MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

THE OHIO STATE UNIVERSITY

High-Performance Broadcast for Streaming and Deep Learning

Ching-Hsiang Chu

chu.368@osu.edu

Department of Computer Science and Engineering

The Ohio State University

Outline

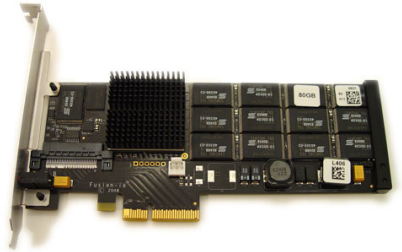
- Introduction
- Proposed Designs in MVAPICH2-GDR
- Performance Evaluation
- Concluding Remarks

Trends in Modern HPC Architecture

Multi-core Processors

High Performance Interconnects – InfiniBand **(IB),** Omni-Path
< 1 µsec latency, 100 Gbps Bandwidth>

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

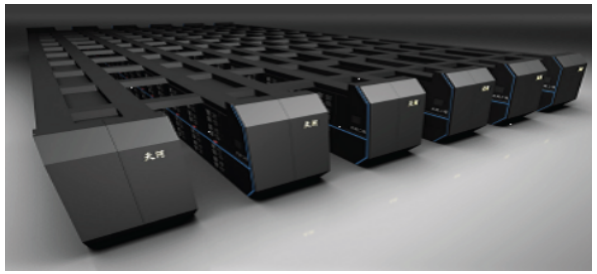


SSD, NVMe-SSD, NVRAM

- **Multi-core/many-core technologies**
- **High Performance Interconnects**
- **Accelerators/Coprocessors are becoming common in high-end systems**
- **High Performance Storage and Compute devices**

Sunway TaihuLight

K - Computer



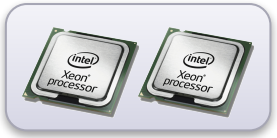
Tianhe – 2

Titan

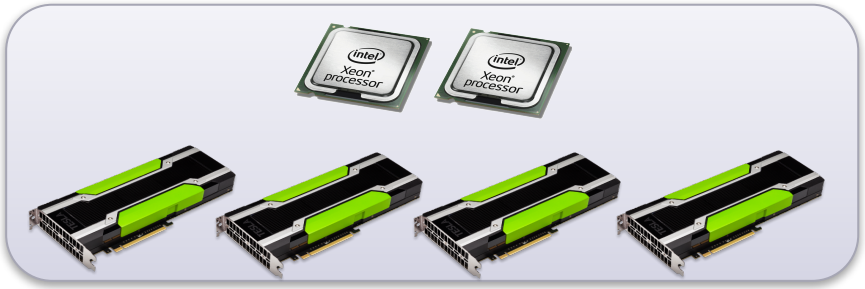
Architectures for Deep Learning (DL)

Past and Current Trend

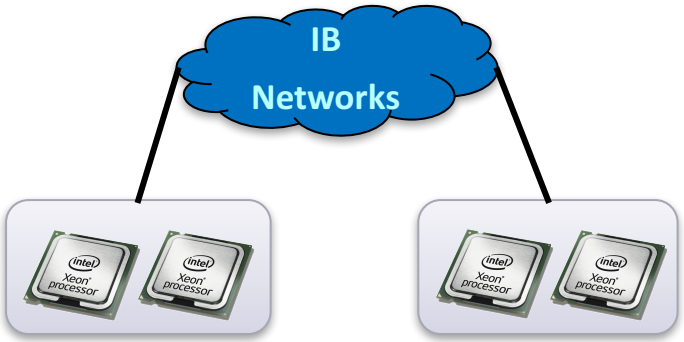
Multi-core CPUs within a node



Multi-core CPUs + Multi-GPU within a node



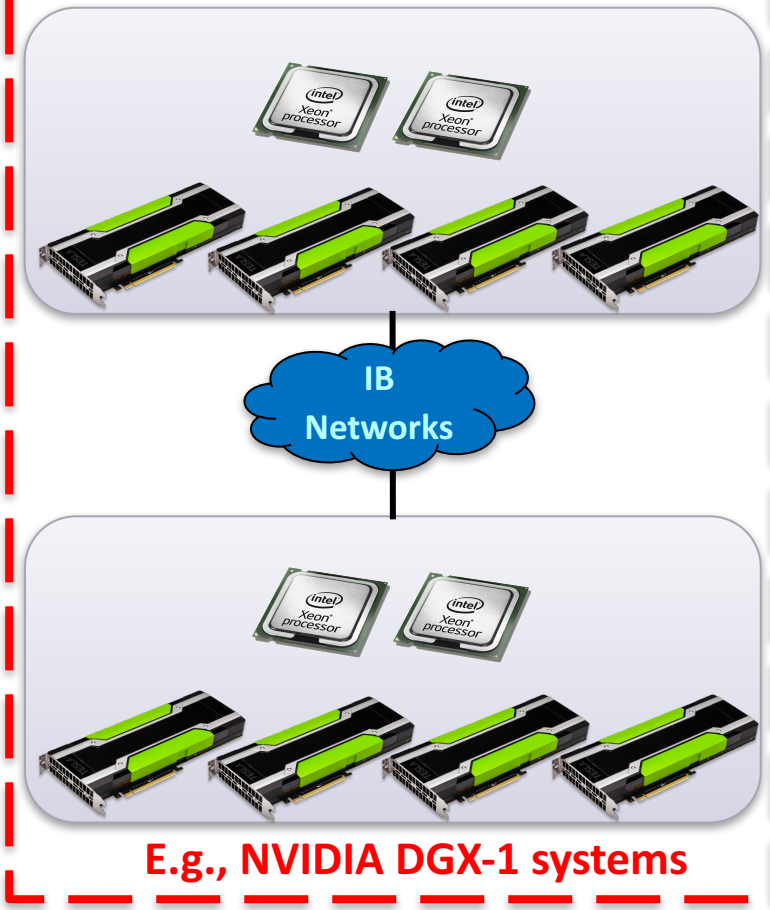
Multi-core CPUs across nodes



Multi-core CPUs + Single GPU across nodes

Near-future

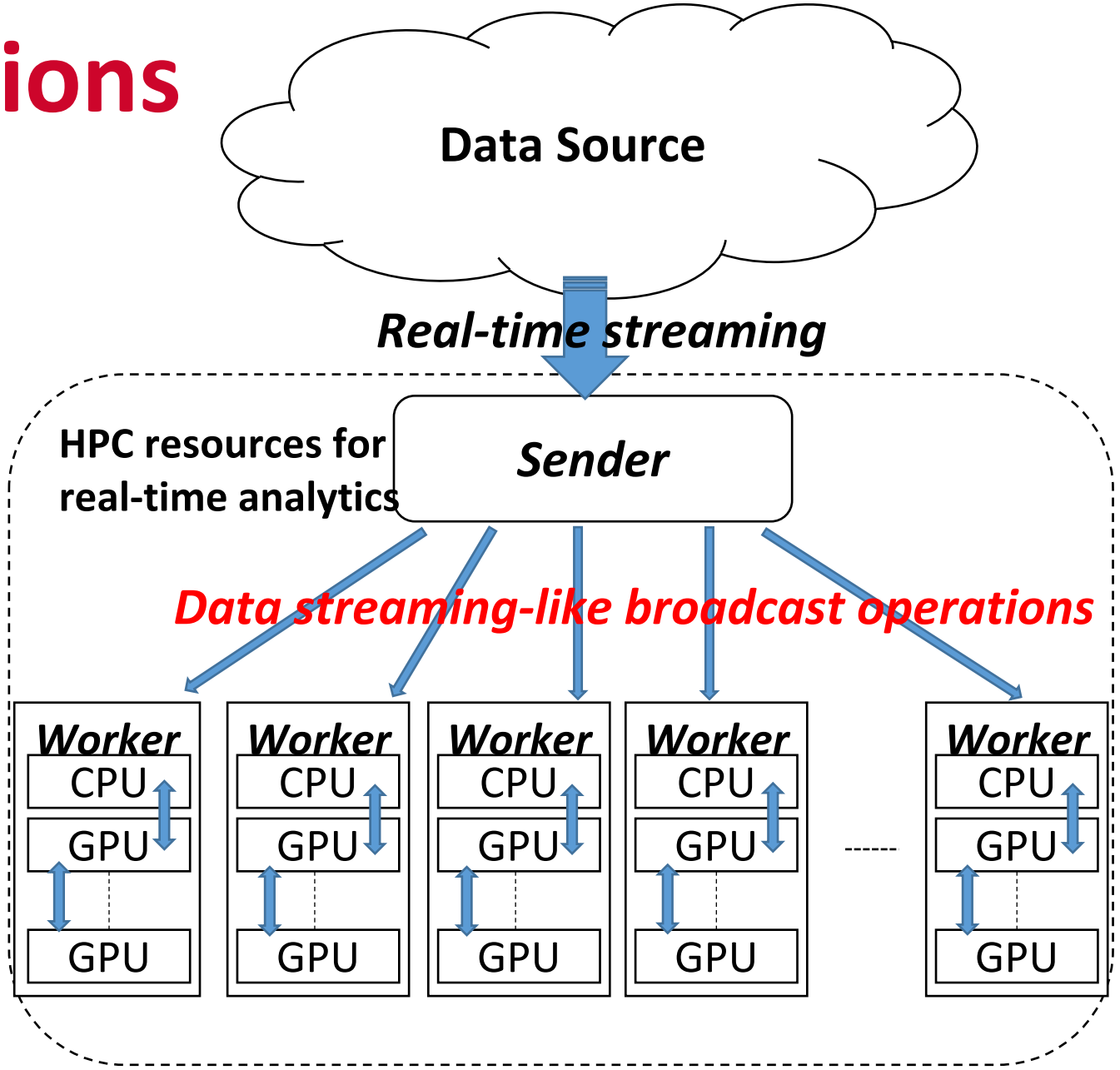
Multi-core CPUs + Multi-GPU across nodes



E.g., NVIDIA DGX-1 systems

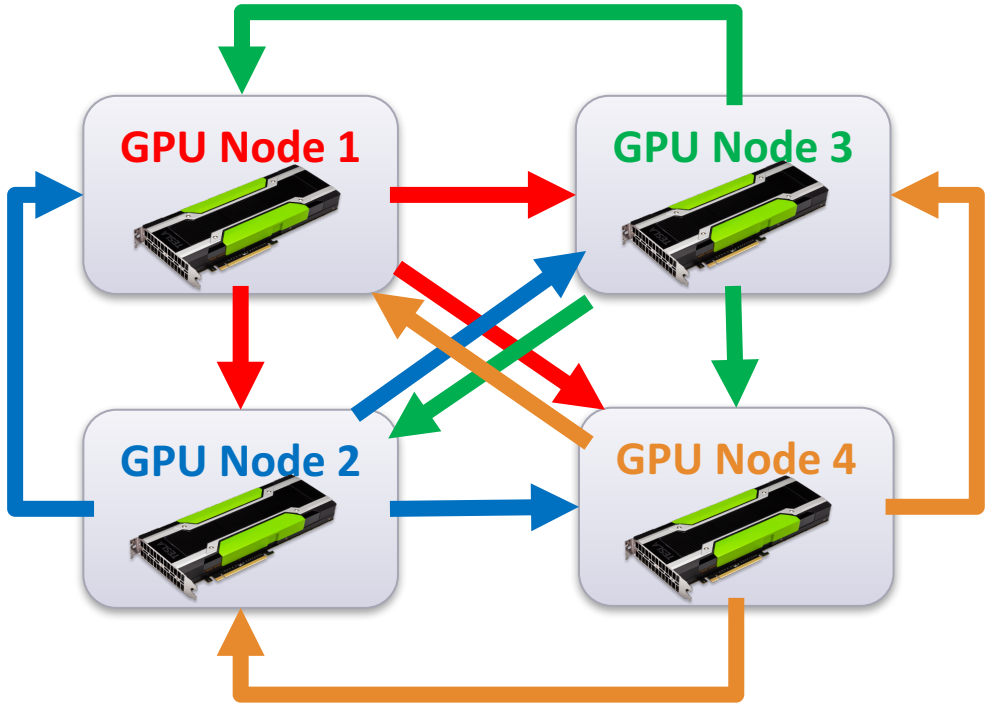
Streaming Applications

- **Streaming applications on HPC systems**
  1. **Communication (MPI)**
     - Broadcast-type operations
  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**
- **Challenges**
  - High computation-communication **overlap**
  - Good **scalability** for upcoming large-scale GPU clusters
  - No application-level modification

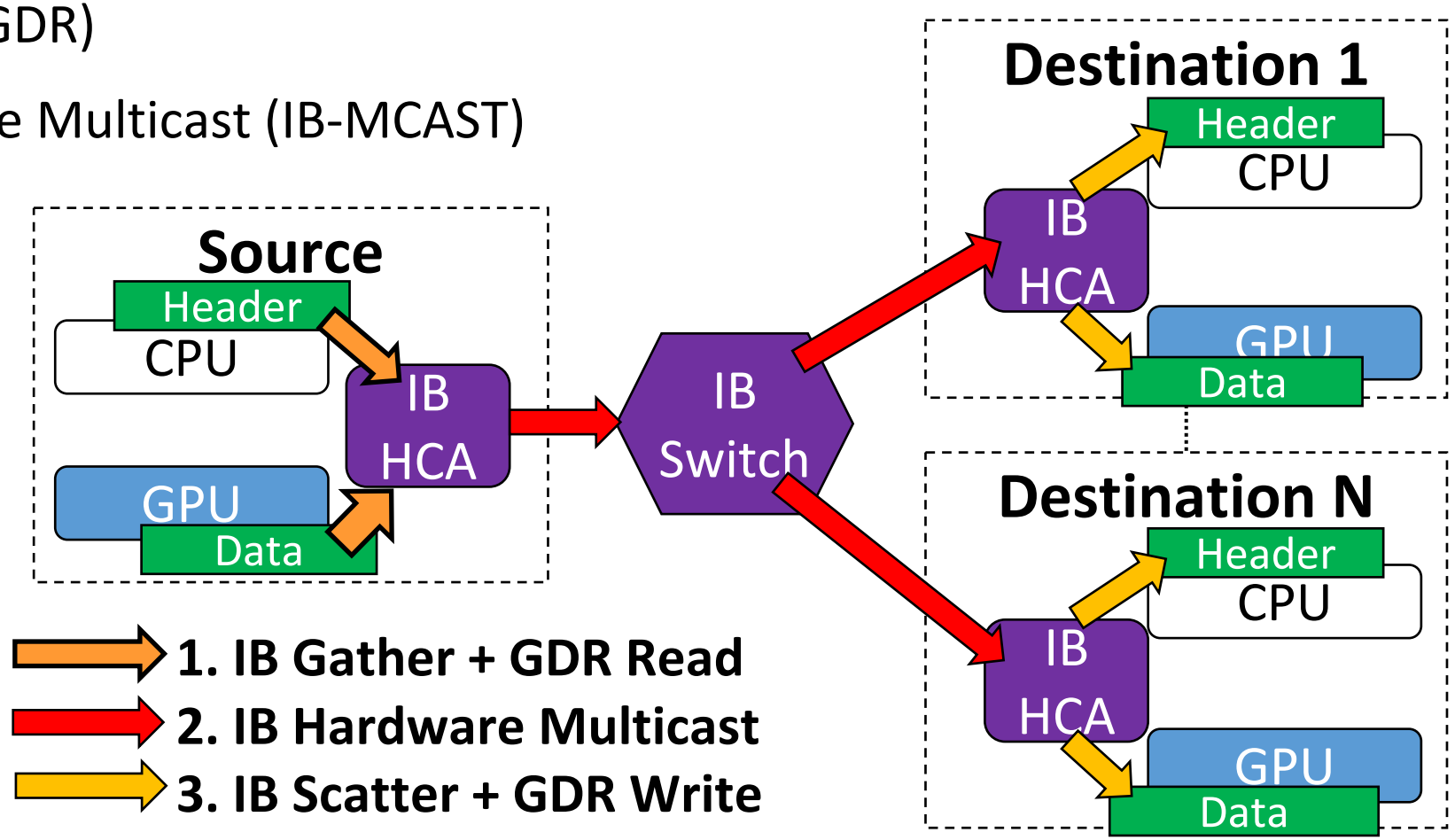


Outline

- Introduction
- Proposed Designs in MVAPICH2-GDR
- Performance Evaluation
- Concluding Remarks

Hardware Multicast-based Broadcast

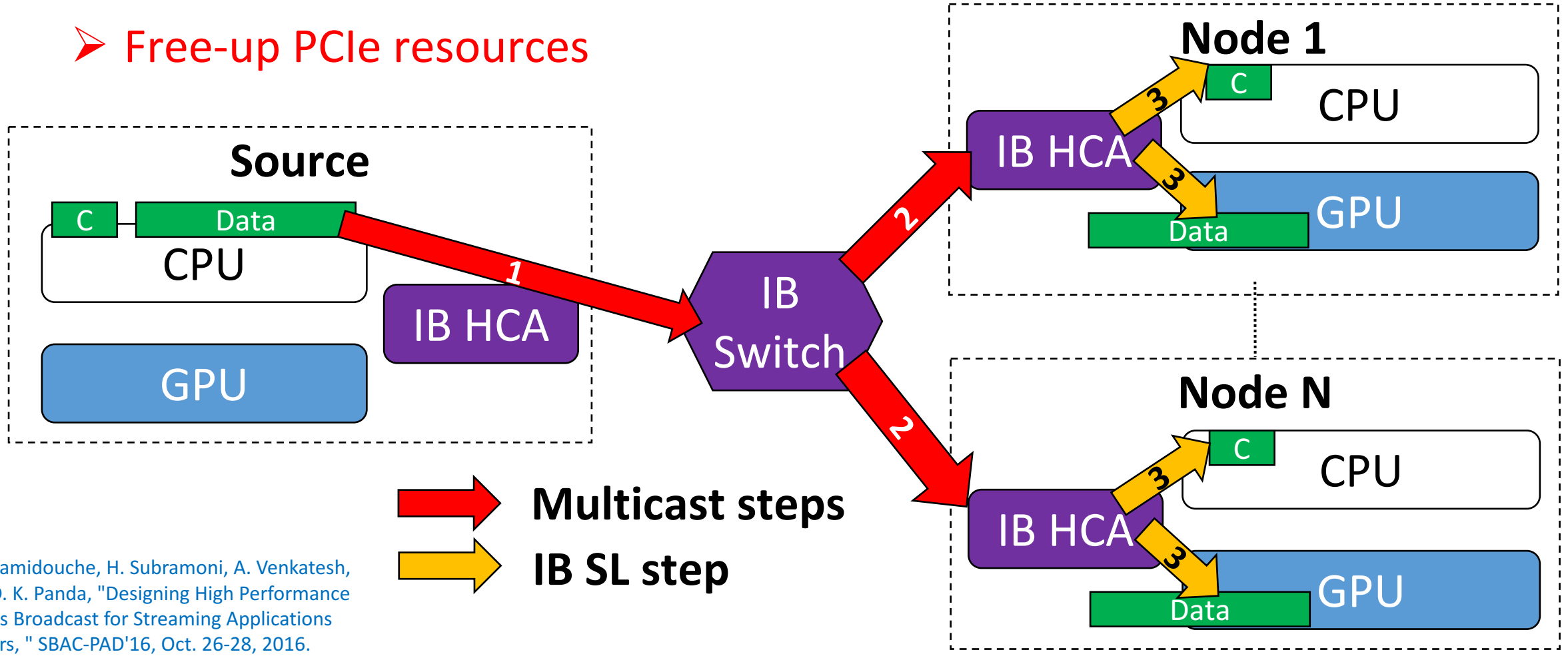
- **For GPU-resident data, using**
  - GPUDirect RDMA (GDR)
  - InfiniBand Hardware Multicast (IB-MCAST)
- **Overhead**
  - IB UD limit
  - GDR limit



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)

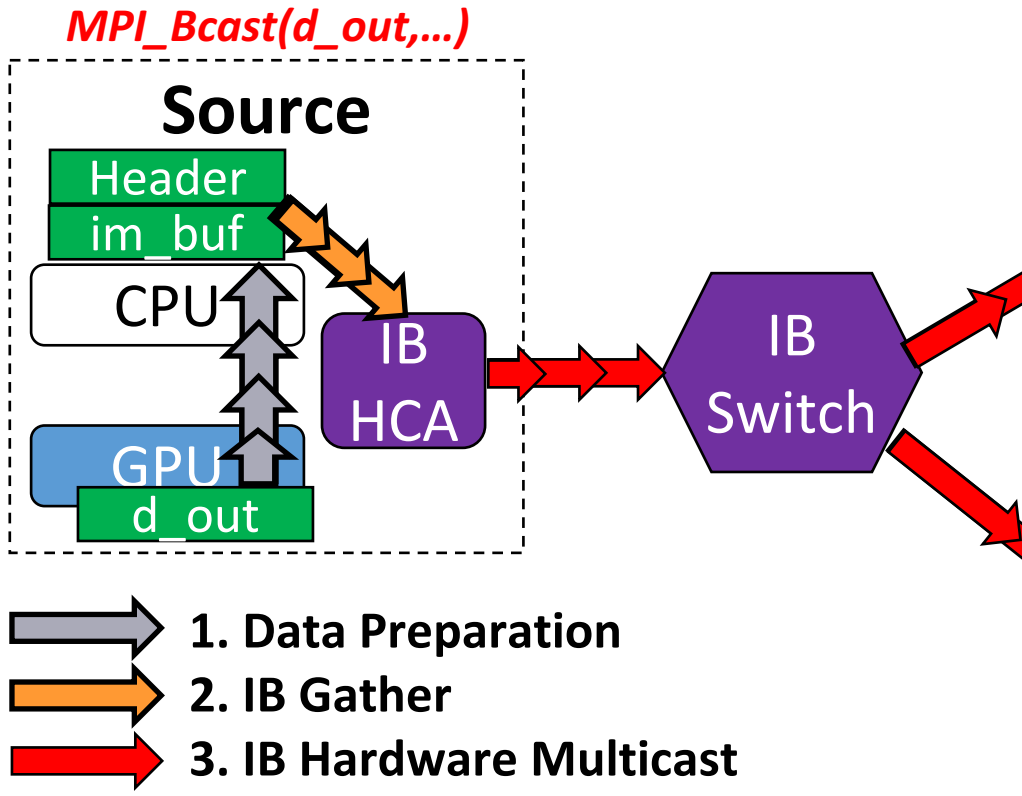
- **Heterogeneous Broadcast for streaming applications**
  - ➢ Free-up PCIe resources



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

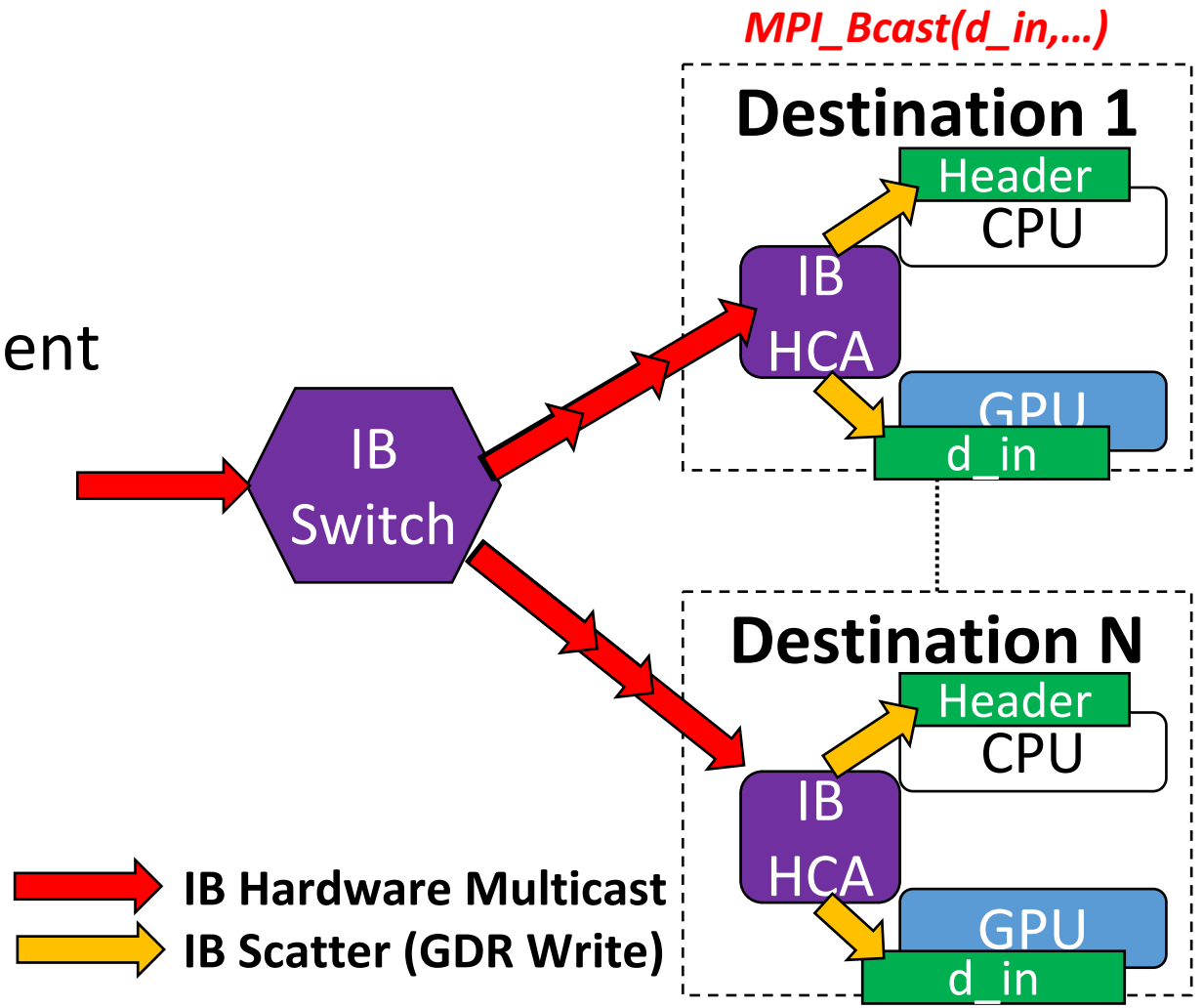
- **Preparing Intermediate buffer** (*im_buf*)
  - Page-locked (pinned) host buffer
    - ➢ Fast Device-Host data movement
  - Allocated at initialization phase
    - ➢ Low overhead
- **Streaming data through host**
  - Fine-tuned chunked data
  - Asynchronous copy operations
  - ➢ Three-stage pipeline



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

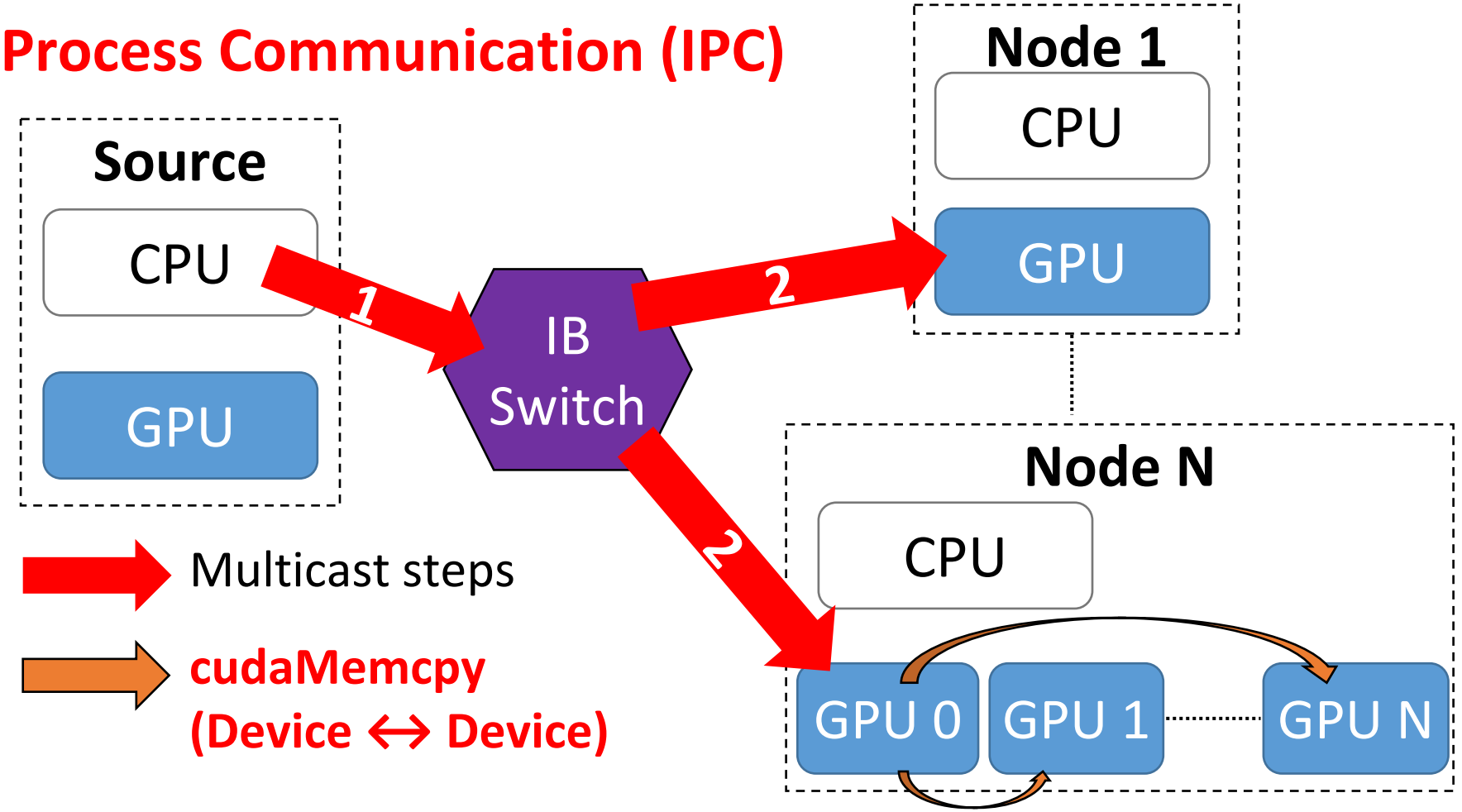
- **Zero-copy broadcast receive**
  - Pre-posted user buffer *(d_in)*
  - Avoids additional data movement
  - Leverages IB Scatter and GDR features
  - ➢ Low-latency
  - ➢ Free-up PCIe resources for applications



C.-H. Chu, X. Lu, A. A. Awan, H. Subramoni, 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.

Broadcast on Multi-GPU systems

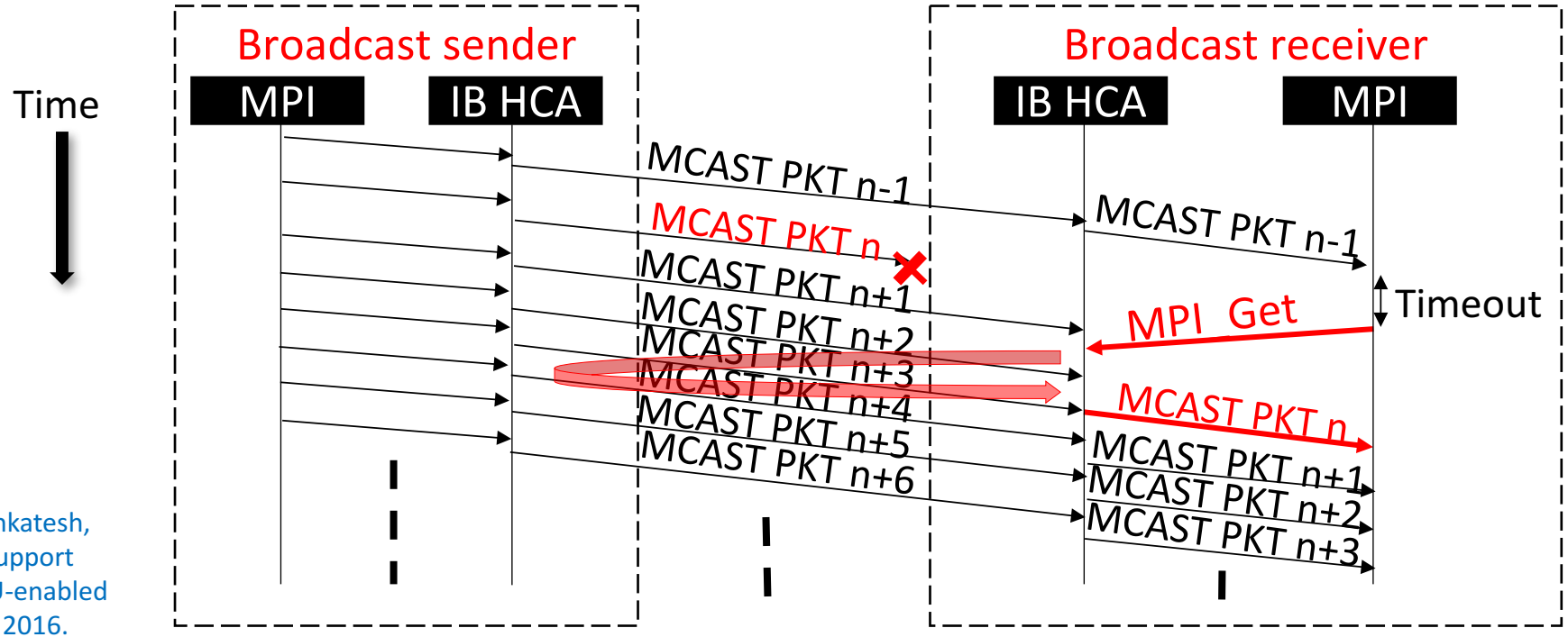
- **Proposed Intra-node Topology-Aware Broadcast**
  - **CUDA InterProcess Communication (IPC)**



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.

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, 2016.

Outline

- Introduction
- Proposed Designs in MVAPICH2-GDR
- Performance Evaluation
- Concluding Remarks

Experimental Environments

- **Ohio State University (OSU) Micro-Benchmark (OMB)**

  http://mvapich.cse.ohio-state.edu/benchmarks/

  - osu_bcast - MPI_Bcast Latency Test
  - osu_bcast_streaming – MPI_Bcast streaming Test

- **Deep learning framework: CUDA-Aware Microsoft Cognitive Toolkit (CA-CNTK)***

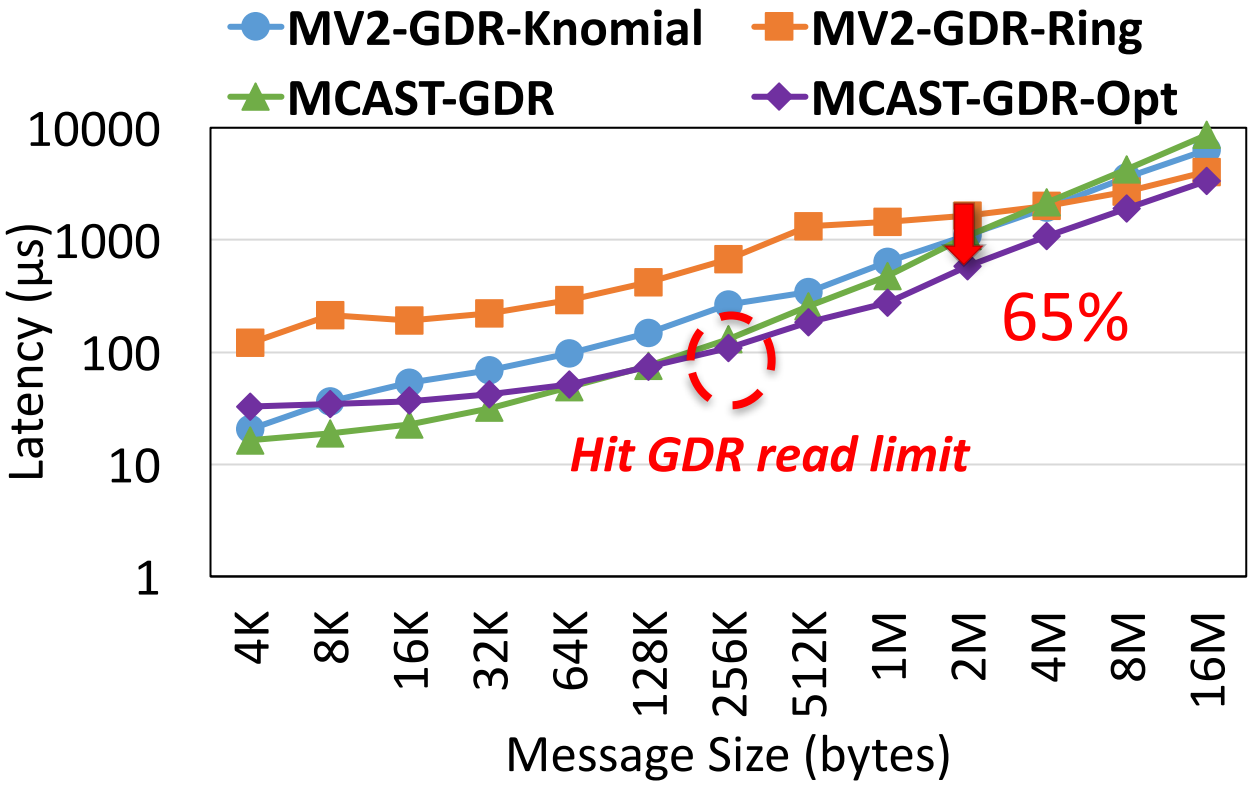
  - AlexNet and VGG models with ImageNet dataset

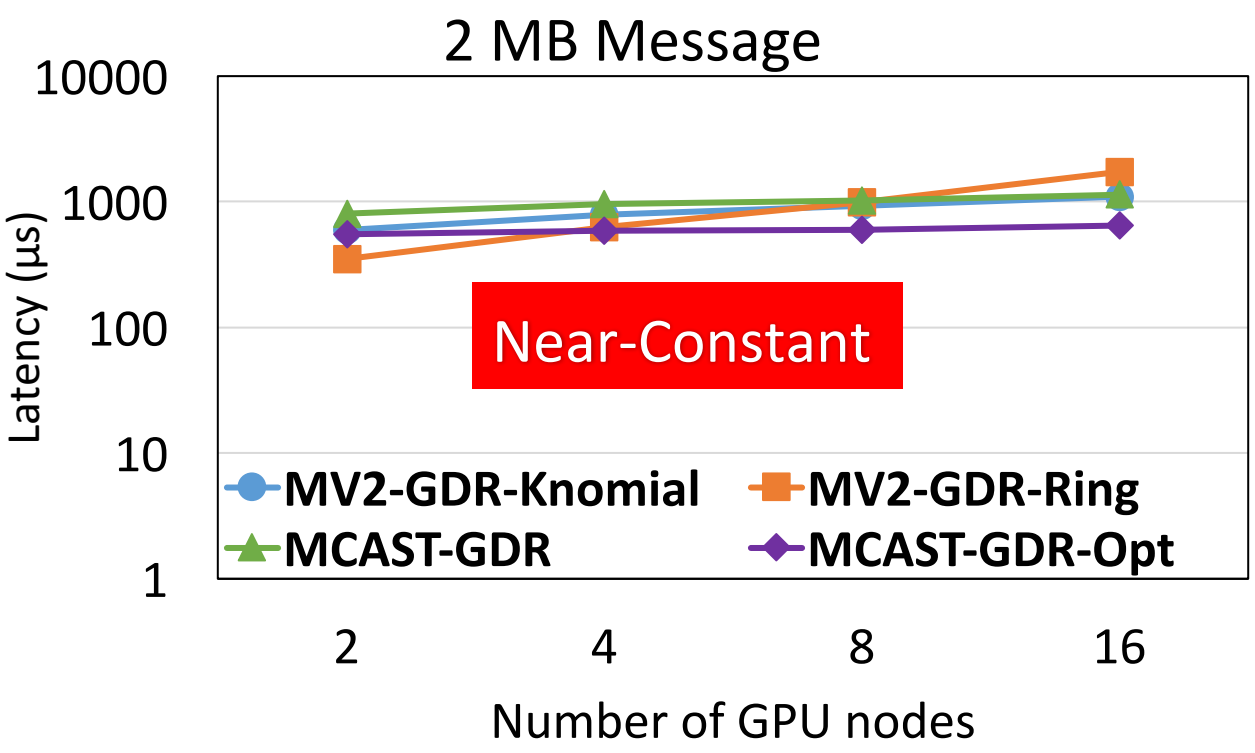
*D. S. Banerjee, K. Hamidouche and D. K. Panda, "Re-Designing CNTK Deep Learning Framework on Modern GPU Enabled Clusters," IEEE CloudCom, Luxembourg City, 2016, pp. 144-151.

Benchmark Evaluation

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

Lower is better

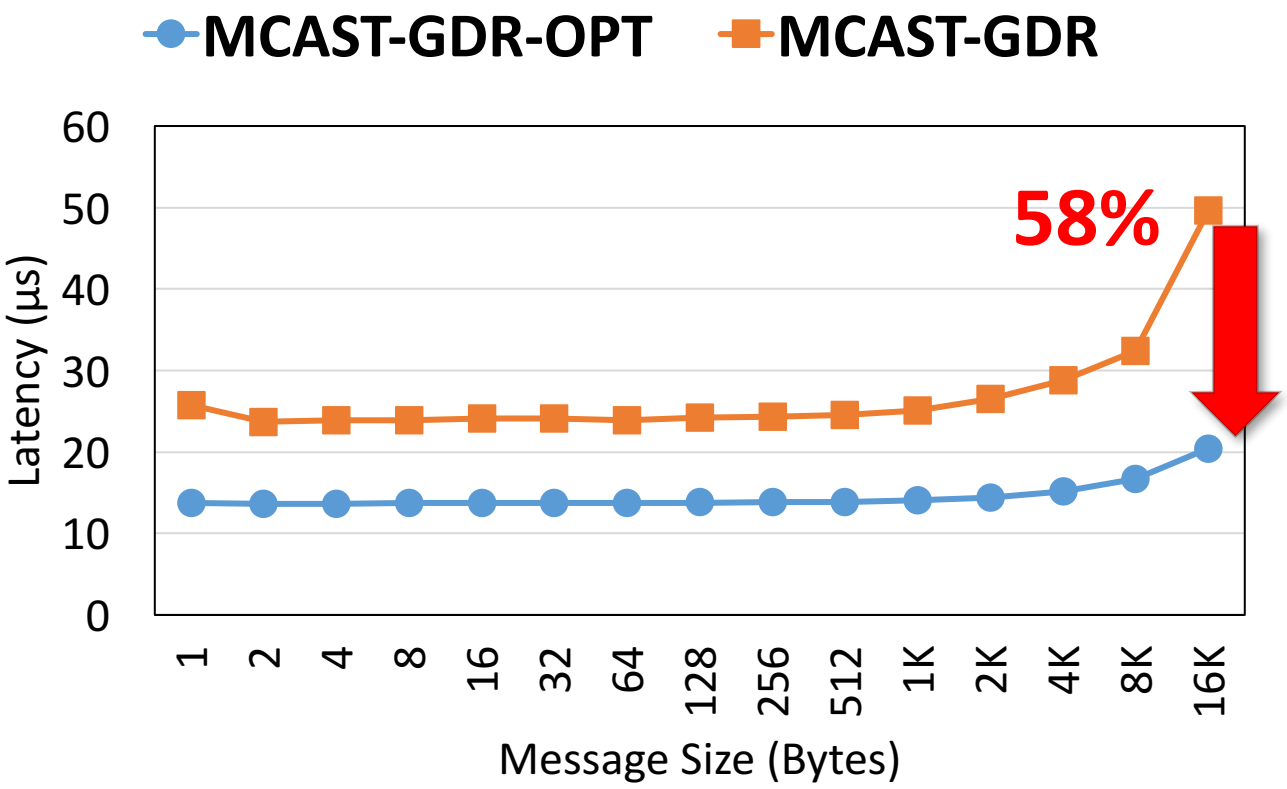


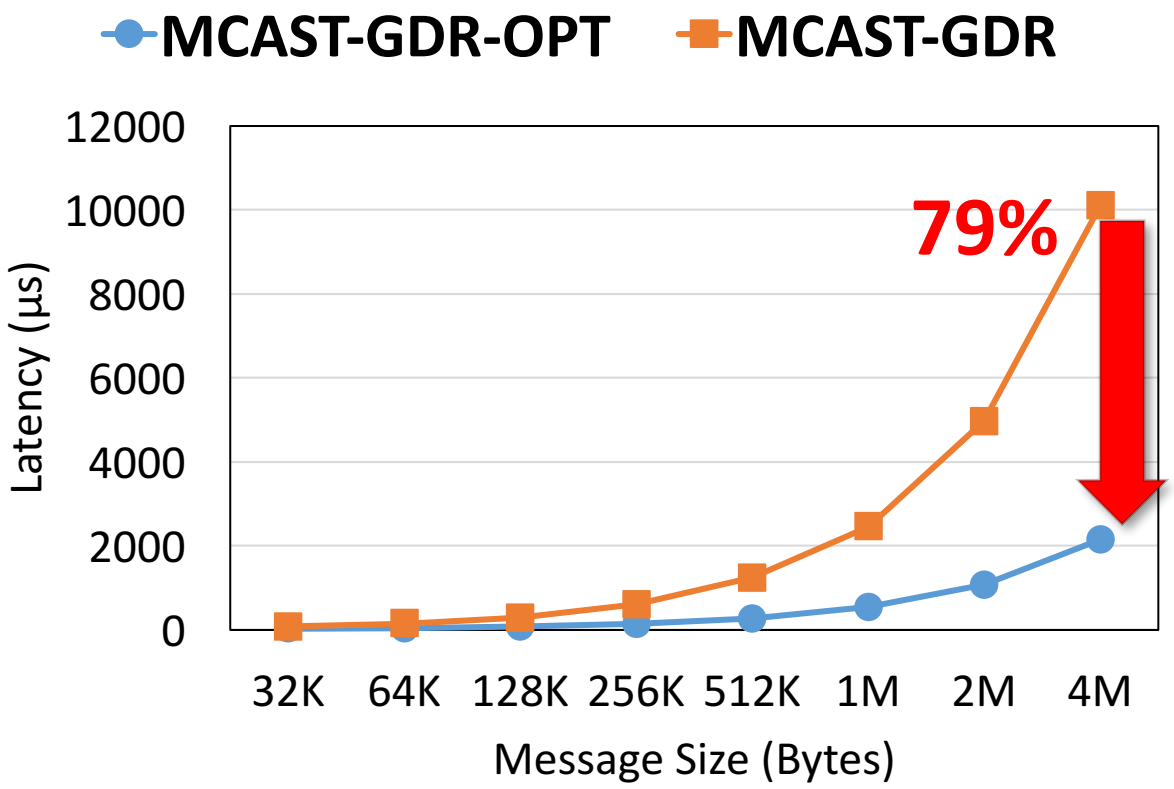


- Provide near-constant latency over the system sizes
- Reduces up to 65% of latency for large messages

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 Benchmark @ CSCS (88 GPUs)



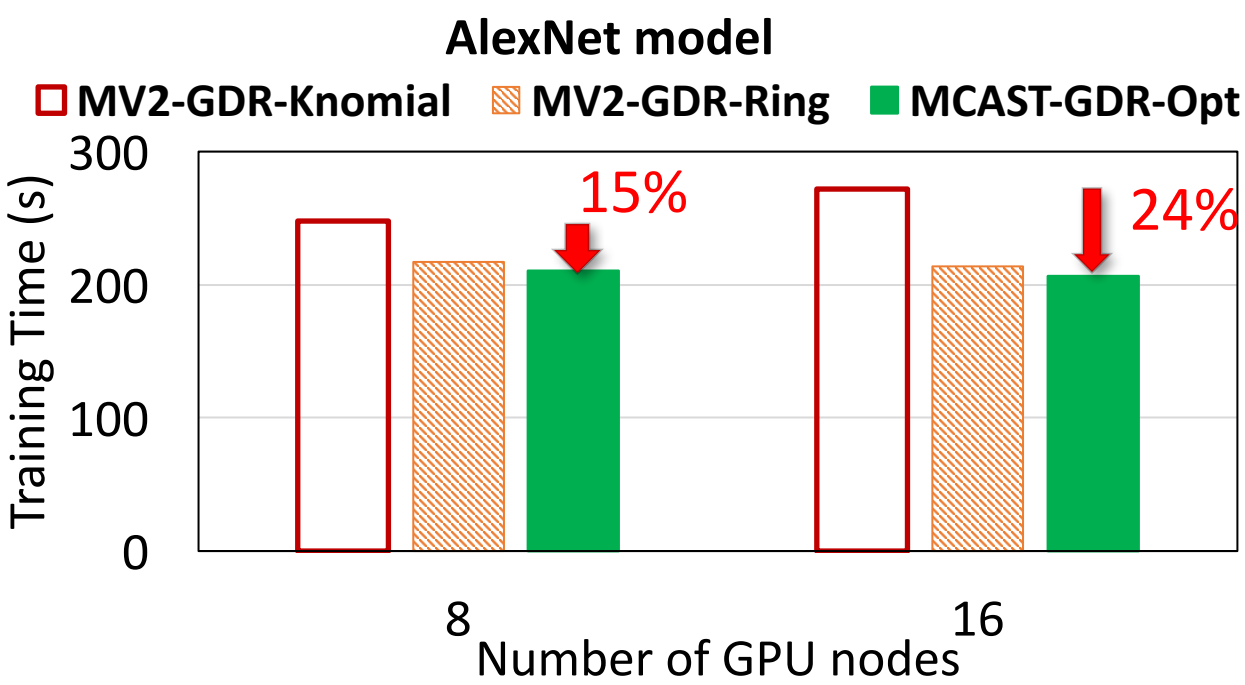


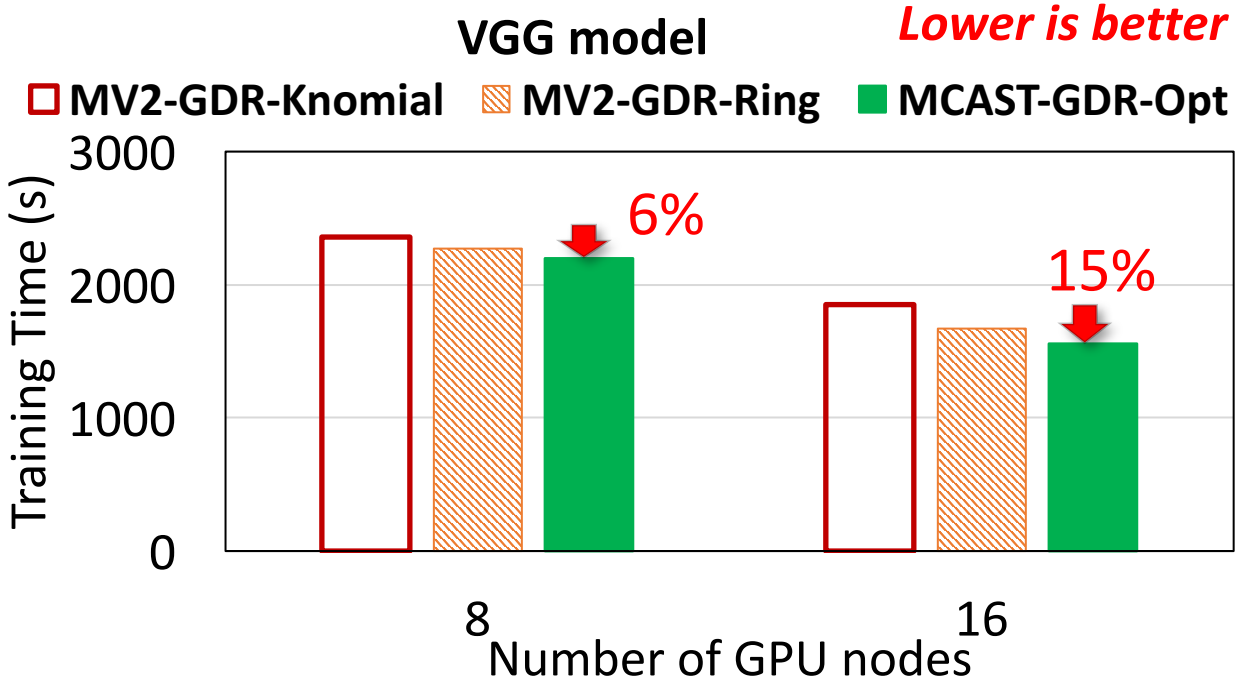
- **IB-MCAST + GDR + Topology-aware IPC-based schemes**
  - Up to **58% and 79% reduction** for small and large messages

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.

Deep Learning Frameworks

- **@ RI2 cluster, 16 GPUs, 1 GPU/node:**
  - CUDA-Aware Microsoft Cognitive Toolkit (CA-CNTK) without modification





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

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 in MVAPICH2-GDR 2.3a**

- ➢ http://mvapich.cse.ohio-state.edu/
- ➢ http://mvapich.cse.ohio-state.edu/userguide/gdr/2.3a/

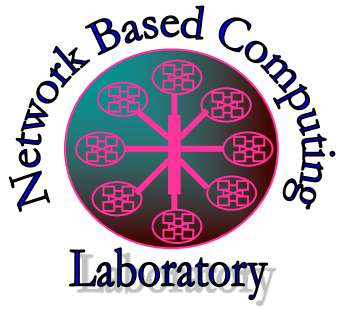
THE OHIO STATE UNIVERSITY

Thank You!

Ching-Hsiang Chu

chu.368@osu.edu

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



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

[1] 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.

[2] 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.

[3] 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*, 2016.

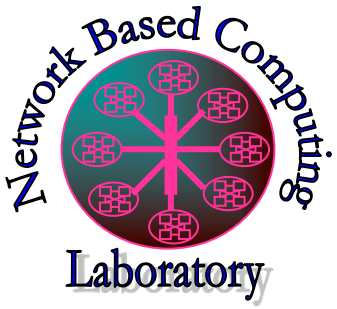
[4] C.-H. Chu, X. Lu, A. A. Awan, H. Subramoni, B. Elton and D. K. Panda, "**Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast** , " *submitted to IEEE TPDS. (Under review)*

THE OHIO STATE UNIVERSITY

Thank You!

- **Join us for more tech talks from MVAPICH2 team**
  - http://mvapich.cse.ohio-state.edu/conference/677/talks/

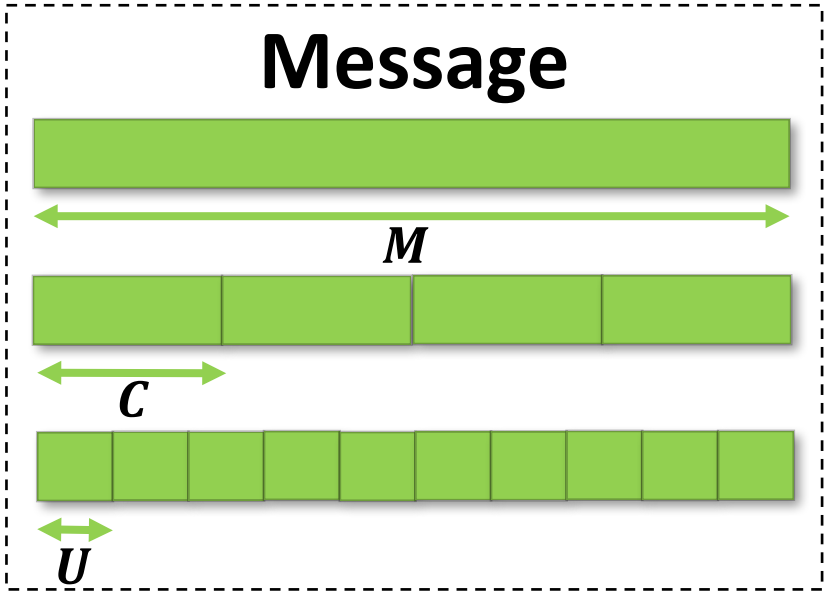
The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

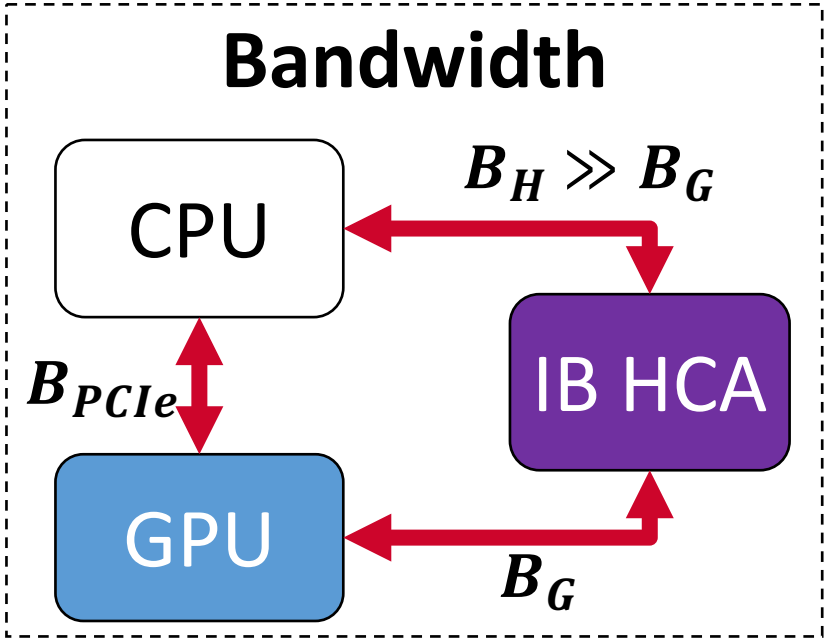


Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

Evaluation Parameters

Notation	Meaning	Unit
n	Number of processes	N/A
m	Number of broadcast sources	N/A
t_s	Set up time for sending data	sec
$t_o(n)$	Overhead for issuing an IB-MCAST packet	sec
M	Original message size	bytes
C	Size of a data chunk	bytes
U	Maximum Transmission Unit for IB-MCAST, provided by hardware manufacturer	bytes
B_H	Bandwidth of reading Host memory	bytes/sec
B_G	Bandwidth of reading GPU memory (NVIDIA GPUDirect RDMA)	bytes/sec
B_{PCIe}	PCIe Bandwidth between Host and GPU memory	bytes/sec





Ring-based Broadcast

- **Direct**

$$(n-1)\times\left(t_s+\frac{M}{B_G}\right)$$

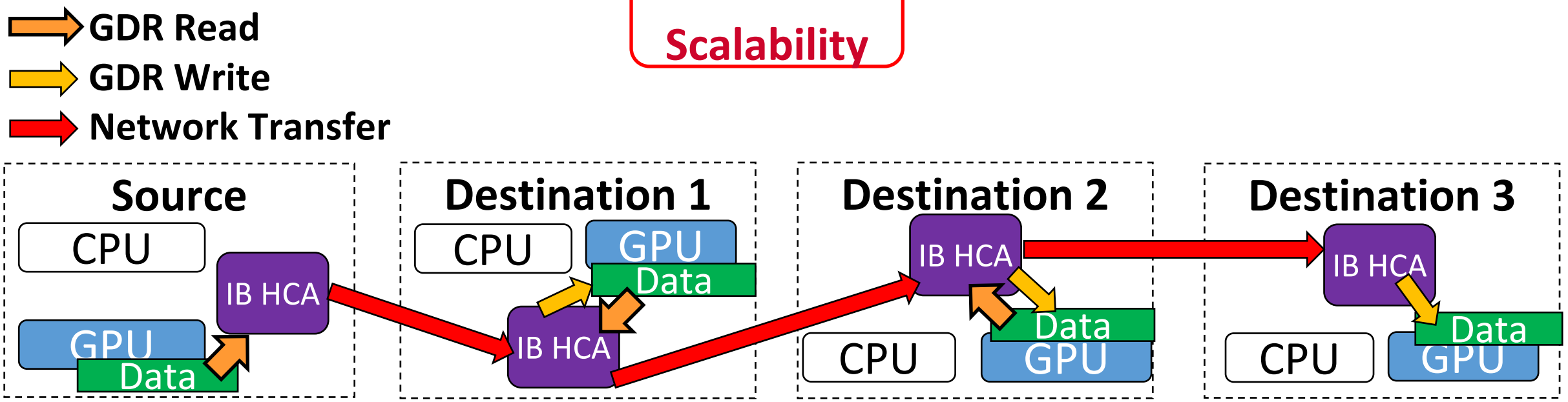
- **Pipeline**

$$\left[\frac{M}{C}+(n-2)\right]\times\left(t_s+\frac{C}{B_G}\right)$$

- **Staging**

$$\frac{M}{B_{PCIe}}+(n-1)\times\left(t_s+\frac{M}{B_H}\right)$$

Poor Scalability



K-nomial-based Broadcast

- **Direct**

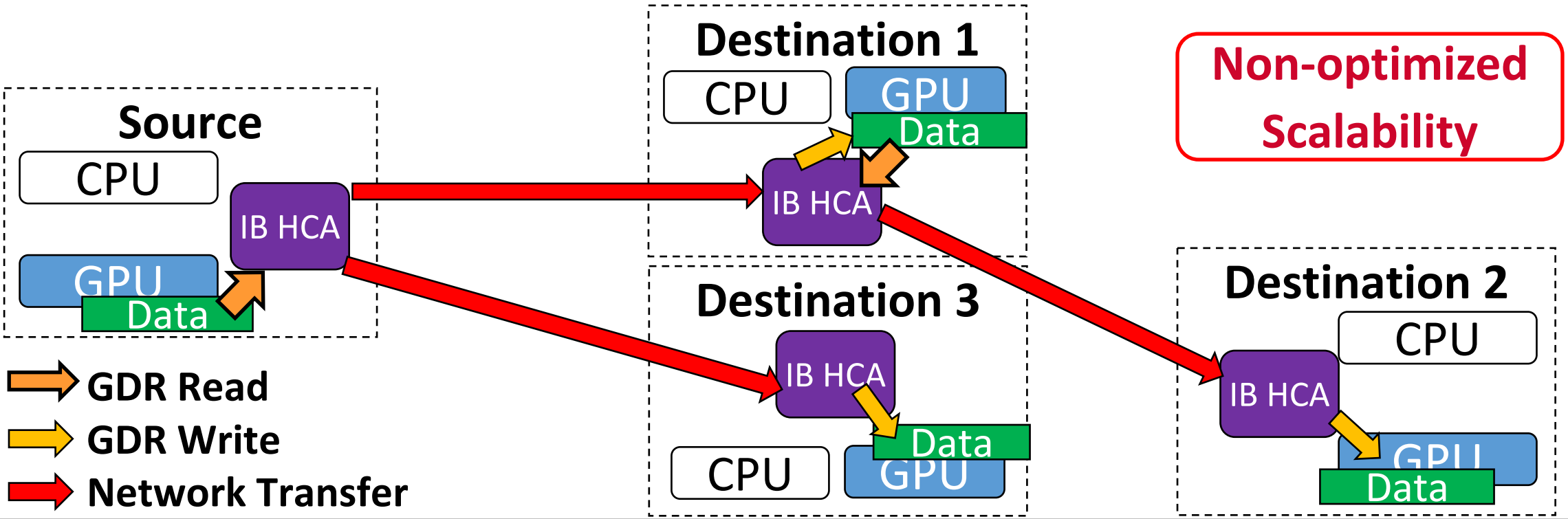
$$\lceil \log_k n \rceil \times \left(t_s + \frac{M}{B_G} \right)$$

- **Pipeline**

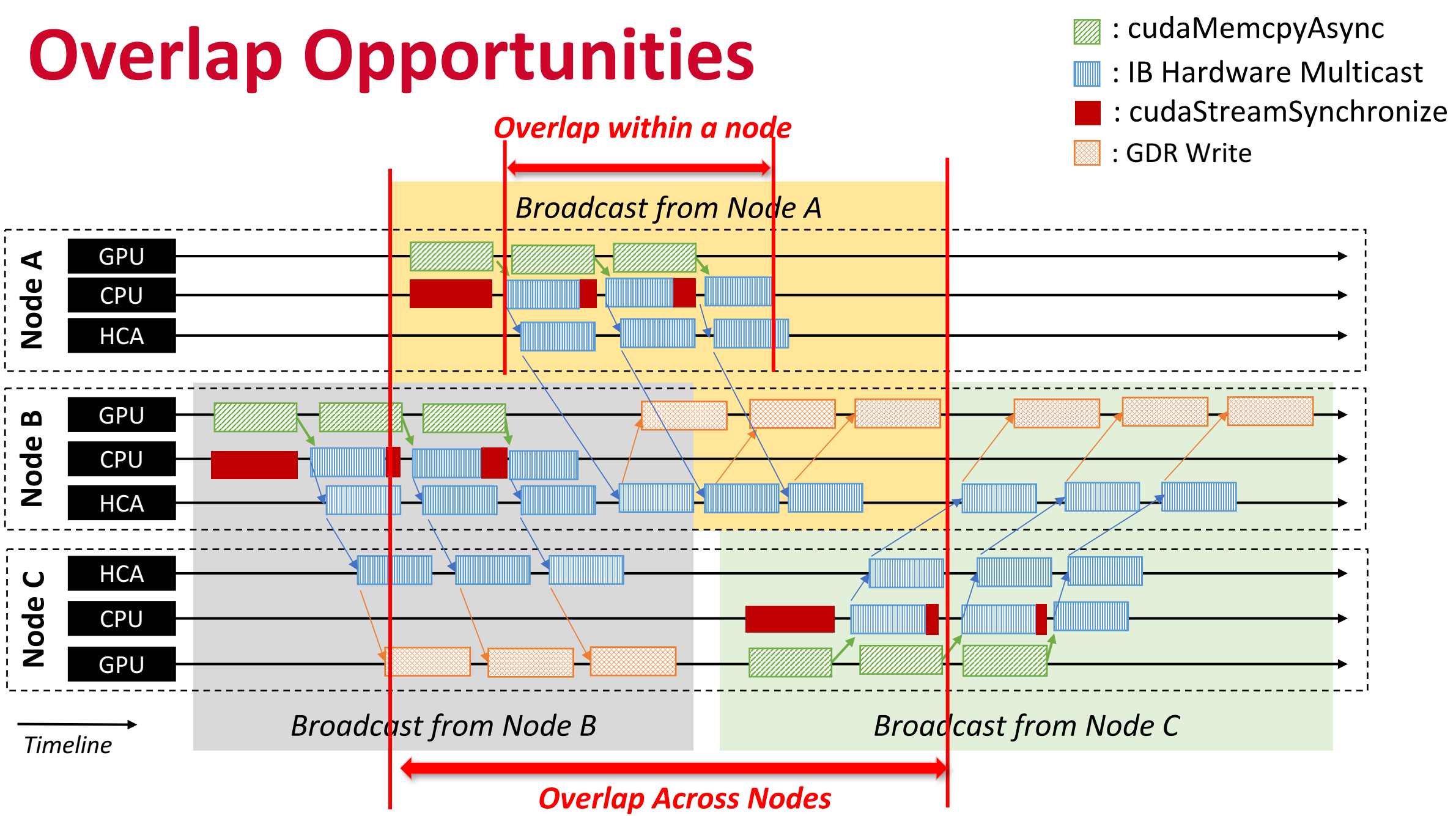
$$\left(\frac{M}{C} \times \lceil \log_k n \rceil \right) \times \left(t_s + \frac{C}{B_G} \right)$$

- **Staging**

$$\frac{M}{B_{PCIe}} + \lceil \log_k n \rceil \times \left(t_s + \frac{M}{B_H} \right)$$



Overlap Opportunities



MCAST-based Broadcast

- **NVIDIA GPUDirect[1]**
  - Remote direct memory access (RDMA) transfers between GPUs and other PCIe devices ⇒ **GDR**
  - and more…
- **InfiniBand (IB) hardware multicast (IB MCAST)[2]**
  - Enables efficient designs of broadcast operations
    - Host-based[3]
    - GPU-based[4]

[1] https://developer.nvidia.com/gpudirect
[2] Pfister GF., "An Introduction to the InfiniBand Architecture. " High Performance Mass Storage and Parallel I/O, Chapter 42, pp 617-632, Jun 2001.
[3] J. Liu, A. R. Mamidala, and D. K. Panda, "Fast and Scalable MPI-level Broadcast using InfiniBand's Hardware Multicast Support," in *IPDPS 2004*, p. 10, April 2004.
[4] 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.

Future Work

- **Extend the design for other broadcast-based collective algorithms as well as non-blocking operations**
  - Allreduce, Allgather, …, and so on