

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

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

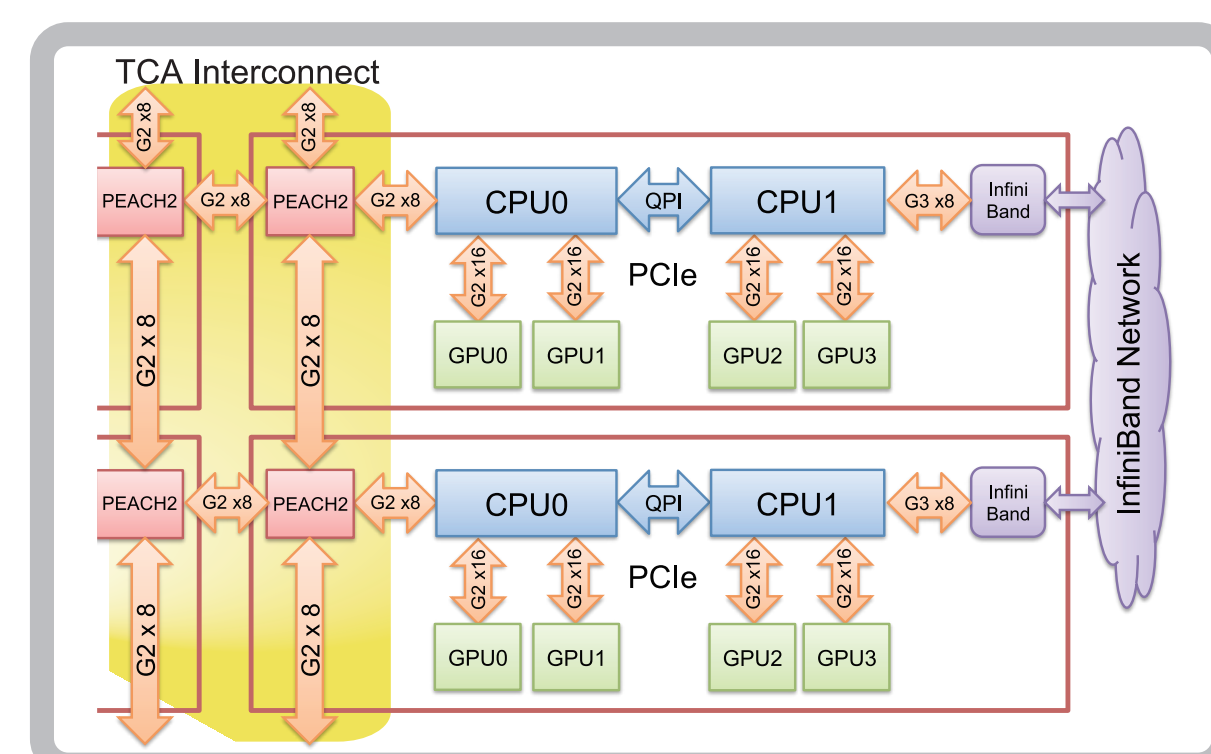
GPGPU is now widely used for accelerating scientific and engineering computing to improve performance significantly with less power consumption.

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.

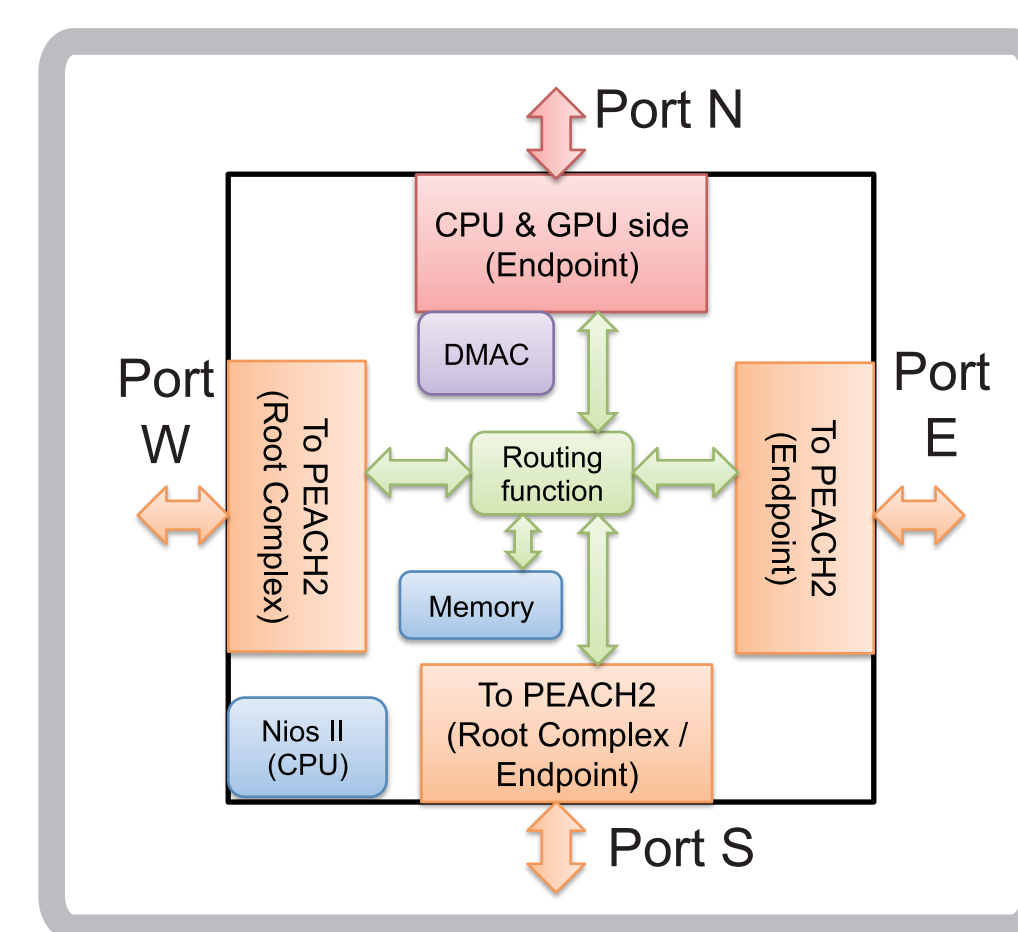
PEACH2 (PCI Express Adaptive Communication Hub ver. 2) chip is developed and implemented by FPGA (Field Programmable Gate Array) for flexible control and prototyping. PEACH2 board is also developed as an PCI Express extension board.

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.

Currently we develop programming APIs for CUDA programmer, evaluate several benchmarks and applications, such as QUDA (Lattice QCD library), Himeno benchmark (3D Poisson Solver), HPCG etc. We also have a poster in technical session, which shows block-stride transfer performance and Himeno benchmark results.



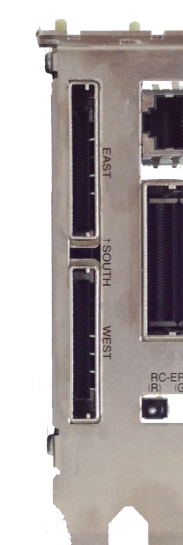
Block diagram of computation node of HA-PACS/TCA



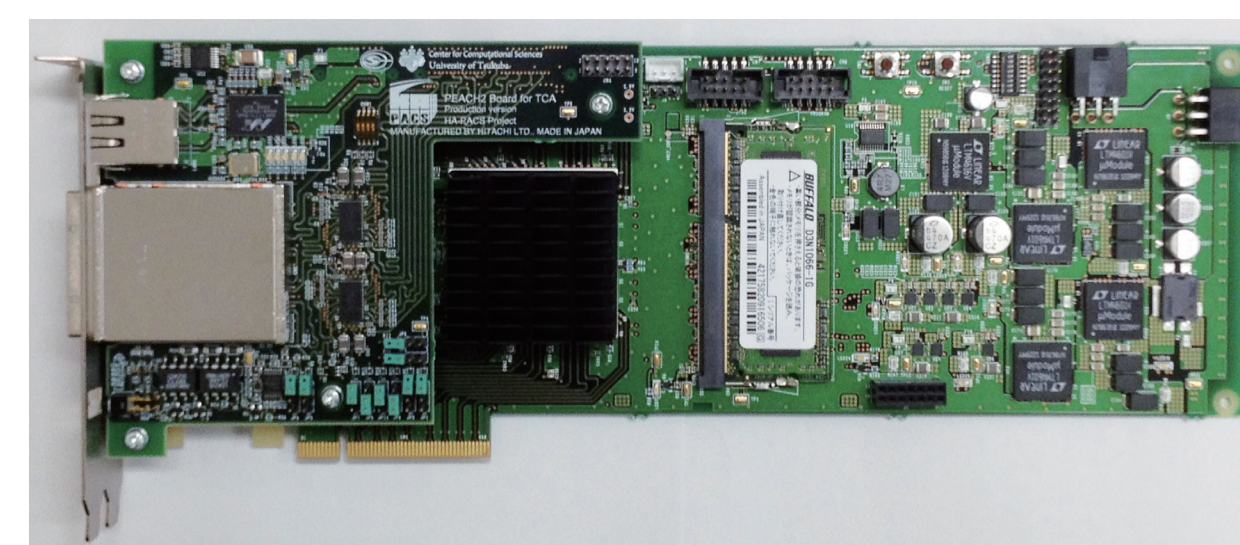
Block diagram of PEACH2 Chip



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



(Side View)



(Top View)

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

HA-PACS/TCA Specification

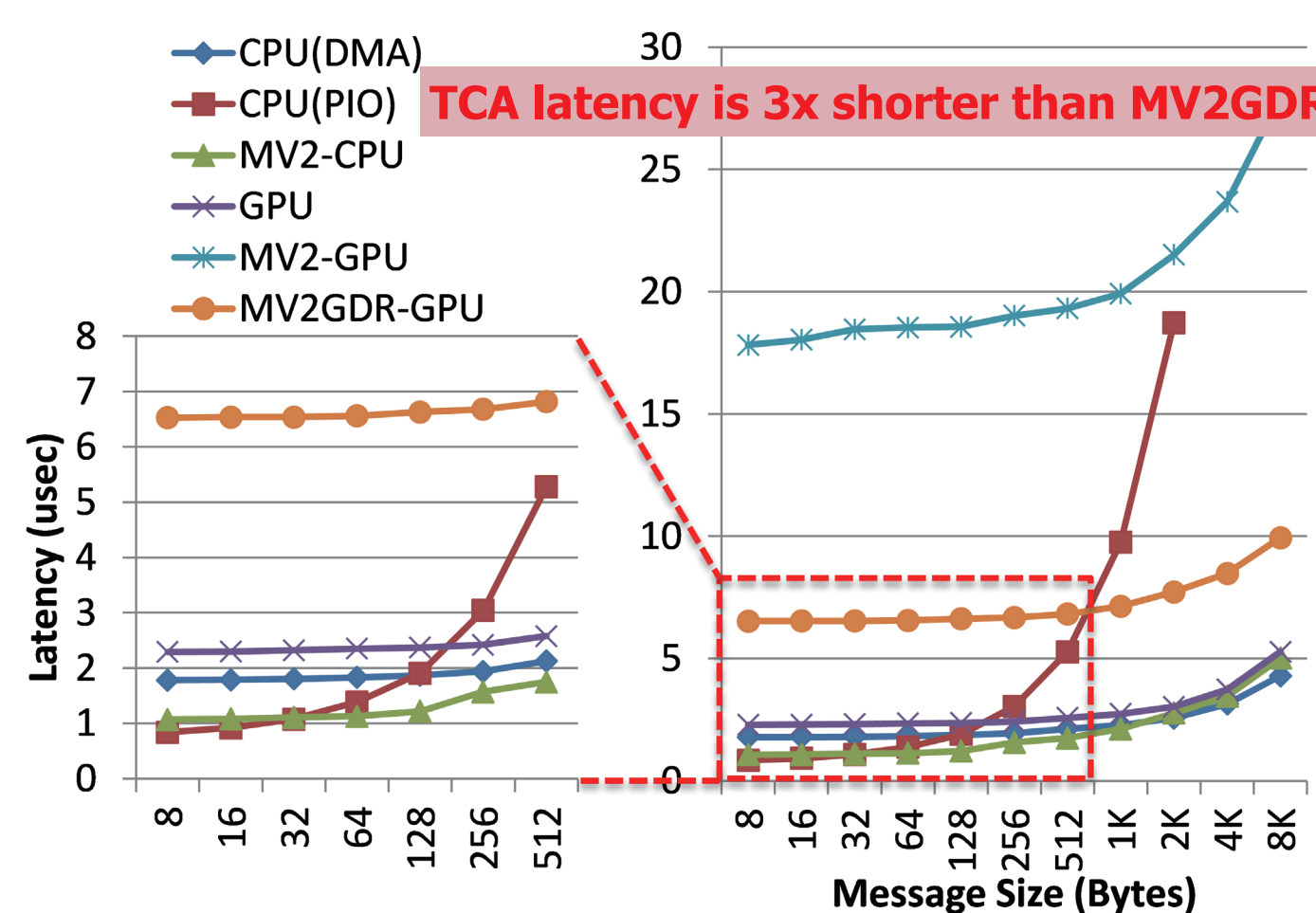
Computation Node	CRAY 3623G4-SM
Motherboard	SuperMicro X9DRG-QF
CPU	Intel Xeon E5 2680 v2 (Ivy Bridge 2.8 GHz, 10 core) x 2 socket
Memory	DDR3-1866MHz 4ch. 128 GB (119.4 GB/s)
Peak Performance	448 GFLOPS/node
GPU	NVIDIA Tesla K20X x 4 GPU
Memory	GDDR5 2600MHz, 6 GB/GPU (250 GB/s/GPU)
Peak Performance	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)
LINPACK Benchmark	277 TFLOPS (Efficiency: 76%) 3.52 GFLOPS/W (3 rd Nov. 2013 Green500)

Basic Performance of TCA Communication

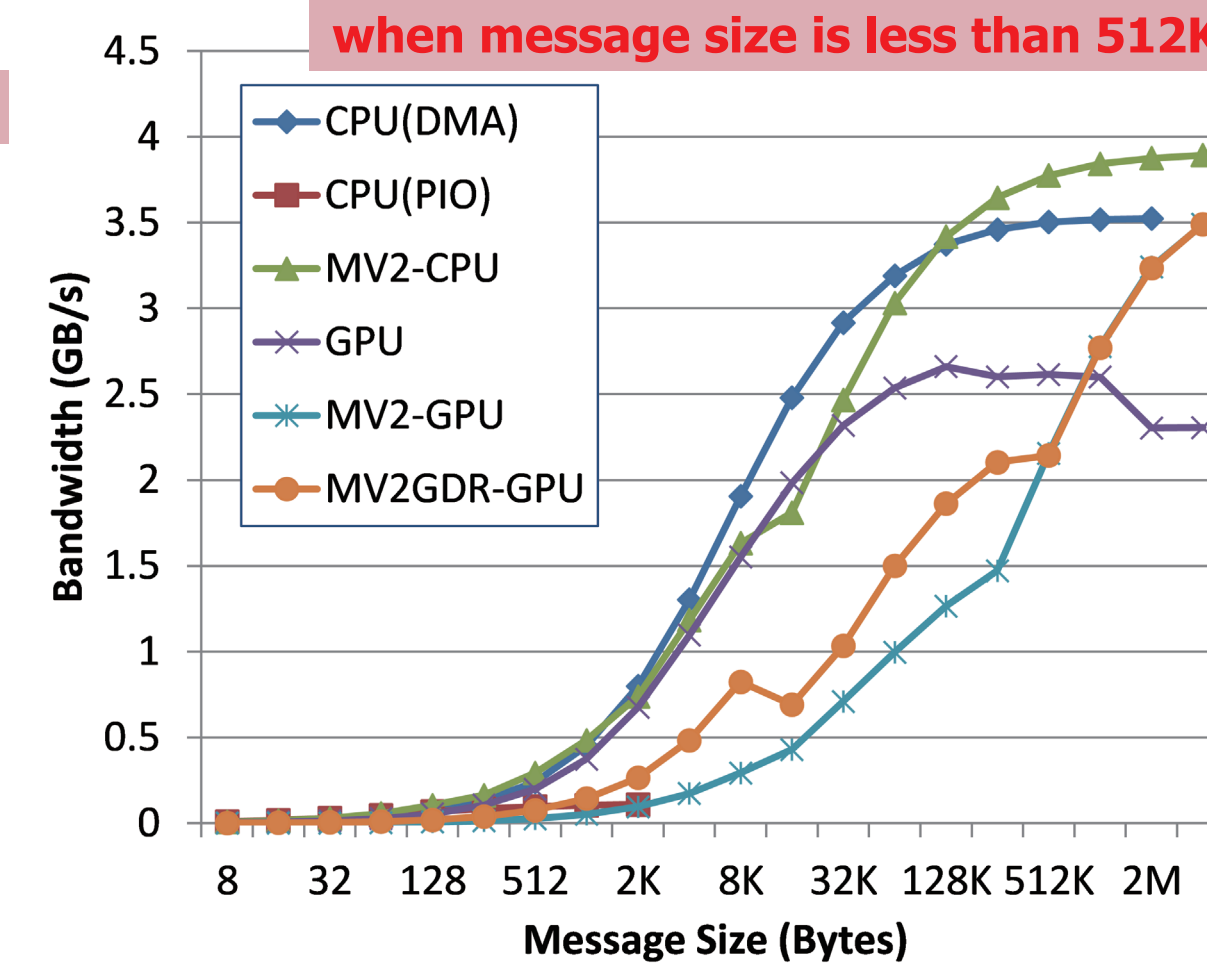
CPU: CPU-to-CPU neighbor communication, GPU: GPU-to-GPU neighbor communication

MV2: MVAPICH2, MV2GDR: with GDR on IB QDRx1

TCA bandwidth is better than MV2GDR when message size is less than 512K



Ping-pong Latency using DMA



Ping-pong Bandwidth using DMA