

KISTI-GSDC Report

Geonmo Ryu, Byungyun Kong
On behalf of KISTI-GSDC



25 September 2025 @ ATCF9

KISTI

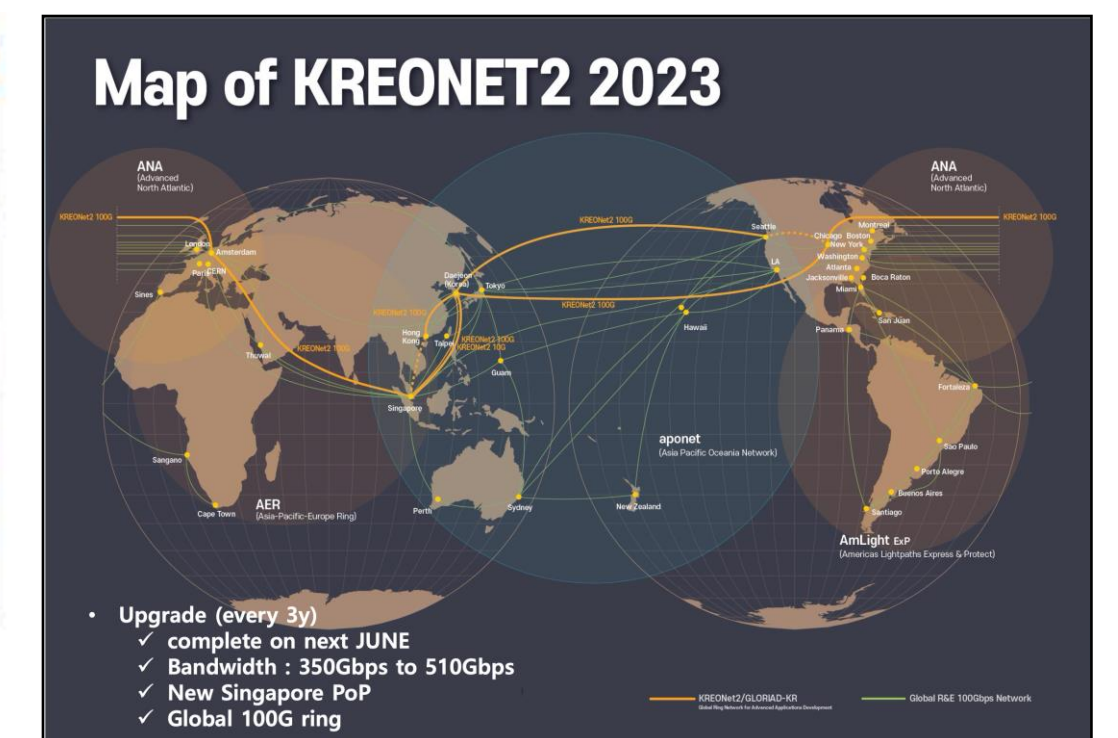
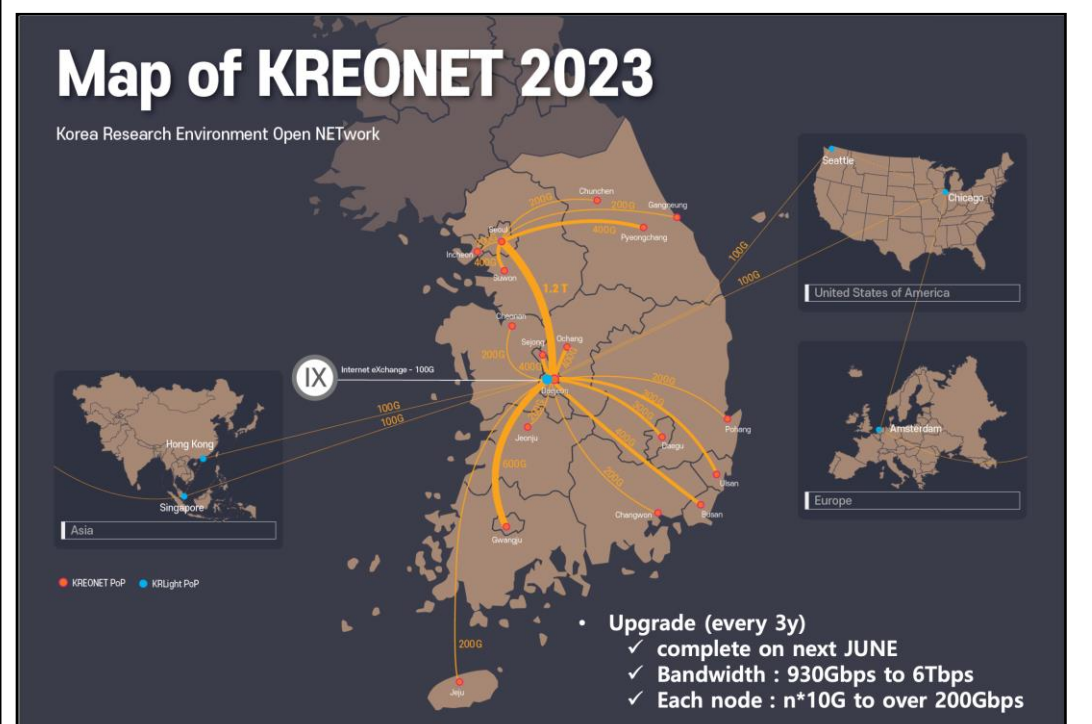
Korea Institute of Science and Technology Information



- Government-funded research institute founded in 1961 for national information services and supercomputing
- National Supercomputing Center
 - **5th Supercomputer**
 - **Nurion** - Cray CS500 system
 - 25.7 PFlops at peak, ranked 11th of Top500 (2018) ⇒ 46th (Nov 2022)
 - **Neuron** - GPU system, 1.24 PFlops
 - **6th Supercomputer (HANGANG)**
 - 600 PFlops at peak, 8,496 GPUs
 - Will start next year
- **KREONet/KREONet2** - National/International R&E network



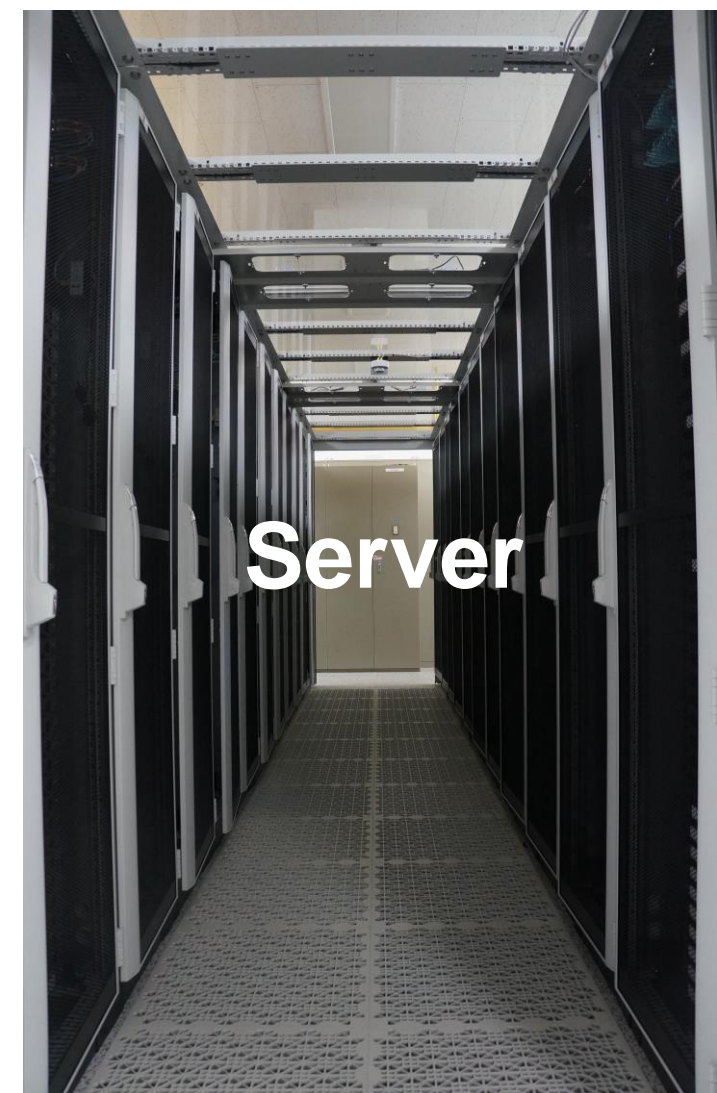
※ CPU(Central Processing Unit) : 복잡한 수치해석 및 다양한 제어와 연산에 적합.
 GPU(Graphics Processing Unit) : 일부 연산에 대해 대규모 데이터의 병렬 처리를 지원하며, 단순한 수치해석 및 AI 분야에 적합



GSDC

Global Science experimental Data hub Center

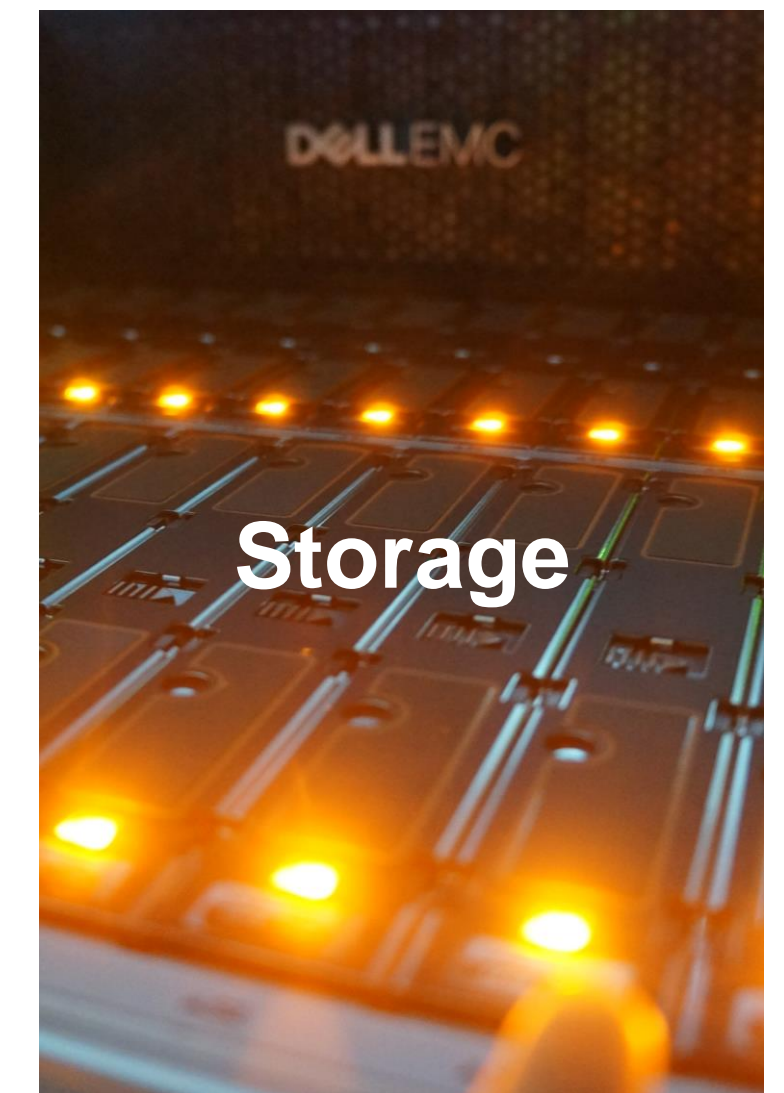
- Government-funded project, started in 2009 to promote Korean fundamental research through providing computing power and data storage
- **Datacenter for data-intensive fundamental research**
 - Preserving data from domestic or overseas large and complex scientific instruments as well as simulation-R&D activities
 - Providing services based on technology development: distributed computing structure, high availability storage system, infra integrated management, disk-based custodial storage



Server

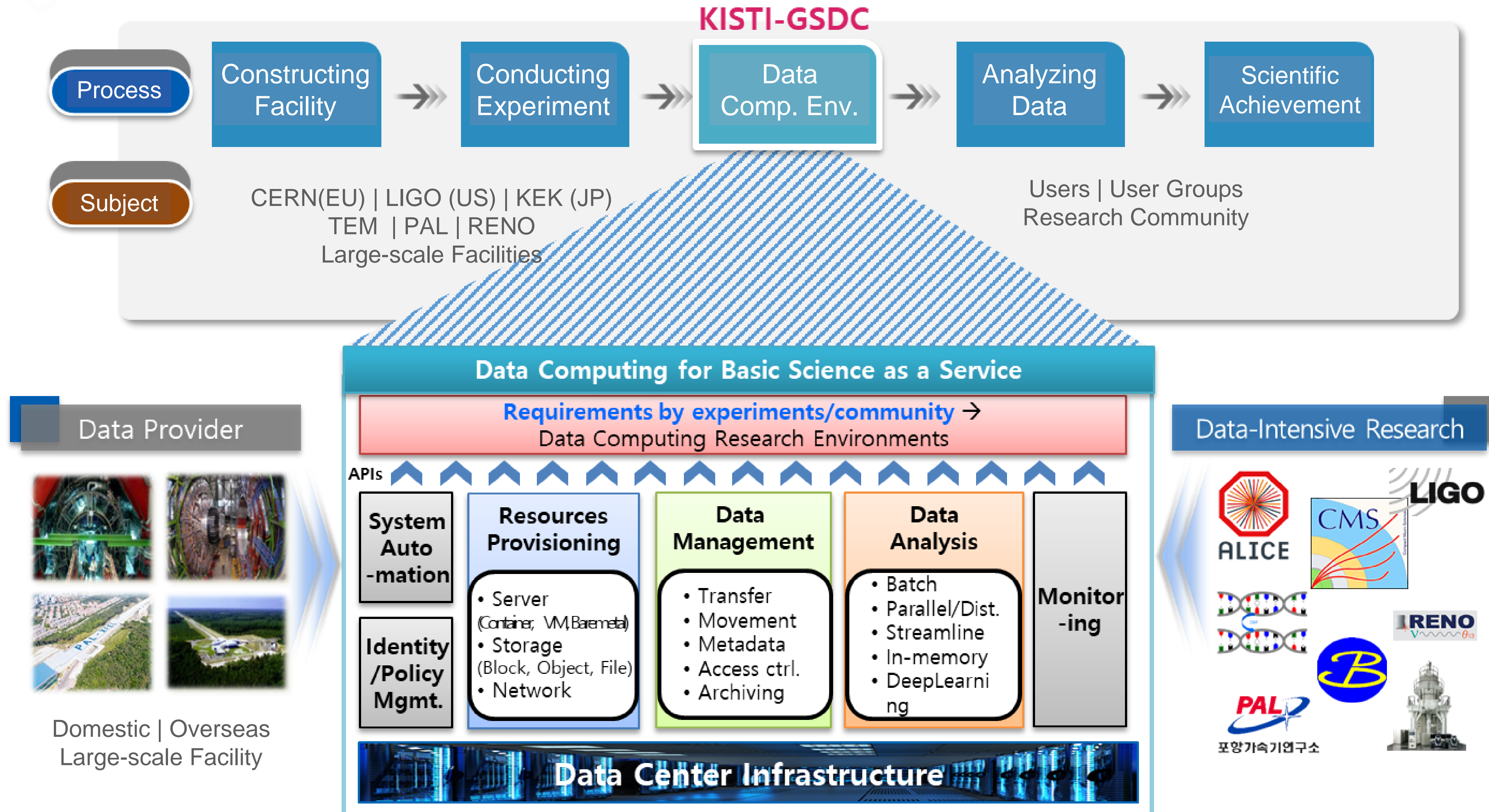


Network

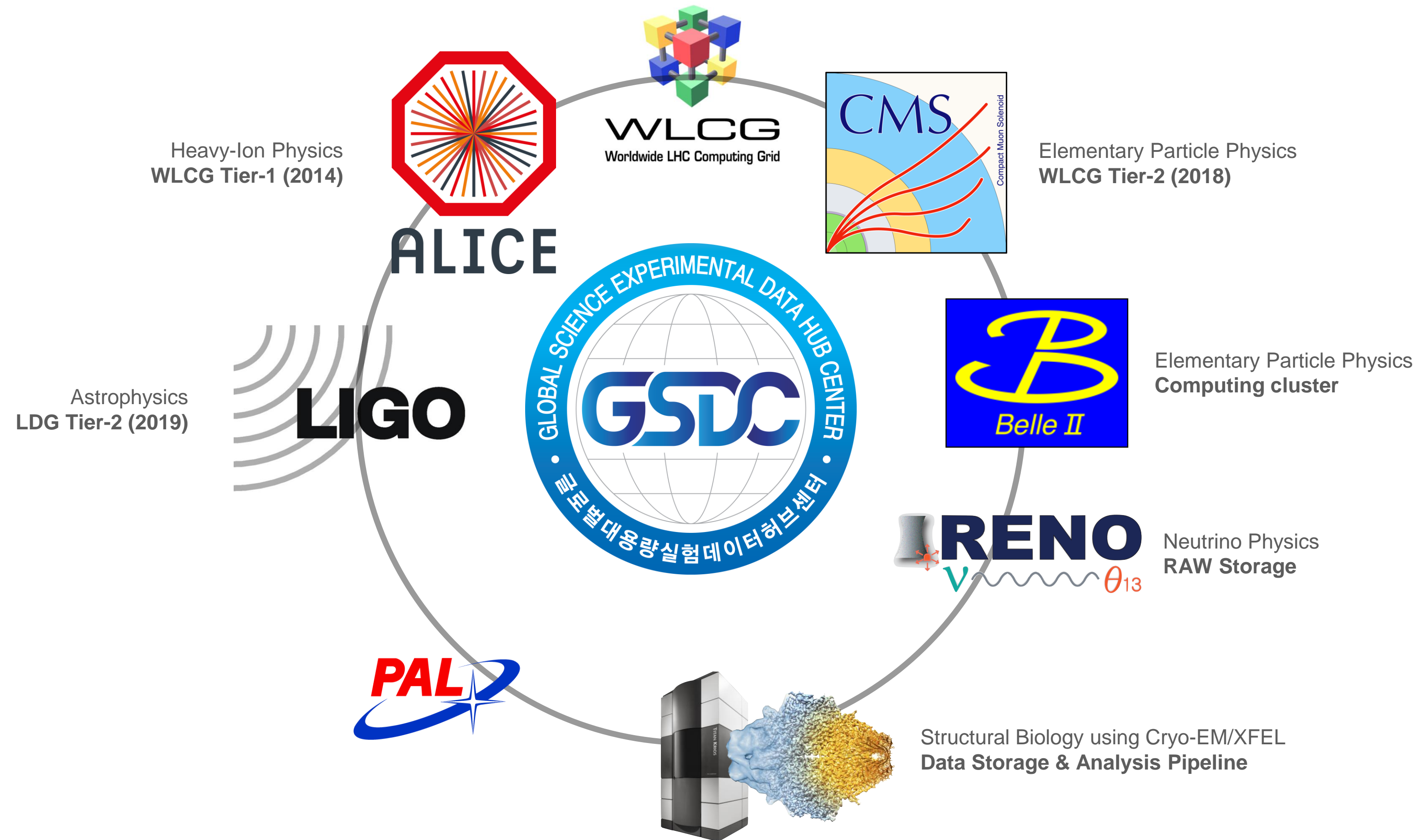


Storage

Role of GSDC for Data-intensive Research



Supporting Experiments

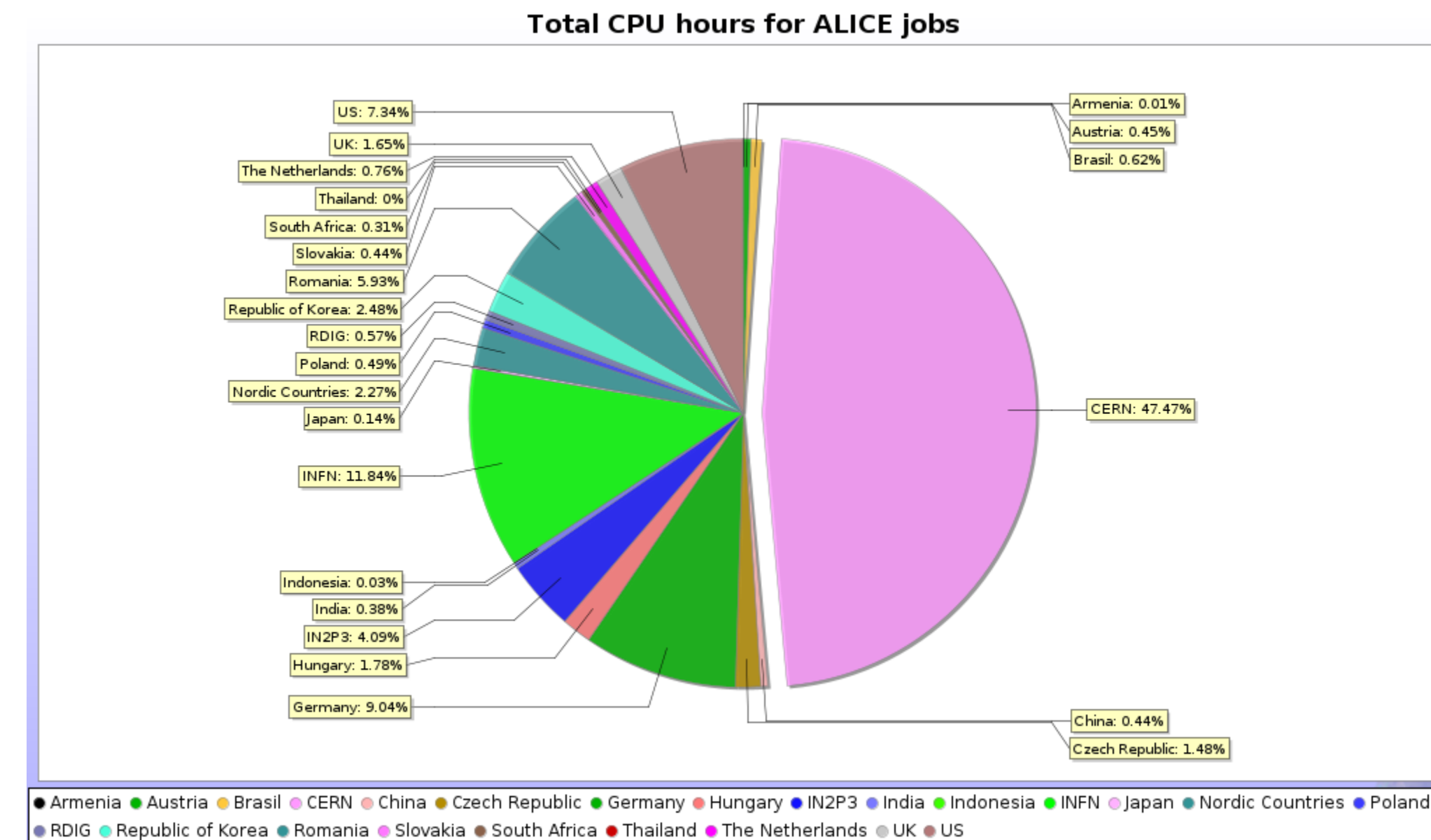


WLCG Tier-1 @ KISTI-GSDC

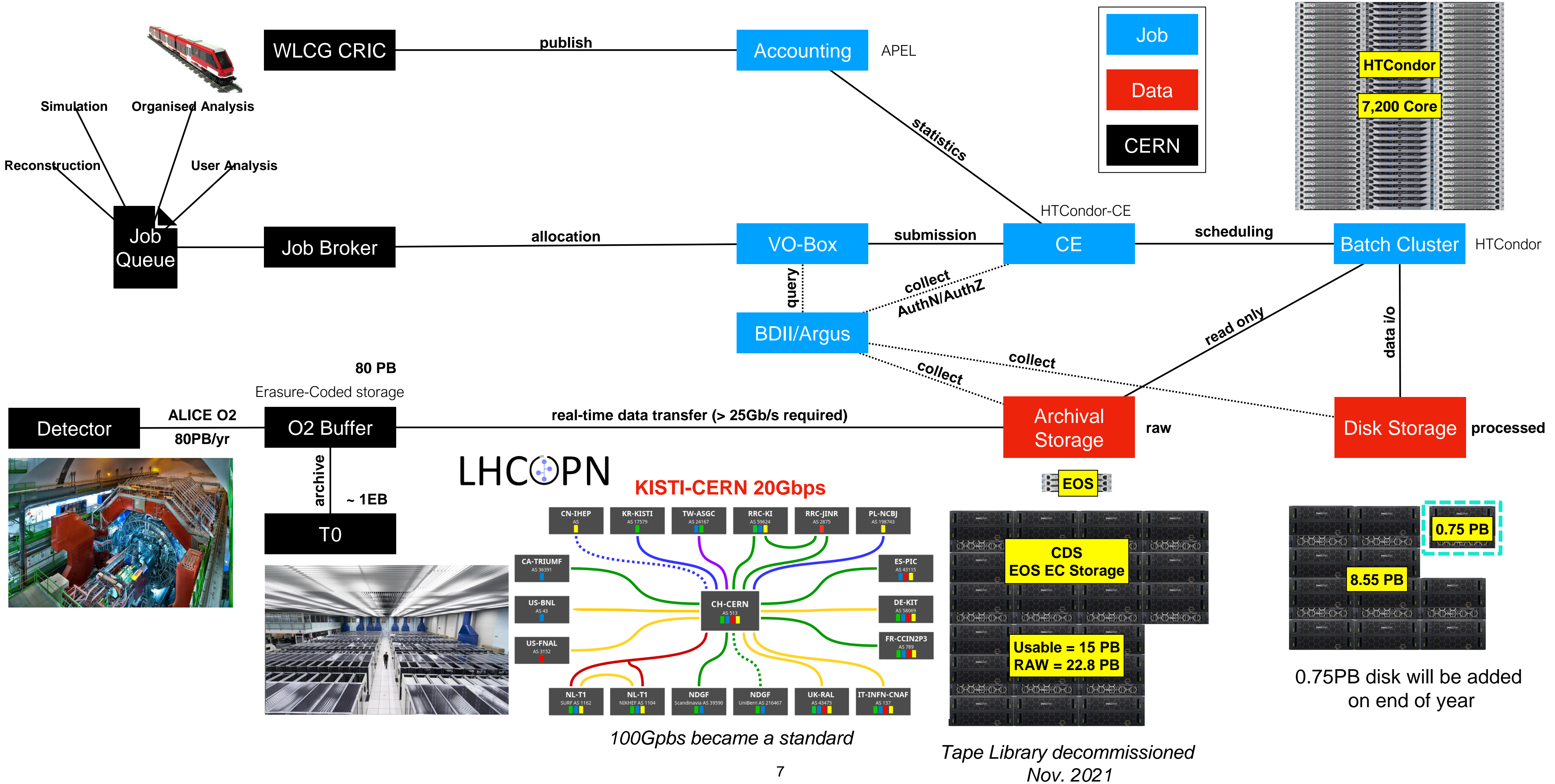
Flagship Service for Data-intensive Computing



- A WLCG Tier-1 in Asia for the ALICE experiment
 - Contributing about 10% of T1 resource requirements of ALICE
 - More than 2% of total (T0+T1+T2+AFs) resource requirements of ALICE
 - Availability/Reliability: 98.88% / 99.13% (2025)
- CE
 - HTCondor-based, whole-node submission enabled (for N-core jobs)
- SE
 - XRootD/EOS based disk storage
 - Archival SE : CDS, the disk-based one powered by EOS
- Networking
 - LHCOPN : 20G dedicated link between Daejeon (KR) and Geneva (CH)
 - LHCONE : 100G provisioned by KREONet connecting to EU, US and Asia (SG/HK)



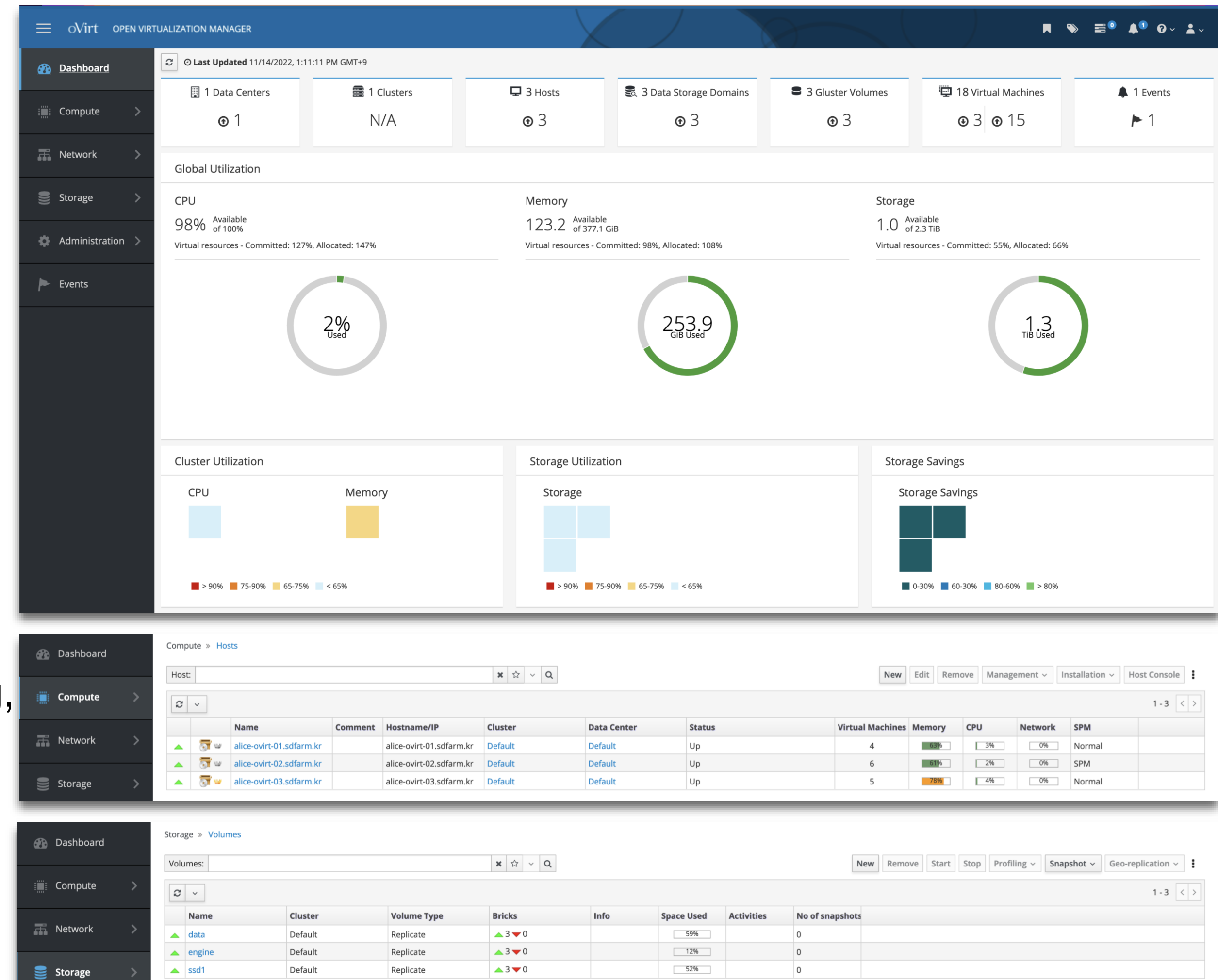
KISTI ALICE T1 Structure Overview



T1 Grid Services

Site Status

- Many grid services running on VMs provided by oVirt cluster
 - oVirt 4.3.8 + GlusterFS 6.10
 - 3 oVirt hosts with 384 GB of RAM and 2.3 TB of Gluster Storage (1.5 TB HDDs, 0.8 TB SSDs)
 - Live migration & load-balancing
- VMs for Grid services
 - VOBOX for job submission
 - Site-BDII & Argus (AuthN & AuthZ)
 - 3 Squid caches for CernVM-FS (Application provisioning, e.g. AliRoot, ROOT, GEANT4, etc.)
 - APEL (WLCG Accounting)
 - 3 HTCondor-CEs (CE & Condor 24.0.8-1)



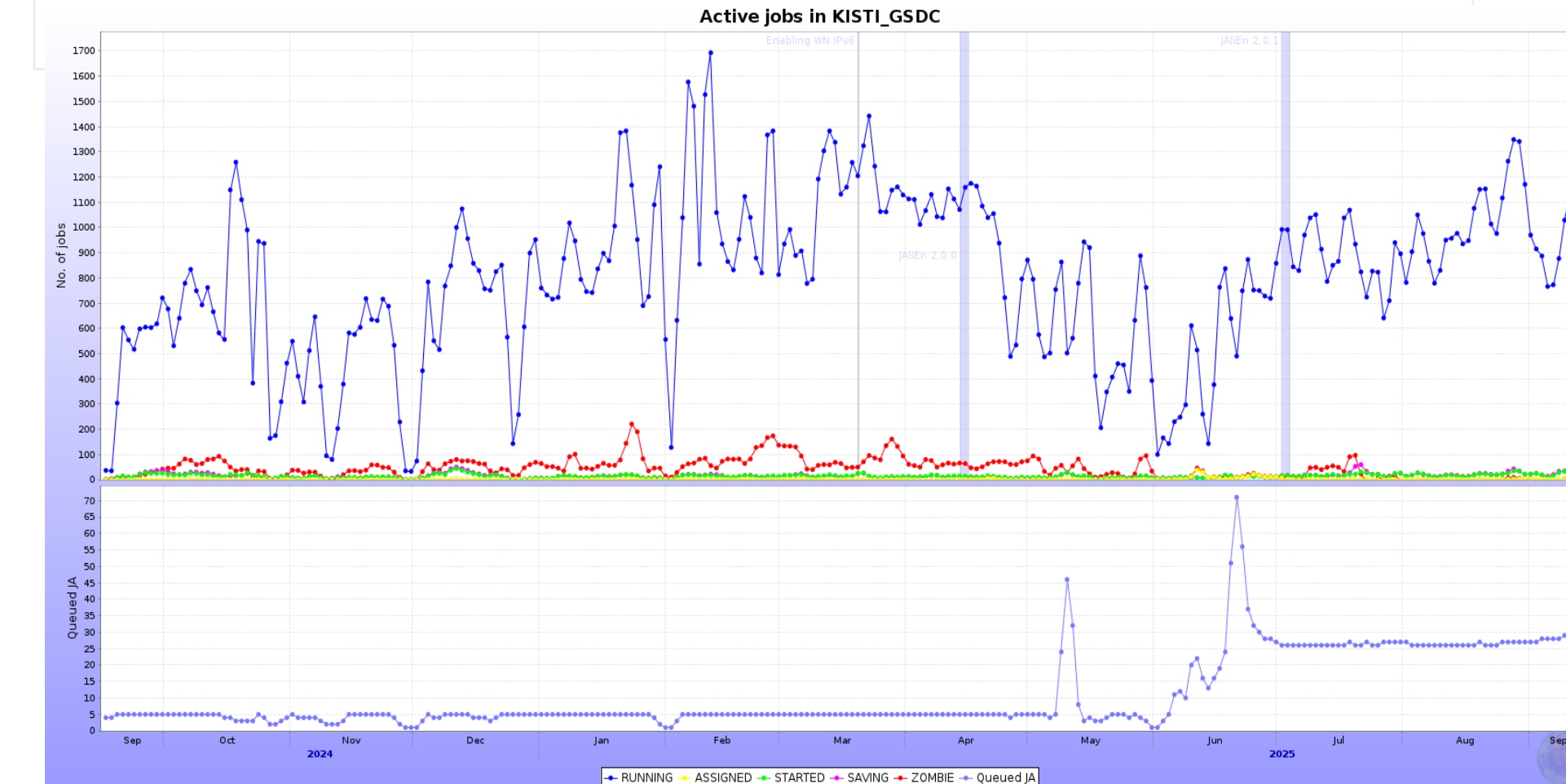
Computing Elements

- VOBOX
 - On the VOBOX machine, JAliEn CE and MonALISA are currently running.
 - The new VOBOX container has successfully replaced the old VM.

- Computing Resource

HTCondor-CE Hostname	Number of Worker Nodes	HepScore 23 Benchmark score
alice-t1-ce04.sdfarm.kr	12	21,029
alice-t1-ce05.sdfarm.kr	35	27,524
alice-t1-ce06.sdfarm.kr	55	25,092

- Total: 73.6 kHS23 / Newer than 2020 : 50.8 kHS23
- Pledge: 69 kHS23 (2025) / 72 kHS23 (2026)
- Current Status
 - Servers without a functioning web console have been excluded from the total count
- Plan
 - Add about 22.8 kHS06 this year by replacing old Worker Nodes
 - New AMD servers prepared with 192 cores each (expected ~4–5k HS23 per server)
 - HS23 benchmark values still require adjustment due to power limit factors
- Goal: phase out all equipment manufactured before 2020



	Series	Last value	Min	Avg	Max
1.	1	2	0	397.1	5001
2.	2	6	0	135.8	2284
3.	4	1	0	9.37	1115
4.	8	621	10	518.9	1016
5.	64	0	0	0	1

EOS Disk Status

```
[root@rbod-mgmt-01 ~]# podman ps
CONTAINER ID  IMAGE                                COMMAND  CREATED  STATUS  PORTS  NAMES
fc893da4099d localhost/eos-local:latest           8 weeks ago Up 8 weeks  mq
d77d4f463b55 localhost/eos-local:latest           8 weeks ago Up 8 weeks (healthy)  qdb
f6bc9cbae6b8 localhost/eos-local:latest           7 weeks ago Up 7 weeks (healthy)  mgm
ade9ef23dbf8 localhost/eos-local:latest           7 weeks ago Up 7 weeks (healthy)  fst
```

- Structure

- Running a container-based EOS server directly with podman
 - Container Image: AlmaLinux 9.6
 - 5 EOS servers (v5.3.8)
 - 2x (MQ + MGM + QDB + FST) + 3x (QDB+FST)
 - MGM DNS Round robin (eos-disk.sdfarm.kr)

- Status

- Total : 8.55 PB (=7.597PiB)
- Used : 5.57 PB (=4.94PiB / 65.14%)

```
EOS Console [root://localhost] | /eos/alicekistigsdc/grid/> node ls
```

type	hostport	geotag	status	activated	heartbeatdelta	nofs
nodesview	jbod-mgmt-10.sdfarm.kr:1095	kisti::gsdc::g04	online	on	1	12
nodesview	rbod-mgmt-01.sdfarm.kr:1095	kisti::gsdc::g05	online	on	1	12
nodesview	rbod-mgmt-03.sdfarm.kr:1095	kisti::gsdc::g05	online	on	1	12
nodesview	rbod-mgmt-04.sdfarm.kr:1095	kisti::gsdc::g06	online	on	1	12
nodesview	rbod-mgmt-06.sdfarm.kr:1095	kisti::gsdc::g06	online	on	1	24

- Plan

- Will be added more space on December
 - + 0.75 PB \approx 9 PB (2026 Pledge)

```
EOS Console [root://localhost] | /eos/alicekistigsdc/grid/> space ls
```

type	wfe	ntx	name	groupsize	groupmod	N(fs)	N(fs-rw)	sum(usedbytes)	sum(capacity)	capacity(rw)
			active	intergroup						
spaceview			default	12	24	72	70	5.57 PB	8.55 PB	8.30 PB

SE Name	AliEn name	Tier	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version	EOS Version
1. KISTI_GSDC - EOS	ALICE::KISTI_GSDC::EOS	1	7.597 PB	4.899 PB	2.698 PB	64.48%	83,927,425	FILE	7.597 PB	4.944 PB	2.653 PB	65.08%	Xrootd 5.7.3	5.3.8
Total			7.597 PB	4.899 PB	2.698 PB		83,927,425		7.597 PB	4.944 PB	2.653 PB			

CDS Archiving Storage

- Structure

- Running a container-based EOS server directly with podman on 9x CentOS7 and 5x AlmaLinux9 hosts
 - Container Image: AlmaLinux 9.6
 - 14 EOS servers (v5.3.17)
 - 3x (MGM) + 5x (QDB+FST) + 6x (FST)
 - MGM DNS Round robin (eos-archive.sdfarm.kr)

- Status (Installed / Usable after EC)

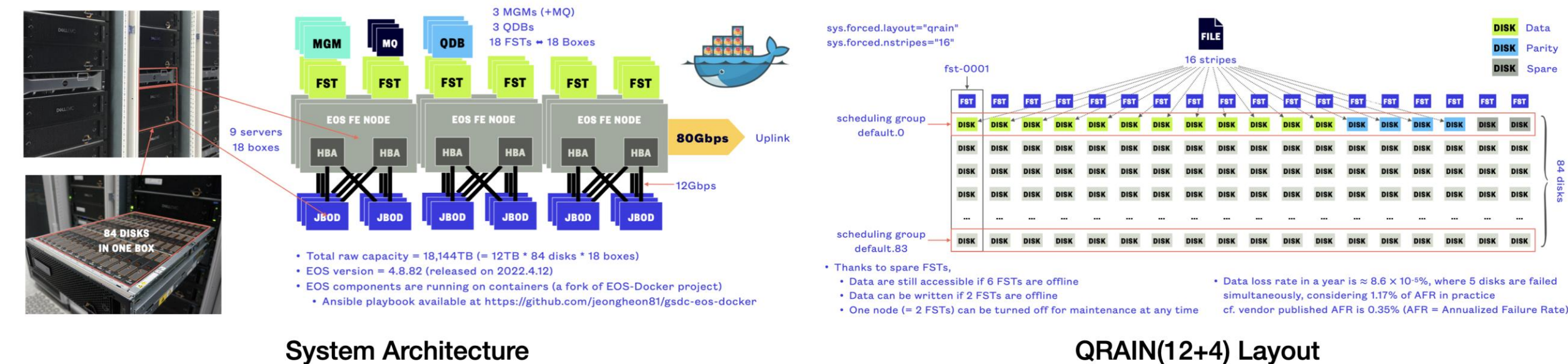
- Total : 22.80 / 15.2 PB (= 20.25 / 13.5PiB)
- Used : 20.28 / 13.5 PB (=18.01 / 11.99PiB) [88.9%]

- Plan

- Replace old storage servers (maintenance expired) with new equipment
- Secure 15 PB usable capacity (2026 pledge)
- Slightly less total space, but improved stability

Custodial Disk Storage Tapeless Archiving

- The first disk-based custodial storage replaced tape for ALICE experiment
- 12 PB usable space with 12+4 erasure coding for data protection (powered by CERN EOS)
- Fully automated deployment of EOS components using Linux containers



Custodial storage elements														
GSDC	AliEn SE			Catalogue statistics (1024-base units)					Storage-provided information (1024-base u					
	SE Name	AliEn name	Tier	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version
1. KISTI_GSDC - CDS	ALICE::KISTI_GSDC::CDS	1	15.79 PB	11.9 PB	3.891 PB	75.36%	13,457,187	FILE	20.25 PB	18 PB	2.248 PB	88.9%	Xrootd 5.8.3	
Total			15.79 PB	11.9 PB	3.891 PB		13,457,187							

Structure of CMS Tier-2 @KISTI

- **Services run on BareMetal servers**

- OS installation by Foreman
- Provisioning with Ansible

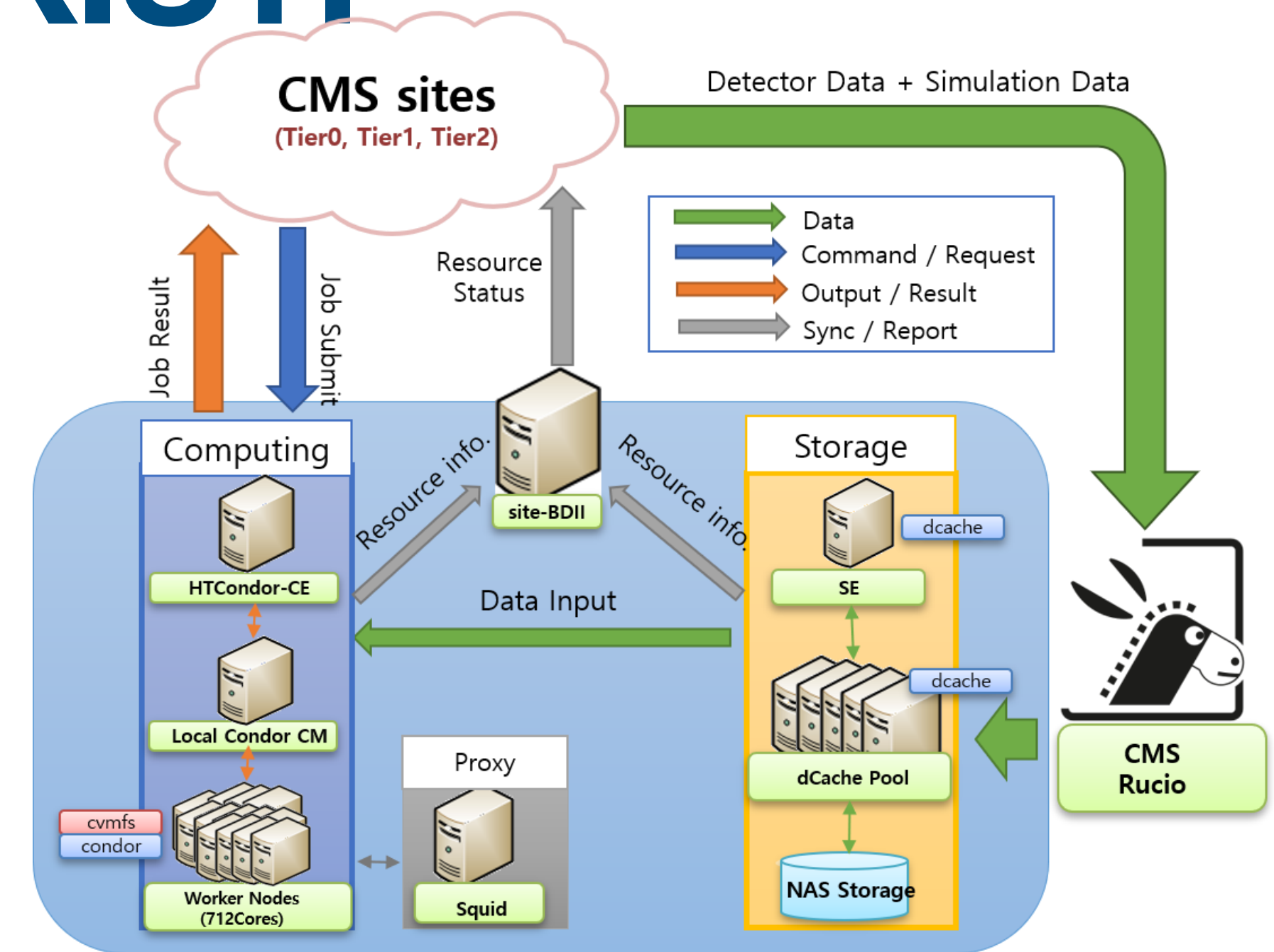
- Ansible Roles:

<https://github.com/orgs/gsdcr/repositories?q=visibility%3Apublic+archived%3Afalse>

- **Main Component**

- Middleware : UMD5 (+ wlcg repo)
- CE : HTCondor-CE v23.0.28
- LRMS : HTCondor (2x CentralManager using HTCondor HA)
 - 46 WNs / 1,424 logical cores / 14.9 kHS06
 - RAM 2,000MB per core
- SE : dCache v9.2 + XRootD PSS server for CMS AAA
 - 9 x dCache Pool Server (10Gbps)
 - 8 JBOD + 9 NFS Pools / 2819TiB (=3100TB)
 - Protocol : XRootD (RO), WebDAV (RW)
- Report : 1x (site-BDII+APEL publisher)
- Cache : 2x Frontier-Squids

- Network: Shared-100Gbps from GSDC to KREONET



Disk Space Usage

CellName	DomainName	Total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/sticky/total free)
cms-t2-wn1055-NFSPool	cms-t2-wn1055-NFSPool-Domain	168032251	25452419	0	
cms-t2-wn1056-JbodPool	cms-t2-wn1056-JbodPool-Domain	209706693	21686181	0	
cms-t2-wn1056-NFSPool	cms-t2-wn1056-NFSPool-Domain	167253788	15719296	0	
cms-t2-wn1057-JbodPool	cms-t2-wn1057-JbodPool-Domain	3383037	0	0	
cms-t2-wn1057-NFSPool	cms-t2-wn1057-NFSPool-Domain	3651671	0	0	
cms-t2-wn1058-JbodPool	cms-t2-wn1058-JbodPool-Domain	3386795	0	0	
cms-t2-wn1058-NFSPool	cms-t2-wn1058-NFSPool-Domain	169411942	26942054	0	
cms-t2-wn1059-JbodPool	cms-t2-se-node01-JbodPool-Domain	209687520	31037469	0	
cms-t2-wn1059-NFSPool	cms-t2-se-node01-NFSPool-Domain	168325171	26313503	0	
cms-t2-wn1060-JbodPool	cms-t2-wn1060-JbodPool-Domain	209689601	43932761	0	
cms-t2-wn1060-NFSPool	cms-t2-wn1060-NFSPool-Domain	169653546	26171361	0	
cms-t2-wn1061-JbodPool	cms-t2-wn1061-JbodPool-Domain	209704854	35397054	0	
cms-t2-wn1061-NFSPool	cms-t2-wn1061-NFSPool-Domain	167700994	17287007	0	
cms-t2-wn1062-JbodPool	cms-t2-wn1062-JbodPool-Domain	172162668	23493010	0	
cms-t2-wn1062-NFSPool	cms-t2-wn1062-NFSPool-Domain	170184951	33074404	0	
cms-t2-wn1063-NFSPool	cms-t2-wn1063-NFSPool-Domain	168550454	24866432	0	

One pool server has been replaced due to failure.

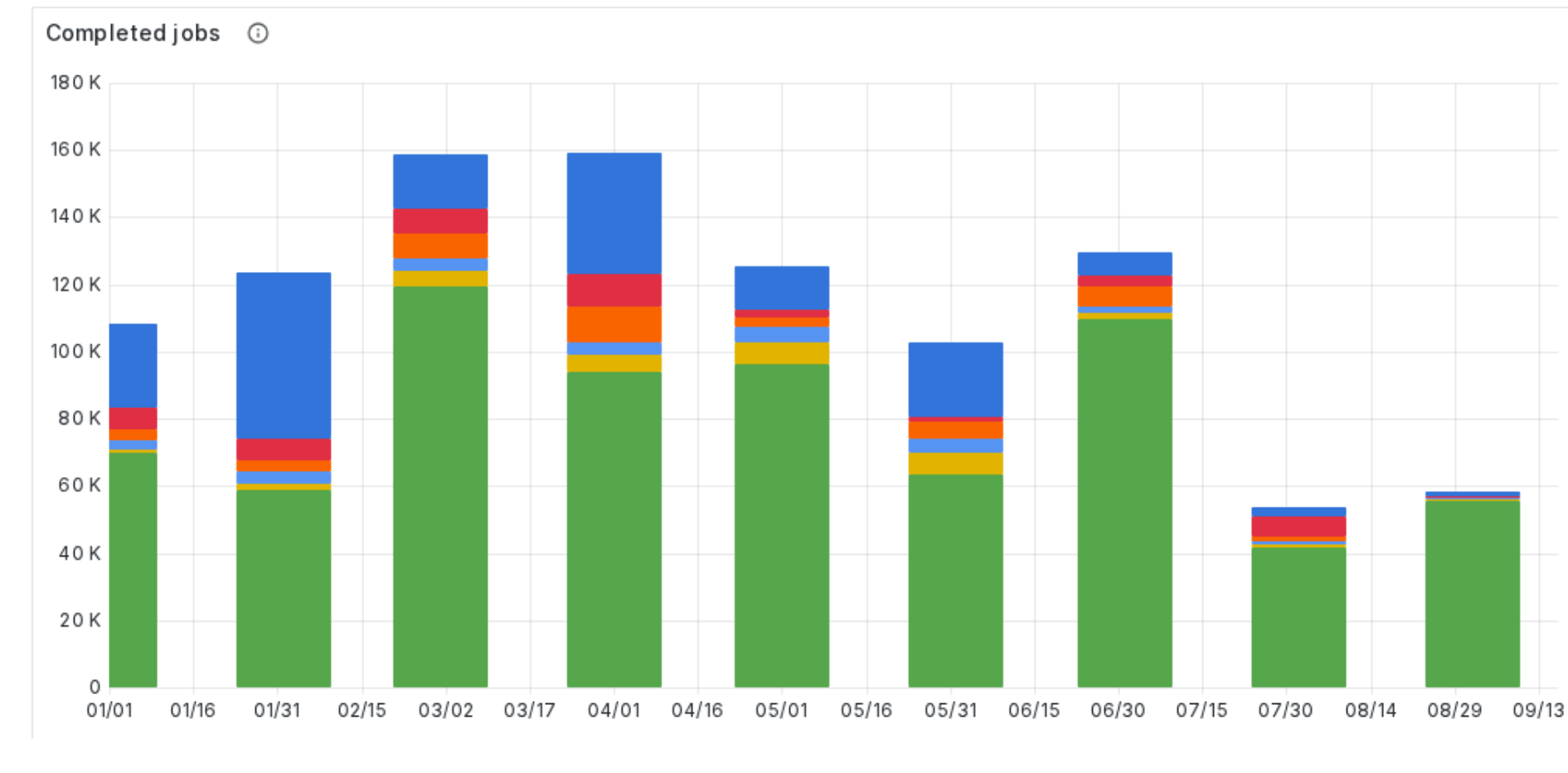
CMS T2 Present and Future

• Site Status

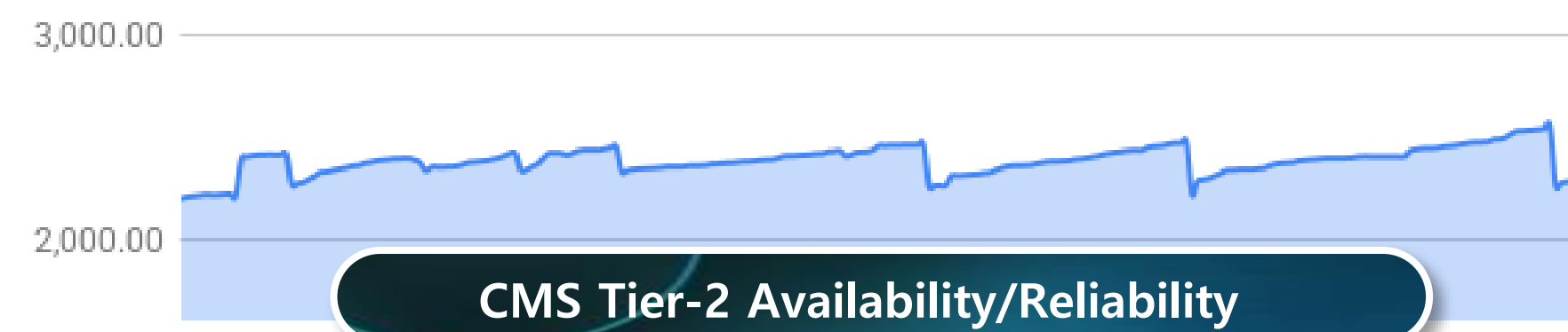
- Computing Resource: 14.9kHS06
 - 1,018,899 jobs Completed
- Storage: 2,555/3,100TB (82.43%) [25-09-16]
- Availability/Reliability : 98.53%

• Plans and Updates

- **By the end of this year: Achieving the Resource Pledge**
 - Replacement of WN (from 15kHS23 → 48kHS23)
- **Early next year: System architecture changes**
 - Managing **resource-intensive services on bare metal**
 - Dedicated WN (CPU-intensive / 48kHS23)
 - dCache Pool (Disk-intensive)
 - Containerization of key services with **IPVLAN**, including:
 - Zookeeper cluster using 3 servers
 - Patroni cluster with 2 or more servers (zookeeper-backend) for PostgreSQL high availability
 - Ephemeral WN / HTCondor-CE / HTCondor Central Manager
 - dCache (Non-Pool)
 - Frontier Squids / site-BDII / APEL Publisher



CMS Tier-2 Storage Usages (TB)

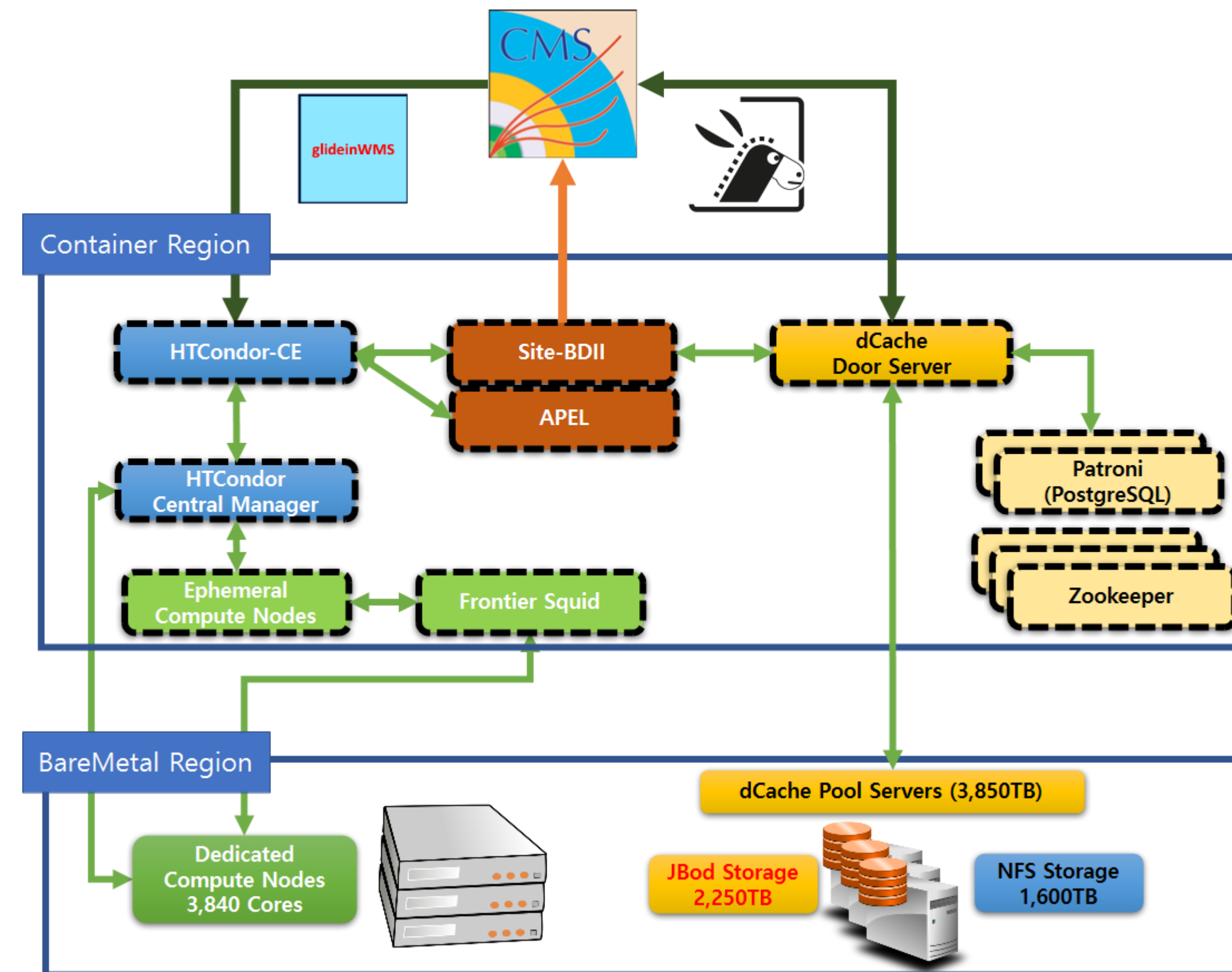


Site	Availability ↓	Reliability
T2_CH_CERN	99.71%	
T2_UK_London_IC	99.59%	
T2_US_Wisconsin	99.41%	
T2_UA_KIPT	98.97%	
T2_KR_KISTI	98.53%	
T2_IT_Legnaro	98.32%	
T2_FL_HIP	98.01%	
T2_HU_Budapest	97.87%	
T2_DE_RWTH	97.81%	
T2_US_Nebraska	97.80%	
T2_AT_Vienna	97.62%	
T2_RU_IHEP	97.37%	
T2_DE_DESY	97.34%	
T2_CN_Beijing	97.32%	
T2_UK_SGrid_RALPP	97.11%	
T2_PL_Swierk	96.84%	
T2_US_Vanderbilt	96.77%	
T2_RU_JINR	96.70%	
T2_BE_IHHE	96.34%	
T2_FR_GRIF	96.26%	
T2_US_Caltech	96.26%	
T2_TR_METU	96.25%	
T2_ES_CIEMAT	96.08%	
T2_BE_UCL	96.08%	
T2_US_UCSD	95.41%	

Future of KISTI CMS Tier-2

- Site Features
 - Pros
 - No need to change the network configuration compared to when using existing bare metal
 - Utilization of IPVLAN as the primary service network (minimizing overhead from overlay networks)
 - No additional programs required for operations beyond Podman
 - Custom agent and manager needed for container HA, but services can operate normally even without them
 - Cons
 - Every container instance must have a fixed IP address
 - This IP must be manually registered in DNS
 - There is no service support for persistent volumes, this must be taken into consideration
 - APEL's Accounting DB volume

Planned Architecture for KISTI CMS Tier-2 Site



Developing for Service HA

- (Zoo)Keeper-based container HA system
 - Motivation
 - k8s on IPVLAN requires extra CNI plugins, complex configuration
 - dCache already uses ZooKeeper, but needs additional ETCD cluster
 - Full k8s may be unnecessary for simple HA → develop a lightweight program
 - Status
 - Testing container restart on host failure
 - Transitioning each containers into Pod to easy managing
 - Limitation
 - Bugs causing redundant container starts under network instability
 - No real-time communication between host and agent

The screenshot shows the 'GSDC HA 시스템 모니터' (GSDC HA System Monitor) interface. It features a navigation bar with '활성 서비스', '호스트 목록', '서비스 목록', '관리 작업', and 'Zookeeper 뷰어'. There are also buttons for 'Zookeeper 전용 페이지' and 'Podman 실행 설정 도움말'. The main content area is divided into three sections: 1. '서비스 생성/업데이트' (Service Create/Update): Includes a '신규 서비스 등록' (New Service Registration) form with fields for '신규 서비스 이름' (New Service Name) and '기존 서비스 선택' (Existing Service Selection) (currently 'patroni'). Below is a 'run_config' JSON snippet. 2. '서비스 삭제' (Service Delete): A form to delete the 'patroni' service. 3. '서비스 환경 변수 설정' (Service Environment Variable Settings): A table for setting environment variables for the 'patroni' service, with columns for '대상 서비스' (Target Service), '변수명' (Variable Name), and '값' (Value). Below this is a '신규 환경 변수 등록' (New Environment Variable Registration) form with an example '예) FOO=BAR BAZ' and a '일괄 추가' (Batch Add) button.

The screenshot shows the '서비스: patroni' (Service: patroni) details page. It is divided into three main sections: 1. '기본 정보' (Basic Information): A table listing '이미지' (Image: quay.io/ry840901/patroni:latest), '복제본' (Replicas: 1), '포트' (Ports: -), '볼륨' (Volumes: /etc/gsdc_ha_system/ca:/etc/gsdc_ha_system/ca), and '네트워크' (Network: storage_net). 2. '환경 변수' (Environment Variables): A table listing variables like 'PATRONI_SCOPE', 'PATRONI_ZOOKEEPER_HOSTS', 'PATRONI_ZOOKEEPER_USE_SSL', 'PATRONI_ZOOKEEPER_CACERT', and 'PATRONI_ZOOKEEPER_VERIFY'. 3. '원본 서비스 정의(JSON)' (Original Service Definition (JSON)): A '복사' (Copy) button and a JSON snippet showing the full service definition, including 'replicas', 'run_config', 'env', 'hostname', 'image', 'podman_network', 'podman_network_ips', and 'volumes'.

Summary

- Site Infrastructure
 - Core Services (CE, site-bdii, apel, vobox)
 - ALICE Tier-1 (T1): Built on an oVirt VM cluster
 - CMS Tier-2 (T2): Running on bare metal
 - Storage System
 - T1: EOS container instances using podman
 - T2: dCache Pool with JBOD and NFS storage
 - Worker Nodes (WNs)
 - Running on bare metal
- Current Status & Plans
 - Sites are operating stably
 - Resource replacement for WNs and storage planned by the end of the year
- Ongoing Work
 - Container Instances with IPVLAN Networking (T2)
 - Work in progress to utilize IPVLAN for container instances
 - Related software development progressing in parallel

Thank you

Backup

Network mode	latency	Throughput perf.	Info
Host (native)	Reference	Reference	No isolation
IPvlan	+30~60% ↑	<5% loss	Not necessary promisc, DHCP is not works
Macvlan	+40~70% ↑	<5% loss	Require promisc, DHCP is works
Overlay (VXLAN, etc)	+200~300% ↑	10~30% loss	encapsulation overhead is larghe

[Comparison of Networking Solutions for Kubernetes — Comparison of Networking Solutions for Kubernetes 2 documentation](#)

Latency percentiles at 150,000 RPS (≈30% of maximum RPS), ms

Setup	95 %ile	99 %ile	99.5 %ile	99.99 %ile	99.999 %ile	Max Latency
IPvlan	0.7	0.9	1	6.7	9.9	18
aws-vpc	0.9	1.1	1.2	6.5	9.8	15.7
host-gw	0.9	1.1	1.2	5.9	9.6	24.3
vxlan	1.2	1.5	1.6	6.6	201.9	405.3
--net=host	0.5	0.6	0.6	4.8	8.9	11.8

IPvlan is slightly better than `host-gw` and `aws-vpc`, but it has the worst 99.99 percentile. `host-gw` performs slightly better than `aws-vpc`.

Latency percentiles at 250,000 RPS (≈50% of maximum RPS), ms

Setup	95 %ile	99 %ile	99.5 %ile	99.99 %ile	99.999 %ile	Max Latency
IPvlan	1	1.2	1.4	6.3	10.1	24.3
aws-vpc	1.2	1.5	1.6	5.6	9.4	27.3
host-gw	1.1	1.4	1.6	8.6	11.2	40.1
vxlan	1.5	1.9	2.1	16.6	202.4	245.5
--net=host	0.7	0.8	0.9	3.7	7.7	16.8

IPvlan again shows the best performance, but `aws-vpc` has the best 99.99 and 99.999 percentiles. `host-gw` outperforms `aws-vpc` in 95 and 99 percentiles.

Latency percentiles at 50k RPS (≈20% of maximum RPS), ms

Setup	95 %ile	99 %ile	99.5 %ile	99.99 %ile	99.999 %ile	Max Latency
IPvlan	0.6	0.8	0.9	5.7	9.6	15.8
aws-vpc	0.7	0.9	1	5.6	9.8	403.1
host-gw	0.7	0.9	1	7.4	12	202.5
vxlan	0.8	1.1	1.2	5.7	201.5	402.5
--net=host	0.5	0.7	0.7	6.4	9.9	14.8

<https://ranger.uta.edu/~jrao/papers/INFOCOM18.pdf>

f