

Sang-un Ahn, Heejun Hahn, Jeong-heon Kim

Global Science experimental Data hub Center (GSDC)
Korea Institute of Science and Technology Information (KISTI)

ATCF6, 2022

Custodial Disk Storage for the ALICE Experiment using EOS

Outline

- Introduction
- CDS Architecture
- QRAIN Layout & Configuration
- Current Status
- Operations: WLCG Tape Challenge, Production for ALICE, Power Consumption
- Plan
- Practices

Introduction

- CDS - a disk based storage designed to store and preserve RAW data from the ALICE experiment by accommodating EOS with its erasure code implementation, a.k.a RAIN configuration
 - Replacing the existing tape library at KISTI (~ 3.2PB)
 - Simplifying architecture hoping for cost reduction
 - Removing additional disk buffers (~ 0.6PB) in front of tape library for I/O
 - Being free from commercial (vendor-specific) software for HSM operations
 - Avoiding vendor lock-in due to monopoly in Tape market
- Provided to the ALICE experiment for commissioning at the early of 2021
- In production since November 2021, replacing the tape storage completely

Recap CHEP2019

<https://doi.org/10.1051/epjconf/202024504001>

- EOS Erasure Coding implementation => RAIN layout
 - 2 FSTs (data server) on a single FE node to maximize usable space (~70%) out of raw capacity
- Upper cap on total throughput limit by PCI-e 3.0 (~6GB/s)
- Power consumption ~ 1.75W/TB (JBOD+Server+S/W)
 - Enterprise Disk storage: 5 ~ 9W/TB
 - Tape library ~ 0.5W/TB

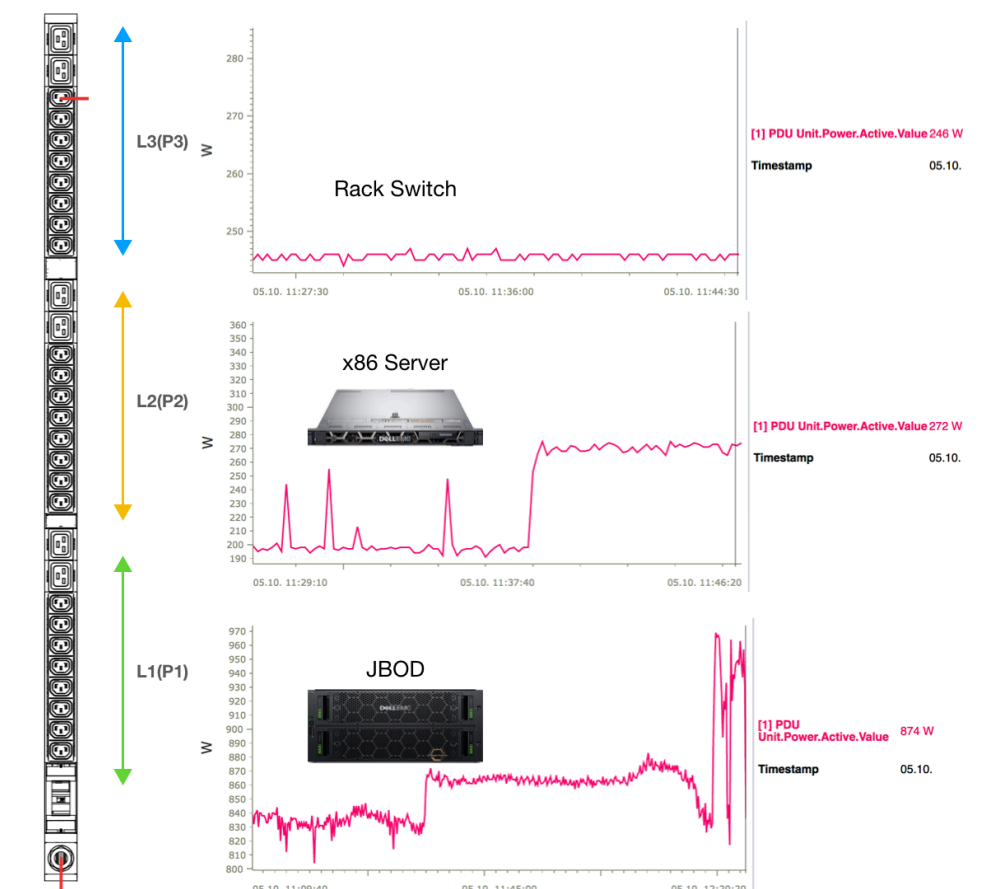
I/O Test: Read/Write

- XFS read/write performance (simultaneous read and/or write from 70 disks)
 - **VDBench** shows full read/write transfer performance @ transfer size >= 2048k (6GB/s)
 - **IOZone** shows full read/write transfer performance @ transfer size ~ 2048k (6GB/s)



Power Consumption

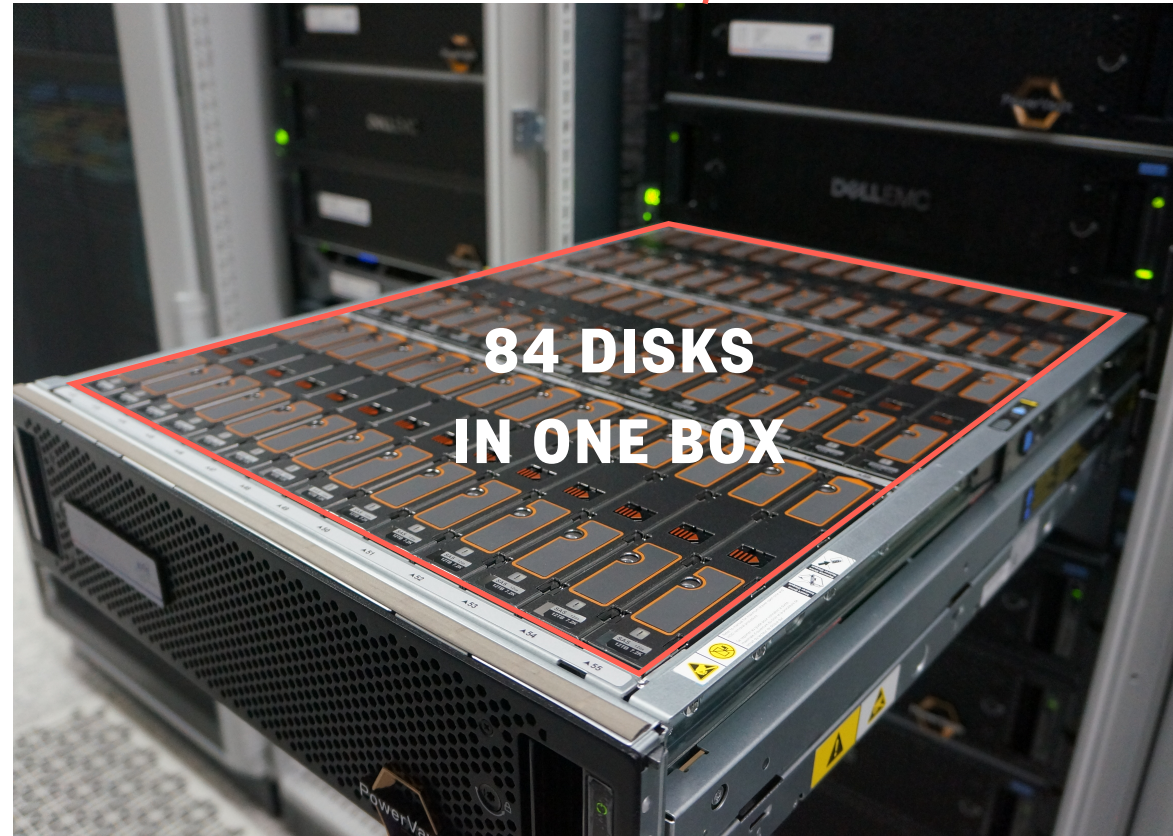
- JBOD Test Equipment (70 Disks)
 - JBOD (DELL ME484): idle = 830W; load = 860W (Max 960) (**1.12W/TB**)
 - Server: idle = 200W; load = 270W
 - Switch: idle = 246W; load = 246W
 - **1.75W/TB** including JBOD, Server and Switch
- Disk Storages (Full Load)
 - DellEMC SC7020, 2.5PB - 12,120W (**4.8W/TB**)
 - EMC Isilon, 16 Nodes, 2.95 PB- 13,730W (**4.6W/TB**)
 - EMC VNX, 12 Nodes, 2.36 PB - 5,100W (**2.2W/TB**)
 - HITACHI VSP, 2 PB - 18,300W (**9.15W/TB**)
 - EMC Isilon, 15 Nodes, 1.43 PB - 12,880W (**9W/TB**)
 - EMC CX4-960, 1.5PB - 14,900W (**9.9W/TB**)
- Tape Library (Full Load)
 - IBM TS3500 5-Frame (3.2PB) - 1,600W (**0.5W/TB**)



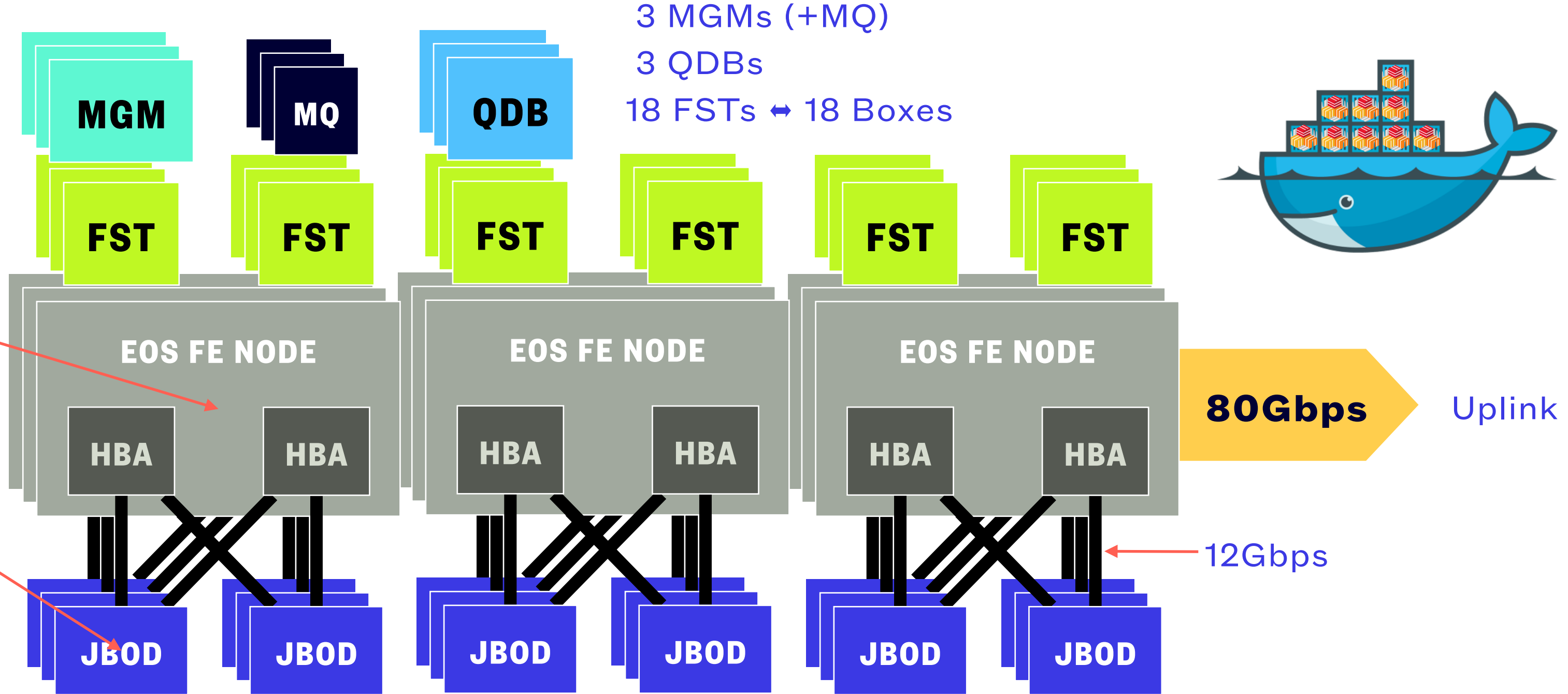
CDS Architecture



9 servers
18 boxes



84 DISKS
IN ONE BOX

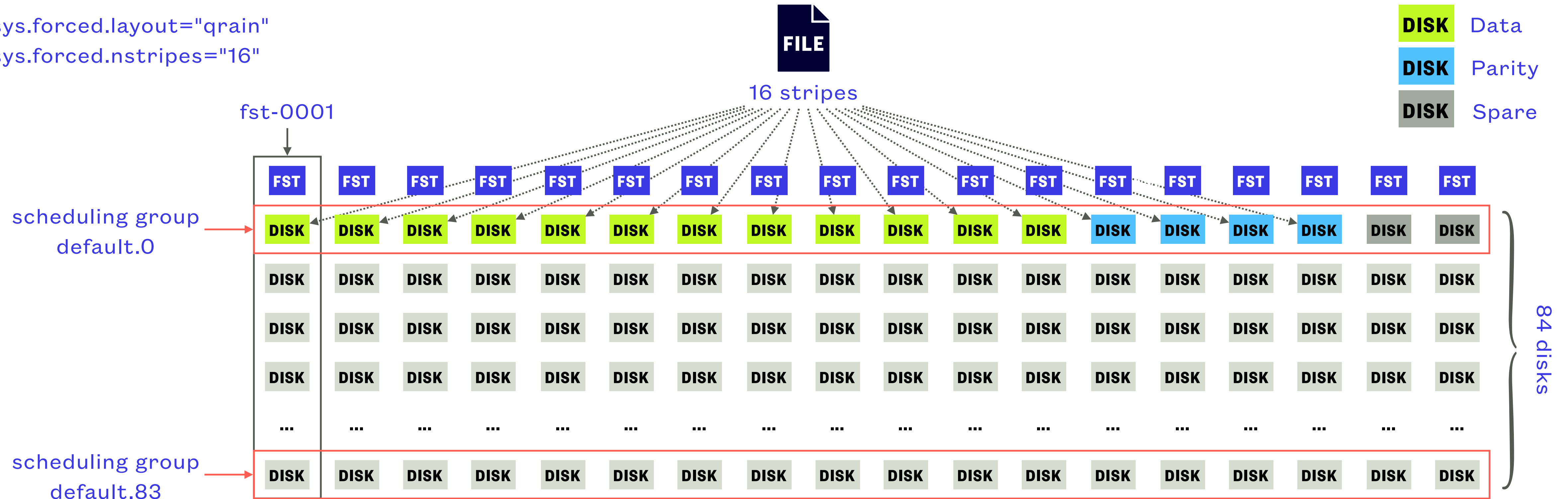


- Total raw capacity = 18,144TB (= 12TB * 84 disks * 18 boxes)
- EOS version = 4.8.82 (released on 2022.4.12)
- EOS components are running on containers (a fork of EOS-Docker project)
 - Ansible playbook available at <https://github.com/jeongheon81/gsd-c-eos-docker>

QRAIN Layout



sys.forced.layout="grain"
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)

QRAIN Configuration

- 'eos attr' command
 - One can have different layouts on different directories (or files) in an EOS instance
 - Available layouts = plain (default, 1 single copy); replica (2 copies); raid6, raiddp (2 parities); archive (3 parities); qrain (4 parities), ...

```
eos attr -r set default=raid6 /eos/gsdctestarea/raid6
eos attr -r set default=archive /eos/gsdctestarea/archive
eos attr -r set default=replica /eos/gsdctestarea/replica
eos attr -r set default=qrain /eos/gsdctestarea/rain12
```

- # of stripes can be changed, e.g. 3 copies, 16 stripes...

```
eos attr -r set default=qrain /eos/gsdctestarea/rain16
eos attr -r set sys.forced.nstripes=16 /eos/gsdctestarea/rain16
```

```
sh-4.2# eos attr ls /eos/gsdctestarea/rain16
sys.eos.btime="1605069261.927407367"
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="qrain"
sys.forced.nstripes="16"
sys.forced.space="default"
sys.recycle="/eos/gsdctestarea/recycle/"
```

Fileinfo



EOS fileinfo command

```
sh-4.2# eos fileinfo /eos/gsdctestarea/rain16/testfile.10G
```

```
File: '/eos/gsdctestarea/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 type
of stripes
of replica

```
Layout: grain Stripes: 16 Blocksize: 1M LayoutId: 40640f52 Redundancy: d5::t0
#Rep: 16
```

File chunk location
Scheduling group
Filesystem status

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

```
*****
```

Issues Fixed (EOS commits)

<https://eos-community.web.cern.ch/>

- When read a file with "eos cp" (or eoscp) from RAIN layout with more than 12 stripes
 - Redirection information over 2kB truncated
 - <https://gitlab.cern.ch/dss/eos/-/commit/4cb0f733650e041a3153da60610a8d5a0e4672f4>
 - <https://gitlab.cern.ch/dss/eos/-/commit/03bcc4b55556bc0d7b3160670a41a98bfa50a941>
- Failed to read a file with "eos cp" through MGM secondaries
 - Quota information propagation corrupting namespace
 - <https://gitlab.cern.ch/dss/eos/-/commit/32b93012459f73409b45fcea3c575cc6c47f421>

SPECIAL THANKS TO ELVIN ALIN SINDRILARU !!!

Current Status

- EOS version installed: 4.8.82
 - Automated deployment via Ansible playbook
- Public DNS name pointing to 3 MGMs (load-balancing)
- 40G NICs Teamed to provide 80G uplink bandwidth
- IPv4/IPv6 dual stack configured
- ALICE Integration
 - Enabling Token-based AuthN/AuthZ
 - Enabling ApMon daemons on all EOS FSTs for ALICE MonALISA monitoring
 - Allowing Third-Party Copy by disabling sss enforcement on FSTs

space view

```
[root@jbod-mgmt-01 MGM_MASTER=true /]# eos space ls
```

type	name	groupsize	groupmod	N(fs)	N(fs-rw)	sum(usedbytes)	sum(capacity)	capacity(rw)	nom.capacity	sched.capacity
spaceview	default	18	85	1512	1512	7.76 PB	17.77 PB	17.77 PB	0 B	10.01 PB

node view

```
[root@jbod-mgmt-01 MGM_MASTER=true /]# eos node ls
```

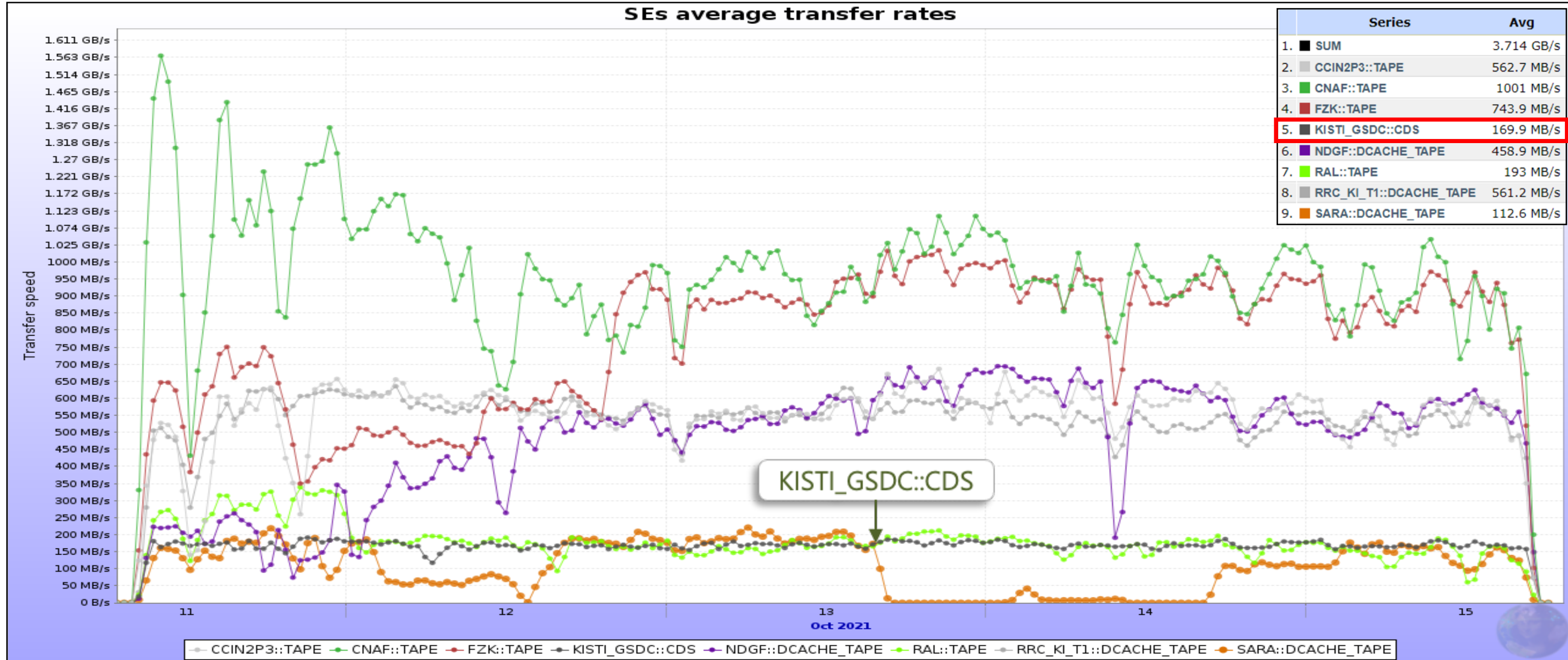
type	hostport	geotag	status	activated	txgw	gw-queued	gw-ntx	gw-rate	heartbeatdelta	nofs
nodesview	jbod-mgmt-01.sdfarm.kr:1095	kisti::gsdc:g01	online	on	off	0	10	120	1	84
nodesview	jbod-mgmt-01.sdfarm.kr:1096	kisti::gsdc:g01	online	on	off	0	10	120	1	84
nodesview	jbod-mgmt-02.sdfarm.kr:1095	kisti::gsdc:g01	online	on	off	0	10	120	1	84
nodesview	jbod-mgmt-02.sdfarm.kr:1096	kisti::gsdc:g01	online	on	off	0	10	120	2	84
nodesview	jbod-mgmt-03.sdfarm.kr:1095	kisti::gsdc:g01	online	on	off	0	10	120	3	84
nodesview	jbod-mgmt-03.sdfarm.kr:1096	kisti::gsdc:g01	online	on	off	0	10	120	1	84
nodesview	jbod-mgmt-04.sdfarm.kr:1095	kisti::gsdc:g02	online	on	off	0	10	120	2	84
nodesview	jbod-mgmt-04.sdfarm.kr:1096	kisti::gsdc:g02	online	on	off	0	10	120	3	84
nodesview	jbod-mgmt-05.sdfarm.kr:1095	kisti::gsdc:g02	online	on	off	0	10	120	2	84
nodesview	jbod-mgmt-05.sdfarm.kr:1096	kisti::gsdc:g02	online	on	off	0	10	120	2	84
nodesview	jbod-mgmt-06.sdfarm.kr:1095	kisti::gsdc:g02	online	on	off	0	10	120	2	84
nodesview	jbod-mgmt-06.sdfarm.kr:1096	kisti::gsdc:g02	online	on	off	0	10	120	2	84
nodesview	jbod-mgmt-07.sdfarm.kr:1095	kisti::gsdc:g03	online	on	off	0	10	120	2	84
nodesview	jbod-mgmt-07.sdfarm.kr:1096	kisti::gsdc:g03	online	on	off	0	10	120	3	84
nodesview	jbod-mgmt-08.sdfarm.kr:1095	kisti::gsdc:g03	online	on	off	0	10	120	1	84
nodesview	jbod-mgmt-08.sdfarm.kr:1096	kisti::gsdc:g03	online	on	off	0	10	120	2	84
nodesview	jbod-mgmt-09.sdfarm.kr:1095	kisti::gsdc:g03	online	on	off	0	10	120	2	84
nodesview	jbod-mgmt-09.sdfarm.kr:1096	kisti::gsdc:g03	online	on	off	0	10	120	2	84

EC attribute

```
[root@jbod-mgmt-01 MGM_MASTER=true /]# eos attr ls /eos/gsdg/grid
sys.eos.btime="1612374338.811408574"
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="grain"
sys.forced.nstripes="16"
sys.forced.space="default"
[root@jbod-mgmt-01 MGM_MASTER=true /]#
```


WLCG Tape Challenge (Oct 2021)

- Participation as a Tape (custodial storage) for the ALICE experiment
- Joined efforts of the WLCG Collaboration preparing for LHC RUN3 data taking
- Successful to meet the target (stable) transfer performance (150MB/s)



170MB/s on average for 5-day of transfer
101.4TB of data (51k files) transferred

Individual files 1.953GB, total transferred 1.766PB

Centre	Files	size
CCIN2P3	143230	279.7TB
CNAF	239913	468.6TB
GridKA	187327	368.9TB
KISTI	51914	101.4TB
RAL	45023	87.9TB
NDGF	100635	196.5TB
RRC_KI	110479	216.8TB
SARA	23566	46TB

CDS for the ALICE experiment

Current snapshot of the CDS in the ALICE monitoring system

<http://alimonitor.cern.ch/stats?page=SE/table>

Custodial storage elements																							
CDS																							
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. KISTI_GSDC - CDS	ALICE::KISTI_GSDC::CDS	1	15.79 PB	4.72 PB	11.07 PB	29.9%	10,856,926	FILE	15.79 PB	6.895 PB	8.89 PB	43.68%	Xrootd v4.12.8						15.11.2022 04:53	24	0	0	
Total			15.79 PB	4.72 PB	11.07 PB		10,856,926		15.79 PB	6.895 PB	8.89 PB				1	1	1	1					1

	Total	Used
Bin	15.79	6.89
Dec	17.77	7.76

ALICE RAW data replication to the CDS

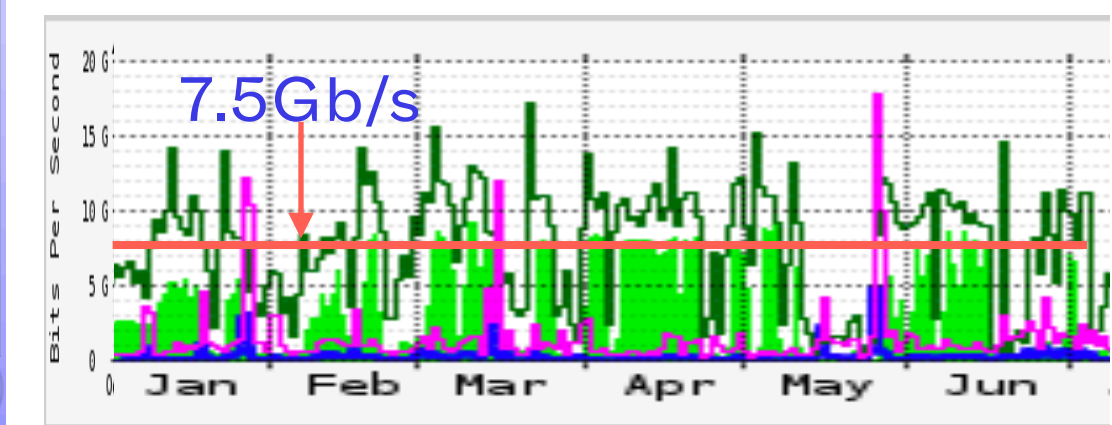
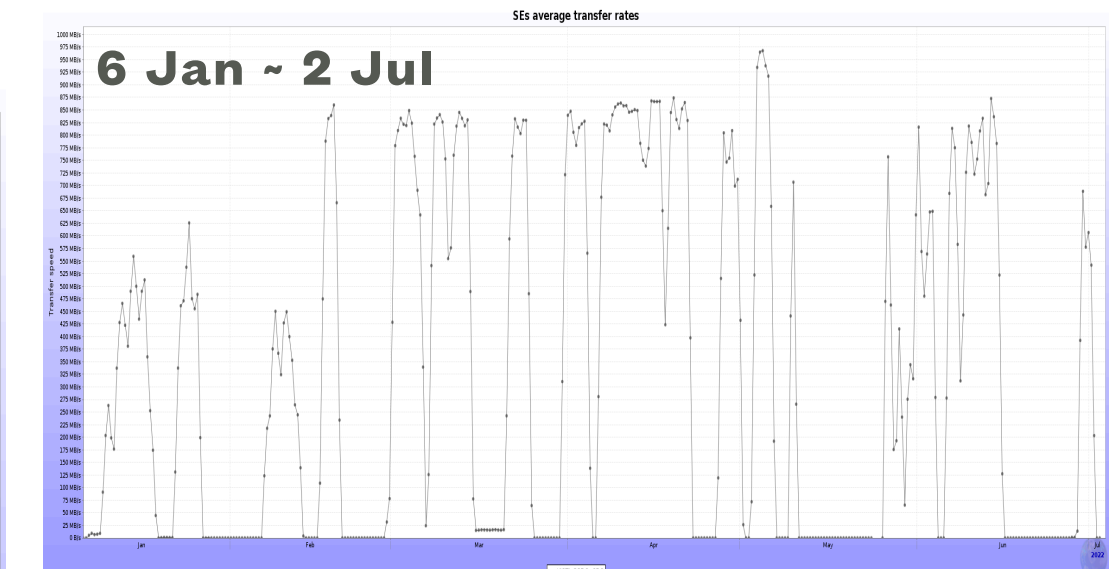
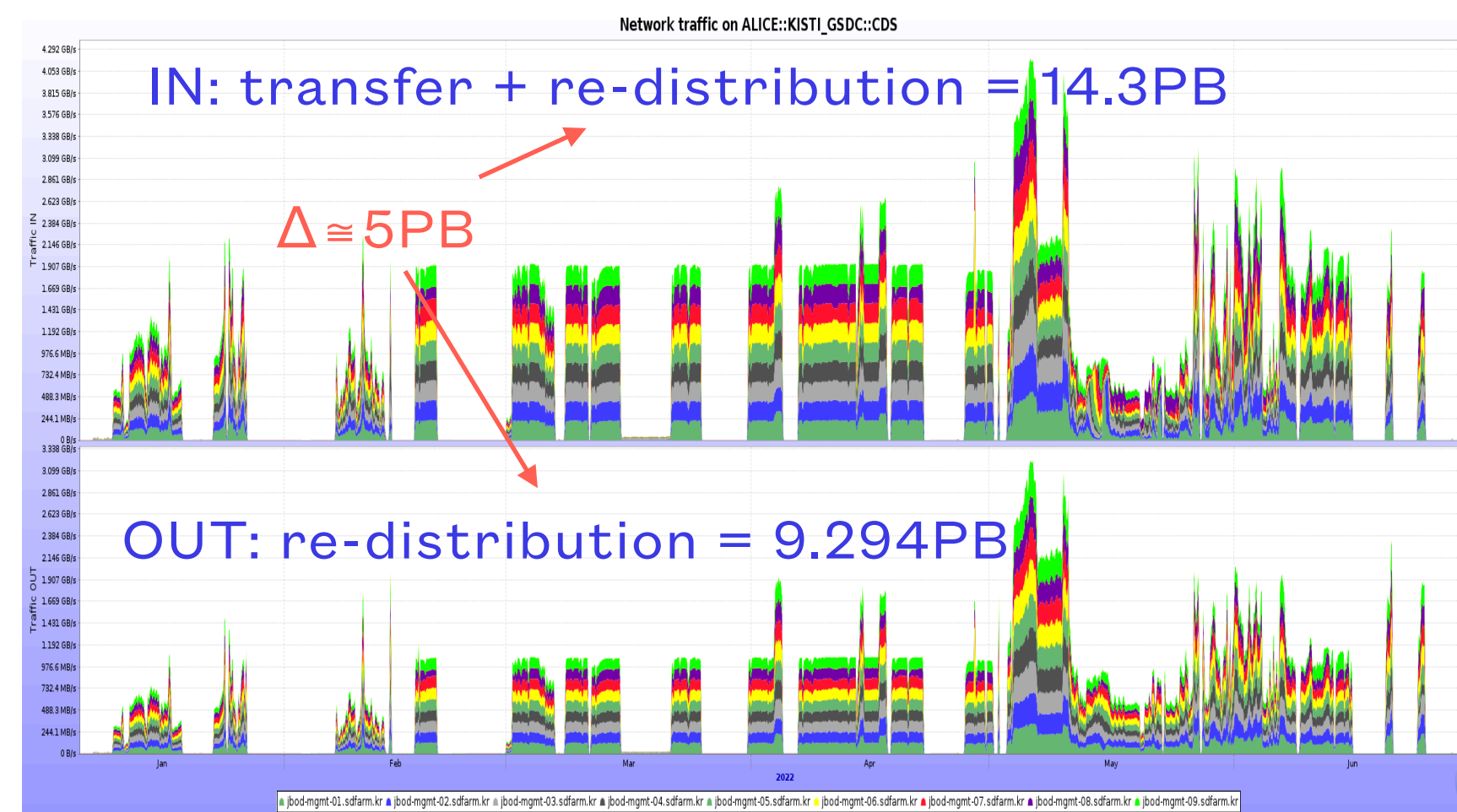
6 Jan ~ 2 Jul

Transfer requests (add new request)								
ID	Path	Target SE	Status	Progress	Files	Total size	Started	Ended
17816	/alice/data/2021/LHC21z/503650/GC/ECS/raw/2021-10-08_21-56/run0503650_2021-10-08T22_00_35Z	ALICE::KISTI_GSDC::CDS	Done		130000	125.3 TB	29 Jun 2022 22:51	02 Jul 2022 04:55
17774	/alice/data/2021/LHC21z/501384/GC/ECS/2021-09-16_13-14/run0501384_2021-09-16T13_15_51Z	ALICE::KISTI_GSDC::CDS	Error		330536	36.67 TB	31 May 2022 23:03	28 Jun 2022 14:50
17773	/alice/data/2021/LHC21z/501376/GC/ECS/2021-09-16_11-01/run0501376_2021-09-16T11_03_21Z	ALICE::KISTI_GSDC::CDS	Error		207776	23.06 TB	31 May 2022 22:49	28 Jun 2022 14:50
17772	/alice/data/2021/LHC21z/501354/GC/ECS/2021-09-16_02-25/run0501354_2021-09-16T02_26_15Z	ALICE::KISTI_GSDC::CDS	Error		1136720	125.7 TB	31 May 2022 22:14	16 Jun 2022 00:49
17771	/alice/data/2021/LHC21z/499999/GC/ECS/2021-09-15_17-06/run0501318_2021-09-15T17_07_23Z	ALICE::KISTI_GSDC::CDS	Error		335962	37.23 TB	31 May 2022 21:44	16 Jun 2022 00:47
17770	/alice/data/2021/LHC21z/499999/GC/ECS/2021-09-11_16-21/run0500983_2021-09-11T16_23_33Z	ALICE::KISTI_GSDC::CDS	Error		906778	101.3 TB	31 May 2022 21:10	16 Jun 2022 00:46
17769	/alice/data/2021/LHC21z/499999/tpc/MW3/tpc-20210330-magnet	ALICE::KISTI_GSDC::CDS	Error		231330	146.3 TB	31 May 2022 20:43	16 Jun 2022 00:44
17768	/alice/data/2021/LHC21z/499999/tpc/MW3/tpc-20210329-magnet	ALICE::KISTI_GSDC::CDS	Error		191409	53.25 TB	31 May 2022 20:31	16 Jun 2022 00:42
17767	/alice/data/2021/LHC21z/499999/tpc/MW2/tpc-xray-20210310	ALICE::KISTI_GSDC::CDS	Error		106417	62.06 TB	31 May 2022 20:25	16 Jun 2022 00:41
17766	/alice/data/2021/LHC21z/499999/tpc/MW2/xray01	ALICE::KISTI_GSDC::CDS	Error		212254	55.08 TB	31 May 2022 20:16	16 Jun 2022 00:38
17765	/alice/data/2021/LHC21z/499999/tpc/MW2/xray02	ALICE::KISTI_GSDC::CDS	Error		1810492	139.8 TB	31 May 2022 19:40	16 Jun 2022 00:56
17466	/alice/data/2021/LHC21z/GC/ECS/raw/2021-10-17_19-30/run0504428_2021-10-17T19_31_12Z	ALICE::KISTI_GSDC::CDS	Done		1720	126.3 GB	31 May 2022 12:37	01 Jun 2022 02:36
17465	/alice/data/2021/LHC21z/GC/ECS/raw/2021-10-17_19-09/run0504425_2021-10-17T19_10_37Z	ALICE::KISTI_GSDC::CDS	Error		96101	231.6 GB	31 May 2022 12:33	01 Jun 2022 02:47
17463	/alice/data/2021/LHC21z/GC/ECS/raw/2021-10-14_20-13/run0504250_2021-10-14T20_14_42Z	ALICE::KISTI_GSDC::CDS	Done		15317	34 GB	31 May 2022 12:24	01 Jun 2022 01:09
17462	/alice/data/2021/LHC21z/GC/ECS/raw/2021-10-14_18-53/run0504242_2021-10-14T18_55_51Z	ALICE::KISTI_GSDC::CDS	Error		45104	101.1 GB	31 May 2022 12:21	01 Jun 2022 02:40
17461	/alice/data/2021/LHC21z/GC/ECS/raw/2021-10-14_18-37/run0504234_2021-10-14T18_38_54Z	ALICE::KISTI_GSDC::CDS	Done		4116	12.69 GB	31 May 2022 12:20	01 Jun 2022 00:12
17460	/alice/data/2021/LHC21z/GC/ECS/2021-08-11_12-24/3c1c9417-fc41-48d5-91a4-b102812e06c2/run00_2021-08-11T12_25_38Z	ALICE::KISTI_GSDC::CDS	Done		75527	333.6 GB	31 May 2022 12:17	01 Jun 2022 00:10
17459	/alice/data/2021/LHC21z/GC/ECS/2021-08-11_11-28/44b8c3f9-beca-4e04-877e-ff8a551f4ecf/run00_2021-08-11T11_29_27Z	ALICE::KISTI_GSDC::CDS	Done		35906	158.5 GB	31 May 2022 12:14	01 Jun 2022 00:00
17458	/alice/data/2021/LHC21z/GC/ECS/2021-08-11_10-16/b602eace-a123-4784-99b4-90c965251151/run00_2021-08-11T10_17_27Z	ALICE::KISTI_GSDC::CDS	Done		65977	692.9 GB	31 May 2022 12:10	31 May 2022 23:00
17457	/alice/data/2021/LHC21z/GC/ECS/2021-08-11_09-37/28525e82-0454-4656-ac84-426a51a11326/run00_2021-08-11T09_38_32Z	ALICE::KISTI_GSDC::CDS	Done		29061	840.1 GB	31 May 2022 12:07	31 May 2022 22:52
17456	/alice/data/2021/LHC21z/GC/ECS/2021-09-24_17-08/run0502238_2021-09-24T17_09_46Z	ALICE::KISTI_GSDC::CDS	Done		147382	448.2 GB	31 May 2022 12:02	31 May 2022 22:52
17455	/alice/data/2021/LHC21z/GC/ECS/2021-09-11_11-12/run0500976_2021-09-11T11_13_33Z	ALICE::KISTI_GSDC::CDS	Done		2000	229.3 GB	31 May 2022 11:59	31 May 2022 21:30
17454	/alice/data/2021/LHC21z/GC/ECS/2021-09-11_11-08/run0500975_2021-09-11T11_10_20Z	ALICE::KISTI_GSDC::CDS	Done		2647	303.1 GB	31 May 2022 11:59	31 May 2022 19:59
17453	/alice/data/2021/LHC21z/tpc/Pre-commissioning_Cleanroom/reco/xray	ALICE::KISTI_GSDC::CDS	Error		6739	407.3 GB	31 May 2022 11:58	28 Jun 2022 14:50
17452	/alice/data/2021/LHC21z/tpc/Pre-commissioning_Cleanroom/reco/pulsar	ALICE::KISTI_GSDC::CDS	Done		1398	29.49 GB	31 May 2022 11:58	28 Jun 2022 14:57
17451	/alice/data/2021/LHC21z/tpc/Pre-commissioning_Cleanroom/reco/pedestal	ALICE::KISTI_GSDC::CDS	Done		981	266.8 MB	31 May 2022 11:58	31 May 2022 15:21
17450	/alice/data/2021/LHC21z/tpc/Pre-commissioning_Cleanroom/reco/laser_xray	ALICE::KISTI_GSDC::CDS	Done		10323	624.7 GB	31 May 2022 11:57	31 May 2022 15:24
17449	/alice/data/2021/LHC21z/tpc/Pre-commissioning_Cleanroom/reco/laser	ALICE::KISTI_GSDC::CDS	Done		4982	301.3 GB	31 May 2022 11:57	28 Jun 2022 15:16

[Total Size]=4.728PB

Peak traffic IN + OUT = 4.172GB/s + 3.218GB/s
= 7.39GB/s ≈ 60Gb/s

Average transfer rate = 328MB/s



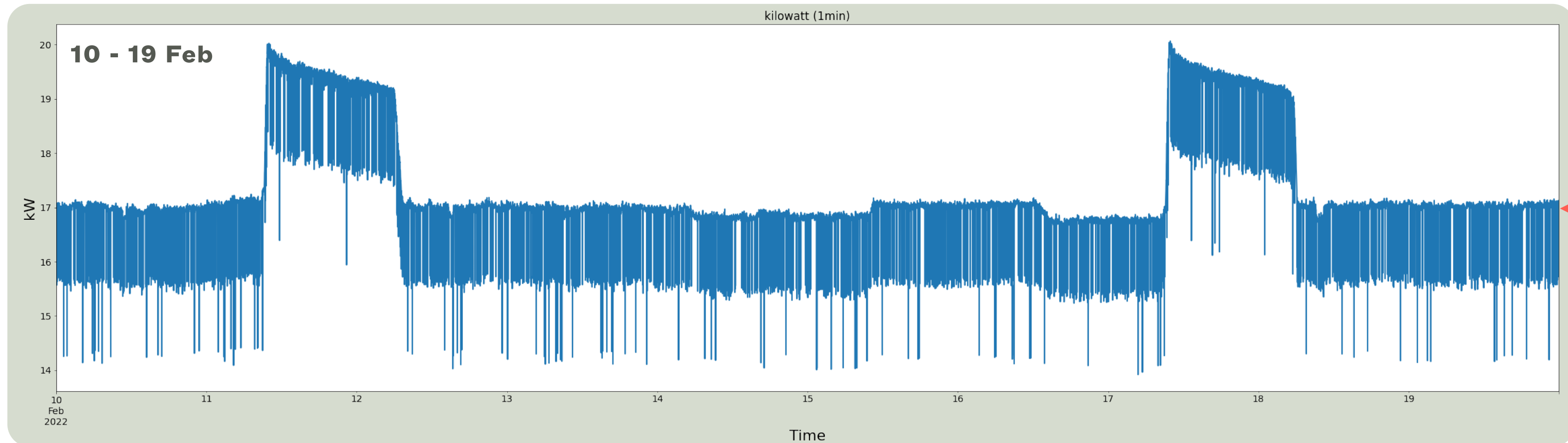
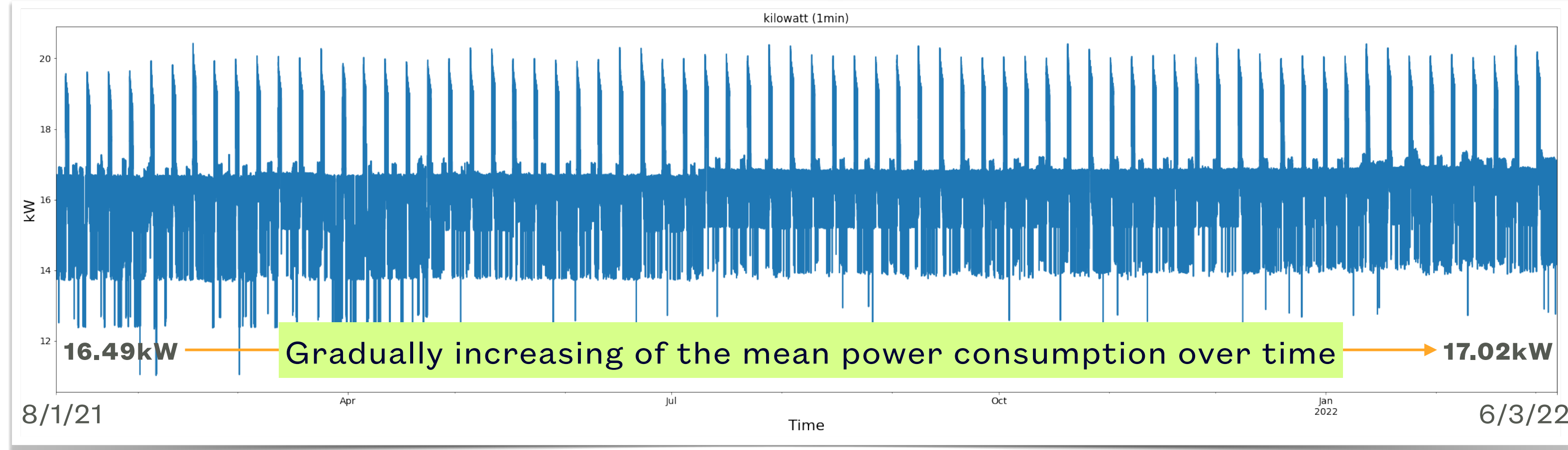
Re-distribution Traffic induced by EC

6 Jan ~ 2 Jul

LHCOPN - KREONet2

Power Consumption

Instantaneous power consumption (kilowatt) per minute (Jan 2021 - Feb 2022)



Comparison with other storage at KISTI
1.125W/TB for full load (cf. 0.5W/TB for Tape)

	Capacity (TB)	Max		Min		Mean	
		kW	W/TB	kW	W/TB	kW	W/TB
CDS	18,144	20.426	1.125	11.015	0.607	16.85	0.923
TS3500	3,200	1.6	0.5	-	-	-	-
SC7020	2,500	12.120	4.8	-	-	-	-
Isilon	2,950	13.730	4.6	-	-	-	-
Isilon	2,360	12.88	9	-	-	-	-
VNX	2,000	5.1	2.2	-	-	-	-
VSP	1,430	18.3	9.15	-	-	-	-
CX4-960	1,500	14.9	9.9	-	-	-	-

Remarkable result for idle state (0.6W/TB)

Periodic full load activities that last 24hours for every 6 days
 (Interval = 518400s) ≠ (EOS scan-interval = 604800s (7 days))

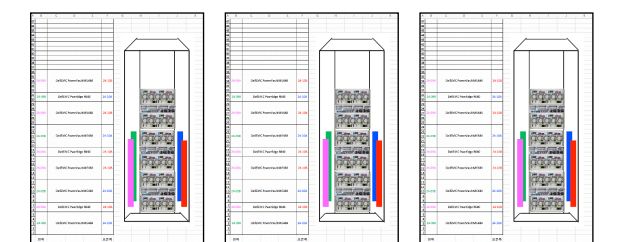
Uncorrelated with data transfers

So far, no clue found for these activities

(Stable electrical characteristics (currents, voltages, etc.)

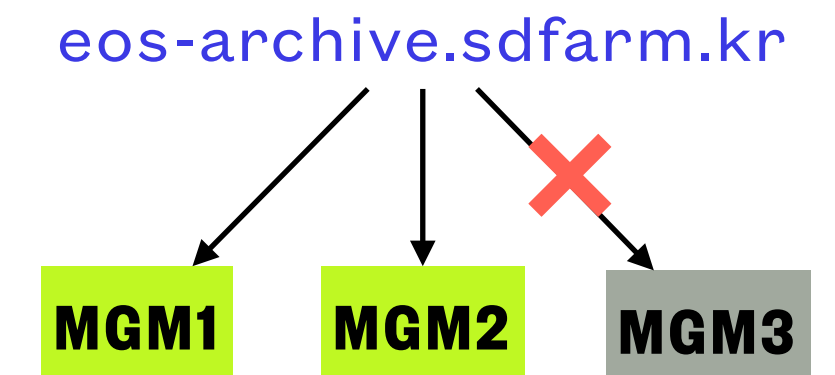
No special features embedded in ME484)

Collected power-related metrics for every minute via SNMP from 12 PDUs in 3 racks



Operations Summary

- Mostly stable - 98% of service availability for the last year
 - Request to a dead MGM went timed-out because DNS name does not dynamically reflect the states of MGMs
 - Dynamic DNS to reflect the states of MGMs would improve service availability
 - New HA scheme of EOS v5 will resolve the issue?
- 29 disks out of 1.5k failed for the last year of operation (2021.11. ~ 2022.11., AFR ~ 1.9%)
 - Replacement is done online without any service discontinuity



Plan

- Upgrading to EOS v5
- Upgrading 80G uplink up to 160G (NIC teaming)
- Developing hardware monitoring system for the enclosures and disks
- Developing filesystem replacement automation (except H/W replacement)
 - Detecting, alerting and excluding problematic filesystem, then adding new filesystem to MGM view
- Expanding CDS to meet the pledges for upcoming years after 2025

Practices for disk replacement online

- Periodic monitoring on EOS filesystems using 'eos status' or 'eos fs ls -e' to identify any faulty filesystems
 - Multi-path related error logs shown in syslog also help predict potential failure of disks or relevant EOS filesystems
- EOS automatically launches draining the filesystem once it detects some errors (but mostly resulted in 'draining error' due to input/output error on the filesystem)
 - Proactively could take action (such as draining) in advance the filesystem actually goes to dead by learning from syslogs
- Any filesystems (disks) to be replaced should be configured as 'empty' then allowed to be removed (note 'fsid')
- FST needs to be restarted once disks or filesystems replacement done, then add the filesystem with 'fsid'
 - Be mindful the minimal number of nodes (FSTs) that sustains an EC layout, e.g. at least 12(16) nodes out of 18 must be active all the time for Read(Write) availability of CDS
 - Note that the disk enclosure or back-end storage should support on-line maintenance feature

KRB5 Authentication

KRB5 CONFIG FOR EOS MGM

```
sh-4.2# cat /etc/xrd.cf.mgm | grep -E 'protocol|protbind' | grep -v '#'
sec.protocol unix
sec.protocol sss -c /etc/eos.keytab -s /etc/eos.keytab
sec.protocol krb5 host/jbod-mgmt-01.sdfarm.kr@SDFARM.KR
sec.protbind localhost.localdomain sss unix
sec.protbind localhost sss unix
sec.protbind * only krb5 sss unix
```

- Testing to get access via FreeIPA KRB5 AuthN/AuthZ
- Enabling KRB5 configuration in EOS MGM
- AuthN/AuthZ successful from internal client
- Mapping GSDC ID & Primary Group information

INTERNAL CLIENT ACCESS VIA KRB5 AUTH

```
~ $ kinit
Password for sahn@SDFARM.KR:
~ $ klist -A
Ticket cache: KEYRING:persistent:556800006:krb_ccache_d440nnV
Default principal: sahn@SDFARM.KR

Valid starting      Expires            Service principal
09/24/2020 11:22:49  09/25/2020 11:22:47  krbtgt/SDFARM.KR@SDFARM.KR
~ $ eos whoami
eos: /cvmfs/alice.cern.ch/x86_64-2.6-gnu-4.1.2/Packages/AlieN/v2-19-395/api/lib/libz.so.1: no version information available (required by e
eos: /cvmfs/alice.cern.ch/x86_64-2.6-gnu-4.1.2/Packages/AlieN/v2-19-395/api/lib/libxml2.so.2: no version information available (required b
eos: /cvmfs/alice.cern.ch/x86_64-2.6-gnu-4.1.2/Packages/AlieN/v2-19-395/api/lib/libxml2.so.2: no version information available (required b
Virtual Identity: uid=556800006 (99,556800006) gid=556800006 (99,556800006) [authz:krb5] host=alice-t1-vobox02.sdfarm.kr domain=sdfarm.kr
```

Access Fusex mount with KRB5

```
[root@fcc ~]# mount -t fuse eosxd /eos
# fsname=
# -o allow_other enabled on shared mount
# -o big_writes enabled
# JSON parsing successful
# File descriptor limit: 524288 soft, 524288 hard
# Disabling nagle algorithm (XRD_NODELAY=1)
# Setting MALLOC_CONF=dirty_decay_ms:0
[root@fcc ~]# df -h
Filesystem                Size      Used Avail Use% Mounted on
/dev/mapper/vg_centos_fcc-lv_root 8.0G  2.2G  5.9G  27% /
devtmpfs                  63G       0   63G   0% /dev
tmpfs                     63G   4.0K   63G   1% /dev/shm
tmpfs                     63G   4.1G   59G   7% /run
tmpfs                     63G       0   63G   0% /sys/fs/cgroup
/dev/sda1                 506M  161M  345M  32% /boot
/dev/mapper/vg_centos_fcc-lv_tmp 1014M    39M  976M   4% /tmp
/dev/mapper/vg_centos_fcc-lv_var  50G   906M   50G   2% /var
/dev/mapper/vg_centos_fcc-lv_home 100G   28G   73G  28% /home
/dev/mapper/vg_centos_fcc-lv_log  5.0G   789M   4.3G  16% /var/log
/dev/mapper/vg_centos_fcc-lv_audit 1014M  430M  585M  43% /var/log/audit
/dev/mapper/vg_centos_fcc-cvmfs_cache 50G    12G   36G  25% /cvmfs_cache
cvmfs2                   42G    12G   30G  28% /cvmfs/cvmfs-config.cern.ch
tmpfs                    13G       0   13G   0% /run/user/556950903
cvmfs2                   42G    12G   30G  28% /cvmfs/sft.cern.ch
cvmfs2                   42G    12G   30G  28% /cvmfs/geant4.cern.ch
cvmfs2                   42G    12G   30G  28% /cvmfs/alice.cern.ch
cvmfs2                   42G    12G   30G  28% /cvmfs/sft-nightlies.cern.ch
pool0.gsn.sdfarm.kr:/ifs/gsd/home/sahn 10T   8.1T   2.0T  81% /share/sahn
tmpfs                    13G       0   13G   0% /run/user/556800006
tmpfs                    13G       0   13G   0% /run/user/556951043
eosxd                    16P  288G   16P   1% /eos
```

- EOS Fusex faster 4x than Fuse
- Tried to validate NAS storage using EOS
 - Kerberos V5 authentication, mounting EOS volume to a host via SSS method
- Validating EOS Fusex for User Home directory
 - Ex: /eos/gsd/user/e/eos_admin1, /eos/gsd/group/eos_admins
 - Secondary GID was not supported

```
> ssh eos_admin1@fcc.sdfarm.kr -p4280
eos_admin1@fcc.sdfarm.kr's password:
Last login: Thu Sep 24 12:55:42 2020
Could not chdir to home directory /eos/gsd/user/e/eos_admin1: Permission denied
-bash: /eos/gsd/user/e/eos_admin1/.bash_profile: Operation not permitted
-bash-4.2$ bash -i
[eos_admin1@fcc /]$ cd
[eos_admin1@fcc ~]$ pwd
/eos/gsd/user/e/eos_admin1
[eos_admin1@fcc ~]$ ls -al
total 2097164
drwx-----. 2 eos_admin1 eos_admins 4096 Sep 23 10:34 .
drwxr-xr-x. 2 root root 4096 Sep 21 16:09 ..
-rw-----. 1 eos_admin1 eos_admins 23 Sep 22 15:59 .bash_history
-rw-----. 1 eos_admin1 eos_admins 18 Sep 22 15:59 .bash_logout
-rw-----. 1 eos_admin1 eos_admins 193 Sep 22 15:59 .bash_profile
-rw-----. 1 eos_admin1 eos_admins 288 Sep 22 16:01 .bashrc
-rw-r-----. 1 eos_admin1 eos_admins 1073741824 Sep 23 10:34 filelg
-rw-r-----. 1 eos_admin1 eos_admins 1073741824 Sep 23 10:34 filelg-localcp
-rw-----. 1 eos_admin1 eos_admins 39 Sep 23 09:02 .lessht
-rwxr-xr-x. 1 eos_admin1 eos_admins 0 Sep 22 13:27 testfile
-rw-----. 1 eos_admin1 eos_admins 658 Sep 22 16:00 .zshrc
[eos_admin1@fcc ~]$
```

Thank you

CDS Deployment Automation

- Fully automated by using Ansible playbook with the following steps:
 - bootstrapping, host ssh-key scanning, host requirement checking, host package upgrading and base service setting (including container runtime engine, e.g. Docker), image building & deploying, and post-processing
- The automation we developed for the CDS includes not only deploying containers but also preparing for host environment so that one can enable the same system as the CDS from the scratch in a non-interactive way
- This approach was efficient and fast to have the CDS start over with different configurations and settings at the pre-production phase however it is completely disruptive and not suitable for service maintenance

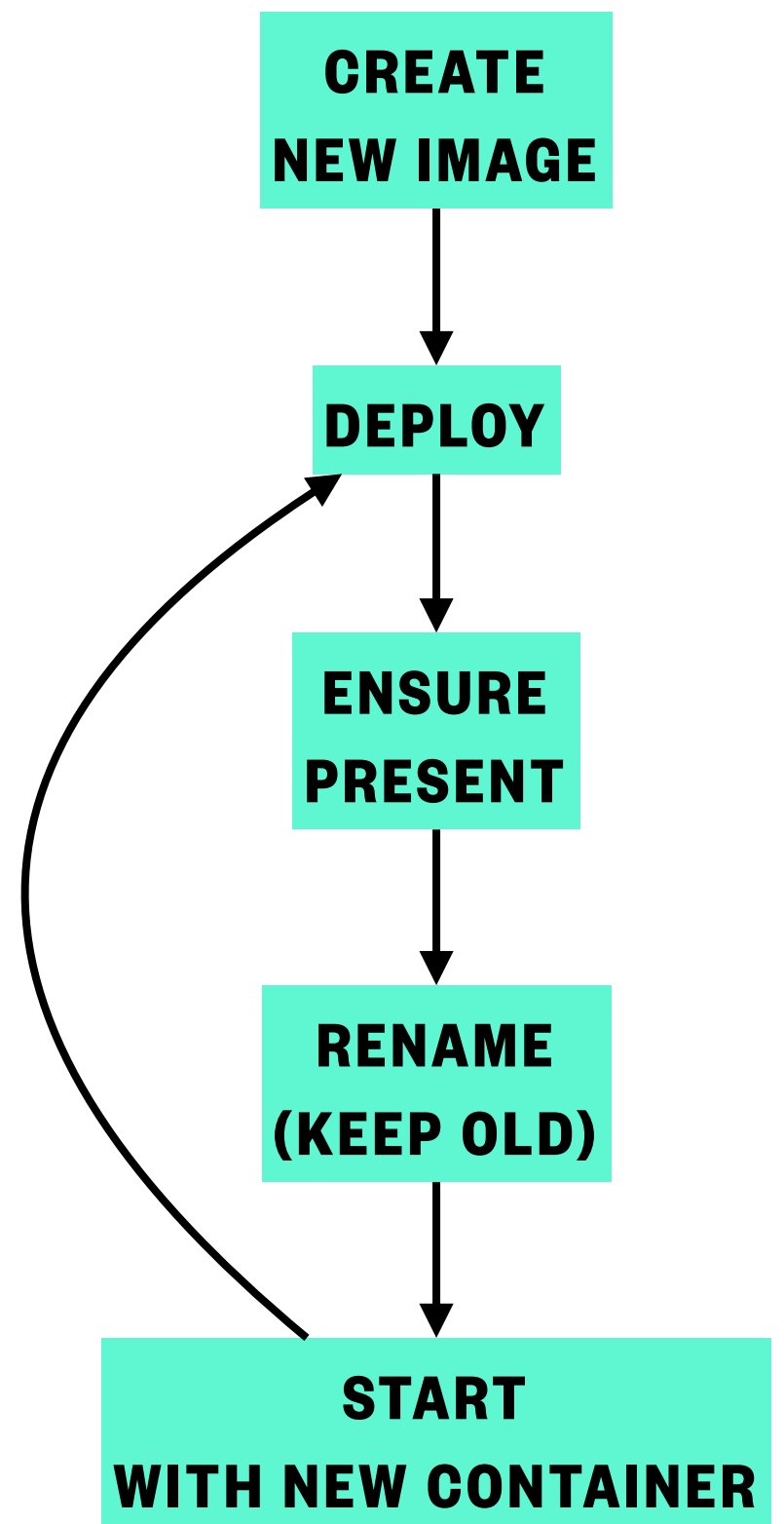
Non-disruptive approach


- A non-disruptive approach was proposed and integrated in the CDS deployment automation
 - Recognizes the already established container environment and running service containers
 - Performs a rolling update so that the containers can be replaced by the new ones incrementally with zero downtime
- Redundant architecture of the CDS allows this approach available
 - Quark-DB cluster (three containers in raft mode) for namespace
 - Masters (MGM) in high-availability mode (one primary + two secondaries)
 - Erasure-coded storage layout (12 data nodes + 4 parity nodes + 2 spare nodes)



CDS Deployment automation (modified by accommodating a non-disruptive approach)

ROLLING UPDATE PROCESS

- Ensure that containers are created, deployed but aren't started so that the existing containers can be kept running =>
- Choose role and node to conduct container update



 config.yaml
 container image tag => { EOS releases + Arbitrary number }
 container state => { present }

 site.yaml  Inventory
 Container role => { mgm + mq | qdb | fst }
 Choose nodes to conduct container update

*Need to check operation status



Practices for Upgrade of the CDS

- Modifications on automation process accommodating the proposed approach as follows
 - Creating a new image including package updates and relevant configuration changes
 - Adding knobs to check the existing container runtime environment and running service containers
 - Defining conditionals to decide container deployment status such as started and present
 - Preserving replaced old containers by renaming them for emergency roll-back
 - Performing the rolling update once all of the conditions are met