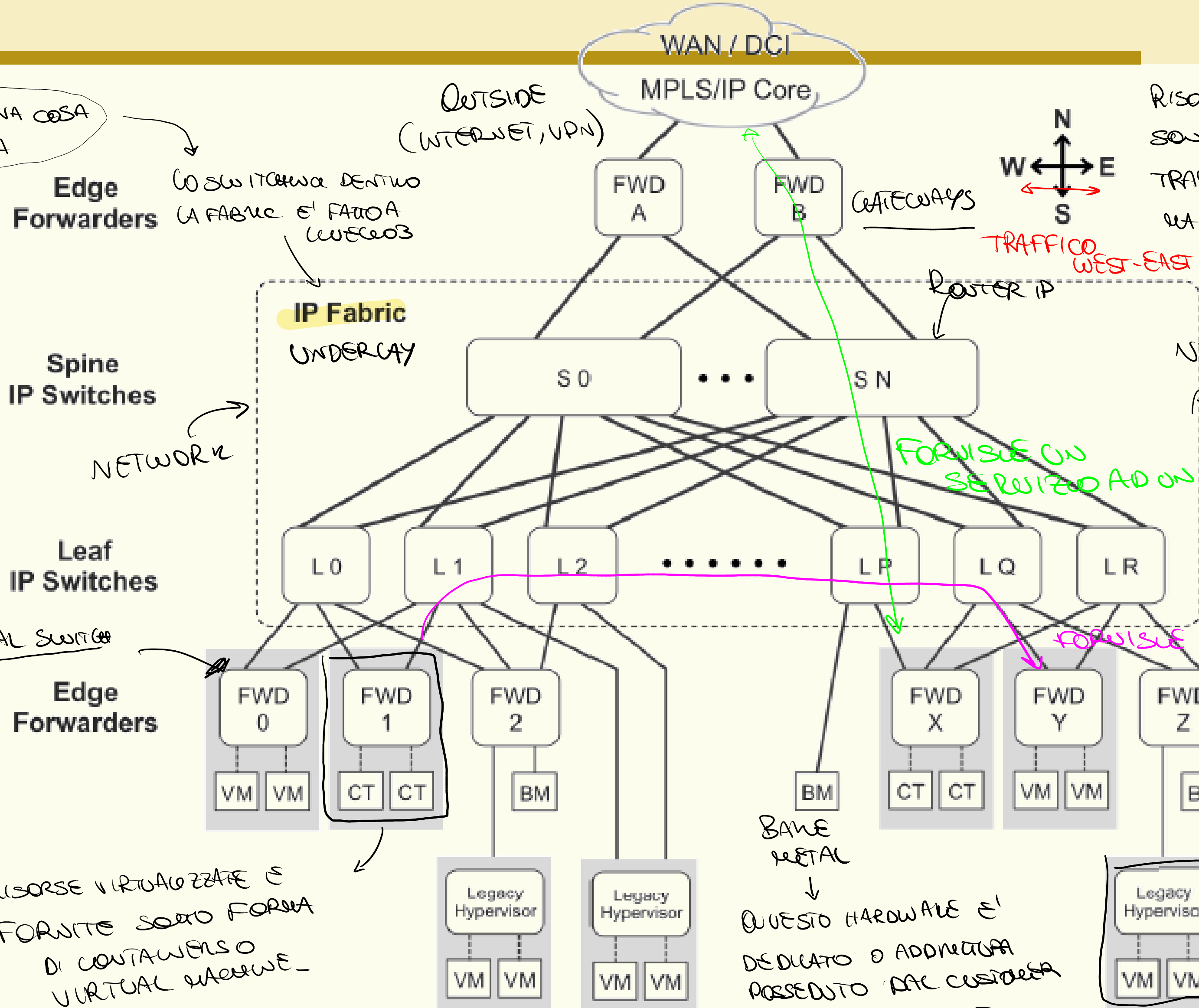
Data Center Networks

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

Data Center Networks

I SERVIZI TIPICI CHE VENGONO FORNITI SONO SERVIZI WEB.



NON E' UNA COSA SCONTATA

LO SWITCHING DENTRO LA FABRIC E' FATTO A LIVELLO3

OUTSIDE (INTERNET, VPN)

GATEWAYS

TRAFFICO WEST-EAST

ROUTER IP

UNDERLAY

NETWORK

RISORSE COMPUTAZIONALI SONO USATE DAI CLIENTI TRAMITE VIRTUALIZZAZIONE, MA SONO ALLOCATE IN DATA CENTER

NEI SERVER ABBIAMO APPLICAZIONI CHE FORNISCONO SERVIZI.

FORNISCE UN SERVIZIO AD UN ALTRO CUSTOMER NORTH-SOUTH TRAFFIC

VIRTUAL SWITCH

FORNISCE UN SERVIZIO AD UN ALTRO SERVIZIO. WEST-EAST TRAFFIC

SERVERS

BARE METAL

QUESTO HARDWARE E' DEDICATO O ADDIRITTURA POSSEDUTO DAL CUSTOMER NON MULTITENANT

RISORSE VIRTUALIZZATE E FORNITE SOTTO FORMA DI CONTAINER O VIRTUAL MACHINE.

SISTEMI LEGACY SENZA VIRTUAL SWITCH

Modern DCN requirements

- **Increased server-to-server communication**
  - modern data center applications involve a lot of server-to-server communication (east-west)

    ↳ MAGGIORE RISPETTO AL TRAFFICO NORTH-SOUTH. (INPUT - OUTPUT)

    MICROSERVICES, SERVERLESS ... AUMENTA LA GRANULARITÀ DI ALLOCAZIONE DELLE RISORSE.

    ↳ QUESTO HA AVUTO IMPATTO SULLA DECISIONE DI PASSARE DA L2 A L3 FABRIC.
- **Scale**
  - modern data centers range from a few hundred to a hundred thousand servers in a single physical location

    ↳ BISOGNA TENERNE CONTO A DESIGN TIME.

    ↳ CI SONO BISOGNO DI CENTINAIA DI MIGLIAIA DI PORTE (ANCHE FISICHE) PER CONNETTERE I SERVER.

    ↳ QUANDO USIAMO CONTAINER OVM I NUMERI CRESCONO ANCORA
- **Resilience**
  - The primary aim is to limit the effect of a failure to as small a footprint as possible

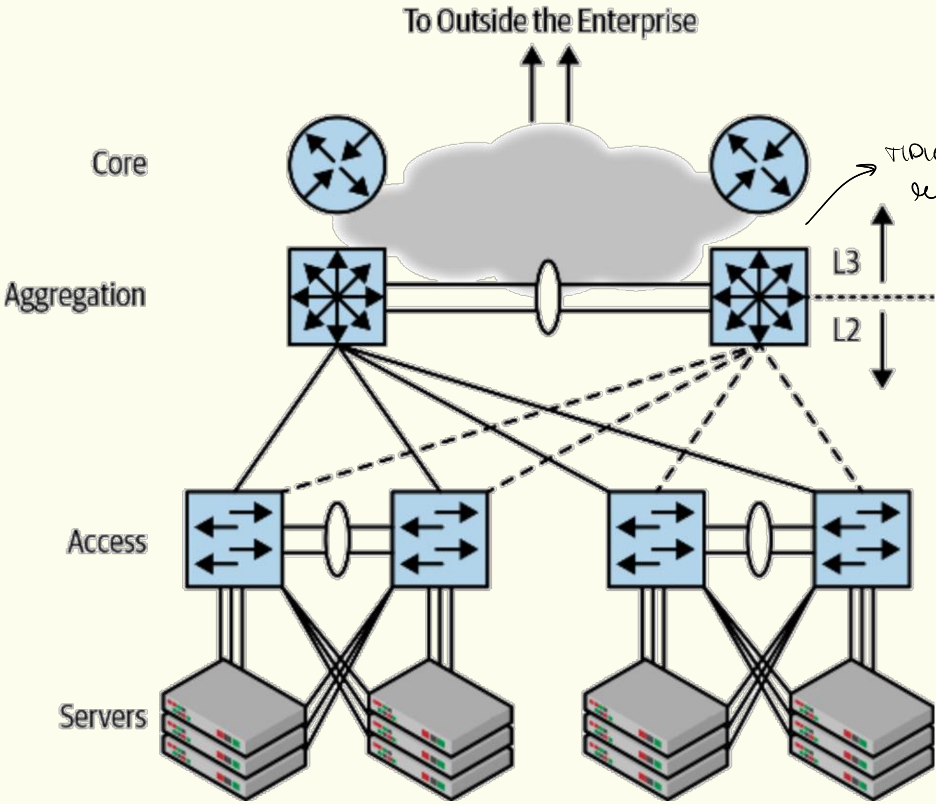
    ↳ DEFINIRE TOOL CHE PERMETTANO DI GESTIRE GUASTI A RUNTIME.

    ↳ GESTIONE E CONFINAMENTO GUASTI

Legacy DCN topologies

Access/Aggregation/Core network design

COSTRUITA COME UNA RETE AZIENDALE_



TIPICAMENTE SONO MULTI LAYER SWITCH

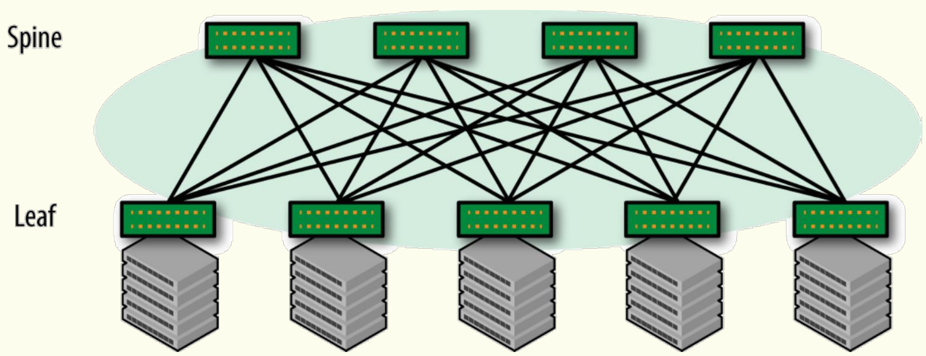
FABRIC

- **Unscalability**
  - Flooding
  - VLAN limitations 12 BIT PER VLAN ID
  - Burden of ARP
  - Limitations of switches and STP → HAI DELLA POTENZA COMPUTAZIONALE CHE NON PUOI USARE PER COME È PENSATA LA RETE DI LIVELLO 2.
- **Complexity**
- **Failure domain**
- **Unpredictability**

Modern DCNs

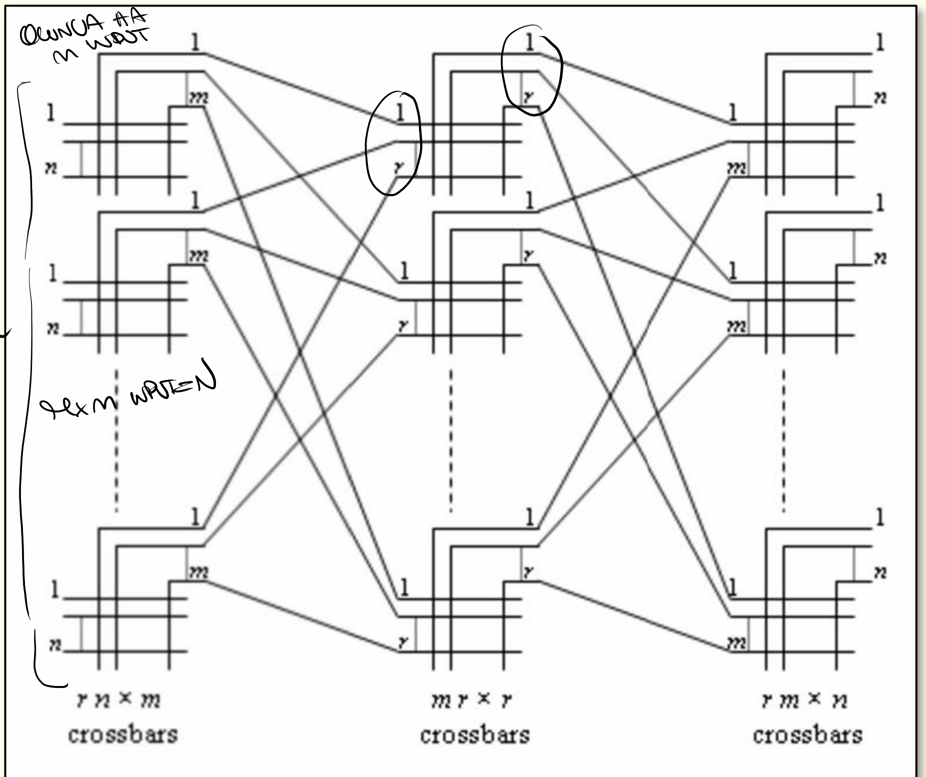
- The flexibility promised by bridging to run multiple upper-layer protocols is no longer needed. The only network-layer protocol that need be supported is IP!
- Modern DCNs are IP-based (**IP fabric**) with a ***Leaf-and-Spine*** topology

QUALE È IL PROBLEMA?
PUÒ ESSERE USATO HARDWARE SPECIALIZZATO, CON SOFTWARE COMPLETAMENTE PROPRIETARIO.



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

RETE UNIDIREZIONALE → ABBIAMO INPUT E OUTPUT SEPARATI E PERCORRIBILI IN UNA SOLA DIREZIONE.

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

- Original matrix: N×N → UN NUMERO DI MATRICE DI SWITCHING PIÙ PICCOLE
- Let us decompose **N = r×n**
- Input stage: **r** (n×m) switches
- Output stage: **r** (m×n) switches
- Intermediate stage: **m** (r×r) switches

NON HO UNO SWITCH GRANDE N×N.

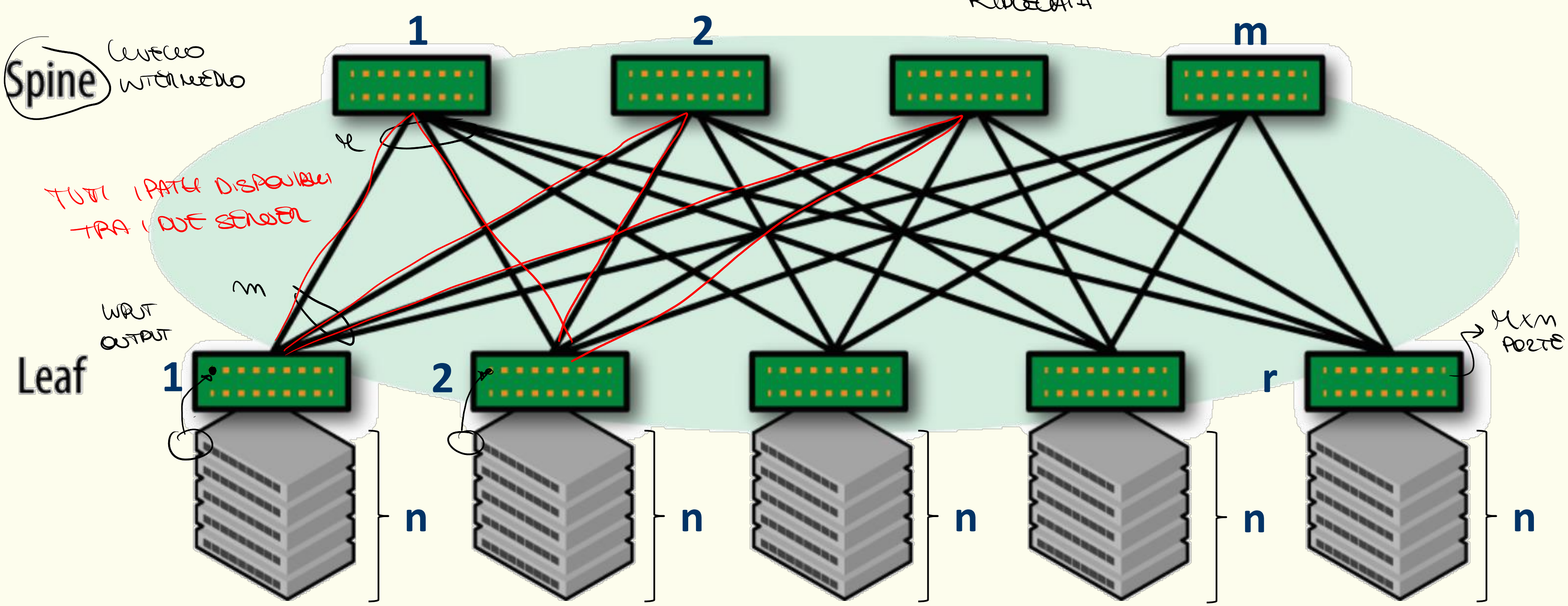
If the number of intermediate switches **m** is not sufficiently high, a blocking condition may occur

- Non-blocking condition (re-routing may be necessary): m ≥ n TROVO QUALCHE CIRCUITO E
- Non-blocking condition without re-routing: m ≥ 2n – 1 (Clos theorem) POSSO CREARE IL CIRCUITO.

Modern DCNs

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

VEDIAMO 2 STADI MA ABBIAMO UNA CLOS NETWORK A 3 STADI RIPIEGATA



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

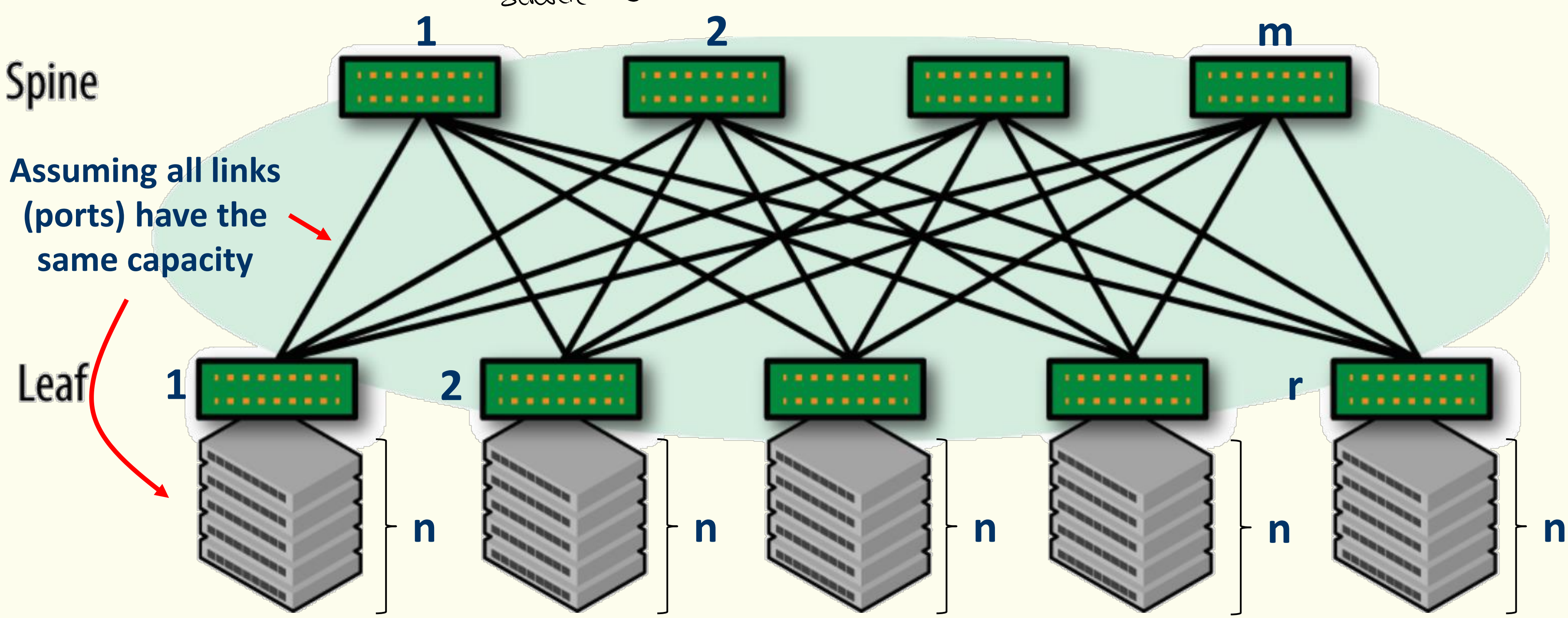
Capacity of the DCN

$m \geq n$

- How many access ports can be available (non-blocking)? N = r×n
- R ports on Spine → r = R → NUMERO MASSIMO DI LEAVES SWITCHES
- K ports on Leaves (n + m = K) → $\boxed{n = m}$ = K/2

SERVER SPINE

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~~ CAPACITY server-side (access) and spine-side on a Leaf switch

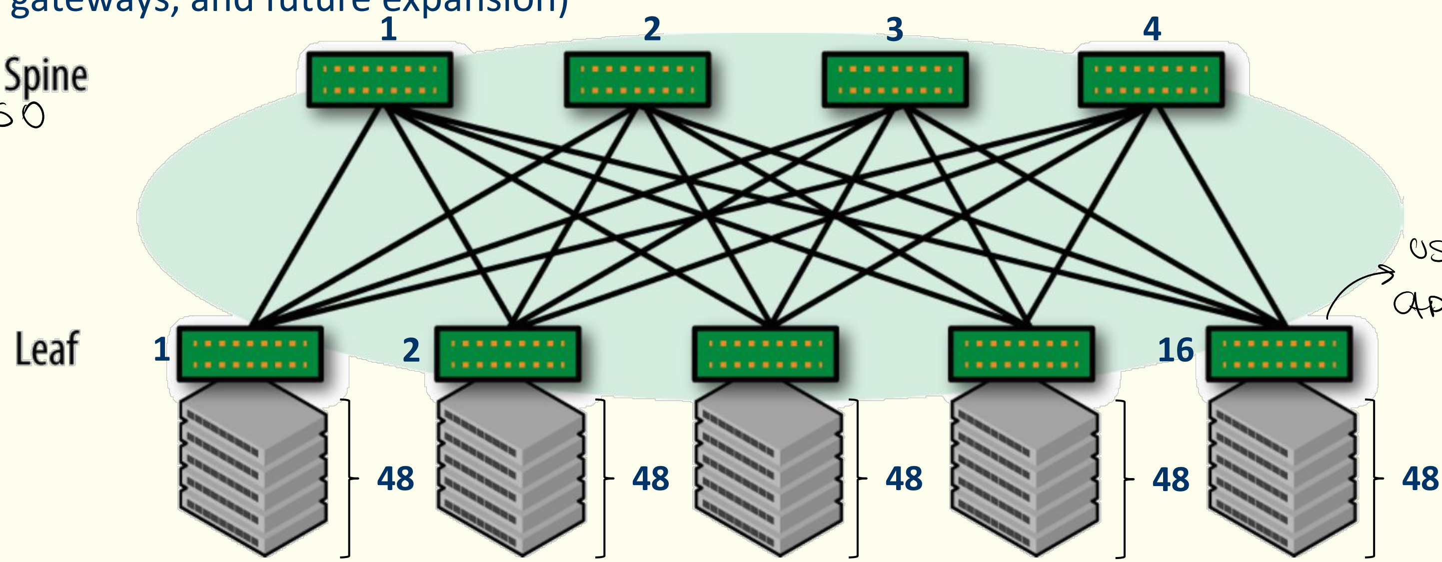
48 × 10 = 480 4 × 40 = 160 → PER COLLEGARE CON LE SPINE VALORE TIPICO

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)

48 × 16 ≥ 750



USO SWITCH CON CAPACITÀ ETEROGENEE

N = 768=48×16 access ports

Capacity of the DCN

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

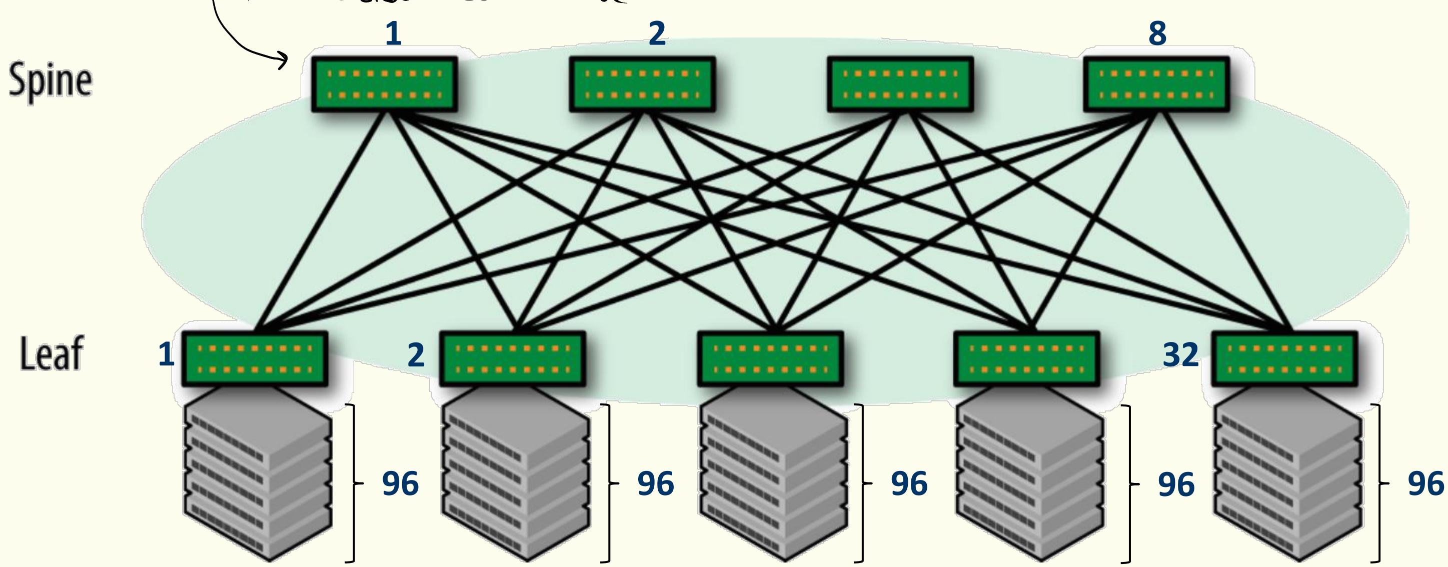
- Leaf switches with **96 10Gb/s** access ports + **8 40Gb/s** uplink ports → oversubscription ratio **3:1**
- Spine switches with **32 40 Gb/s** ports

[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**/2 = 1536 SERVERS

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!!!

NON ESISTONO SWITCH DEL GENERE

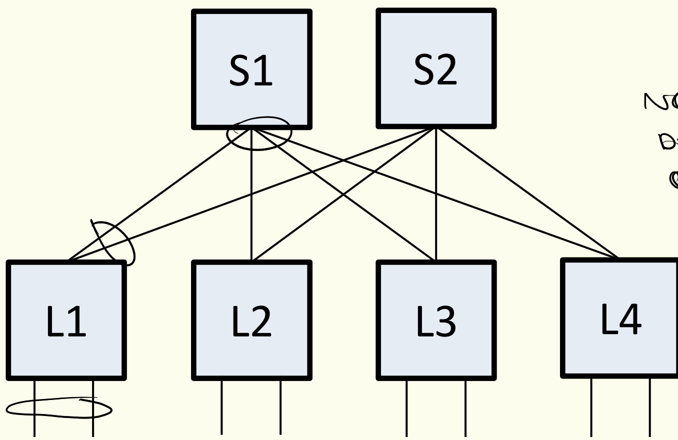


Scaling Clos networks

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

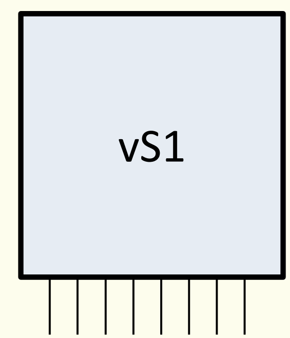
- 6 four-port switches
- 1:1 oversubscription ACCESS CAPACITY = SPINE SIDE CAPACITY

2 + 4 = 6 SWITCH



NON INTRODUCO BOTTLENECK DAL MOMENTO CHE OVERSUBSCRIPTION RATIO È 1:1

8-port *vSpine* switch

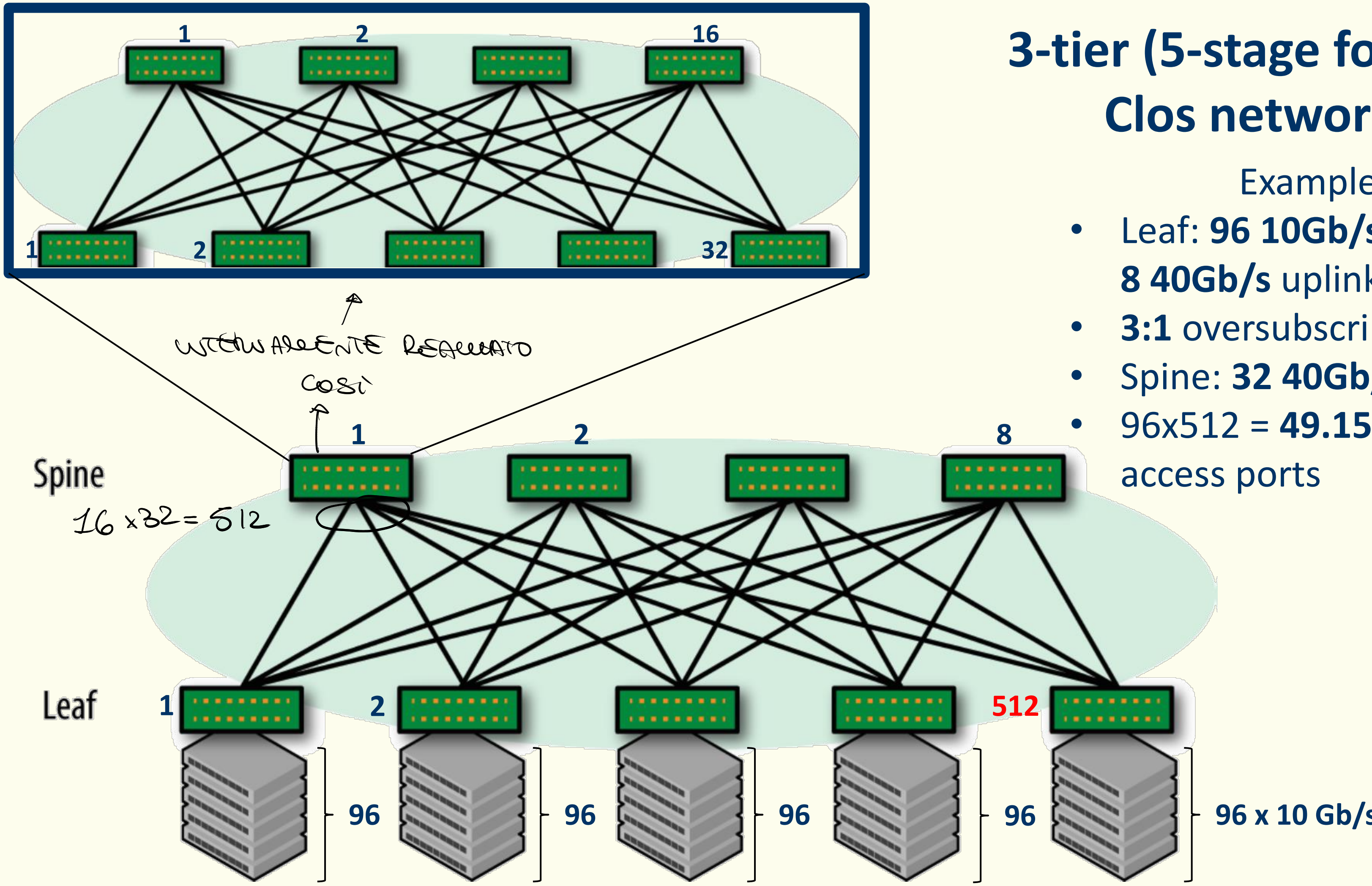


In general

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

$\boxed{(N^2/2)\text{-port } vSpine \text{ switch}}$

Scaling Clos networks



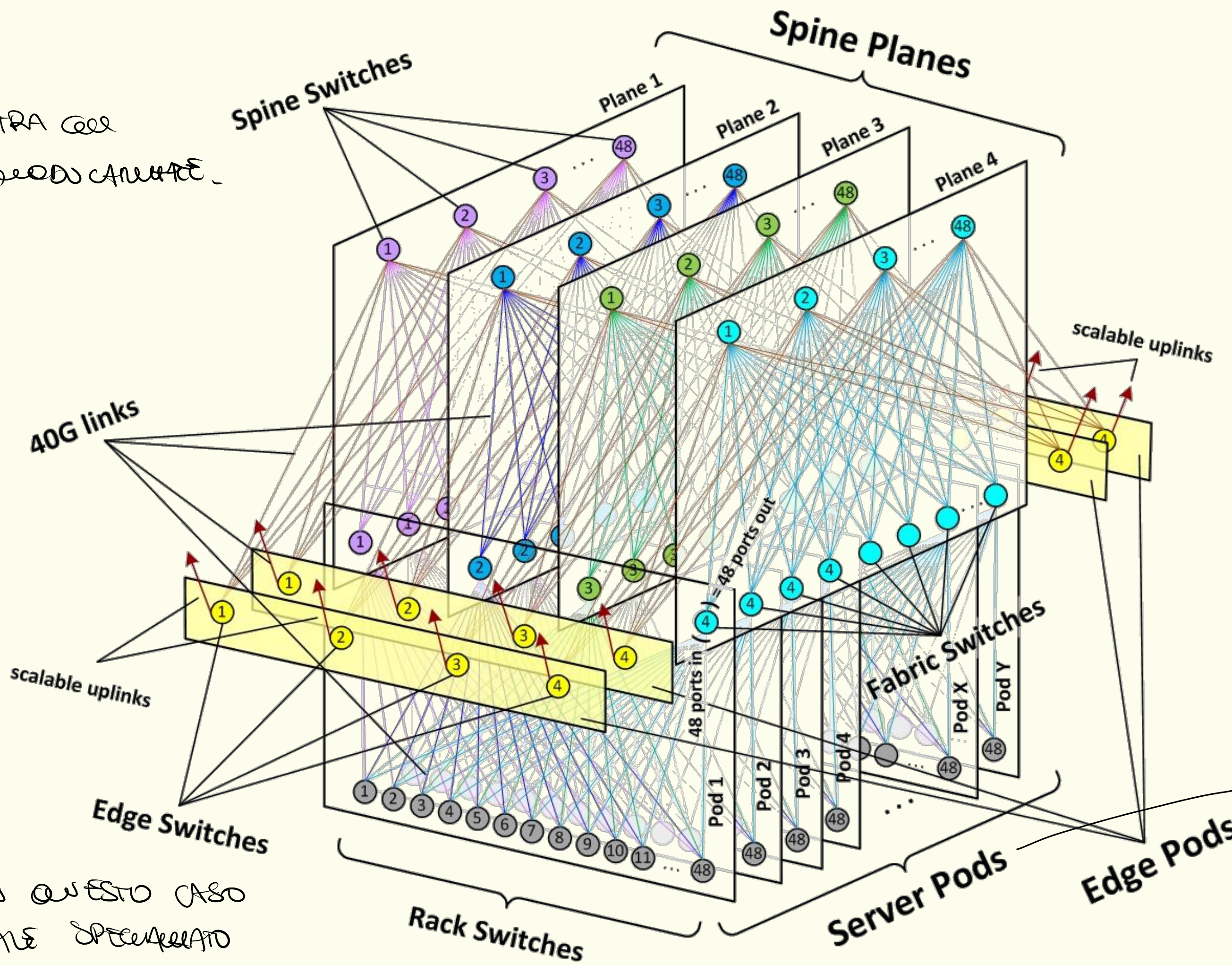
3-tier (5-stage folded) Clos network

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

LE CONNESSIONI TRA GLI SWITCH SONO RIDONDANTE.

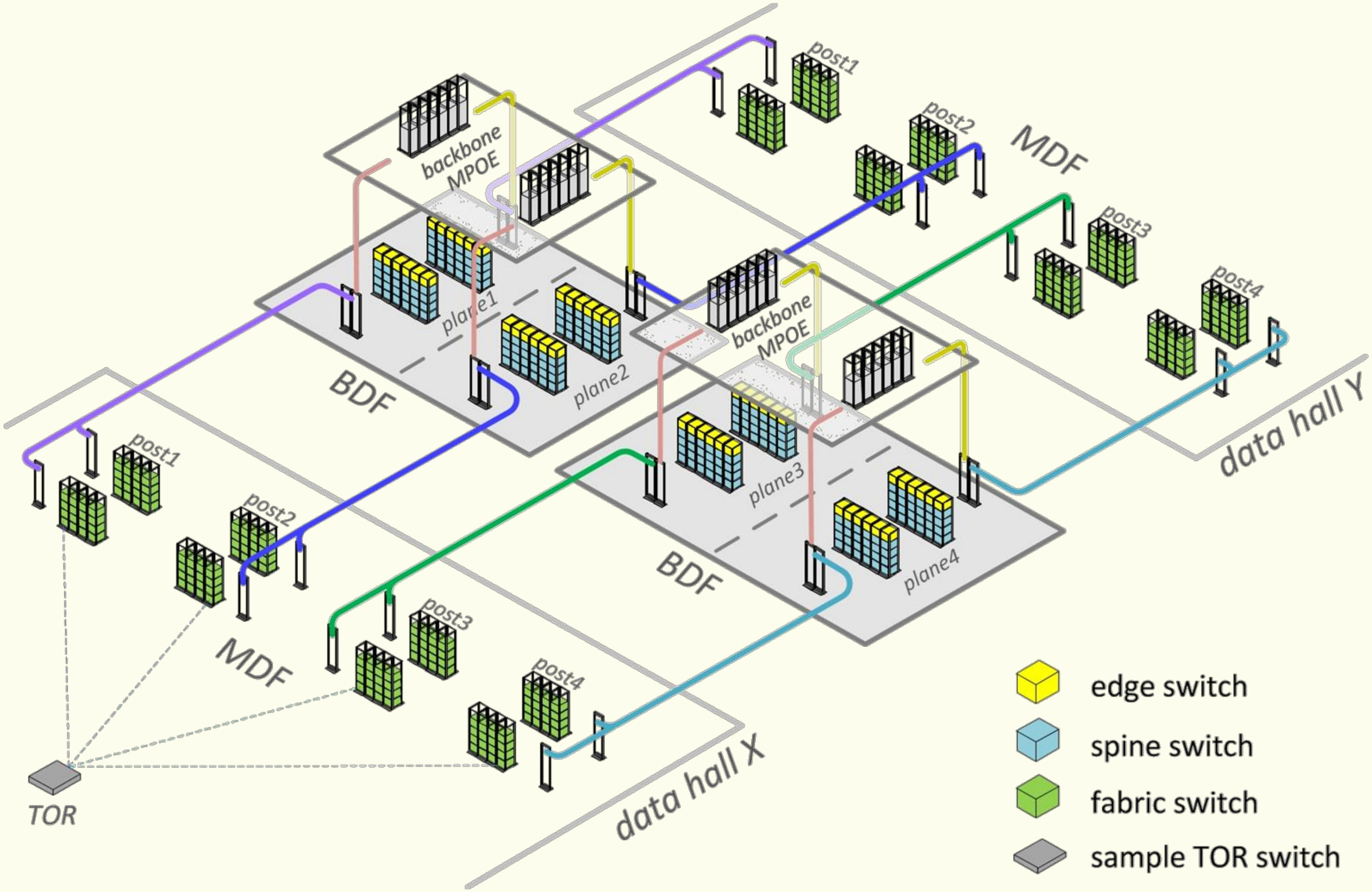


SCALE OUT
=
ADDING POD

GLI SWITCH IN QUESTO CASO
SONO HARDWARE SPECIALIZZATO

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

Data center networks



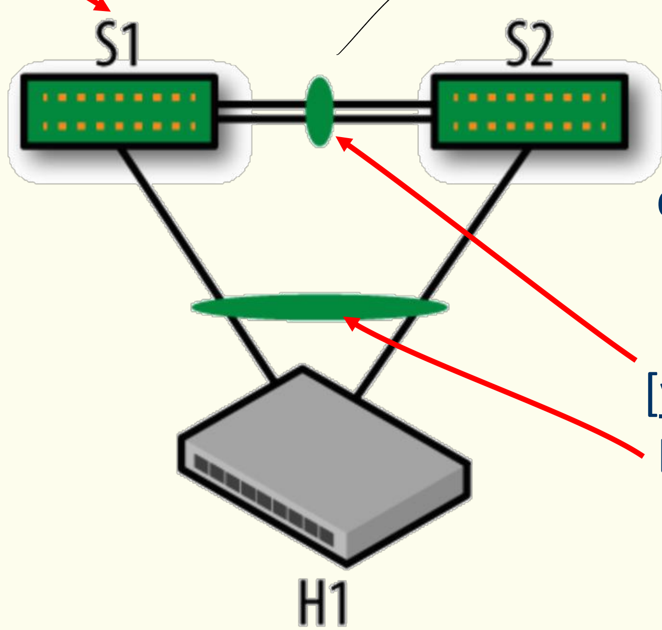
[Introducing data center fabric, the next-generation Facebook data center network - Facebook Engineering (fb.com)](https://engineering.fb.com)

Server attach models

- Single-attach vs. dual-attach server

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

NON VOGLIO CHE UNO DEI DUE VENGA BLOCCATO!
VOGLIO CHE VENGANO USATI IN
ACTIVE-ACTIVE MODE

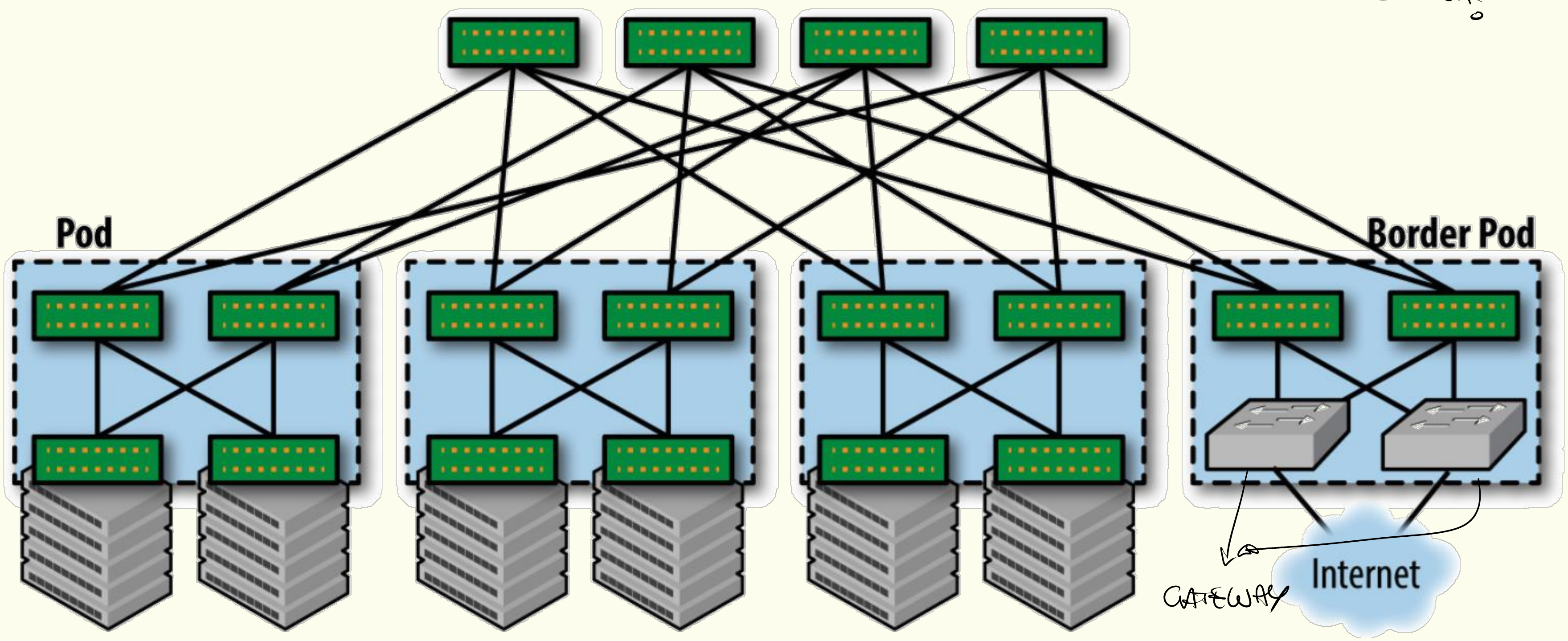


dual links are aggregated into a **single logical link**
This requires [vendor-proprietary protocols +] **LACP** (Link Aggregation Control Protocol)

External connectivity

- Via *border* ToRs or pods

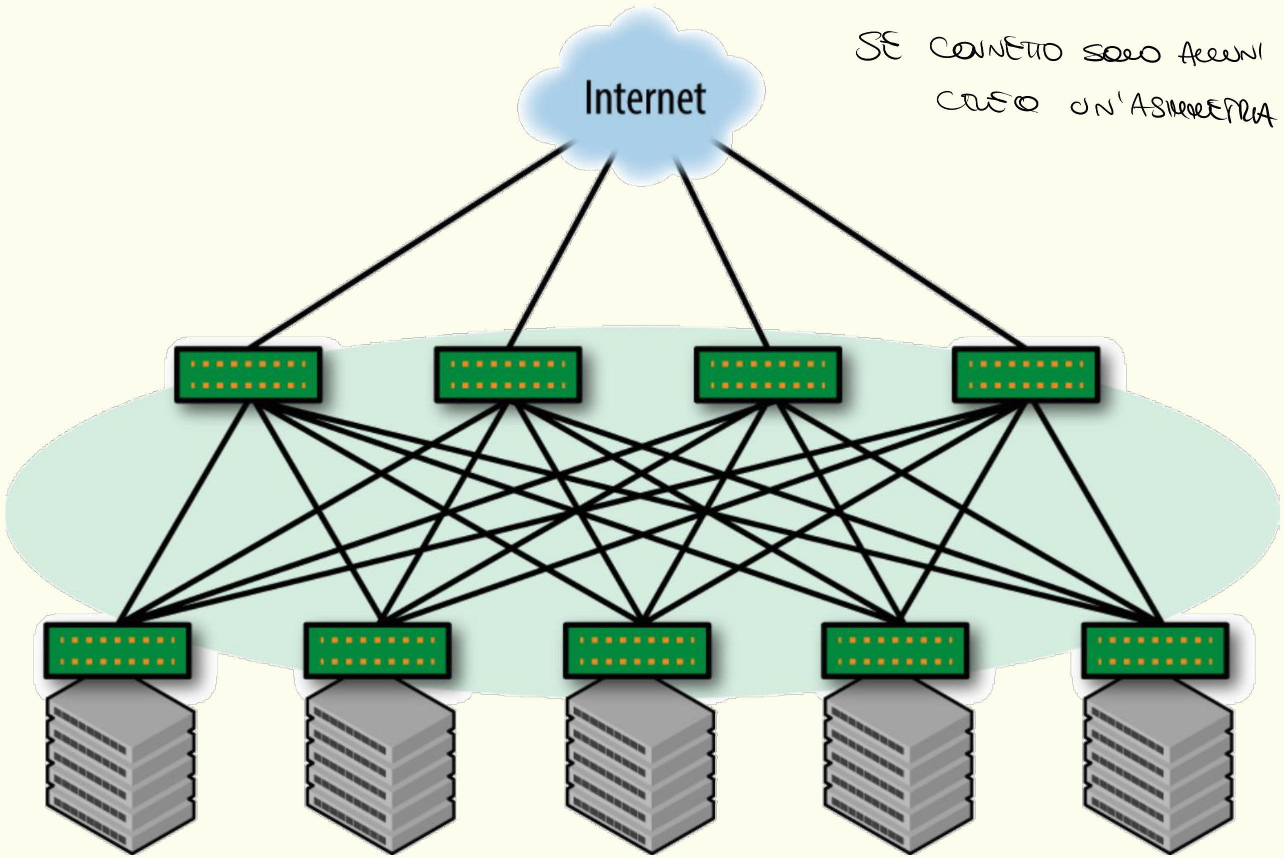
COME CONNETTIAMO, SERVER ALLA RETE ESTERNA?



External connectivity

- Via Spine switches

SE CONNETTO SOLO ALCUNI SPINES
CREO UN'ASIMMETRIA



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**

COME E' IMPLEMENTATO IL CONTROLLO DIPENDE DALLA SCALA!!

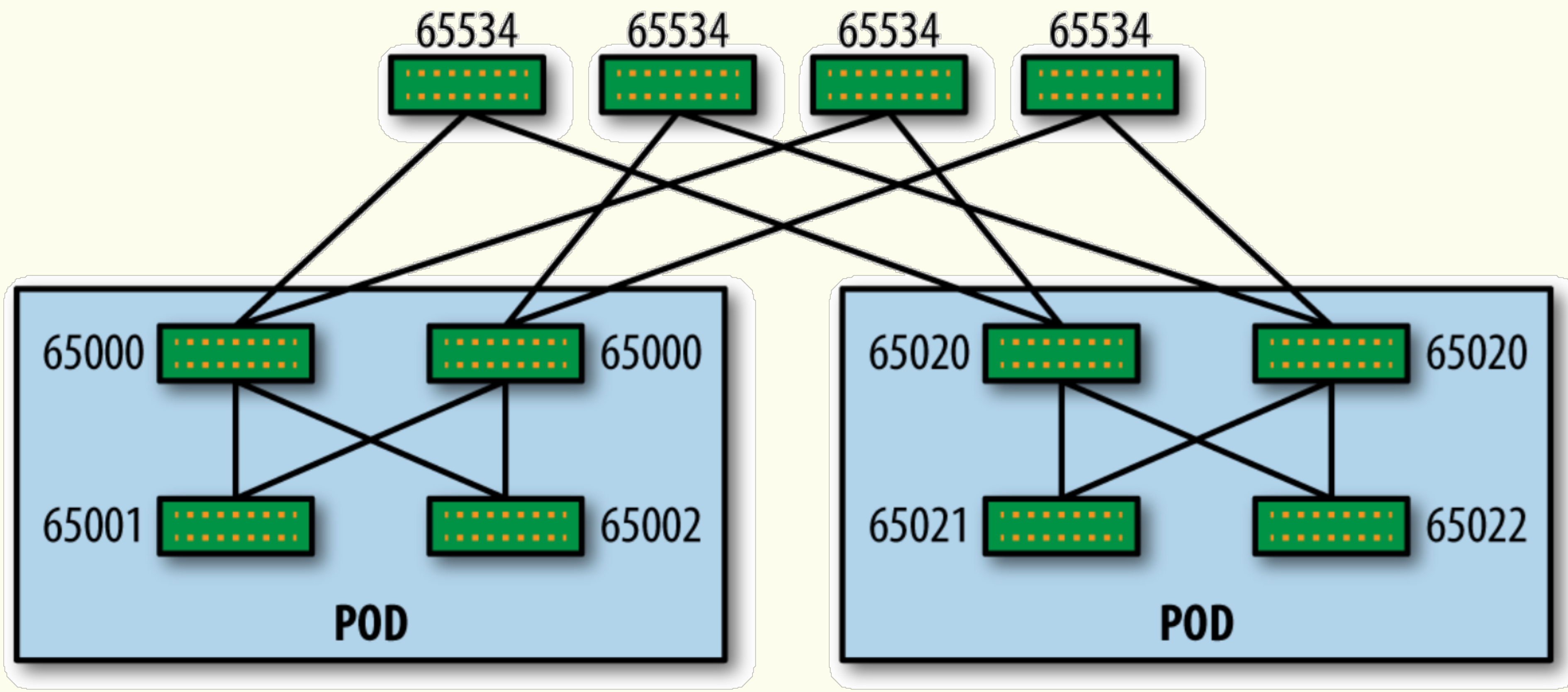
→ EXTERNAL BGP, OGNI ROUTER E' UN AUTONOMOUS SYSTEM.
QUANDO I ROUTER SI PARLANO SI TRATTANO COME AS DIVERSI.

↓

$\boxed{\text{POSSO CONTROLLARE MOLTO MEGLIO}}$

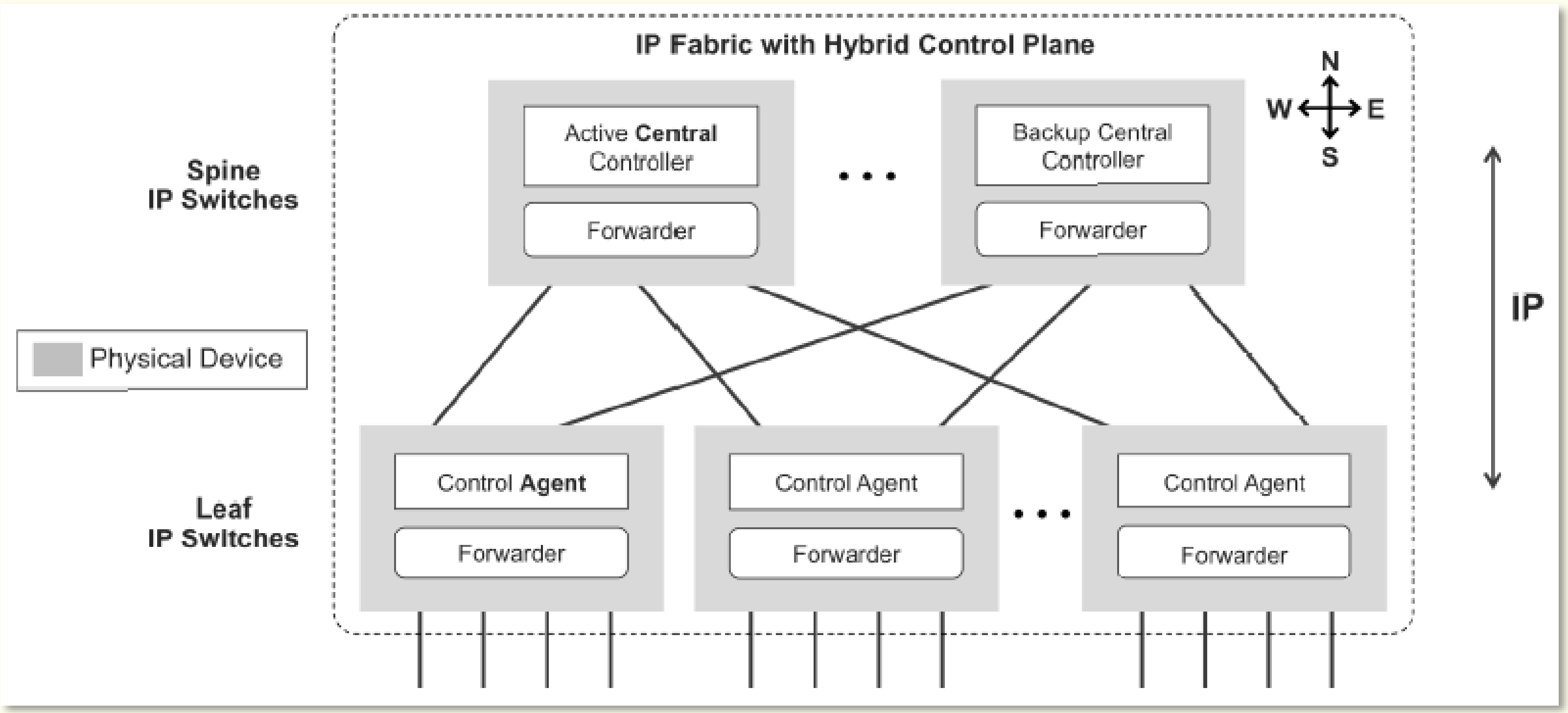
DCN Underlay – IP fabric

- Distributed control plane: eBGP



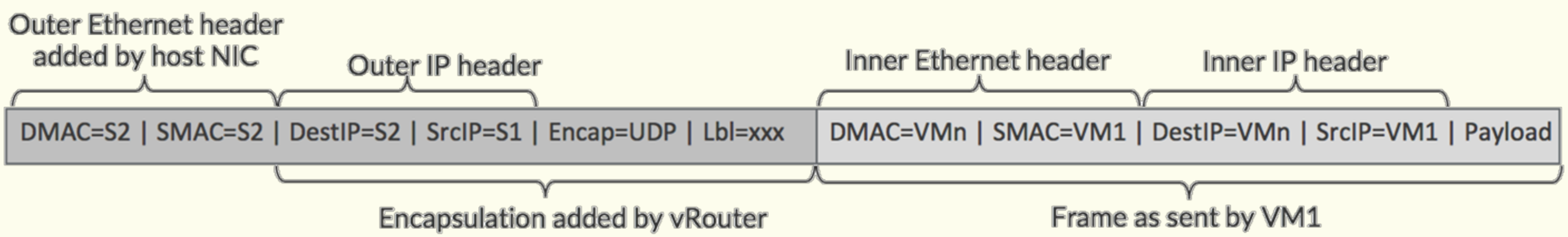
DCN Underlay – IP fabric

- 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, …)

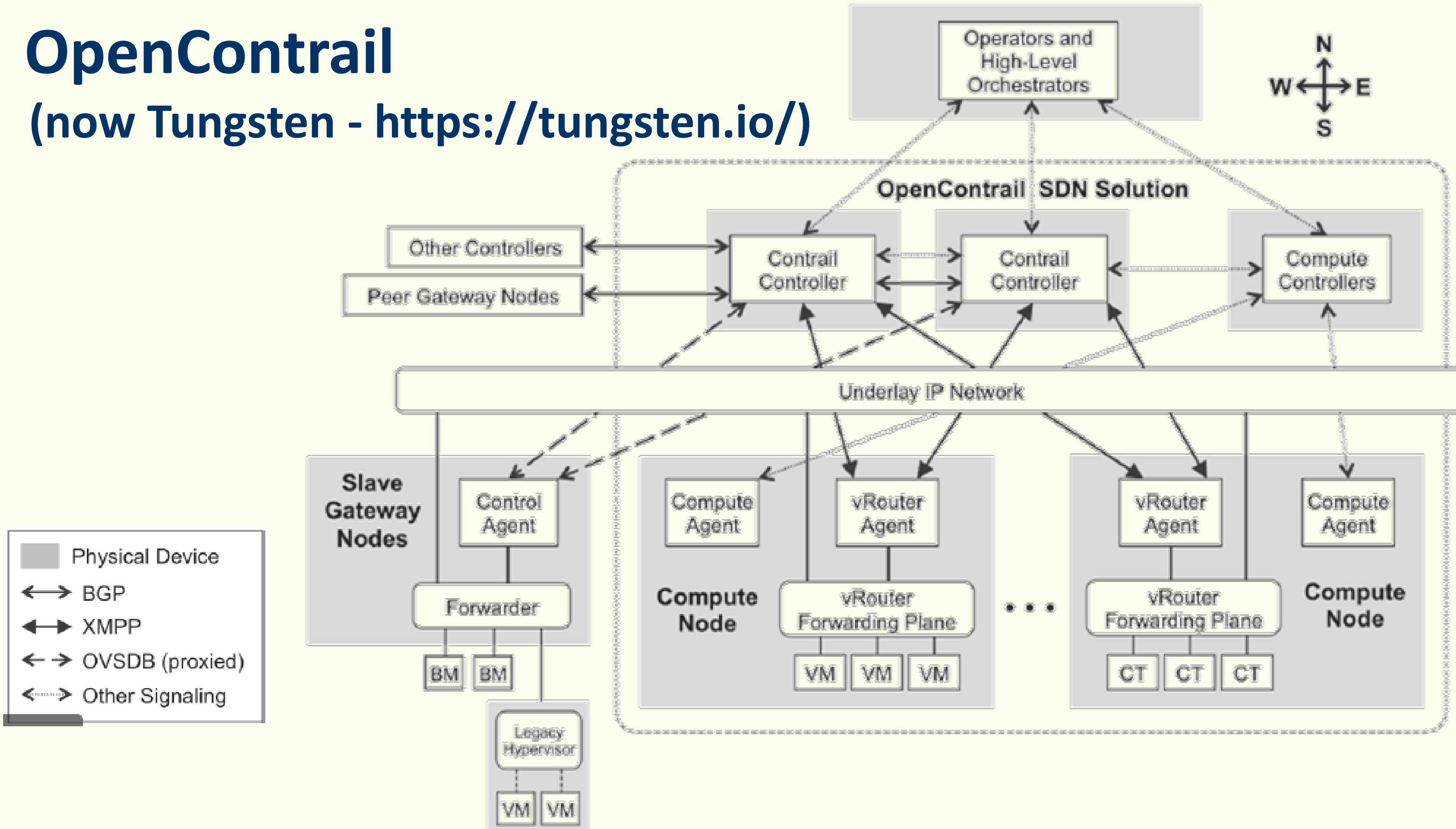


- Control plane: **controller**
  - Centralized: **SDN-based**
  - Protocol-based: **VXLAN + EVPN**

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)**