Exploiting InfiniBand and GPUDirect Technology for High Performance Collectives on GPU Clusters

Ching-Hsiang Chu

chu.368@osu.edu

Department of Computer Science and Engineering

The Ohio State University

Outline

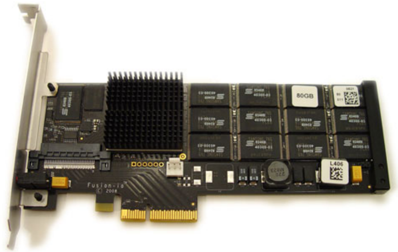
- Introduction
- Advanced Designs in MVAPICH2-GDR
  - CUDA-Aware MPI_Bcast
  - CUDA-Aware MPI_Allreduce / MPI_Reduce
- Concluding Remarks

Drivers of Modern HPC Cluster Architectures - Hardware

Multi-/Many-core Processors

High Performance Interconnects – InfiniBand (with SR-IOV)
<1usec latency, 200Gbps Bandwidth>

Accelerators / Coprocessors
high compute density, high performance/watt
>1 TFlop DP on a chip



SSD, NVMe-SSD, NVRAM

- **Multi-core/many-core technologies**
- **Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)**
- **Solid State Drives (SSDs), NVM, Parallel Filesystems, Object Storage Clusters**
- **Accelerators (NVIDIA GPGPUs and Intel Xeon Phi)**

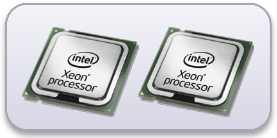
Sierra@LLNL

Stampede2@TACC

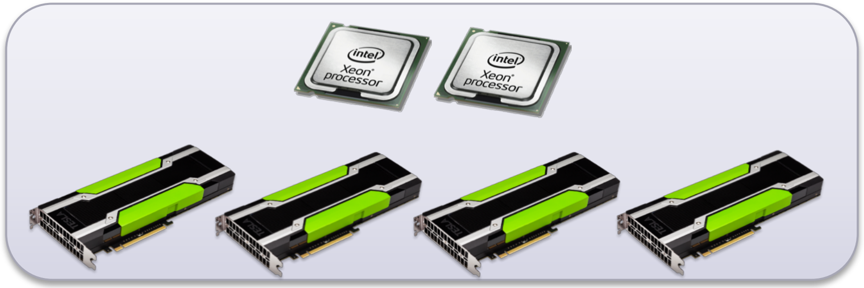
Comet@SDSC

Architectures for Deep Learning (DL)

Multi-core CPUs within a node

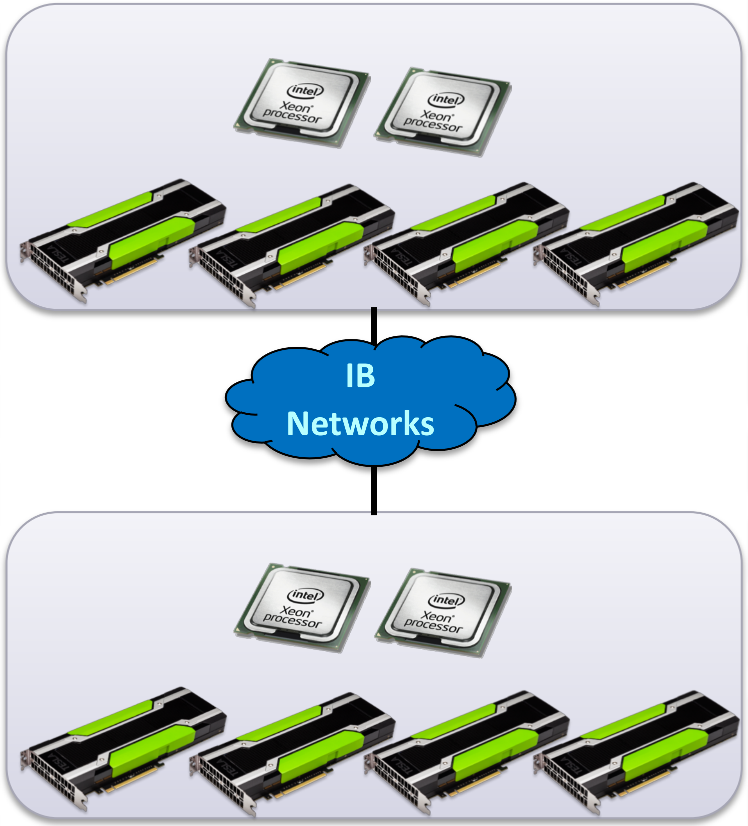


Multi-core CPUs + Multi-GPU within a node

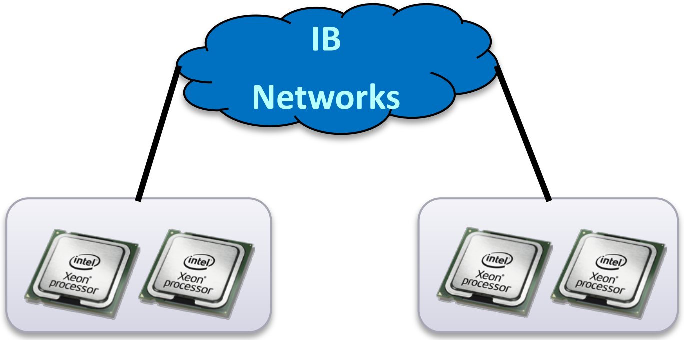


Multi-core CPUs + Multi-GPU across nodes

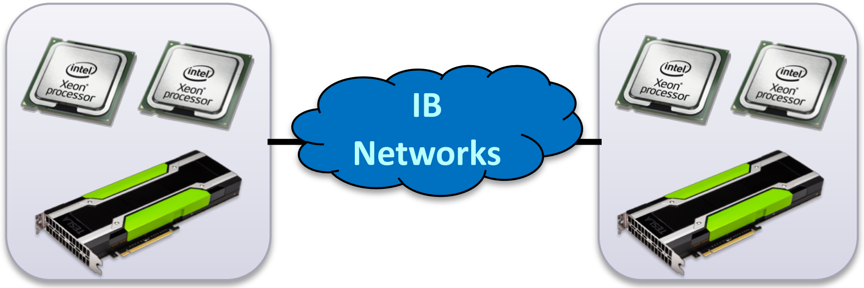
(E.g., Sierra/Summit)



Multi-core CPUs across nodes

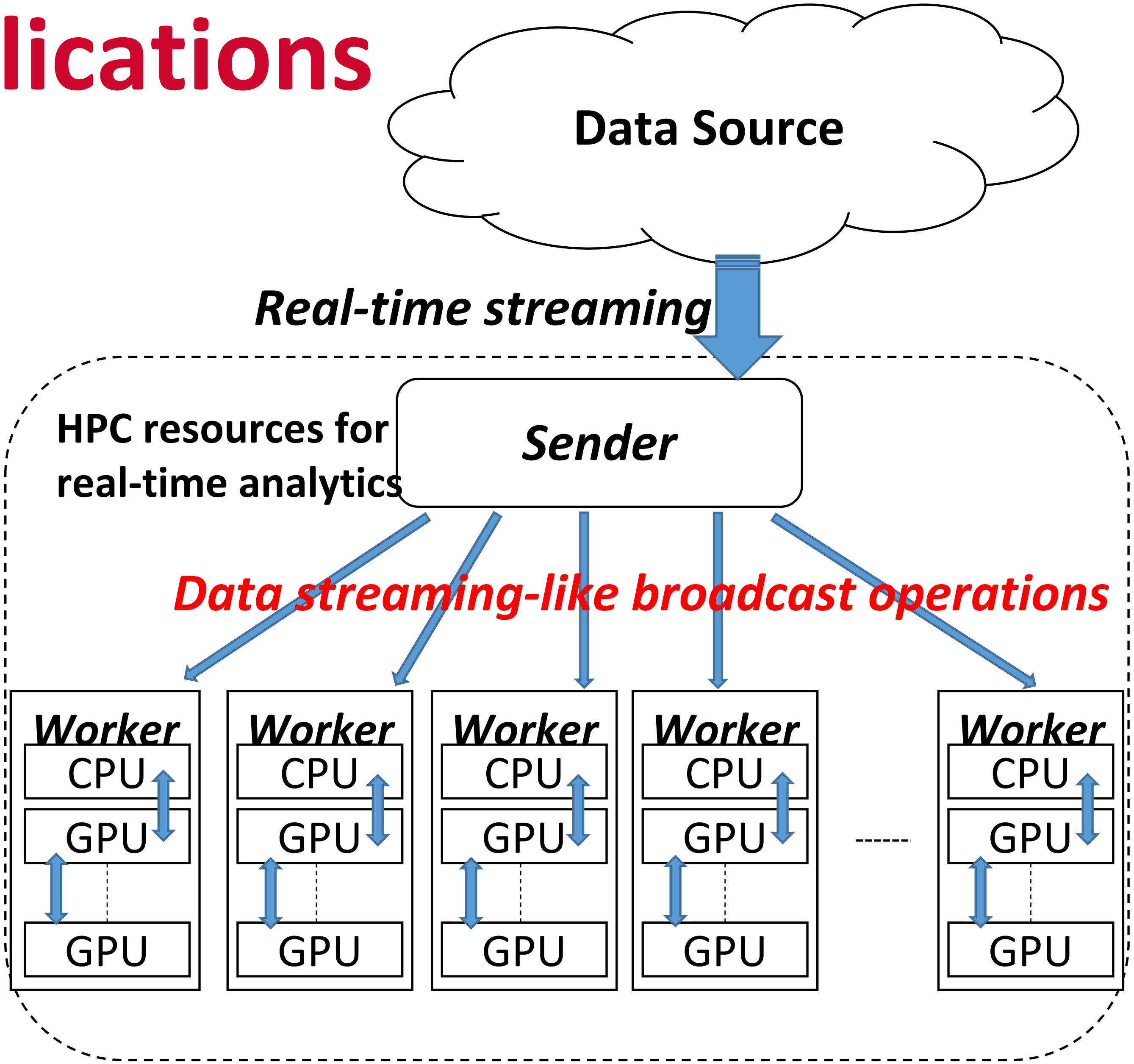


Multi-core CPUs + Single GPU across nodes



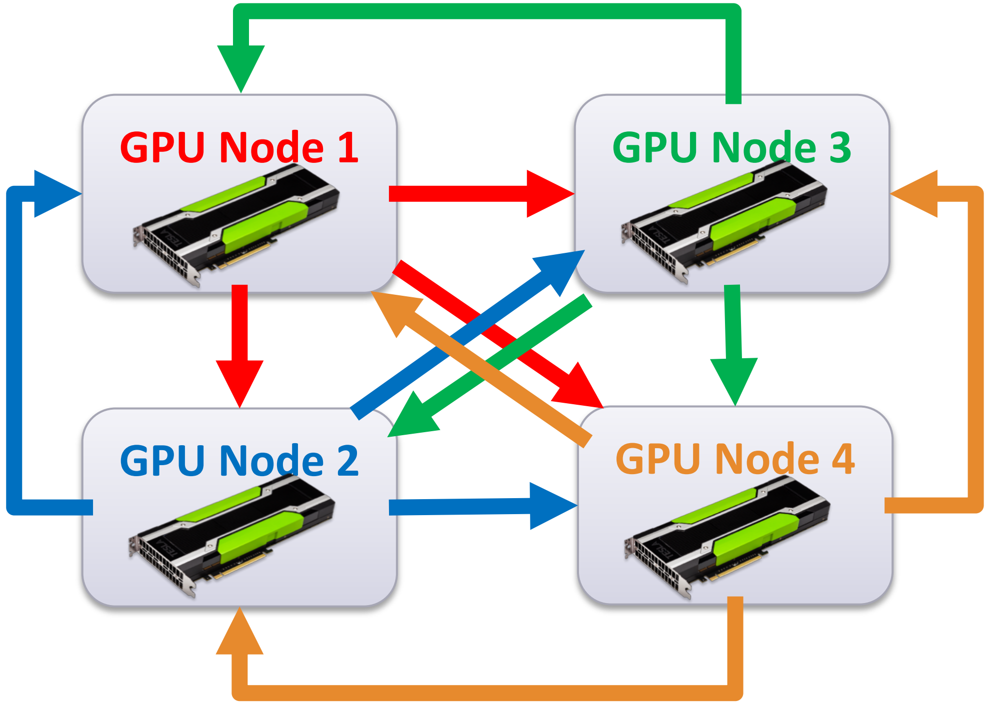
Streaming-like Applications

- **Streaming-like applications on HPC systems**
  1. **Communication (MPI)**
     - Broadcast
     - Allreduce/Reduce
  2. **Computation (CUDA)**
     - Multiple GPU nodes as workers



High-performance Deep Learning

- **Computation using GPU**
- **Communication using MPI**
  - Exchanging partial gradients after each minibatch
  - **All-to-all (Multi-Source) communications**
    - **E.g., MPI_Bcast, MPI_Allreduce**
- **Challenges**
  - High computation-communication **overlap**
  - Good **scalability** for upcoming large-scale GPU clusters
  - No application-level modification

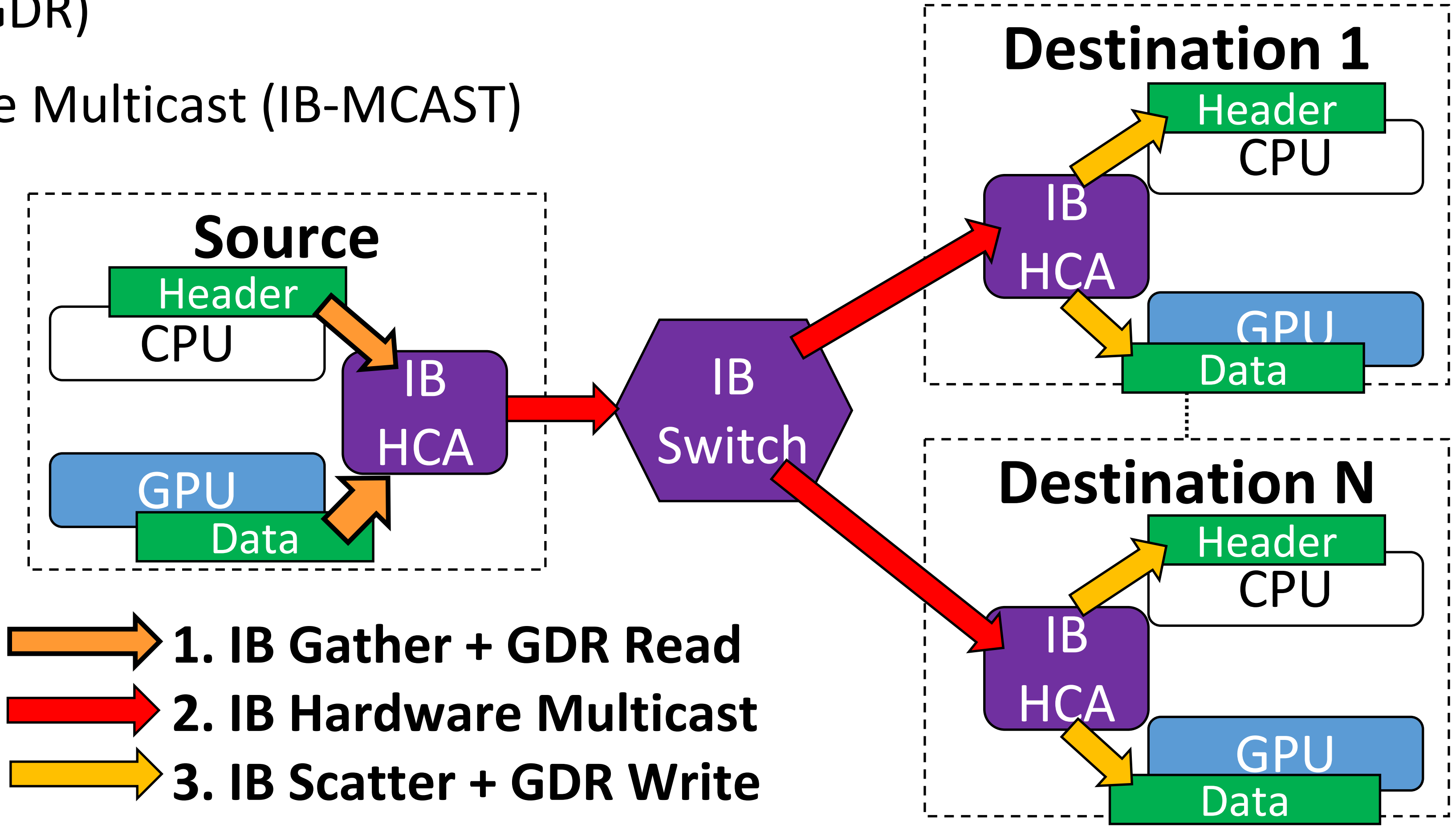


Outline

- Introduction
- Advanced Designs in MVAPICH2-GDR
  - CUDA-Aware MPI_Bcast
  - CUDA-Aware MPI_Allreduce / MPI_Reduce
- Concluding Remarks

Hardware Multicast-based Broadcast

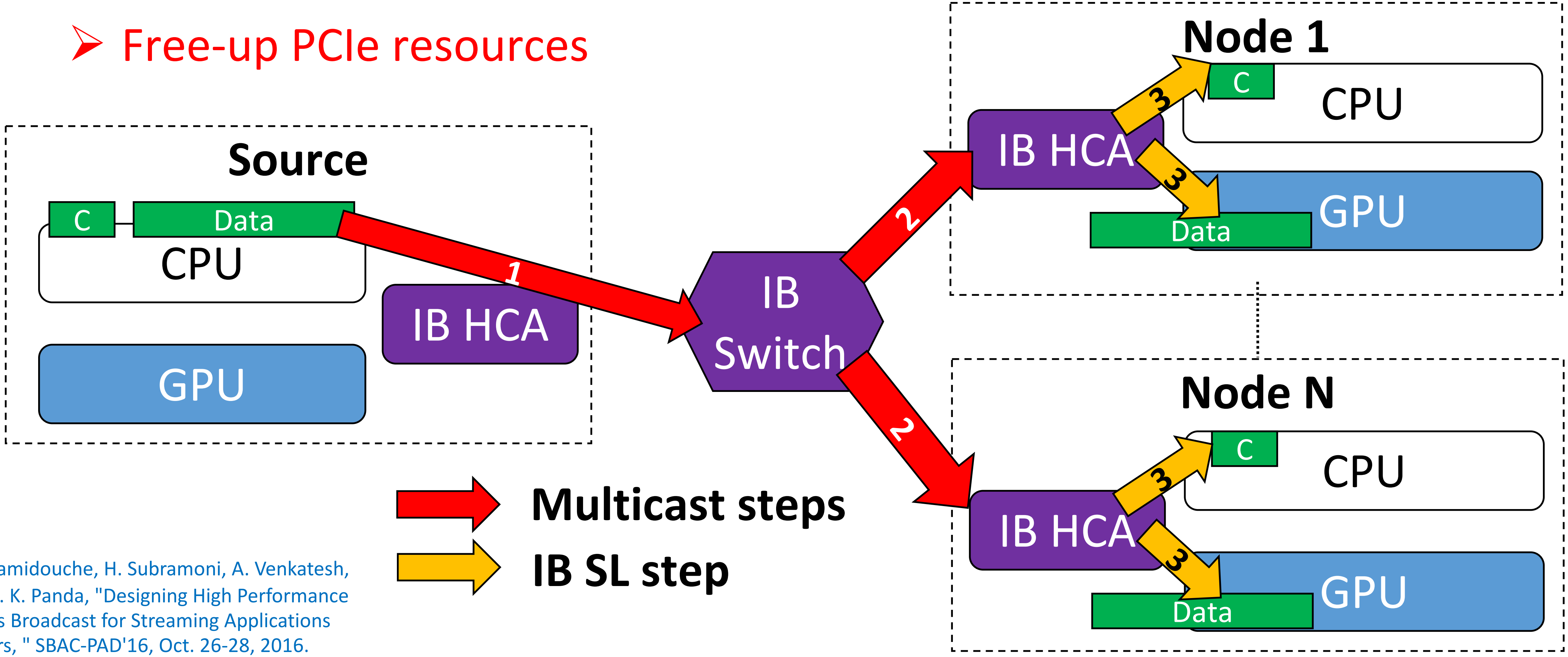
- **For GPU-resident data, using**
  - GPUDirect RDMA (GDR)
  - InfiniBand Hardware Multicast (IB-MCAST)
- **Overhead**
  - IB UD limit
  - GDR limit



A. Venkatesh, H. Subramoni, K. Hamidouche, and D. K. Panda, "A High Performance Broadcast Design with Hardware Multicast and GPUDirect RDMA for Streaming Applications on InfiniBand Clusters," in *HiPC 2014*, Dec 2014.

Hardware Multicast-based Broadcast (con't)

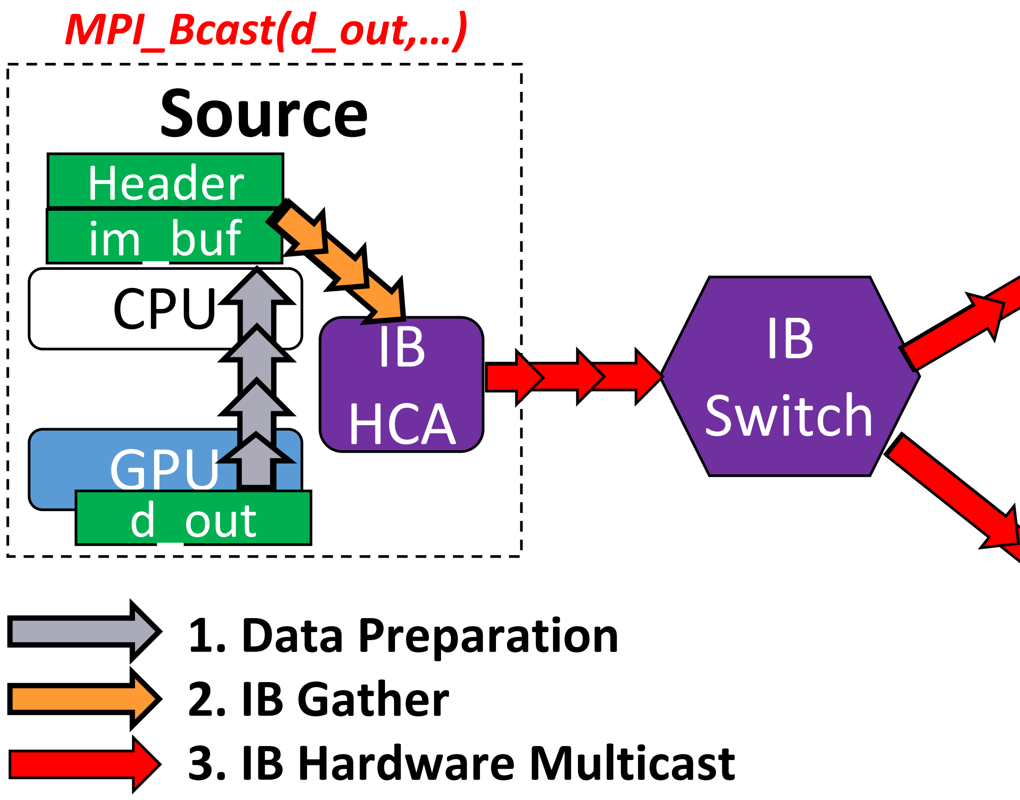
- **Heterogeneous Broadcast for streaming applications**
  - Free-up PCIe resources



C.-H. Chu, K. Hamidouche, H. Subramoni, A. Venkatesh, B. Elton, and D. K. Panda, "Designing High Performance Heterogeneous Broadcast for Streaming Applications on GPU Clusters, " SBAC-PAD'16, Oct. 26-28, 2016.

Optimized Broadcast Send

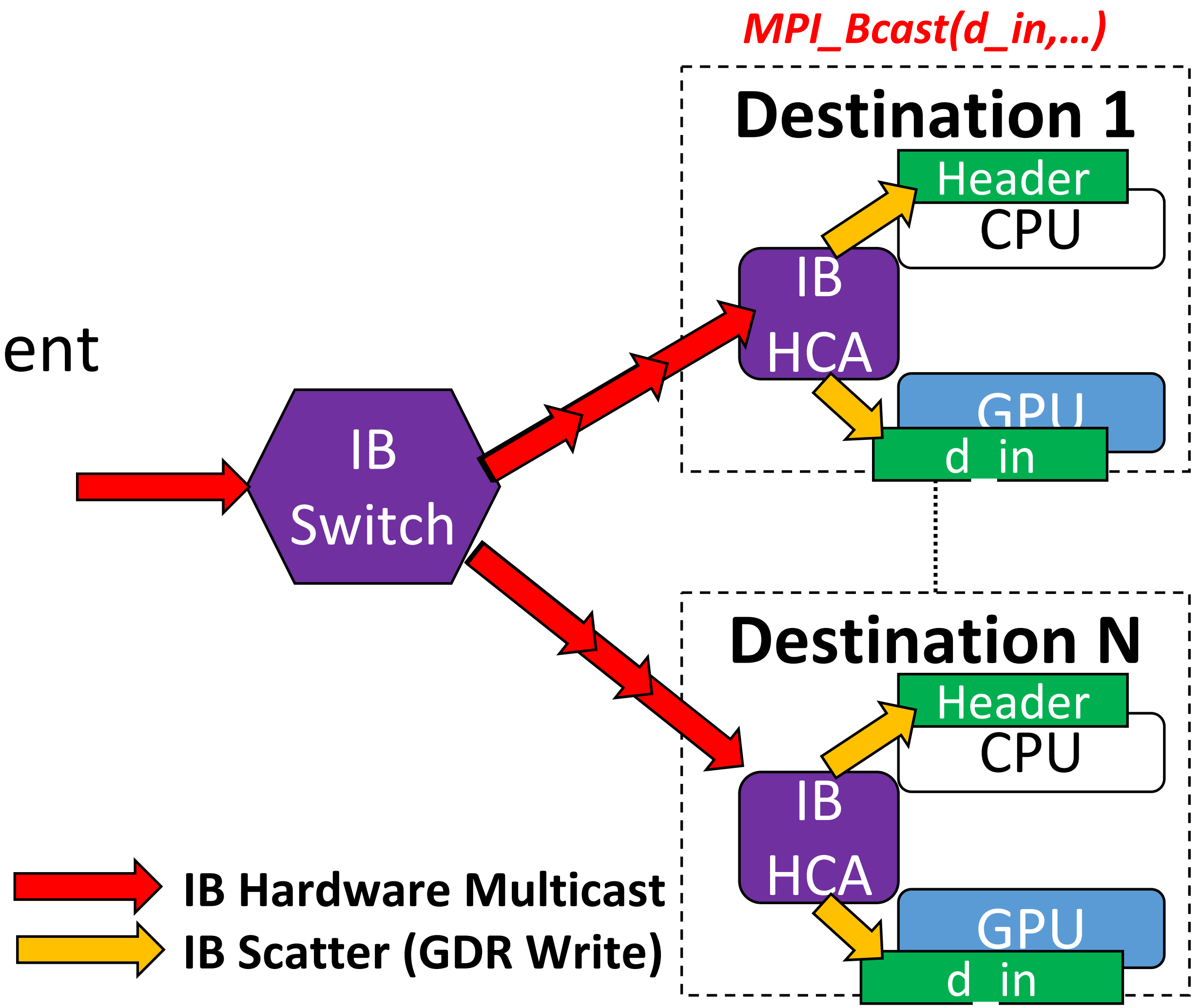
- **Preparing Intermediate buffer** (*im_buf*)
  - Page-locked (pinned) host buffer
    - ➢ Fast Device-Host data movement
  - Allocated at initialization phase
    - ➢ Low overhead, one time effort
- **Streaming data through host**
  - Fine-tuned chunked data
  - Asynchronous copy operations
  - ➢ Three-stage fine-tuned pipeline



C.-H. Chu, X. Lu, A. A. Awan, H. Subramoni, J. Hashmi, B. Elton and D. K. Panda., "Efficient and Scalable Multi-Source Streaming Broadcast on GPU Clusters for Deep Learning, " ICPP 2017, Aug 14-17, 2017.

Optimized Broadcast Receive

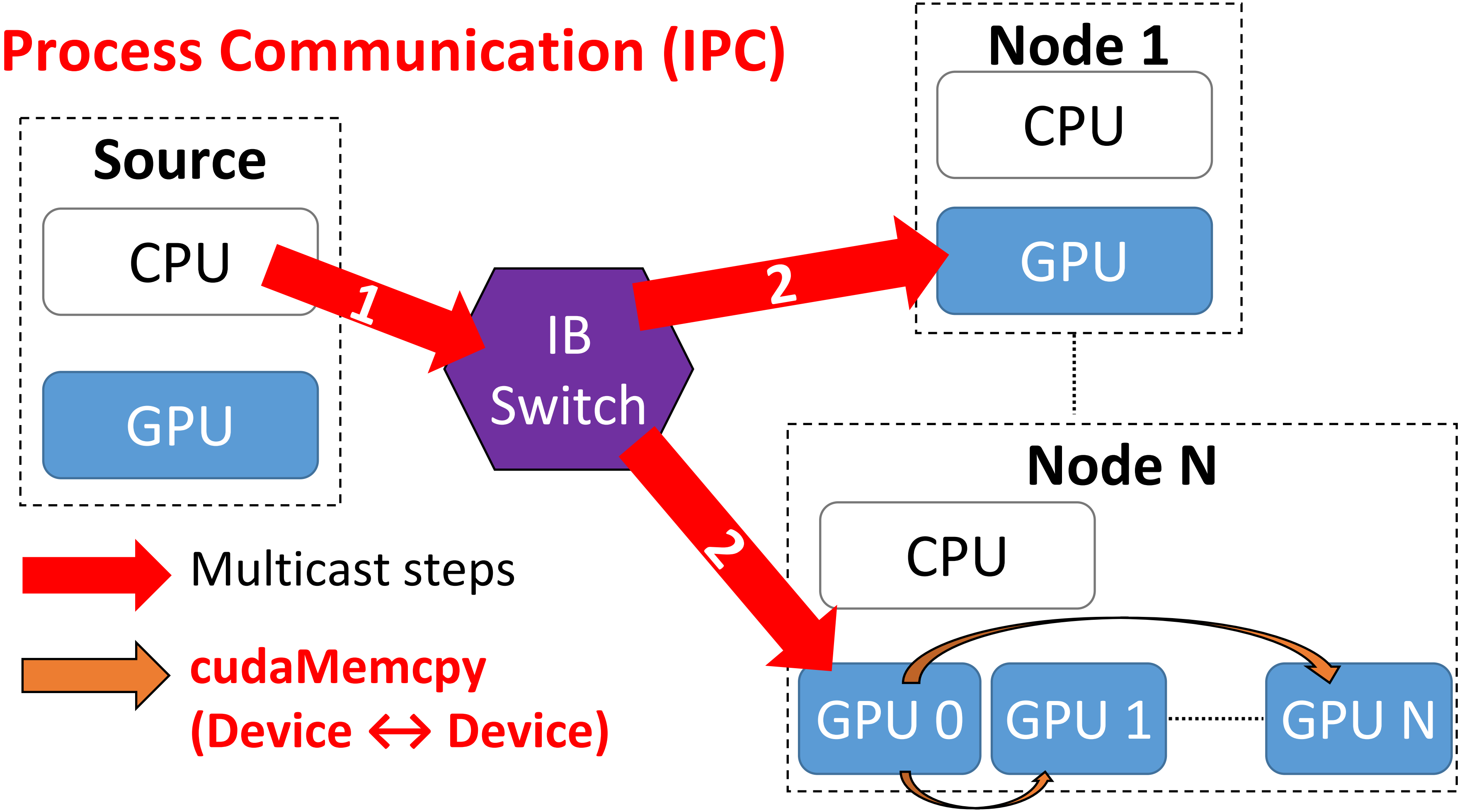
- **Zero-copy broadcast receive**
  - Pre-posted user buffer *(d_in)*
  - Avoids additional data movement
  - Leverages IB Scatter and GDR features
  - ➢ Low-latency
  - ➢ Free-up PCIe resources for applications



C.-H. Chu, X. Lu, A. A. Awan, H. Subramoni, B. Elton, D. K. Panda, "Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast," to appear in IEEE Transactions on Parallel and Distributed Systems (TPDS).

Broadcast on Multi-GPU systems

- **Proposed Intra-node Topology-Aware Broadcast**
  - **CUDA InterProcess Communication (IPC)**

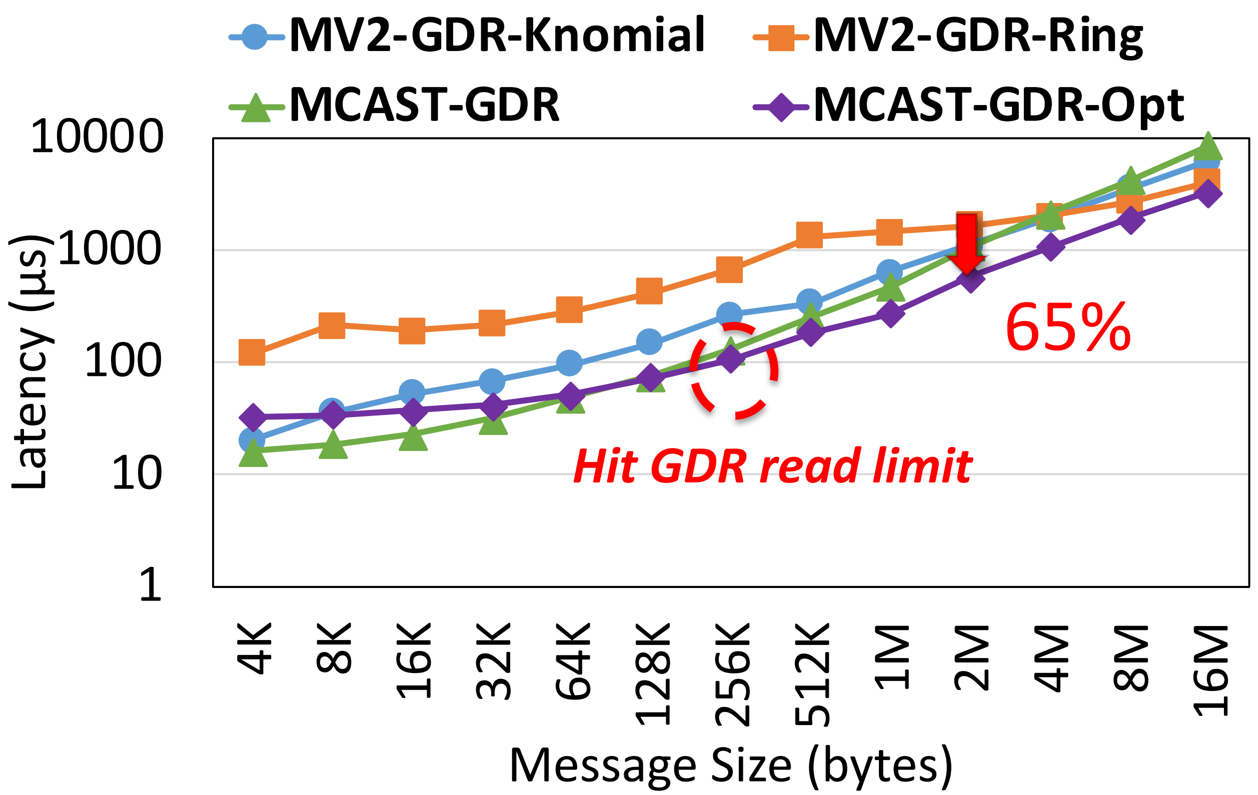


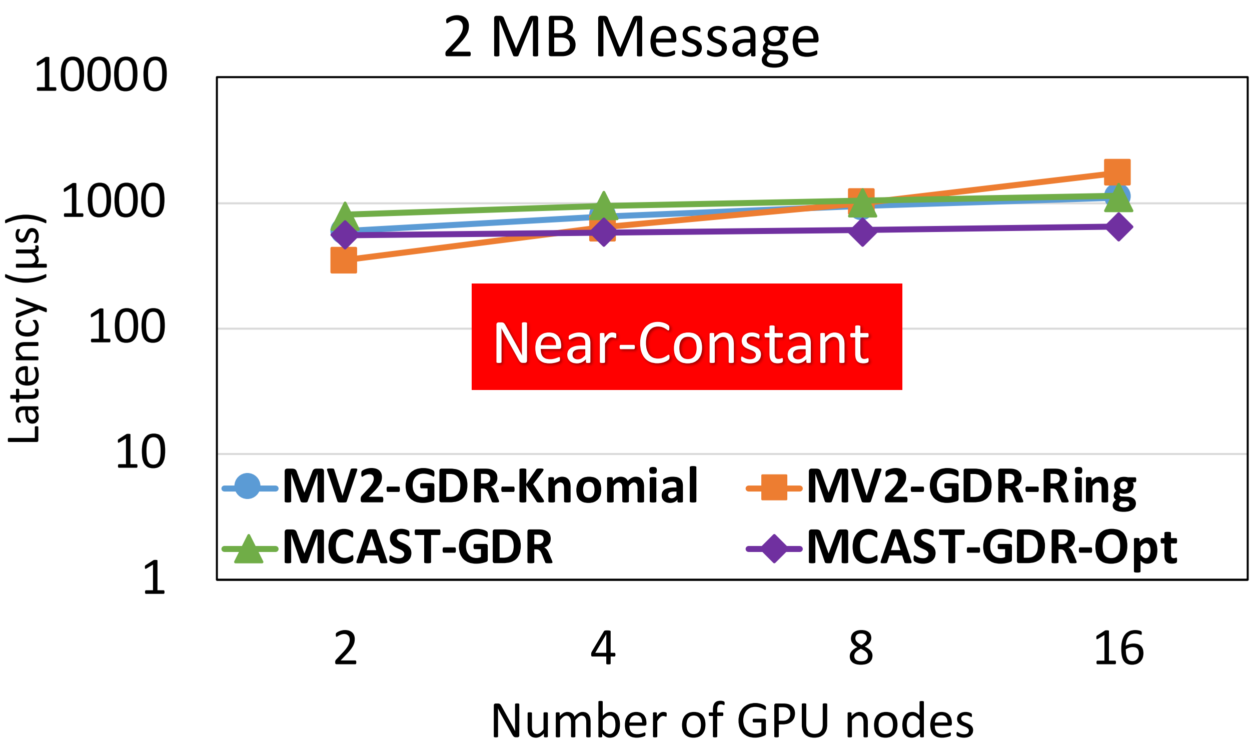
C.-H. Chu, K. Hamidouche, H. Subramoni, A. Venkatesh, B. Elton, and D. K. Panda, "Designing High Performance Heterogeneous Broadcast for Streaming Applications on GPU Clusters, " SBAC-PAD'16, Oct. 26-28, 2016.

Benchmark Evaluation

- **@ RI2 cluster, 16 GPUs, 1 GPU/node**

Lower is better

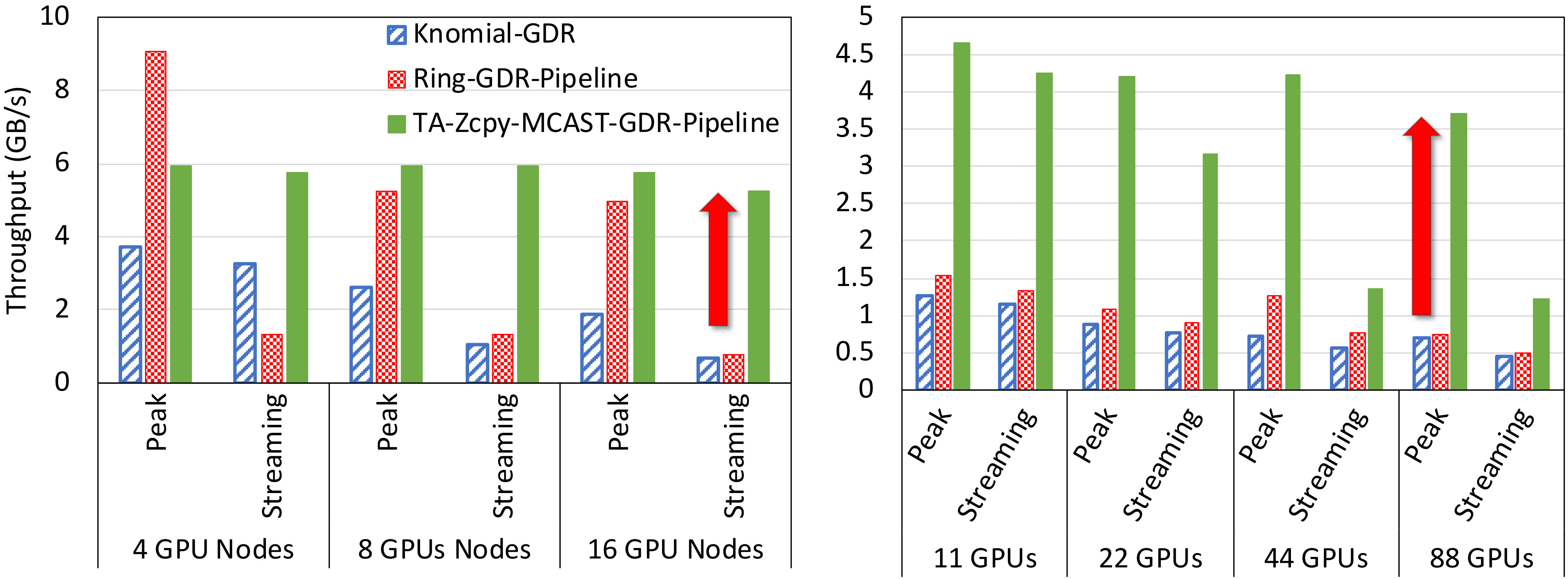




- **Provide near-constant latency over the system sizes**
- **Reduces up to 65% of latency for large messages**

C.-H. Chu, X. Lu, A. A. Awan, H. Subramoni, J. Hashmi, B. Elton and D. K. Panda., "Efficient and Scalable Multi-Source Streaming Broadcast on GPU Clusters for Deep Learning, " ICPP 2017, Aug 14-17, 2017.

Streaming Workload @ RI2 (16 GPUs) & CSCS (88 GPUs)

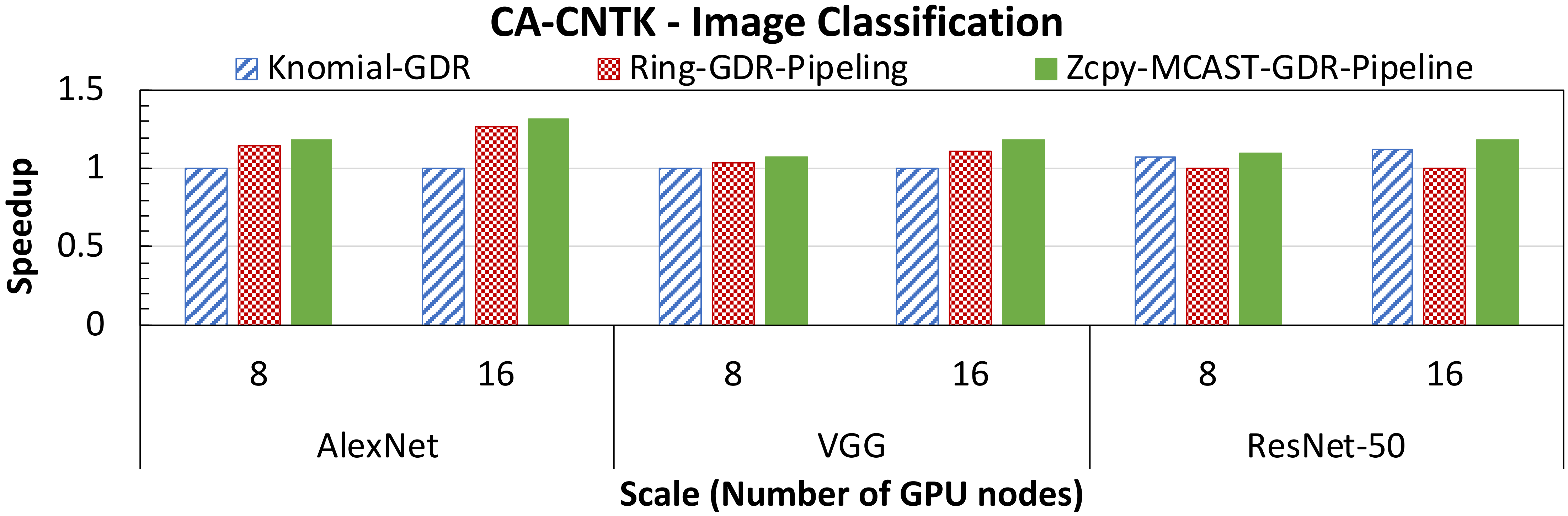


- **IB-MCAST + GDR + IPC-based MPI_Bcast schemes**
  - Stable high throughput compared to existing schemes

C.-H. Chu, X. Lu, A. A. Awan, H. Subramoni, B. Elton, D. K. Panda, "Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast," to appear in IEEE Transactions on Parallel and Distributed Systems (TPDS).

Performance Benefits with CNTK Deep Learning Framework @ RI2 cluster, 16 GPUs

- **CUDA-Aware Microsoft Cognitive Toolkit (CA-CNTK) without modification**



- **Reduces up to 24%, 15%, 18% of latency for AlexNet, VGG, and ResNet-50 models**
- **Higher improvement is expected for larger system sizes**

C.-H. Chu, X. Lu, A. A. Awan, H. Subramoni, B. Elton, D. K. Panda, "Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast," to appear in IEEE Transactions on Parallel and Distributed Systems (TPDS).

CUDA-Aware MPI_Allreduce

- **Existing designs**
  1. Explicit copy the data from GPU to host memory
  2. Host-to-Host communication to remote processes
  3. Perform computation on CPU
  4. Explicit copy the data from host to GPU memory

Expensive!

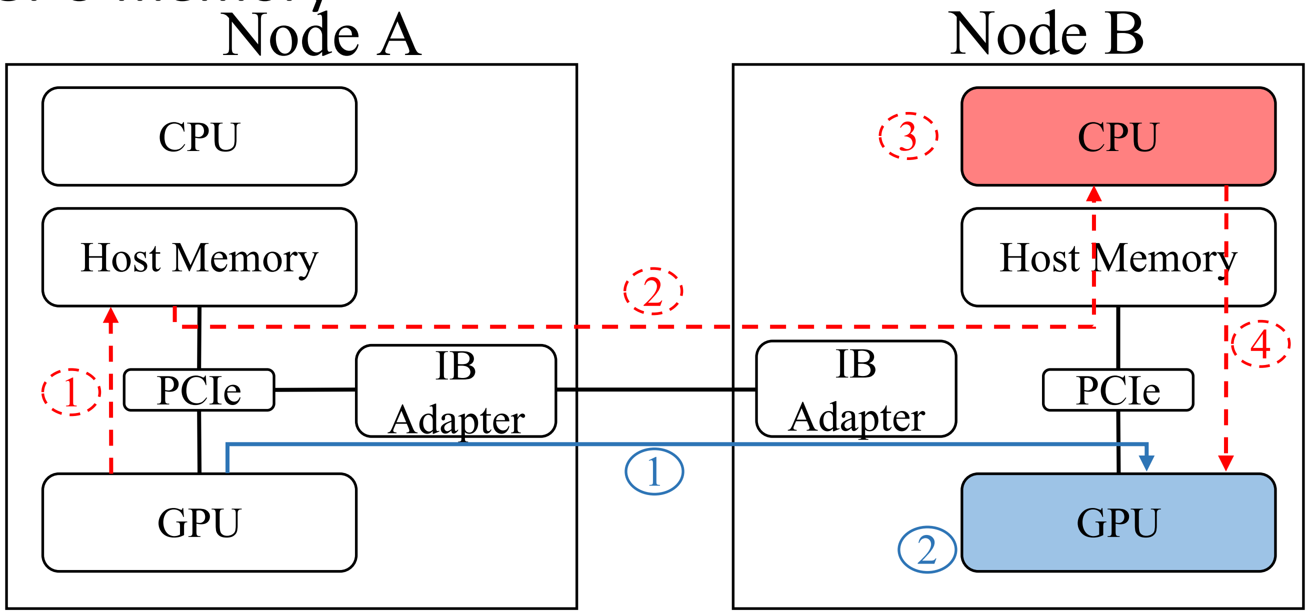
Fast

Relative slow for large data

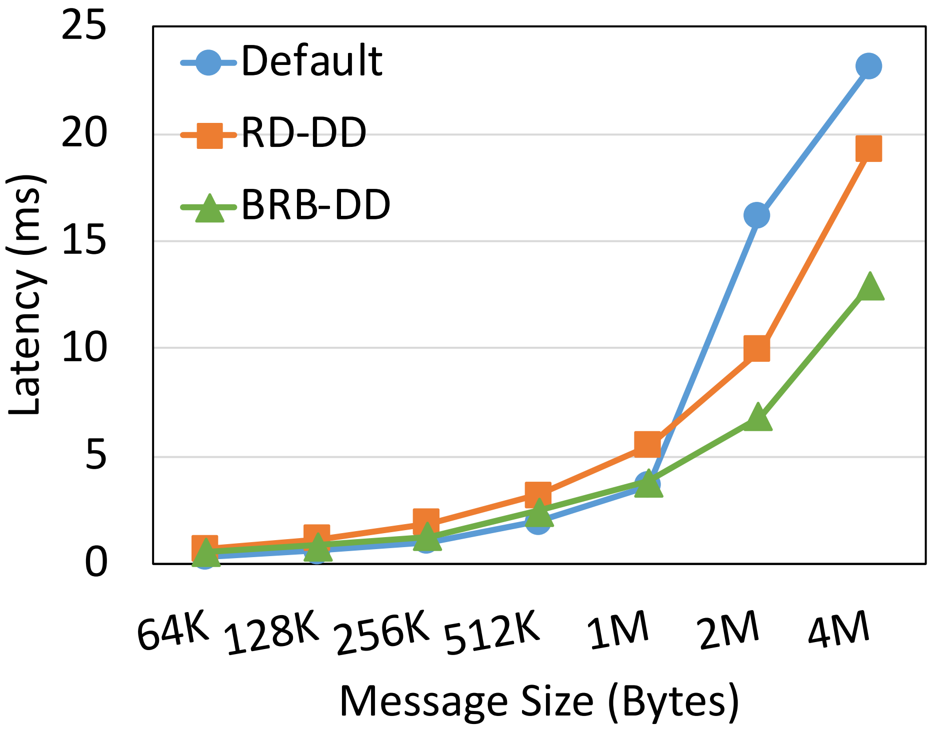
Good for small data

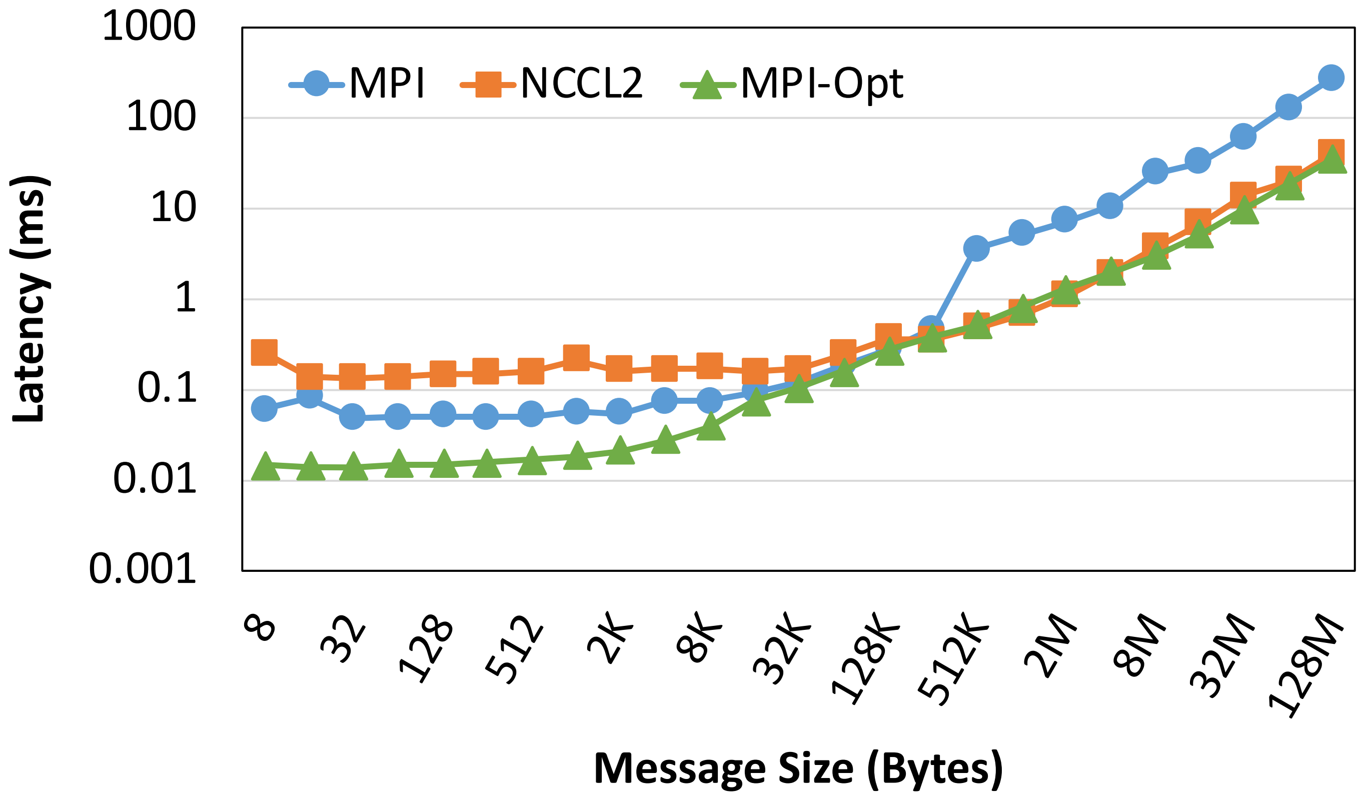
Expensive!

- **Proposed designs**
  1. GPU-to-GPU communication
     - **NVIDIA GPUDirect RDMA (GDR)**
     - Pipeline through host for large msg
  2. Perform computation on GPU
     - Efficient CUDA kernels



Benchmark Evaluation @ RI2 cluster, 16 GPUs





[1] C. Chu, K. Hamidouche, A. Venkatesh, A. A. Awan and D. K. Panda, "CUDA Kernel Based Collective Reduction Operations on Large-scale GPU Clusters," in CCGrid'16, Cartagena, 2016, pp. 726-735.

[2] Awan AA, Bedorf J, Chu CH, Subramoni H, Panda DK. Scalable Distributed DNN Training using TensorFlow and CUDA-Aware MPI: Characterization, Designs, and Performance Evaluation. arXiv preprint arXiv:1810.11112. 2018 Oct 25.

Outline

- Introduction
- Advanced Designs in MVAPICH2-GDR
  - CUDA-Aware MPI_Bcast
  - CUDA-Aware MPI_Allreduce / MPI_Reduce
- Concluding Remarks

Concluding Remarks

- **High-performance broadcast schemes to leverage GDR and IB-MCAST features for streaming and deep learning applications**
  - Optimized streaming design for large messages transfers
  - High-performance reliability support for IB-MCAST
- **High-performance CUDA-Aware Allreduce for deep learning**
  - Efficient reduction kernel on GPUs

➢ **These features are included in MVAPICH2-GDR 2.3**
  - ➢ http://mvapich.cse.ohio-state.edu/
  - ➢ http://mvapich.cse.ohio-state.edu/userguide/gdr/2.3/

Thank You!

- **Join us for more tech talks from MVAPICH2 team**
  - http://mvapich.cse.ohio-state.edu/talks/

The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/