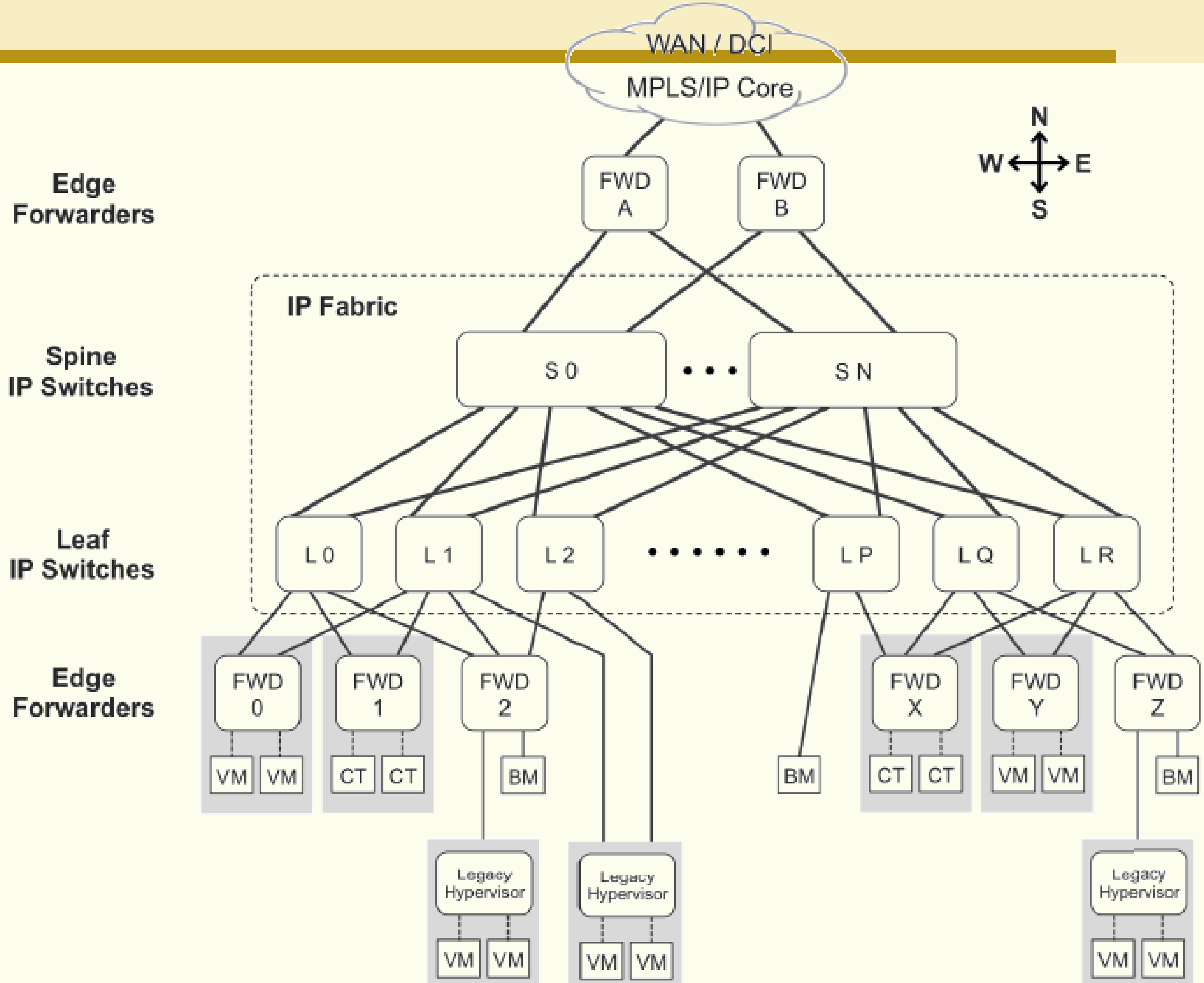
Data Center Networks

Enzo Mingozzi
Professor @ University of Pisa
enzo.mingozzi@unipi.it

Data Center Networks

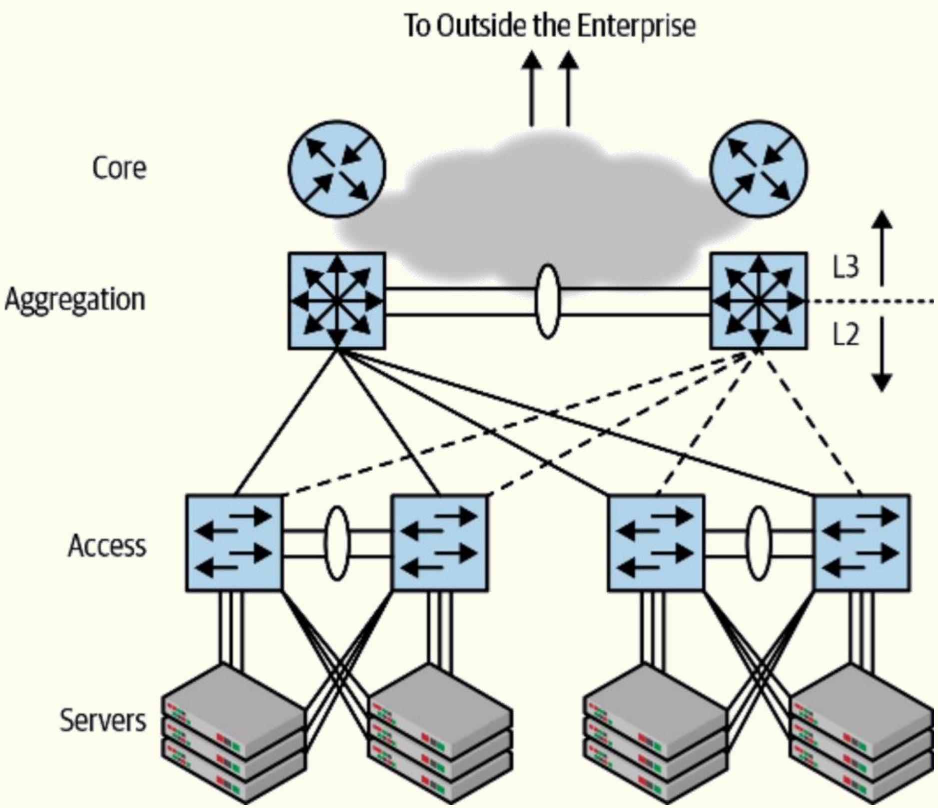


Modern DCN requirements

- **Increased server-to-server communication**
  - modern data center applications involve a lot of server-to-server communication (east-west)
- **Scale**
  - modern data centers range from a few hundred to a hundred thousand servers in a single physical location
- **Resilience**
  - The primary aim is to limit the effect of a failure to as small a footprint as possible

Legacy DCN topologies

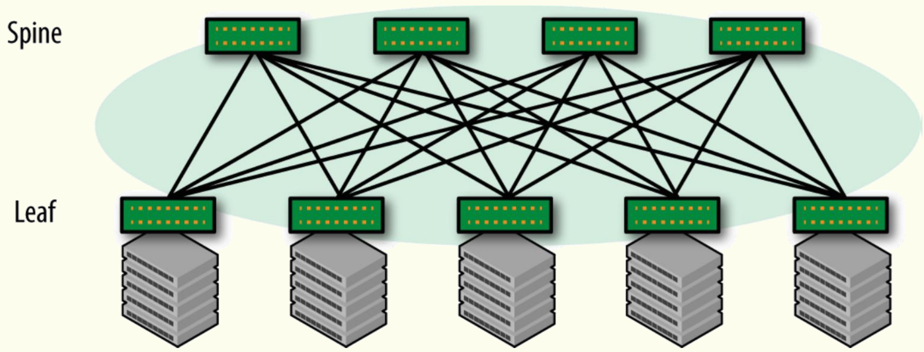
Access/Aggregation/Core network design



- **Unscalability**
  - Flooding
  - VLAN limitations
  - Burden of ARP
  - Limitations of switches and STP
- **Complexity**
- **Failure domain**
- **Unpredictability**

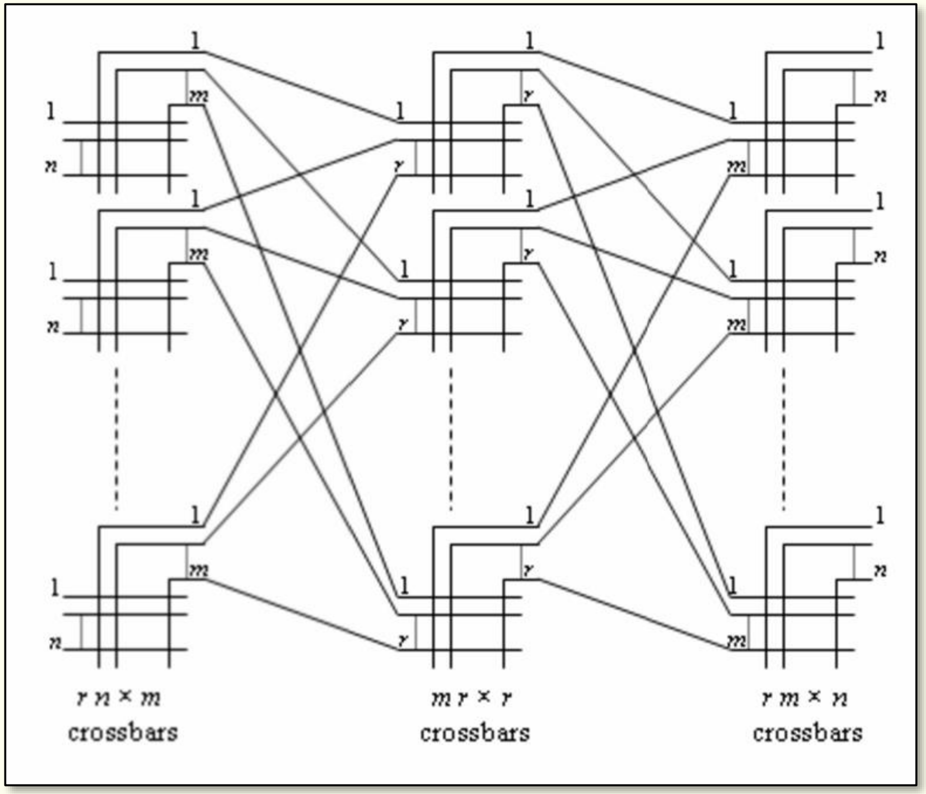
Modern DCNs

- The flexibility promised by bridging to run multiple upper-layer protocols is no longer needed. The only network-layer protocol that need be supported is IP!
- Modern DCNs are IP-based (**IP fabric**) with a ***Leaf-and-Spine*** topology



Clos networks

Originally invented by Charles Clos in the 1950s for old, circuit-switched, telephone networks



Charles Clos. "A Study of Non-blocking Switching Networks". Bell System Technical Journal, March 1953

Scaling a switching matrix by decomposition: An N×N matrix is realized by a multistage network made of smaller switching matrices organized into multiple layers or stages

Let us consider a **3-stage network**

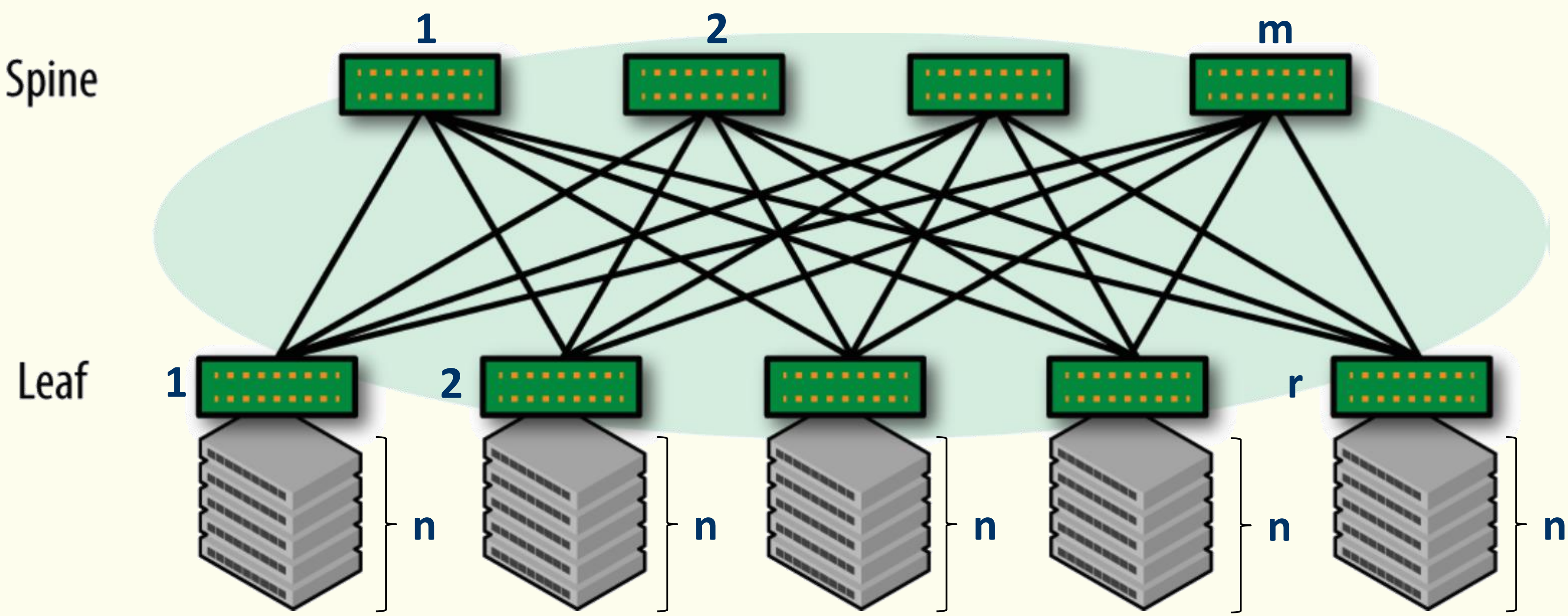
- Original matrix: N×N
- Let us decompose **N = r×n**
- Input stage: **r** (n×m) switches
- Output stage: **r** (m×n) switches
- Intermediate stage: <u>**m**</u> (r×r) switches

If the number of intermediate switches **m** is not sufficiently high, a blocking condition may occur

- <u>Non-blocking condition</u> (re-routing may be necessary): $m \geq n$
- <u>Non-blocking condition without re-routing</u>: $m \geq 2n - 1$ (Clos theorem)

Modern DCNs

- A simple Leaf-and-Spine topology is a folded (three-stage) Clos network



N = r×n servers connected

Modern DCNs

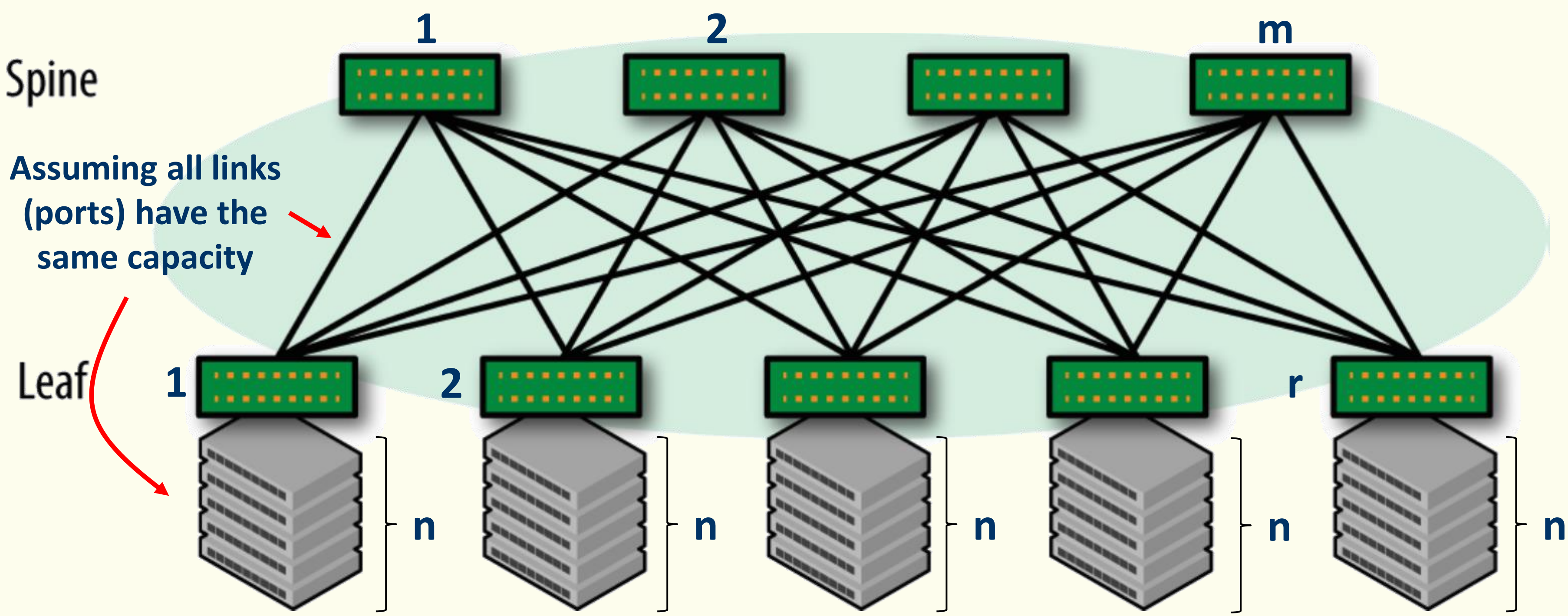
Advantages of Leaf-and-Spine topologies

- **Simple and scalable**
  - Scale-out: More access ports → more leafs (if spines have sufficient downlink ports)
  - Scale-up: More capacity → more spines (and bandwidth on links)
- **Better load distribution**: *m* equal paths between each two leaves → ECMP for east-west traffic
- **Lower CAPEX and OPEX**
  - Economy of scale: switches with fixed configuration
  - Ease of configuration

Capacity of the DCN

- How many access ports can be available (<u>non-blocking</u>)? N = r×n
- R ports on Spine → r = R
- K ports on Leaves (n + m = K) → n = m = K/2

N_{max} = RxK/2



N = r×n servers connected

Capacity of the DCN

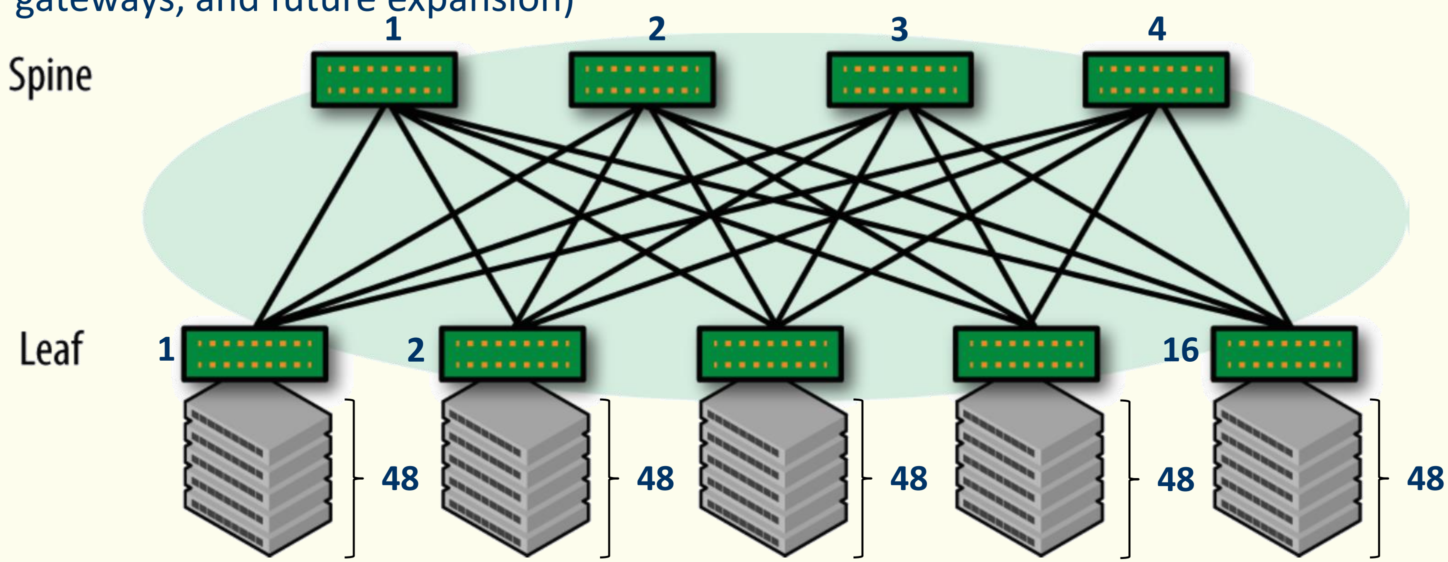
In common deployments, servers are interconnected to leaf switches via lower-speed links (e.g. 10Gb/s), while Leaf switches are interconnected to Spine switches by higher-speed links (e.g., 40Gb/s)

Oversubscription ratio: the ratio between the overall bandwidth server-side (access) and spine-side on a Leaf switch

Example:

- Leaf switches with **48 10Gb/s** (access) ports + **4 40Gb/s** (uplink) ports → oversubscription ratio **3:1**
- Spine switches with **24 40 Gb/s** ports

750 access ports required → 4 Spine switches (using 16 ports per switch, 8 are left for interconnection to router gateways, and future expansion)



N = 768=48×16 access ports

Capacity of the DCN

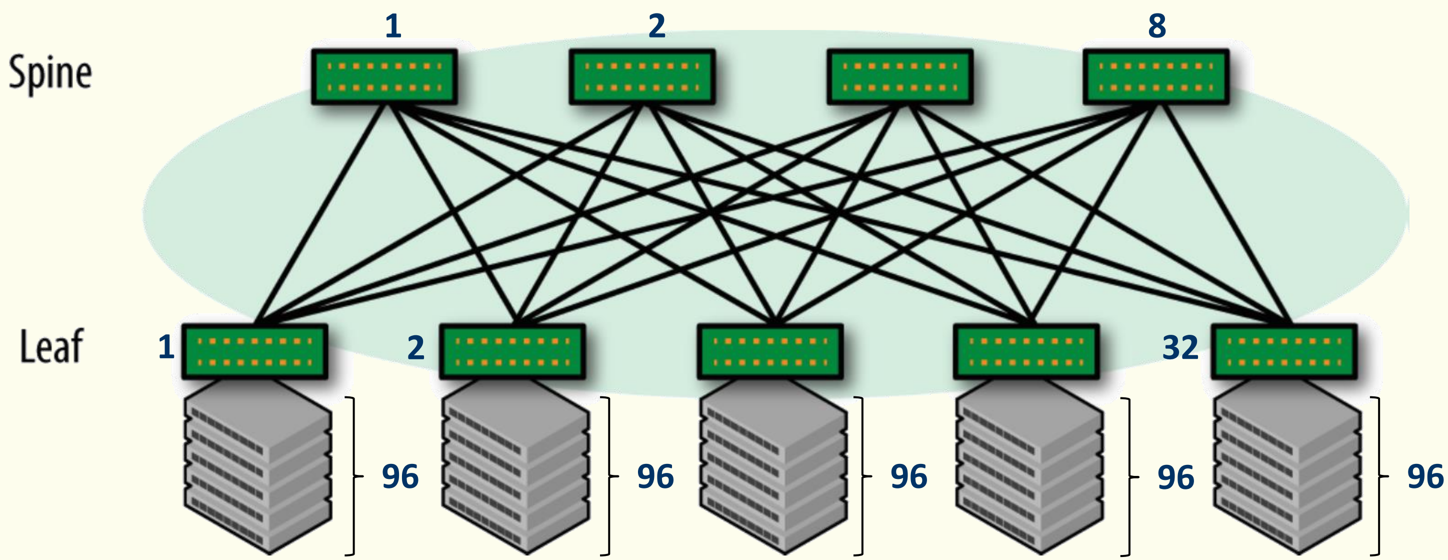
Onother example (realistic figures considering switches available on the market):

- Leaf switches with **96 10Gb/s** access ports + **8 40Gb/s** uplink ports → oversubscription ratio **3:1**
- Spine switches with **32 40 Gb/s** ports

[Today, a 25Gb/s access link coupled with a 100Gb/s uplink is becoming the trend]

How many access ports are available at most? N = 96×32 = **3.072**

What if I need to accommodate 40.000 servers with 10 Gb/s access ports? Spine switches should have 40.000/96 = 417 ports each!!!

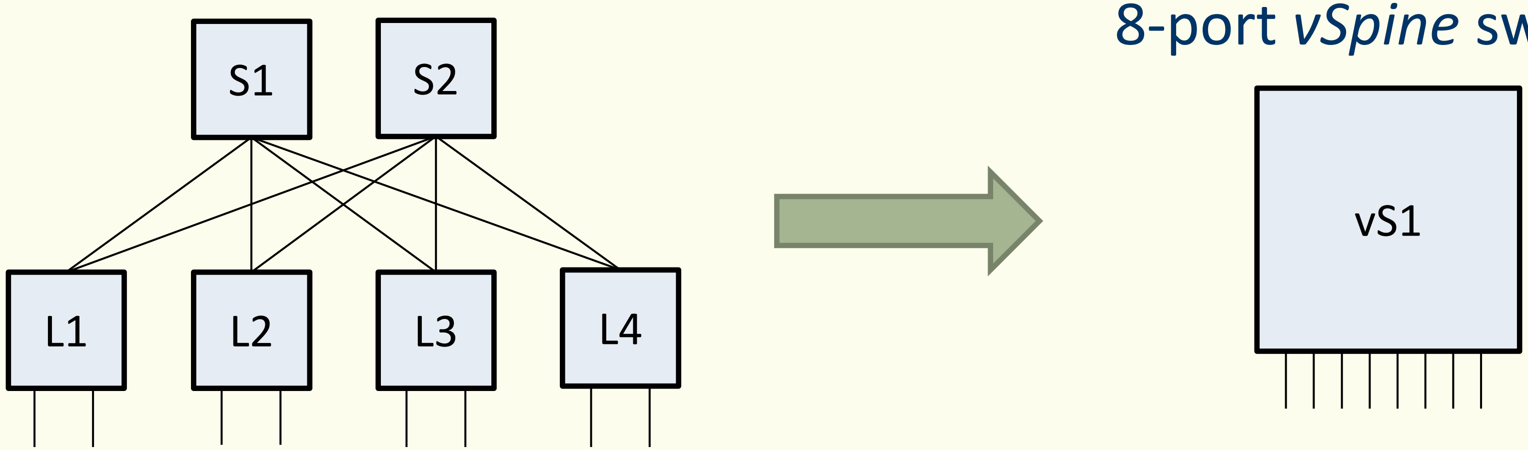


Scaling Clos networks

Example: two-tier (three-stage folded) Clos network

- 6 four-port switches
- 1:1 oversubscription

8-port *vSpine* switch

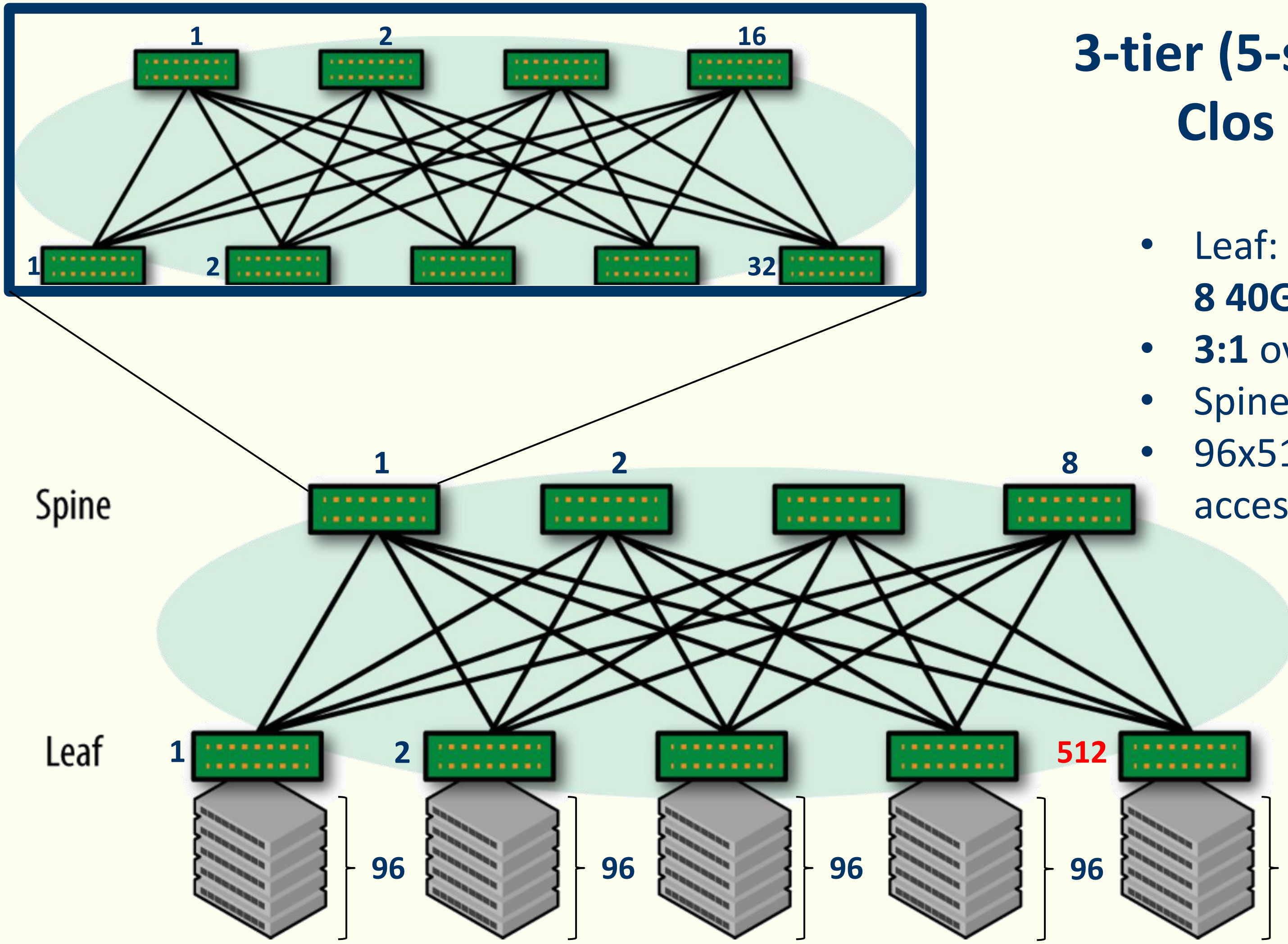


In general

- (N+N/2) N-port switches
- 1:1 oversubscription

$(N^2/2)$-port *vSpine* switch

Scaling Clos networks

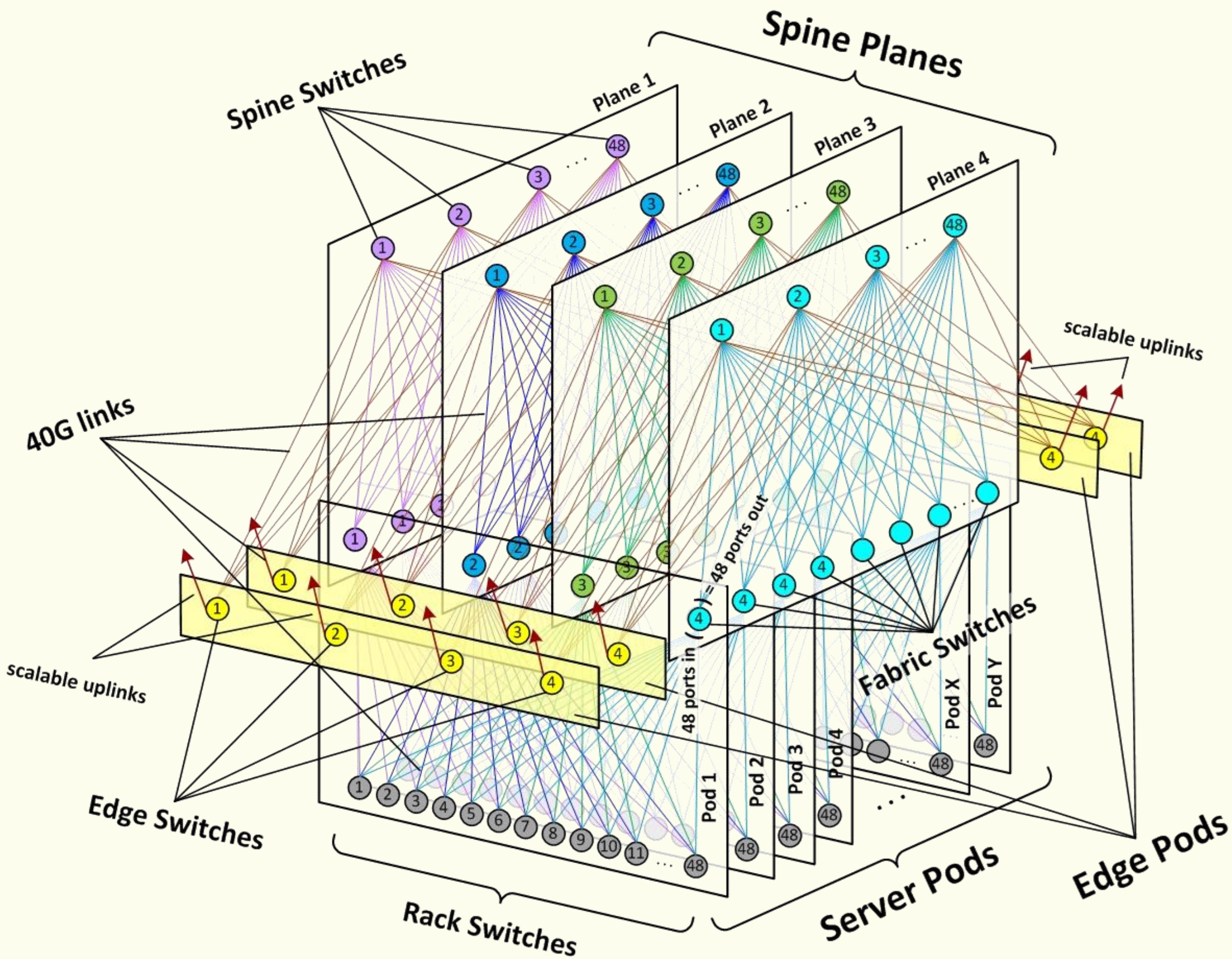


3-tier (5-stage folded) Clos network

Example

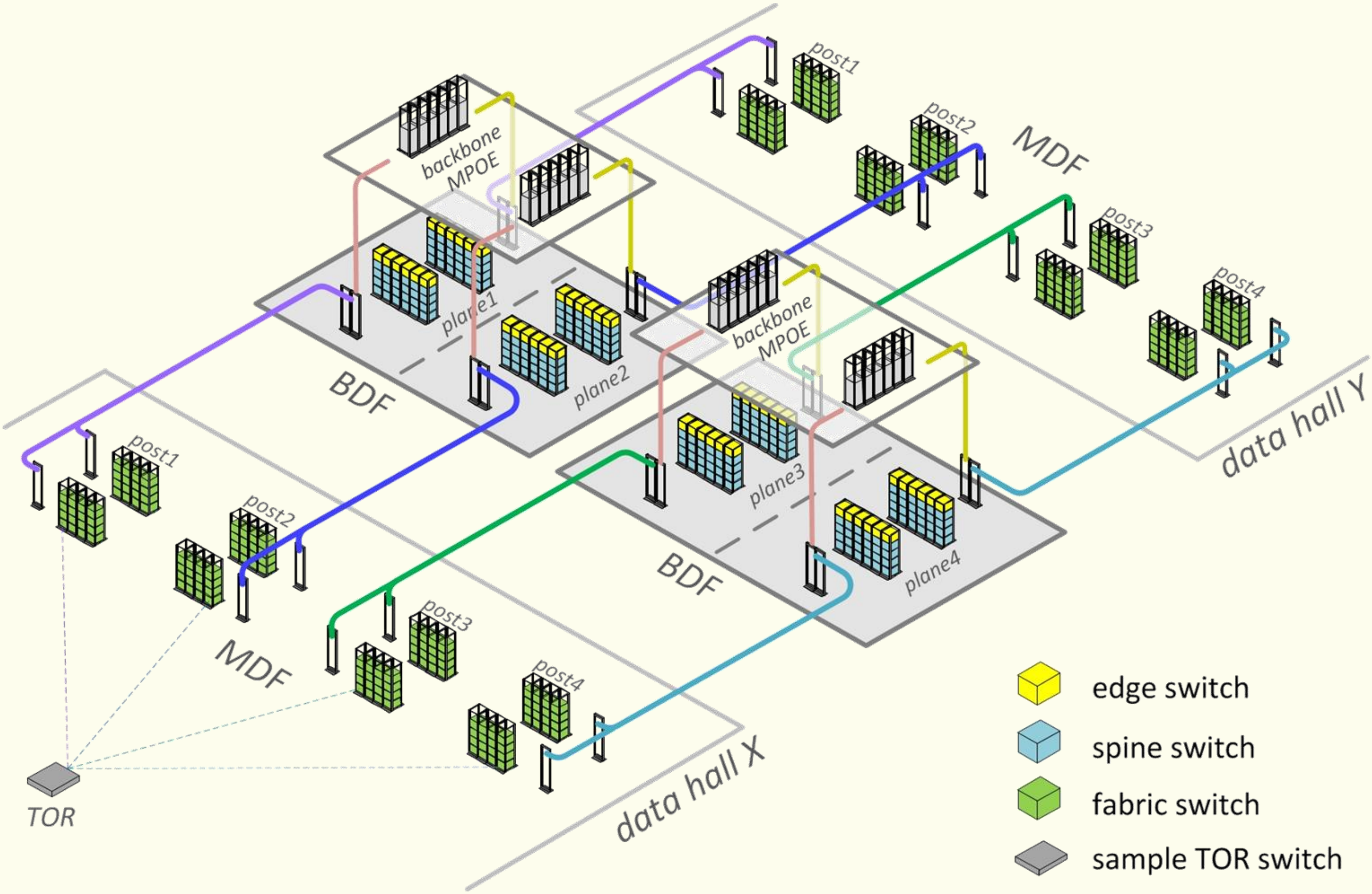
- Leaf: **96 10Gb/s** access + **8 40Gb/s** uplink
- **3:1** oversubscription
- Spine: **32 40Gb/s** ports
- 96x512 = **49.152 10Gb/s** access ports

Data center networks



[Introducing data center fabric, the next-generation Facebook data center network - Facebook Engineering (fb.com)](#)

Data center networks

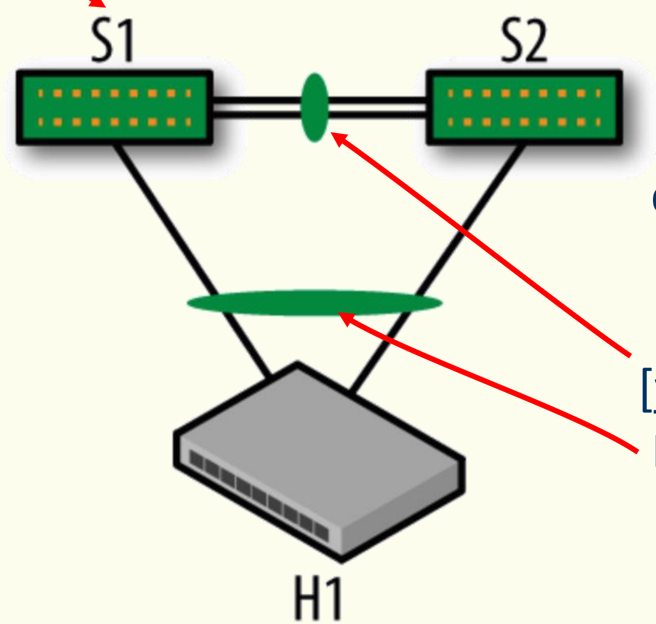


[Introducing data center fabric, the next-generation Facebook data center network - Facebook Engineering (fb.com)](https://engineering.fb.com)

Server attach models

- Single-attach vs. dual-attach server

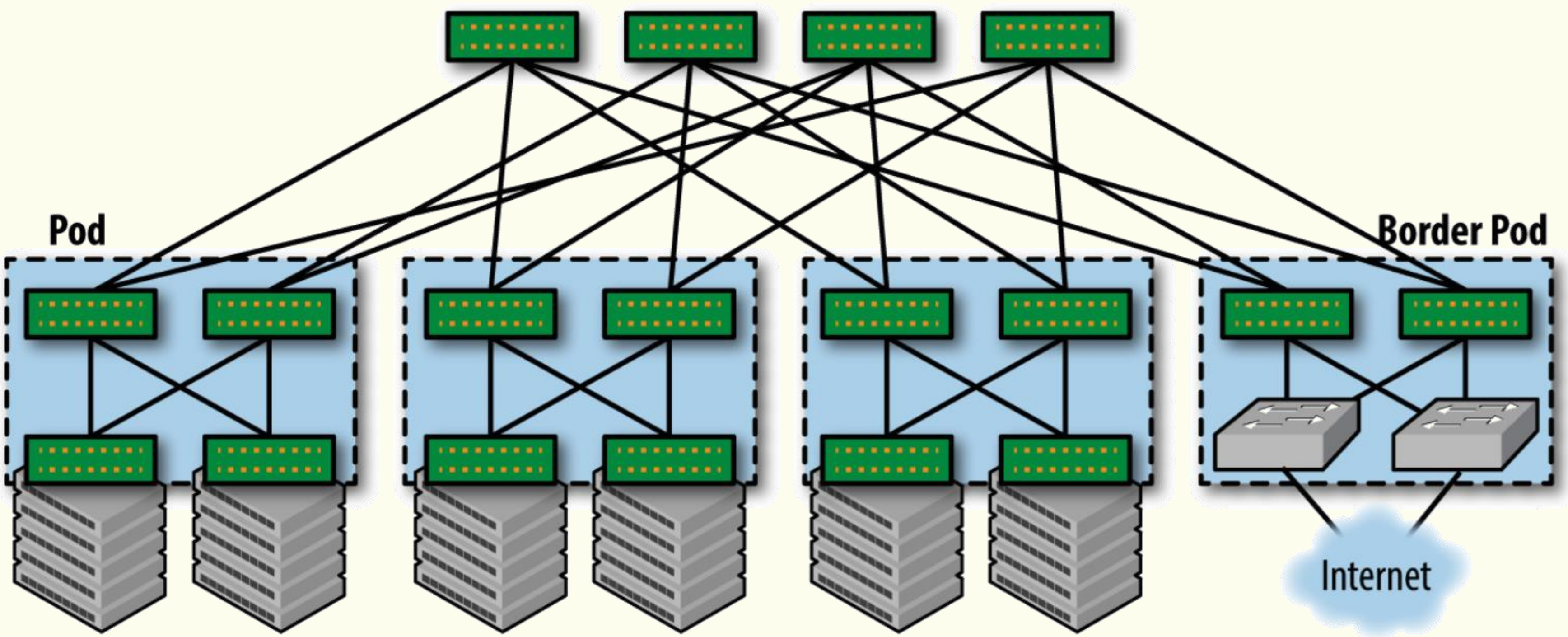
To simplify cabling and facilitate rack mobility, usually both ToRs are in the same rack



dual links are aggregated into a **single logical link**
This requires
[vendor-proprietary protocols +]
LACP (Link Aggregation Control Protocol)

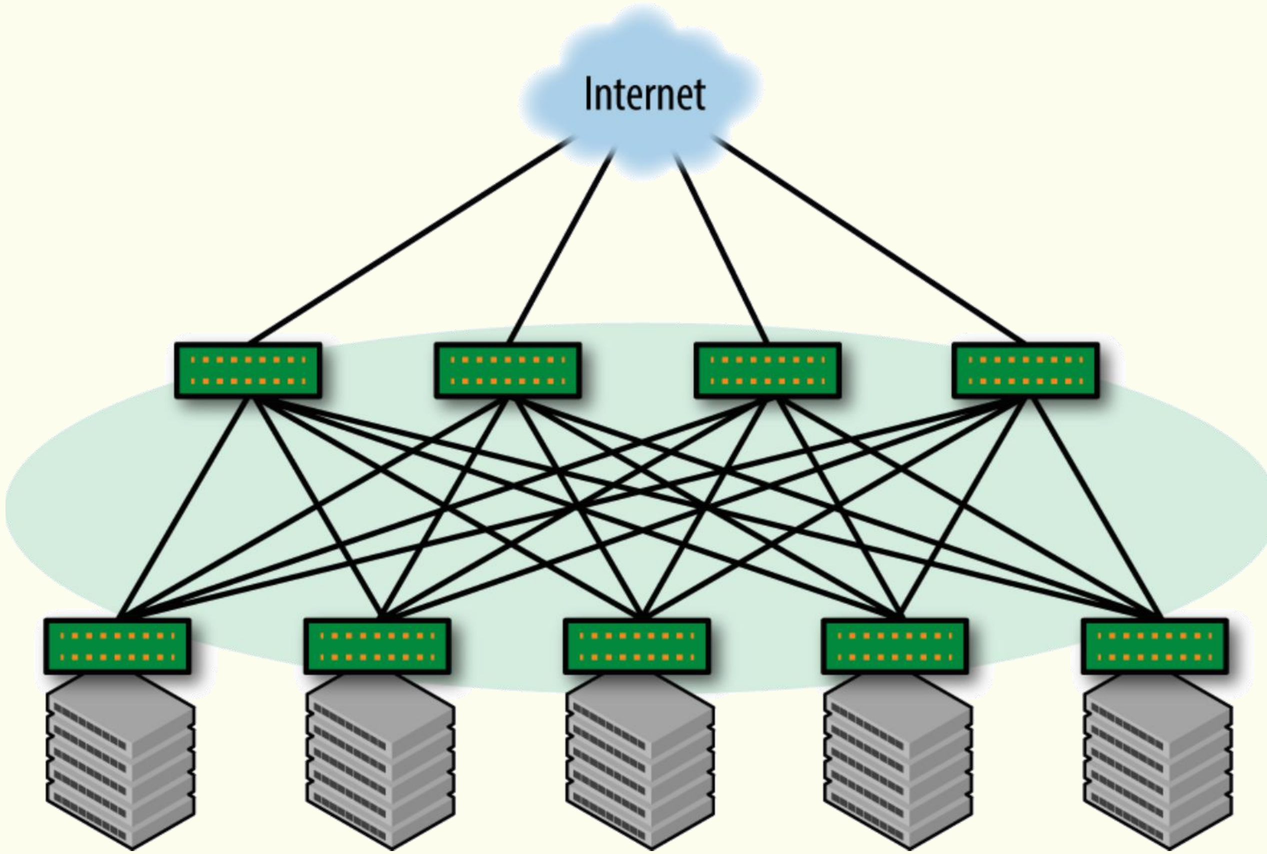
External connectivity

- Via *border* ToRs or pods



External connectivity

- Via Spine switches

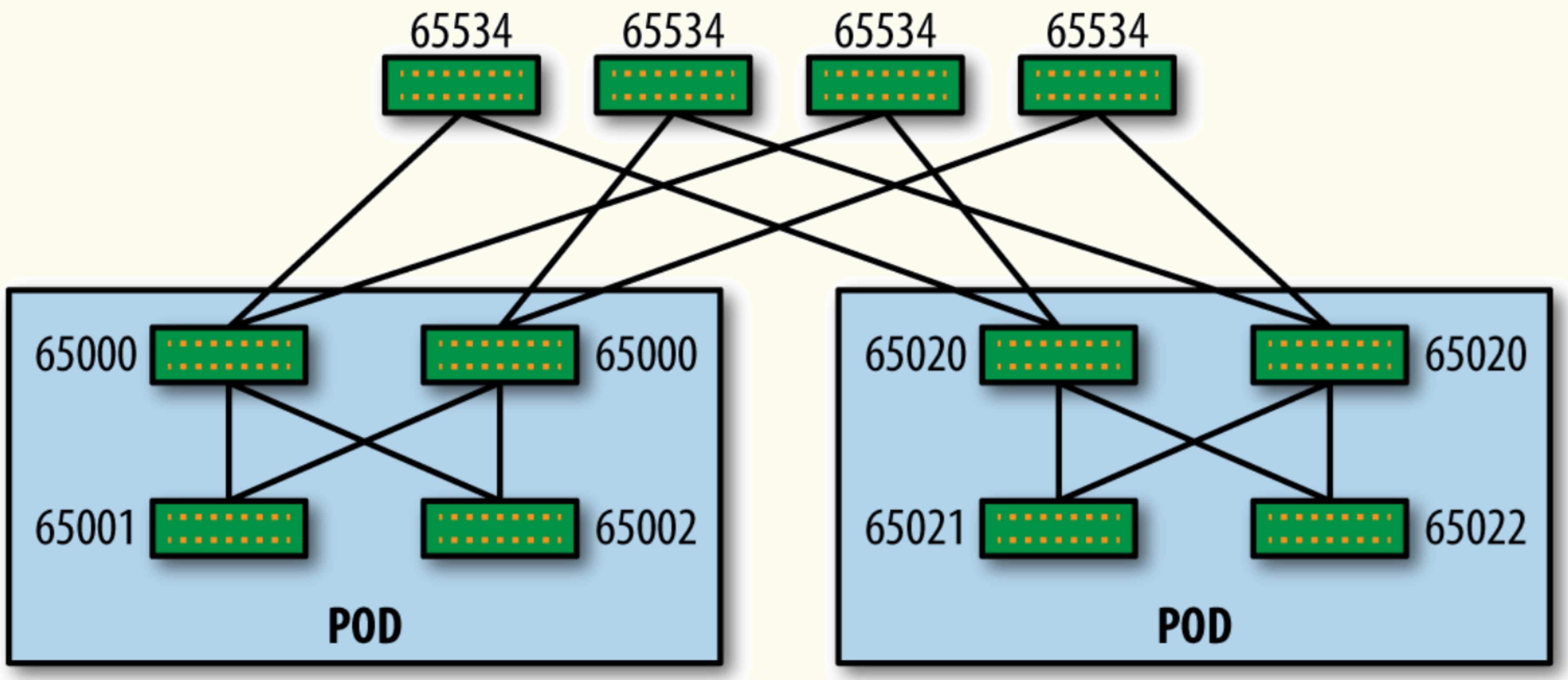


DCN Underlay – IP fabric

- Data plane: **IP**
  - In some cases, **MPLS** may be available
- Control plane
  - Distributed
    - **IGP**: OSPF or IS-IS
    - IGP-free: **eBGP**
  - Centralized/Hybrid
    - **SDN**

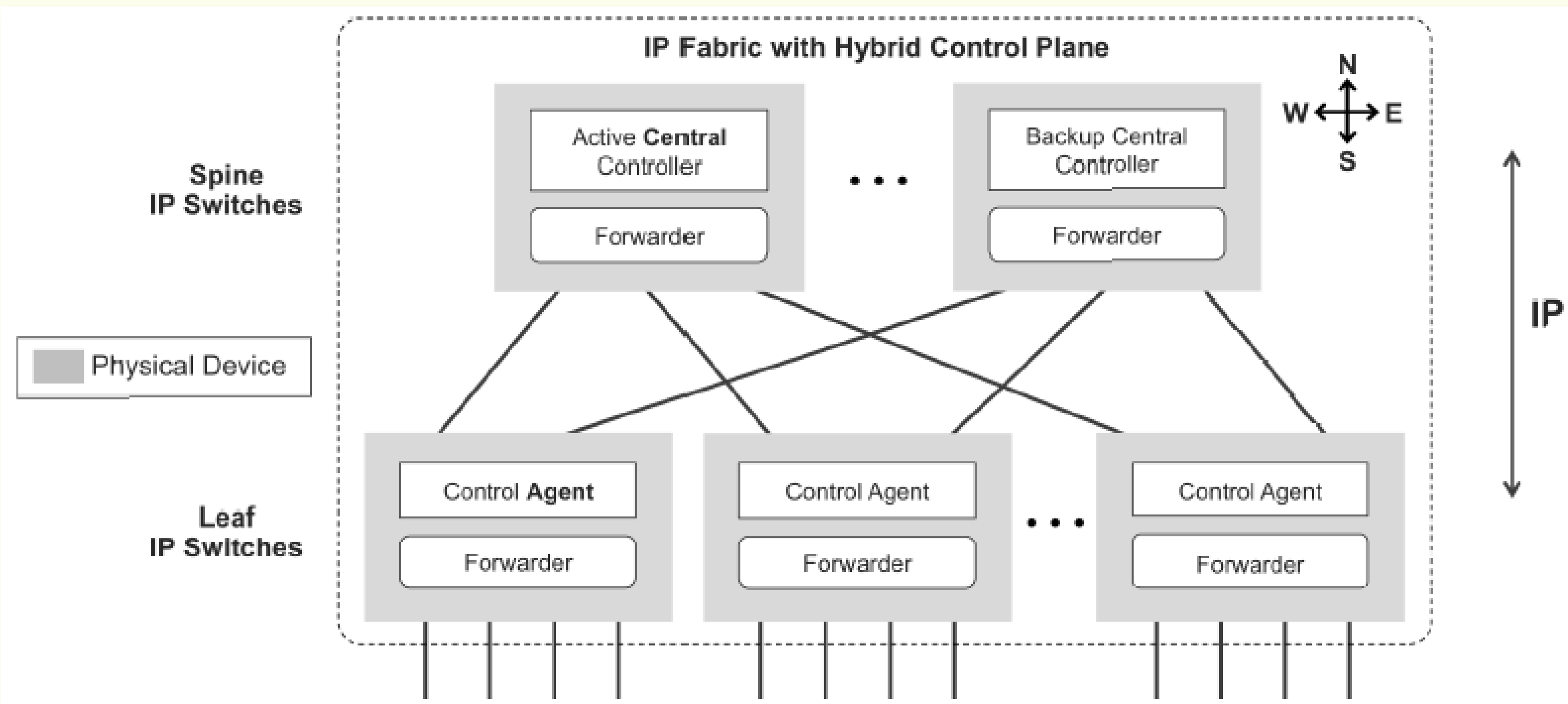
DCN Underlay – IP fabric

- Distributed control plane: eBGP



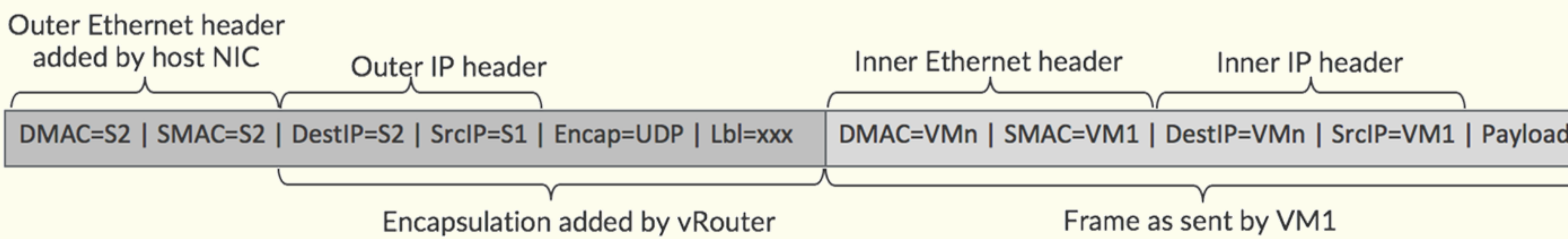
DCN Underlay – IP fabric

- Centralized/Hybrid control plane: either proprietary or use non-standard protocol extensions



Network Virtualization Overlay

- Data plane: **L2**
  - Ethernet frames tunneled over the IP fabric (**VXLAN**, MPLSoUDP, MPLSoGRE, NVGRE, STT, …)

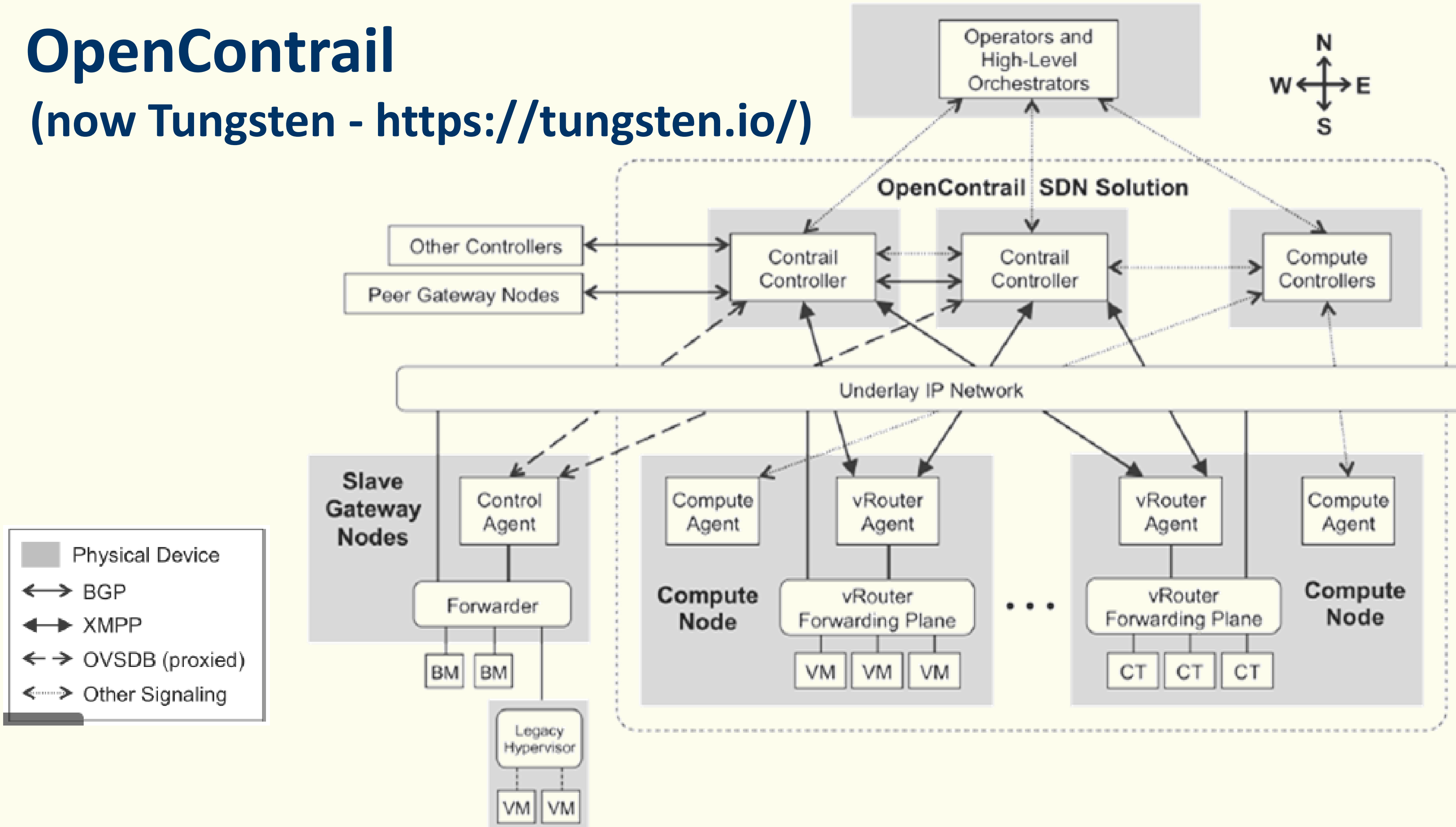


- Control plane: **controller**
  - Centralized: **SDN-based**
  - Protocol-based: **VXLAN + EVPN**

Network Virtualization Overlay

OpenContrail

(now Tungsten - https://tungsten.io/)



References

- Dinesh G. Dutt, **Cloud Native Data Center Networking: Architecture, Protocols, and Tools** 1st ed., O'Reilly, Dec. 2019
- RFC 7938 - **Use of BGP for Routing in Large-Scale Data Centers**
- RFC 8365 - **A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)**