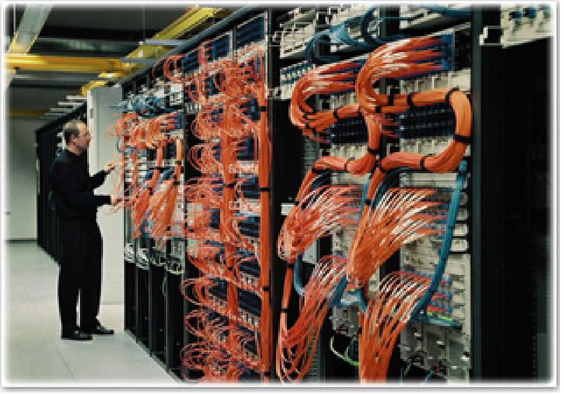
Data Center Network Topologies



Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

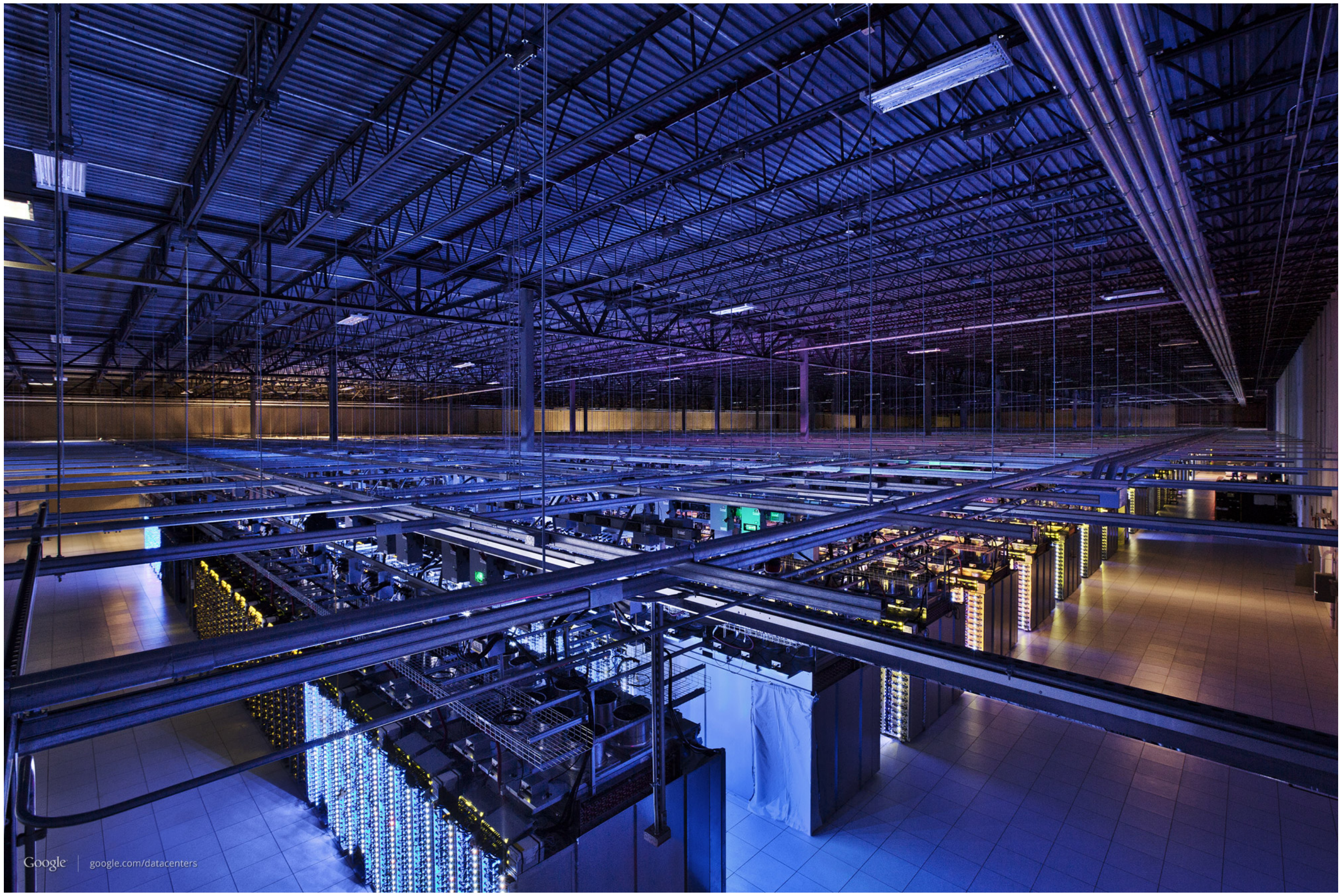
These slides and audio/video recordings of this class lecture are at:

http://www.cse.wustl.edu/~jain/cse570-15/

Overview

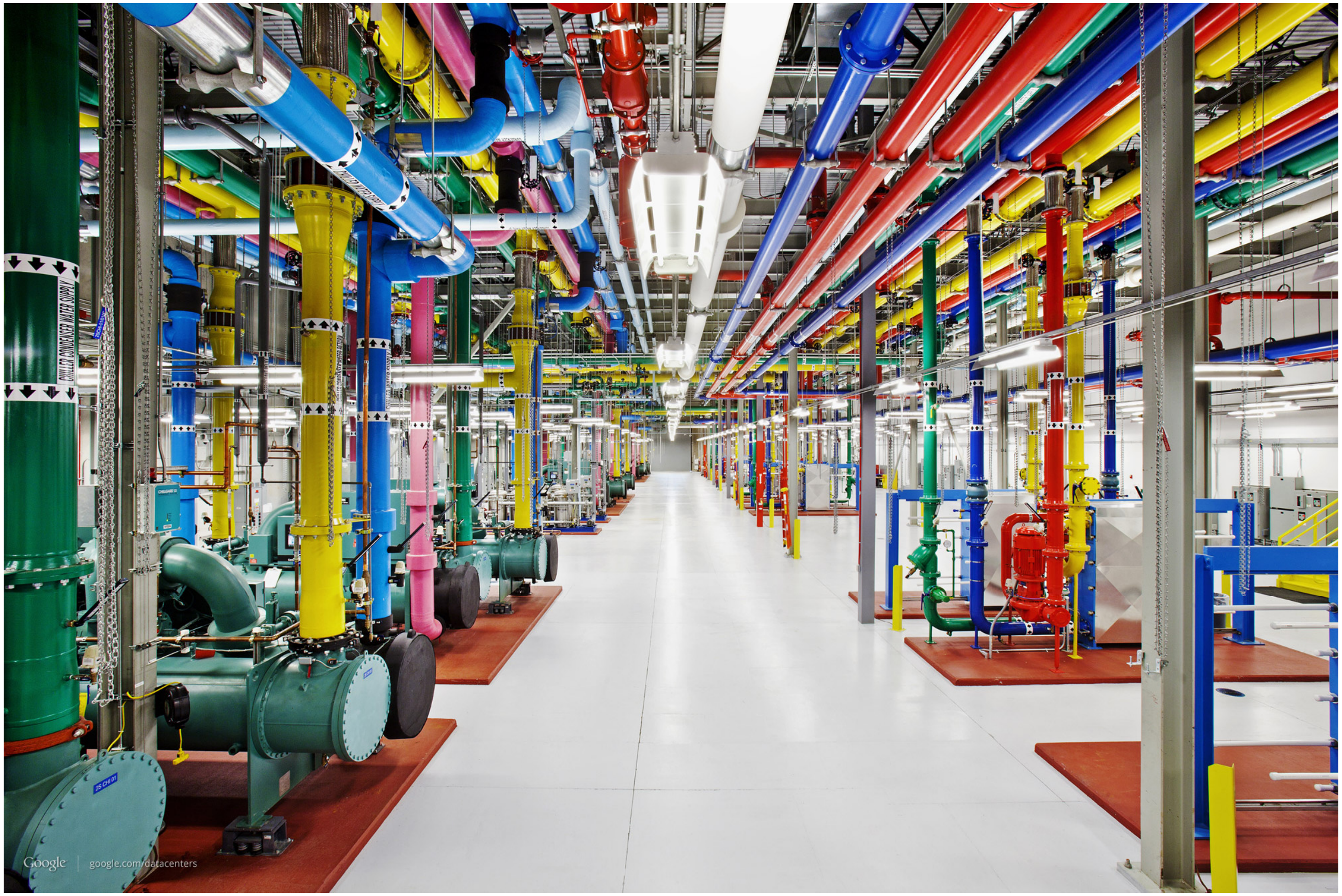
1. Data Center Physical Layout
2. Data Center Network Cabling
3. ToR vs. EoR
4. Clos and Fat-Tree topologies

Google's Data Center



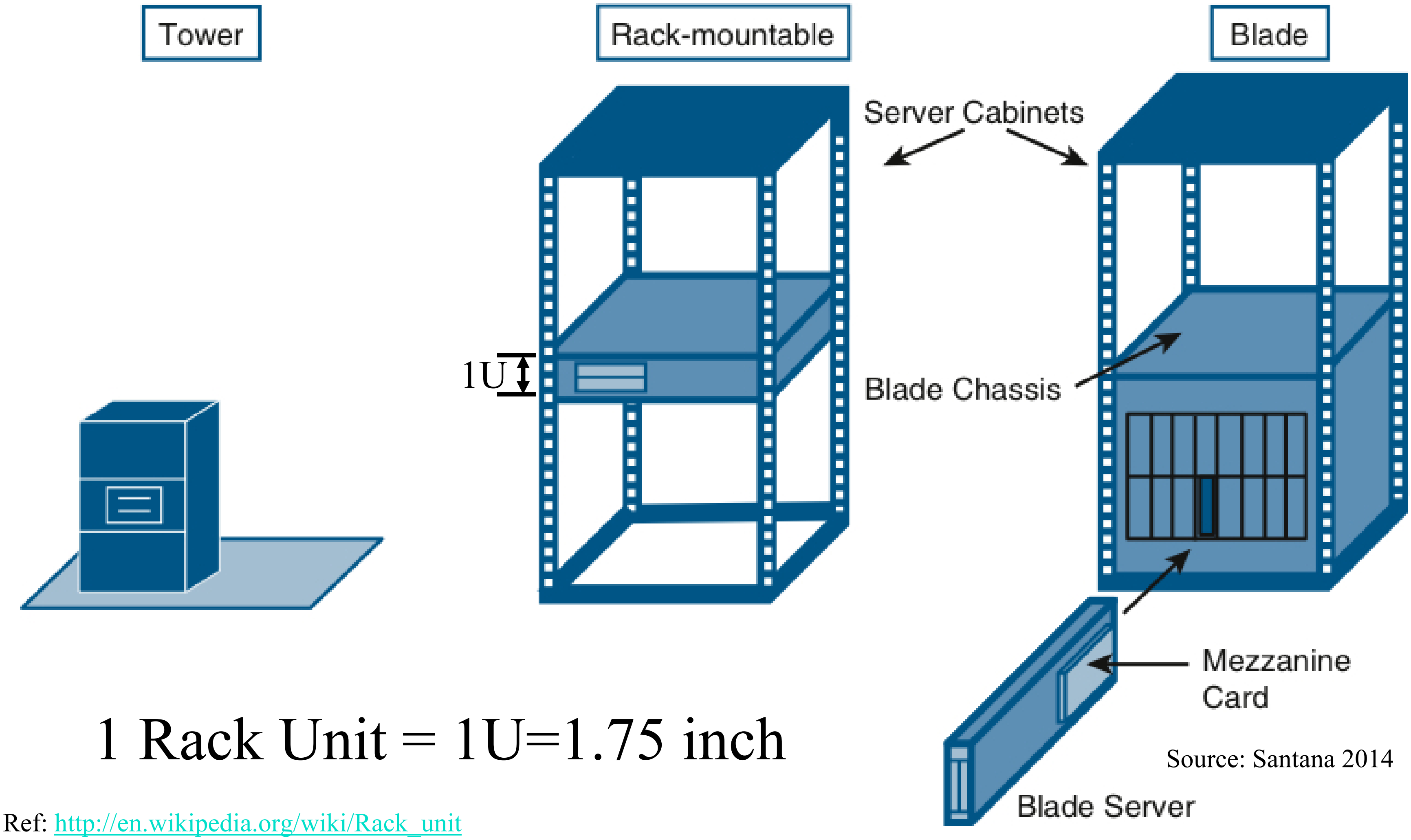
Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/

Cooling Plant



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

Servers



1 Rack Unit = 1U=1.75 inch

Source: Santana 2014

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

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

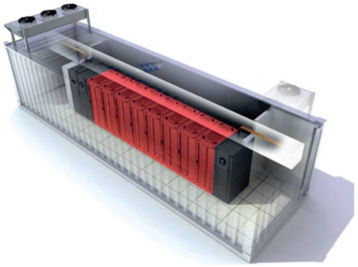
Modular Data Centers

- 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/

Containerized Data Center

EMERSON
Datacenter Infrastructure

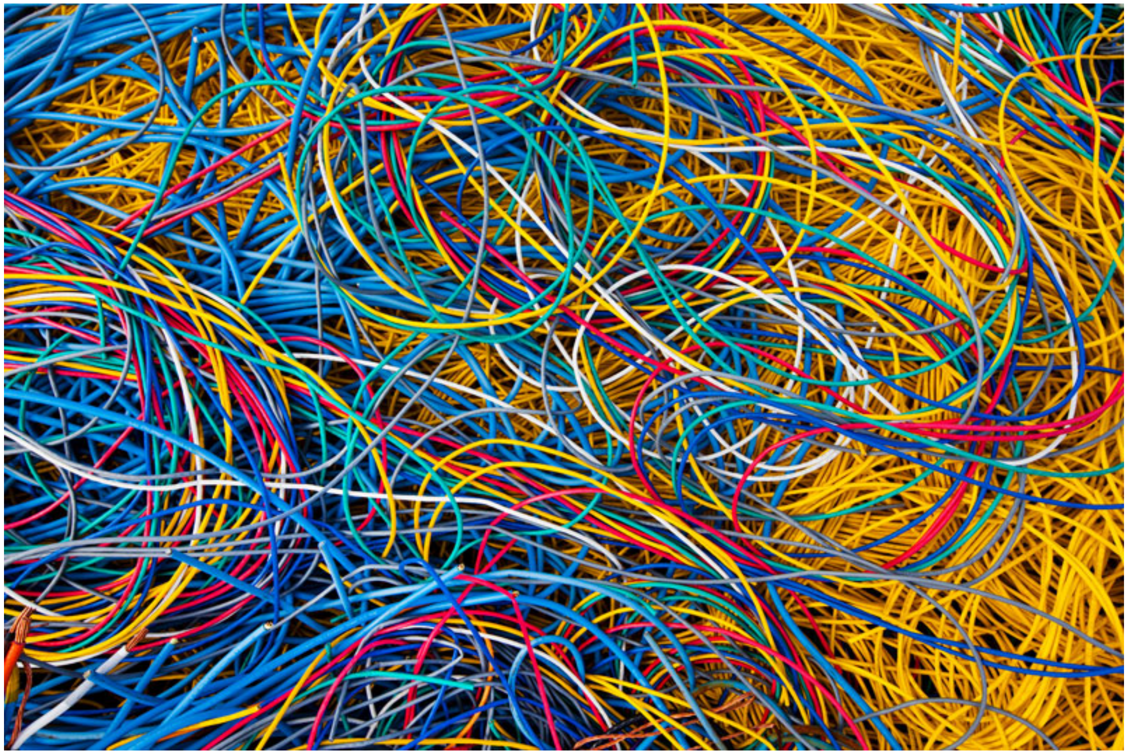


- ❑ Ready to Use. Connect to water and power supply and go.
- ❑ Built in cooling. Easy to scale.
  ⇒ Data Center trailer parks.
- ❑ Suitable for disaster recovery, e.g., flood, earthquake
- ❑ Offered by Cisco, IBM, SGI, Sun/ORACLE,…

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/iij-will-offer-commercial-container-facility/

Unstructured Cabling

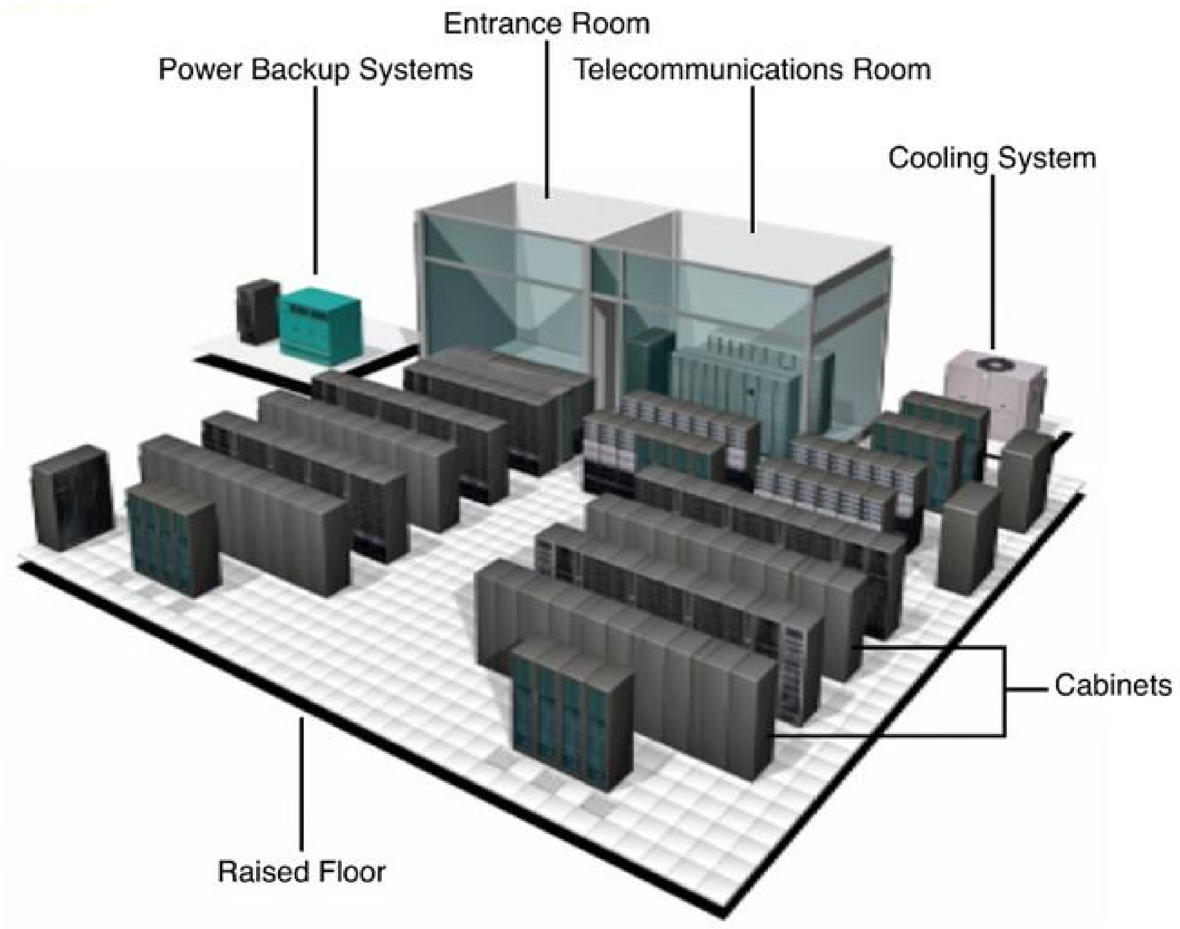


Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/

Structured Cabling

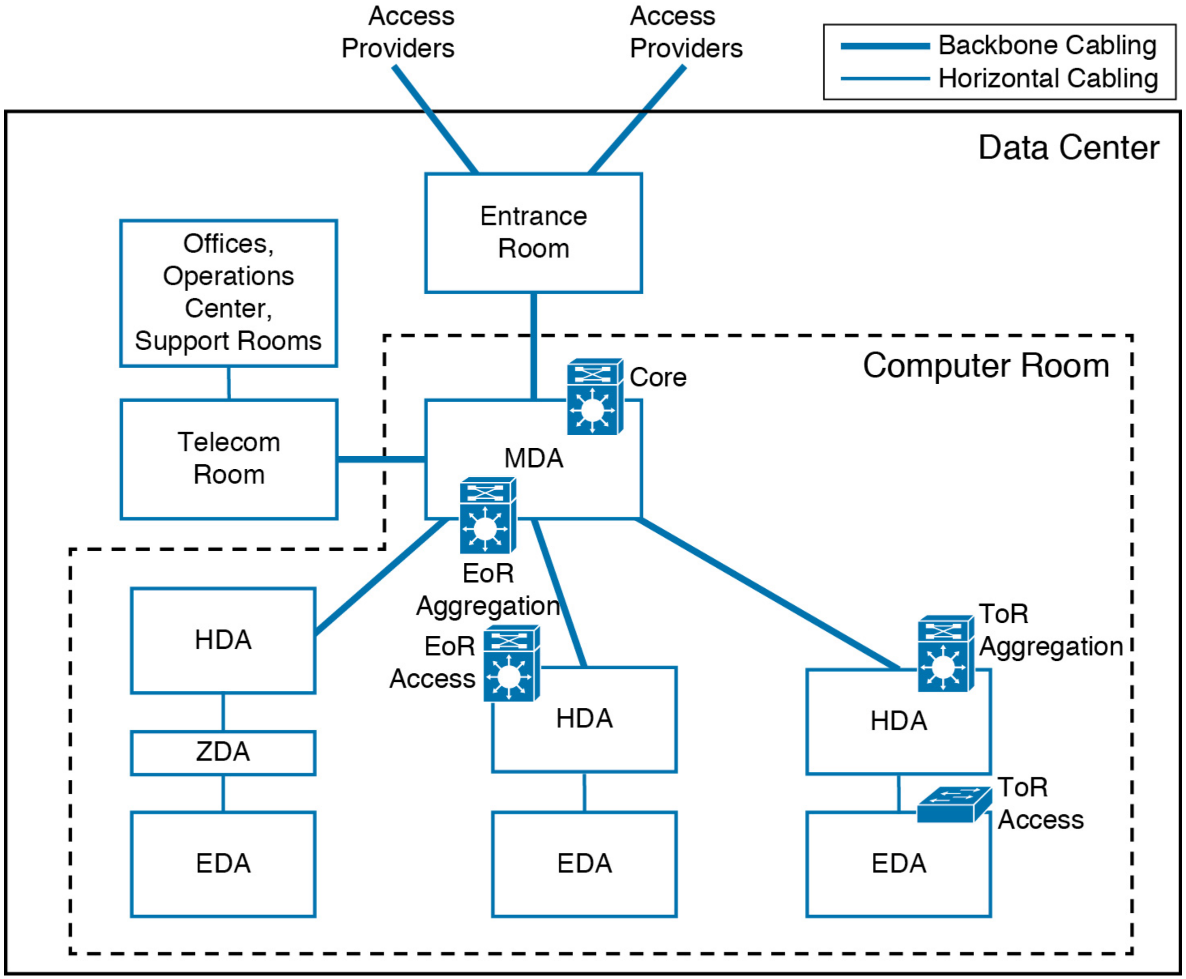
Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/

Data Center Physical Layout



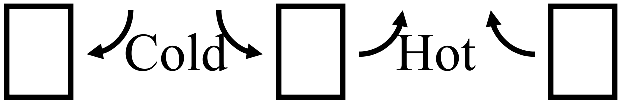
ANSI/TIA-942-2005 Standard

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



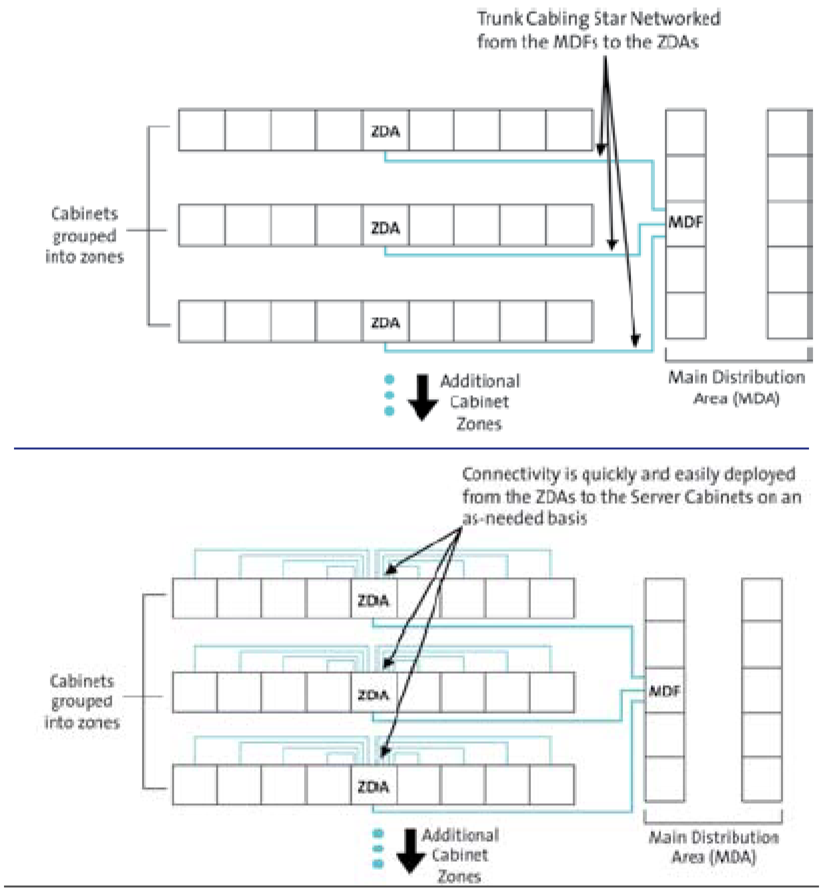
Source: Santana 2014

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

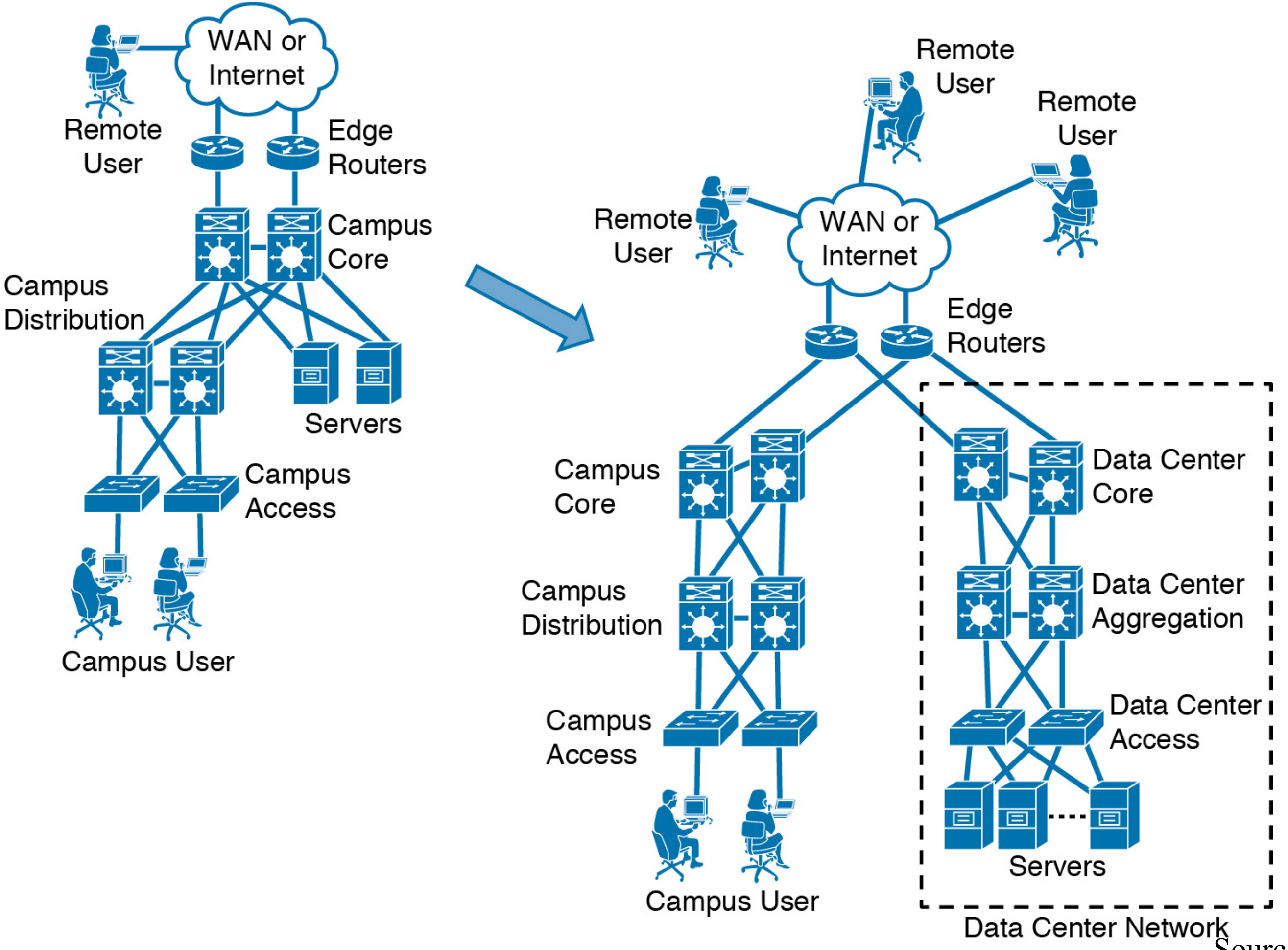

Trunk Cabling Star Networked from the MDFs to the ZDAs
Cabinets grouped into zones
ZDA
ZDA
ZDA
MDF
Additional Cabinet Zones
Main Distribution Area (MDA)
Connectivity is quickly and easily deployed from the ZDAs to the Server Cabinets on an as-needed basis
Cabinets grouped into zones
ZDA
ZDA
ZDA
MDF
Additional Cabinet Zones
Main Distribution Area (MDA)

- ❑ 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



Source: Santana 2014

Data Center Networks

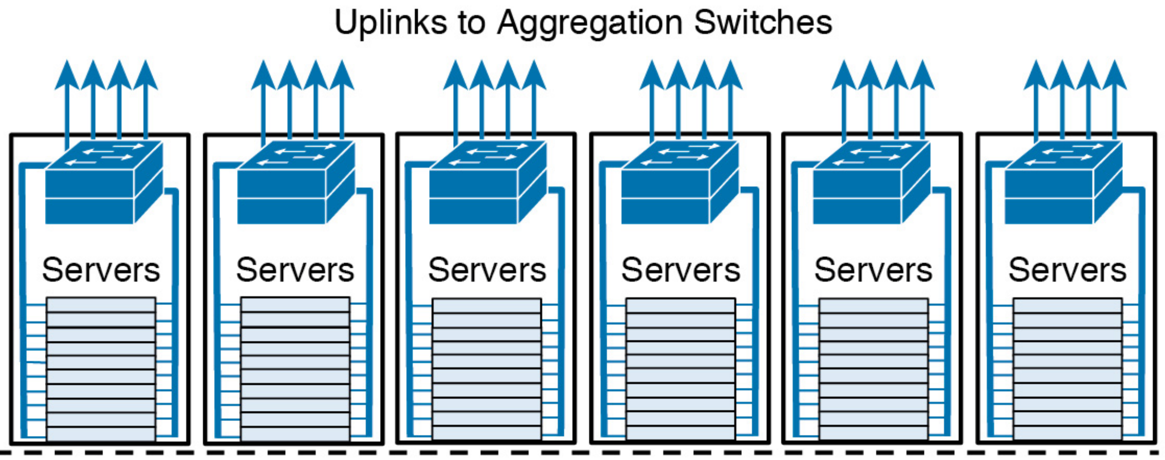
- 20-40 servers per rack
- Each server connected to 2 access switches with 1 Gbps (10 Gbps becoming common)
- Access switches connect to 2 aggregation switches
- Aggregation switches connect to 2 core routers
- Core routers connect to edge routers
- Aggregation layer is the transition point between L2-switched access layer and l3-routed core layer
- Low Latency: In high-frequency trading market, a few microseconds make a big difference.
  ⇒ Cut-through switching and low-latency specifications.

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)

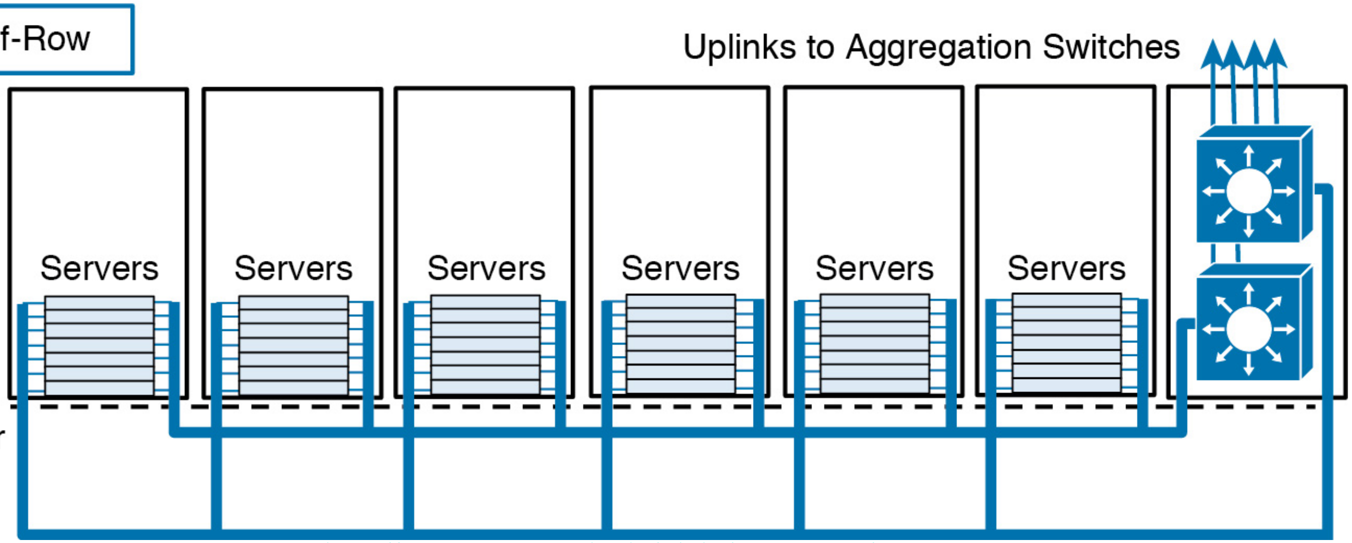
- Core routers manage traffic between aggregation switches and in/out of data center
- All switches below each pair of aggregation switches form a single layer-2 domain
- Each Layer 2 domain typically limited to a few hundred servers to limit broadcast
- Most traffic is internal to the data center.
- 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



Smaller cable between servers and switches

Network team has to manage switches on all racks



All network switches in one rack

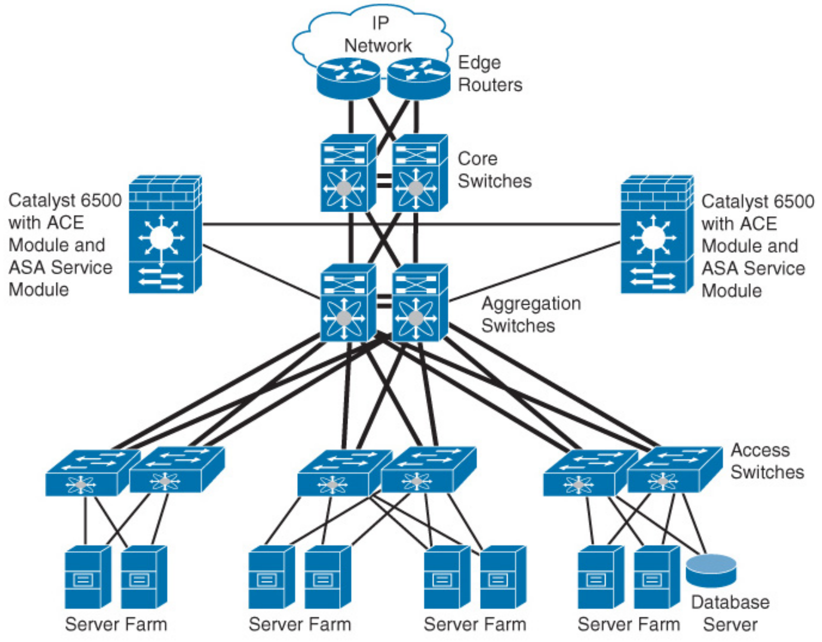
Source: Santana 2014

ToR vs EoR

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



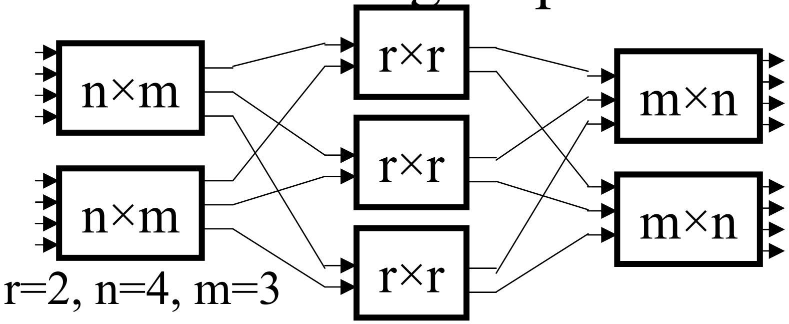
Source: Santana 2014

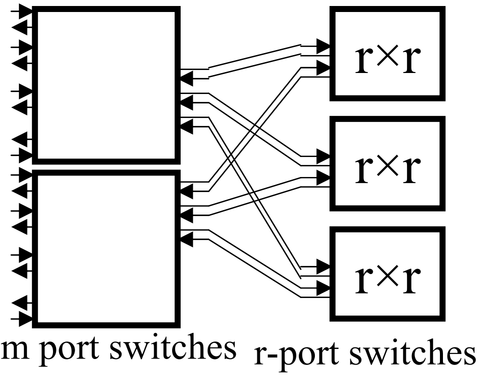
- Stateful devices (firewalls) on Aggregation layer
- Stateful= State of TCP connection

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

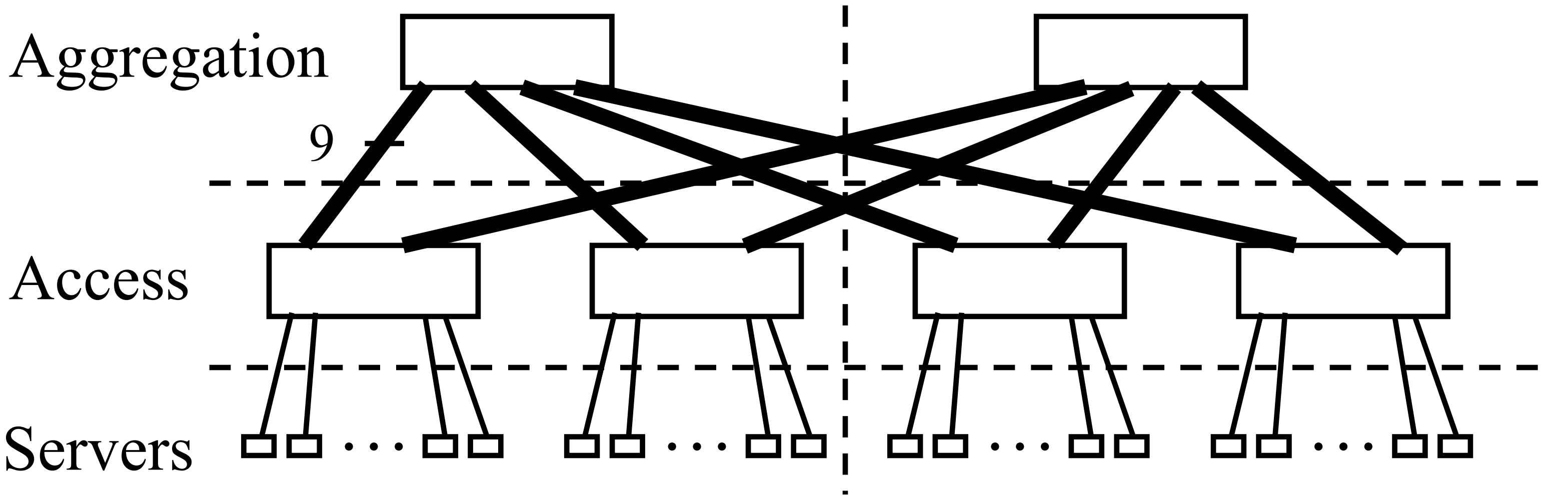
  The number of cross-points is reduced ⇒ Lower cost (then)
- 3-Stage Clos(n, m, r): ingress (r n×m), middle (m r×r), egress (r m×n)
- Strict-sense non-blocking if $m \geq 2n-1$. Existing calls unaffected.
- Rearrangeably non-blocking if $m \geq n$
- Can have any odd number of stages, e.g., 5
- **Folded**: Merge input and output in to one switch = Fat-tree





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

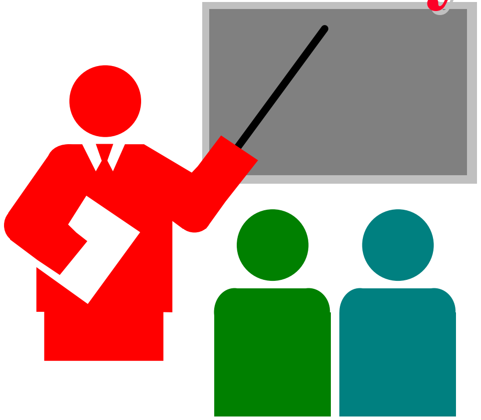
Fat-Tree DCN Example



- 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

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

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

Homework 3

- Draw a 3-stage clos(4,5,3) topology and its folded version.

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
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
NTP	Network Time Protocol

Acronyms (Cont)

NVGRE	Network Virtualization using Generic Routing Encapsulation
PFC	Priority Flow Control
PUE	Power Usage Effectiveness
RADIUS	Remote Authentication Dial-In User Service
RPC	Remote Procedue Call
RSW	Rack switch
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://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*
- 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