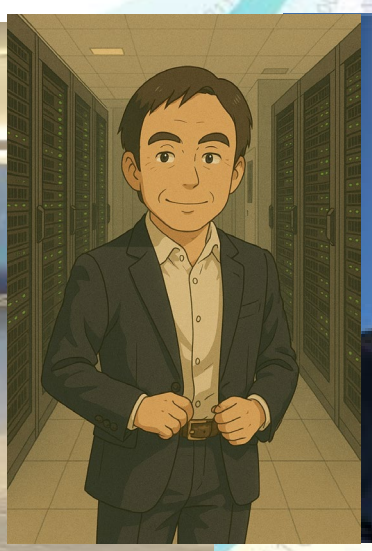
Convergence: HPC, AI, and Quantum towards FugakuNEXT







Satoshi Matsuoka, Director Riken R-CCS
And many other collaborators @ R-CCS and elsewhere

Feb 16 2027
Multicore World, Christchurch, NZ

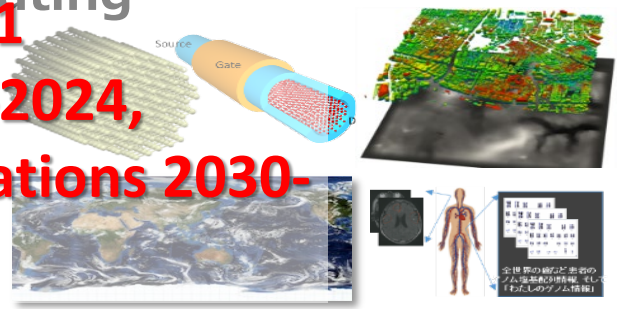
Riken R-CCS Strategy for Innovation by Computing
Future of Science 'of' and 'by' Computing

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

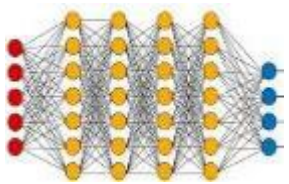
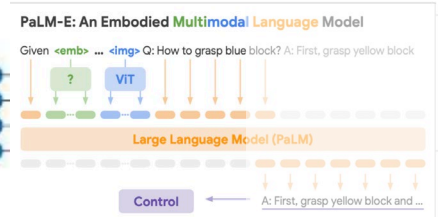



Fugaku: Current until 2030~2031
FugakuNEXT: Feasibility Study 2022-2024,
R&D 2025-2029, Deployment ~2029, Operations 2030-
'Zettascale' @ 40MW

- **Science by High Performance Computing**

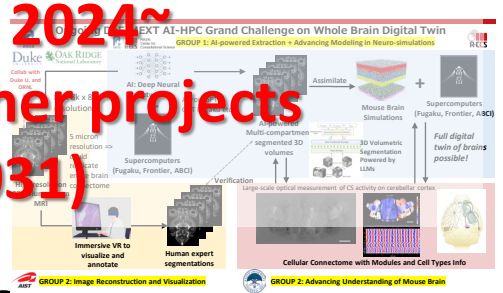
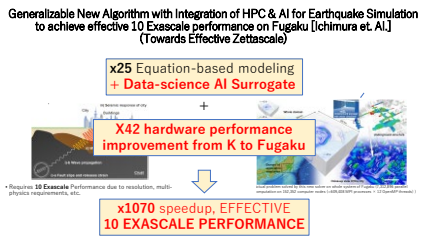


- **Science of High Performance AI**

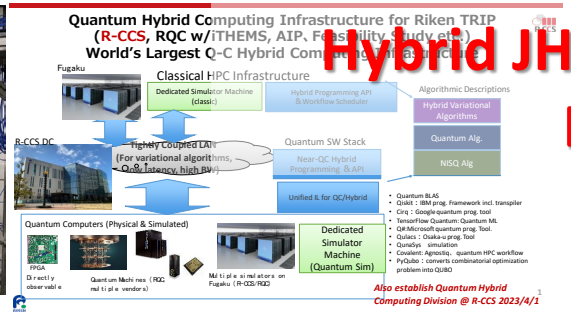




Riken AI for Science FY 2024~
including TRIP-AGIS and other projects
(TRIP-AGIS 2024~2031)

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

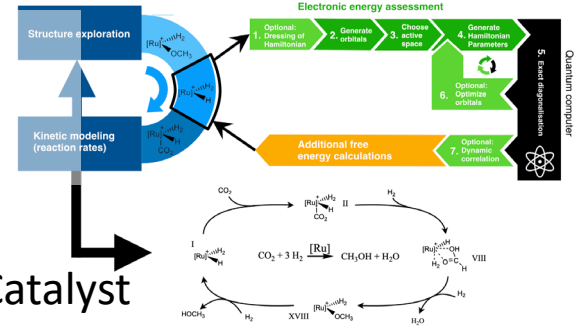



- **Science of Quantum-HPC Hybrid Computing**

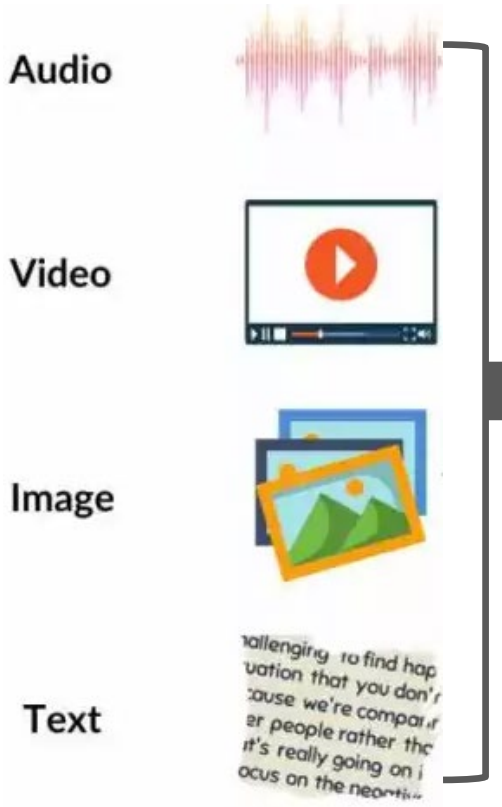



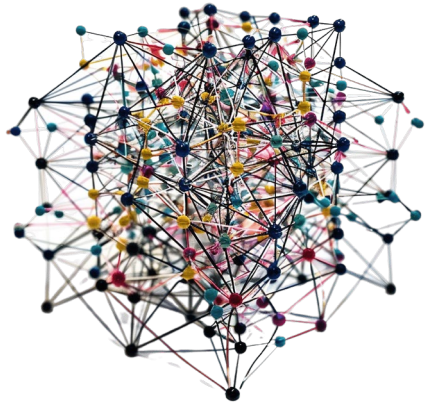
Hybrid JHPC-Quantum Infrastructure Project
Deployment FY2023~2027

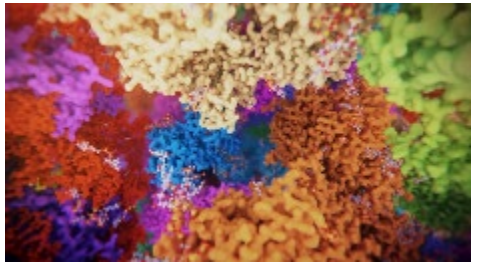
- **Science by Quantum-HPC Hybrid Computing**

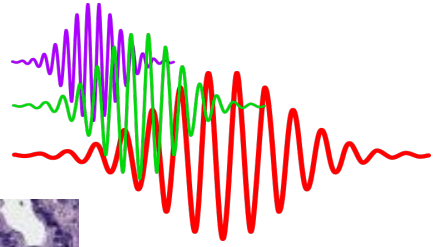


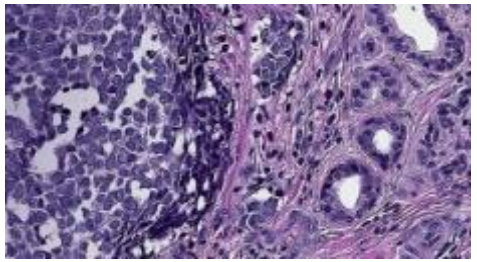
Daily life Data vs. Science Data

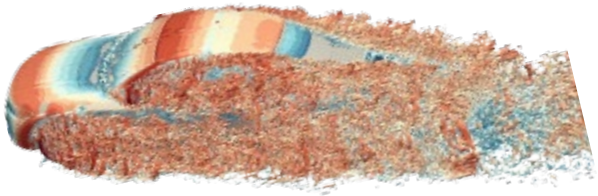












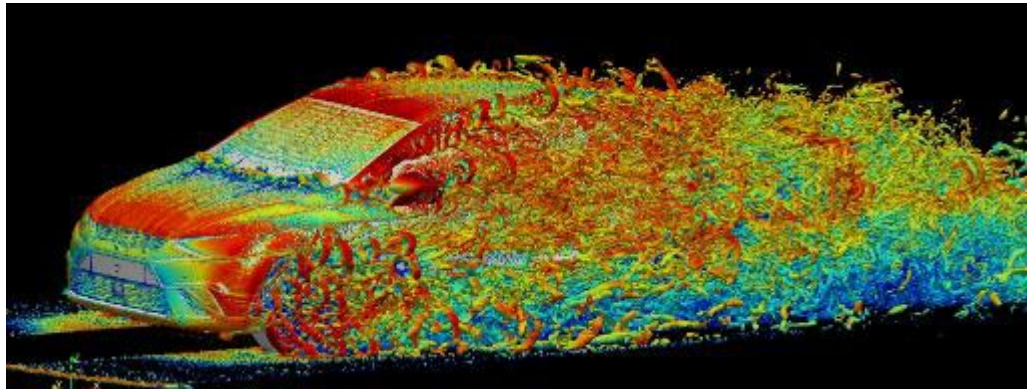
Commercial Models ←GAP→ Scientific Requirements

↑ financial incentive ↓ complexity

↓ financial incentive ↑ complexity

1 Vision AI for Spatiotemporal Applications in Science and Engineering

Complex Spatial-Temporal Data Is Common for Science & Engineering.



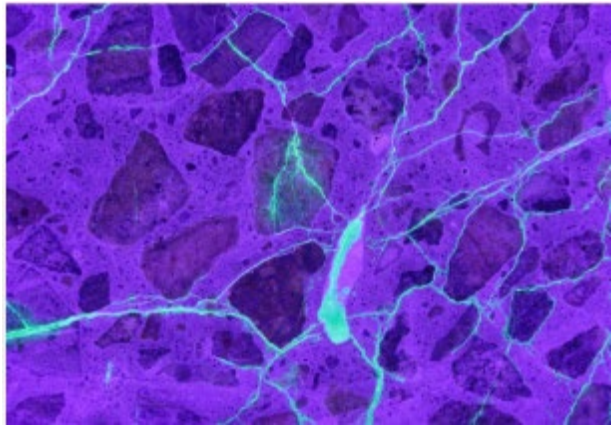
Fluid dynamics



Climate/Weather Prediction



Satellite Remote Sensing



X-ray/electron microscopy



Industrial inspection



Synchrotron beamline

1 Vision AI for Spatiotemporal Applications in Sci&Eng.

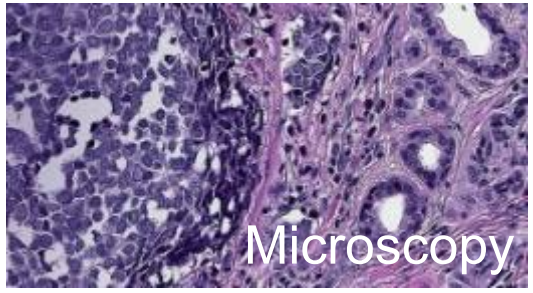
Pattern Discovery

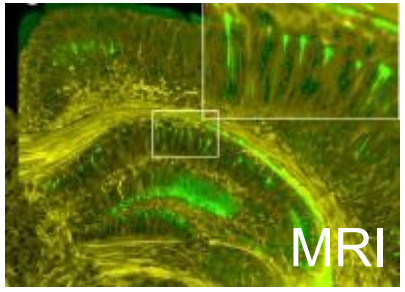
Predictive Modeling

Inverse Problem

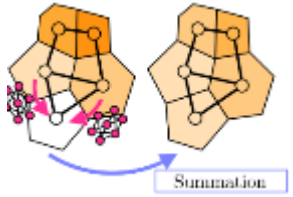
Generative

Ultra-resolution Vision Models



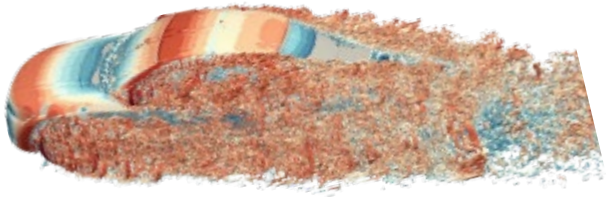


Physics Inspired NNs



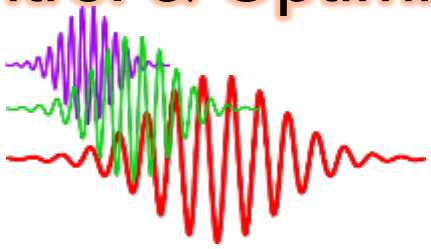


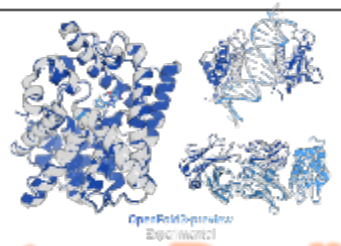




Physical World Models

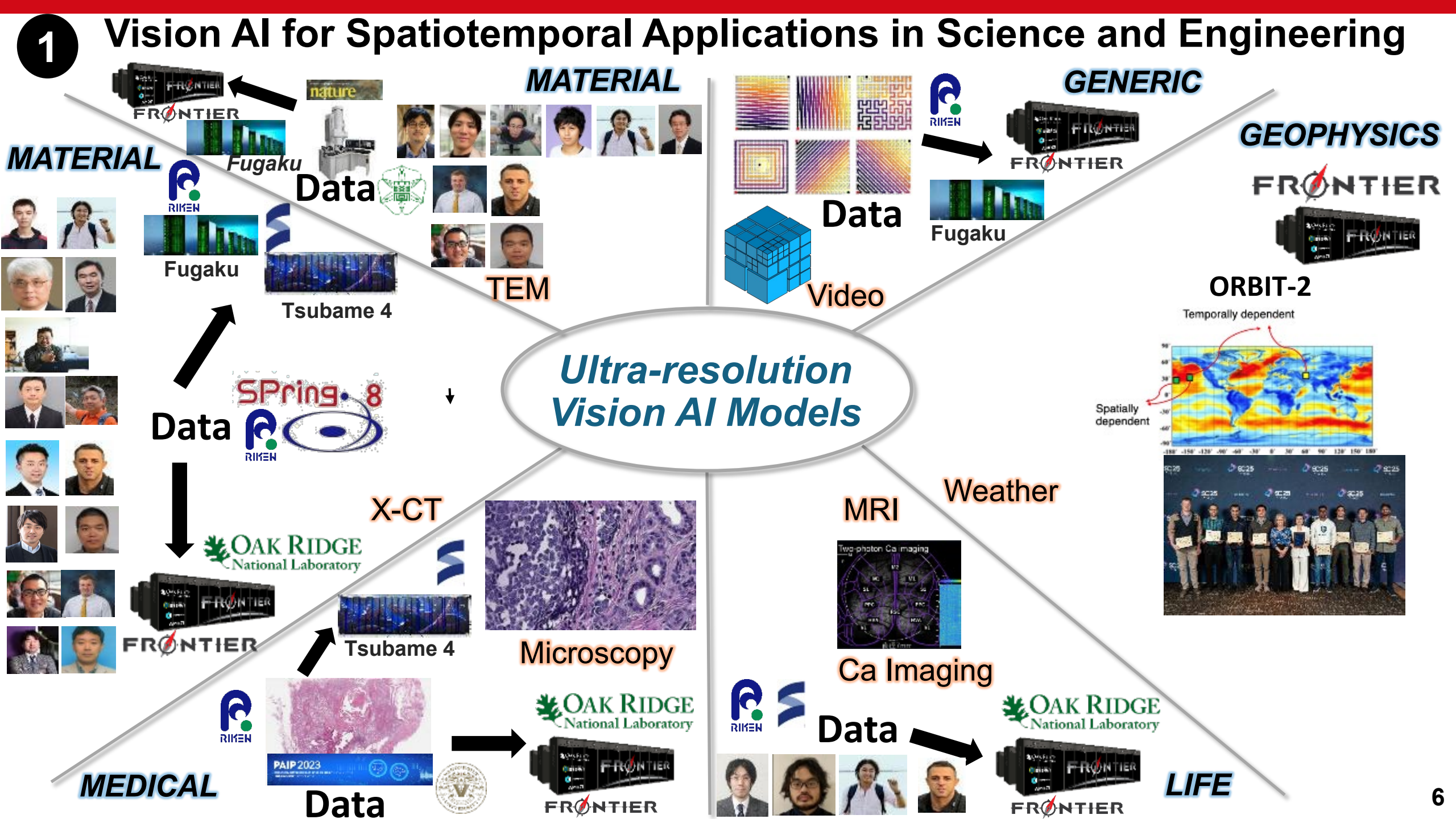
Control & Optimization





Structure Prediction

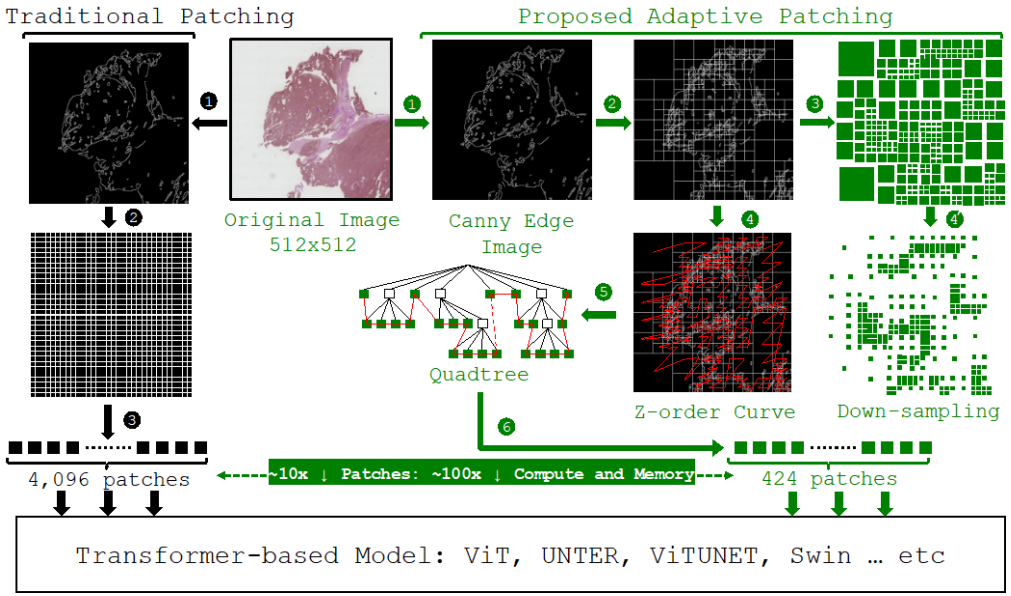
1 Vision AI for Spatiotemporal Applications in Science and Engineering

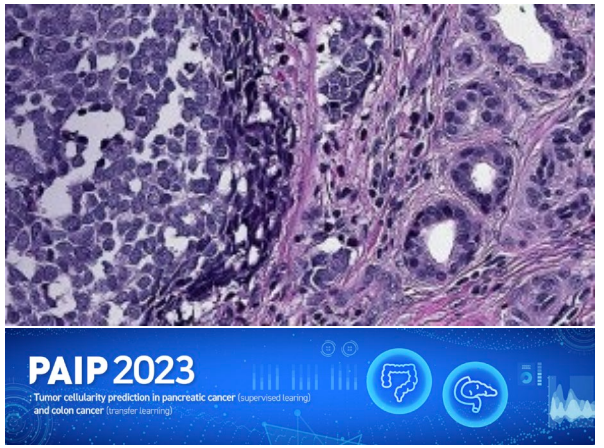


Vision AI Use in Real-world Problems

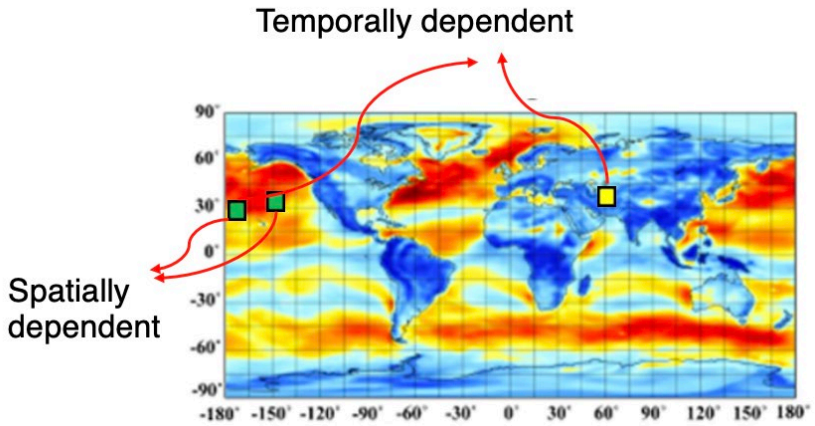
Foundation Vision AI used in production, Examples:

a) microscopic pathology, b) X-ray CT road samples, c) weather prediction

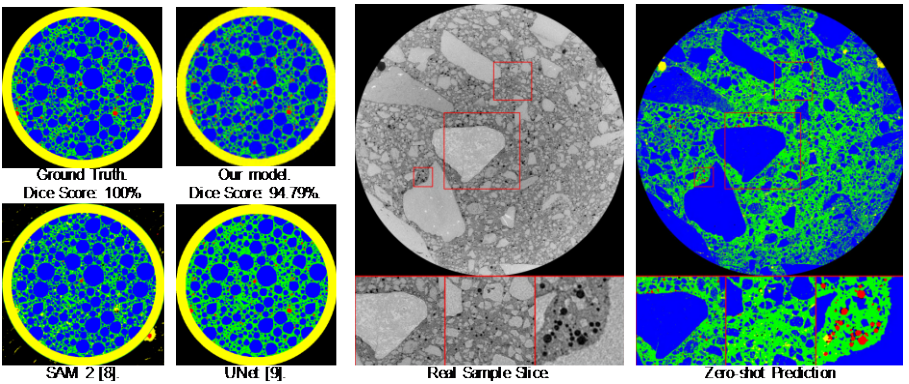


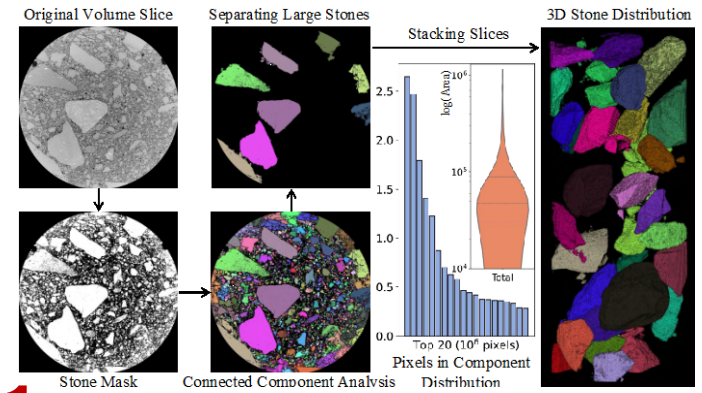


3

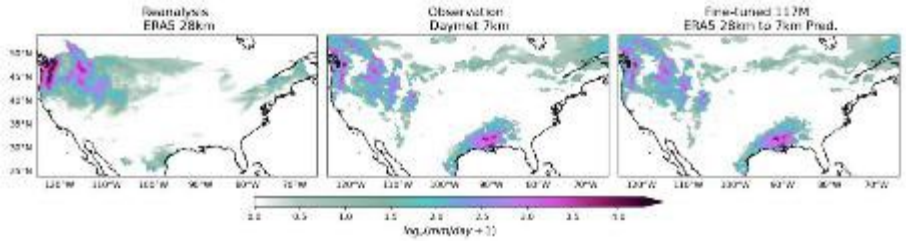


2



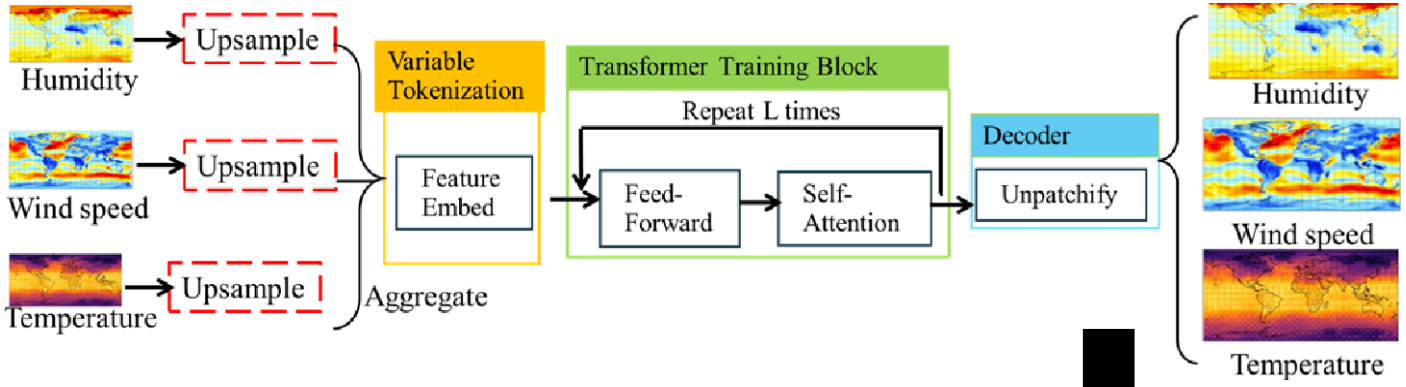


1



1 2 3

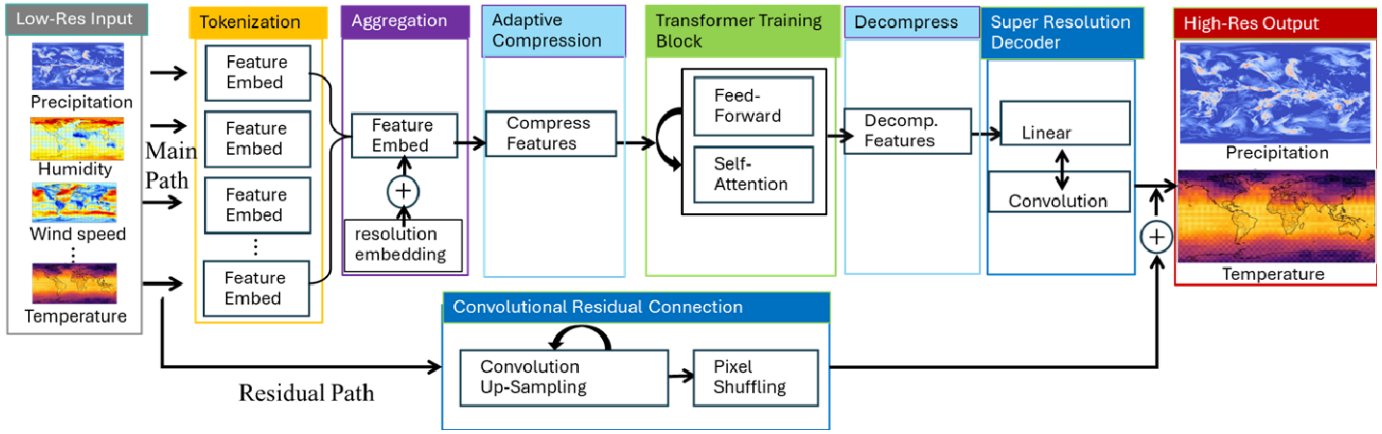
ORBIT-2: Scaling Exascale Vision Foundation Models for Weather and Climate Downscaling

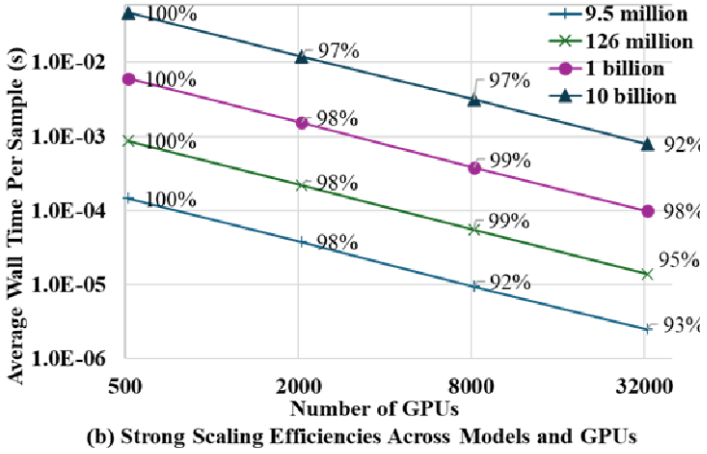




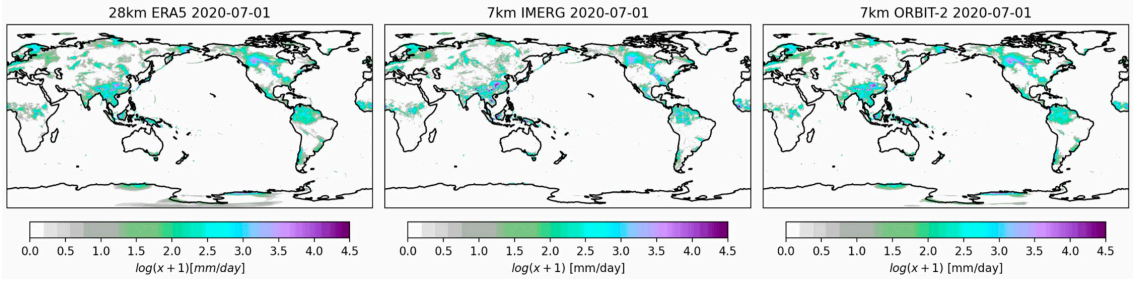
SC'25 Best Paper Award

- Operating directly on compressed inputs
- Condition prediction for reduced uncertainty





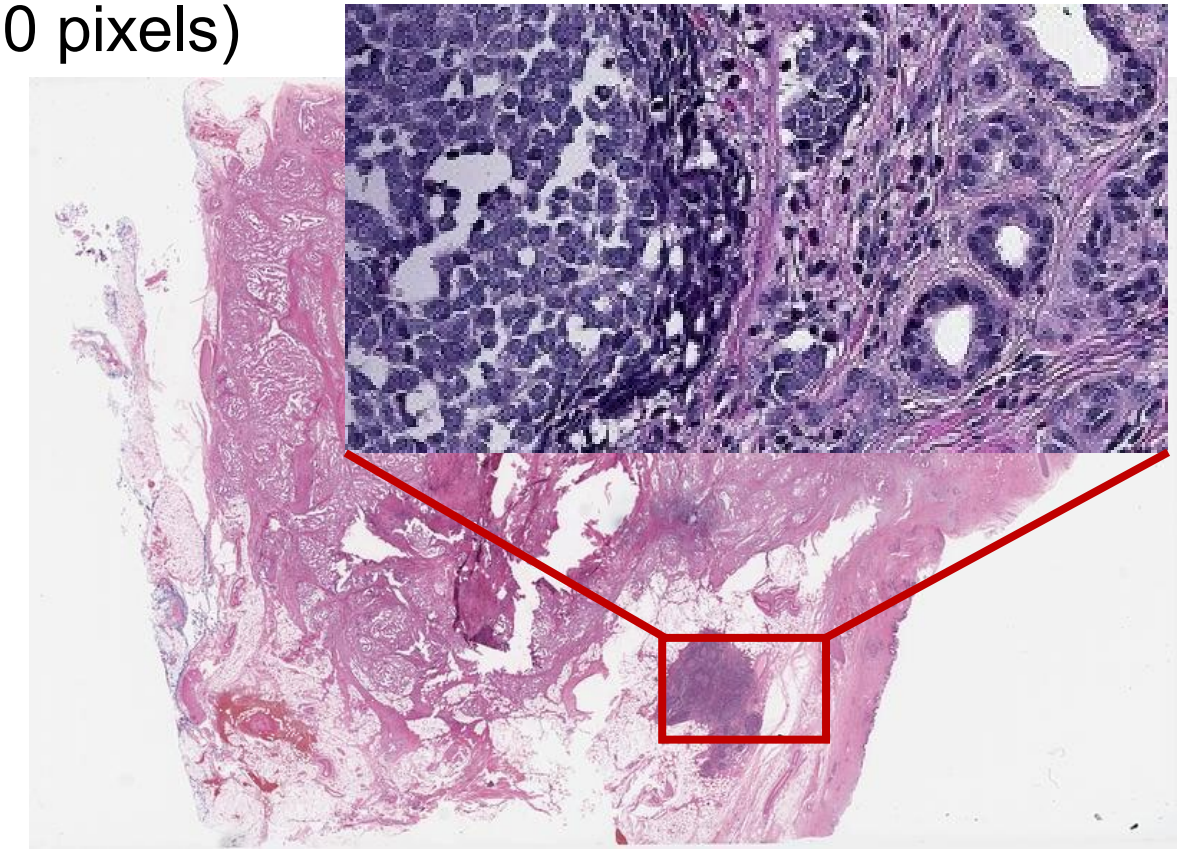
(b) Strong Scaling Efficiencies Across Models and GPUs



Xiao et al. SC'25 [BEST PAPER AWARD]

Tumor Cellularity Prediction in Pancreatic Cancer and Colon Cancer

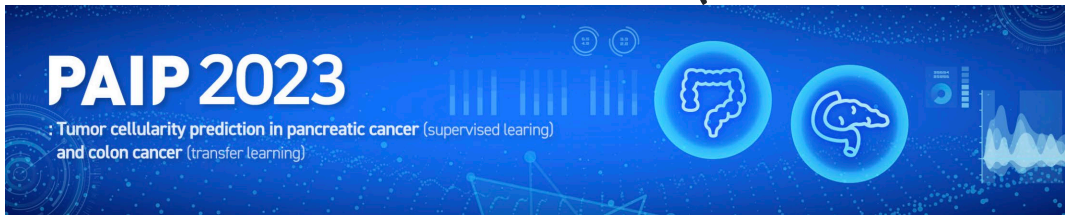
- Very high resolution (up to 100,000 x 100,000 pixels)
  - Used in pathology
  - Ex: PAIP dataset
    - Pancreas
    - **Diagnostic:** Perineural Invasion



- Segmentation with Vision Transformer (ViT)

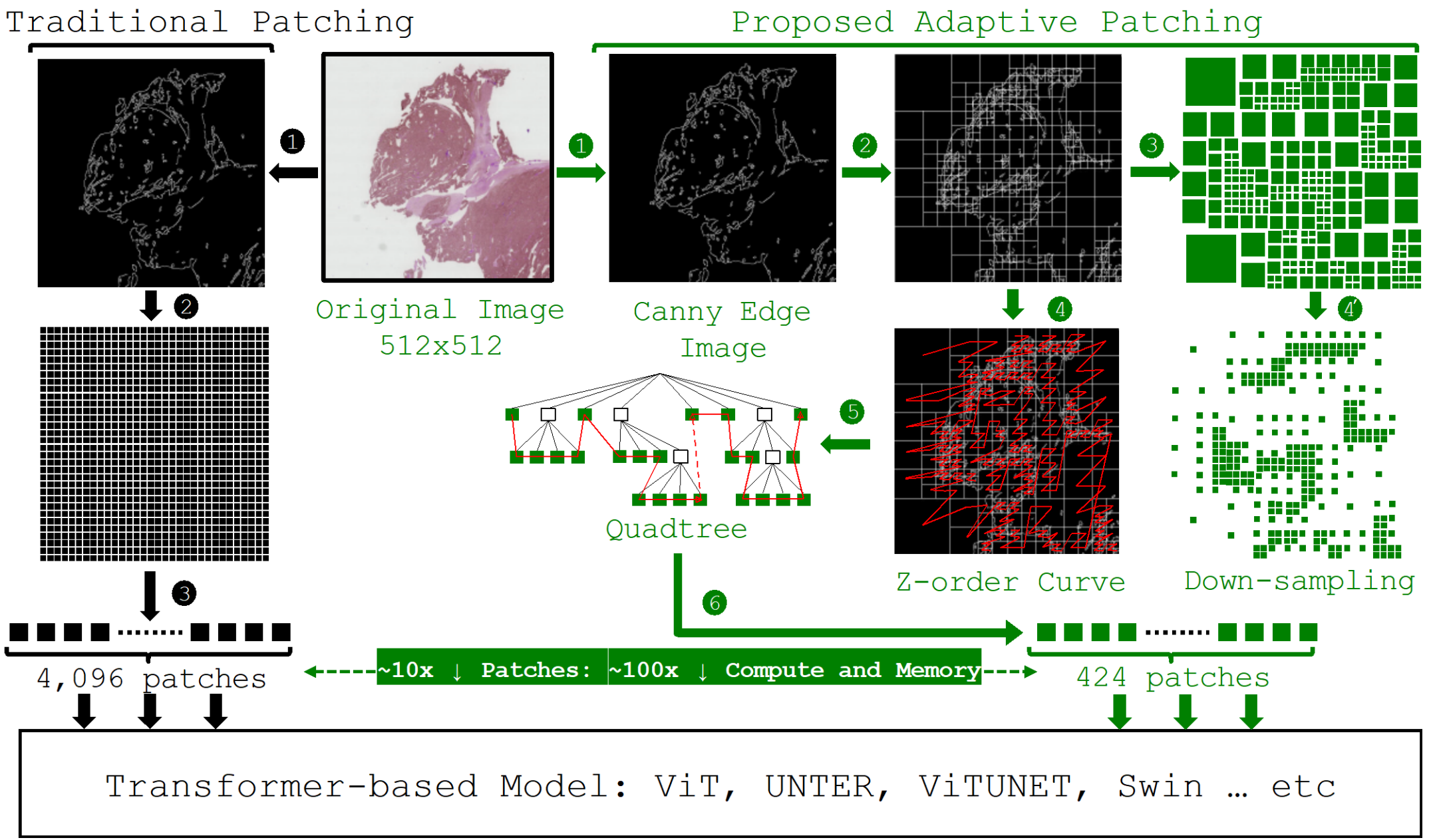
- Challenge:

  PAIP 2023: Tumor cellularity prediction in pancreatic cancer and colon cancer (transfer learning)



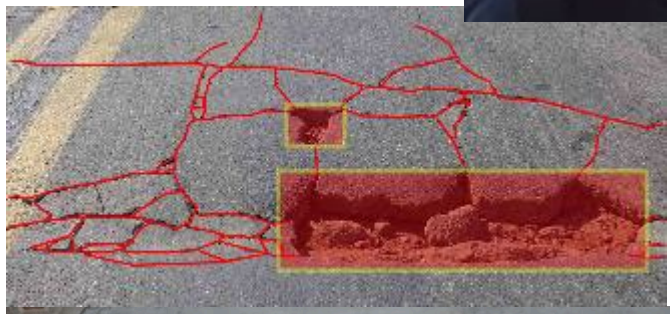
* E. Zhang et al. Adaptive Patching for High-resolution Image Segmentation with Transformers, SC'24

Tumor Cellularity Prediction in Pancreatic Cancer and Colon Cancer



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



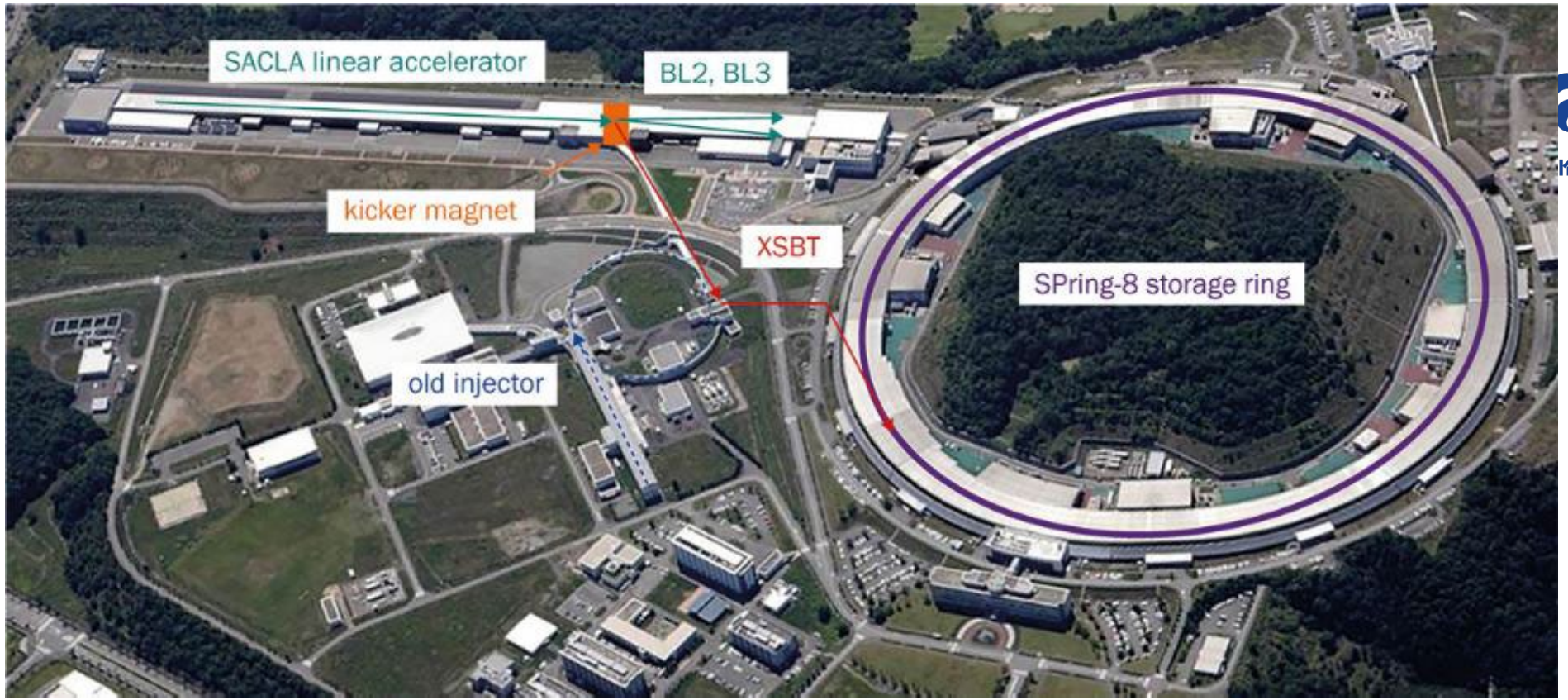




- Machines mounted on vehicles
- Extract cylindrical samples from core of asphalt layers
- Scan (projections) at RIKEN Spring-8 Synchrotron
- Move projections to R-CCS (or other HPC facilities) **for HPC-AI processing**
  - High-performance high-resolution CT image reconstruction
  - 3D volumetric segmentation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

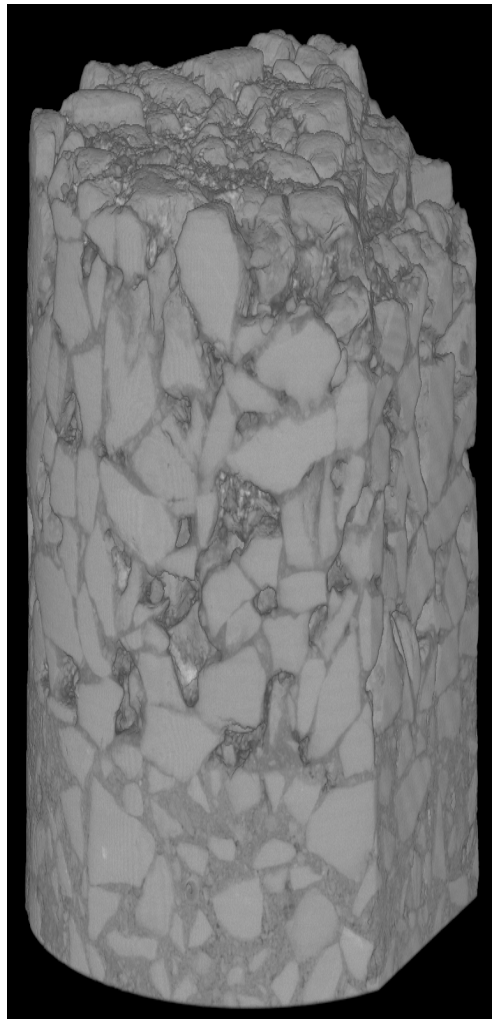
SPring-8

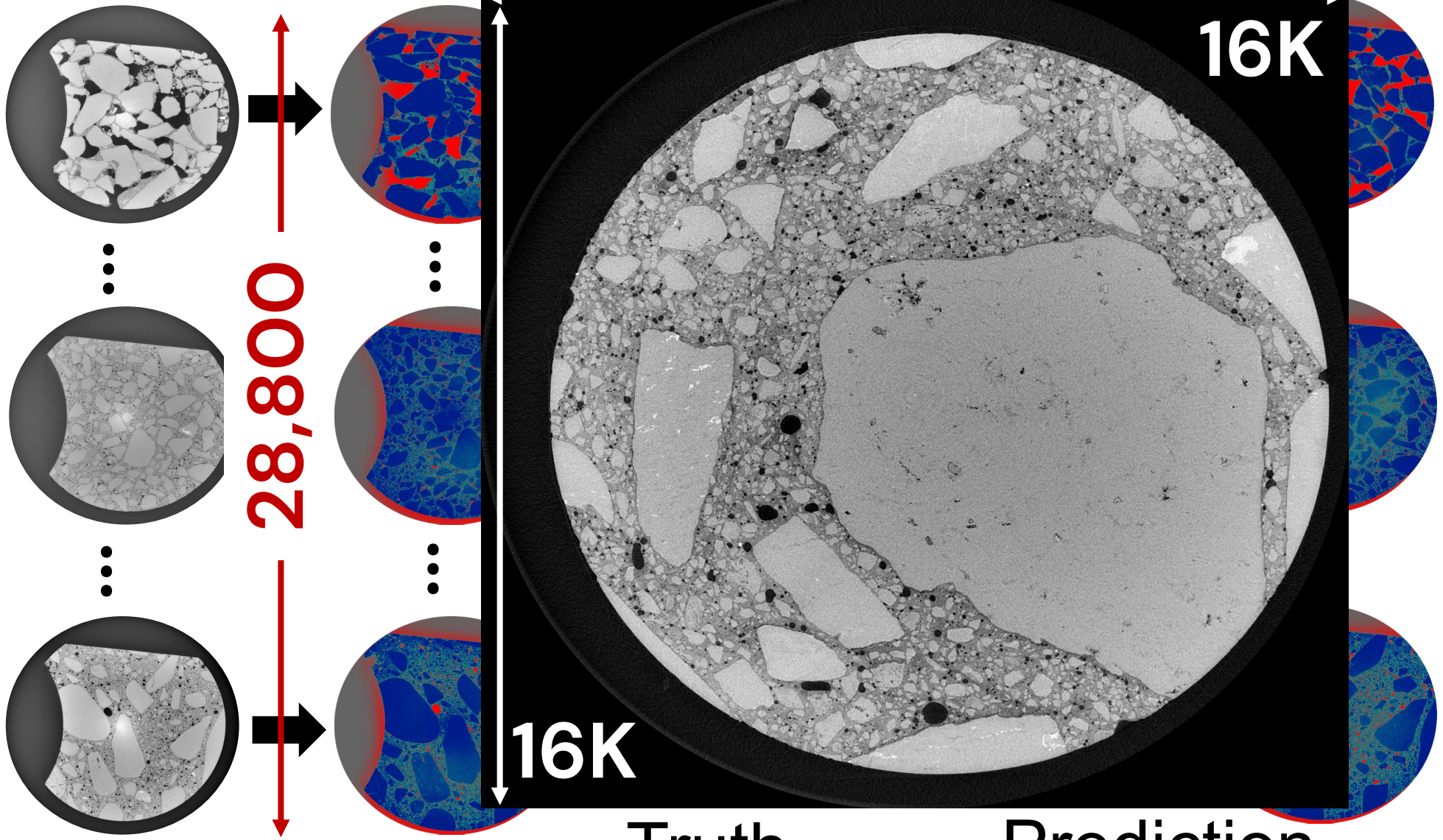
Hanshin Highway Co.

Riken Spring-8 + Sacla Synchrotron Light Source Facility

Why NO LABELD DATA?

x46





Hierarchy-aware Masked Autoencoder as an Effective Ultra-High Resolution Scientific Image Learner

We teach the model by hiding information hierarchically, so it learns micro-structures at multiple scales.

* Zhang et al., Under review

High Quality Segmentation WITHOUT LABELS

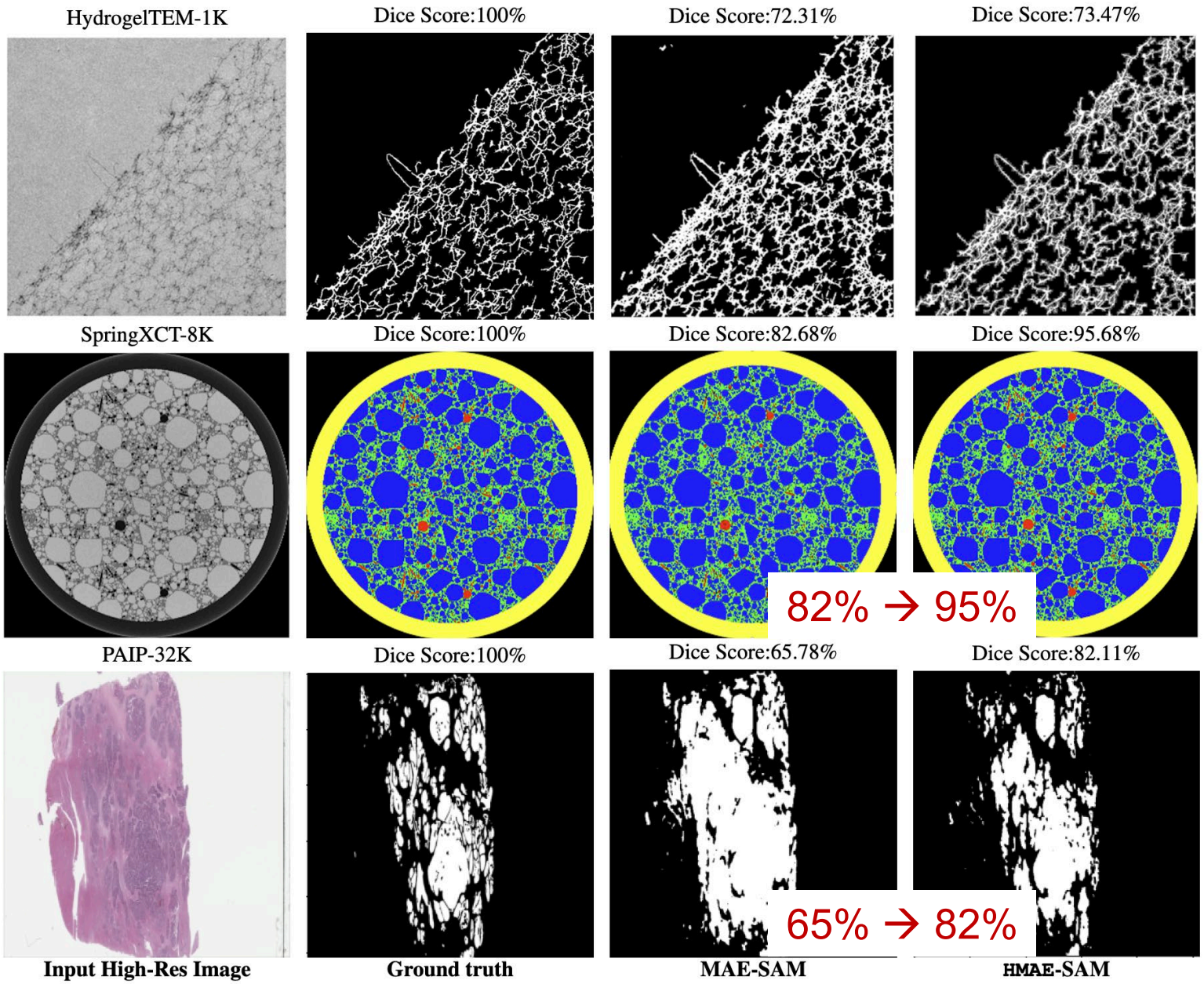
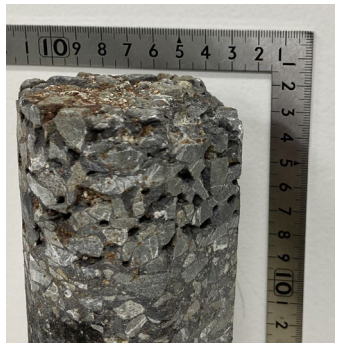


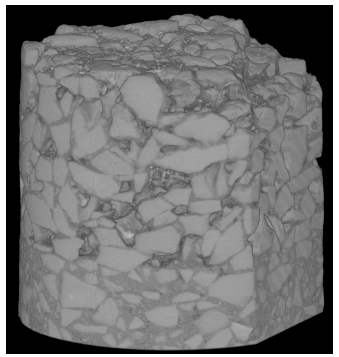
Table 1. Quantitative segmentation results (Dice Score % ↑) on all datasets. Our HMAE pre-training method is compared against standard baselines and MAE pre-training on identical architectures. The "Improvement" column shows the Dice Score gain of our HMAE model over its equivalent MAE-trained baseline (e.g. FT-HMAE-UNETR vs. FT-MAE-UNETR).

Dataset	Model	Patch Size	Seq. Length	Dice (%)	Improvement
HydrogelTEM-1K	SAM 2	8	16384	69.87	-
	UNet	-	-	71.03	-
	UNETR	16	4096	70.59	-
	FT-MAE-SAM	8	16384	72.31	(Baseline)
	FT-MAE-SAM2	8	16384	73.12	(Baseline)
	FT-MAE-UNETR	16	4096	74.56	(Baseline)
	FT-HMAE-SAM	2	16384	73.48	+0.36 (vs. FT-MAE-SAM)
	FT-HMAE-SAM2	2	16384	74.23	+1.11 (vs. FT-MAE-SAM2)
	FT-HMAE-UNETR	2	16384	**75.33**	**+0.77** (vs. FT-MAE-UNETR)
SpringXCT-8K	SAM 2	128	4096	83.81	-
	UNet	-	-	78.73	-
	UNETR	128	4096	82.23	-
	Swin-UNETR	-	-	83.96	-
	FT-MAE-SAM	128	4096	82.68	(Baseline)
	FT-MAE-SAM2	128	4096	85.98	(Baseline)
	FT-MAE-UNETR	128	4096	86.12	(Baseline)
	FT-HMAE-SAM	2	16384	**95.68**	**+9.70** (vs. FT-MAE-SAM)
	FT-HMAE-SAM2	2	16384	93.77	+7.79 (vs. FT-MAE-SAM2)
	FT-HMAE-UNETR	4	8194	91.93	+5.81 (vs. FT-MAE-UNETR)
PAIP-32K	SAM	1024	1024	62.34	-
	UNet	-	-	61.38	-
	UNETR	1024	1024	74.96	-
	TransUNet	-	-	69.88	-
	FT-MAE-SAM	1024	1024	65.78	(Baseline)
	FT-MAE-SAM2	1024	1024	66.37	(Baseline)
	FT-MAE-UNETR	1024	1024	77.24	(Baseline)
	FT-HMAE-SAM	4	16384	82.11	+16.33 (vs. FT-MAE-SAM)
	FT-HMAE-SAM2	4	16384	**83.21**	**+16.84** (vs. MAE-SAM2)
	FT-HMAE-UNETR	8	8194	81.23	+3.99 (vs. FT-MAE-UNETR)

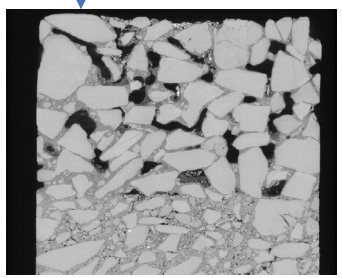
Contribution Towards R-CCS and Spring-8 Mission (X-ray CT)

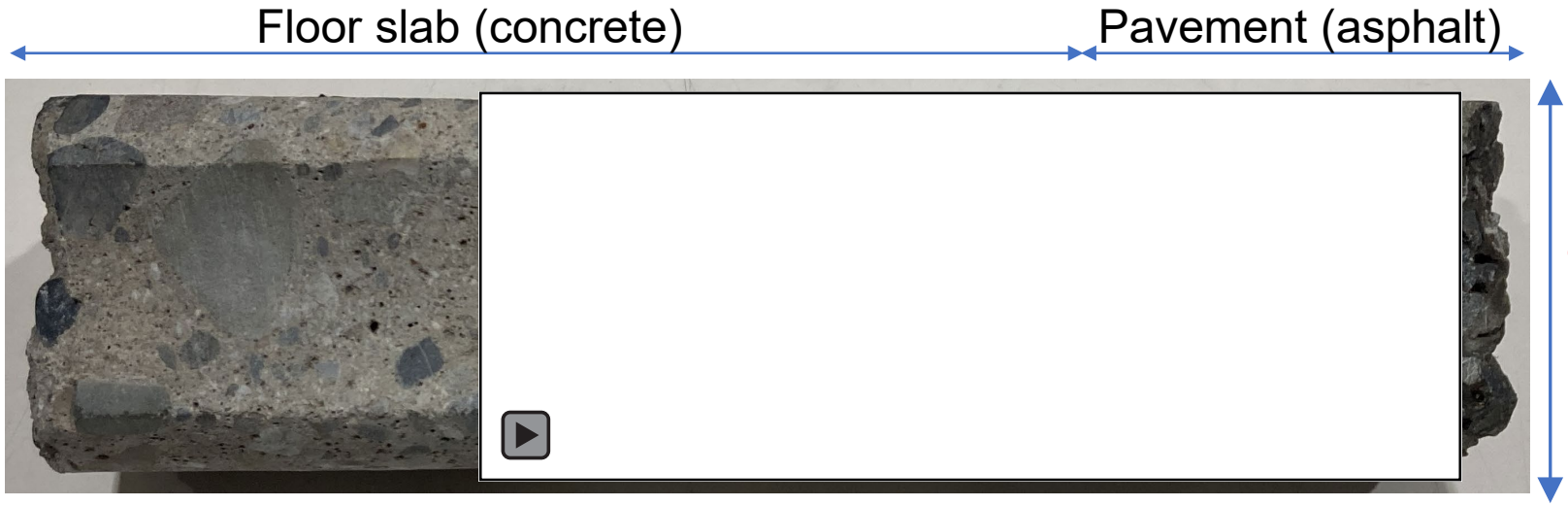


CT撮影

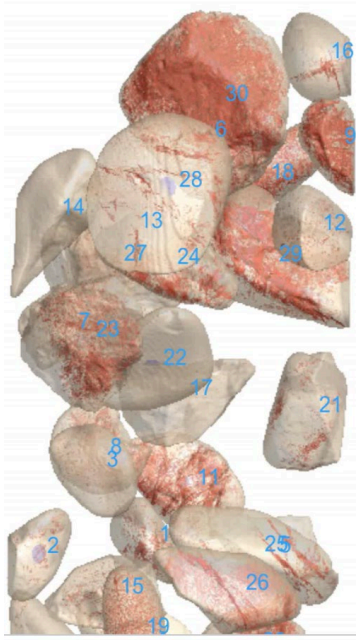


内部断面





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

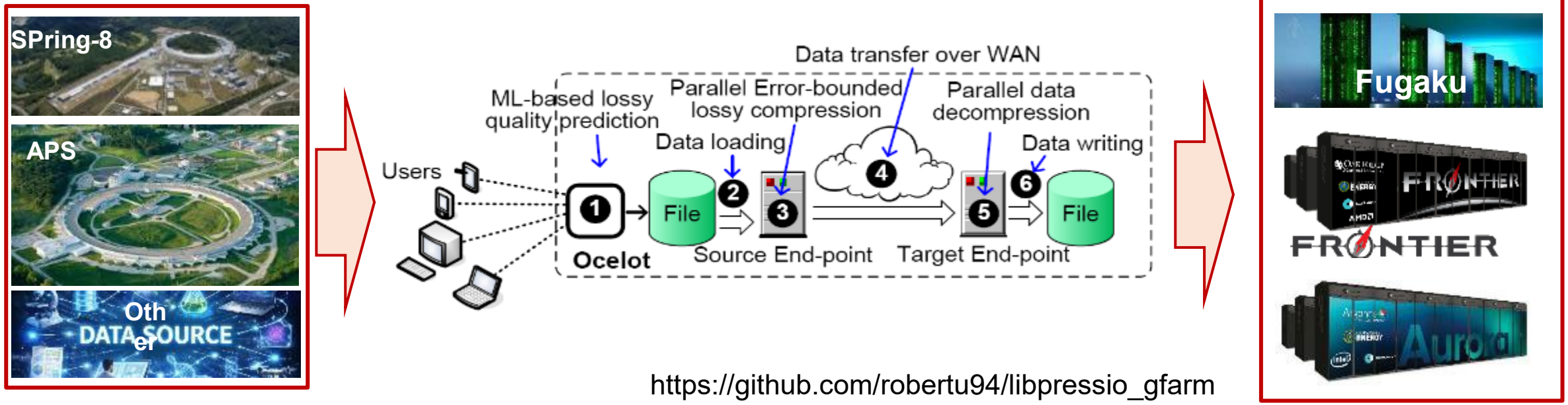
7 seconds per 16^3 volume on ~8% of Fugaku nodes

3 Data transfer infrastructure for AI for Science

LibPressio for SPring-8 and data transfer

Data transfer infrastructure

- **Participants**
  - DOE: R. Underwood, S. Di, F. Cappello (ANL)
  - MEXT: K. Sato, A. Singh (RIKEN)
- **Aim**
  - Integrate Compression into Data Management Solutions for SPring-8 to accelerate transfers and save storage space
- **Technical Approach**
  - For Gfarm, integrate LibPressio as a TAR archive filter
  - For Globus, integrate LibPressio using a Globus Compute Function



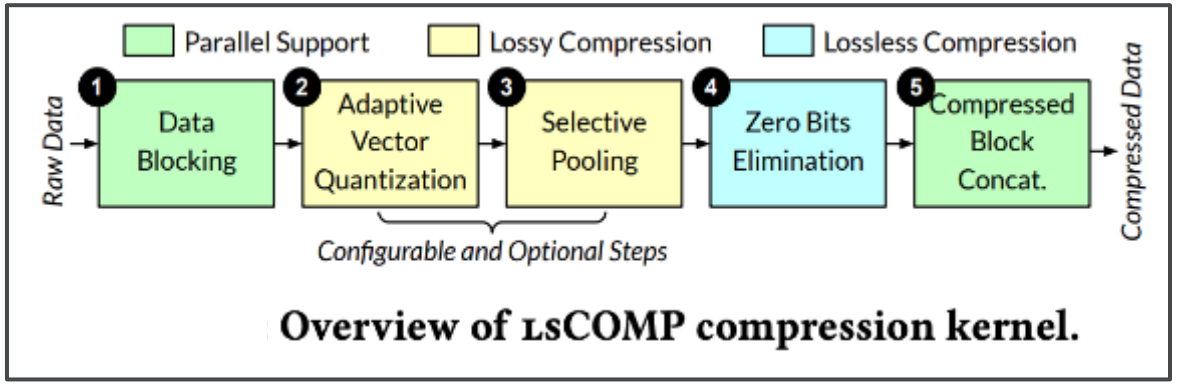
https://github.com/robertu94/libpressio_gfarm

3 Data transfer infrastructure for AI for Science

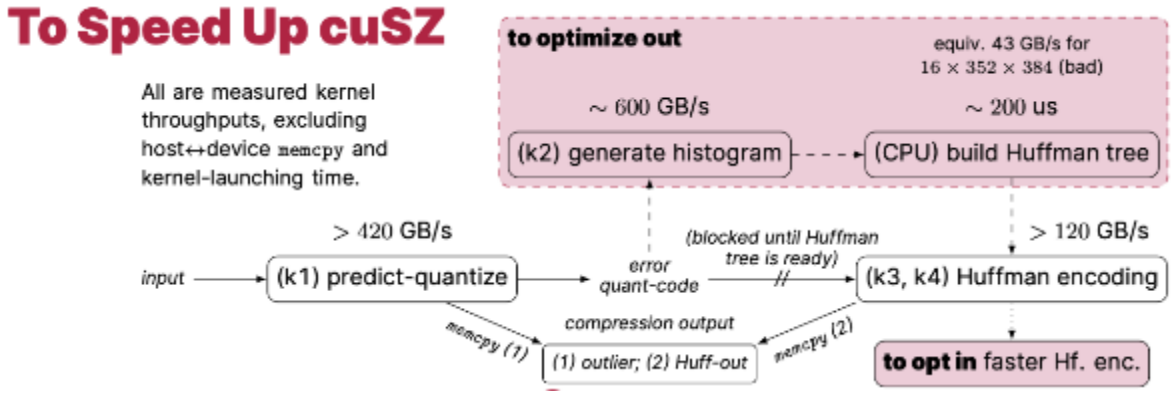
Data compression for SPring-8/SACLA [Ptychography]

Data compression

- **Participants**
  - DOE: R. Underwood, S. Di, F. Cappello (ANL)
  - MEXT: K. Sato, A. Singh, Y. Joti, T. Hatsui (RIKEN)
- **Aim**
  - Significantly reduce the size of SACLA Ptychography dataset
- **Technical Approach**
  - EIP leverages common encoding structures between images (e.g. huffman trees), and improved data management (e.g. concurrent streams, graph kernel launches, etc…) to dramatically improve performance for many small images.
  - lsCOMP improves performance for large images by using adaptive vector encoding + selective pooling to better losslessly and lossily compress sparse aspects of light source data.
  - Integrate the quality metric as a LibPressio external metric



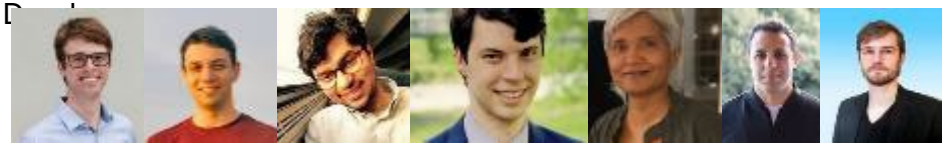
Overview of lsCOMP compression kernel.



lsCOMP: Efficient Light Source Compression: Yafan Huang, Sheng Di, Robert Underwood, Peco Myint, Miaoqi Chu, Guanpeng Li, Nicholas Schwarz, and Franck Cappello. (in Submission to SC25)
EIP: Managing Streaming Data in Heterogeneous Scientific Application Using Error-Bounded Lossy Compression. Jiannan Tian, Robert Underwood, Sheng Di, Chris O Grady, Stefano Merchandini, Xin Liang, Franck Cappello. (In Preparation)

4 AI-HPC SW: RAPTOR – Numerical Profiling of Scientific Applications

Faveo Hoerold, Ivan Ivanov, Akash Dhruv, William Moses, Anshu Dubey, Mohamed Wahib, Jens



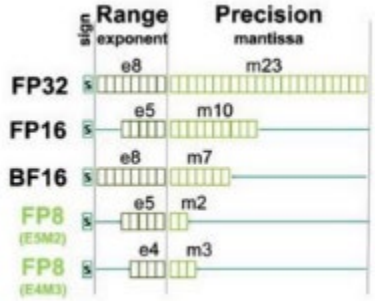


SC'25 Best Reproducibility Advancement Award

Motivation:

- Historically, easy choice: FP32 enough or is FP64 necessary?
- AI induced HW trends: low-precision units & FP64 capabilities reduced



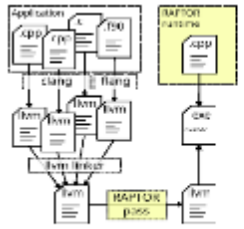


Goals/Features:

- Easily applicable, support arbitrary precisions & numerical error analysis

Design Approach:

- LLVM compiler pass and a supporting runtime using MPFR



(a) RAPTOR's components

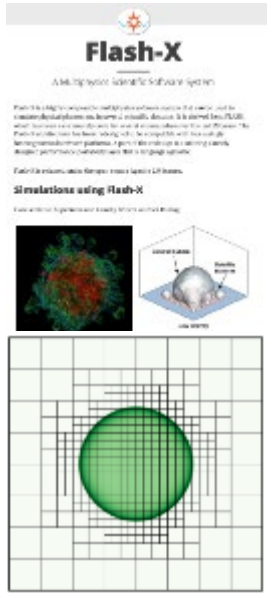
Scope	op-mode	mem-mode
Function	Fully Auto	Semi Auto.
File	Fully Auto	N/A
Program	Fully Auto	N/A

Features	op-mode	mem-mode
Profile	Local	Extensive
Precision Increase	N/A	Supported

(b) Supported scopes and features for each mode

Figure 2: Overview of RAPTOR and config. matrix

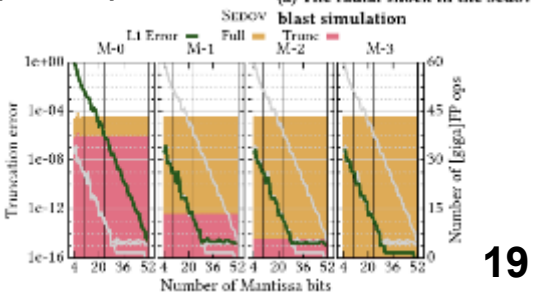
Flash-X: multi-physics multi-domain application software



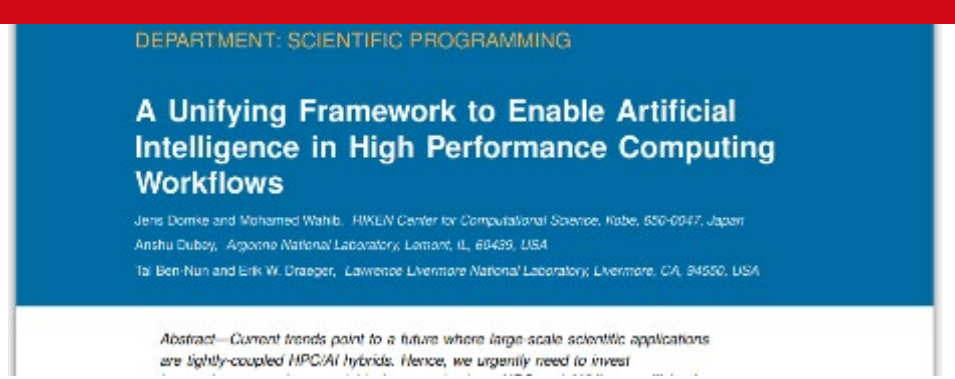
(a) The radial shock in the Sedov blast simulation

- Simulation domain divided into blocks with **AMR**
- **Shock wave** in radial direction
- **Hypothesis 1**: using reduced precision in less-refined AMR blocks w/o quality loss

➔ Sedov problem is very robust to low-precision



5 AI-HPC SW: A Unifying Framework to Enable AI in HPC Workflows


DEPARTMENT: SCIENTIFIC PROGRAMMING

A Unifying Framework to Enable Artificial Intelligence in High Performance Computing Workflows

Jens Domke and Mohamed Wahib, RIKEN Center for Computational Science, Kobe, 650-0047, Japan
Anshu Dubey, Argonne National Laboratory, Lemont, IL, 60439, USA
Tal Ben-Nun and Erik W. Draeger, Lawrence Livermore National Laboratory, Livermore, CA, 94550, USA

Abstract—Current trends point to a future where large scale scientific applications are tightly-coupled HPC/AI hybrids. Hence, we urgently need to invest

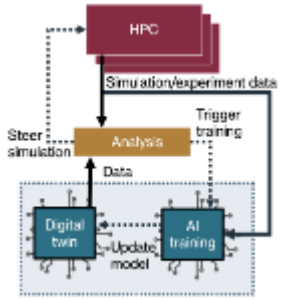
IEEE Computing in Science & Engineering (CiSE)

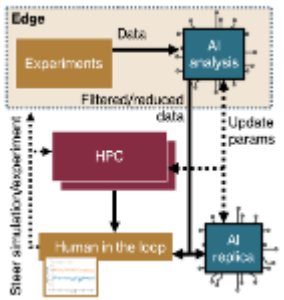
Jens Domke, Tal Ben-Nun, Anshu Dubey, Mohamed Wahib, Erik Draeger, Maximilian Sander, Theresa Pollinger, Youngjun Lee



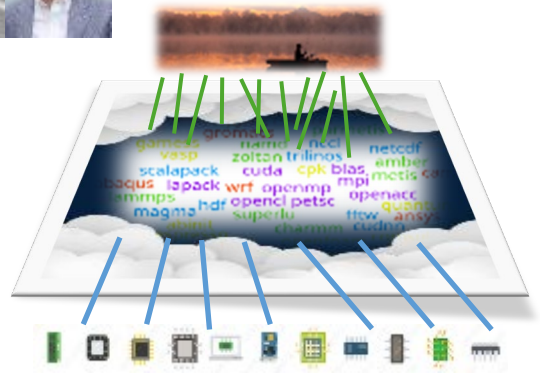
Motivation:

- Future large-scale scientific app are undoubtedly HPC/AI hybrids.
- We urgently need to invest in creating a seamless, scalable





Design → Approach:

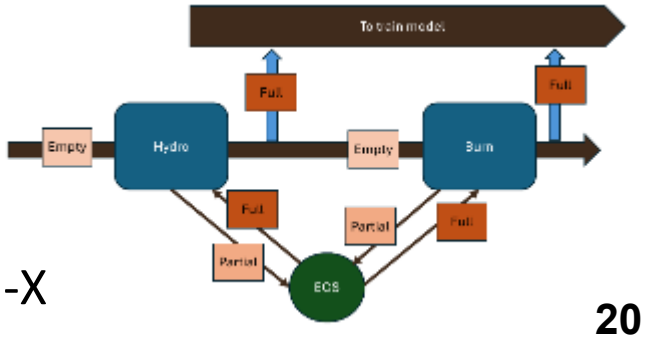


- "Lake" of **user-facing function & data structures**
- Best HPC algorithms combined into **one 'framework' or simulation+AI platform**
- **Effective hardware mapping**

Phase 1 Prototype with FLASH-X, dalotia, ring-buffer, and PyTorch



- Example: **Cellular Detonation**
- Data extraction for **in-situ training** with novel ring-buffer for data staging
- **AI surrogate** model for **equation-of-state** function and **inference via dalotia** inside FLASH-X



The new phase of AI4S Foundation Models



Challenge 1

Semi-structured;
Sparse; High Dim.

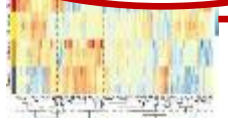
Experiments

Challenge 2

Multi-modality

Mod/Sim

Robots

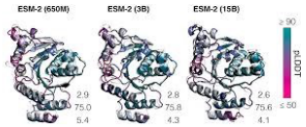


Genome/Transcriptome

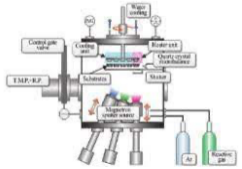


Other Phenotypes

Foundation Model for Drug Design



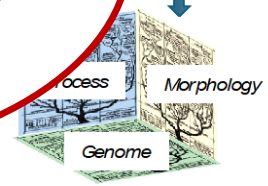
Foundation Model for Materials Science



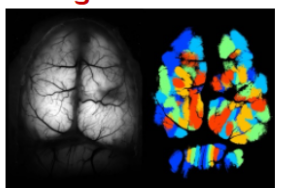
Foundation Model for Cellular Response



on Model for Behavior



Foundation Model for Scientific Imaging 3D Segmentation



Information Retrieval Agent

Information Retrieval Agent

Information Retrieval Agent

Information Retrieval Agent

Challenge 3

LLM Understanding Agents

Frontier LLM

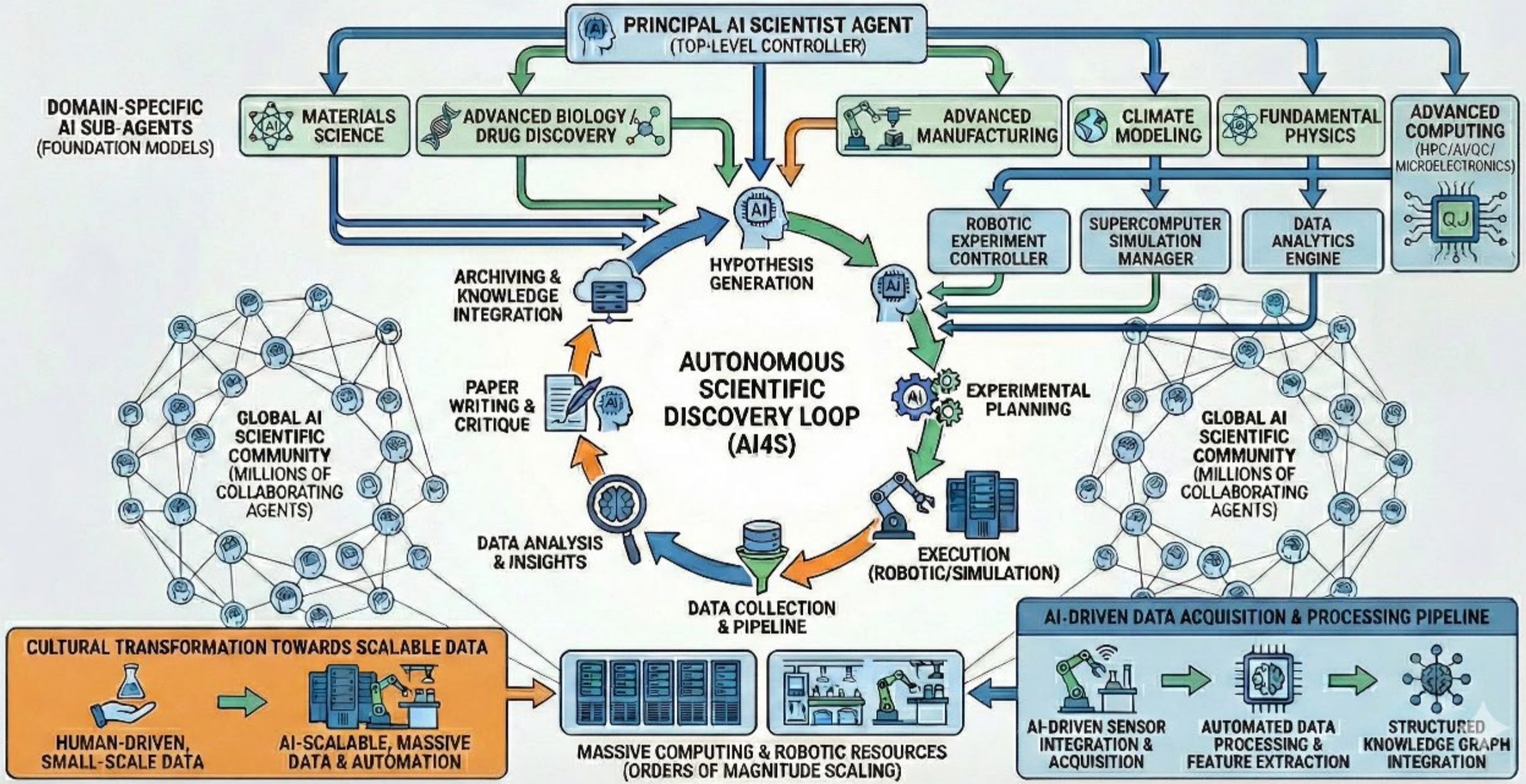
Reasoning



"Identify patterns when neural activity of property X correlates with behavior Y, then suggest how would that manifest as biological spiking neural network with dynamics"



FUTURE OF AI4S: AUTONOMOUS, COLLABORATIVE, AND SCALABLE SCIENTIFIC INNOVATION

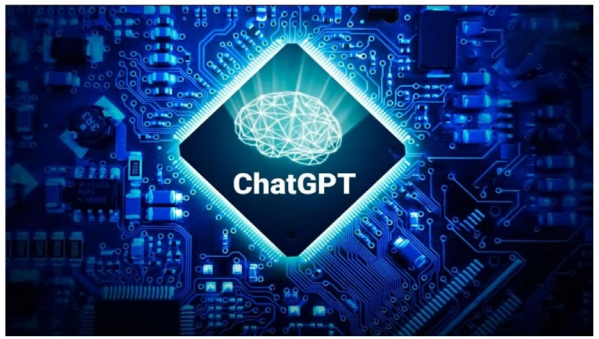


LifeScience-DX AI co-scientist [Okuno HPC/AI Pharma Div.]
including the world's largest ecosystem of 1,164 databases and tools

In response to a user's input, this autonomous computational agent system selects the required tools from an ecosystem of 1,164 databases and tools, then uses a CodeAct agent to plan and execute the workflow.

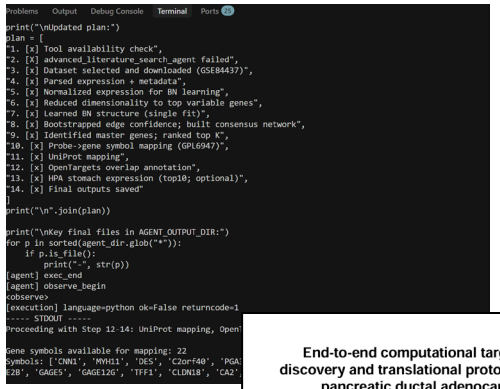


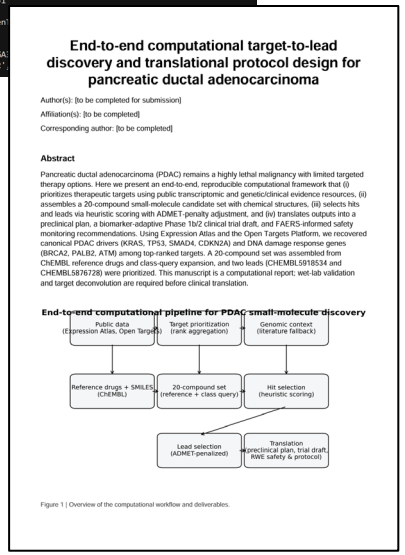
User's input:
Research Hypotheses/Scientific Motivation



Tool Discovery & Retrieval → Workflow Planning → Execution → Output Generation

Output: Calc. Results




End-to-end computational target-to-lead discovery and translational protocol design for pancreatic ductal adenocarcinoma

Author(s): [to be completed for submission]

Affiliation(s): [to be completed]

Corresponding author: [to be completed]

Abstract

Pancreatic ductal adenocarcinoma (PDAC) remains a highly lethal malignancy with limited targeted therapy options. Here we present an end-to-end, reproducible computational framework that (i) prioritizes therapeutic targets using public transcriptomic and genetic/clinical evidence resources, (ii) assembles a 20-compound small-molecule candidate set with chemical structures, (iii) selects hits and leads via heuristic scoring with ADMET-penalty adjustment, and (iv) translates outputs into a preclinical plan, a biomarker-adaptive Phase 1b/2 clinical trial draft, and FAERS-informed safety monitoring recommendations. Using Expression Atlas and the Open Targets Platform, we recovered canonical PDAC drivers (KRAS, TP53, SMAD4, CDKN2A) and DNA damage response genes (BRCA2, PALB2, ATM) among top-ranked targets. A 20-compound set was assembled from ChEMBL reference drugs and class-query expansion, and two leads (CHEMBL5918534 and CHEMBL5876728) were prioritized. This manuscript is a computational report: wet-lab validation and target deconvolution are required before clinical translation.

End-to-end computational pipeline for PDAC small-molecule discovery

Public data (Expression Atlas, Open Targets)
Target prioritization (rank aggregation)
Genomic context (literature fallback)
Reference drugs + SMILES (ChEMBL)
20-compound set (reference + class query)
Hit selection (heuristic scoring)
Lead selection (ADMET-penalized)
Translation (preclinical plan, trial draft, RWE safety & protocol)

Figure 1 | Overview of the computational workflow and deliverables.

LLMs have already read and internalized vast numbers of scientific papers, acquiring knowledge such as:

- **Research motivations**
- **Research processes**
- **Experimental protocols**
- **Known results and interpretations**

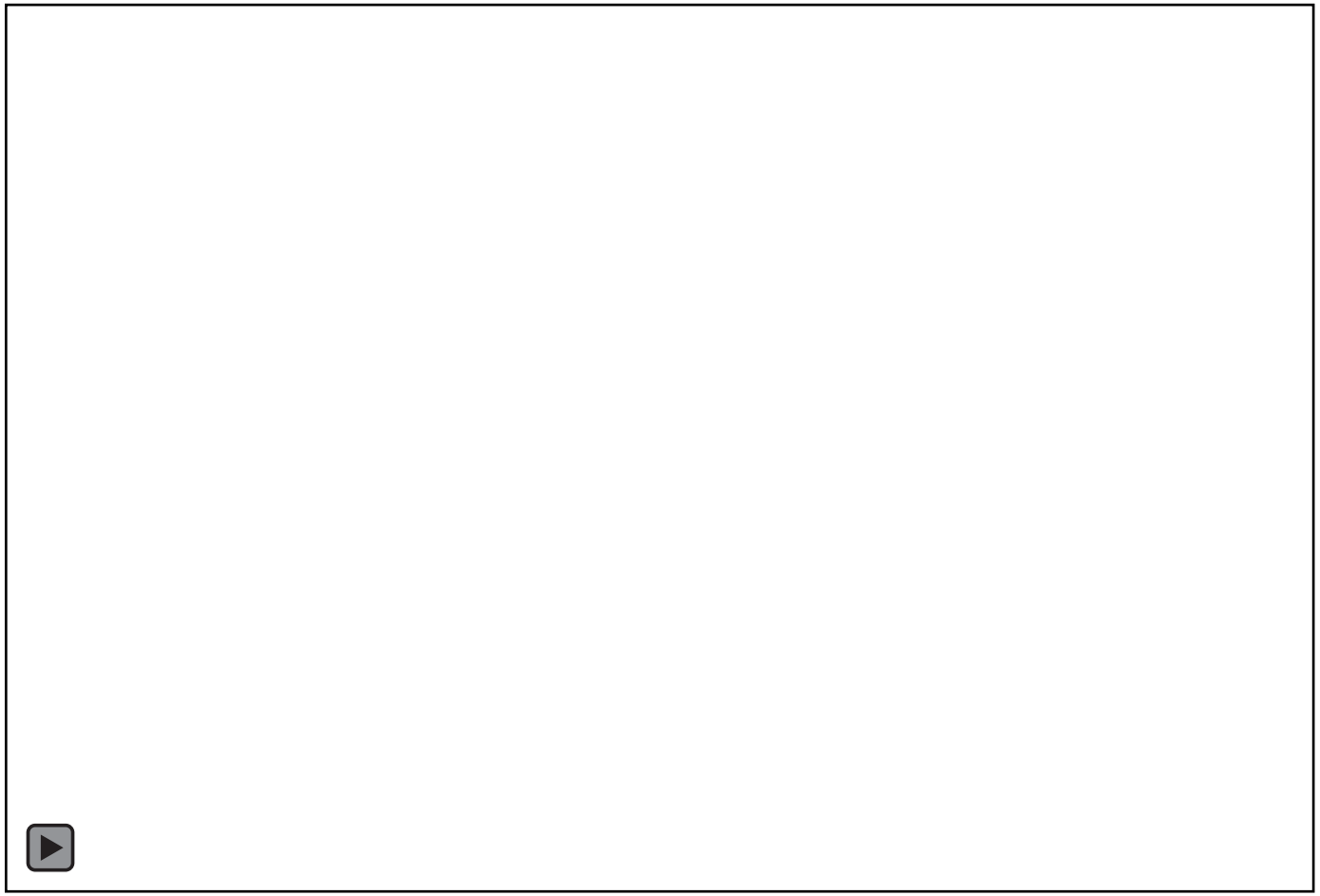
1,164 DBs/tools

including Original tools, and Biomni 150 tools (from Stanford) , ToolUniverse 700 tools (from Harvard and MIT)

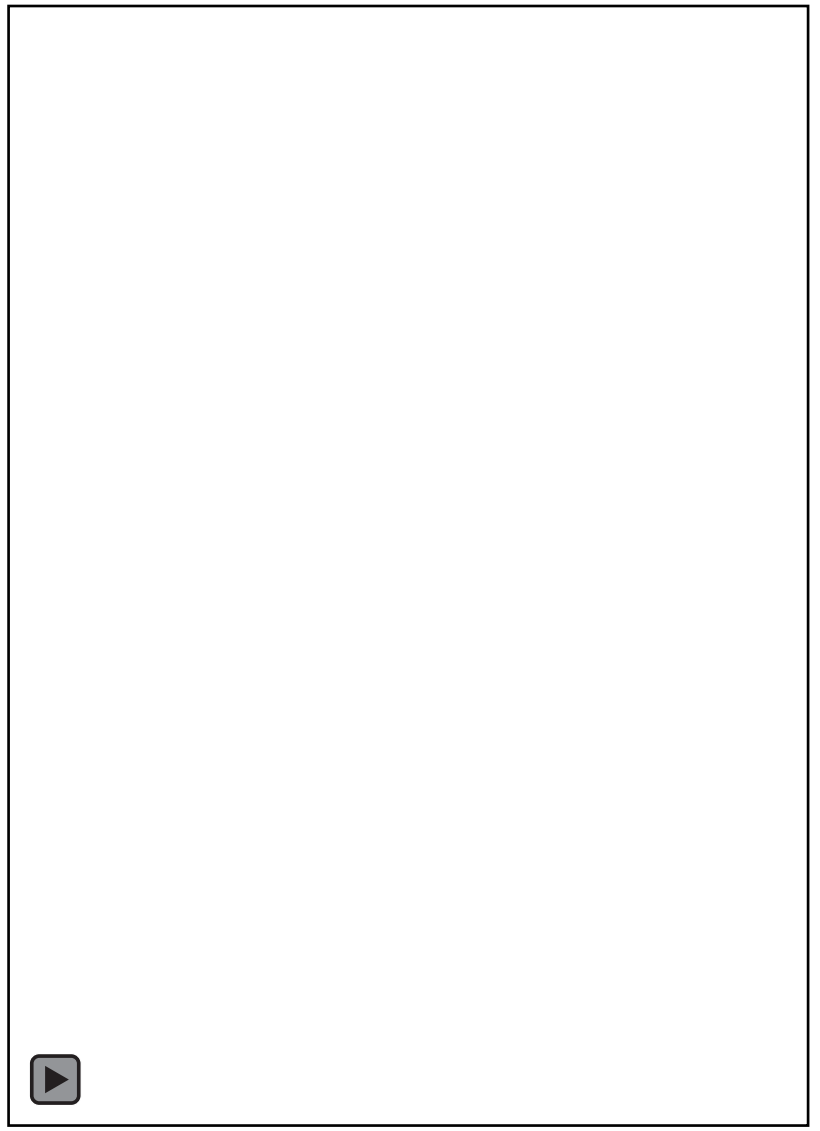
Virtual Drug Development using LS-DX AI co-scientist

User Query: Using multi-omics data and clinical data, we aim to explore and identify drug discovery target molecules for pancreatic cancer; to screen hit compounds from 20 computationally generated molecules; to optimize these compounds based on ADMET predictions; to conduct preclinical evaluation; to design a clinical trial protocol with a high probability of success; to estimate adverse event risks in real-world clinical practice; and ultimately to determine an optimal treatment protocol for real-world clinical use.

Example Execution Flow

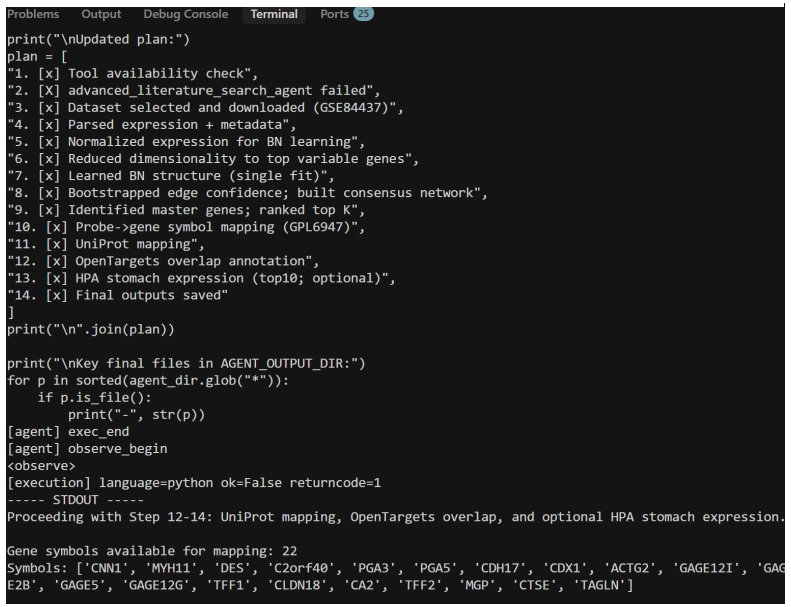


Manuscript-Style Report Generated from the Results



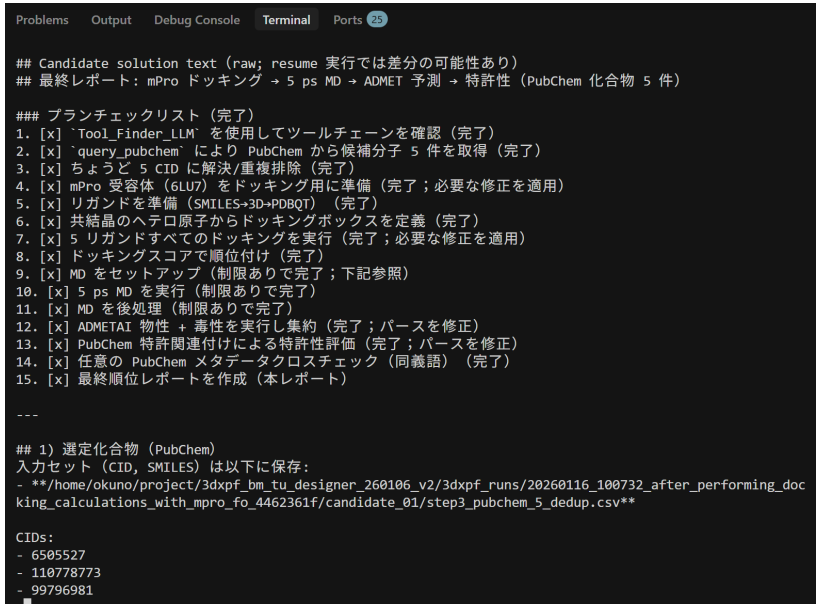
Our co-scientist solves problems across a broad range of domains

Biological Domain
Network-based Target Exploration for Gastric Cancer



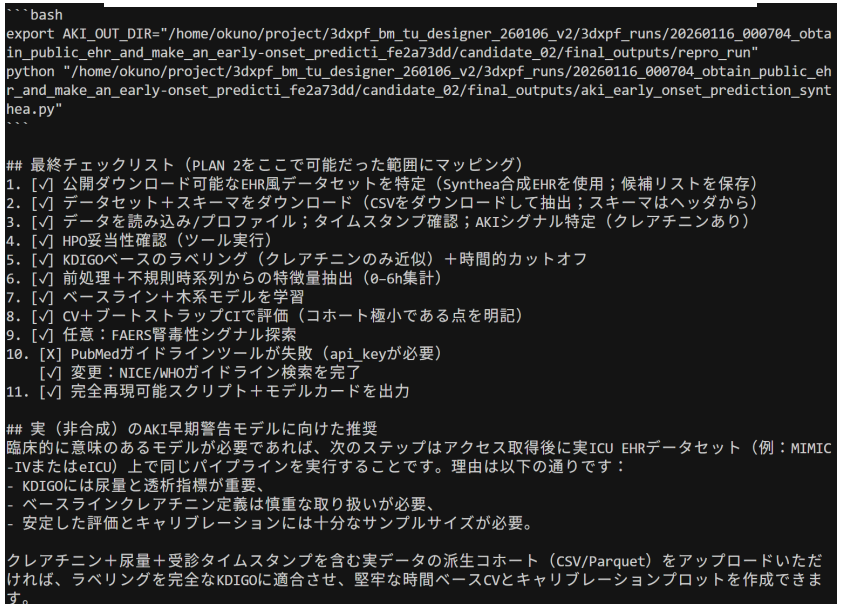
- Gastric cancer RNA data acquisition
- Bayesian network analysis
- Master gene identification
- Target mapping

Chemical & Physical Domain
Novel Compound Discovery for COVID-19



- Compound and protein structure acquisition
- Docking simulations
- Molecular dynamics (MD) simulations
- ADMET prediction
- Patentability assessment

Clinical Domain
Acute Kidney Injury Onset Prediction

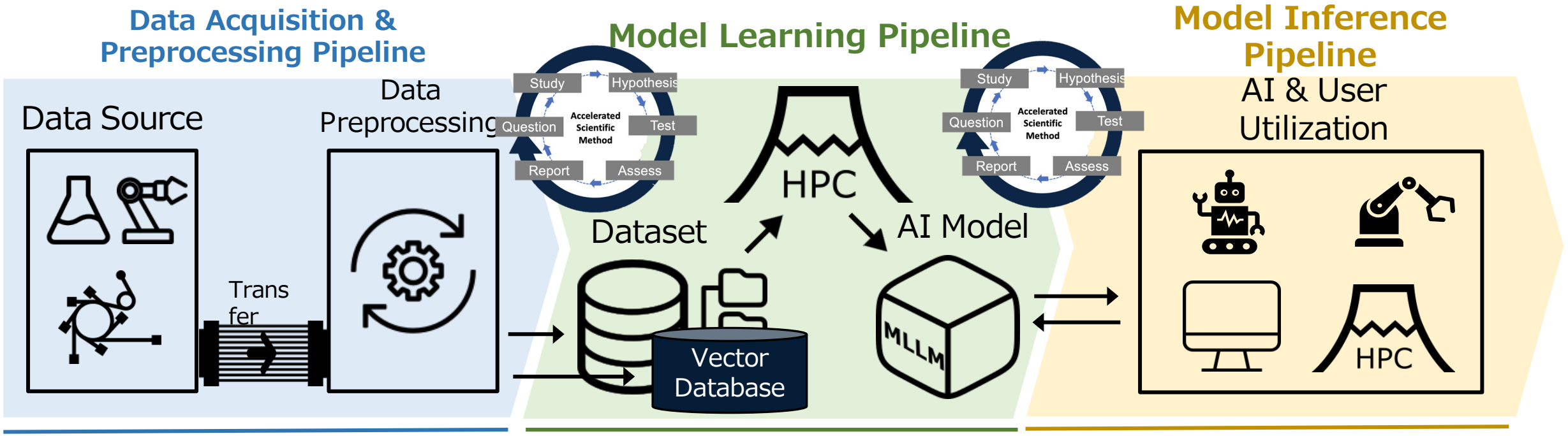


- EHR dummy data acquisition
- Outcome definition and labeling
- Feature selection
- Machine learning model development
- Cross-validation
- FAERS reference analysis

Completed within 1 hour

AI-controlled AI for Science data pipeline

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



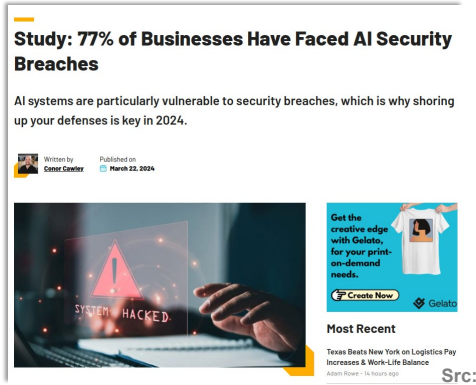
- **Data Generation & Transfer: Compression & high-speed transfer of data from data sources**
- **Data Preprocessing: Data pipeline for preprocessing (vector database construction)**
- **Optimization of model evaluation and AI infrastructure selection: Large-scale pre-training, additional training, fine-tuning, 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**

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

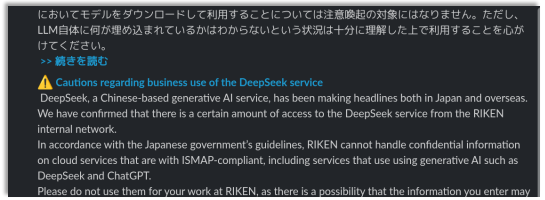
Factor 1: Humans

- 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?



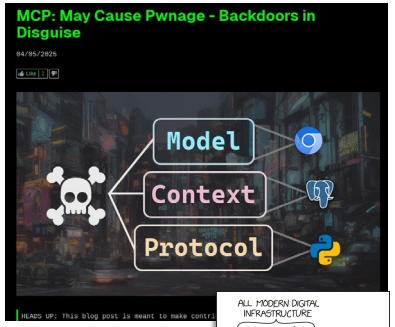
Src: tech.co



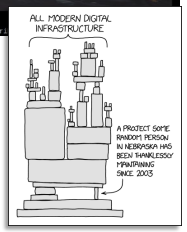
Generative AI – Move fast, ~~break things~~ get hacked!

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

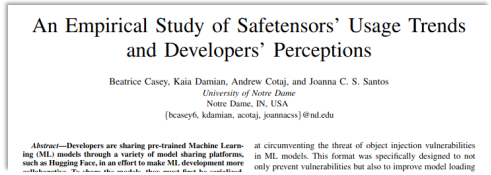
Factor 3: SW "supply chain" (OpenSSH's XZ Backdoor all over again?)

- **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

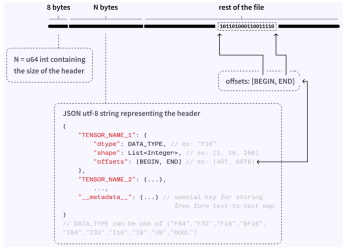

An Empirical Study of Safetensors' Usage Trends and Developers' Perceptions

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



Src: blackfog.com

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

RiVault: Secure and privatizes “OpenAI”-replica for Internal Research

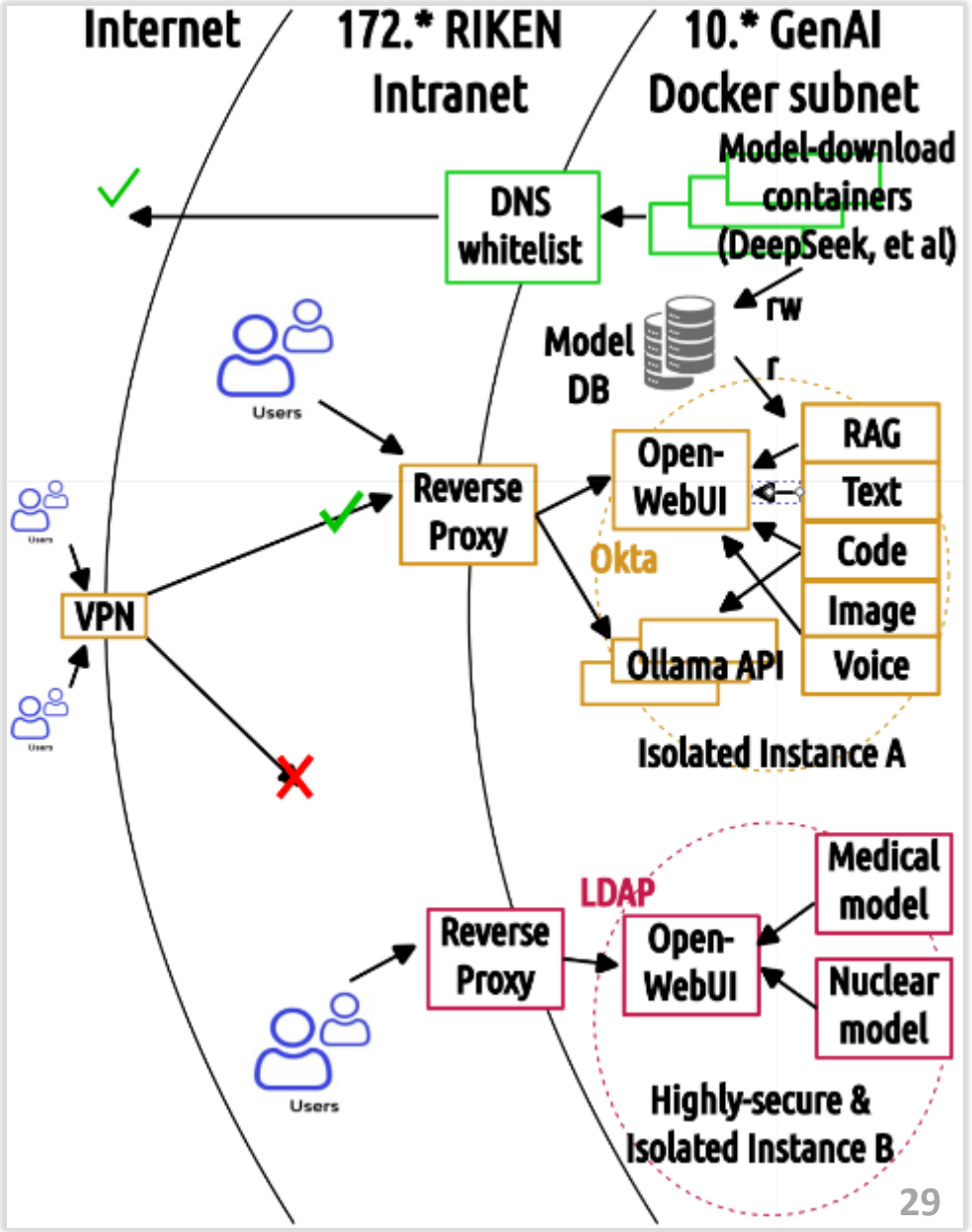
Issue with landscape of serving generative AI models

- 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

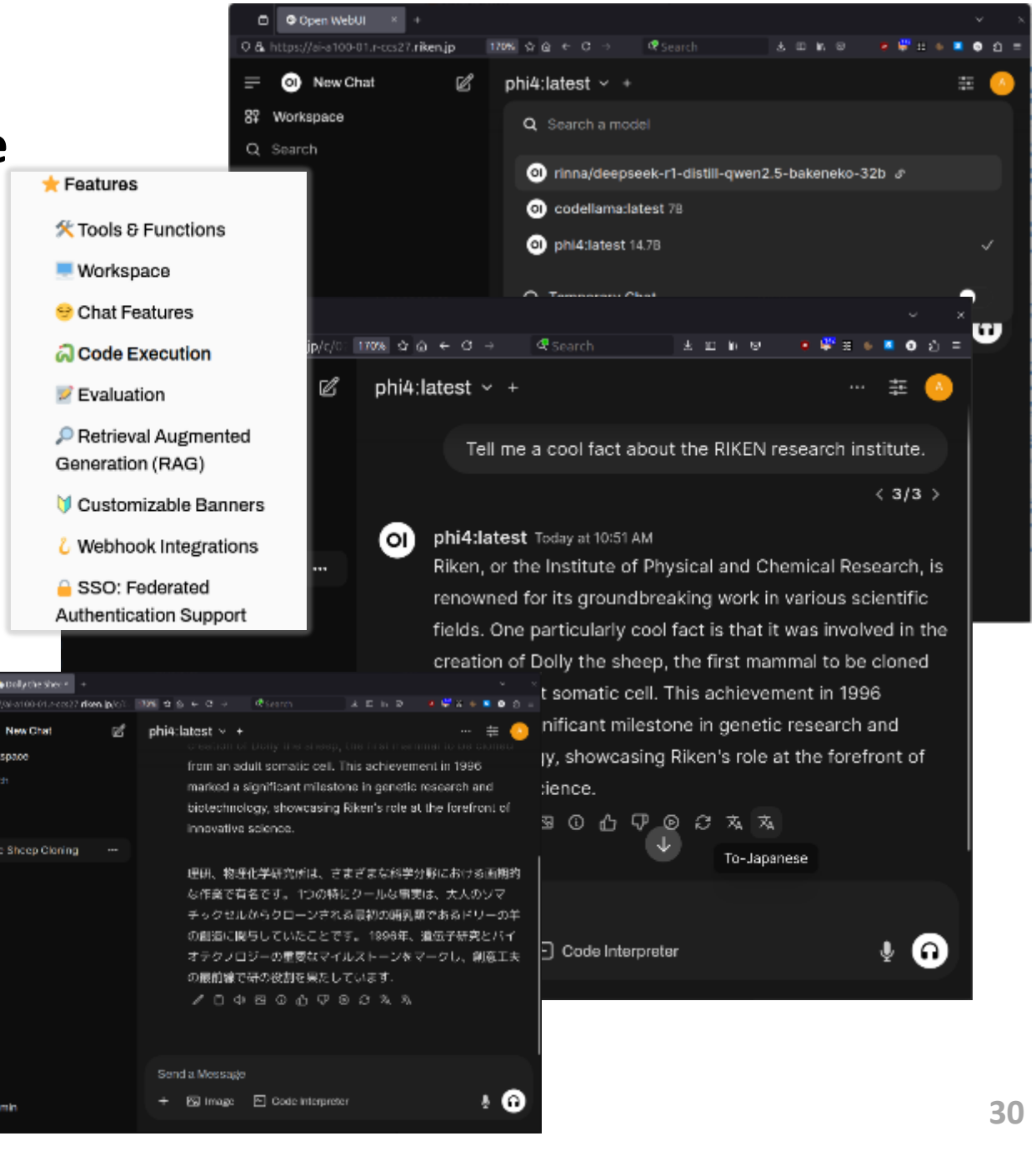
Solution is open-source with adequate security concept

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



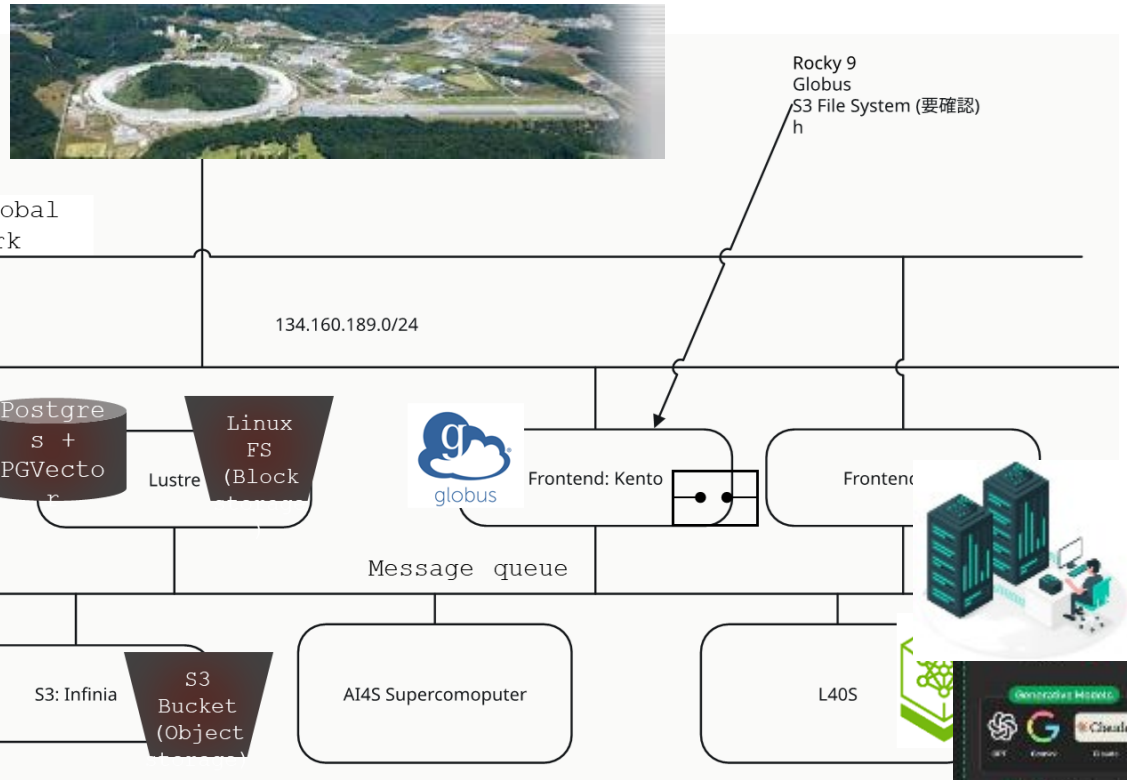
User-facing Open Web UI

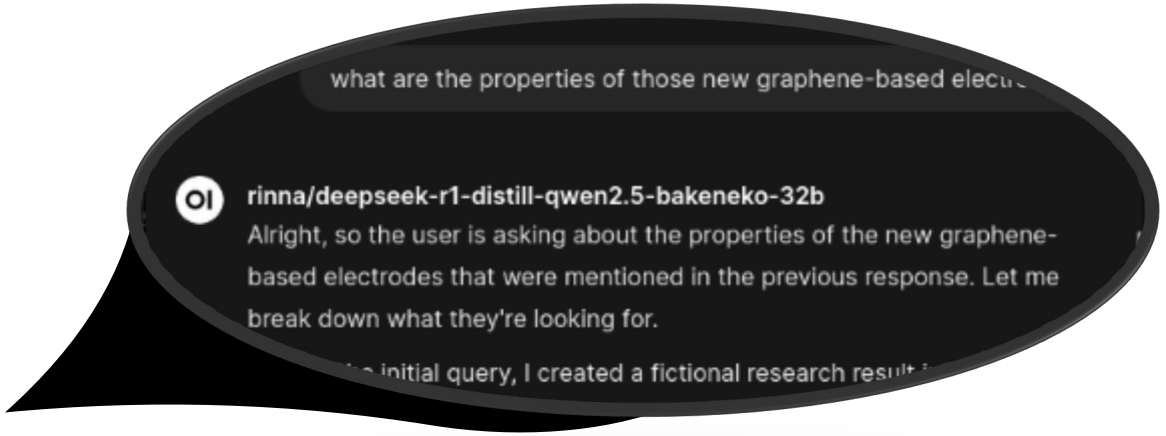
- Very similar design to **OpenAI interface**
- Supports **various features** and **multiple chats/contexts**
- Easy to select from available **models via dropdown**
- **Customizable** via scripts/"functions" (e.g. our eng<->jap translations)
- **Access rights** for models: everyone, groups, or user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

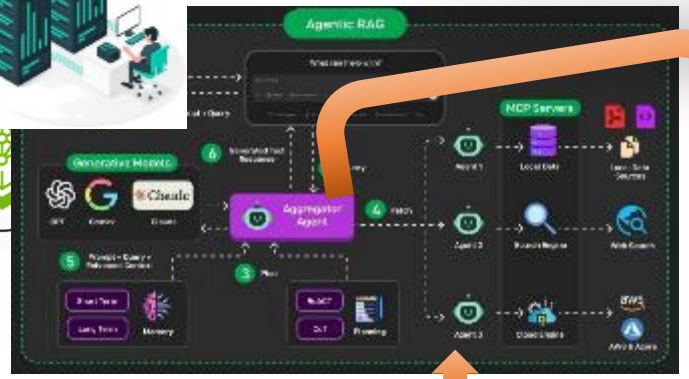


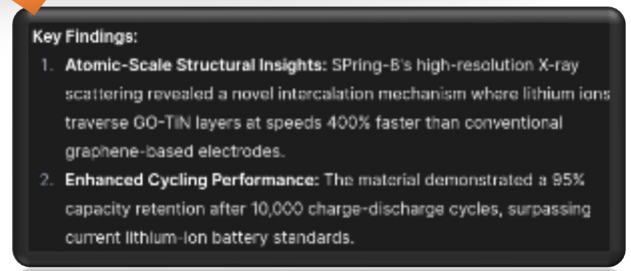
Next steps to enhance our internal/secure Infrastructure

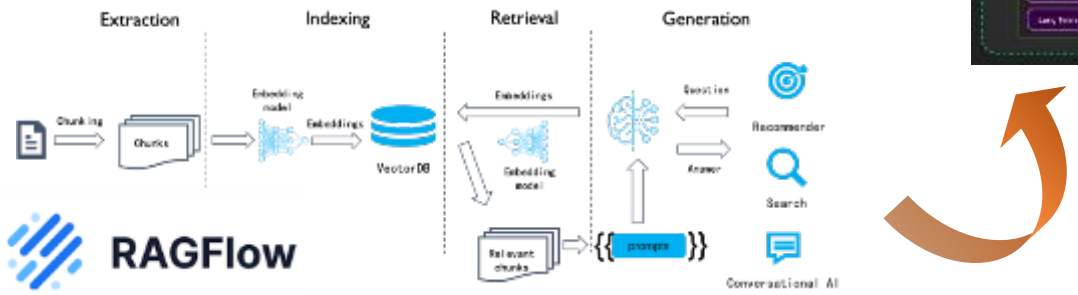
- **Demo: Unified Multimodal Retriever Service for Spring-8**













AI for Science Scientific Research Innovation Program

Supplementary budget for FY2025 37 billion yen

Direction of Issues and Initiatives

- As global investment in AI grows, maintaining domestic R&D is vital for security. In AI for Science, the U.S. and EU strategically advance national initiatives by leveraging resources such as computing power, talent, and data.
- Japan possesses world-class information infrastructure and high-quality experimental data in key fields such as life and materials. By fully leveraging these resources to develop scientific foundation models, AI agents, and next-generation AI-driven lab systems, the nation will drive the research capacity goals of the 7th Science, Technology, and Innovation Basic Plan.

Contents

Project Implementation Period Until Fiscal Year 2028

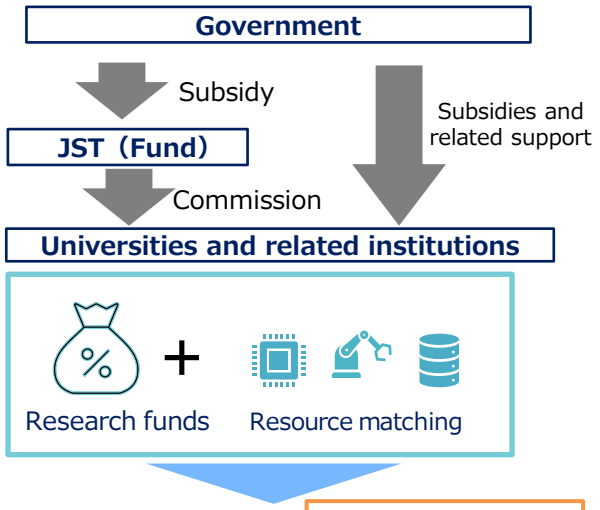
- Under national commitment, Japan will strategically allocate resources, combining project-based focused investment and challenge-driven pioneering research to establish itself as a leading nation in AI for Science.

①**Project-based approach :** In priority fields, Japan will leverage simulation and experimental data to develop AI foundation models and next-generation labs. This will strengthen research capacity, promote industry–academia collaboration, accelerate early implementation and commercialization, and drive innovation transforming scientific research

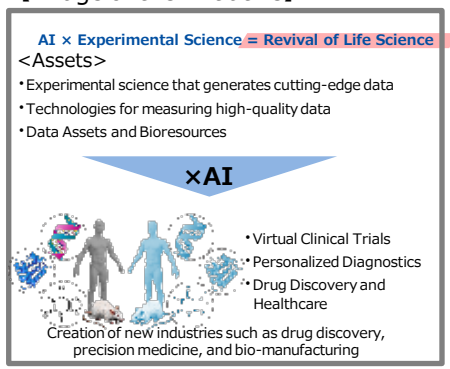
②**Challenge-based approach:** By securing computing resources and fostering AI for Science across academia, Japan supports ambitious researchers in pursuing new ideas, promotes pioneering studies, and builds unique competitive advantages.

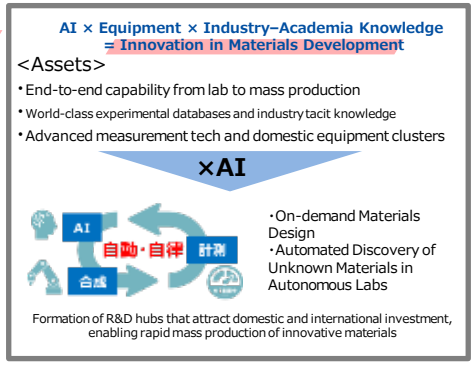
※In addition, 7.6 billion yen is separately allocated for the strategic enhancement of computing resources essential to AI for Science.

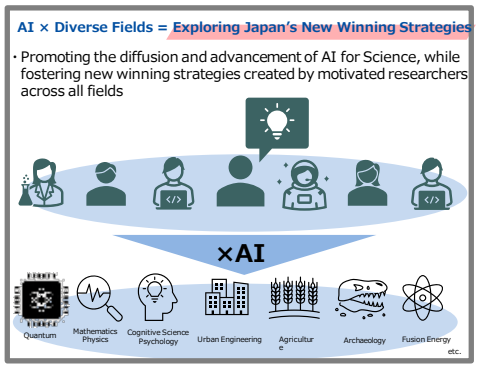
【Scheme】



【Image of the Initiative】







«P-based approach»

32 billion yen

・Number of projects supported: Approximately 5 fields × 3 teams (or individuals)
・Scale of support: Around 2 billion yen per project
・Duration of support: In principle, 3 years

« Challenge-based approach»

5 billion yen

・Number of projects supported: Approximately 1,000 cases
・ Scale of support: Around 5 million yen per case
・ Duration of support: Up to 1 year

Overview of our QC-supercomputer hybrid platform

IBM System 2
"IBM Kobe"
156 qubits
June 2025



R-CCS (Kobe)

Quantiniumm
H1-2
20 qubits => 56 qubits



Wako Campus

Feb~Mar 2026
AI4S
1600 NVIDIA
GB200s
NVL72
+
x00 additional
GPUs
QHPC

PC/Server

GPU Systems

Supercomputer
(Fugaku)

Fast
Low-latency
Network (Infiniband)

QC
Superconductive
qubit



QC
Ion-Trap qubits

Quantinuum > 20
qubits Feb, 2025



Riken RQC 'A'
QC 64 qubits

Internet
(SINET)

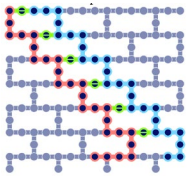
Supercomputer
Osaka U.

Supercomputer
U. Tokyo

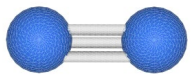


N_2 : Bond breaking on large basis set

Fugaku

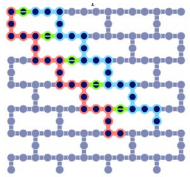


58 qubits

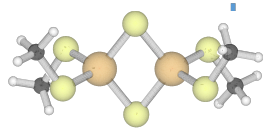


Fe_2S_2: Precision many-body physics

Fugaku

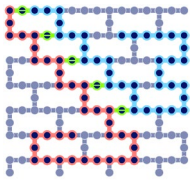


45 qubits

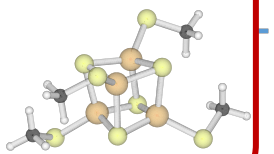


Fe_4S_4: Pushing hardware capabilities

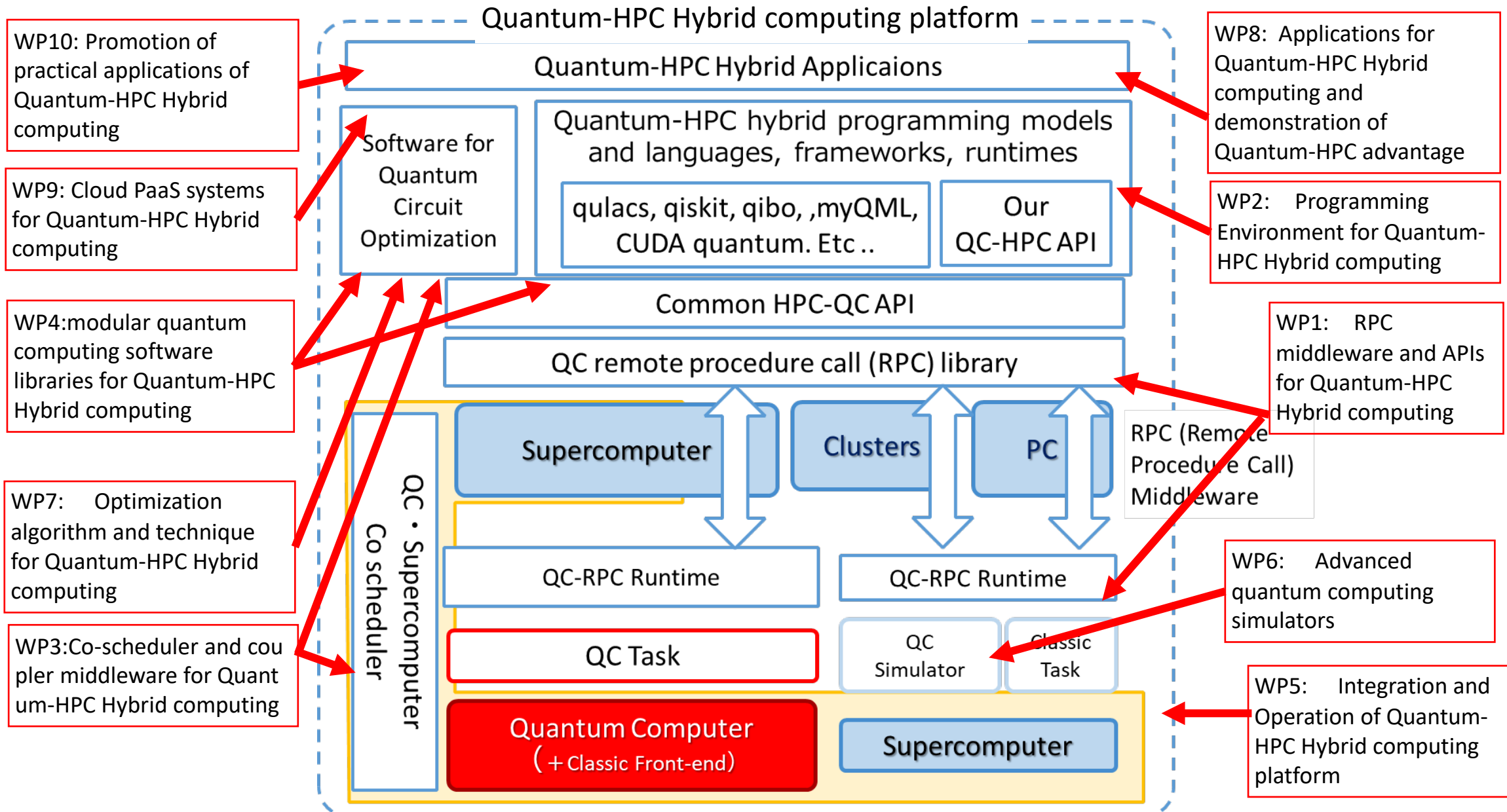
Fugaku



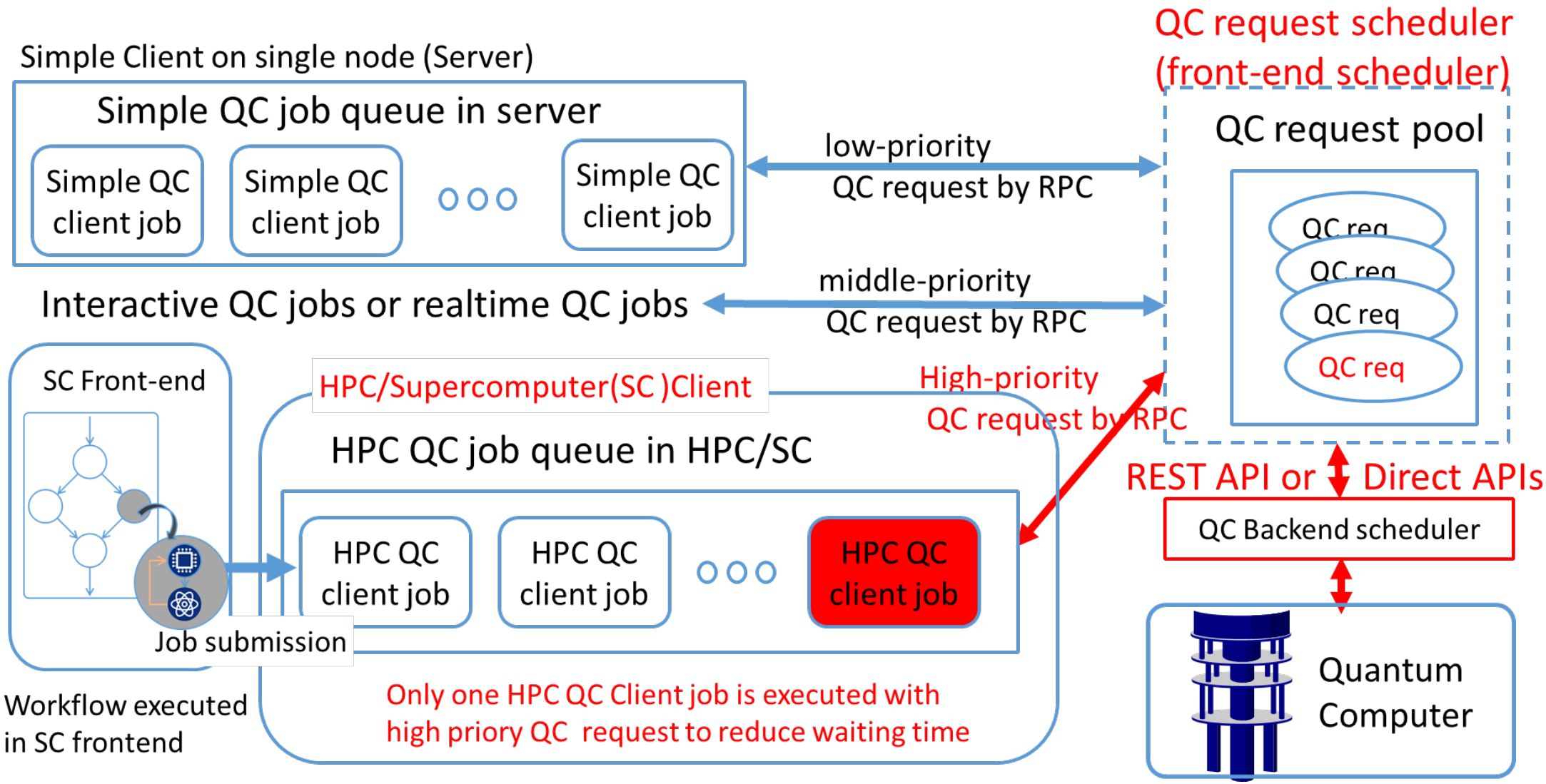
77 qubits



JHPC quantum software structure and work package



Coordinated scheduling with HPC scheduler and QC request scheduler by priority control

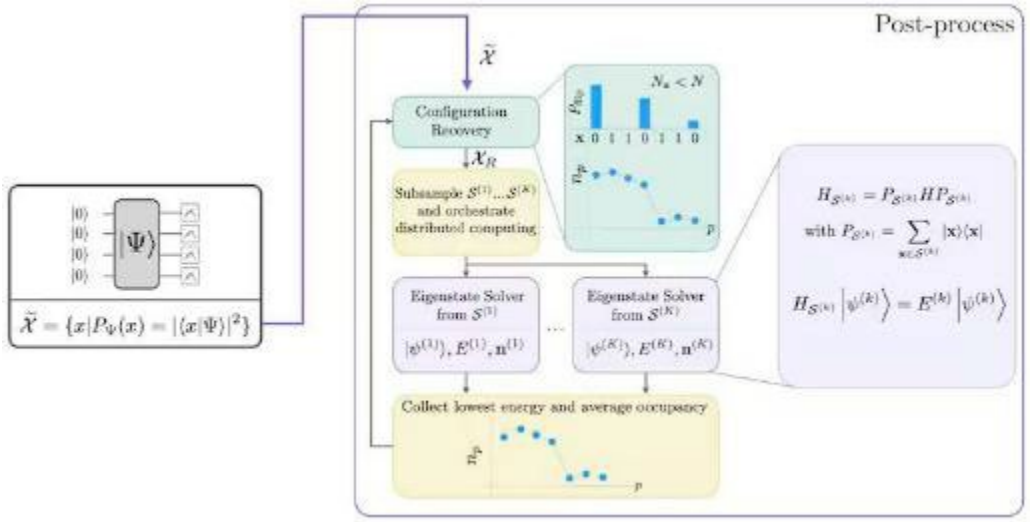


Demonstration of QCSC for quantum chemistry

- 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

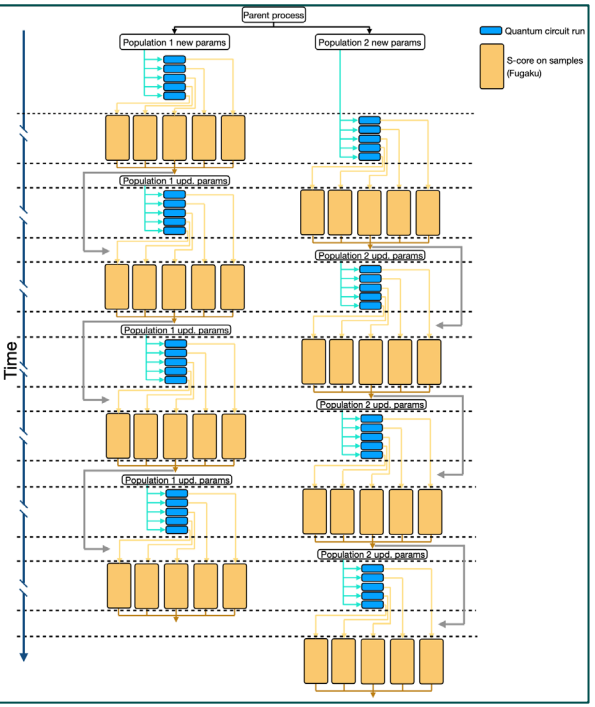
Evaluating energy for [4Fe-4S] with new diagonalization in a quantum subspace by using the latest devices and increasing the complexity of the workflow

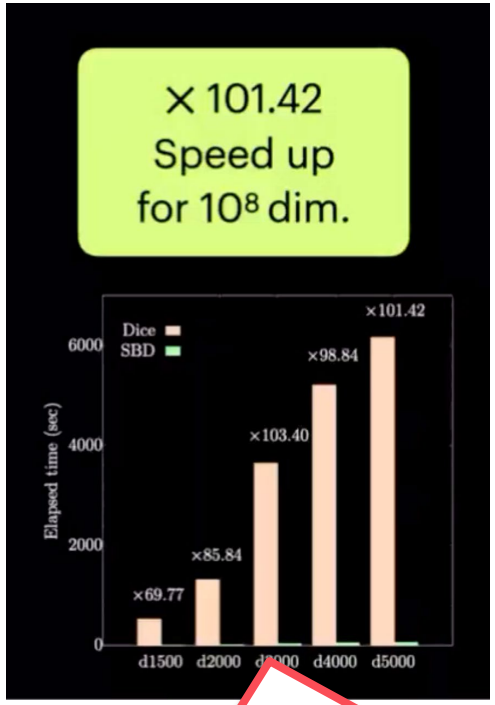
New algorithms
Optimize parameters of LUCJ circuits with iterations

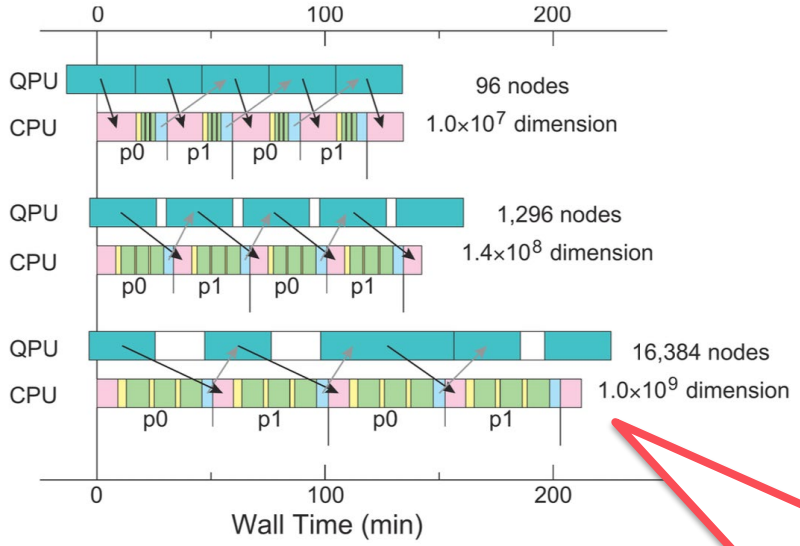
New diagonalization
Fast diagonalization tool.
100x speedup

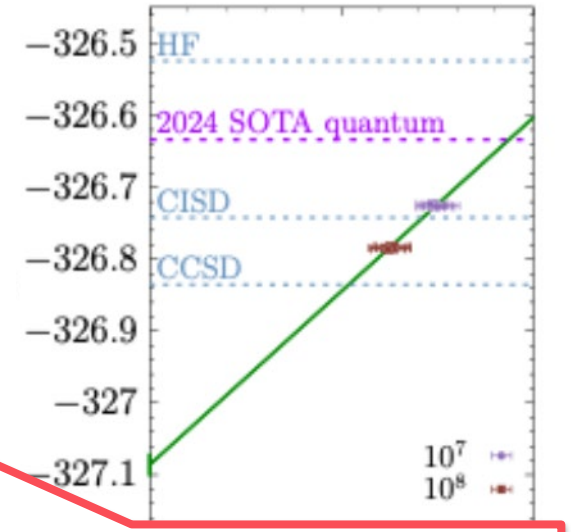
New QCSC workflow
Implementing the algorithms with feedback loop, tight coupling, orchestration of resources

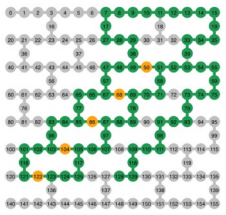
150 mEh improvement over 2024 demo
Efficiently utilize capabilities of Heron r2 (ibm_marrakesh) and 16K Fugaku computation nodes











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**

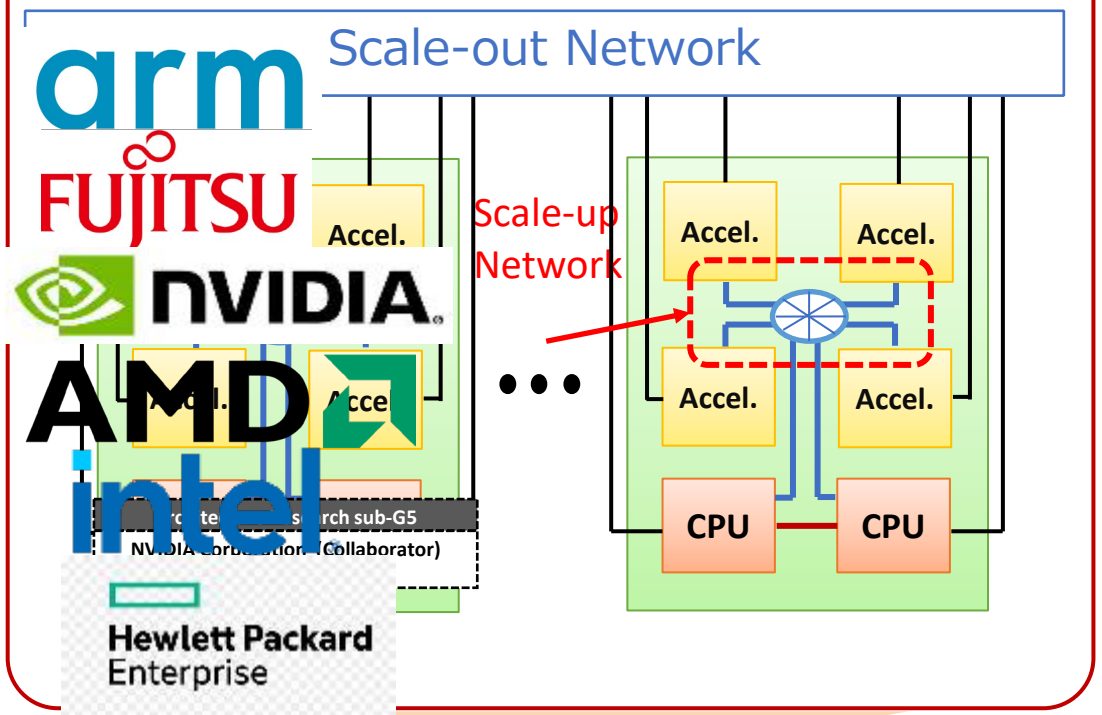
RIKEN

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

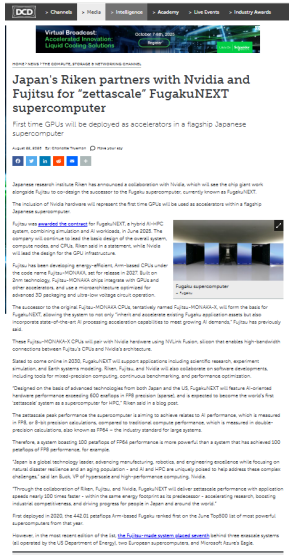
Media	
Web&Newspaper	39
TV Coverage	7
Total	**46**

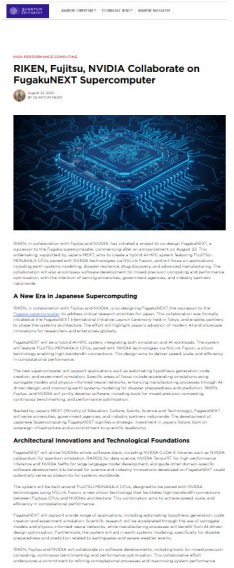
News overages during 2025/8/22-28








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

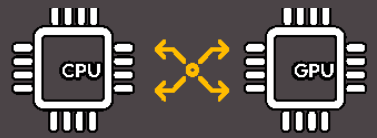
- World's first implementation of low-precision matrix **Arm SME** in a server CPU, enabling low-latency AI processing
- Top AI-CPU performance for standalone & w/GPU
- Expansion of AI acceleration frameworks and libraries

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

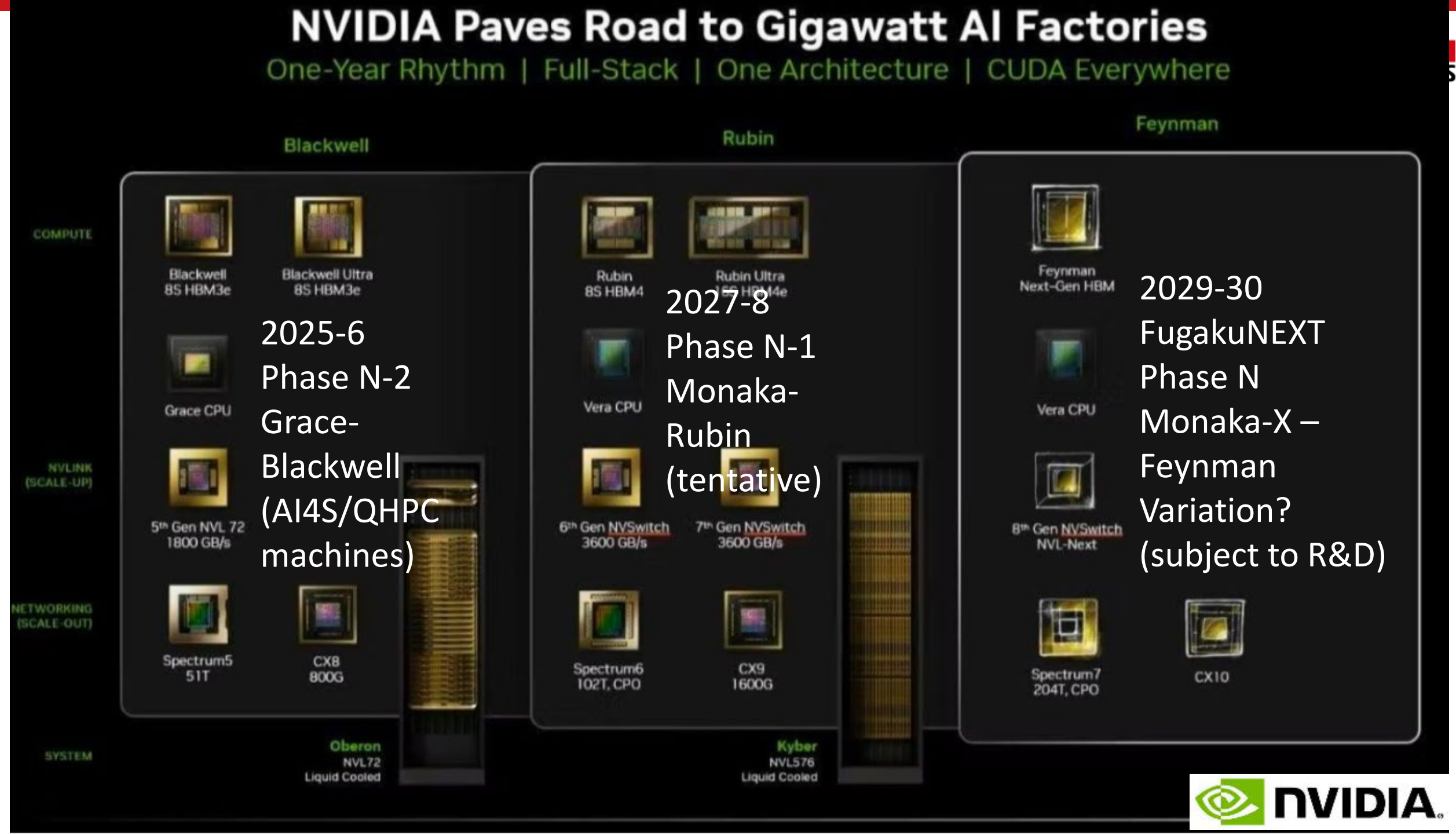
- Energy and space efficiency
- Optimization for AI training and inference infrastructure
- Advanced security, etc.

Edg

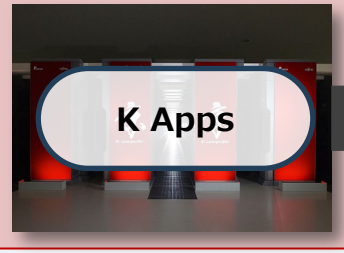
Real-T

- National se
- Telecommu
- Robotics, e

© 2025 Fujitsu Limited


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

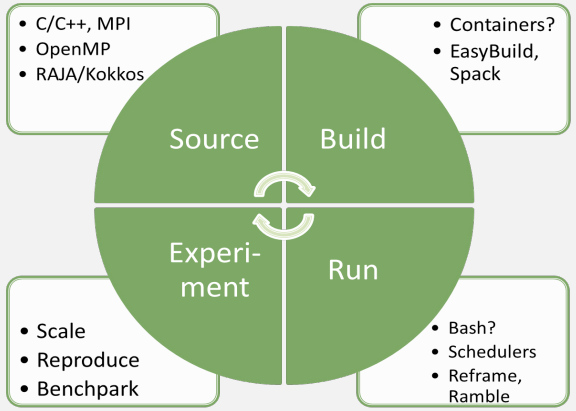
FugakuNEXT 100x Application Performance Leap

	K Computer (2011~)		Fugaku (2020~)		FugakuNEXT (2030~)
Application Performance	 K Apps	~100x	 Fugaku Apps	App Perf **Max100x** Speedup	 FugakuNEXT Apps
Hardware Perf (FP64/Memory)	10.6PFLOPS 5.64PB/s	Vendor HW ~40x	**442.0PFLOPS 163 PB/s**	Vendor HW ~6x	**2.6EFLOPS 以上 800PB/s 以上**
Software Perf (Algorithms, Libraries, Tools)	1	Riken SW ~3x	**3**	Riken SW ~20x	**Max 60x** Max ~1000x
			• Surrogate models / PINNs（10~20x） • Mixed Precision (2~8x) • AI Hardware (E.g. Ozaki Scheme 10x)		
AI Hardware PErf (FP16/FP8)	NA	NEW!	**2.0 EFLOPS**	~300x Vendor HW	**> 600 EFLOPS** (Zettaflops in FP4)

Riken Leadership in FugakuNEXT Applications

① Code Porting & Evaluation

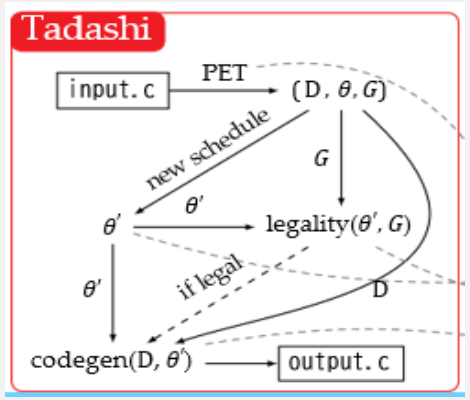
Code Porting Support on Phase 2 machines w/CI/CD/CB



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

② AI-based Code Modernization

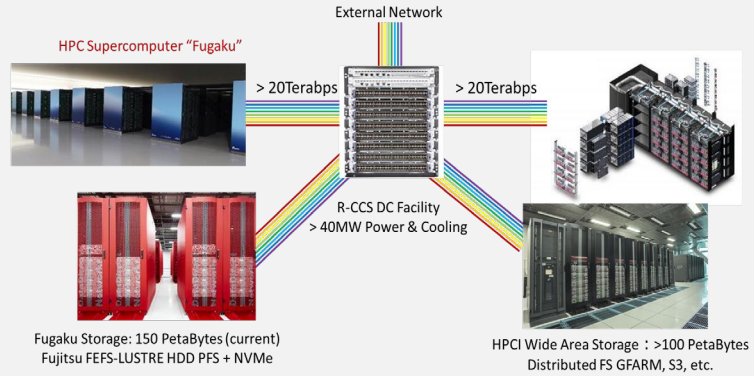
Coding AI to generate & Port Legacy HPC Codes



- 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

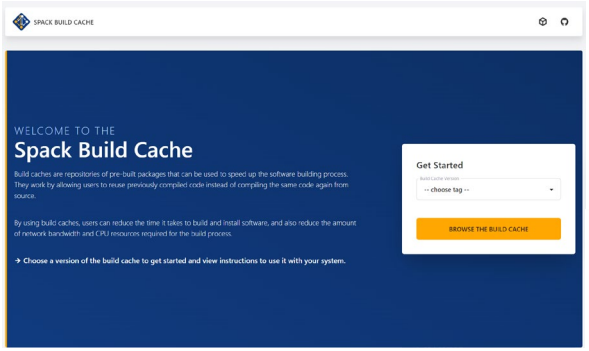


- Phased deployment of N-2, N-1, … platforms
- Utilize production platforms (AI4S, Quantum-HPC) to test system software and apps

FugakuNEXT Partnership Program for Early Access to Development

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

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 2000+GPU in APU configurationVirtual 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



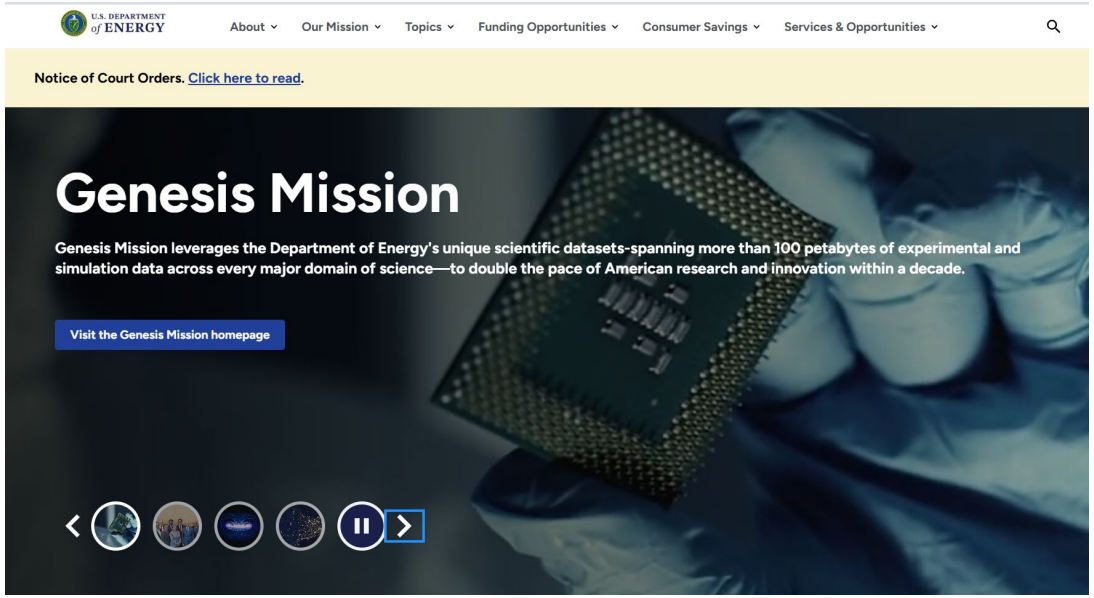
New Datacenter

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

NEW: DOE Genesis – JP AI4S/AI4I Collaboration

US DOE Annouced the new Genesis Mission 2025/11

- **Securing U.S. Leadership in AI for Science** through not only with DOE National Labs and academia, but also with **U.S. industry leaders in Big Tech and semiconductors and select international partners**
- **Dario Gil, DOE Under Secretary for Science**, visited Japan for SCAsia in January; **4-way MOU** signed between **RIKEN, Argonne National Laboratory (ANL), NVIDIA, and Fujitsu**.
- **Govt. talks underway to make Japan as a primary candidate for international cooperation**, featuring key initiatives such as: **Fugaku NEXT,** RIKEN & MEXT **AI4S (AI for Science)**, METI/NEDO/Riken **Quantum-HPC Hybrid**, **Rapidus Fab**, and more.

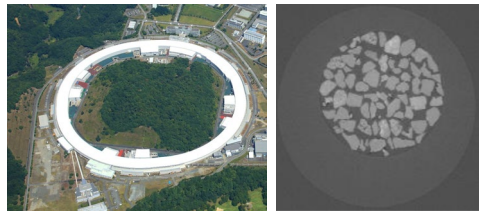




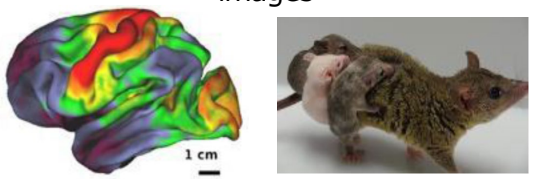
左から理研R-CCS 松岡 センター長、文部科学省 柿田 文部科学審議官、富士通 マハジャン 執行役員副社長 CTO、

Data Acquisition & Preprocessing Pipeline

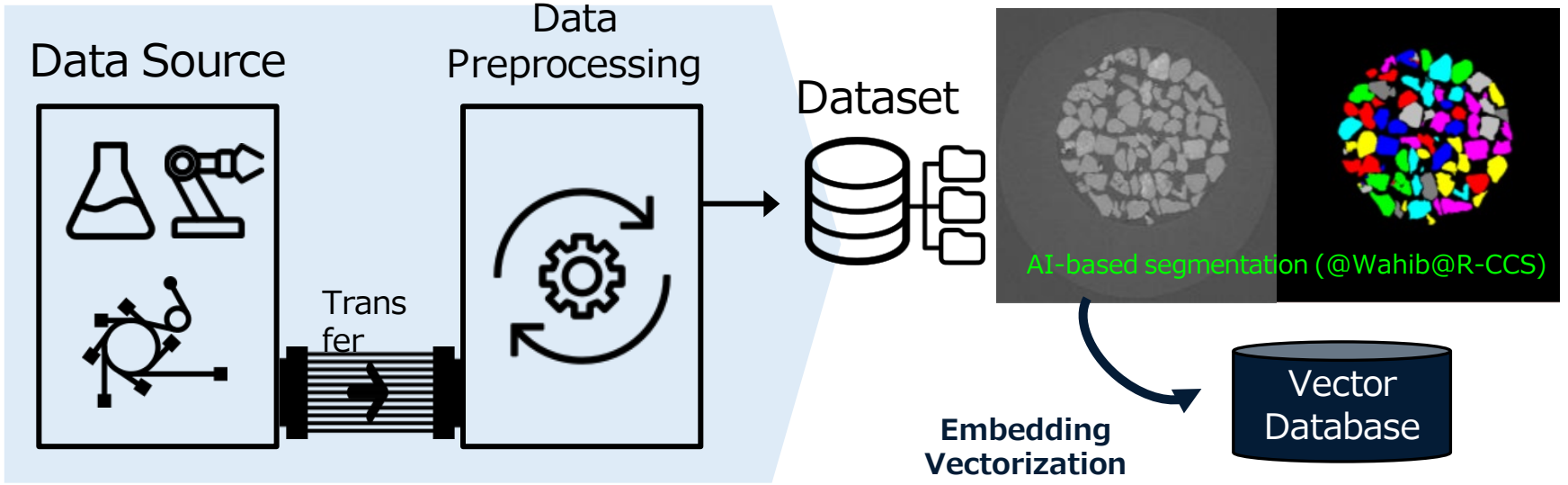
- **Efficient acquisition and management of high-quality and large amounts of data using AI for high-performance model learning**



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