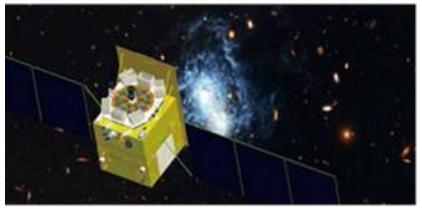
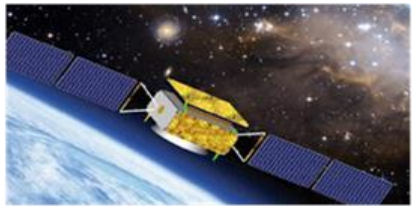
Long term Storage Achieves of IHEP

Qiuling Yao
On behalf of Storage Group
Computing Center, IHEP
2022-03-25

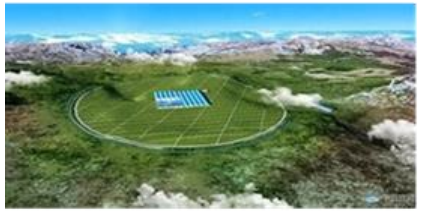
Outline

- Background
- CASTOR at IHEP
- EOSCTA Evaluation and Usage
- Data Storage Workflow
- Next Steps
- Summary

Large-scale Scientific Projects at IHEP

~550KM

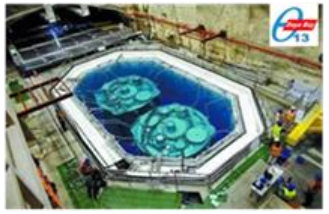
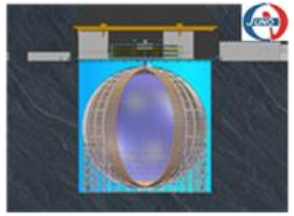



~4400M



中国散裂中子源
China Spallation Neutron Source
高能同步辐射光源
High Energy Photon Source | HEPS

~-5M

~-300M

Astro Physics Experiments
HXMT: 0.5PB/year, 2018-
GECAM:0.5 PB/year, 2020-2023

High Latitude Cosmic Ray Observatory
LHAASO: 6 PB/year, 2020-2030
YBJ: 0.1PB/year, 1990-

Collider Experiments
BESIII: 0.7 PB/year, 2010-2027
CEPC prototype: 0.3 PB/year, 2014-

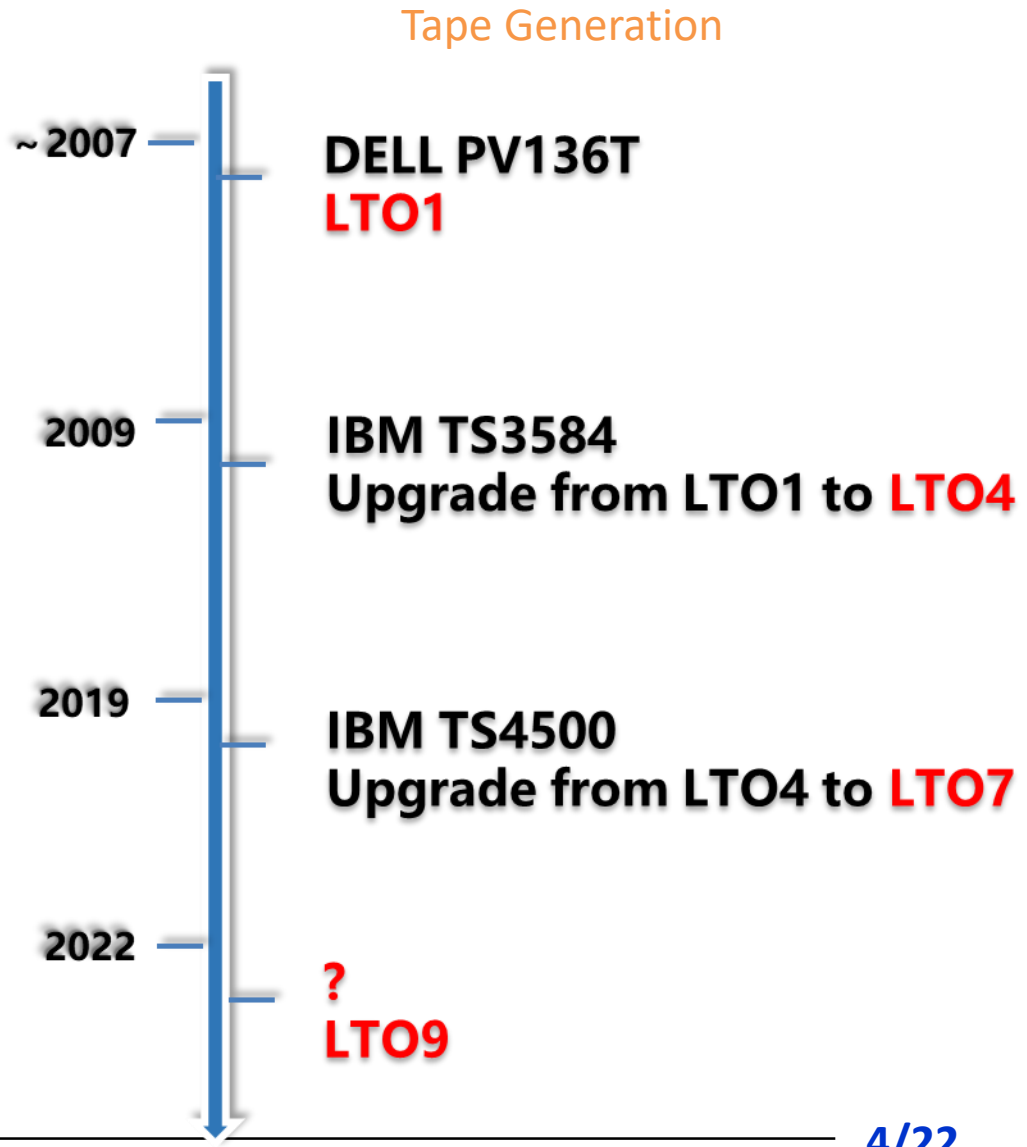
Neutron/Photon Sources
CSNS: 1 PB/year, 2018-
HEPS: 150 PB/year, 2025-

Neutrino Experiments
DayaBay: 0.3 TB/year, 2012-2021
JUNO: 3 PB/year, 2022-2032

- Mass data (~PB/year) from multiple experiments
- Important data will be preserved permanently

Data Storage Management

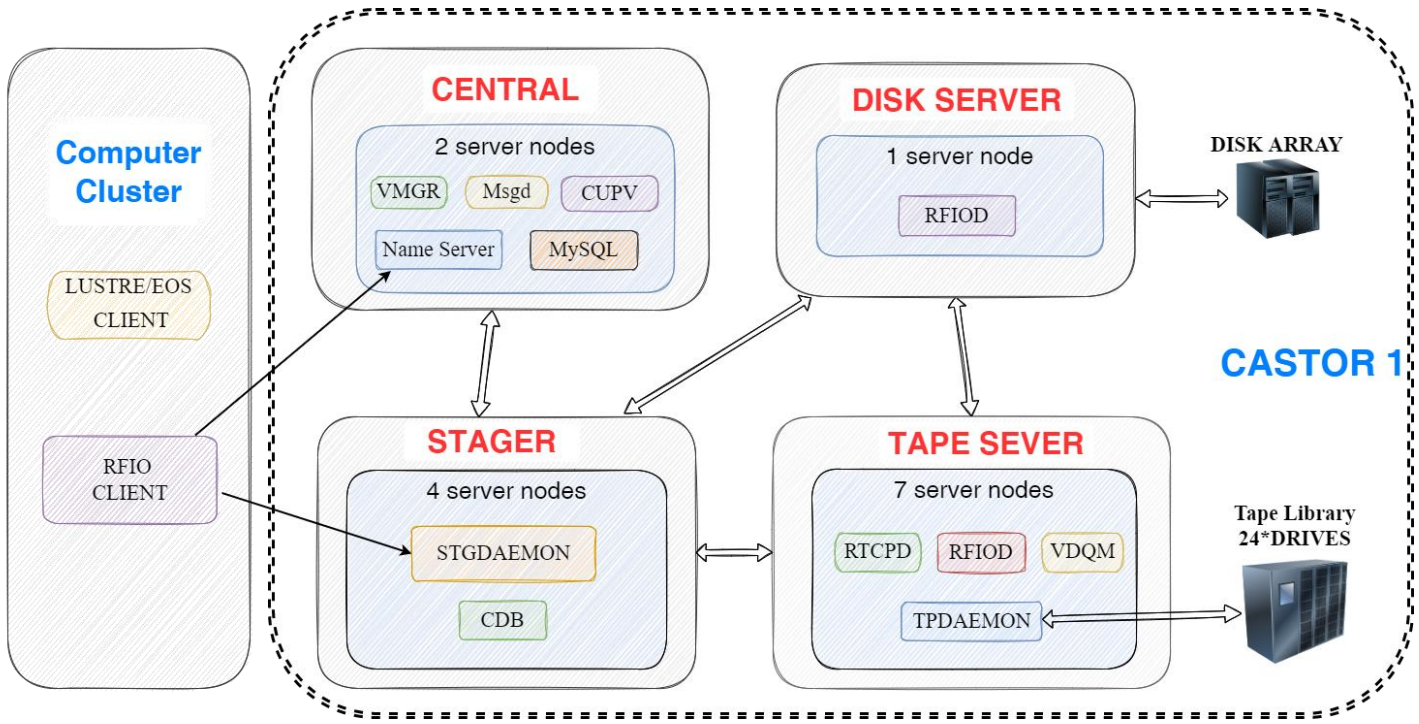
- Objective
  - Massive storage space(PB~EB)
  - High performance
    - Tens of thousands of concurrent job access
  - Dynamic data access frequency
  - Long term reservation
    - The data can be accessed for several years
- Solution
  - Online: Disk Array
  - Offline: Tapes
- Tape usage
  - LTO tapes adopted since 2007
  - Three generations
  - Tape Capacity from GB to TB, now PB



Outline

- Background
- CASTOR at IHEP
- EOSCTA Evaluation and Usage
- Data Storage Workflow
- Next Steps
- Summary

Tape Archival System - CASTOR



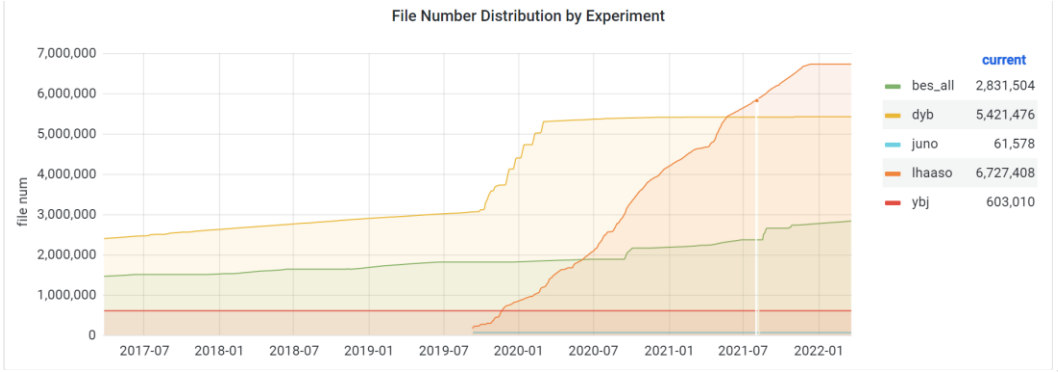
IBM 3584 for BESIII

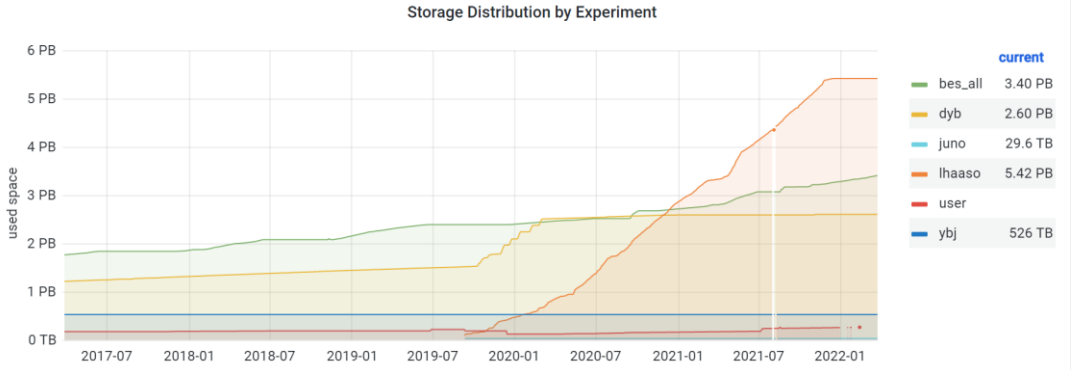
IBM 4500 for LHAASO

- In production since 2007
- Version: 1.7
- 3 tape library & 14 server nodes
- Unified namespace, e.g.: /castor/ihep.ac.cn/bes
- Command lines & programming APIs based on RFIO protocol

CASTOR at IHEP

- Servers for 5 experiments, 2 replicas
  - BESIII/YBJ/DYB/JUNO/LHAASO
  - Backup for user data
- Customized modifications
  - Improve compatibility of LTO generations
  - Database optimization
  - Tape access acceleration
  - Enrich access functions for users





	BESIII	DYB	JUNO	LHAASO	YBJ	BACKUP	TOTAL
FILE NUMBER	2,831,504	5,421,476	61,578	6,727,408	603,010	3,500,271,515	3,515,916,491
USAGE(TB)	3402.505	2599.82	29.644	5415.814	525.551	265.845	12,239.179

CASTOR disadvantages

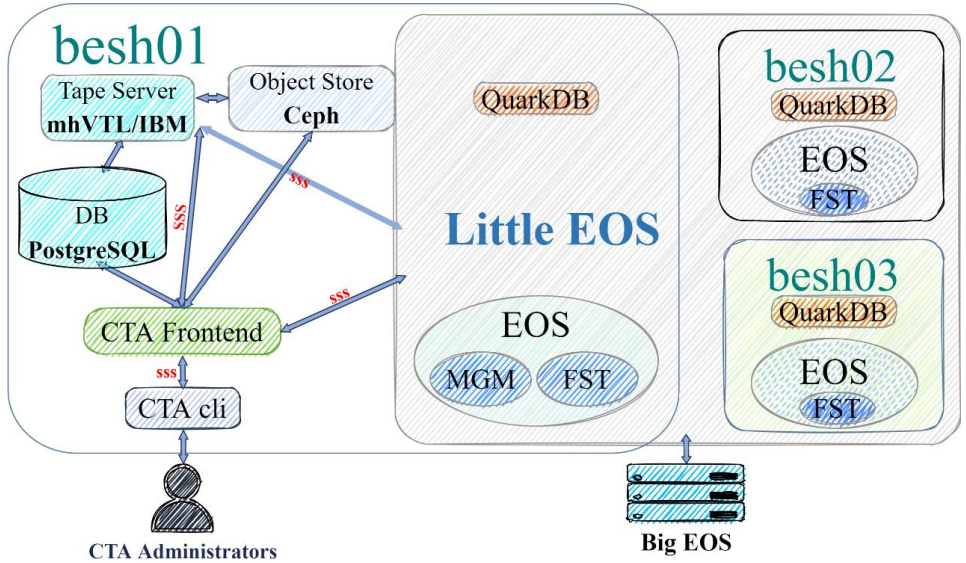
- CASTOR works well but still has problems
  - The failure rate increased when many files (~50K) waiting in queue for migration
  - Limited scalability
  - Only one replica is saved but no alarms
  - No optimized migration mechanism for retrieval requests
  - Operations are complicated
- CASTOR has been replaced by EOSCTA by many Institutions
- EOS is a main online storage system at IHEP

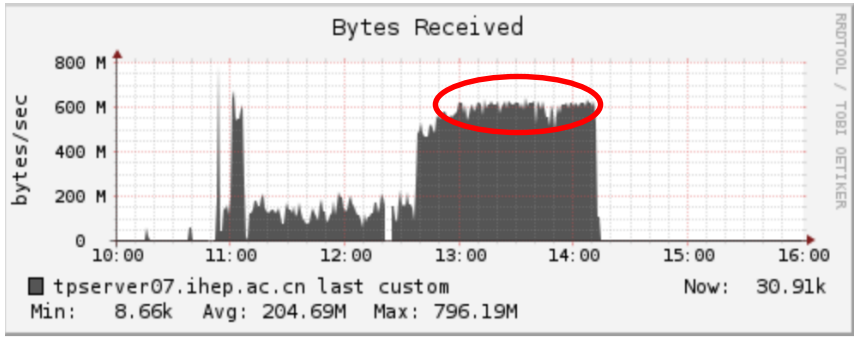
Outline

- Background
- CASTOR at IHEP
- EOSCTA Evaluation and Usage
- Data Storage Workflow
- Next Steps
- Summary

EOSCTA Evaluation

- Testbed
  - EOS : 4.8.34 & 4.8.40, CTA : 3.1-14 & 4.0-1
  - Management of the drives and tapes by CTA
  - CTA commands
  - Data archive and retrieval
- Evaluation results
  - Hardware architecture is very similar to CASTOR
  - All function tests passed except Kerberos authentication
  - The speed of disk and tape is faster than expected

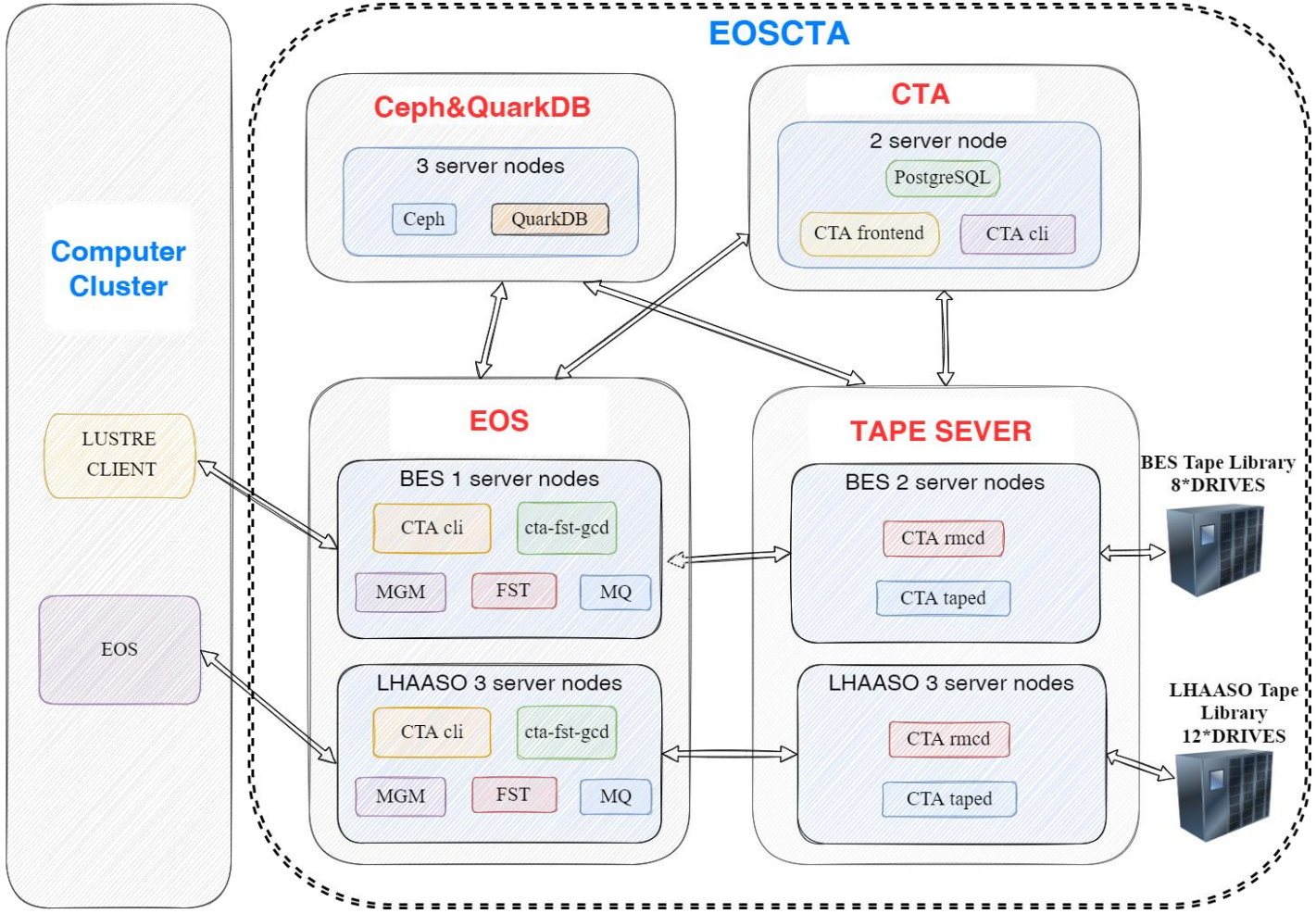




library	drive	host	desired	request	status	since	vid	tapepool	vo	files	data	MB/s	session	priority	activity	age	reason
BESLIB	BES01	tpserver07	Up	ArchiveForUser	Transfer	14399	L71553	dyb-pool	dyb	2572	2.7T	353.6	34263	0	-	13	-
BESLIB	BES02	tpserver07	Up	ArchiveForUser	Transfer	4433	L71687	dyb-pool2	dyb	817	849.3G	282.9	34275	0	-	10	-

EOSCTA Architecture

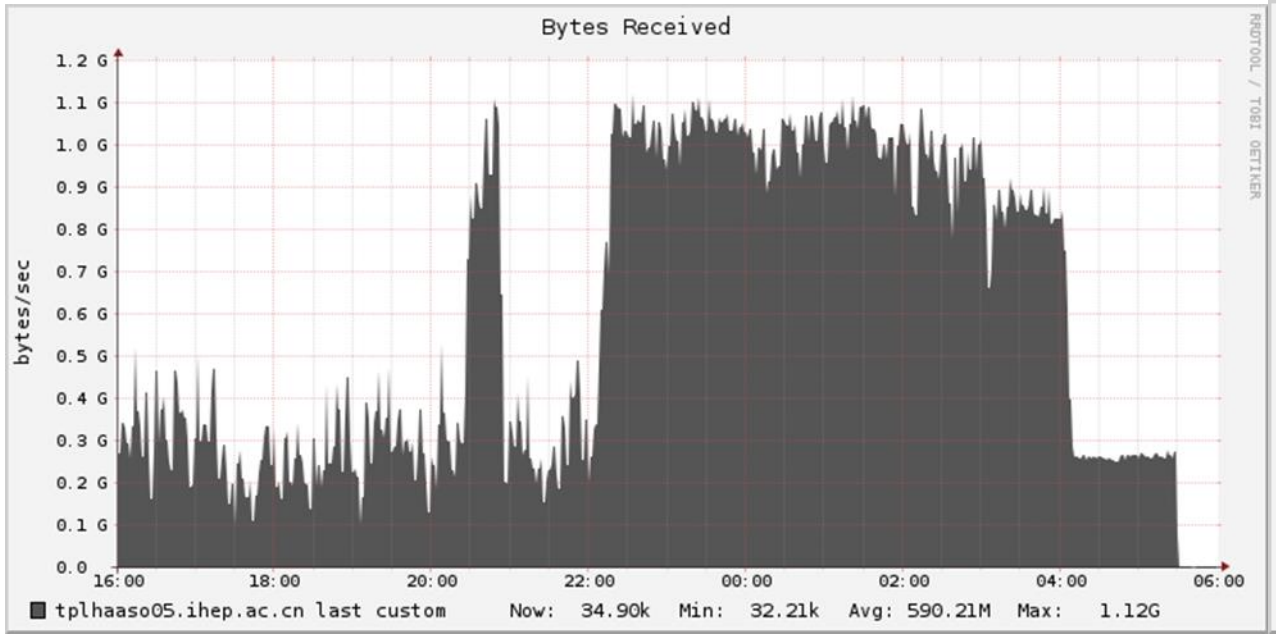
- Hardware
  - 4 EOS storage nodes
    - 12x12TB HDDs
  - 5 Tape Servers for BES & LHAASO
  - 2 Tape library(share with CASTOR)
    - 20 LTO7 drives
  - 1 CTA frontend node and 1 catalogue
    - 2x1TB SATA SSDs
  - 3 Ceph & QuarkDB nodes
    - 8x1TB SATA SSDs
  - Network connection
    - 10G/25G Fiber
- Software
  - Eos:4.8.45 CTA:4.2-1
  - Ceph:14.2.20 QuarkDB:0.4.3

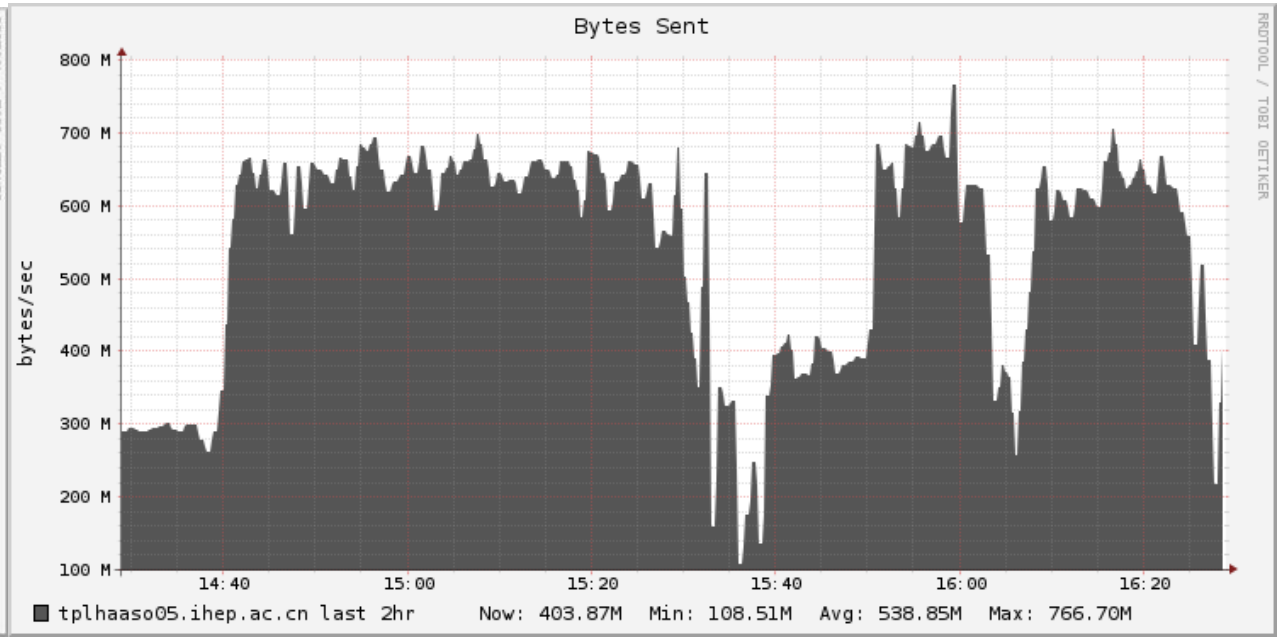


EOSCTA Stress Test in production

- 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 are archived and retrieved successfully.
- The performance of archival and retrieval is quite good.



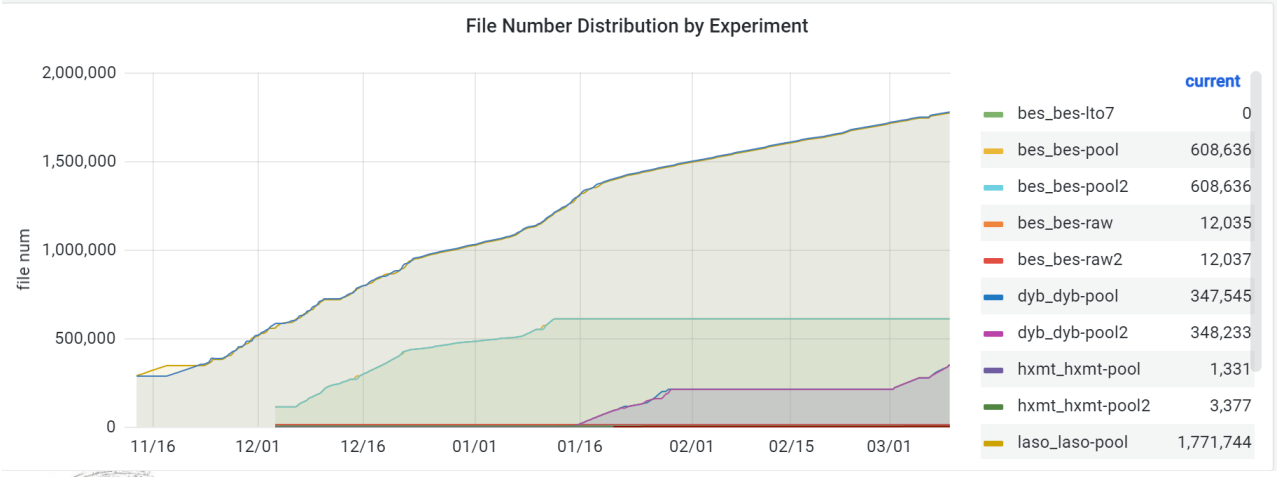


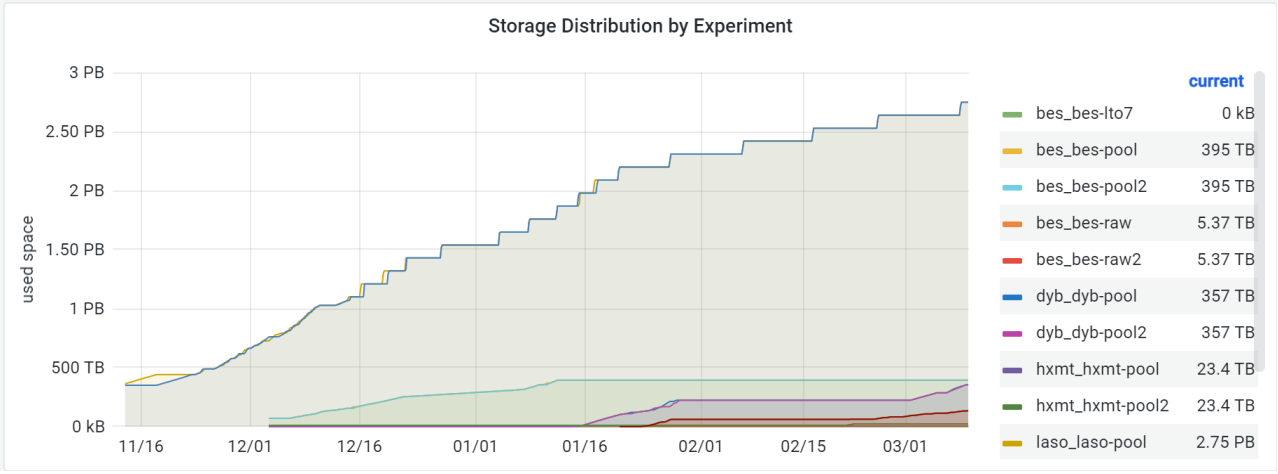
Current status of EOSCTA

- Used in production since 2021
- Has been provided to five experiments
  - HXMT : All long-term storage services switched to EOSCTA
  - BESIII/LHAASO/DYB/YBJ : Data migration in progress

Tape File Number **5,494,101**

Tape Storage Capacity **23.8 PB**





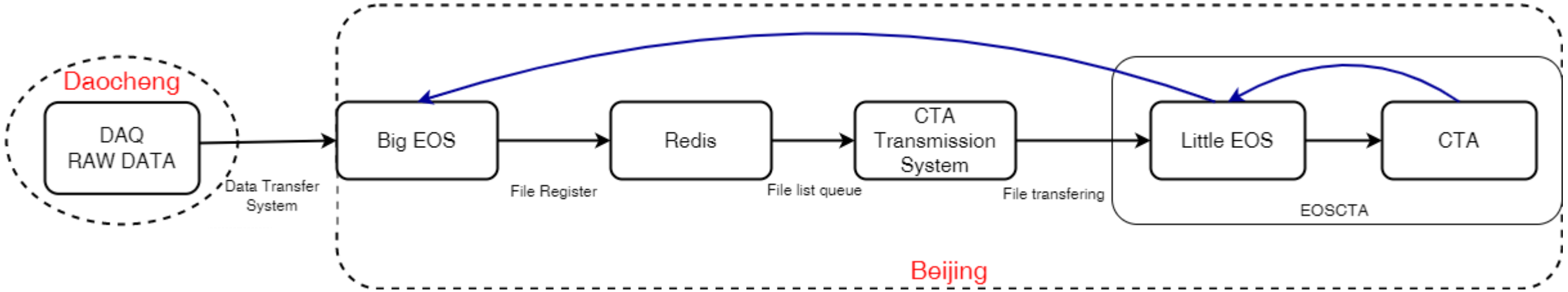
Upgrade of Long-term Storage

- Upgrade plan
  - All existing data of CASTOR should be migrated to EOSCTA
  - Tape upgrade for BESIII/JUNO/YBJ : from LTO4 to LTO7
  - Long-term storage services should be switched from CASTOR to EOSCTA
- Completed jobs
  - Data migration: ½ DYB & 1/5 YBJ
  - New experiment data of LHAASO & HXMT has been archived to EOSCTA since 2021
  - BESIII data except DAQ data has saved to EOSCTA

Outline

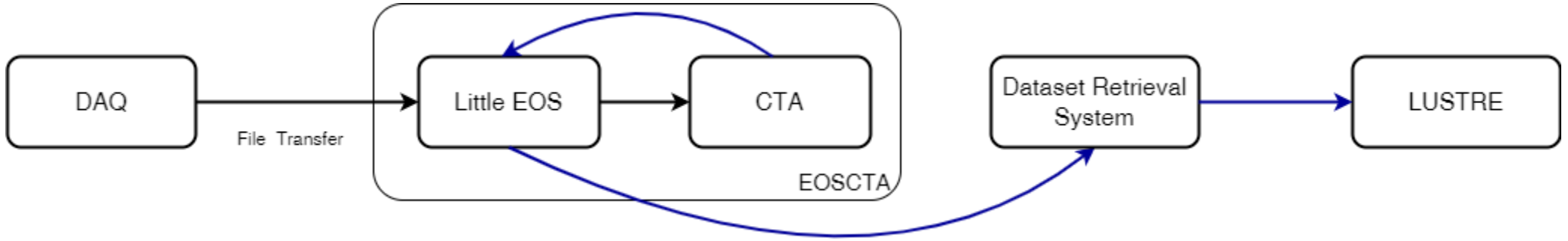
- Background
- CASTOR at IHEP
- EOSCTA Evaluation and Usage
- Data Storage Workflow
- Next Steps
- Summary

LHAASO Storage Workflow



- DAQ(Daocheng) → BIG EOS(Beijing) → EOSCTA(Beijing)
- Files are created all time and recreated occasionally.
- Files are archived automatically but retrieved manually.
- Data integrity check will be done after each transmission.

BESIII Storage Workflow



- DAQ → EOSCTA → LUSTRE, all at Beijing.
- Data is immutable after generation and saved immediately.
- Files are saved to EOSCTA based on Xrootd protocol.
- Files will be retrieved to the online filesystem according to Datasets.

Long-term Storage Policy

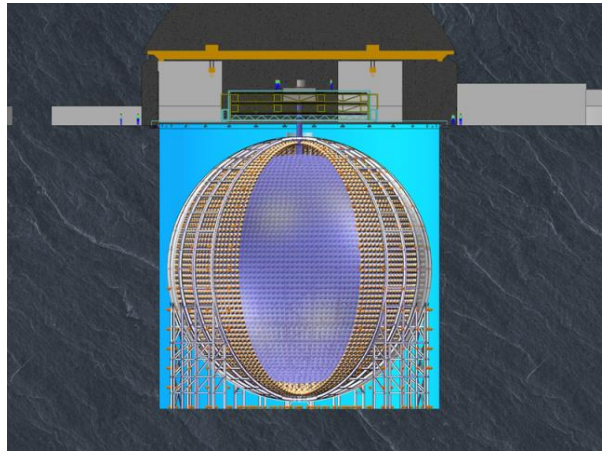
- One-write-all-read data.
- large files(>100MB) preferred.
- Two replicas for experimental data but only one replica for user/backup data.

Next Steps

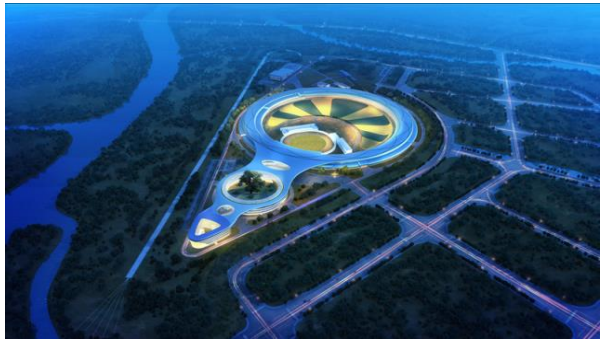
- Try other transmission tools like FTS
- Using Kerberos in CTA
- LUSTRE(HSM) + EOSCTA
- Cooperation with other software
  - DTS(Data Transfer System) and DMS(Data Management System)

Future planning of other experiments

- JUNO(Jiangmen Underground Neutrino Observatory)
  - 3PB/year , 2022~2032
  - EOS + EOSCTA
  - Data workflow is similar to LHAASO
  - LTO9, 20 drives, 5K+ slots



- HEPS(High Energy Photon Source)
  - 150PB/year , 2025~
  - LUSTRE + EOSCTA
  - Has a dedicate DMS(Data Management System) to handle data flow



Summary

- All the experiments at IHEP require long-term storage.
- Long-term storage system upgrade from CASTOR to EOSCTA.
- Different data workflow for experiments.
- Two more experiments will be supported by EOSCTA in near future.

Thank you!