GSDC: Datacenter for Data-intensive Research

e-Science Activities in Korea

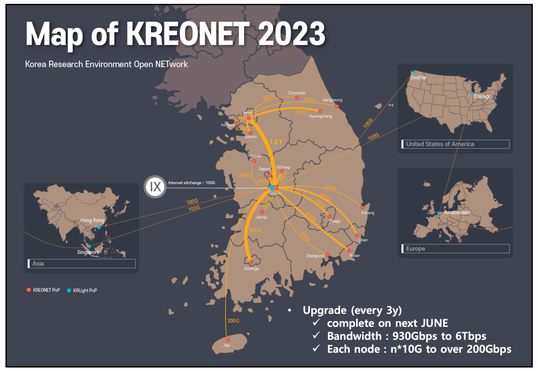
ISGC 2024

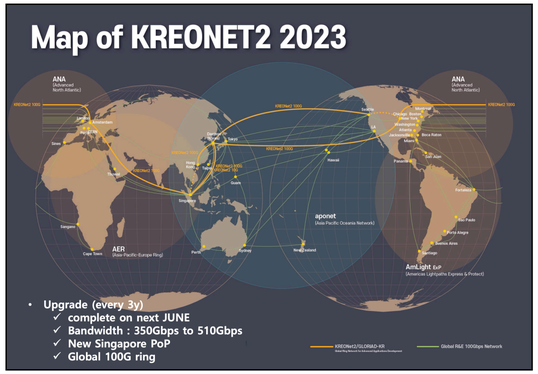
Ahn, Sang-Un

KISTI

Korea Institute of Science and Technology Information

- Government-funded research institute founded in 1961 for national information services and supercomputing
- National Supercomputing Center
  - **Nurion** - Cray CS500 system
    - 25.7 PFlops at peak, ranked 11th of Top500 (2018) ⇨ 46th (Nov 2022)
  - **Neuron** - GPU system, 1.24 PFlops
  - **KREONet/KREONet2** - National/International R&E network





GSDC

Global Science experimental Data hub Center

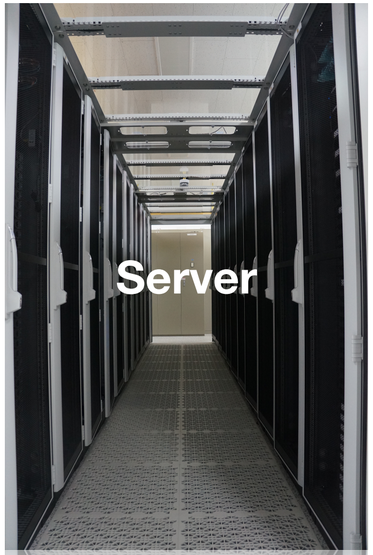
NETMIKO

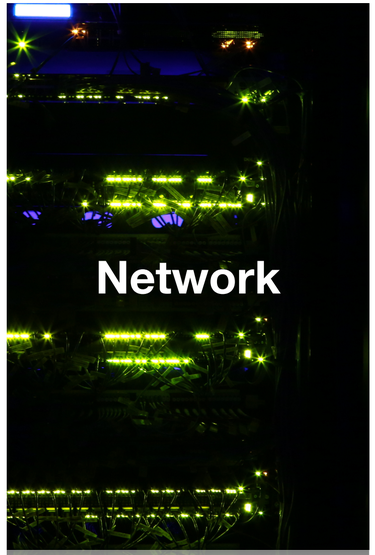
ANSIBLE

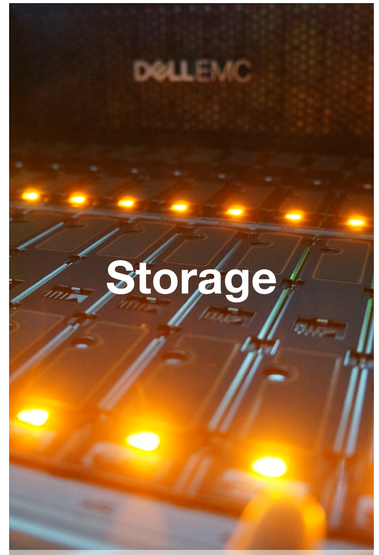
BIRD

puppet

- 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 bio-medical and simulation-R&D activities
  - Providing services based on technology development: distributed computing structure, high availability storage system, infra integrated management, disk-based custodial storage







Supporting Experiments

Heavy-Ion Physics
WLCG Tier-1 (2014)

Elementary Particle Physics
WLCG Tier-2 (2018)

Astrophysics
LDG Tier-2 (2019)

Elementary Particle Physics
B2G Tier-2 (TBD)

Medical Science
ICGC-ARGO Center

Neutrino Physics
RAW Storage

Structural Biology using Cryo-EM/XFEL
Data Storage & Analysis Pipeline

Role of GSDC for Data-intensive Research

KISTI-GSDC

Process

Constructing Facility → Conducting Experiment → Data Comp. Env. → Analyzing Data → Scientific Achievement

Subject

CERN(EU) | LIGO (US) | KEK (JP)
KBSI (KR) | PAL | RENO
Large-scale Facilities

Users | User Groups
Research Community

Data Computing for Basic Science as a Service

Requirements by experiments/community →
Data Computing Research Environments

APIs

System Auto -mation	Resources Provisioning	Data Management	Data Analysis	Monitor -ing
Identity /Policy Mgmt.	• Server (Container, VM, Baremetal) • Storage (Block, Object, File) • Network	• Transfer • Movement • Metadata • Access ctrl. • Archiving	• Batch • Parallel/Dist. • Streamline • In-memory • DeepLearning	

Data Center Infrastructure

Data Provider

Domestic | Overseas
Large-scale Facility

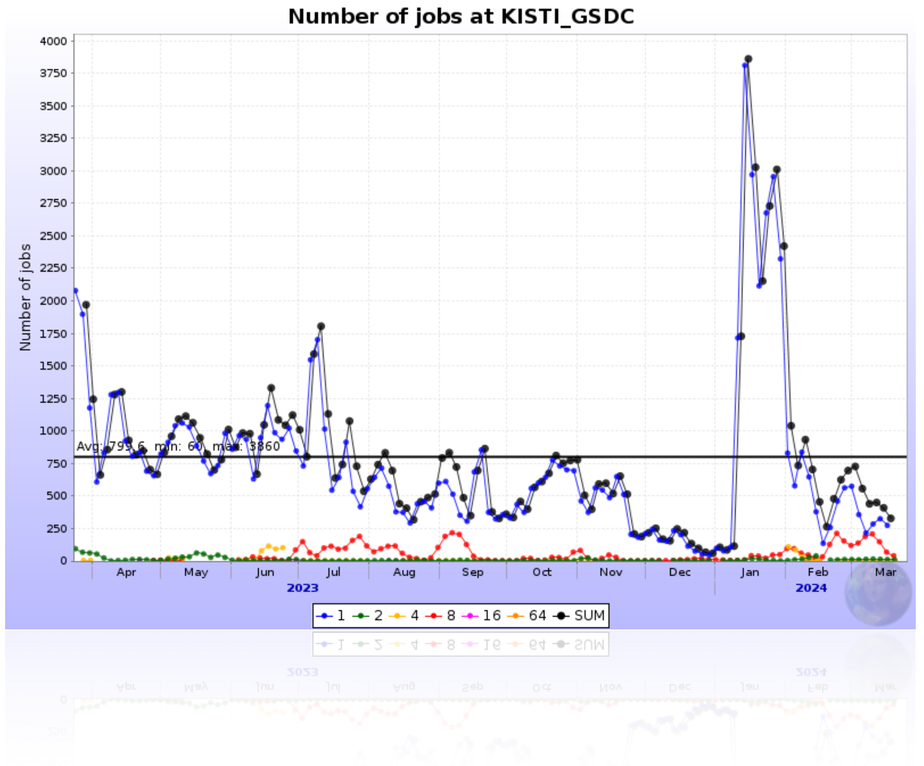
Data-Intensive Research

ALICE, CMS, LIGO, RENO, PAL 포항가속기연구소

WLCG Tier-1 @ KISTI-GSDC

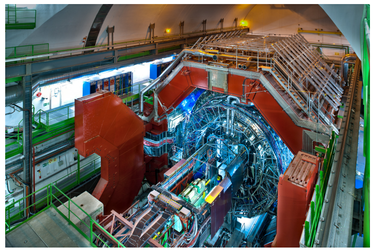
Flagship Service for Data-intensive Computing

- 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
- CE
  - HTCondor-based, whole-node submission enabled (for N-core jobs)
- SE
  - 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)

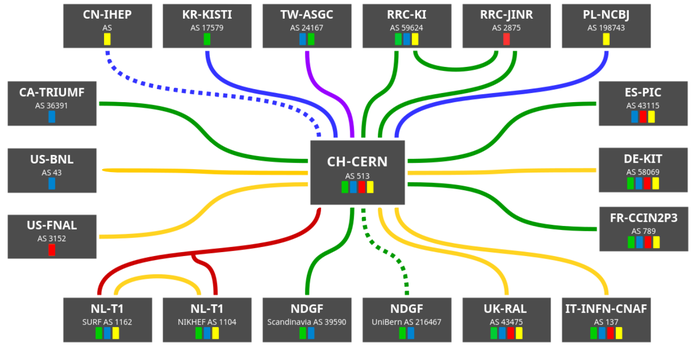


KISTI ALICE T1 Structure Overview

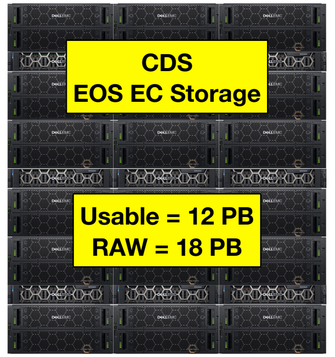
Simulation
Organised Analysis
Reconstruction
User Analysis
Job Queue
Job Broker
allocation
VO-Box
submission
CE
HTCondor-CE
scheduling
Batch Cluster
HTCondor
WLCG CRIC
publish
Accounting
APEL
statistics
Job
Data
CERN
HTCondor
7,000 Core
query
collect
AuthN/AuthZ
BDII/Argus
read only
data i/o
collect
collect
80 PB
Erasure-Coded storage
Detector
ALICE O2
80PB/yr
O2 Buffer
real-time data transfer (> 25Gb/s required)
Archival Storage
raw
Disk Storage
processed
archive
~ 1EB
T0



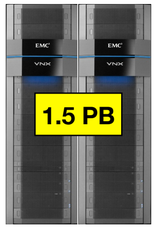
LHCOPN
KISTI-CERN 20Gbps



100Gpbs became a standard



JBOD Storage
Feb. 2024



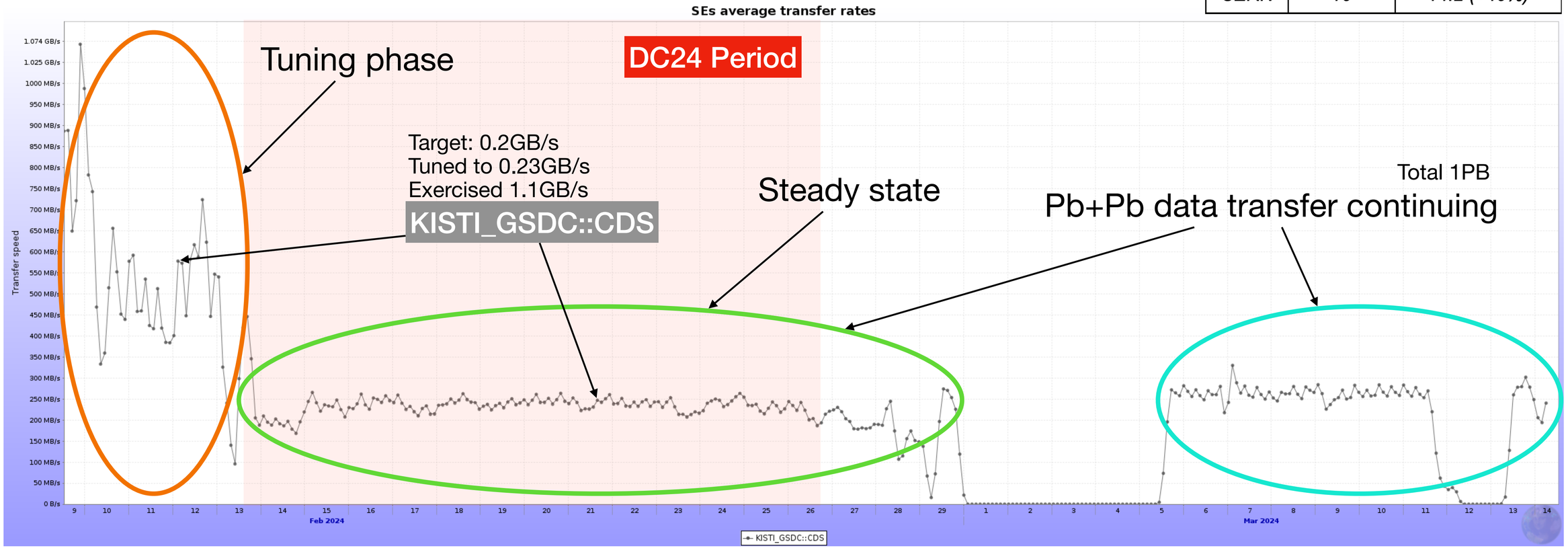
To be decommissioned
Later this year

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

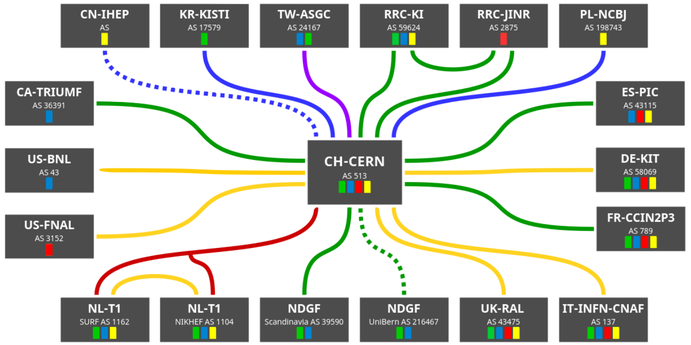
Centre	Target rate GB/s	Average achieved GB/s
CNAF	0.8	0.98 (+20%)
IN2P3	0.4	0.6 (+40%)
KISTI	0.2	0.25 (+22%)
GridKA	0.6	1.12 (+90%)
NDGF	0.3	0.35 (+15%)
NL-T1	0.1	0.25 (+150%)
RAL	0.1	0.58 (+500%)
CERN	*10*	*14.2 (+40%)*



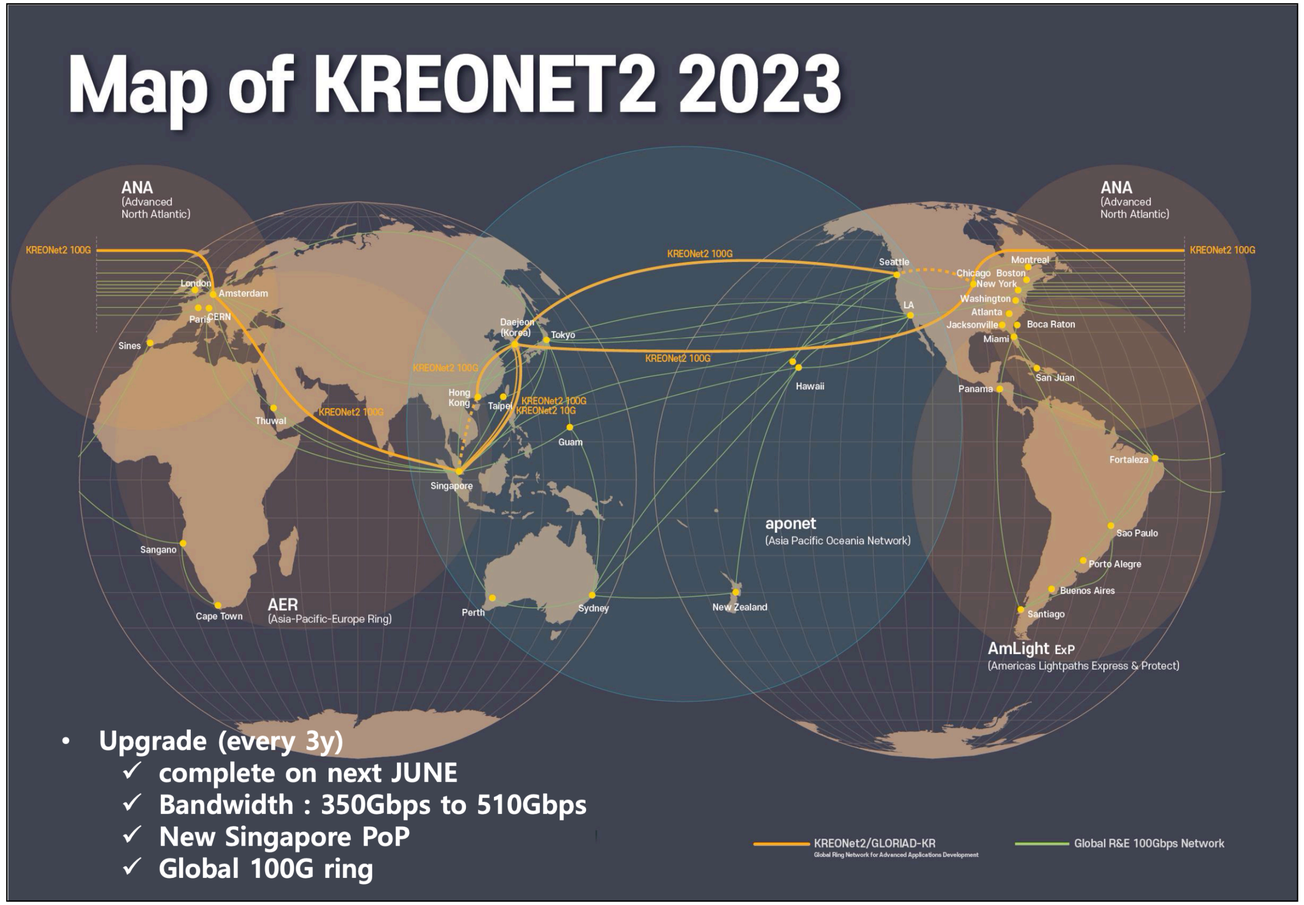
LHC Networking - OPN

Dedication to LHC Raw Data Transfer between T0 and T1s

LHCOPN

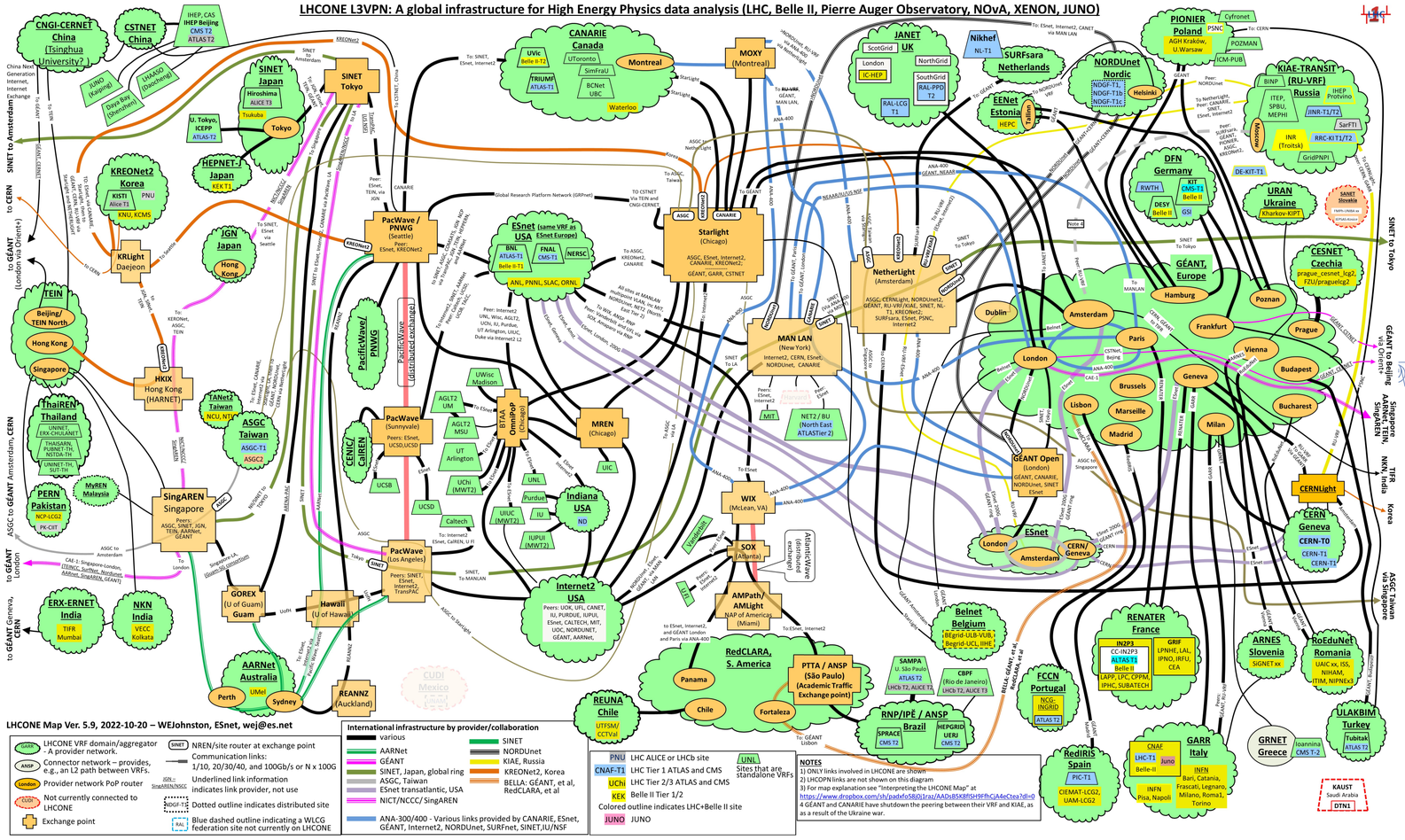


- 20Gbps dedicated links from Daejeon to Geneva provided by KREONet2 with its 100Gbps lambdas
- Primary optical fibers: Daejeon-Chicago-Amsterdam-Geneva (Backup links through Daejeon-Seattle & GLORIAD-consortium)
- KREONet2 directly reaches Geneva from Amsterdam PoP
- Provisioning of 100Gbps by end of LHC RUN3 or before the start of HL-LHC (RUN4)

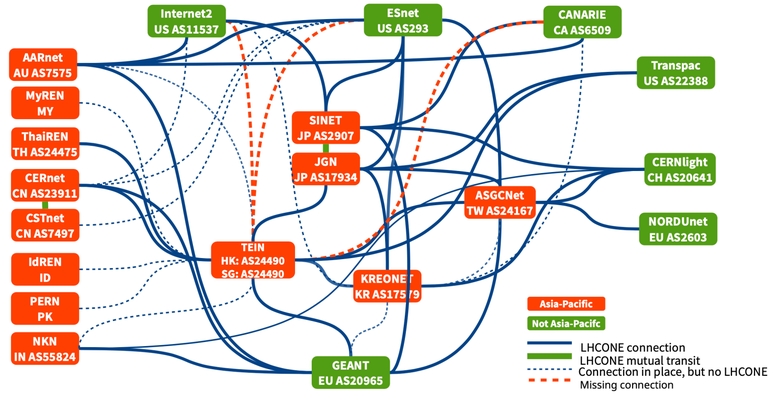


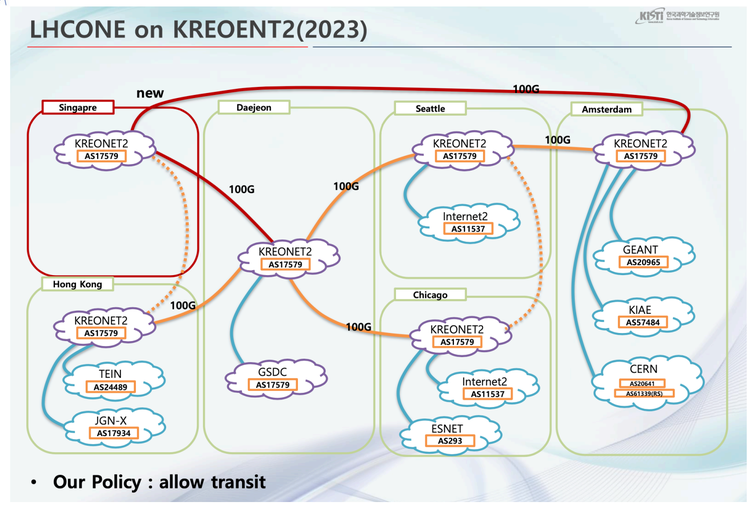
LHC Networking - ONE

Towards full mesh reachability among Tier sites for Big sciences



Asia-Pacific VRFs – Current Status



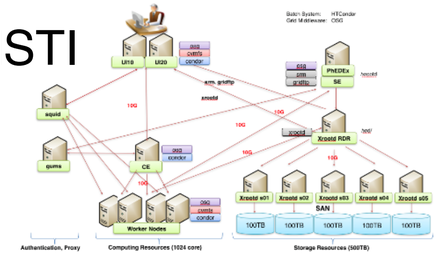


- Policy that allows transit via KREONet2 resolves missing connections in Asia-Pacific region

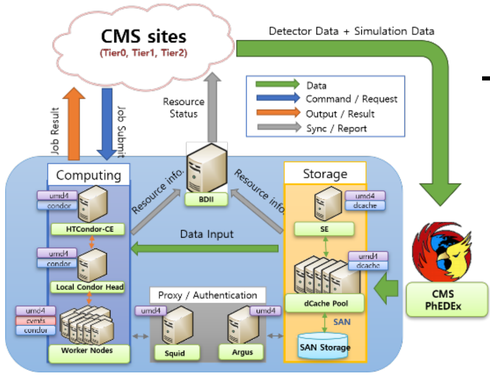
About KISTI CMS Tier-2

- KISTI CMS Tier-2
  - WLCG Tier-2 site for CMS experiment
  - KISTI CMS Tier-2 focuses on providing resources for CMS experiment rather than supporting domestic users
    - Due to the presence of separate CMS Tier-3 site (T3_KR_KISTI)
- CMS Tier-2 History
  - 2017 Mar. : Register as an EGI site (KR-KISTI-GSDC-02)
  - 2017 Aug. : Register as a CMS Site (T2_KR_KISTI)
  - 2017 Sep. : Enable CMS PhEDEx Link (Joining CMS Data Transfer system)
  - 2017 Nov. : Starting CMS T2 Testbed after passing the SAM test stably
  - 2018 Apr. : KISTI-CERN MOU Signing Ceremony for CMS Tier2

T3_KR_KISTI

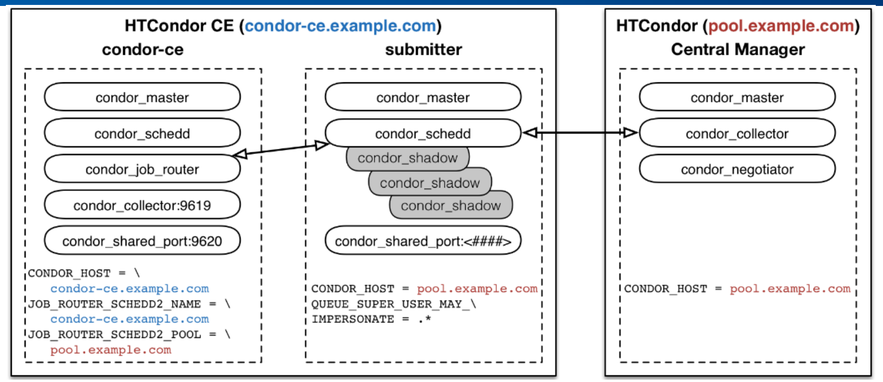


T2_KR_KISTI



T2_KR_KISTI Structure

- Main Component
  - CE : HTCondor-CE 5
    - LRMS : HTCondor 9
      - 1,424 logical cores
      - RAM 3,000MB per core
  - SE : dCache
    - 1 SAN + 1 JBOD + 9 NFS Pools / 1761TB
    - Protocol
      - XRootD, GridFTP(+SRM), pNFS, WebDAV
  - Etc.
    - Report: Site-BDII, APEL
    - Cache : Frontier-Squid
    - CMS AAA
      - 1x Standalone XRootD Server (Forward 1095 ->1094)



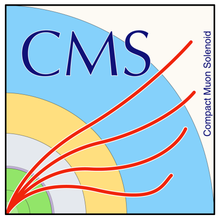
Gridftp WebDaV XRootD +pNFS

Pool Request Queues

CellName	DomainName	Movers			Restores			Stores			P2P-Server			P2P-Client			qsrm_ftp			qsrm_webdav			regular		
		Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued
	Total	632	12620	0	0		0	0		0	0	120	0	0			2	220	0	51	1300	0	579	11100	0
SAMPool	dCacheDomain	0	100	0	0		0	0		0	0	10	0	0			-1	-1		-1	-1		0	100	0
cms-t2-wn1055-NFSPool	cms-t2-wn1055-NFSPool-Domain	53	1120	0	0		0	0		0	0	10	0	0			0	20	0	6	100	0	67	1000	0
cms-t2-wn1055-SANPool	cms-t2-wn1055-SANPool-Domain	79	1120	0	0		0	0		0	0	10	0	0			0	20	0	5	100	0	74	1000	0
cms-t2-wn1056-JbodPool	cms-t2-wn1056-JbodPool-Domain	57	1120	0	0		0	0		0	0	10	0	0			0	20	0	4	100	0	53	1000	0
cms-t2-wn1056-NFSPool	cms-t2-wn1056-NFSPool-Domain	89	1120	0	0		0	0		0	0	10	0	0			0	20	0	4	100	0	85	1000	0
cms-t2-wn1057-NFSPool	cms-t2-wn1057-NFSPool-Domain	47	1120	0	0		0	0		0	0	10	0	0			0	20	0	7	100	0	60	1000	0
cms-t2-wn1058-NFSPool	cms-t2-wn1058-NFSPool-Domain	27	1120	0	0		0	0		0	0	10	0	0			0	20	0	3	100	0	24	1000	0
cms-t2-wn1059-NFSPool	cms-t2-wn1059-NFSPool-Domain	42	1120	0	0		0	0		0	0	10	0	0			1	20	0	4	100	0	37	1000	0
cms-t2-wn1060-NFSPool	cms-t2-wn1060-NFSPool-Domain	58	1120	0	0		0	0		0	0	10	0	0			0	20	0	4	100	0	54	1000	0
cms-t2-wn1061-NFSPool	cms-t2-wn1061-NFSPool-Domain	78	1120	0	0		0	0		0	0	10	0	0			0	20	0	4	100	0	74	1000	0
cms-t2-wn1062-NFSPool	cms-t2-wn1062-NFSPool-Domain	38	1720	0	0		0	0		0	0	10	0	0			1	20	0	7	700	0	30	1000	0
cms-t2-wn1063-NFSPool	cms-t2-wn1063-NFSPool-Domain	64	1720	0	0		0	0		0	0	10	0	0			0	20	0	3	700	0	61	1000	0
	Total	632	12620	0	0		0	0		0	0	120	0	0			2	220	0	51	1300	0	579	11100	0
CellName	DomainName	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued
		Movers			Restores			Stores			P2P-Server			P2P-Client			qsrm_ftp			qsrm_webdav			regular		

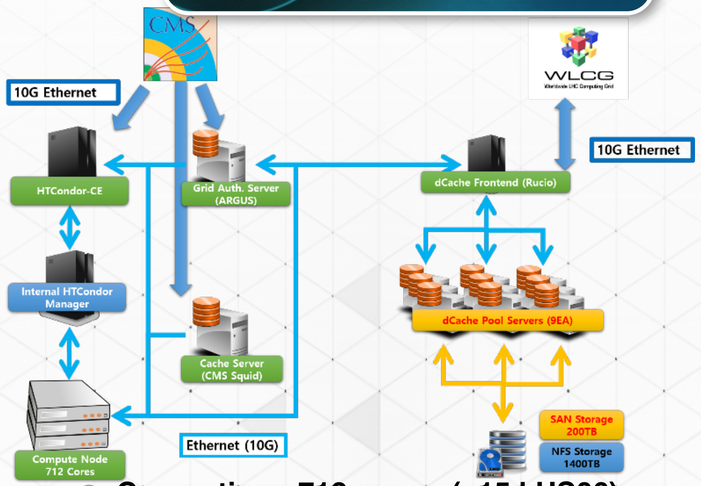
Disk Space Usage

CellName	DomainName	Total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/sticky/other free)
SAMPool	dCacheDomain	20437	2235	0	
cms-t2-wn1055-NFSPool	cms-t2-wn1055-NFSPool-Domain	153411227	17300985	0	
cms-t2-wn1055-SANPool	cms-t2-wn1055-SANPool-Domain	209700851	19690480	0	
cms-t2-wn1056-JbodPool	cms-t2-wn1056-JbodPool-Domain	209701127	46015112	0	
cms-t2-wn1056-NFSPool	cms-t2-wn1056-NFSPool-Domain	156237393	27518364	0	
cms-t2-wn1057-NFSPool	cms-t2-wn1057-NFSPool-Domain	155410193	24222341	0	
cms-t2-wn1058-NFSPool	cms-t2-wn1058-NFSPool-Domain	157334211	31456040	0	
cms-t2-wn1059-NFSPool	cms-t2-wn1059-NFSPool-Domain	153104766	17118567	0	
cms-t2-wn1060-NFSPool	cms-t2-wn1060-NFSPool-Domain	156306536	24808777	0	
cms-t2-wn1061-NFSPool	cms-t2-wn1061-NFSPool-Domain	153410384	17472478	0	
cms-t2-wn1062-NFSPool	cms-t2-wn1062-NFSPool-Domain	165907738	63526508	0	
cms-t2-wn1063-NFSPool	cms-t2-wn1063-NFSPool-Domain	161830347	48410457	0	



KISTI CMS Tier-2 Site Report

CMS Tier-2 Infrastructure



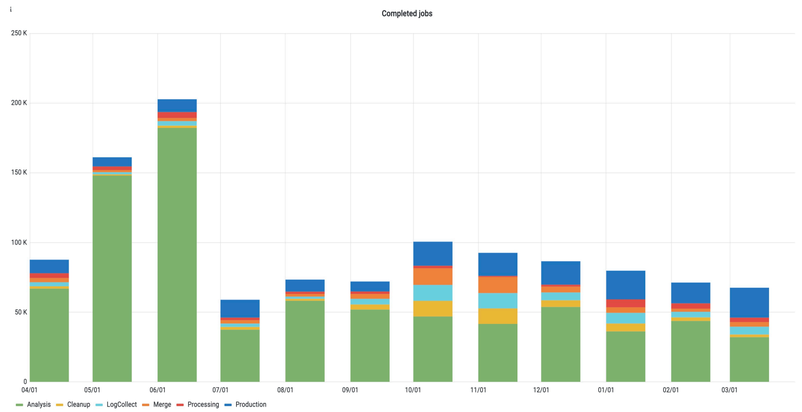
- Computing : 712 cores (~15 kHS06)

Storage Usage

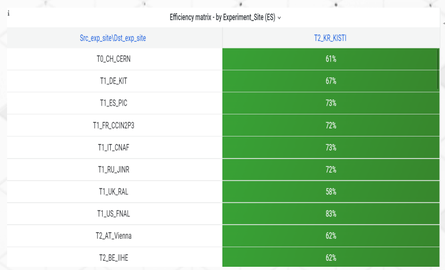
- Disk 1,761 TB (Usage 75.70%)

Job Activities

~1.15 million jobs during this year



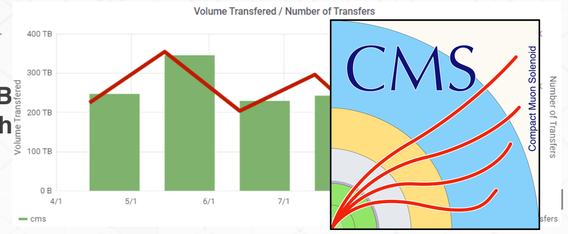
Data Transmission



- KISTI Tier-2 Data Link
  - Tier-0 link : 1
  - Tier-1 link : 7
  - Tier-2 link : 46
  - Tier-3 link : 5

- Data Traffic

Total : 681TB
Average : 68TB /month



WLCG Tier-2 Service Availability / Reliability

	Reliability	Availability
	Overall in 2023	Overall in 2023
CMS	93.51%	94.35%

Monthly target of WLCG : 95%

CMS Tier-2 Availability/Reliability

Site	Availability	Reliability ↓
T2_FR_GRIF_IRFU	94.20%	99.11%
T2_RU_JINR	98.74%	98.74%
T2_DE_DESY	98.69%	98.69%
T2_HU_Budapest	98.39%	98.56%
T2_IT_Legnaro	98.08%	98.53%
T2_DE_RWTH	97.89%	98.41%
T2_UK_London_IC	98.40%	98.40%
T2_FI_HIP	98.33%	98.38%
T2_US_Wisconsin	98.14%	98.14%
T2_KR_KISTI	97.65%	97.68%
T2_US_Caltech	97.42%	97.67%
T2_CH_CERN	97.59%	97.59%
T2_PT_NCG_Lisbon	96.86%	97.42%
T2_FR_GRIF_LLR	97.40%	97.42%
T2_UK_London_Brunel	97.20%	97.23%

GSDC-LDG

- GSDC-LDG (LIGO Data Grid), a gravitational wave data analysis computing environment at the request of the Korea Gravitational Wave Research Foundation (KGWG) in 2010.

- In 2019, the International Gravitational-Wave Observatory Network (IGWN) computing environment was established.

- Currently, the GSDC-LDG system operates as an integrated system that can be used simultaneously by global and domestic users.

History

GSDC-LDG Service Start

2010
- 24 core, 5TB
- HTCondor

2011
- 80 core, 5TB
- PBS

2012
- 200 core, 100TB
- HTCondor

2013
- 420 core, 150TB
- HTCondor

2014
- 420 core, 150TB
- HTCondor

2015
- 636 core, 150TB
- HTCondor
- Puppet, forman

2016
- 864 core, 340TB
- HTCondor

2017
- 864 core, 550TB
- HTCondor

2018
- 996 core, 550TB
- HTCondor
- Added RAM (minimum 72GB)

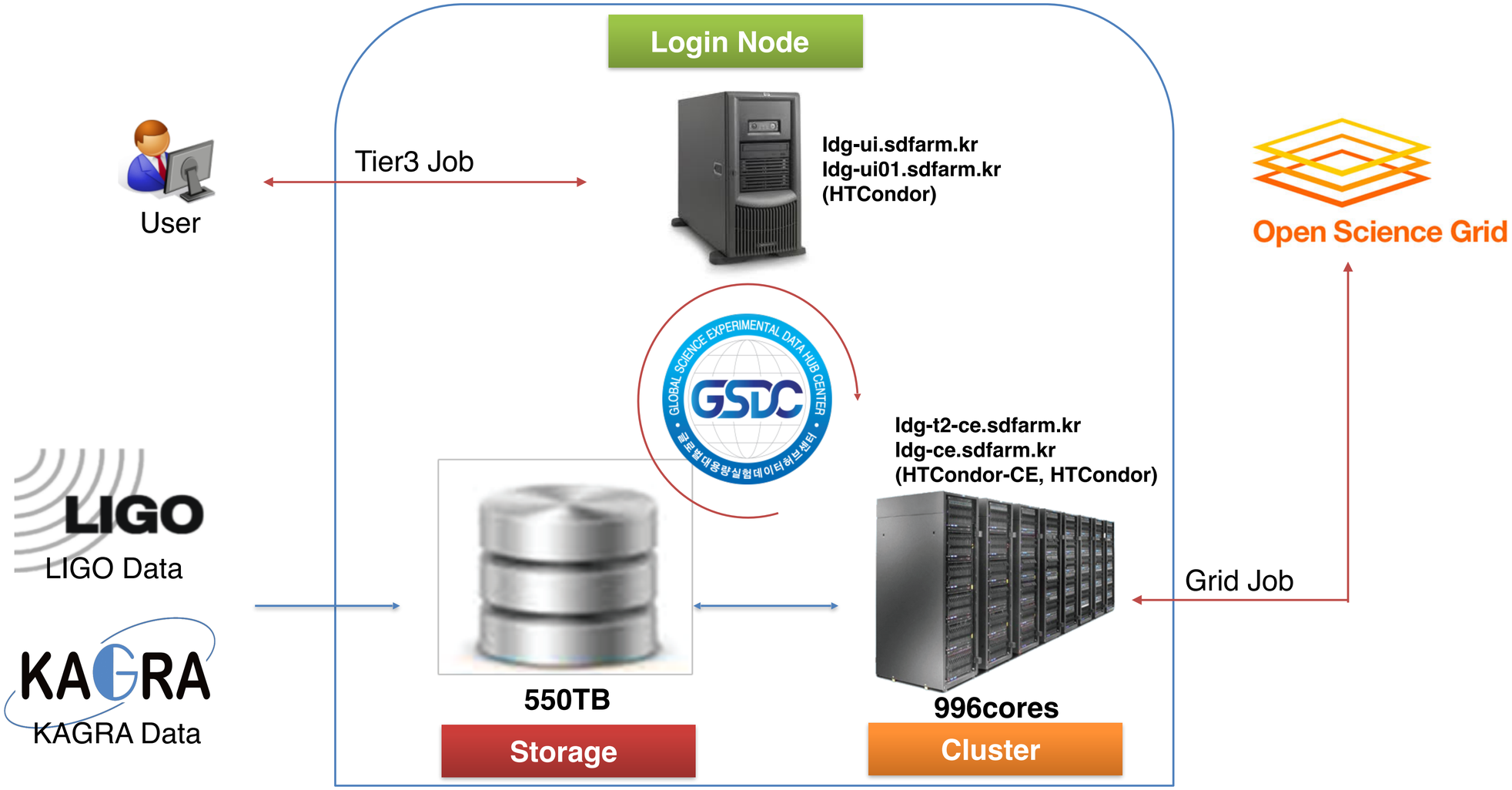
2019
- 996 core, 550TB
- Enabled Tier2 Service
- HTCondor, HTCondor CE
- Added 3 GPU Server (with P40 3 GPU Cards)

2020
- 996 core, 550TB
- HTCondor
- Added 3 GPU cards

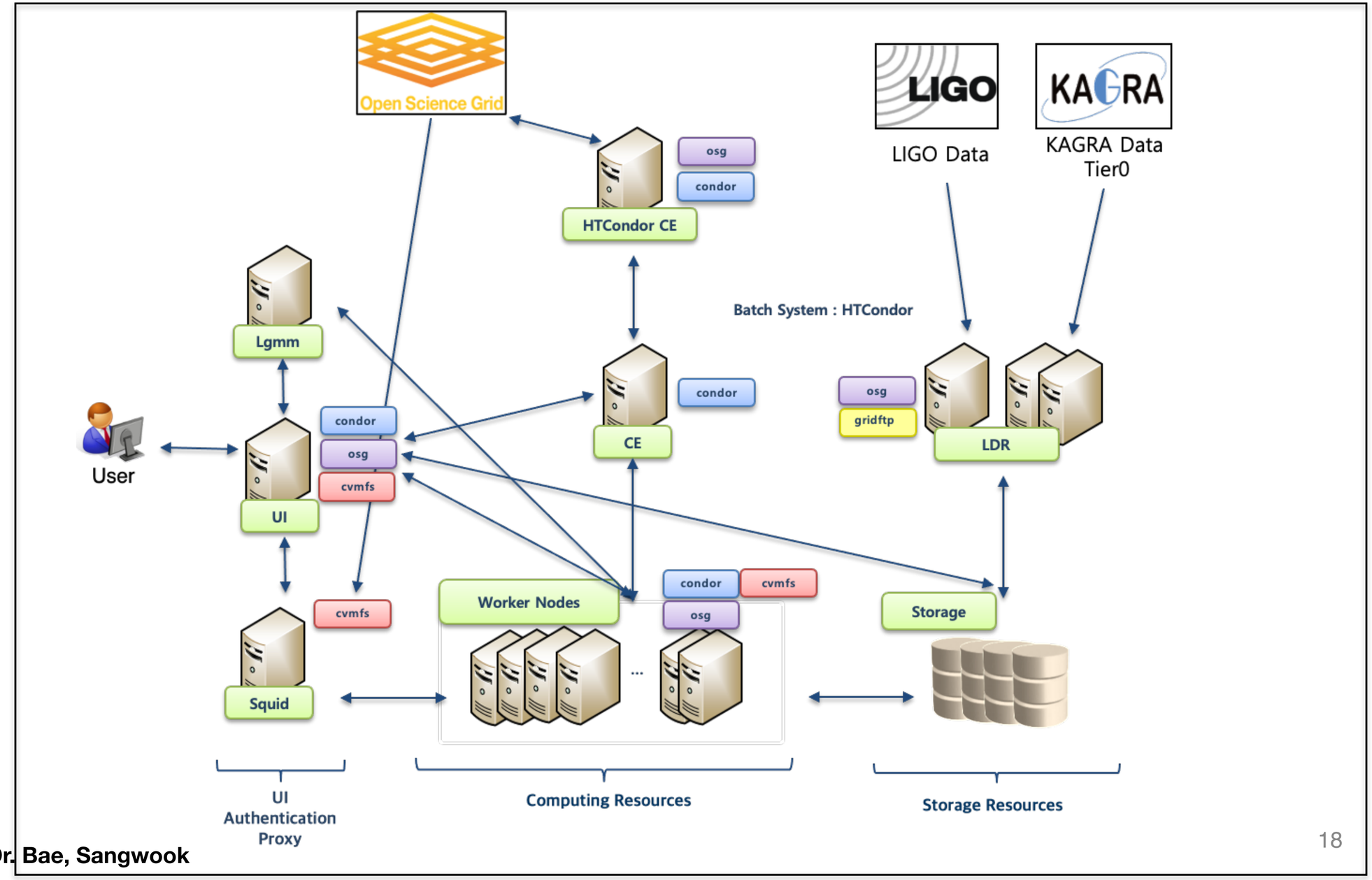
2021
- 996 core, 550TB
- HTCondor

IGWN | Computing

GSDC-LDG Overview



Status of GSDC-LDG



Resources

- Computation Resource

	Physical Core	Memory
Work Node	996 (66 servers)	72GB X 27 96 GB X 33 384 GB X 6
UI,CE,LGM,LDAS,LDR	60 (5 servers)	24GB X 5
Total	**1056**	**7416**

Work Node (GPU)	3 Servers	6 GPU Cards (P40)

- Storage Resources

	Mount on	Size	Used	Avail	Use	Total
LIGO	/data/ligo/	400T	250T	151T	63%	pool0.gsn.sdfarm.kr:/ifs/service/ligo
KAGRA	/data/kagra/	150T	76T	75T	51%	pool0.gsn.sdfarm.kr:/ifs/service/kagra

Available CPU Resources

	HOST NAME	H/W	MIDDLEWARE	OS
CE	belle-ce2.sdfarm.kr	Dell R640	HTCondor-ce	CentOS 7.9
WN	belle-wn[2201~2206].sdfarm.kr		HTCondor	

- **New system has been in operation since July 2022.**

- **Spec. of WN**
  - **# of WN : 6 nodes**
  - **# of cpu/node : 2**
  - **# of core/cpu : 18**
  - **HyperThread ON**
  - **# of Logical core/node : 72**
  - **Total jobslots : 432**
  - **Memory size per node : 384GB**
  - **Memory size per job slot ≈ 5.3GB**
  - **Disk space per job slot : 10GB**

- **HEP-SPEC06**
  - **CPU : Intel Xeon Gold 6245 @3.10GHZ**
  - **HS06/node : 942.76**
  - **TOTAL : 5.7K HEP-SPEC06**

Dr. Yeo, Ilyeon

DEDICATED STORAGE

- **SE Status**
  - **OLD System(Operating) : 100 TB SAN Storage**

	HOST NAME	H/W	MIDDLEWARE	OS
Head Node	belle-se-head.sdfarm.kr	Dell R610	DPM	SL 6.6
Disk Node	belle-se-disk01.sdfarm.kr			

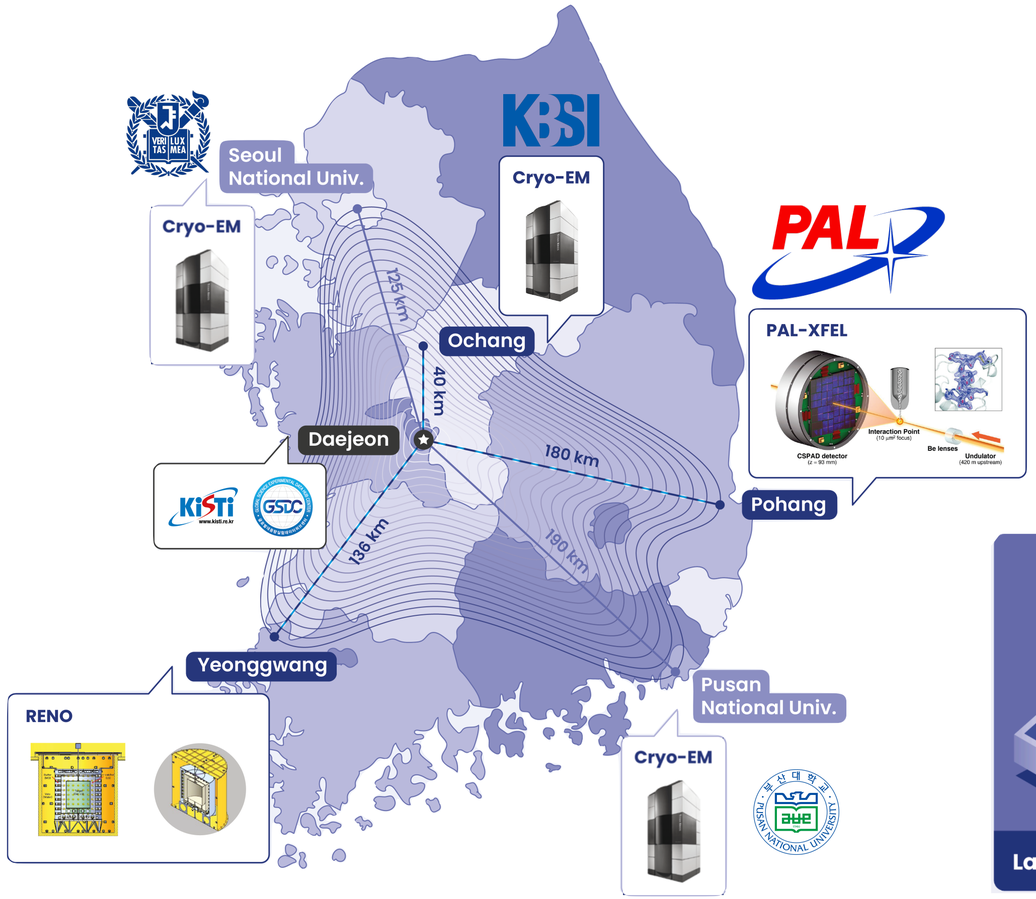
  - **NEW System(will be introduced by May) : 100TB NAS Storage**

	HOST NAME	H/W	MIDDLEWARE	OS
Head Node	belle-se2-head.sdfarm.kr	Dell R640	dCache	CentOS 7.9
Disk Node	belle-se2-disk01.sdfarm.kr			
	belle-se2-disk02.sdfarm.kr			

Dr. Yeo, Ilyeon

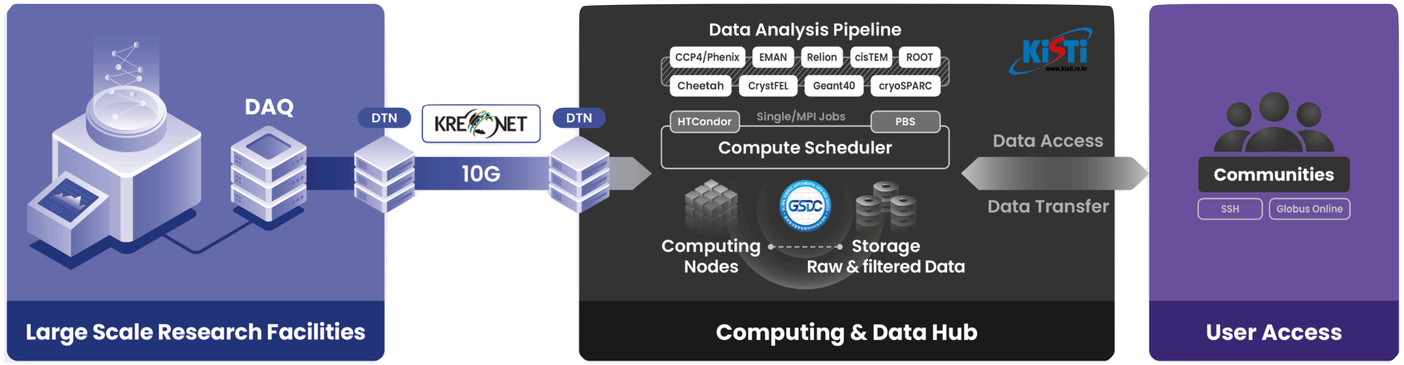
Supporting Domestic Research

Providing data storage, analysis pipeline and access



- Adapting the knowledge learned from operating Grid facilities to domestic region
- Dedicated optical links provided by KREONet for efficient data transfer and sharing
  - No need to move data by using external drives and overseas delivery
- Data analysis pipeline running on compute clusters
  - No need to own and maintain private cluster at individual labs
- User access to data and analysis pipeline without geographical constraints

➡ **Significant reduction of time in research activities**



Dr. Bae, Sangwook, Dr. Na, Sangho & Dr. Yu, Junglok

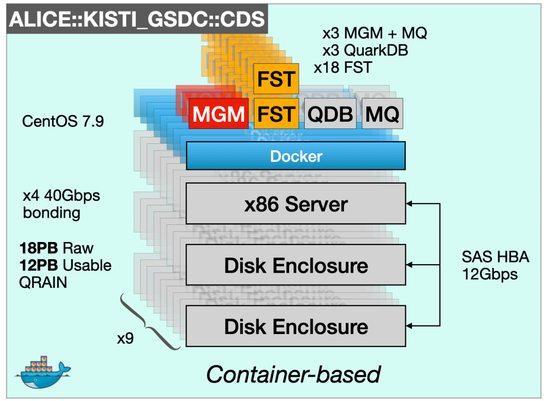
Thank you

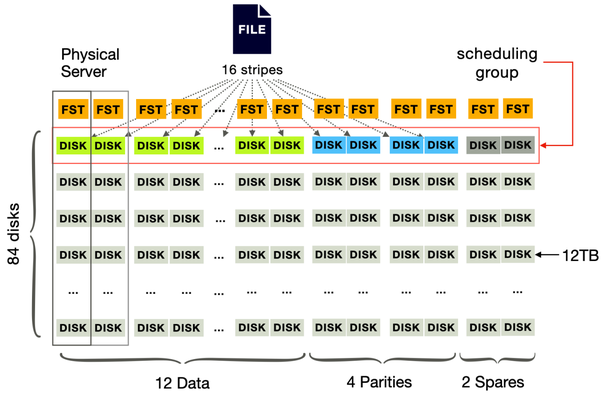
Fully Containerized EOS Storage

Automated Deployment via Ansible Playbook

- Targeting to Podman(EL9)/Docker(EL7) Container Runtimes
- Systemd controlled Container operation (run|rm) with parameters

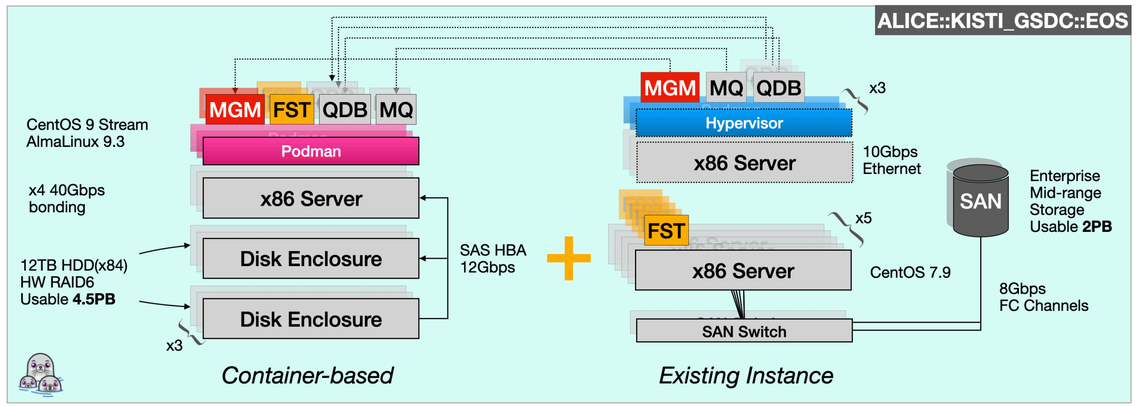
Custodial Storage





- 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)

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