PROCEEDINGS OF SCIENCE

Long-term Storage Achieves of IHEP: From CASTOR to EOSCTA

Qiuling YAO,[a,*] **Yujiang BI**[a,b] **and Yaosong CHENG**[a]

[a] *Institute of High Energy Physics, Chinese Academy of Sciences,*
Beijing 100049, China

[b] *Tianfu Cosmic Ray Research Center,*
Chengdu 610213, China

E-mail: yaoql@ihep.ac.cn, biyujiang@ihep.ac.cn, chengys@ihep.ac.cn

CASTOR is the primary tape storage system of CERN, it has been used for over fifteen years in IHEP. By 2022, the data volume has reached 12PB from the various experiments. More experiments in IHEP, such as JUNO, CEPC, and HEPS require long-term storage, and to handle the quick increase of data, we plan to replace the tape storage system from CASTOR to EOSCTA. From 2021, new data on LHAASO has been saved gradually in EOSCTA. In 2022, BES online data and JUNO raw data will be saved directly in EOSCTA.

In this paper, we describe the current infrastructure of EOSCTA in IHEP. We set up two EOS instances, which are served for four experiments, to support multiple online file systems (LUSTRE and EOS). According to the different data generations, we design different workflows to receive data from remote experimental stations or local disk arrays to EOSCTA.

CASTOR will be replaced by EOSCTA and all existing data of CASTOR will be migrated to EOSCTA. We also will update the generation of tapes, from LTO 4 to LTO 7, including five CASTOR instances, and two types of the tape library. It is planned to complete the most migration by 2023. The paper reports the migration plan, the steps and methods of data migration, and the inspection after ensuring the data integrity.

Based on the previous experience with EOSCTA, we present the outlook on the requirement of experiments in IHEP and discuss the possible way to use EOSCTA to achieve mass data storage from different data sources.

International Symposium on Grids & Clouds 2022 (ISGC 2022)
21 - 25 March, 2022
Online, Academia Sinica Computing Centre (ASGC), Taipei, Taiwan

[*] Speaker

1. Introduction

IHEP is China's biggest laboratory for the study of particle physics and has a broad range of research in related fields such as Astrophysics, accelerator technologies, and nuclear analysis techniques. More projects at IHEP, including BESIII(Beijing Spectrometer experiment III), LHAASO(Large High Altitude Air Shower Observatory), and DYB(DayaBay Neutrino experiment)[1], generate a large amount of experimental data every year.

Since 2007, IHEP has been using CASTOR 1 to provide a long-term storage service for experimental data and backup data. With the rapid growth of physics data, the capacity of the storage system increases from petabyte to exabyte. CASTOR[2] works well but still has some problems handling massive data. EOS[3] is our main online storage system now, in conjunction with CTA[4], and is the long-term storage solution we are considering.

2. CASTOR

2.1 Current Usage

CASTOR is a mass storage management software developed by CERN, which uses disk-tape-based hierarchical storage technology to store physical experiment data and user files. A unified namespace is provided in the cluster. By using the command line or programming APIs, users can easily retrieve files stored on CASTOR.

The current CASTOR system manages 3 tape libraries and 14 servers. The data volume has reached 12PB from the various experiments. Most of the raw data from the multiple experiments have two replicas on tape, so the usage of tape exceeded 20 PB.

	BESIII	DYB	JUNO	LHAASO	YBJ	BACKUP	TOTAL
Files	2831504	5421476	61578	6727408	603010	3500271515	3515916491
Usage(TB)	3402.505	2599.82	29.644	5415.814	525.551	265.845	12239.179

Table 1: Data statistic in CASTOR(Mar.2022)

2.2 Some issues of CASTOR

First, CASTOR includes tapes and disk arrays, but it is outdated and may not be compatible with later tape and server types. Second, when a large amount of data had been archived to the upper disk arrays and waited in queues for migration to tape, the failure rate increased. Third, all raw data is normally kept in two replicas, but a small amount of data was kept in only one replica on tape, and the archival process did not report any errors or alarms. Fourth, since castor was developed a long time ago, there was no optimized migration and tape acceleration for data retrieval. The data retrieval from castor is very slow.

3. EOSCTA

EOSCTA is designed to replace CASTOR[5]. CTA(The CERN Tape Archive) is a tape backend to EOS disk system.Firstly we built an EOSCTA test bed for the basic operation and functionality.

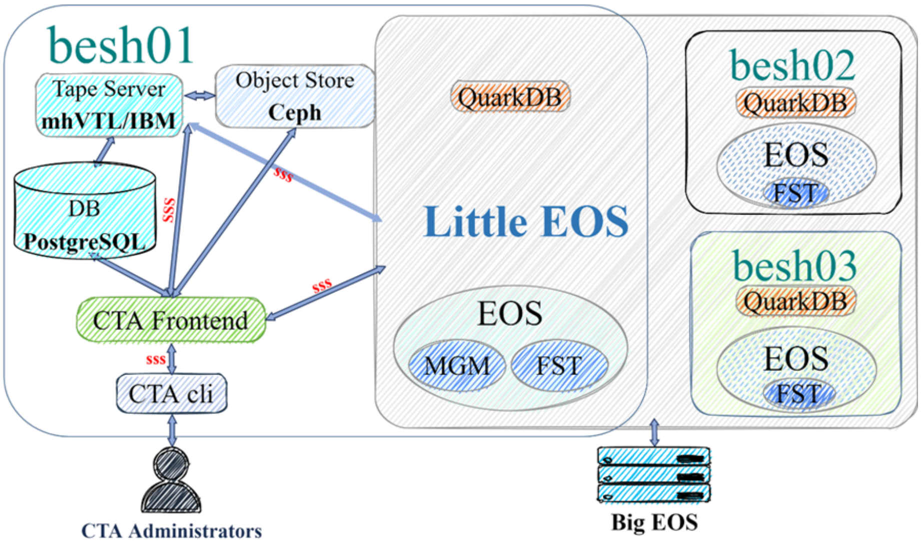

Figure 1: Testbed for EOSCTA at IHEP

3.1 Testing of EOSCTA

The testbed of EOSCTA included a test tape library and three server nodes. Two nodes were used as a little EOS, which can copy data from the big EOS filesystem in production. The other node was used to install the main components of CTA. Using PostgreSQL for CTA Catalogue,the Object store of Ceph for store message queues, and SSS for user authentication.The administrator can implement commands forwarding and execution through the CTA frontend.The structure is illustrated in Fig.1.

We tested the management of LTO drives and tapes by CTA, created an EOSCTA instance, and performed data archival and retrieval tests.

After the functional testing of EOSCTA was completed, we found that the hardware architecture of EOSCTA is very similar to CASTOR, and EOSCTA can share the tape library with CASTOR. The speed of disk array is not very different from CASTOR, but the tape speeds are much faster.

3.2 Stress test

Starting in 2021, we gradually build the official environment for the use of the EOSCTA. CTA splits 4 LTO7 drives and a portion of tapes from CASTOR's tape library. We also build a little EOS, which is completely separate from the big EOS, to act as a disk cache that can serve multiple physical experiments.

We did a stress test before official use. 100 thousand files of 1 GiB were archived to EOSCTA via 20 threads from 2 nodes, and then all the files were retrieved from the tape. There were up to 80 thousand files queued for migration to tape during the data archival progress. Before retrieving the data from the tape, we cleaned up the data replica on the EOS disks. All 100 thousand files were archived and retrieved successfully and the performance is quite good. The results are shown in Fig.2 and Fig.3.

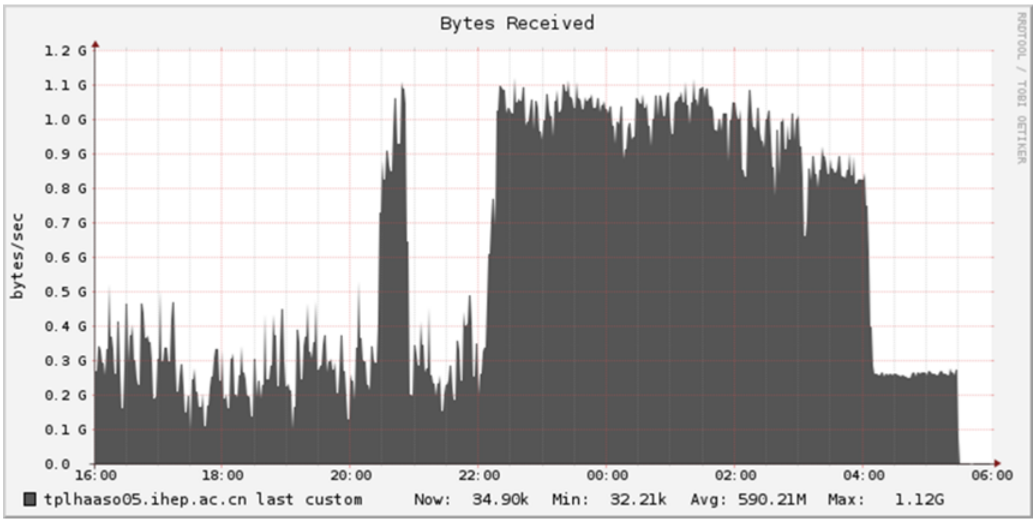

Figure 2: Network traffic of data archival test on a tape server

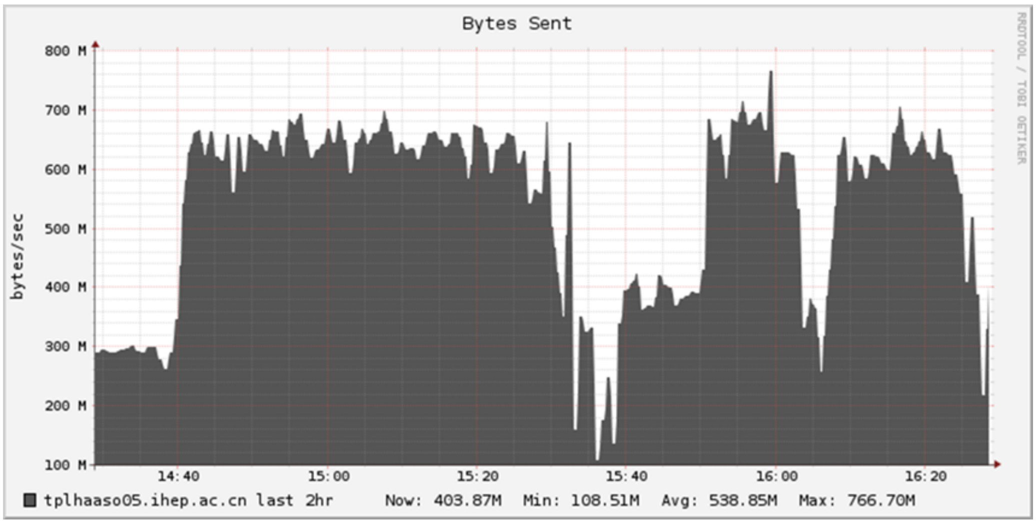

Figure 3: Network traffic of data retrieval test on a tape server

3.3 The current infrastructure of EOSCTA

The storage system in IHEP consists of two parts: the online file systems, LUSTRE and EOS, which are stored experimental data, user scripts, and all job execution processes. These file systems are mounted on computer nodes in the cluster, available both for the physical experiments to archive data and used by all users to process data. The offline file systems are mainly for long-term storage, storing experimental raw data, archived data, and backup data. Some physics experiments archive the raw data directly to the online file system. And some write to the offline system, based on the data set, physicists select the data needed and write them to the online file systems. File exchange between both file systems occurs almost all the time.

EOSCTA is officially in use in IHEP. It consists of several important components, as shown in Fig.4. The little EOS is a component that links EOSCTA to the online file system, similar to a temporary storage directory. Through manual copy or copy scripts, data in the online system can be written to little EOS, and data in little EOS can be read to the online file system. CTA is the back end of little EOS, passing data to the tape server and eventually saving data to tape. The message queue in Ceph controls the operation of the CTA. The Catalogue of EOSCTA maintains file information, tape information, migration policies, etc. Similar to CASTOR, data stored in the EOSCTA is provided externally with a uniform filename starting with /eos, regardless of where the data is stored.

Two experimental instances were defined in EOSCTA, an instance serves for BESIII and DYB experiment, and another is used for LHAASO, HXMT, and YBJ. Each instance has a tape library (BES: IBM TS3584 LHAASO: IBM TS4500), which is shared with CASTOR. CTA consists of 5 tape nodes, completely independent of CASTOR. There are four EOS storage nodes, each node contains 12*12TB HDDs for disk buffer. The central service like Ceph and database is also shared between the two instances. All EOSCTA servers are linked via a 10Gbit or 25Gbit network. In the upper layer of EOSCTA, there is an online file system(LUSTRE or EOS).

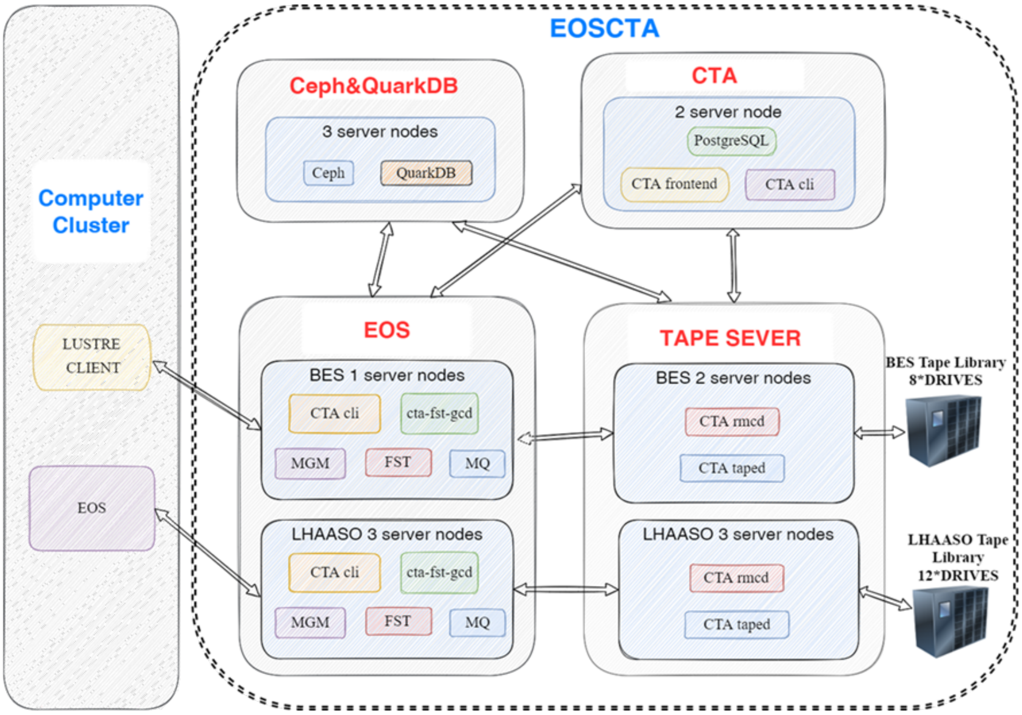

Figure 4: EOSCTA's infrastructure map

3.4 Storage workflow for LHAASO raw data

The LHAASO experimental data is generated at a remote site in Daocheng, Sichuan province. All experimental raw data needs to be transferred from Daocheng to the Computing Centre, Beijing.Occasionally, due to some problems, the raw data need to be re-generated, re-transferred and re-saved, and the previous file with the same name has to be replaced.

In addition to the data archival process described above, data retrieval is a major part of the workflow. Data are copied from tapes, passed through the little EOS, and eventually saved in the big

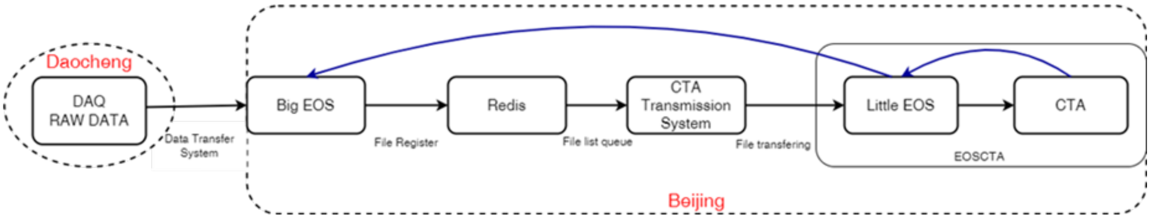

Figure 5: LHAASO storage workflow

EOS for analysis. The storage workflow of the LHAASO experiment is shown in Fig.5, including data archiving and retrieval.

Data archival process:

1. Using the xrootd protocol, files are transferred via DTS (Data Transfer System) to IHEP in real-time and saved to the big EOS of LHAASO.

2. Do the file integrity by checking the file's alder32 code. If no errors, delete the file on the local hard disk in Daocheng.

3. Files register a bit of file information, such as file size, date, checksum, etc in Redis. Redis keeps queues of file list and inserts the archival requests into the queues.

4. CTA Transmission system, a transfer tool we developed, queries the file list queues in Redis and transfer some files to the little EOS.

5. Every directory of the little EOS has a storage policy already defined, including which tape pool the files can be migrated to, the migration interval, and whether the EOS file cache is retained. According to the file storage policy, CTA archives the files from the little EOS to an available tape.

6. Do the file integrity between disk and tape. If the files are consistent, keep or delete the files on disk according to the definition of the garbage collector of EOSCTA.

The entire data archival process is automated and the data retrieval process is manual. Generally, data retrieval is simply a matter of fetching back to the big EOS.

Data retrieval:

1. Check file location definition via EOS commands.

2. If the data replica is still on disks of the little EOS, copy it back to the big EOS.

3. If the data replica of disks has been evicted, use "xrdfs prepare" to inform CTA to copy the files on the tape.

4. Mount the tape into a tape drive, and transfer the files to the disks of little EOS.

5. Query the execution results for "prepare" request. If it is true, do step 2.

3.5 Storage workflow for BESIII raw data

The BESIII experimental facility is at IHEP and the raw data does not go through via file transfer system. The experimental file is immutable after it is generated.

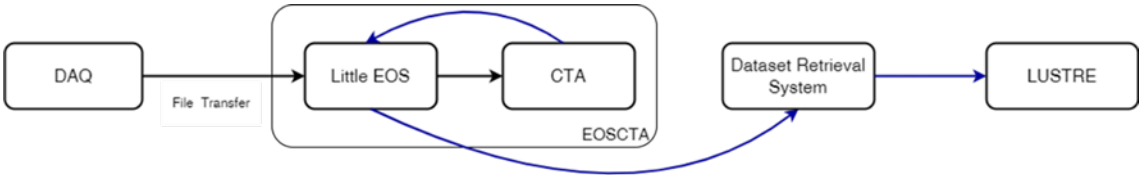

Figure 6: BES storage workflow

Previously, the DAQ program used the rfio protocol in CASTOR to archive the physics raw data to tapes in real-time. It will start converting to xrootd of EOSCTA this year. Within a day or two, depending on the data sets, some required data will retrieve from EOSCTA to LUSTRE, the BES online file system. Due to the short interval, the files are usually retrieved from the disk cache to LUSTRE. The storage workflow of the BESIII experiment is shown in Fig.6

4. Data Migration from CASTOR to EOSCTA

The upgrade of the long-term storage system is not just a software upgrade from CASTOR to EOSCTA. The CASTOR version of IHEP is version 1.7, we cannot just upgrade the metadata information, and even cannot use the migration tools of CERN. We have to do a physical data movement[6]. Upgrade plan includes: all data saved in CASTOR should be migrated to EOSCTA, BES and JUNO tapes need to be upgraded from LTO4 to LTO7, and all long-term storage services at IHEP will be switched to EOSCTA.

4.1 Migration Plan

Since several tens of petabytes of data are already stored in CASTOR, it is necessary to gradually migrate the data from CASTOR to EOSCTA. The data of CASTOR consists of two types: the original data remains on the online file system. The files can be archived to EOSCTA via xrootd. Afterward, delete the CASTOR data records and the data retrieval services change to EOSCTA.

In the other case, the original data is saved only on tape. The data is copied to temporary storage via the CASTOR command and written to EOSCTA, with the same file path behind the data name before and after, except for the different prefix, to facilitate access to the data.

As shown in Fig.7, the overall migration time starts in November 2021 and is expected to end in March 2023. It starts with the migration of the DYB experimental data, which is migrated from LUSTRE to EOSCTA. All other experimental data will need to go through a tape-to-disk, disk-to-tape copy process. The BES raw data migration has to avoid the BESIII online fetch phase. In addition, for the earlier saved YBJ, JUNO, and BACKUP data, each file was too small, so data collation was also required, using TAR to pack the data before storing it in EOSCTA.

4.2 Tape Migration Method

Retrieving data from CASTOR's tapes was the slowest part of the whole migration process. In the beginning, we used multiple streams to retrieve, i.e. copying multiple CASTOR directories

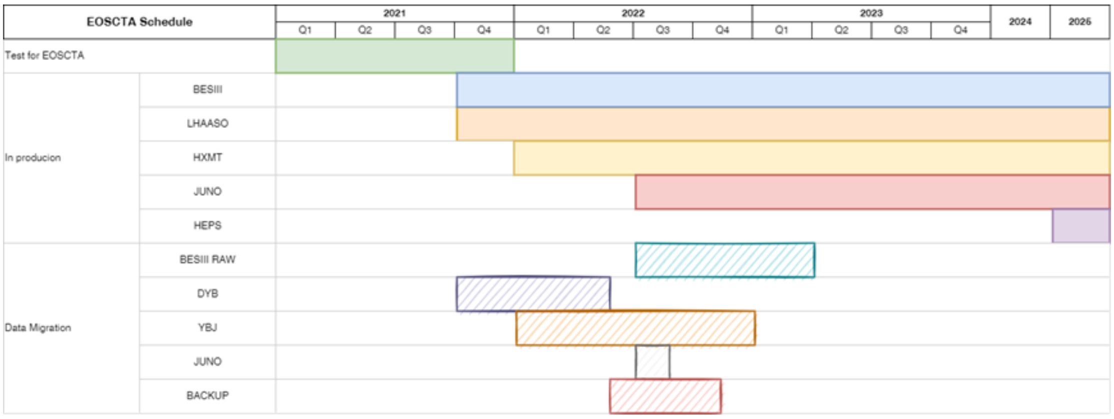

Figure 7: EOSCTA usage and data migration schedule

at once, getting a list of files, and splitting the tape copy process into multiple tapes based on the location of these files. But there were so many tape skips and the same tape being repeatedly mounted, which led to inefficient tape access.

Based on our previous understanding of tape access, and the fact that the data from these LTO4 tapes will always be retrieved, to improve data access efficiency and reduce the number of times a tape is mounted, we have adopted an alternative data migration solution. The specific steps are as follows:

1. For a tape pool need to migrate data, firstly get the tape list from the tape pool and query the drives that are now available, mount as many tapes as possible to all free drives.

2. Get a list of tapes and copy all the files on this tape to a temporary directory at once using “tpread”. Compare the original tape file with the same file in the temporary directory according to the alder32 code. If they do not match, insert the file path to the recopy list. Once the check is complete, the files in the recopy list are migrated again and compared once more. If a file is copied three times with errors, the file is recorded in the failed list.

3. Use "eos cp" to copy the files from the temporary directory to EOSCTA in the original path, e.g. /castor/ihep.ac.cn/bes/raw/round01/file in EOSCTA as /eos/bes/raw/ round01/file. Do the file integrity of temporary files and EOS files. If they are the same, delete the files in the temporary directory.

4. If all the files on a tape has been migrated and checked correctly, remove the tape’s label from the tape list and add it to the completed tape list.

5. Repeat the above process until all the tapes in the tape pool have been processed.

6. Check the CASTOR directory corresponding to the tape pool to make sure that all files have been saved to EOSCTA, and if there are any missing files, record them in the failed list as well.

After this process, most of the files have been migrated from CASTOR to EOSCTA, the unsuccessful files are saved to the failed list directory. For these files, if there is another replica

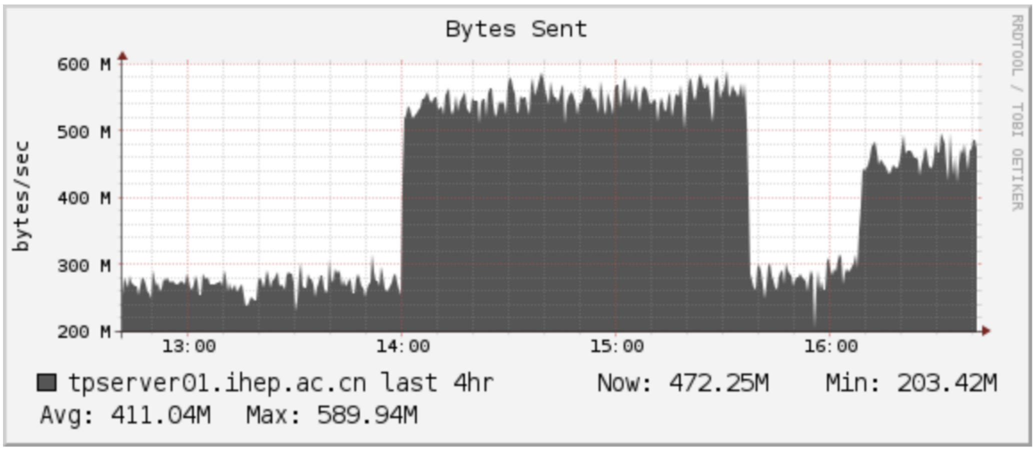

Figure 8: Network traffic of data retrieval from CASTOR

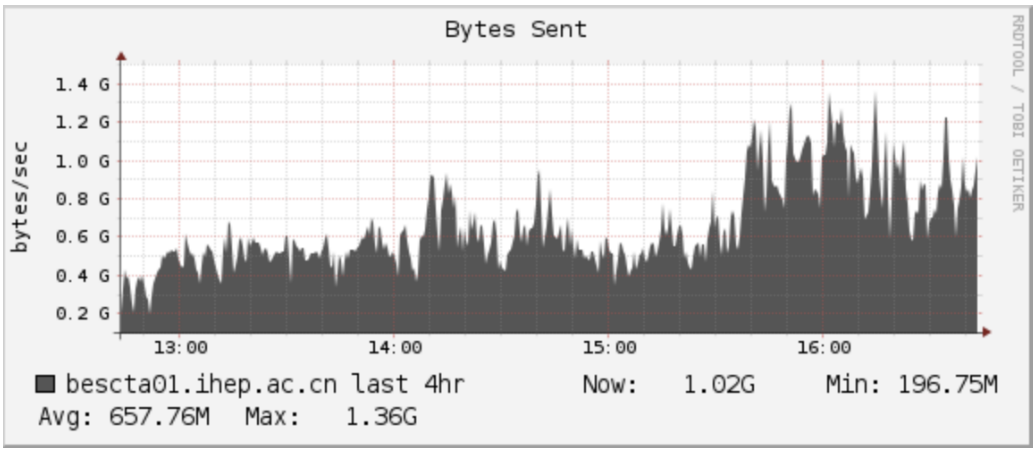

Figure 9: Network traffic of data archive to CTA

on the other tape pool, try to retrieve them from another tape, otherwise, the files can no longer be accessed and are marked with an “unreadable” flag.

The data migration is planned to take a long time, so a portion of the tape drive is specifically allocated for data migration. Fig.8 and Fig.9 show the network traffic for both phases of the data migration process. Fig.8 shows the total traffic from LTO4 tapes with up to four LTO4 drives reading data simultaneously at an average of 411MB/s.Fig.9 shows the data archive to tapes of CTA at an average speed of approximately of 657MB/s.

During the process we found that: 1. "tpread" is more efficient than "rfcp". The sequential read method can read the tapes continuously, almost to the maximum rate of the LTO4 tape. 2. Retrieving data from old LTO4 tapes has more failure rate, not only hardware failures but also some data that was read properly but the adler32 codes did not match and needed to be re-read. The EOS file checks were virtually error-free. 3. CTA accepts data archive requests continuously and distributes the requests to multiple LTO7 tapes, avoiding the previous problem of CASTOR concentrating writes to one tape. 4. Two file checks consume time and the whole migration process is not particularly fast, but we believe that data integrity is more important than speed.

Currently, we have completed the migration of half of the data of DYB and one-fifth of the YBJ data; the new experimental data generated by LHAASO and HXMT has been switched to EOSCTA since 2021. Besides the data from BES DAQ in real time, other BES data storage services have also been switched to EOSCTA.

5. Outlook

Two more experiments are going to use EOSCTA soon. The JUNO experiment will start to run in 2022. It is expected to generate 3 petabytes of raw data per year. It will run for about ten years and all experimental data will be stored in EOSCTA. Its storage workflow will be similar to the LHAASO experiment. We will adopt a new tape library for the JUNO experiment, using LTO 9 drives and tapes, with a total of 20 drives and a capacity of up to 5,000 tapes.

HEPS experiment will start to run in 2025, with about 150 petabytes of data generated per year, and will use LUSTRE and EOSCTA. It will have a dedicated data management system to manage the flow of data, including data location, archival method, and so on.

6. Conclusion

Last year we started an upgrade of long-term storage from CASTOR to EOSCTA. we have designed different data storage workflows to the requirements of the experiments. We will try to use other transmission tools, such as FTS, to implement an online file system for EOSCTA. We will study the Kerberos of CTA, the feature of Hierarchical storage management of LUSTRE, and the new feature of EOSCTA.

All IHEP experimental data require long-term storage and we want to provide efficient storage for the management of these massive amounts of data. Not only archival requirements, but also later data retrieval and analysis. The storage system should also provide high performance. The computing center has blade clusters and a job scheduling system, so some hot data may be accessed by tens of thousands of jobs at the same time. Data access frequency also varies over time. After some time, not all of the data needs to be analyzed. But these cold data should be retrieved at any time, especially the raw data cannot be deleted, the data must be kept for a long time so that scientists can do additional analyses on them years later.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 12075268). Yujiang Bi is partly supported by the Xiejialin Foundation of IHEP (No. E15461U2).

References

[1] IHEP Facilities: http://english.ihep.cas.cn/se/fs/. URL http://english.ihep.cas.cn/se/fs/.

[2] CASTOR homepage: https://castor.web.cern.ch/castor/. URL https://castor.web.cern.ch/castor/.

[3] AJ Peters, EA Sindrilaru, and G Adde. EOS as the present and future solution for data storage at CERN. *Journal of Physics: Conference Series*, 664(4):042042, dec 2015. doi: 10.1088/1742-6596/664/4/042042. URL https://doi.org/10.1088/1742-6596/664/4/042042.

[4] Michael C Davis, Vladímir Bahyl, Germán Cancio, Eric Cano, Julien Leduc, and Steven Murray. CERN Tape Archive - from development to production deployment. *EPJ Web Conf.*, 214:04015. 9 p, 2019. doi: 10.1051/epjconf/201921404015. URL https://cds.cern.ch/record/2701399.

[5] Eric Cano, Bahyl, Vladimír, Caffy, Cédric, Cancio, Germán, Davis, Michael, Keeble, Oliver, Kotlyar, Viktor, Leduc, Julien, and Murray, Steven. Cern tape archive: a distributed, reliable and scalable scheduling system. *EPJ Web Conf.*, 251:02037, 2021. doi: 10.1051/epjconf/202125102037. URL https://doi.org/10.1051/epjconf/202125102037.

[6] Eric Cano, Vladimír Bahyl, Cédric Caffy, Germán Cancio, Michael Davis, Viktor Kotlyar, Julien Leduc, Tao Lin, and Steven Murray. Cern tape archive: production status, migration from castor and new features. 245:04013, 2020.