JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 10(Special): 56-65, 2022

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2022.10.S.6

A Disk-based Archival Storage System Using the EOS Erasure Coding Implementation for the ALICE Experiment at the CERN LHC

Sang Un Ahn*
Korea Institute of Science and Technology Information (KISTI), Global Science experimental Data hub Center (GSDC), Daejeon, Korea
E-mail: sahn@kisti.re.kr

Latchezar Betev
European Organization for Nuclear Research (CERN), Geneva, Switzerland
E-mail: Latchezar.Betev@cern.ch

Eric Bonfillou
European Organization for Nuclear Research (CERN), Geneva, Switzerland
E-mail: Eric.Bonfillou@cern.ch

Heejune Han
Korea Institute of Science and Technology Information (KISTI), GSDC, Daejeon, Korea
E-mail: hjhan@kisti.re.kr

Jeongheon Kim
Korea Institute of Science and Technology Information (KISTI), GSDC, Daejeon, Korea
E-mail: jh.kim@kisti.re.kr

Seung Hee Lee
Korea Institute of Science and Technology Information (KISTI), GSDC, Daejeon, Korea
E-mail: benecia@kisti.re.kr

Bernd Panzer-Steindel
European Organization for Nuclear Research (CERN), Geneva, Switzerland
E-mail: Bernd.Panzer-Steindel@cern.ch

Andreas-Joachim Peters
European Organization for Nuclear Research (CERN), Geneva, Switzerland
E-mail: Andreas.Joachim.Peters@cern.ch

Heejun Yoon
Korea Institute of Science and Technology Information (KISTI), GSDC, Daejeon, Korea
E-mail: k2@kisti.re.kr

ABSTRACT

Korea Institute of Science and Technology Information (KISTI) is a Worldwide LHC Computing Grid (WLCG) Tier-1 center mandated to preserve raw data produced from A Large Ion Collider Experiment (ALICE) experiment using the world's largest particle accelerator, the Large Hadron Collider (LHC) at European Organization for Nuclear Research (CERN). Physical medium used widely for long-term data preservation is tape, thanks to its reliability and least price per capacity compared to other media such as optical disk, hard disk, and solid-state disk. However, decreasing numbers of manufacturers for both tape drives and cartridges, and patent disputes among them escalated risk of market. As alternative to tape-based data preservation strategy, we proposed disk-only erasure-coded archival storage system, Custodial Disk Storage (CDS), powered by Exascale Open Storage (EOS), an open-source storage management software developed by CERN. CDS system consists of 18 high density Just-Bunch-Of-Disks (JBOD) enclosures attached to 9 servers through 12 Gbps Serial Attached SCSI (SAS) Host Bus Adapter (HBA) interfaces via multiple paths for redundancy and multiplexing. For data protection, we introduced Reed-Solomon (RS) (16, 4) Erasure Coding (EC) layout, where the number of data and parity blocks are 12 and 4 respectively, which gives the annual data loss probability equivalent to 5×10^{-14}. In this paper, we discuss CDS system design based on JBOD products, performance limitations, and data protection strategy accommodating EOS EC implementation. We present CDS operations for ALICE experiment and long-term power consumption measurement.

Keywords: Worldwide LHC Computing Grid Tier-1, A Large Ion Collider Experiment, Custodial Disk Storage, Exascal Open Storage, Erasure Coding

Received: April 21, 2022 **Accepted:** May 4, 2022
Published: June 20, 2022

***Corresponding Author:** Sang Un Ahn
https://orcid.org/0000-0001-8847-489X
E-mail: sahn@kisti.re.kr

1. INTRODUCTION

The Large Hadron Collider (LHC), the world's largest particle accelerator built and operated by the European Organization for Nuclear Research (CERN), helps scientists explore the nature of the universe through experiments such as ALICE, ATLAS, CMS, and LHCb. These experiments have produced several tens of petabytes of data in a year while they foresee several hundreds of petabytes of data for the coming years after the upgrade of experiments as well as of the LHC. In order to store, process, and analyze such large amounts of data, the Worldwide LHC Computing Grid (WLCG) provides an inter-continental computing infrastructure to the LHC experiments in collaboration with 170 institutes, universities, and computing centers around the globe. In general, the raw data produced from the experiments are transferred to CERN Data Center at first, which is the Tier-0 of the WLCG, and their shards are replicated to the several Tier-1 centers synchronously or asynchronously for post-processing and long-term preservation.

Providing a custodial storage is crucial for these tiers to protect the data safely, and a typical choice for persistent media is tape (Colarelli & Grunwald, 2002), thanks to its reliable durability and the least price per storage capacity (less than 10 USD per terabyte), compared to other media such as magnetic disk (less than 25 USD per terabyte) and optical technology (about 50 USD per terabyte) (Spectra Logic, 2021). However, to efficiently operate the tape-based archival storage in a production environment and complement its slow data staging process requires significant additional costs such as a Hierarchical Storage Management (HSM) system including cache or buffer storage, organized accesses, and dedicated human efforts with a certain level of expertise on these system complexities. Furthermore, a decreasing number of manufacturers of enterprise-class tape drives and tape cartridges and patent disputes among the suppliers have escalated the risk of the market.

Thus, as a WLCG Tier-1 center, the Global Science experimental Data hub Center (GSDC) at Korea Institute of Science and Technology Information (KISTI) has launched a project in collaboration with CERN Information Technology (IT) department and the ALICE experiment to provide an alternative to the tape-based data preservation strategy and we proposed a disk-only erasure-coded archival storage system, the Custodial Disk Storage (CDS), powered by EOS Open Storage, which is an open-source storage management software developed by the storage group of CERN IT (Peters et al., 2015).

Our goal is to make the CDS simple, reliable, and cost-effective compared to the existing tape library at KISTI before the start of new data taking (LHC Run 3) scheduled in 2022. In this paper, we mainly discuss the CDS system design based on high density Just-Bunch-Of-Disks (JBOD) products available in the domestic market, performance limitations, and its data protection strategy accommodating the erasure coding (EC) implementation of EOS. Then, we present the operations of the CDS as a production custodial storage service for the ALICE experiment as well as a remarkable result on power consumption measurement done in the long-term period.

This paper is organized as follows: design of the CDS in Section 2, deployment of EOS EC storage in Section 3, custodial storage service for the ALICE experiment in Section 4, and long-term power consumption measurement presented in Section 5, followed by conclusions with future planning and perspectives in Section 6.

2. DESIGN OF THE CDS

The principle of the CDS as an alternative to the tape-based archival storage system is to replace the existing 3+ petabytes of tape library, IBM TS3500, at GSDC-KISTI and the commercial solutions for the HSM, such as Spectrum Protect (the former Tivoli Storage Manager) and Spectrum Scale (General Parallel File System), with cheap off-the-shelf equipment and open-source storage solutions. We expect this would substantially simplify setup and reduce operational costs compared to the tape-based system as well as provide an acceptable quality of services in terms of raw data preservation required by the ALICE experiment. In this section, we briefly discuss the high density JBOD products in the domestic market, performance limitations inherited by the interfaces, system setup of the CDS, and its data protection strategy with the EC implementation of EOS.

2.1. High Density JBOD Products

We define the JBOD products capable to host more than 60 disks in a single chassis as *High Density* JBOD here and these products were chosen for higher storage capacity per value in comparison to tape. In South Korea, the price per terabyte for tape including the cost of tape drives, frames, robotics, and commercial software solutions for the HSM were about 160 USD per terabyte in 2015 and 140 USD per terabyte in 2019. Note that this did not consider the cost for 600 TB of disk buffer. For the high density JBOD products, the number of magnetic disks ranged from 60 to 102 in 2019's domestic market and the storage capacity per box easily exceeded 1 PB with 12 TB of hard disks. The price per terabyte of high density JBOD products in 2020 went down to 55 USD including servers and interconnections.

2.2. Bottleneck Study

The high density JBOD enclosures are attached to servers via 12 Gbps SAS HBA interfaces. Despite the fact that the custodial storage does not require high performance on I/O, the throughput of CDS should be at least 2 GB/s, which is the throughput of the tape library at KISTI. The third-generation SAS HBA interface is capable of transferring data in a single lane at the speed of 12 Gbps and each port of a SAS HBA card is composed of four lanes. Therefore, a typical 2-port card can transfer data at a nominal speed of 9.6 GB/s while a 4-port model with 16 lanes reaches at most 19.2 GB/s. However, a study by Broadcom (2013) reported that the transfer speed is barred under 8 GB/s and is precisely 6.4 GB/s when considering 80% of efficiency, due to the maximum bandwidth of PCI-e 3.0 because generally an SAS HBA card is interfaced with a PCI-e 3.0 slot. Ahn et al. (2020) confirmed this cap by observing 6 GB/s of peak I/O performance in multi-bus mode (allowing simultaneous read/write through multiple paths provided by SAS HBA interfaces) with an 84-disk JBOD model hosting 70 spinning disks at moderate speed (Fig. 1).

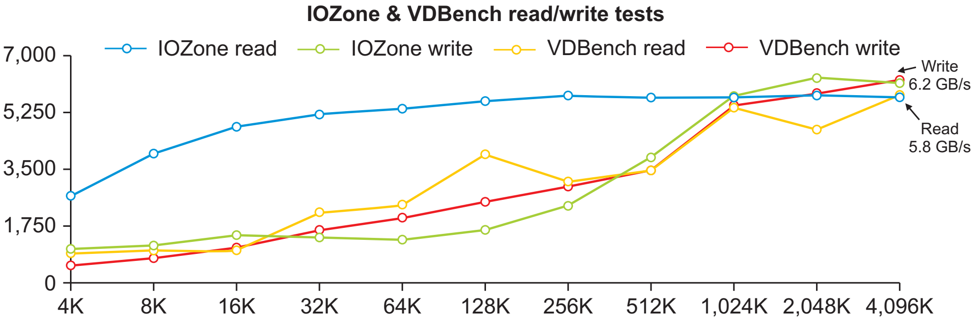


Fig. 1. Both results of IOZone and VDBench on reading and writing through SAS HBA reached performance cap around 6 GB/s due to the maximum bandwidth provided by PCI-e 3.0 (Ahn et al., 2020).

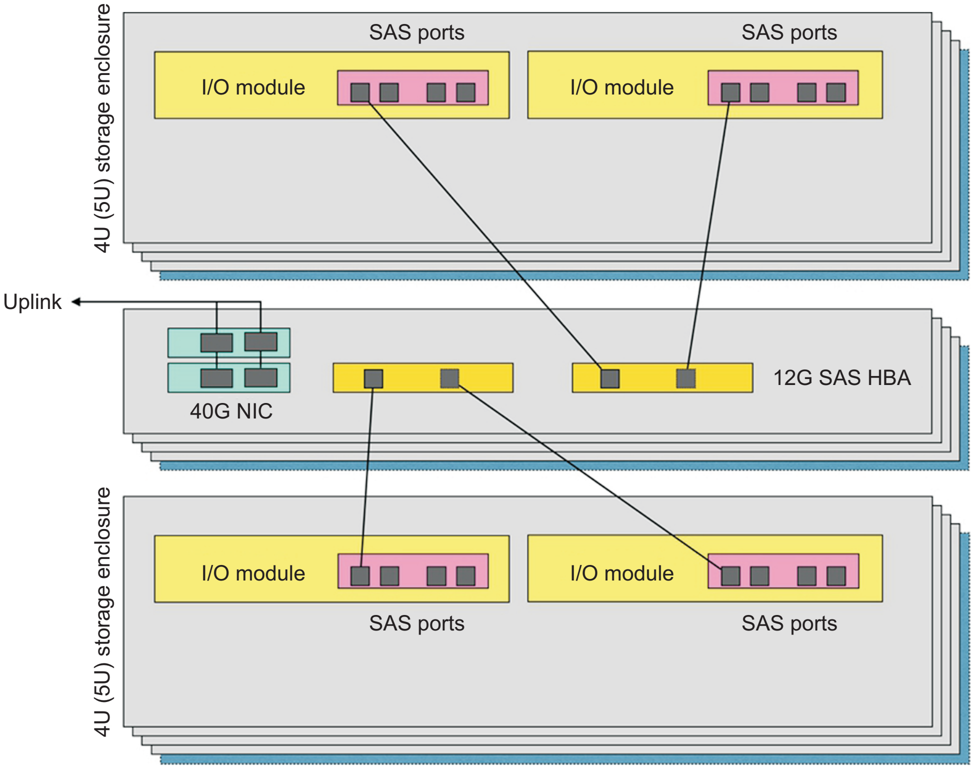


Fig. 2. Schematic view of physical connection of JBOD enclosures and servers. Each of two SAS HBA controllers are connecting to enclosures in dual-domain connection. JBOD, Just-Bunch-Of-Disks.

2.3. Hardware Configuration of the CDS

In addition to take advantage of storage capacity per value and significant throughput performance of high density JBOD products, avoiding any single point of failure in hardware configuration is crucial for data security and service continuity. Thus, the CDS consists of 18 high density JBOD enclosures, which amounts to 18 PB of raw storage capacity, directly attached to 9 servers through each of two 12 Gbps SAS HBA cards, as shown in Fig. 2. It is analogous to a typical disk storage configuration and we anticipate that it simplifies operation of the CDS as an archival storage.

2.4. Data Protection Strategy

EC is a method of data protection in which data is broken into fragments, expanded and encoded with redundant data pieces and stored across a set of different locations of storage media. The EC algorithm implemented in the EOS is the Reed Solomon (RS) variant of JERASURE library (Peters et al., 2020). It allows configuration of EC for data durability on a file basis and supports the following four redundancy levels depending on the quality of services with different parity configurations: *RAID5* (single parity), *RAID6* (double parity), *ARCHIVE* (triple parity), and *QRAIN* (quadruple parity). As the level of data protection increases along with the number of parities, the QRAIN configuration is capable of providing the highest protection on the stored data in an EOS storage system.

For the CDS, we decided to accommodate a QRAIN configuration with total 16 stripes and it can be described as *RS* (m+k, k) = *RS* (16, 4), where k is the number of data and m is the number of parity blocks. A method by Arslan (2014) suggests that data loss probability of erasure coded storage systems can be derived as follows:

$$p = e^{-\lambda}\frac{\lambda^{k+1}}{(k+1)!}$$

, where

$$\lambda = \frac{AFR\ \times (m+k)}{365 \times 24 \div MTTR}$$

For the CDS QRAIN configuration as an example, assuming m+k=16, 2% of Annualized Failure Rate (AFR), 156 hours of Mean-Time-To-Repair (MTTR) for disk failure, the value of λ=0.0057, and with 4 parity disks, the formula gives the probability p as,

$$p = e^{-0.0057}\frac{0.0057^5}{5!} = 5 \times 10^{-14}$$

Note that MTTR depends on best practices in terms of maintenance cycle for disk repair, and AFR can be varied with the conditions such as room temperature, humidity, and electric power stability. Although this is a theoretical estimation, it implies that the risk of data loss in the CDS can be extremely low.

Fig. 3 shows a schematic view of the *RS* (16, 4) layout configuration applied to the CDS. Two JBOD enclosures are attached to one server and 2 FSTs (file servers) of the EOS running on the server manages the two enclosures. This layout stores fragmented data on 16 FSTs with 4 parity nodes while the 2 remaining nodes provide room for daily operations and maintenance by allowing one of 9 servers to be shut down at any convenience without causing service discontinuity. Also, it is cost-effective compared to the case of deploying 18 FSTs on 18 servers. However, one should note that the usable space is reduced by a ratio of the number of parities plus spare nodes to the total number of nodes. The reduction factor of the usable space for *RS* (16, 4) with 2 spares is $R = \frac{4+2}{18} \approx 0.333\ldots$ and

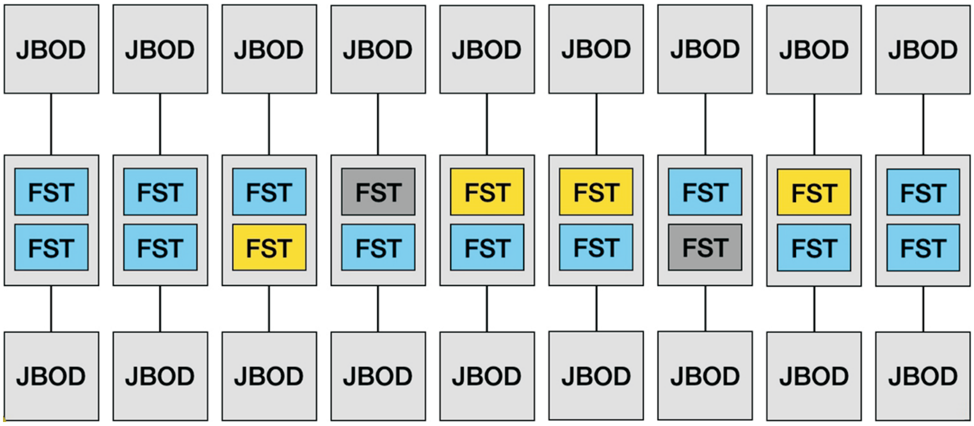


Fig. 3. Schematic view of the *RS* (16, 4) layout configuration with 18 stripes with 4 parities and 2 spares. Blue, yellow, and dark gray squares represent data, parity, and spare nodes, respectively. FST, File Storage Server, JBOD, Just-Bunch-Of-Disks.

therefore, the usable capacity of the CDS is 12 PB, which is about 67.77% of 18 PB.

3. DEPLOYMENT OF EOS EC STORAGE

With the design principles discussed in the previous sections, we deployed the EOS storage system upon the hardware procured in 2019 and Table 1 lists the hardware types and their brief specifications. The procured high density JBOD enclosure is capable of hosting up to 84 disks and it can afford to provide 1 PB of storage capacity per enclosure with 12 TB hard disks. Thus, the total amount of raw storage capacity of the CDS is precisely 18,144 TB. In this section, we discuss some technical points on EOS deployment upon the procured hardware and EC configuration for the CDS in detail.

3.1. EOS Deployment

The essential EOS components are MGM (Manager), MQ (Message Queue), QDB (QuarkDB for namespace), and FST (File Storage Server). MGM exposes an endpoint of the EOS storage system and redirects the data I/O requests from clients to FSTs, in which the requested I/O is handled by the assigned FST directly without going through the MGM. MQ is used for communication among the MGM and FSTs. A cluster of QDB instances running in raft mode provides the namespace for the stored data in a consistent way. The details regarding the EOS open storage system can be found in Peters and Janyst (2011), and Peters et al. (2015). Sindrilaru et al. (2017) describe the developments of EOS in adopting new technologies including distributed namespaces, QDB, for scalability and flexibility.

Table 1. List of procured hardware type and specification for the CDS

Type	Specification
JBOD	DellEMC PowerVault ME484
Disk	Seagate/HGST 7.2k NL-SAS 12TB
Server	Intel Xeon E5-2637 v4 @ 3.50GHz, 4core × 2EA 192GB DDR4 RAM, 480GB SSD × 2EA (RAID-1), 12Gbps SAS HBA × 2EA
Interconnection	40Gbps Ethernet

CDS, Custodial Disk Storage; JBOD, Just-Bunch-Of-Disks.

The schematic view of the system architecture of the CDS with the EOS components deployed is shown in Fig. 4. In order to provide the high availability of archiving storage service, three of the MGMs are running concurrently behind a single domain name service (DNS) endpoint. Normally the MGM master node has the responsibility for all of the data I/O requests received through the other secondary nodes. When the master one is unresponsive, the fail-over to one of the other MGMs is almost instantaneous so that service availability can be unharmed. Essentially each MGM requires its own MQ service for the communication with FSTs and QDB should run in a raft mode that requires at least 3 pairs of nodes to achieve a quorum. Also, 18 FST components are deployed on top of 9 servers (EOS Front-End nodes) in order to accommodate an *RS* (16, 4) QRAIN layout with an additional two spares as discussed in Section 2.4.

All of the EOS components in the CDS run based on Linux container technology with Docker runtime. Automation using provisioning and configuration management tools such as Foreman, Puppet, and Ansible facilitates the deployment and operations. For example, a couple of releases of EOS, including fixes related to some critical issues identified during testing—such as redirected layout information truncation of QRAIN configuration with over a dozen stripes, and namespace corruption due to misconduct of update on quota information during

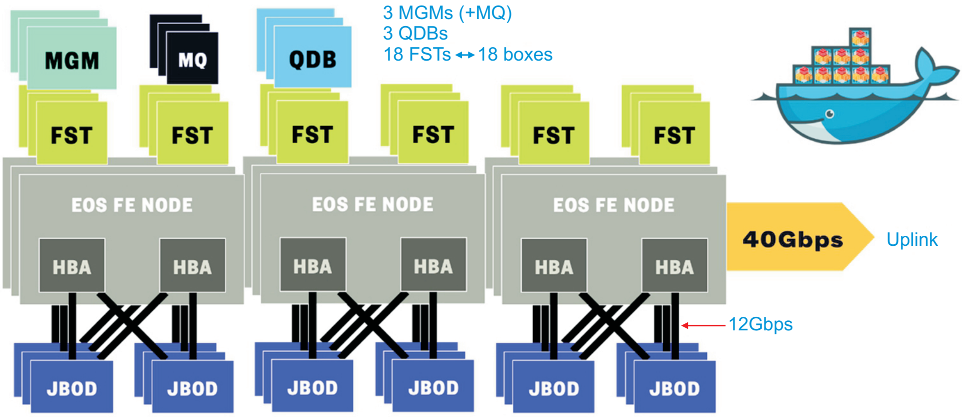


Fig. 4. System architecture of the CDS. CDS, Custodial Disk Storage; MGM, Manager; MQ, Message Queue; QDB, QuarkDB for namespace; FST, File Storage Server, JBOD, Just-Bunch-Of-Disks.

master role transition among MGMs— was easily and promptly deployed thanks to the automation.

3.2. EC Layout for the CDS

An EC layout in EOS works with a scheduling group which is a group of file systems provided by FST(s) so it is important to configure an appropriate number of file systems corresponding to the given layout. For the *RS* (16, 4) QRAIN layout as an example, a single scheduling group must be configured with 16 file systems because the layout requires 12 data and 4 parity blocks. By taking into account the two additional spare blocks, the number of file systems of a scheduling group in the CDS system should be 18. Also, the scheduling group must span file systems across individual FSTs of the CDS system so that the data protection can be effectively realized with the EC layout. There are 84 scheduling groups from *default.0* to *default.83* spanning each of 84 file systems (disks) mounted on 18 FSTs. The detailed view of EC layout for the CDS system is shown in Fig. 5.

EOS allows configuration of the EC layout and the number of stripes through directory attributes such as *layout* and *nstripes*, and those attributes are set to *qrain* and *16* respectively for the CDS system. When a file is written on the layout we configured, it will be fragmented and expanded to 12 (data)+4 (parity) blocks and then distributed to 16 FSTs. The remaining 2 FSTs as spares are involved in neither writing nor reading.

4. CUSTODIAL STORAGE FOR THE ALICE EXPERIMENT

CDS has been provided to the ALICE experiment at CERN as a custodial storage for raw data preservation since early 2021 after the ALICE integration. After successful commissioning for several months and participation in the WLCG Tape Challenges, the existing tape library at KISTI Tier-1 center was decommissioned in November 2021 and completely replaced by the CDS since then.

4.1. Integration as a Storage Element for the ALICE Experiment

The ALICE experiment's specific features should be enabled so that the CDS can work as a storage element for the experiment, and they are—

- Token-based authentication and authorization for data access
- Monitoring agent daemons on all FST components for ALICE monitoring
- Third-Party Copy (TPC)
- IPv6 networking

The first two functionalities can be enabled by installing and configuring *alicetokenacc* and *eosapmon* packages properly. Bypassing Simple Shared Secret (SSS) mechanism for the authentication between MGM and FST with *EOS_FST_NO_SSS_ENFORCEMENT=1* allows TPC to be enabled in EOS. Also, KISTI Tier-1 center is already capable of supporting IPv6 because IPv4/IPv6 dual stack configuration on networking has been recommended to all of the WLCG Tier-1/Tier-2 centers since 2017.

CDS is currently working for the ALICE experiment as a custodial storage in the name of *ALICE::KISTI_GSDC::CDS* and it has passed the periodic functional tests such as *add* (write), *get* (read), *rm* (delete), *3rd* (third-party copy) and *IPv6 add* (write over IPv6 network). The

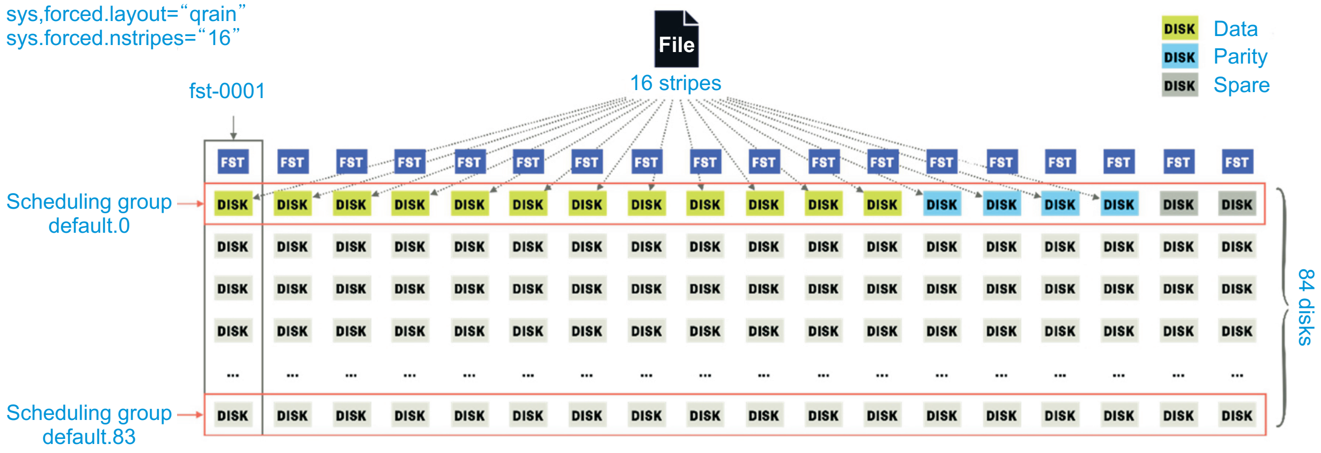


Fig. 5. *RS* (16, 4) QRAIN layout with additional 2 spares for CDS. CDS, Custodial Disk Storage; FST, File Storage Server.

list of all custodial storage elements for the experiment, their capacity, and usage as well as the status of periodic functional tests results are shown in Fig. 6. Note that the raw capacity of the storage reported by ALICE monitoring agents is 15.79 PB in 1024-base while what EOS reports is 17.77 PB, which is 1000-base.

4.2. WLCG Tape Tests Challenges

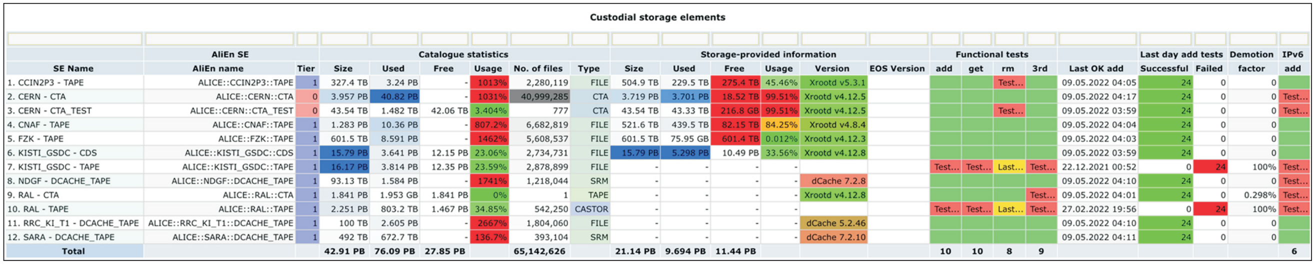
In mid October 2021, the WLCG scheduled a week of an intensive raw data transfer campaign in order for the LHC experiments and the WLCG Tier-1 centers to demonstrate the required performance for the LHC Run 3 goals. The Tier-1 tape network bandwidth needed in reads and writes for the ALICE experiment and the results of the challenges are listed in Table 2.

KISTI Tier-1 takes 5% of share (0.15 GB/s) among 8 Tier-1s contribution, which is 2.8 GB/s. Fig. 7 presents the data transfer rates over time of 8 Tier-1s for the ALICE experiment. The result shows that the overall performance of KISTI Tier-1 custodial storage (*KISTI_GSDC::CDS*) was extremely stable over time during the challenge week (October 11-15, 2021) and we achieved 169.9 MB/s (≈0.17 GB/s) on average.

4.3. Raw Data Replication and EC Impact on Network

In November 2021, the tape library at KISTI was completely replaced by the CDS for the ALICE experiment. Since early 2022, some share of the raw data taken in previous years from 2013 to 2018 has been replicated to the CDS and the amount of data transferred as of April 18 is about 2.9 PB, with more than 2 million files, which were once replicated in the tape library at KISTI. The average transfer rate for the last one month was about 0.5 GB/s and it is three times higher than the target performance set in the WLCG Tape Tests Challenges.

The network traffic captured during the replication (shown in Fig. 8) indicates that the available bandwidth of the CDS, which is 40 Gbps (=5 GB/s), has been fully utilized by the incoming data and outgoing traffic. The peak performances on both directions are 2.6 GB/s (incoming) and 1.9 GB/s (outgoing), and their sum, 4.5 GB/s, almost reaches the available bandwidth. The re-distribution of data fragments from an FST node to others invoked by the EC mechanism dominates both incoming and outgoing traffic (~2.0 GB/s). In order to maximize the utilization of network available between CERN and KISTI, which is a 20 Gbps (=2.5 GB/s) dedicated link, the bandwidth within



Custodial storage elements

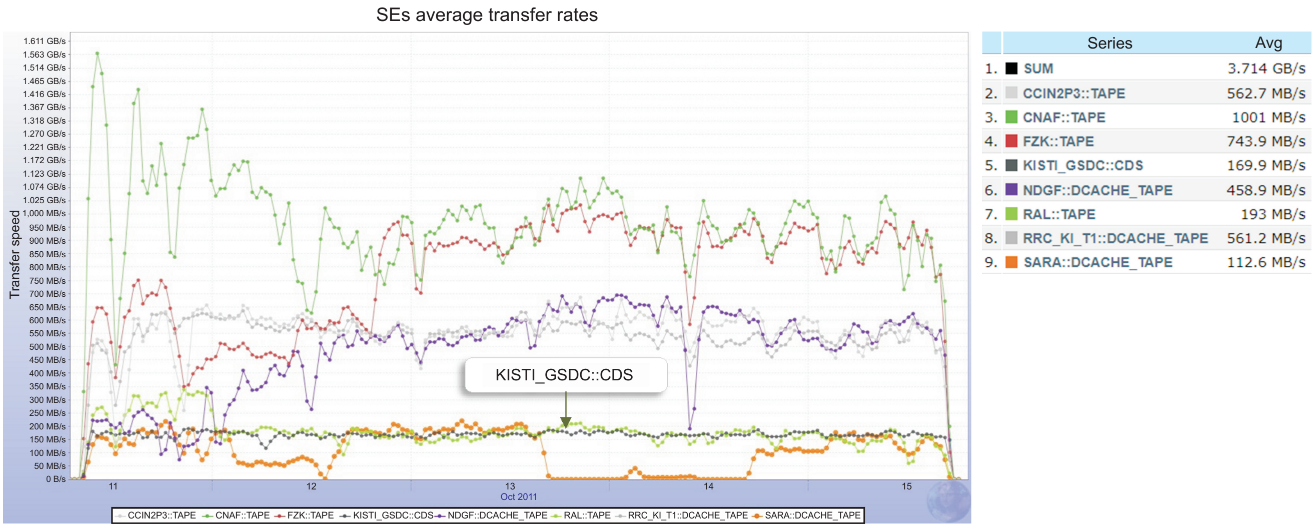
	AliEn SE		Catalogue statistics						Storage-provided information						Functional tests					Last day add tests		Demotion	IPv6
SE Name	AliEn name	Tier	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version	EOS Version	add	get	rm	3rd	Last OK add	Successful	Failed	factor	add
1. CCIN2P3 - TAPE	ALICE::CCIN2P3::TAPE	1	327.4 TB	3.24 PB	-	1013%	2,280,119	FILE	504.9 TB	229.5 TB	275.4 TB	45.46%	Xrootd v5.3.1		🟩	🟩	Test...	🟩	09.05.2022 04:05	24	0	0	🟩
2. CERN - CTA	ALICE::CERN::CTA	0	3.957 PB	40.82 PB	-	1031%	40,999,285	CTA	3.719 PB	3.701 PB	18.52 TB	99.51%	Xrootd v4.12.5		🟩	🟩	🟩	🟩	09.05.2022 04:17	24	0	0	Test...
3. CERN - CTA_TEST	ALICE::CERN::CTA_TEST	0	43.54 TB	1.482 TB	42.06 TB	3.404%	777	CTA	43.54 TB	43.33 TB	216.8 GB	99.51%	Xrootd v4.12.5		🟩	🟩	Test...	🟩	09.05.2022 03:59	24	0	0	Test...
4. CNAF - TAPE	ALICE::CNAF::TAPE	1	1.283 PB	10.36 PB	-	807.2%	6,682,819	FILE	521.6 TB	439.5 TB	82.15 TB	84.25%	Xrootd v4.8.4		🟩	🟩	🟩	🟩	09.05.2022 04:04	24	0	0	🟩
5. FZK - TAPE	ALICE::FZK::TAPE	1	601.5 TB	8.591 PB	-	1462%	5,608,537	FILE	601.5 TB	75.95 GB	601.4 TB	0.012%	Xrootd v4.12.3		🟩	🟩	🟩	🟩	09.05.2022 04:03	24	0	0	🟩
6. KISTI_GSDC - CDS	ALICE::KISTI_GSDC::CDS	1	15.79 PB	3.641 PB	12.15 PB	23.06%	2,734,731	FILE	15.79 PB	5.298 PB	10.49 PB	33.56%	Xrootd v4.12.8		🟩	🟩	🟩	🟩	09.05.2022 03:59	24	0	0	🟩
7. KISTI_GSDC - TAPE	ALICE::KISTI_GSDC::TAPE	1	16.17 PB	3.814 PB	12.35 PB	23.59%	2,878,899	FILE	-	-	-	-			Test...	Test...	Last...	Test...	22.12.2021 00:52	0	24	100%	Test...
8. NDGF - DCACHE_TAPE	ALICE::NDGF::DCACHE_TAPE	1	93.13 TB	1.584 PB	-	1741%	1,218,044	SRM	-	-	-	-	dCache 7.2.8		🟩	🟩	🟩	🟩	09.05.2022 04:10	24	0	0	Test...
9. RAL - CTA	ALICE::RAL::CTA	1	1.841 PB	1.953 GB	1.841 PB	0%	1	TAPE	-	-	-	-	Xrootd v4.12.8		🟩	🟩	🟩	Test...	09.05.2022 04:01	24	0	0.298%	Test...
10. RAL - TAPE	ALICE::RAL::TAPE	1	2.251 PB	803.2 TB	1.467 PB	34.85%	542,250	CASTOR	-	-	-	-			Test...	Test...	Last...	Test...	27.02.2022 19:56	0	24	100%	Test...
11. RRC_KI_T1 - DCACHE_TAPE	ALICE::RRC_KI_T1::DCACHE_TAPE	1	100 TB	2.605 PB	-	2667%	1,804,060	FILE	-	-	-	-	dCache 5.2.46		🟩	🟩	🟩	🟩	09.05.2022 04:10	24	0	0	🟩
12. SARA - DCACHE_TAPE	ALICE::SARA::DCACHE_TAPE	1	492 TB	672.7 TB	-	136.7%	393,104	SRM	-	-	-	-	dCache 7.2.10		🟩	🟩	🟩	🟩	09.05.2022 04:11	24	0	0	🟩
Total			42.91 PB	76.09 PB	27.85 PB		65,142,626		21.14 PB	9.694 PB	11.44 PB				10	10	8	9					6

Fig. 6. List of custodial storage elements provided to the ALICE experiment.

Table 2. ALICE objectives for tape tests challenges (Tier-1) – Writes (DT), Reads (A-DT), Writes (A-DT), and the results

Site	Writes (DT) (GB/s)	Reads (A-DT) (GB/s)	Writes (A-DT) (GB/s)	Transfer rates in tape challenges (GB/s)
CNAF (Italy)	0.8	0.3	0.8	1.0
CC-IN2P3 (France)	0.4	0.1	0.4	0.54
KISTI (Korea)	0.15	0.1	0.15	0.17
KIT (Germany)	0.6	0.3	0.6	0.53
NDGF (Nordic)	0.3	0.1	0.3	0.3
NL-T1 (Netherlands)	0.08	0.05	0.08	0.12
RRC-KI (Russia)	0.4	0.1	0.4	0.6
RAL (UK)	0.08	0.05	0.08	0.23

DT and A-DT stand for Data Taking and After Data Taking, respectively.



	Series	Avg
1.	■ SUM	3.714 GB/s
2.	■ CCIN2P3::TAPE	562.7 MB/s
3.	■ CNAF::TAPE	1001 MB/s
4.	■ FZK::TAPE	743.9 MB/s
5.	■ KISTI_GSDC::CDS	169.9 MB/s
6.	■ NDGF::DCACHE_TAPE	458.9 MB/s
7.	■ RAL::TAPE	193 MB/s
8.	■ RRC_KI_T1::DCACHE_TAPE	561.2 MB/s
9.	■ SARA::DCACHE_TAPE	112.6 MB/s

Fig. 7. Data transfer rates for WLCG Tape Tests Challenges in mid-October 2021. Dark green represents the performance of KISTI Tier-1 center. WLCG, Worldwide LHC Computing Grid.

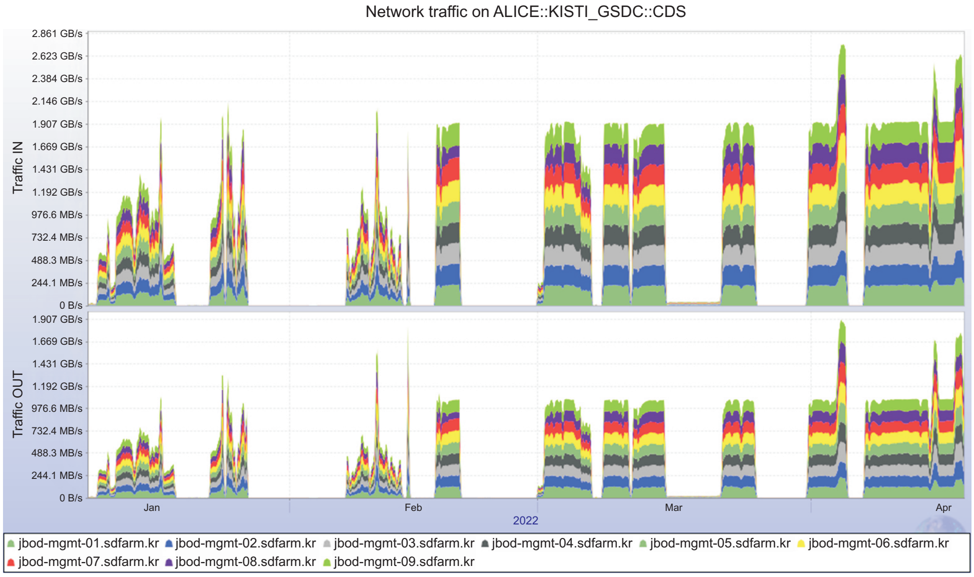


Fig. 8. Network traffic observed by the CDS nodes during the ALICE raw data transfer from January 10 to April 18, 2022. CDS, Custodial Disk Storage.

CDS should be capable of supporting above 6.5 GB/s (=52 Gbps) considering about 4.0 GB/s of the internal traffic. However, this is a rough estimation and the internal traffic also would be increased as the bandwidth grows. Therefore, we plan to enlarge the available bandwidth of the CDS system up to 80 Gbps by teaming the two installed 40 Gbps network interfaces on the servers.

5. LONG-TERM ELECTRIC POWER CONSUMPTION

Electric power rates usually take a significant share of operational costs in the data center infrastructure and it is a factor where tape-based archival storage is advantageous. A rough estimation based on short-term power consumption measurement of disk storage with high density JBOD enclosures showed that the CDS can be significantly power efficient compared to the other mid-range or enterprise class disk storages and is not uncomfortably higher than tape (Ahn et al., 2020).

In order to get a long-term picture, we have collected electric power consumption in every minute for over a year since January 2021 through 12 power distribution units (PDU) installed in 3 racks for the CDS. Fig. 9 shows the physical setup of the installation of PDUs and their connection to power supply units of servers and enclosures in the racks. The instantaneous power consumption measured in kilowatts from 12 PDUs were aggregated over time with some corrections, such as for missing data, duplicates, and none values. The full-scale view on the measurement over time from January 2021 to early March 2022 is shown in Fig. 10. The maximum, minimum, and average values of power consumption are 20.426 kW, 11.015 kW, and 16.85 kW, respectively. In particular, the mean power consumption has gradually increased over time by 3.2% from 16.49 kW (January 8, 2021) to 17.02 kW (March 6, 2022).

To compare with the other storages including tape, the results were normalized such that the power consumption (watt) is represented by unit storage capacity (terabyte). Thus, the normalized results for the CDS are 1.125 W/TB (maximum), 0.607 W/TB (minimum), and 0.923 W/TB (mean). The tape library used at KISTI was a TS3500 from IBM and it was equipped with 8 TS1140 tape drives including tape drive mounting kits, expansion frames, additional accessor, backend switches, and so on. The calculation of total library power consumption at maximum loads is 1.6 kW (IBM Corporation, 2012) and the normalized rate is 0.5 W/TB considering 3,200 TB of total capacity. It turned out that the normalized power consumption of the CDS at maximum load is two times higher than that of the tape; however, it is remarkable that the CDS is as power efficient as tape is at idle states. The comparison between the CDS and tape library at KISTI in power consumption is listed in Table 3.

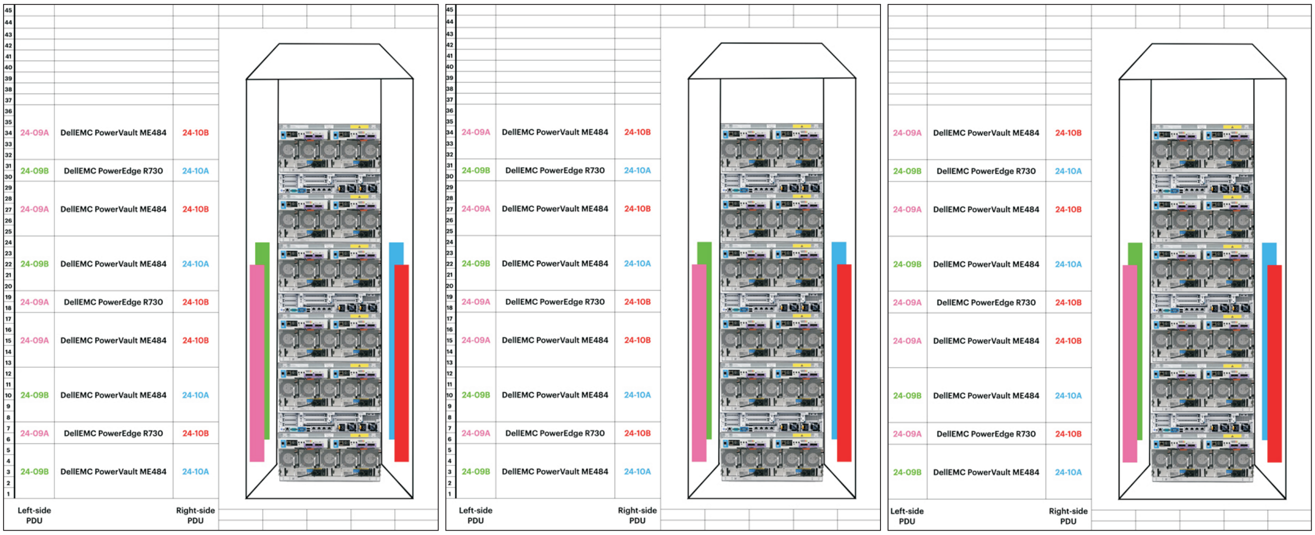

24-09A DellEMC PowerVault ME484 24-10B
24-09B DellEMC PowerEdge R730 24-10A
24-09A DellEMC PowerVault ME484 24-10B
24-09B DellEMC PowerVault ME484 24-10A
24-09A DellEMC PowerEdge R730 24-10B
24-09A DellEMC PowerVault ME484 24-10B
24-09B DellEMC PowerVault ME484 24-10A
24-09A DellEMC PowerEdge R730 24-10B
24-09B DellEMC PowerVault ME484 24-10A
Left-side PDU
Right-side PDU

Fig. 9. Physical setup of 12 power distribution units installed in 3 racks for the CDS. CDS, Custodial Disk Storage.

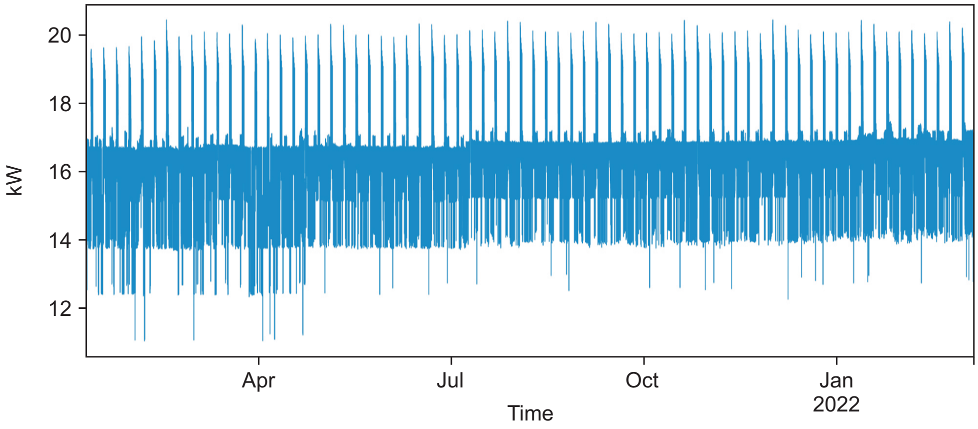


Fig. 10. Instantaneous electric power consumption (in kilowatt) per minute from January 2021 to March 2022.

Table 3. Power consumption comparison between CDS and tape library at KISTI Tier-1 center

Storage model	Capacity (TB)	kW		W/TB	
		Max	Min	Max	Min
CDS	18,144	20.426	11.015	1.125	0.607
TS3500 (Tape)	3,200	1.6	1.3	0.5	0.406

Maximum loads are normalized such that how much power (watt) is consumed by unit storage capacity (terabyte).
CDS, Custodial Disk Storage.

6. CONCLUSIONS

The CDS is designed to replace the existing tape library at KISTI as a custodial storage for data preservation, and simplicity, reliability and cost-effectiveness are the goal of the disk-based custodial storage system. The system consists of 18 high density JBOD enclosures attached to 9 servers through multiple 12 Gbps SAS HBA interfaces for redundancy and multiplexing. With the EOS EC implementation of *RS* (16, 4) plus 2 spares, the data loss probability of the CDS is expected to be extremely low. Integration as a custodial storage element for the ALICE experiment and commissioning with periodic functional tests were successful so that the CDS could participate in the WLCG Tape Tests Challenges in timely manner. The raw replication started in early of 2022 has been on-going until now (as of writing) and has helped understand the impact of the EC mechanism on networking. A long-term view on consumed electric power suggests that the CDS is not only simple and reliable but also cost-effective at a certain level compared to the tape-based custodial storage. Further understanding and analysis on the correlation of long-term measurement of electric power consumption with the EOS EC operations, and upgrading the network bandwidth of the CDS remain as future works. We hope that the work presented in this paper will give a promising option to whomever wants to look for alternatives to tape-based storages or services, particularly to the WLCG Tier-1 centers.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) through contract N-22-NM-CR02 and the Program of Data Computing Service for Large-scale Experimental Data (K-22-L02-C02).

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Ahn, S. U., Betev, L., Bonfillou, E., Han, H., Kim, J., Lee, S. H., Panzer-Steindel, B., Peters, A. J., & Yoon, H. (2020). Seeking an alternative to tape-based custodial storage. *EPJ Web of Conferences*, 245, 04001. https://doi.org/10.1051/epjconf/202024504001.

Arslan, S. S. (2014). *Durability and availability of erasure-coded systems with concurrent maintenance*. http://www.suaybarslan.com/Reliability_Systems_14.pdf.

Broadcom. (2013). *12Gb/s SAS: Busting through the storage performance bottlenecks*. https://docs.broadcom.com/docs/12353459.

Colarelli, D., & Grunwald, D. (2002). *Massive arrays of idle disks for storage archives*. Paper presented at SC '02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing, Baltimore, MD, USA.

IBM Corporation. (2012). *TS3500 tape library power requirements for frames*. https://www.ibm.com/docs/en/ts3500-tape-library?topic=requirements-power-frames.

Peters, A. J., & Janyst, L. (2011). Exabyte scale storage at CERN. *Journal of Physics: Conference Series*, 331(5), 052015. https://doi.org/10.1088/1742-6596/331/5/052015.

Peters, A. J., Simon, M. K., & Sindrilaru, E. A. (2020). Erasure Coding for production in the EOS Open Storage system. *EPJ Web of Conferences*, 245, 04008. https://doi.org/10.1051/epjconf/202024504008.

Peters, A. J., Sindrilaru, E. A., & Adde, G. (2015). EOS as the present and future solution for data storage at CERN. *Journal of Physics: Conference Series*, 664, 042042. https://doi.org/10.1088/1742-6596/664/4/042042.

Sindrilaru, E. A., Peters, A. J., Adde, G. M., & Duellmann, D. (2017). EOS developments. *J Phys: Conf Ser*, 898, 062032. https://doi.org/10.1088/1742-6596/898/6/062032.

Spectra Logic. (2021). *Data storage Outlook 2021*. https://spectralogic.com/wp-content/uploads/DSO_2021.pdf.