Sang-un Ahn
sahn@kisti.re.kr

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

EOS Workshop, 7 - 11 March 2022

Operation status of Custodial Disk Storage for the ALICE experiment

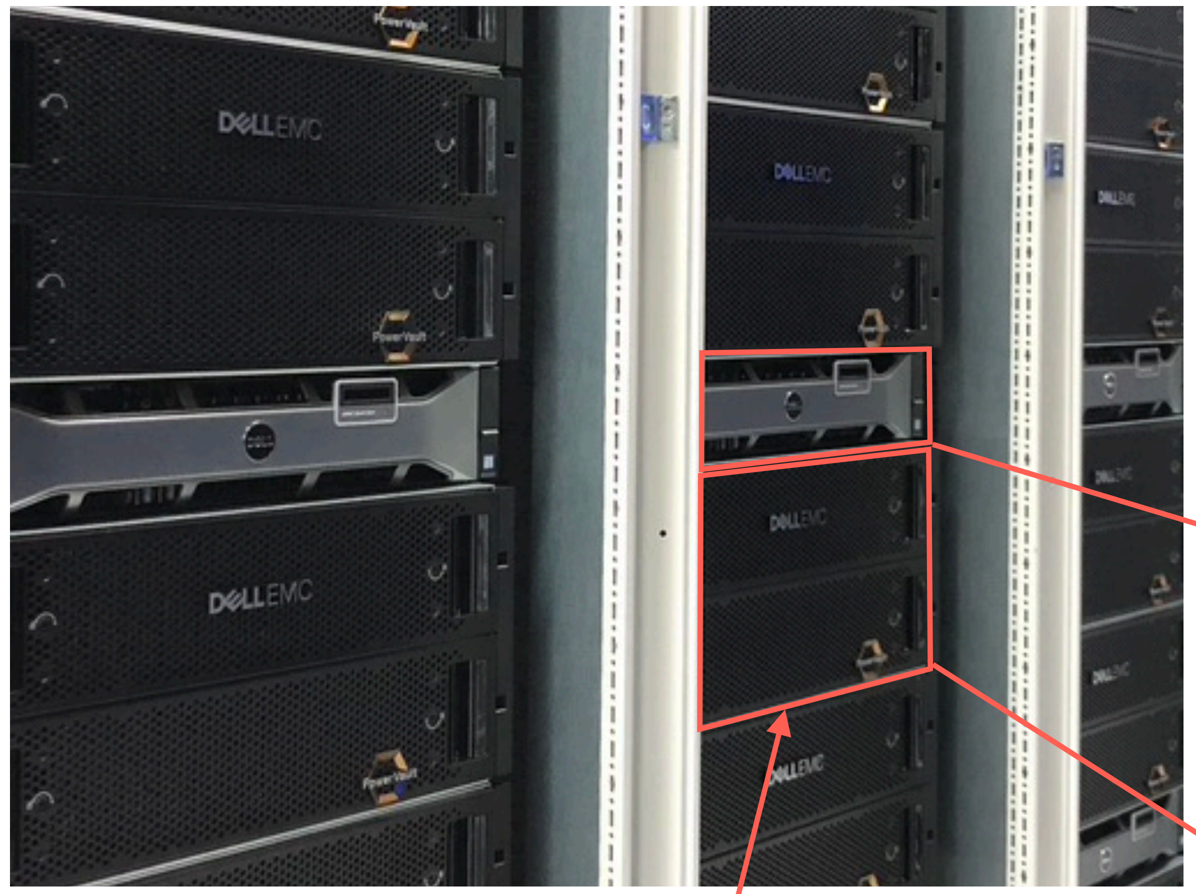
Outline

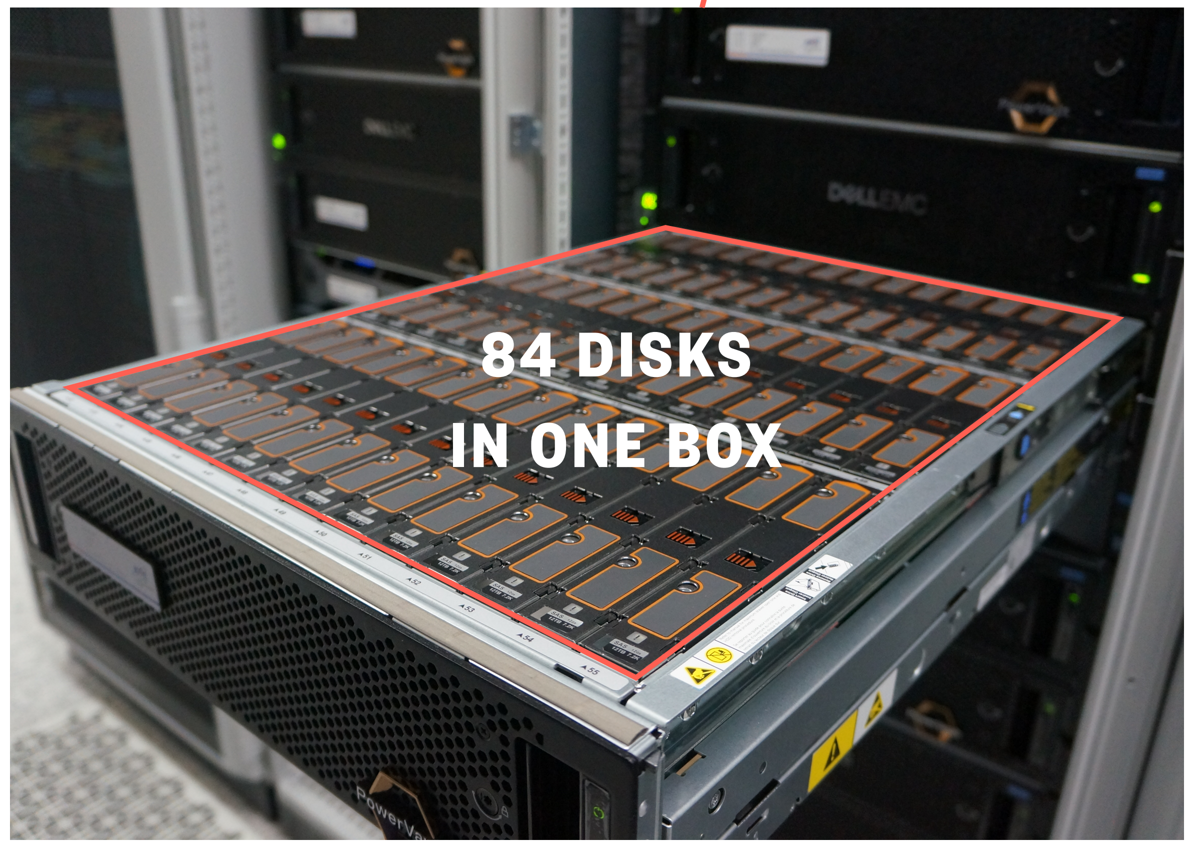
- Introduction
- System Architecture
- QRAIN Layout
- Current Status
- Operations: Incidents, WLCG Tape Challenge, Production service for the ALICE
- Power Consumption
- Plan

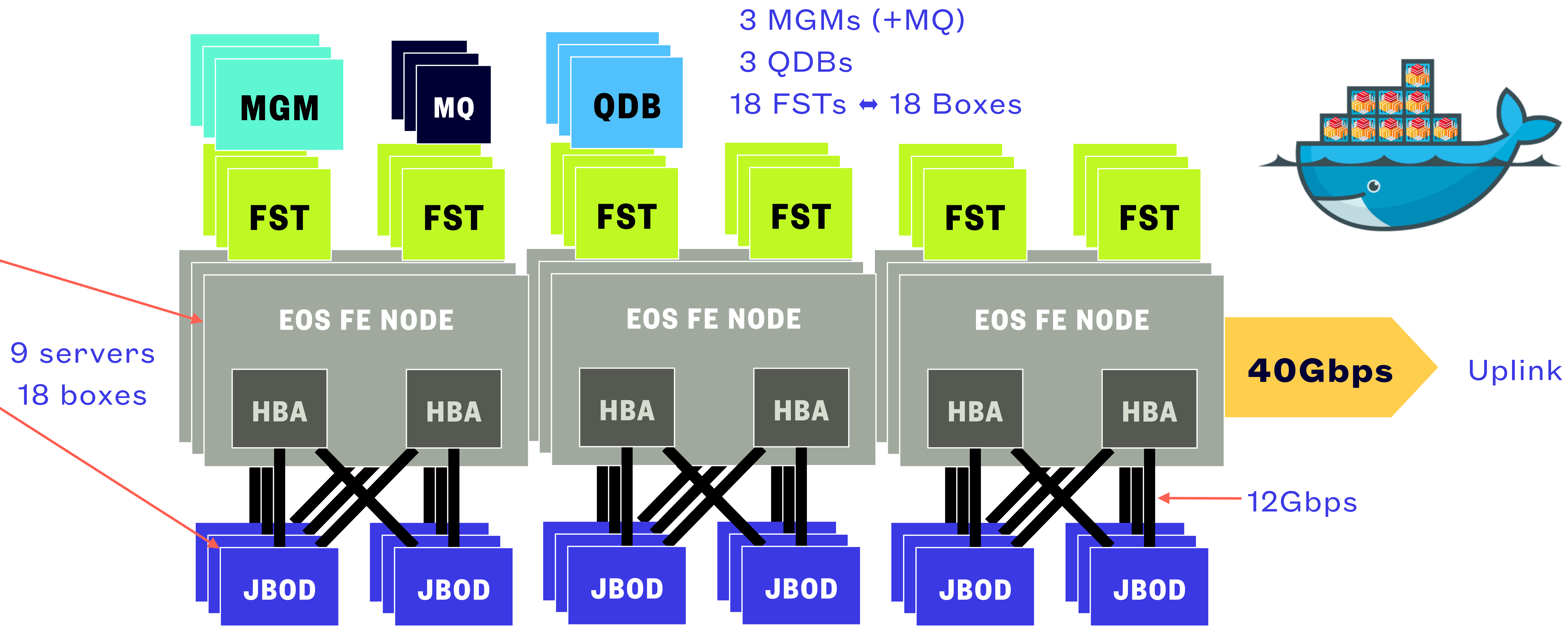
Introduction

- 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
  - 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 by replacing completely the tape storage

System Architecture



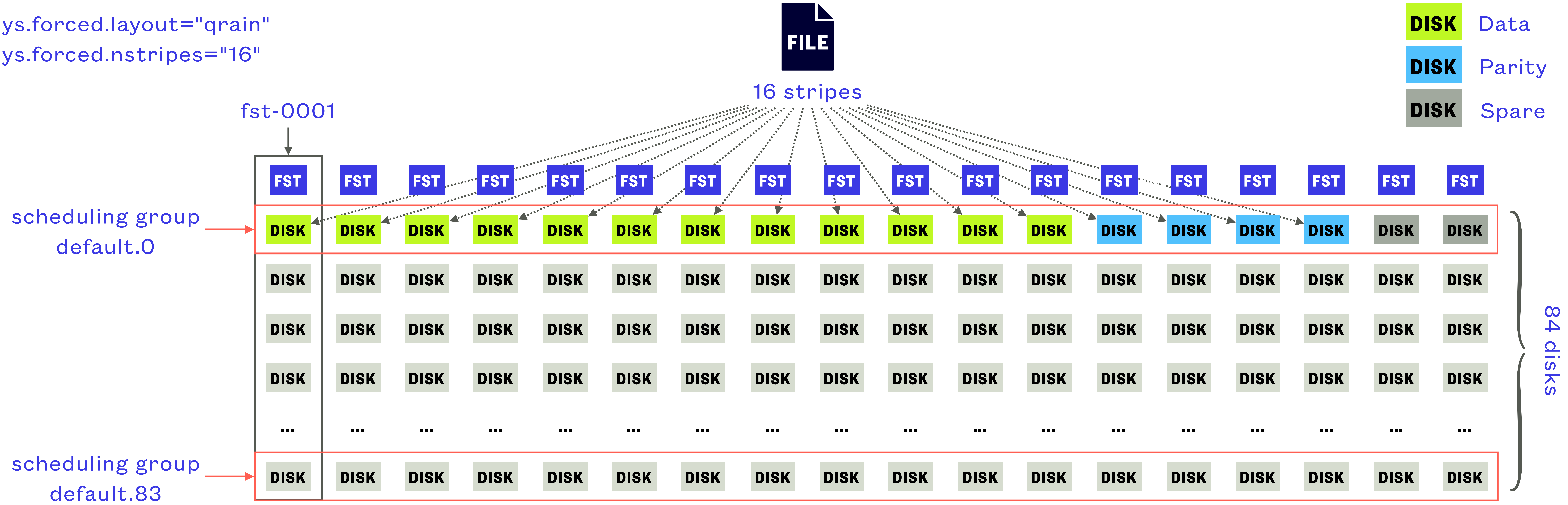




- Total raw capacity = 18,144TB (= 12TB * 84 disks * 18 boxes)
- EOS version = 4.8.31 (released on 2020.12.07)
- EOS components are running on containers (a fork of EOS-Docker project)
  - Ansible playbook available at https://github.com/jeongheon81/gsdc-eos-docker

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 ≈ 8.6×10^{-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)

Current Status

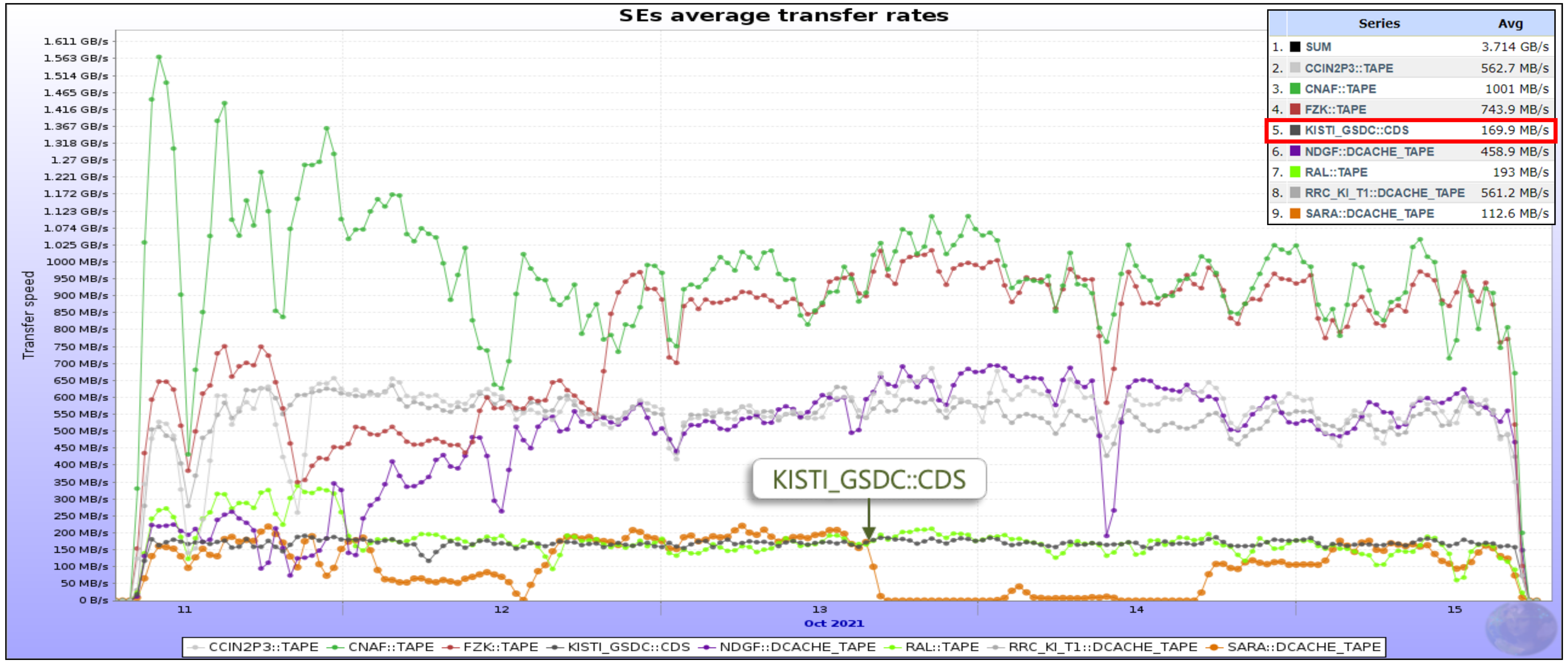
- EOS version installed: 4.8.31
  - Automated deployment via Ansible playbook
- Public DNS name pointing to 3 MGMs
- IPv4/IPv6 dual stack configured
- ALICE Integration
  - Enabling Token-based AuthN/AuthZ
  - Enabling ApMon daemons on all EOS FSTs for ALICE MonALISA monitoring
  - Allowing Third-Party Copy by disabling sss enforcement on FSTs

Operations: Incidents

- Mostly stable
  - An incident induced by the failure on the automatic failover among three MGMs
    - For most cases, the automatic failover to the (randomly chosen) secondaries provoked by the unresponsiveness of the current master works well
  - 2 disks out of 1.5k (0.13%) failed per month on average
    - Replacement is done online without any service discontinuity

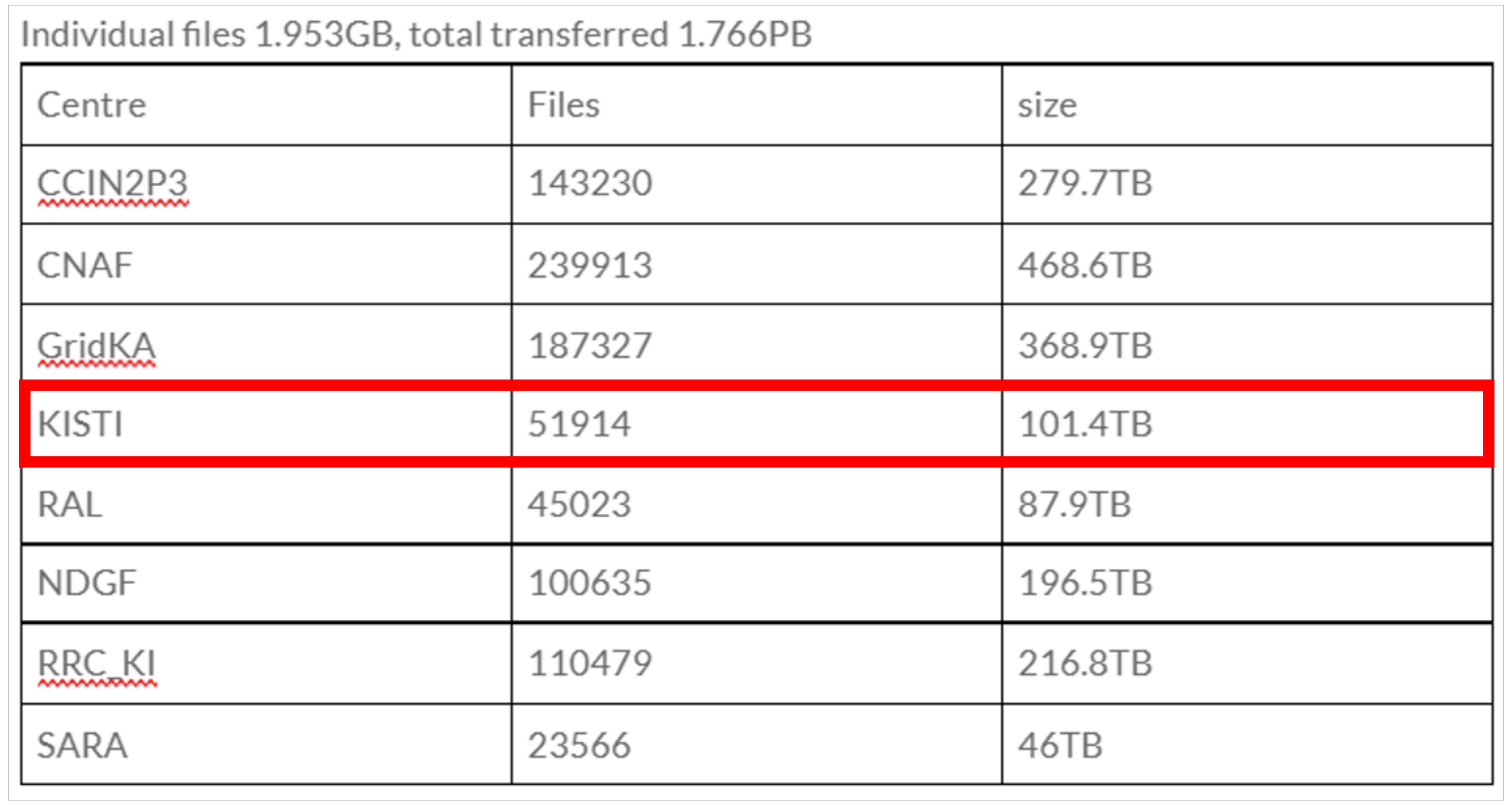
WLCG Tape Challenge (Oct 2021)

- 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)



	Series	Avg
1.	■ SUM	3.714 GB/s
2.	■ CCIN2P3::TAPE	562.7 MB/s
3.	■ CNAF::TAPE	1001 MB/s
4.	■ FZK::TAPE	743.9 MB/s
5.	■ KISTI_GSDC::CDS	169.9 MB/s
6.	■ NDGF::DCACHE_TAPE	458.9 MB/s
7.	■ RAL::TAPE	193 MB/s
8.	■ RRC_KI_T1::DCACHE_TAPE	561.2 MB/s
9.	■ SARA::DCACHE_TAPE	112.6 MB/s

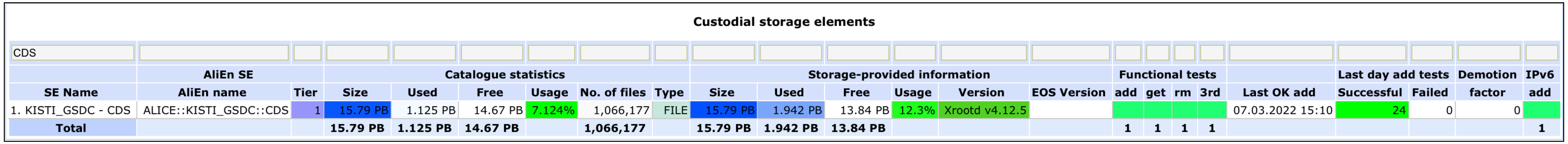
170MB/s on average for 5-day of transfer
101.4TB of data (51k files) transferred


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



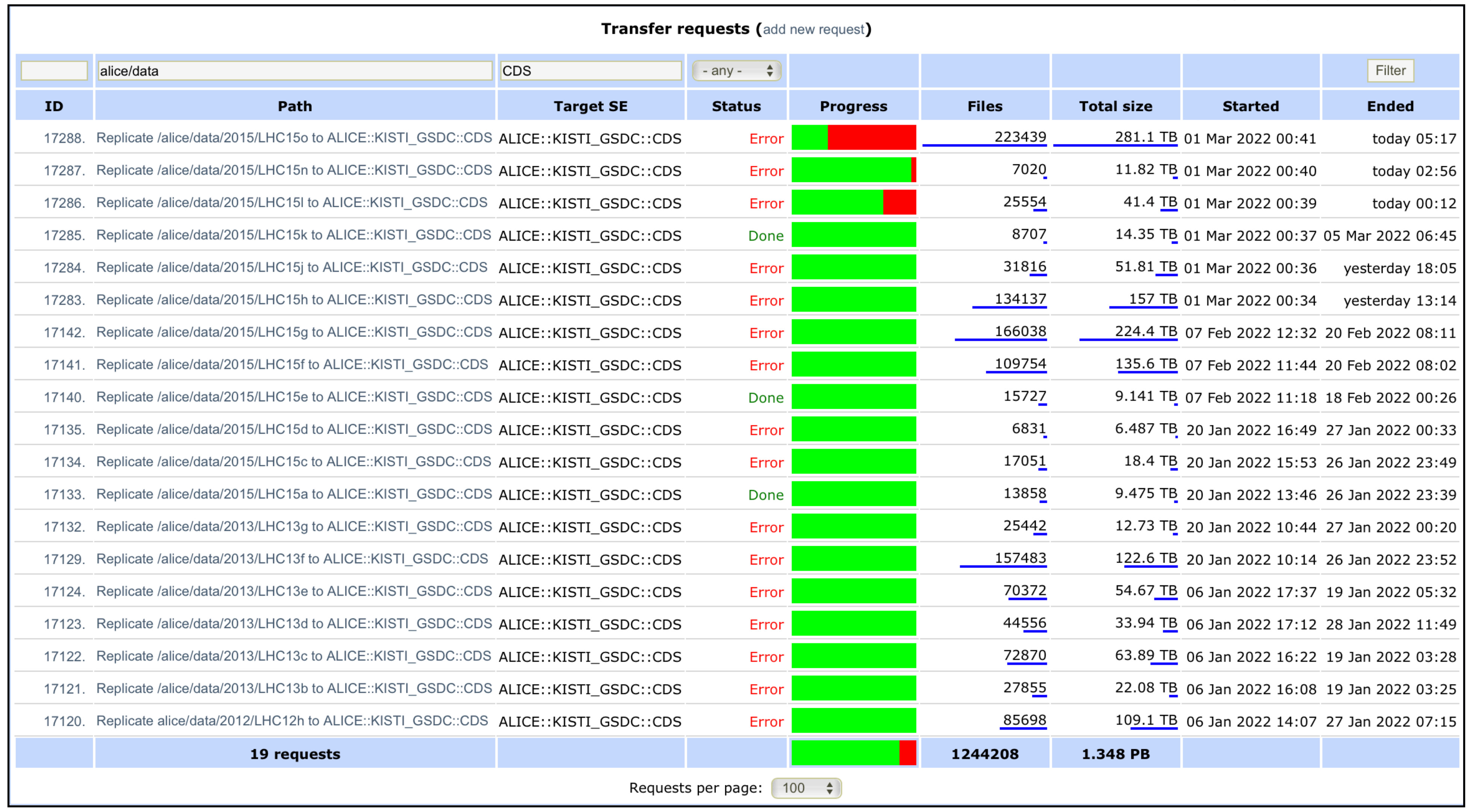
Custodial storage elements

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	1.125 PB	14.67 PB	7.124%	1,066,177	FILE	15.79 PB	1.942 PB	13.84 PB	12.3%	Xrootd v4.12.5						07.03.2022 15:10	24	0	0	
Total			15.79 PB	1.125 PB	14.67 PB		1,066,177		15.79 PB	1.942 PB	13.84 PB				1	1	1	1					1

	Total	Used
Bin	15.79	1.942
Dec	17.77	2.19

ALICE RAW data being replicated to the CDS

6 Jan ~ 7 Mar



Transfer requests (add new request)

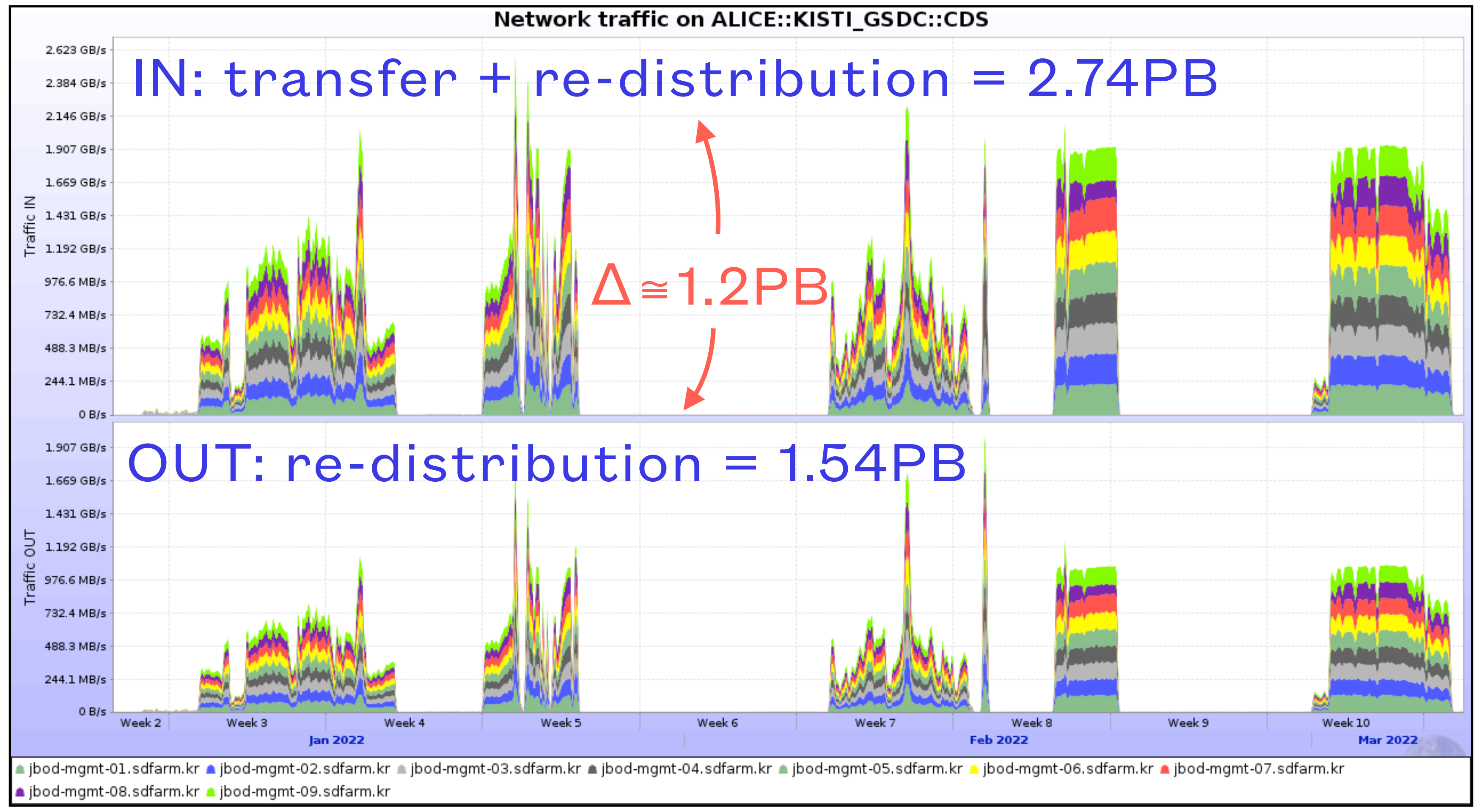
	alice/data	CDS	- any -					Filter
ID	Path	Target SE	Status	Progress	Files	Total size	Started	Ended
17288.	Replicate /alice/data/2015/LHC15o to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		223439	281.1 TB	01 Mar 2022 00:41	today 05:17
17287.	Replicate /alice/data/2015/LHC15n to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		7020	11.82 TB	01 Mar 2022 00:40	today 02:56
17286.	Replicate /alice/data/2015/LHC15l to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		25554	41.4 TB	01 Mar 2022 00:39	today 00:12
17285.	Replicate /alice/data/2015/LHC15k to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Done		8707	14.35 TB	01 Mar 2022 00:37	05 Mar 2022 06:45
17284.	Replicate /alice/data/2015/LHC15j to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		31816	51.81 TB	01 Mar 2022 00:36	yesterday 18:05
17283.	Replicate /alice/data/2015/LHC15h to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		134137	157 TB	01 Mar 2022 00:34	yesterday 13:14
17142.	Replicate /alice/data/2015/LHC15g to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		166038	224.4 TB	07 Feb 2022 12:32	20 Feb 2022 08:11
17141.	Replicate /alice/data/2015/LHC15f to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		109754	135.6 TB	07 Feb 2022 11:44	20 Feb 2022 08:02
17140.	Replicate /alice/data/2015/LHC15e to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Done		15727	9.141 TB	07 Feb 2022 11:18	18 Feb 2022 00:26
17135.	Replicate /alice/data/2015/LHC15d to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		6831	6.487 TB	20 Jan 2022 16:49	27 Jan 2022 00:33
17134.	Replicate /alice/data/2015/LHC15c to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		17051	18.4 TB	20 Jan 2022 15:53	26 Jan 2022 23:49
17133.	Replicate /alice/data/2015/LHC15a to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Done		13858	9.475 TB	20 Jan 2022 13:46	26 Jan 2022 23:39
17132.	Replicate /alice/data/2013/LHC13g to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		25442	12.73 TB	20 Jan 2022 10:44	27 Jan 2022 00:20
17129.	Replicate /alice/data/2013/LHC13f to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		157483	122.6 TB	20 Jan 2022 10:14	26 Jan 2022 23:52
17124.	Replicate /alice/data/2013/LHC13e to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		70372	54.67 TB	06 Jan 2022 17:37	19 Jan 2022 05:32
17123.	Replicate /alice/data/2013/LHC13d to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		44556	33.94 TB	06 Jan 2022 17:12	28 Jan 2022 11:49
17122.	Replicate /alice/data/2013/LHC13c to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		72870	63.89 TB	06 Jan 2022 16:22	19 Jan 2022 03:28
17121.	Replicate /alice/data/2013/LHC13b to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		27855	22.08 TB	06 Jan 2022 16:08	19 Jan 2022 03:25
17120.	Replicate alice/data/2012/LHC12h to ALICE::KISTI_GSDC::CDS	ALICE::KISTI_GSDC::CDS	Error		85698	109.1 TB	06 Jan 2022 14:07	27 Jan 2022 07:15
	19 requests				1244208	1.348 PB		

Requests per page: 100

[Done w/o Error]=1.15PB

Peak aggregated traffic IN + OUT

= 2.5GB/s + 1.9GB/s ≦ 40Gbps (Uplink bandwidth)



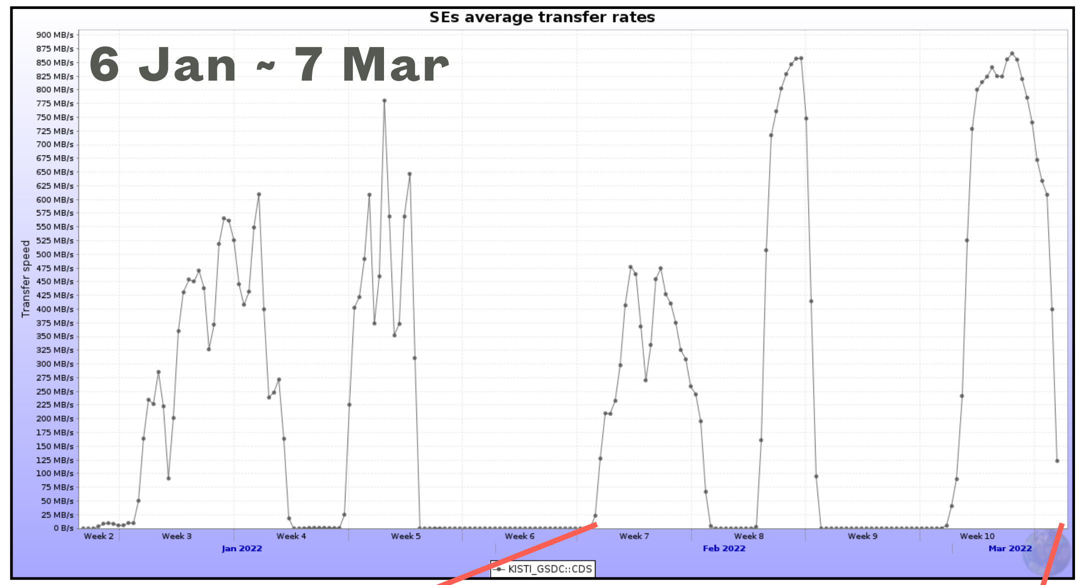
IN: transfer + re-distribution = 2.74PB

Δ≅1.2PB

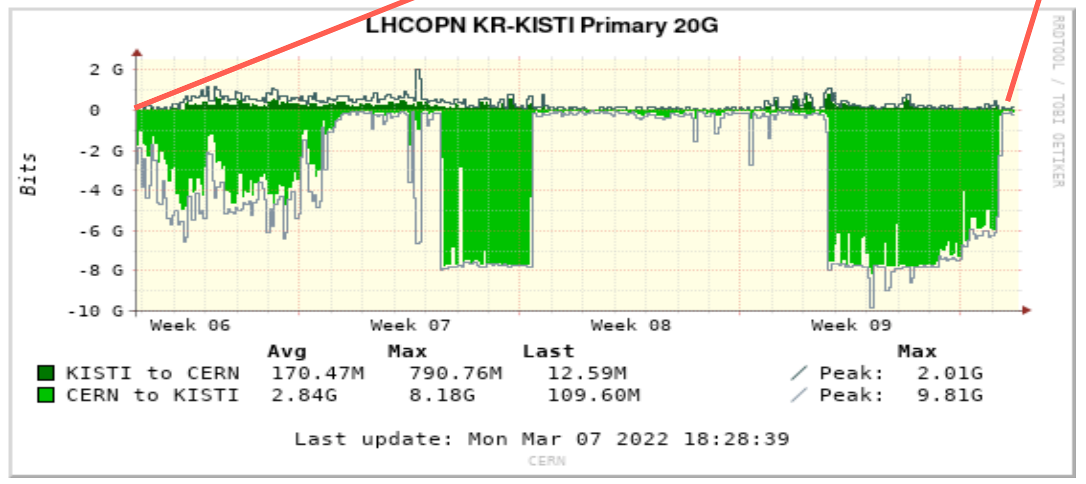
OUT: re-distribution = 1.54PB

6 Jan ~ 7 Mar

Average transfer rate = 236MB/s



6 Jan ~ 7 Mar



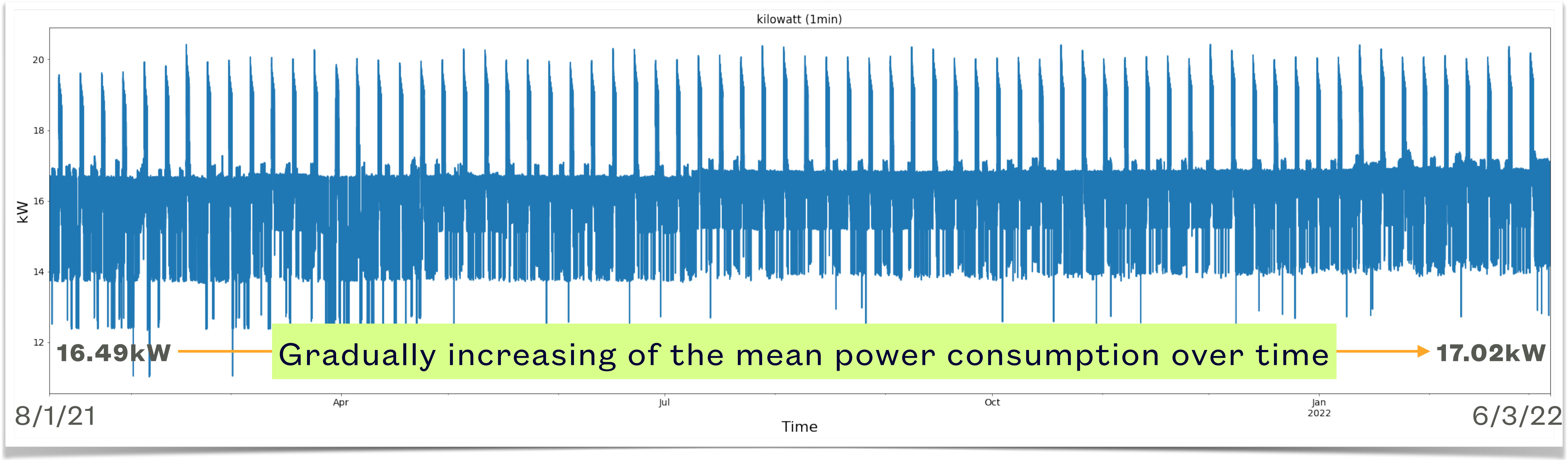
```
              Avg       Max        Last          Max
KISTI to CERN 170.47M   790.76M    12.59M        Peak:  2.01G
CERN to KISTI 2.84G     8.18G      109.60M       Peak:  9.81G

     Last update: Mon Mar 07 2022 18:28:39
```

LHCOPN - Monthly View

Power Consumption

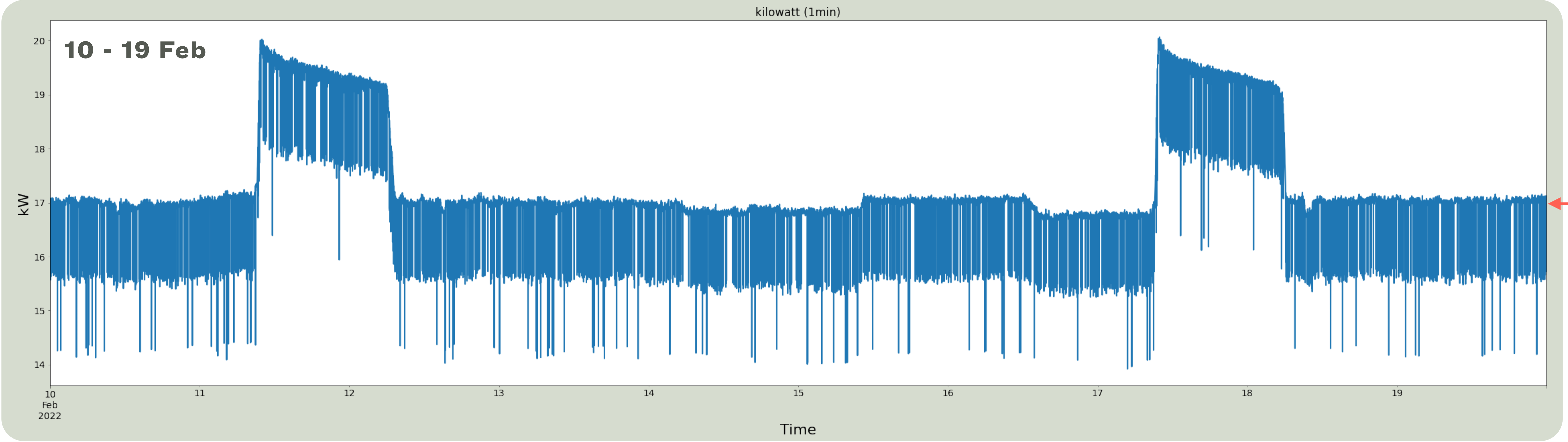
Instantaneous power consumption (kilowatt) per minute (Jan 2021 - Feb 2022)



16.49kW → Gradually increasing of the mean power consumption over time → 17.02kW

8/1/21

6/3/22



10 - 19 Feb

Comparison with other storage at KISTI

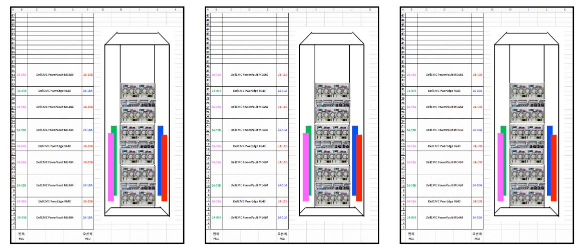
1.125W/TB for full load (cf. 0.5W/TB for Tape)

	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
Any other EOS config parameters related?
OS or H/W-level activities under investigation

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



Plan

- Updating EOS to the latest stable releases
- Developing hardware monitoring system for the enclosures and disks
- Upgrading 40G uplink up to 80G (NIC bonding)

Thank you