[1] The University of Tokyo | [2] Keio University | Collaboration with U. of Tsukuba

Evaluation of Graph Application using Tightly Coupled Accelerators

Toshihiro Hanawa[1], Takahiro Kaneda[2], Hideharu Amano[2]

Tightly Coupled Accelerators (TCA) Architecture, PEACH2, and PEACH3

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 was developed and it has been evaluated using HA-PACS/TCA cluster, which employs PEACH2 board in each node . However, PEACH2 performance was limited by PCI Express Gen2 x8.

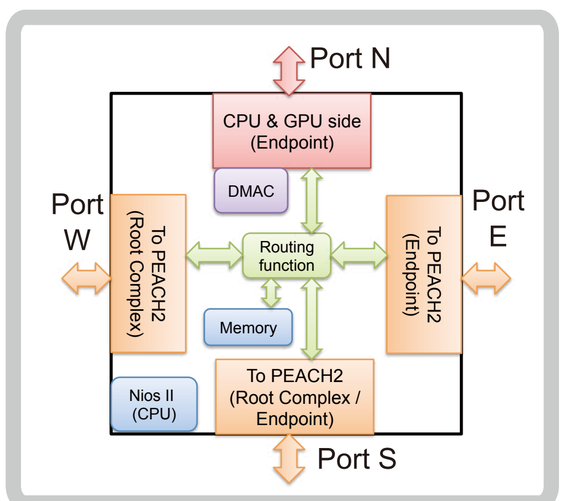
In order to improve the PCI Express performance, we introduce the new FPGA which supports PCI Express Gen3 with the hard IP. We have designed and implemented new hub chip named "PEACH3." PEACH3 board has also been developed as an PCI Express extension board similar to PEACH2 board. PEACH3 basic bandwidth attains twice better than PEACH2 bandwidth,

Evaluation Environment

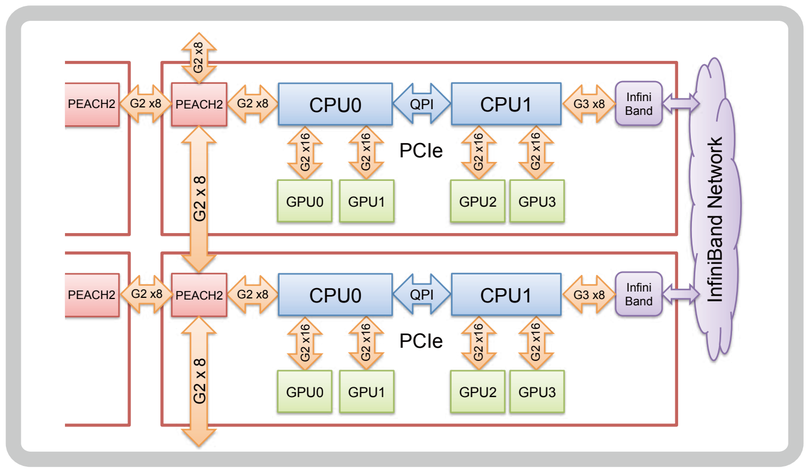
	HA-PACS/TCA	PEACH3 Test Env.
CPU	Intel Xeon CPU E5-2680 v2	
GPU	NVIDIA Tesla K20x (Gen2 x16)	NVIDIA Tesla K40m (Gen3 x16)
# of Nodes	2 of 64	2
CUDA	7.0	
MPI	MVAPICH2-GDR/2.1rc2	MVAPICH2-GDR/2.1
InfiniBand	Infiniband QDR 2-rail	(Infiniband FDR 1-rail)
PEACH	PEACH2	PEACH3

PEACH3 Specification vs. PEACH2 Specification

	PEACH2	PEACH3
FPGA Family	Altera Stratix IV GX	Altera Stratix V GX
FPGA Chip	EP4SGX530NF45C2	ES5GXA7N3F45C2
Process Technology	40nm	28nm
Available LEs	531K	622K
Port	PCIe Gen2 x8	PCIe Gen3 x8
Maximum Bandwidth	4 Gbyte/sec	7.9 Gbyte/sec
Operation Frequency	250 MHz	250 MHz
Internal Bus Width	128 bit	256 bit
Usage of LEs	22%	38%
DRAM on Board (Available)	DDR3 512 Mbyte	DDR3 512 Mbyte



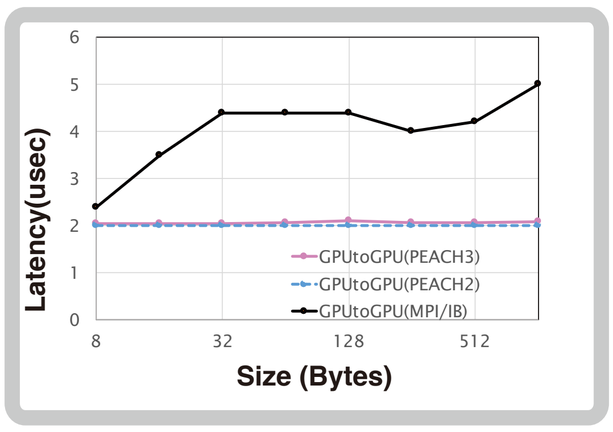
Block diagram of PEACH2/3 Chip (Port S is omitted on PEACH3 Board)



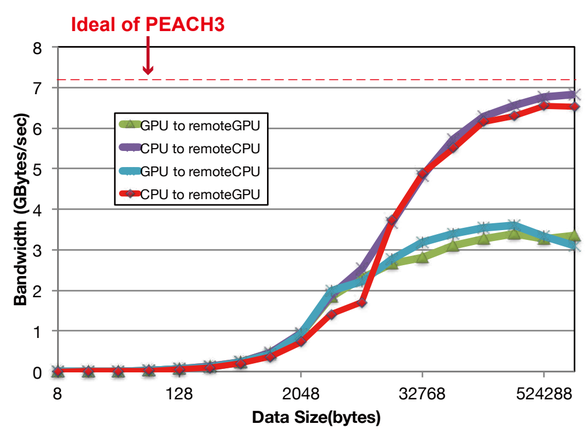
Block diagram of computation node of HA-PACS/TCA (same for PEACH3)

PEACH3 Communication Board (PCIe CEM Spec., single height)

Basic Performance of PEACH2 and PEACH3



Ping-pong Latency using DMA

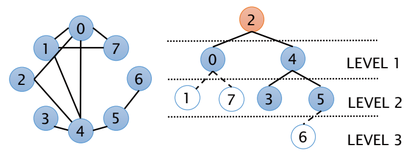


Ping-pong Bandwidth using DMA

Implementation and Evaluation of BFS using PEACH3

Level-Synchronized BFS

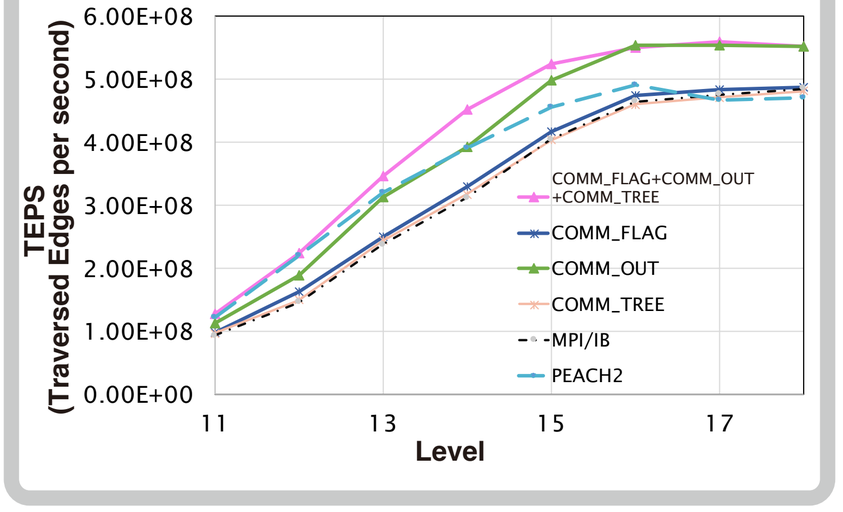
Each thread works on the same depth (**level**) in synchronization



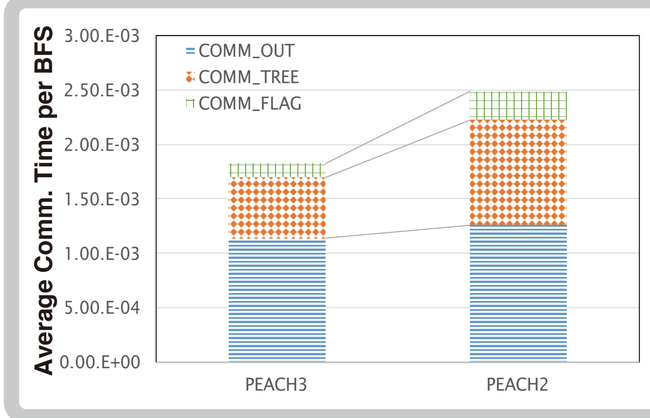
Communications on BFS

Type of Comm.	Message Size (Byte)	Count / 1 BFS	Detail of comm.
COMM_FLAG	4	4 to 7	Flag whether an adjacent vertex is found or not
COMM_OUT	$2^{scale} * 4 / \|P\|$	4 to 7	Vertex information on each level
COMM_TREE	$2^{scale} * 8 / \|P\|$	1	Final results

scale: level (2^{scale} = # of vertices), P: # of GPUs



TEPS on Graph500 application in comparison with MPI/IB, PEACH2, and PEACH3



Break down of Communication Time using PEACH2 and PEACH3 with Level-18

HA-PACS Project was supported by MEXT special fund as a program named "Research and Education on Interdisciplinary Computational Science Based on Exascale Computing Technology Develpment (FY2011-2013)" in U. of Tsukuba, and by the JST/CREST program entitled "Research and Development on Unified Environment of Accelerated Computing and Interconnection for Post-Petascale Era" in the research area of "Development of System Software Technologies for post-Peta Scale High Performance Computing."