EOS 2023 Workshop @ CERN
24-27 April 2023

Operation Status of CDS for the ALICE Experiment

Ahn Sang-Un, Han Heejune, Kim Jeong-heon, Lee Seung Hee, Yoon Heejun

Outline

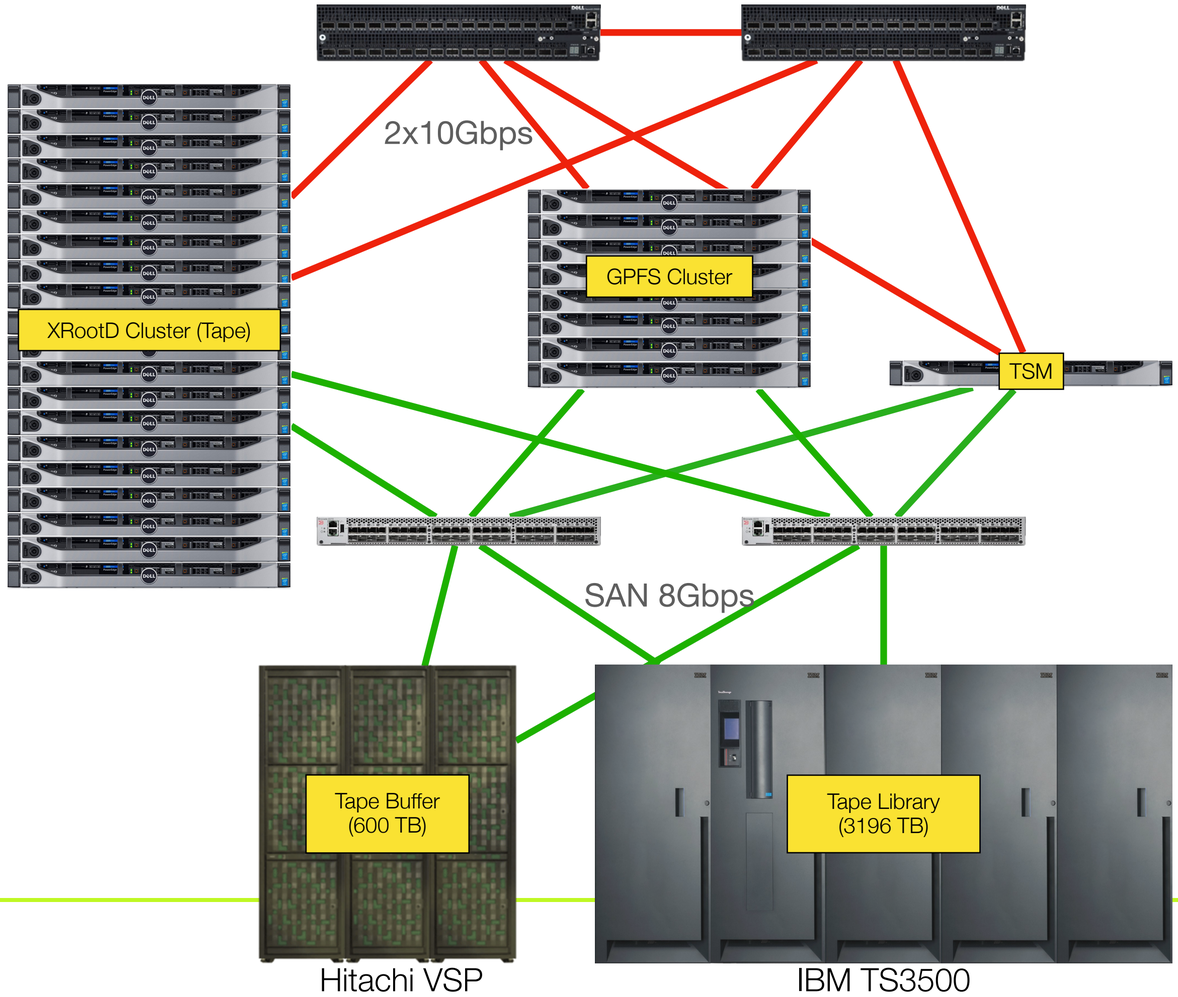
- Introduction
- CDS Architecture
- QRAIN Layout & Configuration
- Current Status
- Operations: WLCG Tape Challenge, Production for ALICE, Power Consumption
- Plan

Introduction

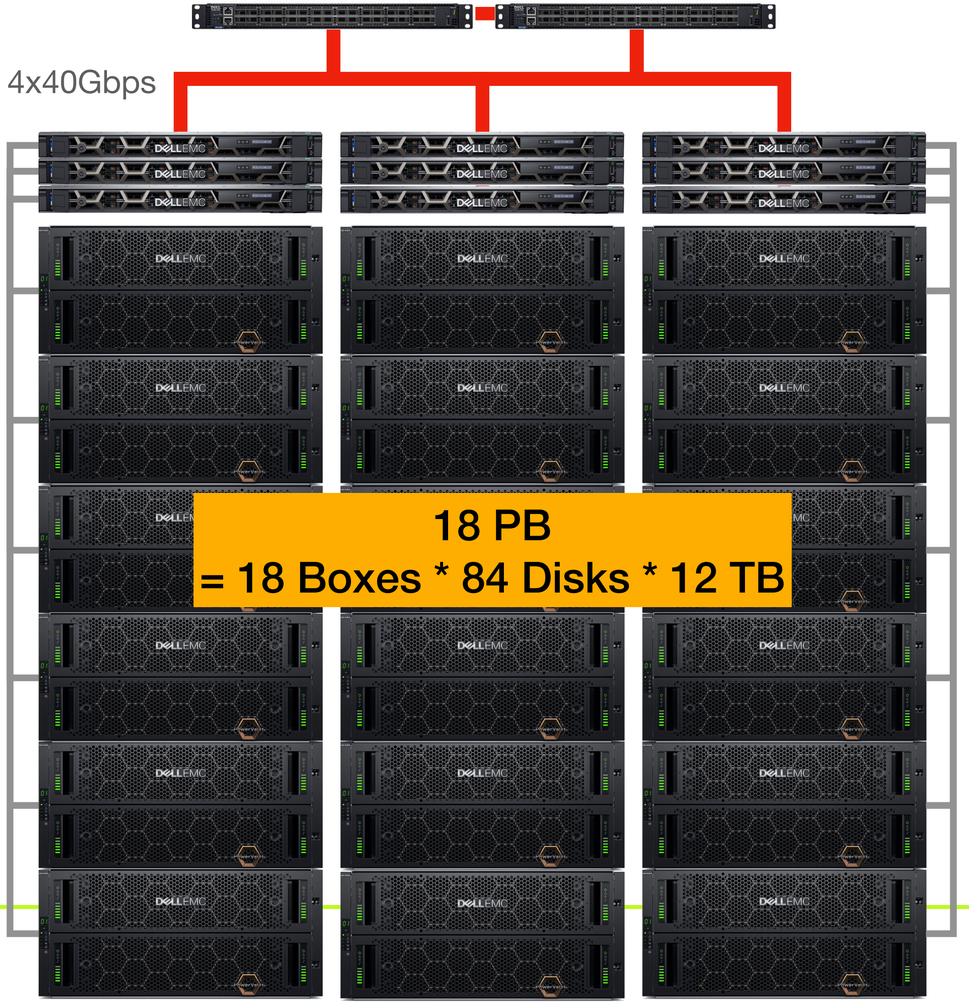
- CDS - a disk based storage designed to store and preserve RAW data from the ALICE experiment by accommodating EOS with its erasure code implementation, a.k.a RAIN configuration
- Simplified architecture : removing additional disk buffers (~ 0.6PB) in front of tape library for I/O and commercial HSM software
- Better data accessibility: any data available at any time
- Avoiding vendor lock-in due to monopoly in tape market
- Provided to the ALICE experiment for commissioning at the early of 2021 and in production since November 2021, replacing the tape storage completely

Architecture (Tape vs. CDS)

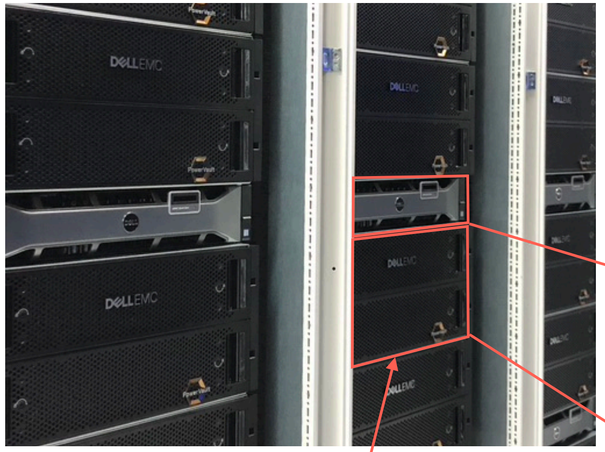
Tape (2012~2021)

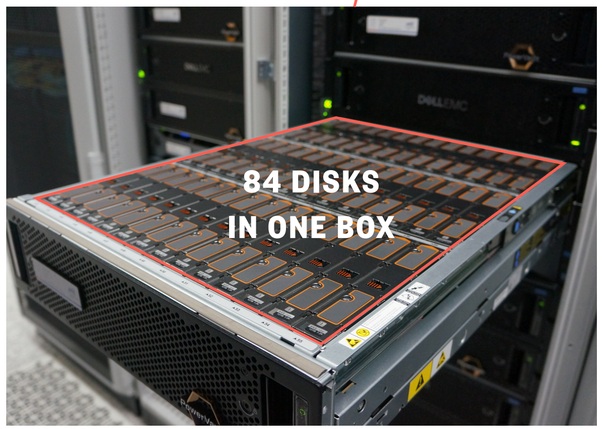


CDS (2021~)



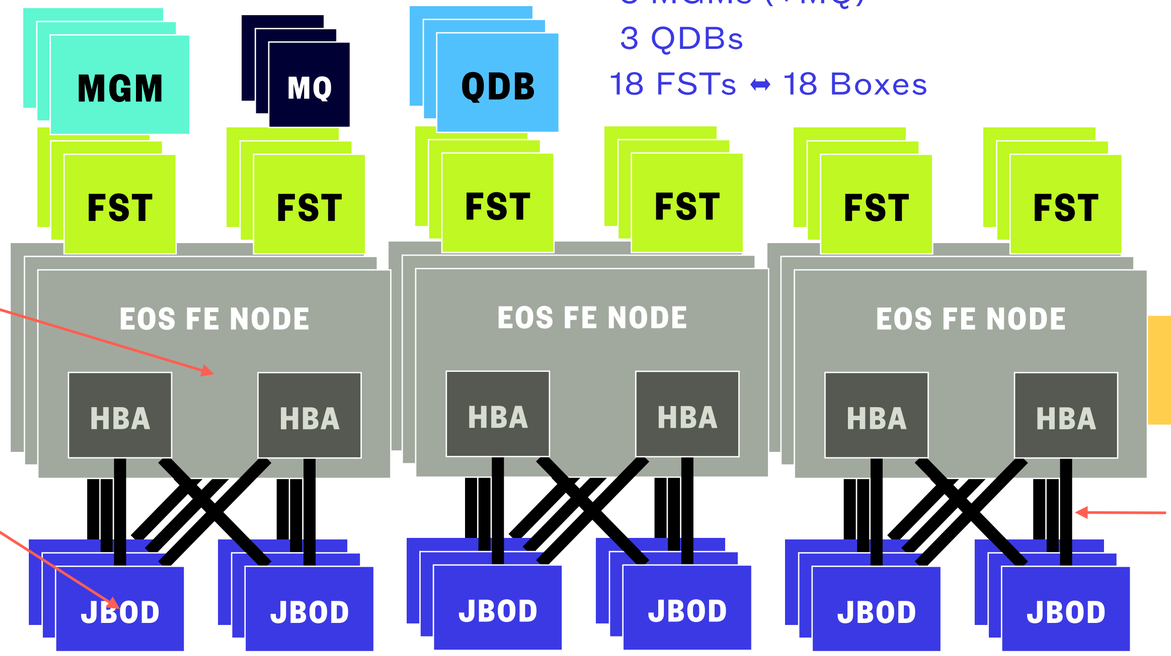
CDS Architecture

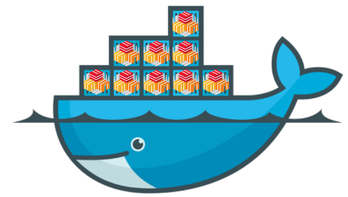




9 servers
18 boxes

3 MGMs (+MQ)
3 QDBs
18 FSTs ⬌ 18 Boxes





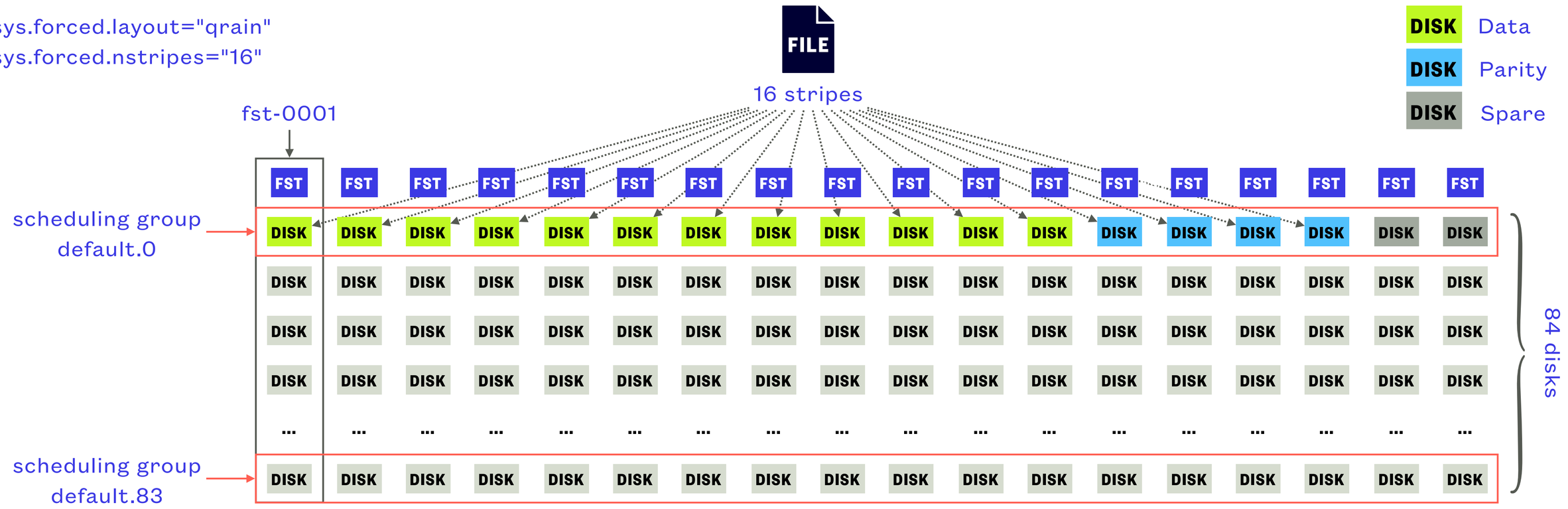
80Gbps Uplink

12Gbps

- 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/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 $\approx 8.6 \times 10^{-5}\%$, if 5 disks fail simultaneously in 5 independent FSTs, 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/gsdc/testarea/raid6
eos attr -r set default=archive /eos/gsdc/testarea/archive
eos attr -r set default=replica /eos/gsdc/testarea/replica
eos attr -r set default=qrain /eos/gsdc/testarea/rain12
```

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

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

```
sh-4.2# eos attr ls /eos/gsdc/testarea/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/gsdc/proc/recycle/"
```

Fileinfo in QRAIN Layout

EOS fileinfo command

Layout type
# of stripes
# of replica

File chuck location
Scheduling group
Filesystem status

```
EOS Console [root://localhost] |/eos/gsdc/grid/15/65278/> fileinfo fedccd04-4f19-11e8-9717-0f23a10abeef
  File: '/eos/gsdc/grid/15/65278/fedccd04-4f19-11e8-9717-0f23a10abeef'  Flags: 0600
  Size: 1864614942
Modify: Sat Apr 16 23:48:08 2022 Timestamp: 1650152888.625957000
Change: Sat Apr 16 23:26:45 2022 Timestamp: 1650151605.137469164
 Birth: Sat Apr 16 23:26:45 2022 Timestamp: 1650151605.137469164
  CUid: 10367 CGid: 1395 Fxid: 00378511 Fid: 3638545 Pid: 13439 Pxid: 0000347f
XStype: adler    XS: 06 b0 69 a9    ETAGs: "976714486251520:06b069a9"
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	871	jbod-mgmt-06.sdfarm.kr	default.30	/jbod/box_11_disk_030	booted	rw	nodrain	online	kisti::gsdc::g02
1	1459	jbod-mgmt-09.sdfarm.kr	default.30	/jbod/box_18_disk_030	booted	rw	nodrain	online	kisti::gsdc::g03
2	283	jbod-mgmt-02.sdfarm.kr	default.30	/jbod/box_04_disk_030	booted	rw	nodrain	online	kisti::gsdc::g01
3	1375	jbod-mgmt-09.sdfarm.kr	default.30	/jbod/box_17_disk_030	booted	rw	nodrain	online	kisti::gsdc::g03
4	1123	jbod-mgmt-07.sdfarm.kr	default.30	/jbod/box_14_disk_030	booted	rw	nodrain	online	kisti::gsdc::g03
5	31	jbod-mgmt-01.sdfarm.kr	default.30	/jbod/box_01_disk_030	booted	rw	nodrain	online	kisti::gsdc::g01
6	535	jbod-mgmt-04.sdfarm.kr	default.30	/jbod/box_07_disk_030	booted	rw	nodrain	online	kisti::gsdc::g02
7	115	jbod-mgmt-01.sdfarm.kr	default.30	/jbod/box_02_disk_030	booted	rw	nodrain	online	kisti::gsdc::g01
8	1207	jbod-mgmt-08.sdfarm.kr	default.30	/jbod/box_15_disk_030	booted	rw	nodrain	online	kisti::gsdc::g03
9	703	jbod-mgmt-05.sdfarm.kr	default.30	/jbod/box_09_disk_030	booted	rw	nodrain	online	kisti::gsdc::g02
10	451	jbod-mgmt-03.sdfarm.kr	default.30	/jbod/box_06_disk_030	booted	rw	nodrain	online	kisti::gsdc::g01
11	955	jbod-mgmt-06.sdfarm.kr	default.30	/jbod/box_12_disk_030	booted	rw	nodrain	online	kisti::gsdc::g02
12	1039	jbod-mgmt-07.sdfarm.kr	default.30	/jbod/box_13_disk_030	booted	rw	nodrain	online	kisti::gsdc::g03
13	367	jbod-mgmt-03.sdfarm.kr	default.30	/jbod/box_05_disk_030	booted	rw	nodrain	online	kisti::gsdc::g01
14	619	jbod-mgmt-04.sdfarm.kr	default.30	/jbod/box_08_disk_030	booted	rw	nodrain	online	kisti::gsdc::g02
15	199	jbod-mgmt-02.sdfarm.kr	default.30	/jbod/box_03_disk_030	booted	rw	nodrain	online	kisti::gsdc::g01

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

Current Status

- EOS version installed: 4.8.82
  - Automated deployment via Ansible playbook
- Public DNS name pointing to 3 MGMs (load-balancing)
- 4x40G NICs teamed to provide 160G 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
```

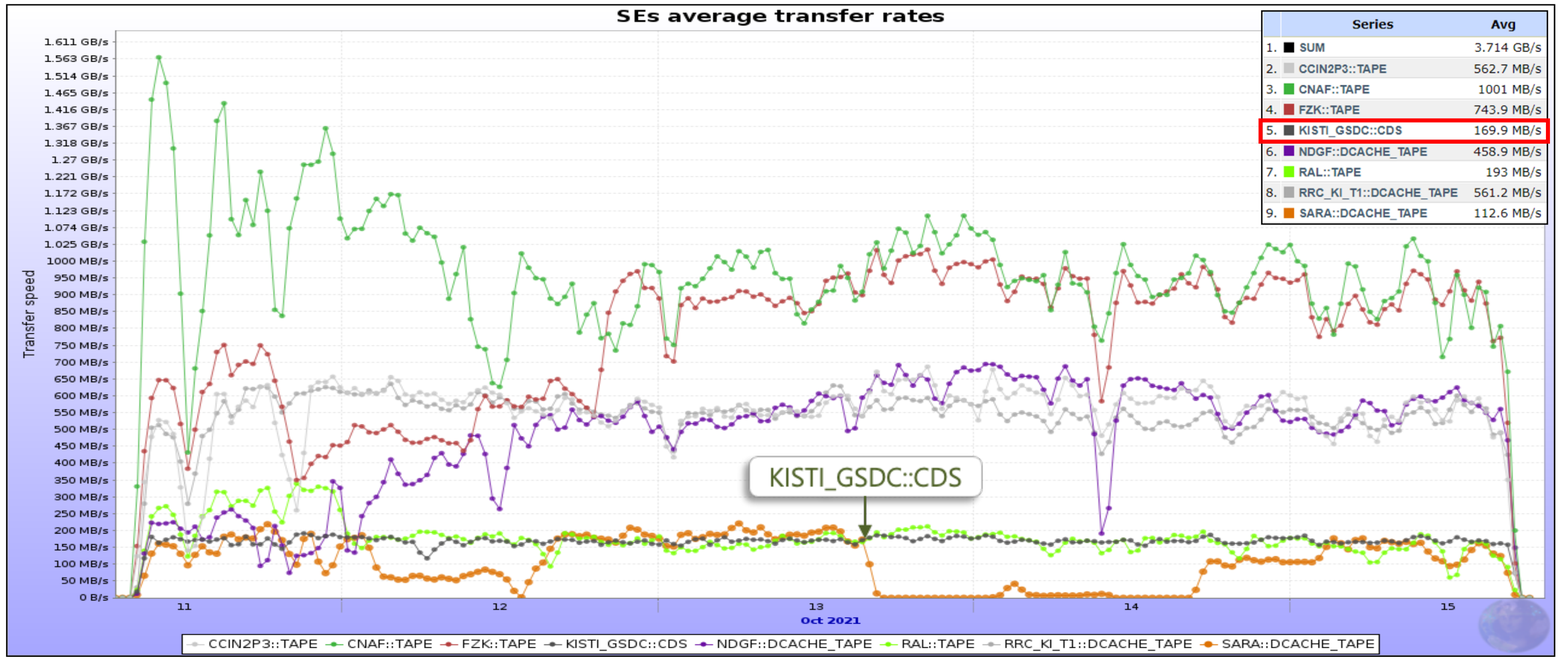
type	hostport	geotag	status	activated	txgw	gw-queued	gw-ntx	gw-rate	heartbeatdelta	nofs
nodesview	jbod-mgmt-01.sdfarm.kr:1095	kisti::gsdc::g01	online	on	off	0	10	120	1	84
nodesview	jbod-mgmt-01.sdfarm.kr:1096	kisti::gsdc::g01	online	on	off	0	10	120	1	84
nodesview	jbod-mgmt-02.sdfarm.kr:1095	kisti::gsdc::g01	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/gsdc/grid
sys.eos.btime="1612374338.811408574"
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="qrain"
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

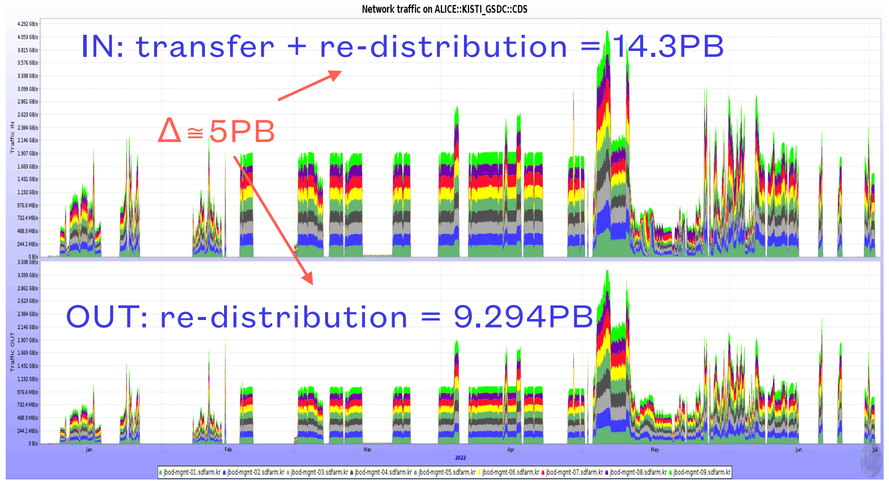
Transfer requests (add new request)

alice/data | CDS | - any - | Filter

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-b102612e06c2/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-fffa55154ecf/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_09-37/28525e82-045d-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/pulser	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

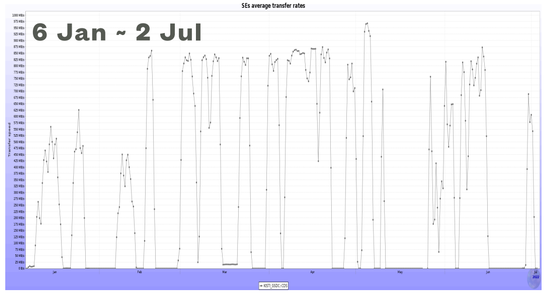
Peak traffic IN + OUT = 4.172GB/s + 3.218GB/s
= 7.39GB/s ≅ 60Gb/s

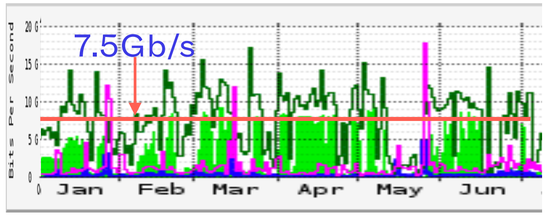


Re-distribution Traffic induced by EC

6 Jan ~ 2 Jul

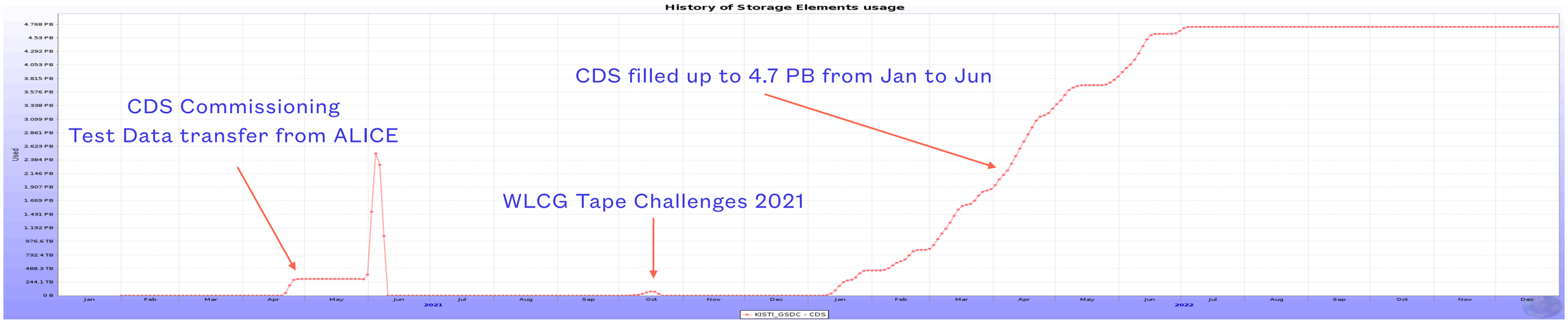
Average transfer rate = 328MB/s

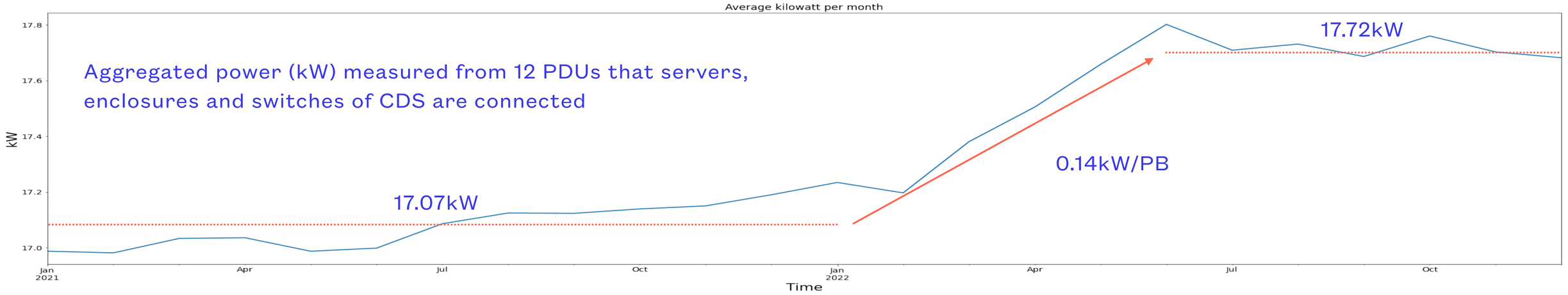




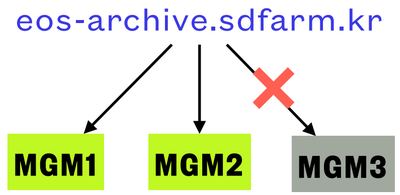
LHCOPN - KREONet2

Power Consumption





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 would improve service availability
    - New HA scheme of EOS v5 will resolve this issue?
- 27 disks out of 1.5k failed for the last year of operation (2022.01. ~ 2022.12., AFR ~ 1.78%)

cf. Vendor published AFR on average ~ 0.35%

  - Replacement is done online without any service discontinuity thanks to spare nodes
  - Note that EOS storage with RAIN will go read-only (not writable) when the number of available nodes is below the total number of RAIN nodes, e.g. we have 16 (12+4) stripes configuration while the storage will go read-only in anyway once any one of nodes is under maintenance (16 → 15, i.e. 11+4 or 12+3)

Plan

- Upgrading to EOS v5 (roll-out of new containers with updated image)
- Network tuning to fully exploit 160 Gbps bandwidth (for OPN 100G foreseen at the end of RUN3)
- Developing hardware monitoring system for the enclosures and disks
  - Developing filesystem (disk) replacement automation
  - Detecting, alerting and excluding problematic filesystem (disk), then adding new filesystem to MGM view (still human intervention is required for emptying EOS filesystem and physical disk replacement)
- Expanding CDS to meet the pledges for upcoming years after 2025
  - Moving to denser disk boxes (84→106) || disks with greater capacity (> 20TB) || etc.

Thank you