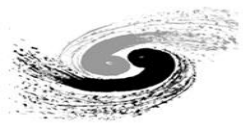


中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

EOS

CERN
Tape Archive

CTA Status and Future at IHEP

Yujiang Bi, IHEPCC

EOS Workshop 2023

26/04/2023

Outline

- CTA at IHEP
- Tape SE Practice
- CTA Roadmap
- Summary

Tape Storage Overview

- 3 physical libraries and 2 under construction at IHEP
  - BESIII & DYB： LTO4/LTO7， IBM TS3500
    - 24 drives, ≥ 5000 tapes， ≥ 22PB
  - LHAASO&HXMT： LTO7, IBM TS4500
    - 20 drives, 4500 tapes, 25PB
  - JUNO: LTO9， IBM TS4500
    - 3 drives, 120 tapes, 2.2PB
  - LHCb TIER1(middle of this year): LTO9
    - 4 drives, 560 tapes
  - HEPS(late this year?): LTO9?
- 1 library phased out – YBJ
  - Free up more space to expand LHASSO tape library

Tape Storage Overview

- ALL experiments managed by CTA
  - LHAASO since late 2021
  - BESIII & JUNO ready since late 2022
- Some old data managed by Castor
  - LHAASO, HXMT, BESIII ……
  - Will be migrated to CTA in late this year(?)

- CTA deployment
  - 3 production instances: BESIII&LHAASO, JUNO, LHCb Tier 1
  - 1 testbed with IBM TS2900 for test and evaluation
- EOS deployment
  - 4 instances: BESIII, LHAASO, JUNO and LHCb Tier 1
  - 1 testbed for upgrading test and evaluation

Setup for LHAASO & BESIII

Hardware

- LHAASO & HXMT
  - Library: TS4500
  - 12 LTO 7 drives & 3 tape servers
  - ~ 45000 LTO 7 tapes
  - Servers:
    - 2 nodes each with 12x12TB HDDs
    - Network: 25 Gb/s Fibric

No SSD

- BESIII & DYB
  - Library: TS3500
  - 8 LTO7 drives & 2 tape serves
  - ~ 1600 LTO7 tapes
  - Server:
    - 1 node with 12x12 TB HDDs
    - Network： 25Gb/s Fibric

Software

- CTA 4.7.7 & Ceph 15.2.15
- EOS 4.8.86 & XRootD 4.12.8
- PostgreSQL 14 with active-passive setup
- Authentication
  - KRB5 added to LHAASO & BESIII

Shared services between LHAASO & BESIII
Frontend, catalogue, objectstore

Setup for JUNO & LHCb Tier 1

Hardware

- Tape
  - Library: TS4500
  - 3(JUNO) + 4(LHCb) LTO 9 drives
  - 120(JUNO) + 590(LHCb) LTO 9 tapes

Shared tape hardware between JUNO and LHCb **temporarily**

- Server
  - JUNO: 2 nodes and 1 DELL 4084 Disk Array with 84x20TB HDDs
  - LHCb: 2 nodes and 1 DELL 584 Disk Array with 84x12TB HDDs
    - RAID DDP mode instead of JBOD
  - Network: 25 Gb/s Fibric

No SSD

Software

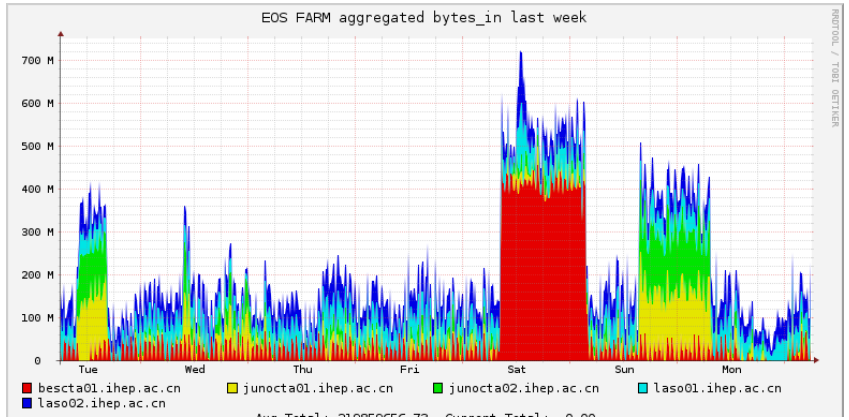
- CTA 5.7.14 and Ceph 15.2.15
- EOS 4.8.86 with XRootD 4.12.8
- PostgreSQL 14 with active-passive setup

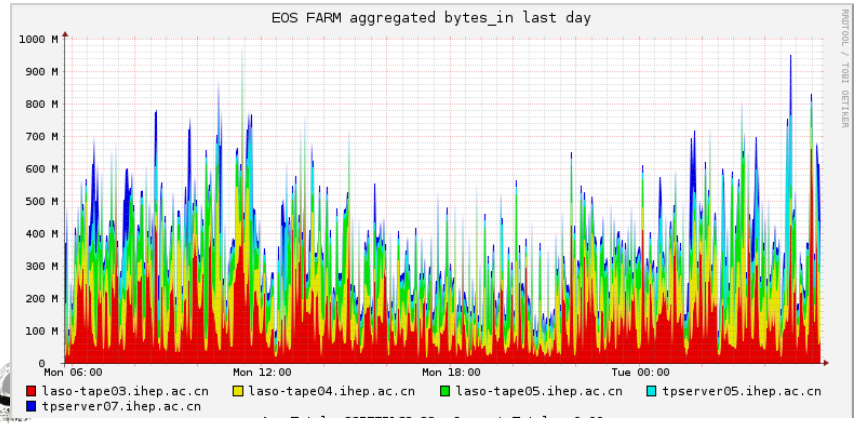
Seperated services between JUNO and LHCb

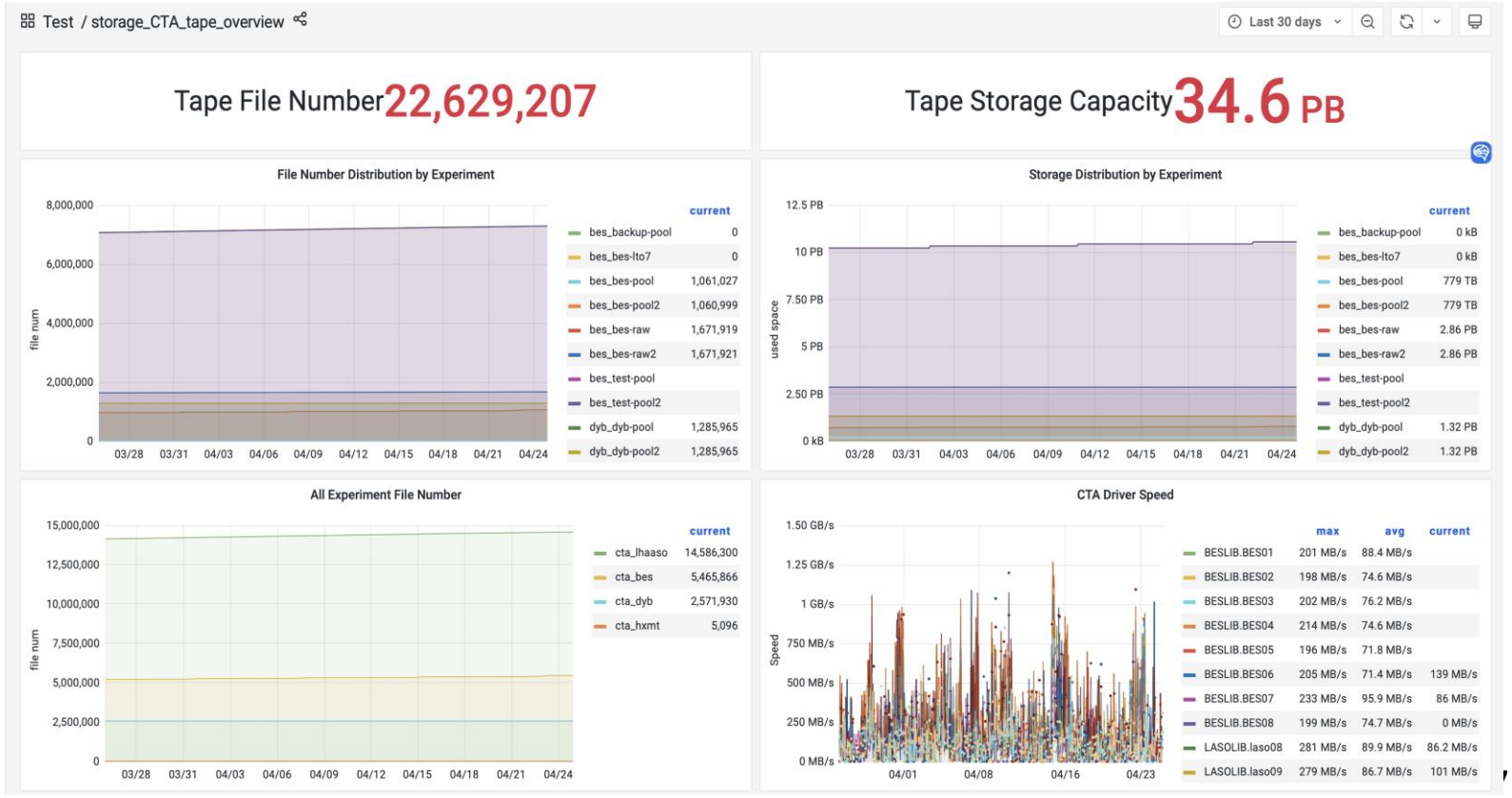
Usage Statistics

- Experiment usage overview
- Compression is turned on
  - More space than expected
- Service monitoring : Ganglia + Grafana

Experiment	LHAASO	HXMT	YBJ	BESIII	DYB	JUNO
Used/Capacity	9.6P/10.6P	22T/30T	185T/600T	3.3P/3.6P	1.2P/1.0P	800T/2.0P
Files	7.3M	3K	2.5K	600K	300K	170K
Drives	12 LTO7			8 LTO7		3 LTO9







Some Problems

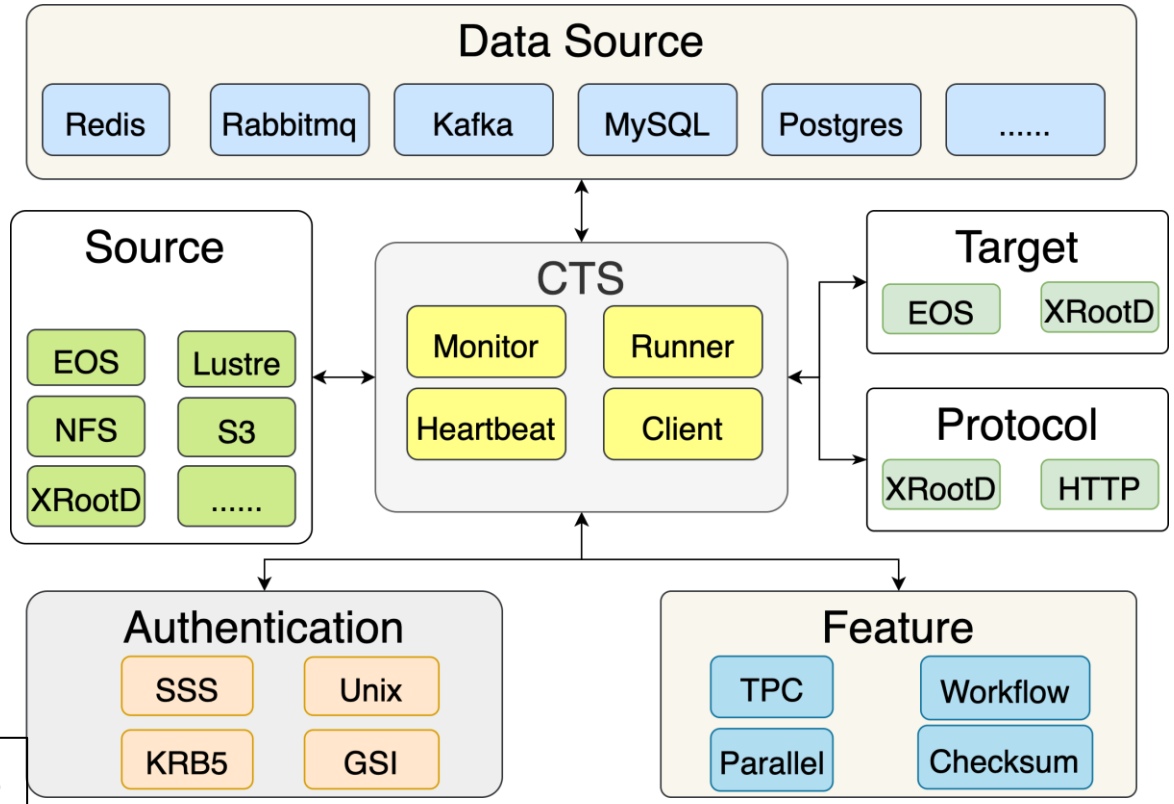
- EOS
  - WFE will lose triggered entries when MGM crashes
    - A cron job to filter missed files and add to archival queue
  - Atomic copy mode with Workflow for CTA
    - Temporary files beginning with **.sys.#** will trigger CTA Workflow and be archived
- CTA
  - Fail to recover files deleted by accident using **cta-restore-deleted-files**
    - Cannot get correct current container ID and file ID
  - cta-taped crashes on LTO9 drive when RAO is turned on
    - CTA 5.7.14 and IBM firmware P370/P380

```
CTA sent two consecutive Generate RAOs with a large amount of UDS (User Data
Segments) to drive during reading, causing the drive to stop responding.

Normally, after sending a Generate RAO, backup software would use Receive RAO
obtain the recommended read order from the drive, and then process all UDS
according to the order.

CTA only processed a small amount of UDS before sending another Generate RAO,
which caused the drive to spend a lot of resources processing the Generate
RAO and subsequently stopped responding.
```

```
[root@cta ~]# export EOS_MGM_URL=root://laso01.ihep.ac.cn/
[root@cta ~]# cta-restore-deleted-files -I 12386910 -i eoslaso  -v B02810
Created namespace endpoint laso01.ihep.ac.cn:50051 with token 263a2ef5-4a72-42f7-956b-7495dd616870
Created namespace endpoint bescta01.ihep.ac.cn:50051 with token 0bd92d00-e479-11eb-8e7d-3a68dd5ca0e7
Apr 24 19:54:55.976376 cta.ihep.ac.cn cta-restore-deleted-files: LVL="INFO" PID="2929675" TID="2929675" MSG="Listing dele
ted file in cta catalogue" userName="root" tapeVid="B02810" diskInstance="eoslaso" archiveFileId="12386910"
Apr 24 19:54:55.992476 cta.ihep.ac.cn cta-restore-deleted-files: LVL="INFO" PID="2929675" TID="2929675" MSG="Listed delet
ed file in cta catalogue" userName="root" tapeVid="B02810" diskInstance="eoslaso" archiveFileId="12386910" nbFiles="1"
Apr 24 19:54:56.227605 cta.ihep.ac.cn cta-restore-deleted-files: LVL="INFO" PID="2929675" TID="2929675" MSG="Restoring fi
le in the eos namespace" userName="root" diskInstance="eoslaso" archiveFileId="12386910" diskFileId="8902740" diskFilePat
h="/eos/laso/raw/wcda/2019/1226/ES.35556.WCDA_EVENT.P1GRBM20M.20191226111349.035.dat.gz"
Aborting: FATAL ERROR: attempt to inject file with id=0, which exceeds EOS current file id=0
[root@cta ~]# eos whoami
Virtual Identity: uid=2 (2) gid=2 (2) [authz:unix] host=cta.ihep.ac.cn domain=ihep.ac.cn
```

CTA Transmission System (CTS)

- A simple system to handle CTA archive and recall
  - To satisfy the need of JUNO, LHAASO and BESIII
    - JUNO use Kafka to pass messages between subsystems
- Designed to support various source types and protocols
  - RabbitMQ, Kafka support only
  - Redis/Postgres under development
- Features
  - Using WFE to register files to SQL
  - TPC support and parallel mechanism



Yujiang Bi › CTA Transmission System

C CTA Transmission System

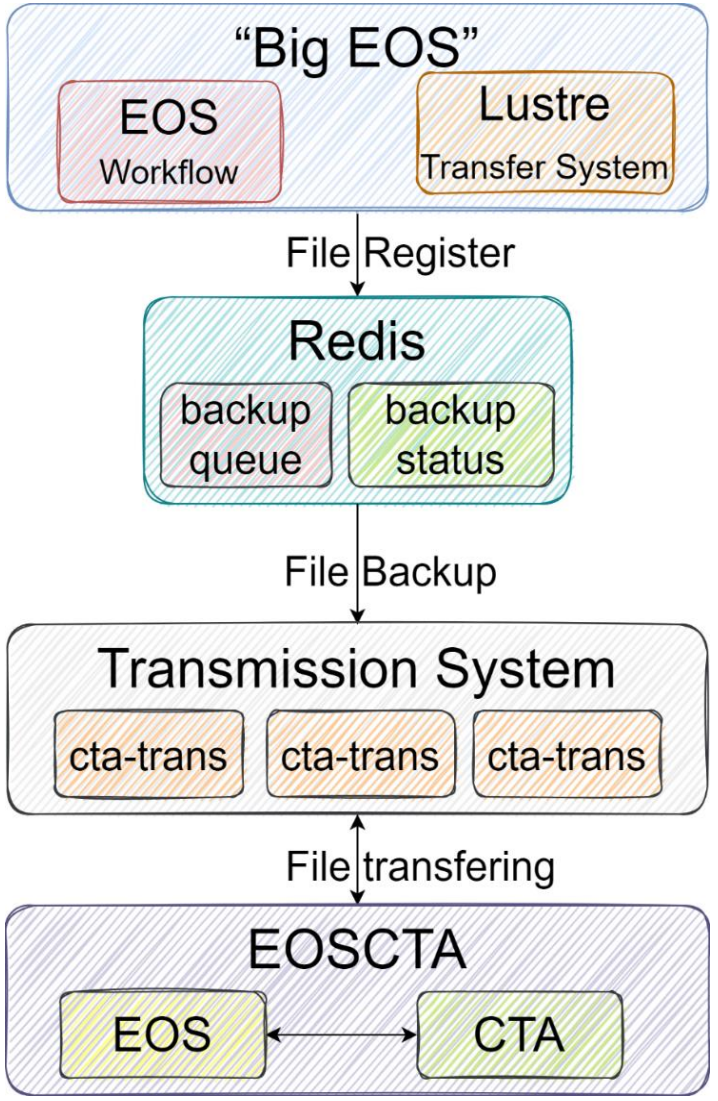
Project ID: 1356

54 Commits 2 Branches 0 Tags 307 KB Project Storage

https://code.ihep.ac.cn/biyujiang/cts

CTS Situation at IHEP

- Two version: shell & python
  - Shell version used for LHAASO CTA and data transfer
  - Python version under benchmark for JUNO
- Components of Python version
  - Runner: archival routine
  - cli: a simple script to retrieve files
  - Monitor(not yet): monitoring threads
  - Checker(not yet): check if files are archieved
- Developing Plan
  - Fully implement Redis support for LHAASO
  - Implement monitor and check function
  - Add a Redis-like DB to track instance status



Solution is not CTS but FTS

- CTS is a temporary solution for LHAASO and other experiments
  - Works well so far but limited usage scenarios
  - Immature development and lack of manpower
  - Lack of many functions like user requests management
- FTS is more reliable and stable, but more efforts are needed
  - Some instances deployed at IHEP
  - Backup Workflow to be integraded
  - Monitor and alart to be developed

Tape Storage SE

- Tape storage is necessary accessed by WAN safely for HEP
  - JUNO、Herd、LHCb Tier1 at IHEP ……
- EOS as SE for both disk and tape storage?
  - Supporting various auth like krb5, **GSI** and **scitokens**
  - Supporting XRootD natively and HTTP(s) by XrdHTTP
  - Providing tape restful API to access CTA
- Things to do before into production
  - SE deployment and configuration
  - Necessary SE workflows to evaluation
- Testbed Setup
  - CTA 5.7.14 + EOS 5.1.9 + IBM TS2900
  - Protocol + Authentication : **XRootD + GSI** | **HTTP(s) + GSI/Token**

Tape SE Practice

- Deployment & configuration
  - Using Herd's & JUNO's IAM for testing

More details in **X.T Zhang**'s report

- Workflows to evaluate
  - File archival, staging, status query and recall
  - Replica evict, request cancel and delete
- Tools
  - XRootD: xrdfs, xrdcp or eos cp
  - HTTP(s): curl, wget and gfal
- Result
  - All workflows were passed with

XRootD + GSI & HTTPS + GSI/Token

```
$ cat recall.submit.sh
#!/bin/bash -

source $HOME/cc/cta/test/token/se.env.sh

json='
{
        "files":[
          {
                "path":'"\"$1\""'
          }
        ]。
}
'

echo $json | curl -L -v -H "Accept: application/json" -H "Authorization: Bearer $BEARER_TOKEN" \
-X POST $mgm/api/v1/stage -d @- 2>/dev/null | jq
$ ./recall.submit.sh /eos/test/juno/http/http.test.img
{
  "requestId": "680ff528-b700-11ed-925a-141877378c22"
}
```

  - Full test progress can be found at here in case any interest.

Tape SE Practice – Todo list

- More function to evaluate
  - Third-Party-Copy with XRootD and HTTP(S) protocols
  - XRootD GSI Delegation Proxy for TPC and auth for FST(?)
- Authentication to test and support
  - ztn: Tokens over XRootD protocol
- Automatically user mapping configuration
  - Scitokens and GSI gridmap-file
- IAM services for various experiments if possible
- Performance benchmark
- Join the joint DOMA Tape Restful API test(?)

Roadmap for CTA

- Phasing out Castor services
  - LHAASO, BESIII and Backup
- Monitoring and virtualization
  - Automatically and intelligently recovering CTA services
  - Refining Grafana dashboard for CTA
  - Alarming with Mattermost, Wechat or other IMs
- CTA Upgrading strategy
  - Upgrading all instances to EOS/CTA 5 in the middle of this year
  - Following the stable release pace of CTA

Roadmap for CTA

- Deploying and managing tape storage SE
  - Preparing tape SE for LHCb TIER1, JUNO, Herd...
  - Monitor, management, communication, collaborations...
- Deploying FTS for LHAASO, BESIII and JUNO CTA
  - Automatically handling retrieval requests
  - Virtualizing service status and statistics
- **Separating CTA service for LHAASO and BESIII**
  - Frontend, catalogue, objectstore ……
- Service containerization……

Summary

- CTA is the dominant tape system at IHEP
  - Managing LHAASO, BESIII, JUNO, LHCb Tier 1, and other HEPs
- CTS takes care of migration between big and little EOSs
  - FTS shall be the solution for CTA migration
- Tape SE instances deployed and in testing
  - Function evaluation and performance stress test
- Tasks ahead to do
  - Castor retirement, CTA upgrading, FTS for CTA
  - Service monitoring and containerization……

Q & A

Thanks!