THE 2023 INTERNATIONAL WORKSHOP ON THE HIGH ENERGY CIRCULAR ELECTRON-POSITRON COLLIDER (CEPC)

Status and Planning of High Energy Physics Data Storage System

Haibo Li

On behalf of Computing center, IHEP, CAS

2023-10-26

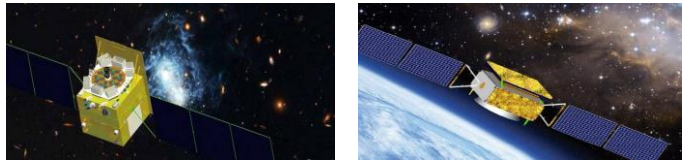
中国科学院高能物理研究所
Institute of High Energy Physics, Chinese Academy of Sciences

Outline

- Data storage requirements for HEP experiments
- Data storage system and services at IHEP
- Future Development Trends
- Summary

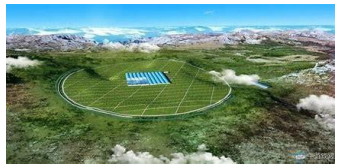
Data Volume of HEP Experiments

- BECPII/BESIII
  - ~1PB of raw data per year, accumulating 10PB+
- JUNO
  - Generate 3PB of data annually
- LHAASO
  - ~10PB of data annually
- HEPS
  - >150PB/year raw data
- HXMT/HERD/eXTP/GeCAM
  - ~10PB of data annually
- AliCPT, CSNS
- The data volume of future CEPC will be even large



~550KM

Space astronomy satellite
(HXMT, Gecam)

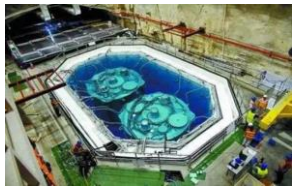



~4400M

High Altitude Cosmic Ray Observatory
(LHAASO, YBJ)


~-5M

Particle collider
(BEPCII, CSNS, HEPS)




~-700M

Underground Neutrino Experiment
(JUNO, Dayabay)

Challenges Faced by Storage System

- The storage system is full all the time
- Unlike the computing system, "opportunistic storage" is almost non-existent
- Simplifying the data structure would yield some benefits, but not significant, roughly around "30%"
- Therefore, storage is the most expensive part and biggest challenge!
- Currently, using tape storage appears to be the most cost-effective solution, but, in reality, it poses significant challenges
  - Tape usage is highly complex and slow
  - The number of tape vendors is decreasing, leading to substantial procurement risks

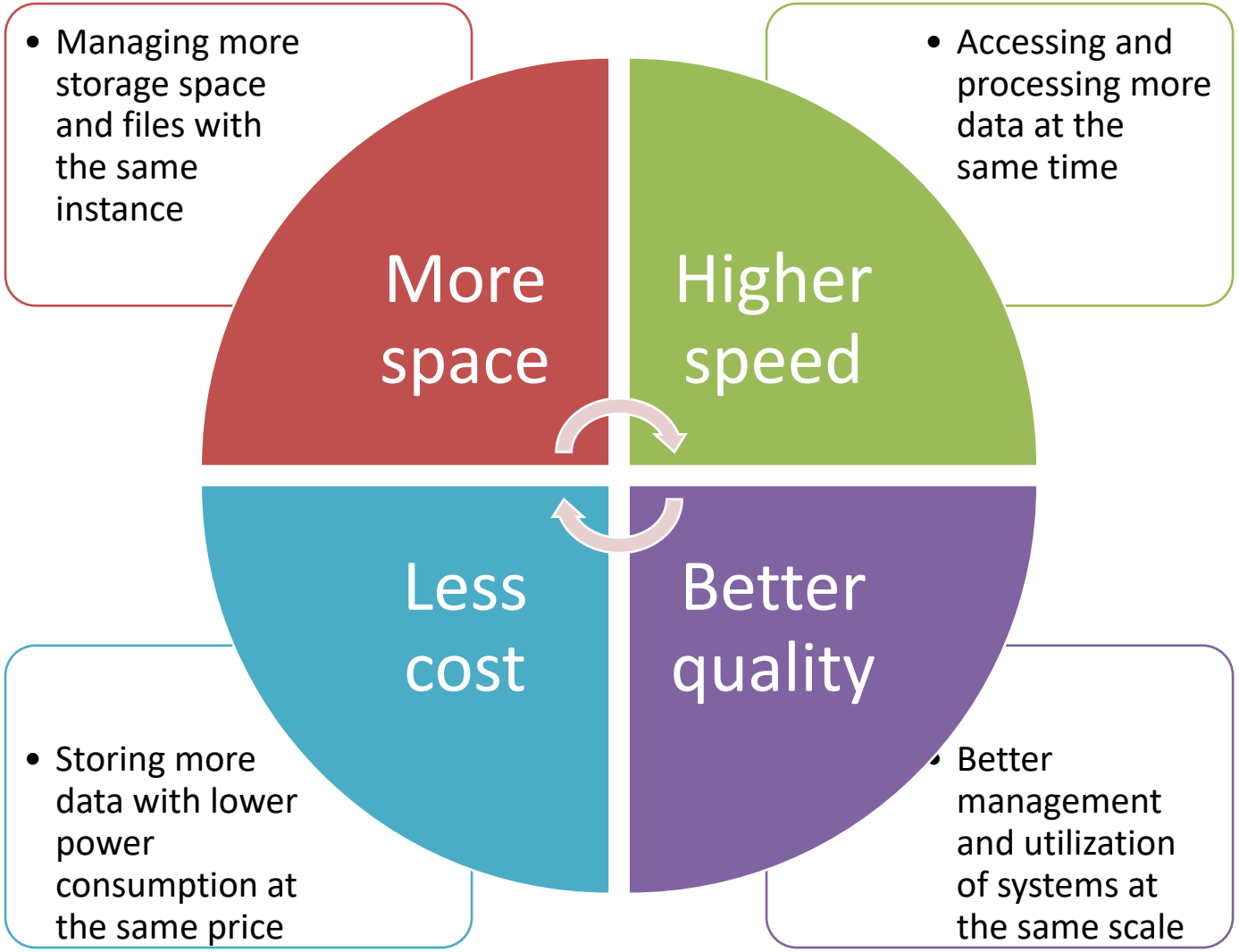
> Storage is our biggest cost component and biggest challenge.
> We need new approaches!
> ——Torre Wenaus, BNL

Storage Capacity Evaluation Indicators

Primary indicators	Secondary indicators	Tertiary indicators
Storage capacity	Inventory	Storage capacity
	Increment	Storage capacity growth rate
		Storage utilization rate
Performance metrics	Balance	Savings balance
	Agility	Advanced storage ratio
Security and reliability	Security and reliability	Disaster recovery coverage rate
		RTO * Time recovery goal
		RTO restores point goals
Green and low-carbon	Economic	Unit storage ownership cost
	Green	Storage device energy consumption level

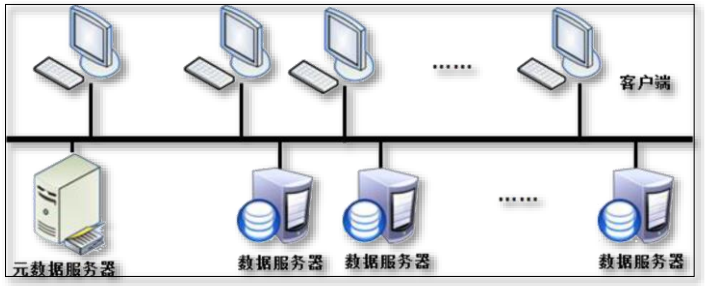
China Academy of Information and Communications Technology, China Storage White Paper, 2022.

Main Goals of A Storage System



Capacity Management

- Single instance manages more space and files, solves the problem of “fitting in”, and requires high scalability
- The essence of the problem is the architecture, including: centralized metadata server, distributed metadata server and no metadata server

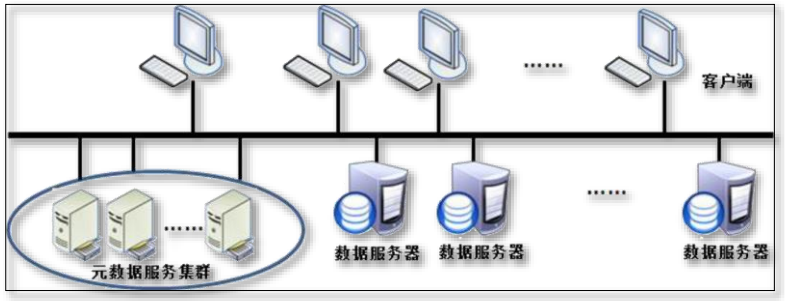


Centralized metadata server

Pros: Simple to implement, low complexity

Cons: Single point of failure, performance bottleneck

Lustre, EOS, HDFS, etc

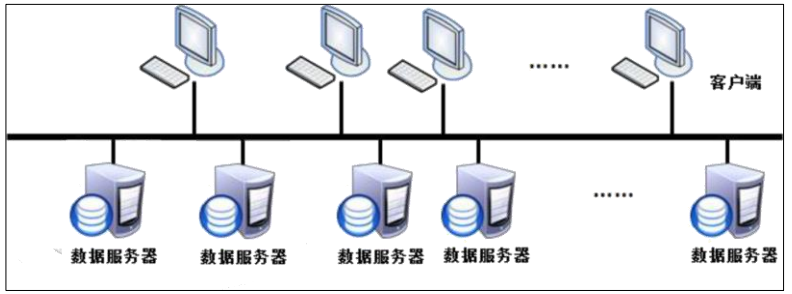


Distributed metadata server

Pros: good performance and scalability

Cons: design is complex, implementation is difficult

GPFS,Panasas,etc



No metadata server

Pros: No single point of failure and bottleneck

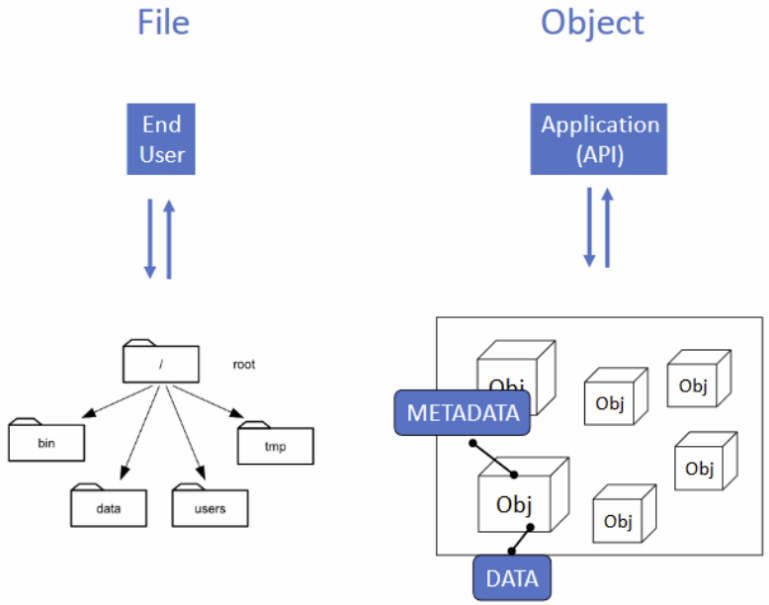
Cons: Low efficient of directory management

Glusterfs, Minio, etc

Most of the existing storage systems commonly use centralized metadata servers, which can scale to 100PB level per instance. Single instance often cannot natively support cross-data center, so it needs to rely on data transfer system FTS and other technologies

Object Storage

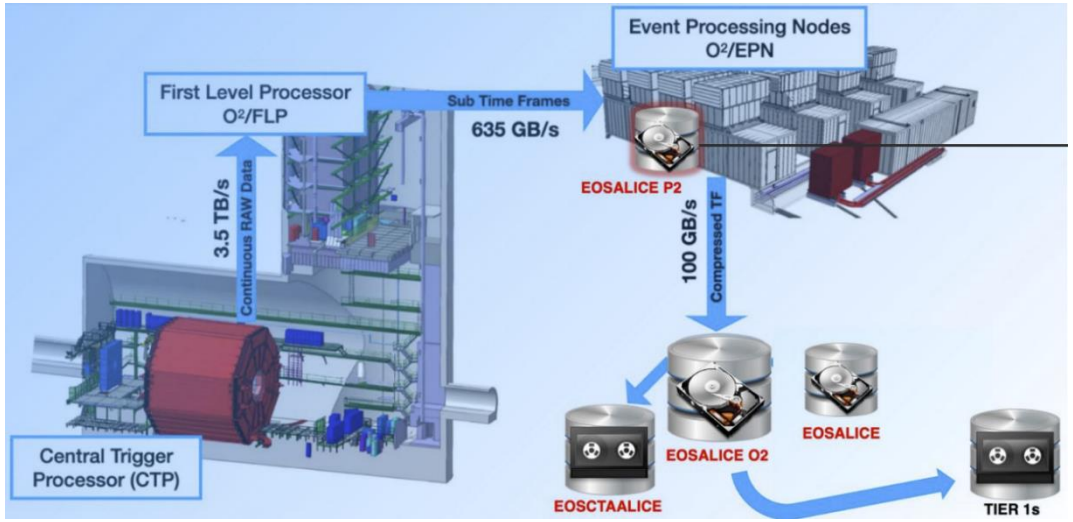
- Object-based Storage is a widely used technology in the cloud computing era, which simplifies the semantics of file system and has very high scalability
- S3(Simple Storage Service) was first proposed by Amazon, has become the de facto standard
- Main advantages of object Storage
  - Unlimited capacity: flat structure, nodes, clusters, sites independently expand capacity
  - High data reliability: redundant storage; Access control; Storage encryption; Transmission encryption
  - Easy to use: simple interface, compatible with S3 standard, rich tools
  - WAN storage and access: multi-site synchronization

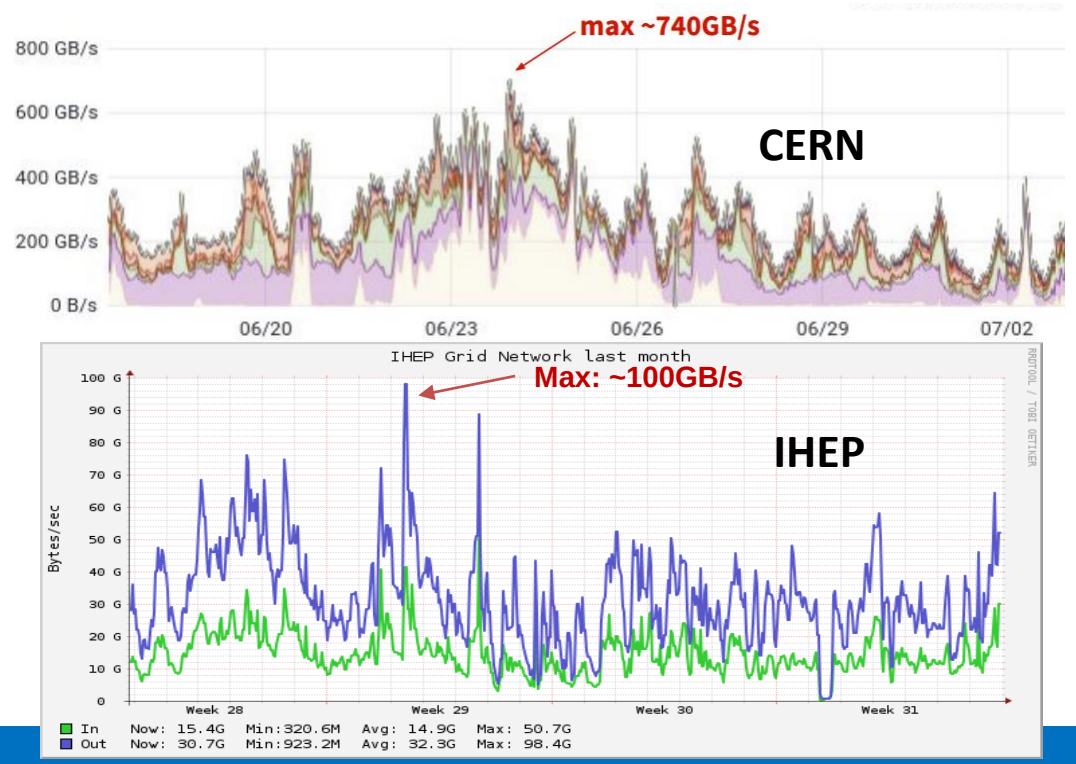


Most of the existing storage systems in the field of high energy physics use file system semantics and cannot natively support object storage. They need to rely on third-party systems such as MINIO as gateways

Data access

- Access and process more data in the same time to solve the "compute fast" problem
- The access speed is a great challenge for high-speed data acquisition (DAQ) and large-scale data analysis
- The essence of the problem is data access and performance optimization, including **storage media, storage and computing integration** and other technologies

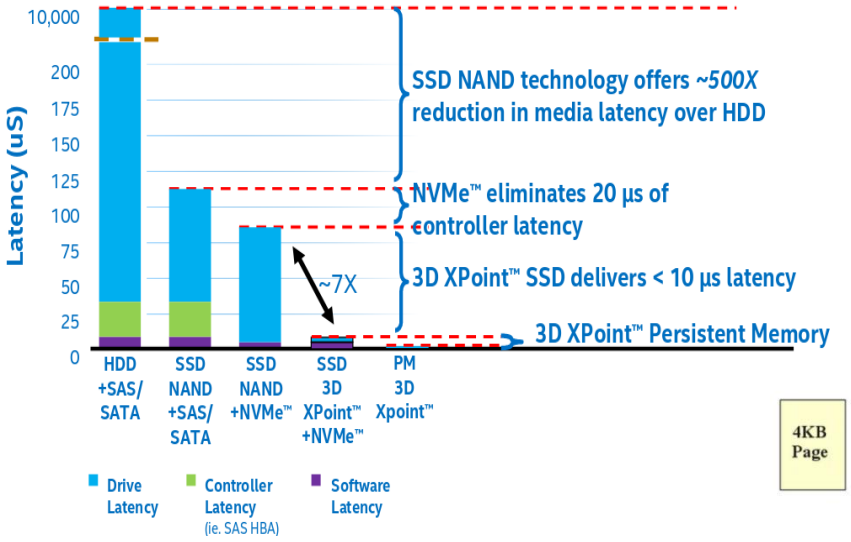


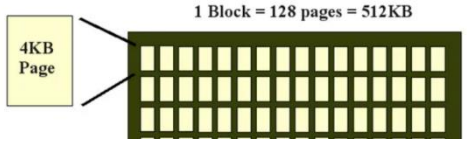


ALICE's DAQ data output is 100GB/s, and it is set 13.5PB buffer for 1.5 days

Storage media

- SSD replace HDD, bring a series of changes in storage hardware, operating system, file system, software and algorithm
  - SSD has obvious advantages in performance and power consumption, and it is estimated that the price of SSD will be equal to that of mechanical hard disk in 2026
  - "write amplification" problem, useless data must be "erased" before writing, "write a wrong word, tear the whole paper"





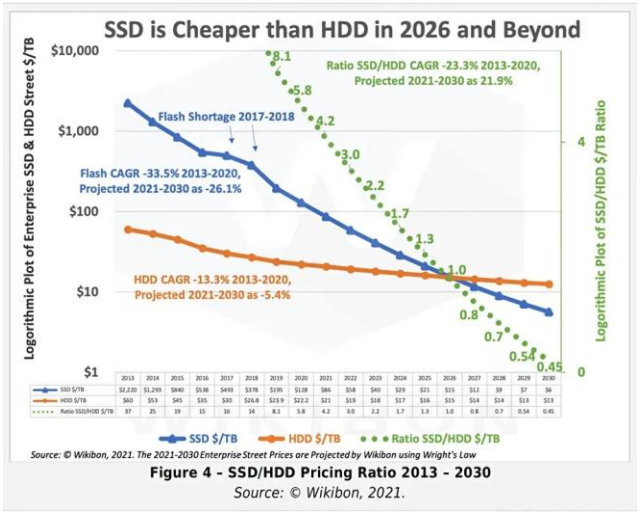


Figure 4 - SSD/HDD Pricing Ratio 2013 - 2030
Source: © Wikibon, 2021.

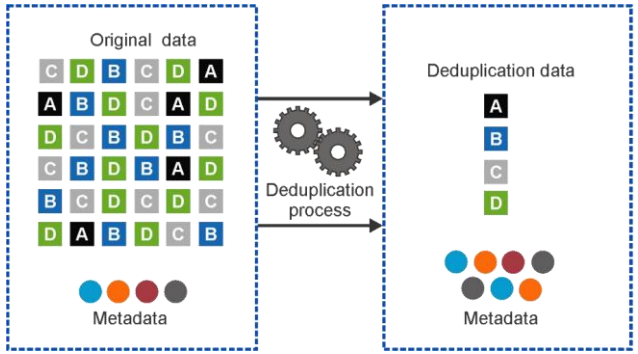
Most of the storage systems commonly used in the field of high energy physics are designed based on mechanical hard drives, and lack of optimization for SSD hard drives

Good to use

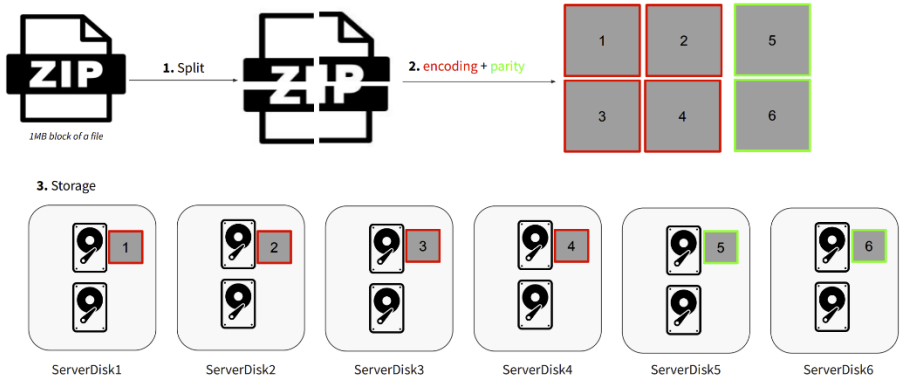
- Easy to manage, easy to use, safe and reliable, solving the problem of "good to use"
- Manageability is often overlooked and is critical in large scale storage systems
- Standardized file system interface and application program interface
  - POSIX interfaces: strongly consistent semantics, fully compatible with filesystem interfaces
  - XRootD: supports ROOT data analysis software, which can remotely open and access files
  - MPI-IO: Parallel file read/write IO
  - HTTP/S3: Remote access or object storage interface, mainly supports upload and download, difficult to support random read and write
- Secure and reliable
  - Data redundancy is commonly used, including RAID/RAIN/ replica, etc
  - Security certification: Kerberos, X.509, SciTokens, etc
  - Data validation, data encryption, etc

Data reduction

- Data deduplication, data compression, erasure coding, etc
- Deduplication and erasure coding try to remove redundant information from data
- Erasure code is a kind of redundant coding, which can realize any N+M data coding
  - It is more flexible than traditional RAID technology, for example, it can use 6+2, 5+3, 12+2 and so on
  - After disk failure, it can be rebuilt faster than RAID
- IHEP and CERN have begun to use erasure code technology



Data deduplication



Erasure codes

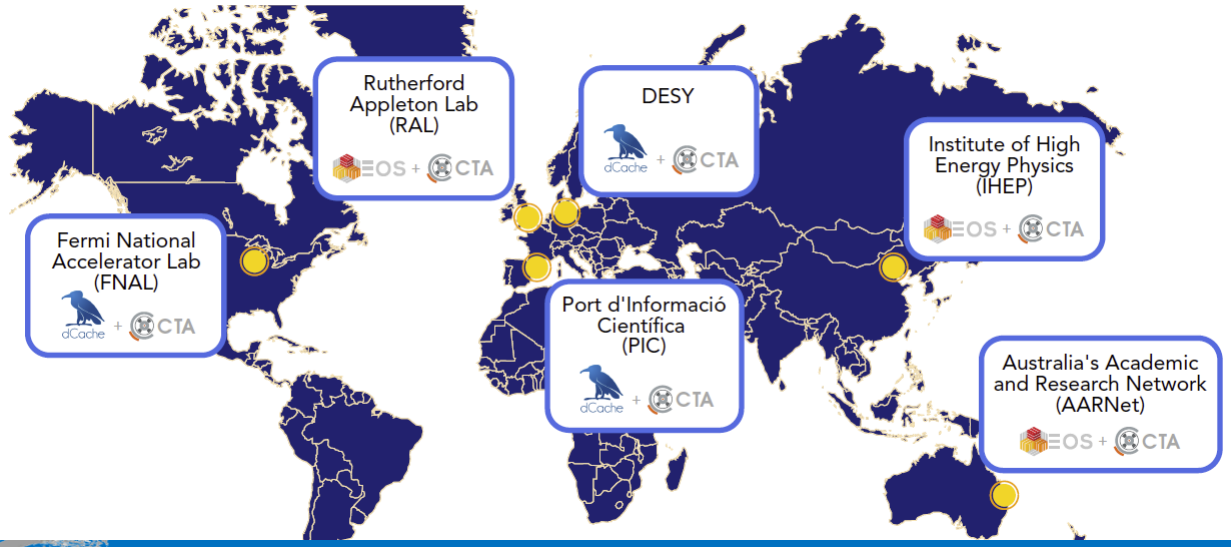
	Value
Aliceo2 Savings with Erasure Coding 10+2	55.3 PB
AMS Savings with Erasure Coding 8+2	7.37 PB
CMS Savings with Erasure Coding 10+2	1.12 PB

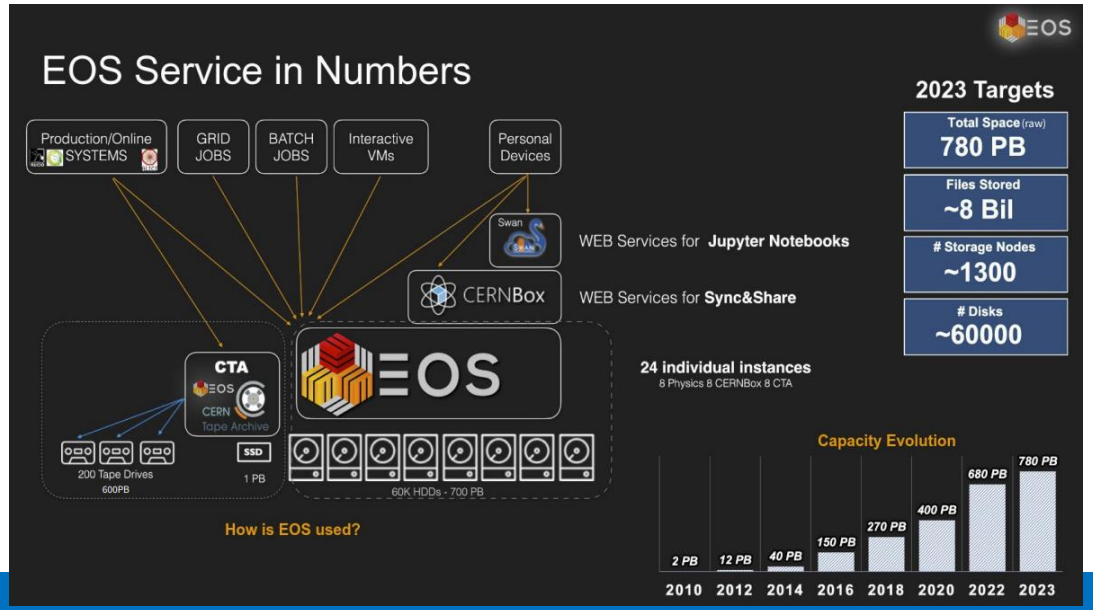
Total: 63.79PB

CERN uses erasure codes to save space

Storage systems commonly used in high-energy physics

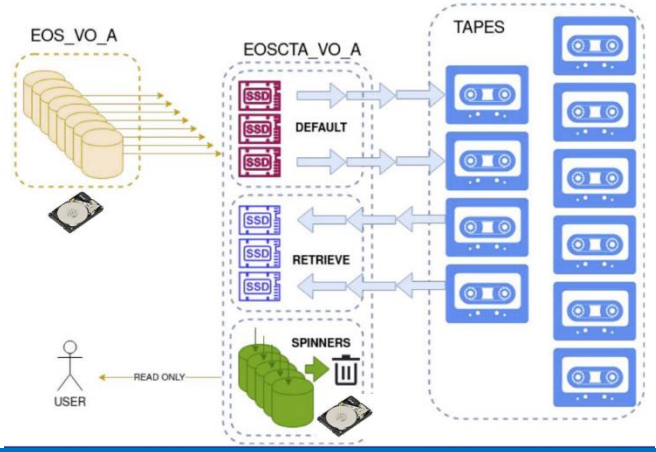
- Lustre: an open source parallel file system for industry
  - Adopted by more than 70% of the world's supercomputing centers
  - IHEP and GSI are the largest units deploying LUSTRE in the field of high energy physics
- EOS: open source storage system developed at CERN
  - CERN's main disk storage system, up to 780 petabytes
  - Based on XRootD protocol, with good scalability and performance
  - Besides CERN, IHEP is the largest EOS deployment site, approaching 100PB

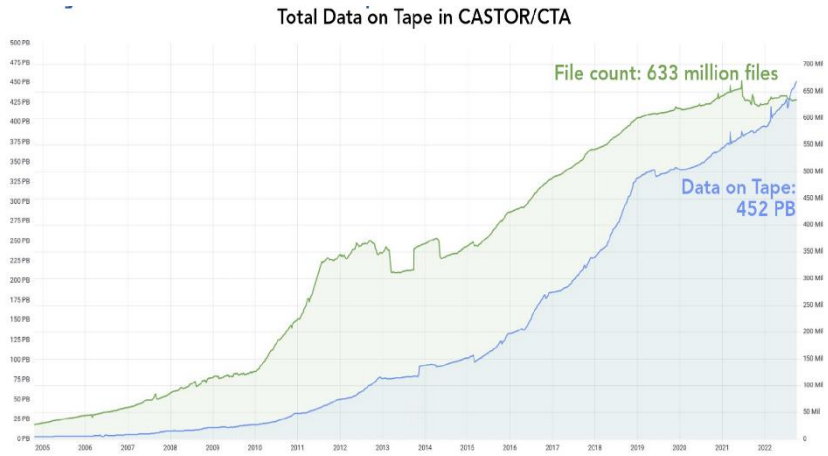




Storage Systems Commonly used in High Energy Physics (II)

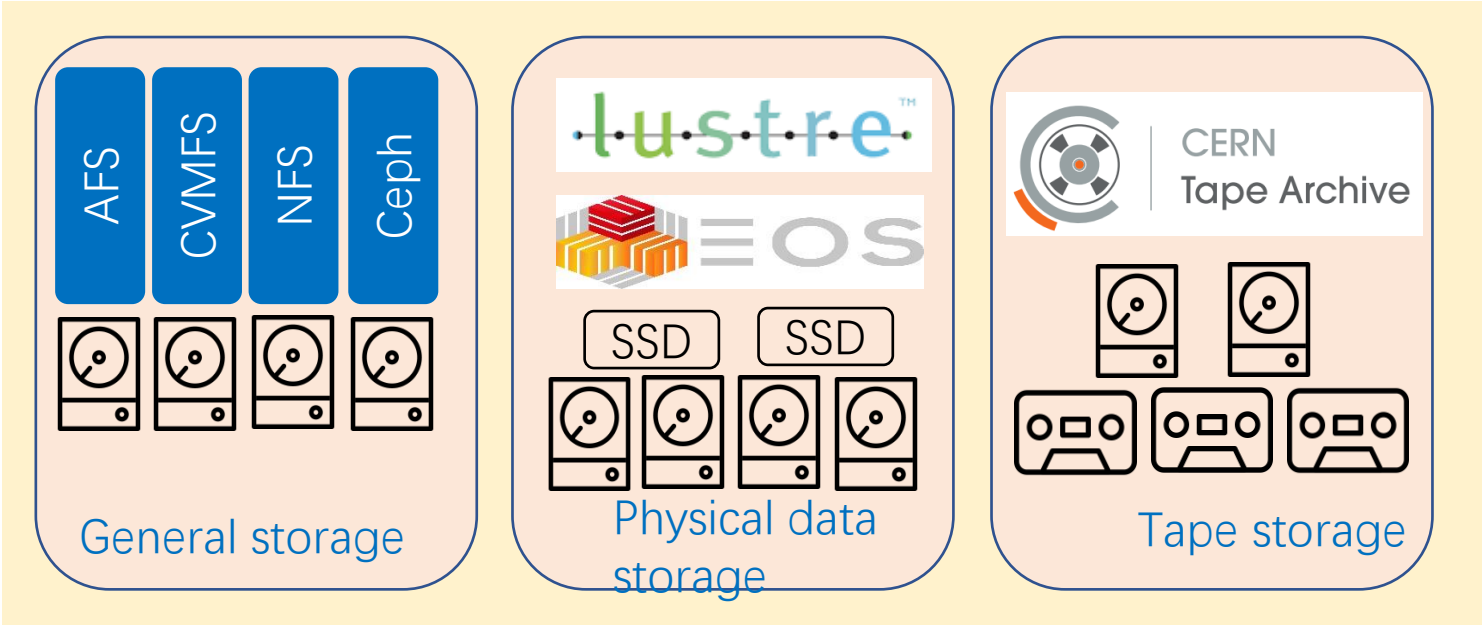
- dCache is an open source storage system developed by DESY and FNAL
  - For data-intensive science, it is widely used in LHC grid sites
  - It supports high-speed data acquisition, batch processing, interactive analysis and wide area network transmission
- CTA: An open source tape library storage system developed by CERN
  - Based on the previous generation system CASTOR
  - It supports EOS/ dCache disk storage system
  - Almost the only open source tape library management system software in the field of high energy physics





Stauts of IHEP Storage Service

- General Data Storage
  - HOME directory, software storage service, personal cloud disk, block storage
- Experimental data storage
  - Disk data storage
- Long-term storage and backup
  - Tape library storage
  - Backup system

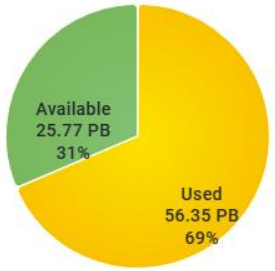


An overview of IHEP data storage system

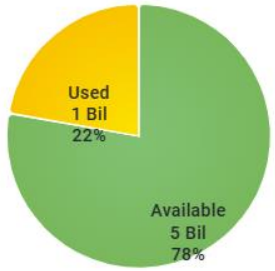
storage overview

Read Throughput	Write Throughput	Current Metadata IOPS	Total Space	Used Space	Total Files
8.53 GB/s	846 MB/s	18.1 K	102 PB	68.9 PB	1,545,620,533

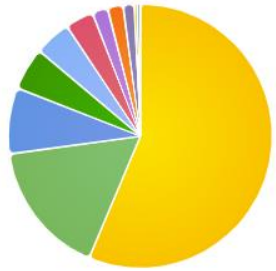
Space Usage



Inode Usage

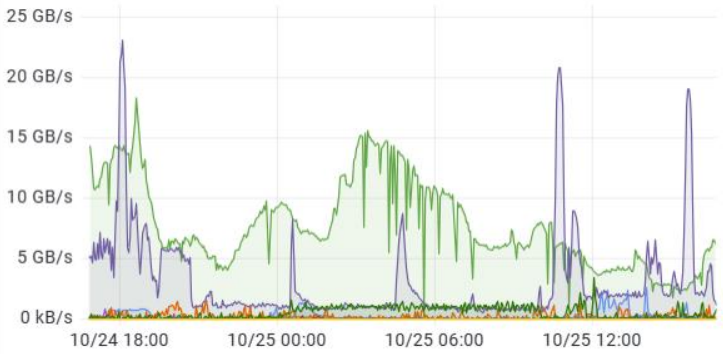


Space Distribution



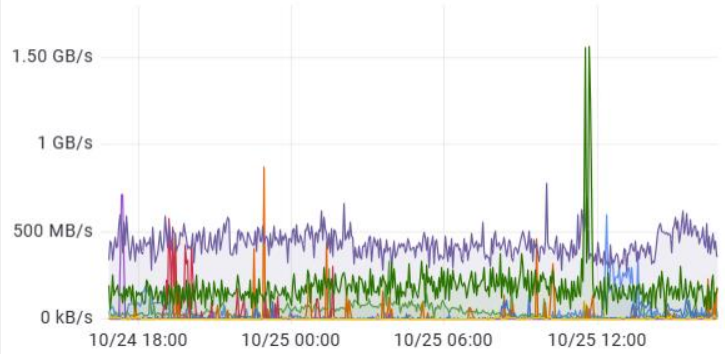
	Value
LHAASO	38 PB
BES	11 PB
JUNO	5 PB
OTHER	3 PB
HXMT	3 PB
CEPC	2 PB
HPC	1 PB
DYB	1 PB

Read Throughput



	Mean	Last *	Max
BES	7.74 GB/s	6.24 GB/s	18.4 GB/s
DYB	41.6 MB/s	200 MB/s	253 MB/s
JUNO	246 MB/s	32.3 MB/s	2.64 GB/s
Ali	237 MB/s	0 kB/s	1.56 GB/s
MBH	8.71 MB/s	172 kB/s	400 MB/s
HXMT	29.5 MB/s	71.5 kB/s	216 MB/s
HEPS	3.75 MB/s	0 kB/s	849 MB/s
LHAASO-ihep	3.08 GB/s	1.18 GB/s	23.2 GB/s
LHAASO-lhmt	543 MB/s	879 MB/s	3.41 GB/s
PUBLICEOS	2.33 MB/s	649 kB/s	105 MB/s

Write Throughput



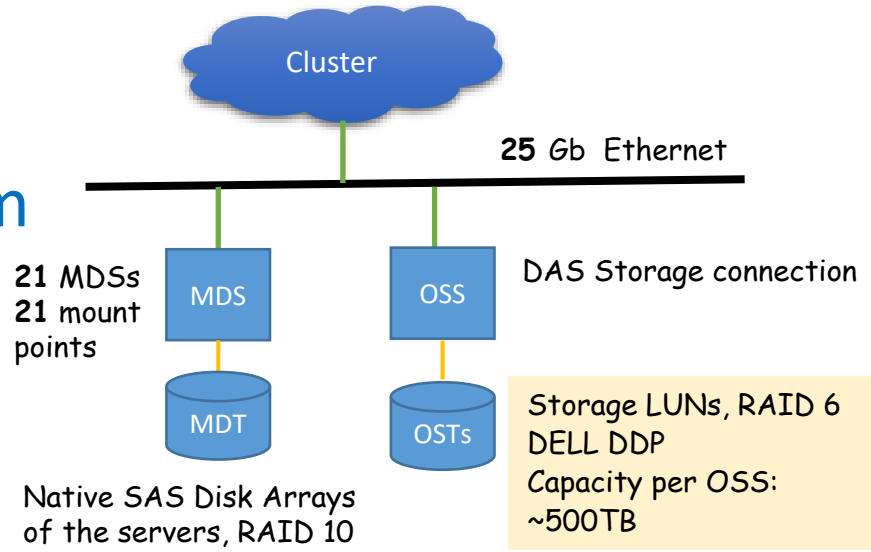
	Mean	Last *	Max
BES	42.1 MB/s	4.07 MB/s	333 MB/s
DYB	1.12 MB/s	20.5 MB/s	100 MB/s
JUNO	32.4 MB/s	19.7 MB/s	596 MB/s
Ali	15.5 MB/s	177 MB/s	872 MB/s
MBH	17.8 MB/s	299 kB/s	576 MB/s
HXMT	10.3 MB/s	13.8 MB/s	135 MB/s
HEPS	6.09 MB/s	36 kB/s	718 MB/s
LHAASO-ihep	434 MB/s	369 MB/s	782 MB/s
LHAASO-lhmt	182 MB/s	179 MB/s	1.56 GB/s
PUBLICEOS	2.08 MB/s	9.99 MB/s	21.4 MB/s

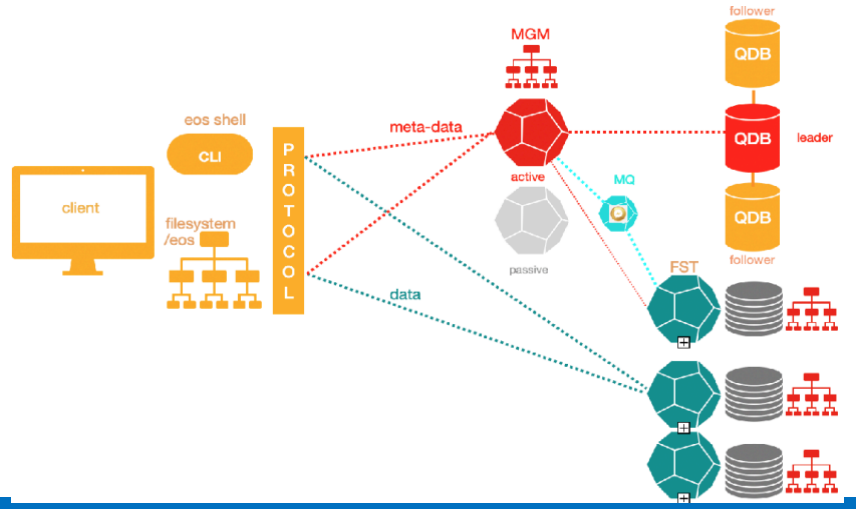
General Storage Service

- HOME Directory
  - Small files, high concurrent access
  - AFS file system, used for many years, official maintenance has stopped
  - Explore trying to use a commercial file system
- CVMFS software storage system
  - 6 warehouses, 14 terabytes
  - Deploy software repositories for high energy physics experiments
  - Synchronize CERN and EGI software repositories
- Cloud platform storage service
  - CEPH is used to provide block storage for Openstack platforms

Experiment with data storage services

- Experimental data storage system is the main storage system for high energy physics data
  - Large capacity, high bandwidth and high reliability
- The software uses open source distributed file system
  - Lustre
    - 40PB+ capacity
    - Mainly used for BESIII, JUNO, HXMT,GeCAM, etc
  - EOS
    - 50PB+ raw capacity
    - Mainly used for LHAASO and JUNO
- Hardware
  - Adopt storage server + disk array mode
  - Disk array
  - Disk expansion cabinet

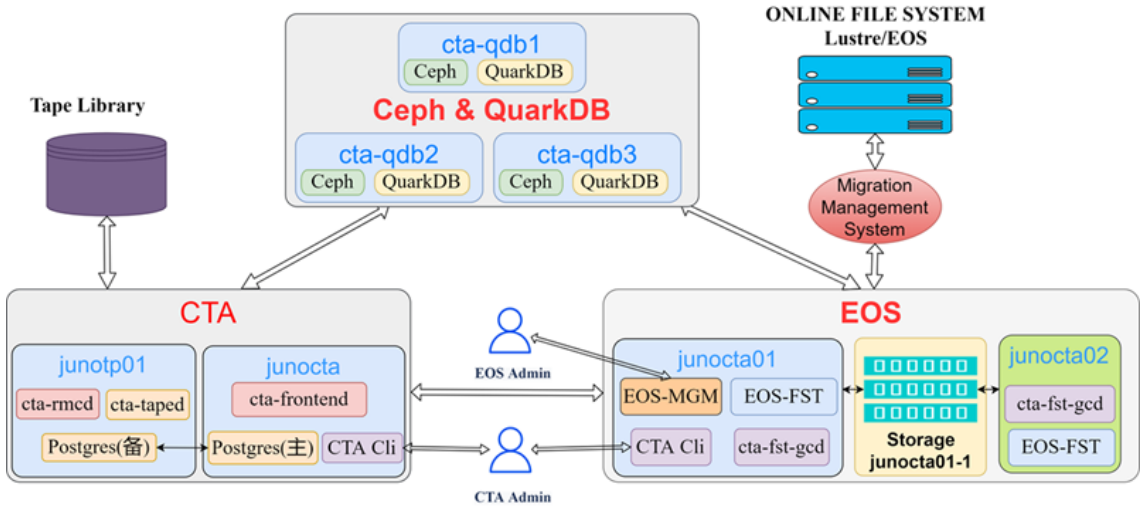




Tape storage and backup services

- CTA: A new generation open source tape storage system developed at CERN
  - Supports BESIII, LHAASO, JUNO, HEPS
  - Supports LHC b Tier-1
- Backup services
  - Amanda, Restic
  - HOME directory, mail data and other key data

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



Future outlook

- Computing and storage software and hardware are rapidly evolving and changing
  - NVMe SSD, heterogeneous computing, parallelization, distributed,...
  - Object storage is the de facto standard of cloud storage worldwide
- Data storage systems need to change with them
  - SSD-optimized, distributed, and object storage
- High-energy physics case models need to evolve
  - TTree → RNtuple, smaller storage space, better parallelization
  - Supports DAOS, 8GB+ write and 4GB+ read per second on a single node
  - S3 protocol is supported for efficient cross-data center migration
- We have the opportunity to develop the next generation of storage technologies and systems
  - PaticleFS is being designed and planned
  - SSD-optimized storage, natively supported object storage, and across data center sites, computable storage, parallel data access model,...

Summary

- With the development of large-scale experiments in high energy physics, the data explosion has become one of the biggest technical challenges in the future
- The storage hardware is undergoing a major transformation, and the basic software such as operating system and storage system needs to change with it, which is both a challenge and an opportunity
- Focusing on the needs of the next generation of high energy physics big science and engineering, such as CEPC, to realize the technology innovation and guidance of IT

Thank you for your attention

lihaibo@ihep.ac.cn