

Data Center Network Topologies



Raj Jain

Washington University in Saint Louis

Saint Louis, MO 63130

Jain@cse.wustl.edu

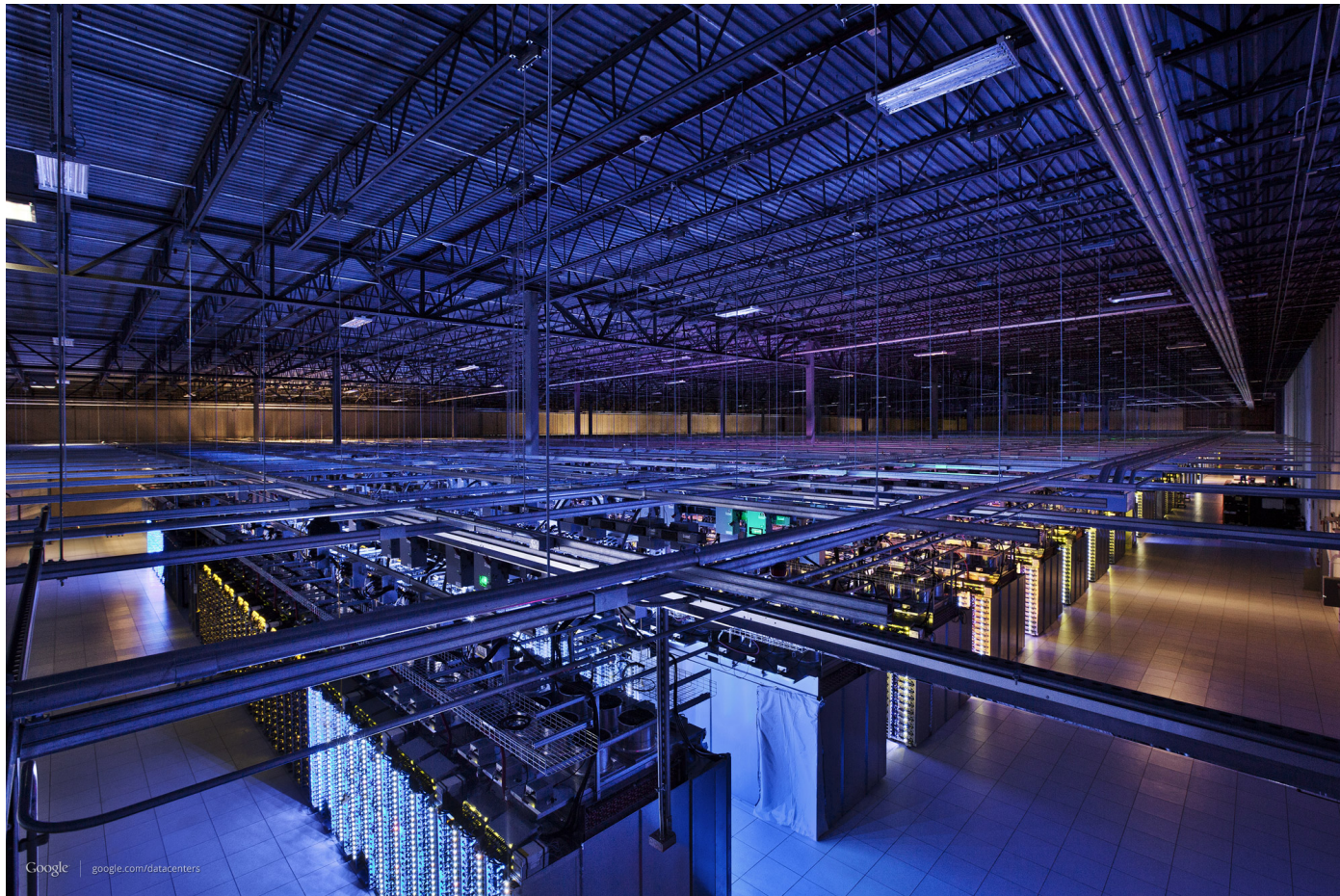
These slides and audio/video recordings of this class lecture are at:

<http://www.cse.wustl.edu/~jain/cse570-18/>



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

Google's Data Center



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

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 Raj Jain

Cooling Plant



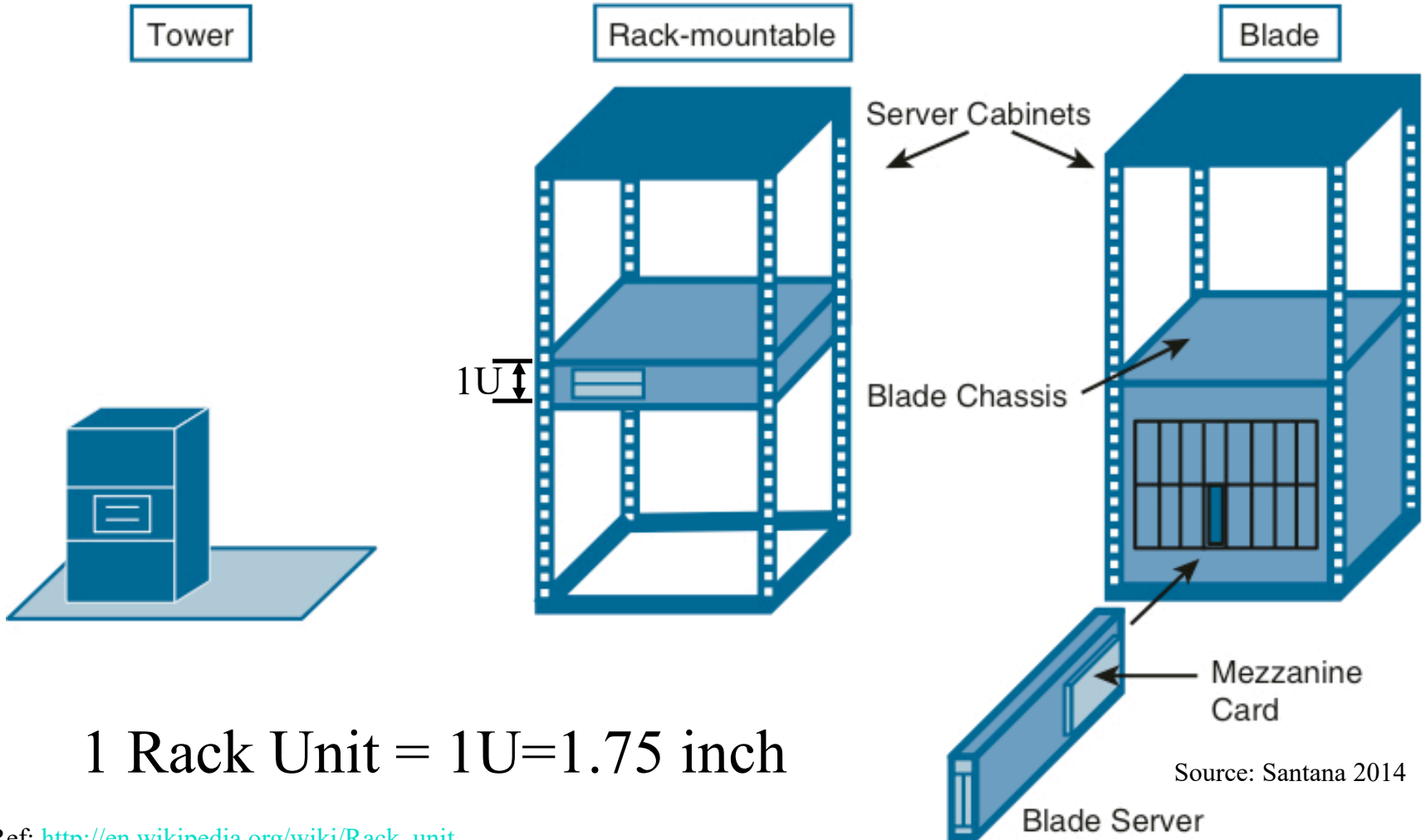
Source: <http://webodyssey.com/technologyscience/visit-the-googles-data-centers/>

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 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/cse570-18/>

Source: Santana 2014

©2018 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

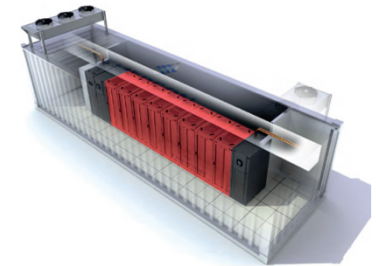
Ref: http://www.sgi.com/products/data_center/ice_cube_air/

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 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: Datacenter Infrastructure – mobile Data Center from Emerson Network Power

, <http://en.m-info.ua/180-container-data-center/755-datacenter-infrastructure-mobile-data-center-from-emerson-network-power>

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

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 Raj Jain

Unstructured Cabling



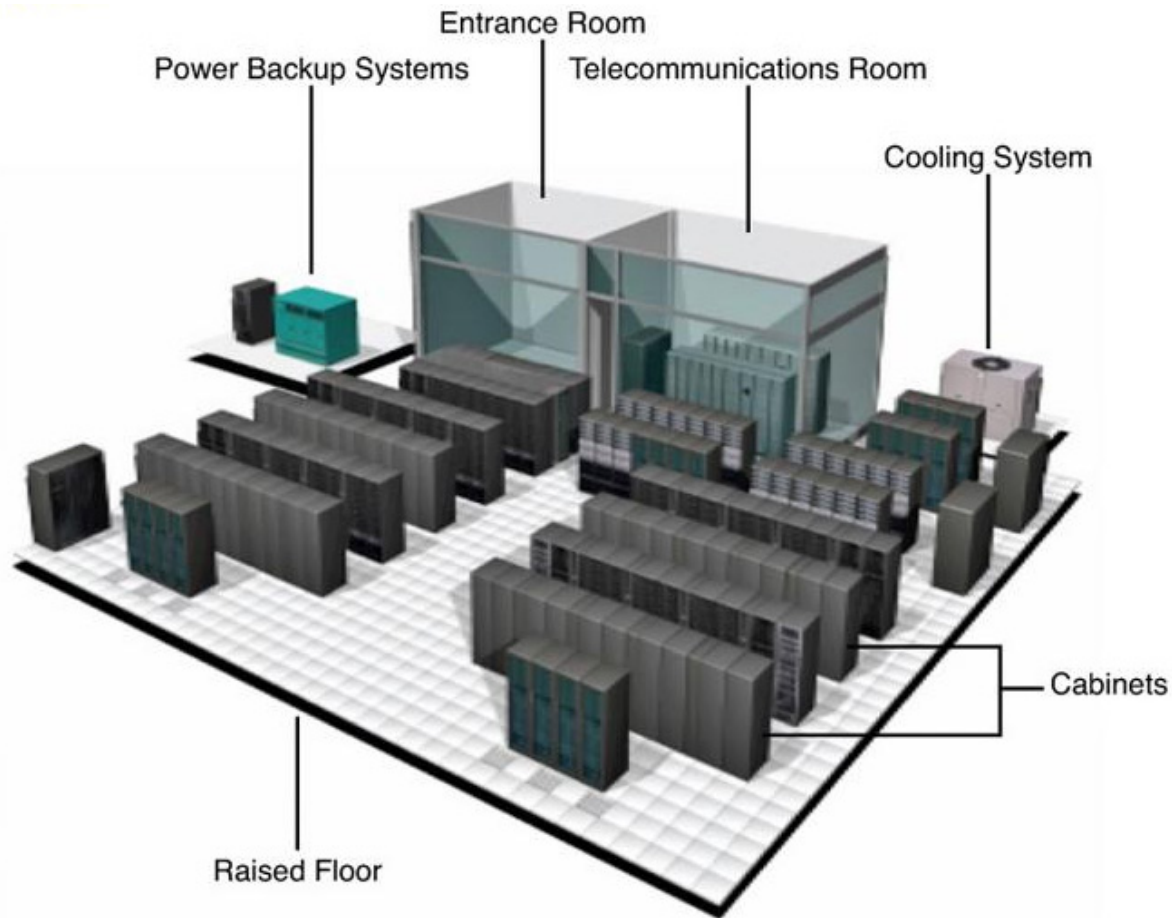
Source: <http://webodyssey.com/technologyscience/visit-the-googles-data-centers/>

Structured Cabling



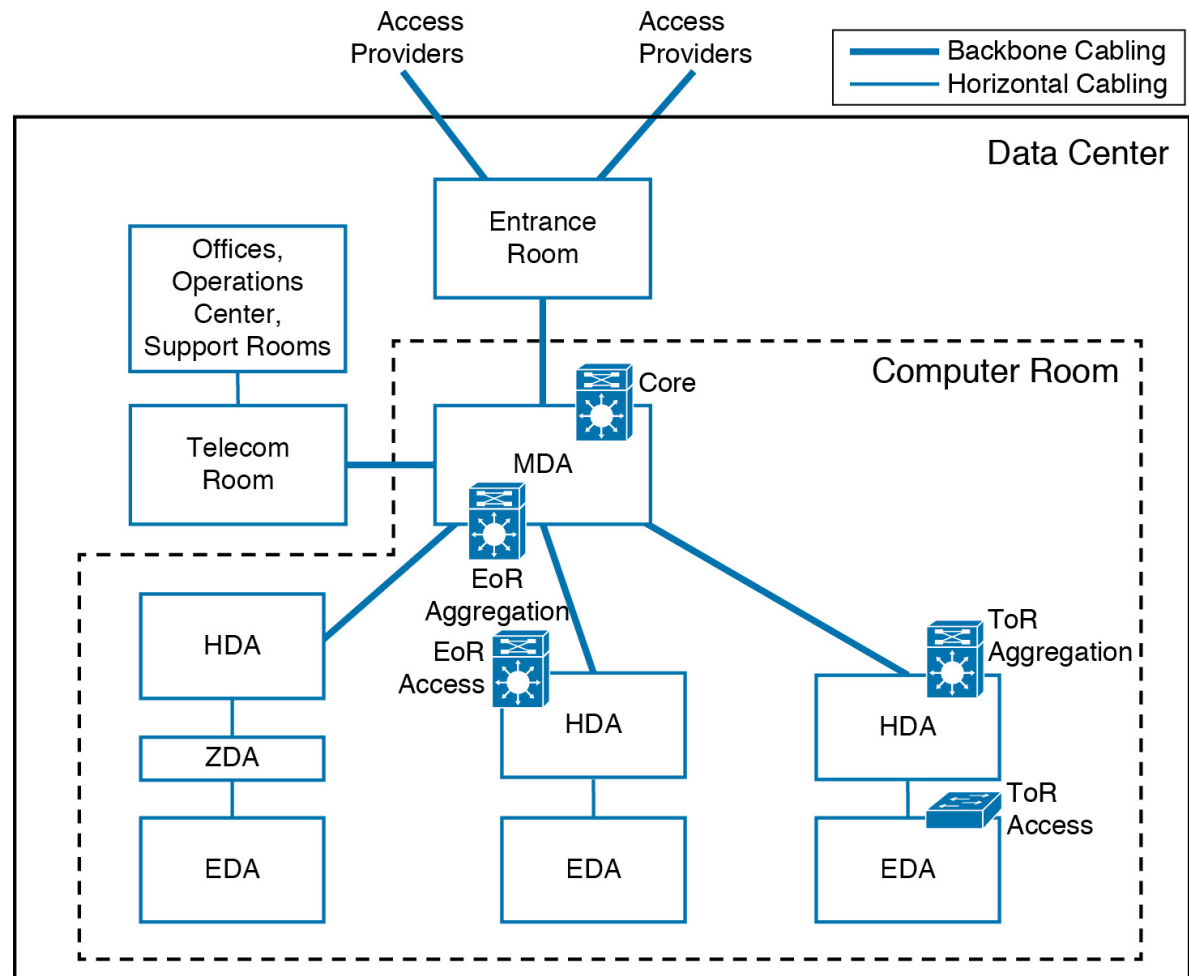
Source: <http://webodyssey.com/technologyscience/visit-the-googles-data-centers/>

Data Center Physical Layout



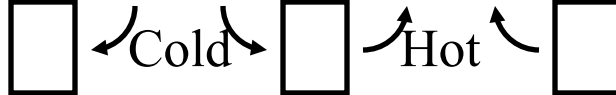
ANSI/TIA-942-2005 Standard

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

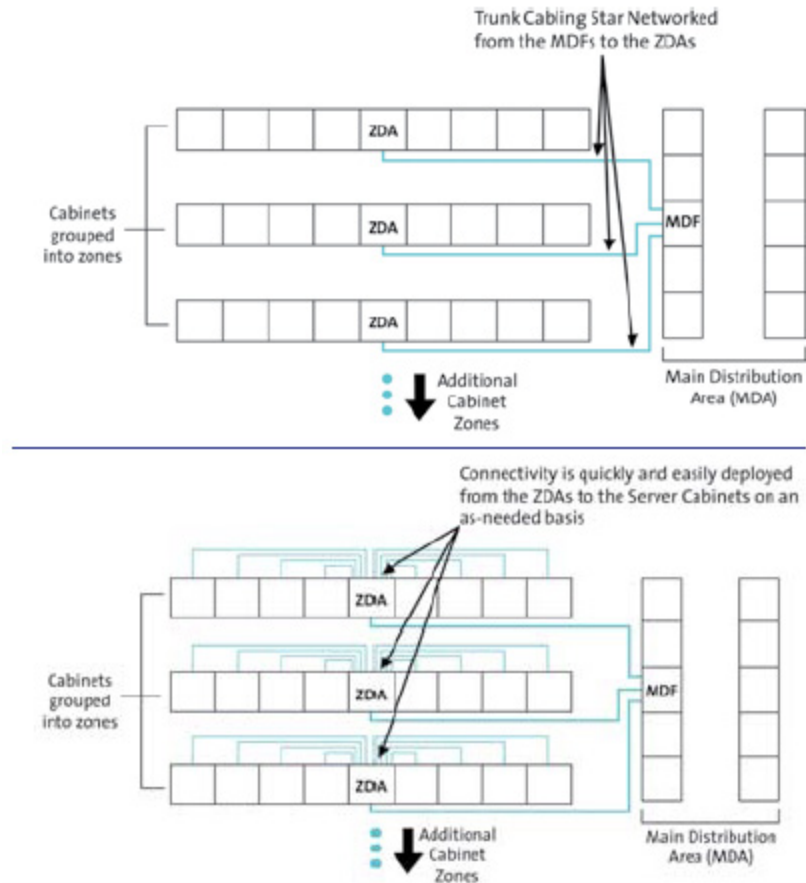


Source: Santana 2014

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

Zone Distribution Area



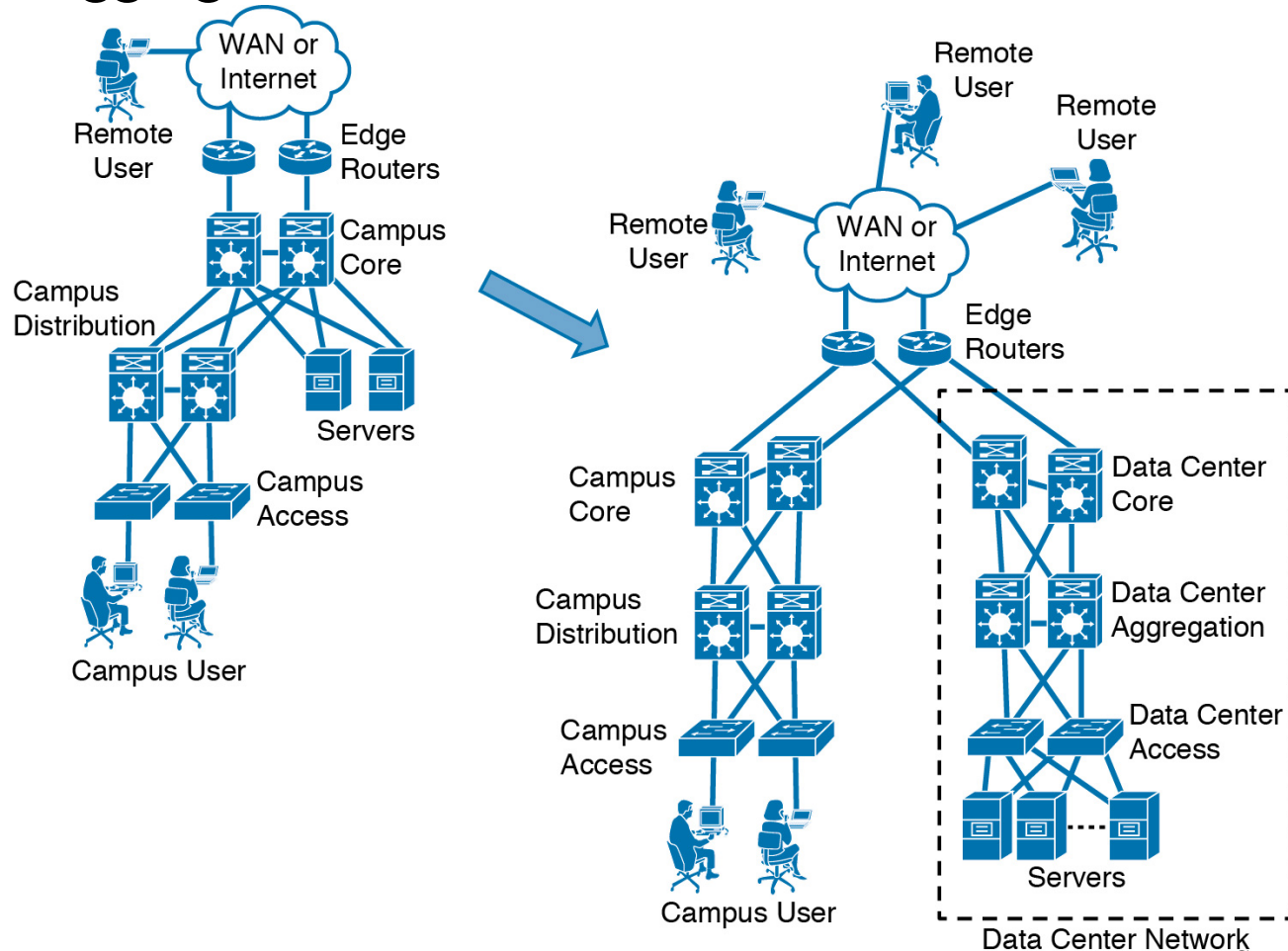
- ❑ High-fiber count cables connect ZDA to MDA or HDA.
Low-fiber count cables connect ZDA to EDA as needed.

Ref: Jennifer Cline, "Zone Distribution in the data center,"

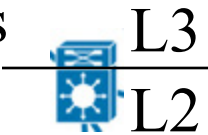
<http://www.graybar.com/documents/zone-distribution-in-the-data-center.pdf>

Data Center Network Topologies

□ Core, Aggregation, Access



Data Center Networks

- ❑ 20-40 servers per rack
- ❑ Each server connected to 2 access switches with 1 Gbps (10 Gbps becoming common)
- ❑ Access switches connect to 2 aggregation L3 switches (Switches that implement routing functions) 
- ❑ Aggregation switches connect to 2 core L3 switches
- ❑ Core L3 switches connect to edge routers
- ❑ Core layer forwards data center ingress and egress traffic
- ❑ Aggregation layer forwards server-to-server traffic in the data center
- ❑ Aggregation layer is also a place to put middleboxes, such as, firewalls, load balancers

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>.

Data Center Networks (Cont)

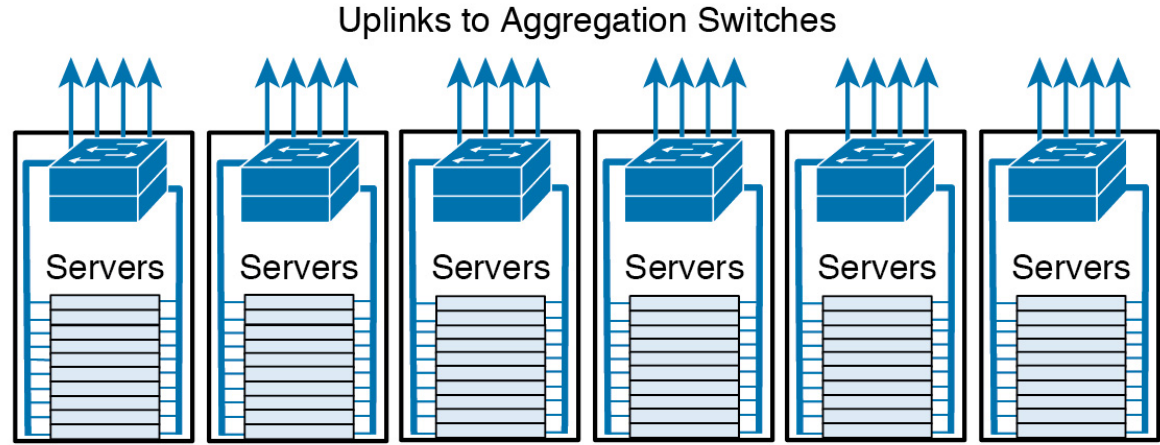
- ❑ Access Layer provide high number of ports for connectivity.
- ❑ Low Latency: In high-frequency trading market, a few microseconds make a big difference.
 - ⇒ Cut-through switching and low-latency specifications.
- ❑ All switches below each pair of aggregation switches form a single layer-2 domain
- ❑ Each Layer 2 domain typically limited to a few hundred servers to limit broadcast
- ❑ Most traffic is internal to the data center.
- ❑ Network is the bottleneck.
 - Uplinks utilization of 80% is common.
- ❑ Most of the flows are small.
 - Mode = 100 MB. DFS uses 100 MB chunks.

Switch Locations

Top-of-Rack

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

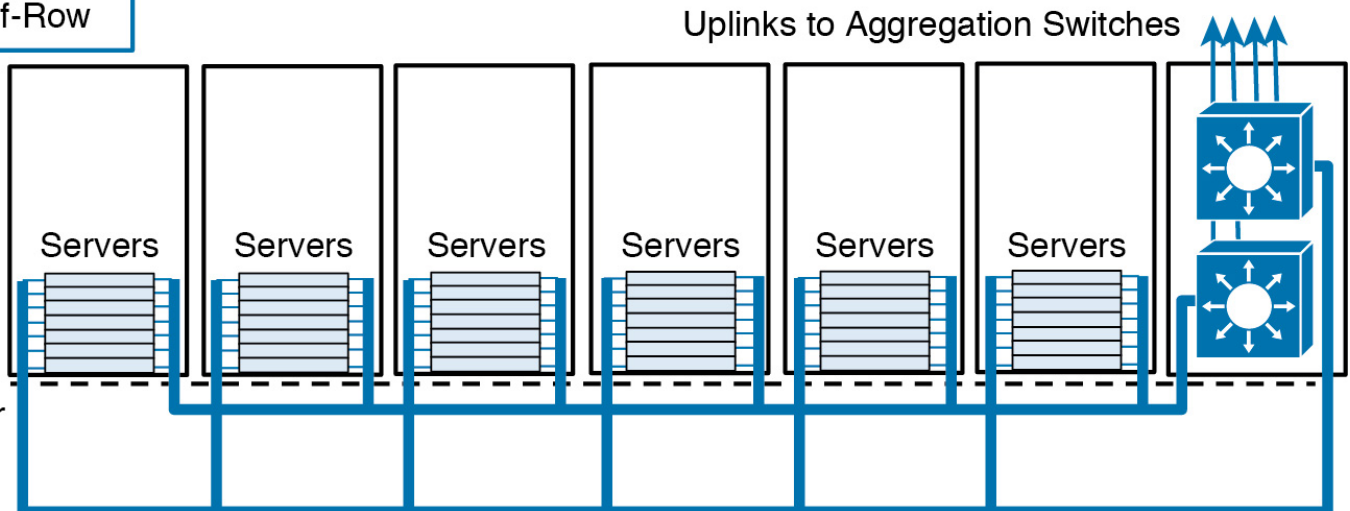
Raised Floor



End-of-Row

All network
switches in
one rack

Raised Floor



Source: Santana 2014

ToR vs EoR

□ ToR:

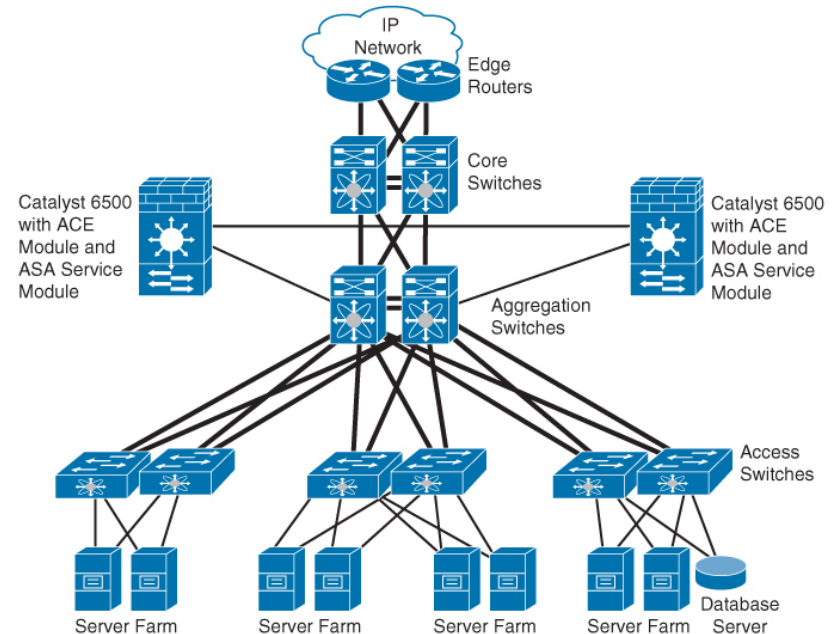
- + Easier cabling
- - If rack is not fully populated \Rightarrow unused ToR ports
- - If rack traffic demand is high, difficult to add more ports
- - Upgrading (1G to 10G) requires complete Rack upgrade

□ EoR:

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

Hierarchical Network Design

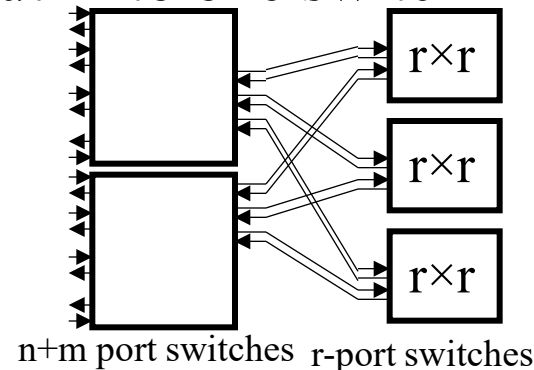
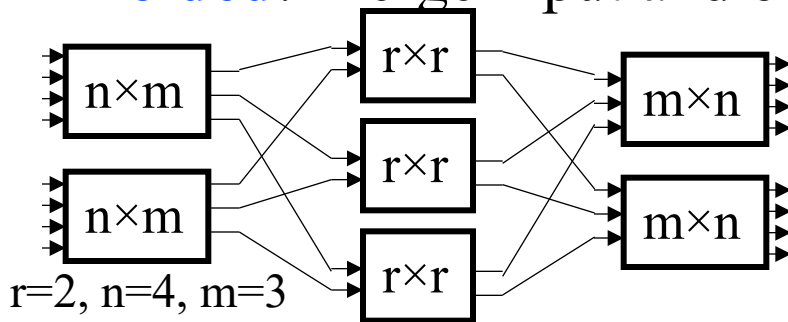
- ❑ All servers require application delivery services for security (VPN, Intrusion detection, firewall), performance (load balancer), networking (DNS, DHCP, NTP, FTP, RADIUS), Database services (SQL)
- ❑ ADCs are located between the aggregation and core routers and are shared by all servers
- ❑ Stateful devices (firewalls) on Aggregation layer
- ❑ Stateful= State of TCP connection



Source: Santana 2014

Clos Networks

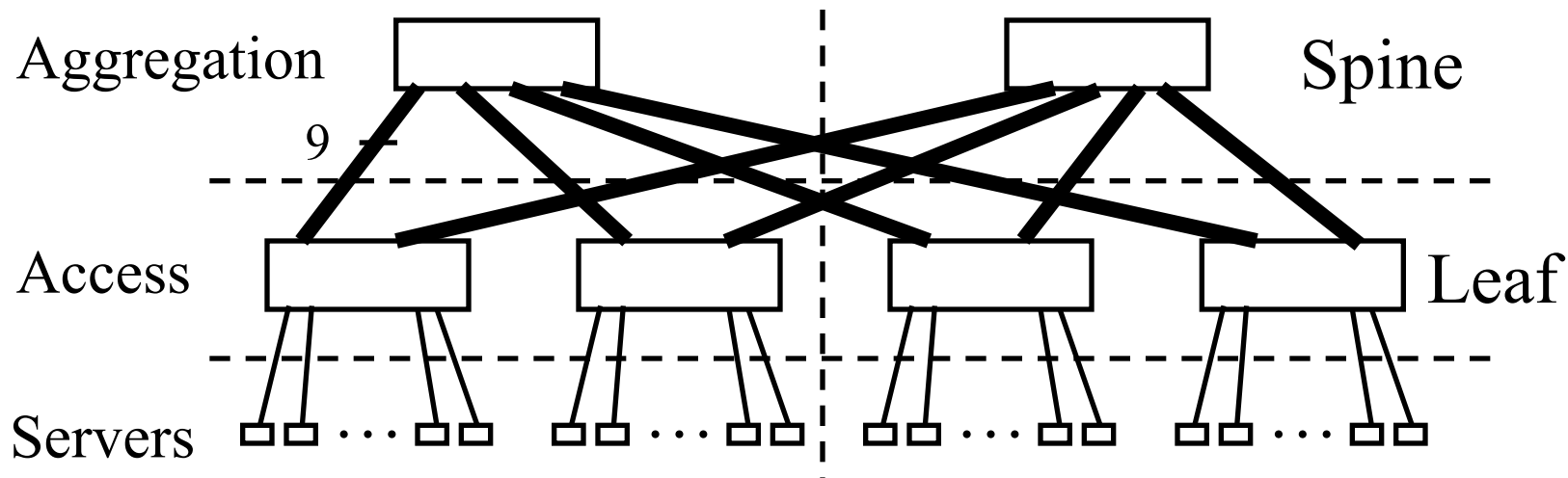
- ❑ Multi-stage circuit switching network proposed by Charles Clos in 1953 for telephone switching systems
- ❑ Allows forming a large switch from smaller switches
The number of cross-points is reduced \Rightarrow Lower cost (then)
- ❑ 3-Stage Clos(n, m, r): ingress ($r \times n \times m$), middle ($m \times r \times r$), egress ($r \times m \times n$)
- ❑ Strict-sense non-blocking if $m \geq 2n-1$. Existing calls 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 = Fat-tree



Homework 3A

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

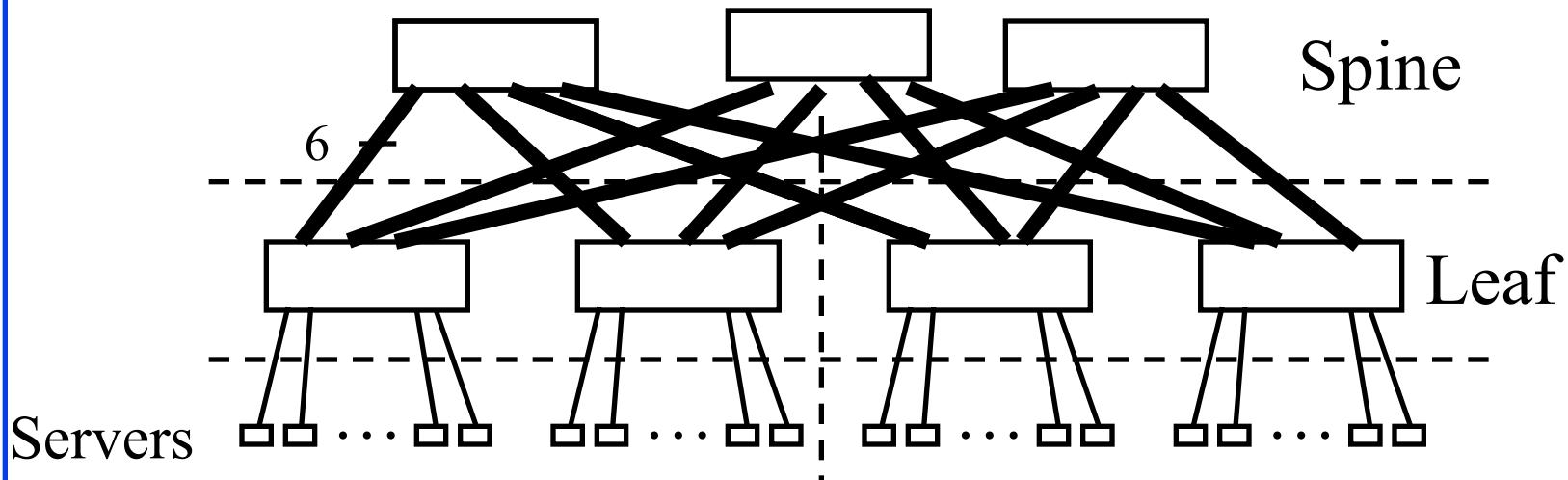
Fat-Tree DCN Example



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

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 = $\frac{1}{2}$ # of ports on leaf switches

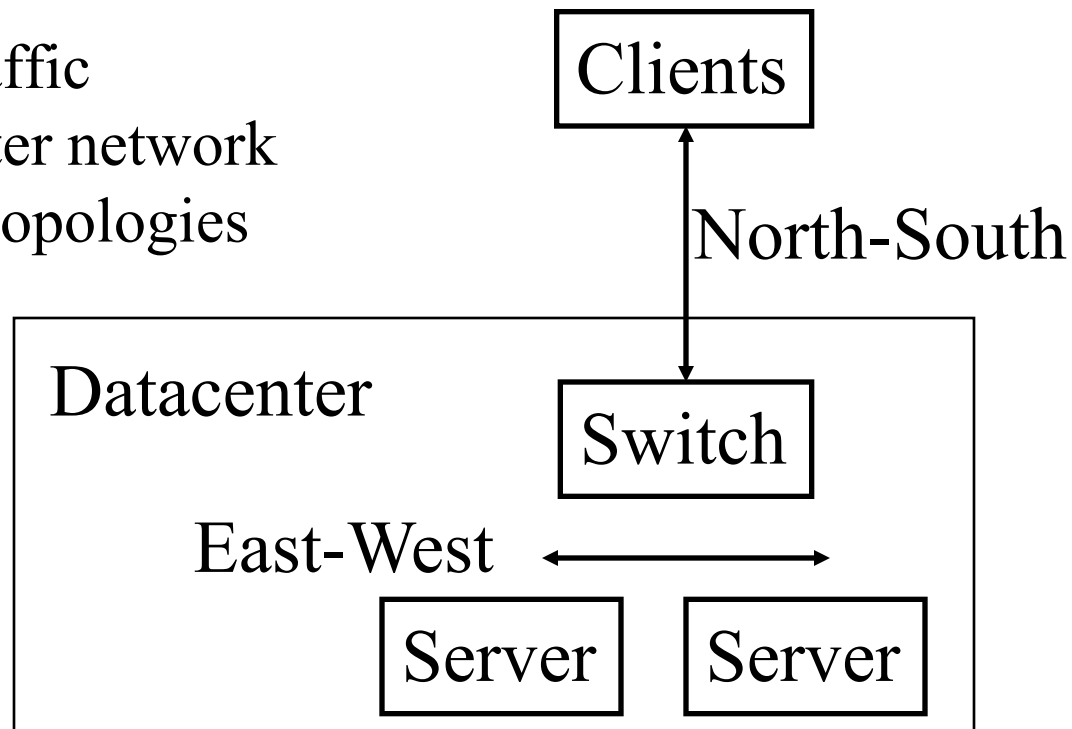


Homework 3B

Draw a Fat-tree topology using four 4-port switches. Assume each server is connected to a single access switch while the access switches are dual homed to aggregation switches. There is no core tier.

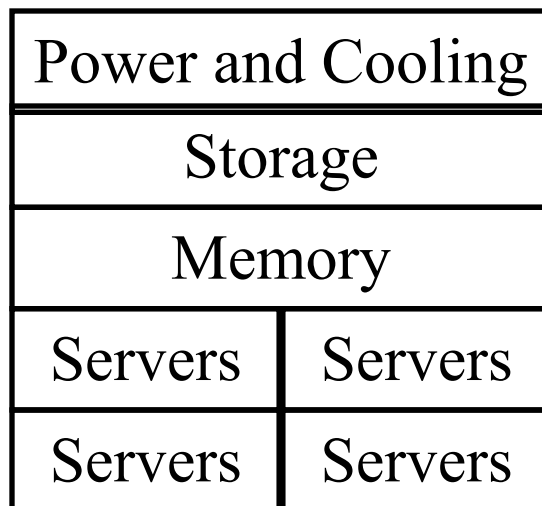
North-South vs. East-West Traffic

- ❑ 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



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)



Micro-Servers

- ❑ Microserver = a small system on a chip (SOC) containing CPU, memory and multiple NICs
- ❑ Many microservers on a board (look like memory DIMMs)
- ❑ Microserver sleds can replace server sleds in rack scale architecture

Summary



1. Modular data centers can be used for easy assembly and scaling
2. Three tiers: Access, Aggregation, Core
3. Application delivery controllers between Aggregation and core
4. Need large L2 domains
5. Fat-tree topology is sometimes 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 |

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 |

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 |

Acronyms (Cont)

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

Reading List

- ❑ <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/ijj-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>
- ❑ G. Santana, “Data Center Virtualization Fundamentals,” Cisco Press, 2014, ISBN:1587143240 (Safari book)
- ❑ 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/>

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
- ❑ http://en.wikipedia.org/wiki/Fat_tree
- ❑ http://en.wikipedia.org/wiki/Hierarchical_internetworking_model
- ❑ http://en.wikipedia.org/wiki/Clos_network

Scan This to Download These Slides



Raj Jain

<http://rajain.com>

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>