HPC and AI :
Accelerating Discovery
Enabling Trust

NATIONAL INSTITUTES OF APPLIED RESEARCH
National Center for
High-performance Computing

NCHC Website

HPC @ NCHC

HPC and AI :

Accelerating Discovery

Enabling Trust

AI HPC in TAIWAN

Nation-wide Computing and Data Centers
Public-Private Partnership

National Science and Technology Council

TAIWANIA 2	9.0 PF
TAIWANIA 3	2.7 PF
Forerunner 1	3.5 PF
Taiwan Chip-based Industrial Innovation Program	Estimated 280 PF
NANO 5	13.06 PF
NANO 4	H200: 81.55 PF GB200: 4.5 PF
Southern Taiwan Silicon Valley Program	Estimated 200 PF

Public Sector

- Central Weather Bureau (dedicated)

Private Sector

- Wistron
- UBILINK
- FOXCONN
- AMD
- NVIDIA Taipei-1
- Konsttech
- Enlight AI Computing Center
- GMI Cloud
- Vantage Data Center
- ASE AI Energy Saving Data Center

Public Cloud

- Azure
- AWS
- Google

SOVEREIGN AI in TAIWAN

NCHC Strategy for Trusted AI

AI HPC Computing Power
AI from training to inference

Domestic Data
Traditional Chinese
Indigenous languages
Hokkien language

Model
Local deployment of models:
TAIDE, Gemma, Llama...

AI HPC in Taiwan
Public-Private Partnership

TAIWAN AI RAP
TAIWAN Resilient and high-performance AI Platform (RAP)

晶創26 NANO 4

Next generation AI Platform

TOP 500 29th | Green 500 73rd | GREEN 500

System Overview

2.21 MW

- **220 NVIDIA H200 Nodes + 2 NVIDIA GB200 NVL72 Racks**
- **Interconnect :** InfiniBand NDR 400
- **GPUs :** 1,760 NVIDIA H200 GPUs+ 144 NVIDIA Blackwell GPUs
- **Storage :** 25 PB
- **Cooling :** Direct Liquid Cooling
- **PUE :** 1.18

H200 :
HPL(FP64) : 117.92 PF(Rpeak) / 81.55 PF(Rmax)
AI-FLOPS(FP16) : 1740 PF(Rpeak)

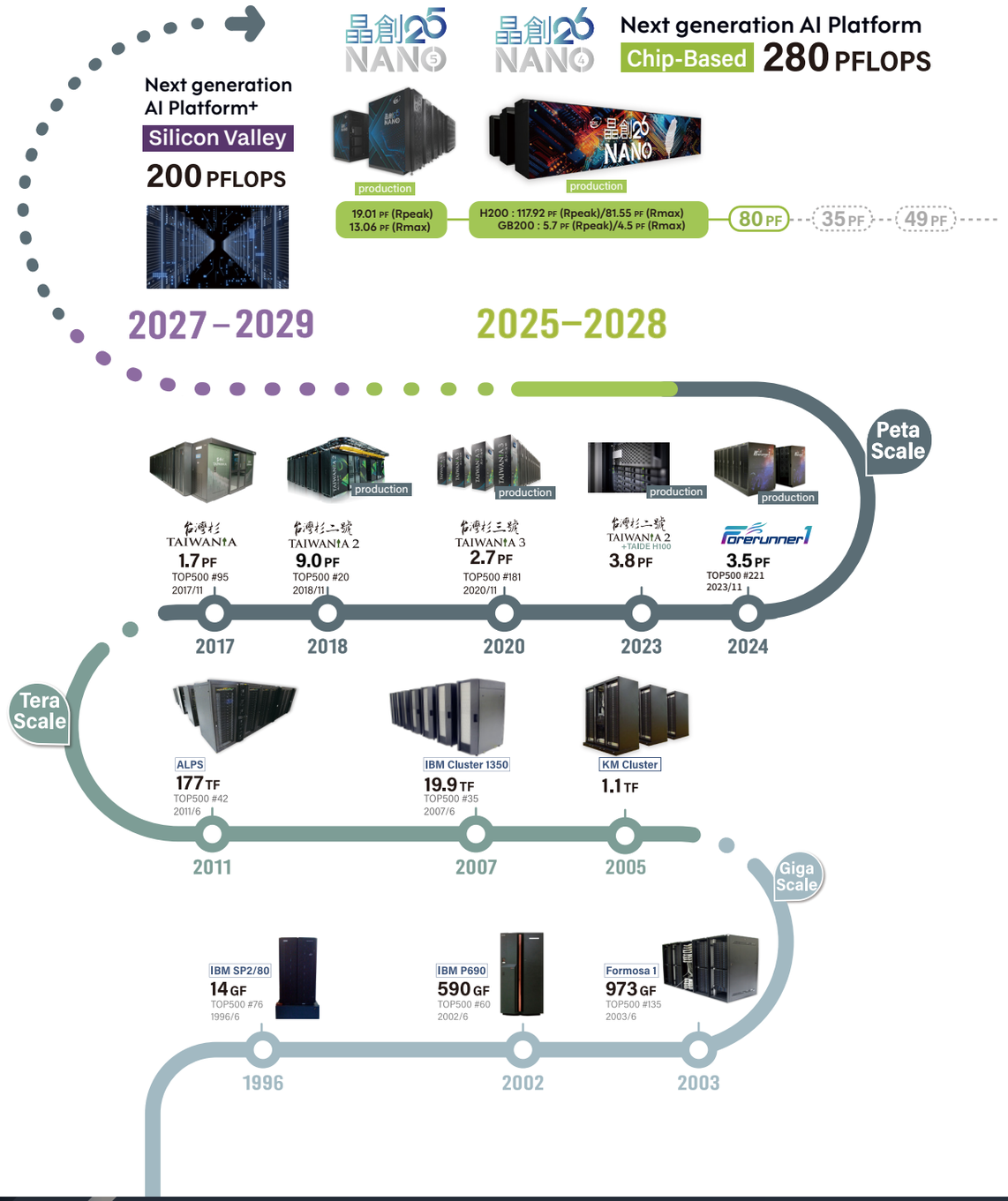
1) NVIDIA H200
CPU: Intel Xeon Platinum 8480+ x2
GPU: NVIDIA HGX H200 x8
Memory: DDR5 2TB

GB200 :
HPL (FP64) : 5.7 PF(Rpeak) / 4.5 PF(Rmax)
AI-FLOPS(FP16) : 360 PF(Rpeak)

2) NVIDIA GB200 NVL72 Rack
CPU：NVIDIA Grace CPU x 36
GPU：NVIDIA Blackwell GPU x 72
Memory：HBM3e 13.4TB

NCHC

HPC MILESTONES



National Large-Scale Computing Infrastructure Initiative

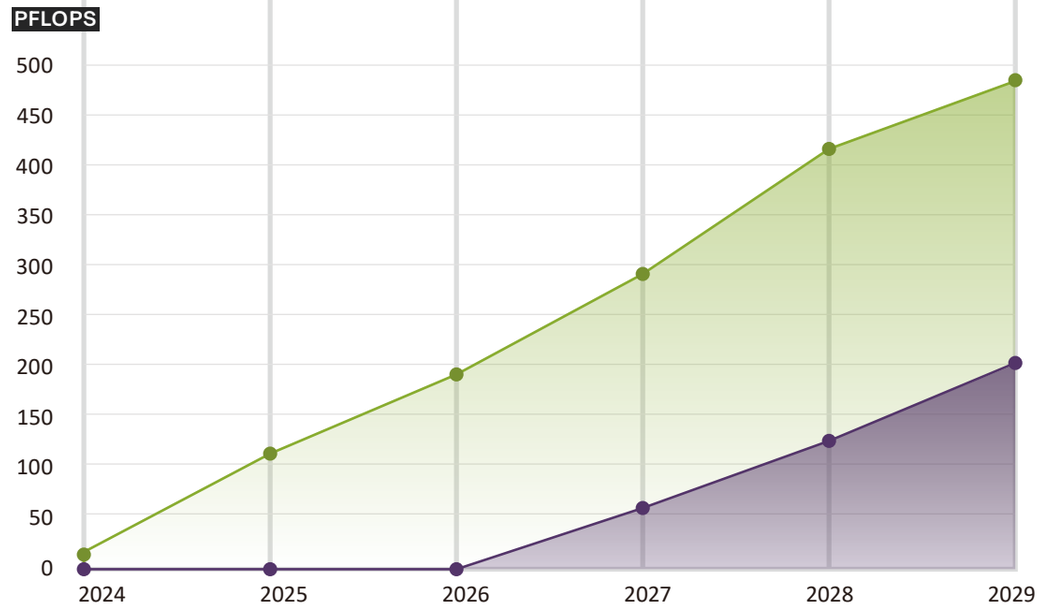
Taiwan CHIP-BASED Industrial Innovation Program

- Build a high-performance computing and AI evaluation environment
- Integrate AI, cloud, and IoT technologies
- Promote generative AI R&D and applications
- Develop a shared heterogeneous architecture supercomputer (GPU / CPU / future quantum computing)
- Create a user-friendly cloud service platform to improve application efficiency

Southern Taiwan Silicon Valley Program

- Develop the AI industry with Tainan Shalun as the core hub
- Integrate semiconductor, biomedical, precision machinery, and optoelectronic green energy industries
- Upgrade AI R&D and system development capabilities
- Promote industrial digital transformation and net-zero transition
- Increase national computing power to 480 PF

2024	2025	2026	2027	2028	2029
16 PF	116 PF	196 PF	291 PF	415 PF	480 PF
● 16 PF	● 116 PF	● 196 PF	● 231 PF	● 280 PF	● 280 PF
			● 60 PF	● 135 PF	● 200 PF
0.8 MW	5.1 MW	8.7 MW	13.6 MW	19.9 MW	23 MW

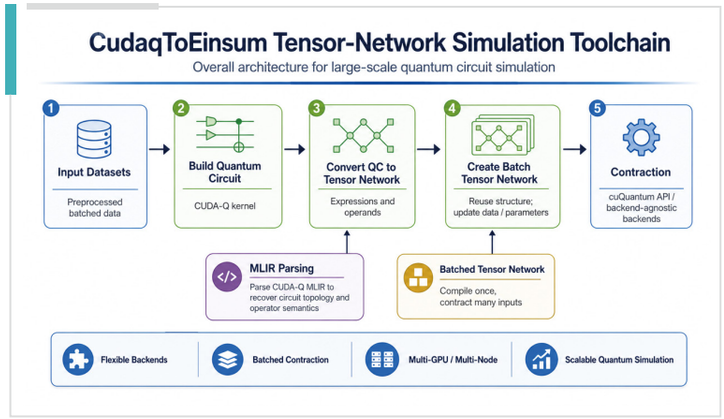


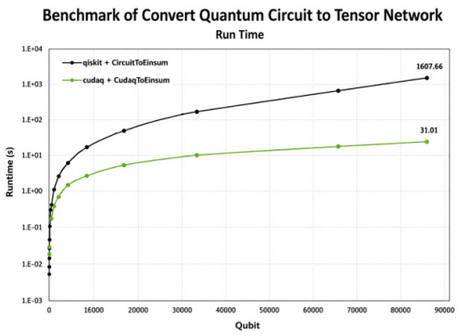
Efficient Tensor - Network Conversion
for Large-Scale Quantum Circuit Simulation

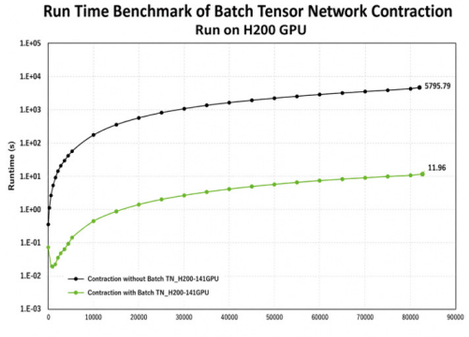
Tai-Yue Li | 2503001@niar.org.tw

NCHC is developing CudaqToEinsum, an efficient tensor-network simulation toolchain for large-scale quantum circuit simulation. It converts CUDA-Q circuits into flexible tensor-network representations by parsing CUDA-Q MLIR and generating standardized Einstein summation equations and operands. The generated tensor networks are backend-agnostic and compatible with cuQuantum, PyTorch, and other contraction backends.

With batched contraction, CudaqToEinsum can reuse the same network structure while updating only input data or circuit parameters, reducing memory usage and computational costs. By combining fast circuit conversion, batched contraction, and multi-GPU/multi-node HPC support, it provides a scalable platform for quantum simulation, quantum machine learning, and HPC scientific computing, helping researchers accelerate algorithm testing and prepare experiments for quantum hardware.







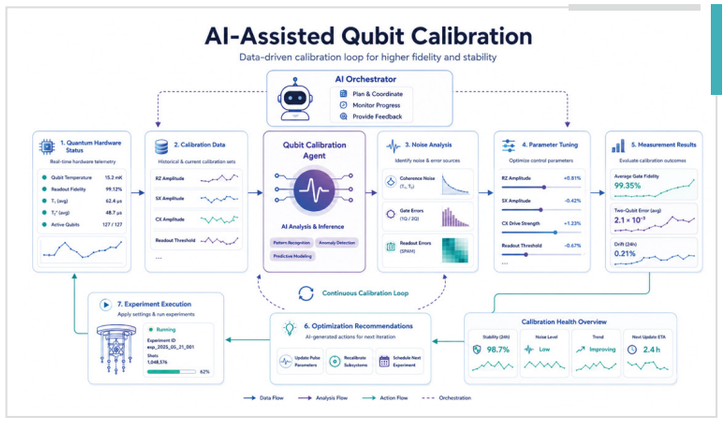
Agentic AI for Quantum Computing
Automating Quantum Workflows from Calibration to Execution

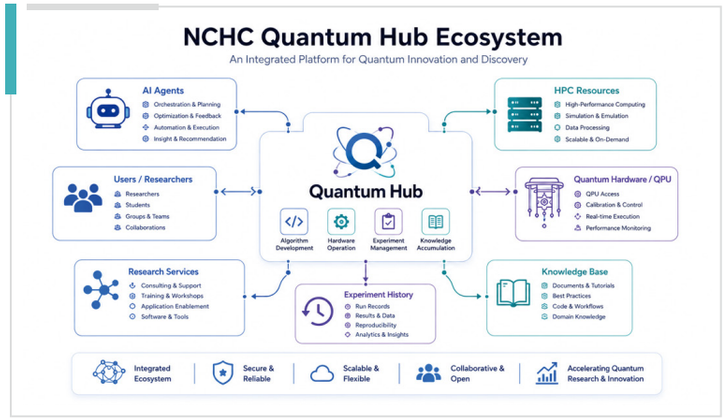
Tai-Yue Li | 2503001@niar.org.tw

NCHC is developing an Agentic AI workflow that combines high-performance computing (HPC), artificial intelligence (AI), and quantum computing to support more automated and intelligent quantum operations. The workflow uses AI agents to assist with algorithm design, parameter tuning, job execution, and result analysis, helping accelerate quantum research.

For qubit calibration, the workflow links hardware status, calibration data, and experimental results. AI agents can analyze qubit stability, noise behavior, and system performance, then suggest calibration adjustments and optimization strategies. This helps reduce manual work and improve experimental efficiency.

This workflow will serve as a key foundation for NCHC's future Quantum Hub. By integrating AI automation, HPC acceleration, and quantum infrastructure, NCHC can build a smarter platform for quantum algorithm development, hardware operation, and research services.



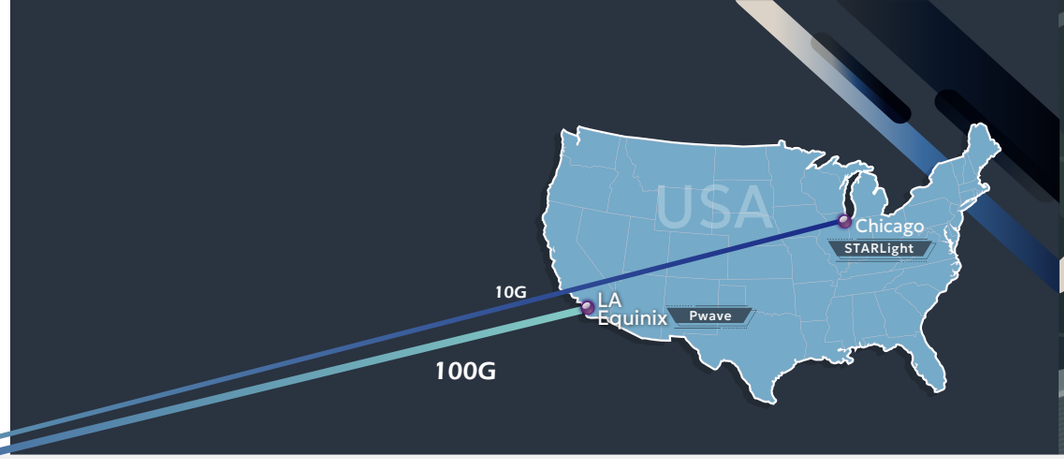


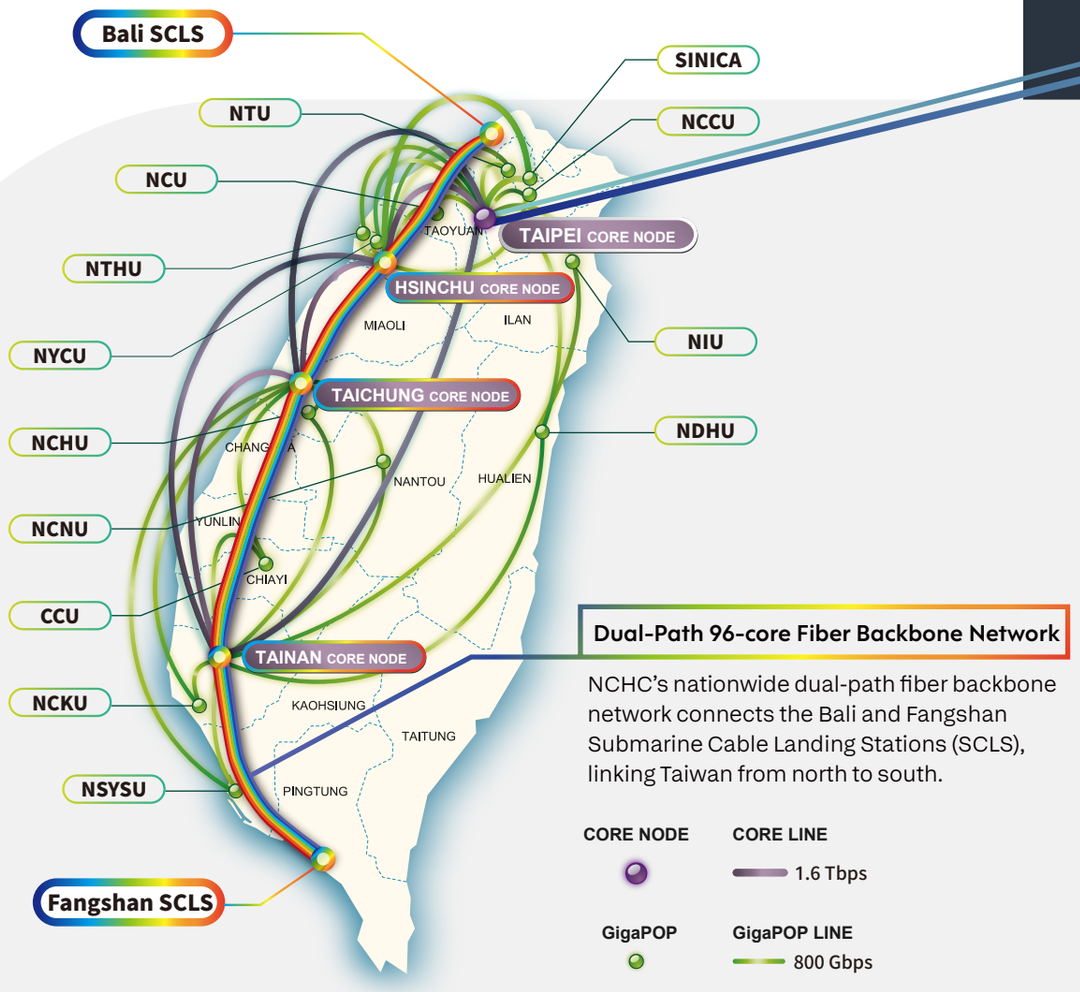
TWAREN

Accelerating Research Connectivity:

TWAREN's Leap to 400G

TWAREN's advanced 400Gbps network backbone delivers up to 1.6Tbps nationwide connectivity, with 110Gbps links to Los Angeles and Chicago.





TWAREN takes a major step forward with its next-generation 400G backbone upgrade, delivering higher speed and resilience for research and innovation. Enhanced with a 3D Digital Twin Dashboard, the new architecture enables real-time monitoring and smarter network management.

TWAREN's international circuits have also been upgraded to 110G, strengthening global connectivity, while Taiwan's nationwide fiber backbone interlinks north and south through terrestrial and submarine cables for a more reliable research infrastructure.

IOWN APN

lower power consumption | ultra-low latency high-capacity transmission

NCHC is an Academic & Research member of the IOWN Global Forum

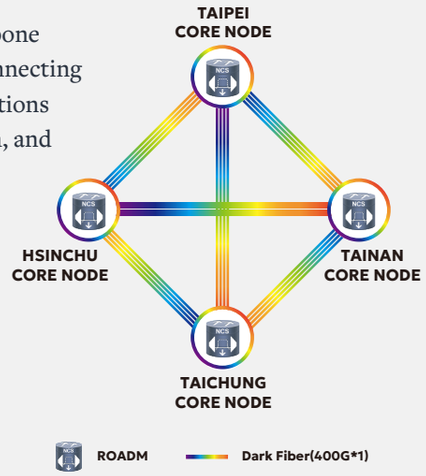
NCHC IOWN APN–Aligned All-Photonics Network Architecture

Integrating All-Photonics Networking for AI, HPC, and Smart City Applications

NCHC is deploying an IOWN APN-aligned optical network by integrating its nationwide fiber backbone with the TWAREN research network. The all-photonic architecture provides ultra-low latency, high bandwidth, and energy-efficient transmission for AI, HPC, and smart city applications. Through a distributed DCI framework and regional core nodes, NCHC enables scalable and resilient cross-domain collaboration while strengthening international connectivity and next-generation digital infrastructure.

NCHC's IOWN Architecture for Smart City Applications

NCHC has deployed a dual-path, 96-core fiber backbone network along the Taiwan High Speed Rail track, connecting the Bali and Fangshan Submarine Cable Landing Stations located at the northern and southern ends of Taiwan, and linking to key hubs.



TWAREN NGI: IOWN APN–Ready

TWAREN NGI (Next-Generation Infrastructure) plans to launch a 400Gbps NREN service in 2026. The network is being designed as IOWN APN-ready to support future high-capacity, low-latency research networking.

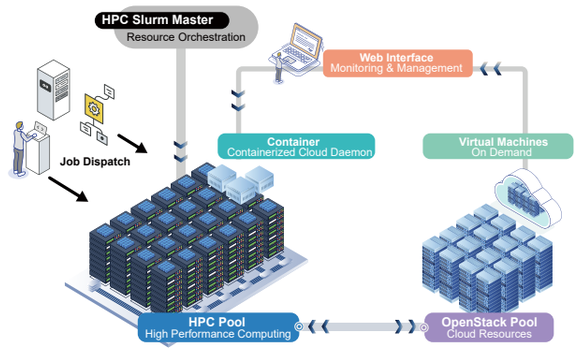
Dynamic GPU Resource Orchestration Between HPC and Cloud in Data Centers

Chia-Chuan Chuang | jimmy_chuang@niar.org.tw

This architecture presents a **Slurm**-centered architecture for sharing GPU and compute capacity between traditional HPC workloads and cloud-style services. Rather than treating these environments as separate resource silos, the system extends the role of Slurm beyond batch scheduling and uses it as the control point for node transition. In this way, existing HPC infrastructure can respond more flexibly to changing demand without abandoning familiar operational practices.

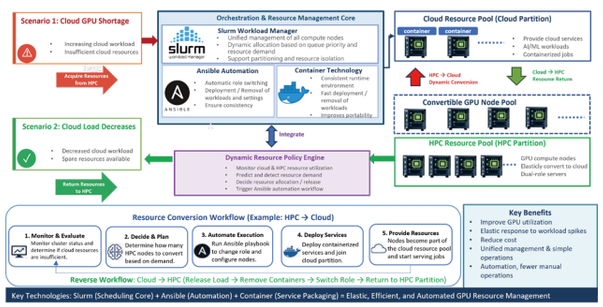
The main contribution is a controlled and reversible handoff mechanism governed by Slurm. Nodes are reassigned only after they have been explicitly allocated through the scheduler. When additional service capacity is needed, selected nodes can be repurposed; when demand decreases, they can be returned cleanly to the Slurm cluster for normal scientific workloads.

The key advantage of this approach is that it improves utilization without weakening operational safety. It reduces the time and manual effort required to bring extra capacity online, supports consistent deployment across environments, and provides clear integration points for local policies and site-specific procedures. For HPC centers facing dynamic AI and simulation demand, this architecture offers a practical way to make Slurm-managed infrastructure more adaptive, efficient, and valuable.



Dynamic GPU resource orchestration between HPC and cloud environments

Centered on Slurm, this work enables HPC nodes to transition between batch scheduling and cloud services. The architecture improves GPU utilization, preserves scheduling discipline reduces manual reconfiguration effort, and helps data centers adapt shared infrastructure more efficiently to changing service demands.



Dynamic Compute Resource Alloation in Data Center

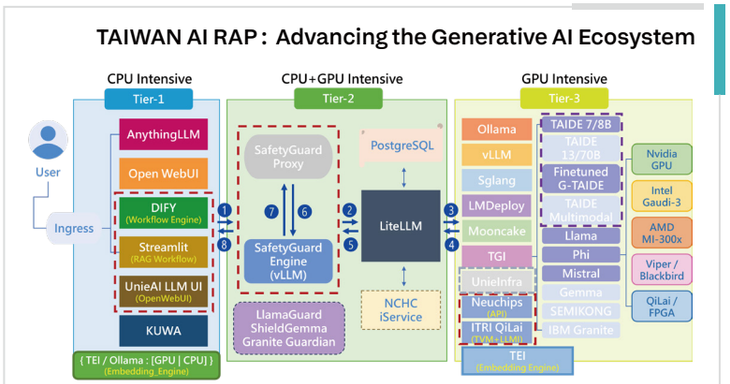
When cloud GPU capacity is insufficient, HPC GPU nodes are automatically converted to cloud resources.

When cloud demand decreases, resources are returned to HPC.

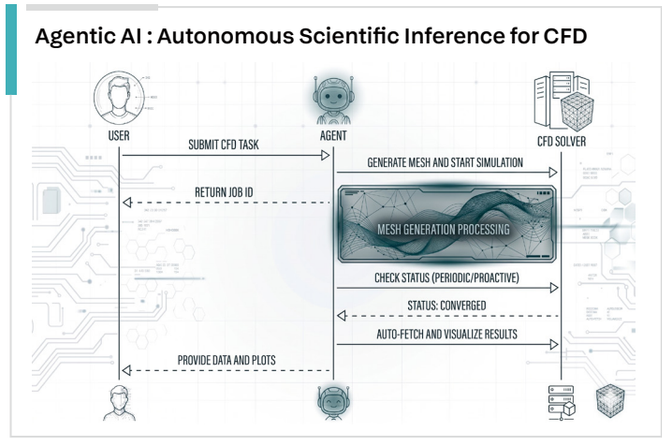
Agentic AI driven Inference Service for Scientific Applications - using numerical simulation as example

Shao-Jia Fan | 2503127@niar.org.tw

TAIWAN AI RAP is a next-generation platform driving Taiwan's leadership in generative AI. It integrates robust computing infrastructure with a user-friendly, no-code workflow designer, enabling the rapid creation of customized AI services. By offering a unified multi-model API and advanced fine-tuning pipelines, RAP lowers technical barriers and shortens development cycles. It serves as a cornerstone for innovation, fostering collaboration between research and industrial sectors to scale AI solutions effectively.



Introducing an AI-driven Model Context Protocol (MCP) server for Computational Fluid Dynamics (CFD). Integrated with OpenWebUI, this agentic AI autonomously orchestrates solving strategies and precise parameter configurations. By analyzing real-time simulation convergence and dynamically refining mesh density through data-driven feedback, the agent revolutionizes scientific inference, ensuring high-fidelity results while significantly optimizing complex engineering workflows.



Workload Forecasting and Scheduler Optimization Framework for HPC System

Weicheng Huang | whuang@niar.org.tw, whuang91362@gmail.com

Yen-Zhen Lai†, Weicheng Huang*, Wei-Chu Wang †, Te-Min Chen ♦, Ying-Yu Shih ♦ , Li-Wei Shih ♦
† : Dept. of System Engineering and Naval Architecture, National Taiwan Ocean University
♦, * : National Center for High-performance Computing, National Institutes of Applied Research

A data-driven mechanism that integrates workload forecasting with adaptive optimization is attempted to improve scheduler performance. Historical workload data is modeled using XGBoost to predict future demands. An enhanced optimization algorithm, PSO$^+$, is then adopted to dynamically adjust scheduler parameters for better resource partitioning to reduce overall job turnaround time.

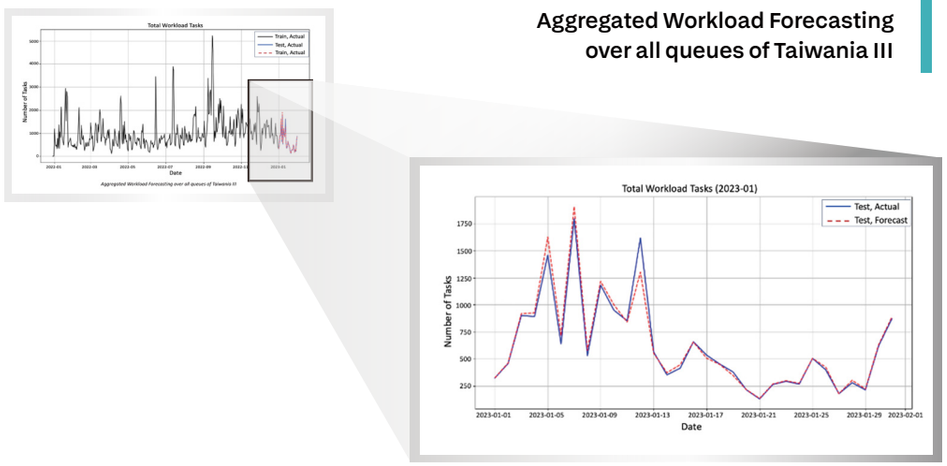
Experimental results demonstrate improved utilization and scheduling efficiency, highlighting the effectiveness of combining machine learning and optimization for HPC workload forecasting.

Data

The workload data is acquired from Taiwania III of NCHC with 6 job queues. This time series data is normalized for both temporal and resource space, via z-score.

Forecast of workload

The forecasting model leverages XGBoost with Optuna to optimize hyperparameters and to find the case with lowest RMSE and MAE. The data is partitioned into training and testing portion with ratio of 11:1.



Aggregated Workload Forecasting over all queues of Taiwania III

Prediction error of workload

Queue	Actual jobs	Predicted jobs	Error	Error%
ctest	1047	851	-196	18.72%
ct56	11622	12111	489	4.21%
ct224	4925	4647	-278	5.64%
ct560	1343	1330	-13	0.97%
ct2k	564	422	-142	25.18%
ct8k	219	151	-68	30.05%
Total	19720	19512	-208	1.05%

Optimization of scheduler configuration :

The PSO$^+$, PSO with “clustering” and “escaping” strategy, is chosen, after testing various optimization schemes, for the task, with objective as minimizing total job waiting time. The optimization involved 16 design variables. The results yield an overall performance improvement over 21%, with recommended configurations over initial ones.

Total Job waiting time before and after optimization (minutes)

Queue	Original	PSO$^+$	Improvement	Improvement %
ctest	2,351,728,040	2,344,139,138	7,588,902	0.32%
ct56	67,047,176,856	54,273,824,292	12,773,352,564	19.05%
ct224	37,917,792,346	26,551,429,051	11,366,363,295	29.98%
ct560	4,325,723,595	4,639,894,004	-314,170,409	-7.26%
ct2k	1,770,910,971	1,624,889,683	146,021,288	8.25%
ct8k	235,262,620	326,353,985	-941,091,365	-38.72%
Total	113,648,594,428	89,760,530,153	23,888,064,275	21.02%

Initial vs. Optimized scheduler configurations

Queue	ctest		ct56		ct224		ct560		ct2k		ct8k	
optimization	Init	PSO$^+$	Init.	PSO$^+$	Init.	PSO$^+$	Init.	PSO$^+$	Init.	PSO$^+$	Init.	PSO$^+$
MaxJobPU	2	2	50	69	25	63	15	61	6	8	2	9
GrpJobs	15	21	200	238	40	56	30	31	18	26	4	6
Min. cores available			1	1	57	195	225	310	561	1104	2241	4264

Future works (international partners for FL)

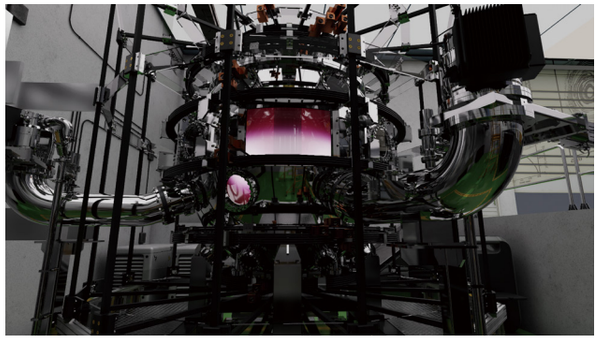
To further enhance the performance of this approach, the authors would like to seek the possibility of international collaboration via Federated Learning to enhance the privacy and security of data in the collaboration.

Toward a Digital Twin for Taiwan's FIRST Spherical Tokamak: Modeling and Visualization

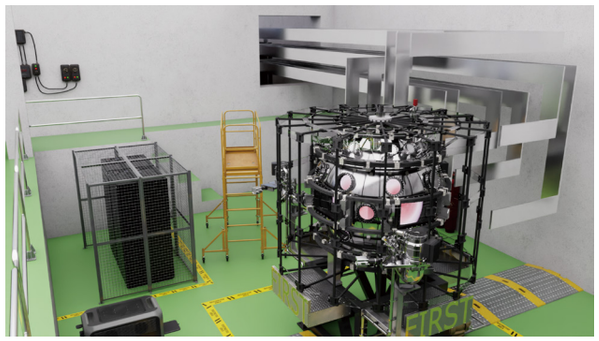
Tsung-Che Tsai | tctsai@niar.org.tw

This work also extends the tokamak digital twin beyond static modeling. JOREK simulation data have already been integrated into the platform, enabling physics results to be explored in an immersive visual environment. In the future, the framework will be expanded to incorporate GTC, GENE, and experimental data, forming a more comprehensive digital twin for visualization, analysis, planning, and collaborative research.

The Omniverse-based digital model of the FIRST tokamak laboratory is developed from the tokamak engineering model and rendered with materials that reflect the appearance of the actual chamber and surrounding structures. By combining geometric accuracy with realistic visualization, it provides an intuitive digital environment for presenting the device and its laboratory context, while also supporting equipment layout planning and spatial arrangement assessment.



Close-up Omniverse rendering of the FIRST tokamak digital model, showing the engineered chamber geometry, surrounding support structures, and realistic material appearance.



Overview of the Omniverse-based FIRST tokamak laboratory digital twin, presenting the device within its laboratory environment for visualization, layout planning, and spatial assessment.

Digital Twin Tokamak Video

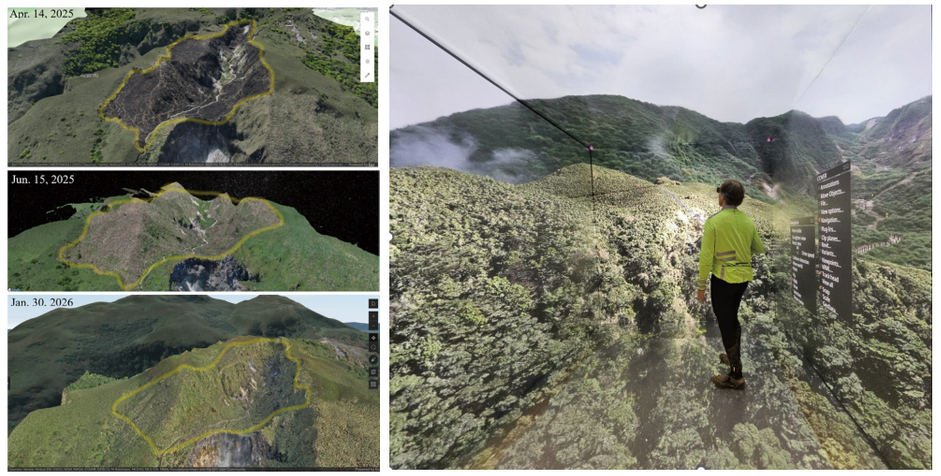
Digital Twin of National Park - Study of Environmental Sustainability

Hsiu-Mei Chou | hmchou@niar.org.tw

Restoring the unique ecosystem of Yangmingshan's Xiaoyoukeng is a race against time. Following a recent brush fire, recovery is challenged by acidic volcanic soil and volatile weather. Monitoring this process requires a balance between broad observation and scientific depth.

To meet this challenge, we are building a sophisticated digital twin of Xiaoyoukeng. Using High-Performance Computing (HPC), we fuse high-resolution UAV imagery with traditional expert ground survey data. By integrating manual field observations to "ground-truth" our virtual model, we avoid the complexity of permanent on-site hardware while ensuring scientific accuracy.

This hybrid approach allows our advanced AI to recognize and classify plant species from above, distinguishing native pioneers from invasive threats. By transforming aerial data and field expertise into actionable intelligence, we are documenting the return of life to ensure a resilient future for Taiwan's iconic volcanic landscape.



Supercomputing for Nature's Survival. We are building a living digital twin of the Xiaoyoukeng volcanic site to monitor its recovery following a recent brush fire. By harnessing massive computing power, we can track how the landscape heals, protect emerging wildlife, and shield our ecosystems from future climate threats.

AI Computing Center

The NCHC Cloud Compute Center (C³), located in the Tainan Science Park, is a next-generation computing facility designed to support the growing demands of AI, HPC, and data-intensive research. With enhanced computing and storage capabilities, energy-efficient infrastructure, and sustainable design, C³ serves as a national computing hub for both academia and industry. By providing scalable computing resources, C³ accelerates scientific discovery, AI innovation, and digital transformation, helping strengthen Taiwan's position in the global technology landscape.

Cloud Compute Center (C³)

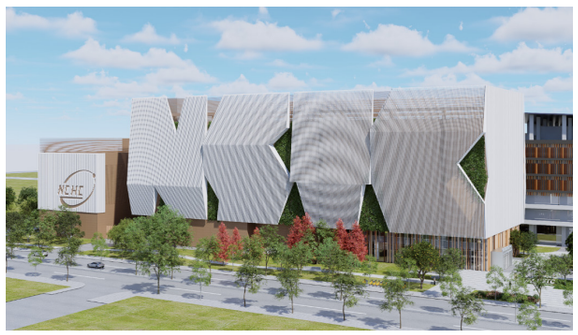
A national strategic hub for AI factories with broadband network resources

- 15MW Power Capacity
- Resistant to magnitude 6 earthquakes
- Floor Load Capacity：1,450 kg/m²
- Green building (silver level certificate)
- Average power usage effectiveness (PUE) below 1.3
- Located in Tainan Science Park

Compliant with the government policy of proximity to power sources for data centers

Intelligent Innovation Hub

- 90MW Power Capacity
- Earthquake Resistance
- Floor Load Capacity: 2,500 kg/m²
- Located in Shalun, Tainan



Scheduled to launch in 2029