Data center network topologies

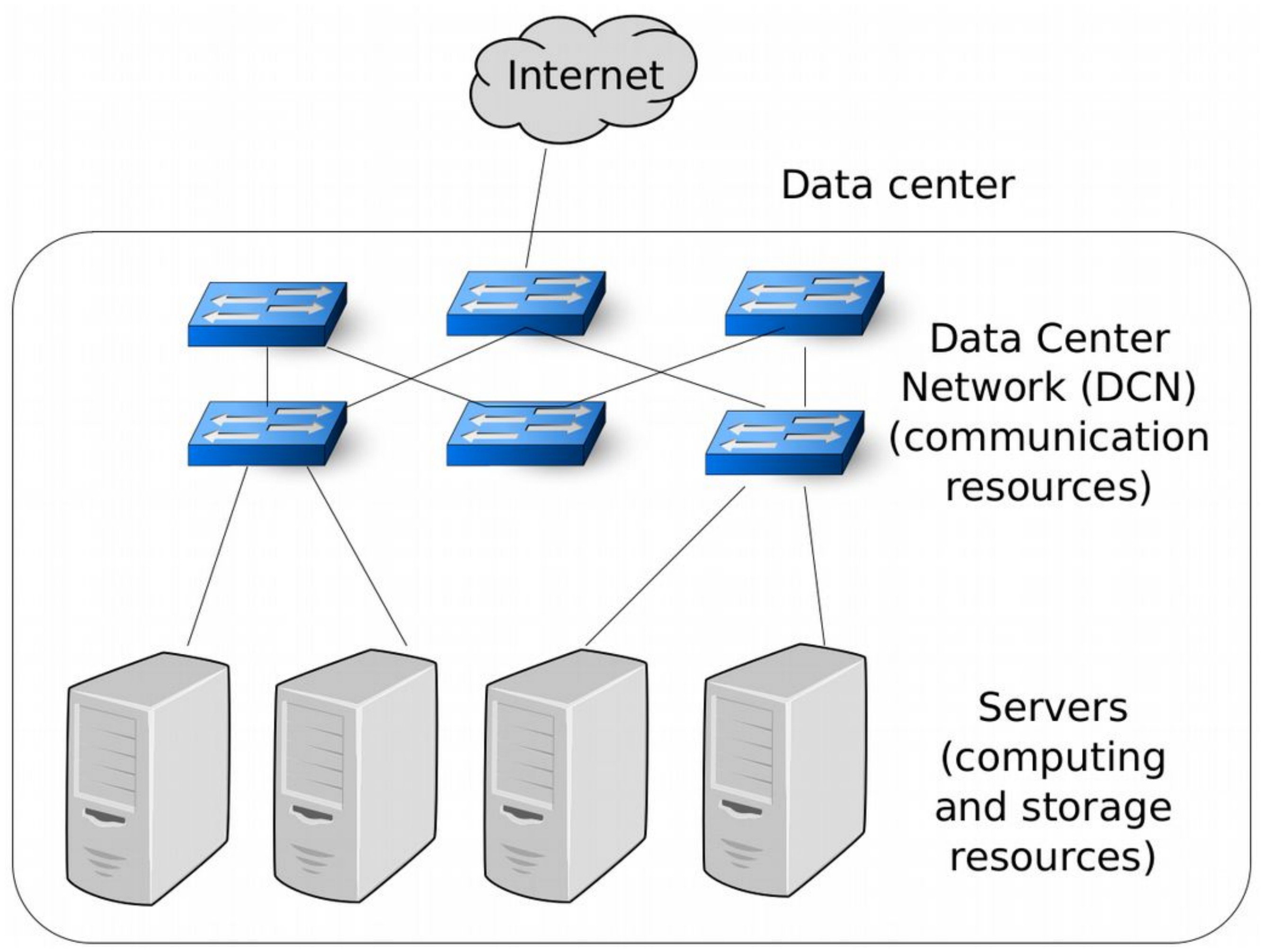
Eduardo Grampín Castro
grampin@fing.edu.uy
Universidad de la República
Uruguay

based on several academic and industrial presentations

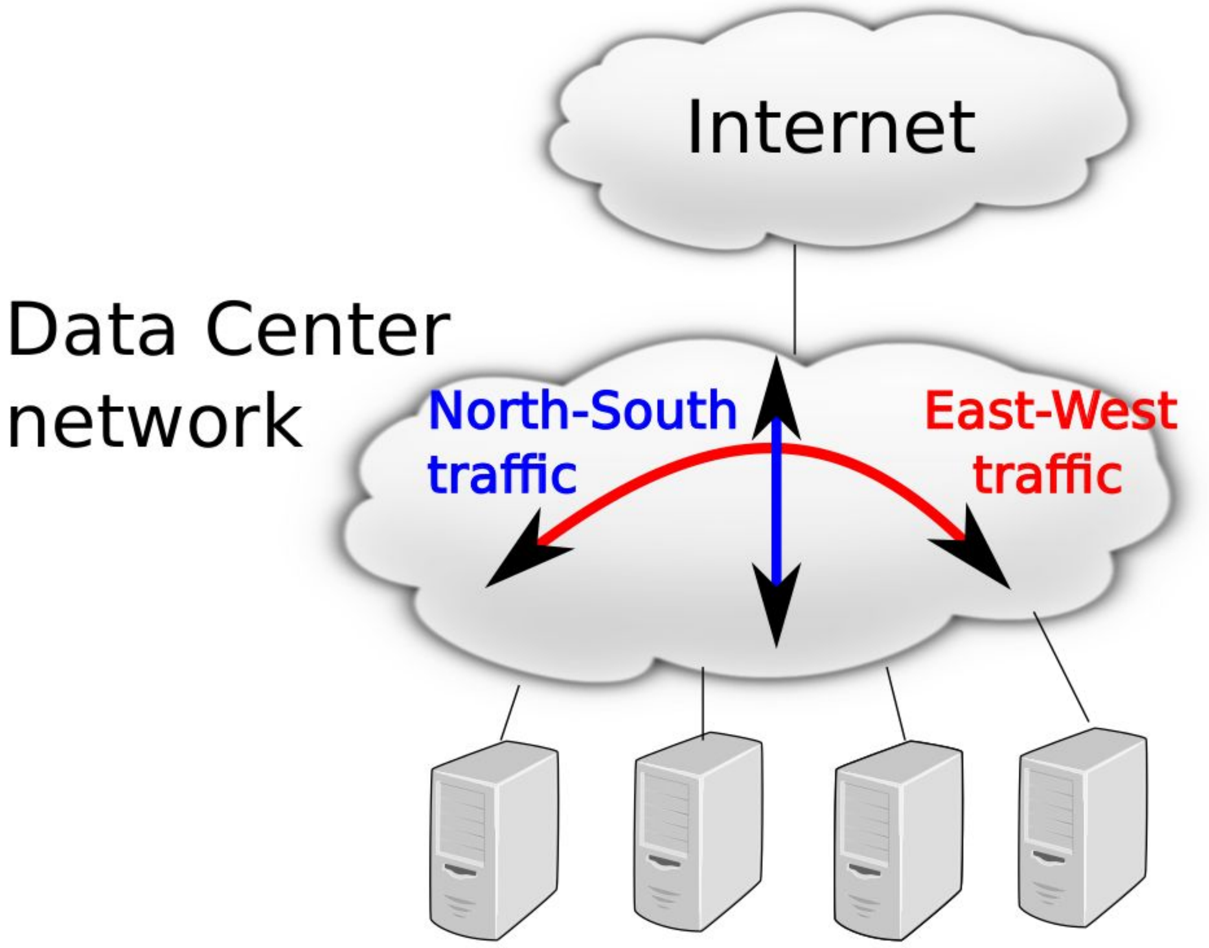
Outline

- The legacy data center infrastructure
- Fat-Tree/Clos topologies
- Routing in the fat-tree
- Real world examples

Logical view of a data center


Internet
Data center
Data Center Network (DCN) (communication resources)
Servers (computing and storage resources)

Data traffic within a data center



Intra-data center 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

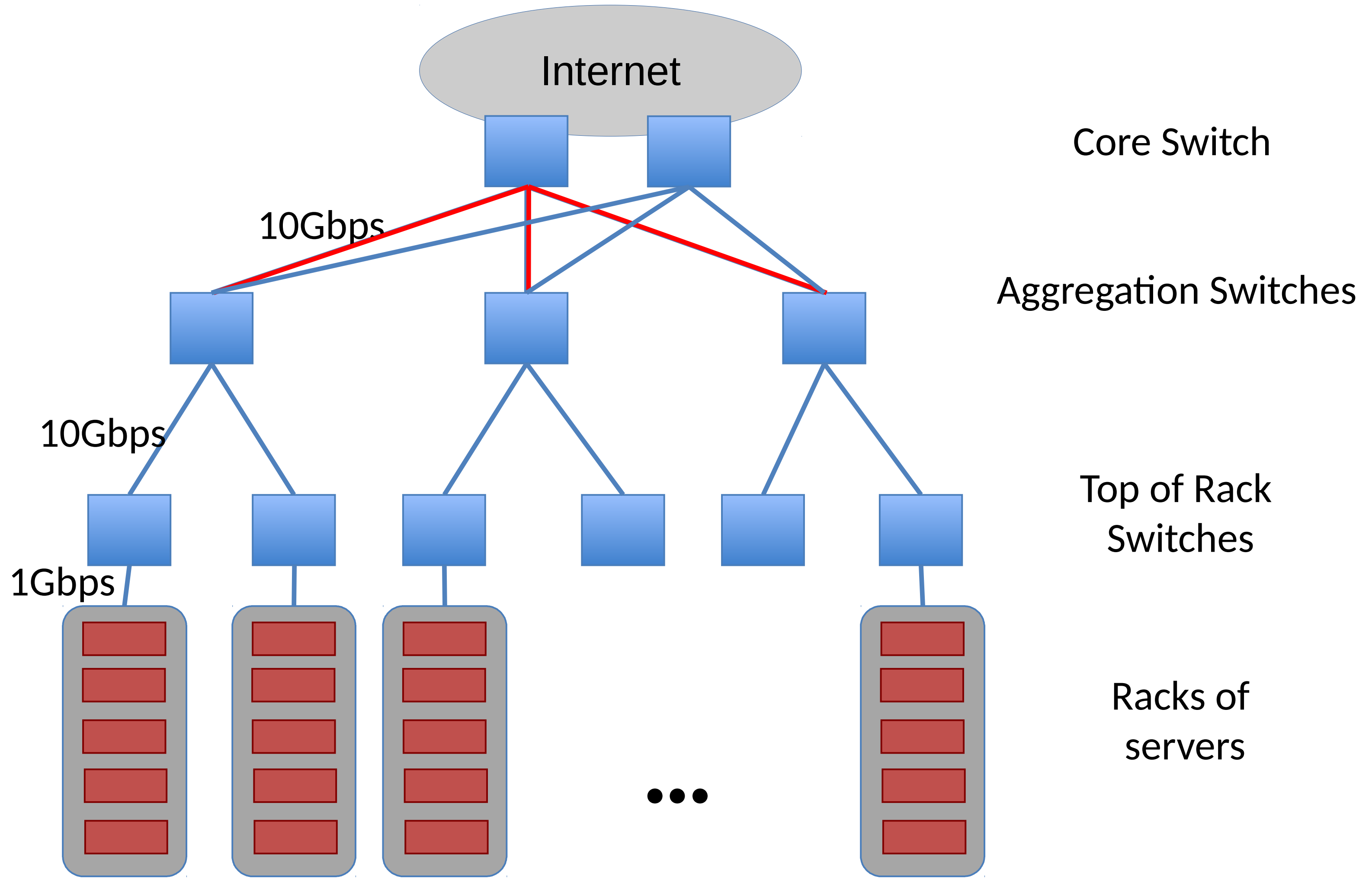
“Overall, east-west traffic (traffic within the data center and traffic between data centers) will represent 85 percent of total data center by 2021, and north-south traffic (traffic exiting the data center to the Internet or WAN) will be only 15 percent of traffic associated with data centers.”

Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper

Intra-data center 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

Traditional Data Center Topology



Traditional Data Center Topology

- 20-40 servers per rack
- Each server connected to 2 access switches (1 to 10 Gbps)
- Access switches connect to 2 aggregation switches
- Aggregation switches connect to 2 core routers
- Core routers connect to edge routers
- Aggregation layer is the transition point between L2-switched access layer and L3-routed core layer

Traditional Data Center Topology

- Core routers manage traffic between aggregation routers and in/out of data center
- All switches below each pair of aggregation switches form a single Layer 2 domain
- In this architecture, each Layer 2 domain typically limited to a few hundred servers to limit broadcast
- Remember: most traffic is internal to the data center
  - Network is the bottleneck.
  - Uplinks utilization of 80% is common

Traditional Data Center Topology

- Higher layers oversubscribed:
  - Other servers in the same rack 1:1
  - Uplinks from ToR: 1:2 to 1:20
    - (e.g., 32x10Gb down, 8X10Gb up = 4:1 oversubscription)
  - Core Routers: 1:240
    - Generally keep services in one tree
    - Can't arbitrarily move servers
- Moving across subnets is painful
  - Requires reconfiguration of IP addresses and VLAN trunks
- Services walk over each-other
  - Overuse by one service affects others
- Poor reliability
  - One access switch failure doubles the load on the other

Traditional Data Center Topology

- Under-utilization
  - Even when multiple paths exist only one is used
- ECMP (Equal Cost Multipath) is used by routers to spread traffic to next hops using a hash function. However, only 2 paths exist
- Moreover, traditional switches/routers are expensive!

Requirements for data center networks

- Backwards compatible with existing infrastructure
  - No changes in application
  - Support of layer 2 (Ethernet)
- Cost effective
  - Low power consumption & heat emission
  - Cheap infrastructure
- Host communication at line speed
  - Full-bisection bandwidth
  - No over-subscription
- Fault tolerant

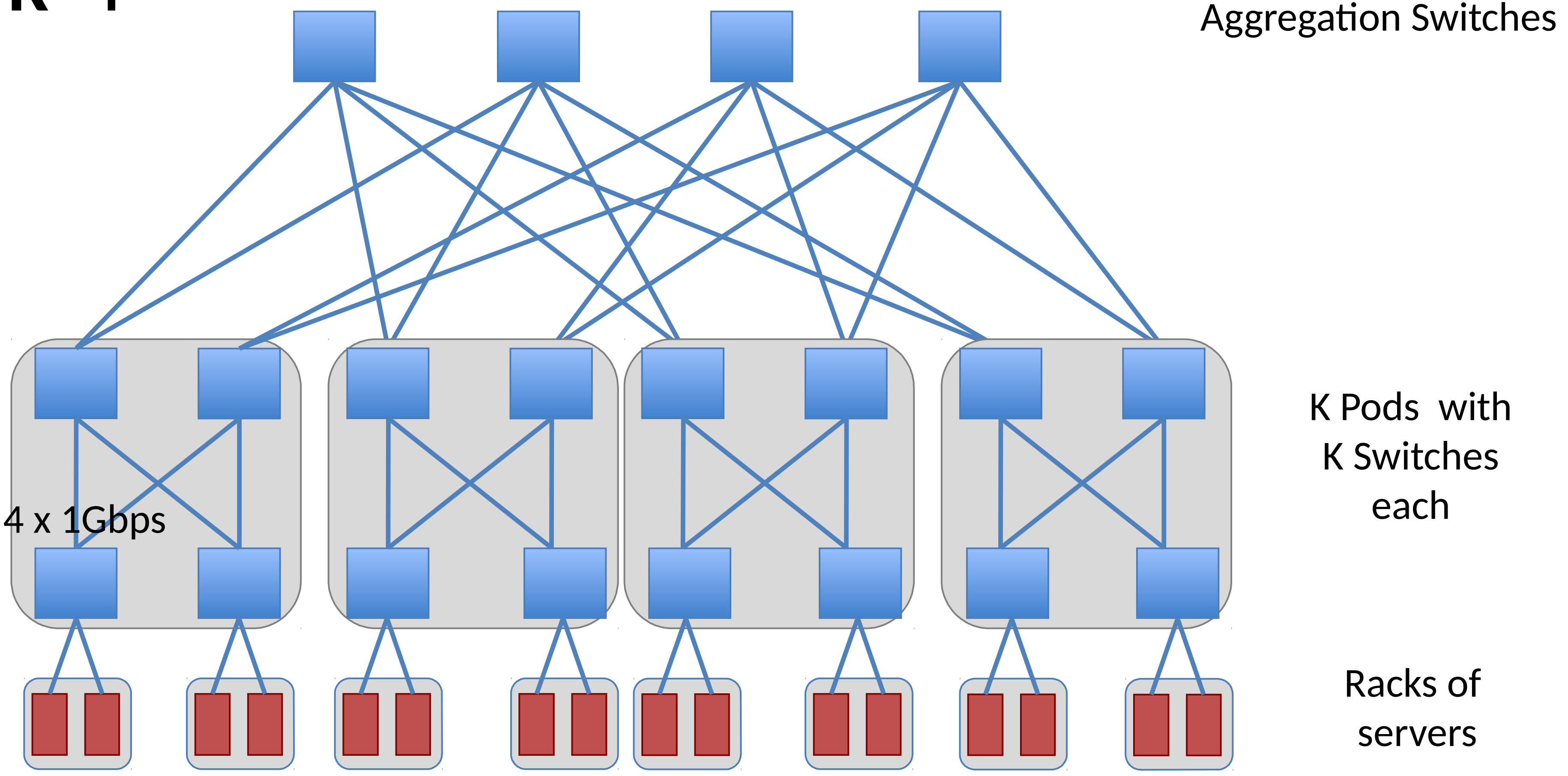
The Fat Tree

[Al Fares et al, Sigcomm2008, inspired by Clos, 1953]

- Inspired from the telephone networks of the 50’s -> Clos networks
- Uses cheap, commodity switches -> all switches are the same
- Lots of redundancy
- Single parameter to describe the topology:
  - **k** -> the number of ports in a switch

Fat Tree topology

K=4



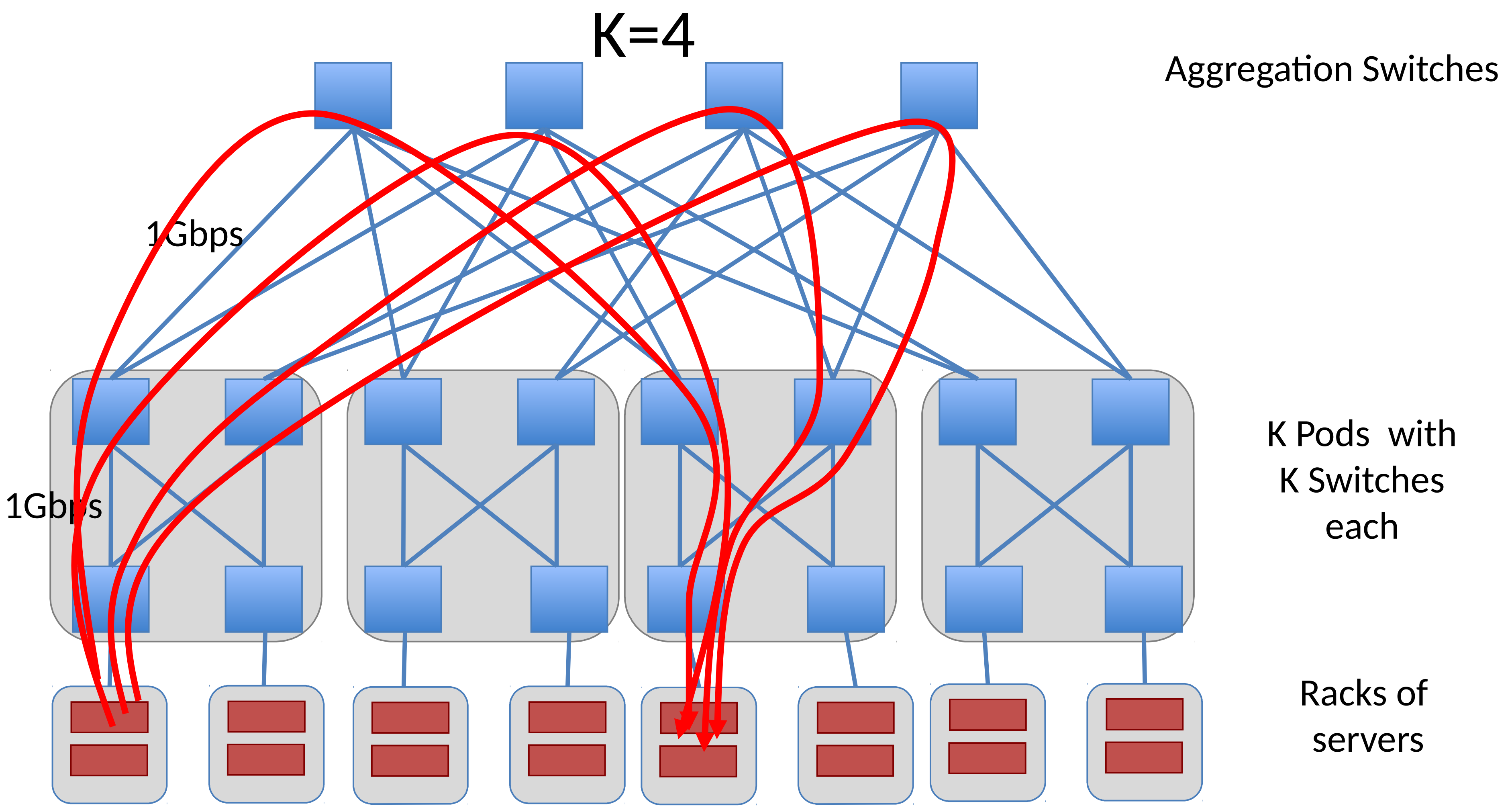
Fat Tree topology

- Inter-connect racks (of servers) using a fat-tree topology
- 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

Fat Tree topology

- Why 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
  - All devices can transmit at line speed if packets are distributed uniform along available paths
- Great scalability: k-port switch supports $k^3/4$ servers

The Fat Tree Topology has k*k/4 paths between any two endpoints



Addressing and routing

- Usual scenario: 100000 servers, >30 VMs per server → 3 million IP addresses!

- Layer-2 addressing
  - One single LAN
  - Drawbacks
    - very large forwarding tables in switches
    - lots of broadcast (BUM) traffic (e.g. ARP)
    - STP avoids loops but waste resources; even if a multi-path L2 solution would exist, BUM traffic is a problem
  - Layer 2 still might be desirable, though
    - Some apps expect servers in the same LAN

Addressing and routing

- Layer-3 addressing
  - One subnet per VLAN
  - Drawbacks
    - many DHCP servers and VLANs
    - very large number of switches and routers (around 10,000)
    - IGP sacalability?
  - Canonical VM migration
    - when changing LAN, a new IP address is required and existing TCP connections break

Addressing and routing

- Multipath Routing at Layer 3
  - Run a link-state routing protocol on the switches (routers) (e.g. OSPF)
  - Compute shortest-path to any destination
  - Drawback: must use smarter, more expensive switches!
- Equal Cost Multipath Routing (ECMP)
  - shortest path IP routing will typically use only one path despite the path diversity in the topology
  - if using equal-cost multi-path routing at each switch independently and blindly, packet re-ordering may occur; further load may not necessarily be well-balanced
- Aside: control plane flooding!

Addressing and routing

- Practical solutions
  - VXLAN/EVPN
  - LISP
    - provides IP address mobility across layer-3 subnets
  - Proprietary schemas
- Other solutions: FabricPath, TRILL, NVGRE, …
- Still, underlying routing protocol with good ECMP support is needed!

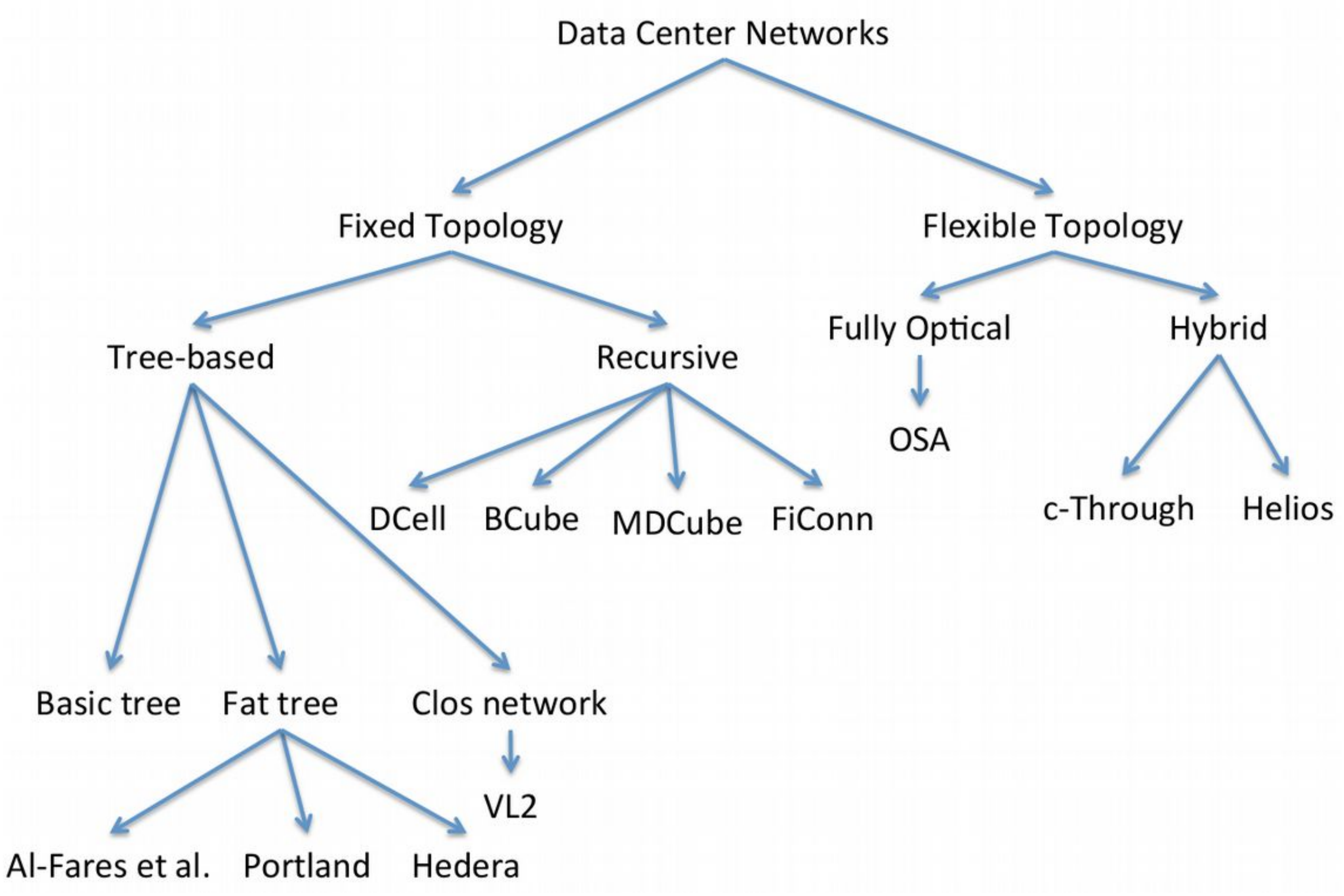
Routing proposals

- RFC7938 Use of BGP for Routing in Large-Scale Data Centers. P. Lapukhov, A. Premji, J. Mitchell, Ed.. August 2016. (Status: INFORMATIONAL) (DOI: 10.17487/RFC7938)
- Link State Vector Routing (lsvr). https://datatracker.ietf.org/wg/lsvr/about/
- Routing In Fat Trees (rift). https://datatracker.ietf.org/wg/rift/about/

Discussion

- Bandwidth is the scalability bottleneck in large scale clusters
- Existing solutions are expensive and limit cluster size
- Fat-tree topology with scalable routing and backward compatibility with TCP/IP and Ethernet
- Large number of commodity switches have the potential of displacing high end switches in DC the same way clusters of commodity PCs have displaced supercomputers for high end computing environments

Taxonomy of solutions



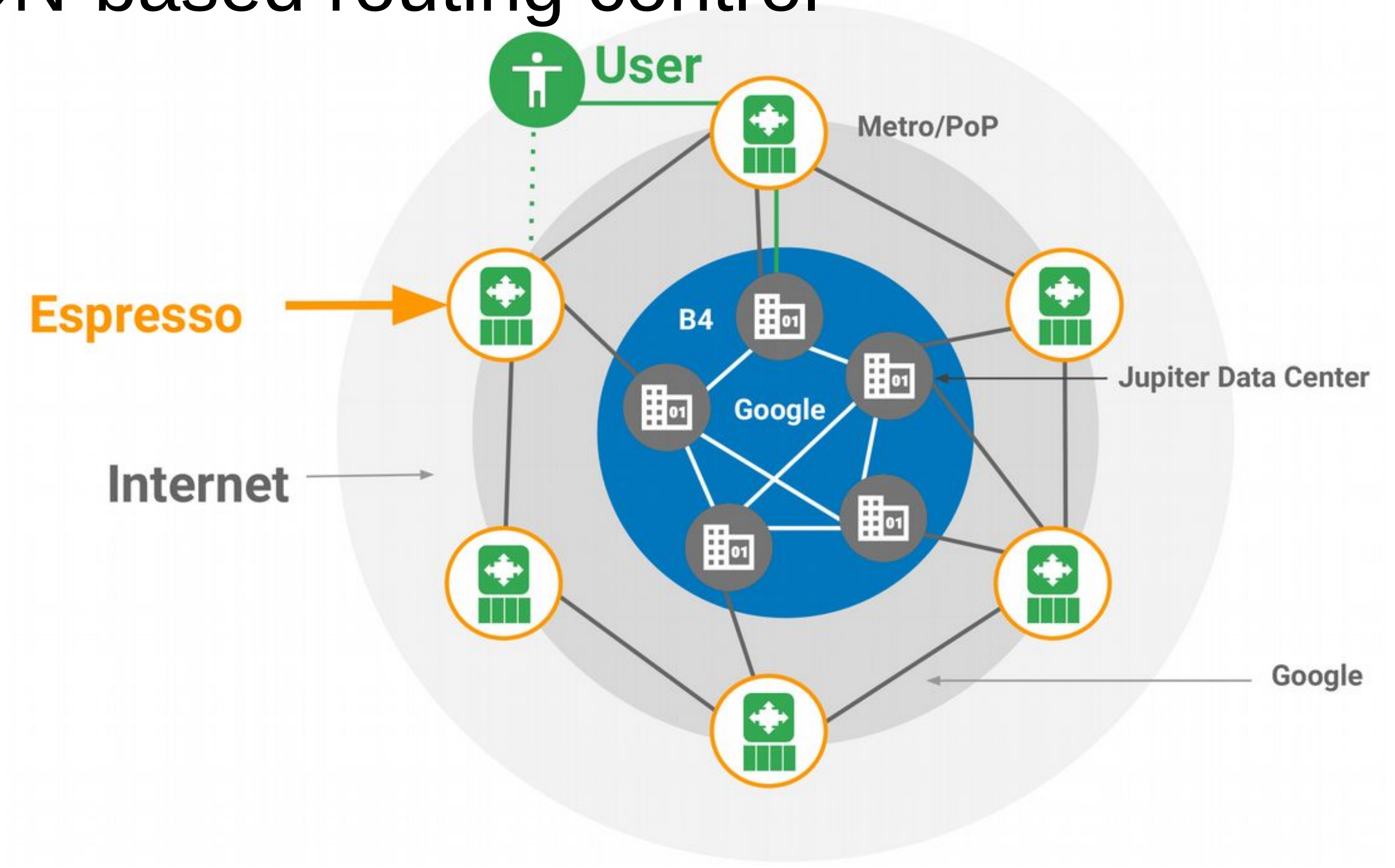
Some proposals

- A Scalable, Commodity Data Center Network Architecture
  - a new Fat-tree inter-connection structure (topology) to increase bi-section bandwidth
  - needs new addressing, forwarding/routing
    - Topological/geographical numbering
    - Two stages forwardings lookup
- VL2: A Scalable and Flexible Data Center Network
  - Link state ECMP among switches
  - Valiant Load Balancing (random) among servers
  - Naming and location (addressing) separation with directories
- Other Approaches:
  - PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric
  - BCube: A High-Performance, Server-centric Network Architecture for Modular Data Centers
- Hybrid optical data center network architectures

Real world examples

Google

- Clos-based evolving architecture
- Home-made hardware
- SDN-based routing control



Google

- Four components of its SDN strategy:
  - Jupiter: Google employed SDN principles to build Jupiter, a data center interconnect capable of supporting more than 100,000 servers. As of 2013 it supports more than 1 Pb/s of total bandwidth to host its services.
  - B4 WAN interconnect: Google constructed B4 to connect its data centers to one another to replicate data in real-time between individual campuses. Built on white boxes with proprietary software.
  - Andromeda: network functions virtualization (NFV) stack that allows Google to deliver the same capabilities available to its native applications all the way to containers and virtual machines running on the Google Cloud Platform.
  - Espresso extends SDN to the peering edge of Google’s network where it connects to other networks across the planet. The Espresso technology allows Google to dynamically choose from where to serve individual users based on measurements of how network connections are performing in real time.

Google

- Chi-Yao Hong et al., "B4 and after: managing hierarchy, partitioning, and asymmetry for availability and scale in google's software-defined WAN," In Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication, Pages 74-87

- M. Dalton et al., "Andromeda: Performance, Isolation, and Velocity at Scale in Cloud Network Virtualization," In Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI ’18)

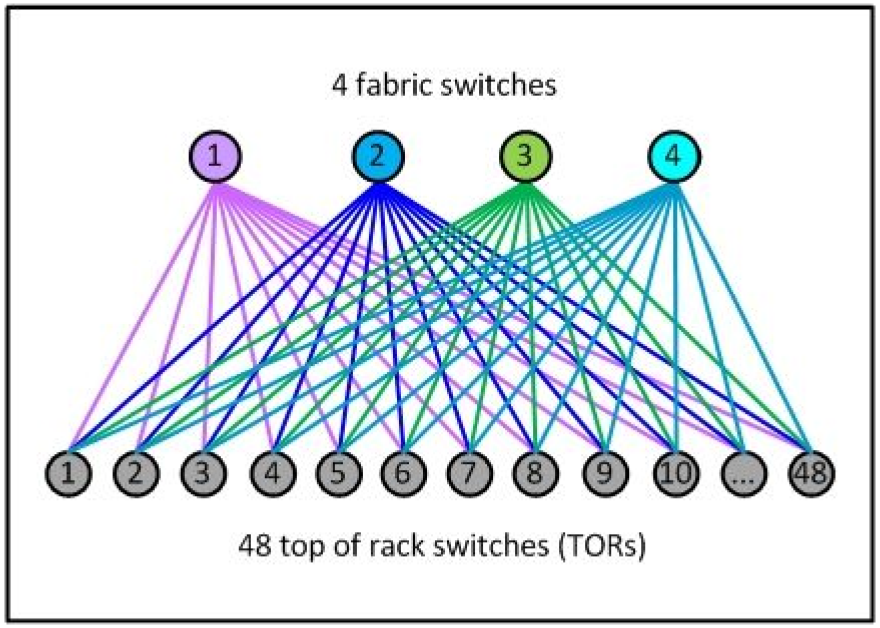
- Kok-Kiong Yap et al. "Taking the Edge off with Espresso: Scale, Reliability and Programmability for Global Internet Peering," In Proceedings SIGCOMM '17 Conference of the ACM Special Interest Group on Data Communication, Pages 432-445

- Arjun Singh et al. "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," In Proceedings of the ACM SIGCOMM 2015 Conference on Data Communication, Pages 183-197

- S. Jain et al., "B4: Experience with a Globally-deployed Software Defined Wan," In Proceedings of the ACM SIGCOMM 2013 Conference, pages 3-14

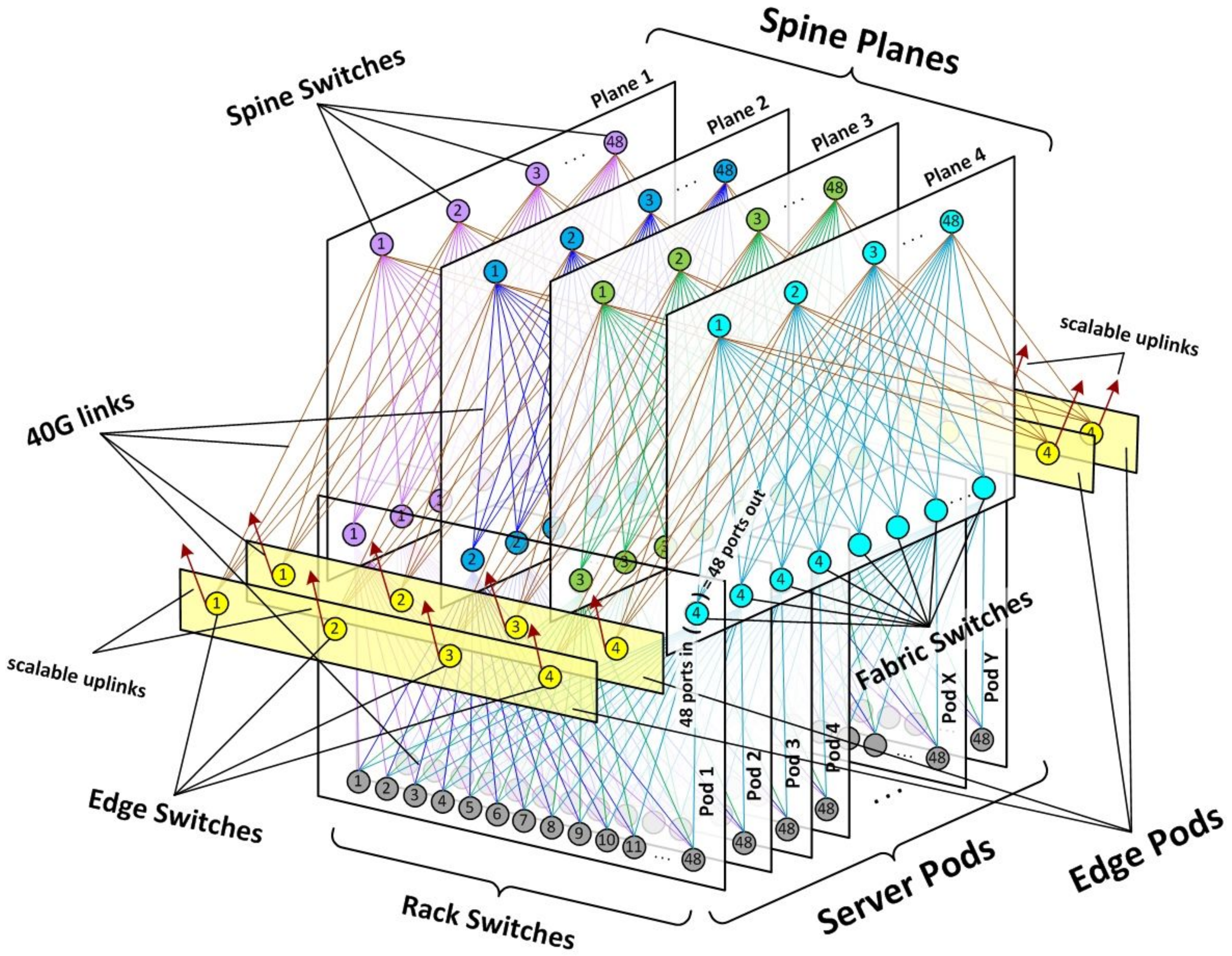
Facebook

- Data center fabric



A sample pod – network unit

Facebook



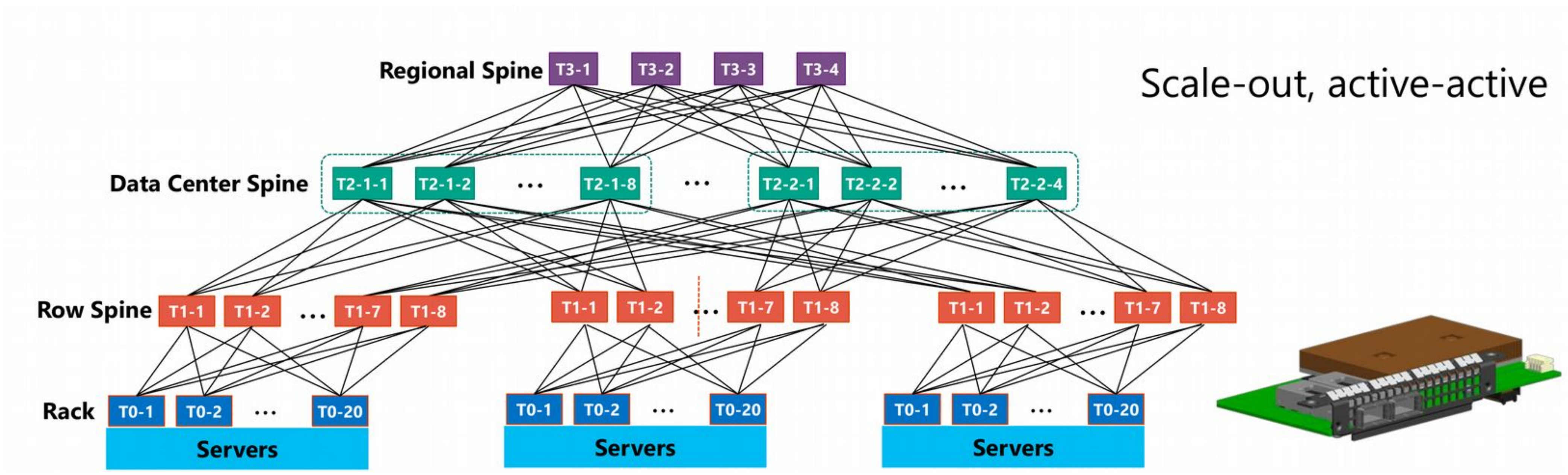
Schematic of Facebook data center fabric network topology

Facebook

- Introducing data center fabric, the next-generation Facebook data center network. Online: https://code.fb.com/production-engineering/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/
  - Video: https://youtu.be/mLEawo6OzFM
- Open/R: Open routing for modern networks. Online: https://code.fb.com/connectivity/open-r-open-routing-for-modern-networks/
  - Code online: https://github.com/facebook/openr
- Open Compute Project (OCP). Online: http://www.opencompute.org/

Microsoft

- VL2 Clos network topology
  - Home-made NICs (Azure SmartNIC)
  - Azure Cloud Switch open Switch Abstraction Interface (SAI): https://github.com/opencomputeproject/sai
  - SDN control/management



Play time

https://clusterdesign.org/fat-trees/

References

[Torino2018] Paolo Giaccone, The design of data center networks. Notes for the class on “Switching technologies for data centers”. Politecnico di Torino, November 2018. Online: https://www.telematica.polito.it/course/switching-technologies-for-data-centers/

[Cornell2014] Hakim Weatherspoon, Data Center Network Topologies: FatTree. CS 5413: High Performance Systems and Networking, Cornell University, 2014. Online: https://www.cs.cornell.edu/courses/cs5413/2014fa/lectures/08-fattree.pdf

[RaiciuRo] Costin Raiciu, Datacenter Network Topologies. Computer Science and Engineering Department of University Politehnica of Bucharest. Online: http://andrei.clubcisco.ro/cursuri/5master/aac-atds/lecture3-dcnet.pptx

[FatTree2009] M. Al-Fares et al., "A scalable, commodity data center network architecture," In Proceedings of the ACM SIGCOMM 2008 conference on Data communication. Pages 63-74

[VL2-2009] A. Greenberg et al., "VL2: a scalable and flexible data center network," In Proceedings of the ACM SIGCOMM 2009 conference on Data communication, Pages 51-62

[Clos1953] Charles Clos. "A Study of Non-Blocking Switching Networks," First published: March 1953. https://doi.org/10.1002/j.1538-7305.1953.tb01433.x