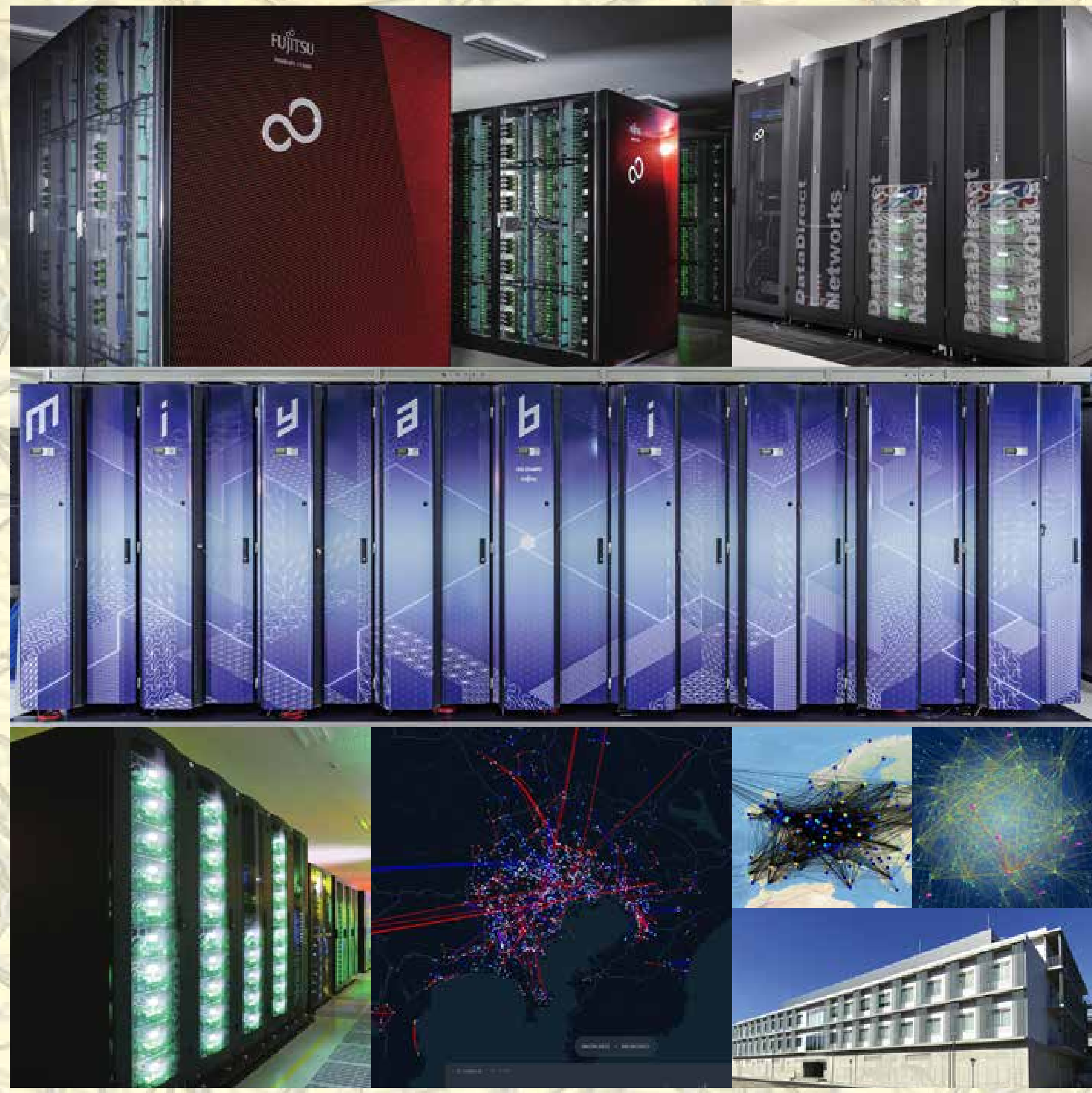
東京大学
THE UNIVERSITY OF TOKYO

Supercomputing & Data Science

Information Technology Center, The University of Tokyo



Welcome

ITC, The University of Tokyo

The Information Technology Center (ITC) (https://www.itc.u-tokyo.ac.jp/en/) was organized in 1999, and it consists of 5 research divisions (Campus-wide Computing, Data Science, Network, Supercomputing and Inter-disciplinary Information Science). ITC is also a core organization of the "Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures (JHPCN)", and a part of HPCI (the High-Performance Computing Infrastructure) operated by the Japanese Government.

SCD/ITC, The University of Tokyo, Japan

The Supercomputing Research Division, Information Technology Center, The University of Tokyo (http://www.cc.u-tokyo.ac.jp/) was originally established as the Supercomputing Center of the University of Tokyo in 1965, making it the oldest academic supercomputer center in Japan, and it became the Supercomputing Research Division (SCD) of ITC after 1999. The three main missions of SCD/ITC are (i) Operations of Supercomputers & Services, (ii) Research & Development, and (iii) Education & Training. Currently, SCD/ITC consists of more than 10 faculty members. SCD/ITC is now operating the "Integrated Supercomputer System for Simulation, Data and Learning (Wisteria/BDEC-01)" by Fujitsu with 33.1 PFLOPS. The system operated by SCD/ITC contain 3,000+ users; 55+% of these users are from outside the university. Hands-on tutorials for parallel programming (on-line) are held 20+ times per year. Up to 10% of the total computational resources of each system are open to users from the industry.

Data Science Research Division

Nowadays, a huge volume of digital data is gathered from natural as well as artificial sources, including weather and seismic monitoring data, human and vehicle mobility data, and social activity data including business transactions, medical care, and so on. Digitally archiving historic documents and records at risk of dissipation also produces digital data. In these cases, digitization secures and enhances the value of knowledge by making it accessible regardless of physical distance. There is also a lot of new digital data being created, including web pages, social networks, academic papers, and so on. Advances in data analysis and modeling techniques, most notably machine learning, allow us to extract more meaningful and interpretable information from data and networking technology. This makes it possible to combine information from various sources. Data science is about turning raw data from a stream of digits into valuable insights and knowledge. Data science is also closely related to advances in high performance computing technologies, including high performance processors, storage and networking, big data analytics, deep learning numerical algorithms, and so on. The Data Science Research Division was established at the end of 2018, replacing the Academic Information Science Research Division. It takes a leadership role not only for the research on data science but also for designing and building a national infrastructure for the data science research community. It has been leading the development of the "mdx" platform, a national-wide academic cloud computing service in partnership with 10 universities and research institutes.

Computational Science Alliance, the University of Tokyo

At the University of Tokyo, we established the Computational Science Alliance (http://www.compsci-alliance.jp/) in 2015 by collaborating with 14 departments, including ITC. The primary purpose of this alliance is to provide an interdisciplinary education program for High-Performance Computing (HPC). The alliance started lectures in April 2017.

Computational Science Alliance
The University of Tokyo

JHPCN: Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures

"JHPCN (https://jhpcn-kyoten.itc.u-tokyo.ac.jp/en/)" comprises academic supercomputer centers in Japan associated with eight national universities (Hokkaido, Tohoku, Tokyo, Tokyo Tech, Nagoya, Kyoto, Osaka, and Kyushu). This began in April 2010. The total performance of the supercomputer systems involved is 185+ PFLOPS (April 2024). JHPCN promotes collaborative research projects using the facilities and human resources of these eight centers, including the supercomputers, storage systems, and networks; interdisciplinary projects using multiple facilities are particularly encouraged. 40 or more projects have been accepted each year. New frameworks for international and industry collaborations have been initiated since 2017. Moreover, a new category focusing on data science/data analytics has been introduced in 2022, in addition to computational science. We have 63 projects (55: Computational Science & Engineering(CSE) , 8: Data Science/Data Analytics (DATA)) in 2022, 68 (51:CSE, 17:DATA) in 2023, and 77 (57:CSE, 20:DATA) in 2024.

JHPCN

JCAHPC: Joint Center for Advanced High Performance Computing

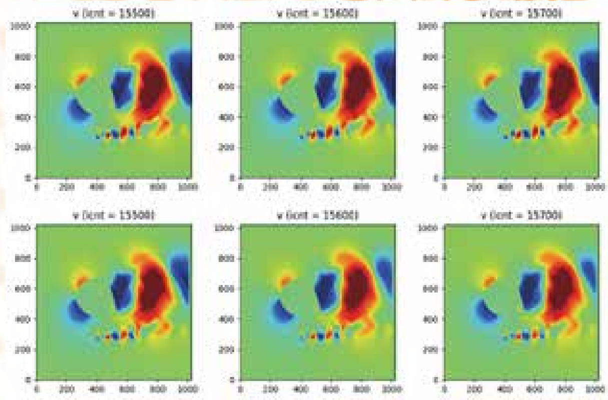
In 2013, Center for Computational Sciences, University of Tsukuba (CCS) and ITC agreed to establish the Joint Center for Advanced High-Performance Computing (JCAHPC). JCAHPC consists of more than 20 faculty and staff members of CCS and ITC. Originally, the primary mission of JCAHPC is designing, installing, and operating the Oakforest-PACS system (OFP). In addition, CCS and ITC develop system software, numerical libraries, and large-scale applications for OFP in collaboration. JCAHPC is a new model for collaboration for research and development between multiple supercomputer centers. Oakforest-PACS (OFP) has contributed significantly to the development of computational science in Japan and around the world, especially after shutdown of the K computer in August 2019. Moreover, three proposals were adopted for the "HPCI Urgent Call for Fighting against COVID-19" in 2020. OFP retired on March 31, 2022. JCAHPC is introducing Miyabi (OFP-II) system with 80+ PFLOPS, the successor of OFP, which starts operation in January 2025. Miyabi has two groups of nodes: Miyabi-G with 1,120 nodes of NVIDIA GH200, and Miyabi-C with 190 nodes of Intel Xeon CPU Max 9480.

JCAHPC

Scientific Computing & Numerical Algorithms

Prediction of CFD Simulations by Deep Learning

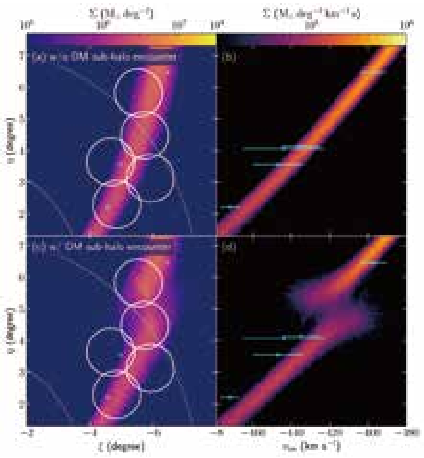
Computational fluid dynamics (CFD) is widely used in science and engineering. However, since CFD simulations requires a large number of grid points and particles for these calculations, these kinds of simulations demand a large amount of computational resources such as supercomputers.
Recently, deep learning has attracted attention as a surrogate method for obtaining calculation results by CFD simulation approximately at high speed. We are working on a project to develop a parallelization method to make it possible to apply the surrogate method based on the deep learning to large scale geometry. Unlike the model parallel computing, the method we are currently developing predicts large-scale steady flow simulation results by dividing the input geometry into multiple parts and applying a single small neural network to each part in parallel. This method is developed based on considering the characteristics of CFD simulation and the consistency of the boundary condition of each divided subdomain. By using the physical values on the adjacent subdomains as boundary conditions, applying deep learning to each subdomain can predict simulation results consistently in the entire computational domain. For time evolution simulations, we are developing a model that predicts from three frames at 100-step intervals to three frames at the following 100-step intervals. Patch-based CNN inference is being developed to enable predictions for computational domains with arbitrary sizes.



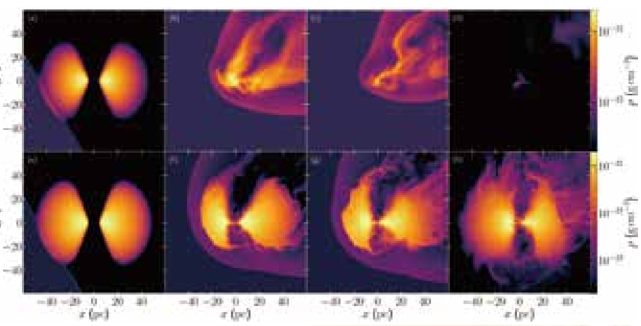
Comparison of the flow velocity results obtained by the conventional simulation (upper) and the prediction of these results by deep learning (lower)

GPU-accelerated *N*-body code and application to galactic archaeology

Collisionless *N*-body simulations are frequently employed to explore the formation and evolution of galaxies. We have developed a gravitational *N*-body code optimized for GPU: GOTHIC (Gravitational Oct-Tree code accelerated by HIerarchical time step Controlling). GOTHIC includes both the tree method and the hierarchical time step. The code runs entirely on GPU and is optimized for from the Fermi to the NVIDIA Hopper GPU architectures. NVIDIA H100 (PCIe) achieves a 1.4-fold acceleration compared to NVIDIA A100. The observed speed-up is consistent with the ratio of the theoretical peak performance of integer operations.
Stellar halos of galaxies provide fossil records of the formation and evolution of galaxies through galactic mergers. The Andromeda galaxy (M31) is an attractive laboratory for galactic archaeology due to its proximity and external perspective. *N*-body simulations using GOTHIC have reproduced stellar structures observed in the M31's halo. Stellar streams are promising probes to detect interactions between the streams and invisible dark matter (DM) sub-halos. The further synergy of *N*-body simulations on GPU-accelerated supercomputers and dedicated observations using the Subaru telescope will unveil the nature and assembly history of the DM halo.

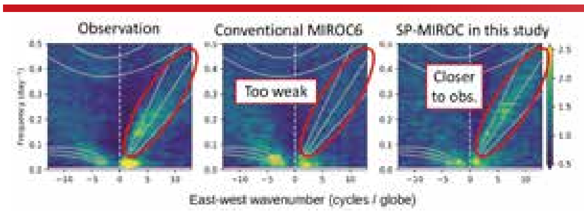


Imprints of past DM sub-halo collision to stellar stream

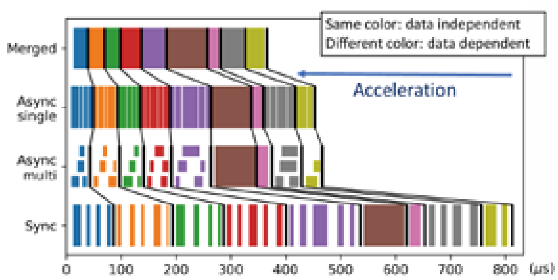


Time evolution of torus-shaped gas surrounding the massive black hole during a head-on galaxy collision

Adapting an overhead-sensitive Atmospheric Simulator to GPUs



SP-MIROC better reproduces spatiotemporal spectra of tropical cloud abundance (OLR) compared to the conventional MIROC6



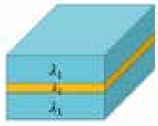
Time series of GPU kernel execution in various launch configurations(Kernels in the core component only)

In atmospheric simulations, it is important to appropriately handle small-scale processes, such as clouds, even if one is only interested in simulating large-scale phenomena like El Nino. Super-parameterization (SP) is one of methods to accomplish this. In a SP model, a large number of small 2D domains of a high-resolution atmospheric model run in parallel, calculating small-scale influences for a single low-resolution global model. In this study, SP-MIROC, an SP version of a global climate model MIROC6 coupled with a high-resolution model SCALE-RM, was ported to NVIDIA GPUs primarily using OpenACC.
GPU offloading was carried out only for the resource-intensive high-resolution component (SCALE-RM). Atmospheric codes are consists of a large number of small kernels. As a result, most of the simulation time is spent for an enormous number of GPU kernels running only around 10 µs, which exposes kernel launch overheads.
To hide overheads of the numerous kernel launches, most of the GPU kernels were launched asynchronously with respect to the host CPU. Kernels were launched in a single GPU queue to simplify code management. To push the performance further, some data-independent kernels were merged to a single kernel, switching operations depending on CUDA block indices. Because OpenACC does not support such task parallism inside a single kernel, the kernel merging was implemented in CUDA. Combining these optimizations, the performance of the core fluid dynamics solver doubled.

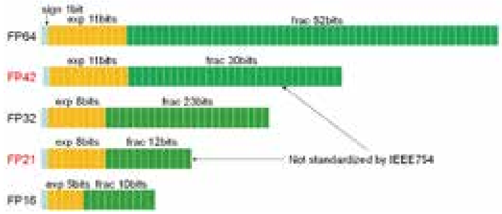
System, Tools & Hardware

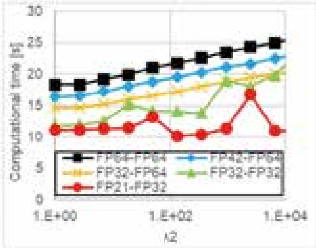
Key to Green Computing: Adaptive Precision Computing with FP42/FP21

Double precision computing by FP64 has been traditionally used for scientific computing. Recently single precision (FP32) is widely used in certain class of problems for efficiency. Moreover, half precision (FP16) is also available on A64FX architecture used in the Fugaku and Wisteria/BDEC-01 (Odyssey). While FP16 is originally used for machine learning computation, it is not suitable for scientific computing due to instability. We developed *adaptive precision*, such as FP21 (between FP16 and FP32) and FP42 (FP32 and FP64). More robust and efficient scientific computing is expected by utilizing adaptive/mixed precision computing. In the present work, mixed precision computing was applied to ICCG solver for large-scale linear equations with sparse coefficient matrices derived from finite-volume applications for 3D steady-state heat conduction problems in heterogeneous media. Mixed precision computing by FP21-FP32 (Preconditioning:FP21, Others such as SpMV:FP32) provides 40.5% faster than FP64-FP64 for homogeneous cases ($\lambda_1=\lambda_2=1.0$) on a single node of Odyssey with A64FX. While careful validation for accuracy is essential, it was found that computation time could be reduced dramatically by adopting low/adaptive/mixed precision according to the conditions of target problems.



Finite Volume Model for 3D Steady-State Heat Conduction in Heterogeneous Media





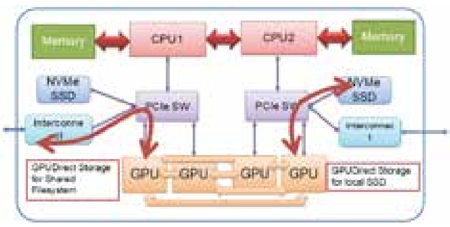
Performance is based on the total comp. time of ICCG solver in each case of mixed precision computing

Preconditioning	Others (SpMV etc.)	Performance
FP64	FP64	1.000
FP21	FP32	1.405
FP32	FP32	1.353
FP32	FP64	1.199
FP42	FP64	1.105

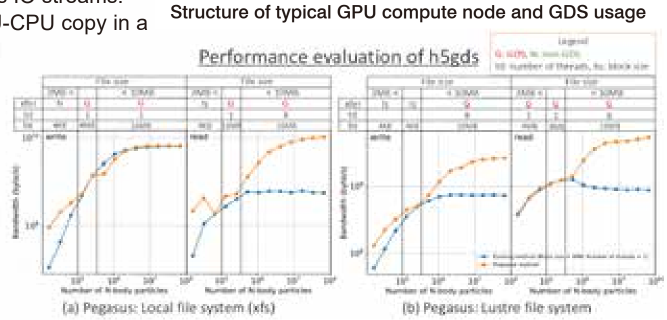
Relationship between comp. time for ICCG solver and λ_2 for various combinations of mixed prec. computing

Direct File IO at GPU using GPUDirect Storage

In the last decade, GPU clusters have been widely used to achieve the highest performance with limited power. However, in order to utilize GPUs in practical applications, since it is necessary to input/output data handled in independent memory for GPU, either the data transfer time itself needs to be reduced or the data transfer must be hidden by overlapping with the computation. "GPUDirect Storage (GDS)", an extension of "GPUDirect for RDMA" to realize RDMA transfer of GPU memory such as MPI, is now available for file IO. In fact, GDS cannot always achieve the best performance, and according to the characteristic of read/write by GPUDirect, the preference between direct transfer manner by GDS and pipeline transfer via CPU memory copy should properly differ depending on such as the data size and the number of simultaneous IO streams.
In this study, we aim to establish a sophisticated method to switch GDS and GPU-CPU copy in a pipeline manner properly. To realize efficient file access from/to GPU in practical applications, we apply such strategies to the HDF5 plugin for GDS "vfd-gds" based on observing various file IO characteristics. We also implement a file IO-aware benchmark suitable for GPU clusters. As a result, the performance improvement is 6.09x in maximum compared with the original implementation. The results will be helpful for the users on "Miyabi (OFP-II)" of the Joint Center for Advanced HPC (JCAHPC) in collaboration with the Center for Computational Sciences, Univ. of Tsukuba.



Structure of typical GPU compute node and GDS usage

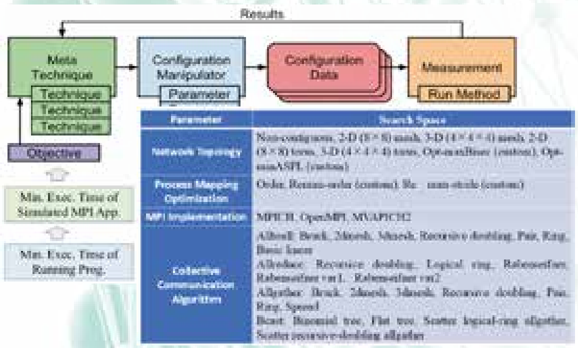


Comparison of file IO performance

This work is supported by "Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (JHPCN)" and "High Performance Computing Infrastructure (HPCI)" in Japan, and supported (in part) by Multidisciplinary Cooperative Research Program in CCS, University of Tsukuba.

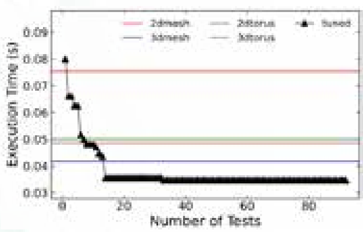
Autotuning for Communication Optimization in Parallel Computing Systems

In communication-intensive parallel applications, a significant portion of execution time is spent exchanging numerical data between processes. While enhancing processor performance and network bandwidth is crucial, reducing communication overhead (an inherent side effect of parallelization) requires significant attention. In this context, our objective is to develop a performance tuning function that alleviates the burden on developers of high-performance parallel programs. We proposed an autotuning method to find the optimal environment from a large search space for a target parallel application. This automation can explore larger optimization spaces more efficiently than manual tuning. This approach significantly optimized program execution time within a reasonable time interval by automating the search for optimal implementations and configurations.
In this work, we leveraged OpenTuner to define a search space of potential implementations and optimizations, which are specific assignments of parameters such as network topologies, mapping algorithms, MPI implementations, collective communication algorithms, and other settings. The evaluation on SimGrid (v3.26) demonstrated that MPI applications achieve accelerated and near-optimal execution performance, with speedups of up to 2.29x after only a few search trials.

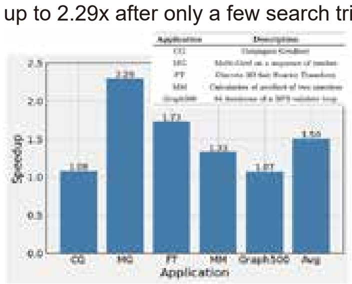


Parameter	Search Space
Network Topology	Non-contiguous, 2-D (8 × 8) mesh, 3-D (4 × 4 × 4) mesh, 2-D (8 × 8) torus, 3-D (4 × 4 × 4) torus, OptimalBisec (custom), OptimASPL (custom)
Process Mapping Optimization	Order, Random-order (custom), Re[illegible]num-stride (custom)
MPI Implementation	MPICH, OpenMPI, MVAPICH2
Collective Communication Algorithm	Alltoall: Bruck, 2dmesh, 3dmesh, Recursive doubling, Pair, Ring, Basic linear Allreduce: Recursive doubling, Logical ring, Rabenseifner, Rabenseifner-rsr1, Rabenseifner-rsr2 Allgather: Bruck, 2dmesh, 3dmesh, Recursive doubling, Pair, Ring, Spread Bcast: Binomial tree, Flat tree, Scatter logical-ring allgather, Scatter recursive-doubling allgather

Configuration Parameters and Autotuning Objectives in the MPI Application Autotuning Framework



Tuning of MG App.

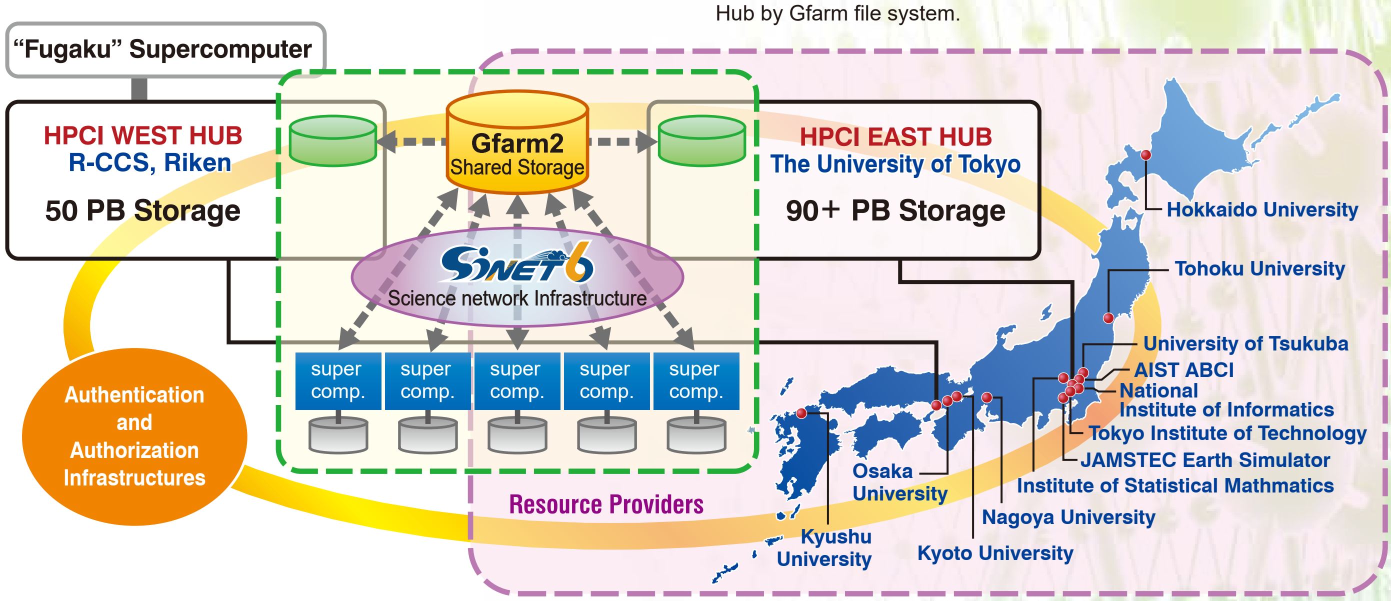


Speedup of tuned MPI Apps

Supercomputers at SCD/ITC

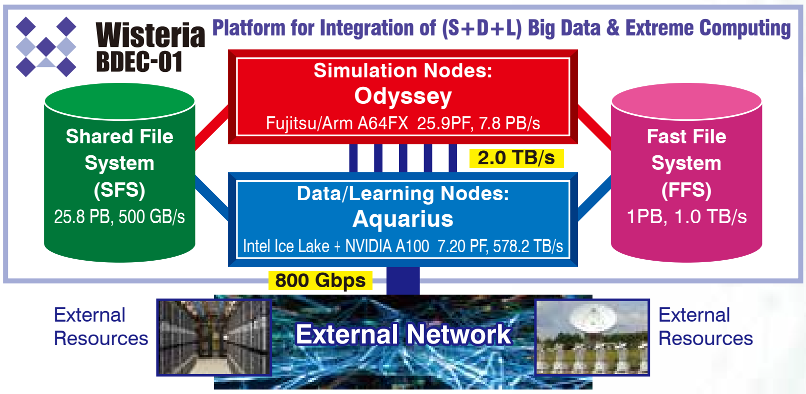
HPCI: High Performance Computing Infrastructure

High performance computing infrastructure (HPCI) is an environment that enables easy usage of flagship “Fugaku” supercomputer and other computation resources (tier-2) in Japan. In addition, HPCI is expected to match a user’s needs and computational resources to accelerate exploratory research, large-scale research, and industrial use of HPC. HPCI comprises 13 computational resource providers; nine are supercomputing centers at national universities, and four are governmental research institutes. These resource suppliers are connected via SINET6, which is a high-speed academic backbone network with 400 Gbps. SCD/ITC participates in this project as a hub resource provider in the Kanto region (the HPCI EAST Hub). The HPCI EAST Hub provides a 90+ PB storage system integrated with the WEST Hub by Gfarm file system.



Wisteria/BDEC-01 (Fujitsu PRIMEHPC FX1000 & Fujitsu PRIMERGY)

We started discussions on the BDEC system (Big Data & Extreme Computing) as a platform for integration of (Simulation (S) +Data (D) +Learning (L))(S+D+L) in 2015. Wisteria/BDEC-01, which started its operation in May 2021, is the first BDEC system. Wisteria/BDEC-01 is a Hierarchical, Hybrid, Heterogeneous (h3) system, and it consists of two types of node groups for computing, Simulation Nodes (Odyssey) and Data/Learning Nodes (Aquarius), Shared File System (25.8 PB) and Fast File System (1.0 PB). The total peak performance is 33.1 PFLOPS, and aggregated memory bandwidth is 8.38 PB/sec. Simulation nodes for HPC (Odyssey) with more than 25 Peta FLOPS is based on Fujitsu’s PRIMEHPC FX 1000 with A64FX with High Bandwidth Memory. This part has the same architecture as that of the Fugaku supercomputer. Data/Learning nodes (Aquarius) are a GPU cluster consisting of Intel Xeon Ice Lake and NVIDIA A100 Tensor Core GPUs, with 7.2 Peta FLOPS for Data Analytics, AI and Machine Learning workloads. Some of Data/Learning nodes are directly connected to external resources through SINET, Japan. Odyssey and Aquarius are mutually connected through InfiniBand-EDR network with 2 TB/sec.



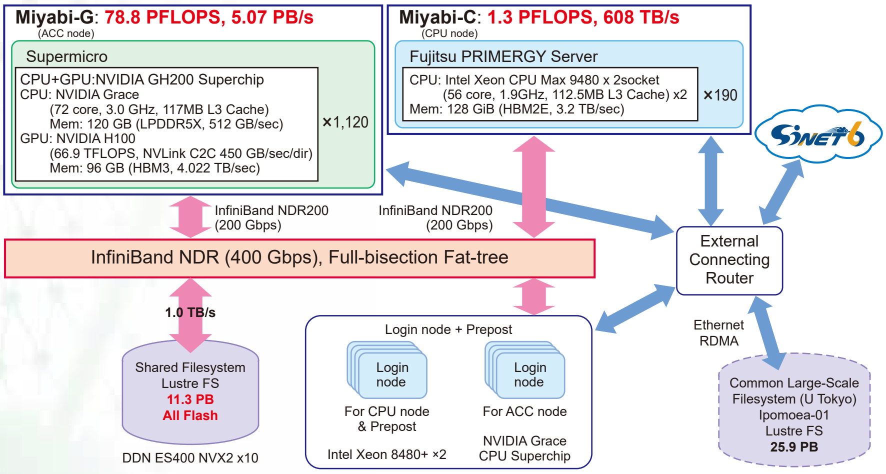
Wisteria/BDEC-01		Simulation Nodes: Odyssey	Data/Learning Nodes: Aquarius
Number of nodes		7680	45
Peak performance		25.9 PFlops	7.2 PFlops
Total memory Capacity		240.0 TiB	36.5 TiB
Network		Tofu-D: 6D-Mesh/Torus	InfiniBand HDR: Full-bisection Fat Tree
Compute node		FUJITSU Supercomputer PRIMEHPC FX1000	FUJITSU Server PRIMERGY GX2570 M6
CPU		Fujitsu/ARM A64FX, 48 cores, 2.2 GHz, 3.38TFLOPS, 32GB HBM2, 1,024 GB/sec	Intel Xeon Platinum 8360Y (Ice Lake), 36 cores x 2, 2.4GHz, 5.53TFLOPS, 512 GiB DDR4, 409.6 GB/sec
GPU		—	NVIDIA A100 GPUs, 108 SM's, 19.5 TFLOPS, 40GB Memory, 1.555 TB/sec, 8 GPUs on each node
Shared File System	Name	FEFS (Fujitsu Exabyte File System)	
	OSS	DDN SFA7990XE x 16	
	Capacity & Performance	25.8 PB, 504 GB/sec	
Fast File System	Name	FEFS (Fujitsu Exabyte File System)	
	OSS	DDN SFA400NVXE x 16	
	Capacity & Performance	1.0 PB, 1.0 TB/sec	

Supercomputer at JCAHPC & Common Storage System

Miyabi: New Supercomputer System of JCAHPC

Miyabi is the second supercomputer system introduced by JCAHPC (Joint Center for Advanced HPC), established by SCD/ITC and Center for Computational Sciences, U. Tsukuba (CCS), as a successor to Oakforest-PACS.
The system consists of 1,120 nodes for Accelerator node, each employing NVIDIA GH200 Grace Hopper Superchip (Arm Neoverse V2 core, cache-coherent with GPU), 190 nodes for General-purpose CPU node with two sockets of Intel Xeon CPU Max, 11.3 PB of DDN Lustre storage by all flash, and NVIDIA/Mellanox InfiniBand NDR200 as a high-performance interconnect. In addition, “Ipomoea-01” is also available as an external large-scale filesystem to be coupled to Miyabi.
We will start operating it in January 2025. This system is currently being installed at the Kashiwa campus of the University of Tokyo, at the exact location as the Oakforest-PACS system.
Inheriting the philosophy of the introduction of Oakforest-PACS, Miyabi aims to promote novel computational science methods leveraging AI, such as AI for HPC/Science, to offer a platform supporting Society 5.0 by integrating simulation, data analysis, and machine learning, in addition to support users of large-scale applications. Miyabi will be offered to researchers in Japan and their international collaborators through various programs including HPCI, MEXT’s Joint Usage/Research Centers (JHPCN), and each organization’s original programs.

			Miyabi-G (Accelerator node)	Miyabi-C (General-Purpose CPU node)
Entire system	Theoretical Peak Performance		78.8 PFLOPS	1.29 PFLOPS
Entire system	Number of Nodes		1,120	190
Entire system	Total Memory Capacity		241.9 TB	23.75 TiB
Entire system	Interconnect		InfiniBand NDR200 (200 Gbps) Full-bisection Fat Tree	
Entire system	Shared Filesystem		Lustre FS	
Entire system	# of inodes		appx. 23.5 B	
Entire system	Capacity		11.3 PB (All Flash)	
Entire system	Theoretical BW		1.0 TB/sec	
Node	Server		Supermicro ARS-111GL-DNHR-LCC	FUJITSU Server PRIMERGY CX2550 M7
Node	CPU	Processor	NVIDIA Grace	Intel Xeon CPU Max 9480 (Sapphire Rapids)
Node	CPU	# of CPUs (Core)	1 (72)	2 (56+56)
Node	CPU	Frequency	3.0 GHz	1.9 GHz
Node	CPU	Theoretical Peak Perf.	3.45 TFLOPS	6.8 TFLOPS
Node	CPU	Memory capacity /node	120 GB	128 GiB
Node	CPU	Memory BW / node	512 GB/s	3.2 TB/s
Node	GPU	Processor	NVIDIA Hopper H100	N/A
Node	GPU	# of GPUs	1	
Node	GPU	Theoretical Peak Perf.	66.9 TFLOPS	
Node	GPU	Memory capacity	96 GB	
Node	GPU	Memory BW	4.02 TB/s	
Node	GPU	CPU-GPU conn.	NVLink C2C 450 GB/sec/dir Cache-coherent	
Node	SSD	NVMe SSD	1.92 TB, PCIe Gen4 x4	N/A



Ipomoea-01: Large-Scale Common Storage System

As the processing power of supercomputers has increased, the amount of data handled has also grown. SCD/ITC/UTokyo has traditionally installed storage attached to each supercomputer system, and storage for each system was independent. For this reason, the files had to be moved every time the system was replaced, which caused great inconvenience to the users. There was a strong need to introduce a common storage system that could be accessed from all systems at ITC/UTokyo. Since it is necessary to move files in 26 PB of shared storage when OFP ends its operation in March 2022, we decided to install a "Large-Scale Common Storage System (Ipomoea)" that can be accessed from each system in ITC/UTokyo including mdx.
A new storage system will be installed, and replaced about every three years. Ipomoea-01 (Total Capacity: 25.9PB, Data Transfer Rate: 125 GB/sec) by Fujitsu started its operation in January 2022. The transfer of files from OFP was completed at the end of May 2022, and it has been available to the public since June.

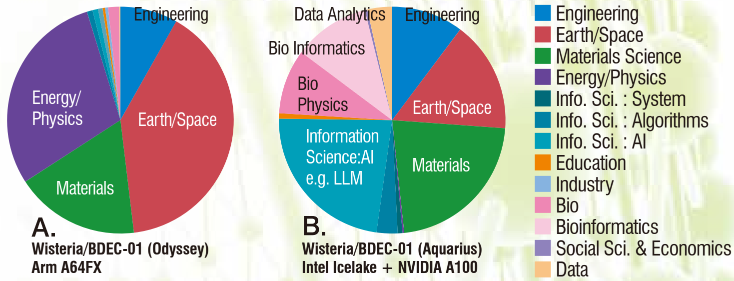
Operation	Fujitsu
Location	2F in Research Complex II,Kashiwa Campus
Filesystem	Lustre
Capacity	25.9 PB, 16.8 B i-nodes
Peak BW	125 GB/sec
Storage Servers	DDN ES7990X x5 set
Interconnect	100 GbE RoCE Ethernet

Ipomoea-01

Innovative Supercomputing by Integration of "Simulation/Data/Learning (S+D+L)" and Beyond

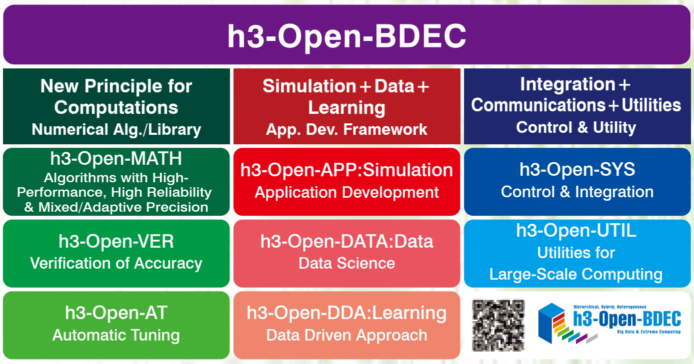
New Directions in Supercomputing

The majority of SCD/ITC supercomputer users are in Computational Science & Engineering (CSE), including engineering simulations, earth sciences, and materials science, as shown in Pie Chart A, which details usage on Wisteria/BDEC-01 (Odyssey) in FY 2023. Recently, there has been a rise in users from data science, machine learning, and AI, as depicted in Pie Chart B, showing usage on Aquarius with GPUs in FY 2023. Emerging research topics include real-time weather prediction, medical image recognition, and human genome analyses. We are pioneering a novel approach to scientific problem-solving by integrating "Simulation (S)", "Data (D)", and "Learning (L)" (S+D+L).

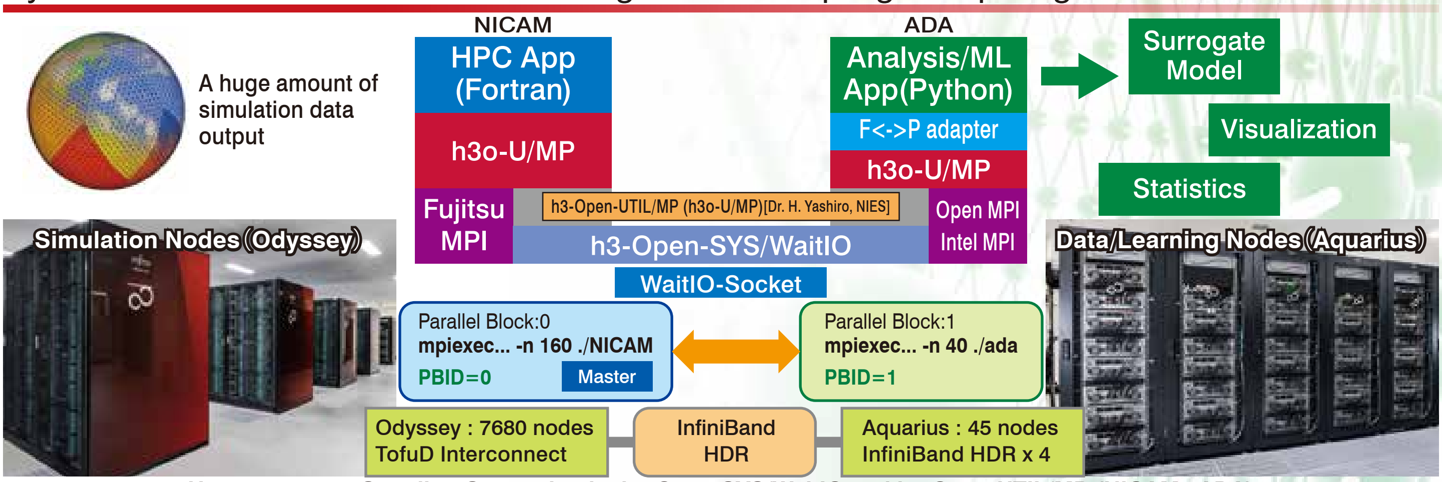


Wisteria/BDEC-01 & h3-Open-BDEC: Innovative Software Platform

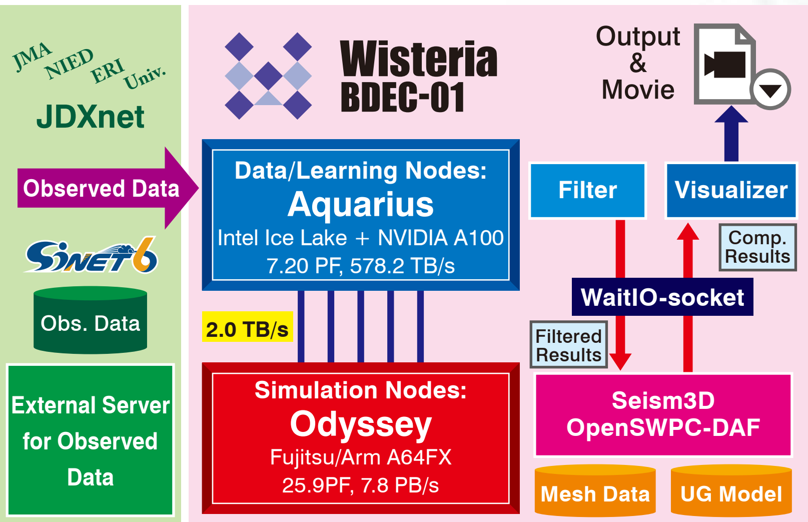
The integration of (S+D+L) on Wisteria/BDEC-01 involves optimizing parameters for large-scale simulations on Odyssey using data analyses, data assimilation, and machine learning on Aquarius. We have developed the innovative software platform "h3-Open-BDEC" for this purpose, evaluating its effects on Wisteria/BDEC-01. Designed to maximize supercomputer performance with minimal energy consumption, h3-Open-BDEC focuses on: (1) innovative numerical analysis methods with high performance, reliability, and power-saving through adaptive precision, accuracy verification, and automatic tuning; (2) a new data-driven approach for (S+D+L) integration; and (3) software and utilities for heterogeneous environments like Wisteria/BDEC-01. Supported by the Japanese Government (FY 2019-2023), the source codes and documents are publicly available. This integration significantly reduces computations and power consumption compared to conventional simulations.



h3-Open-SYS/WaitIO-Socket & h3-Open-UTIL/MP: System-wide Communication for Heterogeneous Coupling Computing



Heterogeneous Coupling Computing by h3-Open-SYS/WaitIO and h3-Open-UTIL/MP (NICAM+ADA)



Integration of 3D Earthquake Simulations with Real-Time Data Observation & Assimilation on Wisteria/BDEC-01 using h3-Open-BDEC

The h3-Open-SYS/WaitIO(WaitIO) is a system-wide communication library to couple multiple MPI programs for heterogeneous environments, such as Wisteria/BDEC-01. The WaitIO provides an inter-program communication environment among MPI programs and supports different MPI libraries with various interconnects and processor types using an MPI-like API. As a higher-level software based on this library, a coupler for coupling multiple MPI programs called h3-Open-UTIL/MP(U/MP) has been developed simultaneously with WaitIO as part of the h3-Open-BDEC project. The U/MP provides capabilities for multi-physics coupling between different grid systems. Furthermore, it is equipped with a function for combined ensemble, and with an interface for coupling codes for large-scale simulations and applications written in Python.

The NICAM-ADA coupling utilizes U/MP (h3o-U/MP) for converting NICAM data in a high-resolution grid on Odyssey to a low-resolution grid for machine learning by GPUs on Aquarius. While such coupled computing has been only possible by MPI on a single/homogeneous system, U/MP with WaitIO provides a more flexible interface for integration of (S+D+L) on the heterogeneous system, such as Wisteria/BDEC-01.

Innovative Supercomputing by Integration of “Simulation/Data/Learning (S+D+L)” and **Beyond**

Long but “Straight” Road to “Miyabi”

To achieve maximum performance under power constraints, transitioning to accelerators like GPUs is crucial over the next decade. Since February 2022, energy costs have risen, ITC/UTokyo and the University of Tsukuba decided to introduce a GPU-based system, Miyabi, as OFP’s successor at JCAHPC in Fall 2021. Preparing to port over 3,000 OFP user codes to GPUs is estimated to take 18-30 months. In June 2022, NVIDIA’s GPU (H100 or successor) was chosen for its high computational performance and Fortran application portability. Users are encouraged to port their applications to GPUs themselves, with support options like 1-week hackathons, monthly consultation meetings via Zoom, and a comprehensive portal site. Additionally, we support 19 community code groups and OpenFOAM. Mostly, our users’ codes are parallelized by MPI+OpenMP, therefore OpenACC and Standard Parallelism are recommended.

Category	Name (Organizations)	Target, Method etc.	Language
Engineering (5)	FrontISTR (U.Tokyo)	Solid Mechanics, FEM	Fortran
	FrontFlow/blue (FFB) (U.Tokyo)	CFD, FEM	Fortran
	FrontFlow/red (AFFr) (Advanced Soft)	CFD, FVM	Fortran
	FFX (U.Tokyo)	CFD, Lattice Boltzmann Method (LBM)	Fortran
	CUBE (Kobe U./RIKEN)	CFD, Hierarchical Cartesian Grid	Fortran
Biophysics (3)	ABINIT-MP (Rikkyo U.)	Drug Discovery etc., FMO	Fortran
	UT-Heart (UT Heart, U.Tokyo)	Heart Simulation, FEM etc.	Fortran, C
	Lynx (Simula, U.Tokyo)	Cardiac Electrophysiology, FVM	C
Physics (3)	MUTSU/iHallMHD3D (NIFS)	Turbulent MHD, FFT	Fortran
	Nucl_TDDFT (Tokyo Tech)	Nuclear Physics, Time Dependent DFT	Fortran
	Athena++ (Tohoku U. etc.)	Astrophysics/MHD, FVM/AMR	C++
Climate/ Weather/ Ocean (4)	SCALE (RIKEN)	Climate/Weather, FVM	Fortran
	NICAM (U.Tokyo, RIKEN, NIES)	Global Climate, FVM	Fortran
	MIROC-GCM (AORI/U.Tokyo)	Atmospheric Science, FFT etc.	Fortran77
	Kinaco (AORI/U.Tokyo)	Ocean Science, FDM	Fortran
Earthquake (4)	OpenSWPC (ERI/U.Tokyo)	Earthquake Wave Propagation, FDM	Fortran
	SPECFEM3D (Kyoto U.)	Earthquake Simulations, Spectral FEM	Fortran
	hbi_hacapk (JAMSTEC, U.Tokyo)	Earthquake Simulations, H-Matrix	Fortran
	sse_3d (NIED)	Earthquake Science, BEM (CUDA Fortran)	Fortran

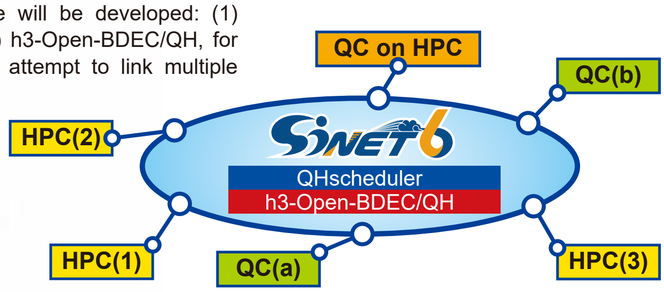
19 applications for “Supported Porting”: 17 of them are written in Fortran

JHPC-quantum

In November 2023, a new project on “Infrastructure for QC (Quantum Computers)-HPC (Supercomputers) Hybrid Computing (JHPC-quantum)” commenced. This five-year project, funded by the Japanese Government (NEDO), involves RIKEN R-CCS, Softbank, the University of Tokyo, and Osaka University. The University of Tokyo team is developing system software for efficient QC-HPC hybrid environments, where QC’s are utilized as accelerators of HPC’s. Following two types of software will be developed: (1) QHscheduler, a job scheduler for multiple remote computer resources, and (2) h3-Open-BDEC/QH, for real-time QC-HPC communication and data transfer. This is the world's first attempt to link multiple supercomputers and quantum computers installed at different sites in real time.

JHPC-quantum

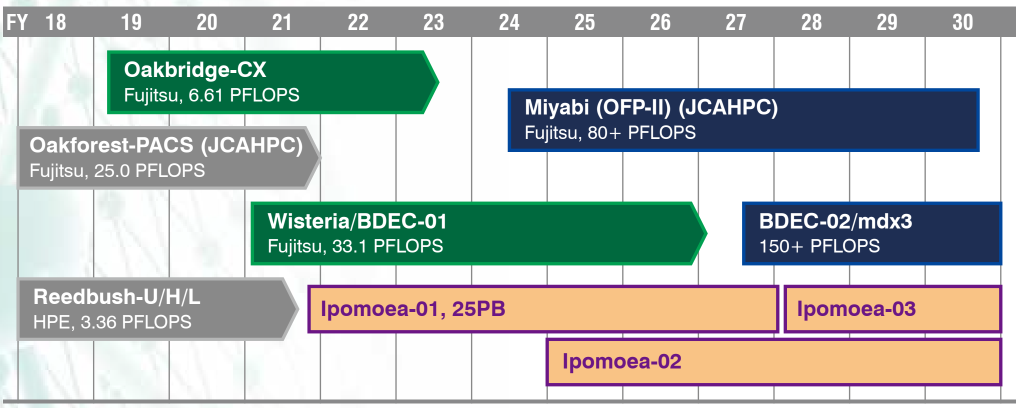
QC-HPC Hybrid: Quantum Computer as Accelerator of Supercomputers



QC-HPC Hybrid Environment over SINET6 using QHscheduler/h3-Open-BEDC/QH

Near Future Plan

Miyabi with NVIDIA GH200 and Intel Xeon Max starts its operation in January, 2025. The Wisteria/BDEC-01 will retire in the end of April 2027. At present, its successor (BDEC-02) is scheduled to be introduced in FY.2027, as an integrated system BDEC-02/mdx3 with the successor to mdx (mdx). Basic design is currently underway. We plan to leverage the respective characteristics of supercomputers and cloud-based data platforms to further promote integration of (S+D+L), and support a wide variety of applications. Furthermore, BDEC-02/mdx3 plans to promote hybrid collaboration with QC’s.



Cloud Platform and Infrastructure

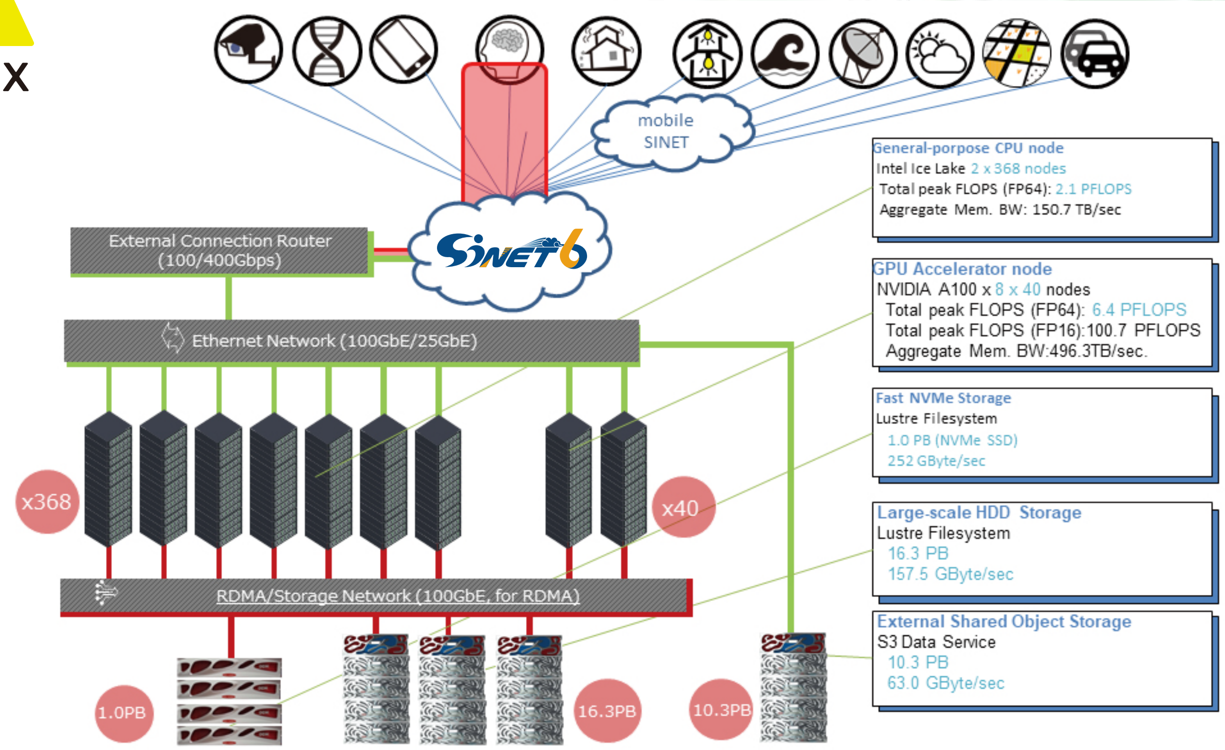
mdx: A Cloud Platform for Supporting Data Science and Cross-Disciplinary Research Collaborations

The growing amount of data and advances in data science have created a need for a new kind of cloud platform that provides users with flexibility, strong security, and the ability to couple with supercomputers and edge devices through high-performance networks. We have built such a nation-wide cloud platform, called "mdx" to meet this need. The mdx platform' s virtualization service, jointly operated by 9 national universities and 2 national research institutes in Japan, launched in 2021, and 50+ projects have been using it. Currently mdx is used by researchers in a wide variety of domains, including materials informatics, geo-spatial information science, life science, astronomical science, economics, social science, and computer science. Through the mdx platform, we anticipate more collaborations for such problems that need interdisciplinary approaches by bringing knowledge and skills in data science and application domains.

The mdx platform currently provides a VM hosting service (IaaS) and will provide data sharing and collaboration tools (PaaS) in the future. It consists of Fujitsu PRIMERGY, which is equipped with 368 CPU nodes and 40 GPU nodes as computing resources, and 1.0 Petabyte NVMe disk storage and 16.3 Petabyte hard disk for the Lustre file system, and 10.3 Petabyte S3-compliant object storage (DDN S3 Data Services) as storage resources. Each compute node employs 2 sockets of Intel Xeon Platinum 8368 with 38 cores of 2.4GHz clock, and each GPU node contains 8 GPUs of NVIDIA A100. These are virtualized by VMware vSphere virtualization platform and co-designed with SINET6, the ultra high-speed/low-latency academic backbone network in Japan. The network is virtualized via L2VPN to achieve high-performance and secure networks connecting to edge devices.

		General-purpose CPU node	GPU Accelerator node
Number of nodes		368	40
Compute node		Fujitsu PRIMERGY CX2550M6	Fujitsu PRIMERGY GX2570M6
CPU		Intel Xeon Platinum 8368 (Ice Lake SP, 38 cores, 2.4 GHz) x 2 sockets 5.83 TFLOPS	
Memory		256 GB (DDR4-3200 x 8ch x 2), 409.6 GB/s	
GPU		None	NVIDIA A100 GPUs (19.5 TFLOPS, 40 GB, 1.555 TB/s, SXM4, NVlink3) x 8
Interconnect (Front)		Ethernet (25 Gbps)	Ethernet (25 Gbps) x 2 link
Interconnect (RDMA / Storage)		Ethernet w. RoCEv2 (100 Gbps)	Ethernet w. RoCEv2 (100 Gbps) x 4 link
Storage	Fast	1.0 PB (NVMe SSD), 252 GB/s	
	Large-scale	16.3 PB (HDD), 157.5 GB/s	
	External shared object	10.3 PB (HDD), 63.0 GB/s, S3 Data service	
Virtualization		VMware vSphere, Overlay network w. EVPN-VXLAN	

mdx



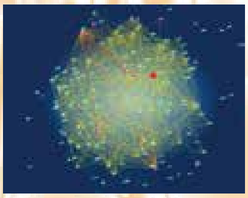
Data Science Research

The Data Science Research Division of the Information Technology Center at the University of Tokyo is engaged in a variety of research themes in data science and artificial intelligence. These areas include **graph neural networks, large language models, materials informatics**, and **system software**. Additionally, the division is at the forefront of a significant Japanese academic cloud initiative, "**mdx**".

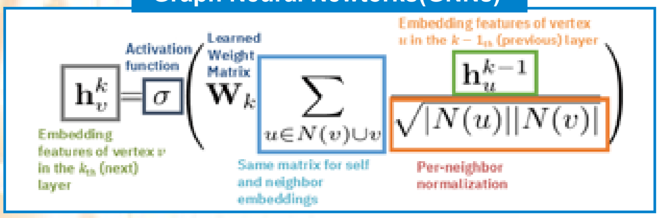
Graph Foundation Model for Artificial Intelligence

Foundation models mark a significant evolution in deep learning, trained on expansive datasets to enhance versatility across a range of tasks. This approach has particularly revolutionized the language domain, with GPT-4 exemplifying its capabilities. Extending this concept, we introduce "graph foundation models" that utilize graph data—a ubiquitous and versatile data structure. These models leverage graph neural networks as their core architecture, enriched further by integrating additional modalities like language, images, and time-series data. We focus on employing graph foundation models in practical, real-world applications, including e-commerce and news recommendation systems, job matching platforms, financial analytics, neuroscience, and materials informatics. This exploration underscores the widespread potential of this innovative paradigm.

Finance



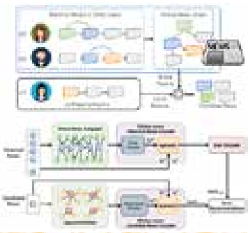
Graph Neural Newtorks(GNNs)



E-Commerce

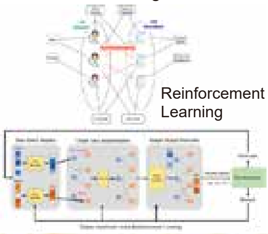
News Recommendation

User sequence modeling with GNN and its recommendation



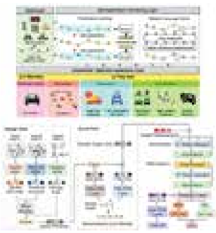
Job Matching

GNN and Reinforcement Learning Based Job Matching



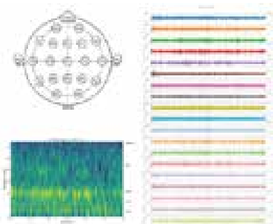
Mobility

Transformer/GNN-based Pretrained Models for Mobility



Brain Science

Building Robust Pretrained Models for EEG Data



Machine learning from large-scale electronic-structure simulation database

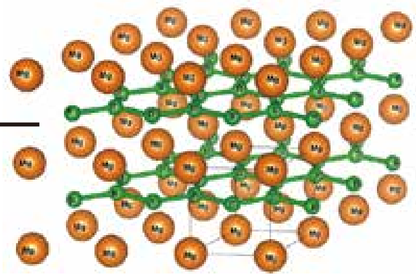
Recent progress in first-principles calculation enables us to perform high-throughput simulation of materials properties.

Bulk properties
- Band gap, optical property
- Thermal conductivity, thermoelectric figure of merit
- Curie temperature, magnetic cohesivity
- Superconducting transition temperature

Atom-decomposed properties
- Charge-transfer from isolated atom
- Magnetic moment
- Partial density of states

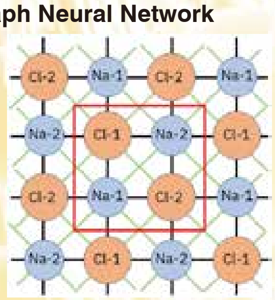
$$\left(\frac{\nabla^2}{2} + V_{KS}(\mathbf{r})\right)\psi_{nk}(\mathbf{r}) = \varepsilon_{nk}\psi_{nk}(\mathbf{r})$$

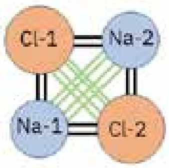
Kohn-Sham eq. (density functional theory)



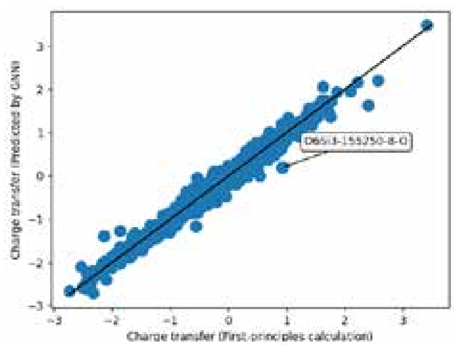
Electronic Structure Database

High-throughput simulation + node-attribution prediction with Graph Neural Network





Describe infinite periodic crystal with graph

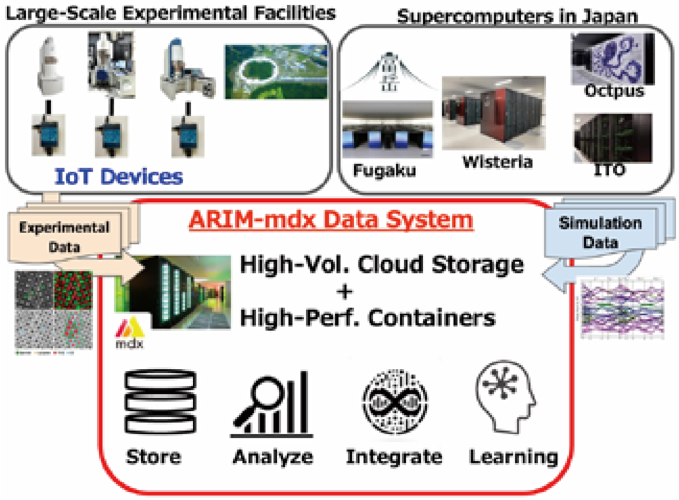


Data Science Research

ARIM-mdx Data System:

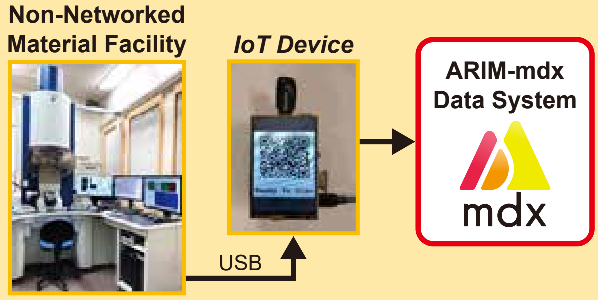
M. Hanai, M. Kawamura, R. Ishikawa, T. Suzumura, and K. Taura,
The 16th IEEE/ACM International Conference on Utility and Cloud Computing (UCC 2023).

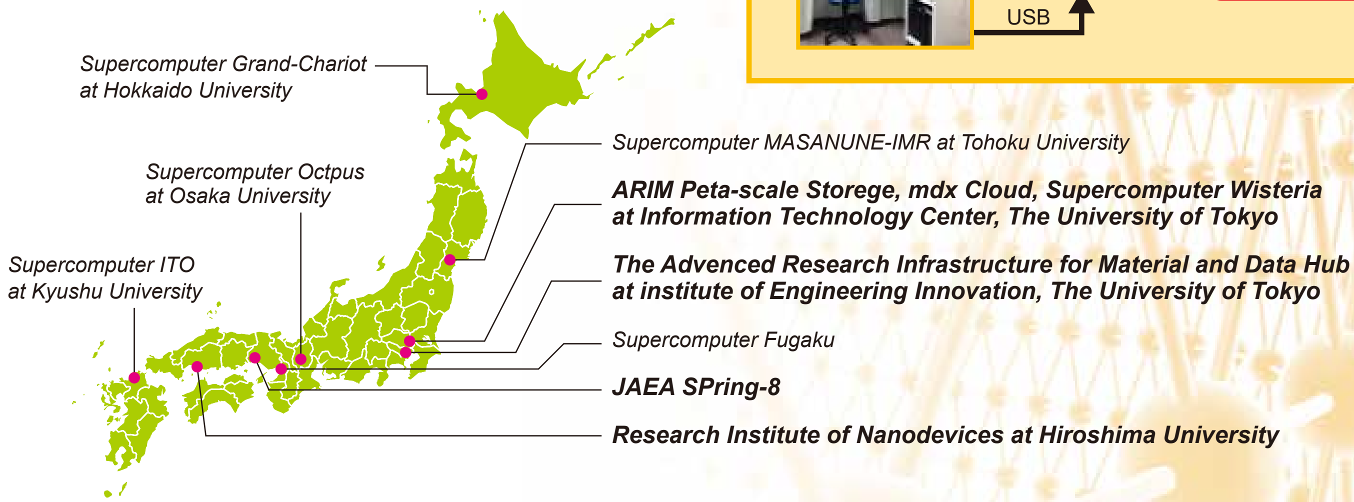
Comprehensive and Nation-wide Data Platform for Experimental, Theoretical, and Data-Driven Material Science and Development.



IoT-based Cloud Data Acquisition from Non-Network Facilities

- Efficient data transfer from experimental equipment that cannot connect to the Internet
- IoT devices emulating USB memory
- User management and coordination for automatic distribution to individual folders





Multi-threaded scp

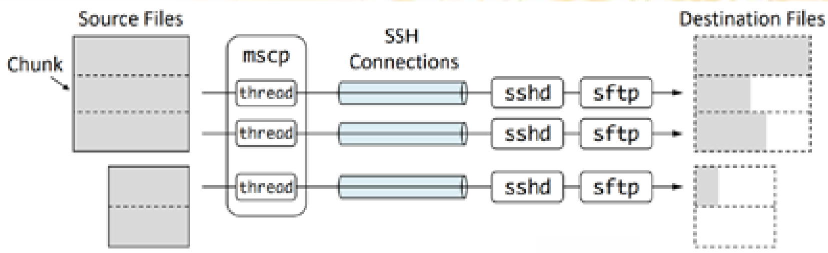
mscp, a variant of scp, copies files over multiple SSH (SFTP) connections by multiple threads. It enables transferring (1) multiple files simultaneously and (2) a large file in parallel, reducing the transfer time for a lot of/large files over networks.

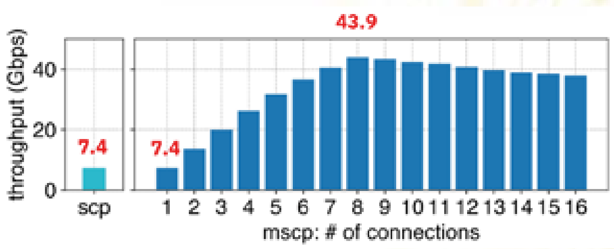
https://github.com/upa/mscp

Available for major Linux distributions and macOS

```
$ mscp srcfile example.com:dstfile
```

R. Nakamura and Y. Kuga, PEARC '23: Practice and Experience in Advanced Research Computing, 320 (2023).

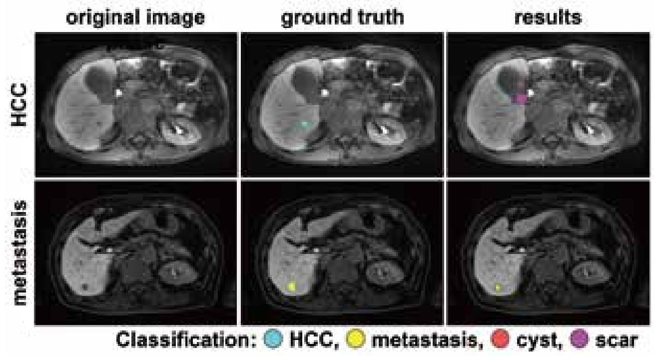


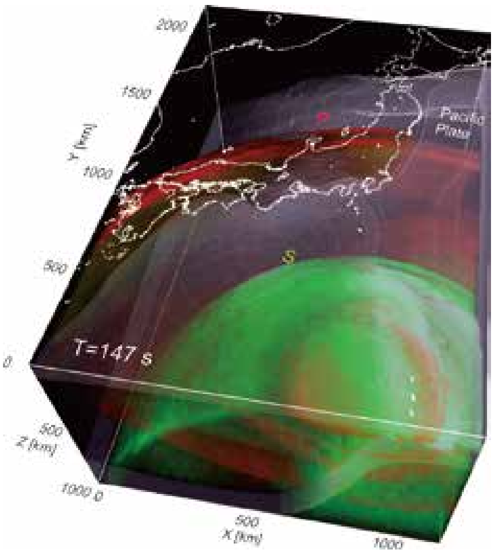


mscp won **the Most Innovative for HPC Uses Award** at Data Mover Challenge 2023

User Applications

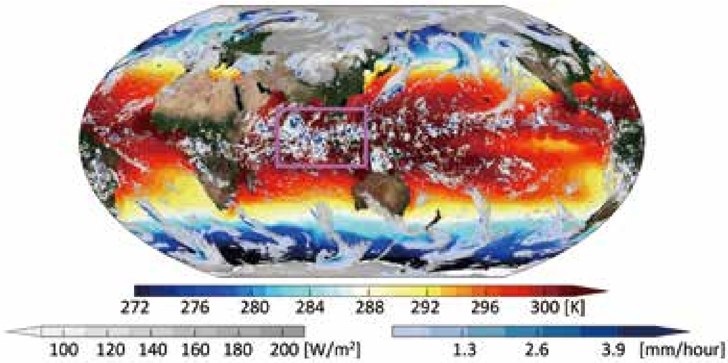
Focal Liver Lesion Detection by Deep Learning
c/o T. Takenaga (U.Tokyo Hospital)

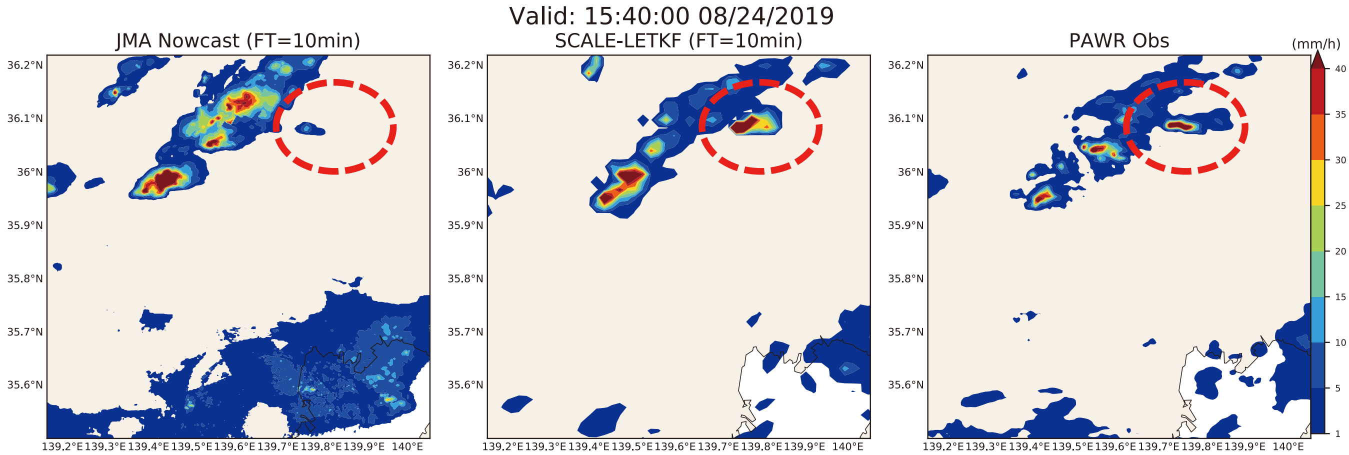




Seismic Wave Propagation Simulation
c/o T. Furumura (ERI/U.Tokyo)

Atmosphere-Ocean Simulations: NICOCO
c/o T. Miyakawa (AORI/U.Tokyo)





Numerical Weather Prediction System: SCALE-LETKF **c/o T. Miyoshi & T. Honda (RIKEN R-CCS)**

Wisteria BDEC-01

miyabi

mdx

Hierarchical, Hybrid, Heterogeneous h3-Open-BDEC Big Data & Extreme Computing

JHPC-quantum

東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO
6-2-3 Kashiwanoha, Kashiwa-shi, Chiba 277-0882, JAPAN TEL: +81-4-7133-4658

Visit our website for more information
https://www.cc.u-tokyo.ac.jp/en/public/pamphlet.php

2024.09