

Regional strategies around convergence of HPC, AI, and Quantum - General roadmaps and projects in Japan -

**Masaaki Kondo
(Division Director)**

**Kento Sato
(Team Principal)**

**Miwako Tsuji
(Unit Leader)**

RIKEN Center for Computational Science

Agenda

- **Overview of High-Performance Computing Infrastructures in Japan**
- **Introduction of FugakuNEXT Development Project**
- **Japanese AI-for-Science Roadmap and RIKEN TRIP AGIS Project**
- **Quantum-HPC Hybrid Platform in Japan**

Major Nation-Wide Programs for Supercomputer Usage

- Mostly proposal-based and free of charge
 - Accept international projects and industry program
 - Paid use also available based on policy of individual institutions
- HPCI (2012-)
 - For large-scale computing (simulations)
 - RIKEN R-CCS
 - Supercomputing center in several national universities
 - Hokkaido U. / Tohoku U. / U. Tsukuba / U. Tokyo / Sci. Tokyo / Nagoya U. / Kyoto U. / Osaka U. / Kyushu U.
 - National Labs
 - AIST, NII, ISM, JAMSTEC
 - Large Storage Systems (West: Kobe by R-CCS, East: Kashiwa by U.Tokyo)
- JHPCN (2010-)
 - For fundamental research
 - 8 National Universities (Core: U.Tokyo)

HPCI High Performance
Computing Infrastructure



Overview of HPCI

Courtesy: RIST

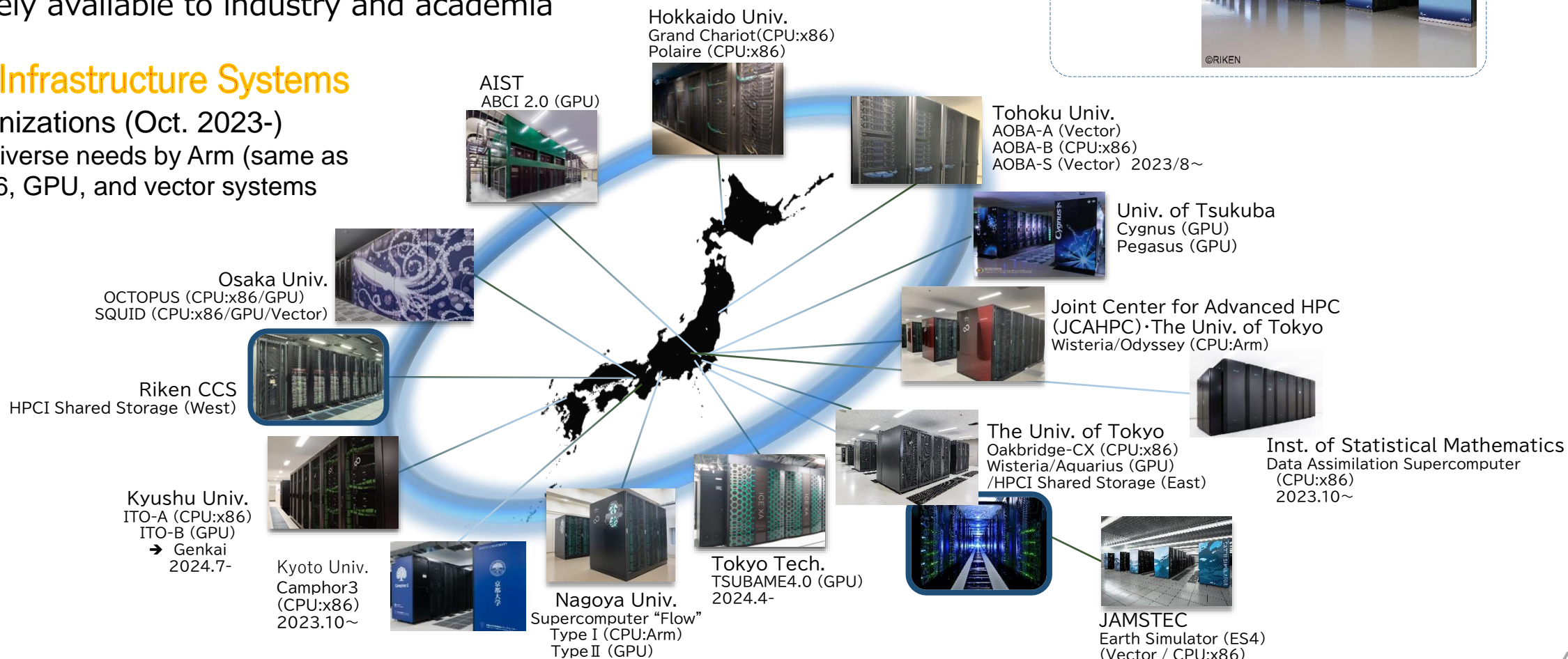
● HPCI: High Performance Computing Infrastructure

- Connects state-of-the-art supercomputers and storage at universities and research institutes in Japan via the high-speed network SINET6, enabling their integrated use and making them widely available to industry and academia

National Infrastructure Systems

11+2 Organizations (Oct. 2023-)

To support diverse needs by Arm (same as Fugaku), x86, GPU, and vector systems



- Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (2010-)
- Alliance of Supercomputer Centers of eight National Universities
 - 7 “Imperial” Universities + Tokyo Tech
 - Core Institute: ITC/U.Tokyo
 - Total 140+PFLOPS (as of May 2022)
- Promotion of collaborative (fundamental, interdisciplinary) research projects using facilities & human resources of 8 Centers



International Corroboration Activities

- MoU between DOE & MEXT on HPC incl. AI
- ANL-Riken MOU on AI for Science
- HANAMI Project
- :




DOE-MEXT

David Turk (DoE Deputy Secretary)
Masahito Moriyama (MEXT Minister)
(April, 2024)



ANL-Riken

Paul Kerns & Rick Stevens (ANL)
Makoto Gonokami, Makiko Naka,
Satoshi Matsuoka & Makoto Taiji (Riken)
(April, 2024)




About HANAMI


HANAMI is a project funded by the EuroHPC Joint Undertaking, the first of its kind regarding international collaboration

HANAMI implements the European Japan Digital Partnership regarding the High Performance Computing area

The project focuses on:

- **Climate Modeling**
- **Materials Science**
- **Biomedical Science**
- High-Performance Computing and Artificial Intelligence







About HANAMI

HANAMI promotes scientific projects involving both Europe and Japanese institutes, and will assist the researchers to access supercomputers in both Japan and Europe

14 Organizations from Europe



10 Organizations from Japan

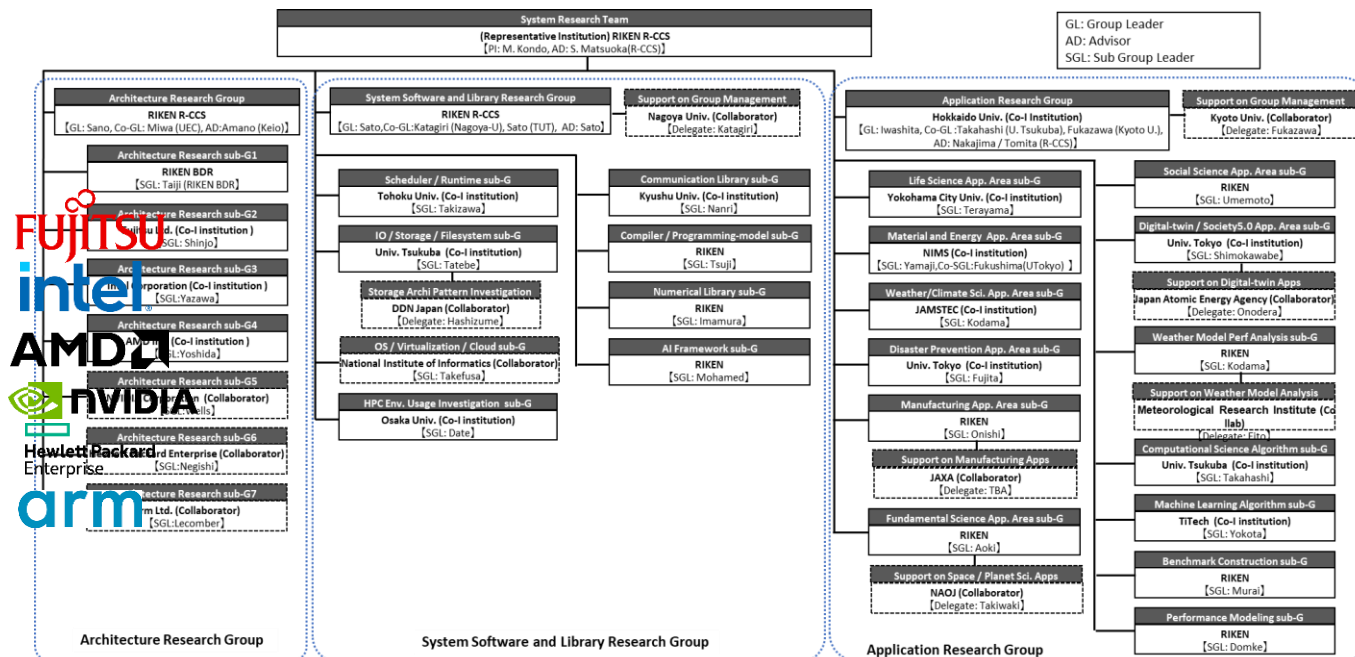


FugakuNEXT: From Feasibility Study To R&D Project

Overview of Feasibility Study (by MEXT)



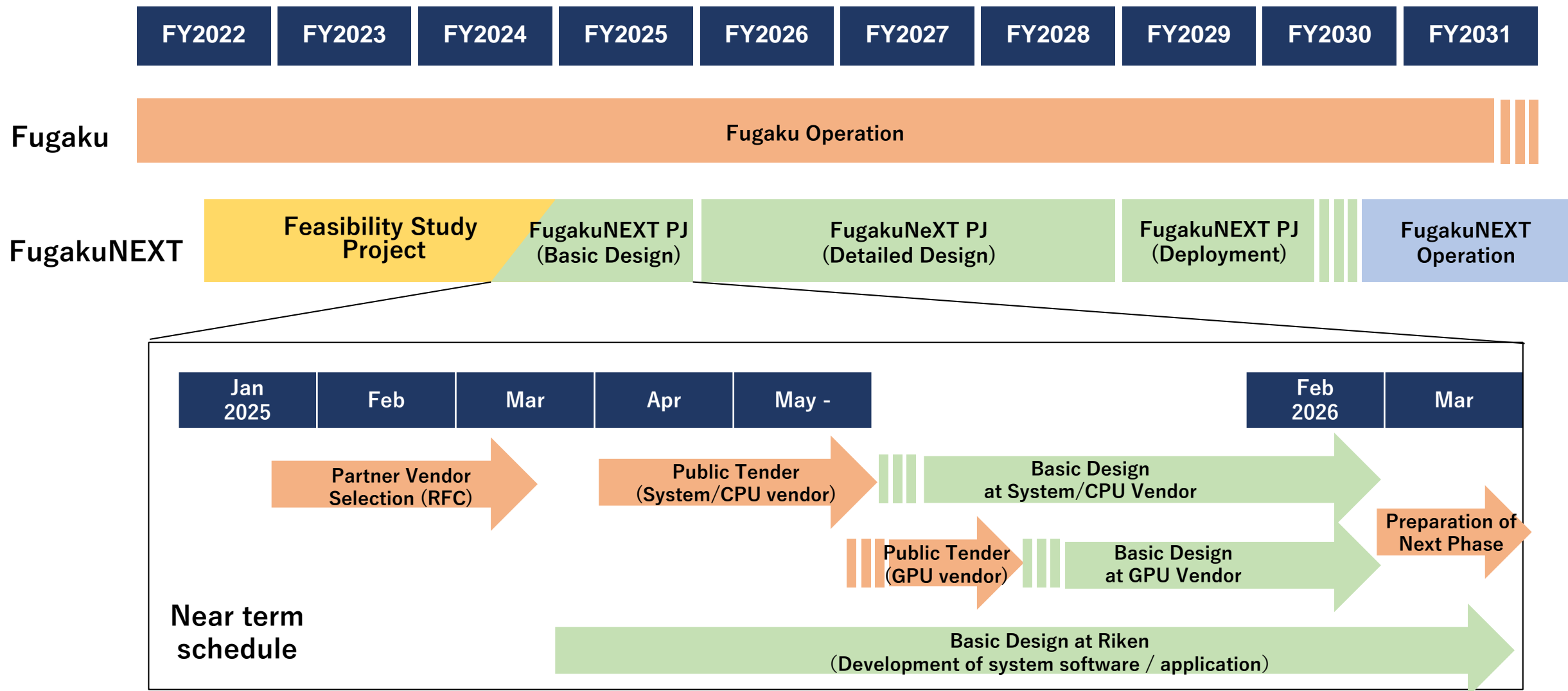
Organization Structure of RIKEN Team



- Establishing a project organization built upon the existing collaboration structure
 - Multiple companies in the RIKEN team have been expressing strong interest in participation
 - Nearly 200 comments were received to the draft specification for RFC of FugakuNEXT basic design
- Continuously studying HPC-Quantum hybrid platform for extending computable problems

Expected 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

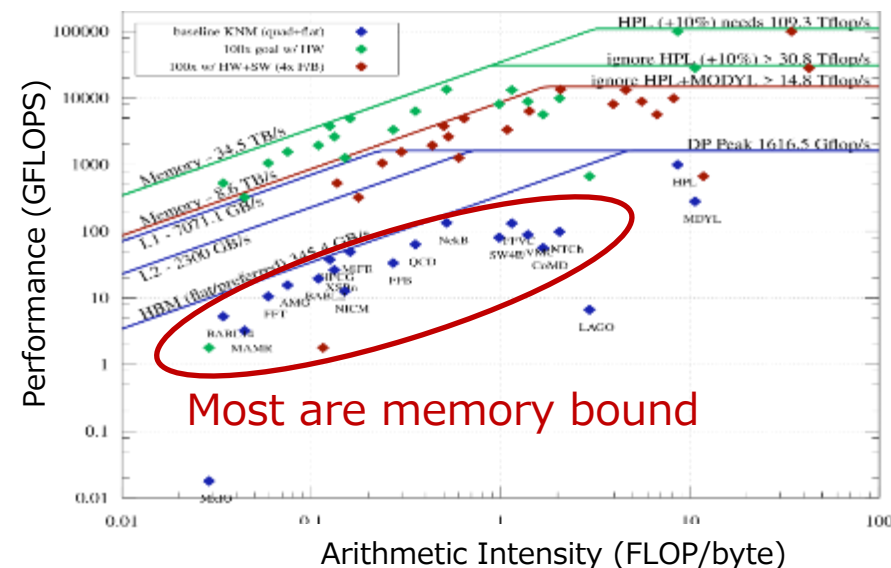


- **Expected to be a globally distinctive hub for computational science**
 - Quantum-HPC Platform with a state-of-the-art quantum computer (IBM-Q)
 - Introduction of a new AI-for-Science machine
 - FugakuNEXT as an HPC-AI integration platform

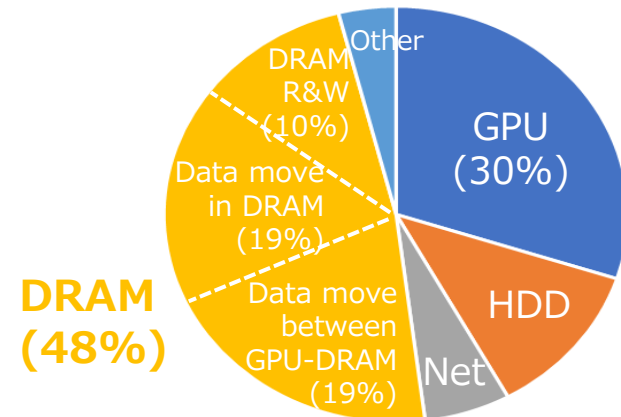
Architectural Direction for AI-for-Science Era

- Realizing a **world-class AI-HPC infra** for advancing science through the 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 surrogate models
- **Optimized data movement** for perf. & power efficiency
 - Using advanced memory tech. available at deployment
 - Interconnection NW design for both simulation and AI
- **Heterogeneous & tightly coupled architecture**
 - CPU+GPU architecture with “Made with Japan” concept
- System configuration that prioritizes the **ecosystem** and incorporates **open standards**
 - Ensuring compatibility with existing system SW, such as AI frameworks, programming env., and file systems
 - System that can be leveraged for cloud

Roofline Analysis



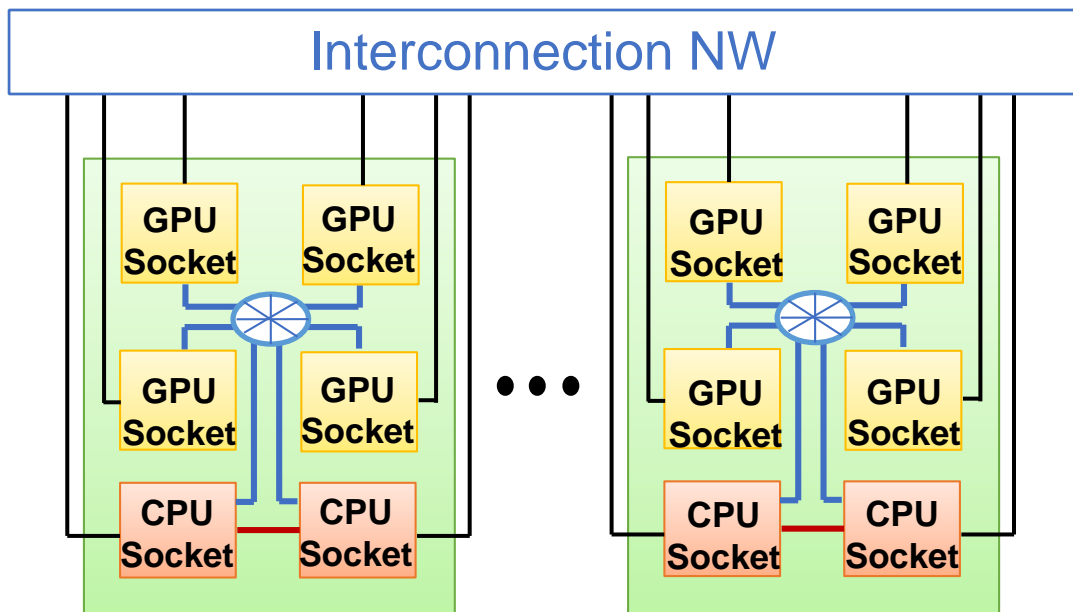
Power breakdown of a GPU server



Based on: J. Zhao et al., “Optimizing GPU energy efficiency with 3D die-stacking graphics memory and reconfigurable memory interface,” in *ACM Trans. Architecture and Code Opt.* vol. 10, Issue 4, pp. 1-25, 2013.

Architectural Outlook Proposed by Feasibility Study

High BW & heterogeneous node arch and whole system overview



Performance target of the entire system

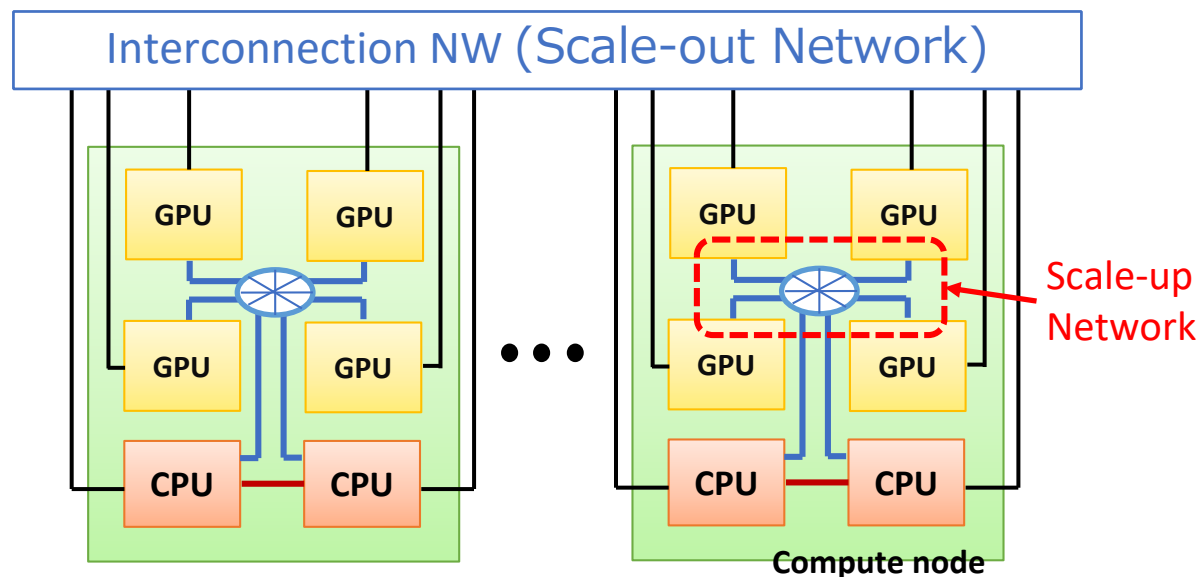
	CPU	GPU
Total Num. of Nodes	>= 3400 Nodes	
FP64 Vector FLOPS	>= 48PFLOPS	>= 3.0EFLOPS
FP16/BF16 AI FLOPS	>= 1.5EFLOPS	>= 150EFLOPS
FP8 AI FLOPS	>= 3.0ELOP	>= 300EFLOP
FP8 AI FLOPS (w/ sparsity)	—	>= 600EFLOPS
Memory Size	>= 10PiB	>= 10PiB
Memory Bandwidth	>= 7PB/s	>= 800PB/s
Total power consumption	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

FugakuNEXT R&D Strategy

- **Technological Innovation**
 - 10-100x apps performance improvement by AI accelerating technologies
 - High BW and heterogeneous node arch with advanced memory technologies
 - System design for emerging computational demands such as "AI for Science"
- **Continuity and Sustainability**
 - Ensure compatibility with existing ecosystems and standard specifications enabling continuous development of software environment
 - Establishing a system environment for continuous R&D and operation enhancement
 - Achieving energy efficiency through advanced operational technologies
- **Made with Japan**
 - Made with international collaboration as well as advancing domestic technologies
 - Driving the project through technological and human resource partnerships both domestically and internationally

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**

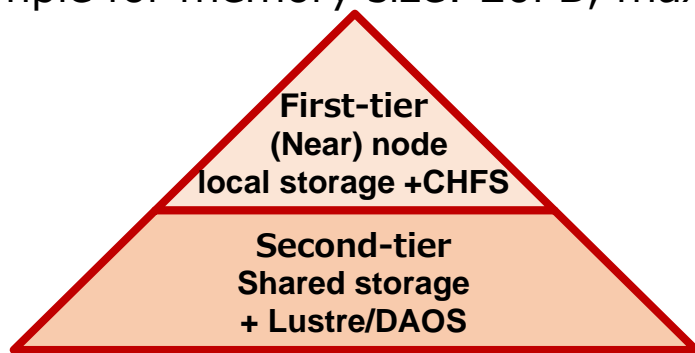
- **Compute node with CPU and Acc (GPU)**
 - Compute node: dist. shared mem & single OS
 - CPU and GPUs are connected by high-speed & low latency link with cache coherency.
- **CPU**
 - Many-core architecture with ARM inst. set
 - FP64 Vector computing performance for HPC, and Low-precision Matrix computing 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

Expectation of Storage System (Under Consideration)

- Direction to storage system for FugakuNEXT
 - Need advanced storage system that can treat with new I/O request for data science, large scale checkpoint, and AI-for-Science
 - Requirement of storage system performance and size from users *SSF: Single Shared File

	architecture	file system	Bandwidth (Effective performance of SSF*)	input-output per second	capacity
Level 1	(near)node local Storage	Under consideration (CHFS, etc.)	Total memory dump time: less than 1 minute	Metadata processing time: less than 1 second	More than twice the total memory size
Level 2	PFS	Lustre, DAOS	Total memory dump time: less than 5 minutes	1/10th IOPS of the first tier	Total memory size: More than 30 times

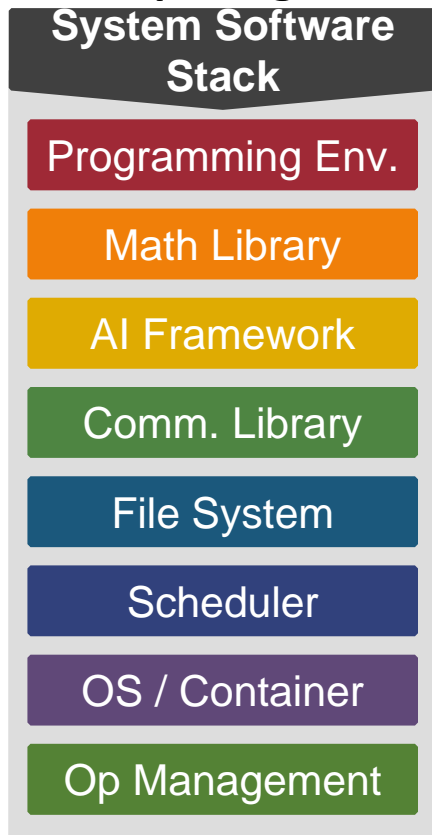
- Data migration from Fugaku to FugakuNEXT (Continuous operation and usage)
- Sustainable development of file-system and system software (needs OSS-based)
- An example of FugakuNEXT storage system (subject to change based on further assessment)
(example for memory size: 20PB, max num. of I/O processes: a few tens millions processes)



First tier	bandwidth: 350 TB/s (stable perf. by SSF) IOPS: More than 100M IOPS (more than 1 IOPS per process) Size: 40 PB
Second tier	bandwidth: 70 TB/s (stable perf. by SSF) IOPS: More than 10M IOPS (more than 0.1 IOPS per process) Size: 600 PB

Direction of System Software R&D

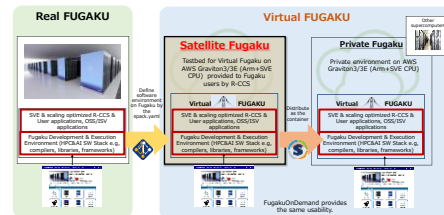
- Developing evolutionary SW that balances cutting-edge innovation and sustainable use
 - SW ecosystem in collaboration with domestic and international communities with effective use of OSS
 - Programming and software environment to efficiently utilize CPUs and accelerators
 - Advanced software stack that enables not only scientific simulations but Also AI for Science, Quantum-HPC Computing, and their Integration



Continuous R&D toward "Fugaku NEXT" and beyond
(Expanding to domestic/international communities for feedback)

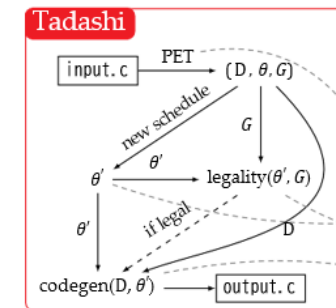
Utilization and Extension of SW stack developed for Fugaku

- Virtual Fugaku: Fugaku as part of cloud infra and implementing Fugaku app and SW env on AWS
- Open OnDemand : Using Fugaku via GUI (Many apps can be run easily)
- (and others)



Modernization of HPC apps and operation by AI

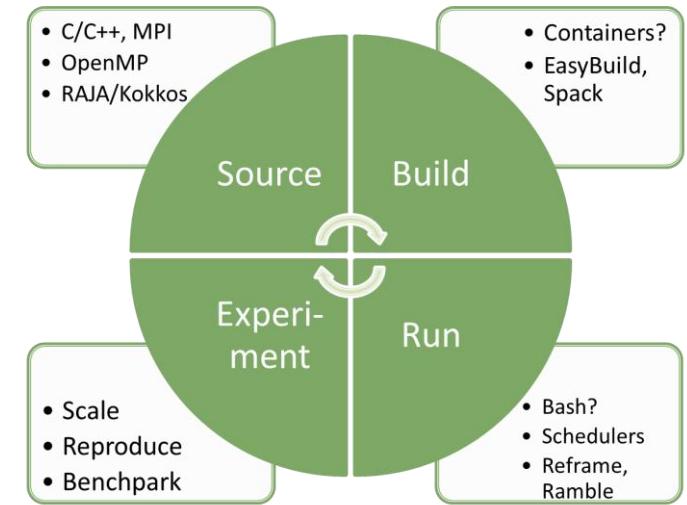
- On-line help desk by generative AI
- Accelerating traditional HPC code utilizing AI hardware
- Porting/Parallelization/Optimization by coding AI



Application First System Design

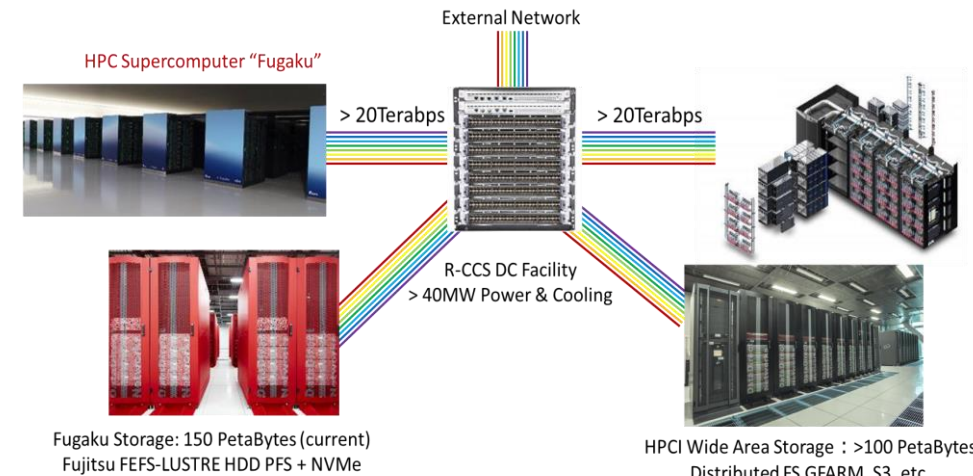
- Co-design strategy

- Benchmarking a wide variety of apps with continuous verification through AI-driven CI/CD/CB
 - Leveraged by Benchpark effort (collaboration with LLNL)
- Timely and appropriate dissemination of info on system development status and architectural perspectives to a wide range of application stakeholders



- AI for Science: utilizing concept tried out in RIKEN TRIP-AGIS project

- Tight coupling High GPU (throughput) and CPU (latency) performance for both AI and HPC performance
- Extensive mixed precision and emulation support
- Convergence of Scale-up and Scale-out network beyond standard HPC network
- Apps/SW will be pre-developed/tested on the precursor TRIP-AGIS machine

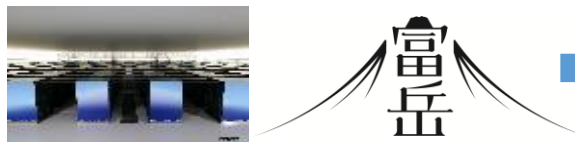


Science Target in FugakuNEXT Era

2011~ 「K computer」



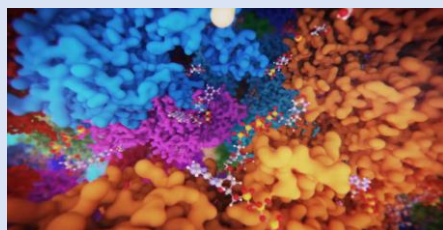
2020~ 「Fugaku」



2030

「Fugaku
NEXT」

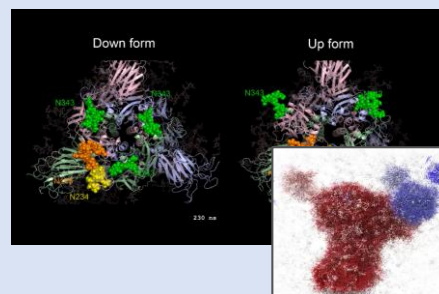
Simulation of Subcellular Sequence Dynamics



Faster all-atom
molecular dynamics
calculations (>100x)

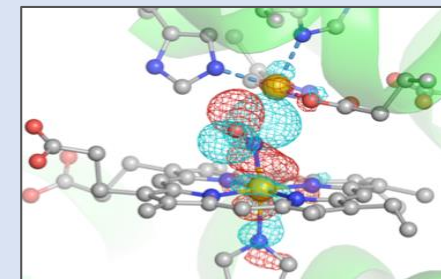
Long-term dynamics
and cellular function
multi-scale models

The “K computer” achieves short time
dynamics of 100 million atoms system.



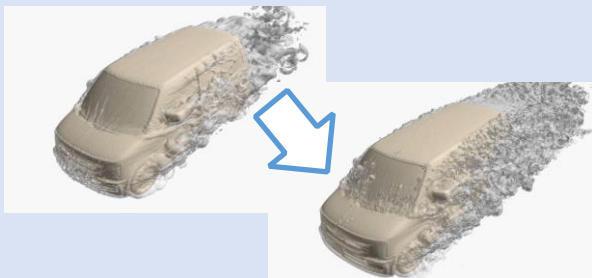
“Fugaku” allows for longer dynamics
of even larger systems.

Parallel evolution
of machines and
algorithms
(coarse-grained)
accelerated x10~

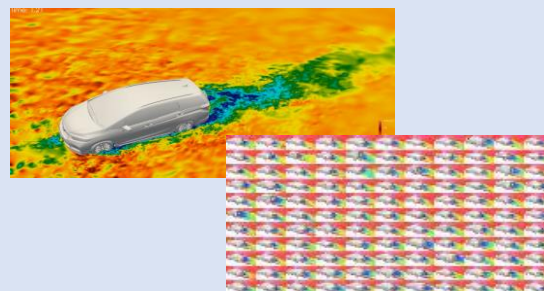


Enables dynamics considering
electronic states (applied to bio-
digital twin antibody drug
discovery, etc.)

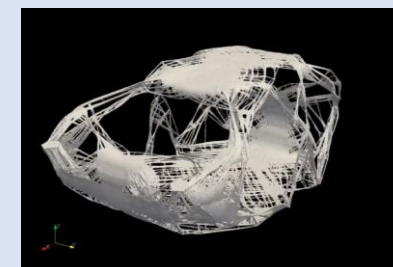
Automobile aerodynamics



Wind tunnel replacement by high-resolution LES
Fundamental research



Digital Twin (Upper)
AI-Assisted Multi-Objective Optimization (Lower)
to Shorten Automotive Design Time



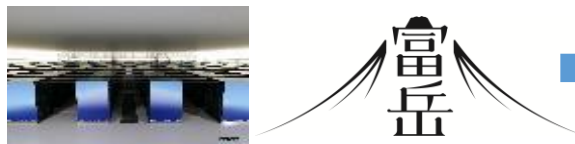
Automation of automobile design
by proposing optimal shapes using
generative AI
Establishment of self-driving tech.

Science Target in FugakuNEXT Era

2011~ 「K computer」



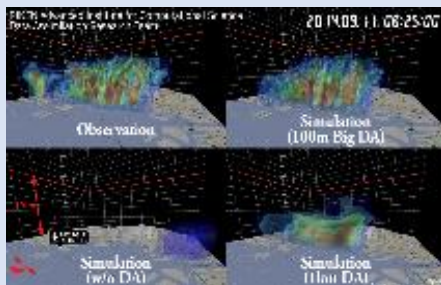
2020~ 「Fugaku」



2030

「Fugaku
NEXT」

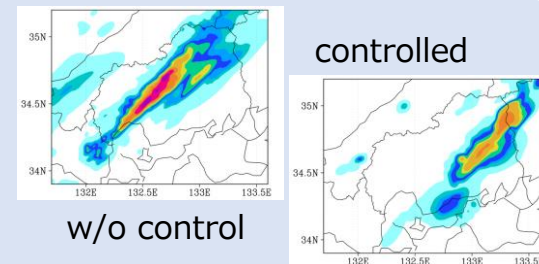
Weather and Climate



Development of a guerrilla rainfall forecasting method using the "K computer"



World's first Real-time guerrilla rainstorm forecast by "Fugaku" during 2021 Tokyo Olympic & Paralympic Games

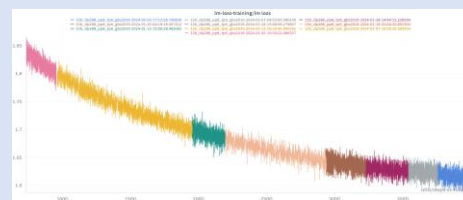


Solving the global climate crisis
Integrate with social and urban digital twin and AI to virtual trial and recommendation of policies

Fugaku LLM (13 billion parameters)

Target models	Number of tokens learned
13B Transformer models	230B Token

It takes about "10 -15 years" to train Fugaku LLM in advance.



Fugaku LLM pre-study completed in "a month" using "Fugaku"'s 1/11th scale



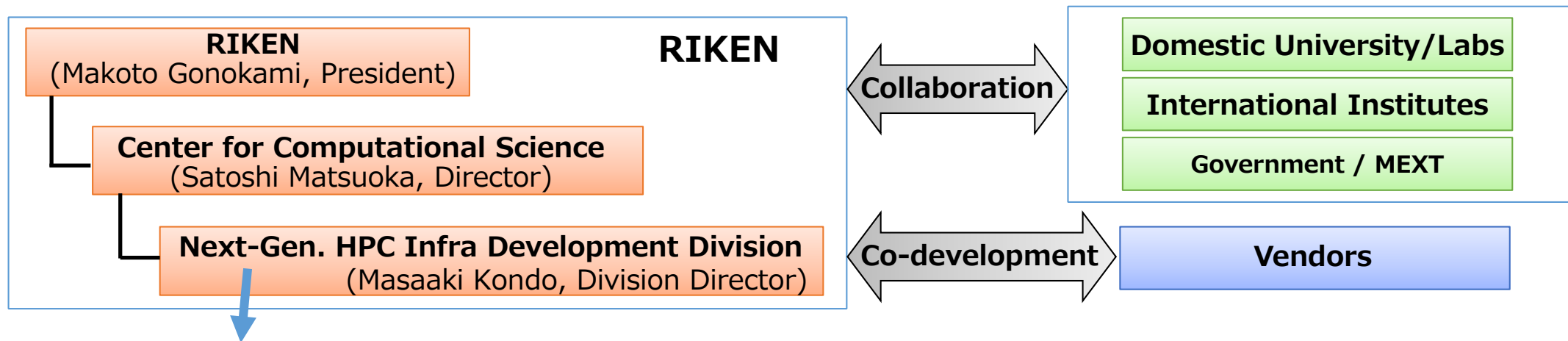
Available free of charge on the Fujitsu Research Portal SambaNova of the U.S. provides a commercial platform.








<https://portal.research.global.fujitsu.com/>

Pre-training of state-of-the-art trillion-level parameter infrastructure models in 2 months

Dramatic evolution of the innovation cycle through AI for Science acceleration

New Division for FugakuNEXT project and Beyond



Division	Units		Tasks
Next-Generation HPC Infrastructure Development Division   Masaaki Kondo, Division Director Fumiyoshi Shoji, Deputy Division Director	Next-Generation HPC Infrastructure System Development Unit	Kentaro Sano, Unit Leader 	Development of FugakuNEXT system including architecture and system SW
	Next-Generation HPC Application Development Unit	Yasumichi Aoki, Unit Leader 	Development, support, and co-design of next generation apps for FugakuNEXT
	Next-Generation HPC Operation Technologies Unit (Under Consideration)	Keiji Yamamoto, Unit Leader 	Development of environment and operation technologies for FugakuNEXT
	Advanced HPC Technologies Development Unit	Kento Sato, Unit Leader 	Feasibility study and elementally technology research for advanced system
	Next-Generation HPC Management Office	Yoji Shimada, Office manager 	Responsible of the projet management of FugakuNEXT and beyond

Agenda

- Overview of High-Performance Computing Infrastructures in Japan
- Introduction of FugakuNEXT Development Project
- **Japanese AI-for-Science Roadmap and RIKEN TRIP AGIS Project**
- Quantum-HPC Hybrid Platform in Japan

- **Develop generative AI models (scientific foundation models) specialized for scientific research** by building collaborative frameworks with research institutions that have strengths in scientific research fields and using the foundation models to conduct **fine-tuning, multimodalization of scientific research data**, etc.
- **By widely opening up the use of our AI models from scientific research to industry and academia**, we aim to innovate scientific research in various fields (dramatically accelerate the scientific research cycle and expand the exploratory space for scientific research).

High quality data

- Collect and maintain high quality data for training, fine tuning, etc.
- Collaboration and joint development with related research institutions that accumulate data
- Target science fields:
 - (1) Life and Medical Sciences (e.g., predicting differences due to dynamic changes and genetic mutations caused by drugs, etc.)
 - (2) Materials/physical properties science (e.g., prediction of physical properties of novel materials)

Advanced model

- Develop, operate, and share scientific multi-modal foundation models for the target science fields
- In parallel, research and development necessary to read, learn, and generate multimodal data

Computing resource

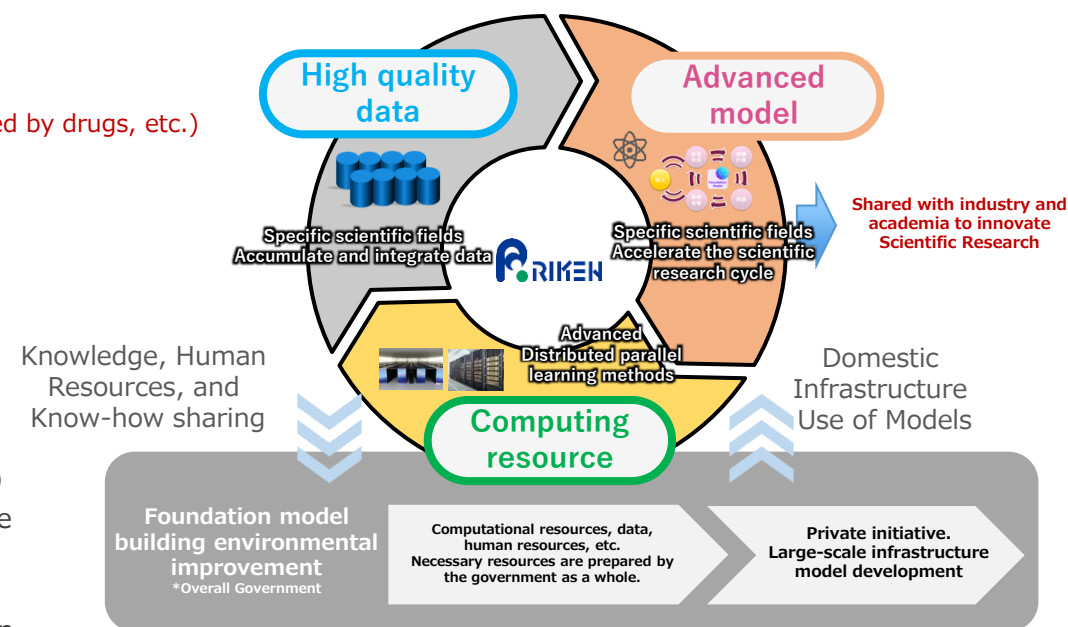
- Define requirements for AI-for-Science research, and procure and operate an AI-for-science system (GPU)
- Combine Supercomputer "Fugaku" with the AI-for-Science system through high-speed network to facilitate interplay between these systems for developing the AI models
- Develop software for accelerating fine-tuning and inference
- Research on new AI architectures (dedicated computing machines other than matrix computation) that can handle multimodal foundation model

Tentative Discussion Points on AI (AI戦略会議, May 26, 2023)

AI Development Capability

- It is also almost certain that the results of AI research will contribute to the acceleration of R&D in areas other than AI.
- As the world is about to be revolutionized by generative AI, it is important to foster fundamental research and development capabilities for generative AI in Japan as quickly as possible.
- It is expected to build an environment for research and human resources development where top talents from all over the world can gather and compete with each other in friendly competition, and to strengthen the basic development capabilities of industry, academia and government.

Innovation of Research through "Generative AI Models for Scientific Research"



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

TRIP-AGIS ①: Common Platform Technology

Develop common infrastructure for creating and sharing generative AI models for scientific research

TRIP-AGIS ②: Generative AI models for scientific research in specific fields

Develop generative AI models for target scientific research areas (Life and Medical Sciences / Material Science)

TRIP-AGIS ③: Innovative Computational Infrastructure

Develop pioneering innovative computational infrastructure for AI-for-Science computation

TRIP-AGIS ③-1: Operations for Innovative Computational Infrastructure (AI4S system)

Advance operation technologies for the innovative computational infrastructure enabling large-scale training and inference

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

Develop fundamental software for advancing the development environment of generative multimodal AI models for scientific research

TRIP-AGIS ③-3: New AI architecture technologies

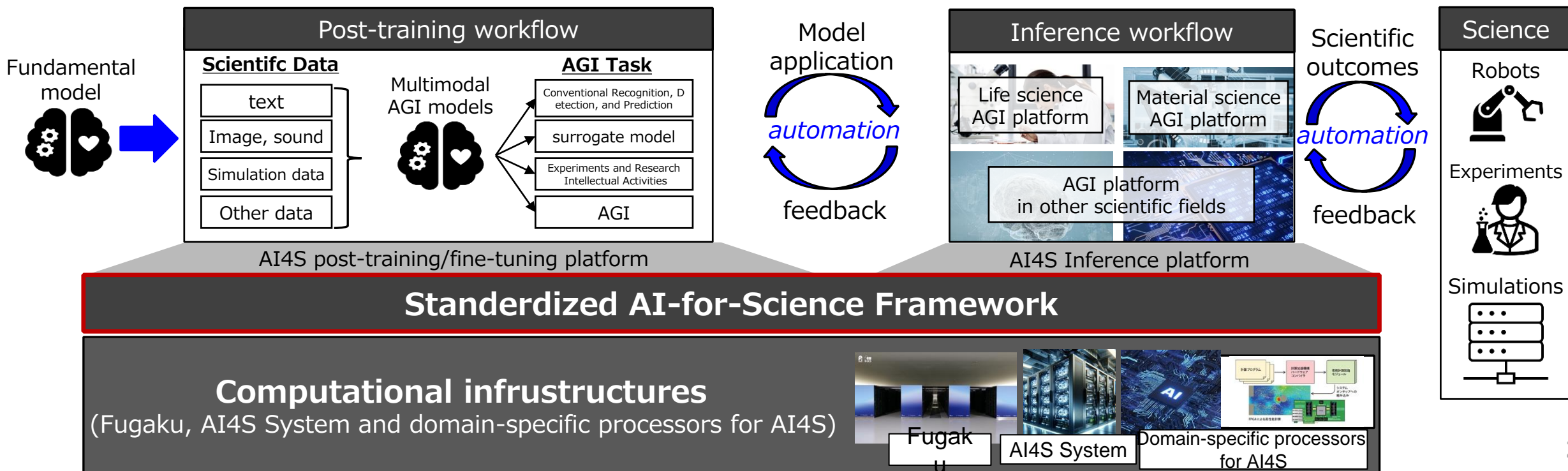
Develop new domain-specific architectures for AI training and inference

BDR

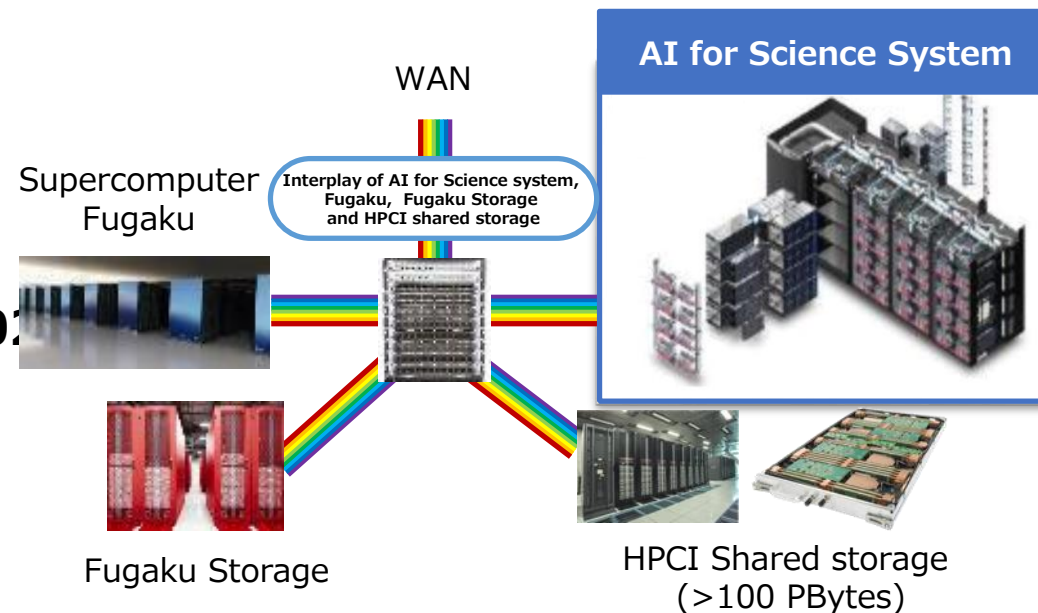
R-CCS

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

- **TRIP-AGIS ③-2-(i): Learning Optimization Platform Development Unit**
 - Analyze performance and advance system software technologies in post-training/inference for interplay between Fugaku and the AI4S system
 - Explore new software development for the new domain-specific AI processors
- **TRIP-AGIS ③-2-(ii): Data Management Platform Development Unit**
 - Analyze storage system and enhance I/O performance for post-training/inference of multimodal AI models requiring a variety of data types (e.g., text, images, sound, videos, other scientific data)
 - Explore new data management and curation (e.g., collection, compression, organization, encryption etc.)
- **TRIP-AGIS ③-2-(iii): Application Interface Platform Development Units**
 - Develop post-training/inference platforms enabling automation of model application and feedback in life and material science



- **Overview:** RIKEN is procuring a new supercomputer system to support large-scale training and inference for AI for Science. This system will complement Fugaku and be optimized for deep learning workloads, including generative AI models.
- **Objectives:**
 - Provide computing capabilities beyond Fugaku, focusing on AI for Science
 - Enable seamless integration with Fugaku
 - Support large-scale model training and inference, especially in FP8 precision
- **Installation Site:**
 - Located in RIKEN R-CCS, Kobe
- **Procurement Schedule:**
 - Start of Public Bidding: March 28, 2025
 - Public Bidding Result Announcement: By June 2025
 - Expected System Delivery: By March 2026



Key performance features

- **Architecture:**
 - Supports AArch64 or x86-64 ISAs
 - Equipped with high-performance accelerators (e.g., GPUs)
- **Performance:**
 - ≥ 20 PFLOPS in FP64 (double precision)
 - ≥ 8 EFLOPS in FP8 precision
- **Memory, Interconnect, Storage:**
 - **Memory:** ≥ 0.16 (CPU) & 6.3 (GPU) PB/s memory bandwidth / ≥ 135 TiB total memory
 - **Interconnect:** RDMA-enabled interconnect with ≥ 400 Gbps/GPU injection bandwidth
 - **Storage:** >1 PB usable capacity w/ Lustre-based parallel file system
- **Cooling & Power:**
 - Direct Liquid Cooling (DLC) used for $\geq 70\%$ of total power
 - Max system power: 4.5 MW

AI for Science Roadmap in Japan

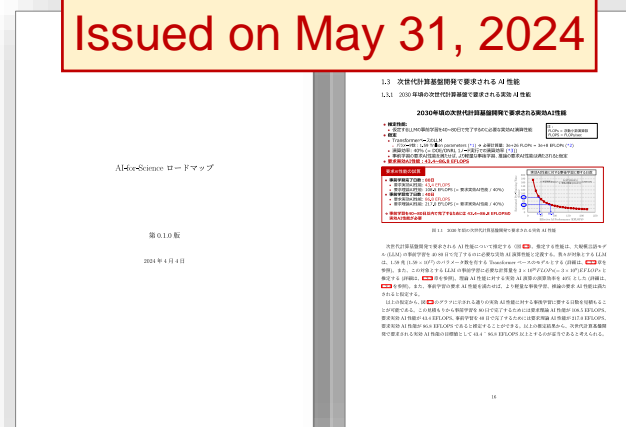
(Issued on May 31, 2024)

AI for Science Roadmap - Overview

- **Abstract:**
 - Summary of a roadmap to drive future AI-for-science researchers in Japan
 - A roadmap is written to clarify examples, guidelines and new challenges on the application of cutting-edge AI technologies such as surrogate modeling and the use of generative AI in Science
 - We estimated required AI computational performance to the next-gen supercomputer
- **Steering Committee:.**
 - Rio Yokota (Professor, Tokyo Institute of Technology), Takashi Shimokawabe (Associate Professor, The University of Tokyo), Masaaki Kondo (Professor, Keio University), Shinji Todo (Professor, The University of Tokyo)
 - (RIKEN R-CCS) Mohamed Wahib, Hirofumi Tomita, Kento Sato, Akiyoshi Kuroda
- **Target Fields: 11 fields listed in the HPCI Consortium Computational Science Roadmap**
 - Elementary Particle Physics & Nuclear Physics, Nanoscience & Devices, Energy & Materials, Life Sciences, Brain & Neuroscience, Drug Discovery & Medicine, Design & Manufacturing, Social Sciences, Earthquakes & Tsunami, Weather & Climate, Astrophysics
- **Authors : 59 (including 8 promoters)**
 - HPCI users searched by a keyword such as “AI” in their proposals
 - Authors of the HPCIC Computational Science Roadmap
 - Priority applications in FY2023 (成果創出加速プログラム)
 - RIKEN R-CCS



Issued on May 31, 2024



Expansion of AI application 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

We are currently translating the work into English and preparing it for publication on arXiv

Agenda

- Overview of High-Performance Computing Infrastructures in Japan
- Introduction of FugakuNEXT Development Project
- Japanese AI-for-Science Roadmap and RIKEN TRIP AGIS Project
- **Quantum-HPC Hybrid Platform in Japan**

Digital Annealers

Fujitsu Superconductive

QuEra Neutral-Atom

OptQC Optical

ABCI-Q system H

IBM Superconductive

Quantinuum Trapped ION

RIKEN Superconductive

National Flagship System

Riken R-CCS "Fugaku"

(CPU: Arm)



GPU Cluster in R-CCS

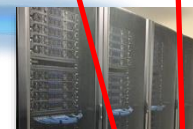
HOKKAIDO UNIV.
Grand Chariot(CPU:x86)
Polaire (CPU:x86)



AIST
ABCI 2.0 (GPU)



Tohoku Univ.
AOBA-A (Vector)
AOBA-B (CPU:x86)
AOBA-S (Vector) 2023/8~



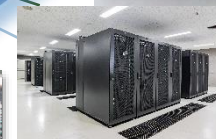
Univ. of Tsukuba
Pegasus (GPU)



Joint Center for Advanced HPC (JCAHPC)・The Univ. of Tokyo
Wisteria/Odyssey (CPU:Arm) Miyabi



The Univ. of Tokyo
Oakbridge-CX (CPU:x86)
Wisteria/Aquarius (GPU)
/HPCI Shared Storage (East)



Inst. of Statistical Mathematics
Data Assimilation Supercomputer (CPU:x86)
2023.10~



JAMSTEC
Earth Simulator (ES4)
(Vector / CPU:x86)



Tokyo Tech.
TSUBAME4.0 (GPU)
2024.4~



Nagoya Univ.
Supercomputer "Flow"
Type I (CPU:Arm)
Type II (GPU)



Kyoto Univ.
Camphor3 (CPU:x86)
2023.10~



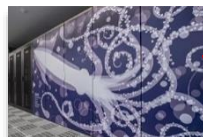
Kyushu Univ.
ITO-A (CPU:x86)
ITO-B (GPU)
→ Genkai
2024.7~



RIKEN R-CCS
HPCI Shared Storage (West)



Osaka Univ.
OCTOPUS (CPU:x86/GPU)
SQUID (CPU:x86/GPU/Vector)

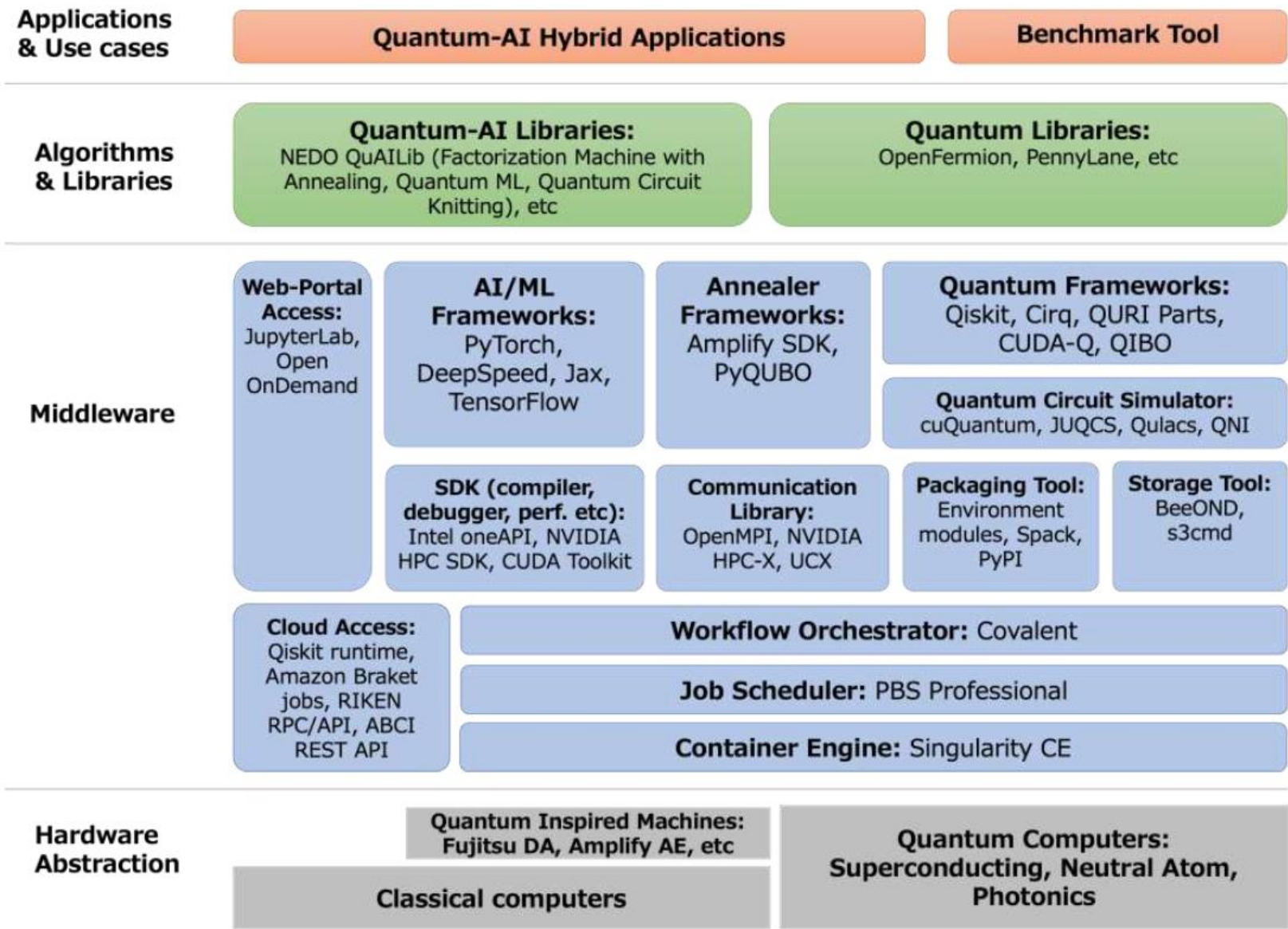


ABCI-Q: Quantum HPC Hybrid in National Institute of Advanced Industrial Science and Technology (AIST)

- System H: (Intel Xeon Platinum 8558 x2 + NVIDIA H100 SXM5 (80GB) x4) x 505 nodes
 - Fujitsu Digital Annealer, Fixstars Amplify AE, Toshiba SQBM+
- System F: Fujitsu Superconducting Quantum Computer, 64 qubits
- System Q: QuEra Neutral-Atom Quantum Computer, 256 qubits
- System O: OptQC Optical Quantum Computer



ABCI-Q: Quantum HPC Hybrid in National Institute of Advanced Industrial Science and Technology (AIST)



Deploy quantum and AI software libraries based on the widely adopted software stack in HPC systems

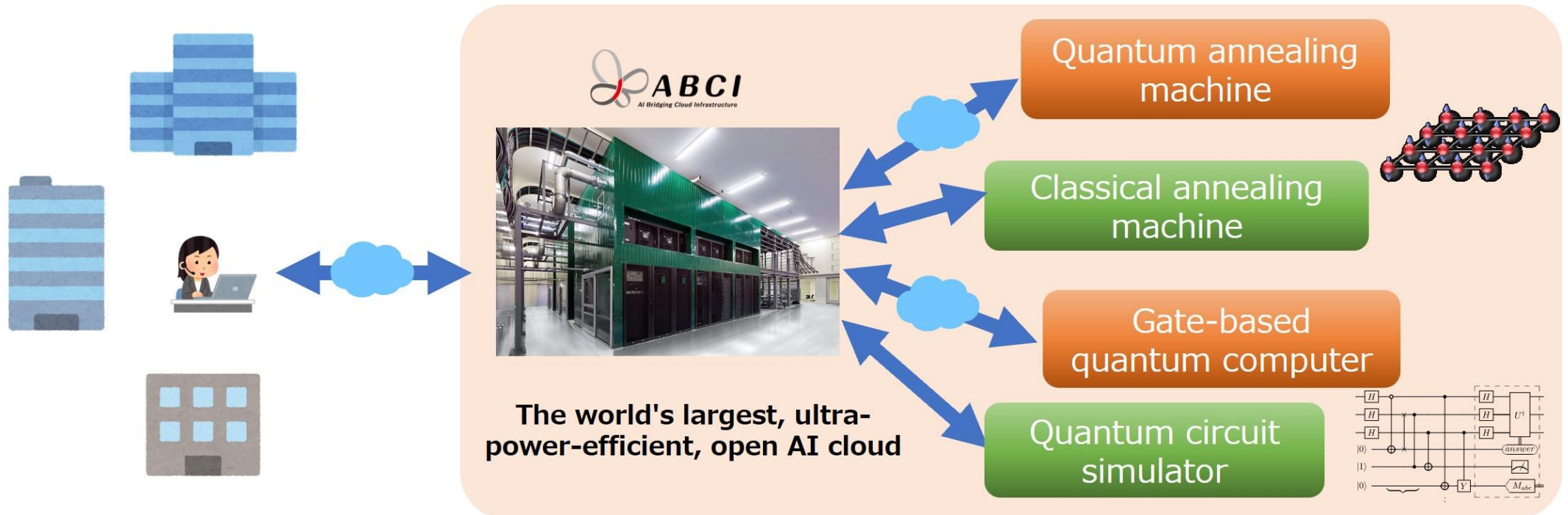
- GPU-based quantum circuit simulators and quantum annealing engines
- Simulator: NVIDIA cuQuantum Appliance
- Annealing engine: Fujitsu Digital Annealer, Fixstars Amplify AE
- a workflow tool for developing quantum-classical hybrid applications
- a web-based development environment for beginners in HPC systems and quantum computing

Schedule

- System H: (Intel Xeon Platinum 8558 x2 + NVIDIA H100 SXM5 (80GB) x4) x 505 nodes
 - Fujitsu Digital Annealer, Fixstars Amplify AE, Toshiba SQBM+
- System F: Fujitsu Superconducting Quantum Computer, 64 qubits
- System Q: QuEra Neutral-Atom Quantum Computer, 256 qubits
- System O: OptQC Optical Quantum Computer

Installed
will be
available soon

2026



Quantum HPC related Projects in RIKEN R-CCS

TRIP3 (2024.04-

Expansion of Computationally Viable Region,
Transformative Research Innovation Platforms of
RIKEN Platforms,

A cross-disciplinary project to establish
connections among RIKEN's cutting-edge research
platforms such as *supercomputers*, *quantum
computing*, large synchrotron radiation facilities,
bioresource-projects.

R-CCS and Center for Quantum Computing (RQC)
are tasked with “the value creation by quantum-
HPC hybrid computing for Accelerating Research
DX”.

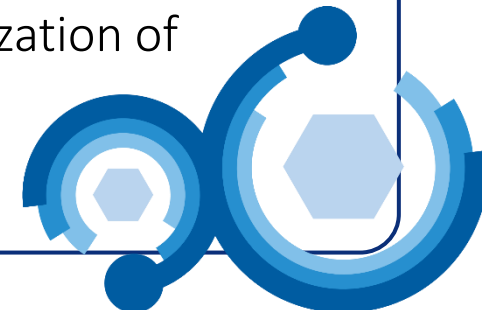


JHPC-Quantum (2024.11-

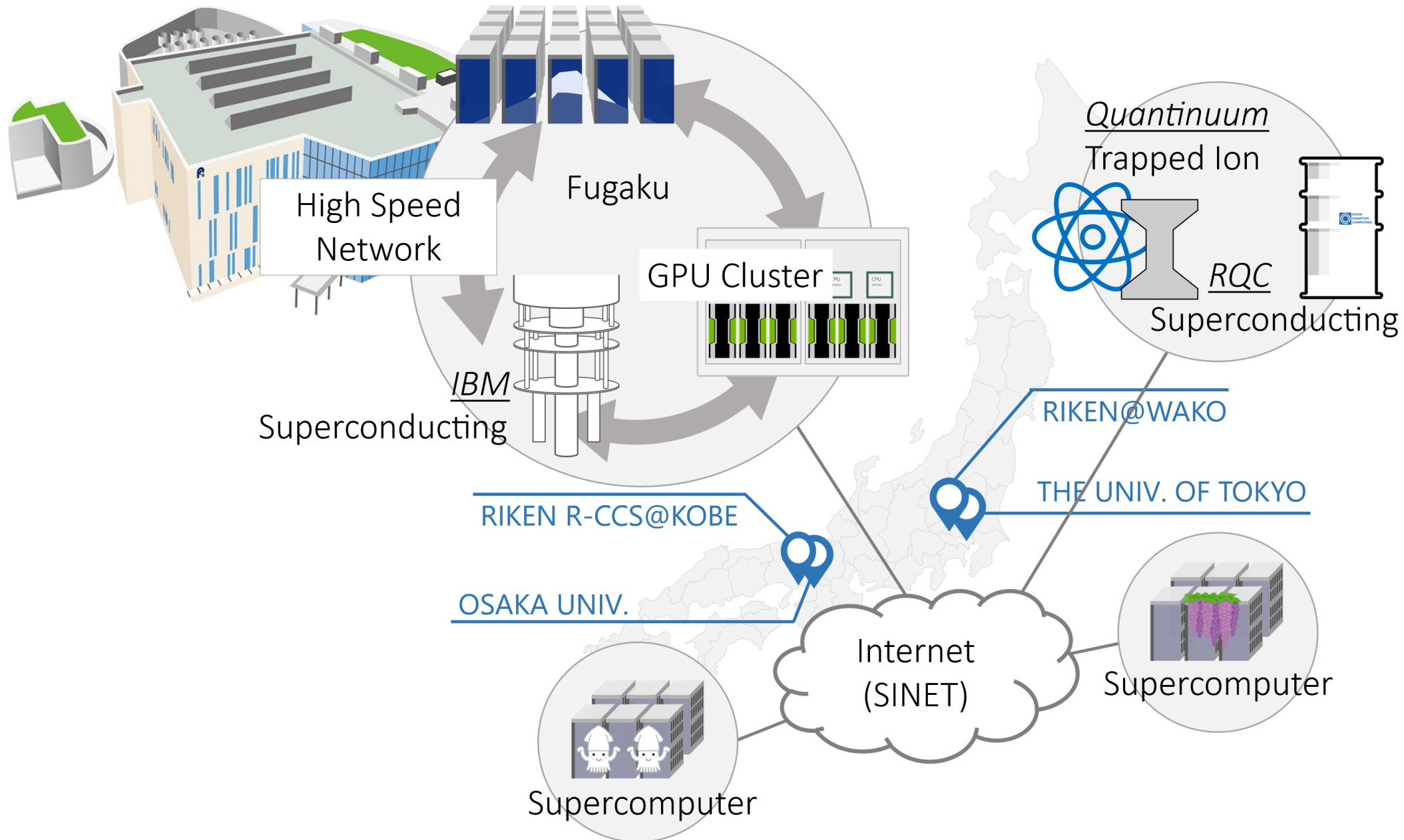
Research and Development of Quantum-
Supercomputers Hybrid Platform for Exploration of
Uncharted Computable Capabilities

One of research projects in “Research and
Development Project of the Enhanced
Infrastructures for Post-5G Information and
Communication Systems” in NEDO, an agency
under the Ministry of Economy, Trade and Industry.
R-CCS, Softbank Corp., the university of Tokyo, and
Osaka university

A platform that connects supercomputers and
different types of quantum computers
to promote the early commercialization of
quantum computers.

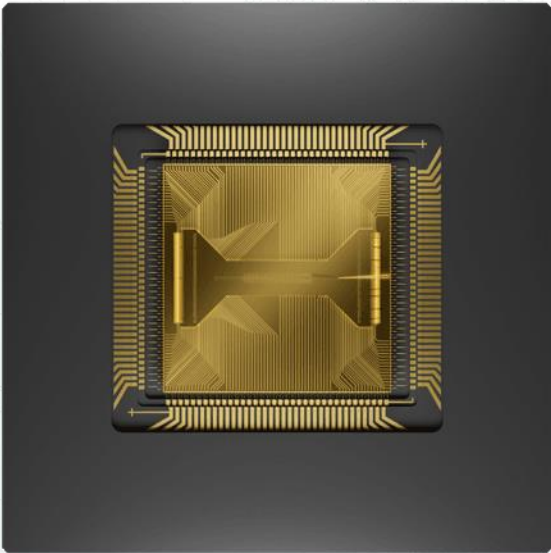


Overview (JHPC-Quantum & TRIP)



Recent Status of Quantum Systems

	IBM Kobe	Quantinuum 黎明 Reimei	RIKEN 叡 A
Name	Heron	System Model H1	A
Type	Superconductive	Trapped Ion	Superconductive
Qubit	156	20	64
Status	Installing	Running	Running



Pictures
<https://jp.newsroom.ibm.com/2023-12-05-IBM-Debuts-Next-Generation-Quantum-Processor-IBM-Quantum-System-Two,-Extends-Roadmap-to-Advance-Era-of-Quantum-Utility>
<https://www.quantinuum.com/products-solutions/quantinuum-systems/system-model-h1>
https://www.riken.jp/medialibrary/riken/pr/news/2023/20231005_1_photo1.jpg

Recent S

Name
Type
Qubit
Status

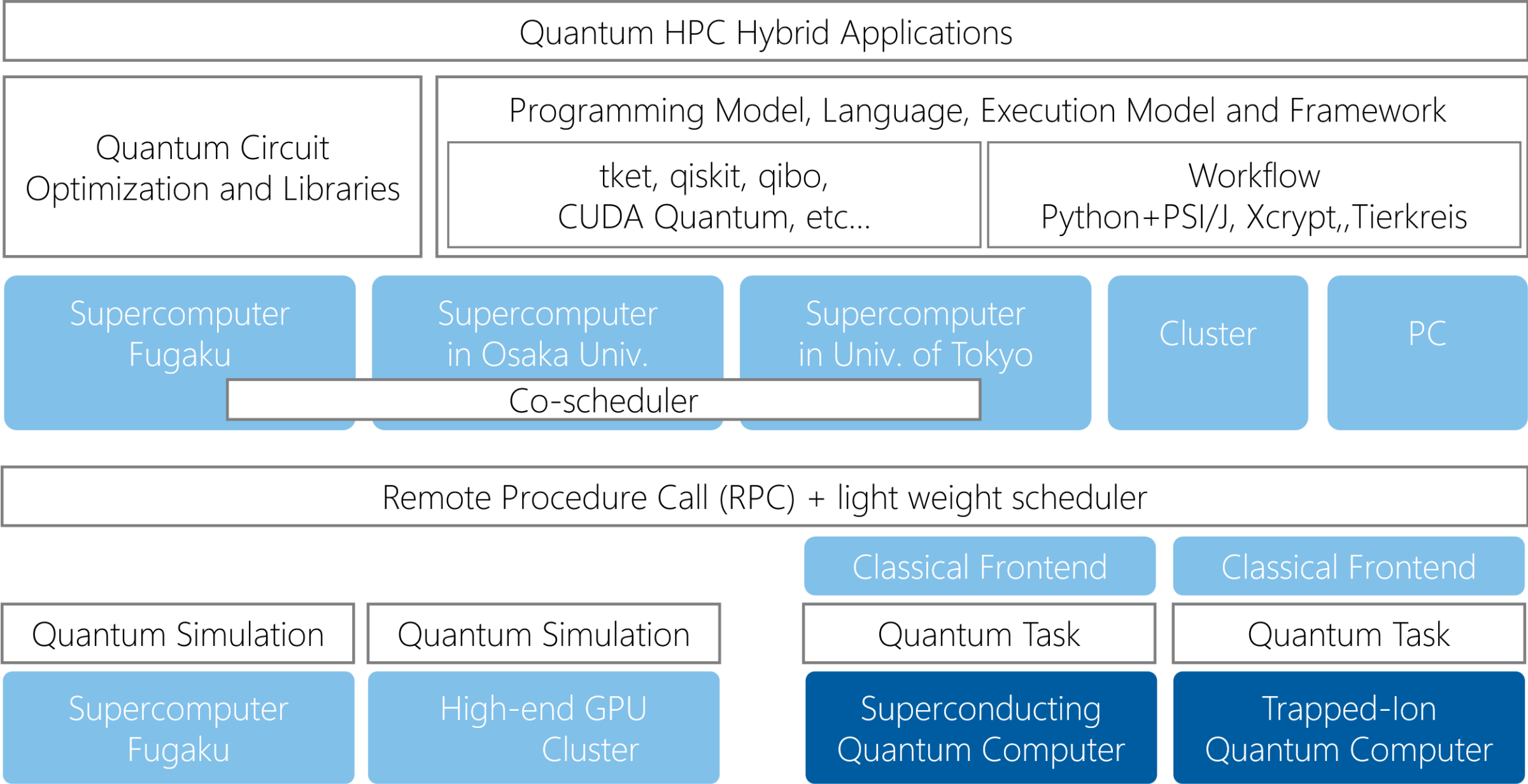


Pictures
<https://jp.newsro>
<https://www.quan>
<https://www.riken>

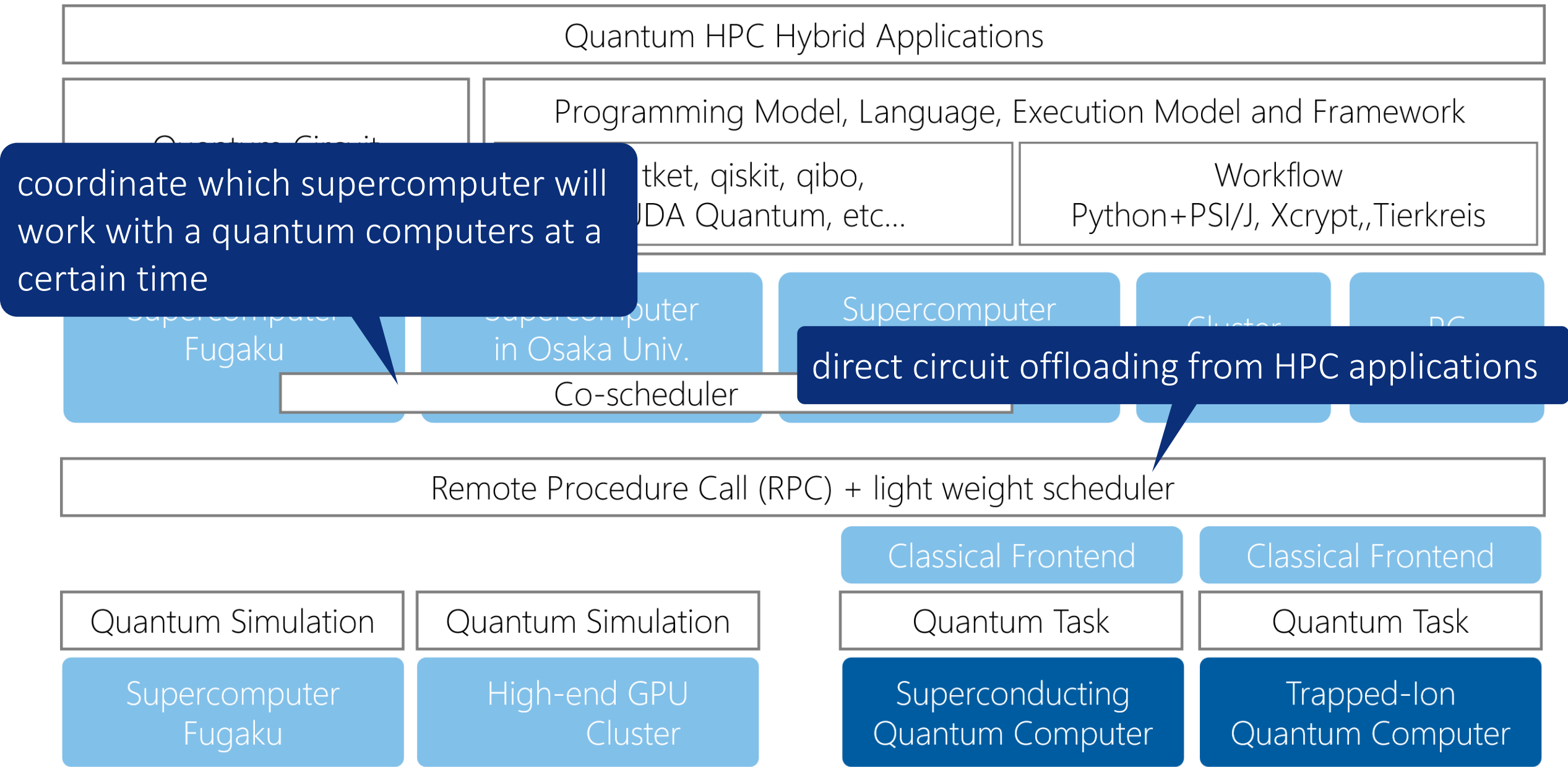




Software Stack

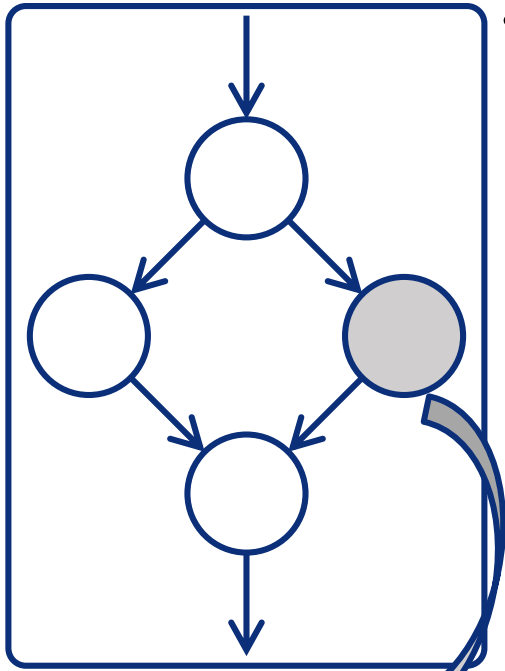


Software Stack






Quantum HPC Middleware: Two Level Programming Model

- Workflow & Task based parallelism



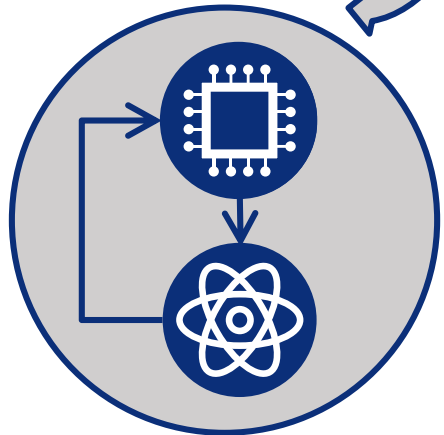
- Workflow

- Tasks (different kernels) are managed by a workflow engine
- The tasks are “**jobs**” in supercomputers and/or quantum computer
- The workflow engine submit each of jobs in a certain order based on the descriptions about dependencies between jobs

Simple HPC Job 	Simple QC Job 	HPC-QC Job 
running a kernel only on HPC	offloads a quantum circuit to quantum computer	includes both of HPC and quantum kernels

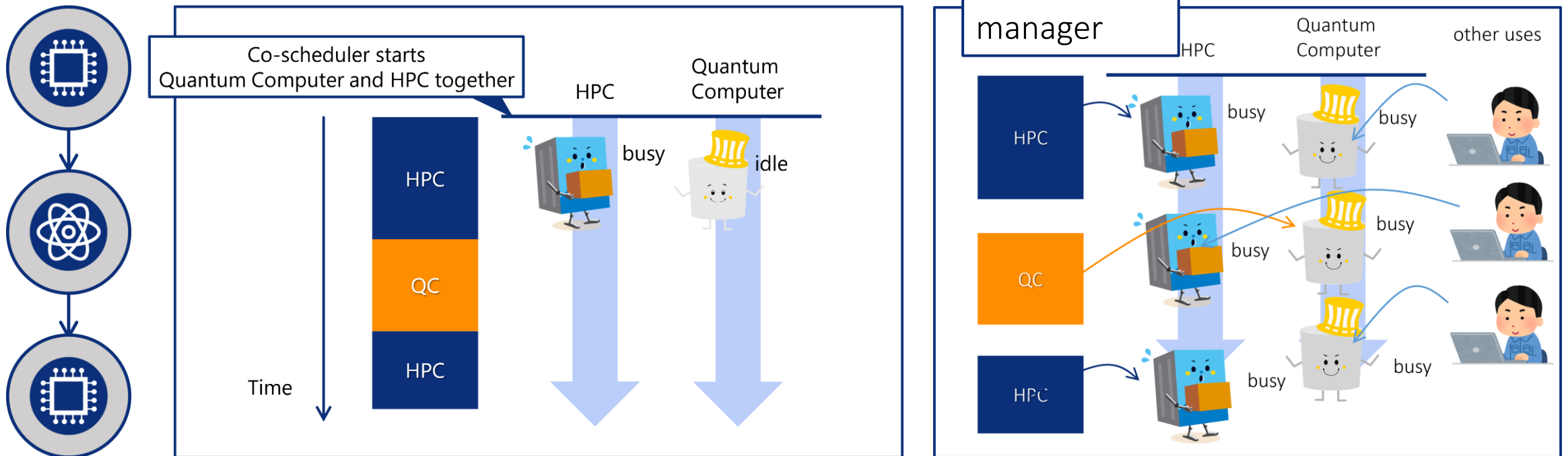
- Task based Parallelism in HPC-QC Jobs

- offloads quantum kernel(s) to a quantum computer via remote procedure call (RPC)
- A single HPC-QC job may offload multiple kernels
- RPC is an asynchronous mechanism; Other classical kernels will be executed on CPU simultaneously



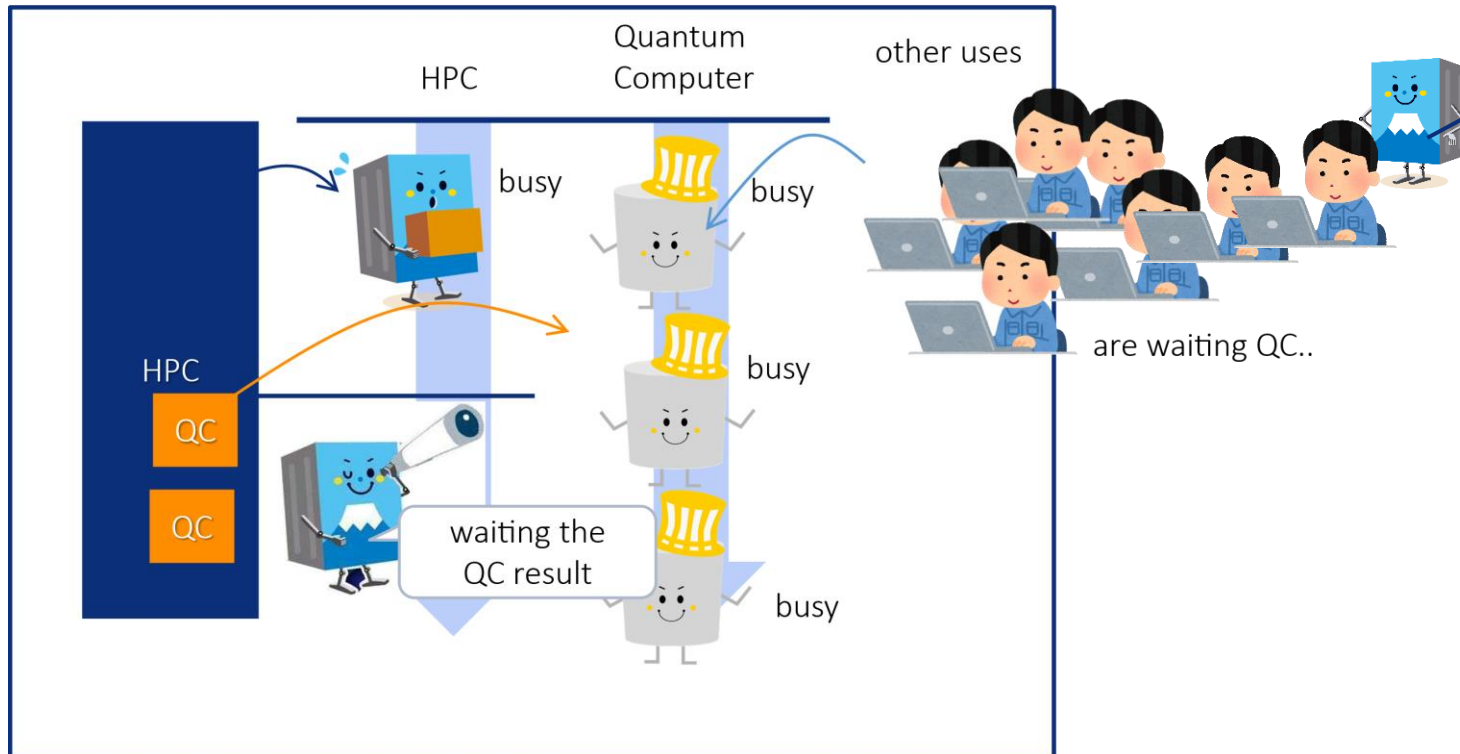
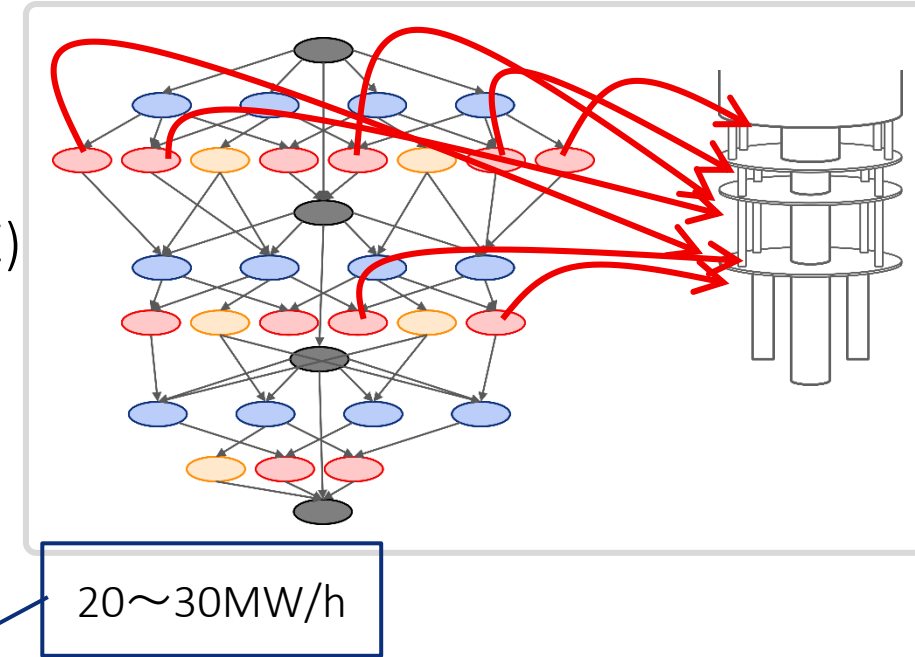
Why do we need the 1st Layer, Job-Scheduler Based Workflow?

- Avoid the waiting times in quantum computer and supercomputer
 - Uses separate quantum kernels and HPC kernels in their applications
 - Define the order of their execution
- Job-schedulers in both computers may arrange jobs to maximize the throughput in their systems.
- A job may have to wait the result from another job, but the waiting time does not consume any computational resource



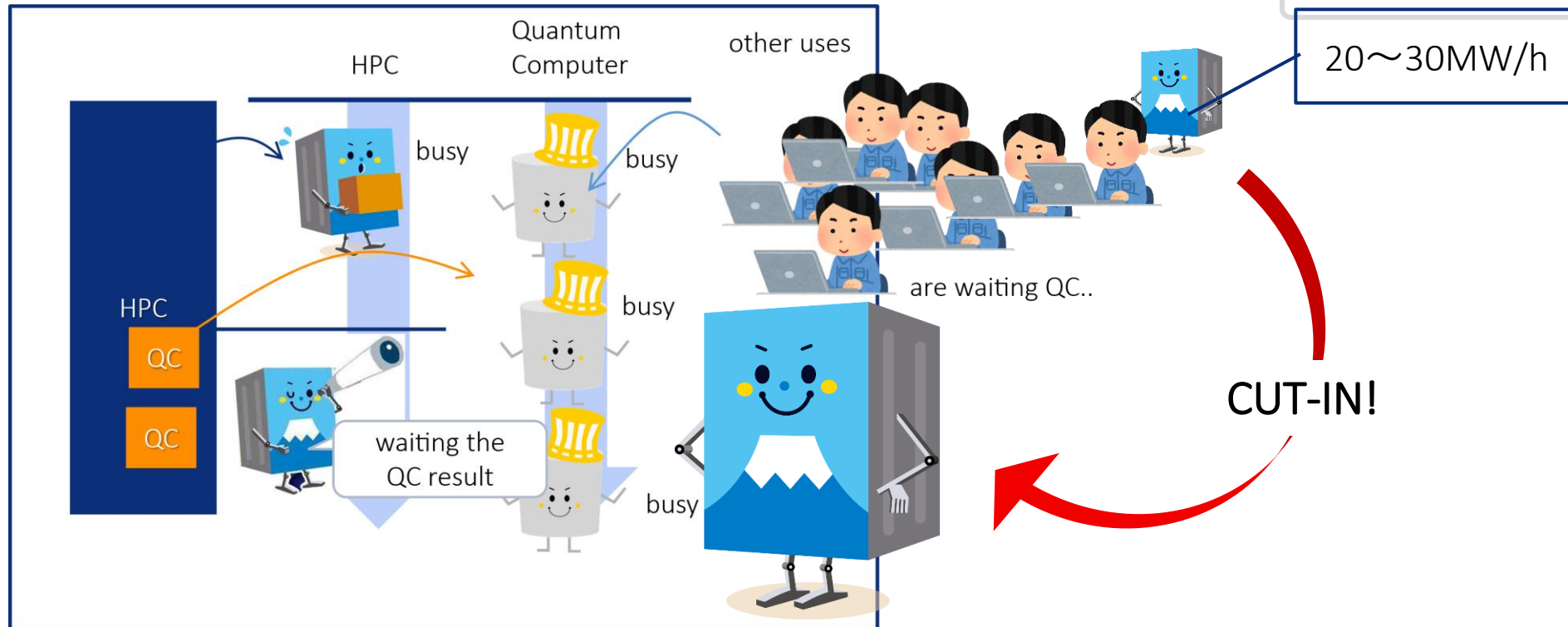
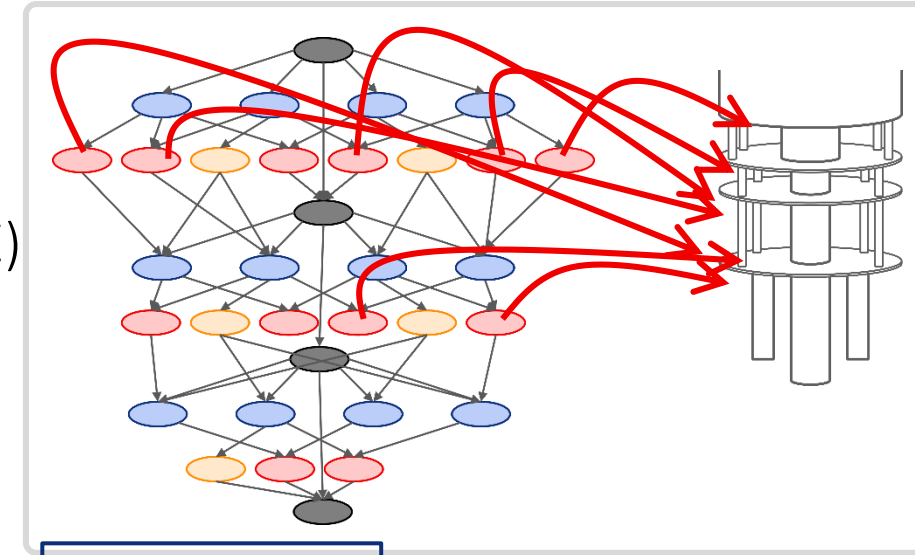
Why do we need 2nd Layer, task-based programming

- Tightly coupled quantum and HPC kernels in a single job
 - Difficult to separate quantum and HPC kernels completely
- Multiple quantum kernel calls from HPC in a short time (ex. VQE)
 - Difficult to wait each of the quantum circuit executions
- (Still) We must avoid the waiting time in HPC, especially for large scale HPC jobs



Why do we need 2nd Layer, task-based programming

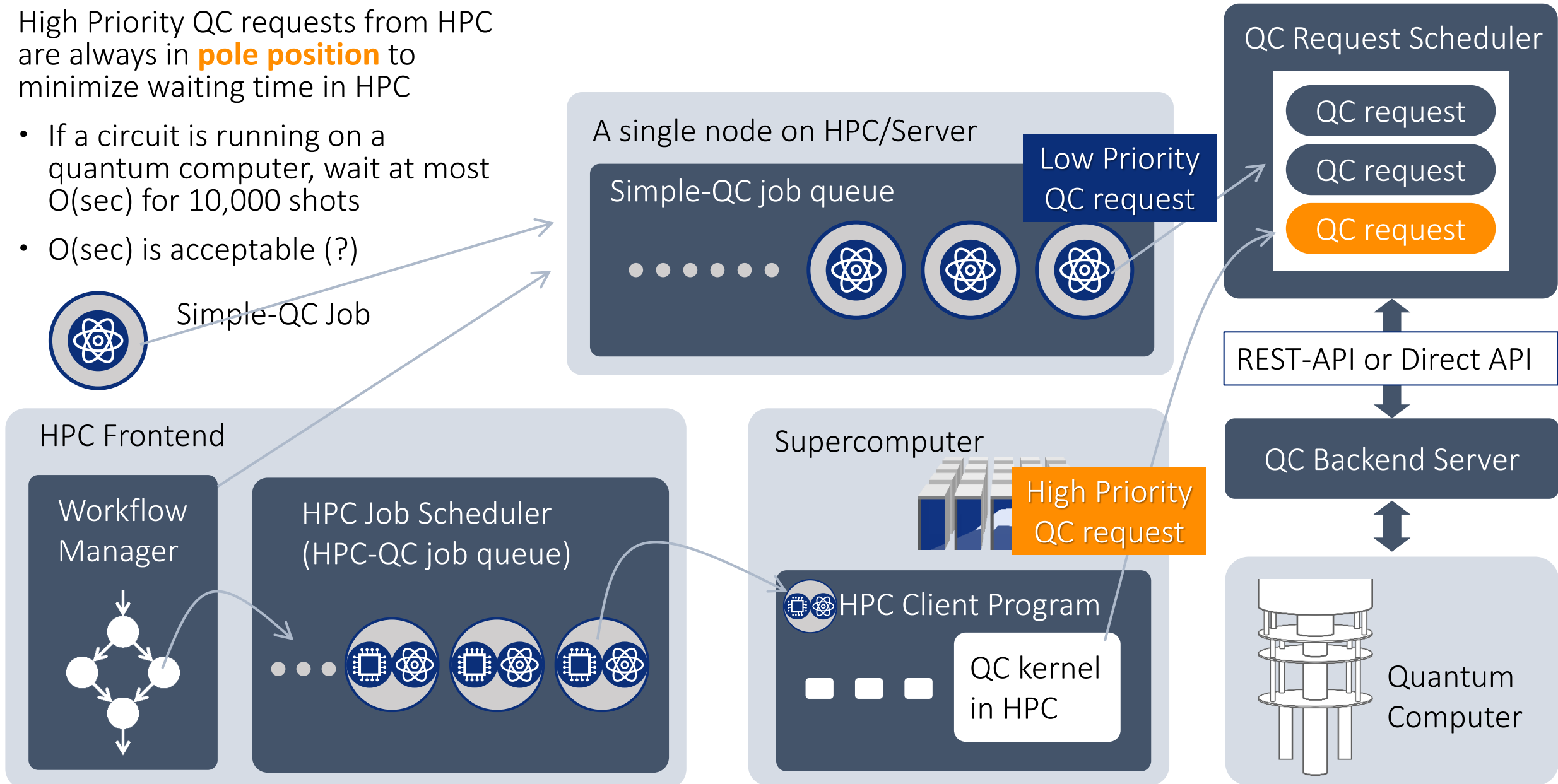
- Tightly coupled quantum and HPC kernels in a single job
 - Difficult to separate quantum and HPC kernels completely
- Multiple quantum kernel calls from HPC in a short time (ex. VQE)
 - Difficult to wait each of the quantum circuit executions
- (Still) We must avoid the waiting time in HPC, especially for large scale HPC jobs



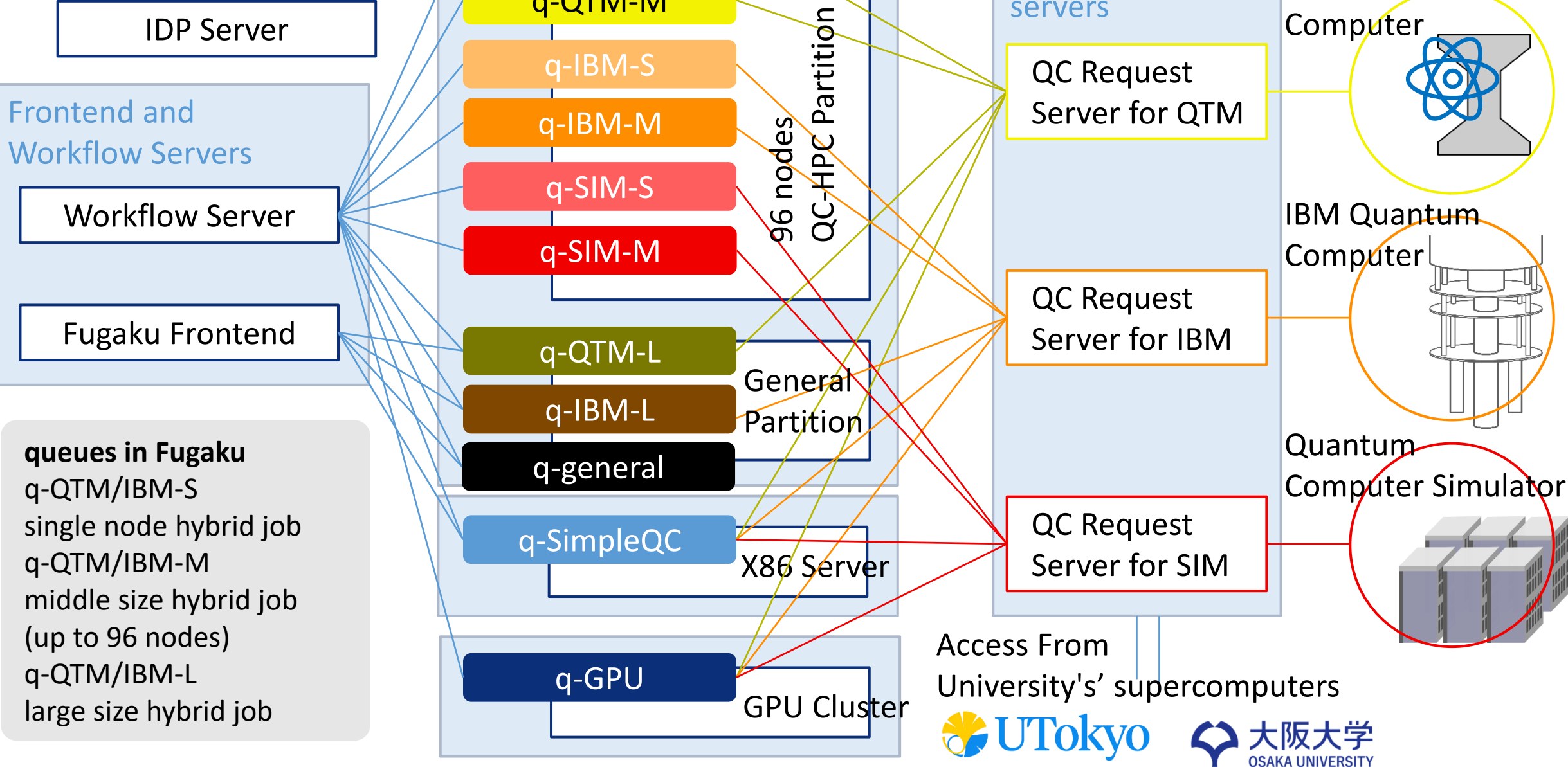
Execution Environment

High Priority QC requests from HPC are always in **pole position** to minimize waiting time in HPC

- If a circuit is running on a quantum computer, wait at most $O(\text{sec})$ for 10,000 shots
- $O(\text{sec})$ is acceptable (?)

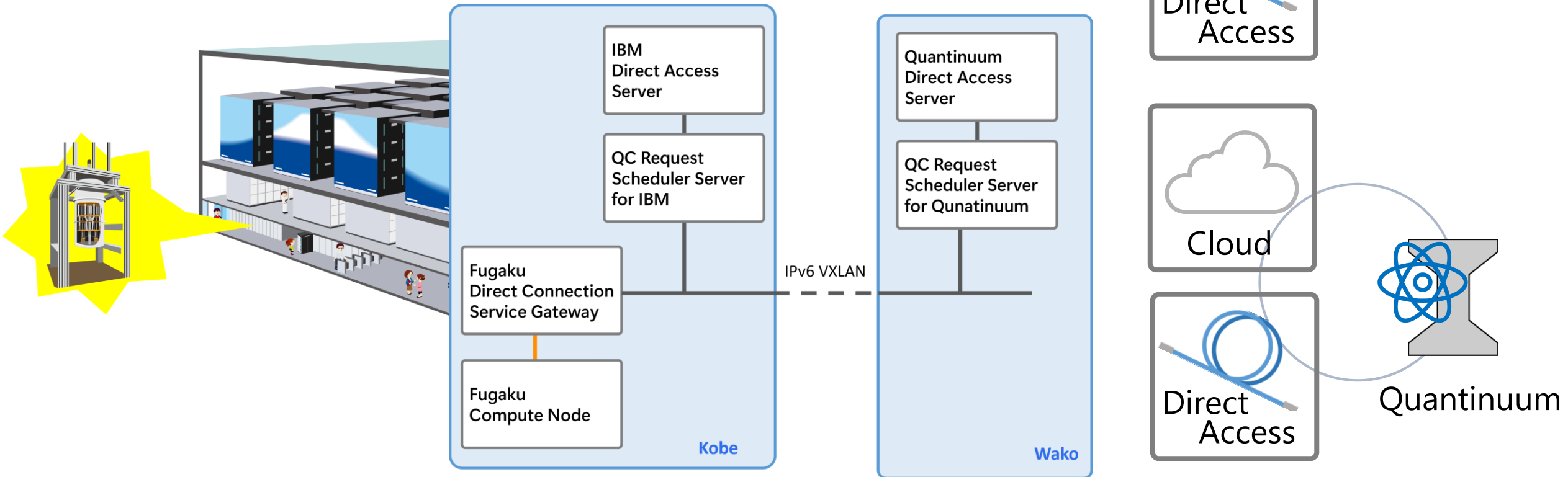


System Configuration



Network and Configurations

- Quantum Computer Backend Server
 - Cloud: Existing cloud-based service. Servers are in *somewhere in the world*.
 - Direct Access: Faster connection via direct local connection or virtual local connection. The server for IBM will be installed in Kobe, the server for Quantinuum is installed in Wako.



Schedule

