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/cse5700-25/

Student Questions

©2025 Raj Jain



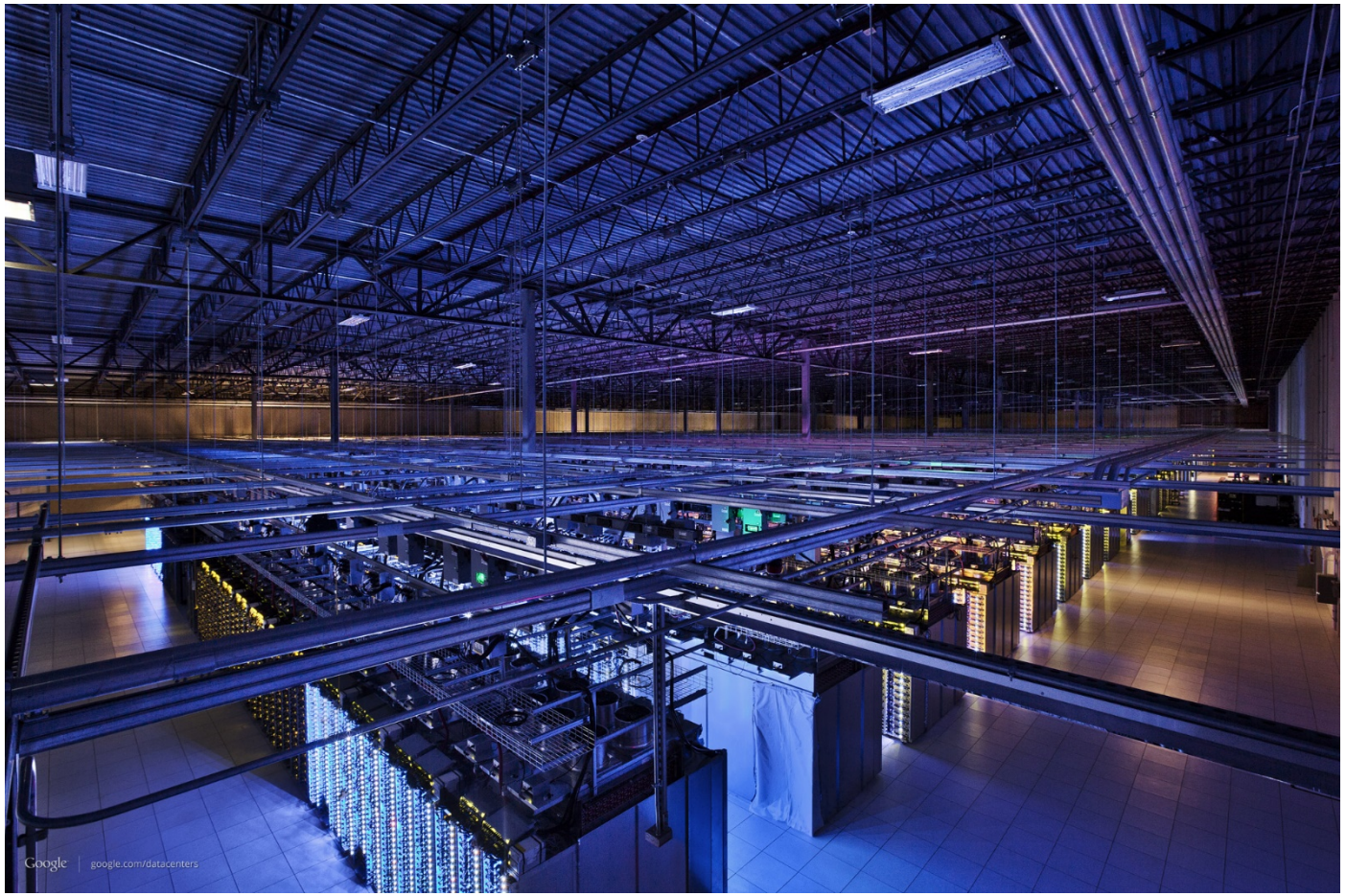
Overview

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

Student Questions

©2025 Raj Jain

Google's Data Center

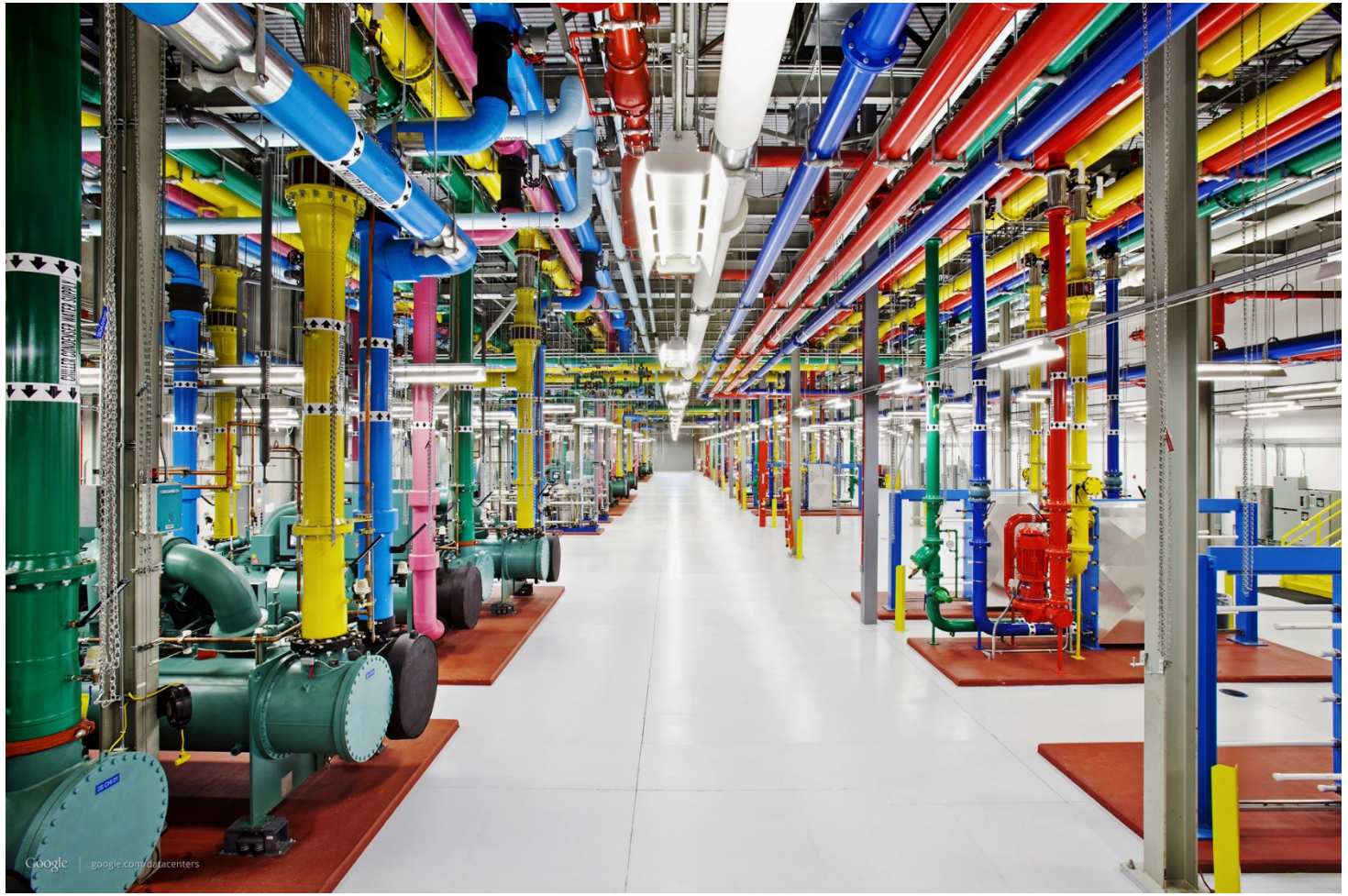


Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/

Student Questions

©2025 Raj Jain

Cooling Plant

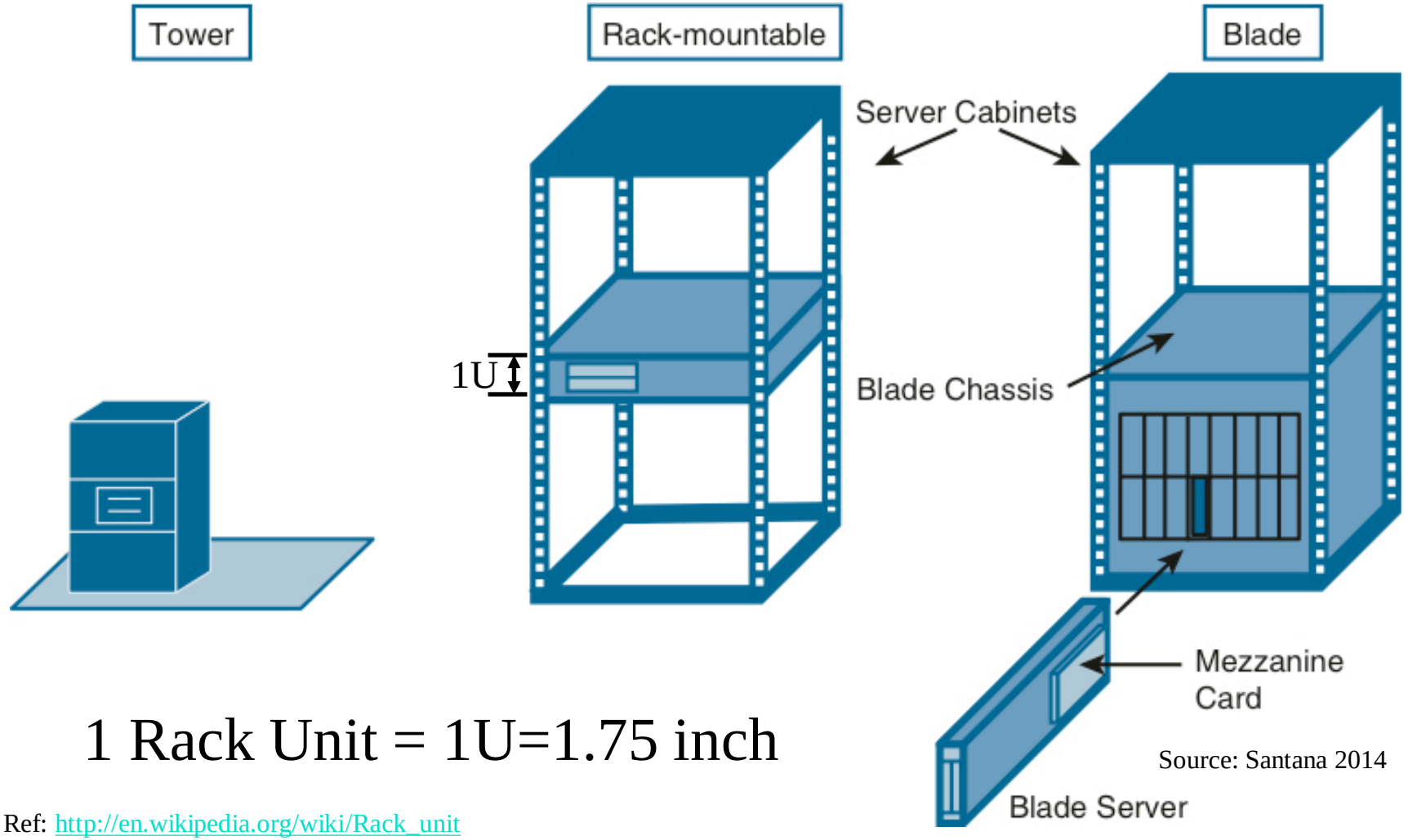


Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/

Student Questions

©2025 Raj Jain

Servers



Source: Santana 2014

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

Student Questions

©2025 Raj Jain

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/

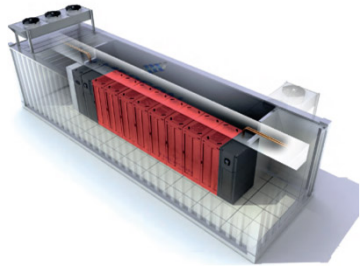
Student Questions

©2025 Raj Jain

Containerized Data Center









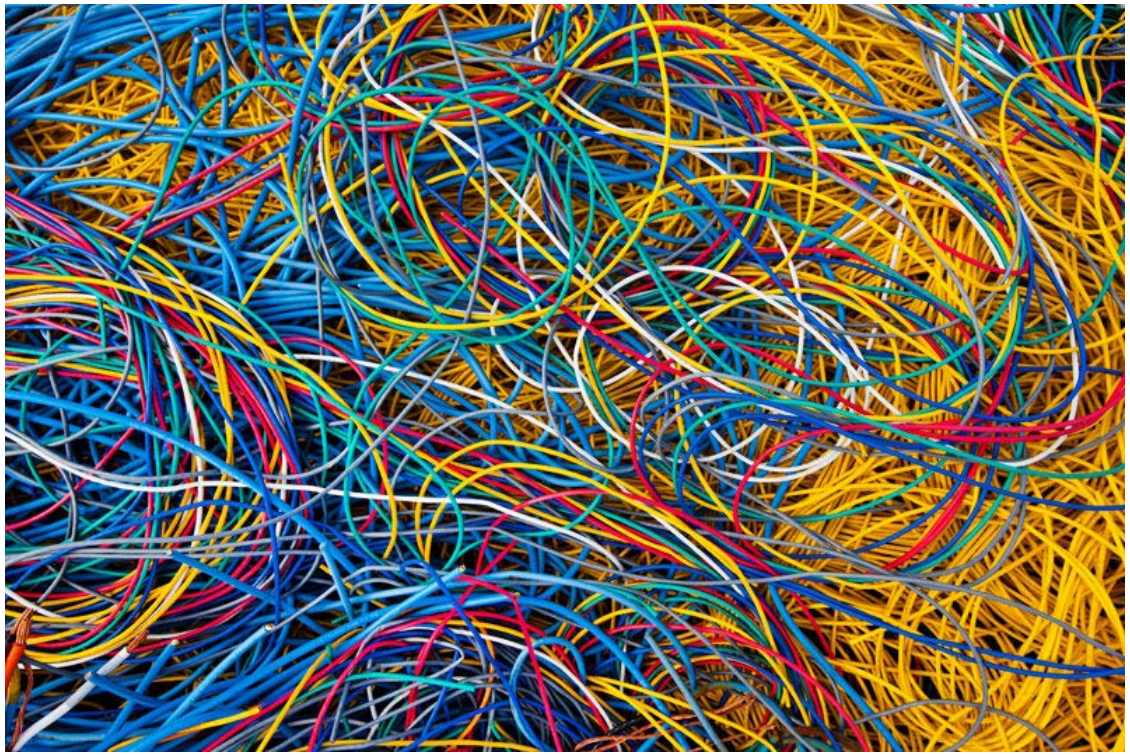
- Ready to Use. Connect to water and power supply and go.
- Built-in cooling. Easy to scale.
  ⇒ Data Center trailer parks.
- Suitable for disaster recovery, e.g., flood, earthquake
- Offered by Cisco, IBM, SGI, Sun/ORACLE,…

Ref: http://www.datacenterknowledge.com/archives/2010/05/31/iij-will-offer-commercial-container-facility/

©2025 Raj Jain

Student Questions

Unstructured Cabling



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

Student Questions

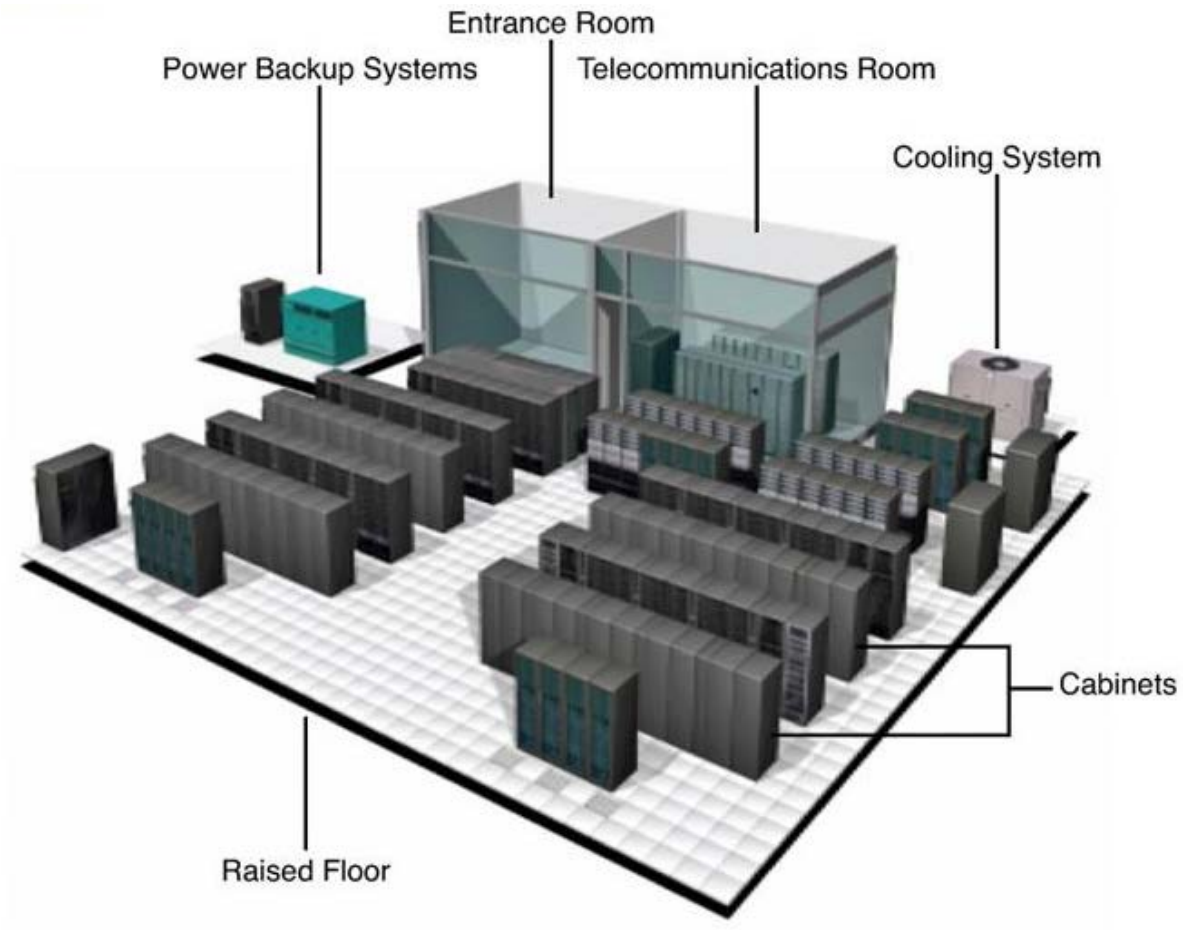
Structured Cabling



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

Student Questions

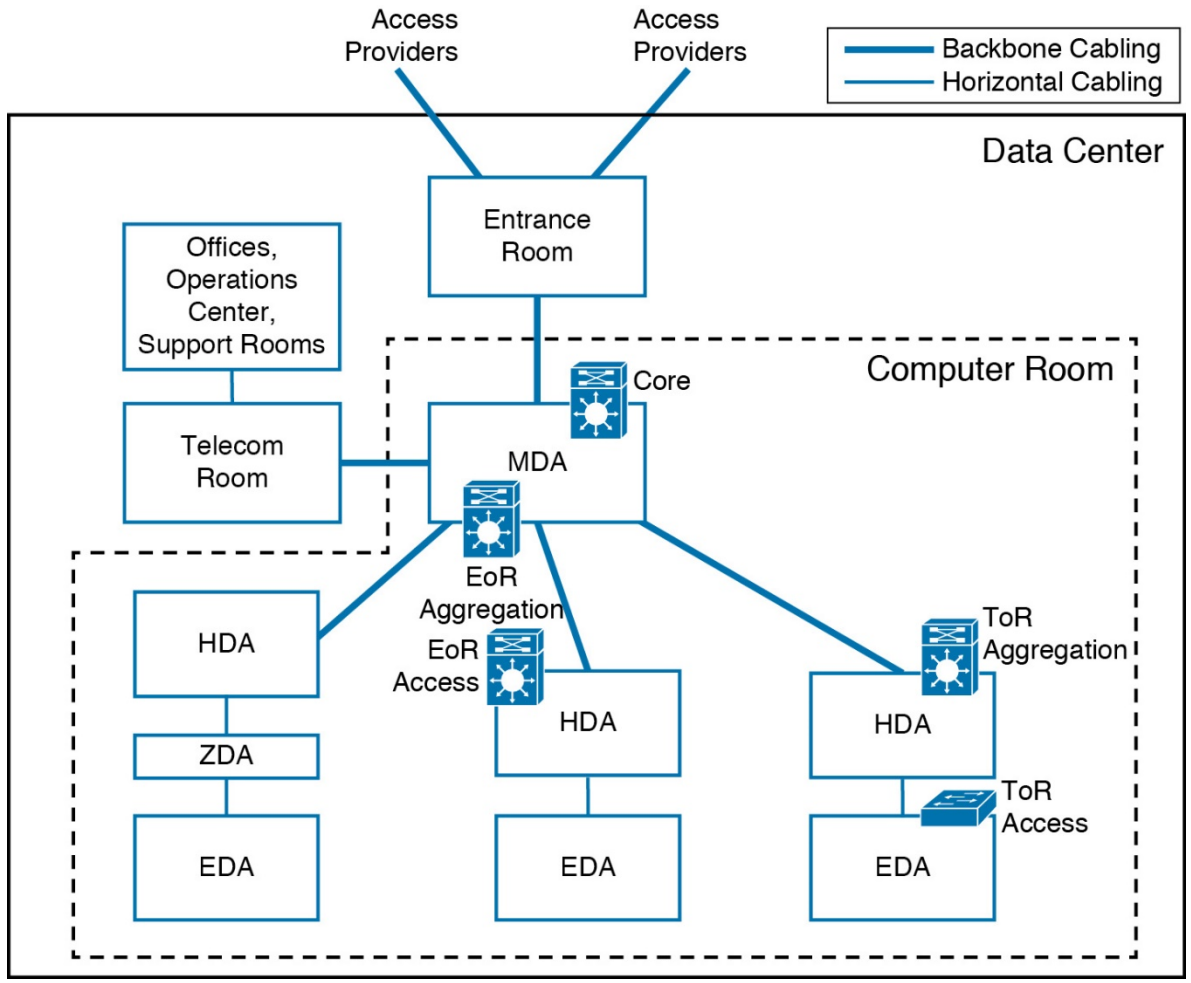
Data Center Physical Layout



Student Questions

ANSI/TIA-942-2005 Standard

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



Source: Santana 2014

Student Questions

©2025 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

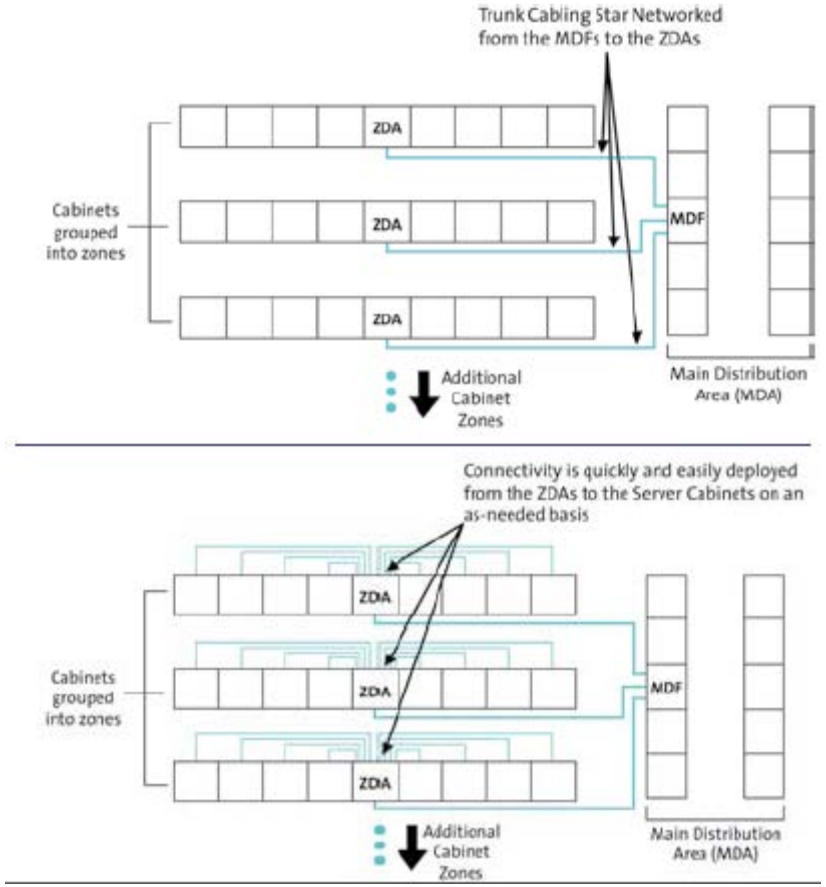
Student Questions

- Why do we need to allow hot air to circulate at all? Wouldn't it be better to make the entire data center a cold chamber and avoid hot aisle complexity? (In case Air-cooling is not enough, liquid cooling can also help)?

Air has to flow in one direction. It will be very hot if the air flows in all directions. Yes, liquid cooling is used too in some data centers.

©2025 Raj Jain

Zone Distribution Area



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

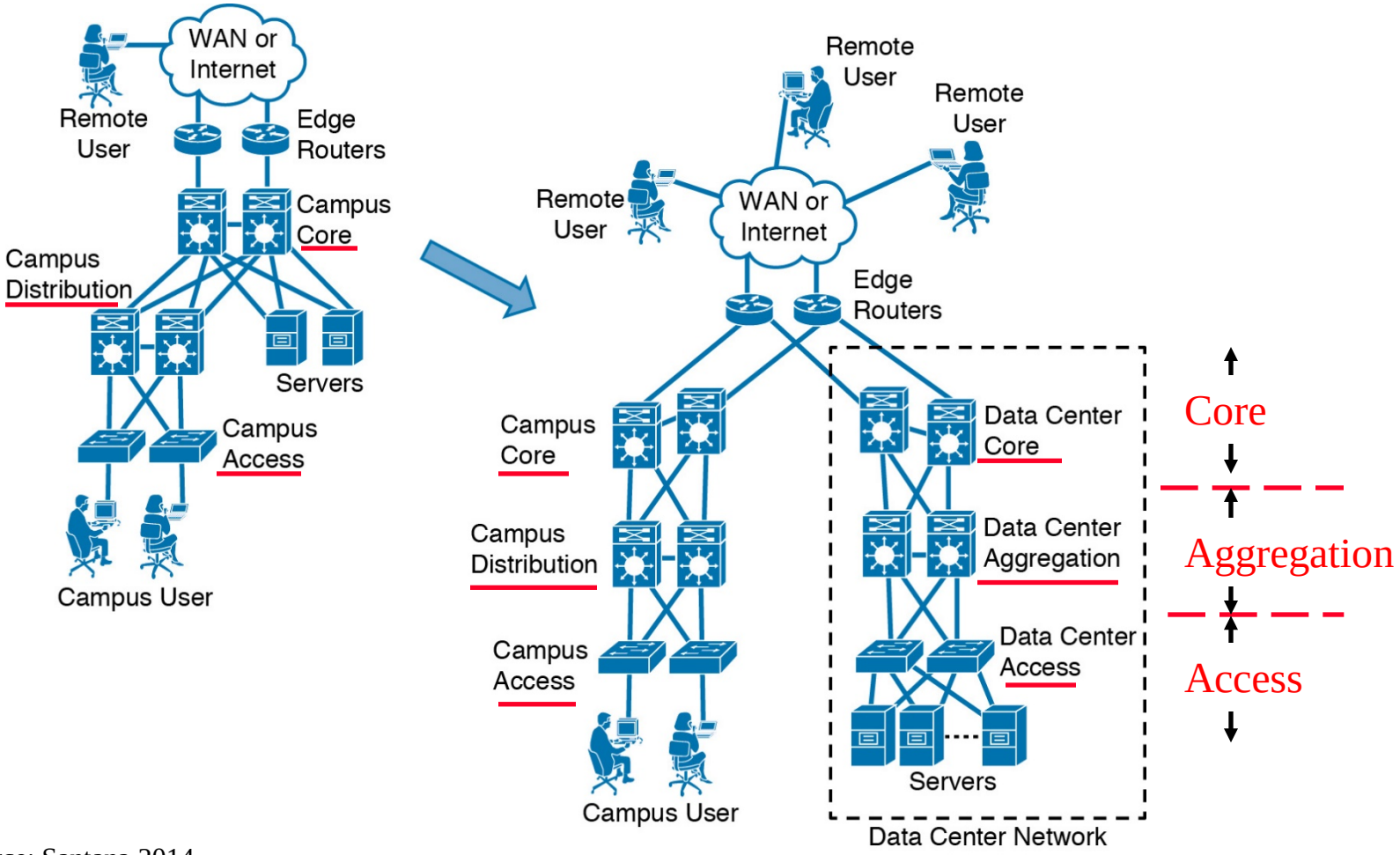
Ref: Jennifer Cline, "Zone Distribution in the data center,"
http://www.graybar.com/documents/***zone-distribution***-in-the-data-center.pdf

Student Questions

©2025 Raj Jain

Data Center Network Topologies: 3-Tier

- Core, Aggregation, Access



Source: Santana 2014

Student Questions

- In the classic 3-tier DC, where do access, aggregation, and core sit, and at which tier does L3 routing typically begin?

They all sit in the same room. L3 routing begins in L3 aggregation switches.

©2025 Raj Jain

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

- The slide mentions "20–40 servers per rack". Question: I checked some online sources, which stated that a typical full-size rack is 42U (https://en.wikipedia.org/wiki/Rack_unit). Is that the actual practical server count is lower? And is this because of power/heat, or are there other reasons?

Some servers may require more or less than 1U of space.

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*.

©2025 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

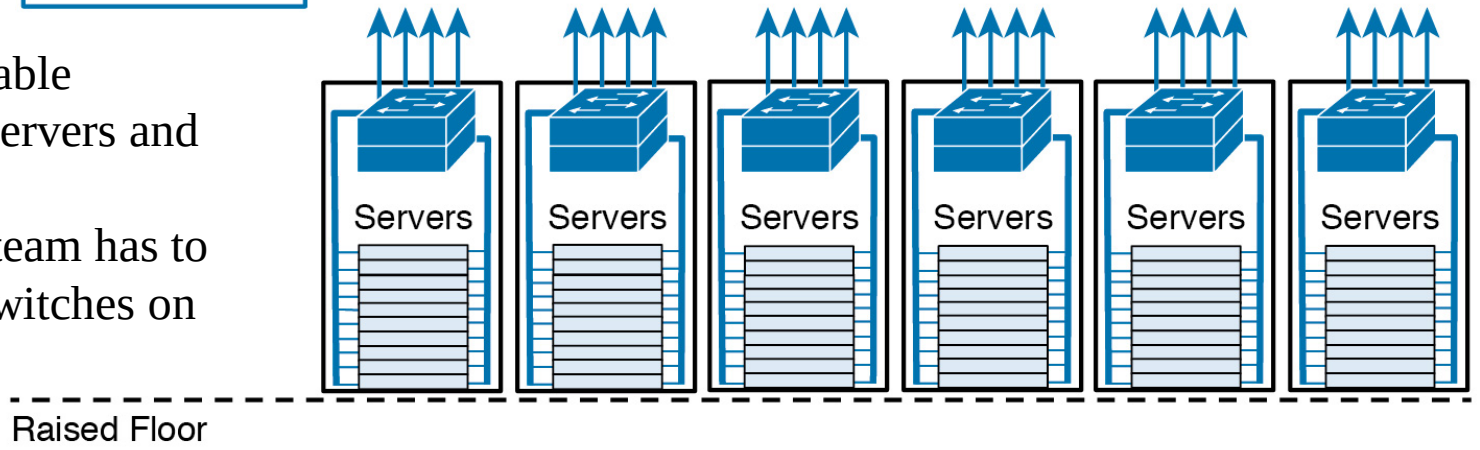
©2025 Raj Jain

Switch Locations

Top-of-Rack

Uplinks to Aggregation Switches

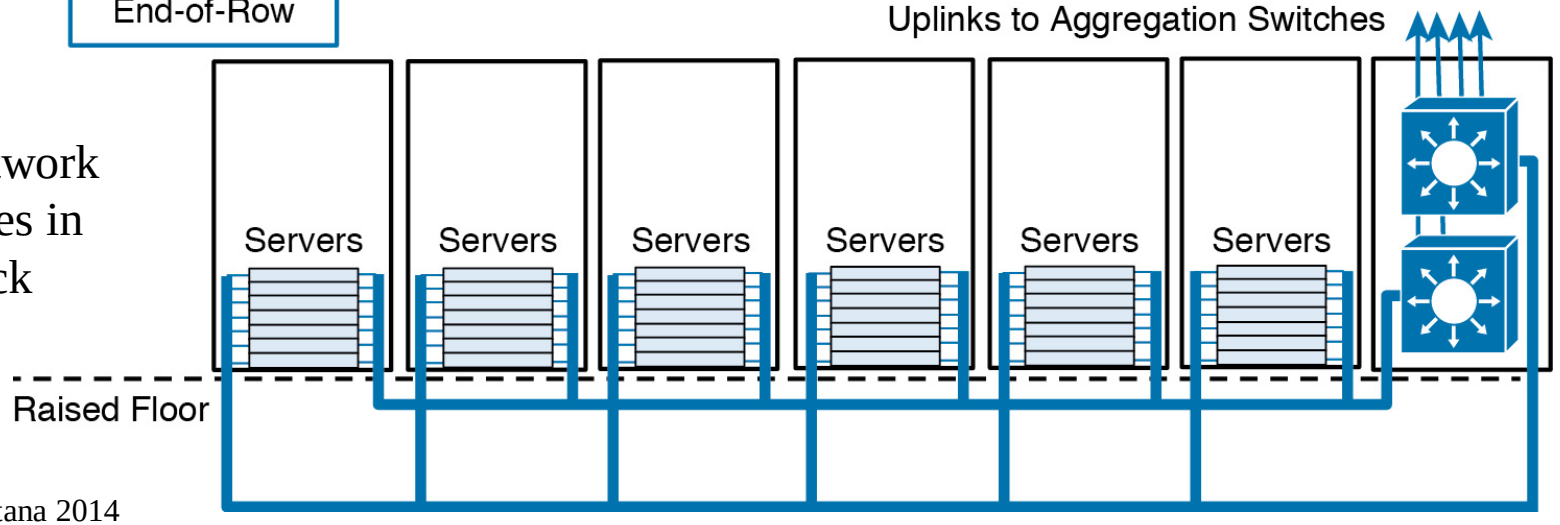
Smaller cable between servers and switches
Network team has to manage switches on all racks



End-of-Row

Uplinks to Aggregation Switches

All network switches in one rack



Source: Santana 2014

©2025 Raj Jain

Student Questions

ToR vs EoR

- ToR:
  - + Easier cabling
  - - If a rack is not fully populated ⇒ 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

- If a ToR switch has unused ports when a rack is partially filled, can't we eventually purchase switches based on need?

Each rack will need a switch.

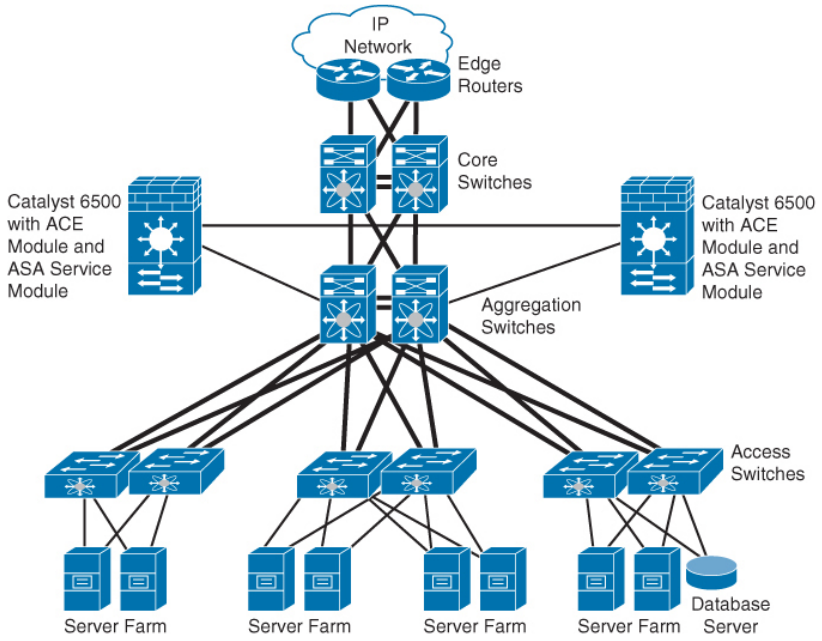
- In the case of EoR, isn't it possible that if we have, say, six racks with 42 servers each, but the large, shared switch has more ports than needed, then it too can have unused ports, right? So, why is this considered a specific disadvantage for ToR, but not for EoR?

ToR has unused ports on each rack that can not be used if needed by other racks.

©2025 Raj Jain

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



Source: Santana 2014

- Stateful devices (firewalls) on Aggregation layer
- Stateful = State of TCP connection
- Stateless, e.g., DNS

Student Questions

©2025 Raj Jain

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

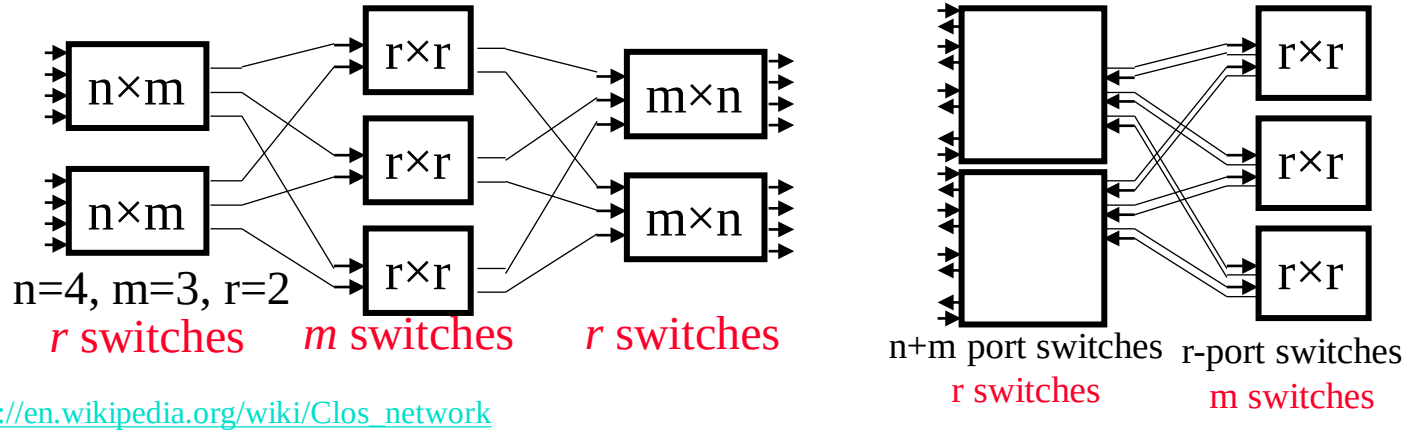
Student Questions

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

©2025 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 are unaffected.
- if m ≥ n
- Can have any odd number of stages, e.g., 5
- **Folded**: Merge input and output in to one switch



Student Questions

- How to calculate the number of bidirectional ports in a folded version of a topology?

Count. Or use the formula shown at the bottom of the diagram.

- In the folded Clos diagram, each link is shown as one port that sends and receives, unlike the separate input/output in the non-folded view. Is this a purely aesthetic choice, or does it alter the port model?

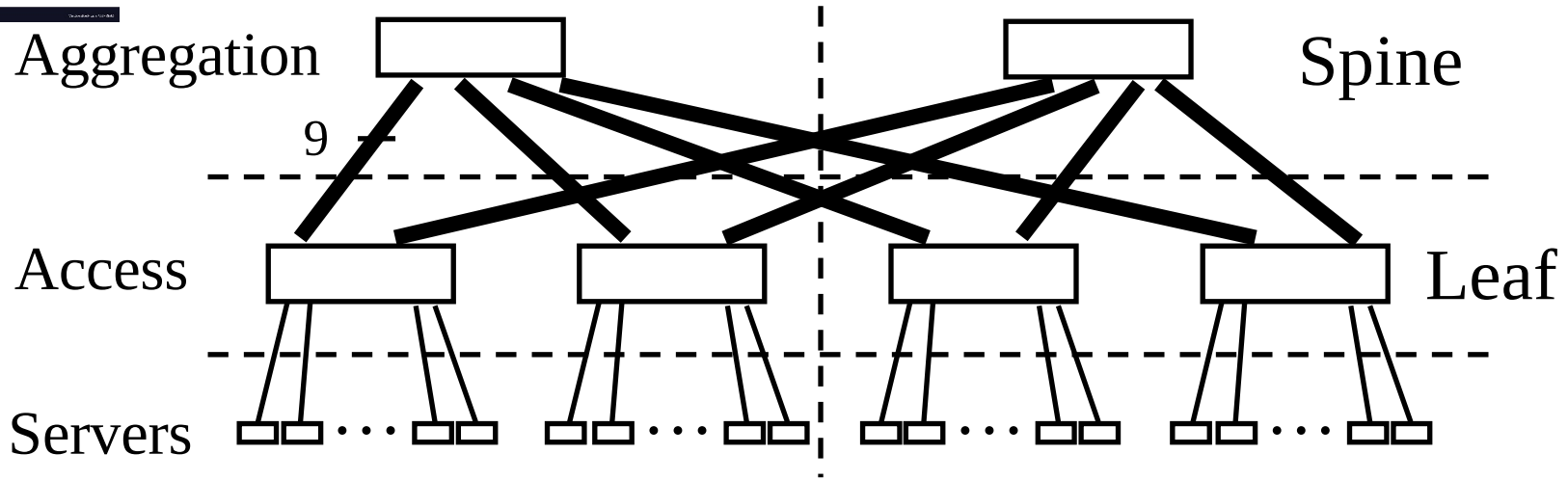
Bidirectional ports need more electronics than unidirectional. Actually, unidirectional ports are rarely, if ever, used. They are shown here for simplicity.

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

©2025 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

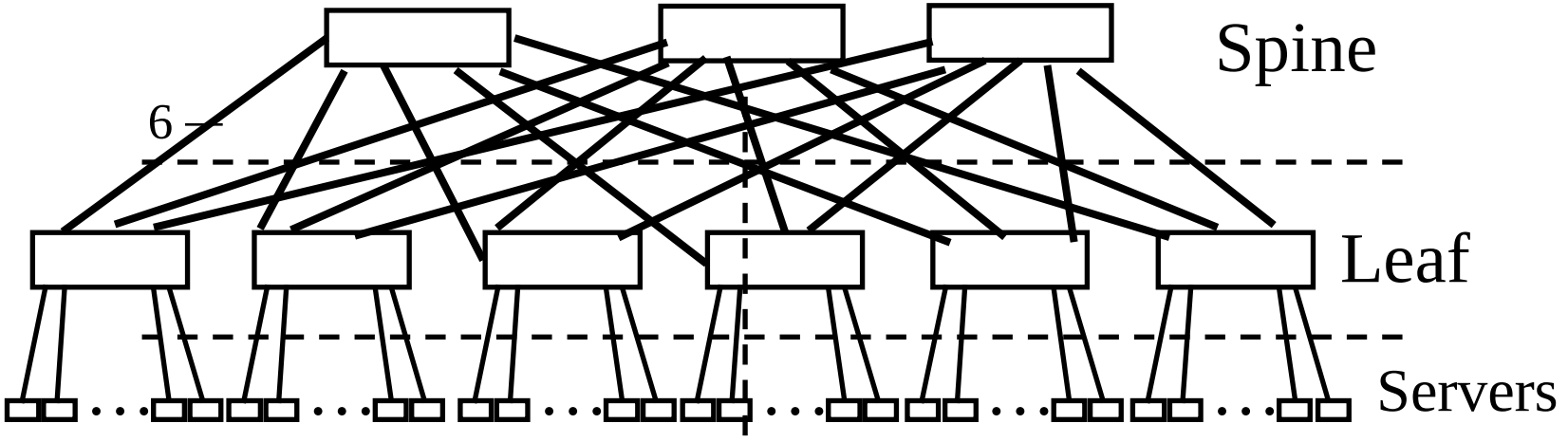
Ref: Teach yourself Fat-Tree Design in 60 minutes, http://clusterdesign.org/fat-trees/

©2025 Raj Jain

Student Questions

Fat-Tree Topology (Cont)

- Half of the leaf switch ports are toward servers, and the other half toward the spine
- With 36 port switches ⇒ 18 ports to the spine
  ⇒ 2, 3, 6, 9, 18 spine switches
- Maximum # of spine switches = ½ # of ports on leaf switches



- Largest configuration with n-port switches: $n^2/2$ servers can be connected using $n+n/2$ switches=(n leaf's + $n/2$ spine) switches).

Student Questions

- What are the disadvantages of maximizing services in the Fat-Tree topology?

Servers are not dual ported.

- If each leaf switch has n ports, how do I estimate the maximum number of servers and required spine count?

Half ports up, half down.

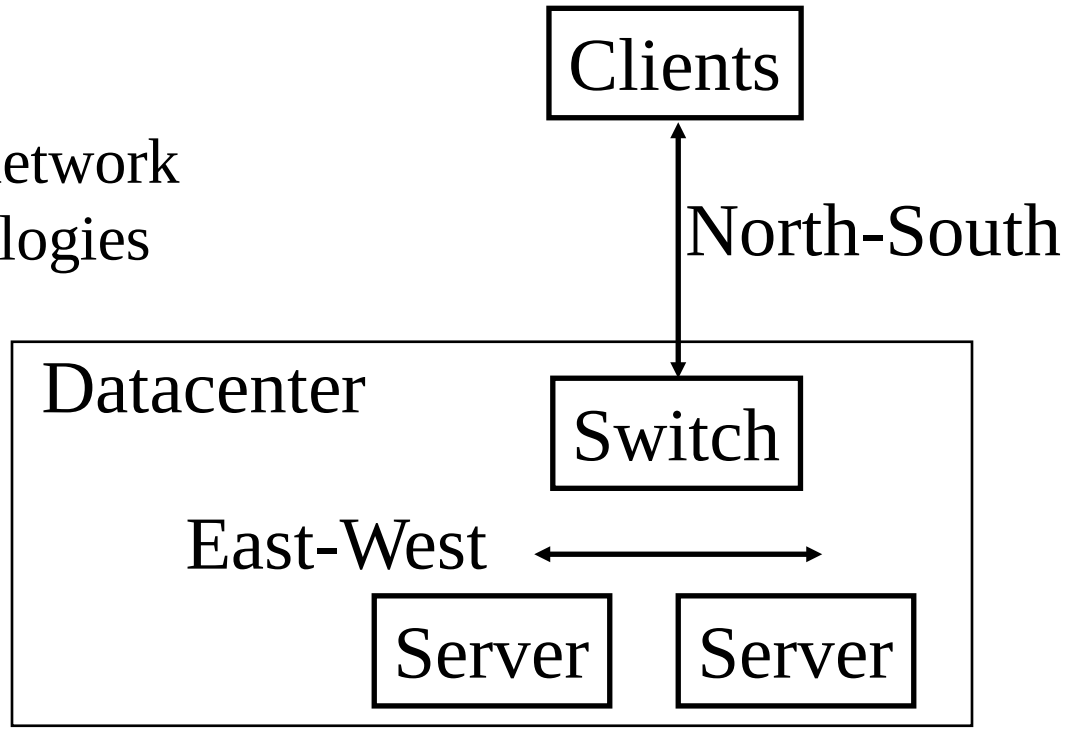
- The slide shows 36-port switches connecting to 18 servers and 18 uplinks. Will this always be a 50-50 split?

Yes.

©2025 Raj Jain

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

©2025 Raj Jain

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

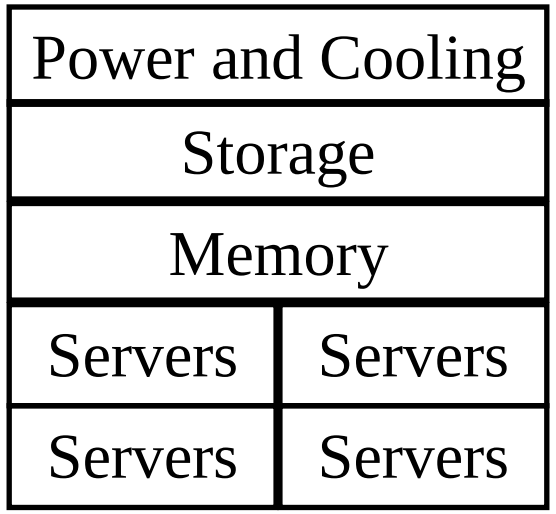
Student Questions

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

©2025 Raj Jain

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)

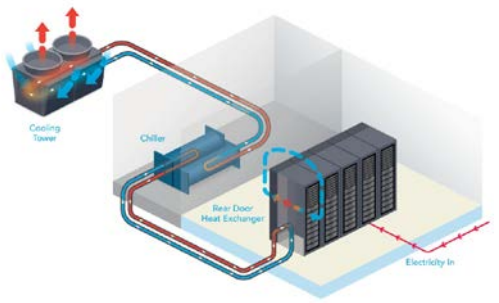


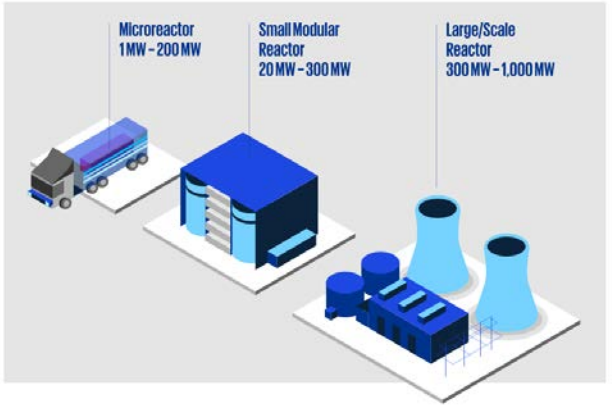
Student Questions

©2025 Raj Jain

Recent Developments

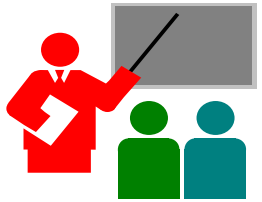
- AI Centric Topologies: Air-cooled designs are being replaced by liquid cooling to support AI.
- Edge Expansion: Data centers in secondary hubs. Large centralized hubs are difficult to power.
- Power-Aware Designs: Near renewable energy parks
- Small Modular Reactors (SMRs) Powered: pre-fabricated and assembled on location. ≈300 MW.
- Google's Tensor Processing Unit (TPU) datacenter for AI workloads







Student Questions



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
4. ANSI TIA-942 Standard for data center layout
5. Open Rack Scale Architecture to improve energy efficiency

Student Questions

©2025 Raj Jain

Homework 3A

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

Student Questions

- When you ask what the number of ports used in a CLOS topology is, are you asking what port switches we are using, or does it indicate how many of these port switches we used?

The first is the number of ports, and the second is the number of switches.

- Can you explain why bidirectional ports is 9?

See the slide 3.21: m + n port switches.

©2025 Raj Jain

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

©2025 Raj Jain

Acronyms

ADC	Application Delivery Controller
AI	Artificial Intelligence
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

©2025 Raj Jain

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

©2025 Raj Jain

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., 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

©2025 Raj Jain

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

©2025 Raj Jain

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

©2025 Raj Jain

Scan This to Download These Slides



Raj Jain
http://rajjain.com

http://www.cse.wustl.edu/~jain/cse5700-25/m_03dct.htm

Student Questions

©2025 Raj Jain

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

©2025 Raj Jain