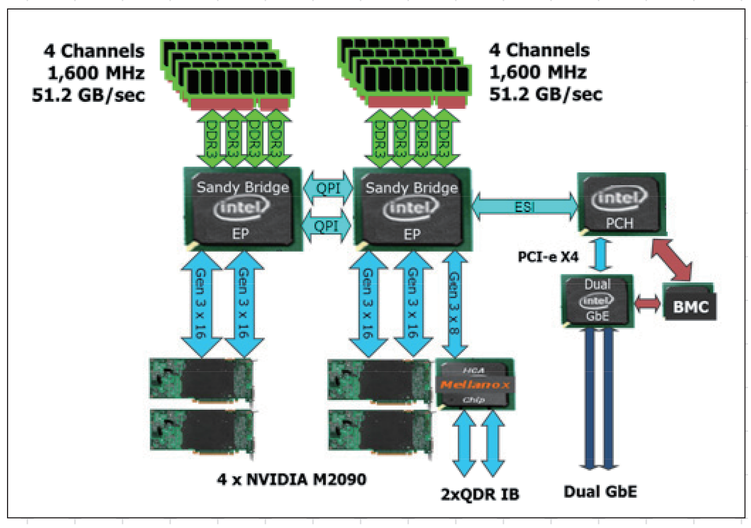
HA-PACS: Massively Parallel GPU Cluster for Accelerated HPC

System architecture and configuration

HA-PACS (Highly Accelerated Parallel Advanced Computer System for Computational Sciences) is the 8th generation of PACS/PAX series supercomputer in CCS, University of Tsukuba. For the development and product-run on cutting edge scientific computations toward next generation accelerated computing, it is equipped with the latest GPUs and CPUs connected by new generation of PCI-express to provide rich I/O bandwidth. On each computation node, two sockets of Intel Sandy Bridge-EP CPUs support full bandwidth connection of four NVIDIA M2090 GPUs without performance bottleneck. This is the first computation node in the world to support four GPUs with full bandwidth on PCIe bus without depending on PCIe switch to make a bottleneck.

Interconnection network employs dual-rail Infiniband QDR with a full bisection bandwidth Fat-Tree configuration to connect 268 of computation nodes. The system started the operation on February 2012 with 802 TFLOPS of peak performance and ranked at #41th on TOP500 list issued on June 2012.



Block diagram of computation node of HA-PACS

Item	Specification
Computation node	Appro Xtreme-X with four GPUs
CPU	Intel E5 (Sandy Bridge EP)
# of cores	8 cores/socket x 2 sockets = 16 cores/node
Clock	2.6 GHz
Peak performance	332.8 GFLOPS/node
PCI-express	generation 3 x 80 lanes (40 lanes/CPU)
Memory	128 GB, DDR3 1600MHz, 4 channel/socket, 102.8 GB/s/node
GPU	NVIDIA M2090
# of GPUs/node	4
Peak performance	2660 GFLOPS/node (665 GF/GPU)
Memory	24 GB/node (6 GB/GPU)
Interconnection	Infiniband QDR x 2 rails (Mellanox ConnectX-3 dual port)

Computation node of HA-PACS

Item	Specification
Peak performance	802 TFLOPS (GPU: 713TF, CPU: 89TF)
# of nodes	268
File system	Lustre, 504 TB user area (DDN SFA10000 ExaScaler)
Infiniband network switch	288 port QDR x 2 (Mellanox IS5300)
Total network bandwidth	2.14 TB/s
Language	Fortran90, C, C++
MPI	MVAPICH2, Intel MPI, OpenMPI
System Management	Appro Cluster Engine, SGE

System Specification

Two aspects of research on HA-PACS

Each computation node of HA-PACS employs two of advanced Intel Xeon processor providing 16 CPU cores and four of NVIDIA Fermi GPU with 2048 GPU cores in total. They are connected with x64 lanes of PCIe gen.2 technology for rich I/O bandwidth to support high performance GPU computing. This is an ideal platform for development and product-level running of large scale scientific and engineering codes and it is the main purpose of the system to develop new generation of accelerated HPC codes.

Another research topic of HA-PACS project is the development of new technology for GPU-to-GPU direct communication over nodes. This technology is named TCA (Tightly Coupled Accelerators) Architecture which is based on the idea to use PCIe link itself as the communication network link between computation nodes and GPUs rather than just as I/O bus within a node. We are developing a dedicated I/O card with FPGA to support four ports of PCIe interface and embedded multicore CPU for firmware. This feature provides ultra low latency direct communication among GPUs in the system to avoid the communication bottleneck in various parallel accelerated computing codes.

HA-PACS Project is supported by MEXT special fund as a program named “Research and Education on Interdisciplinary Computational Science Based on Exascale Computing Technology Develpment”.