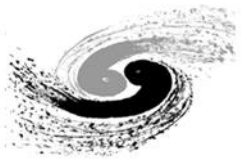


中國科學院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

国家高能物理科学数据中心
National HEP Data Center

高能所計算中心
IHEP Computing Center

Scientific Data Transfer System for High Energy Physics HiDTS

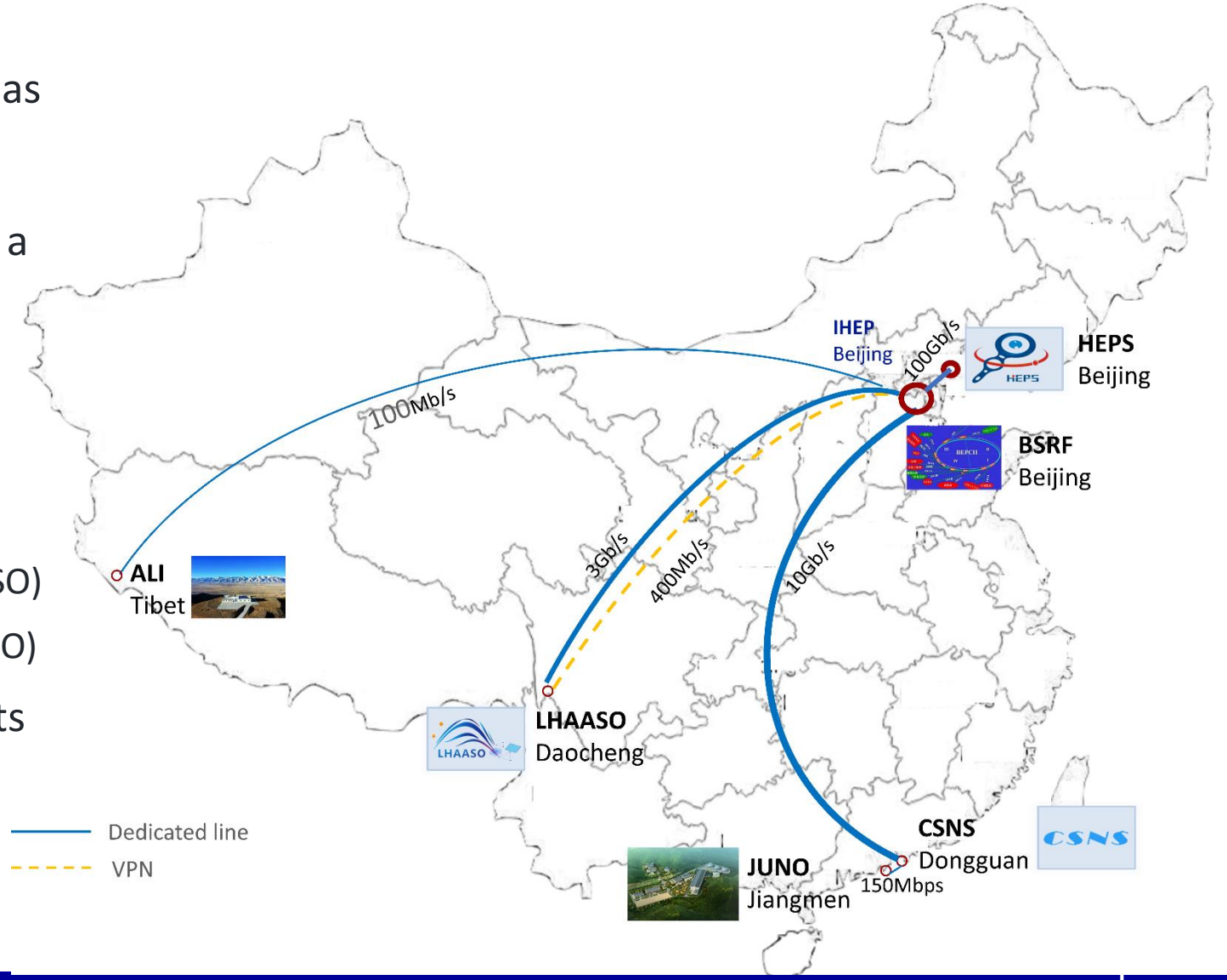
Bo Zhuang 、Hao Hu 、Shan Zeng ， Fazhi Qi
IHEPCC/HEPSCC
Institute of High Energy Physics, CAS

Outline

1. Background of data transfer
2. Challenge of data transfer
3. Technical architecture
4. System functional module.
5. System deployment and application status
6. Summary & Plan

Background of data transfer

- The Institute of High Energy Physics (IHEP) has established multiple large scientific facilities distributed across various locations, generating a large volume of data.
  - Beijing Synchrotron Radiation Facility(BSRF)
  - High Energy Proton Spectrometer(HEPS)
  - Ali CMB Polarization Telescope (AliCPT)
  - Large High Altitude Air Shower Observatory(LHAASO)
  - Jiangmen Underground Neutrino Observatory(JUNO)
- This data requires transfer to IHEP and assists scientists in analyzing experiments in a timely manner.



Challenge of data transfer

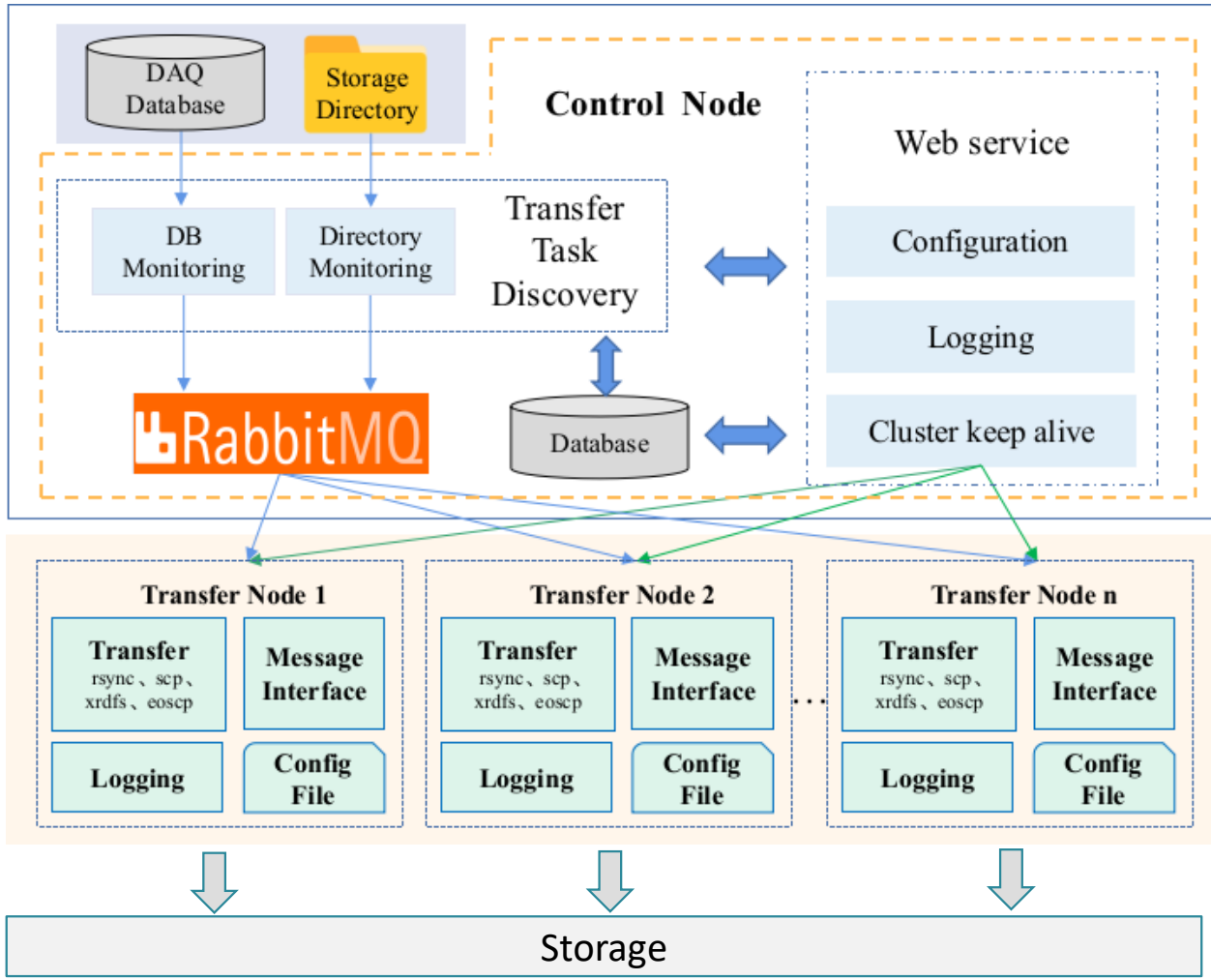
- Huge amount of data transfer Challenge
  - The data volume is rapidly increasing
  - 805.4TB raw data will be produced per day in HEPS
  - 10TB raw data will be produced per day in JUNO
  - 10.2TB raw data is produced per day in LHAASO
- Reliability challenge
  - AliCPT is in high-altitude
    - power outages
    - network disruptions
- Metadata Integrate with other system
  - DOMAS in BSRF and HEPS
  - Rucio in Herd
  - Pipeline in Juno
- Multiple file storage systems in IHEP

Storage System

- Lustre
- EOS
- 9950/9550
- Ceph
- CTA

Technical architecture

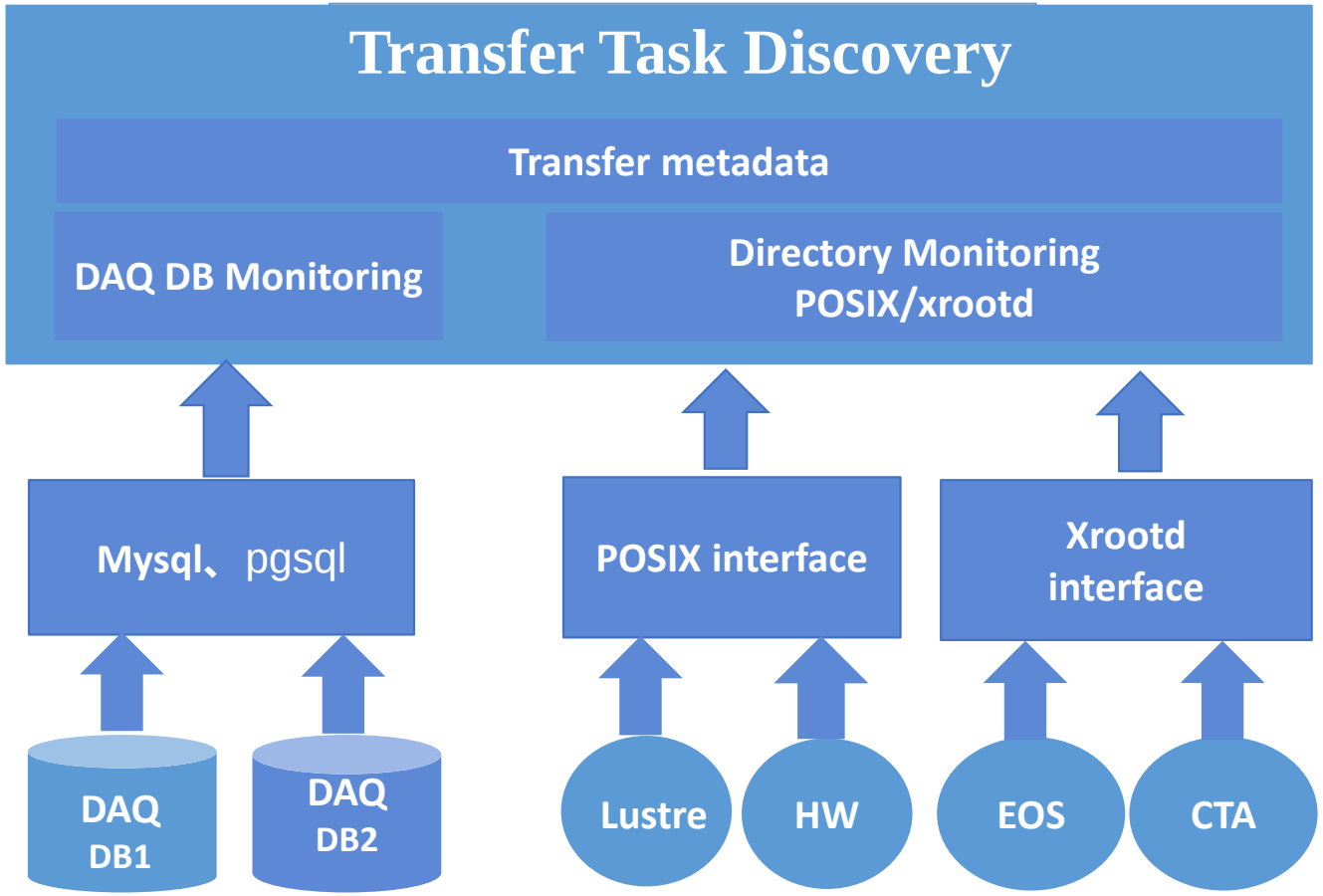
- One control node and multiple transfer nodes
- Control Node
  - Unified web portal for configuration and display
  - Transfer Task Discovery responsible for discovering files
  - RabbitMQ for transfer message buffer
  - Redis for Cluster keep alive management
- Transfer Nodes
  - Transfer Data in different storage
  - Interacting message with other systems
  - File transfer logging
  - Share the same rabbitMQ、Redis、transfer DB and mount the same storage



Transfer Task Discovery

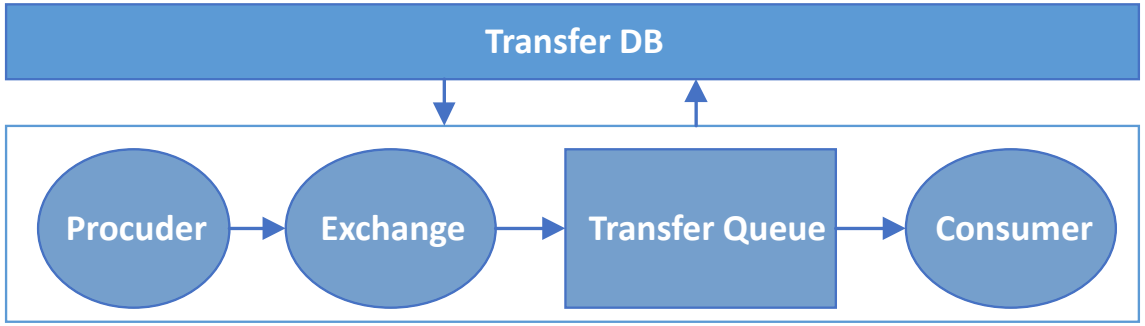
- Monitor transfer files and send messages to the transfer buffer
- Based on web service configuration
- Experiment -> Transfer Task -> DAQ DB monitoring

  -> Directory Monitoring
- DAQ DB monitoring
  - Mysql/Postgresql
  - DAQ DB interface information
- Directory monitoring
  - POSIX/Xrootd
  - Directory information
  - Regex-based matching for directories and files
- We use the last_watched_time as an identifier to discover the newest files

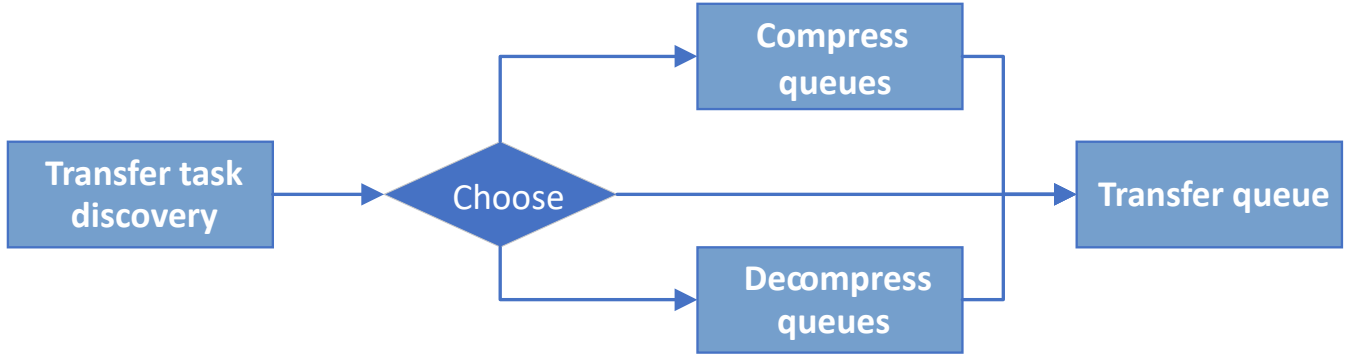


Transfer buffer

- Responsible for managing asynchronous transfer message from transfer task discovery module
- Using RabbitMQ message queue
- Design approach
  - Transfer message queue to decouple the messages
    - Task discovery produce messages to transfer queue.
    - Transfer nodes consume messages from queue
  - File information is stored in the transfer database
  - Compress and decompress messages queues
    - Configure it in web task page
    - Support Compressed data transfer
    - Manage compress and decompress messages



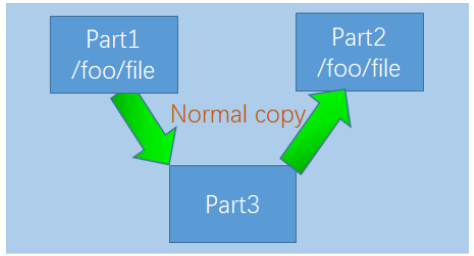
Transfer message queue

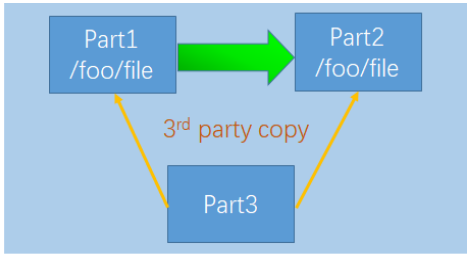


Transfer system queue

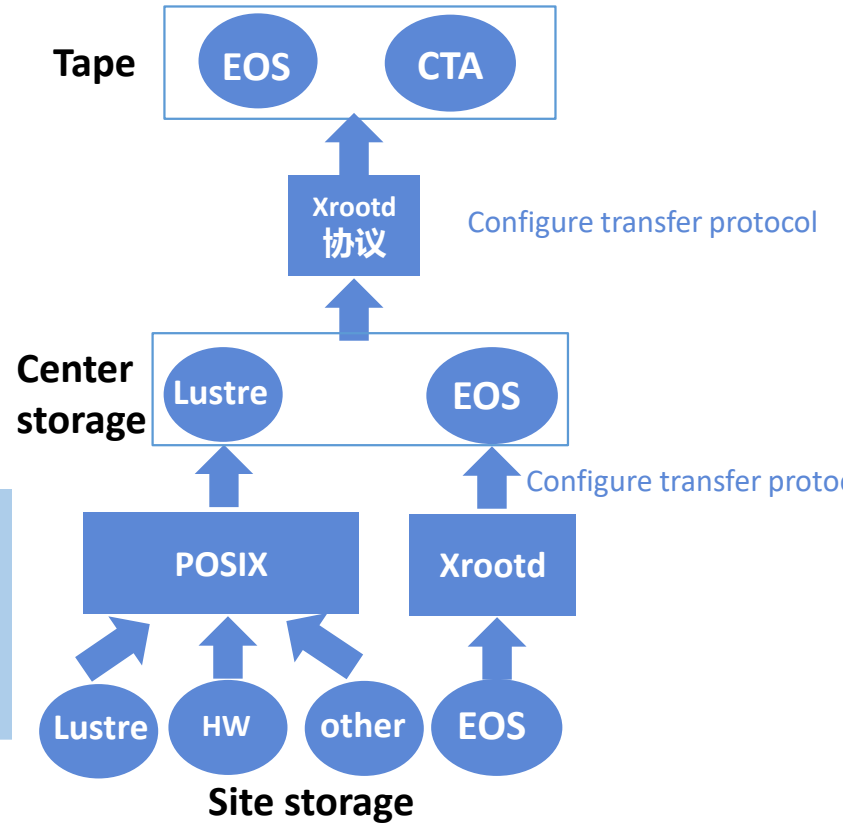
Data transfer

- Consume transfer message from transfer buffer and transfer files.
- support multiple data transfer protocols
  - rsync、scp、eos cp、xrdcp
- High Performance
  - Cluster deployment
  - Multi threads of each transfer node
  - Support Xrootd TPC
- Reliability
  - Checksum verification
  - Automatically retransfer failed file
  - failed files are stored in an failed buffer and retransferred after the issue is resolved
  - Logging and monitoring





Xrootd TPC

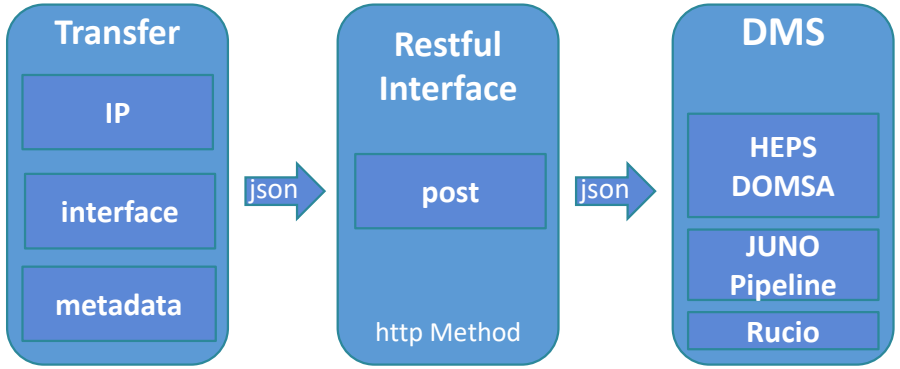


Raw data transfer

Message Interface

- provides Universal interfaces to interact with other systems.
- Generic Design
  - Based on configuration: IP, port, metadata information
  - Metadata is transmitted through RESTful interface
    - Loose coupling: Using standard RESTful
    - Standardization: Interface design based on JSON
    - Compatibility: platform-independent, sending agreed-upon JSON metadata messages
- Reliability Design
  - 5 attempts for message notification
  - Exception message queue, re-sending after server recovery

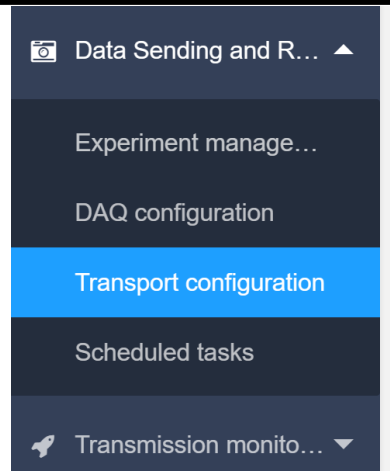
Metadata format to DOMAS

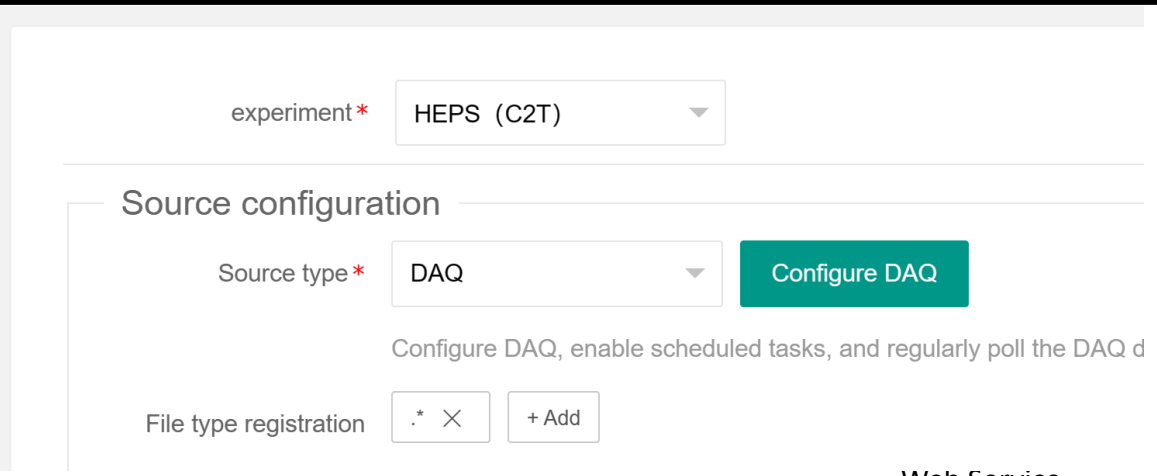


Message interface

Web service and Config

- Web Service
  - Experiment management
  - Transfer tasks management
  - Data Acquisition (DAQ) configuration
  - Scheduled tasks
- Backend configuration（application.yml ）
  - Server
  - Logging
  - Spring Boot
  - MySQL
  - Redis
  - Quartz
  - RabbitMQ





Web Service

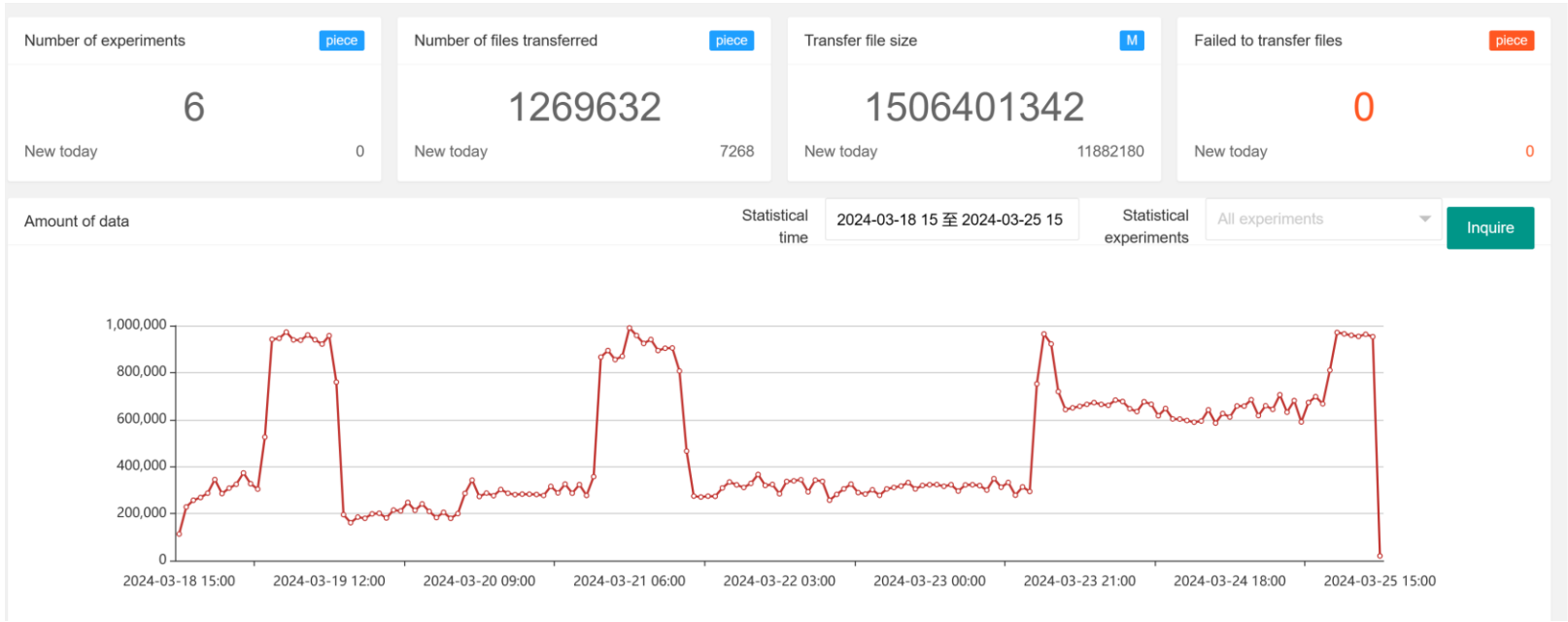
```yaml
##spring相关配置
spring:
  application:
    name: cosmoseye-boot
  profiles:
    active: dev
  messages:
    basename: static/i18n/messages
  main:
    allow-circular-references: true
  web:
    resources:
      static-locations:
      - classpath:/static
      chain:
        enabled: true
        cache: true
        compressed: true
      cache:
        period: 604800
  data:
    web:
      pageable:
        ##layui表格页码数从1开始
        one-indexed-parameters: true
```

Logging and Monitoring

- Logging module
  - The log records the entire lifecycle of data transfer
  - View the logs through the web service.

```
DEBUG 2022.11.30 09:08:05.748 cn.cosmoseye.action.transfer.FileTra... : 开始进行第3次文件传输，处理服务器:192.168.60.162
DEBUG 2022.11.30 09:08:05.944 cn.cosmoseye.ssh.SshServer.send : 开始传输文件：/hepsfs/transfertest/4W1A/2022/20220721test-cw/sample1/sam/sam_89.000.tif，信号文件：无
ERROR 2022.11.30 09:08:36.019 cn.cosmoseye.action.transfer.FileTra... : 文件传输失败
文件不存在：/hepsfs/transfertest/4W1A/2022/20220721test-cw/sample1/sam/sam_89.000.tif
```

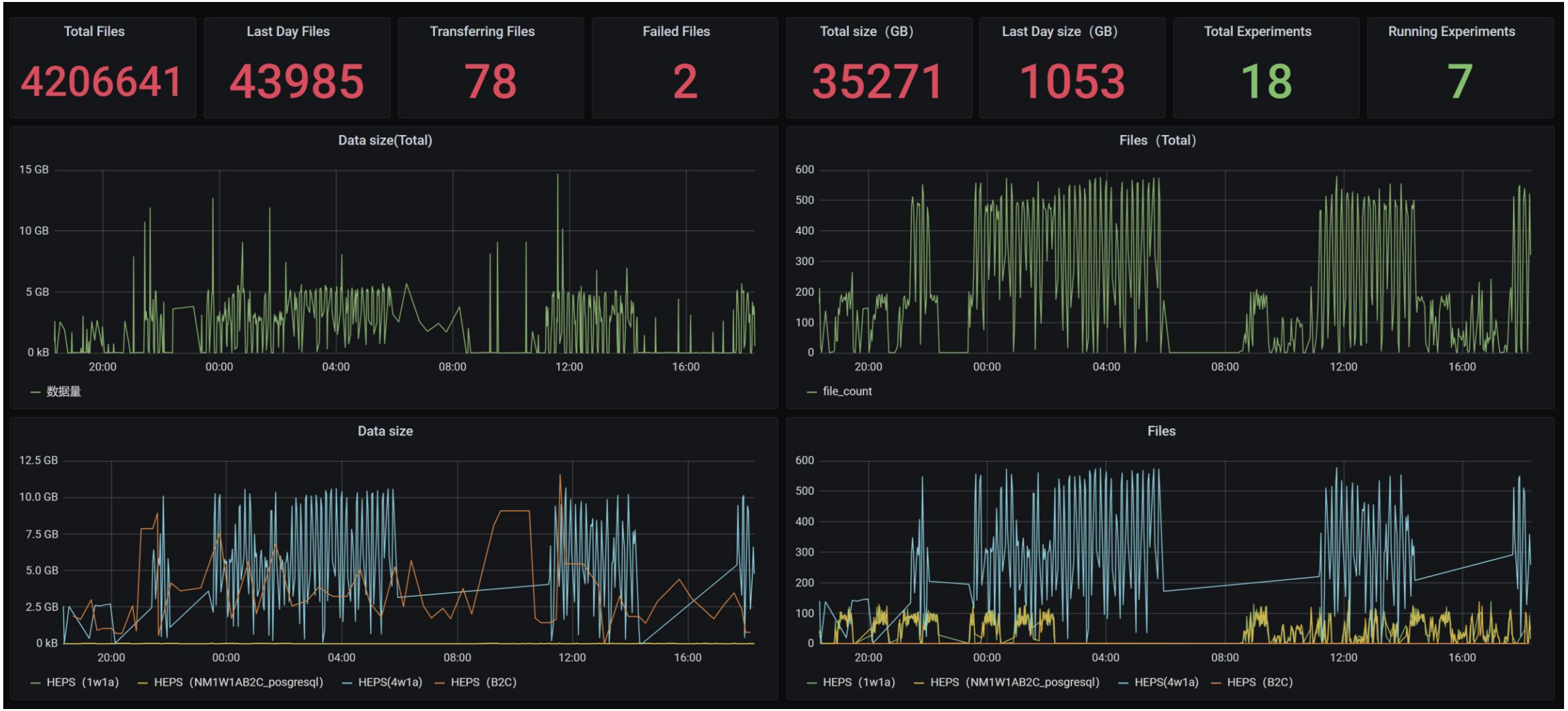
- Monitoring Module
  - Monitor the status of data transfers, including statistics..
  - Web-based graphical representation
  - WEB-based list representation
    - Transfer list
    - To be transferred list
    - Failed files list



Dashboard for users

国家高能物理科学数据中心
National HEP Data Center

高能所計算中心
IHEP Computing Center

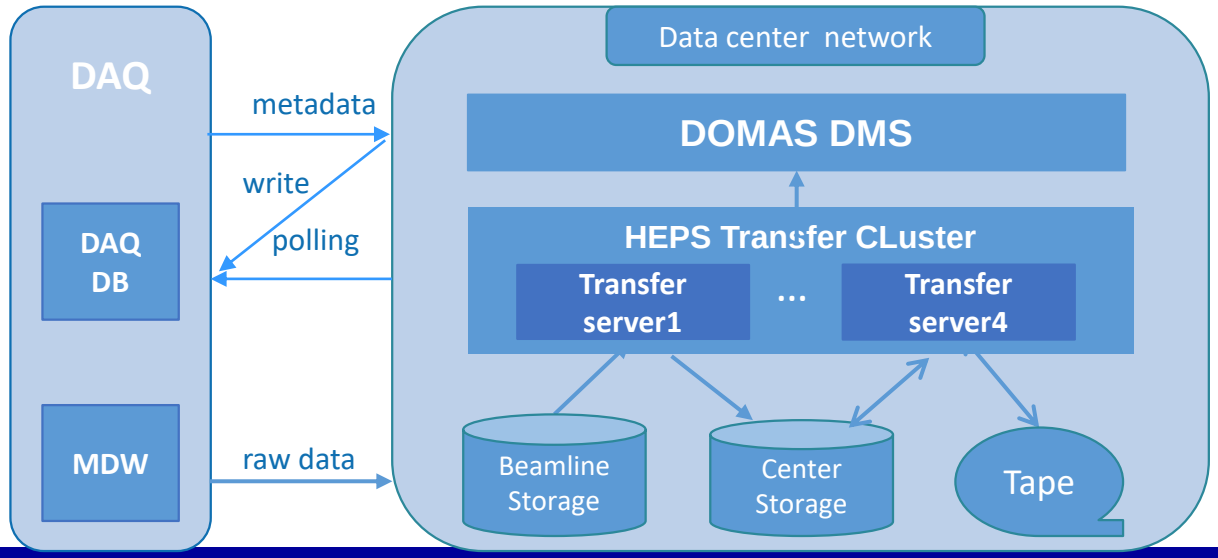


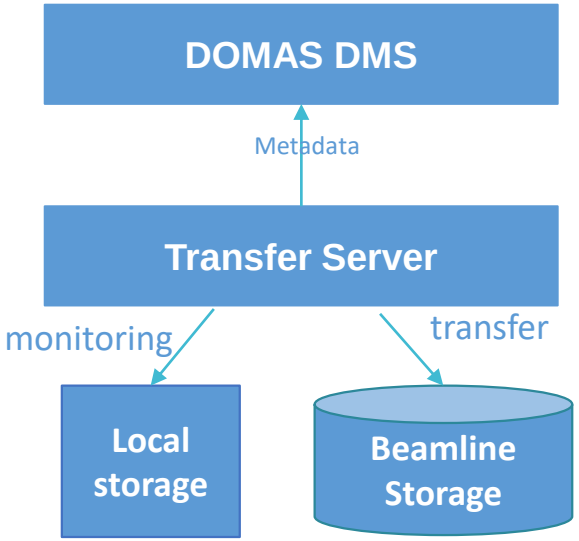
Data transfer at BSRF

- BSRF
  - Beijing Synchrotron Radiation Facility
  - Located in IHEP, Beijing, China
  - 14 beamlines are in operation
- The data transfer at BSRF is for the preparation of HEPS
  - Transfer process : Local storage → Beamline storage → Center storage→ CTA
  - Interacting metadata with DOMAS
  - Multiple storage systems
- Infrastructure status at BSRF
  - Two data transfer servers with 10Gb network cards.
  - 10/25Gbps data center network link

The deployment at BSRF

- Deployed in 4W1B and 3W1 beamline in 2022.06 , 1W1A and 4W1A beamline in 2023.06.
- The optimization we have done.
  - Support transferring File and Directory Access Control Lists (Facl)
  - Configure directory filtering rules for the beamtimeID directory
  - Support data transfer to beamline storage without Mamba（DAQ）
  - According to the beamline station experimental mode, support continuous writing of file transfer
  - Updated the DAQ (Data Acquisition) database from MySQL to PostgreSQL for improved performance

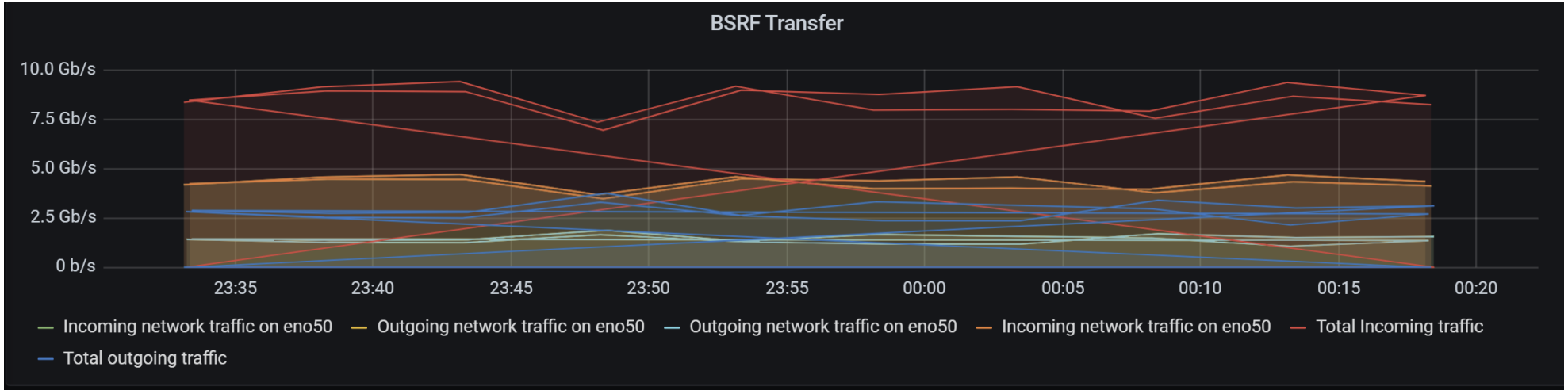


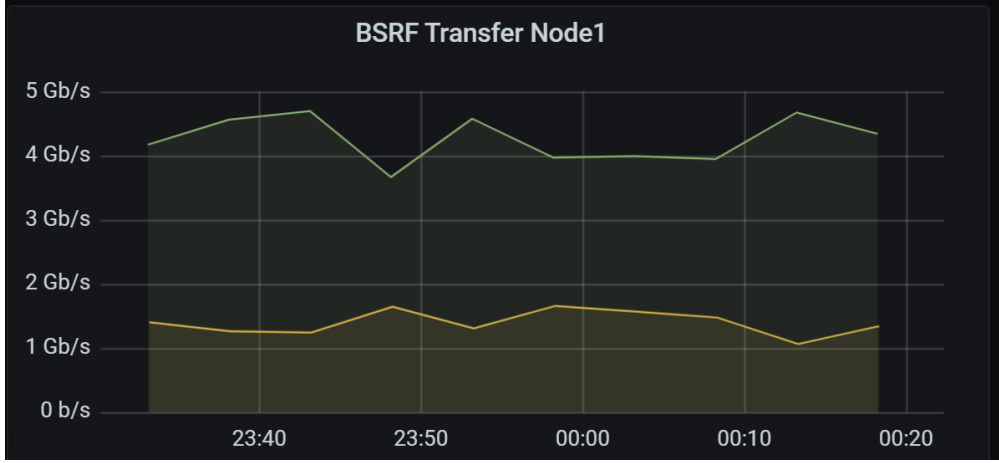


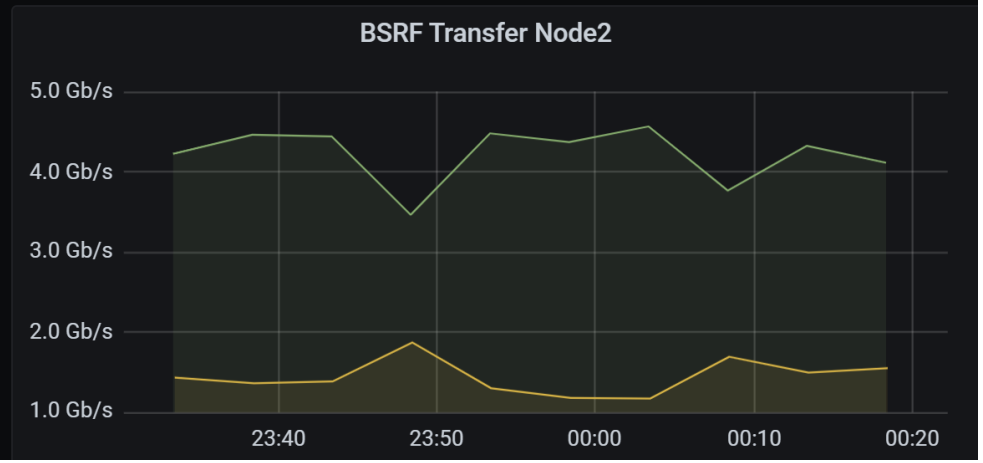
Transfer performance at BSRF

国家高能物理科学数据中心
National HEP Data Center

- In the case of sharing a 10Gb bandwidth with other systems, the transfer performance is good, and the total transfer throughput can reach around 9Gb

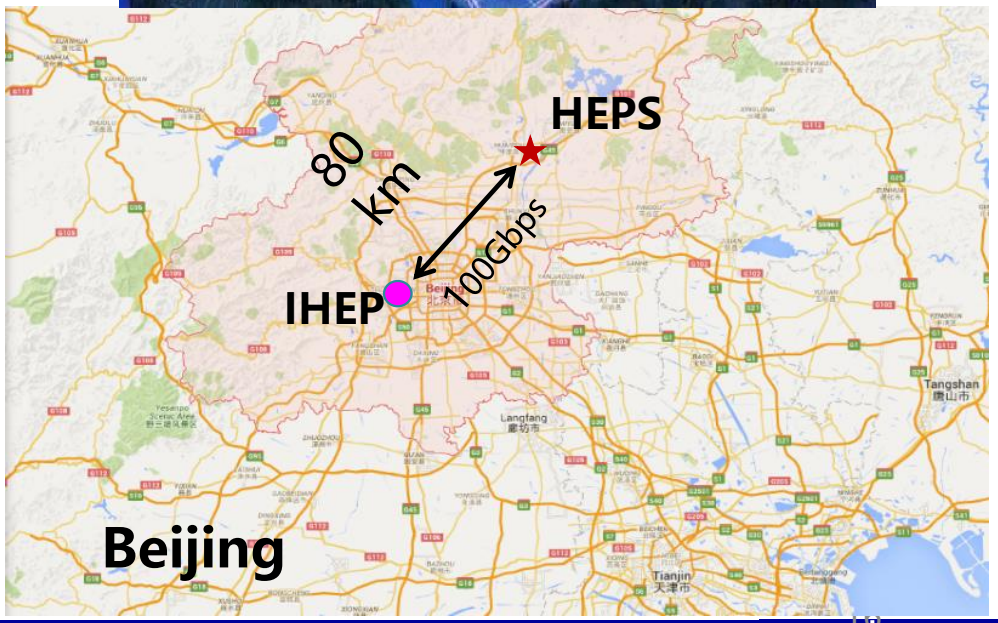






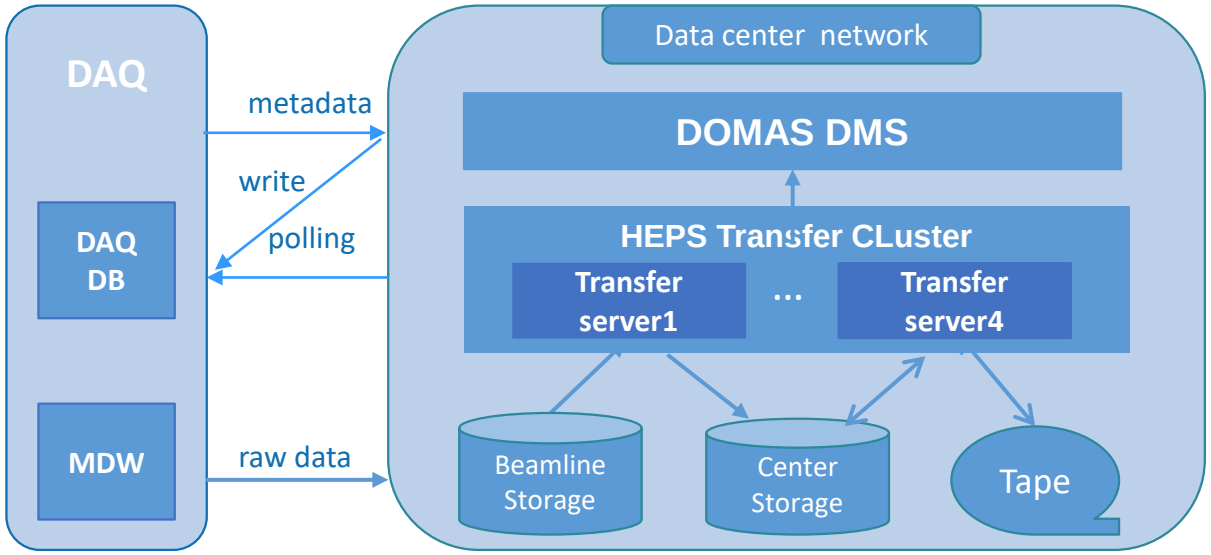
Data transfer at HEPS

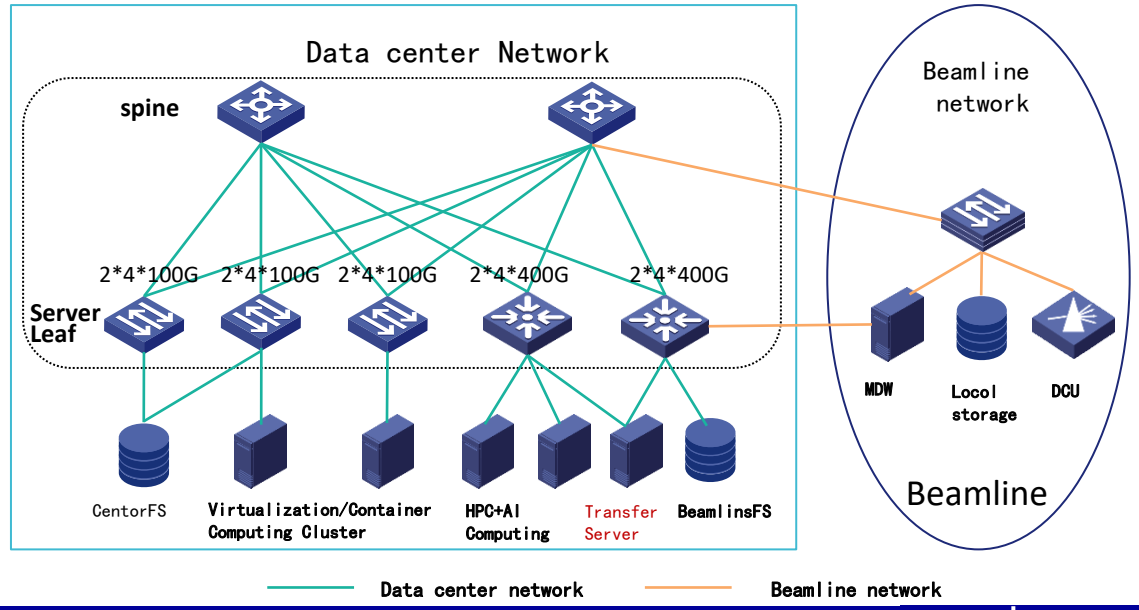
- HEPS
  - High Energy Proton Spectrometer
  - Located in Beijing - about 80KM from IHEP
  - The whole project will be finished in mid-2025
  - 14 public beamlines + 1 optics test beamline in Phase I
- Huge amount of data is a big challenge for data transfer
  - 805.4TB raw data will be produced per day
  - Data needs to be transmitted in a timely manner
- Infrastructure status at HEPS
  - Four data transfer servers with 100Gbps network cards
  - 100Gbps/25Gbps data center network
  - Network support RoCE and Ethernet



The deployment at HEPS

- Deployed at HEPS data center in 2023.09.
  - Deployed four data transfer servers as transfer cluster now
  - The data transfer scheme is the same as BSRF
  - The transfer scheme were tested and work well
  - High performance data transfer will test through data center network next（RoCE and Ethernet）



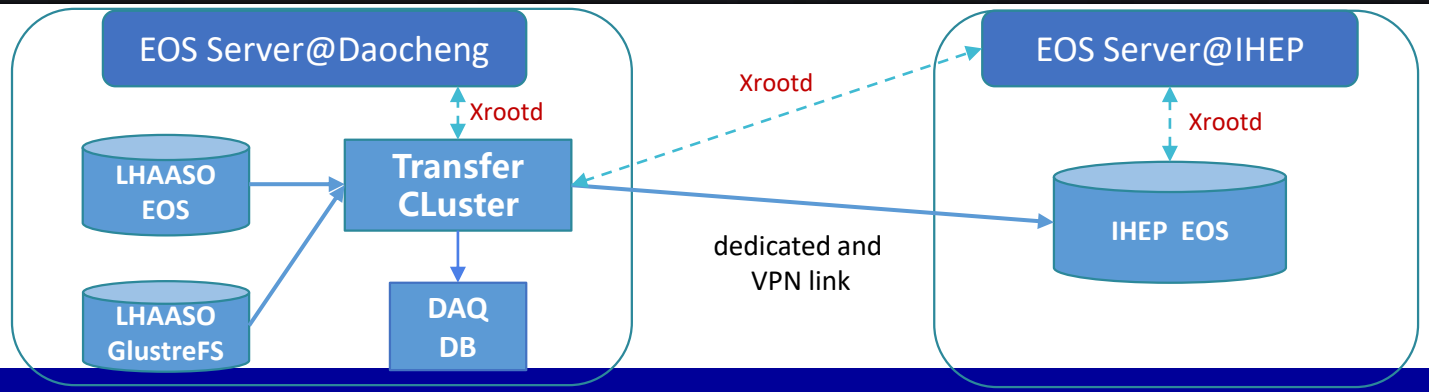


Data transfer at LHAASO

- LHAASO
  - Large High Altitude Air Shower Observatory
  - Located in Daocheng , Sichuan - about 1738KM from IHEP
  - Altitude 4410 meters
- LHAASO is a remote site that has a large amount of data
  - 10.2TB raw data is produced per day now
  - Data needs to be transmitted to IHEP in a timely manner
- Infrastructure status at LHAASO
  - Two data transfer servers with 10Gbps network cards
  - 3Gbps dedicated link and 400Mbps VPN from LHAASO to IHEP
  - Use EOS storage both in LHAASO and IHEP

The deployment at LHAASO

- Deployed in LHAASO
  - Upgraded to HiDTS in 2023.10.
    - Currently two transfer nodes onsite, each with a 10Gb/s network interface.
  - An automatic retransfer Module was deployed
    - Retransfer the files in the transfer failed list every hour
- Transfer performance is good, Continuously and fully utilize the 2G bandwidth before.

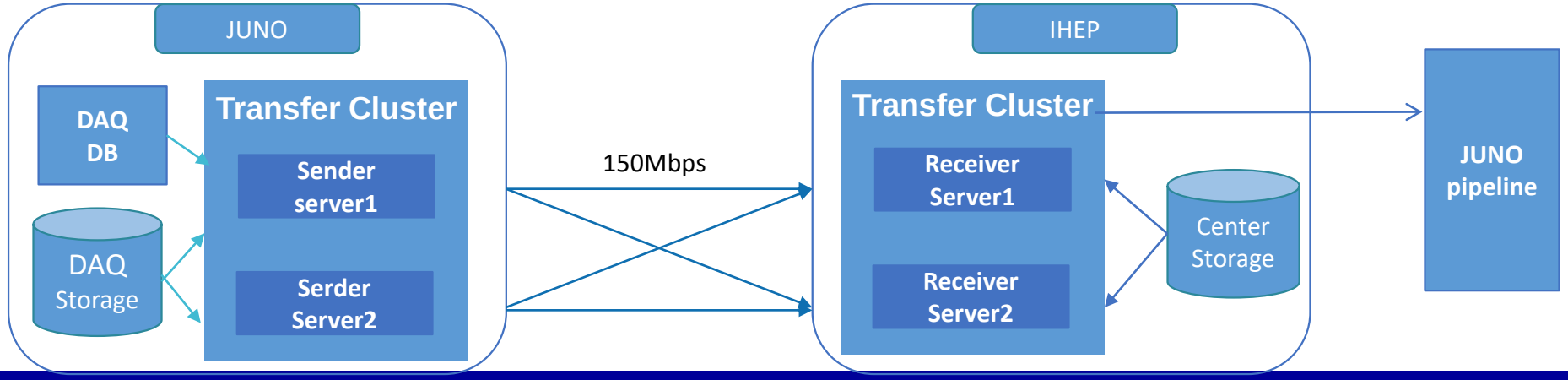


Data transfer at JUNO

- JUNO
  - Jiangmen Underground Neutrino Observatory
  - Located in Jiangmen , Guangdong - about 2000KM from IHEP
  - 700 meters underground
- JUNO is a remote site and has a large amount of data
  - Estimated data volume: 200 files and 10TB per day
  - Data needs to be transmitted in a timely manner
  - Interacting metadata with Pipeline system
- Infrastructure status at JUNO
  - Two data transfer servers with 1Gbps network cards in JUNO
  - Two data transfer servers with 10Gbps network cards in IHEP
  - 150Mbps dedicated link from JUNO to IHEP , 2Gbps in future

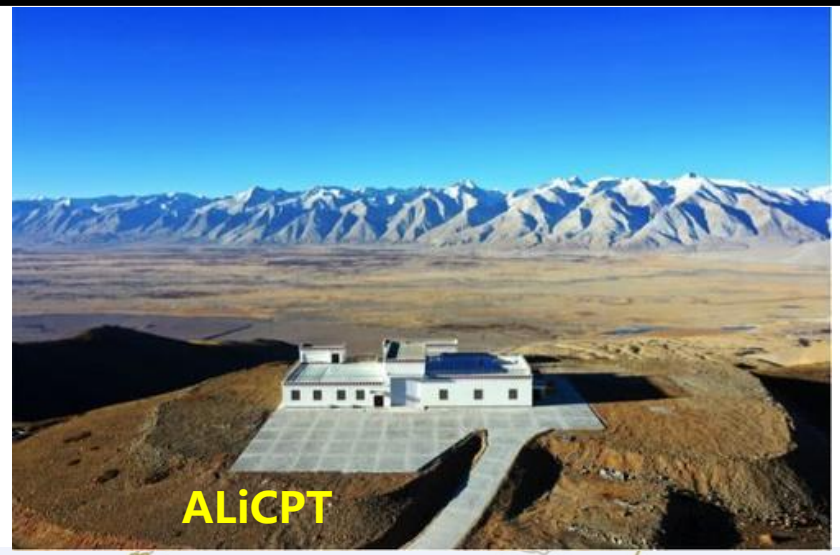
The deployment at JUNO

- The deployment at JUNO follows a similar to the deployment way used in our previous experiments at DayaBay
- Deployment In JUNO
  - Two send server deployed in 2023
  - Transfer data and semaphore file to receiver server buffer
- Deployment In IHEP
  - Two receiver server deployed in 2023
  - Transfer raw data from buffer to JUNO EOS
  - Send metadata Pipeline
- The transfer process has been verified



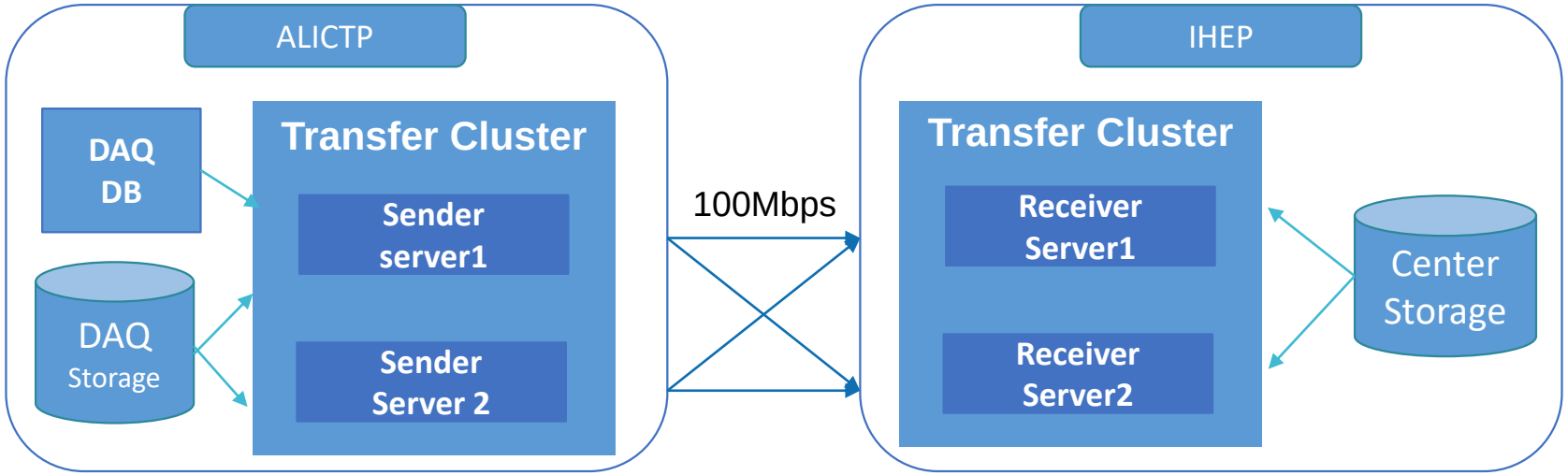
Data transfer at ALICPT

- ALICPT
  - Ali CMB Polarization Telescope
  - Located in Ali , Tibet - about 3000KM from IHEP
  - Altitude 5250 meters
- High reliability challenge due to high-altitude, power outages, and network disruptions.
- Estimated data volume: 150TB/Year
- Infrastructure status at ALICTP
  - Two data transfer servers with 1Gbps network cards in Ali B1
  - Two data transfer servers with 1Gbps network cards in IHEP
  - 100Mbps dedicated link from ALICPT to IHEP.



The deployment at AliCPT

- Deployed in ALiCPT and IHEP
  - We used spade transfer system for testing before
  - Upgraded to HiDTS in 2024.01
- high reliability
  - Two servers both in ALiCPT and IHEP
  - Retransfer for 5 times and automatic retransfer Module was deployed
  - Establish data replicas at Ali B1 site to prevent network interruptions



Summary & Plan

- HiDTS is developed and will be released as open-sourced progressively
- HIDTS is using in BSRF、HEPS、LHAASO、JUNO、ALICTP and NHEPSDC
- Promote the application of HiDTS at SHINE、Herd and multi Storage Center
- It can be integrated into RUCIO as a plugin
- High speed transfer test Based RoCE network in HEPS Data Center
- Keep on Improving reliability and high performance
- Keep on cooperating with other big Scientific facilities and communities

Thank you for your attention!

Comments or suggestions?