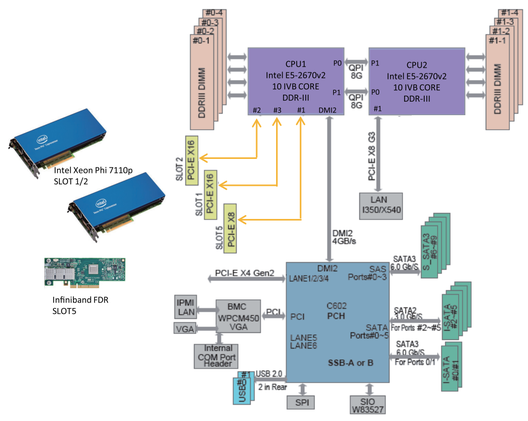
COMA (PACS-IX): Massively Parallel Many-Core Accelerated Cluster

System architecture and configuration

COMA (Cluster of Many-core Architecture processors) is the 9th generation of PACS/PAX series supercomputer in CCS, University of Tsukuba. The processor with many-core architecture is one of promised basic computation resource for next generation of HPC. Currently, Intel Xeon Phi is the most practical and productive co-processor, but many-core processor will be a strong main processor for HPC applications in very near future. Toward development of wide variety of scientific applications based on many-core concept, we introduced the largest many-core based PC cluster in Japan with Intel Xeon Phi co-processor.

COMA consists of 393 of high-end computation nodes, where two Intel Xeon Phi co-processors are equipped on each node as well as dual socket IvyBridge CPUs. All the nodes are connected by FDR InfiniBand network under Fat-Tree configuration with full bisection bandwidth by Mellanox HCA and switch.



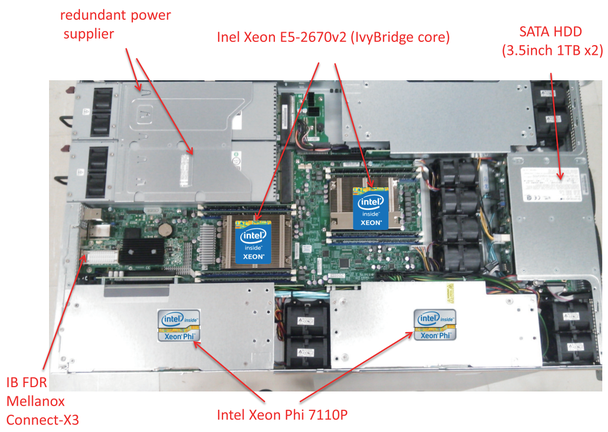
Block diagram of computation node of COMA

Item	Specification
Computation node	Cray CS300 Cluster Unit with two Xeon Phi
CPU	Intel E5-2670v2 (Ivy Bridge EP)
# of cores	10 cores/socket x 2 sockets = 20 cores/node
Clock	2.5 GHz
Peak performance	400 GFLOPS/node
PCI-express	generation 3 x 80 lanes (40 lanes/CPU)
Memory	64 GiB, DDR3 1866MHz, 4 channel/socket, 119 GB/s/node
MIC	Intel Xeon Phi 7110P
# of MICs/node	2
Peak performance	2.14 TFLOPS/node (1.07 TF/MIC)
Memory	16 GiB/node (8 GiB/MIC)
Interconnection	Infiniband FDR (Mellanox ConnectX-3)

Computation node of COMA

Item	Specification
Peak performance	1.001 PFLOPS (MIC: 843.9TF, CPU: 157.2TF)
# of nodes	393
File system	Lustre, 1.5 PB user area (DDN SFA12000)
Infiniband network switch	393 port FDR (Mellanox SX6536)
Total network bandwidth	2.75 TB/s
Language	Fortra90, C, C++
MPI	MVAPICH2, Intel MPI
System Management	Cray Advanced Cluster Engine, SLURM

System Specification



Chases inside of computation node

Node architecutture

A computation node of COMA is equipped with two set of Xeon Phi co-processor. These co-processors and InfiniBand FDR HCA require 40x lanes of PCIe gen-3. We decided to concentrate all of them to be attached to single CPU while there are two sockets of E5 CPUs. When Xeon Phi co-processor works in “native mode” where it runs its own Linux Operating System as like as ordinary Xeon processor, it accesses another Xeon Phi co-processor or InfiniBand HCA directly through PCIe switch inside Xeon CPU. It is known that PCIe device-to-device direct communication performance is drastically degraded through Intel QPI path, then all these devices are connected to one side of CPU in COMA.

JCAHPC plan and COMA

Currently COMA is dedicated to several supercomputer utilization programs to support nation-wide computational sciences.

University of Tsukuba and University of Tokyo plan to introduce a joint supercomputer system under a new organization named JCAHPC (Joint Center for Advanced High Performance Computing) which is located in Kashiwa Campus of University of Tokyo. We will introduce a new supercomputer based on many-core architecture to achieve 30 PFLOPS class of peak performance. COMA is dedicated to the application development and performance tuning as a testbed for this new machine.