Seeking an alternative to tape-based custodial storage

Sang Un Ahn[1,*], *Latchezar* Betev[2], *Eric* Bonfillou[2], *Heejune* Han[1], *Jeongheon* Kim[1], *Seung Hee* Lee[1], *Bernd* Panzer-Steindel[2], *Andreas Joachim* Peters[2], and *Heejun* Yoon[1,]

[1]KISTI, GSDC, 245 Daehak-ro Yuseong-gu 34141, Daejeon, South Korea
[2]CERN, Esplanade des Particules 1, 1211 Geneva 23, Switzerland

Abstract. In November 2018, the KISTI Tier-1 centre started a project to design, develop and deploy a disk-based custodial storage with error rate and reliability compatible with a tape-based storage. This project has been conducted in collaboration with KISTI and CERN; especially the initial design was laid out from the intensive discussion with CERN IT and ALICE. The system design of the disk-based custodial storage consisted of high density JBOD enclosures and erasure coding data protection, implemented in EOS, the open-source storage management developed at CERN. In order to balance the system reliability, data security and I/O performance, we investigated the possible SAS connections of JBOD enclosures to the front-end node managed by EOS and the technology constraints of interconnections in terms of throughput to accommodate large number of disks foreseen in the storage. This project will be completed and enter production before the start of LHC Run3 in 2021. In this paper we present the detailed description on the initial system design, the brief results of test equipment for the procurement, the deployment of the system, and the further plans for the project.

1 Introduction

The Worldwide LHC Computing Grid (WLCG) is a global cyber-infrastructure formed to store and process large-scale data produced from the experiments at the Large Hadron Collider (LHC), the world's largest accelerator at the European Organisation for Nuclear Research (CERN). The WLCG consists of a number of research institutes, universities and computing centres around the world. They are classified into three tiers – Tier-0, Tier-1 and Tier-2 – depending on their roles such as raw data preservation and processing, simulation, organised/unorganised analyses, etc. There is only one Tier-0 at CERN, which receives and processes data from the experiments, while copies the data to custodial storage and a second copy is sent to several Tier-1s. Like Tier-0 at CERN, providing custodial storage is the one of the requirements for all WLCG Tier-1 sites in order to securely preserve a portion of raw data generated and distributed from CERN.

The custodial storage must be cheap, cost-effective and reliable for long-term preservation of raw data. In general, tape is considered as a good candidate for archive storage [1] thanks to its reliable nature and cheap media cost among the different substrates such as magnetic disk, solid-state, and optical. The prices of the media that could be considered for archive storage

[*] e-mail: sahn@kisti.re.kr

are tape at less than 10 USD per terabyte, magnetic disk at less than 25 USD per terabyte, and optical technology at around 60 USD per terabyte [2]. However, when reading and writing data using tape it requires a Hierarchical Storage Management (HSM) including buffer or cache, typically disk storage. An organised access is mandatory to efficiently complement slow tape staging time and the serial access to the tape media. This implies dedicate human efforts with a certain level of expertise are available to operate the HSM and tape library in production.

Lately, the tape market has been facing potential risks from single vendor lock-in issues since 2017, when Oracle dropped the production of enterprise-class tape drives (i.e. T10KD and its successor) from its hardware product development roadmap [3]. Today IBM is the only manufacturer of enterprise-class tape drives and Linear Tape-Open (LTO). Also the number of tape cartridge manufacturers has been shrunk from six to two since 2015 [4]. The two remaining companies have had patent dispute in the first half of 2019 which led to the suspension of entire LTO-8 tape cartridge supply throughout the year [5, 6]. These events have raised the community's concerns about the exclusive control of tape supply since 2017.

2 Alternative to Tape Archive Storage Project

On 30th November 2018, we, a WLCG Tier-1 centre for the ALICE experiment at, Korea Institute of Science and Technology Information (KISTI) in South Korea, proposed to the 32nd Meeting of WLCG Overview Board that, with the endorsement of ALICE, we plan to pursue an alternative custodial storage to tape - a disk-only system. The goal aims to replace the existing 3+ PB tape library, IBM TS3500, and commercial solutions such as Spectrum Protect (the former Tivoli Storage Manager, TSM) and Spectrum Scale (General Parallel File System, GPFS) with cheap off-the-shelf equipment and open-source storage solutions. This will simplify substantially the setup and reduce operational costs as well as provide an acceptable quality of service in terms of raw data archive required by the ALICE experiment. Thereafter we launched Alternative to Tape Archive Storage (ATAS) project to design, procure, implement and validate a custodial storage system based on high density Just-Bunch-Of-Disks (JBOD) enclosures and EOS [7], the CERN developed highly scalable storage management software, to be deployed on top of front-end nodes hosting JBODs. In particular, the recent releases of EOS implemented erasure coding that enables to configure Redundant Array of Independent Nodes (RAIN) [8]. This feature provides a high level of reliability in terms of data protection with the combination of many data nodes and a few parity nodes analogous to the Redundant Array of Independent Disks (RAID).

3 High Density JBOD Products

The term, 'High Density JBOD', is defined here as a JBOD enclosure that can host more than 60 disks in a single chassis. There are several high density JBOD products available in the market from Dell/EMC, HPE, QCT, WD, etc. They have different dimensions, 4U or 5U, and the number of disk drives, ranging from 60 up to 102. The list of selected products from those manufacturers is shown in Table 1. In the table, the capacity per enclosure was calculated by multiplying the maximum number of disks in a JBOD enclosure by 12TB (the largest HDD space available in the market at the time of writing). The JBOD enclosures listed in Table 1 are models without hardware-level RAID controllers. Products shipped with entry-level RAID cards offering low-level data protection are also available in the market.

In general, storage vendors provide hardware management software suites bundled with their products to be installed on host servers to which the storage enclosures are attached.

Table 1: List of selected high density JBOD products. These models are not equipped with hardware-level RAID controllers.

Model	Unit	# Disks	Capacity per box
Dell/EMC PowerVault ME484	5U	84	1,008 TB
HPE D6020	5U	70	840 TB
QCT JB4602 JB9T	4U	60	720 TB
WD Ultrastar Data60 H4060-J SE4U60-60	4U	60	720 TB
WD Ultrastar Data102 H4102-J SE4U102-102	4U	102	1,224 TB

Table 2: Sample storage configurations showing the I/O rate limiting element and the number of drives supported at their peak throughput [9]. Note that SAS x4(x8) denotes the number of lanes embedded in one port of SAS HBA controller.

Configuration	Limiting element (MB/s)	# of HDDs (MB/s)	# of SSDs (MB/s)
6Gb/s SAS x8 / PCIe 2.X	PCIe (3200)	14 (230)	6 (550)
12Gb/s SAS x4 / PCIe 2.X	PCIe (3200)	14 (230)	6 (550)
12Gb/s SAS x4 / PCIe 3.0	SAS (4400)	19 (230)	8 (550)
12Gb/s SAS x8 / PCIe 3.0	PCIe (6400)	28 (230)	12 (550)
12Gb/s SAS x8 / PCIe 3.0	PCIe (6400)	56 (170)	12 (550)

These are essentially a useful CLI tool for administrators to check the hardware-level information such as installed HBA cards, disks, enclosures and to monitor the status of disks, enclosures, power supplies, fans, temperature. Also, some manufacturers provide REST APIs that may help the integration with the existing monitoring framework at sites. Note that to fully manipulate these features the JBODs must be compatibility matched with the SAS HBAs.

4 State-of-the-art SAS HBA and PCIe 3.0 Limitation

The state-of-the-art SAS HBA is third-generation SAS and has 12Gbps bandwidth in a single lane. One SAS HBA card can support up to 1,024 non-RAID disks. Each port of a SAS HBA card is composed of four 12Gbps lanes therefore a typical 2-port card is capable to transfer data at a nominal 9,600MB/s. There is a 4-port model on the market with 16 lanes of 12Gbps allowing for at most 19,200MB/s in half duplex mode.

Normally the SAS HBA card is interfaced with PCI Express 3.0 slot. A white paper [9] published by LSI (now Broadcom) in 2013 reported that PCIe 3.0 can be a bottleneck as its upper limit of transfer speed is under 8,000MB/s, precisely 7,877MB/s. Considering 80% efficiency, the typical transfer speed is reduced to 6,400MB/s. Considering I/O speed of disk drives, fast-spinning SAS disks with a 6Gbps interface can deliver a data rate of up to 230MB/s, whereas SATA drives with a 6Gbps interface normally deliver from 100 to 170MB/s. SSDs are capable of delivering data rates closer to the peak throughput of the 6Gbps SAS interface, or around 550MB/s. Due to the limitation of transfer performance inherit to PCIe 3.0, 28 fast-spinning SAS disks (or 12 SSDs) or about 60 slower SATA disks are enough to saturate the bandwidth provided by a 2-port 12Gbps SAS HBA card. This is shown in Table 2.

5 Initial System Design

The main consideration of the design is the balance between data protection and usable capacity within the project budget, which is approximately one million US dollars. Through intensive market searches in South Korea and worldwide and in meetings with domestic suppliers, we concluded that we can attain approximately 20PB of disk storage in few tens of JBOD enclosures. The number of JBOD enclosures and thus the total raw capacity are dependent on the manufacturers and resellers. The initial system design comprises twenty high density JBOD enclosures and ten front-end nodes interconnecting with two 12Gbps SAS HBA cards as shown in Figure 1.

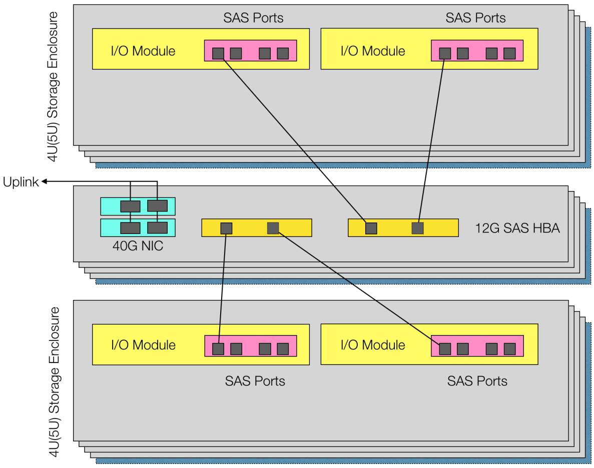


Figure 1. The schematic view of physical connection of JBOD enclosures and front-end nodes. Each of two SAS HBA controllers are connecting to JBOD storage enclosures in dual-domain connection.

As briefly mentioned in Section 2, the data protection of the system relies on RAIN layout of EOS across the front-end nodes. The available RAIN layouts in EOS are `raid5`, `raid6`, `raiddp`, `archive` and `qrain`. EOS also provides simple layouts such as `plain` (no erasure coding or replication) and `replica`. The difference of the RAIN layouts is the number of parity nodes enforced in the configuration. For example, `raid6` enforces two parity nodes out of at least 6 nodes in one pool while `archive` and `qrain` can have three and four parity nodes, respectively. The higher the number of parity, the less likely the data loss and the usable space. With high parity, 30-40% of the raw capacity has to be sacrificed. If we have sufficient data nodes, we can balance the data protection and usable capacity.

For example, in our design, we have two FSTs – a FST is referred to as a storage server component in EOS – on each front-end node, resulting in 20 FSTs across ten nodes as shown in Figure 2. One can enable two or more FSTs on a single host by starting relevant daemons with different port numbers. We decided to use container technologies, which allows us to have substantial flexibility in deployment, operations and maintenance. We developed and published [10] an automated deployment procedure of of EOS components using Red Hat Ansible platform referring to EOS Docker project [11] and AARNet's deployment practices [12].

The RAIN layout applied in this design consists of 14 data nodes and 4 parity nodes out of 20 FST nodes, i.e. $EC(M, K) = (14, 4)$, where M is the number of data nodes and K is the number parity nodes. The RAIN layout of this configuration is shown in Figure 2. Data, parity and spare blocks are determined randomly by EOS. The usable space out of this configuration is 77.7% of physical capacity and the data loss probability is about 5.02×10^{-8} which is acceptable for ALICE. A method [13] to estimate the data loss probability is shown

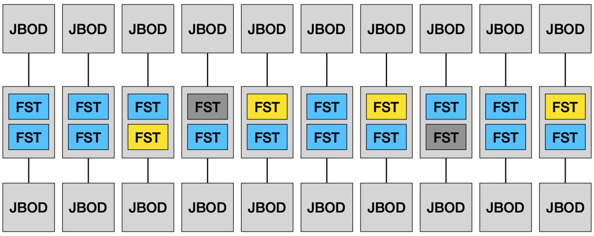


Figure 2. The layout of the initial system design for disk-based archive storage and EOS FST deployment with quad-parity RAIN layout. Blue, yellow and dark squares represent data, parity and spare nodes respectively.

below:

$$p = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1}$$

where k = *number of parity disks + 1* and

$$\lambda = \frac{AFR \times (Number\ of\ disks)}{365 \times 24 \div MTTR}. \tag{2}$$

For example, assuming 1,680 disks out of total 20 JBODs, 2% of Annualized Failure Rate (AFR) and 24 hours of Mean Time To Repair (MTTR), the value of $\lambda = 0.092$ and with 4 parity disks, the data loss probability p is,

$$p = e^{-0.092}\frac{0.092^5}{5!} = 5.01 \times 10^{-8}. \tag{3}$$

Note that this is a theoretical estimate based on our best practice in terms of maintenance cycle for disk repair (24 hours) and the average value of AFR, which is derived from the ones periodically published by hard disk manufacturers. In reality, the life time of any device depends on the environmental conditions such as temperature, humidity, electric power stability. With this RAIN layout – $EC(M, K) = (14, 4)$ – across 20 FSTs, there are always 2 spare (extra) nodes that are used neither as data node nor as parity node. This allows operational flexibility allowing one front-end node (including 2 FSTs) to be shut down for maintenance at any time.

6 Demo Equipment Test

Based on the design discussed in the previous section, we started a procurement in May 2019. The call for tender was announced at the end of September. In the meantime, we tested one of the high density JBOD models listed in Table 1, a Dell/EMC ME484 with 70 disks (12TB per drive) installed. The specification of the test equipment is shown in Table 3. The read/write performance test and electrical power consumption measurement were done at the same time. As a result, we confirmed the I/O rate limiting element discussed in Section 4, which is unavoidable for any JBOD as long as PCIe 3.0 interface is used. The tests were conducted using two popular I/O test suites, VDBench and IOZone, with increasing the size of transfer blocks from 4 to 2,048 or 4,096KB. The target 70 filesystems (equivalent to 70 physical disks) in the tests were formatted with XFS, the default filesystem in popular Linux distributions such as Red Hat Enterprise Linux, CentOS, Ubuntu.

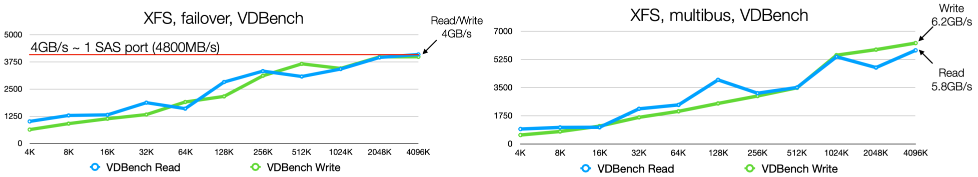
6.1 Multipath Mode

The SAS interface allows multipath configuration so that directly attached storage (i.e., JBOD enclosures) can be accessed via multiple physical links allowing for redundancy and multiplexing through those links for larger bandwidth. Typically two multipath modes are available: *failover* (active-standby) and *multibus* (active-active). The *multibus* mode allows one to use multiple paths simultaneously whereas the *failover* does not.

Table 3: The specification of test equipment including high density JBOD enclosure and host server used as the front-end node.

Model	Specification
Dell/EMC PowerVault ME484	HGST 12TB 7.2k NL-SAS (840TB) x70
Dell/EMC PowerEdge R640	Intel Xeon Scalable 6150 2.7GHz 18core x2 DDR4 16GB 2666MHz x24 Dell/EMC PowerEdge 12Gb/s SAS HBA x2 QLogic 4x10GE QL41164HMCU CNA x2

The read/write test results in two different multipath modes shown in Figure 3 showing the anticipated behavior. As the *failover* allows to use only single path while the other path remains in standby, the peak transfer speed is capped by the peak performance of 12Gbps SAS HBA at 4,800MB/s. In Figure 3a, both read and write speeds reach 4GB/s at 2,048KB transfer block size. The test with *multibus* mode confirmed that the mode allows the use of both links simultaneously and the peak performance of read and write reaches around 6GB/s at 4,096KB transfer block size as shown in Figure 3b. The I/O limit of PCIe 3.0 at around 6GB/s is seen and it persists throughout the test results.



(a) failover mode.

(b) multibus mode.

Figure 3: The read/write test results in two different multipath modes: *failover* and *multibus*.

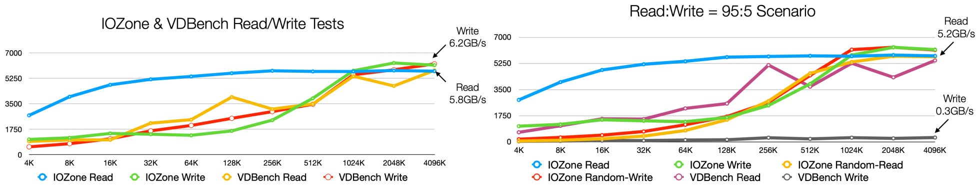
6.2 Read/Write Testing with Scenarios

We conducted read/write performance tests in *multibus* mode based on different scenarios: a) read-only (100%), b) write-only (100%) and c) write-once-read-many (read:write=95:5). Specifically, scenario c) write-once-read-many is the behavior of archive storage operations in which the data is repeatedly being read for reprocessing.

In Figure 4, the read/write performance test results using two different I/O benchmark suites are presented. For read-only and write-only, the results from VDBench and IOZone are approaching 6GB/s at 2,048KB transfer block size (Figure 4a) and the results of write-once-read-many are similar (Figure 4b). Although the PCIe 3.0 I/O limit (~ 6GB/s) persists, this is a higher result by a factor of 3 compared to the throughput of current tape system at KISTI, at 2GB/s delivered by 8 tape drives (~ 250MB/s per drive).

6.3 Power Consumption

The advantage of having disk-only system as archive storage in terms of performance is obvious compared to the existing tape system as shown in the previous section. In addition,



(a) read/write only (b) write-once-read-many

Figure 4: The read/write test results on different scenarios

disk storage provides direct random access to data while tape system only allows serial access. Nevertheless, the strength of tape for archiving purpose is that tape is reputed to be the cheapest media, in terms of electrical power consumption. Thus it is crucial to measure the power consumption of the disk-only archive storage and to compare with tape as well as the other types of disk storage.

To simplify the measurement, we spared an entire rack to host only one network switch, one front-end node and one JBOD enclosure. A power distribution unit (PDU) installed in the rack can measure power consumption in three different vertical zones. We placed the rack switch at the top, the host server in the middle and the JBOD enclosure at the bottom. Figure 5 shows the configuration of the test setup and the power consumption curves as a function of time representing both idle and full load during read/write performance tests.

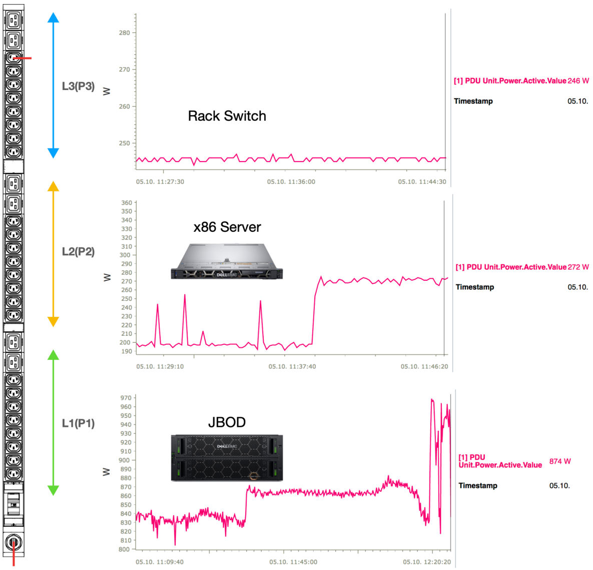


Figure 5. The electrical power consumption in idle status and full load measured during read/write tests. In this configuration, only one network switch, one x86_64 front-end server and one JBOD enclosure are placed in a rack.

A single JBOD enclosure with 70 disks consumes 960W at full load. Given that the installed capacity of the box is 840TB (= 70 disks × 12TB), it can be normalised to watts per terabyte resulting in 1.12W/TB. Adding the server and switch power, the test setup of high density JBOD archive storage consumes 1,476W at full load, which can be normalised to 1.75W/TB. This is not uncomfortably higher value than tape when comparing with the other type of storage at KISTI computing centre as shown in Table 4. The normalised results of high-end enterprise class disk storage such as EMC Isilon, Hitachi VSP are from about

Table 4: Power consumption of disk storage and tape in production at KISTI and comparison to high density JBOD archive storage.

Model	Capacity (TB)	Full Load (W)	W/TB
Disk Storage			
Dell/EMC SC7020	2500	12120	4.8
EMC Isilon (16 Nodes)	2950	13730	4.6
EMC VNX (12 Nodes)	2360	5100	2.2
Hitachi VSP	2000	18300	9.1
EMC Isilon (15 Nodes)	1430	12880	9.0
EMC CX4-960	1500	14900	9.9
Tape			
IBM TS3500 (5 Frames)	3200	1600	0.5
Disk-only Archive Storage			
JBOD (Dell/EMC ME484)	840	960	1.12
Server	-	270	-
Switch	-	246	-
JBOD+Server+Switch	840	1476	1.75

5W/TB up to 10W/TB. As anticipated, tape uses relatively little power (0.5W/TB) since most of power is consumed by 8 tape drives and robotics during read/write requests while the rest of the tape cartridges are stationed in slots without power.

7 Conclusions

We have investigated a disk-based system as an alternative to tape-based custodial storage. The benefits of having disk-only archive system are obvious: avoiding single-vendor dependency and taking advantage of common expertise for all storage systems in the computing centre. We focused on designing a reliable and economic storage for archive that meets the ALICE requirement. The initial system design is based on high density JBOD enclosures and EOS with 4-parity erasure coding RAIN layout. A demo equipment test conducted while the procurement process was on-going confirmed the I/O limits of PCIe 3.0 interface at ∼ 6GB/s on transfer speeds that a two-(or four-)port 12Gbps SAS HBA card can sustain. This is higher by a factor of 3 than the throughput of the tape system currently in use at KISTI. The normalised result of power consumption measurement on the demo setup shows 1.75W/TB, which is not uncomfortably higher than tape. There needs a significant improvement of electrical power consumption in order for disk-only archive storage to be competitive with tape.

This project has been conducted in collaboration between KISTI and CERN with a goal to have a production system before the start of LHC Run3, scheduled to start in June 2021. The procurement of high density JBOD enclosures including front-end nodes and necessary interconnection devices within the project budget were successfully finished by January 2020. The total physical capacity is 18PB and the usable space after the setup of quad-parity RAIN layout should be around 14PB. In the coming years, we plan to evaluate this disk-only archive storage in terms of reliability through metrics such as disk failure rate, data loss or corruption rate, recovery capability. We aim to provide a good input to the community as an alternative option for archiving solutions at the conclusion of the ATAS project.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) through contract N-19-NM-CR01-S01 and the Program of Construction and Operation for Large-scale Science Data Center (K-19-L02-C03-S01).

References

[1] D. Colarelli, D. Grunwald, In: SC'02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing (47-47) 2002.

[2] Spectra Logic Corporation, Digital Data Storage Outlook 2019, White Paper, https://spectralogic.com/data-storage-outlook-report-2019

[3] https://www.theregister.co.uk/2017/02/17/oracle_streamline_tape_library_future

[4] https://www.bloomberg.com/news/articles/2018-10-17/the-future-of-the-cloud-depends-on-magnetic-tape

[5] https://www.theregister.co.uk/2019/05/31/lto_patent_case_hits_lto8_supply

[6] https://www.theregister.co.uk/2019/08/06/sony_fujifilm_storage_patent_lawsuit_settled

[7] EOS Open Storage, https://eos.web.cern.ch

[8] V. Bohossian, C. Fan, P. LeMahieu, M. D. Riedel, L. Xu, J. Bruck, IEEE Transactions on Parallel and Distributed Systems, 12(2), (99-114) 2001.

[9] Broadcom, 12Gb/s SAS: Busting Through the Storage Performance Bottlenecks, White Paper, https://docs.broadcom.com/docs/12353459

[10] https://github.com/jeongheon81/gsdc-eos-docker

[11] https://gitlab.cern.ch/eos/eos-docker

[12] https://github.com/AARNet/eos-docker

[13] https://blog.synology.com/data-durability