DCI e SD-WAN

Stefano Zani (INFN CNAF)

Workshop GARR

Roma, 29-31 Maggio 2018

Data Center Interconnection (Casi d'uso)

- **Datacenter Extension**
- **Business Continuity**

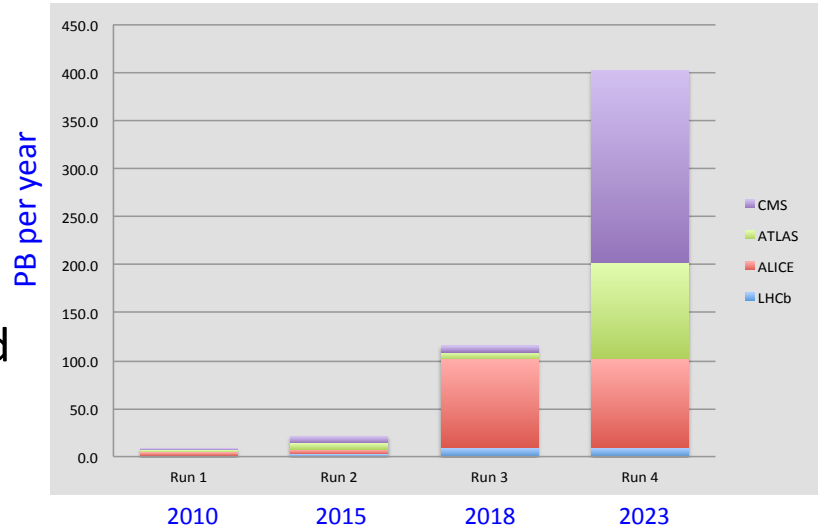
Interesse specifico per INFN:

La quantità di dati che verranno prodotti da LHC e dovranno essere analizzati, nei prossimi RUN subirà un aumento esponenziale (ad investimento costante).

→Occorre utilizzare tutte le risorse disponibili ovunque si trovino.

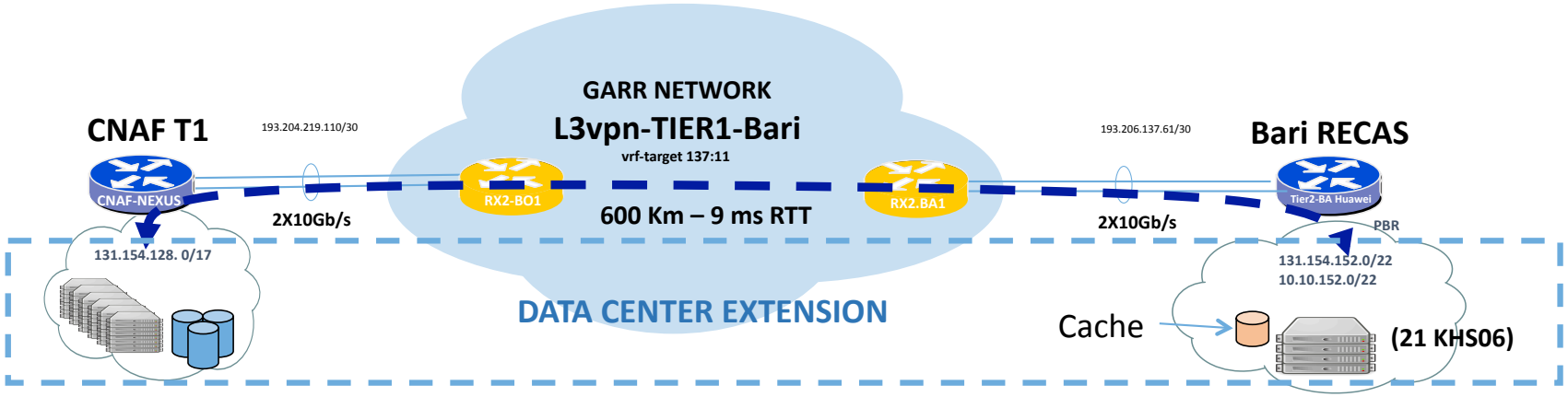
DCI realizzati negli ultimi anni:

- Estensione del TIER1 su altri centri di calcolo scientifico gestendo tutte le risorse di calcolo con lo stesso batch system e con accesso ai medesimi dataset.
- Cloud Bursting: Test sull' utilizzo di risorse di cloud provider pubblici per assorbire picchi di richiesta di CPU.



Estensione TIER1 su nodi di ReCas Bari (dal 2016)
Overlay (L3 VPN) con banda dedicata

Caso d'uso: 21 KHS06 di RECAS BARI ad uso degli esperimenti LHC come estensione del TIER1 (CNAF Bologna)
40 nodi da 64 Core (AMD Opteron), 256 GB RAM
Banda teorica ideale verso lo Storage : Circa 100 Gb/s (Throughput 5MB/s per core).
Banda disponibile per il DCI:20 Gb/s (L3 VPN del GARR) su 2x10Gb/s (Domini di broadcast separati)
I Nodi di Bari hanno IP del CNAF e sono **installati e gestiti come gli altri WN del CNAF** (stesso batch system del TIER1).



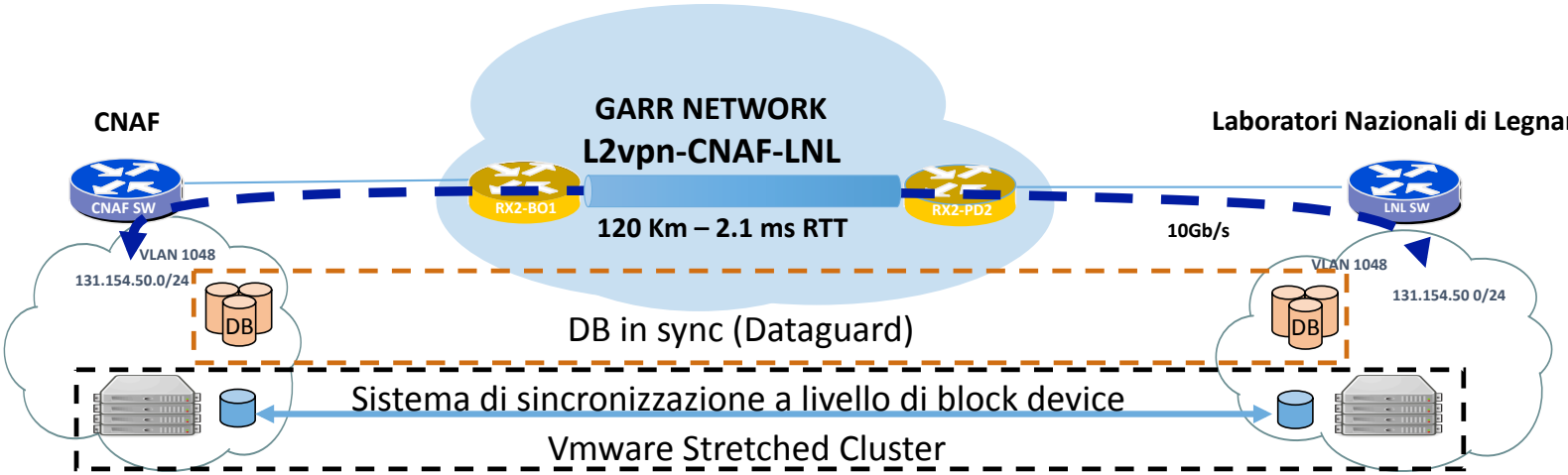
Efficenza legata al tipo di Job (CPU intensive o I/O Intensive).
Job CPU intensive girano con efficienza uguale a quelli interni, **I job più I/O Intensive, come prevedibile hanno una efficienza minore.**

20 Gb/s disponibili su 100 Gb/s teorici →cache (300 TB AFM)

- Squid per CVMFS (*FS per sw di esperimento*)
- Frontier Squid per Atlas e CMS *(Data transfer cache)*
- DNS per mappare squid del T1 sugli squid locali

DCI fra CNAF ed LNL
(L2 VPN per Business Continuity)

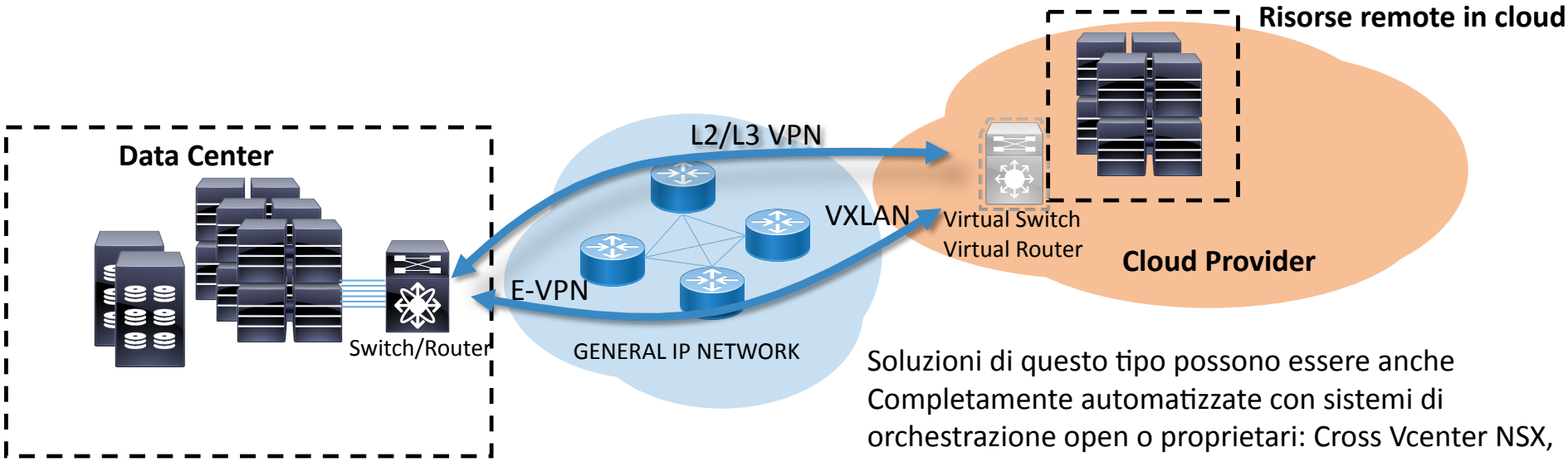
Caso d'uso: Infrastruttura di business-continuity per le applicazioni IT e gestionali dell'INFN.
Requisiti di rete: Link livello 2 fra CNAF ed i Laboratori Nazionali di Legnaro con RTT<5ms .



DCI in pre produzione realizzato da **GARR** utilizzando un LSP (Label Switched Path) MPLS ed una L2 VPN fra i due router GARR presso le sedi utente (in modo da ridurre al minimo la latenza).
Su questo link è possibile propagare le VLAN necessarie estendendo i domini di broadcast fra i due centri.
Anche in questo caso oltre alla configurazione della VPN in Overlay c'è anche l'utilizzo do un path a banda dedicata.
GARR sta anche valutando l'ipotesi di realizzare la connessione punto a punto direttamente a livello trasmissivo.

DCI o DCE con NFV (Overlay)

Nel caso di utilizzo di ingenti quantità di risorse di calcolo su Cloud, *Analogamente a come "Noleggio" VM per fare elaborazione dati , posso realizzare DCI utilizzando NFV istanziando vRouter, vSwitch, VPN Concentrator o vFirewall* tramite i quali realizzare i "Tunnel" necessari.

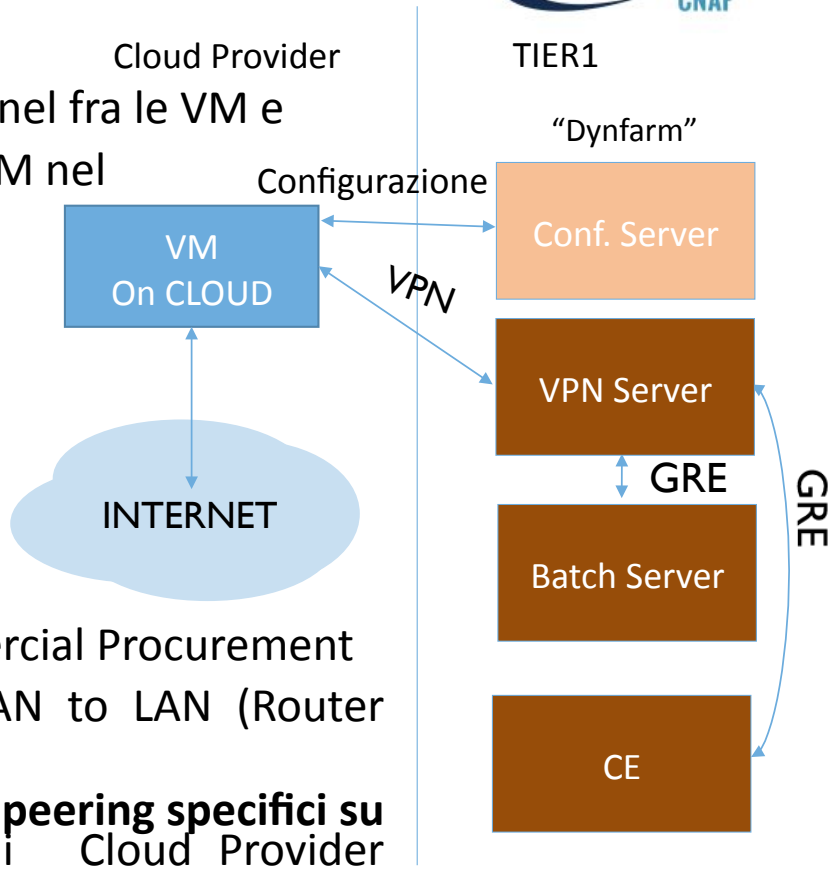


Soluzioni di questo tipo possono essere anche Completamente automatizzate con sistemi di orchestrazione open o proprietari: Cross Vcenter NSX, Contrail, ACI, Big Cloud Fabric, IP Infusion ,ecc.

Test di Cloud Bursting su cloud pubbliche
(Overlay su General IP)

Sperimentazioni su cloud pubbliche →Utilizzo di Tunnel fra le VM e gli elementi strettamente necessari ad integrare le VM nel Batch system.

- **Aruba**
- **Unicredit**
- **Clouditalia**
- **Microsoft**



Partecipazione al progetto HNSciCloud: Pre Commercial Procurement

- Sperimentazione Cloud di IBM → VPN IPsec LAN to LAN (Router Vyatta (NFV)) – (Attività terminata).
- T-System e REA (1200 Core nel 2018)→ Dynfarm + **peering specifici su vrf dedicato a banda riservata** fra GEANT ed i Cloud Provider coinvolti.

Data Center Interconnection (Ottico) ad alte prestazioni

Caso d'uso: Accordo INFN - CINECA per l'utilizzo di un sottoinsieme della partizione (A1) del super calcolatore Marconi (basato su , Intel® Xeon® E5-2697 v4) in phase out come calcolatore HPC di CINECA ma molto efficace per I workflow tipicamente HTC del TIER1 del CNAF.

216 Server (36 physical cores, 256 GB RAM)

Throughput teorico verso lo storage del TIER1: >300 Gb/s

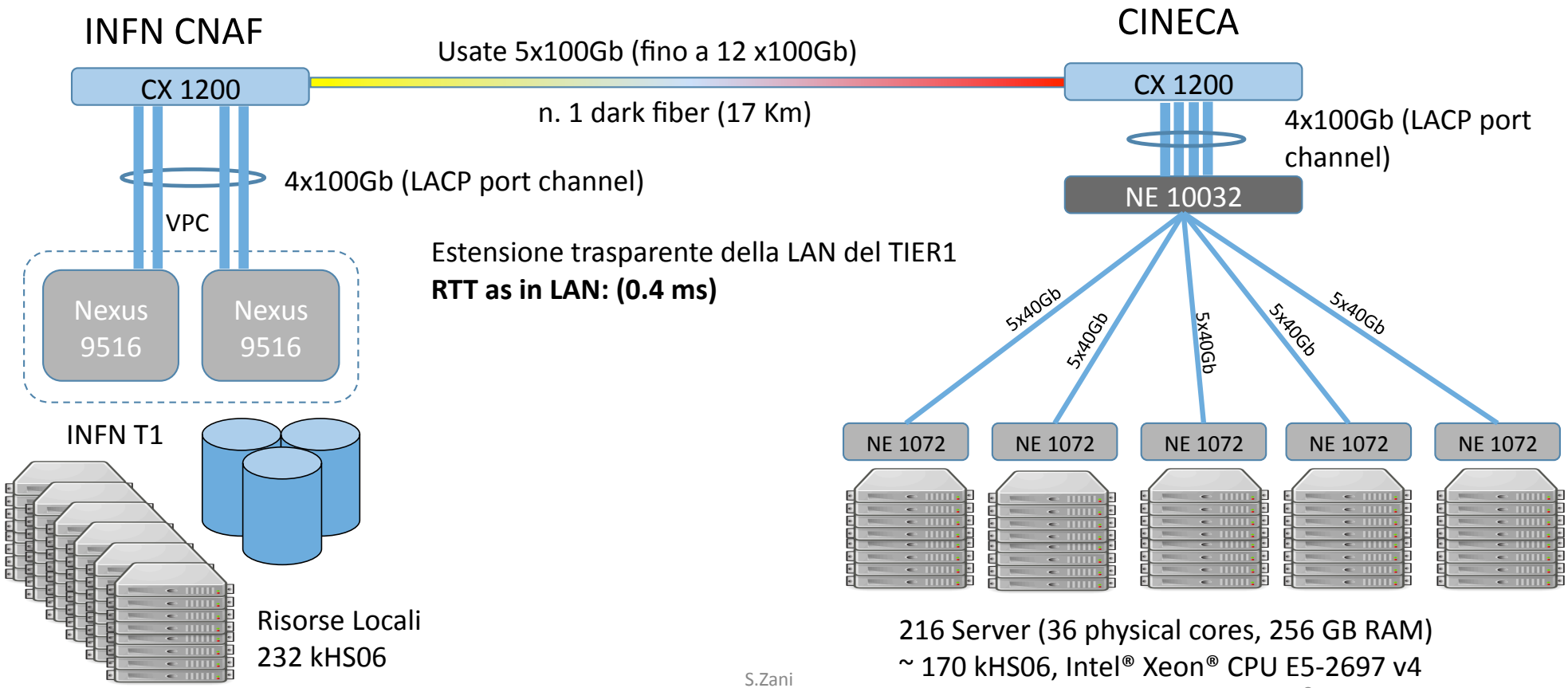
La distanza fra CINECA e CNAF è di circa 17 Km

Problema: **Garantire l' I/O di accesso ai dati che risiedono sullo storage del TIER1**

DCI con Infinera Cloud Express 2© (CX1200):

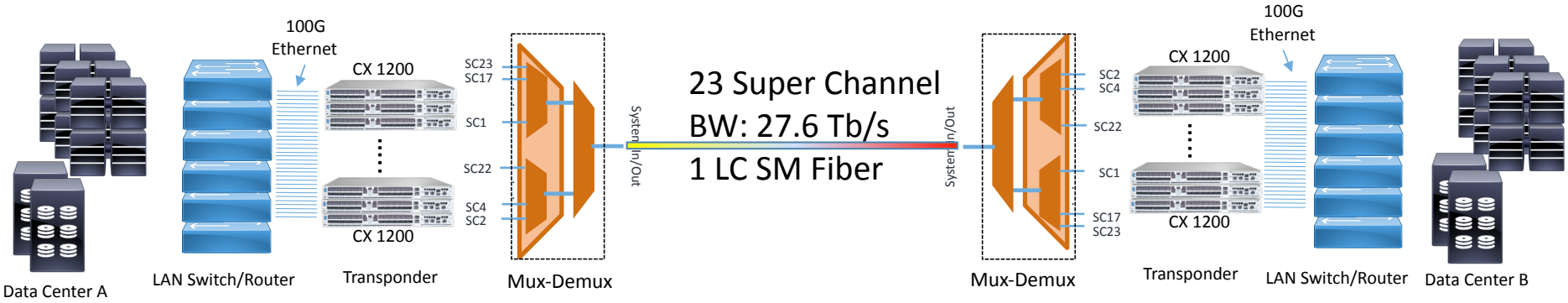
- 1.2 Tb/s in 1 U (12x100Gb interfacce Ethernet)
- Massima distanza 100-150Km
- 5,5 µs latenza
- Fino a 27 Tb/s su una fibra monomodale (stacking CX1200)
- Altamente configurabile: CLI, DNA (Sistema software di gestione di Infinera) o in configurazione SDN (usando le API) RESTCONF, gRPC

CNAF- CINECA Data Center Interconnection
In collaborazione con GARR



DCI – Super channel.

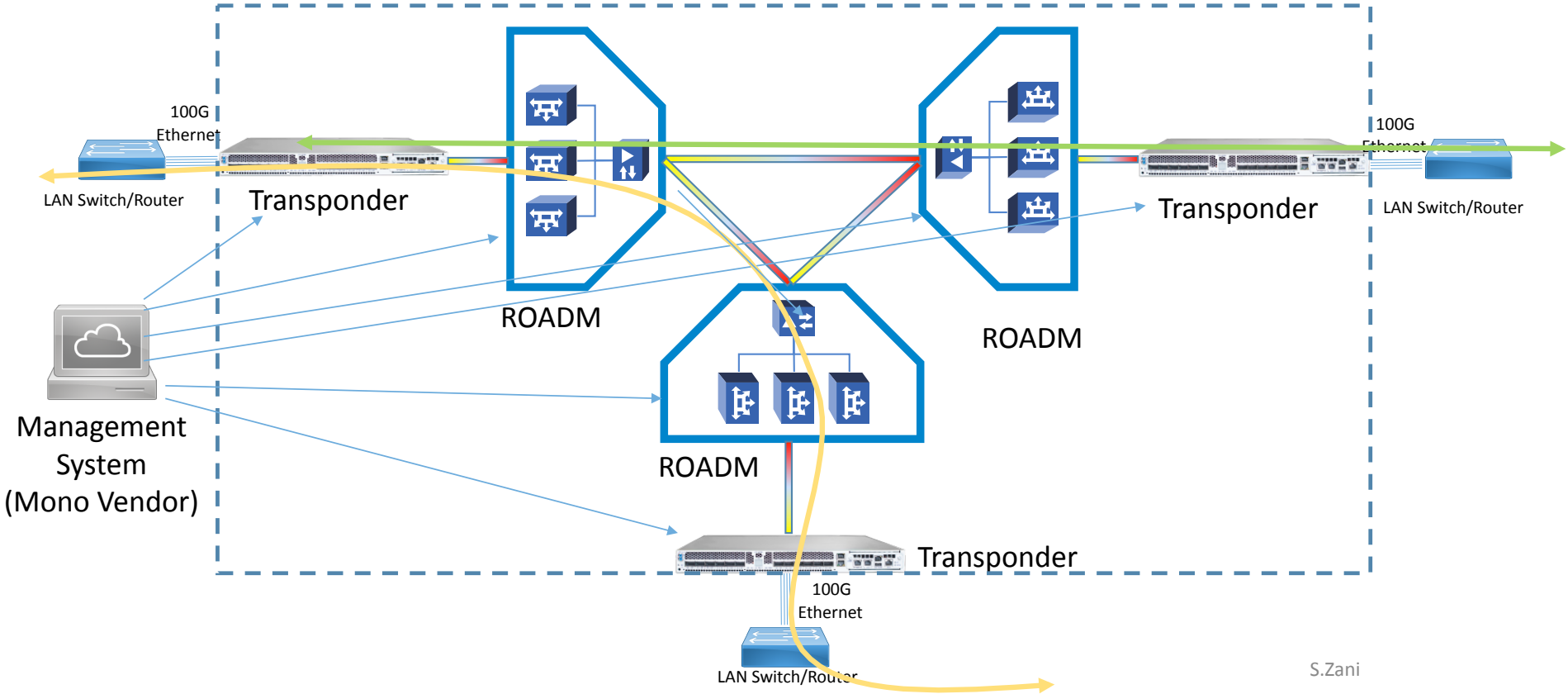
Il transponder utilizzato per la connessione CNAF – CINECA apre una nuova strada: si riescono a sfruttare le caratteristiche tipiche degli apparati trasmissivi direttamente con apparati da “Datacenter” eliminando i costi di complesse infrastrutture per la gestione di trasmissivi o router carrier class.



Esistono apparati “Transponder” in grado di gestire le stesse capacità integrandosi in un sistema trasmissivo che gli consenta di coprire distanze più ampie e gestire collegamenti verso destinazioni differenti.

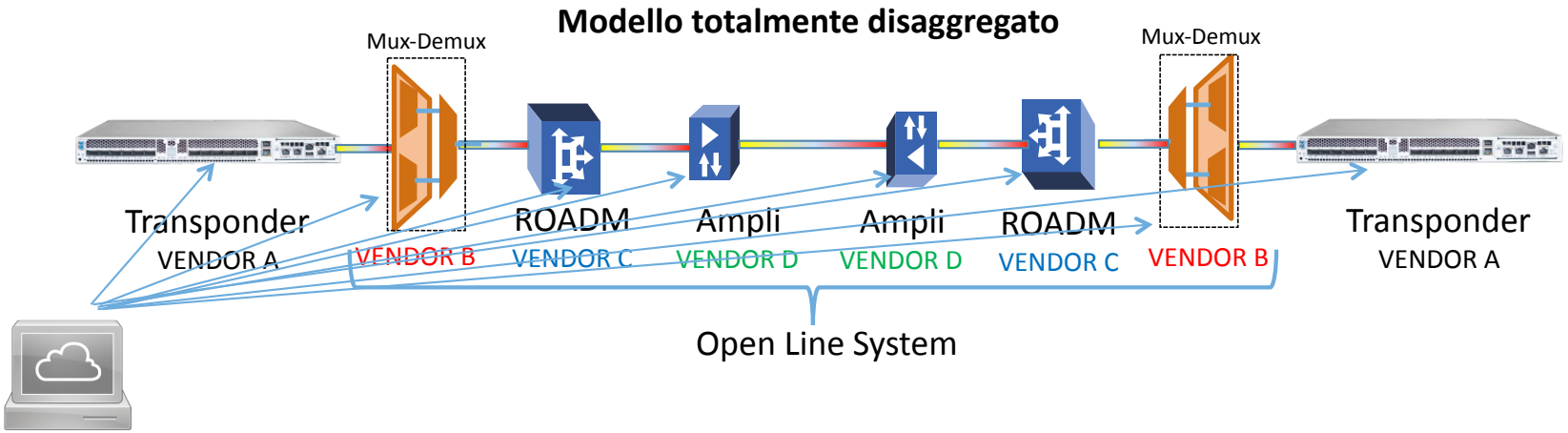
ROADM: *(Reconfigurable Optical Add/Drop Multiplexer)*

Questi OADM “Programmabili” sono composti da WSS (Wavelength Selective Switch) che consentono di gestire instradamento ottico in maniera flessibile e “ Programmabile ”. Moduli di amplificazione ottica consentono di coprire distanze più ampie.



Disaggregazione del sistema di trasporto ed Open Line System

I vendor che dispongono di soluzioni che vanno dall'IP all'ottico hanno da tempo sviluppato control plane multilivello ma i provider cercano di uscire dal "Vendor lock" e quindi si cerca di andare verso un modello disaggregato:



- Cicli di vita differenti dei vari elementi → possibile cambiare gli elementi in tempi diversi
- Soluzioni multi vendor → Maggiori performace e/o minori costi

Per riuscire a gestire un sistema disaggregato occorre un control plane che riesca ad interagire con tutti gli elementi.

Una soluzione SDN possibilmente basata su standard aperti costituisce la base per ottenere ***programmabilità ed automazione della rete geografica***.

Varie tecnologie per SD-WAN e DCI con la parte programmabile ROADM *(Reconfigurable Optical Add Drop Multiplexer)* **Open Line System**

Coriant Groove G30

3,2 Tbs/s in 1U

19,2 Tb/ on 1 Fiber

Adva FSP 3000 Cloud Connect

3,6Tb/s in 1U

34,4 Tbs on 1 fiber

Infinera XT 3300

“Meshponder” up to 6000Km

1,2 Tb/s in 1 U

27Tb/s on 1 fiber

Juniper TCX1000

Programmable ROADM

25.6 Tb/s

…….

NON SOLO ETHERNET

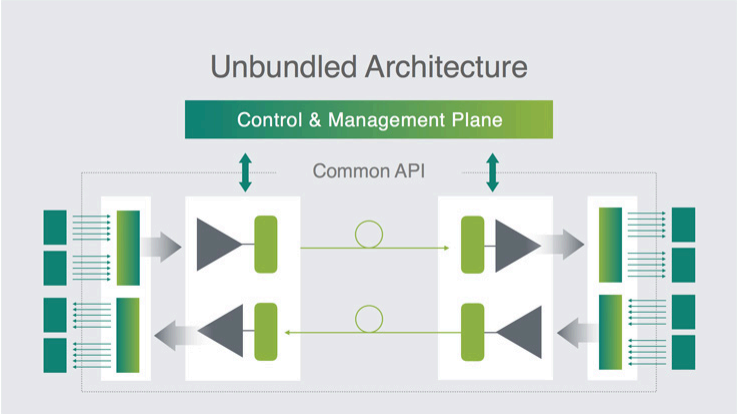
Alcuni transponder (Pacchetto/Ottici)
Consentono il trasporto di :
Ethernet, Fibre Channel ed Infiniband

Open Optical Packet Transport white box (Materia di sperimentazione con GARR)

- **Voyager**

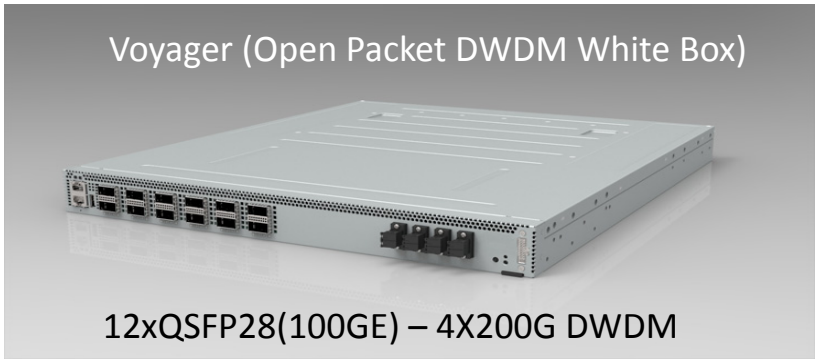
  In ambito TIP su iniziativa di Facebook è nato il progetto Voyager che rappresenta il primo white box che integra le fuzinalità di DWDM, “Transponder” e Packet Switch/Router (Open Packet DWDM).

  https://telecominfraproject.com/open-optical-packet-transport/



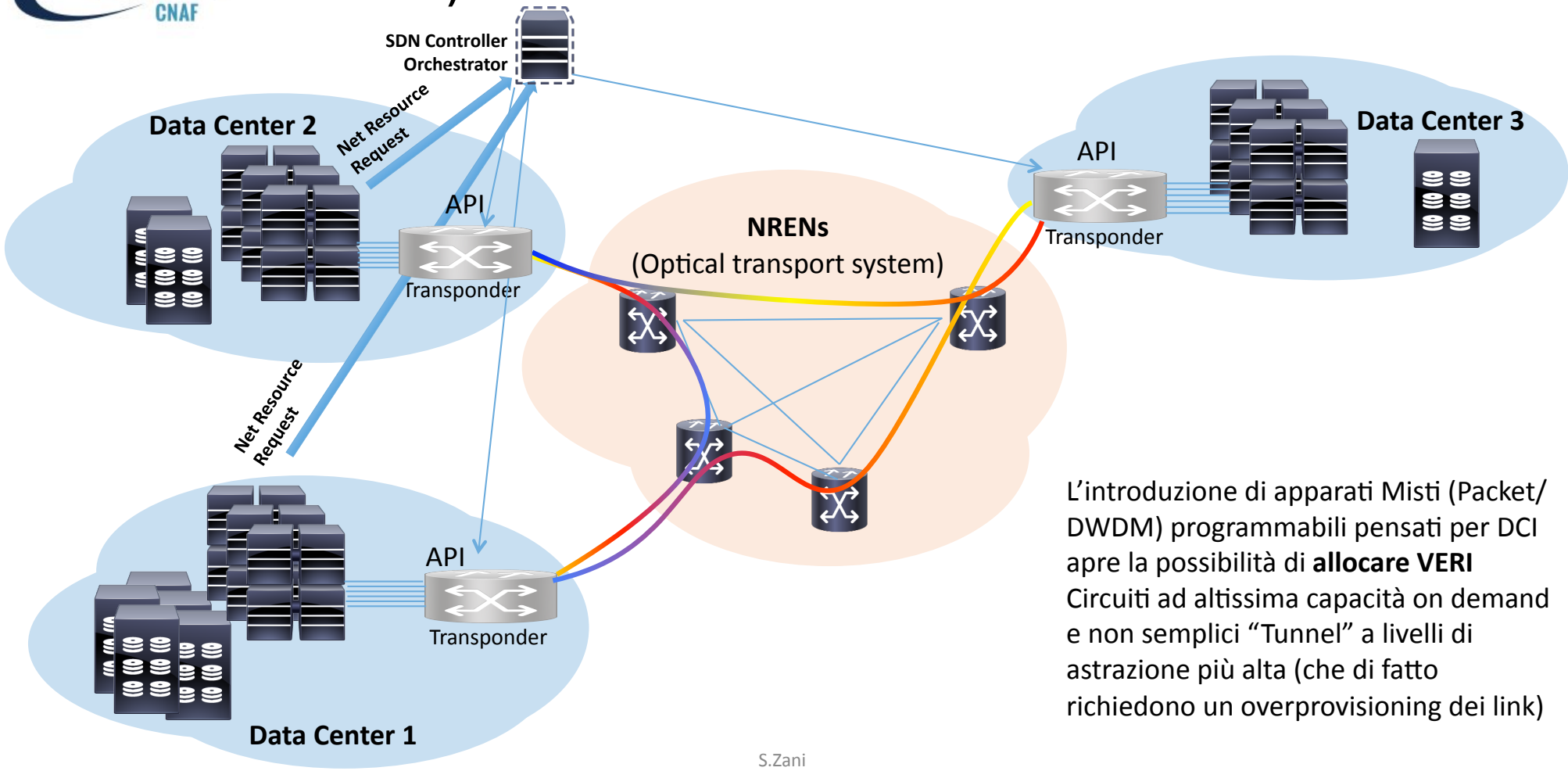
Partecipanti TIP

ACACIA
ADVA
Ciena
Cisco
Coriant
Facebook
Infinera
Juniper
Lumentum



Broadcom Tomahawk, AC 400 (DSP ASIC Optics)
Facebook FBOSS, Cumulus Linux …

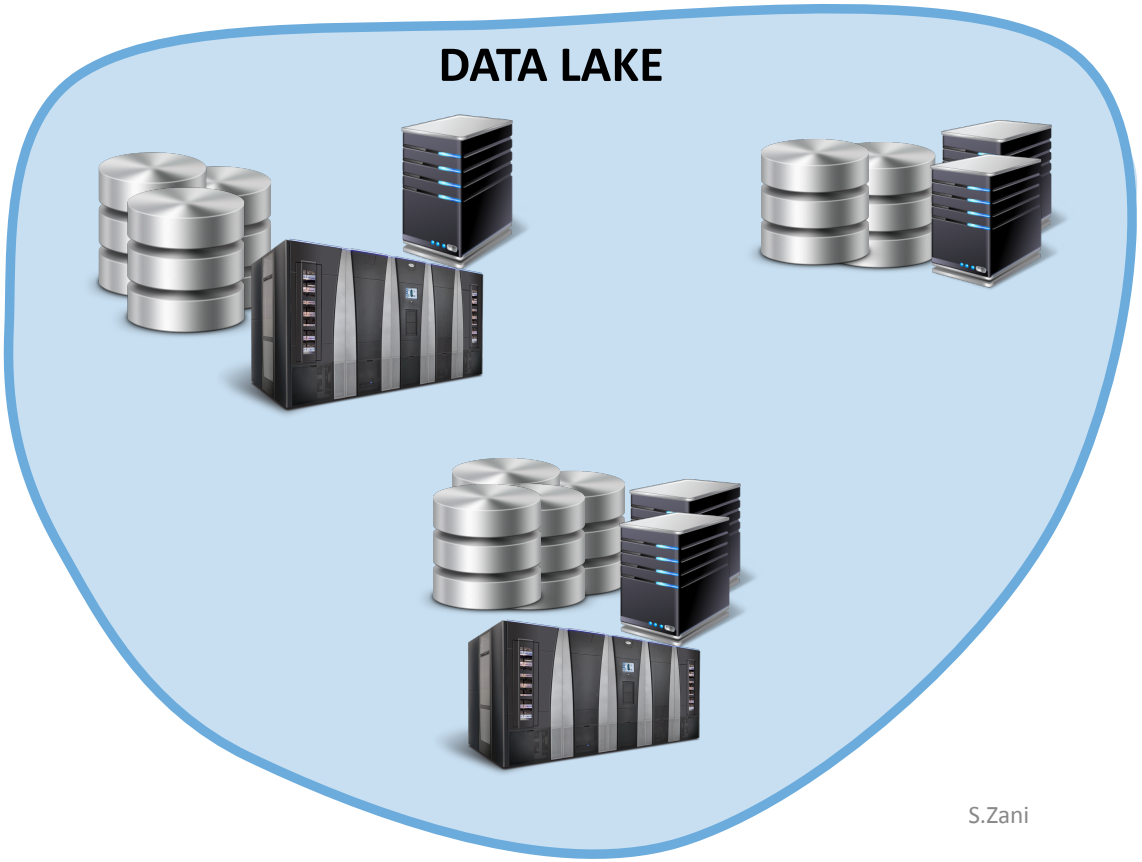
DCI (Software Defined – WAN) (Sperimentazine in collaborazione con GARR)



L'introduzione di apparati Misti (Packet/ DWDM) programmabili pensati per DCI apre la possibilità di **allocare VERI** Circuiti ad altissima capacità on demand e non semplici "Tunnel" a livelli di astrazione più alta (che di fatto richiedono un overprovisioning dei link)

Possibile usecase per DCI: il Data Lak

Gli esperimenti HEP stanno valutando un modello di data management che prevede l'accesso a pochi ma grandi "Data Lake".

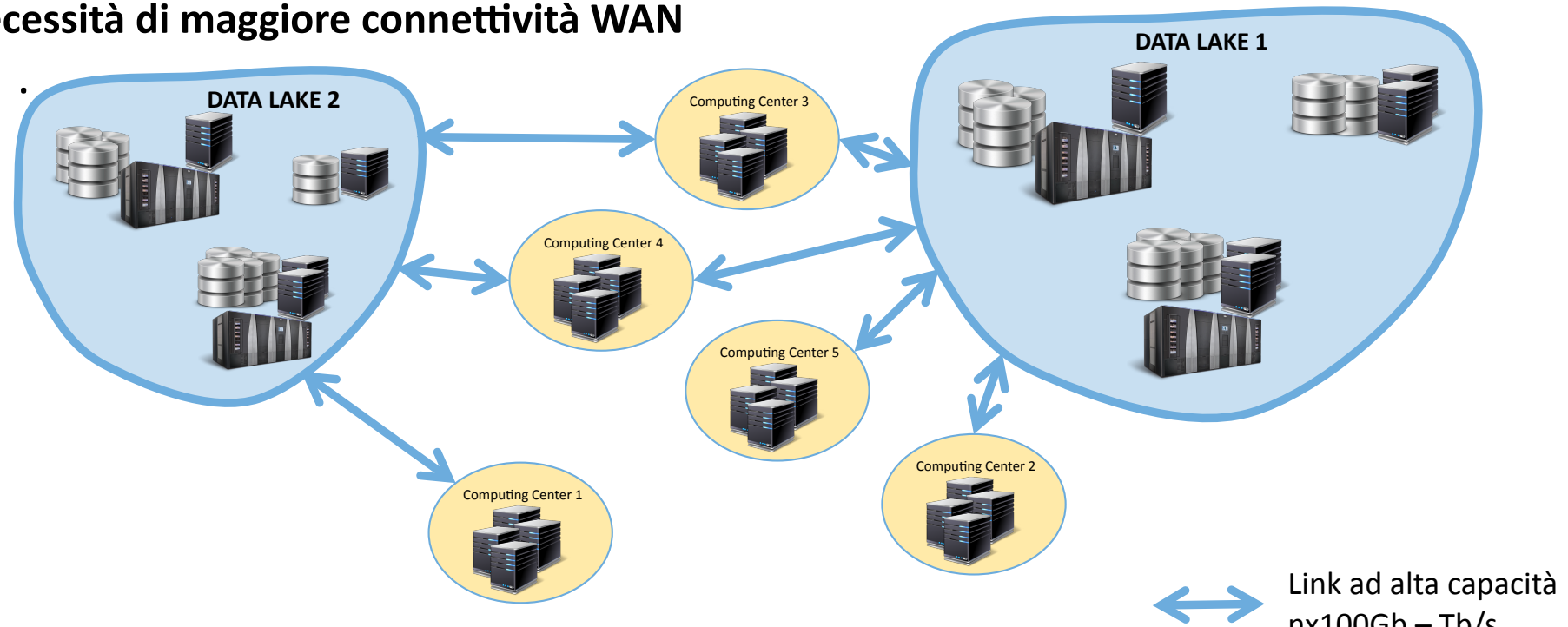


"I data lake sono gruppi di Storage Center geograficamente distribuiti che, potenzialmente utilizzano differenti tecnologie di storage ma sono gestiti ed acceduti come singole entità"

Possibile “Usecase” per SD-WAN : Accesso ai Data Lake

Il modello basato sui “Data Lake” dovrebbe:

- Ridurre il numero di copie dei dati rispetto ai modelli attuali
  - Minori costi per storage e minori costi per la gestione.
- **Necessità di maggiore connettività WAN**



DCI ed SD-WAN: Conclusioni

Le tecniche di DCI stanno diventando sempre più comuni e le reti della ricerca stanno evolvendo le proprie infrastrutture per fornire servizi on demand di connettività ad alta capacità .

Il concetto di rete geografica “Governata” via software non è nuovo agli operatori di rete, probabilmente oggi però è realmente possibile per l’utente, arrivare a “Programmare” lo strato di rete che costituisce il trasporto per le proprie applicazioni utilizzando interfacce e protocolli aperti .

C’è molto da imparare, molto da sperimentare e molto da fare assieme ai colleghi del GARR e delle altre reti della ricerca .

…Credo sarà molto divertente.

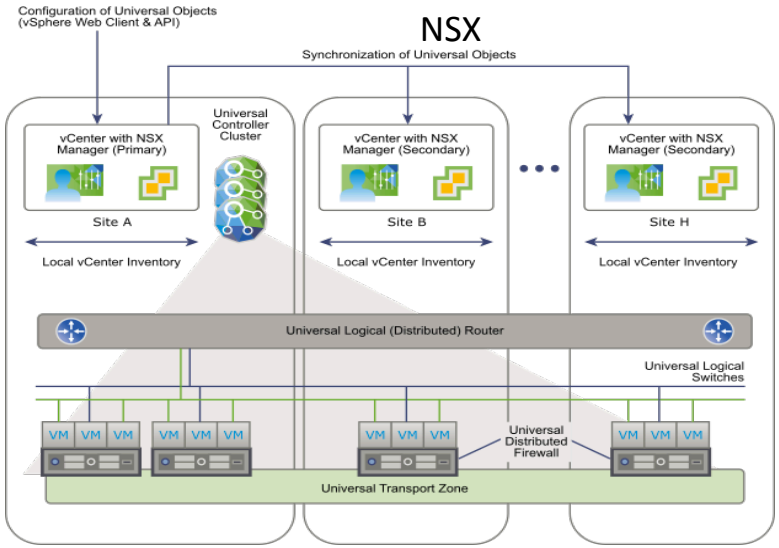
FIN
E

Backup slides

Alcune soluzioni overlay di DCI basate su SDN ed NFV

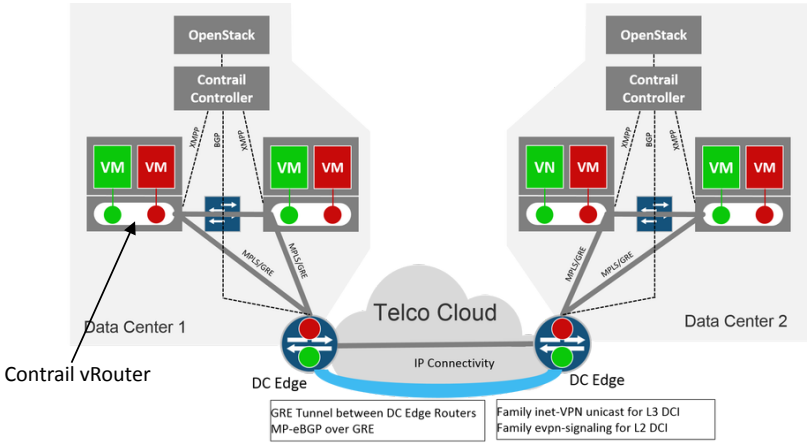
Soluzioni di estensione geografica dei datacenter, sono realizzabili anche a livello di ambienti di virtualizzazione ed orchestrazione: Alcune esempi di soluzioni: Cross Vcenter NSX, Contrail, ACI, Big Cloud Fabric, IP Infusion ,ecc.

NSX



NSX: La tecnologia di overlay sottostante sono le VXLAN.

Open Stack + Contrail



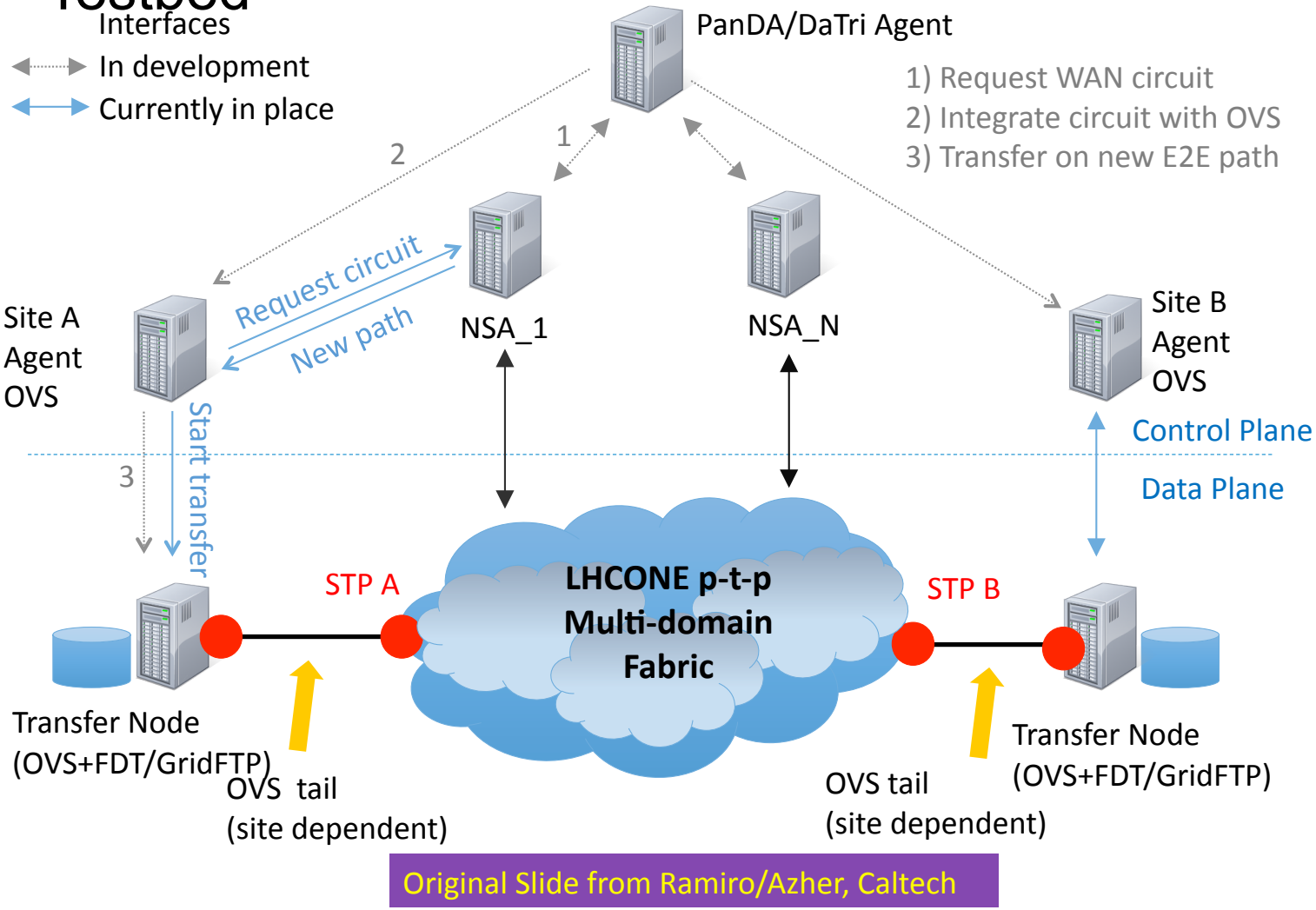
Con Contrail si utilizzano E-VPN per l'estensione L2

R&D su SDN e NFV in ambito HEP

E’ stato creato un gruppo di lavoro sulla virtualizzazione della rete in seno ad HEPiX (https://www.hepix.org)
HEPIX NFV WG (Mailing list: https://listserv.in2p3.fr/cgi-bin/wa?SUBED1=hepix-nfv-wg)
Il gruppo si propone di esplorare tecnologie di NFV e SDN che possano aiutare a gestire gli enormi flussi di dati previsti dai prossimi RUN di LHC tenendo presente che esperiment di altre scienze competeranno pesantemente sulle stesse infrastrutture di rete.
Raccoglie le varie esperienze fatte in ambito WLGC **mettendo assieme Esperimenti, Siti, NREN**

Michigan University: Progetto ATLAS (AGLT2/Michigan, MWT2/Chicago e KIT) consiste nel deployment di OVS all’interno di istanze di produzione di Storage (dCache in questo caso) e prevede di collegare tutti gli OVS ad un unico controller per gestire una sorta di traffic shaping fra i diversi storage element dei siti.
https://www.aglt2.org/wiki/bin/view/Main/Open_vSwitch/WebHome

Diagram of Possible Future SDN Dev-Ops Testbed



Original Slide from Ramiro/Azher, Caltech

The suggested LHC computing model