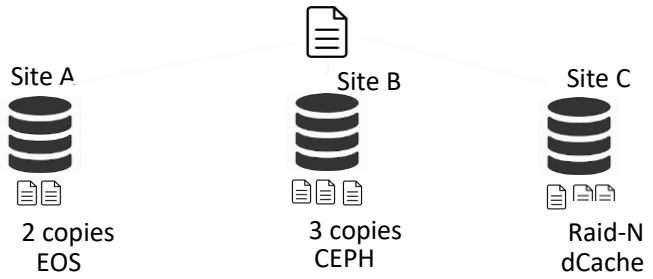
IDDLS PROJECT STATUS UPDATE

DANIELE CESINI – INFN-CNAF

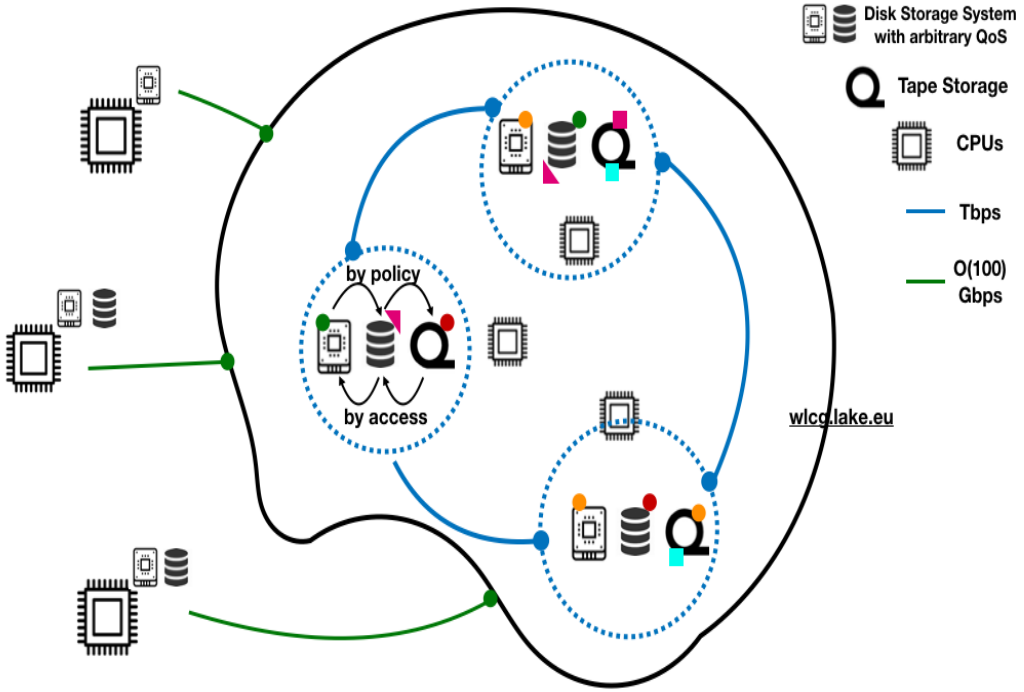
INFN-CCR WS 26TH MAY 2021

SOME IDEAS ON REDUCING STORAGE COSTS WITHIN WLCG

- Reduce hardware cost: better exploiting the concept of QoS(Quality of Service)
  - Probably today we replicate more than we need
  - Reducing the number of copies
- Identify the impact on costs of deploy fewer (larger) storage sites maintaining high standards in availability and reliability
  - Create large storage repositories that "look like one, but it is composed of many" ➔ the DataLake
- Co-location of Storage and CPU will not be guaranteed anymore
  - Need technologies for quasi-transparent data access from remote locations
    - Caching systems needed

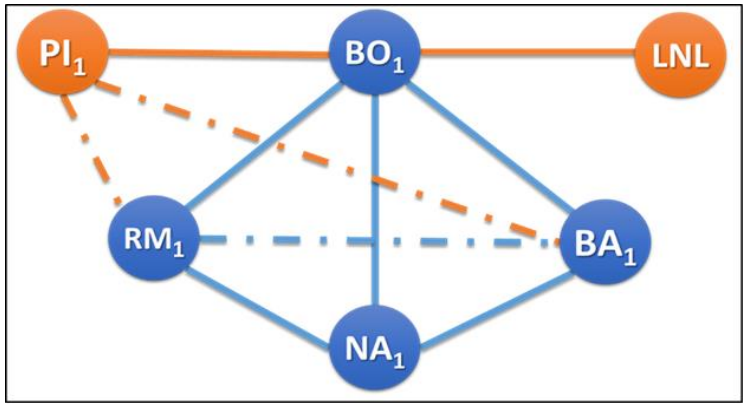


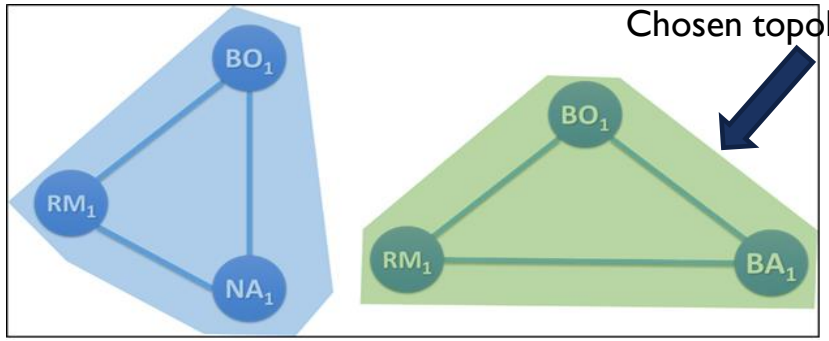
A stronger integration of sites could lead to a reduction of the number of copies



THE IDDLS PROJECT

- INFN-GARR collaboration to realize a prototype of an Italian Data Lake exploiting:
  - Last generation networking technologies provided by GARR
    - DCI (Data Center Interconnection) equipment
    - SDN (Software Defined Network) deployment
  - Software for creating **scalable storage federations** provided by INFN
    - eXtreme-DataCloud project
    - DOMA in WLCG
    - SCoRES project (INFN-NA)
  - Real life use cases for testing
    - CMS
    - ATLAS
    - BELLE-II
    - Possibly involving LNGS experiments (XENON) and VIRGO
  - Funded by CSN5 and GARR
  - Participating sites: BA, CNAF, LNF, LNL, NA, PI, PG, RM1 + GARR
  - Three years project – started in 2019
    - Extension under discussion





Possible IDDLS topologies

PROJECT TIMELINE

- First year
  - identify the networking technology – Done
  - complete tenders – Done
- Second year
  - test the networking equipment in the lab – Ongoing
  - deploy at the sites – foreseen in September 2021
  - definition of the Data Lake architecture and middleware - Ongoing
- Third year
  - Performance tests using VO use cases

→ This was supposed to be 2020:
1 year COVID delay

TEST OF TRANSPONDERS AT NX100G

1. **Transponder TP100GOTN XTM Infinera**

   Client: 100G Ethernet

   Line: CFP1 100G ACO - QPSK coherent

2. **Juniper ACX6360 (and MX240)**

   Client: 100-Gbps (QSFP28) pluggable interfaces

   Line: 200-Gbps CFP2-DCO coherent DWDM pluggable interfaces, which support 100-Gbps QPSK and 200-Gbps 8QAM, and 16QAM modulation options

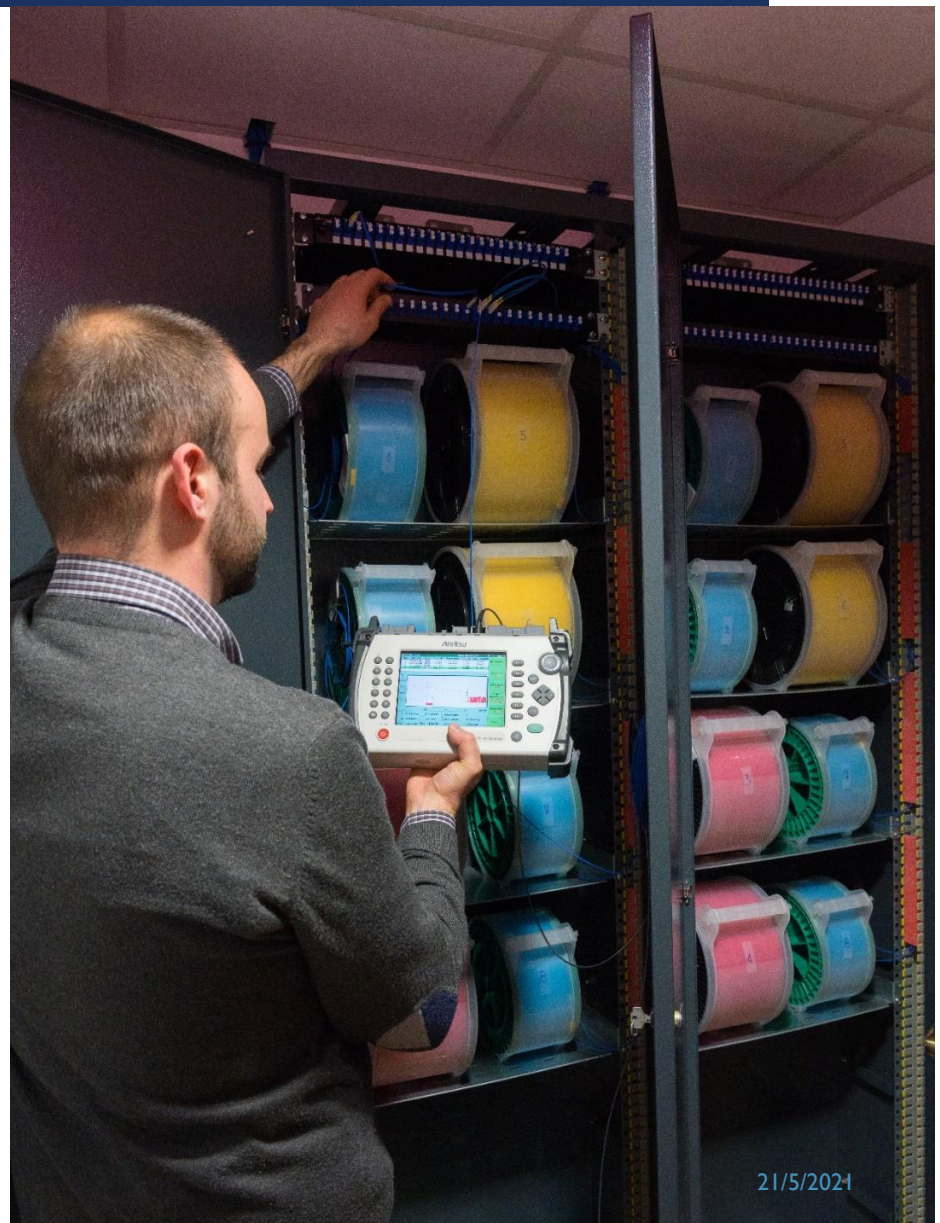
3. **Infinera Groove G30**

   Client: up to 4x100GbE QSFP28

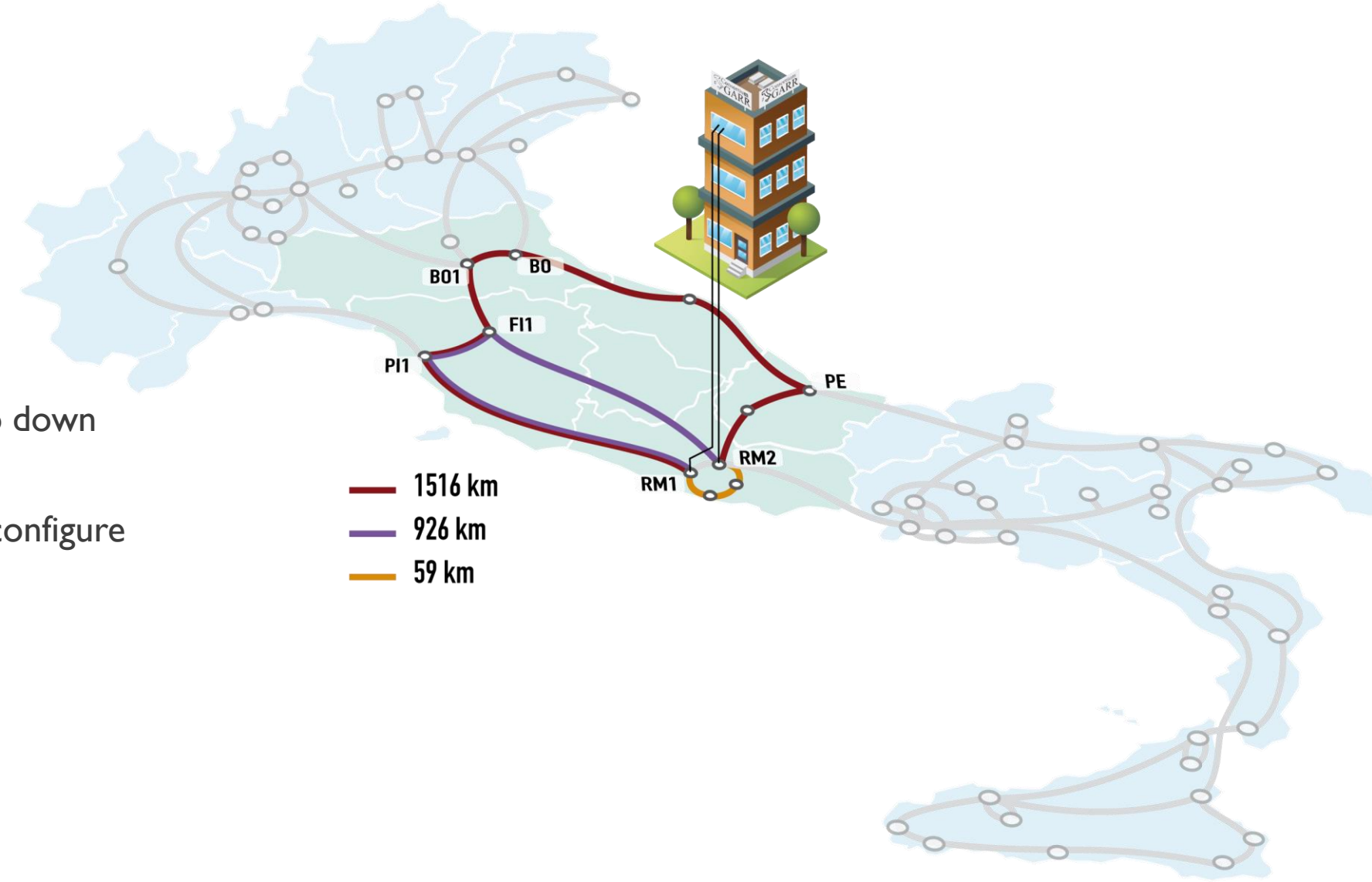
   Line: CHM1 sled – 400G 2xCFP2-ACO (100G QPSK, 150G 8QAM, 200G 16QAM)

GARR LAB: FIBER SPOOLS

- 16 spools of single G.652d fiber
  - vendor: Fujikura
  - 8 spools 50km -> 4 pairs
  - 8 spools 25 km -> 4 pairs
  - In total 600km single fiber
  - The **flexible topology** changes through a patch panel on the rack top



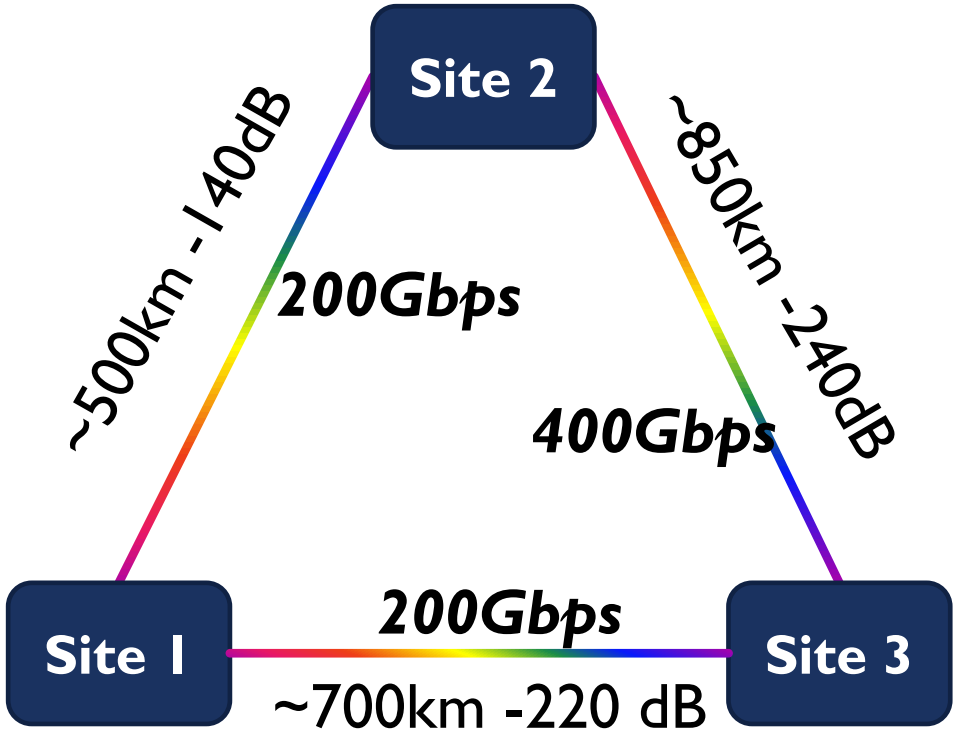
LINKS TOWARD THE PRODUCTION ENVIRONMENT



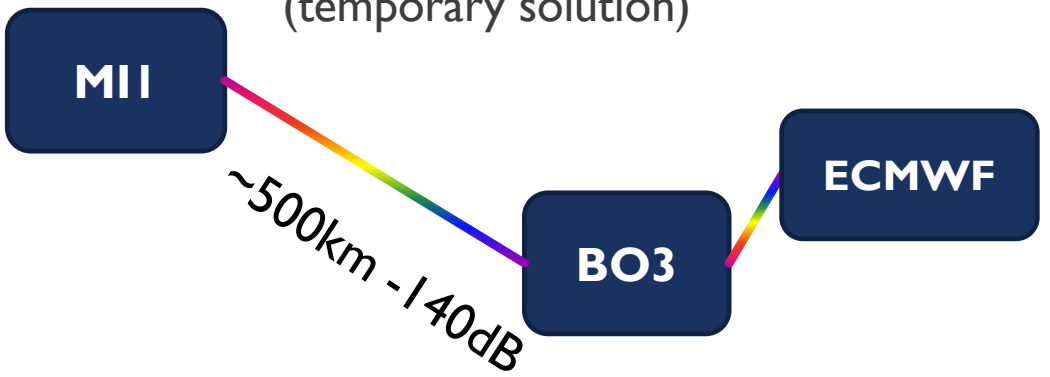
- From the GARR lab two cables go down and reach the RM2 and RM1 PoPs
- From these PoPs it is possible to configure links that are as long as:

- ~60 km
- ~1000 km
- ~1500 km

- **Italian Distributed Data Lake for Science**
  - Large bandwidth over hundreds of kms



- ECMWF access to GARR (temporary solution)



	ECMWF	IDDLS
INFINERA XTM	OK	NOK
Juniper ACX	OK	NOK
INIFNERA GROOVE	OK	OK

See full test details
https://agenda.infn.it/event/19934/
© Gloria Vuagnin @CCR

INFINERA GROOVE G30

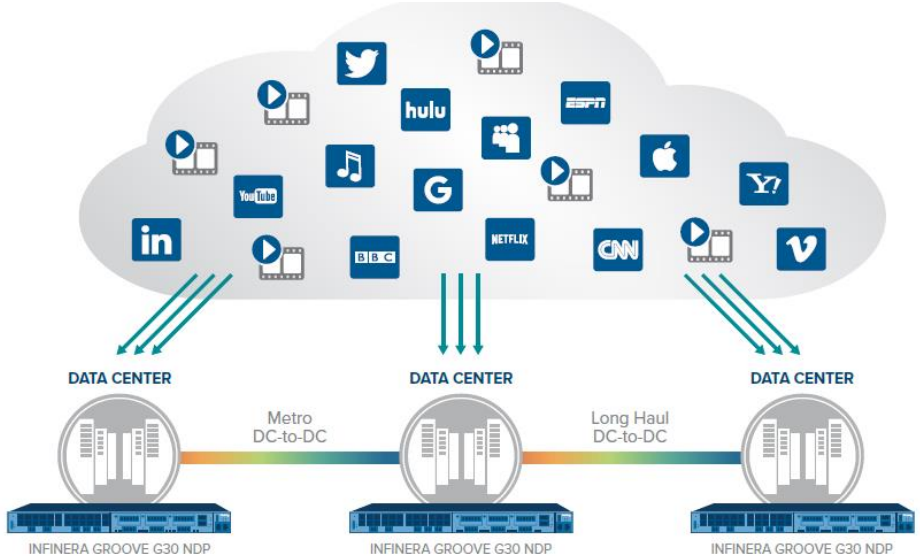


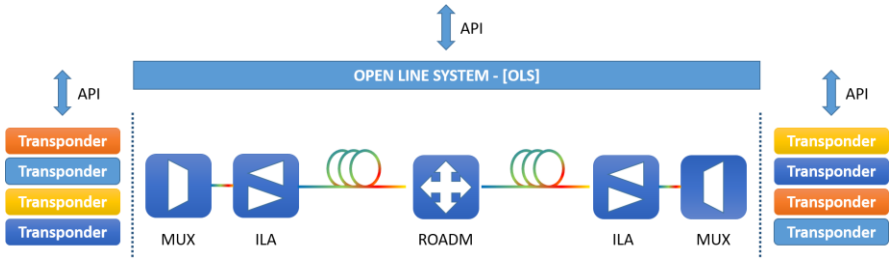
Figure 1: Powering High Performance, Cost-efficient Data Center Connectivity

THE PURPOSE-BUILT INFINERA GROOVE G30 NETWORK DISAGGREGATION PLATFORM

The Infinera Groove G30 Network Disaggregation Platform (NDP) is an innovative 1RU modular open transport solution for cloud and data center networks that can be equipped as a muxponder terminal solution and as an Open Line System (OLS) optical layer solution. Purpose-built for interconnectivity applications, the disaggregated Groove G30 delivers industry-leading density, flexibility, and low power consumption.

Infinera Groove G30 Open Line System (OLS)

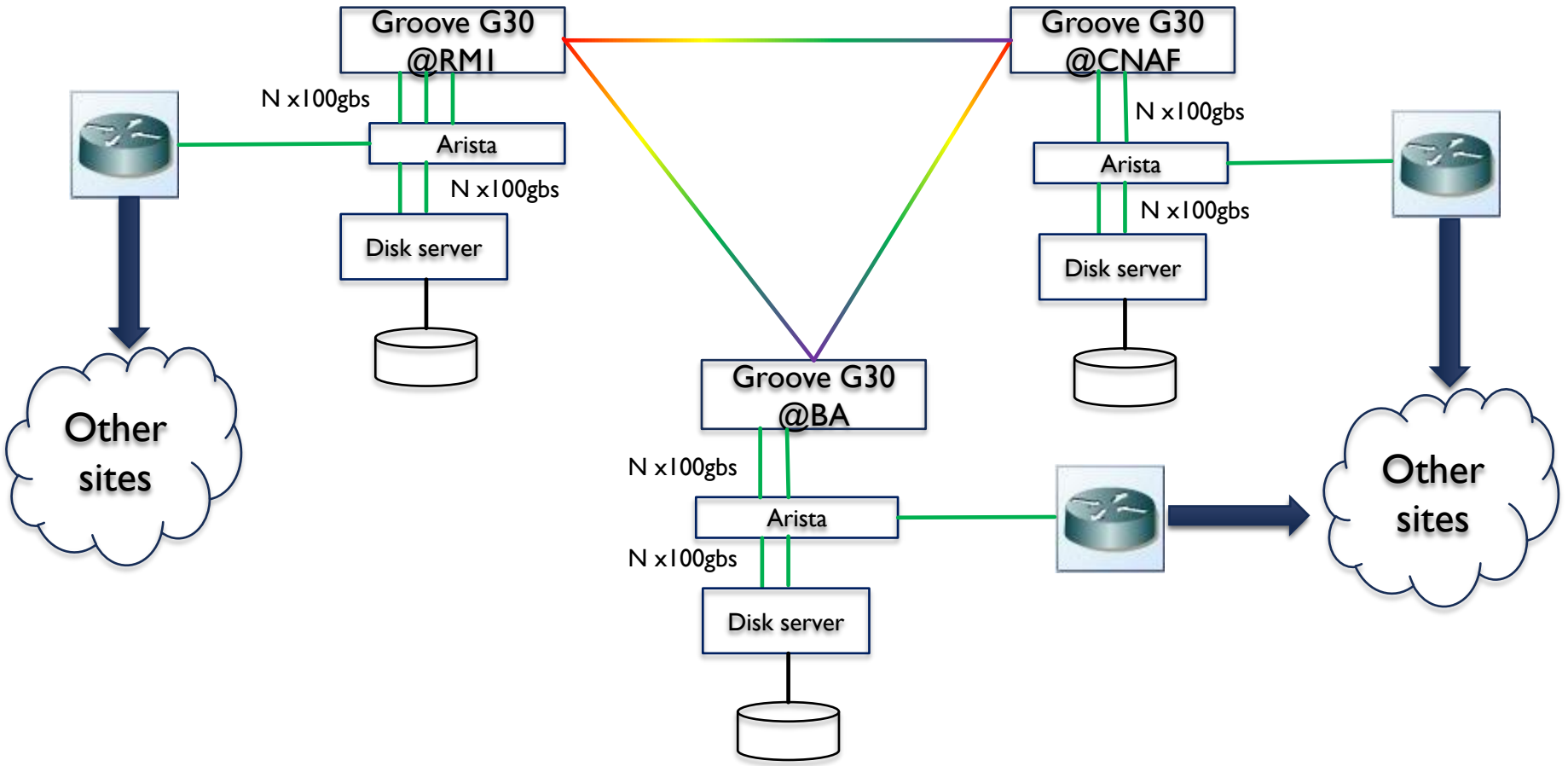
Infinera Groove G30 Muxponder (MUX)



partially disaggregated ➔ decoupling of photonic elements and transponders

The GX G30 supports management, automation, and streaming telemetry via open interfaces

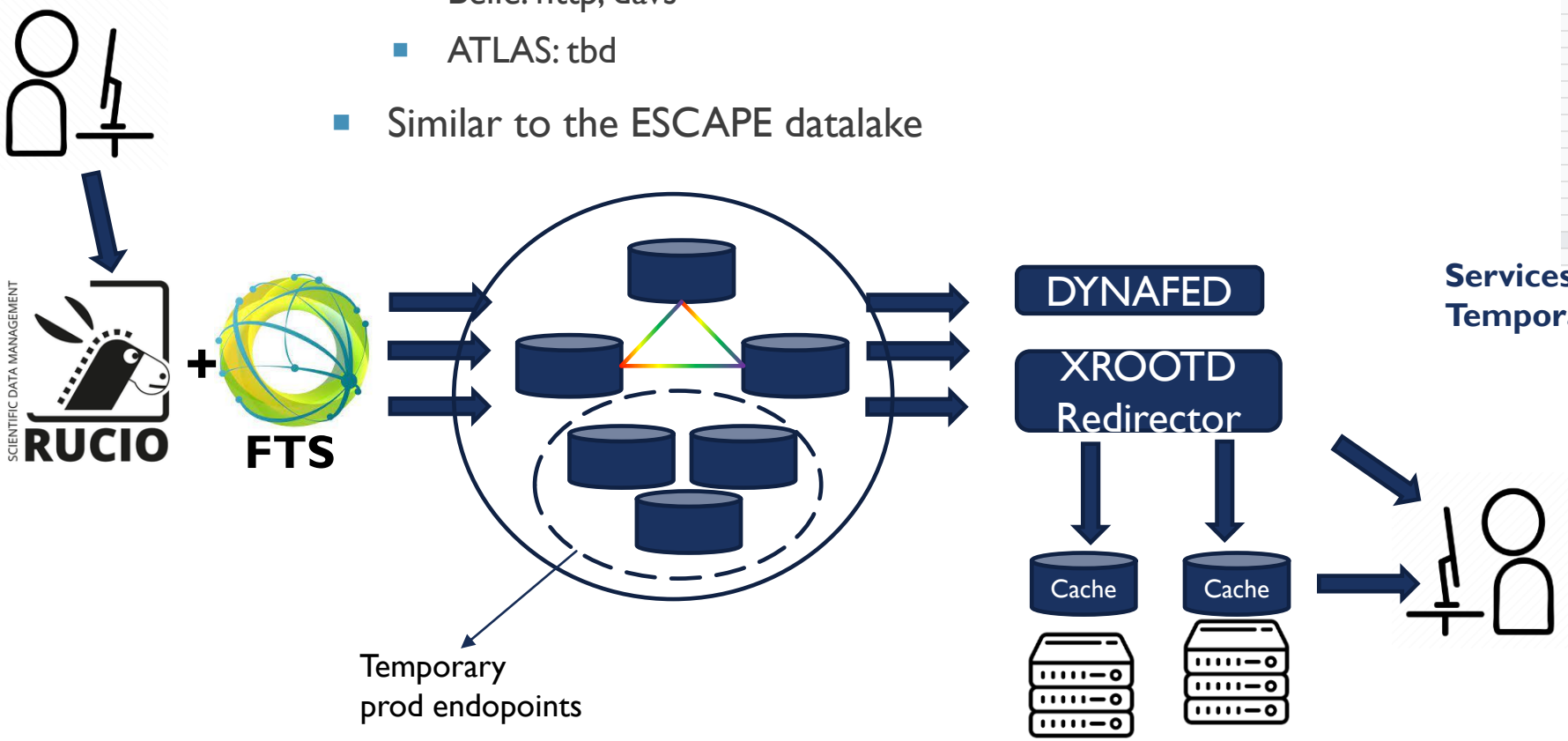
IDDLS ARCHITECTURE



- Link at L3 as base configuration
- Possible L2 through overlay technologies to realize data center extension
  - Under testing
- SDN capabilities under testing
  - Bandwidth modulation between the sites (multiple of 100G
  - Zero Touch Provisioning on the ARISTA switches
    - Remote interaction via API+REST already setup

DEPLOYMENT SCENARIO: QOS BASED DATALAKE WLCG+ BELLE

- VO writes using RUCIO
- Users read via federators
  - CMS: xrootd
  - Belle: http, davs
  - ATLAS: tbd
- Similar to the ESCAPE datalake

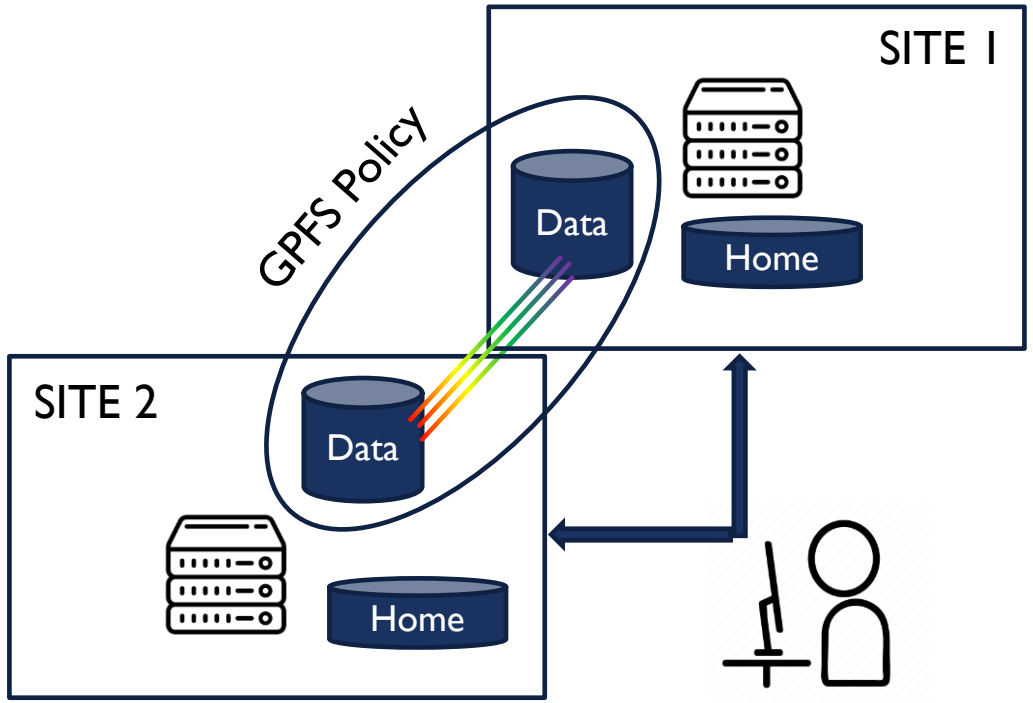


	A	B	C	D	E
1	Site	Storage Technology	VO	Auth Type	Endpoint
2			Endpoints		
3	CNAF	xrootd	CMS		xrootd-cms
4	CNAF	xrootd	ATLAS		xrootd-atlas
5	CNAF	gftp/davs	BELLE		gridftp-storm-archive
6	NA	gftp/davs	BELLE		belle-dpm-01.na.infn.it
7	NA	davs	BELLE		recas-dpm-01.na.infn.it
8	NA		ATLAS		
9	NA		ATLAS		
10	BA	xrootd	CMS		ss-01.recas.ba.infn.it
11	BA	xrootd	CMS		ss-02.recas.ba.infn.it
12	BA	xrootd	CMS		ss-02.recas.ba.infn.it
13	PI		CMS		
14					
15					
16					
17					
18			Federators		
19	CNAF	xrootd redir	CMS		??
20	CNAF	RUCIO	CMS	x509	rucio1.cloud
21	CNAF	RUCIO	ATLAS	x509	rucio2.cloud
22	CNAF	RUCIO	BELLE	x509	rucio3.cloud
23	CNAF	FTS	??	x509	https://fts3-cnaf.cloud.cnaf.infn.it:8446
24	NA	DYNAFED	BELLE	davs	dynafed-prod01.na.infn.it
25					
26					
27					
28			Caches		
29					

Services almost complete
Temporary endpoints identified for BELLE, CMS

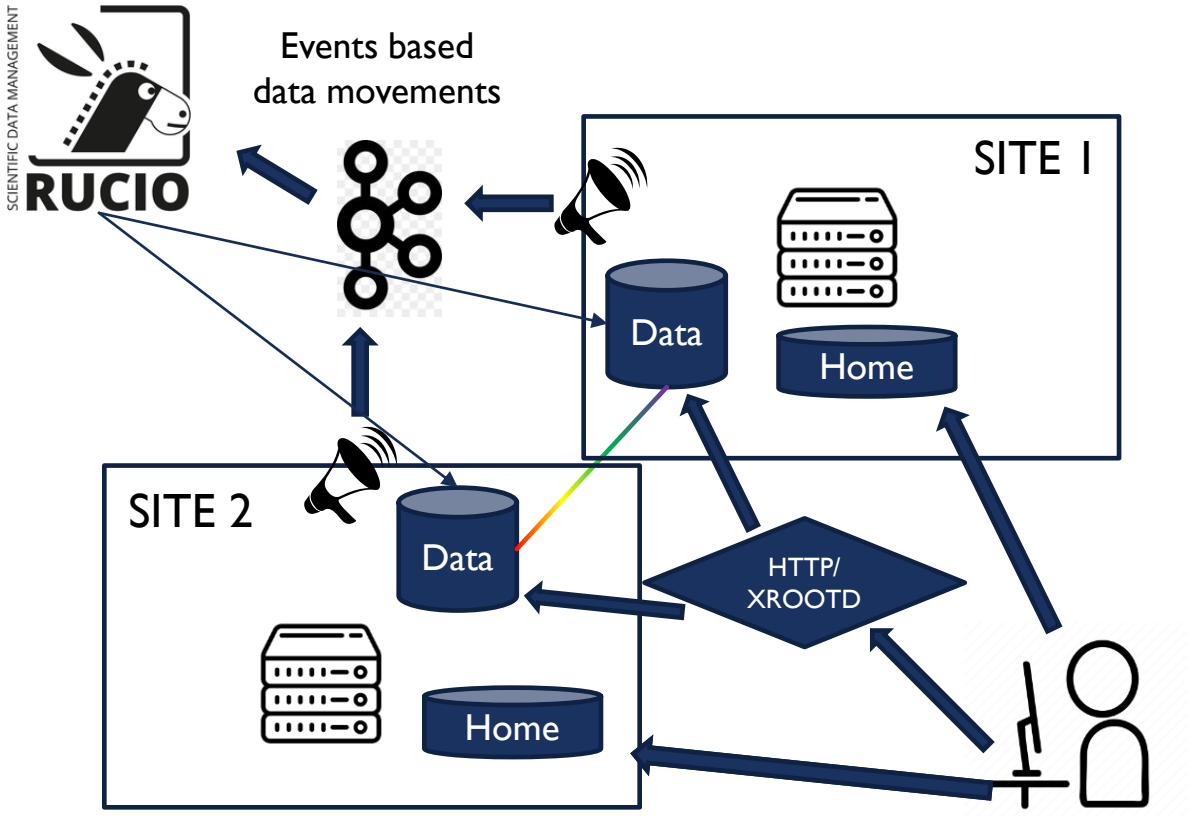
DEPLOYMENT SCENARIO: HA FOR VOS WITHOUT A DISTRIBUTED COMPUTING MODEL

- Allow users to work in HA using at least two geographically distributes sites
- Federation at the level of the filesystem
- Synchronous replicas performed at the infrastructure level by the filesystem i.e. GPFS Policies
- Transparent to the users by same FS solution needed at the sites



- Tests already performed by CNAF, PI, GE outside IDDLS (©Vladimir):
- Distrubuted GPFS cluster
  - CNAF➔10Gb Pisa and Genova
  - Genova: quorum node 1 GbE
  - CNAF and Pisa: two identical servers equipped with: 6 dischi SATA disks and 6 SSD disks
  - FS configured to host two storage pools - SATA and SSD with replica 2
  - net.ipv4.tcp_syncookies = 0 + jumbo frame double the performances
- Test results using Iozone: no differences between local and remote IO for SATA pool if blocksize > 1MB

DEPLOYMENT SCENARIO: HA FOR EXP WITHOUT A DISTRIBUTED COMPUTING MODEL

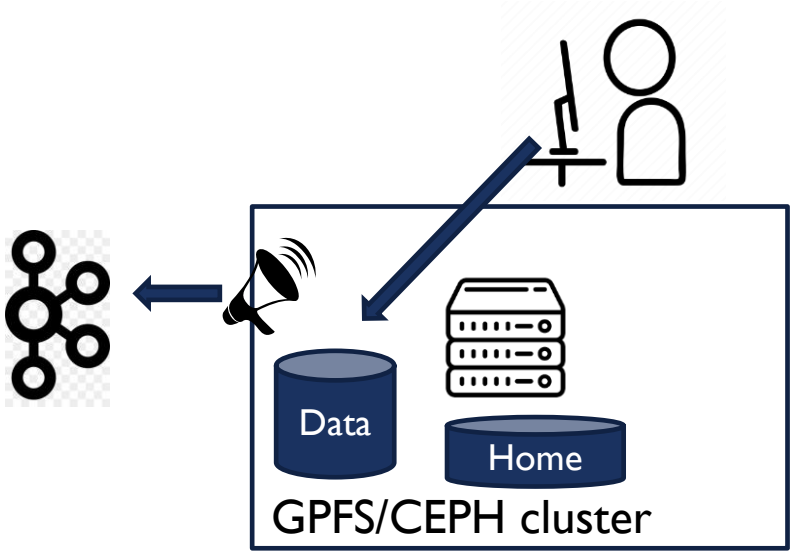


- Transparent creation of redundant federated storage resources – XDC approach
  - automatic replication in different, geographically distributed, availability zones
  - Replica performed at the infrastructure level, without the user intervention via RUCIO+FTS
  - Asynchronous copy but not bounded to a particular filesystem at the sites

EXPERIENCE WITH STORAGE EVENTS @ CNAF

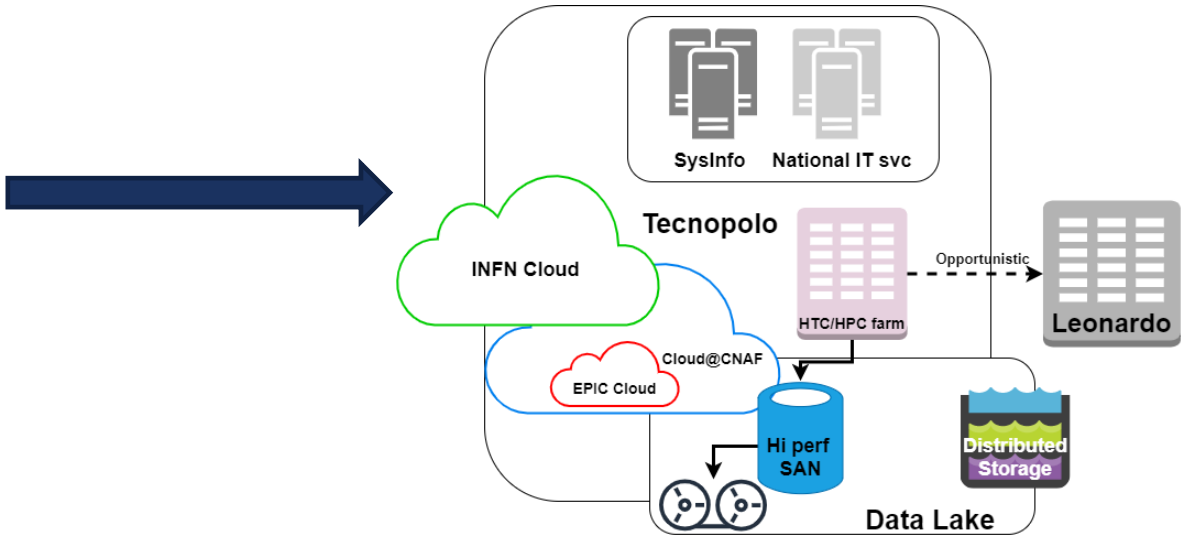
- **Intercept relevant events on the filesystem and receive corresponding notification**
  - File Creation
  - File Write
  - Directory access
- **Inject the notifications into the policy manager system** to trigger data movements
  - Within the site or the Lake, including QoS change
    - QoS change without scanning the buffer
- FS technologies tested
  - GPFS
    - an internal utility that needs further licenses
  - CEPH
    - script developed ad hoc to parse log files
- Preliminary test successfully completed
  - Mainly on write operations
    - **0.8 Hz (gpfs)**
    - **2.8 Hz (ceph)**
  - Need to test at higher rate

© E.Fattibene et al.



CONCLUSION

- IDDLS is proceeding even if highly impacted by the lockdowns
- Data Center Interconnect technologies for multi100G over 1000km identified after lab testing
- Tenders for networking equipment and servers have been completed and hw acquired
- Data Lake architecture and testing scenario identified
  - WLCG VOs
    - QoS based data management based on RUCIO
  - Experiments without a distributed computing model
    - Transparent replication done at the infrastructure level
- Configuration tests started in the GARR lab
- Deployment at sites foreseen in September 2021
- Project extension under discussion



CNAF Tecnopolo building blocks