EOS and CTA Status at IHEP

Yujiang Bi

On Behave of IHEPCC Storage Team

EOS Workshop 2022

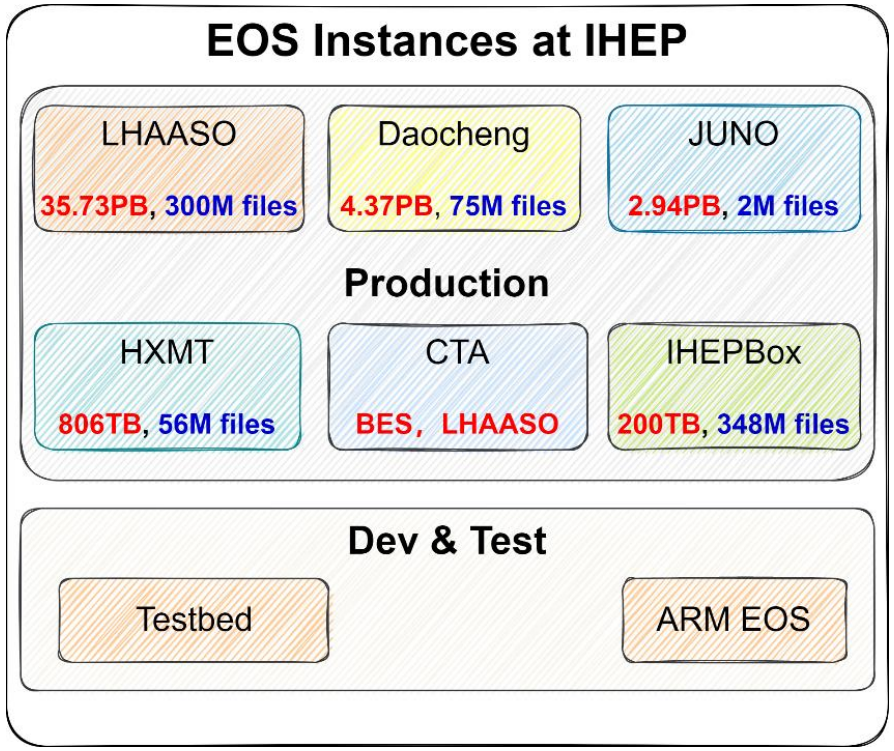
3/9/2022

Outline

- Storage Overview
  - EOS Storage
  - Tape Storage
- CTA Practice
  - Evaluation
  - Production
  - Migration
  - Problems
- CTA Roadmap
- Summary

EOS Storage

- One of the major storage systems at IHEP
  - Serves for LHAASO,HXMT,JUNO,CTA and IHEPBox
  - Version history: 0.3 -> 4.2 -> 4.7 -> 4.8
  - RAID Array (before 2019) and JBOD Array(after 2019)
- Storage element
  - **JUNO EOS** acts as DIRAC SE, providing DIRAC service
  - GSI auth for external access only and unix/sss/krb5 for local
- ARM EOS cluster
  - ARM machines are cheaper than X86 usually
  - Test on Kunpeng 920 ARM Chips (48C/64C)
- Computational storage services(CSS)
  - Based on customized ARM devices (with FPGA)
  - Extending EOS with callable functions on MGM/FTS
    - TFile::Open("root://eos_url//eos/aaa.root&**css_app=decode**", O_RDONLY)



Tape Storage

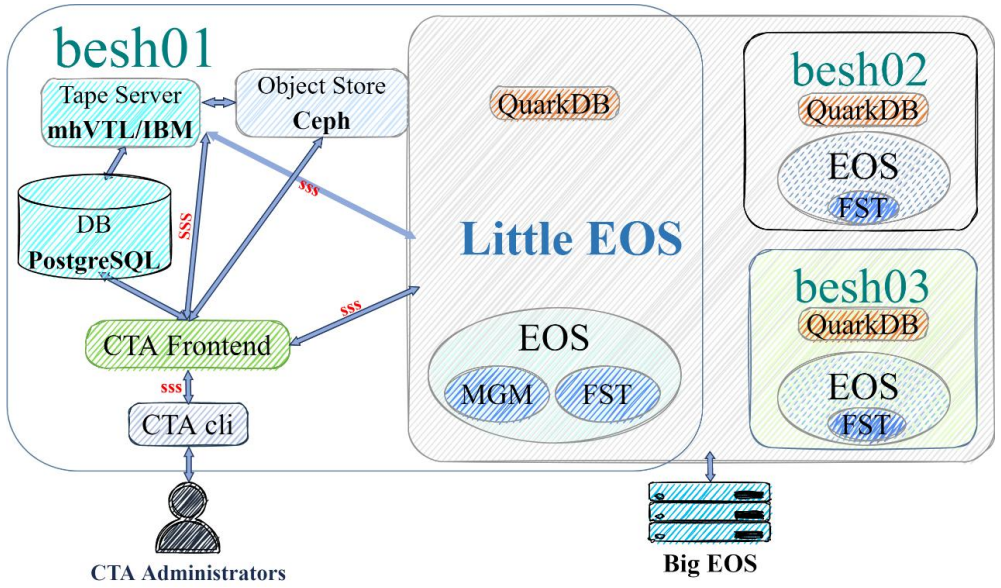
- Castor has been in service for over 20 years at IHEP
  - Modified Castor v1.7 instead of v2
  - Serves for BESIII, LHAASO and HXMT and other experiments
- CTA will manage all libraries after late 2022
- 3 Physical tape libraries and 1 under construction
  - BESIII, DYB: LTO4 & LTO 7, IBM TS3500
    - 24 drives, 5000 tapes, 22 PB
  - LHAASO, HXMT: LTO7, IBM TS4500
    - 20 drives, 4200 tapes, 25 PB
  - YBJ: LTO4, IBM TS3500
    - 4 drives, 1100 tapes, 880 TB
  - JUNO: LTO9 (this year)
    - 20 drives, ~4000 tapes, ~36 PB

Castor Situation

- Castor works well but is outdated
  - Low performance for massive LHAASO data archive
    - Stager get staled when too many files (~50K) in queue waiting for migration
    - No migration scheduling mechanism to optimizing retrieval requests
  - Operation and maintenance is complicated and not friendly
  - Not supported anymore long ago
- CTA is ready for production
  - All experiments at CERN managed by CTA
  - Many other institutes like AARNet, RAL, Fermilab
- Time to switch to CTA?
  - Understand rules before playing with CTA

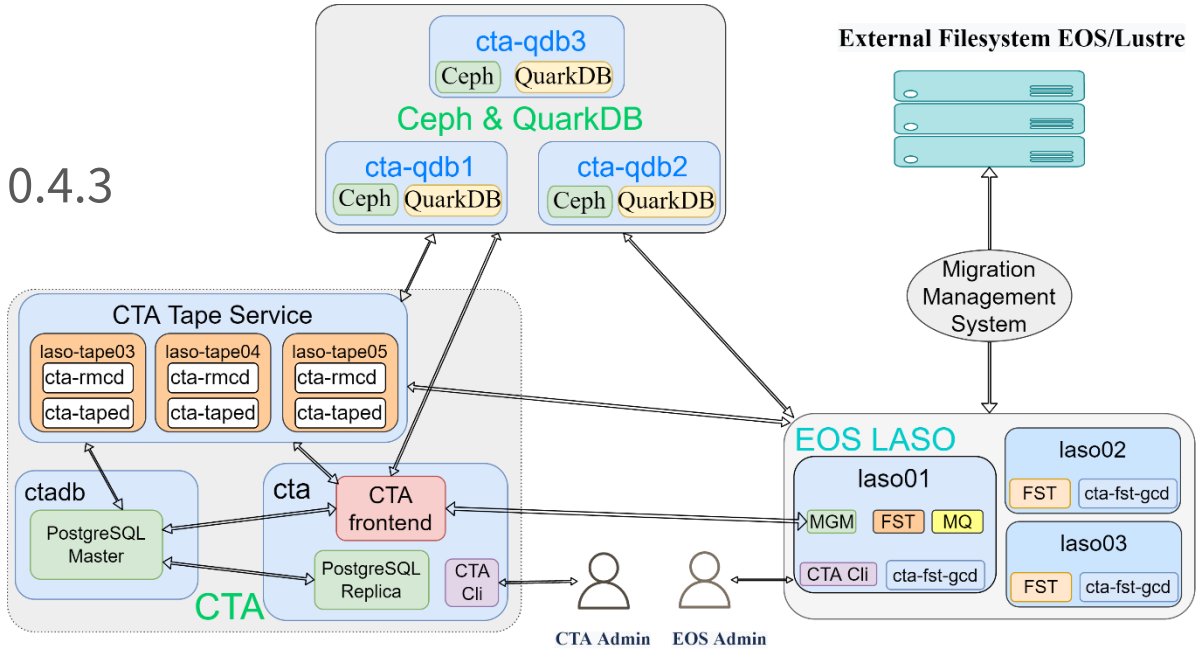
CTA Evaluation

- Testbed setup
  - Hardware
    - Tape Libraries: mhVTL and IBM TS2900
    - 3 nodes for EOS, QuarkDB and Ceph
    - 1 node for CTA frontend, catalogue and tape server
  - Software
    - CTA : 3.1-14 & 4.0-1， EOS : 4.8.34 & 4.8.40
- Evaluation
  - Building and deploying CTA from scratch
  - CTA admin and tape operations
  - Data archival and retrieval workflow
- All items passed except Kerberos auth(06/2021)
  - Ready for production use?



CTA in Production

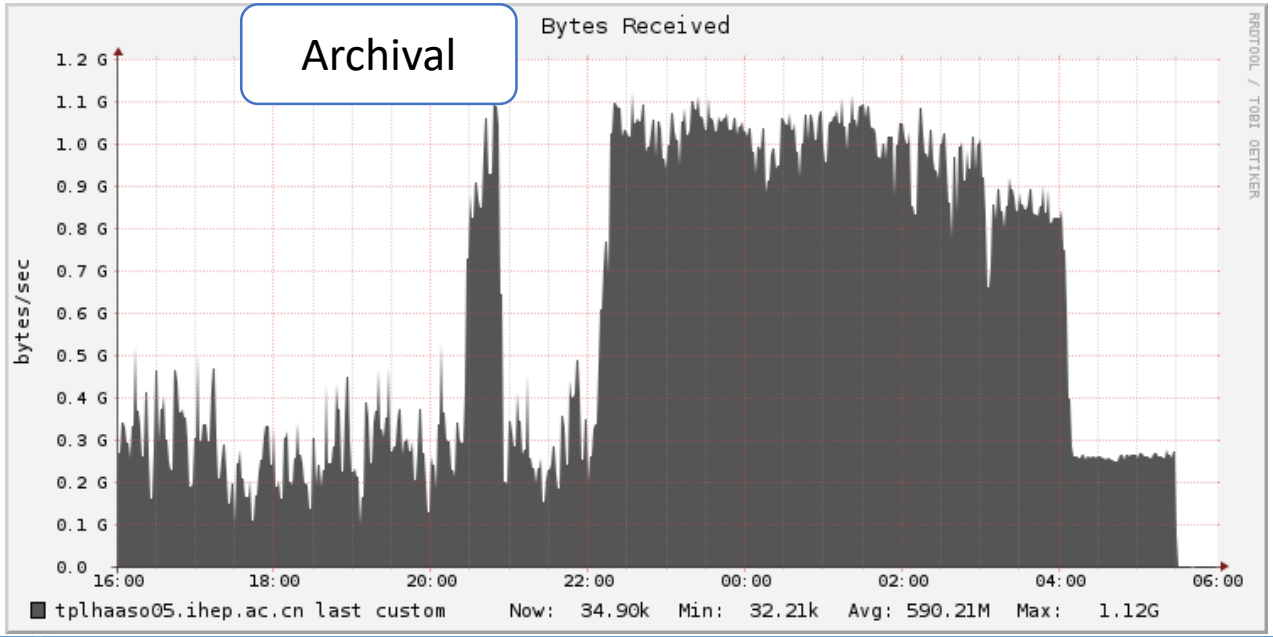
- Hardware
  - 1 CTA frontend node and 1 catalogue: 2x1TB SATA SSDs
  - 5 Tape Servers for BES & LHAASO
    - Rest reserved for Castor
  - 3 Ceph & QuarkDB nodes with 8x1TB SATA SSDs
  - 3 EOS storage nodes : 12x12TB HDDs
    - **No SSDs** ☺
  - Network connection: 10G/25G Fiber
- Software
  - CTA 4.2.1, Ceph 14.2.20, EOS 4.8.45, QuarkDB 0.4.3
- Architecture
  - Two little EOS instances
    - BESIII & DYB & Public
    - LHAASO & HXMT & YBJ
  - A central CTA frontend and PostgreSQL
  - SSS auth between CTA frontend and Admins

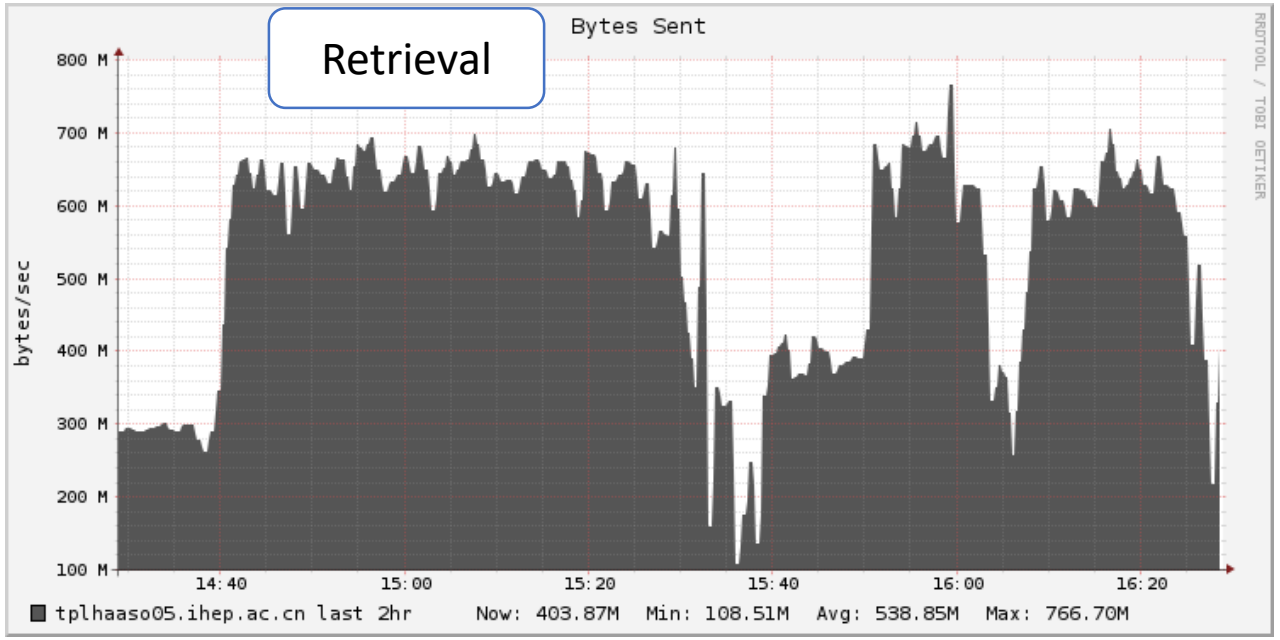


Stress Test in Production Environment

- Archiving 100K ~1GiB Files
  - Archiving from 2 nodes with 20 threads
  - ~ 80K files maximum in queue
- Retrieving these 100K Files
  - Disk replicas cleared beforehand
  - Preparing all files in one request

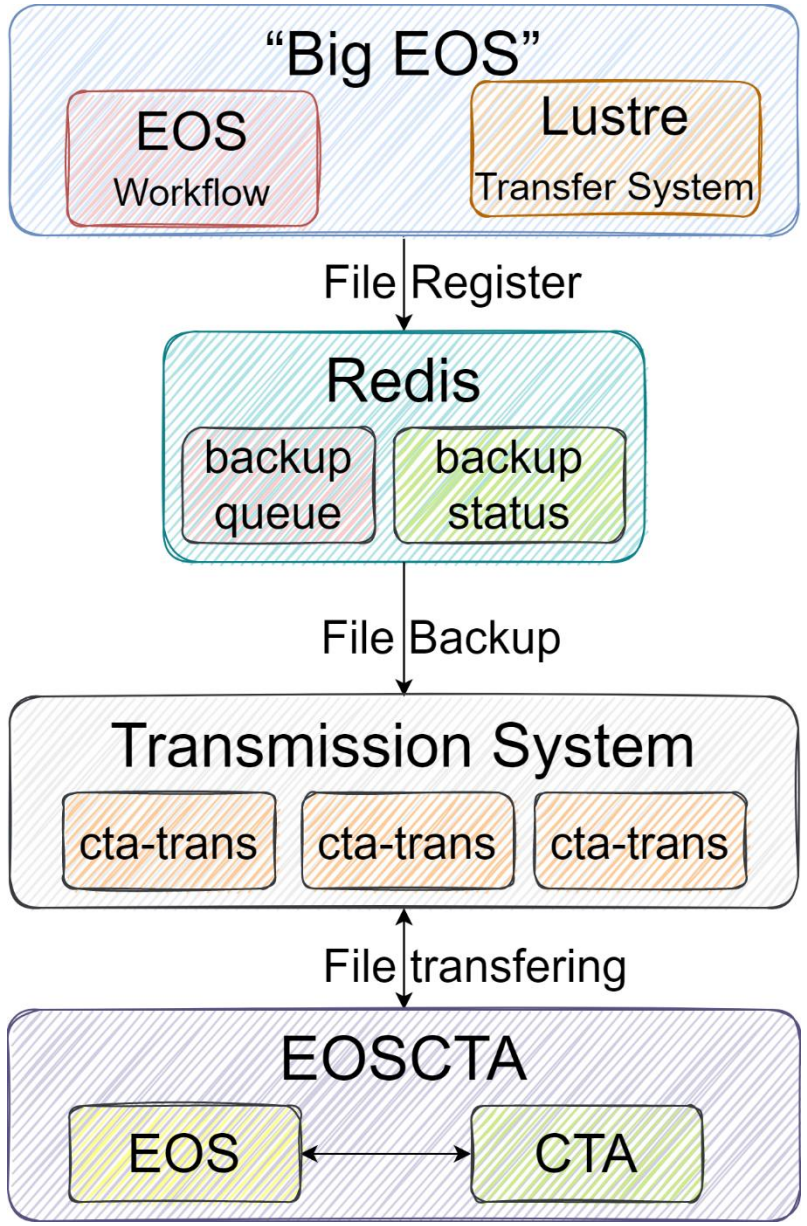
- All 100K files archived and then retrieved successfully
- Performance of archival and retrieval are quite good





CTA Transmission System (CTS)

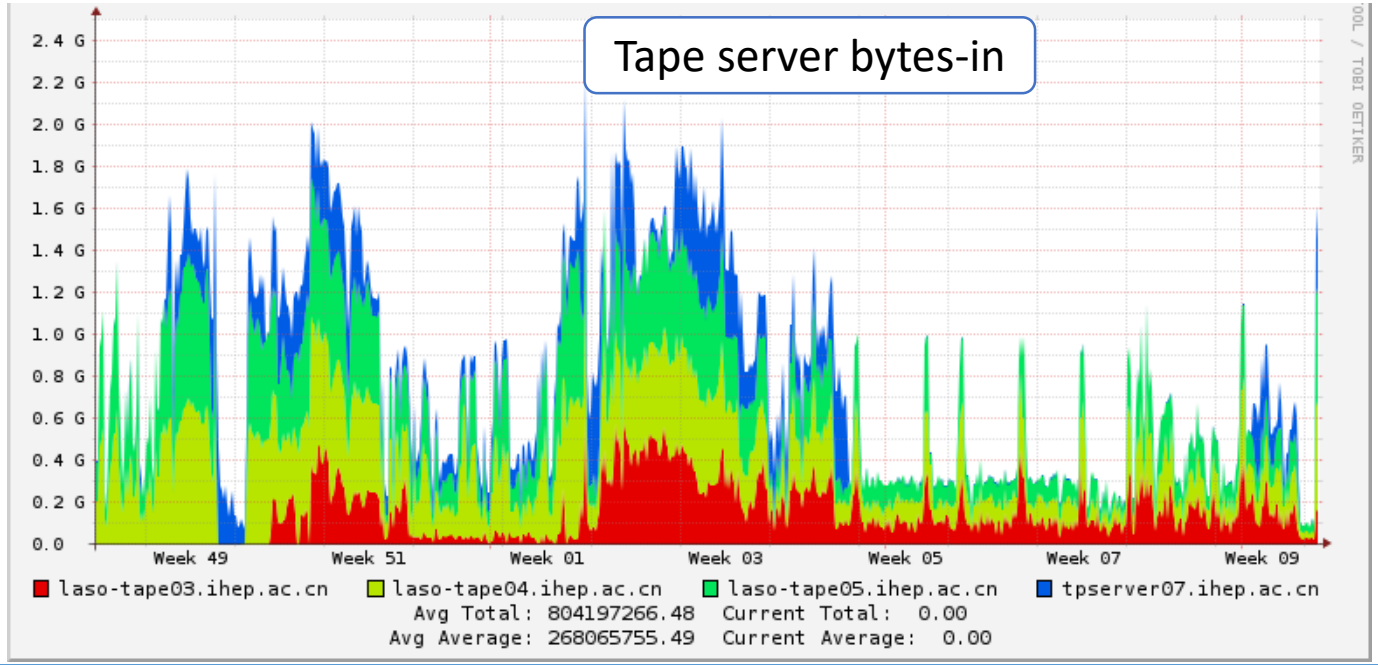
- Files needs to be archived to CTA with short delay
  - Files created all the time and recreated occasionally
  - Files should be archived automatically and safely
- CTS - a simple but robust transmission system
  - EOS WFE + Redis + Shell Scripts
- EOS WFE is great
  - Register file into Redis automatically
  - Submit file processing job to job scheduler
- Just shell scripts
  - Transferring files in parallel
- Works well but not good enough
  - Just archiving files from “big EOS” to CTA
  - New transmission system like FTS needed



CTA Current Status

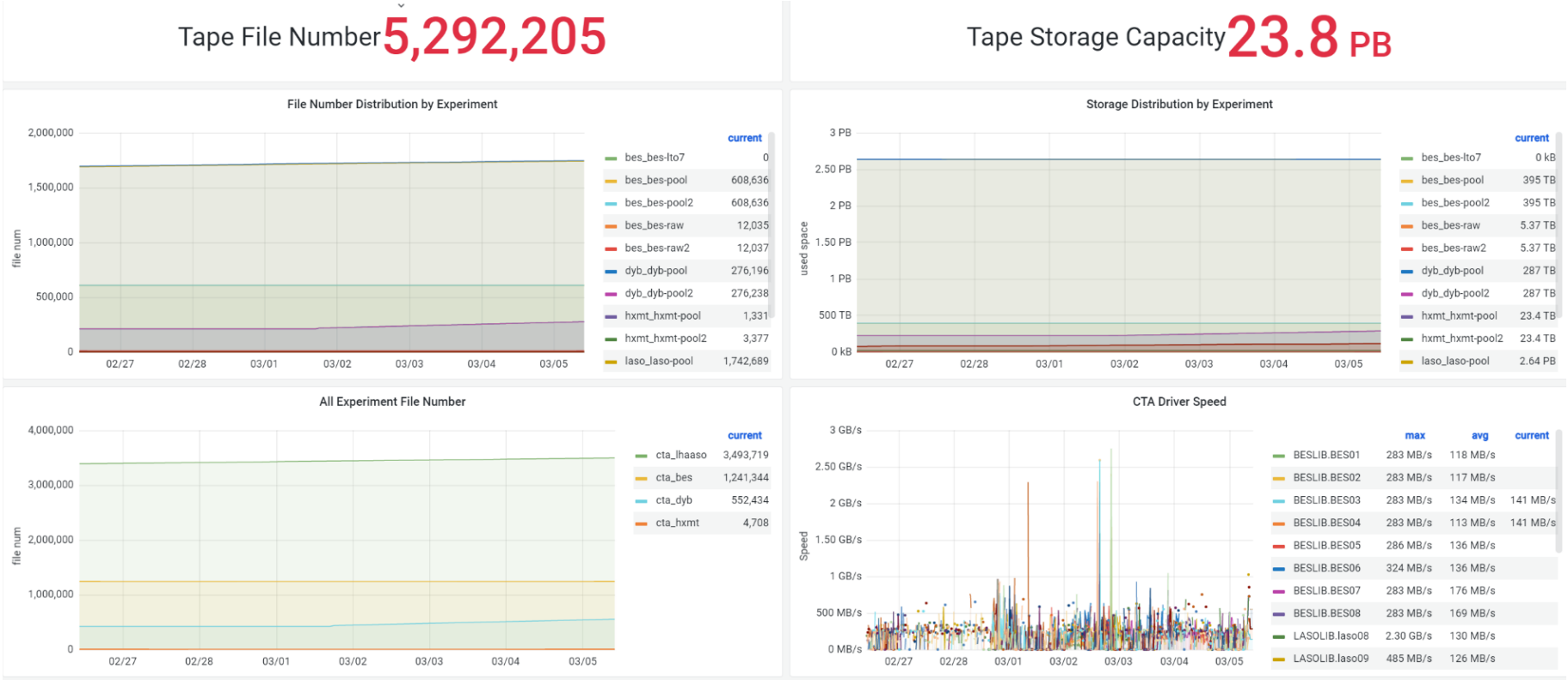
- All Archive Service Switched to CTA
  - BESIII still use castor due to DAQ technics, and will switch to CTA this year
- Works Well since the beginning (10/2021)
  - No big faults except operation errors
- Backup overview
- Performance during last 3 months

Experiments	LHAASO	HXMT	YBJ	BESIII	DYB
Used/Capacity	2.4P/6.2P	22T/30T	113T/600T	370T/420T	288T/600T
Files	1.7M	3K	2K	600K	300K
Drives	12 LTO7			8 LTO7	



Service Monitoring

- Important to ensure CTA services stable
  - Backup status presentation and rapid response to failover
- Grafana dashboard to display service status
  - Storage Capacity, total files archived, drive speed, failed request ……
- Nagios monitoring scripts combined with WeChat
  - Drive status
  - failed request
  - Tape pools



Castor Migration

- Castor is too old to use the migration tools
  - Still 1.7.x and using MySQL
  - LTO4 and LTO7 tapes coexist
- A Clumsy but Safe Migration Strategy
  - Data on Castor and disks : Disks → CTA
    - LHAASO, HXMT, DYB, BESIII and JUNO in future
  - Data on castor only: Castor → Disks → CTA
    - LHAASO, YBJ, HXMT, BESIII
  - A long and dull task
    - Checking integrity of castor data simultaneously via **xrdadler32**
- Current Migration Progress
  - YBJ: 4 LTO4 drives for migrating, 1/5 done, will be finished this year(?)
  - BESIII & LHAASO: will be done by this year or 2023(?)


```
[root@castor ~]# showqueues |grep tpserver08
drv25@tpserver08 (35743045 MB) jid 11109 KJ9403(read) user (12015,580) 730 secs.
drv26@tpserver08 (35517823 MB) jid 11101 KJ9428(read) user (12015,580) 747 secs.
drv27@tpserver08 (34704154 MB) jid 11817 A01019(read) user (12015,580) 40 secs.
drv28@tpserver08 (35355304 MB) jid 9448 A01055(read) user (12015,580) 779 secs.
QUEUED: A01059 ReqID: 79169 user (12015,580)@tpserver08.ihep.ac.cn received at Nov 09 22:36:43
QUEUED: A01689 ReqID: 79170 user (0,0)@tpserver08.ihep.ac.cn received at Mar 08 08:13:54
QUEUED: A01689 ReqID: 79171 user (0,0)@tpserver08.ihep.ac.cn received at Mar 08 08:14:45
QUEUED: A01689 ReqID: 79172 user (0,0)@tpserver08.ihep.ac.cn received at Mar 08 08:15:03
[root@castor ~]#
```

Problems of EOS & CTA

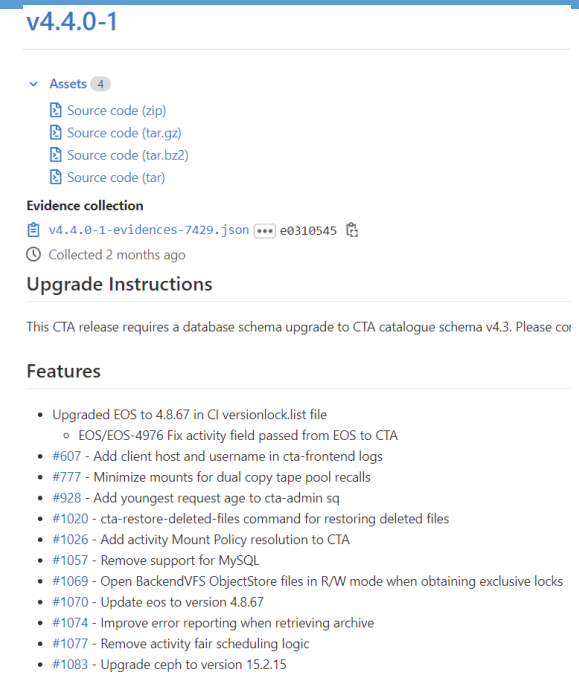
- EOS
  - Fail to copy files when filenames contain special char like “?”, “’”, “,” and “ ”
  - Fail to adjust file’s replica to normal when EOS has two or more spaces
  - EOS SE security: how to avoid attacks from WANs
- CTA
  - CTA taped service went down several times with no obvious errors
    - cta-taped got stuck and no file archived for a long time

```
LASOLIB laso08 laso-tape03      Up              -     Free 125045      -          -   -     -      -     -       -        0        - 117371 [STALE] -
LASOLIB laso09 laso-tape03    Down              -     Down 117363      -          -   -     -      -     -       -        0        - 117363 [STALE] [cta-taped] ERROR getLogicalLibraries: create[...]
LASOLIB laso10 laso-tape03      Up              -     Free 117947      -          -   -     -      -     -       -        0        - 117371 [STALE] -
LASOLIB laso11 laso-tape03      Up              -     Free 128487      -          -   -     -      -     -       -        0        - 117370 [STALE] -
LASOLIB laso12 laso-tape04      Up ArchiveForUser    Mount 117418 B01607 laso-pool2 laso     -      -     -   25714        0        - 117418 [STALE] -
LASOLIB laso13 laso-tape04      Up ArchiveForUser Transfer 117491 B01462  laso-pool laso    91 105.2G 268.2   25713        0        - 117009 [STALE] -
LASOLIB laso14 laso-tape04    Down              -     Down 117361      -          -   -     -      -     -       -        0        - 117361 [STALE] [cta-taped] ERROR getLogicalLibraries: create[...]
LASOLIB laso15 laso-tape04    Down              -     Down 117361      -          -   -     -      -     -       -        0        - 117361 [STALE] [cta-taped] ERROR getLogicalLibraries: create[...]
LASOLIB laso16 laso-tape05      Up              -     Free 126706      -          -   -     -      -     -       -        0        - 117368 [STALE] -
LASOLIB laso17 laso-tape05    Down              -     Down 117364      -          -   -     -      -     -       -        0        - 117364 [STALE] [cta-taped] ERROR getLogicalLibraries: create[...]
LASOLIB laso18 laso-tape05      Up              -     Free 120201      -          -   -     -      -     -       -        0        - 117371 [STALE] -
LASOLIB laso19 laso-tape05      Up              -     Free 124615      -          -   -     -      -     -       -        0        - 117370 [STALE] -
```

  - Upgrading Strategy
    - Stable version? LTS version?
    - Public CTA repo not accessible

Roadmap

- SSS Authentication is not flexible
  - Not convenient to add new access keys to EOS & CTA
  - Adding Kerberos auth to EOS & CTA operations
- New transmission system like FTS
- Containerized CTA testbed
  - CTA is under rapid development
  - Catalogue and dependencies like Ceph change
  - Agile to packaging and upgrading CTA via container
- Containerized production CTA
  - Isolate from OS and deploy easily
  - Migrate fast to for server failover



Summary

- EOS is one of major filesystems at IHEP
  - Servicing for LHAASO, JUNO, HXMT, IHEPBox, CTA ……
  - Exploring new technics like CSS based on ARM
- Immature CTA practice
  - Evaluation and stress tests on testbed and production passed
  - A small production instance for LHAASO, HXMT, YBJ, BESIII
  - Castor migration started and is still under way
- CTA roadmap at IHEP
  - Containerizing production instance and testbed
  - Testing modern transmission system like FTS
  - Using Kerberos in CTA

Thanks!