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/cse570-13/
Overview

1. Data Center Physical Layout
2. Data Center Network Topologies
3. ToR vs. EoR
4. Data Center Networking Issues
5. Data Center Networking Requirements
Google’s Data Center

Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/
Source: http://www.cse.wustl.edu/~jain/cse570-13/

©2013 Raj Jain
Cooling Plant

Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/
Washington University in St. Louis http://www.cse.wustl.edu/~jain/cse570-13/
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

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
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/cse570-13/ ©2013 Raj Jain
Unstructured Cabling

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

Source: http://webodyssuem.com/technologyscience/visit-the-googles-data-centers/
Washington University in St. Louis http://www.cse.wustl.edu/~jain/cse570-13/ ©2013 Raj Jain
Data Center Equipment Cabinets

Three Layers: Bottom: Signal, Middle: Power, Top: Fiber

Minimize patching between cabinets and racks

Cabling under raised floors provides better appearance and cooling

Washington University in St. Louis http://www.cse.wustl.edu/~jain/cse570-13/ ©2013 Raj Jain
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. ZDA allows easy
- 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,”
Data Center Network Topologies

- Core, Aggregation, Access

http://www.cse.wustl.edu/~jain/cse570-13/
©2013 Raj Jain
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.

Data Center Networks (Cont)

- Core routers manage traffic between aggregation routers 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

**Top-of-Rack**

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

**End-of-Row**

All network switches in one rack

Source: Santana 2014

Washington University in St. Louis

http://www.cse.wustl.edu/~jain/cse570-13/

©2013 Raj Jain
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
Access Aggregation Connections

1. **Looped Triangle:**

2. **Looped Square:**
   Oversubscription doubles if failure.

3. **Loop-Free U:**
   No L2 communication between aggregation switches if any switch links fail

4. **Loop-Free Inverted U:**
   Black-holes on some failures

Source: Santana 2014
Data Center Networking Issues

- Higher layers oversubscribed:
  - Other servers in the same rack 1:1
  - Uplinks from ToR: 1:2 to 1:20
    (e.g., 32x10Gb down, 8X10Gb up ⇒ 4:1 oversubscription)
  - Core Routers: 1:240
    ⇒ Generally keep services in one tree
    ⇒ Can't arbitrarily move servers

- Moving across Subnets is painful
  ⇒ Requires reconfiguration of IP addresses and VLAN trunks

- Service trample on each-other.
  Overuse by one service affects others

- Poor reliability.
  One access switch failure doubles the load on the other.
Data Center Networking Issues (Cont)

- Under-utilization.
  Even when multiple paths exist only one is used.

- ECMP (Equal Cost Multipath) is used by routers to spread traffic to next hops using a hash function. However, only 2 paths exist.
DCN Requirements

- Needs to be Scalable, Secure, Shared, Standardized, and Simplified (5 S's)
- Converged Infrastructure: Servers, storage, and network have to work together
- Workload Mobility: Large L2 domains required for VM mobility
- East-West Traffic: Significant server-to-server traffic as compared to server to user. One Facebook request required 88 cache looks, 35 database lookups, 392 backend RPC calls. Internet traffic 935X the http request/response [Farrington]
- Storage traffic on Ethernet: Congestion management on Ethernet

4-Post Architecture at Facebook

- Each rack switch (RSW) has up to 48 10G downlinks and 4-8 10G uplinks (10:1 oversubscription) to cluster switch (CSW)
- Each CSW has 4 40G uplinks – one to each of the 4 FatCat (FC) aggregation switches (4:1 oversubscription)
- 4 CSW’s are connected in a 10G×8 protection ring
- 4FC’s are connected in a 10G×16 protection ring
- No routers at FC. One CSW failure reduces intra-cluster capacity to 75%.


Washington University in St. Louis  http://www.cse.wustl.edu/~jain/cse570-13/  ©2013 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 (r n×m), middle (m r×r), egress (r m×n)
- Strict-sense non-blocking if m ≥ 2n-1. Existing calls unaffected.
- Rearrangeably non-blocking if m ≥ n
- Can have any odd number of stages, e.g., 5
- **Folded**: Merge input and output into one switch = Fat-tree

Washington University in St. Louis

Fat-Tree DCN Example

- 6 identical 36-port switches. All ports 1 Gbps. 72 Servers.
- Each edge switch connects to 18 servers.
  9 Uplinks to first core switch. Other 9 links to 2nd core switch.
- Throughput between any two servers = 1 Gbps using ECMP.
  Identical bandwidth 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/cse570-13/
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

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADC</td>
<td>Application Delivery Controller</td>
</tr>
<tr>
<td>ANSI</td>
<td>American National Standards Institute</td>
</tr>
<tr>
<td>BPE</td>
<td>Business Process Engineering</td>
</tr>
<tr>
<td>CSW</td>
<td>Core Switch</td>
</tr>
<tr>
<td>DCBX</td>
<td>Data Center Bridging eXtension</td>
</tr>
<tr>
<td>DCN</td>
<td>Data Center Network</td>
</tr>
<tr>
<td>DFS</td>
<td>Distributed File System</td>
</tr>
<tr>
<td>DHCP</td>
<td>Dynamic Host Control Protocol</td>
</tr>
<tr>
<td>DNS</td>
<td>Domain Name System</td>
</tr>
<tr>
<td>ECMP</td>
<td>Equal Cost Multipath</td>
</tr>
<tr>
<td>EDA</td>
<td>Equipment Distribution Area</td>
</tr>
<tr>
<td>EoR</td>
<td>End of Row</td>
</tr>
</tbody>
</table>
# Acronyms (Cont)

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETS</td>
<td>Enhanced Transmission Selection</td>
</tr>
<tr>
<td>EVB</td>
<td>Edge Virtual Bridge</td>
</tr>
<tr>
<td>FC</td>
<td>Fibre Channel</td>
</tr>
<tr>
<td>FSW</td>
<td>Fabric switch</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>HDA</td>
<td>Horizontal Distribution Area</td>
</tr>
<tr>
<td>LACP</td>
<td>Link Aggregation Control Protocol</td>
</tr>
<tr>
<td>LAG</td>
<td>Link Aggregation</td>
</tr>
<tr>
<td>LLDP</td>
<td>Link Layer Discovery Protocol</td>
</tr>
<tr>
<td>MAC</td>
<td>Media Access Control</td>
</tr>
<tr>
<td>MDA</td>
<td>Main Distribution Area</td>
</tr>
<tr>
<td>MW</td>
<td>Mega-Watt</td>
</tr>
<tr>
<td>NTP</td>
<td>Network Time Protocol</td>
</tr>
</tbody>
</table>
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 Procedure 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

Reading List (Cont)

- 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/Cable_management