High Throughput Datacenter Network

Stefano Zani

INFN CNAF

WORKSHOP GARR 2021

NET MAKERS

09/11/2021

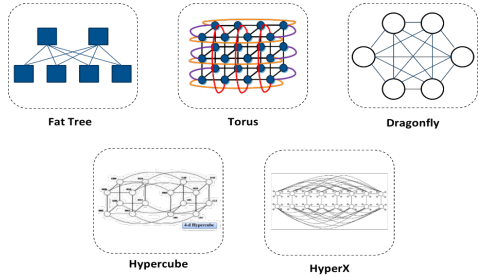
HTC

High Throughput Computing Center (Cosa significa?, Che rete serve?)

HPC→ Calcolo parallelo

Interprocess communication

Reti a bassa Latenza



HTC→ Job indipendenti

Accesso ad **alta velocita'** a grandi quantità di dati

Reti ad alto throughput

La rete di un Data Center HTC prende forma dal flusso dei dati che la attraversano analogamente a come un'auto di formula 1 viene plasmata dal flusso dell'aria.

Data Flow definito da:

Modelli computazionali degli esperimenti

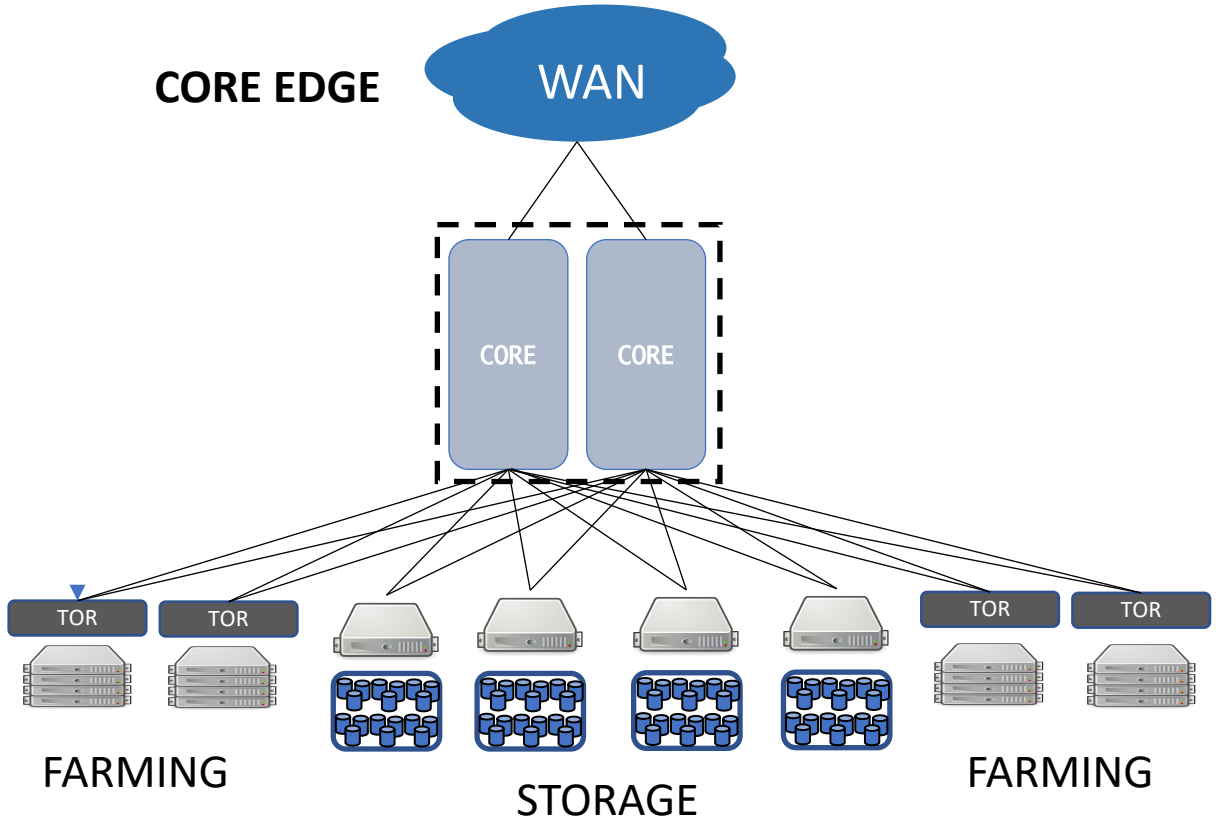
Soluzioni architetturali adottate per Farming (CPU) e Storage

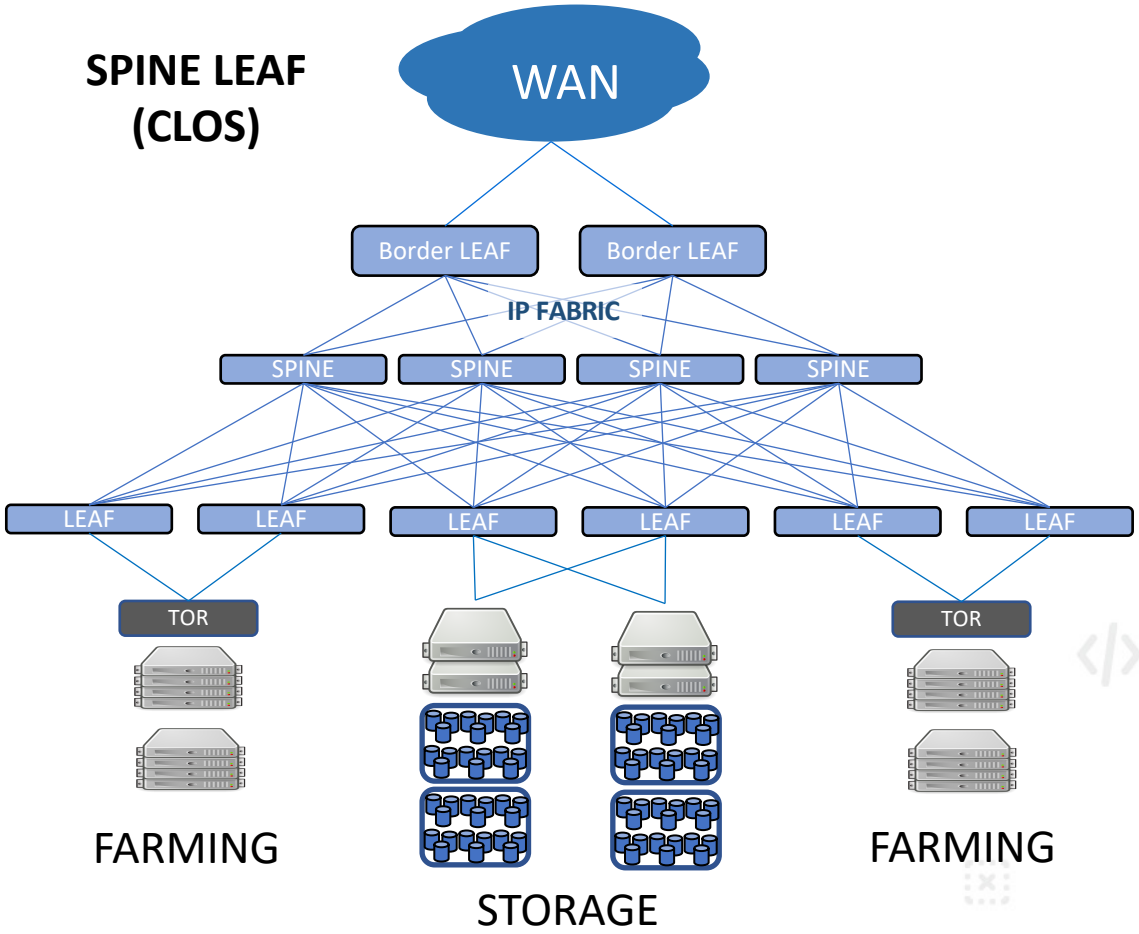
Parametri di un DC HTC che influiscono sul disegno della rete

- Caratteristiche fisiche del DC
  - Dimensioni
  - Eventuali **vincoli** legati a potenza elettrica e raffreddamento
    - Zone del DC: Isole a bassa o alta densità: condizionate in aria , Rear Door, DLC (Direct Liquid Cooling)
- Numero e tipo di CPU/GPU
- Quantità di spazio disco e nastro
- Tipologia di cluster filesystem utilizzato
  - Eventuali Storage Area Network
- Throughput interno CPU<->STORAGE (Accesso ai Dati)
  - Modello di accesso ai dati (Simulazione, Ricostruzione ed Analisi)
  - Matrice di accesso ai dati:
    - Tutte le CPU devono potere accedere a tutti i filesystem
    - Ci sono «Isole» distinte per specifici utenti
- WAN Throughput: Deve essere garantito il rate di trasferimento ed accesso ai dati in concorrenza con alle attività di elaborazione interna al DC
- Eventuali collegamenti ad alta velocità con altri Datacernter (DCI).

TOPOLOGIE

Possibili topologie considerate

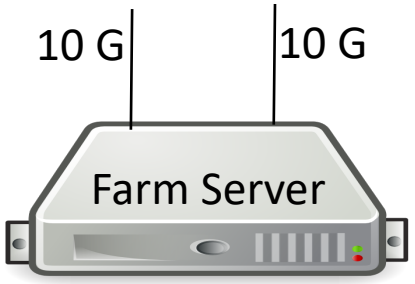




FARMING

Ipotesi di dimensionamento di rete per il Farming (CPU)

- CPU: 1 M HEPSPEC 06
  - Circa 50000 core fisici (100.000 Job slot)
  - Throughput xJob: 5MB/s
  - Banda teorica circa 4 Tbps
  - 2 TOR/Rack (a 2x100Gbps)=400G
  - Circa 13 Rack su un paio di «ROW»

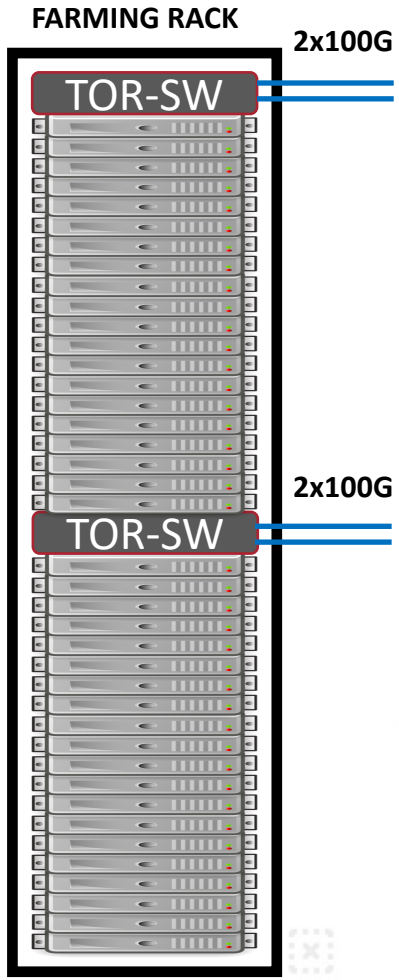


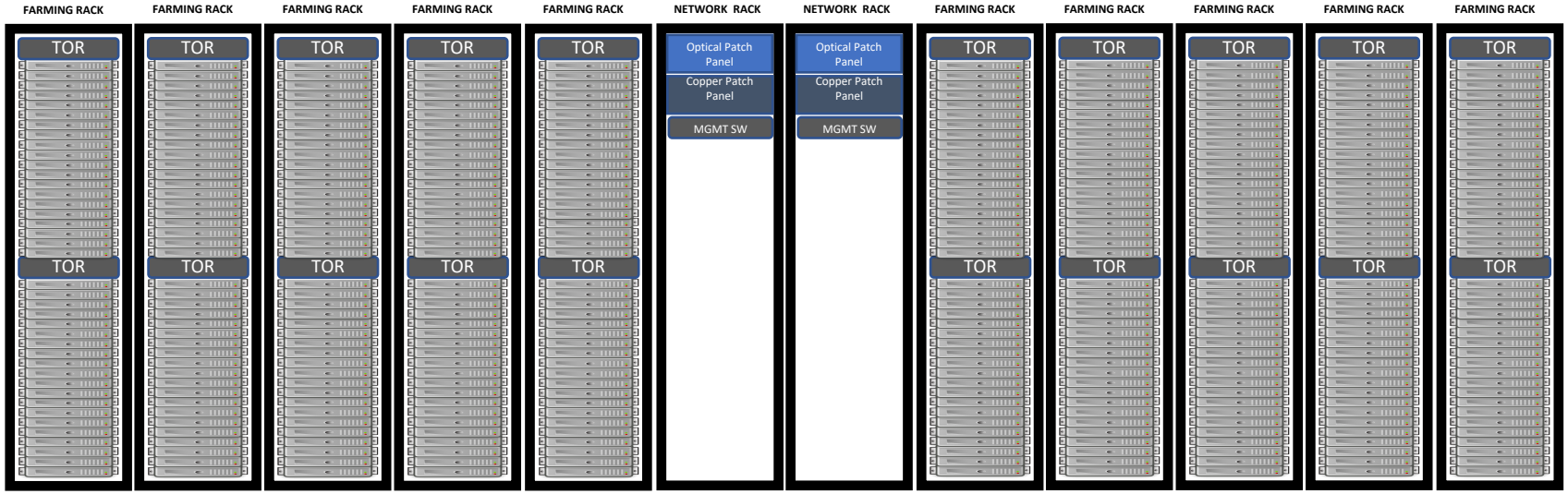
Consumo <40kw/Rack

2 Motherboard 1 Rack Unit

1 CPU per Motherboard

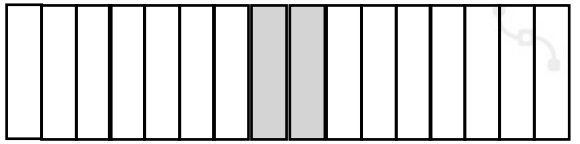
50 CORE/CPU→100Thread /CPU (Job)→4 Gbps per Motherboard-> 320Gbps/Rack





Farming ROW

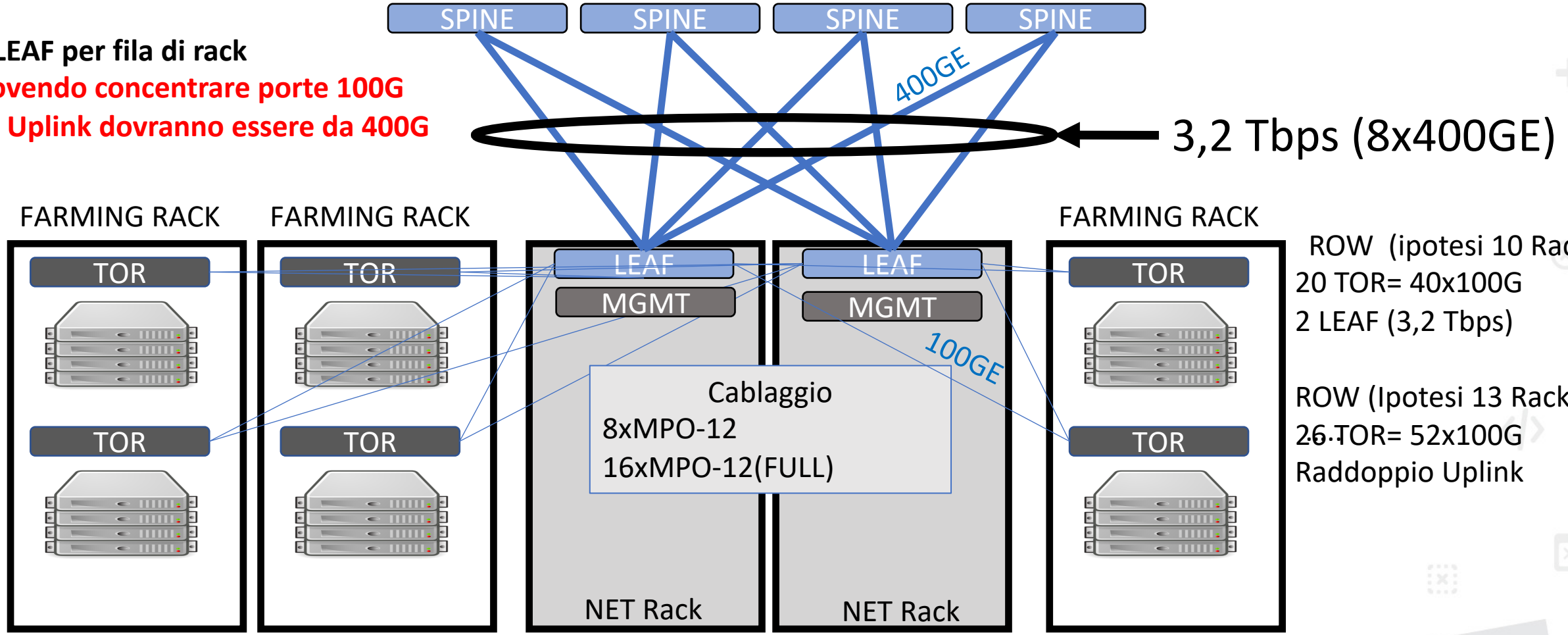
(Up to 14 Racks/row)



2 LEAF per fila di rack

Dovendo concentrare porte 100G

gli Uplink dovranno essere da 400G

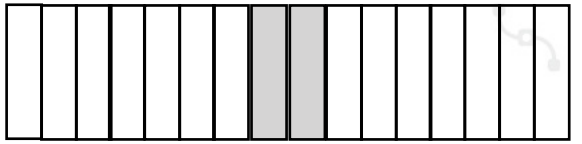


ROW (ipotesi 10 Rack)
20 TOR= 40x100G
2 LEAF (3,2 Tbps)

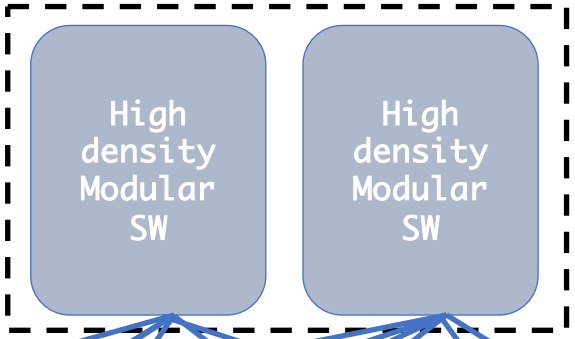
ROW (Ipotesi 13 Rack)
26 TOR= 52x100G
Raddoppio Uplink

Farming ROW

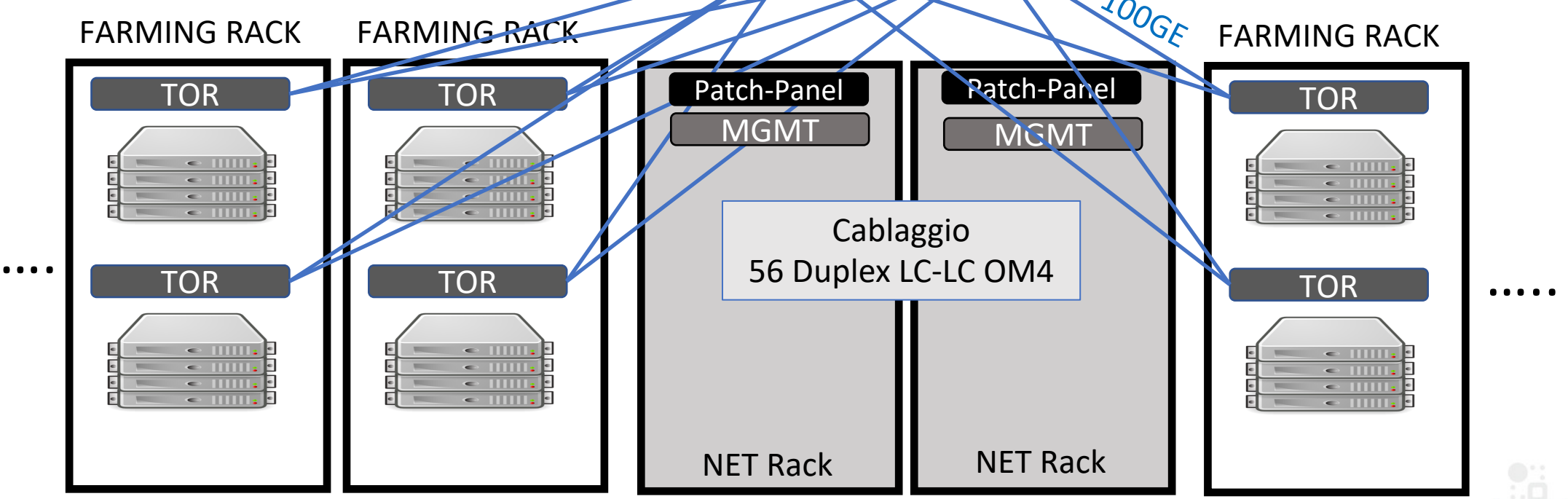
(Up to 14 Racks/row)



Maggiore numero di connessioni verso il CORE della rete ma di tipo LC su fibra MM (Commodity).
No oversubscription fra TOR e CORE



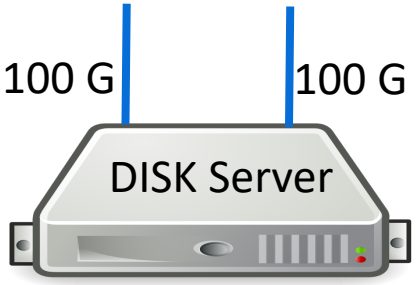
Impatto sui CORE(1000KHS06)
13 Racks (80 CPU/Rack)
Circa 60x100Gbps
Total: 2 Moduli (36Ports 100G)

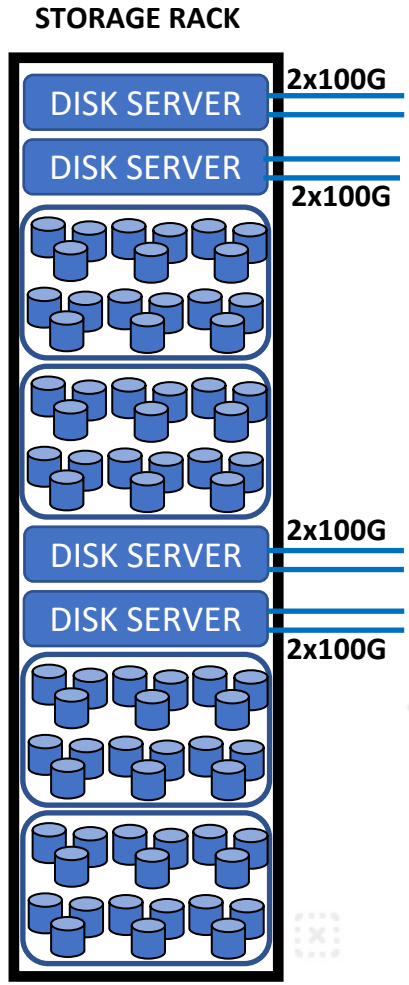


Ipotesi di dimensionamento di rete per lo Storage

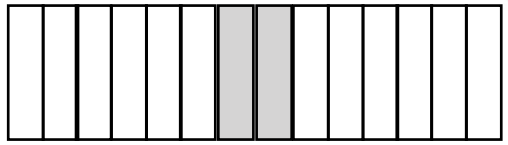
STORAGE (DISCO)120 PB

- Circa 60 Disk Server collegati a 2x100 Gbps
- Densità ipotizzata: 8 PB per Rack
- Ipotesi di dispiegamento di 15 Rack per il cluster Filesystem





Storage ROW



STORAGE RACK (8PB/Rack)

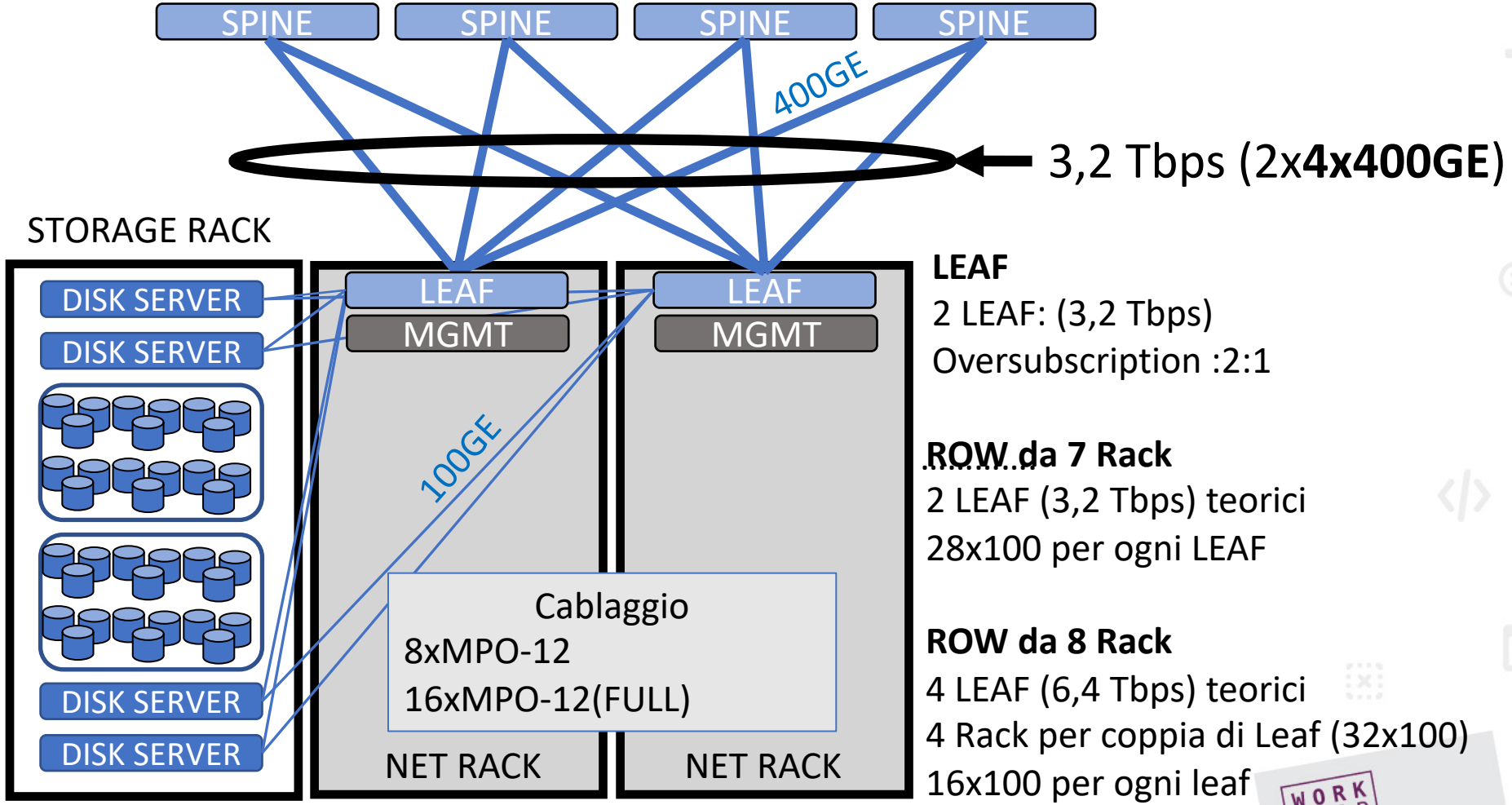
Storage totale:120PB
Banda netta/Rack:320 Gbps
(5MBps al TB)

4x DS/Rack
2X100G/Disk Server (Ridondanza)
8x 100G ports /Rack

.....

Installazione in 2 file
15 Racks: (8 Racks ROW1
+7 Racks ROW 2)

ROW1: 32 Servers (64x100G)
ROW2: 28 Servers (56x100G)



LEAF
2 LEAF: (3,2 Tbps)
Oversubscription :2:1

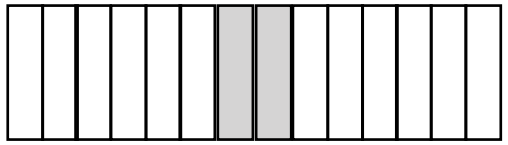
ROW da 7 Rack
2 LEAF (3,2 Tbps) teorici
28x100 per ogni LEAF

ROW da 8 Rack
4 LEAF (6,4 Tbps) teorici
4 Rack per coppia di Leaf (32x100)
16x100 per ogni leaf

Storage ROW

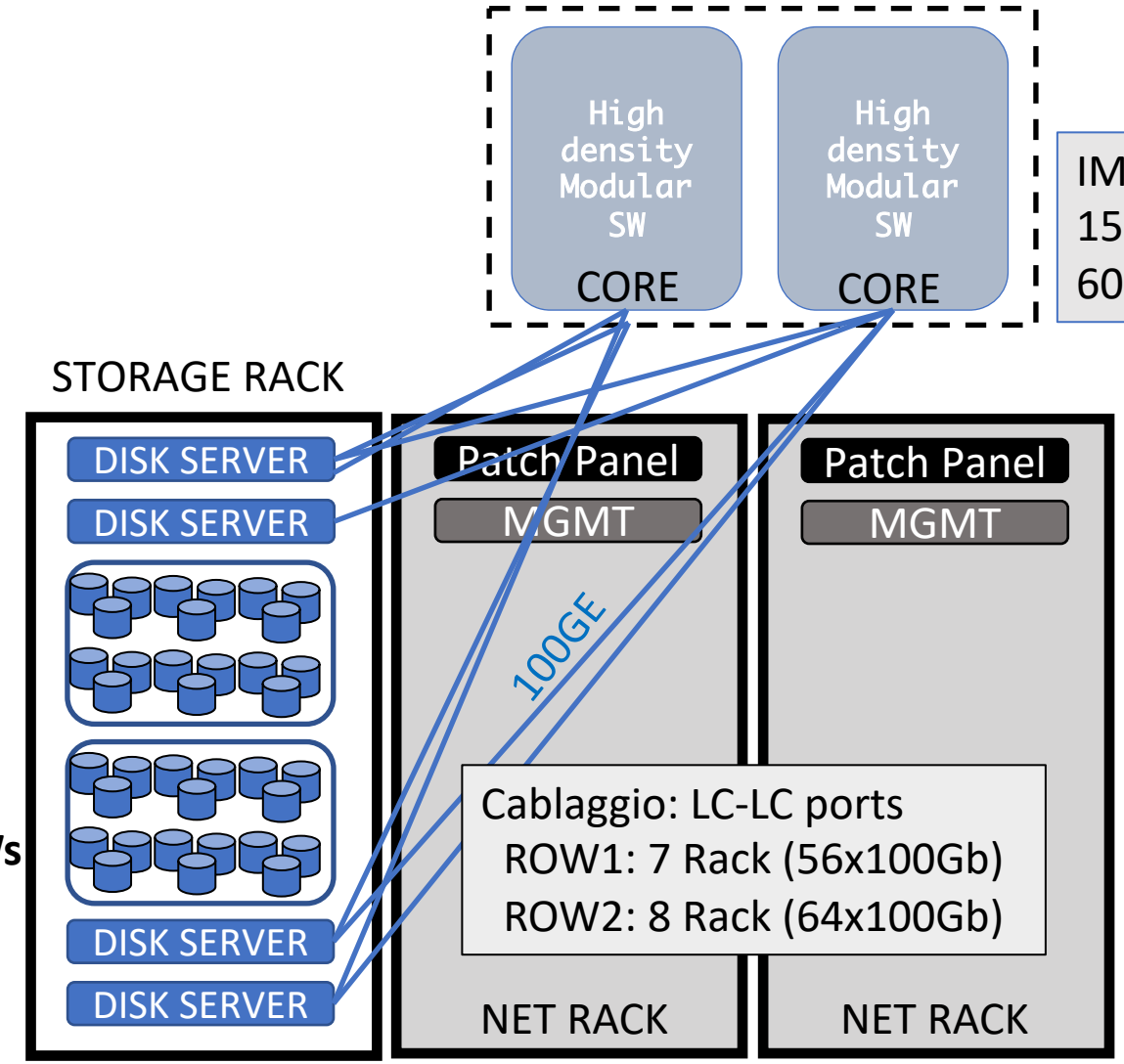
STORAGE Edge-CORE

UP to 12 Rack per Row



STORAGE RACK

Storage:120PB (8PB/Rack)
4x DS/Rack
2X100G/Disk Server (Redundancy)
8x 100G ports /Rack



IMPATTO SUI CORE
15Racks=15*8 100G Ports=(120 Ports)
60x100G ports per Chassis (2 modules)

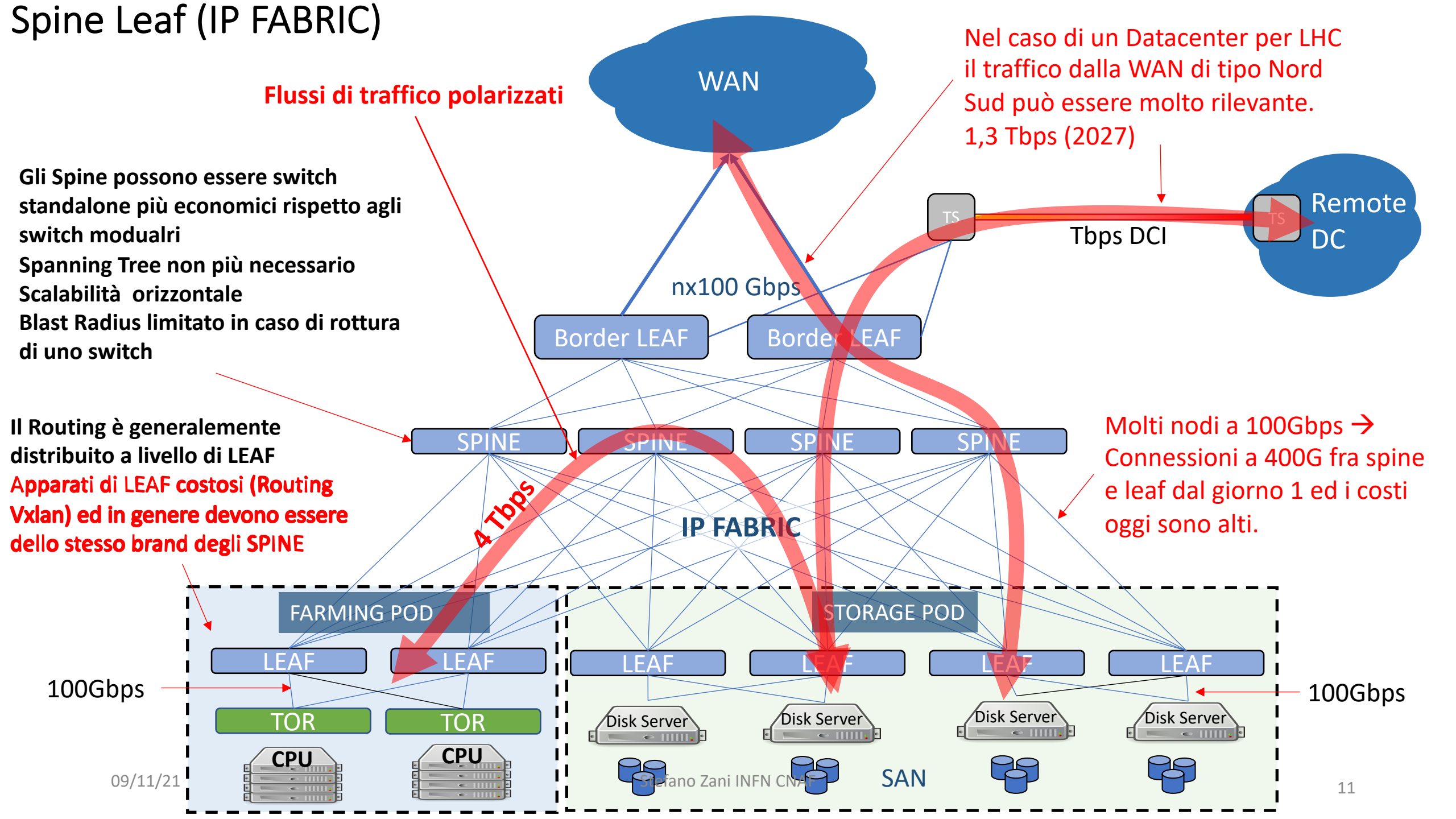
.....

.....

Cablaggio: LC-LC ports
ROW1: 7 Rack (56x100Gb)
ROW2: 8 Rack (64x100Gb)

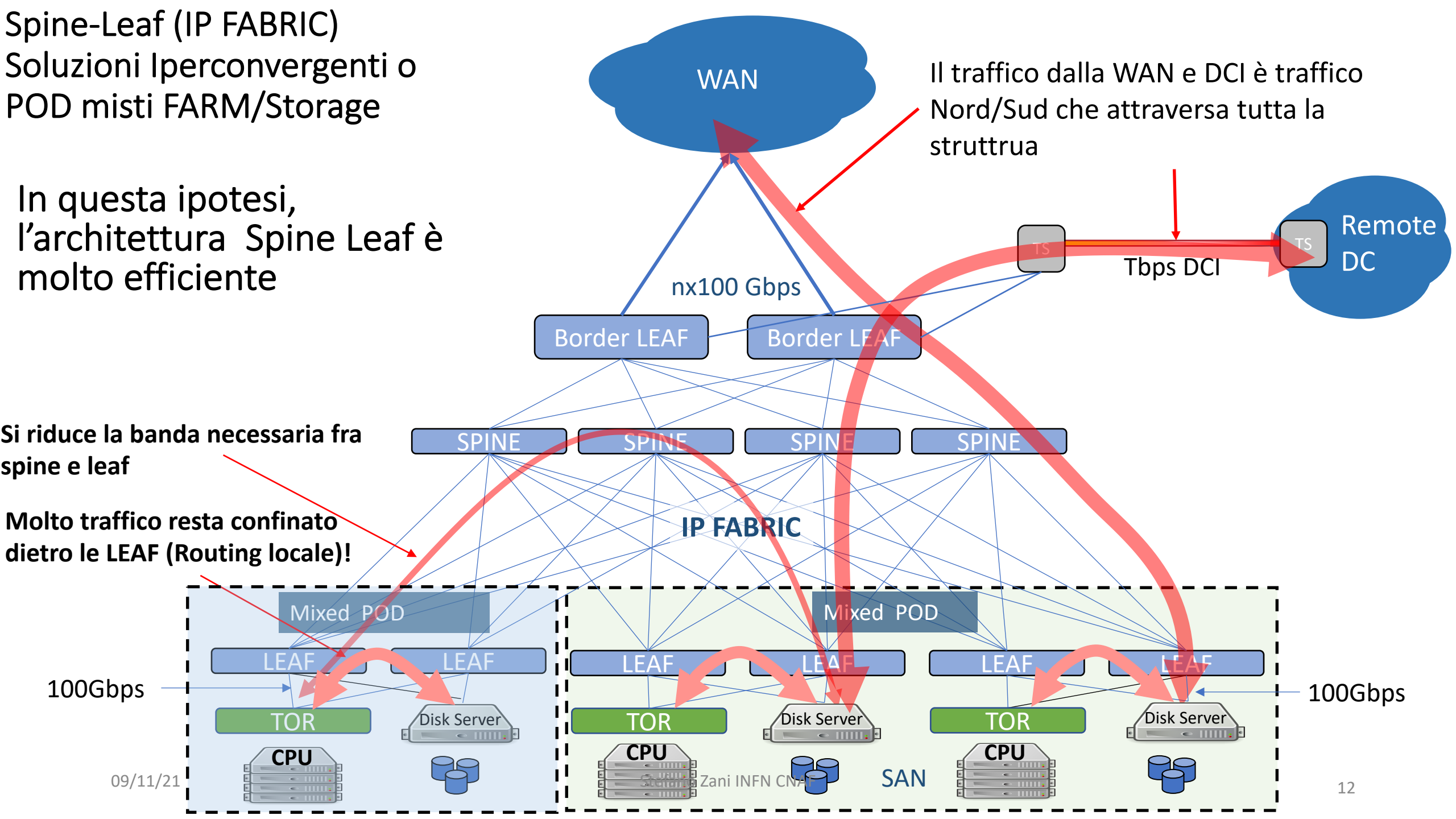
TOTAL INSTALLATION in 2 ROWs
15 Racks: (8 Racks ROW1
+7 Racks ROW 2)

Spine Leaf (IP FABRIC)



Spine-Leaf (IP FABRIC)
Soluzioni Iperconvergenti o POD misti FARM/Storage

In questa ipotesi, l'architettura Spine Leaf è molto efficiente



CORE EDGE

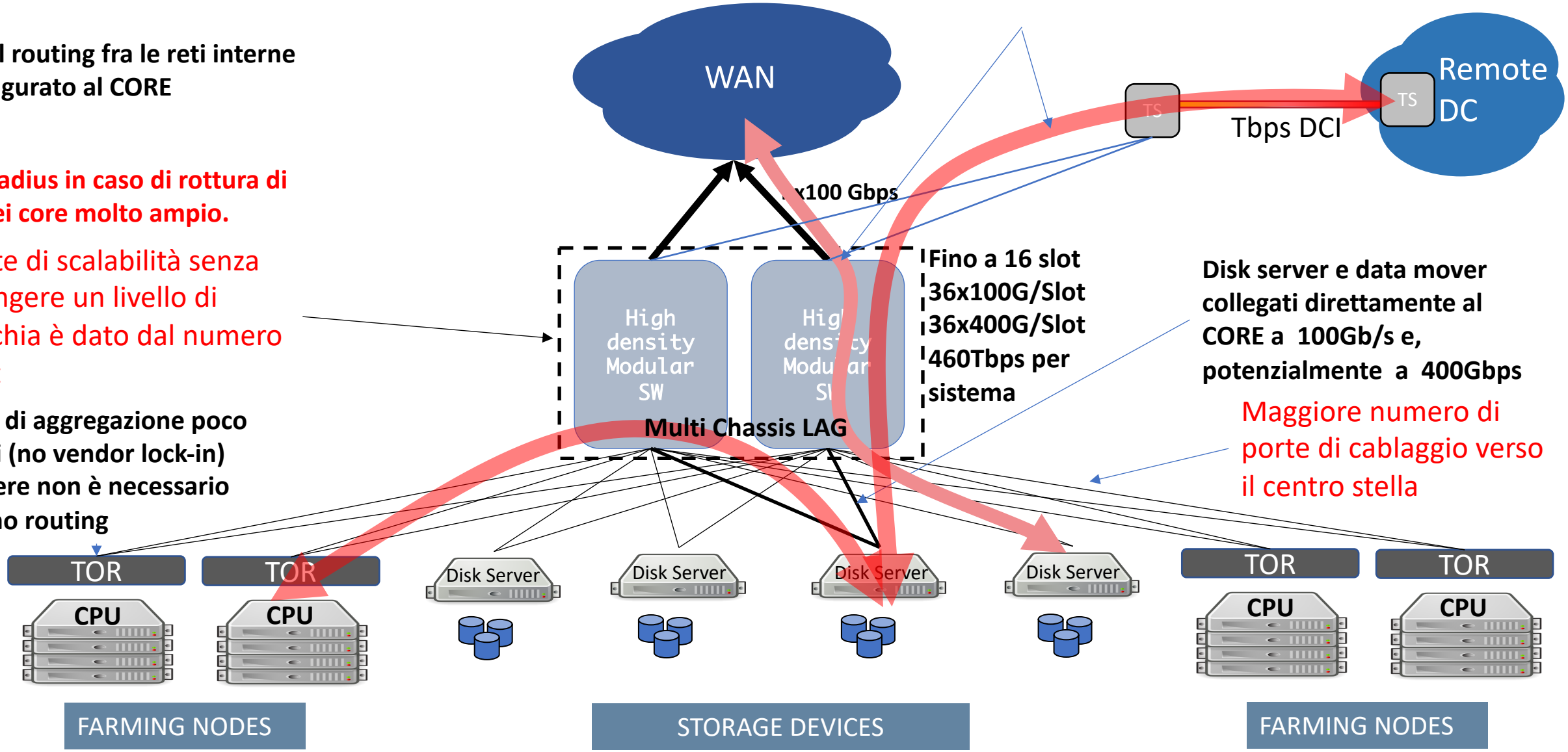
Traffico dalla WAN e DCI attraversano i CORE

Tutto il routing fra le reti interne È configurato al CORE

Blast radius in caso di rottura di uno dei core molto ampio.

Il limite di scalabilità senza aggiungere un livello di gerarchia è dato dal numero di slot

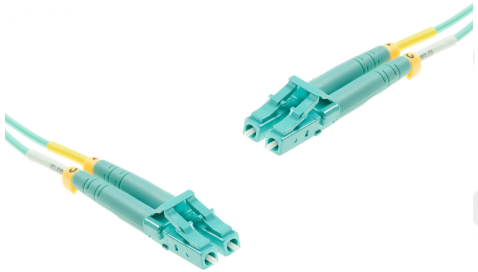
Switch di aggregazione poco costosi (no vendor lock-in)
In genere non è necessario facciano routing



Considerazioni sulla tecnologia 400G Ethernet

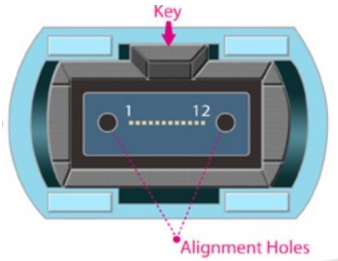
Tecnologie dei Transceiver e dei cablaggi sono molto importanti in termini economici per Datacenter di grandi dimensioni con centinaia o migliaia di transcaivers.

Cablaggi duplex sono molto più economici dei cablaggi MPO sia che si tratti di fibra monomodale o di fibra multimodale.



Tecnologie BiDi e SWDM4 supportano link **100GE link su una singola coppia di fibre multimodali** (Soluzione commodity per distanze < 100 m).

Sfortunatamente non c’è in roadmap nessuna soluzione su duplex multimodale per il 400 GE



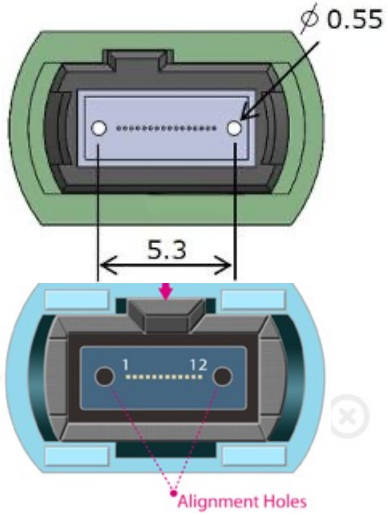
400G

Considerazioni sulla tecnologia 400G Ethernet

Transceiver	Tipo Fibra	Distanza	Connettore	Stima Costo
400GBASE-SR-8	Multi Mode 8 pairs	100m	MPO-16	(€) 9k€→0,6K€
400GBASE-SR-4.2	Multi Mode 4 pairs	100m	MPO-12	(€??) Q1 2022?
400GBASE-DR4-S	Single Mode 4 pairs	500m	MPO-12	(€€) 21k€→1K€
400GBASE-FR4	Single Mode 1 pair	2Km	LC	(€€€)14K€→1,6K€

MPO-16

MPO-12



Quale transceiver 400G short reach diventerà "Commodity" ?

400GBASE-SR 4.2: tecnologicamente dovrebbe costare meno di DR4 ma non è ancora disponibile sul mercato

400GBASE-DR4: La tecnologia è intrinsecamente più costosa ma è già acquistabile ed il prezzo potrebbe abbassarsi per logiche di mercato.

Oggi i listini sono fantascienza ed i costi reali non sono facilmente identificabili.

Automazione nella gestione della rete

Per reti con più di 100 switch, si rendono necessari strumenti di automazione.

ZTP (Zero touch provisioning): Installazione automatica di switch da zero

Netconf ed Ansible: Propagazione delle configurazioni su gruppi di switch in modo automatico

- E' necessario scrivere e mantenere un po' di codice
- Strumenti come Ansible o Puppet sono già utilizzati anche per la gestione dei nodi di calcolo (Stesso metodo di deployment)

Strumenti di gestione proprietari: Molto completi ma in genere limitati alla gestione degli apparati di un singolo vendor.

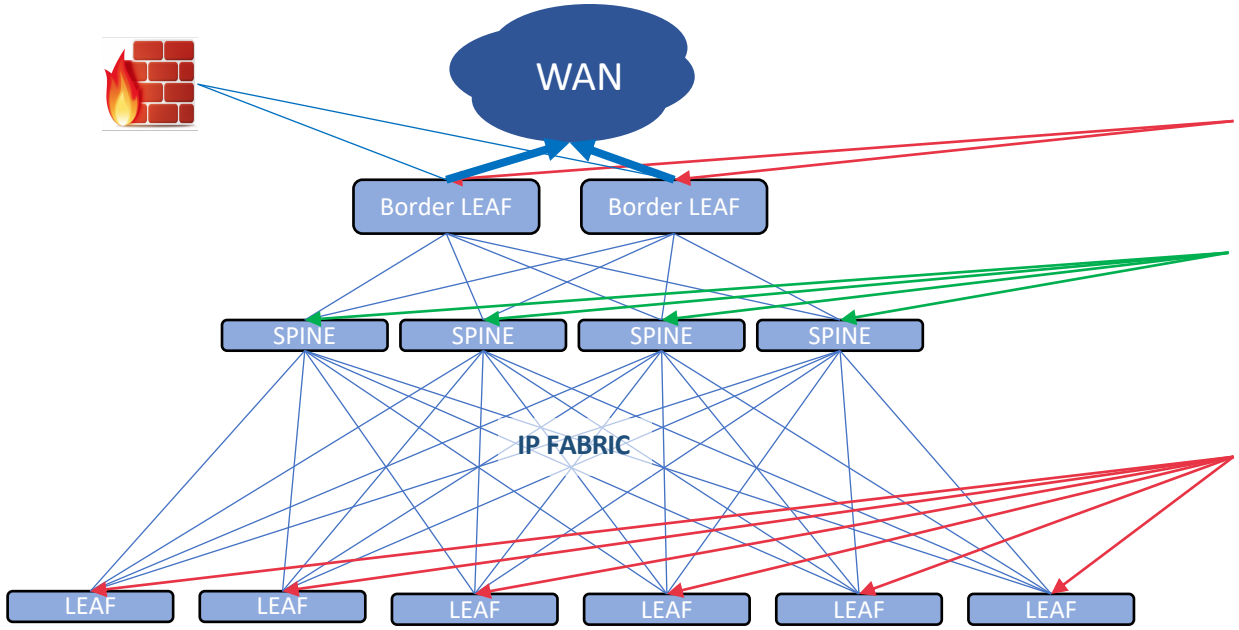
Automation

Gestione della complessita' soluzini IP Fabric

Topologie «closs» ed IP fabric

Gestione di diversi piani di routing Underlay, Overlay (protocolli di routing dinamici).

La necessità di distribuzione delle configurazioni di routing e di sicurezza a livello di leaf, suggerisce l'adozione di una piattaforma software di gestione spesso basata di fatto su un controller SDN.



Routing verso WAN, reti esterne, Firewall/loadbalancer...
ACL

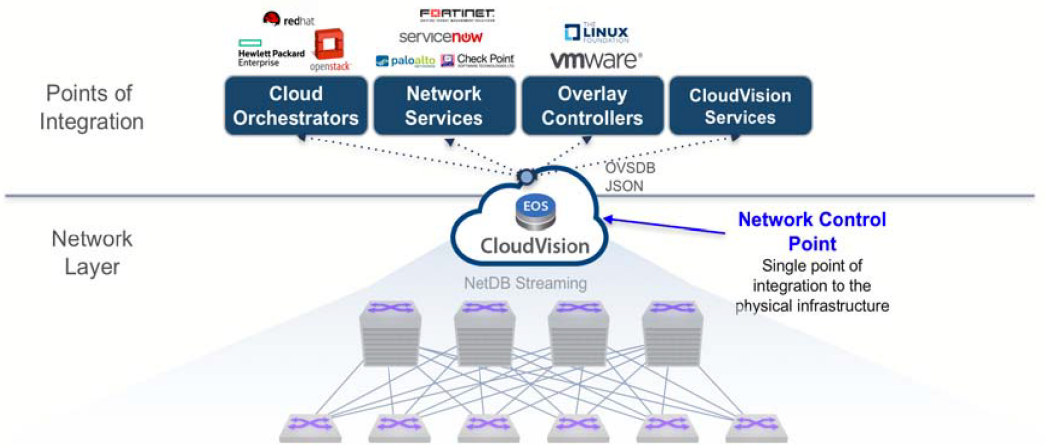
Routing dinamico della IP Fabric underlay

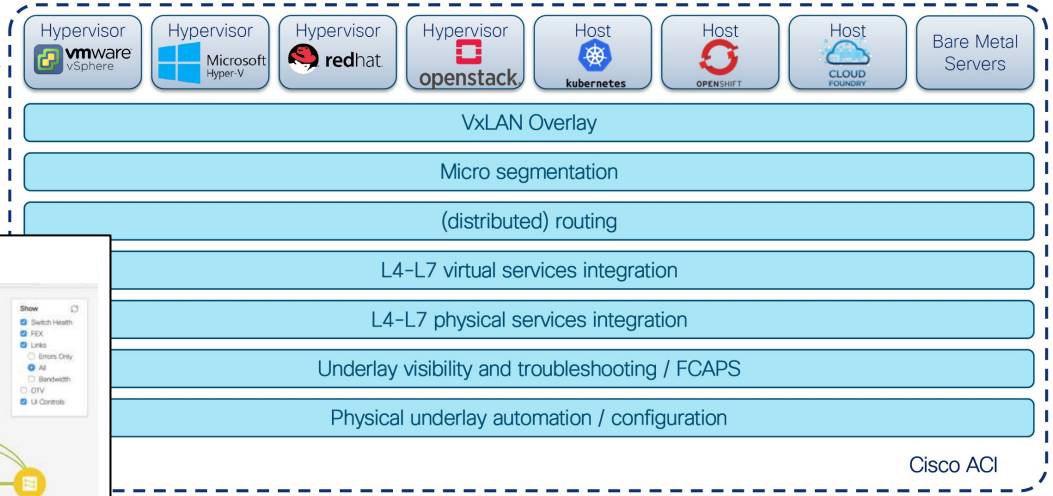
Routing delle reti interne replicato a livello di singola LEAF
(VXLAN EVPN routing)
ACL di sicurezza e segregazione

Automation

Alcuni strumenti avanzati (proprietari)

ACI , DCNM *(Cisco),* ***Cloud Vision*** *(Arista),* ***iMaster NCE*** *(Huawei)…*





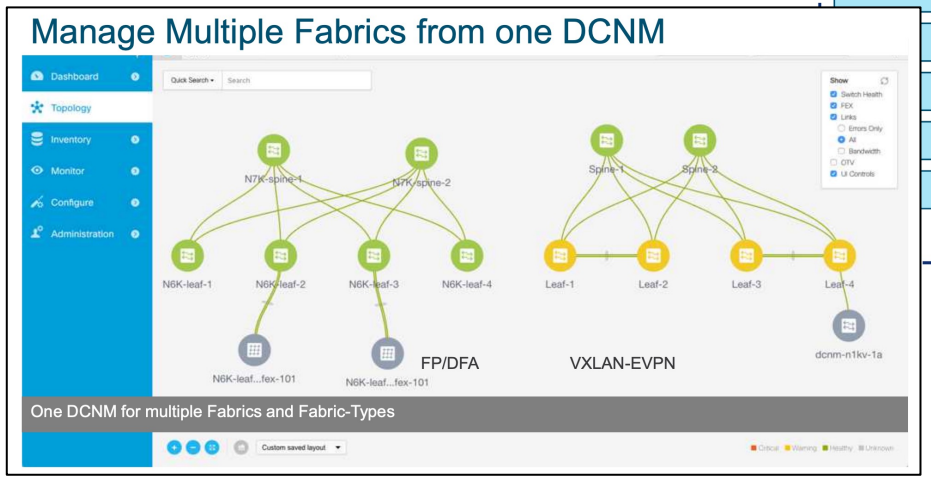
Vantaggi:

- Interazione via API (Con strumenti Open)
- integrazione con i sietemi di orchestrazione (vmware, Openstack, Qubernetes)
- Simulazione dei cambiamenti, deployment automatico delle configurazioni su tutta la struttura di rete, configuration rollback.
- Telemetria e reportistica automatica,

Svantaggi:

Costi ricorrenti

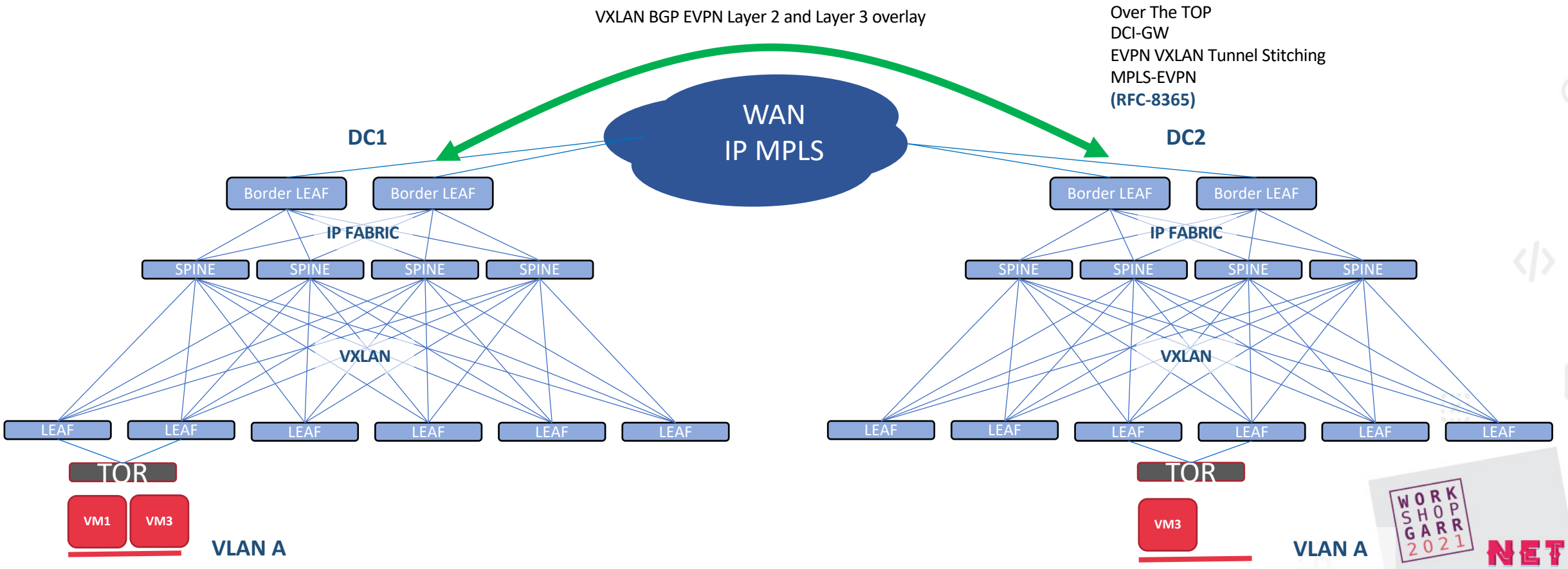
Vendor Lock in



Overlay Networks e DCI

VXLAN EVPN Multisite (MP-BGP)

La costruzione di reti overlay (L2 su L3) tramite l'utilizzo di VXLAN EVPN, aggiunge flessibilita' alla rete consentendo di estendere reti fra due DC utilizzando la normale rete IP come trasporto.



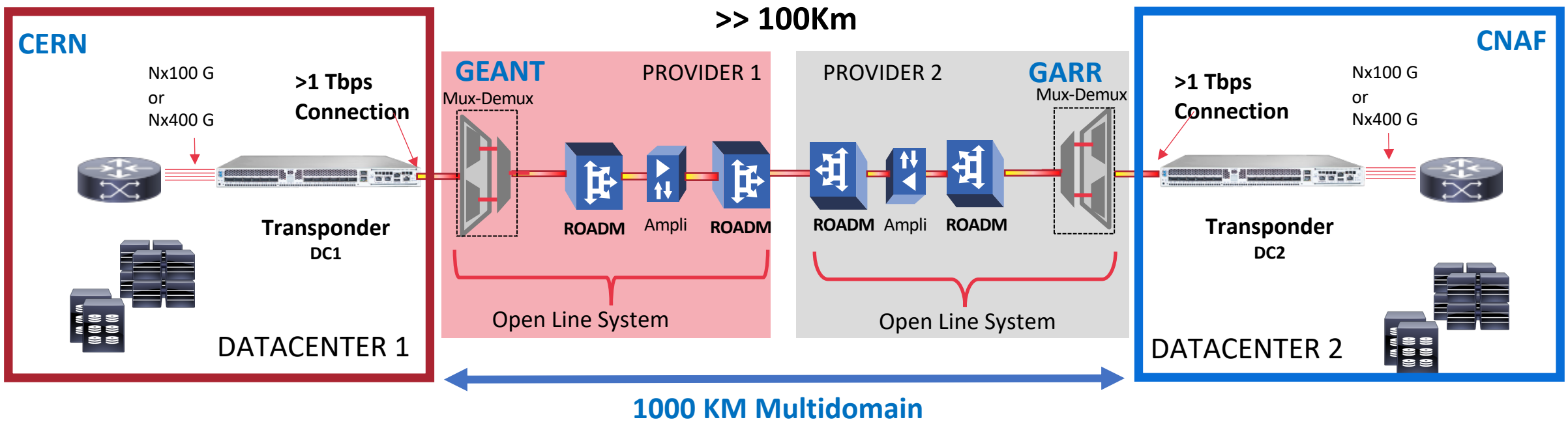
Datacenter Extension (DCI FISICO)

DCI ad alta velocità fra due datacenter relativamente vicini

Oggi esistono soluzioni per DCI a «breve» distanza anche integrate in apparati di rete che arrivano a velocità superiori ai 3,2 Tbps (Per esempio OSFP – Line System module di Arista)

Datacenter Extension (DCI FISICO)

DCI ad alta velocità fra due datacenter distanti > 100 KM



Attività di sperimentazione *(INFN, GARR, GEANT, CERN)* Sarting Soon!

Alcune considerazioni finali

NON c'è un modello di rete «Per tutte le stagioni»

Nella progettazione della rete di un centro HTC, giocano un ruolo fondamentale sia le dimensioni fisiche , sia le dimensioni dei flussi dati.

La tecnologia 400G Ethernet è ancora molto costosa a tutti i livelli: Porte, transceiver, cablaggio.

Indipendentemente dalla topologia scelta, il concetto di Overlay Network va implementato per la flessibilità e le possibilità che offre in termini di «Multi Tenancy»

Nella scelta di una soluzione di rete comoplessa, oltre alle performance del silicio, il software e gli strumenti di Network Management sono fondamentali.

Alcune considerazioni finali

ANSIBLE
Playbooks

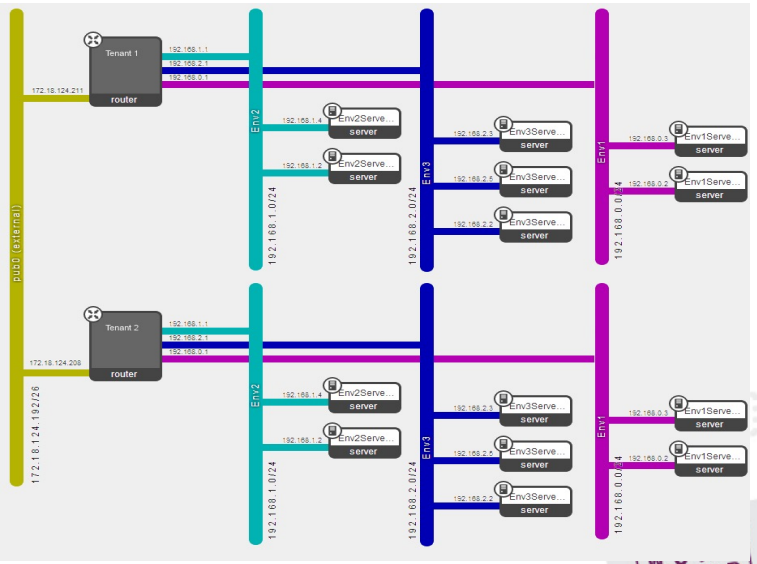
L'utilizzo di strumenti di automazione come Ansible o Puppet è necessario per questioni di scala.

```
---
- name: Run Arista command
  arista.eos.eos_command:
    commands: show ip int br
  when: ansible_network_os == 'arista.eos.eos'

- name: Run Cisco NXOS command
  cisco.nxos.nxos_command:
    commands: show ip int br
  when: ansible_network_os == 'cisco.nxos.nxos'

- name: Run Vyos command
  vyos.vyos.vyos_command:
    commands: show interface
  when: ansible_network_os == 'vyos.vyos.vyos'
```

L'interazione della rete con i sistemi di orchestrazione delle risorse di calcolo (come OpenStack, vmware) è un aspetto fondamentale da tenere presente e c'è ancora molto da fare.



Fine