Data Centers

Nelson L. S. da Fonseca
IEEE ComSoc Summer Scool
Albuquerque, July 17-21, 2017

Acknowledgement

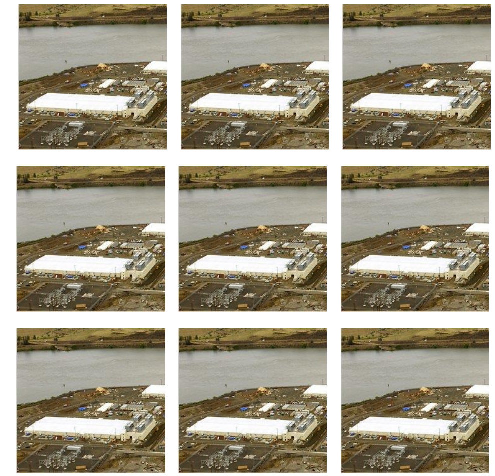
- Some slides in this set of slides were kindly provided by:
  - Raj Jain, Washington University in St. Louis
  - Dzmitry Kliazovich, University of Luxenbourg
  - EMC Corporation

Ever increasing processing needd

Server

Ever increasing processing need

Ever increasing processing needs

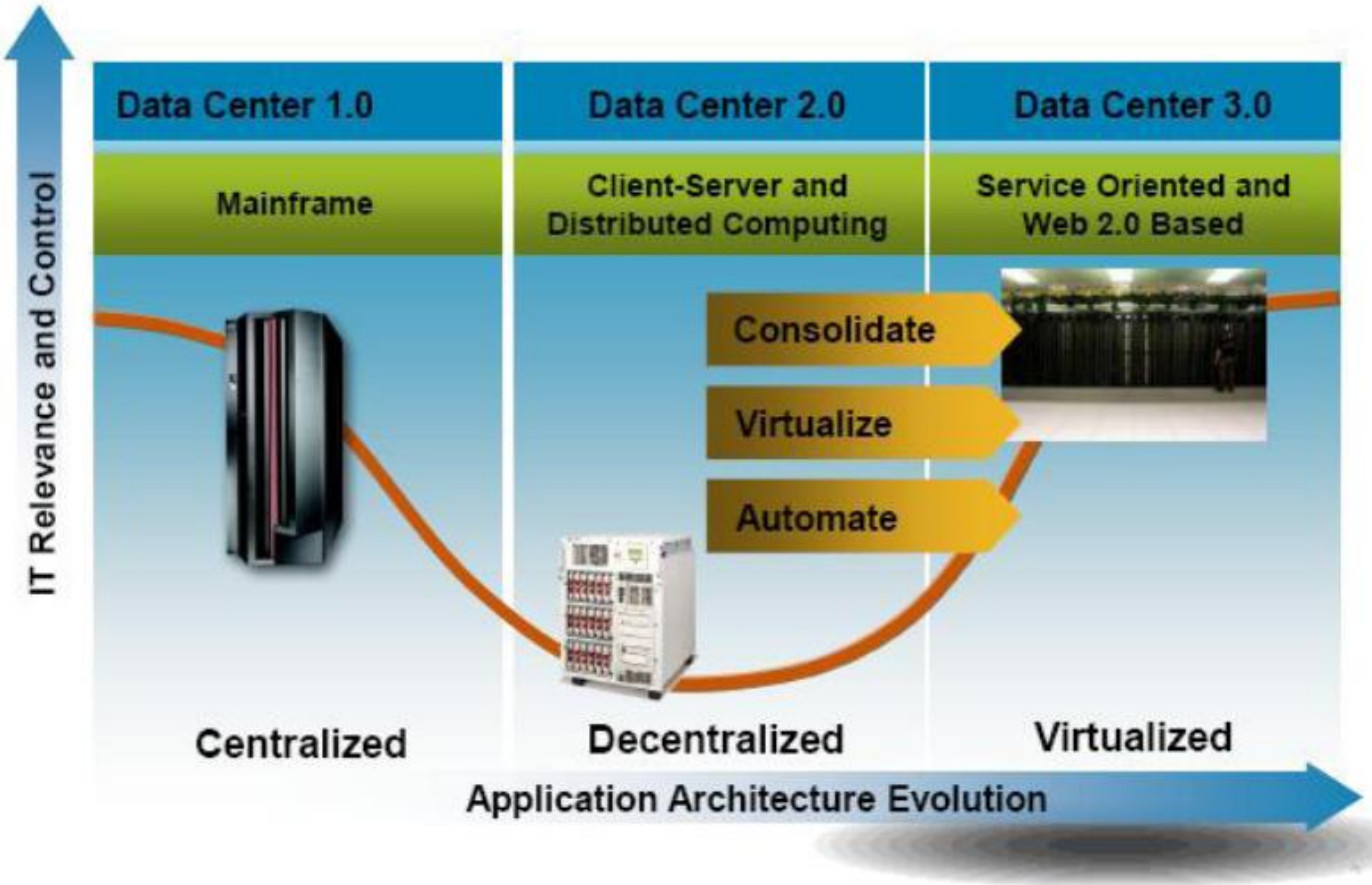


What is a Data Center?

- A data center is a facility used to house computer systems and associated components, such as networking and storage systems, cooling, uninterruptable power supply, air filters…
- A data center typically houses a large number of heterogeneous networked computer systems
- A data center can occupy one room of a building, one or more floors, or an entire building

Data Center

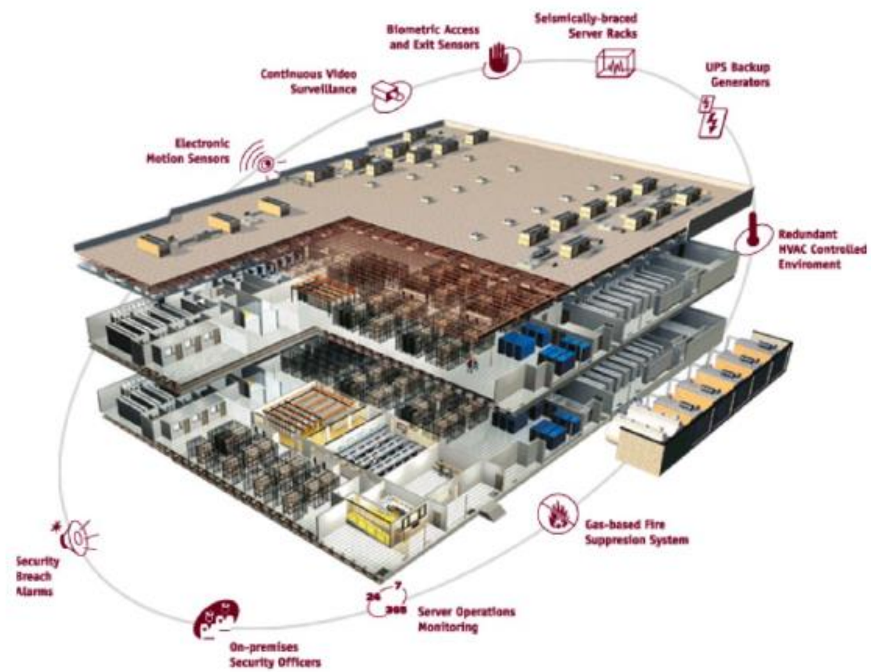
Data Center Evolution



Fonte: Cisco Data Center 3.0

Data Center Components

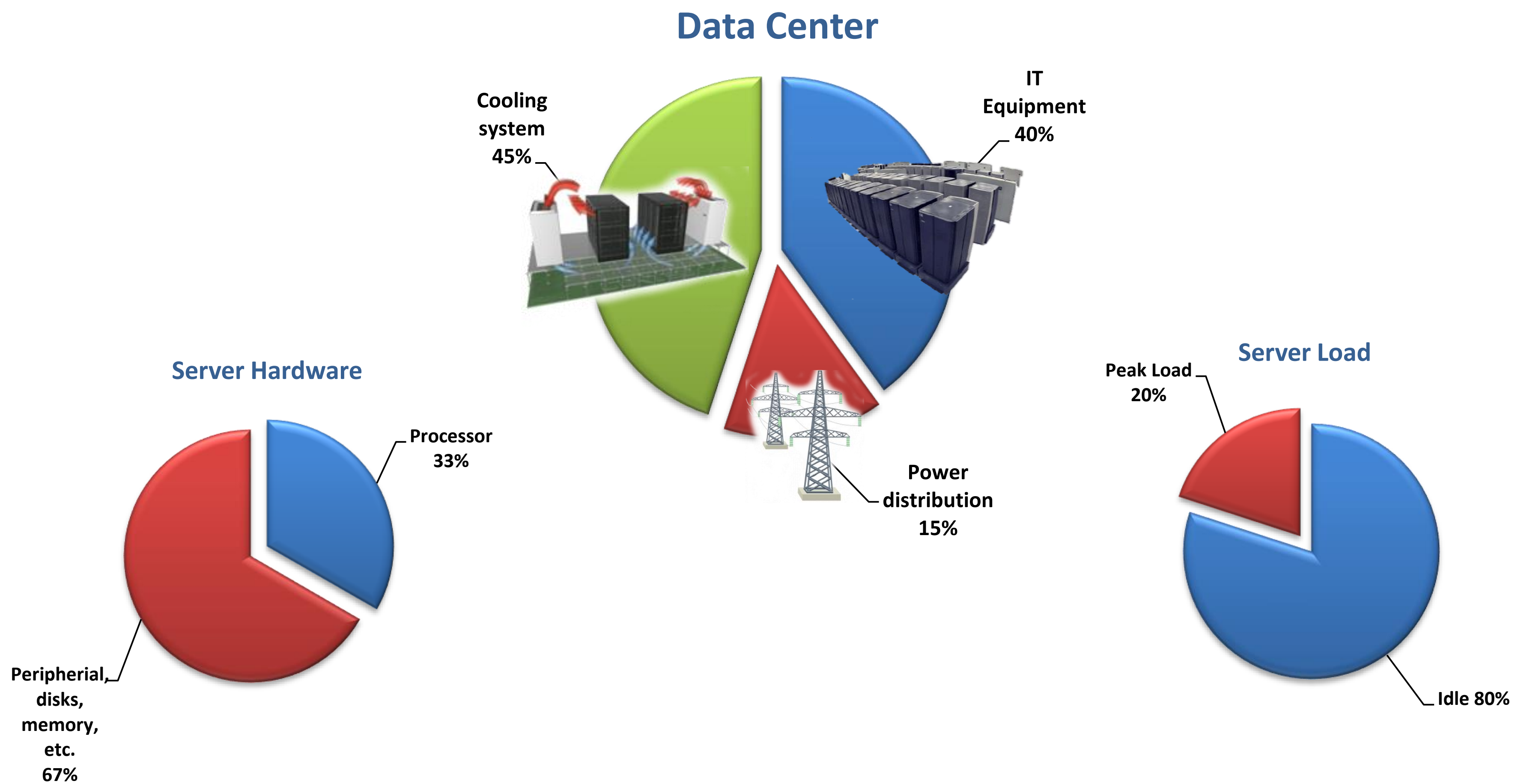
- Air conditioning
  - Keep all components in the manufacturer's recommended temperature range
- Redundant Power
  - UPS/Generators
  - Multiple power feeds
- Fire protection
- Physical security
- Monitoring Systems
- Connectivity



Energy Efficiency of Data Center

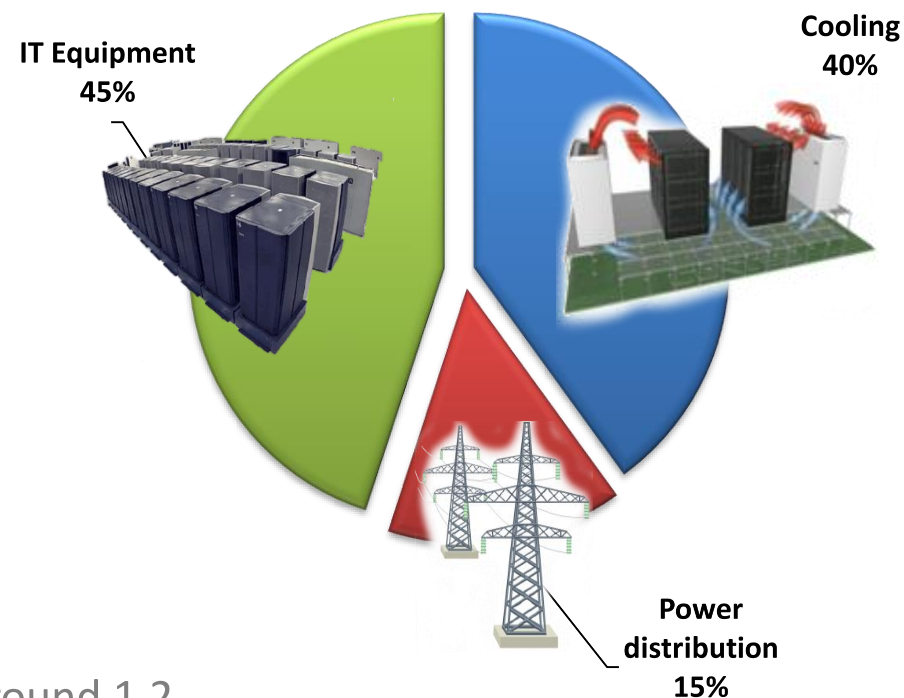
Data centers consume 1.5% of all electricity consumed in the world, but only 15-30% efficient

Energy Efficiency of Data Centers



Power Usage Effectiveness (PUE)

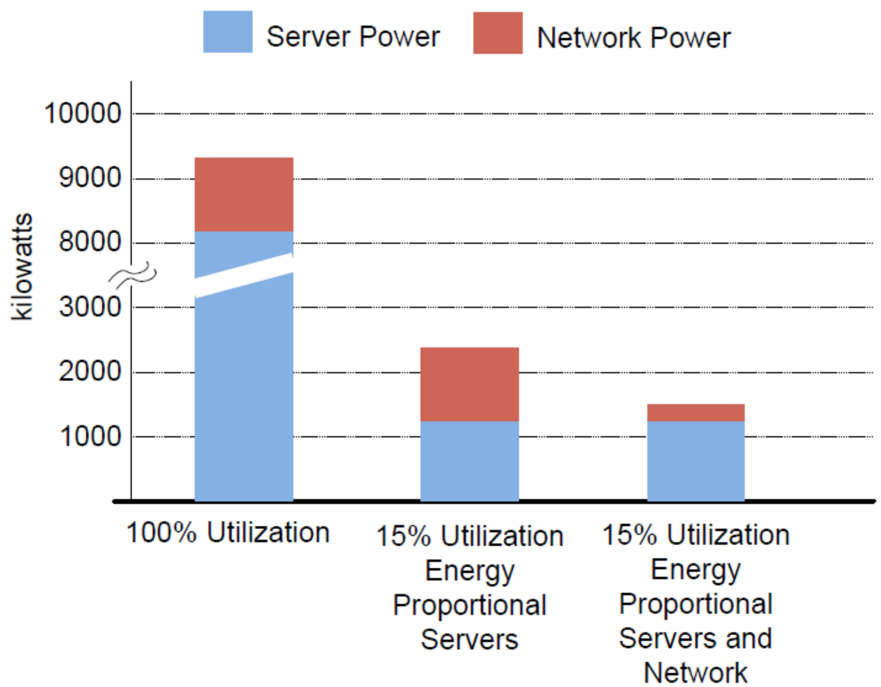
$$PUE = \frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy}$$



Modern data centers report PUE around 1.2

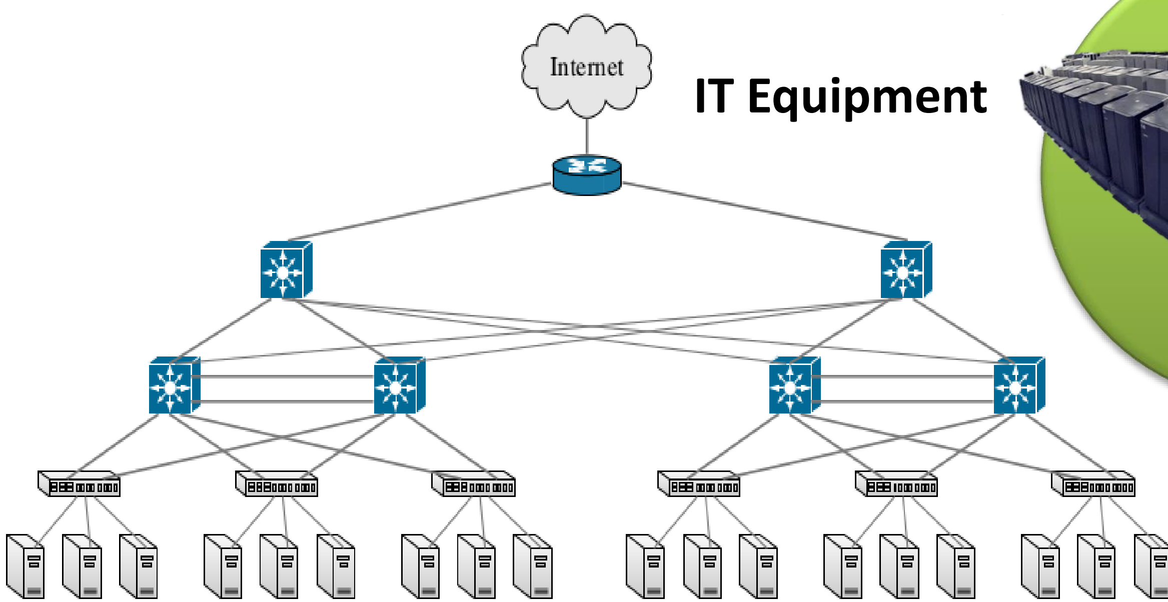
Energy Efficiency of Data Centers

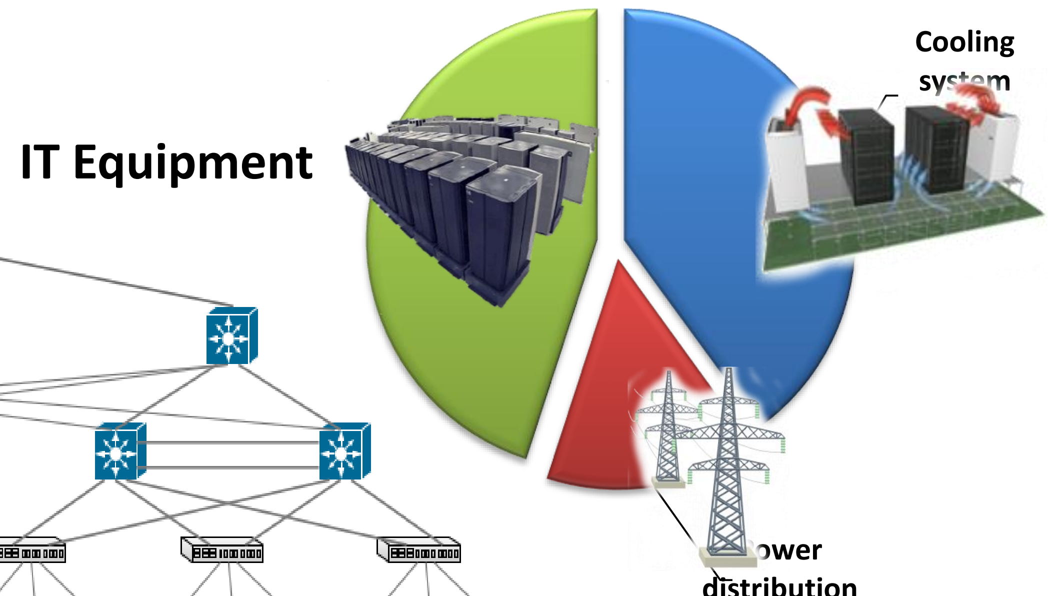
- Communication network consumes 30% to 50% of the total power used by the IT equipment



Power Usage Effectiveness (PUE)

$$PUE = \frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy}$$

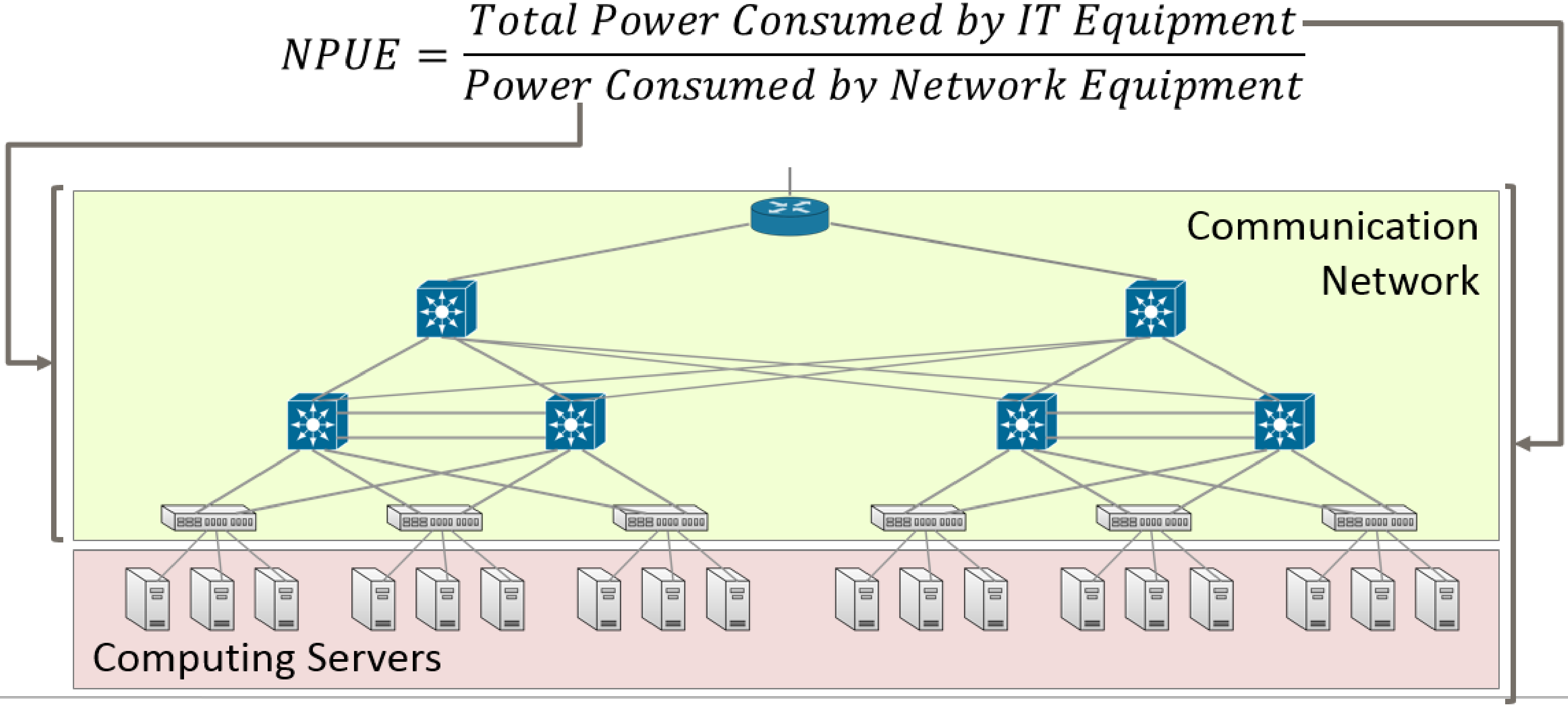




Network Power Usage Effectiveness (NPUE)

– Fraction of IT power spent to operate the network

$$NPUE = \frac{Total\ Power\ Consumed\ by\ IT\ Equipment}{Power\ Consumed\ by\ Network\ Equipment}$$



Communication Network Energy Efficiency (CNEE)

- Energy to Deliver a Single Bit of Information

$$CNEE = \frac{Power\ Consumed\ by\ Network\ Equipment}{Effective\ Network\ Throughput\ Capacity}$$

Joule/Bit

Data Center Classification

- Macro data centers
- Micro data centers
- Nano data centers
- Container data centers

Macro Data Centers

- 100.000 or more servers
- Energy consumption 10s Mega Watts
- Applications that demand large computational and storage capacity
- Amazon EC2, Windows Azure, Google AppEngine

How big are data centers?

Data Center Site	Sq footage
Facebook (Santa Clara)	86,000
Google (South Carolina)	200,000
HP (Atlanta)	200,000
IBM (Colorado)	300,000
Microsoft (Chicago)	700,000

Wembley Stadium:**172,000** square ft

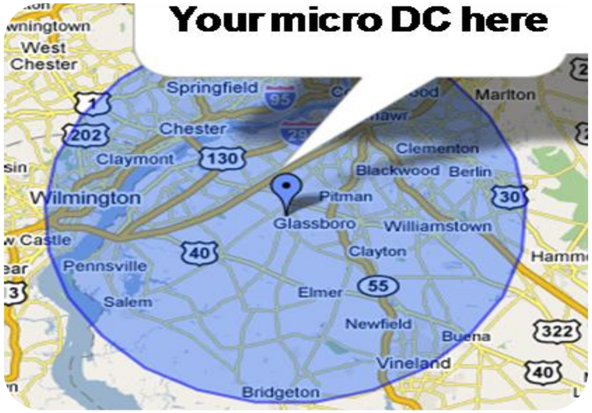
How big are data centers?

- Estimates on the size of global infrastructures of companies
  - Google: ~900,000 servers
  - Amazon: ~450,000 servers
  - Microsoft: ? servers (Chicago data center can hold up to 300k servers)
  - Rackspace: 79,805 servers (*officially disclosed*)

- List of known data center locations (*officially disclosed*)
  - North-America: 6 Google, 11 Amazon, 8 Microsoft and 5 Rackspace
  - Europe: 3 Google, 3 Amazon, 7 Microsoft and 2 Rackspace
  - Asia: 3 Google, 5 Amazon, 5 Microsoft and 1 Rackspace
  - Latin America: 1 Google, 2 Microsoft and 1 Amazon

Micro Data Centers

- 1000 or more servers
- Energy consumption order of 10s Kwatts
- Built close to urban areas
- Applications that require large exchange of data



Nano Data Center

- At end user facility
- P2P concept
- Distribute the data center functionality to several distributed users equipament
- Not the classical use of data center as we know it

Modular Data Centers

- Speed of deployment
- Lower capital and operational costs
- High mobility
- Increased cooling efficiency

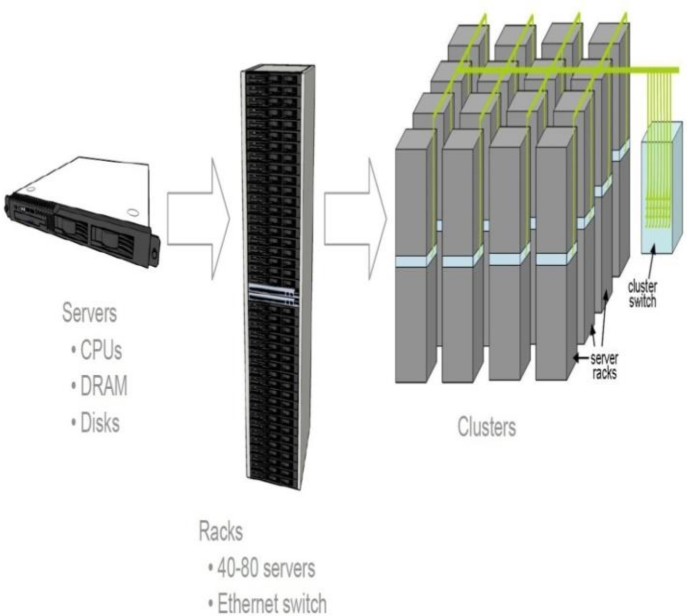
Modular Data Centers

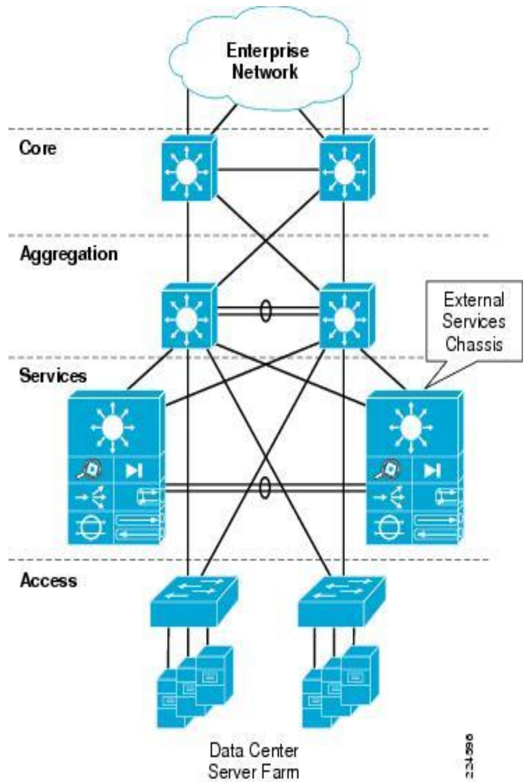
In a single container

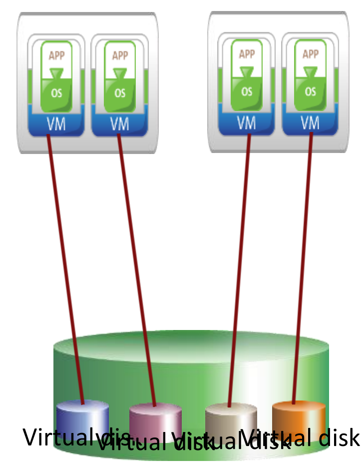
- Up to 46,080 cores
- 30 petabytes of storage
- Low cooling and energy costs

Data Center Components

- Servers + Network + Storage







VMFS Volume

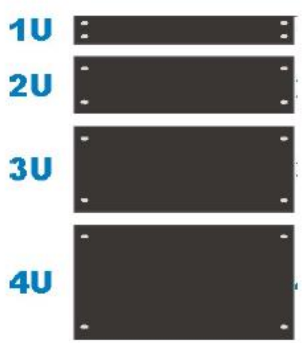
What is a Server?

- Servers are computers that provide “services” to “clients”
- They are typically designed for reliability and to service a large number of requests
- Organizations typically require many physical servers to provide various services (Web, Email, Database, etc.)
- Server hardware is becoming more powerful and compact

Racks

- Equipment (e.g., servers) are typically placed in racks
- Equipment are designed in a modular fashion to fit into rack units (1U, 2U etc.)
- A single rack can hold up to 42 1U servers

1U Server

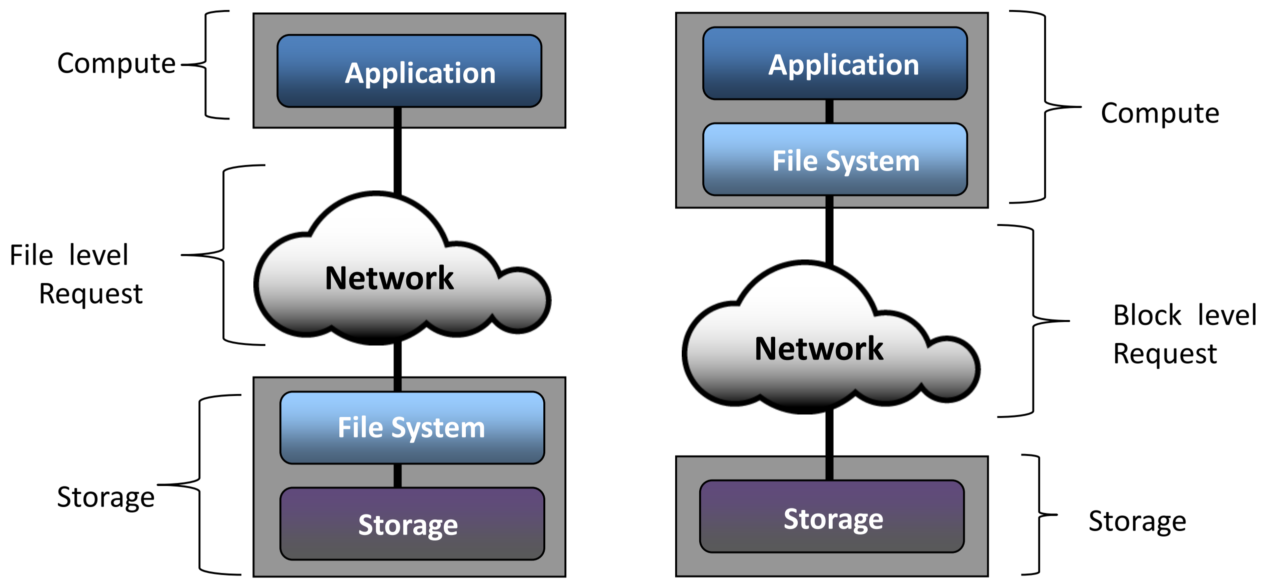


Blades and Blade Enclosures

- A blade server is a stripped down computer with a modular design
- A blade enclosure holds multiple blade servers and provides power, interfaces and cooling for the individual blade servers

Storage

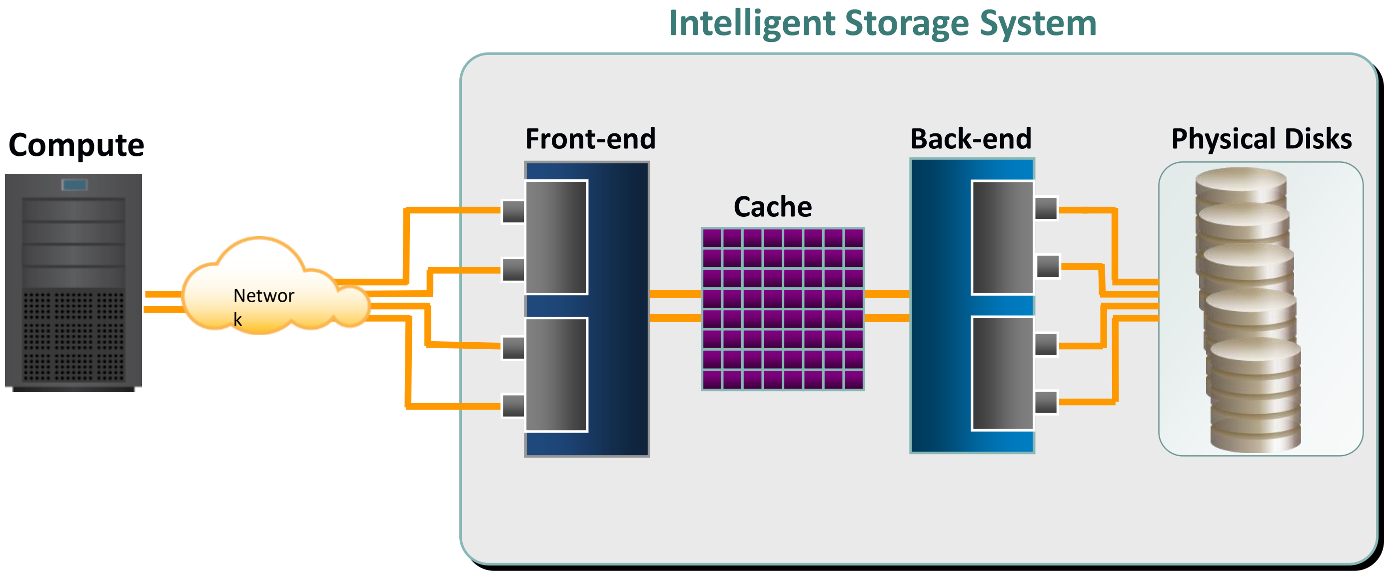
Data Access by Compute



File level Access

Block level Access

Components of an Intelligent Storage System



Compute to Storage Communication

Channel Technology	Network Technology
Compute system and peripheral devices are connected through channel	Compute system and peripheral devices are connected over a network
Provides low protocol overhead due to tight coupling	High protocol overhead due to network connection
Supports transmission only over short distances	Supports transmission over long distances
Protocol examples: PCI, IDE/ATA, SCSI, etc.	Protocol examples: iSCSI(SCSI over IP), FCoE (Fibre Channel over Ethernet), and FC

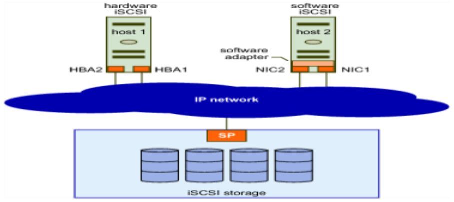
Communication Protocols

- Peripheral Component Interconnect (PCI)
  - Provides interconnection between CPU and attached devices
  - Latest PCI Express bus provides throughput of 133 MB/sec
- Integrated Device Electronics/Advanced Technology Attachment (IDE/ATA)
  - Popular protocol to connect to disk drives
  - Supports 16-bit parallel transmission
  - Serial version is called Serial ATA (SATA)
  - Both versions offer good performance at a relatively low cost

Small Computer System Interface (SCSI)

- Preferred storage connectivity option for high-end environments
- Improved performance, scalability, and high cost when compared to ATA
- Serial version is called Serial Attached SCSI (SAS)
- Used to connect disk drives and tapes to computer
- 8-16 devices on a single bus. Any number of hosts on the bus
- At least one host with host bus adapter (HBA)
- Each device on the SCSI bus has a "ID".
- Each device may consist of multiple logical units (LUNs).
- A direct access (disk) storage is addressed by a Logical Block

TCP/IP



- Transmission Control Protocol/Internet Protocol (TCP/IP)
  - Now used for compute to storage communication also
    - iSCSI (SCSI over IP) and FCoE (Fibre Channel over Ethernet)

Fiber Channel

- **Fibre Channel**, or **FC**, is a high-speed network technology (commonly running at 2-, 4-, 8- and 16-Gbps rates) primarily used to connect computer data storage
- primarily used in supercomputers, but has become a common connection type for storage area networks (SAN) in enterprise storage.

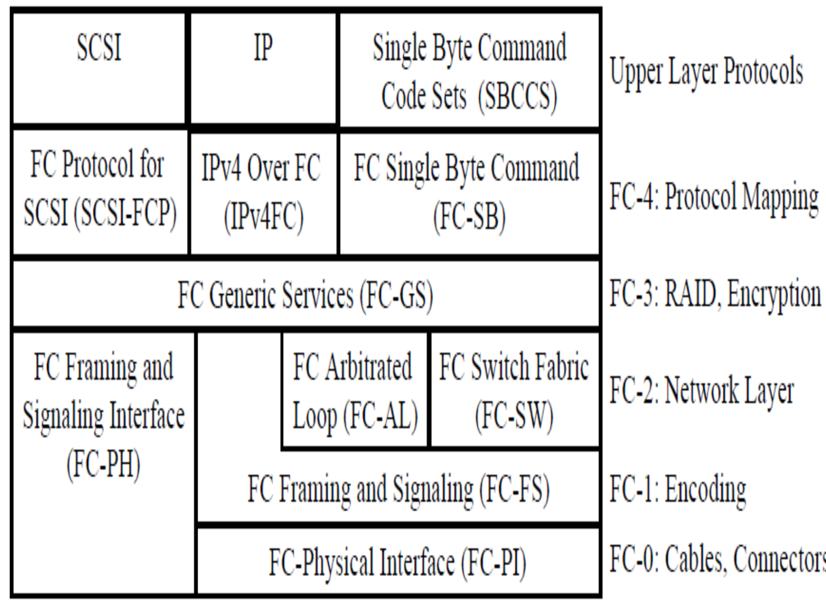
Fibre Channel Variants

NAME	Line-rate (gigabaud)	Line coding	Nominal throughput full duplex; MB/s	Net throughput per direction; MB/s[v 1][v 2]	Efficiency[v 3][v 2]	Availability
1GFC	1.0625	8b10b	200	99.6	78.6%	1997
2GFC	2.125	8b10b	400	199	78.6%	2001
4GFC	4.25	8b10b	800	398	78.6%	2004
8GFC	8.5	8b10b	1,600	797	78.6%	2005
10GFC	10.52	64b66b	2,400	1,195	95.3%	2008
16GFC	14.025	64b66b	3,200	1,593	95.3%	2011
32GFC	28.05		6,400			2016 (projected)
128GFC	4x28.05		25,600			2016 (projected)

Fiber Channel Protocols

Fibre Channel Protocol Layers

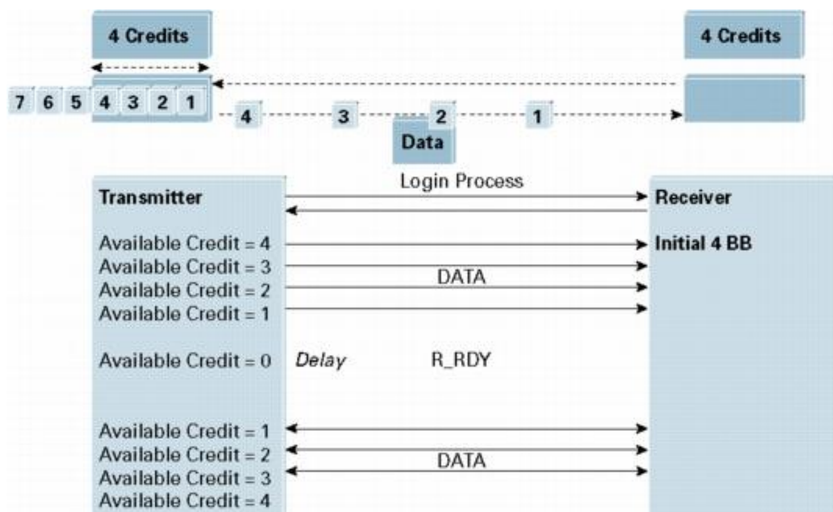
FC-4	Protocol Mapping	Existing and future network and channel protocols e.g. SCSI-3, IP, HIPPI, FC-AV, FISCON
FC-3	Common Services	Application specific layer for encryption, compression, RAID striping etc. – still under construction.
FC-2	Data Delivery	Framing, flow control protocols and classes of service
FC-1	Byte Encoding	IBM's 8b/10b encoding logic with guaranteed error rate of 10^{-12}, at gigabit speeds is less than 1 error every 16 seconds.
FC-0	Physical Layer	Either fiber optic or copper cabling



FC Flow Control

Transmitter sends frames only when allowed by the receiver - Credit-based flow control

- Both Hop-by-Hop and End-to-End



Data Center Networks

Requirements

- High Throughput
- High Availability
- Wide Scalability
- Low Latency
- Robustness

Communication In Data Centers

- Communication in data centers are most often based on networks running the IP protocol suite

- Data centers contain a set of routers and switches that transport traffic between the servers and to the outside world

- Traffic in today's data centers:
  - 80% of the packets stay inside the data center
  - Trend is towards even more internal communication

- Typically, data centers run two kinds of applications:
  - Outward facing (serving end-users)
  - Internal computation (data mining and indexing)

Communication Latency

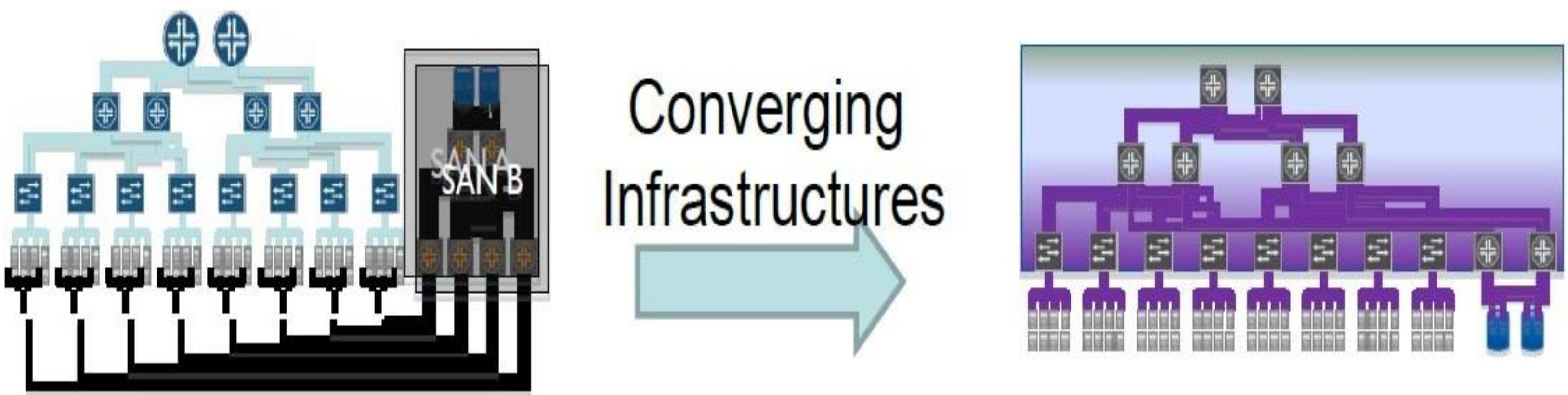
- Propagation delay in the data center is essentially 0
  - Light goes a foot in a nanosecond

- End to end latency comes from
  - Switching latency
    - 10G to 10G:~ 2.5 usec (store&fwd); 2 usec (cut-thru)
  - Queuing latency
    - Depends on size of queues and network load

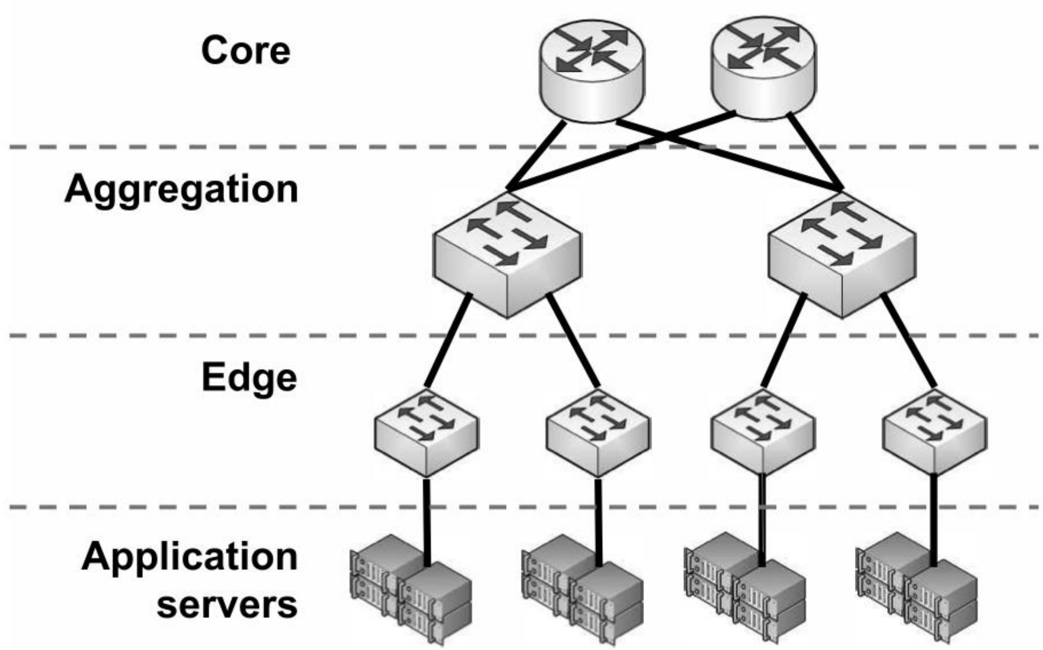
- Typical times across a quiet data center: 10-20usec

Converged Infrastructure

- Servers, storage, and network have to work together



Traditional DCN Topology



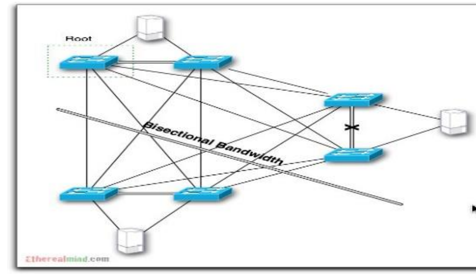
Traditional DCN

- Access switches connect to 2 aggregation switches
- Aggregation switches connect to 2 core routers
- 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

Limitations of traditional DCN

- Not suited for East-West traffic
- Incremental expansion hindered by rigid structure
- Coarse-grained failure domain
- Poor server-to-server capacity, capped by oversubscription
- Higher layers oversubscribed:
  - ✓ Servers in the same rack 1:1
  - ✓ Uplinks from ToR: 1:2 to 1:20
  - ✓ Core Routers: 1:240
- Limited bisection bandwidth (overloaded network core)
- Poor exploitation of multiple paths

Bissection Bandwidth



- Split N nodes into two groups of N/2 nodes such that the bandwidth between these two groups is minimum: that is the bisection bandwidth
- Why is it relevant: if traffic is completely random, the probability of a message going across the two halves is ½ – if all nodes send a message, the bisection bandwidth will have to be N/2

Current Application Requirements

Traffic Pattern
- Between servers (East-West) instead of client-server (North-South)

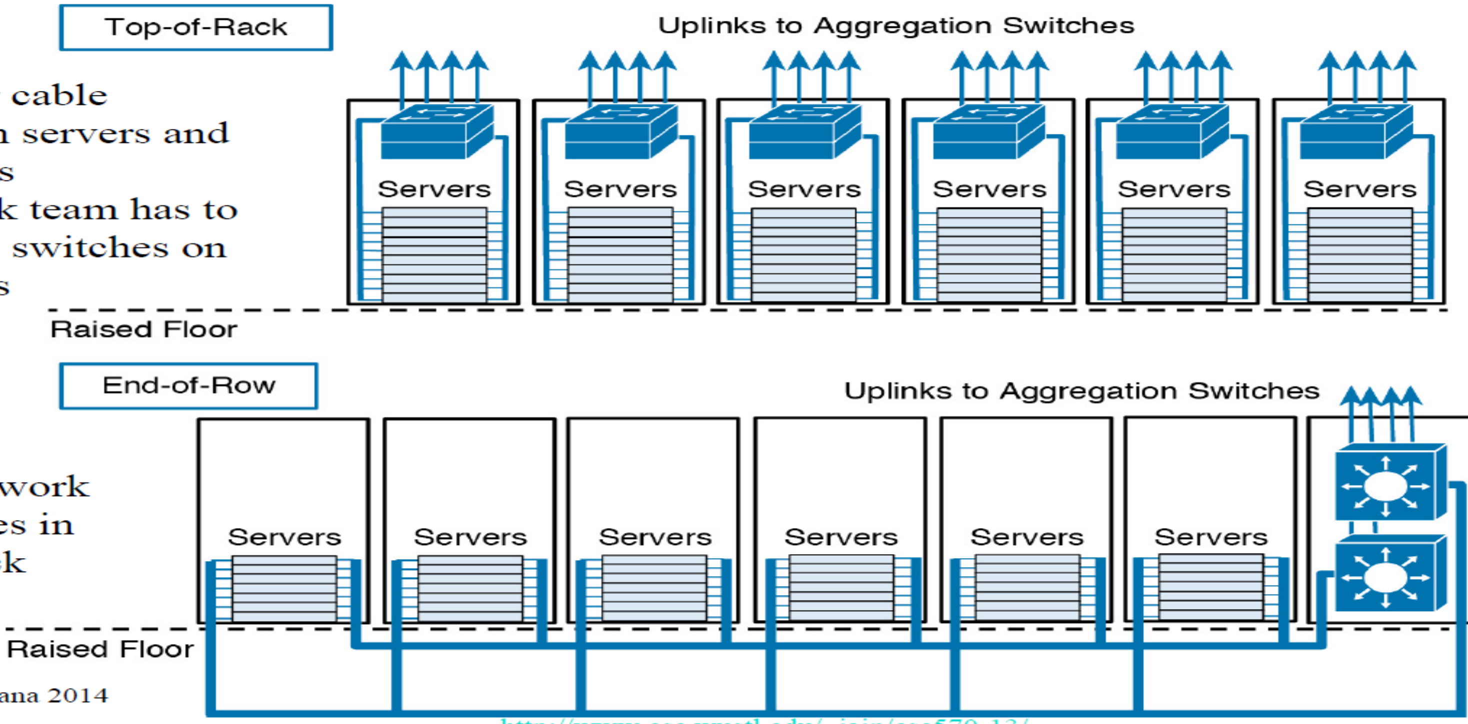
Scale
- 10s of thousands to 100s of thousands of endpoints

Agility
- New endpoints and racks powered up in hours instead of weeks
- New networks spun up in seconds instead of weeks

Flexibility
- Ability to reuse same infrastructure for different applications

Resilience
- Fine grained failure domain

Switch Locations

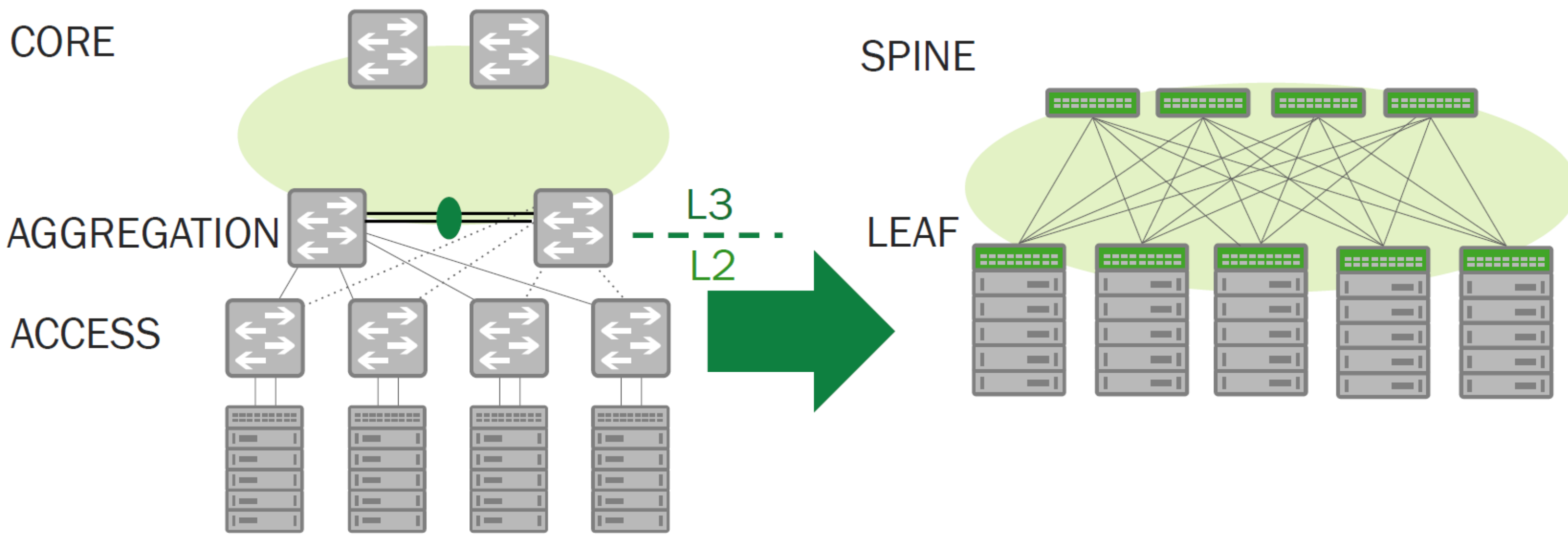


Smaller cable between servers and switches
Network team has to manage switches on all racks

All network switches in one rack

Source: Santana 2014

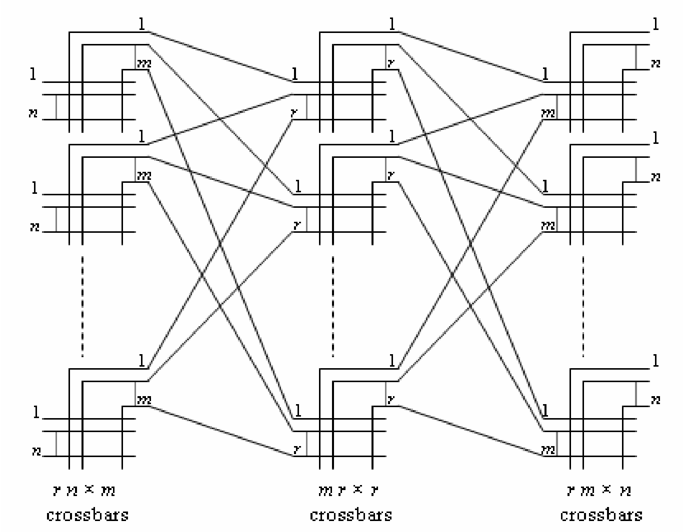
DCN Transformation



Dinesh Dutt, "The House That CLOS Built Network Architecture For the Modern Data Center"

Clos Networks

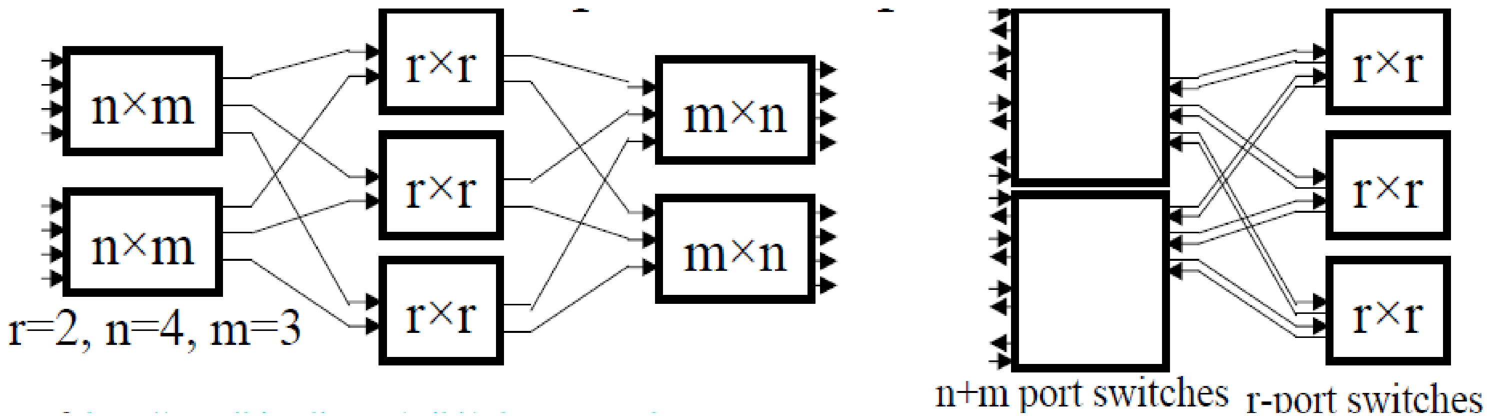
- Multi-stage circuit switching network proposed by Charles Clos in 1953 for telephone switching systems
- 3-Stage Clos(n, m, r): ingress (r n×m), middle (m r×r), egress (r m×n)



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

Fat Tree

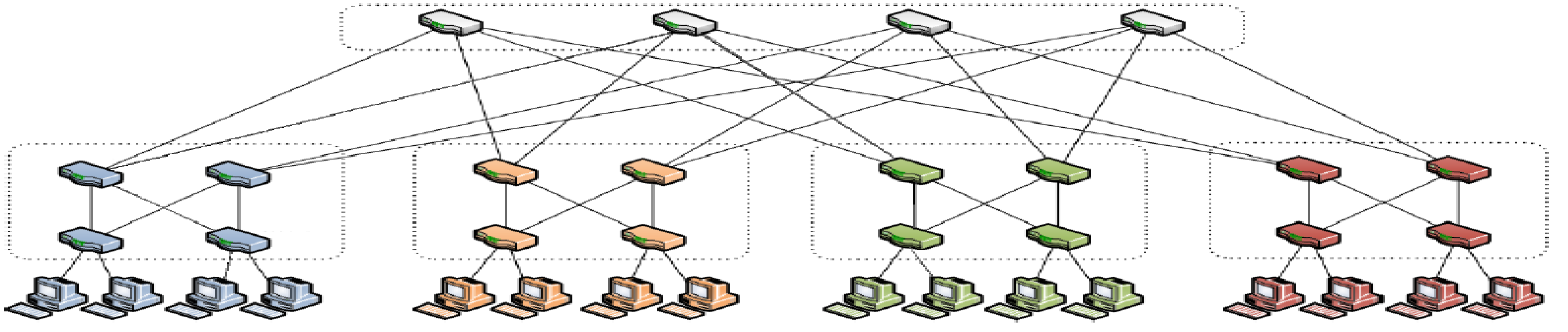
- Merge input and output in one switch



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

Fat Tree

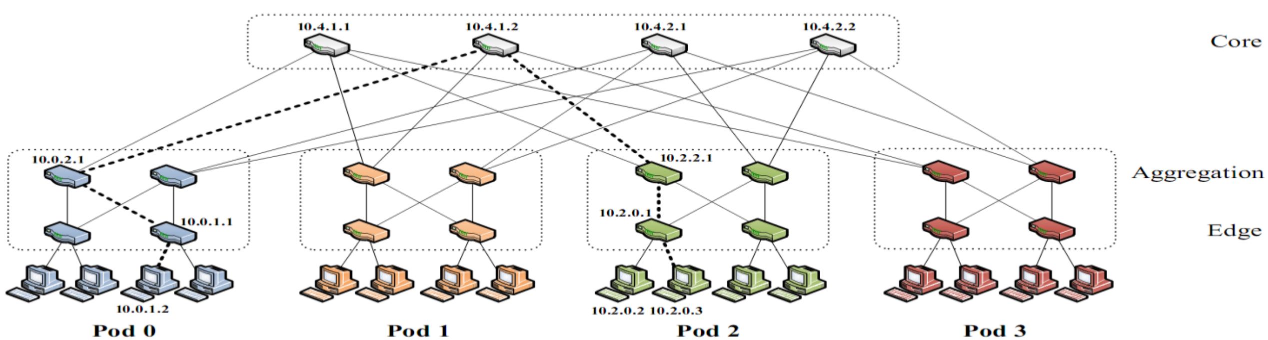
- Every level is fully connected to lower and upper levels
- Provides higher fault-tolerance and richer connectivity



Fat-Tree topology (adapted from [Al-Fares et al., 2008])

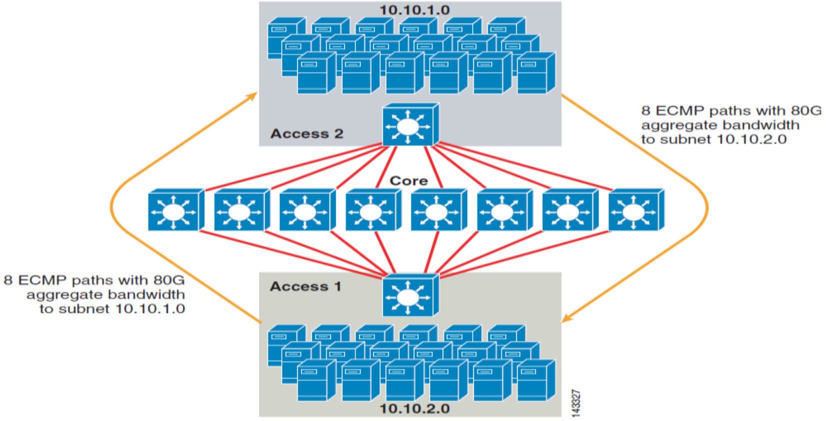
Fat-Tree Based DC Architecture

- K-ary fat tree: three-layer topology (edge, aggregation and core)
- Each pod consists of $(k/2)^2$ servers & 2 layers of k/2 k-port switches
- Each edge switch connects to k/2 servers & k/2 aggr. switches
- Each aggr. switch connects to k/2 edge & k/2 core switches
- $(k/2)^2$ core switches: each connects to k pods



Equal cost multi-path (ECMP) routing

- Equal cost multi-path (ECMP) routing
  - Load balancing technology that optimizes flows across multiple IP paths between any two subnets
  - Applies load balancing for TCP and UDP packets on a per-flow basis
  - ICMP is distributed on a packet-by-packet basis
  - ECMP is based on RFC



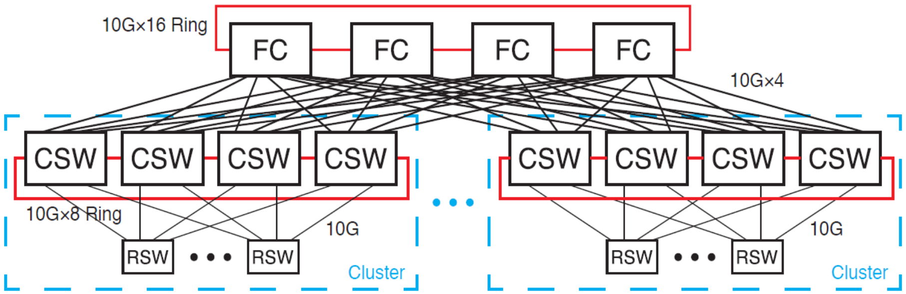
Fat-Tree

- Fat tree has identical bandwidth at any bisections
- Each layer has the same aggregated bandwidth
- Can be built using cheap devices with uniform capacity
- Each port supports same speed as end host
- Scalability: k-port switch supports $k^3/4$ servers/hosts: (k/2 hosts/switch * k/2 switches/pod * k pods)
  - For commonly available 24, 36, 48 and 64 ports commodity switches, such a Fat-tree structure is limited to sizes 3456, 8192, 27648 and 65536, respectively.
- switch-oriented, which might not be sufficient to support 1-to-x traffic.

M Al-Fares , A. Loukissas , A. Vahdat, "A Scalable, Commodity Data Center Network Architecture"

Facebook DCN

- 4-post DCN: Rack Switch (RSW), Cluster Switch (CSW), Fatcat Switch (FC)
- Each RSW has up to 48 10G downlinks and 4-8 10G uplinks(10:1 oversubscription) to CSW
- Each CSW has 4 40G uplinks – one to each of the 4 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



Google´s DCN

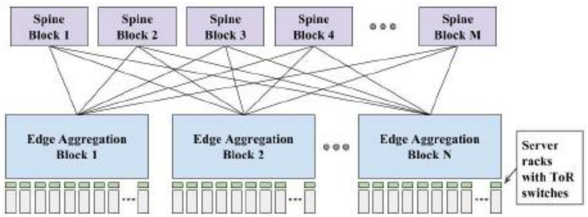


Figure 4: A generic 3 tier Clos architecture with edge switches (ToRs), aggregation blocks and spine blocks. All generations of Clos fabrics deployed in our datacenters follow variants of this architecture.

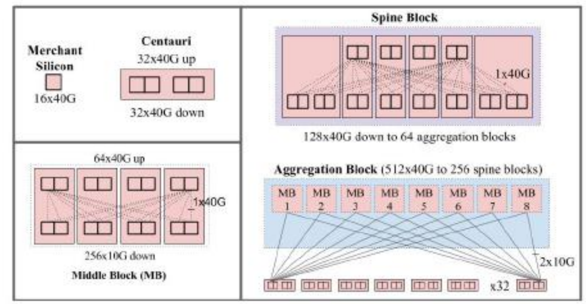


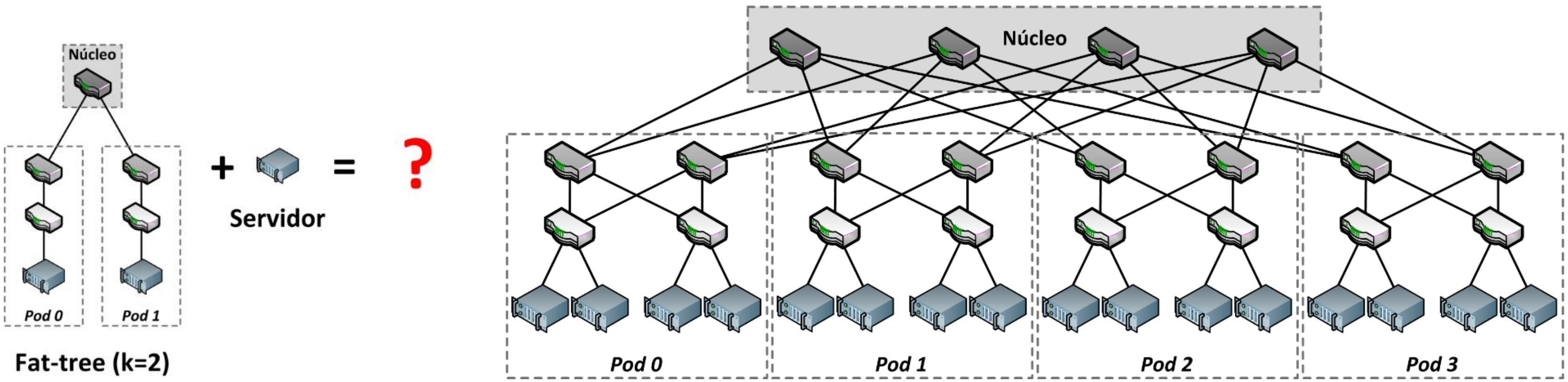
Figure 13: Building blocks used in the Jupiter topology.

Datacenter Generation	First Deployed	Merchant Silicon	ToR Config	Aggregation Block Config	Spine Block Config	Fabric Speed	Host Speed	Bisection BW
Four-Post CRs	2004	vendor	48x1G	-	-	10G	1G	2T
Firehose 1.0	2005	8x10G 4x10G (ToR)	2x10G up 24x1G down	2x32x10G (B)	32x10G (NB)	10G	1G	10T
Firehose 1.1	2006	8x10G	4x10G up 48x1G down	64x10G (B)	32x10G (NB)	10G	1G	10T
Watchtower	2008	16x10G	4x10G up 48x1G down	4x128x10G (NB)	128x10G (NB)	10G	nx1G	82T
Saturn	2009	24x10G	24x10G	4x288x10G (NB)	288x10G (NB)	10G	nx10G	207T
Jupiter	2012	16x40G	16x40G	8x128x40G (B)	128x40G (NB)	10/40G	nx10G/ nx40G	1.3P

Table 2: Multiple generations of datacenter networks. (B) indicates blocking, (NB) indicates Nonblocking.

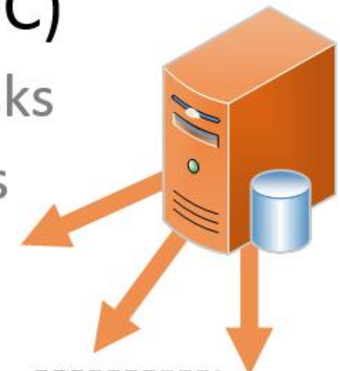
Singh, Arjun, et al. "Jupiter rising: a decade of clos topologies and centralized control in Google's datacenter network." *Communications of the ACM* 59.9 (2016): 88-97.

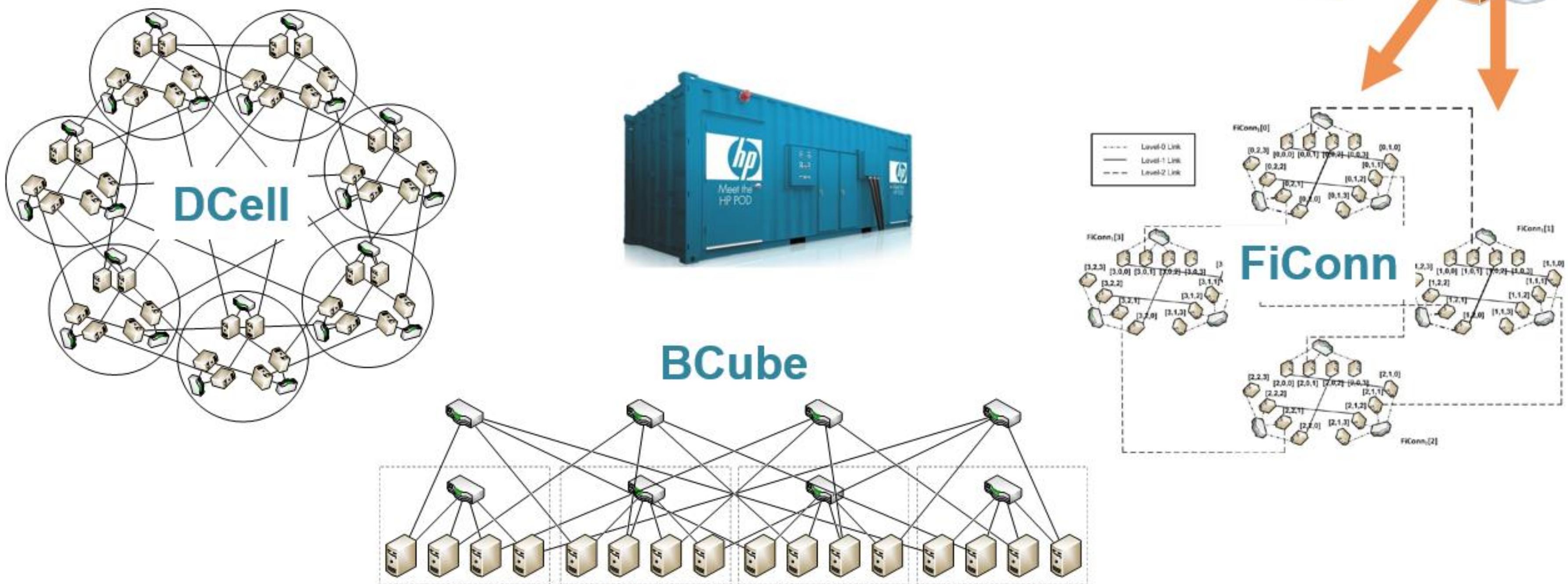
DCN Expansion



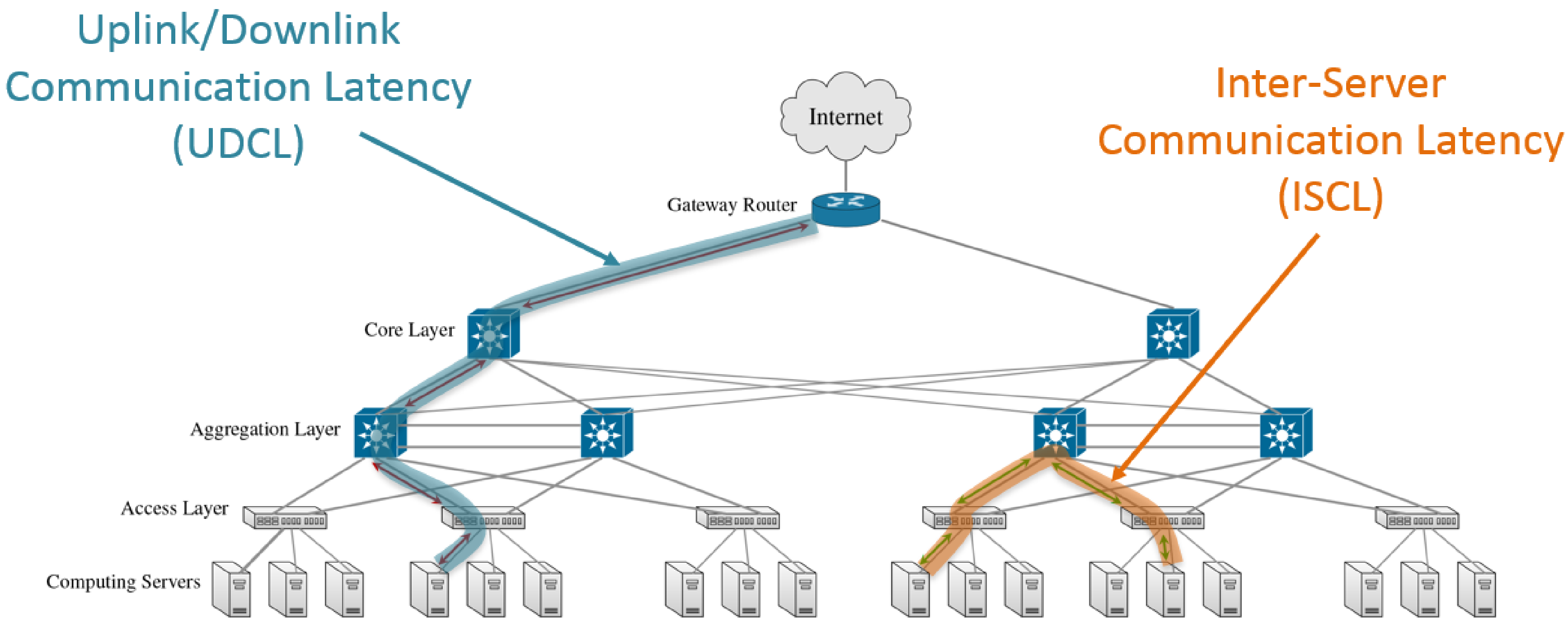
L. H. M. K. Costa, M. D. de Amorim,M. El. M. Campista, M. G. Rubinstein,P. Florissi and O. C. M. B. Duarte, Grandes Massas de Dados na Nuvem: Desafios e Técnicas para Inovação, SBRC, 2012

- Average Server Degree of Connectivity (ASDC)
  - Relevant for Hadoop and parallel distributed tasks
  - Effective in distributed data center architectures

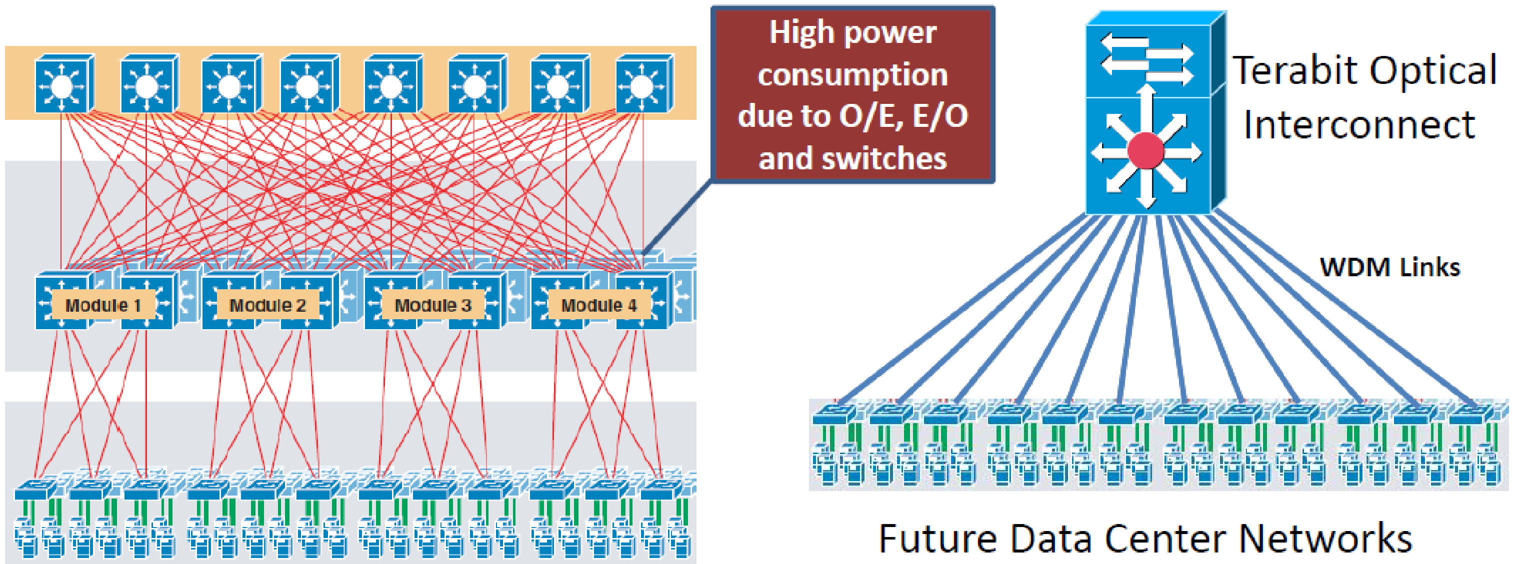




- Communication Latency in Data Centers



Hybrid Data Center Networking



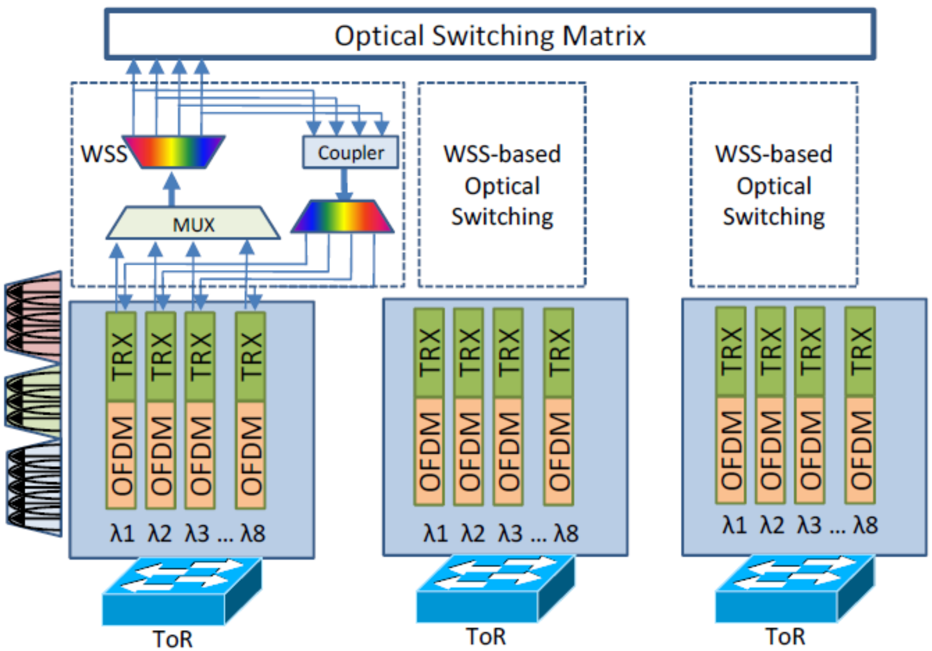
Current

Future Data Center Networks

[Source: Cisco, “Petabit Optical Switch for Data Center Networks”, “Scaling Networks in Large Data Centers”, Facebook

Christoforos Kachris and Ioannis Tomkos "Optical interconnection networks for data centers", ONDM 2013

Hybrid Data Center Network



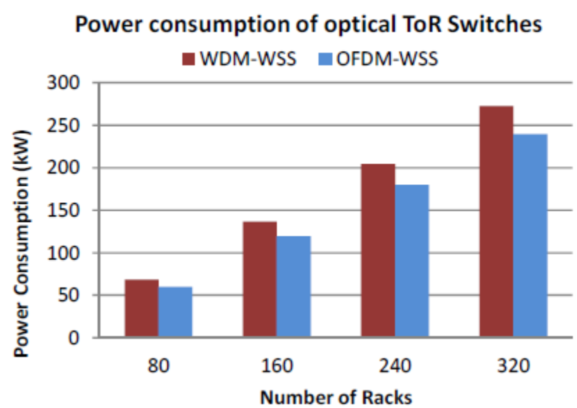


Figure 5. Power consumption of OFDM-WSS

Christoforos Kachris and Ioannis Tomkos "Optical interconnection networks for data centers", ONDM 2013

Research Challenge

Need of high-radix, scalable, energy efficient Data Centers that can sustain the exponential increase of the network traffic.

Review questions

- What is “oversubscription” in DCNs? What are the implications?

- Identify two limitations with current data-center topologies.

Ethernet in Data Centers

Raj Jain´s quiz

T F

□ □ Ethernet uses CSMA-CD

□ □ Ethernet bridges use spanning tree for packet forwarding

□ □ Ethernet does not provide any delay guarantee

□ □ Ethernet has no congestion control

□ □ Ethernethas strict priority

Raj Jain´s quiz

T F

☐ ☑ Ethernet uses CSMA-CD

☐ ☑ Ethernet bridges use spanning tree for packet forwarding

☐ ☑ Ethernet does not provide any delay guarantee

☐ ☑ Ethernet has no congestion control

☐ ☑ Ethernethas strict priority

Ethernet

	Residential	Data Center
Distance	Up to 200m	No limit
Scale	Few MAC address 4096 VLANs	Millions of MAC address Millions of VLANs Q-in-Q
Protection	Spanning tree	Rapid Spanning tree
Path	Determined by spanning tree	Traffic engineered path
Service	Simple	Service Level Agreement
Priority	Priority	Per flow/per class QoS
Performance	No monitoring	Needs monitoring

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

Spanning Tree Protocol

Algorhyme (por Radia Perlman)

"I think that I shall never see a graph more lovely than a tree.

A tree whose crucial property is loop-free connectivity.

A tree that must be sure to span so packets can reach every LAN.

First, the root must be selected.

By ID, it is elected.

Least-cost paths from root are traced.

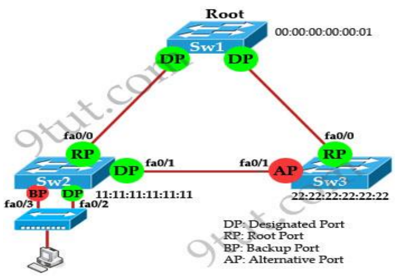
In the tree, these paths are placed.

A mesh is made by folks like me,

then bridges find a spanning tree."

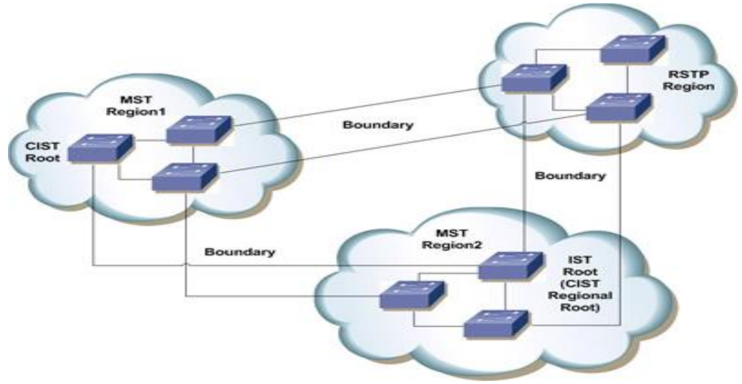
https://www.youtube.com/watch?v=iE_AbM8ZykI

Rapid Spanning Tree Protocol (RSTP)



- IEEE 802.1w
- In the Spanning Tree Protocol (IEEE 802.1D), a change can cause 1 minute of traffic loss which implies the loss of all TCP connections
- RSTP sends Hellos every 2 second rather than on topology change as in STP
- RTSP merges three port states (Disabled, blocking, listening) in to one (discarding)
- RTSP usus only full-duplex links
- New Bridge Protocol Data Unit (BPDU) fields allows rapid configuration to change

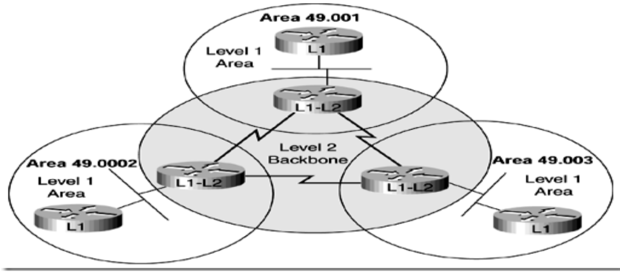
Multiple Spanning Tree Protocol



- Each tree serves a group of VLANs
- A bridge port can be in forwarding state for some VLANs and blocked state for others
- IEEE 802.1aq

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

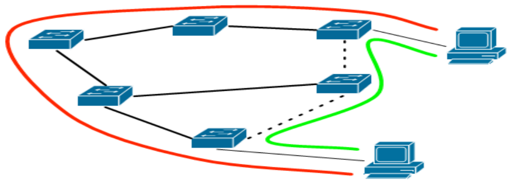
IS-IS Protocol



- Intermediate System to Intermediate System (IS-IS) is a protocol to build routing tables. Link-State routing protocol
- Dijkstra's algorithm is used by each node to build its routing table.
- OSPF is designed for IPv4 and then extended for IPv6. IS-IS is general enough to be used with any type of addresses
- OSPF is designed to run on the top of IP IS-IS is general enough to be used on any transport - Adopted by Ethernet

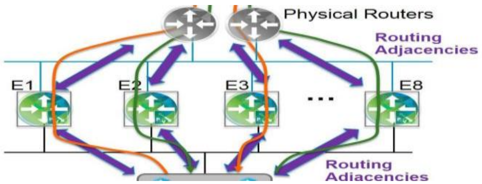
http://www.cse.wustl.edu/~jain/cse570-13/

Shortest Path Bridging



- Shortest Path Bridging is the replacement for the older spanning tree protocols (STP, MSTP, RTSP) that permitted only a single path toward the root bridge and blocked any redundant paths that could result in a layer 2 loop.
- SPB allows all paths to be active with multiple equal cost paths, and provides much larger layer 2 topologies (up to 16 million compared to the traditional virtual local area network (VLAN) limit of 4,096 specified in the IEEE standard 802.1Q)

Shortest Path Bridging



- IS-IS link state protocol is used to build shortest path trees for each node to every other node within the SPB domain
- Uses Equal-cost multi-path routing (ECMP) which is a routing strategy where next-hop packet forwarding to a single destination can occur over multiple "best paths" which tie for top place in routing metric calculations. It potentially offers substantial increases in bandwidth by load-balancing traffic over multiple paths;
- This symmetric and end to end ECMT behavior gives IEEE 802.1aq a highly predictable behavior and off line engineering tools can accurately model exact data flows

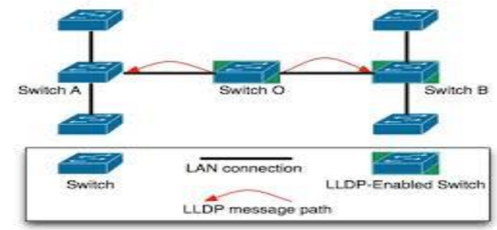
Shortest Path Bridging

- During the 2014 Winter games this fabric network was capable of handling up to 54 Tbit/s of traffic
- In 2013 and 2014 SPB was used to build the InteropNet backbone with only 1/10 the resources of prior years

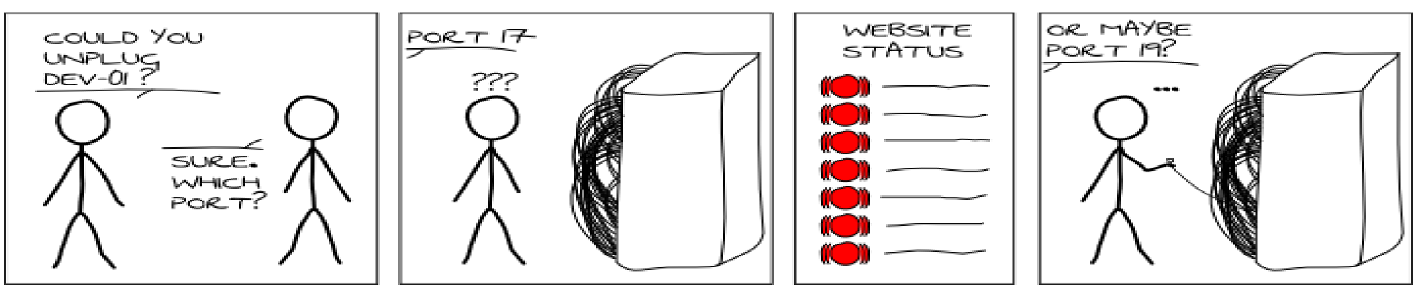
Shortest Path Bridging

- Advantages:
  - the ability to use all available physical connectivity, because loop avoidance uses a Control Plane with a global view of network topology
  - fast restoration of connectivity after failure, again because of Link State routing's global view of network topology
  - under failure, the property that only directly affected traffic is impacted during restoration; all unaffected traffic just continues
  - rapid restoration of broadcast and multicast connectivity, because IS-IS floods all of the required information in the SPB extensions to IS-IS, thereby allowing unicast and multicast connectivity to be installed in parallel

Link Layer Discovery Protocol



- Neighbor discovery by periodic advertisements
- Every minute a LLC frame is sent on every port to neighbors
- LLDP frame contains information in the form of Type-Length-Value (TLV)
- IEEE 802.1AB-2009



http://www.eetimes.com/document.asp?doc_id=1272069

Data Center Bridging

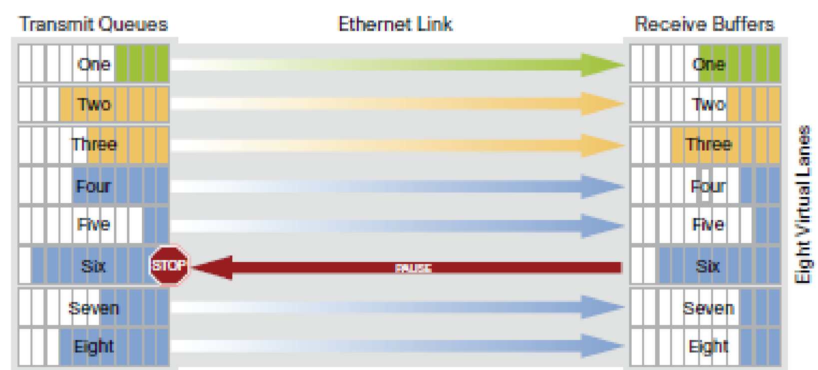
- Expand Ethernet networking and management to provide next-generation infraestructure to data centers

Feature	Benefit
Priority-based Flow control (PFC; IEEE 802.1 Qbb)	Provides capability to manage bursty, single traffic source on a multiprotocol link
Enhanced transmission selection (ETS; IEEE 802.1 Qaz)	Enables bandwidth management between traffic types for multiprotocol links
Congestion notification (IEEE 802.1 Qau)	Addresses the problem of sustained congestion by moving corrective action to the network edge
Data Center Bridging Exchange (DCBX) Protocol	Allows autoexchange of Ethernet parameters between switches and endpoints

http://www.cisco.com/c/dam/er

Priority Based Flow Control (PFC)

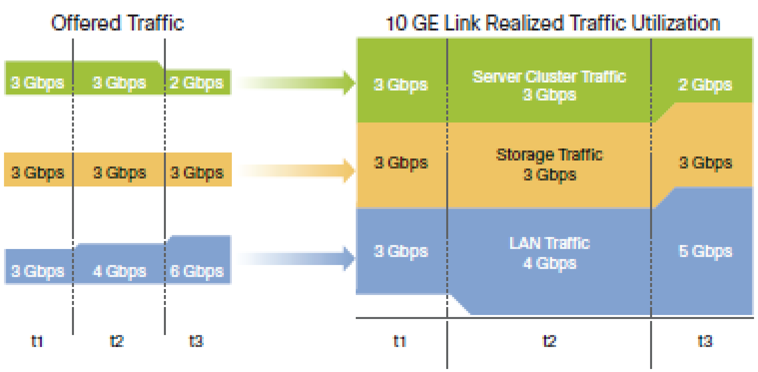
- Link sharing is crucial to I/O consolidation
- PFC divides a link in eight virtual links; their flow can be individually controlled so that there is no inter-flow performance interference.
- IEEE 802.Qbb



http://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/ieee-802-1-data-center-bridging/at_a_glance_c45-460907.pdf

Enhanced Transmission Selection

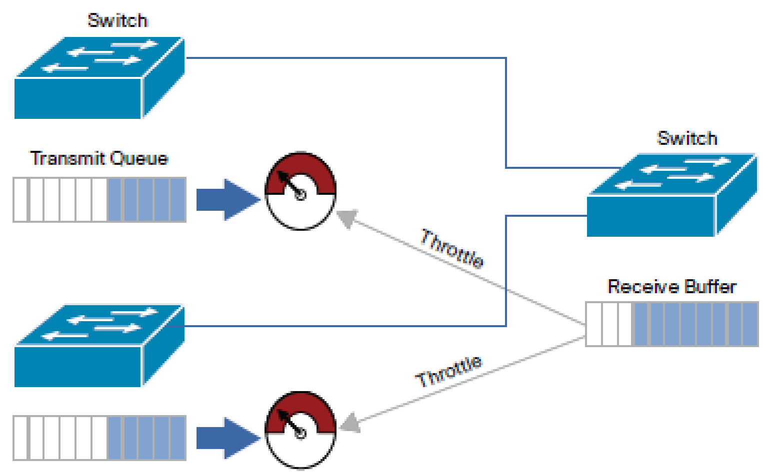
- One PFC flow can be divided in eight classes for optimal bandwidth management
- Classes can be grouped
- IEEE 802.11Qaz



http://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/ieee-802-1-data-center-bridging/at_a_glance_c45-460907.pdf

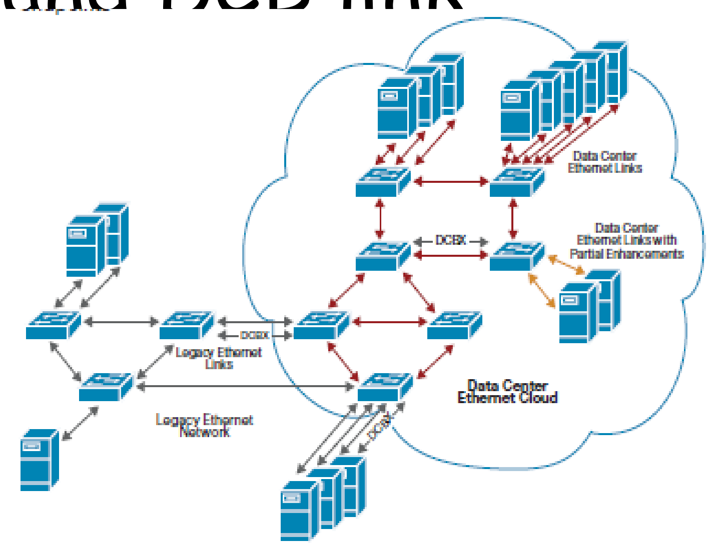
Congestion Notification

- Source quench messages are sent by the congeste switch for source rate limiting
- Implemented in switches and not in host, not effective
- IEEE 802.1Qau



Data Center Bridging eXchange (DCBX)

- Allows autoexchange of Ethernet Parameters and Discovery Functions between switches and endpoints such as DCB peer dicovery, mismatched configuration detection and DCB link configuration of peers.
- Uses the LLDP protocol
- IEEE 802.11Qaz



Review questions

- What are the requirement diferences between residential and data center ethernet?
- What enables the rapid spanning tree protocol to react faster to changes than the spanning tree protocol?
- What can be achieve by the use of shortest path bridging?
- Why ISO IS-IS became popular in data center?
- How can Data Center Bridging facilitate multi-tenancy in data centers?
- How can Data Center Bridging eXchange facilitate federated data centers?

Data Center Traffic

Data Center Traffic

17%

Data center to users traffic

7%

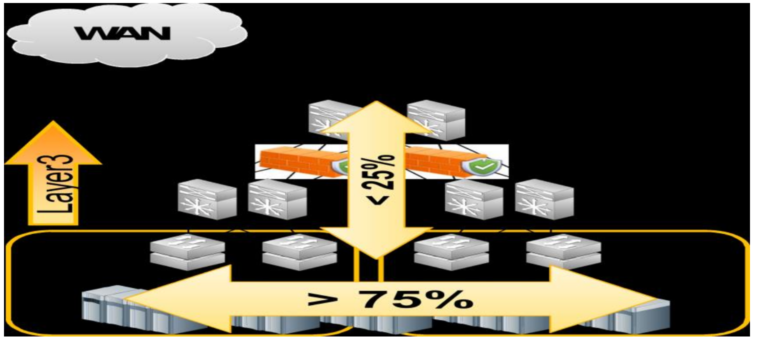
Between data centers

76%

Traffic within data center

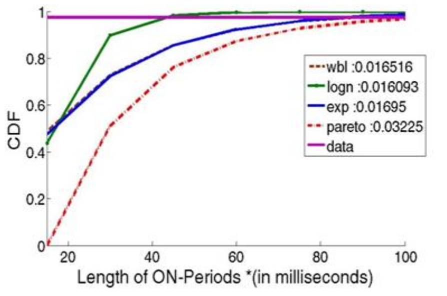
Data Center Traffic

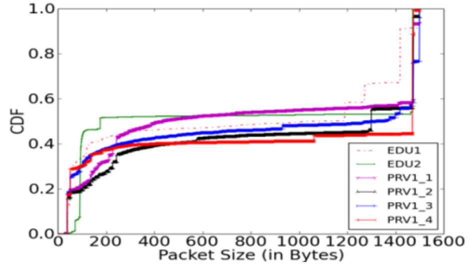
- North-South: extra-cloud communication (to/from the Internet)
- East-West: intra-cloud communication (inter-VMs)
- Depends on the kind of data center/mix of applications
- North-South traffic is increasing, but the East-West portion of overall traffic is getting much larger
- Inter-data center (D2D) traffic is a growing concern



Data Center Traffic

- Most of the flows are small in size (< 10 KB)
- Most of the bytes in top 10% large flows
- Traffic leaving edge switches ON-OFF, lognormal distributions
- Packet size distribution bimodal (200 to 1400 B)

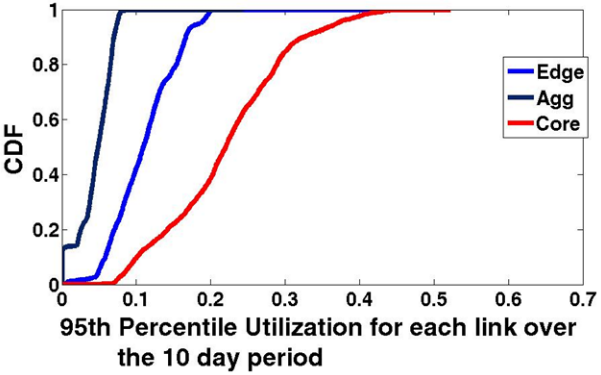


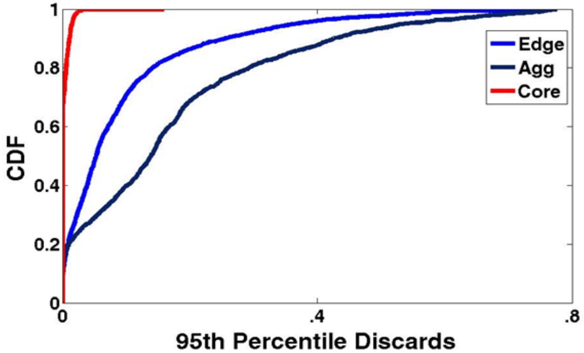


T. Benson, A. Akella, and D. A. Maltz. 2010. Network traffic characteristics of data centers in the wild. In Proc of the 10th ACM SIGCOMM conference on Internet measurement (IMC '10). ACM, New York, NY, USA, 267-280

Data Center Traffic

- In cloud data center majority of flows stay in rack (80%) while in enterprise and university data center it varies from 40% to 90%
- Core layer most utilized, edge layer lightly utilized
- Core layer contain hot spot but less than 25% of links
- No need for more bisection bandwidth
- Most of losses occur in links with low utilization due to bursty traffic





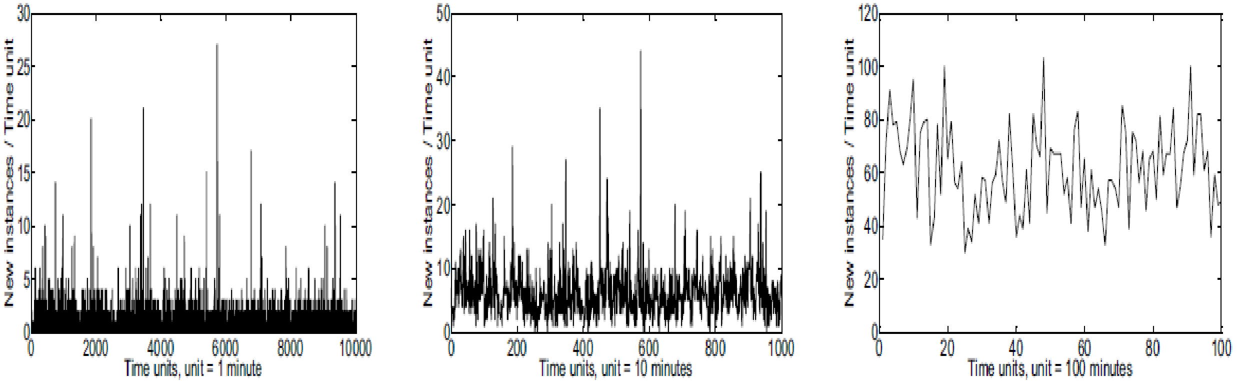
T. Benson, A. Akella, and D. A. Maltz. 2010. Network traffic characteristics of data centers in the wild. In Proc of the 10th ACM SIGCOMM conference on Internet measurement (IMC '10). ACM, New York, NY, USA, 267-280

Data Center Traffic

- Bimodal and Skewed
- Elephant flows: long-lived, bandwidth hungry, and scarce/bursty,
  - less than 1% of all flows,
  - generate more than half of the data volume
- Mice flows: very small and short-lived
- Mixing types of traffic may cause adverse effects
- Elephants create hot-spots, dropping several mice packets

VM Processes

- VM arrival and departure processes - self similar, power law
- VM in the system: ARIMA model



Yi Han, Jeffrey Chan and Christopher Leckie. Analysing Virtual Machine Usage in Cloud Computing. In Proceedings of the IEEE 2013 3rd International Workshop on Performance Aspects of Cloud and Service Virtualization, 2013