High Performance & Scalable Broadcast Schemes for Deep Learning in GPU Clusters

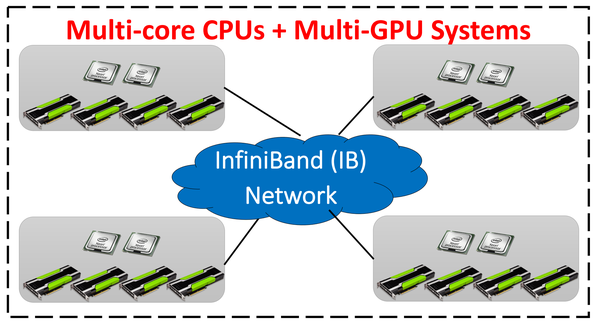
THE OHIO STATE UNIVERSITY

Ching-Hsiang Chu and Dhabaleswar K. Panda (Advisor), Department of Computer Science and Engineering

chu.368@osu.edu, panda@cse.ohio-state.edu

MOTIVATION

- Dense-GPU clusters are increasingly common
  - — e.g., Cray CS-Storm, NVIDIA DGX-1 systems
- Broadcast is heavily used in deep learning apps



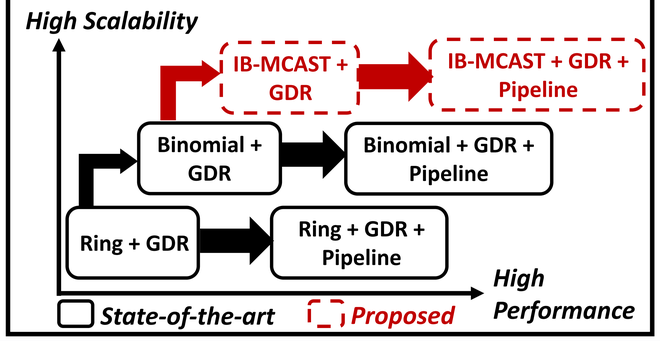
CHALLENGES

- How can we design a scalable heterogeneous broadcast for large-scale GPU clusters? [1,4]
- Can we design an efficient intra-node broadcast for Dense-GPU systems? [1,4]
- Is it possible to address low PCIe bandwidth of NVIDIA GPUDirect RDMA (GDR) read operations for GPU-to-GPU broadcast? [2,4]
- How to provide efficient reliability support for unreliable IB hardware multicast (IB-MCAST)? [3]

CONTRIBUTIONS

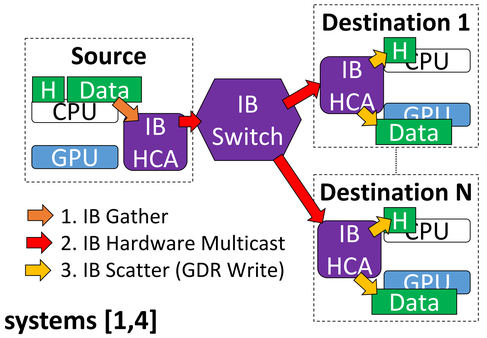
- Proposes analytical models to capture and predict performance behavior of alternative broadcast schemes on GPU clusters
- Designs scalable and reliable zero-copy homogeneous and heterogeneous broadcast schemes
  - — Leverages IB-MCAST and NVIDIA GDR features

➠ **No application code changes**

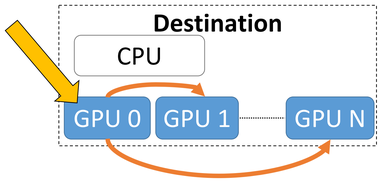


PROPOSED BROADCAST DESIGNS

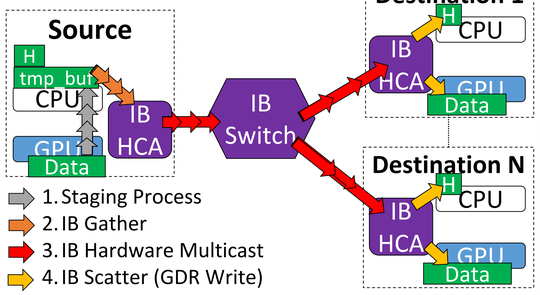
- **IB-MCAST & NVIDIA GDR for heterogeneous broadcast [1,4]**
  - — Multicast two separate addresses (control header + data)—**in one IB message**
  - — **Benefits**
    - ★ Directly IB write to GPU buffers using GDR feature
    - ★ Frees up PCIe bandwidth resource for application needs



- **Intra-node broadcast for Dense-GPU systems [1,4]**
  - — Leverages CUDA Inter-Process Communication (IPC)
  - — **Benefits**
    - ★ Direct read/write GPU buffers
    - ★ Bypasses CPU
    - ★ Frees up PCIe bandwidth resource between CPU and GPU



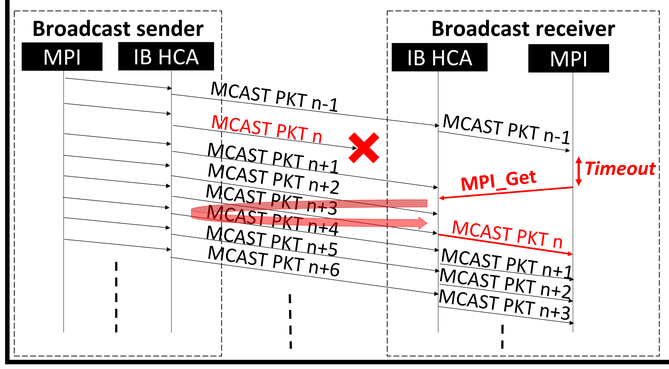
- **Streaming-based Optimization for GPU-to-GPU broadcast [2,4]**
  - — Streaming GPU-resident data through host memory
    - ★ Three-stage pipeline
  - — **Benefits**
    - ★ Avoids low-bandwidth GDR read operations
    - ★ Overlapping data transfers within and across nodes



PROPOSED RELIABILITY SUPPORT

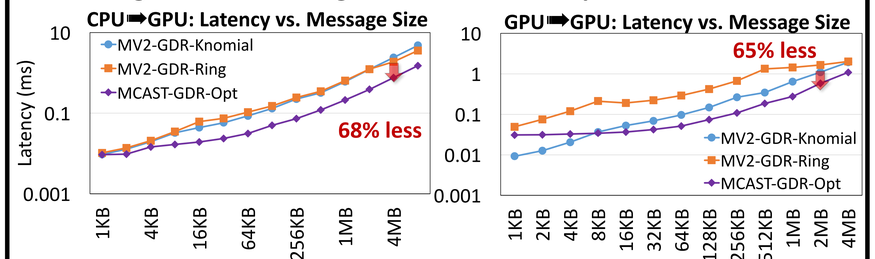
- Allows receivers to retrieve lost IB-MCAST packets through RMA operations without interrupting sender [3]

➠ **Benefit:** Maintains pipelining of broadcast operations

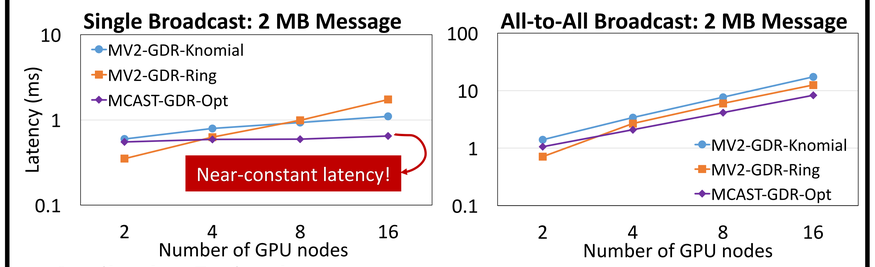


PERFORMANCE EVALUATION

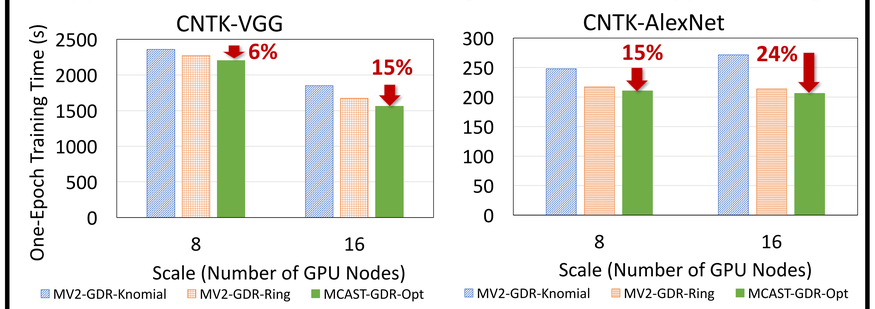
- **Experimental environment:**
  - — Two 14-core Intel (Broadwell) Xeon E5-2680 V4 processors
  - — 1 NVIDIA K80 GPU (i.e., 2 GPU chips) per node—used up to 16 GPU nodes
  - — InfiniBand EDR HCA, and Mellanox SB7790 and SB7800 switches
- **Heterogeneous & Homogeneous Broadcast Operations**



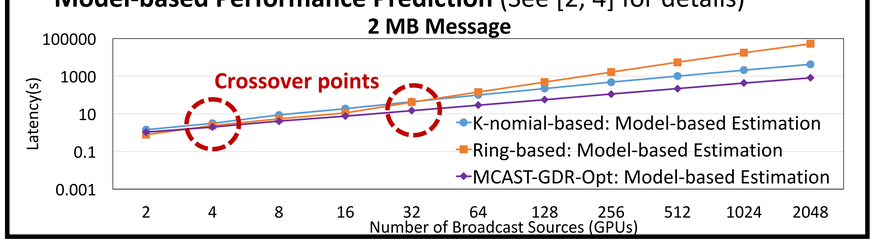
- **Scalability Evaluation:** **OSU Micro-benchmark**



- **Application Evaluation:** **Microsoft Cognitive Toolkit—No App Code Changes**



- **Model-based Performance Prediction** (See [2, 4] for details)



PUBLICATIONS

[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)*

This material is based upon work supported by, or in part by, the Department of Defense (DOD) High Performance Computing Modernization Program (HPCMP) under User Productivity, Technology Transfer and Training (PETTT) contract number #GS04T09DBC0017 through Engility Corporation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DOD, the DOD HPCMP, or the employer of the author.