RIKEN FUTURE COMPUTING PLAN

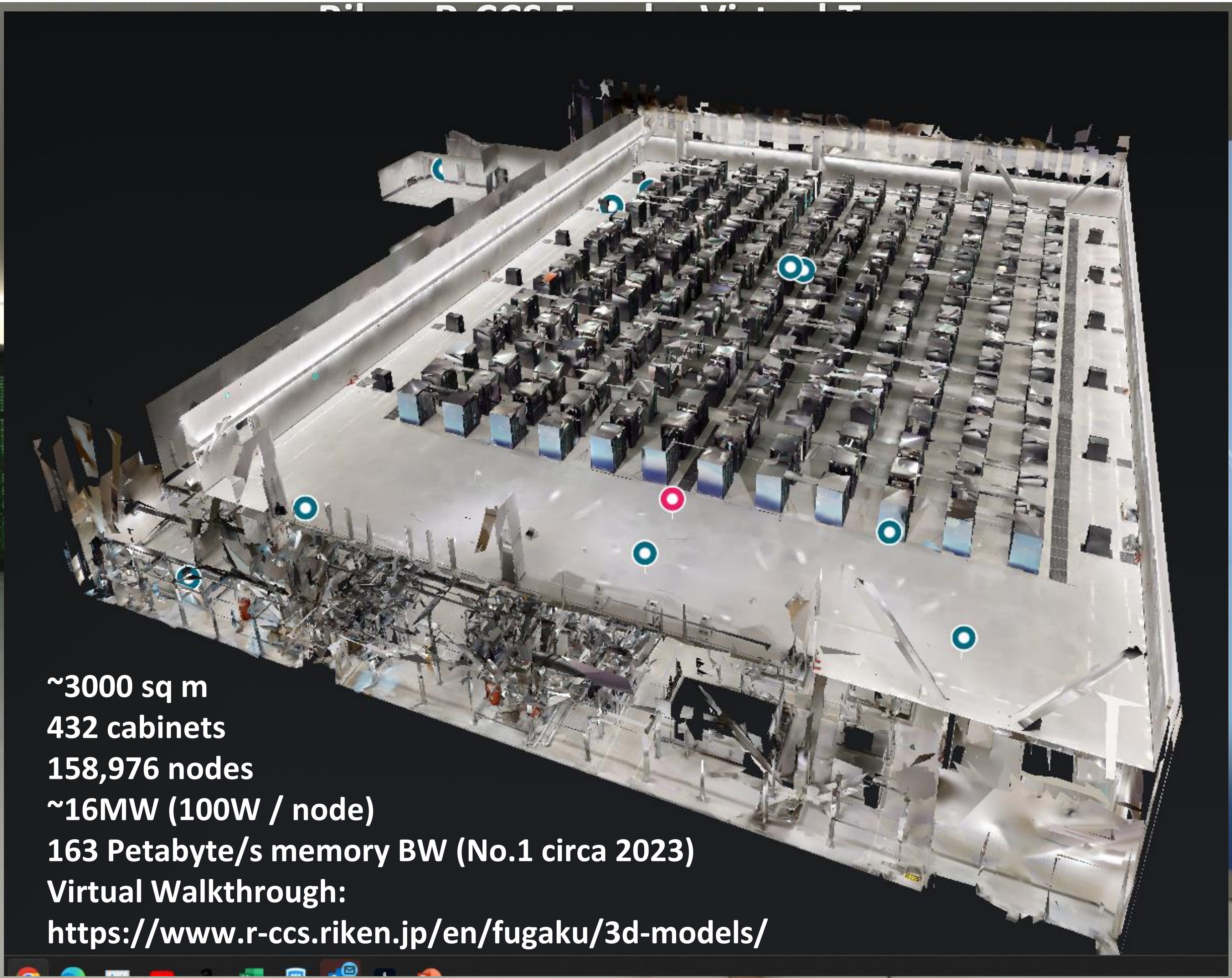
Fugaku Towards AI “Zettascale” FugakuNEXT (@40MW) (long version)

Satoshi Matsuoka, Director Riken R-CCS

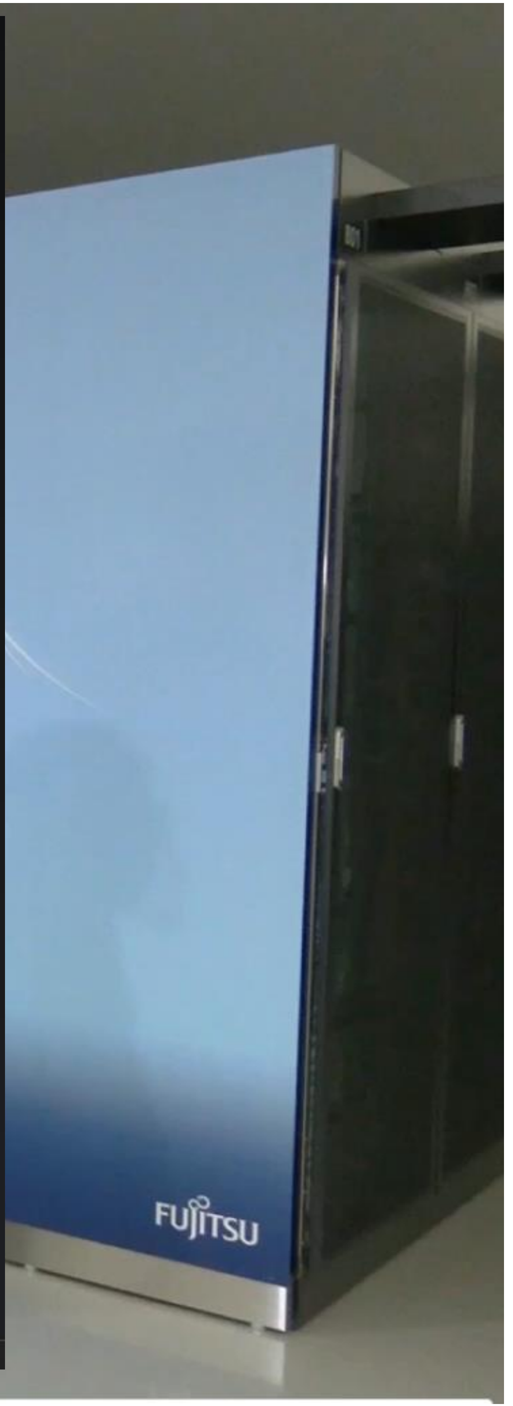
DoE ASCAC Presentation

Bethesda, MD, USA, Jan 17th 2025

RIKEN R-CCS



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



Organization of RIKEN R-CCS as of 1st April 2024

RIKEN

R-CCS

We are recruiting researchers, postdocs, interns, …

R-CCS Director
S. Matsuoka

R-CCS Deputy Director
Science of Computing
K. Nakajima

**R-CCS Deputy Director
Science by Computing
Y. Sugita (April 2024)**

Science of Computing (Computer Science)

- Advanced Processor Architectures
  K. Sano
- Large-scale Parallel Numerical Computing Technology
  T. Imamura
- Next Generation High Performance Architecture
  M. Kondo
- High Performance Big Data
  K. Sato
- High Performance AI Systems
  M. Wahib
- Supercomputing Performance
  J. Domke
- **Large-Scale Digital Twin
  H. Yamaguchi
  (April 2024)**
- **Cloud & Security
  A. Takefusa
  (Sep 2024)**

Science by Computing (Computational Science)

- Field Theory
  Y. Aoki
- Discrete Event Simulation
  N. Ito
- Computational Molecular Science
  T. Nakajima
- Computational Materials Science
  S. Yunoki
- Computational Biophysics
  Y. Sugita
- Computational Climate Science
  H. Tomita
- Complex Phenomena Unified Simulation
  M. Tsubokura
- Data Assimilation
  T. Miyoshi
- Computational Structural Biology
  F. Tama
- Computational Disaster Mitigation & Reduction
  S. Oishi

Office of the Fugaku Society 5.0 initiative

- Office Director
  S. Matsuoka
- Office Deputy Director
  Y. Watanabe
- Office Coordinator
  H. Shirai

HPC and AI driven Drug Development Platform Division

- Division Director & Biomedical Computational Intelligence
  Y. Okuno
- Deputy Division Director & Medicinal Chemistry Applied AI
  T. Honma
- Molecular Design Computational Intelligence
  M. Ikeguchi
- AI driven Drug Discovery Collaborative
  Y. Okuno

（＊Now recruiting：Biomedical Computational Intelligence, Medicinal Chemistry Applied AI）

Quantum-HPC Hybrid Platform Division

- Division Director
  M. Sato
- **Quantum-HPC Hybrid Software Environment
  M. Tsuji
  (April 2024)**
- Quantum Computing Simulation
  N. Ito
- Quantum-HPC Hybrid Platform Operations
  S. Miura

（＊Now recruiting：Deputy Division Director）

AI for Science Platform Division (April 2024)

Considering strengthening the system at the start of the next medium- to long-term plan

- Division Director
  S. Matsuoka
- AI Development Computing Environment Operation Technologies
  S. Miura
- Advanced AI Device Development
  K. Sano
- Learning Optimization Platform Development
  M. Wahib
- Data Management Platform Development
  K. Sato
- Life and Medical Science Application Interface Platform Development
  Y. Sugita
- Material Science Application Interface Platform Development
  T. Nakajima

Operations and Computer Technologies

- Division Director
  F. Shoji
- Deputy Division Director & System Operations and Development
  Y. Iguchi
- Facility Operations and Development
  S. Miura
- Software Development Technology
  H. Murai
- **Data Interaction Technology Development
  T. Kai
  (March 2024)**
- Advanced Operation Technologies
  K. Yamamoto

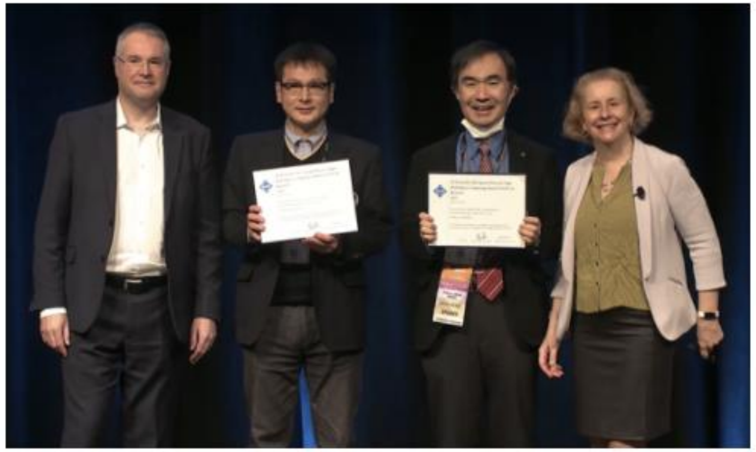
（＊ Now recruiting：System Operations and Development Unit UL）

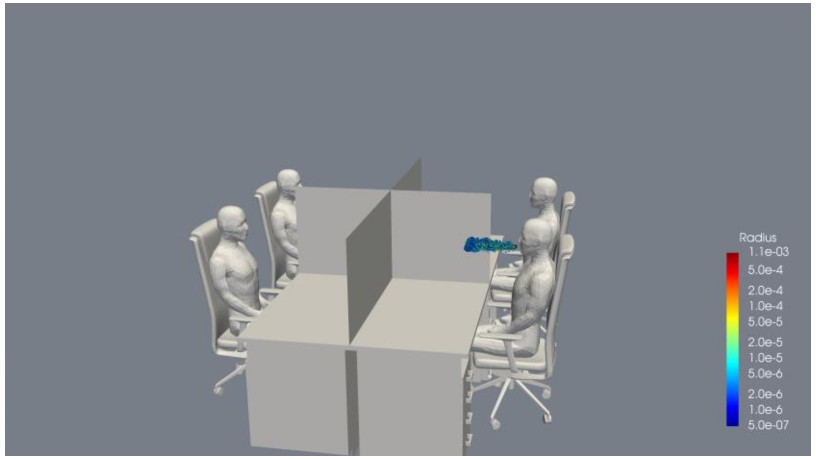
Major achievements of Fugaku

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

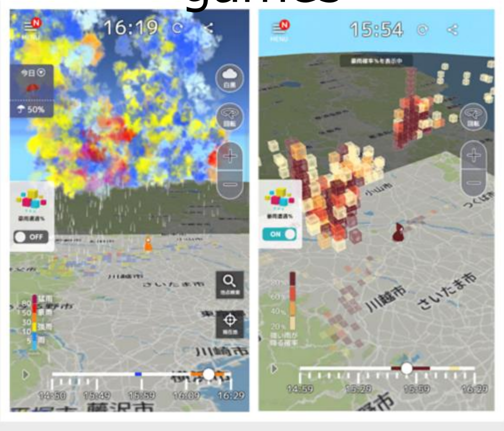
#1 in MLPerf HPC(Nov.2021-)

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





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



今回の実証実験で表示される「3D雨雲ウォッチ」アプリイメージ

The Gordon Bell Prize for Climate Modelling 2023

powered by FUGAKU

Finalists!

“Big Data Assimilation: Real-time 30-second-refresh Heavy Rain Forecast Using Fugaku During Tokyo Olympics and Paralympics”

The Gordon Bell Prize for Climate Modelling

Nominations will be selected based on their impact on climate modelling, and on wider society by applying high-performance computing to climate modelling applications. In 2023, the first year, three finalists have been selected.

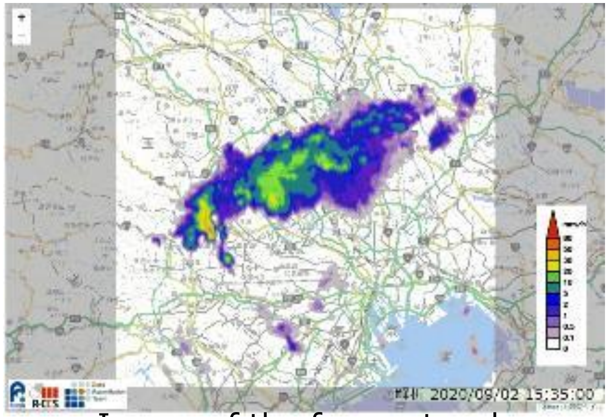


Image of the forecast web

Data Assimilation Research Team
Takemasa Miyoshi, Team Leader

Computational Climate Science Research Team
Hirofumi Tomita, Team Leader

2013: Start with "K computer"
2021: Achieve with "Fugaku"

The work presents a real-time 30-second-refresh numerical weather prediction (NWP), during the 2021 Tokyo Olympics and Paralympics. It revealed the effectiveness NWP for rapidly evolving convective rainstorms. This endeavor stands as a testament to the value of engaging advanced computational methodologies to advance understanding of intricate meteorological phenomena.

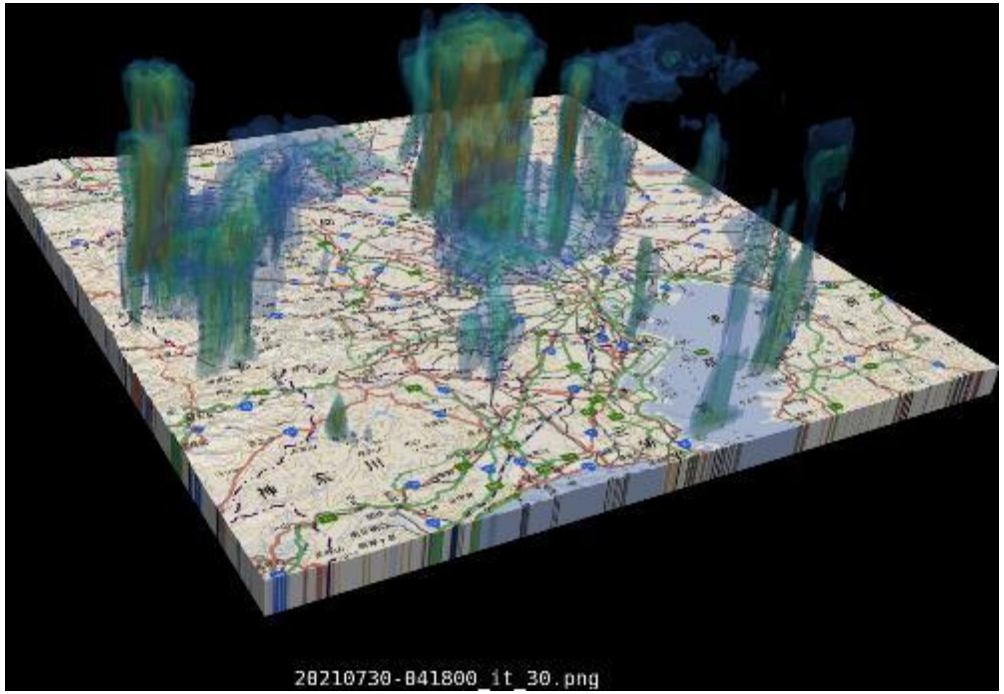
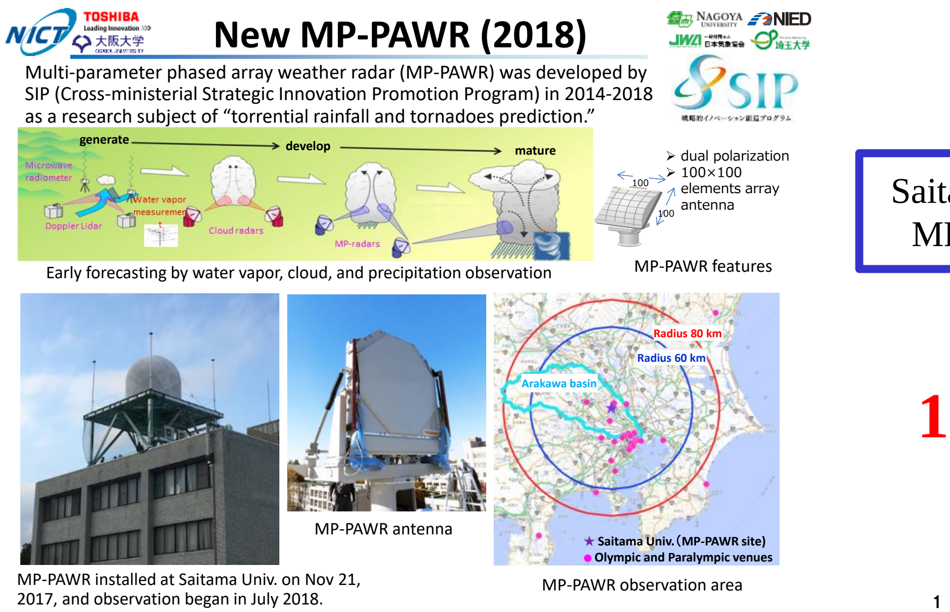


Figure: Bird's-eye view of 15-minute forecast rain distributions at 04:33:00 UTC, July 30, 2021, initialized at 04:18:00 UTC. Colors represent rain intensity. Vertical scale is stretched by three times. Map data courtesy of the Geospatial Information Authority of Japan

Real-time data transfer & data assimilation for Tokyo Olympics 2020 and **Osaka Expo 2025 (new!) – Weather is a huge business now --**



NICT
Saitama Univ.
TOSHIBA

Saitama Univ. MP-PAWR → NICT ds01

JIT-DT
106 MB per obs.
in 3 seconds

data monitor
auto-restarter

Fugaku login1 → SCALE-LETKF → weather.riken.jp

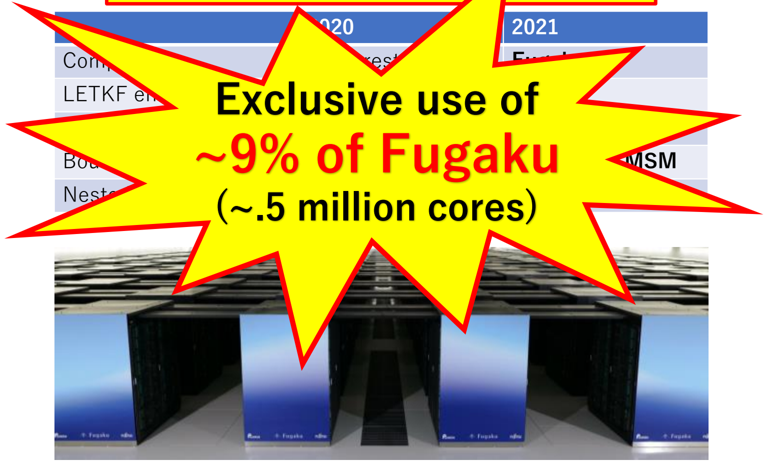
webpage

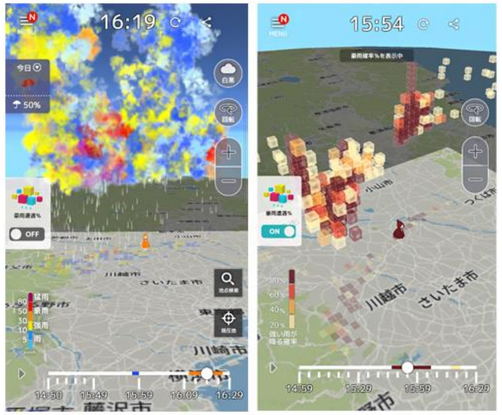
MTI
Amazon AWS

smartphone

https://weather.riken.jp/

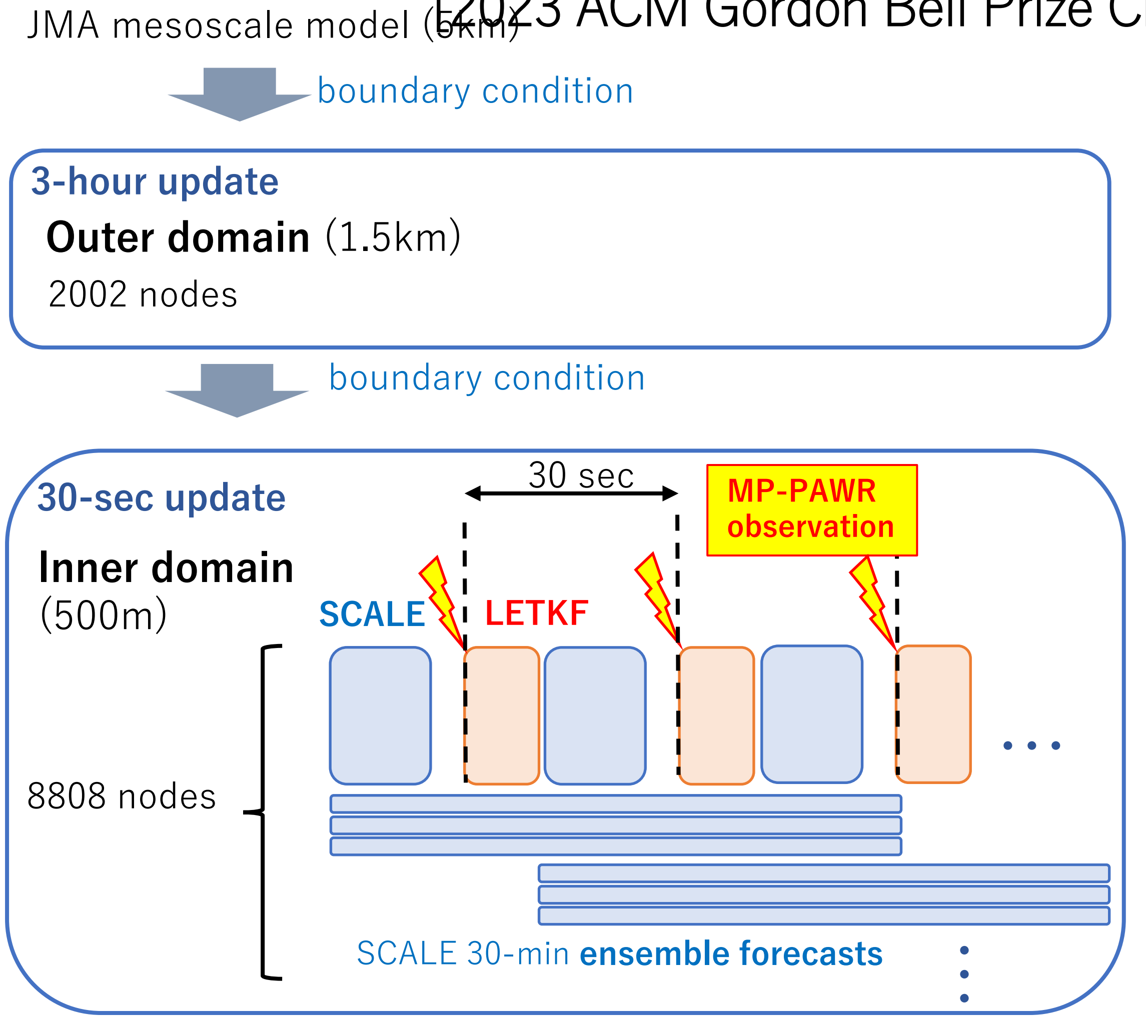
Real-time experiments in 2021
- July 20-August 8 (Olympic)
- August 24-September 5 (Paralympic)

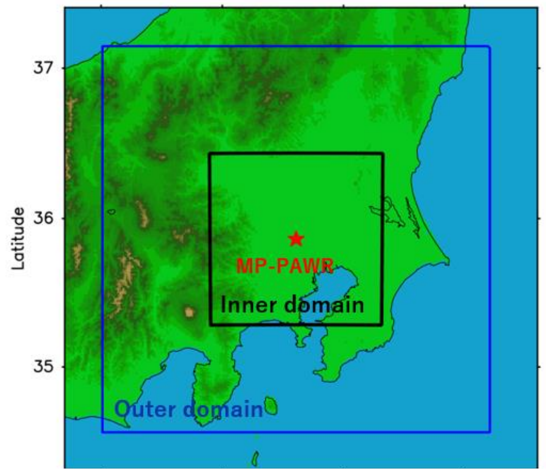




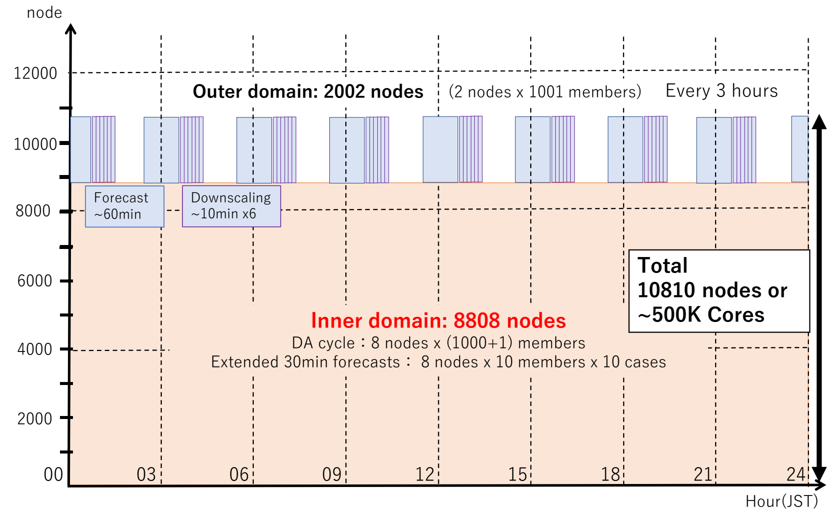
Real-time workflow of 30 sec, 500m weather forecast for 2020 Tokyo Olympics

[2023 ACM Gordon Bell Prize Climate Prize Finalist]





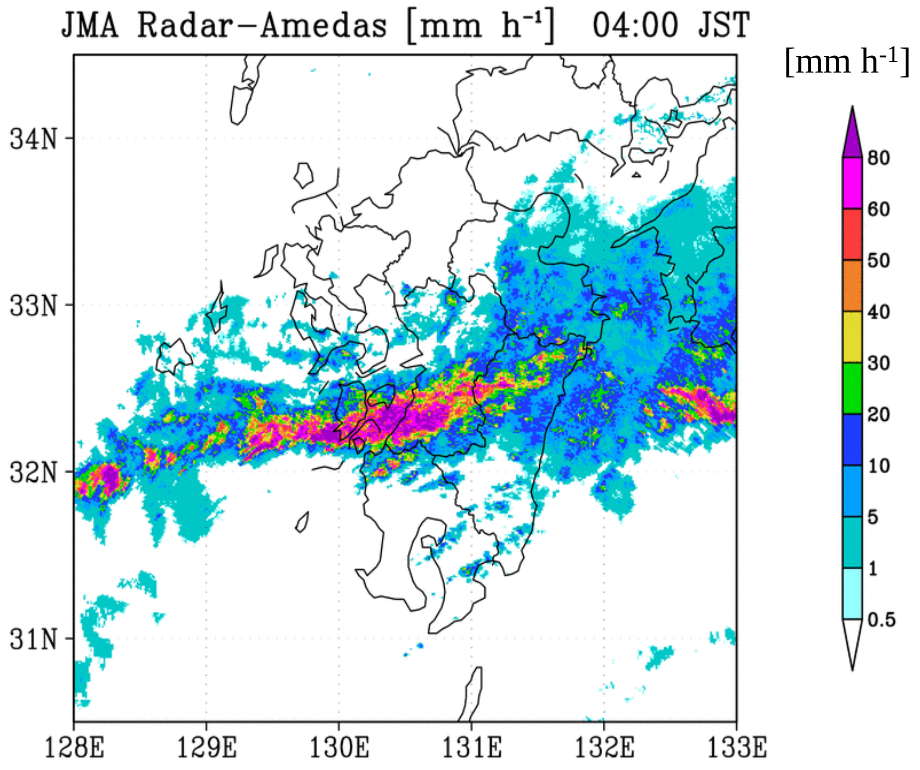
Real-time job scheduling of 1/2 million cores



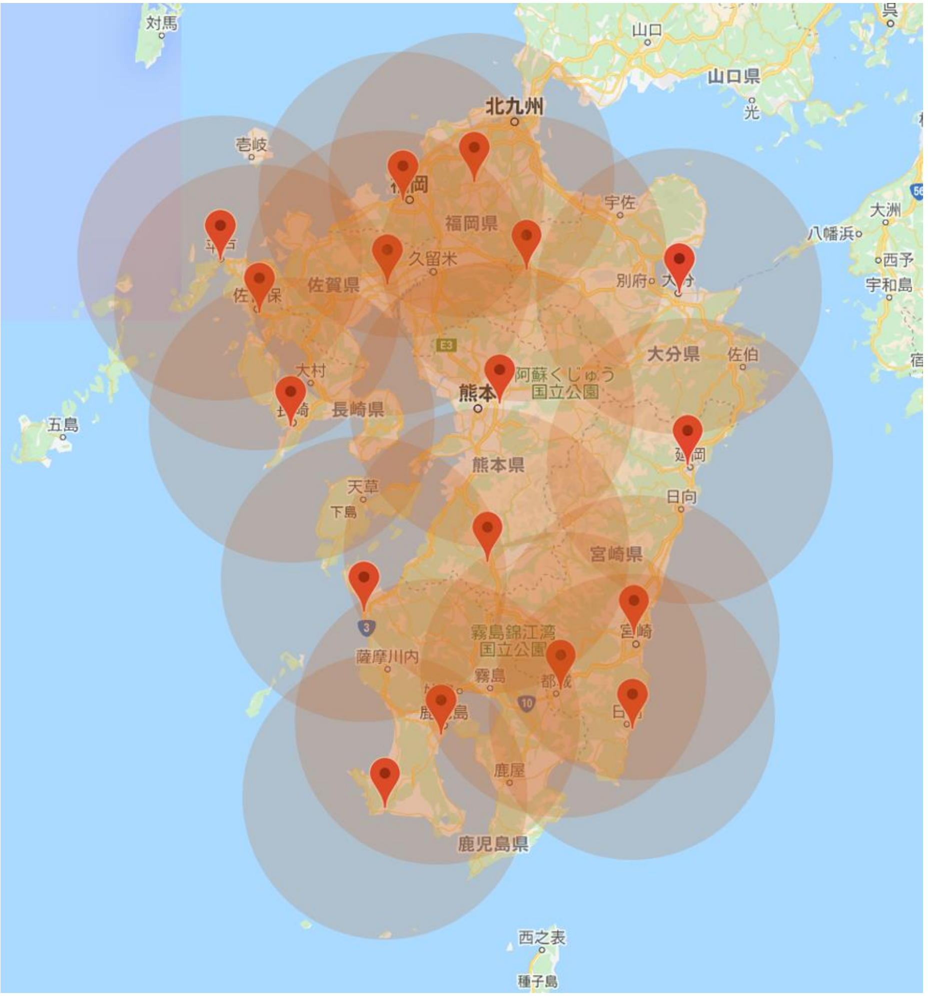
What if we had many PAWRs?

An Observing System Simulation Experiment (OSSE)

July 2020 heavy rain



A virtual PAWR network



Maejima et al. (2022, SOLA, doi:10.2151/sola2022-005)

Fugaku Siblings Preventing Natural Disasters

- Japan Meteorological Agency(JMA) utilized large scale externa supercomputer for the first time to simulate torrential rain band causing catastrophic damages
- Critical research advances were made such that they acquired a smaller version of Fugaku (15PF x 2) as a research SC, separate from their production SC for forecast
- JMA Started production 12-hour ahead torrential rain forecast with its twin Fugaku-compatible machines from May 2024

図１　線状降水帯予測スーパーコンピュータ

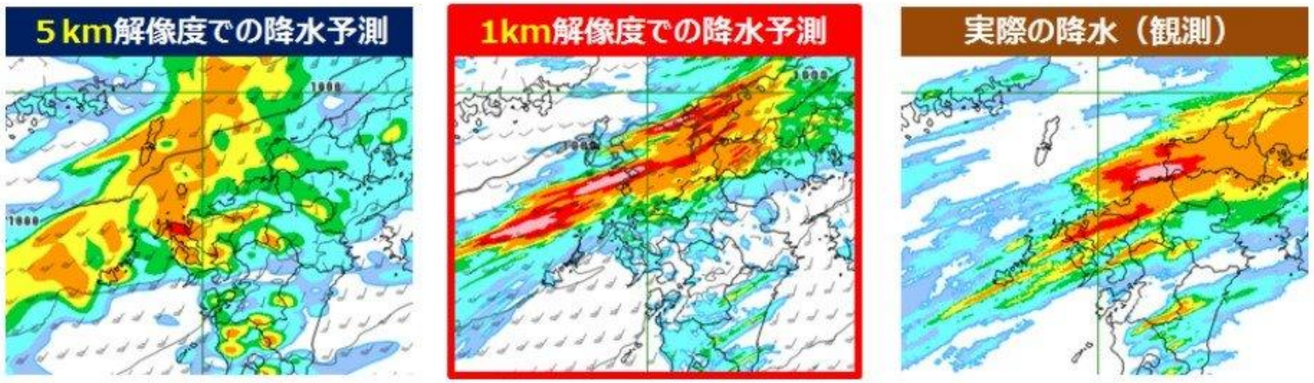
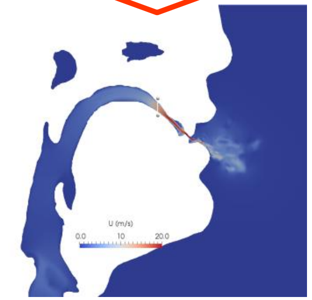
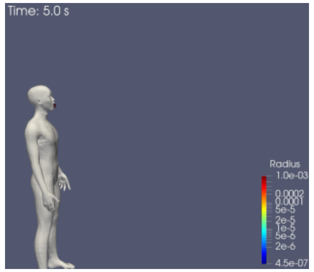


図２　水平解像度１kmに高解像度化した局地モデルのイメージ

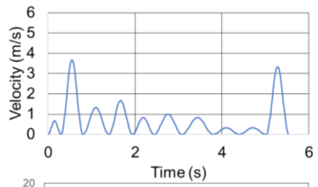
Sub-Task C :Indoor-Environment Design Robust for the Infectious Diseases

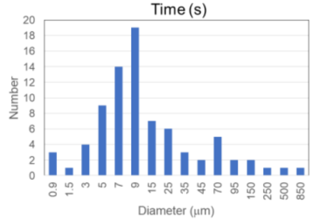
Generation of droplet/aerosol inside human body

Condition of droplet/aerosol generation (breathing, speaking, coughing, sneezing…)

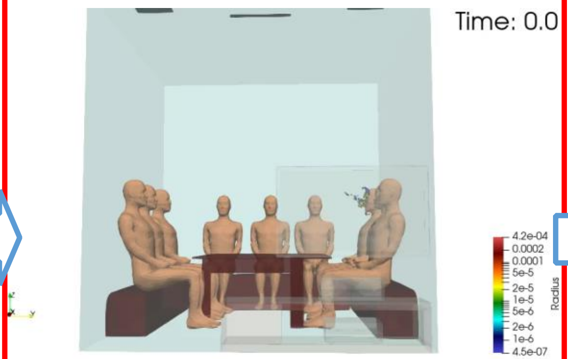
Breath flow rate, droplet size distribution



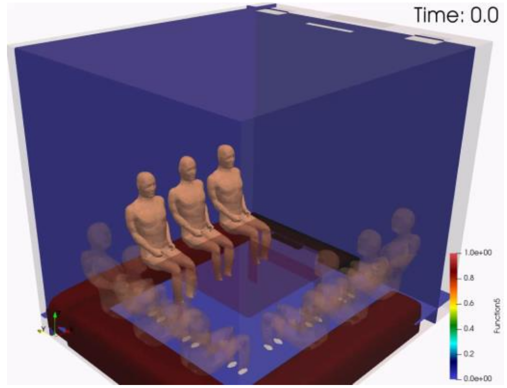


Droplet/aerosol dispersion in indoor environments

Indoor environment and human allocation



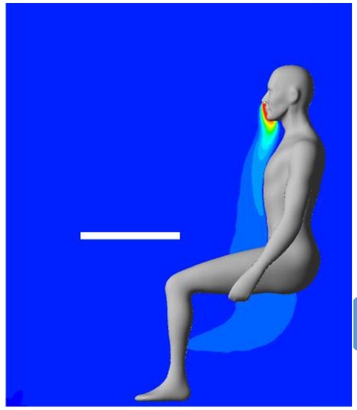
Coupling simulation of droplet/aerosol and indoor flow



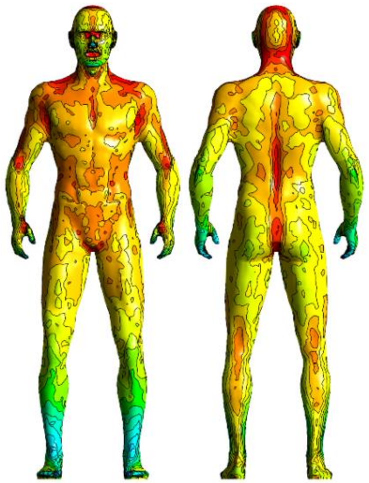
Indoor environment evaluation based on HPC simulation

Numerical human body

Biological information of an at-risk person



Precise reproduction of human breathing

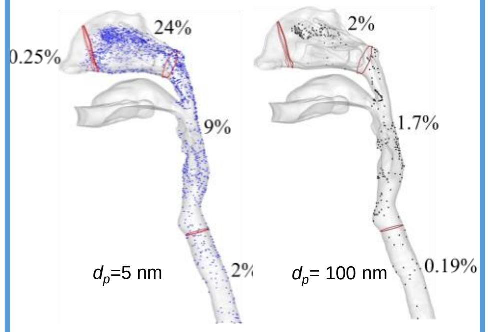


Precise reproduction of body temperature

Numerical respiratory tract

Biological information of an at-risk person

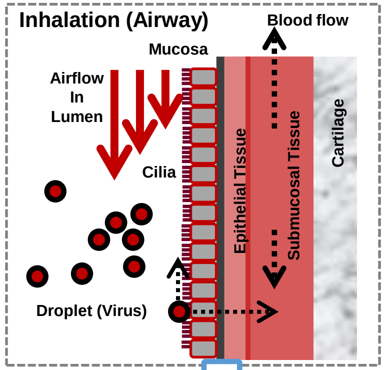
Reproduction of nasal/oral cavity and respiratory tract



Prediction of deposit distribution of droplet/aerosol on the airway surface and its dependence of droplet size.

Infection risk assessment based on the bio-regulation model

Biological information of an at-risk person and target virus



Bioregulation

(Host cells, Pathogen, Adaptive Immune System)

$$\frac{dT_T}{dt} = -\beta_T T_T V - \phi F T_T + \xi R \frac{dR}{dt}$$ (Target Cells)

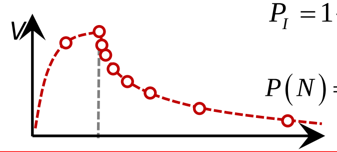
$$\frac{dI}{dt} = \beta_T T_T V - \kappa_F I F - \kappa_E I T_C - \delta_X I$$ (Infected Cells)

$$\frac{dV}{dt} = \beta_E I - \delta_V V - \kappa_V V A$$ (Virus)

$$\frac{dF}{dt} = \beta_F I - \kappa_A F$$ (Interferon)

$$\frac{dT_H}{dt} = \left[\frac{\pi_{H2} D_M}{\pi_{H2} + D_M}\right](1 - T_H / K_H) - \left[\frac{\delta_{H2} D_M}{\delta_{H2} + D_M}\right] T_H$$ (Helper T Cells)

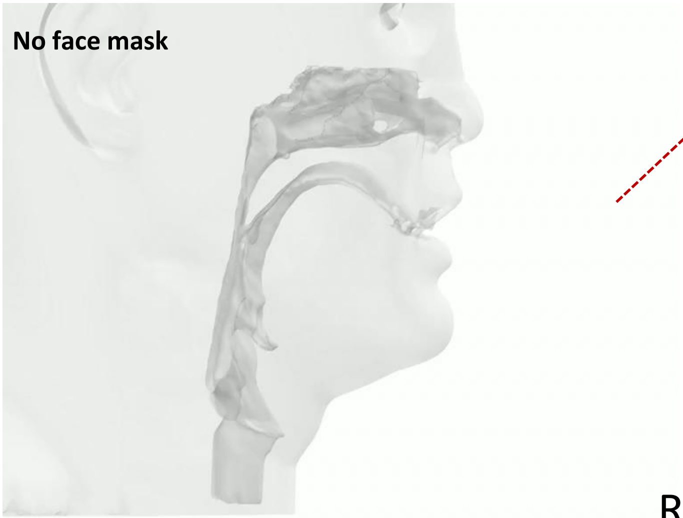
Quantitative evaluation of infection risk

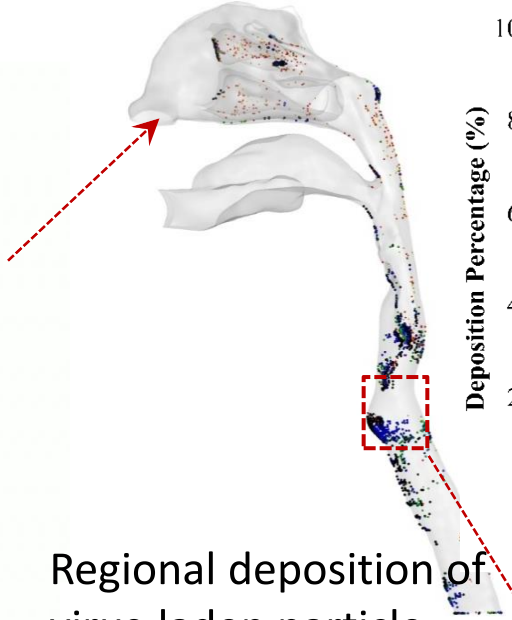


$$P_I = 1 - \exp\left(-\frac{Iqpt}{Q}\right)$$

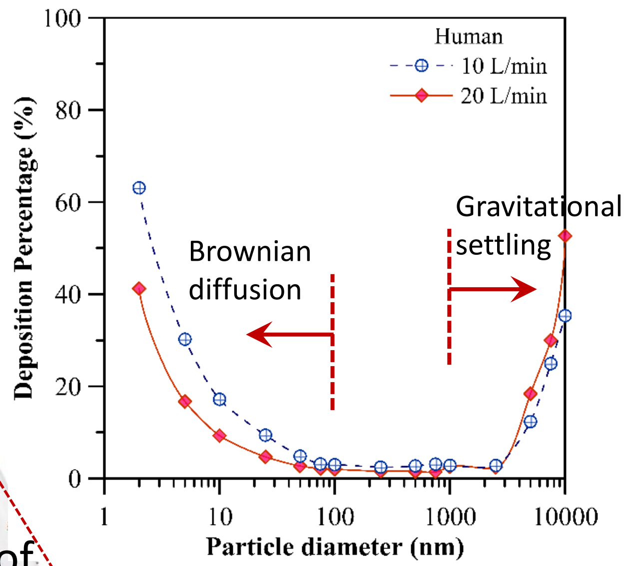
$$P(N) = 1 - \exp\left(-\frac{N}{N_0}\right)$$

Host cell dynamics coupled with numerical respiratory tract

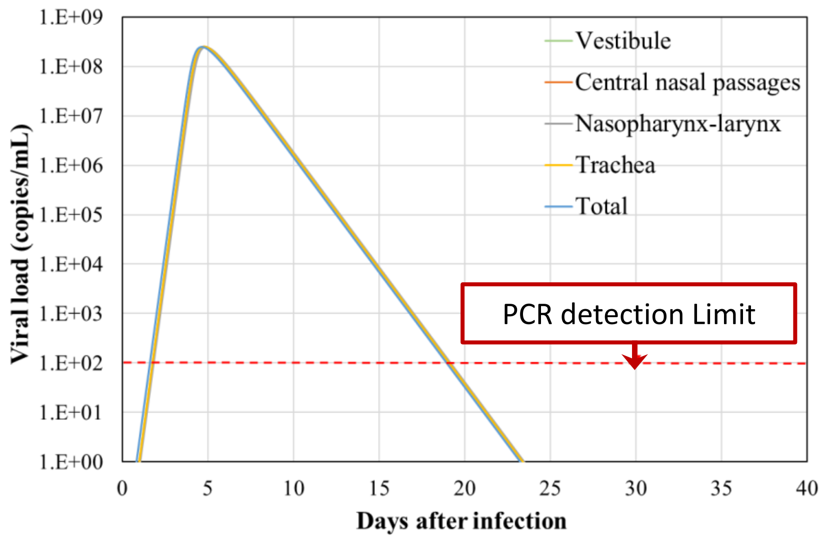




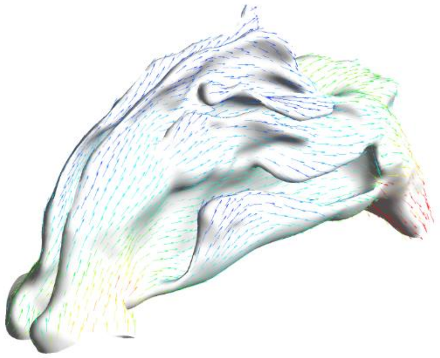
Regional deposition of virus-laden particle



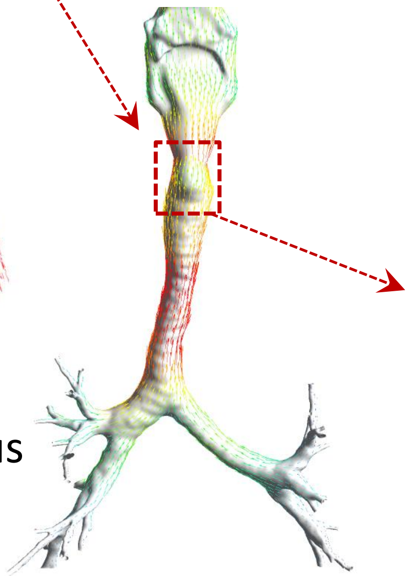
Deposition fraction

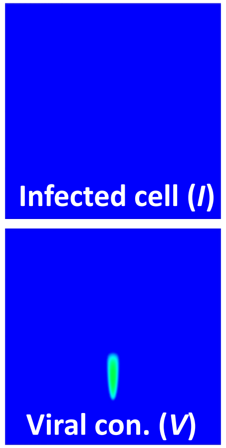


Resultant viral replication



Mucociliary/mucus transport



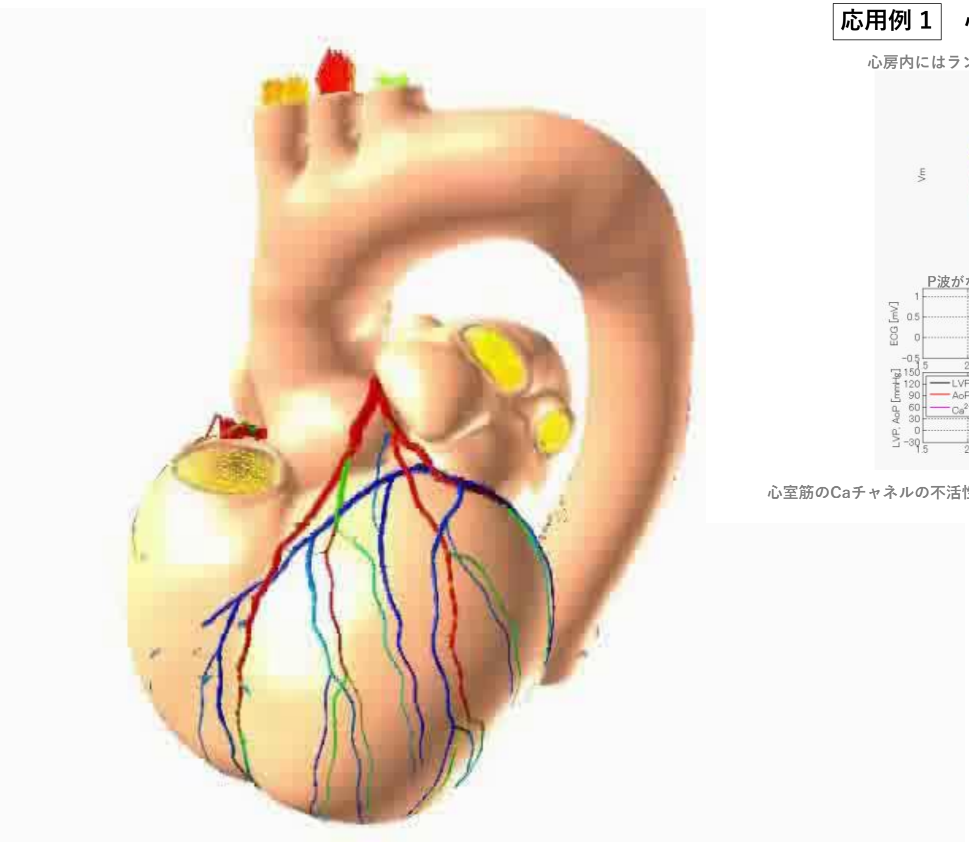


Bioregulation – Host Cell Dynamics

(Host cells, Pathogen, Adaptive Immune System)

$$\begin{cases} \dfrac{dT(t)}{dt} = -\beta T(t)V(t) \\ \dfrac{dI(t)}{dt} = \beta T(t)V(t) - \delta I(t) \\ \dfrac{dV(t)}{dt} = pI(t) - cV(t) \end{cases}$$

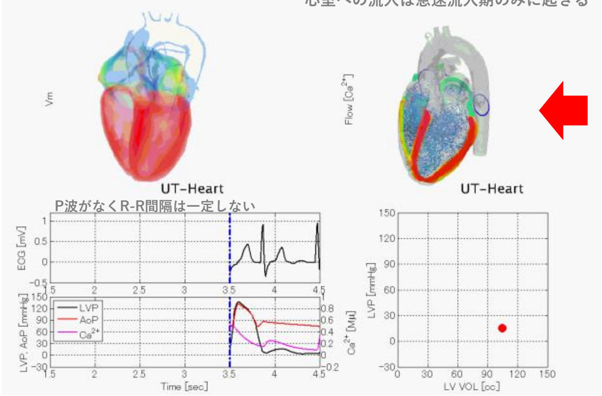
UT-Heart: "Personalized" Precise Heart Digital Twin Platform



応用例１　心房細動のシミュレーション

心房内にはランダムな興奮(re-entry)

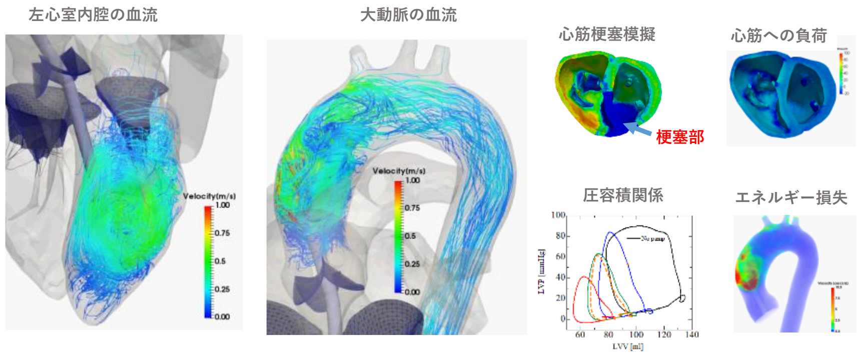
心房内のCa濃度は低く細かく振動
心室への流入は急速流入期のみに起きる



薬剤や治療機器の効果を計算機上で自在に試すことができる

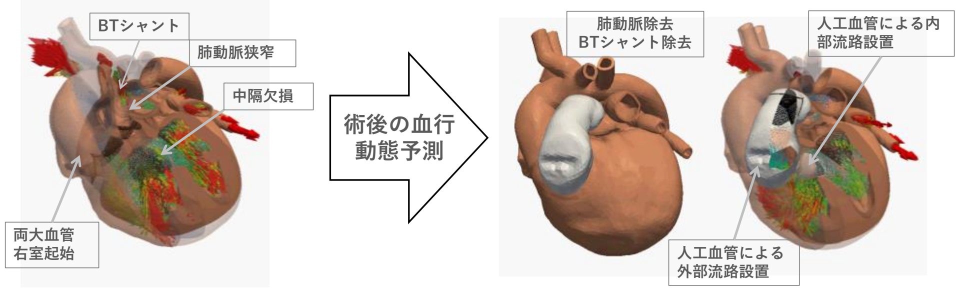
心室筋のCaチャネルの不活性化中に興奮波が到達⇒Ca放出量↓⇒収縮力↓⇒LV圧↓⇒大動脈弁開かないことがある

応用例２　補助循環用ポンプカテーテルIMPELLAの性能評価



任意の状態の心臓に対し各種医療機器の性能評価を計算機上で行うことが可能

応用例３　小児先天性心疾患の手術シミュレーション



各種術式による血行動態の改善を事前に計算機上で予測し、最適な手術を実施

現在、国立循環器病研究センター主導で多施設前向き臨床研究を実施中
次年度は医師主導治験を予定

NEW! Can create personalized digital twin of individual hearts via non-intrusive CT Scans via AI techniques

↓

Now being applied to real medical apps

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

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

The Fugaku potential returns are very strong

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

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

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

The Fugaku potential returns are very strong

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

2025 report for FugakuNEXT

Expect > 100x ROI

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

Fugaku researchers are addressing a broad set of SDG's

Projects in these areas include:

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

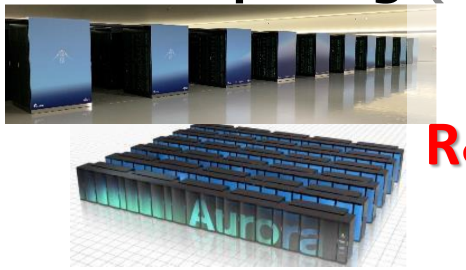
#4 Research Finding: Fugaku Is Focused On Creating Industrial Economic Growth

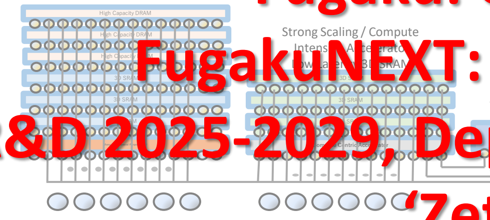
By directly supporting industry with a strong outreach program

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

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

- **Science of High Performance Computing** (towards ‘Zettascale’)



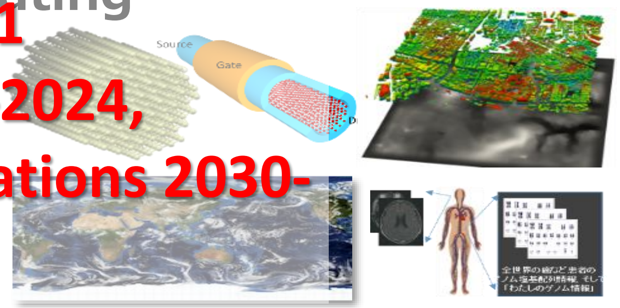


Fugaku: Current until 2030~2031

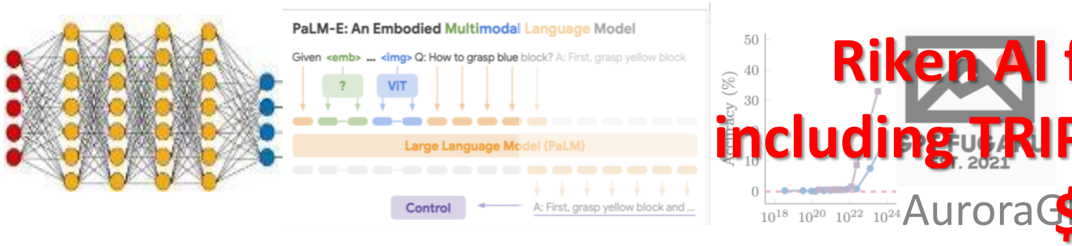
FugakuNEXT: Feasibility Study 2022-2024, R&D 2025-2029, Deployment ~2029, Operations 2030-

‘Zettascale’ @ 40MW

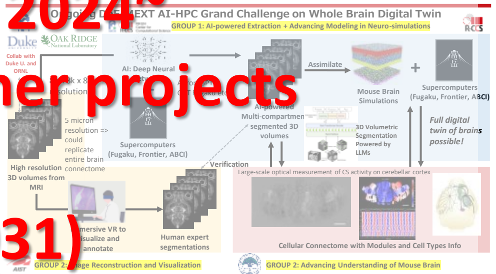
- **Science by High Performance Computing**

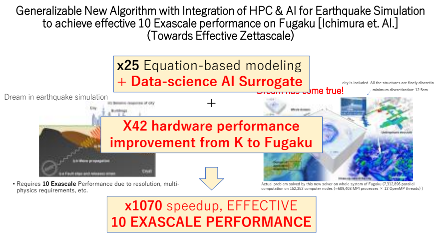


- **Science of High Performance AI**



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



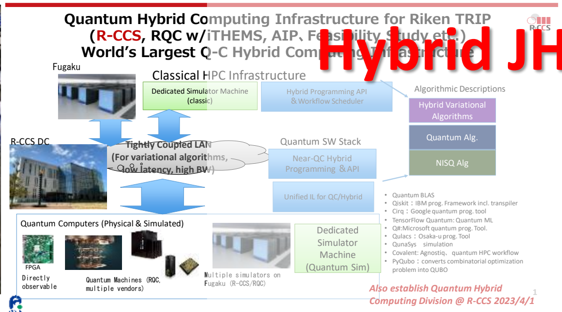


Riken AI for Science FY 2024~ including TRIP-AGIS and other projects

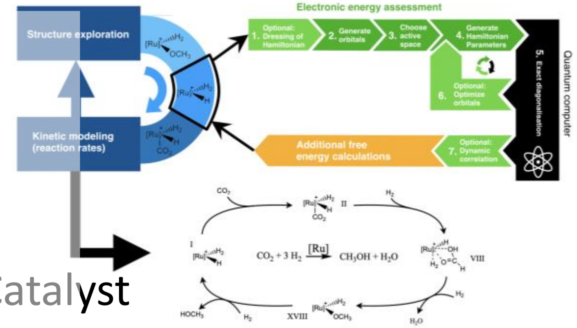
$X00 million

(TRIP-AGIS 2024~2031)

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



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



Hybrid JHPC-Quantum Infrastructure Project

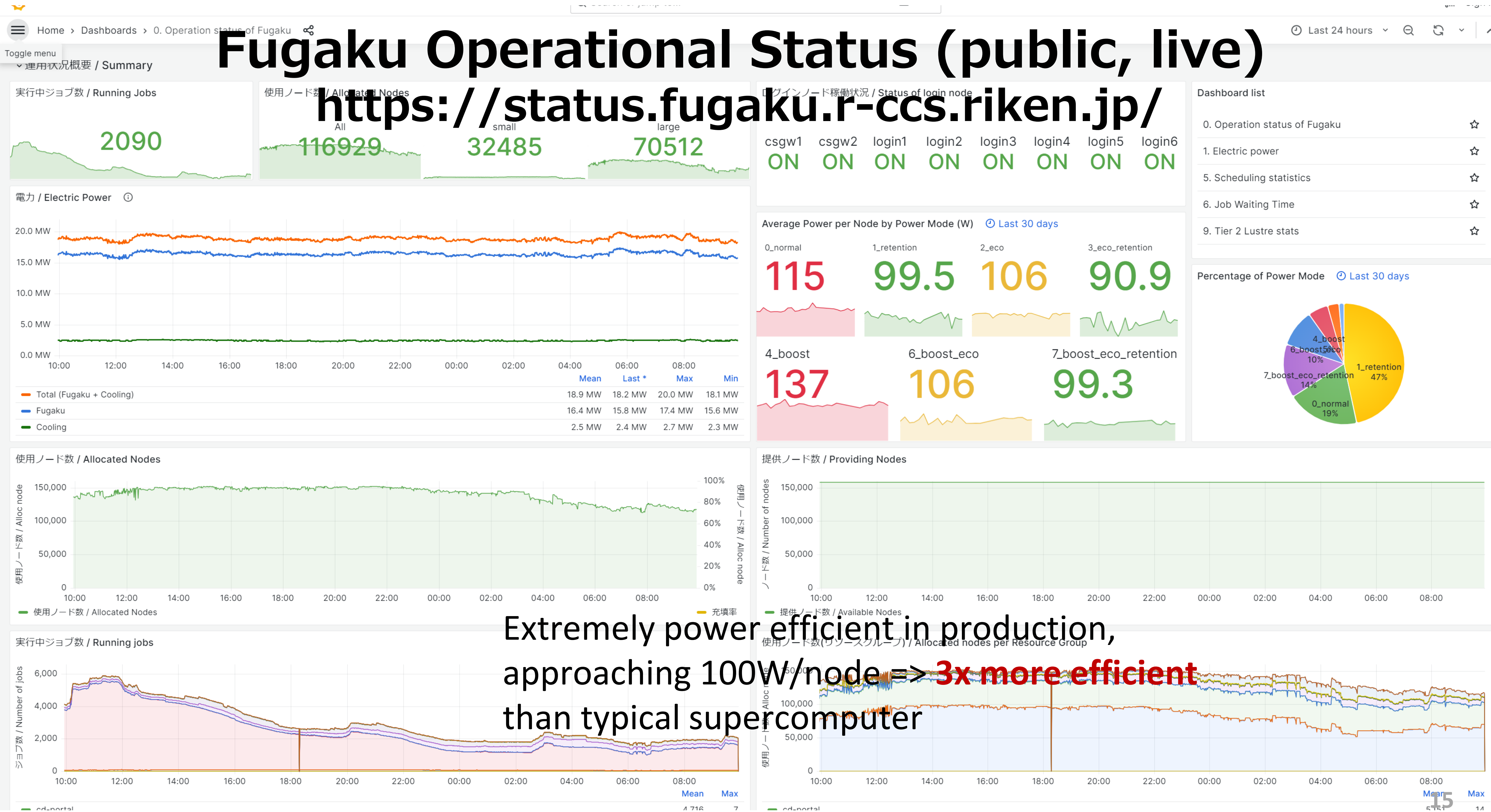
Deployment FY2023~2027

~$150 million+

(2023~2027)

Fugaku Operational Status (public, live)

https://status.fugaku.r-ccs.riken.jp/



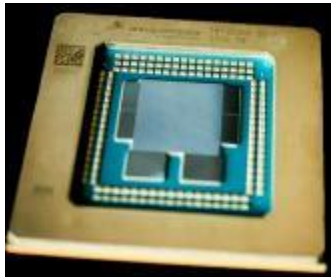
Extremely power efficient in production, approaching 100W/node => **3x more efficient** than typical supercomputer

January 2023 MoU Between AWS & R-CCS
Expanding the Scientific Platforms of Fugaku to the Cloud

Fujitsu-Riken A64FX HPC (2018) Arm+SVE CPU

High ISA (Arm+SVE) & Performance

Compatibility



AWS Graviton3/3E (2022) Arm+SVE CPU

Fugaku/FX1000

富岳

‘Cloudifying Fugaku”

“Cloud APIs on Fugaku”
Fugaku as part of cloud infra
e.g. Support S3 protocol (done)

Amazon EC2
C7g/C7gn instance

Riken R-CCS SC

‘Fugaku-fying the Cloud’

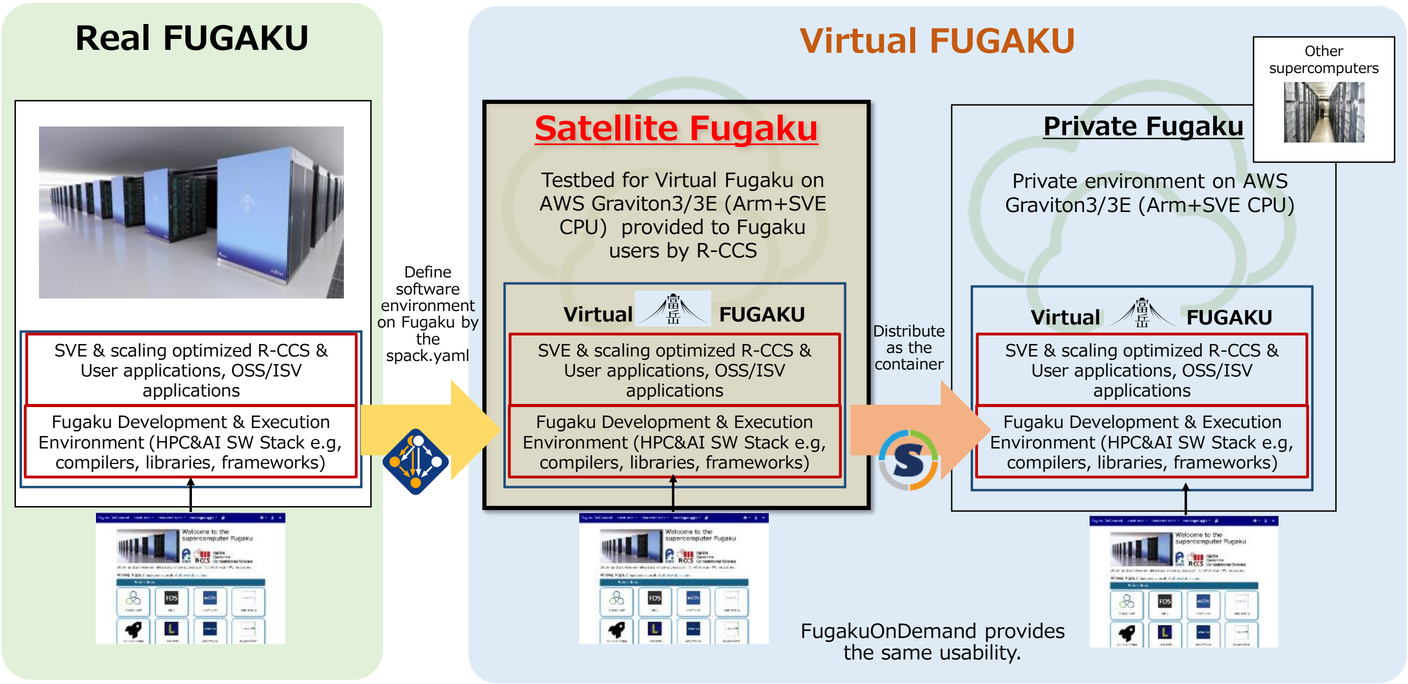
“Virtual Fugaku”
Implementing Fugaku Applications and Software Environment on AWS

aws

Virtualizing the Domain Specific Platform to utilize both
E.g. Companies develop methods using massive Fugaku Resource, production run on AWS, allow immediate propagation of latest research results onto production

Overview of 'Virtual Fugaku' Ver.1 (release Aug. 8, 2024)

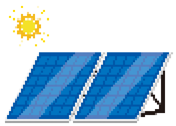
- **Two environments targeted at AWS Graviton CPUs:**
  - **Satellite Fugaku**: A test environment for 'Virtual Fugaku' (for Fugaku users).
  - **Private Fugaku**: A Singularity container for AWS users.
- **Both environments share the same software configuration (defined and containerized by SPACK).**
- **Basis for fully vendor–independent ready-made OSS stack for HPC/AI**



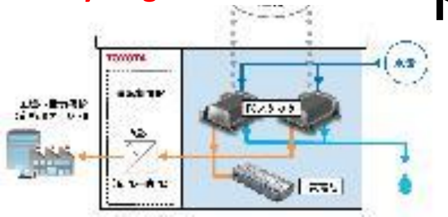
Target Study of Carbon Neutralization for Fugaku-next and A sustainable HPC center operation

(1) Use of Renewable Energy

DC's own operations can respond to future energy price hikes.



Hydrogen and Biofuels



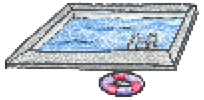
Effect: Stable reuse of electrical energy

Hydrogen
biomass

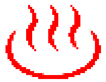
Carbon Neutral
Procurement of Electricity

Carbon neutral
In-House Generation of Electricity

(3) Reusing Waste heat

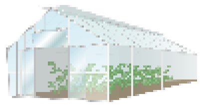


(swimming) pool

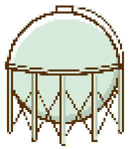


bathing facilities

https://datacenterfrontier.com/waste-heat-utilization-data-center-industry/



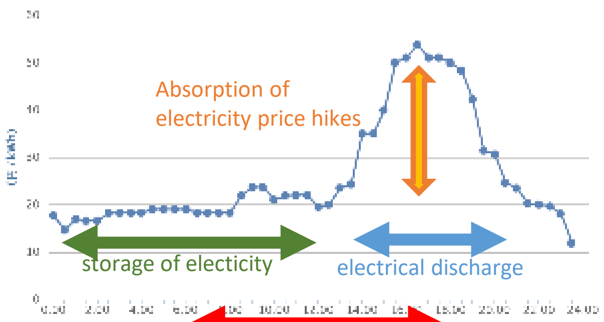
greenhouse



Heat for liquefied hydrogen naturalization

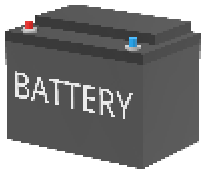
(2) Responding to electricity price fluctuations through the use of renewable energy

Market Price of Electricity



Time when sunlight is available

Introduction of large storage batteries



Supercomputer Simulation
Return of results to society

Waste heat

Sustainable power

(4) Energy-efficient supercomputer operation

Power saving of the main body of the calculator
- Power efficient programming
- Allocation of computing resources in response to chan ges in electricity prices

Improved efficiency of cooling facilities
- Higher cooling water temperature
- Highly efficient cooling facilities for free cooling
- automation

(2) Use of Large Energy Storage System

- Organize the concept of using storage batteries for each use case assumed in the use of storage batteries.
- Survey of storage battery types and examples in five categories organized according to the concept of storage battery use.

Storage Battery Use Cases	Concept of Storage Battery Use	output (e.g. of dynamo) (MW)	time capacity (h)
Used as load fluctuation absorption (Assistance for private power generation)	Absorbs minute-to-minute load fluctuations	5 to 20	0.1 to 0.5
Leveling of renewable energy sources	Absorb hourly fluctuations in renewable energy generation	100	3-12
Used as load fluctuation absorption	Absorbs minute-to-minute load fluctuations (Institutional, not yet supported is also acceptable)	-	-
Electricity from peak shaving Reduction of basic fee	Discharge when power setting is exceeded (limited number of discharge days)	1-10	1-3
Electricity prices by time of day	Charging and discharging linked to market prices (Discharge is from 15:00 to 21:00)	1-20	3-6
Use of raw green electricity (Re-energy and consumption are matched on an hourly basis)	Absorb load fluctuations on an hourly basis in conjunction with the amount of renewable energy generation	1-20	1-6

The 24/7 Carbon Free Energy Compact, an international initiative, provides 100% carbon-free power supply in accordance with hourly power consumption 24 hours a day, 365 days a year.

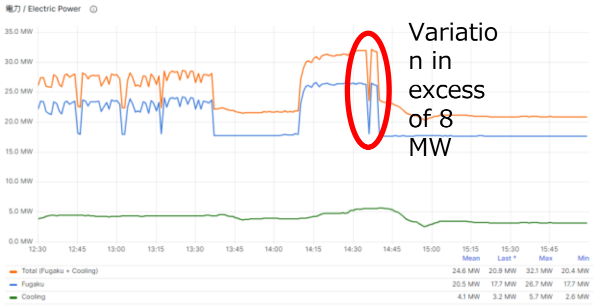


Figure . Maximum load fluctuation results (2023.7.27)

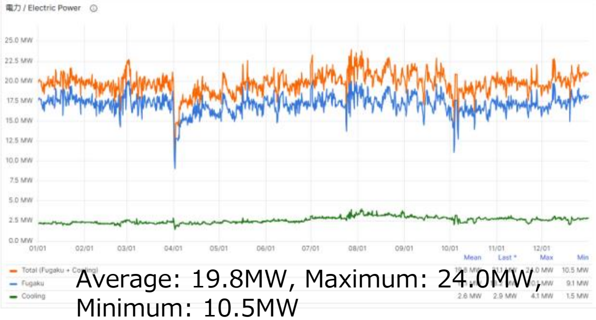


Figure 1: Actual annual load changes in 2023.

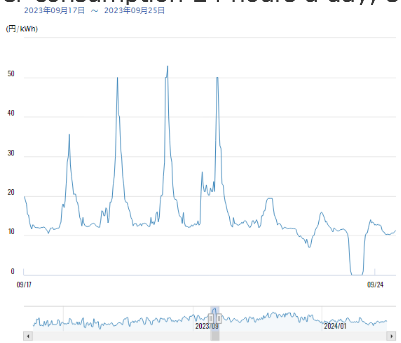


Fig. JEPX contract prices (2023.9.17-25)

Large Energy Storage System Case Studies

- **Installation Examples in JAPAN**

Figure 1: Redox flow battery at the Minamihayarai substation of Hokkaido Electric Power Co.

Figure 2: Lithium batteries at Tohoku Electric Power Company's Nishi-Sendai substation

Figure 3: Lithium batteries at Tohoku Electric Power Company's Minamisoma substation

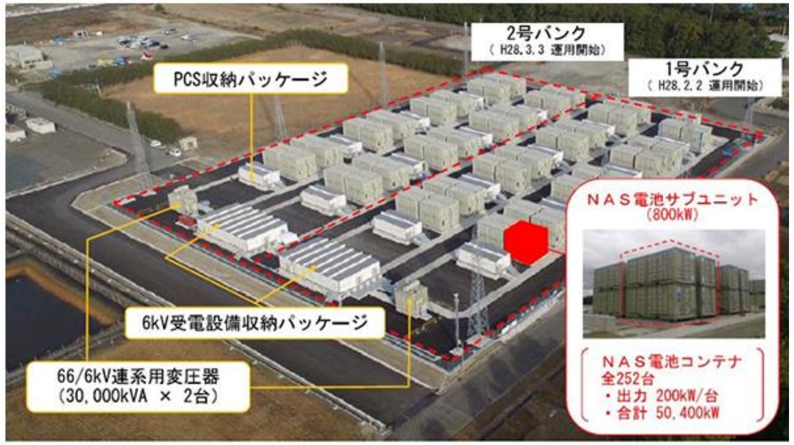
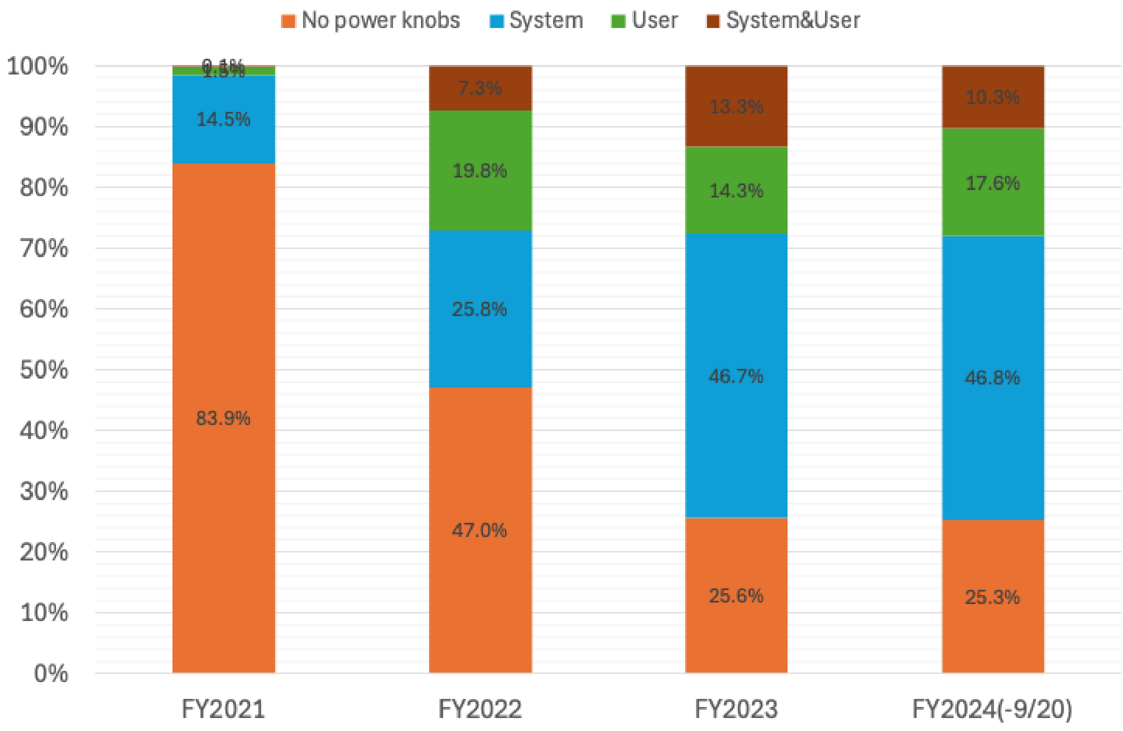


Figure 4: Sodium-sulfur battery at the Toyomae substation of Kyushu Electric Power Co.

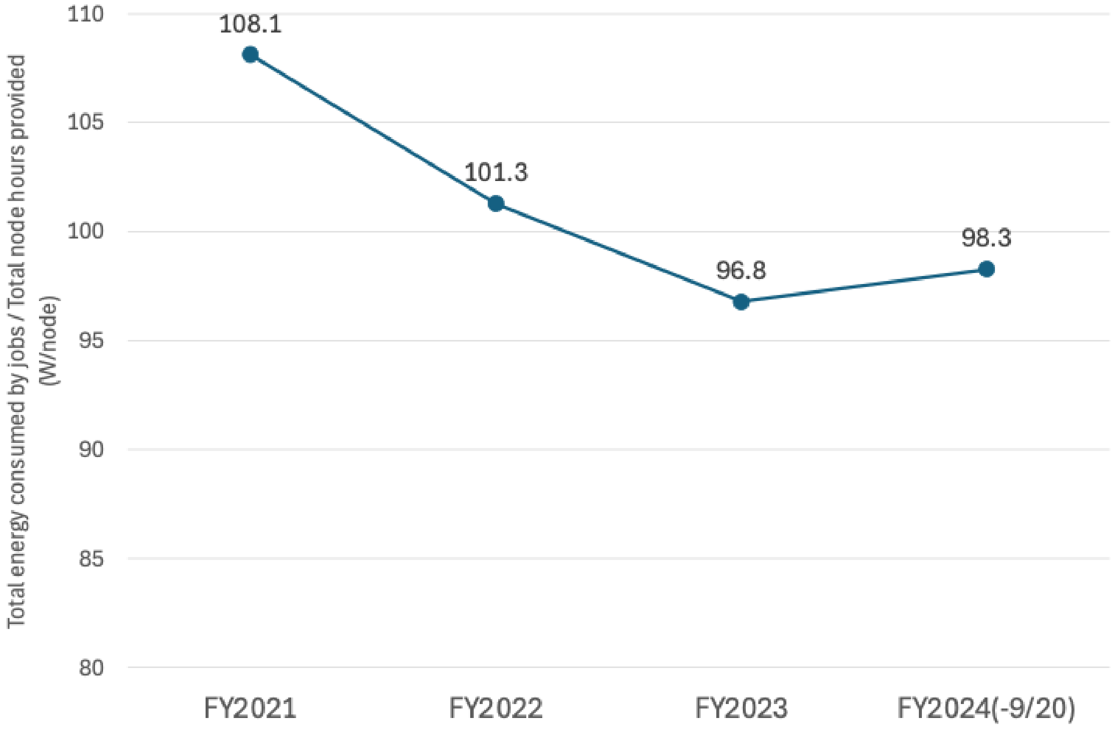
Source: New Energy Foundation website, New Energy "Recent Topics/Keyword" Explanation Corner
Figure 1 https://www.nef.or.jp/keyword/sa/articles_sa_03.html Figure 2 https://www.nef.or.jp/keyword/sa/articles_sa_03_02.html
Figure 3 https://www.nef.or.jp/keyword/sa/articles_sa_03_04.html Figure 4
https://www.nef.or.jp/keyword/sa/articles_sa_03_03.html

(4) Energy-efficient HPC system operation by incentivizing user cooperation

- **<u>"Fugaku Point" program since 2023</u>**
  - Fugaku has several functions for power saving, called "power knobs." However, it was one of the significant issues for us to facilitate users to use the functions.
  - The "Fugaku point" quantifies user cooperation for energy-efficient operations and is awarded for jobs with lower power consumption than a standard.
  - User can execute their jobs with higher priority by redeeming the points.



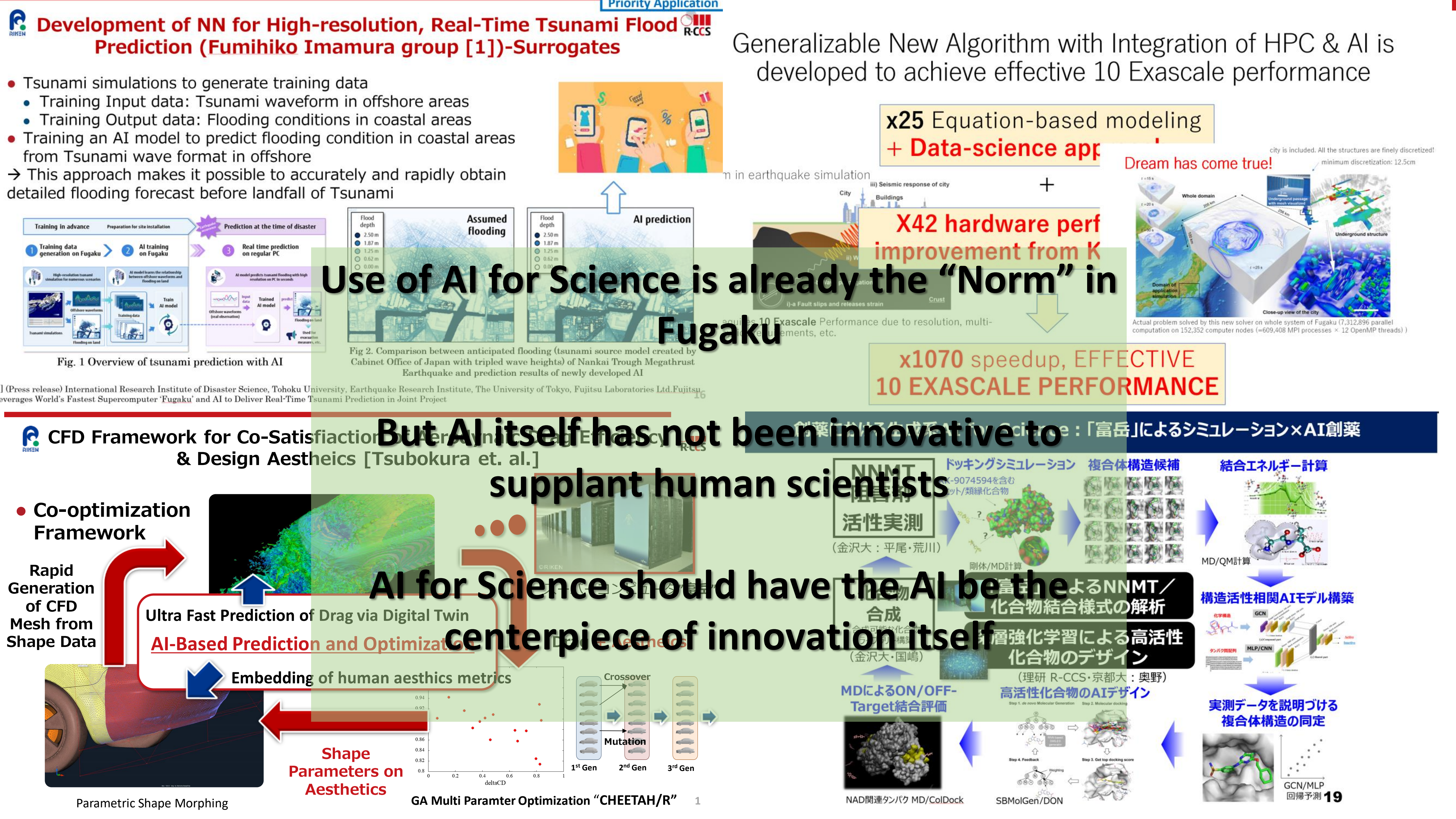
The percentage of jobs that use power knobs is increasing



Consequently, the watts per node have been reduced gradually

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: > x60**
- (Fugaku ROI according to Hyperion: 60x~80x. Expect greater ROI for FugakuNEXT of over 100x)
- 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**

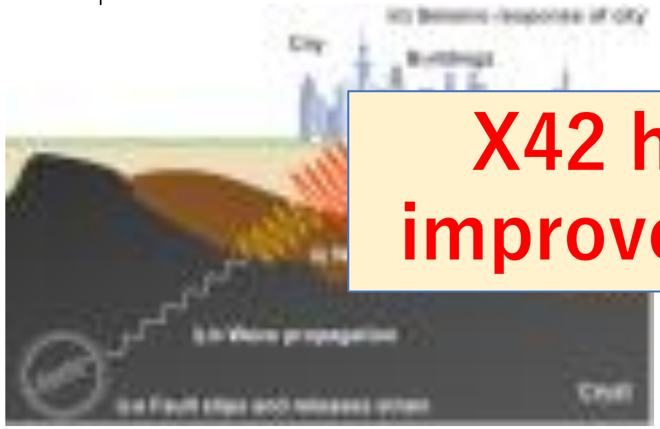


Use of AI for Science is already the "Norm" in Fugaku
But AI itself has not been innovative to supplant human scientists
AI for Science should have the AI be the centerpiece of innovation itself
Priority Application
Development of NN for High-resolution, Real-Time Tsunami Flood Prediction (Fumihiko Imamura group [1])-Surrogates
Tsunami simulations to generate training data
Training Input data: Tsunami waveform in offshore areas
Training Output data: Flooding conditions in coastal areas
Training an AI model to predict flooding condition in coastal areas from Tsunami wave format in offshore
→ This approach makes it possible to accurately and rapidly obtain detailed flooding forecast before landfall of Tsunami
Training in advance
Prediction at the time of disaster
Training data generation on Fugaku
AI training on Fugaku
Real time prediction on regular PC
Assumed flooding
AI prediction
Fig. 1 Overview of tsunami prediction with AI
Fig 2. Comparison between anticipated flooding (tsunami source model created by Cabinet Office of Japan with tripled wave heights) of Nankai Trough Megathrust Earthquake and prediction results of newly developed AI
] (Press release) International Research Institute of Disaster Science, Tohoku University, Earthquake Research Institute, The University of Tokyo, Fujitsu Laboratories Ltd.Fujitsu
everages World's Fastest Supercomputer 'Fugaku' and AI to Deliver Real-Time Tsunami Prediction in Joint Project
16
Generalizable New Algorithm with Integration of HPC & AI is developed to achieve effective 10 Exascale performance
x25 Equation-based modeling
+ Data-science app
X42 hardware perf improvement from K
x1070 speedup, EFFECTIVE 10 EXASCALE PERFORMANCE
Dream has come true!
Actual problem solved by this new solver on whole system of Fugaku (7,312,896 parallel computation on 152,352 computer nodes (=609,408 MPI processes × 12 OpenMP threads))
CFD Framework for Co-Satisfiaction of Aerodynamic Drag Efficiency & Design Aestheics [Tsubokura et. al.]
Co-optimization Framework
Rapid Generation of CFD Mesh from Shape Data
Ultra Fast Prediction of Drag via Digital Twin
AI-Based Prediction and Optimization
Embedding of human aesthics metrics
Shape Parameters on Aesthetics
Crossover
Mutation
1st Gen
2nd Gen
3rd Gen
Parametric Shape Morphing
GA Multi Paramter Optimization "CHEETAH/R"
「富岳」によるシミュレーション×AI創薬
ドッキングシミュレーション
複合体構造候補
結合エネルギー計算
活性実測
(金沢大：平尾・荒川)
MD/QM計算
構造活性相関AIモデル構築
化合物のデザイン
(理研 R-CCS・京都大：奥野)
MDによるON/OFF-Target結合評価
高活性化合物のAIデザイン
実測データを説明づける複合体構造の同定
NAD関連タンパク MD/ColDock
SBMolGen/DON
GCN/MLP 回帰予測
19

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

x25 Equation-based modeling
+ Data-science app

Dream in earthquake simulation



+

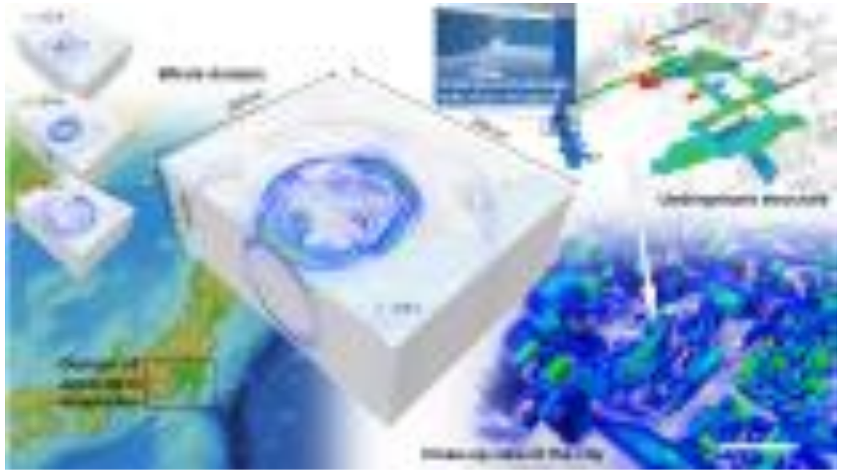
X42 hardware perf improvement from K

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

city is included. All the structures are finely discretized!

Dream has come true!

minimum discretization: 12.5cm



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

x1070 speedup, EFFECTIVE
10 EXASCALE PERFORMANCE

Fugaku-LLM – Massive LLM Training on Fugaku

Rio Yokota

Collaborators

GPT-Fugaku Team

Noriyuki Kojima, Kazuto Ando, Koji Nishiguchi, Jungo Kasai, Keisuke Sakaguchi, Shukai Nakamura

GPT FUGAKU EST. 2021

DL4Fugaku Team @ R-CCS

Aleksandr Drozd, Mohamed Wahib, Kento Sato, Jens Domke, Emil Vatai

R-CCS

DL4Fugaku Team @ LLNL

Nikoli Dryden, Tal Ben Nun

Fujitsu

FUJITSU

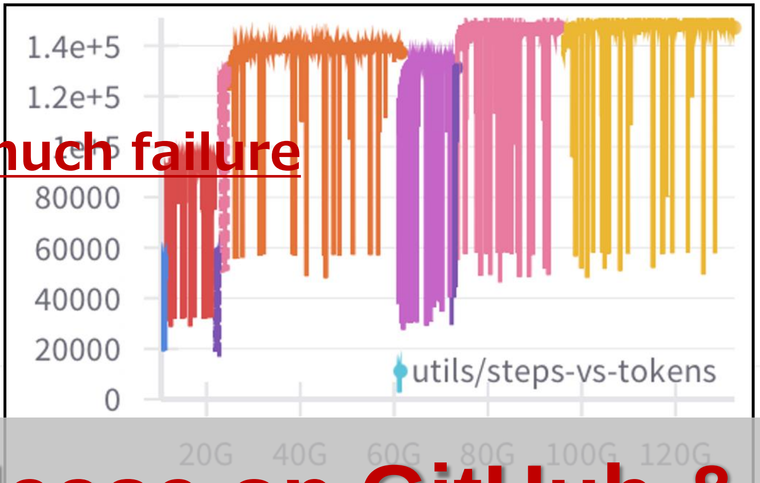
Kenichi Kobayashi, Naoto Fukumoto, Akihiko Kasagi

Koichi Shirahata, Kentaro Kawakami, Masafumi Yamazaki, Hiroki Tokura, Takumi Honda, Tsuguchika Tabaru

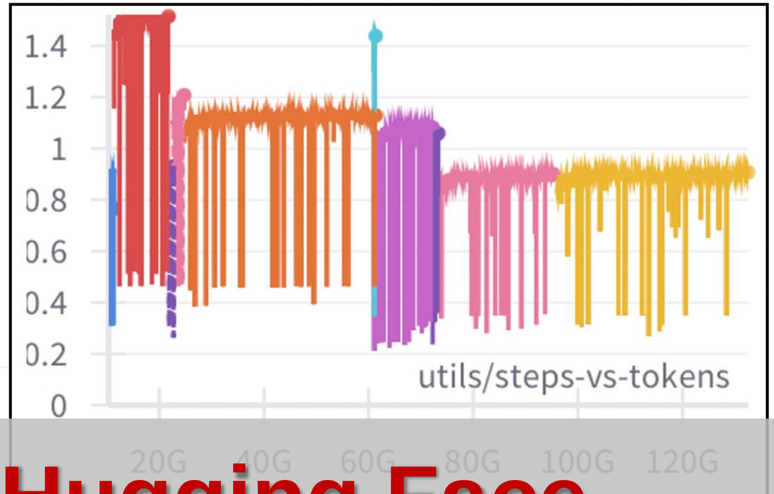
FugakuLLM – training on 14,000 nodes

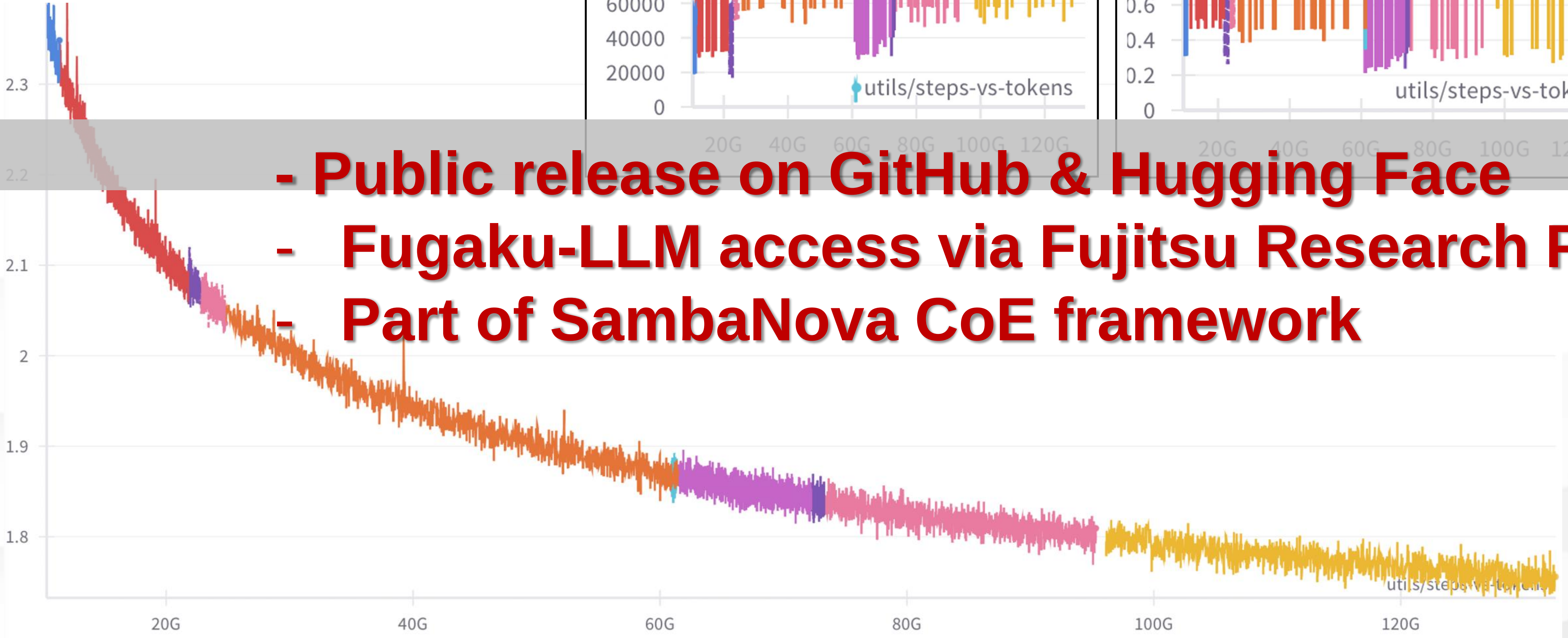
- **Data size: 400B**
- **Model size: 13B**
- **Fugaku: 13,824 nodes**
- **<u>Weeks of training without much failure</u>**
- **1-1.4 Tflops/s / node**

Tokens per sec.



TFLOPS per node





- Public release on GitHub & Hugging Face

- Fugaku-LLM access via Fujitsu Research Porta

- Part of SambaNova CoE framework

AI for Science Roadmap in Japan

(Issued on May 31, 2024)

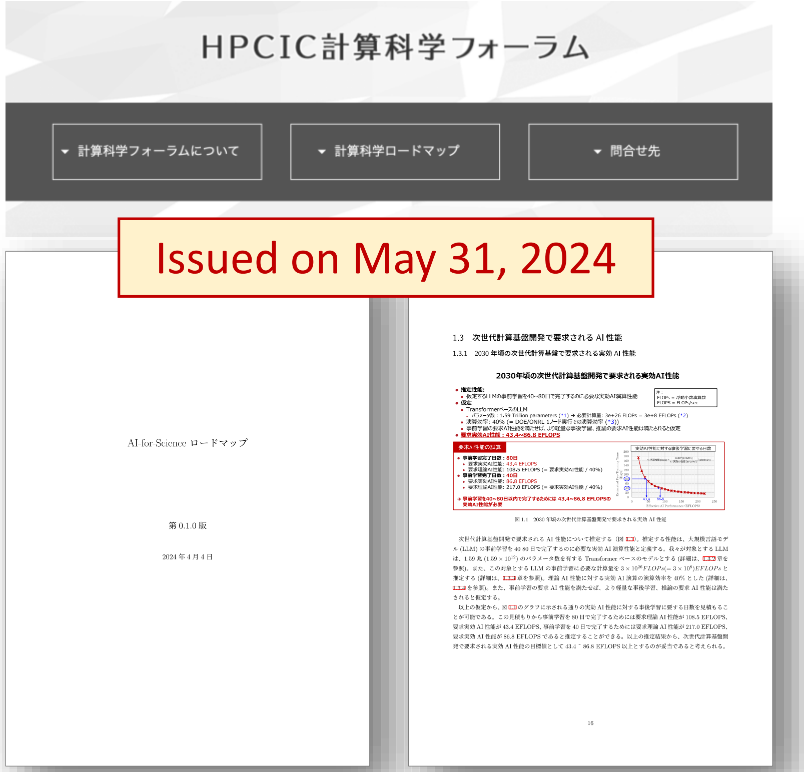
AI for Science Roadmap - Overview

- **Abstract:**
  - Summary of efforts to drive future AI-for-science researchers in Japan
  - A roadmap is being developed that includes examples, guidelines and new challenges on the application of cutting-edge technologies such as surrogate modeling and the use of generative AI to research areas, potential use cases, and possibilities.
  - Estimation of required AI computational performance to the next-gen supercomputer based on the roadmap and by identifying issues related to AI governance

- **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)**
  - Researchers extracted from keyword searches such as AI from HPCI proposals
  - Authors of the HPCIC Computational Science Roadmap in their respective fields
  - FY2023 Accelerated Program for the Creation of Tomiyama PI
  - RIKEN R-CCS



AI for Science Roadmap in Japan
(Issued on May 31, 2024)

Expansion of AI application areas in various scientific fields

2. 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

3. 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

4. elementary particles and nuclei

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

5. 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

6. 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

7. design and manufacturing

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

8. social sciences (to be written after 2024)

9. brain science and artificial intelligence

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

10. 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

11. 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**

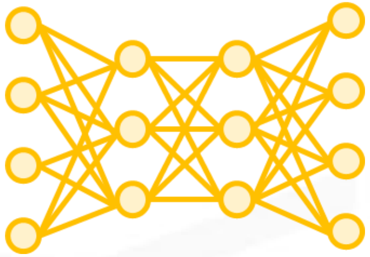
12. space and astronomy

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

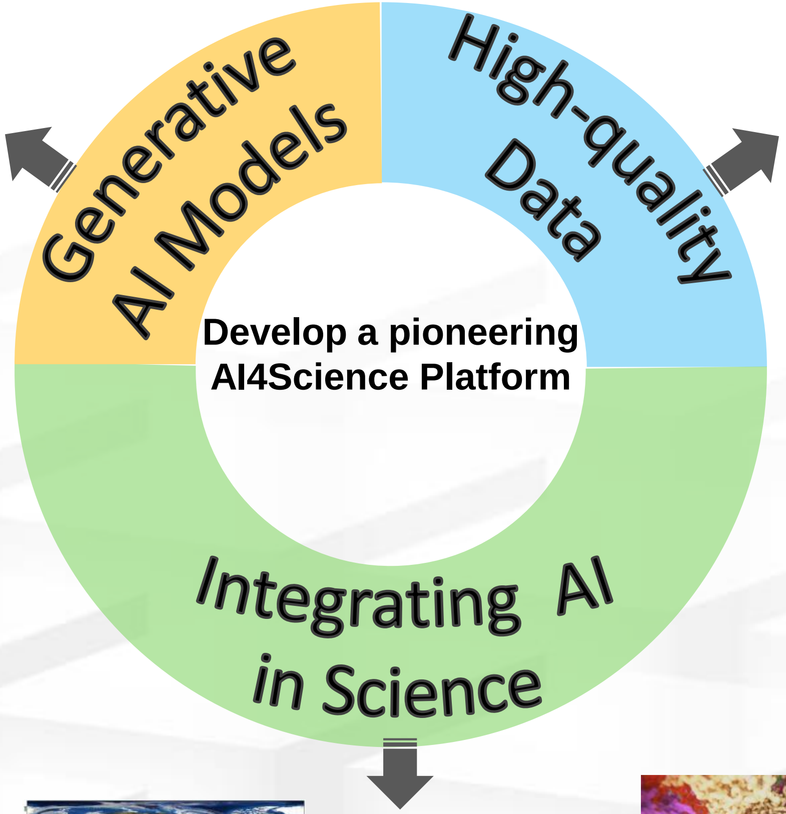
RIKEN's Initiatives ～TRIP-AGIS～

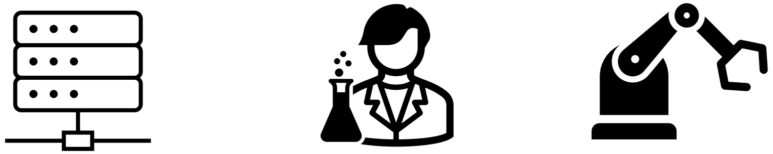
Artificial General Intelligence for Science of Transformative Research Innovation Platform (TRIP-AGIS)

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



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



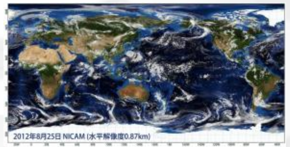


Simulations Experiments Robots

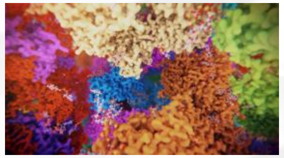
Produce large amounts of high-quality data through RIKEN's and its parternerships/collaborations. Strengths in measurement techniques and experiment automation

Purpose and Challenge

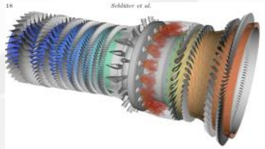
- **Solve intractable science problems**
- **Lead advanced science**
  - **Starting from basic science**
  - **To societal impact (GX, inclusive society, etc.)**



Physical/Earth



Life/Medical



Engineering

Overview of Riken TRIP-AGIS AI for Science Project (2024-2031)

① Common platform technology

Advanced model Development of fundamental technology that enables training of multimodal generative AIs.

High-quality data Automation and acceleration of experiments that enable both (1) generation of massive data essential for multimodal foundation models and (2) automatic execution of the experiments designed by the AI model.

② Generative AI models for scientific research in specific scientific fields

Life and medical sciences

High-quality data Time course of drug responses of cells, effects of diseases on the animal's behavior and body, etc.

Advanced model Model that enables comprehensive interpretation and prediction of phenomena from genomes, cells to whole organisms.

Materials sciences

High-quality data Material structure, properties, electronic state, manufacturing method, etc.

Advanced model Model that can generate data based on integrated interpretation of properties, material structures, fabrication methods, etc., both inorganic and organic.

③ Innovative Computational Infrastructure

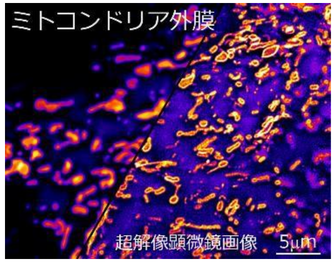
Develop and operate a computer system for the development and sharing of generative AI models for scientific research that are optimized for inference, training, and generation of various types of scientific research data.

Research on novel computing principles with high computing and power performance beyond conventional GPUs.

- We try to integrate various data as **multimodal foundation models (FMs)**

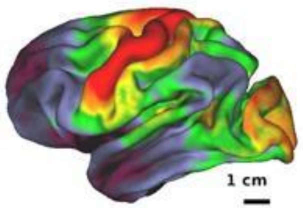
Genome/
Transcriptome

Proteins



Images

Other Omics
lipidome etc.

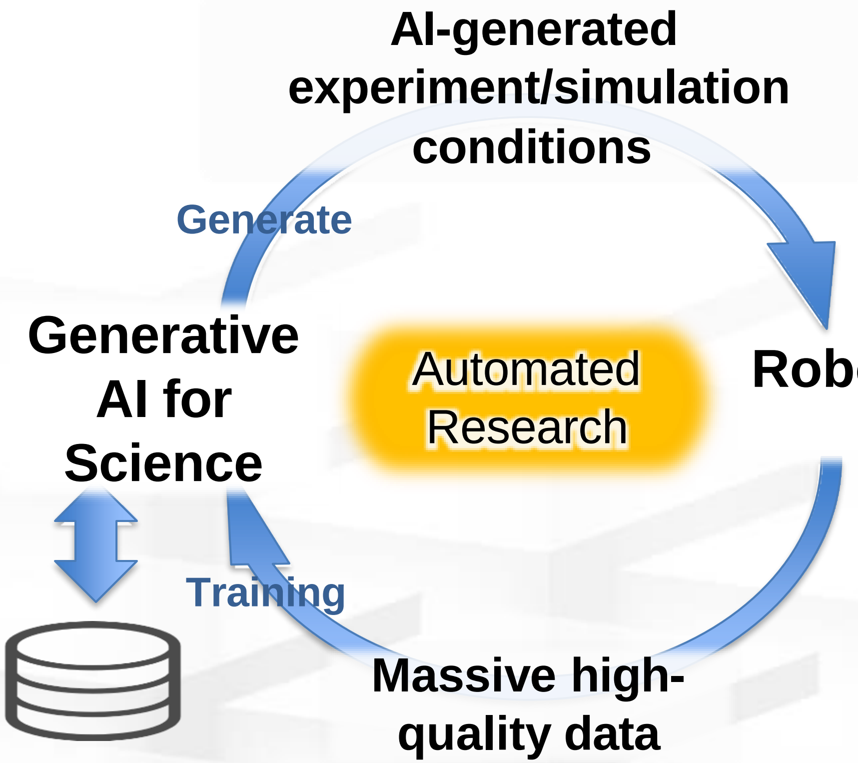


Neural
activities

Other
Phenotypes

- We especially focused on multimodal FMs of dynamical behaviors based on **systematic data acquisition of simultaneous multimodal measurements.**
  - ▷ Dynamic / spatial transcriptome and super-resolution imaging
  - ▷ Animal behaviors (motions and voices) with genetic backgrounds / neural activities
- RIKEN can cover measurements of many modalities in life science.

AI-driven automatic research and massive data production using robotic experiments and large-scale simulations



Robotic experiment / Simulation

Parallel robot lab.

Large-scale simulation by Fugaku

Acceleration of scientific research by AI

Development of cellular response atlas

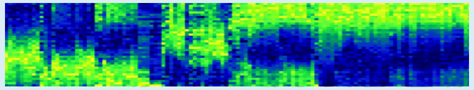
Drugs etc.
～5,000 types

iPS-derivied cells etc.
～100 types

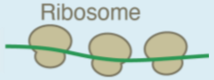
@TogoTV

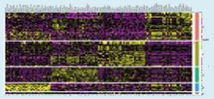
Systematic measurements of time-series of multimodal omics and image data for
>5,000 types of stimulus X >100 Cell types
= 500K combinations

- Large-scale single-cell transcriptome

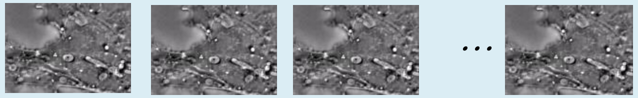


- Translation control by Ribosome profiling





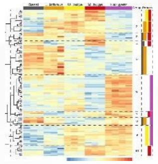
- High-speed live cell imaging



Resolution: λ/3
10 ms/frame

+ More Modalities in future

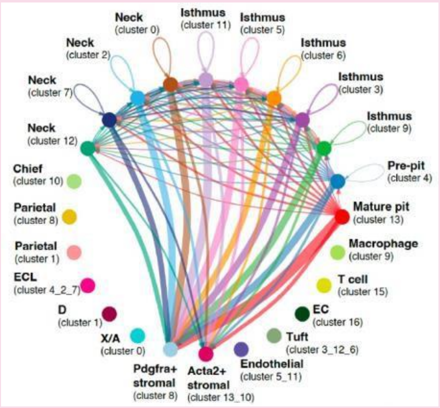
Genome Sequence



ATGCACGTCAGC
GACCGACGTAAT

Foundation model of cellular response

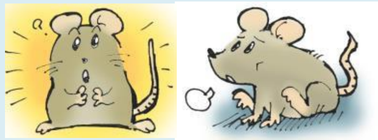
- ✓ A model to predict cellular dynamics on stimulus
- ✓ Predictions of time-series of cellular status by drugs etc.
- ✓ Applications to pharmaceutical developments, organoid developments, regenerative medice, and so on.



Lifetime home-cage monitoring of animals

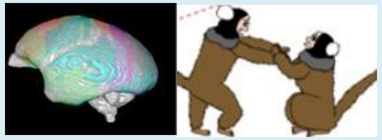
AI-based analysis of long-term motional and vocal behavior

Mouse

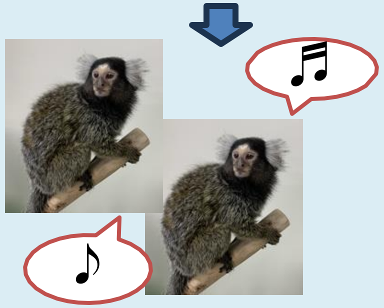


Normal and disease models

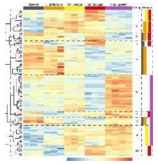
Marmoset



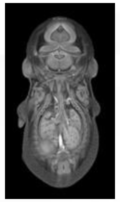
Human-like social behavior



ATGCACGTCAGC
GACCGACGTAAT



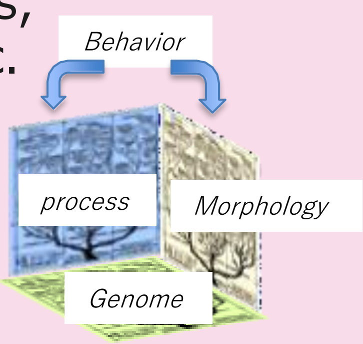
Genomes



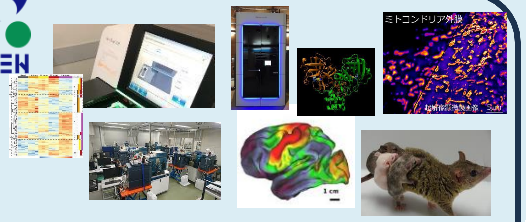
Morphology By CT

Foundation model of animal behavior

- ✓ Develop a model to reproduce variety and complexity of animal behavior
- ✓ Predict diseases or their precursors from behavior, and vice versa.
- ✓ A connection with human behavior at diseases, social situations etc.



Mutimodal data in public DB (and in RIKEN)



In-house data in RIKEN

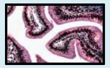
ToMMo

BIOBANK JAPAN

NBDC

PubMed

LINC Life Intelligence Consortium

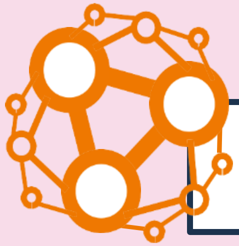


Clinical data

AI/ML-based approach

Extract data as Knowledge-graph

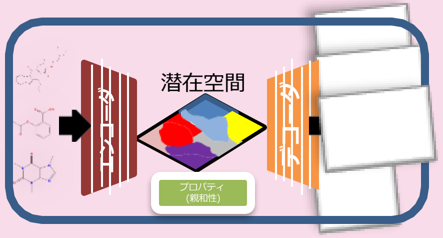
Extract non-graph data (table etc.)



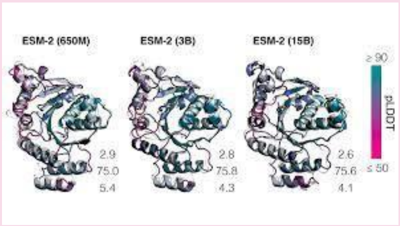
Foundation model of Knowledge-graph

Clinical applications

Generative AI for Chemical structure



Foundation models for proteins



ESM-2

+ Molucular simulation, Molecular docking

Multimodal Foundation model for drug design

Multimodal Foundation Model for Drug Developments

Purpose

◆Propose candidate materials and synthetic methods to achieve desired material functions.

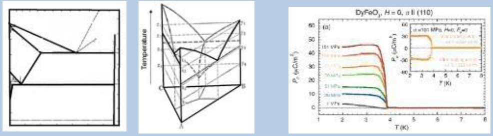
◆Accelerate and enhance materials science research in basic science and industry by allowing users to train additional machine learning models using their own data

Our Approach

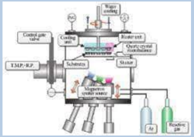
High quality material data from literature, experiments, and simulations

Propose materials with the required properties and predict their synthesis and processing methods.

- Material data and information on synthesis and processing as described in the literature
- High quality experimental data to be newly acquired



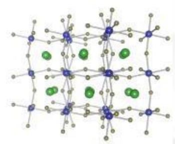
Material data by computational materials science



Material data by combinatorial synthesis method

Generative AI model for materials science based on a generic LLM →

- Prediction of physical properties by combining computational physics and machine learning

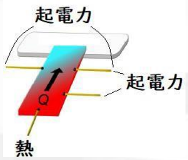


Crystal structure

Electronic structure

$$\mathcal{H} = \sigma_{ij} t_{ij} a_i^{\dagger} a_j + \sigma_i U_i n_{i\uparrow} n_{i\downarrow}$$

Model Hamiltonian



Material properties

Computational Physics

Machine Learning

→ Accurate prediction of material properties based on physical laws

Foundation Model for Materials Science

◆ Generation of 3D arrangement information of atoms to achieve desired material properties by AI model

◆ Generation of synthesis and processing methods for proposed material.

→ **Accelerate development of innovative materials**

Step 1: Magnetic materials
Step 2: Polymer materials, and others

RIKEN
Center for
Computational Science

AI4Science Software: Models, Data, and Integration

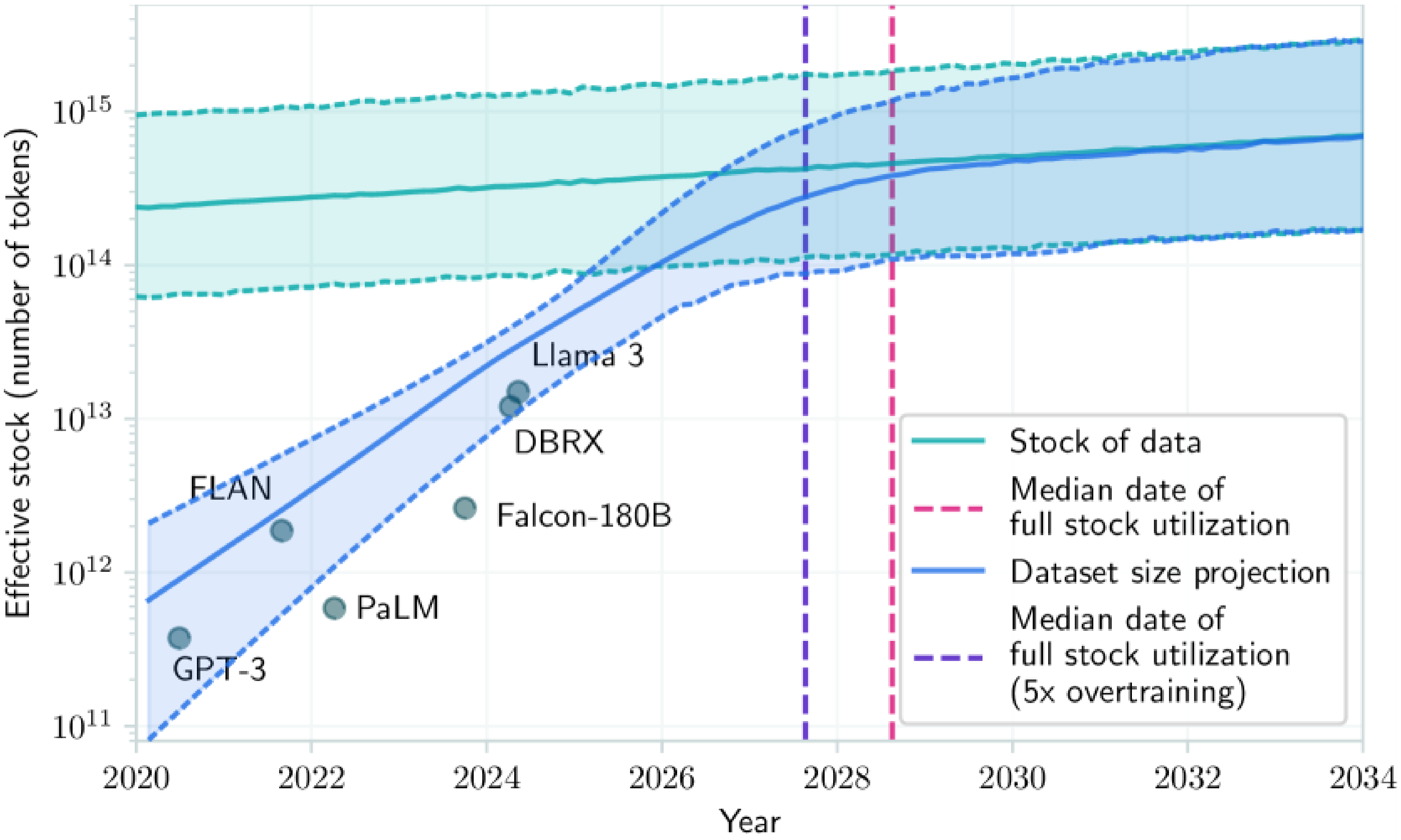
Mohamed Wahib[1,2]

1. High Performance Artificial Intelligence Systems Research Team

2. Learning Optimization Platform Development Unit

August 2024

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



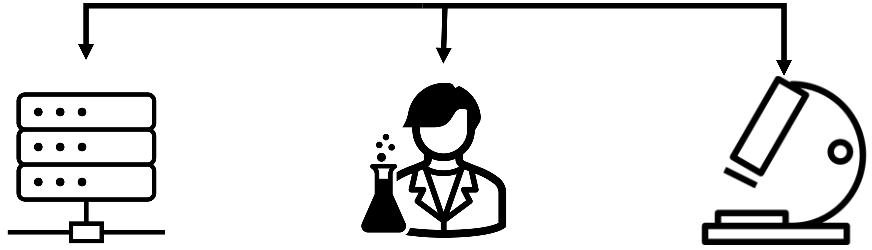
* *Villalobos et al., "Will we run out of data? Limits of LLM scaling based on human-generated data", ICML'24*

AI for Science will Innovate Modern AI - Data

Compute → Money and time problem

Data → Sourcing problem

Sources of Scientific Data



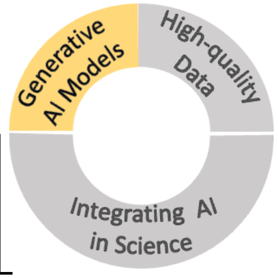
- **Now**: models pre-trained on traditional AI applications data → tuned on science data
- **Future??**: models continually or pre-trained on scientific data → tuned on traditional AI applications data

	Traditional AI Applications Data	Scientific Data
Properties	Structured; low dim; ubiquitously-used formats; low-quality +	Semi-structured; high dim; arbitrary or complex formats; high-quality −
Tooling	Rich ecosystem +	Abysmal −
Volume	O(100TB) excluding videos −	Arguably more than your storage budget +
Growth	Existing data est. to run out ~2028*; new data grows ~linearly −	Exponential/linear (based on science area) +
Authenticity	Sources contaminated with Generated data** −	Clean Source +
Ownership	Courts still deciding! −	Usually open +
Lineage	Ever tried to track a photo source on the Internet? −	Clear lineage & trackability +

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

** *Shumailov et al., "AI models collapse when trained on recursively generated data", Nature 631, 755–759 (2024)*

Multi-dimensional Images in Science/Engineering



Dimensions	Resolution	Tokens/Sample Patch = $16^2/16^3$	Dataset Sizes	Example
3 Spatial + 1 Temporal + N Channels	- $100s^3$ - 10s channels (ERA5 dataset)	~ 300K	~10 PB	Weather\Climate Simulations
2 Spatial + 1 Temporal + N Channels	- $1000s^3$ - 10s channels	~5M	~ 10s TB	Satelliate Images
2 Spatial + 1 or N Channels	- $100K^2$	~100Ks (4x4 patch)	~ 10s TBs	Microscopic (Ex: Pathology)
2 Spatial + 1 Temporal + N Channels	- 100s2 ~ Hours (24 f/s) (YouTube-8m)	~1M	~1 PB	Video
3 Spatial + 1 Channel	~$8\text{-}12K^3$ >16^3 new beam	~1B	~100s TB	X-Ray CT (Ex: SP-µCT)
3 Spatial + N Channels	~$4K^3$ (sub 5-micron)	~ 30M	~ 10s TB	MRI (Ex: dMRI)

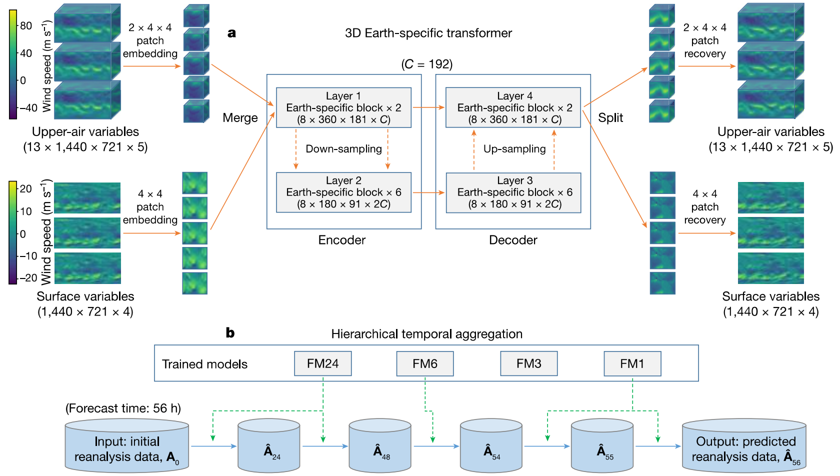
Weather Forecasting with Vision Transformer

R-CCS

Oak Ridge National Laboratory

nature

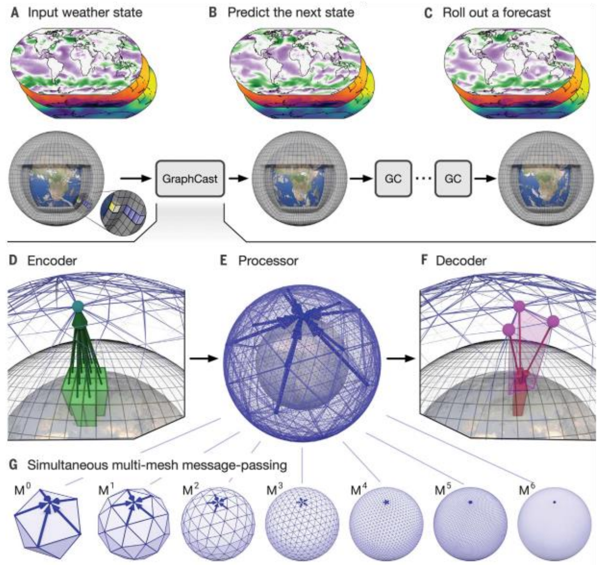
Accurate medium-range global weather forecasting with 3D neural networks



Pangu (by Baidu)

Science

Learning skillful medium-range global weather forecasting



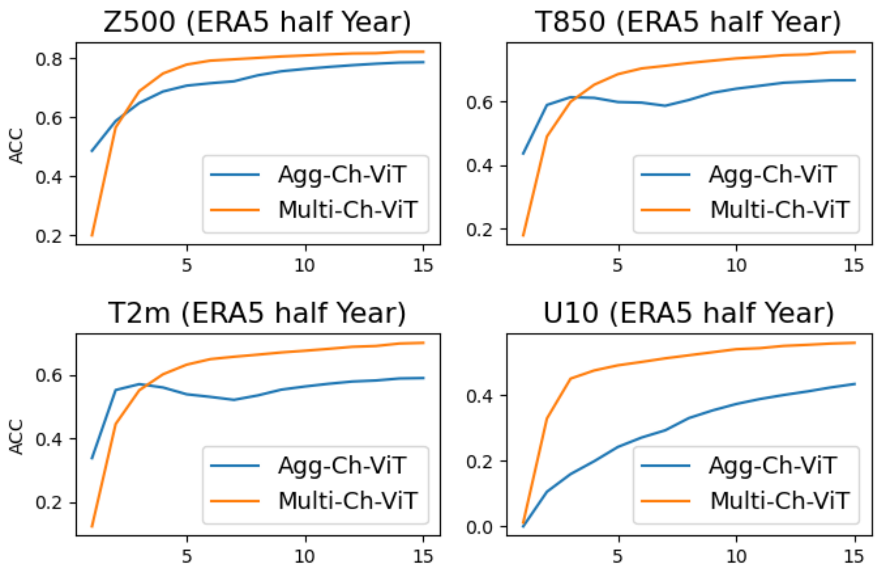
GraphCast (by Google Deepmind)

- Impressive results despite not training on the ENTIRE dataset (ERA5 dataset)
  - 1940 to present: each year at full resolution and all parameters ~ 100TB → 8.4 Petabytes
  - For reference, GPT4 trained on 20T tokens = 15 Terabytes (1/560 of ERA5)
- Could we train a weather prediction foundation model with entire dataset?

Weather Forecasting with Vision Transformer

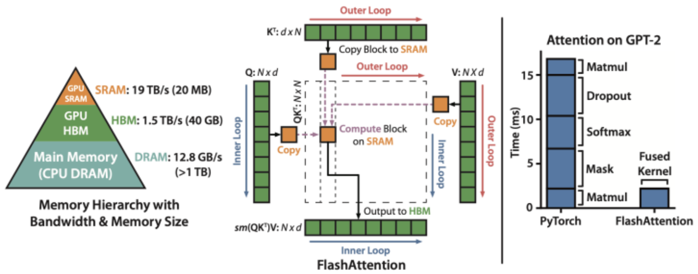
[ACM Gordon Bell Prize Finalist 2024]

- To train with entire ERA5 → Solve the long sequence problem
  - Combine different methods
  - Train on Frontier supercomputer (in collab. w/ ORNL)
  - ½ year at ~6% resolution 10K node-hours per epoch
    - Entire dataset (84 year) @100% resolution → Full Frontier 8 years

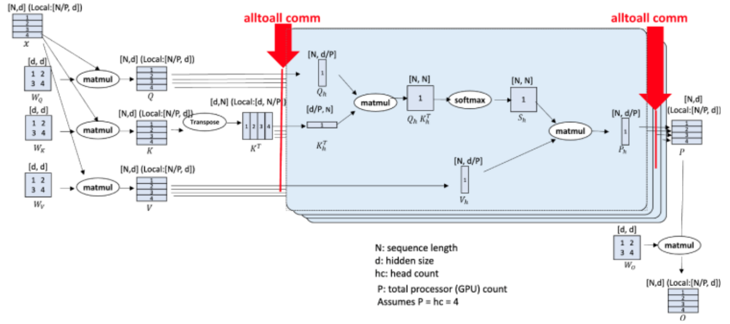


Validation accuracy on half-year of the ERA5 dataset. The effect in the accuracy is shown for including all 92 variables in the model.

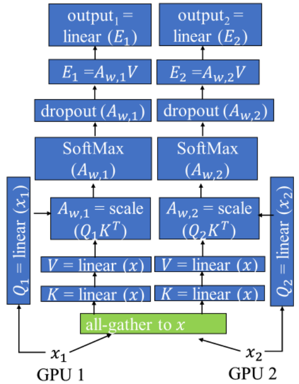
FlashAttention



MS DeepSpeed-Ulysses



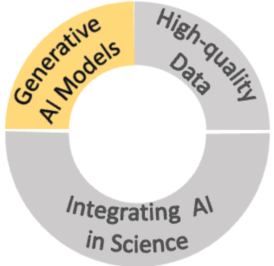
Fully Distributed Sequence



* Under review: Tsaris et al, *"Sequence Length Scaling in Vision Transformers for Scientific Images on Frontier"*

Models

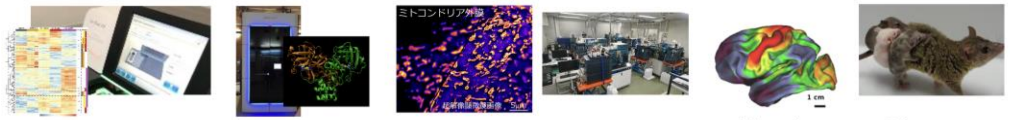
"generate python code to generate this cat drawn by polygons"




```
# Draw the cat's ears
t.penup()
t.goto(-30, 80)
t.pendown()
t.fillcolor("pink")
t.begin_fill()
t.circle(20)
t.end_fill()
```

Consumer facing AI:
- Switching between modalities
- Injecting from modality A to modality B

AI-based Science:
- Extracting knowledge from combined view of modalities



Different partially/completely aligned (or not) view of same target phenomenon (animal behavior) → Extract knowledge

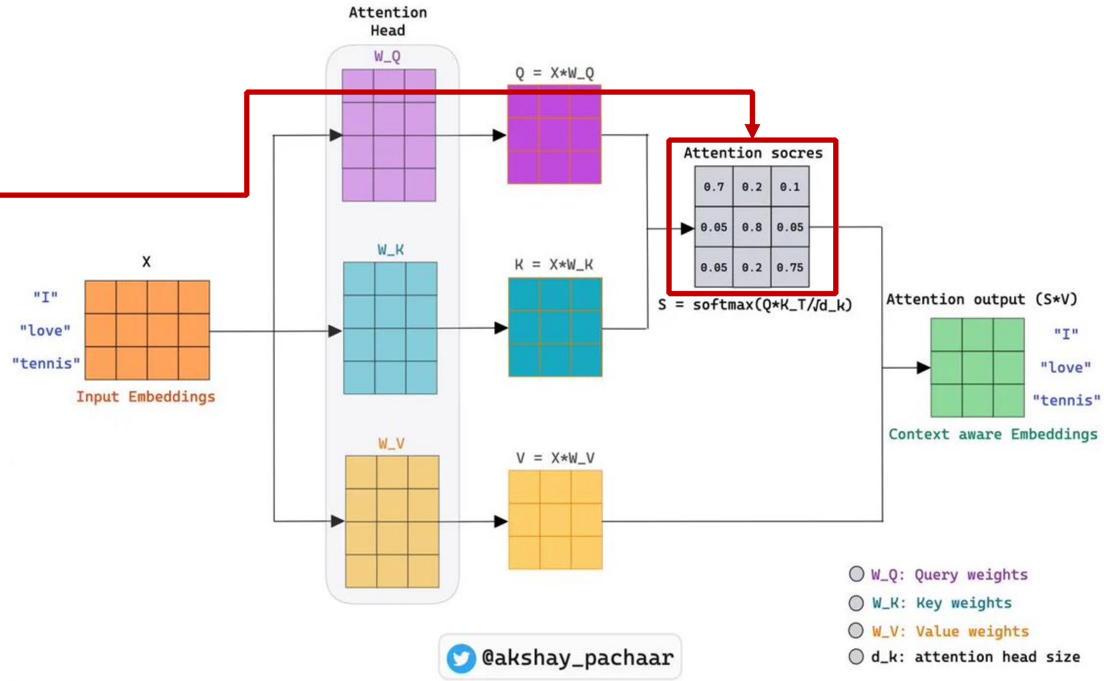
- **Multi-modalities**
  - Complex and custom encoding schemes are often required
  - AI for Science: much more variety in modalities vs. consumer facing AI
    - Different encoders for different modalities: share common latent space (ex: concat)

Multimodality in Science
    - Aligning representations of different modalities
    - Mask and predict information about modality A from modality B
    - Advanced multimodal fusion to combine features from different modalities

Longer Sequence: a Challenge

- The longer the sequence, the more the **context** that can be extracted
  - Ex: feeding an LLM entire books, library of papers, RAG, or **segmentation**
  - GPT-4-turbo → 128,000 tokens – GPT4-32k → 32,768 tokens (1 Token = ¾ Word)
  - Gemini supports 1 million tokens but…

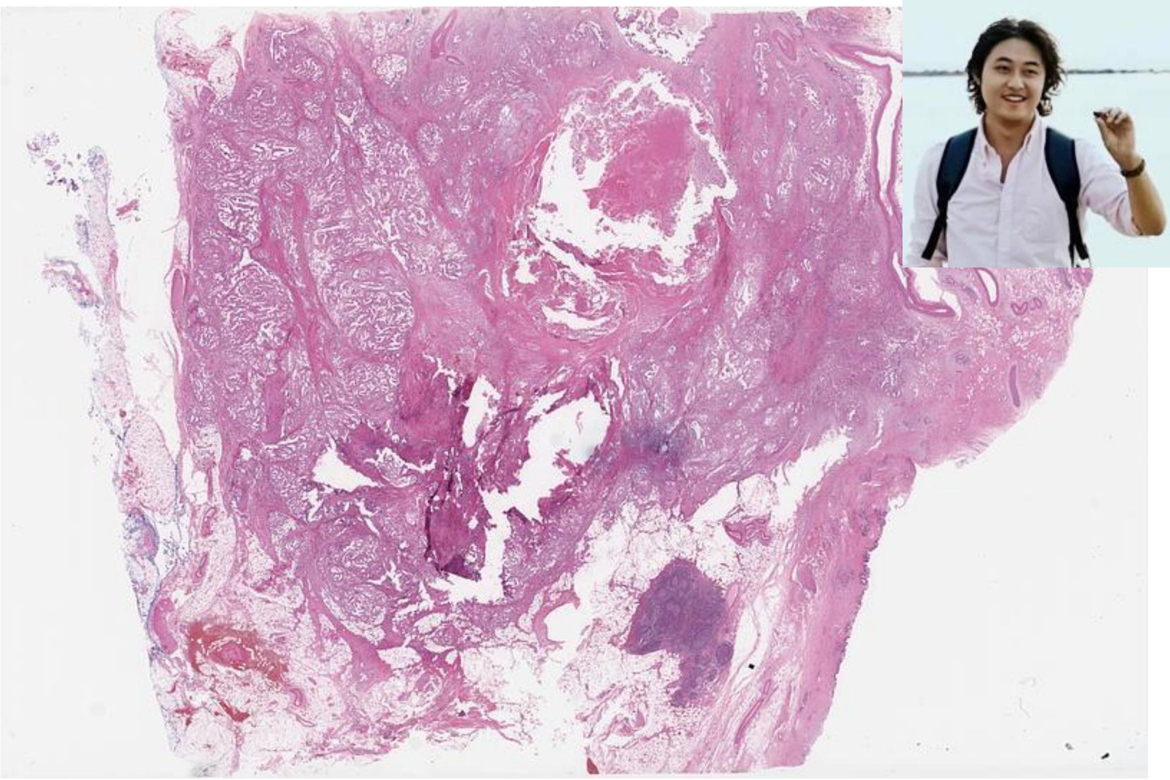
- Compute and memory cost ∝ sequence^2



Tumor Cellularity Prediction in Pancreatic Cancer and Colon Cancer

Enzhi Zhang (PhD Student @Hokkaido U.)

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



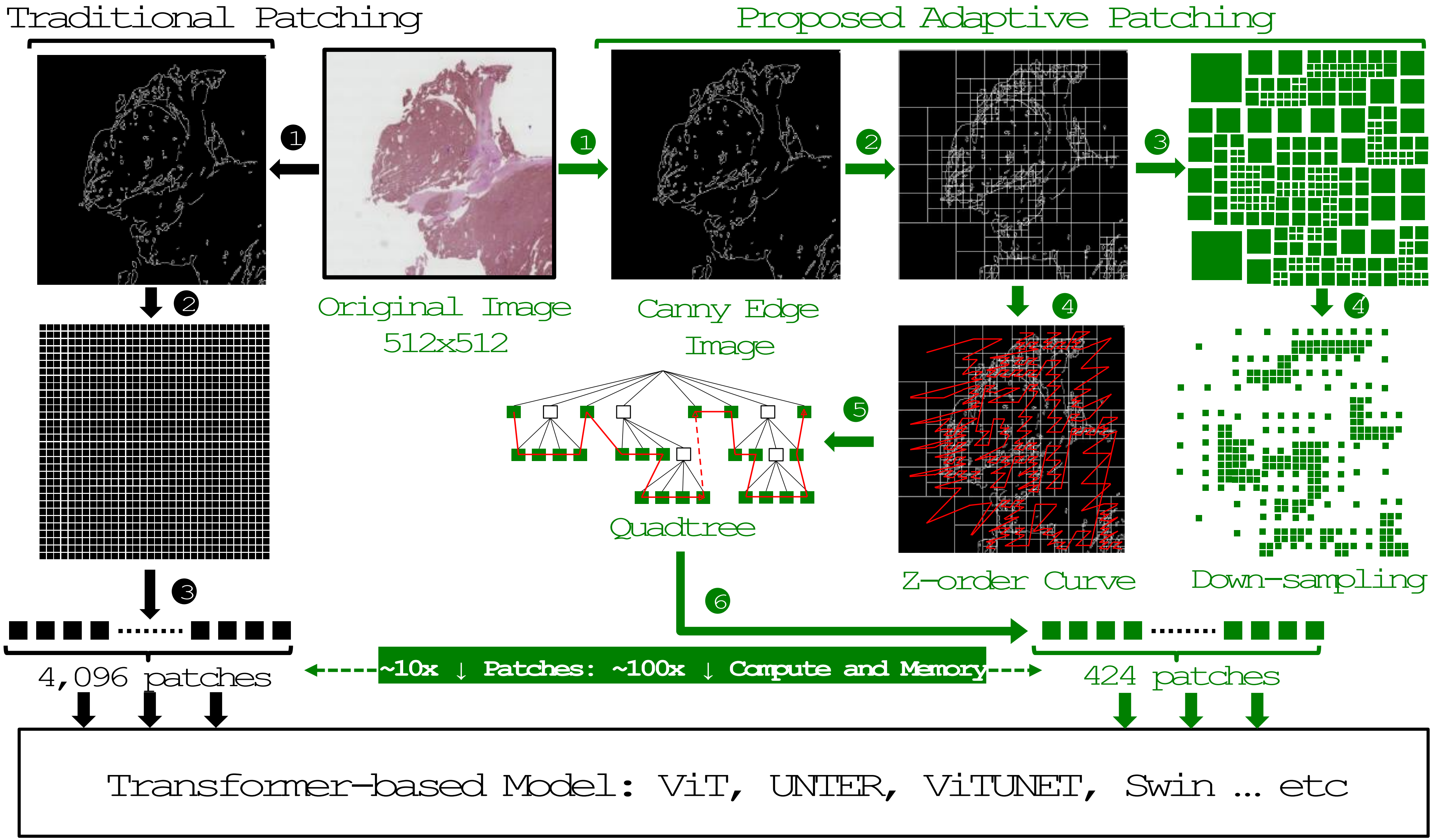
- Segmentation with Vision Transformer (ViT)
- Might require 1 billion input tokens(!)

- Challenge:

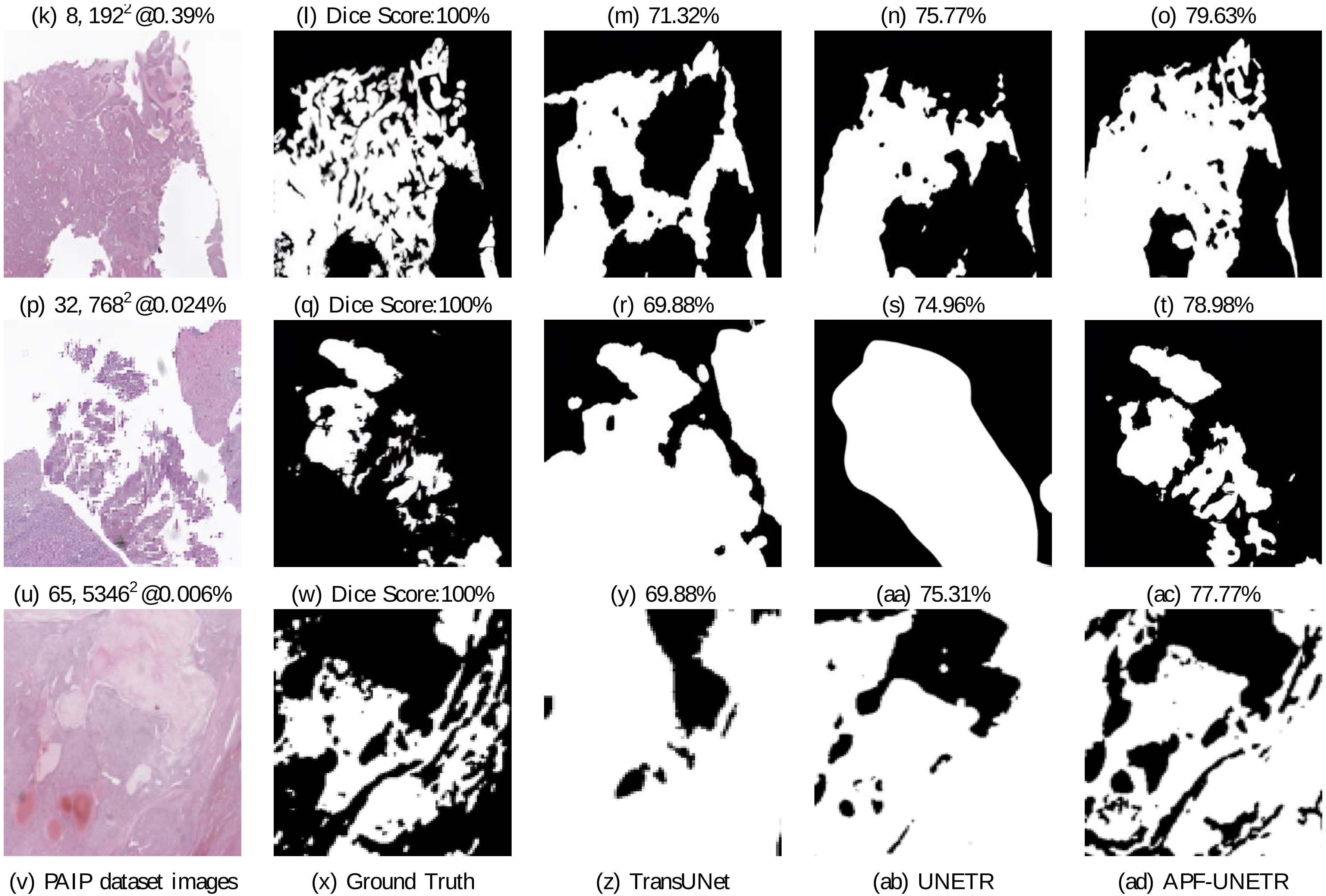
PAIP 2023: Tumor cellularity prediction in pancre **g)**

* https://arxiv.org/pdf/2404.09707

Tumor Cellularity Prediction in Pancreatic Cancer and Colon Cancer

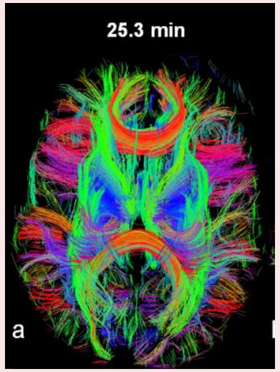


Tumor Cellularity Prediction in Pancreatic Cancer and Colon Cancer

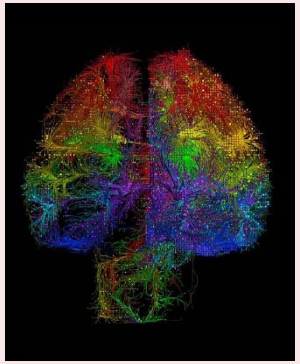


Resolution of Mouse-brain MRI Images

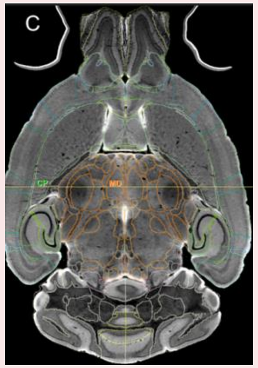
1-2mm Human connectome, and atlas (HCP)

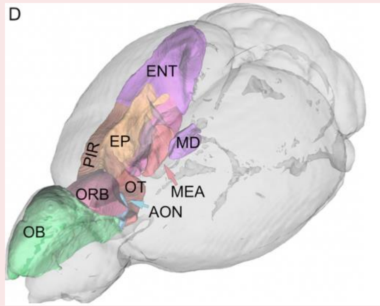


100 micron connectome
(Knox et al., 2019)

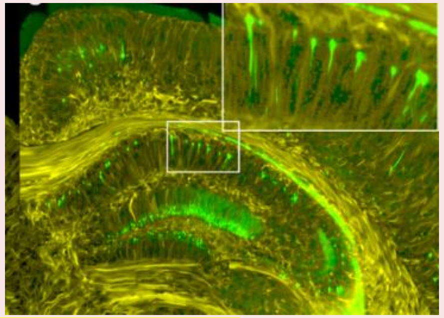


15 micron iso voxel dMRI
(Johnson et al., 2022)





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



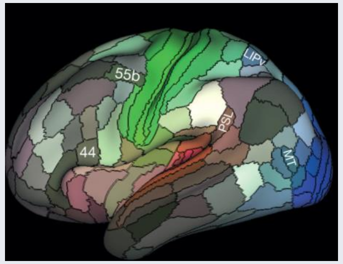
Brain Data (collab. ORNL/Duke U.)

>100 microns

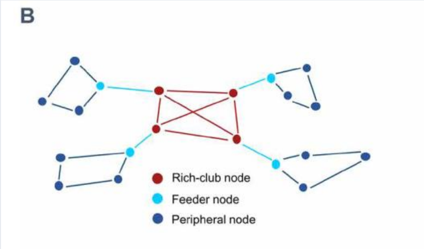
100 microns

15 microns

<5 microns

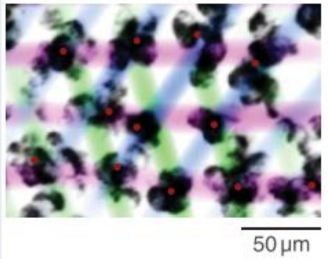


Classified regions
(Glasser et al., 2016)



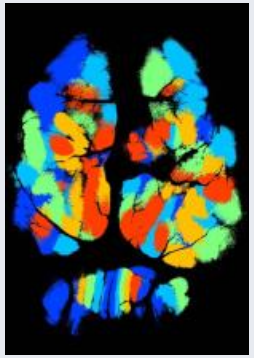
Network structure
(Colleta et al., 2020)

Find Functional Module Structure



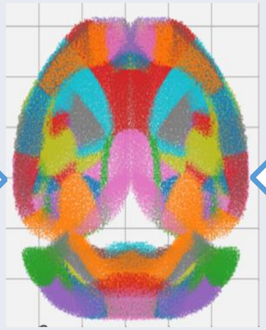
Columns (ex: Maruoka et al., 2017, Zeeuw et al., 2020)

Calcium Imaging



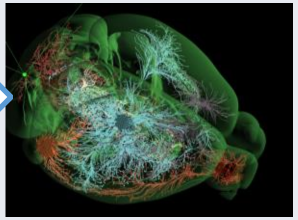
Michikawa et al., in prep

Brain Simulation



Igarashi et al., in prep

Cellular Connectome



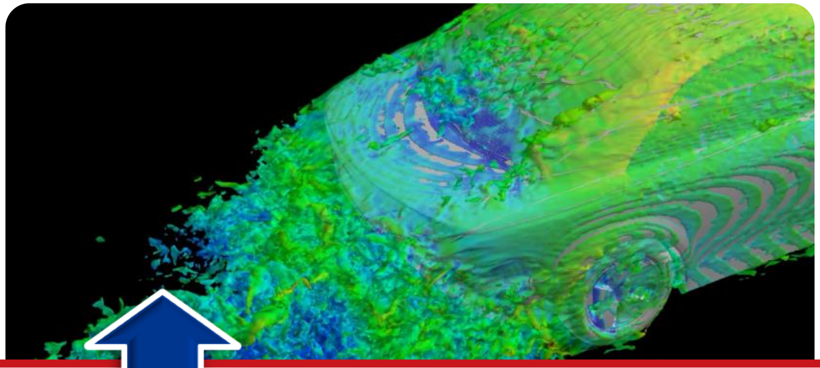
EXPECTED OUTPUT IN THIS PROJECT

Capability of Understanding Mouse Brain

CFD+AI Design framework Aerodynaic Drag Efficiency & Design Aestheics => Better EV Design [Tsubokura et.al.]

- Co-optimization Framework

Rapid Generation of CFD Mesh from Shape Data



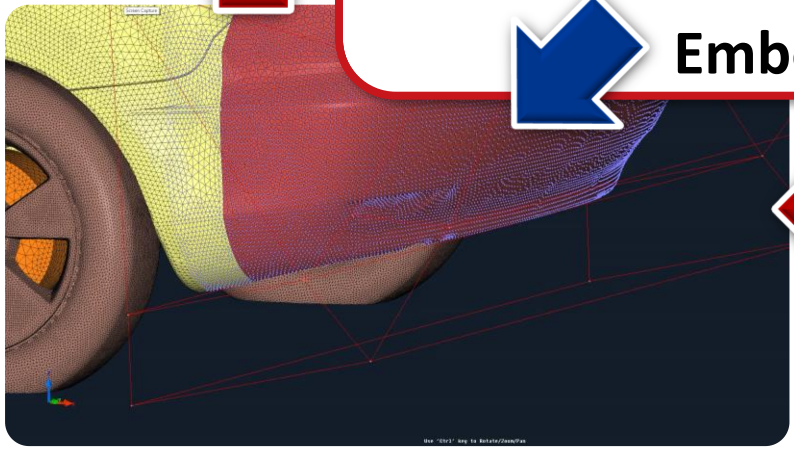
Supercomputer Fugaku

Ultra Fast Prediction of Drag via Digital Twin

<u>AI-Based Prediction and Optimization</u>

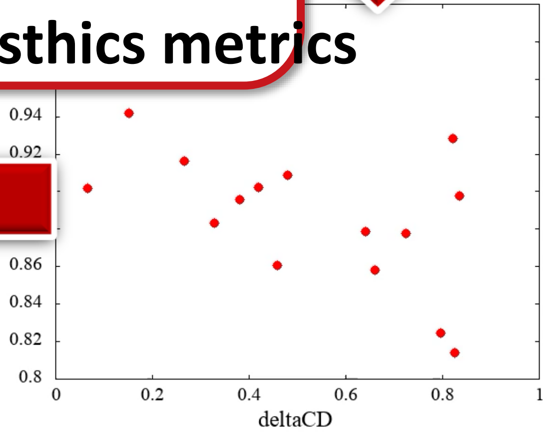
Embedding of human aesthics metrics

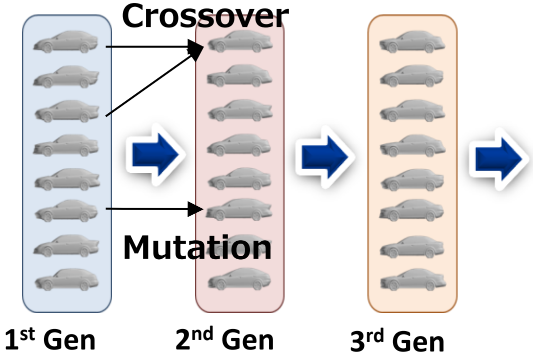
Drag + Aestheics



Parametric Shape Morphing

Shape Parameters on Aesthetics





GA Multi Paramter Optimization “CHEETAH/R”

NAGOYA UNIVERSITY

Towards Foundational Models for Structural Engineering
[Koji Nishiguchi](Nagoya-U/Riken R-CCS)

Innovating vehicle structure with a giant aluminum die-casting

30% weight reduction
40% manufacturing cost reduction

Giga-press (Tesla)

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

Magic3D （NVIDIA, 2022）

Shap-E （OpenAI， 2023）

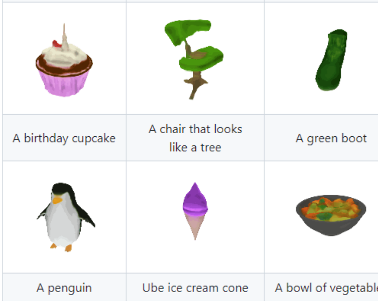
Rapid performance improvement of 3D generative AI

Nagoya University

Recent studies of 3D generative AI

- From 2022 onwards, not only **2D generative AI** but also **3D generative AI** have been emerging one after another.
  - —Lack of 3D datasets
  - **—No dataset that can be applied to structural mechanics has been proposed.**

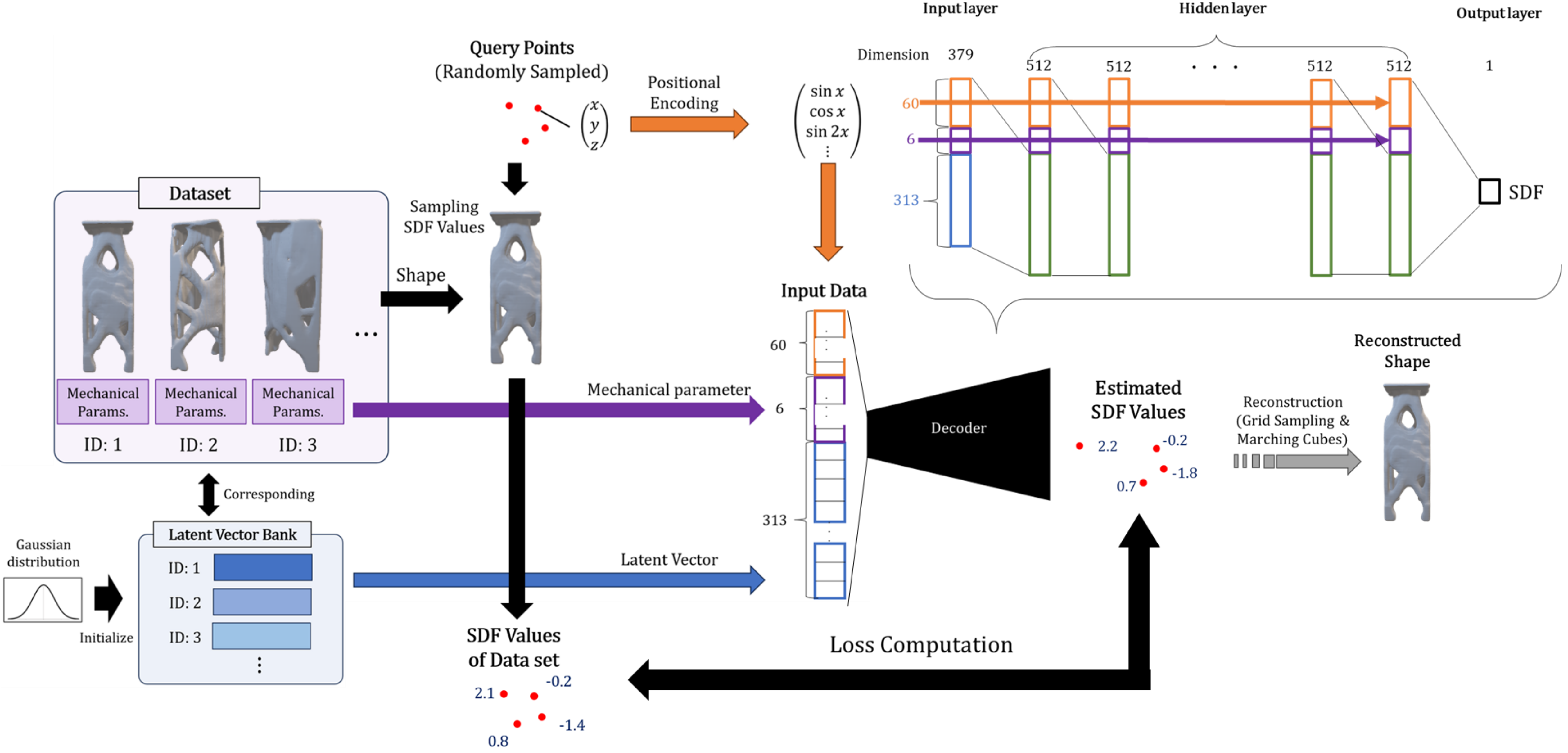
Model name	Release date	Research group	3D representation	Model architecture	Data set	Number of 3D data
Shap-E	May 2023	OpenAI	Implicit function	Transformer-based diffusion model	ShapeNet（3D）, WebImageText（2D）	Several millions
Point-E	December 2022	OpenAI	3D point cloud	Transformer-based diffusion model	ShapeNet（3D）, WebImageText（2D）	Several millions
Magic3D	November 2022	NVIDIA	3D mesh	NeRF, diffusion model	COCO（2D）, ImageNet（2D）	None
DreamFusion	September 2022	Google, UCB	Implicit function	NeRF, diffusion model	COCO（2D）, ImageNet（2D）	None



Shap-E

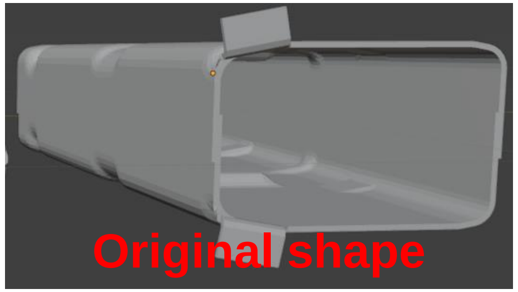
Magic3D

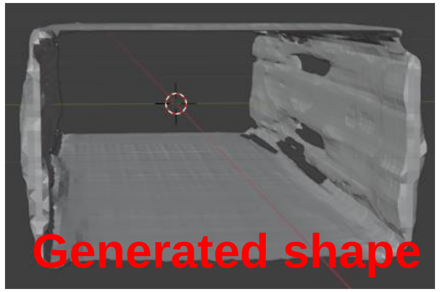
DeepSDF incorporating structural dynamics



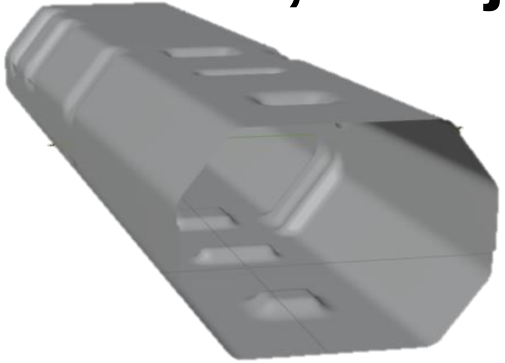
Parameter-to-3D foundation model

- **Future Challenges: Model for thin-walled structures**
  - —Almost all automotive structures and civil engineering structures are composed of thin-walled structures.
  - —In our present model, generating thin-walled structures is difficult.

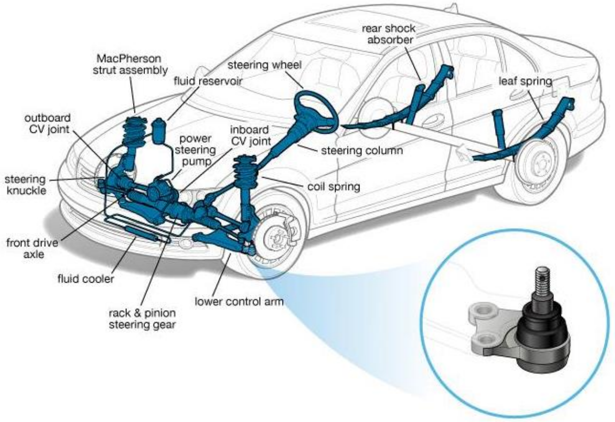




- **Future Challenges: Model for structures including local features (beads, spot welds, bolt joints)**



https://images.app.goo.gl/S7DhP33XuP58gnfJ6



https://www.cars.com/auto-repair/glossary/ball-joint/

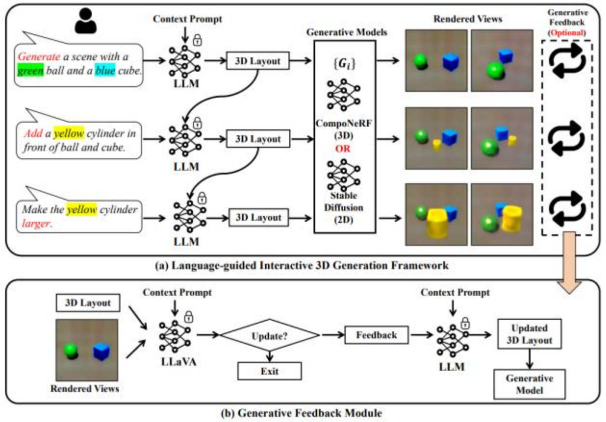
Final goal: Automation and democratization of structural design

NAGOYA UNIVERSITY

Human feedback by Non-experts

Designer **Marketer**

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



Natural language

Mechanical parameters

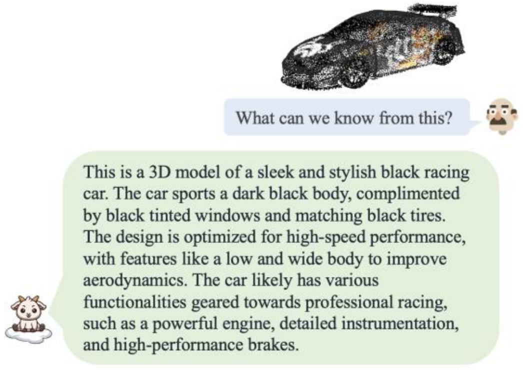
Paramter-to-3D model

3D structure

Natural language

Human feedback

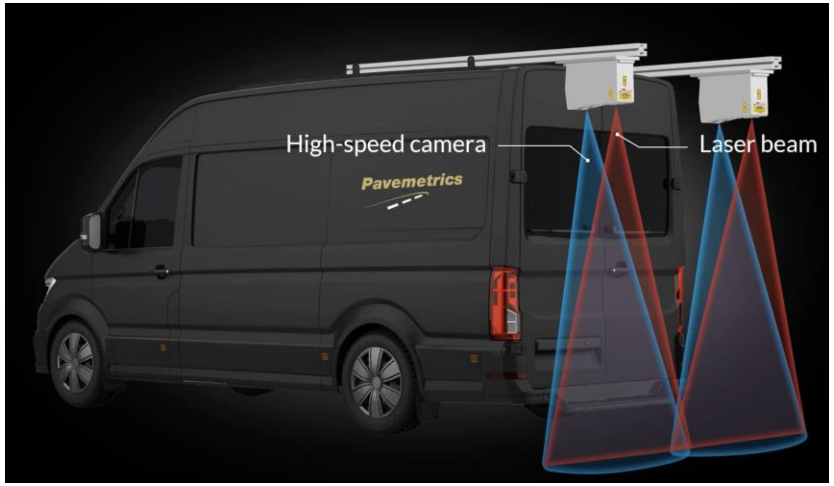
3D-to-text model
(LLMs to understand 3D structure)

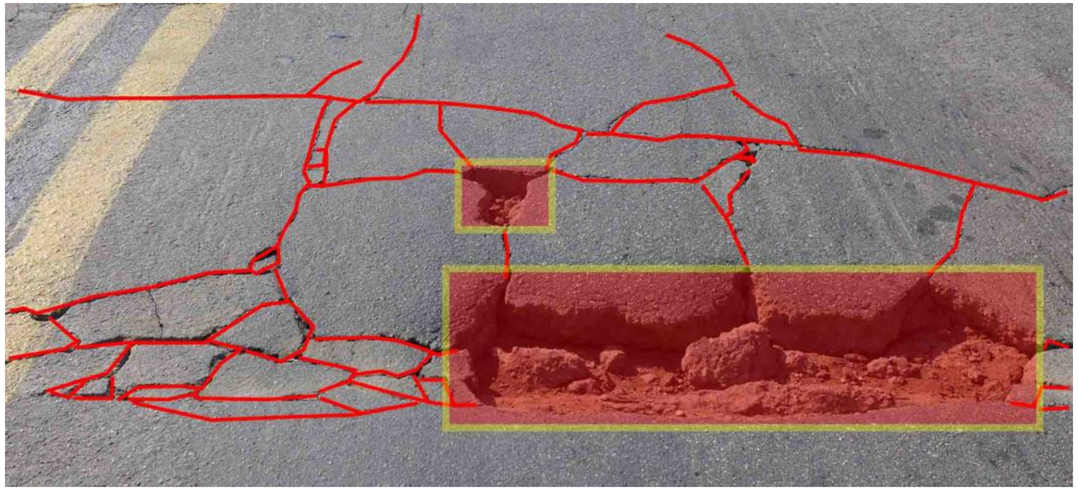


Another Real-world Problem:
How to Inspect Roads for Maintenance?

- Manual inspection
  - Time: O(Decades)
  - Cost: O($ Billions)

- Camera/laser Imaging technology
  - Good for fast screening of visible surface cracks, depressions etc
  - Not a reliable technology for understanding sub-surface conditions



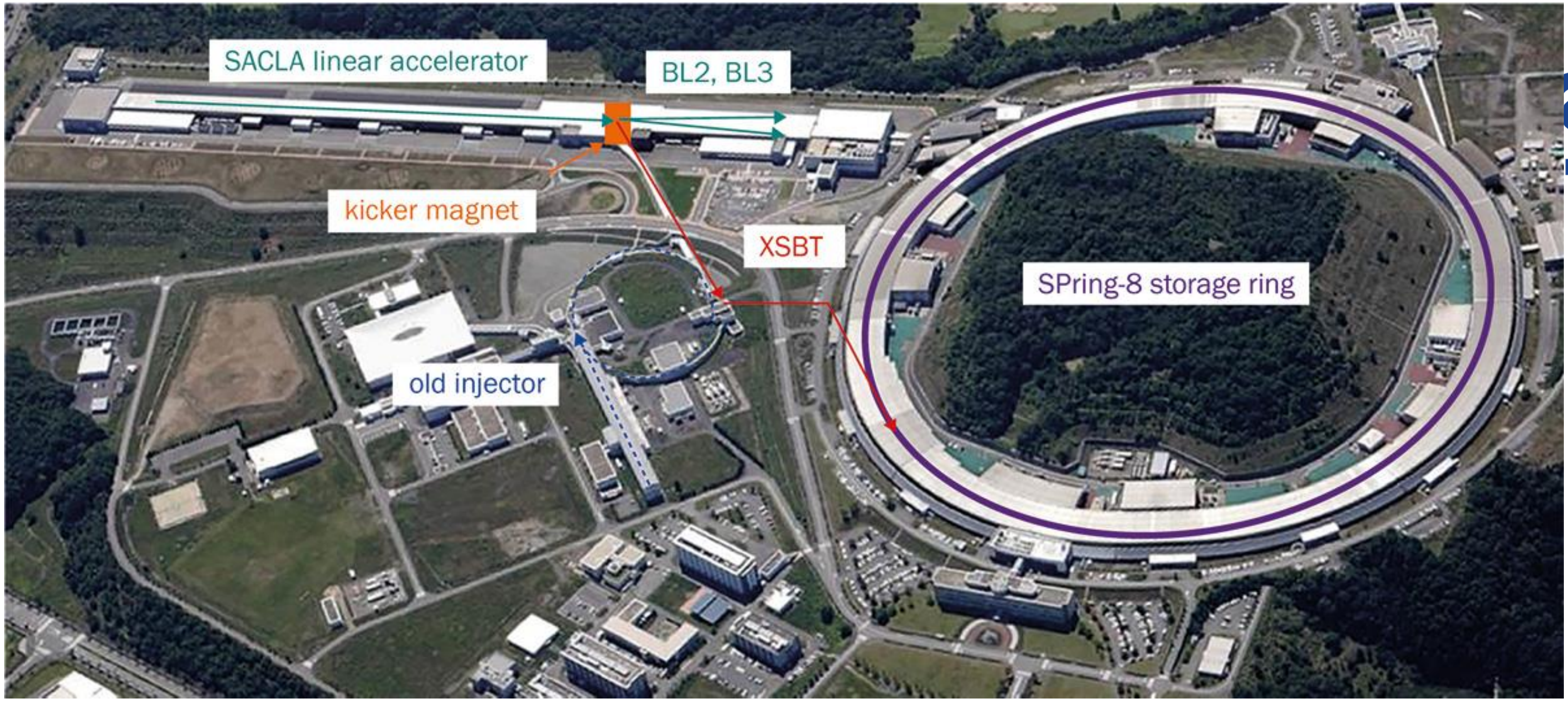


How?

- 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)
  - 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?



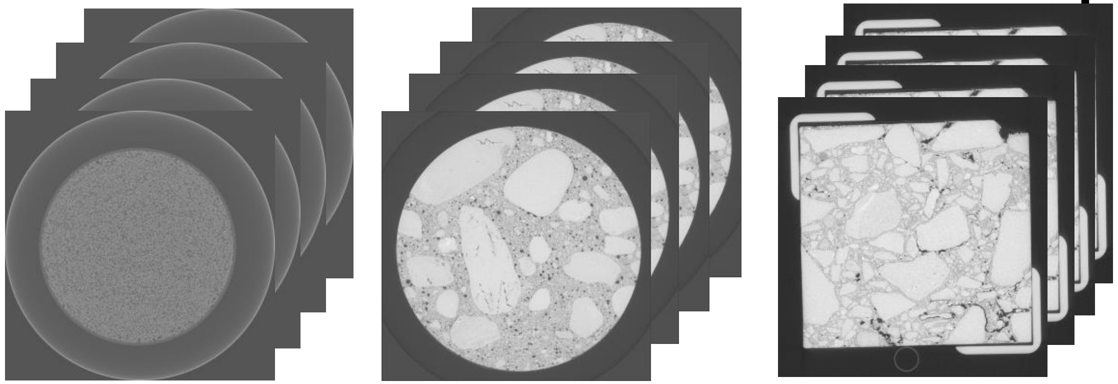
Riken Spring-8 + Sacla Synchrotron Light Source Facility

RIKEN
Center for
Computational Science

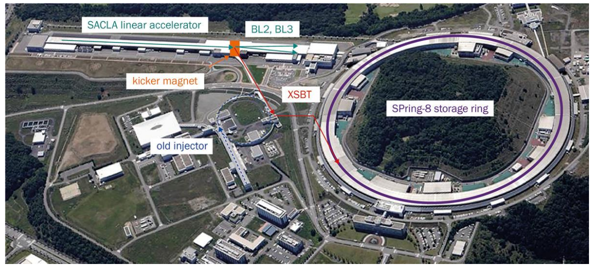
阪神高速
Hanshin Highway Co.

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

Sand particles (4Kx4Kx6K) | Concrete & Asphalt (6Kx6Kx11K) | Concrete (6Kx6Kx3K)

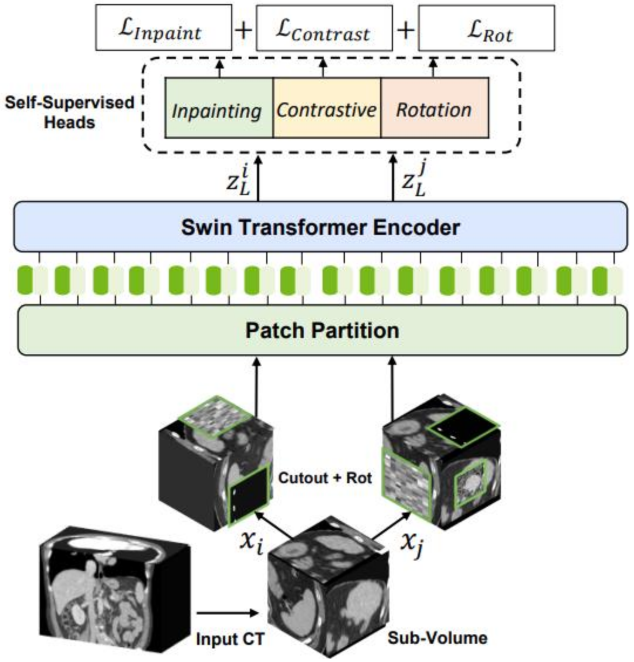


End-to-end High-resolution CT Powered by Supercomputing



+

Supercomputers
(Fugaku/ABCI/ Frontier/AWS)



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

State-of-the-art scale of resolution

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

Reconstruction + AI + Analytics

↓ Cost: O($ Billions)

TRIP-AGIS AI & HPC Infrastructure 2025

Extensive re-use of Existing Fugaku Assets=>FugakuNEXT

Current Fugaku Resources

HPC Supercomputer “Fugaku”

HPC: 163PetaBytes/s memory bandwidth (No.1 currently)

Foundation model training: 2 Exaflops FP16

Operational Power: 16~20MW

Inference to be enhanced exploiting world’s top mem BW

> 20Terabps

External Network> 3.2 Terabps
NTT IOWN, to Clouds,
Instruments, other SCs,
etc.

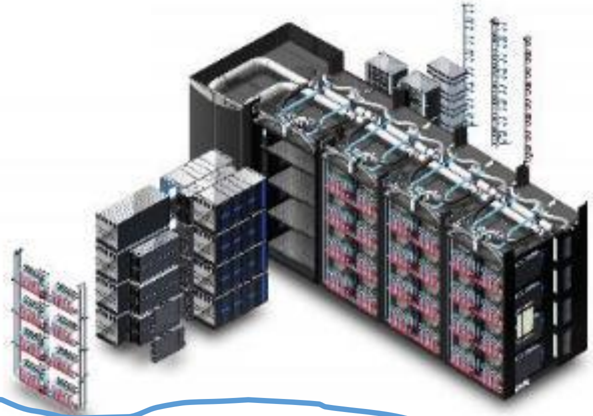
> 20Terabps

AI for Science Supercomputer Accelerator

AI Training 8+ Exaflops 8bits (4~5x Fugaku)

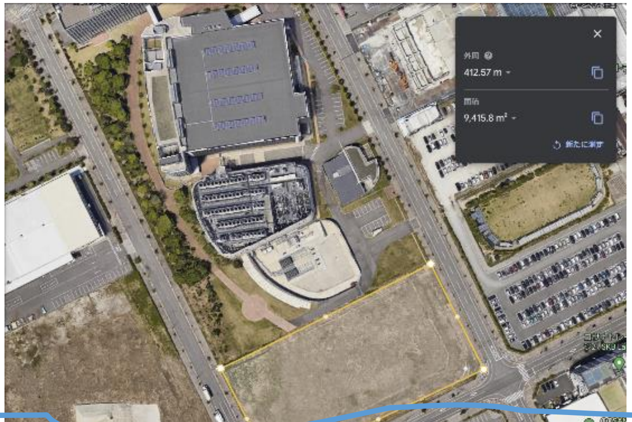
AI Inference 8+ Exaflops, 15PB/s Mem BW (1/10 Fugaku()

Operational Power 5~10MW (1/4 Fugaku)



Fugaku Storage: 150 PetaBytes (current)
Fujitsu FEFS-LUSTRE HDD PFS + NVMe

R-CCS DC Facility
> 40MW Power & Cooling



HPCI Wide Area Storage：>100 PetaBytes
Distributed FS GFARM, S3, etc.

MoU between DOE & MEXT on HPC (incl. AI) as well as ANL-Riken MOU on AI for Science April, 2024

DOE-MEXT

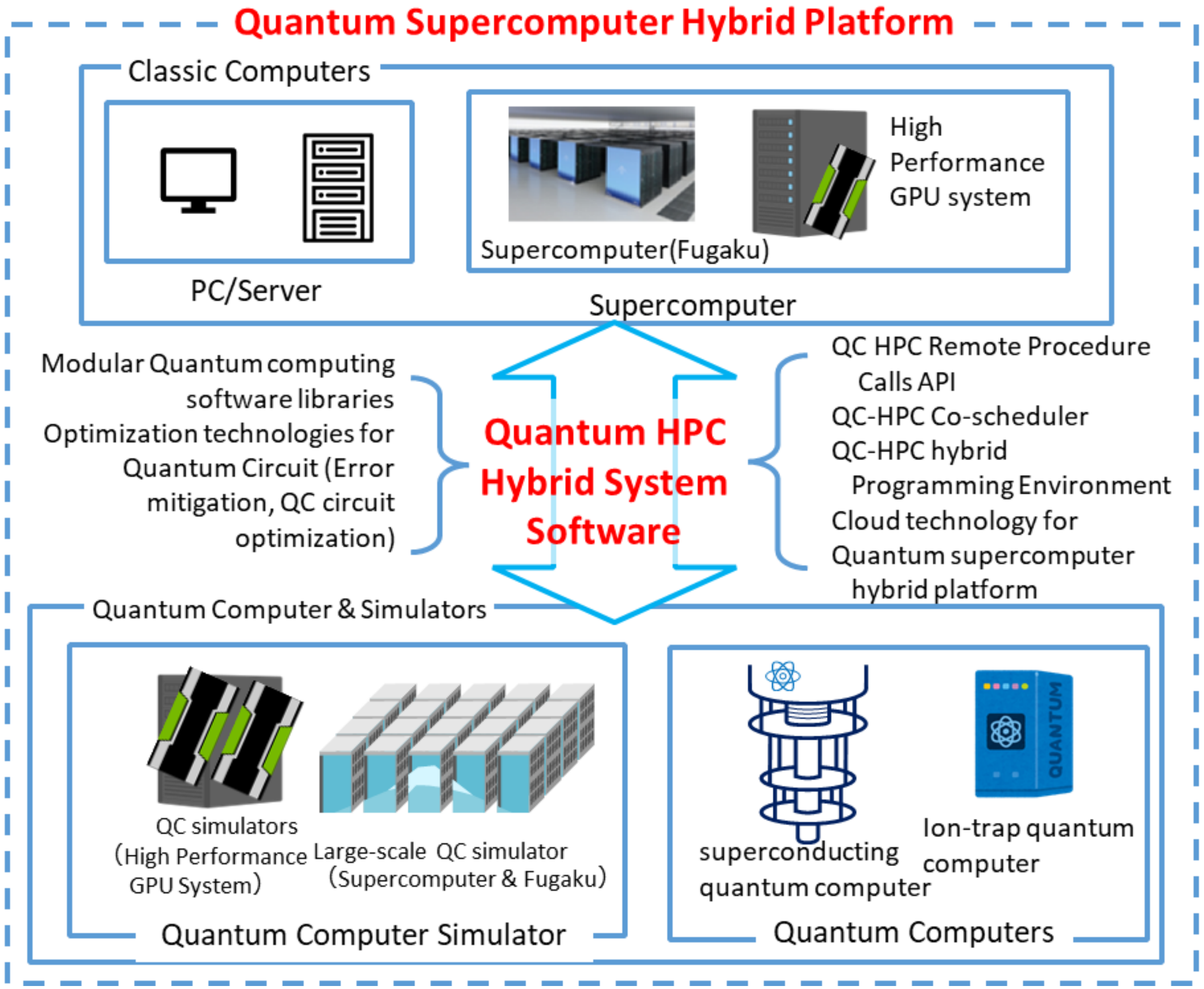
David Turk (DoE Deputy Secretary)
Masahito Moriyama (MEXT Minister)

ANL-Riken

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

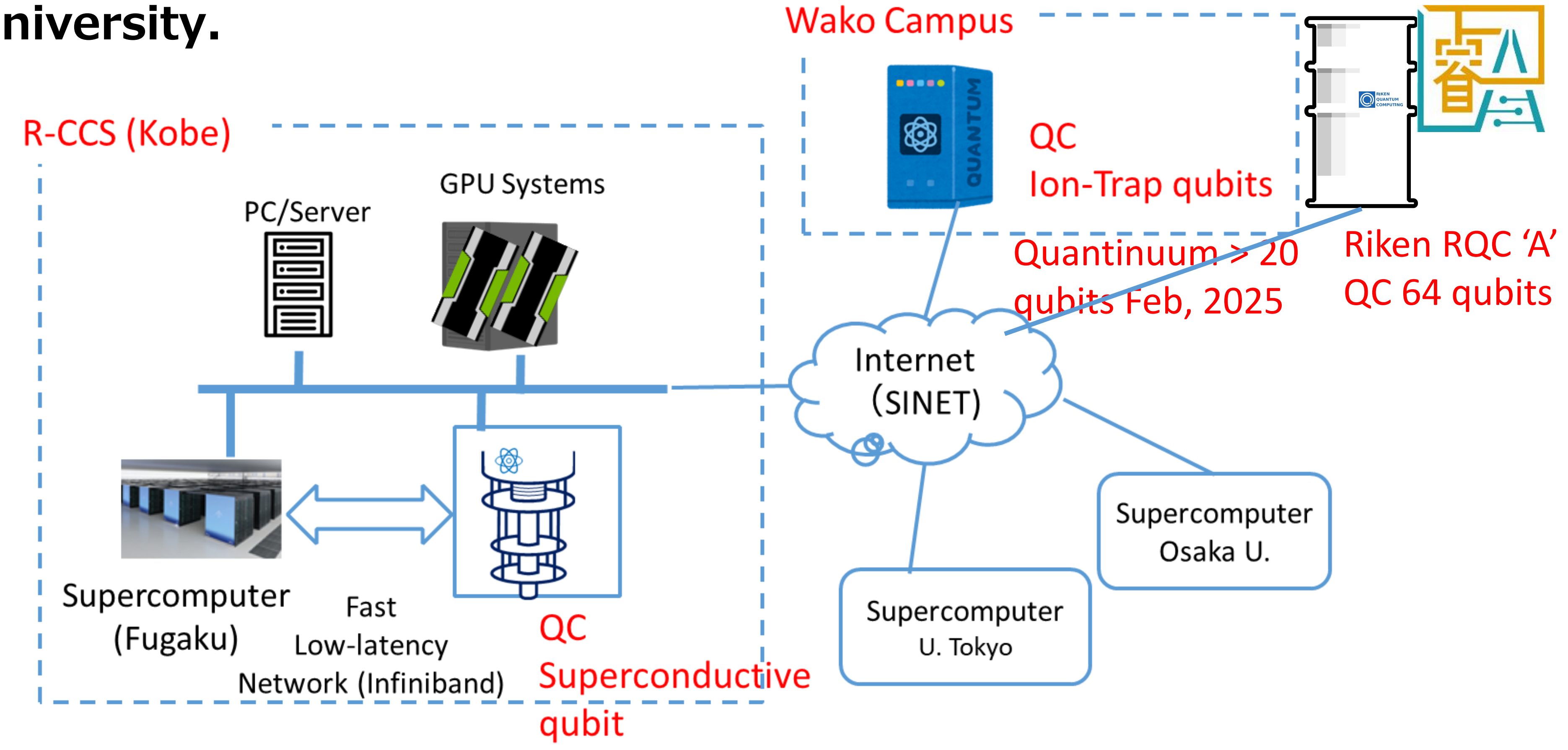
JHPC Quantum Project: R&D Topics

- **<u>Quantum HPC hybrid software:</u>** Development of system software for seamless and efficient use of quantum computers and supercomputers by coordinating computing resources optimally.
- **<u>Modular quantum software libraries:</u>** Developing modular software tailored to application fields and developing high-level software libraries for error mitigation and circuit optimization processing specialized to the characteristics of quantum computers. The software enables to develop advanced quantum applications by combining them as modules.
- **<u>Cloud computing technology for quantum supercomputer hybrid platform:</u>** Develop cloud infrastructure software to support the use of quantum applications for business development using quantum computer for post-5G era.

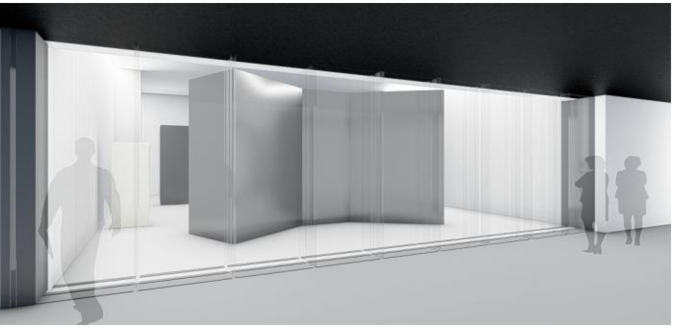


Overview of our QC-supercomputer hybrid platform

- **Two types of quantum computers with different characteristics will be installed at on-premises the RIKEN Center for Computational Science (Kobe) and (Wako). Planned quantum supercomputers hybrid platform consist of these quantum computers, Fugaku supercomputer, and supercomputers of the University of Tokyo and Osaka University.**



IBM > 100qubits
May-June 2025



JHPC quantum software structure and work package

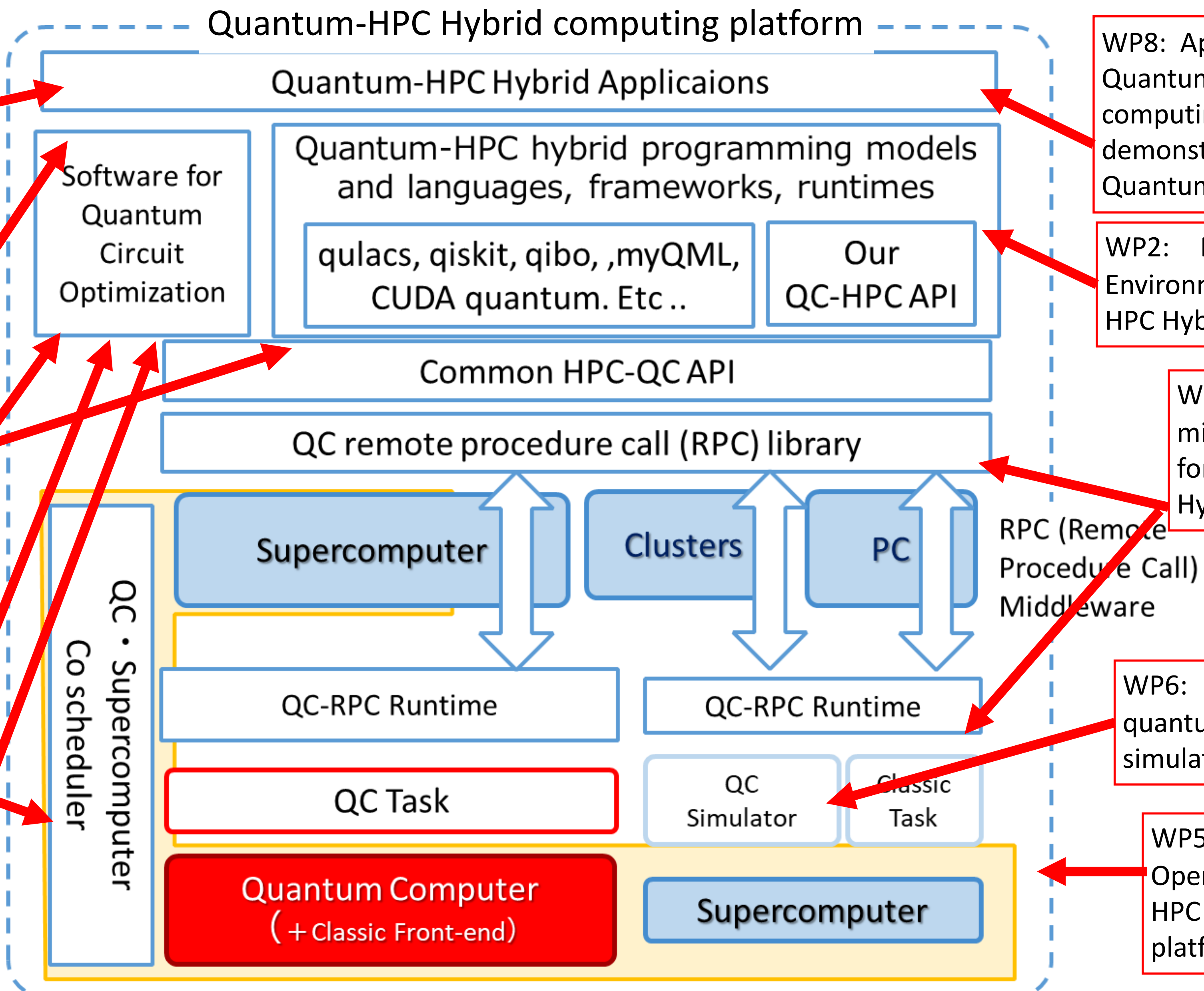
WP10: Promotion of practical applications of Quantum-HPC Hybrid computing

WP9: Cloud PaaS systems for Quantum-HPC Hybrid computing

WP4:modular quantum computing software libraries for Quantum-HPC Hybrid computing

WP7: Optimization algorithm and technique for Quantum-HPC Hybrid computing

WP3:Co-scheduler and cou pler middleware for Quant um-HPC Hybrid computing



WP8: Applications for Quantum-HPC Hybrid computing and demonstration of Quantum-HPC advantage

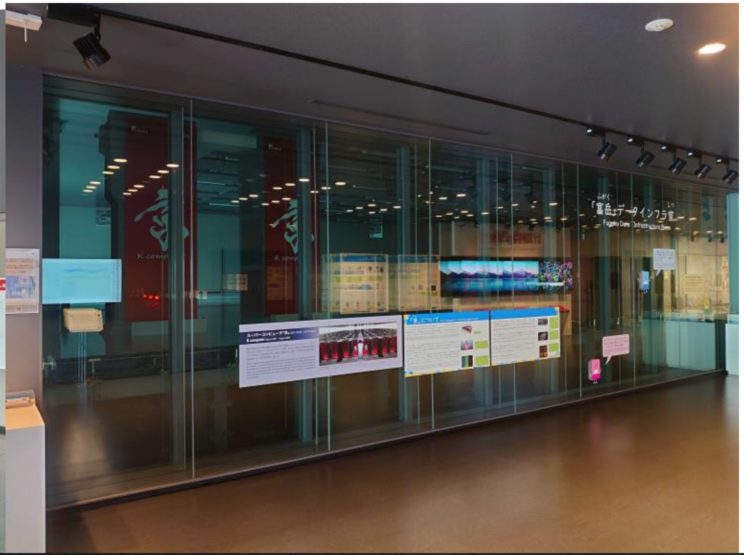
WP2: Programming Environment for Quantum-HPC Hybrid computing

WP1: RPC middleware and APIs for Quantum-HPC Hybrid computing

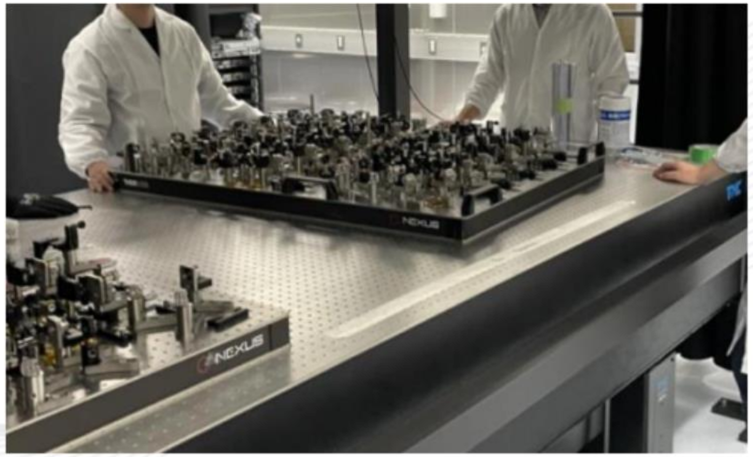
WP6: Advanced quantum computing simulators

WP5: Integration and Operation of Quantum-HPC Hybrid computing platform

IBM Quantum System 2 (Heron, 133Qubits) Installation Prep @ Riken R-CCS Kobe (Prodution by May 2025)



Quantinuum H1-2 Installation @ Riken Wako Campus (Production Feb 2025)



Why IBM and Quantinuum?

QC qubits	Characteristics	Targets
Superconducting Qubits (IBM and 'A'	Medium qubit count (100 qubits or more) Fast operating speed (a few ns). Medium Fidelity.	Development of utilization technology and system software for the utilization and practical use of large- and medium-scale NISQ machines.
Trapped Ion Qubits (Quantinuum)	High fidelity, the number of qubits is not large.(about 20 qubits). Slow operation speed (a few ms). Efficient all-to-all qubit operation.	Software development using small scale but high fidelity. Use of quantum computers with properties different from superconducting qubits.

- **System software for QC-HPC integration should be able to support different kinds of QCs.**
  - Quantum computers differ in their characteristics such as speed, fidelity, etc.
- **Superconducting quantum computers are reaching the scale of several hundred qubits. In order to aim practical use of QC including NISQ, we should explore use-case using large qubits for practical use.**

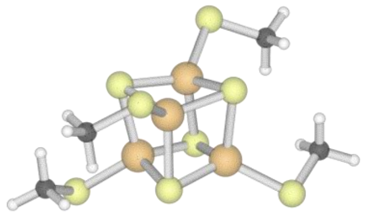
Quantum-centric Supercomputing for quantum chemistry

IBM Quantum RIKEN

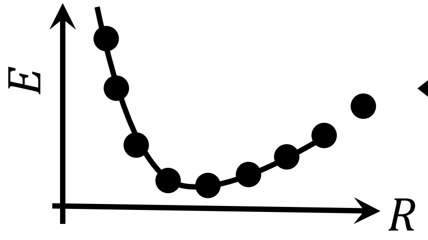
J. Robledo-Moreno et al., arXiv:2405.05068

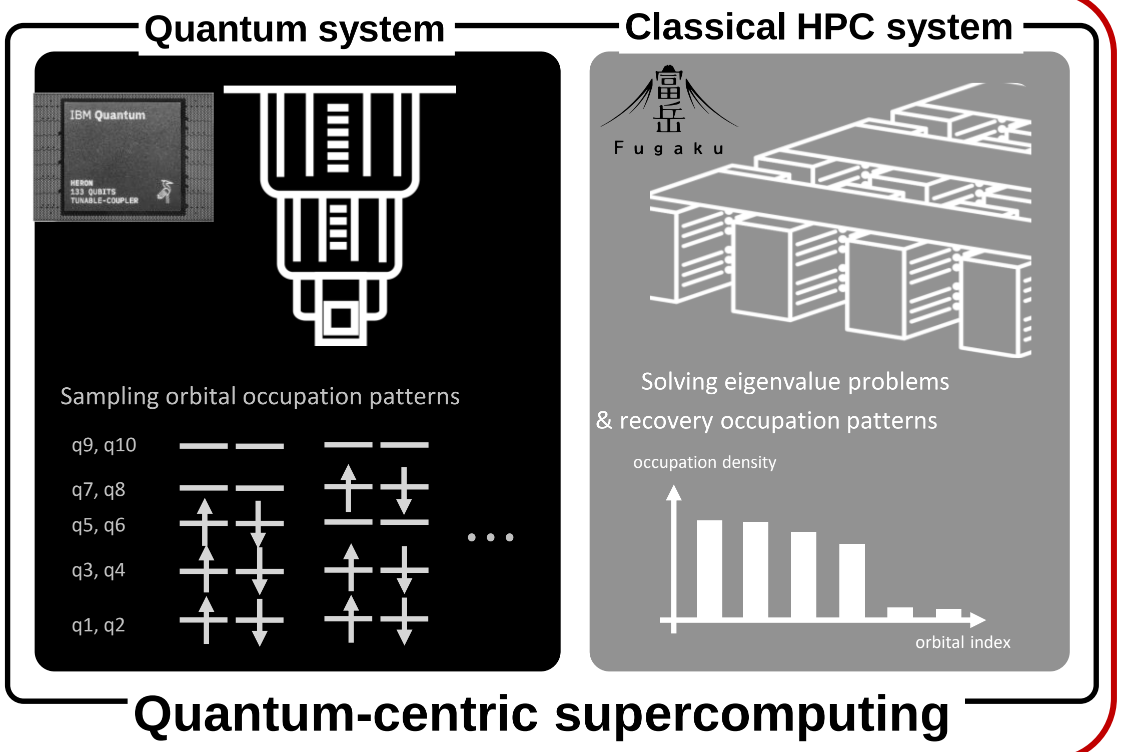
“**Chemistry Beyond Exact Solutions on a Quantum-Centric Supercomputer**” Although universal quantum computers are promising for predicting electronic structure problems in quantum chemistry, the deep circuits and huge amount of measurements required by current quantum computers make realistic quantum chemistry calculations difficult. In this study, **the 6400 nodes of the supercomputer "Fugaku"** are used to assist **IBM's latest quantum processor, Heron**, to study large molecules that cannot be handled by conventional quantum-classical hybrid calculations, and molecules that are difficult to calculate only by HPC-based classical computers (N_2 triple bond breaking and the electronic structure of iron-sulfur clusters), which are difficult to calculate using only HPC computers. As a result, it was shown that the combination of supercomputer and quantum processors **(quantum-centric supercomputing)** can provide good approximate solutions for practical quantum chemical calculations. In this study, the quantum circuits representing the quantum states of molecules were fixed, and large data were transferred only from the quantum computer to the supercomputer. For more accurate computation, future tasks include the improvement of quantum circuits by data transfer between the quantum computer and the supercomputer, and the development of algorithms on the classical computer side that are suitable for quantum-centric supercomputing.

target material

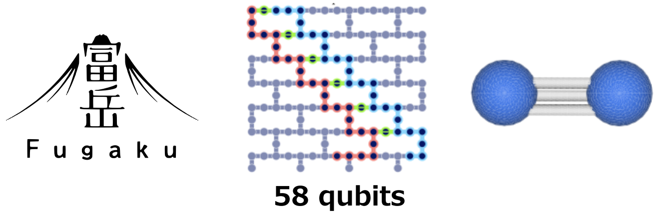


chemical properties

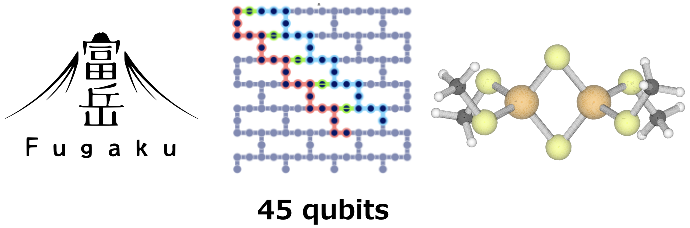




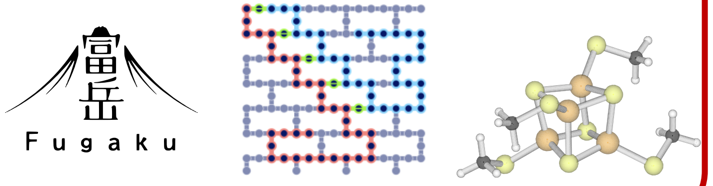
N_2 : Bond breaking on large basis set



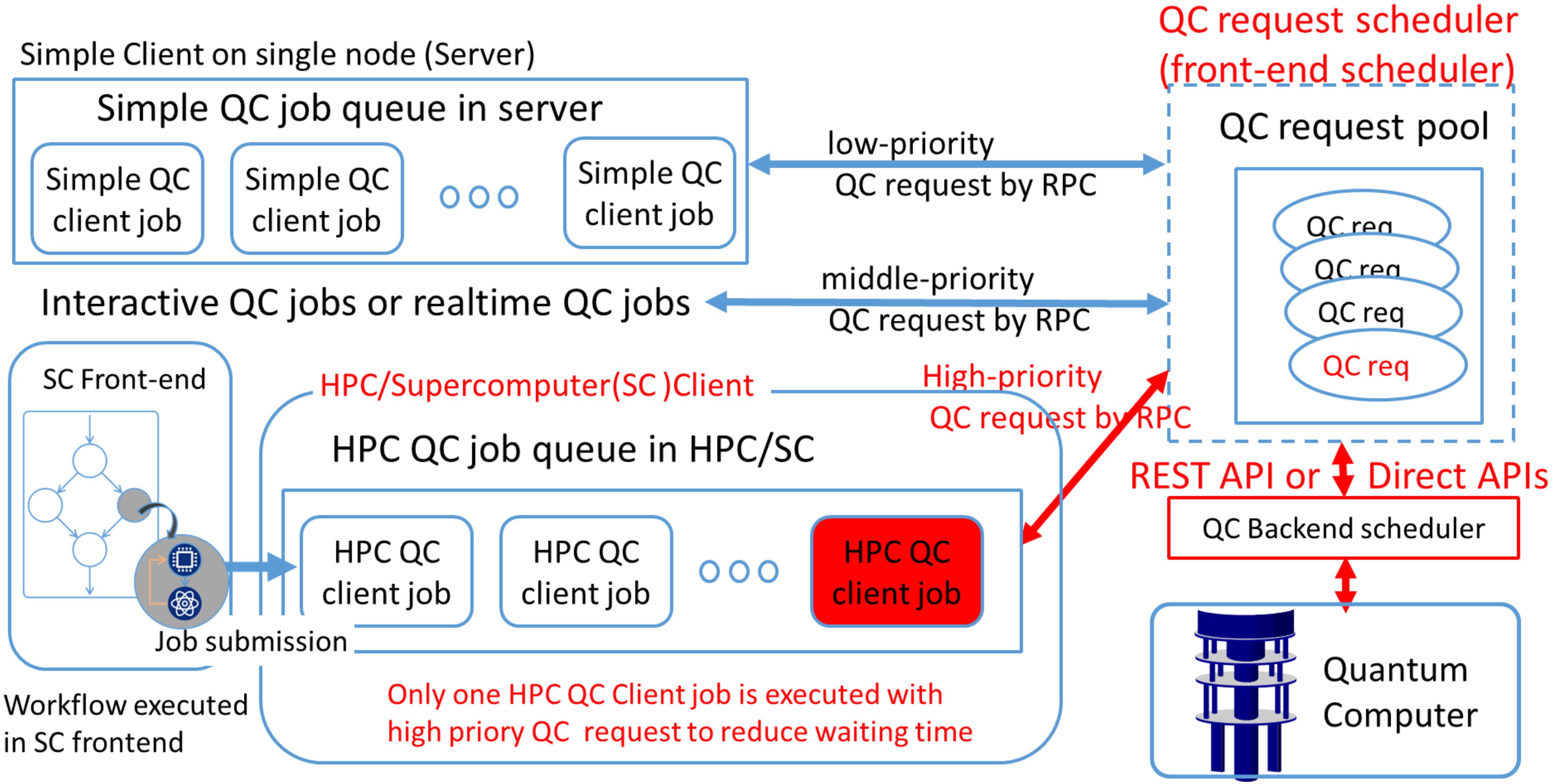
Fe_2S_2: Precision many-body physics



Fe_4S_4: Pushing hardware capabilities



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

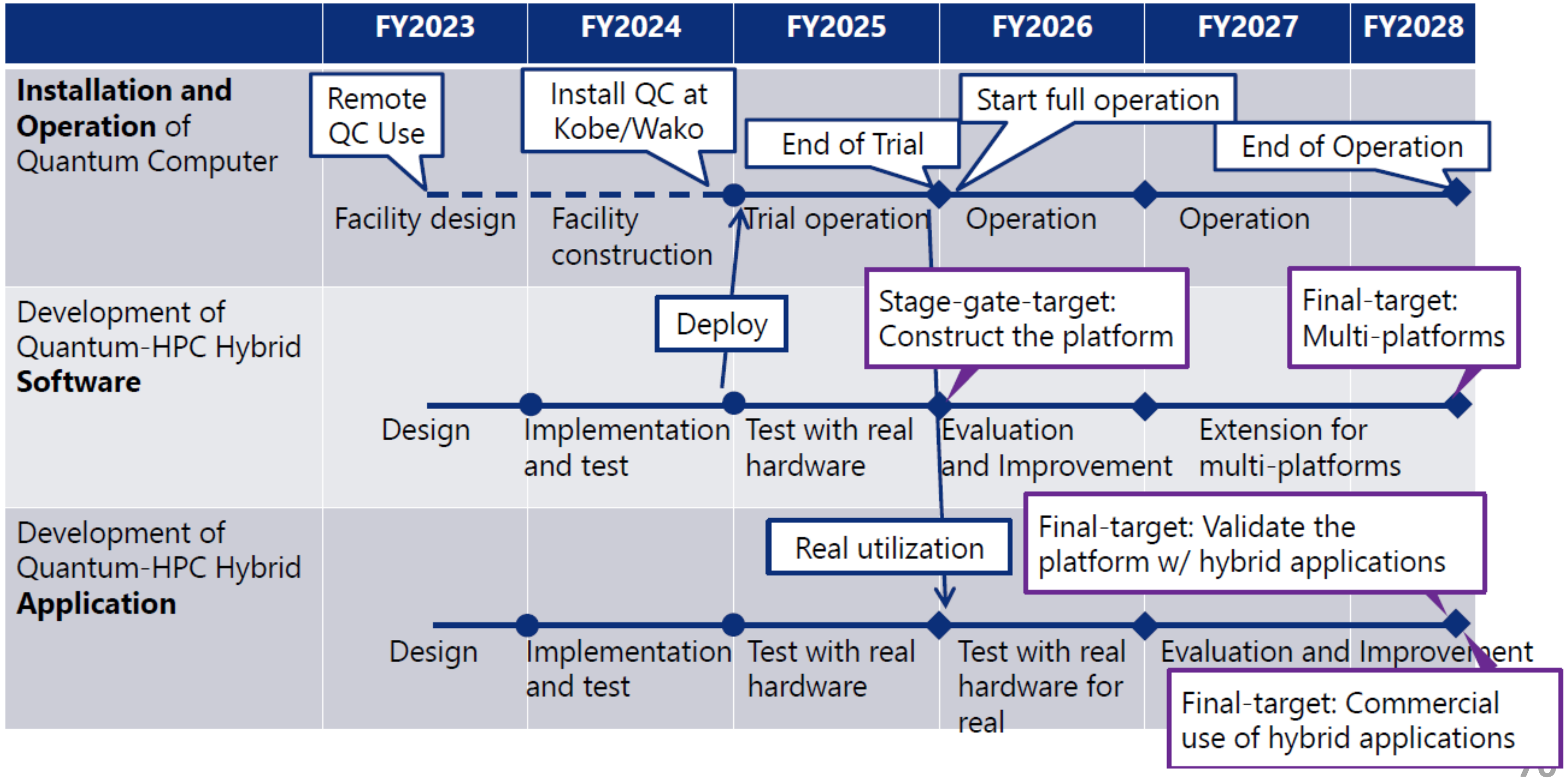


JHPC quantum project schedule

- Our project, **JHPC quantum**, was accepted and started from Nov. 2023.
- Installation of QC hardware in 2Q 2025
- In 1st Q of 2026, operation of the quantum supercomputer hybrid platform will be started and used to demonstrate the effectiveness of quantum and HPC hybrid applications in the later half of our project.

We will start “test-user program” to invite external users who are interested in QC-HPC hybrid computing.

International collaboration is welcome



FugakuNEXT Feasibility Study (Towards “Zetta-scale” AI&HPC)

Project Overview

The next-generation computational infrastructure is expected to become a platform for realizing SDGs and Society 5.0 by **providing advanced digital twins** that will bring "Research DX" in the science. Aiming to realize a versatile computing infrastructure that can **execute entire workflow by making full use of wide range of computational methods, such as simulation techniques, AI, and BigData** at scale, we conduct a holistic investigation on architecture, system software and library technologies through co-design with applications.

As a basic principle of system design, we **practice the "FLOPS to Byte" concept** from architecture development to algorithm or application design to streamline data transfer and computation under power constraints, while taking necessary computing accuracy into consideration. Under the ALL JAPAN team composition, we will investigate system configurations and elementary technologies which improve effective performance of the next-generation computing infrastructure.

Subject of Investigation

Research on Architecture

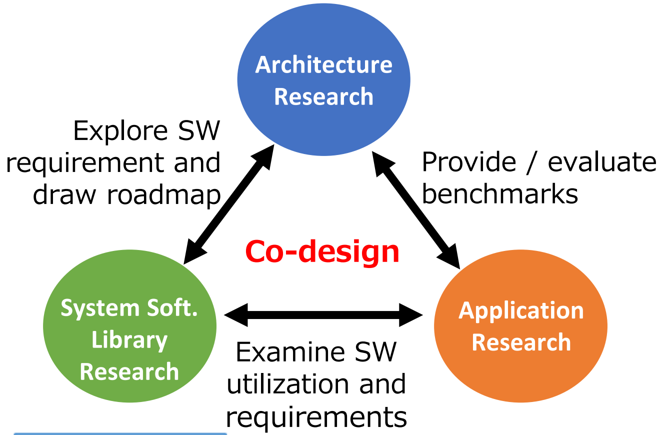
- Investigating technological possibilities (such as 3D stacked mem, accelerators, chip-to-chip direct optical link) and performance of the entire system or its components based on trends in semiconductor and packaging technologies
- Predicting future system performance based on performance analysis of benchmark sets provided by Application Research Group, and feeding back to next-generation application development

Research on System Software and Library

- Drawing roadmap for future system software development in Japan, specially considering data utilization enhancement, integration of AI technology with first-principles simulation, real-time data processing, and assurance of high security

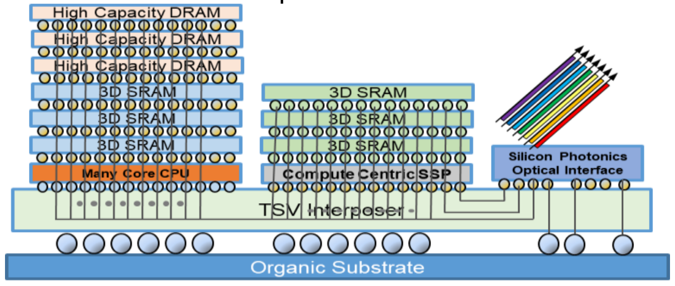
Research on Applications

- Building a broad benchmark set to evaluate multiple architecture choices while considering improvements in algorithms and parameters of application based on the results of architectural evaluations and exploratory "what-if" performance analysis
- Investigating what classes of algorithms are expected to evolve significantly for future systems



Investigation Schedule

	2022 H2	2023 H1	2023 H2	2024 H1	2024 H2
Architecture	Explore device/arch technology		Performance estimation with benchmarks		Arch selection and their R&D
System Software		Examine existing SW and its utilization	Identify requirement of SW development		Draw roadmap
Application		Examine existing apps and benchmark design	Perf. analysis by benchmark evaluation		Study for target science



Strawman processing element architecture

Organization Chart of System Research by RIKEN

System Research Team
(Representative Institution) RIKEN R-CCS
【PI: M. Kondo, AD: S. Matsuoka(R-CCS)】

GL: Group Leader
AD: Advisor
SGL: Sub Group Leader

Architecture Research Group

Architecture Research Group
RIKEN R-CCS
【GL: Sano, Co-GL: Miwa (UEC), AD:Amano (Keio)】

- **Architecture Research sub-G1** — RIKEN BDR 【SGL: Taiji (RIKEN BDR)】

System Software and Library Research Group

System Software and Library Research Group
RIKEN R-CCS
【GL: Sato,Co-GL:Katagiri (Nagoya-U), Sato (TUT), AD: Sato】

Support on Group Management
Nagoya Univ. (Collaborator)
【Delegate: Katagiri】

- **Scheduler / Runtime sub-G** — Tohoku Univ. (Co-I institution) 【SGL: Takizawa】
- **IO / Storage / Filesystem sub-G** — Univ. Tsukuba (Co-I institution) 【SGL: Tatebe】
  - **Storage Archi Pattern Investigation** — DDN Japan (Collaborator) 【Delegate: Hashizume】
- **OS / Virtualization / Cloud sub-G** — National Institute of Informatics (Collaborator) 【SGL: Takefusa】
- **HPC Env. Usage Investigation sub-G** — Osaka Univ. (Co-I institution) 【SGL: Date】
- **Communication Library sub-G** — Kyushu Univ. (Co-I institution) 【SGL: Nanri】
- **Compiler / Programming-model sub-G** — RIKEN 【SGL: Tsuji】
- **Numerical Library sub-G** — RIKEN 【SGL: Imamura】
- **AI Framework sub-G** — RIKEN 【SGL: Mohamed】

Application Research Group

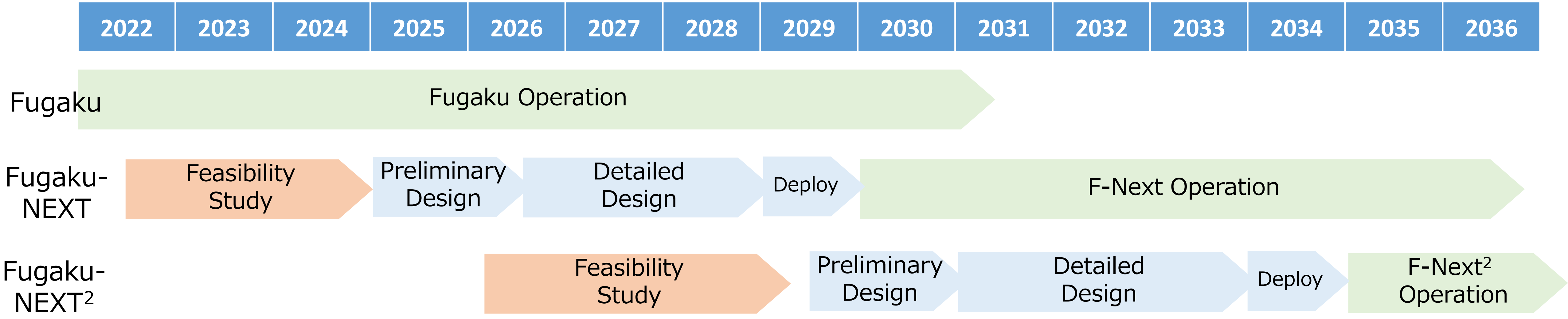
Application Research Group
Hokkaido Univ. (Co-I Institution)
【GL: Iwashita, Co-GL :Takahashi (U. Tsukuba), Fukazawa (Kyoto U.), AD: Nakajima / Tomita (R-CCS)】

Support on Group Management
Kyoto Univ. (Collaborator)
【Delegate: Fukazawa】

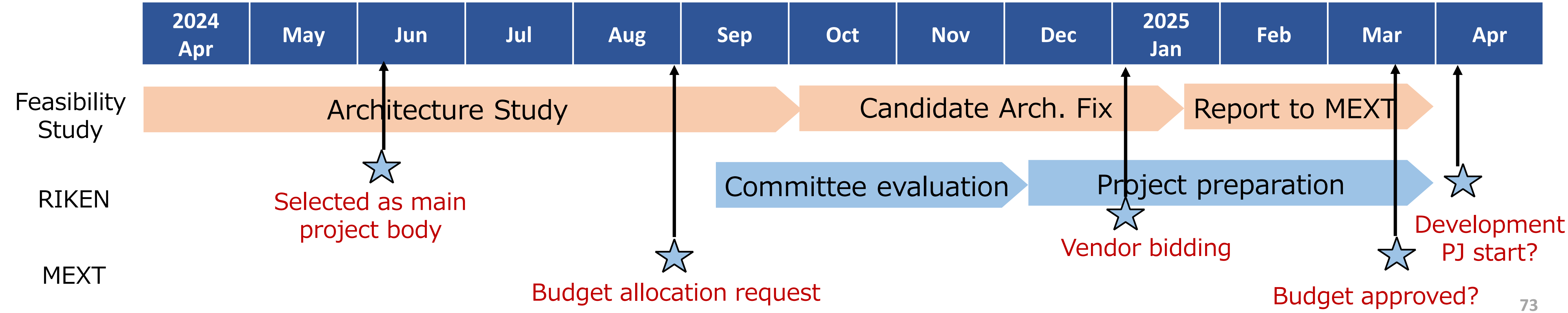
- **Life Science App. Area sub-G** — Yokohama City Univ. (Co-I institution) 【SGL: Terayama】
- **Material and Energy App. Area sub-G** — NIMS (Co-I institution) 【SGL: Yamaji,Co-SGL:Fukushima(UTokyo) 】
- **Weather/Climate Sci. App. Area sub-G** — JAMSTEC (Co-I institution) 【SGL: Kodama】
- **Disaster Prevention App. Area sub-G** — Univ. Tokyo (Co-I institution) 【SGL: Fujita】
- **Manufacturing App. Area sub-G** — RIKEN 【SGL: Onishi】
  - **Support on Manufacturing Apps** — JAXA (Collaborator) 【Delegate: TBA】
- **Fundamental Science App. Area sub-G** — RIKEN 【SGL: Aoki】
  - **Support on Space / Planet Sci. Apps** — NAOJ (Collaborator) 【Delegate: Takiwaki】
- **Social Science App. Area sub-G** — RIKEN 【SGL: Umemoto】
- **Digital-twin / Society5.0 App. Area sub-G** — Univ. Tokyo (Co-I institution) 【SGL: Shimokawabe】
  - **Support on Digital-twin Apps** — Japan Atomic Energy Agency (Collaborator) 【Delegate: Onodera】
- **Weather Model Perf Analysis sub-G** — RIKEN 【SGL: Kodama】
  - **Support on Weather Model Analysis** — Meteorological Research Institute (Collab) 【Delegate: Eito】
- **Computational Science Algorithm sub-G** — Univ. Tsukuba (Co-I institution) 【SGL: Takahashi】
- **Machine Learning Algorithm sub-G** — TiTech (Co-I institution) 【SGL: Yokota】
- **Benchmark Construction sub-G** — RIKEN 【SGL: Murai】
- **Performance Modeling sub-G** — RIKEN 【SGL: Domke】

Expected Timeline of Fugaku-NEXT R&D and Future Plan

● Expected schedule



● What's going on in FY2024 for Fugaku-NEXT development



Organization for FugakuNEXT Development

In order to promote research and development of Japanese new flagship supercomputer, “Next-Generation HPC Infrastructure Development Division (tentative name)” will be established at the RIKEN Center for Computational Science (R-CCS) in April 2025. This division will coordinate and promote the development effort for the next-generation flagship supercomputer system collaborating with research organizations both within and outside R-CCS.

RIKEN

President

Makoto Gonokami (Ph.D.)

Center for Computational Science

Director

Satoshi Matsuoka (Ph.D.)

Next-Generation HPC Infrastructure Development Division

Division Director Masaki Kondo (Ph.D.)

Next-Generation HPC Infrastructure System Development Unit

Development of architectures, system software and other systems related to next-generation HPC infrastructures

Kentaro Sano (Ph.D.)

Kento Sato (Ph.D.)

Next-Generation HPC Application Development Unit

Develop and support applications related to the next-generation HPC infrastructure, and study co-design (co-design) using these applications.

Yoshiyasu Aoki (Ph.D.)

Other Units / Project Management Office

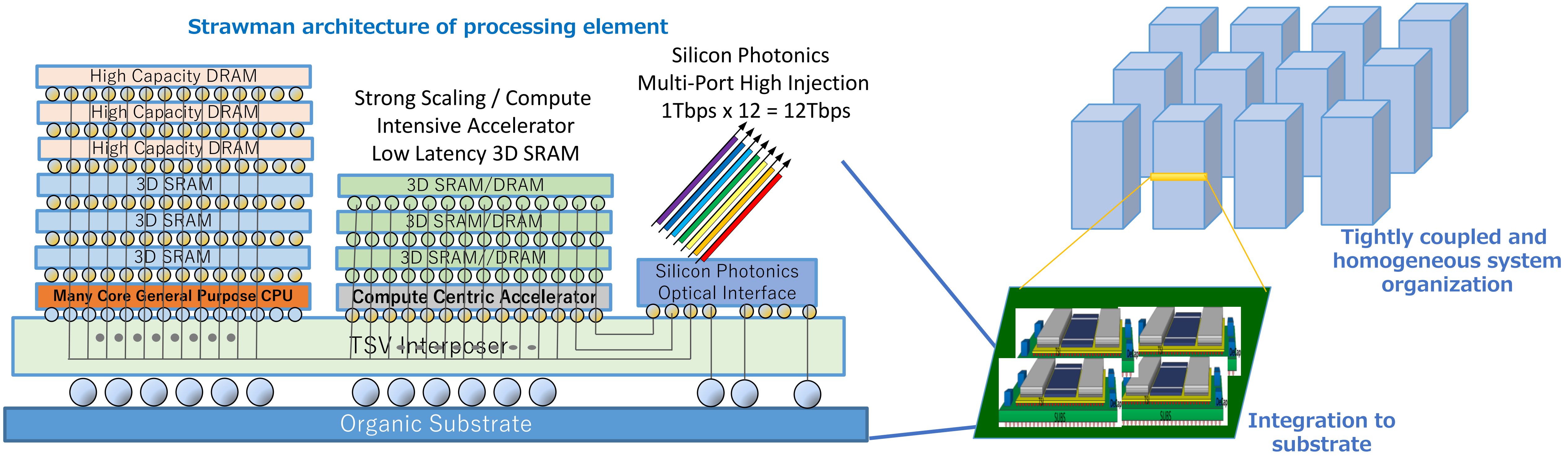
System Performance Requirement in RFP

- **Performance requirement for FugakuNEXT 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 node and storage)	

A Direction toward Next-Generation Computational Infrastructure

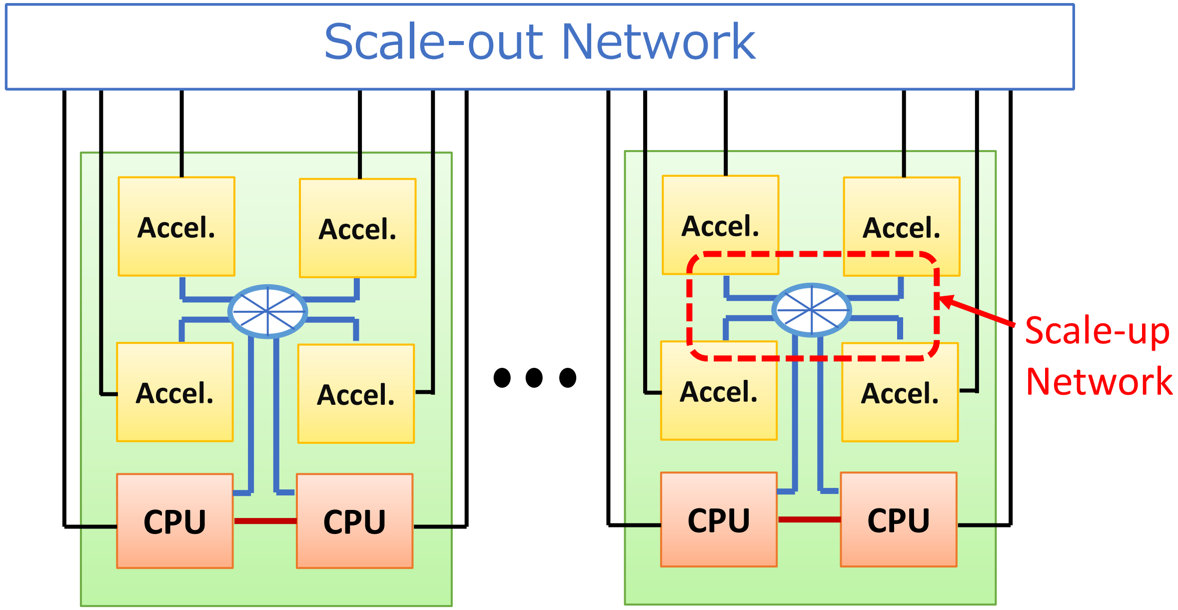
- **Initial vision of architectural directions**
  - Paradigm shift in architecture-algorithm toward “FLOPS to Byte (data movement efficiency)”
  - Significant increase in relative memory bandwidth using 3D stacked memories and processors
  - Silicon photonics to ensure high bandwidth for remote memory accesses
  - Ensure execution efficiency in strongly scaled problems with low latency execution, etc.



“3D stacked memory” & “Photonics” technologies: Post-Fugaku technology driver

Example Node Architecture for the AI-for-Science Machine

System Architecture for AI-for-Science Computing Infrastructure



- **System network which is good for both strong/weak scaling**
  - Combination of scale-up/scale-out NW
- **Having more than 10K accelerator sockets in the system**
  - NW among accelerator sockets

- **Heterogeneous node architecture**
  - CPU + GPU architecture
  - Tentatively 2-CPU and 4-GPU configuration
    - Subject to Scale-up/Scale-out and chiplet integration technologies
  - High BW with advanced memory technology
- **Scale-up NW (intra-node socket NW)**
  - P2P or switched connection w/ UALink
- **Scale-out NW (inter-node NW)**
  - Fat-tree topology w/ Ultra-Ethernet

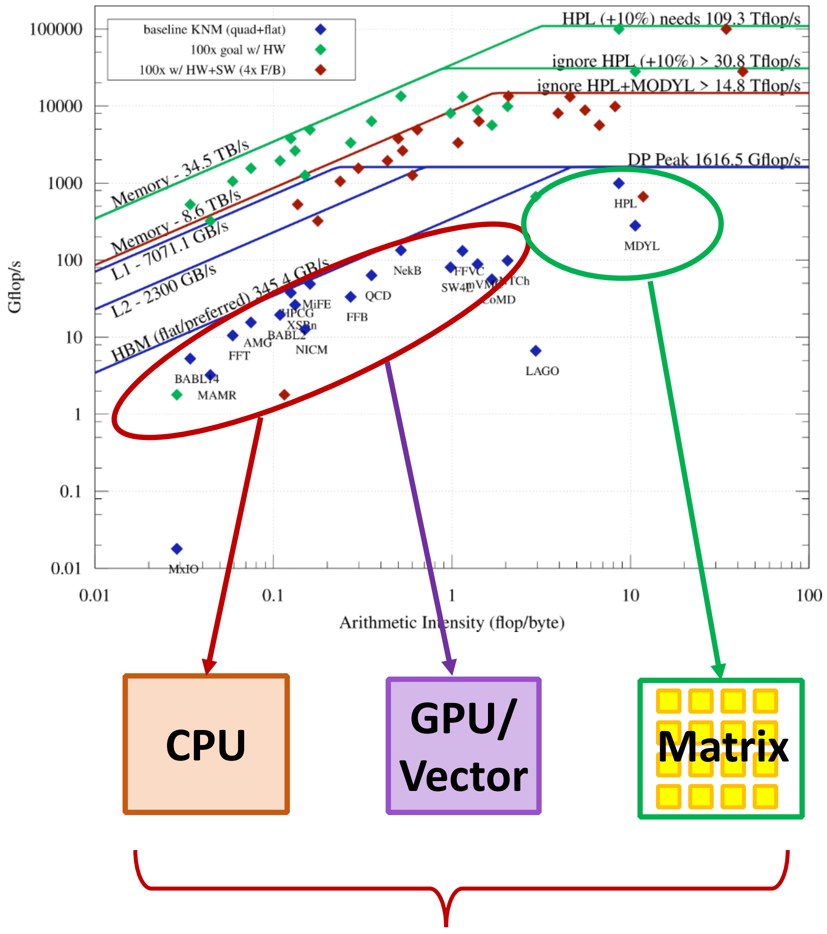
System target: More than 5-10x effective performance improvement in HPC applications and more than 50EFLOPS AI training performance (needs Zetta-scale low-precision arithmetic perf.)

Key Research Item for Node Architecture Selection

- **Needs for a power-efficient compute node**
  → Exploration of accelerators
  - Truly useful accelerator for HPC and AI workloads
  - HPC→Memory bound, AI→Compute & Memory bound
- **Characteristics of current processing element**
  - CPU: high generality, low-latency, low compute density
  - GPU (SP): vector processing, middle compute density
  - Matrix: dedicated for dense algebra, high compute density
    (ex. Tensor core, XMM, SME, AMX, TPU, CGRA, ···)
- **What to study in node architecture exploration**
  - What and how to integrate them
  - Effective memory bandwidth + data movement with high programming productivity

Quantitative benchmarking analyses is necessary

Roofline analysis on A64FX



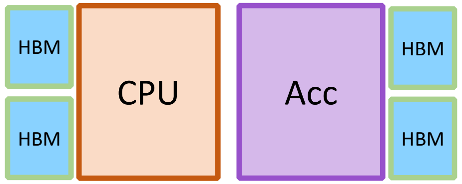
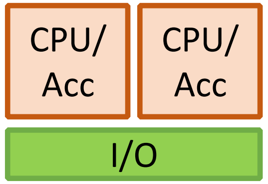
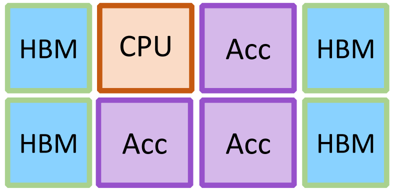
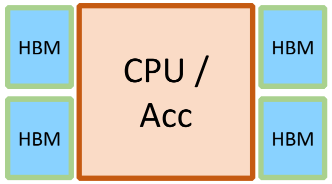
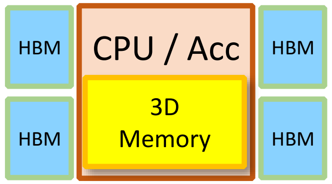
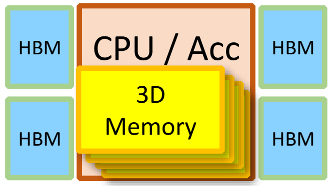
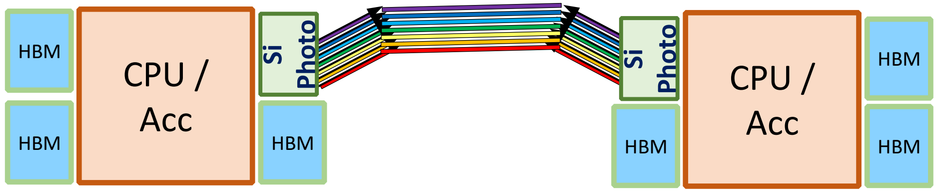
Need to find the optimal balance

Implementation Approaches for Node Architectures

- **Candidates of packaging technologies**

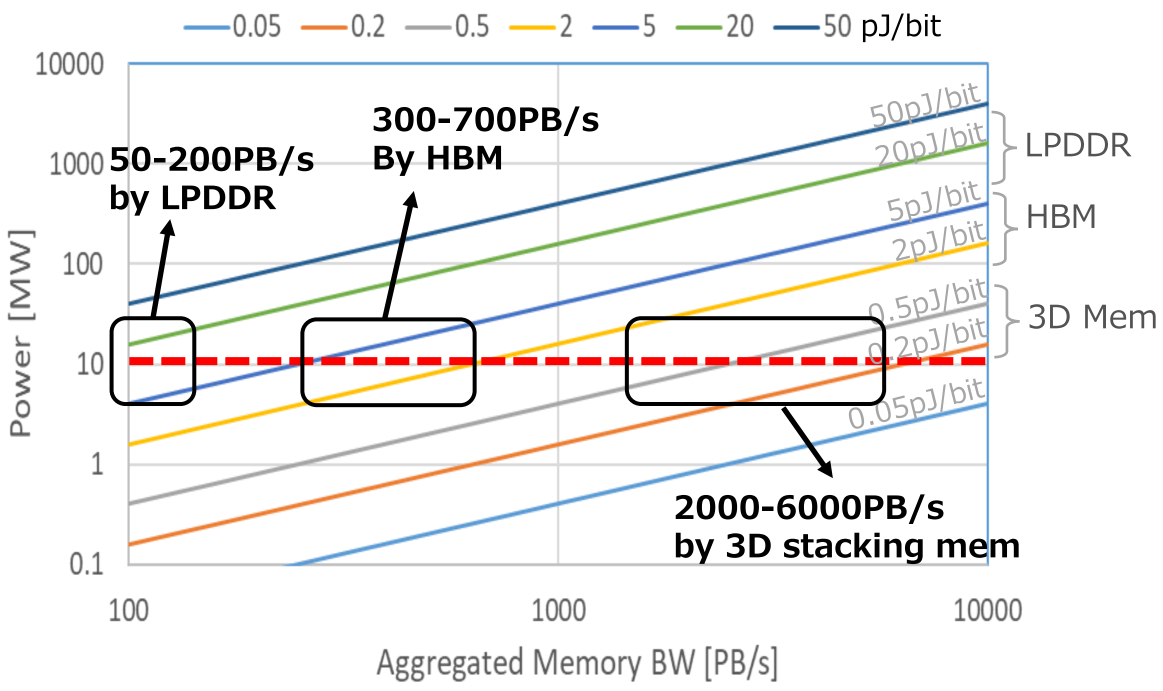
Technical difficulty: Low ⟵⟶ High

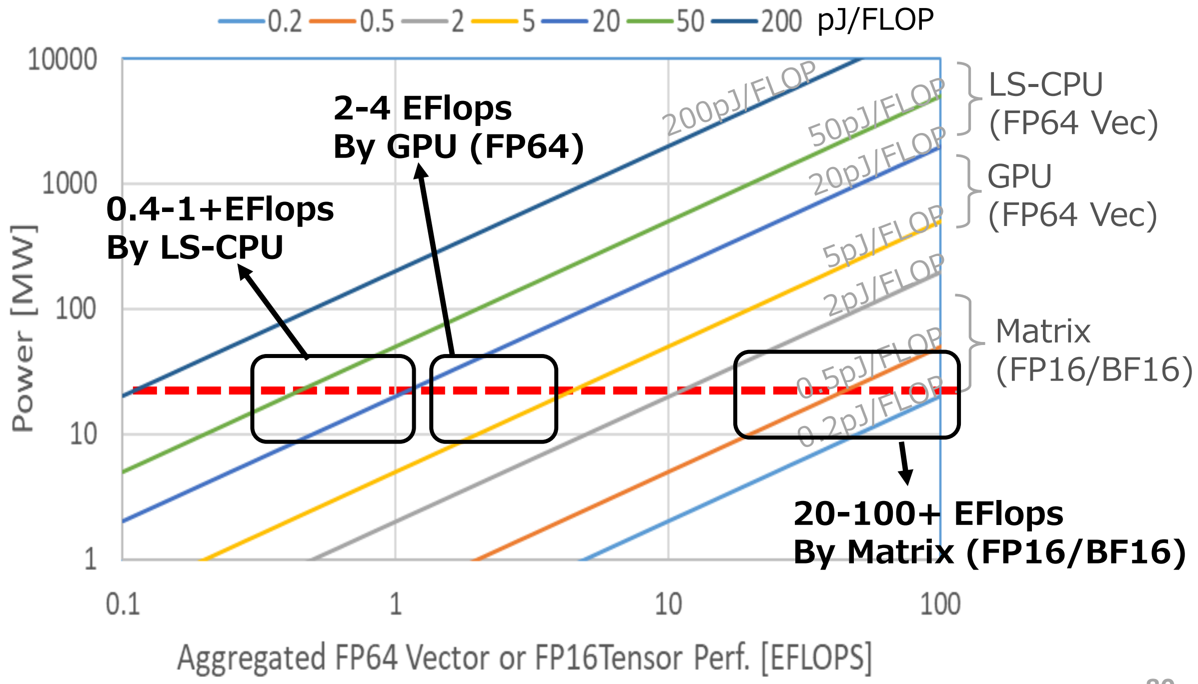
Power efficiency of data movement: Low ⟵⟶ High

chip-to-chip connection (chiplets)	**Monolithic die (conventional)** 	**Chiplet-based (becoming main-stream)** 	**More aggressive chiplet-based (Future direction)** 
3D stacking approaches	**2.5D connection (conventional)** 	**3D - Hybrid Bonding (single chip stacked)** 	**3D implementation (multi chips stacked)** 
Optics	**AOC (conventional)** 	**Silicon-Photonics – co-packaged optics connection (various technology candidates incl. WDM)** 	

Performance Projection in Power Constrained Scenarios

- **Estimated energy per operation on current and future technologies**
  - Based on historical trend obtained by publically available data
  - Not related to any partner vendors' perspective
- **Case for 30MW power budget (10MW for memory and 20MW for compute)**
  - Network is omitted for simplicity but it is very important
  - May not be realistic due to other constraint such as cost and thermal issues





Performance Projection in Power Constrained Scenarios

- **Estimated energy per operation on current and future technologies**
  - Based on historical trend obtained by publically available data
  - Not related to any partner vendors' perspective
- **Case for 30MW power budget (10MW for memory and 20MW for compute)**
  - Network is omitted for simplicity but it is very important
  - May not be realistic due to other constraint such as cost and thermal issues

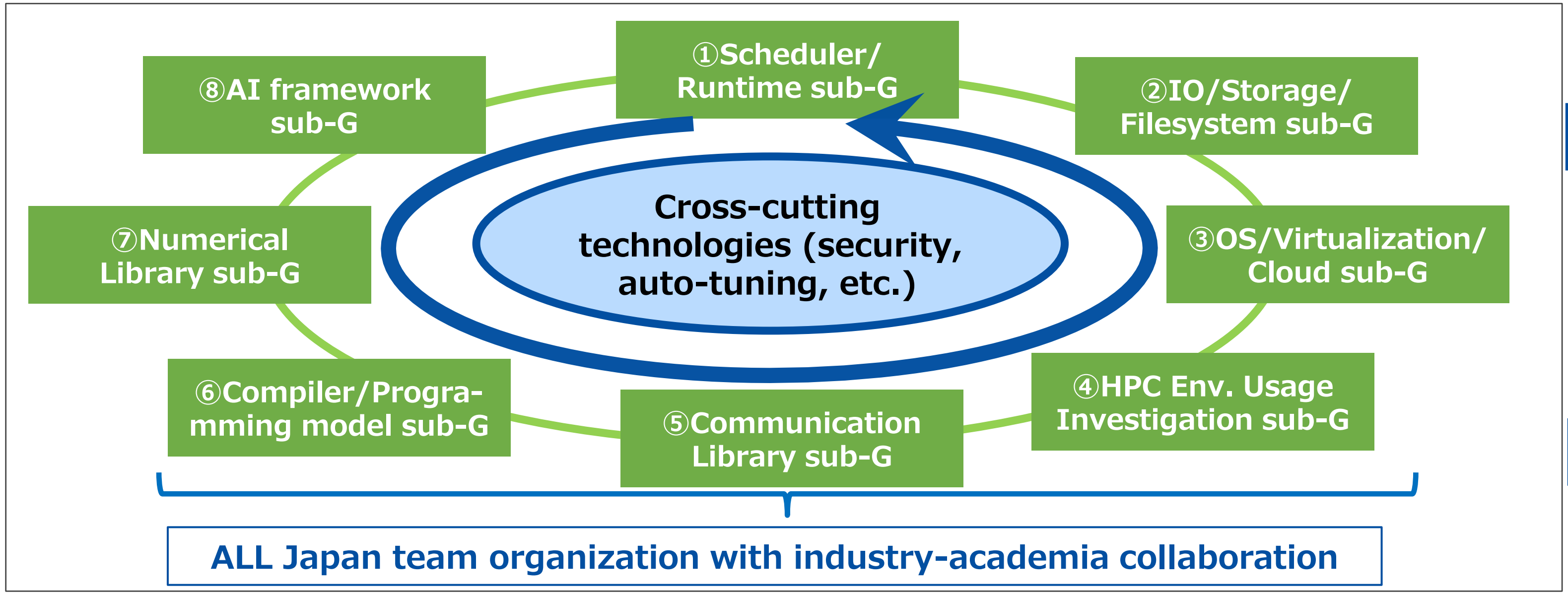
Summary of system performance projection

	LPDDR	HBM	3D Staking Mem.
LS CPU (FP64 Vec.)	1EFlops, 100PB/s (B/F = 0.1)	1EFlops, 500PB/s (B/F = 0.5)	1EFlops, 4000PB/s (B/F = 4.0)
GPU (FP64 Vec.)	4EFlops, 100PB/s (B/F = 0.025)	4EFlops, 500PB/s (B/F = 0.13)	4EFlops, 4000PB/s (B/F = 1.0)
Matrix (FP16 Tensor)	100EFlops, 100PB/s (B/F = don't care)	100EFlops, 500PB/s (B/F = don't care)	100EFlops, 4000PB/s (B/F =don't care)

System Software and Library Research

Objective and Overview

- **Objective**
  - Investigate technological trend of system software and draw R&D roadmap based on it
- **Research overview**
  - Item 1: Investigates System Software Trends
    - Study existing system software and future trends in terms of portability, productivity and performance
    - Study current usage status of system software in the HPCI systems and major supercomputing centers in the world
  - Item 2: Collects information to decide software development strategies
    - Define strategies for software development (proprietary or open-source software?)
  - Item 3: Comparison of similar software
    - Select best software and clarification of alternative software



Survey of system SW trend & draw development roadmap

Examine new system SW areas for industrialization

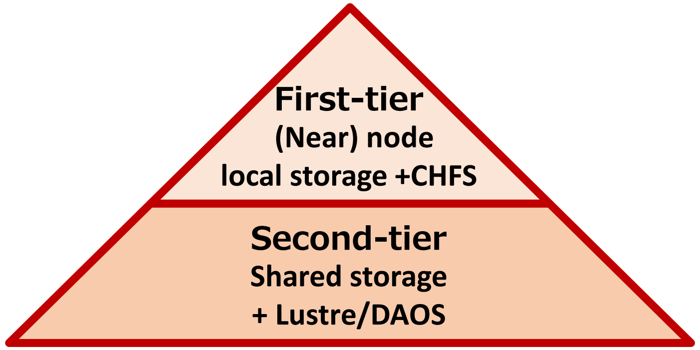
Expectation of Storage System (Under Consideration)

- **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)	IOPS	Amount
First Tier	(Near) node local storage	Now consideration (such as CHFS)	Time for dumping all memory: Less than 1min	Time for meta-data processing of max I/O processes: less than 1s	Twice as total memory size
Second Tier	Shared storage	Lustre, DAOS	Time for dumping all memory: Less than 5min	1/10 of first tier storage	30x of total memory size

  - Data migration from Fugaku to FugakuNEXT (Continuous operation and usage)
  - Hardware/Software design for stable performance
  - 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

Application Research

Objective

- **Surveying computational resources requirement to realize cutting-edge research results by next-generation computing infrastructure**
  - Not only in general performance but also in various indices such as programming productivity
- **Constructing (micro)benchmarks that reflect the characteristics of representative applications to estimate application performance**

Overview and Current Status

- **Pure apps group (Life science, Materials and energy, Weather and climate, Earthquake/tsunami disaster prevention, Manufacturing, Fundamental science, Social science, Digital-twin & Society 5.0)**
  - Completed a survey on application analysis on current supercomputers
  - Studying expected results in each application field and the computer resources required for them around 2030
  - Developed benchmark programs reflecting the characteristics of programs in each application area (GENESIS, qNET_kernel, QWS, SCALE, CUBE, QWS, ISPACK)
- **CS group (computational science/ML algorithms, benchmark building, performance modeling)**
  - Decided to use MLPerf as a machine learning benchmark and completed model selection
  - Studying benchmarks with variable problem size and amount of memory per core

Hardware and application co-design for post Exascale computing is important

Science Target in FugakuNEXT Era

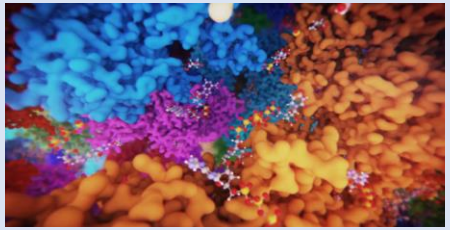
2011~「K computer」

2020~「Fugaku」

2030

「Fugaku NEXT」

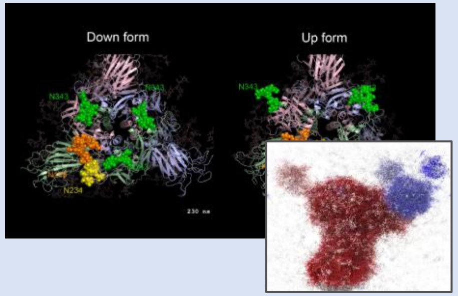
Simulation of Subcellular Sequence Dynamics



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

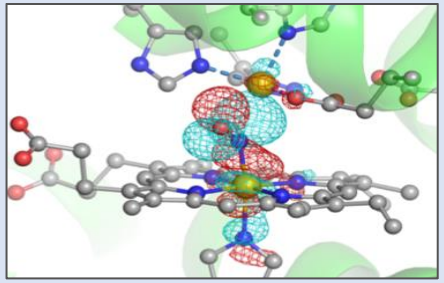
Long-term dynamics and cellular function multi-scale models

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



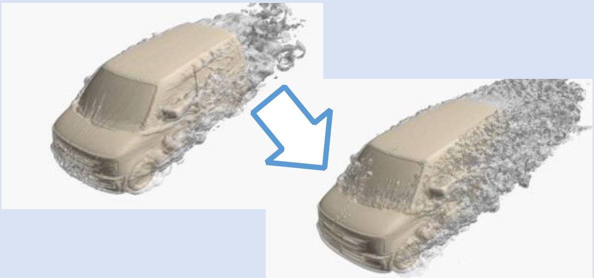
"Fugaku" allows for longer dynamics of even larger systems.

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

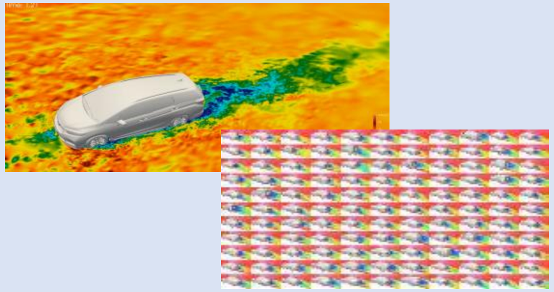


Enables dynamics considering electronic states (applied to bio-digital twin antibody drug discovery, etc.)

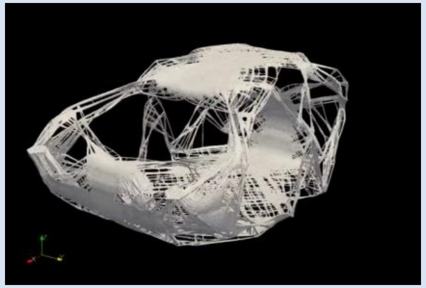
Automobile aerodynamics



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 usin g generative AI,
Establishment of automatic dri ving technology

Science Target in FugakuNEXT Era

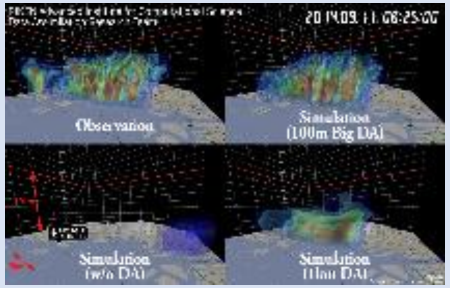
2011～「K computer」

2020～「Fugaku」

2030

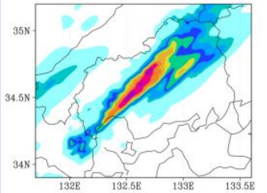
「Fugaku NEXT」

Weather and Climate



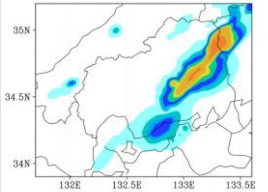
Development of a guerrilla rainfall forecasting method using the “K computer”

World’s first Real-time guerrilla rainstorm forecast by “Fugaku” during 2021 Tokyo Olympic & Paralympic Games



w/o control

controlled



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 learn 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

AI Hardware Trends

- As pretraining models becomes ever expensive with super-quadratic complexity, and LLM usage spreads, training market will confined to a few players while market emphasis will shift to inference chips that can be made much more power efficient.
- Also LLM training improvement is saturating with lack of data; emphasis is now shifting to reinforcement learning at inference time as per ChatGPT-o1
- Inference of heavy-duty LLMs will not happen at the edge as it will be much cheaper to send the data over 5G/6G, not sacrificing battery life and other resources such as memory
- Thus inference at IDC will be the largest infrastructure as well as consumer of societal energy (e.g., ChatGPT-o1)
- ‘Zettascale’ in AI with 40MW power budget on FugakuNEXT contributes to this with emphasis on low precision (FP/INT 4/8 bits)

Modern GPUs accelerated by Low Precision Matrix Engines

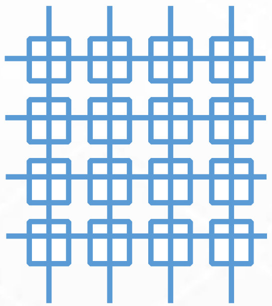
Computer simulations create the future

	H100	B200	Mi300A
FP64	67TF	40TF	123TF (60+TF)
FP32	67TF	40TF	123TF
TF32	495TF	1100TF	490TF
FP16/BF	990TF	2200TF	981TF
FP8	1980TF	4500TF	1960TF
INT8	1980TOPS	4500TOPS	1960TOPS
FP4	NA	9000TF	NA

What about Dense Linear Algebra?

Precision Depending Analysis – what and how matrix engines provide good ROI relative to their silicon occupancy?

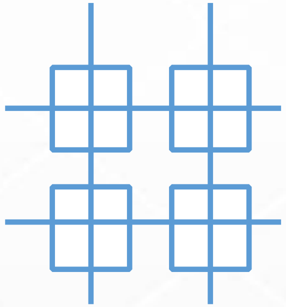
- Energy = compute (multipliers, volume) + data movement (between units, surface)
  - Low precision – low surface:volume, optimize to minimize data movement, matrix engines to minimize wire distance
  - High precision – high surface:volume, data transfer less problem, performance & energy gain small, dark silicon of unused multipliers wasteful, **wide vectors sufficient.**
- 4~16 bit apps: Deep Learning/AI training
- 19~ (TF32) ~ 32 bit apps: DL/AI, molecular dynamics, higher order methods (**mixed precision**)
- 64 bit apps: first-principle material science eg DFT => **Emulation of "64 bit" apps with "Ozaki Scheme" => with 1/20 slowdown we expect effective 10 Exaflops from 200 INT8 ExaOps "Zettascale" AI machine (20x Fugaku)**



Low precision MM

Low volume (compute) : surface (comm) ratio

Matrix units help to reduce data transfer energy



High precision MM

high volume (compute) : surface (comm) ratio

Vector units may be sufficient as benefit of matrix may be low

(Slide Courtesy NVIDIA)

FP64 Emulation Using INT8 Tensor Cores

Algorithm Description

DGEMM on Integer Matrix Multiplication Unit

Hiroyuki Ootomo
ootomo.h@rio.gsic.titech.ac.jp
Tokyo Institute of Technology
Tokyo, Japan

Katsuhisa Ozaki
ozaki@sic.shibaura-it.ac.jp
Shibaura Institute of Technology
Saitama, Japan

Rio Yokota
rioyokota@gsic.titech.ac.jp
Tokyo Institute of Technology
Tokyo, Japan

- We implemented this on NVIDIA Ada, **Hopper**, and **Blackwell** GPUs
- Various applications were tested to determine accuracy and performance impact:
  - HPL
  - Materials Science
  - Electronic Structure
  - Molecular Dynamics
  - Computational Chemistry
  - Sparse Direct Solvers

https://arxiv.org/abs/2306.11975

- Input and output matrices are IEEE FP64 (C = A x B)
- Structure of DGEMM leveraging INT8 Tensor cores
  - Prologue:
    - Find max(A[i,:]), max(B[:,j])
    - Align mantissa values of A and B elements to the same exponent
    - Slice up A and B mantissas in integer buckets
  - Compute:
    - Compute-accumulate dot products of slices using integer arithmetic
    - Structurally similar to FP64 hardware MAC, just 8 bits at a time but using IMMA tensor cores
  - Epilogue:
    - Assemble FP64 results from sliced representation and the exponent information

Acceleration of Quantum Chemistry using Combinatios of Emulation (Ozaki) & Mixed Precision utilizing AI-Centric GPUs

Reducing Numerical Precision Requirements in Quantum Chemistry Calculations

William Dawson,*,† Jens Domke,† Takahito Nakajima,† and Katsuhisa Ozaki‡

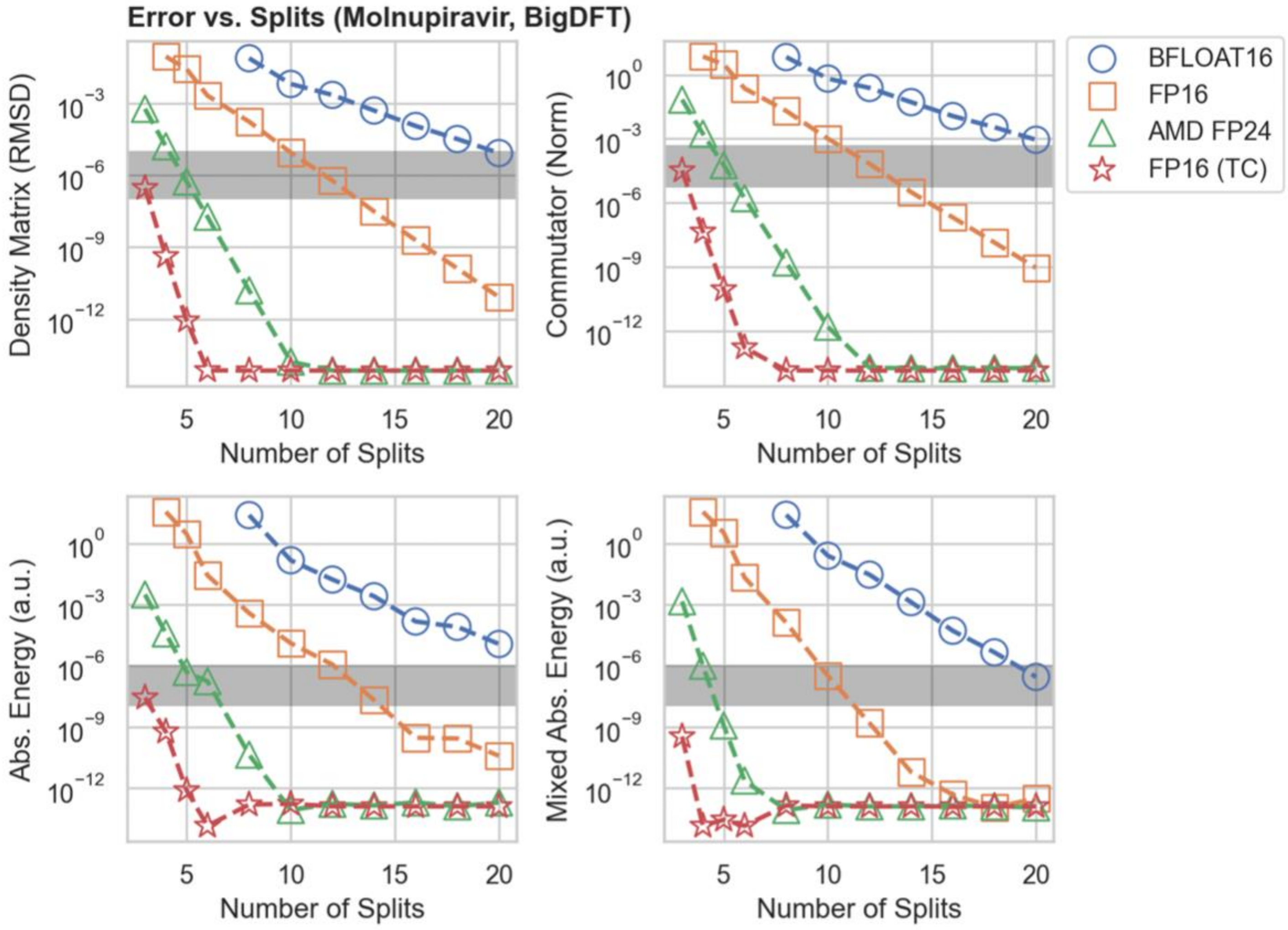
†*RIKEN Center for Computational Science, Kobe, Japan*

‡*Shibaura Institute of Technology, Saitama, Japan*

E-mail: william.dawson@riken.jp

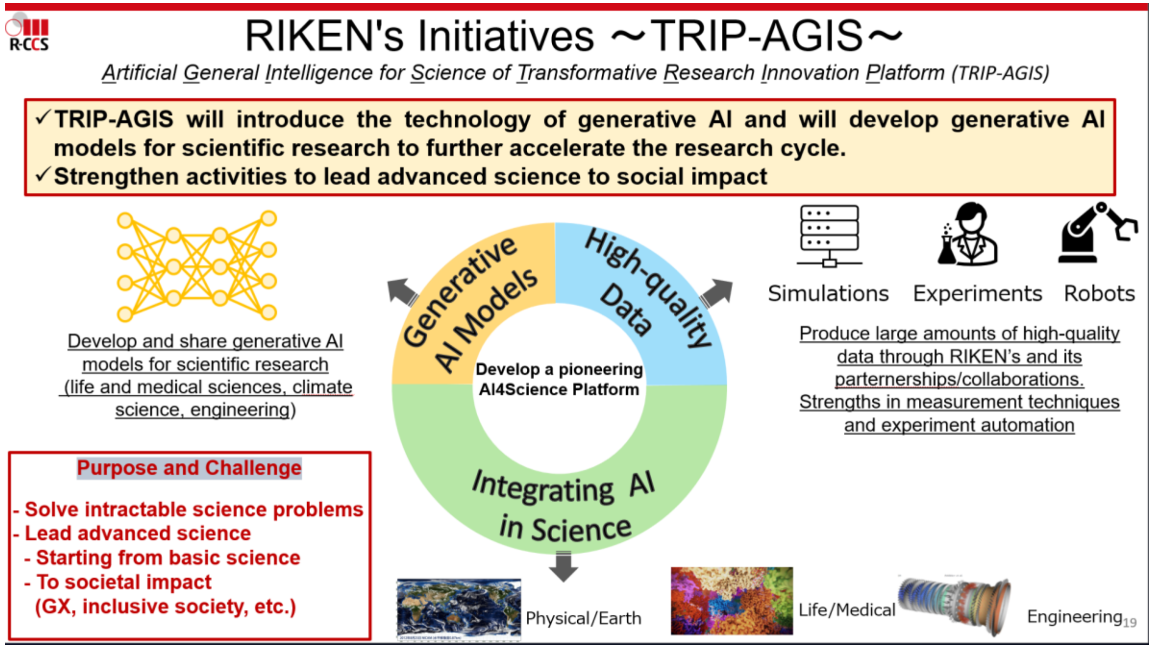
Abstract

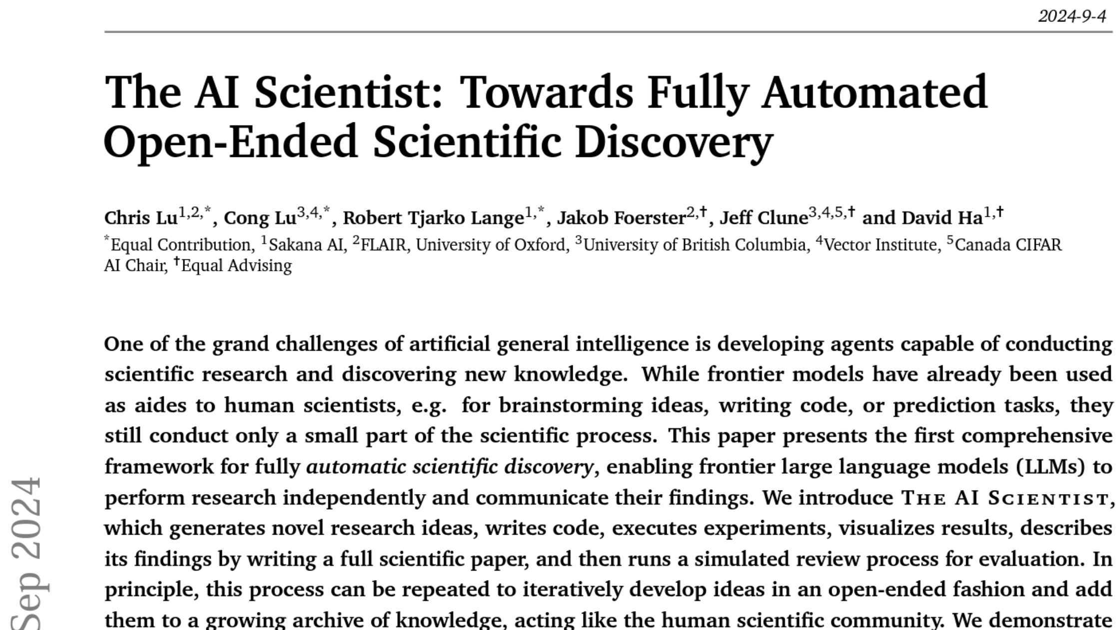
The abundant demand for deep learning compute resources has created a renaissance in low precision hardware. Going forward, it will be essential for simulation software to run on this new generation of machines without sacrificing scientific fidelity. In this paper, we examine the precision requirements of a representative kernel from quantum chemistry calculations: calculation of the single particle density matrix from a given mean field Hamiltonian (i.e. Hartree-Fock or Density Functional Theory) represented in an LCAO basis. We find that double precision affords an unnecessarily high level of precision, leading to optimization opportunities. We show how an approximation built from an error-free matrix multiplication transformation can be used to potentially accelerate this kernel on future hardware. Our results provide a road map for adapting quantum chemistry software for the next generation of High Performance Computing platforms.



AI for Science Needs to Be "Scientifically Creative"

- **Science needs to be accelerated by AI via innovations, not merely by streamlining**
  - Just getting rid of the mundane admin work for the scientists has limited value due to Amdahl's Law
- **The ultimate goal of AI for Scientist is for the AI to have sufficient scientific creativity that would rival or even exceeded human scientists, thus solving the true energy crisis (of having too many human scientists)**





Sep 2024

2024-9-4

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Chris Lu[1,2,*], Cong Lu[3,4,*], Robert Tjarko Lange[1,*], Jakob Foerster[2,†], Jeff Clune[3,4,5,†] and David Ha[1,†]
[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Canada CIFAR AI Chair, [†]Equal Advising

One of the grand challenges of artificial general intelligence is developing agents capable of conducting scientific research and discovering new knowledge. While frontier models have already been used as aides to human scientists, e.g. for brainstorming ideas, writing code, or prediction tasks, they still conduct only a small part of the scientific process. This paper presents the first comprehensive framework for fully *automatic scientific discovery*, enabling frontier large language models (LLMs) to perform research independently and communicate their findings. We introduce The AI Scientist, which generates novel research ideas, writes code, executes experiments, visualizes results, describes its findings by writing a full scientific paper, and then runs a simulated review process for evaluation. In principle, this process can be repeated to iteratively develop ideas in an open-ended fashion and add them to a growing archive of knowledge, acting like the human scientific community. We demonstrate

4D Parallelism: TP+CP+PP+DP

- **Tensor Parallel [TP]**
  - The more you split the layer via TP → less compute and more comm
  - TP → strong scaling
  - Conclusion: do TP inside the node (on a multi-GPU system)
  - In practice, we observe TP = 2~8
    - Depends on intra-node interconnect
- **Context Parallel [CP]**
  - Necessary evil
- **Pipeline Parallel [PP]**
  - Necessary evil
    - There is always inefficiency (bubble in a pipeline)
    - Used when running into the limits of TP, CP, and DP
- **Data Parallel [DP]**
  - Use to the maximum possible

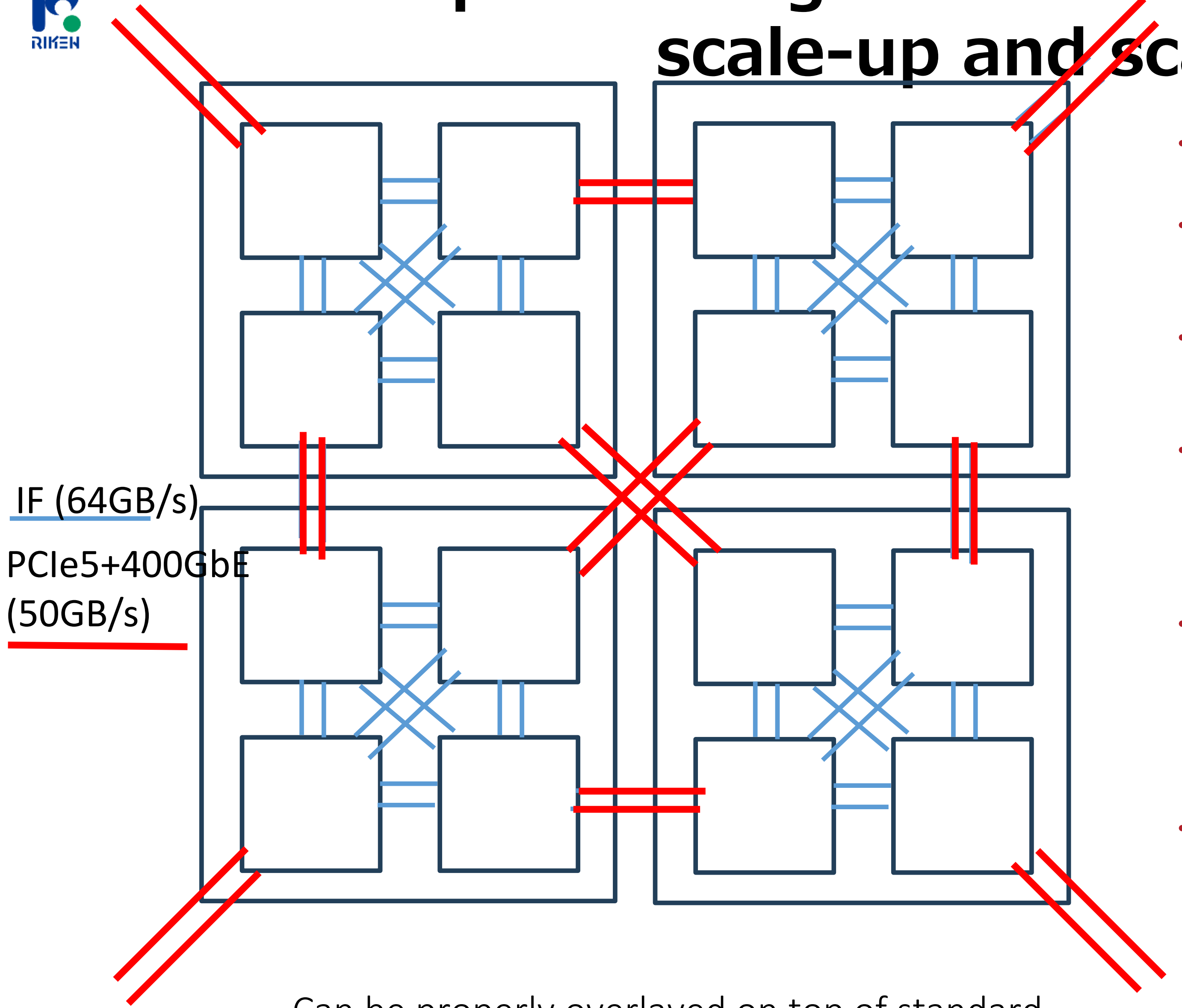
DP [CP [PP [TP

GPT Compute & communication estimate [by M. Wahib & A. Drozd, R-CCS]

- “Compute and Communication cost per an iteration of GPT3-175B parameterized as: B = 16, E = 12K, S = 32K, N_p = 175B, L =96, W = 2. Model_FLOPS is empirically measured (ModelFLOPS = 467.9 x 96 TF)”
- **Given 2PF compute FP8 w/50%utilization, and 400GByte/s injection BW, TP transfer time would be less than compute.**

	FLOPS per Worker	Total FLOPS	Payload Size per Worker (Bytes: logical)	Agg. Payload Workers (Bytes: logical)	Rounds of Communication (Communication Pattern)
TP only (T workers)	Model_FLOPS / T	Model_FLOPS (constant to T)	W x B x E x S x 4 x L (constant to T)	N/A	Per Layer = 4x = 2x in forward + 2x in backward (AllReduce)
Example: T = 8	5,614.8 TF	44,918.4 TF	4,718.5 GB	N/A	
PP only (P workers)	Model_FLOPS / P	Model_FLOPS (constant to P)	W x B x E x S x 2 x P (Linear to stages)	W x B x E x S x 2 x P x (P-1)	Per layer: 2 x P (P2P) [note: assumption that number of stages = P]
Example: P = 8	5,614.8 TF	44,918.4 TF	196.8 GB	1,377.6 GB	
DP only (D workers)	Model_FLOPS	Model_FLOPS x D (linear to D)	(W x N_p) (constant to D)	N/A	Single update per model = 1x OR Segmented update per layer = L (AllReduce)
Example: D = 8	44,918.4 TF	359,347.2 TF	350 GB	N/A	
TP+PP+DP (T x P x D workers)	Model_FLOPS / (T x P)	Model_FLOPS x D (linear to D)	AllReduce = W x B x E x S x 4 x L / P + W x N_p / (T x P) P2P = (W x B x E x S x 2 x P) / T	N/A	Per worker = L/P + T x P (AllReduce) Per worker = 2 x P (P2P)
Example: 8 x 8 x 8	701.85 TF	359,347.2 TF	AllReduce = 595.2 GB P2P = 24.6 GB	N/A	

Isomorophic Tree gather-scatter network ‘merging’ scale-up and scale-out



Can be properly overlayed on top of standard HPC networks e.g. Fattree, prioritarizing shortcuts to reducing latency

- Quad APU (Mi300A, GB200 etc.) x 4 node as a unit
- 8 high bandwidth intra node links tightly connecting APUs, 6 links intra node and 2 links inter node (as PCIe5-400GbE)
- This creates an isomorphic quad-tree with almost same bandwidth for IF (64GB/s x2) and 400GbE(50GB/s x 2)
- So the tree is 4, 16, 64, 256, … There are shortcut links as in practice the 400GbE links are connected to a fat switch, allowing shortcuts but we will ignore those for the moment
- Given such a tree, there is a classic collective algorithm for reduction, whose runtime is exactly the amount of data that are injected into the network / bandwidth, sans a small startup overhead. This does not change for arbitrary tree size
- For example, to do a word-wise collective summation of 100GB data on every node in this network will always take one second, which is equivalent to the time it takes to inject 100GB of data into the network. There is a small amount of logarithmic overhead but can be ignored for a large payload
- In a nutshell, gather-scatter time ~= injection time

Macro-scale terascale memory within Scale-up Network

Training

- ✓ Aggressive offloading
- ✓ w/o affecting performance

Per DeepSpeed: ~20-25% Offloadable

Inference (caching)

- ✓ Very long decoding jobs (ex: CoT, GoT)
- ✓ KV-cache: perf penalized, job stays local
- ✓ Prompt caching (90% cost saving)

Simple perf model: ~10-20x ↑ KV-cache ~5-10x ↓ slowdown

Model Swapping

- ✓ Commercial: pooling/comparing different models
- ✓ Science: Different models at different phases in simulation

Swapping-in 500B Parameter Model: ~1 Second

Tandem sim/training

- ✓ Simulation does in-Situ training
- ✓ Swap model and simulation data

Swapping 1TB: ~1 Second

Checkpoint/ Logging

- ✓ Avoid jitter or interruptions when checkpointing or logging

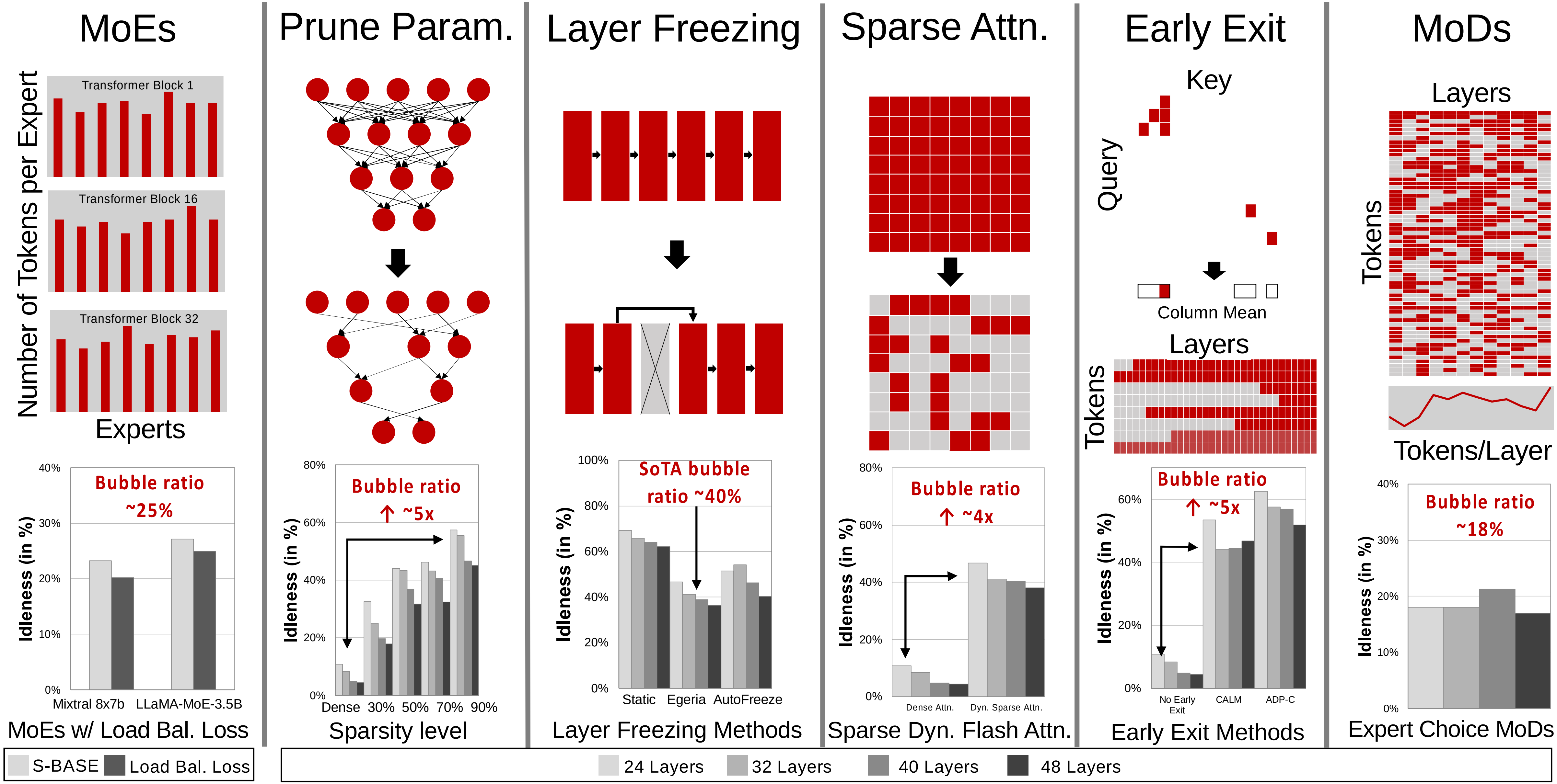
Zero Overhead: checkpointing 500B Parameter Model (Very important for GPUs whose RAS are not up to A64FX level)

Datasets

- ✓ Stream data when training
- ✓ Free shuffle (memory is byte addressable)

Staging Dataset From Storage

Research: Dynamic LLMs



Towards 'Zettascale' HPC Performance for FugakuNEXT

- **Simulation Workloads**
  - Raw HW Performance Gain: 10x ~ 20x
  - Mixed precision or emulation: 2x ~ 8x
  - Surrogates / PINN: 10x ~ 25x
  - Total: 200x ~ 1000x or more over Fugaku => 'Zettascale'

- **Raw AI HW performance**
  - Low precision, sparsity, new models…
  - Expect 'Zettascale' AI performance

- **With 40MW Limit (not GigaW e.g., hyperscalars)**

Many of the FugakuNEXT Concepts will be tried out in TRIP-AGIS 2025 AI machine… Stay tuned

- **Both AI and HPC (simulation) performance & tight coupling**
  - High GPU (throughput) & CPU (latency) performance
- **Extensive mixed precision and emulation support**
- **Convergence of Scale-up and Scale-out network beyond standard HPC network**
  - Low cost bearing in mind AI and HPC communication patterns
- **High capacity memory within scale-up network for PIM-like processing – performance, resilience, …**
- **DLC Ultra high-density configuration (> 100KW/OCP rack) despite massive cabling and water**
- **Compliant to industry standards (e.g., OCP)**

SCA/HPC Asia 2026 will be held in Japan!

- Co-hosted by **SCA and HPC Asia**
- Showcase of cutting-edge HPC, AI, Big Data, Cloud Storage and Quantum Computing
- Science and Innovation through HPC, AI, Big Data and QC
- Opportunity to attract international talents from Asia and other countries

- **Date:** January 26 – 29, 2026
- **Venue:** Osaka International Convention Center
- **Co-located events:** in progress
  - ・Asian International HPC School,
  - ・Trillion Parameter Consortium, etc.
- **Expected number of participants:** 1500～3000
- **In collaboration with NSCC Singapore**

グランキューブ大阪
Osaka International Convention Center
（大阪府立国際会議場）