University of BRISTOL

Lecture 2

Architectures of Data Center Networks

Outline

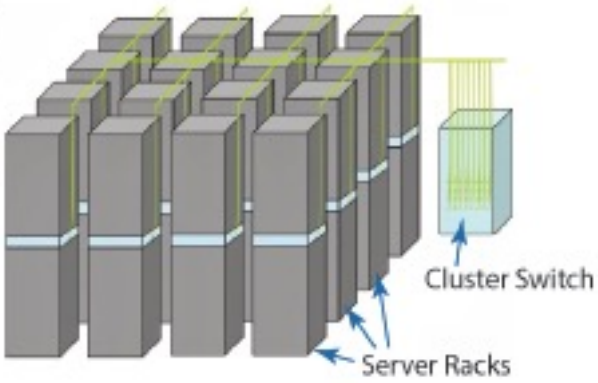
University of BRISTOL

- Requirements of data center design
- Current DCN architecture
- Trend of DCN design

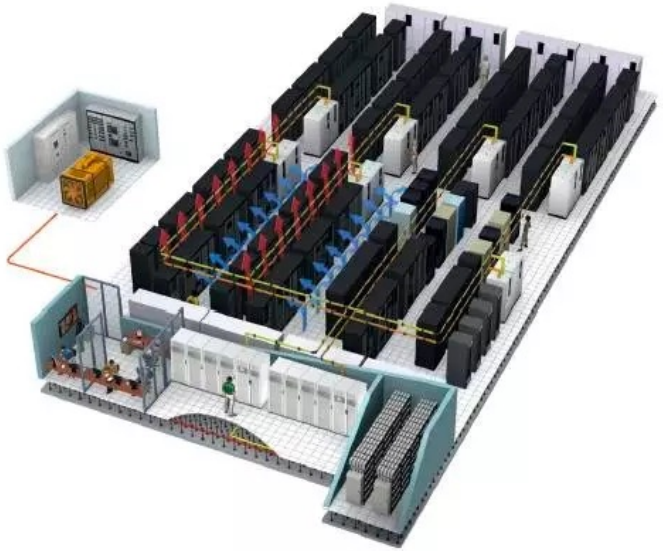
Building blocks for DCN

Server

Rack



Cluster



Data Center

Basics of DCN

University of BRISTOL

- A rack has ~20-40 servers

Front of a rack

Rear of a rack

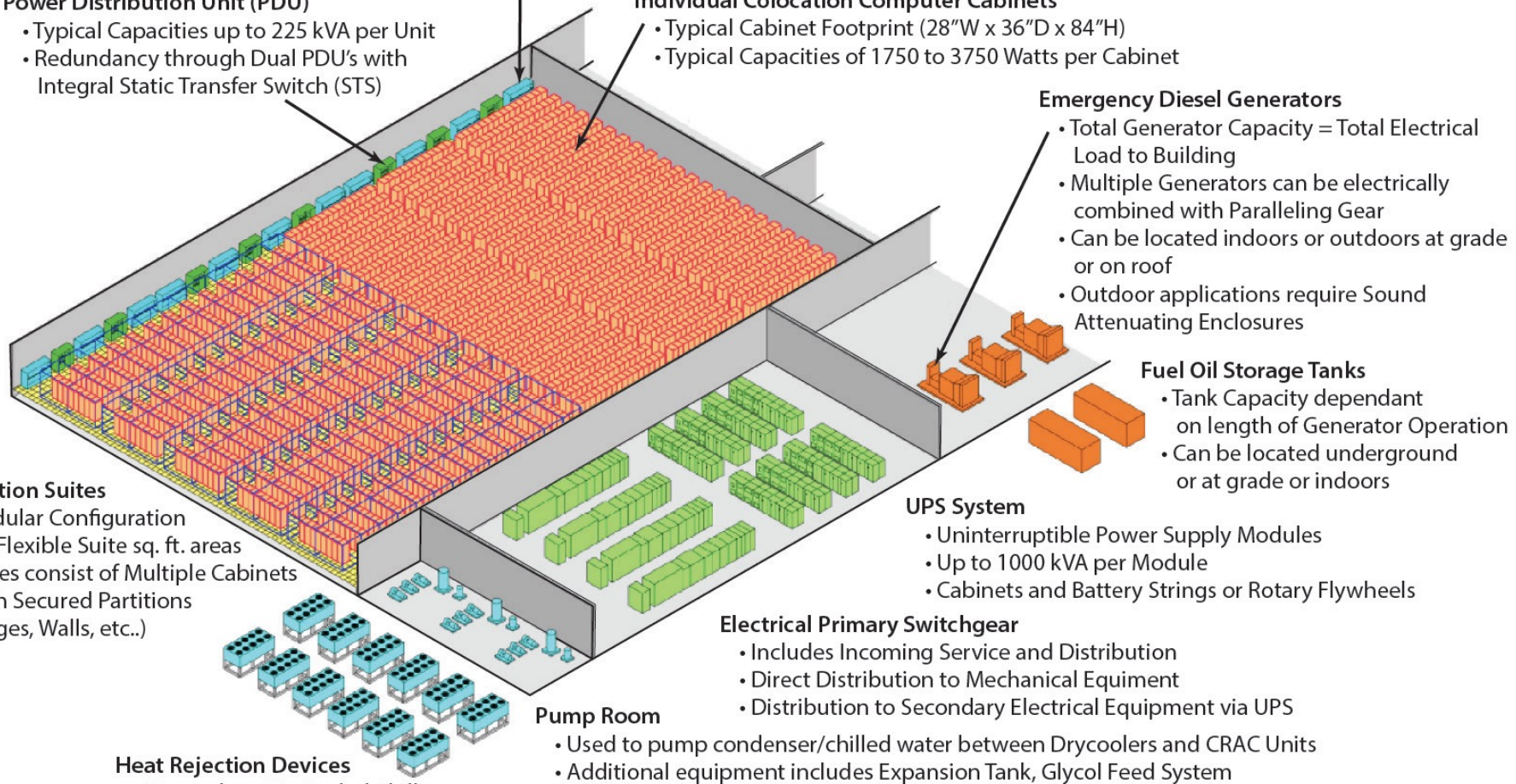
Transceivers

Cabling

- Example of A TOR switch with 48 ports

"Top of Rack" switch

Main components of a typical DCN



Size of one google data center

- Location: Douglas county
- Estimated server count: 417600

- https://www.google.com/about/datacenters/inside/locations/douglas-county/index.html

Data centre network requirements

University of BRISTOL

- Uniform high capacity
  - Capacity between servers limited only by their Network Interface Cards (NICs)
  - High capacity between any two servers
- Performance isolation
  - Traffic of one service should be unaffected by others
- Flexibility and scalability
  - Agility – Any service, Any Server
  - Turn the servers into a single large fungible pool
    - dynamically expand and contract their footprint as needed
    - Unlimited workload mobility
- Resiliency and security
  - Fault tolerant

Overall DCN design goals

University of BRISTOL

- Cost-effective
- Ease of management: “Plug-&-Play” (layer-2 semantics)
  - Flat addressing, so any server can have any IP address
  - Server configuration is the same as in a LAN
  - Legacy applications depending on broadcast must work

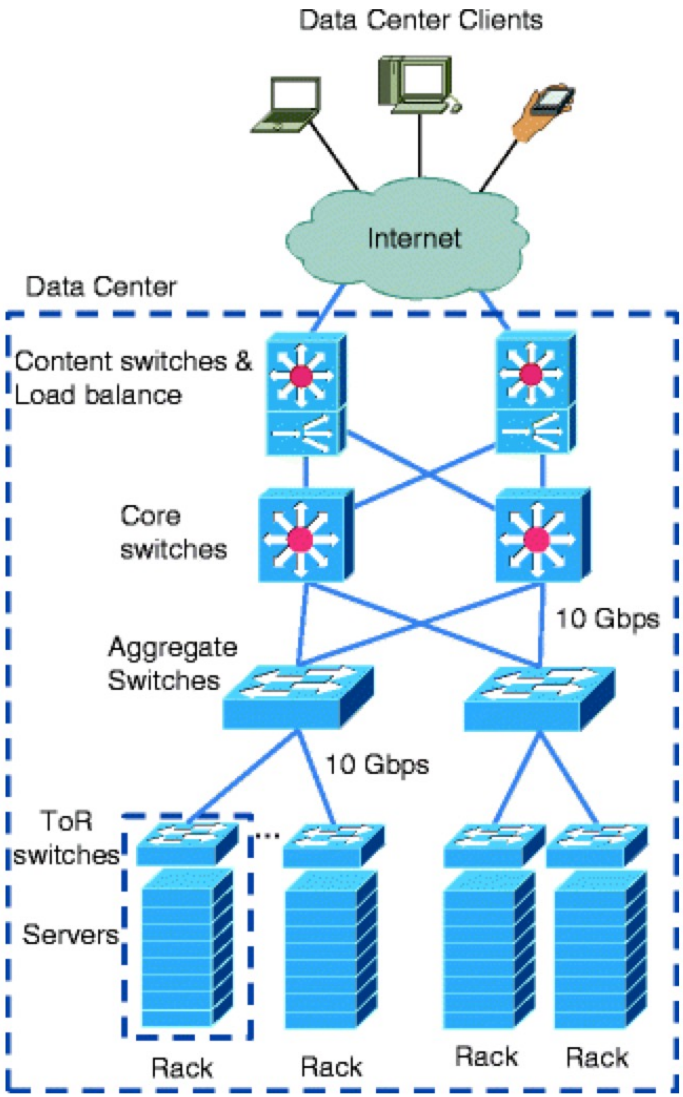
Applications in Data Centers

University of BRISTOL

- Data centers typically run two types of applications
  - Outward facing (e.g., serving web pages to users)
  - Internal computations (e.g., MapReduce for web indexing)
- Workloads often unpredictable:
  - Multiple services run concurrently within a DC
  - Demand for new services may spike unexpected
- Failures of servers are the norm

Typical architecture of current DCN

- **Servers** (usually up to 48 in the form of blades) are accommodated into racks and connected to a **Top-of-the-Rack (ToR)** switch using either 1Gbps or more recently 10Gbps links.
- ToR switches connect to one or more end of row (EoR) switches or ***aggregate switches*** using 10 Gbps links in a tree topology.
- ***Core switches*** *for intercnnection of aggregate switches*

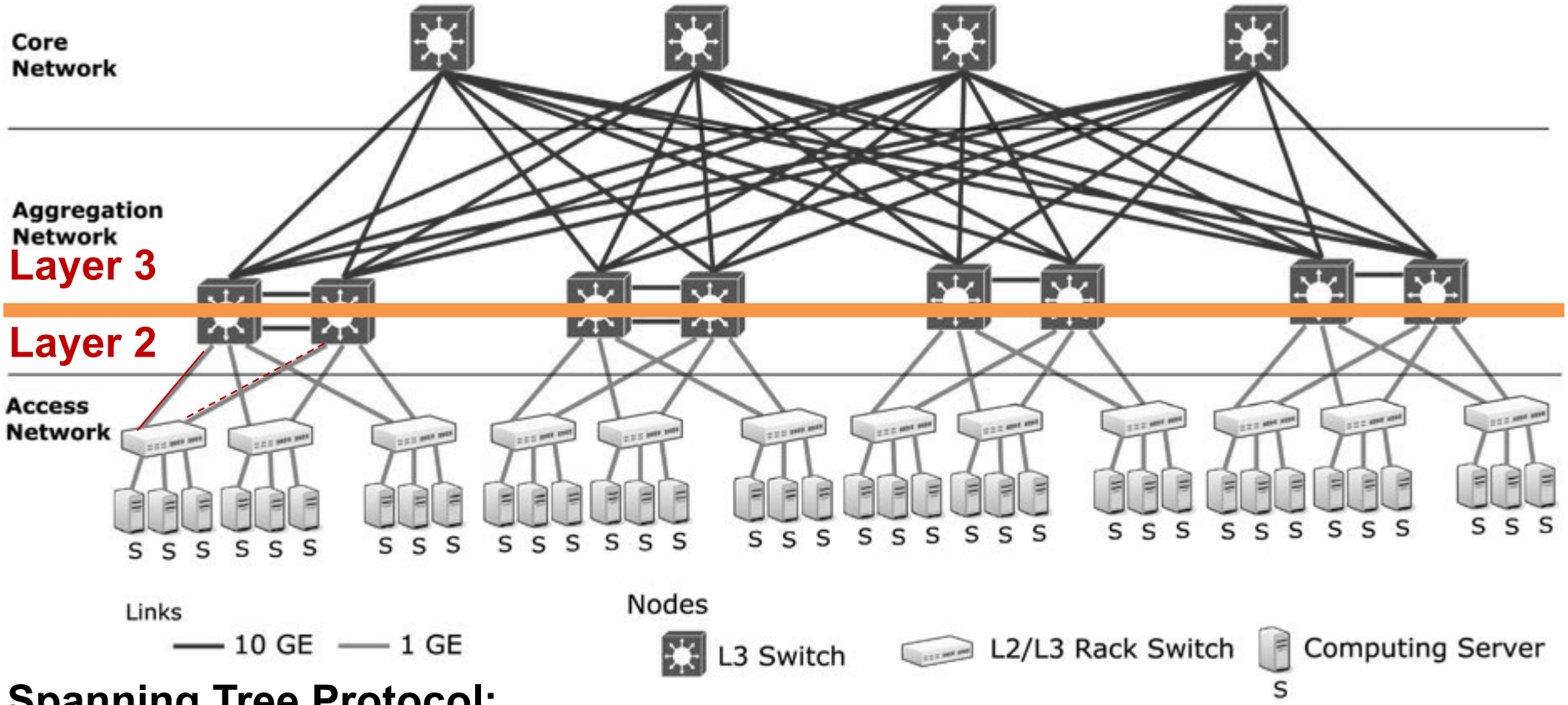


Forwarding in typical data center topology

University of BRISTOL

- Layer 3 approach:
  - Assign IP addresses to hosts hierarchically based on their directly connected switch.
  - Use standard intra-domain routing protocols, eg. OSPF.
  - Large administration overhead
- Layer 2 approach:
  - Forwarding on flat MAC addresses
  - Less administrative overhead
  - Bad scalability
  - Low performance
- Middle ground between layer 2 and layer 3:
  - VLAN: a technology which defines broadcast domains in a layer 2 network
  - Feasible for smaller scale topologies
  - Resource partition problem

Three-tier Data Center architecture

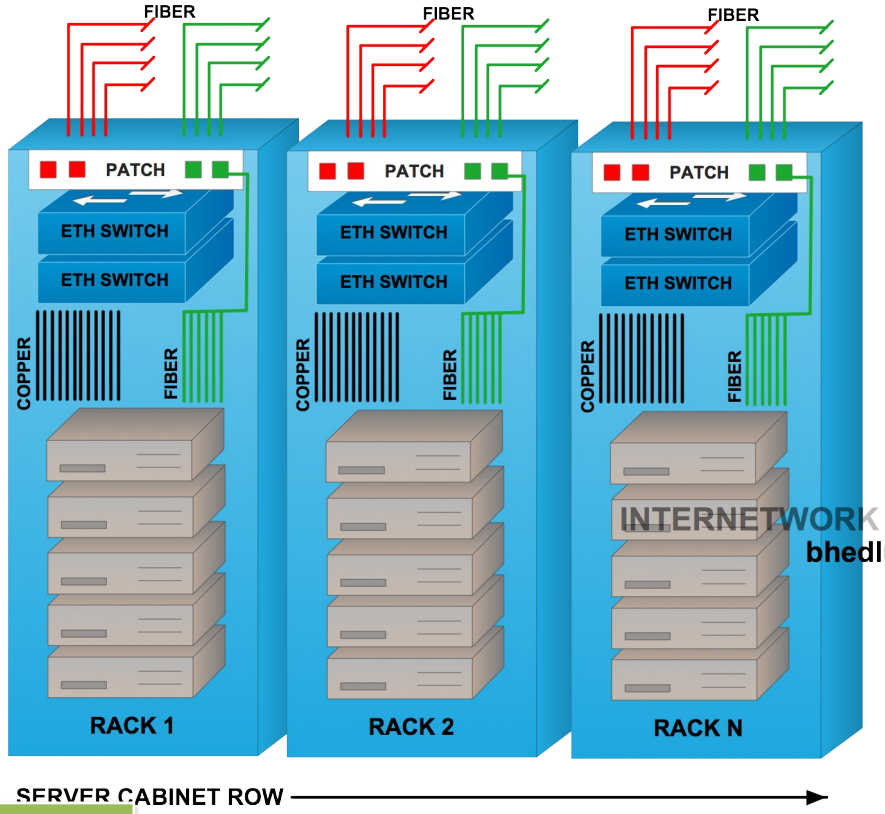


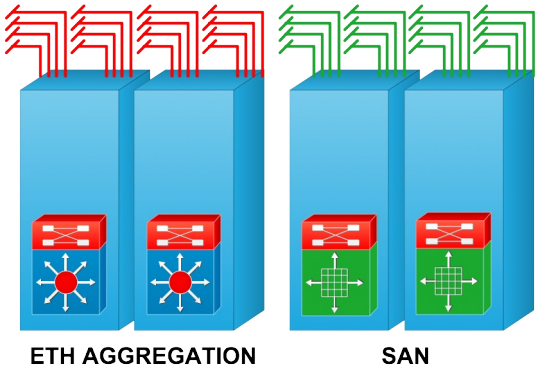
Spanning Tree Protocol:

- Simple ☺
- plug-and-play technology requiring little configuration ☺
- Loop-free ☺
- Can not use parallel forwarding ☹

https://en.wikipedia.org/wiki/Spanning_Tree_Protocol

Top of Rack (ToR) VS. End of Rack (EoR) switch





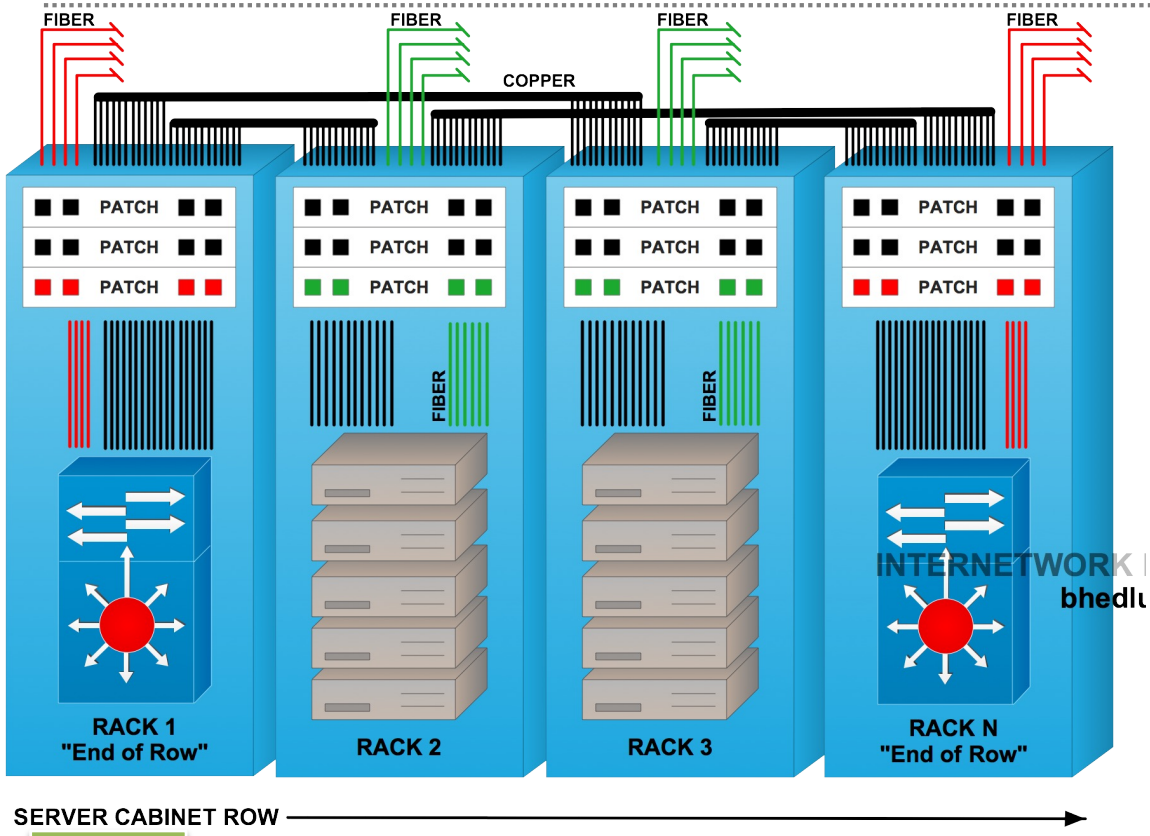
Pro's:

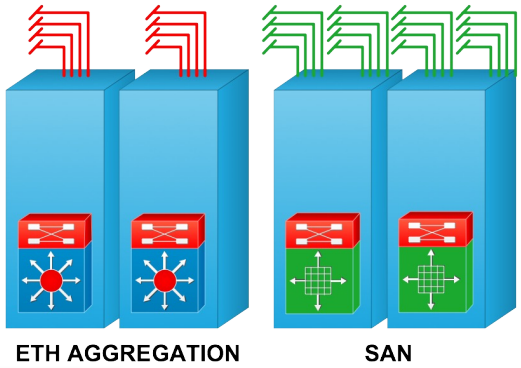
- Copper stays "In Rack"
- Short copper cabling to servers allows for low power, low cost 1oGE, 40G in the future
- Ready for Unified Fabric today

Con's:

- More switches to manage. More ports required in the aggregation.
- More Layer 2 server-to-server traffic in the aggregation.
- Racks connected at Layer 2.
- Unique control plane per 48-ports (per switch), higher skill set needed for switch replacement.

EoR Switch





Pro's:

- Fewer switches to manage.
- Fewer ports required in the aggregation.
- Longer life, high availability, modular platform for server access.
- Unique control plane per hundreds of ports (per modular switch)

Con's:

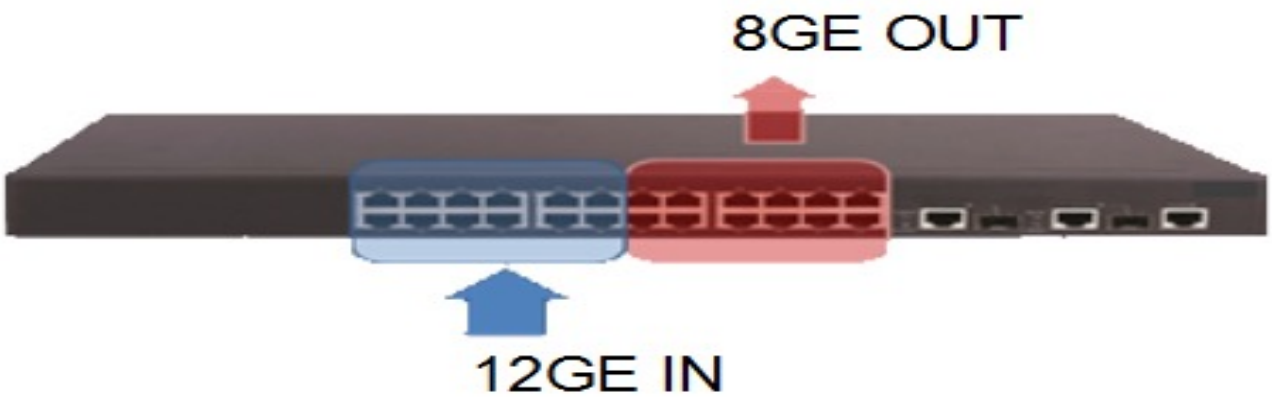
- Requires an expensive, bulky, rigid, copper cabling infrastructure.
- More infrastructure required for patching and cable management.
- Long twisted pair copper cabling limits the adoption of lower power higher speed server I/O.
- More future challenged than future proof.
- Less flexible "per row" architecture. Platform upgrades/changes affect entire row.

ToR vs. EoR

University of BRISTOL

Physical Architecture	TOR Architecture	EOR/MOR Architecture
Advantages	• Simple cabling, convenient cable maintenance, and high scalability • Cabinet-based modular management and small-scale fault impact	• Simple management and high reliability • High interface use
Disadvantages	• Port waste • Complex management and maintenance of TOR switches	• Complex cabling and difficult maintenance • Large-scale fault impact

Oversubscription



Network oversubscription refers to a point of bandwidth consolidation where the ingress bandwidth is greater than the egress bandwidth.

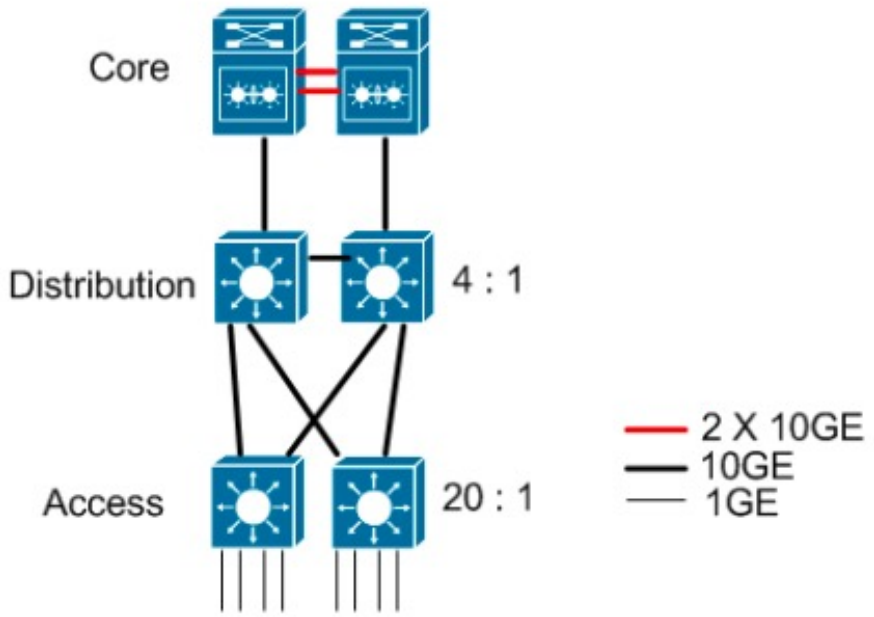
- High-bandwidth-port switches are expensive
  - cost increasing non-linear with number of ports and bandwidth
- Traffic aggregation of multiple input traffics
- Not all bandwidth are fully utilized.

Oversubscription

Oversubscription ratio of a switch:

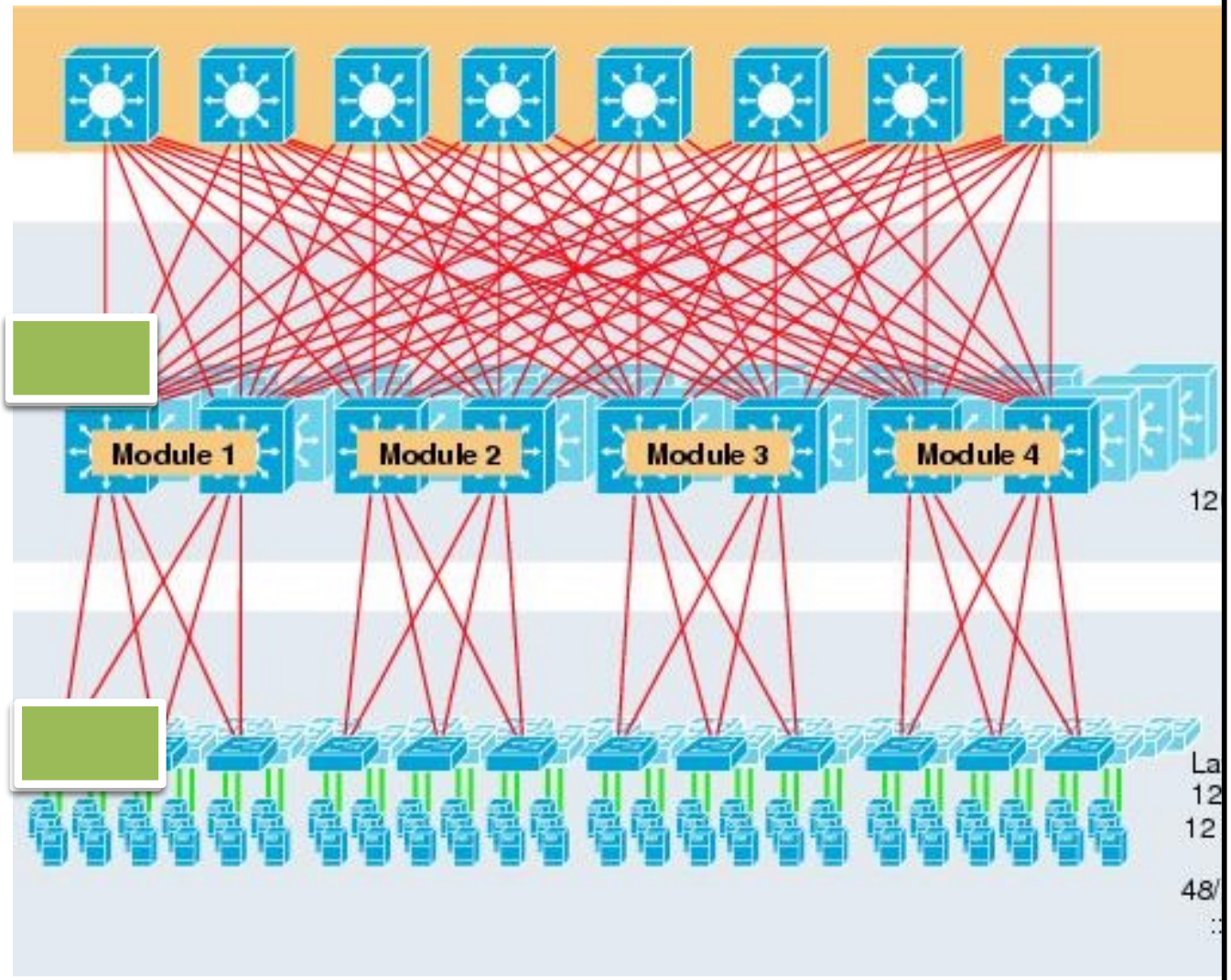
Ratio of ports facing downwards vs. ports facing upwards

- Reduce port numbers or port bandwidth for aggregate/core switches
- Improve hardware utilization
- Lower the total costs



- Poor bisection bandwidth
- Switch up-links get heavily loaded

Calculating Oversubscription



8 Core Nodes
32- 10GE ports

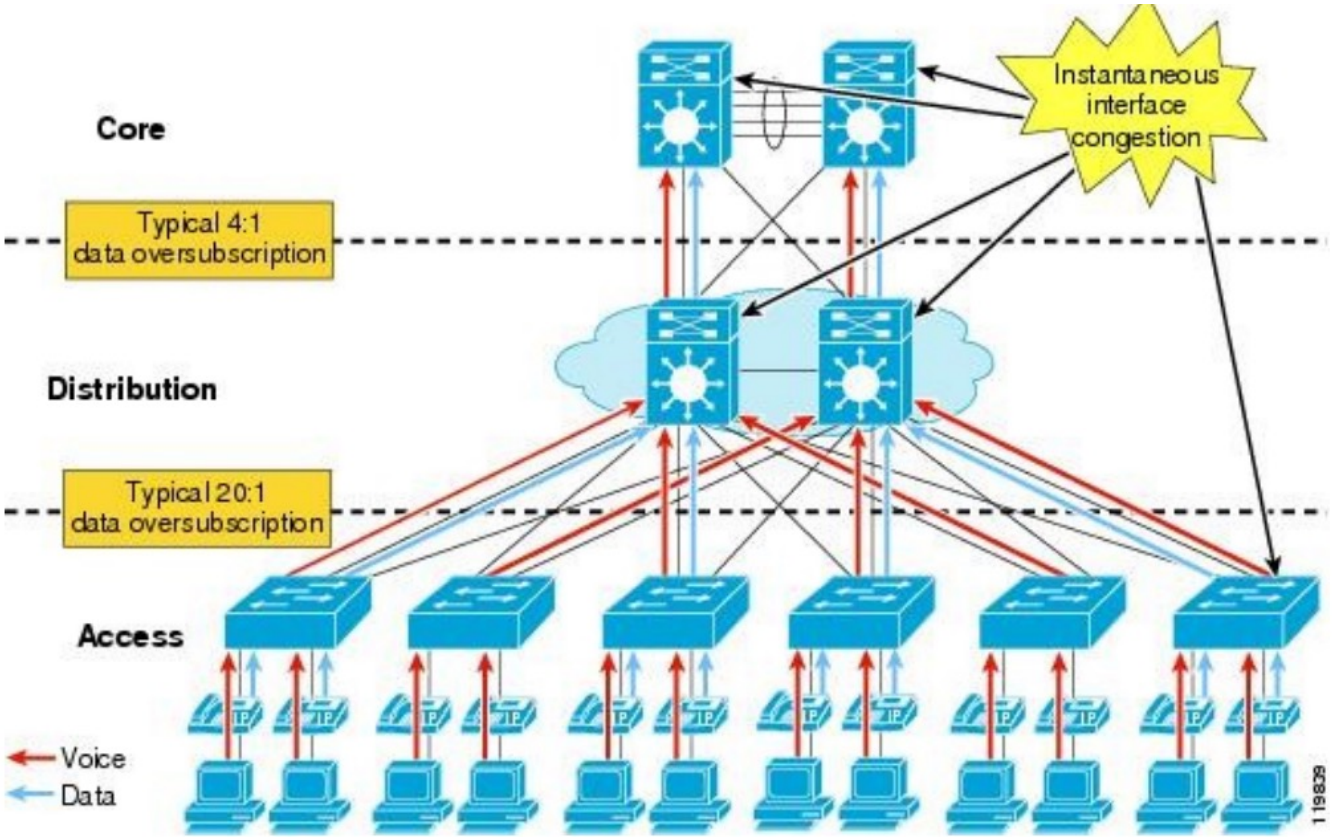
16 Aggregation Modules
32 Aggregation switches
8×10GE uplinks
12×10GE to access layer

12 TOR switches per Aggregation Modules
48 Server per ToR
48 ×1GE tp server ports
2×10GE to aggregation modules

The combined oversubscription ratio: 1.5×2.4 = 3.6:1
The bandwidth per server is: ~277 Mbps

Oversubscription

- Typical over-subscription ratios
  - ToR uplinks oversubscription 4:1 to 20:1
  - EoR uplinks oversubscription 4:1



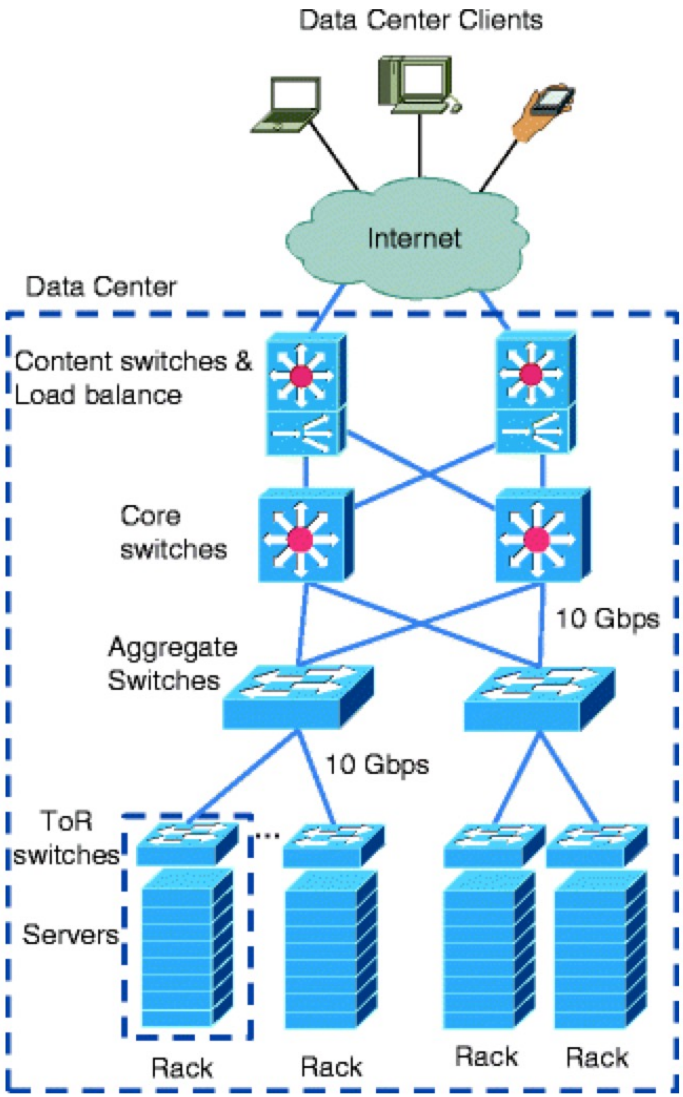
Example Configuration

University of BRISTOL

- Data center with 11'520 machines
- Machines organized in racks and rows
  - Data center with 24 rows
  - Each row with 12 racks
  - Each rack with 40 blades
- Machines in a rack interconnected with a ToR switch (access layer)
  - ToR Switch with 48 GbE ports and 4 10GbE uplinks
- ToR switches connect to End-of-Row (EoR) switches via 1-4 10GigE uplinks (aggregation layer)
  - For fault-tolerance ToR might be connected to EoR switches of different rows
- EoR switches typically 10GbE
  - To support 12 ToR switches EoR would have to have 96 ports (4*12*2)
- Core Switch layer
  - 12 10GigE switches with 96 ports each (24*48 ports)

Advantages of 3-Tier Data center architecture

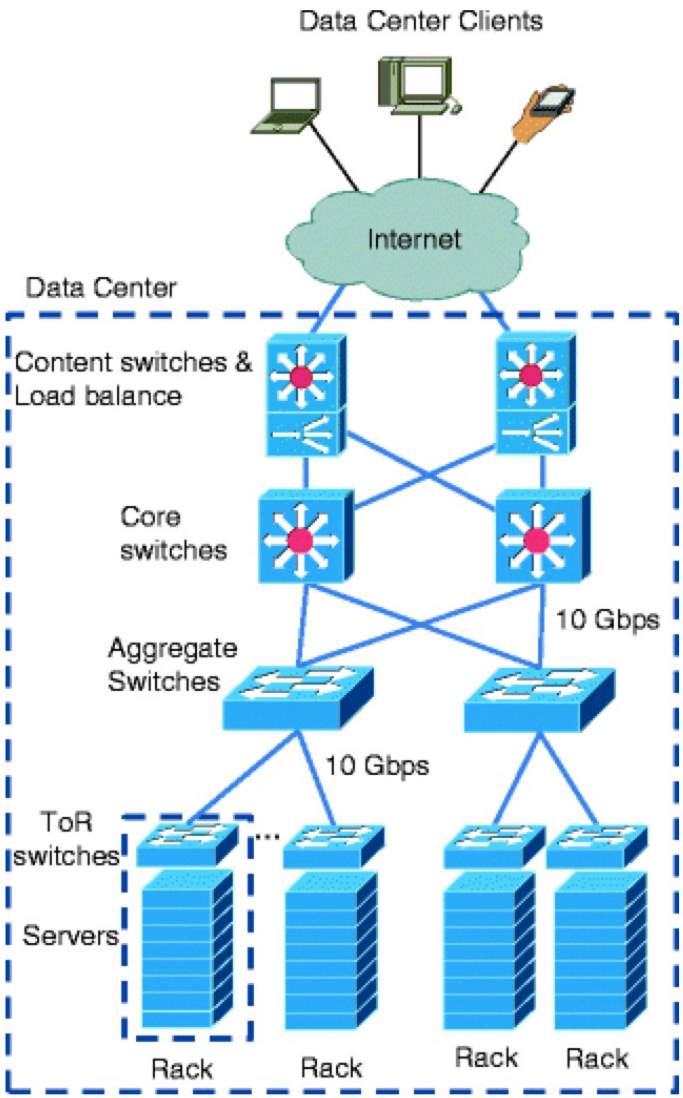
- Easily scalable
- Fault tolerant (e.g. a ToR switch is usually connected to 2 or more aggregate switches).



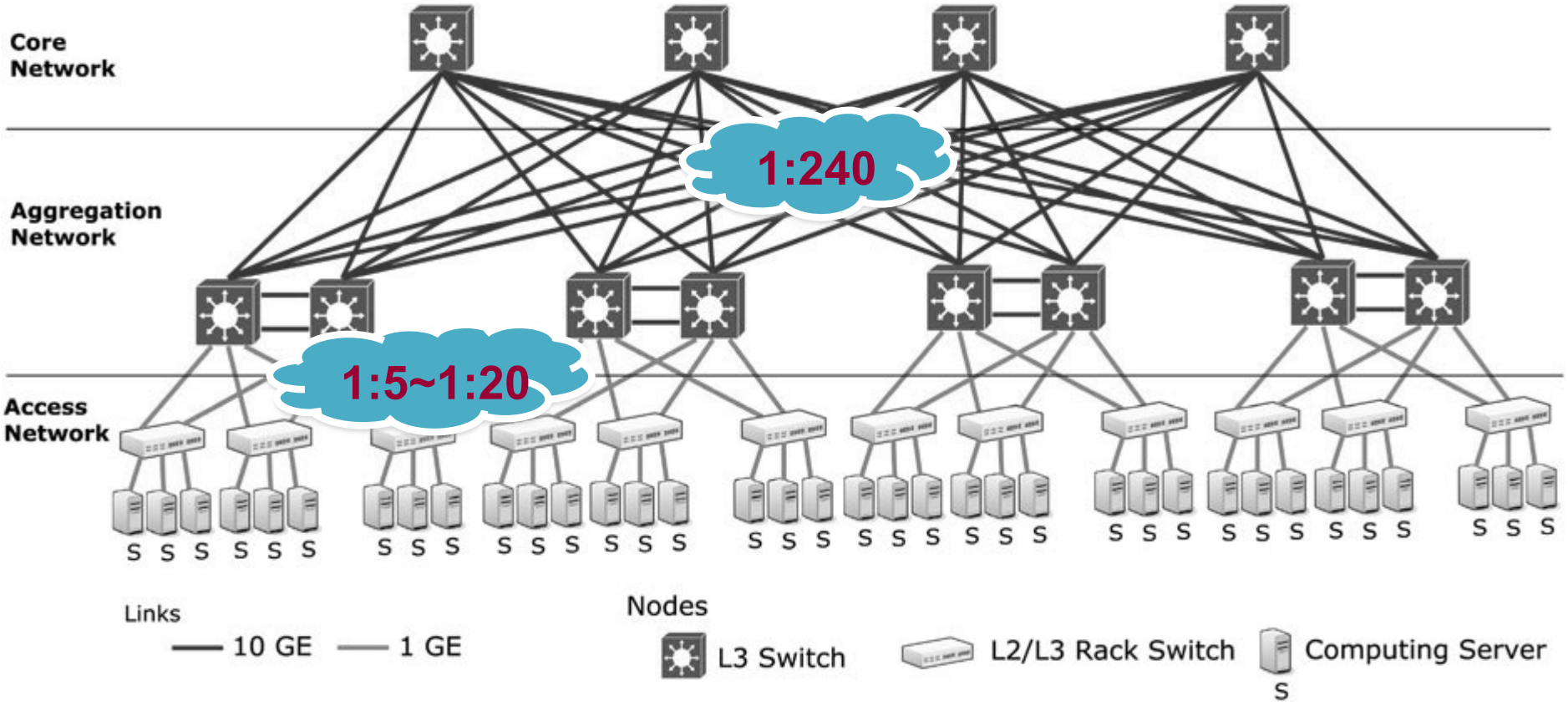
Disadvantages of 3-Tier Data center architectur

University of BRISTOL

- High power consumption of the ToR, aggregate switch and core switches
  - Mainly caused by the power consumed by the Optical-to-Electrical (O-E) and E-O transceivers and the electronic switch fabrics (crossbar switches, SRAM-based buffers, etc.)
- High number of links
- High latency introduced due to multiple store-and-forward processing.
  - Significant queuing and processing delays as a packet travels from one Server to another through the ToR, the aggregate and the core switch.
- Oversubscription
  - Network bottleneck at aggregate and core swiches

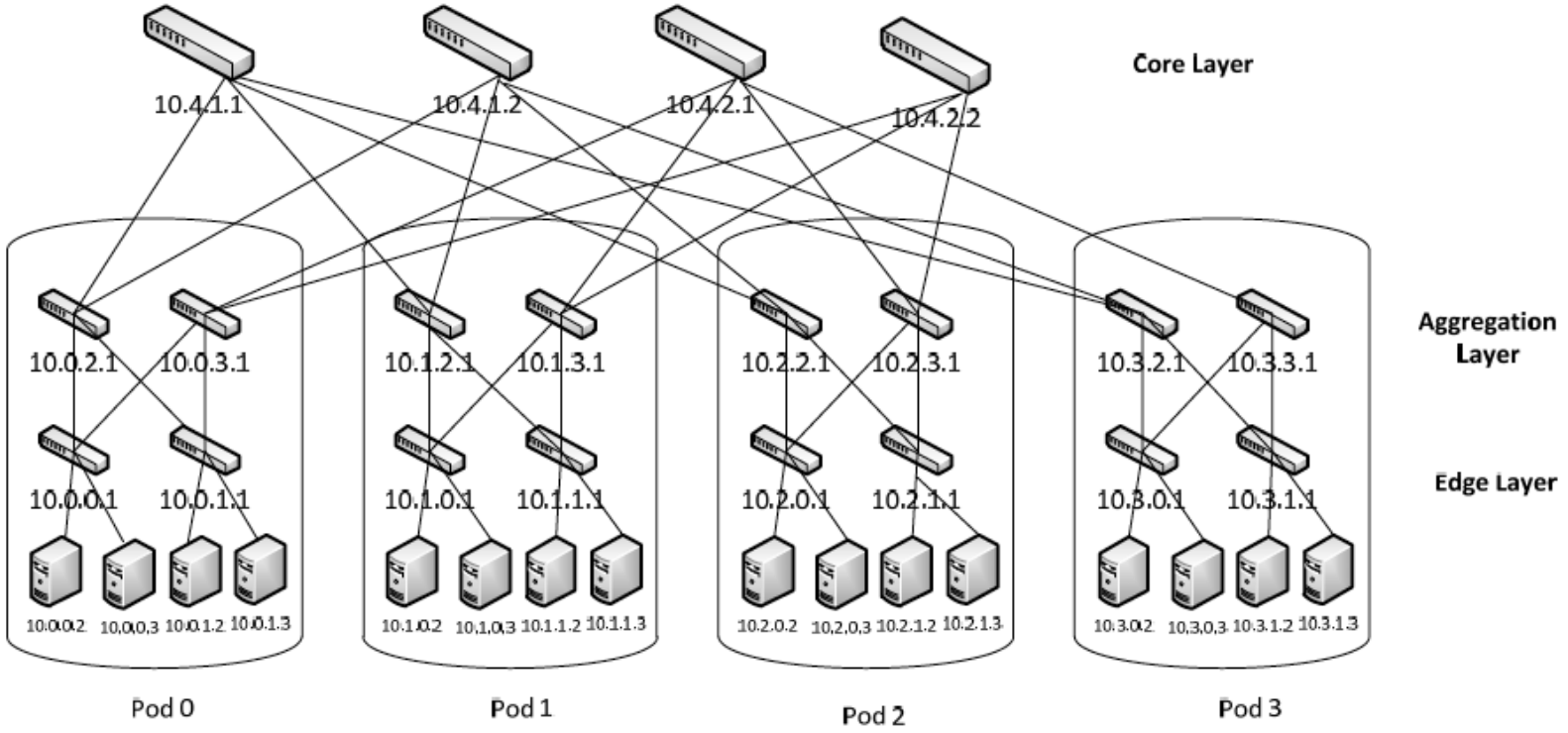


Three-tier Data Center architecture



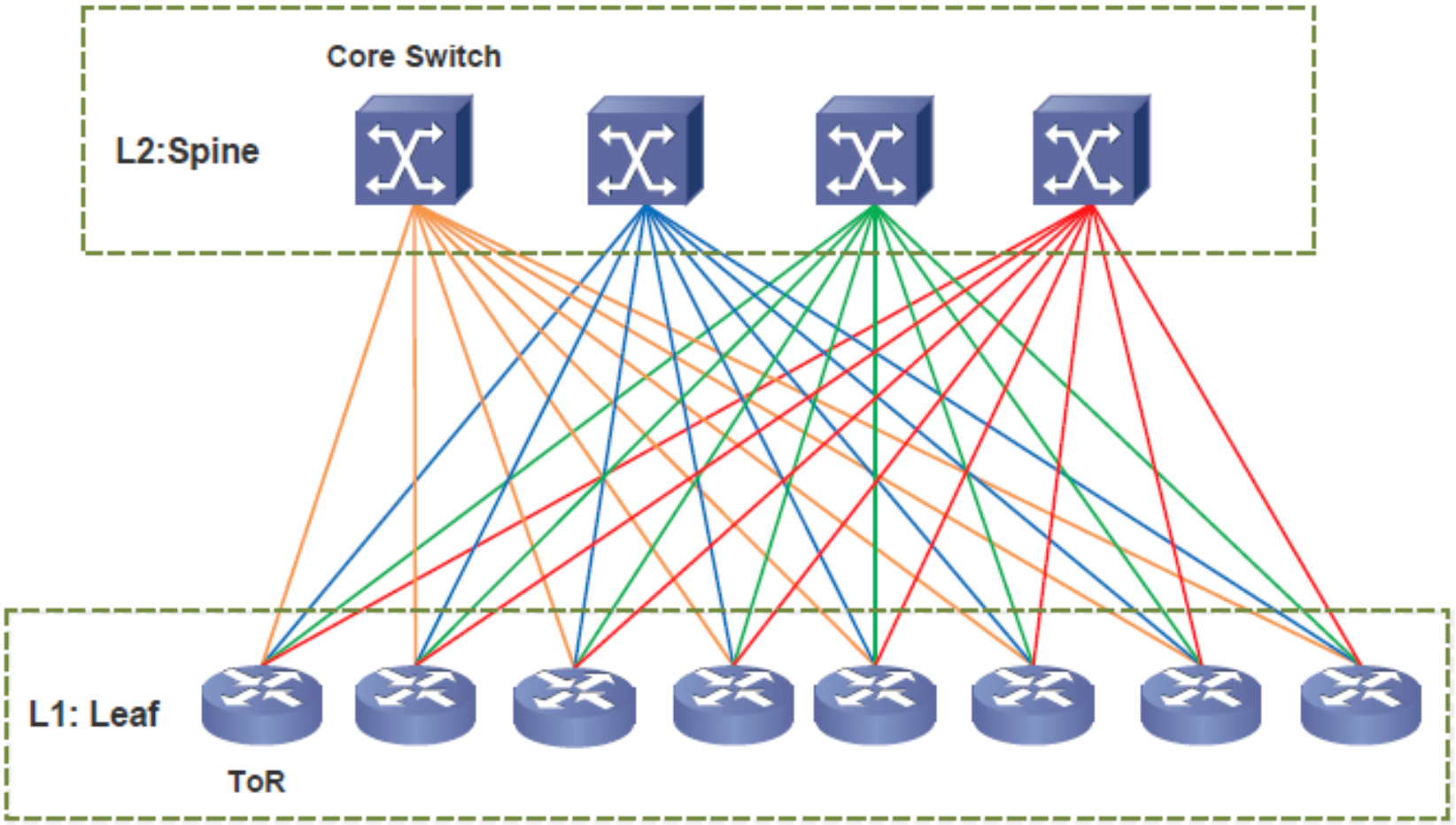
Serious communication bottleneck caused by over-subscription in access and aggregation and core networks

Fat-tree DCN architecture



- Over-subscription ratio: 1:1
- Customized routing algorithms

Leaf-spine network

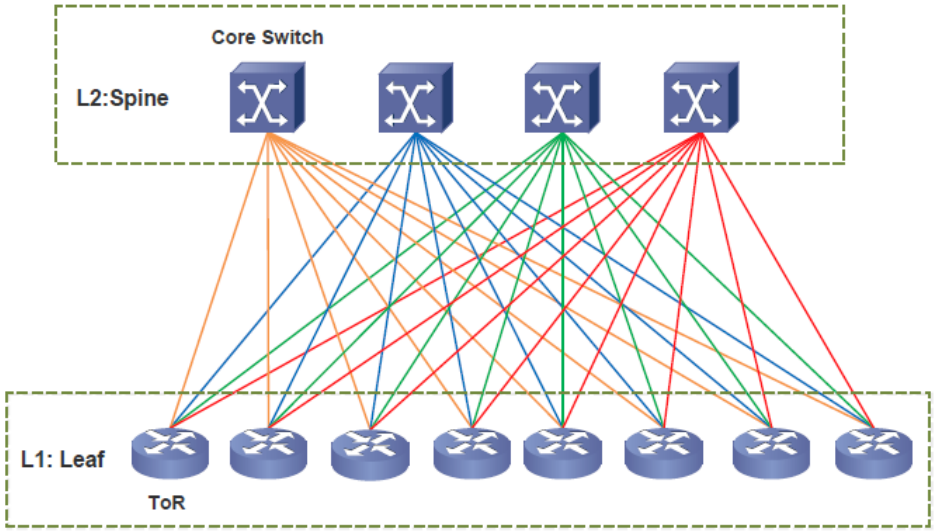


Leaf-spine network

University of BRISTOL

Advantages ☺

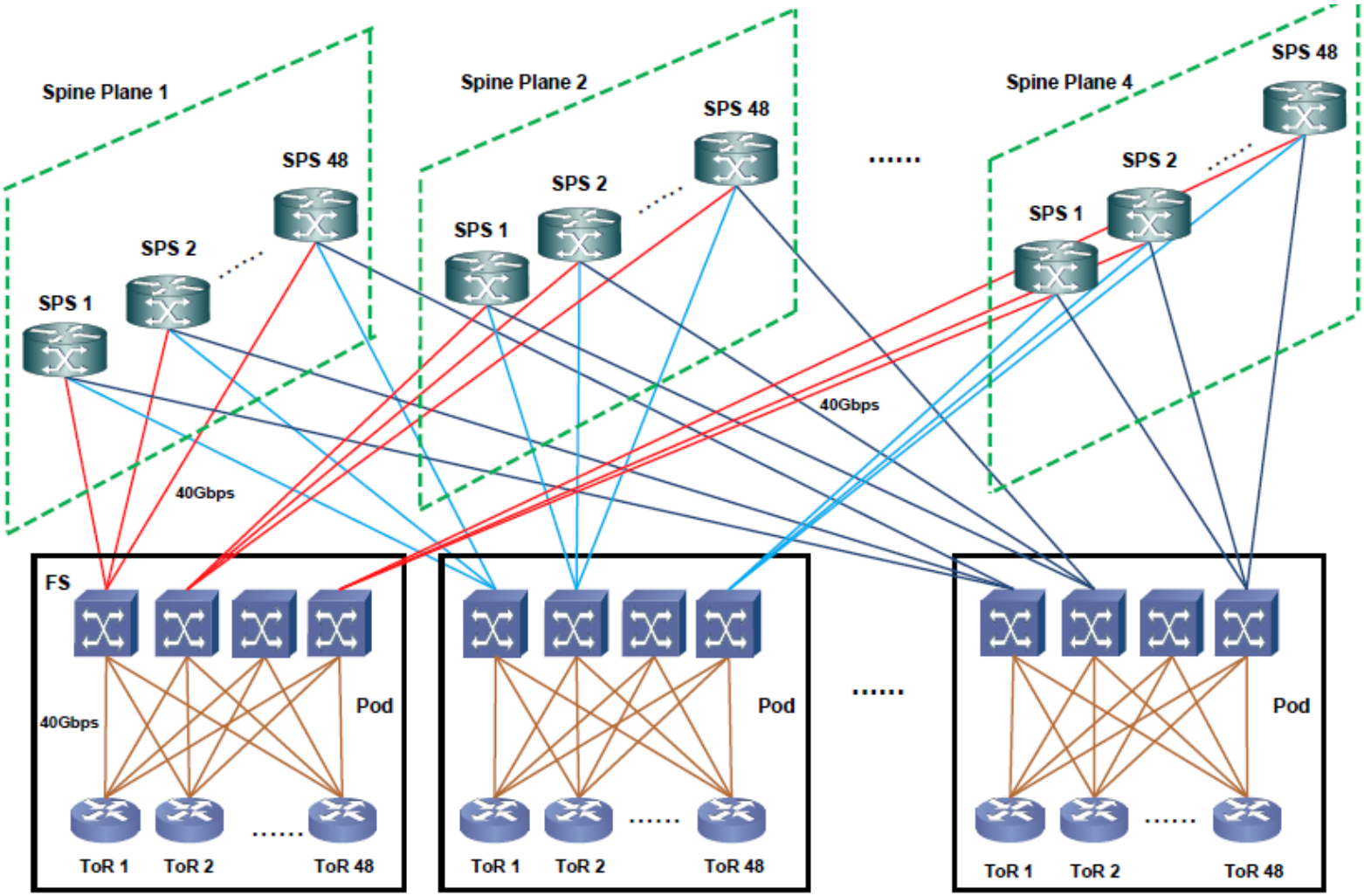
- Simplified network cabling
- Clos-based network with same latency
- Easily scale up for both spine and leaf switches
- Traffic crosses the same number of devices with predictable latency



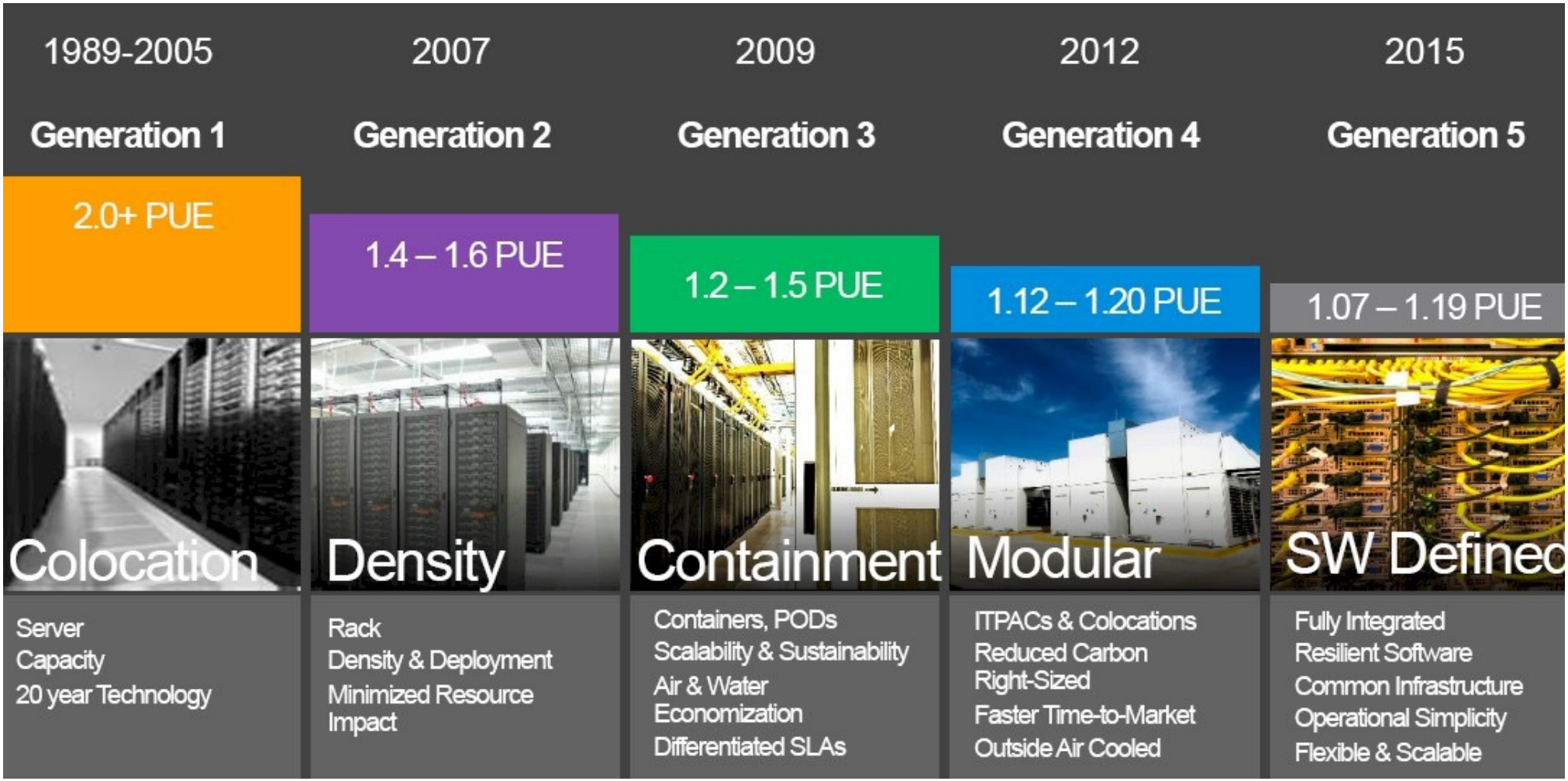
Disadvantages ☹

- Horizontal scaling determined by the design of the high-port count spine switches

Next-generation Facebook DCN (Leaf and Spine



Data Center evaluation (Microsoft)

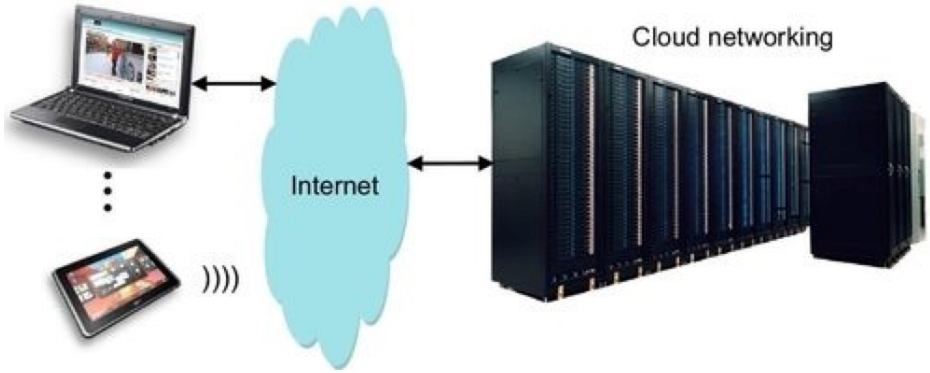


Power usage effectiveness (PUE) = (Facility power) / (IT Equipment power)

Modular Data Centers

- Microsoft is using modular Performance Optimized Data Center (PODs) modules as basic building blocks.

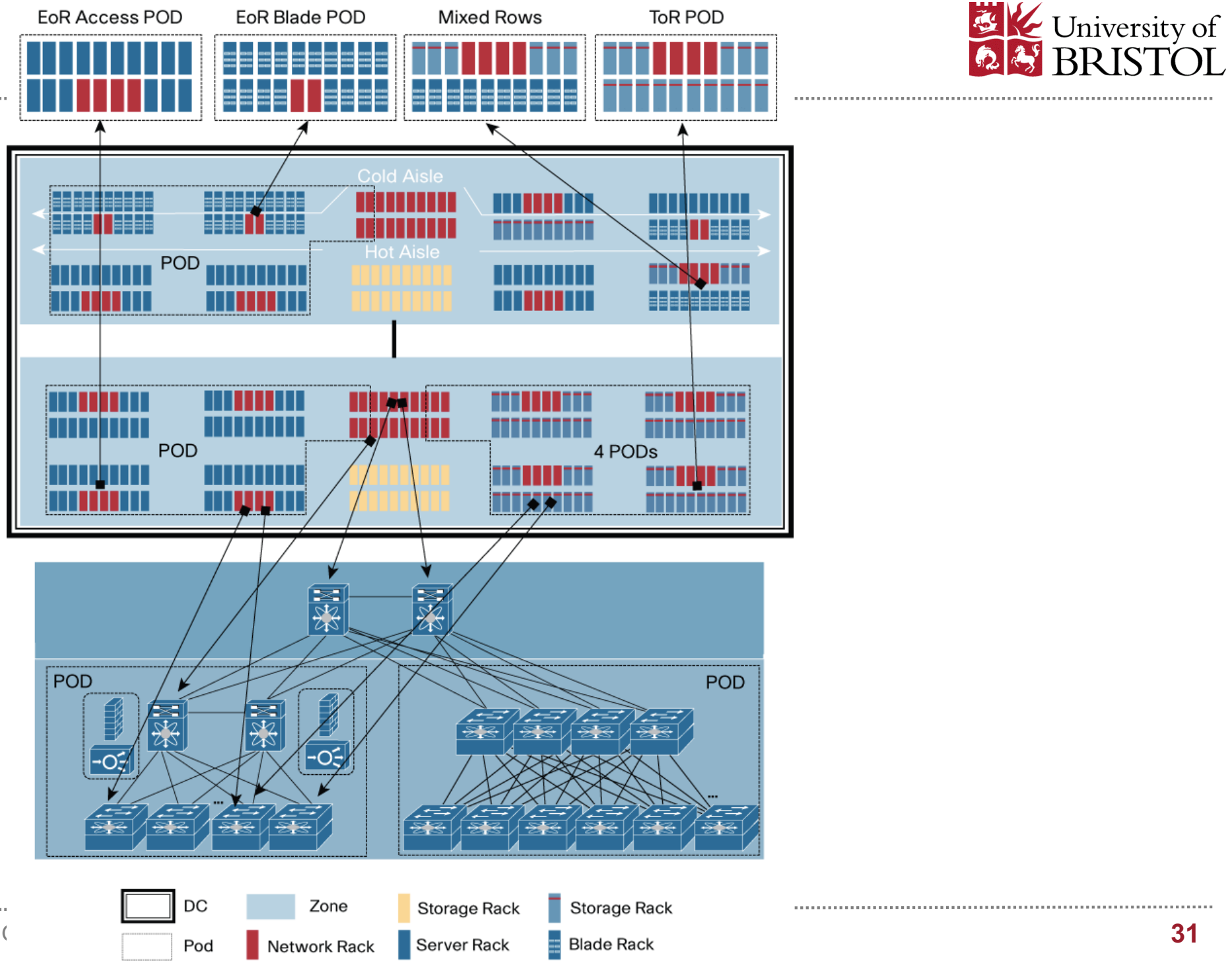
- A shipping container includes servers, storage, networking, power, and cooling.
- Simply stack the containers, connect external networking, power and cooling and the DC is ready to run.



Data center PODs

University of BRISTOL

- Advantages
  - Repaired on site in case of failure.
  - If a POD fails, a container truck moves it out and a new one comes in.
- Disadvantages
  - Commonly built using equipment from single vendor=> doesn't allow the administrator to optimize performance.
  - Single rack failure requires replacement of multiple racks that make up the POD, taking down several operational racks.


EoR Access POD
EoR Blade POD
Mixed Rows
ToR POD
Cold Aisle
Hot Aisle
POD
POD
4 PODs
POD
POD
DC
Pod
Zone
Network Rack
Storage Rack
Server Rack
Storage Rack
Blade Rack

Server-centric DCN architectures

- Bcube
- Dcell

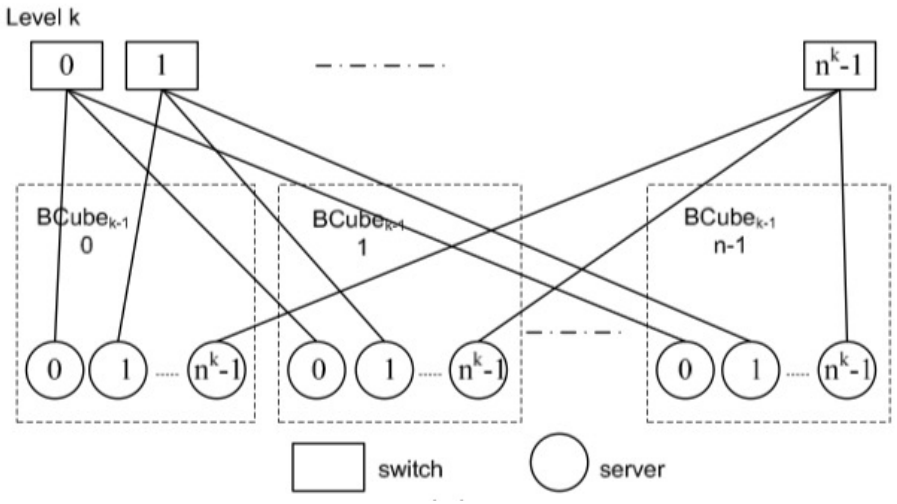
BCube

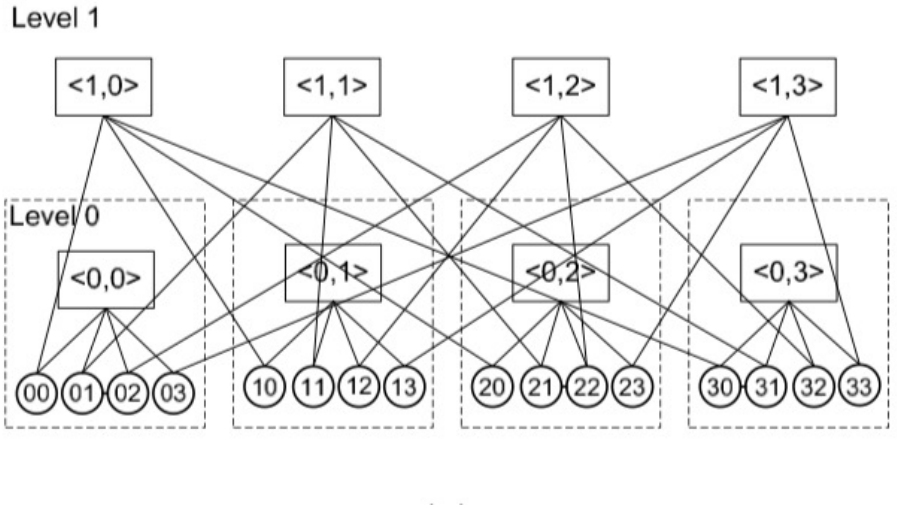
University of BRISTOL

- Main Goal: network architecture for shipping-container based modular data centers
- Designed for shipping-container modular DC
- BCube construction: level structure
  - $BCube_k$ recursively constructed from $Bcube_{k-1}$
- server-centric:
  - servers perform routing and forwarding
- Consider a variety of communication patterns
  - one-to-one, one-to-many, one-to-all, all-to-all
  - single path and multi-path routing

Bcube Topology Construction

- Recursive structure: $BCube_k$ is recursively constructed from *n* $BCube_{k-1}$ and *nk* *n*-port switches

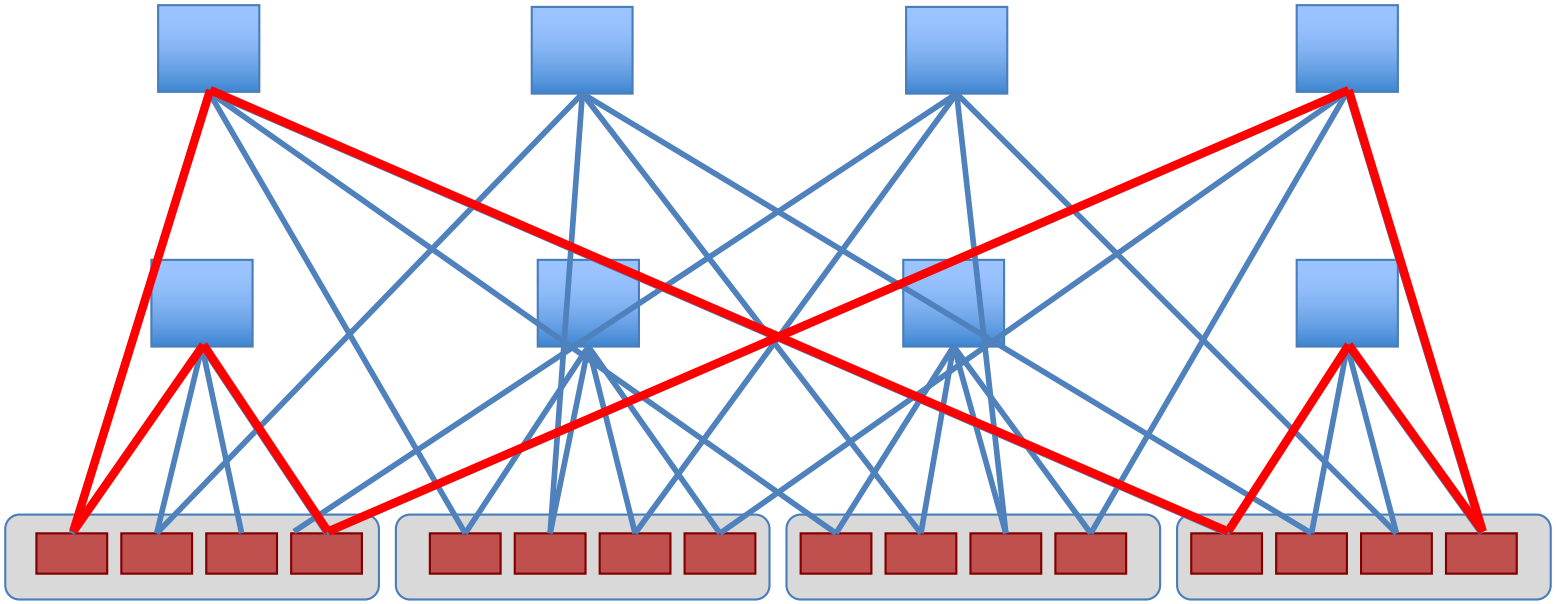




Bcube Topology

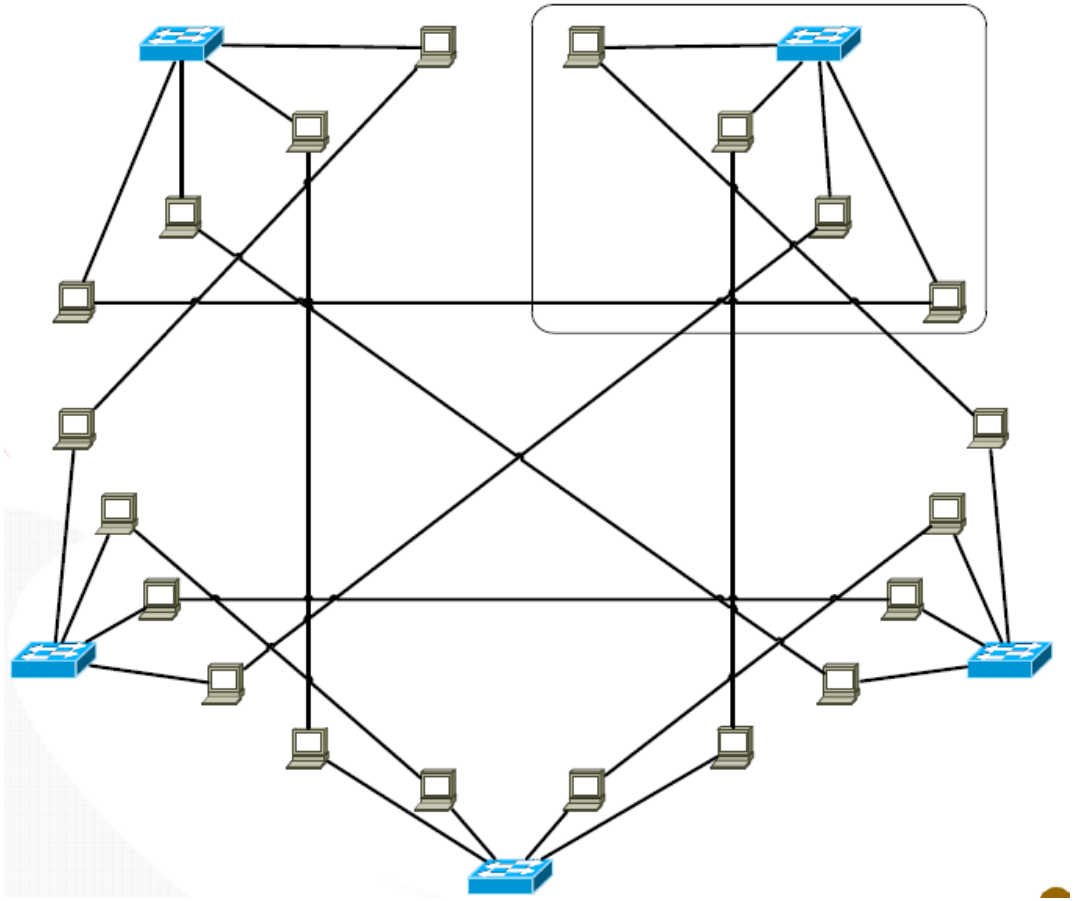
University of BRISTOL

BCube (4,1)



C. Guo *et al.*, “BCube: a high performance, server-centric network architecture for modular data centers,” *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4, pp. 63–74, 2009.

Dcell



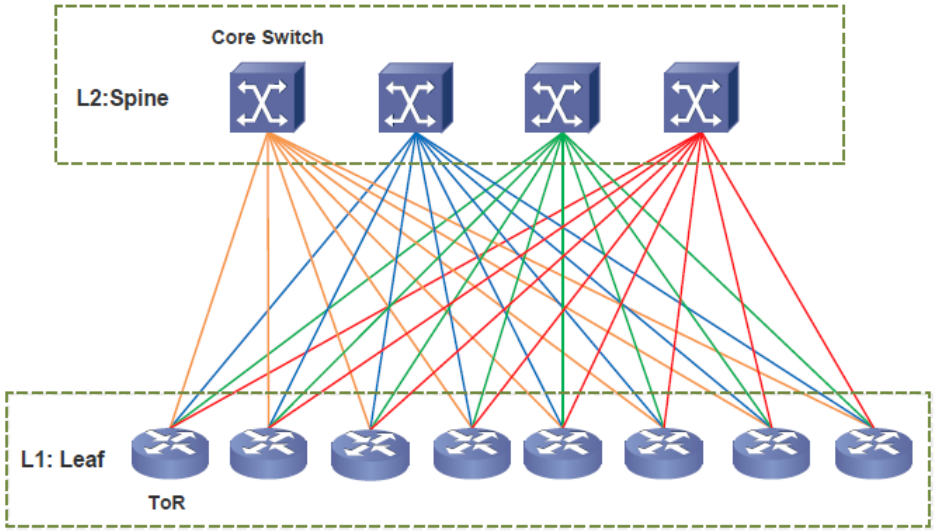
Level-2 Dcell

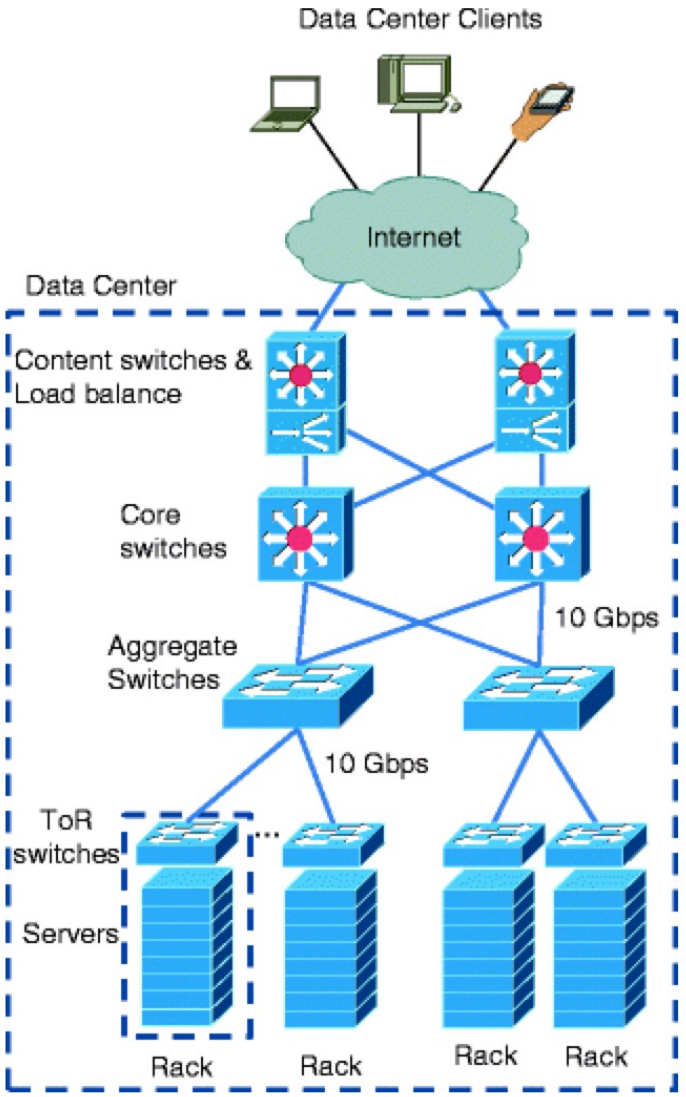
- An extremely scalable architecture and a 3 level DCell having 6 servers in DCell0 can accommodate around 3.26 Million servers.
- A hybrid routing and data processing protocol

C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, “Dcell: a scalable and fault-tolerant network structure for data centers,” *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4, pp. 75–86, 2008.

Summary

- Typical 3-Tier DCN architecture
- Spine-leaf DCN architecture
- The trends of DCN





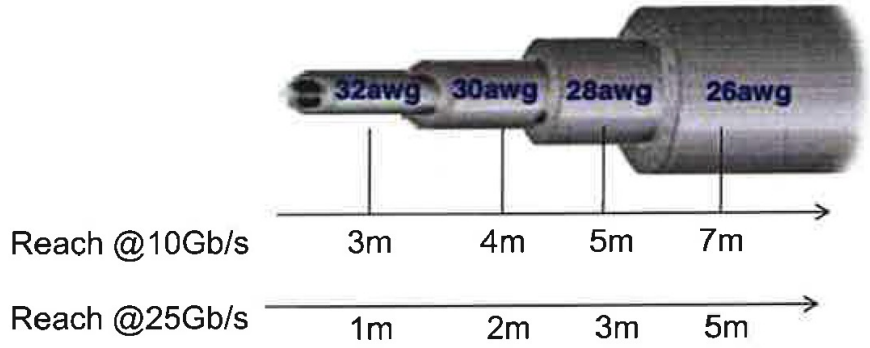
Links

System packaging: Copper Cables

University of BRISTOL

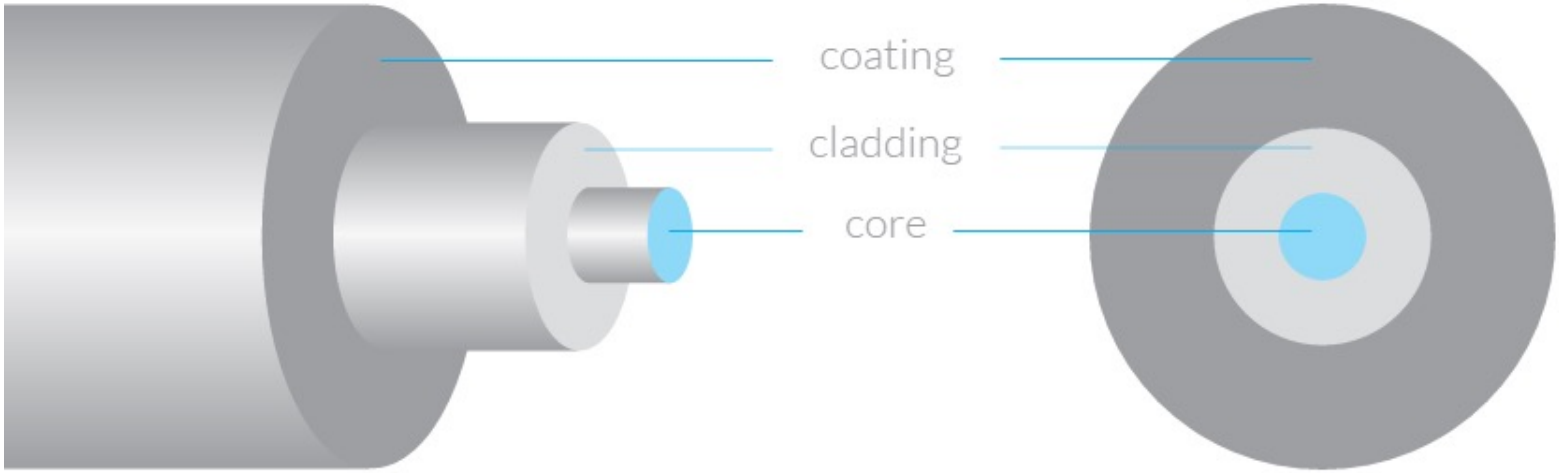
-

	32AWG	30AWG	28AWG	26AWG
Wire diameter	0.20	0.25	0.32	0.41
8x Cable outer diameter (mm)	4.4	5.5	6.5	7.1
Attenuation @ 5GHz (dB/m)	-4.0	-3.3	-2.4	-2.0
Attenuation @ 12.5GHz (dB/m)	-6.4	-5.2	-3.8	-3.2



- A single copper cable can support a single channel.
- The attenuation is dependent on the signal bandwidth.
- Copper cables experience 3dB per meter attenuated @ 12.5 GHz.
- They can offer up to 5meters reach for 25 Gb/s data channels.

Multimode fiber: low cost platform

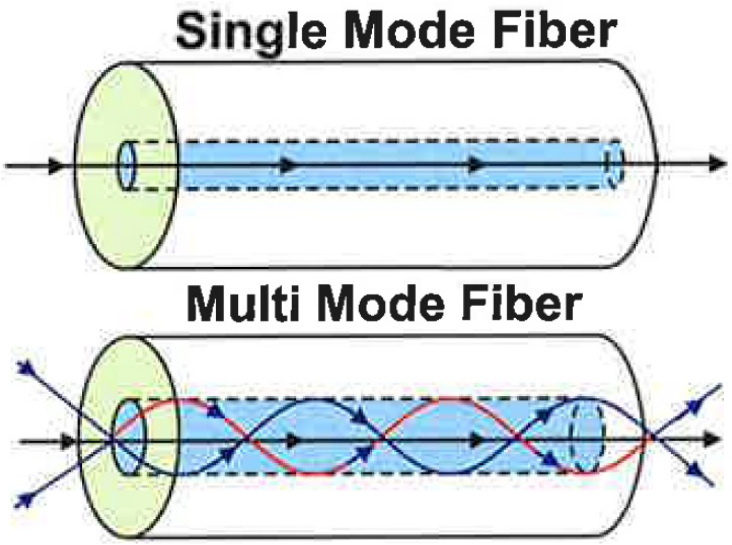


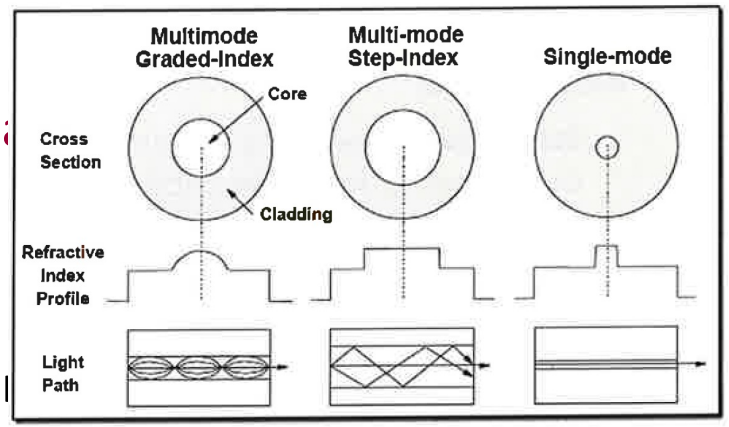
MMF: core diameter: 50 µm, 62.5 µm

Application	Standard	IEEE Reference	Media	Speed	Target Distance	
10-Gigabit Ethernet	10GBASE-SR	802.3ae	MMF	10Gb/s	33m (OM1) 300m (OM3)	82m (OM2) 400m (OM4)
	10GBASE-LR		SMF		10km	
	10GBASE-LX4		MMF		300m	
	10GBASE-ER		SMF		40km	
	10GGABE-LRM	802.3aq	MMF		220m (OM1/OM2)	300m (OM3)
40-Gigabit Ethernet	40GBASE-SR4	802.3bm	MMF	40Gb/s	100m (OM3)	150m (OM4)
	40GBASE-LR4		SMF		10km	
	40GBASE-FR		SMF		2km	
	40GBASE-ER4		SMF		10km	
100-Gigabit Ethernet	100GBASE-SR10		MMF		100m (OM3)	150m (OM4)
	100GBASE-LR4		SMF		10km	
	100GBASE-SR4		SMF		70m (OM3)	100m (OM4)
	100GBASE-ER4		SMF		40km	

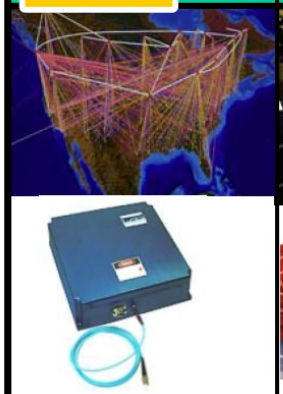
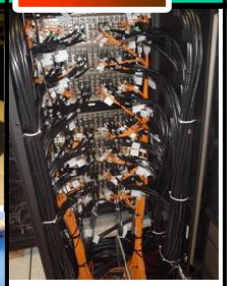
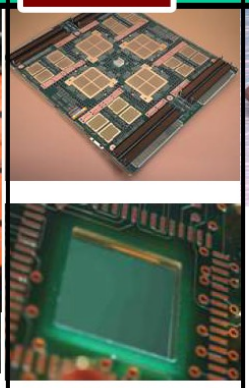
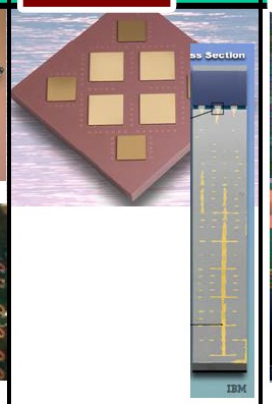
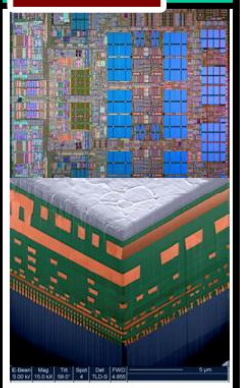
Fibers

- SMF: core diameter: 9 µm (single core)
  - Attenuation: ~ 0.2 dB/Km
- MMF: core diameter: 50 µm, 62.5 µm
- MMF (Multi Mode Fiber) has limited bandwidth
  - Modal dispersion limited: Multiple modes propagate at different velocities causing pulse spreading
  - Graded parabolic index profile reduces mod dispersion
    - Requires careful control of doping concentration layer by layer from the ce of the core



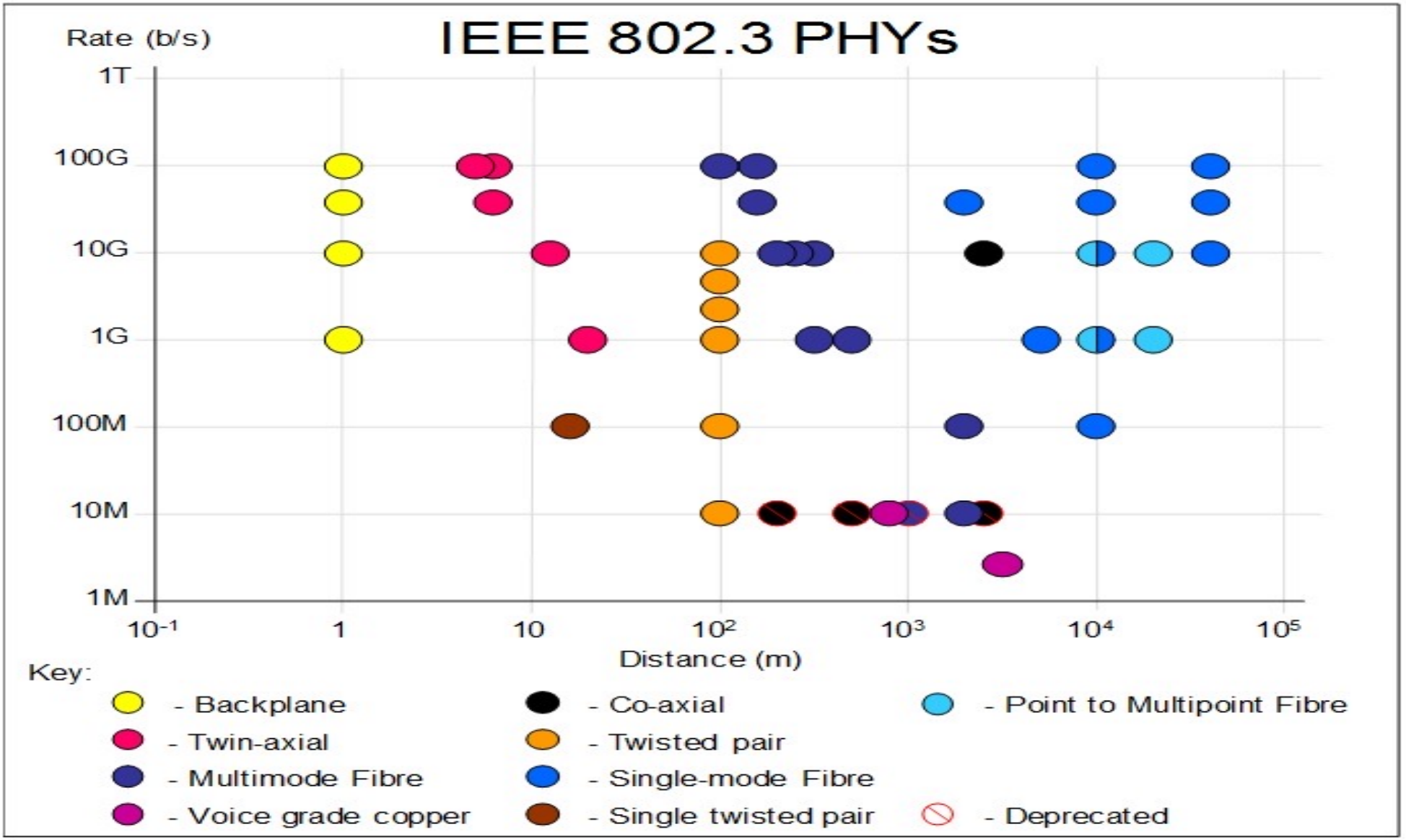


Optical Interconnect Evolution

PHYSICAL Link Types	MAN & WAN	Cables – Long	Cables – Short	Backplane / Card-to-Card	Intra-Card	Intra-Module	Intra-chip
Distinguished by Length & Packaging							
Length	Multi-km	10, - 300 m	1 m - 10 m	0.3 m - 1 m	0.1 m - 0.3 m	5 mm - 100 mm	0 mm - 20 mm
Typical # lanes per link	1	1 - 10s	1 - 10s	1 - 100s	1 - 100s	1 - 100s	1 - 100s
Use of optics	Since 80s	Since 90s	Since late 00's	Since 2010-2011	2012-2015	After 2015	Later

Ref: A. Benner, "Optical Interconnect Opportunities in Supercomputers and High End Computing, OFC 2012

Speed and reach for various IEEE Std 802.3 MAUs and PHYs



The rise of optical interconnects

University of BRISTOL

- Currently the optical technology is utilized in data centers only for point-to-point links. These links are based on low cost multi-mode fibers (MMF) for short reach communication.
- These MMF links are based for connections of the switches using fiber-based Small Form-factor Pluggable transceivers (SFP for 1Gbps and SFP+ for 10 Gbps)
- Higher bandwidth transceivers will be adopted (for 40 Gbps and 100 Gbps Etherent) such as 4x10 Gbps QSFP modules with four 10 Gbps parallel optical channels and CXP/CFP2/… for 100 Gbps using VCSELs or DFBs.