国家高能物理科学数据中心
National HEP Science Data Center

高能所計算中心
IHEP Computing Center

The Beijing LHCb T1 Status and JUNO Distributed Computing System

Jingyan SHI, Xiaowei Jiang &Xiaomei Zhang

IHEP Computing Center

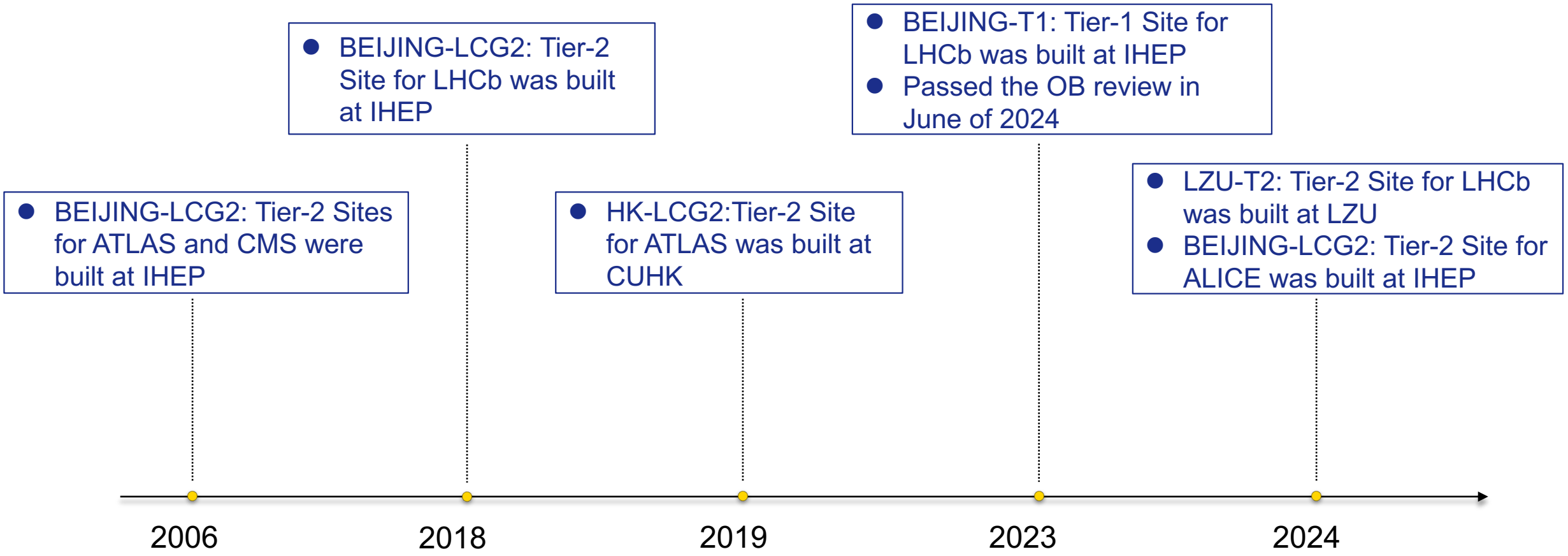
2024-12-04

Outline

1 The Beijing LHCb T1 Status

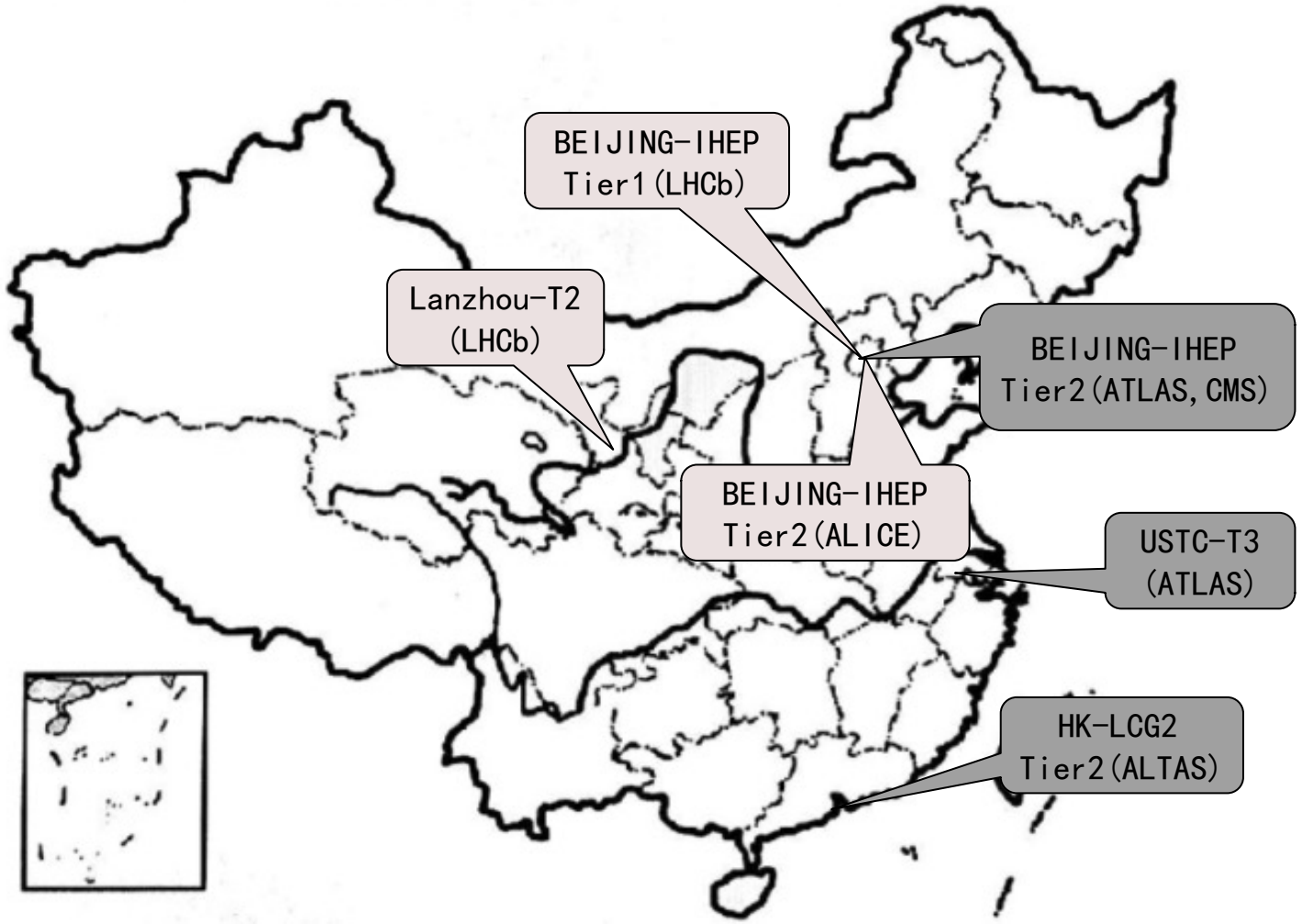
2 JUNO Distributed Computing System

WLCG History in China



WLCG Sites in China Mainland

- Tier-2 sites
  - BEIJING-IHEP (ATLAS, CMS, LHCb)
  - HK-CUHK(ATLAS)
- New sites in recent two years
  - Tier-1: BEIJING-IHEP (LHCb)
  - Tier-2: LZU-T2 (LHCb)
  - Tier-2: BEIJING-IHEP (ALICE)

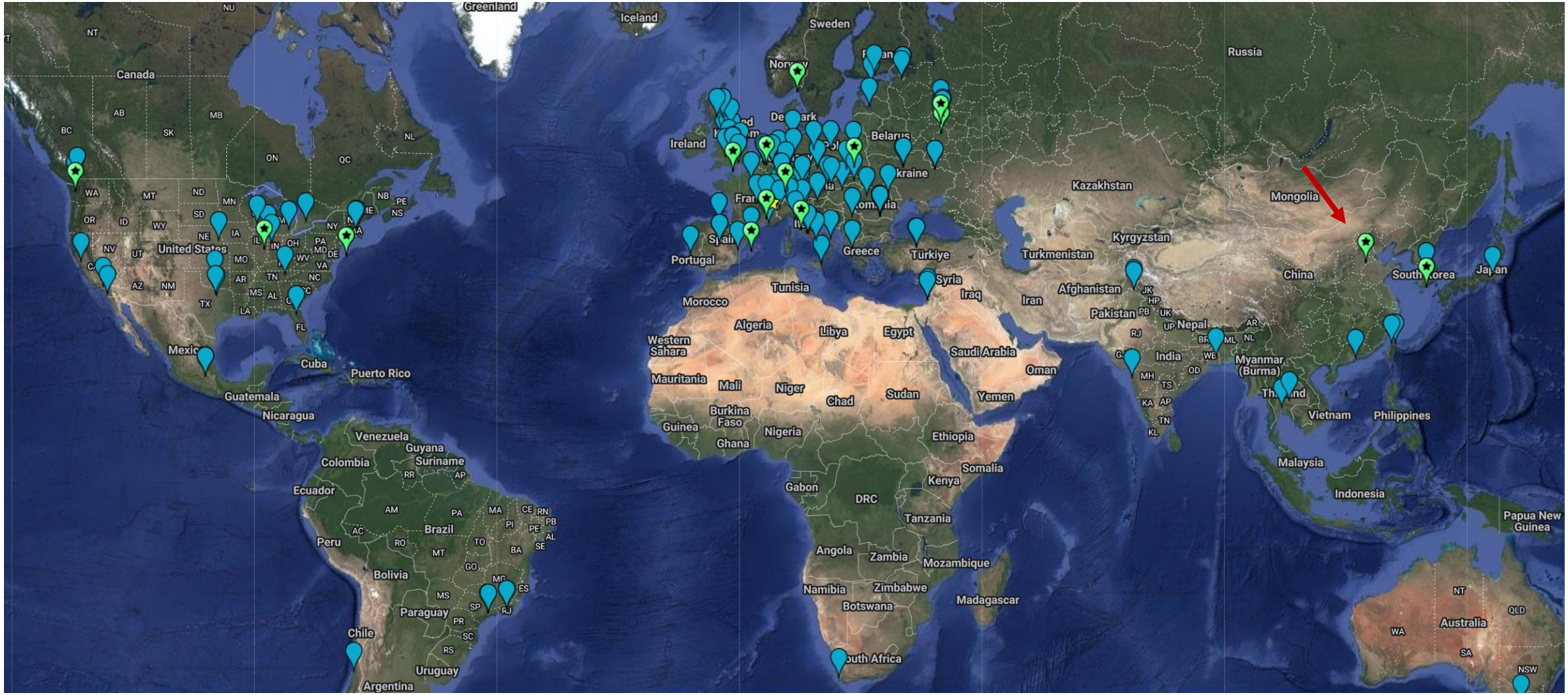


Beijing LHCb Tier-1 Site Construction

- In Sep. 2022, formulate the construction plan and estimated funding
- In Oct. 2022, LHCb Chinese Group decides to build a Tier-1 site
- In Dec. 2022, WLCG officially confirmed the establishment of the LHCb Tier1 Beijing site
- In Dec. 2022, begin procurement of equipment.
- In May. 2023, all equipment procurement was ready.
- In Jun. 2023, finish the site service setup and deployment
- In Oct. 2023, establish of the LHCOPN network link
- In Nov. 2023, complete basic functional test
- In Jun. 2024, complete the data challenge test
- In Jun. 2024, The“BEIJING-T1”site was passing the WLCG OB review and officially joining the WLCG

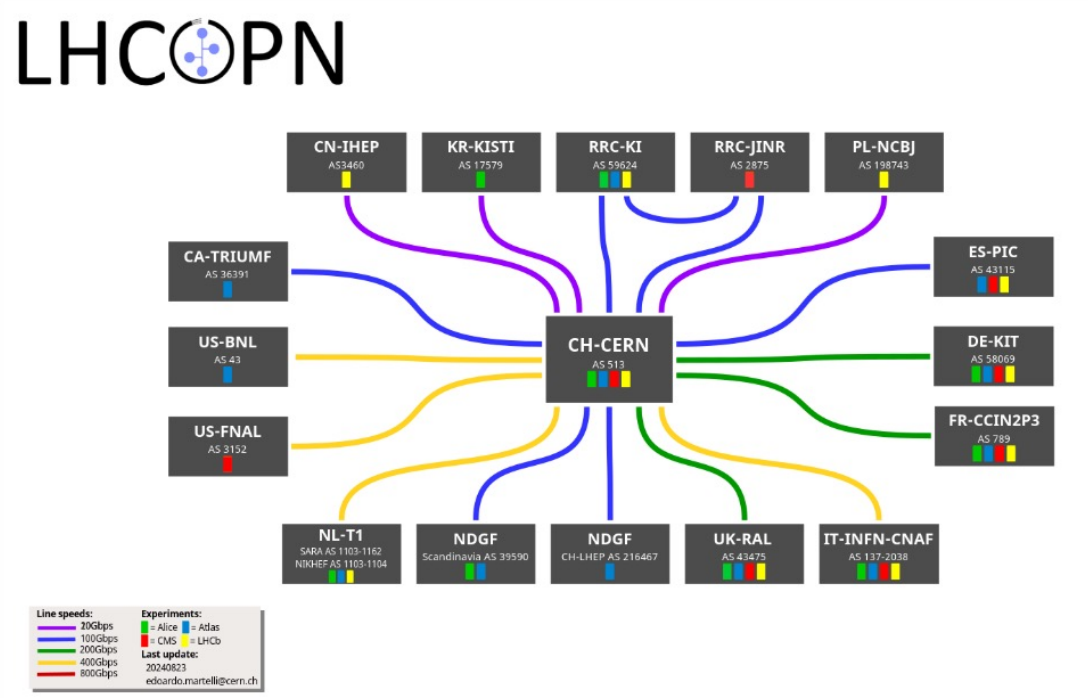
- All the LHCb coordinators and WLCG coordinator were coming to visit the Tier-1 site and keep caring about the progress during the construction

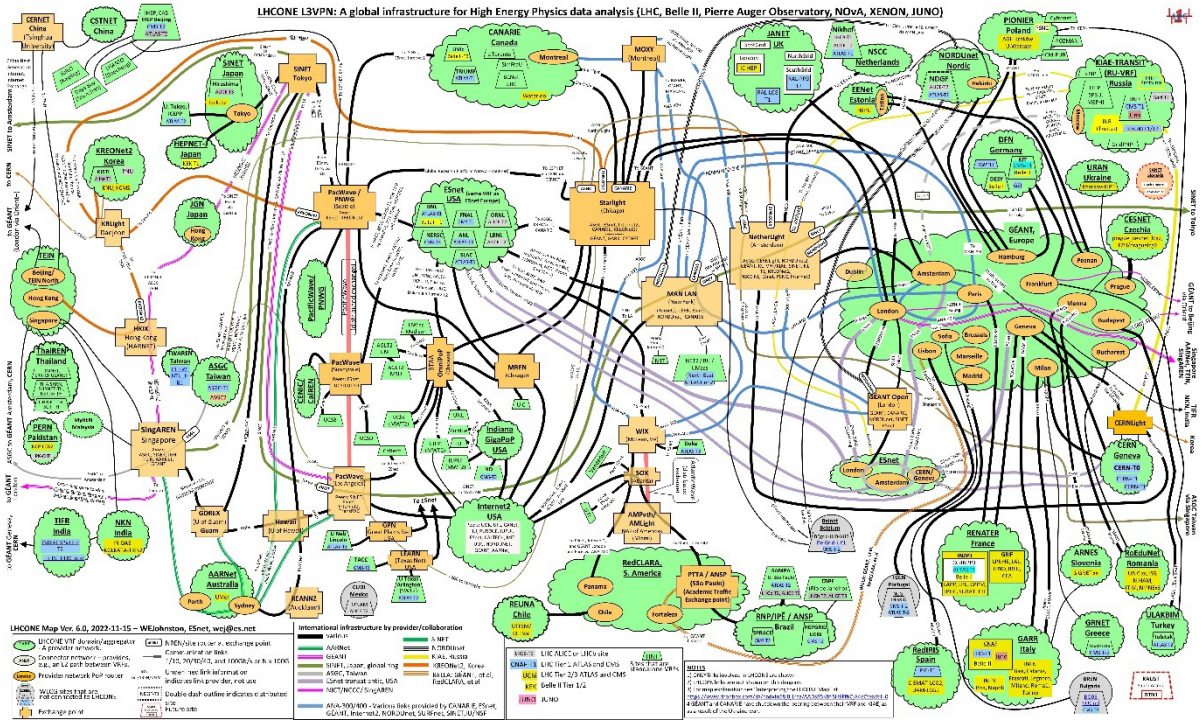
BEIJING-T1 site is in production now



LHC Network Organization

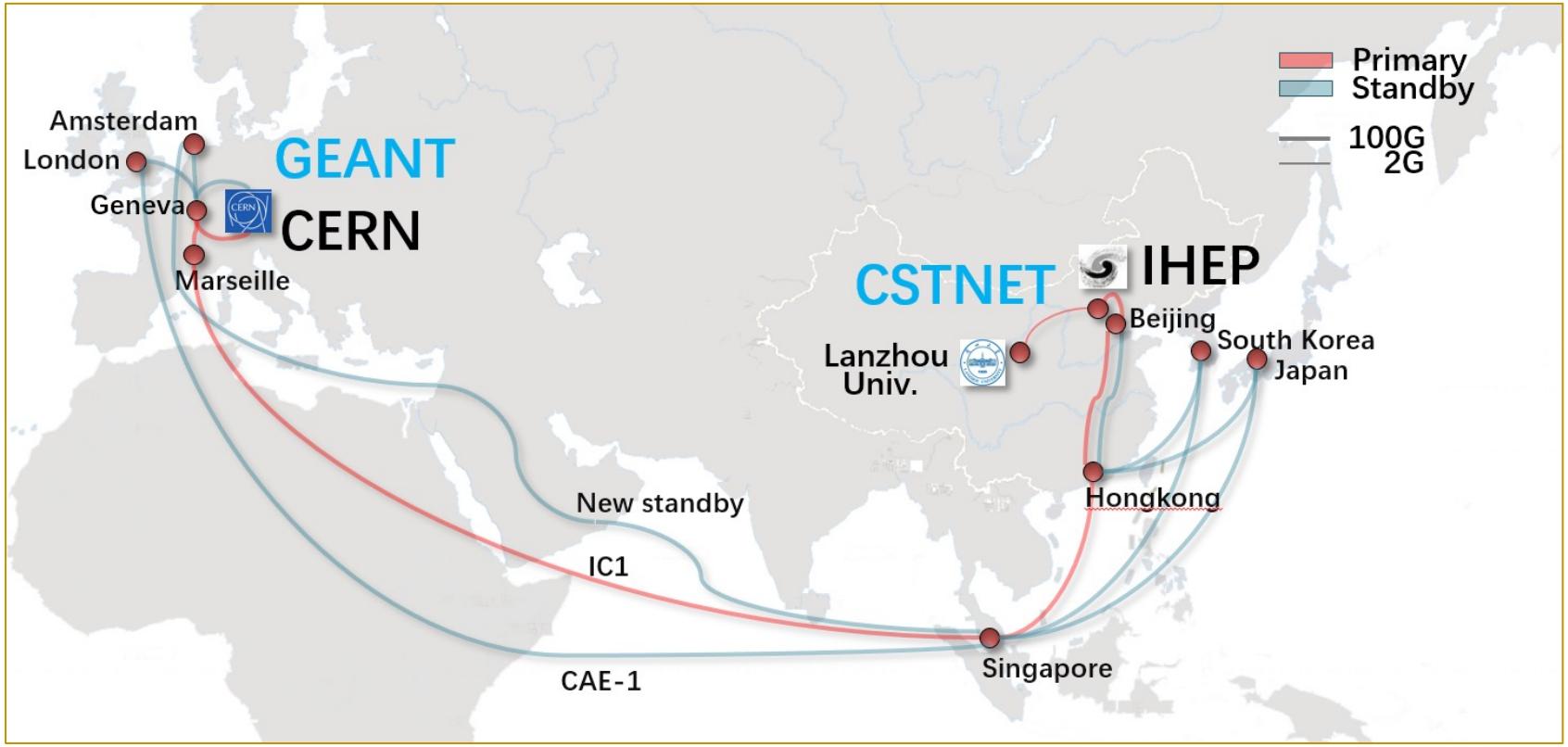
- LHC is using LHCOPN and LHCONE to support the data transfer needs of the LHC community and serve the networking requirements of the distributed computing models
  - LHCOPN (LHC Optical Private Network): linking Tier 0 and the Tier 1s
  - LHCONE (LHC Open Network Environment: linking the Tier 2 community





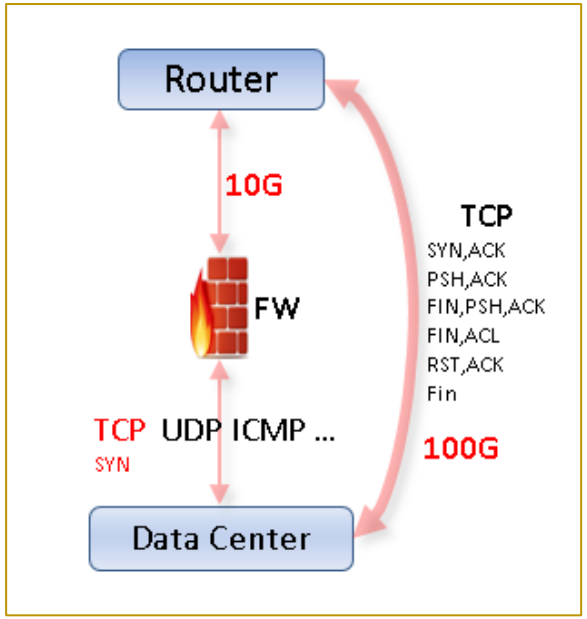
The International Network for LHC at IHEP

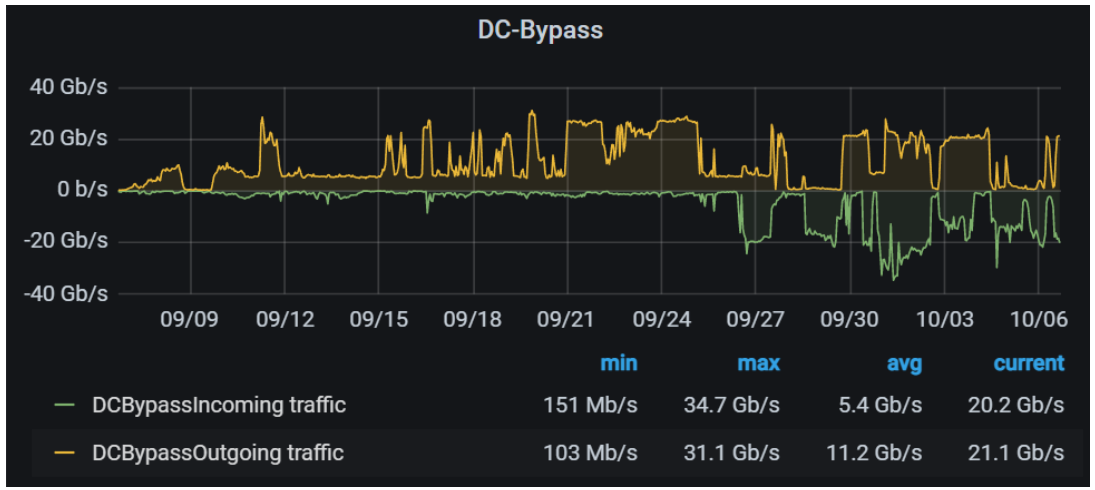
- A new international 100Gbps network link: established by CSTNET, cooperated with GEANT by the end of 2023
  - 20Gbps bandwidth dedicated (3 links redundancy)
  - 100Gbps bandwidth shared (WLCG is the largest user currently)

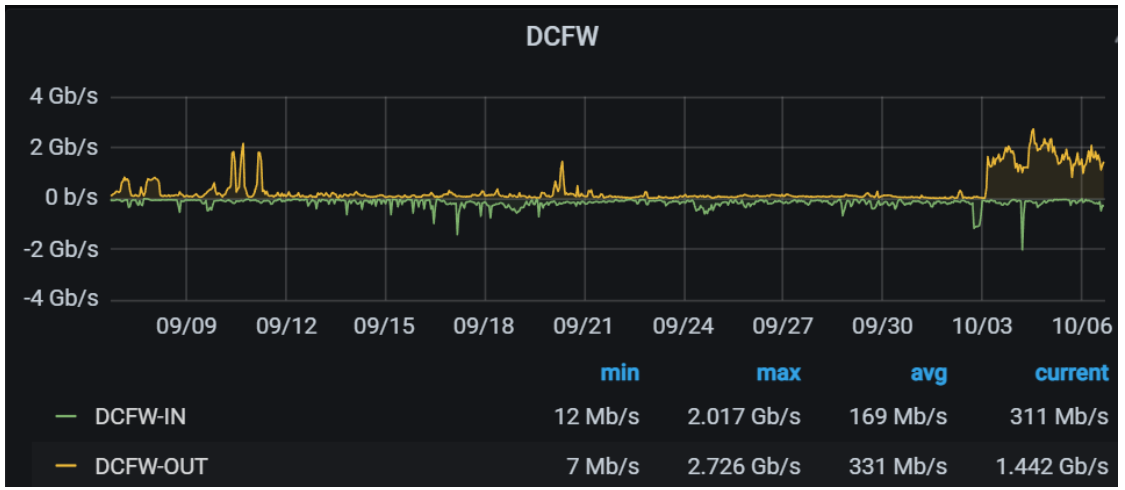


Scientific Data Traffic Bypass Policy

- Bypass policy
  - Bypass TCP communication except TCP SYN, Support Public-IPv4 and IPv6
  - The Firewall assumes normal security protection for all hosts
  - The Firewall traffic includes TCP-SYN, UDP, IPv4-NAT and so on
- Current Status
  - The MAX bandwidth > 35Gbps and Firewall bandwidth <3Gbps
  - Running well

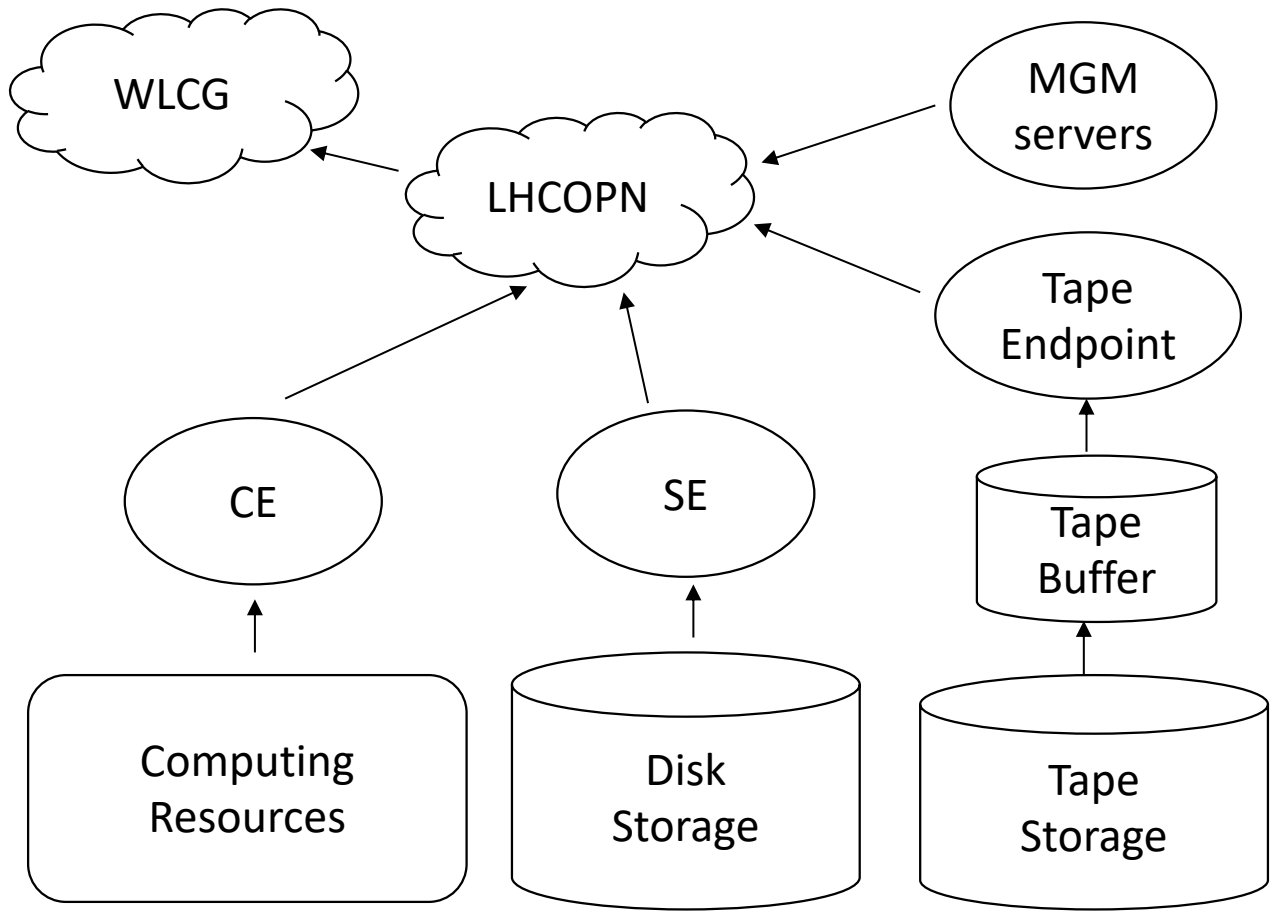






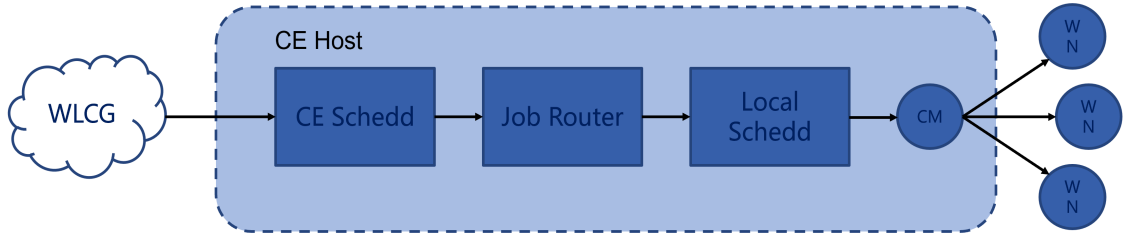
The First WLCG Tier1 Site

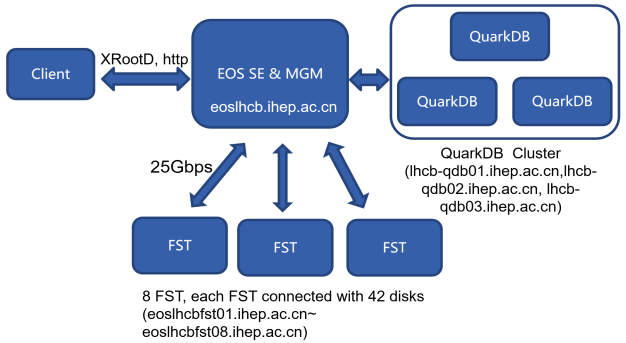
- Building LHCb Tier1 site starts since October 2022
  - Cooperated by Chinese LHCb Collaboration, IHEPCC and CNIC
- First Step
  - Computing: 3216 CPU cores (67 kHS23), 40 worker nodes (Intel & AMD)
  - Disk storage: 3.2 PB, 4 storage arrays
  - Tape storage: 3PB, 4 drivers (IBM) and 170 tapes, LTO-9
- This Week
  - Disk storage updated to ~13PB
  - Tape storage updated to ~10PB

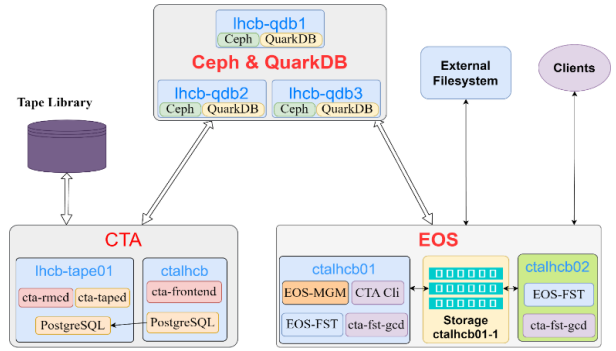


Middleware Deployment for Tier-1

- All the middleware deployments have been done
  - Computing System: HTCondorCE and HTCondor
  - Disk Storage System: EOS
  - Tape Storage System: EOS-CTA
  - Other services
    - Certification service: Argus
    - Grid information service: BDII
    - Accounting service: APEL



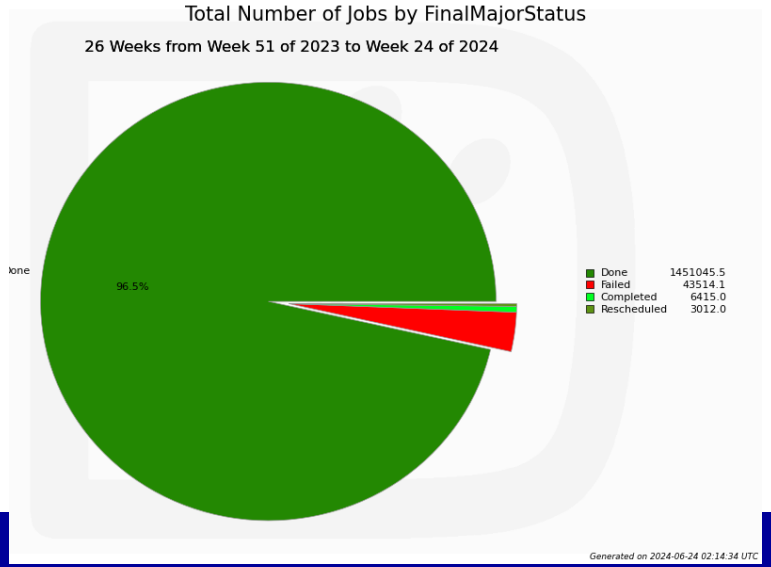


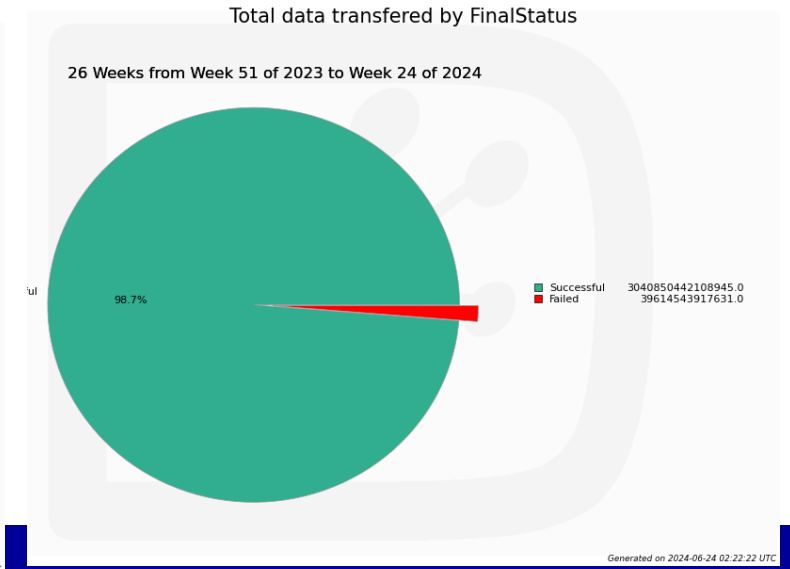


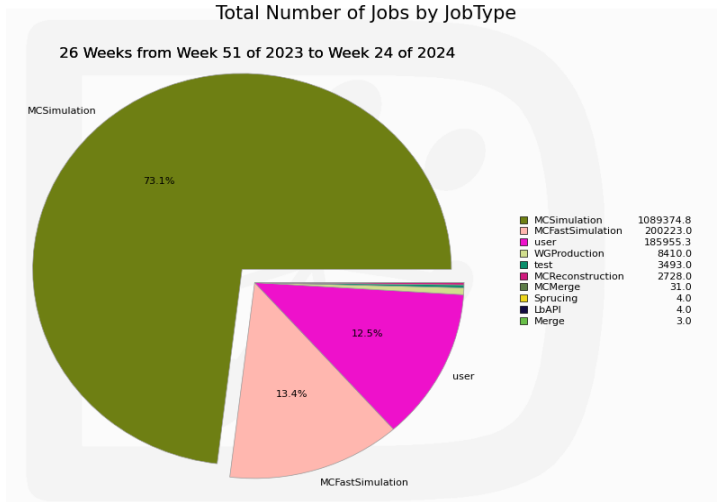
- All the middleware are deployed following WLCG requirements
- OS has been upgraded to Alma Linux 9 in Aug. 2024

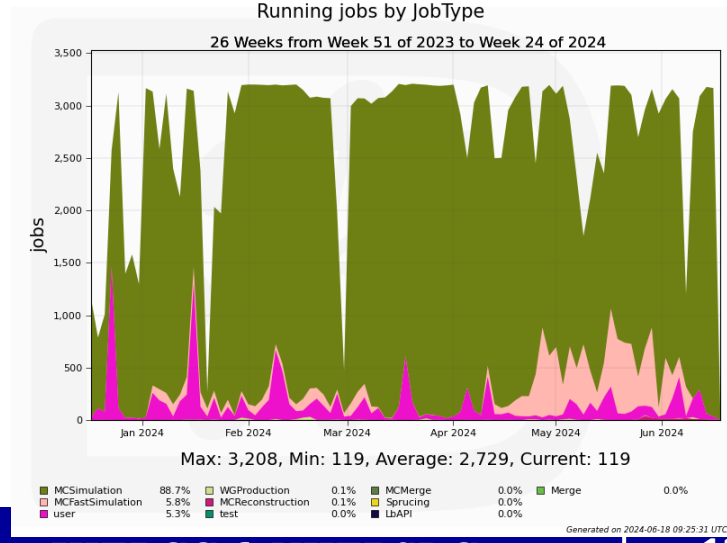
Statistics on Beijing LHCb T1-- Computing Service

- Computing system was integrated to LHCb Distributed Computing System in Dec. 2023
- 1.49 million jobs and 12.18 million CPU core-hours in the first half year
- 96.5% job succeed and 98.7% transfer succeed



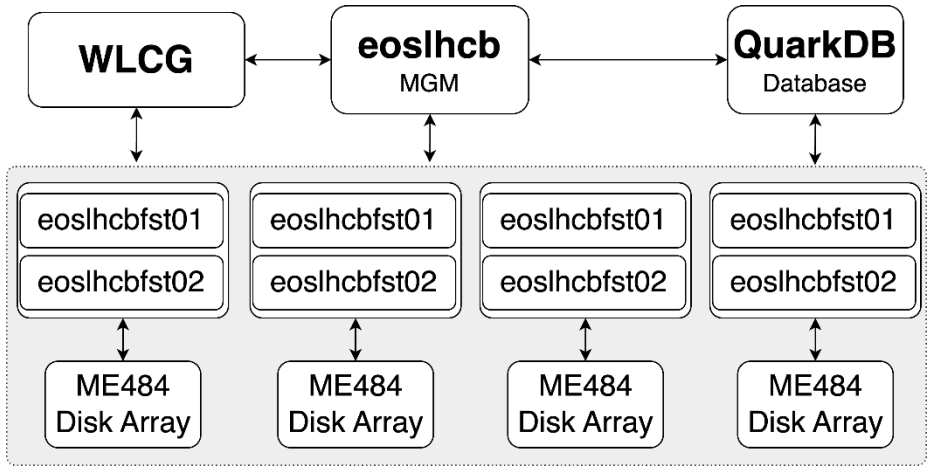


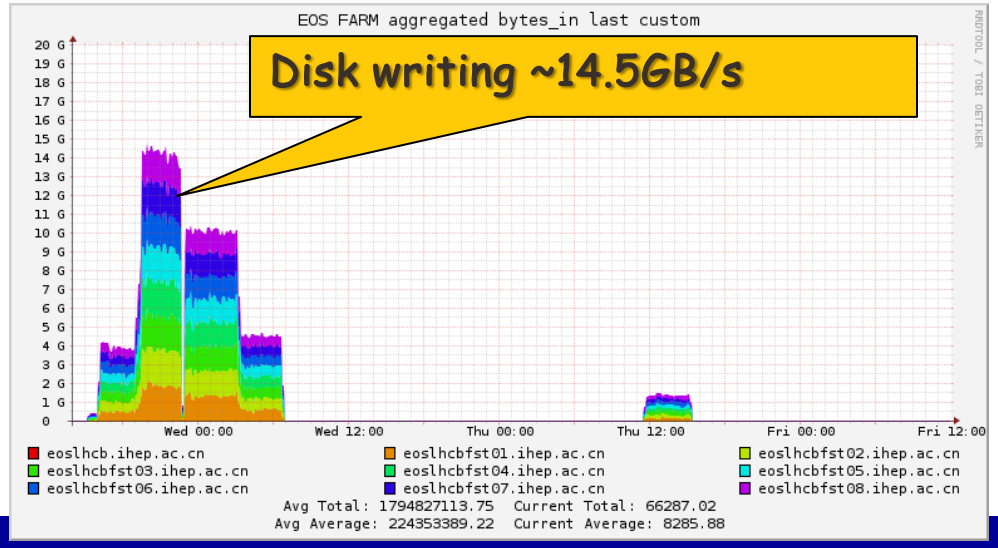


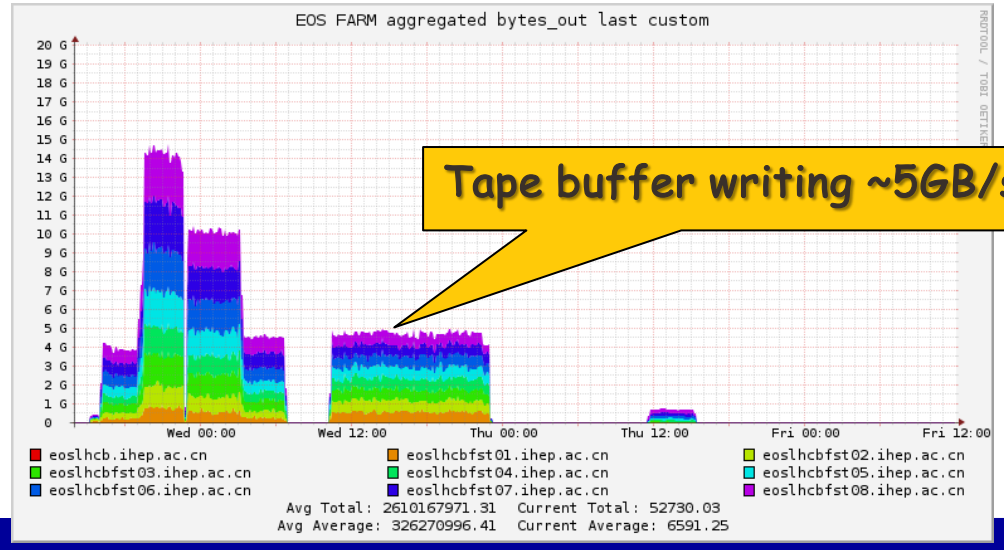


Disk Storage -- 13PB

- Disk storage structure
  - 4 management servers: MGM and DB
  - 24 storage nodes with 21 disk arrays
- Performance Test with a 200TB dataset
  - Write into SE: speed by ~14.5GB/s
  - Read from SE and write into tape buffer: speed by ~5GB/s

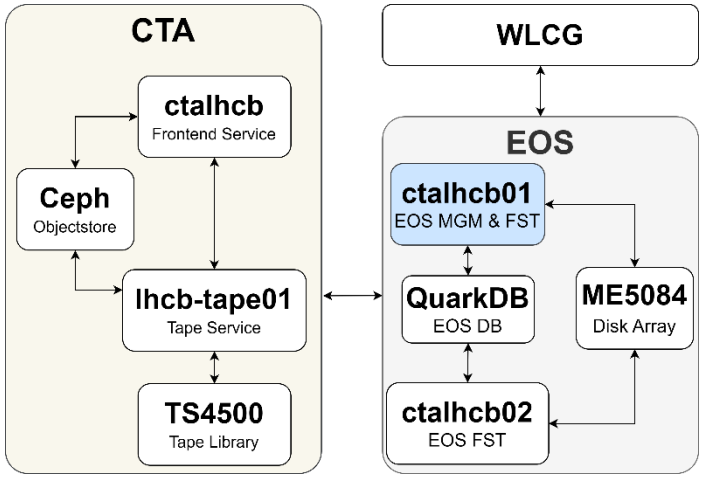


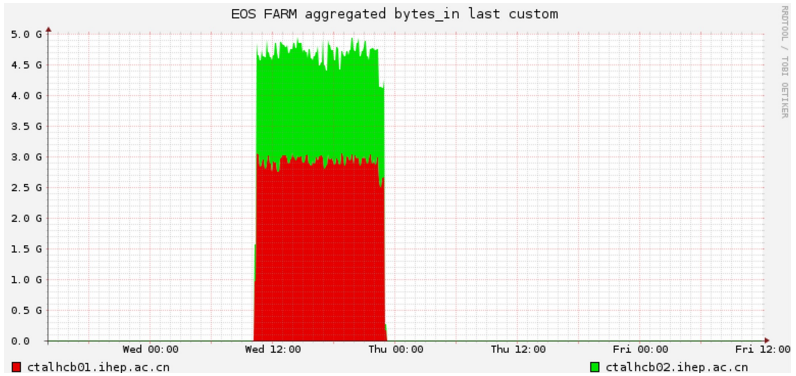


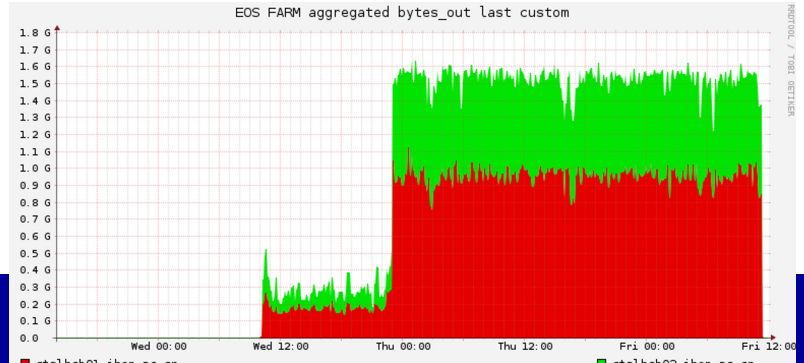


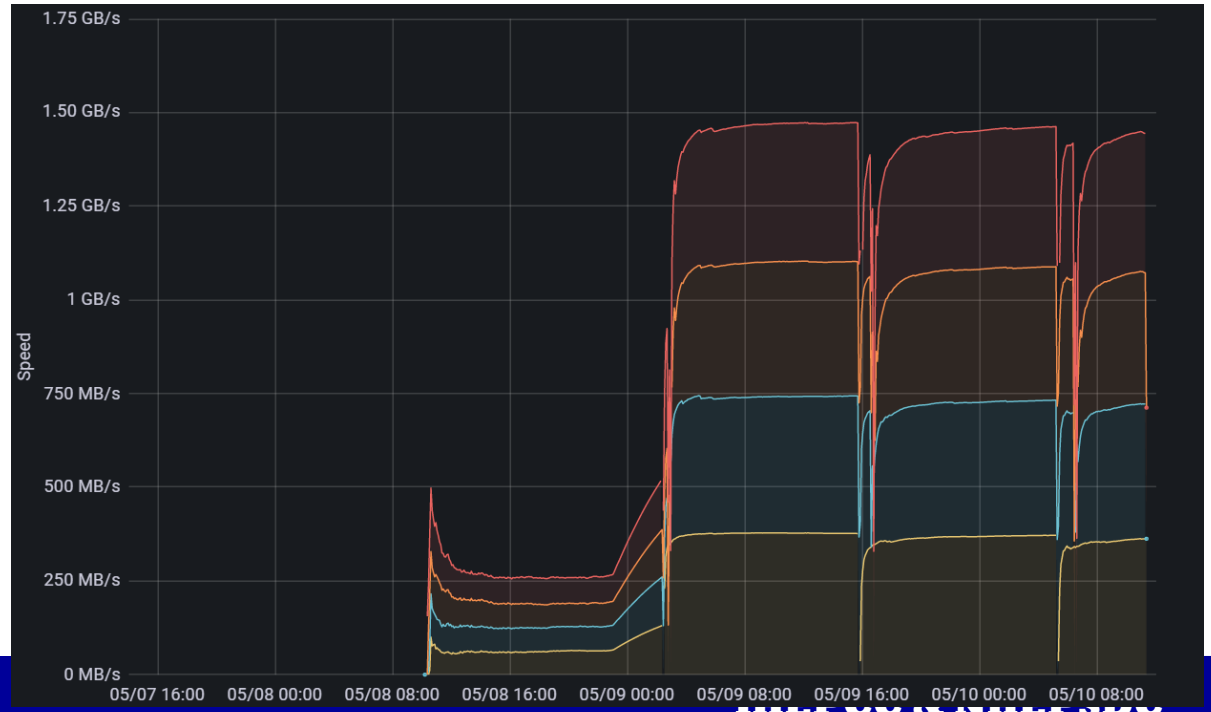
Tape Storage

- Tape storage structure
  - 2 buffer servers with a disk array (672TB), can accommodate 3 days of data
  - Tape library with 4 drivers for the moment
- Performance test with a 200TB dataset
  - Tape writing speed is ~1.2GB/s



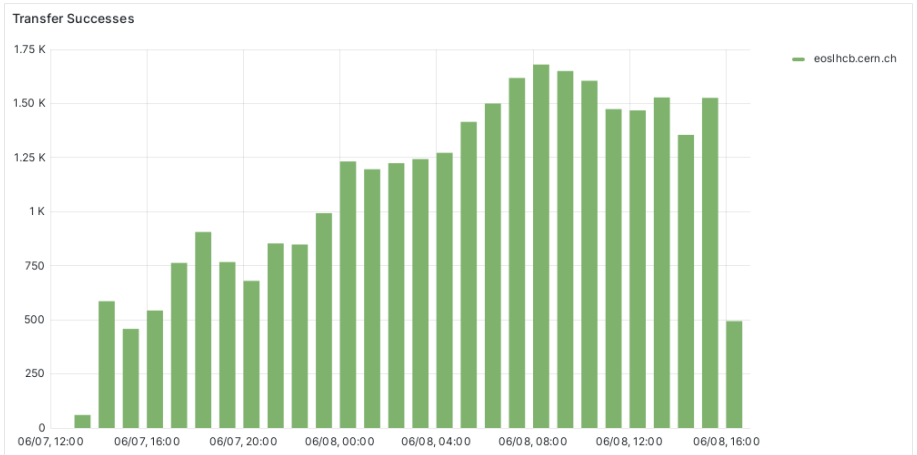


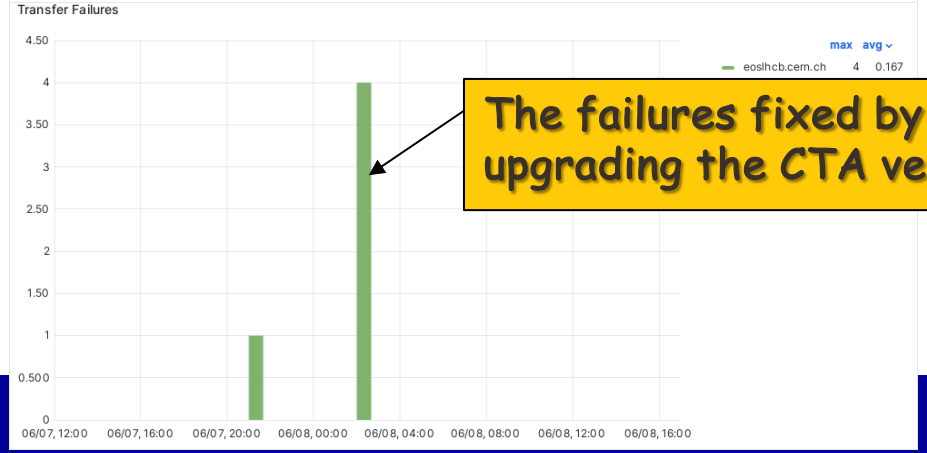




Data Challenge Before Site Review

- LHCb data challenges carried out in Jun. 2024
- Average transfer throughput is about 1.75GB/s, with almost no failures





The failures fixed by upgrading the CTA version

List | Find | Login | Help

07-Jun-2024 15:04 from Alexander Rogovskiy : Test transfers for Beijing->CERN link

07-Jun-2024 15:30 from Alexander Rogovskiy : Test transfers for Beijing->CERN link

10-Jun-2024 10:06 from Alexander Rogovskiy : Test transfers for Beijing->CERN link

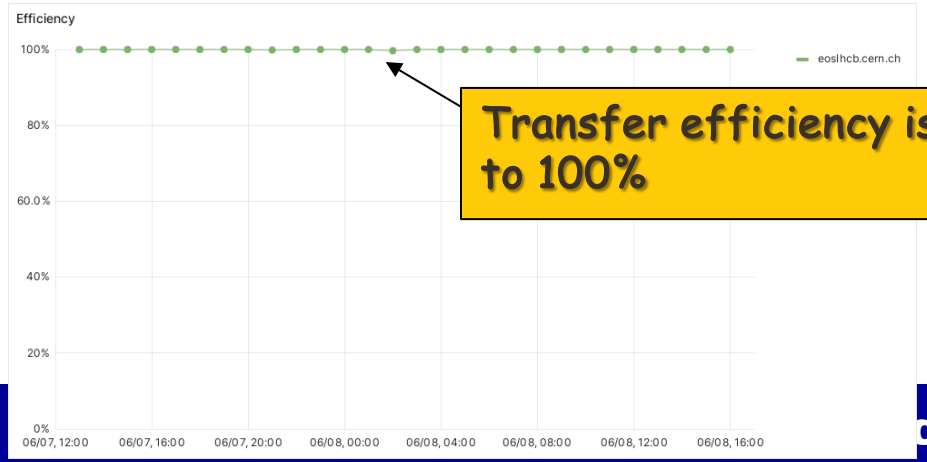
10-Jun-2024 17:29 from Alexander Rogovskiy : Test transfers for Beijing->CERN link

Message ID: **38431** Entry time: **10-Jun-2024 10:06** In reply to: **38425** Reply to this: **38440**

Author:	Alexander Rogovskiy
System:	Data Management
Production number:	
Site:	CERN | Beijing
GGUS Ticket:	
Trello/JIRA ticket:	
CC:	
Subject:	Test transfers for Beijing->CERN link

The test is over. On average 1.75GB/s was achieved (very close to EOS->Beijing disk result achieved during the Data Challenge), with almost no failures.
Plots are attached.

Attachment 1:	Beijing-Disk_EOS.zip 145 kB



Transfer efficiency is close to 100%

The Tier-1 site is scaling up

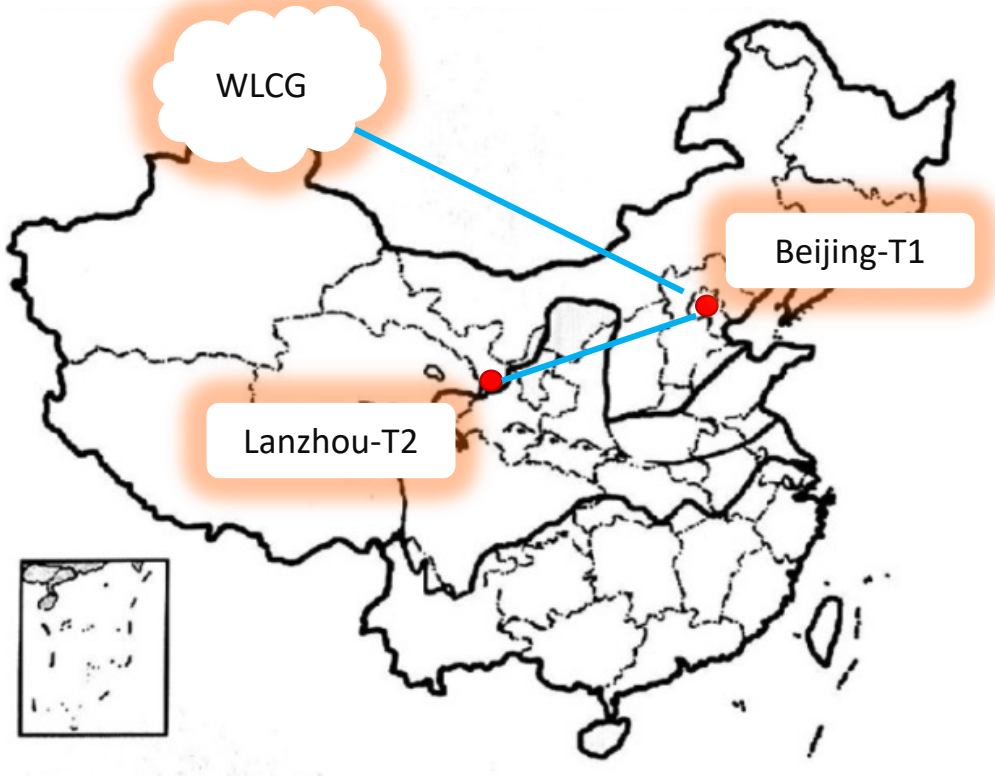
- LHCb needs that site contribution should be match to the number of authors

LHCb Demands	Beginning of BEIJING-T1	The number will be from next week
• Computing: ~5500 CPU cores • Disk Storage: 10.8 PB • Tape Storage: 23 PB	• Computing: 3280 CPU cores • Disk Storage: 3.2 PB • Tape Storage: 3 PB	• Computing: 3280 CPU cores • Disk Storage: 13 PB • Tape Storage: 10 PB

- LHCb is looking for more storage due to the storage problem of Italian site
- Gap still exist on the computing and tape storage

New LHCb Tier-2 Site at LanZhou University

- A new LHCb Tier-2 site is ready at Lanzhou University (LZU) in Sept. 2024
  - 3520 CPU cores, ~77,000 HS23
  - 3.4PB Disk Storage
  - Dedicated 2Gbps link between IHEP and LZU
- The biggest LHCb Tier-2 site (except CERN)
- Jointly maintained by CC-IHEP and LZU
  - Hardware maintenance: Lanzhou University
  - Software deployment and maintenance: IHEPCC



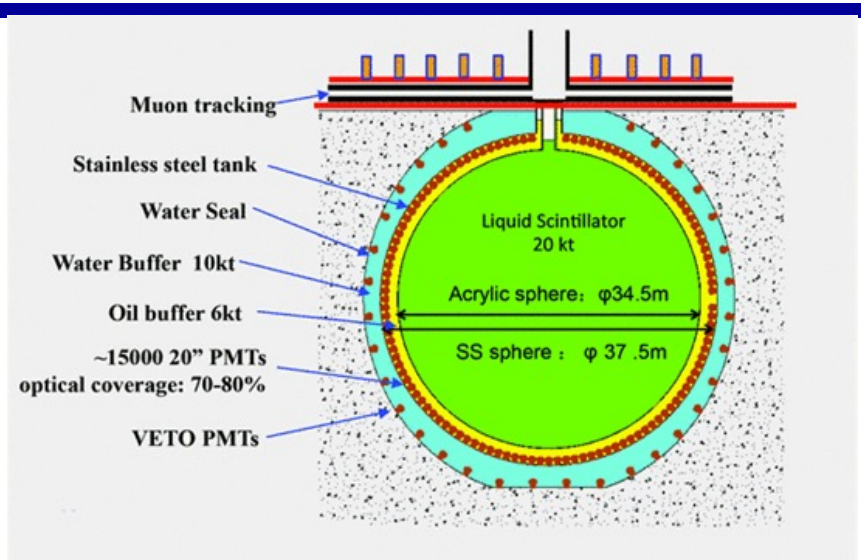
Outline

1 The Beijing LHCb T1 Status

2 JUNO Computing System

JUNO

- A multi-purpose neutrino experiment
  - Measure neutrino mass hierarchy and mixing parameters
  - Located at Guangzhou, China
  - Expect to take data in 2024
  - 20 kt Liquid Scintillator detector, 700m deep underground
  - 2-3% energy resolution
- JUNO-TAO is a satellite detector
  - Improve sensitivity of JUNO on mass hierarchy study
- Data volume expected
  - Raw: 2.4PB/year (JUNO+TAO)
  - MC+Rec: 600TB



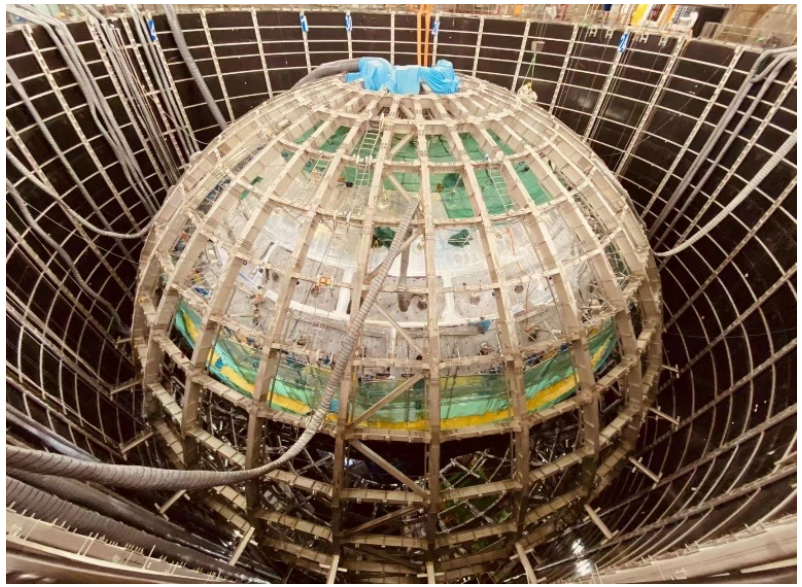
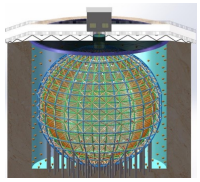


Photo taken in 2023.2

Data centers and Computing model



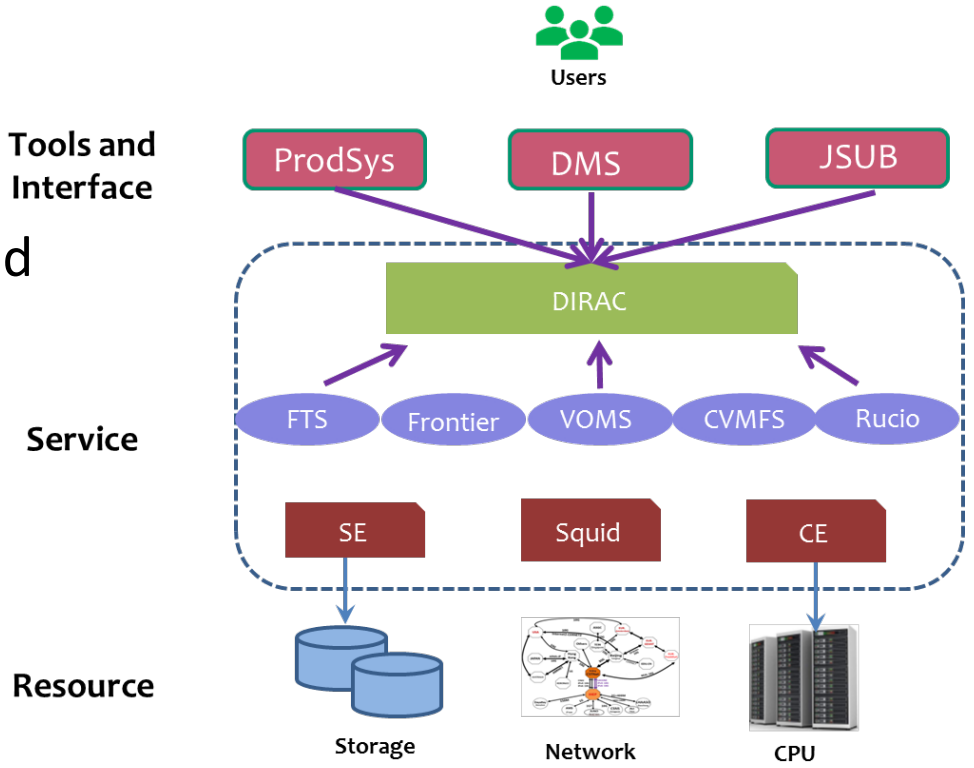
- Five data centers: IHEP, CC-IN2P3, INFN-CNAF, JINR, MSU
- Raw data flows from Online to IHEP which then distributes to other centers
- 1st Reconstruction and Calibration will run at IHEP
- MC Simulation, 2nd Reconstruction and Analysis are expected to run at all data centers
- Other centers provide a backup to JUNO data (CNAF/JINR 100%, IN2P3/MSU 1/3)

Resource Status

	IHEP	CC-IN2P3	INFN-CNAF	JINR	MSU	Total
Tape pledged (PB)	4.0	2.0	1.0	10.0	NA	17.0
Tape really present (PB)	2.7	2.0	1.0	5.1	NA	10.8
Tape available (PB)	**1.35**	**no quota**	**1.0**	**5.1**		**9.45**
Disk pledged (PB)	14.0	0.2	3.0	10.0	NA	27.2
Disk really present (PB)	4.0	0.2	3.0	1.0	NA	8.2
Disk available (PB)	**3.2**	**0.2**	**2.0**	**0.6**		**6.0**
CPU pledged (kHS23)	110.0	15.0	20.0	120.0	NA	265.0
CPU slots in grid batch system	1052	500	1282	2000	NA	4834
CPU really present (kHS23)	102.6	15.0	20.0	48.0		185.6

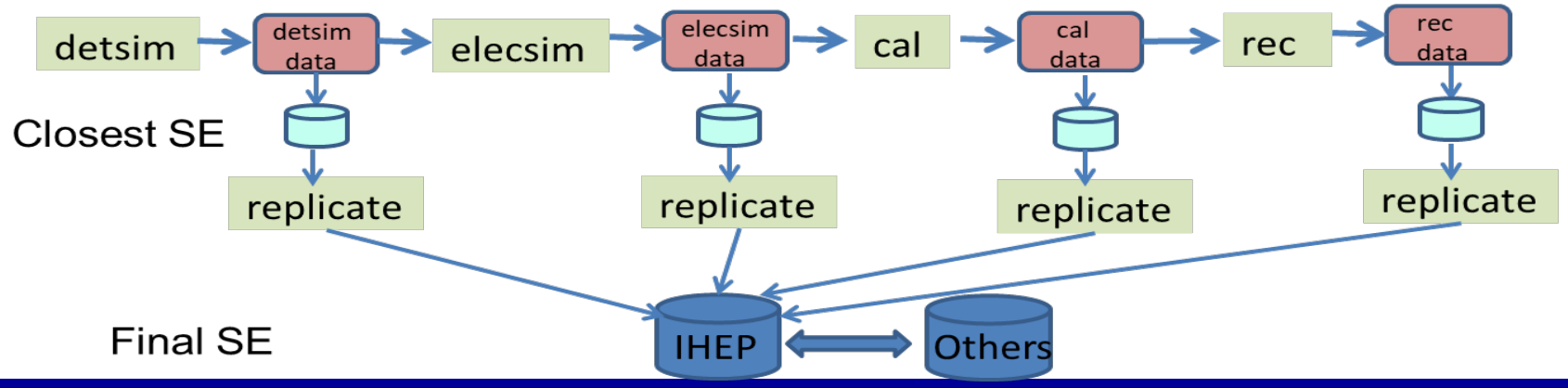
JUNO Distributed Computing System

- The system was built to deal with data processing activities and data distribution in grid environment
- DIRAC is core of the system
  - Organize heterogeneous resources
  - Integrate necessary services
  - Provide framework for workload management (WM) and data management (DM)
- Other WLCG services used
  - VOMS/IAM, authentication and authorization
  - FTS, file movement
  - CVMFS, software distribution
- Application tools and Interface (details in later slides)
  - JUNO-specific systems developed to meet the requirements of JUNO data placement and processing



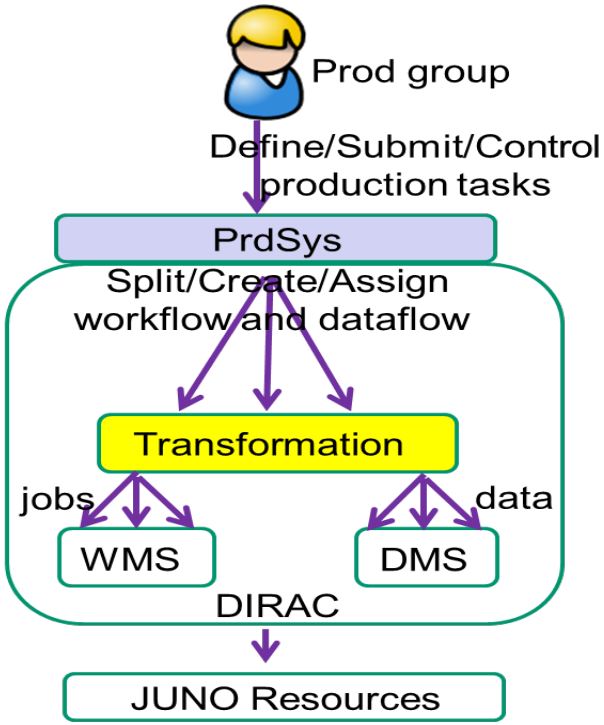
ProdSys

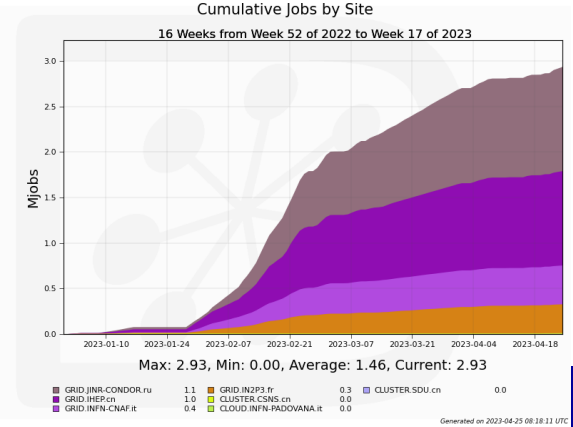
- ProdSys implemented as a data-driven pipeline system
  - submit JUNO production tasks (simulation, re-reconstruction…) in grid env
  - manage workflow and dataflow in tasks automatically
- Each JUNO production task is composed of several steps
  - Detector simulation (detsim), Electronics simulation (elecsim) , PMT Reconstruction (cal), Event Reconstruction (rec), Replication of output to destination sites
- All above steps can be connected to each other with data to form a pipeline, chained and started through ProdSys

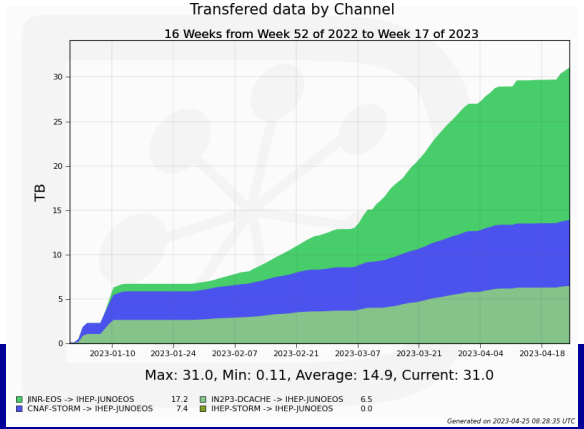


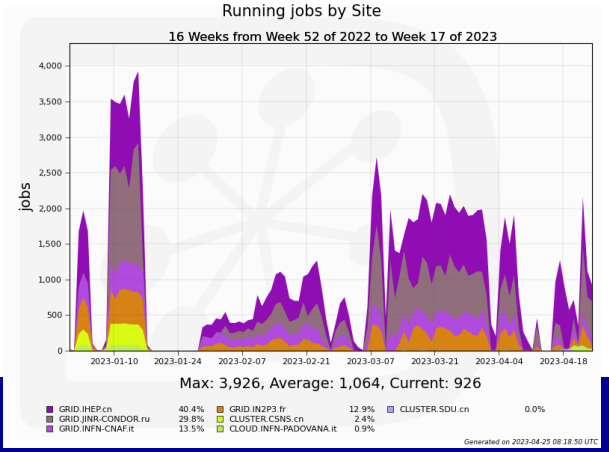
MC simulation tasks with ProdSys

- ProdSys is implemented based on DIRAC
  - Interface developed for users to define tasks
  - Transformation system allow to define and transform JUNO workflow and dataflow into a pipeline
  - DIRAC File Catalogue records data status, triggering data processing steps
  - Generated jobs and replicating tasks are submitted to DIRAC WMS and DMS



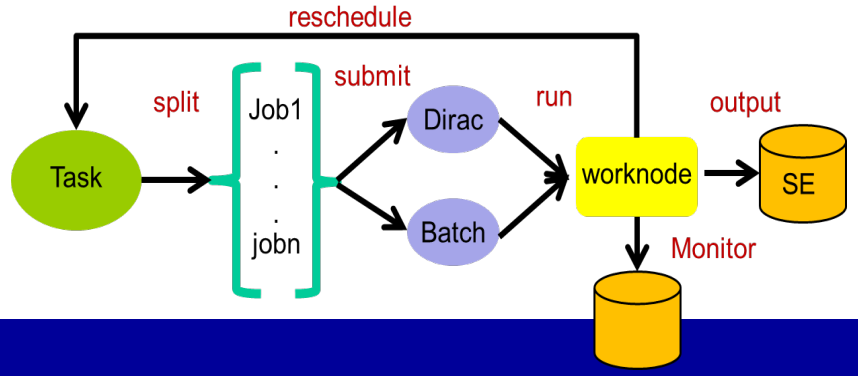


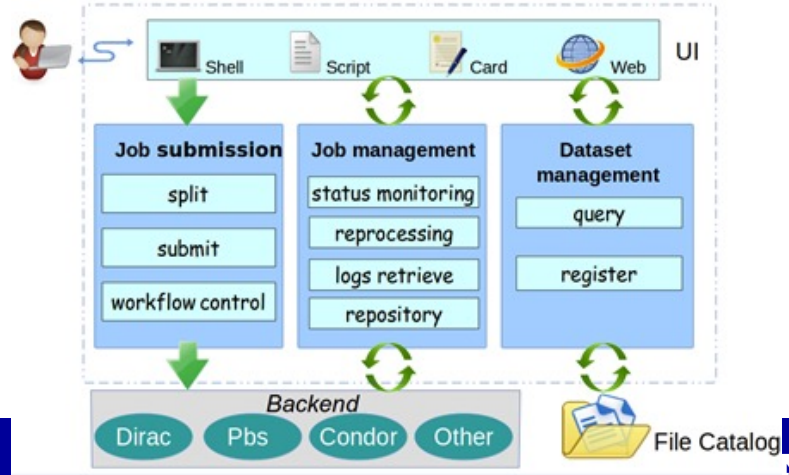




JSUB

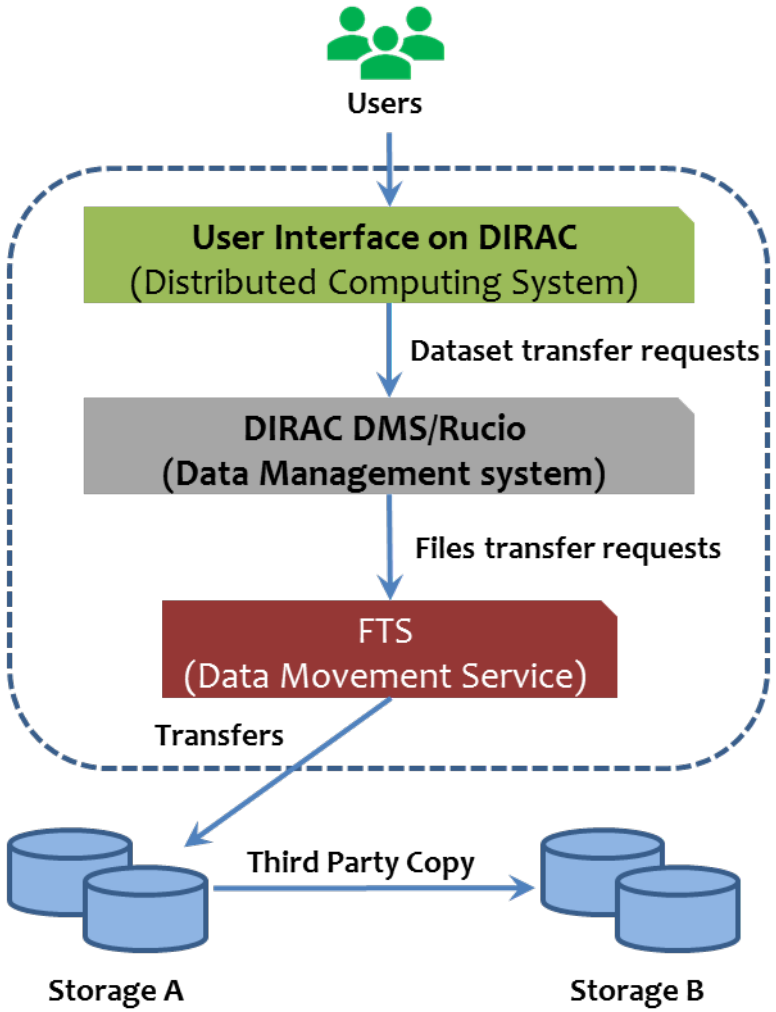
- JSUB – lightweight user job submission tool, developed in python
  - Ease process of physics analysis and small number of simulation for JUNO users
  - Automatically take care of life cycle of analysis tasks in grid env
- Main common function packages are provided include
  - Job submission, Job management, dataset management, backend, UI
- Extensible with experiment plugins and allow support of multi experiments
- Extensible with one more type of backends: DIRAC, Condor…..
- User Steering file is written in YAML
- Provide fast submission with DIRAC parameter job submission feature instead of one by one submission





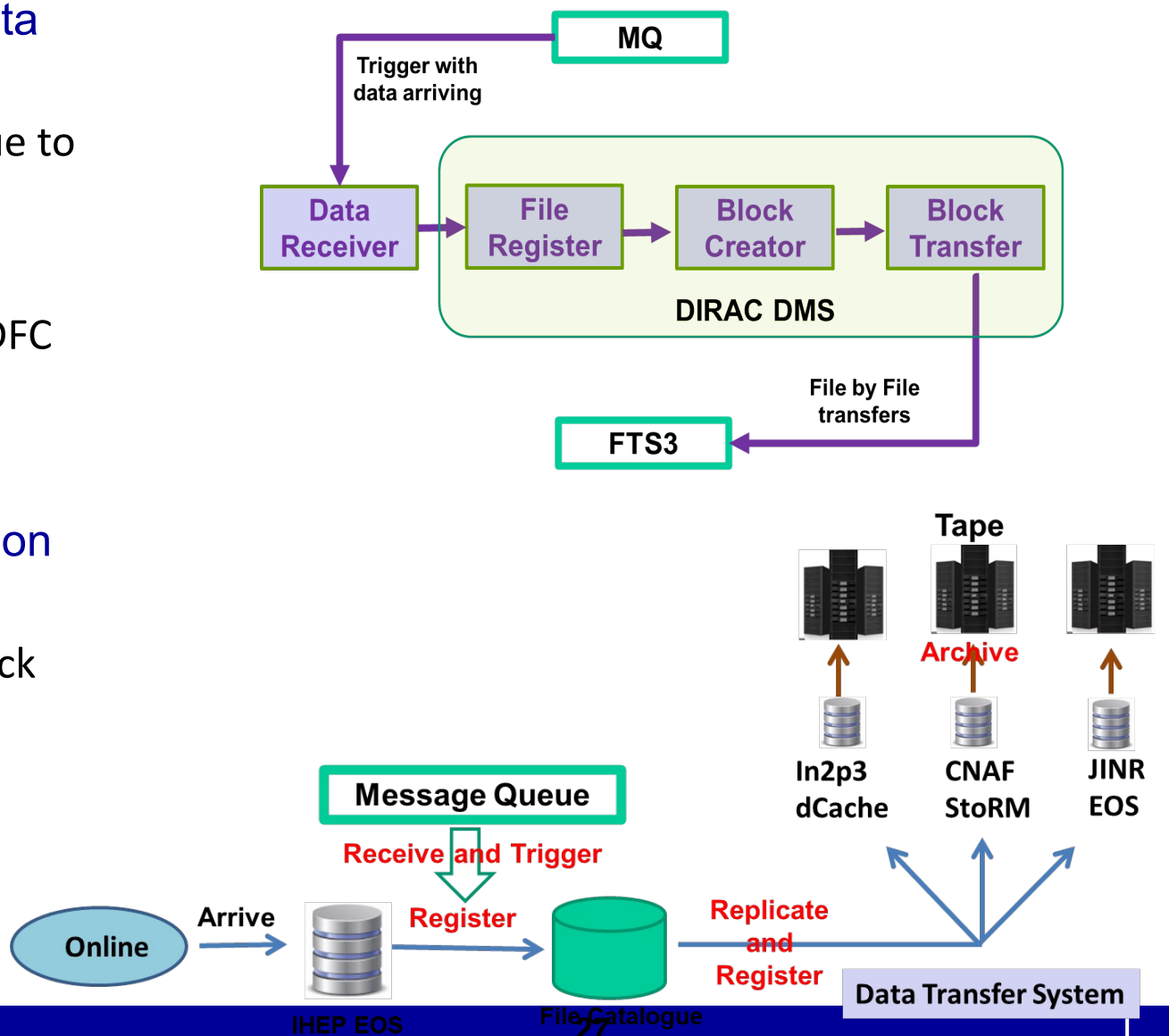
Data management

- User Interface
  - Provide a global data view for users
  - Provide dataset management
  - Provide tools on submission and management of bulk transfer requests
- DIRAC Data Management System (DIRAC DMS)
  - DFC: metadata and replicas catalogue
  - RMS: arrange requests in queue
  - Interface to file transfer tools
- Data Movement Service – FTS
  - Take care of file transfers



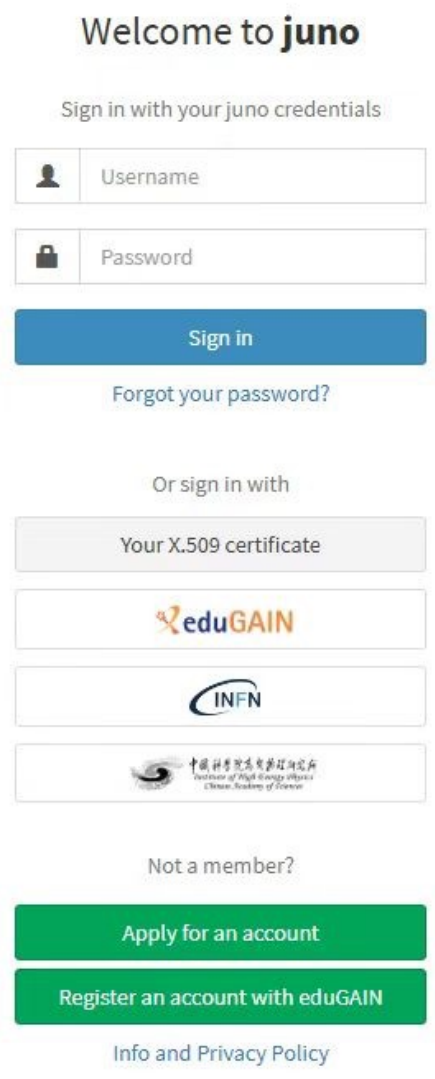
Raw Data Transfer System

- Aims to take care of raw data distribution to data centers
  - Receive data information from Message Queue to trigger the whole process
  - Register data in DFC
  - Replicate data to data center and register to DFC
  - Archive in tape and register in DFC
  - Validate data and monitor status
- Consists of four modules, implemented based on DIRAC DMS
  - Data Receiver, File Register, Block Creator, Block Transfer
  - Blocks are grouped based on date
  - Transfers and validation are based on blocks



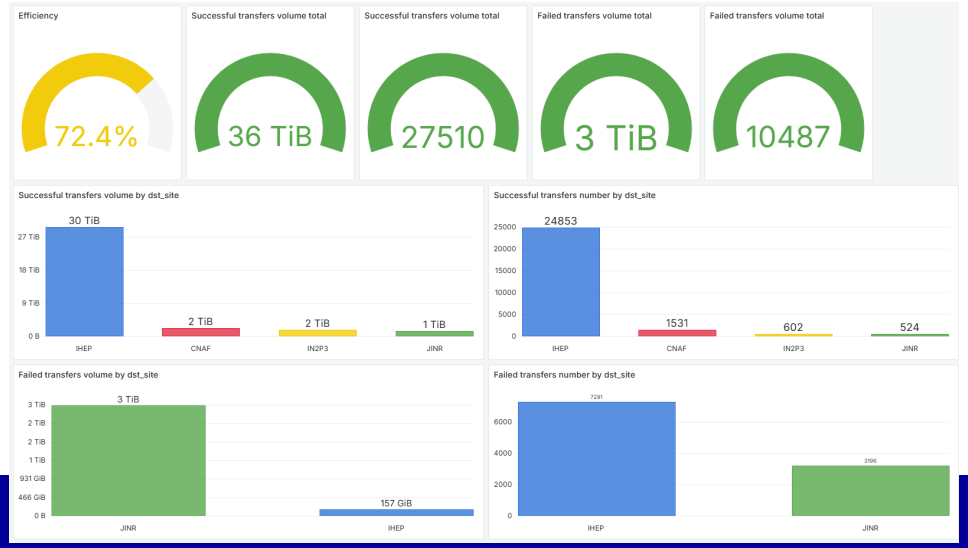
Token-based IAM

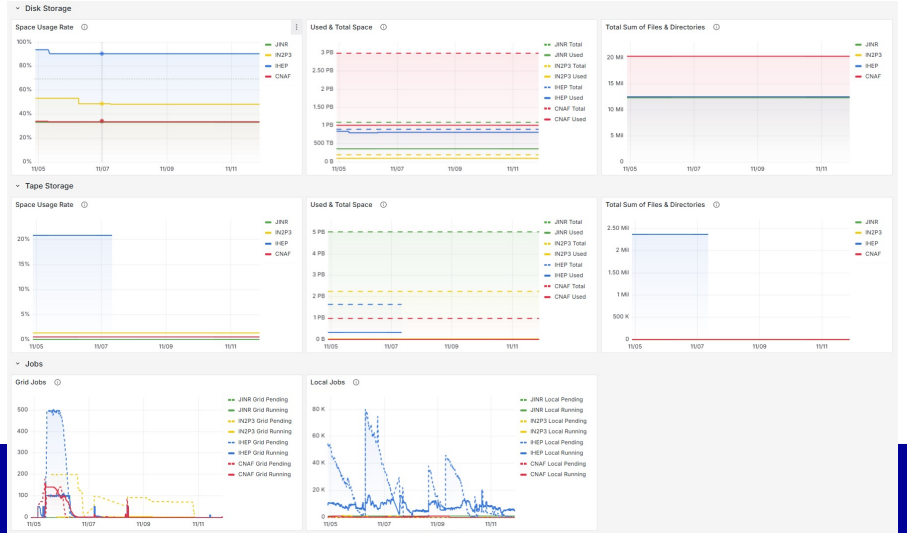
- Status:
  - IAM service has been set up: https://iam-juno.cloud.cnaf.infn.it/login
  - Connections to eduGAIN
  - IAM service IS in production in parallel with VOMS
    - New DIRAC version supporting both certificate and token
    - Enable site CEs and SEs to support token



JUNO DCI Monitoring System

- To monitoring all JUNO grid sites, middleware infrastructure, distributed computing systems running status.
- Based on workflow system with developed fine-grained probers to actively detect each component status.
- Data collection and visualization services for monitoring panels,
  - Including site operational status, data transfer status, data transmission statistics, and service operational status.
  - >2 billion prober logs collected since 2024.
  - Planning to extend our service to monitor other JUNO's system.

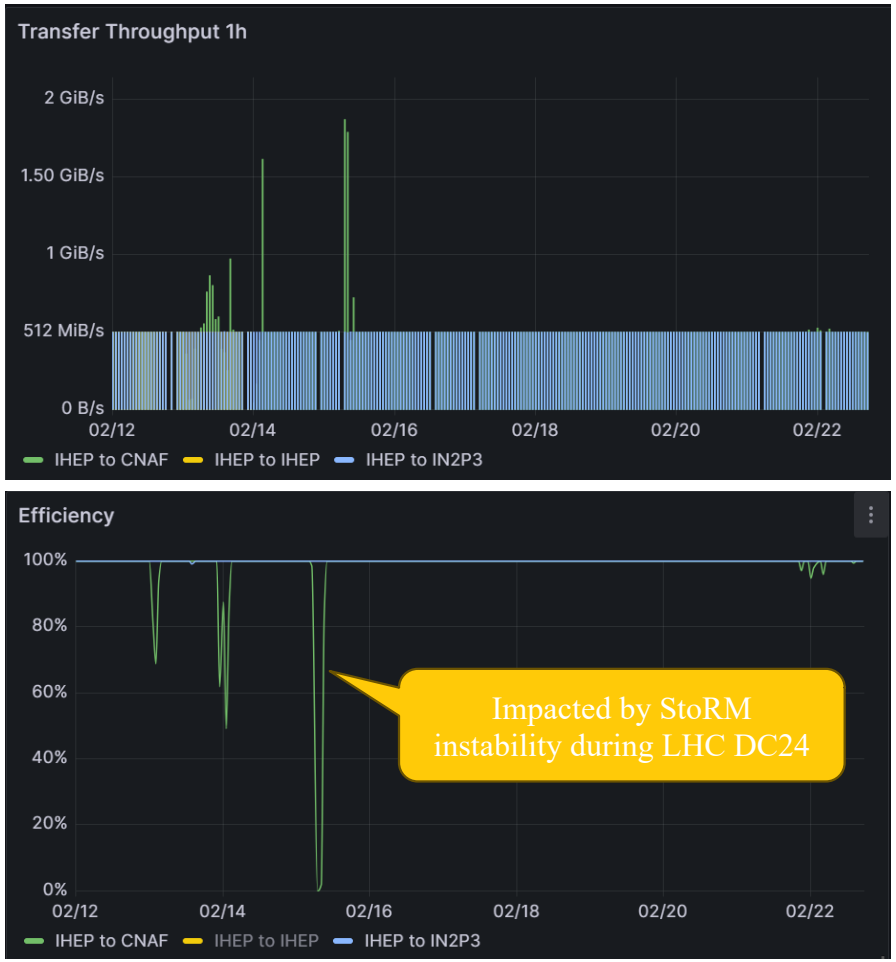




Data Challenge 2024

- JUNO and LHC DC24 are carried out on same period since 12th Feb 2024.
  - To challenge bandwidth limits.
- Optimization based on pre-transfer knowledge,
  - IHEP->CNAF, IHEP->IN2P3 directions are challenged. IHEP->JINR transfer needs to be optimized for data challenge.
  - Stress challenge with 12 PB/y mimic data throughput.
    - =180 transfers with 5 GB per file in 1 hour in each direction at the same time.
  - 3 time retries for failure transfer file.
  - FTS3 service is still in 200 max active streams but will be upgraded in next time data challenge.
- A Data injection tools are developed by Yifan Li,
  - Could conveniently control data injection frequency and injection size.

Data Center 2024 Results



- IHEP->CNAF met some connection failure and triggered re-transfer.
  - Known issue, CNAF StoRM met some running instability during LHC DC24, JUNO got impacted.
  - Come back to normal since 16th Feb.
- IHEP->IN2P3 transfer worked well.
  - 12 PB/year throughput is OK and a 24 PB/year throughput is in plan.

Summary

- IHEP have been running WLCG grid site for nearly 20 years and all the 4 LHC Experiments supported. **Thanks to IN2P3 for the continuous help!**
- Beijing LHCb Tier 1 is the first T1 in Asia. As the new Tier 1, **IHEP hopes to strengthen collaboration with IN2P3 continuously!**
- JUNO will be in production after nearly 10 years construction. Stable operation of JUNO Distributed Computing System is becoming **increasingly crucial for the experiment!**