ASGC Site Report

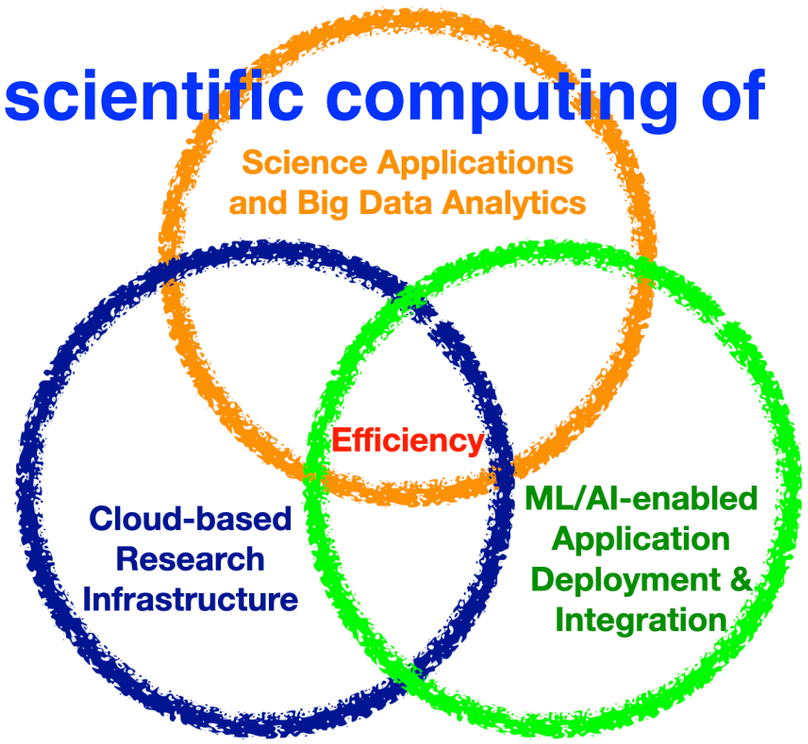
Eric Yen and Felix Lee

Academia Sinica Grid Computing Centre (ASGC)
Taiwan

HEPiX Autumn 2023
Victoria, Canada
17 Oct. 2023

ASGC Is Accelerating Discovery and Innovation

- ASGC joined WLCG development and deployment for the Large Hadron Collider grand challenges since 2001
  - ASGC T1 and WLCG Asian Regional Operation Centre has been operational from 2005
  - Reorganized as an ATLAS T2 from Oct. 2023
- ASGC has been supporting multi-disciplinary e-Science applications of Academia Sinica from 2006, based on WLCG core technologies
  - The research infrastructure, platform and services are improved progressively along with growing scientific applications of various disciplines
- System efficiency optimization (including power, thermal, system and applications, etc.) is also a strategic goal of ASGC aided by machine learning technologies
- ASGC becomes the Core Facility for big data and scientific computing of AS & Taiwan from 2023



Enabling Innovations by Integrated Research Infrastructure - Connecting Instruments, Data, Minds, and Computing

- **Fast growing needs of big data analysis and scientific computing is the primary challenge - cloud-based research infrastructure**
  - **Integrating experiment/instruments and analysis facility**
  - **Batch and interactive job submission**
  - **Optimization of Data analysis pipeline and system efficiency**
  - **Collaborations: ATLAS, CMS, AMS, KAGRA, ICECube, Proton Therapy, CryoEM/Synchrotron Source, Astronomy, Condense Matter, Lattice QCD, NGS, Bioinformatics, Earth Science, Environmental Changes, etc.**
- **Resources: 20,090 CPU Cores; 236 GPU Cards; 30 PB Disk Storage**
- **Leverage the WLCG core technology and develop capacity to support broader scientific applications**
- **24/7/365 services since 2006**
  - **Data Center availability: 99%+**
  - **Scientific Computing Service reliability: 97%+**
  - **Daily average power consumption:10,326 KWH (2023), >20% reduction than 2022**
  - **Power saving efficiency: ~ 20% (cluster-based)**
- **Reliability and Performance are the key objective**
  - **User Scale：(#Groups, #Users) = (90, 350)**
  - **Finished #Jobs (2023 estimated): > 5,000,000 (40% for WLCG)**
  - **International Data Transmission (Inbound + Outbound, WLCG): > 21PB (2022)**
  - **Inside Data Center Traffic (Inbound + Outbound) > 1PB daily**
  - **Training and workshop: 5 events a year**

WLCG Activities

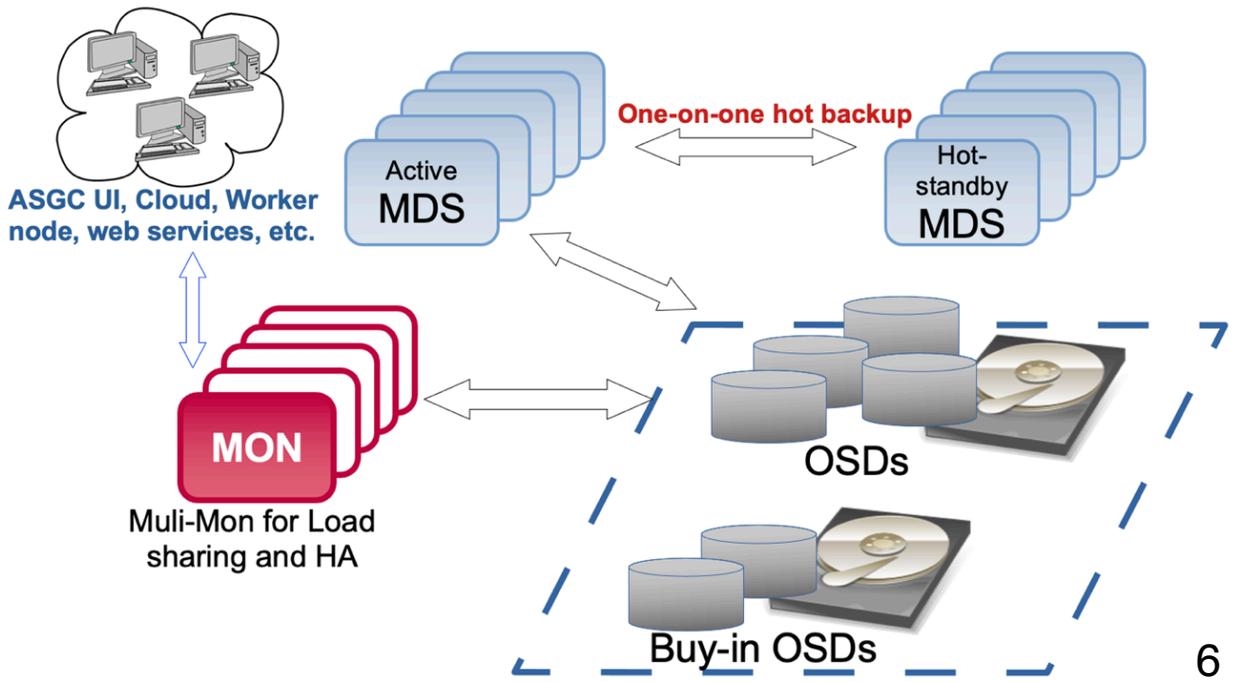
- Migrating to an ATLAS T2 Site
  - WLCG T1 mission ended since 1st Oct. 2023
  - Will associate with USATLAS T1
  - Will have 2,208 cores, ~36k HEPSpec2k6, and 5PB dis storage
    - 1.6x Computing power and 5x disk storage space growth of existing T2
- Migrating storage from DPM (9PB) to EOS
  - Thanks for the great help of ATLAS - distribute data sets to other sites first and purge dark data as well
  - Will have 5PB after migration
- Networking for LHCONE: supported by USATLAS and ESNet - discussion is ongoing
  - Local network service will be provided by Academia Sinica and TWAREN
  - LHCONE services will be supported by ESNet and TEIN/APAN
- Supporting ATLAS High Granularity Timing Detector (HGTD) DB operation and backup
- CMS T3 analysis facility setup and operation are supported at ASGC

Migration From DPM to EOS

- **No DPM to EOS direct migration tool is in place**
  - **Short of disk storages for one-to-one migration of 9PB data**
  - **Prepared 1PB as buffer on EOS side, then using RUCIO to do high level data transfer from DPM to EOS.**
  - **Gradually moving disk server from DPM to EOS as long as there are enough space released at DPM side.**
- **Everything was working just fine until significant high disk failure rates happening at EOS**
  - **Several zfs RAIDZ-2 crashed due to too many hard disks got failed at once (94 HDDs failed in recent 4 years) —> 15x higher than average failure rates**
  - **The root cause is still unclear, but vendor(HPE) replaced disk back-plane, and we were forced to replace ZFS by hardware RAID-6.**
- **Entire migration was delayed because of this disaster…, now, it's still few PB to go**

ASGC Science Cloud Storage Architecture

- Scalable and reliable online storage system based on Ceph mainly
- Ceph Configurations: ~9PB
  - 6 MDS + 6 hot-standby (one-on-one backup); 7 MONs
  - 462 OSDs, 51 hosts.
- Services
  - 3 TB/PI Group setup by default; PI could extend the space through management UI flexibly
- Reached 2GB/s R/W throughput so far
- Tape-based remote backup system (4PB) will be established and integrated in late 2023, supported by EOS
- Providing big pool for HPC, HTC, AI and various applications concurrently
- Capacity will be growing to 13PB by end of 2023
  - Plan to procure new 4PB disk servers for Ceph System in 2024 and 2025 respectively

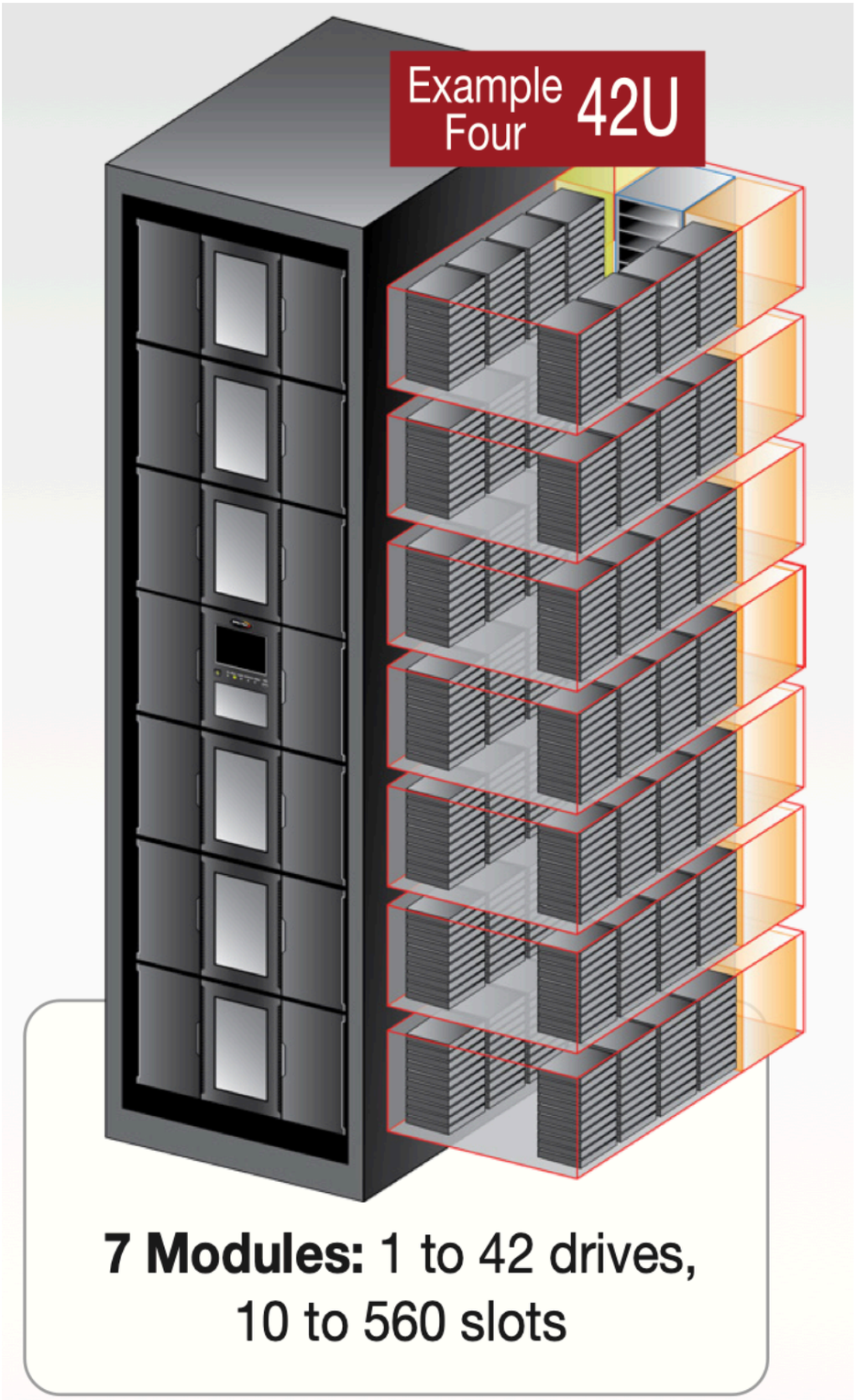


Lessons Learned From Ceph Operation

- Customization: for enhanced reliability and HA
  - Increased MDS and hot backup according to utilization/workloads
  - Grouping MDS for specific applications/directories
    - To avoid split/merge subtree across the MDS (performance degradation)
    - Could equip powerful H/W for I/O intensive services. e.g. bigger MDS memory
- ML-enabled analysis of Ceph for improved rebalancing, reliability and performance is under development
- Data lost of CephFS caused by multiple disk failures during recovery
  - The disks were failing one after another and eventually four of them hit a single PG (8+3 erasure coding)
    - Because of insufficient free/safe space especially
  - Remedies
    - Ensure/improve recovery efficiency:
      - Reconfiguring Ceph by smaller-scale storage server and disk drive, larger quantity of storage nodes, in order to enhance the reliability and performance of storage services (36x18TB —> 12x12TB/14TB)
    - Maintain 15-20% free capacity (OSD) for automatic recovery and surges of demands
- Bluestore corruption or unable to boot by unknown reason
  - No physical disk failure or what.
    - Purge OSD and recreating OSD from the same disk, then service get recovered
  - Suffered 5 times so far, ever since we switched from filestore to bluestore.
  - Nothing impacted if we have enough acting OSD in PG, but it’s annoying

Tape Remote Backup System Will be setup by end of 2023

- **Serving as 2nd layer remote backup system**
  - **Plan to setup at a separate data center**
  - **For cold data, or 2nd-copy backup**
  - **For backup of users' core data on Ceph**
- **Scalability: capacity on demand**
  - **Max 7 modules x 6u, 42 drives, 560 tape slots**
  - **LTO-9 tape: 1.44PB (native)/3.6PB (compressed) per module (80x 18TB/tapes)**
- **Tape drive performance: 300 - 400 MB/s using fibre network**
- **Integration and services: based on EOS and CTA**
- **Tape-related data services should be operational by end Q1 2024.**
- **Will be extended to 12PB capacity in 2024**



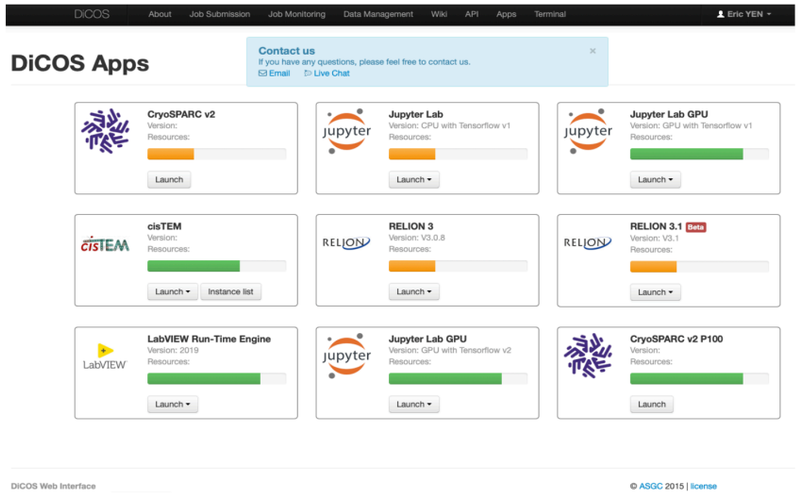
Support and Service of ML-Enabled Data Analytics by ASGC

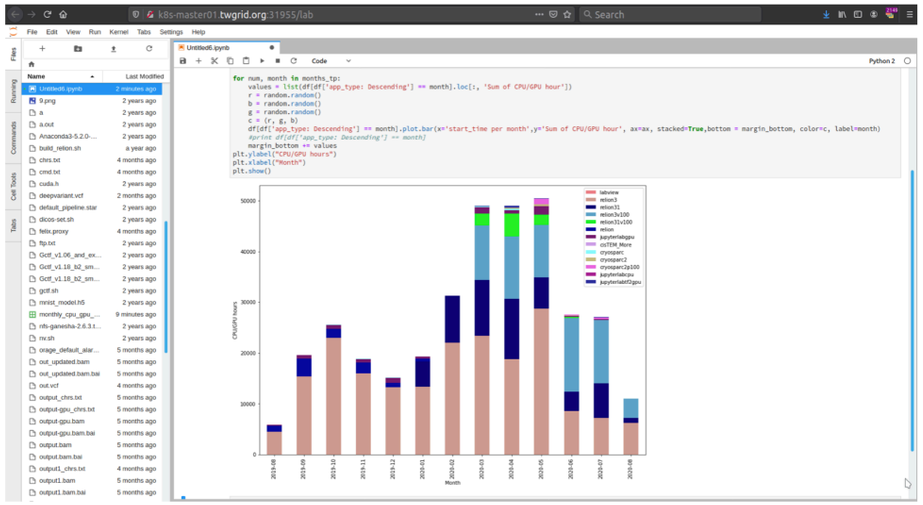
- **ML/AI application platform service is available NOW - SW library, HW, integration and application**
  - **Build up customized ML platforms for user specified projects - Deploy ML packages ready envirnonment in order to help ML development smoothly and provide on-demand computing power**
  - **Upkeep of the application framework**
  - **Workflow and data pipeline integration**
  - **Efficiency Improvement**
- **Potential use cases**
  - **Users who bring existing source code - ASGC could help to setup a virtual environment and confirm source code running normally**
- **Approaches**
  - **Supporting Kubernetes/Jupyter lab for development purpose**
    - **Create Kubernetes/Jupyter lab environment with user specified ML packages ready.**
    - **Support on-demand scalable CPU/GPU computing power.**
  - **Supporting containerized environment (e.g, Docker image) for deployment purpose**
    - **Create takeout images in Docker format as an option for user who wants to train/predict model**
    - **Docker images could be downloaded from ASGC server and deployed on users' Docker Desktop on Windows/Linux.**

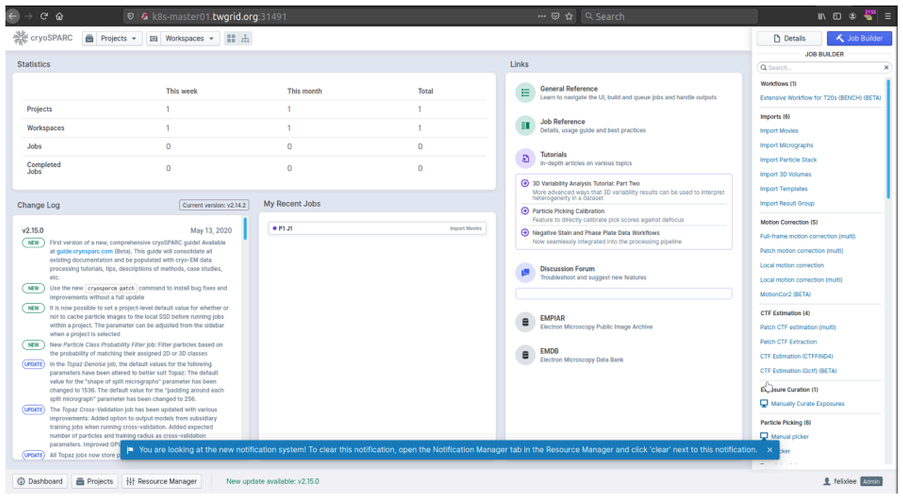
Available Hardware, Software & Use Cases

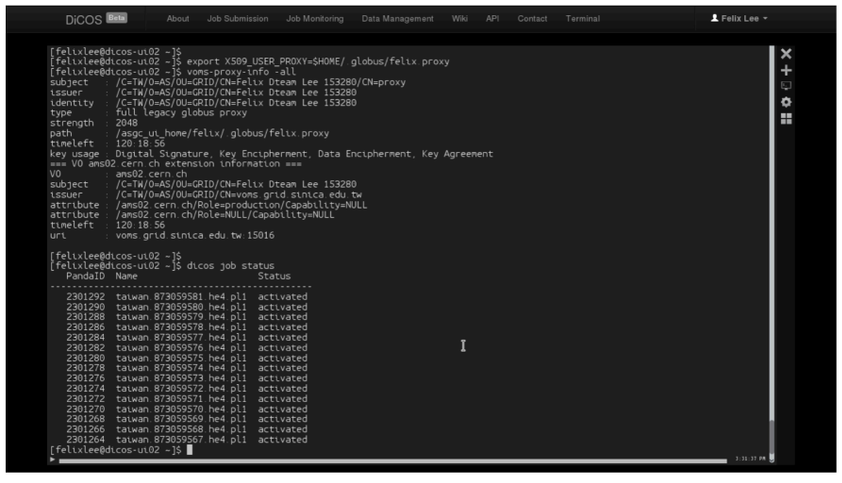
- GPU Servers (with local SSD enhanced)
  - A100 (8xboards/server, 80GB RAM/board) * 3
  - V100 (8xboards/server, 24GB RAM/board) * 6
  - 3090 (8xboards/server, 11GB RAM/board) * 4
- ML related framework and tools
  - TensorFlow, PyTorch, Keras, NVIDIA Triton, Scikit Learn
- Large-scale storage /file system
  - 9 Petabyte+ disk-based storage system managed by CephFS
  - Tape-based backup storage will be available by end of 2023
- Use Cases
  - CryoEM - ML-enabled bioimage processing
  - Deployment of ML-enabled protein simulation tools - AlphaFold, RosettaFold & Diffusion, DiffDock
  - Deployment of ML-enabled packages (by IOP PABS group): DeepMD-kit (with interface with LAMMPS)
  - AMS & KAGRA - programs developed by local groups
  - Data Center intelligent monitoring & control (ASGC projects): Air Handler, power saving, etc.

Supporting Big Data & AI in Innovations









CLI | **Web Portal** | **DiCOS APP** | **Jupyter Notebook** | **Science Portal** | **Web Browser/ Terminal**

Application-specific/ Generic Learning Engines

scikit learn

ROOT Data Analysis Framework

Deep Learning Engines/Frameworks

NVIDIA TRITON INFERENCE SERVER

PyTorch

TensorFlow

Keras

Computing Resource (Cloud/Grid/Slurm)

Storage Resource (Ceph/EOS)

Distributed Data Management & Cloud Storage Services

Network & Data Transmission Services

50+ Web Applications Provided

- Web Portal
- Application over Cloud
- Jupyterlab
- Web Terminal

PHYS

Deepmd-kit
Version: GPU with A100
Resources: 12%
Launch

Deepmd-kit
Version: GPU with V100
Resources: 80%
Launch

MAML
Version: GPU with A100
Resources: 12%
Launch

MAML
Version: GPU with V100
Resources: 80%
Launch

PVserver
Version: 5.8.0 (GPU 1080Ti)
Resources: 66%
Launch

Paraview Client
Version: 5.8.0
Resources: 97%
Launch

PyRoot
Version: GPU with 1080ti
Resources: 66%
Launch

Jupyter

Jupyter Lab
Version: CPU with Tensorflow v1
Resources: 97%
Launch

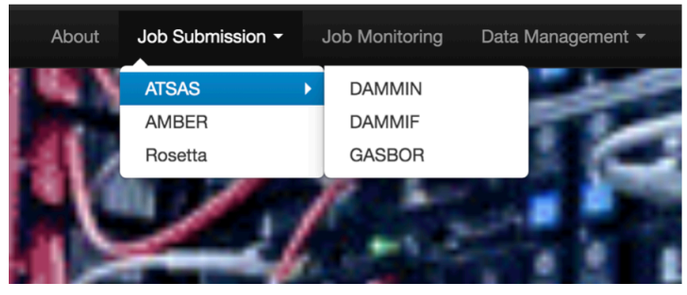
Jupyter Lab gpu 3090
Version: GPU with Tensorflow 3090
Resources: 51%
Launch

Jupyter Lab GPU V100
Version: GPU with Tensorflow V100
Resources: 80%
Launch

Jupyter Lab GPU A100
Version: GPU with Tensorflow A100
Resources: 12%
Launch

LabVIEW Run-Time Engine
Version: 2019
Launch

DiCOS-BioSAXS Platform



Other

spyder cpu/eman2
Version:
Resources: 97%
Launch

Octave
Version: V5.2
Resources: 66%
Launch

Transfer Data
Version:
Resources: 97%
Launch

cisTEM
Version:
Resources: 100.0%
Launch

Ovito
Version:
Resources: 97%
Launch

OpenACC
Version: GPU P100
Resources: 50%
Launch

Triton
Version: 22.01-py3 (GPU P100)
Resources: 50%
Launch

AlphaFold
Version: GPU with V100
Resources: 80%
Launch

AlphaFold
Version: GPU with A100
Resources: 12%
Launch

IMOD
Version:
Resources: 66%
Launch

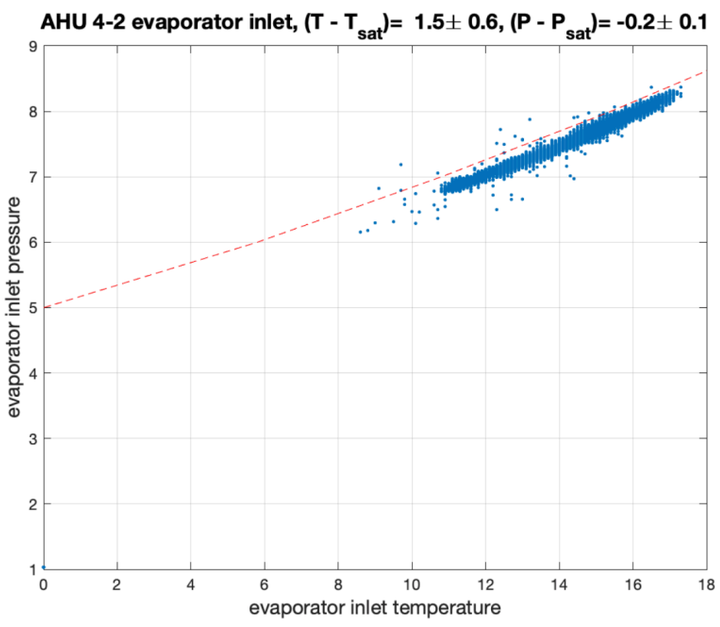
RoseTTAFold
Version:
Resources: 51%
Launch

Dynamo
Version:
Resources: 66%
Launch

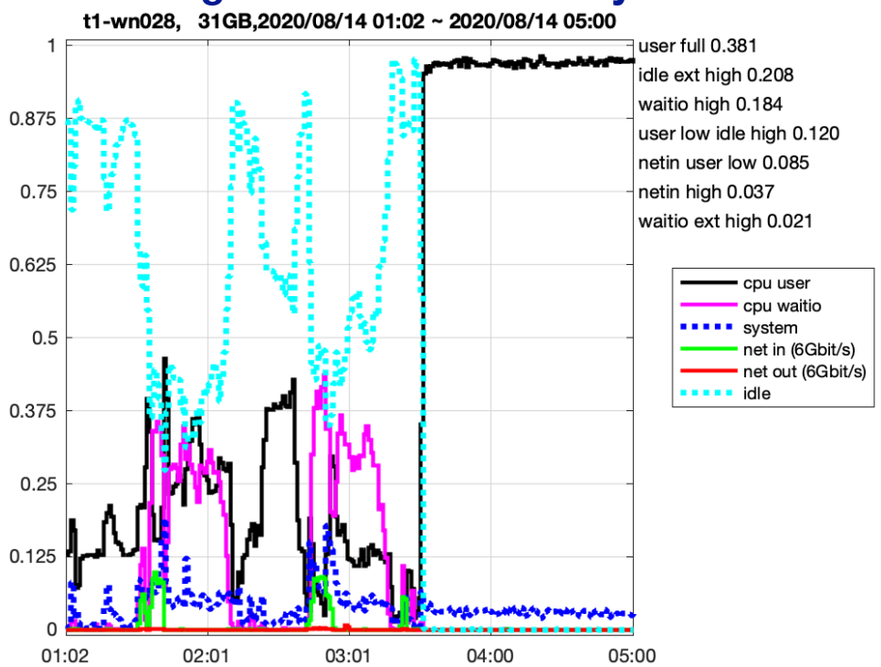
System Efficiency Optimization

- Goals: maximize application performance by available resources dynamically, in terms of power, thermal and system (Comp, Storage, Network, application) efficiency
- Scope: Power, Thermal, research infrastructure (Compute, Storage, Network), Cloud system, and applications
- Strategy: intelligent monitory and control assisted by ML
- Example: Thermal management, Compute/storage/network anomaly detection, Power saving of work nodes
- AHU monitoring and control
  - Detection of refrigerant operating issues and abnormal components; Efficiency optimization
  - 13 sensors x 16AHU; 18K data points/day;
  - Realtime monitoring, adjustment and diagnostics: refrigerant operating issue; abnormal components detection; efficiency tuning; ML-based automatic detection of critical problems;

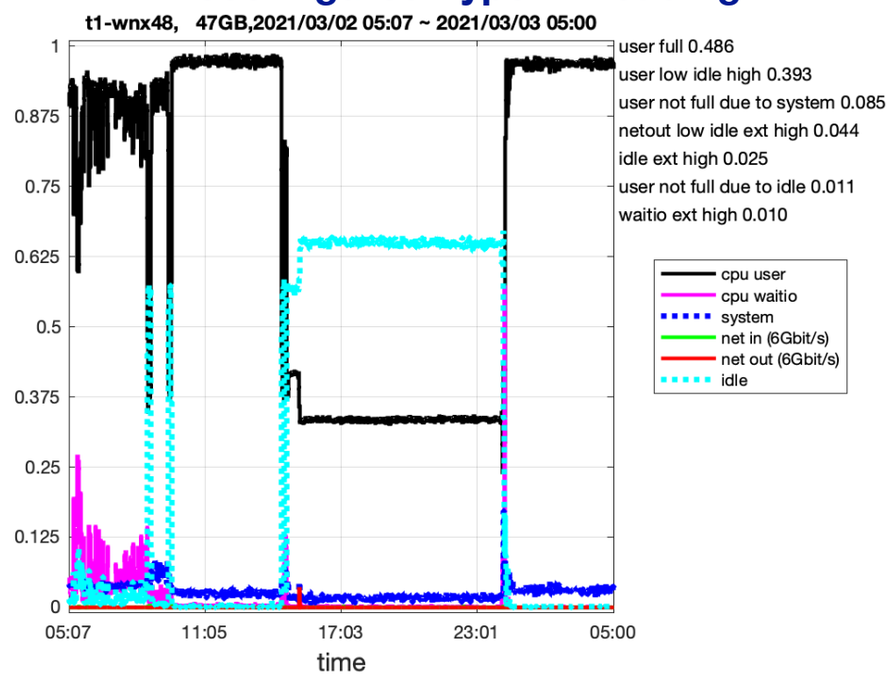
AHU Performance Monitoring



Worknode Monitoring: High ratio of WaitIO & System

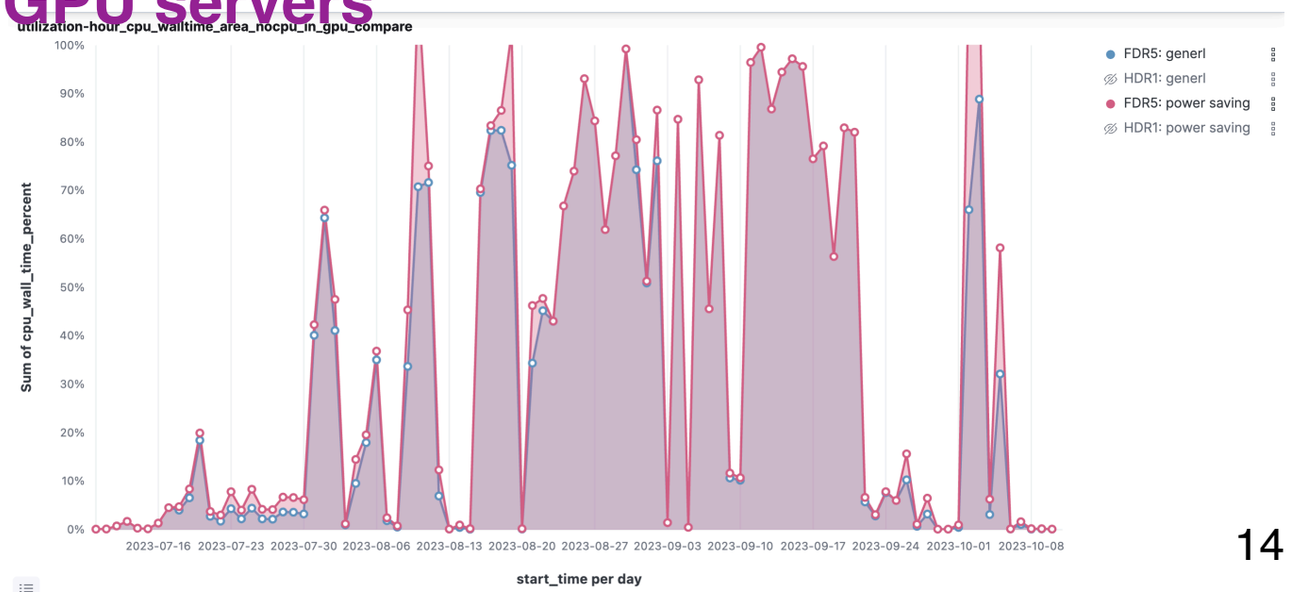


Worknode Monitoring: Misconfigured hyper-threading



Power Saving for Work Nodes

- Shutdown unused machines according to work load —> 10-20% power saved !
- Effective on 3 CPU clusters (> 3,000 CPUCores) from May 2023
- Turn on/off of a single work node
  - Decision making based on key metrics (e.g, utilization, waiting time)
    - Minimize the idle WNs without sacrificing job waiting time
  - according to Resource status, execution job status, queue status, etc., as well as historical data
  - Check each WNs (whole cluster) in every 15min
- Future work
  - Able to apply to other CPU clusters
  - Fine tuning the efficiency
  - Investigate the solutions to control GPU servers



Welcome To ISGC2024 in Taipei

- Schedule: 24-29 March 2024
- Venue: Academia Sinica, Taipei, Taiwan
- Call for Abstract/ Session will be open on 20 Oct. until 30 Nov 2023
- Event Web site: https://indico4.twgrid.org/event/33/
- Contact: ISGC Secretariat
  - vic@twgrid.org

ASGC Services

- ASGC Web Site: https://www.twgrid.org
- Access to ASGC Resources
  - https://dicos.grid.sinica.edu.tw/
- Contact point: DiCOS-Support@twgrid.org