EOS Status at IHEP

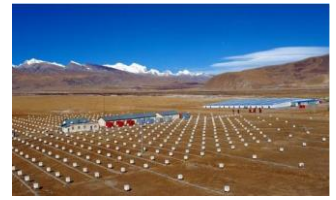
Haibo LI<lihaibo@ihep.ac.cn>
On behalf of the IHEP CC storage team

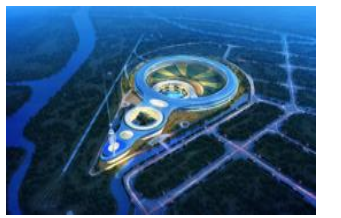
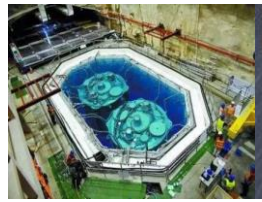
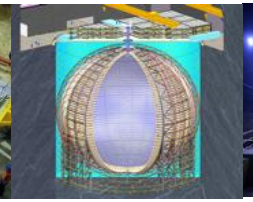
March 15, 2024

中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences
高能所計算中心
IHEP Computing Center

IHEP at a glance

- Institute of **H**igh **E**nergy **P**hysics (IHEP) is a premier research institute in China, dedicated to the study of particle physics, accelerator physics, and related fields.
- Data storage requirements for HEP experiments at IHEP
  - Beijing Electron Positron Collider BEPCII/BESIII
    - 1PB/year, with an accumulation of over 10PB
  - Jiangmen Neutrino Experiment
    - 2PB/year
  - Large High Altitude Air Shower Observatory (LHAASO)
    - 10PB/year
  - High Energy Synchrotron Radiation Light Source (HEPS)
    - 100PB/year
  - High-Energy Space Astronomy Experiments such as HXMT/HERD/eXTP/GeCAM, etc.
    - 10PB/year
  - WLCG: CMS, ATLAS, LHCb, BelleII, etc.


~550KM

~4400M

~-5M

-2500M
~-300M

Space astronomy satellite
(HXMT, HERD)

High Altitude Cosmic Ray Observatory
(LHAASO, YBJ)

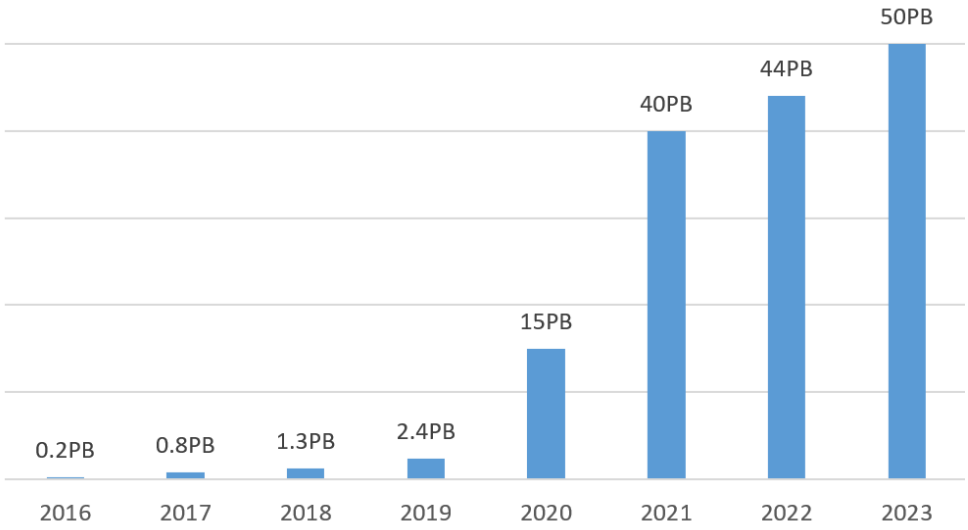
Particle collider
(BEPCII, HEPS,CSNS)

Underground Neutrino Experiment
(JUNO, Dayabay)

EOS deployment at IHEP

- EOS is one of the major storage systems at IHEP
  - Serves for LHAASO, JUNO, CTA and IHEPBox
- EOS in production since 2016
- 11 EOS instances
  - 6 instances for physics
  - 4 instances for CTA
  - 1 instance for IHEPBox
- EOS versions
  - MGM: 5.1.25, 5.2.X FST: 5.1.25, 5.2.X
  - Computing nodes: 5.1.25
- Replica 2
- Access via xrootd at batch nodes
- Fusex at login node

Raw Capacity	~ 50PB
Disk server	148
Number of fs	3906
Number of files	~ 700 Mil
Number of directories	~ 15 Mil
Peak throughput	>100 GB/s



EOS Hardware

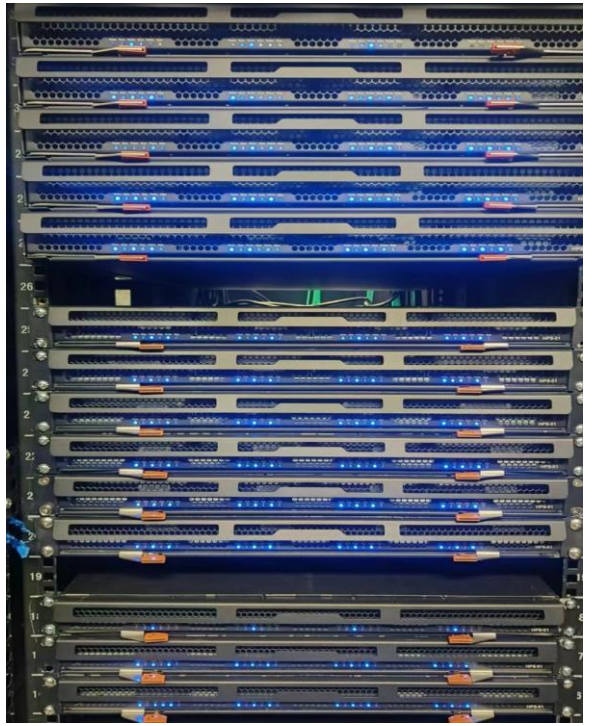
- RAID array
  - Disks per RAID: 84
  - Disk sizes:12TB, 20TB
  - Mainly used in Daocheng site, Sichuan province (at the altitude of 4410m)
- JBOD array
  - Disks per JBOD: 84
  - Disk sizes: 12TB, 16TB，20TB
  - Connected with two FSTs, configured with 25Gbs network
- FST servers
  - 16-40 cores, 128GB-256 GB RAM, 960GB SSD

Updates in 2023 – SE from DPM to EOS

- EOS SE replace DPM as WLCG T2 storage
  - One instance servers for ATLAS, CMS, BelleII
  - Capacity: 2 PB
  - Use RAID array
  - One replica
  - EOS version: 5.1.25
- EOS SE and CTA used for LHCb T1 storage
  - 1 instance for SE, 1 instance for CTA
  - SE Capacity: 6.69 PB
  - Use JBOD array
  - Replica 2
  - EOS version: 5.1.25
  - Ready, undergoing data challenge now

Updates in 2023 – ARM EOS in Production

- ARM EOS in production since Oct. 2023
  - 2 PB gross capacity, 310 TB used
  - OS: Ubuntu 18.04.6
  - Recompile EOS
  - EOS version: 5.1.27
  - CPU: 4*ARM@1.8G Neon
  - NIC: 10 Gbps
  - Disk: 12 disks * 16 TB



Updates in 2023 – Capacity Expansion and Upgrade

- Expansion of 6 PB of storage in 2023
- Upgrade EOS version to EOS5
  - 4.8.48 -> 5.1.25, 5.2.0, 5.2.16
  - Use eosxd instead of eosd
- Migrated from LevelDB to extended attributes
- Verified EOS on Alma Linux 9.3

Issues in 2023 – MGM Stuck

- LHAASO EOS instances have been experiencing MGM stuck since we update to EOS V5.1
  - 'eos ViewRWMutex' shows high latency，related to the EOS global namespace lock
  - Ask for help from CERN, and back to normal after upgrading to version 5.2.0
- Characteristics of LHAASO EOS instance
  - Currently 46 PB, 450 M files, 90% used
  - Many small files, like 1KB, 1MB, 10MB files
  - A large number of files are read and written within a short period of time
  - Frequent use of 'eos newfind' and 'eos stat' commands
- Some measures
  - Limit user threads: threads:* => 200
  - Optimize user jobs
- Current status
  - The occurrence of MGM unresponsiveness has significantly decreased, though they still occur occasionally
  - Generally caused by individual users, who can be identified using '**eos ipc:// ns**'
  - Restricting the access for the identified user

Issues in 2023 – Others

- Data deletion stress
  - Storage runs full for long periods of time and deleting data is a routine task
  - When user use the 'eos rm -rf' command to delete a directory, it may fail with timeout
  - ''eos rm'' on administrator's side will cause MGM to be unresponsive
  - Currently we use 'eos newfind -f' to generate a list of files first, and then delete them one by one in the background, which takes a long time
- Fusex client crashed
  - Fusex client sometimes generates coredump, the /root/core.* file needs to be cleaned regularly
  - Sometimes the Fusex cache is not cleared automatically, resulting in full space and crashed
  - Sometimes files are not synchronized
  - EOS client version: 5.1.25, we will update it to 5.2.19 later

Plans in 2024 – New EOS instance for LHAASO

- Add new instances for LHAASO to reduce the pressure on MGM
- Be transparent to experiments and provide unified MGM endpoint
  - Use ‘eos route link‘ to federate these two instance
- Functionality has been verified to work
- New instances and old instances are distinguished by different directories

Plans in 2024 - Others

- Capacity expansion 6PB of gross space
- Migrate to Alma Linux 9.3 by June 30 of this year
- Construction of T2 EOS SE sites for other universities in China
- Update EOS version to the latest 5.2.x

Summary

- EOS SE Replaces DPM as WLCG SE at IHEP
- ARM EOS is currently in production and running stable
- As the data increases, the EOS systems face more and more demands and challenges, and we need do more efforts
- EOS provides good storage guarantee for LHAASO experimental data processing and ensures the output of LHAASO scientific results

Thanks for great support from the CERN EOS team!

Thanks for your attentions!

谢谢！