FugakuNEXT Development Project – Paving the Wat to AI-for-Science Era

Masaaki Kondo

Division Director, RIKEN Center for Computational Science
Professor, Keio University

Supercomputer “Fugaku”

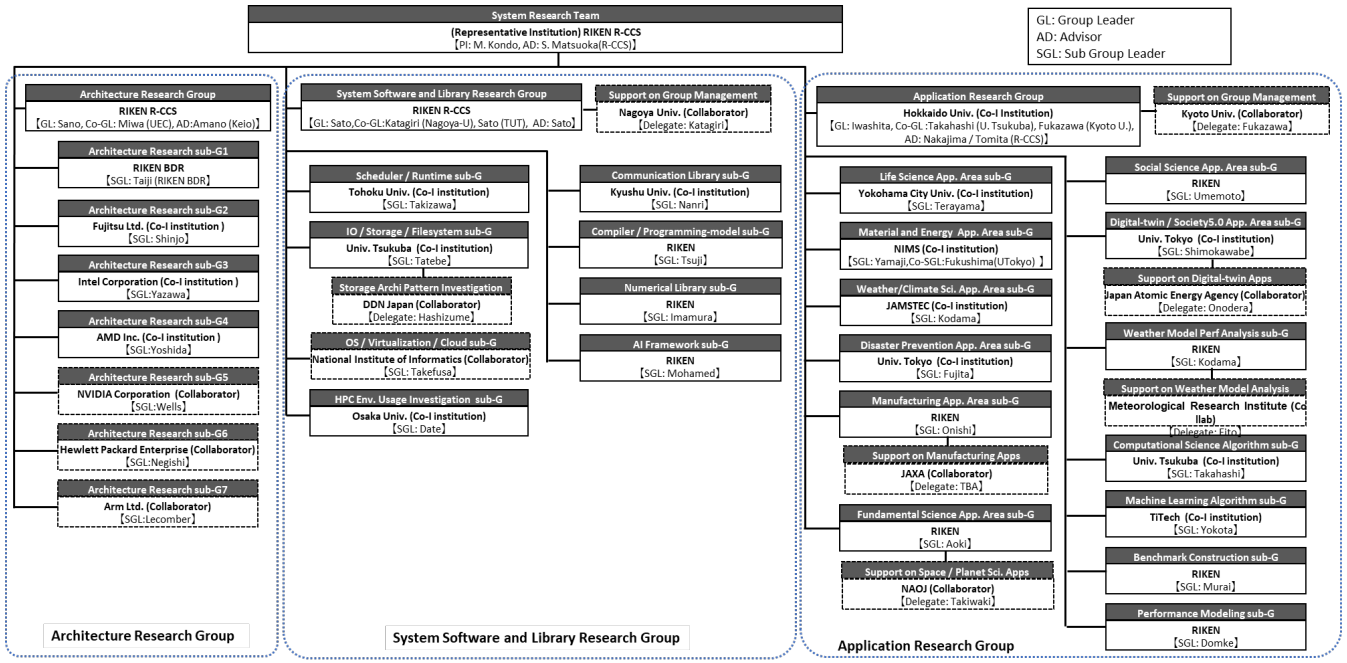
- A flagship supercomputer system in Japan
  - Installed and operated at Riken Center for Computational Science in Kobe
  - Peak performance : 537 PFLOPS (double precision)

富岳

From Feasibility Study to FugakuNEXT Development

Overview of Feasibility Study (by MEXT)

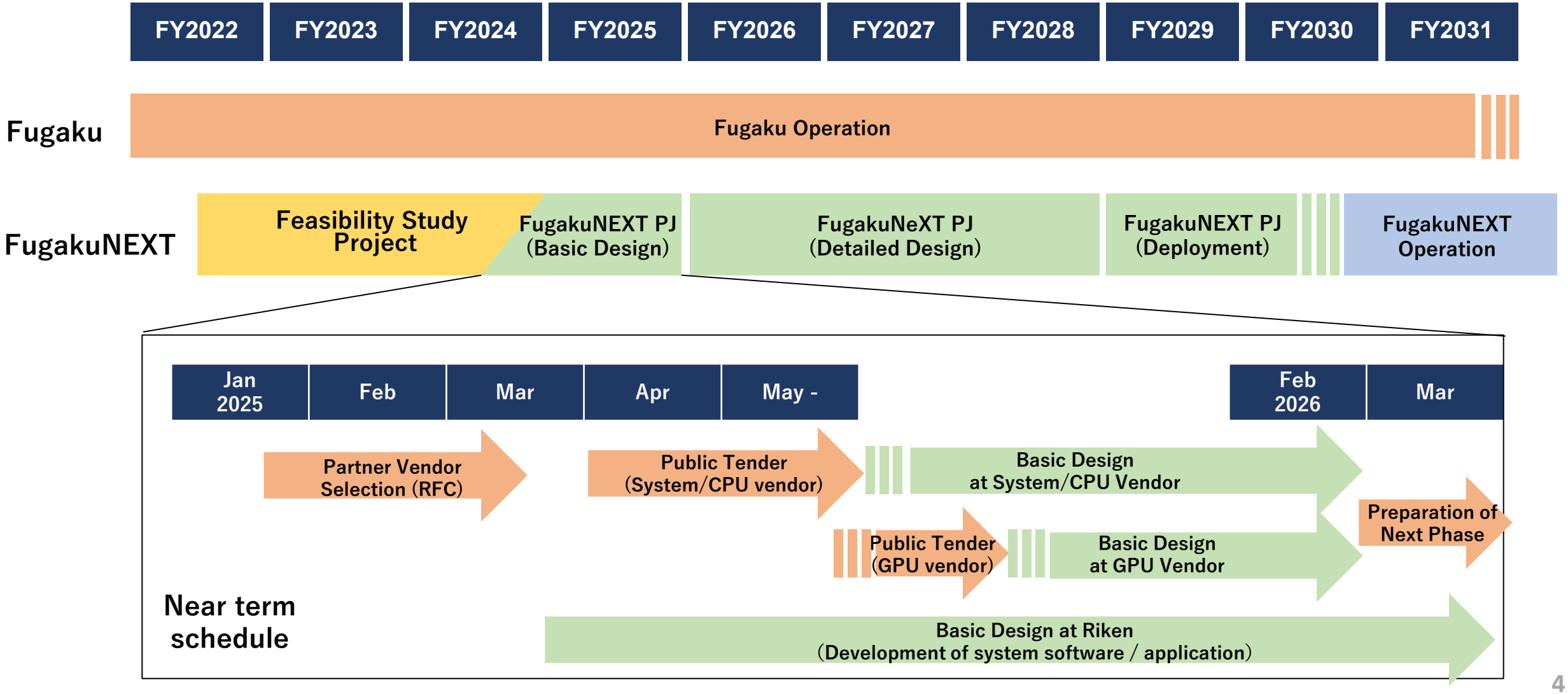
Organization Structure of RIKEN Team



- FugakuNEXT (successor to Fugaku) project started January 2025
  - Realizing a world-class AI-HPC infra. for advancing science through the integration of HPC and AI
  - Accelerating scientific discovery by AI-for-Scicnece
- Establishing a project organization built upon the existing collaboration relationship

Update of Timeline of Fugaku-NEXT R&D

- **FugakuNEXT development and deployment schedule**



Extension of R-CCS Facility towards FugakuNEXT

- **FugakuNEXT will be deployed on the site adjacent to R-CCS**
  - A new building for datacenter facility is planned for construction
  - Fugaku and FugakuNEXT can be operated in parallel for a certain period of time
  - Possible to collaborate with Fugaku by leveraging the assets of the supercomputer

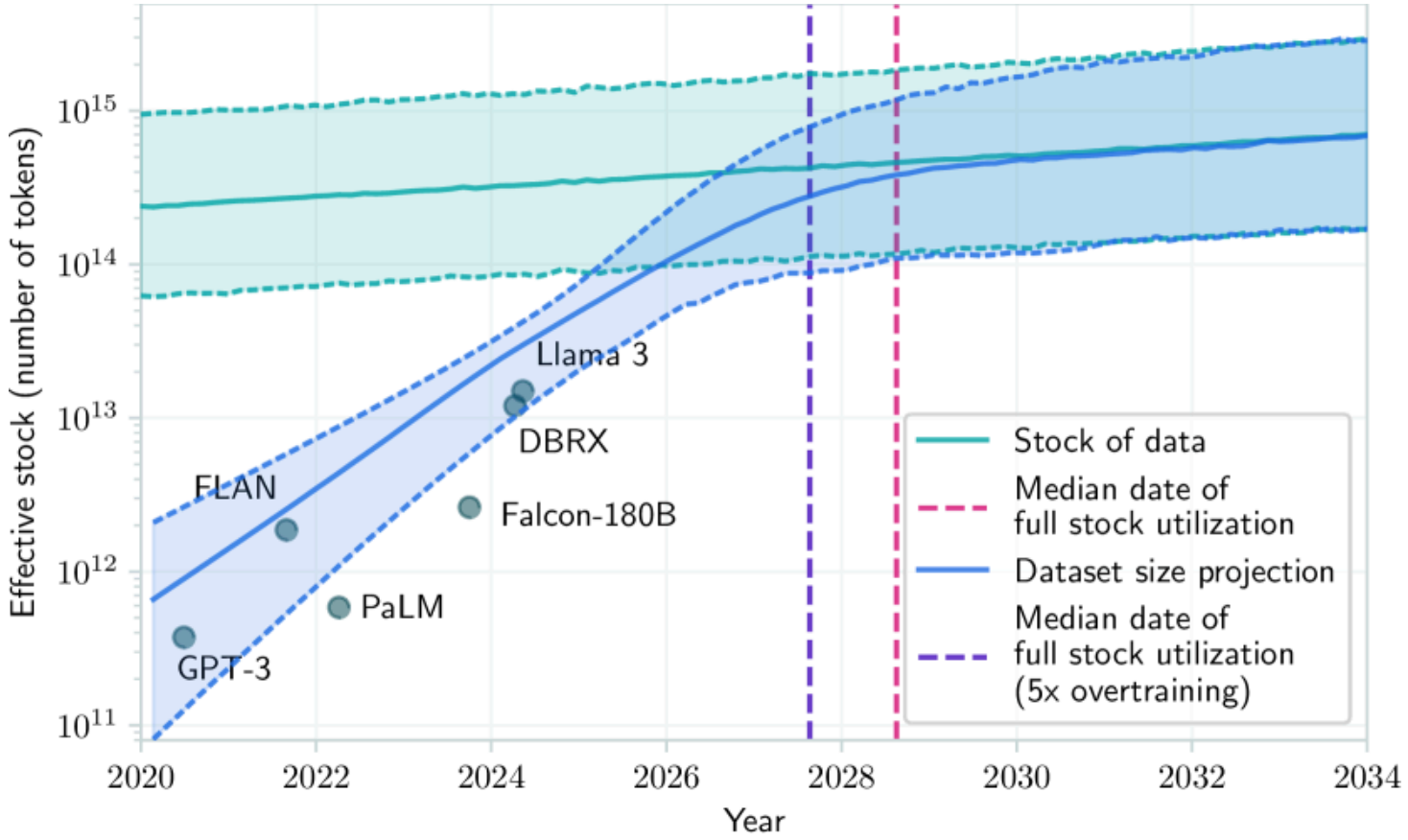
- **Expected to be a globally distinctive hub for computational science**
  - Quantum-HPC Platform with a state-of-the-art quantum computer (IBM-Q)
  - Introduction of a new AI-for-Science machine
  - FugakuNEXT as an HPC-AI integration platform

AI for Science - Important for Societal Innovation

- Goldman Sachs (Data as of December 31, 2023):
  - The percentage of macro productivity upside relative to no technology breakthrough baseline
    - 30.2% for steam engine (1769)
    - 30.6% for electricity (1880)
    - 12.6% for PCs/Internet (1981)
    - 17.5% for AI (2023)
- Recent Gartner talk: AI will increase GDP by 8~9%
  - Moreover, such productivity increase could be a one-time effect
- **GDP increase from 1960s to 2023: > 60x**
- Thus the effect of Science and Engineering to induce new technologies rather than being productivity gains should have profound effect
  - But right now AI for Science usage is still very limited, overshadowed by consumer-facing AI investments

Courtesy: Satoshi Matsuoka (Keynote talk at R-CCS Symposium 2025)

Consumer Facing LLMs may run out of data in 2028..



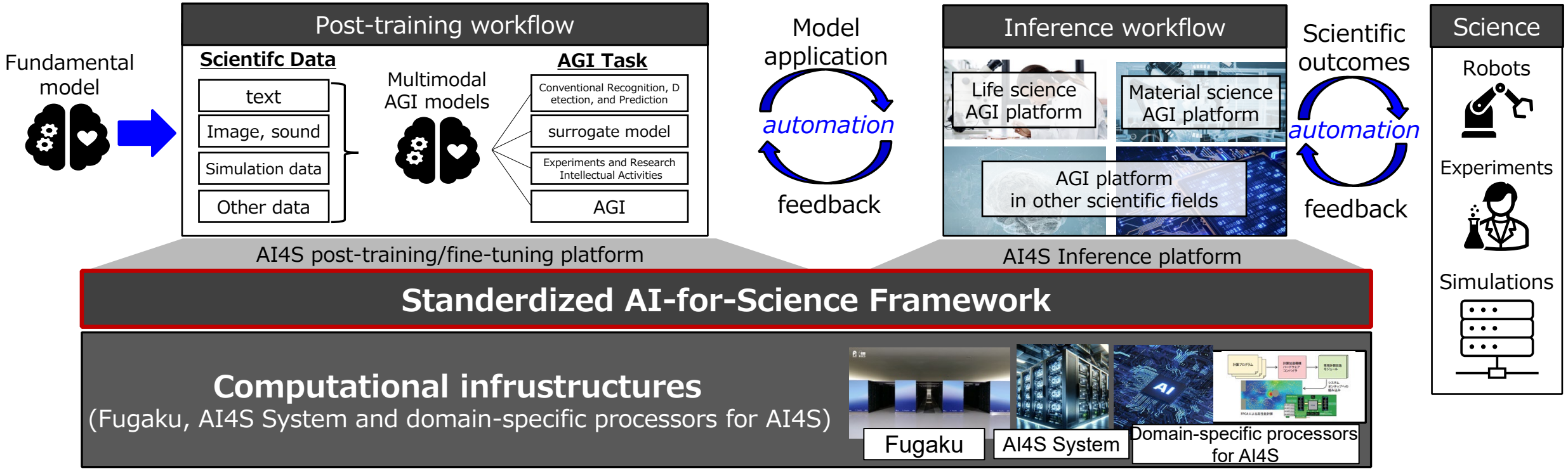
Source: Villalobos et al., "Will we run out of data? Limits of LLM scaling based on human-generated data", ICML'24

- **Scaling generative AIs with high-quality data created by digital twins (simulation) and AI-driven experiments**

RIKEN's Current Effort for AI-for-Science: TRIP-AGIS

TRIP-AGIS ③-2 Software technologies for Innovative Computational Infrastructure

- **Developing workflow infrastructures to facilitate post-training, inference and its applications**



Expansion of AI App Areas in Various Scientific Fields

1. Nanoscience devices

- AI Applications in Materials Research: Machine Learning Potential Molecular Dynamics
- Construction of material analysis flow by integrating data science and spectroscopic experiments
- Machine Learning Model Building Using Quantum Computers and its Application to Computing of Physical Properties
- AI Application in New Materials Development
- Data-driven approach to the analysis of strongly correlated quantum matter
- Numerical solution of quantum many-body problems and its applications
- Integrated analysis of experimental data
- AI Application to Amorphous Material Dynamics - From GNN to Generative Modeling

2. Energy and Resources

- Materials Design and Exploration by Simulation and Informatics
- High-precision molecular dynamics simulation of molecular systems using machine learning potentials
- Description of quantum many-body system by artificial neural network
- Quantum Chemistry Accelerated by High Performance Computing and Artificial Intelligence

3. Elementary Particles and Nuclei

- Structure and reaction calculations for nucleon many-body systems
- Analysis of quantum many-body problems using artificial neural networks

4. Life Science

- 3D structure analysis of biomolecules based on machine learning
- Searching for reaction coordinates of biomolecules using machine learning
- Conducting medical and biological research through reinforcement learning that incorporates "world models
- Fragment Molecular Orbital Calculations and AI/Data Science
- Optimization of Molecular Dynamics Force Field Using Difference Simulation
- Coarse-grained molecular dynamics (CGMD) force field development using AI
- Development and Prospects of Machine Learning Potential
- Dimensionality reduction for describing biopolymer dynamics
- Expression learning of protein dynamics by extending VAE

5. Drug discovery and Medical care

- Language Models and Multimodal Infrastructure Models in Medicine
- Current Status and Issues of Protein Language Models
- Large-scale language models for genome sequencing
- Base model for gene expression data
- Molecular Design by Generative Modeling
- Prediction of compound-protein interactions
- Protein Structure Prediction
- AI Accountability and Intervention Simulation in Healthcare

6. Design and Manufacturing

- Flow feature extraction using CNN-AE and its application
- Application of 3D Generation AI to Optimal Structural Design

7. Social Sciences (to be written after 2024)

8. Brain science and Artificial intelligence

- Neuroscience and AI Techniques and Large-scale Detailed Neural Circuit Simulation

9. Earthquakes and Tsunamis

- Examples of PINN in inverse problems in seismology and its applicability to large-scale problems
- Accelerating Large-Scale Simulations with Data Science Methods

10. Weather and Climate

- **Surrogate modeling:** application of AI to cloud microphysical processes, gravitational wave parameterization, RC learning for Navier-Stokes turbulence
- **Weather applications:** Global Numerical Climate Model (GCM) emulation, AI data assimilation fusion/precipitation nowcasting, reservoir computation and weather forecasting applications
- **Platform for dataset and model sharing, intercomparison, and analysis**

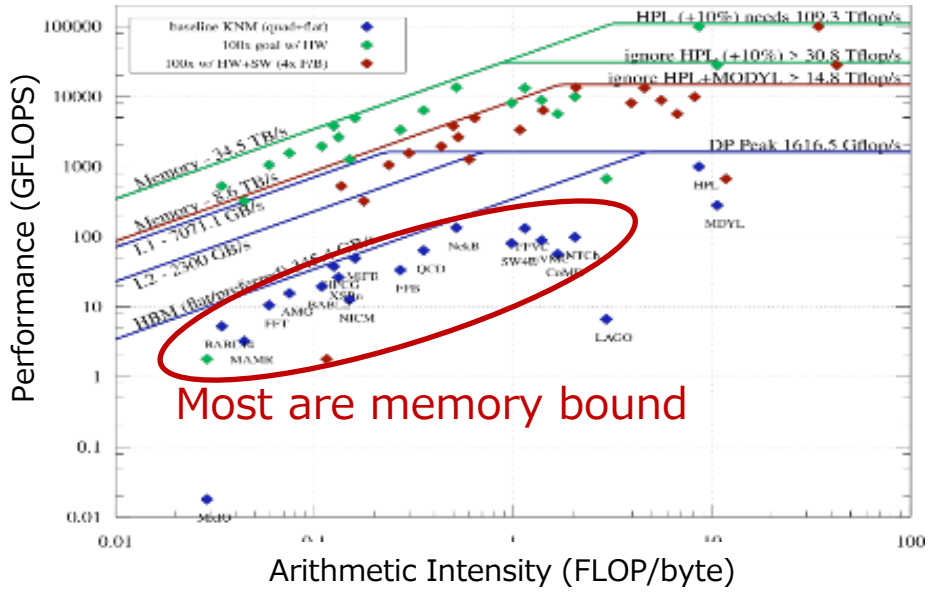
11. Space and Astronomy

- Deep Learning to Study High Energy Astronomical Phenomena
- Extracting Cosmological Information from Astronomical Big Data

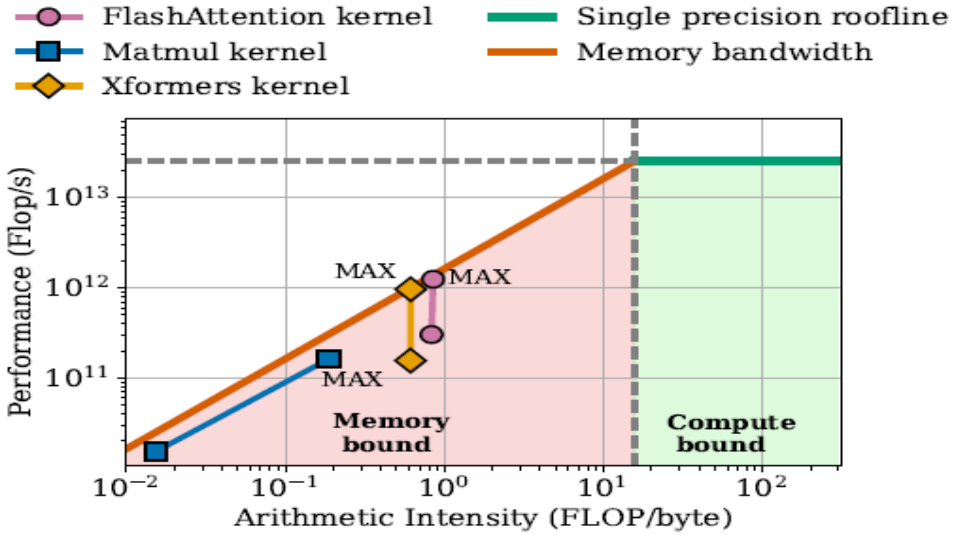
Architectural Direction of FugakuNEXT

- **Realizing a world-class AI-HPC infrastructure by integration of simulation and AI**
  - Accelerating scientific discovery by automation and advancement of science, including hypothesis generation and validation
  - Significant HPC perf. improvement (beyond HW limit) by mixed-precision computing and surrogates
- **Optimized data movement for performance & power efficiency**
  - **Using advanced memory tech. available at deployment**
- **Heterogeneous & tightly coupled architecture**
  - **CPU+GPU architecture with "Made with Japan" concept**

Roofline analysis of HPC apps

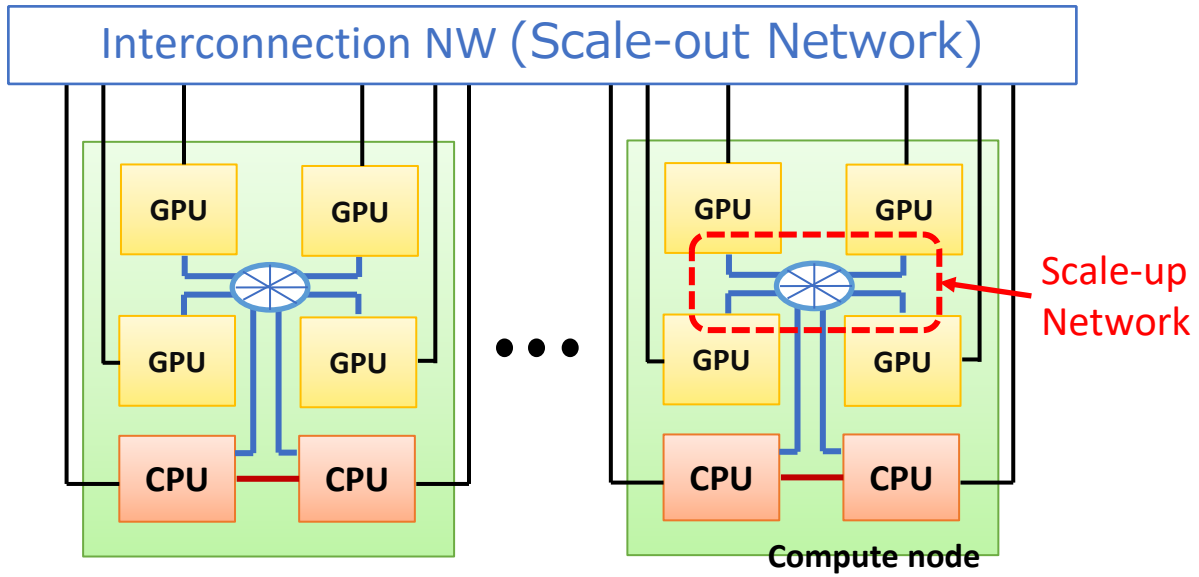


Roofline analysis of LLMs



Source: P.G. et al. Mind the Memory Gap: Unveiling GPU Bottlenecks in Large-Batch LLM Inference, arXiv:2503.08311, Mar. 2025.

System Architecture Overview (Under Consideration)

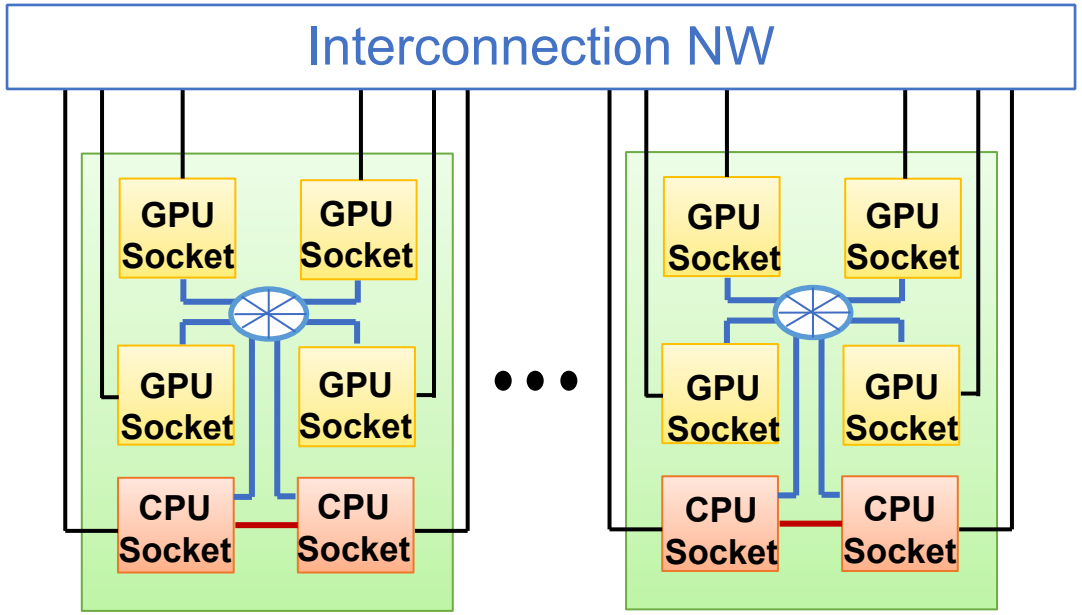


- **Compute node with CPU and accelerators**
  - CPU compatible with "Fugaku" at the binary level
  - GPUs as accelerators
- **Network both for strong and weak scaling**
  - Combination of Scale-up and Scale-out networks
- **Tens of thousands of accelerator sockets throughout the system**

- **Compute note with CPU and Acc (GPU)**
  - **Compute node: dist. shared mem & single OS**
  - **CPU and GPUs are connected by high-speed & low latency link with cache coherency.**
- **CPU**
  - **Many-core architecture with ARM inst. set**
  - **FP64 Vector computing performance for HPC, and Low-precision Matrix computing for AI inference (FP16/BF16/FP8/INT8, etc.)**
- **Accelerators（GPUs）**
  - **FP64 Vector computing as well as low-precision Matrix computing (FP16/BF16/FP8/INT8, etc.)**
  - **HBM or more advanced memory technology**
  - **Also DDR memory for more capacity**
  - **Support multiple GPU instance and virtualization**

Architectural Outlook Proposed by Feasibility Study

High BW & heterogeneous node arch and whole system overview



Performance target of the entire system

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

Goal: More than 5-10x effective performance gain in HPC apps, more than 50EFLOPS effective AI performance (needs Zetta-scale low-precision perf.) , and 10-100x apps performance improvement by combining simulation and AI