MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

The Ohio State University

Scalable and Reliable Broadcast using InifiniBand and NVIDIA GPUDirect Technology in MVAPICH2-GDR

Ching-Hsiang Chu

chu.368@osu.edu

Ph.D. Candidate

Department of Computer Science and Engineering

The Ohio State University

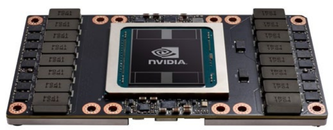
Outline

- Introduction
- Advanced Broadcast Designs in MVAPICH2-GDR
- Concluding Remarks

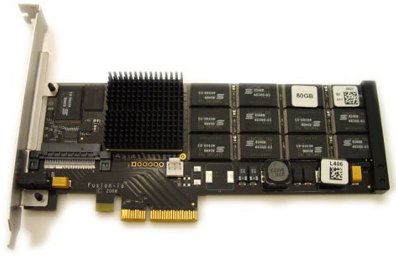
Trends in Modern HPC Architecture: Heterogeneous

Multi/ Many-core Processors

High Performance Interconnects
InfiniBand, Omni-Path, EFA
<1usec latency, 200Gbps+ Bandwidth



Accelerators / Coprocessors
high compute density,
high performance/watt



SSD, NVMe-SSD, NVRAM
Node local storage

- **Multi-core/many-core technologies**
- **High Performance Interconnects**
- **High Performance Storage and Compute devices**
- **Variety of programming models (MPI, PGAS, MPI+X)**

#1 Summit
(27,648 GPUs)

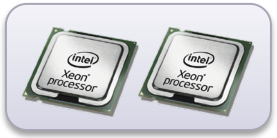
#2 Sierra (17,280 GPUs)
#10 Lassen (2,664 GPUs)

#8 ABCI
(4,352 GPUs)

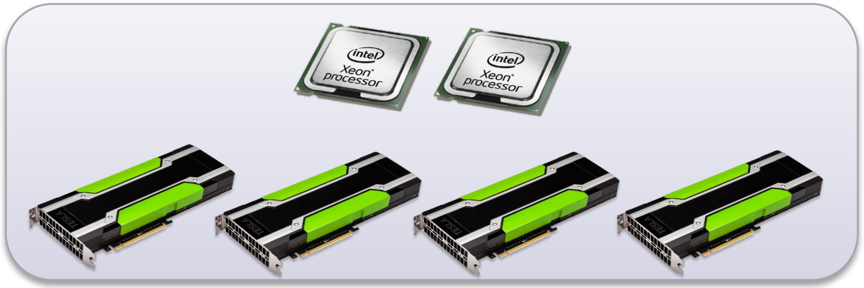
#22 DGX SuperPOD
(1,536 GPUs)

Architectures: Past, Current, and Future

Multi-core CPUs within a node

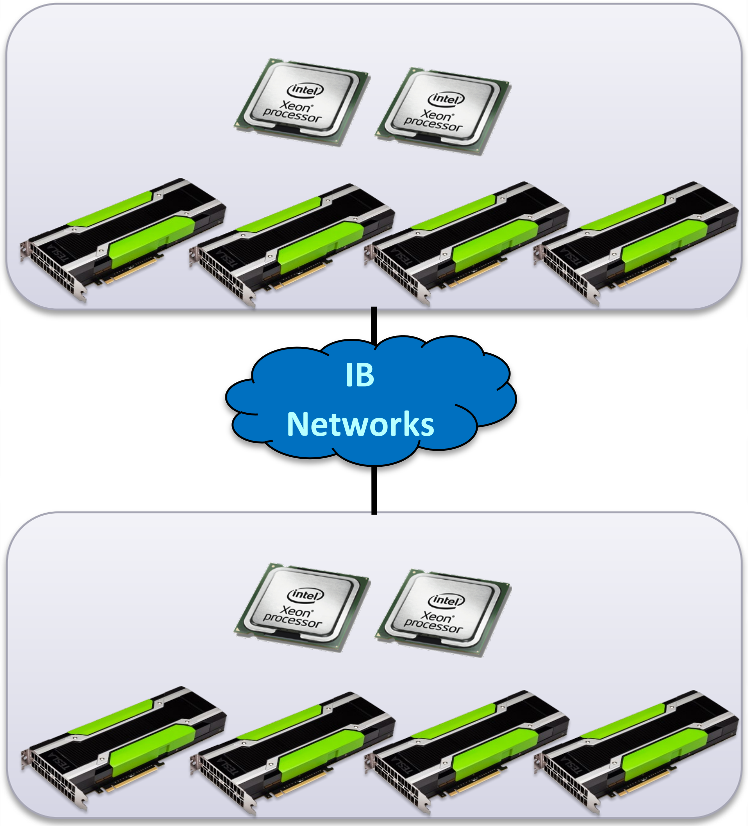


Multi-core CPUs + Multi-GPU within a node

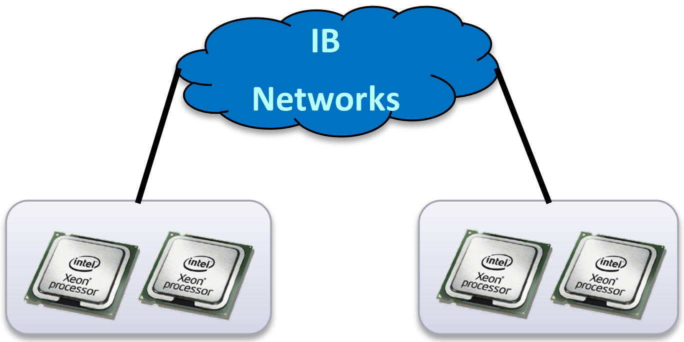


Multi-core CPUs + Multi-GPU across nodes

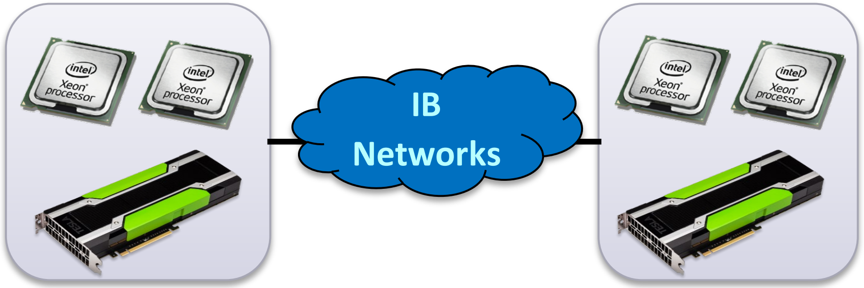
(E.g., Sierra/Summit, Frontier)



Multi-core CPUs across nodes

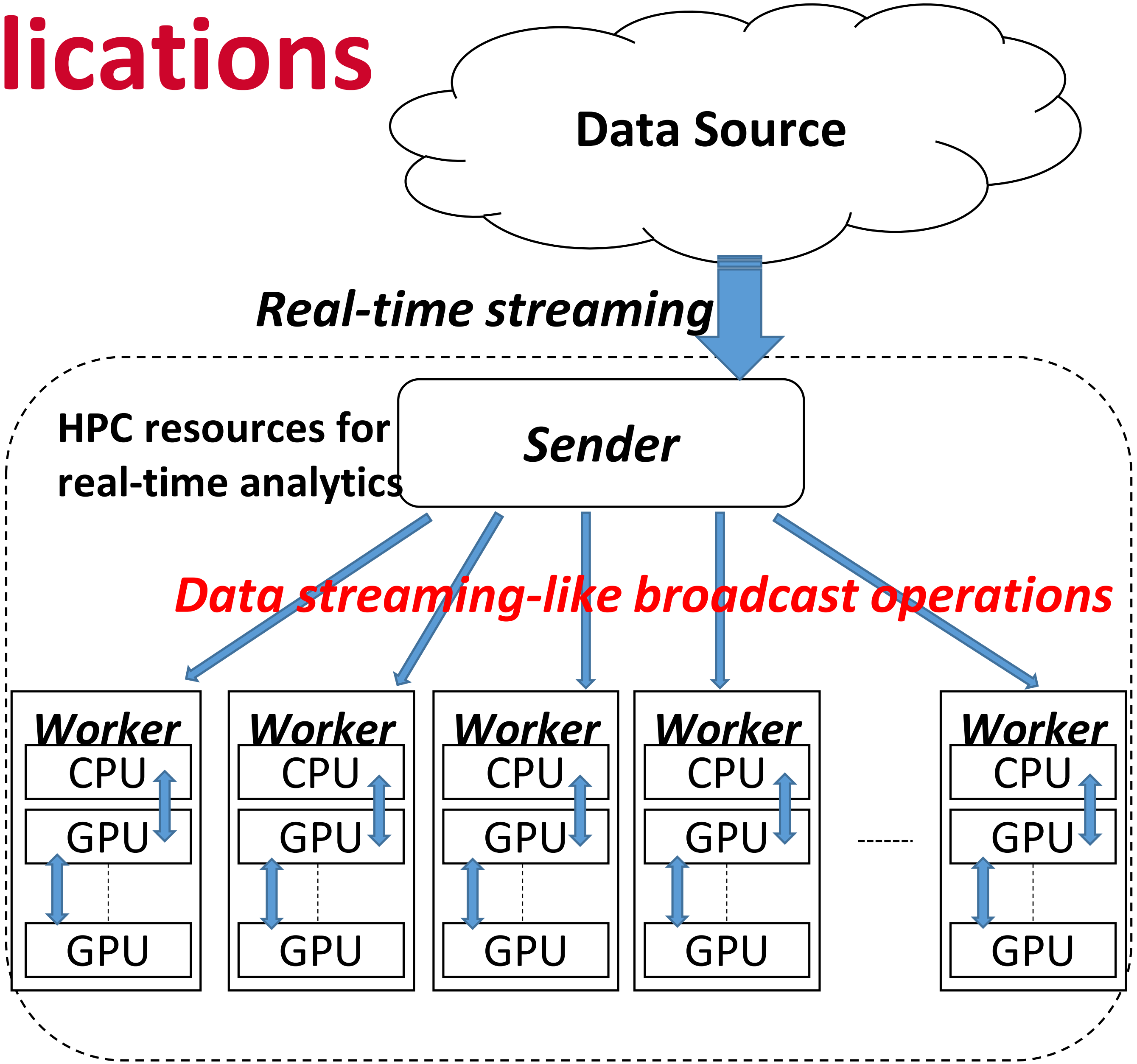


Multi-core CPUs + Single GPU across nodes



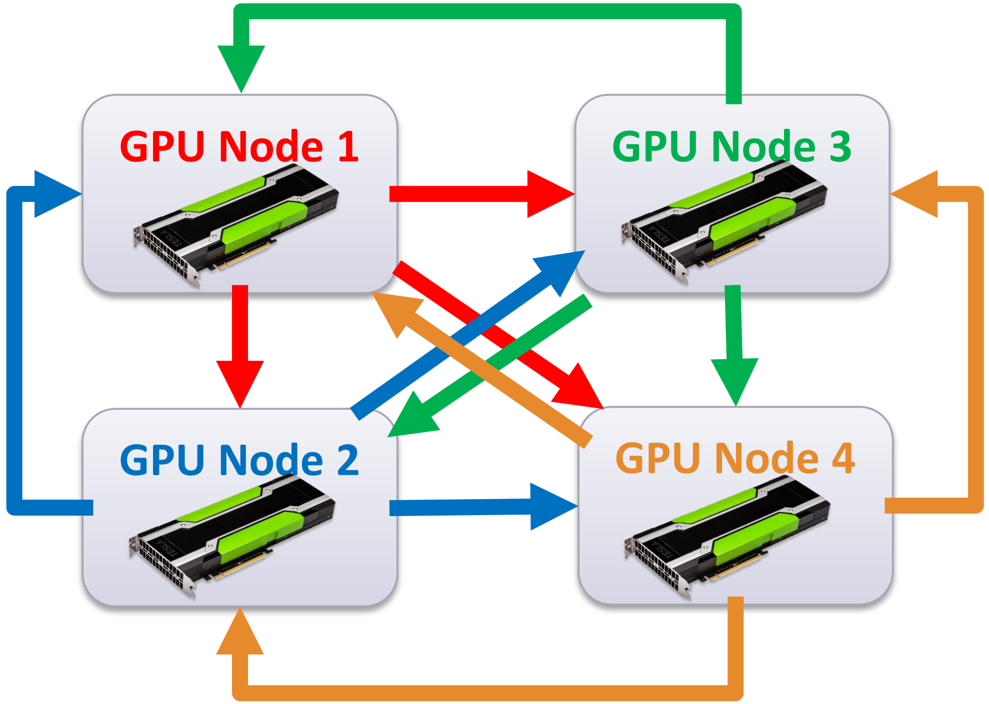
Streaming-like Applications

- **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

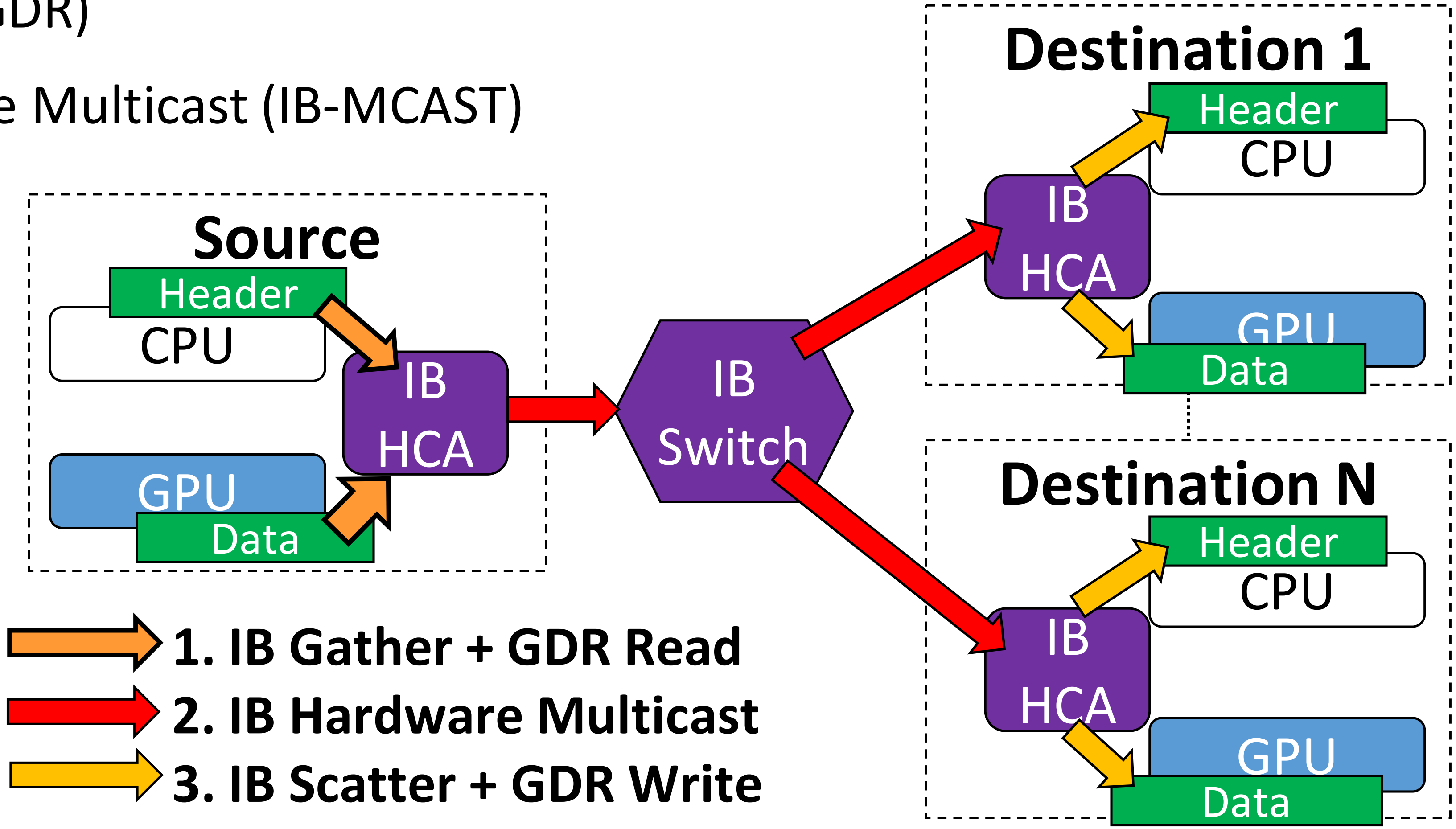


Outline

- Introduction
- Advanced Broadcast Designs in MVAPICH2-GDR
- Concluding Remarks

Hardware Multicast-based Broadcast

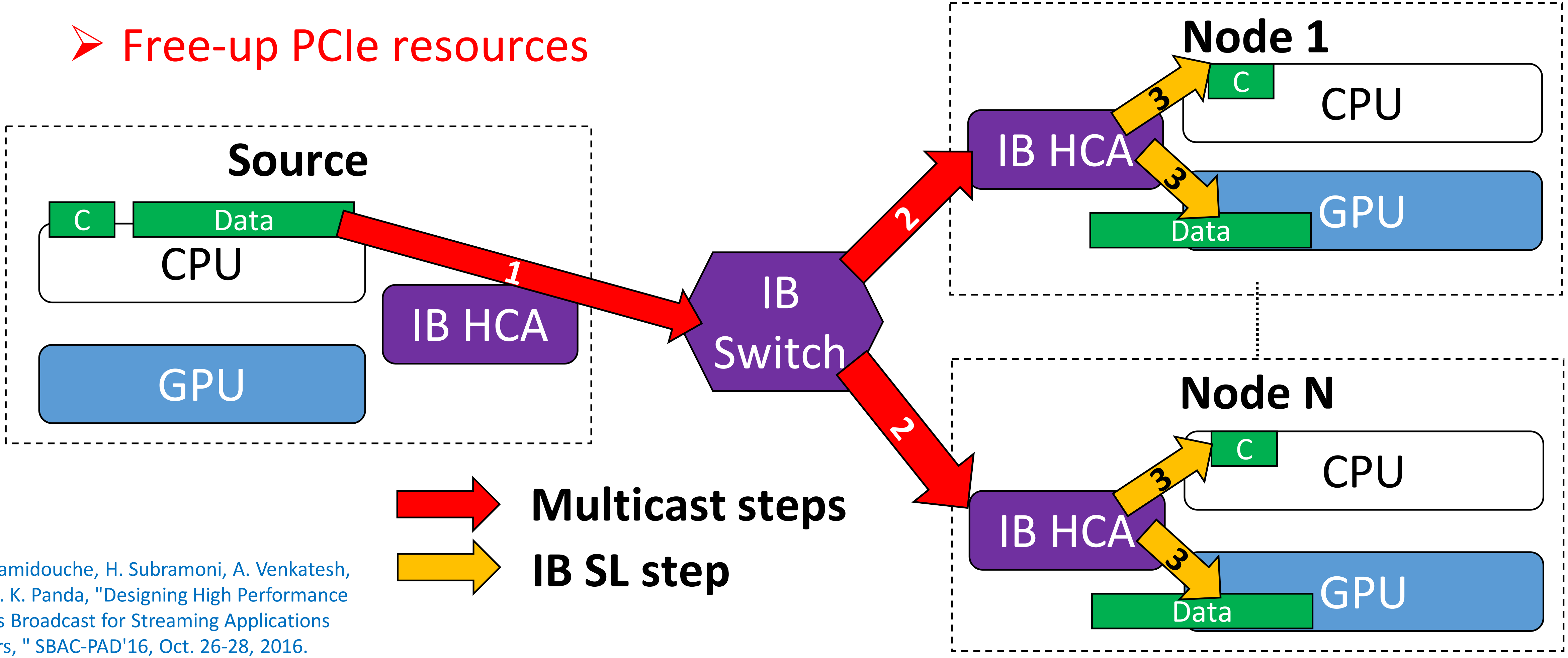
- **For GPU-resident data, using**
  - GPUDirect RDMA (GDR)
  - InfiniBand Hardware Multicast (IB-MCAST)
- **Overhead**
  - IB UD limit
  - GDR limit



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)

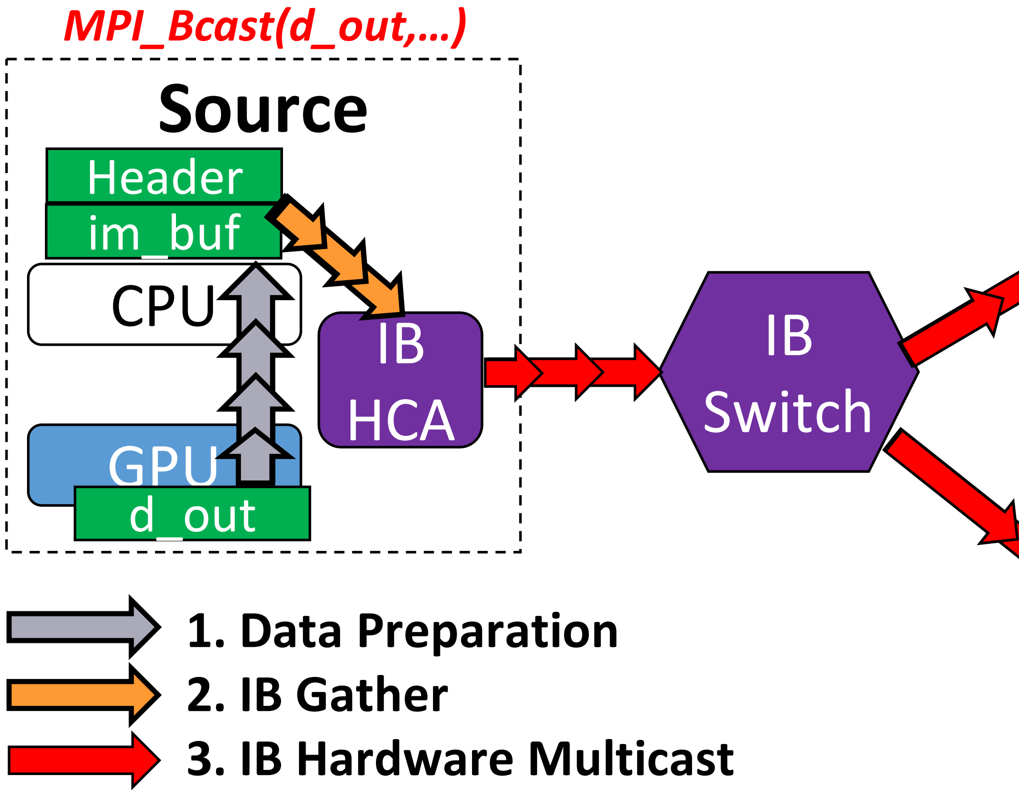
- **Heterogeneous Broadcast for streaming applications**
  - Free-up PCIe resources



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

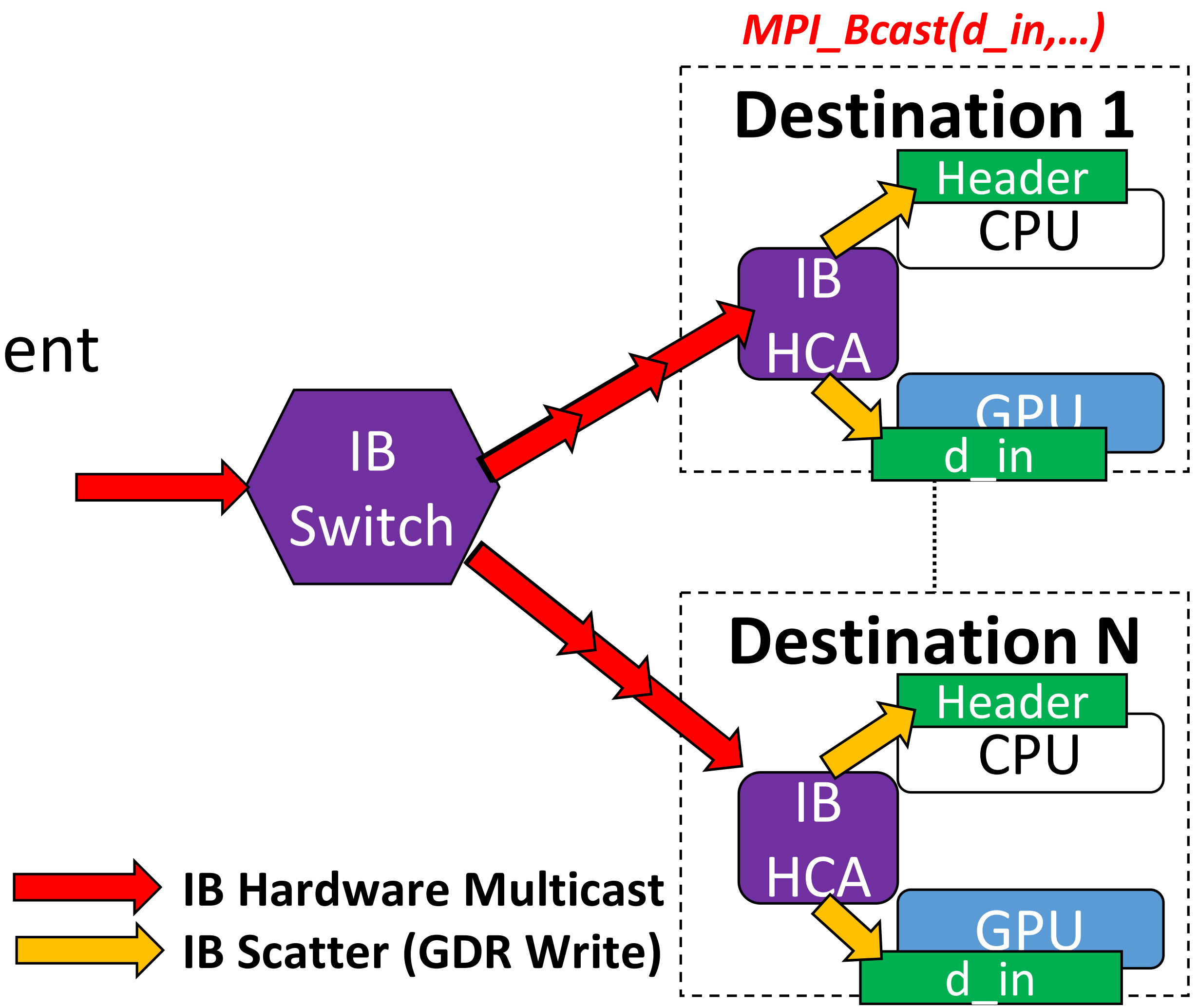
- **Preparing Intermediate buffer** (*im_buf*)
  - Page-locked (pinned) host buffer
    - ➢ Fast Device-Host data movement
  - Allocated at initialization phase
    - ➢ Low overhead, one time effort
- **Streaming data through host**
  - Fine-tuned chunked data
  - Asynchronous copy operations
  - ➢ Three-stage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

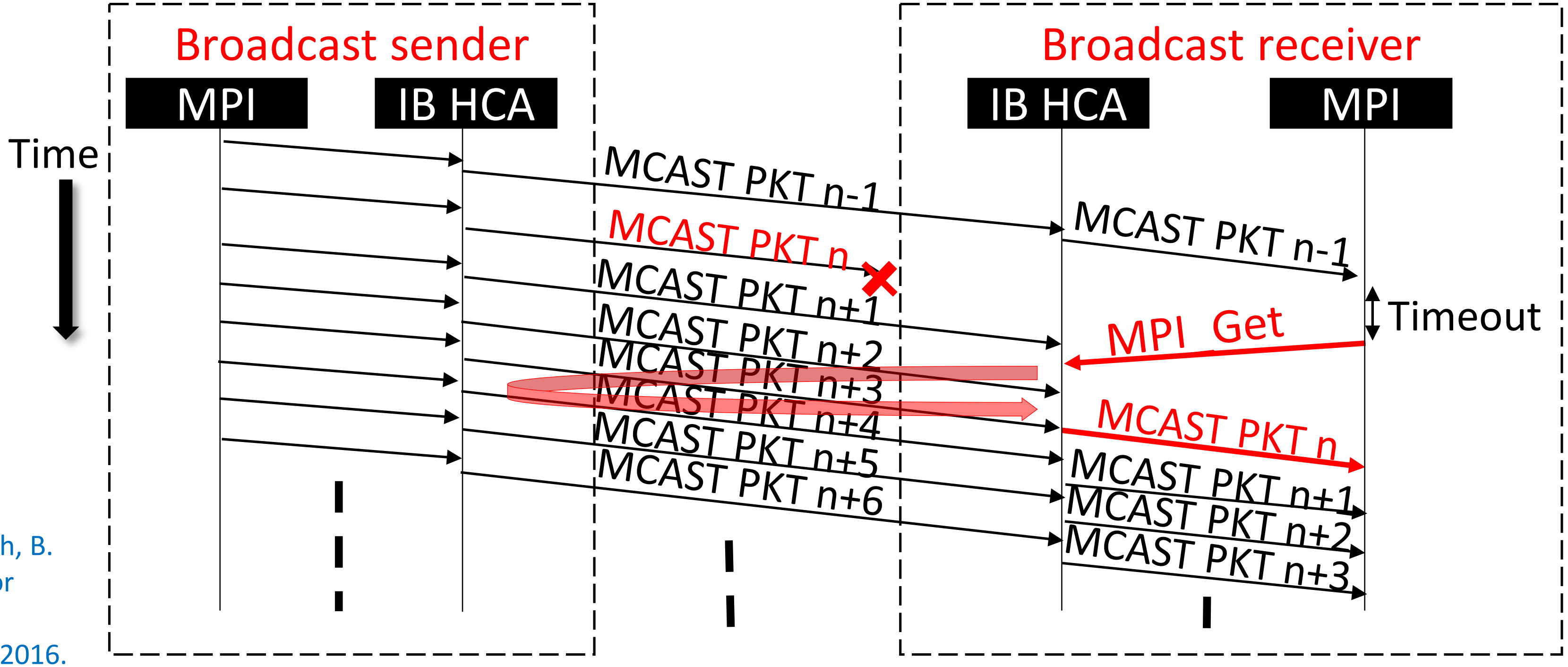
- **Zero-copy broadcast receive**
  - Pre-posted user buffer *(d_in)*
  - Avoids additional data movement
  - Leverages IB Scatter and GDR features
  - ➢ Low-latency
  - ➢ Free-up PCIe resources for applications



C.-H. Chu, X. Lu, A. A. Awan, H. Subramoni, B. Elton, D. K. Panda, "Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast," in IEEE Transactions on Parallel and Distributed Systems (TPDS), vol. 30, no. 3, pp. 575-588, 1 March 2019..

Efficient Reliability Support for IB-MCAST

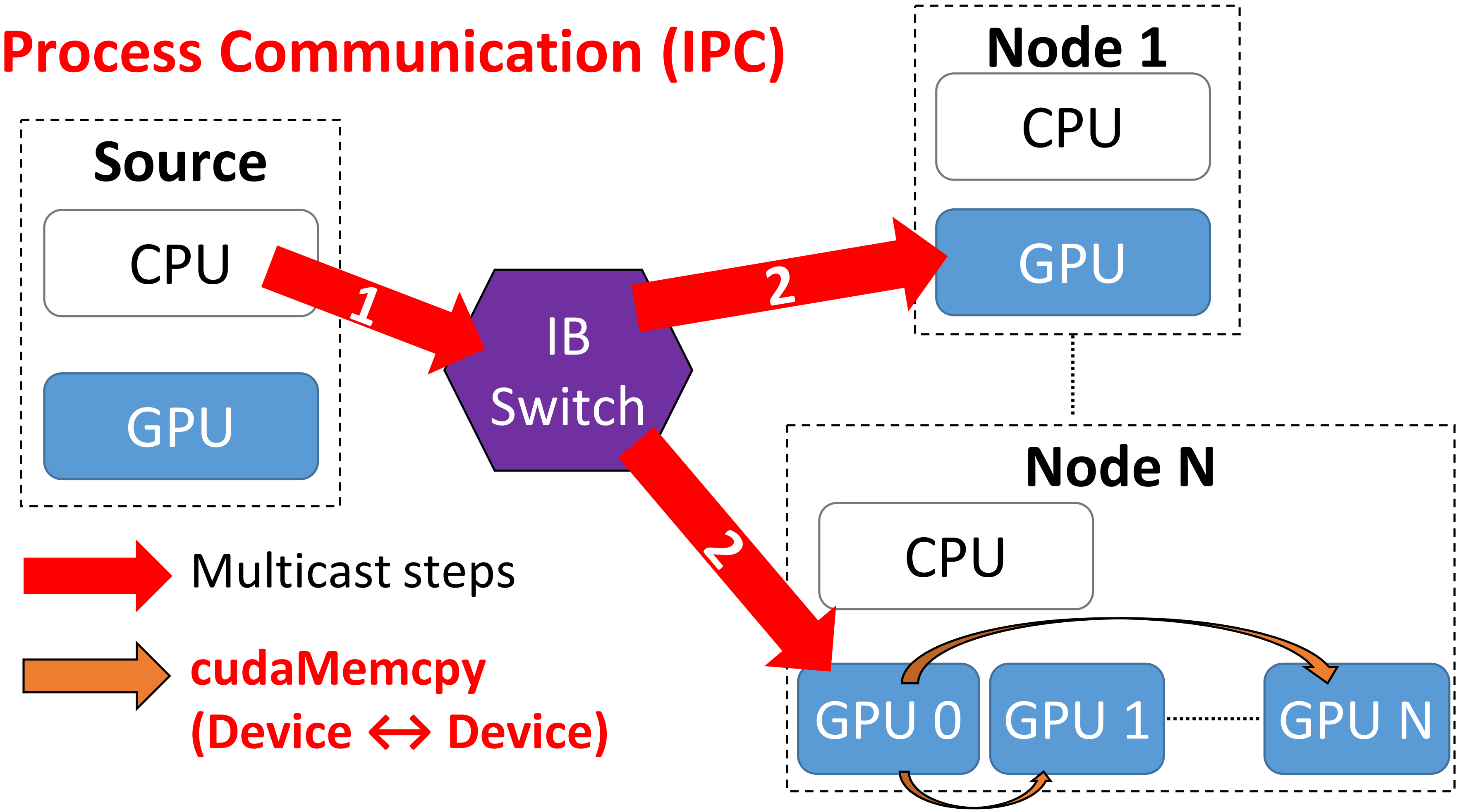
- **When a receiver experiences timeout (lost MCAST packet)**
  - Performs the RMA Get operation to the sender's backup buffer to retrieve lost MCAST packets
  - **Sender is not interrupted**



C.-H. Chu, K. Hamidouche, H. Subramoni, A. Venkatesh, B. Elton, and D. K. Panda, "Efficient Reliability Support for Hardware Multicast-based Broadcast in GPU-enabled Streaming Applications, " COMHPC Workshop @ SC, 2016.

Broadcast on Multi-GPU systems

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

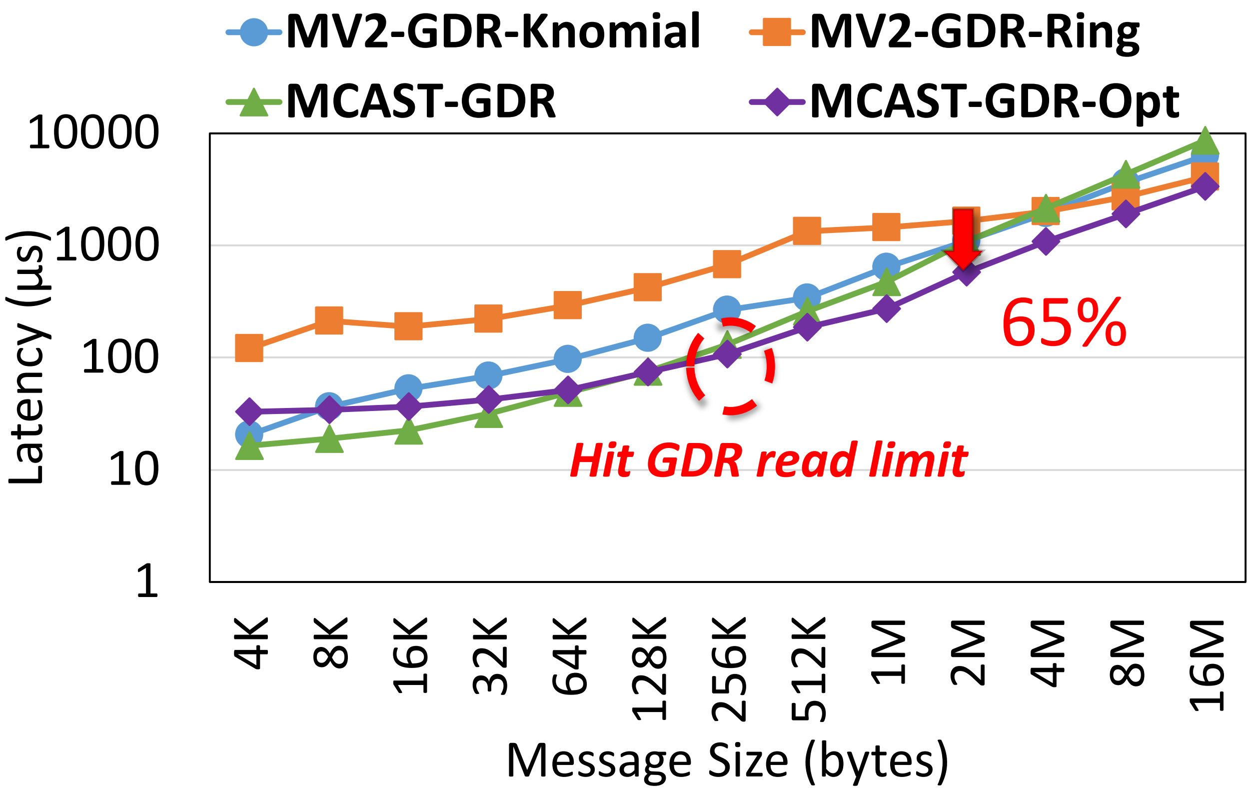


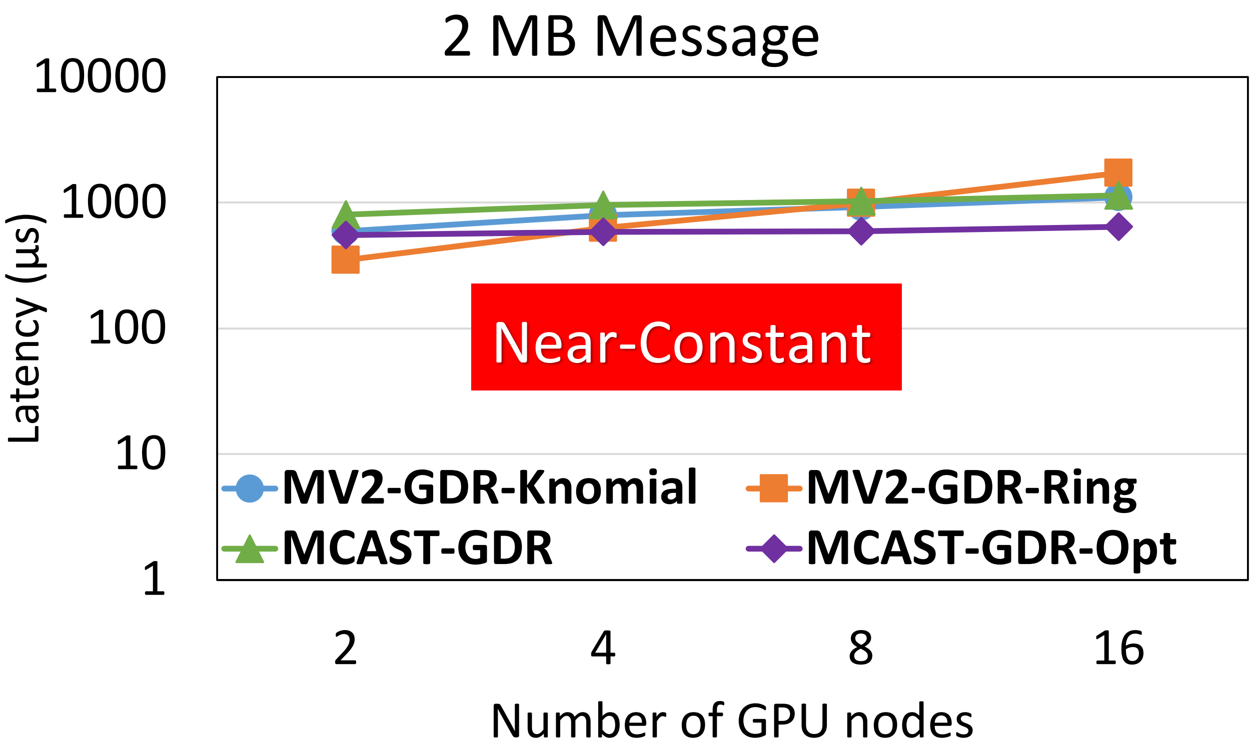
C.-H. Chu, K. Hamidouche, H. Subramoni, A. Venkatesh, B. Elton, and D. K. Panda, "Designing High Performance Heterogeneous Broadcast for Streaming Applications on GPU Clusters, " SBAC-PAD'16, Oct. 26-28, 2016.

Benchmark Evaluation

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

Lower is better

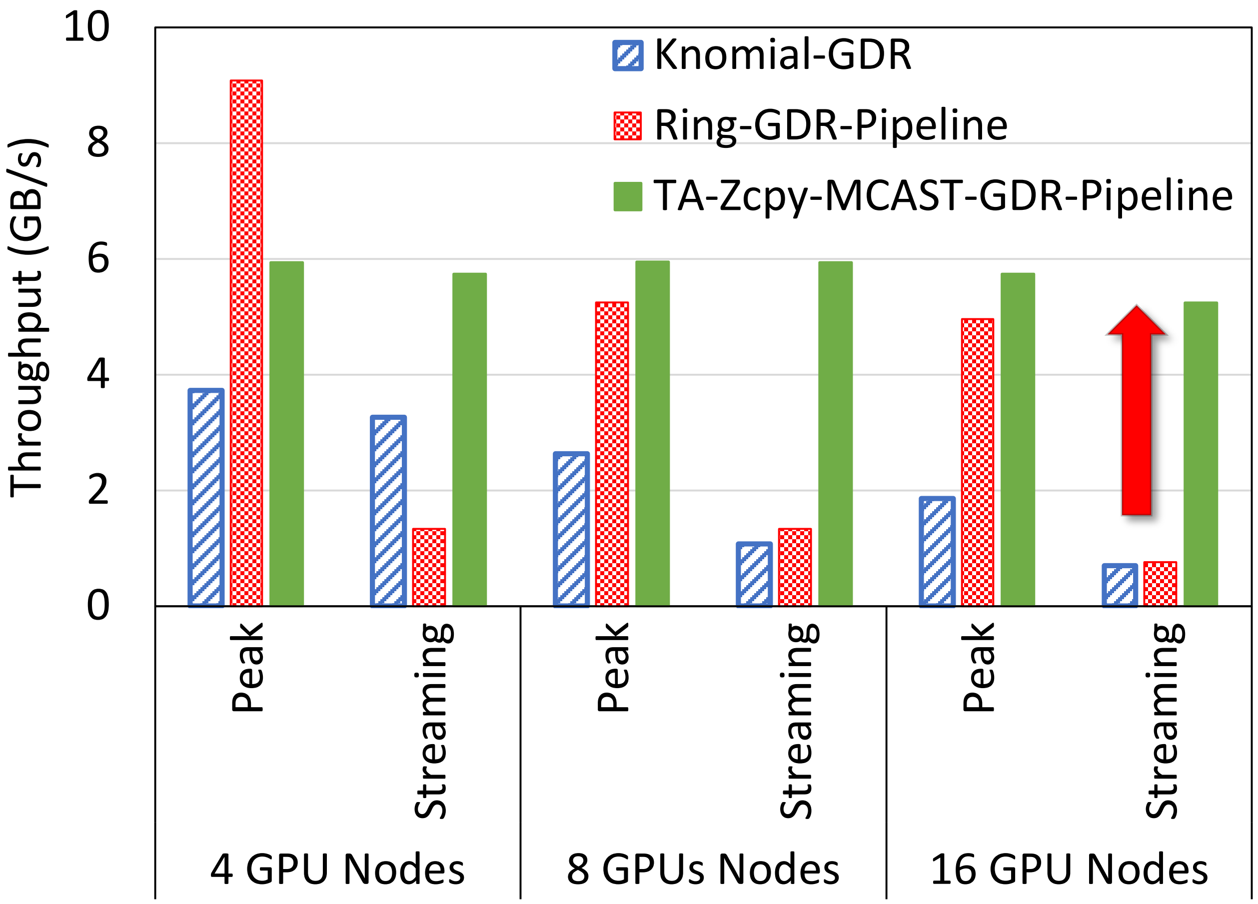


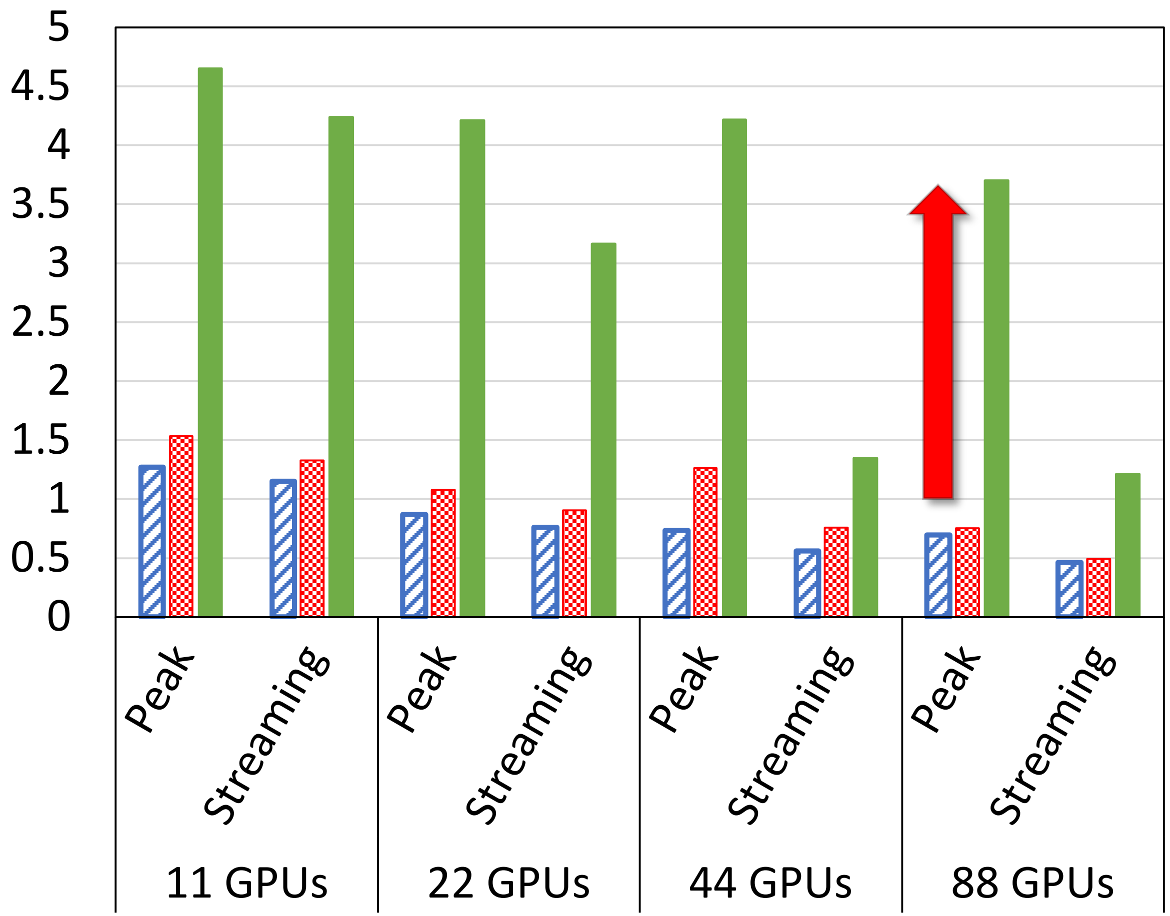


- **Provide near-constant latency over the system sizes**
- **Reduces up to 65% of latency for large messages**

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)



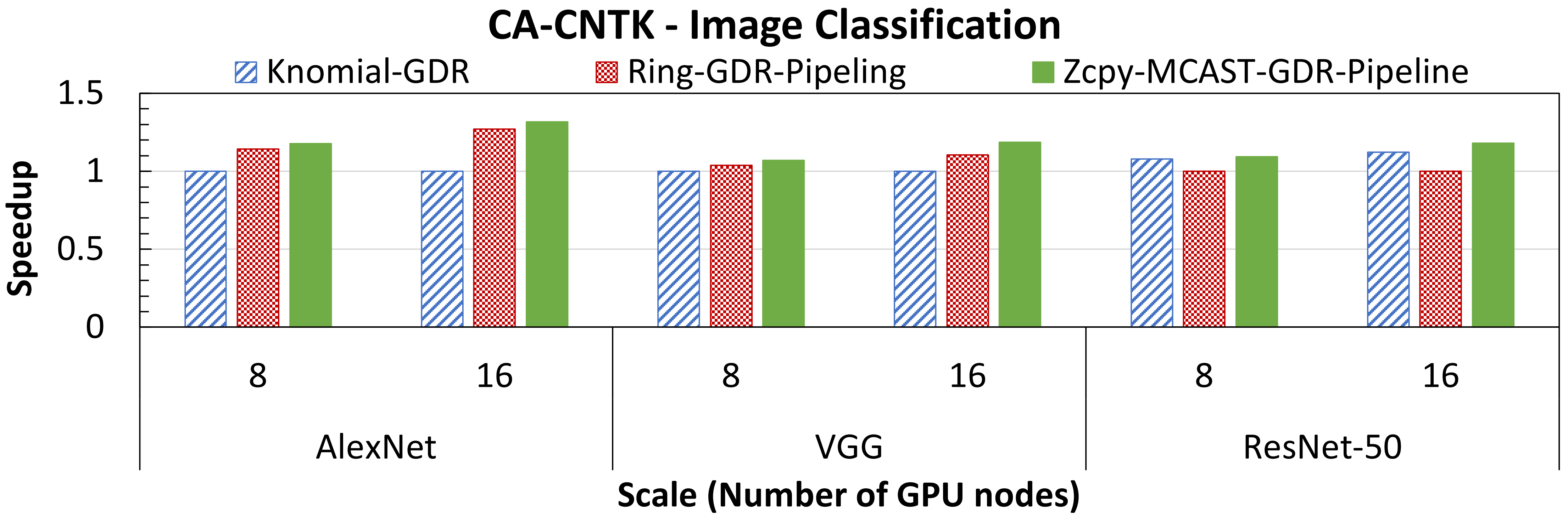


- **IB-MCAST + GDR + IPC-based MPI_Bcast schemes**
  - Stable high throughput compared to existing schemes

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," in IEEE Transactions on Parallel and Distributed Systems (TPDS), vol. 30, no. 3, pp. 575-588, 1 March 2019..

Outline

- Introduction
- Advanced Broadcast Designs in MVAPICH2-GDR
- 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

➢ **These features are included since MVAPICH2-GDR 2.3**

  ➢ http://mvapich.cse.ohio-state.edu/
  ➢ http://mvapich.cse.ohio-state.edu/userguide/gdr/

THE OHIO STATE UNIVERSITY

Thank You!

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

MVAPICH

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

Network Based Computing Laboratory

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