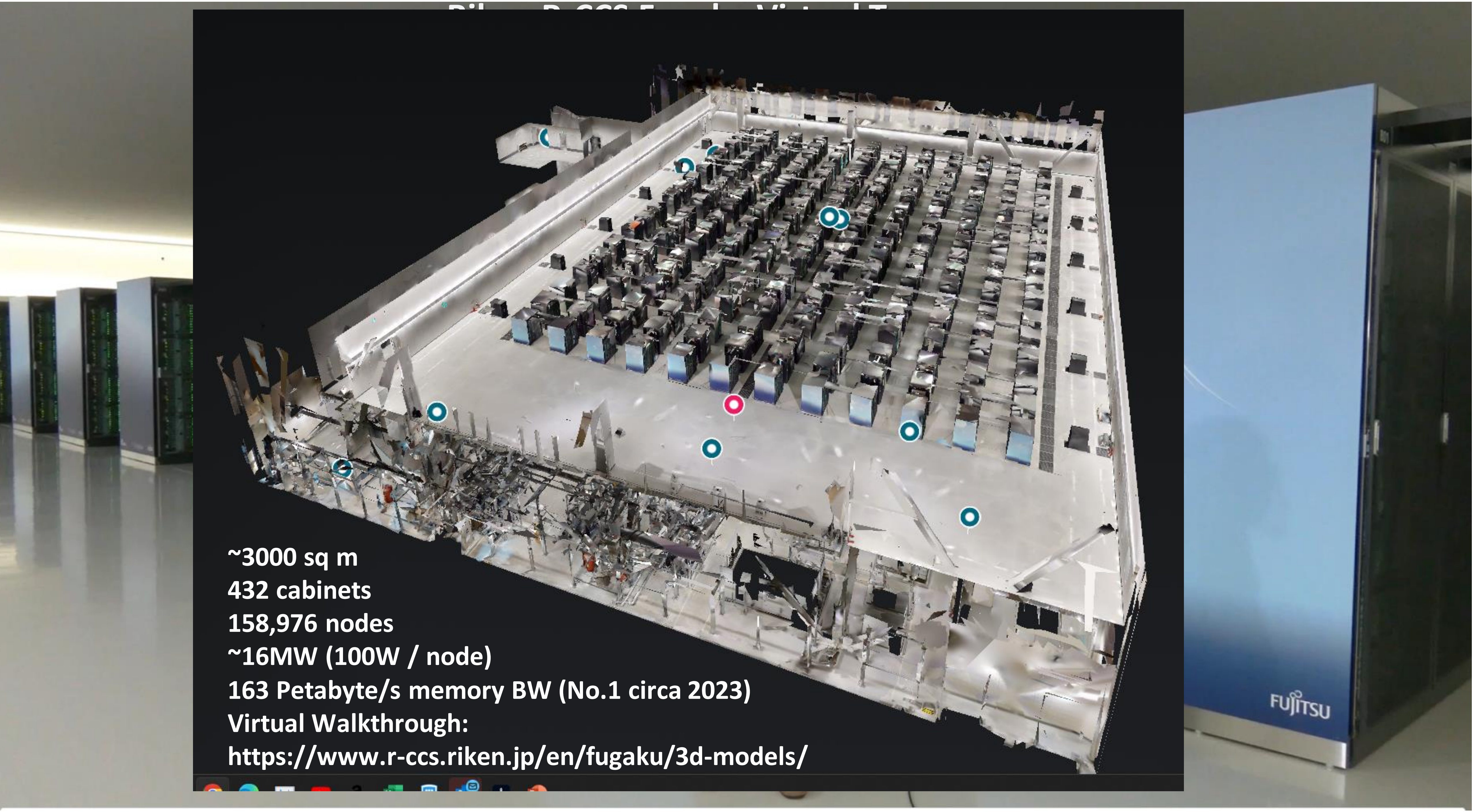
Computing for the Future at RIKEN R-CCS: AI for Science, Quantum-HPC

Satoshi Matsuoka, Director Riken R-CCS

Multicore World, Christchurch, NZ

Feb 14, 2024



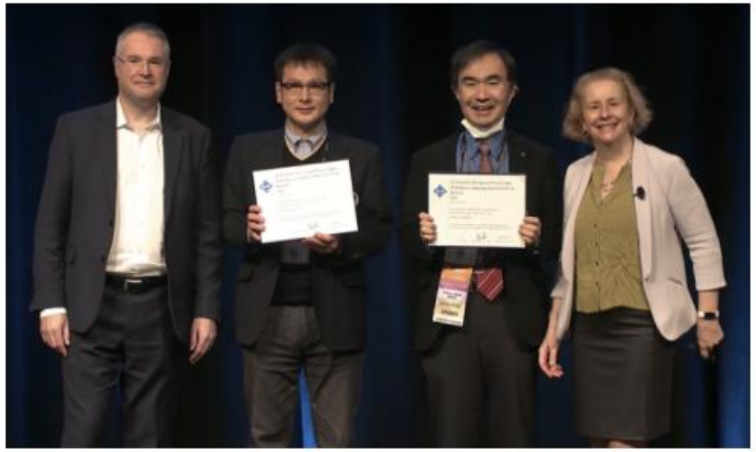
~3000 sq m
432 cabinets
158,976 nodes
~16MW (100W / node)
163 Petabyte/s memory BW (No.1 circa 2023)
Virtual Walkthrough:
https://www.r-ccs.riken.jp/en/fugaku/3d-models/

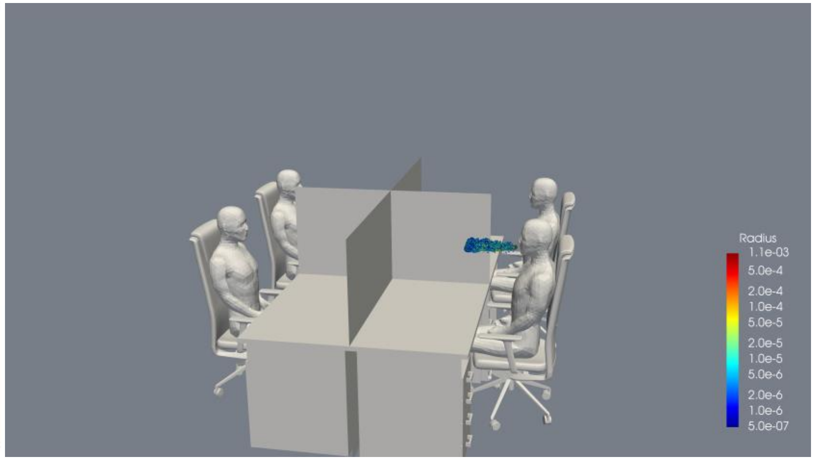
Major achievements of Fugaku

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

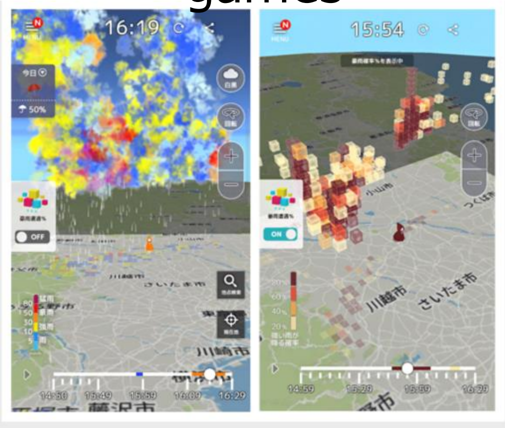
#1 in MLPerf HPC(Nov.2021-)

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





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



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

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

Finalists!

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.

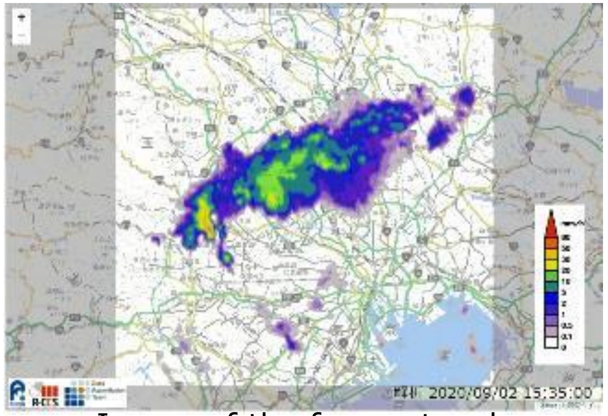


Image of the forecast web

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.

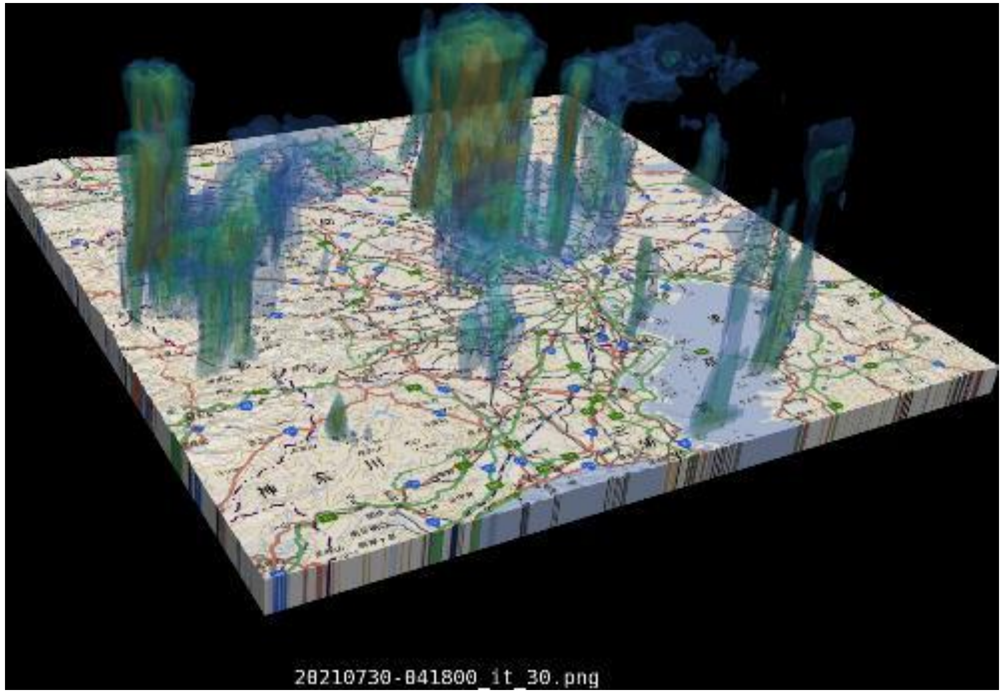
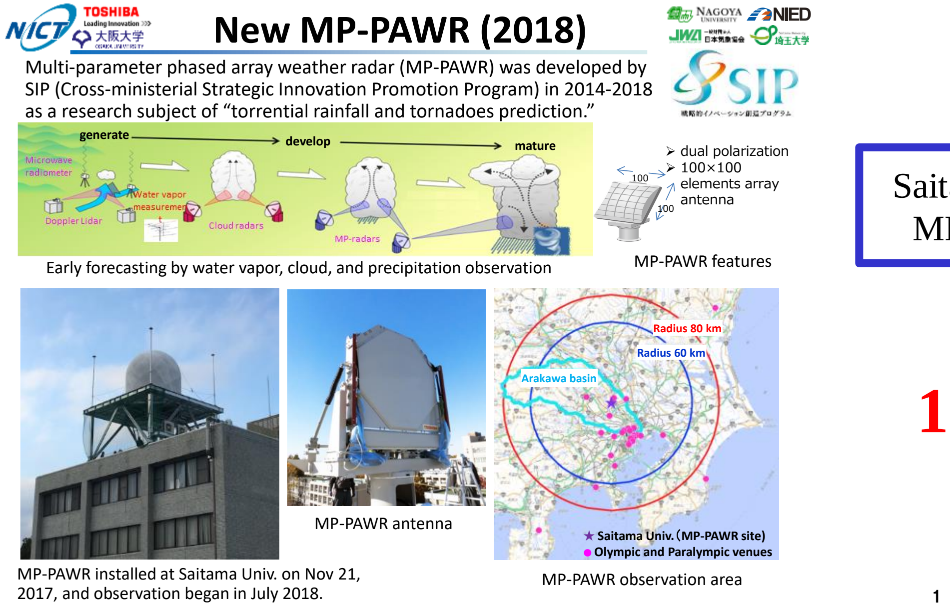


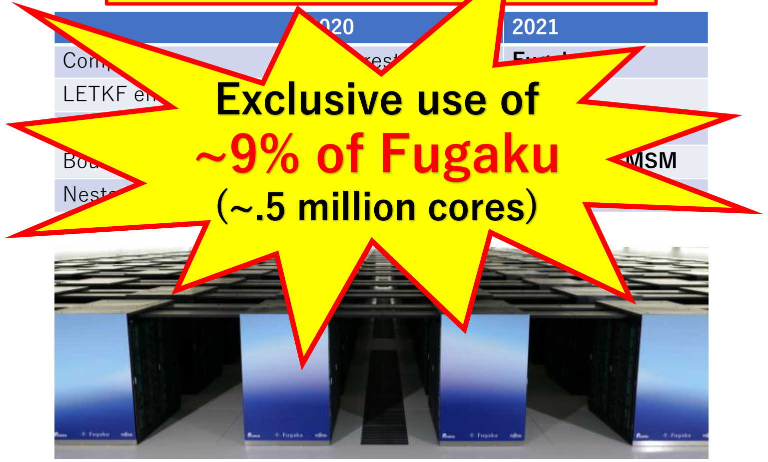
Figure: Bird's-eye view of 15-minute forecast rain distributions at 04:33:00 UTC, July 30, 2021, initialized at 04:18:00 UTC. Colors represent rain intensity. Vertical scale is stretched by three times. Map data courtesy of the Geospatial Information Authority of Japan

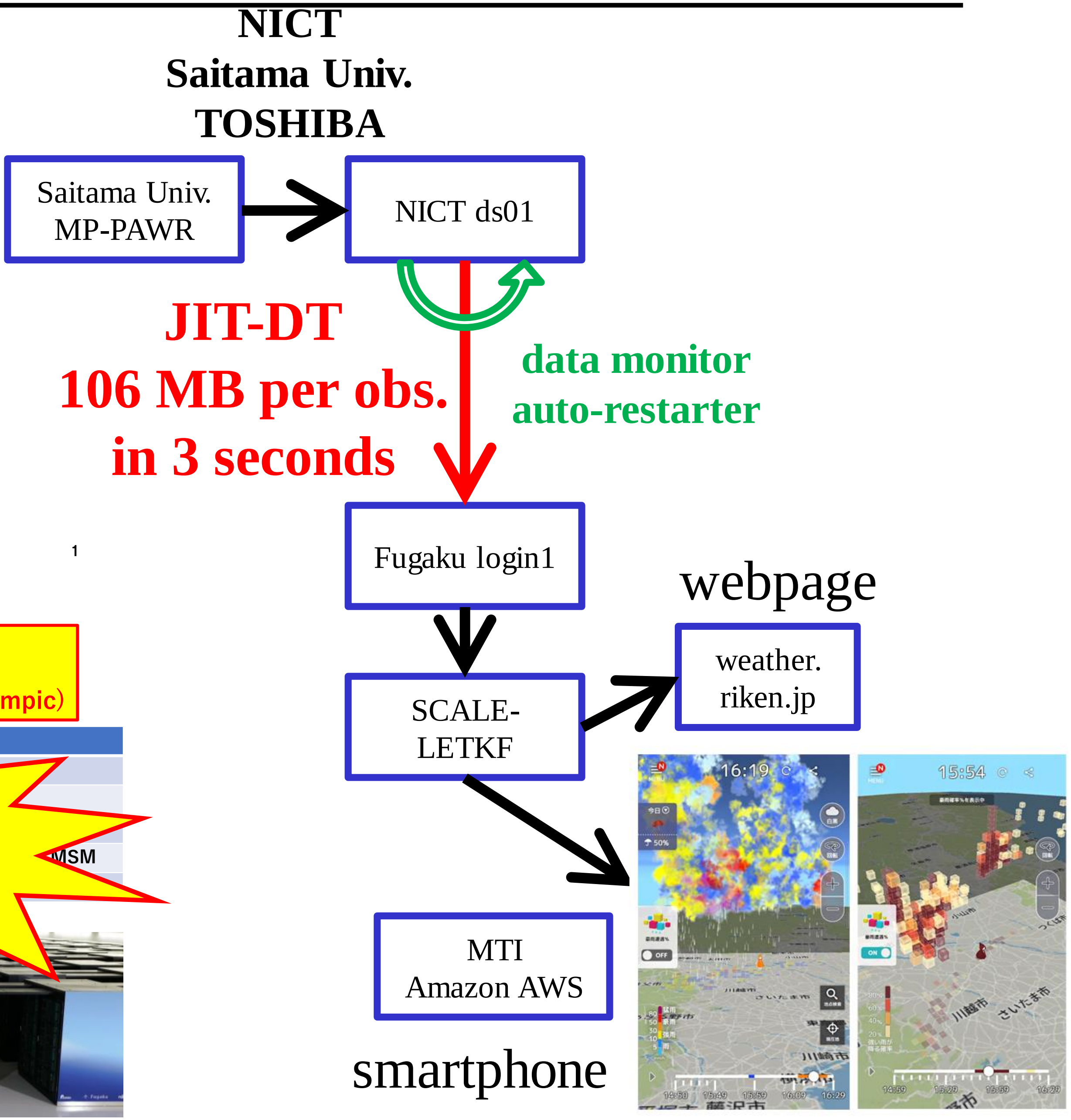
Real-time data transfer & data assimilation for Tokyo Olympics 2020



https://weather.riken.jp/

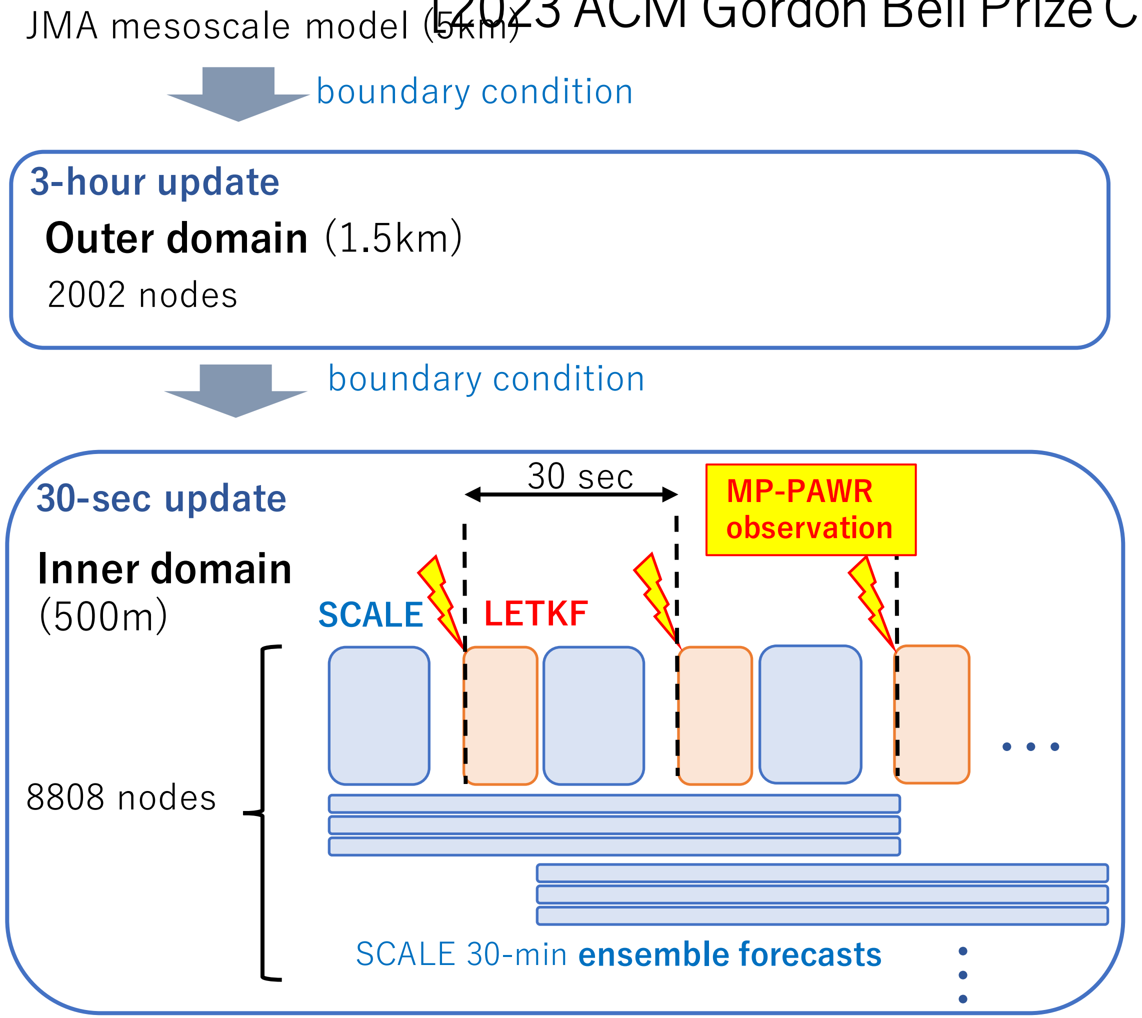
Real-time experiments in 2021
- July 20-August 8 (Olympic)
- August 24-September 5 (Paralympic)

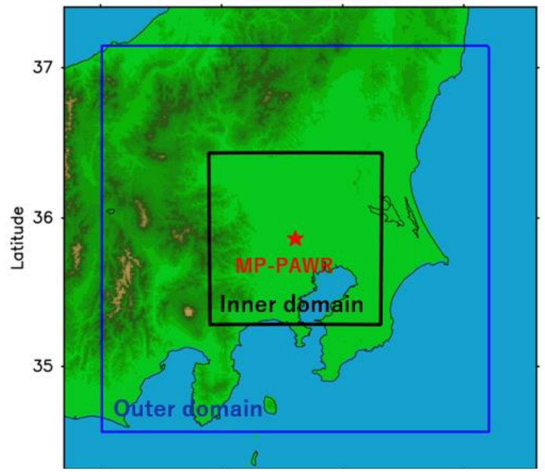




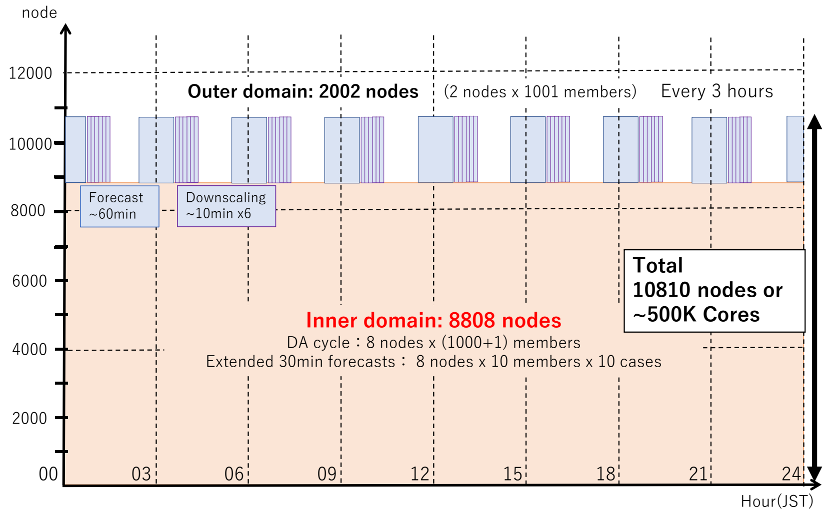
Real-time workflow of 30 sec, 500m weather forecast for 2020 Tokyo Olympics

[2023 ACM Gordon Bell Prize Climate Prize Finalist]





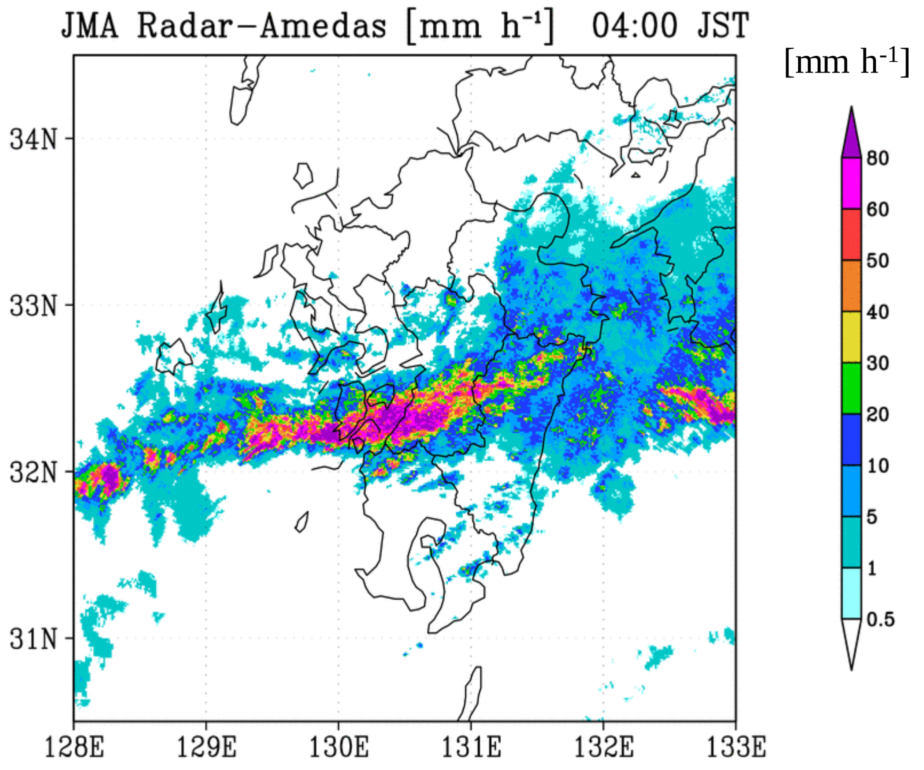
Real-time job scheduling of 1/2 million cores



What if we had many PAWRs?

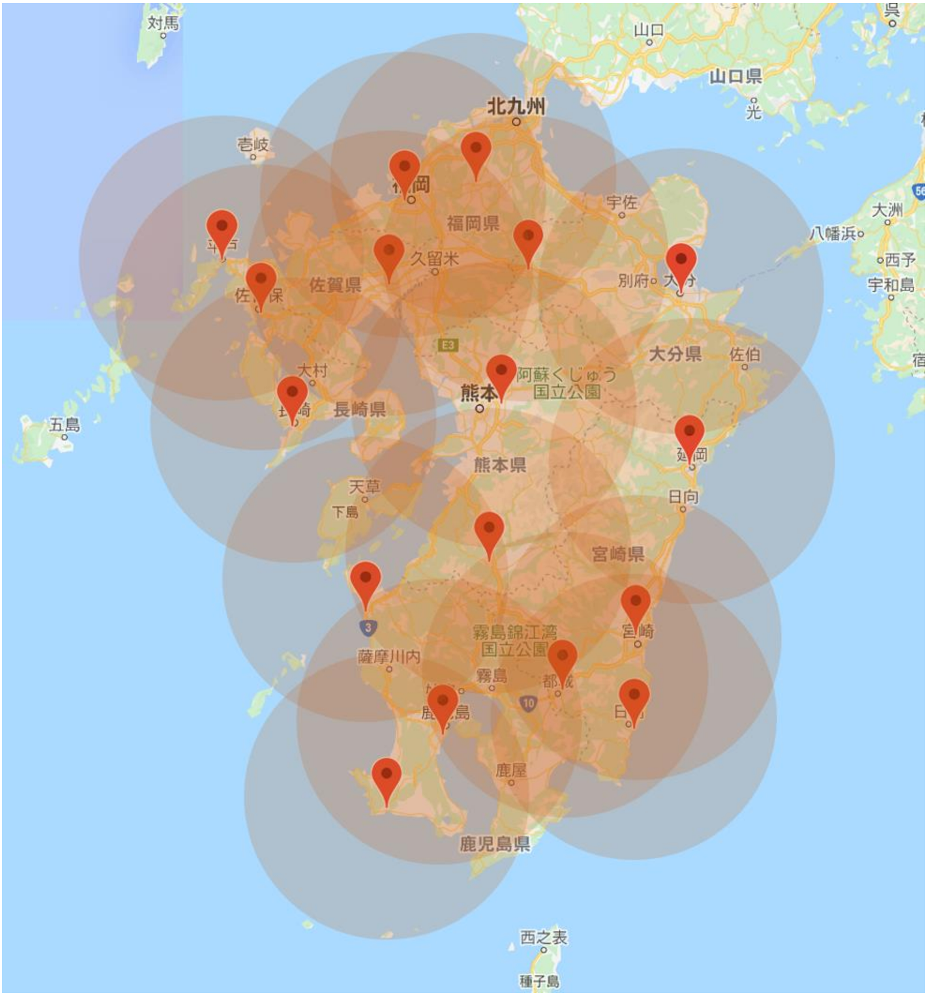
An Observing System Simulation Experiment (OSSE)

July 2020 heavy rain



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

A virtual PAWR network



Fugaku Siblings Preventing Natural Disasters

- Japan Meteorological Agency utilized large scale external 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

図１　線状降水帯予測スーパーコンピュータ

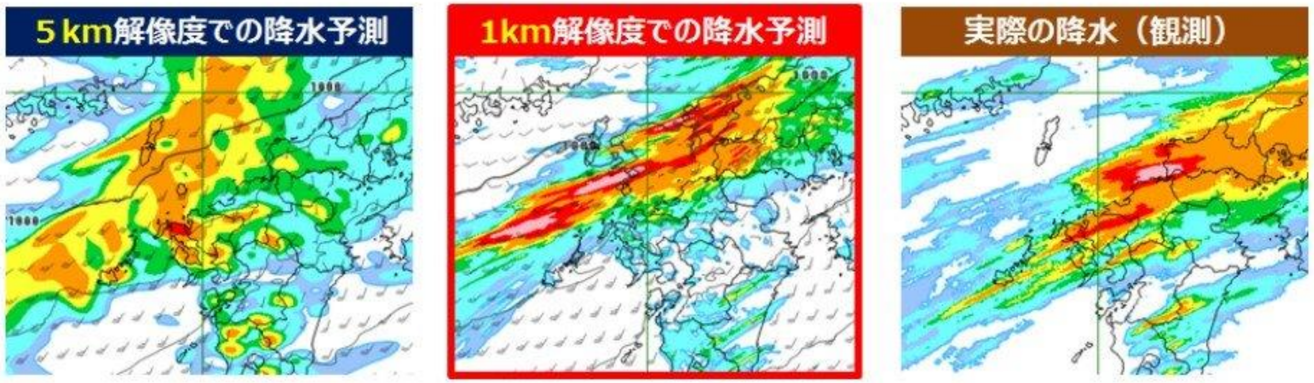


図２　水平解像度１kmに高解像度化した局地モデルのイメージ

2023 Hyperion Report on Fugaku Values

2 years into full production since Mar 2021 (3 years since pre-production)

#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.

#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.

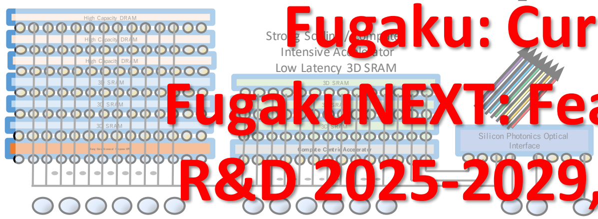
#4 Research Finding: Fugaku Is Focused On Creating Industrial Economic Growth

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).

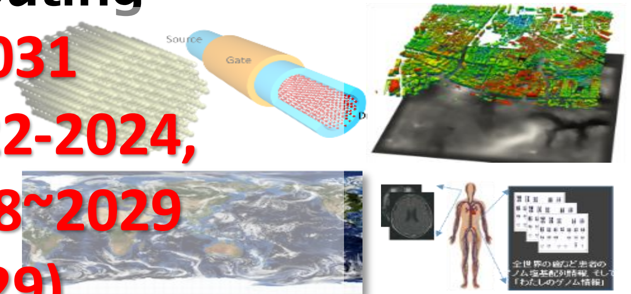
Riken-Intel Strategy for Innovation by Computing
Scientific Innovations are the 'Blue Ocean' in Computing

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

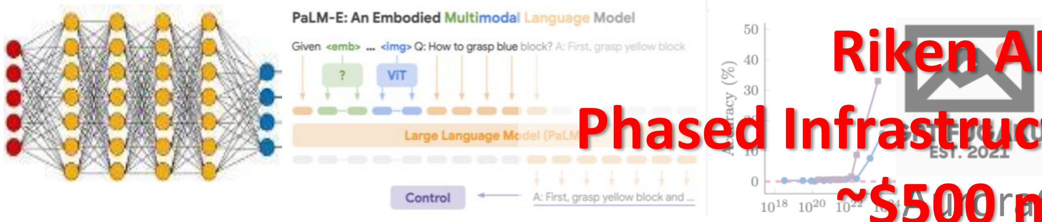


Fugaku: Current until 2030~2031
FugakuNEXT: Feasibility Study 2022-2024, R&D 2025-2029, Deployment 2028~2029
$Billion R&D&D (FY 2022~2029)

- Science **by** High Performance Computing

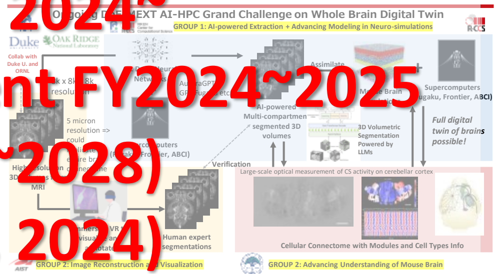


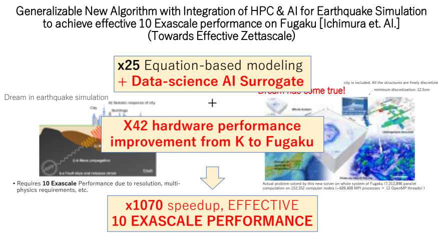
- Science **of** High Performance AI



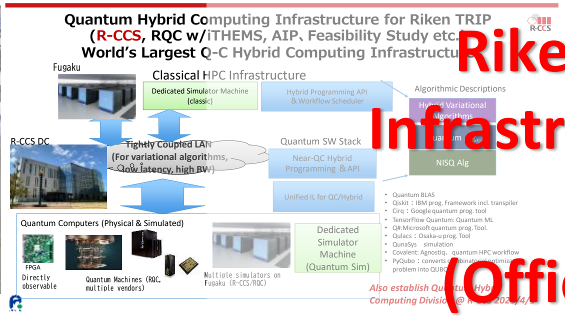
Riken AI for Science FY 2024~
Phased Infrastructure Deployment FY2024~2025
~$500 million (FY2024~2028)
(Officially Start April 1, 2024)

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



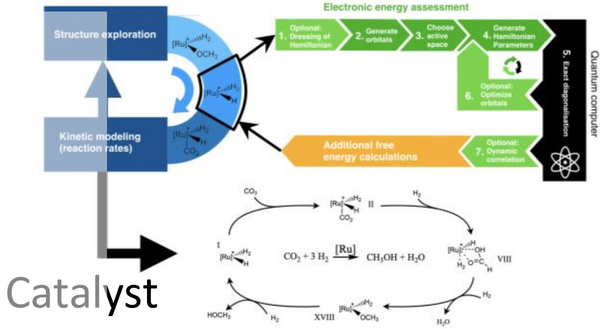


- Science **of** Quantum-HPC Hybrid Computing



Riken 'TRIP' Hybrid Quantum-HPC Infrastructure Deployment FY2023~2027
~$150 million+
(Officially announced Nov. 1, 2023)

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



Organization of RIKEN R-CCS as of 1st April 2024 (draft as of 29th Jan 2024)

R-CCS Deputy Director
Science of Computing
K. Nakajima

R-CCS
Director
S. Matsuoka

**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
Mohamed WAHIB

Supercomputing Performance
Jens DOMKE

**Large-Scale Digital Twin
H. Yamaguchi (April 2024)**

Recruitment for female PIs in progress, new teams 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

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
(expected April 2024)**

Quantum Computing Simulation
N. Ito

Quantum-HPC Hybrid Platform Operations
S. Miura

Office of the Fugaku Society 5.0 initiative

Office Director
S. Matsuoka

Office Deputy Director
Y. Watanabe

Office Coordinator
H. Shirai

Operations and Computer Technologies

Division Director
F. Shoji

Deputy Division Director & System Operations and Development
Y. Iguchi

Facility Operations & Development
S. Miura

**Data Interaction Technology Development
T. Kai
(March 2024)**

Software Development Technology
H. Murai

Advanced Operation Technologies
K. Yamamoto

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

AI for Science Platform Division

To be launched 1st April *We are hiring! (incl. postdocs, interns…)*

"The one who rules the platform rules the world"
The change of Science will be the same as Shopping

Shopping knowledge confined to individual merchant & consumer

DX

Consumer shopping behavior, product supply chain, delivery …

Amazon Storefront | Merchant A Storefront | Merchant B Storefront

Digitized shopping activities

Shopping knowledge *encoded & accumulated digitally* as digital twins - data, algorithms, programs, trained NN - in the platform

Scientific knowledge encoded as papers by individual labs

DX

Scientific research activities

Pharma Platform | Smart City Platform | Power Grid Platform

Digitized on each platform

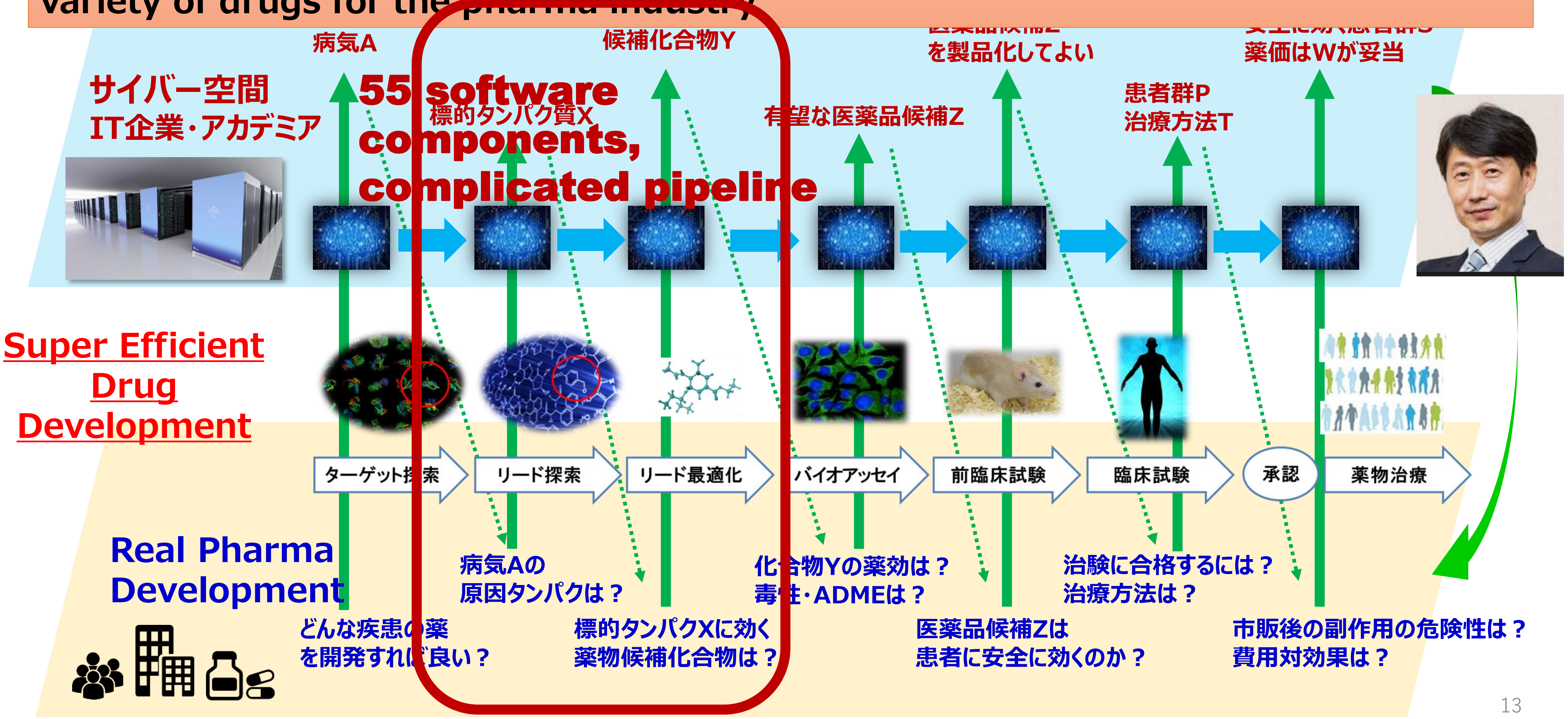
FugakuNEXT

Leadership in General Platform technologies will lead the entire IT and society

Scientific knowledge *encoded & accumulated digitally* as digital twins - data, algorithms, programs, trained NN - in the platform

Fugaku Pharma Platform [Okuno et al.]

Combining Simulation, AI and Big Data, construct a through Pharma pipeline on Fugaku and other IT resources, significantly decreasing time to development for a variety of drugs for the pharma industry



Data Processing Platform for SmartCity (Oishi, R-CCS)

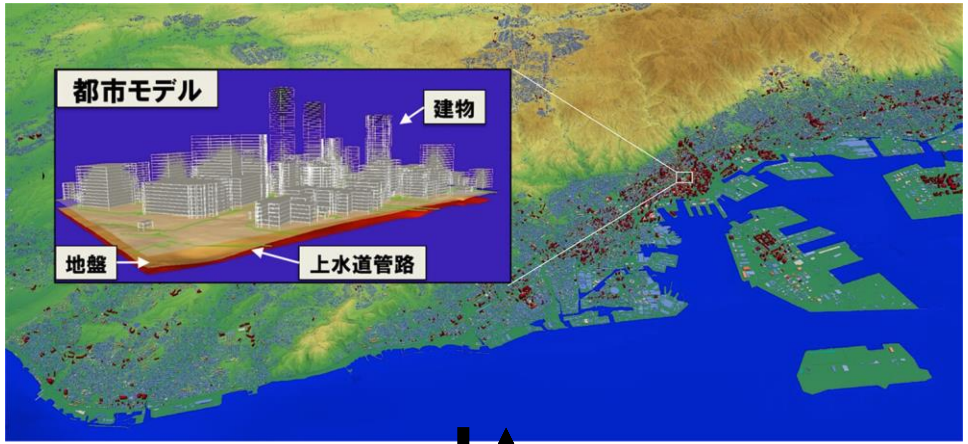
A variety of data and simulations will be connected by DPP and form a framework

Commercial Resource

Official Database

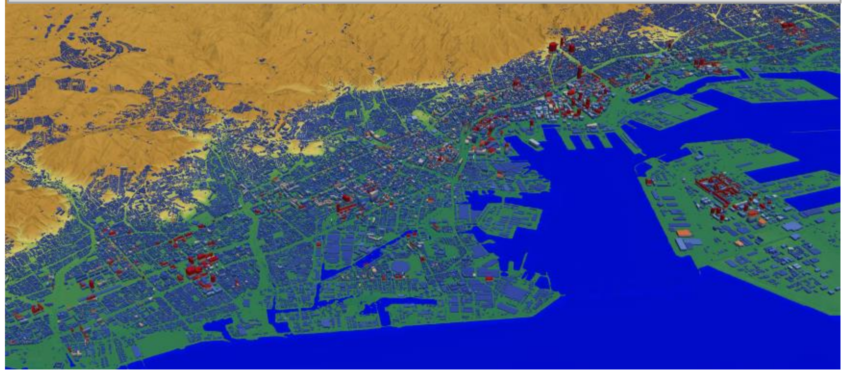
Monitoring Data

Sensor Data



Data Processing Platform (DPP)

Earthquake Simulation



Tsunami Simulation

Evacuation

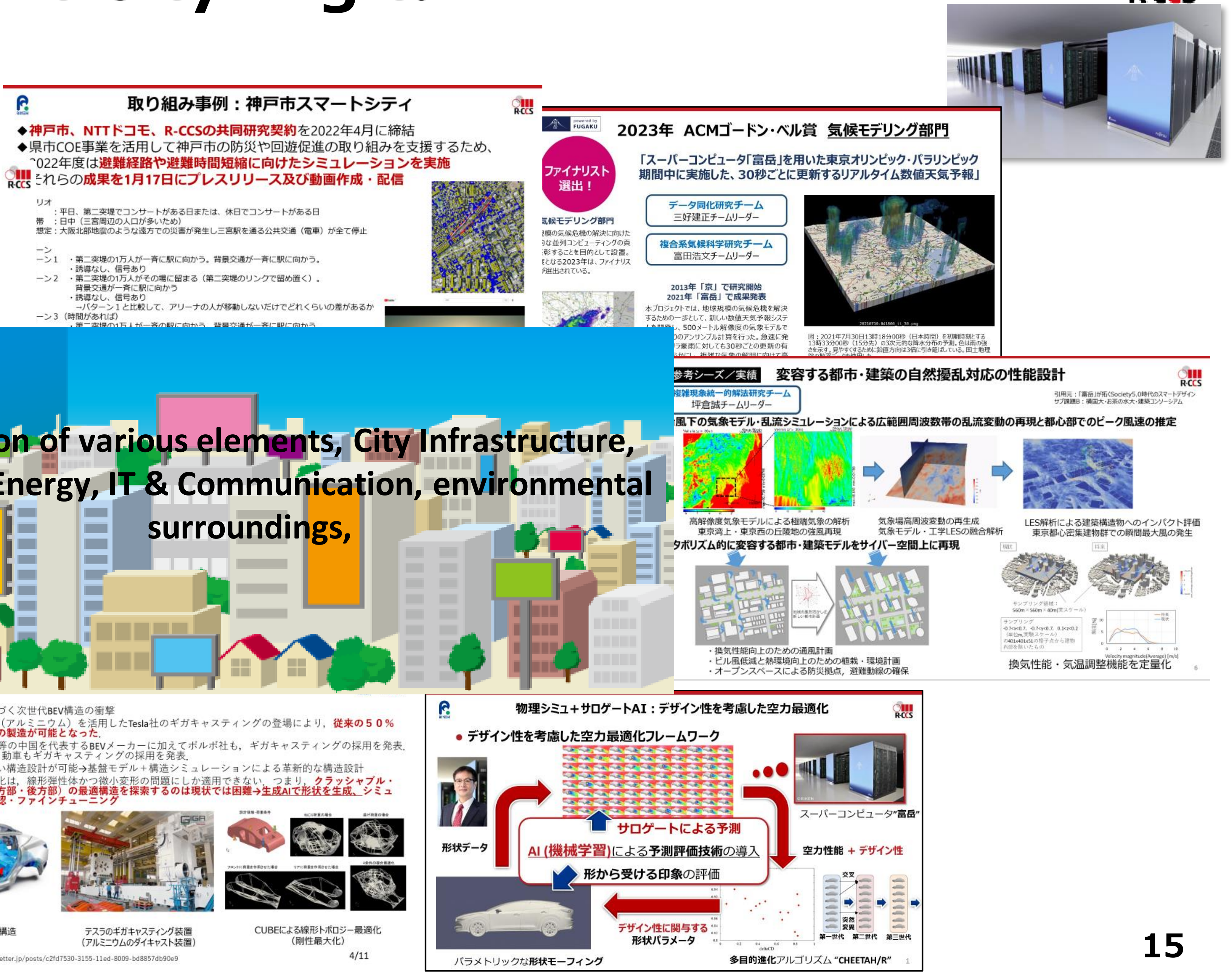
Disaster Response

Social Simulation

Disaster Recovery

Smart City Digital Twin

Integrate the various digital twin elements for smart cities

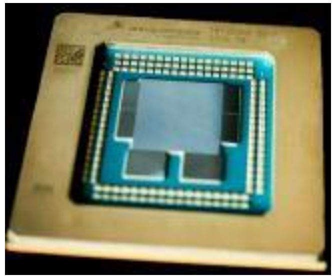


January 2023 MoU Between AWS & R-CCS
Expanding the Scientific Platforms of Fugaku to the Cloud

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’

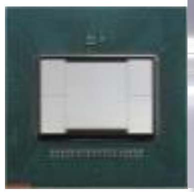
“Virtual Fugaku”
Implementing Fugaku Applications and Software Environment on AWS

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

From Futaku to Virtual Fugaku: De Facto Software Distro for HPC

Widespread distribution of Software and Application outcomes of Fugaku

Fugaku as a 'de facto', ease of user environment for sup
MOU with AWS




User

Fugaku Open On Demand GUI Applications Portal

job

job

R-CCS Fugaku
✓ A64FX

Fugaku Distro VM Image
- Highly Tuned Arm SVE Apps developed for Fugaku Project @R-CCS & others
- OSS/ISV HPC Apps, Desktops, Vis & Workflow, Tools > 50 apps
- Fugaku system SW (LLVM Compilers, HPC libraries, MPI, OneDNN, …)

Fugaku Distro VM Image
- Highly Tuned Arm SVE Apps developed for Fugaku Project @R-CCS & others
- OSS/ISV HPC Apps, Desktops, Vis & Workflow, Tools > 50 apps
- Fugaku system SW (LLVM Compilers, HPC libraries, MPI, OneDNN, …)

R&D to R-CCS, HPC-OSS Community Via Spack

'Satellite Fugaku'

AWS Cloud aws
✓ Graviton3 etc

Fugaku Distro VM Image
- Highly Tuned Arm SVE Apps developed for Fugaku Project @R-CCS & others
- OSS/ISV HPC Apps, Desktops, Vis & Workflow, Tools > 50 apps
- Fugaku system SW (LLVM Compilers, HPC libraries, MPI, OneDNN, …)

AWS Virtual Cluster on VM

Other Supercomputers
✓ Arm, x86, GPU

User

Fugaku Distro VM Image
- Highly Tuned Arm SVE Apps developed for Fugaku Project @R-CCS & others
- OSS/ISV HPC Apps, Desktops, Vis & Workflow, Tools > 50 apps
- Fugaku system SW (LLVM Compilers, HPC libraries, MPI, OneDNN, …)

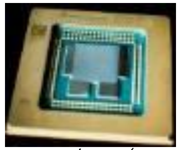
'Private Fugaku'

AWS Cloud aws
✓ Graviton3 etc

Fugaku Distro VM Image
- Highly Tuned Arm SVE Apps developed for Fugaku Project @R-CCS & others
- OSS/ISV HPC Apps, Desktops, Vis & Workflow, Tools > 50 apps
- Fugaku system SW (LLVM Compilers, HPC libraries, MPI, OneDNN, …)

AWS Virtual Cluster on VM

User



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

Fugaku OnDemand

powered by
OPEN OnDemand

Original slide courtesy Rick Stevens, ANL, 2023

Riken AI for Science w/HPC

Large-Scale Observational Models
- Segmentation
- Recognition
- Data Assimilation

Surrogate Models (PINN)
- Parameter Search
- Augmented Simulations
- Optimization

Foundational Models for Science
- Distillation
- Integration
- Protocols
- Writing
- Synthesis
- Q+A

Developing Scientific Codes
- Debugging
- Optimization
- Security
- Translation
- Coding

Controlling complex systems and simulations of digital twins
- Surgery
- Simulation
- Reactors
- Mobility
- Manufacturing

Science Target, Experiment Design & Optimization
- Materials
- Proteins
- Devices
- Industrial Structures

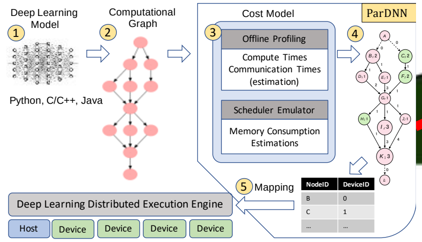
AGI Scientist w/Foundational Models
- Experiments& Simulations Planning
- Instruments & HPC Control
- Discovery & Evaluation

Difference between (traditional) Science of AI vs. AI for Science

- **Solving traditional AI problems:** image recognition, natural language understanding, …
- Focus on **traditional datasets out of traditional AI problems**: natural language text, Internet images that are **abundant in nature**
- Research credit on improving training accuracy etc. => heavy focus on **training**
- Data usually available in corpuses, data pre-scavenged from Internet => **heavy batching** possible in training, making training compute-oriented
- Due to above, machine architecture will focus heavily on compute, making supercomputers w/**GPUs** that facilitate **matrix (tensor) engines that allow dense BLAS** preferred HW
- **Trustworthy AI is less of an issue** for many apps – people accept the inaccuracies
  - Such due to **inherent 'human in the loop'** such as RLHF for LLMs
  - Quick fix e.g. RAG

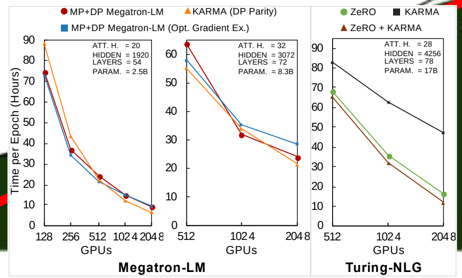
- **Solving difficult scientific problems with AI,** not resolvable with experiments & simulations
- Focus on **new semi-structured scientific data specific to domain**: genomic sequences, molecular structures, physical structures
  - Natural Language serving as semantic 'glue' in multimodal models
- Research credit on making new scientific discoveries => heavy focus on **inference**
- Data may be available from real-time instruments, and/or may not be storable in archive => **real time training** which is bandwidth-oriented
- Due to above, machine architecture will focus heavily on data movement, making traditional supercomputers with **high memory&network bandwidth & capacity** preferred HW
- **Trustworthy AI is of high priority** due to scientific accuracy required
  - **May be automated via loop where simulation will be driving the RL loop**, as is with AlphaGO

Exploring and Merging Different Routes to O(100,000s) Nodes Deep Learning



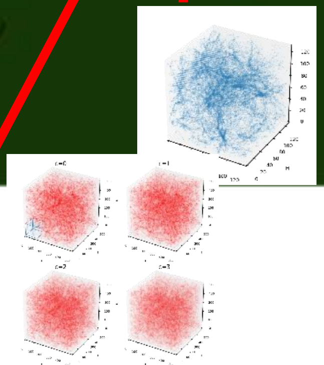
Non-intrusive graph-based partitioning strategy for large DNN models achieving superlinear scaling [1]

AIST, Koc U.



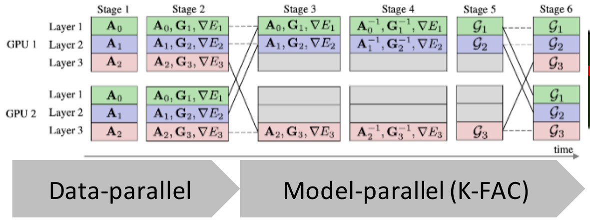
KARMA: Out-of-core distributed training (pure data-parallel) outperforming SoTA NLP models on 2K GPUs [2]

AIST, Matsuoka-lab, RIKEN



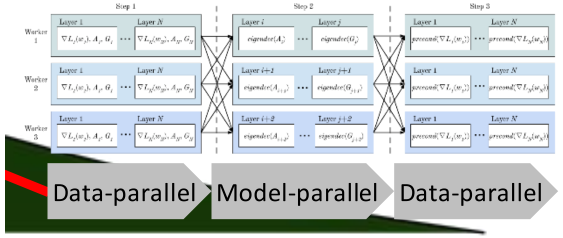
Model-parallelism enables 3D CNN training on **2K GPUs** with 64x larger spatial size and better convergence [3]

Matsuoka-lab, LLNL, LBL, RIKEN



A model-parallel **2nd-order method (K-FAC)** trains ResNet-50 on **1K GPUs** in 10 minutes [4]

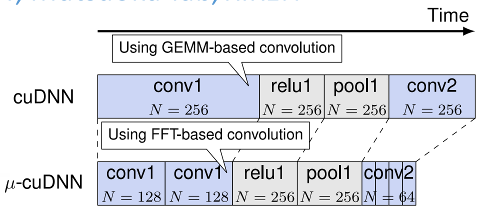
TokyoTech, NVIDIA, RIKEN, AIST



Layer-wise distribution and inverse-free design further accelerate K-FAC [5]

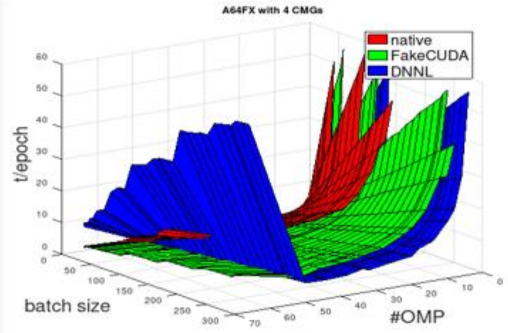
UT Austin, UChicago, ANL

Merging Theory and Practice



Layer-wise loop splitting accelerates CNNs [6]

Matsuoka-lab, ETH Zurich



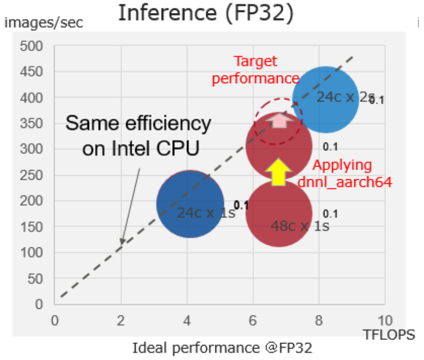
MocCUDA: Porting CUDA-based Deep Neural Network Library to A64FX and (other CPU arch.)

RIKEN, Matsuoka-lab, AIST

Engineering for Performance Foundation

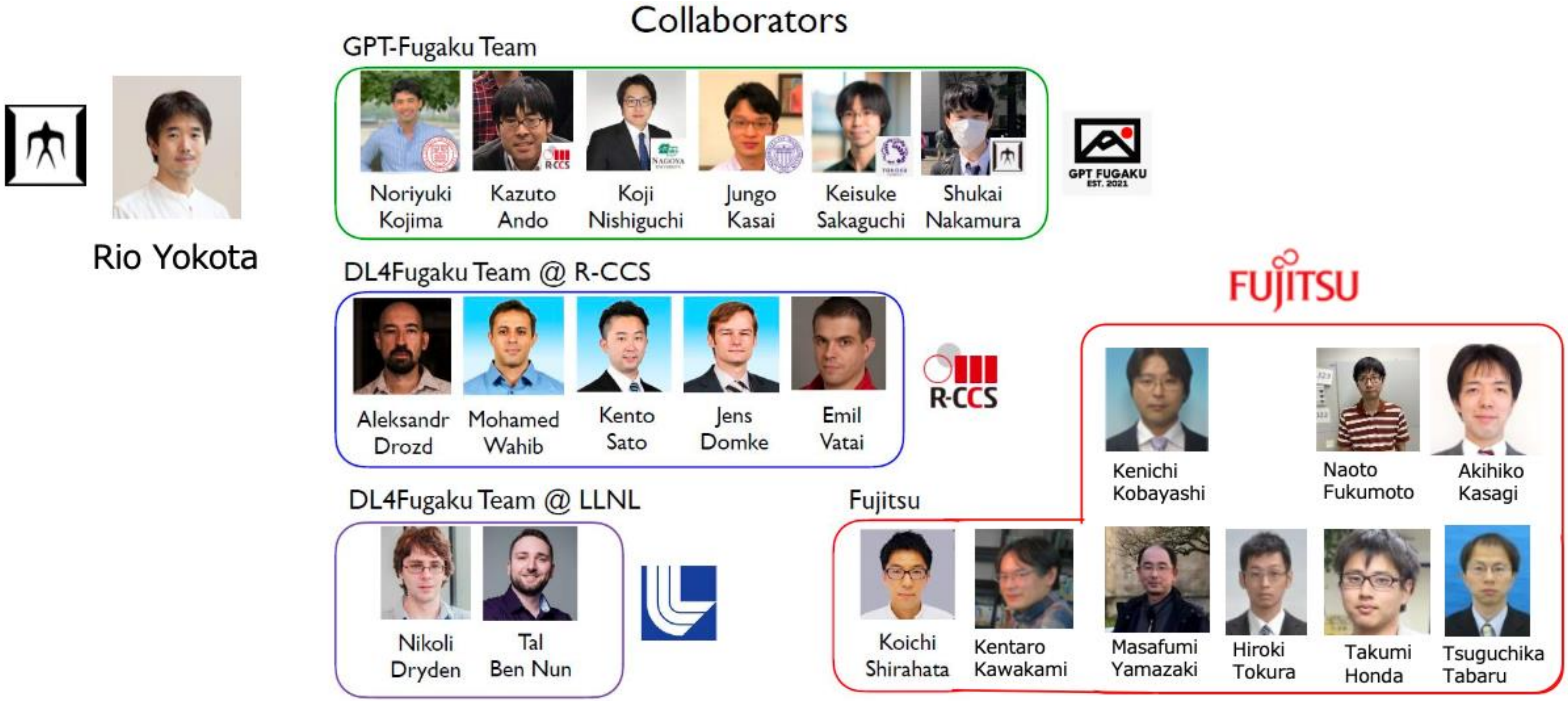
Porting High Performance CPU-based Deep Neural Network Library (DNNL) to A64FX chip

Fujitsu, RIKEN, ARM



[1] M. Fareed et al., "A Computational-Graph Partitioning Method for Training Memory-Constrained DNNs", Submitted to PPoPP21
[2] M. Wahib et al., "Scaling Distributed Deep Learning Workloads beyond the Memory Capacity with KARMA", ACM/IEEE SC20 (Supercomputing 2020)
[3] Y. Oyama et al., "The Case for Strong Scaling in Deep Learning: Training Large 3D CNNs with Hybrid Parallelism," arXiv e-prints, pp. 1–12, 2020.
[4] K. Osawa, et al., "Large-scale distributed second-order optimization using kronecker-factored approximate curvature for deep convolutional neural networks," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 12351–12359, 2019.
[5] J. G. Pauloski, Z. Zhang, L. Huang, W. Xu, and I. T. Foster, "Convolutional Neural Network Training with Distributed K-FAC," arXiv e-prints, pp. 1-11, 2020.
[6] Y. Oyama et al., "Accelerating Deep Learning Frameworks with Micro-Batches," Proc. IEEE Int. Conf. Clust. Comput. ICCC, vol. 2018-September, pp. 402–412, 2018.

GPT-Fugaku


Collaborators
GPT-Fugaku Team
Noriyuki Kojima
Kazuto Ando
Koji Nishiguchi
Jungo Kasai
Keisuke Sakaguchi
Shukai Nakamura
GPT FUGAKU EST. 2021
Rio Yokota
DL4Fugaku Team @ R-CCS
Aleksandr Drozd
Mohamed Wahib
Kento Sato
Jens Domke
Emil Vatai
R-CCS
FUJITSU
Kenichi Kobayashi
Naoto Fukumoto
Akihiko Kasagi
DL4Fugaku Team @ LLNL
Nikoli Dryden
Tal Ben Nun
Fujitsu
Koichi Shirahata
Kentaro Kawakami
Masafumi Yamazaki
Hiroki Tokura
Takumi Honda
Tsuguchika Tabaru

Japanese LLM benchmark (by Weights & Biases)

	run name	Overall	llm-jp-eval	QA	NLI	FA	RC	MR	EL	MC	MT-bench	coding	extraction
25	gpt-4-0613	0.7622	0.6463	0.415	0.768	0.2199	0.891	0.97	0.3004	0.96	8.781	7.8	9.65
31	gpt-4-1106-preview	0.7479	0.6295	0.4002	0.74	0.2388	0.8206	0.96	0.317	0.93	8.663	7.5	9.5
22	gpt-3.5-turbo	0.6701	0.5161	0.2723	0.56	0.1886	0.8406	0.67	0.2913	0.79	8.241	7.9	9
2	anthropic.claude-v2:1	0.6682	0.5188	0.2882	0.676	0.129	0.6575	0.84	0.201	0.84	8.175	8.6	8.7
1	anthropic.claude-v1	0.6387	0.4911	0.3326	0.694	0.1613	0.4951	0.75	0.2244	0.78	7.863	7.1	7.85
3	anthropic.claude-v2	0.6345	0.4834	0.2888	0.642	0.1454	0.4977	0.82	0.2	0.79	7.856	7.45	8.4
4	gemini-pro	0.5851	0.4415	0.421	0.68	0.1745	0.7095	0	0.1856	0.92	7.287	5.55	8.85
12	stabilityai/StableBeluga2	0.5284	0.4111	0.2123	0.654	0.1158	0.8839	0.72	0.2915	0	6.456	4.3	7.1
16	mistralai/Mixtral-8x7B-Instruct-v0.1	0.5006	0.2774	0.1155	0.672	0.0793	0.7563	0.13	0.189	0	7.238	6.9	8.8
15	tokyotech-llm/Swallow-70b-instruct-hf	0.4712	0.5036	0.4803	0.642	0.1797	0.8506	0.7	0.0927	0.58	4.387	2.4	6.95
19	stabilityai/japanese-stablelm-instruct-beta-70b	0.3732	0.2432	0.2975	0.062	0.0558	0.7055	0.55	0.0117	0.02	5.031	3	5.65
29	tokyotech-llm/Swallow-13b-instruct-hf	0.373	0.3716	0.3946	0.454	0.1623	0.7811	0.28	0.049	0.48	3.744	1.2	5.45
21	tokyotech-llm/Swallow-7b-instruct-hf	0.3689	0.3734	0.3947	0.454	0.1591	0.7871	0.27	0.049	0.5	3.644	1.25	5.6
10	rinna/nekomata-14b-instruction	0.3644	0.4375	0.3402	0.494	0.1651	0.8663	0.42	0.0067	0.77	2.912	2.5	2.45
34	stabilityai/StableBeluga-13B	0.3626	0.2965	0.1893	0.572	0.06	0.8114	0.44	0.0029	0	4.288	2.85	7.4
26	stabilityai/StableBeluga-7B	0.3284	0.2567	0.09	0.5	0.052	0.655	0.5	0	0	4	2	5
6	elyza/ELYZA-japanese-Llama-2-13b-instruct	0.3278	0.1506	0.1741	0.128	0.0668	0.5352	0.15	0	0	5.05	2.9	5.3
11	meta-llama/Llama-2-70b-chat-hf	0.3004	0.0783	0.0471	0.152	0.0117	0.3373	0	0	0	5.225	4.05	6.75
20	llm-jp/llm-jp-13b-instruct-lora-jaster-v1.0	0.2947	0.4687	0.5239	0.928	0.0059	0.9229	0.01	0	0.89	1.206	1	1.3
30	llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	0.2927	0.4698	0.5371	0.93	0.0016	0.91	0	0	0.91	1.156	1	1.6

https://wandb.ai/wandb-japan/llm-leaderboard/reports/Nejumi-Leaderboard-Neo--Vmlldzo2MTkyMTU0

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

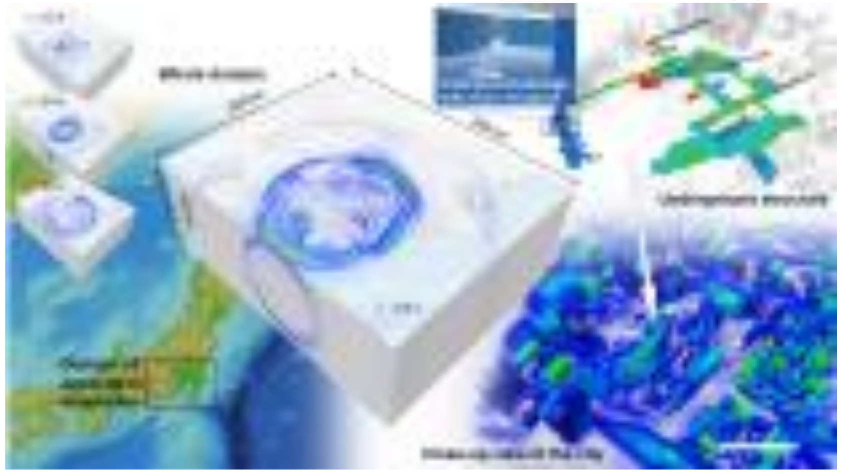
Dream in earthquake simulation

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

Dream has come true!

city is included. All the structures are finely discretized!

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

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

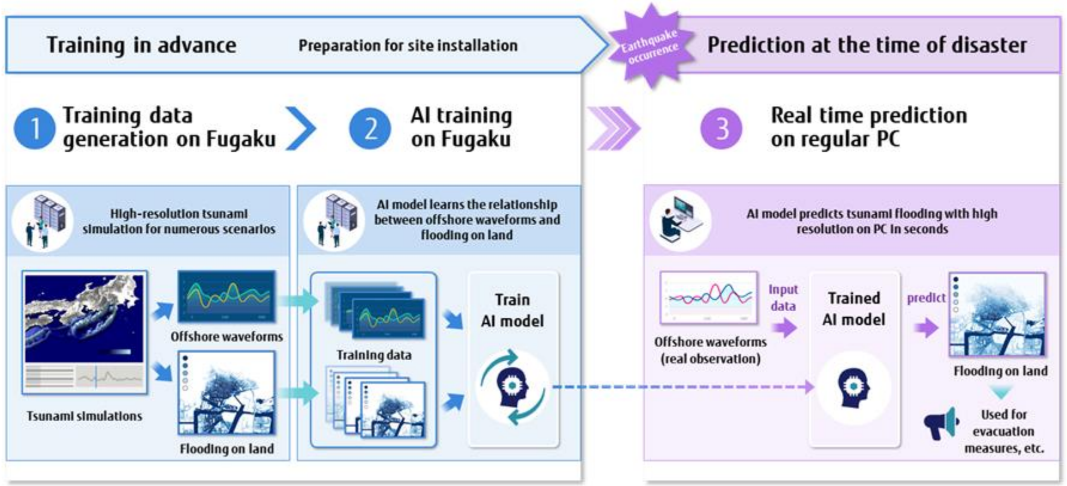


Fig. 1 Overview of tsunami prediction with AI

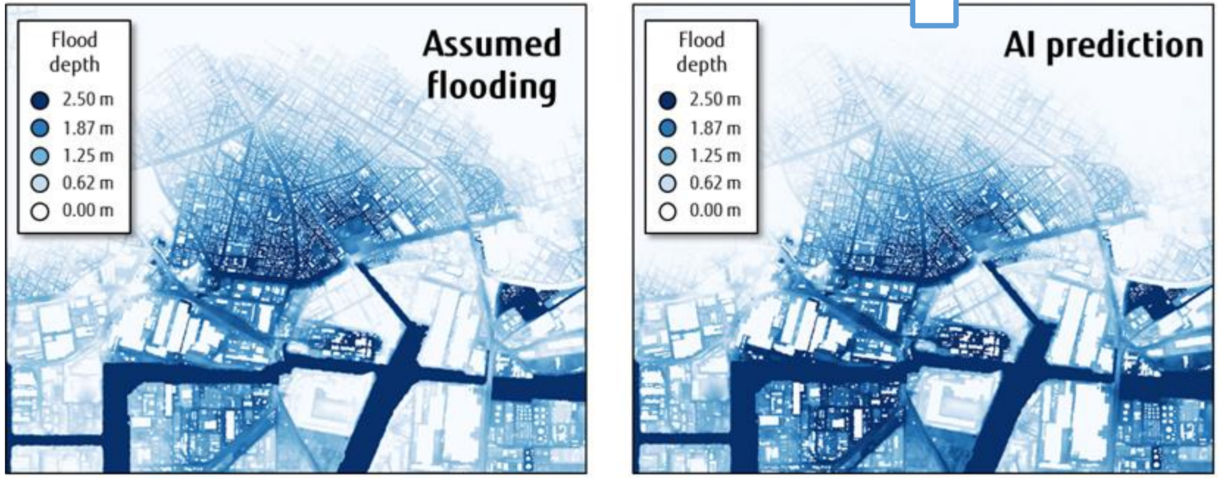
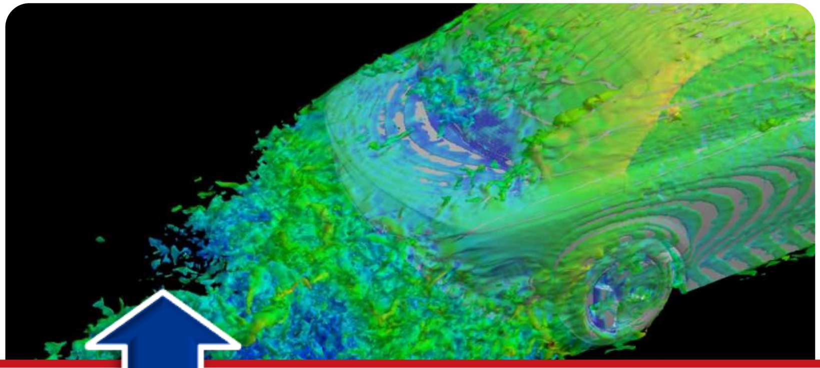


Fig 2. Comparison between anticipated flooding (tsunami source model created by Cabinet Office of Japan with tripled wave heights) of Nankai Trough Megathrust Earthquake and prediction results of newly developed AI

[1] (Press release) International Research Institute of Disaster Science, Tohoku University, Earthquake Research Institute, The University of Tokyo, Fujitsu Laboratories Ltd.Fujitsu Leverages World's Fastest Supercomputer 'Fugaku' and AI to Deliver Real-Time Tsunami Prediction in Joint Project

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

- **Co-optimization Framework**



Supercomputer Fugaku

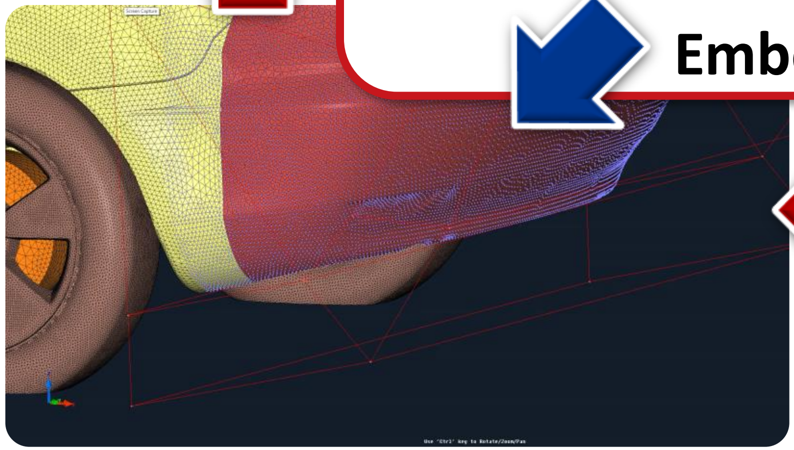
Rapid Generation of CFD Mesh from Shape Data

Ultra Fast Prediction of Drag via Digital Twin

AI-Based Prediction and Optimization

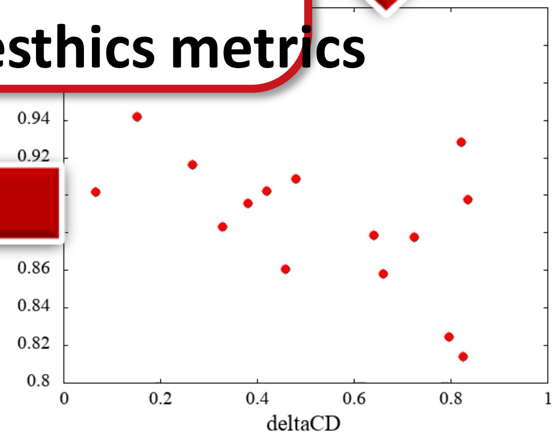
Embedding of human aesthics metrics

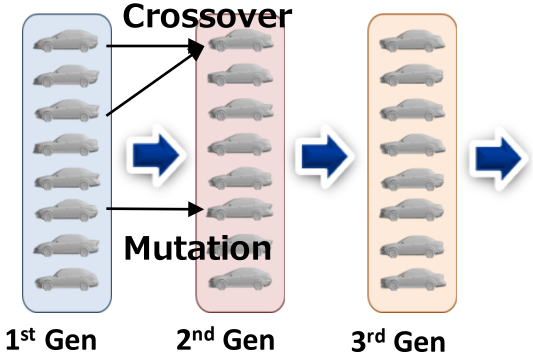
Drag + Aestheics



Parametric Shape Morphing

Shape Parameters on Aesthetics





GA Multi Paramter Optimization "CHEETAH/R"

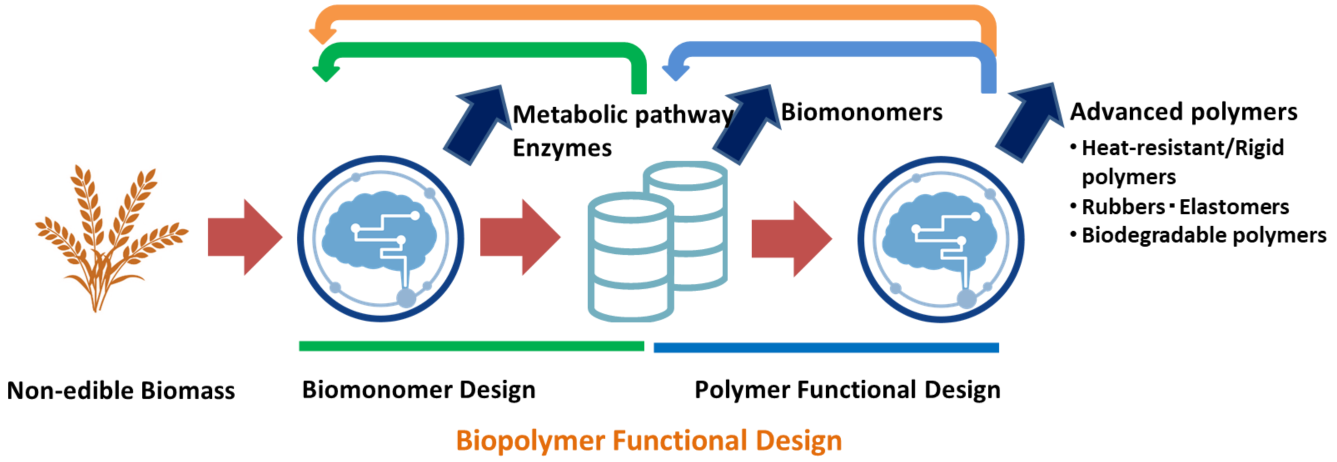
Advanced Material Science Contributing Sustaibability
Simulation + AI + Synthesis

Development of Functional-Design and Production Technologies for Innovative Bio-Materials and Products

Cross-ministerial Strategic Innovation Promotion Program (SIP), CAO, FY2018–FY2022

Objectives: to provide cyber-physical technology to design and rationally produce highly functional materials as high-value products using low-cost sugars obtained from non-edible parts that have been discarded as raw materials

=> Innovative Biodegradable Plastics, Extreme heat resistant polymars

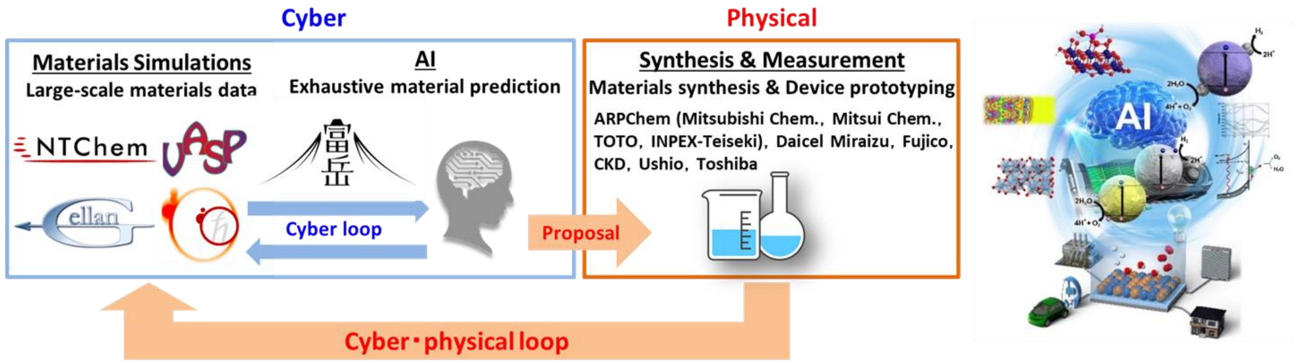


Realization of Innovative Light Energy Conversion Materials utilizing the Supercomputer Fugaku

Program for Promoting Researches on the Supercomputer Fugaku, MEXT, FY2021–FY2025

Objectives: to realize the social implementation of innovative light energy conversion materials by utilizing massive materials simulations and informatics on "Fugaku“

=> New toxicity(Pb)-free perovskite photovoltaic cell material with 25% efficiency (photovoltaic cells everywhere)



Realization of Innovative Photocatalyst for Hydrogen Production
Kobe U・NAIST

Countermeasures against Infectious Diseases by Virus Inactivation
Kobe U・Yokohama CU

Novel Materials Design for High Efficiency Lead-Free Perovskite Solar Cells
RIKEN・Yokohama CU・ENEOS

Riken AI for Science exemplar – AI Pharma Research

「富岳」を用いて創薬研究の最上流である標的探索から化合物創出に至るまで、AI創薬の各種要素技術を統合したプラットフォームを構築。例：標的タンパクに高活性かつ安定な化合物を「富岳」とAI専用計算機を活用し創出。

Pretrained Foudational Model (FM, Training)

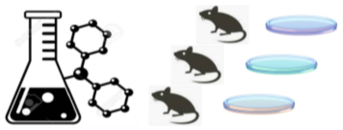
Accelerated FM Training

Training Dataset

Real-time & Automated

Domain Specic Additionally Pretrained Model

Human Feedback



New Domain Specific Training Data

AIの推論

Auto Labeling

(Automated) Preprocessing of Training Data

Reinforcement Feedback

Inference via FM

Test Inference

新規薬剤分子の創出

Highly Precise FM for Science For Innovation (Inference + Sim)

化合物、タンパク質及び、それらの結合情報のDB（ChEMBL, PubChem, DrugBankなど）

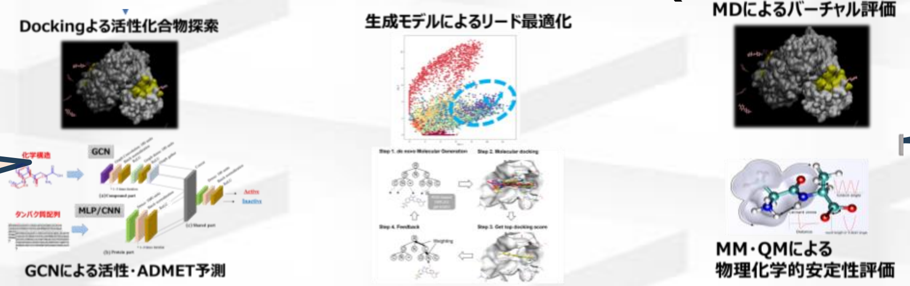
ヒト、ラット、マウス等各生物の全タンパク質配列に対して作用する薬剤分子候補の生成

疾患に関連する標的タンパク質に対して、高活性かつ安定な新規薬剤分子を創出

Reinforcement Learning for Precision

AIの推論

Validation via Simulation (incl. surrogate)



タンパク質の配列から立体構造を予測、活性部位に薬剤候補分子が結合するかどうかをシミュレーション実験により評価

薬剤候補分子の物理的性質（安定性など）の予測・評価

AI for Science Roadmap Urgent Calls You!

Cf Fugaku Feasibility Study 2012-2013

- https://cs-forum.github.io/hpci-aplfs/roadmap-2014/
- **Feasibility Study of 100x Speedup over K by Fugau in 9 areas**

Computational Science Roadmap
-Overview-

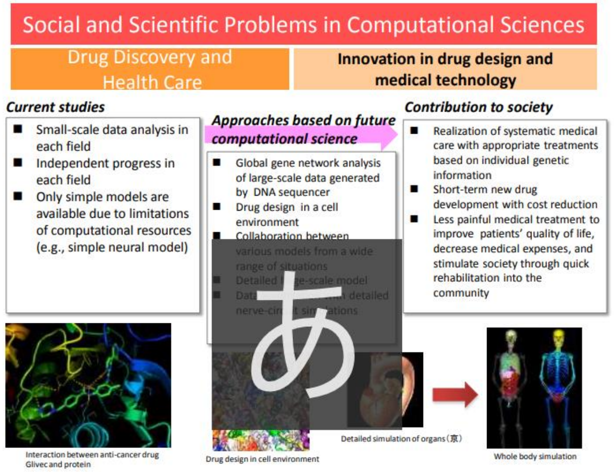
Social Contributions and Scientific Outcomes Aimed for by Innovations through Large-Scale Parallel Computing

May, 2014

Feasibility Study on Future HPC Infrastructures

(Application Working Group)

mechanisms, such as blood clot formation in the heart or brain infarctions, and will be effective in improving patients' Quality of Life (QOL) through the development of minimally invasive treatments, which only pose a slight burden to the patient, and of the medical devices required for these treatments. It will further be effective in revitalizing society through patients' early re-entry into the community and in reducing costs of medical treatment.



The supercomputer's vast computational power will undoubtedly greatly contribute to the development of various aspects in the field of life science, such as detailed neural and cellular simulations, simulations over extended periods of time and space, and almost real-time assimilation[4] of those data. Eventually it could form an important scientific basis for innovative drug design and medical technologies.

The table below lists the computational performance required in the future for the respective areas of drug discovery and healthcare.

[4] One of the methods to merge different observational and experimental data into a numerical model at a high degree.

Subject	Perfor-mance (PFLOPS)	Memory bandwidth (PB/s)	Memory size per case (PB)	Storage size per case (PB)	Elapse Time /Case (hour)	Number of Cases	Total operation count (EFLOP)	Summary and numerical method	Problem size	Notes
Personal Genome Analysis	0.0054	0.0001	1.6	0.1	0.7	200000	2700	Sequence matching	Cancer Genome Analysis: Short read mapping and mutation identification of 200,000 people's genome	1 case = 1 person Integer operations are dominant. "Total operation count"= total instruction count (Total FLOP = 46 EFLOP)
Gene Network Analysis	25	89							s x	
MD and Free-energy calculation for drug design and so on	1000	400							10	B/F=0.4. Supposed to run 100-1000 cases simultaneously. Memory size per case is estimated for a 100 node run.
MD simulations under cellular environments or MD simulations of Virus	490	49							cles	B/F=0.1
Simulations of cellular signaling pathways	42	100							ells	integer operations
Precise Structure-Based Drug Design	0.83	0.14							s in	1TB/s IO speed required to dump 1TB dataset per second
Design of Biological Devices	1.1	0.19							0	1TB/s IO speed required to dump 1TB dataset per second
Multi-scale simulation of a blood clot	400	64	1	1	170	10	2500000	structure interaction with chemical factors	tion size:0.1um, Velocity:10^-2m/s, Delta T:1us	
High Intensity Focused Ultrasound	380	460	54	64	240	10	3300000	Explicit FDM simulation of sound wave and heat transfer	Calculation Area:400mm^3, Grid: 225x10^12, Steps: 1459200, FLOP/grid/step: 1000	
Simulations of Brain and Neural Systems	* 6.9	* 7.6	* 56	* 3600	0.28	100	700	Single compartment model	100 billion neuons, 10000 synapses/neuron, 10^5steps	
Data assimulation of whole insect brain via communication between a phisological experiment and a simulation, Parameter estimator in insect brain simulation	* 71	* 60	* 0.2	* 20	28	20	140000	Multi-compartment HH model with local Crank-Nicolson method, evolutionary algorithm	1000 neurons, 10^6 genes, 100 generations	Supposing 100 MB/s communication to external environment will be required

*Figures marked with a * are still under examination. The website will show more accurate figures as they become available.*

Similar Top-Down AI for Science Fea sibility Study needed for Riken AI for Science and FugakuNEXT

We start immediately to identify the AI (+ simulation) needs for future Science driven by AI in a common format => R-CCS researchers expected to be main contributor

Example HPC & AI Infrastructure Phase 1
Extensive re-use of Existing Fugaku Assets=>FugakuNEXT

Current Fugaku Resources

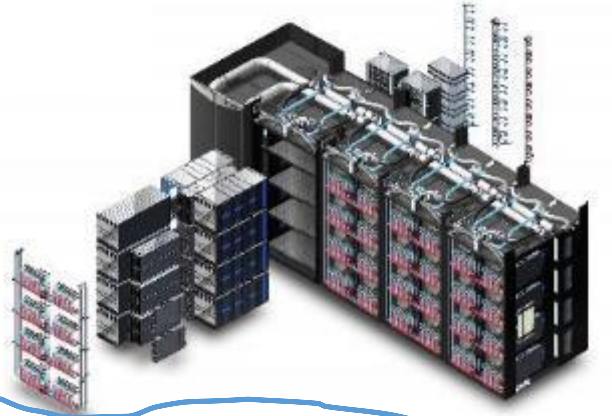
HPC Supercomputer "Fugaku"
HPC: 163PetaBytes/s memory bandwidth (No.1 currently)
Foundation model training: 2 Exaflops FP16
Operational Power: 16~20MW
Inference to be enhanced exploiting world's top mem BW

> 20Terabps

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

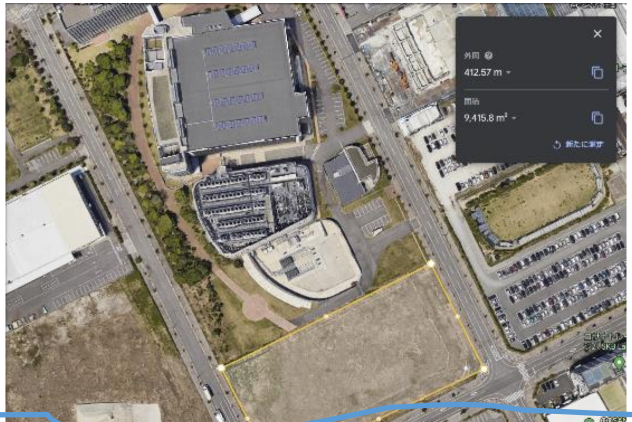
> 20Terabps

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

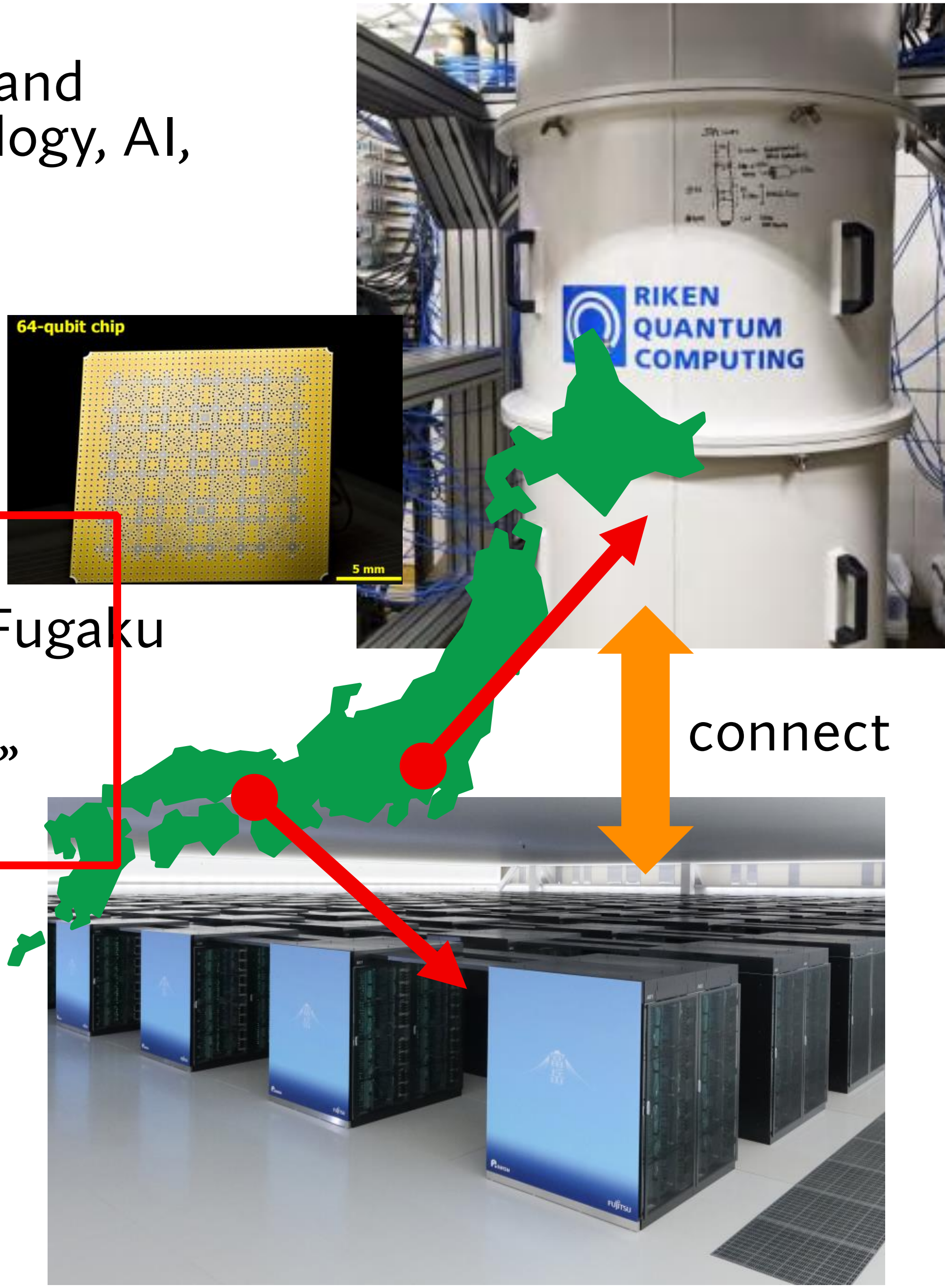
R-CCS DC Facility
> 40MW Power & Cooling



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

RIKEN & Quantum Computing Research in RIKEN

- **RIKEN** is a comprehensive research organization for basic and applied science , founded in 1917. 10+ centers, Physics, Biology, AI, etc..
- RQC: Center for Quantum Computing (since 2021)
  - Superconducting Quantum Computer
  - Optical Quantum Computing
  - Theoretical Computing Theory of Quantum
- **R-CCS**: Center for Computational Science (since 2010)
  - Quantum Computer Simulator on the supercomputer Fugaku
  - Hybrid of Quantum computer and Fugaku
  - Feasibility Study of the Quantum computing and “next” Fugaku systems
- iTHEMS: Interdisciplinary Theoretical and Mathematical Sciences Program
  - Theoretical Computing Theory of Quantum
- AIP: Center for Advanced Intelligence Project
  - Quantum Machine Learning



(Slide partly courtesy of Rick Stevens, ANL)

How will QC and (Classcial) SC collaborate?

- How to resolve if quantum can contribute meaningfully to solving real problems faster than CLASSICAL supercomputers
- Current state small QC machines, unreliable, “circuit” model for programming, lack of error correction, lack of a good number of killer apps (and superpolynomial speed up candidates), ad hoc integration strategies

We need machines with 1,000’s of virtual-reliable qubits (1K-10K) able to run programs/circuits of depth $O(10^{10})$-$O(10^{12})$ $\Longrightarrow$ > 1M physical qubits and ~ 2 weeks of running at ~ns clocks

We need algorithms for problems better than quadratic speedups)

Scientific Analysis (not Hype) of Utility of Quantum Computing

- **For practical 'quantum supremacy', exponential speedup cf classical algorithm is necessary**
  - Many algorithms only achieve quadratic speedup, thus will lose to classical in practice
    - E.g., Shor's algorithm – exponetional => Good
    - E.g., Grover's algorithm – quadratic=>NG
- **For 'pure' quantum algorithms, none exist that exhibit quadratic speedup & can be executed practically on current NISQ machines w/~100 qubits**
  - Shor's algorithm may break RSA 2048 in the far future but will require 20~200mil NISQ qubits https://arxiv.org/pdf/1905.09749.pdf
- **Hybrid algorithms e.g., variational algorithms (e.g. VQE) might be useful in much closer future**
- ***Require platform to conduct scientific analysis of QC, as large qubits as possible, using real state-of-the art real machines and simulators!***

(To be published in *Communication of the ACM*)

Disentangling Hype from Practicality: On Realistically Achieving Quantum Advantage

TORSTEN HOEFLER, Microsoft Corporation, USA and ETH Zurich, Switzerland
THOMAS HÄNER and MATTHIAS TROYER, Microsoft Corporation, USA

Quantum computers offer a new paradigm of computing with the potential to vastly outperform any imagineable classical computer. This has caused a gold rush towards new quantum algorithms and hardware. In light of the growing expectations and hype surrounding quantum computing we ask the question which are the promising applications to realize quantum advantage. We argue that small data problems and quantum algorithms with super-quadratic speedups are essential to make quantum computers useful in practice. With these guidelines one can separate promising applications for quantum computing from those where classical solutions should be pursued. While most of the proposed quantum algorithms and applications do not achieve the necessary speedups to be considered practical, we already see a huge potential in material science and chemistry. We expect further applications to be developed based on our guidelines.

ACM Reference Format:
Torsten Hoefler, Thomas Häner, and Matthias Troyer. 2022. Disentangling Hype from Practicality: On Realistically Achieving Quantum Advantage. 1, 1 (September 2022), 7 pages. https://doi.org/XXXXXXX.XXXXXXX

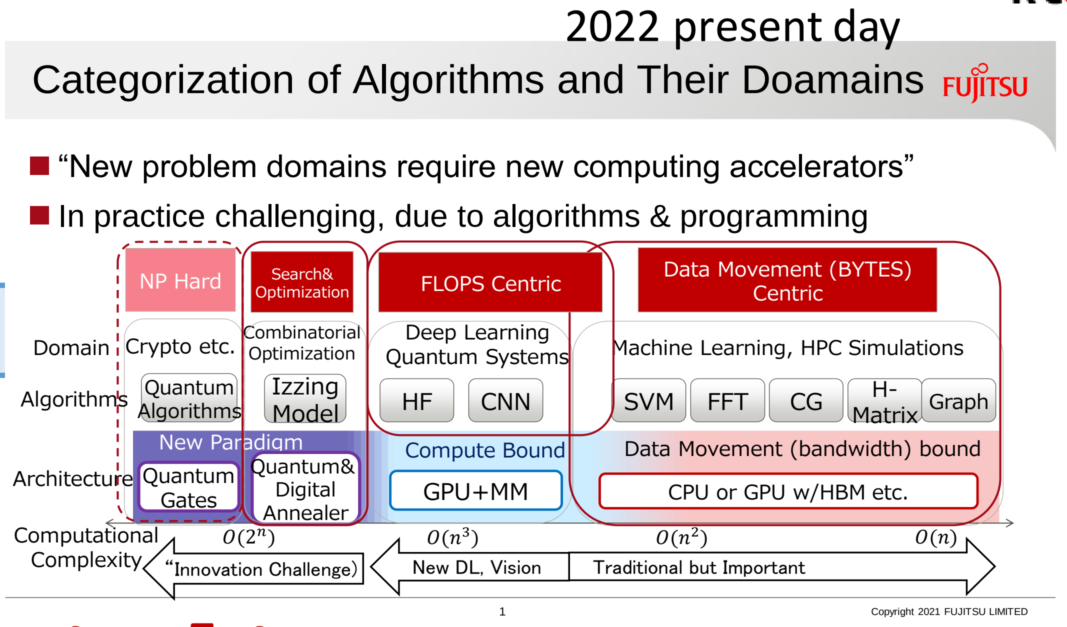
Practical and impractical applications. We can now use the above considerations to discuss several classes of applications where our fundamental bounds draw a line for quantum practicality. The most likely problems to allow for a practical quantum advantage are those with exponential quantum speedup. This includes the simulation of quantum systems for problems in chemistry, materials science, and quantum physics, as well as cryptanalysis using Shor's algorithm [13]. The solution of linear systems of equations for highly structured problems [10] also has an exponential speedup, but the I/O limitations disc and undo this advantage if knowledge of the full solution is required (as opp obtained by sampling the solution).

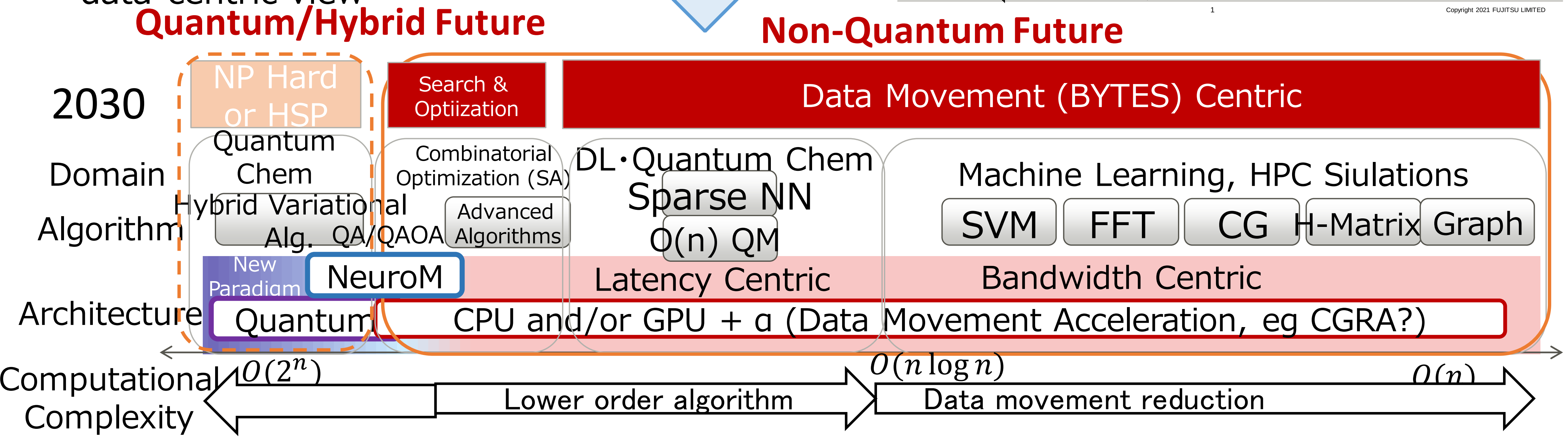
Equally importantly, we identify dead ends in the maze of applications. quadratic quantum speedups, such as many current machine learning tra design and protein folding with Grover's algorithm, speeding up Monte walks, as well as more traditional scientific computing simulations includ systems of equations, such as fluid dynamics in the turbulent regime, weat achieve quantum advantage with current quantum algorithms in the fores the identified I/O limits constrain the performance of quantum computing linear systems, and database search based on Grover's algorithm such that

These considerations help with separating hype from practicality in the can guide algorithmic developments. Specifically, our analysis shows that to focus on super-quadratic speedups, ideally exponential speedups and 2 bottlenecks when deriving algorithms to exploit quantum computation be *quantum practicality are small-data problems with exponential speedup, an problems in chemistry and materials science.*

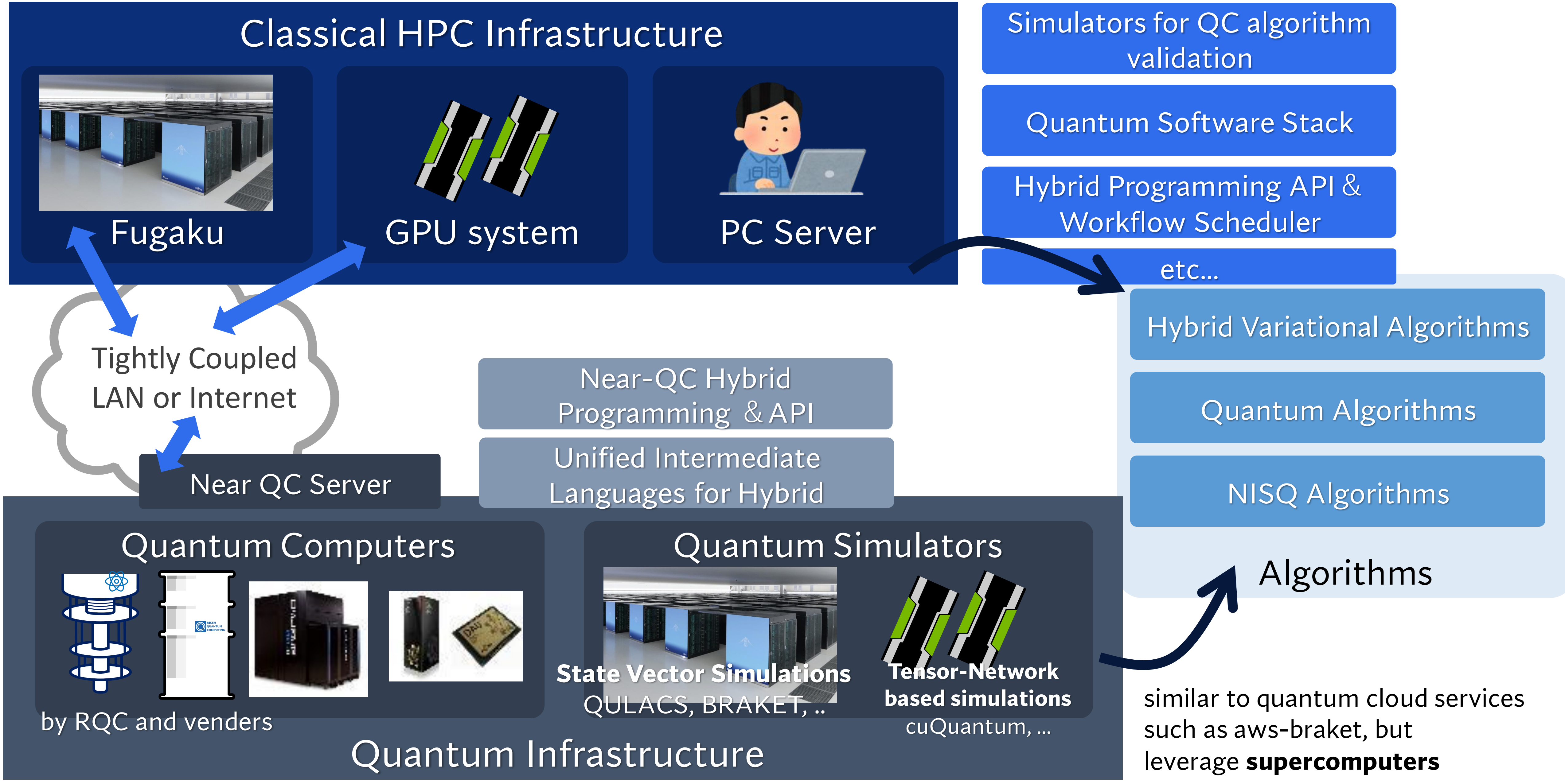
Non-Quantum and Quantum Future Workload Characterization

- **Towards 2030 Post-Moore era**
  - End of ALU compute (FLOPS) advance
  - Disrupritve reduction in data movement cost with new devices, packaging
  - Algorithm advances to reduce the computational order (+ more reliance on data movement)
  - Unification of BD/AI/Simulation towards data-centric view





Quantum-HPC hybrid platform in R-CCS (2024~)




IBM
IBM Quantum
System Two


IBM Quantum
HERON
133 QUBITS
TUNABLE-COUPLER

IBM Research technology schematic

IBM Quantum: the path to Blue Jay system

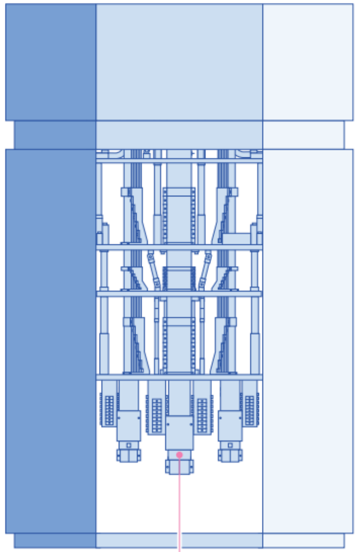
Abstract

Scaling to achieve universal quantum computation.

Development team

IBM Quantum
Yorktown Heights, NY

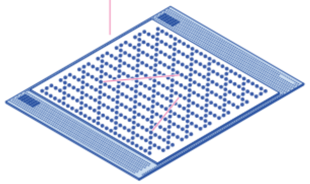
2026



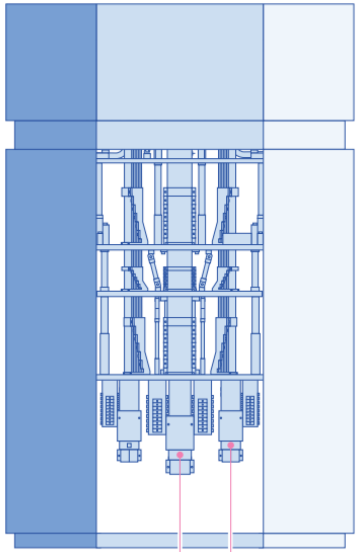
Kookaburra system

Logical memory

→ C-coupler
→ Degree 6 gates
→ 1,200 control lines



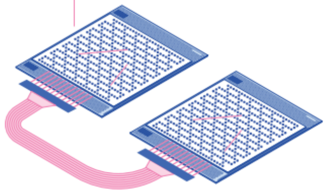
2027



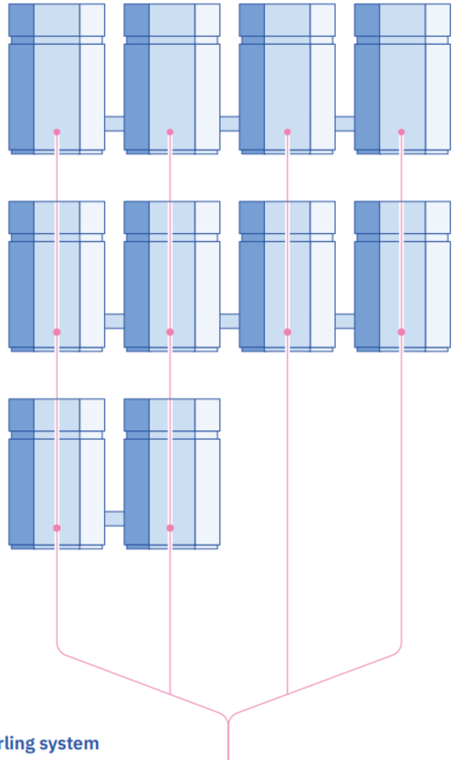
Cockatoo system

Logical operations

→ L-coupler
→ Logical commiunication
→ 3,600 control lines



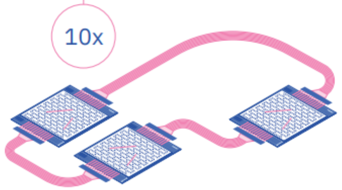
2029



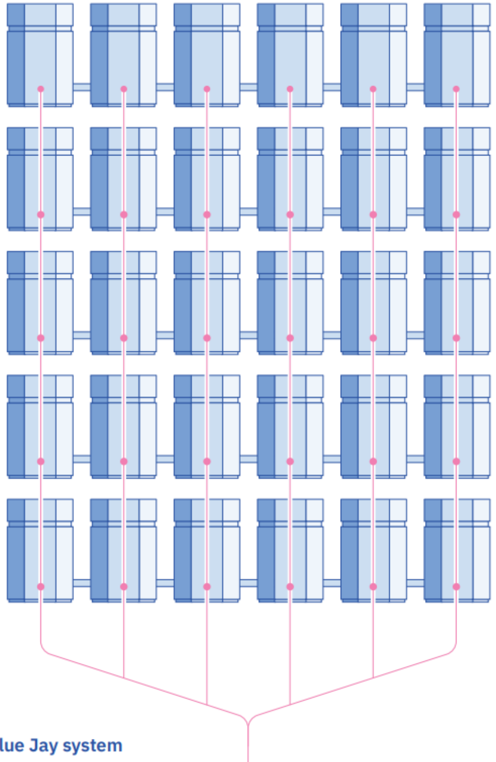
Starling system

100 million gates

→ Gen-3 flex
→ FPGA control
→ Universal computation
→ 40K control lines



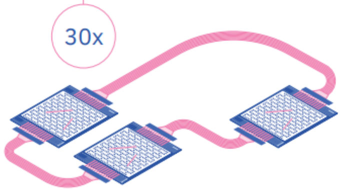
2033



Blue Jay system

1 billion gates

→ Gen-4 flex
→ ASIC control
→ Universal computation
→ 400K control lines



FugakuNEXT Feasibility Study (System Research by RIKEN)

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, simulation techniques, 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

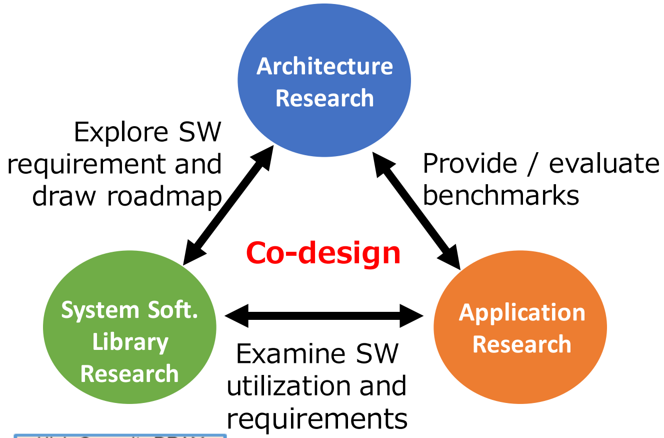
- Investigating technological possibilities (such as 3D stacked mem, accelerators, chip-to-chip direct optical link) and performance of the entire system or its components based on trends in semiconductor and packaging technologies
- Predicting future system performance based on performance analysis 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

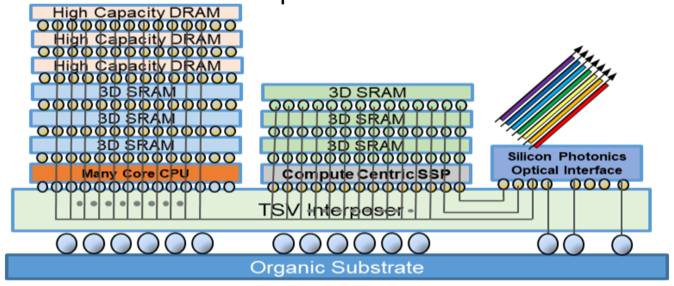
Research on Applications

- 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 Q3	2022 Q4	2023 Q1	2023 Q2	2023 Q3	2023 Q4	2024 Q1
Architecture	Explore device/architecture technology			Performance estimation with benchmarks		Architecture study	
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 algorithm improvement	



Strawman processing element architecture

Organization Chart of System Research by RIKEN

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

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

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)】

Architecture Research sub-G2
Fujitsu Ltd. (Co-I institution)
【SGL: Shinjo】

Architecture Research sub-G3
Intel Corporation (Co-I institution)
【SGL:Yazawa】

Architecture Research sub-G4
AMD Inc. (Co-I institution)
【SGL:Yoshida】

Architecture Research sub-G5
NVIDIA Corporation (Collaborator)
【SGL:Wells】

Architecture Research sub-G6
Hewlett Packard Enterprise (Collaborator)
【SGL:Negishi】

Architecture Research sub-G7
Arm Ltd. (Collaborator)
【SGL:Lecomber】

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】

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
- **Collaboration with other groups**
  - Responding to surveys from Architecture and System Software research groups

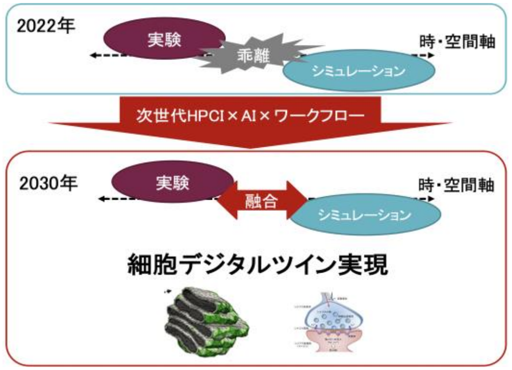
List of Benchmark Applications in RIKEN Team

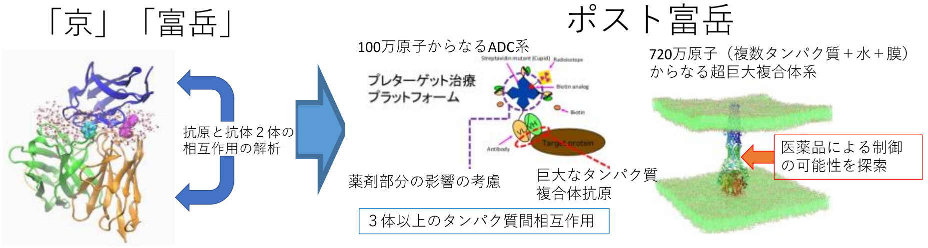
- **Initial application list for estimating performance of future architectures**
  - More benchmark applications will be evaluated at a later stage

Area	Application	Type	Language	GPU	Note
Life Science	GENESIS	MD (particle)	Fortran	working	strong-scalability oriented Mixed precision
New Material & Energy	SALMON	DFT, Stencil, FFT	Fortran	✓	high-precision GEMM required Possible Emulation w/ME
Weather and Climate	SCALE-LETKF	CFD (structured mesh)	Fortran	working	
Earthquake & Tsunami Disaster Prevention	EbE-method	FEM (unstructured mesh)	C++	✓	
Manufacturing	FrontFlow/blue	FEM (unstructured mesh)	Fortran	working	
Fundamental Science	LQCD-HMC-DWF	Stencil, SpMV	C++	working	
AI	Hugging Face GPT-2 XL	Transformer	PyTorch	✓	1.5B parameters Single node
AI	Megatron-LM DeepSpeed	Transformer	PyTorch	✓	70B parameters Multi node
AI	???	Transformer (Inference)	PyTorch	✓	Unbatched

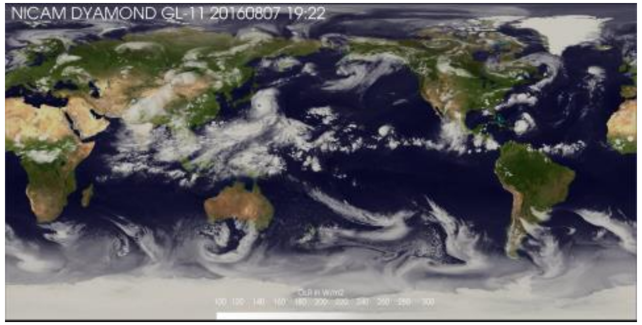
Roadmap of Target Sciences in FugakuNEXT Era

- **Case for life science area**
  - **Cell digital-twin by simulation x AI x experiment**
    - Now takes 8333 days with 16386 nodes in Fugaku for 10us simulation -> shortening to 2-3 months by 100x performance improvement.
  - **Fully automated drug discovery**
    - Mutual interactions analysis of two particles in Fugaku. -> analysis of multi particles for large complex antigens protein etc. in FugakuNEXT towards a practical antigen design framework.

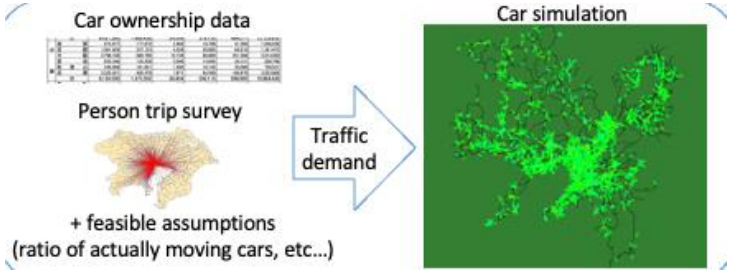




- **Case for weather/climate science area**
  - **Atmospheric digital-twin by high-resolution prediction model**
    - Analysis of Japan area for 10h ahead of time with 2km horizontal resolution -> 18h ahead of time with 200m horizontal resolution in 2030.
  - **Global Cloud-Resolving and Ocean-Eddy-Resolving Models for 100-Year Climate Simulation**
    - Atmospheric horizontal resolution of 3.5km and vertical resolution of 78 layers with 100 year integration. Refine understanding and prediction of El Niño, typhoons, etc. associated with climate change. Reducing uncertainty in climate sensitivity.



- **Case for social science area**
  - **Traffic simulation of entire Japan**
    - Now only Kinki-region simulation -> Simulation for whole Japan including prediction of disaster impact propagation with economical mutual interactions.

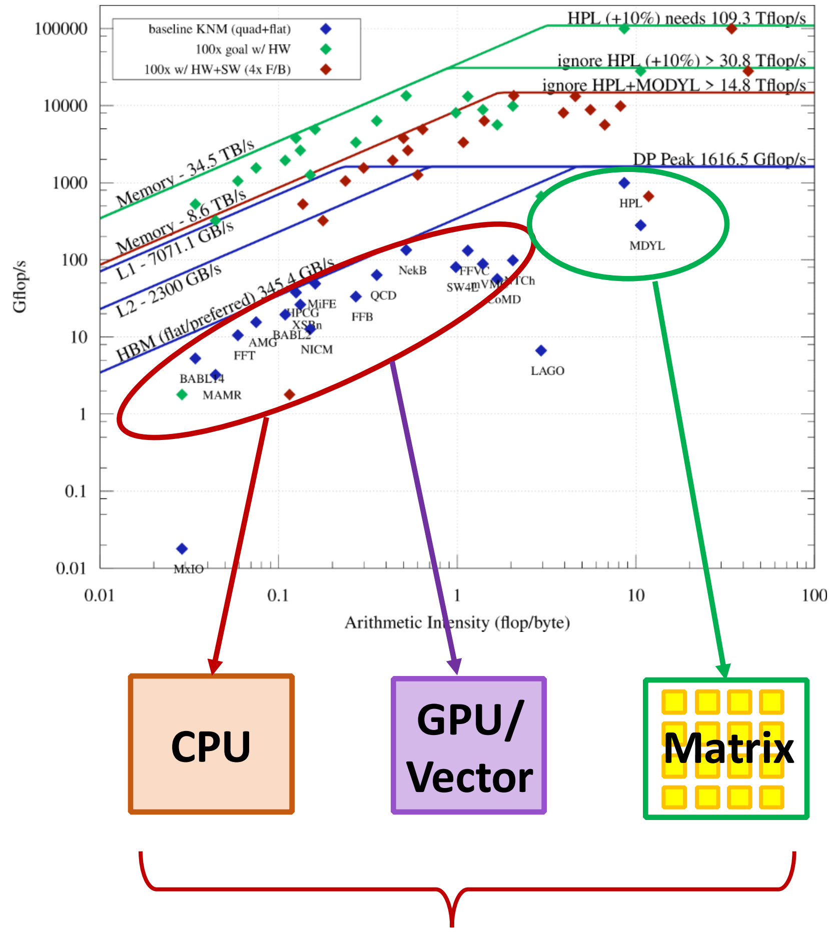


Key Research Item for Node Architecture Selection

- **Needs for a power-efficient compute node**
  → Exploration of accelerators
  - Truly useful accelerator for HPC and AI workloads
  - HPC→Memory bound
  - AI Training→Compute bound, AI Inference→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

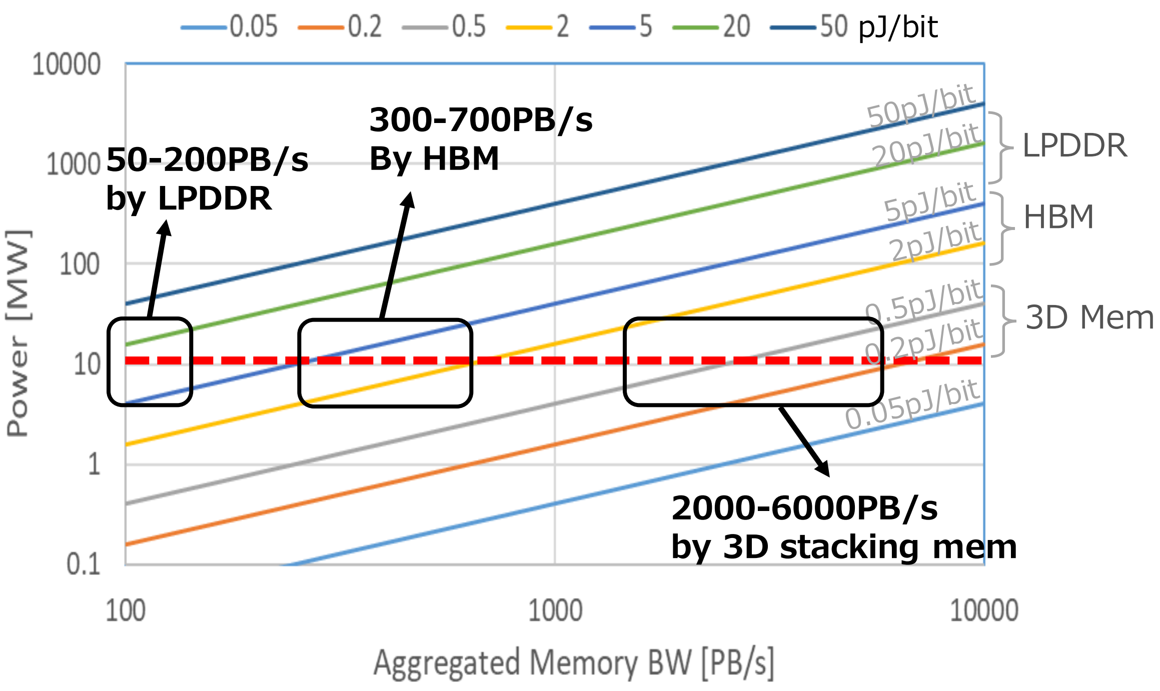
Roofline analysis on A64FX

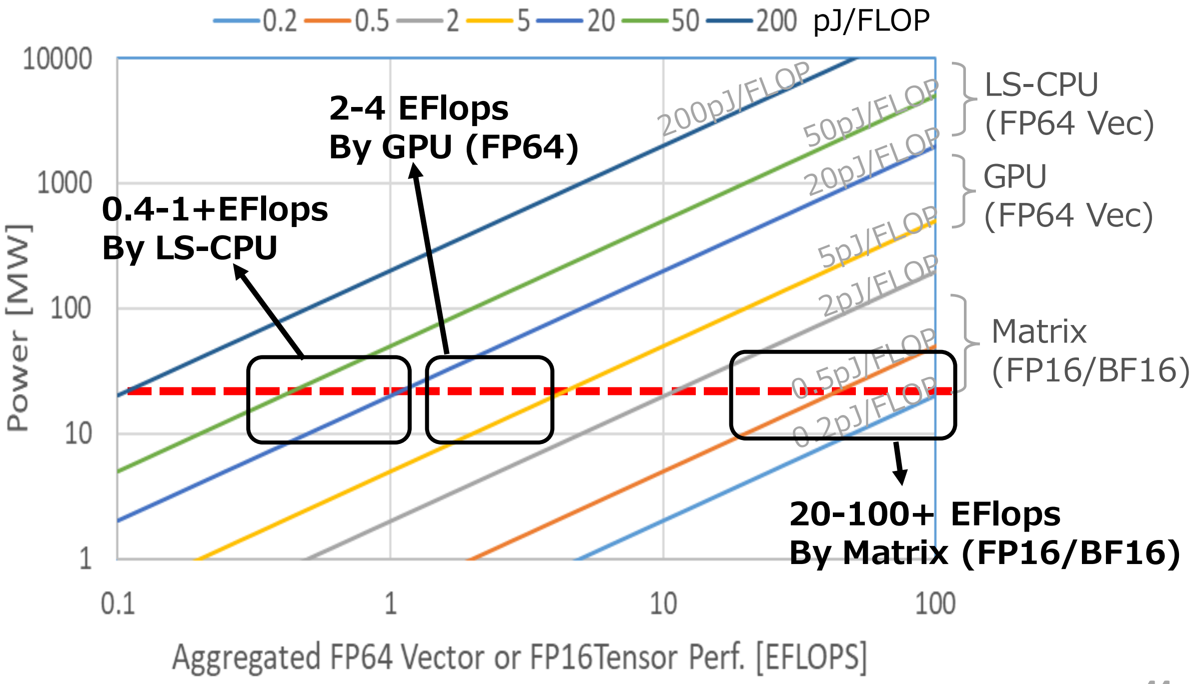


Need to find the optimal balance

Performance Projection in Power Constrained Scenarios

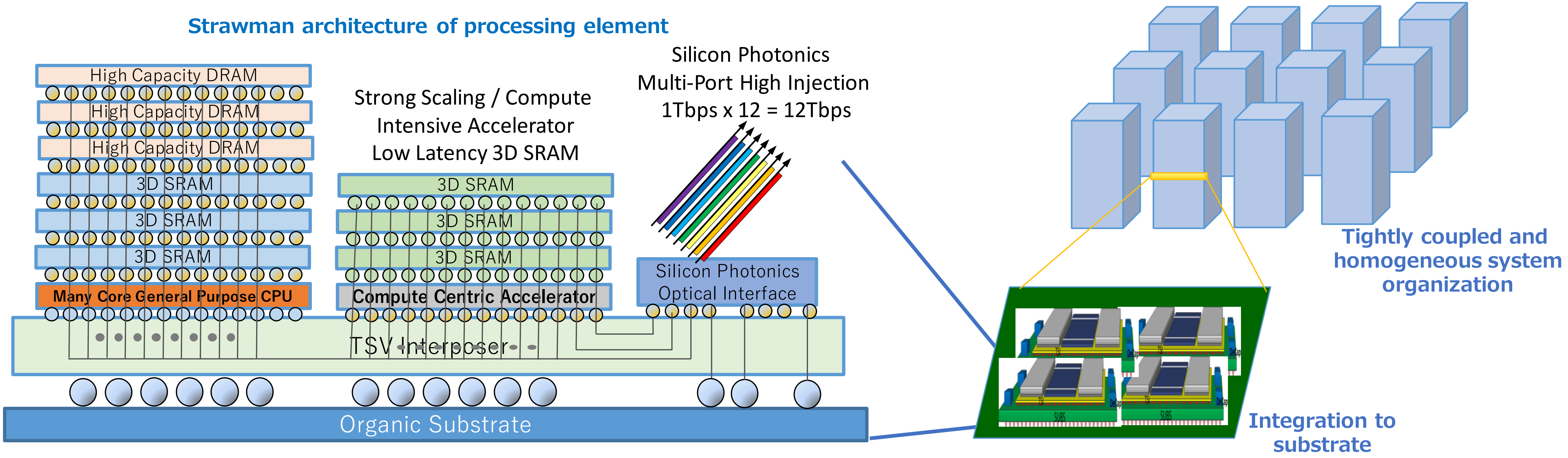
- **Estimated energy per operation on current and future technologies**
  - Based on historical trend obtained by publically available data
  - Not related to any partner vendors' perspective
- **Case for 30MW power budget (10MW for memory and 20MW for compute)**
  - Network is omitted for simplicity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





A Direction toward Next-Generation Computational Infrastructure

- **Initial key architectural directions**
  - Paradigm shift in architecture-algorithm toward "FLOPS to Byte (data movement efficiency)"
  - Significant increase in relative memory bandwidth using 3D stacked memory technology
  - 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 as a technology driver

Implementation Approaches for Node Architectures

- **Candidates of packaging technologies**

Technical difficulty: Low ⟷ High

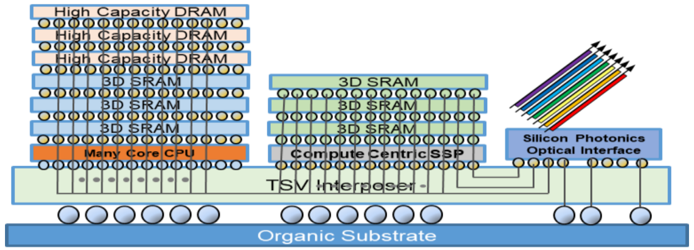
Power efficiency of data movement: Low ⟷ High

chip-to-chip connection (chiplets)	**Monolithic die (conventional)** HBM, HBM, CPU, Acc, HBM, HBM	**Chiplet-based (becoming main-stream)** CPU/ Acc, CPU/ Acc, I/O	**More aggressive chiplet-based (Future direction)** HBM, CPU, Acc, HBM HBM, Acc, Acc, HBM
3D stacking approaches	**2.5D connection (conventional)** HBM, HBM, CPU / Acc, HBM, HBM	**3D - Hybrid Bonding (single chip stacked)** HBM, HBM, CPU / Acc, 3D Memory, HBM, HBM	**3D implementation (multi chips stacked)** HBM, HBM, CPU / Acc, 3D Memory, HBM, HBM
Optics	**AOC (conventional)** 	**Silicon-Photonics - chip-to-chip optical connection (various technology candidates incl. WDM)** HBM, HBM, CPU / Acc, Si Photo, HBM — Si Photo, HBM, CPU / Acc, HBM, HBM	

In fact, inference is GEMV (Albeit in low precision)

- **For one-shot LLM inference, more than 80% of time is low precision GEMV (source Samsung)**
- **For very large models, memory capacity (1 billion parameters = 1TB) & memory bandwidth (for 30ms respond per token for real-time response to natural language queries, 30TB/s) are the bottlenecks**
  - It is reported that, OpenAI uses 128 A100 GPU supercomputer to do the GPT inference

- **Challenge: Can we build a single chip with memory system with 1TByte mem capacity & 30TByte/s mem BW ?**

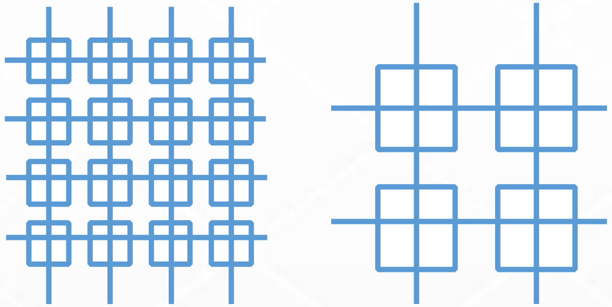


Innovative memory subsystem for FugakuNEXT

What about Dense Linear Algebra?

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

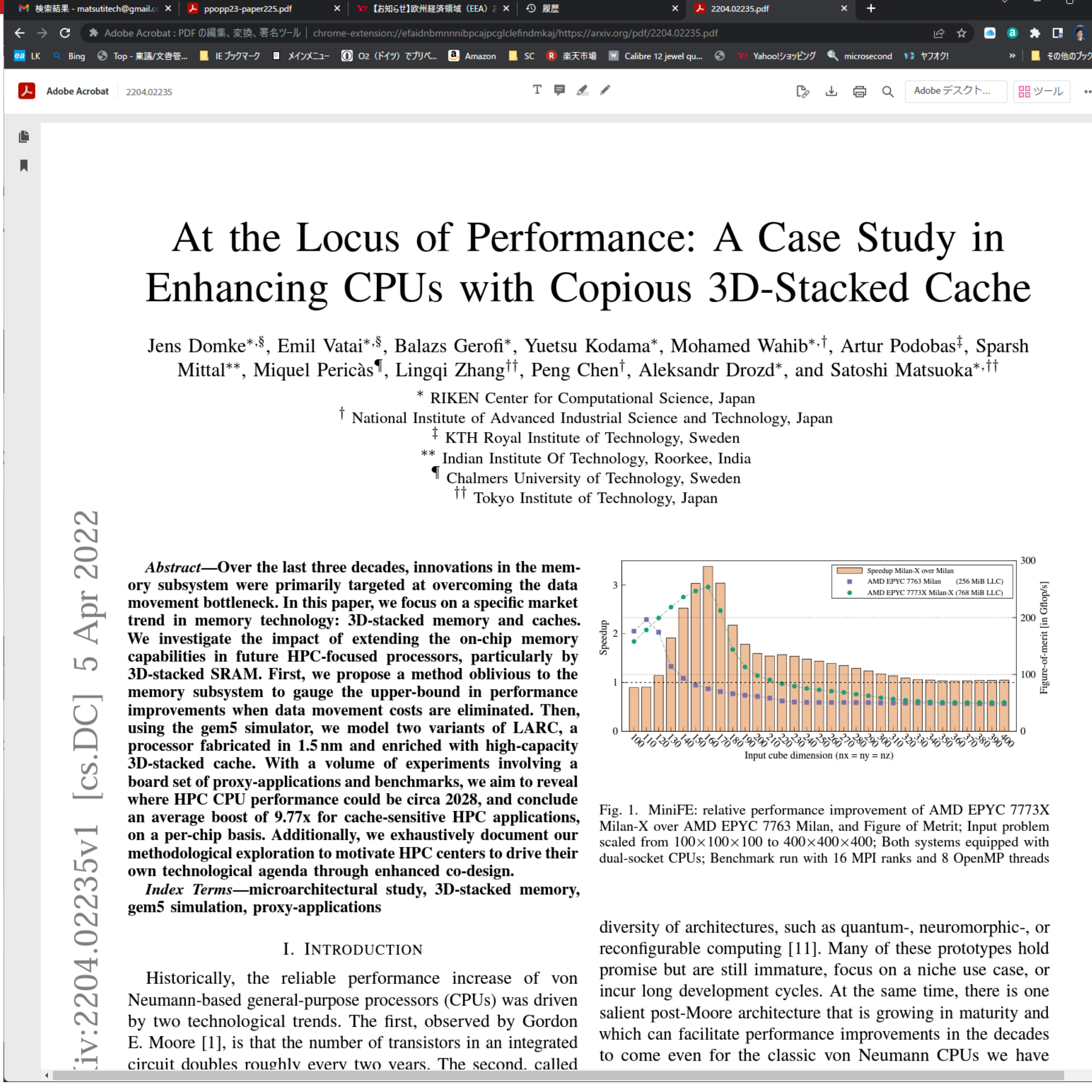
- Energy = compute (multipliers, volume) + data movement (between units, surface)
  - Low precision – low surface:volume, worthwhile to optimize to minimize data movement, matrix engines helpful to minimize wire distance
  - High precision – high surface:volume, data transfer less problem, performance & energy gain small, dark silicon of unused multipliers wasteful, wide vectors sufficient?
- 8~16 bit apps: Deep Learning/AI training, some higher order methods? => Emulation of "64 bit" apps with various methods
- 19~ (TF32) ~ 32 bit apps: DL/AI, molecular dynamics, higher order methods (mixed precision)
- 64 bit apps: first-principle material science e.g., DFT



Low precision MM	High precision MM
Low volume (compute) : surface (comm) ratio	high volume (compute) : surface (comm) ratio
Matrix units help to reduce data transfer energy	Vector units may be sufficient as benefit of matrix may be low

'LARC' Next Gen Mammoth BW CPU

- https://arxiv.org/abs/2204.02235
  - (new version under review)
- **Performance study of future processors w/10~20x cores & 10~20x memory BW as 3D-SRAM**
- **Various benchmarks, Riken Fibre, ECP, SPEC, etc.**
- **~10x speedup possible over A64FX**



arXiv:2204.02235v1 [cs.DC] 5 Apr 2022

At the Locus of Performance: A Case Study in Enhancing CPUs with Copious 3D-Stacked Cache

Jens Domke[*,§], Emil Vatai[*,§], Balazs Gerofi[*], Yuetsu Kodama[*], Mohamed Wahib[*,†], Artur Podobas[‡], Sparsh Mittal[**], Miquel Pericàs[¶], Lingqi Zhang[††], Peng Chen[†], Aleksandr Drozd[*], and Satoshi Matsuoka[*,††]

[*] RIKEN Center for Computational Science, Japan
[†] National Institute of Advanced Industrial Science and Technology, Japan
[‡] KTH Royal Institute of Technology, Sweden
[**] Indian Institute Of Technology, Roorkee, India
[¶] Chalmers University of Technology, Sweden
[††] Tokyo Institute of Technology, Japan

Abstract—Over the last three decades, innovations in the memory subsystem were primarily targeted at overcoming the data movement bottleneck. In this paper, we focus on a specific market trend in memory technology: 3D-stacked memory and caches. We investigate the impact of extending the on-chip memory capabilities in future HPC-focused processors, particularly by 3D-stacked SRAM. First, we propose a method oblivious to the memory subsystem to gauge the upper-bound in performance improvements when data movement costs are eliminated. Then, using the gem5 simulator, we model two variants of LARC, a processor fabricated in 1.5 nm and enriched with high-capacity 3D-stacked cache. With a volume of experiments involving a board set of proxy-applications and benchmarks, we aim to reveal where HPC CPU performance could be circa 2028, and conclude an average boost of 9.77x for cache-sensitive HPC applications, on a per-chip basis. Additionally, we exhaustively document our methodological exploration to motivate HPC centers to drive their own technological agenda through enhanced co-design.

Index Terms—microarchitectural study, 3D-stacked memory, gem5 simulation, proxy-applications

Fig. 1. MiniFE: relative performance improvement of AMD EPYC 7773X Milan-X over AMD EPYC 7763 Milan, and Figure of Merit; Input problem scaled from 100×100×100 to 400×400×400; Both systems equipped with dual-socket CPUs; Benchmark run with 16 MPI ranks and 8 OpenMP threads

I. Introduction

Historically, the reliable performance increase of von Neumann-based general-purpose processors (CPUs) was driven by two technological trends. The first, observed by Gordon E. Moore [1], is that the number of transistors in an integrated circuit doubles roughly every two years. The second, called

diversity of architectures, such as quantum-, neuromorphic-, or reconfigurable computing [11]. Many of these prototypes hold promise but are still immature, focus on a niche use case, or incur long development cycles. At the same time, there is one salient post-Moore architecture that is growing in maturity and which can facilitate performance improvements in the decades to come even for the classic von Neumann CPUs we have

From A64FX to hypothetical LARC Processor w/ 3D SRAM

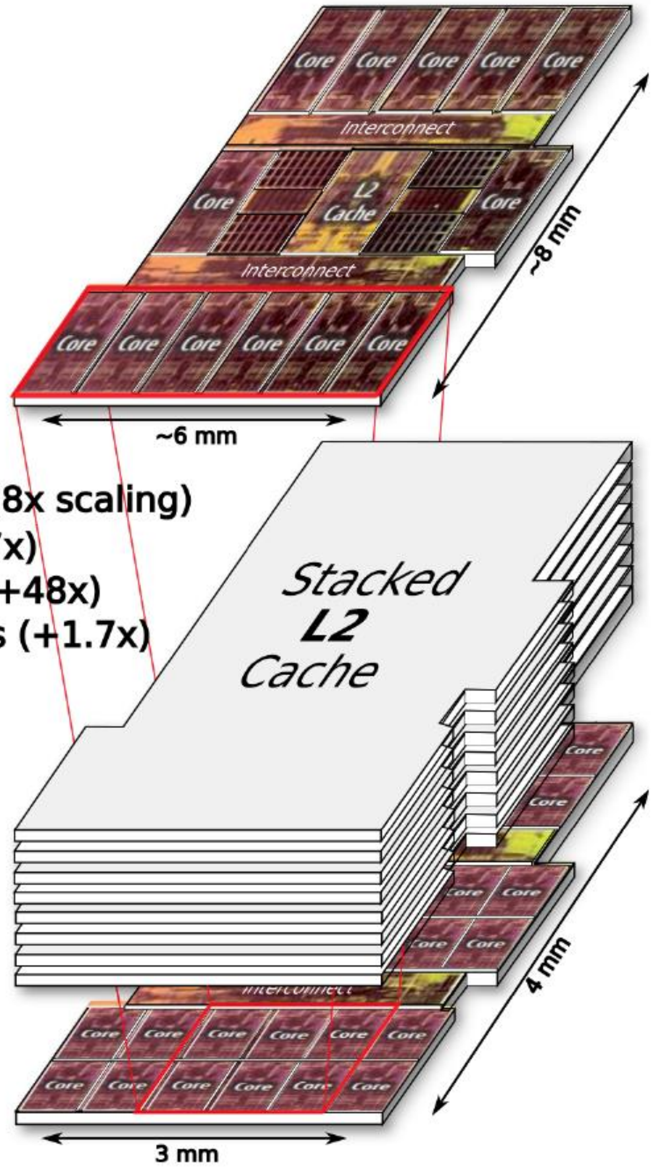
- New **LARC CMG** in 2028 timeframe
  - 32 A64FX-like cores w/ 64 KiB L1i and 64 KiB L1d, total of ≈2.3 Tflop/s
  - 384 MiB L2 with eight SRAM layers
  - (keep HBM2 to isolate perf. gains)
- New/**hypothetical LARC** CPU
  - die size similar to A64FX
  - 512 processing cores and **6 GiB of stacked L2 cache** with peak L2 bandwidth **of 24.6 TB/s**
  - peak HBM2 bandwidth of 4.1 TB/s
  - total **≈36 Tflop/s** in IEEE-754 FP64

A64FX CMG @7nm

CMG Area:	48 mm²
# Cores:	12
L2 Cache:	8 MiB
L2 B/W:	900 GB/s
HBM B/W:	256 GB/s

LARC CMG @1.5nm

CMG Area:	12 mm² (8x scaling)
# Cores:	32 (+2.67x)
L2 Cache:	384 MiB (+48x)
L2 B/W:	1536 GB/s (+1.7x)
# Dies:	8+1
# TCI Chan./Die:	384
# TCI Channels:	3072
TCI Channel Cap.:	128 KiB
HBM B/W:	256 GB/s



A64FX vs. LARC Core Memory Group Layout Comparison

Next Steps in the Feasibility Study Project

- **Selecting architecture/system candidates for a next-generation system**
  - Accelerator, memory technology, photonics technology, and packaging
  - Consider effective accelerator architecture based on quantitative benchmarking analyses
  - Optimizing balance or fusion between HPC and AI performance

- **Creating R&D roadmap for system software**
  - Being strongly conscious of software ecosystem
  - Optimized workflow execution specially for HPC and AI cooperation

- **Application first system design**
  - Design a system target for science breakthrough NOT just for ranking such as Top500
  - Building benchmark framework for fair architectural comparison
  - Blushing up future science roadmap including roadmap on “AI for Science”

- **Collaborating operation technique and new computing-paradigm teams**
  - Data framework, realtimeness, carbon neutrality, ･･･
  - Extending computable areas by HPC-Quantum hybrid platforms