Rete al Tecnopolo

(Aggiornamenti e roadmap)

CNAF Reloaded
11/2/2021

Stefano Zani

Per il Gruppo Rete del CNAF

Novità sul collegamento geografico del Tecnopolo

GARR ha già collegato ECMWF nell'area del Tecnopolo (Al momento lo ha collegato al POP Bo1 che si trova all'interno del CNAF)

Il POP Bo4 sarà realizzato in prossimità del Tecnopolo all'interno del Datacenter già esistente di Lepida (Il gestore della MAN regionale) distante meno di 1 Km *(Stima: Q1 2022)*

GARR Installerà cassetti ottici con raccordi in fibra ed apparati trasmissivi all'interno del nostro Datacenter in modo da potere attivare tutte le connessioni esterne senza dovere effettuare operazioni di posa ad ogni salto tecnologico.

Utilizzando queste fibre, si realizzerà il link DCI fra CNAF e Tecnopolo con una tecnologia simile a quella già in produzione tra CNAF e CINECA (Transponder based) e, ovviamente, si collegheranno tutti i link geografici verso la WAN.

Roadmap in termini di banda esterna ed interna al Datacenter

Roadmap connettività esterna:

WAN (LHC-OPN/ONE): 2x100 Gbps (2022) → 4x100Gbps (2023) →fino ad 800Gbps/1Tbps (2025 - 2027)

WAN (General Internet): **2x10 Gbps (2022) →4x10Gbps (2023) →**scalabilità potenziale fino a 2x100Gbps (2025 - 2027)

WAN (Cloud@INFN - Backbone): **1-2x10 Gbps (2022) →** scalabilità fino ad un fattore 10 in base alle esigenze (2025 - 2027)

DCI CNAF-TECNOPOLO (Transizione): **1.2 Tbps (12x100GbE)**

- 8-10x100 Gbps (T1)
- 2x100G (Estensione altri servizi)

DCI CNAF-LNL: 1-2x10Gbps (Secondo link auspicabile)

DCI IDDLS: Infrastruttura ottica programmabile a capacità variabile (Multipli di 100Gbps) fra CNAF, Bari e Roma (Prototipo di connettività per il DataLake italiano)

Connettività interna:

Throughput di accesso allo storage (120PB): **600 GByte/s** (60-80 server)

Throughput server farming del CNAF (70.000 core): **350 GByte/s** (2.8Tbps netti) – 60x100Gbps (Oversubscription 1:2)

Throughput nodi di calcolo di Leonardo (35.000 core): **175 GByte/s** (1.4 Tbps)

- Collegamento tramite Skyway (Probabilmente avrà porte a 100G)

Stima del numero di interconnessioni necessarie

- Esprimendo le necessità di throughput interno ed esterno in quantità di porte a 100Gbps, servirebbero circa **200 porte a 100Gb** sugli apparati di aggregazione per il collegamento delle risorse elencate nella slide precedente.

- Il numero e la tipologia delle porte si stabilirà in modo preciso appena avremo identificato **la densità delle installazioni e la distribuzione all'interno delle isole** (Esercizio da fare al più presto) in modo da determinare l'eventuale livello di over subscription che si intende tollerare per le varie installazioni.

- Il cablaggio strutturato, sarà concentrato in un'area dedicata alla rete al centro della sala. In ogni fila di rack saranno installati 2 rack di rete per ospitare i patch panel di distribuzione nelle isole e gli eventuali apparati attivi.

Architetture di rete

- I modelli di accesso ai dati del TIER1 si stanno a definendo con più precisione mentre sono in fase di definizione i modelli di deployment e gestione delle risorse di calcolo.
- Si stanno valutando architetture di rete che rispettino i requirement prestazionali e funzionali di WLCG che siano abilitanti nei confronti delle tecnologie di virtualizzazione (**VM** e **Container**) e gli strumenti di orchestrazione che si utilizzeranno.
- Si pensa di mantenere infrastrutture di rete distinte ma interconnesse fra TIER1 e gli altri contesti (SGSI, Cloud@CNAF, Servizi Nazionali, Sysinfo) per:
  - Ottimizzazione a livello prestazionale (Utilizzo dell'architettura più performante per ogni workflow)
  - Evitare che la propagazione di un problema o un errore di configurazione in un singolo contesto possa compromettere l'operatività di altri settori.
  - Segregazione dei contesti per migliore gestione della security ed autonomia operativa (SLA differenti)
- Gli apparati di rete infrastrutturali dovranno supportare tecnologie di network overlay (VXLAN Bridging e routing, EVPN)

Firewall e segregazione dei contesti

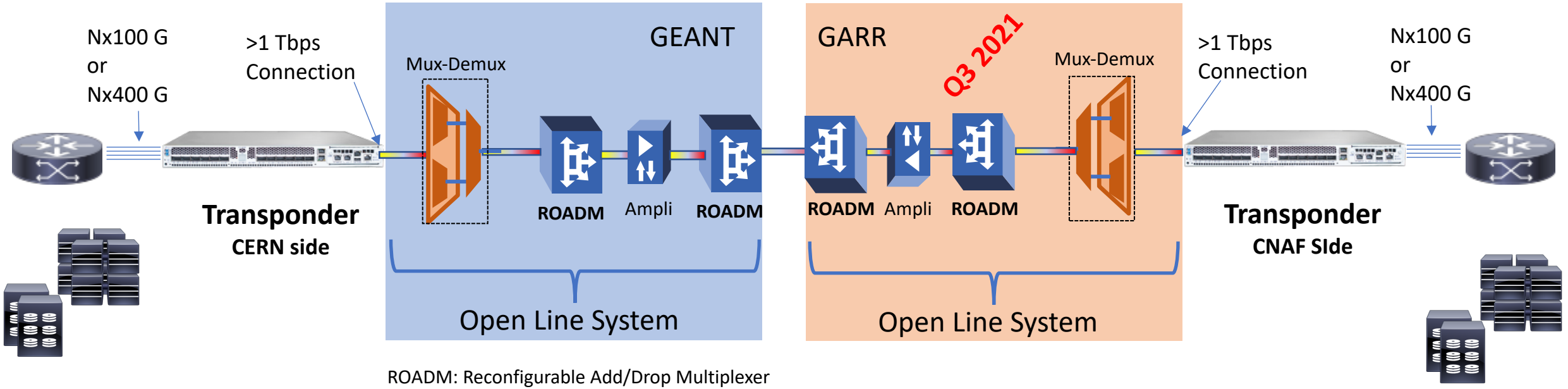
- Il next generation firewall di accesso alla rete General IP dovrà essere in grado di effettuare deep inspection e threat prevention in hardware seguendo l'aumento della banda di accesso e dovrà essere ridondato.
- Il dimensionamento andrà fatto accuratamente visti i costi molto elevati di questo tipo di apparati.
- I contesti virtuali e la multitenancy (supportati anche dalla attuale piattaforma) permetterebbero di utilizzare gli stessi apparati per proteggere diversi rami della rete senza dovere acquisire un firewall fisico per ogni ambito di applicazione.
- Si adotteranno, dove applicabile, soluzioni di Firewall as a Service (Necessita sperimentazione)

DCI fra CERN e CNAF (Attività di R&D)
Connettività per DataLake CERN-CNAF?

All'interno della collaborazione CERN-CNAF, l'8 Ottobre (https://indico.cern.ch/event/936998/) abbiamo proposto di avviare una attività di sperimentazione per implementare un collegamento dell'ordine del Terabit per secondo fra CERN e CNAF.

Packet/optical **transponder** all'interno dei datacenter connessi da una infrastruttura di trasporto di tipo **Open Line System** fornita da GARR+GEANT.

GEANT GN4 **Spectrum Connection Service (SCS).**

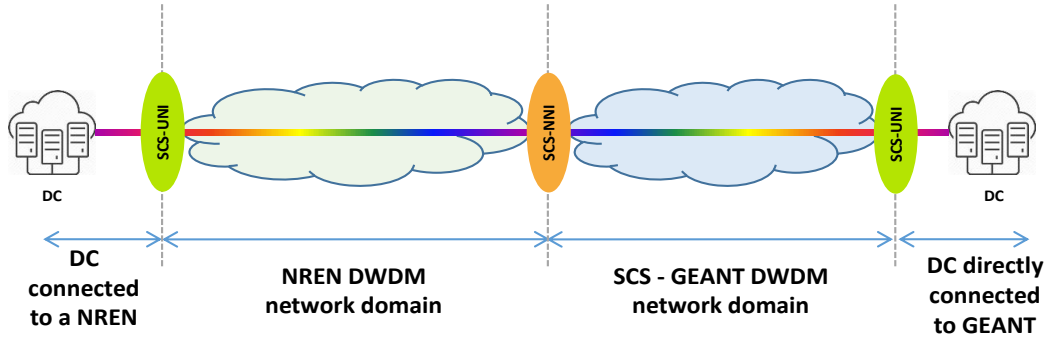


ROADM: Reconfigurable Add/Drop Multiplexer

GEANT CSC (Spectrum Connection Service) proposed usecase

Credit Guy Roberts (GEANT Senior Network Architect) TNC STF 26 Oct.2020

Generic Use Case: Data Centers Interconnection by SCS



- Easiest set-up with two photonic domains: one NREN and Géant
- Two Data Center as end-sites, providing DCI boxes to generate optical signal
- No 3R regeneration-> max 1000km path.
- Focus on APIs to provision, manage and monitor the connection
- Simulation tools test
- Candidate use case: CNAF (IT) Tier1 to CERN Tier0, through GARR and Géant SCS service providers

Technology tracking e POC (in preparazione dei tender)

Le attività di technology tracking stanno proseguendo e si stanno focalizzando sempre di più con il defnirsi dei requirement e dei workflow. Si svolgeranno le **attività di POC** con i principali produttori in modo da avere tutti gli elementi per la definizione dei tender per gli apparati attivi e per il cablaggio passivo.

Schedula delle attività propedeutiche alla preparazione dei tender (Le deadline si adatteranno in base alla data di consegna degli stabili)

- Raccolta dei requirement da Farm/Storage T1, SDDS, Servizi (SSNN, Sysinfo, Progetti) Febbraio 2021
- POC e test Luglio 2021
- Stima dei costi delle differenti architetture e scelta di quelle da adottare Luglio 2021
- Scrittura capitolati per i tender (Cablaggio ed apparati di rete) Settembre 2021

Stima del man power per la progettazione :circa **1,5 FTE su 5 persone (Tutti i componenti del reparto)** + Vincenzo Rega (nuovo collaboratore afferente al supporto utenti) + collaborazione da parte di altri reparti per test e verifiche sul supporto delle funzionalità richieste dai vari reparti.

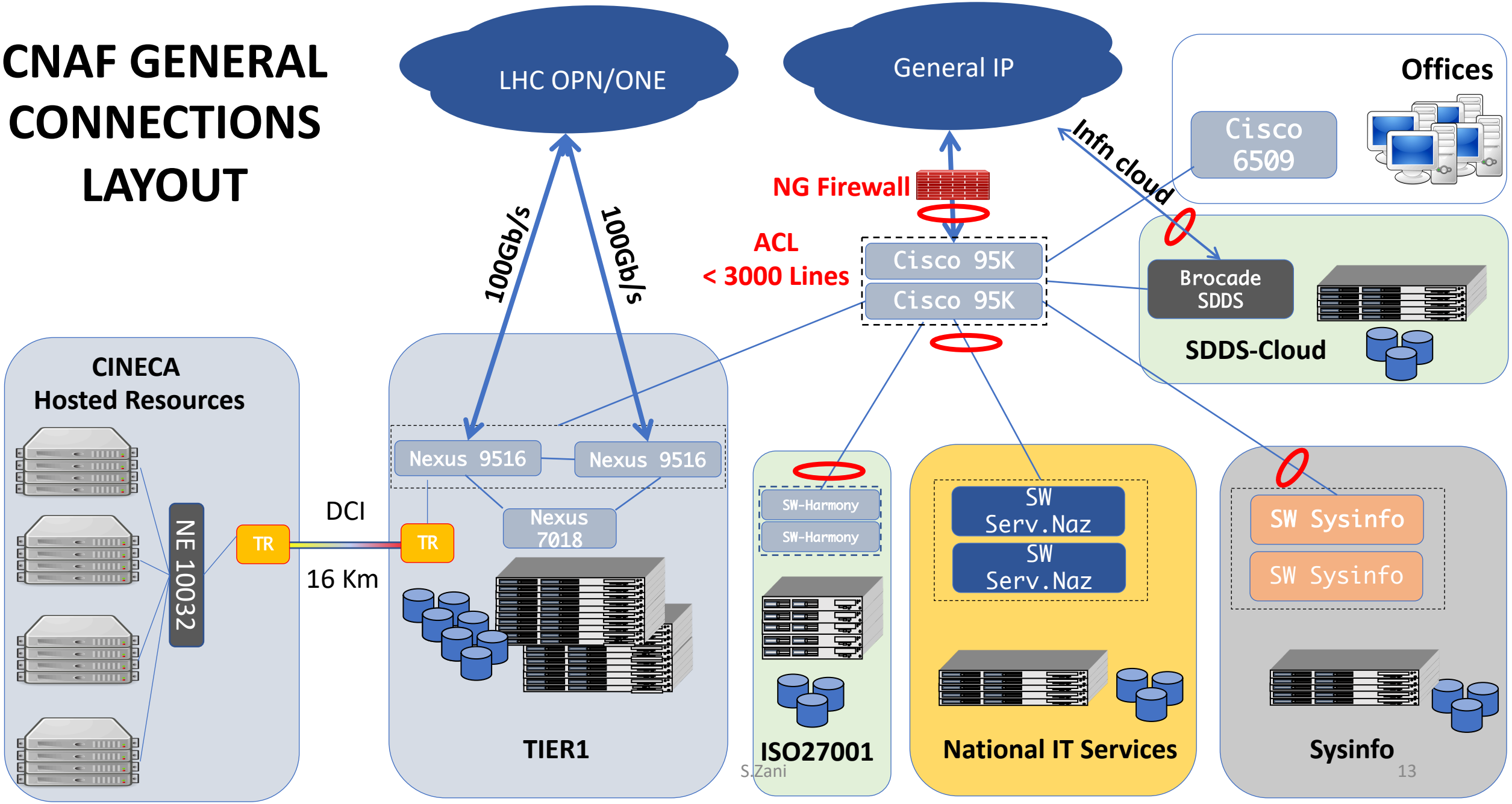
FINE

Backup SLides

Note sul cablaggio (Teck track sta proseguendo)

- Il cablaggio strutturato dovrà essere 400Gbit Ethernet ready.
  - Al momento si pensa ad una distribuzione di MPO-12 OM4.
  - Parte di questi MPO verranno collegati a moduli che renderanno disponibili porte in formato Duplex LC per connessioni a velocità inferiori.
  - Si sta valutando anche una piccola quantità di porte in rame (2 per rack) per la rete di management.

CNAF GENERAL CONNECTIONS LAYOUT



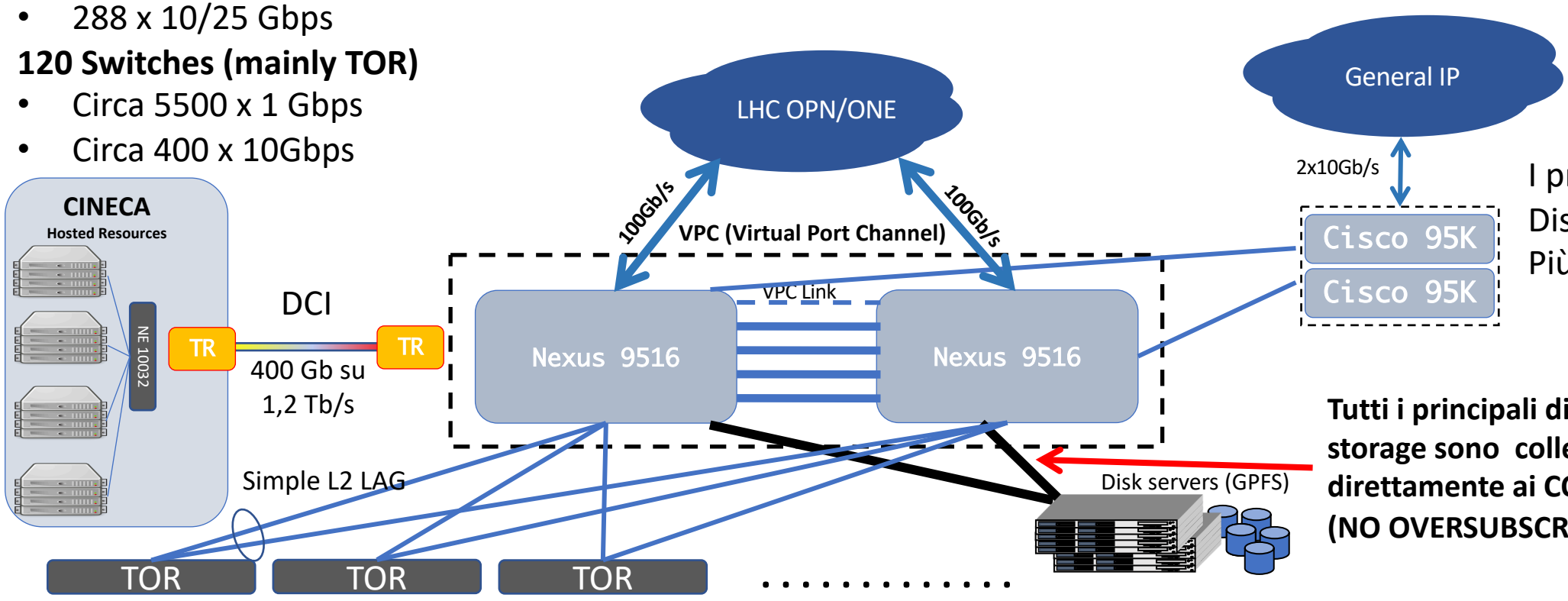
INFN CNAF TIER1 (Architettura attuale)

2 Core Switch Routers

- 152 x 100 Gbps (*64 in acquisto)*
- 288 x 10/25 Gbps

120 Switches (mainly TOR)

- Circa 5500 x 1 Gbps
- Circa 400 x 10Gbps



I principali flussi di dati:
Disk Server (GPFS) →WNs
Più di 600 Gbps sostenuti.

Tutti i principali disk server dello storage sono collegati direttamente ai CORE 2x 100Gb/s (NO OVERSUBSCRIPTION)

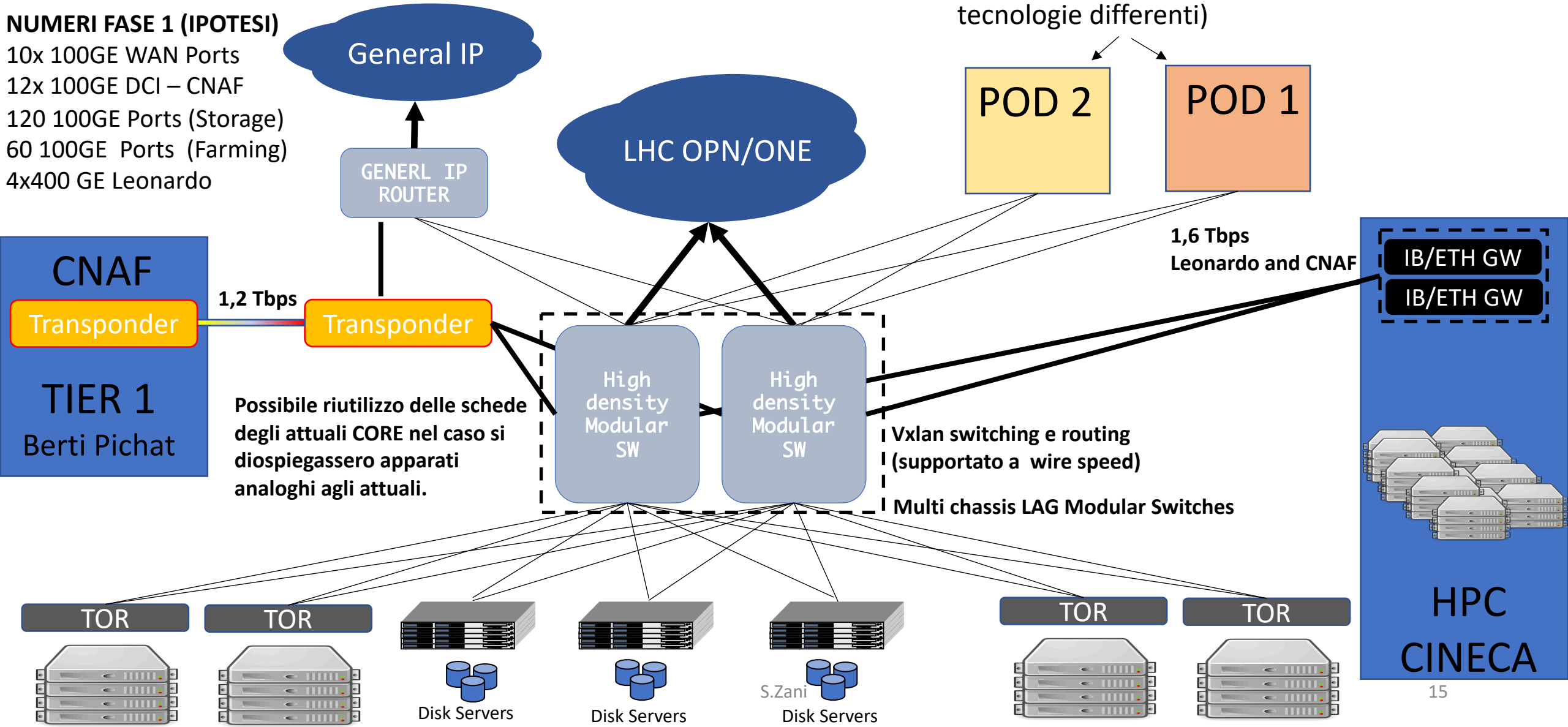
Farming Resources

Tutte le risorse sono collegate ad entrambi i Core Switch
I TOR switche devono gestire **solo le VLAN (802.1Q) – Routing al CORE** (Conf. Semplice e standard dei TOR) ci consente di potere utilizzare **switch di qualunque vendor** (Cisco, Juniper, Huawey, Lenovo, DELL, Extreme Networks, Super Micro).

Ipotesi di architettura (Topologia analoga alla attuale)

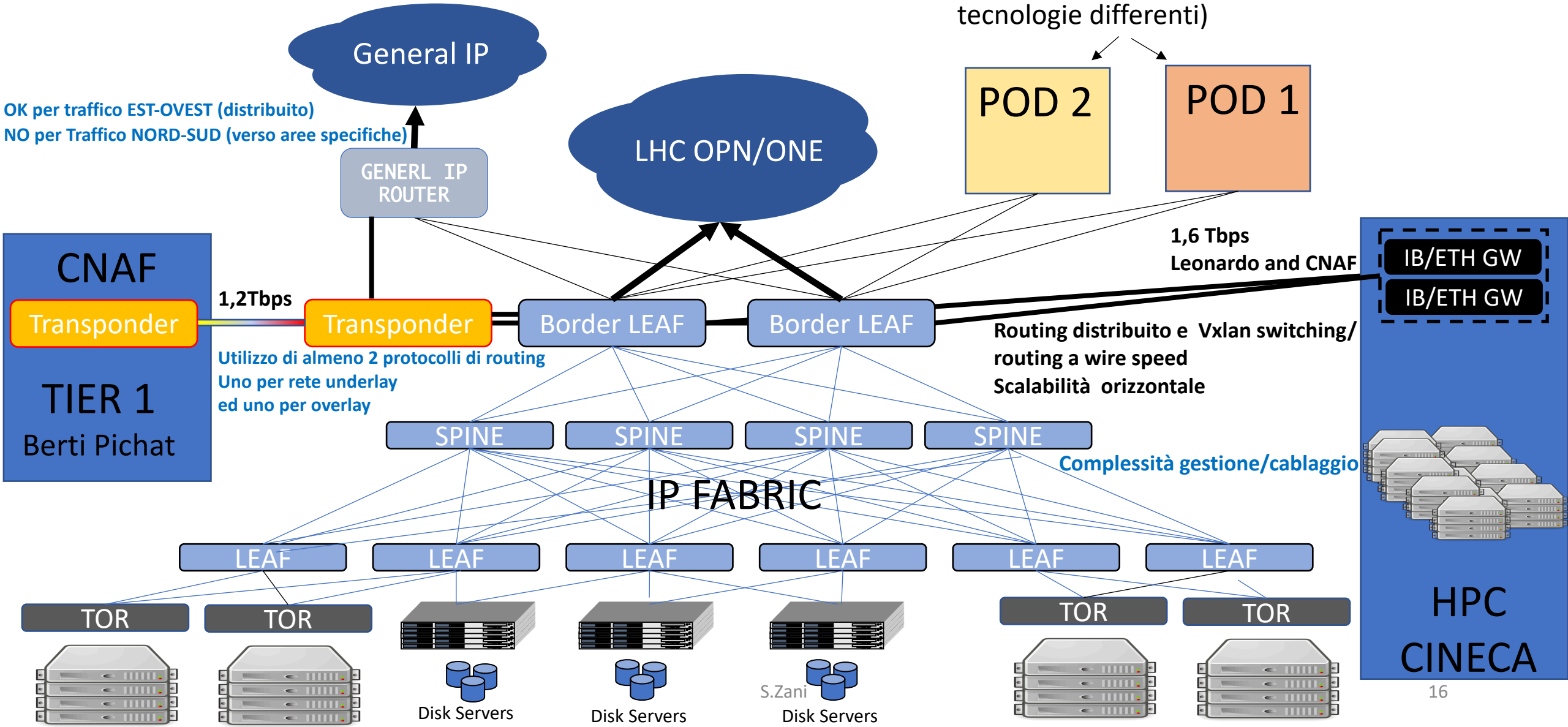
Infrastrutture di rete per servizi differenti (Cocentrazione di nodi a velocità diverse o con tecnologie differenti)

NUMERI FASE 1 (IPOTESI)
10x 100GE WAN Ports
12x 100GE DCI – CNAF
120 100GE Ports (Storage)
60 100GE Ports (Farming)
4x400 GE Leonardo

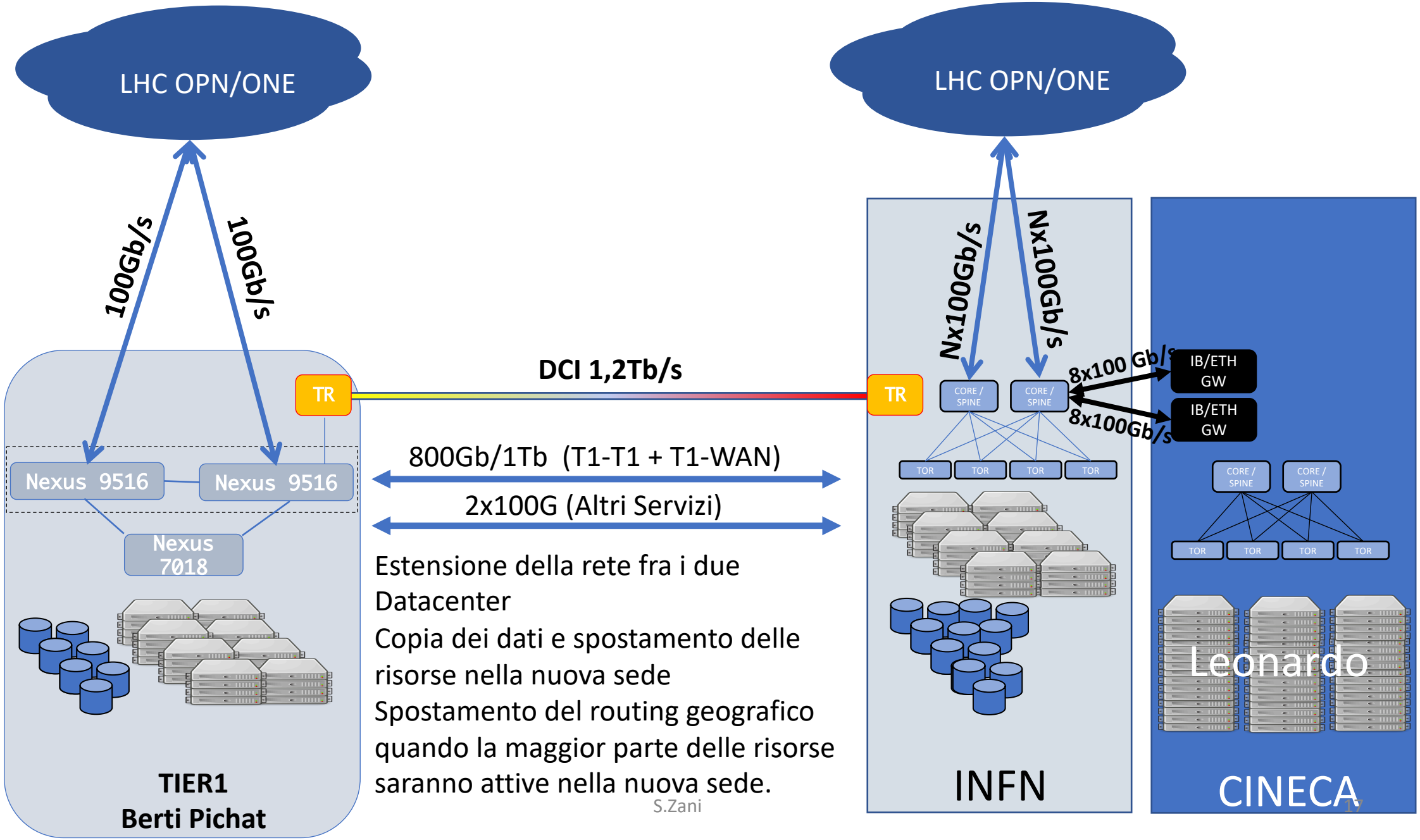


Ipotesi di architettura (Spine-Leaf)

Infrastrutture di rete per servizi differenti (Cocentrazione di nodi a velocità diverse o con tecnologie differenti)



Interconnessioni di rete per la transizione

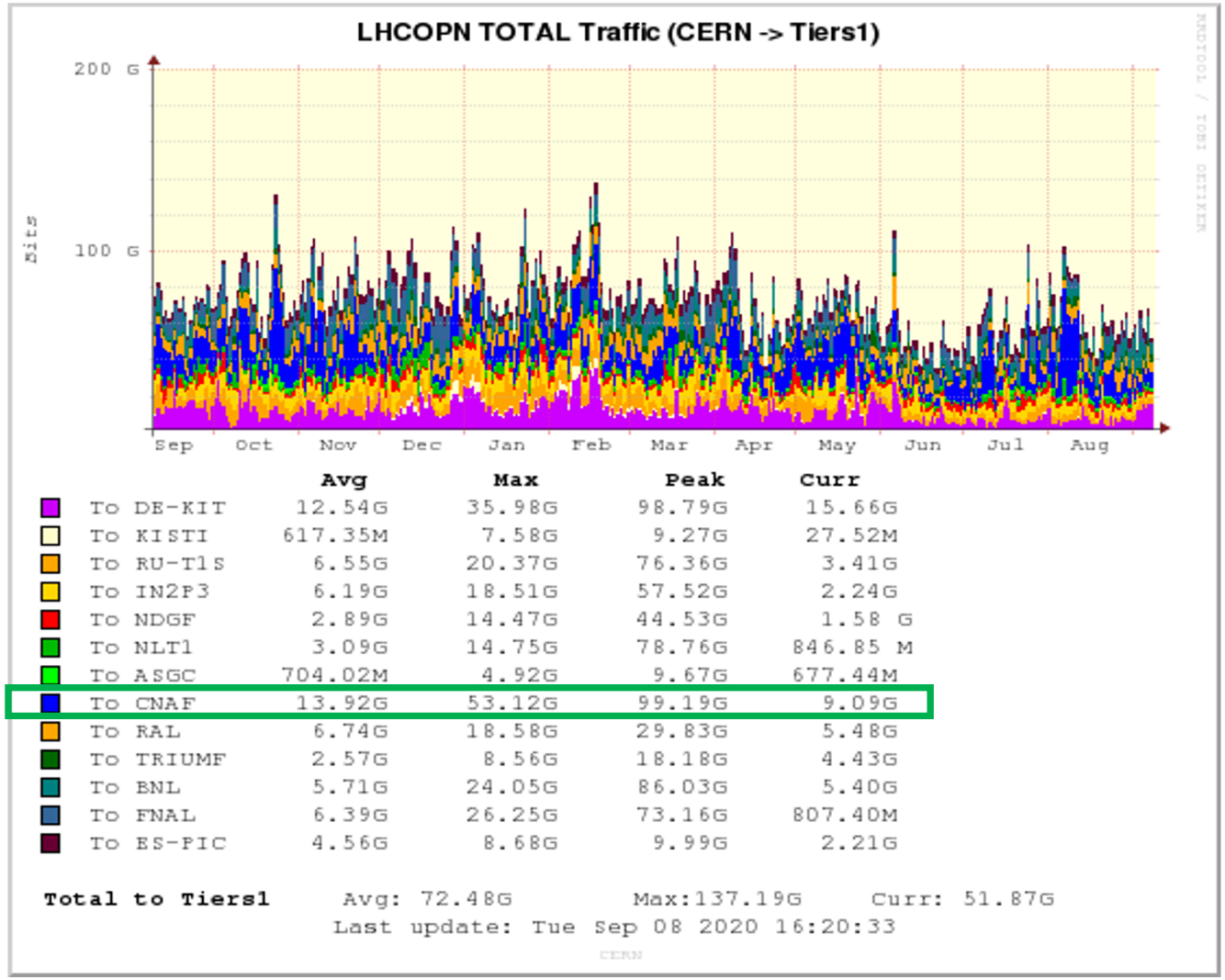


Estensione della rete fra i due Datacenter

Copia dei dati e spostamento delle risorse nella nuova sede

Spostamento del routing geografico quando la maggior parte delle risorse saranno attive nella nuova sede.

CNAF TIER1 9/2020



	Avg	Max	Peak	Curr
To DE-KIT	12.54G	35.98G	98.79G	15.66G
To KISTI	617.35M	7.58G	9.27G	27.52M
To RU-T1S	6.55G	20.37G	76.36G	3.41G
To IN2P3	6.19G	18.51G	57.52G	2.24G
To NDGF	2.89G	14.47G	44.53G	1.58 G
To NLT1	3.09G	14.75G	78.76G	846.85 M
To ASGC	704.02M	4.92G	9.67G	677.44M
To CNAF	13.92G	53.12G	99.19G	9.09G
To RAL	6.74G	18.58G	29.83G	5.48G
To TRIUMF	2.57G	8.56G	18.18G	4.43G
To BNL	5.71G	24.05G	86.03G	5.40G
To FNAL	6.39G	26.25G	73.16G	807.40M
To ES-PIC	4.56G	8.68G	9.99G	2.21G

Total to Tiers1 Avg: 72.48G Max:137.19G Curr: 51.87G

Last update: Tue Sep 08 2020 16:20:33

https://netstat.cern.ch/monitoring/network-statistics/ext/?q=LHCOPN&p=LHCOPN&mn=00-Total-Traffic&t=Yearly

Some considerations

- Need to better understand if the Data flow will be compatible with an IP Fabric Spine Leaf solution (In this moment we have more north-south traffic than east-ovest). How the Storage will be “Served” and where the Computing Resources will be located can impact the network design.
- Network management and automation tools will be essential
  - An integration with the orchestration tools used for VM deployment is desirable
  - Possibly avoid vendor lock in solutions.
- More Techtrack activity with vendors like Arista, Mellanox, Juniper, Cisco and cabling vendors will be done.