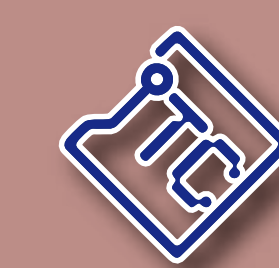


# Tightly Coupled Accelerators Architecture for Low-latency Inter-node Communication between Accelerators

Toshihiro Hanawa Yuetsu Kodama Taisuke Boku Mitsuhsa Sato



Information Technology Center



THE UNIVERSITY OF TOKYO



Center for Computational Sciences



University of Tsukuba

## 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.

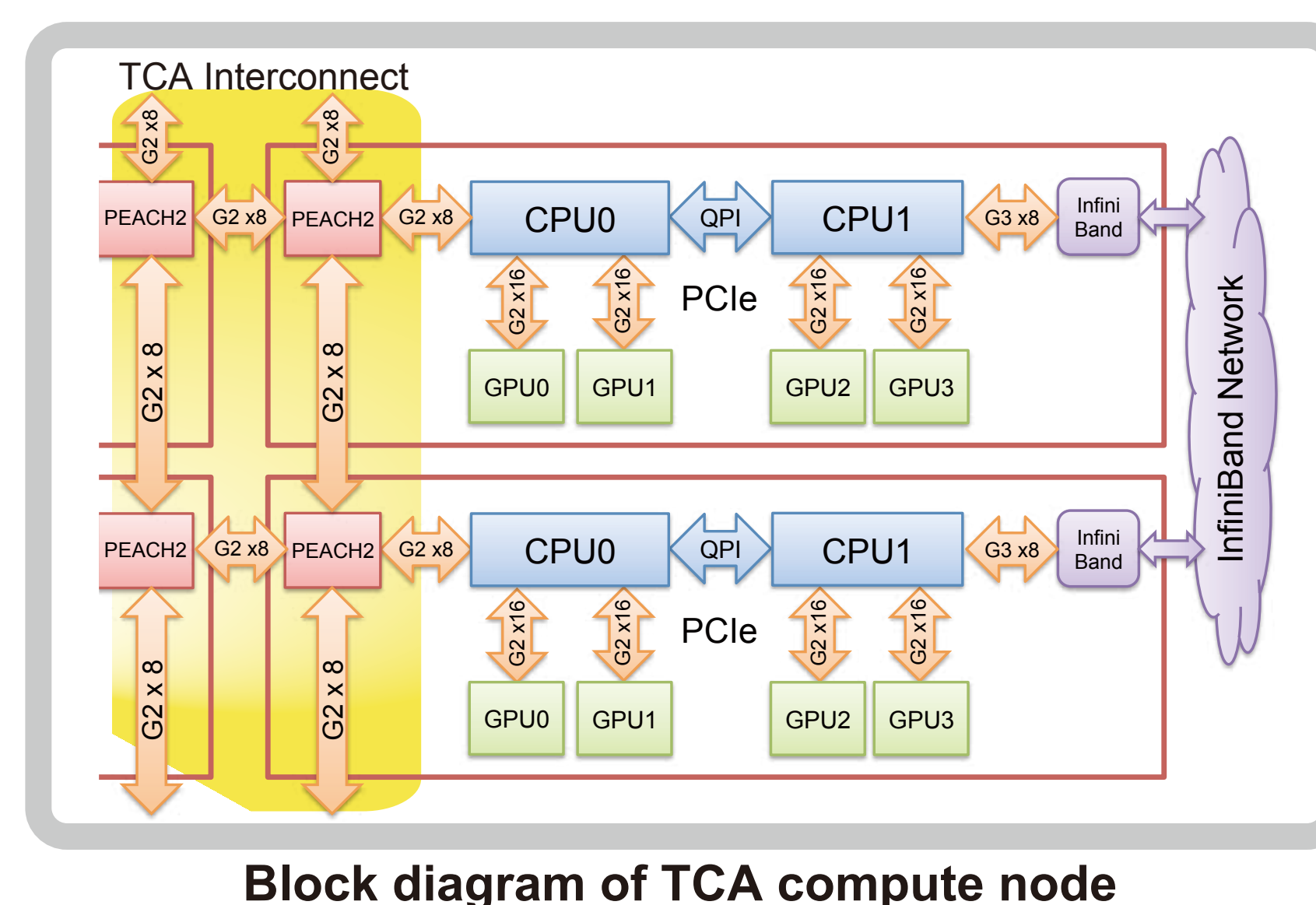
However, I/O bandwidth bottleneck causes serious performance degradation on GPGPU computing. Especially, latency on inter-node GPU communication significantly increases by several memory copies. To solve this problem, **TCA (Tightly Coupled Accelerators)** enables direct communication among multiple 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.

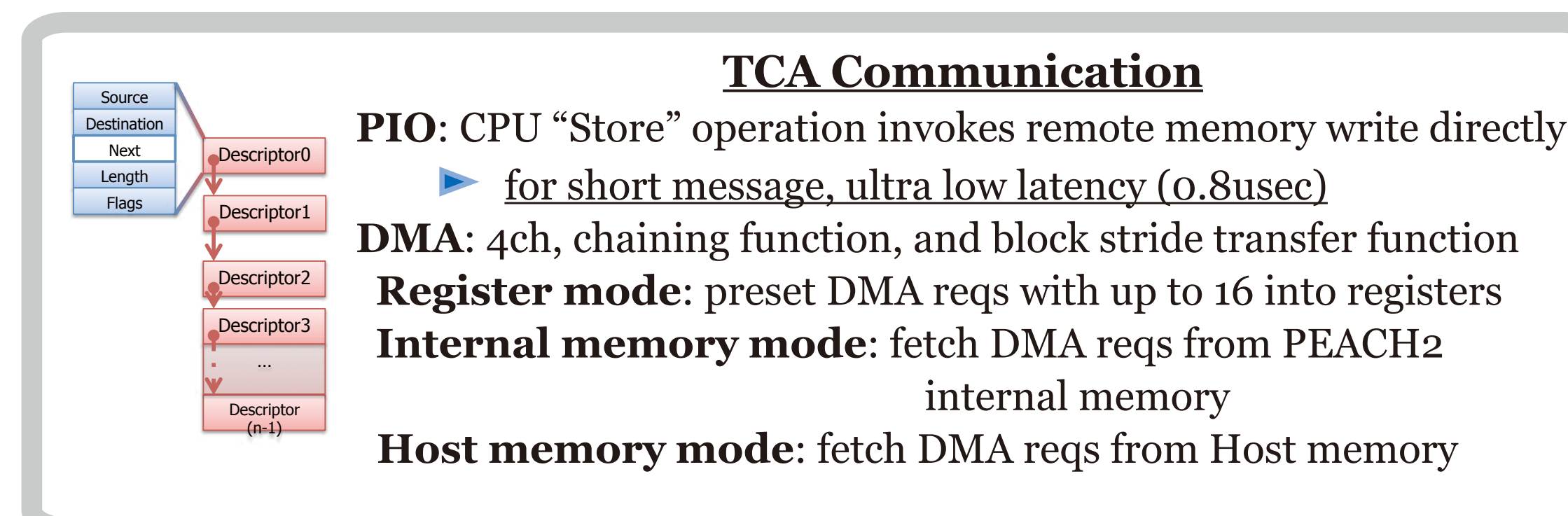
TCA provides the following benefits:

- Direct I/O among GPU memory over nodes (PCIe Gen2 x8 = 40Gbps)
  - Reduce the overhead
- Shared PCI Express address space among multiple nodes
  - Ease to program

HA-PACS/TCA is a proof-of-concept GPU cluster based on TCA architecture equipped with not only **TCA Interconnect** but also **InfiniBand QDR**.

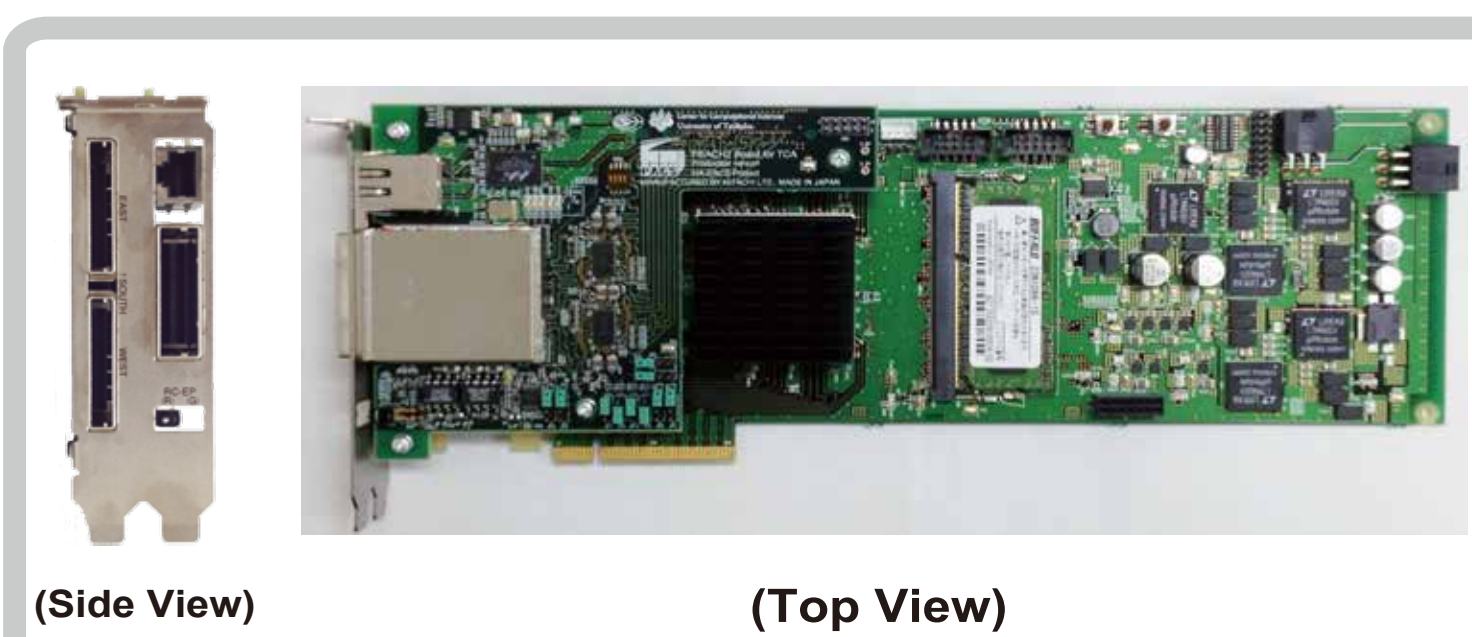


Block diagram of TCA compute node



### TCA Communication

- PIO:** CPU "Store" operation invokes remote memory write directly for short message, ultra low latency (0.8usec)
- DMA:** 4ch, chaining function, and block stride transfer function
- Register mode:** preset DMA reqs with up to 16 into registers
- Internal memory mode:** fetch DMA reqs from PEACH2 internal memory
- Host memory mode:** fetch DMA reqs from Host memory



TCA Communication Board using Altera FPGA (Stratix IV) (PCIe CEM Spec., double height)

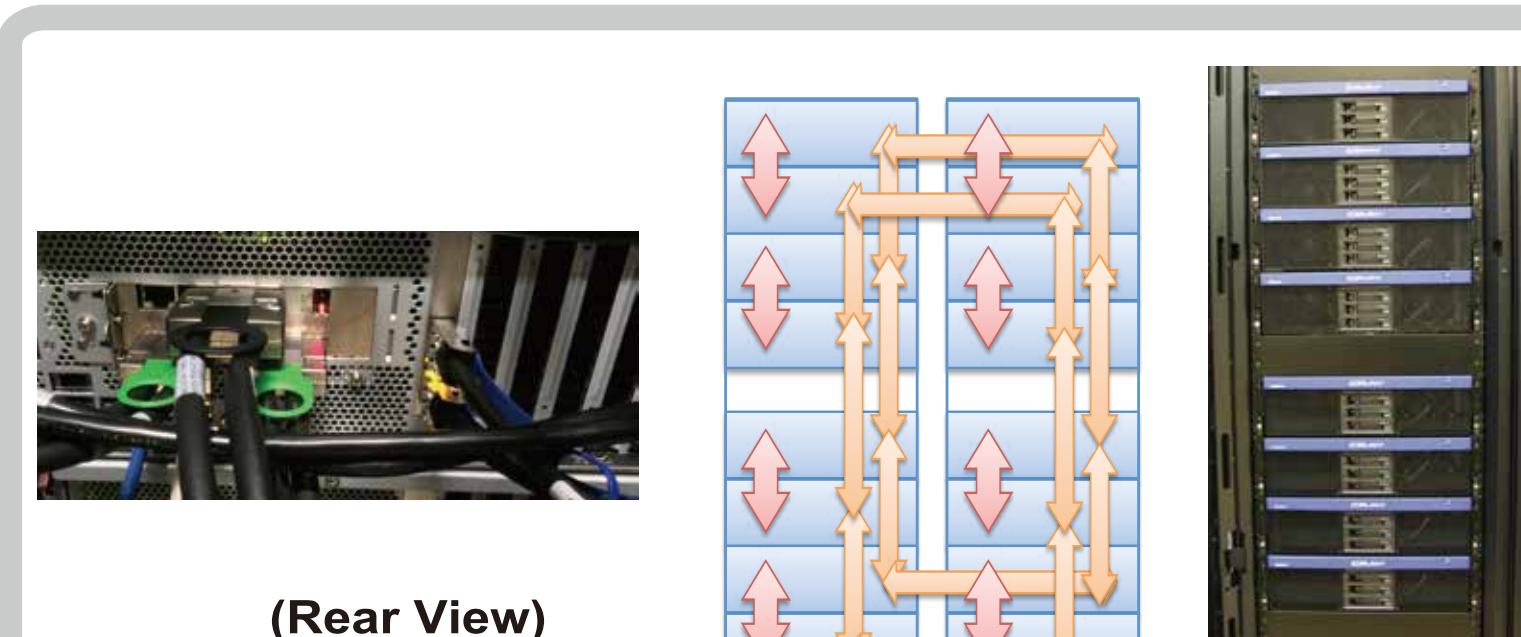
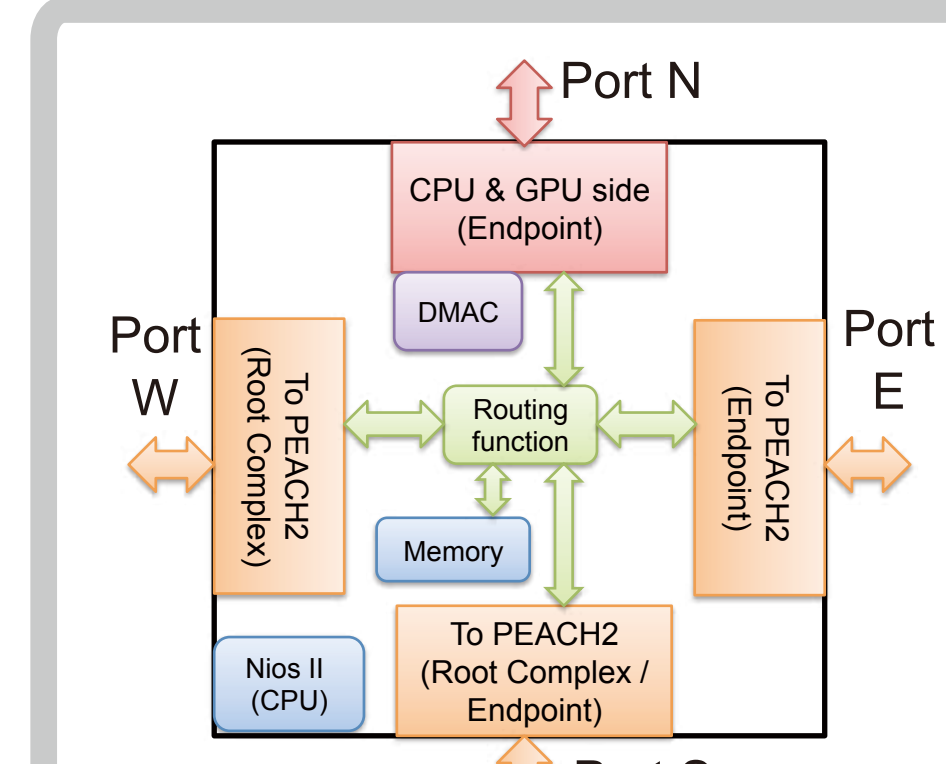


Photo of HA-PACS/TCA Compute Node, and Cabling configuration in TCA sub-cluster (16 nodes/group)



Block diagram of PEACH2 Chip

### HA-PACS/TCA Specification

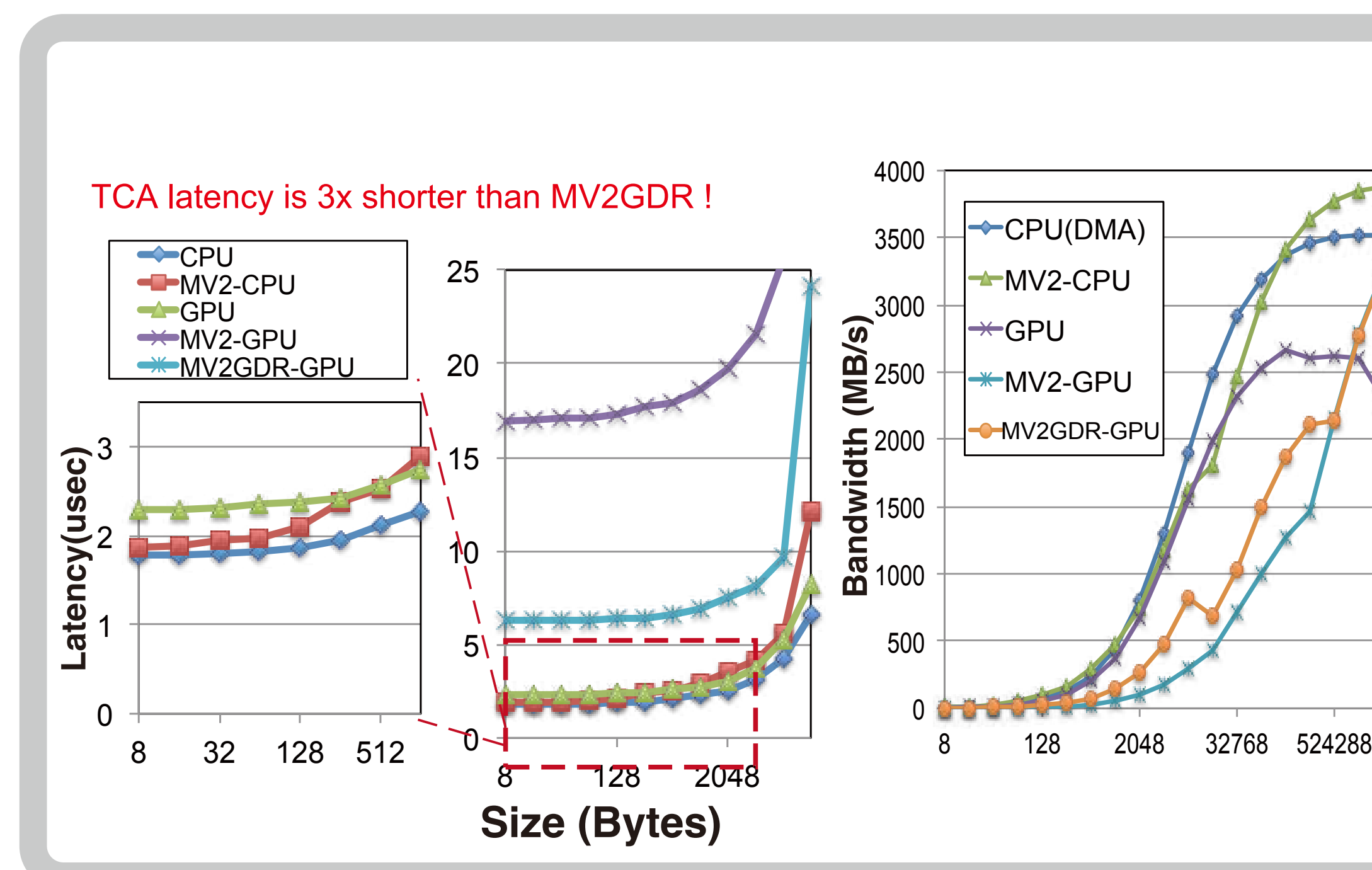
Computation Node	CRAY 3623G4-SM
Motherboard	SuperMicro X9DRG-QF
CPU	Intel Xeon E5 2680 v2 (Ivy Bridge 2.8 GHz, 10 core) x 2 socket
Memory	DDR3-1866MHz 4ch, 128 GB (119.4 GB/s)
Peak Performance	448 GFLOPS/node
GPU	NVIDIA Tesla K20X x 4 GPU
Memory	GDDR5 2600MHz, 6 GB/GPU (250 GB/s/GPU)
Peak Performance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



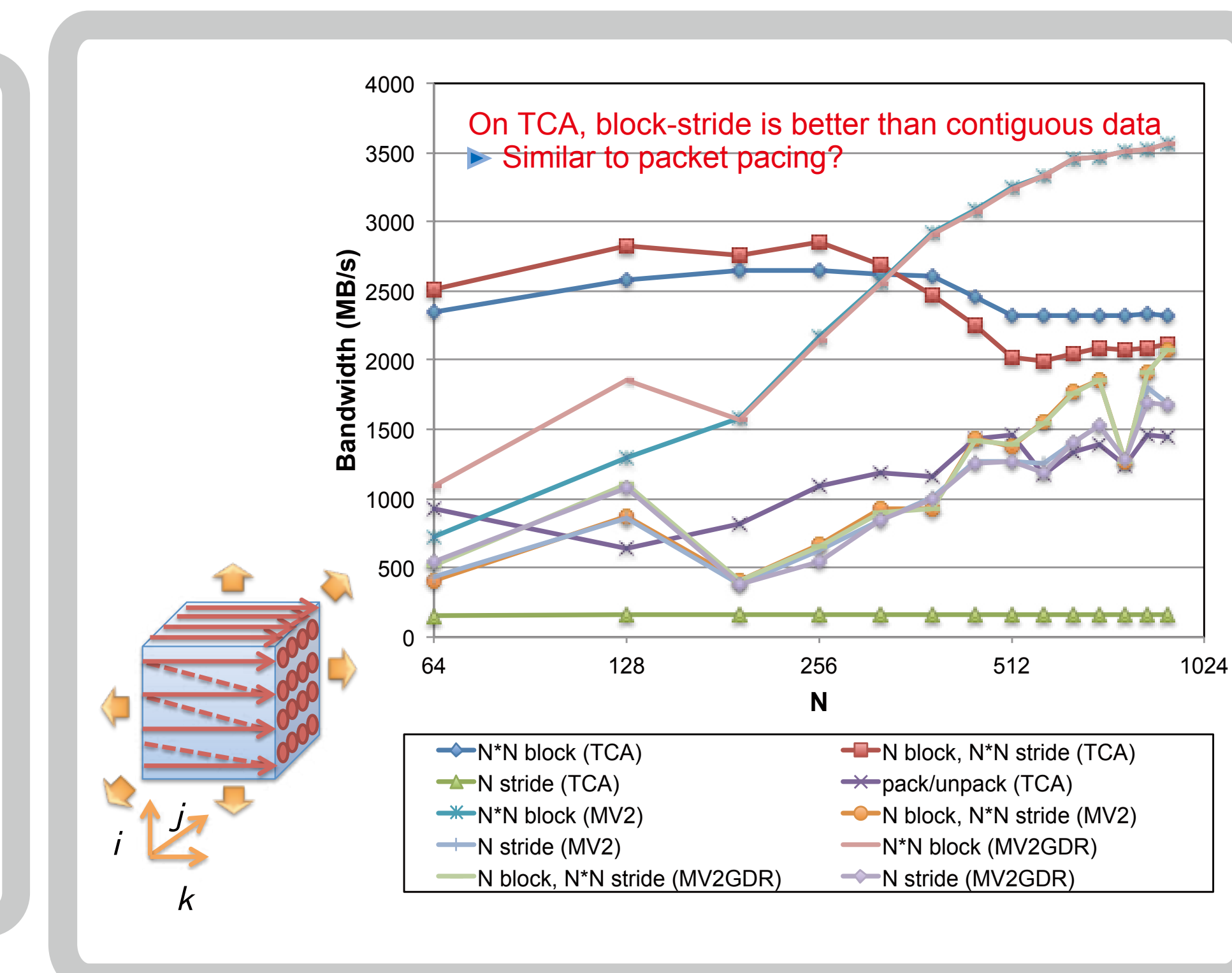
Entire HA-PACS System Including HA-PACS/TCA (5 racks x 2 rows) [Located at Univ. of Tsukuba]

## Performance of TCA Communication

**CPU:** CPU-to-CPU neighbor communication, **GPU:** GPU-to-GPU neighbor communication  
**MV2:** MVAPICH2, **MV2GDR:** with GDR

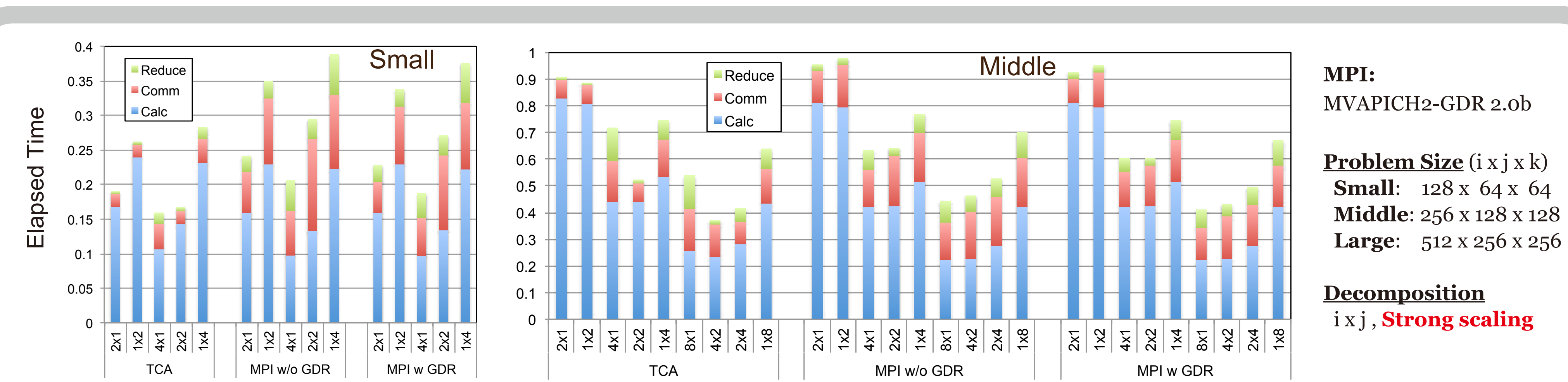


Ping-pong Latency and Bandwidth using DMA

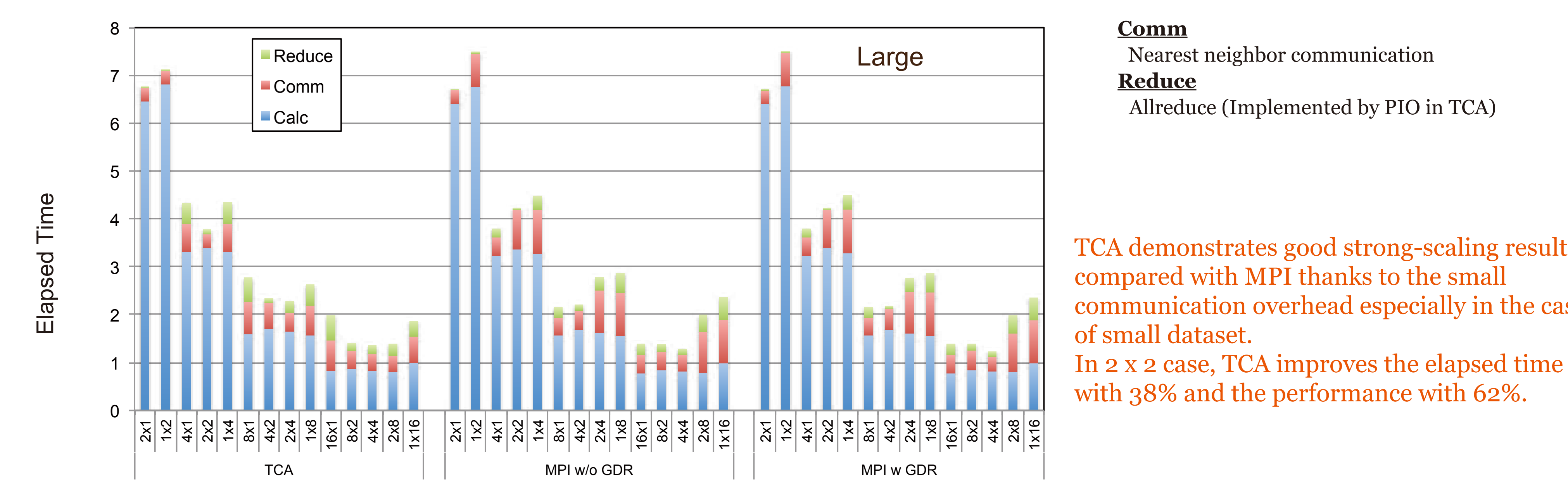


DMA performance of halo exchange on each plane in stencil computation

## Himeno Benchmark Results



**MPI:** MVAPICH2-GDR 2.0b  
**Problem Size** (i x j x k)  
**Small:** 128 x 64 x 64  
**Middle:** 256 x 128 x 128  
**Large:** 512 x 256 x 256  
**Decomposition** i x j, **Strong scaling**



**Comm** Nearest neighbor communication  
**Reduce** Allreduce (Implemented by PIO in TCA)

TCA demonstrates good strong-scaling results compared with MPI thanks to the small communication overhead especially in the case of small dataset. In 2 x 2 case, TCA improves the elapsed time with 38% and the performance with 62%.

HA-PACS Project is partially supported 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."