KISTI-GSDC Report

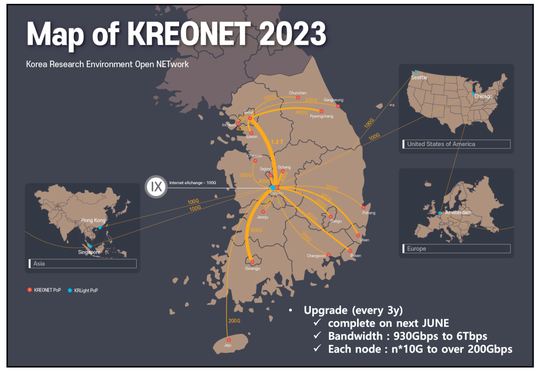
Geonmo Ryu, Sang-Un Ahn, Sangwook Bae
On behalf of KISTI-GSDC

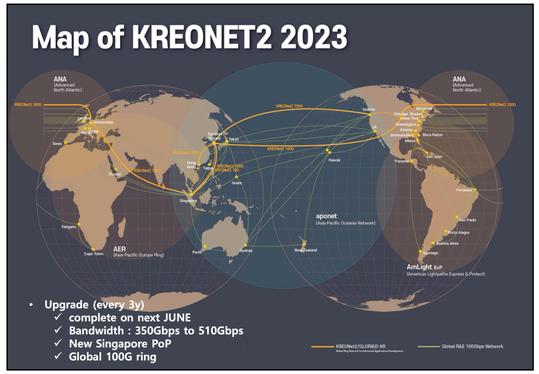
1-3 November 2023 @ ATCF7

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

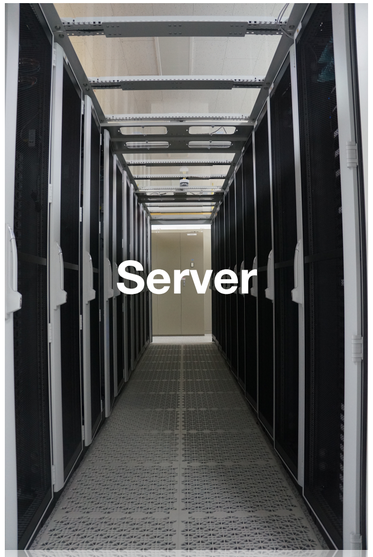


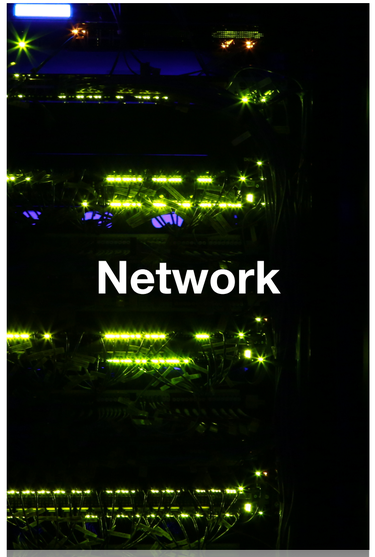


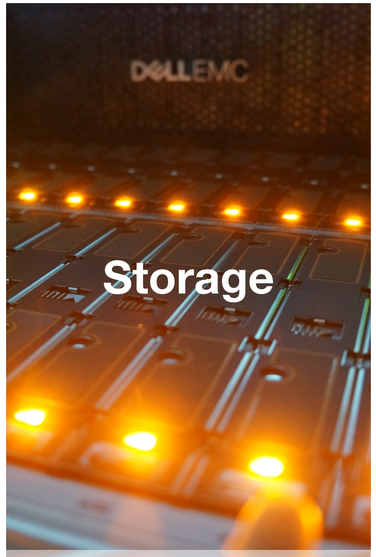
GSDC

Global Science experimental Data hub Center

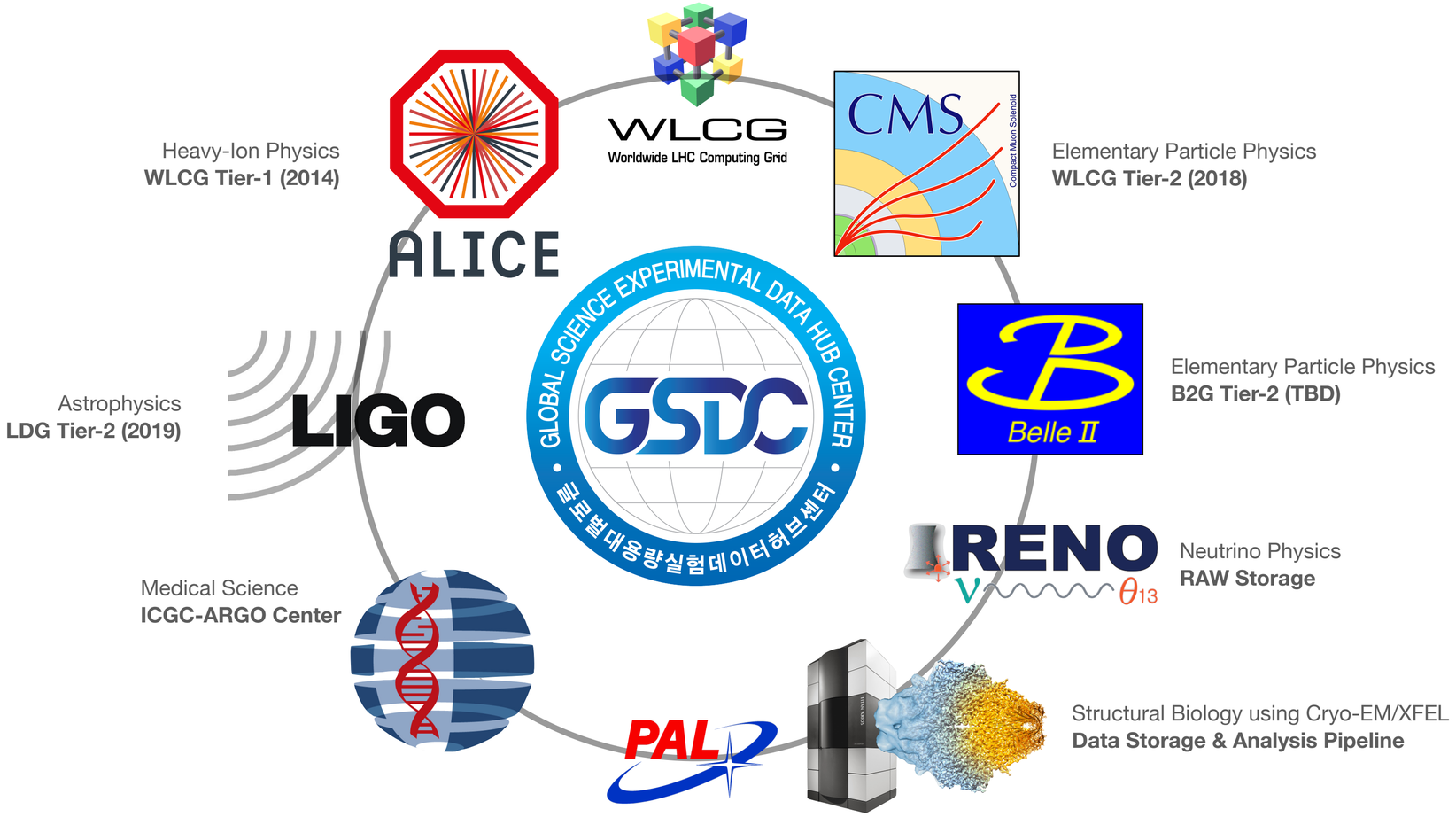
- Government-funded project, started in 2009 to promote Korean fundamental research through providing computing power and data storage
- **Datacenter for data-intensive fundamental research**
  - Preserving data from domestic or overseas large and complex scientific instruments as well as bio-medical 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



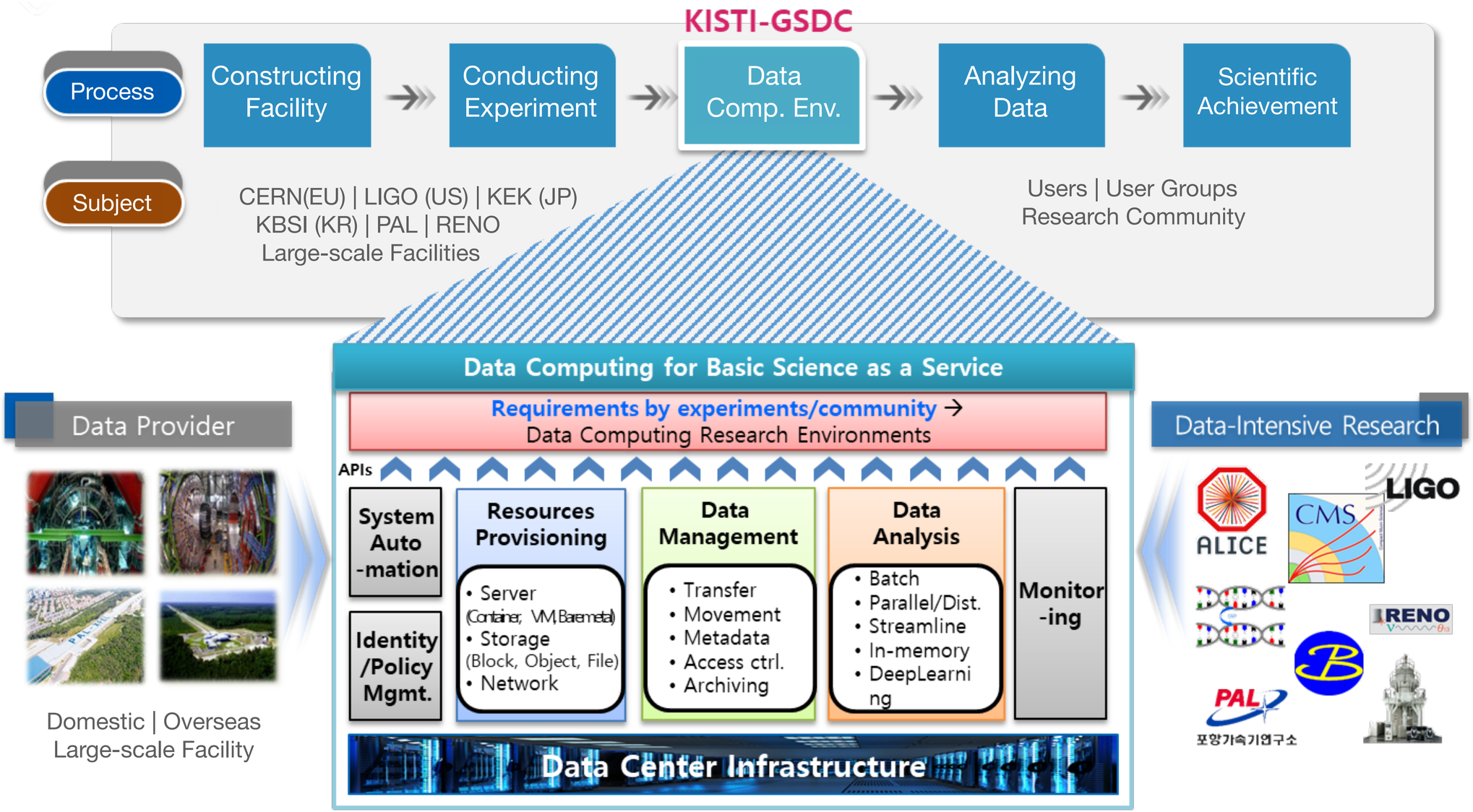




Supporting Experiments



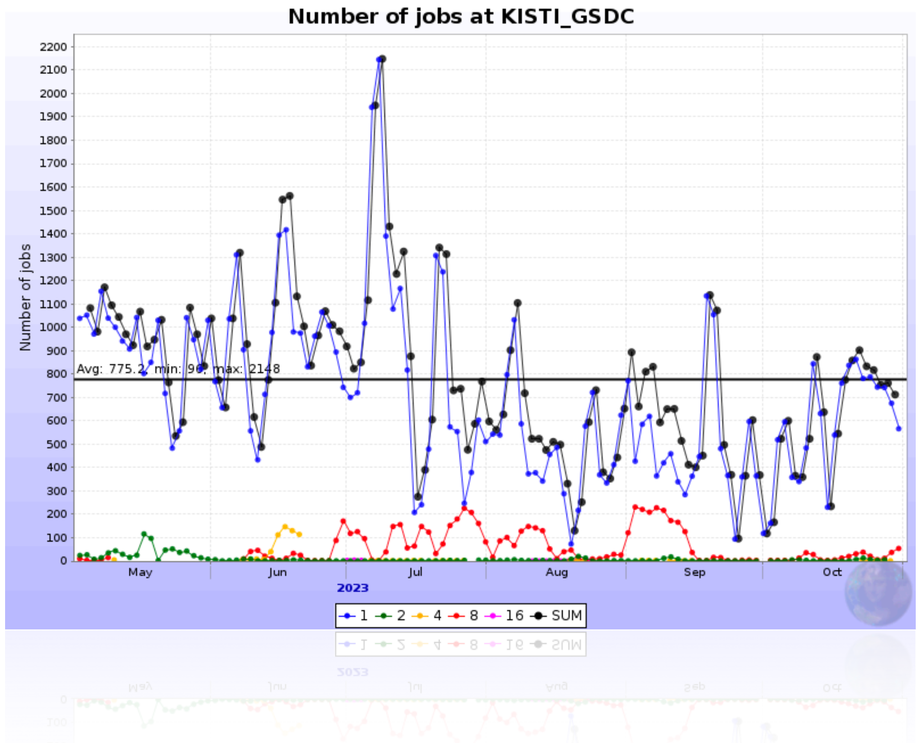
Role of GSDC for Data-intensive Research



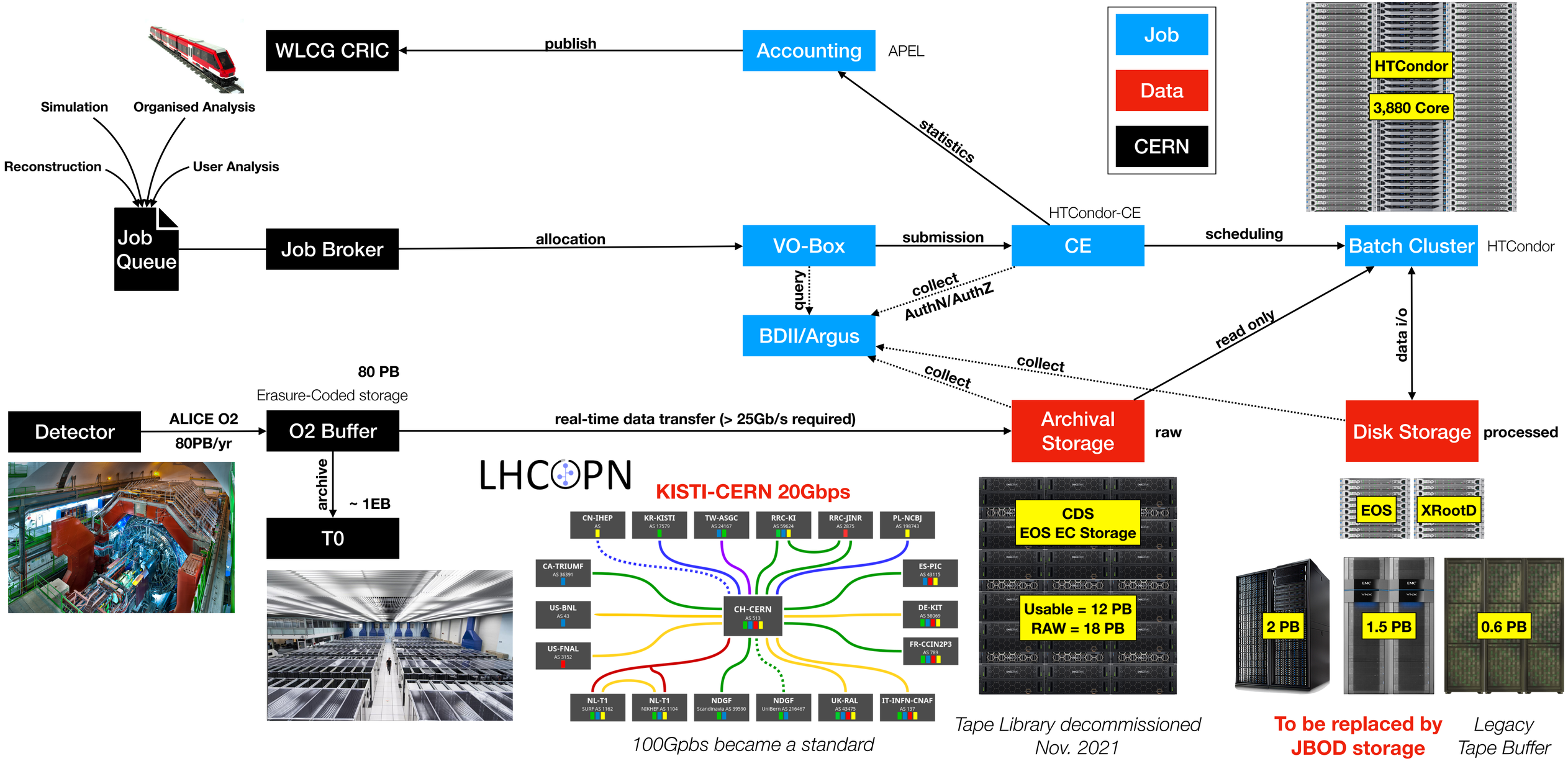
WLCG Tier-1 @ KISTI-GSDC

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
  - 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



100Gpbs became a standard

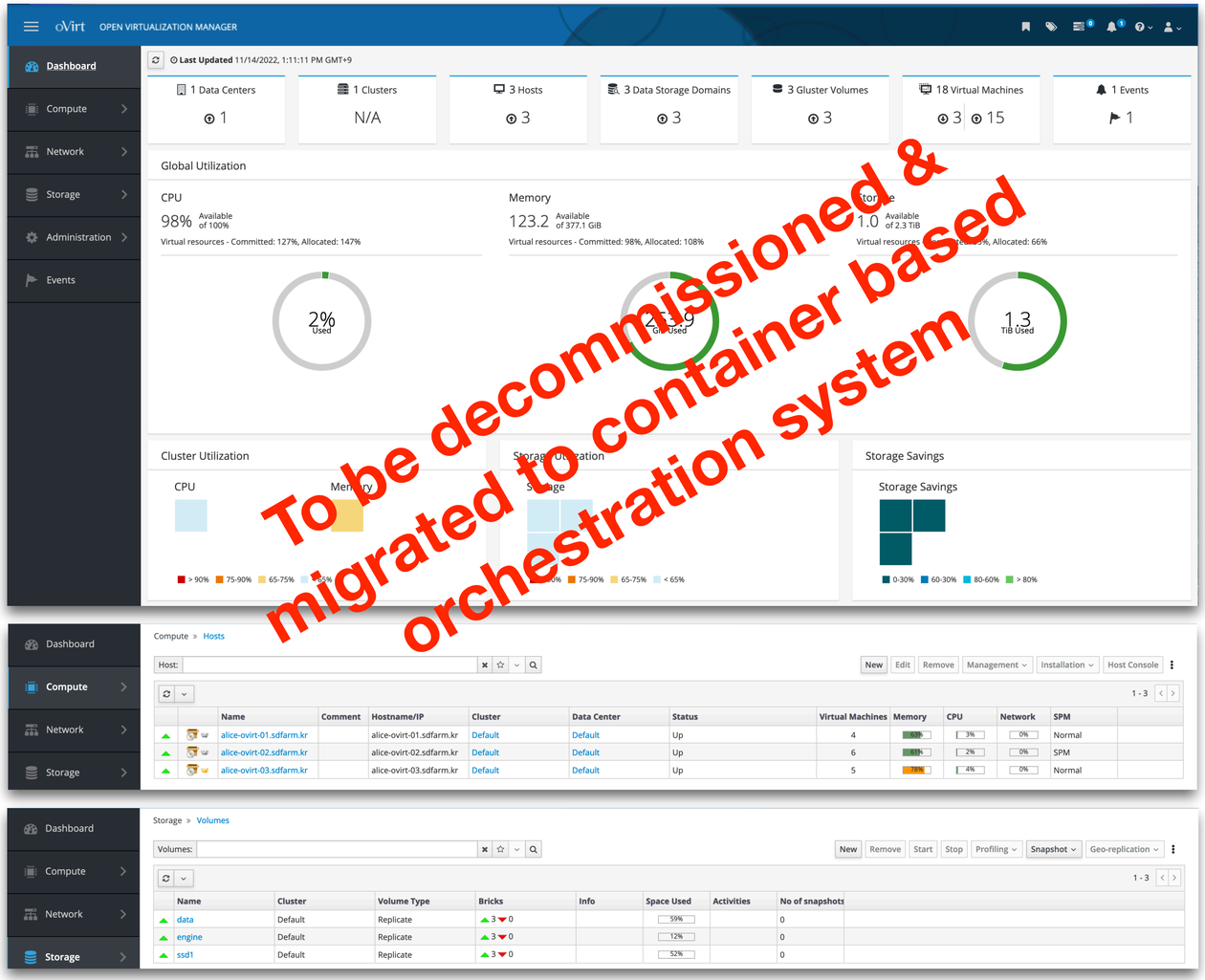
Tape Library decommissioned Nov. 2021

To be replaced by JBOD storage

Legacy Tape Buffer

T1 Grid Services

- 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
  - VO-Box (ALICE Job Submission, JAliEn enabled)
  - 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 5.1.5, Condor 9.0.14)
  - EOS MGM nodes & XRootD redirectors
  - EOS QuarkDB clusters (deployed upon SSD disk groups)



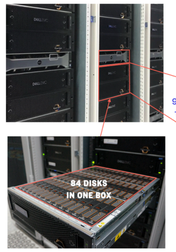
CDS in one slide

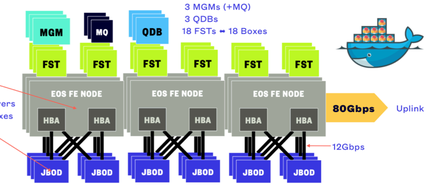
Custodial Disk Storage

Tapeless Archiving

EC Layout using 4 parity nodes

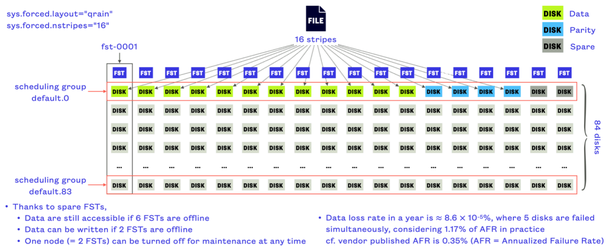
- The first disk-based custodial storage replaced tape for ALICE experiment
- 12 PB usable space with 12+4 erasure coding for data protection (powered by CERN EOS)
- Fully automated deployment of EOS components using Linux containers





- 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

System Architecture

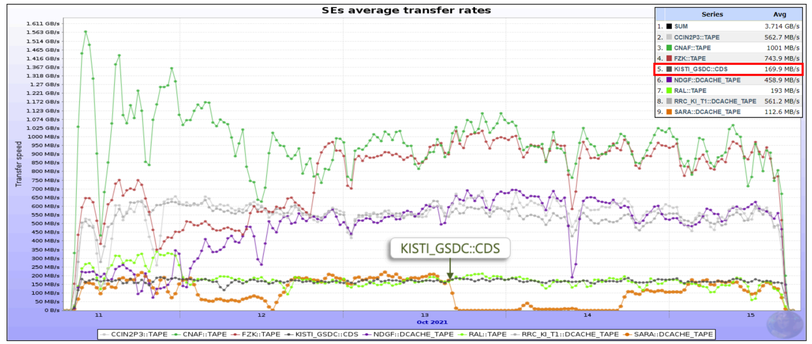


QRAIN(12+4) Layout

WLCG Data Challenges (Oct 2021)

Preparation for LHC RUN3 raw data transfer

- 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 Operation for ALICE

Fully commissioned since Nov 2021

Significant but endurable EC induced traffic observed

Current snapshot of the CDS in the ALICE monitoring system

http://alimonitor.cern.ch/stats?page=SE/table

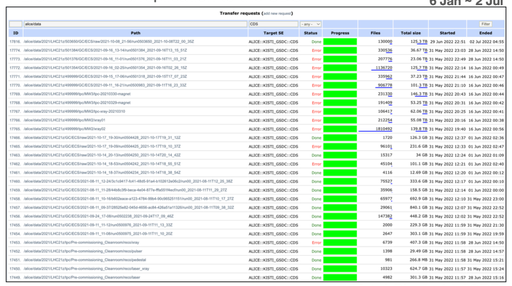
Custodial storage elements

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	Demotion factor	IPv6 add
CDS																							
Catalogue statistics									Storage-provided information						Functional tests					Last day add tests			
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

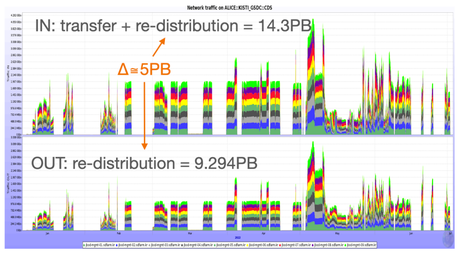
ALICE RAW data replication to the CDS

6 Jan ~ 2 Jul



[Total Size]=4.728PB

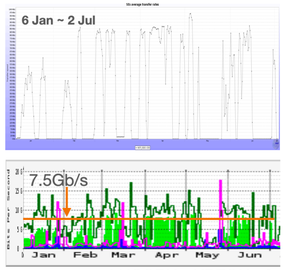
Peak traffic IN + OUT = 4.172GB/s + 3.218GB/s = 7.39GB/s ≈ 60Gb/s



Re-distribution Traffic induced by EC

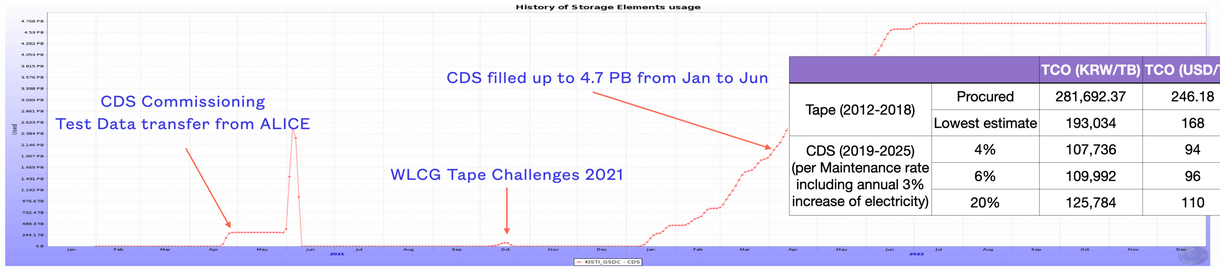
6 Jan ~ 2 Jul

Average transfer rate = 328MB/s

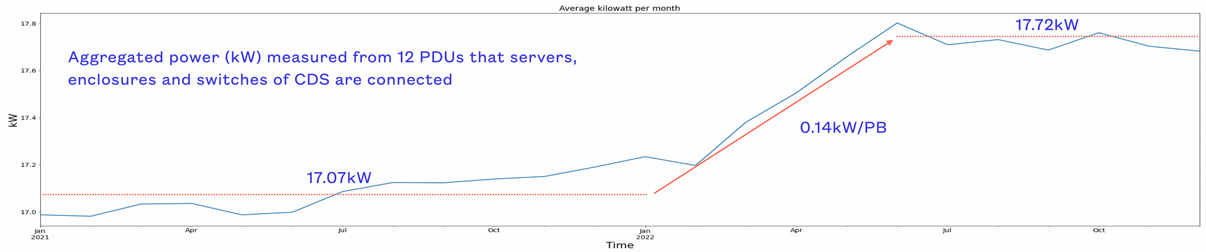


LHCOPN - KREONet2

CDS Power Consumption (2021-2022)



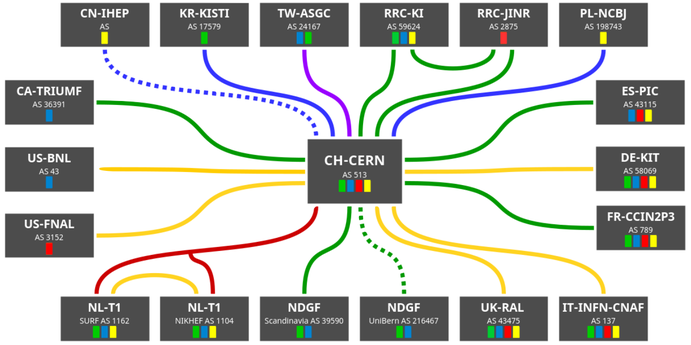
		TCO (KRW/TB)	TCO (USD/TB)
Tape (2012-2018)	Procured	281,692.37	246.18
	Lowest estimate	193,034	168
CDS (2019-2025) (per Maintenance rate including annual 3% increase of electricity)	4%	107,736	94
	6%	109,992	96
	20%	125,784	110



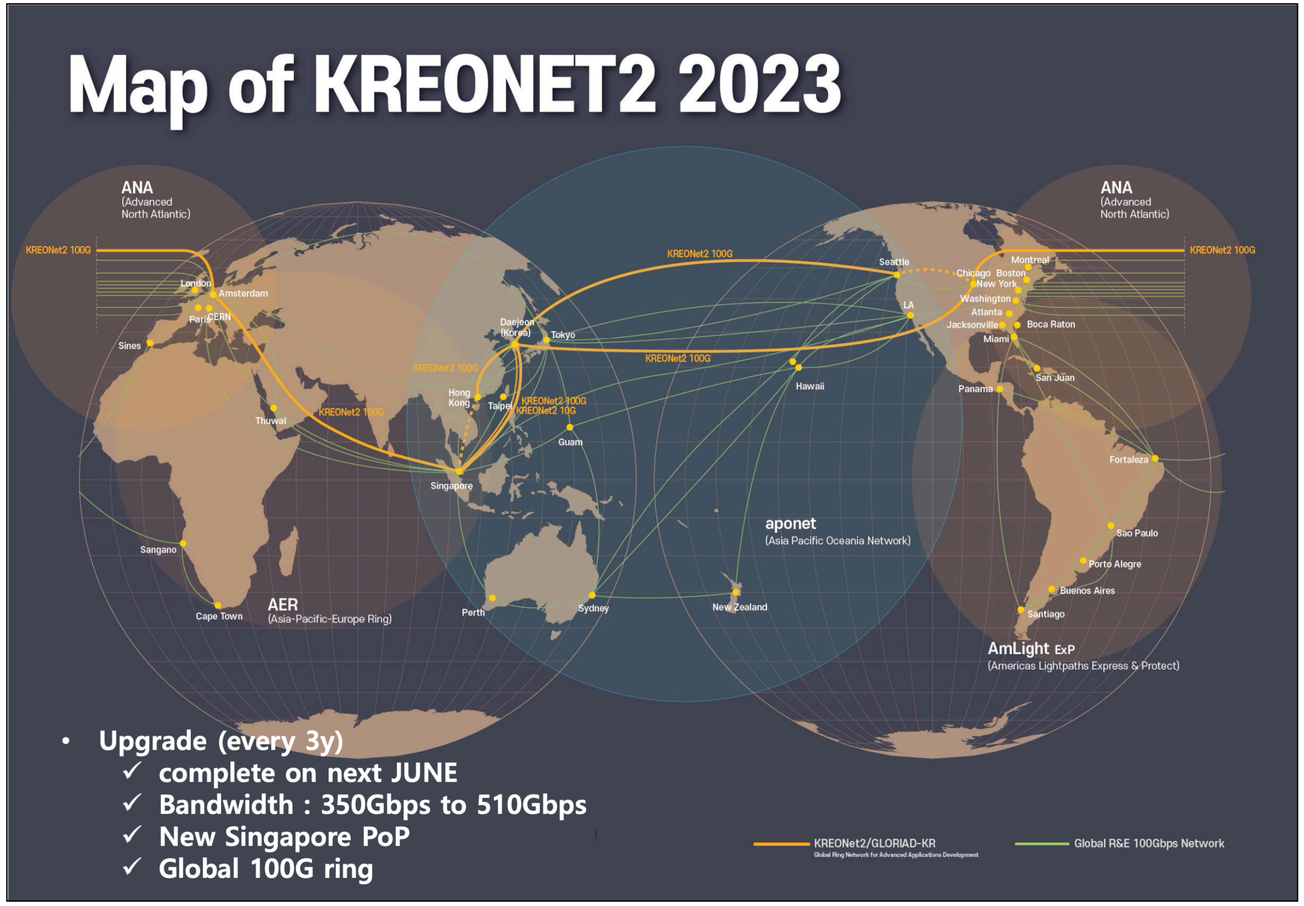
LHC Networking - OPN

Dedication to LHC Raw Data Transfer between T0 and T1s

LHCOPN

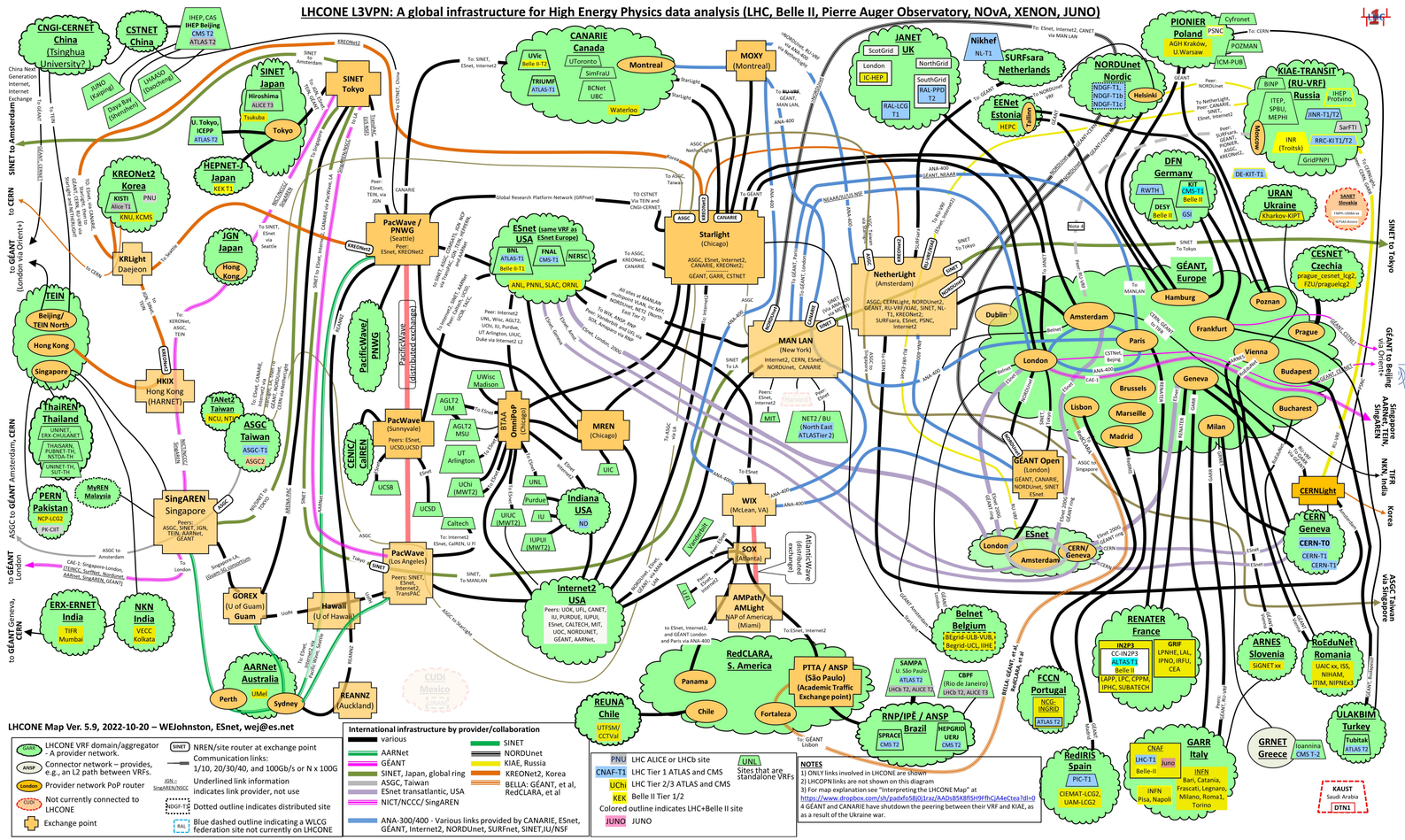


- 20Gbps dedicated links from Daejeon to Geneva provided by KREONet2 with its 100Gbps lambdas
- Primary optical fibers: Daejeon-Chicago-Amsterdam-Geneva (Backup links through Daejeon-Seattle & GLORIAD-consortium)
- KREONet2 directly reaches Geneva from Amsterdam PoP
- Provisioning of 100Gbps by end of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

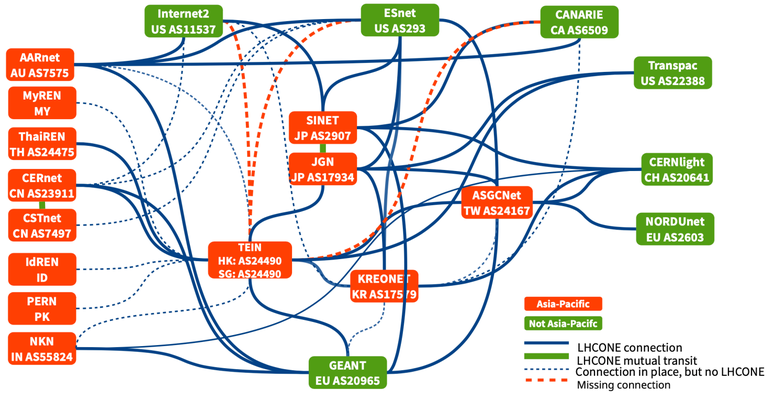


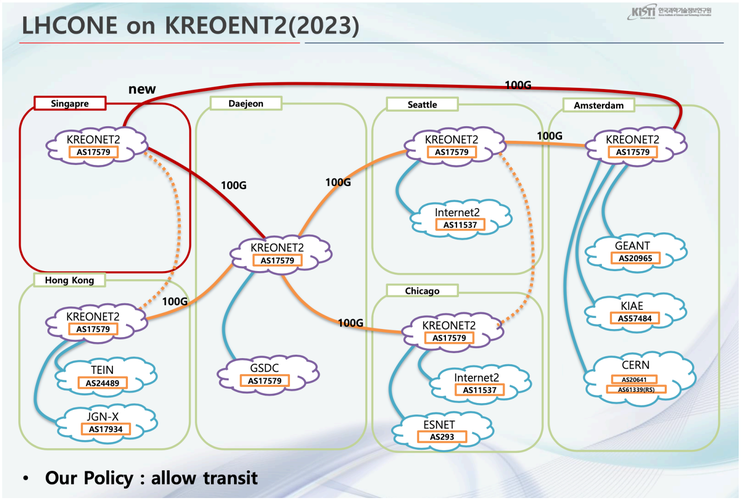
LHC Networking - ONE

Towards full mesh reachability among Tier sites for Big sciences



Asia-Pacific VRFs – Current Status



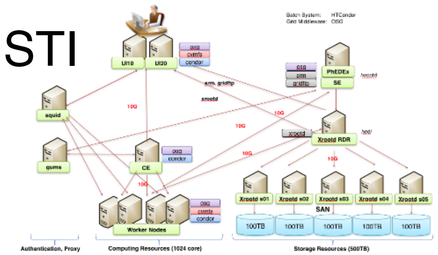


- Policy that allows transit via KREONet2 resolves missing connections in Asia-Pacific region

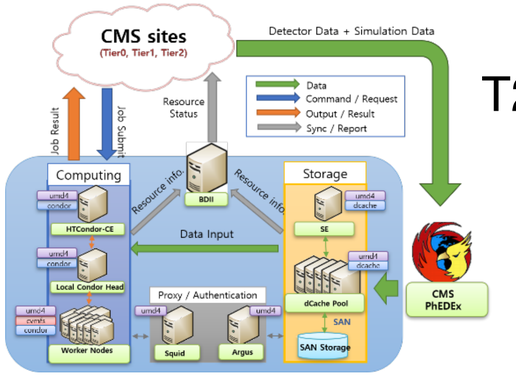
About KISTI CMS Tier-2

- 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

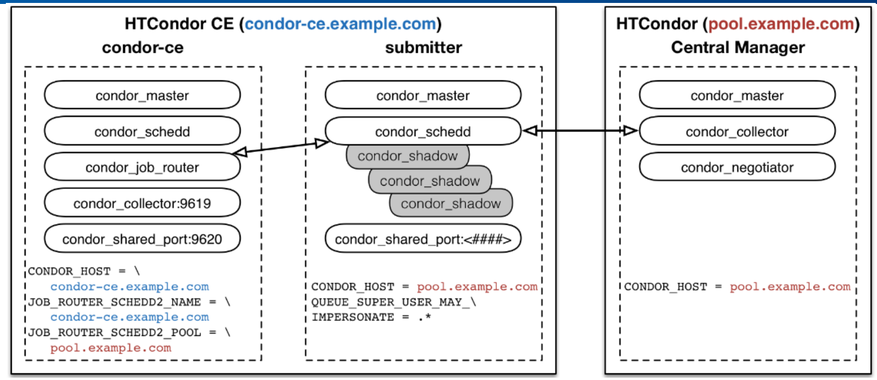


T2_KR_KISTI



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

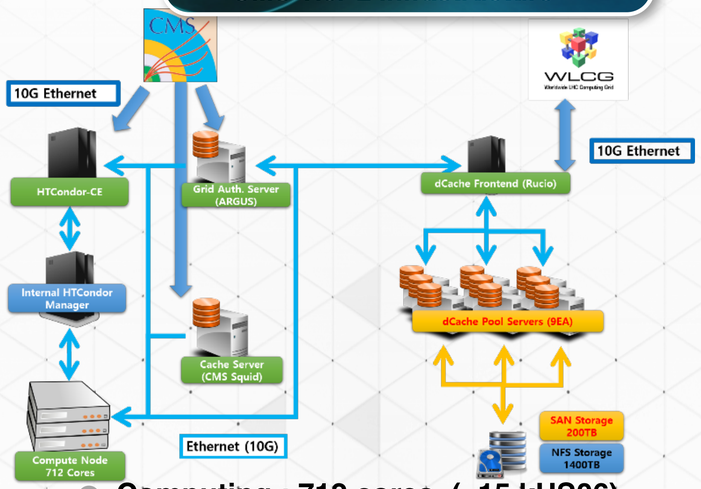
CellName	DomainName	Movers			Restores			Stores			P2P-Server			P2P-Client			queue ftp			queue webdav			regular		
		Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued	Active	Max	Queued
	Total	632	12620	0	0		0	0		0	0	120	0	0			2	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	47	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	40	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	1120	0	0		0	0		0	0	10	0	0			1	20	0	7	200	0	30	1000	0
cms-t2-wn1063-NFSPool	cms-t2-wn1063-NFSPool-Domain	64	1120	0	0		0	0		0	0	10	0	0			0	20	0	3	200	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			queue ftp			queue webdav			regular		

Disk Space Usage

CellName	DomainName	Total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/sticky/[illegible] 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



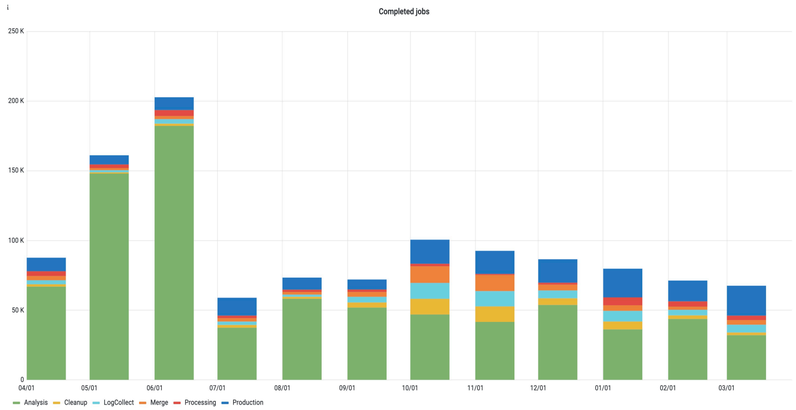
- Computing : 712 cores (~15 kHS06)

Storage Usage

- Disk 1,761 TB (Usage 75.70%)

Job Activities

~1.15 million jobs during this year



Data Transmission

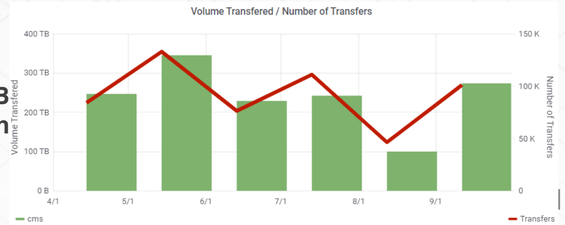
Efficiency matrix - by Experiment_Site (ES)

Src_exp_site\Dst_exp_site	T2_KR_KISTI
T0_CH_CERN	61%
T1_DE_KIT	67%
T1_ES_PIC	73%
T1_FR_CCIN2P3	72%
T1_IT_CNAF	73%
T1_RU_JINR	72%
T1_UK_RAL	58%
T1_US_FNAL	83%
T2_AT_Vienna	62%
T2_BE_IIHE	62%

- 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



Dr. Ryu, Geonmo

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)

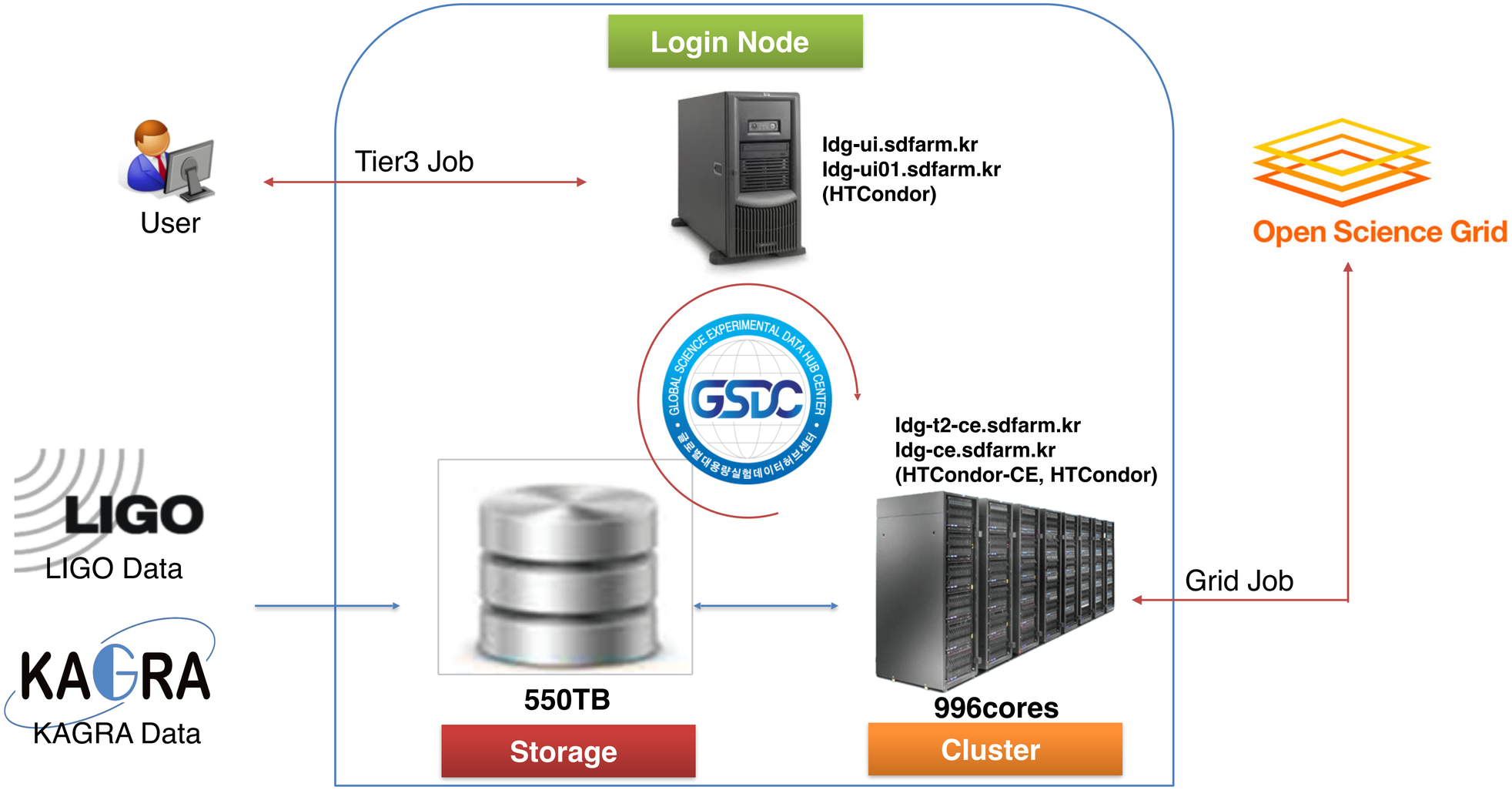
2020
- 996 core, 550TB
- HTCondor
- Added 3 GPU cards

2021
- 996 core, 550TB
- HTCondor

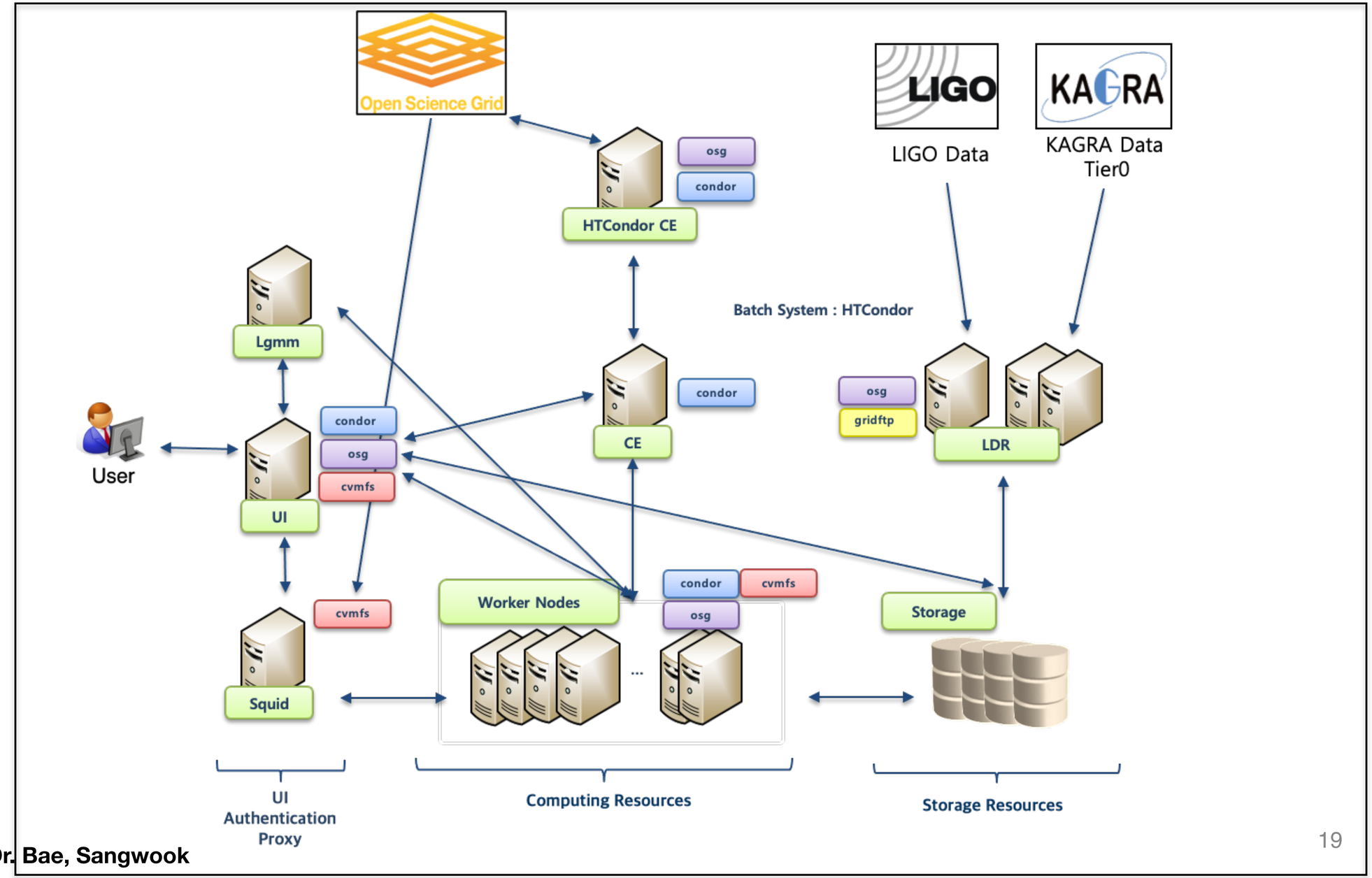
IGWN | Computing

Dr. Bae, Sangwook

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

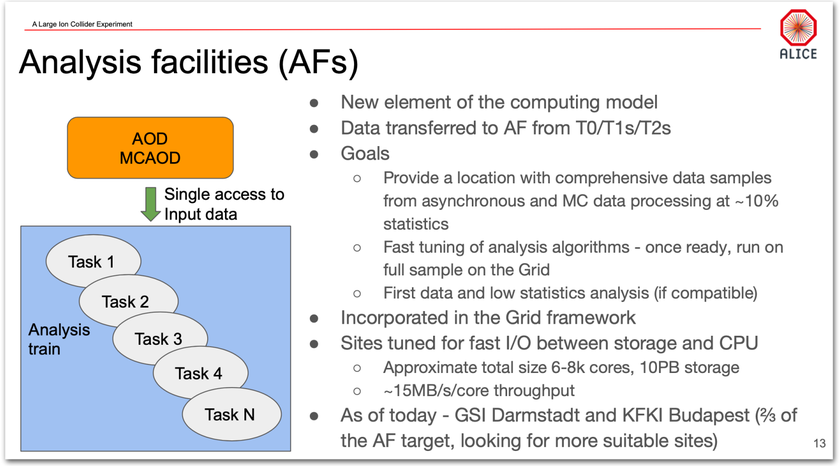
	Mount on	Size	Used	Avail	Use	Total
LIGO	/data/ligo/	400T	250T	151T	63%	pool0.gsn.sdfarm.kr:/ifs/service/ligo
KAGRA	/data/kagra/	150T	76T	75T	51%	pool0.gsn.sdfarm.kr:/ifs/service/kagra

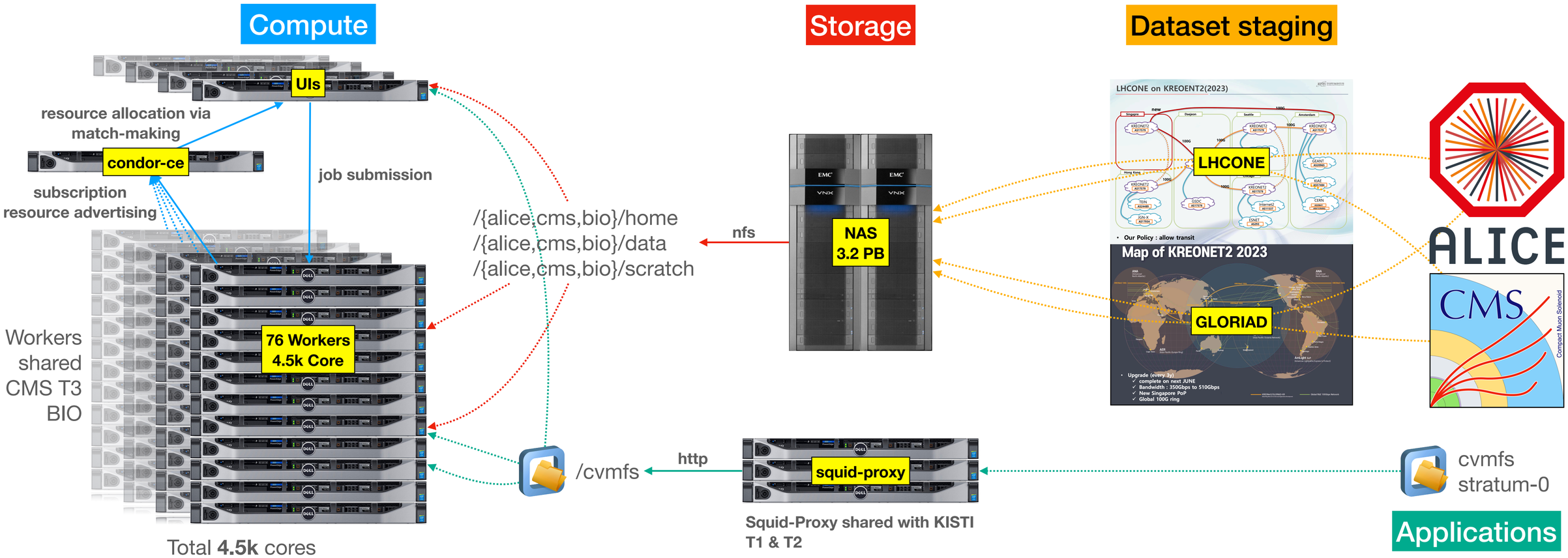
KISTI Tier-3

For Domestic Researchers

- Unified HTCondor Clusters provided for different experiments
- Quota and groups are managed by HTCondor Negotiator
- Application distributed by CernVM-FS

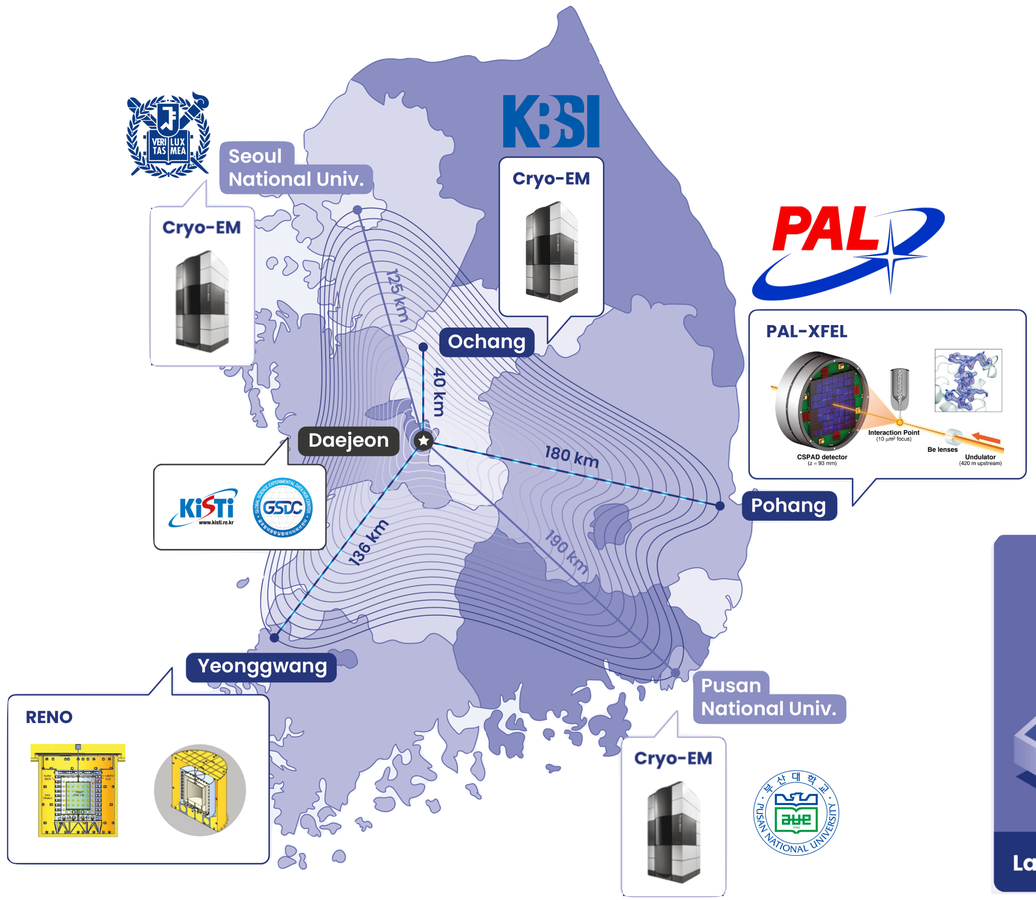
≠





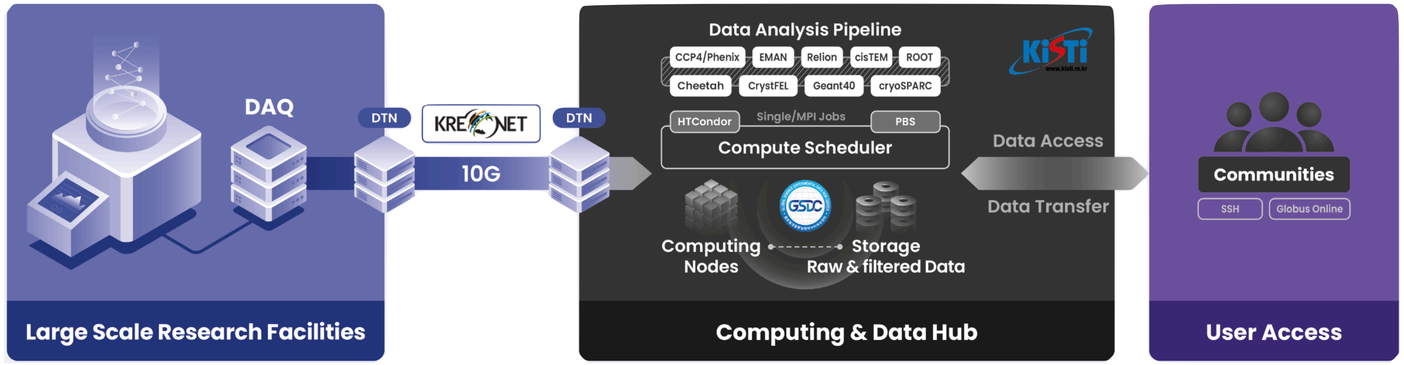
Supporting Domestic Research

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**



Thank you