Tightly Coupled Accelerator for HA-PACS

Yuetsu KODAMA

Center for Computational Sciences

University of Tsukuba

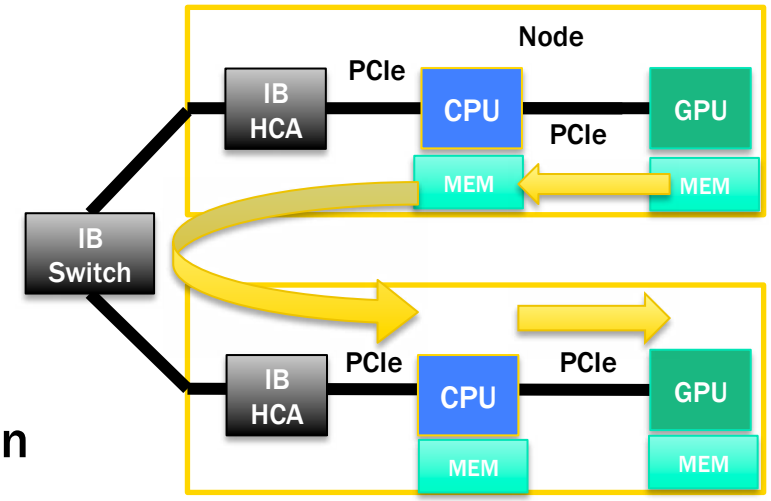
kodama@cs.tsukuba.ac.jp

HA-PACS (Highly Accelerated Parallel Advanced system for Computational Sciences)

- **Two parts**
  - **Base Cluster**:
    - for development of GPU-accelerated code for target fields, and performing product-run of them
  - **TCA**: (**Tightly Coupled Accelerators**)
    - for elementary research on new communication technology for reducing latency between accelerators.
    - Our original communication system based on PCI-Express named “PEARL”, and a prototype communication chip named “PEACH”
    - Improve PEACH for TCA : PEACH2


Problems of GPU Cluster

- **Problems of GPGPU for HPC**
  - Data I/O performance limitation
    - Ex) GPGPU: PCIe gen2 x16 (8GB/s)
      ⇔ 665 GFLOPS (NVIDIA M2090)
  - Memory size limitation
    - Ex) M2090: 6GByte vs CPU: 128GByte
  - Communication between accelerators:
    no direct path (external) ⇒ communication latency via CPU becomes large



- **Researches for direct communication between GPUs are required**

The target of TCA is developing a direct communication system between external GPUs for a feasibility study for future accelerated computing

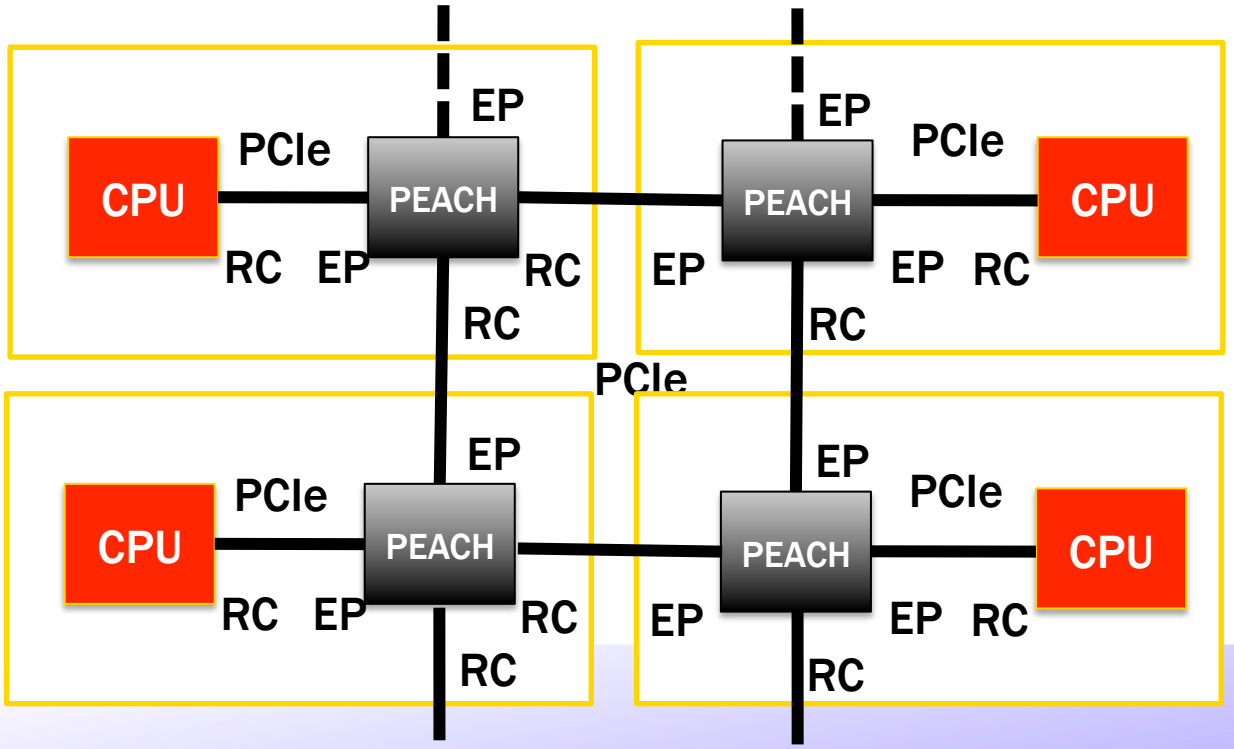
HA-PACS: TCA (Tightly Coupled Accelerator)

- **TCA: Tightly Coupled Accelerator**
  - Direct connection between accelerators (GPUs)
  - Using PCIe as a communication device between accelerator
    - Most acceleration device and other I/O device are connected by PCIe as PCIe end-point (slave device)
    - An intelligent PCIe device logically enables an end-point device to directly communicate with other end-point devices
- **PEARL: PCI Express Adaptive and Reliable Link**
  - We already developed such PCIe device (PEACH, PCI Express Adaptive Communication Hub) on JST-CREST project "low power and dependable network for embedded system"

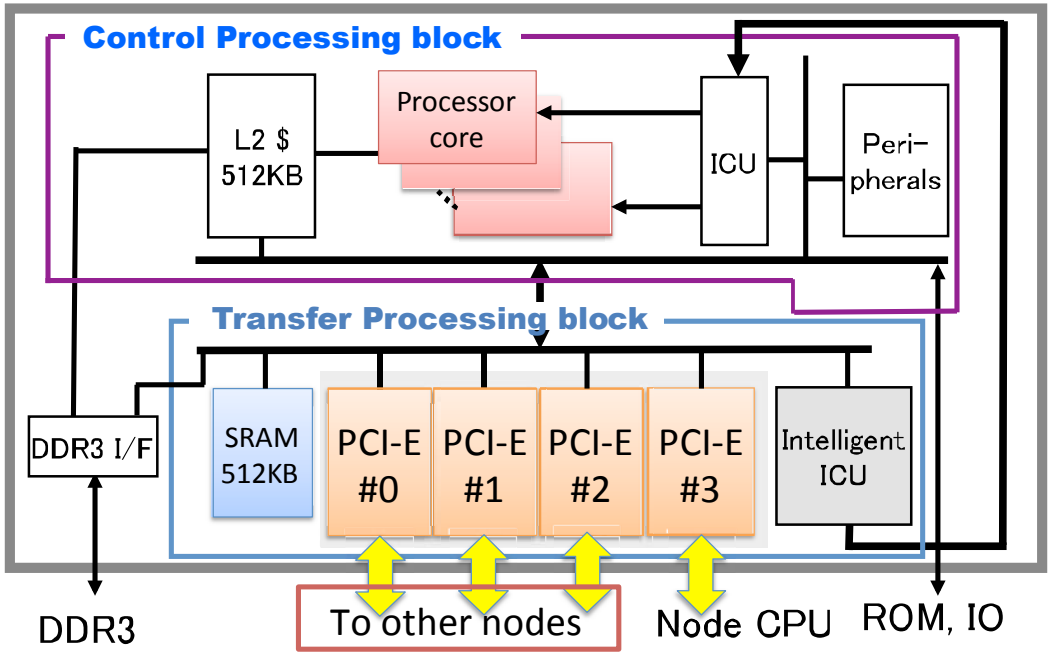
⇒ Improving PEACH for HPC to realize TCA : PEACH2

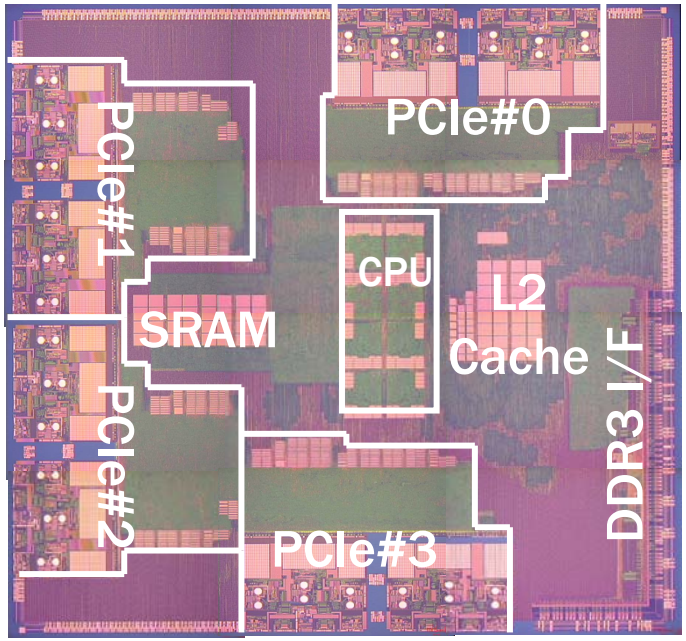

PEACH, PCI Express Adaptive Communication Hub

- Applying PCIe link as “a direct link between nodes”
- PCI-E link edge control feature: “root complex” and “end points” are automatically switched (flipped) according to the connection handling
- Each PCIe is an independent bus, and Intelligent controller on PEACH transfers each packet between PCIe buses.



PEACH chip [Otani et al., ISSCC2011]



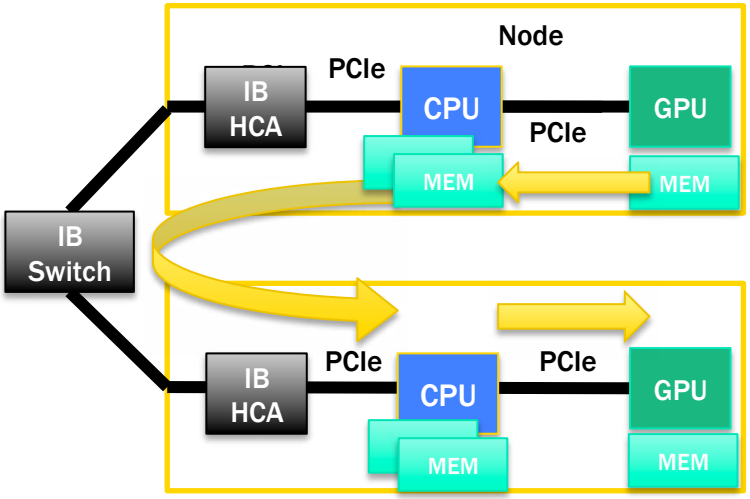


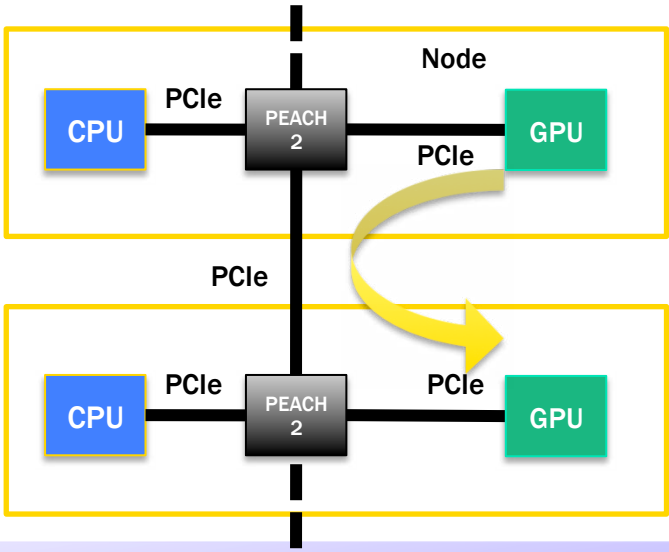
- **CPU: Renesas M32R 4core SMP (max. 400MHz)**
  - small core size, low power
  - SMP
  - Controlling PCIe 4 port
  - health check for the node, link
  - communication link management
  - route reconfiguration

- comm. link:
  PCI Express Gen2 x4 lanes (20Gbps)
  x4 port
- SuperHyway, DMAC
- Low power
  - Controlling #lanes for each port
  - gen1/gen2 switching
  - core frequency control

HA-PACS/TCA (Tightly Coupled Accelerator)

- **True GPU-direct**
  - current GPU clusters require 3-hop communication (3-5 times memory copy)
  - For strong scaling, Inter-GPU direct communication protocol is needed for lower latency and higher throughput
- Enhanced version of PEACH
  ⇒ ***PEACH2***
  - x4 lanes -> x8 lanes
  - hardwired on main data path and PCIe interface fabric



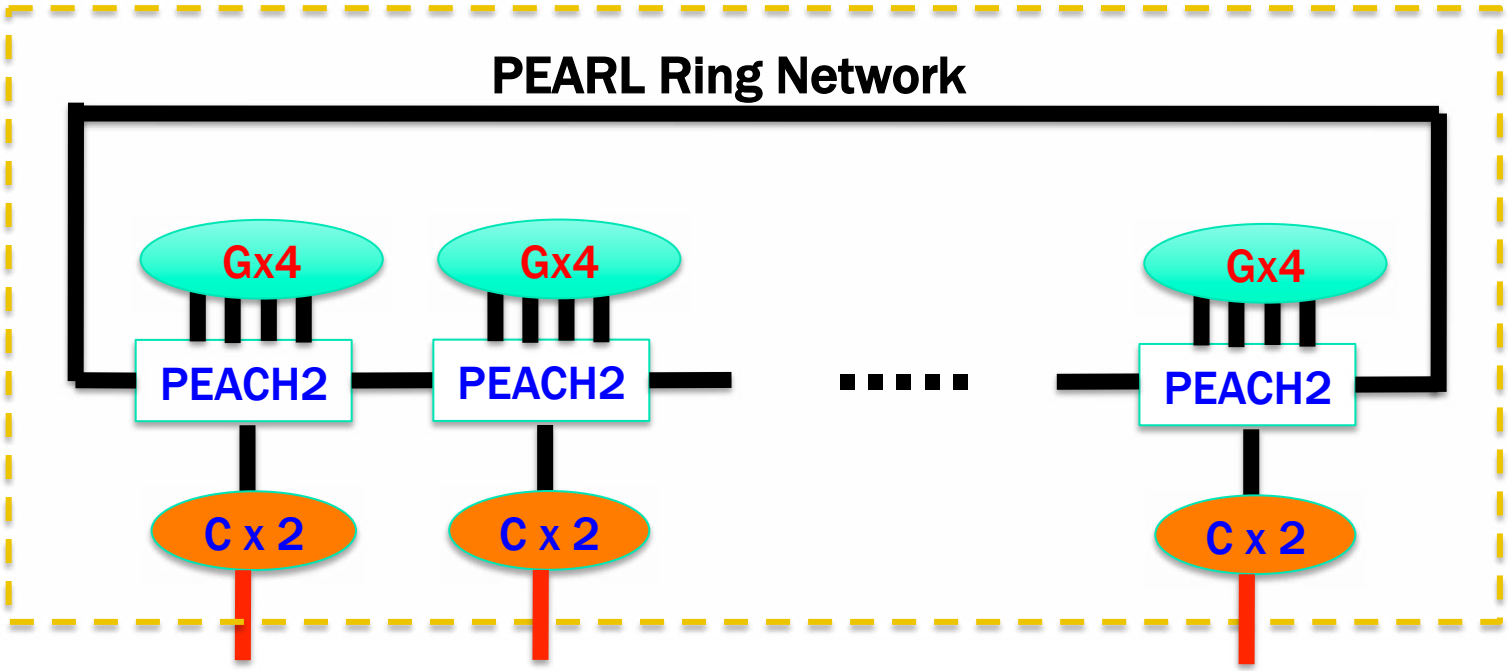


Implementation of PEACH2: ASIC⇒FPGA

- **FPGA based implementation**
  - today's advanced FPGA allows to use PCIe hub with multiple ports
  - currently gen2 x 8 lanes x 4 ports are available
    ⇒ soon gen3 will be available (?)
  - easy modification and enhancement
  - fits to standard (full-size) PCIe board
  - internal multi-core general purpose CPU with programmability is available
    ⇒ easily split hardwired/firmware partitioning on certain level on control layer
- **Controlling PEACH2 for GPU communication protocol**
  - collaboration with NVIDIA for information sharing and discussion
  - based on CUDA4.0 device to device direct memory copy protocol


HA-PACS/TCA

Node Cluster = NC

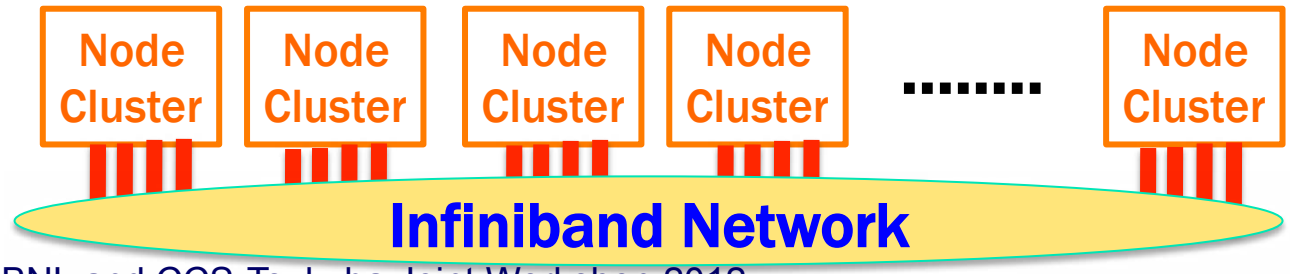


Node Cluster with 16 nodes
- GPUx64 (G)
- CPUx32 (C)
- GPU comm with PCIe
- IB link / node
- CPU: Xeon E5
- GPU: Kepler

Infiniband Link

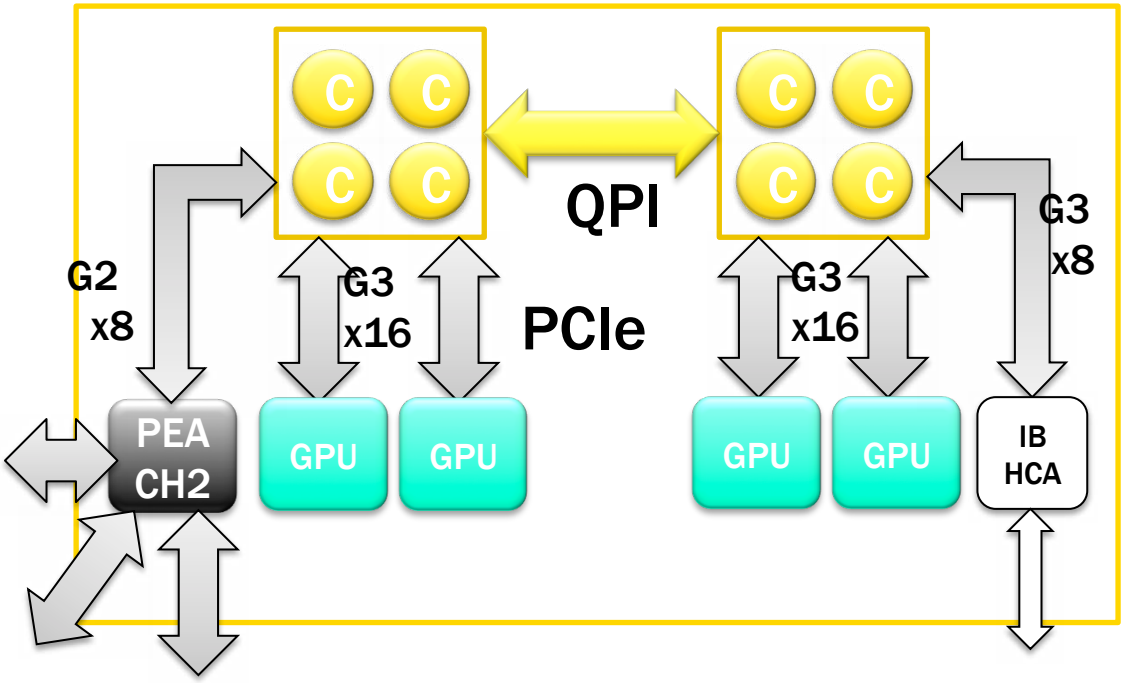
- High speed GPU-GPU comm. by PEACH within NC (PCI-E gen2x8 = 4GB/s/link)
- Infiniband QDR (x2) for NC-NC comm. (4GB/s/link)

4 NC with 16 nodes, or 8 NC with 8 nodes = 360 TFLOPS extension to base cluster



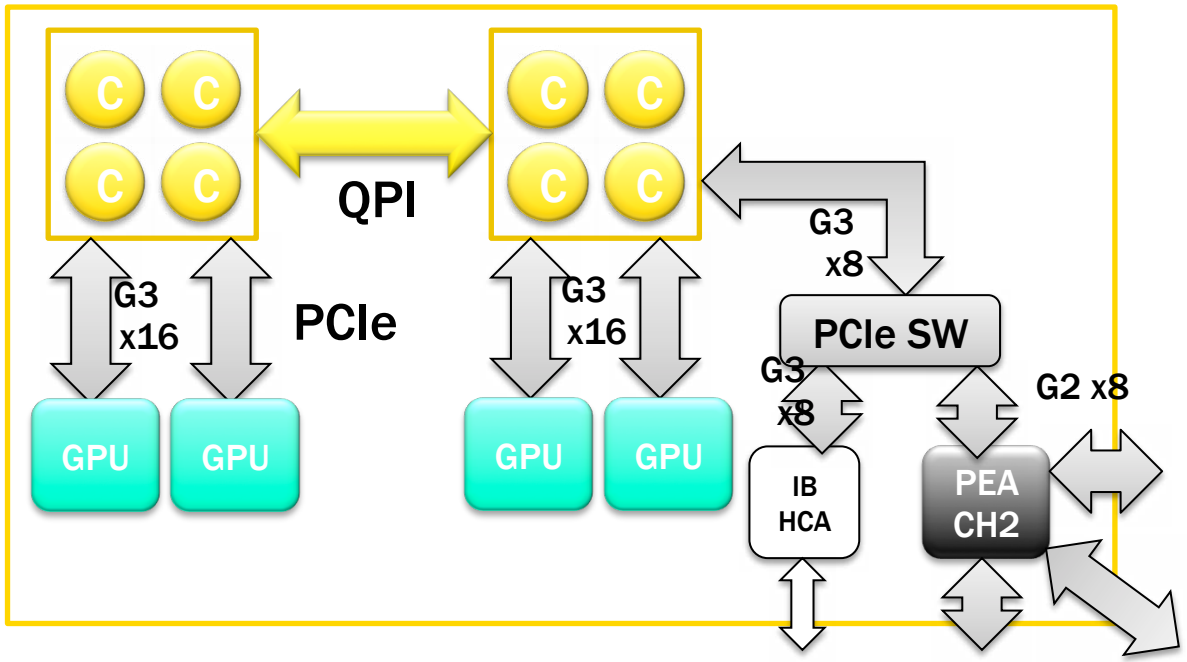
PEARL/PEACH2 variation (1)

- **Option 1:**
  - GPU host driver is available as-is
  - 4 GPUs are equivalent from PEACH2



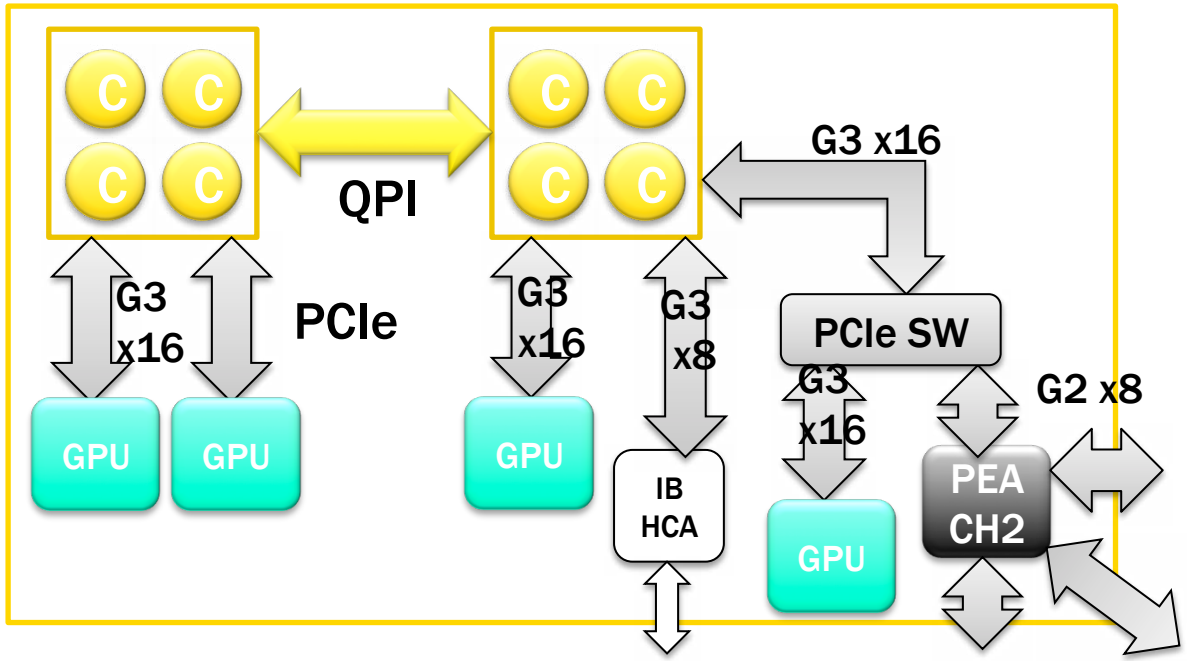
PEARL/PEACH2 variation (2)

- **Option 1:**
  - **Performance comparison among IB and PEARL can be evenly compared**
  - **Additional latency by PCIe switch**



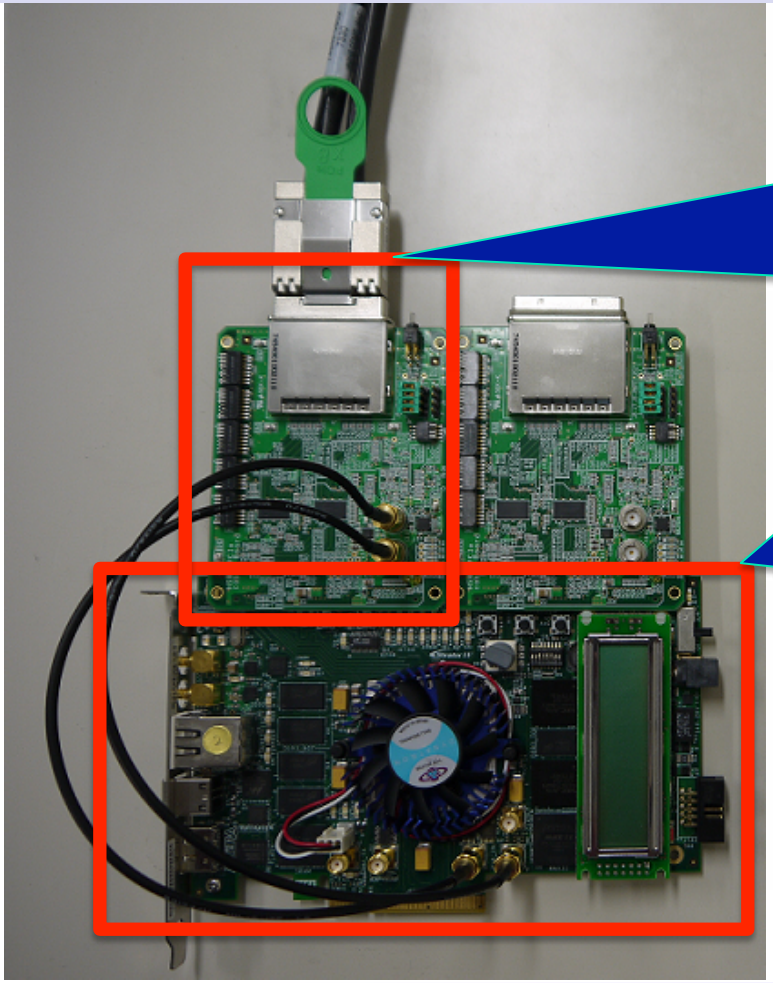
PEARL/PEACH2 variation (3)

- **Option 3:**
  - **Requires only 72 lanes in total**
  - **asymmetric connection among 3 blocks of GPUs**



PEACH2 FPGA test bed

Evaluate PCI-Express hard IP on FPGA



HSMC-PCIe converter board (newly developed)

- HSMC (High-Speed Mezzanine Card) General I/O port
- PCIe x8 cable connector
- RootComplex / Endpoint switchable
- both x4 / x8 available

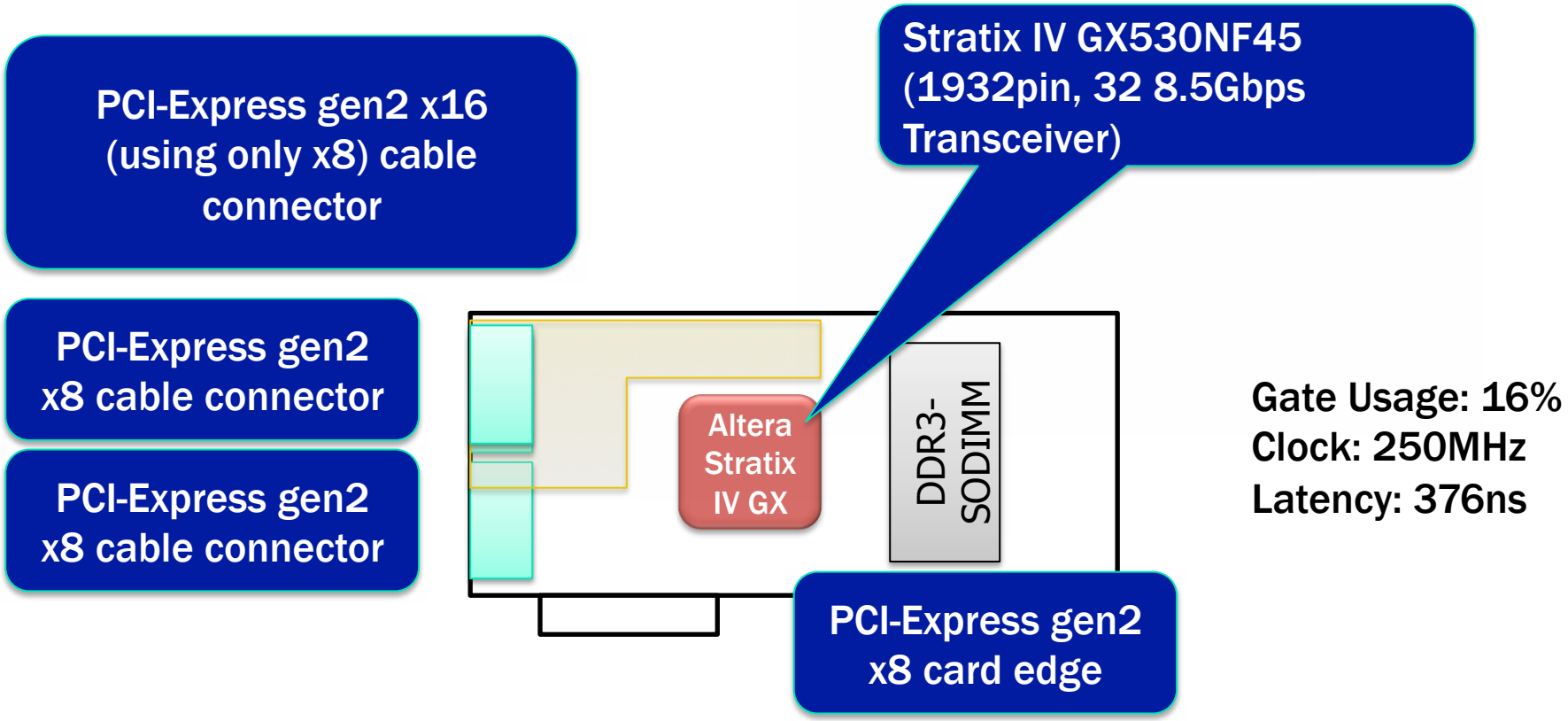
generic FPGA evaluation board (DEV-4SGX530N)

- PCIe x8 endpoint
- HSMC support PCIe x 8 Gen2
- HSMC support PCIe x 4 Gen2
- Stratix IV GX530KH40 (1517pin, 531K LE, 20Mbit, 4 PCIe IP, 24 8.5Gbps Transceiver)

Read: 3.2GB/s, Write 3.3GB/s in 64K

PEACH2 PCIe compatible board(~ Mar. 2012)

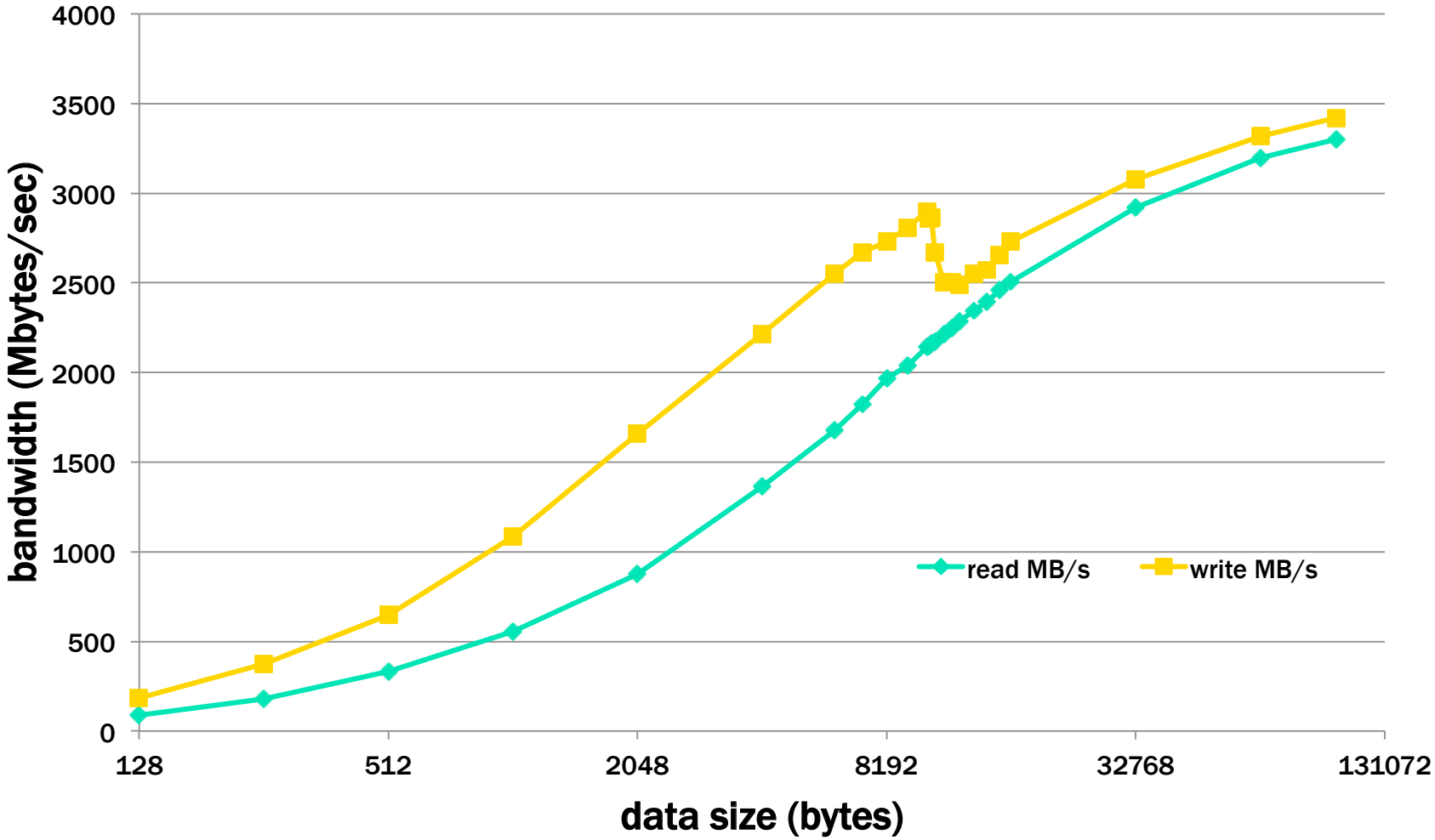
Real PCIe board for HA-APCS/TCA (and for any other platform)



Summary

- HA-PACS/TCA for elementary study for advanced technology on direct communication among accelerating devices (GPUs)
- FPGA implementation of PEACH2 will be finished for the prototype version on Mar. 2012 and enhanced for final version in following 6 months
- HA-PACS/TCA with at least 200 TFLOPS additional performance will be installed around Mar. 2013


PCI express IP core performance (Gen2 x 8)



PCI Express

- I/O Interface for PC after PCI, PCI-X
  - Standard by PCI-SIG
  - Used in connecton between PC and NIC
- Serial communication
  - Speed of 1 lane: 2.5Gbps(Gen1), 5.0Gbps(Gen2)
  - Multilane is supported for high throughput
    - x1, x2, x4, x8, x16
  - Line length is limited
    - 30cm for Gen2 on board
    - 2m with external cable
    - Enough for Embeded or small scale cluster