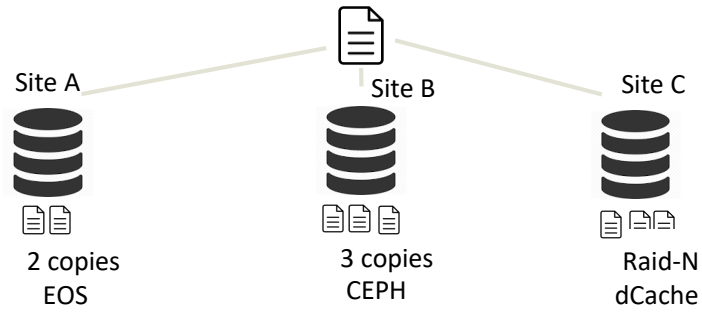
IDDLS: Italian Distributed Data Lake for Science

Stato progetto 20/12/2019

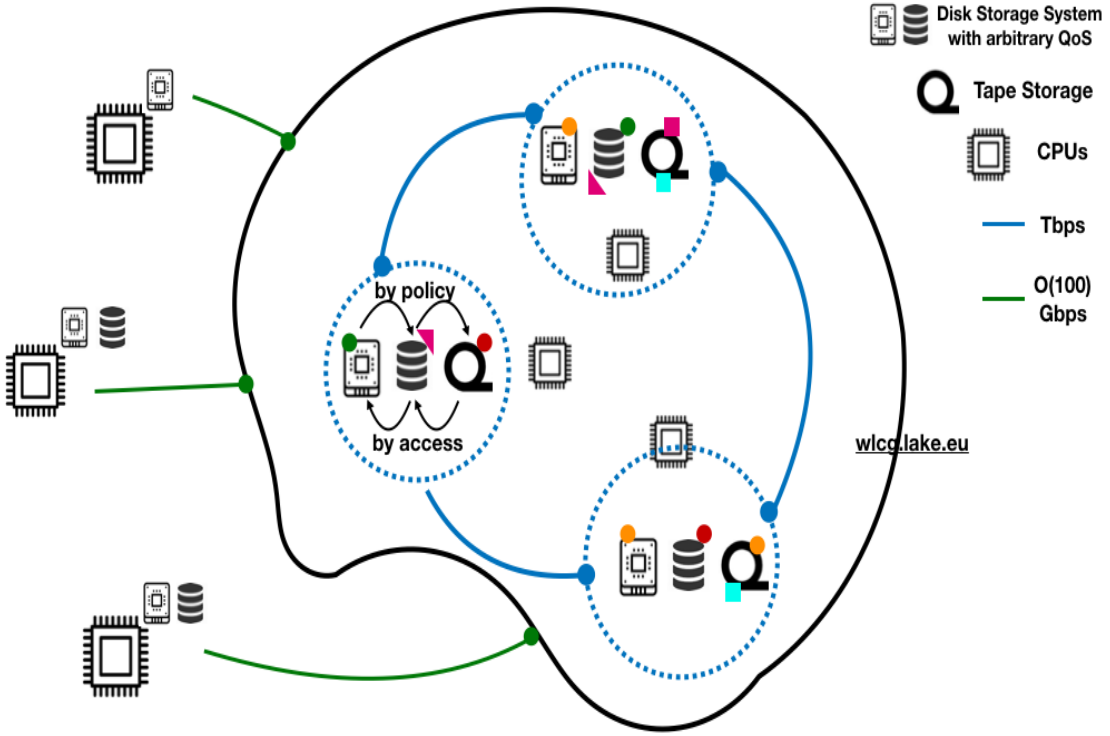
Some ideas on reducing Storage costs

- 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



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

- Reduce Operational Cost: deploy fewer (larger) storage services 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
    - Smart Caching

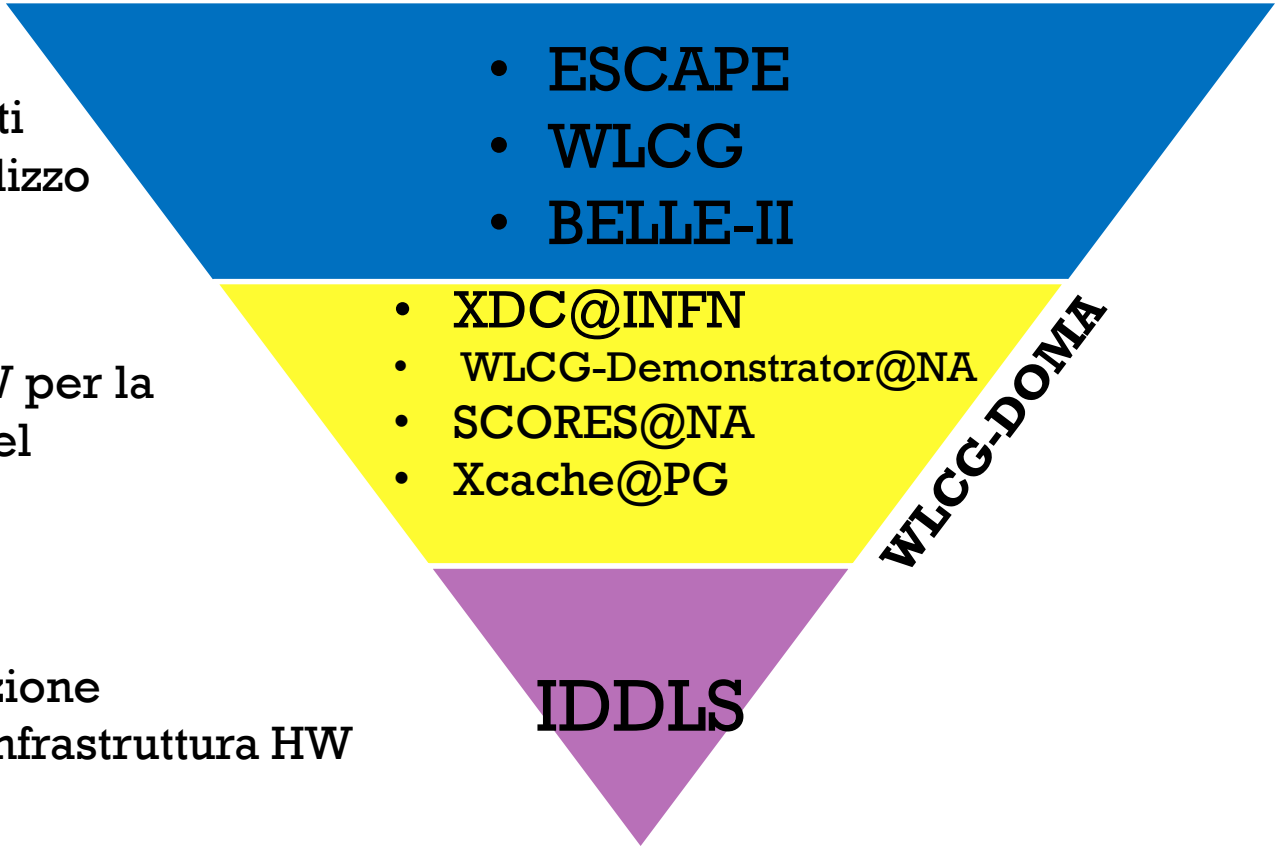


Sinergie in Europa

Comunità di utenti interessate all'utilizzo della tecnologia

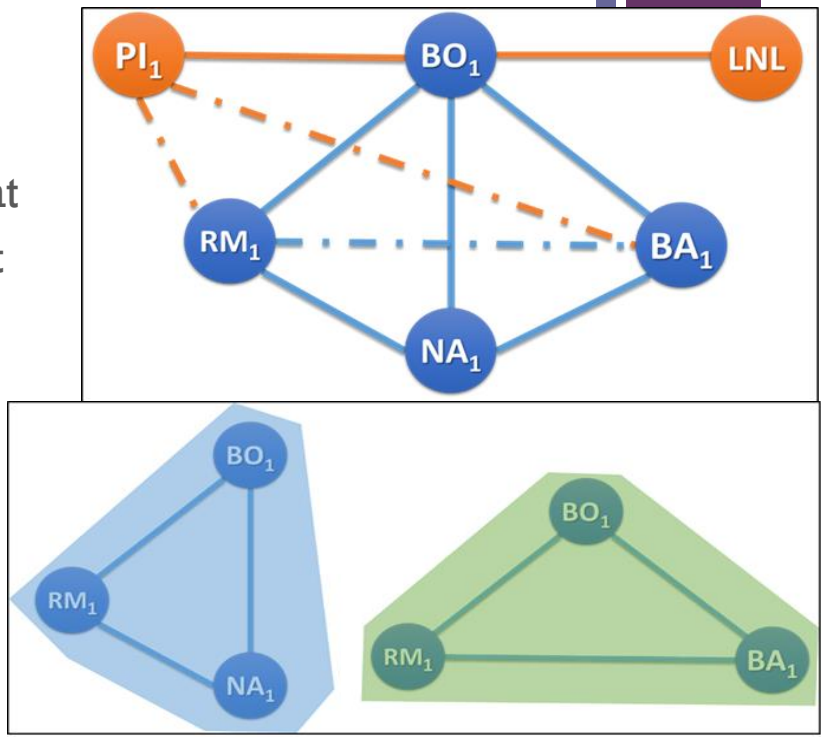
Sviluppo SW per la creazione del DataLake

Creazione dell'infrastruttura HW



The project

- INFN-GARR collaboration to realize a prototype of an Italian DataLake 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
    - SCoRES project (INFN-NA)
  - Real life use cases for testing
    - CMS
    - ATLAS
    - BELLE-II
    - Possibly involving LNGS experiments (XENON) and VIRGO



Possible topologies of the GARR Network with DCI and SDN for the DataLake

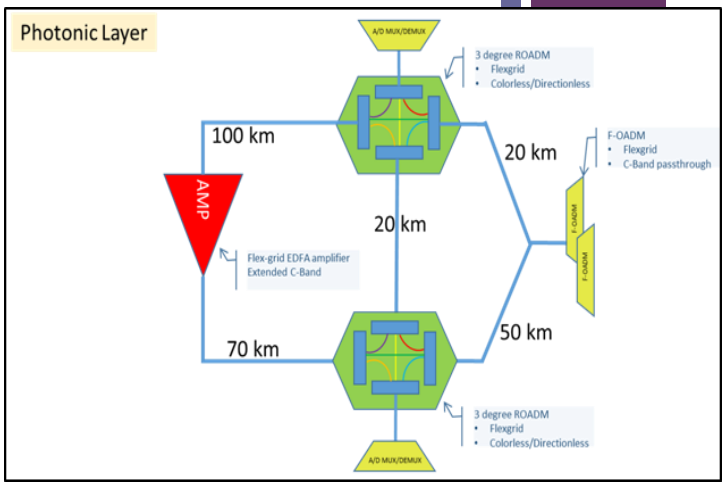
Timeline

- **3 years project**

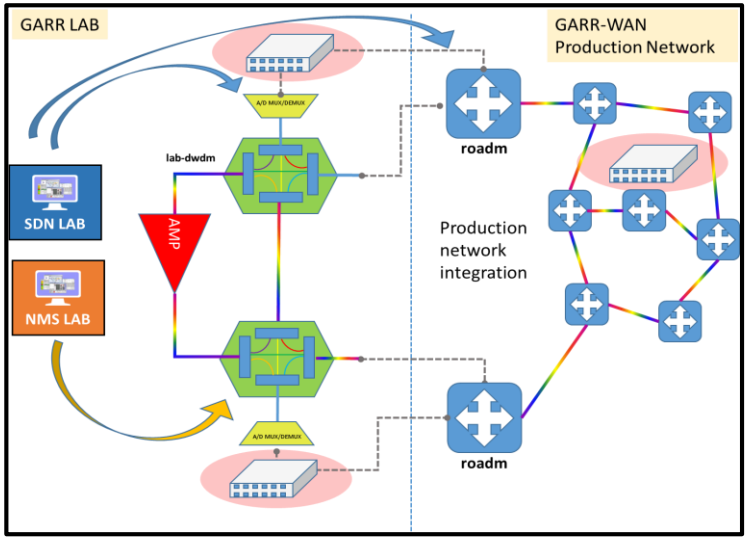
- First year
  - Technology scouting for DCI equipment to be deployed by GARR
  - Application (INFN) requirements analysis
  - Network equipment acquisition (INFN and GARR) and Lab testing
  - Deployment on mixed Lab+WAN environment of the networking equipment
  - Creation of the DataLake on sites connected with standard networking and first prototype using DCI

- Second year
  - Testing of the mixed (Lab+WAN) configuration
  - Final creation of the DataLake on the 3 INFN sites with DCI systems
  - Performance evaluation and comparison
  - Possible acquisition of new equipment with increased performance

- Third year
  - Deployment only on WAN of the networking equipment
  - Optimization of the DataLake
  - Performance evaluation
  - Final consideration



Lab deployment at GARR for testing



Mixed Lab+WAN deployment

Milestone 2019 e stato

- **Gennaio 2019: Kickoff meeting @GARR**
  - https://agenda.infn.it/event/17957/
- **Gennaio – Giugno:** test apparati trasmissivi in Lab GARR (slide successive)
- **30/06/2019: Scelta degli apparati di networking per la creazione del DataLake**
  - Apparati GARR testati
    - Infinera Groove G30
    - Transponder TP100GOTN XTM Infinera
    - JunipAer CX6360 (and MX240)
  - Switch con porte a 100Gb per connessione lato datacenter valutati:
    - Arista 7050SX3-48YC8, Arista716032CQ, DCS-7050CX3-32S-F
    - Extreme Networks SLX930
    - NEXSUS 93180
  - Gara in corso – Da assegnare – hanno partecipato Arista e Cisco
- **31/12/2019: Deployment degli apparati** di rete in una configurazione Lab+WAN
  - primi portotipi DataLake su apparati DCI e standard
- **31/12/2019: Primi run per la valutazione delle performance sui prototipi**

Apparati lato GARR

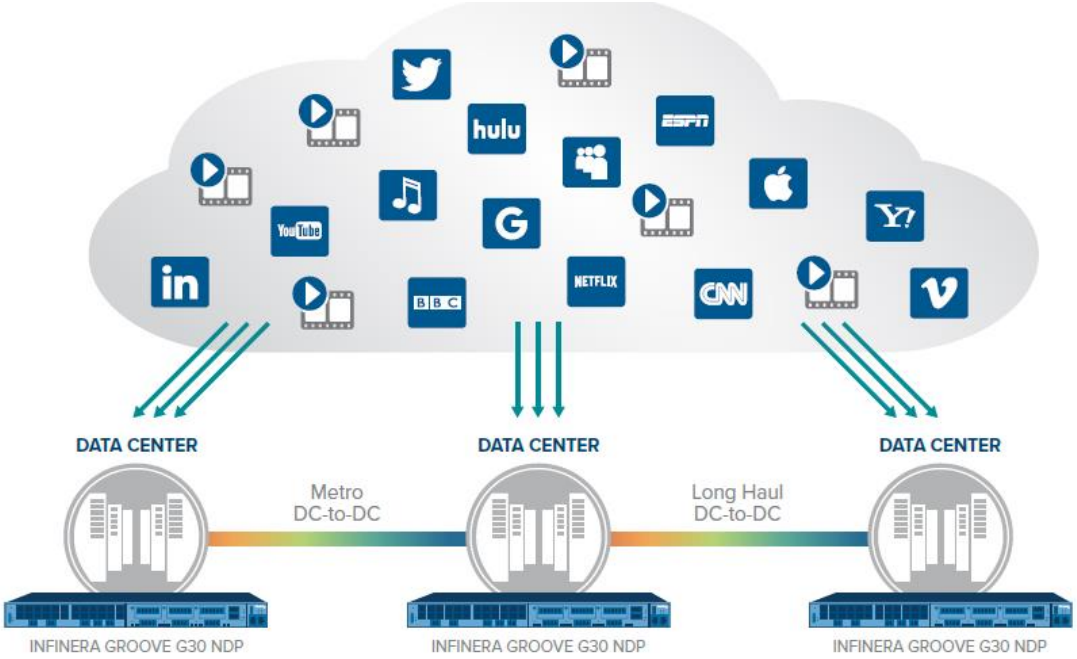


Figure 1: Powering High Performance, Cost-efficient Data Center Connectivity

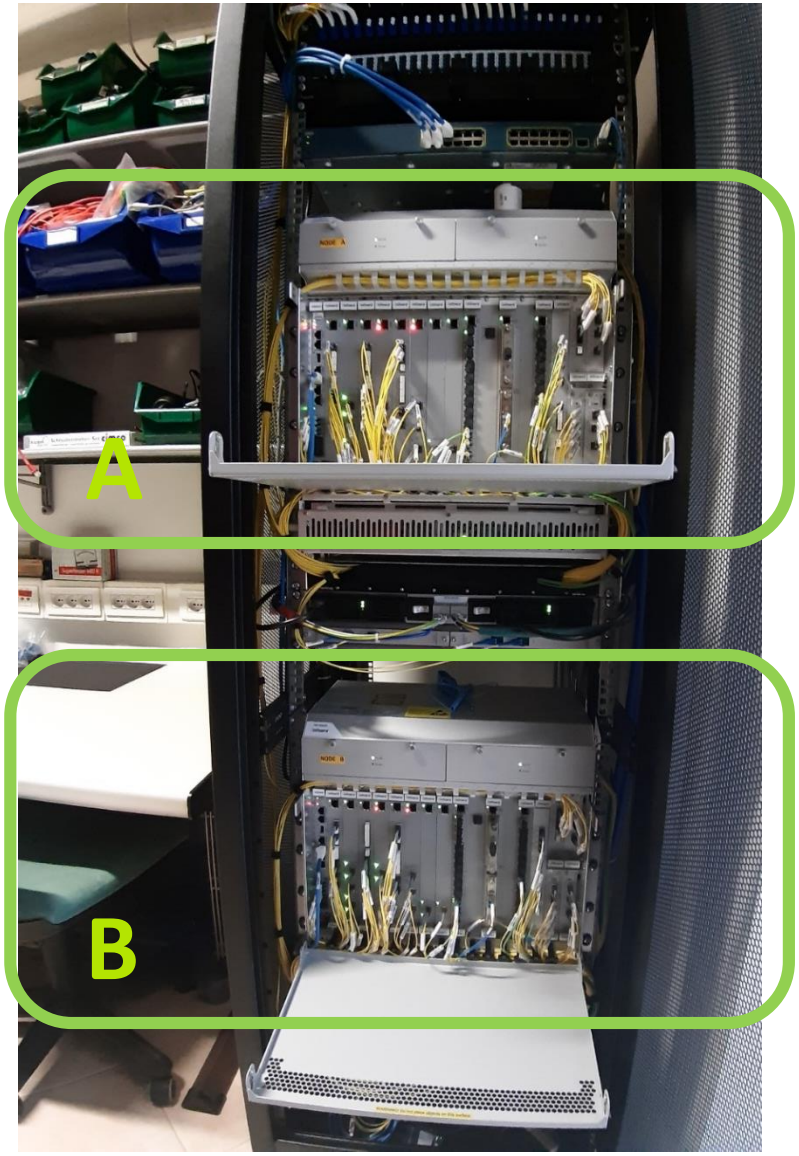
THE PURPOSE-BUILT INFINERA GROOVE G30 NETWORK DISAGGREGATION PLATFORM

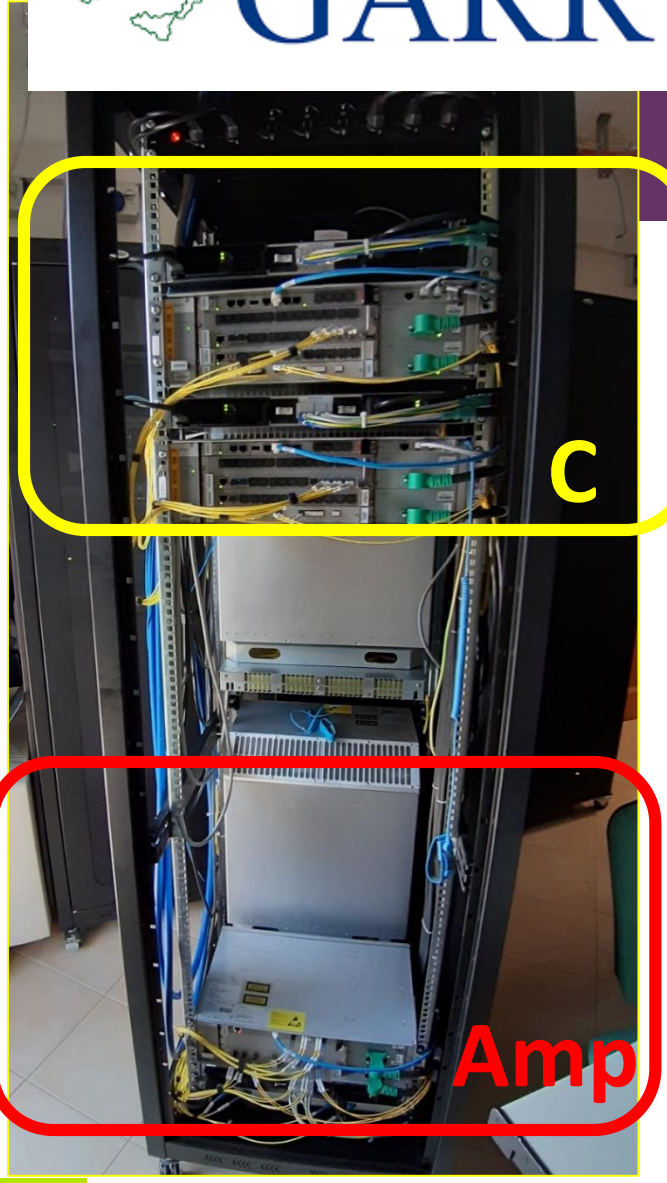
The Infinera Groove G30 Network Disaggregation Platform (NDP) is an innovative 1RU modular open transport solution for cloud and data center networks that can be equipped as a muxponder terminal solution and as an Open Line System (OLS) optical layer solution. Purpose-built for interconnectivity applications,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

Infinera Groove G30 Open Line System (OLS)

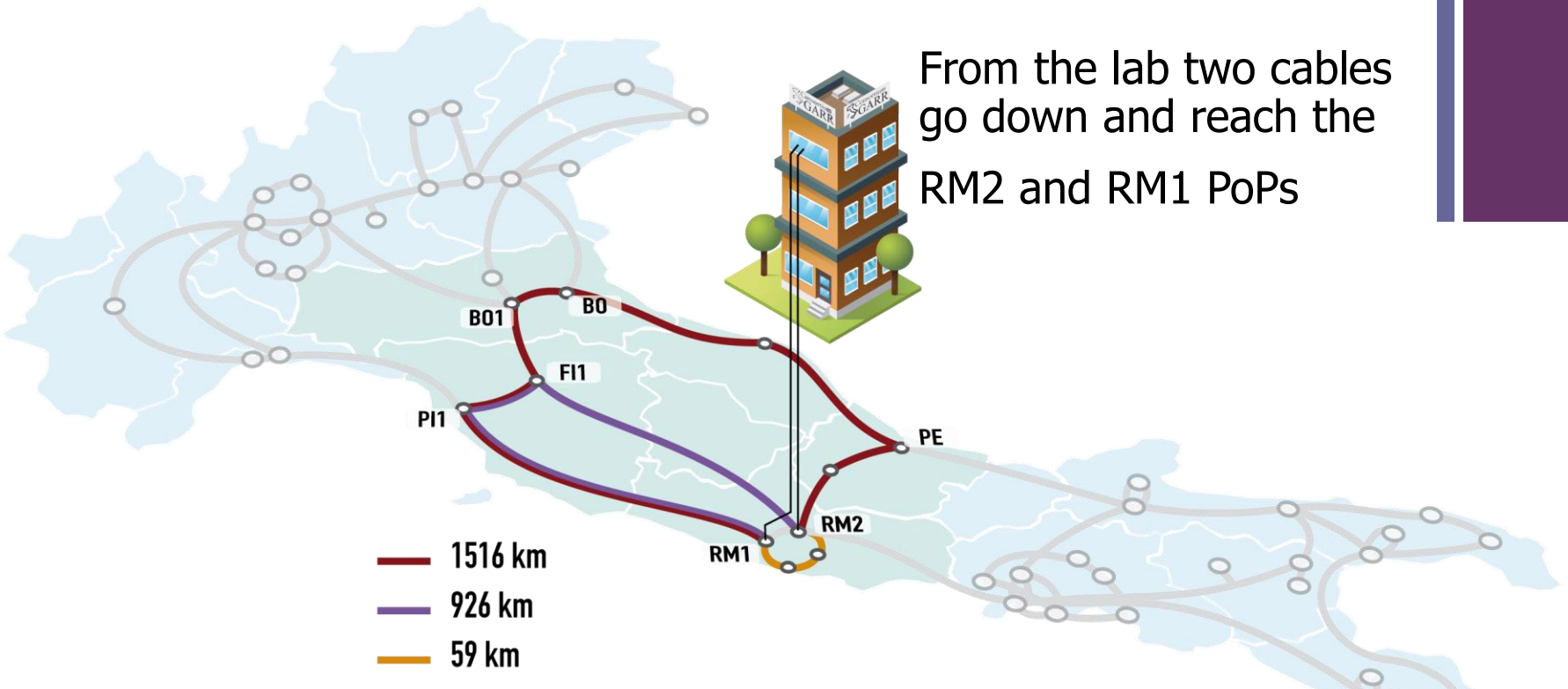
Infinera Groove G30 Muxponder (MUX)

Slides from GARR…





Links towards the production environment



- From these PoPs we can configure links that are as long as:
  - ~60 km
  - ~1000 km
  - ~1500 km

Consortium GARR

Test transponders 100G

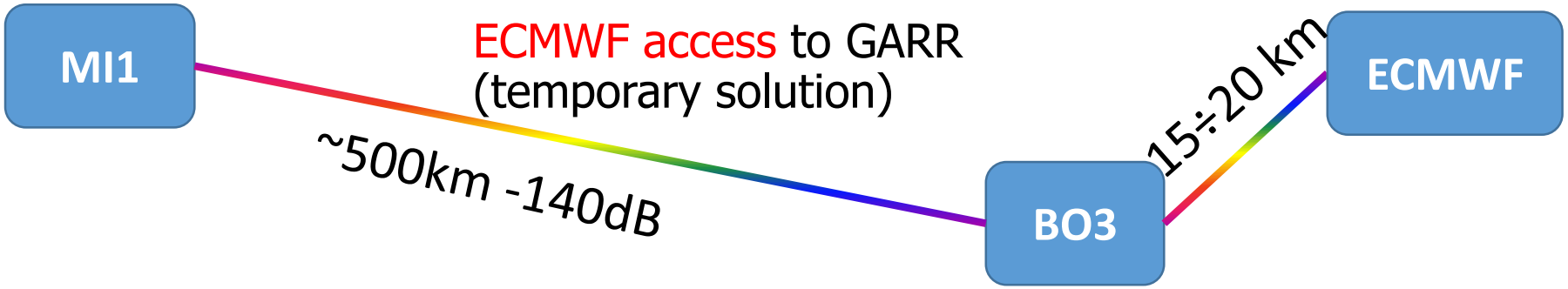
1. Transponder TP100GOTN XTM Infinera
   Client: 100G Ethernet
   Line: CFP1 100G ACO - QPSK coherent

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)

Consortium GARR

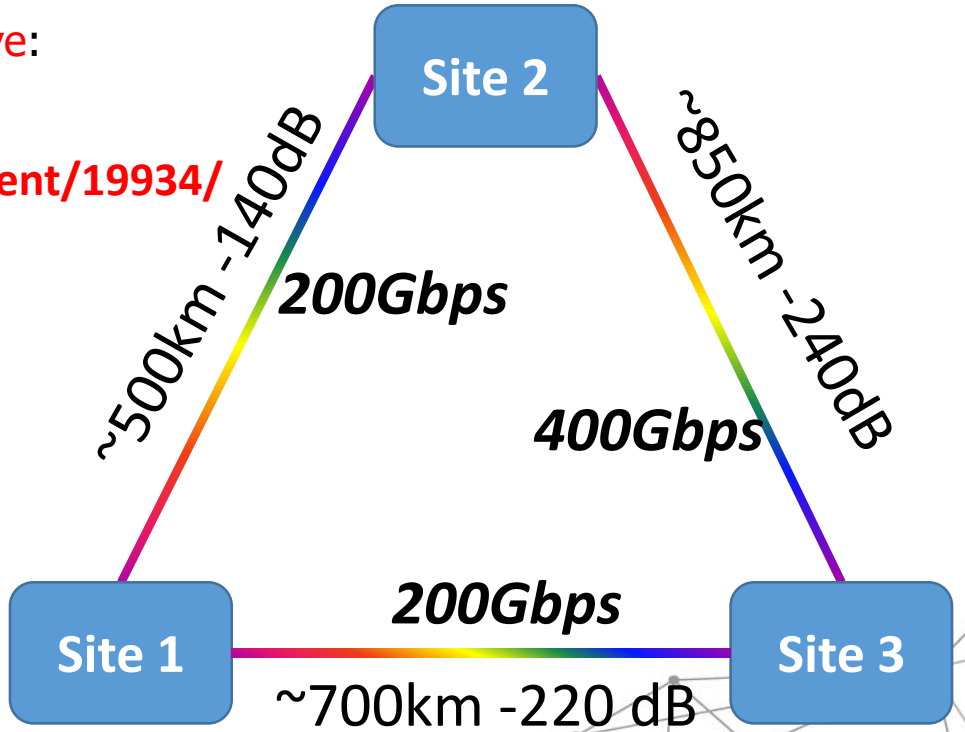
Which transponder for 100GE



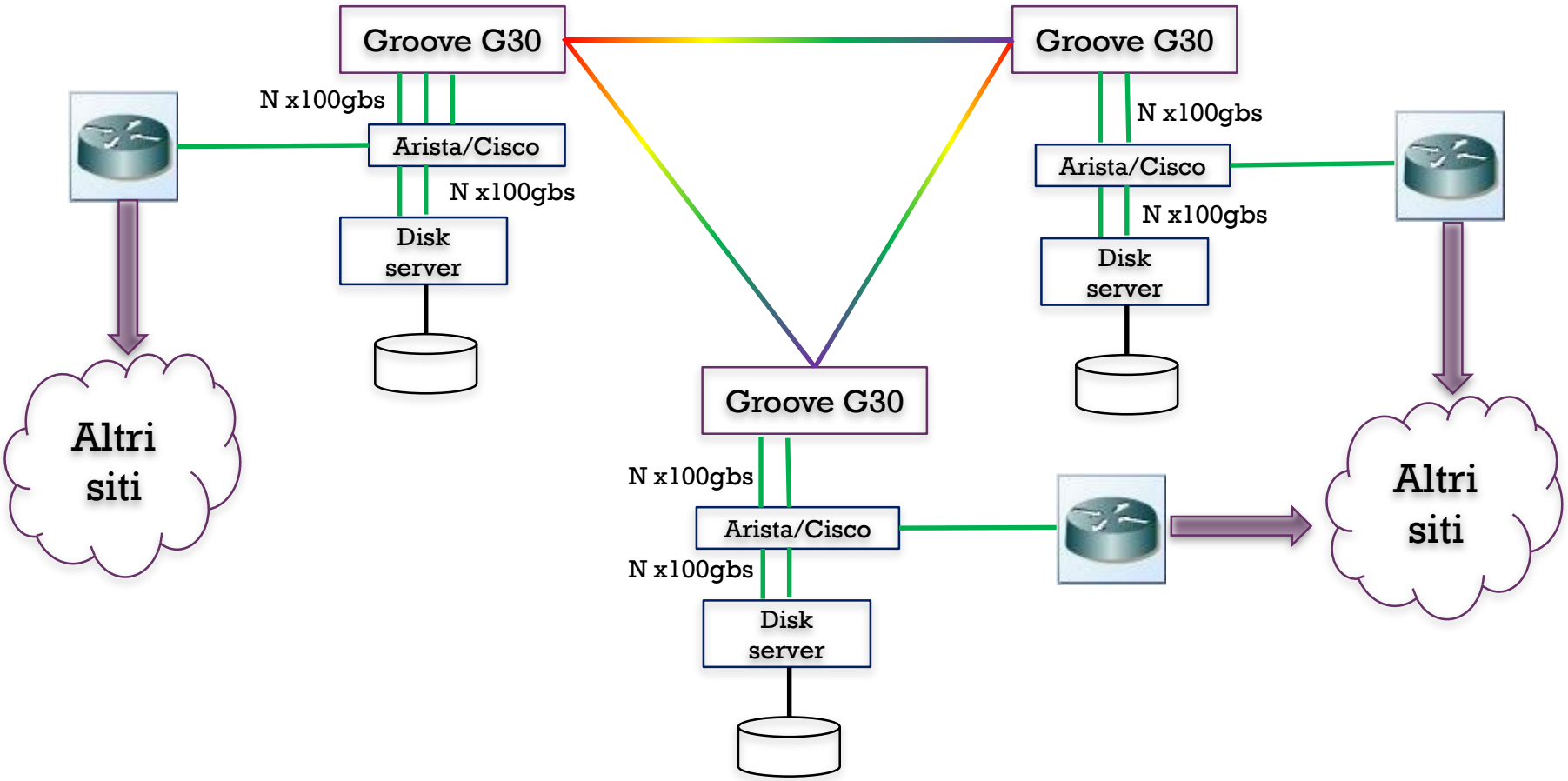
For the Italian Distributed Data Lake initiative:
Large bandwidth over hundreds of kms

See full test details: https://agenda.infn.it/event/19934/

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



IDDLS Network Architecture



Apparati lato INFN

- 2 opzioni per Switch (gara da asseganre):
  - 3xArista 7050SX3-48YC8
    - 32x100Gb/s QSFP+ + 2x10Gb/s SFP+
    - + 9 ottiche 100Gb/s
    - manutenzione
  - 3xCisco Catalyst9500-32C-A
    - 32x100Gb/s QSFP+ + 2x10Gb/s SFP+
    - + 9 ottiche 100Gb/s
    - manutenzione
  - Da definire apparati di storage per test scala seconda parte da progetto
    - 3 disk server finanziati dal progetto
  - Storage da definire
    - recuperarlo da dismissioni/produzione
    - Piccoli sistemi SSD

Budget Assegnato INFN CSN5

2019

Sez. & Suf.	MISS				INV			
		Sj	Dot.	Ant.		Sj	Dot.	Ant.
BA	1.5 0.5							
CNAF	3.0 1.5				65.0 5.0	40.0		
LNF	1.5 0.0							
LNL	1.5 0.5							
NA	1.5 0.5	0.5 0.0						
PI	1.5 0.5							
RM1	1.5 0.0							
TOTALE	12	0.5			65			
	12.5				65			
	3.5		0	0	5	40	0	0
	3.5				45.0			

2 switch + 2 server + 2 NIC

2020

Sez. & Suf.	MISS				APP		
		Sj	Dot.	Ant.	Sj	Dot.	Ant.
BA	1.5 0.0						
CNAF	3.0 1.5				24 15.0		9.0
LNF	1.5 0.0						
LNL	1.5 0.0						
NA	2.0 0.0						
PG	1.5 1.5						
PI	1.5 1.5						
RM1	1.5 0.0						
TOTALE	14				0	24	
	14				24		
	4.5		0	0	0	15	0	9
	4.5				15.0		

Anticipo 2019 per 1 server

Terzo switch

1. 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§
2. 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§
3. 4x100gbe optical tranceiverÂ§

Budget GARR: 200k in-kind per 3 apparati

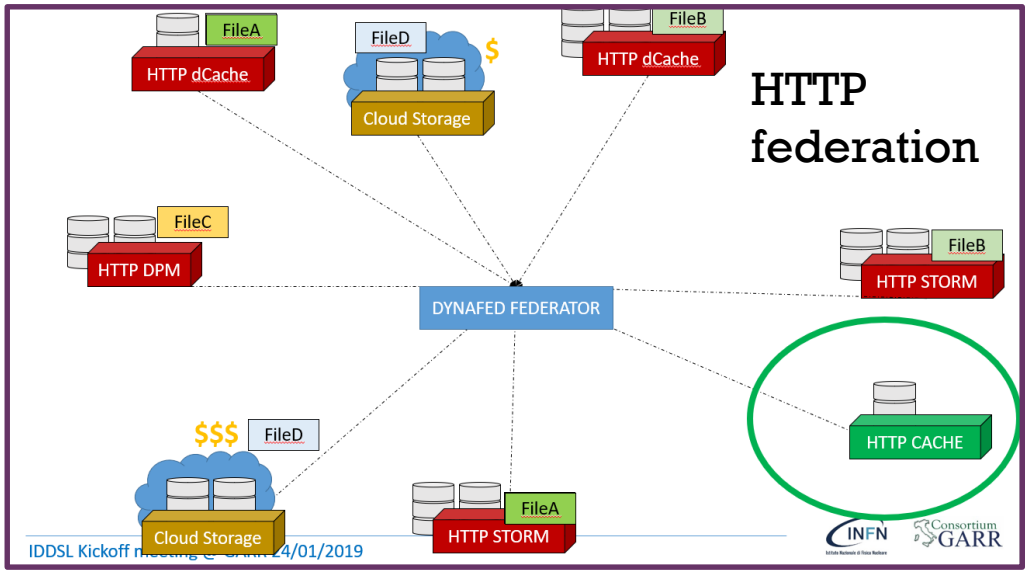
Gruppi di lavoro

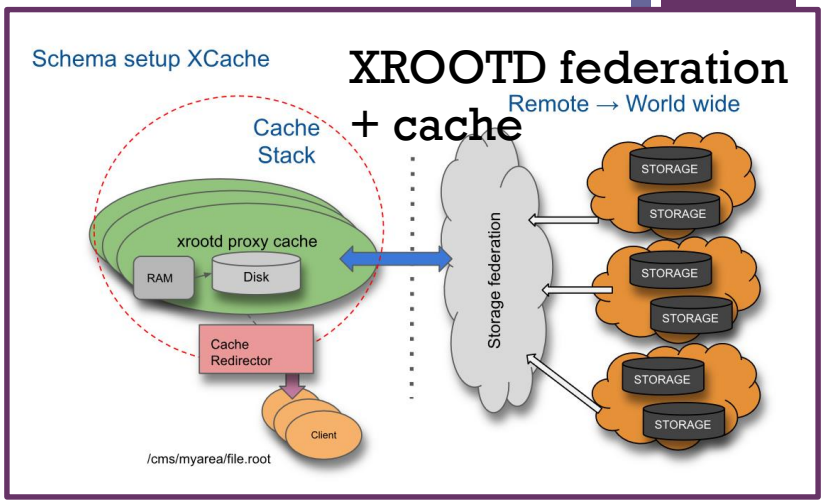
- WP1 – Management
  - Coordinamento, rapporti CSN5 e referee, organizzazione meeting
  - Progress report periodici
  - Procedure acquisti
- WP2 - Studio, definizione e implementazione dei link ad alta velocità
  - Scouting tecnologico delle soluzioni Data Centre Interconnect (DCI)
  - Identificazione dei requisiti degli esperimenti INFN
  - Integrazioni delle componenti HW e SW delle tecnologie DCI
  - Sperimentazione mista laboratorio e infrastruttura di rete geografica
  - Condivisione dello spettro in ambiente protetto;
  - Modelli di provisioning
  - Modelli di gestione e controllo
  - Sperimentazione su infrastruttura geografica su 3 siti

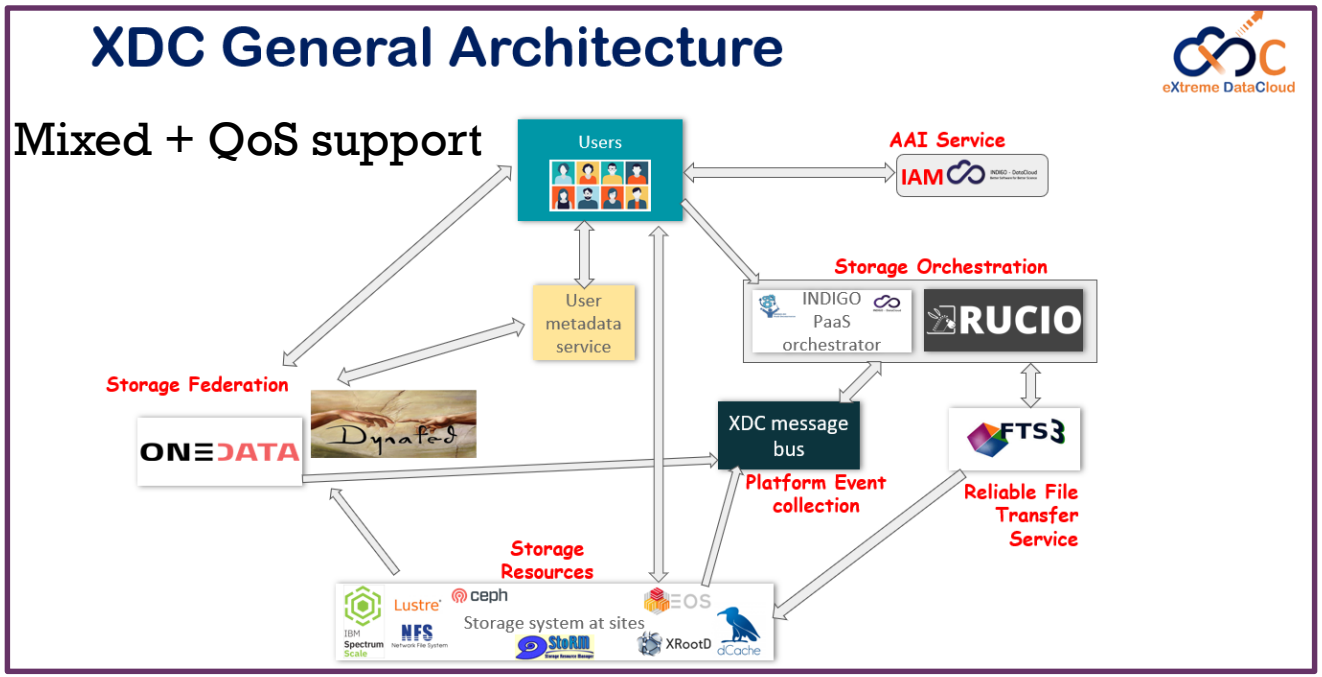
Gruppi di lavoro

- WP3 – Creazione del DataLake
  - Definizione dello stato dell'arte delle tecnologie esistenti
  - Implementazione del DataLake con tecnologie basate su protocollo HTTD/XROOTD con e senza sistemi di caching
  - Implementazione del DataLake con tecnologie differenti (eventuali)
- WP4 – Testing del DataLake
  - Definizione della testsuite del progetto basata sul software degli esperimenti rappresentati, almeno CMS, ATLAS e BELLEII
  - Esecuzione della testsuite sul DataLake sfruttando sia i siti interconnessi con tecnologie di tipo DCI che di tipo legacy
  - Interazione con sedi INFN o legate all'ente produttrici di dati (i.e. LNGS, CASCINA) che possano essere interessate a testare le soluzioni dell progetto
  - Valutazione delle performance ottenute

Lake Architecture(s)







Datalake infrastructure (WP3 next milestones)

- Storage endpoints deployed on all participating sites
- Covering different technologies: DPM, dCache?, SToRM
- **Data transfer and data management services**
  - Dedicated RUCIO instance with on Rucio Storage Element (RSE) per storage endpoint
    - QoS
    - Need to add tape endpoints RUCIO team about how to implement it
  - File Transfer Service (FTS) integration
- Dedicated IAM instance based on x509, token integration WIP
- Ability to move files with xrootd/http/gsiftp with RUCIO+FTS and CLI/WebUi
- **Monitoring**: FTS metrics, RUCIO events collected on an ES cluster, visualized through Kibana
- **Caching technologies**: first investigations on XCache