POLITECNICO DI MILANO

Computing Infrastructures





Networking _ Guido Maier

Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano

Ph: +39 022399 3575, Fax +39 022399 3413

guido.maier@polimi.it

Acknowledgements

- This set of slides has been prepared by Guido Maier, thanks to a generous and ample contribution by Prof. Paolo Giaccone of Politecnico di Torino, Italy, who offered some teaching material of his course "Switching technologies for data centers", a.a. 2021/22
- I would like to acknowledge also the contributions of Proff. Manuel Roveri and Achille Pattavina, Politecnico di Milano, and Prof. Roberto Rojas-Cessa, New Jersey Institute of Technology, USA

Outline

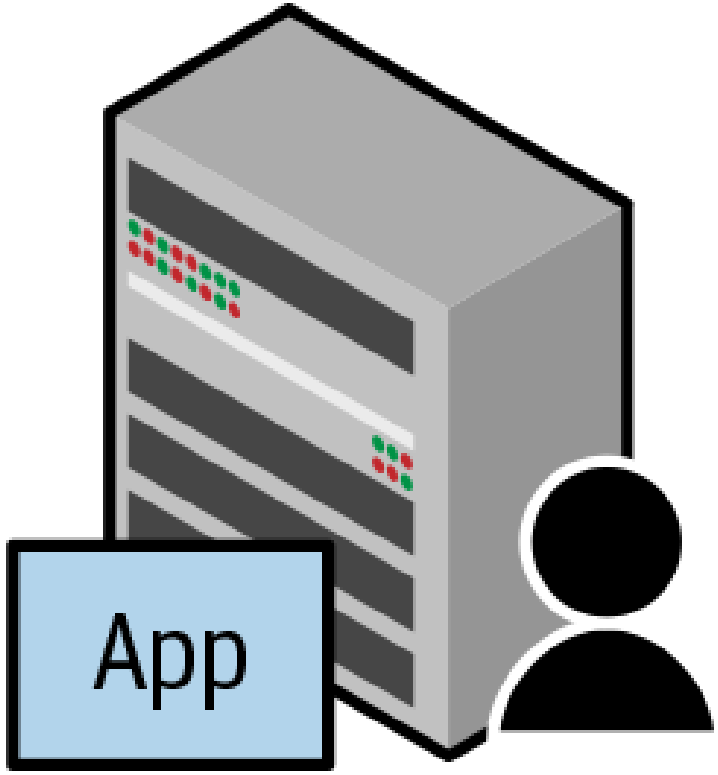
- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures

Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures

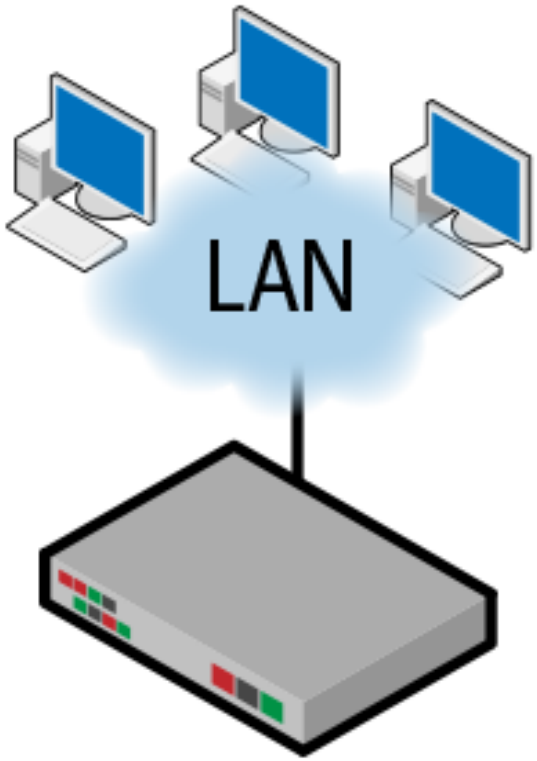
Stages of enterprise infrastructures
1 – Monolithic app

- Minimal network demands
- Proprietary protocols



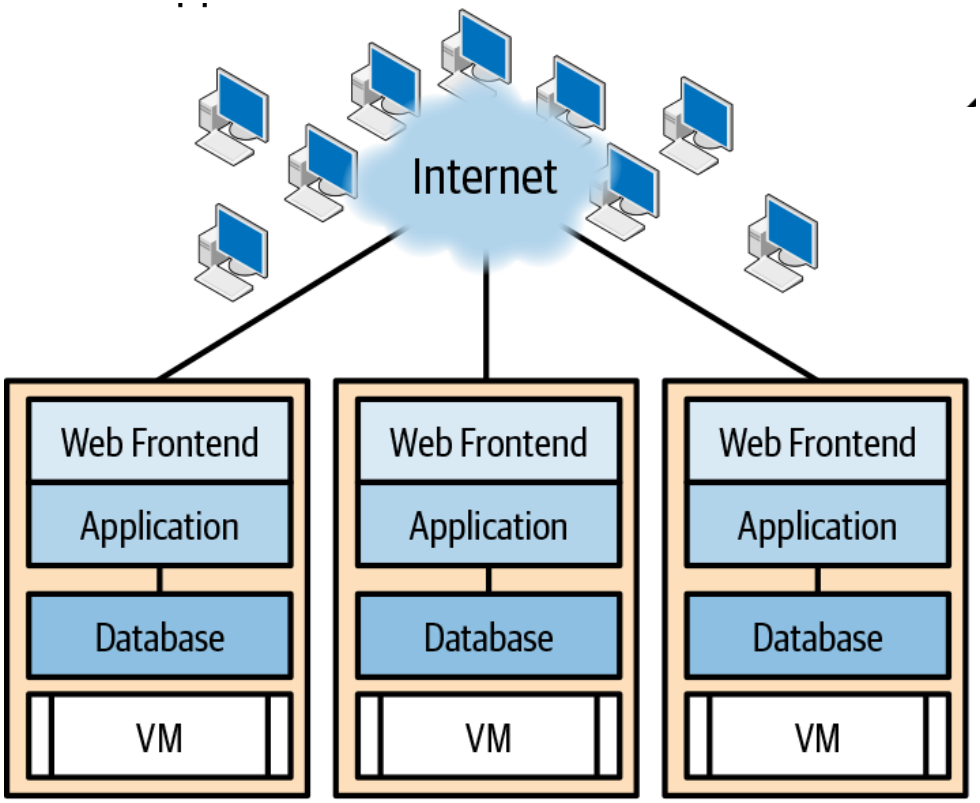
Stages of enterprise infrastructures
2 – Client server

- High network demands inside the enterprise
- Applications walled within the enterprise
- TCP/IP + Proprietary protocols



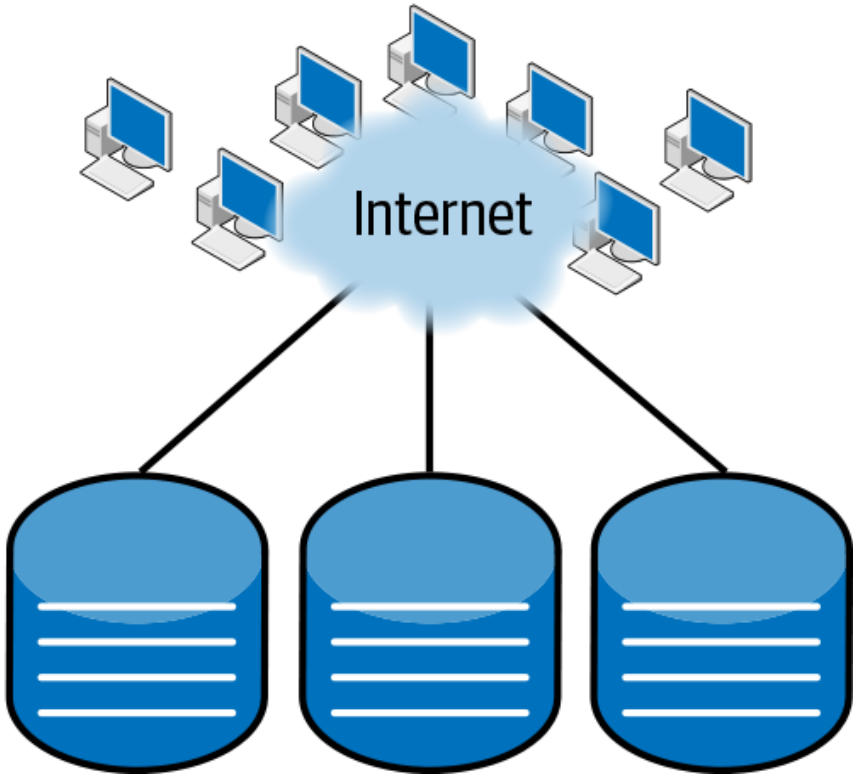
Stages of enterprise infrastructures
3 – Web applications

- Ubiquitous TCP/IP
- Access from anywhere
- Servers are broken into multiple units



Stages of enterprise infrastructures
4 - Microservices

- Infrastructure moved to cloud providers
- Servers broken into microservices
- Increase of server-to-server traffic



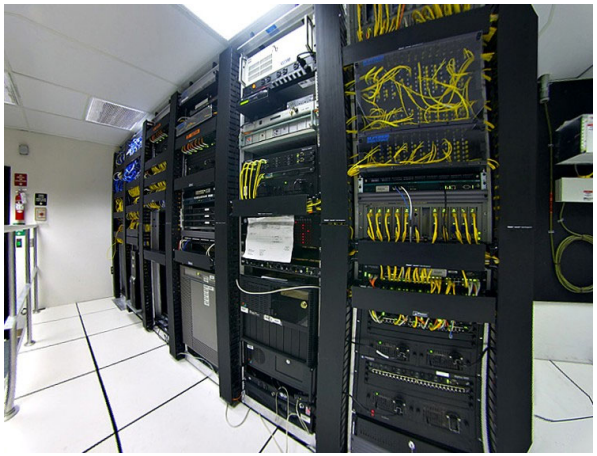
Datacenters – aka Warehouse Scale Computing (WSC)

- 100k servers!
- Cooling
- Power supply
- Renewable energy

Data centers

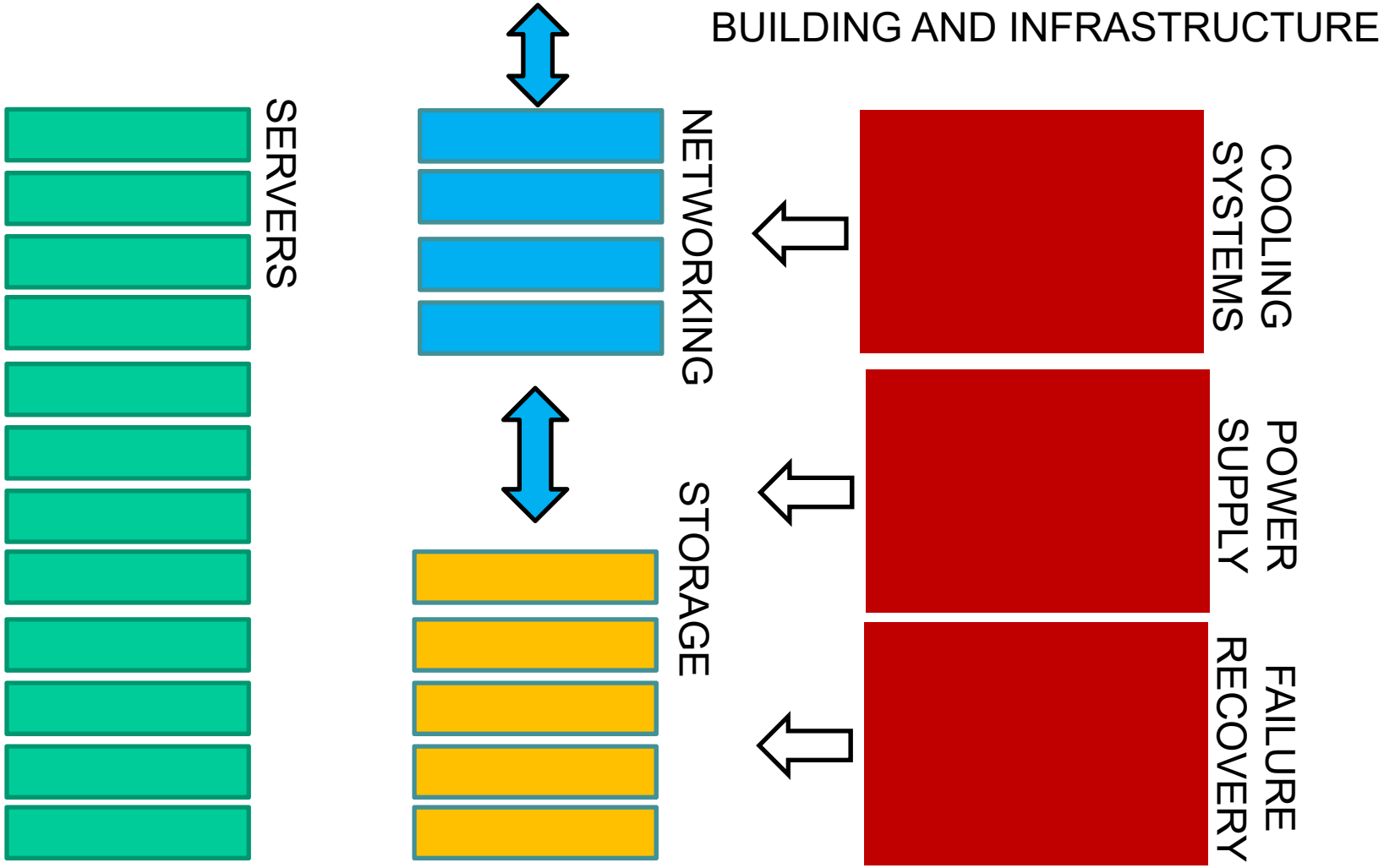
Data center

- Set of all the physical infrastructure required to support a cloud computing service
- The whole infrastructure is co-located either in a room, or in a building, or in a set of adjacent buildings

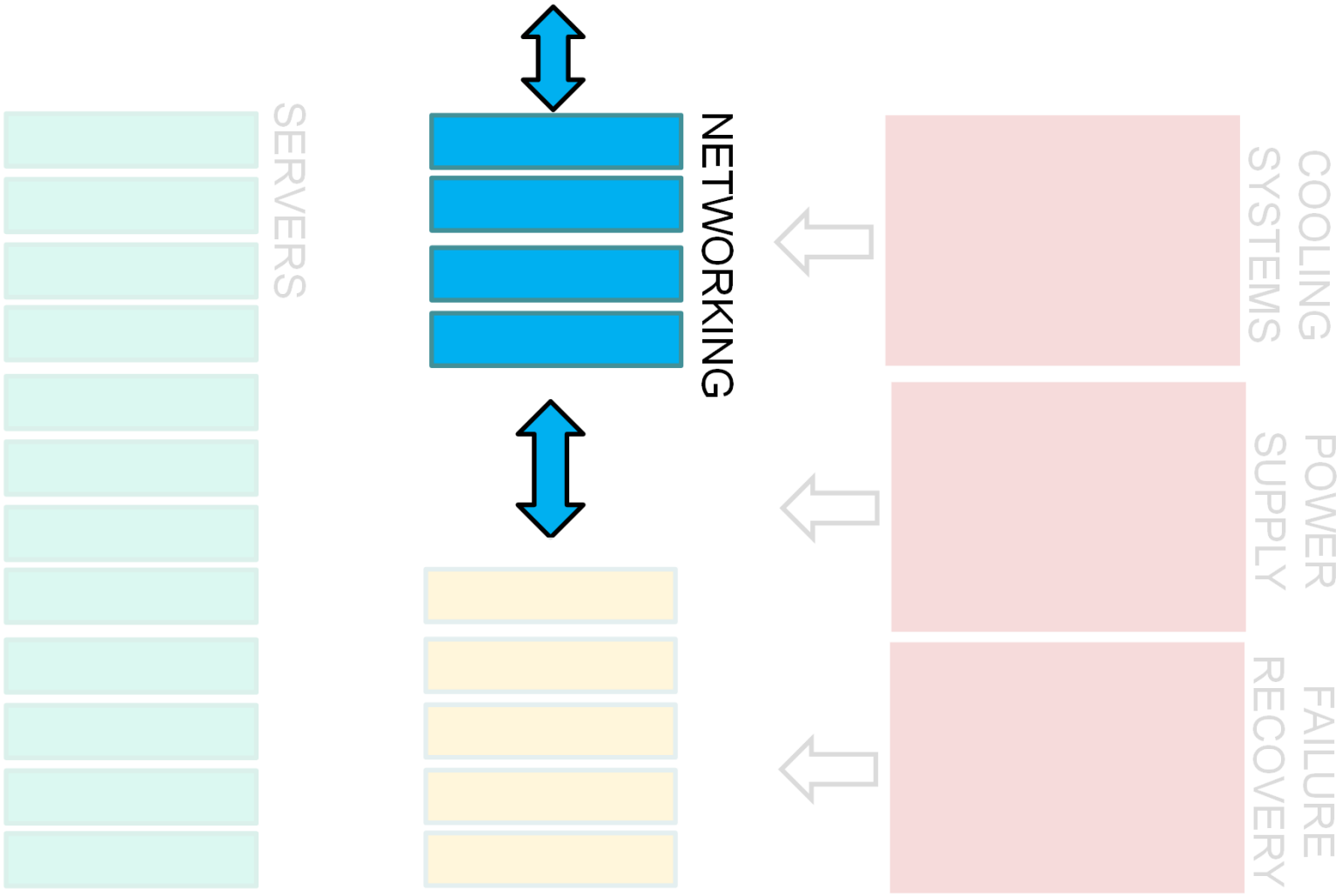


- Applications
  - cloud computing
  - cloud storage
  - web services
- Consolidation of computation and network resources
- Very large data centers
  - 1 000 - 10 000 - 100 000 - 1 000 000 servers

Data center infrastructure

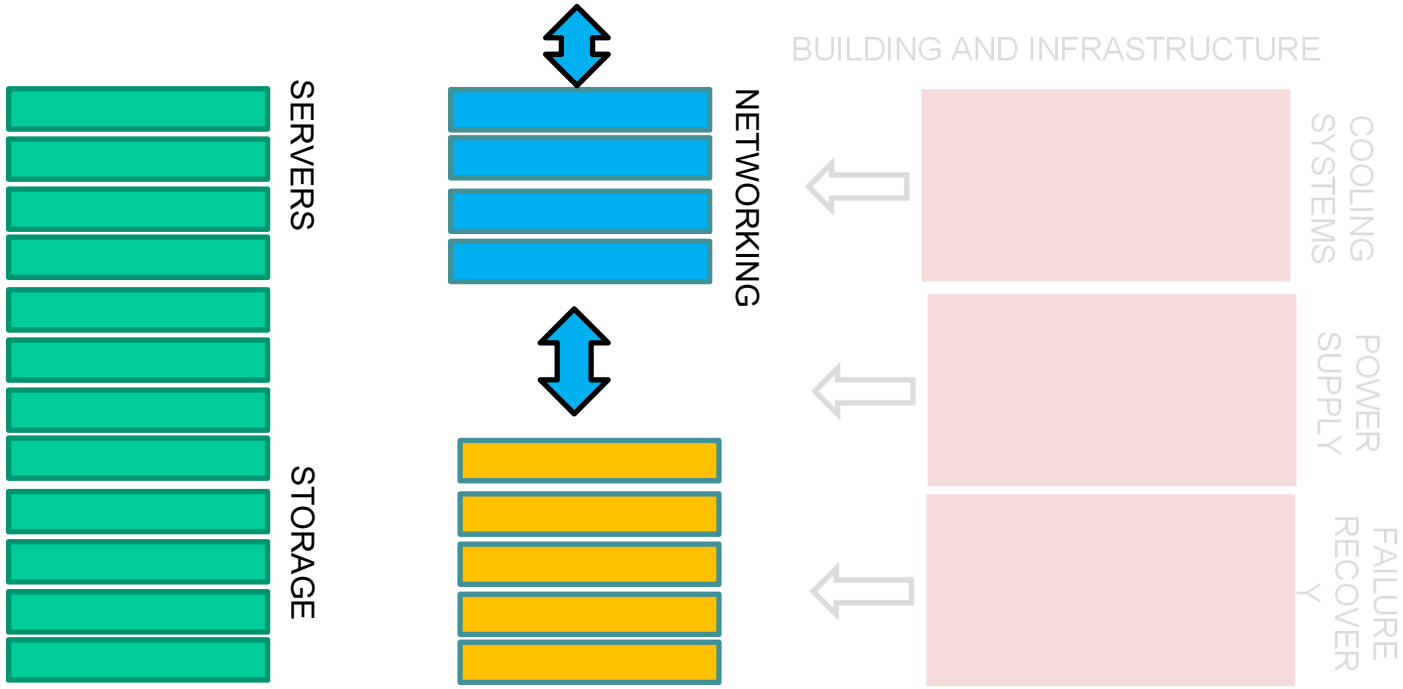


Intra-Datacenter networking
(aka DataCenter [interconnection] Network – DCN)



The need for effective networking in WSCs

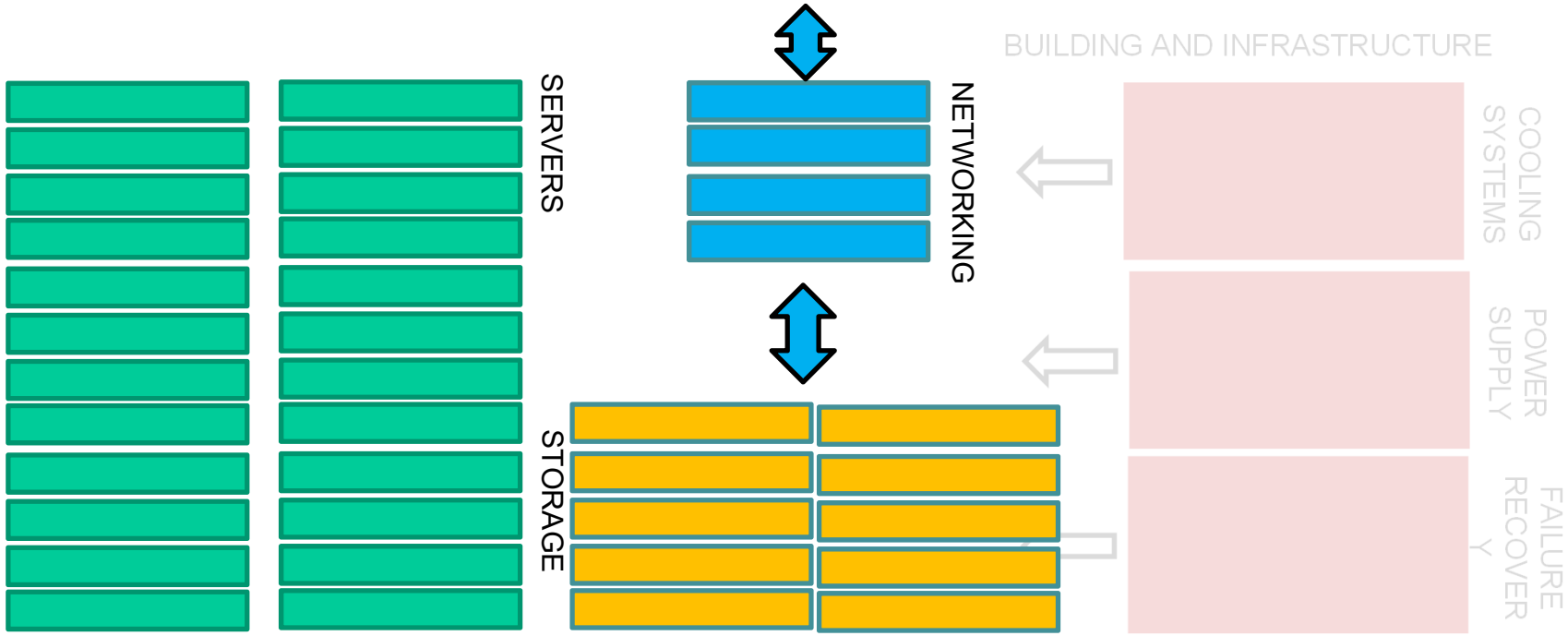
- The performance of servers increases over time, the demand for inter-server bandwidth naturally increases as well!!!



The need for effective networking in WSCs

- The performance of servers increases over time, the demand for inter-server bandwidth naturally increases as well!!!

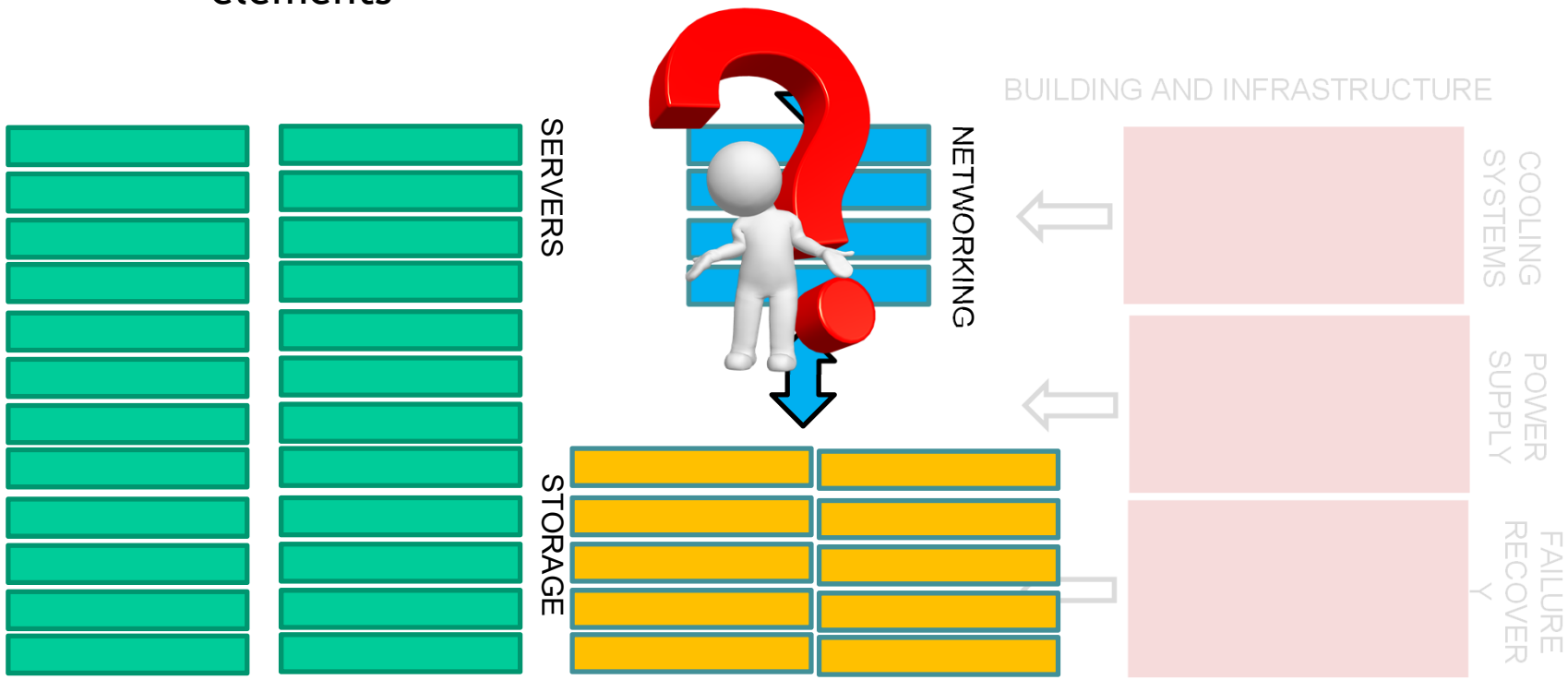
  ⇒ We can double the aggregate compute capacity or the aggregate storage simply by doubling the number of compute or storage elements



The need for effective networking in WSCs

- The performance of servers increases over time, the demand for inter-server bandwidth naturally increases as well!!!

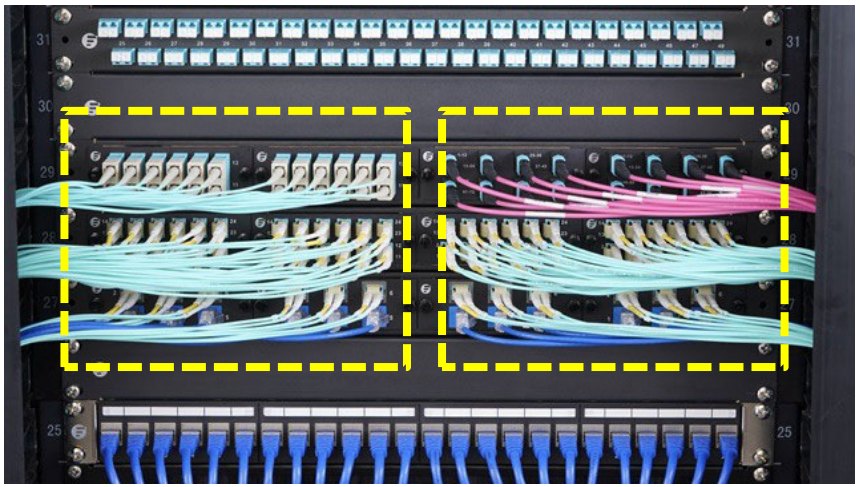
  ➡ We can double the aggregate compute capacity or the aggregate storage simply by doubling the number of compute or storage elements



The need for effective networking in WSCs

- Networking has no straightforward horizontal scaling solution.
- Doubling leaf bandwidth is easy:
  - with twice as many servers, we'll have twice as many network ports and thus twice as much bandwidth.

x 1

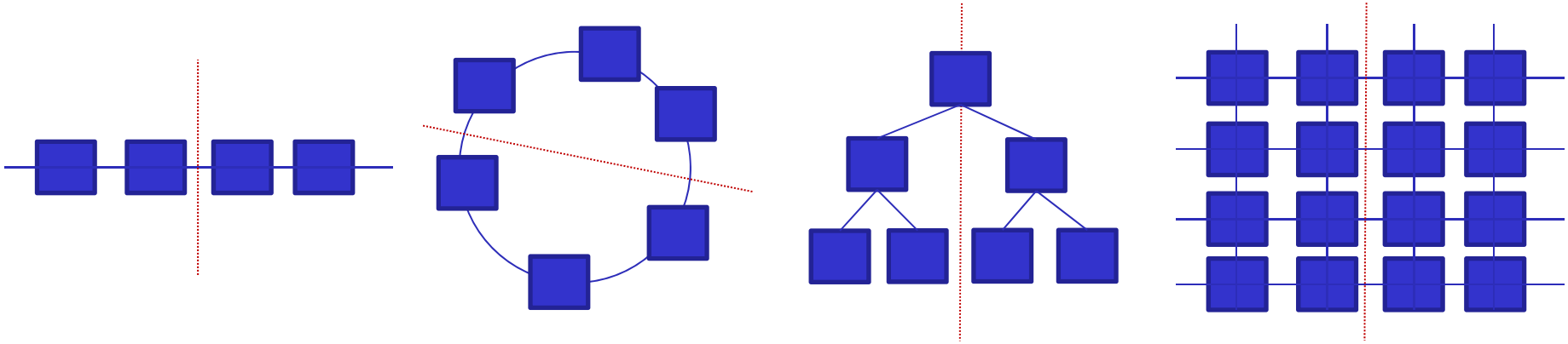


x 2

- But if we assume that every server needs to talk to every other server, we need to deal with bisection bandwidth

Bisection bandwidth

- The bandwidth across the narrowest line that equally divides the cluster into two parts
- Characterizes network capacity since randomly communicating processors must send data across the “middle” of the network



If we assume that every server needs to talk to every other server, we need to double not just leaf bandwidth but *bisection bandwidth*

Classes of DCN

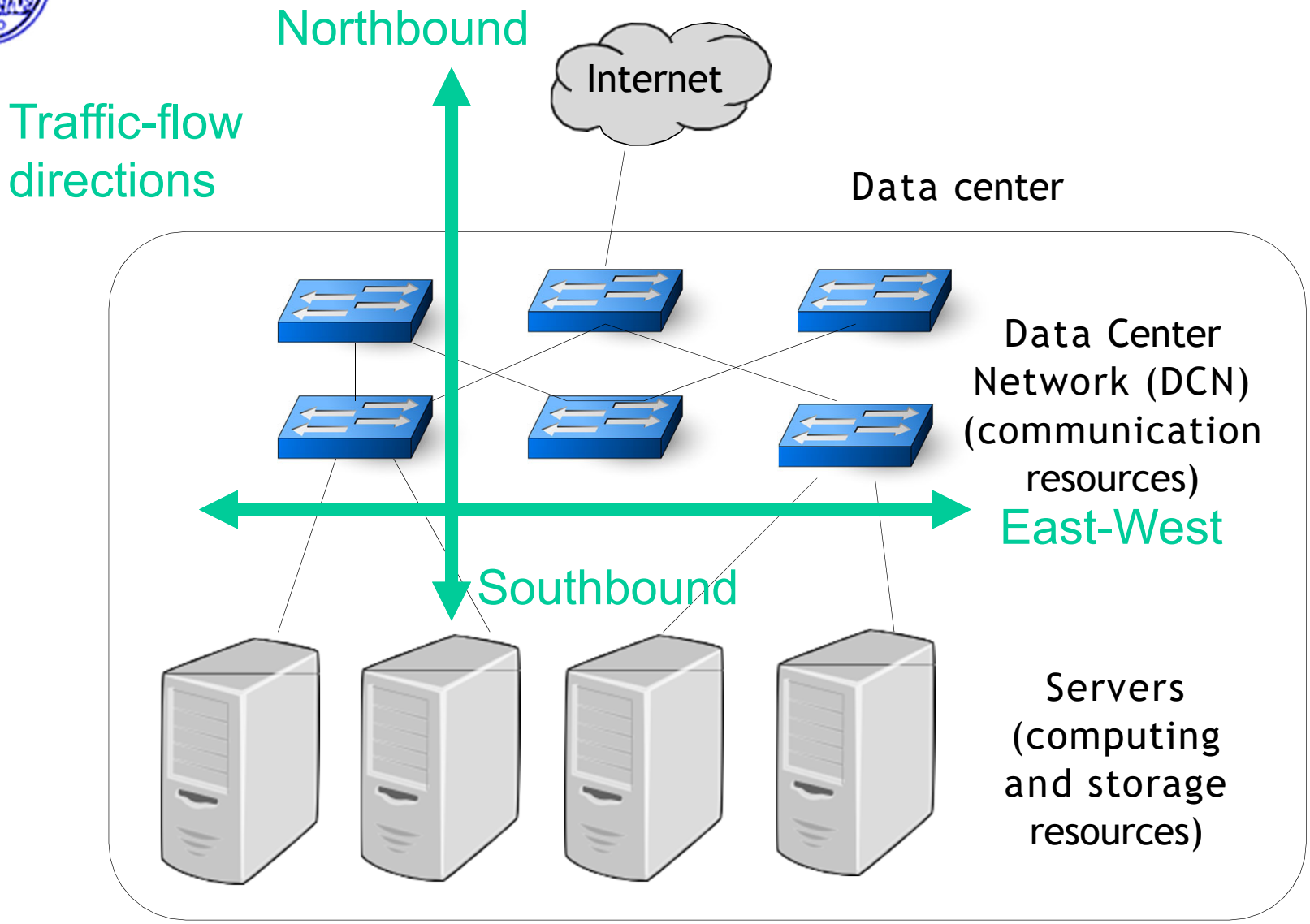
DCNs can be classified into three main categories:

- **Switch-centric** architectures
  - Uses switches to perform packet forwarding
- **Server-centric** architecture
  - Uses servers with multiple Network Interface Cards (NICs) to act as switches in addition to performing other computational functions
- **Hybrid** architectures
  - Combine switches and servers for packet forwarding

Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures

Server-centric architectures



East-West traffic

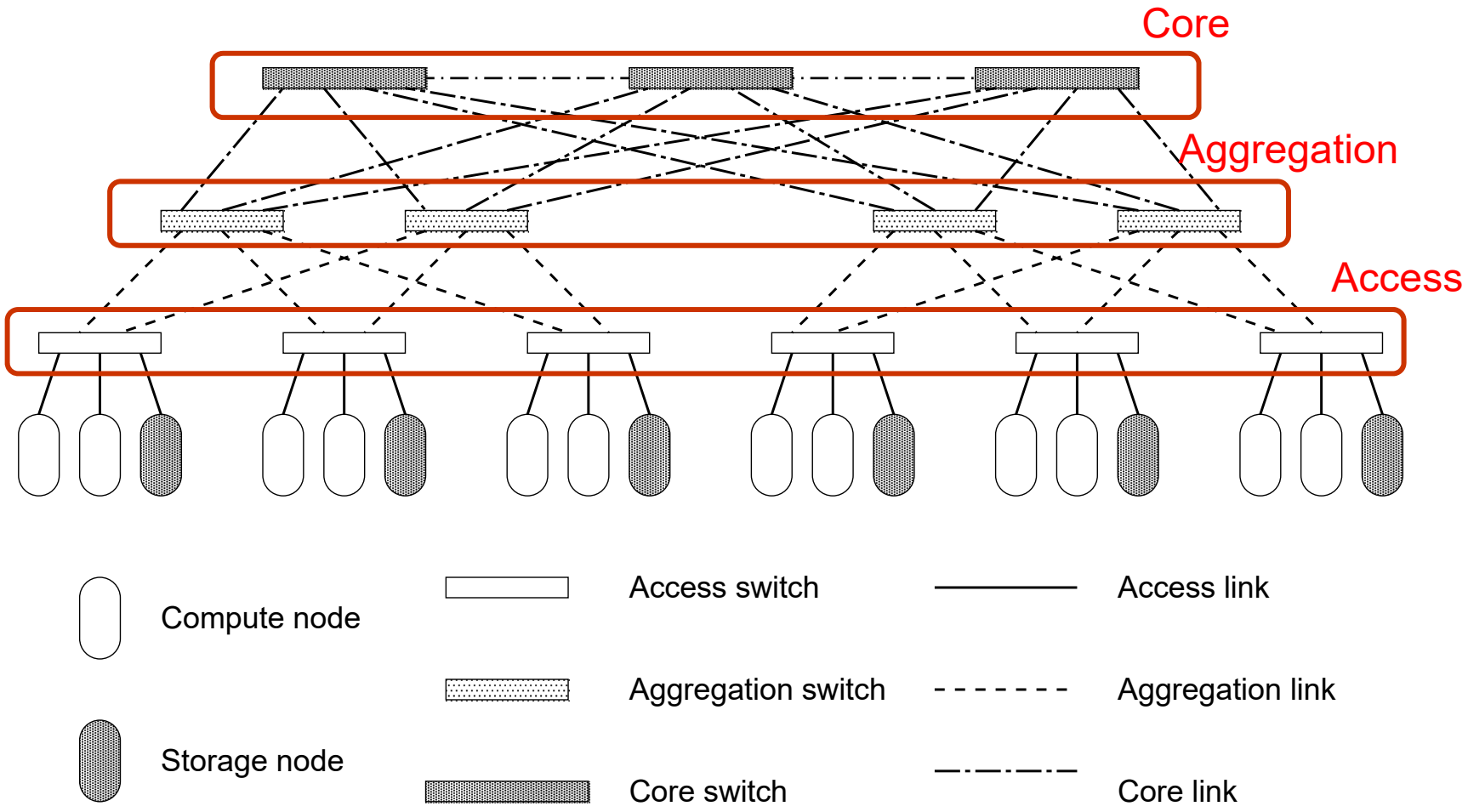
- East-West traffic
  - storage replication (few flows, many data)
    - in Hadhoop distributed filesystem, at least 3 copies of the same data, usually two in the same rack and one in another rack
  - VM migration
  - Network Function Virtualization (NFV)
    - data is processed through a sequence of VMs (e.g., firewall, web server, parental control, accounting server)
- East-West traffic usually larger than North-South traffic
  - Some citations:
    - A 1 byte transaction in North-South traffic generates on average a 100 bytes transaction in East-West traffic
    - According to Cisco's Global Cloud Index (http://blogs.cisco.com/security/trends-in-data-center-security-part-1-traffic-trends, May 2014):
      - In a data center: East-West traffic (76%), North-South traffic (17%), inter-data center traffic (7%).
      - In campus networks: North-South traffic (>90%).

East-West traffic

- Unicast
  - point-to-point communication
  - e.g., VM migration, data backup, stream data processing
- Multicast
  - one-to-many communication
  - e.g., software update, data replication (≥ 3 copies per content) for reliability, OS image provision for VM
- Incast
  - many-to-one communication
  - e.g., reduce phase in MapReduce, merging tables in databases

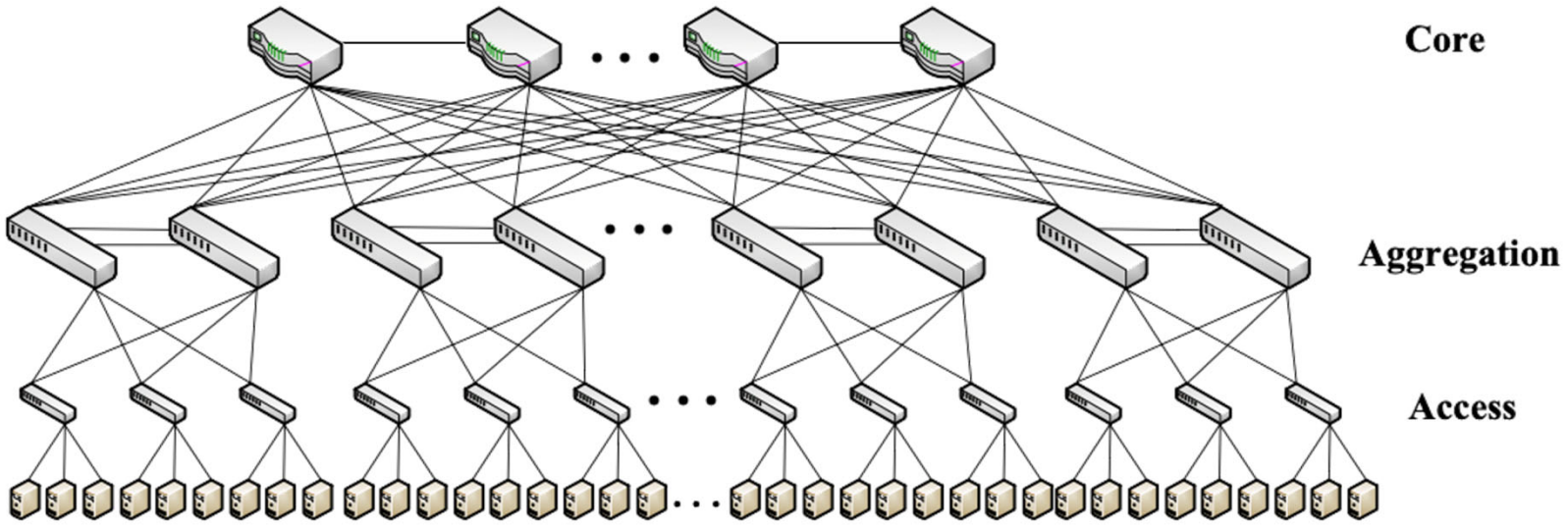
Three-Tier (or layer) “Classical” Network

- Three layer architecture configures the network in three different layers:



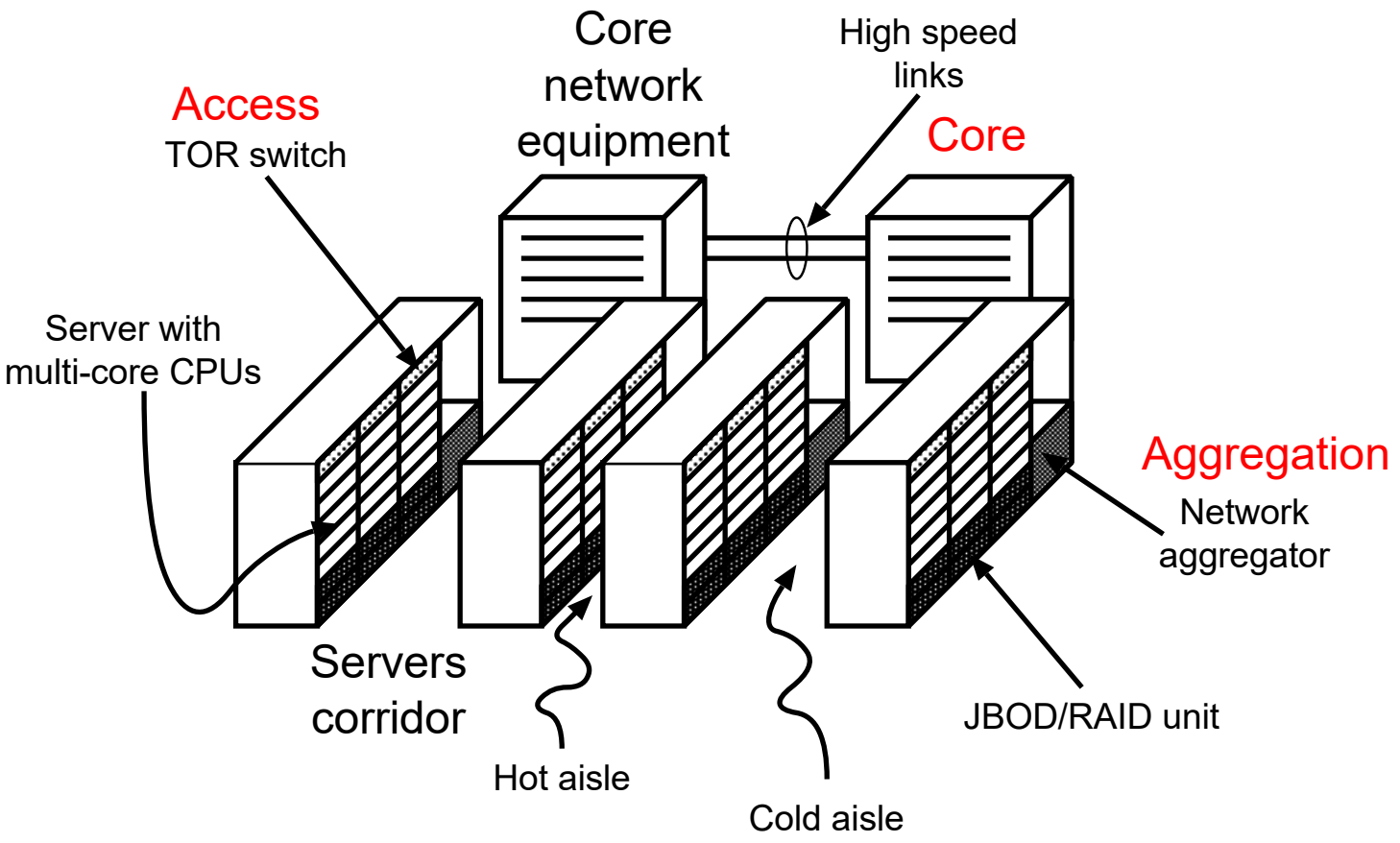
Three-Tier (or layer) “Classical” Network

- A simple DCN topology
- Servers are connected to the DCN through access switches.
- Each access-level switch is connected to at least two aggregation-level switches.
- Aggregation-level switches are connected to core-level switches (gateways).



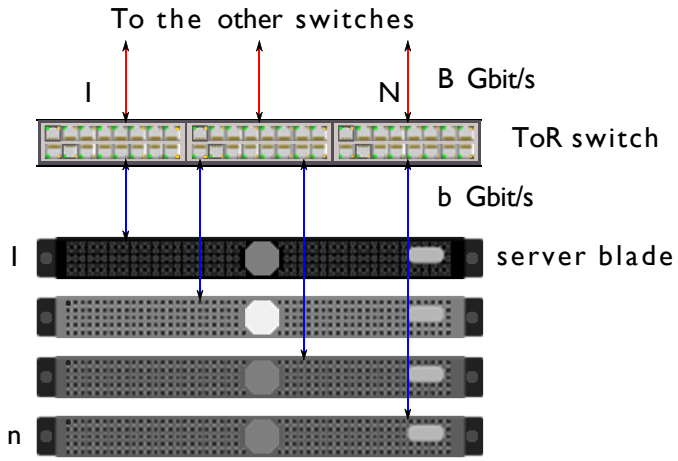
Data-center network architectures

- Three layer architecture reflects the topology of the data center:



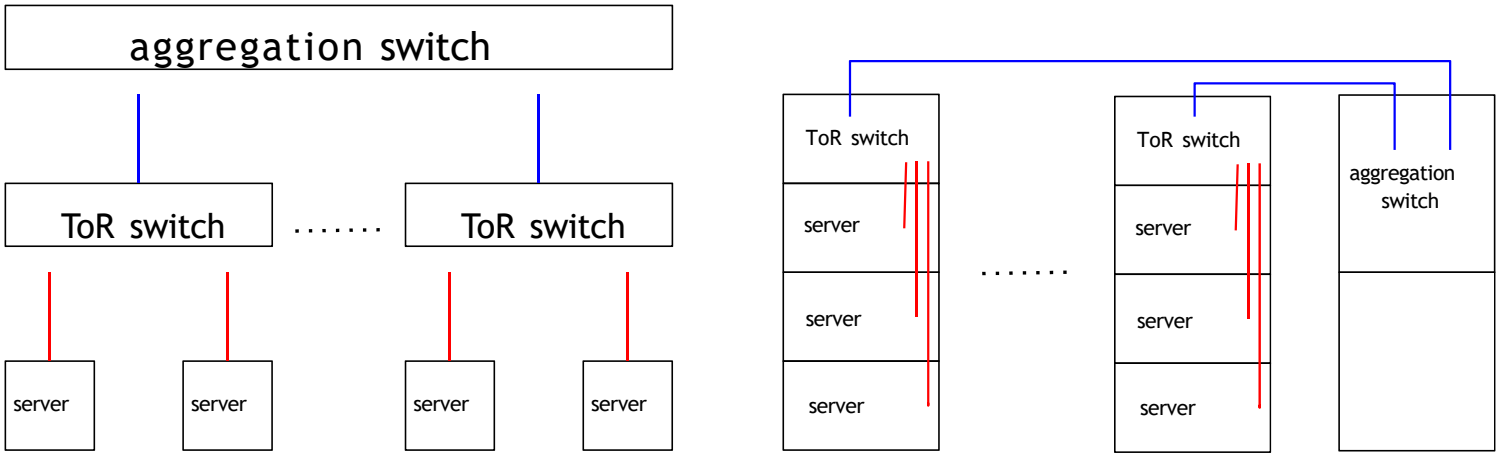
Server packing in a rack

- Standard 19 inch rack 42 EIA Units (pizza box)
  - 40 server blades
    - possible single /26 subnet
  - 1 ToR (Top of Rack) switch
- without oversubscription: NB = nb
  - example
    - 40 ports @ 1 Gbit/s to the servers
    - 4 ports @ 10 Gbit/s to the other switches
- with oversubscription: NB < nb
  - example with oversubscription 1:4
    - 40 ports @ 1 Gbit/s to the servers
    - 1 ports @ 10 Gbit/s to the other switches



ToR vs EoR architectures

- ToR (Top-of-Rack) architecture
  - in a rack, all servers are connected to a ToR access switch
  - the servers and the ToR switch are colocated in the same rack
  - aggregation switches are in dedicated racks or in shared racks with other ToR switches and servers
  - The number of cables is limited → simpler cabling. The number of ports per switch is also limited (lower costs)
  - Limited scalability, higher complexity for switch management (high number of switches)

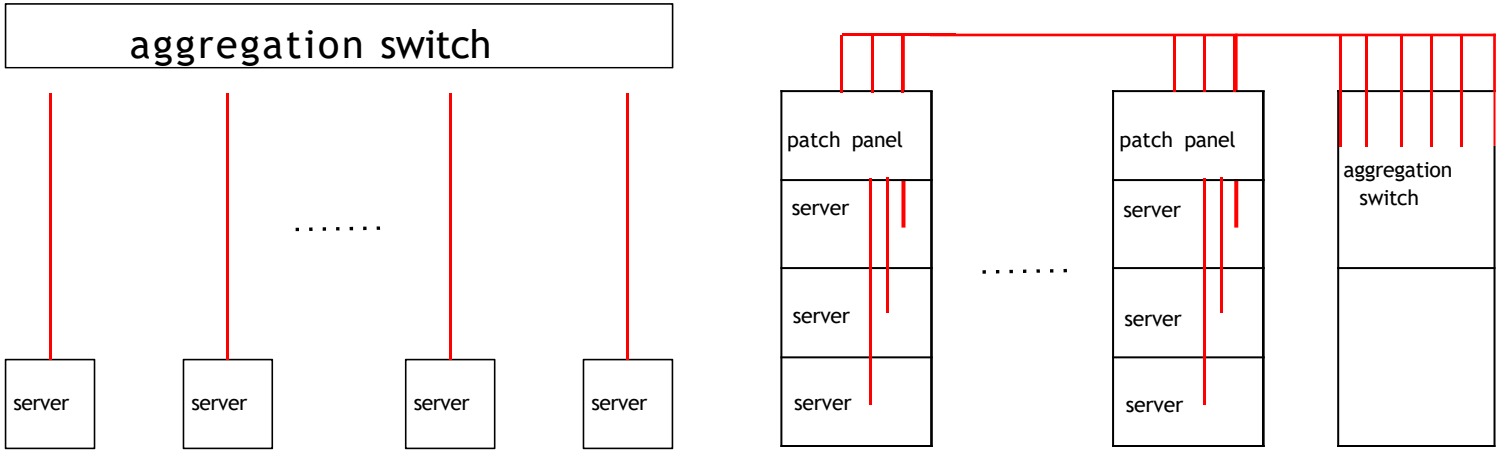


logical topology

physical topology

ToR vs EoR architectures

- EoR (End-of-Row) architecture
  - Aggregation Switches are positioned one per corridor, at the end of a line of rack.
  - servers in a racks are connected directly to the aggregation switch in another rack
  - Aggregation switches must have a larger number of ports,
  - more complex cabling, longer cables are required (higher costs)
  - patch panel to connect the servers to the aggregation switch
  - simpler switch management (lower number of switches)

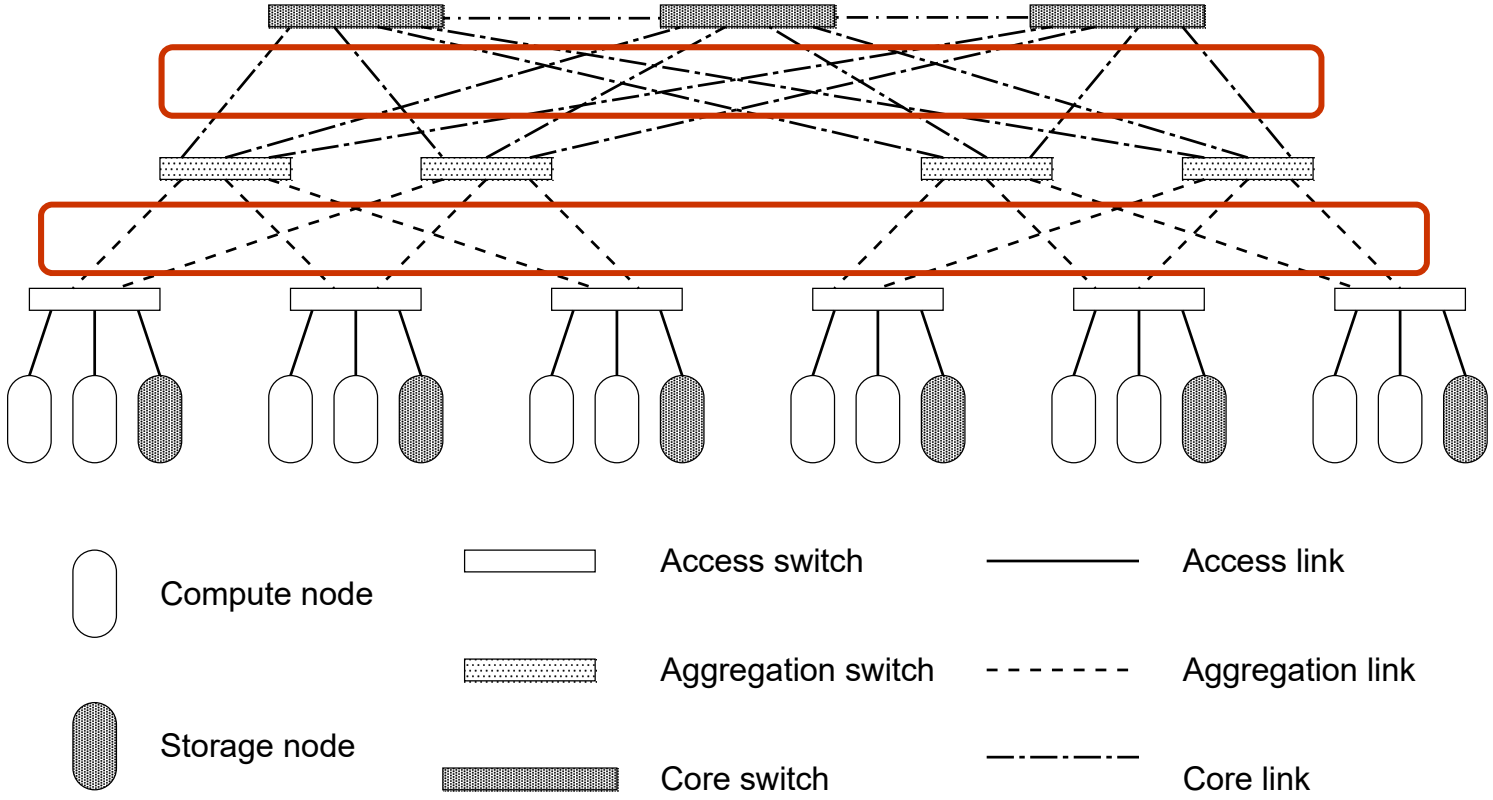


logical topology

physical topology

Three-Tier (or layer) "Classical" Network

Bandwidth can be increased by increasing the switches at the core and aggregation layers, and by using routing protocols such as Equal Cost Multiple Path (ECMP) that equally shares the traffic among different routes



Three-Tier (or layer) “Classical” Network

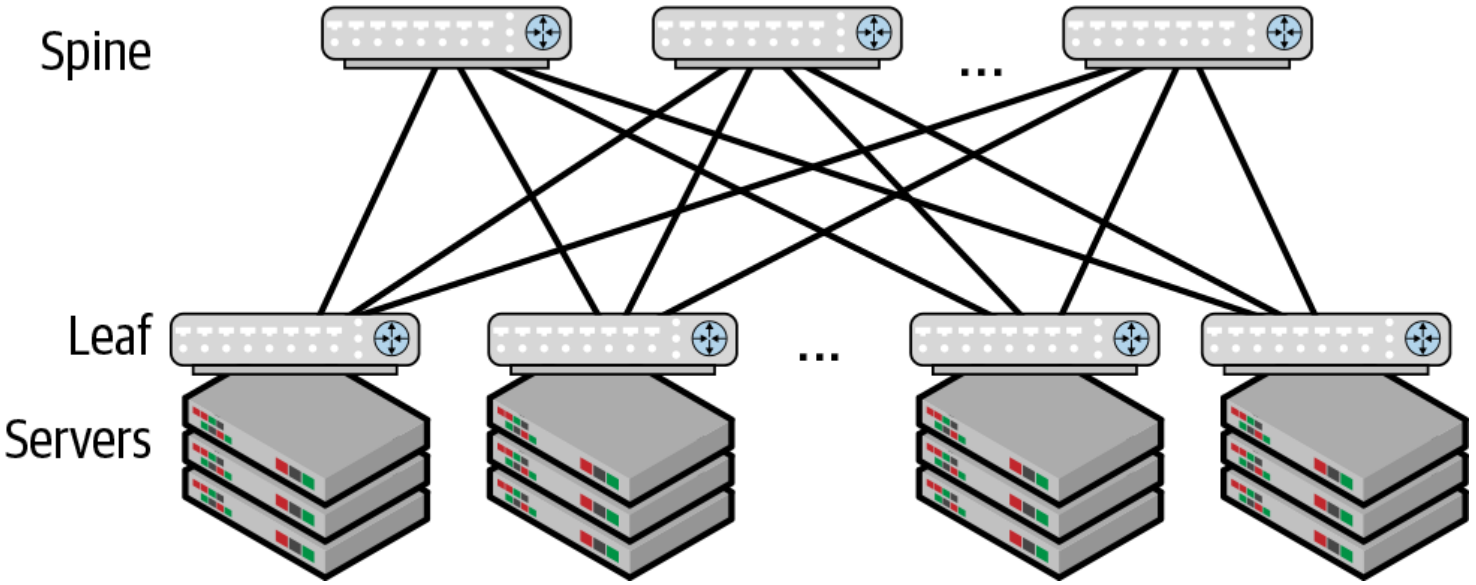
- This solution is very simple, but can be very expensive in large data-centers since:
  - Upper layers require faster network equipments. For example:
    - 1 GB Ethernet at the access layer
    - 10 GB Ethernet at the aggregation layer
    - 25 GB Optical connections at the core layer
  - Each layer is implemented by switches of a different kind
  - The cost in term of acquisition, management, spare-part stocks and energy consumption can be very high

Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures

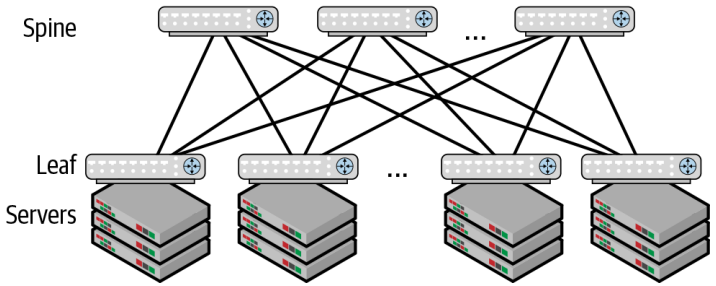
DNC Clos topology (a.k.a. Leaf-Spine)

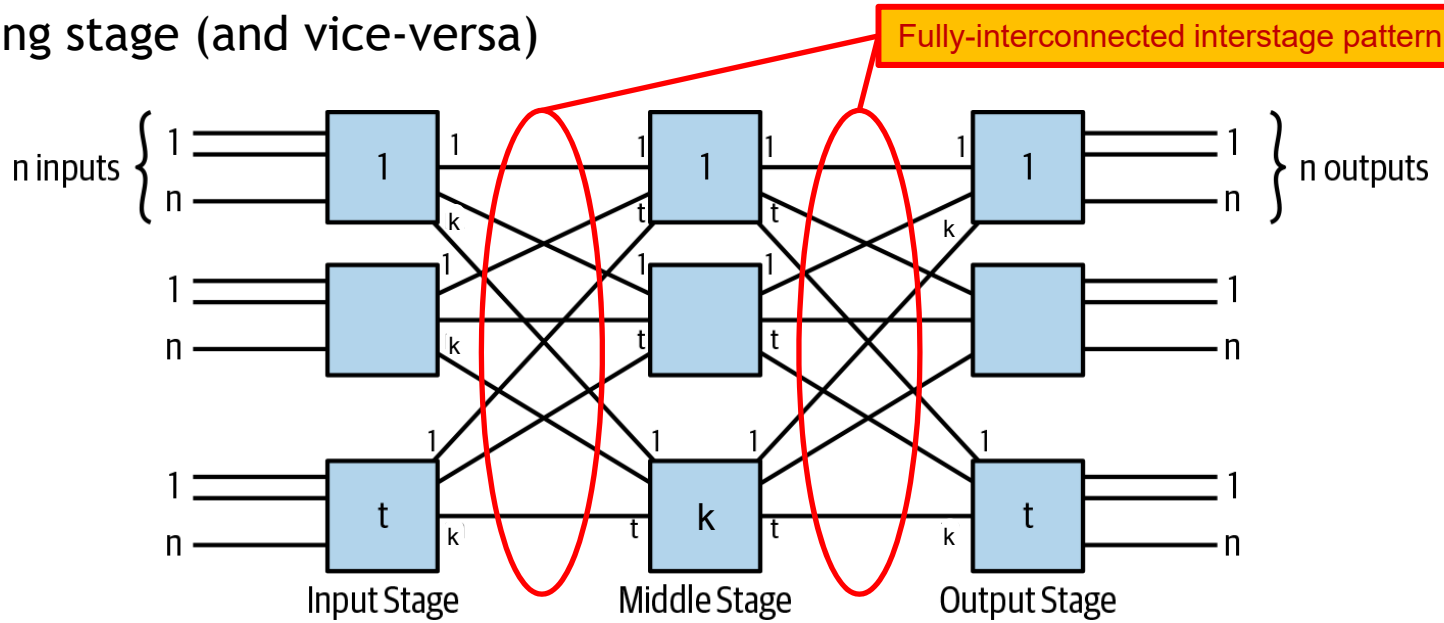
- Two stage interconnections
  - Leaf: ToR switch
  - Spine: dedicated switches (aggregation switches)



Clos networks

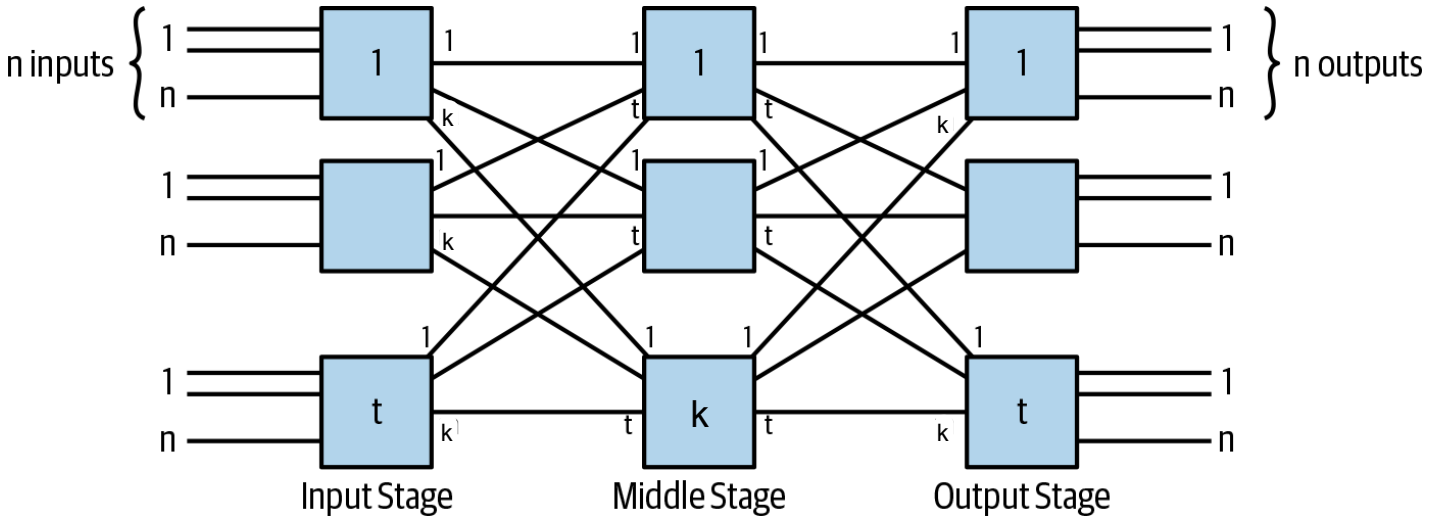
- Spine-leaf topologies are borrowed from the telephone world
- Non-folded Clos structure
- Fully-interconnected stages
  - Each matrix of one stage is connected (at least once) with each matrix of the following stage (and vice-versa)





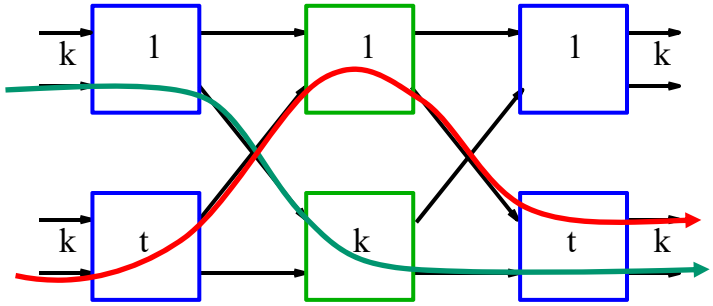
Clos networks in the telephone world

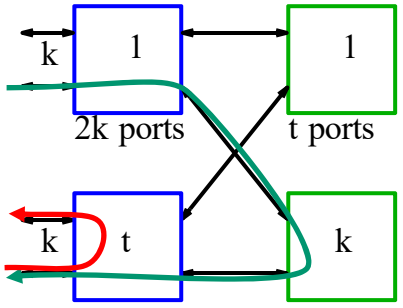
- Let k be the number of middle stage switches
- Let n be the number of input and outputs of switches of side stages
  - If k ≥ n there is always a way to rearrange communications to free a path between any pair of idle input/output
  - If k ≥ 2n - 1 there is always a free path between any pair of idle input/output
    - Notice that t is a free design parameter → the total number of input/output N=t·n can scale freely (by increasing the size of middle-stage switches)
- But a DCN is a PACKET-SWITCHED network!!



From Clos to "Leaf and Spine" topology

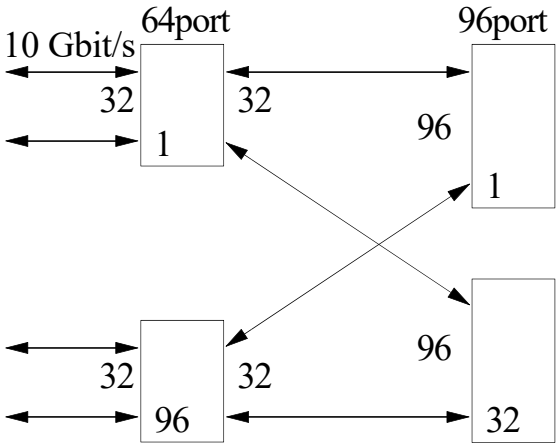
- Clos topology (n = m = k)
  - each switching module is unidirectional
    - k input + k output ports per module
  - k matrices in the central stage
  - t matrices in the side stages
  - each path traverses 3 modules
- Leaf and spine topology
  - each switching module is bidirectional
    - Leaf: t switching modules with 2k bidirectional ports per module
    - Spine: k switching modules with t bidirectional ports per module
  - each path traverses either 1 or 3 modules

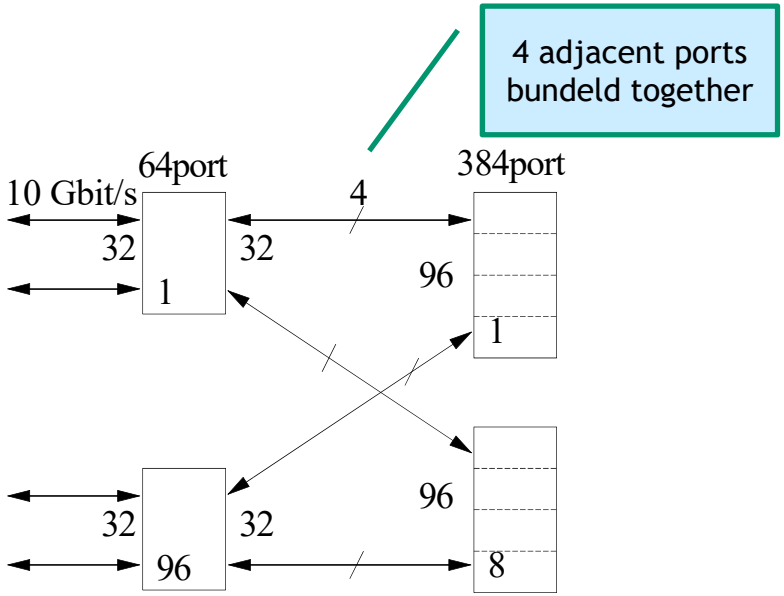




Examples of DCN design

- 3072 servers
  - 3072 ports at 10 Gbit/s ⇒ 30.72 Tbit/s
    alternative designs
    - t=96 switches with 64 ports and k=32 switches with 96 ports
    - t=96 switches with 64 ports and k/4=8 switches with 384 ports

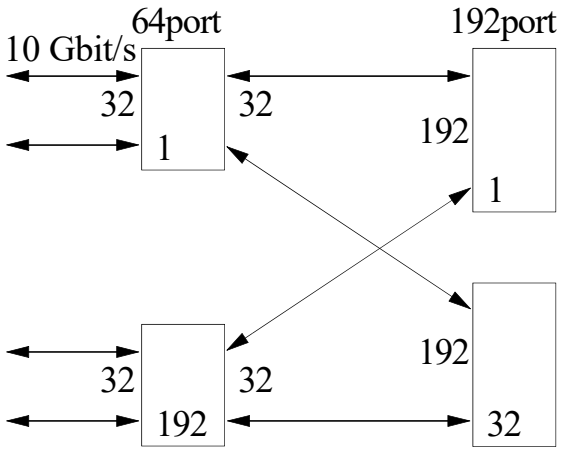


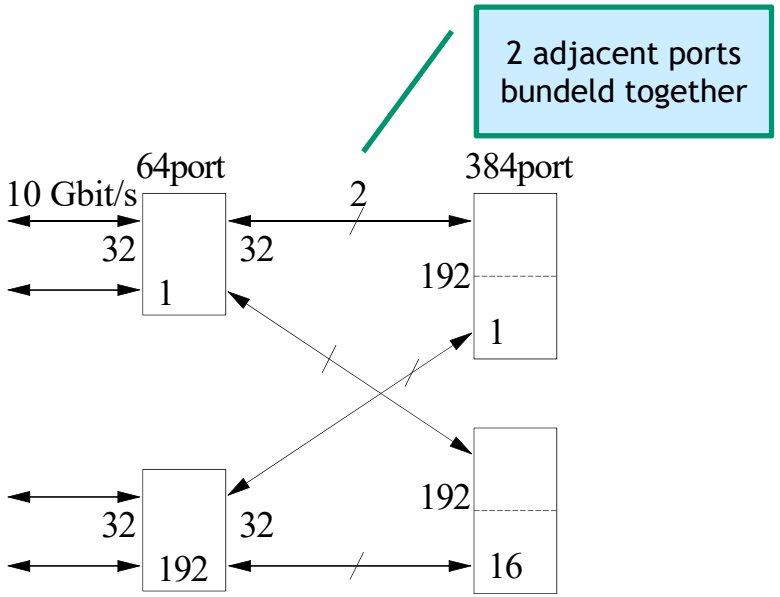


Example taken from “Cisco’s Massively Scalable Data Center”, 2009

Examples of DCN design

- 6144 servers
  - 6144 ports at 10 Gbit/s ⇒ 61.44 Tbit/s
    alternative designs
    - t=192 switches with 64 ports and k=32 switches with 192 ports
    - t=192 switches with 64 ports and k/2=16 switches with 384 ports

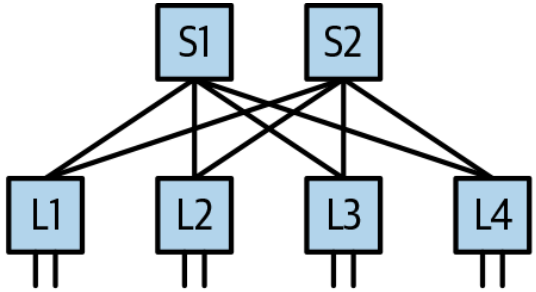




Example taken from “Cisco’s Massively Scalable Data Center”, 2009

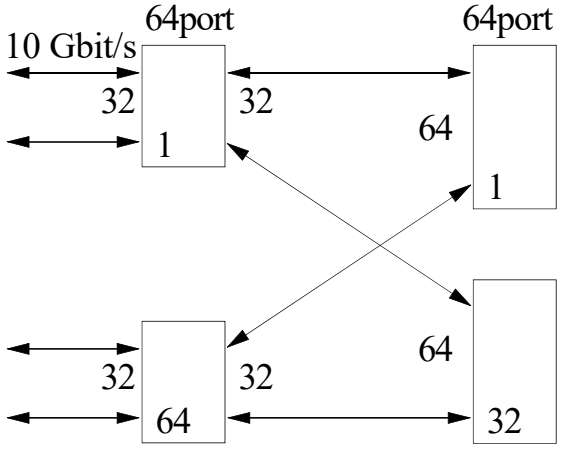
Examples of DCN design

- Interesting case: all switches have an equal number of ports 2k
  - Leaves: k x k switches (2k ports), k servers per TOR
  - Spine: 2k-port switches; # spine switches = k
  - → # leaves = t = 2k → $2k^2$ servers



(a) Two-tier Clos using four-port switches

- k = 2 → 8 servers
  - 8 ports at 10 Gbit/s ⇒ 80 Gbit/s
  - t=4 leaf switches with 4 ports and k=2 spine switches with 4 ports



- k = 32 → 2048 servers
  - 2048 ports at 10 Gbit/s ⇒ 20.48 Tbit/s
  - t=64 leaf switches with 64 ports and k=32 spine switches with 64 ports

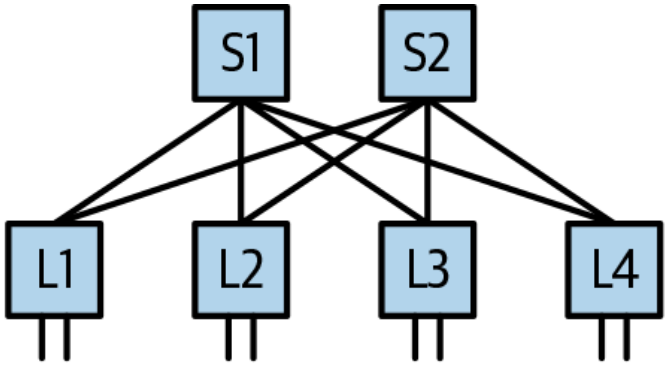
Advantages of Clos design in DNC

- Use of homogeneous equipment
- Routing as the *Fundamental Interconnect Model*
  - no Learning & Forwarding / no STP
  - Equal Cost Multipath (ECMP) strategy with a routing protocol (IS-IS, SPB, TRILL)
- Number of hops is the same for any pair of nodes
- Small blast radius

Scaling of Clos networks

- Can we scale to multi-tier designs?

Start with a two-tier network and add an additional row of switches

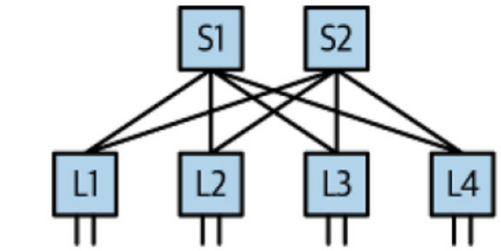


(a) Two-tier Clos using four-port switches

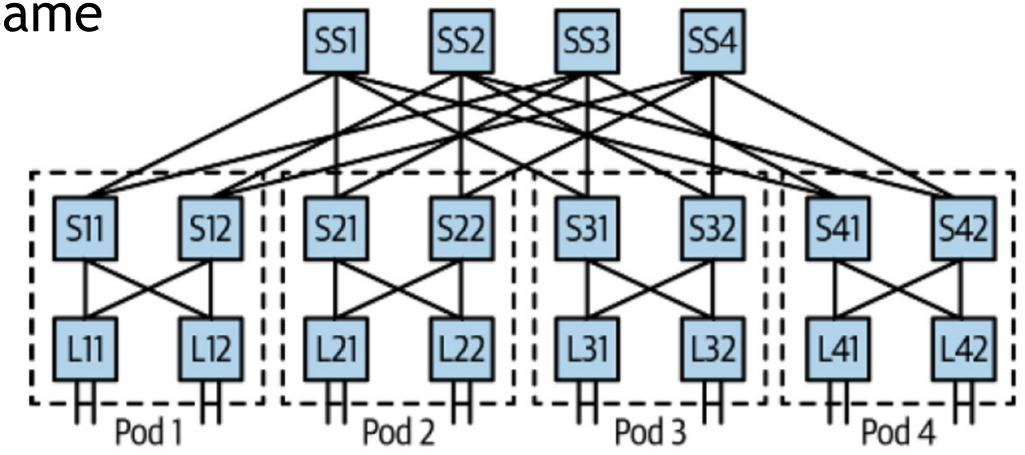
Scaling to a three-tier network

PoD-based model, aka the Fat Tree

- An option: transform each spine-leaf group into a «pod» and add a super-spine tier
- A highly scalable and cost-efficient DCN architecture that aims to maximize bisection bandwidth.
- It can be built using commodity Gigabit Ethernet switches with the same number of ports.
- Used by Microsoft, Amazon



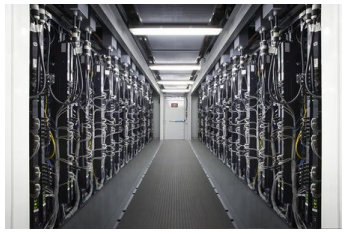
(a) Two-tier Clos using four-port switches

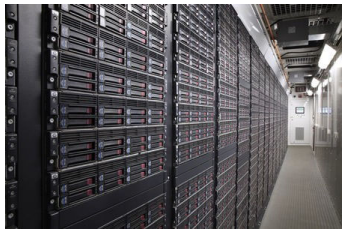


(c) Pod-based three-tier Clos using four-port switches

Point Of Delivery (POD)

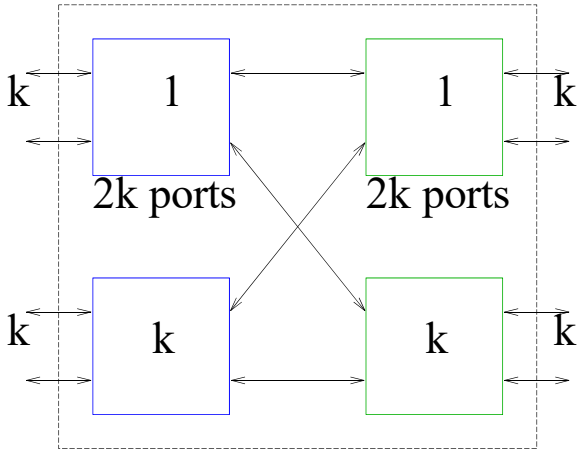
- A module or group of network, compute, storage, and application components that work together to deliver a network service
- The PoD is a repeatable pattern, and its components increase the modularity, scalability, and manageability of data
  - (taken from Wikipedia)





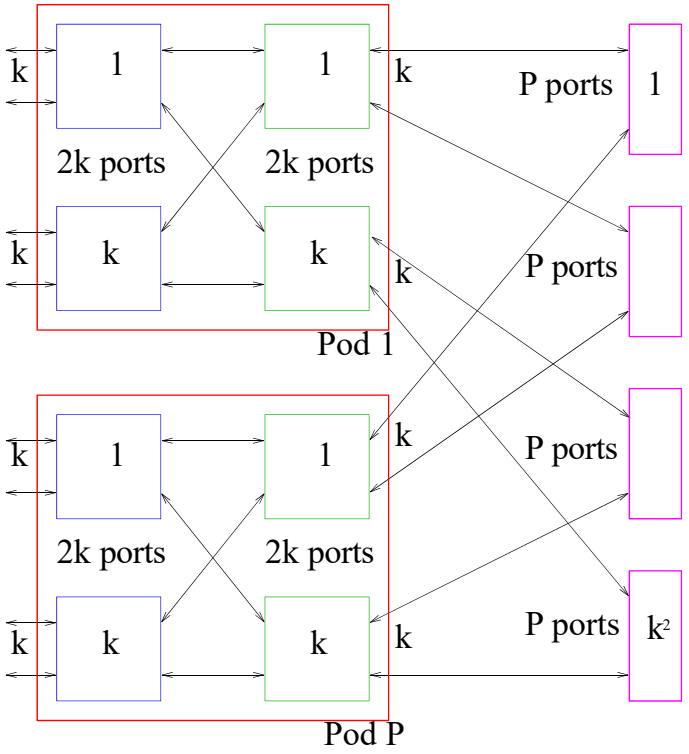
Leaf-spine PoD architecture

- Leaf with $2k^2$ bidirectional ports
  - k^2 ports to the servers and k^2 ports to the data center network
    - note that this cannot be used to interconnect directly $2k^2$ servers since the network would be blocking
  - built with 2k switches with 2k ports



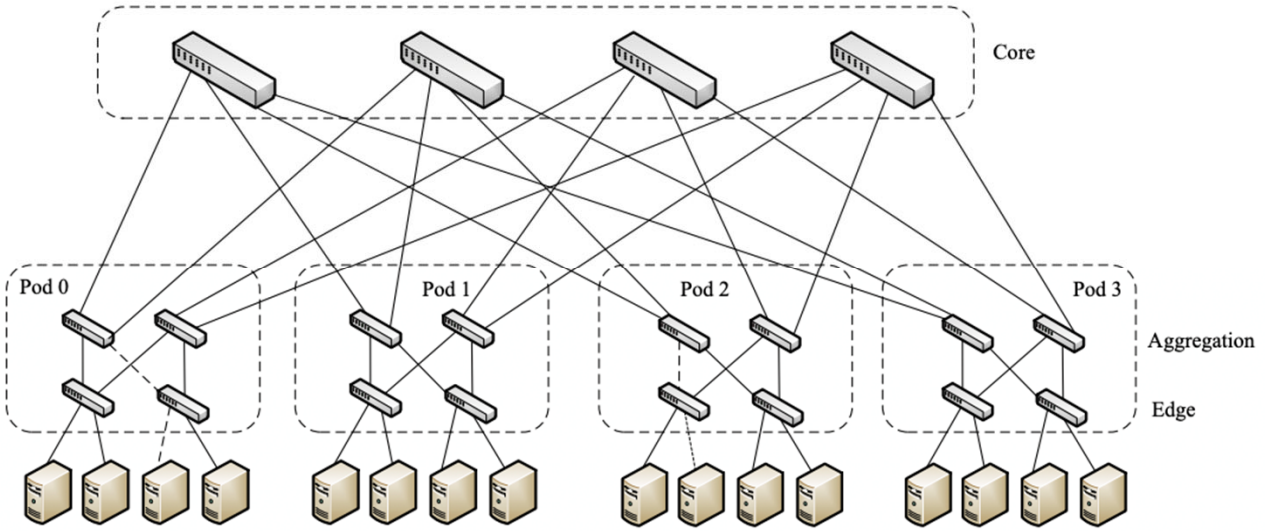
Intra-PoD and inter-PoD communications

- The PoD is replicated P times
- k^2P servers
  - 2kP switches with 2k ports
  - k^2 switches with P ports
- Fat-tree: choose P = 2k
  - $2k^3$ servers
  - $(k^2 + 2k \cdot 2k) = 5k^2$ switches with 2k ports



The Fat-Tree Network

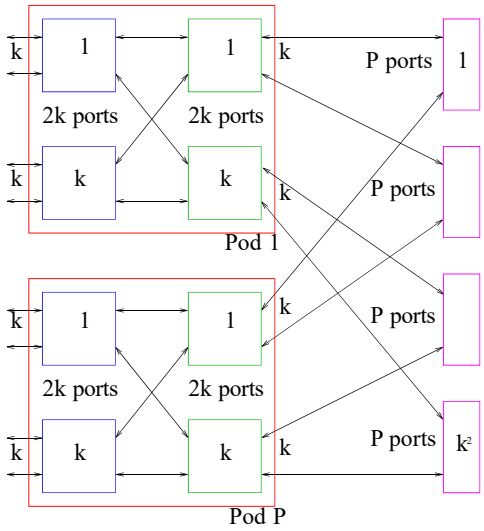
- At the edge layer, there are 2*k* PoDs (groups of servers), each with k^2 servers.
- Each edge switch is directly connected to k servers in a pod and k aggregation switches.
- A fat-tree network with 2*k*-port commodity switches can accommodate $2k^3$ servers in total
- k^2 core switches with 2*k*-port each, each one connected to 2*k* pods
- Each aggregation switch is connected to *k* core switches
  - Note the partial connectivity at switch level



k = 2
4 pods
16 servers
20 switches

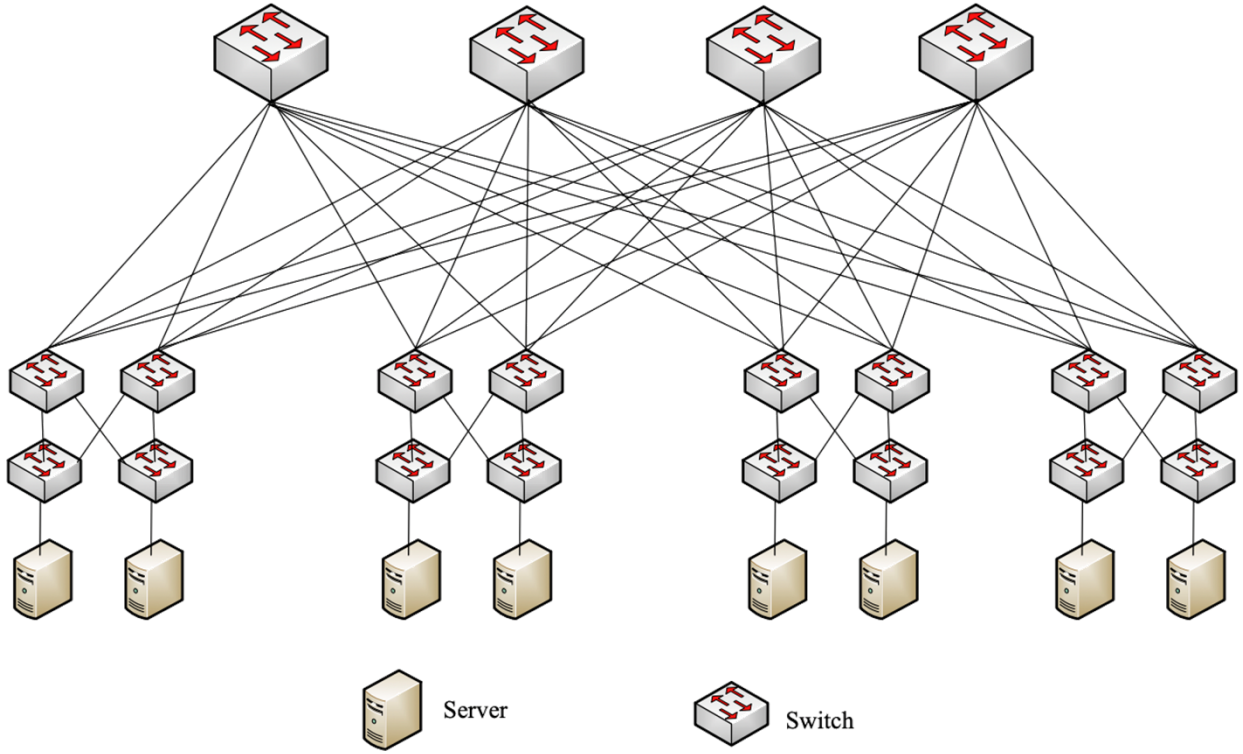
Fat tree example

- Data center with 65 536 servers in 64 PoDs
  - 65536 ports at 10 Gbit/s ⇒ 655 Tbit/s
  - P = 64 PoDs, k = 32
  - in total 5120 switches with 64 ports

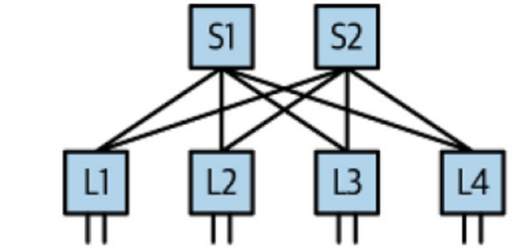


VL2 Network

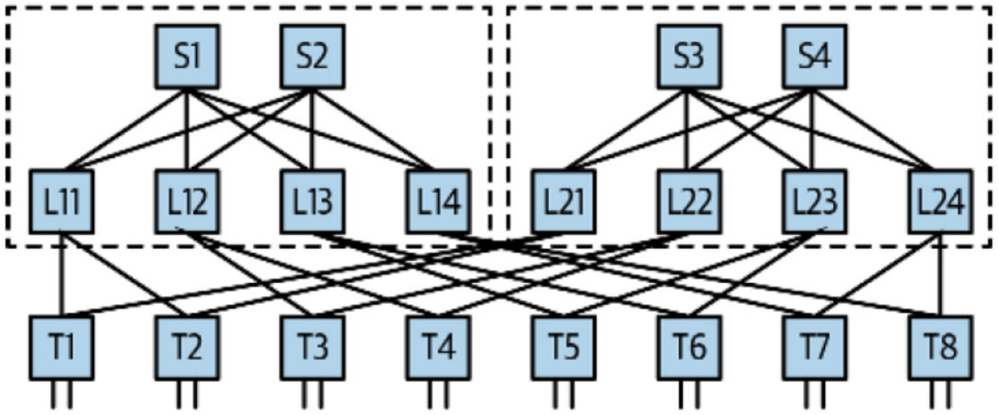
- A cost-effective hierarchical fat-tree-based DCN architecture with high bisection bandwidth.
- It uses three types of switches: intermediate, aggregation, and top-of-rack (ToR) switches.
- It uses $D_A/2$ intermediate switches, D_I aggregation switches and $D_A \cdot D_I/2$ ToR switches.
- The number of servers in a VL2 network is $20(D_A \cdot D_I)/4$
- It uses a load-balancing technique called valiant load balancing (VLB).



Virtual chassis model (e.g. Facebook)



(a) Two-tier Clos using four-port switches



(b) Virtual chassis-based three-tier Clos using four-port switches

Google scenario → Goolge Datacenter

- World-wide coverage with tens of sites
- Data center traffic
  - Bandwidth demand doubles every 12-15 months (faster than Internet)
    - larger datasets (photo/video content, logs, Internet-connected sensors, etc.)
    - web services
    - internal applications (index generation, web search, serving ads, etc.)
- Google Design approach
  - multistage Clos topologies on commodity switch silicon
  - centralized control
    - one configuration pushed to all the switches
    - SDN approach
  - modular hardware design with simple, robust software

A. Singh, et al., "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network", ACM SIGCOMM Computer Communication Reviews, Oct. 2015

Hardware alternative choices

- Commodity switches
  - cheap and simple, fast evolving
  - intermittent capacity
  - few available protocols, suitable for a single operator scenario
- WAN switches
  - complex and expensive, slow evolving
  - highest availability, but here intermittent capacity is allowed
  - many available protocols to support interoperability among multivendor WANs
- Google's choice
  - general-purpose, off-the-shelf switch components for commodity switches
  - 5 generations in the period 2004-2015

Google' Jupiter Datacenter

- Jupiter is the latest generation of operational data centers
- Clos topology
  - basic switching module with small number of ports scale to any size
    - limited by the control plane scalability path diversity and redundancy complexity of multiple equal cost paths
  - complex fiber interconnection
- Design
  - multistage, recursively factorized Clos network
  - basic block: 16 ports @ 40 Gbit/s, but each interface can be split in 4 ports @ 10 Gbit/s
    - e.g. 48 ports @ 10 Gbit/s plus 4 ports @ 40 Gbit/s
  - overall bisection bandwidth: 512 × 64 × 40 = 1 310 720 Gbit/s
  - (1.3 Pbit/s)
  - server: 192 × 32 × 64 = 393 216 servers @ 10 Gbit/s
  - 3:1 oversubscription ratio between server capacity and network capacity

Jupiter topology

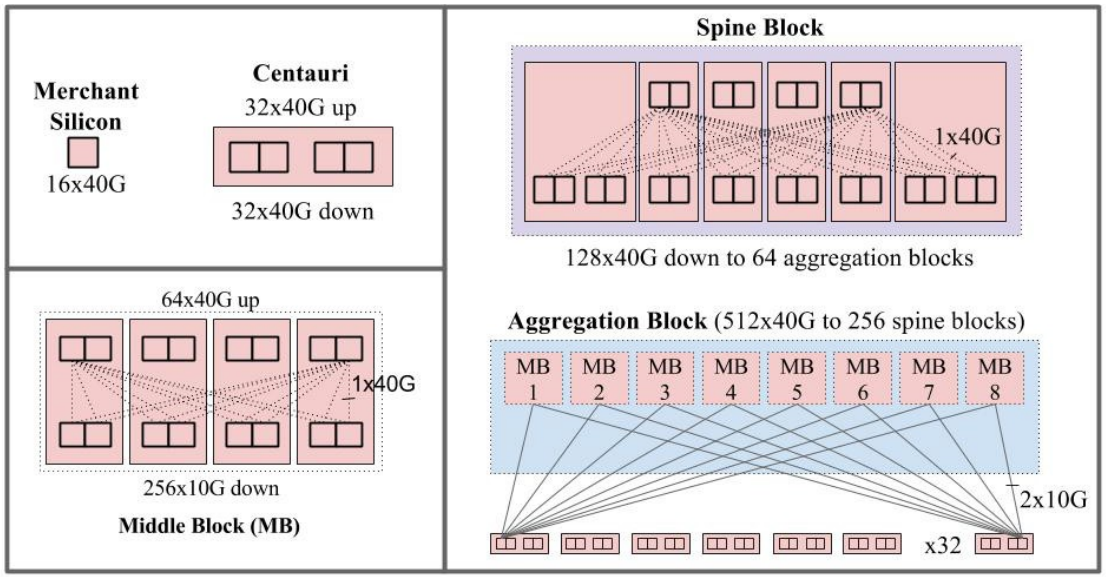


Figure 13: Building blocks used in the Jupiter topology.

Figures reproduced from [*Google*].

Figure 14: Jupiter Middle blocks housed in racks.

Connecting Jupiter to Internet

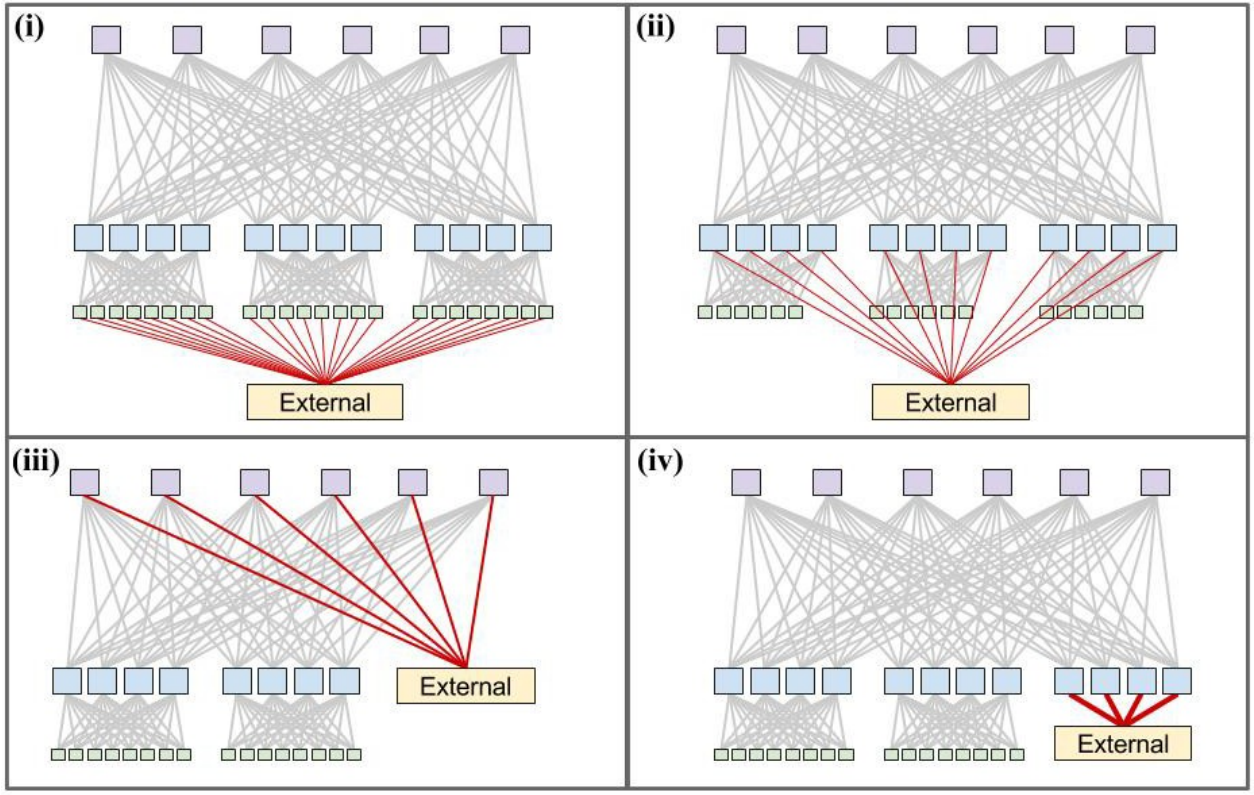


Figure 15: Four options to connect to the external network layer.

Figures reproduced from [*Google*].

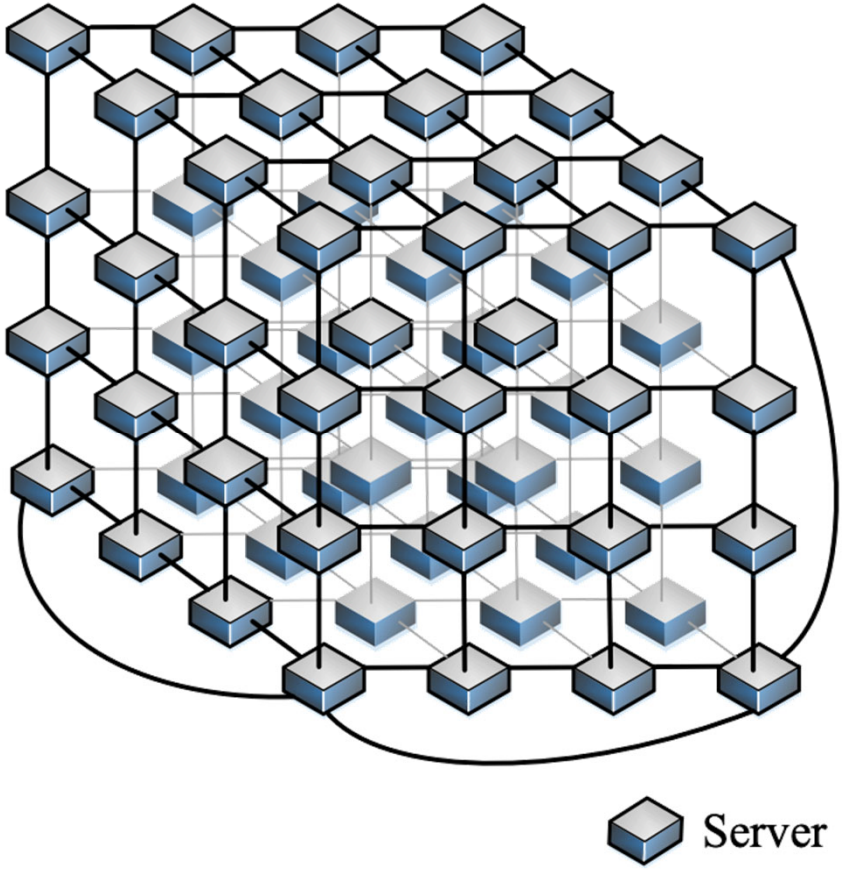
Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures

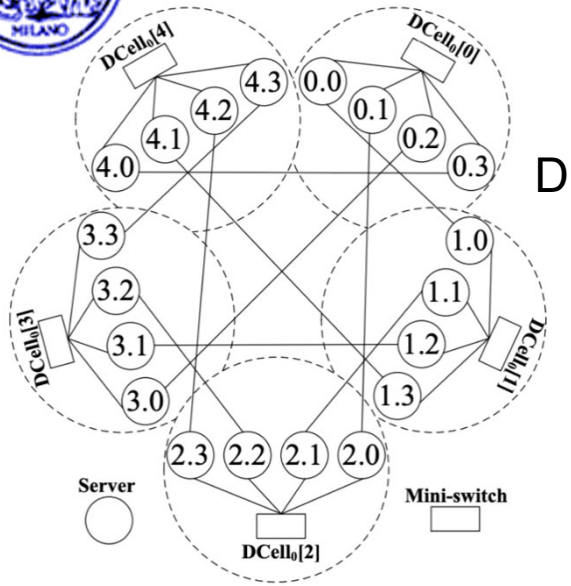
Server-centric architectures
CamCube

- A server-centric architecture proposed for building container-sized data centers.
- It may reduce implementation and maintenance costs by using only servers to build the DCN.
- It uses a 3D-Torus topology to interconnect the servers directly.
- As a torus-based architecture, it exploits network locality to increase communication efficiency.
- Drawbacks: CamCube requires servers with multiple NICs to assemble a 3D Tours network, long paths, and high routing complexity

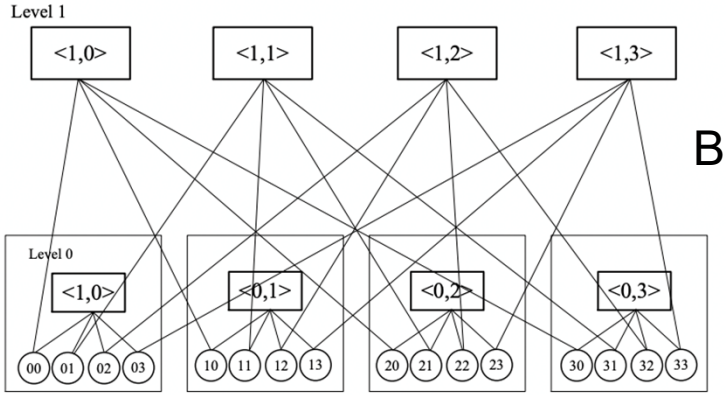
3D Torus with 64 servers



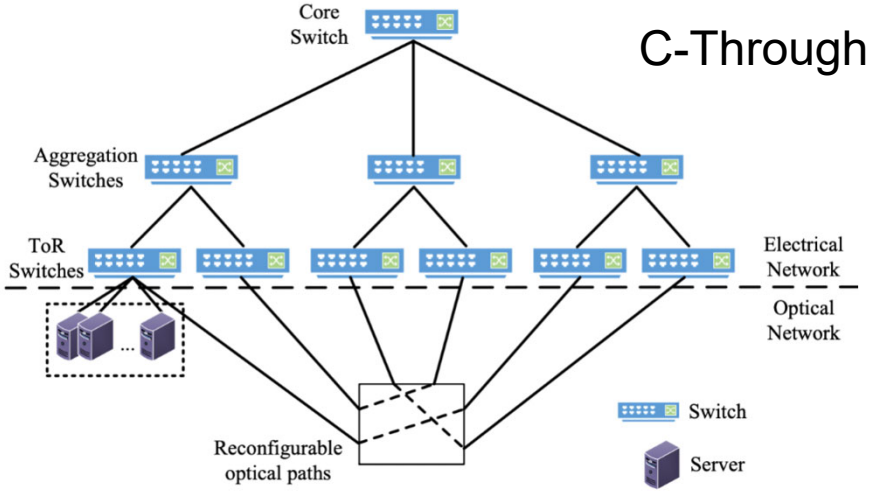
Hybrid architectures



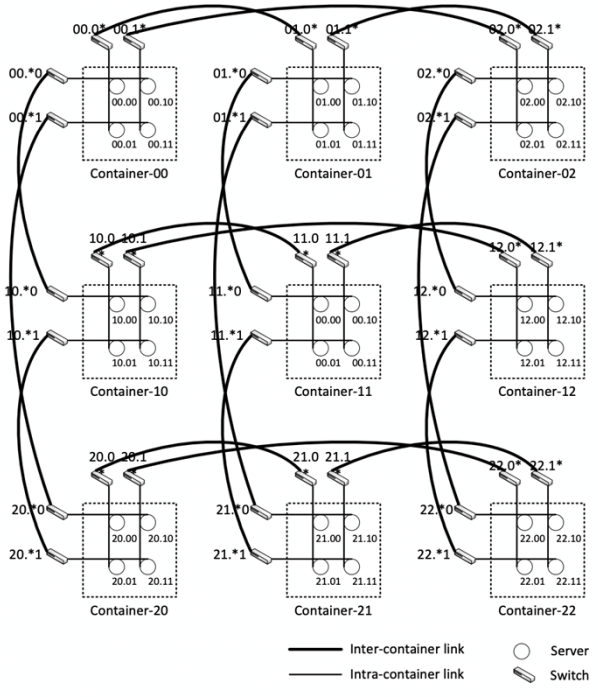
DCell



BCube



C-Through



MDCube

Addressing and Routing in DCN

- Scenario
  - 100,000 servers, 32 VM each → $3 \cdot 10^6$ MAC and IPs
- Addressing and routing schemes challenging
  - standard schemes do not provide efficient solutions
- Specific solutions for DCN have been developed
  - At Layer-2:
    - Overcoming of classical Ethernet-based Spanning Tree-based routing to exploit equal-cost multipaths
      - Rapid and Multiple Spanning Tree Protocols (RSTP), Multi-Chassis Link-Aggregation (MC-LAG)
  - At Layer-3:
    - Overcoming the complexity of classical Interior Gateway routing Protocols (e.g., OSPF) and supporting VM migration
      - FabricPath, TRILL, NVGRE, OTV, Shortest Path Bridging (SPB), Segment Routing (SRv6), + DCN-oriented BGP extensions
- A complex topic that would need a dedicated course…

Thanks!

- guido.maier@polimi.it