JCAHPC

University of Tsukuba Center for Computational Sciences

INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

Oakforest PACS

Overview

JCAHPC (Joint Center for Advanced HPC), which is a cooperative organization by the University of Tokyo and University of Tsukuba for joint procurement and operation of the largest scale of supercomputer in Japan, introduced a new supercomputer system “Oakforest-PACS” with **25 PFLOPS** peak performance and started its operation from December 1st, 2016. The Oakforest-PACS system is ranked at #6 in TOP500 List of November 2016 with 13.55 PFLOPS of Linpack performance, and also recognized as Japan's fastest supercomputer. The system is installed at the Kashiwa Research Complex II building in the Kashiwa-no-Ha campus, the University of Tokyo.

The Oakforest-PACS system has 8,208 compute nodes, each of which consists of the latest version of Intel Xeon Phi processor (code name: Knights Landing), and Intel Omni-Path Architecture as the high performance interconnect. The Oakforest-PACS system is the largest cluster solution with Knights Landing processor as well as also the largest configuration with Omni-Path Architecture in the world. The system is integrated by Fujitsu Co. Ltd, and its PRIMERGY server is employed as each of compute node. Additionally, the system employs the Lustre shared files system (capacity: 26 PB), and IME (fast file cache system, 940 TB), both of which are provided by DataDirect Network (DDN).

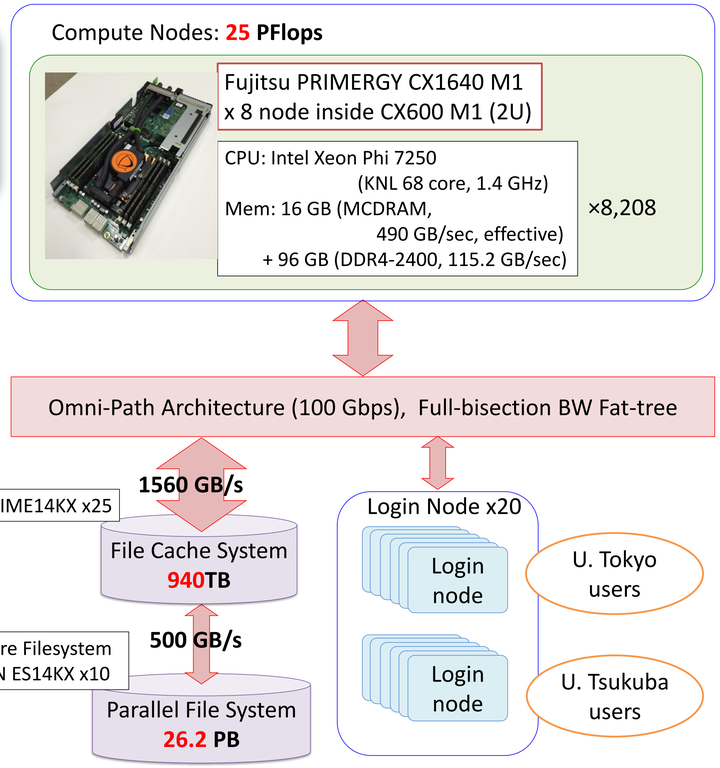
All the computation nodes and servers including login nodes, Lustre servers and IME servers are connected by a full bisection bandwidth of Fat-Tree interconnection network with Intel Omni-Path Architecture to provide highly flexible job allocation over the nodes and high performance file access.

TOP 500 #6 (#1 in Japan), HPCG #3 (#2), Green 500 #6 (#2) @Nov. 2016
IO 500 #1 @Nov. 2017, Jun. 2018
IO-500 BW #1 @Jun. 2019

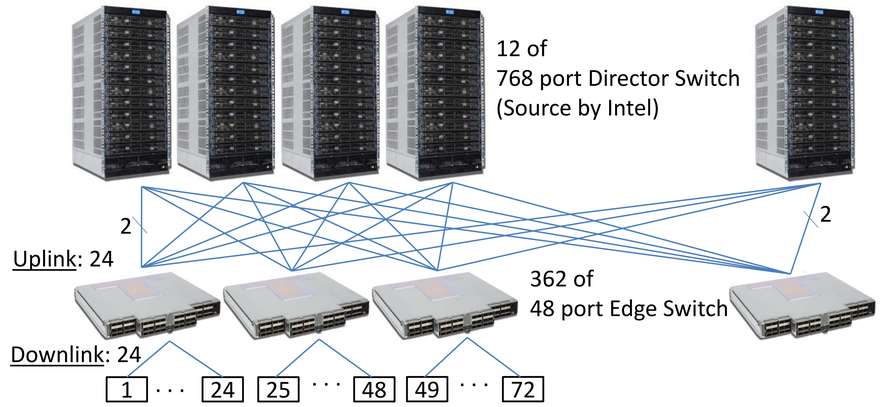
Research & Education

The Oakforest-PACS is offered to researchers in Japan and their international collaborators through various types of programs operated by HPCI under MEXT, and by original supercomputer resource sharing programs by two universities.

It is expected to contribute to dramatic development of new frontiers of various field of studies. The Oakforest-PACS will be also utilized for education and training of students and young researchers. We will continue to make further social contributions through operations of the Oakforest-PACS.



System Configuration

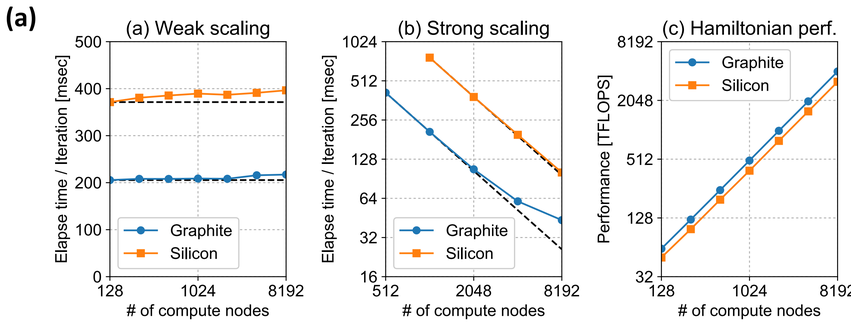


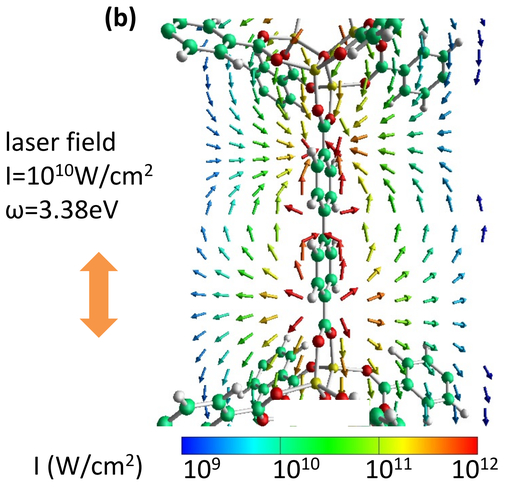
Total peak performance			25 PFLOPS
Total number of compute nodes			8,208
Power consumption			4.2 MW (including cooling)
# of racks			102
Cooling system	Compute Node	Type	Warm-water cooling Direct cooling (CPU) Rear door cooling (except CPU)
		Facility	Cooling tower & Chiller
	Others	Type	Air cooling
		Facility	PAC

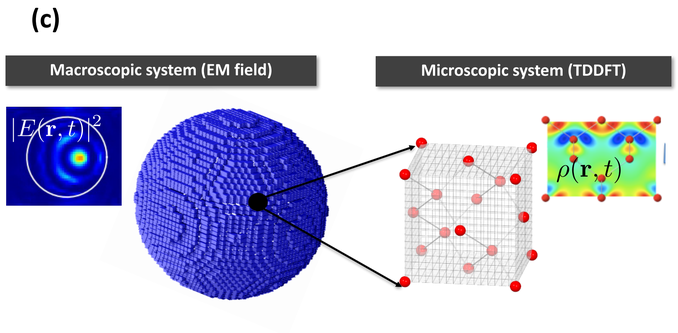
Applications with the Oakforest-PACS

SALMON: Scalable Ab-initio Light-Matter simulator for Optics and Nanoscience

We are developing a computer code SALMON, Scalable Ab-initio Light-Matter simulator for Optics and Nanoscience (http://salmon-tddft.jp). It is based on first-principles time-dependent density functional theory and describes electron dynamics in molecules, nanostructures, and solids induced by optical electric fields by solving the time-dependent Kohn-Sham equation in real time and real space. Recently, we have successfully achieved large-scale simulations for nano-optics phenomena solving a coupled equation of 3D Maxwell for light electromagnetic fields and 3D time-dependent Kohn-Sham for light-induced electron dynamics. It provides an accurate and precise platform of numerical experiments that will be indispensable in forefront optical sciences.





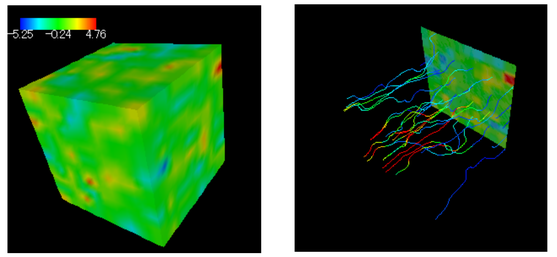


(a) Performance of SALMON using full nodes of Oakforest-PACS.
(b) Optical near-field generated in metal-organic framework, IRMOF10.
(c) Intense pulsed light irradiates on silicon nano-sphere of 250nm radius by coupled Maxwell-TDDFT calculation.

Parallel Multigrid Methods on the Oakforest-PACS System with IHK/McKernel

Parallel Multigrid Method

- Although the parallel multigrid method is expected to be one of the most powerful numerical algorithms in the exascale era due to its scalable features, there are many things to be done.
- Target Application: pGW3D-FVM for 3D Groundwater Flow through Heterogeneous Porous Media
  - MGCG Solver with IC(0) Smoother, V-Cycle
  - Sliced ELL Format for Storage of Sparse Matrices
- CGA (Coarse Grid Aggregation) [KN IEEE Cluster 2012]
  - Switching to coarse-grid solver earlier to avoid comm. Overhead
  - Significant overhead for coarse grid solver if number of MPI proc. is O(104)
- hCGA (Hierarchical CGA) [KN IEEE ICPADS 2014]
  - Hierarchical version of CGA (2-levels)
  - Number of MPI processes is reduced
  - Processes are repartitioned in an intermediate level before the final coarse grid solver on a single MPI process
  - Significant improvement of performance on Fujitsu PRIMEHPC FX10 with 4,096 nodes (65,536 cores) for a problem with 17+B DOF: Weak Scaling-1.61x, Strong Scaling 6.27x



Groundwater flow through heterogeneous porous media.
(Left) Distribution of water conductivity; (Right) Streamlines

AM-hCGA: Adaptive Multilevel hCGA

- If the number of MPI proc. is $O(10^6\text{-}10^7)$, number of proc. at the 2nd level of hCGA could be $O(10^4)$.
  - 2-Levels might not be enough for more processes
  - More levels are needed ? -> AM-hCGA (Adaptive Multilevel hCGA)

IHK/McKernel

- Light Weight Multi Kernel OS for HPC by RIKEN [BG IPDPS 2016]
  - McKernel implements only a small set of performance sensitive system calls and the rest of the OS services are delegated to Linux.
  - Same binary on pure Linux can be used
  - Lower Noise, Lower Communication Overhead
  - to be available on K, Fugaku (Post K)

Summary of Preliminary Results

- IHK/McKernel provided more **efficiency** and **stablility**
  - Reduction of Noise, Communication Overhead
- Improvement is more significant, if the prob. size/proc. is smaller
  - 4%: Medium, 17%: Small, 22%: Tiny
- Improvement by AM-hCGA = 10% for Tiny Case at 2,048 nodes with IHK/McKernel
  - Effect of AM-hCGA was not clear without IHK/McKernel, because fluctuation of computation time was very large.