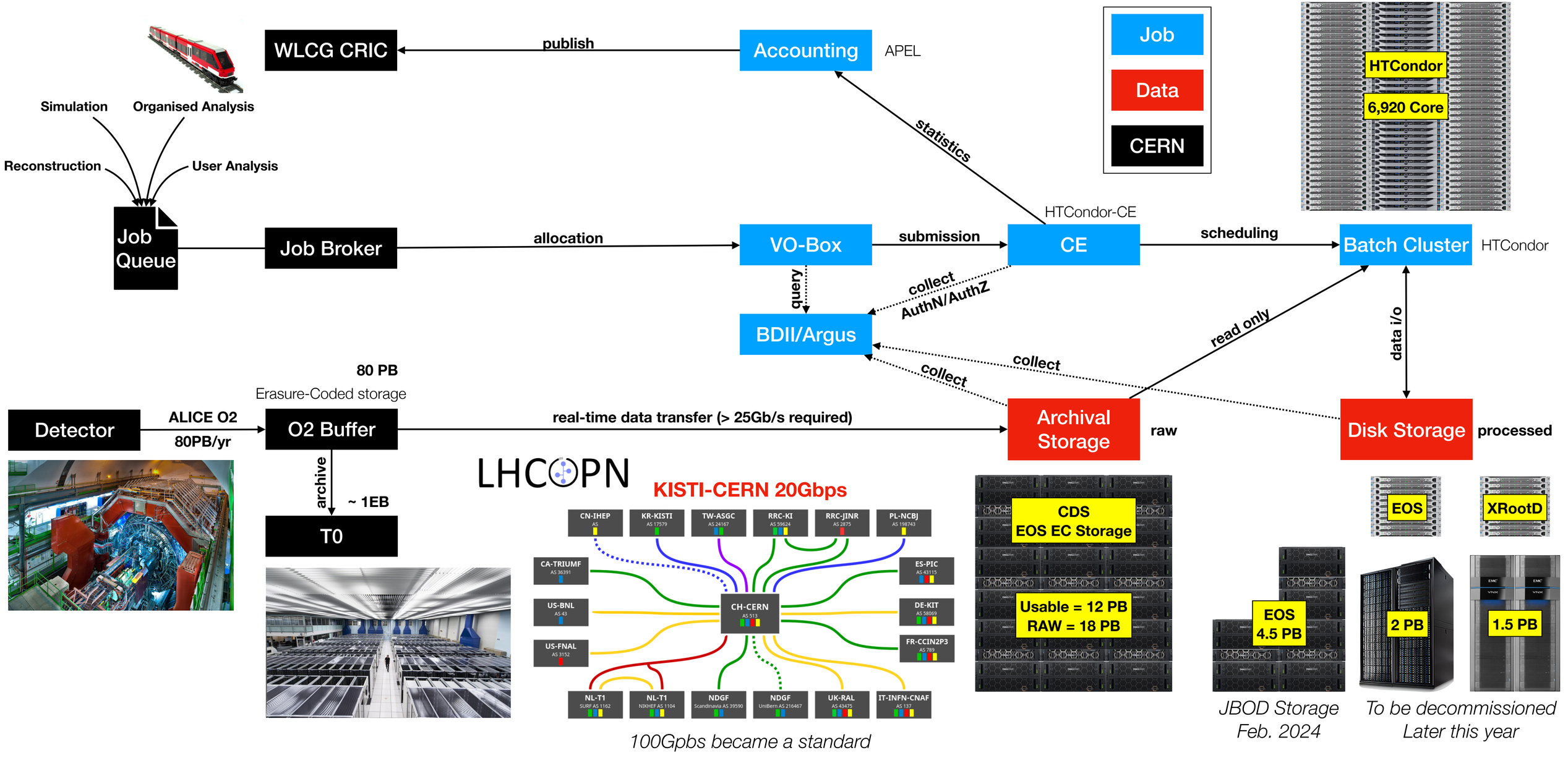
KISTI-GSDC Report

Sang-Un Ahn
On behalf of KISTI-GSDC

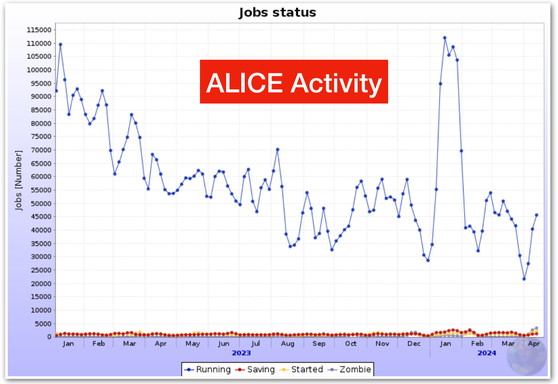
16-18 April 2024 @ ALICE Tier-1/Tier-2 Workshop in Seoul

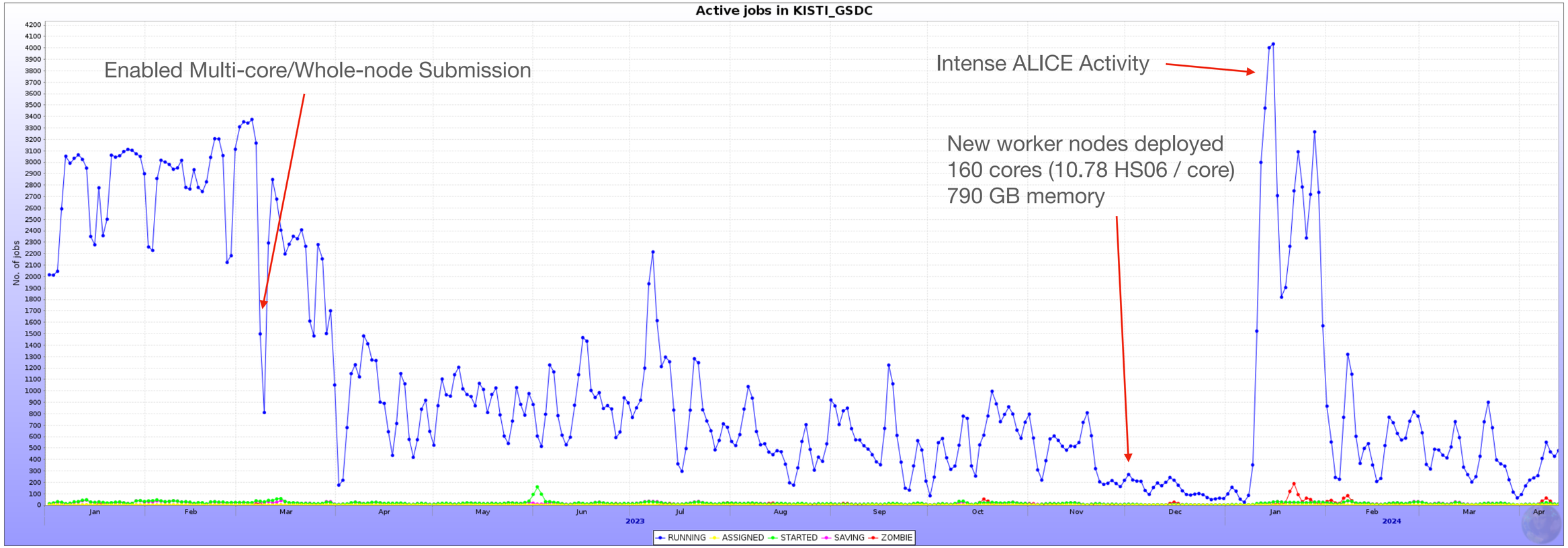
Operations

KISTI ALICE T1 Structure Overview



Active Jobs

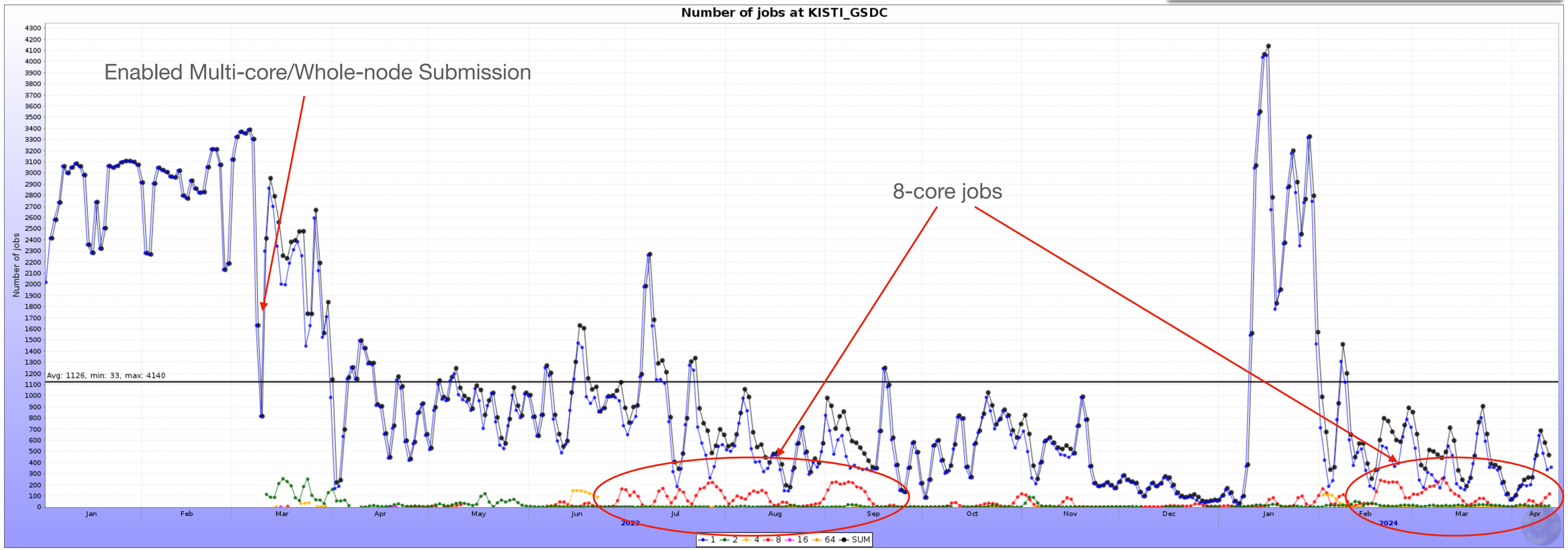

Jobs status
ALICE Activity
Jobs [Number]
Running
Saving
Started
Zombie


Active jobs in KISTI_GSDC
Enabled Multi-core/Whole-node Submission
Intense ALICE Activity
New worker nodes deployed
160 cores (10.78 HS06 / core)
790 GB memory
No. of jobs
RUNNING
ASSIGNED
STARTED
SAVING
ZOMBIE

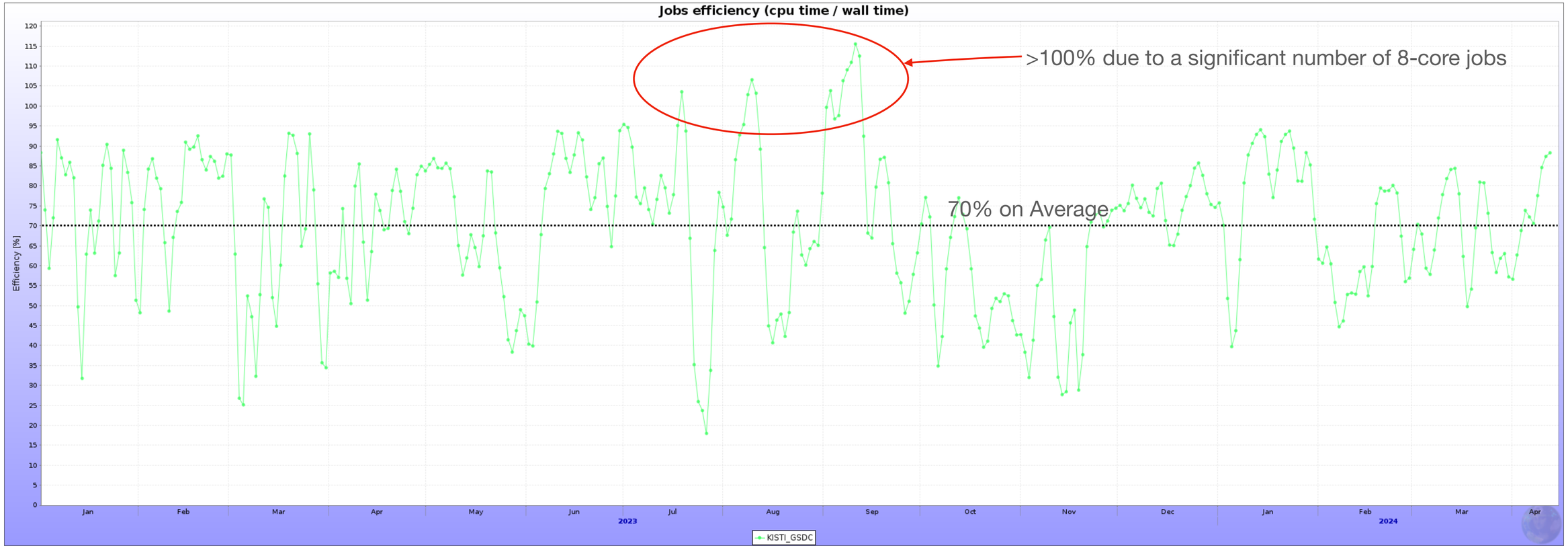
Multicore Jobs

Number of jobs at KISTI_GSDC

	Series	Last value	Min	Avg	Max
1.	1	358	0	1069	5801
2.	2	12	0	18.39	874
3.	4	1	0	15.16	300
4.	8	119	0	54.35	549
5.	16	1	0	0.619	11
6.	64	2	0	4.2	20

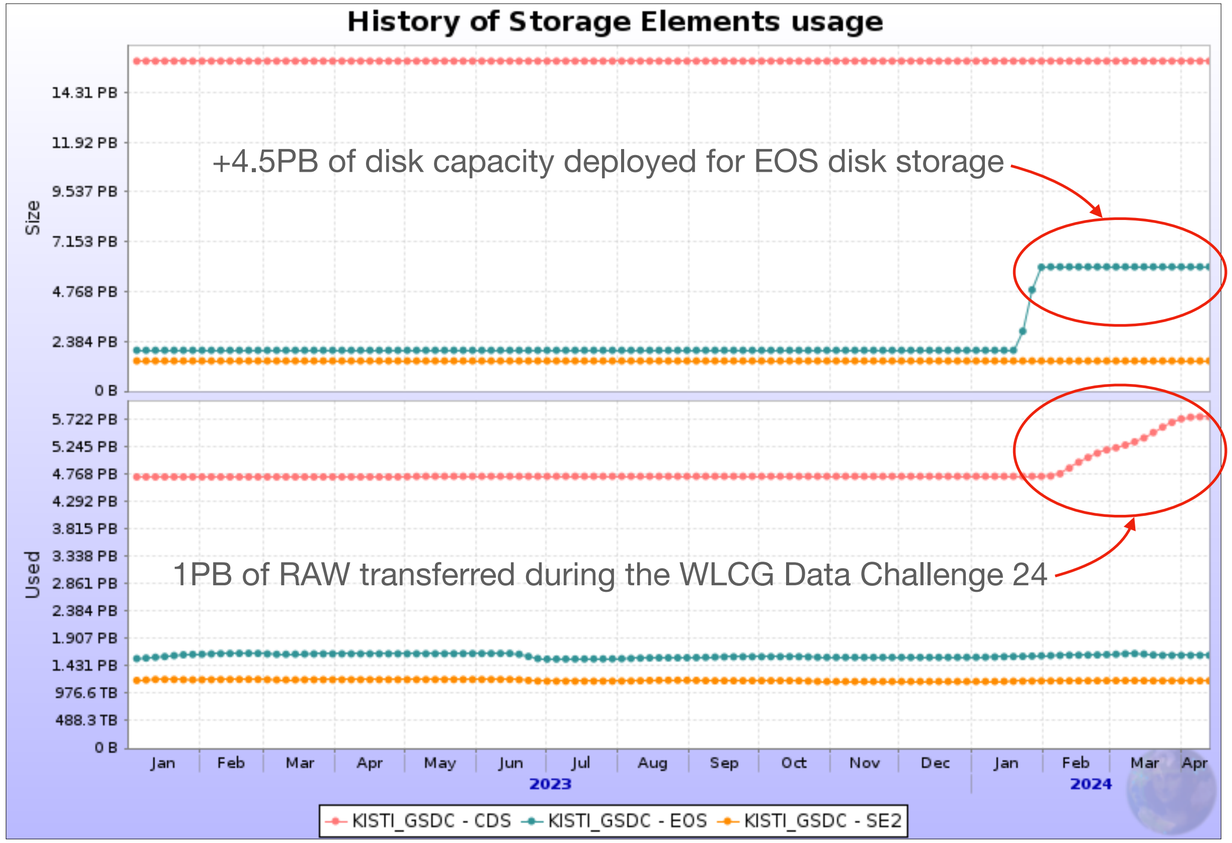


Jobs Efficiency

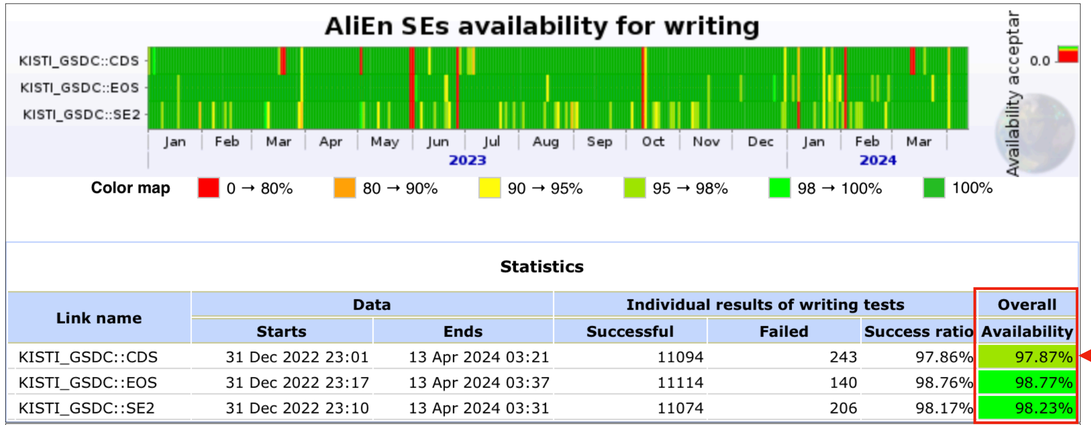


SE Status

	Used (PiB)	Total (PiB)	Usage (%)
CDS	5.766	15.78	36.54
EOS	1.62	5.948	27.24
SE2	1.172	1.446	81.05

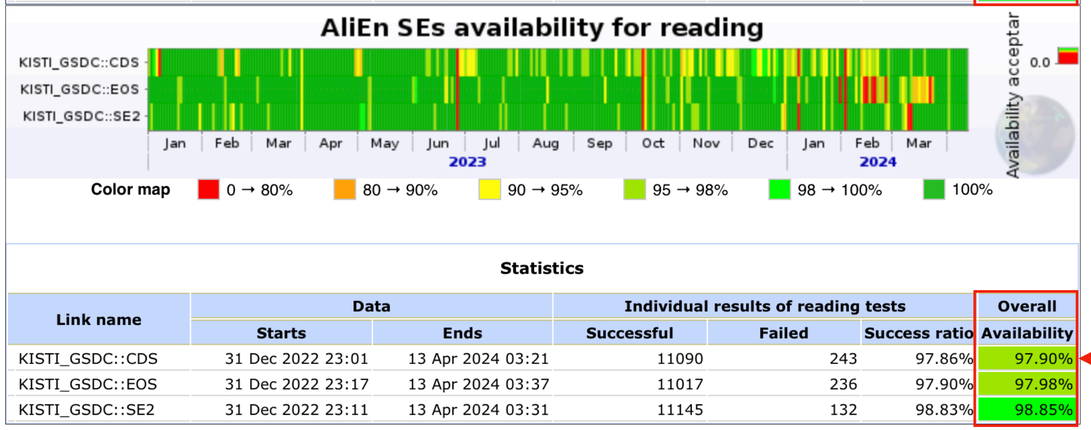


> 97% of overall SE availability for writing/reading



Statistics

Link name	Data		Individual results of writing tests			Overall
	Starts	Ends	Successful	Failed	Success ratio	Availability
KISTI_GSDC::CDS	31 Dec 2022 23:01	13 Apr 2024 03:21	11094	243	97.86%	97.87%
KISTI_GSDC::EOS	31 Dec 2022 23:17	13 Apr 2024 03:37	11114	140	98.76%	98.77%
KISTI_GSDC::SE2	31 Dec 2022 23:10	13 Apr 2024 03:31	11074	206	98.17%	98.23%



Statistics

Link name	Data		Individual results of reading tests			Overall
	Starts	Ends	Successful	Failed	Success ratio	Availability
KISTI_GSDC::CDS	31 Dec 2022 23:01	13 Apr 2024 03:21	11090	243	97.86%	97.90%
KISTI_GSDC::EOS	31 Dec 2022 23:17	13 Apr 2024 03:37	11017	236	97.90%	97.98%
KISTI_GSDC::SE2	31 Dec 2022 23:11	13 Apr 2024 03:31	11145	132	98.83%	98.85%

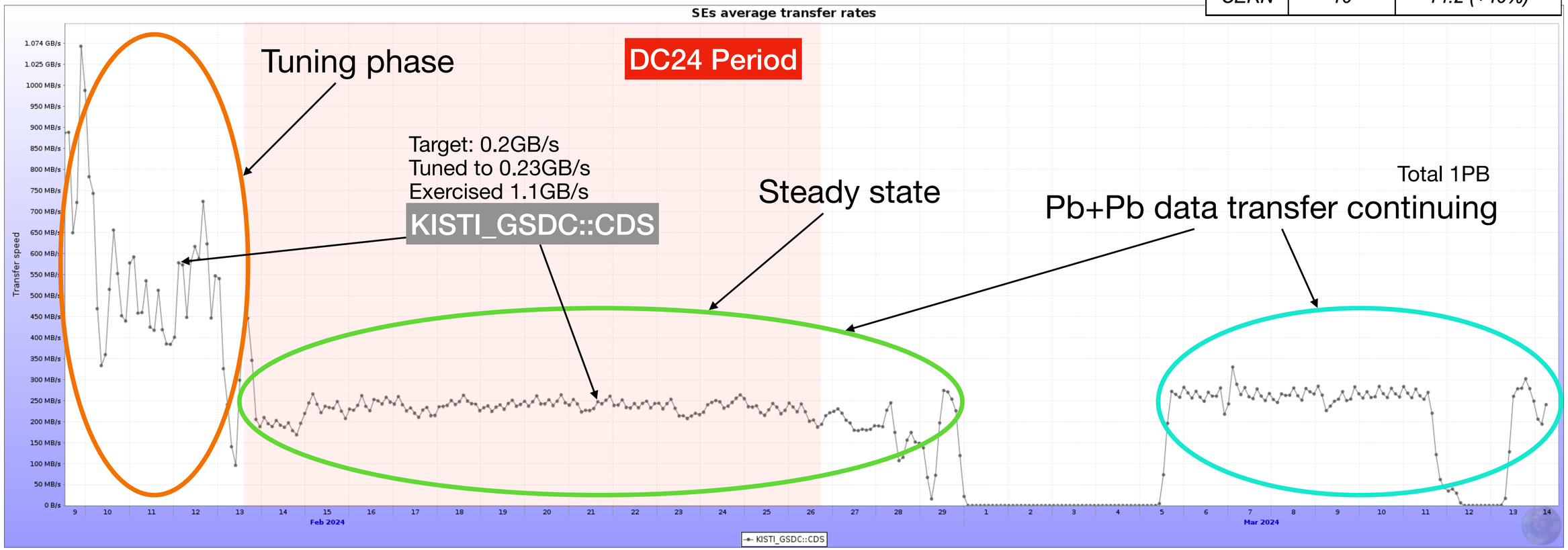
CDS : Custodial Disk Storage
EOS : EOS Disk Storage
SE2 : XRootD Disk Storage

WLCG Data Challenge 24

CDS Participation as a Tape

- Transfer of real Pb+Pb data collected in 2023, 34PB in total
- 1PB of data being transferred after the challenge, ETA end of March

Centre	Target rate GB/s	Average achieved GB/s
CNAF	0.8	0.98 (+20%)
IN2P3	0.4	0.6 (+40%)
KISTI	0.2	0.25 (+22%)
GridKA	0.6	1.12 (+90%)
NDGF	0.3	0.35 (+15%)
NL-T1	0.1	0.25 (+150%)
RAL	0.1	0.58 (+500%)
CERN	*10*	*14.2 (+40%)*

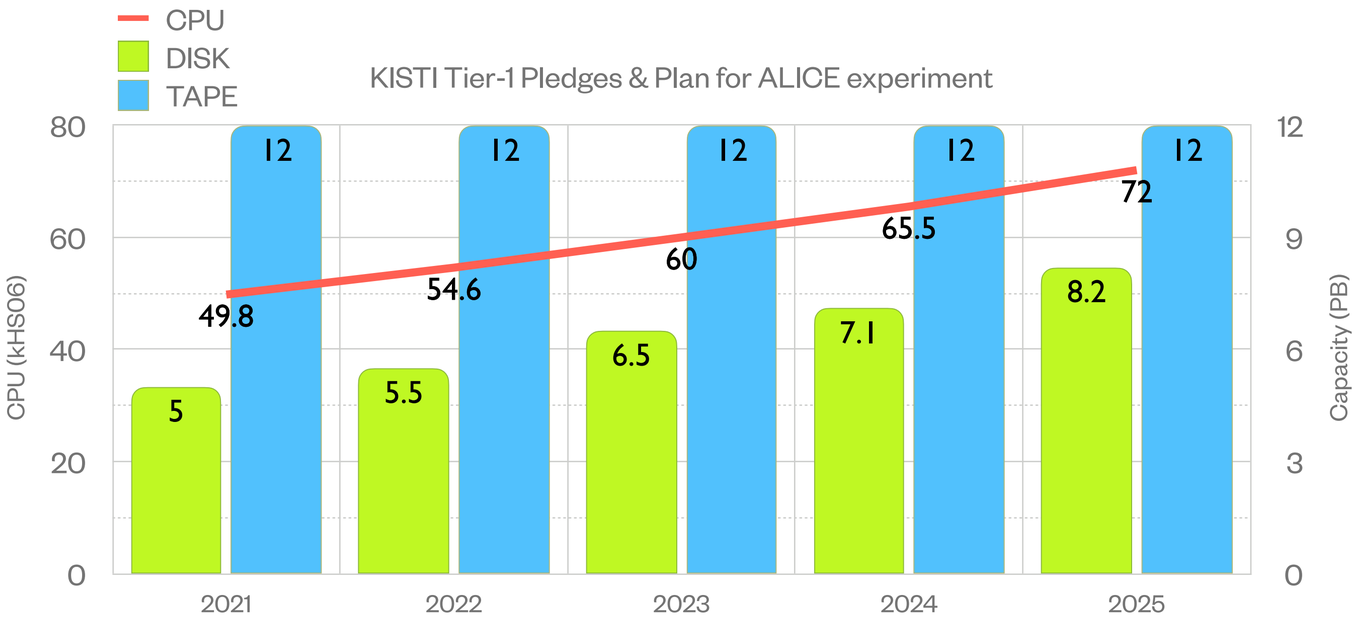


Site Availability

	2023												2024	
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**
Availability	98	97	99	100	98	94	98	98	99	98	95	97	97	97
Overall	97.5												97	

- Misconfiguration on LHCOPN backup link has affected significantly the site availability, persisting from the second half of last year until recently
  - Close collaboration with the KREONet service team to address the issue
- An unknown timeout issue during the EOS deployment on CentOS 9 Stream degraded availability in February and March of this year
  - Migrating to AlmaLinux 9.3 resolved the issue, attributed to the newer kernel on CS9 and complementary packages accompanying infrastructure provisioning for security, accounting, auditing and monitoring
- Availability re-computation requests vis GGUS tickets were approved and applied to WLCG monthly reports

Pledges



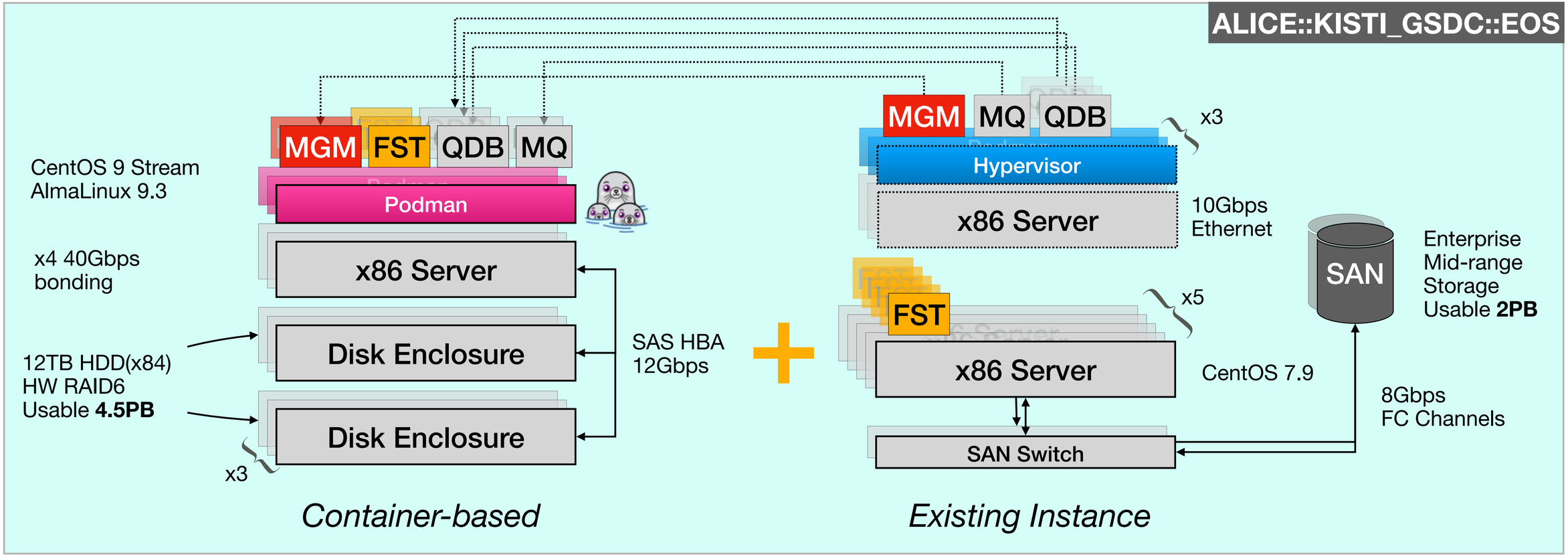
Installed (Plan)	2021	2022	2023	2024	(2025)
CPU (cores)	3,880	3,880	3,880	6,920	8,040
DISK (TB)	4,500	4,500	6,500	7,100	8,200
TAPE (TB)	12,000	12,000	12,000	12,000	12,000

- Current deployment
  - CPU : 72 kHS23 (6.9k Cores)
  - Disk : 8 PB (7.4 PiB)
  - CDS (TAPE) : 12 PB
- Plan
  - CPU : 84 kHS23 (+12 kHS23 later this year)
  - Disk : 7.5 PB
    - SE2 -1.5 PB (decommissioning this year)
      8 PB → 6.5 PB
    - EOS +3.0 PB to replace old HW -2.0 PB
      6.5 PB → 7.5 PB
  - CDS : 15 PB (tentatively)
    - Expanding in 2025 with 2024 procurement

EOS

EOS Deployments (1/2)

Disk Storage

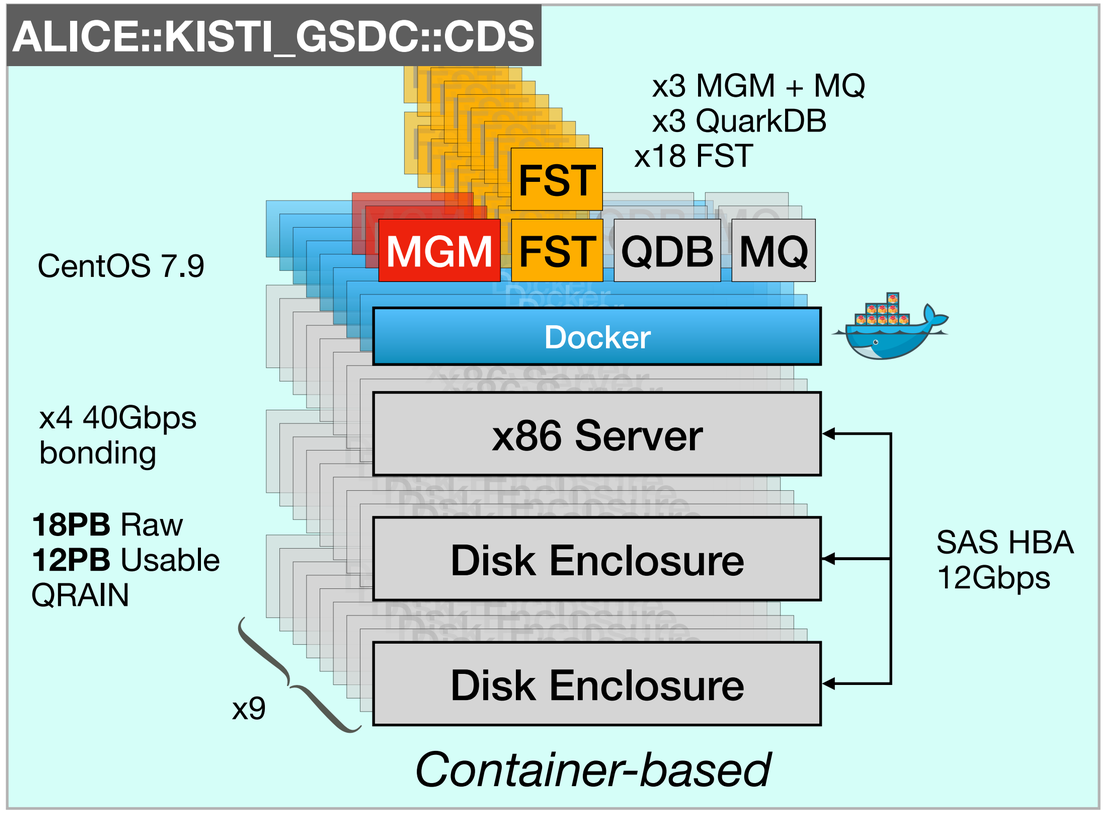


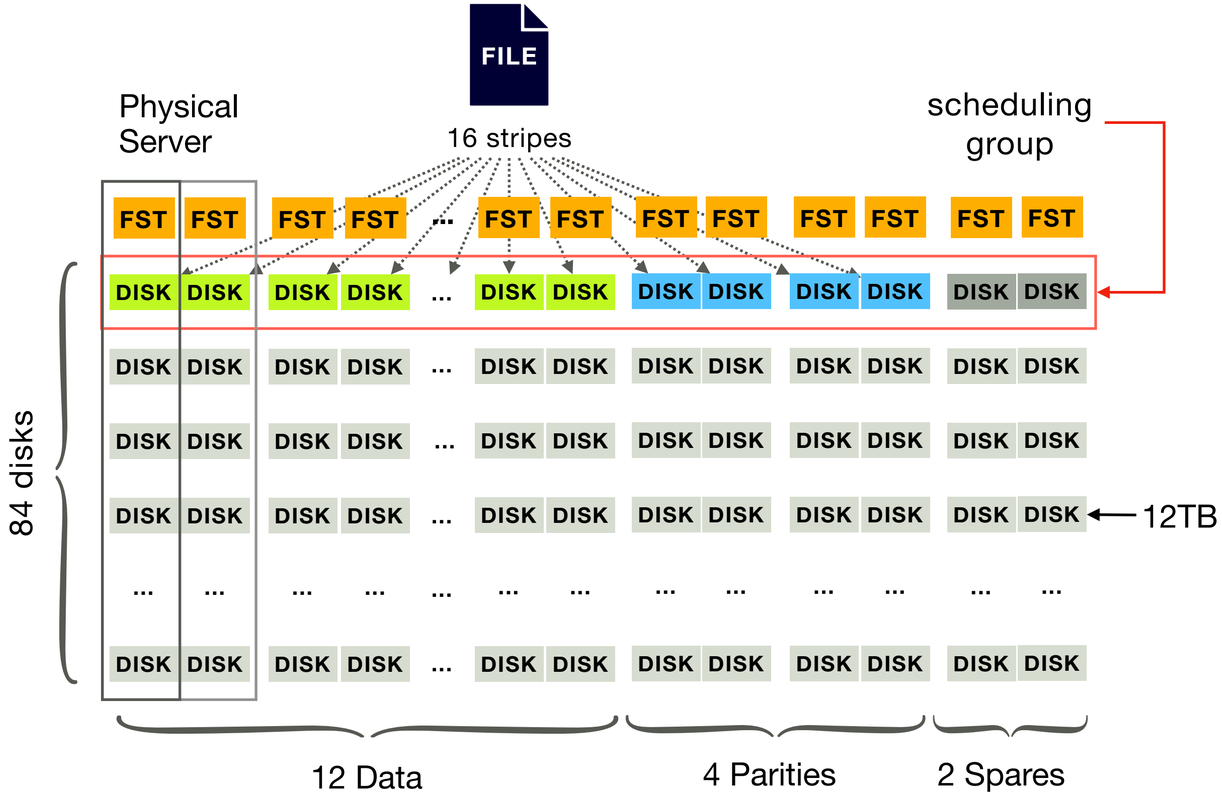
- Transparent transition of MGM and QuarkDB clusters from VMs to Containers
- EOS upgrade from 5.1.22 to 5.2.16 for existing setup, FMD migration from LevelDB completed beforehand
- Expanded to 6.5PB

EOS Deployments (2/2)

Custodial Storage

```
[root@jbod-mgmt-07 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"
```





- Disk-based Raw Archive storage for ALICE in production since 2021 deployed using Docker Container
- Comparable level of data protection provided by QRAIN Layout (12 stripes + 4 parities + 2 spares)
- Successful upgrade to v5.1.22 from v4.8.82 (May 2023)

EOS @ KISTI for ALICE

ALICE::KISTI_GSDC::EOS

Disk storage elements

KISTI_GSDC - EOS																							
	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 - EOS	ALICE::KISTI_GSDC::EOS	1	5.948 PB	1.639 PB	4.309 PB	27.55%	50,149,564	FILE	5.948 PB	1.74 PB	4.208 PB	29.25%	Xrootd 5.6.7	5.2.16					14.03.2024 14:43	25	0	0	
Total			**5.948 PB**	**1.639 PB**	**4.309 PB**		**50,149,564**		**5.948 PB**	**1.74 PB**	**4.208 PB**				**1**	**1**	**1**	**1**					**1**

ALICE::KISTI_GSDC::CDS

Custodial storage elements

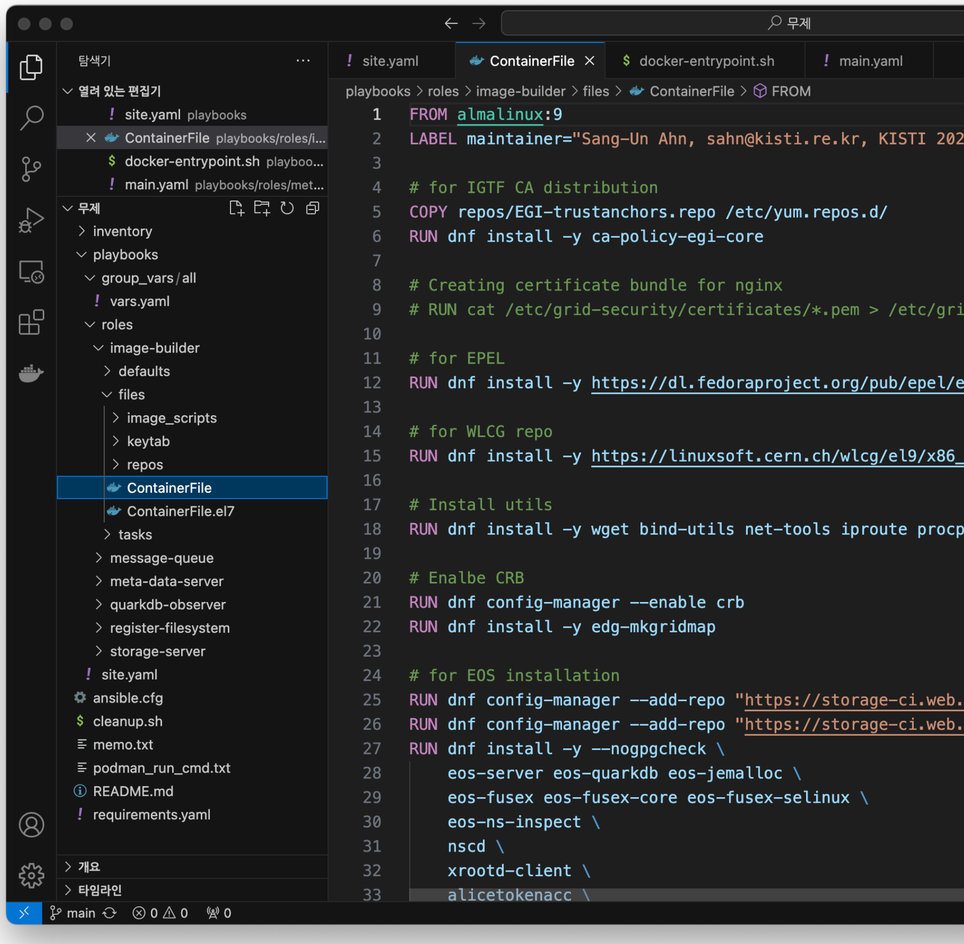
CDS																							
	AliEn SE		Catalogue statistics						Storage-provided information						Functional tests					Last day add tests		Demotion	IPv6
SE Name	AliEn name	Tier	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version	EOS Version	add	get	rm	3rd	Last OK add	Successful	Failed	factor	add
1. KISTI_GSDC - CDS	ALICE::KISTI_GSDC::CDS	1	15.79 PB	5.378 PB	10.41 PB	34.06%	10,959,791	FILE	15.76 PB	7.909 PB	7.856 PB	50.17%							14.03.2024 14:27	24	0	4.706%	
Total			**15.79 PB**	**5.378 PB**	**10.41 PB**		**10,959,791**		**15.76 PB**	**7.909 PB**	**7.856 PB**				**1**	**1**	**1**	**1**					**1**

- IPv4/IPv6 Dual Stack
- ALICE-Specific Token Authentication/Authorization, HTTP(S), Third-Party Copy enabled
- MLSensor (successor of EOS Apmon) deployed for monitoring (not yet for CDS)

EOS v5 Container on EL9: Practices (1/3)

Podman container runtime

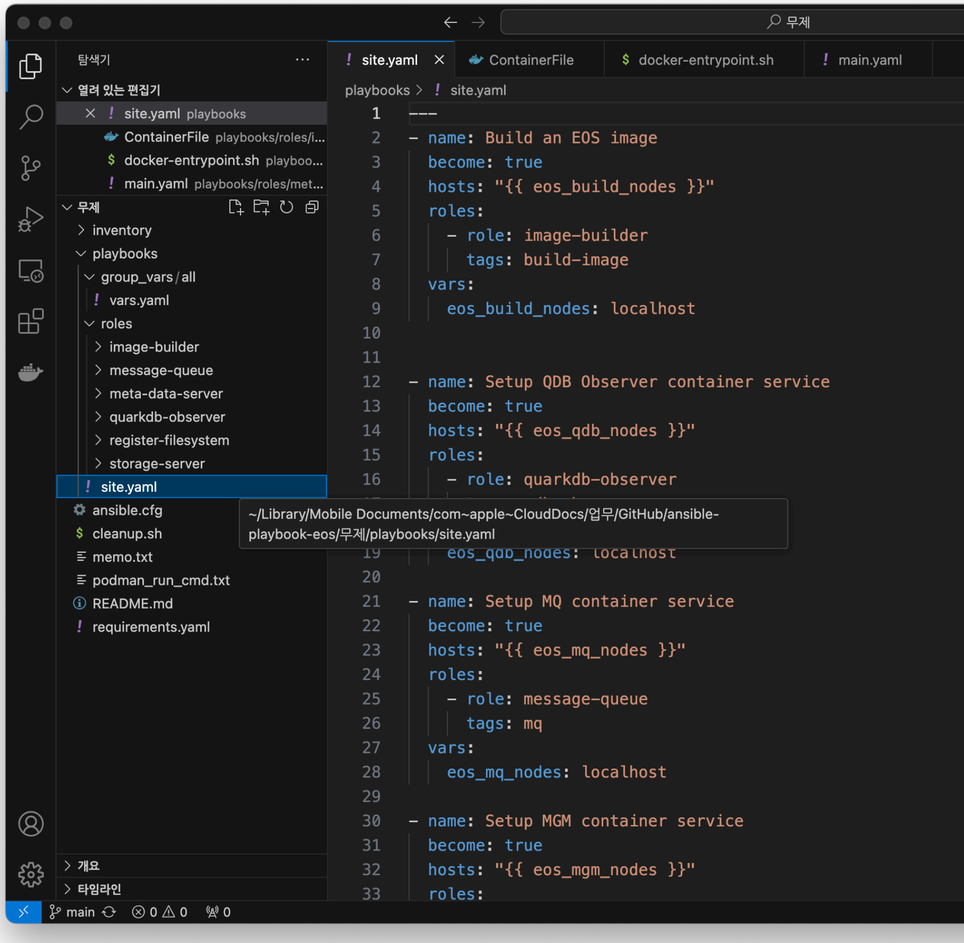
- Exploiting native support of EL9 for Podman (daemonless)
- ContainerFile (cf. *Dockerfile*)
  - EL9 base images - { CentOS 9 Stream | AlmaLinux 9.3 }
  - EOS v5 EL9 release installation
  - Reused Container entry script for CDS deployment (Docker based)
    - A few modification made to accommodate different monitoring scheme of ALICE: *eosapmond →mlsensor*



EOS v5 Container on EL9: Practices (2/3)

Automation via Ansible Playbook

- Playbook structure:
  - *site.yaml* - tags defined to perform specific operation (role) in an automated way
  - group_vars
    - *vars.yaml* - key-value variables for group parameters such as *eos_instance_name*, *eos_geotag*, ports, master/slave MGM FQDNs, QDB cluster and FST data directories
  - roles
    - *image-builder* | *message-queue* (MQ) | *meta-data-server* (MGM) | *quarkdb-observer* (QDB) | *register-filesystem* | *storage-server* (FST)
    - *handlers* defined to invoke *firewalld* policy implementation and *systemd* integration
    - Creating essential configuration files by templating *xrd.cf.{qdb|mq|mgm|fst}*, *eos_env*, *scitokens*, ALICE-specific (*TkAuthz.Authorization & mlsensor)*, etc.



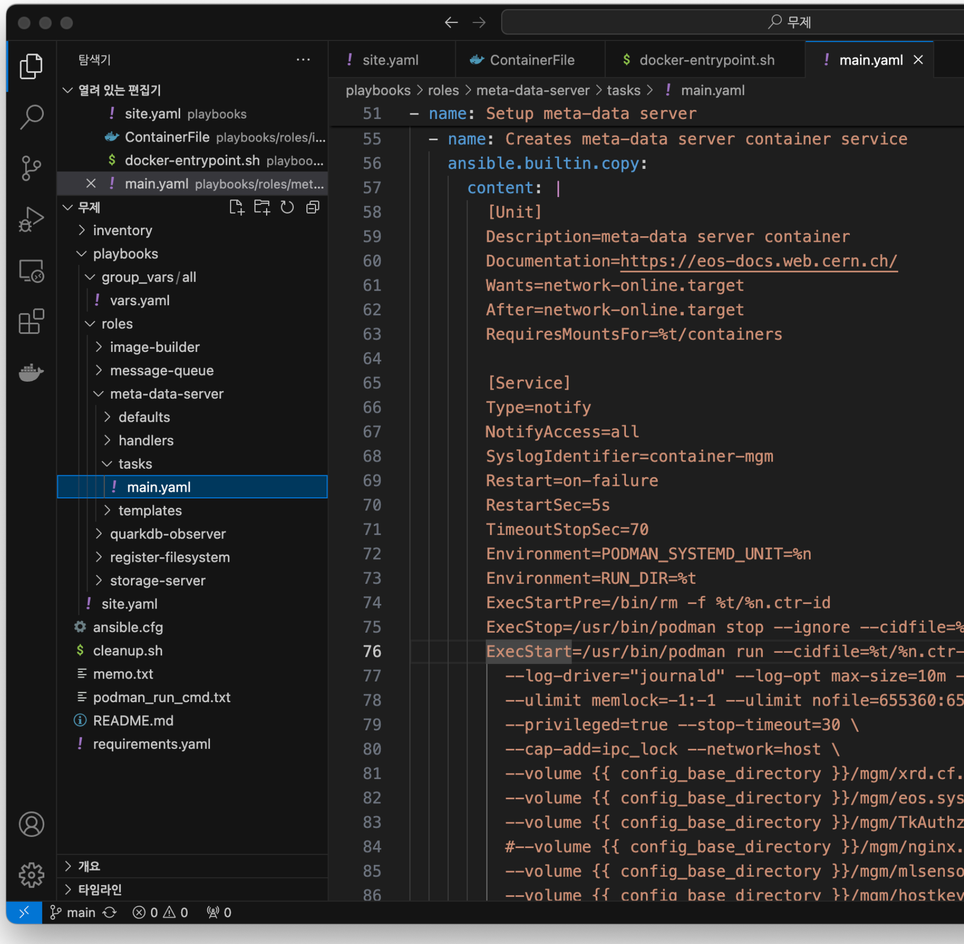
EOS v5 Container on EL9: Practices (3/3)

Systemd Integration

- Systemd service file for each of EOS components manipulating podman commands in such a way that it invokes *podman {run|rm|stop|...}*
  - E.g. */etc/systemd/system/{qdb|mq|fst|mgm}-container.service*

    <...>
    ExecStart=/usr/bin/podman run < parameters >
    ExecStop=/usr/bin/podman stop
    ExecStopPost=/usr/bin/podman rm
    <...>

  - *systemdctl {start|stop|restart} {qdb|mq|fst|mgm}-container.service*
    - *syslog* (*journalctl*) traces container logs (= podman logs)
  - Service update as well as roll-back can be quick and easy
    - Update images (pulling from registries or uploading from local one)
    - *systemctl restart *-container.service*



Plan

- Further work on EOS deployment playbook to run on AWX system
- Expanding EOS Disk for ALICE further up to 7.5 PB to meet pledges
  - FST nodes running on bare metals (2PB) to be decommissioned
    - Group draining could help to vacate there FSTs
- Updating EOS CDS to v5 as well as upgrading to EL9 flavour
  - Heavy revisions required on CDS Docker deployment

HPC

HPC for ALICE @ KISTI

Collaboration with KISTI Nurion Team & CBNU (a member of KoALICE)

HPC (Nurion)
- 25.7 PFlops
- 8,305 Knight Landing nodes
- 21PB Lustre Disk storage
- OPA Interconnection
- PBS Pro

Payload Requests / Monitoring

Payloads

Enqueuing JA
CernVM-FS Caching

Large Bandwidth
40-100Gbps

Data Transfer

Central Services
- TaskQueue
- Catalog

ALICE

JA submission

Monitoring

GLOBAL SCIENCE EXPERIMENTAL DATA HUB CENTER
GSDC
글로벌 대용량실험데이터허브센터

NOVA APERIO
1951

Site Services
- VO-Box
- Compute Entry (CE)
- Storage Element (SE)
- Squid-proxies for CernVM-FS

Considerations

- A "*Knight Landing*" node has 68 cores and 96GB memory (1.4GB RAM per core)
- Assuming 100MB/s bandwidth per single job slot, (may) start with 1,020 cores (15 nodes) for testing

ALICE HPC Requirements

- Preemption policy allowing backfilling
- Java compatibility (ALICE Job Agent runs on JVM)
- CernVM-FS
- 2GB RAM & 10GB space per single core job (resources proportional to # of cores)

- Bypassing external authentication procedure that allows (direct) submission from CE (or VO-Box) @ GSDC
- Pool accounts (or dedicated one to be mapped)
- Scratch storage (spooling input files, storing output)
- autofs for CernVM-FS mount
- Java Runtime to run JVM binaries on CernVM-FS
- Allowing out-going connectivity
- ...

Payload Requests / Monitoring

Payloads

Enqueuing JA
CernVM-FS Caching

Large Bandwidth
40-100Gbps

Data Transfer

Central Services

- TaskQueue
- Catalog

ALICE

JA submission

Monitoring

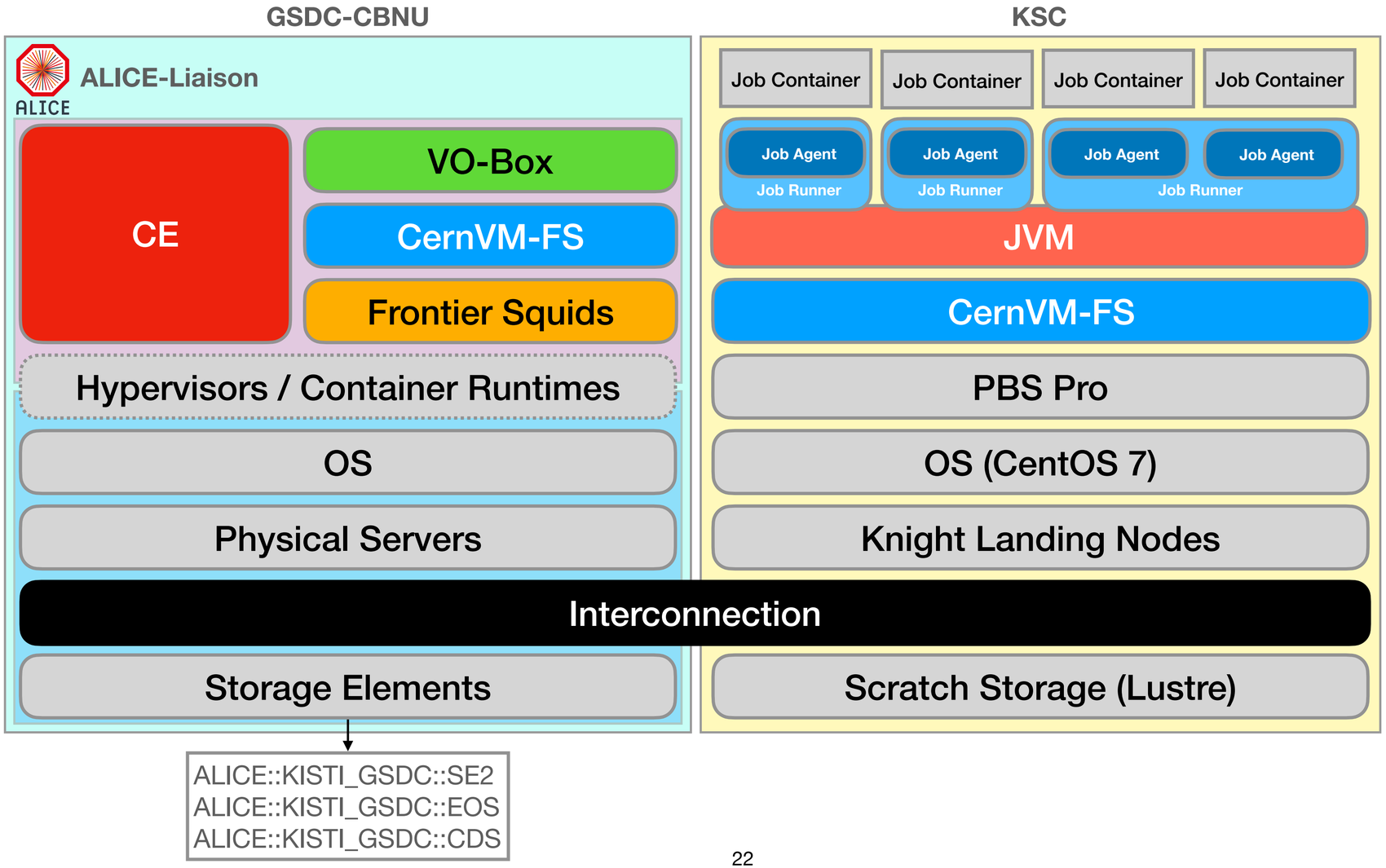
GLOBAL SCIENCE EXPERIMENTAL DATA HUB CENTER
GSDC
글로벌 대용량실험데이터허브센터

NOVA APERIO
1951

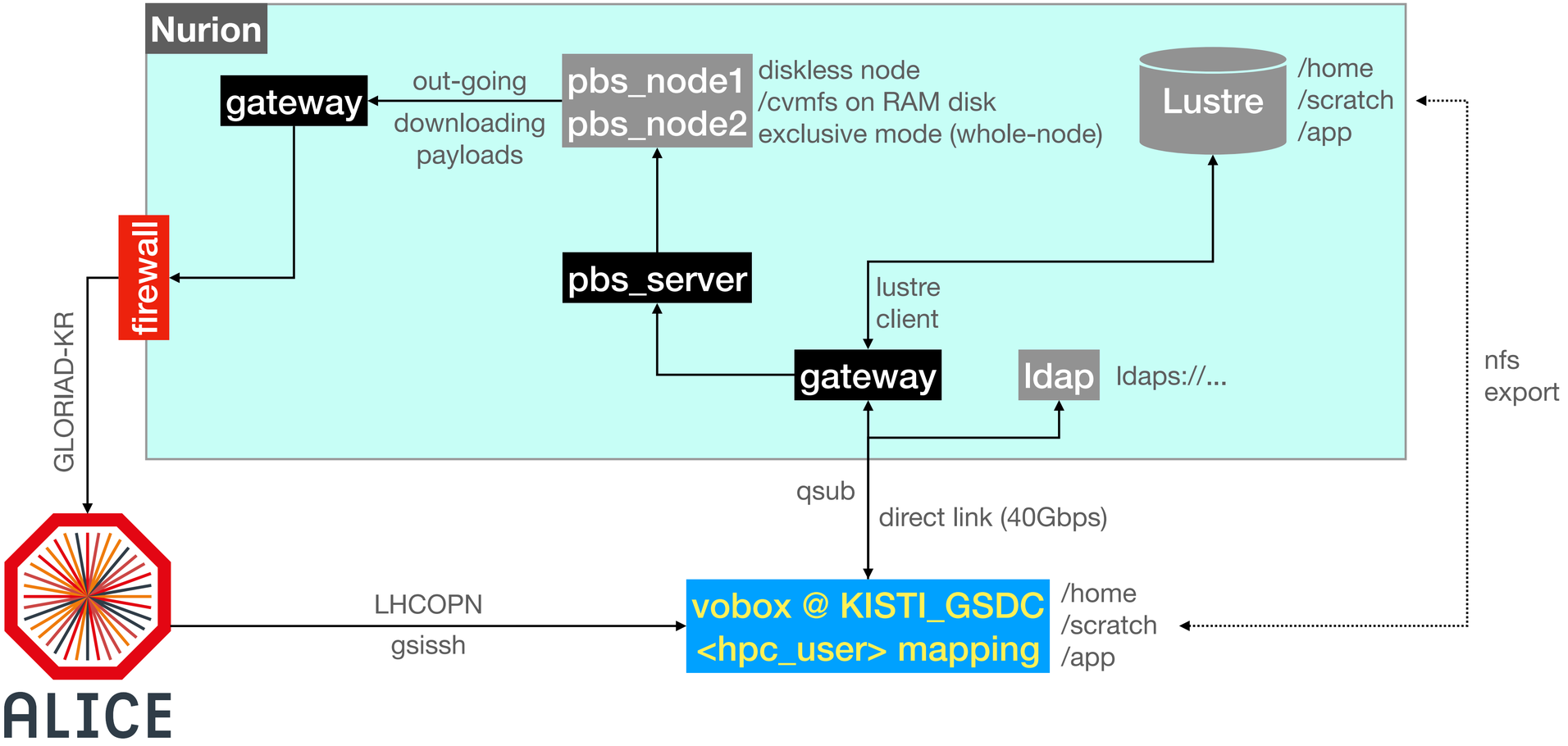
- Standalone VO-Box (independent of T1)
  Unique Site identifier (e.g. KISTI_Nurion)
- Standalone CE (independent of T1)
  Compatible to HPC batch scheduler
  PBS Pro + ARC-CE (or direct submission?)
- Accessible SEs @ GSDC
- Standalone (load-balanced) Squid-proxy
- ...

Prototyping Site Services

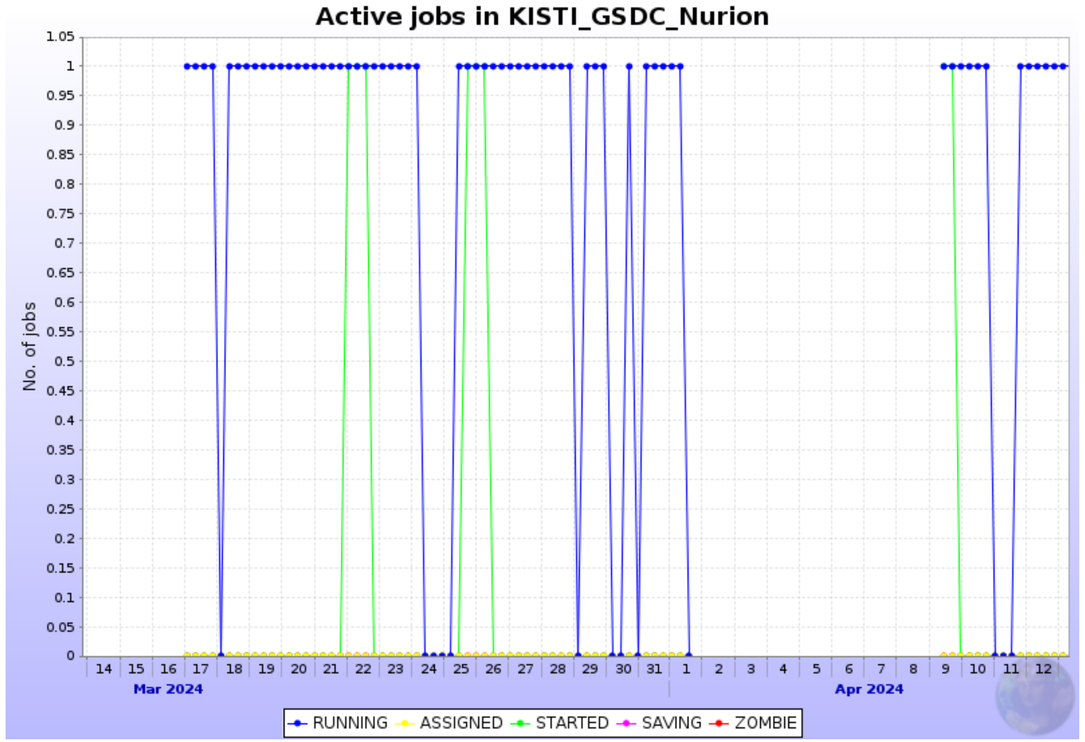
KISTI_GSDC_Nurion

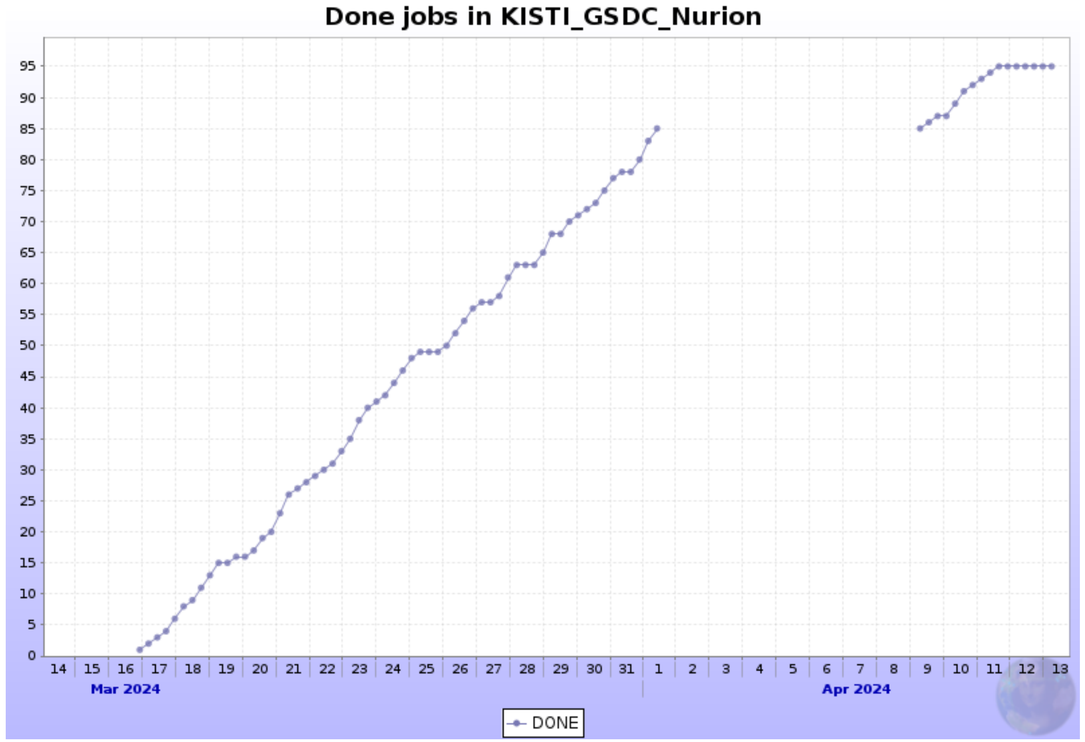


Current Test-bed Status



Successful Jobs





Summary

Summary

- KISTI Tier-1 for ALICE experiment has been operating without critical issues
  - Configuration change made for HTCondor to accept multi-core as well as whole-node submission jobs
  - New and powerful machines were deployed to meet CPU pledges
  - Successful participation to WLCG DC24 in early this year
  - OPN backup network and storage OS issues affected site availability (re-computation approved)
- Major EOS disk deployment with container technology on EL9 flavour has been successful
  - With efforts on deployment automation and systemd integration
- KISTI HPC Project has conducted in collaboration with a KoALICE member and KISTI Nurion Team
  - Successful jobs have been observed since March this year