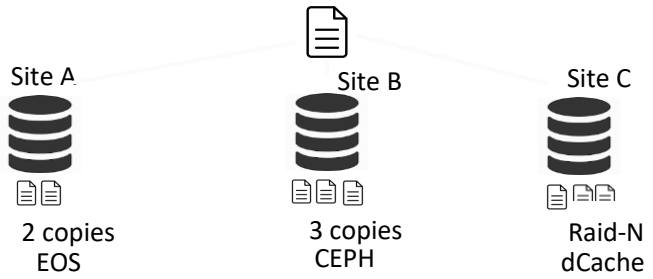
IDDLS PROJECT STATUS UPDATE

DANIELE CESINI – INFN-CNAF

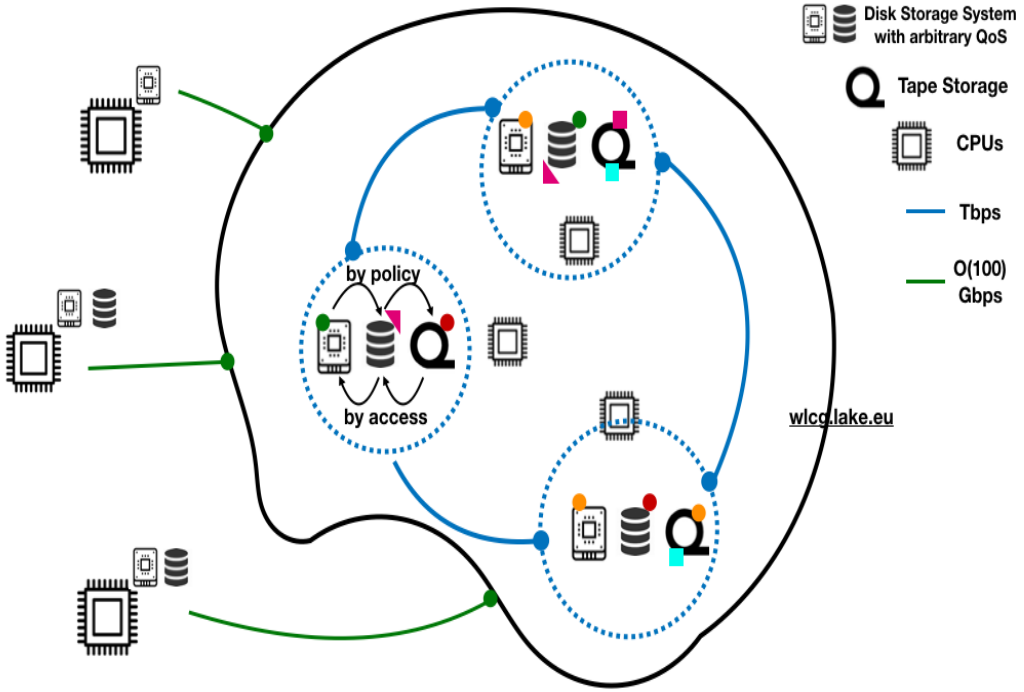
REFEREE 07TH JULY 2021

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

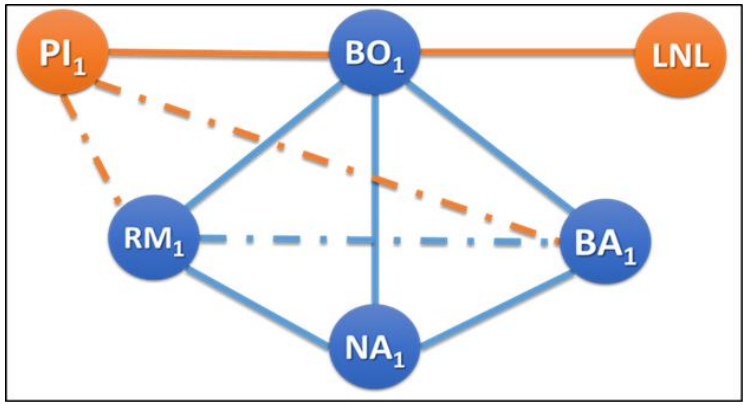


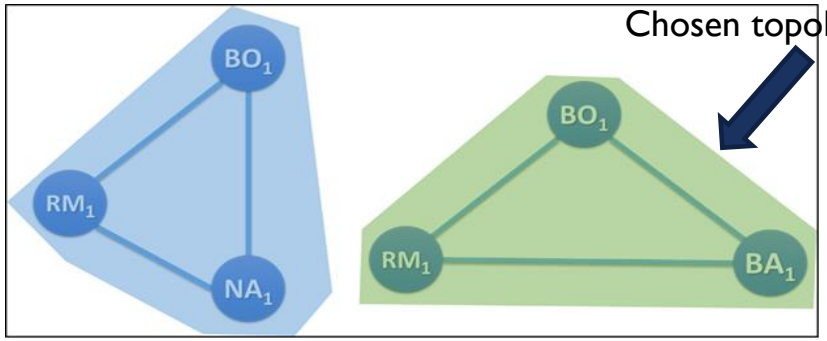
A stronger integration of sites could lead to a reduction of the number of copies



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

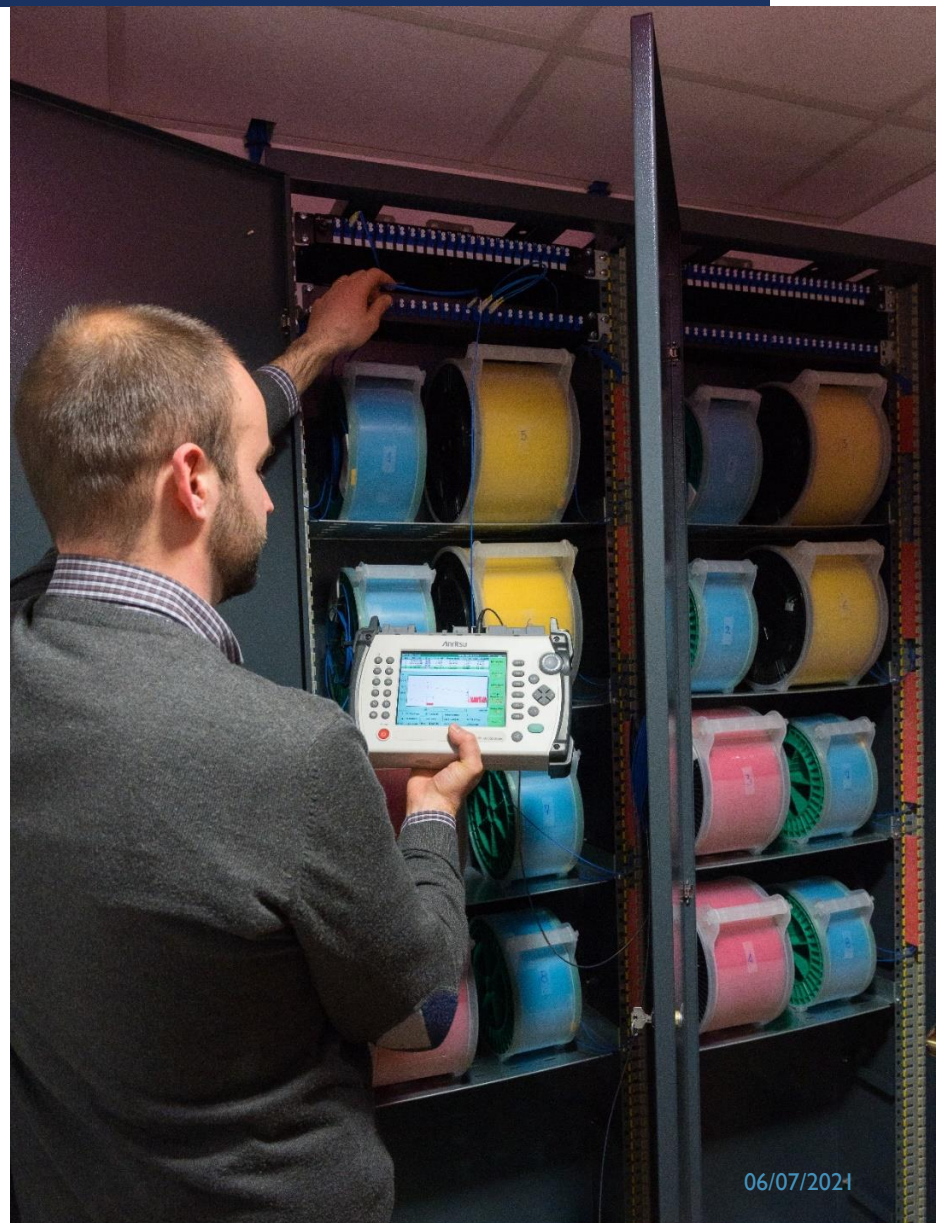
3. **Infinera Groove G30**

   Client: up to 4x100GbE QSFP28

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

GARR LAB: FIBER SPOOLS

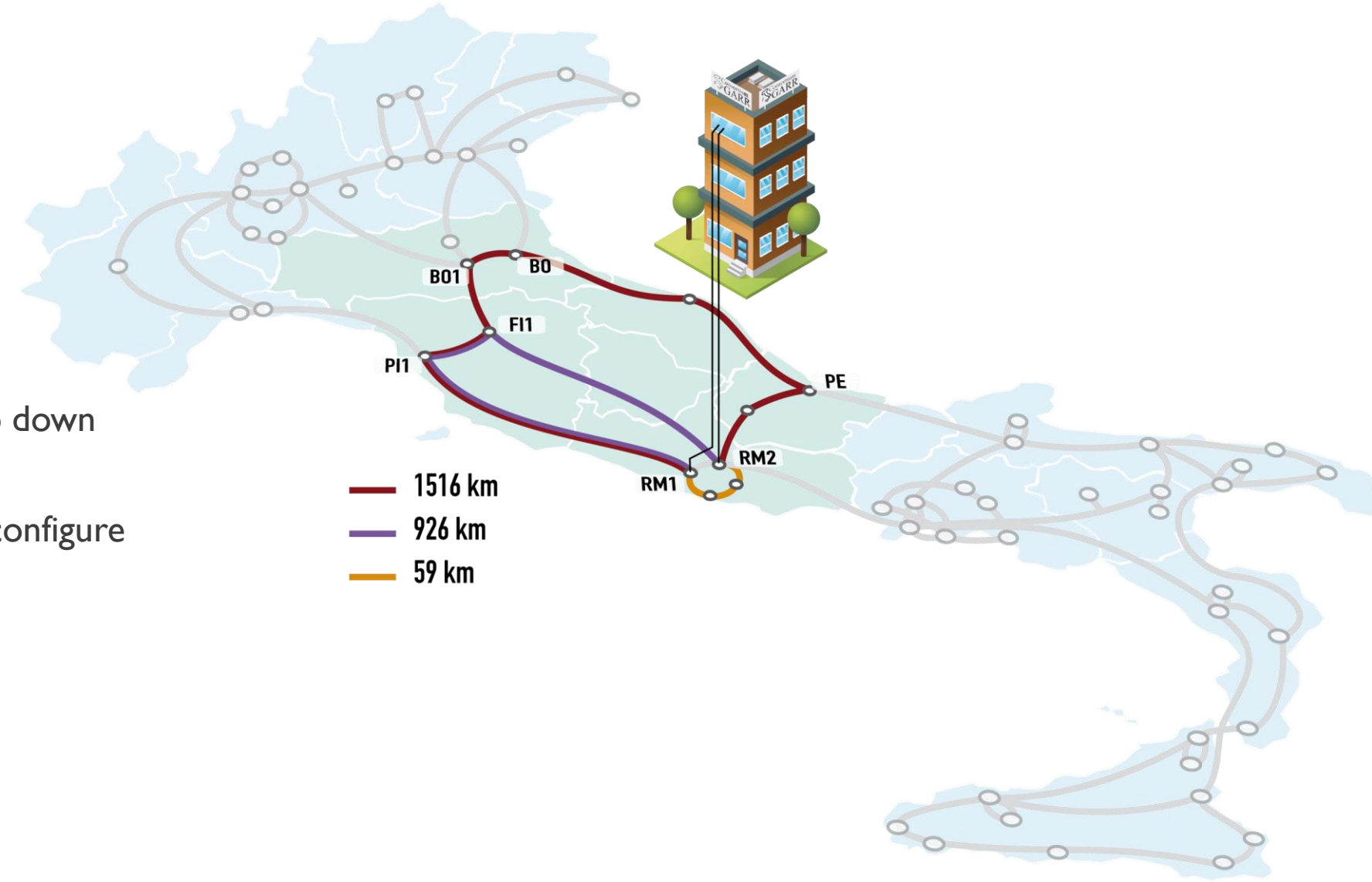
- 16 spools of single G.652d fiber
  - vendor: Fujikura
  - 8 spools 50km -> 4 pairs
  - 8 spools 25 km -> 4 pairs
  - In total 600km single fiber
  - The **flexible topology** changes through a patch panel on the rack top



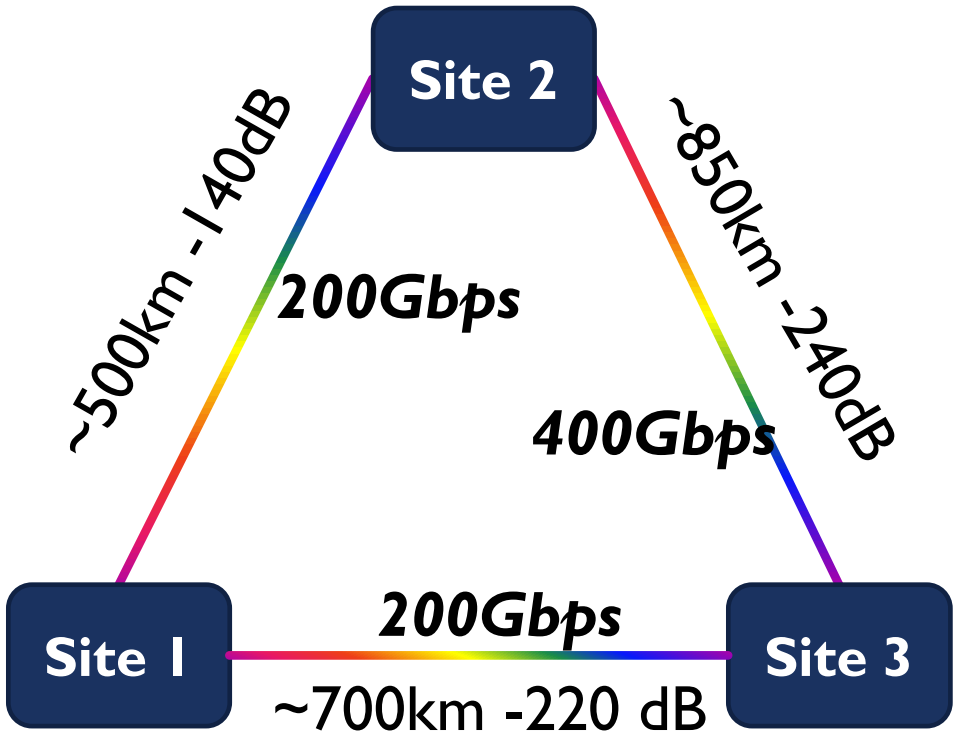
LINKS TOWARD THE PRODUCTION ENVIRONMENT

- 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:

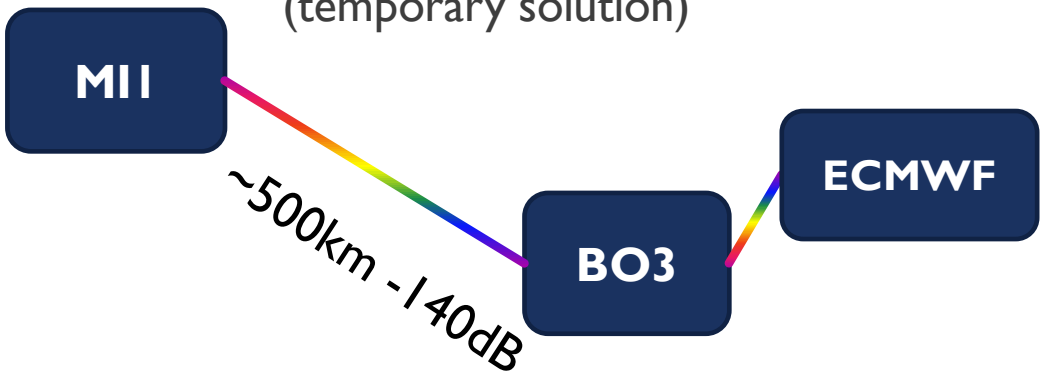
- ~60 km
- ~1000 km
- ~1500 km



- **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

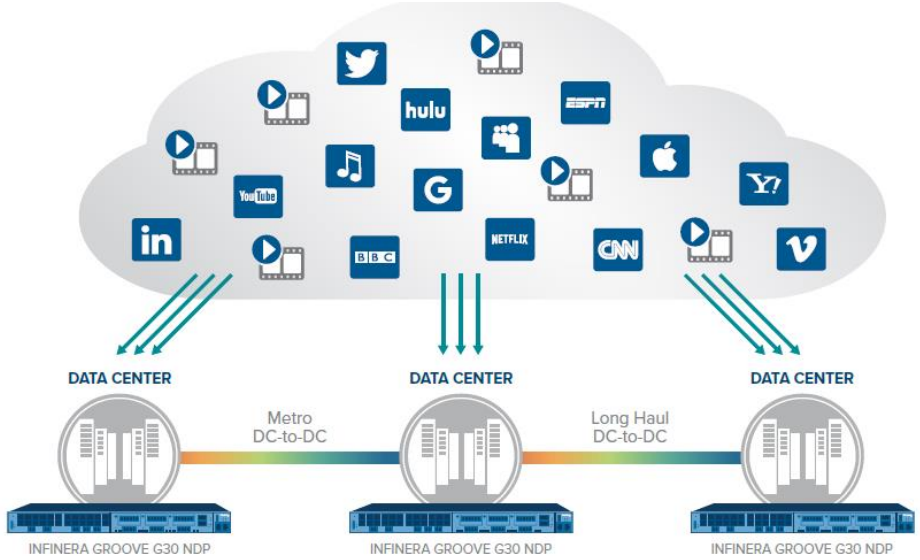


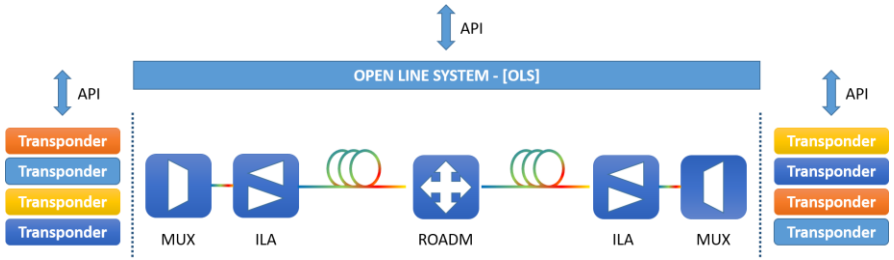
Figure 1: Powering High Performance, Cost-efficient Data Center Connectivity

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)

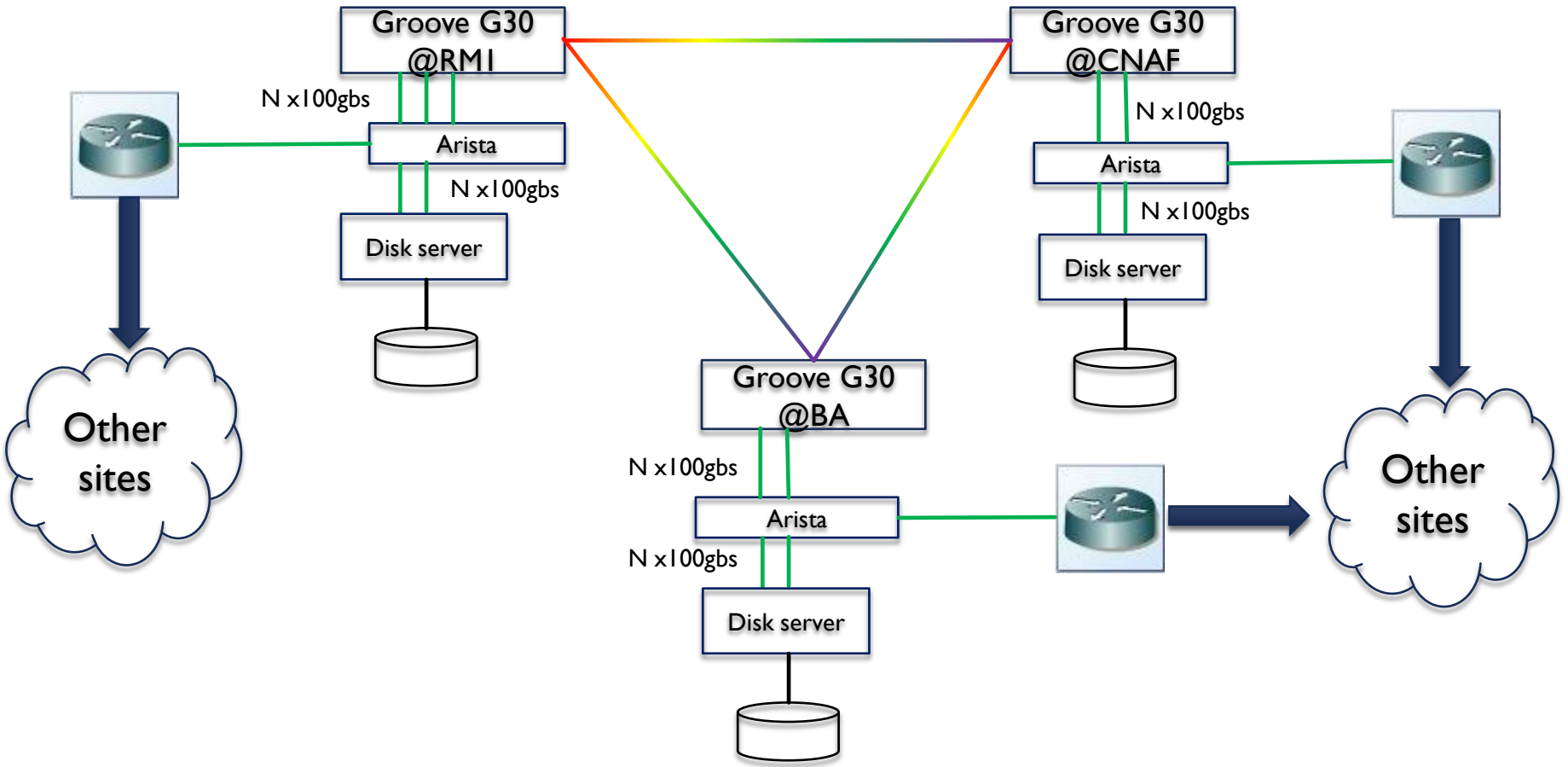
Infinera Groove G30 Muxponder (MUX)



partially disaggregated ➔ decoupling of photonic elements and transponders

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

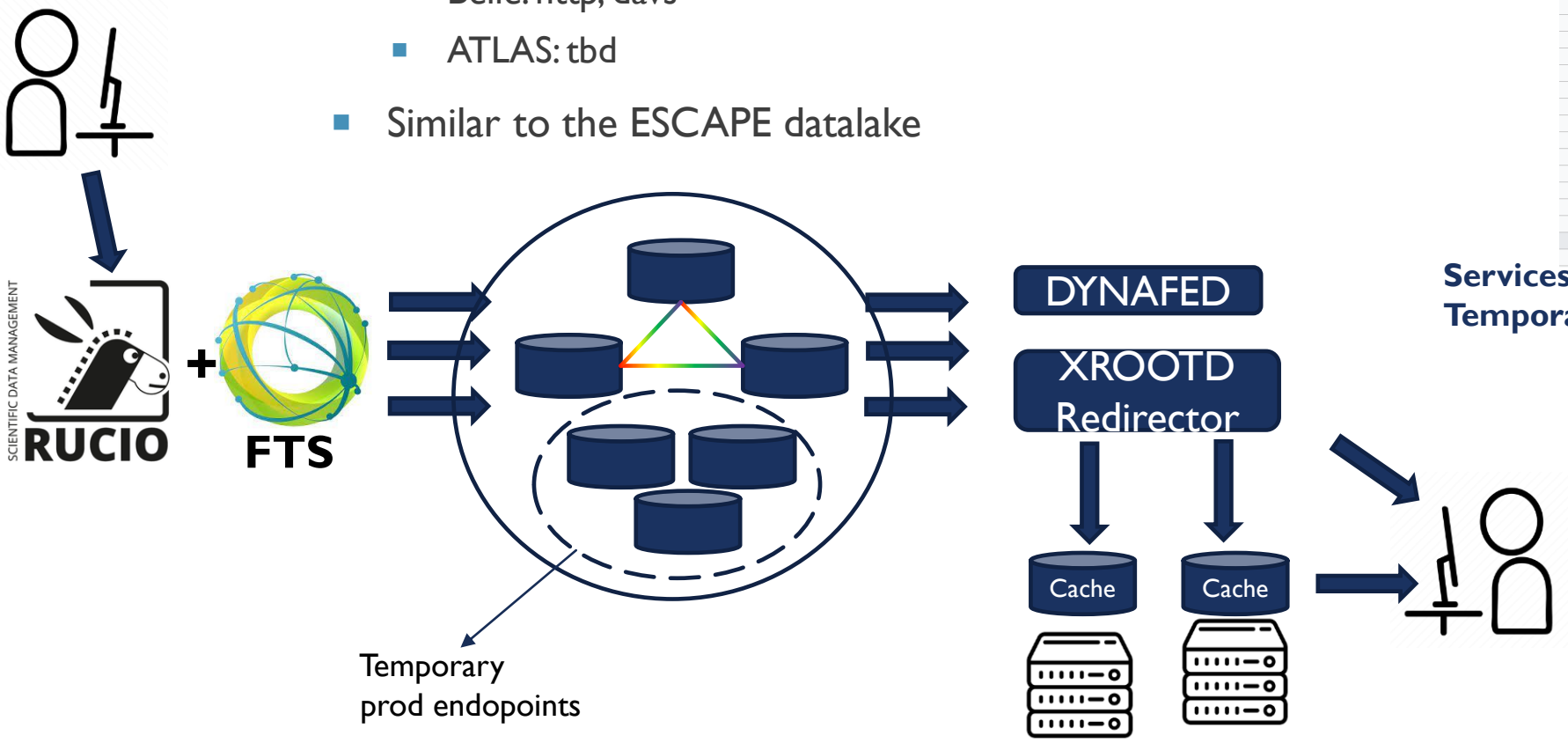
IDDLS ARCHITECTURE



- 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

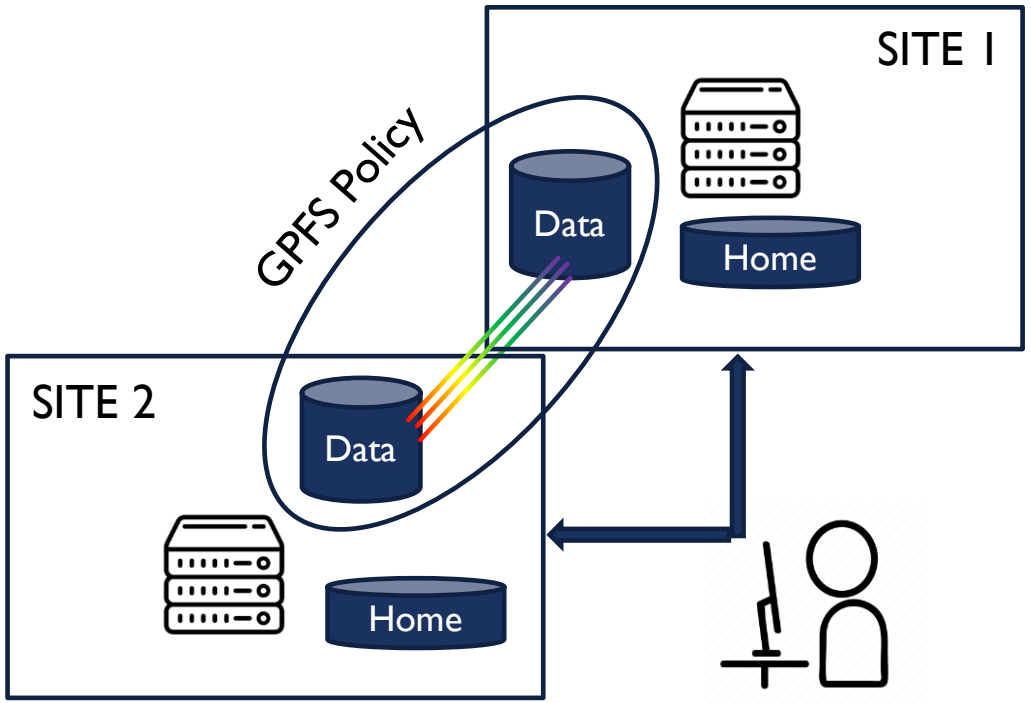


	A	B	C	D	E
1	**Site**	**Storage Technology**	**VO**	**Auth Type**	**Endpoint**
2			**Endpoints**		
3	CNAF	xrootd	CMS		xrootd-cms
4	CNAF	xrootd	ATLAS		xrootd-atlas
5	CNAF	gftp/davs	BELLE		gridftp-storm-archive
6	NA	gftp/davs	BELLE		belle-dpm-01.na.infn.it
7	NA	davs	BELLE		recas-dpm-01.na.infn.it
8	NA		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

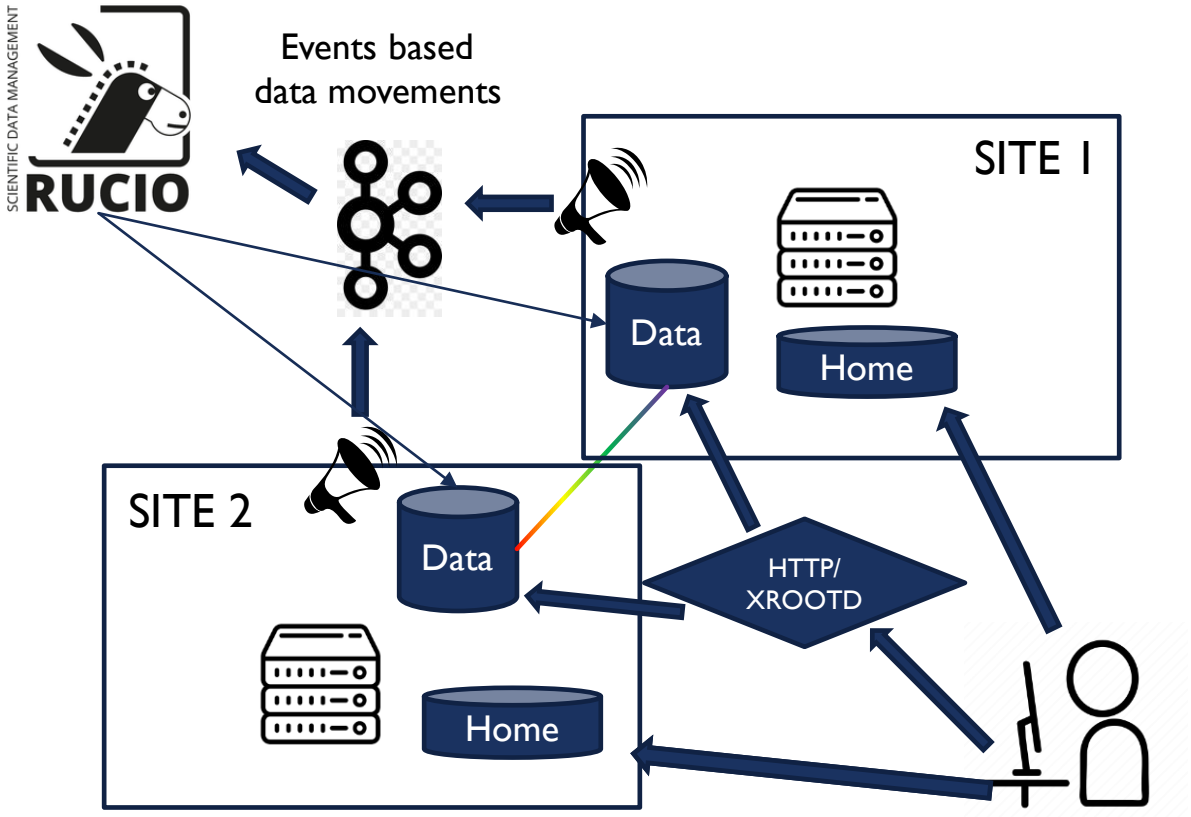
DEPLOYMENT SCENARIO: HA FOR VOS WITHOUT A DISTRIBUTED COMPUTING MODEL

- 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

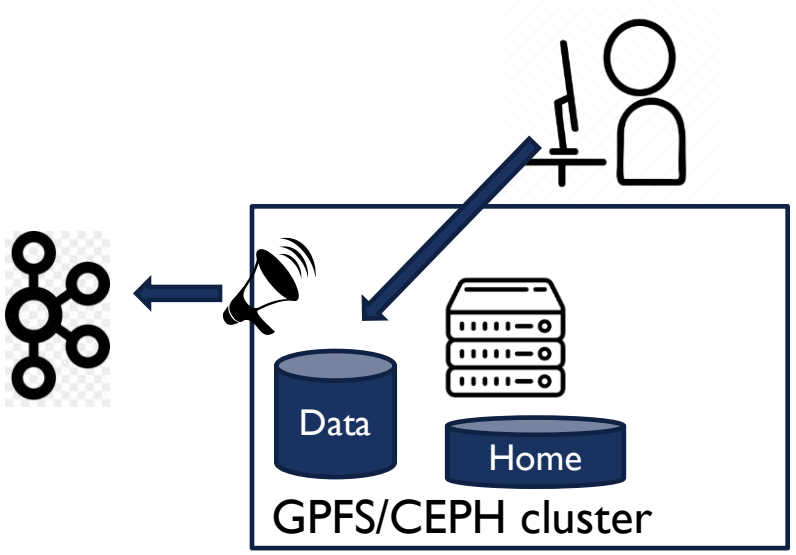


- Transparent creation of redundant federated storage resources – XDC approach
  - automatic replication in different, geographically distributed, availability zones
  - Replica performed at the infrastructure level, without the user intervention via RUCIO+FTS
  - Asynchronous copy but not bounded to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

EXPERIENCE WITH STORAGE EVENTS @ CNAF

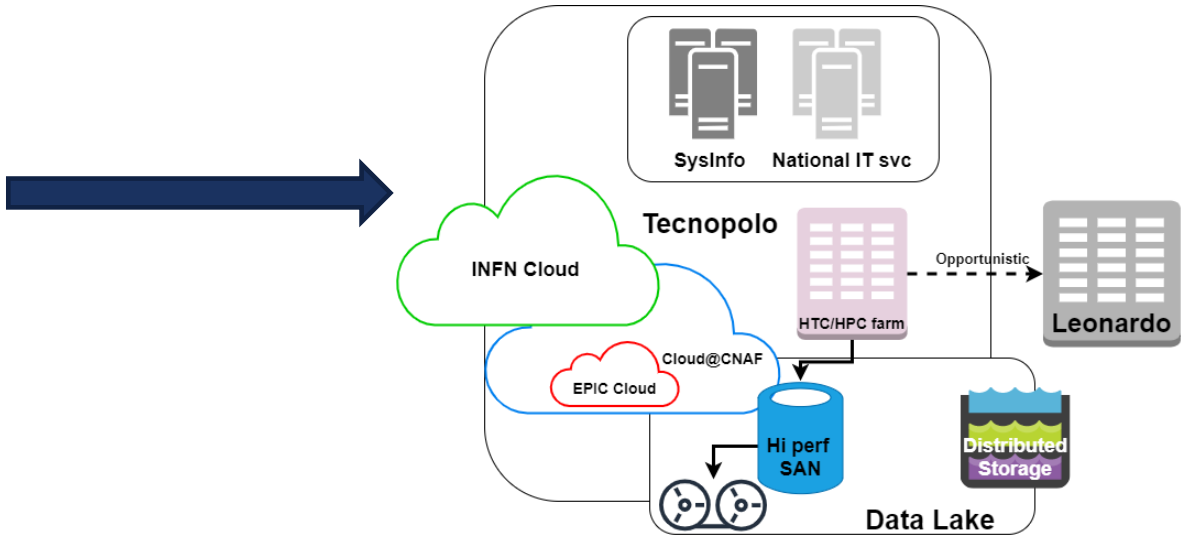
- **Intercept relevant events on the filesystem and receive corresponding notification**
  - File Creation
  - File Write
  - Directory access
- **Inject the notifications into the policy manager system** to trigger data movements
  - Within the site or the Lake, including QoS change
    - QoS change 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

STATO ASSEGNAZIONI 2021

Sez. & Suf.	MISS				CON				ALTRICONS				TRA				SEM				PUB				MAN				INV				LIC-SW				APP				SPSERVIZI				TOTALE			
		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.		Sj	Dot.	Ant.
BA	1.5																																									1.5				
	0.0																																									0				
CNAF	1.5																													30												1.5	30			
	1.0																												10.0	0.0		5.0										11.0			5	
LNF	1.5																																									1.5				
	0.0																																									0				
LNL	1.5																																									1.5				
	0.0																																									0				
NA	1.5																																									1.5				
	0.0																																									0				
PG	1.5																																									1.5				
	1.0																																									1.0				0
PI	1.5																																									1.5				
	0.0																																									0				
RM1	1.5																																									1.5				
	0.0																																									0				

5k Ant. Completata infrastruttura per siti datalake, 3x switch + 3x server + ottiche

Sigla Loc.	Capitolo	Riunione	Note Alla Richiesta	Rich.	Rich. SJ	Assegn.	Assegn. SJ	Assegn. Dot.	Commento Alla Assegnazione
CNAF	MISS	Assegnazioni	Missioni italiane presso i partner del progettoÂ§	1.5	0.0	1.0			
	Totale MISS			**1.5**	**0.0**	**1.0**	**0.0**	**0.0**	
	INV	Assegnazioni	15k Tasca per acquisto almeno 3 server CONSIP per installazione su siti esterni al datalake. SJ a creazione del datalake. Â§si veda allegato per dettagli	0.0	15.0	0.0			per ora no siti esterni
		Assegnazioni	10k tasca per potenziamento server storage sulle sedi. SJ all'installazione e configurazione dei server acquistati su fondi 2020. Â§si veda allegato per dettagli	0.0	10.0	10.0			alta priorita, anticipo
		Assegnazioni	3 x transceiver 100gb arista per potenziamento link del datalake. SJ alla creeazione del datalakeÂ§si veda allegato per dettagli	0.0	5.0	0.0			serve solo per arrivare a 200GB. siamo lontani ancora. anticipo
	Totale INV			**0.0**	**30.0**	**10.0**	**0.0**	**0.0**	
Totale CNAF				**1.5**	**30.0**	**11.0**	**0.0**	**0.0**	
Totale Generale IDDLS				**1.5**	**30.0**	**11.0**	**0.0**	**0.0**	

Ancora disponibili

STATO ASSEGNAZIONI 2020

Sigla Loc.	Capitolo	Riunione	Note Alla Richiesta	Rich.	Rich. SJ	Assegn.	Assegn. SJ	Assegn. Dot.	Commento Alla Assegnazione
CNAF	MISS	Assegnazioni	missioni di progetto inclusa una missione per ws escapeÂ§	1.5	0.0	0.0			
		Assegnazioni	Assegnazione per mobilità di Francesco Giacomini, Osservatore in CSN5	1.5	0.0	1.5			
		Totale MISS		**3.0**	**0.0**	**1.5**	**0.0**	**0.0**	
	INV	Anticipi	3 x transceiver 100gb arista per potenziamento link del datalake. SJ alla creeazione del datalakeÂ§si veda allegato per dettagli	5.0	0.0	5.0			serve solo per arrivare a 200GB. siamo lontani ancora. anticipo
		Totale INV		**5.0**	**0.0**	**5.0**	**0.0**	**0.0**	
	APP	Assegnazioni	Switch per connessione banda larga per terzo sito (anno I finanziato per due sedi) SJ alla finalizzazione degli acquisti del primo anno di porgetto. In allegato datasheet ed offerta su MEPA della soluzione sceltaÂ§	0.0	15.0	0.0	15.0		a valle dell'esito della gara del 2019
		Assegnazioni	Server per connessione a banda larga munito di scheda ethernet 100gbs e ottiche rlative. sj all'acquisto contemporaneo dello switch per il terzo sito.Â§	0.0	7.5	0.0			Anticipo al 2019
		Assegnazioni	4x100gbe optical tranceiverÂ§	0.0	1.5	0.0			Anticipo al 2019
		Giugno	sblocco sub-judice a valle della gara per acqusito switch effettuata nel 2019. Con lo sblocco potremo acquisire i server da connettere agli switch come da programma per completare l'infrastruttura IDDLS:	15.0	da sj 15.0	15.0	da sj 15.0		
		Totale APP		**15.0**	**9.0**	**15.0**	**0.0**	**0.0**	
Totale CNAF				**23.0**	**9.0**	**21.5**	**0.0**	**0.0**	
Totale Generale IDDLS				**23.0**	**9.0**	**21.5**	**0.0**	**0.0**	

SPESE 2020

Dettaglio capitolo U2020105001 - Attrezzature Scientifiche/

Filtra le richieste

Tipo Pratica: Seleziona | Numero Pratica: | Tipo Documento: Seleziona

Data Iniziale: | Data Finale:

Descrizione: | Intestatario:

Filtra

Data	Tipo	Num	Doc	Intestatario	Descrizione	Assegn.	SubJudice	Iniziale	Variaz.	Totale
14/12/2020	Ordini	2260	Impegno contabile	MACTRONICS.IT S.R.L.	2 x Nodi di calco	0,00	0,00	20.000,00	0,00	20.000,00
09/10/2020			Variazione di bilancio		Variazione di Bil	5.000,00	0,00	0,00	0,00	5.000,00
04/11/2020			Variazione di bilancio		Variazione di Bil	15.000,00	0,00	0,00	0,00	15.000,00

Esporta

1/1 10

→ I due server mancanti

SPESE 2019

Data	Tipo	Num	Doc	Intestatario	Descrizione	Assegn.	SubJudice	Iniziale	Variaz.	Totale
07/01/2019			Stanziamento iniziale (assegnazioni)		Stanziamento iniz	45.000,00	0,00	0,00	0,00	45.000,00
04/11/2019			Variazione di bilancio		Variazione di Bil	49.000,00	0,00	0,00	0,00	49.000,00
31/07/2019			Variazione di bilancio		Variazione di Bil	-40.000,00	0,00	0,00	0,00	-40.000,00
07/01/2019			Stanziamento iniziale (assegnazioni)		Assegnazione Vinc	0,00	40.000,00	0,00	0,00	40.000,00
31/07/2019			Variazione di bilancio		Assegnazione Vinc	0,00	-40.000,00	0,00	0,00	-40.000,00
18/12/2019	Ordini	2081	Impegno contabile	CONVERGE SPA	4x Server per sto	0,00	0,00	5.322,00	0,00	5.322,00
10/01/2020	Ordini	2098	Impegno contabile	VISTA TECHNOLOGY SRL	3x Switch/Router	0,00	0,00	48.667,02	0,00	48.667,02

Uno dei tre server Partecipazione in altro acquisto consip

3 switch per le sedi

EVENTUALI AFFERENZE 2022

- Tutte le sedi INFN hanno (circa) confermato le percentuali 2021
  - FTE: BARI: 0.5 LNF:0.15, LNL:0.60, NA:1, PG:0.95, PI:0.35, RM1:0.05

CNAF – Proposta TBD

Nome	Aff %
Cesini	50
Falabella	10
Fattibene (resp loc.)	10
Sapunenko	20
Vianello	10
Soares	10
Vistoli	10
Rendina	30
Rega	20
TOT	170%

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

MILESTONE 2020-2021

30-06-2020: Installazione apparati lato GARR nelle sedi INFN ==> 50%

Per fine Settembre è previsto il completamento con il deployment sulle sedi

Al momento gli apparati sono in test al GARR

Tutti gli apparati HW necessary per il prototipo (con link a 100gbit/s sulle sedi) sono stati acquisti

31-07-2020 : Creazione del prototipo DataLake su apparati DCI e standard solo su WAN 30%

Iniziata su apparati standard in attesa del deployment su apparati del DataLake

31-12-2020: run per la valutazione delle performance sui prototipi ==> 0%

Di fatto sono diventate milestone 2021.

SHIFT MILESTONE 2021 A 2022

Descrizione	Data completamento	
Potenziamento infrastrutture storage nei siti del datalake	01-05-2021	31/12/2021
Deployment dei server su siti esterni al datalake e configurazione sistemi di caching sugli stessi per accesso trasparente ai dati	31-07-2021	31/12/2022
Valutazione delle performance del sistema IDDLS con applicazioni relai degli esperimenti	31-12-2021	31/12/2022

RICHIESTE 2022

- Avendo a disposizione ancora I fondi non anticipati 2021 (10k) non prevediamo richieste hw significative
  - In fase di verifica se abbiamo tutti I component per la rete (ottiche + cavi)
- Riproporremmo le richieste missioni come in 2021
- Resta in sospeso la questione potenziamento infrastruttura siti esterni al datalake
  - Da investigare con CCR come discusso lo scorso anno