Center for Computational Sciences, University of Tsukuba
www.ccs.tsukuba.ac.jp

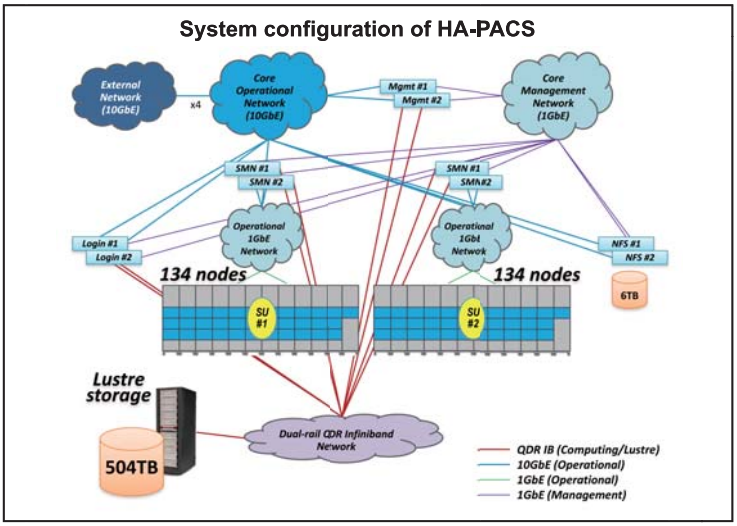
HA-PACS : High Bandwidth GPU Cluster for Computational Sciences

New Generation GPU Cluster with Rich I/O Bandwidth

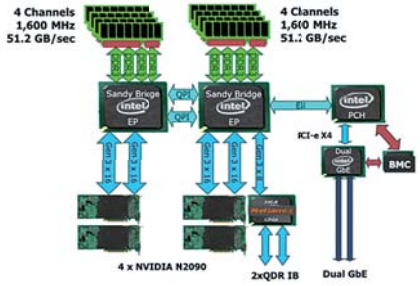
HA-PACS (Highly Accelerated Parallel Advanced System for Computational Sciences) is the 8th generation of PACS/PAX series supercomputer in CCS, University of Tsukuba. For the development and product-run on cutting edge scientific computations toward next generation accelerated computing, it is equipped with the latest GPUs and CPUs connected by new generation of PCI-express to provide rich I/O bandwidth. Two sockets of Intel Sandy Bridge-EP CPUs support full bandwidth connection of four NVIDIA M2090 GPUs without performance bottleneck. Interconnection network employs dual-rail Infiniband QDR with a full bisection bandwidth Fat-Tree configuration.

The system will be delivered on January 2012 with 802 TFLOPS of peak performance.

System configuration of HA-PACS



Block diagram of computation node of HA-PACS



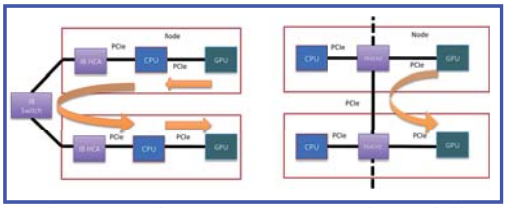
Computation node of HA-PACS

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 head)

System Specification

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	Fortra90, C, C++
MPI	MVAPICH2, Intel MPI, OpenMPI
System Management	Appro Cluster Engine, SGE

Tightly Coupled Accelerators: True GPU-Direct Technology



Ordinary inter-GPU communication over Infiniband (left) and True GPU-direct by TCA (right)

PEACH2 test board with FPGA and PCIe gen.2 external connector

To solve the I/O bandwidth bottleneck causing serious performance degradation on multi-GPU on computation node, we have been developing a new technology named TCA (Tightly Coupled Accelerators), which enables direct communication among multiple GPUs over computation nodes. PCIe extended connection is used as an interconnection medium, and intelligent PEACH2 (PCI-Express Adaptive Communication Hub ver.2) chip controls the device-to-device direct communication without through CPU memory. This technology greatly reduces the latency on parallel GPU applications. We call this feature as True GPU-Direct.

PEACH2 is developed as an enhanced version of PEACH/PEARL (PCI-Express Adaptive and Reliable Link) solution originally developed in JST-CREST research project “Computation Platform for Power-aware and Reliable Embedded Parallel Processing System” in the research area of “Dependable Embedded Operating Systems for Practical Use” . While the prototype of PEACH chip is implemented as a custom LSI, new PEACH2 chip is implemented on FPGA for flexible programming on communication control.

Any PCIe-ready device can be directly attached to PEARL with PEACH2 chip, which enables low-latency direct device access for next generation of accelerated computing.