EPJ Web of Conferences **251**, 02051 (2021) https://doi.org/10.1051/epjconf/202125102051
*CHEP 2021*

# The first disk-based custodial storage for the ALICE experiment

*Sang Un* Ahn[1,*], *Jeongheon* Kim[1], *Heejune* Han[1], *Seung Hee* Lee[1], and *Heejun* Yoon[1]

[1]KISTI, GSDC, 245 Daehak-ro Yuseong-gu 34141, Daejeon, South Korea

**Abstract.** We proposed a disk-based custodial storage as an alternative to tape for the ALICE experiment at CERN to preserve its raw data. The proposed storage system relies on Redundant Array of Independent Nodes (RAIN) layout – the implementation of erasure coding in the EOS storage suite, which is developed by CERN – for data protection and takes full advantage of high-density Just-Bunch-Of-Disks (JBOD) enclosures to maximize storage capacity as well as to achieve cost-effectiveness comparable to tape. The system we present provides 18 PB of total raw capacity from the 18 set of high-density JBOD enclosures attached to 9 EOS front-end servers. In order to balance between usable space and data protection, the system will stripe a file into 16 chunks on the 4-parity enabled RAIN layout configured on top of 18 containerized EOS FSTs. Although the reduction rate of available space increases up to 33.3% with this layout, the estimated annual data loss rate drops down to $8.6 \times 10^{-5}$%. In this paper, we discuss the system architecture of the disk-based custodial storage, 4-parity RAIN layout, deployment automation, and the integration to the ALICE experiment in detail.

## 1 Introduction

The ALICE experiment [1] is one of the four gigantic experiments being performed using the Large Hadron Collider (LHC) at European Organization for Nuclear Research (CERN), to study the nature of primordial matter, the Quark-Gluon Plasma, believed to be formed just after the Big Bang. Like the other giant experiments at CERN such as ATLAS, CMS, and LHCb, the ALICE experiment mainly relies on the Worldwide LHC Computing Grid (WLCG) [2] for its asynchronous processing of data produced from proton-proton or proton-nucleus or nucleus-nucleus collisions at the LHC. In particular, 14 centres in the WLCG categorized as Tier-1 share their role, together with CERN Data Centre – the Tier-0, of the experimental raw data preservation in their custodial storage systems. Tape library has been widely used as the custodial storage system for several decades thanks to its cheap cost and reliable nature of the medium. However additional effort and cost are required to operate the tape based system efficiently. A Hierarchical Storage Management (HSM) system including disk buffer should be placed in front of tape to read and write data, and an organized way of access requests is mandatory to complement the serial access of tape through drives. More recently the decreasing number of enterprise-class tape drive and cartridge manufacturers [3,

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

4] followed by the patent disputes between tape media suppliers [5, 6] have increased the risk of consistent tape supply.

Therefore in 2018 at the Korea Institute of Science and Technology Information (KISTI) in South Korea - a WLCG Tier-1 centre for the ALICE experiment - we proposed a disk-based custodial storage system for the experiment as an alternative to tape. A preliminary design has been presented [7], based on high-density Just-Bunch-Of-Disks (JBOD) enclosures accommodating erasure coding implementation in EOS, together with the study on performance limitation of 12 Gbps SAS HBA, and demo equipment test results including I/O performance and power consumption compared to the tape library as well as other enterprise-class disk storage running at KISTI Tier-1 centre.

In this paper, we discuss in detail the system architecture, targeting a production service for ALICE experiment in mid-2021, based on EOS implementation of erasure coding for data protection.

## 2 System Architecture

The first principle of the system architecture design for the disk-based custodial storage was to maximize usable space out of cheap high-density JBOD enclosures within the available budget while keeping at the same time the level of data protection high enough respect to tape. Based on this principle, we deployed EOS on top of cheap high-density JBOD enclosures for the custodial storage system. EOS is an open-source project developed by CERN for large-scale data management and it has been used in production for the LHC experiments [8].

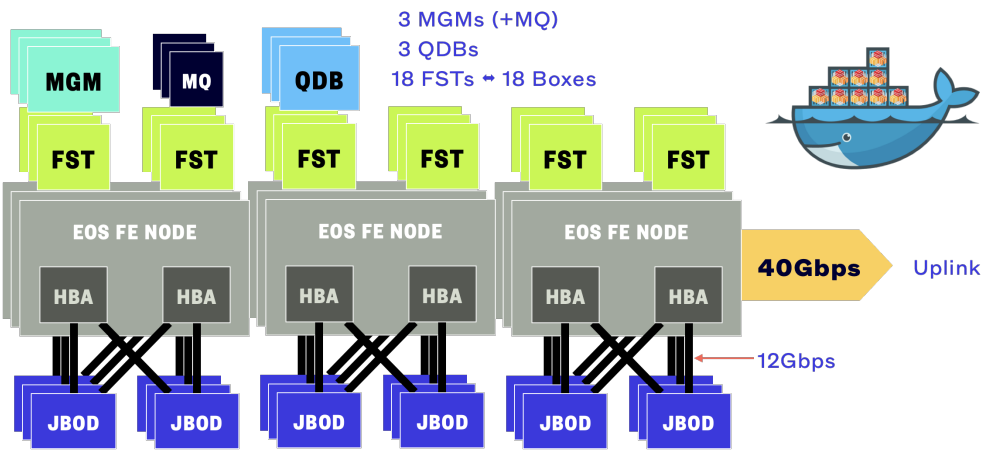



**Figure 1.** The system architecture of the disk-based custodial storage for the ALICE experiment.

As shown in Figure 1, the installed system for the disk-based custodial storage at KISTI includes 9 x86_64 servers (to be used for EOS front-end node) equipped with two SAS HBA cards and two 40 Gbps Ethernet NICs for each, 18 high-density JBOD enclosures capable to accommodate up to 84 disks in each box attached to the servers in pars, and two 40 Gbps Ethernet switches for the interconnection. The unit capacity of each installed disk is 12 TB and the nominal raw capacity of the storage in total is summed up 18,144 TB. Each of JBOD enclosures is attached to EOS FE node via two 12 Gbps SAS HBA cards and the maximum transfer rate is about 6 GB/s [7]. To minimize the reduction on usable space because of RAIN configuration (to be discussed in the following section), we deployed two FST components

– the EOS component managing attached storage – on a single EOS FE node using container technology so that we can increase the total FST nodes up to 18. The other key EOS components such as MGM, MQ and QDB are as well deployed upon containers. In order to achieve the high availability of the storage system, we have three MGM nodes including MQ daemons – the manager nodes using message queue system for the communication across the EOS components – and three QDB nodes in cluster for the namespace.

## 3 QRAIN Layout

In this section, among other useful features of EOS, we discuss its erasure coding implementation across multiple nodes, which is known as Redundant Array of Independent Nodes (RAIN). RAIN is a type of file storage layouts provided by EOS. Like Redundant Array of Inexpensive Disks (RAID), RAIN provides a certain level of data protection with multiple nodes by providing some number of parity nodes. For example, `raid5`, `raid6` or `raiddp` layout provides the same level of data protection with RAID5 or RAID6 by providing 1 or 2 parity nodes. The EOS even provides 3 or 4 parity configuration, which are named as `archive` or `qrain` respectively, that allows us to achieve higher level of data protection expected to be comparable to tape. One more remarkable feature of EOS we should mention here is that it is allowed for all of layouts to increase the number of stripes, i.e. data nodes plus parity nodes. For example, `raid6` can be extended to 10 nodes (or more) from 6 while the number of parity node is unchanged. In such a way, one can expect more usable space however note that some level of data protection must be compromised. There is also a flexibility on RAIN configuration in which one can configure different layouts on different directories in a single EOS storage instance.

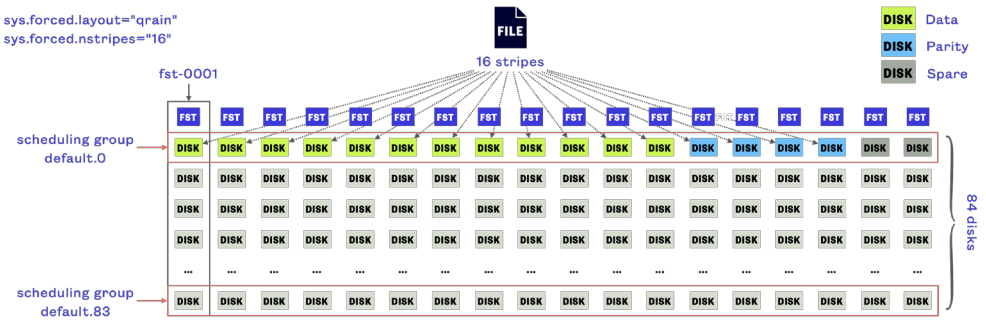



**Figure 2.** QRAIN layout configuration with 16 stripes and 2 spares.

The RAIN layout we configured for the disk-based custodial storage is depicted in Figure 2. We chose `qrain` layout configuration, the 4 parity mode, in line with the first principle of the system architecture described in Section 2. In general, RAIN configuration is applied to a scheduling group which consists of each of filesystems provided by FSTs in the EOS storage. For example, as shown in Figure 2, `default.0` scheduling group consists of 18 disks (filesystems) across 18 FSTs, which are the first disks of each JBOD enclosures. We configured to distribute a file in 16 stripes to keep the data protection level high enough while not to sacrifice much of usable capacity. Note that we spare 2 disks in a scheduling group so that it can help facilitate operations and maintenance, i.e. if needed, we can turn off one EOS FE node, where 2 FSTs are running, at any time without affecting the system availability.

The reduction rate of the usable capacity by the `qrain` configuration with 16 stripes is about 33.3% which gives 12 PB of usable space. The estimated data loss rate in a year is

approximately $8.6 \times 10^{-5}$%, which is equivalent to having 5 disks out of 16 failing simultaneously, if we consider 1.17% of the Annualized Failure Rate (AFR) in practice (c.f. the vendor published AFR is 0.35%) [9].

## 4 EOS Deployment

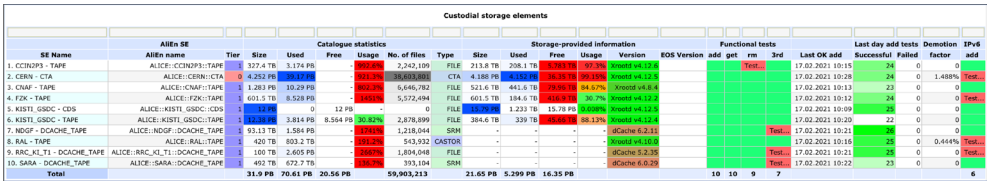
To ease operations and maintenance, like many other Tier centres in WLCG, we rely on configuration management and automation frameworks such as Foreman [10], Puppet [11] and Ansible [12]. The essential system configurations of EOS FE nodes, e.g. kernel parameters, host network scripts and IPv4/IPv6 information, local disk partitions and JBOD mounts, are centrally managed and applied on the fly when provisioning new physical node through the Foreman and the Puppet. Those crucial parameters are kept synchronized via the Puppet agents when they check and validate host configurations periodically.

In addition, we utilize the Ansible, which is a simple but powerful automation tool for infrastructure configuration and management, to deploy EOS components upon containers. We defined essential tasks as Ansible playbook and wrote a configuration manuscript file in `yaml` format [13]. The playbook contains common tasks – essential packages and services setup, docker installation, base image build and distribution – and selective tasks to run EOS components properly upon containers. The roles of each of EOS FE nodes, i.e. which host to run MGM and 2 FSTs, and which host to run QDB and 2 FSTs, and so on, are defined in the configuration file above-mentioned. Common tasks followed by selective tasks referenced the configuration file were executed on EOS FE nodes through the Ansible to deploy the EOS components as described in Section 2. Since we started deploying and testing EOS with `qrain` layout for the custodial storage system, there has been tens of new EOS releases that include fixes and introduce new features. Thanks to the automation via the Ansible, we could easily and repeatedly re-deploy the EOS components from scratch with a few changes.

A few critical issues were identified and fixed during the testing: 1) the layout information shared among nodes locating chunks of a file easily exceeds hard limit of 2kB when we have 12 or more stripes 2) the namespace was likely corrupted due to misconduct of quota information update during the master-slave transition. The EOS version 4.8.31 or later have the fixes on those issues.

## 5 ALICE integration

Additional steps were required to be integrated as storage element for the ALICE experiment. The steps include enabling ALICE token-based authentication and authorization, enabling ApMon daemon on all of EOS FST components for the ALICE MonALISA monitoring, allowing Third-Party Copy (TPC) in EOS, and finally enabling IPv6 networking.



Custodial storage elements

| | AliEn SE | | Catalogue statistics | | | | | | Storage-provided information | | | | | | Functional tests | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | Last day add tests Successful | Failed | Demotion factor | IPv6 add |
| 1. CCIN2P3 - TAPE | ALICE::CCIN2P3::TAPE | 1 | 327.4 TB | 3.174 PB | - | 992.6% | 2,242,109 | FILE | 213.8 TB | 208.1 TB | 5.783 TB | 97.3% | Xrootd v4.12.6 | | | | Test... | | 17.02.2021 10:15 | 24 | 0 | 0 | |
| 2. CERN - CTA | ALICE::CERN::CTA | 0 | 4.252 PB | 30.77 PB | - | 621.3% | 38,603,083 | CTA | 4.188 PB | 4.155 PB | 36.35 TB | 99.13% | Xrootd v4.12.5 | | | | | | 17.02.2021 10:28 | 24 | 0 | 1.488% | Test... |
| 3. CNAF - TAPE | ALICE::CNAF::TAPE | 1 | 1.283 PB | 10.29 PB | - | 902.3% | 5,646,782 | FILE | 521.6 TB | 441.5 TB | 78.95 TB | 84.57% | Xrootd v4.8.4 | | | | | | 17.02.2021 10:13 | 23 | 0 | 0 | |
| 4. FZK - TAPE | ALICE::FZK::TAPE | 1 | 601.5 TB | 8.528 PB | - | 1451% | 5,572,494 | FILE | 601.5 TB | 184.6 TB | -816.9 TB | 30.7% | Xrootd v4.12.2 | | | | | | 17.02.2021 10:12 | 24 | 0 | 0 | Test... |
| 5. KISTI_GSDC - CDS | ALICE::KISTI_GSDC::CDS | 1 | 12 PB | 0 | 12 PB | - | 0 | FILE | 15.29 PB | 1.233 TB | 15.78 PB | 0.008% | Xrootd v4.12.5 | | | | | | 17.02.2021 10:09 | 25 | 0 | 0 | |
| 6. KISTI_GSDC - TAPE | ALICE::KISTI_GSDC::TAPE | 1 | 12.31 PB | 3.814 PB | 8.564 PB | 30.82% | 2,878,899 | FILE | 384.6 TB | 339 TB | -45.66 TB | 88.13% | Xrootd v4.12.4 | | | | | | 17.02.2021 10:20 | 22 | 0 | 0 | |
| 7. NDGF - DCACHE_TAPE | ALICE::NDGF::DCACHE_TAPE | 1 | 93.13 TB | 1.584 PB | - | 1741% | 1,218,044 | SRM | - | - | - | - | dCache 6.2.11 | | | | | Test... | 17.02.2021 10:21 | 24 | 0 | 0 | |
| 8. RAL - TAPE | ALICE::RAL::TAPE | 1 | 420 TB | 803.2 TB | - | 191.2% | 543,932 | CASTOR | - | - | - | - | Xrootd v4.10.0 | | | | | | 17.02.2021 10:15 | 25 | 0 | 0.444% | Test... |
| 9. RRC_KI_T1 - DCACHE_TAPE | ALICE::RRC_KI_T1::DCACHE_TAPE | 1 | 100 TB | 2.605 PB | - | 2667% | 1,804,048 | FILE | - | - | - | - | dCache 5.2.35 | | | | | Test... | 17.02.2021 10:21 | 25 | 0 | 0 | Test... |
| 10. SARA - DCACHE_TAPE | ALICE::SARA::DCACHE_TAPE | 1 | 492 TB | 672.7 TB | - | 136.7% | 393,104 | SRM | - | - | - | - | dCache 6.0.29 | | | | | Test... | 17.02.2021 10:22 | 23 | 0 | 0 | |
| Total | | | 31.9 PB | 70.61 PB | 20.56 PB | | 59,903,213 | | 21.65 PB | 5.299 PB | 16.35 PB | | | | 10 | 10 | 9 | 7 | | | | | 6 |

**Figure 3.** The list of ALICE custodial storage element shown in MonALISA monitoring site.

The disk-based custodial storage is currently integrated as Tape Storage Element, which is now renamed Custodial Storage Element, for the ALICE experiment in the name of

`ALICE::KISTI_GSDC::CDS` as shown in Figure 3 [14]. The raw capacity of the storage reported by ApMon daemons running on EOS FSTs and the usable space given by 4-parity RAIN layout are 15.79 PB (1024-base) and 12 PB, respectively. Note that the actual size of the installed 12 TB disk is 11,478,771,712 bytes which equals to 10.7 TB in 1024-base and this gives 10.7 TB × 84 disks × 18 enclosures ÷ 1024 = 15.79 PB, while the EOS reports 17.7 PB (probably 1000-base). Since the end of January 2021, `ALICE::KISTI_GSDC::CDS` has passed the periodic functional tests without any issues. The tests include `add` (write), `get` (read), `rm` (delete) and `3rd` (TPC), which are running hourly and their results are correlated with WLCG Availability/Reliability test framework [15]. The same tests are running for IPv6 network as well.

## 6 Monitoring

As our custodial storage system is based on cheap JBOD enclosures purchased without any vendor enterprise-class solutions or any relevant technical support contracts for the system administration and management for monitoring and alerting, we rely on S.M.A.R.T (Self-Monitoring, Analysis and Reporting Technology) information provided by an open-source tool such as smartmontools [18] or a Linux CLI tool such as `sg_ses` for hardware-level monitoring and fault detection. We collect SCSI information of all disks and JBOD enclo-

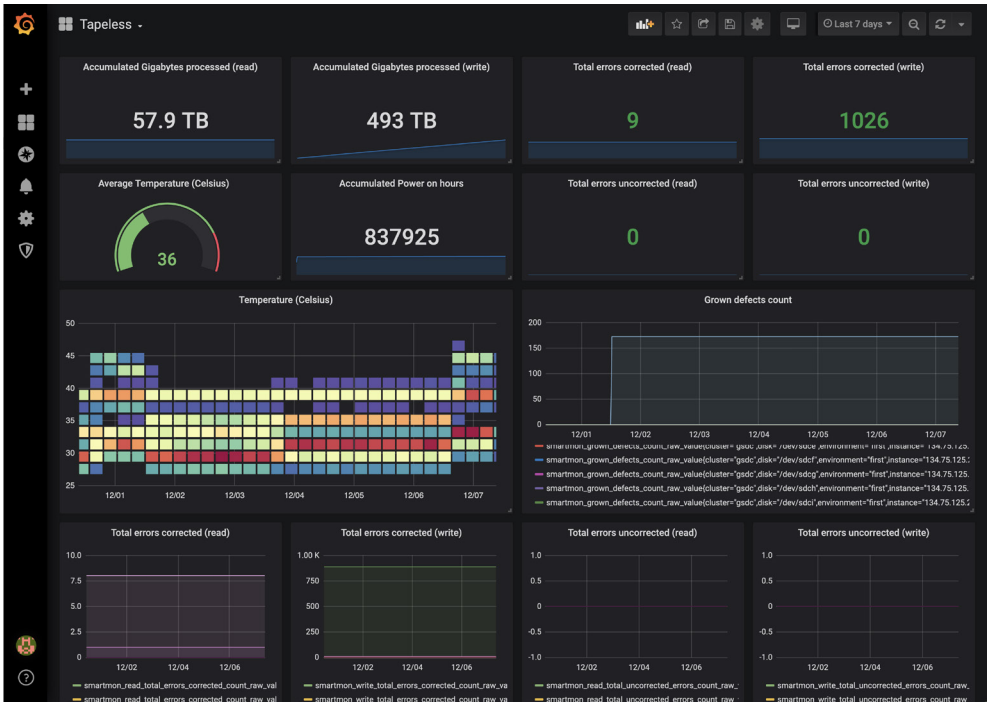



**Figure 4.** Grafana dashboard for hardware-level monitoring.

sures from these tools and transmit the information to a time-series database provided by Prometheus [16] through its *node_exporter*, and Grafana [17] is used for visualizing queries on the collected information such as corrected or uncorrected error counts, defect counts and temperature as shown in Figure 4. We anticipate that we could learn and predict to some

extent when and which disks or enclosures will fail based on this monitoring framework. Also we use the Grafana to facilitate log monitoring on all of EOS components running on containers through Loki [19] plugin.

## 7 Conclusion

In order to avoid the potential tape market risks, we proposed a disk-based custodial storage system as an alternative to tape for preserving data produced from the ALICE experiment at CERN. The proposed storage relies on the cheap high-density JBOD enclosures to reduce cost and the EOS QRAIN layout to achieve high enough data protection comparable to tape. Thanks to the system architecture consisting of 18 high-density JBOD enclosures attached to 9 EOS FE nodes where all necessary EOS components are running upon containers, the estimated annual data loss rate is down to $8.6 \times 10^{-5}$%. The deployment of EOS components upon containers are fully automated to facilitate operations and maintenance, and a home-grown monitoring is being built for hardware-level error detection and prediction.

The proposed storage is currently integrated and is being commissioned as custodial storage element for the ALICE experiment. We plan to make the storage fully operational by July 2021, originally the start of LHC RUN3 which has been postponed until next year due to the COVID-19 pandemic. And finally we will try to extract useful metrics to be compared with those of tape such as long-term power consumption, real-world data loss rate, and total-cost of ownership that might be helpful for those who look for alternatives.

## Acknowledgement

The authors thank the EOS developers for their dedicated effort, kind help, and prompt actions to identify and fix the issues found in the deployment and testing. This work was supported by the National Research Foundation of Korea (NRF) through contract N-21-NM-CR01-S01 and the Program of Construction and Operation for Large-scale Science Data Center (K-21-L02-C04-S01).

## References

[1] The ALICE Collaboration, Journal of Instrumentation **3**, S08002 (2008)

[2] *Worldwide LHC Computing Grid*, https://wlcg.web.cern.ch (2021), accessed: 2020-02-28

[3] *Did Oracle just sign tape's death warrant? Depends what 'no comment' means*, https://www.theregister.co.uk/2017/02/17/oracle_streamline_tape_library_future (2017), accessed: 2021-02-28

[4] *The Future of the Cloud Depends on Magnetic Tape*, https://www.bloomberg.com/news/articles/2018-10-17/the-future-of-the-cloud-depends-on-magnetic-tape (2018), accessed: 2021-02-28

[5] *LTO-8 tape media patent lawsuit cripples supply as Sony and Fujifilm face off in court*, https://www.theregister.co.uk/2019/05/31/lto_patent_case_hits_lto8_supply (2019), accessed: 2021-02-28

[6] *Sony, Fujifilm storage patent lawsuit is all taped up: Better LTO-8 than never, right?*, https://www.theregister.co.uk/2019/08/06/sony_fujifilm_storage_patent_lawsuit_settled (2019), accessed: 2021-02-28

[7] S. U. Ahn et al, EPJ Web of Conferences **245**, 04001 (2020)

[8] *EOS Open Storage*, https://eos.web.cern.ch (2021), accessed: 2021-02-28
[9] *Backblaze Hard Drive Stats Q2 2020*, https://www.backblaze.com/blog/backblaze-hard-drive-stats-q2-2020/ (2020), accessed: 2021-02-28
[10] *Foreman*, https://theforeman.org (2021), accessed: 2021-02-28
[11] *Puppet*, https://puppet.com (2021), accessed: 2021-02-28
[12] *Ansible*, https://www.ansible.com (2021), accessed: 2021-02-28
[13] *YAML Ain't Markup Language*, https://yaml.org (2021), accessed: 2021-02-28
[14] *MonALISA Repository for ALICE*, http://alimonitor.cern.ch/stats?page=SE/table, accessed: 2021-02-28
[15] *WLCG SITEMON*, https://monit-wlcg-sitemon.web.cern.ch/monit-wlcg-sitemon/ (2021), accessed: 2021-02-28
[16] *Prometheus*, https://prometheus.io (2021), accessed: 2021-02-28
[17] *Grafana*, https://grafana.com/oss/grafana/ (2021), accessed: 2021-02-28
[18] *smartmontools*, https://www.smartmontools.org (2021), accessed: 2021-02-28
[19] *Grafana Loki*, https://grafana.com/oss/loki/ (2021), accessed: 2021-02-28