end of this module.

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.

- ❑ How would an SDN impact the core layers architecture?

SDN is a management technique. It does not change the architecture.

- ❑ What key factors should organizations consider when choosing a data center topology?

Size is the key. Topology, as described here, is for large data centers. You can reduce the number of layers for small datacenters to one or two.

3-Tier Data Center Networks

- ❑ 20-40 servers per rack. Limited by power/cooling
- ❑ Each server is 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)
⇒ Aggregation switches are L3 switches ⇒ implement routing
- ❑ Aggregation switches connect to 2 core L3 switches
- ❑ Core L3 switches connect to edge routers
- ❑ The core layer forwards data center ingress and egress traffic



Student Questions

- ❖ Will we need to design a network with a core layer on the exam?

No.

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>.

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

3-Tier Data Center Networks (Cont)

- ❑ The aggregation layer is also a place to put middleboxes, such as firewalls and load balancers.
- ❑ The access Layer provides a high number of ports for connectivity.
- ❑ Low Latency: In high-frequency trading market, a few microseconds make a big difference.
⇒ Cut-through switching and low-latency specifications.
- ❑ Each Layer 2 domain is 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.
- ❑ The aggregation layer forwards server-to-server traffic in the data center ⇒ 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."

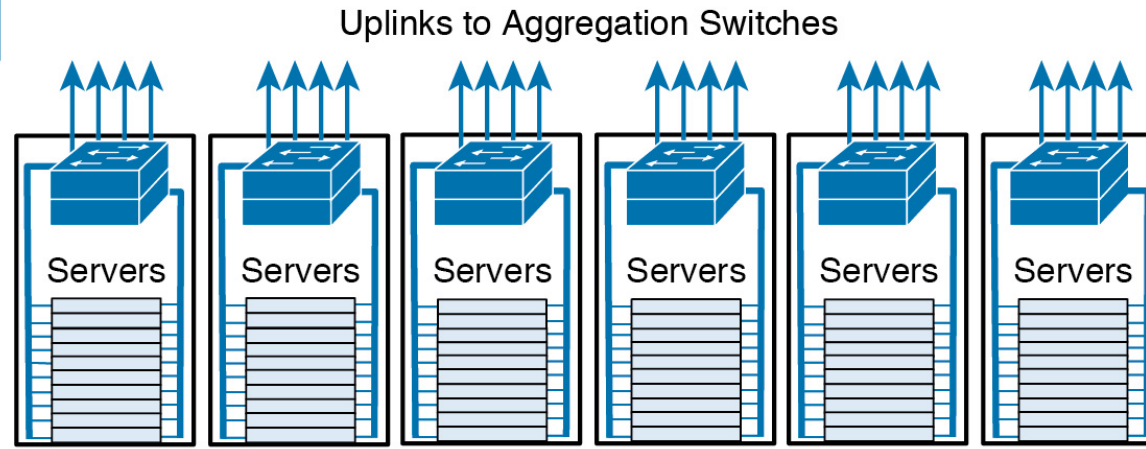
- ❑ Has IoT had any impact on the way access layers are designed?

Yes, the number of devices has increased significantly, requiring carriers to use smaller cells.

Switch Locations

Top-of-Rack

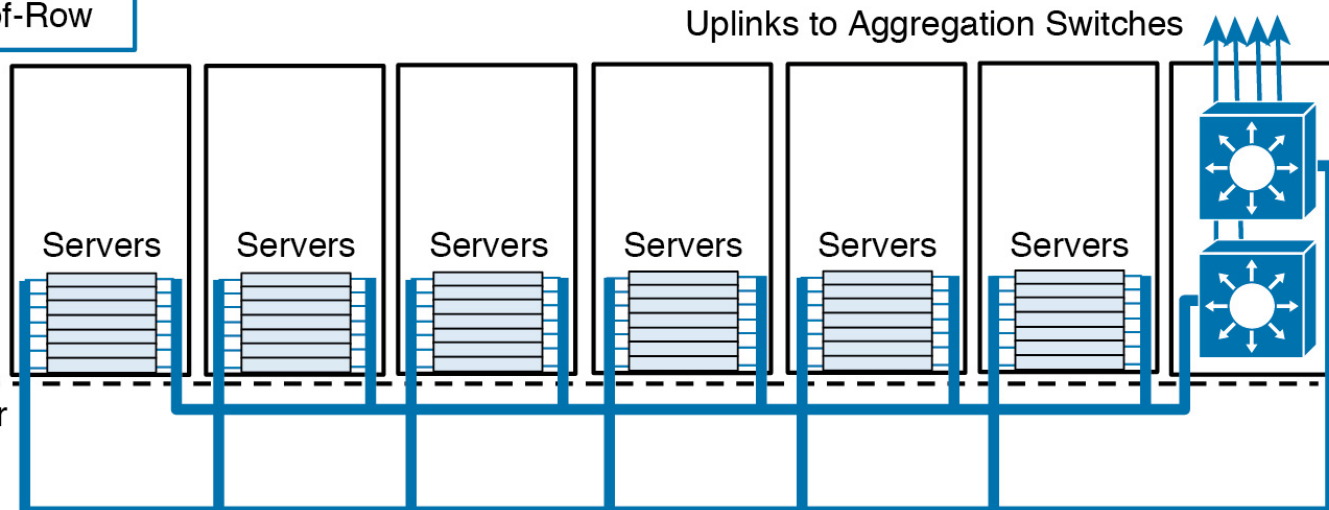
Smaller cable between servers and switches
Network team has to manage switches on all racks



Raised Floor

End-of-Row

All network switches in one rack



Raised Floor

Source: Santana 2014

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

Student Questions

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

“All switches on ToR” is false.

- EOR is the end of the rack or the end of the 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.

ToR vs EoR

□ ToR:

- + Easier cabling
- - If a 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 a complete Rack upgrade

□ EoR:

- - Longer cables
- + Servers can be placed on any rack
- + Ports can easily be added, upgraded

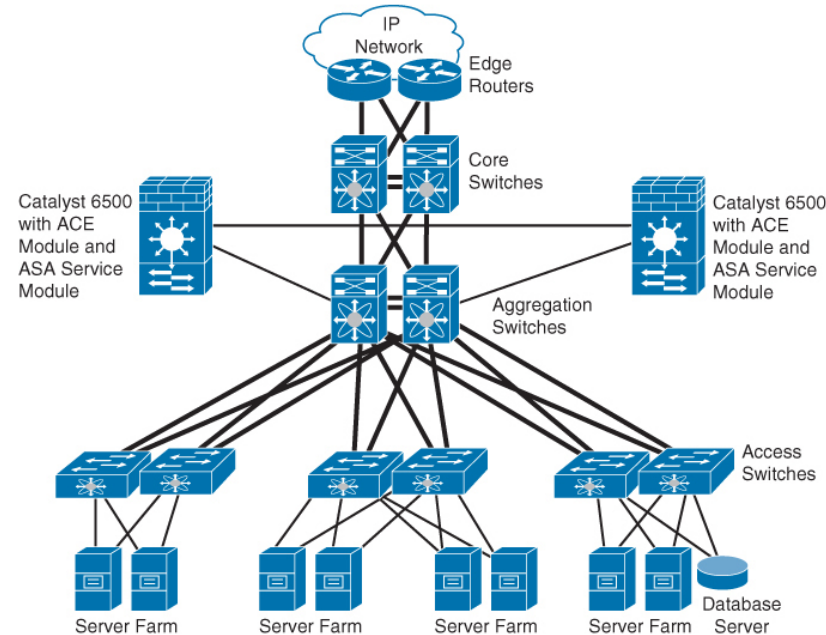
Student Questions

- You stated that ToR is the favored solution. Is that because optimizing cable latency is the most critical factor?

Easier cabling is the main reason.

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

- ❑ Failure of a single link can reduce the available bandwidth by half
- ❑ With more than two aggregation switches, the spanning tree becomes unpredictable in case of certain failures.
- ❑ Two aggregation switches => 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

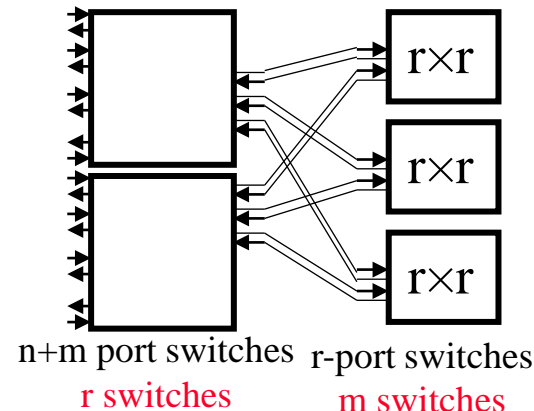
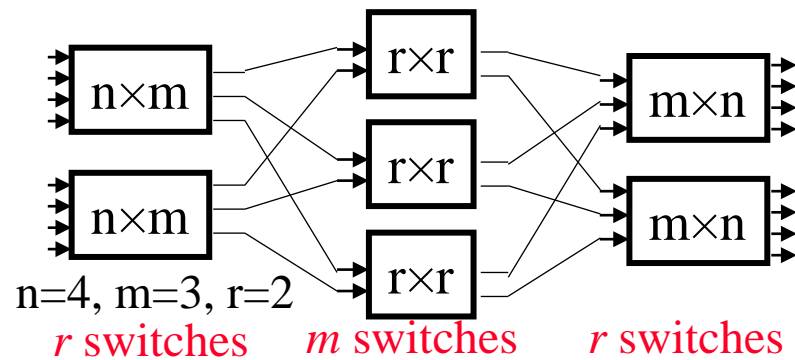
Student Questions

- ❑ Why one link failure would reduce bandwidth by half?
Two cables connect each component. Hence, each provides 1/2th of the bandwidth.
- ❑ Could you clarify why VLANs don't fit with the structure again?

VLAN is a module by itself. It is covered in later modules.

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 \Rightarrow Lower cost (then)
- 3-Stage Clos(n, m, r): ingress ($n \times m$), middle ($m \times r$), egress ($r \times m$)
- Strict-sense non-blocking* if $m \geq 2n-1$. Existing calls are unaffected.
- if $m \geq n$
- Can have any odd number of stages, e.g., 5
- Folded**: Merge input and output in to one switch



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

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain

Student Questions

- Can you please provide an example of 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 piece of paper and fold it in the center.

- How to define the number of input/output ports for the r switches? for the figure on the right? is it $n \times 2$?

Yes. Each bidirectional line pair is one port.

- In the case where $m > n$ what happens after folding? do we need to rearrange?

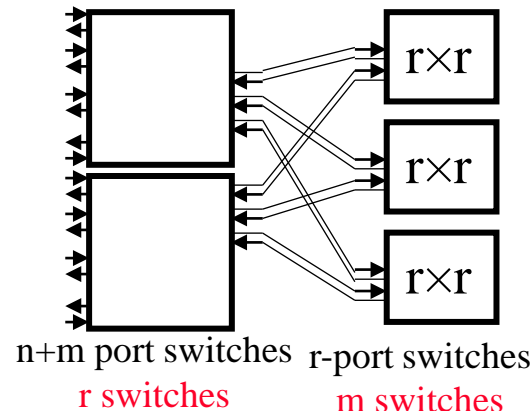
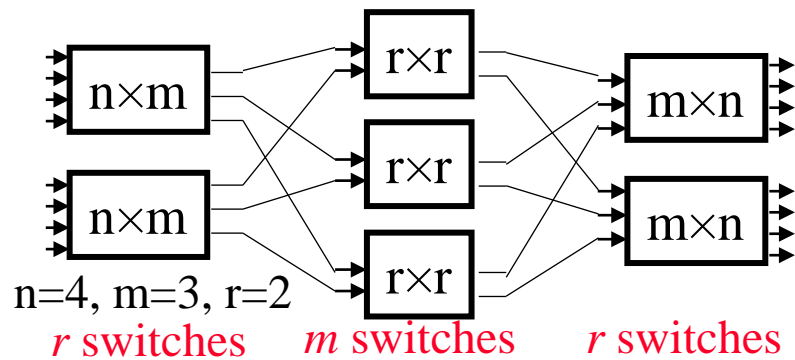
You cannot have $m > n$. If there are n streams coming in, they will have to find a way to select n out of m output ports inside the switch, and so the extra ports will be wasted.

- What is "rearrangeably non-blocking"?

An empty input can be connected to an empty output without rearranging existing calls.

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 \Rightarrow Lower cost (then)
- ❑ 3-Stage Clos(n, m, r): ingress ($n \times m$), middle ($m \times r$), egress ($r \times m$)
- ❑ *Strict-sense non-blocking* if $m \geq 2n-1$. Existing calls are unaffected.
- ❑ *Rearrangeably non-blocking* if $m \geq n$
- ❑ Can have any odd number of stages, e.g., 5
- ❑ **Folded**: Merge input and output in to one switch



Student Questions

- ❑ What is meant by the number of stages in the Clos Network? Is this just the stages of stacked switches?

Number of vertical stacks in the unfolded form (See left figure)

- ❖ Is the primary advantage of a Clos network to transform multiple cheap switches into "larger" ones?

Yes.

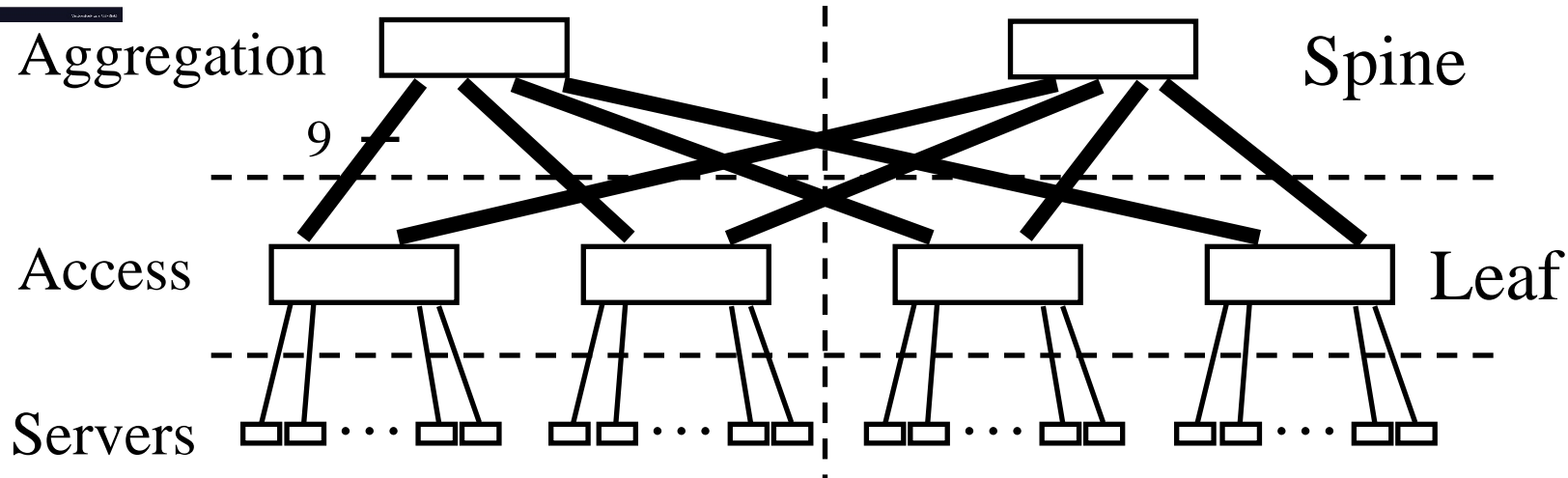
Homework 3A

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

Student Questions



Fat-Tree DCN Example



- ❑ 6 identical 36-port switches. All ports 1 Gbps. 72 Servers.
- ❑ Each access switch connects to 18 servers.
9 Uplinks to the 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

Student Questions

- ❑ What do we need to take into consideration when we want to migrate a data center from a 3-tier DCN to a 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 increases 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 the 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 the virtualization module.

Depends on the virtualization level: L2 or L3.

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

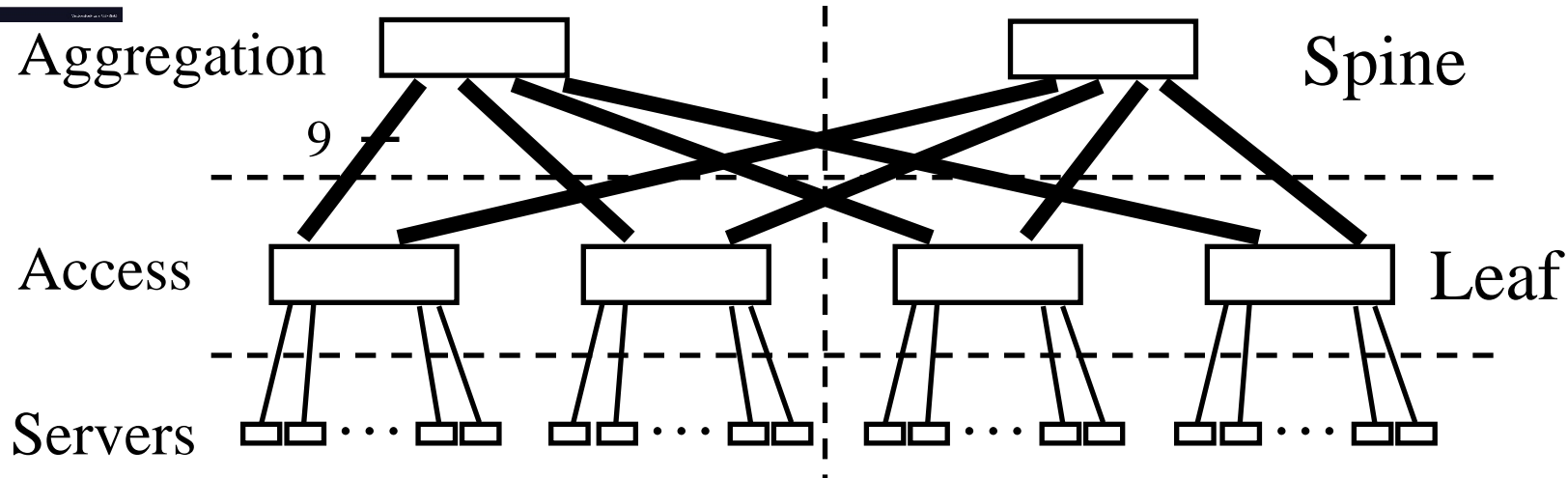
Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain



Fat-Tree DCN Example



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

Student Questions

- ❑ What are those free ports for?
There are no free ports. Thick lines represent 9 ports.
- ❑ What is the "Identical bandwidth at any bisection"?
Why is it 36Gbps?
Look at the three dotted lines. The bandwidth lost is 36-72 Gbps
- ❑ Do the spine switches have all 36 ports available (instead of 18) because they don't "uplink" to anything and instead serve as a connecting backbone?
Only in this diagram. In practice, you will need to save some ports for uplinks.

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

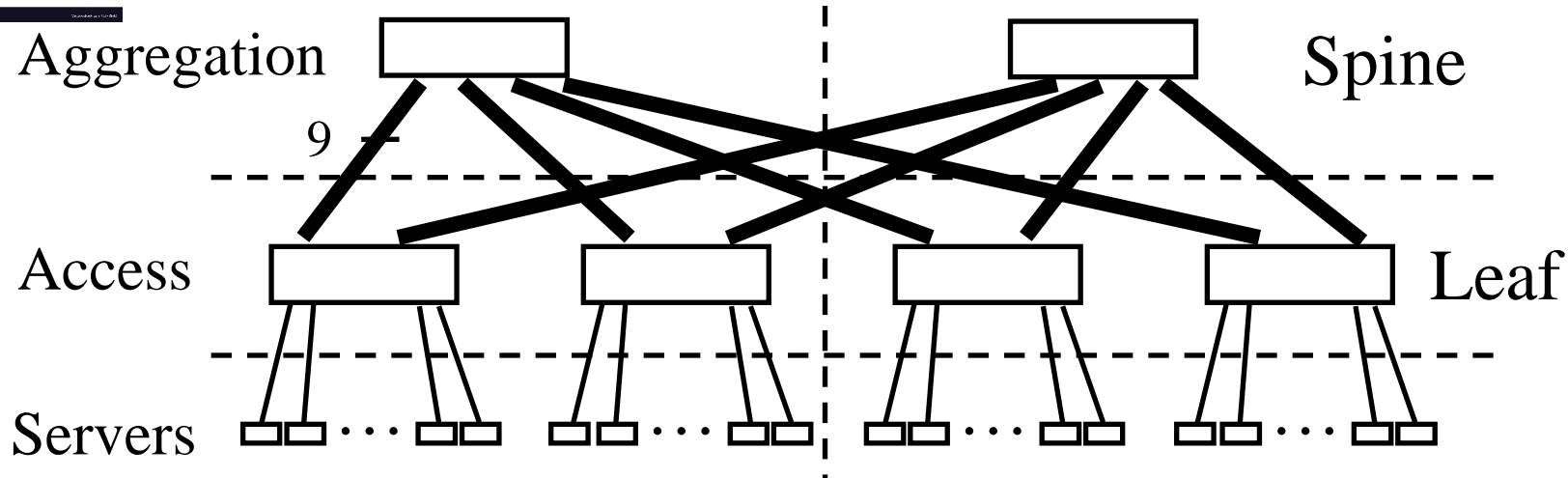
Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-23/>

©2023 Raj Jain



Fat-Tree DCN Example



- ❑ 6 identical 36-port switches. All ports 1 Gbps. 72 Servers.
- ❑ Each access switch connects to 18 servers.
9 Uplinks to the 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/cse570-23/>

©2023 Raj Jain

Student Questions

- ❑ Are all the switches in a data center uniform like they are in the example? Is this setup optimal for throughput?

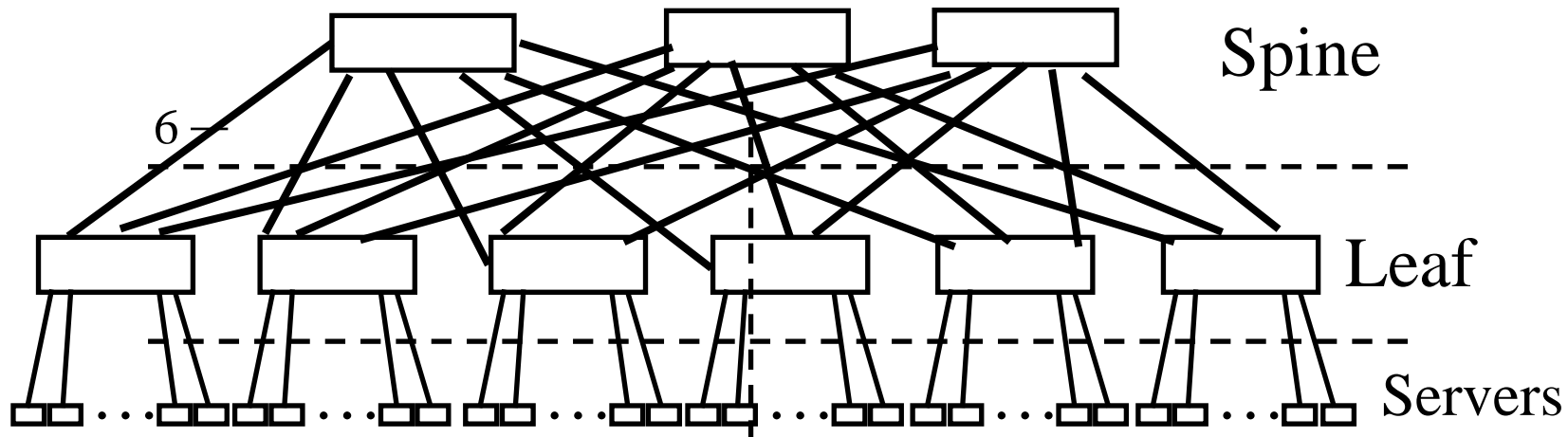
No. There is no standard for connectivity. This is just a proposal by someone.

- ❑ Could you go over again the process of calculating the number of spine and leaf switches?

Sure.

Fat-Tree Topology (Cont)

- Half of the leaf switch ports are toward servers, and the other half toward the spine
- With 36 port switches \Rightarrow 18 ports to the spine
 \Rightarrow 2, 3, 6, 9, 18 spine switches
- Maximum # of spine switches = $\frac{1}{2}$ # of ports on leaf switches



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

Student Questions

- Is the leaf switch always TOR? what about the spine switch?

All switches can be ToR or EoR.

- Can we always assume identical switches in the test?

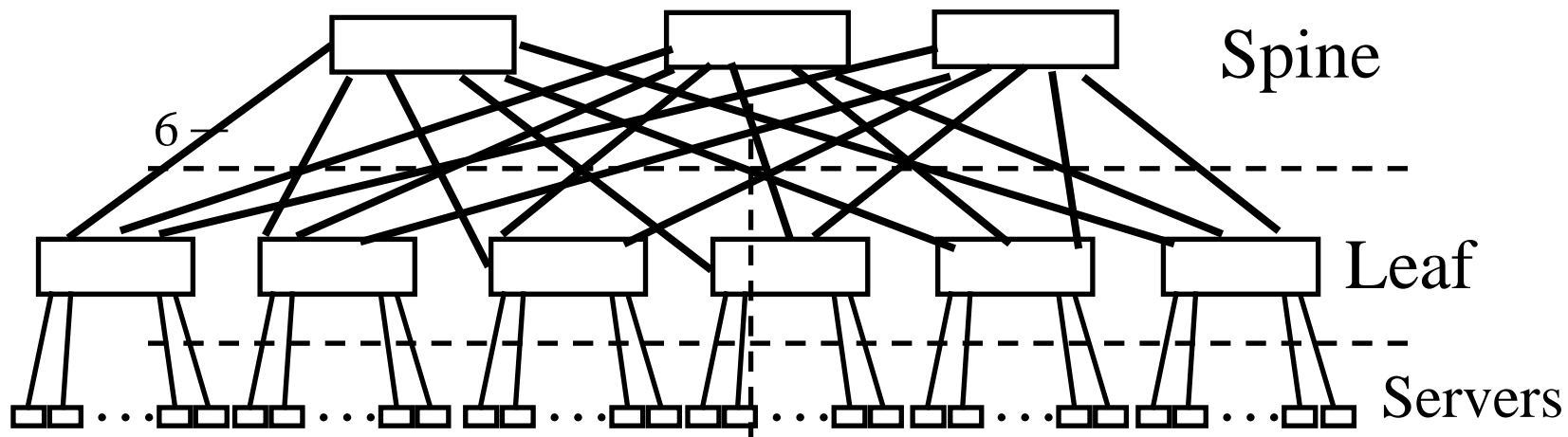
Yes. All access switches are identical. All aggregation switches are identical.

- Many of the online materials said that fat tree has 3 layers (aggregation, access, core). Why here we only say 2 layers (spine and leaf) ?

There is no standard. This is just an example.

Fat-Tree Topology (Cont)

- ❑ Half of the leaf switch ports are toward servers, and the other half toward the spine
- ❑ With 36 port switches \Rightarrow 18 ports to the spine \Rightarrow 2, 3, 6, 9, 18 spine switches
- ❑ Maximum # of spine switches = $\frac{1}{2}$ # of ports on leaf switches



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

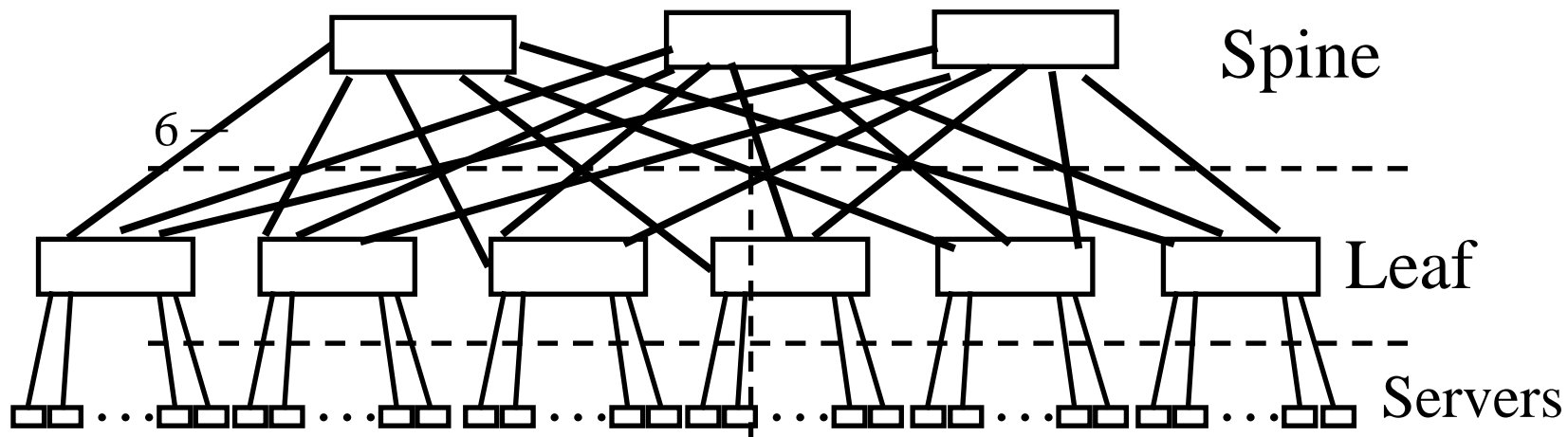
Student Questions

- ❑ What does it mean for "topology with 6"? The video said it means three spine switches and six ports going up; what's the relationship between these numbers, and why only six ports are going up?

See the updated figure on the left. Each spine has six 6-Gbps links going down to six leaves.

Fat-Tree Topology (Cont)

- ❑ Half of the leaf switch ports are toward servers, and the other half toward the spine
- ❑ With 36 port switches \Rightarrow 18 ports to the spine \Rightarrow 2, 3, 6, 9, 18 spine switches
- ❑ Maximum # of spine switches = $\frac{1}{2}$ # of ports on leaf switches



- ❑ Largest configuration with n -port switches: $\frac{n^2}{2}$ servers can be connected using $n + \frac{n}{2}$ switches.

Student Questions

- ❖ Can we always assume that for n -port switches, we can have $\frac{(n^2)}{2}$ servers and $\frac{(n+n)}{2}$ switches for the leaf and $\frac{n}{2}$ switches for the spine?

One-half of leaf ports go down, and the other half go up. Depending upon the number of spine switches, you should divide the uplinks from leaves. E.g., 12 ports, six up, three each to 2 spines, or two to 3 spines.

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

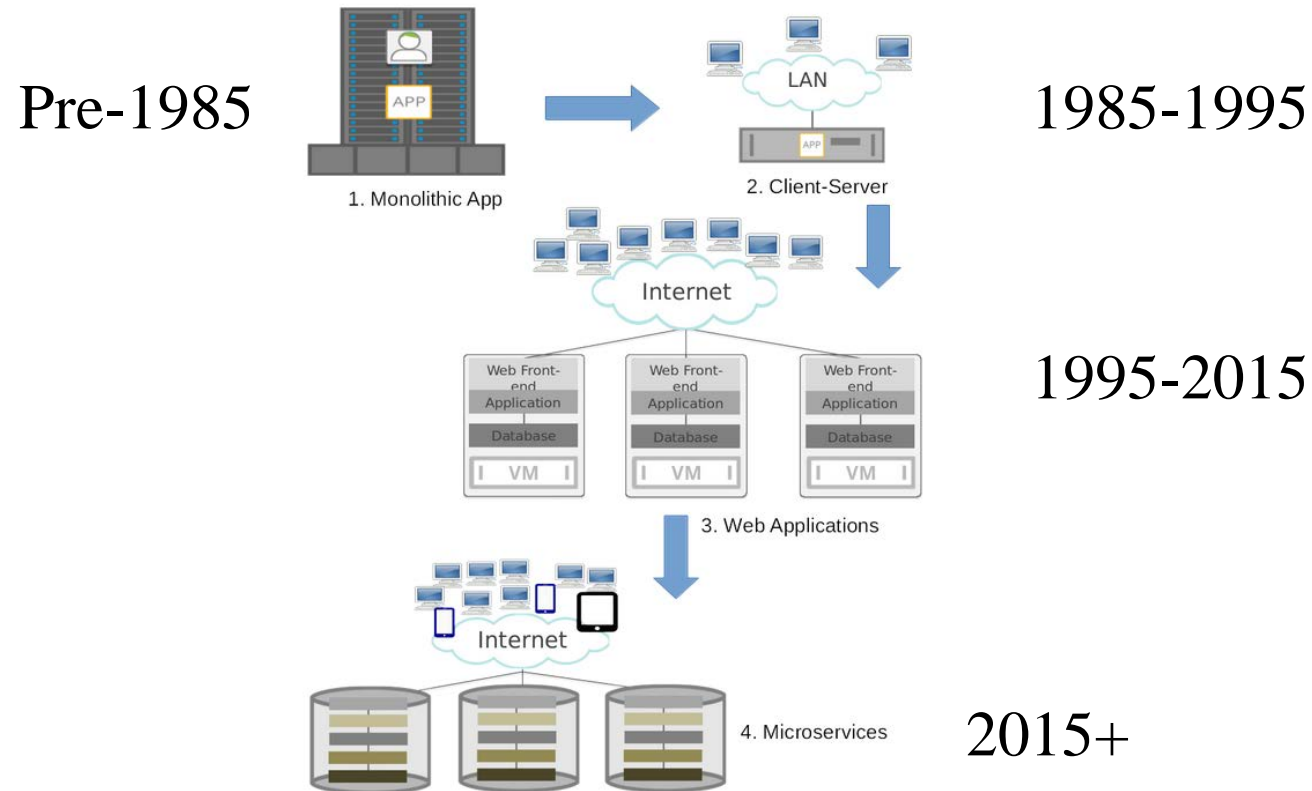
- Question three, "in all" means (spine + leaf)?

Yes

- Can you mark the chapters we need to read?

There are no textbooks for this course. Some Safari books are included in the reading list. The chapters are marked there.

Evolution of Applications



- ❑ 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.

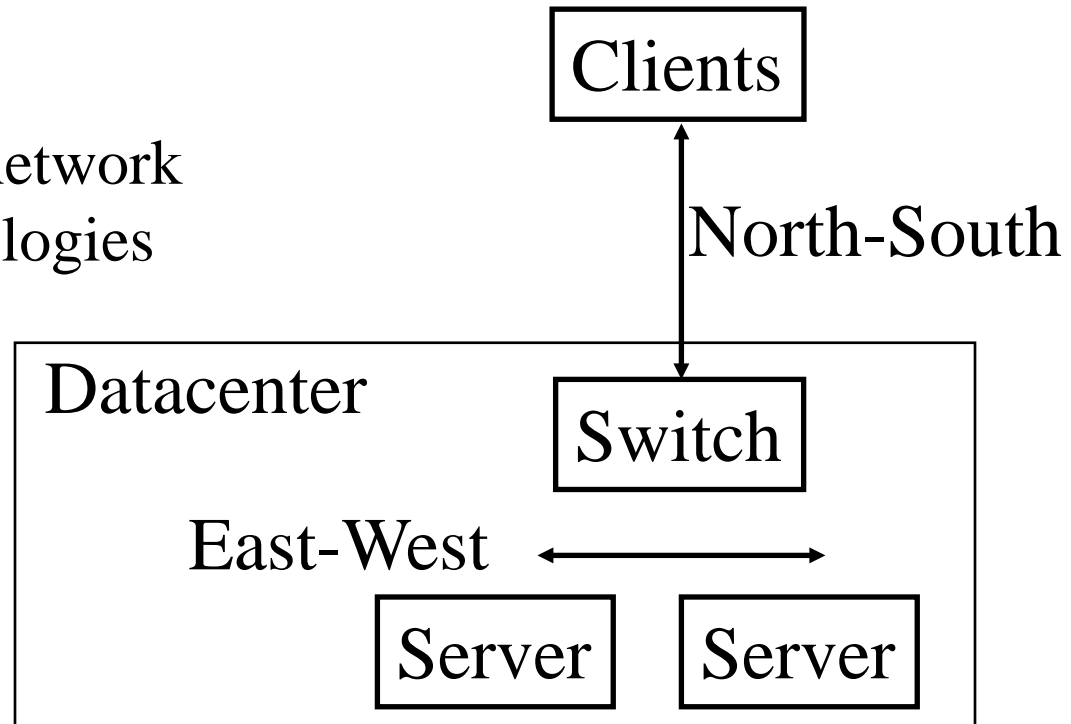
Student Questions

- ❑ 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 toward traffic between servers for big data analysis
 - ⇒ East-West traffic
 - ⇒ Requires flatter network
 - ⇒ Fat-tree like topologies



Student Questions

- ❑ Suppose there's traffic between two servers from different data centers like one server is in California, and the other is in New York. Is this traffic north-south traffic or east-west traffic?

If all data centers are on one Ethernet, their physical location does not matter. Server-to-server traffic is still inside one virtual data center and is, therefore, east-west.

- ❑ Do East-West topologies provide any security benefits?

East-West refers to the traffic direction, not the topology. Compare this to North-South traffic.

- ❖ Can we go over East-west vs. North-South (incast?) and the uses/purposes of each?

North-south traffic going in/out. East-west traffic circulates inside.

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 are 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

Student Questions

- ❑ Will some data packets be lost? How will it be fixed if it is lost?
- ❑ Besides VXLAN, are companies using some proprietary protocols for server communication?

L4 protocols take care of lost packets.
There are no proprietary protocols.

- ❑ Can you please elaborate more on how hashing is used to select the 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 switches?
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.

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 are 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

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

Student Questions

- ❑ Are the hashes for a particular flow ever cached to speed up routing (so that ECMP is not recomputed for every packet of the same flow)? If so, where is the caching done?

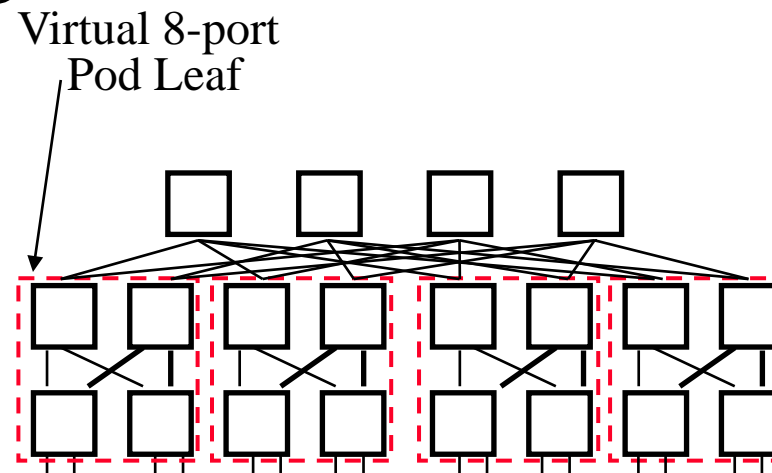
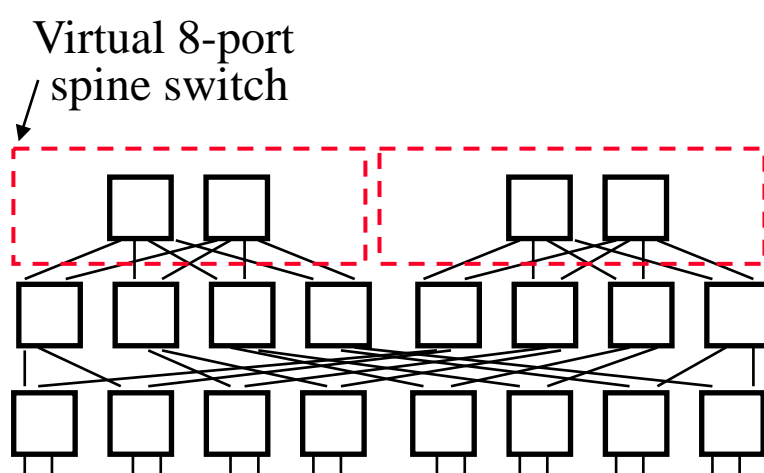
ECMP routers may classify flow packets and direct all flow packets to one link. This is remembered in the router.

- ❑ Does the Same path mean using one leaf and one spine switch? Or can it still have multiple switches?

You can have multiple switches. The exact path means all flow packets go on the same outgoing link.

Variations

- ❑ Higher-speed Inter-Switch Links (ISLs) may be used:
 - 1 Gbps server/10 Gbps ISL, 10 Gbps Server/40 Gbps ISL
 - Reduces the 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



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

Student Questions

- ❑ In a Clos network, increasing the number of ports on switches has a 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.

- ❑ 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 Core, Edge and Aggregation.

[1] [10.1145/1402958.1402967](https://doi.org/10.1145/1402958.1402967).

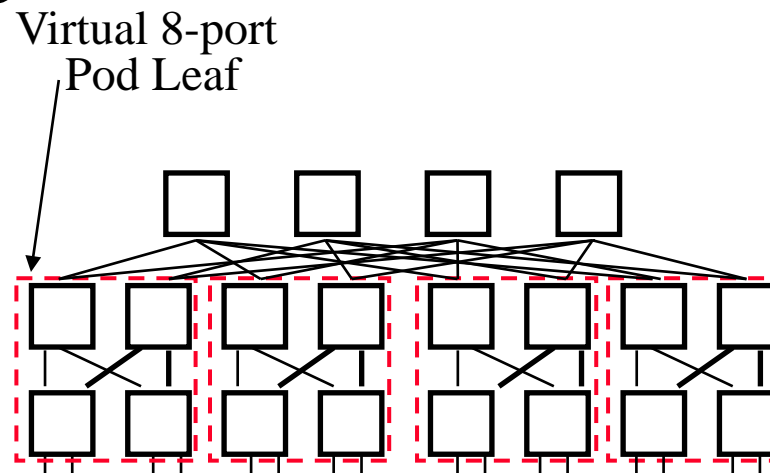
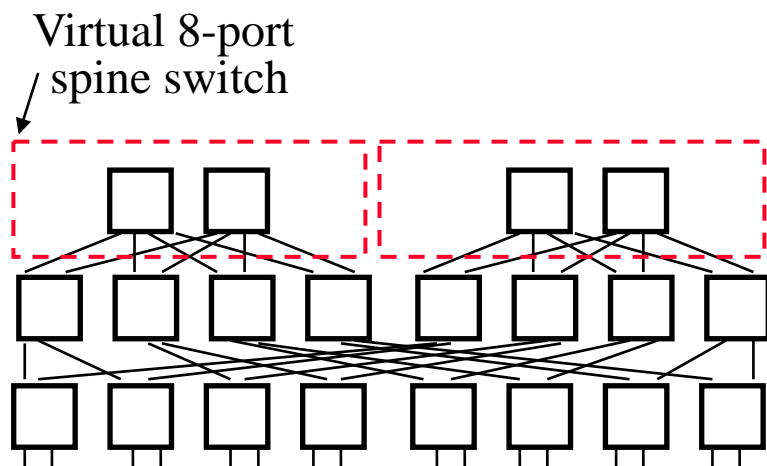
Yes, each tier or switch could be a Clos network.

- ❑ In practice, how many ports are there on the switches in Clos network?

It could be several thousand.

Variations

- ❑ Higher-speed Inter-Switch Links (ISLs) may be used:
 - 1 Gbps server/10 Gbps ISL, 10 Gbps Server/40 Gbps ISL
 - Reduces the 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

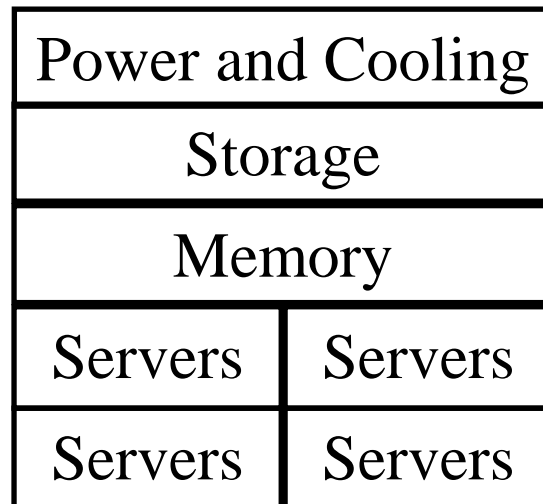


Ref: Dinesh G. Dutt, "Cloud-Native Data Center Networking," O'Reilly Media, Inc., December 2019,
ISBN: 9781492045595, Safari Book.

Student Questions

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 into the networking board on the back
- ❑ Buy complete racks rather than individual servers
- ❑ Being standardized by Open Compute Project (OCP)



Student Questions

- ❑ Is it possible for such an architecture to have such a disadvantage that if an accident occurs in one of the shared resources, for example, the cooling systems break down, will it affect all servers in the same rack?

Yes, this is possible.

- ❑ If this disadvantage exists, how to fix it?

Redundant servers are located in different racks.

- ❑ What is the network architecture for connecting the redundant servers from rack to rack?

Any architecture can be used.

Availability will depend on the resources being shared among redundant servers.

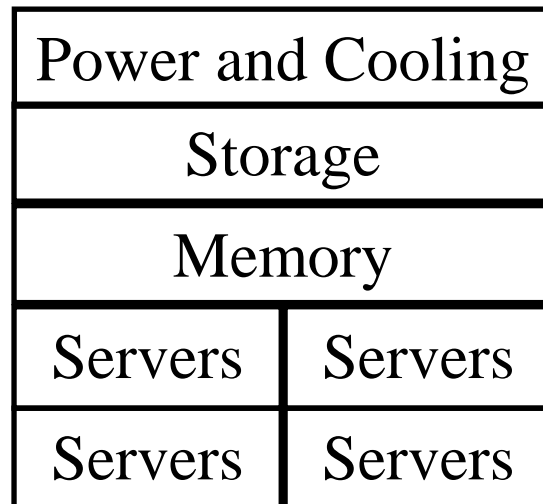
- ❑ What is the difference between storage and memory?

Storage = Disks (Slower)

Memory = Memory (faster)

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 into the networking board on the back
- ❑ Buy complete racks rather than individual servers
- ❑ Being standardized by Open Compute Project (OCP)



Student Questions

- ❑ Why would the large companies want to coordinate on OCP when they usually like to develop their proprietary standards?

Standards sell more.

- ❑ As they share the memory and storage, how do they deal with multiple servers writing to the exact location or file simultaneously? Will it have race conditions?

All shared resources have a race condition and need a queueing/serving discipline.

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 requested and produce money.

- ❑ What is memory DIMM?

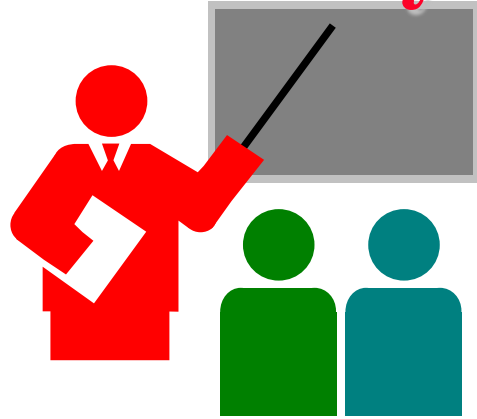
Dual In-line Memory Module have 64 contacts for 64-bit access.

Single in-line memory modules (SIMMs) have 32-bit.



Photo: Wikipedia

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

Student Questions

Is Leaf switch a ToR?

No, it could be ToR or EoR.

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



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., December 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://webdysseum.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



Raj Jain

<http://rajjain.com>

http://www.cse.wustl.edu/~jain/cse570-23/m_03dct.htm

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

- ❖ Can you go over some example true and false questions?

See Quiz 0 in Module 01.

- ❖ What is an example of a strictly point-to-point LAN?

Point-to-Point is a link. Using Ethernet format frame on point-to-point links is what we do today.

- ❖ Would you give us the answer to the homework that we have done?

As announced on Piazza by the TA, he has provided correct answers to each of you on your answer sheet. If you did not submit the answers, please try now and get the correct answer.

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=PLjGG94etKypJWOSPMh8Azcg5e_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

- ❖ How will questions related to the reading list appear on the exam?

They could be any of the three formats: numeric, fill-in-the-blank, or True and False.