EOS at IHEP

Yaodong Cheng, Lu Wang, Haibo Li, Yujiang Bi, Qiuling Yao, Yaosong Cheng, Minxing Zhang

Institute of High Energy Physics, CAS

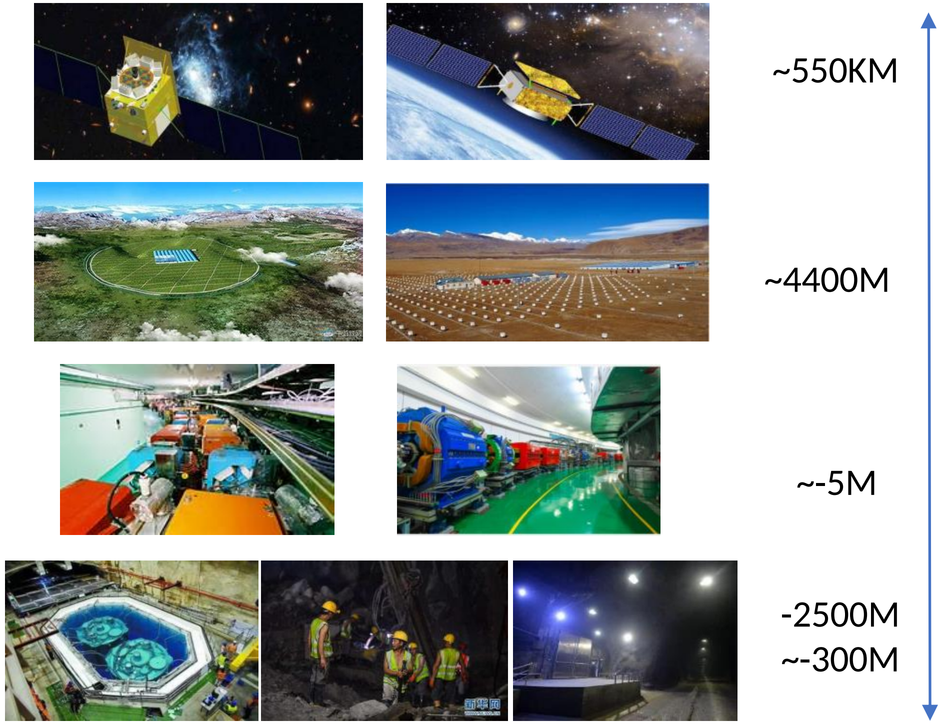
2020-03

Outline

- IHEP and IHEP CC
- EOS deployment status at IHEP
- Updates in 2020
- Some issues
- Wisher for the future of EOS
- Next plan
- Summary

IHEP at a Glance

- Institute of High Energy Physics, Chinese Academy of Sciences
- ~1500 staffs, ~1200 scientists and engineers
- The largest fundamental research center in China with following research fields
  - Experimental Particle Physics
  - Theoretical Particle Physics
  - Astrophysics and cosmic-rays
  - Accelerator Technology and applications
  - Synchrotron radiation and applications
  - Nuclear analysis technique
  - Computing and Network applications
  - …



~550KM

Space astronomy satellite

（HXMT, Gecam）

~4400M

High Altitude Cosmic Ray Observatory

（LHAASO, YBJ）

~-5M

Particle collider

（BEPCII, CSNS, HEPS）

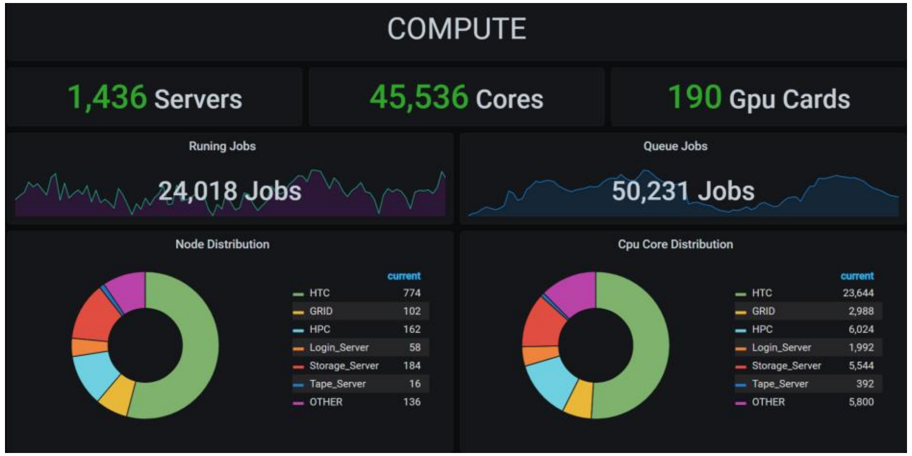
-2500M
~-300M

Underground Neutrino Experiment

（JUNO, dayabay）

IHEP CC

- IHEP Computing Center provides data services and informatization support for experiments in IHEP, and is responsible for the construction, operation and maintenance of large-scale high-performance scientific computing and network environments
- 45,536 cpu cores, 190 GPU cards for more than 10 experiments
  - HTCondor cluster runs for HTC jobs
  - Slurm cluster runs for HPC jobs
  - WLCG tier 2 site
- 45.6 PB storage
  - Lustre and EOS are two main file systems
  - Castor for tape storage
- Network
  - IP V4/ IP V6 dual stack
  - Ethernet(100Gb) / IB (100Gb) supported
  - LHCOne joint



COMPUTE

1,436 Servers | 45,536 Cores | 190 Gpu Cards

Runing Jobs: 24,018 Jobs

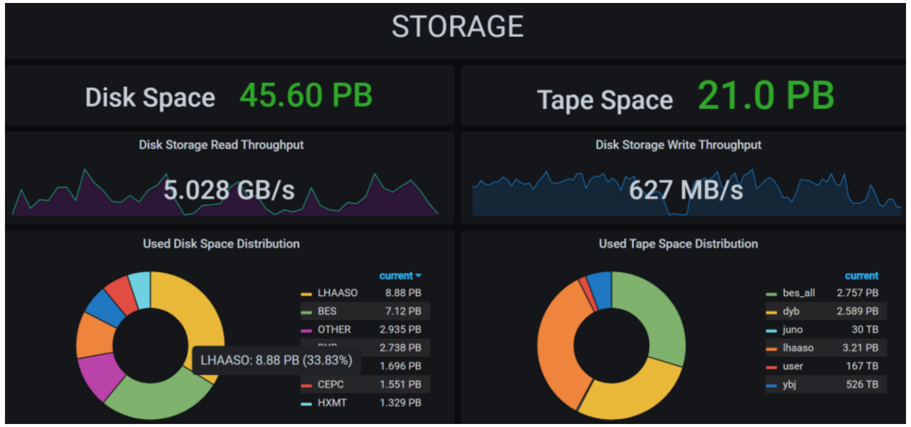
Queue Jobs: 50,231 Jobs

Node Distribution

	current
HTC	774
GRID	102
HPC	162
Login_Server	58
Storage_Server	184
Tape_Server	16
OTHER	136

Cpu Core Distribution

	current
HTC	23,644
GRID	2,988
HPC	6,024
Login_Server	1,992
Storage_Server	5,544
Tape_Server	392
OTHER	5,800



STORAGE

Disk Space 45.60 PB

Tape Space 21.0 PB

Disk Storage Read Throughput: 5.028 GB/s

Disk Storage Write Throughput: 627 MB/s

Used Disk Space Distribution

	current ▾
LHAASO	8.88 PB
BES	7.12 PB
OTHER	2.935 PB
[illegible]	2.738 PB
[illegible]	1.696 PB
CEPC	1.551 PB
HXMT	1.329 PB

LHAASO: 8.88 PB (33.83%)

Used Tape Space Distribution

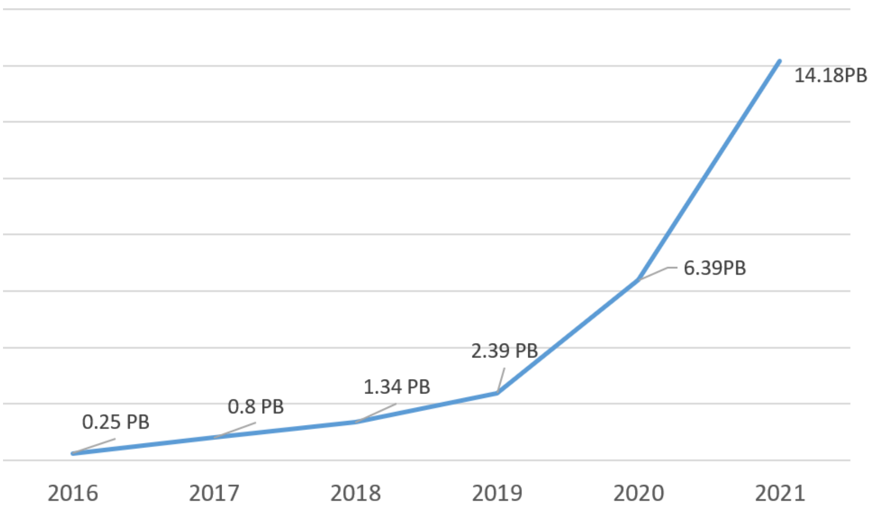
	current
bes_all	2.757 PB
dyb	2.589 PB
juno	30 TB
lhaaso	3.21 PB
user	167 TB
ybj	526 TB

EOS deployment at IHEP

- 4 instances since 2016

Instance	Capacity	Number of fs	Number of Files	Number of directories
LHAASO-Beijing	14.5PB	1048	155 Mil	20 Mil
LHAASO-Daocheng	2.5PB	78	44 Mil	4 Mil
HXMT	806TB	13	40 Mil	2 Mil
IHEPBox	200TB	5	30 Mil	2.5 Mil

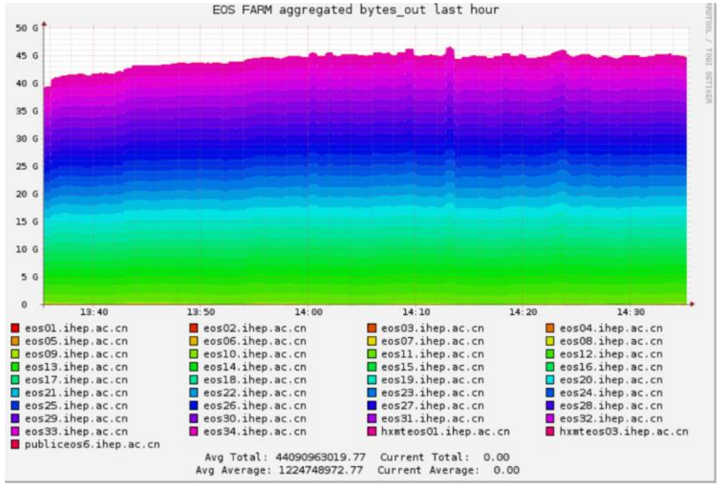
Raw Capacity	~ 18 PB+26 PB incoming
Disk server	~49
Number of fs	1144
Number of files	~266 Mil
Number of directories	~10 Mil
Peak throughput	>50 GB/s



A typical instance at IHEP

- LHAASO-Beijing instance
  - Largest instance at IHEP
  - 14.5 PB capacity, 43 fsts, 1048 fs
  - Configured with two space

Space	Usage	Quota	Capacity	Hardware	Layout
/eos/user	user data	1TB/user	2.5PB	RAID Array	plain
/eos/lhaaso	experiment data	No limit	12PB	JBOD	2 copies



- How user use it?
  - Provide an additional Lustre mount point for users to submit jobs
  - Users must use xrootd to access the EOS data at worker node
  - For CORSIKA simulation jobs that do not support ROOT, the lustre mount point is used instead of EOS, and the data is migrated to EOS after job finished

Hardware changes

- Before 2019
  - RAID array
    - DELL MD3860f (60x8TB)
    - DELL ME4084 (84x12TB）
- Since 2019
  - all new storage devices use high density JBOD array
  - DELL ME484 (84x12TB)
  - Each ME484 connected with two FSTs, configured with 25Gbs network

Updates in 2020– eos upgrade

- 4.4.23->4.7.7, Migrate namespace to QuarkDB during summer maintenance(July 2020)
  - MGM boot time is greatly reduced (1s)
  - The number of files stored in a single instance exceeds 150 million
- 4.7.7->4.8.31, December 2020
  - MGM crash decreased
  - The format of the returned results of some eos commands has changed, correspondingly, the job script will also change accordingly
    - Such as ‘eos newfind’ command doesnot show “path=” prefix

Updates in 2020– Storage expansion

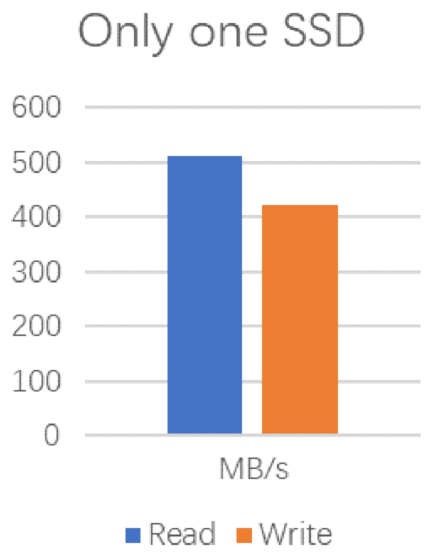
- LHAASO-Beijing instance increases 8PB, now 14.5PB
- LHAASO-Daocheng instance increases 1PB, now 2.5PB

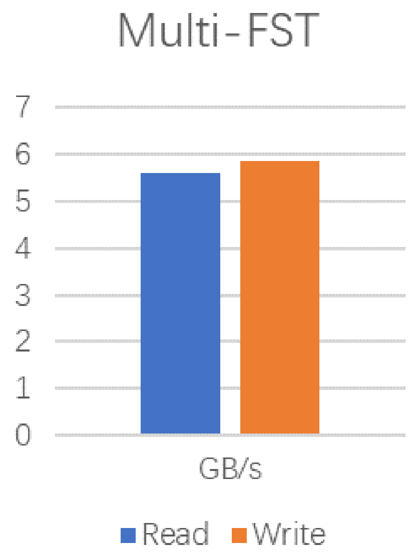
Updates in 2020– from fuse to xrootd

- Spend a lot of time training users to submit jobs using xrootd
- Encourage users use xrootd instead of fuse in jobs
    - The compute node unmounts the /eos mount point
    - The login nodes retains /eos mount point
- Currently LHAASO users have accepted the xrootd method, and the stability of operations has been improved a lot
- However, text files and non-supported ROOT programs have poor performance in xrootd access

Updates in 2020– EOS on ARM

- To evaluate EOS on ARM
- Ported EOS to ARM
  - Inline assembly: redefined to aarch64 assembly functions
  - EOS & dependent packages: dynamically compiled and linked
- Performance evaluation

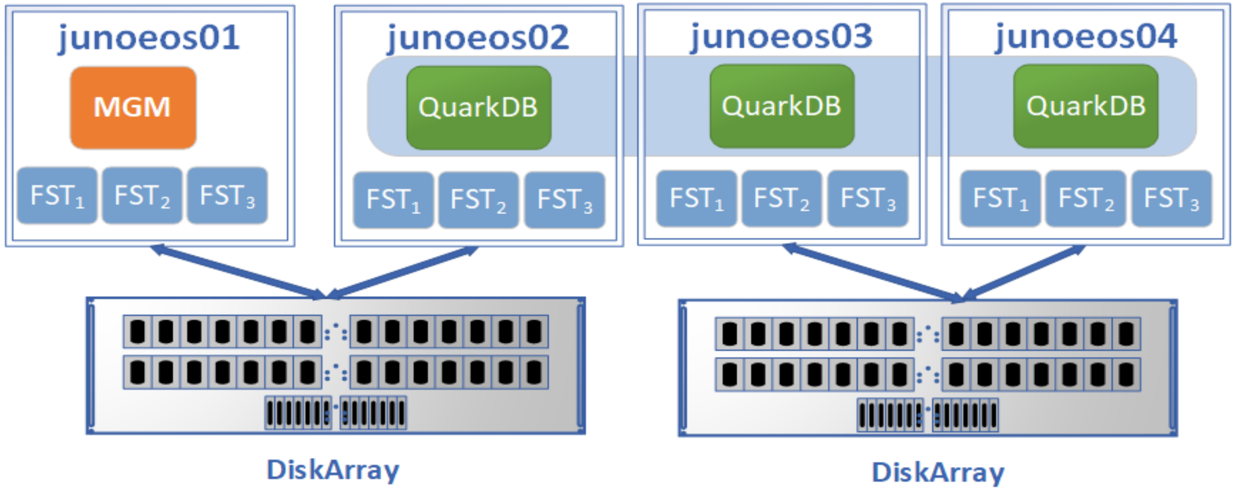




Testing environment
CPU: HiSilicon Kunpeng 920-6426 CPU @ 2600MHz 64cores * 2EA
NIC: 25 Gbps
SSD: SATA 3.2, 6.0 Gb/s
Operating System: CentOS Linux release 7.6.1810
EOS version: 4.7.7

Updates in 2020– EOS for JUNO Testbed

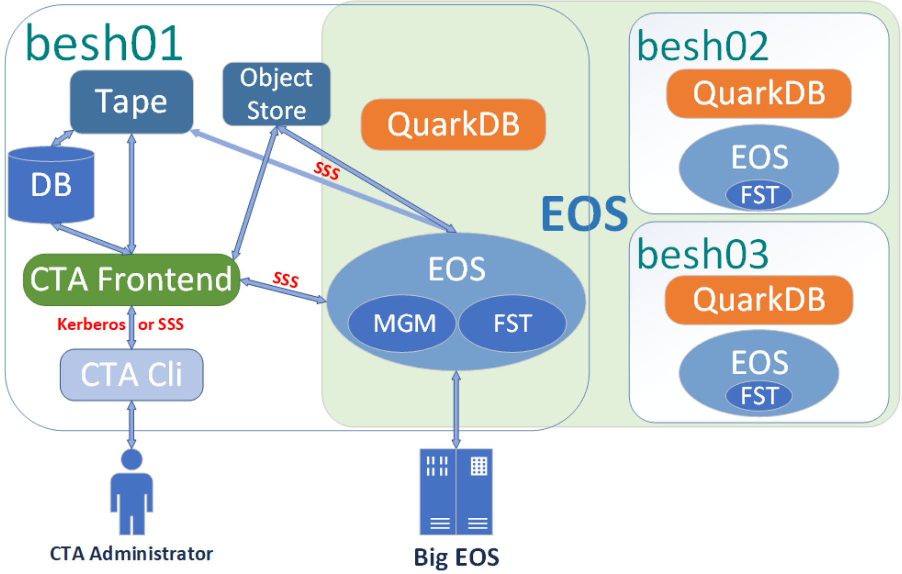
- EOS evaluation for JUNO(ready in 2022)
  - EOS I/O System Performance
  - Application Performance with EOS
  - To replace Lustre and as main filesystem?
- JUNO EOS testbed status
  - Ready for user test since 2020/08
  - 980TB in total, 732TB used, 1.2M files
- Hardware
  - 4 nodes, 2 JBOD disk arrays
- Software
  - EOS: 4.8.25, QuarkDB: 0.4.2
  - 1 MGM, 4 FST nodes, 84 FSs in total, 3 QuarkDB nodes



- Usage
  - Pure XRootD protocol and EOS command to access EOS
  - No fuse or fusex mountpoint /eos on login/worker nodes

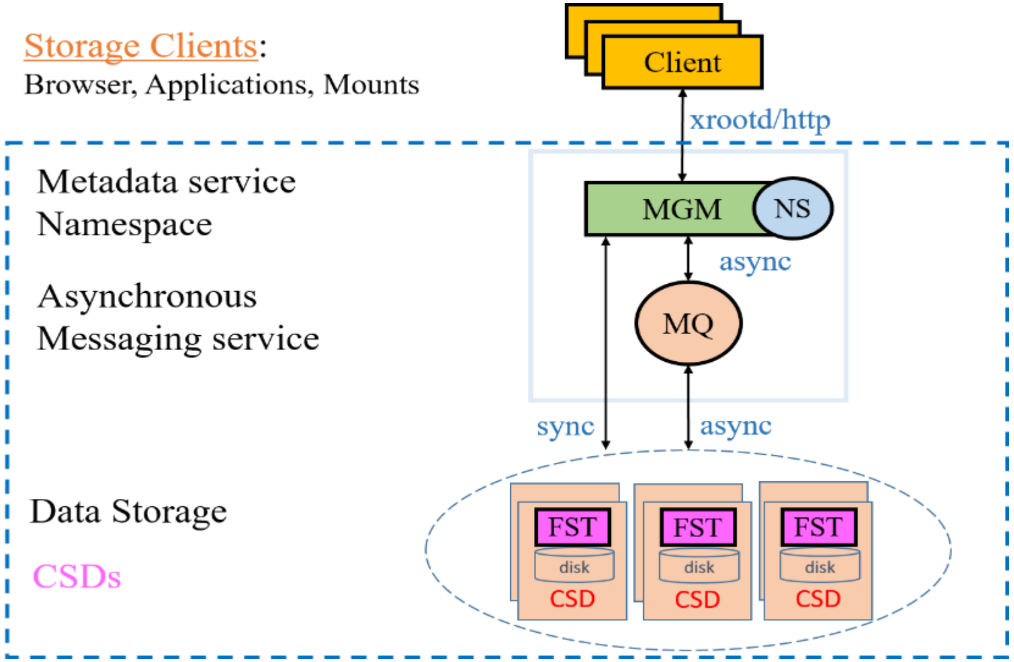
Updates in 2020– EOS CTA

- CTA
  - The tape back-end of EOS
  - Evaluation for JUNO and others
- Testbed Setup
  - Hardware
    - 3 nodes
    - Virtual tape Library
  - Software
    - CTA : 4.0
    - EOS : 4.8.37
    - QuarkDB: 0.4.3
    - PostgreSQL: 9.6
    - mhVTL: 1.5.3
- Status
  - EOS&CTA Ready for test
- Todo
  - Kerberos authentication
  - Ceph as ObjectStore
  - FTS as transfer system?
  - Migration from Castor to CTA?
- Plan
  - Ready for production deployment for JUNO this year



Updates in 2020– EOS computational storage

- Computational storage lowers part of the computing power to the data storage part, reducing data movement and improving computing efficiency
- Application areas
  - Mainly suitable for scenarios where large amounts of data are read in and out, and low CPU consumption, such as decode, reconstruct, etc
- Current status
  - Developed a demo system, which can realize computable storage services by adding xrootd plug-ins to EOS



More details refer to *minxing zhang's presentation*

Some issues

- MGM crash
  - Caused by the operation of "eos newfind *subdirectories in linked directories*", occurred in 4.7.7, resolved in 4.8.31
  - Caused by unknown reason still occurs in 4.8.31
- Unable to store file, file incomplete or file lost
- Client disconnected from FST because of connection timeout
  - Adding XRD_STREAMTIMEOUT=600 into job scripts
  - EOS > 4.8.31 solved this problem in server side
    - Setting EOS_FST_ASYNC_CLOSE=1 in EOS sysconfig file
- Memory occupied by the MGM process is still very large by using QuarkDB
  - ~60GB for an instance with 150 million files

Wishes for the future of EOS

- Fuse/fusex is more mature
- Erasure code detection and automatic recovery
- EOS operation and maintenance experience and document sharing

Next Plan

- Capacity extension in 2021 for LHAASO
  - 24 PB for LHAASO-Beijing, 2PB for LHAASO-Daocheng
- JUNO EOS SE deployment
- EOS + Kerberos deployment
- EOS CTA Migration
- Promote EOS computational storage service to HEP experiment, such as LHAASO decode

Summary

- EOS storage capacity is growing rapidly, and has now become the main storage system for the LHAASO experiment
- The JUNO experiment will use EOS as its main storage system
- We will further explore the new possibilities of EOS and contribute to the development of EOS community
- As the scale increases, the pressure on operation and maintenance increases
- Thanks for support from the CERN EOS team

Thanks for your attentions!

谢谢！