THE UNIVERSITY OF TOKYO

SCD/ITC

Supercomputing Division,
Information Technology Center,
The University of Tokyo

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, a Hitachi SR16000/M1 based on Power7 architecture with 54.9 TFLOPS of peak performance (Yayoi), a Fujitsu PRIMEHPC FX10 System (Oakleaf-fx) at 1.13 PFLOPS, and another Fujitsu PRIMEHPC FX10 System (Oakbridge-fx) at 136.2 TFLOPS for long-time execution.

Joint Center for Advanced High Performance Computing (JCAHPC)

In 2013, the Center for Computational Sciences, University of Tsukuba (CCS) and ITC agreed to establish the Joint Center for Advanced High Performance Computing (JCAHPC, http://jcahpc.jp/). JCAHPC consists of more than 20 faculty and staff members from CCS and ITC. The primary mission of JCAHPC is designing, installing and operating **the Post T2K System, a system** based on many-core architectures. The Post T2K System is expected be able to achieve 20-30 PFLOPS of peak performance, and will be installed in FY.2015 at the Kashiwa-no-Ha (Oakleaf) Campus of the University of Tokyo. The budget for the T2K supercomputer systems operated at CCS and ITC will be spent for installing and operating the Post T2K. In addition, CCS and ITC will develop system software, numerical libraries, and large-scale applications for the Post T2K system in collaboration made possible by the establishment of JCAHPC. JCAHPC is a new model for collaboration in research and development between supercomputer centers.

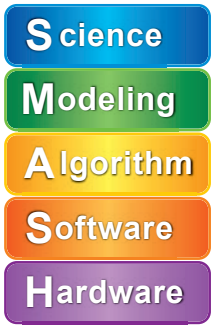
JCAHPC

Services for Academia and Industry

There are approximately 1,500 users on the three supercomputer systems operated by SCD/ITC, and 50% of them are from outside of the university. All of the systems are quite busy, and their average utilization ratio is approximately 90%. Providing services to support these users is one of our most important responsibilities. Hands-on tutorials for parallel programming are held 5-6 times per year, and individual on-site consulting is also available. Up to 10% of the total computational resources of the Oakleaf-FX system is open for users from industry.

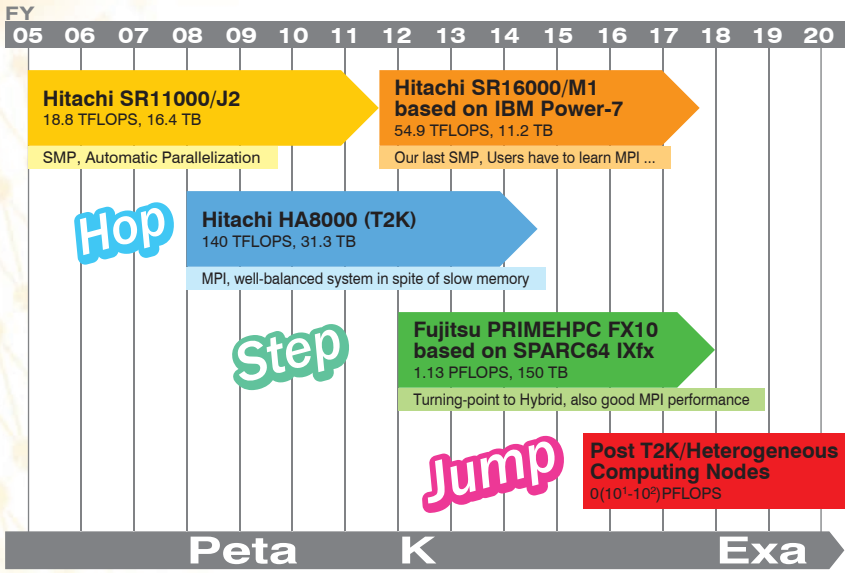
Interdisciplinary HPC Education Program for CS&E

Experience and knowledge of parallel programming are key advantages for the development of code for complicated, large-scale problems on massively parallel computers. At the University of Tokyo, we started a special "Interdisciplinary HPC Education Program for CSE" from FY2008 with the collaboration of four graduate schools, two research institutes, and ITC. Flexible and comprehensive classes and courses are provided based on the **SMASH** (Science -Modeling-Algorithm-Software-Hardware) model.



Supercomputer Systems in SCD/ITC

Recently, SCD/ITC has been installing a new supercomputing system every three years. **The Post T2K System**, based on many-core architectures and promising 20-30 PFLOPS of peak performance, will be installed in FY.2015 in collaboration with the University of Tsukuba. The Post T2K System is considered a *post-petascale* system, and is a really important milestone to an *exascale* system, which is expected to be developed by the end of the decade. **Oakleaf-FX** also plays an important role in the paradigm shift from single-level parallel programming models (*e.g.* pure MPI) to multi-level hybrid parallel programming models (*e.g.* MPI+OpenMP /CUDA /OpenCL /OpenACC, etc.), which will be used in post-petascale and exascale systems. ppOpen-HPC (see the back cover of this pamphlet) provides a parallel programming environment for scientific computing on the Post T2K System.

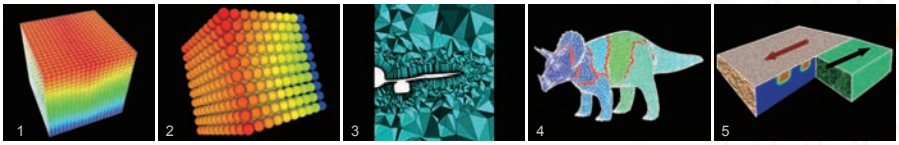


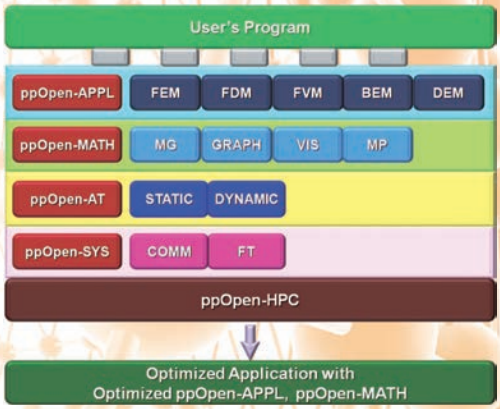
International & Domestic projects

ppOpen-HPC: Open Source Infrastructure for Development and Execution of Large-Scale Scientific Applications on Post-Petascale Supercomputers with Automatic Tuning (AT)

Recently, high-end parallel computer systems have become larger and more complex. Yet, it is very difficult for scientists and engineers to develop efficient application code that can take advantage of the potential for performance improvement of these systems. We propose an open source infrastructure for development and execution of optimized and reliable simulation code on large-scale parallel computers. We have named this infrastructure "ppOpen-HPC," where "pp" stands for "post-peta." The target system is the Post T2K System based on many-core architectures, which will be installed in FY2015. "ppOpen-HPC" is part of a five-year project (FY2011-2015) spawned by the "Development of System Software Technologies for Post-Peta Scale High Performance Computing" project funded by JST-CREST. The infrastructure consists of various types of libraries for scientific computations. Source code developed on a PC with a single processor is linked with these libraries, and the parallel code generated is optimized for post-peta-scale systems. The framework covers various types of procedures for scientific computations, such as parallel I/O of data-sets, matrix-formation, linear-solvers with practical and scalable pre-conditioners, visualization, adaptive mesh refinement and dynamic load-balancing, in various types of computational models, such as FEM, FDM, FVM, BEM and DEM. This type of framework will provide dramatic efficiency, portability, and reliability in the development and execution of scientific applications. It reduces both the number of steps in the source code and the time required for parallelization and optimization of legacy code. Automatic tuning (AT) technology enables automatic generation of optimized libraries and applications under various types of environments. We release the most updated version of ppOpen-HPC as open source software every year in November (2012-2015).

1:**FDM** [Finite Difference Method]
2:**DEM** [Discrete Element Method]
3:**FVM** [Finite Volume Method]
4:**FEM** [Finite Element Method]
5:**BEM** [Boundary Element Method]





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

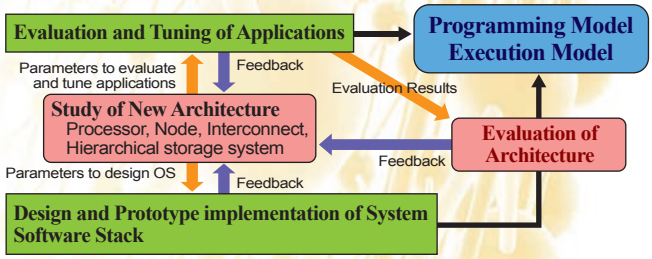
JHPCN (Japan High Performance Computing & Networking plus Data Analysis and Information Systems, http://jhpcn-kyoten.itc.u-tokyo.ac.jp/) is a 6-year project carried out by the "Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures," which consists of eight academic supercomputer centers in Japan: those of Hokkaido University, Tohoku University, the University of Tokyo, Tokyo Tech, Nagoya University, Kyoto University, Osaka University, and Kyushu University (Core Organization: U. Tokyo). The project was started in April of 2010. The total performance of the supercomputer systems involved is approximately 5 PFLOPS (April 2012). JHPCN promotes collaborative research projects using the facilities and human resources of these eight centers, including supercomputers, storage systems, and networks. Four main research areas have been defined: scientific computation, data analysis, networks, and large-scale IT systems. Interdisciplinary projects utilizing multiple facilities over networks are especially encouraged. So far, 35-40 projects have been accepted each year since 2010 (FY.2010 (37), 2011 (39), 2012 (35)).

Optimization of preconditioned iterative solvers for new Intel architecture

In December 2013, SCD/ITC joined **IPCC** (Intel® Parallel Computing Centers, *https://software.intel.com/en-us/ipcc#centers*). It is one of three Japanese institutes in IPCC. Our primary target as a member of IPCC is intensive optimization of preconditioned iterative solvers for structured/unstructured sparse coefficient matrices in *UTbench* for the new Intel Xeon and Intel Xeon Phi processors, and to construct general strategies for optimization of these procedures for the new processors.
UTbench consists of two codes, **GeoFEM-Cube/CG** and **Poisson3D-OMP**. GeoFEM-Cube/CG is a benchmark code based on GeoFEM, and it solves 3D static linear-elastic problems in solid mechanics. It contains typical procedures in finite-element computations, such as matrix assembling and preconditioned iterative solvers. Two types of parallel programming models (Flat-MPI and OpenMP/MPI Hybrid) are implemented to GeoFEM-Cube/CG. Poisson3D-OMP is a finite-volume based 3D Poisson equation solver using ICCG iterative method. Poisson3D-OMP is parallelized by OpenMP. Poisson3D-OMP supports a variety of reordering methods, methods for matrix storage (CRS and ELL), and coalesced/sequential memory access models.
Iterative solvers of GeoFEM-Cube/CG and Poisson3D-OMP also utilized as iterative solvers of ppOpen-HPC. Moreover, UTbench will be adopted as one of the benchmarks for procurement of Post T2K system in JCAHPC.

Feasibility Study on Future High Performance Computing Infrastructures

The Japanese government selected four interdisciplinary research teams (one for applications, and three for systems), running in FY2012 and 2013, for a feasibility study of future advanced HPC infrastructures. The "Feasibility study on advanced and efficient latency-core based architecture," led by SCD/ITC, is the responsibility of one of the three system study teams. Through this feasibility study, we will design supercomputer systems for scientific problems, identify R&D issues for development, evaluate the system using the selected applications, and estimate the cost of the system. Our team is focusing on general purpose supercomputers based on the K computer and FX10 systems. Target applications are ALPS, RSDFT, NICAM, COCO, QCD, FrontFlow/blue, and Modylas. Four groups (architecture design, application tuning, architecture evaluation, and system software design) are intensively involved in "co-design."



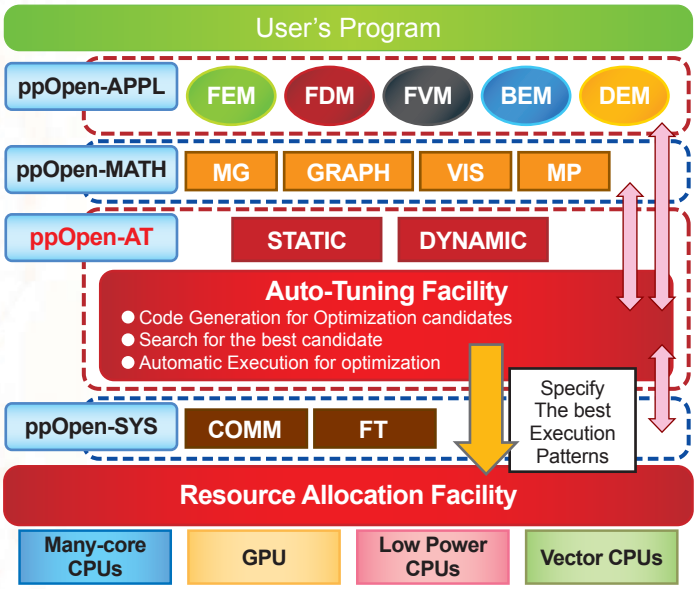
Scientific Computing

ppOpen-AT: An Auto-tuning Description Language for ppOpen-HPC

Computer architectures are becoming more and more complex due to non-uniform memory accesses and hierarchical caches. It is very difficult for scientists and engineers to optimize their code to extract potential performance improvements on these architectures.

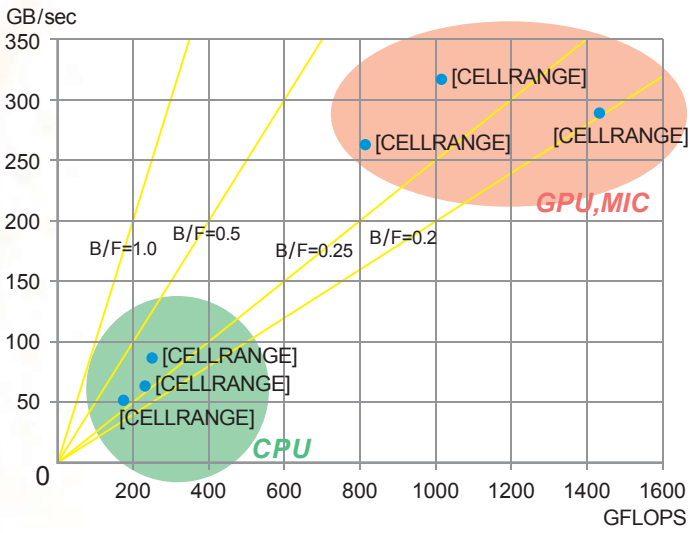
We propose an open source infrastructure for development and execution of optimized and reliable simulation code on large-scale parallel computers. We have named this infrastructure “ppOpen-HPC,” where “pp” stands for “post-peta.”

An auto-tuning (AT) capability is important and critical technology for further development of new architectures and maintenance of the overall framework. ppOpen-AT is an AT language for code optimization in five crucial numerical methods provided by the ppOpen-HPC project. The functions and software layers are shown in the figure below. New AT functions for the AT function in ppOpen-AT are summarized as follows. (1) Loop fusion (loop collapse) and loop fission functions for kernels of explicit methods; (2) re-ordering of sentences in loops; (3) an optimal code selection function between manycore and multicore CPUs; (4) a code generation function for the libraries of ppOpen-HPC. All AT functions are tested and evaluated with real code for ppOpen-HPC.

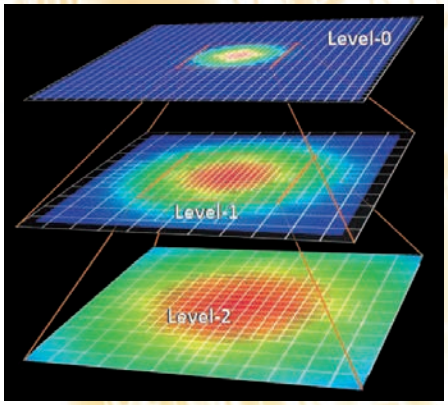


Parallel computing on current parallel processors

GPUs and many-core processors (MICs) are utilized in various HPC applications. In order to utilize these processors and obtain high performance, users have to use special programming environment and/or algorithms to make parallel programs. In the case of GPU, many HPC users use NVIDIA's GPU today and CUDA is required to obtain enough performance. CUDA programming is not very difficult but it is not easy to obtain high performance. The optimization techniques of CUDA are not similar to CPU fashion and users have to modify the source code dramatically. On the other hand, in the case of MIC, users can use same programming environment to CPU. However, optimal parameters and algorithms are not similar in many cases. Today, users can use OpenACC on CPU, GPU, and MIC. OpenACC provides uniform parallel programming environment to these processors. But, in order to obtain high performance, users have to consider the characteristics of target hardware and choose the appropriate implementation. In this research, we aim to develop high performance algorithms and implementation of scientific applications for various parallel processors. This activity contains improving performance of specific applications, and developing libraries and frameworks.



Adaptive Mesh Refinement Technique for ppOpen-HPC



An example of a computational domain with the adaptive mesh refinement technique.

We have developed an adaptive mesh refinement (AMR) technique for ppOpen-HPC applications. The demands of multi-scale and multi-physics simulations will be met with the advent of post-peta scale super computer systems. To achieve such simulations with a reasonable cost of computer resources, the spatial and temporal resolutions have to be adjusted locally and dynamically, depending on the local scales of physical phenomena. In the AMR code, computational grids with different spacing are dynamically created in hierarchical layers according to the local conditions of phenomena. Fine grids suitable to the local domain which need high resolution are applied only there, and other regions are simulated by using moderate size grids. Therefore, increments to the numerical cost due to the localized region are not serious if the AMR technique is adopted.

System, tools & hardware

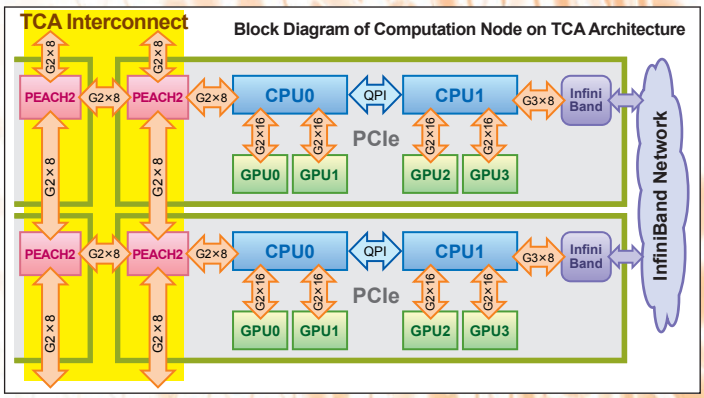
Tightly Coupled Accelerators(TCA)

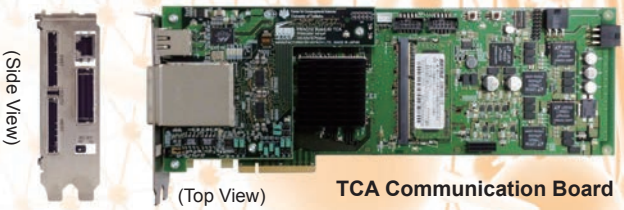
GPGPU is now widely used for accelerating scientific and engineering computing to improve performance significantly with less power consumption. However, I/O bandwidth bottleneck causes serious performance degradation on GPGPU computing. Especially, latency on inter-node GPU communication significantly 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. 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

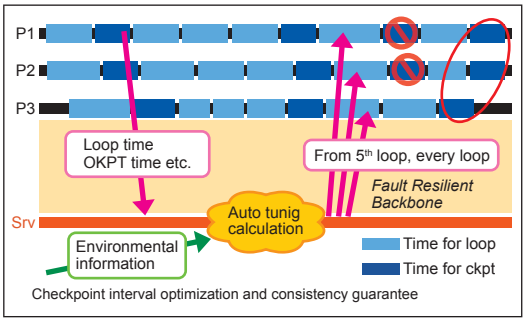
PEACH2 can transfer not only GPU memory but also host memory seamlessly since PEACH2 relies on the PCIe protocol. The DMA controller in the PEACH2 chip provides a chaining DMA function in order to transfer multiple data segments using the chained DMA descriptors automatically via hardwired logic, and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

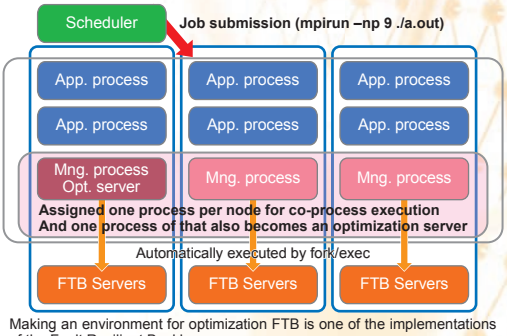




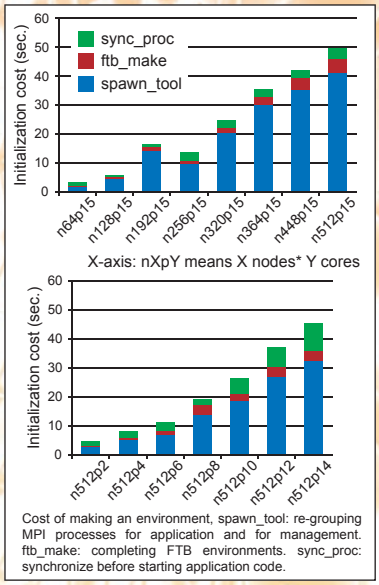
Fault Tolerance for Large-scale systems

This application level checkpoint is frequently implemented within an application that has time stepping. However the checkpoint interval tends to depend on the application programmer's ad hoc decisions. Essentially, the checkpoint interval is determined based on execution environment information such as the failure rate of hardware and the checkpoint time. We propose a directive-based application-level checkpoint/restart framework that includes optimizing the checkpoint interval automatically. The subject of this study is an application that has time stepping and utilizes the SPMD model. The optimization and renewal of the checkpoint interval are done asynchronously. A prototype implementation that includes cooperation with the job submitting system has been designed.





Making an environment for optimization FTB is one of the implementations of the Fault Resilient Backbone.

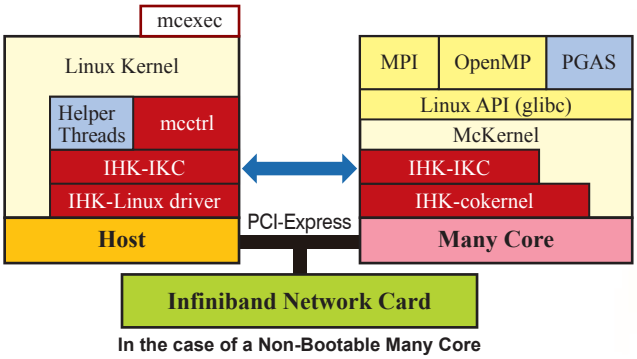


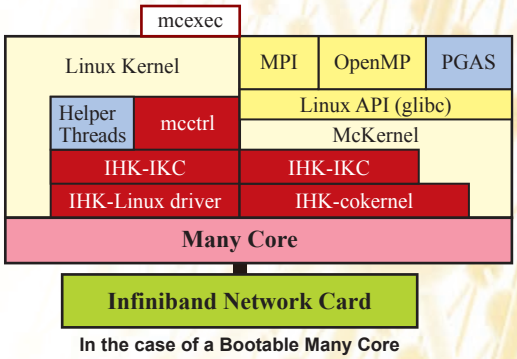
Cost of making an environment, spawn_tool: re-grouping MPI processes for application and for management. ftb_make: completing FTB environments. sync_proc: synchronize before starting application code.

Hetero Manycore Cluster

The Information Technology Center has been designing and developing a new scalable and cache-aware system software stack for many-core based supercomputers in cooperation with RIKEN AICS Hitachi, NEC, and Fujitsu. The many-core units are partitioned into two parts in which a Linux kernel and a light-weight micro kernel called a McKernel run. The GNU libc library for Linux and other Linux libraries run on the McKernel.

The McKernel provides basic OS services such as process/thread invocations, signaling, memory management, and inter-kernel communication. Other OS services such as file systems and networking are delegated to the Linux kernel. Thus, all applications running on a Linux kernel run on the McKernel without modification. The system has been implemented using the Intel Xeon Phi.





- IHK (interface for Heterogeneous Kernel)
  - Provides interface between Linux kernel and micro kernels
  - Provides generic-purpose communication and data transfer mechanisms
- McKernel
  - lightweight micro kernel

Supercomputers in SCD/ITC

HPCI: High Performance Computing Infrastructure

The HPCI is intended to be an environment enabling a user to easily use the “K” supercomputer and other top level computation resources in Japan. Also it is expected to match user’s needs and computation resource for accelerating an HPC scenario that includes exploratory research, large-scale research, and industrial use. The HPCI has eleven computational resource providers, of which are nine are universities and two are governmental research centers. And these resources suppliers are loosely connected with SINET-4, the high speed academic backbone network. SCD/ITC participates in this prioject as a hub resource provider in the Kanto region (HPCI EAST hub). Our resources include two cluster systems which are for the exclusive use of the HPCI, one cluster system which is shared with our business service for supercomputer operations (refer to Oakleaf-fx), two storage systems, one tape archive and few hosting nodes. Some of these resources are provided for constructing HPCI EAST hub which is core system on HPCI.
The HPCI EAST Hub consists of a PC cluster, a PC+GPU cluster, a 12PB storage system, a 5PB sub-storage system, and a tape archiver. Each cluster node has one Infiniband (4xQDR) interface, and is connected to the storage system via a large Infiniband switch (Voltare Grid Director 4700: MAX 648 ports). PC cluster has four Infiniband switches and they are connected to each other by one Infiniband. Every switch is connected to storage by two Infinibands. In addition the PC+GPU cluster is connected to storage by one Infiniband.

PC Cluster system

Node (87 nodes)	
CPUs (Cores)	2 (12)
Main Memory	96 GB
Processor	
Processor	Intel X5680 (3.33GHz)
Cache Memory	L2: 256KB/core L3: 12MB

PC+GPU Cluster system

Node (32 nodes)	
CPUs (Cores)	2 (12)
GPU (Cores)	1(448)
Main Memory	48 GB
Processor	Intel X5670 (2.93GHz)
GPU Card	
Model	Tesla M2050 (1.15GHz)
Memory	3GB GDDR5

12PB storage system

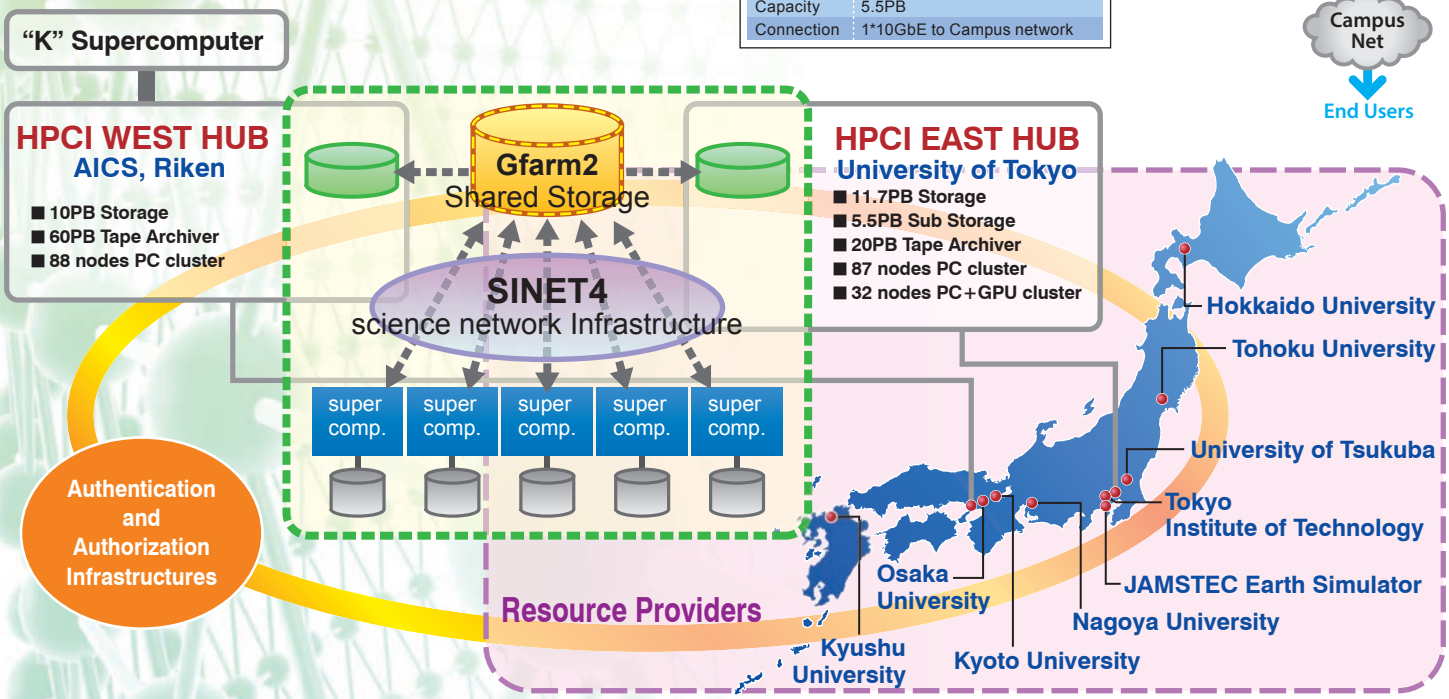
Capacity	11.7PB
Allocation	7.2PB for HPCI 3.6PB for Oakleaf-FX 0.9PB for JLDG
Connection	8*IB 4xQDR to PC cluster system 1*IB 4xQDR to PC+GPU cluster system 40*IB 4xQDR to Oakleaf-FX 2*10GbE to Campus Net

20PB Tape Archive

Capacity	20PB(MAX 25PB)
Connection	4*IB 4xQDR to 12PB Storage system via 4 head nodes 4*10GbE to Campus network via 4 head nodes

5PB Sub storage system

Capacity	5.5PB
Connection	1*10GbE to Campus network



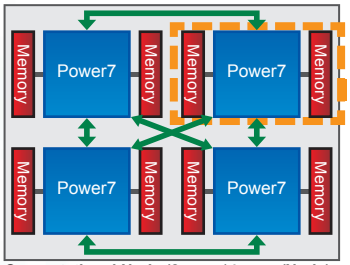
SR16000 M1 (Yayoi)

The SR16000 M1 (Yayoi) consists of Power7 computational nodes. Each node has four Power7 processors and 200GB of shared memory. Each of eight nodes are connected to each other via a fast network We began providing computing service (only for the “personal course”) in October, 2011 as the successor system of the SR11000. This system is expected to achieve research outcomes for many existing programs which require large shared memory.

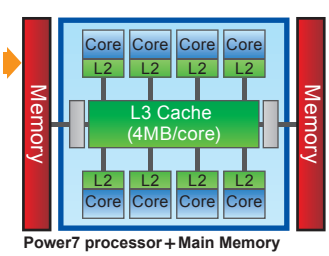

SR16000 M1 (Yayoi)

Entire	Theoretical Peak	54.906 TFLOPS
	Main Memory	11200 GB
Node	Theoretical Peak	11200 GB
	CPUs (Cores)	4 (32)
	Main Memory	11200 GB
Processor	Processor	IBM Power7 (3.83Hz)
	Cache Memory	L2: 512KB/Core L3: 2MB/Processor
	Theoretical Peak pre Core	30.64 GFLOPS

SR16000 (Yayoi) Specification



Computational Node (8cores*4procs/Node)



Power7 processor+Main Memory

Supercomputers in SCD/ITC

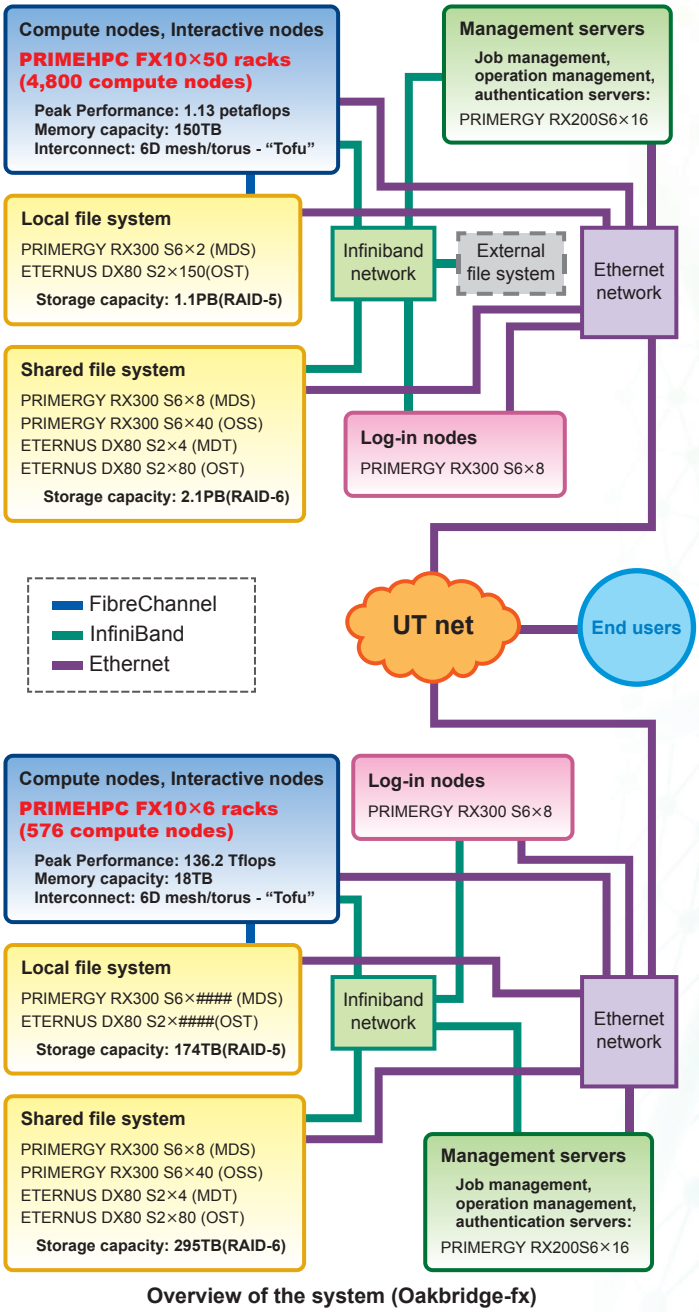
Oakleaf-fx (PRIMEHPC FX10)

Oakleaf-fx (PRIMEHPC FX10) is the first PFLOPS supercomputer system in SDC/ITC. We began computing service with it in April, 2012. The system has 4800 compute nodes with SPARC64IXfx CPUs and all nodes are connected by 6-Dimension Mesh/Torus Interconnects (Tofu). Well-balanced computational performance and power consumption are achieved. Because the architecture is compatible with the K computer, great contributions are expected for computer science in Japan.
We provide two kinds of computing services: a "personal course" for individual researchers, and a ”group course" for research groups. We also provide various types of special services such as services for educational purposes, services for young users, services for commercial users, and a program called the “large HPC challenge.”

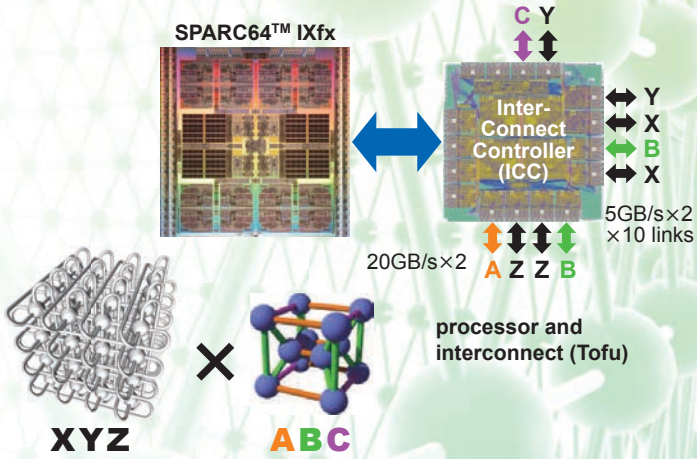
Entire	Theoretical Peak	1.13 PFLOPS
	Main Memory	150 TB
Node	Theoretical Peak	236.5 GFLOPS
	CPUs (Cores)	1 (16)
	Main Memory	32GB
Processor	Processor	Fujitsu SPARC64IXfx (1.848GHz)
	Cache Memory	L2: 12MB/Processor
	Theoretical Peak pre Core	14.8 GFLOPS

Oakleaf-fx (PRIMEHPC FX10) Specification

Overview of the system (Oakleaf-fx)



Overview of the system (Oakbridge-fx)



Oakbridge-fx (Fujitsu PRIMEHPC FX10)

The Oakbridge-fx is another PRIMEHPC FX10 system for long-time execution. The system has 576 nodes that is same architecture as Oakleaf-fx’s one. The users on Oakleaf-fx can also use this system with job class: long. This job class can use 24-576 nodes for up to 1week.

Entire	Theoretical Peak	136.2 TFLOPS
	Main Memory	18 TB
Node	Theoretical Peak	236.5 GFLOPS
	CPUs (Cores)	1 (16)
	Main Memory	32GB
Processor	Processor	Fujitsu SPARC64IXfx (1.848GHz)
	Cache Memory	L2: 12MB/Processor
	Theoretical Peak pre Core	14.8 GFLOPS

Oakbridge-fx (Fujitsu PRIMEHPC FX10) Specification

SCD/ITC

Supercomputing Division,
Information Technology Center,
The University of Tokyo

東京大学情報基盤センター
Information Technology Center, The University of Tokyo

Information Technology Center, The University of Tokyo
2-11-16 Yayoi, Bunkyo, Tokyo 113-8658, JAPAN
TEL:03-5841-2710 FAX:03-5841-2708(G3) http://www.itc.u-tokyo.ac.jp/index-e.html