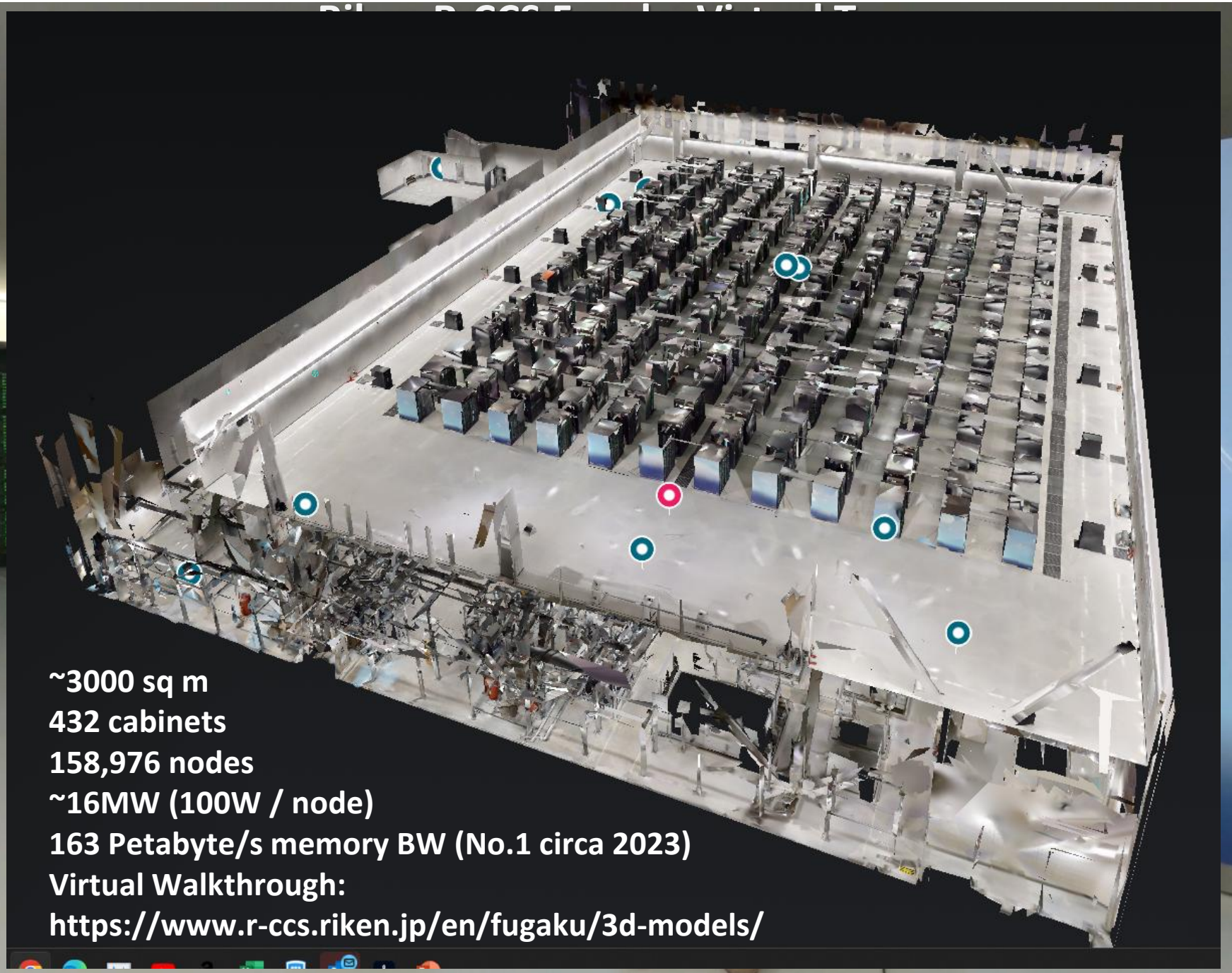
The Convergence of HPC, AI, and Quantum: Towards cutting-edge eScience on FugakuNEXT

Satoshi Matsuoka, Director Riken R-CCS
And many other collaborators @ R-CCS and elsewhere
Australian eResearch Conference
Oct. 22, 2025

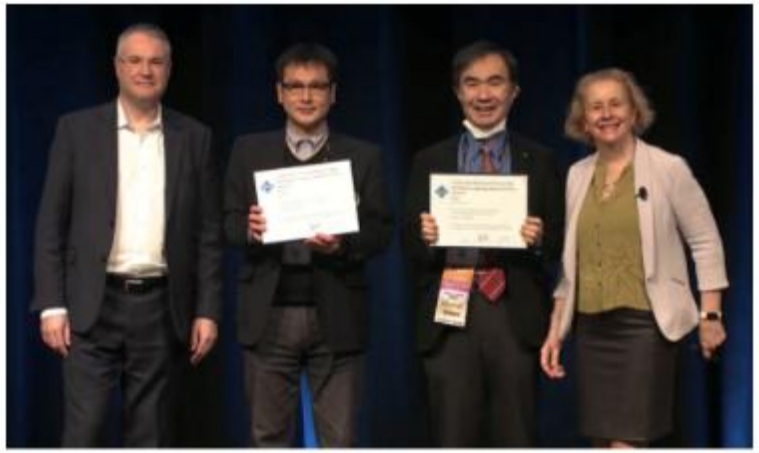

~3000 sq m
432 cabinets
158,976 nodes
~16MW (100W / node)
163 Petabyte/s memory BW (No.1 circa 2023)
Virtual Walkthrough:
https://www.r-ccs.riken.jp/en/fugaku/3d-models/

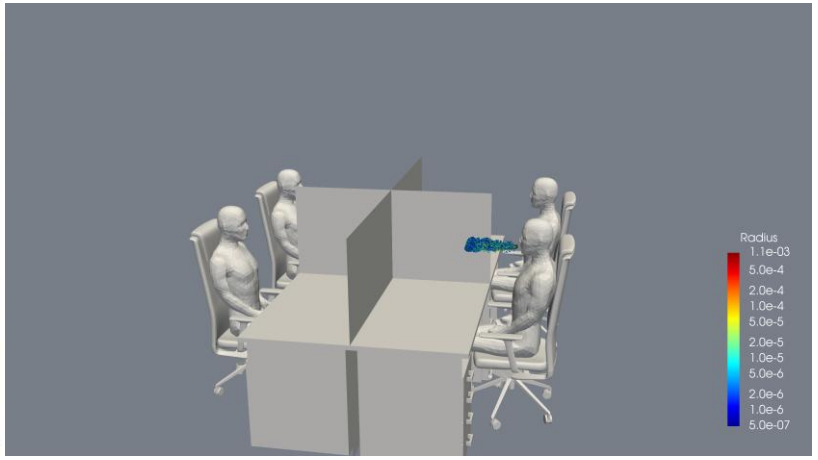
Major achievements of Fugaku

#1 in major benchmark rankings:TOP500 and HPL-AI(Jun.2020-Nov.2021), Graph500 and HPCG (Jun.2020-)

#1 in MLPerf HPC(Nov.2021-)

ACM Gordon Bell Special Prize for HPC based COVID-19 research(Nov.2021), also 2022





Weather forecasting trial for “guerrilla downpour” in TOKYO2020 Olympic/Paralympic games

2023 Hyperion Report on Fugaku Values (2025 report forthcoming to include AI for Science)

#1 Research Finding: Fugaku Will Likely Return 68 to 90 Times Its Costs

The Fugaku potential returns are very strong

1. **The potential economic value:**
   - $15 billion from projects like those that were done on the K system ($4 billion plus has already been accomplished on 6 projects)
   - $50 to $75 billion from keeping Japan from shutting down its economy
   - $10 to $22.5 billion for large value industrial projects
   - And a potential of $22.5 billion or more from addressing important SDG goals

- **For a total of $102 to $135 billion in financial value – this represents a return of 68 to 90 times the investment in Fugaku**

#2 Research Finding: Researchers Are pleased with The Design and Operations of Fugaku

The Fugaku potential returns are very strong

2. **The percentage of the researchers that like the Fugaku system design and operations is one of the highest seen in our studies with only a few that aren't pleased with the system design.**
   - Most sites around the world typically have only 60% to 75% of the researchers pleased with their system design & approach.

2025 report for FugakuNEXT

Expect > 100x ROI

#3 Research Finding: Fugaku Is Focus On High Value SDG's

Fugaku researchers are addressing a broad set of SDG's

Projects in these areas include:

- Disaster prevention, resilience to urban wind disasters and heat islands, wind resistance safety of bridges, realization of Society 5.0, availability of large-scale computers and entry of non-professionals into computation, increased international competitiveness in automobiles/manufacturing, safe behavior criteria for COVID-19, preventing spread of COVID-19, drug discovery, research and development of new materials, new products, fuel cells, efficiency in combustor and furnace design, and the efficiency of large offshore wind power generation.

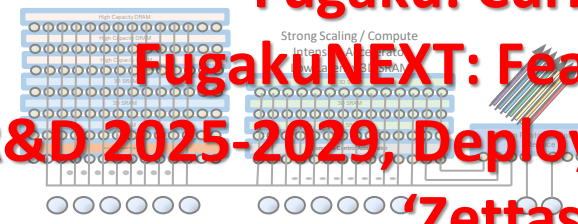
#4 Research Finding: Fugaku Is Focused On Creating Industrial Economic Growth

By directly supporting industry with a strong outreach program

4. **Fugaku is more focused on supporting industrial growth and helping companies create economic value vs. focusing more heavily on pre-competitive R&D. Riken has a strong industrial outreach program which is more industry-friendly than most other nations.**
   - The focus is more directly on increasing Japanese companies' economic growth and competitiveness (and not only on longer term R&D).

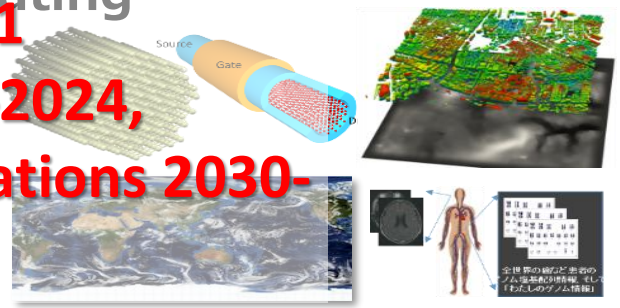
Riken R-CCS Strategy for Innovation by Computing Future of Science 'of' and 'by' Computing

- Science **of** High Performance Computing (towards 'Zettascale')

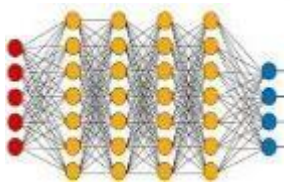


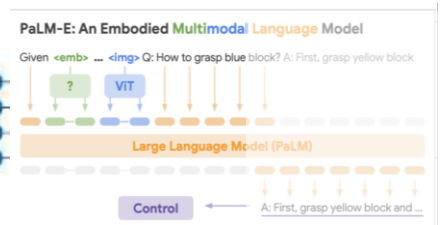
Fugaku: Current until 2030~2031
FugakuNEXT: Feasibility Study 2022-2024, R&D 2025-2029, Deployment ~2029, Operations 2030-
'Zettascale' @ 40MW

- Science **by** High Performance Computing



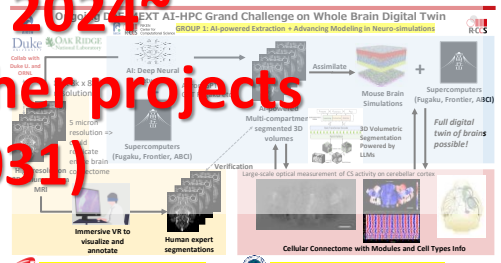
- Science **of** High Performance AI

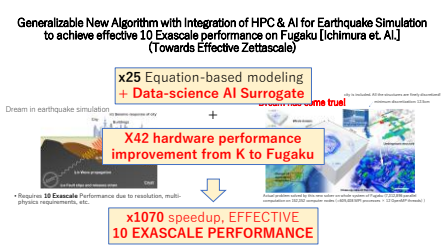




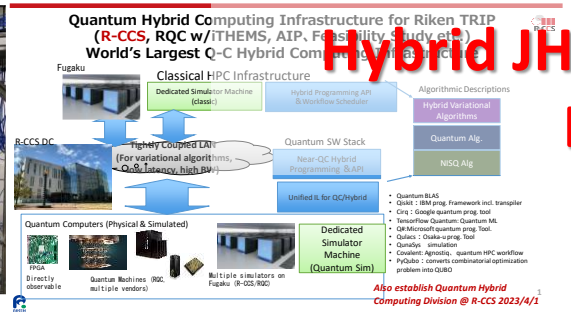
Riken AI for Science FY 2024~
including TRIP-AGIS and other projects
(TRIP-AGIS 2024~2031)

- Science **by** High Performance AI (AI for Science) w/HPC Simulations



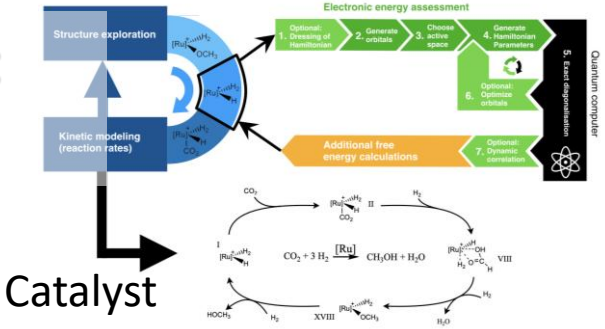


- Science **of** Quantum-HPC Hybrid Computing



Hybrid JHPC-Quantum Infrastructure Project
Deployment FY2023~2027

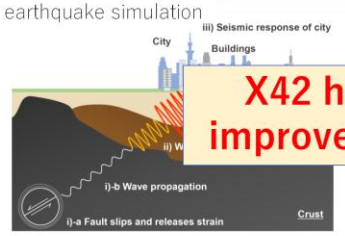
- Science **by** Quantum-HPC Hybrid Computing




Generalizable New Algorithm with Integration of HPC & AI is developed to achieve effective 10 Exascale performance

x25 Equation-based modeling
+ Data-science app

Dream in earthquake simulation

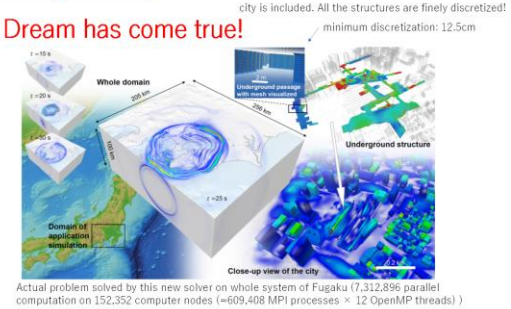


+

X42 hardware perf
improvement from K

- Requires **10 Exascale** Performance due to resolution, multi-physics requirements, etc.

Dream has come true!



Actual problem solved by this new solver on whole system of Fugaku (7,312,896 parallel computation on 152,352 computer nodes (=609,408 MPI processes × 12 OpenMP threads))

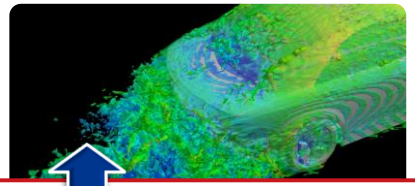
x1070 speedup, EFFECTIVE
10 EXASCALE PERFORMANCE

CFD Framework for Co-Satisfiaction of Aerodynaic Drag Efficiency & Design Aestheics [Tsubokura et. al.]

R-CCS

- **Co-optimization Framework**

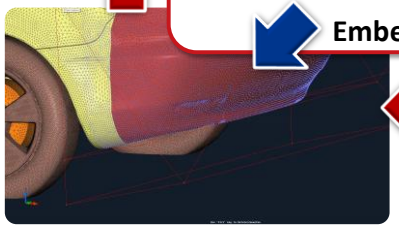
Rapid Generation of CFD Mesh from Shape Data



スーパーコンピュータ"富岳"

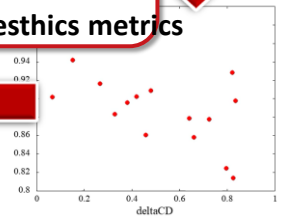
Ultra Fast Prediction of Drag via Digital Twin
AI-Based Prediction and Optimization
Embedding of human aesthics metrics

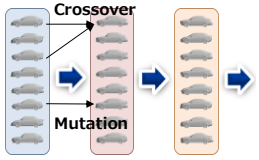
Drag + Aestheics



Parametric Shape Morphing

Shape Parameters on Aesthetics

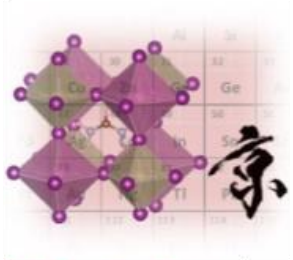




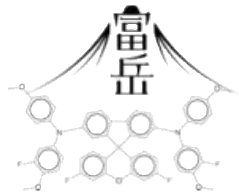
GA Multi Paramter Optimization "CHEETAH/R"

Development of Perovskite Solar Cell Materials through the Integration of Physical Simulations on Fugaku and AI Surrogates (Nakajima, R-CCS)

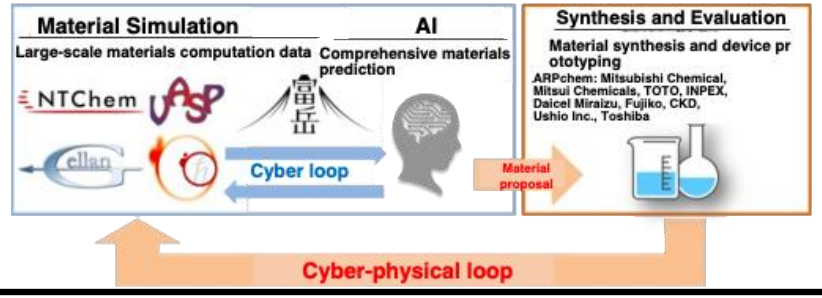
Material simulations on the K supercomputer enabled the design of novel materials for organic solar cells and photocatalysts. Building on this knowledge, we aim to realize higher-efficiency organic solar cells and photocatalysts through advanced simulations on Fugaku combined with surrogate AI models, paving the way for industrial-scale implementation



During the era of the K supercomputer, simulations alone were used to screen 11,025 compounds, leading to the proposal of 51 low-toxicity perovskite solar cell candidates.



Cyber Technologies **Physical Technologies**



Conversion Efficiency: 24.4%

In the Fugaku era, the integration of big data, AI, and simulations enabled the design of high-efficiency hole transport materials from among millions of candidates.

Leveraging Generative AI for Scientific Innovation in Drug Discovery
AI-Driven Drug Discovery Powered by Fugaku Simulations

NNMT Inhibitor and Experimental Activity Measurement
Kanazawa University: Hirao & Arakawa

Docking Simulation
Hit and Analog Compounds Including
Hit and Analog Compounds Including RK-9074594

Rigid-Body and Molecular Dynamics (MD) Calculations

Candidate Complex Structures

Quantum Mechanical (QM) Calculations

AI-Driven Drug Discovery Powered by Fugaku Simulations

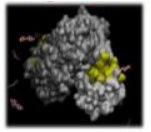
Chemical Synthesis of Compounds
Construction of a Synthesizable Compound Library
Kanazawa University: Kunishima

Analysis of NNMT/Compound Binding Modes Using Fugaku Supercomputer

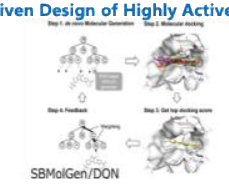
Design of Highly Active Compounds via Deep Reinforcement Learning
RIKEN R-CCS ·U. of Kyoto : Dr. Okuno

AI-Driven Modeling of Structure–Activity Correlations

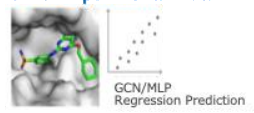
ON/OFF-Target Binding Evaluation via Molecular Dynamics (MD)



AI-Driven Design of Highly Active Compounds



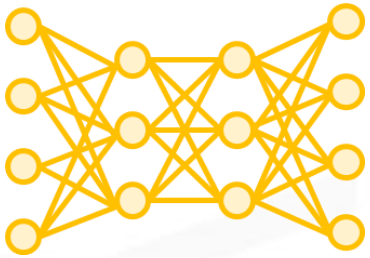
Determination of Complex Structures Consistent with Experimental Data



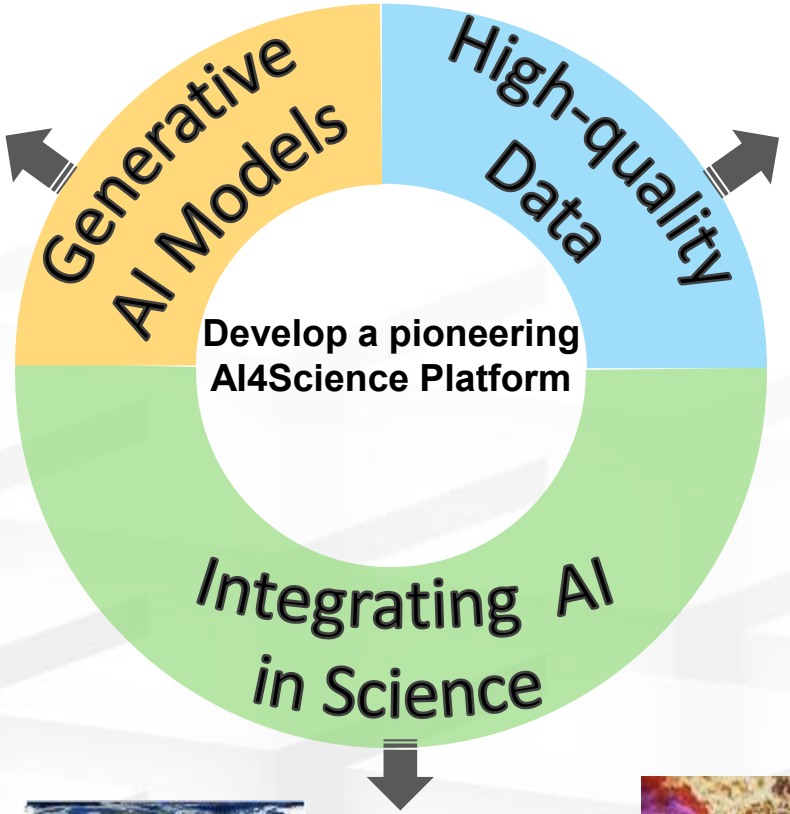
RIKEN's Initiatives ～TRIP-AGIS～

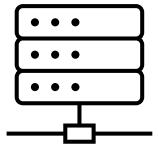
Artificial General Intelligence for Science of Transformative Research Innovation Platform (TRIP-AGIS)

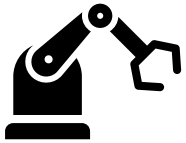
- ✓**TRIP-AGIS will introduce the technology of generative AI and will develop generative AI models for scientific research to further accelerate the research cycle.**
- ✓**Strengthen activities to lead advanced science to social impact**



Develop and share generative AI models for scientific research (life and medical sciences, climate science, engineering)







Simulations Experiments Robots

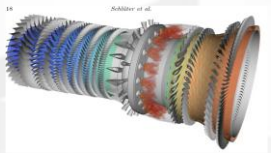
Produce large amounts of high-quality data through RIKEN's and its parternerships/collaborations. Strengths in measurement techniques and experiment automation

Purpose and Challenge

- **Solve intractable science problems**
- **Lead advanced science**
  - **Starting from basic science**
  - **To societal impact (GX, inclusive society, etc.)**

Physical/Earth

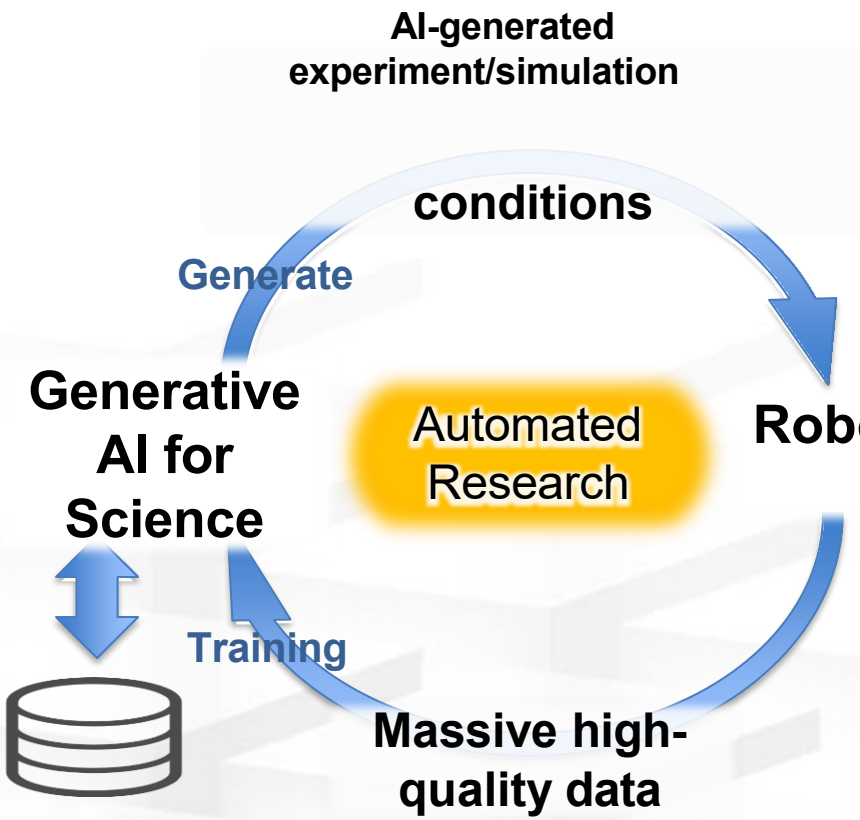
Life/Medical



Engineering

Riken TRIP-AGIS AI4S Project

AI-driven automatic research and massive data production using robotic experiments and large-scale simulations to create Science Foundation Models



Parallel robot lab.

Large-scale simulation by Fugaku

Acceleration of scientific research by AI

NAGOYA UNIVERSITY

Towards Foundational Models for Structural Engineering

[Koji Nishiguchi](Nagoya-U/Riken R-CCS)

Innovating vehicle structure with a giant aluminum die-casting

30% weight reduction
40% manufacturing cost reduction

Giga-press (Tesla)

3D generative AI (Parameter-to-3D model) for nonlinear structural engineering

Magic3D （NVIDIA, 2022）

Shap-E （OpenAI, 2023）

Rapid performance improvement of 3D generative AI

NAGOYA UNIVERSITY

Final goal: Automation and democratization of structural design

Human feedback by Non-experts

Designer **Marketer**

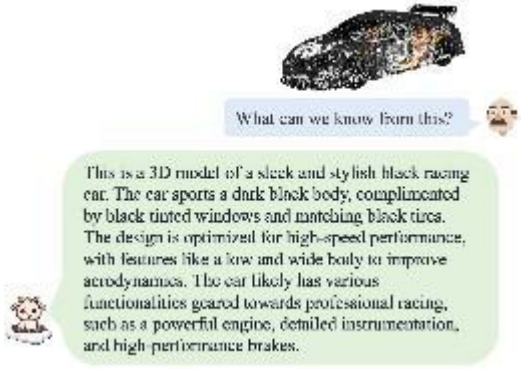
Natural language

Mechanical parameters

3D structure

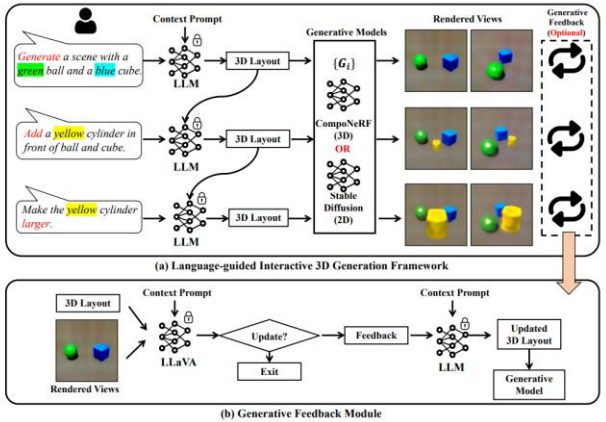
Natural language

3D-to-text model
(LLMs to understand 3D structure)



Human feedback

Text-to-parameter model
(LLMs as Parameter Interpreter)



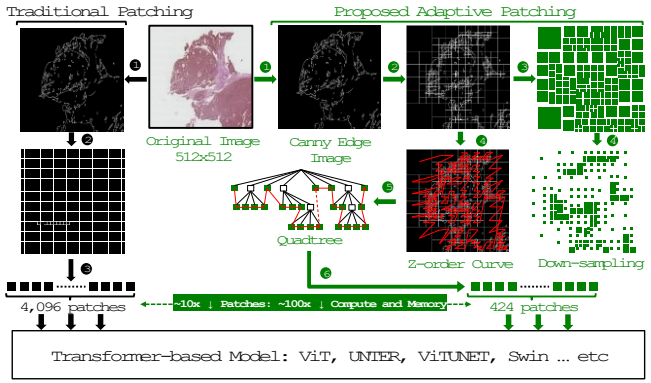
Paramter-to-3D model

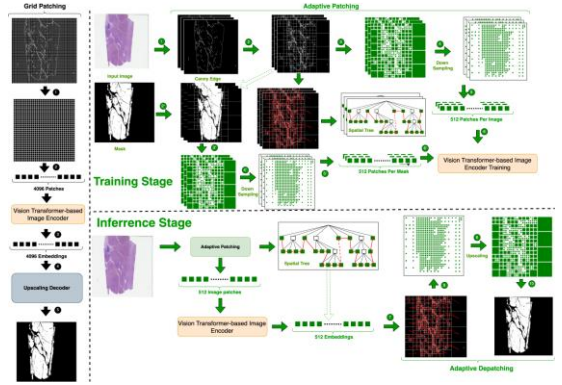
Adaptive Patching: Spatial, Temporal, Physics-inspired [2/3]

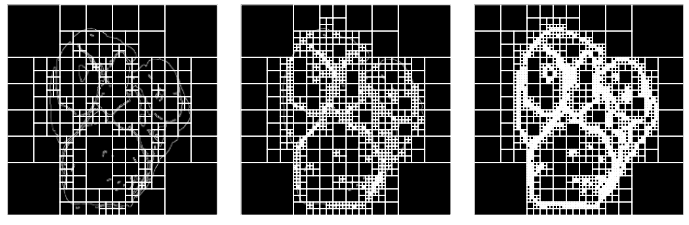
Summary of findings/results

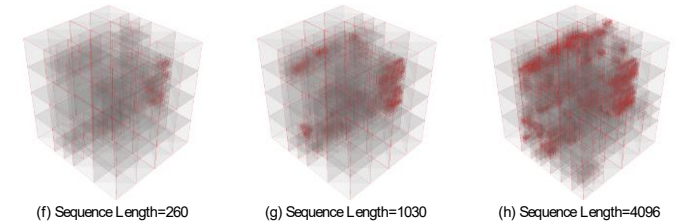
- High-quality segmentation on real-world datasets of WSI.
- 12.7x speedup; +7% classification accuracy for microscopic pathology

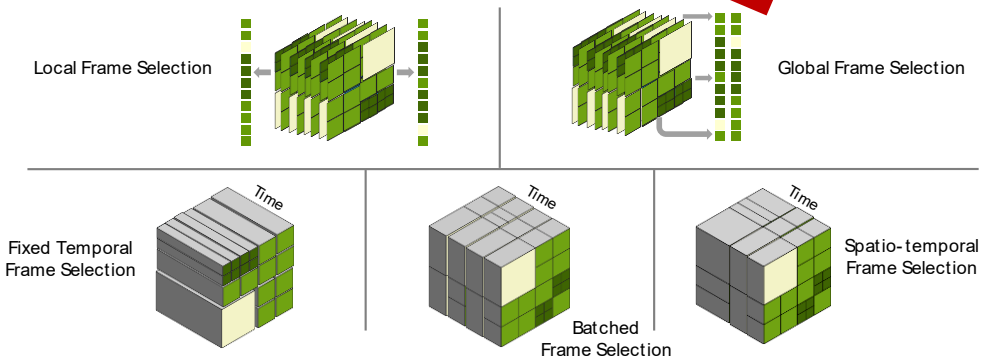
- Expanding to temporal and and physics-inspired adaptive patching

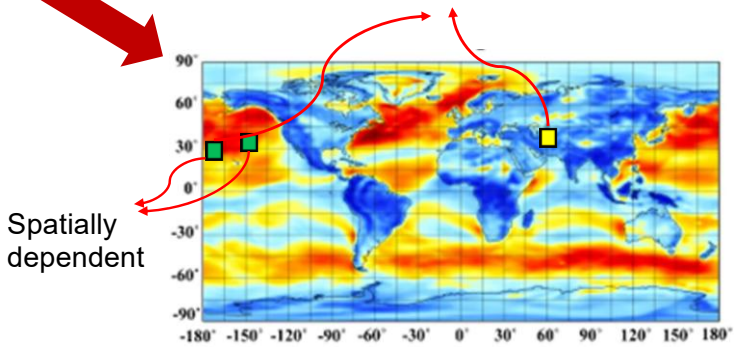








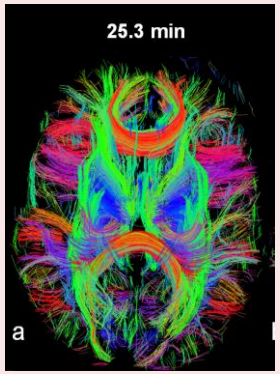




ORBIT Foundation Model

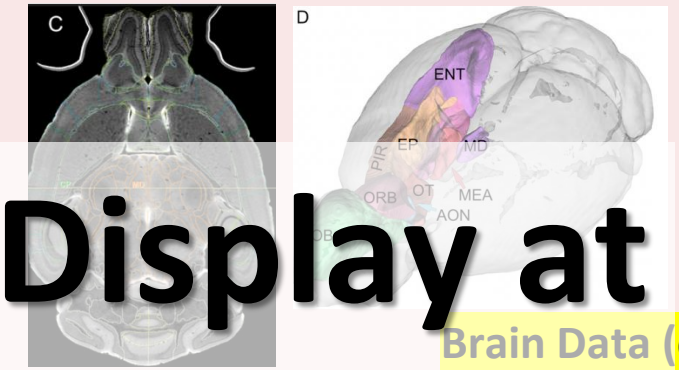
Resolution of Mouse-brain MRI Images using Advanced AI on Fugaku & ABCI

1-2mm Human connectome, and atlas (HCP)

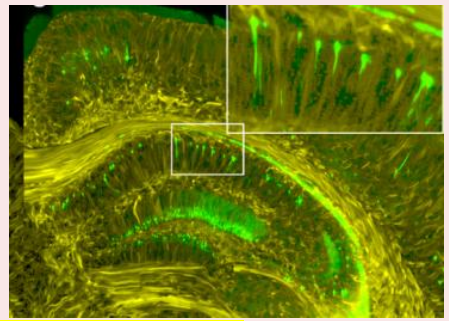


100 micron connectome (Knox et al., 2019)

15 micron iso voxel dMRI (Johnson et al., 2022)



5 micron tractography and single cell level registration with LSM (Johnson et al., 2023)



Brain Data (collab. ORNL/Duke U.)

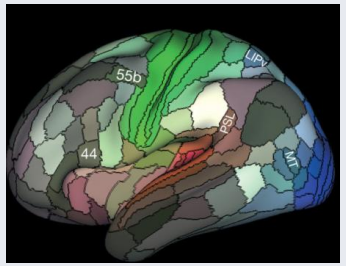
>100 microns

100 microns

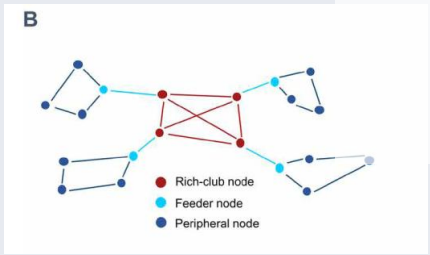
15 microns

<5 microns

On Display at EXPO 2025 @ Osaka



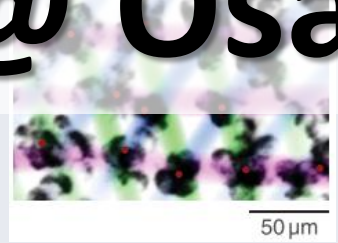
Classified regions (Glasser et al., 2016)



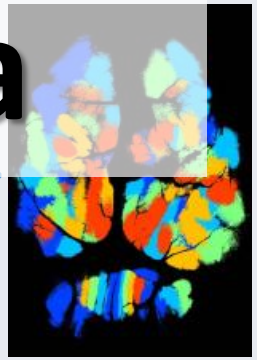
Network structure (Colleta et al., 2020)

Find Functional Module Structure

Calcium Imaging

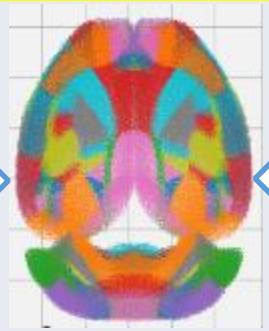


Columns (ex: Maruoka et al., 2017, Zeeuw et al., 2020)



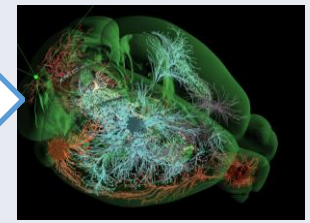
Michikawa et al., in prep

Brain Simulation



Igarashi et al., in prep

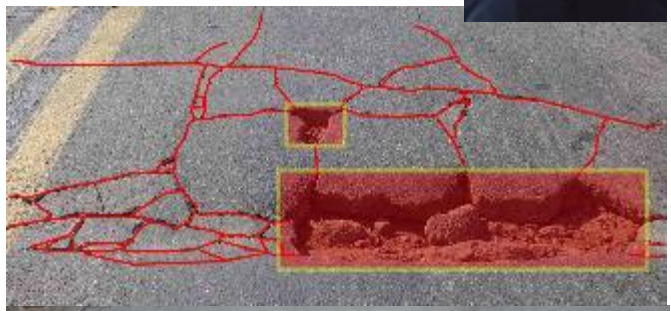
Cellular Connectome



EXPECTED OUTPUT IN THIS PROJECT

Capability of Understanding Mouse Brain, full brain simulation of Fugaku

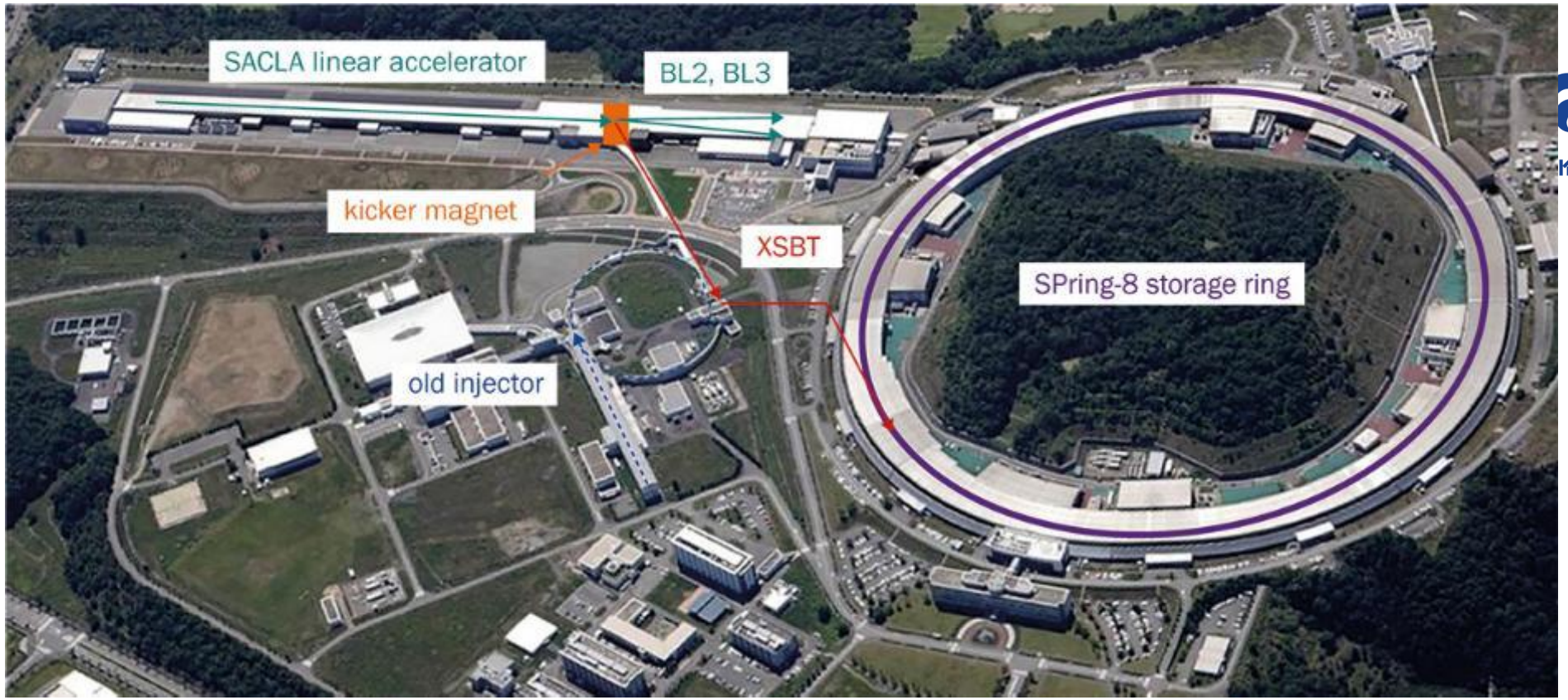
Another Real-world Problem: How to Inspect Roads for Maintenance?



- Machines mounted on vehicles
- Extract cylindrical samples from core of asphalt layers
- Scan (projections) at RIKEN Spring-8 Synchrotron
- Move projections to R-CCS (or other HPC facilities) **for HPC-AI processing**
  - High-performance high-resolution CT image reconstruction
  - 3D volumetric segmentation (~$8K^3$)
- Provide resulting data for experts to analyze

- **Radically changes how road infrastructure is inspected**

Can Imaging + HPC + AI Solve this Intractable Problem?



RIKEN Center for Computational Science

SPring-8

KYOTO UNIVERSITY FOUNDED 1897

阪神高速 Hanshin Highway Co.

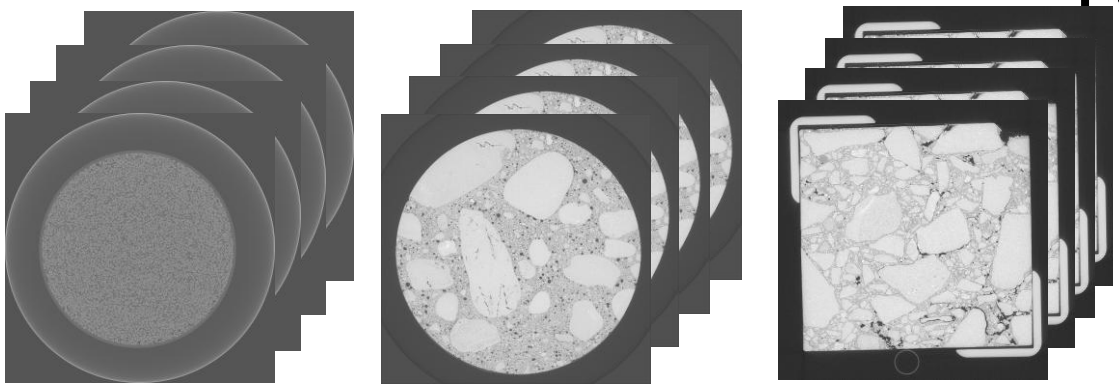
Riken Spring-8 + Sacla Synchrotron Light Source Facility

Analyzing and Solving the Science of Infrastructural Decays [Wahib et.al.]

Sand particles (4Kx4Kx6K)

Concrete & Asphalt (6Kx6Kx11K)

Concrete (6Kx6Kx3K)

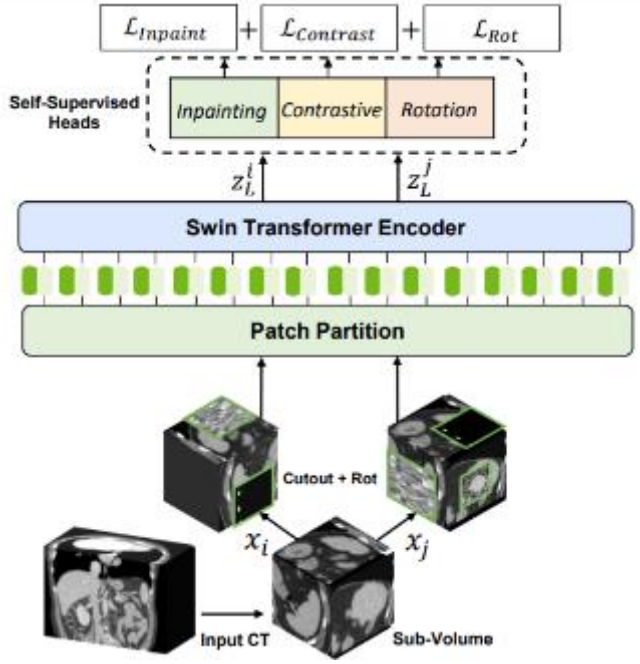


End-to-end High-resolution CT Powered by Supercomputing

+

富岳

Supercomputers (Fugaku/ABCI/ Frontier/AWS)



3D Volumetric Segmentation Powered by LLMs

(Image from https://developer.nvidia.com/blog/novel-transformer-model-achieves-state-of-the-art-benchmarks-in-3d-medical-image-analysis/)

State-of-the-art scale of resolution

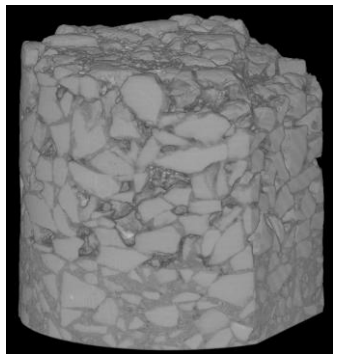
LLM powering 3D segmentation technology at unprecedented level of detail and accuracy

Reconstruction + AI + Analytics

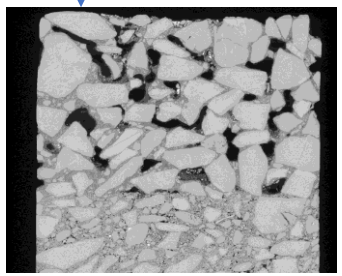
↓ Cost: O($ Billions)

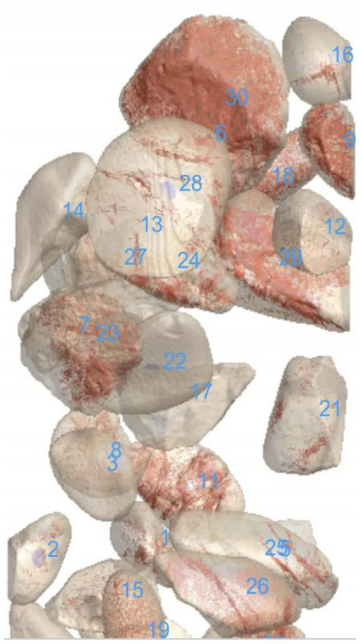
Contribution Towards Center Mission (X-ray CT) [3/3]

CT撮影



内部断面





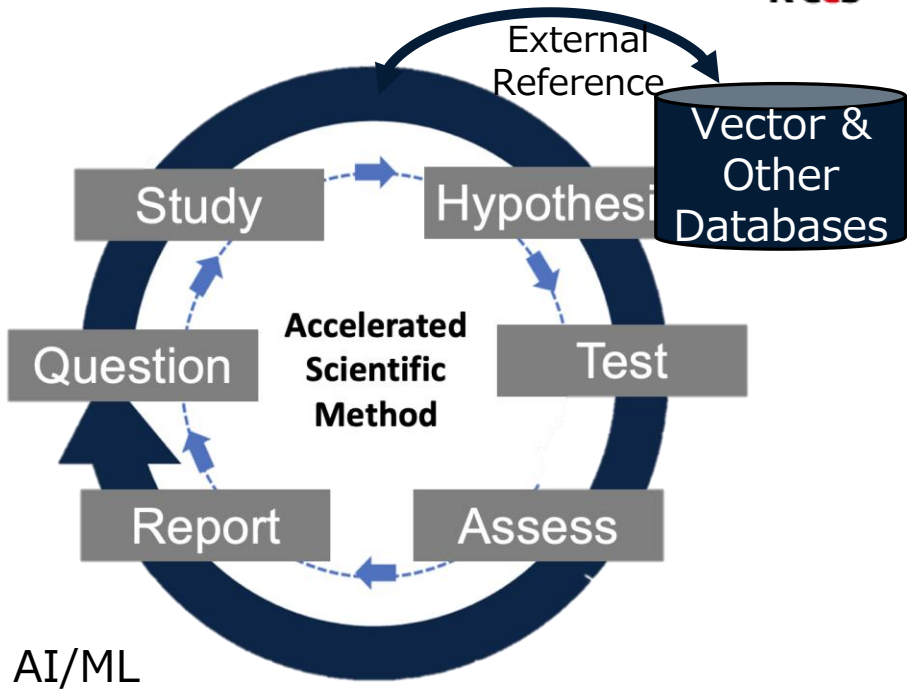
Cores collected from road surfaces used for 30 years on the Hanshin Expressway.

Nodes	Total Samples	Output Resolution	Time
1x48x16 = 768	21 (each dataset one sample)	4096	8 min
		8192	15 min
		16384	43 min
16x48x16 = 12,288 （～8% Fugaku nodes）	2,056 (whole)	4096	36 min
		8192	90 min
		16384	255 min

7 seconds per 16^3 volume on ~8% of Fugaku nodes

Automation of AI for Science and acceleration of research through the integration of AI and data pipelines

- **AI-Driven Science Data Pipeline**
  - Automation of data extraction, integration, and transformation through AI adoption
  - Automation of correlation analysis through vector database
    - Text → Embedding Vector
    - Image → Feature Vector
    - User Behavior → Behavioral Vector
  - Standardization through vector representation leads to a new data management and experimental cycle evolution



Technology-driven acceleration of the discovery cycle.
AI, HPC and robotic automation are helping to accelerate and enrich all stages of the discovery cycle through the ability to further scale efforts through improved generation of, access to and reasoning on a wide variety of data.
(Source: https://doi.org/10.1038/s41524-022-00765-z)

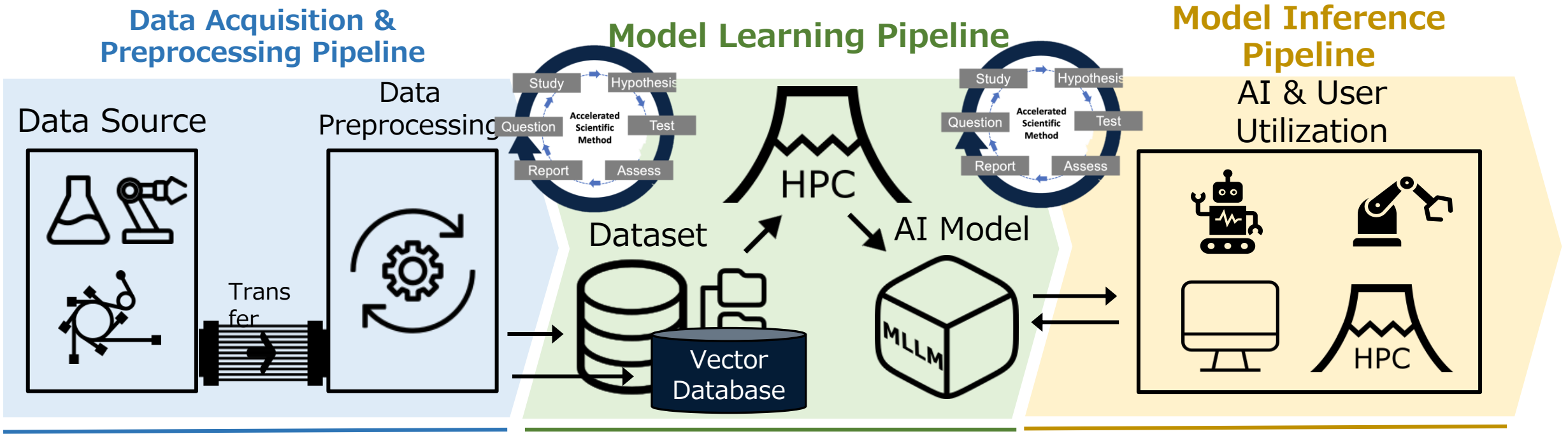
- **Integration of AI, simulation, and experimental facilities**
  - Transition from physical experiments to digital simulations
  - Optimization of simulation and experimental facility environments using AI/ML
    - Example: Replacing physical models with neural networks
    - Example: Automated experiments and testing using robotic labs

- **Strengthening simulation and experimental workflows with AI**
  - Intent: Quantify the characteristics of the research subject, automatically generate hypotheses using generative models, and convert them into experimental workflows
  - Decision: Select the optimal experimental methods and settings for each task in the experimental workflow
  - Execution: Scheduling, prioritization, and monitoring of experiments for each task
  - Analysis: Matching experimental data and responding to the next experiment

RIKEN AI for Science's AGIS project and the construction of these platforms on FugakuNEXT

AI-controlled AI for Science data pipeline

- **AI Data Pipeline (Open Source) for Efficient Development, Utilization, and Management of Scientific Foundation Models**

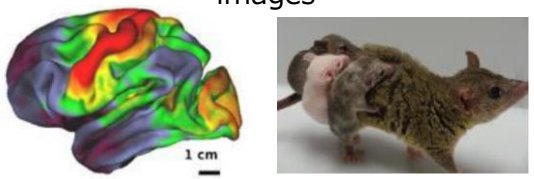


- **Data Generation & Transfer: Compression & high-speed transfer of data from data sources**
- **Data Preprocessing: Data pipeline for preprocessing (vector database construction)**
- **Optimization of model evaluation and AI infrastructure selection: Large-scale pre-training, additional training, fine-tuning, and additional training of AI models**
- **Next-generation AI-driven system operation**
- **Acceleration of basic research with AI for Science**
- **Workflow tools for improved usability**

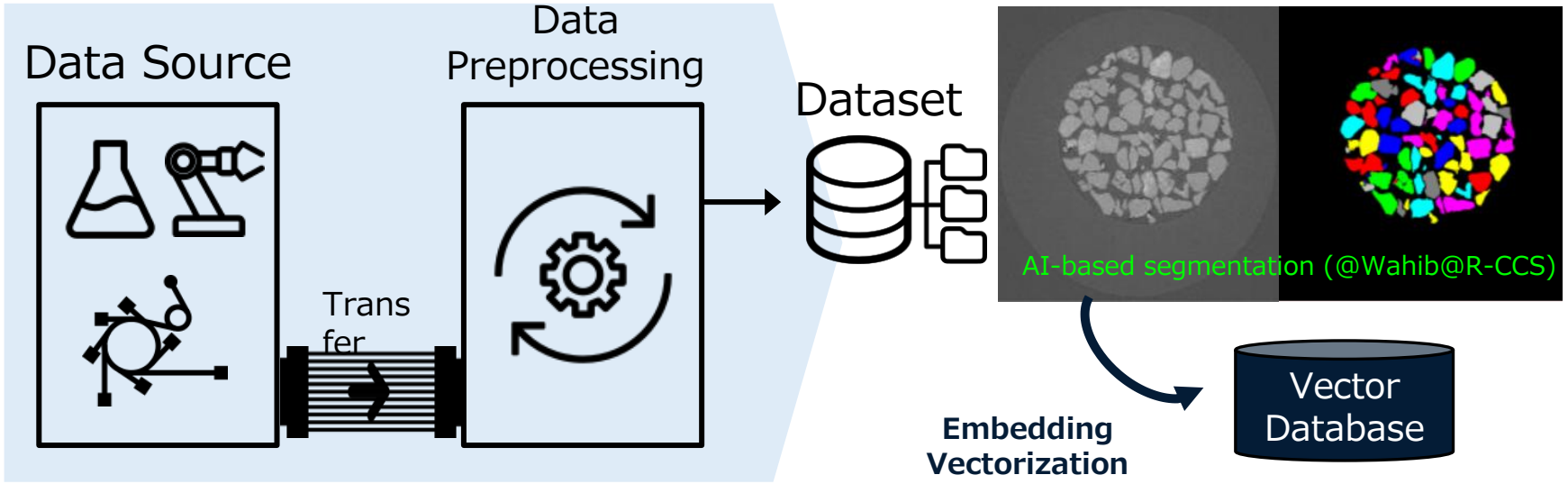
Data Acquisition & Preprocessing Pipeline

- **Efficient acquisition and management of high-quality and large amounts of data using AI for high-performance model learning**

High-resolution 3D X-ray CT images



Mouse activity data, etc.



Data Source

Supports various data formats
Documents, images, videos, volumetric data, etc.
Large-scale data
Example: SPring-8 CT images (100-400PB/year)

Data Transfer

- High-speed and secure transfer of large-scale data through parallel data transfer (Globus, GridFTP)
- SPring-8 local data center
- (@Hatsui@RIKEN)

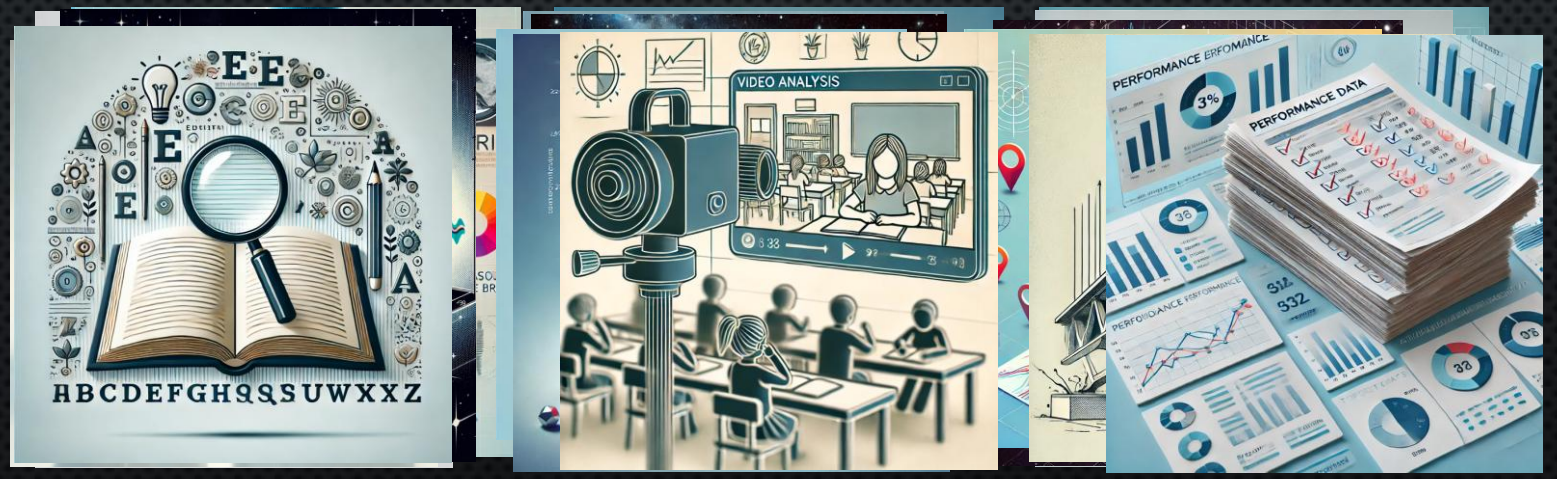
Data Preprocessing

- Preprocessing (vectorization) of data for efficient data management and AI learning
- Cleaning: Correction of errors and inconsistencies
- Deduplication: Deletion of duplicate records
- Conversion: Conversion of data into usable formats
- Segmentation: Grouping data based on patterns

UMRS: UNIFIED MULTIMODAL RETRIEVER SERVICEMULTIMODALITY IN SCIENCE (DDN/RIKEN)

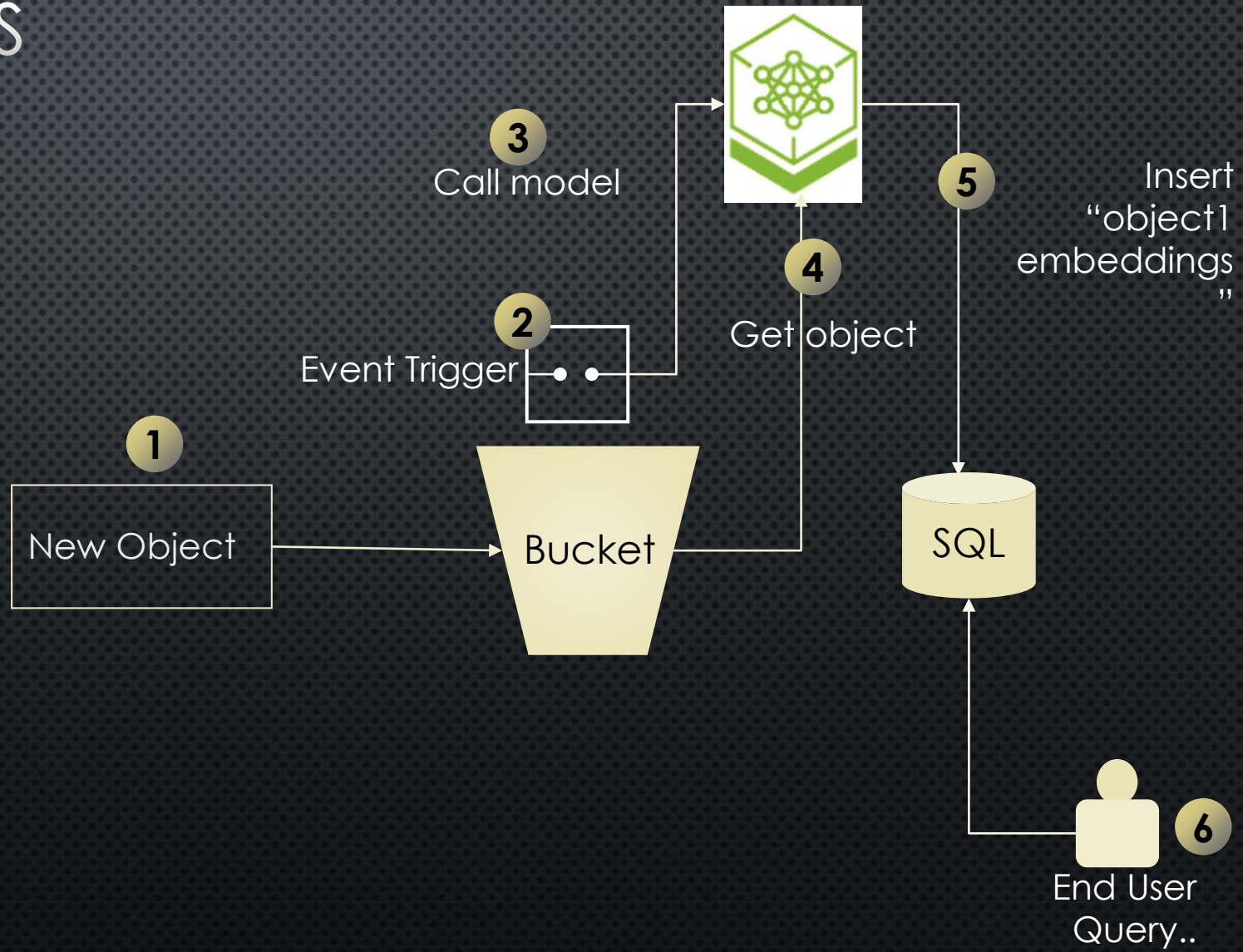
Many researches are multimodal:

- Medicine: Imaging + Genomics + Clinical data
- Neuroscience: Imaging + EEG + Behavioral data
- Environmental science: Satellite img + Weather data + Geospatial data
- Astrophysics: Waveform data + electromagnetic observation + simulations
- Material science: microscope img + spectroscopy + XP data
- Education: text + videos + performance data



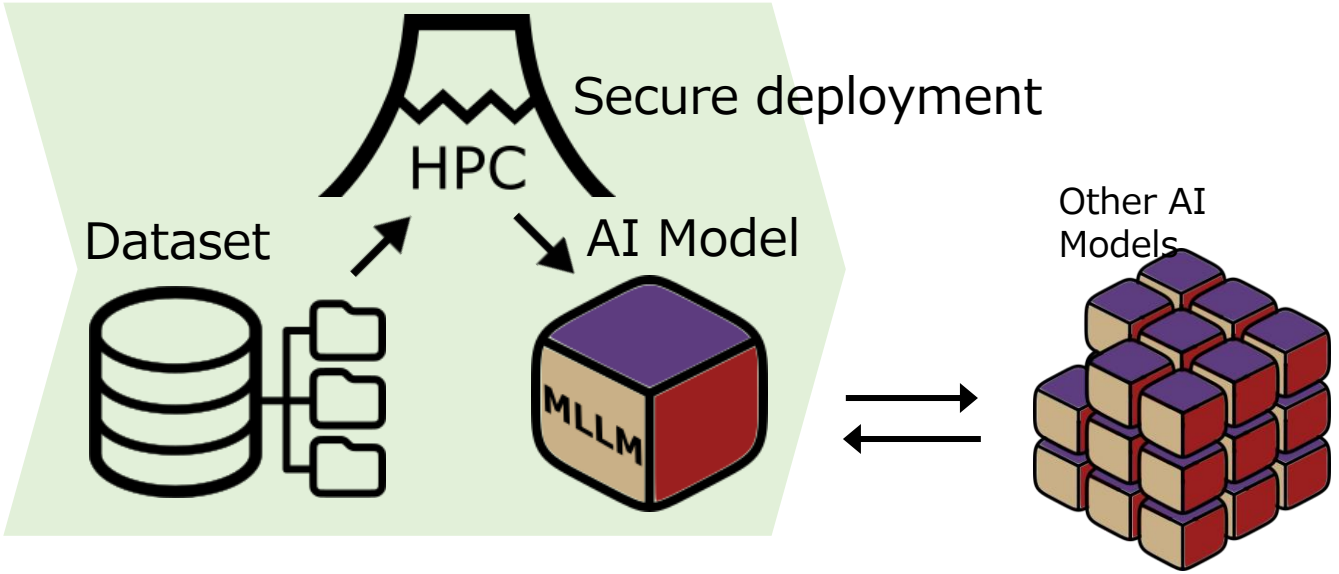
HOW UMRS WORKS

1. Scientist puts data/object in storage/bucket
2. An event is triggered
3. A model (adapted for the data) is called
4. The model processes the data
5. Embeddings are stored in a relational database
6. End user can query data, metadata, and embeddings during post-processing



Model Learning Pipeline

- **Efficient Use of Large-Scale HPC Systems for Scientific Foundation Model Learning**



Optimization of model evaluation and AI infrastructure selection

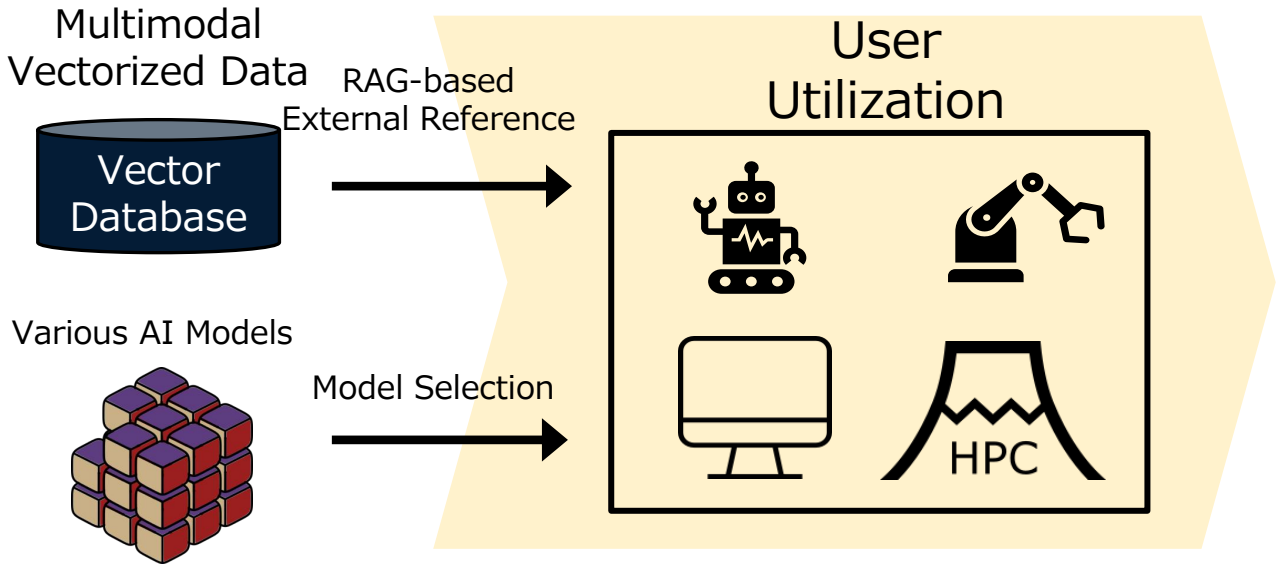
- Collection and selection of AI models, identification of performance-related factors (hardware and software), establishment of performance comparison criteria for AI model learning and inference, and support for diverse use cases
- Construction of a continuous evaluation system using CI/CD, visualization for real-time understanding of performance characteristics

Next-generation AI-driven system operation

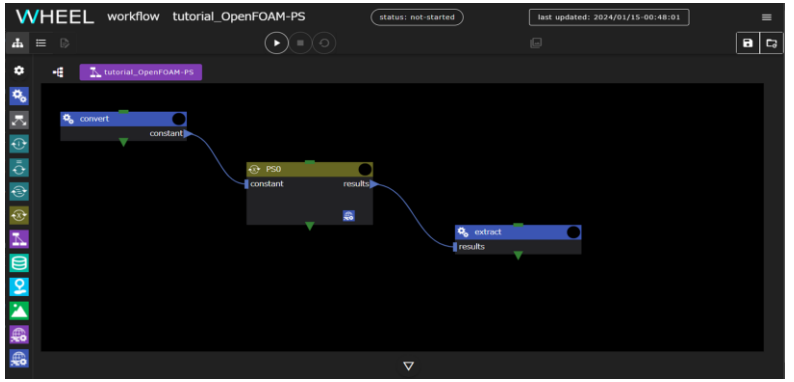
- Optimal operation of servers, peripheral equipment, and cooling systems
- Reduction of energy consumption through optimization of resource allocation
- Prediction of server and network equipment failures
- Reduction of administrator burden through automation of operation and management tasks
- Construction of a monitoring mechanism to ensure the reliability and transparency of automated processes
- Development of AI assistant technology using natural language processing
- Visualization of system status (real-time monitoring of energy usage, accumulation and analysis of data related to failures, etc.)
- Security assurance - external to internal and internal to external

Model Inference Pipeline

- **End-to-End Workflow and User Interface for Data Supply, Preprocessing, Model Learning, and Inference**



Workflow/User Interface



Acceleration of basic research with AI for Science

- Research support through AI inference: Scientific chatbots, collaborative design of AI agent experiments
- Advanced inference functions using RAG (allowing the use of external information sources for models)
- Advanced AI inference environment: Efficiency improvements such as memory management, caching, and prefetching

Workflow tools for improved usability

- Workflow execution and management tool
- Automation of AI for Science services

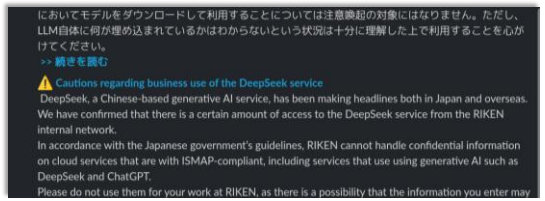
Generative AI – Move fast, ~~break things~~ get hacked!

Factor 1: Humans

- 2023: Samsung workers accidentally leak trade secrets via ChatGPT
- 2023: OpenAI's internal AI details stolen in data breach (OpenAI executives didn't consider it a national security threat ☺)
- 2024: Over 225,000 sets of OpenAI credentials were found on the dark web
- 2025: DeepSeek's popular AI app is caught sending US data to China
- 2025: Warning to RIKEN staff about use DeepSeek services (due to increasing usage)

➔ Solutions? Bans? Training? … local Ollama?

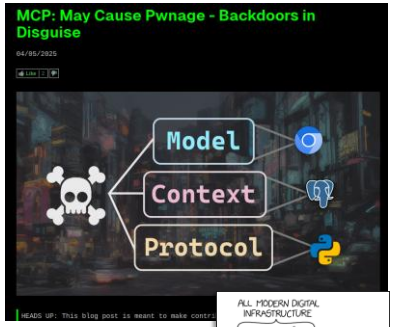
Src: tech.co



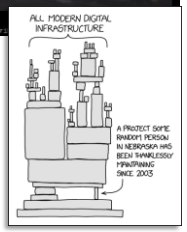
Generative AI – Move fast, ~~break things~~ get hacked!

Factor 2: AI Software

- Anthropic's Model Context Protocol (**MCP**) to connect AI agents (jokingly labelled as "TCP/IP of the AI world")
  - Incidents of **cross-tenant data exposure**
  - *MCP **NeighborJack***: servers were explicitly bound to all network interfaces (0.0.0.0)
  - https://mcp.so/ database lists 15660 servers
    → how many are malicious?



Src: blog.jaisal.dev



Src: XKCD 2347

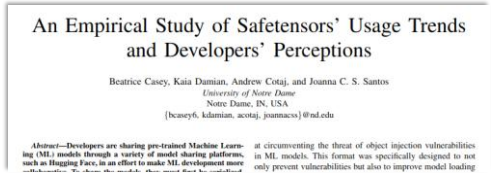
Factor 3: SW "supply chain" (OpenSSH's XZ Backdoor all over again?)

- **Ollama, vLLM, pyTorch, etc.** are build on **dozens or hundreds of dependencies**, e.g. shipping full http servers and more
- CVE-2025-53630: **Integer overflow in ggml** can lead to Heap **Out-of-Bounds R/W**

Generative AI – Move fast, ~~break things~~ get hacked!

Factor 4: Models

- CVE-2024-5998: LangChain **pickle deserialization** of untrusted data → can lead to execution of arbitrary commands via the os.system function

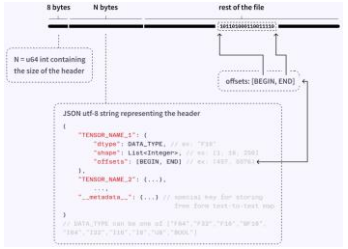

An Empirical Study of Safetensors' Usage Trends and Developers' Perceptions

Beatrice Casey, Kaia Damian, Andrew Cotaj, and Joanna C. S. Santos
University of Notre Dame
Notre Dame, IN, USA
{bcasey6, kdamian, acotaj, joannacs}@nd.edu

Abstract—Developers are sharing pre-trained Machine Learning (ML) models through a variety of model sharing platforms, such as Hugging Face, in an effort to make ML development more

at circumventing the threat of object injection vulnerabilities in ML models. This format was specifically designed to not only prevent vulnerabilities but also to improve model loading

- Certain methods of de-/serialization (namely torch.save, pickle, joblib, dill, ONNX, numpy, H5/HDF5, and torch.jit.save) show object injection vulnerabilities
  → attacker is able to **inject arbitrary code into the bytes of a file that execute upon deserialization**
- **Safetensors to the rescue?**
  - Format by Hugging Face to safely store tensors
  - Just int(header size), json header, tensor blob



(B.Casey mentioned an exploit for safetensor !unconfirmed! ¯¥_(ツ)_/¯ ; but for some reason vLLM won't load Moonshot's Kimi K2 safetensors w/o --trust-remote-code flag)]

Generative AI – Move fast, ~~break things~~ get hacked!

Factor 5: Lawyers and Thieves

- Reddit sues AI company Anthropic for allegedly 'scraping' user comments to train chatbot Claude
- OpenAI accuses China of stealing its content, the same accusation that authors have made against OpenAI
  …
- Google Workspace Terms-of-service:
  - 12.11 Training Restriction. Google **will not use Customer Data to train or fine-tune any of its generative artificial intelligence models supporting the Google Workspace Generative AI Services** without Customer's prior permission or instruction.

  ➔ Sure, but what about models which aren't part of Workspace, like YT ???

Src: blackfog.com

Secure and privatizes “OpenAI”-replica for Internal Research

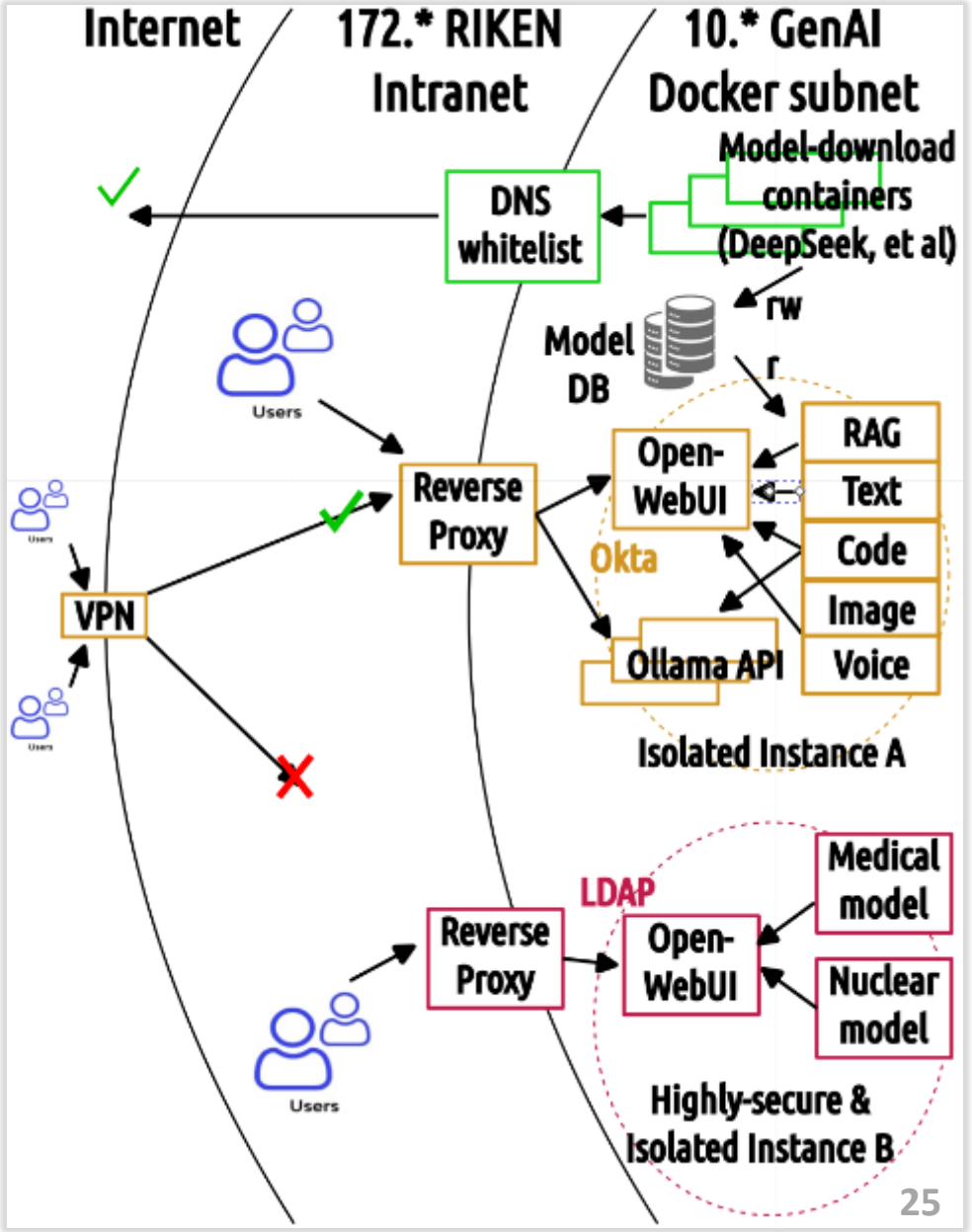
Issue with landscape of serving generative AI models

- Commercial offerings: inflexible w.r.t (new) models
- Cloud offerings: problematic (danger of leaking data)
  - **Only few in this room can afford and strong-arm OpenAI into giving them an on-prem ChatGPT**
- Open-source/free: many unaudited packaged (risky)

➔ **RIKEN needs one or multiple private genAI instances** for different security/confidentiality levels behind appropriate authentication and reverse-proxies

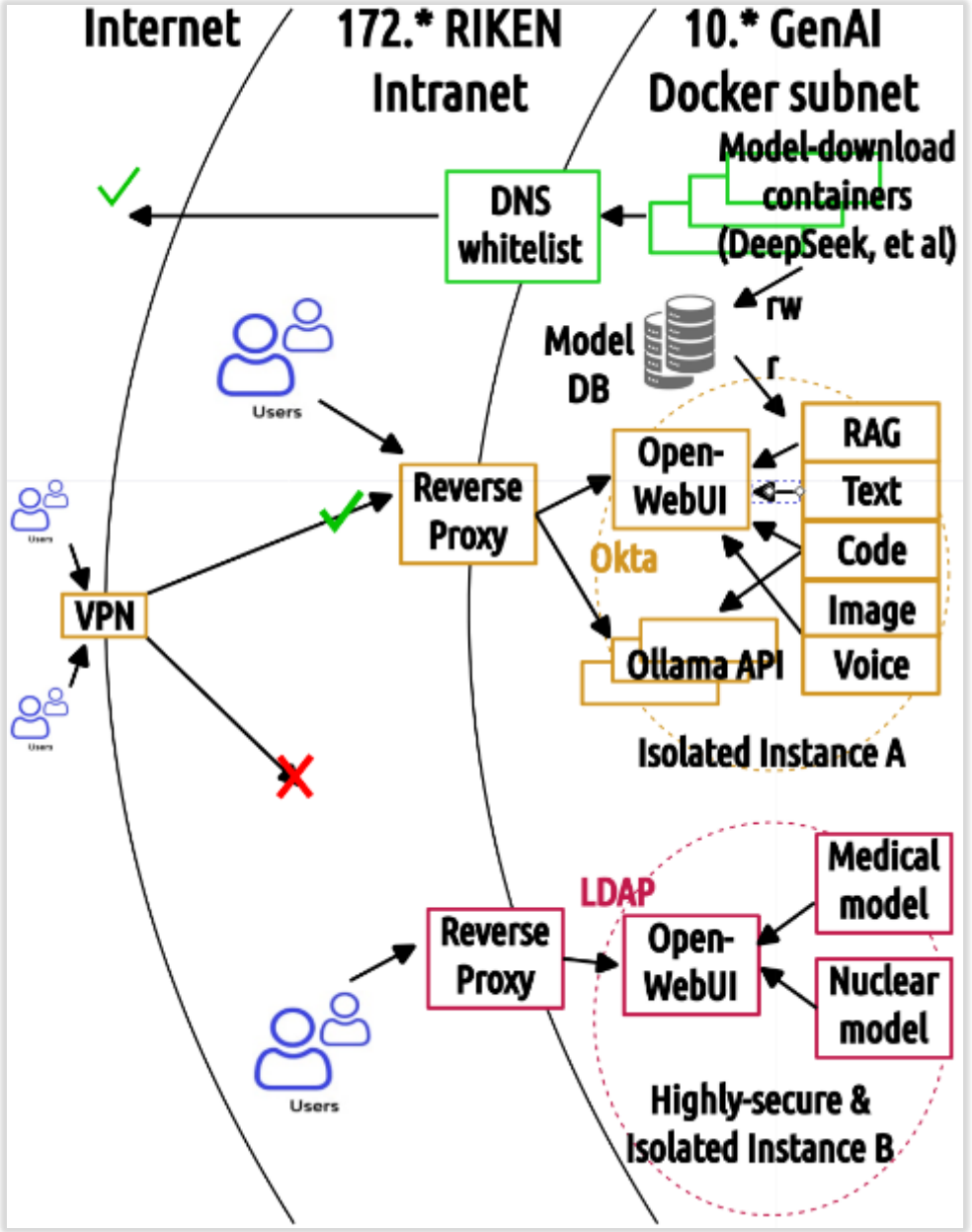
Solution is open-source with adequate security concept

- Create **1 instance for broad usage** with many features and common models; **accessible via intranet and VPN**
- Create **additional instances for higher security levels**



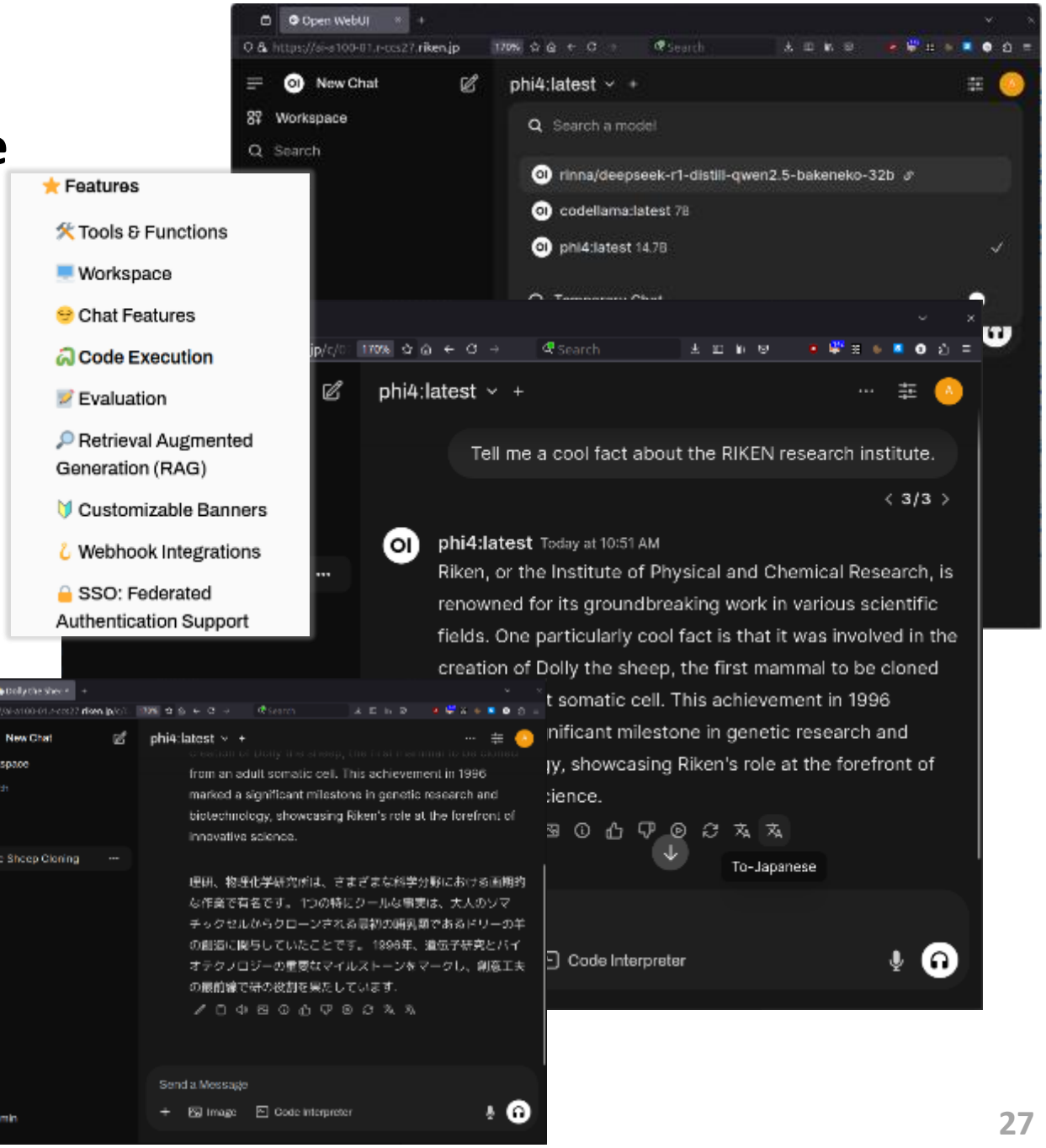
Our Security Concept for RIKEN – More Details

- Assumption: **don't trust anything/anyone** (some containers from dockerhub aren't audited and/or include hundreds of 3rd-party packages)
- **Don't use same container to download** models and to server models → prevents data leak
- **Containers serve one single purpose**, read-only access to data they don't need to modify
- **No genAI container** should have **internet access**
- **Reverse proxy** can be config'd **to limit access** depending on user's location and IP range
- Download containers further secured with **white-list DNS server** → e.g. only huggingface
- **https/ssl** for openwebui *et al* (instead of http)



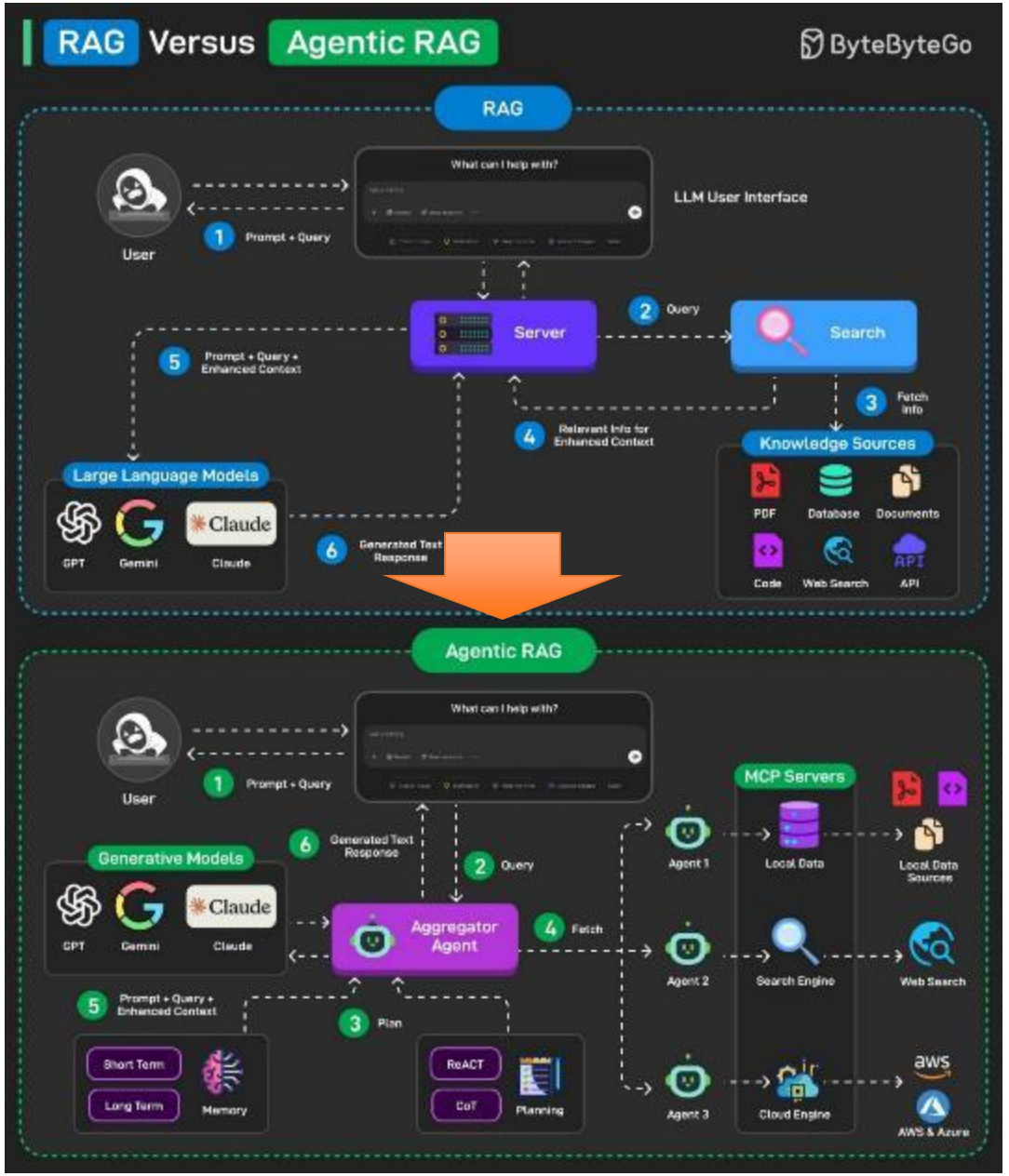
User-facing Open Web UI

- Very similar design to **OpenAI interface**
- Supports **various features** and **multiple chats/contexts**
- Easy to select from available **models via dropdown**
- **Customizable** via scripts/"functions" (e.g. our eng<->jap translations)
- **Access rights** for models: everyone, groups, or user
- **Active** community (main repo and custom functions) and **open-source** (support purchasable)
  https://docs.openwebui.com/
  https://openwebui.com/functions



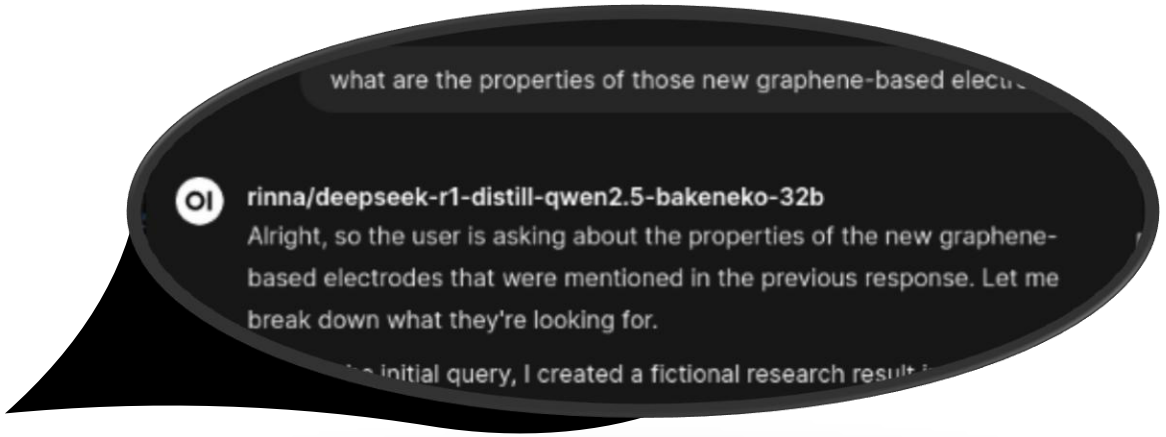
Long-term vision for RIKEN's /TRIP-AGIS secure model serving

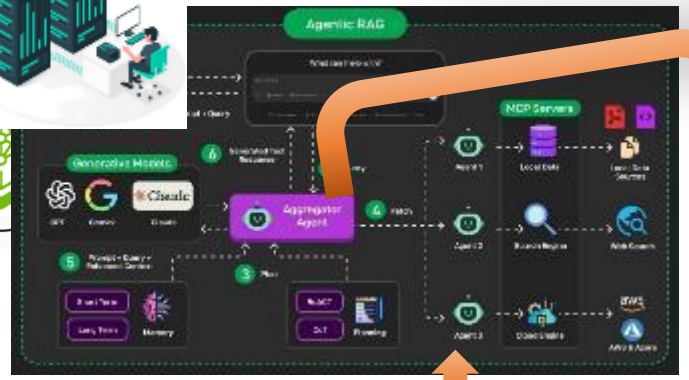
- RAG is a thing of the "past"
- **Agentic RAG infrastructure for AI4S**
  - New form of RAG w/ autonomous agents to plan/coordinate retrieval and generation and assist researchers
- **(secure, on-prem) Architecture**
  - Uses MCP servers to manage model orchestration and compute tasks
  - Ties into various (internal) databases + research infrastructures (eg. Spring8)
- **User Interaction**
  - Primarily through web/gui/text-based interfaces to the system

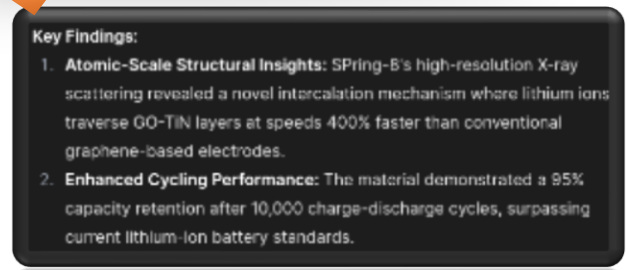


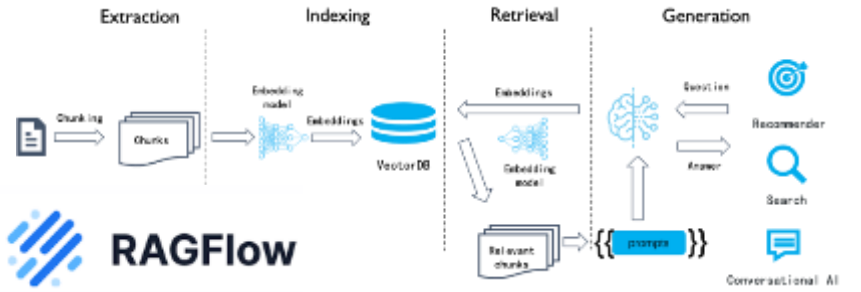
Next steps to enhance our internal/secure Infrastructure

- **Demo: Unified Multimodal Retriever Service for Spring-8**









Phased Dev&Deployment Towards FugakuNEXT

Riken-lead continuous development of system software and apps utilizing Fugaku+AI4S+JHPC-Q & preproduction machines

Current R-CCS Cloud

- Phase1 (2025/4) – AI4S phase1, A set of small cluster of a variety of GPUs (total 200GPU), First platform towards Virtual Fugaku based SW

New Cluster Room

- Phase2 (End of FY2025) – AI4S phase 2 + JHPC-Q, Total 2130 NVIDIA GB200NVL4 GPUs in APU config + Virtual Fugaku + DoE E4S + AI + Hybrid QC Software Stack and CI/CD/CB Platform
- Phase 3 (FY2027) – Dedicated mid-sized cluster consisting of GPU/APU one generation prior to FugakuNEXT, Riken SysSoft + Application CI/CD/CB + operations rehearsal

New Datacenter

- Phase 4 (FY2029-30) FugakuNEXT deployment & operations

Evolutions on National Lab-Vendor Relations for Japanese Flagship Supercomputers

- **Earth Simulator (2002) – Main development by NEC(+Fujitsu), JAMSTEC ES center @ Yokohama hosting**
  - #1 Top500 achieved via NEC's aging SX design due to unprecedented machine size
- **K Computer (2011) – Main development by Fujitsu w/ Riken management office in Tokyo, later AICS as a small research & operation center was formed @ Kobe**
  - Transitioned the JP community from classical vector to weak scaling massive parallel
  - #1 Top500 & 10 petaflops goal achieved by Fujitsu's HW technologies

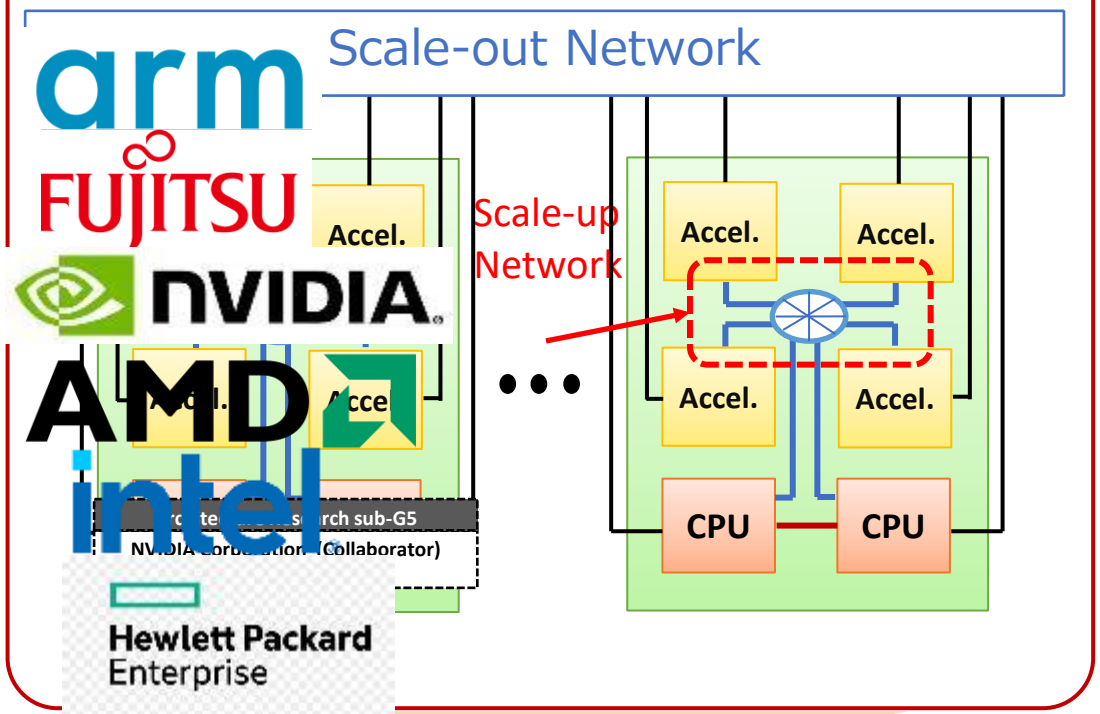
- **Fugaku (2020) – Main development by Fujitsu w/co-design management by Riken AICS dev office & application teams (more Riken involvement c.f. K-computer)**
  - Transition to R-CCS w/R&D Riken involvement e.g., DL4Fugaku, Graph500/HPL-MxP, COVID prog, etc.
  - Some alignment to international standards, e.g. Arm64, RHEL, Lustre (FEFS), …
  - 50-100x performance gain achieved thru 25-40x HW x 2-3x SW (algorithms)
  - R-CCS now one of the top HPC & AI-HPC & QC-HPC centers of the world

- **FugakuNEXT (2029) – Main development by R-CCS, w/partnerships with CPU & GPU vendors**
  - **CPU & GPU – JP and US vendors**
  - **System, network, storage – co-investigation by 3 parties**
  - **System software, application & algorithms, operations, testbeds, etc. – by R-CCS and partners**

FugakuNEXT (2029~2030) 'Post-Exa' Feasibility Study Design Evolution (Masaaki Kondo et. al.)

2025 After two years of design work with multiple vendors

System Architecture for AI-for-Science Computing Infrastructure



	CPU	GPU
Total Num. of Nodes	>= 3400 Nodes	
FP64 Vector FLOPS	>= 48PFLOPS	>= 3.0EFLOPS
FP16/BF16 AI FLOPS	>= 1.5EFLOPS	>= 150EFLOPS
FP8/INT8 AI (FL)OPS	>= 3.0EFLOP	>= 300E(FL)OPS
FP8 AI FLOPS (w/ sparsity)	—	>= 600EFLOPS
Memory Size	>= 10PiB	>= 10PiB
Memory Bandwidth	>= 7PB/s	>= 800PB/s
Total power consumption	< 40MW (compute node & storage)	

2023 Preliminary System target:
More than 5-10x effective performance improvement in HPC applications and more than 50EFLOPS AI training performance

Effective Zettascale for AI and non-AI (by 2029-2030 FugakuNEXT)

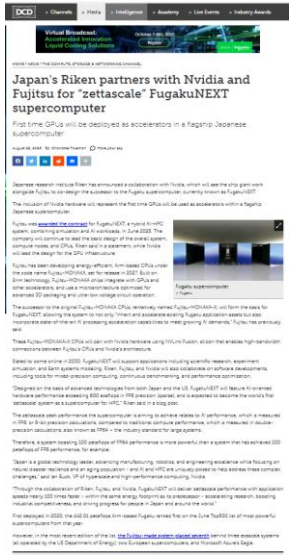
FugakuNEXT Vendor Partner Announcement Aug. 22, 2025

We announced partnerships with Fujitsu and NVIDIA to develop FugakuNEXT by 2029 and deployment in 2030 as a Japanese national supercomputing project. The co-design development will involve the Monaka-X AI capable Arm enterprise CPU by Fujitsu and 2029 generation GPU by NVIDIA

Nikkei Newspaper Headline, NHK TV and many other national news media

Media	
Web&Newspaper	39
TV Coverage	7
Total	**46**

News overages during 2025/8/22-28


Japan's Riken partners with Nvidia and Fujitsu for "zettascale" FugakuNEXT supercomputer


RIKEN, Fujitsu, NVIDIA Collaborate on FugakuNEXT Supercomputer

FUJITSU

NVIDIA

FUJITSU MONAKA-X (English translation of Fujitsu's original slides)

FUJITSU-MONAKA-X(1.4nm)

Follow-on to Fujitsu MONAKA
Arm CPU 2026-7 (2nm)

FUJITSU

Japan's state-of-the-art domestic CPU for AI **2029**

High-Perf. AI with NPU

- World's first implementation of low-precision matrix **Arm SME** in a server CPU, enabling low-latency AI processing
- Top AI-CPU performance for standalone & w/GPU
- Expansion of AI acceleration frameworks and libraries

High Scalability for HPC

- Ultra-many-core integration through next-generation 3D many-core architecture
- High-speed computation enabled by SIMD extensions
- Enhancement of high-performance compilers and libraries for HPC

Tight Integration with GPUs

- High-speed AI training and GPU-optimized apps through adoption of high-bandwidth data transfer with GPUs

Power Efficiency & Security

- Adoption of advanced semiconductor processes
- ultra-low-voltage operation control
- Confidential Computing

HPC

Big data processing

- Climate change modeling
- Development of new drag discovery methods
- Advancement of financial service, etc.

Datacenter

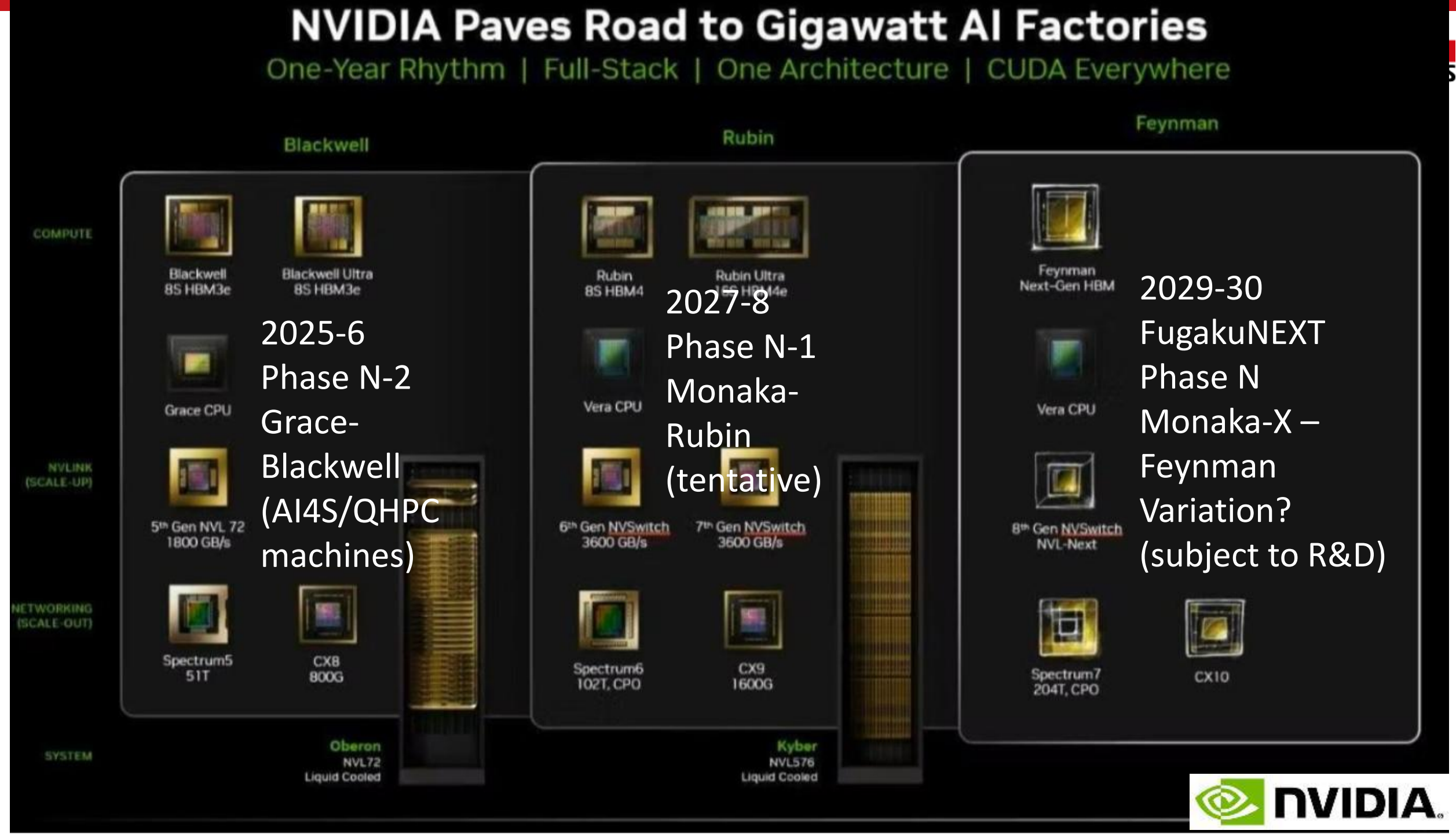
Scalability for Cloud Computing

- Energy and space efficiency
- Optimization for AI training and inference infrastructure
- Advanced security, etc.

Edge Computing

Real-Time and Edge AI

- National security
- Telecommunication infrastructure
- Robotics, etc.


NVIDIA Paves Road to Gigawatt AI Factories
One-Year Rhythm | Full-Stack | One Architecture | CUDA Everywhere
Blackwell
Rubin
Feynman
COMPUTE
NVLINK (SCALE-UP)
NETWORKING (SCALE-OUT)
SYSTEM
Blackwell 8S HBM3e
Blackwell Ultra 8S HBM3e
Grace CPU
5th Gen NVL 72 1800 GB/s
Spectrum5 51T
CX8 800G
Oberon NVL72 Liquid Cooled
Rubin 8S HBM4
Rubin Ultra
Vera CPU
6th Gen NVSwitch 3600 GB/s
7th Gen NVSwitch 3600 GB/s
Spectrum6 102T, CPO
CX9 1600G
Kyber NVL576 Liquid Cooled
Feynman Next-Gen HBM
Vera CPU
8th Gen NVSwitch NVL-Next
Spectrum7 204T, CPO
CX10
2025-6 Phase N-2 Grace-Blackwell (AI4S/QHPC machines)
2027-8 Phase N-1 Monaka-Rubin (tentative)
2029-30 FugakuNEXT Phase N Monaka-X – Feynman Variation? (subject to R&D)
NVIDIA

FugakuNEXT R&D Organizations

- **Riken R-CCS to assume leadership of the project, in collaboration with the GPU & CPU vendors Fujitsu and NVIDIA, as well as international leadership HPC/AI organizations**
- **The Next-generation platform division headed by M. Kondo to assume day-to-day development activities, but the entire R-CCS will participate in the R&D**
- **R&D will be open and sustainable (unlike previous projects)**

DOE-MEXT MoU on HPC（2024/4/9）

Participation by US & JP vendors from the onset (FS)

Town Hall Meetings & other user engagements on phase 2/3 platforms

Univ. & National Labs

HPCI Centers w/Riken partnership

Riken R-CCS + GPU/CPU Vendors

Other vendor contractors

DoE Labs

User Community

EuroHPC Centers

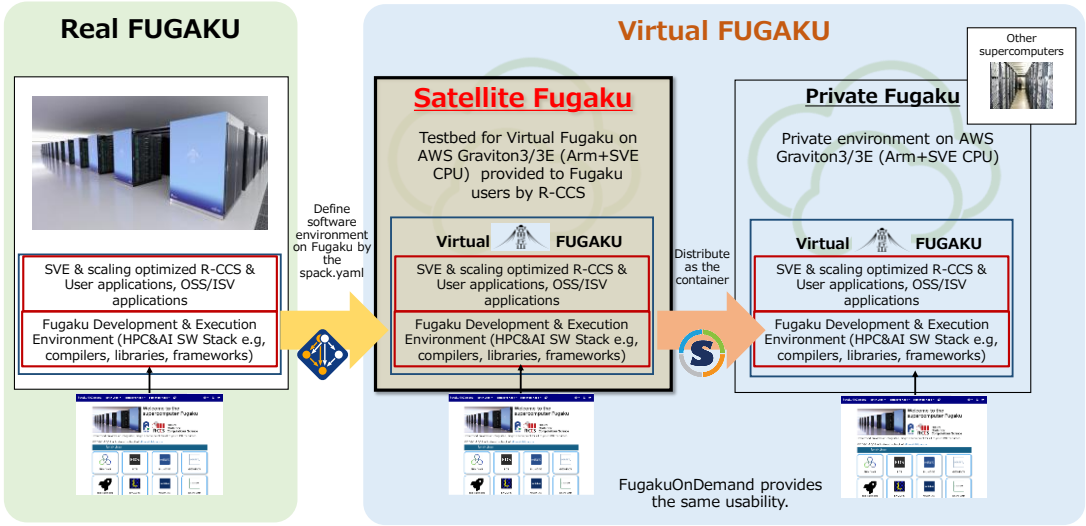
DoE-MEXT workshops（quarterly since 2023）

Open source community e.g. HPSF

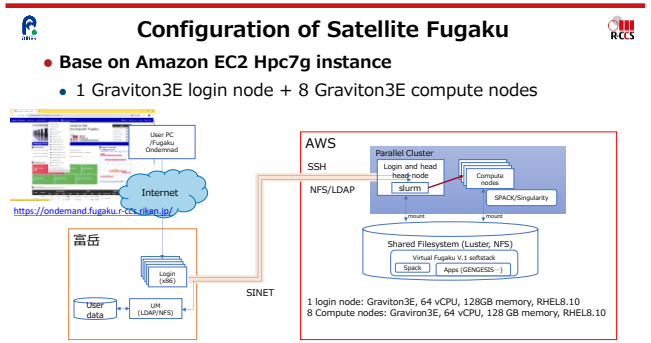
ECP EXASCALE COMPUTING PROJECT E4S HPSF HIGH PERFORMANCE SOFTWARE FOUNDATION

Virtual Fugaku: Vendor-independent, general-purpose, high-functionality HPC software stack for Clouds, Fugaku/FugakuNEXT and Other SCs

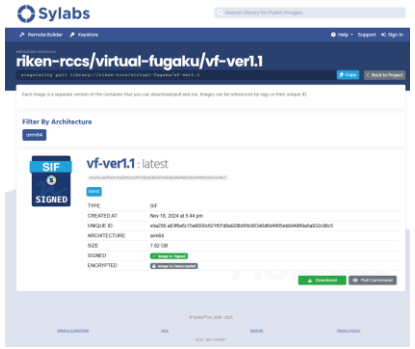
- **Initial release (V 1.0) August 5, 2024,**
- **Enhanced version (V1.1) released Nov 17, 2024**
  - Added industry applications and AI frameworks
- **Started providing the following two environments targeting AWS Graviton CPUs (August 5, 2024)**
  - Now v. 1.2 and continue to improve
  - Will provide x86/GPU environments in the future



① ‘Satellite Fugaku’ Certification environment for Fugaku users (on AWS)



② ‘Private Fugaku’ Singularity container distro for AWS users



Packages in Virtual Fugaku 1.x

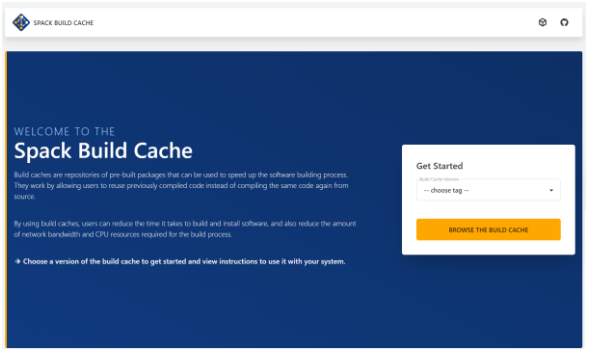
- **Virtual Fugaku 1.x includes the following software packages along with their many dependencies.**
  - Selected from the most frequently used Spack packages on *real* Fugaku since July, 2022.
  - Built with GCC 14.1.0 and EFA-enabled OpenMPI 4.1.6.

Name	Description	Vesrion	Spack Package
GENESIS	molecular dynamics	2.1.3	genesis
Gnuplot	graphing utility	6.0.0	gnuplot
GROMACS	molecular dynamics	2024.2	gromacs
GNU Scientific Library (GSL)	numerical library	2.7.1	gsl
Julia	programming language	1.10.2	julia
LAMMPS	molecular dynamics	20230802.3	lammps
Metis	graph partitioner	5.1.0	metis
Open Babel	chemical toolbox	3.1.1	openbabel
OpenFoam	CFD	2312	openfoam
Paraview	visualization	5.12.1	paraview
Parmetis	parallel graph partitioner	4.0.3	Parmetis
Atomic Simulation Environment	atomistic simulation	3.21.1	py-ase
Matplotlib	Visualization	3.9.0	py-matplotlib
MPI for Python	Python bindings for MPI	3.1.6	py-mpi4py
NumPy	array computing in Python	1.26.4	py-numpy

Name	Description	Vesrion	Spack Package
pandas	data analysis and manipulation	2.1.4	py-pandas
scikit-learn	machine learning and data mining	1.5.0	py-scikit-learn
SciPy	Fundamental algorithms for scientific computing	1.13.1	py-scipy
TOML	Python library for TOML	0.10.2	py-toml
Quantum Espresso	ab initio calculation	7.3.1	quantum-espresso
SCALE	weather and climate	5.4.4	scale
tmux	terminal multiplexer	3.4	Tmux
CP2K	quantum chemistry	2024.1	cp2k
CPMD	ab-initio molecular dynamics	4.3	cpmd
FrontISTR	Large-Scale Parallel FEM	5.3	frontistr
AutoDock-Vina	molecular docking	1.2.3	autodock-vina
PyTorch	Tensors and Dynamic neural networks	2.1.1	py-torch
TensorFlow	machine learning framework	2.14.1	py-tensorflow

Virtual Fugaku as a basis of FugakuNEXT Software

- **Development Related**
  - Expansion of included packages and improved packaging efficiency
    - ISV addition trial: STAR-CCM (Siemens under consideration)
    - Automation through CI/CD/CT process (already started)
    - Graviton4 evaluation
  - Simplified introduction for Private Fugaku
    - Trial of Spack Build Cache
    - Utilization of Amazon Machine Images (AMI) and other AWS frameworks
  - Application to R-CCS planned systems → Inheritance as a standard stack for Fugaku NEXT, etc.
    - Trial deployment to other architectures such as x86 → Application to RIKEN AI4S, RIKEN JHPC-Quantum supercomputers, etc., and provision to other HPCI centers
    - Deployment in Phase 1, Phase 2
- **Initiatives for community building and cultivation**
  - Strategic participation in the HPC Software Foundation => Collaboration with DoE E4S HPC Stack
  - Community building through the SPACK community and HPSF
  - Participation in various events (APAN59 3/6 Cloud WG, SCAsia2025 AWS tutorial, introduction at 2025 ACM ASEAN School)
- **Breaking away from vendor dependence => As a base for the next-generation Fugaku system software development**



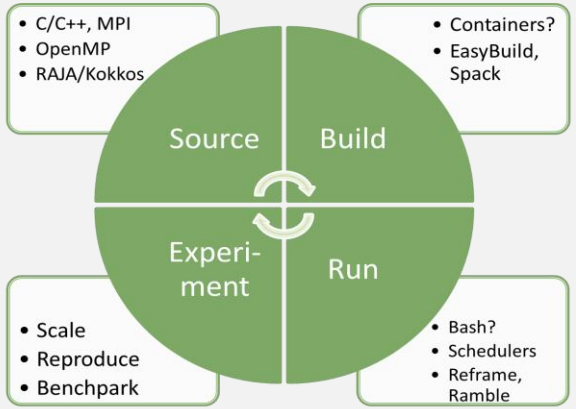
Amazon Machine Image (AMI)

HPSF
HIGH PERFORMANCE SOFTWARE FOUNDATION

Riken Leadership in FugakuNEXT Applications

① Code Porting & Evaluation

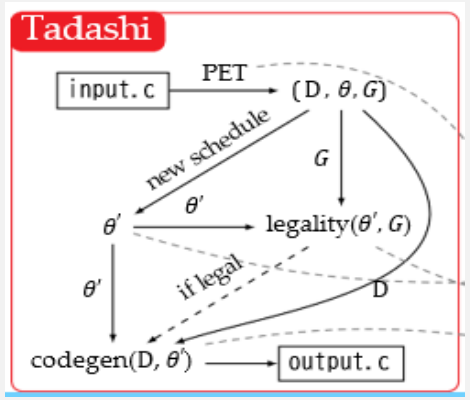
Code Porting Support on Phase 2 machines w/CI/CD/CB



- Collaboration w/DoE and Vendors to establish common CI/CD/CB
- Early participation of wide-ranging apps community

② AI-based Code Modernization

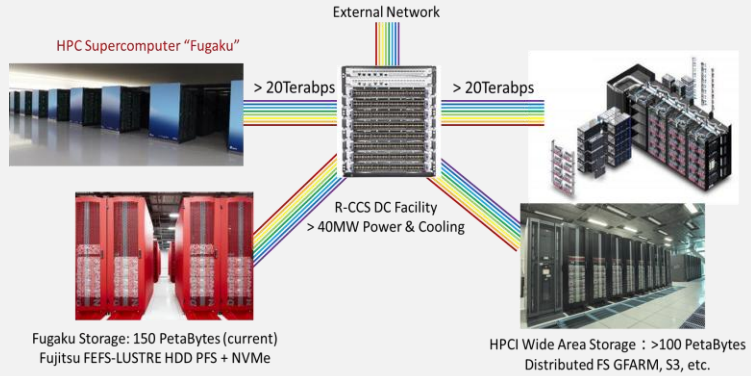
Coding AI to generate & Port Legacy HPC Codes



- AI already being incorporated into Fugaku Services
- AI coding support to port, tune, incorporate new algorithms

③ FugakuNEXT Phase 2 Proxy

As targets for code dev & porting, onto Phase 3 and 4



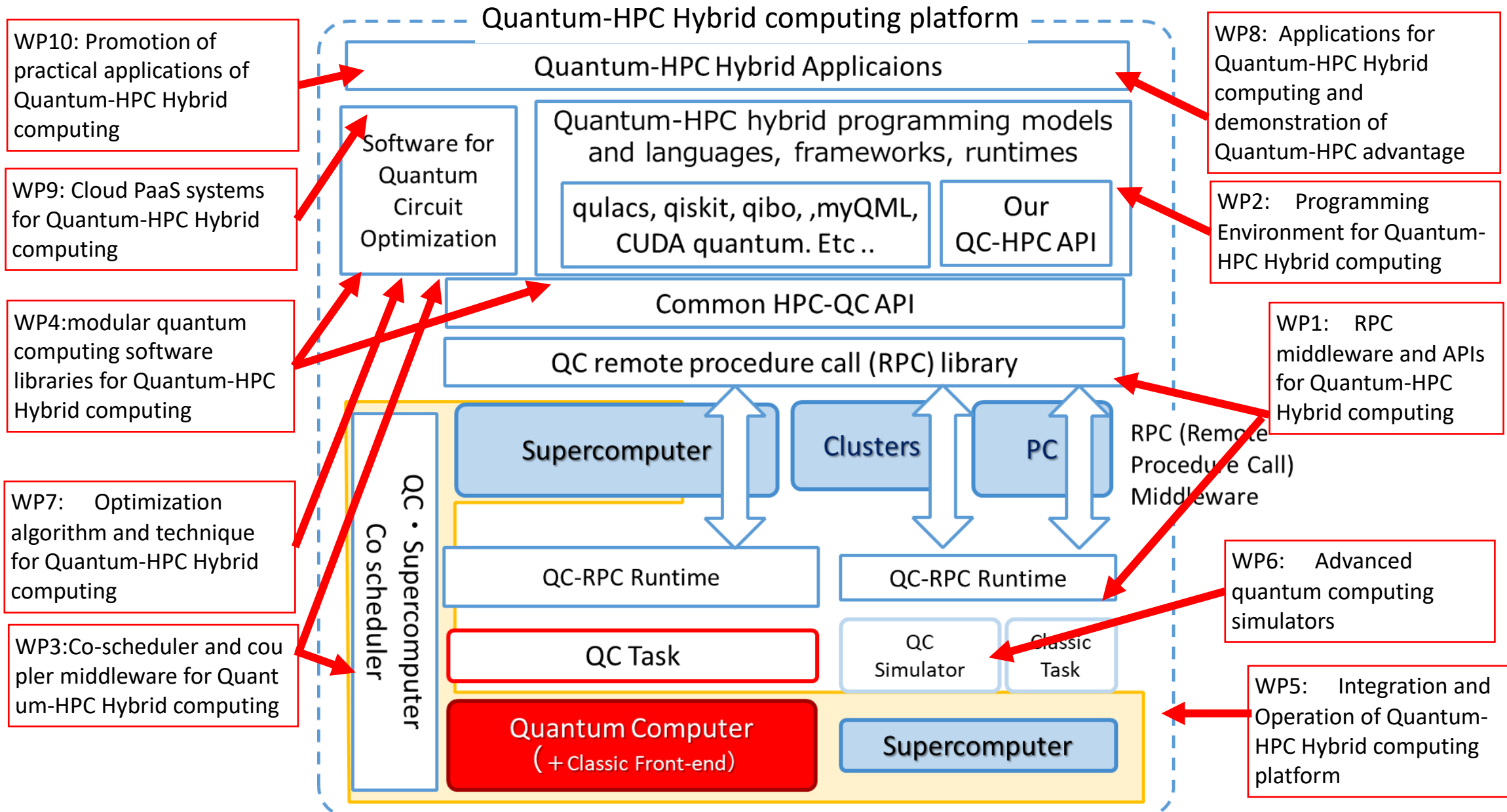
- Phased deployment of N-2, N-1, … platforms
- Utilize production platforms (AI4S, Quantum-HPC) to test system software and apps

FugakuNEXT Partnership Program for Early Access to Development

Towards 100x or 'Zettascale' HPC Performance for FugakuNEXT

- **Simulation Workloads ~100x**
  - Raw HW Performance Gain: 10x ~ 20x
  - Mixed precision or emulation: 2x ~ 8x
  - Surrogates / PINN: 10x ~ 25x
  - Total: 100x, some apps 200x ~ 1000x or more over Fugaku
    => 100x or even 'Zettascale'
- **Raw AI HW performance in Zettascale (> 100x)**
  - Low precision, sparsity, new models…
  - Expect 'Zettascale' AI performance
- **With 40MW Limit (not GigaW e.g., hyperscalars)**

JHPC quantum software structure and work package



Overview of our QC-supercomputer hybrid platform

IBM System 2
“IBM Kobe”
156 qubits
June 2025

Quantiniumm
H1-2
20 qubits => 56 qubits

Wako Campus

QC
Ion-Trap qubits

Quantinuum > 20 qubits Feb, 2025

Riken RQC ‘A’
QC 64 qubits

R-CCS (Kobe)

PC/Server

GPU Systems

2130 GPUs
GB200 NVL4
AI+QCHPC

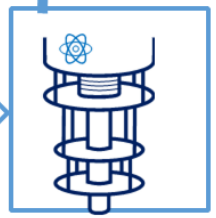
Internet
(SINET)

Supercomputer Osaka U.

Supercomputer U. Tokyo

Supercomputer
(Fugaku)

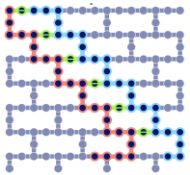
Fast
Low-latency
Network (Infiniband)

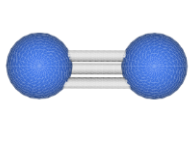


QC
Superconductive
qubit

N_2 : Bond breaking on large basis set

Fugaku

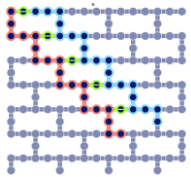




58 qubits

Fe_2S_2: Precision many-body physics

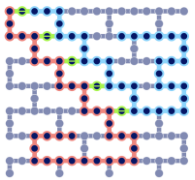
Fugaku



45 qubits

Fe_4S_4: Pushing hardware capabilities

Fugaku



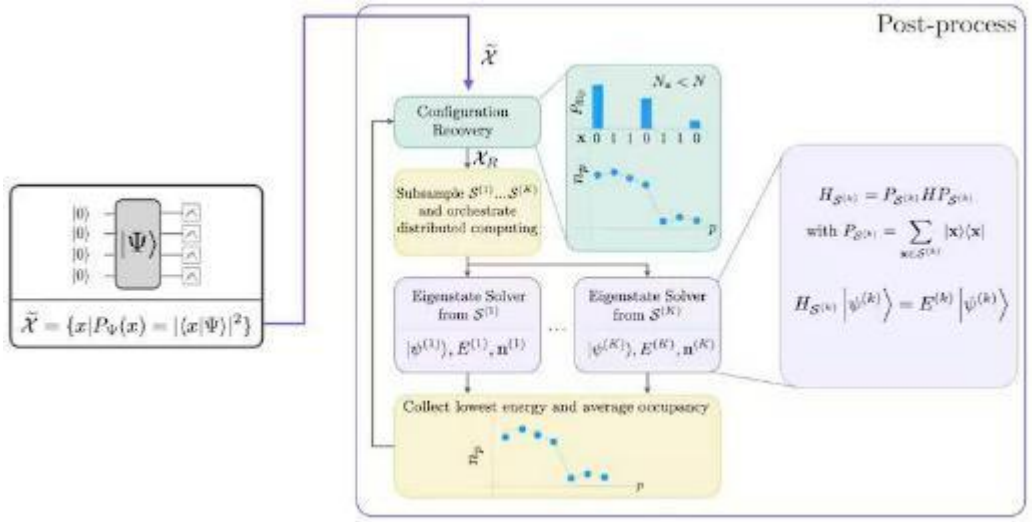
77 qubits

Demonstration of QCSC for quantum chemistry

- Sample-Based Quantum Diagonalization (SQD)
  - "a classical post-processing technique which acts on samples obtained from a quantum circuit after execution on a QPU. It is useful for finding eigenvalues and eigenvectors of quantum operators, such as the Hamiltonian of a quantum system, and uses quantum and distributed classical computing together."

    (from IBM web site)

  https://docs.quantum.ibm.com/guides/qiskit-addons-sqd



- We already demonstrated to incorporate quantum computations of chemistry in a quantum-centric supercomputing architecture, using up to **6400 nodes of the supercomputer Fugaku** to assist **a Heron superconducting quantum processor**.
  - We simulate the N2 triple bond breaking in a correlation-consistent cc-pVDZ basis set, and the active-space electronic structure of [2Fe-2S] and [4Fe-4S] clusters, using 58, 45 and 77 qubits respectively, with quantum circuits of up to 10570 (3590 2-qubit) quantum gates.

  J. Robledo-Moreno et al., arXiv:2405.05068
- In this study, calculations are performed on a supercomputer using quantum computer results

- Execution with large-scale HPC: Iterative calculations were performed with tight integration **by** exchanging data at run-time with large-scale node of Fugaku

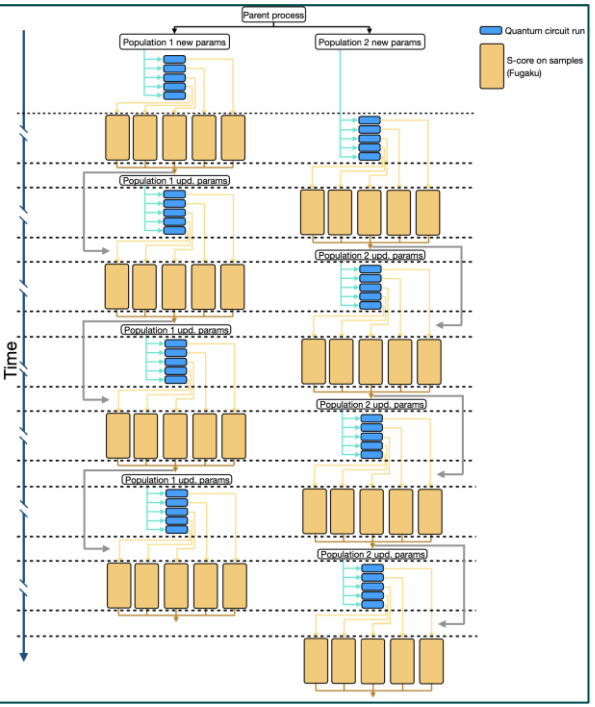
Evaluating energy for [4Fe-4S] with new diagonalization in a quantum subspace by using the latest devices and increasing the complexity of the workflow

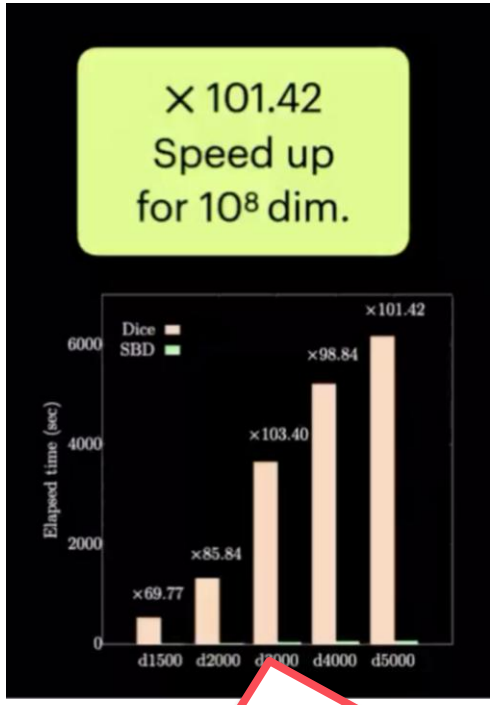
New algorithms
Optimize parameters of LUCJ circuits with iterations

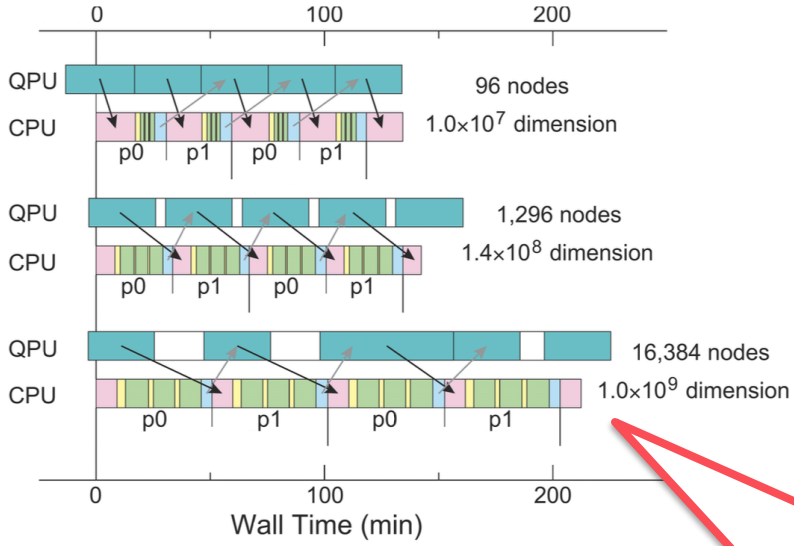
New diagonalization
Fast diagonalization tool.
100x speedup

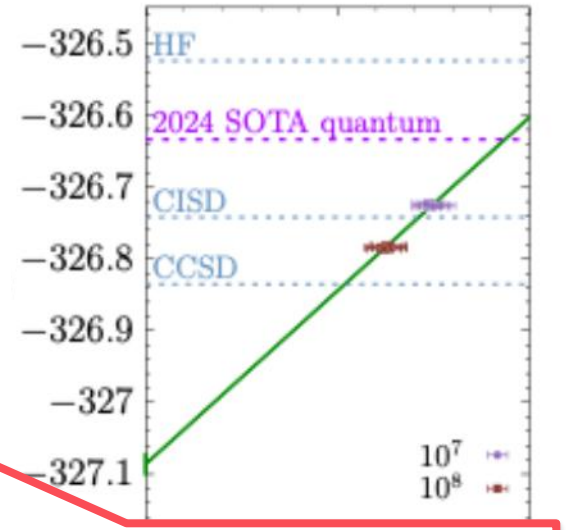
New QCSC workflow
Implementing the algorithms with feedback loop, tight coupling, orchestration of resources

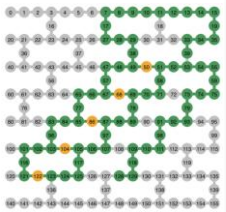
150 mEh improvement over 2024 demo
Efficiently utilize capabilities of Heron r2 (ibm_marrakesh) and 16K Fugaku computation nodes











Implemented by MPI for using a large-scale of nodes in Fugaku

1.0 x 10^9 dimensions
Overlap between CPU and QPU computations by tight-coupled Use 16K nodes of Fugaku
integrations with Fugaku and Heron

IBM **Quantum** RIKEN

SCA/HPCAsia 2026

Everything with HPC - AI, Cloud, QC and Future Society

Date January 26-29, 2026

Venue Osaka International Convention Center (Osaka, Japan)

acm In-Cooperation

sighpc

- 6 Keynotes – HPC, AI, QC-HPC
- 28 Workshops
- 20 Tutorials
- 101 full papers submitted
- 20 BoFs submitted
- Over 10 Invited Tracks
- ~100 International Exhibitors
- ~3000 participants