富岳nextに向けた数値計算ライブラリ調査状況について

--Feasibility Study report about numerical libraries towards FugakuNEXT --

Toshiyuki Imamura
RIKEN R-CCS
理化学研究所 計算科学研究センター
今村俊幸

26, Dec. 2024

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

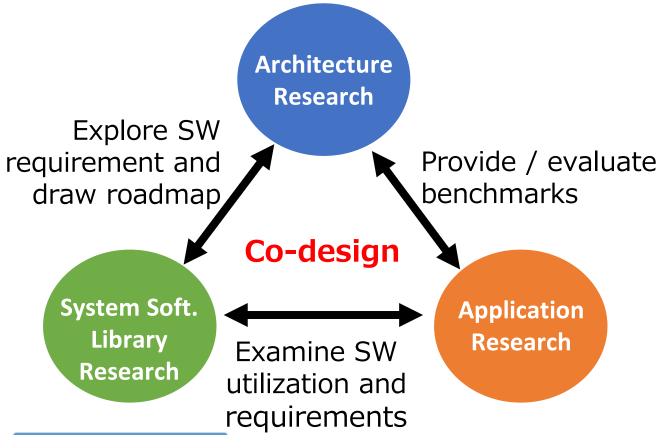
- Investigating technological possibilities (such as 3D stacked mem, accelerators, chip-to-chip direct optical link) and performance of the entire system or its components based on trends in semiconductor and packaging technologies
- Predicting future system performance based on performance analysis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

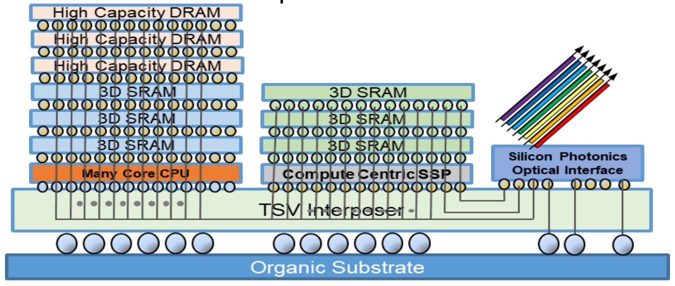
Research on Applications

- 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 – 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

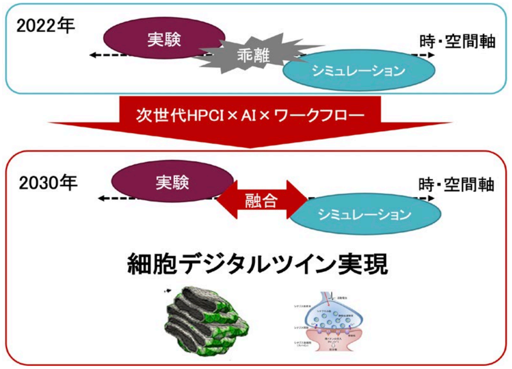
List of Benchmark Applications in RIKEN Team

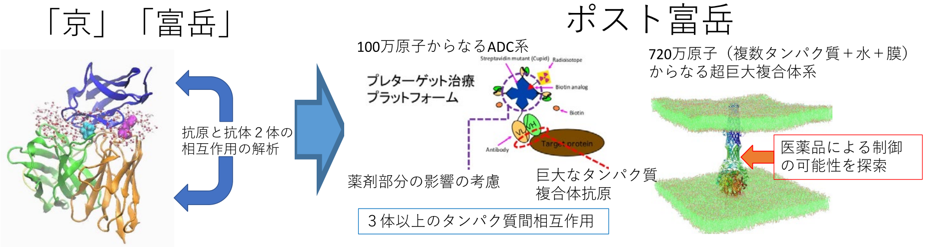
- **Initial application list for estimating performance of future architectures**
  - More benchmark applications will be evaluated at a later stage

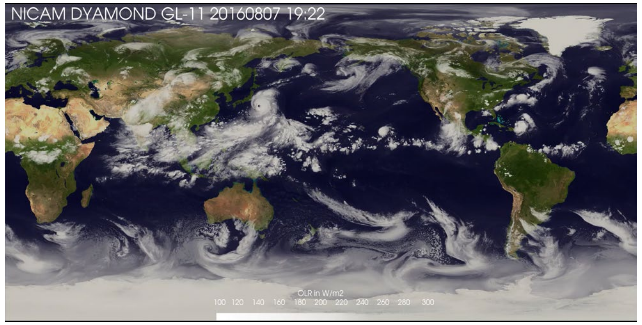
Area	Application	Type	Language	GPU	Note
Life Science	GENESIS	MD (particle)	Fortran	working	strong-scalability oriented Mixed precision
New Material & Energy	SALMON	DFT, Stencil, FFT	Fortran	✓	high-precision GEMM required Possible Emulation w/ME
Weather and Climate	SCALE-LETKF	CFD (structured mesh)	Fortran	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

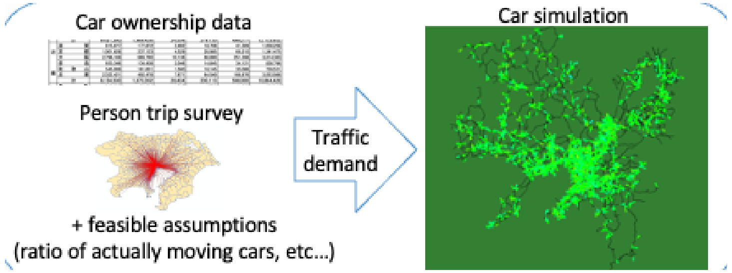
Roadmap of Target Sciences in FugakuNEXT Era

- **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.







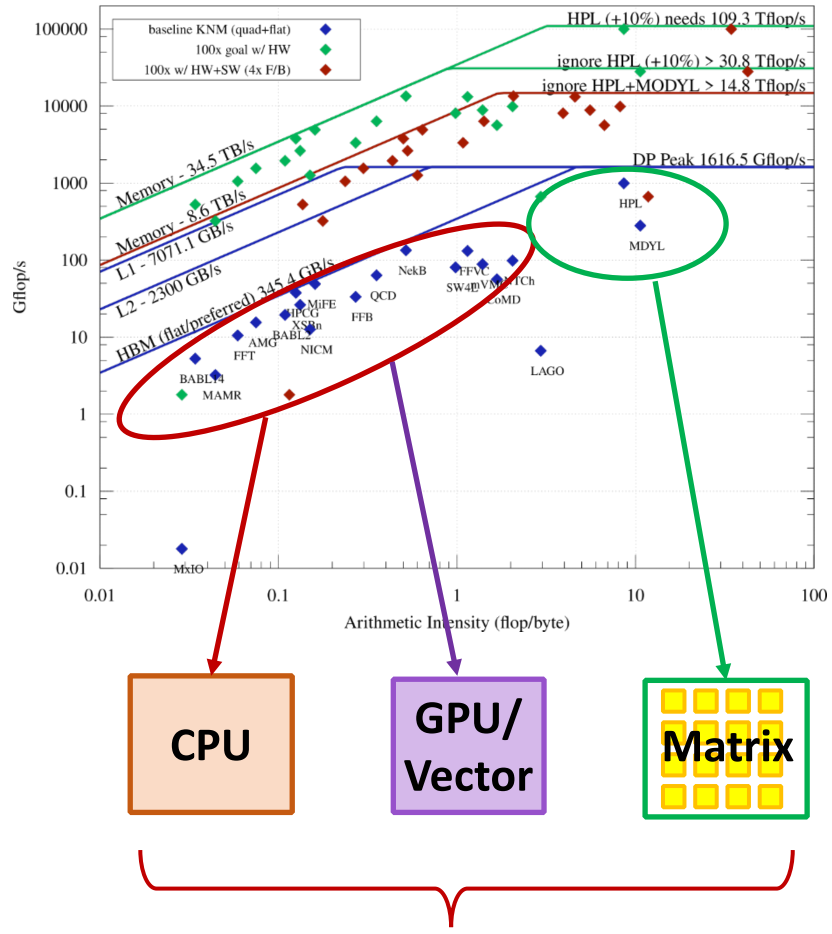


Key Research Item for Node Architecture Selection

- **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

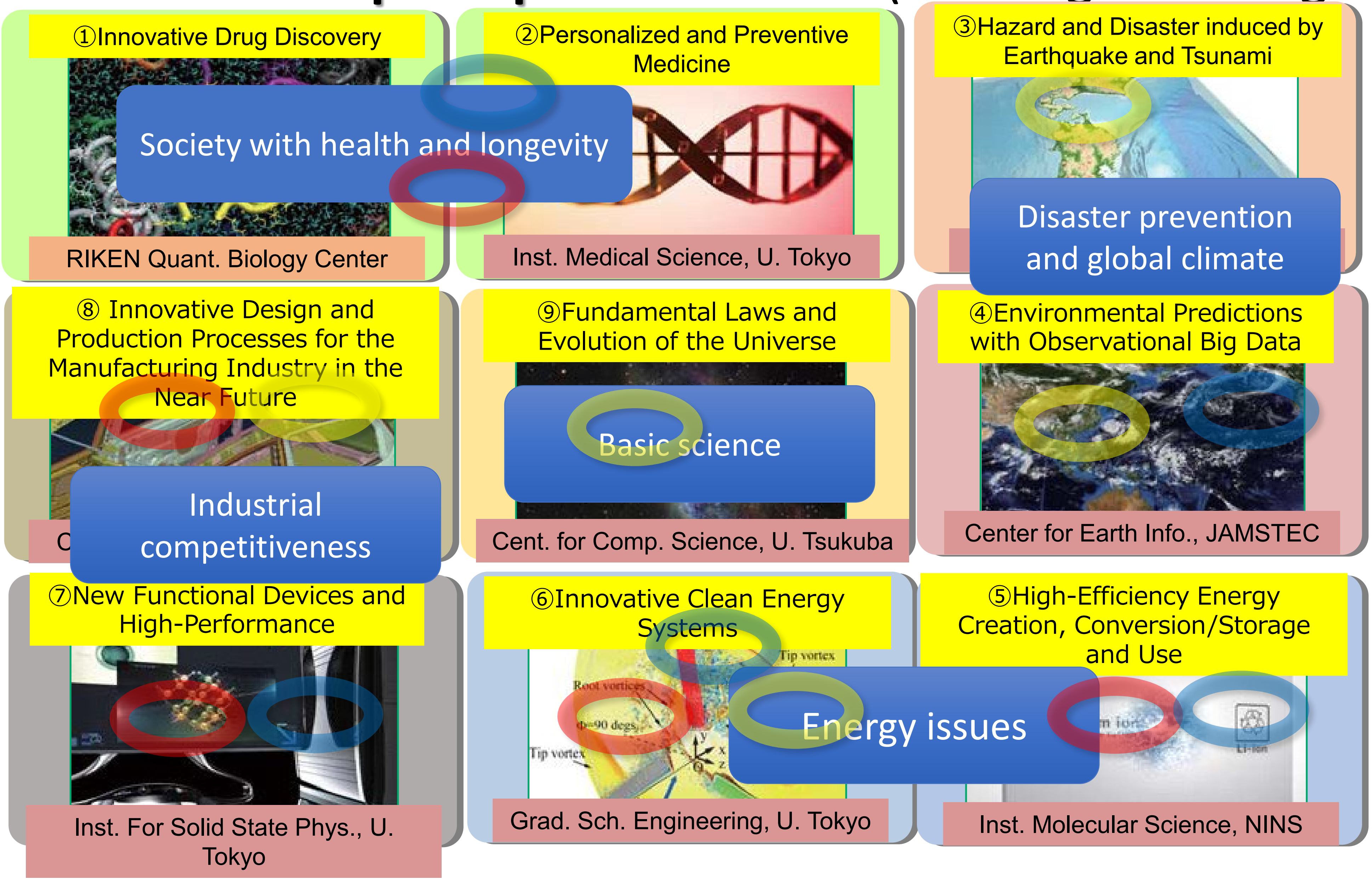
Roofline analysis on A64FX



Need to find the optimal balance

- 利用者アンケート
- 既存ソフトウェアサーベイ
- 想定されるシステムでの機能調査
- 重要性の高いソフトウェア、国産ソフトウェアの調査
- 提言としてまとめる

The mathematical **principle** is invariant (K →Fugaku→FugakuNext)



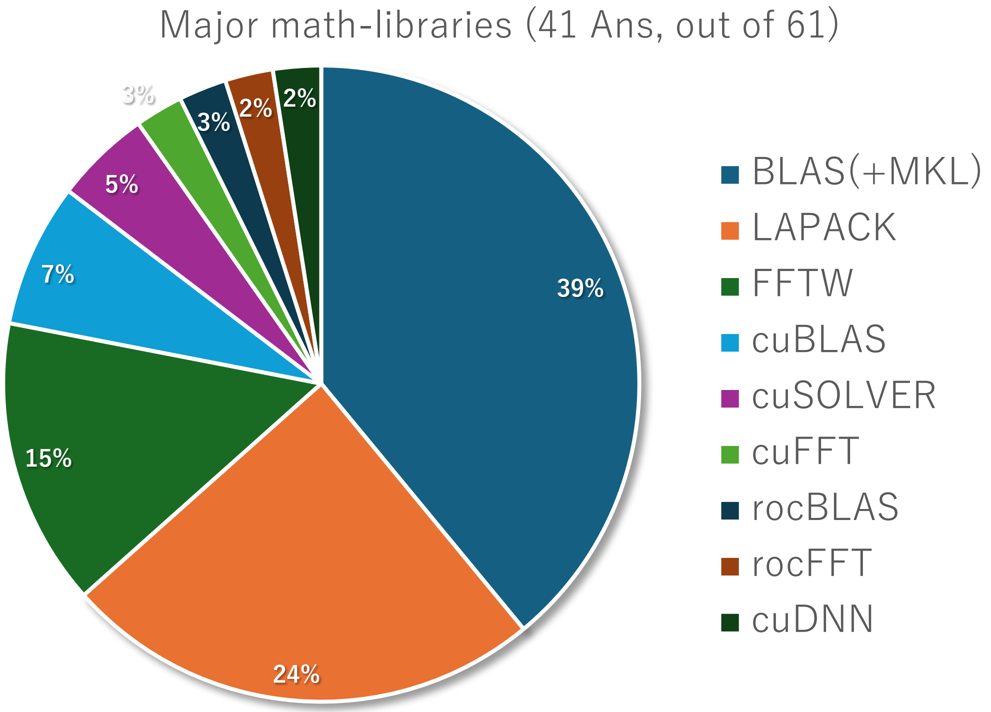
FFT

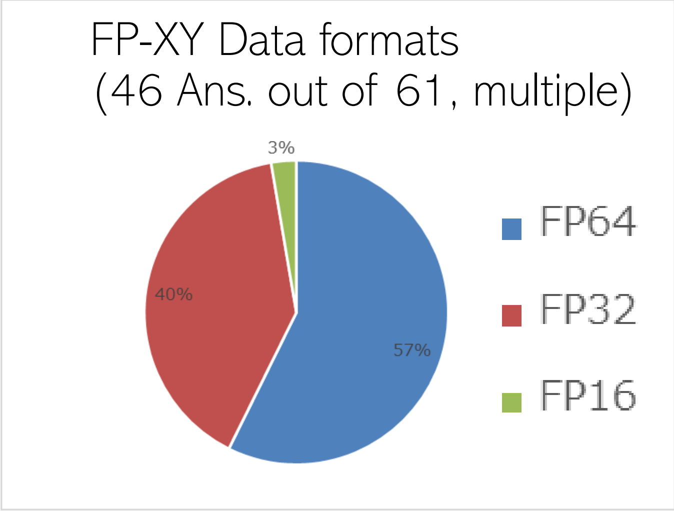
EVD&SVD

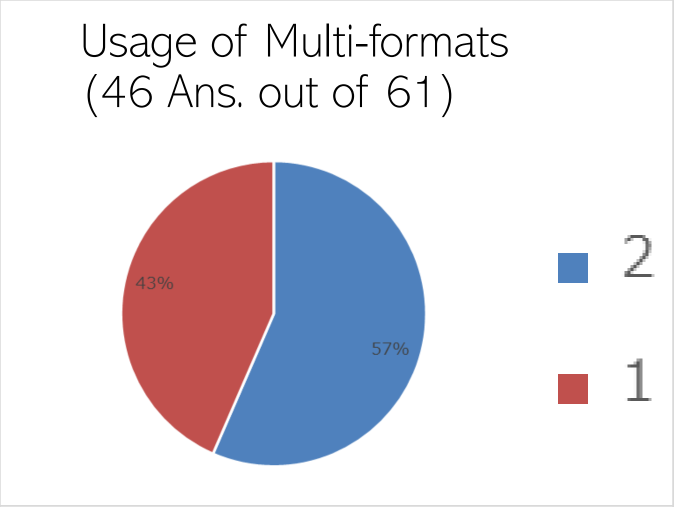
LinSys

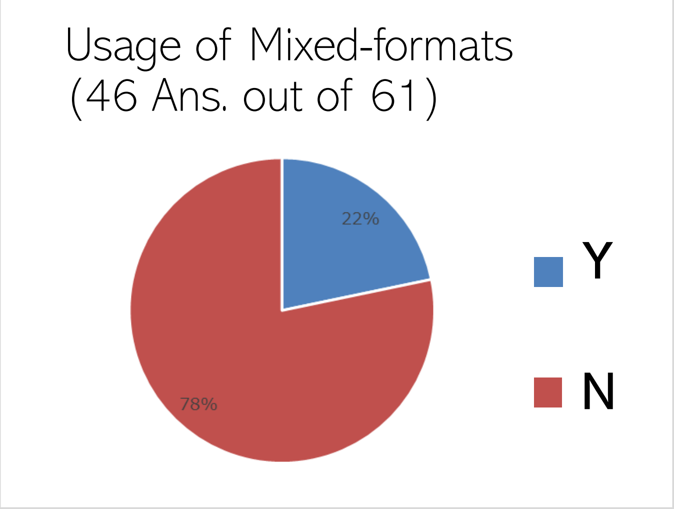
Sustainability for next-Fugaku

Which math libs are the most required?





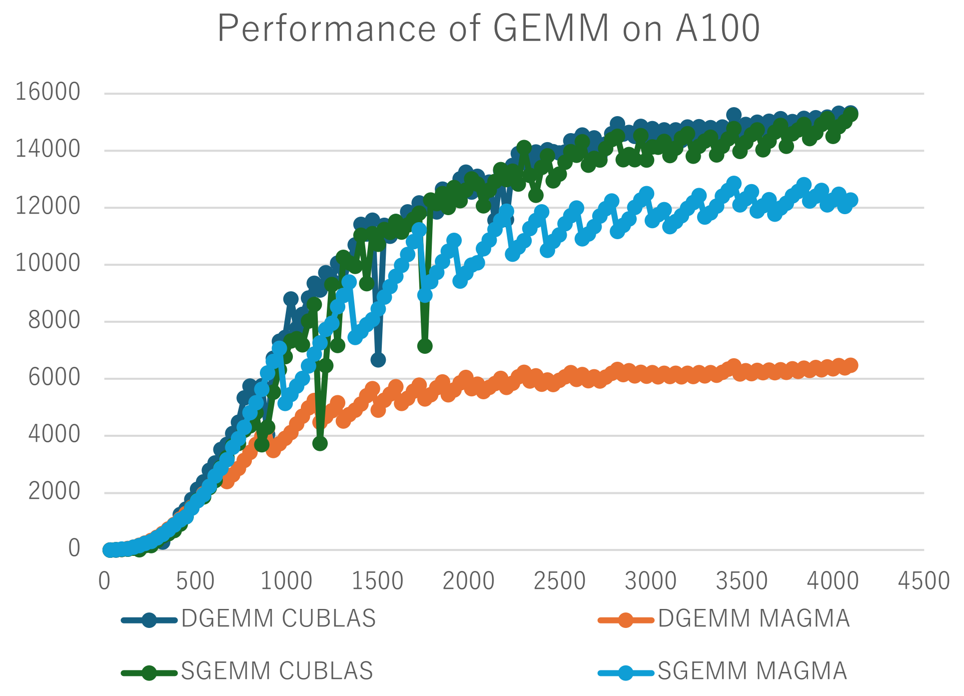


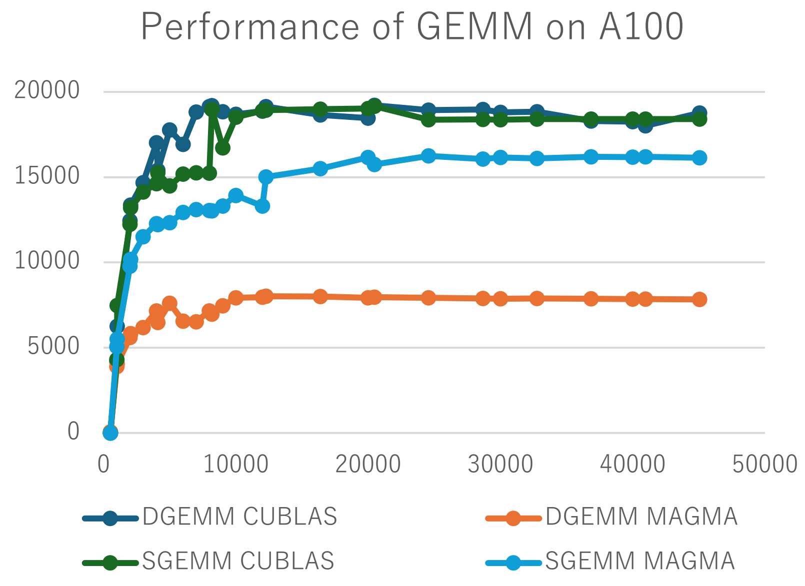


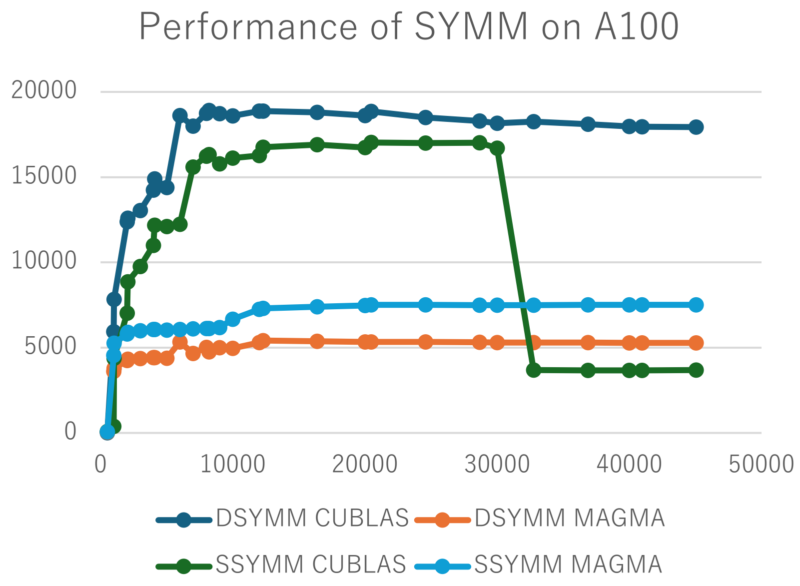
- Almost 100% of required kernels only focus on **BLAS, LAPACK, FFTW**, or their **variant optimized for target architectures.**
  - 1/3 of users said they need **in-house coded routines** and **no public libraries.**
  - Uncertain of the significance of the distributed parallel.
- User-level Multi or Mixed-precision arithmetic is ready-employed
  - **Math libraries must support multi-mixed-precision APIs in future systems!**

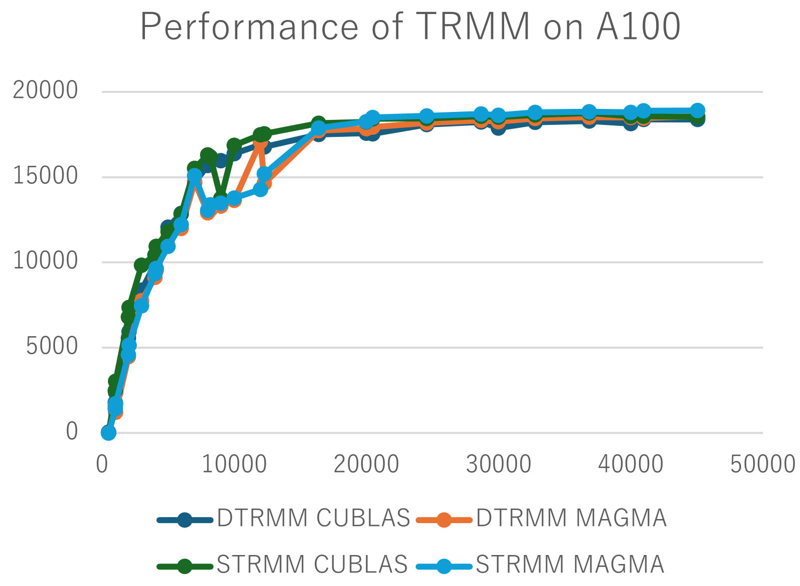
NVIDIA A100 80GB-SXM

- FP64: 9.7T(core)/19.5T(TC)
- FP32: 19.5T(core)
- TF32: 156T, FP16, BF16: 312T
- INT8: 624T
- GPU mem: HBM2e 2039GB/s
- CUDA 12.5, MAGMA-2.8.0



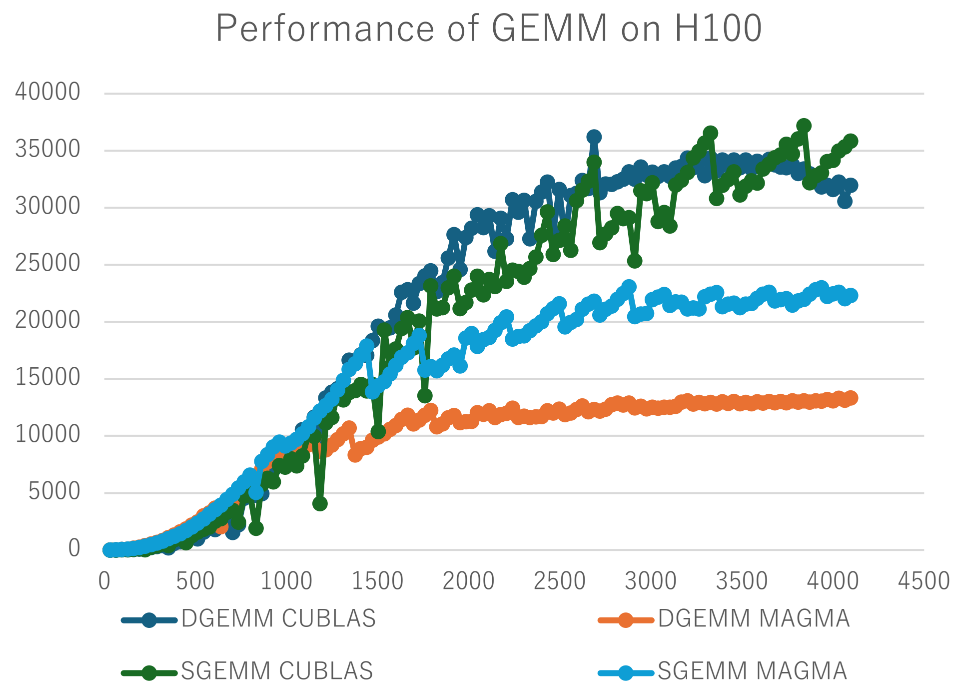


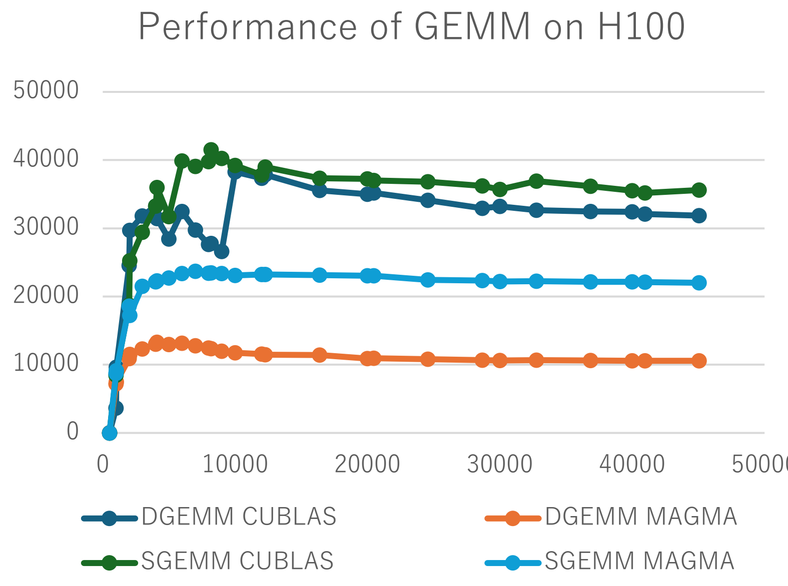


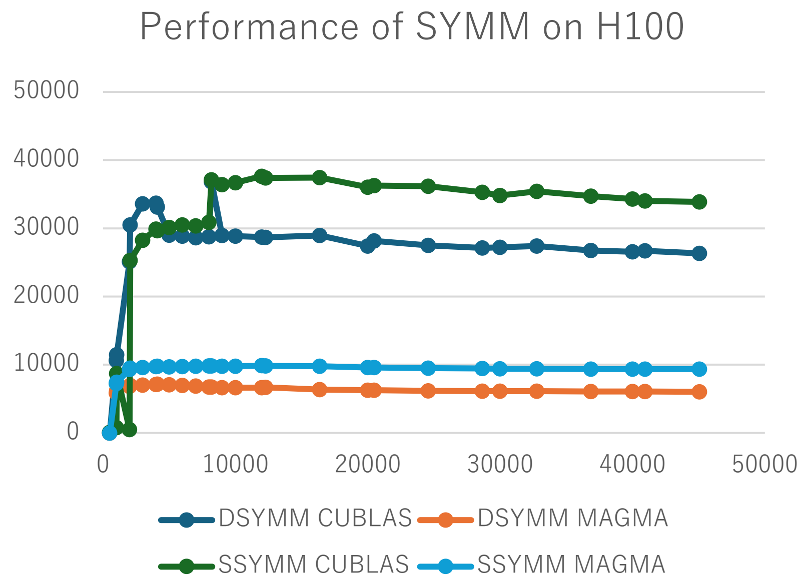


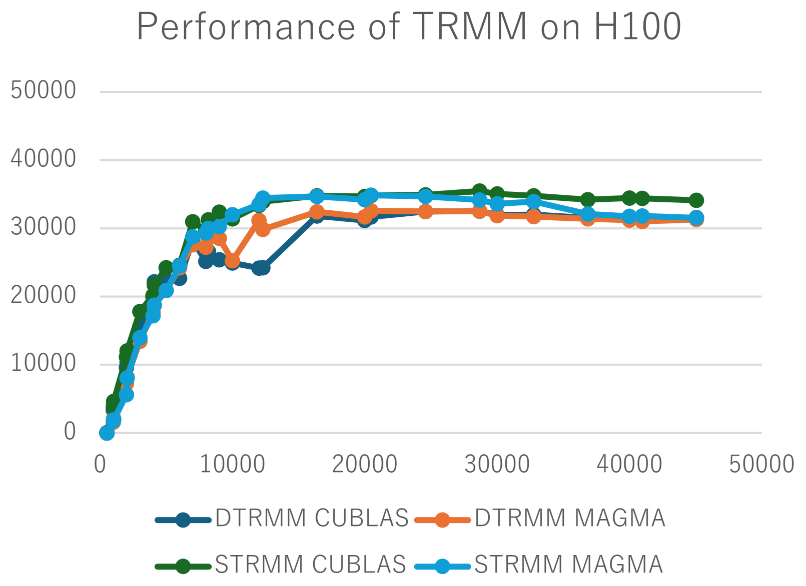
NVIDIA H100 80GB-PCIe

- FP64: 24T(core)/48T(TC)
- FP32: 48T(core)
- TF32: 400T, FP16, BF16: 800T
- INT8: 1600T
- GPU mem: HBM2e 2000GB/s
- CUDA 12.5, MAGMA-2.8.0



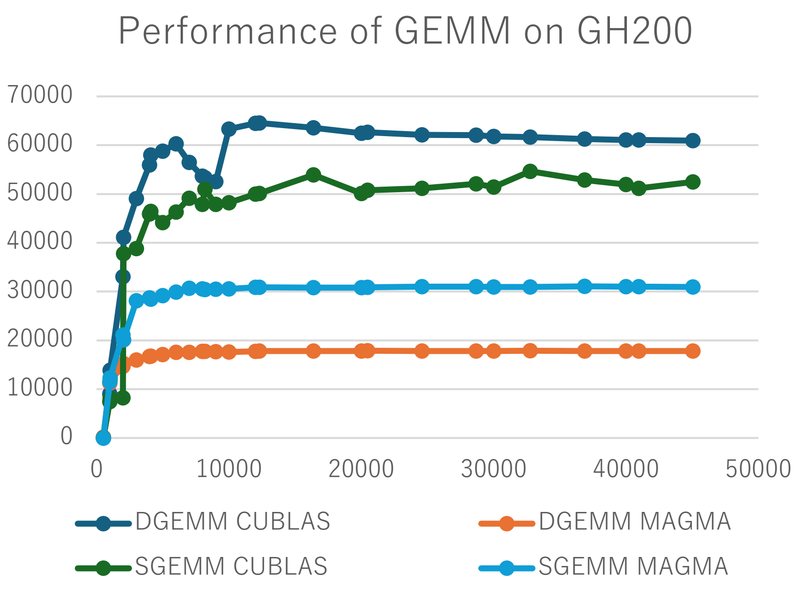


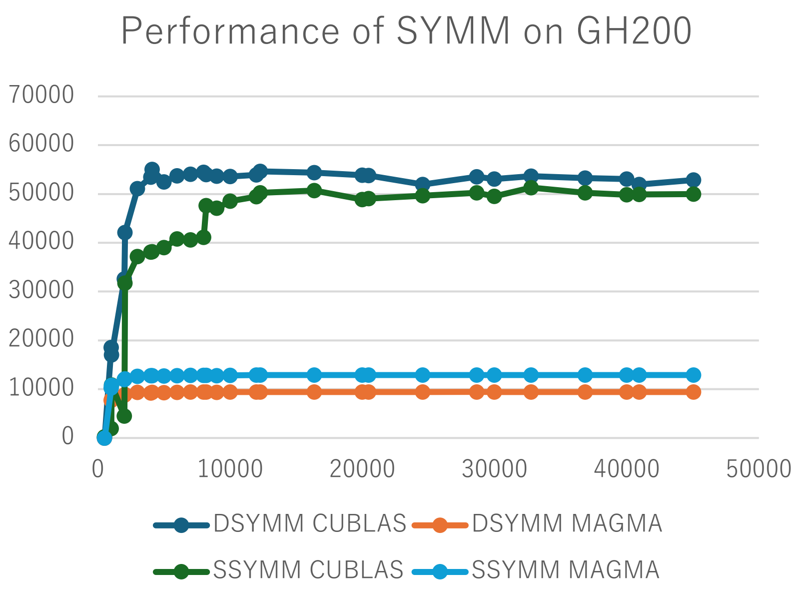


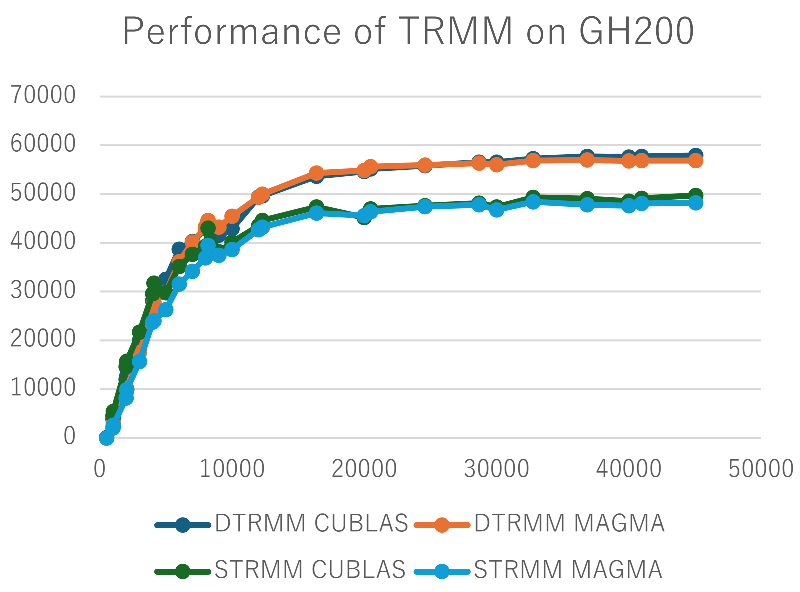


NVIDIA GH200(GPU part)

- FP64: 34T(core)/67T(TC)
- FP32: 67T(core)
- TF32: 494T, FP16, BF16: 990T
- INT8: 1979T
- GPU mem: HBM3 96GB, 4000GB/s
- CUDA 12.5, MAGMA-2.8.0



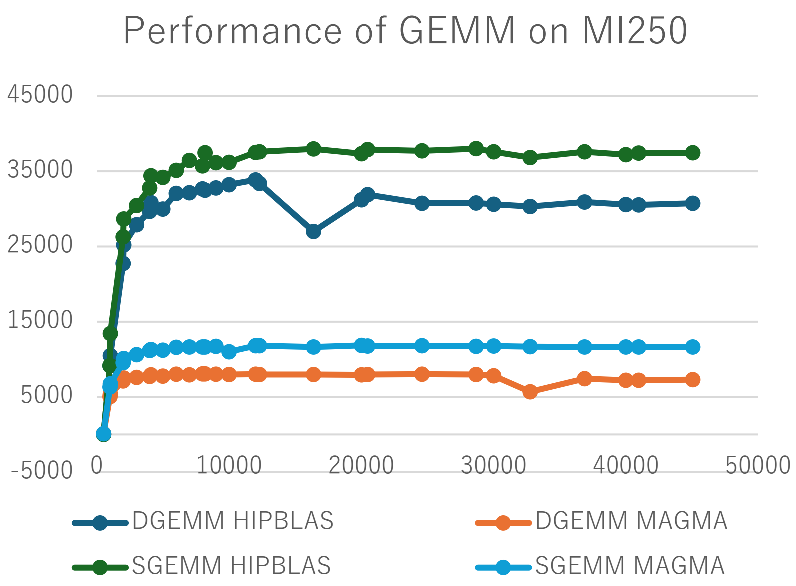


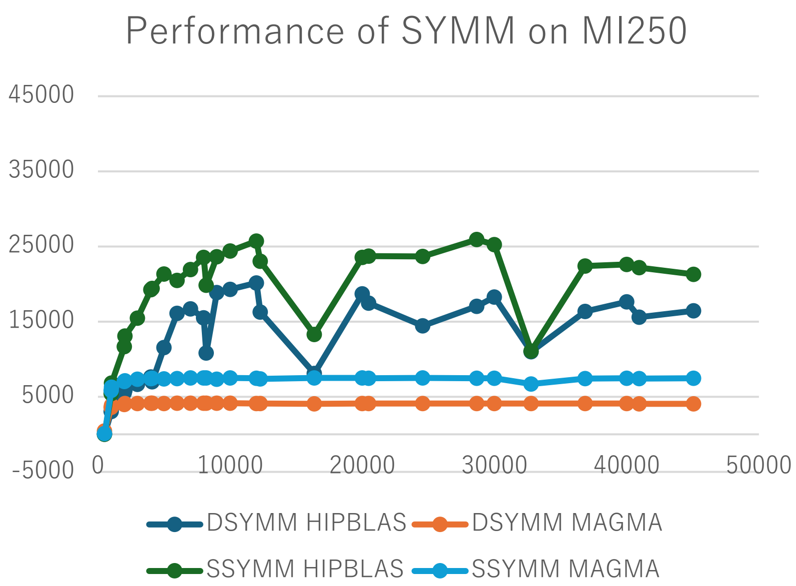


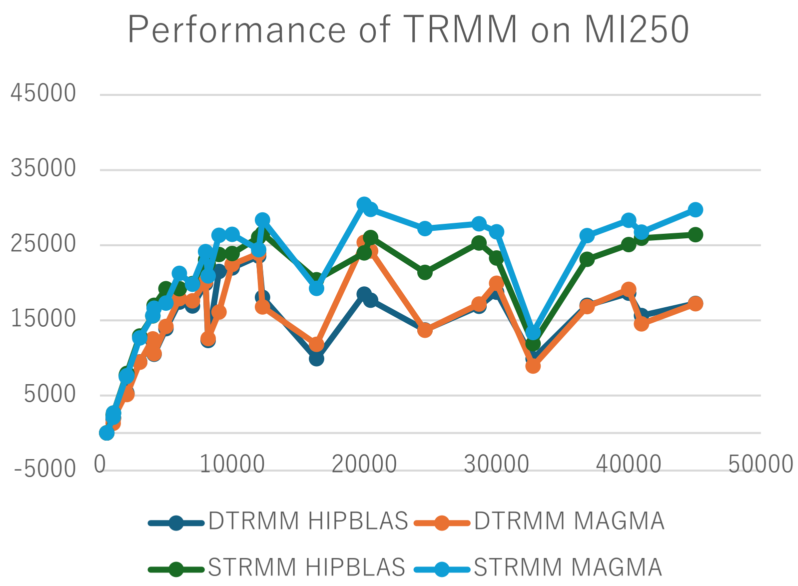
AMD MI250

- FP64: 45.3T(core)/90.5T(Matrix)
- FP32: 45.3T(core)/90.5T(Matrix)
- FP16, BF16: 362.1T(core)
- INT8: 362.1T
- GPU mem: HBM2e 128GB, 3200GB/s
- Rocm 6.2(BLASは2GCDうち1GCD動作), MAGMA-2.8.0

hipBLASは
RocBLASを呼ぶ構造
ROCm/rocBLAS: Next generation BLAS implementation for ROCm platform (github.com)
GEMMカーネル情報はyamlで記述







Preliminary Performance summary on BLAS Level3 kernels

- NVIDIA cuBLAS
  - cuBLAS自身の性能は極めて良好
  - MAGMAは独自ソースで性能は1/2 (Tensor機能は利用していない?), cuBLASが呼ばれているものもおそらくある
  - 90%近い性能（GH200でSGEMMが遅い傾向）
  - cuBLAS-EX版が低精度版ベンチマークで必要(int8)
  - 1万次元以下の性能詳細
    - 性能低下がみられる
    - 1~4000次元では大きな劣化ではない。
    - ~10000次元での詳細を再測定中

- AMD Rocm, hipBLAS
  - GEMMは70~80%程度、更なる高性能化が望まれる(NNはよいがTNが遅い傾向もみられる)
  - 2GCDを使う拡張も必要
  - GEMM以外は、動作不安定（2ベキ次元なのでキャッシュの問題？）
  - SYMMは1/2程度に劣化
  - OSSであるのでソースコードを調査(読んで)してみるが, GEMMのパラメターはYAMLで管理
  - hipBLAS/rcoBLASの上位LAPACK版は、最適化が不足している傾向…

Sustainability for next-Fugaku

Required Items of Math libraries

- **Numerical Lib1 : BLAS**
  - A standard library of numerical linear algebra that handles basic operations on vectors and sequences
  - It is a top priority for AI (DNN) performance as well as for scientific and technological simulations.
  - Due to the abstraction on the functional level, there are many variants in terms of language, execution environment, target architecture, implementation method, etc.
- **Numerical Lib2 : Dense matrix solvers**
  - **LAPACK**: Numerical library for realizing such as a sys of linear eqs., eigenvalue calculations, singular value calculations, etc.
    - Numerical library for generalized numerical computation, with many variants and implementations on shared memory, GPUs, etc., in addition to single CPUs
  - **SLATE, DPLASMA**: for distributed parallel computing in ECP as a successor to ScaLAPACK distributed parallel version of LAPACK. We need to port and optimize under international cooperation
    - SLATE incorporates the capability of abstract program description and the flexible distributed data structure of C++.
    - DPLASMA is a library that promotes high speed in the direction of strongly promoting task parallel execution

Sustainability for next-Fugaku

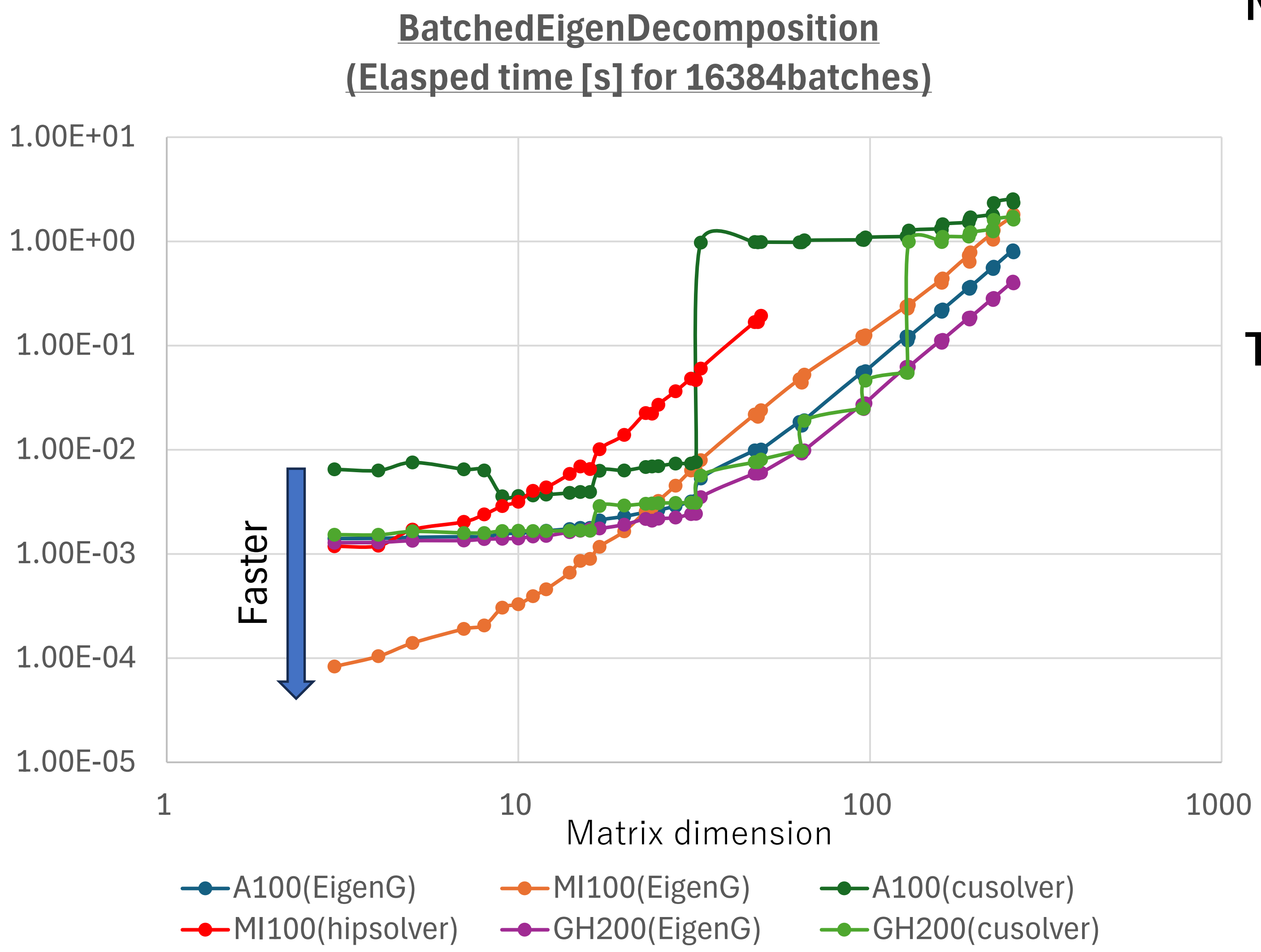
Required Items

- **Numerical Lib3 : EigenSolver EigeExa(Dense-parallel eigensolver)**
  - One of the **Japanese products**. RIKEN has been developed on the continuous projects of ‘K’, ‘Fugaku’, and next-Fugaku. Dedicated not only to supercomputers but also to general CPUs, extracting capabilities on batched and GPU-spined-off versions. Strengthened on homogeneous environments such as Fugaku
- **Numerical Lib4 : Sparse linear/eigen solvers**
  - Includes direct method routines, eigenvalue-specific sparse iterative method libraries, as well as general iterative method support software for sparse matrices
  - **PETSc/SLEPc**: Distributed Parallel PDE, **MUMPS/SuperLU/Dissection**, **ARPACK-NG/FEAST**
- **Numerical Lib5：FFT**
  - **FFTX/SPIRAL**, the successor to the industry-standard FFTW, a numerical library for high-speed Fourier transforms, and **FFTE, a domestic product, also a SPIRAL kernel, have been imported**

- **Other than the above, the following are medium priorities. However, the SWG reminds us that higher-priority software must be developed within the international community**
  - [Preprocessing tools] (**Hypre/GAMG**)
  - [Ordering tools] Tools for converting node information and the shape of the sequence into an appropriate format (**METIS/SCOTCH**)
  - [Pseudo-random number generators] (very-long-period, parallelizable generators are extremely important)

Performance issue (updated on 3rd of May 2024)

- Batched FP32-EVD (MI100 vs A100(12.2) vs GH200(12.4))



Note:

Test matrix has a condensed-spectrum form such as $\mathrm{Norml}(Q\mathrm{diag}\{1+\epsilon_i\}_i Q^\top)$

hipsolver breaks down numerically when N>49.

Theoretical Peak Performance:

The **A100** offers the double-precision FP64 performance up to 9.7 TFLOPS. The single-precision FP32 performance is **19.5 TFLOPS (wo TensorCores)**.

The **GH200** offers up to 34 TFLOPS and **67 TFLOPS** of peak FP64 and FP32, respectively (wo TensorCores).

The **MI100** offers up to 11.5 TFLOPS of peak FP64 performance for HPC and up to **23.1 TFLOPS** peak FP32 performance (wo AI Matrix engine).

Other topics for next-gen xPUs

Required Items

- **Energy and parallelism**
  - Hierarchical hardware mapping
  - Advanced PARALLEL programming model and languages
  - Borderless memory models, such as unified memory
- **Next-generation Numerical Linear algebra software?**
  - cuSolverMP on NVIDIA GPU cluster
  - AMD (which model? Hip+MPI? Or kokkos+HPX?)
- **Low-precision arithmetic**
  - GEMMFP64 emulation by int8 based on the Ozaki-scheme: Ootomo-Ozaki-Yokota, and Uchino-Ozaki-Imamura are good examples (50TFLOPS on a GH200, expected to >90TFLOPS overperforming FP64 matrix-TCs on GB200)
  - Other possibility? Like multi-component floating point format and arithmetic by Ozaki-Imamura (PA, QTW, QQW, and mX_real)
- **In the Japanese SC community,**
  - H-matrix, Approximation/Randomized algorithm
  - FPGA
  - LLML and AI
  - Proxy model toward Quantum system?

We welcome you to the 7th R-CCS International Symposium on January 23-24, 2025, Kobe, Japan!

R-CCS is the world's top-level research center for HPC and the core research center in Japan upholding "science of computing, by computing, and for computing." With

INTERNATIONAL HPC SUMMER SCHOOL 2025

Details about the 2025 International HPC Summer School

Graduate students and postdoctoral scholars from institutions in Australia, Canada, Europe[1], Japan and the United States are invited to apply for the 15th International High Performance Computing (HPC) Summer School, to be held July 6-11, 2025 in Lisbon, Portugal. **The deadline for application is 23:59 AOE (Anywhere On Earth), January 31, 2025.**