EOS Workshop 2020

EOS status at IHEP

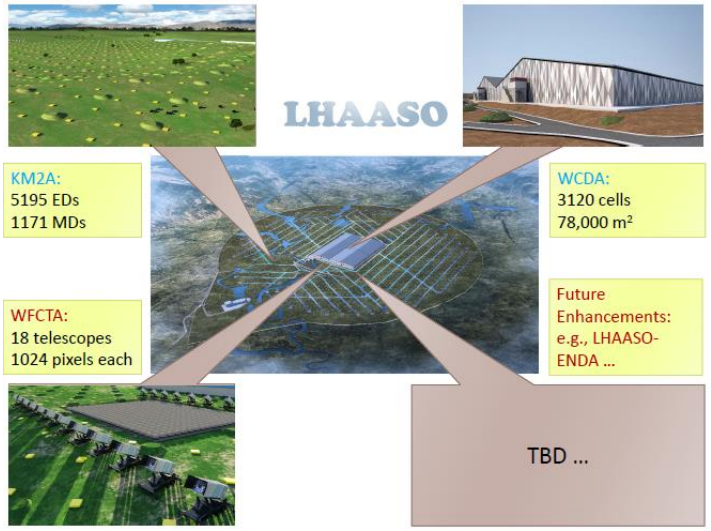
Yaodong Cheng, Lu Wang, Haibo Li, Yujiang Bi, Mengyao Qi
Institute of High Energy Physics, CAS
2020-02-04

Contents

- Requirements
- Current deployment
- Experience and Issues
- Next Plan
- Summary

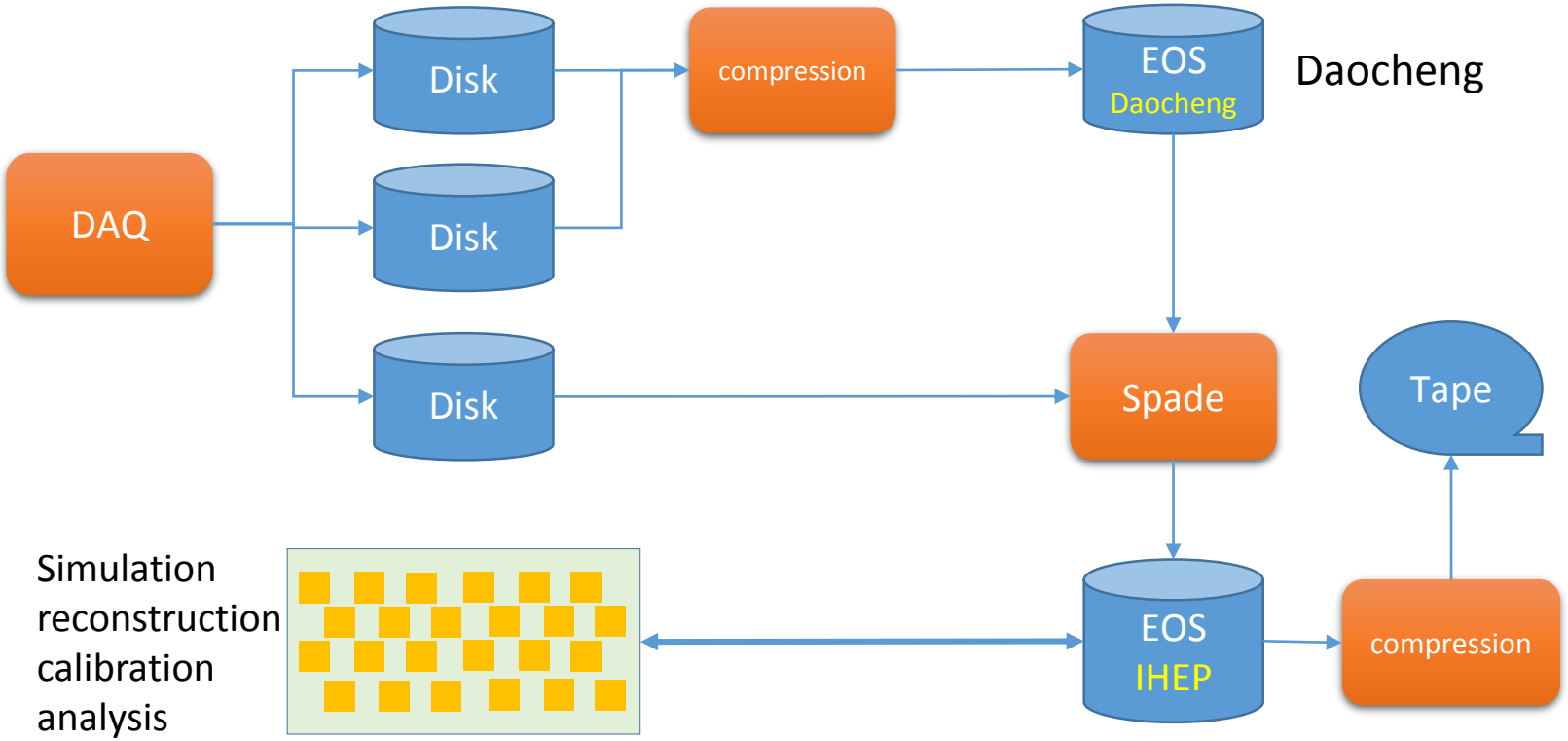
The LHAASO Project

- Large High Altitude Air Shower Observatory (LHAASO)
  - Located in Daocheng, Sichuan province (at the altitude of 4410 m)
  - Maintain multiple sites, including Daocheng cluster, Beijing cluster, and incoming Chengdu cluster
  - Network delay and stability of Internet private line network (400Mbps) between daocheng and Beijing
- Status
  - 6 PB of raw data each year
  - Planning to provide 20 PB+ capacity storage system
  - Start taking data in 2018
  - 1/2 construction completed in the end of 2019
  - Fully completed in 2020



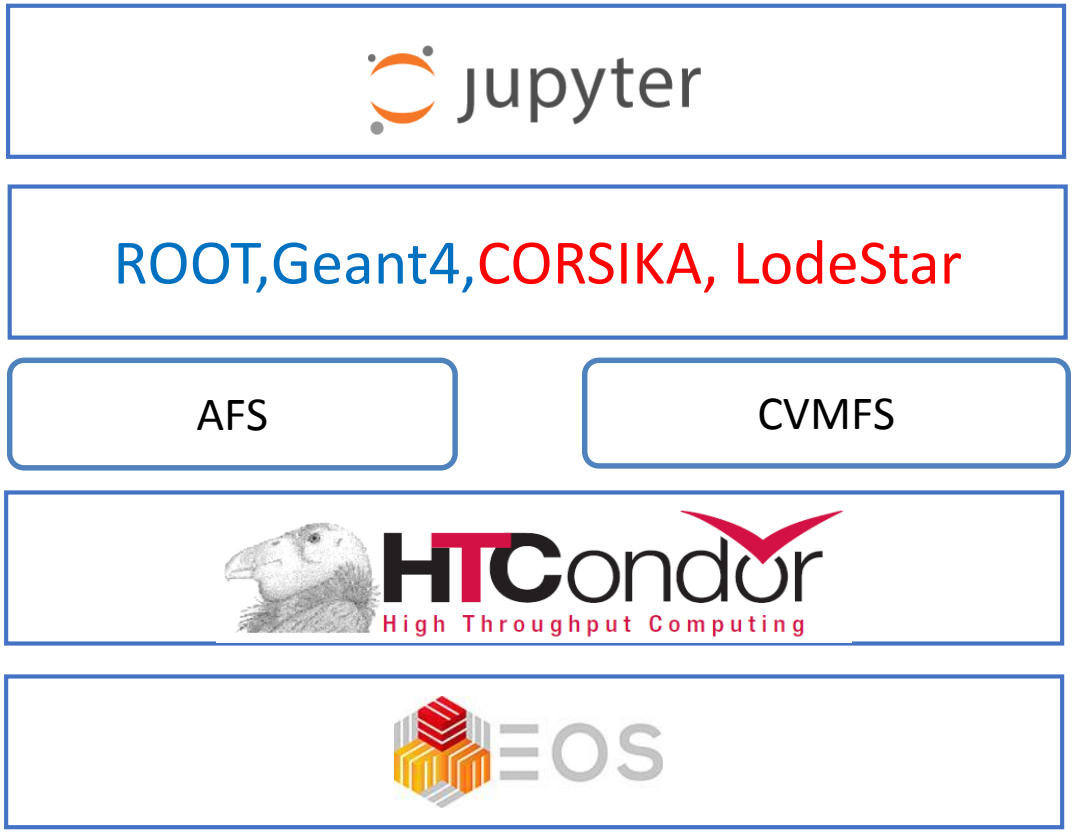
Challenges in LHAASO data processing

- Large amount of data, large number of files, complex data processing process
- Each step of processing depends on traversing the file and manually recording the status
- Any step error may cause data inconsistency
- Traversing files leads to high pressure on the storage system and low access efficiency



Current LHAASO computing platform

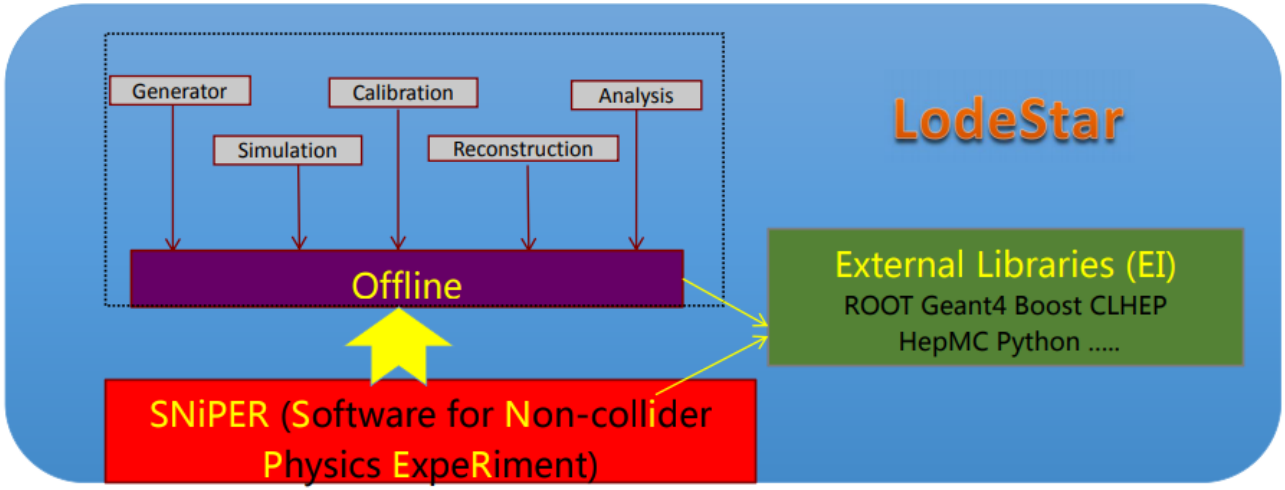
- Using EOS as main disk storage
- Using HTCondor as job management



Corsika: air shower simulation software
LodeStar: LHAASO Offline Software

Software

- CORSIKA
  - Air shower simulation software
  - Not based on ROOT framework, so not support xrootd access
- Lodestar
  - LHAASO Offline Data Processing Software Framework
  - Based on ROOT framework ,so support xrootd access



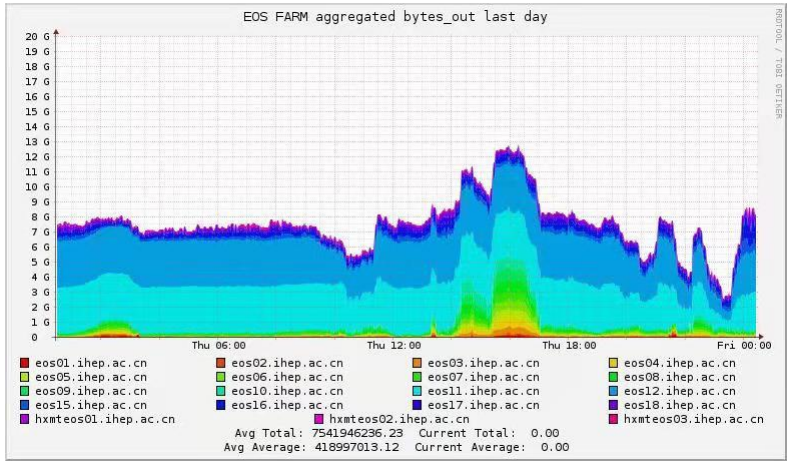
Hardware

- Before
  - RAID array
    - DELL MD3860f (60x8TB)
    - DELL ME4084 (84x12TB)
- Now
  - Begin to use high density JBOD array
    - DELL ME484 (84x12TB)

EOS deployment at IHEP

- 4 instances since 2016
  - LHAASO-Beijing
  - LHAASO-Daocheng (Remote)
  - HXMT (Hard X-ray Modulation Telescope)
  - IHEPBox
- LHAASO-Beijing instance
  - 2 space
    - No replica, 2.39 PB (RIAD array)
    - Replica, 4PB (new JBOD's 84X12TB)
- Characteristics
  - Mostly small files
  - Mostly fuse access, little XRootD access
- EOS Citrine release v4.4.23

Raw Capacity	**~8 PB**
Disk server	~30
Number of files	~166M
Number of directories	~55K
Peak throughput	>13GB/s



Issues with the current system

- Fuse
  - Poor read performance
  - 5%-10% job error rate
- Read performance limitation on RAID array
  - ME484 array performance is limited to 3GB/s
- Remote data access
  - High latency and bandwidth limitations

ROOT based software

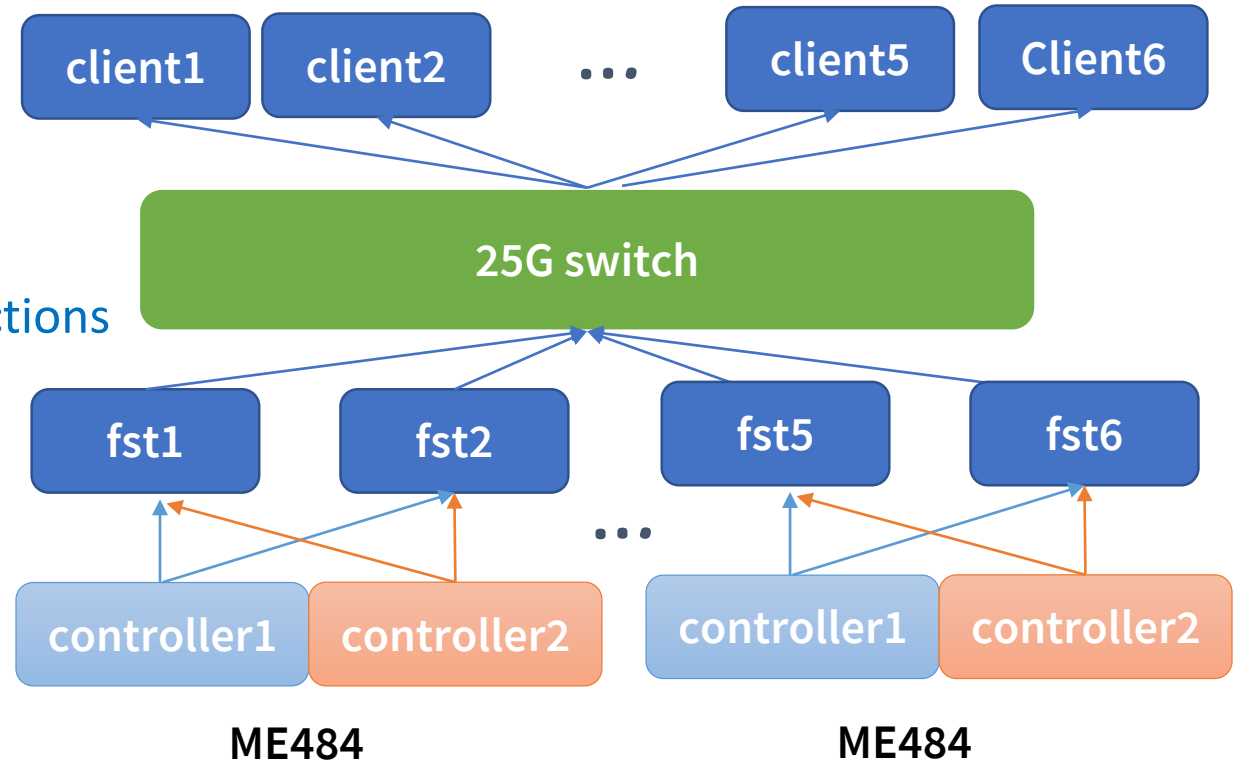
- The existing user program does not support the xrootd access and needs to be modified
  - Three methods to generate a TFile object
    - Declaration: TFile(PATHNAME); ✘
    - New method: new TFile (PATHNAME) ✘
    - Open method: TFile::Open(PATHNAME) ✔
- Results
  - ROOT5 has problems on writing data when using XRootD
  - ROOT6 in JUNO experiment runs well

Non ROOT software

- Use Lustre as home directory and scratch
  - Mainly used on Beijing cluster
- Copy data to local disk
  - Firstly copy the data to the local directory of the computing node with `eos cp` command, and then read it directly
  - Mainly used on daocheng cluster which only have AFS directory

Evaluation of new EOS+JBOD architecture

- EOS version: 4.5.6
- QuarkDB: 0.3.9
- 6 FSTs, 6 clients
- Network: all 25Gbps connections
- Each ME484 connected to 2 FSTs by multipath



New hardware

- JBOD: DELL PowerVault ME484
  - Disk: 84*12TB SAS
- server: DELL PowerEdge R740xd
  - CPU: Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz 12cores * 2EA
  - Memory: DDR4 32GB 2933MHz * 4EA
  - HBA: DELL PowerEdge 12Gbps SAS HBA firmware version (FW version: 16.17.00.03)
  - NIC: 25Gbps

DELL ME484 JBOD

5U84 drive expansion

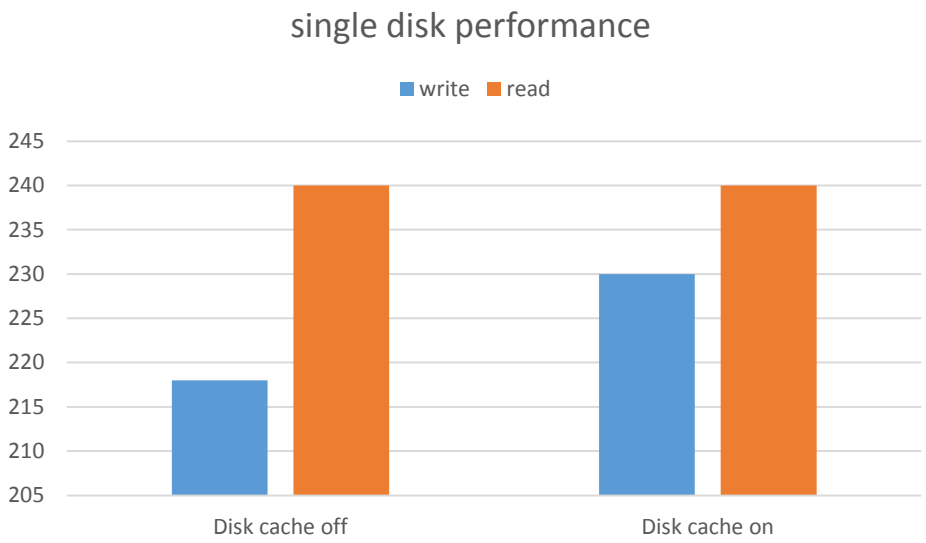
12TB *84, 4PB raw capactiy

Support for direct attach SAS

using 12Gb SAS HBA

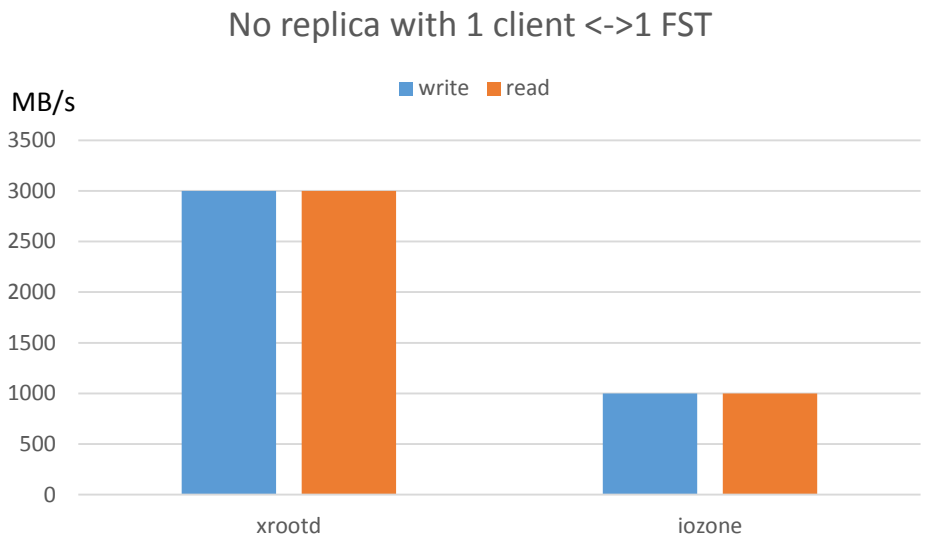
Local Performance

- **Sequential read and write**



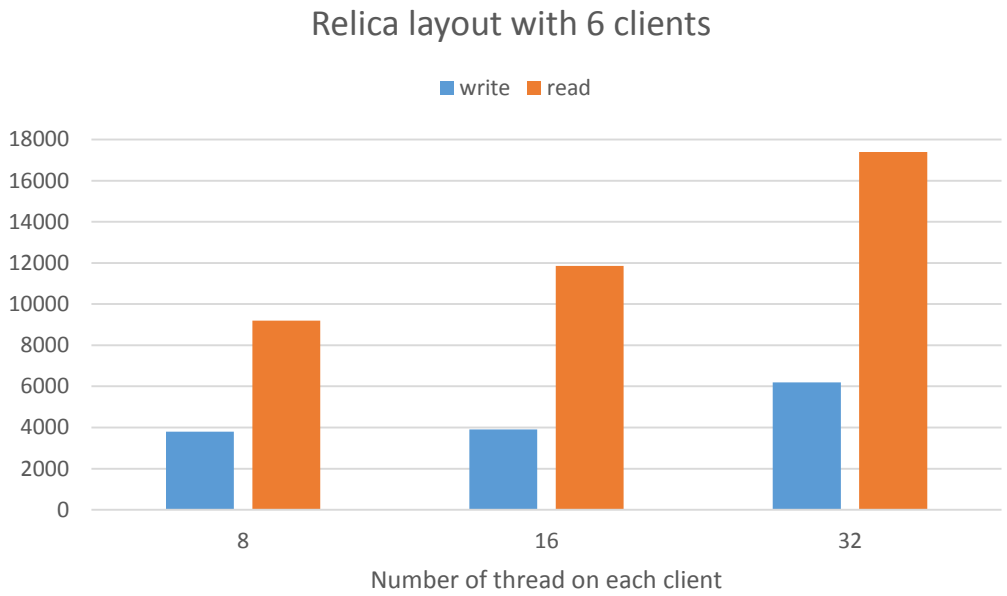
- The sequential reading and writing of single disk are above 200 MB/s
- The performance is up to 8GB/s for the whole array
- The performance of 42 disks is 7GB/s, which is higher than that of the HBA card (6GB/s)

No replica test



- The throughput can reach to 3GB/s when using xrootd
- One fuse client has a maximum performance of 1GB/s

Replica test



- The read performance could reach to 3 GB/s for each FST
- Maximum bandwidth can reach to 18 GB/s with 6 FST
- The write performance just reaches to 6GB/s, may be related to the replica write method of EOS?

Some issues

- Write performance on replica layout
  - Need as many FSTs as possible
  - For one FST configured with 42 disks, it is not very suitable?
- Throughput of 25G Network bottleneck
  - 25Gb cards bonding on FST
- Auto repair function for replica does not work well
- There are other issues such as the delay of replica synchronization

XRootD Proxy for Caching (XCache) deployment

- XCache@IHEP
  - Cache for remote storage (for LHAASO)
    - reduce latency & avoid WAN access on subsequent access
    - single server for caching data from EOS Daocheng to Beijing
    - XCache clusters with redirectors
  - Cache on node (in plan)
    - minimize data accessing latency
    - reduce EOS network traffic
    - improve job success rate.
  - XCache for operative institutions (in plan)
    - LHAASO
    - JUNO
- Gateway for Grid with EOS as backend (in progress)
  - Proxy for EOS TPC (EOS support TPC)

Next plan

- Add 10 PB raw capacity in 2020 for LHAASO
- JUNO experiments evaluation on EOS
  - JUNO will begin to generate data in 2021
- EOS + Kerberos deployment
- Plan to upgrade to new stable EOS version
- Migrate namespace to QuarkDB during summer maintenance(July 2020)

Summary

- LHAASO data process is a typical use case
  - large amount of data, large number of file, remote site maintenance
- Promote EOS to other experiments
- Larger JBOD, fewer FST and 25Gbps network brings new challenges
- Thanks for strong support from the CERN EOS team
- Need more support in future

Thanks for your attentions!

谢谢！