HA-PACS/TCA: Tightly Coupled Accelerators for Low-Latency Communication

Overview of Tightly Coupled Accelerators (TCA) Architecture and HA-PACS/TCA

GPGPU is now widely used for accelerating scientific and engineering computing to improve performance significantly with less power consumption.

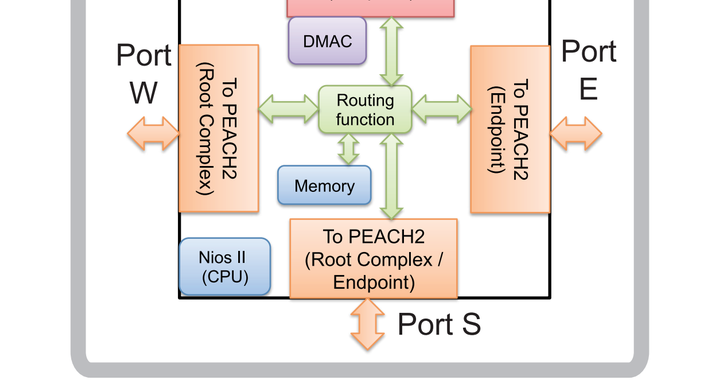
However, I/O bandwidth bottleneck causes serious performance degradation on GPGPU computing. Especially, latency on inter-node GPU communication significantly increases by several memory copies. To solve this problem, TCA (Tightly Coupled Accelerators) enables direct communication among multiple GPUs 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
- Shared PCI Express address space among multiple nodes
  - Ease to program

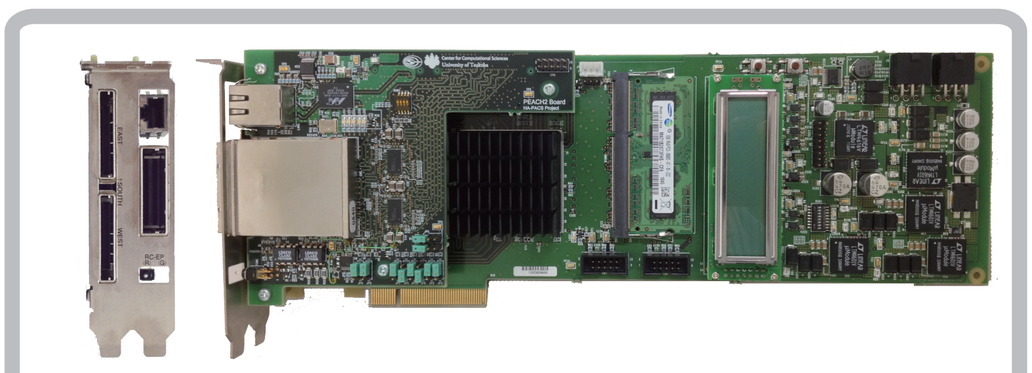
HA-PACS/TCA, which is an extended part of HA-PACS base cluster, was installed with PEACH2 board in each node on Oct. 2013. HA-PACS/TCA is operated with HA-PACS base cluster, and entire HA-PACS system becomes over 1.1 PFLOPS GPU cluster.



Block diagram of computation node of HA-PACS/TCA

Block diagram of PEACH2 Chip

Entire HA-PACS System Including HA-PACS/TCA (5 racks x 2 rows)



(Front View) (Top View)

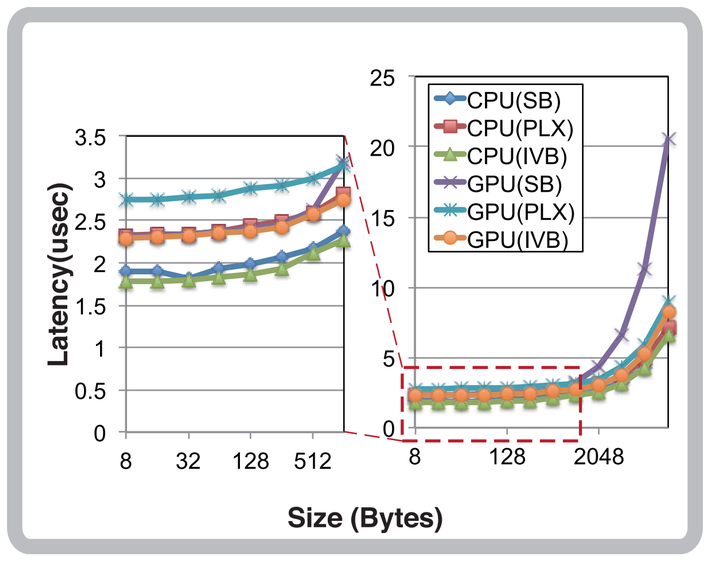
TCA Communication Board (PCIe CEM Spec., double height)

HA-PACS/TCA Specification

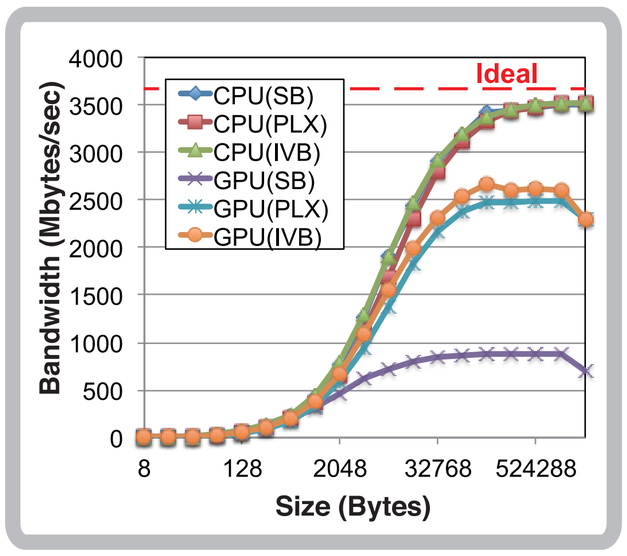
Computation Node	CRAY 3623G4-SM
Motherboard	SuperMicro X9DRG-QF
CPU	Intel Xeon E5 2680 v2 (Ivy Bridge 2.8 GHz, 10 core) x 2 socket
Memory	DDR3-1866MHz 4ch. x2 128 GB (119.4 GB/s)
Peak Performance	448 GFLOPS/node
GPU	NVIDIA Tesla K20X x 4 GPU
Memory	GDDR5 2600MHz, 6 GB/GPU (250 GB/s/GPU)
Peak Performance	5.24 TFLOPS/node
Interconnect	IB QDR x 2 rails (Mellanox Connect X-3)
TCA Interconnect	PEACH2 (FPGA: Altera Stratix IV 530GX)
# of Nodes	64
Peak Performance	364 TFLOPS (CPU: 28.7 TF, GPU: 335.3 TF)

Basic Performance of TCA Communication

CPU: CPU-to-CPU neighbor communication, **GPU**: GPU-to-GPU neighbor communication
SB: SandyBridge, **PLX**: PCIe Switch, **IVB**: IvyBridge (HA-PACS/TCA)



Ping-pong Latency using DMA



Ping-pong Bandwidth using DMA

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 (FY2011-2013)".