IHEP Site Report

LU WANG, LU.WANG@IHEP.AC.CN
ON BEHALF OF COMPUTING CENTER, IHEP
HEPIX FALL 2022

Outline

- Brief Introduction
- Computing platform
  - Computing
  - Storage
  - Network
- Supports and R&Ds
- Summary

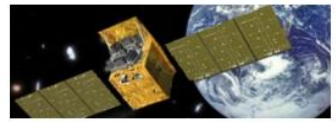
Brief Introduction

- 52.8 K CPU cores, 210 GPU cards to for more than 10 experiments
  - HTCondor cluster runs for HTC jobs (38K CPU cores)
  - Slurm cluster runs for HPC jobs (10K CPU cores + 210 GPU)
  - Distributed computing, WLCG,DIRAC etc., (4.8K CPU cores)
- 76.4 PB disk storage, 51.8 PB tape storage
  - Lustre (30.4 PB, POSIX) and EOS (46 PB, XRootD,+7.6 PB)
  - Castor for tape storage (19 PB, will be retired)
  - EOSCTA for tape storage (32.8 PB, +9.8 PB)
- Network
  - IPV4/ IPV6 dual stack
  - Ethernet / IB protocols supported
  - LHCONE member
  - WAN Bandwidth: 40Gbps (LHCONE 20Gbps)

Chinese local or IHEP driven

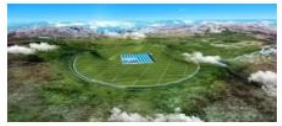
BESIII (Beijing Spectrometer III at BECPII)

JUNO (Jiangmeng Underground Neutrino Observatory)



HXMT(Hard X-Ray Moderate Telescope)

China Spallation Neutron Source



LHAASO (Large High Altitude Air Shower Observatory)

HEPS (High Energy Photon Source)

International collaborated

CMS

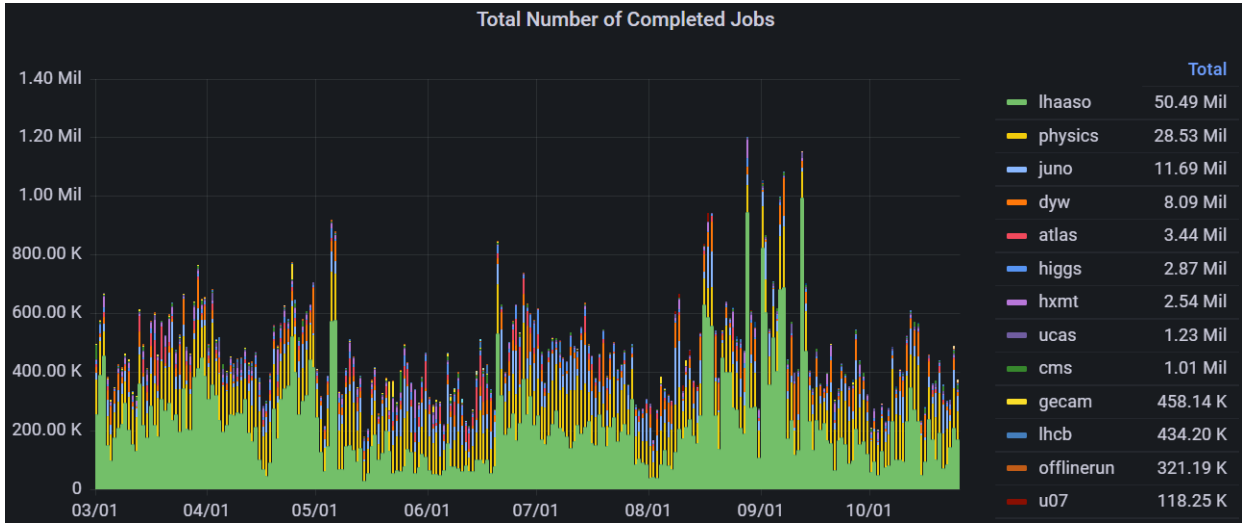
ATLAS EXPERIMENT

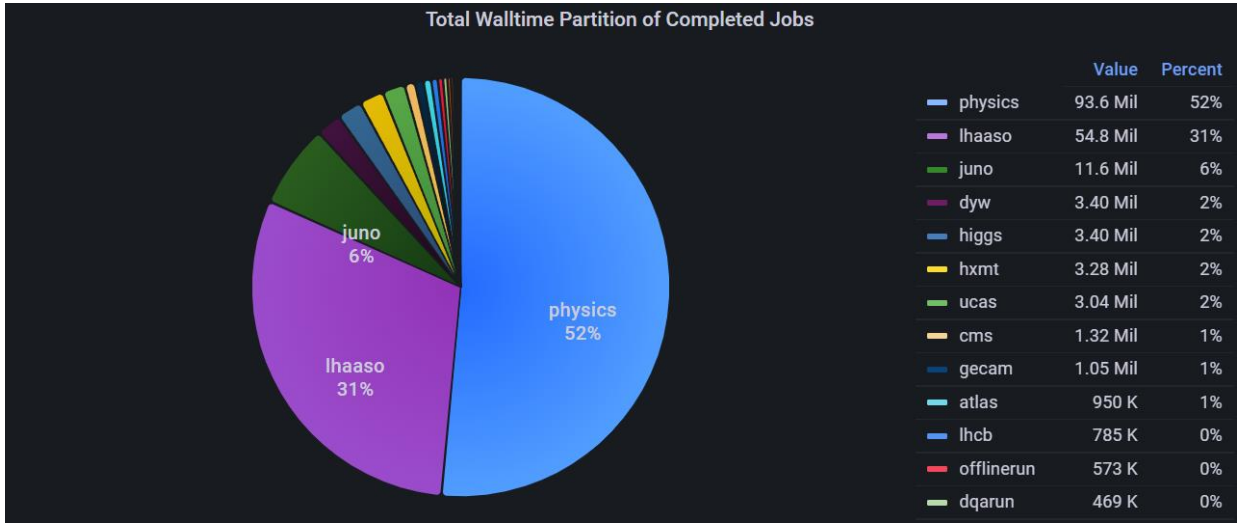
LHCb

Belle II

HTCondor/dHTC

- Job Statistics (2022.3-2022.10)
- Total Job Number: 111,503,370 Jobs
- Total Walltime in hours: 179,153,459 CPU Hours

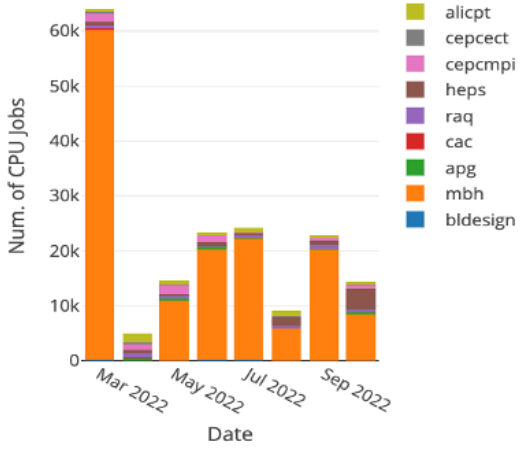




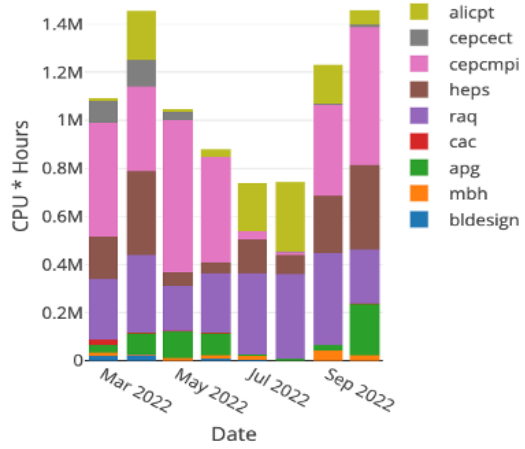
Slurm Cluster Status

- Resources:
  - 228 worker nodes
  - ~ 10K CPU cores, 210 GPU cards
- 9 CPU APPs, 11 GPU APPs
- Job Statistics (since 2022.03):
  - 177.2K CPU jobs, 8.7M CPU Hours
  - 253.2K GPU jobs, 742.7K GPU Hours

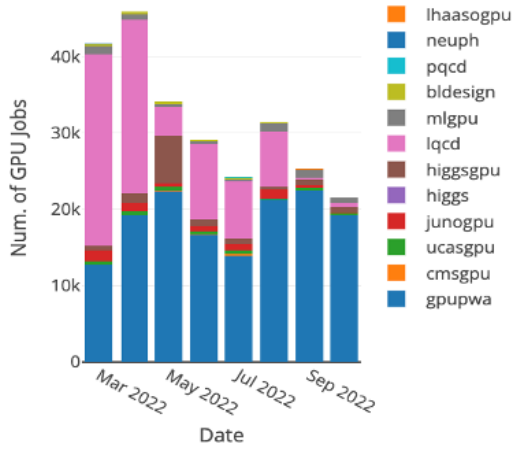
Num. of CPU Jobs of CPU_APP groups



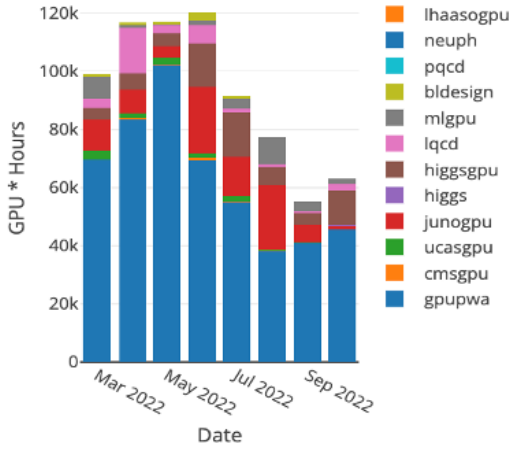
CPU Hours of CPU_APP groups



Num. of GPU Jobs of GPU_APP groups

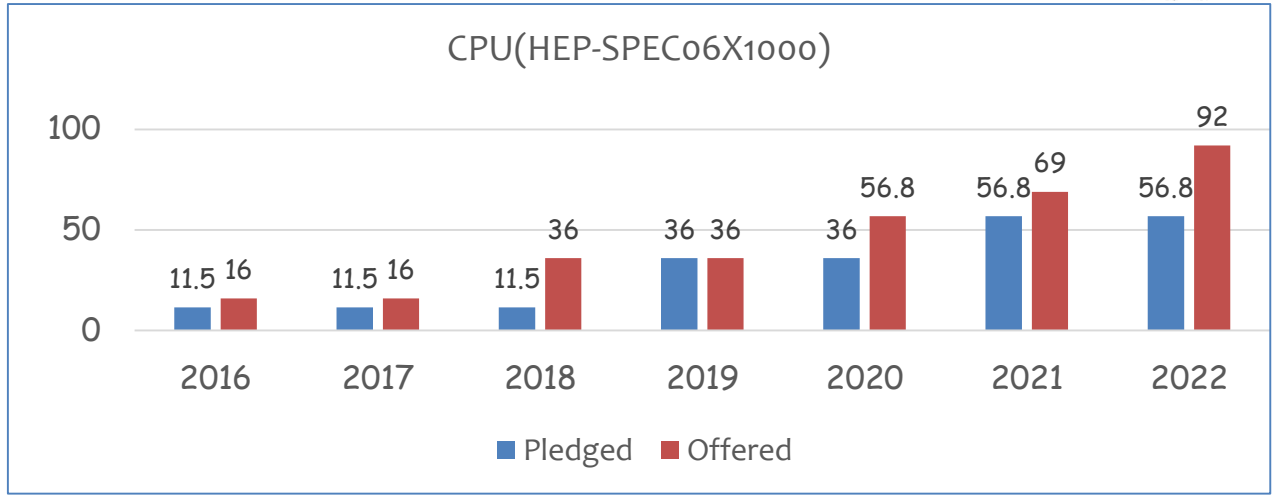


GPU Hours of GPU_APP groups

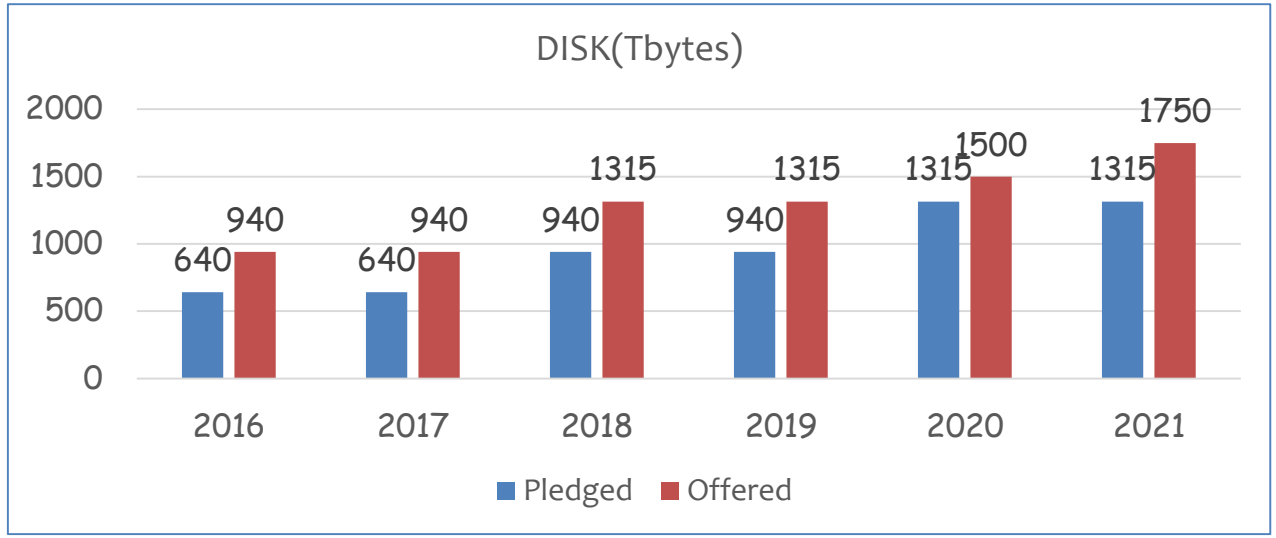


BEIJING-LCG2 Tier2

- Computing : 4232 CPU cores
  - CE & Batch: HTCondorCE & HTCondor
  - VO: ATLAS, CMS, LHCb, BelleII, JUNO, CEPC
  - Upgrade Linux kernel to get avoid potential risks of vulnerability

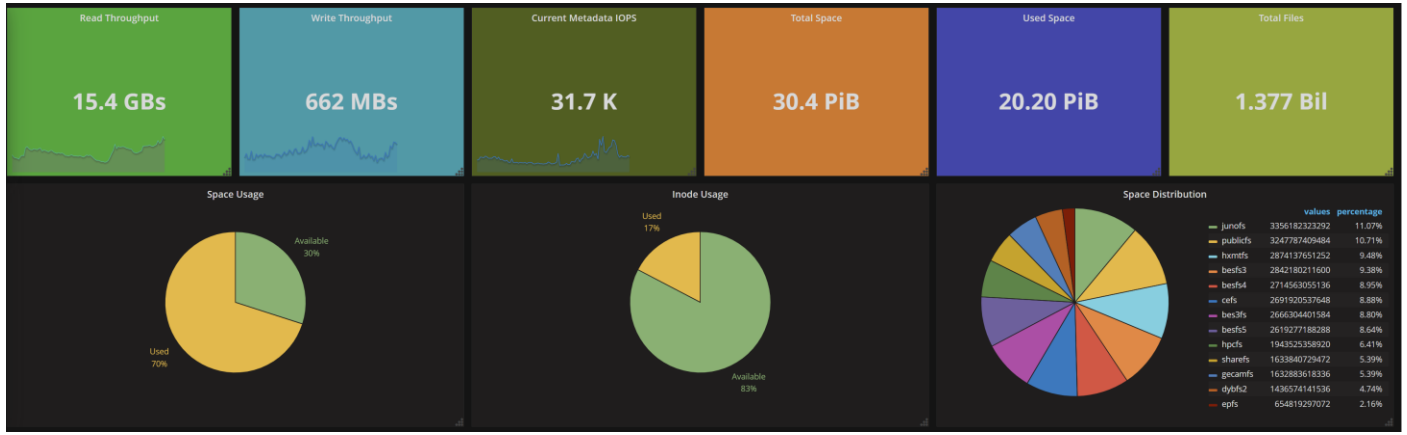


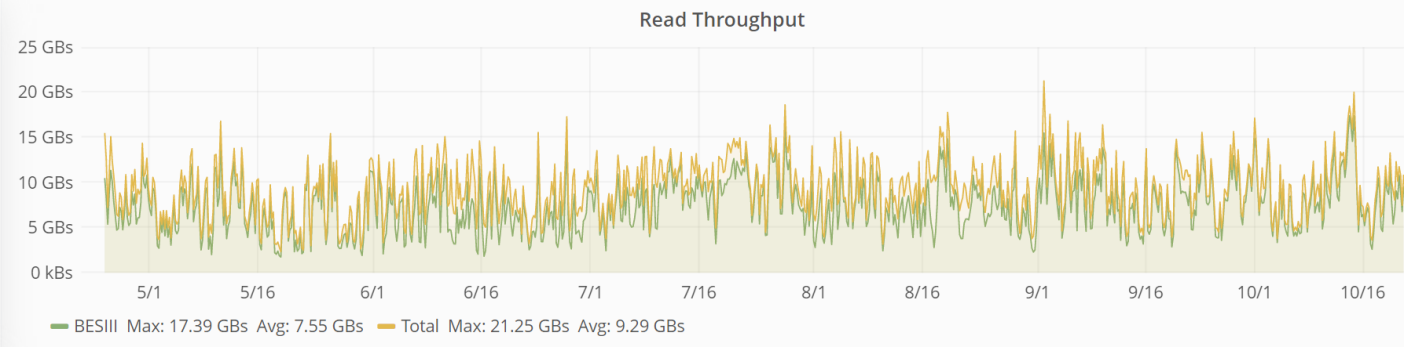
- Disk Storage: 1750TB
  - upgrade to DPM-1.15.2
  - + ICL VO support
  - Migrations from DPM-> EOS/dCache?

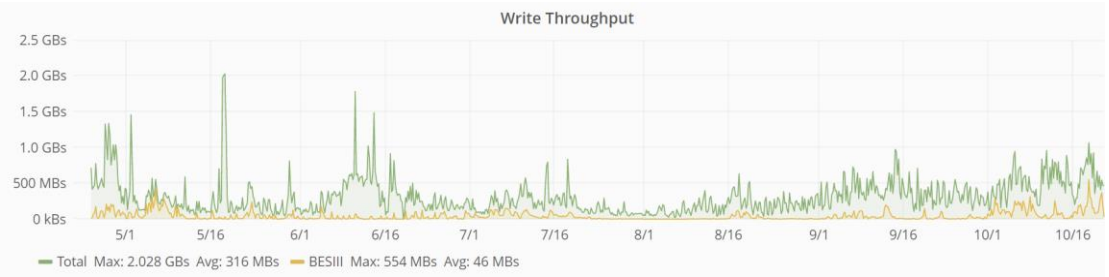


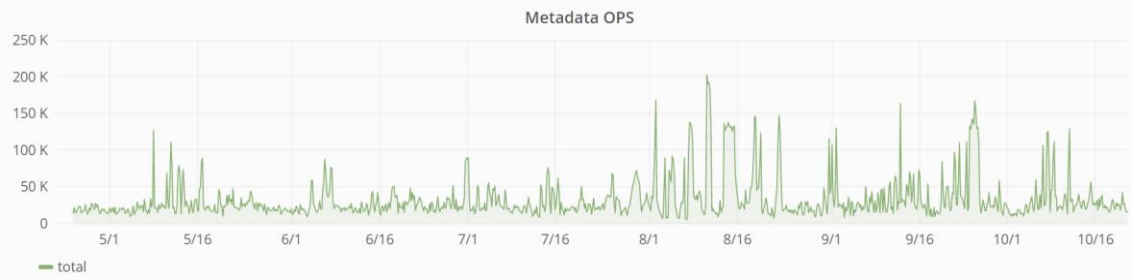
Disk Storage: Lustre

- Software: community Lustre (2.12.5)
  - 21 instances(+2 new , -1 retired)
  - +2PB capacity, +112M files
  - Read: 9.29 GB/s avg, 21.21 GB/s peak
  - Write: 316 MB/s avg, 2 GB/s peak
  - Metadata OPS: 18.2k avg, 192k peak





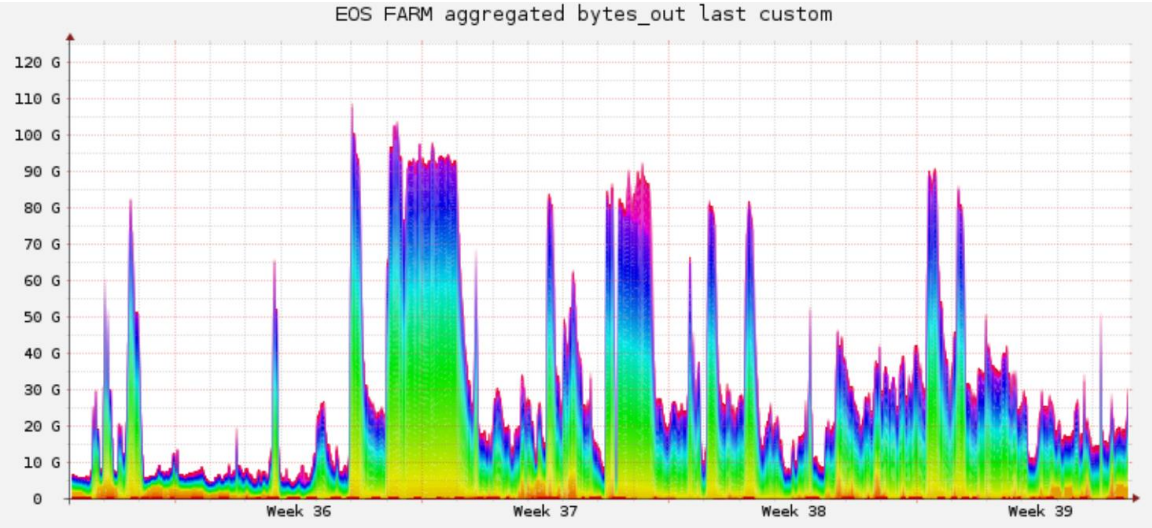




Disk Storage: EOS

- 5 instances, 4 for physics experiments, 1 for Box
  - Main data storage for LHAASO and JUNO

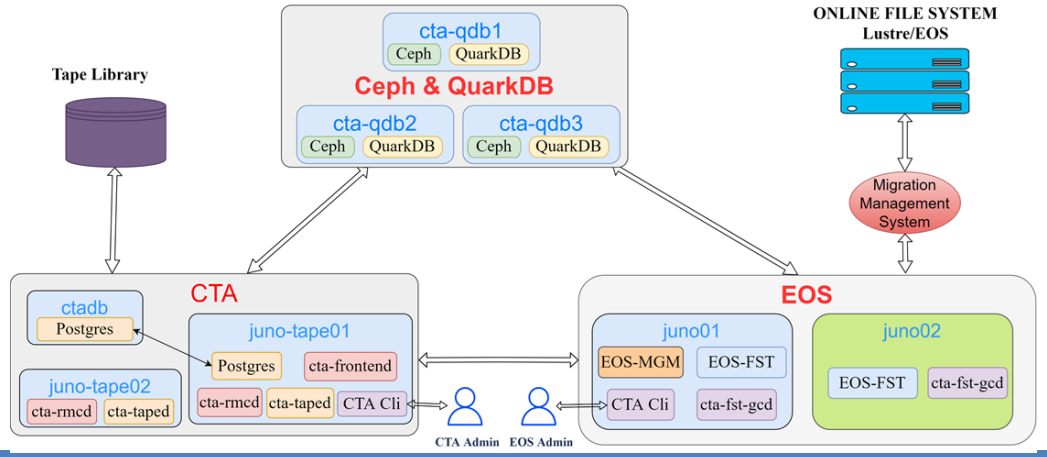
Raw Capacity	~ 46 PB
Disk server	~110
Number of EOS fs	3465
Number of files	~555 (+33)Mil
Number of directories	~15Mil
Peak Read throughput	~110(+30)GB/s



Tape Storage: EOS CTA

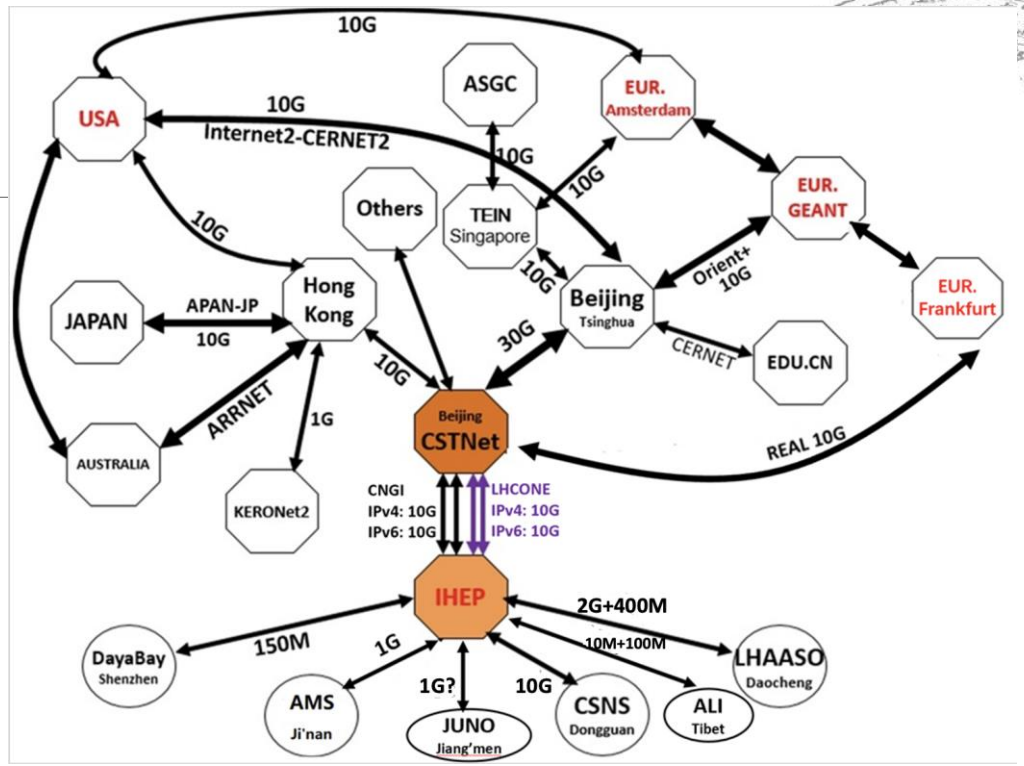
- Current Status
  - 18.14 PB used / 32.8 PB in total
- Serving LHAASO, DYB, YBJ, HXMT and BESIII
- New tape Library for JUNO experiment
  - LENOVO/IBM TS4500
  - 6 frames, 3 LTO9 drives
  - 120 tapes
- Data Migration from Castor to CTA

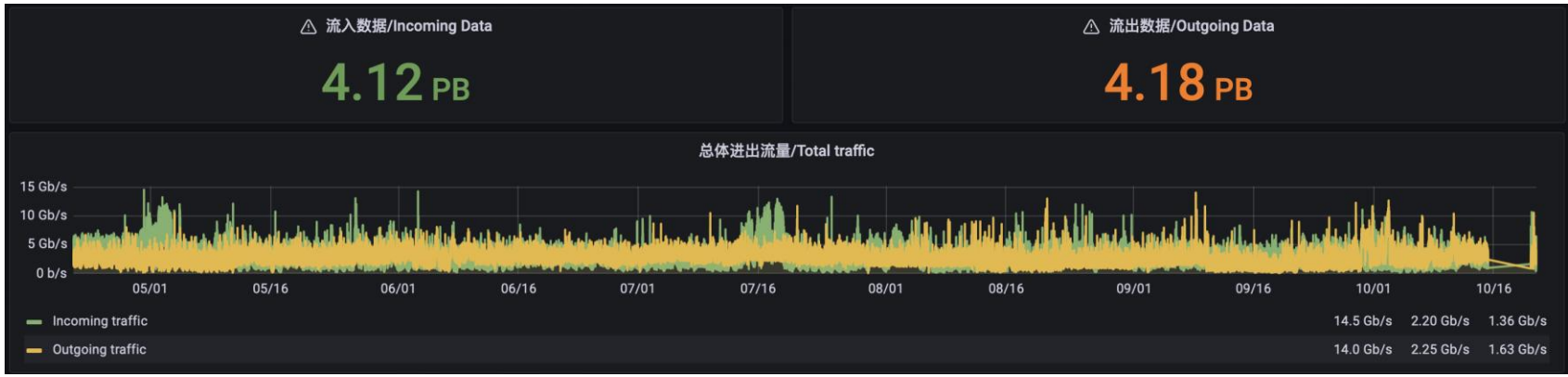
	Total	Completed	Completion Rate
BESIII	2.5 PB	1.3 PB	52%
DYW	1.2 PB	1.2 PB	100%
YBJ	525 TB	504 TB	96%



Network

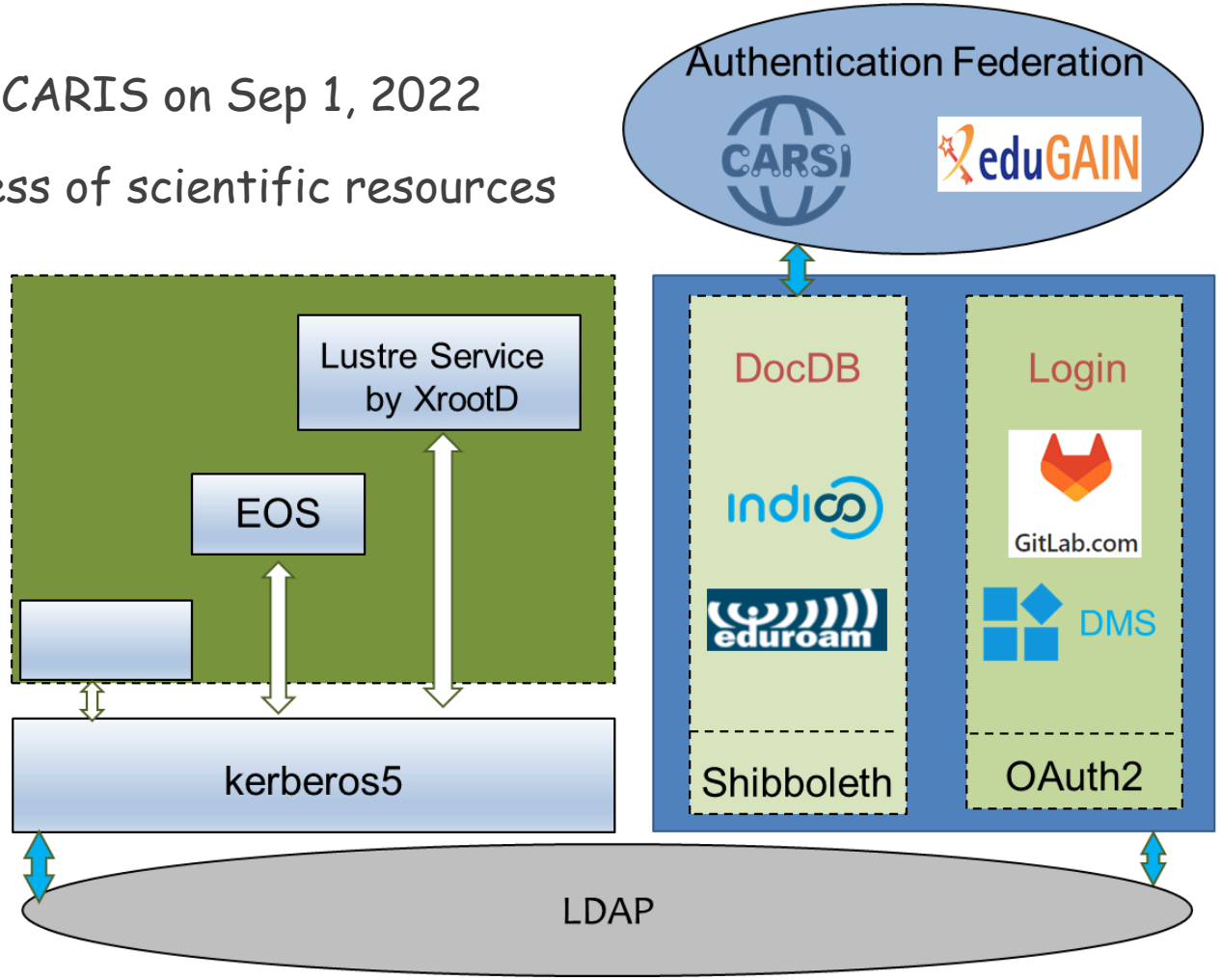
- Upgrade the bandwidth between IHEP and CSNS
  - To support future JUNO raw data transferring
  - Bandwidth 5Gbps → 10Gbps
- Upgrade data center network
  - Backbone bandwidth: 800Gbps(Storage),600G(compute)
  - Added 25G TOR switches: count (14 → 19)
  - 25G port Usage: 71.2% to 62.9%
- Peak traffic of IHEP WAN
  - Incoming: 14.5 Gbps
  - Outgoing: 14 Gbps





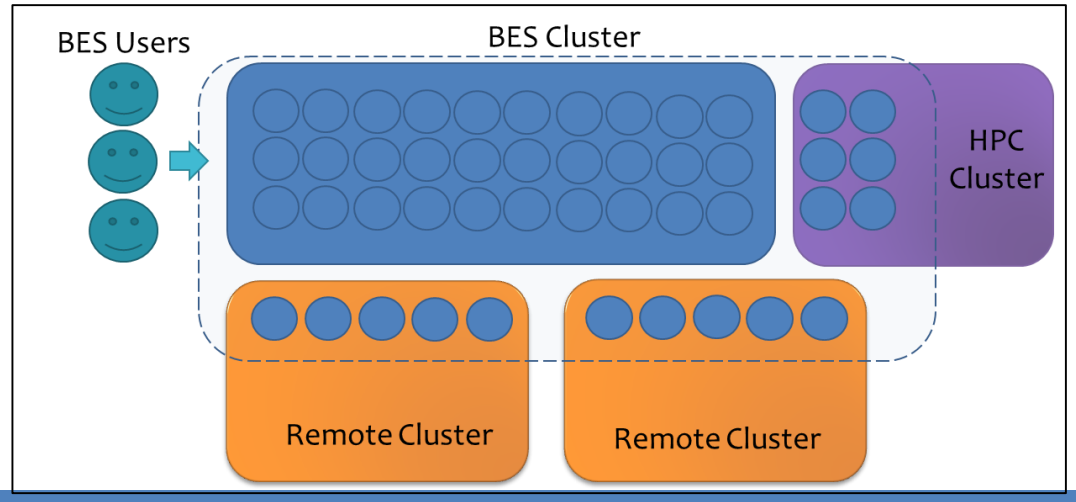
eduGAIN

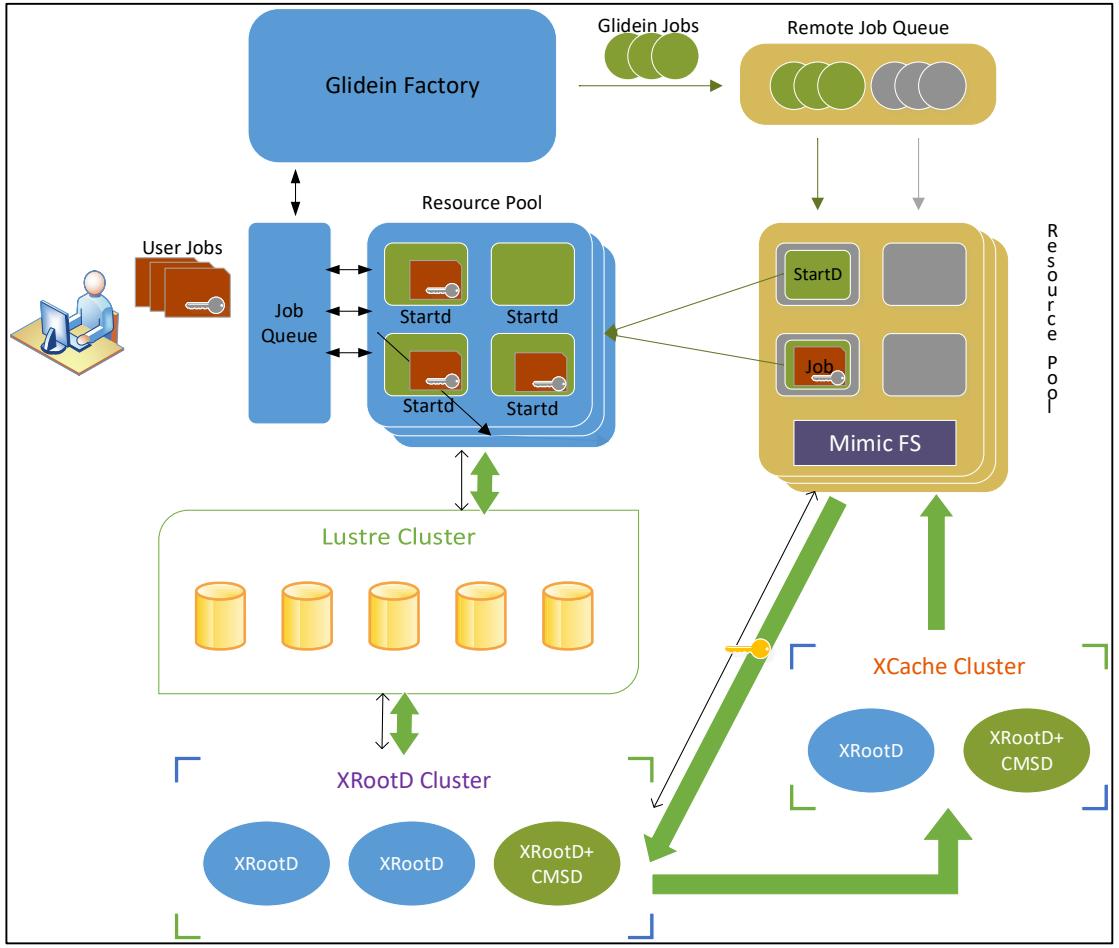
- IHEP joined eduGAIN identity federation through CARIS on Sep 1, 2022
- IHEP SSO users benefited a lot from the easy access of scientific resources via eduGAiN
  - 21000 user accounts
  - 200 OAuth applications
  - 160 collaborations
  - 120 cluster groups
- Next step, external users will be allowed to sign-on some IHEP web services via eduGAIN
  - Indico, DocDB …



One Platform, Multi-Centers (1/2)

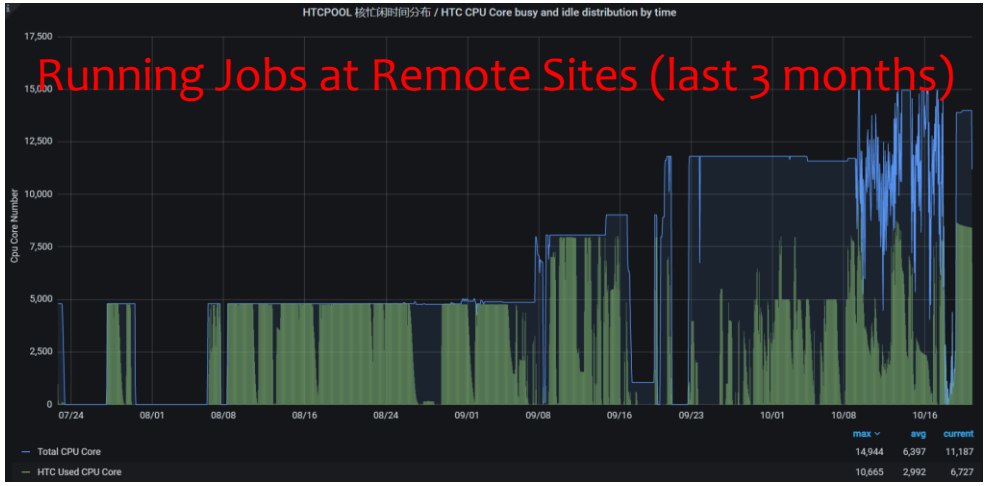
- OPMC aims to extend the IHEP-CC site by integrating the domestic distributed resources
  - Remote Resources: 8K X86 Cores, 10K ARM Cores
  - HTCondor Glidein (Scheduling & Resource Mgmt.)
  - Home-developed Mimicfs (POSIX Data Access)
  - XRootD Cluster (Lustre exportation)
  - XCache Cluster (remote read-only data cache)
  - Kerberos tokens (user mapping and authentication)

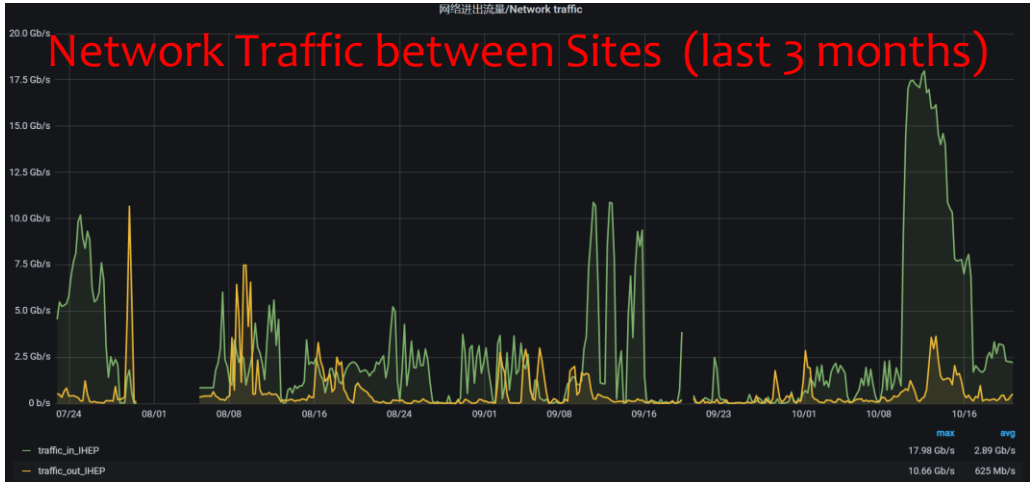




One Platform, Multi-centers (2/2)

- Simulation jobs with small I/Os can be transparently scheduled to the remote sites
  - BESIII、LHAASO and HERD experiments, typical I/O of O(10) KB/s





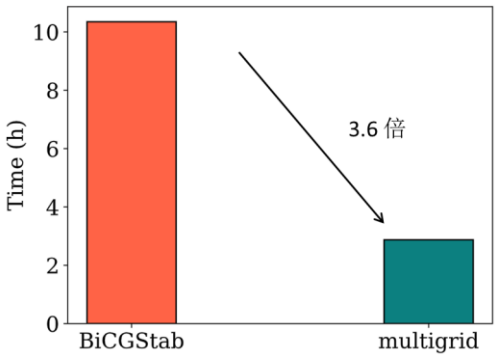
- Scheduling of BESIII reconstruction jobs to is under development
  - concatenating the simulation and reconstruction stages into one job to reduce network traffics of intermediated data
  - Cache the RandomTrigger files on remote site to improve reconstruction performance

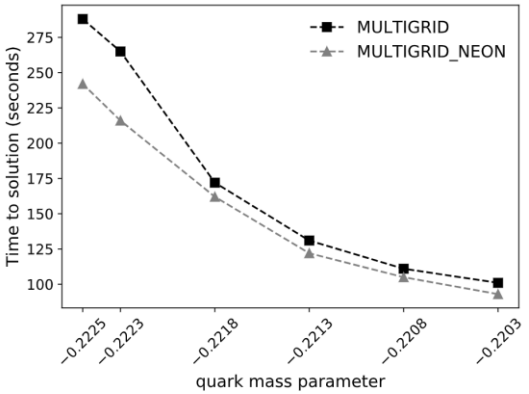
R&D: Simulations on the ARM architecture

- The largest ARM computing cluster in HEP at DongGuan, Guangdong
  - Huawei Taishan 200K server with Kunpeng 920 CPU
  - 100 worker nodes & ~10K CPU cores
  - 100 Gb RoCE interconnection

Taishan 200K server

- Lattice QCD
  - 3.6x speed up of LQCD simulation with multigrid algorithm
  - Another 20% performance improvement with matrix multiplication compression and NEON vectorization
  - Test->Operation
    - 4136 jobs, 2508376.5 machine hours, 200TB data was generated



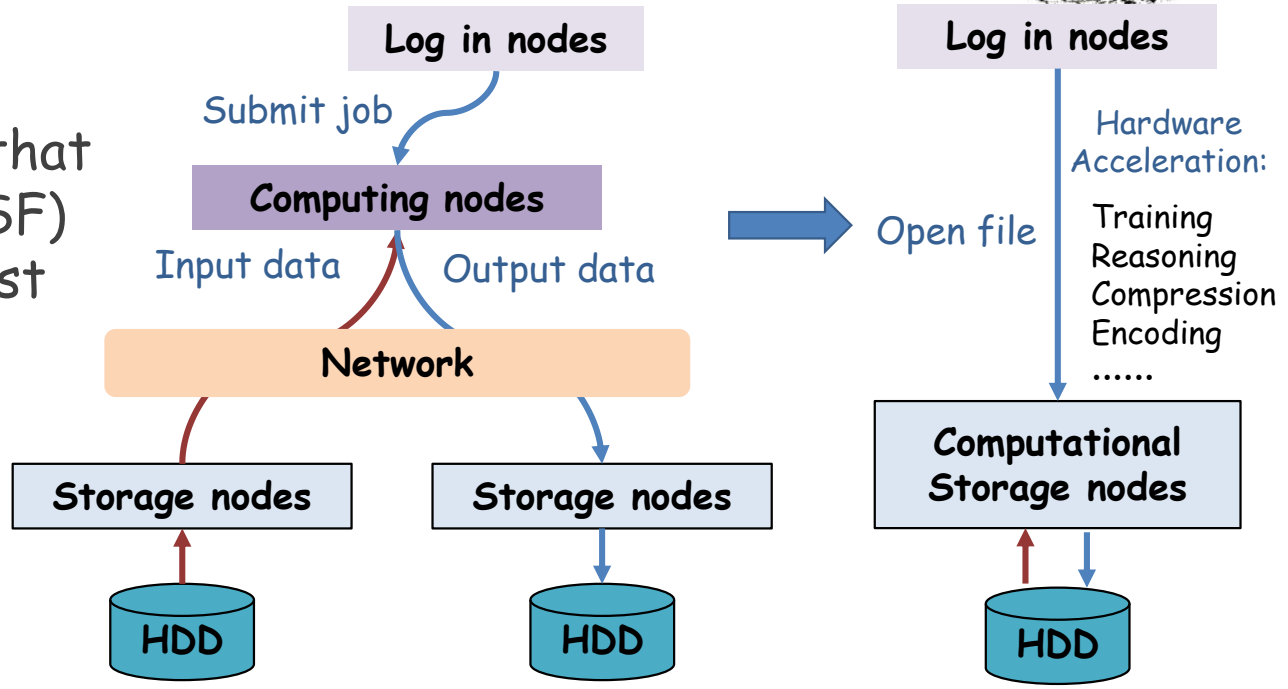


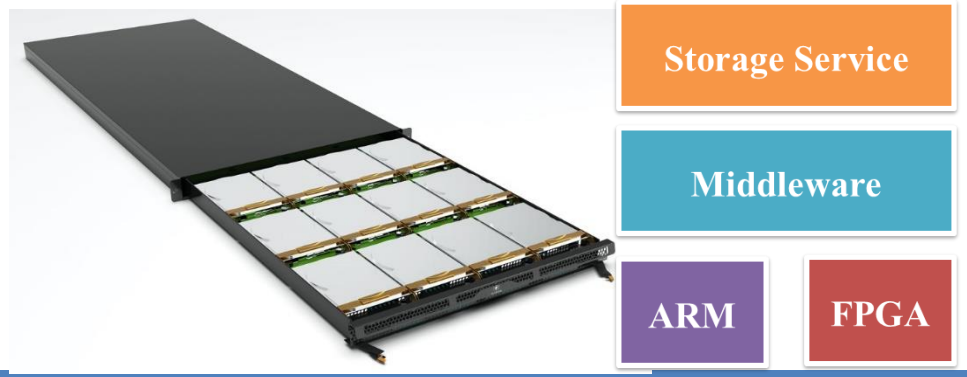

```
[root@condor06 config.d]# condor_status -const "Arch==\"aarch64\""
Name                                      OpSys      Arch    State     Activity LoadAv Mem     ActvtyTime

slot1@ihep_glidein_89961@acn001.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:46
slot1@ihep_glidein_52313@acn002.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:45
slot1@ihep_glidein_55427@acn056.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:45
slot1@ihep_glidein_55264@acn059.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:44
slot1@ihep_glidein_35588@acn064.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:44
slot1@ihep_glidein_63095@acn065.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:44
slot1@ihep_glidein_29280@acn066.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:46
slot1@ihep_glidein_66396@acn067.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:45
slot1@ihep_glidein_57463@acn068.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:45
slot1@ihep_glidein_66966@acn069.csns.ihep.ac.cn LINUX      aarch64 Unclaimed Idle      0.000 247947  0+00:04:39

                     Total Owner Claimed Unclaimed Matched Preempting Backfill  Drain

 aarch64/LINUX    10     0       0         10       0          0        0      0

        Total     10     0       0         10       0          0        0      0
```

- Data Simulation of HERD Experiment (newly added)
  - Software libs have been compiled on ARM archtecture and published in CVMFS
  - Jobs can be automatically forwarded to Dongguan site through HepJob
  - Transparent data I/O through the Mimicfs to Lustre fs at IHEP

R&D: Computational Storage

- Computational Storage is the architecture that provide Computational Storage Functions (CSF) coupled to storage, into order to offload host processing's or to reduce data movements.
- Status:
  - Designed and customized a 1U server
    - Integration of ARM & FPGA
    - High density: 1U & 12 HDD
  - Two applications are implemented
    - data processing of LHAASO: >10PB /year
      - Auto-Decoding & Data compression
    - Lossless compression of X-Ray CT image for HEPS: >100PB/year
      - CNN training\reasoning, Huffman encoding



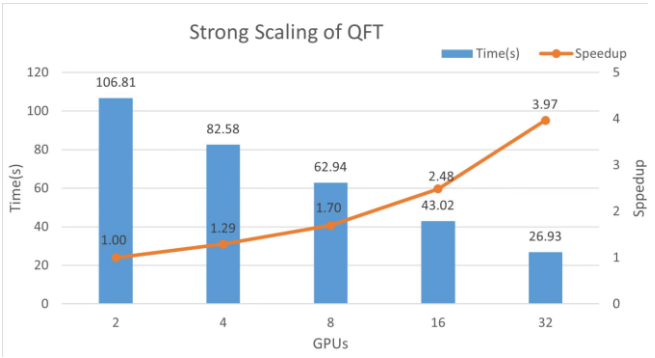


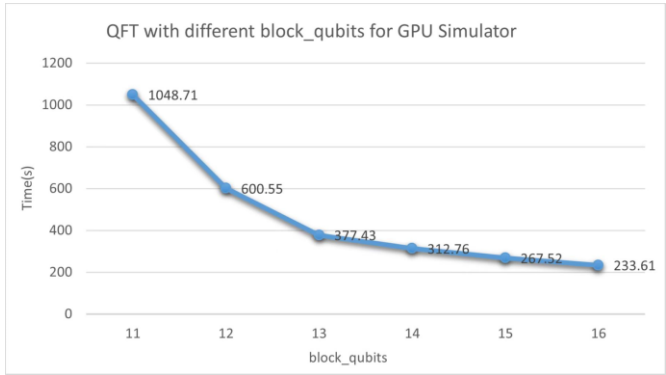
Self-developed Computational storage server

R&D: Quantum Computing Simulation Platform

- A distributed heterogeneous interactive platform to facilitie the explorations of quantum algorithms in multiple experiments
  - L-QCD, CEPC,BESIII …

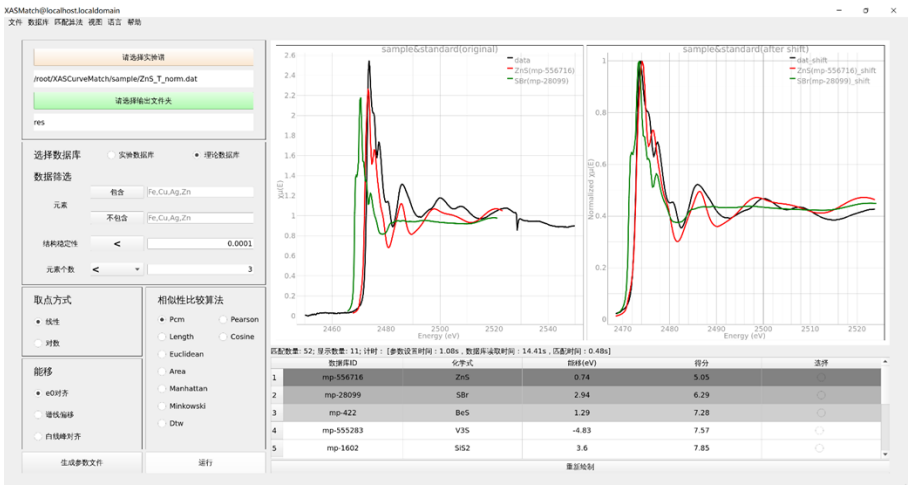
- Jupyter-based Interactive Developing & Analysis Platform
  - supports simulators like IBM Qiskit, Google Cirq, D-Wave Solver…
  - Support CPU and GPU environment
  - Support interacting with Slurm Cluster (under development)
  - Unified user authentication with IHEP SSO
- Drag-and-drop Programming Interface
  - Support basic quantum gates, and QASM translation
- Heterogeneous Computing Cluster
  - Simulating up to **38** qubits on IHEP GPU Cluster
  - Support Qiskit with cuQuantum and OPENMPI enabled

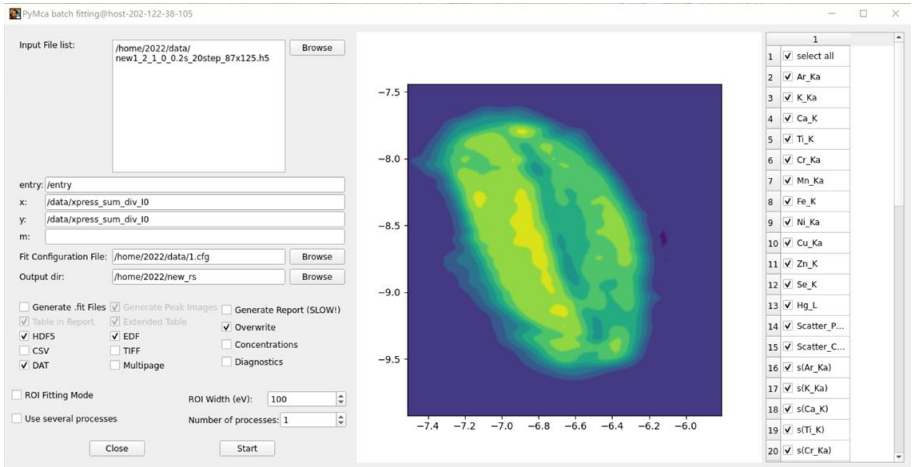




R&Ds of HEPS

- HEPS, High Energy Photon Source, will be put into service in 2025
- Computing & Communication system (HEPSCC)
  - Network, Computing, Storage, Data analysis framework, Data management, Database & Public Service , Monitoring, Security
- Data analysis framework(Under Development)
  - Integrate methods and algorithms:Alphafold, PDFgetX3, PyMca
  - Develop PyMca and X-ray absorption matching interface of Daisy workbench based on PyQt5
  - Develop the PDF workflow based on PyFai and PDFgetX3
- Data management system(Finished)
  - develop data management module for 3-Tiers data storage (beamline storage → Central storage → Tape)
  - Data transfer module moves data between different storage media automatically





R&Ds of HERD

- A working group for the HERD distributed computing system was launched on August 2022.
  - It would serve data production and data insert checking to bookkeeping system in the HERD data flow.
  - The computing model is based on the DIRAC-Rucio Integration Model of BelleII experiment.
  - The Data management Policy would follow the policy of JUNO experiment.
- R&D Status:
  - Rucio and DIRAC testbed was deployed, and test RSE and basic SE function was ensured.
  - Customized HERD data policy has been finished.
  - Data registering API tools for production with auto-registering and register-checking function is under developed.
  - DIRAC-Rucio integration is nearly finished.

Summary

- The Local cluster is running smoothly during last 6 months
  - increase of disk I/O throughput, extension of tape and network capacity according to experiment requirements
- Experiments have already benefited from project "One Platform, Multi-Centers"
  - BESIII, LHAASO and HERD simulation jobs can be scheduled to remote site without notification of users
- Simulations over the ARM computing cluster have seen performance improvements via adoptions of new algorithms
- Other R&D works are progressed as planed
  - Computational storage server and application
  - Quantum Computing Simulation Platform
  - Daisy & Data management system of HEPSCC
  - distributed computing system for HERD (newly launched since August)

Questions & Thanks!

Backup: Grid Sites: WLCG & DIRAC

- Resources and operation: (03/2022-10/2022)

	ATLAS	CMS	LHCb	BelleII	JUNO
CPU cores	444	444	1680	576	576
Storage	400TB	684TB	375 TB	255 TB	820 TB
Jobs	81k	82.7K	325.7K	361 K	515K
CPU times	1,592k hours	1,105K hours	8,930 K hours	1,594K hours	811K hours
Upload	124TB	129.4 TB	2.81 PB	1.54 PB	47.1 TB
Download	108TB	88.5 TB	43.2 TB	67.0 TB	340 TB