Scientific Computing Facilities at INFN-LNF and INAF-OAR

IL PROGETTO IBiSCo

Università di Napoli «Federico II»

18/04/2024

Elisabetta Vilucchi INFN-LNF
Stefano Gallozzi INAF-OAR

Il calcolo Scientifico nei Laboratori Nazionali di Frascati dell'INFN

- I Laboratori Nazionali di Frascati dell'INFN (LNF) ospitano uno dei 4 Tier2 italiani di ATLAS nel DC dedicato al calcolo scientifico
  - WLCG (Worldwide LHC Computing Grid)
  - INFN-DataCloud
- Progetti regionali, nazionali, europei: IBiSCo, PNRR, …
- CTA DC (uno di quattro) grazie alla partecipazione dei LNF al PON IBiSCo in collaborazione con INAF-OAR
- PNRR@LNF: realizzazione di un nuovo DC in grado di ospitare anche sistemi di calcolo ad alta densità/HPC
- Negli ultimi anni: investimento in personale dedicato al calcolo scientifico (IBiSCo, PNRR)

PON IBiSCo, Infrastruttura per BIg data e Scientific Computing

- Cherenkov Telescope Array Observatory (CTAO): le risorse IBiSCo sono confluite in una infrastruttura a disposizione dell'INAF per il calcolo dell'esperimento CTA.
- La collaborazione tra INFN-LNF e INAF-OAR per lo sviluppo del calcolo del progetto ASTRI-miniarray nel PON IBiSCo ha fatto sì che il Tier2 di Frascati divenisse la sede di uno dei 4 DC dell'esperimento CTA in collaborazione con l'INAF-OAR
  - 24 server ~ 17 kHS06
  - 1.5 PB disk storage
  - 10 server per servizi, upgrade infra rete T2 a 10/25/100Gbps

CIR IBiSCo, Infrastruttura per BIg data e Scientific Computing

- Personale: fondi per 4 AdR
- Un AdR 3y per attività nel servizio biblioteca+attività INFN Open Access (presentazione domani)
  - Irene Piergentile*: Open Access repositories for scientific literature and research data*
- Due AdR per attività tecnologica informatica
  - L. Gondgaze: Sistema di monitoring basato su Grafana (Telgraf, InfluxDB, Fifemon)
  - M. Behtouei: Ceph Storage Cluster

PNRR ICSC Spoke 0: il centro nazionale di ricerca in HPC, Big Data e Quantum Computing

- Frascati partecipa al Centro Nazionale ICSC: National Centre for HPC, Big Data and Quantum Computing con due Data Centre (DC)
- ICSC - Spoke 0 – Consolidamento infrastruttura del Tier2
  - 300k€ infrastruttura sala T2 (realizzazione isola corridoio freddo, impianto elettrico, di condizionamento, nuovi armadi)
  - HW per il Tier2 (esperimenti e INFN-DataCloud)
- ICSC – Spoke 0 – Realizzazione DC Space Economy
- 5 M€ per la realizzazione di un DC con raffreddamento ad aria e DLC (Direct Liquid Cooling) nel nuovo edificio recentemente acquisito.
- Personale: 4 TD 2y: 3 T3 e un C6

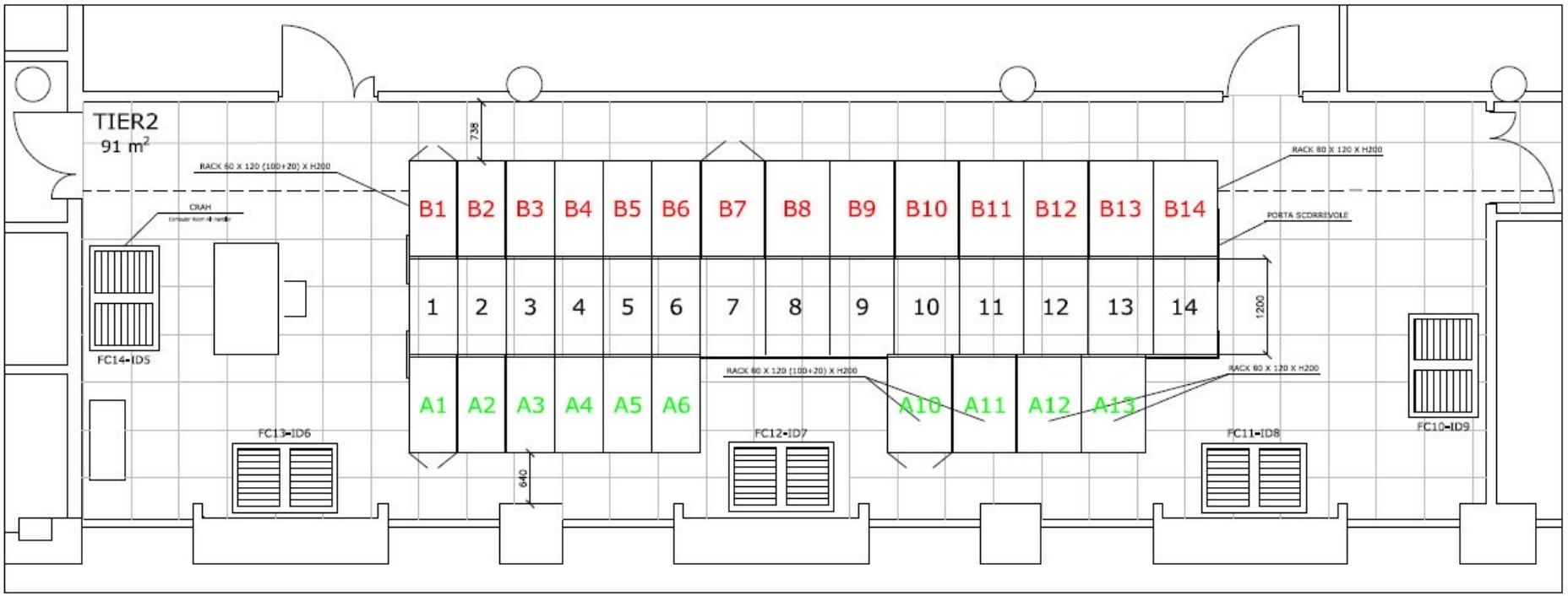
Il Tier2 di Frascati

- Ospita risorse degli esperimenti ATLAS e PADME come parte dell'infrastruttura distribuita WLCG (middleware Grid)
- Accesso opportunistico per altri esperimenti: Belle-II, LHCb…
- Ospitata una farm di CTA
  - Uno dei quattro off-site DC dell'esperimento
- In fase di sviluppo uno dei nodi dell'INFN-DataCloud con tecnologia OpenStack
- Ospita una serie di risorse/servizi per gruppi locali
- Partecipazione a progetti di calcolo regionali, nazionali, europei

Il DC del TIER2

- Spazio e potenza elettrica potenzialmente usabili a regime a fine lavori PNRR
  - ~100mq per un totale di 24 rack (mixed 60X120, 100X120)
  - 9 rack in più in grado supportare anche machine con maggior assorbimento
    - recupero ed ampliamento dei rack esistenti
    - ri-cablaggio di tutti i rack
  - aumento della potenza disponibile senza aggiunta di nuove macchine di raffreddamento grazie alla compartimentazione completa
  - Il carico totale installabile raggiungerà 160 kW, compatibile con la distribuzione elettrica ed il sistema di raffreddamento.
    - costituzione del gruppo frigo di emergenza che assicurerà la copertura del carico completo
  - Aumento del recupero termico in virtù del maggior carico IT installabile.

Il DC del Tier2

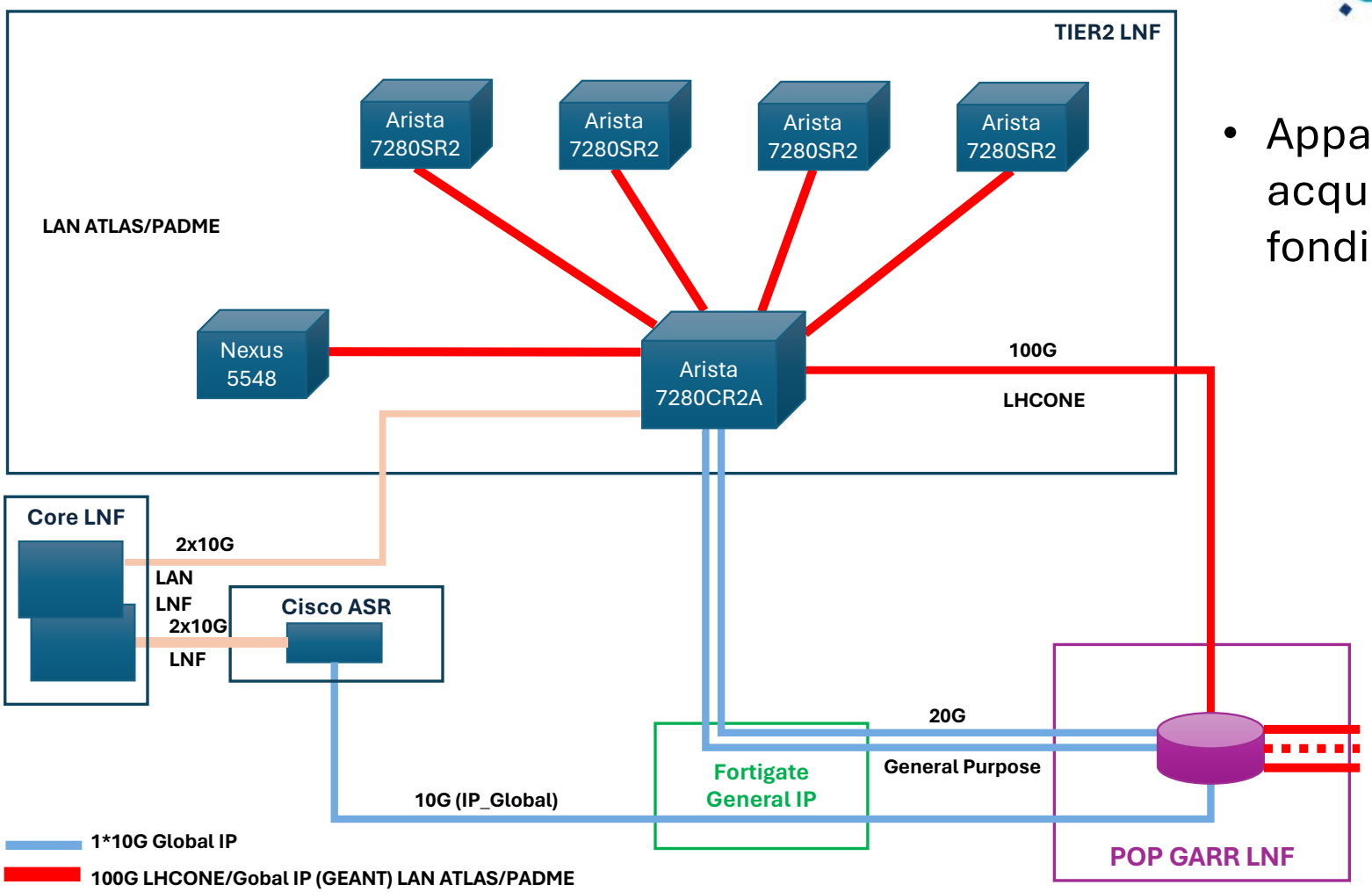


SALA TIER2 – 1:50

Risorse del Tier2

- Attuale carico IT del Tier2: 70 kW
  - 5 Computer Room Air Handler: Uniflair Leonardo (ora VERTIV), 4 CRAH accesi
  - UPS utilizzata da tutte sale calcolo ed. 14
- Disco pledged: ~ 7 PB netti
- CPU (anche extra pledge): ~ 125 kHS06 compresa la farm di CTA
- ~ 50 server
- LAN 10/25/100Gbps
  - Infrastruttura spine-leaf
  - 5 switch 10-25/100G
  - Diversi Cisco 2960 1/10G
- Tape Library nell'edificio 11
  - Nel DC dell'esp. KLOE, non integrata nel T2

Infrastruttura di rete



- Apparati Arista acquisiti con fondi IBISCO

CDZ edificio calcolo green: recupero energia termica

- Sistema di recupero dell'energia termica: in estate il PUE è 1.74, in inverno è 1.1. Media annuale PUE: 1.46.
- Il sistema copre il fabbisogno termico di 12.000 mq di uffici, laboratori e officine (33% della superficie totale riscaldata). Il riutilizzo dei reflui termici dei refrigeratori condensati ad acqua consente, durante la stagione fredda, di distribuire agli edifici acqua a 42 °C.
- Il carico termico stabile del DC assicura un calore economico, affidabile e rinnovabile. Dal 2016 viene prodotto circa 1 GWh termico l'anno, evitando il consumo di circa 100.000 metri cubi di gas naturale all'anno (equivalenti a circa 250-300 k€ con i prezzi attuali, 65 k€ nel 2019). In termini di parametri ambientali, questo significa un risparmio di 200 tonnellate di CO2 all'anno.
- Per il sistema in funzionamento ordinario il DC lavora con temperatura aria di mandata a 24 -26 C, conforme alla classe meno severa della norma che definisce gli standard climatici per le apparecchiature informatiche (ASHRAE TC9.9 classe A1), e non a 20 C come da progetto.

Nuovo DC ICSC: National Centre for HPC, Big Data and Quantum Computing

- PNRR: un nuovo DC in fase di realizzazione nei capannoni recentemente acquisiti dall'INFN vicino ai Laboratori.
- Riservati 400 mq per ospitare fino a 50 rack di calcolo.
- Cabina elettrica da 1.2MW di potenza IT.
- Sistema di raffreddamento misto: aria e DLC (Direct Liquid Cooling)
  - Esempio: ~400kW di sistemi di media densità (raffreddati ad aria) e 800kW in DLC.
- Prevista espansione degli spazi e della potenza IT installabile per ospitare in futuro altri progetti.

Data Centers for CTAO

- **PIC** in Barcelona, Spain
- **DESY** in Zeuthen, Germany
- Swiss National Supercomputing Centre (**CSCS**) in Lugano, Switzerland
- INAF/INFN in **Frascati**, Italy
- **SDMC in Zeuthen, Germany**

Status of the Italian CTAO DC

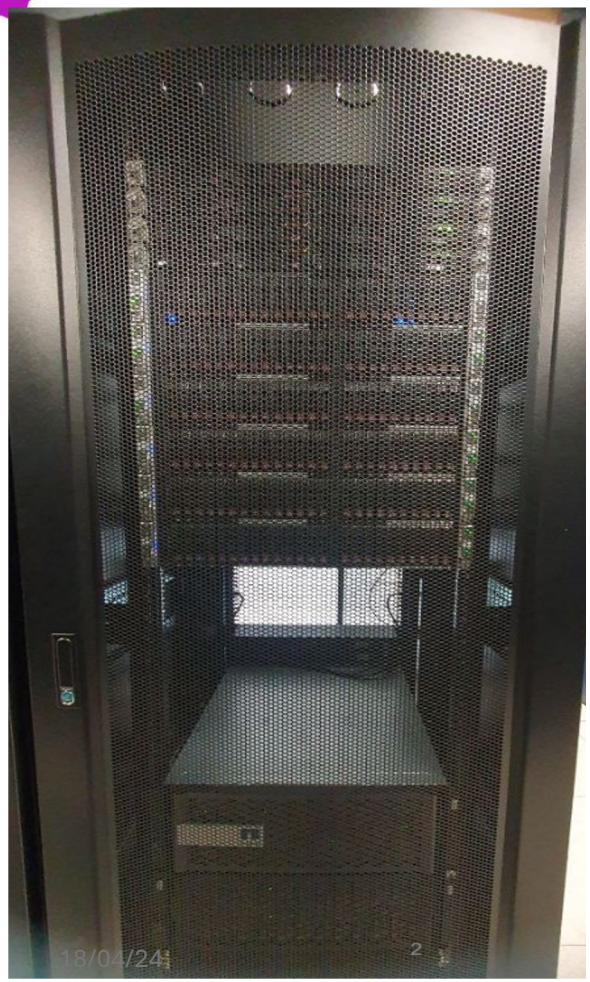
- Resources available @ 2023
  - 10 kHS06 (24 worker nodes with 48 cores@2.3GHz and 768 RAM DDR4 each) @ INFN LNF
  - 1 PBn disk NetApp E5760 (expandible on demand) @ INFN LNF
  - 3 PBn disk @ INAF OAR (available if needed)
  - 11 PBn tape library @ INFN LNF (up to be 20 PB without expansion modules)
    - Presently connected @ 1 Gbps. Connection to be upgraded with PNNR funds within 2023

delayed

	2021	2022	2023	2024	2025
CPU (KHS06)	9.57	0.11	0.29	0.48	0.62
DISK (TB)	1448	1472	1469	1469	784
Tape (TB)	3265	3542	3517	3506	2279

Total
11.07
6.642
16.469

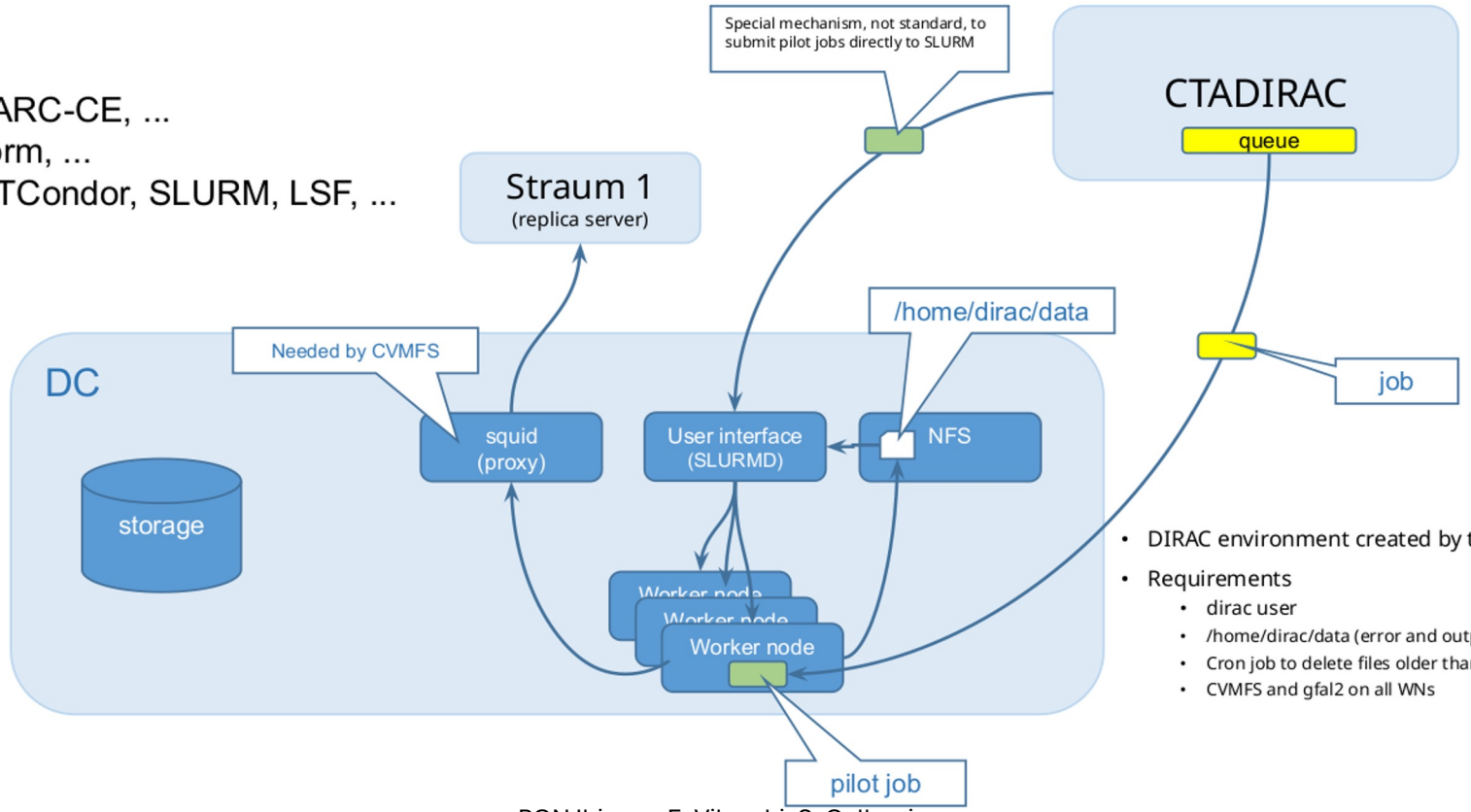
Hardware Status - LNF



- 24 physical hosts
  - AMD EPYC 7352 24-Core Processor (96 cores HT)
  - ~800GB RAM
  - ~15 TB (disk SSD)
  - Rocky Linux release 8.9 (Green Obsidian)
  - 4 hosts: proxmox cluster
    - services: slurmd, nfs, mariadb, prometheus + grafana
  - 20 hosts: worker nodes (slurm)
- Storage:
  - 4 hosts (64 cores HT, 270GB RAM)
  - NetApp E5760 (1PB)
- 5 networks
  - 1 public (outgoing connectivity for all nodes) 10GBit
  - 4 private (storage int, storage ext, management, ipmi) 10GBit

Current
High level CTAO-ITA DC

CE: HTCondor, ARC-CE, ...
SE: dCache, Storm, ...
batch system: HTCondor, SLURM, LSF, ...

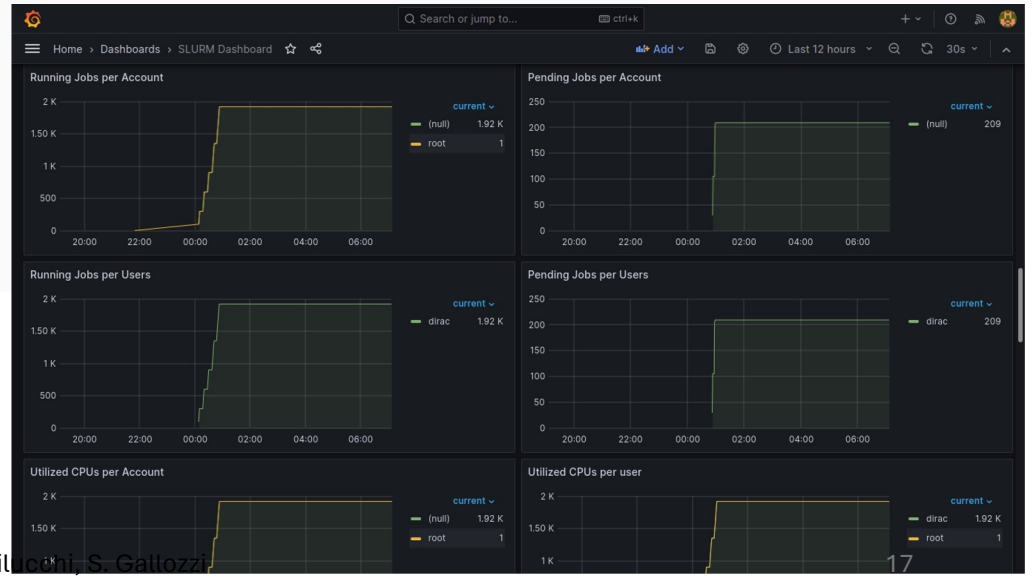


- DIRAC environment created by the pilot job
- Requirements
  - dirac user
  - /home/dirac/data (error and output files) shared with NFS
  - Cron job to delete files older than 1 week
  - CVMFS and gfal2 on all WNs

Current Status of CTA-ita node

1) tests with DIRAC WMS successful

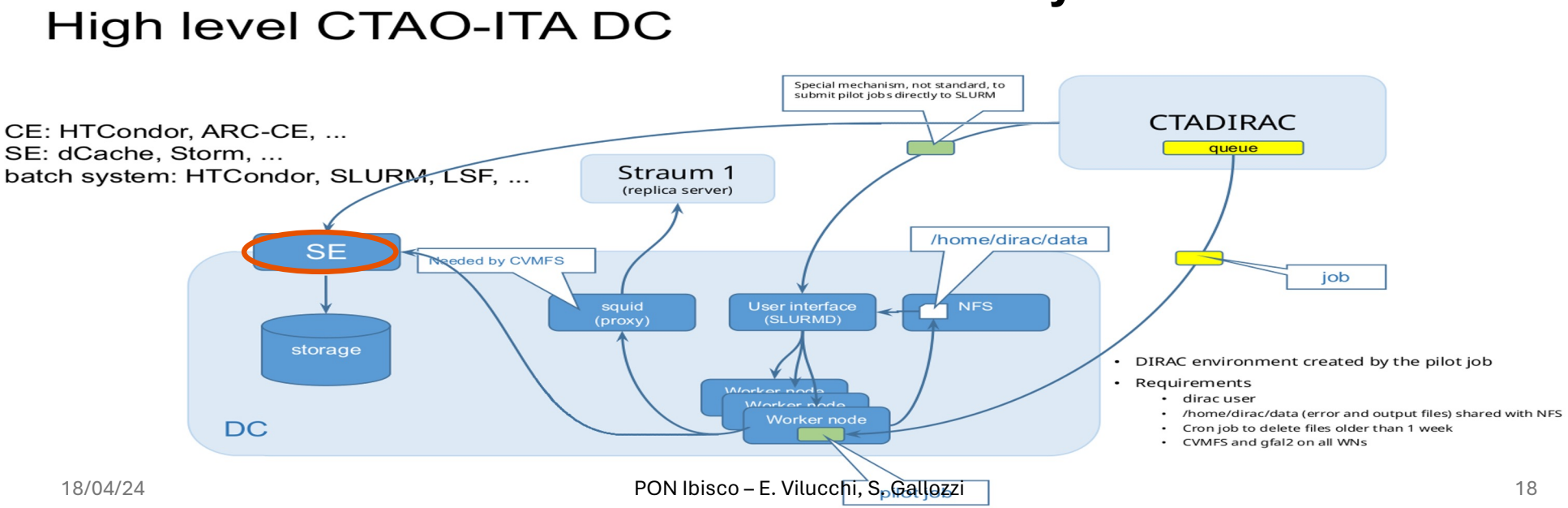
- firewalls are open P2P to 4 CTA DC (FF)
- 20/24 working nodes ready with a Slurm queue available to DIRAC client to run 1920 pilots and jobs for prod6 (FF and LZ)
- 4/24 nodes used to manage storage catalogs and other VM (FF and SG)
- 1PB Storage directly accessible by WNs through Lustre FS



Current Status of CTA-ita node

2) currently working on build-up StorageElement

- a new Storage Element (~1PB) is going to be installed using dCache (L.Zangrando and D.Miceli)
  - need DCs specifications to enable Storage accessible by DIRAC and BDMS?



BACK-UP

Hardware Status - OAR

- Prototipo per calcolo Astri-Miniarray
- Servizi centrali di virtualizzazione HPC, Database,
- Sistemi di storage,
- LDAP e gateway offsite per il progetto astri-miniarray.
- Sistemi e potenziamenti hardware integrati nell'environment del progetto ASTRI.
- Nr. 12 server di storage DELL-EMC Server PowerEdge R740XD
  - Chassis in grado di alloggiare 16x3.5" HDD

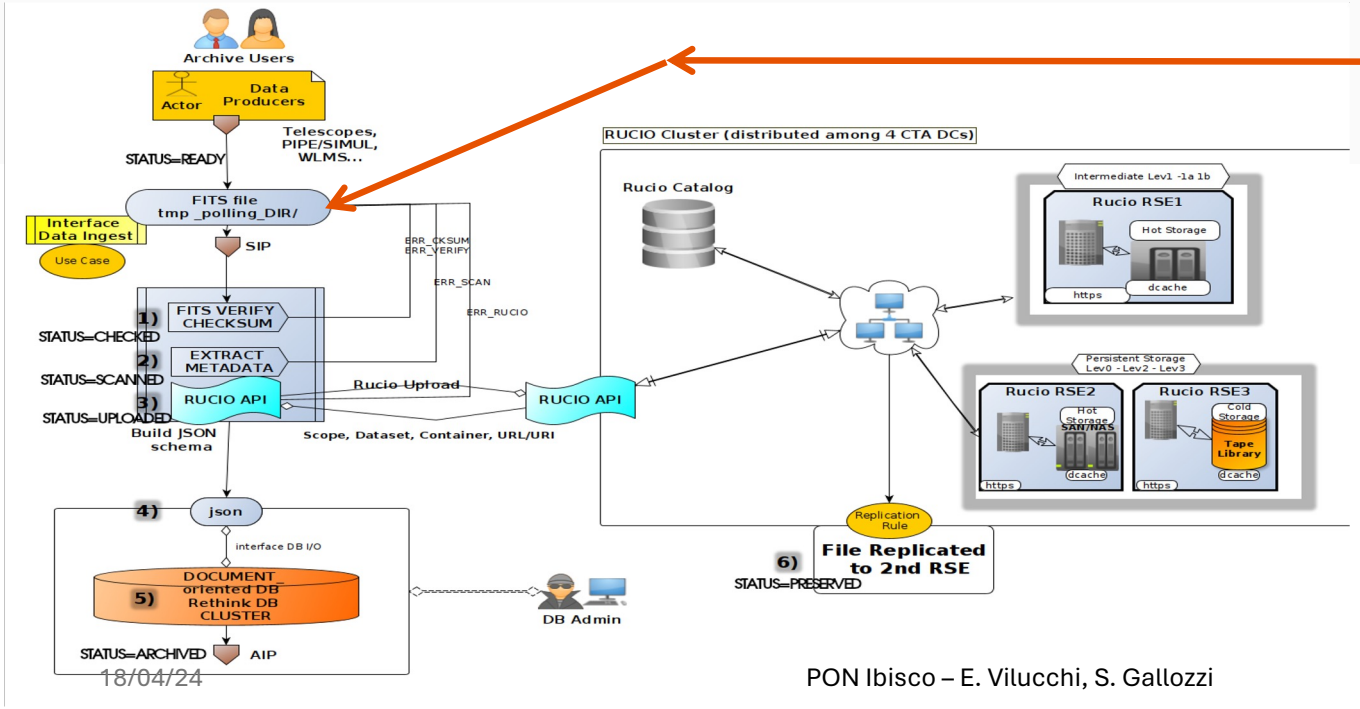
Tape Library @ LNF

- A tape library (one of tape libraries at LNF) is available in the KLOE DC in bld 11, but it is not integrated in the Tier2 storage system in Grid (even if it's used by PADME exp.) .
  - Software: IBM Specrtum Protect (Ex Tivoli Storage Manager)
- 6 Tape-drives, 1600 cartridges with 500 free slots (1100 10 TB cartridges each not compressed)
  - Standard IBM tape media: 3592 Type D
  - 3 drives always used
  - Possible to add drivers up to 16 with no expansion
- 2 modules (up to 16) 1600 cartridges in total
  - +500 tapes 10TB each with the current drives,
  - Available space up to 5PB with no additional drivers
  - it's possible to add 14 modules.

Next Steps for CTA-ita node

3) use the DC VMs to test and enable RUCIO_vs_DIRAC

- K8s cluster to enable RUCIO instance
  - BDMS interface to watch and monitor SE



The /tmp_path should be DIRAC & RUCIO accessible within a CTA-DataCenter Object Storage Element

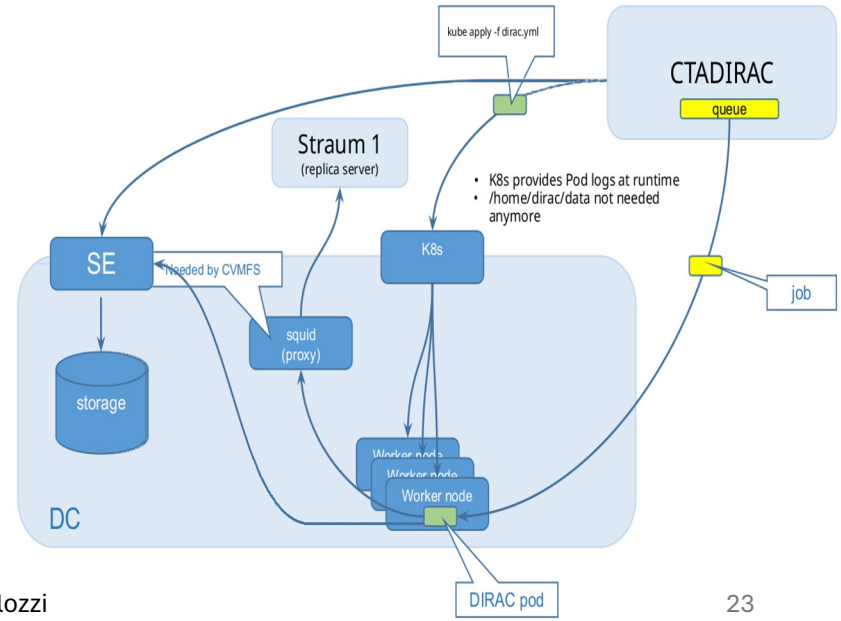
Next Steps for CTA-ita node

4) implement a CE with K8s technology

- K8s cluster to substitute SLURM pilots queue and enable DIRAC job pod on worker nodes

- K8s as substitute of CE + batch system
- LHCb already deploy DIRAC in K8s (https://gitlab.cern.ch/lhcb-dirac/diracchart)
- new K8s operator may be developed for implementing new CTAO custom resources (api)
  - e.g. "create at least 10 DIRAC instances on WNs and autoscale if needed"
  - there is no need to submit pilot jobs periodically
  - to investigate if DIRAC already provides this feature
- almost all our DC services can be deployed in K8s

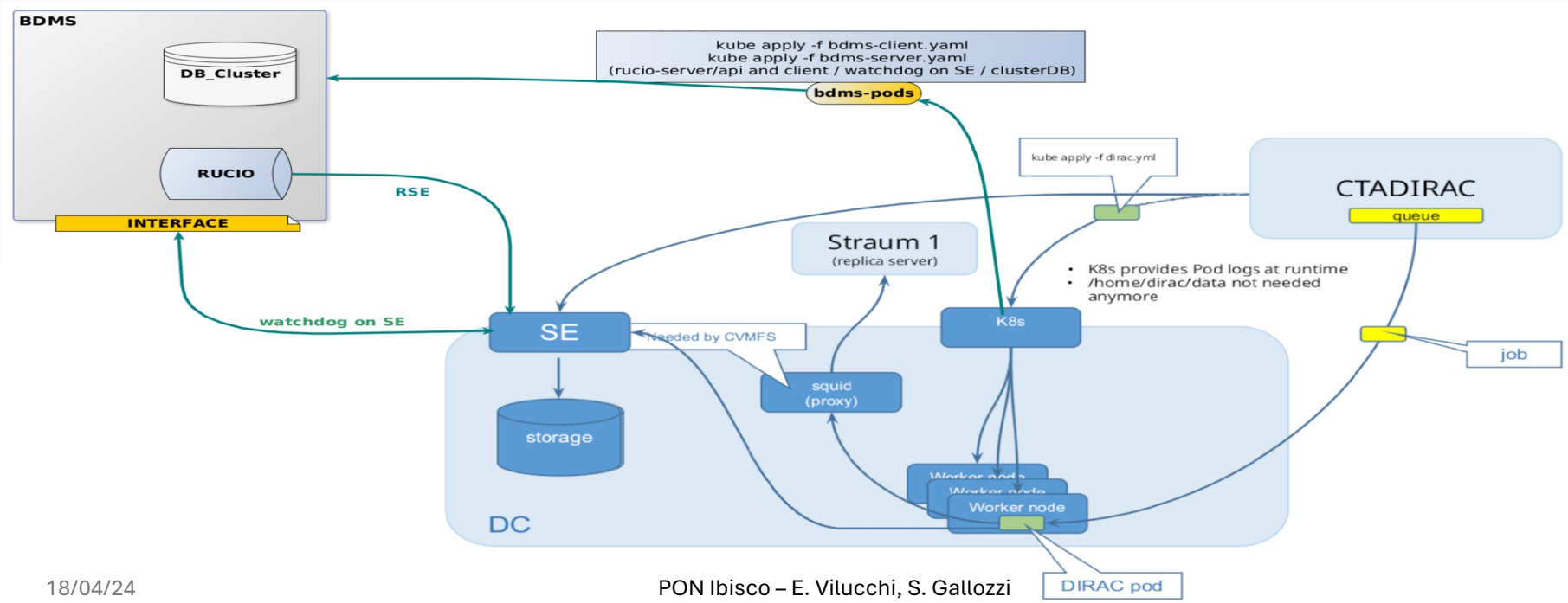
Kubernetes based architecture



Next Steps for CTA-ita node

5) use K8s technology to implement BDMS and DIRAC

- K8s cluster will serve the BDMS interfaces and services (RUCIO and DBs) as well as DIRAC WMS



Sistemi in fase di acquisizione per il nuovo DC HPC

- Rete:
  - Core switch che sarà connesso 2X400G al core switch del Tier2
- HPC:
  - 8 Server Lenovo con 2 AMD EPYC 9654 2.4 GHz (192 core fisici), 1.5 TB RAM, 4 GPU Nvidia H100 94GB RAM, 1porta NDR 400G IB, 4 porte 25Gb Eth
  - 6 Storage server DELL PowerEdge R760xd2 con 2 processori Intel Xeon Gold di quarta generazione (Sapphire Rapids), modello 6428N, frequenza 1.8GHz, 32Core/64Thread, 60M Cache, DDR5-4800, 185W TDP, 1024GB RAM per un totale di 384TB “raw” per nodo su dischi HDD, IB e un totale di 25.6TB “raw” per nodo su dischi SSD con connessione NVMe