Status of Tape-less Archive Storage project @ KISTI-GSDC

AHN SANG-UN @ HSF-WLCG VIRTUAL WORKSHOP, 19-23 NOVEMBER 2020

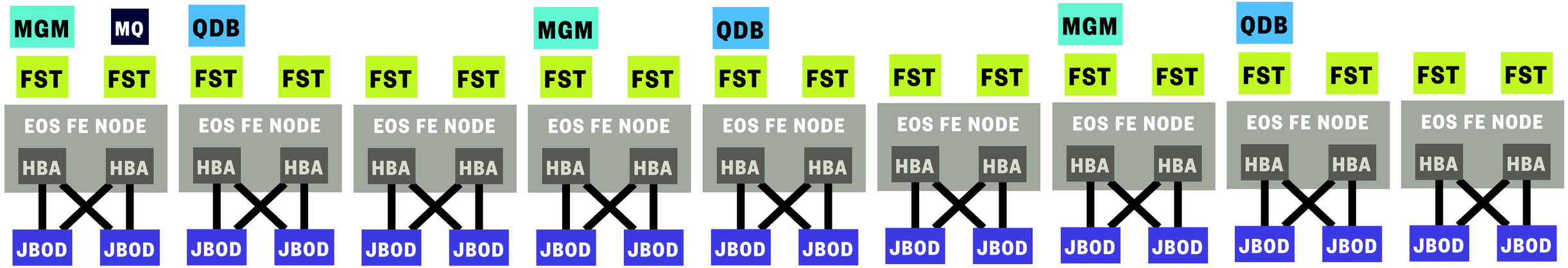
Outline

- Introduction
- System Architecture
- QRAIN Layout
- Current Status
- Monitoring
- Plan
- Summary

Introduction

- Replacing tape library (+3PB) with disk-only storage for archiving @ KISTI for ALICE experiment
  - Simpler architecture, less operational efforts, cost-effective comparable to tape
- Found domestic suppliers of high-density(> 60 disks/box) JBOD models
- Relying on EOS erasure coding implementation (RAIN layout) for data protection
- About 1M CHF budget (2019) included
  - 18 High-density JBOD boxes (84 disks/box ≃ 18PB raw capacity)
  - 9 Servers for EOS front-end nodes (12Gbps SAS HBAs, 40Gbps uplinks + switches)
- Providing production service to ALICE before the start of RUN3 (by June 2021) ← POSTPONED

System Architecture



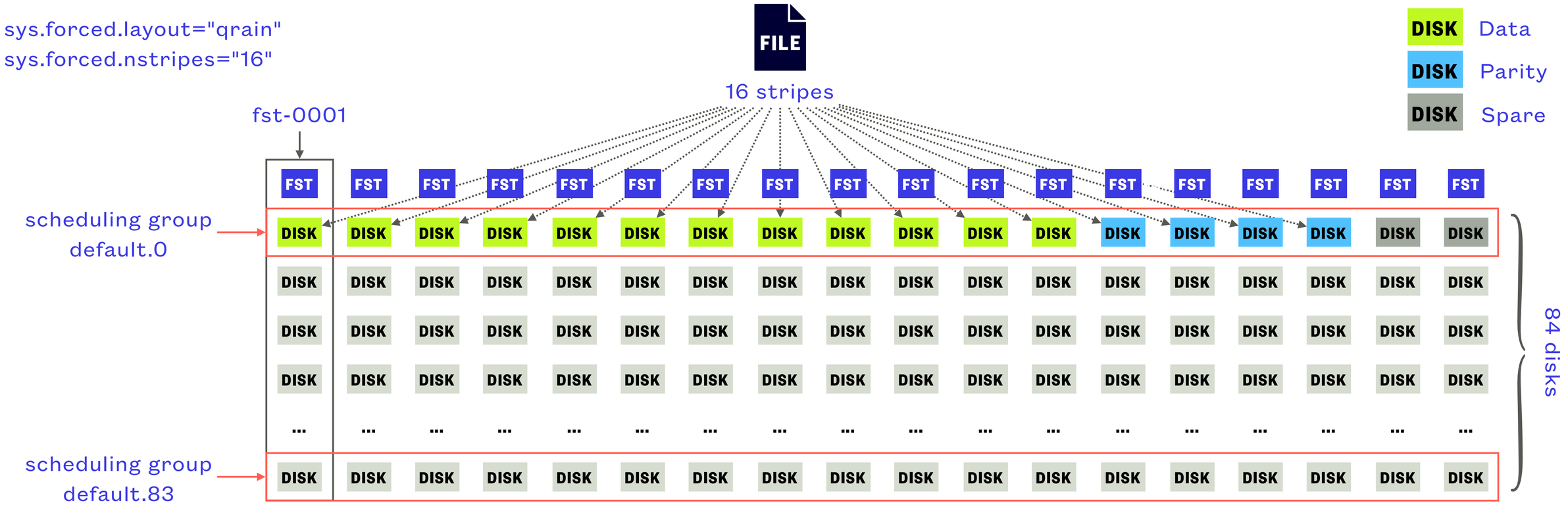
9 servers, 18 boxes

84 Disks in one box

- Total raw capacity = 18,144TB (= 12TB * 84 disks * 18 boxes)
- EOS version = 4.8.12 (20200907174735gitcf98311), cf. the latest tag(stable) = 4.7.7
  - Including 2 fixes for redirection info longer than 2kB (critical for layouts with many stripes)
- EOS components are running on containers (a fork of EOS-Docker project)
  - Ansible playbook available at https://github.com/jeongheon81/gsdc-eos-docker

QRAIN Layout

sys.forced.layout="qrain"
sys.forced.nstripes="16"



- Thanks to spare FSTs,
  - Data are still accessible if 6 FSTs are offline
  - Data can be written if 2 FSTs are offline
  - One node (= 2 FSTs) can be turned off for maintenance at any time
- Data loss rate in a year is $\approx 8.6 \times 10^{-5}\%$, where 5 disks are failed simultaneously, considering 1.17% of AFR in practice
  cf. vendor published AFR is 0.35% (AFR = Annualized Failure Rate)

Fileinfo

EOS fileinfo command

Layout type
# of stripes
# of replica

File chuck location
Scheduling group
Filesystem status

```
sh-4.2# eos fileinfo /eos/gsdc/testarea/rain16/testfile.10G
  File: '/eos/gsdc/testarea/rain16/testfile.10G'  Flags: 0640
  Size: 10485760000
Modify: Thu Oct 22 00:01:35 2020 Timestamp: 1603324895.724750000
Change: Thu Oct 22 00:00:51 2020 Timestamp: 1603324851.619542497
 Birth: Thu Oct 22 00:00:51 2020 Timestamp: 1603324851.619542497
  CUid: 0 CGid: 0 Fxid: 0000159b Fid: 5531 Pid: 40 Pxid: 00000028
XStype: adler    XS: a1 1c 00 01    ETAGs: "1484716507136:a11c0001"
Layout: qrain Stripes: 16 Blocksize: 1M LayoutId: 40640f52 Redundancy: d5::t0
  #Rep: 16
```

no.	fs-id	host	schedgroup	path	boot	configstatus	drain	active	geotag
0	995	jbod-mgmt-06.sdfarm.kr	default.70	/jbod/box_12_disk_070	booted	rw	nodrain	online	kisti::gsdc::g02
1	1499	jbod-mgmt-09.sdfarm.kr	default.70	/jbod/box_18_disk_070	booted	rw	nodrain	online	kisti::gsdc::g03
2	659	jbod-mgmt-04.sdfarm.kr	default.70	/jbod/box_08_disk_070	booted	rw	nodrain	online	kisti::gsdc::g02
3	407	jbod-mgmt-03.sdfarm.kr	default.70	/jbod/box_05_disk_070	booted	rw	nodrain	online	kisti::gsdc::g01
4	827	jbod-mgmt-05.sdfarm.kr	default.70	/jbod/box_10_disk_070	booted	rw	nodrain	online	kisti::gsdc::g02
5	491	jbod-mgmt-03.sdfarm.kr	default.70	/jbod/box_06_disk_070	booted	rw	nodrain	online	kisti::gsdc::g01
6	1079	jbod-mgmt-07.sdfarm.kr	default.70	/jbod/box_13_disk_070	booted	rw	nodrain	online	kisti::gsdc::g03
7	71	jbod-mgmt-01.sdfarm.kr	default.70	/jbod/box_01_disk_070	booted	rw	nodrain	online	kisti::gsdc::g01
8	743	jbod-mgmt-05.sdfarm.kr	default.70	/jbod/box_09_disk_070	booted	rw	nodrain	online	kisti::gsdc::g02
9	1247	jbod-mgmt-08.sdfarm.kr	default.70	/jbod/box_15_disk_070	booted	rw	nodrain	online	kisti::gsdc::g03
10	155	jbod-mgmt-01.sdfarm.kr	default.70	/jbod/box_02_disk_070	booted	rw	nodrain	online	kisti::gsdc::g01
11	1415	jbod-mgmt-09.sdfarm.kr	default.70	/jbod/box_17_disk_070	booted	rw	nodrain	online	kisti::gsdc::g03
12	911	jbod-mgmt-06.sdfarm.kr	default.70	/jbod/box_11_disk_070	booted	rw	nodrain	online	kisti::gsdc::g02
13	1331	jbod-mgmt-08.sdfarm.kr	default.70	/jbod/box_16_disk_070	booted	rw	nodrain	online	kisti::gsdc::g03
14	239	jbod-mgmt-02.sdfarm.kr	default.70	/jbod/box_03_disk_070	booted	rw	nodrain	online	kisti::gsdc::g01
15	575	jbod-mgmt-04.sdfarm.kr	default.70	/jbod/box_07_disk_070	booted	rw	nodrain	online	kisti::gsdc::g02

```
*******
```

For a single file,
Read: 800-1200MB/s
Write: 200-300MB/s

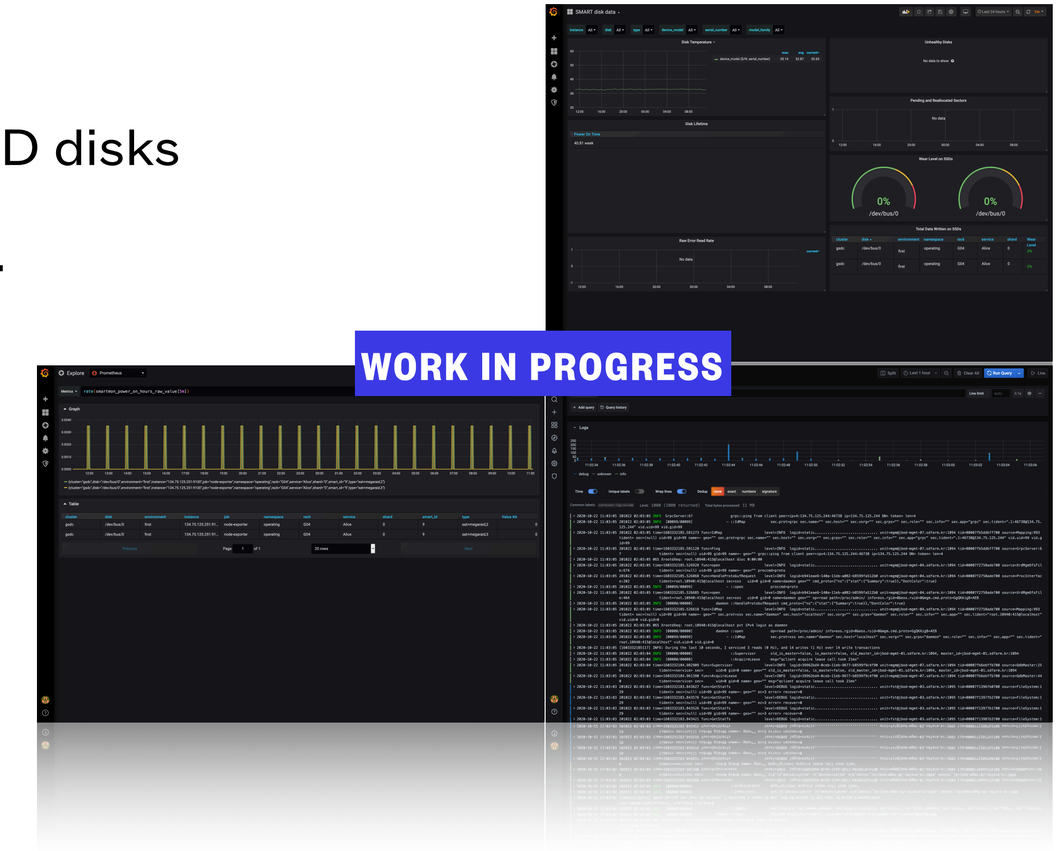
cf. KISTI tape total throughput (w/ 8 drives) ≃ 2GB/s

Current Status

- Focused on ensuring that QRAIN layout is working as expected
  - Great thanks to EOS developers for supporting this project
    - Identifying an issue regarding file access failure with eoscp protocol due to the exceed of hard limit on size of redirection request URL (= 2kB), which can be easily happening on any RAIN layout with many stripes ($\geq$ 12)
  - Helpful posts in EOS Community (https://eos-community.web.cern.ch/)
- Working on maintenance and operation schema, and maintenance automation code
  - Disk replacement, JBOD and/or server maintenance
  - Rolling update/upgrade of EOS components such as QDB, MGM, and FST

Monitoring

- Prometheus node_exporter + Grafana dashboard
  - Hardware level monitoring using *smartctl* on JBOD disks
    - Health check, temperature, error counters, etc.
  - Docker container health check
  - EOS services log dump using *loki, promtail*
  - Server monitoring
  - Alerting



Plan

- Updating EOS to the latest commit release that includes recent fixes and improvements
  - Ensuring that QRAIN layout is working well
- Creating a public end-point with a proper redundancy (recognizing multi-MGMs underneath)
  - Dynamic update of DNS records
- Enabling token based authentication for ALICE
- Integrating as a new ALICE tape storage element and performing periodic functional tests
- Monitoring and improving stability and reliability

Summary

- Working on providing a disk-only archive storage for ALICE experiment with the help of EOS erasure coding implementation for data protection
- Successfully deployed QRAIN layout with 16 stripes including 4 parities and made it working by fixing a couple of issues
- Working on establishing a dedicated monitoring framework for the archive storage
- On track of schedule to provide production service by June 2021, even though the start of RUN3 has been postponed

Thank you