

Supercomputing Division, Information Technology Center, The University of Tokyo











ITC

Welcome

SCD/ITC, The University of Tokyo, Japan

The Supercomputing 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. The Information Technology Center (ITC) was organized in 1999, and the Supercomputing Center became the Supercomputing Division (SCD) of the ITC, joining three other divisions at that time. ITC is also a core organization of the "Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures" project, and a part of HPCI (the High-Performance Computing Infrastructure) operated by the Japanese Government. The three main missions of SCD/ITC are (i) providing services for supercomputer operations and supporting supercomputer users, (ii) doing research, and (iii) providing education and training. Currently, SCD/ITC consists of more than 10 faculty members. SCD/ITC is now operating three supercomputer systems, "Integrated Supercomputer System for Data Analyses & Scientific Simulations (Reedbush-U/H)" by HPE with 1.93 PFLOPS, "Supercomputer System with Accelerators for Long-Term Executions (Reedbush-L)" by HPE with 1.43 PFLOPS and "Manycore-based Large-scale Supercomputer System (Oakforest-PACS)" by Fujitsu with 25 PFLOPS as JCAHPC.

Services for Academia and Industry

The three supercomputer systems operated by SCD/ITC contain >2,000 users; 50% of these users are from outside the university. All the systems are extremely busy, and their average utilization ratio is ~90%. Providing services to support these users is one of our most important responsibilities. Hands-on tutorials for parallel programming are held ~10 times per year, and individual on-site consulting is also available. Up to 10% of the total computational resources of the Reedbush-U/H/L systems and the Oakforest-PACS system are open to users from the industry.

Computational Science Alliance, the University of Tokyo

Experience and knowledge concerning parallel programming are key advantages in the development of code for complicated, large-scale problems on massively parallel computers. At the University of Tokyo, we established the Computational Science Alliance (http://www.compsci-alliance.jp/) in 2015 by collaborating with 13 departments, including ITC. The primary purpose of this alliance is to provide an high-performance computing (HPC) interdisciplinary education program for CS&E with flexible and comprehensive classes and courses. The alliance started lectures in April 2017. In addition, we are conducting a series of international symposiums (International Symposium on Research and Education of Computational Science, RECS) in 2016, 2017, and 2018 (in planning).

Joint Center for Advanced High Performance Computing (JCAHPC)

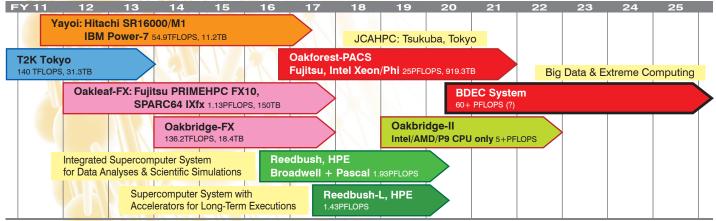
In 2013, Center for Computational Sciences, University of Tsukuba (CCS) and ITC agreed to establish the Joint Center for Advanced High Performance Computing(JCAHPC). JCAHPC consists of more than 20 faculty and staff members of CCS and ITC. Primary mission of JCAHPC is designing, installing and operating Oakforest-PACS system. In addition, CCS and ITC will develop system software, numerical libraries, and large-scale applications to for Oakforest-PACS system in collaboration made possible by the establishment of JCAHPC. JCAHPC is a new model for collaboration for research and development between supercomputer centers.

http://jcahpc.jp/

Supercomputer Systems in SCD/ITC: Oakforest-PACS and Reedbush-U/H/L

SCD/ITC started the operation of two new systems in FY.2016. The first system is JCAHPC's Oakforest-PACS by Fujitsu, which comprises 8,208 compute nodes with Intel Xeon Phi processors and started full operation on December 1, 2016. Oakforest-PACS has been offered to researchers in Japan and their international collaborators through various programs operated by the HPCI, MEXT's Joint Usage/Research Centers, and CCS and ITC under their original supercomputer resource sharing programs. Oakforest-PACS is contributing to dramatic developments in new frontiers of various fields of study, including computational science and engineering (CSE). Oakforest-PACS is also used for education and the training of students and young researchers in both CSE and HPC. The second system is the "integrated supercomputer system for data analyses and scientific simulations (Reedbush-U/H)" by HPE with Intel Broadwell-EP and NVIDIA Tesla P100 (Pascal) at 1.93 PFLOPS.

Reedbush-U is a traditional cluster with only Intel Broadwell-EPs (420 compute nodes), and Reedbush-H is our first GPU cluster with 120 compute nodes, each including two NIDIA Pascal GPUs. In addition, a third system (supercomputer system with accelerators for long-term wxecutions (Reedbush-L)) started operation in October 2017. The Reedbush-L system is developed by HPE with Intel Broadwell-EP and NVIDIA Tesla P100 (Pascal) at 1.43 PFLOPS; it has 64 compute nodes, and each node has four NVIDIA Pascal GPUs. In March 2019, our new system (Oakbridge-II) will begin its operation; the Oakbridge-II homogeneous compute nodes have multicore CPUs and are expected to achieve >5.0 PFLOPS at peak performance. In addition, we plan to introduce a new system (BDEC, Big Data & Extreme Computing) with 60+ PFLOPS after the fall of 2020.

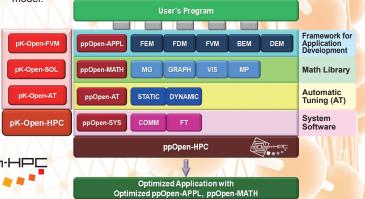


nternational & Domestic projects

ppOpen-HPC & ESSEX-II

"ppOpen-HPC" is an open source infrastructure for the development and execution of optimized and reliable simulation code on post-peta-scale (pp) parallel computers based on many-core architectures comprising various types of libraries that cover general procedures for scientific computations. Source code developed on a PC with a single processor is linked to these libraries, and the generated parallel code is optimized for post-peta-scale systems. The target post-peta-scale system is the Post T2K System. "ppOpen-HPC" is part of a five-year project (FY.2011-FY.2015) spawned from the "Development of System Software Technologies for Post-Peta Scale High-Performance Computing" project funded by JST-CREST. The framework covers various types of procedures for scientific computations, such as the parallel I/O of datasets, matrix assembly, linear solvers with practical and scalable preconditioners, visualization, adaptive mesh refinement, and dynamic load balancing, in various types of computational models, such as FEM, FDM, FVM, BEM, and DEM. Automatic tuning (AT) technology enables the automatic generation of optimized libraries and applications under various types of environments. We released the most updated version of ppOpen-HPC as an open source software every year in November from 2012 to 2015 (available at http://ppopenhpc.cc.u-tokyo.ac.jp/ppopenhpc/).

In 2016, the ppOpen-HPC team joined the Equipping Sparse Solvers for Exascale (ESSEX-II) project (led by P.I. Professor Gerhard Wellein of the University of Erlangen-Nuremberg, http://blogs.fau.de/essex/), which is funded by JST-CREST and the German DFG priority program 1648 "Software for Exascale Computing" (SPPEXA) under a Japan (JST)–Germany (DFG) collaboration, which continues until FY.2018. In the ESSEX-II project, we are developing pK-Open-HPC (an extended version of ppOpen-HPC, which is a framework for exa-feasible applications), preconditioned iterative solvers for quantum sciences, parallel reordering methods, and a framework for AT with a performance model.

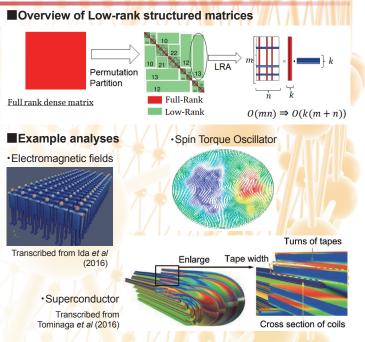


HACApK : Distributed Memory H-matrices Library

Low-rank structured matrices represented by hierarchical matrices (\mathcal{H} -matrices) have recently received attention as a fast computational technique for dense matrices arising from scientific simulations. For a matrix size *N*, the memory complexity of low-rank structured matrices is at worst *O*(*N* log *N*), which is much lower than that of dense matrices *O*(N^2).

We have been developing an open-source \mathcal{H} -matrices library called $\mathcal{H}ACApK$. The development of the library began in 2012 as a part of the ppOpen-HPC project. To exploit recent supercomputer systems, $\mathcal{H}ACApK$ is designed for SMP clusters. Since 2017, our proposals to enhance the library functions have been accepted as an international joint research project of JHPCN, and JSPS KAKENHI projects. In these projects, we address a wide spectrum of issues, such as porting of the library to GPU and FPGA, increasing the available matrix arithmetic functions, and improving the \mathcal{H} -matrices for massively parallel processing. The latest version of the $\mathcal{H}ACApK$ library is available on the webpage of the ppOpen-HPC project.

The $\mathcal{H}ACApK$ library is employed for practical simulations of electric fields, earthquake cycles, superconductors, and micromagnetics. The use of the library enables large-scale simulations to be conducted. In addition, we have been challenging new frontier studies to apply this library. For example, we have been exploring the application of $\mathcal{H}ACApK$ to the elastodynamic boundary integral equation method to investigate earthquake rupture dynamics as a general JHPCN project since 2017.



JHPCN: Japan High Performance Computing & Networking plus Data Analysis and Information Systems



Japan high-performance computing and networking plus data analysis and information systems (JHPCNs) are developed by the "Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures," which comprises eight academic supercomputer centers in Japan associated with Hokkaido

University, Tohoku University, the University of Tokyo, Tokyo Tech, Nagoya University, Kyoto University, Osaka University, and Kyushu University (core organization: the University of Tokyo). This project began in April 2010. The total performance of the supercomputer systems involved is ~70 PFLOPS (April 2018). 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. Since 2013, the JHPCN centers have been responsible for the operation of these joint research resources, which is called the HPCI-JHPCN system because it is part of the HPCI system. So, far, 35–40 projects have been accepted each year since 2010. Since 2017, JHPCN has initiated new frameworks for collaborative research, international collaborative research, and industry collaborative research.

Scientific Computing

High-productivity Framework for Stencil Applications

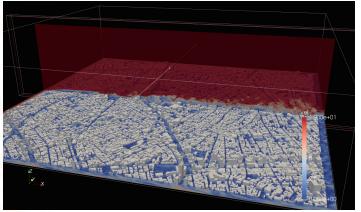
Stencil applications such as computational fluid dynamics are the major applications in high-performance computing. These applications have successfully obtained high performance on modern supercomputers equipped with accelerators such as GPU and Xeon Phi. Obtaining high-performance using thousands of accelerators often needs skillful programming.

We are currently developing the high-productivity framework. The framework is designed for stencil applications with explicit time integration running on regular structured grids. Our framework is implemented in C++ and CUDA languages. It automatically translates user-written stencil functions that update a grid point and generates both GPU and CPU codes. A stencil function can be defined as a C++ functor. The programmers write user code just in the C++ language, and it can be executed on multiple GPUs with the auto-tuning mechanism and the overlapping method to hide communication cost by computation. It can be also executed on multiple CPUs with OpenMP without any change of code. In addition, our framework provides a data structure that supports element-wise computations, which allow us

data structure that supports element-wise computations, which allow us to write GPU kernel codes as inline codes.

We are introducing the mechanism for enabling the computations beyond the capacity of the GPU device memory into this stencil framework. We realize this by a combination of a temporal blocking method for locality improvement and an automatic swapping between GPU and CPU. The temporal blocking technique can suppress performance degradation caused by frequent memory swapping between GPU and CPU. The automatic swapping is based on a MPI/CUDA wrapper run-time library called HHRT. By using the framework-based approach, computation exceeding the capacity of the GPU device memory is realized without complicated modification of the structure of the time integration loop accompanying data movement between GPU and CPU. The framework-based application for the airflow in an urban city preserves 80% performance of the maximum performance obtained by the original version even with the twice larger than the GPU memory capacity.

An example of a stencil function



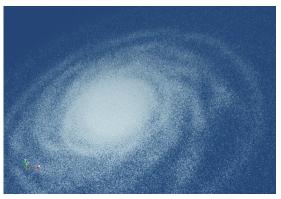
A snapshot of simulation results of the airflow in an urban city.

Software development for numerical astrophysics

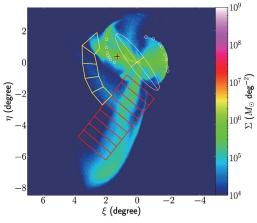
Collisionless *N*-body simulations are frequently employed to explore the formation and evolution of gravitational many-body systems, such as galaxies or large-scale structure of the Universe. An initial-condition generator and a gravitational *N*-body code are standard toolkits must be developed and optimized for such systems.

The initial conditions for idealized galaxies in *N*-body simulations that resemble observed systems should be dynamically stable. However, generating a galaxy model as a system in dynamical equilibrium is difficult because a galaxy contains several components, including a bulge, disk, and halo. Moreover, most disk galaxies represented by the Milky Way possess multiple disk components having different thicknesses. MAny-component Galaxy Initializer (MAGI) is a newly developed initial-condition generator that satisfies these requirements. The developed generator supports various types of density models, their superposition, and the presence of multiple disks. We tested the dynamical stability of systems generated by MAGI representing elliptical and disk galaxies and confirmed that the model galaxies maintained their initial distributions for over a billion years. The execution times required to generate particle distributions are negligible compared to the typical execution times for *N*-body simulations.

The tree method is a fast algorithm for collisionless *N*-body simulations in astrophysics that is well suited for GPU (graphics processing units) implementations. Adopting hierarchical time stepping can accelerate *N*-body simulations; however, it is infrequently implemented and its potential remains untested in GPU implementations. We have developed a Gravitational Oct-Tree code accelerated by HIerarchical time step Controlling called GOTHIC, which includes both the tree method and the hierarchical time step. The code incorporates several adaptive optimizations by monitoring the execution time of each function and minimizes the time-to-solution by balancing the measured time of multiple functions. The code is optimized for Fermi, Kepler, Maxwell, and Pascal generation GPUs. Results of performance measurements using the Andromeda galaxy model show that the hierarchical time step achieves a speedup by a factor of approximately three to five times compared to a shared time step. The averaged performance of the code corresponds to 10%–30% of the theoretical single precision peak performance of a GPU.



An example of disk galaxy model generated by MAGI.



A snapshot of galactic merger simulation.

System, tools & hardware

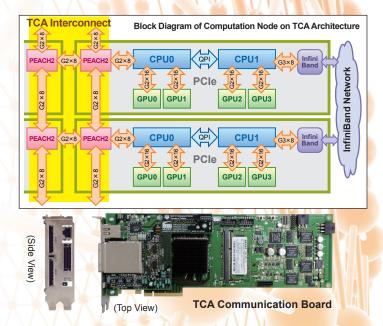
Tightly Coupled Accelerators(TCA)

GPGPU is now widely used for accelerating scientific and engineering computing to improve performance significantly with less power consumption. However, I/O bandwidth bottleneck causes serious performance degradation on GPGPU computing. Especially, latency on inter-node GPU communication significantly increases by several memory copies. To solve this problem, TCA (Tightly Coupled Accelerators) enables direct communication among multiple GPUs over computation nodes using PCI Express. PEACH2 (PCI Express Adaptive Communication Hub ver. 2) chip is developed and implemented by FPGA (Field Programmable Gate Array) for flexible control and prototyping in cooperation with University of Tsukuba. PEACH2 board is also developed as an PCI Express extension board.

TCA provides the following benefits:

- Direct I/O among GPU memory over nodes
- Reduce the overhead, obtain good scaling
- Shared PCI Express address space among multiple nodes
 - Ease to program

PEACH2 can transfer not only GPU memory but also host memory seamlessly since PEACH2 relies on the PCIe protocol. The DMA controller in the PEACH2 chip provides a chaining DMA function in order to transfer multiple data segments using the chained DMA descriptors automatically via hardwired logic, and also supports a block-stride transfer which can be specified with a single descriptor.



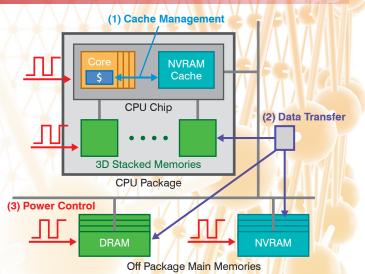
Improving Energy Efficiency using Emerging Memory Technologies

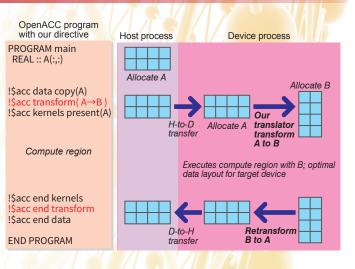
It is critical to improve the energy efficiency (performance per watts) of supercomputing nodes to build future exa-scale systems. This is because recent top-class supercomputers have already reached the power supply limit of a few tens of megawatts; therefore, we cannot merely scale the number of nodes to gain performance. In addition, VLSI technology scaling, which is the major contributor to energy-efficiency improvements, is coming to an end; therefore, alternative approaches must be developed.

To this end, we are focusing on the node architecture which comprises emerging memory technologies (e.g., 3D stacking and NVRAM), and developing software/hardware techniques to exploit the energy efficiency of the nodes. These memory technologies are helpful to scale bandwidth or capacity with smaller power budgets, and are indispensable to improve the performance of various memory intensive applications. Particularly, we are developing (1) data management on NVRAM-based cache hierarchies, (2) data transfer optimizations on hybrid main memories which comprise multiple different memory technologies, and (3) power control techniques for such systems. By combining these techniques, the energy efficiency can be considerably improved.

OpenACC Extension for Performance Portability

OpenACC is gaining momentum as an implicit and portable interface in porting legacy CPU-based applications to heterogeneous, highly parallel computational environment involving many-core accelerators such as GPUs and Intel Xeon Phi. OpenACC provides a set of loop directives similar to OpenMP for the parallelization and also to manage data movement, attaining functional portability across different heterogeneous devices; however, the performance portability of OpenACC is said to be insufficient due to the characteristics of different target devices, especially those regarding memory layouts, as automated attempts by the compilers to adapt is currently difficult. We are currently working to propose a set of directives to allow compilers to have better semantic information for adaptation; here, we particularly focus on data layout such as Structure of Arrays, advantageous data structure for GPUs, as opposed to Array of Structures, which exhibits good performance on CPUs. We propose a directive extension to OpenACC that allows the users to flexibility specify optimal layouts.

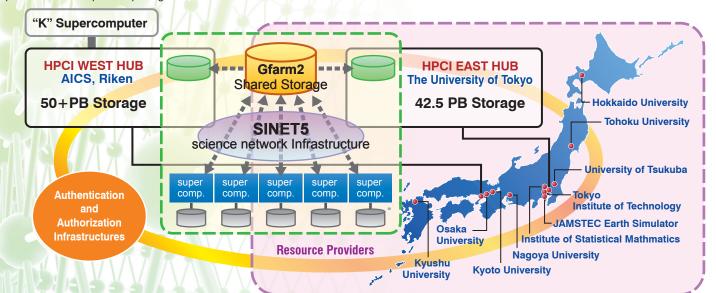




Supercomputers in SCD/ITC

HPCI: High Performance Computing Infrastructure

High performance computing infrastructure (HPCI) is an environment that enables an easy usage of flagship "K" 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 12 computational resource providers; nine of these providers are supercomputing centers at national universities and three are governmental research institutes. These resource suppliers are connected via SINET5, which is a high-speed academic backbone network with 100 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 42.5-PB storage system in combination with the WEST Hub.

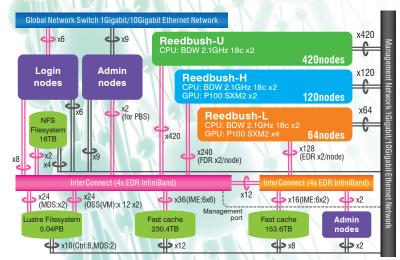


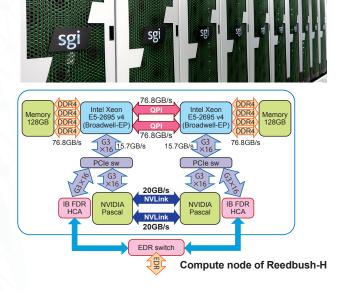
Reedbush (SGI Rackable system)

Reedbush is the first supercomputer system that introduced accelerators in SCD/ITC, and its total peak performance is up to 3.3 PFLOPS. We started computing service with Reedbush-U (CPUs only) in July 2016 and with the full system, including Reedbush-H (with 2 GPUs per node), in March 2017. In addition, the Reedbush-L (with 4 GPUs per node) subsystem has been available since July 2017. This system is installed on the Asano campus (our main campus) and is operated by HPE (ex-SGI). This system has the following missions.

- •Development of new research field, and promotion for new users, such as big data and deep learning researchers
- Development of a pilot system of a next-generation supercomputer system for the integration and fusion of data analyses and scientific simulations

Peak performance	Reedbush-U	Reedbush-H	Reedbush-L
Peak performance			
	509 TFlops	1417 TFlops	1433 TFlops
Number of nodes	420	120	64
Total memory size	105 TByte	30 TByte + 3.75 TByte	16 TByte + 4 TByte
Compute node	SGI Rackable C2112-4GP3	SGI Rackable C1102-GP8	
CPU	Intel Xeon E5-2695v4 (Broadwell-EP, 18 core, 2.1 GHz) x 2 socket 1209.6 GFlops		
Memory	256 GB (DDR4-2400 x 4ch x 2), 153.6 GB/sec		
GPU	None	NVIDIA Tesla P100 (Pascal, 5.3 TFlops, 16 GB, 720 GB/sec) x 2	NVIDIA Tesla P100 (Pascal, 4.8-5.3 TFlops, 16 GB, 720 GB/sec) x 4
Interconnect	InfiniBand EDR 4x (100 Gbps)	InfiniBand FDR 4x 2 link (56 Gbps x2)	InfiniBand EDR 4x 2 link (100 Gbps x2)
Interconnect topology	Full-bisection BW Fat Tree		Full-bisection BW Fat Tree
Parallel file system	Lustre Filesystem (DDN SFA14KE x3) 5.04 PB		145.2 GB/sec
File cache system	Burst buffer (DDN IME14K x6) 230.4 TB, 385.2 GB/sec		Burst buffer (DDN IME240 x8) 153.6 TB, 166.4 GB/sec
	Total memory size Compute node CPU Memory GPU Interconnect Interconnect topology Parallel file system	Total memory size 105 TByte Compute node SGI Rackable C2112-4GP3 CPU Intel Xeon E5-269 Memory 256 GB GPU None Interconnect InfiniBand EDR 4x (100 Gbps) Interconnect topology Full-bisection Parallel file system Lustre Filesyst File cache system Burst buffer (DI	Total memory size 105 TByte 30 TByte + 3.75 TByte Compute node SGI Rackable C2112-4GP3 SGI Rackable CPU Intel Xeon E5-2695v4 (Broadwell-EP, 18 core, 2. 1209.6 GFlops Memory 256 GB (DDR4-2400 x 4ch x 2), 153.6 GPU None NVIDIA Tesla P100 (Pascal, 5.3 TFlops, 16 GB, 720 GB/sec) x 2 Interconnect InfiniBand EDR 4x (100 Gbps) InfiniBand FDR 4x 2 link (56 Gbps x2) Interconnect topology Full-bisection BW Fat Tree Parallel file system Lustre Filesystem (DDN SFA14KE x3) 5.04 PB, File cache system





Supercomputers in SCD/ITC

Director Switch

362 of 48 port Edge Switch

Downlink:24

Compute Nodes

Fuiitsu PRIMERGY

CX-600 + CX-1640

25 PFlops

×8,208

system

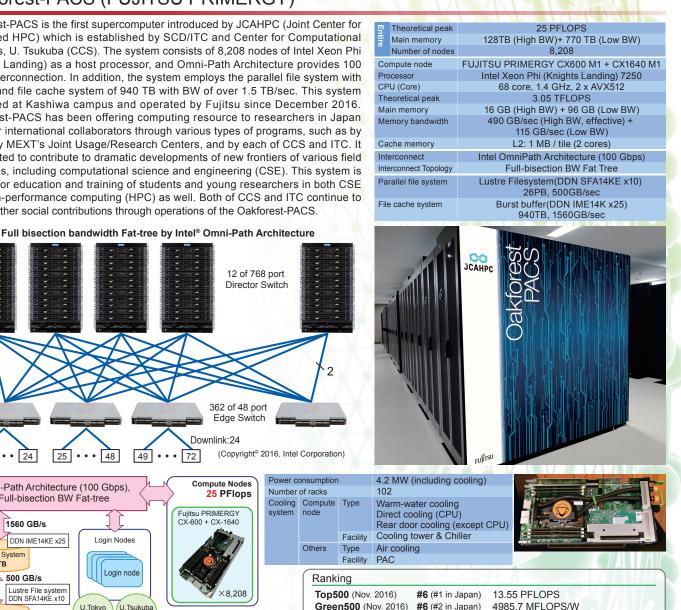
HPCG (Nov. 2016)

IO500 (Nov. 2017)

72

Oakforest-PACS (FUJITSU PRIMERGY)

Oakforest-PACS is the first supercomputer introduced by JCAHPC (Joint Center for Advanced HPC) which is established by SCD/ITC and Center for Computational Sciences, U. Tsukuba (CCS). The system consists of 8,208 nodes of Intel Xeon Phi (Knights Landing) as a host processor, and Omni-Path Architecture provides 100 Gbps interconnection. In addition, the system employs the parallel file system with 26 PB, and file cache system of 940 TB with BW of over 1.5 TB/sec. This system is located at Kashiwa campus and operated by Fujitsu since December 2016. Oakforest-PACS has been offering computing resource to researchers in Japan and their international collaborators through various types of programs, such as by HPCI, by MEXT's Joint Usage/Research Centers, and by each of CCS and ITC. It is expected to contribute to dramatic developments of new frontiers of various field of studies, including computational science and engineering (CSE). This system is utilized for education and training of students and young researchers in both CSE and high-performance computing (HPC) as well. Both of CCS and ITC continue to make further social contributions through operations of the Oakforest-PACS.



12 of 768 port

49

Future Plans of SCD/ITC

• • • 48

Login Nodes

Login node

U.Tsukuba

U.Tokyo

2

24

1560 GB/s

500 GB/s

File Cache System

940 TB

Parallel File System

26.2 PB

DDN IME14KE x25

Lustre File system DDN SFA14KE x10

25

Omni-Path Architecture (100 Gbps),

Full-bisection BW Fat-tree

Uplink²

Majority of SCD/ITC supercomputer system users belong to the fields of CSE, including engineering simulations (fluid dynamics, structural dynamics, and electromagnetics), earth sciences (atmosphere, ocean, solid earth, and earthquakes), and material sciences. Recently, the number of users related to data science, machine learning, and artificial intelligence (AI) has been increasing. Examples of new research topics are weather prediction by data assimilation, medical image recognition, and human genome analyses. Moreover, traditional CSE users are integrating data science, machine learning, and AI into their work to achieve efficient and accurate computations. In general, real-world applications are nonlinear and require numerous case studies to find accurate and reasonable solutions. A data driven approach (DDA) based on deep learning technology can reduce the number of cases Required to find such solutions. SCD/ITC is now developing a new research area to integrate computational science and data science using DDA. To develop this new research area, we plan to introduce a new system called "big data and extreme computing (BDEC)" after the fall of 2020. The peak performance of the BDEC system is expected to be 60+ PFLOPS, and it will comprise two types of compute nodes, internal nodes (INN) for traditional supercomputing applications, and external

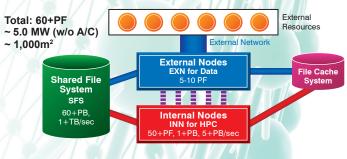
nodes (EXN) for data science. Each node of the EXN will be connected to external resources, such as data storage, directly through an external network (e.g., SINET, Japan). INN and EXN will share large-scale storage and a fast file cache system. Each node of INN and EXN could be based on a different architecture, and the computational resources of EXN will be 10% of the entire system. In particular, EXN may comprise CPU, GPU, FPGA, quantum/neuromorphic chips, and other custom chips. The Reedbush-U/H/L systems and the Oakbridge-II system are prototypes of the BDEC system.

385.5 TFLOPS

BW 471.25 GB/s · MD 21.85 kIOP/s

#3 (#2 in Japan)

#1





Information Technology Center, The University of Tokyo

