THE OHIO STATE UNIVERSITY
COLLEGE OF ENGINEERING

Efficient and Scalable Communication Middleware for Emerging Dense-GPU Clusters

Ching-Hsiang Chu

Advisor: Dhabaleswar K. Panda

Network Based Computing Lab
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH

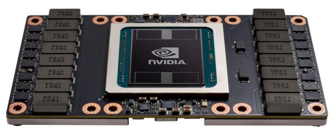
Outline

- **Introduction**
- Problem Statement
- Detailed Description and Results
- Broader Impact on the HPC Community
- Expected Contributions

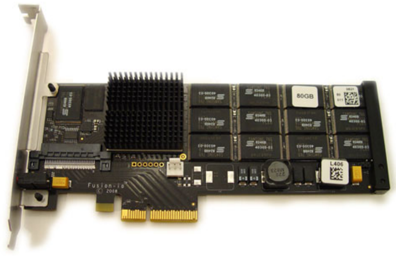
Trends in Modern HPC Architecture: Heterogeneous

Multi/ Many-core
Processors

High Performance Interconnects
InfiniBand, Omni-Path, EFA
<1usec latency, 100Gbps+ 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)

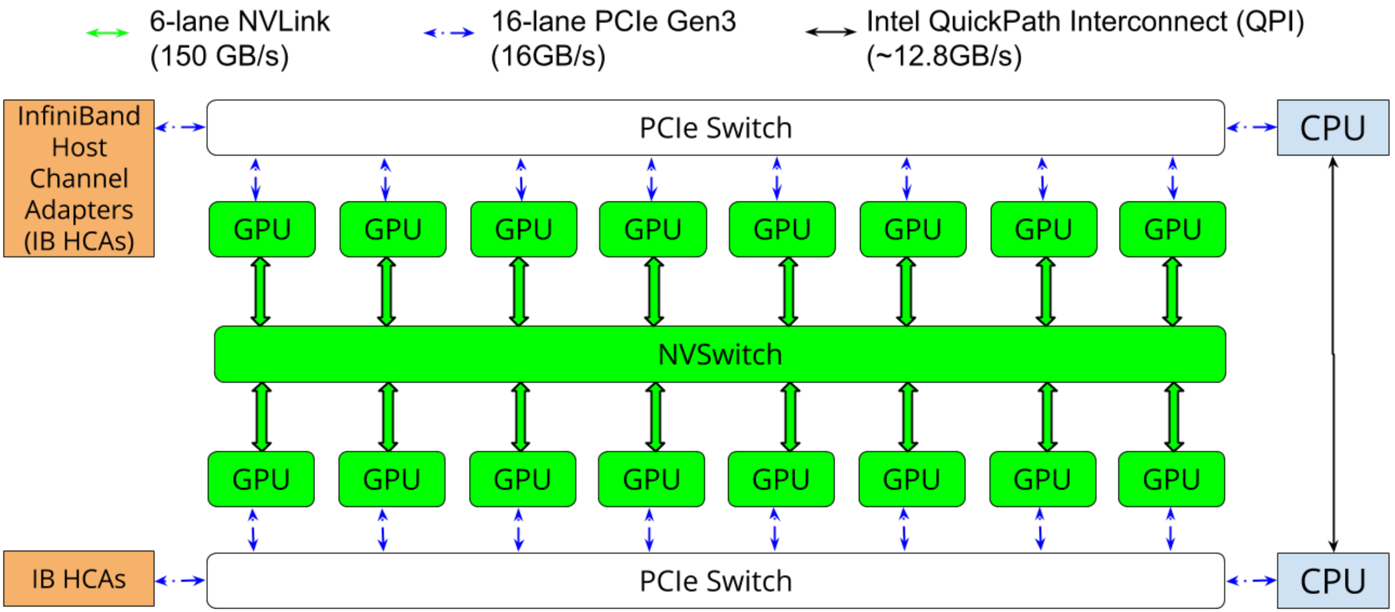
#2 Sierra (17,280 GPUs)
#10 Lassen (2,664 GPUs)

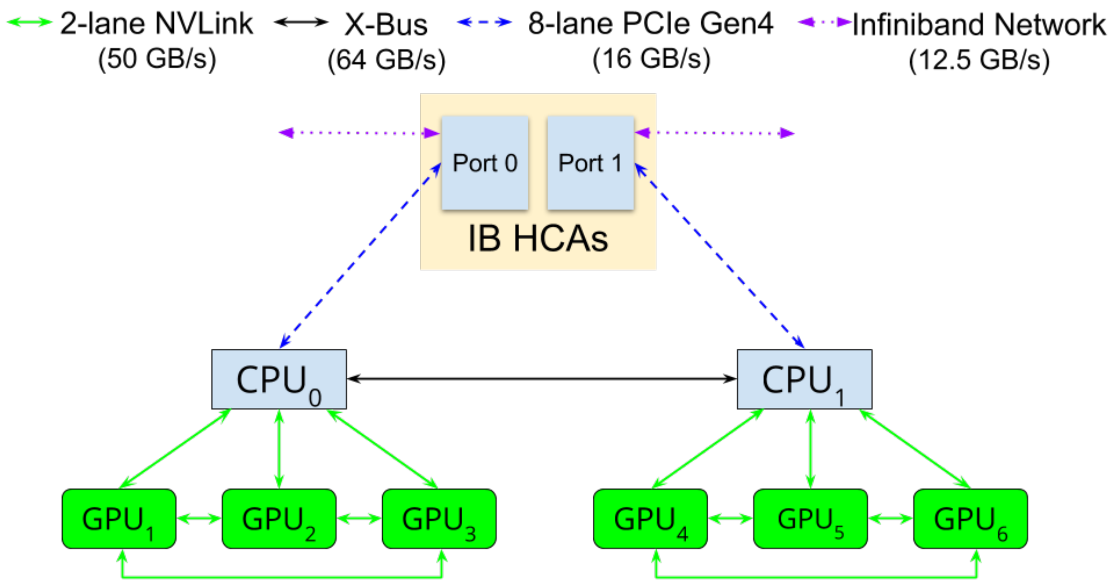
#8 ABCI
(4,352 GPUs)

#22 DGX SuperPOD
(1,536 GPUs)

Trends in Modern Large-scale Dense-GPU Systems

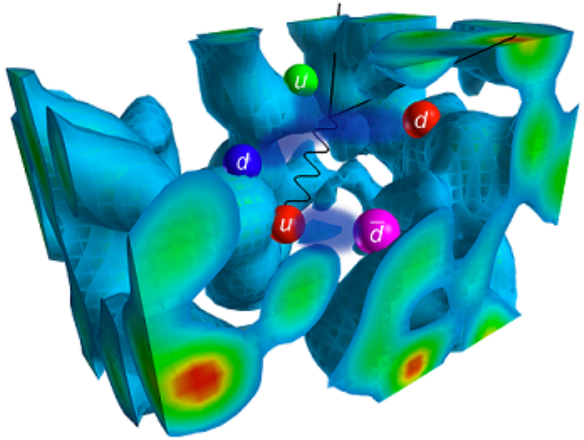
- **Scale-up** (up to 150 GB/s)
  - PCIe, NVLink/NVSwitch
  - Infinity Fabric, Gen-Z, CXL
- **Scale-out** (up to 25 GB/s)
  - InfiniBand, Omni-path, Ethernet
  - Cray Slingshot





GPU-enabled HPC Applications

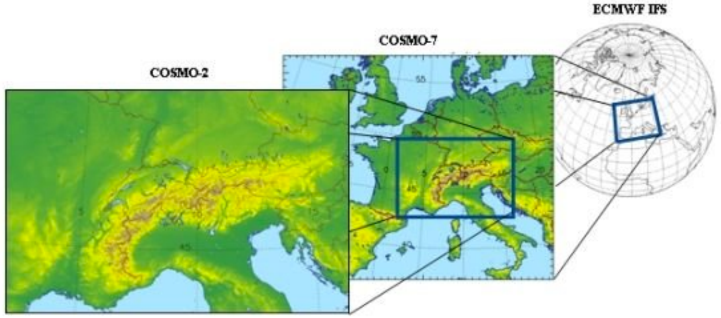
Lattice Quantum Chromodynamics



Derek Leinweber, CSSM, University of Adelaide

23x faster than CPU

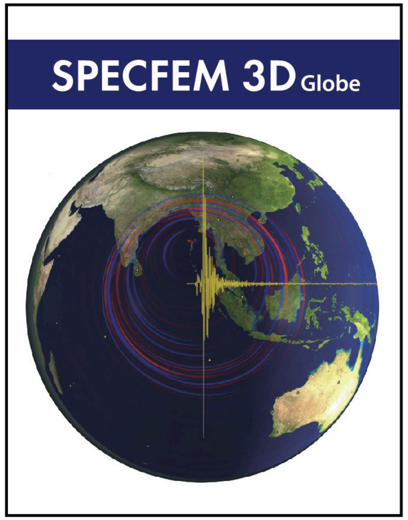
Weather Simulation



Fuhrer O, Osuna C, Lapillonne X, Gysi T, Bianco M, Schulthess T. Towards GPU-accelerated operational weather forecasting. GTC 2013.

2.8x faster than CPU

Wave propagation simulation



https://geodynamics.org/cig/software/specfem3d_globe/

25x faster than CPU

- Various scientific applications are ported to GPU
  - Reportedly significant speedup compared to CPU version
  - High-resolution/precision results

GPU-enabled Emerging Deep Learning Applications

- **Easy-to-use and high-performance frameworks**

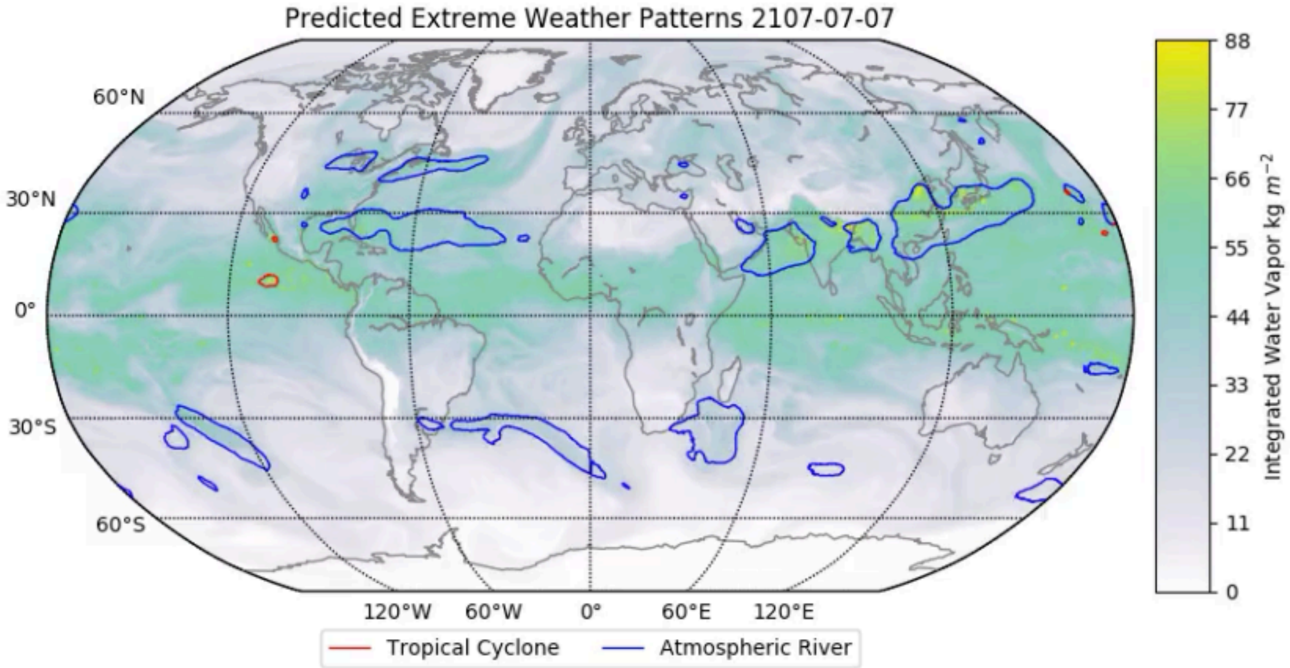
PYTORCH

TensorFlow

Microsoft Cognitive Toolkit
aka.ms/CognitiveToolkit
Microsoft

- **Wide range of applications**
  - Image Classification
  - Speech Recognition
  - Self-driving car
  - Healthcare
  - Climate Analytic

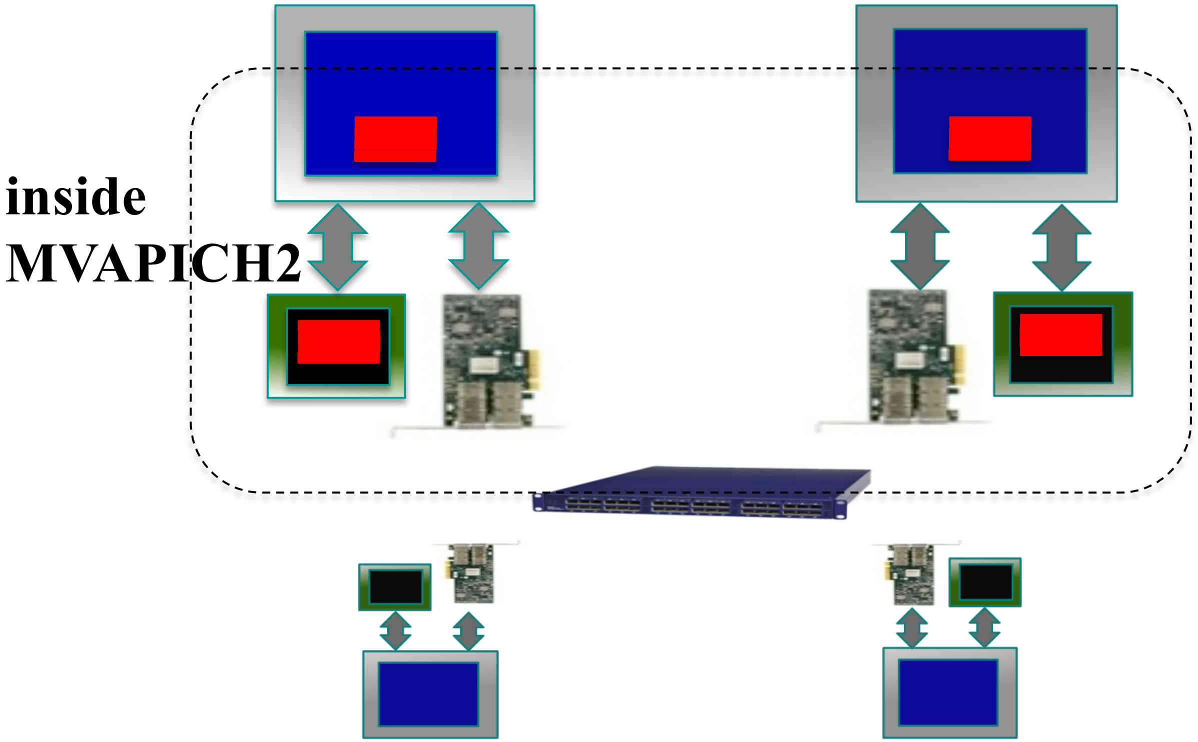
999 PetaFlop/s sustained, and 1.13 ExaFlop/s peak
FP 16 performance over 4560 nodes (27,360 GPU)



Kurth T, Treichler S, Romero J, Mudigonda M, Luehr N, Phillips E, Mahesh A, Matheson M, Deslippe J, Fatica M, Houston M. Exascale deep learning for climate analytics. SC 2018 Nov 11 (p. 51). (Golden Bell Prize)

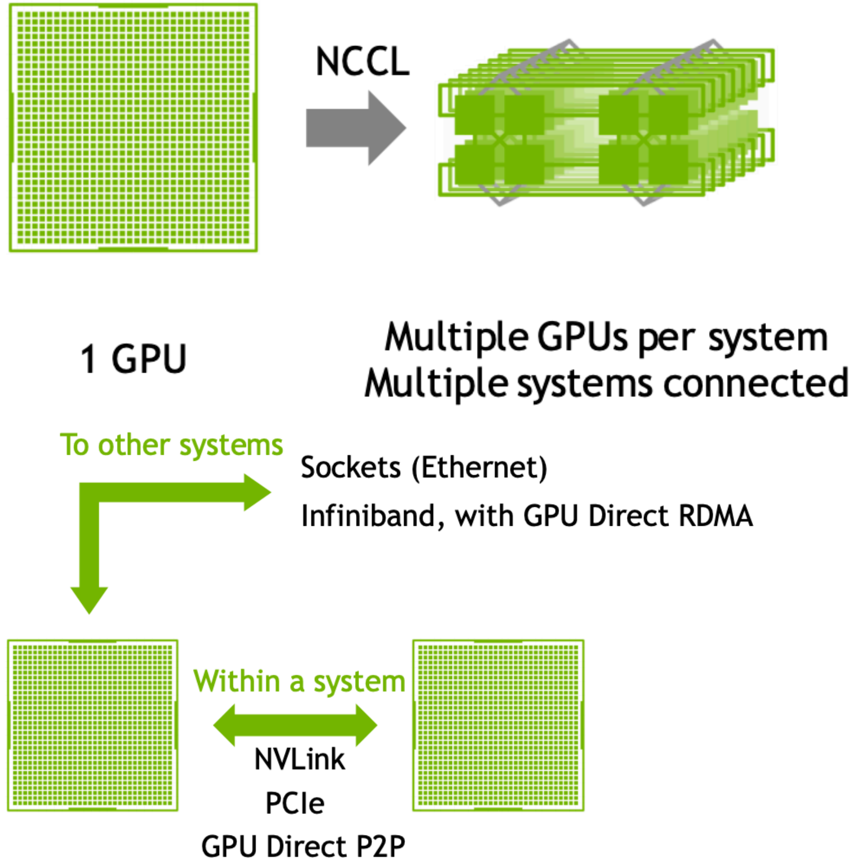
GPU-Aware (CUDA-Aware) Communication Middleware

MPI-based Generic Communication Middleware



- Supports and optimizes various communication patterns
- Overlaps data movement from GPU with RDMA transfers

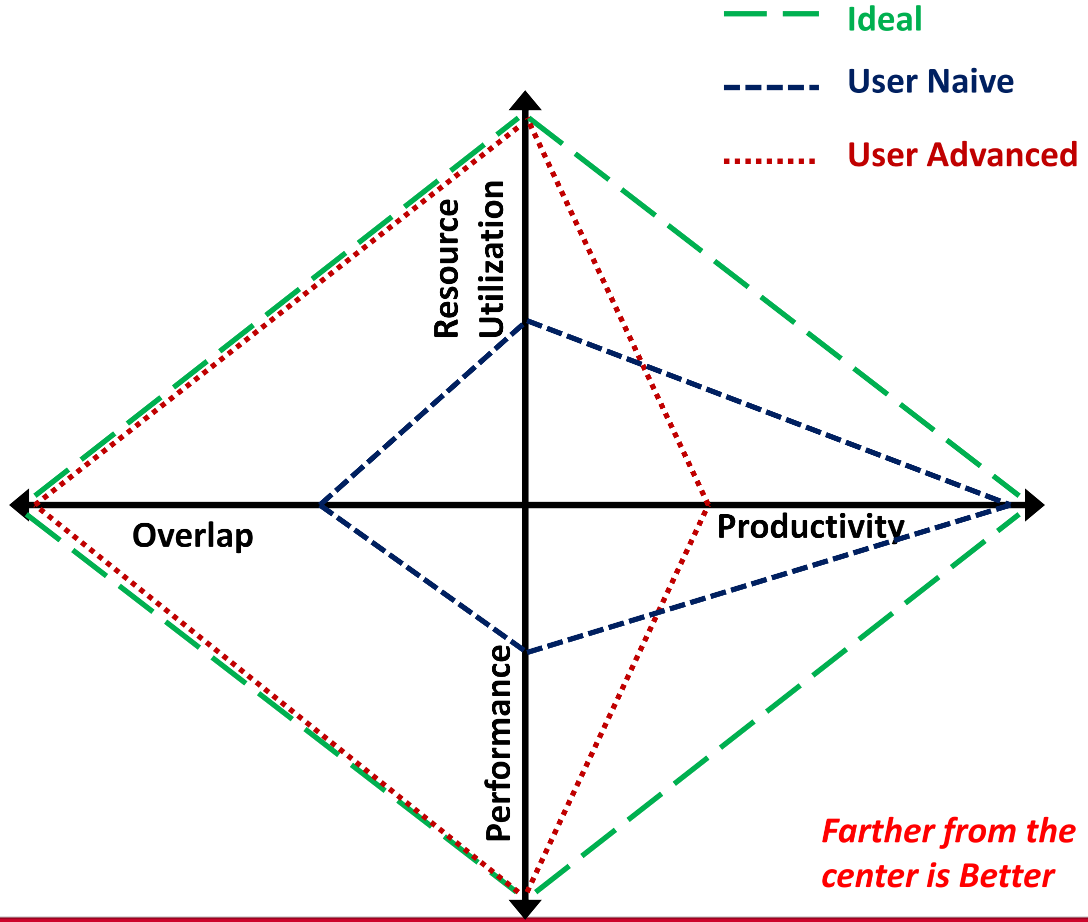
DL-Specific Communication Middleware



- Ring-based collective operations
- Optimized for DL workloads on GPU systems

Broad Challenge

Can we design a generic GPU-enabled communication middleware to fully exploit GPU resources and interconnects for traditional HPC and emerging ML/DL applications?



Farther from the center is Better

Outline

- Introduction
- **Problem Statement**
- Detailed Description and Results
- Broader Impact on the HPC Community
- Expected Contributions

Problem Statements

- What kind of hardware capabilities can be leveraged to fully exploit the modern interconnects deployed in GPU clusters?
- What is the potential scope of communication patterns can be benefited from the proposed GPU-enabled communication middleware?
- How to leverage GPU resources such as high-bandwidth memory (HBM) and massive streaming multiprocessors (SM) to accelerate communication?
- What are design considerations for a GPU-enabled communication middleware to efficiently utilize the hardware features?
- What kind of performance benefits can be expected with the proposed GPU-enabled communication middleware?
- How can the traditional HPC and DL applications can take advantage of the proposed communication middleware without application-level modifications?

Research Framework

Applications
- Benchmarks: OMB, Streaming
- Traditional HPC Applications: MILC, COSMO, AWP-ODC, HOOMD-Blue
- Machine/Deep Learning: TensorFlow, CNTK

Programing Models
- Message Passing Interface (MPI)

GPU-Enabled Communication Middleware
- Efficient Non-contiguous Data Process: Scheduling and Transfer
- Scalable Streaming Broadcast
- GPU-enabled Cooperative and Link-efficient reduction operations

Compute Models
- CUDA
- OpenACC

Hardware Capability
- GPUDirect RDMA
- GDRCOPY
- Hardware Multicast
- Inter-GPU Direct Load/Store

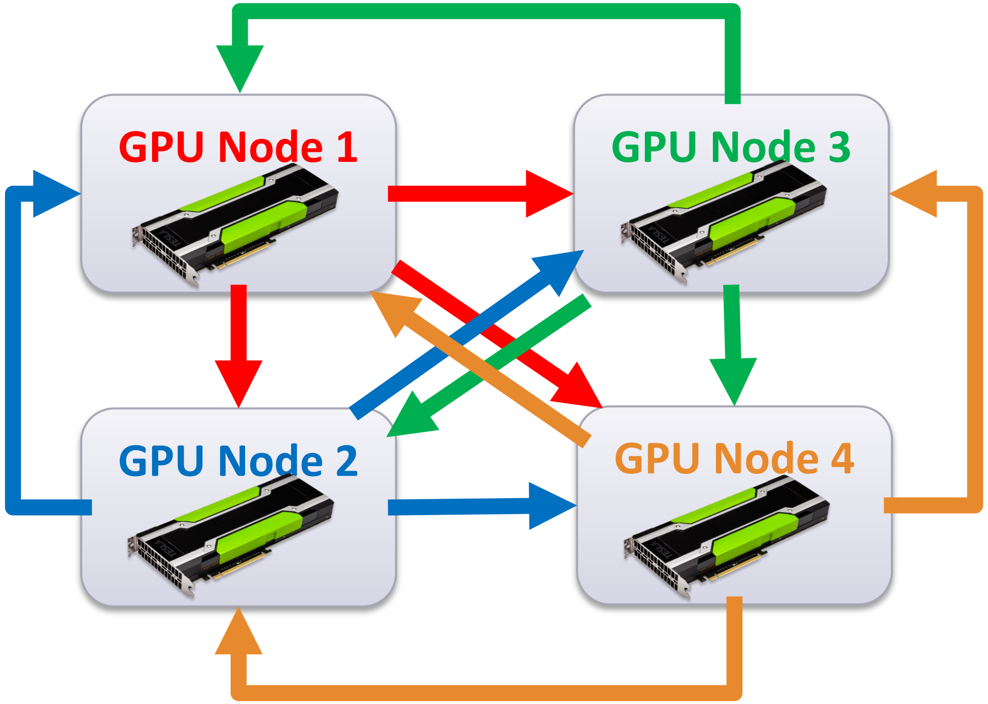
Modern HPC Hardware
- Multi-Core Processors: Xeon, OpenPOWER
- Accelerator: GPU
- Interconnects: InfiniBand, PCIe, NVLink/NVSwitch

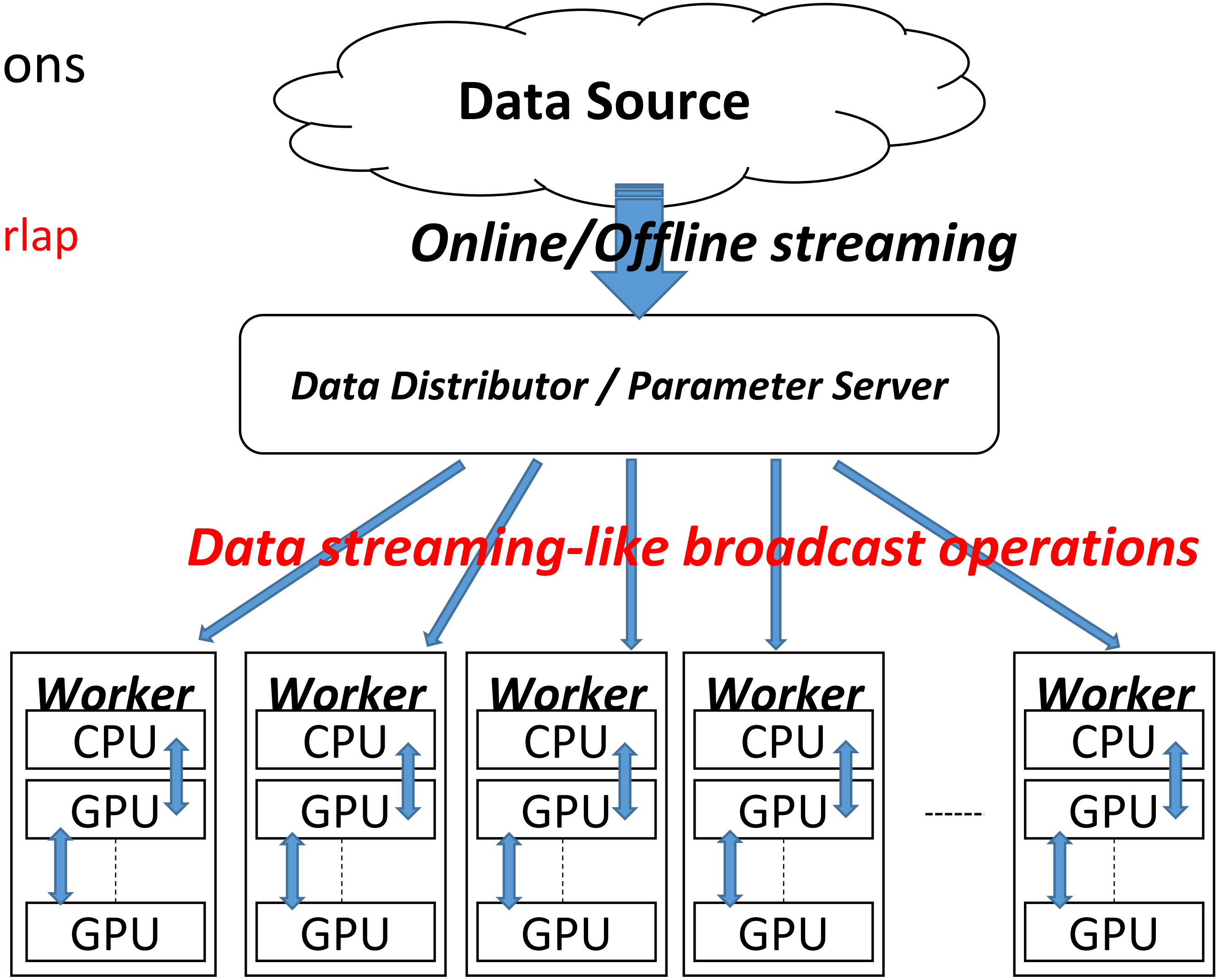
Outline

- Introduction
- Problem Statement
- **Detailed Description and Results**
  - Scalable Streaming Broadcast for InfiniBand Networks
  - Efficient Scheduling of Non-contiguous Data Transfer
  - GPU-enabled Zero-copy Transfer for Non-contiguous Data
  - GPU-enabled Reduction operations
- Broader Impact on the HPC Community
- Expected Contributions

Motivated Example #1 – Need of scalable broadcast

- Streaming & Deep Learning applications
  - Large-scale broadcast operations
  - High computation-communication overlap
  - No application-level modification

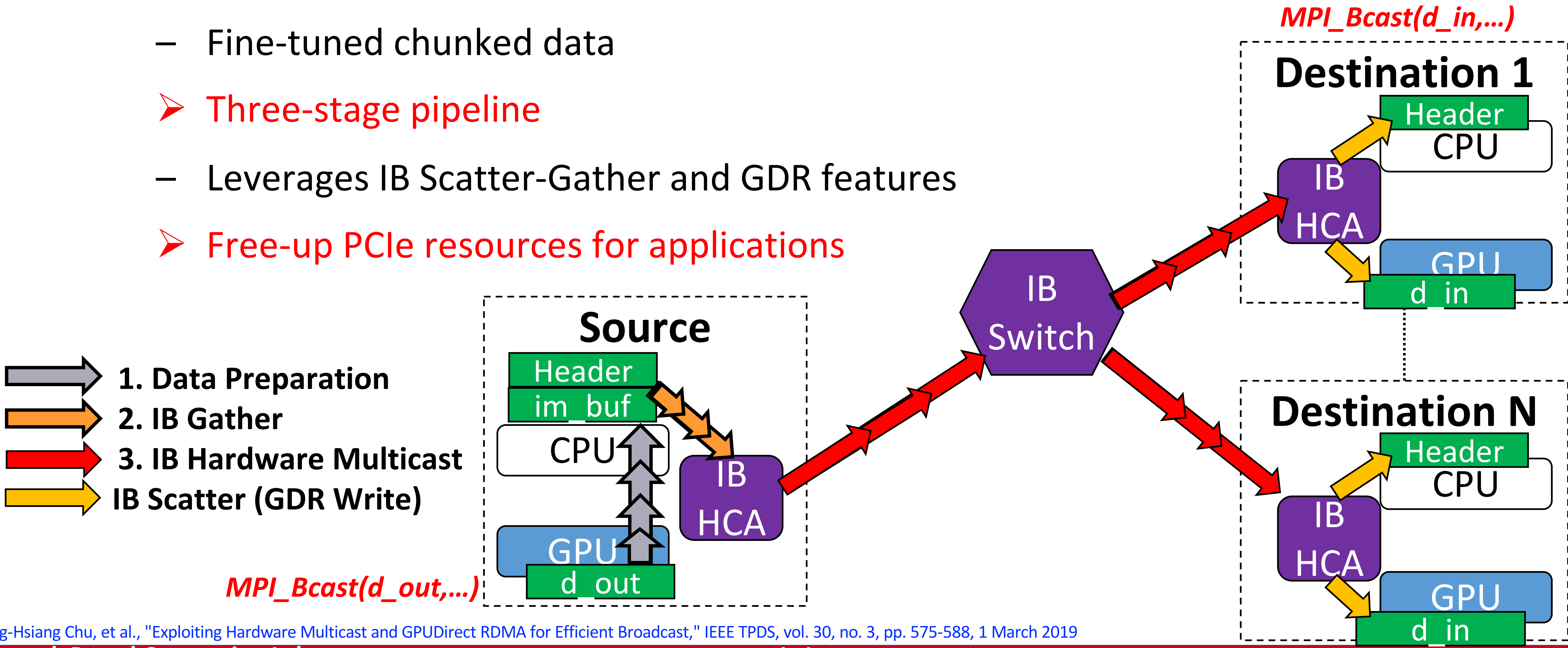




Ching-Hsiang Chu, et al., "Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast," IEEE TPDS, vol. 30, no. 3, pp. 575-588, 1 March 2019

Proposed Efficient and Scalable Solution

- Streaming data through host
  - Fine-tuned chunked data
  - ➢ Three-stage pipeline
  - Leverages IB Scatter-Gather and GDR features
  - ➢ Free-up PCIe resources for applications

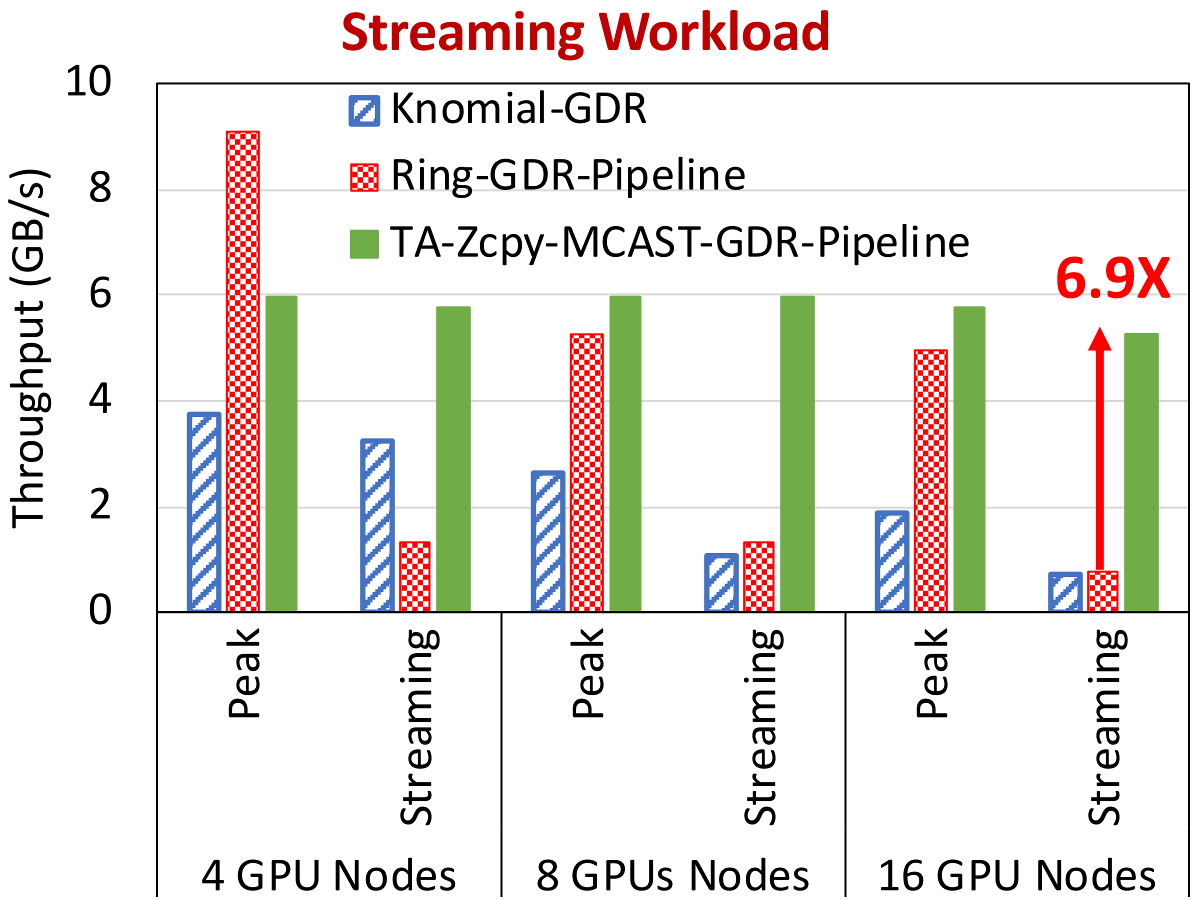


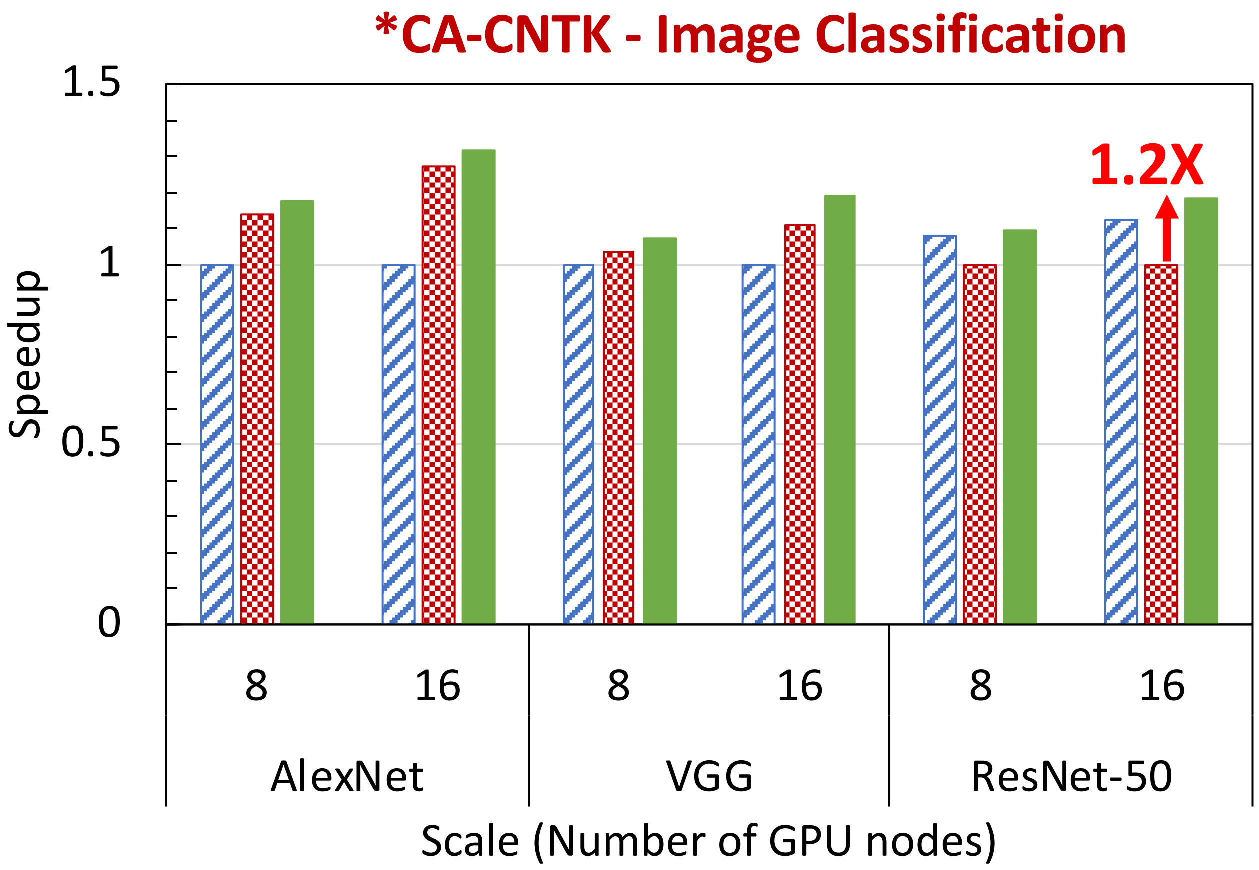
Ching-Hsiang Chu, et al., "Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast," IEEE TPDS, vol. 30, no. 3, pp. 575-588, 1 March 2019

Performance Evaluation

- Evaluated @ RI2 GPU nodes

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



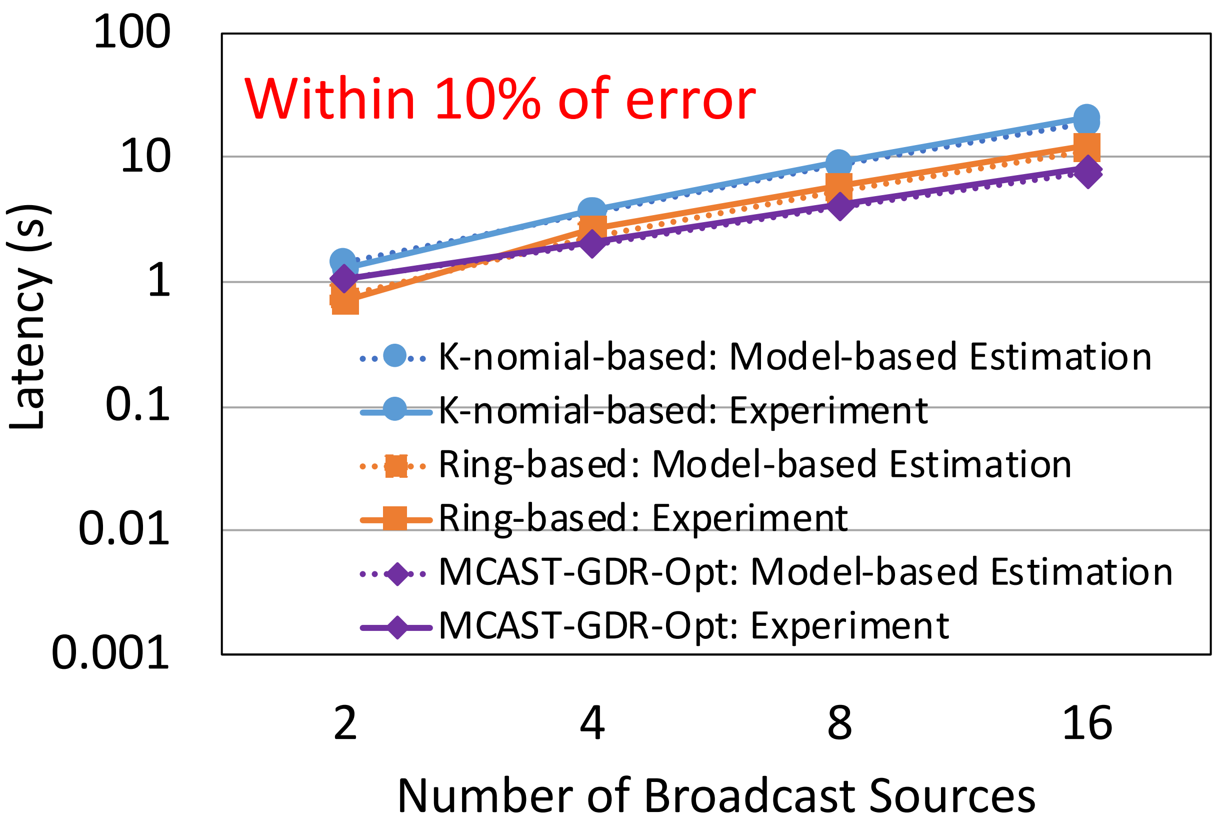


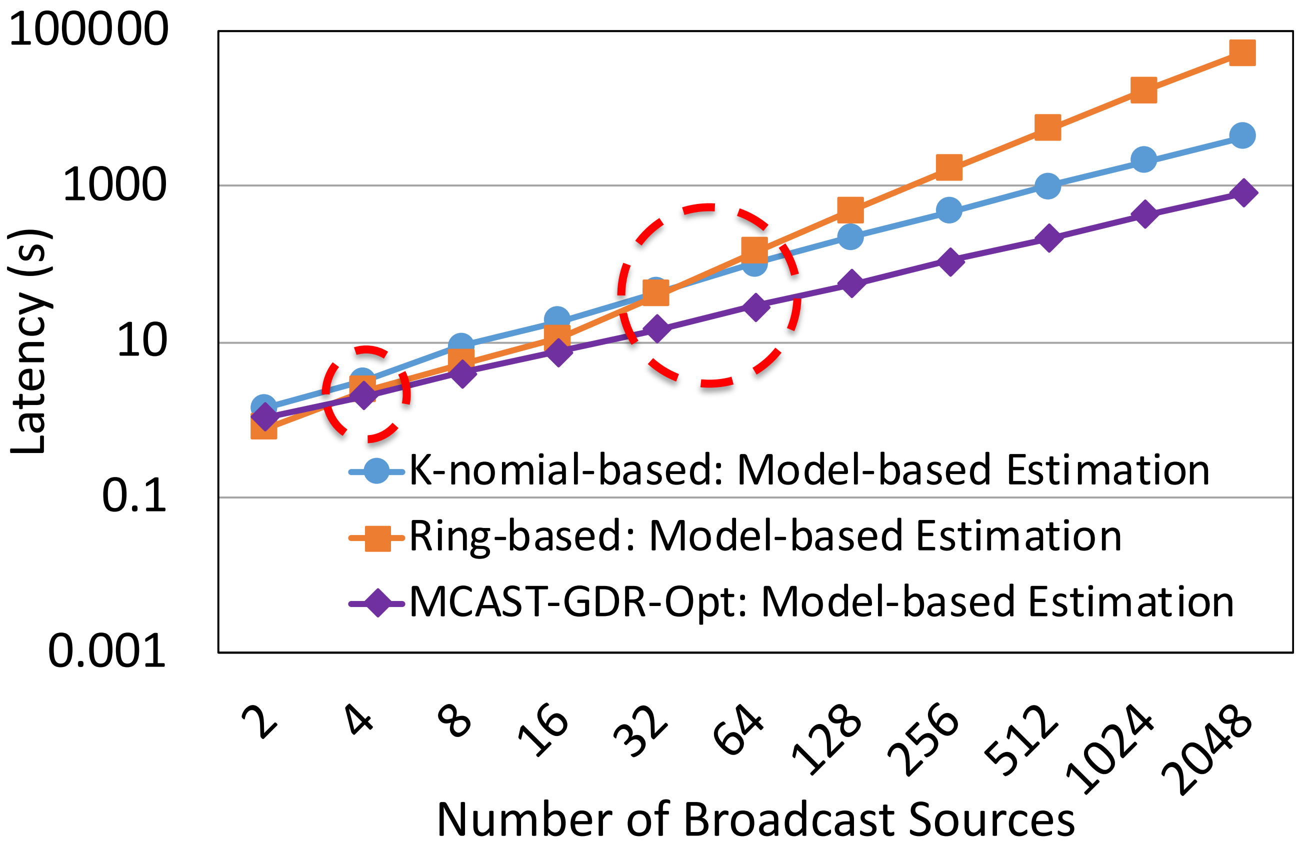
Ching-Hsiang Chu, et al., "Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast," IEEE TPDS, vol. 30, no. 3, pp. 575-588, 1 March 2019

Performance Model Validation and Prediction

- Based on the architecture on RI2 cluster

$$M = 2MB;\ C = 512\ KB;\ U = 4\ KB;\ B_H \approx 100\ Gbps;\ B_{PCIe} = 8\ Gbps;\ t_o(n) \approx \frac{1}{\alpha} \times \ln(n), 15 \le \alpha \le 20$$





Ching-Hsiang Chu, et al., "Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast," IEEE TPDS, vol. 30, no. 3, pp. 575-588, 1 March 2019

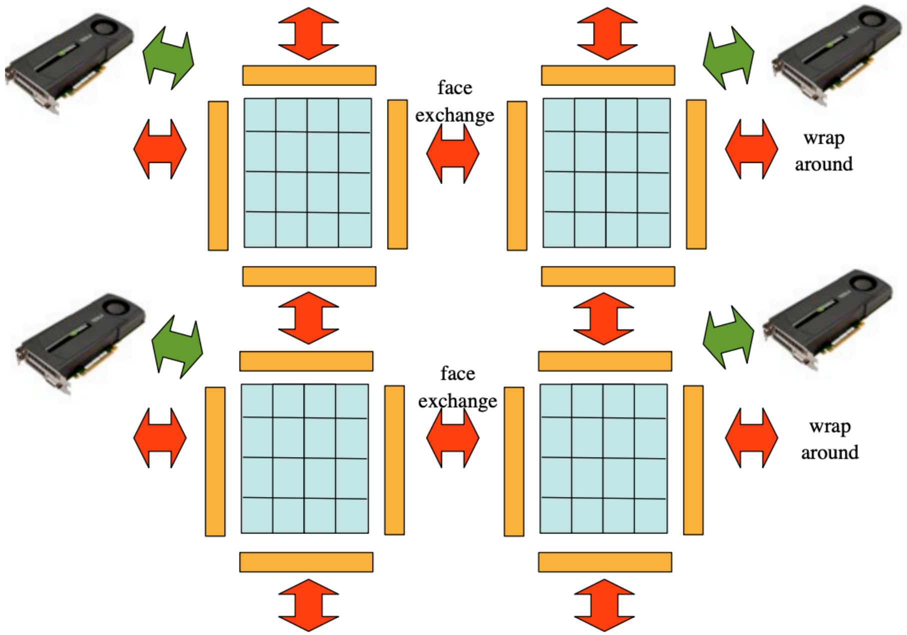
Outline

- Introduction
- Problem Statement
- **Detailed Description and Results**
  - Scalable Streaming Broadcast for InfiniBand Networks
  - **Efficient Scheduling of Non-contiguous Data Transfer**
  - GPU-enabled Zero-copy Transfer for Non-contiguous Data
  - GPU-enabled Reduction operations
- Ongoing Work and Future Research Directions
- Broader Impact on the HPC Community
- Expected Contributions

Motivated Example #2 – Non-contiguous Data Transfer

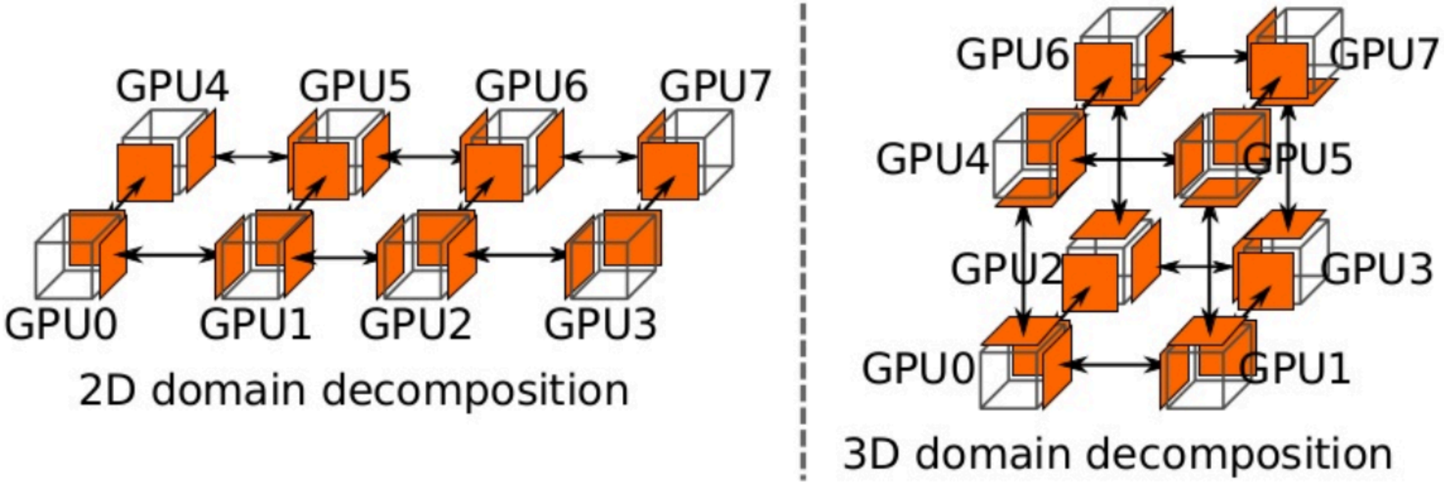
- Wide usages of MPI derived datatype for Non-contiguous Data Transfer
  - Requires Low-latency and high overlap processing

Quantum Chromodynamics: MILC with QUDA



Mike Clark. "GPU Computing with QUDA, "Developer Technology Group, https://www.olcf.ornl.gov/wp-content/uploads/2013/02/Clark_M_LQCD.pdf

Weather Simulation: COSMO model



M. Martinasso, G. Kwasniewski, S. R. Alam, Thomas C. Schulthess, and T. Hoefler. "A PCIe congestion-aware performance model for densely populated accelerator servers. " SC 2016

Existing GPU-enabled MPI Datatype Processing

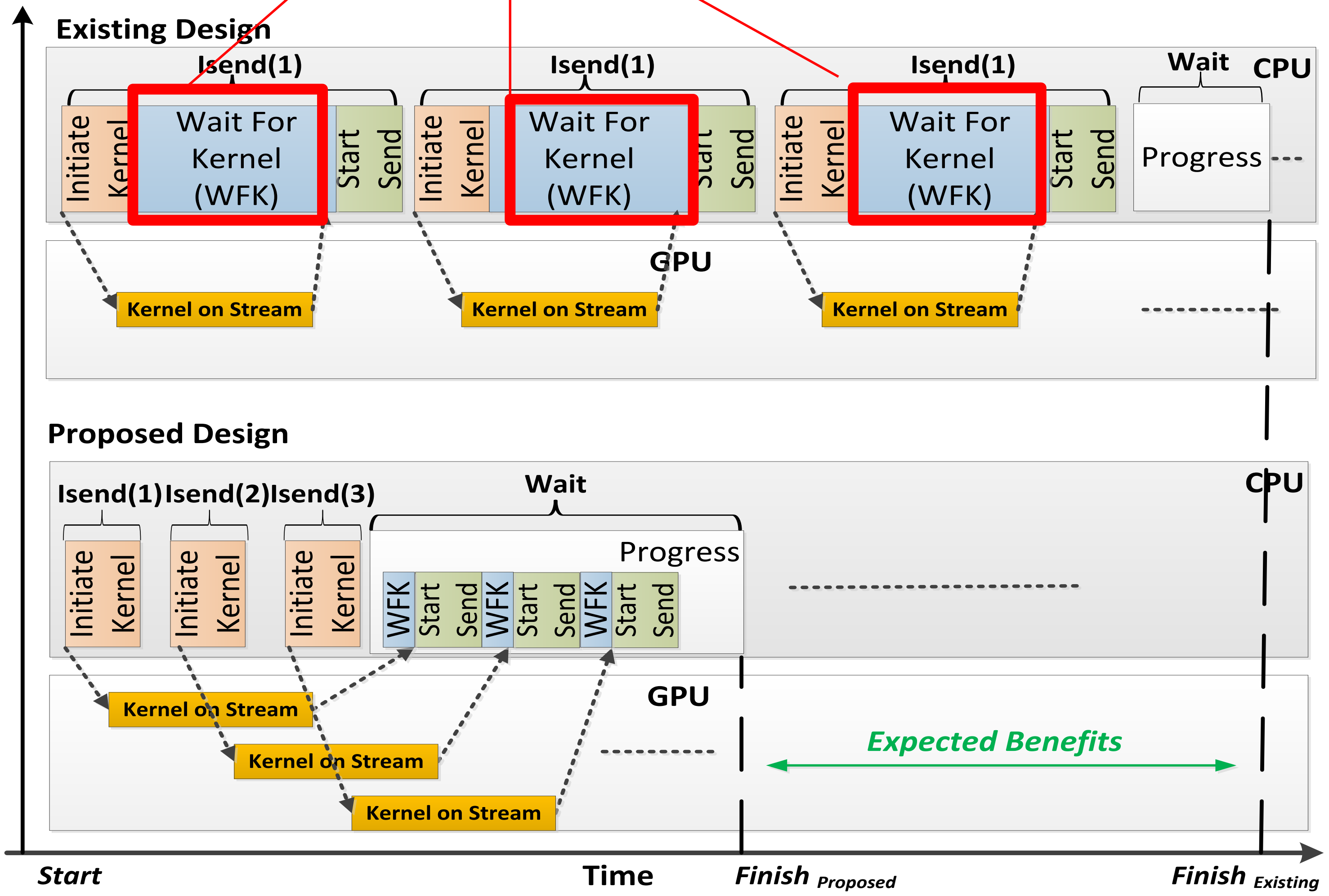
Common Scenario

MPI_Isend (A,.. Datatype,…)
MPI_Isend (B,.. Datatype,…)
MPI_Isend (C,.. Datatype,…)
MPI_Isend (D,.. Datatype,…)
…

MPI_Waitall (…);

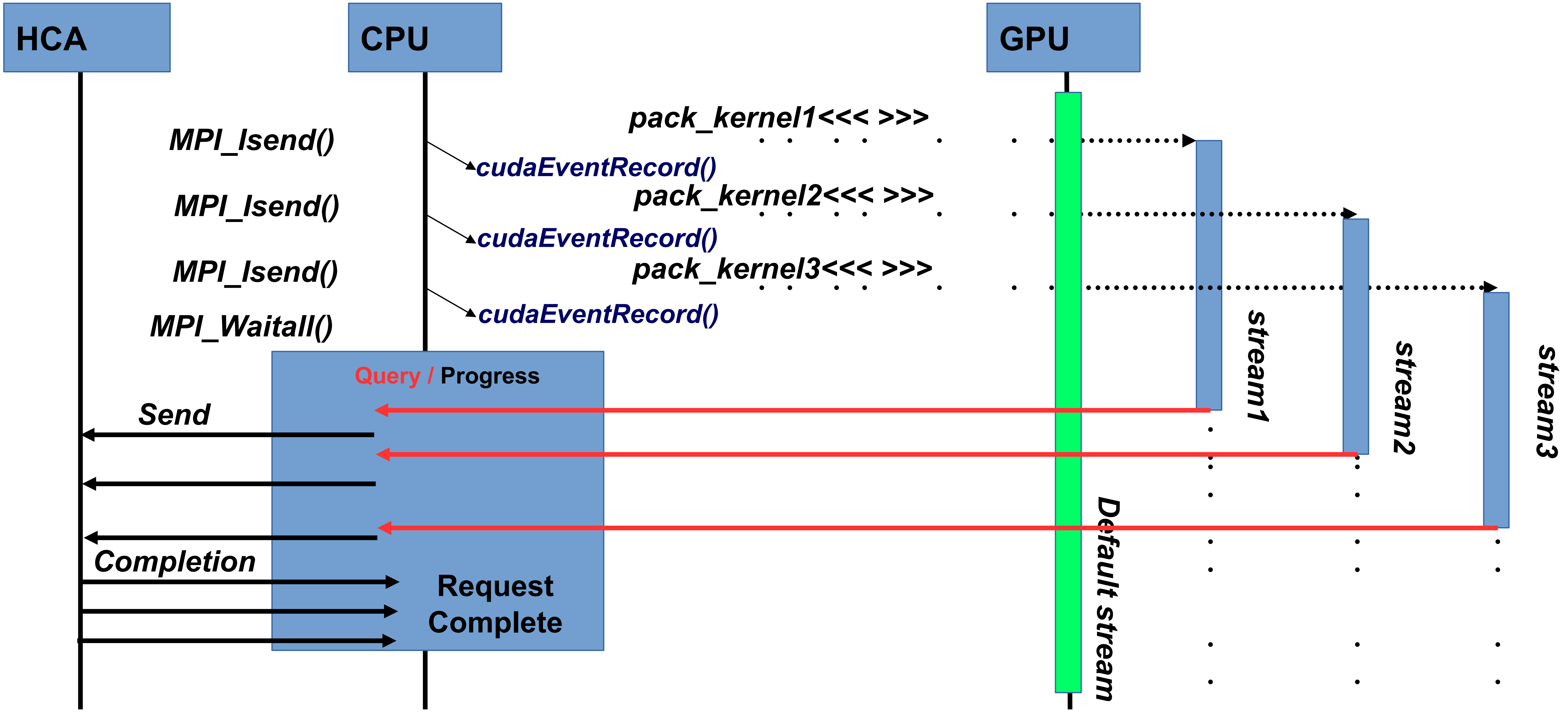
*A, B…contain non-contiguous MPI Datatype

Waste of computing resources on CPU and GPU

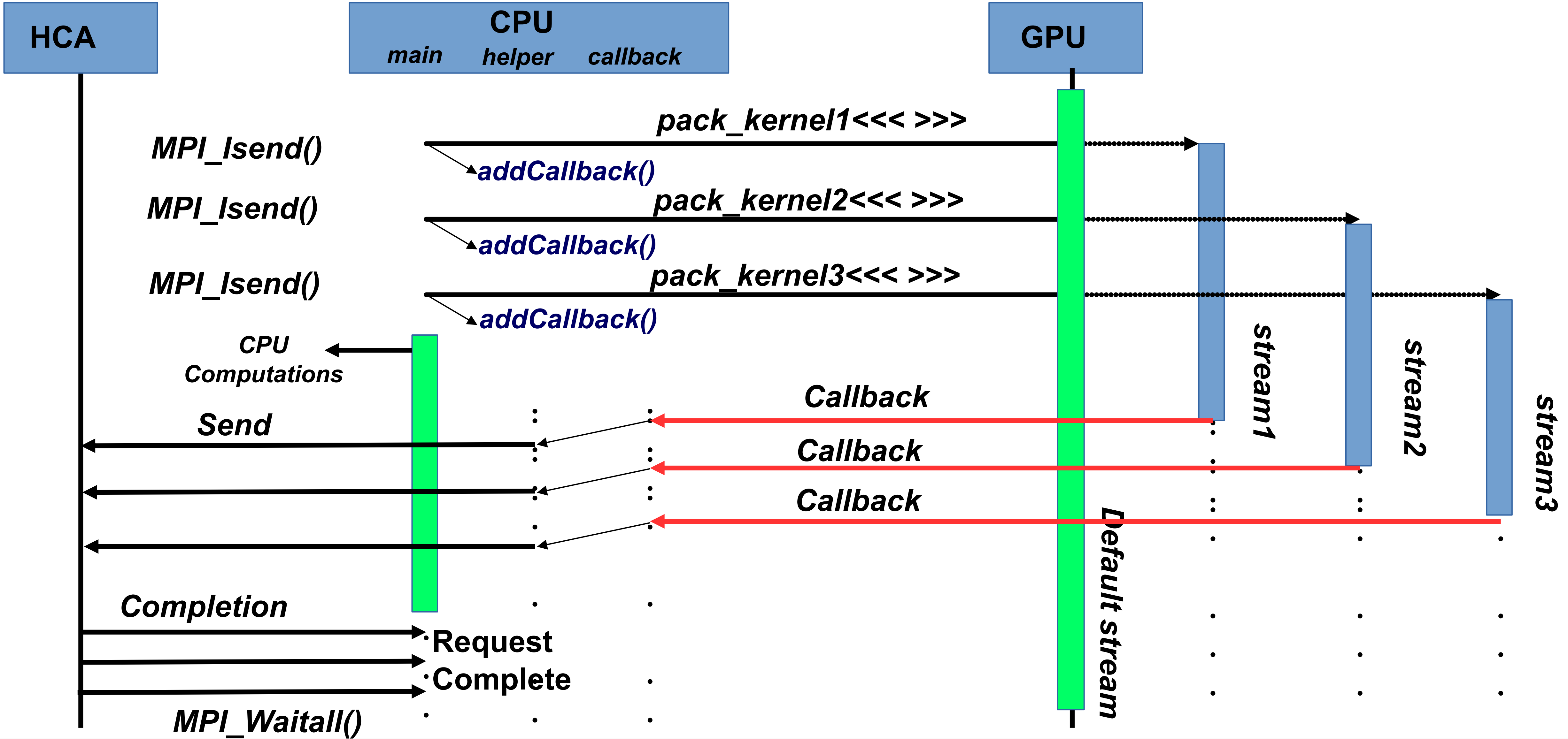


Ching-Hsiang Chu et al., "Exploiting Maximal Overlap for Non-Contiguous Data Movement Processing on Modern GPU-enabled Systems, " IEEE IPDPS 2016.

Proposed Event-based Design – Low Latency

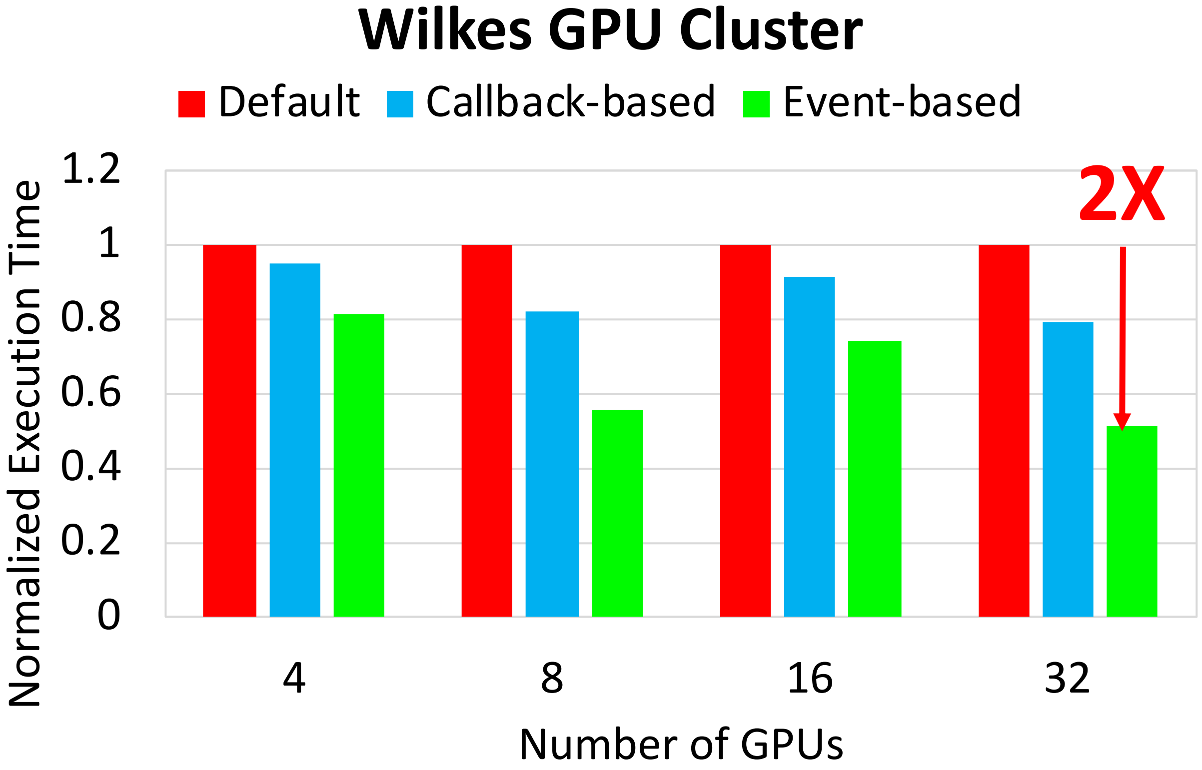


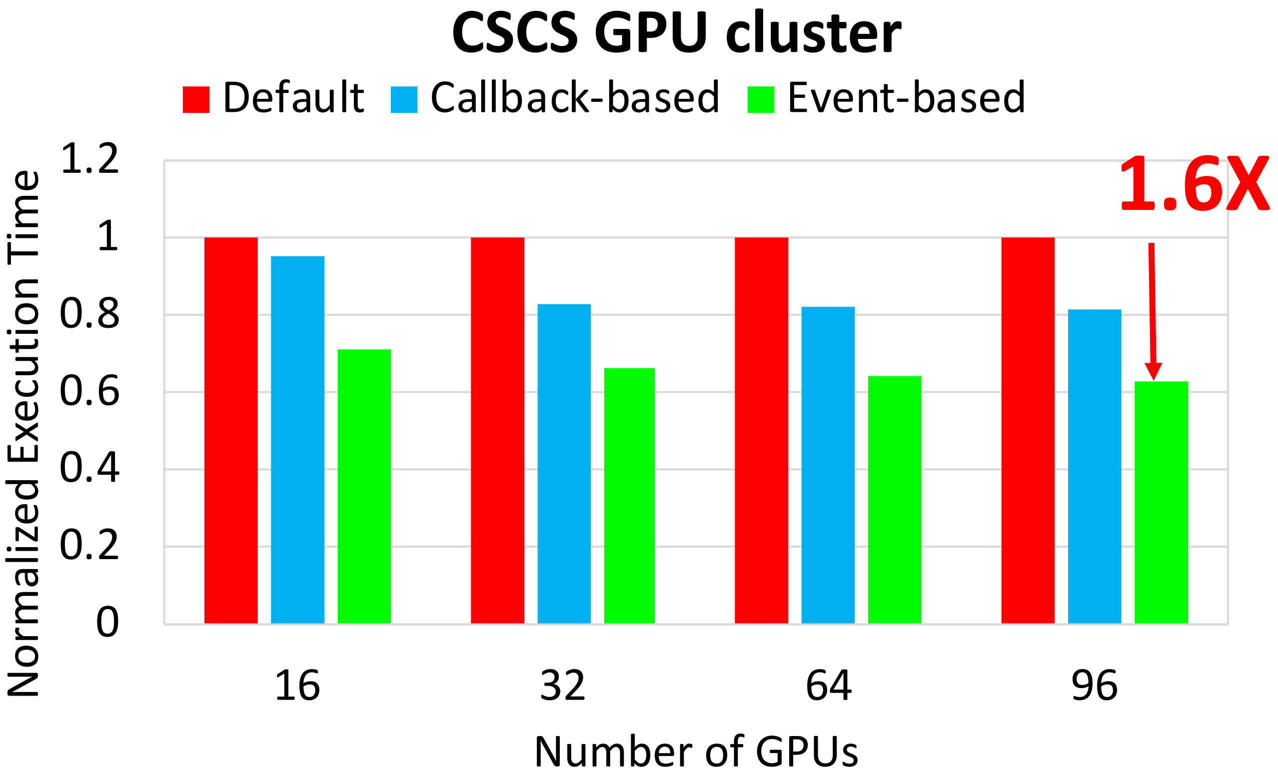
Proposed Callback-based Design – High Overlap



Application-level (COSMO HaloExchange) Evaluation

```
MPI_Isend(Buf1, ...,request1);
MPI_Isend(Buf2, ...,request2);
MPI_Wait (request1, status1);
MPI_Wait (request2, status2);
```





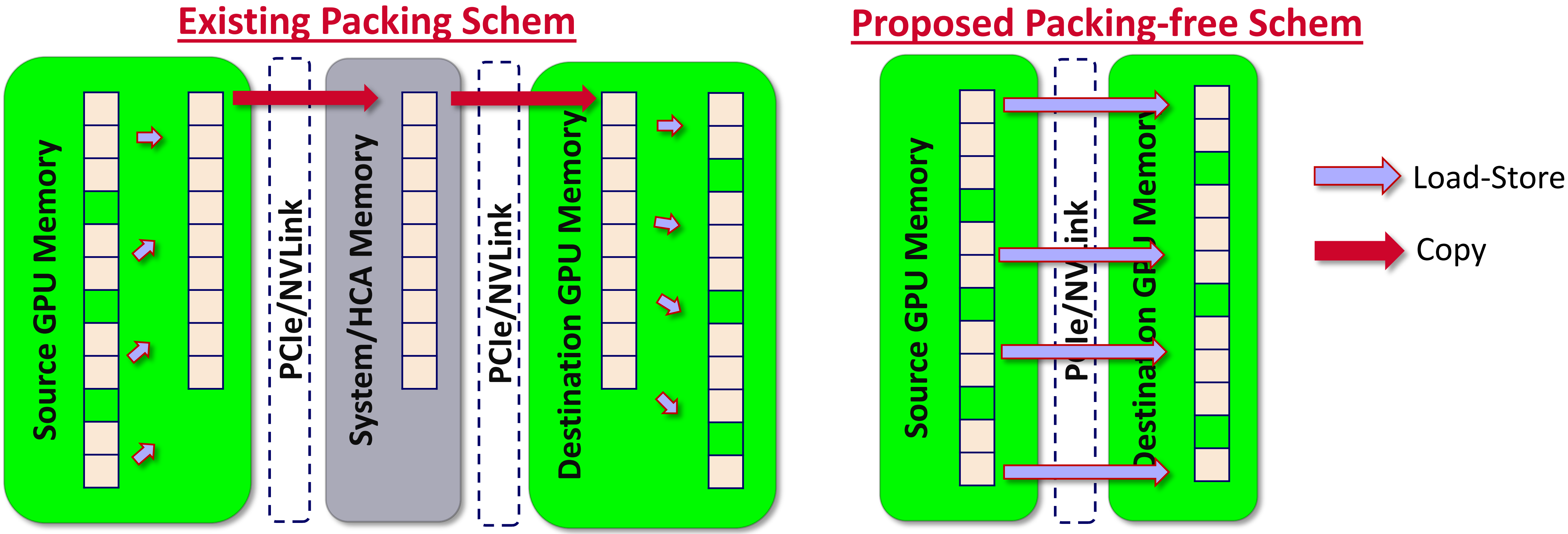
Ching-Hsiang Chu et al., "Exploiting Maximal Overlap for Non-Contiguous Data Movement Processing on Modern GPU-enabled Systems, " IEEE IPDPS 2016.

Outline

- Introduction
- Problem Statement
- **Detailed Description and Results**
  - Scalable Streaming Broadcast for InfiniBand Networks
  - Efficient Scheduling of Non-contiguous Data Transfer
  - **GPU-enabled Zero-copy Transfer for Non-contiguous Data**
  - GPU-enabled Reduction operations
- Ongoing Work and Future Research Directions
- Broader Impact on the HPC Community
- Expected Contributions

Proposed Zero-copy (packing-free) Datatype Transfer

- Exploiting load-store capability of modern interconnects
  - Eliminate extra data copies and expensive packing/unpacking processing

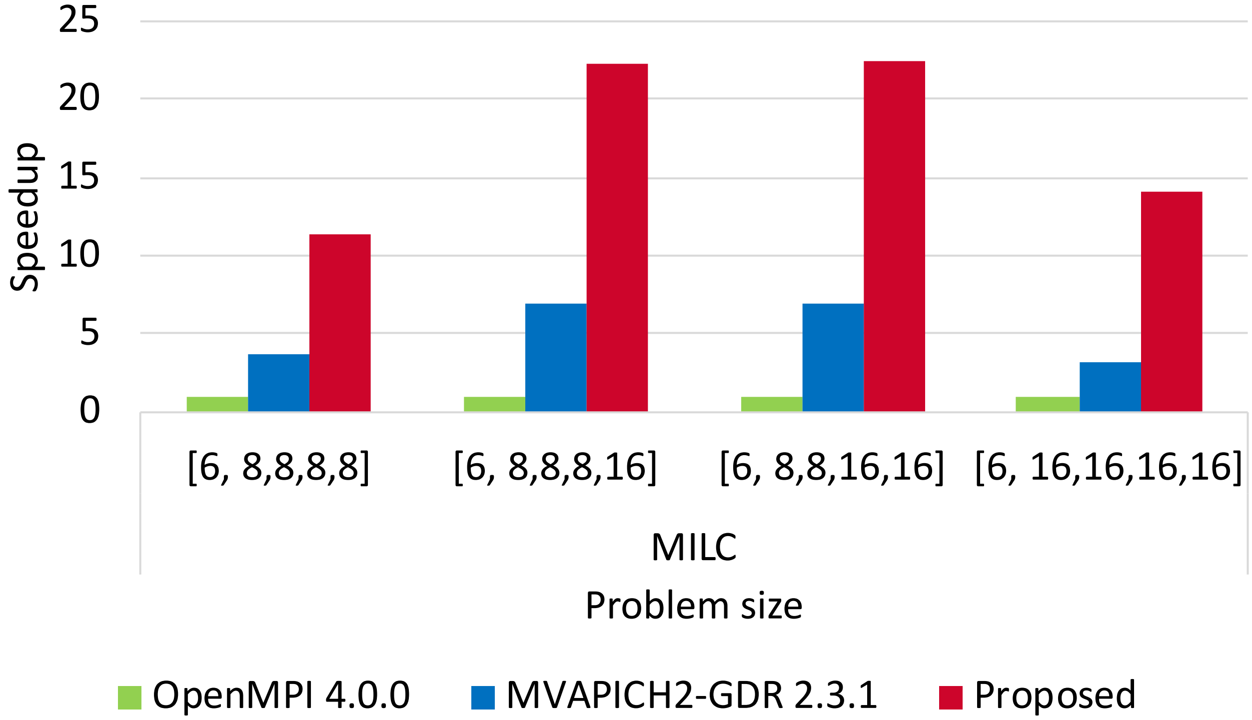


Ching-Hsiang Chu et al., "High-Performance Adaptive MPI Derived Datatype Communication for Modern Multi-GPU Systems", HiPC 2019.

Performance Evaluation

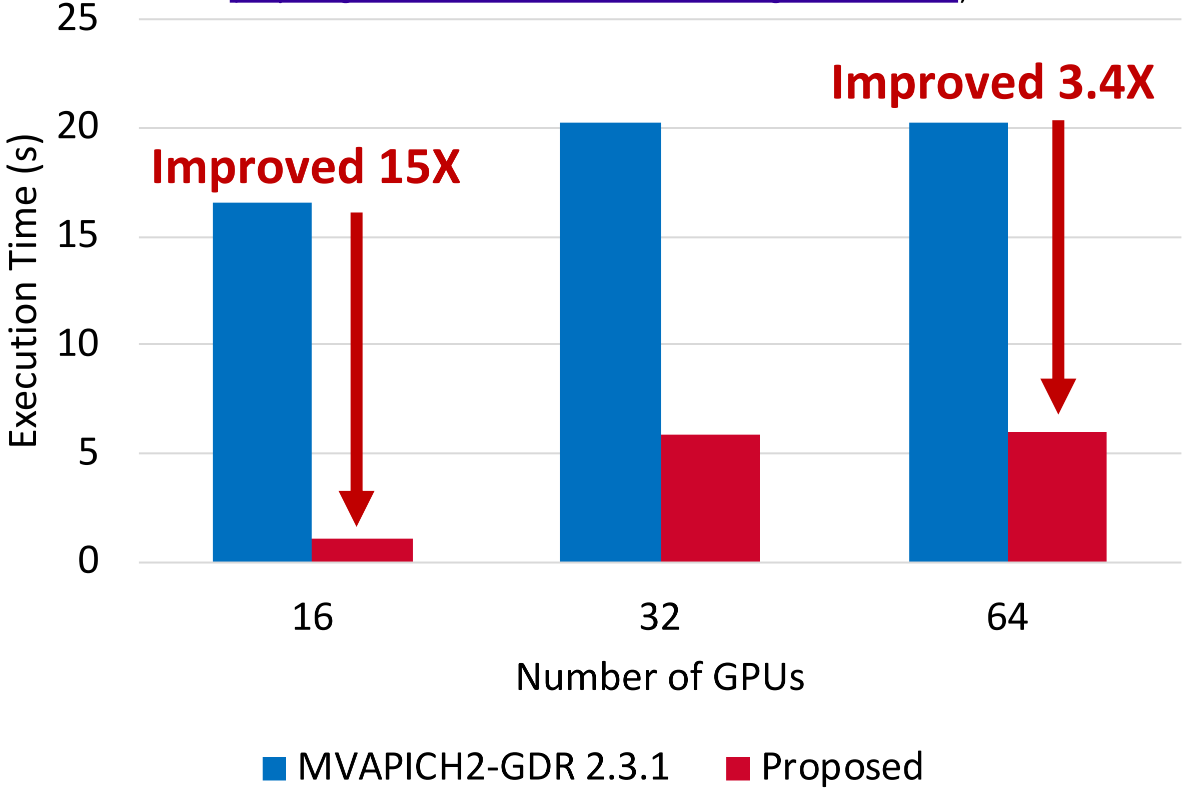
- Zero-copy (packing-free) for GPUs with peer-to-peer direct access over PCIe/NVLink

GPU-based DDTBench mimics MILC communication kernel



Platform: Nvidia DGX-2 system

Communication Kernel of COSMO Model
(https://github.com/cosunae/HaloExchangeBenchmarks)



Platform: Cray CS-Storm

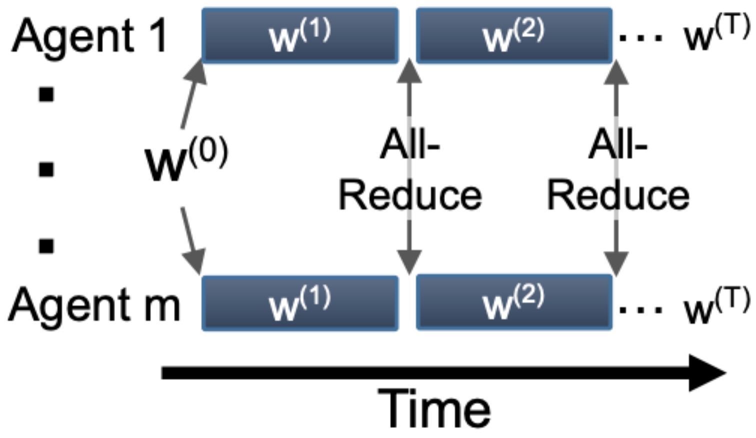
Ching-Hsiang Chu et al., "High-Performance Adaptive MPI Derived Datatype Communication for Modern Multi-GPU Systems", HiPC 2019.

Outline

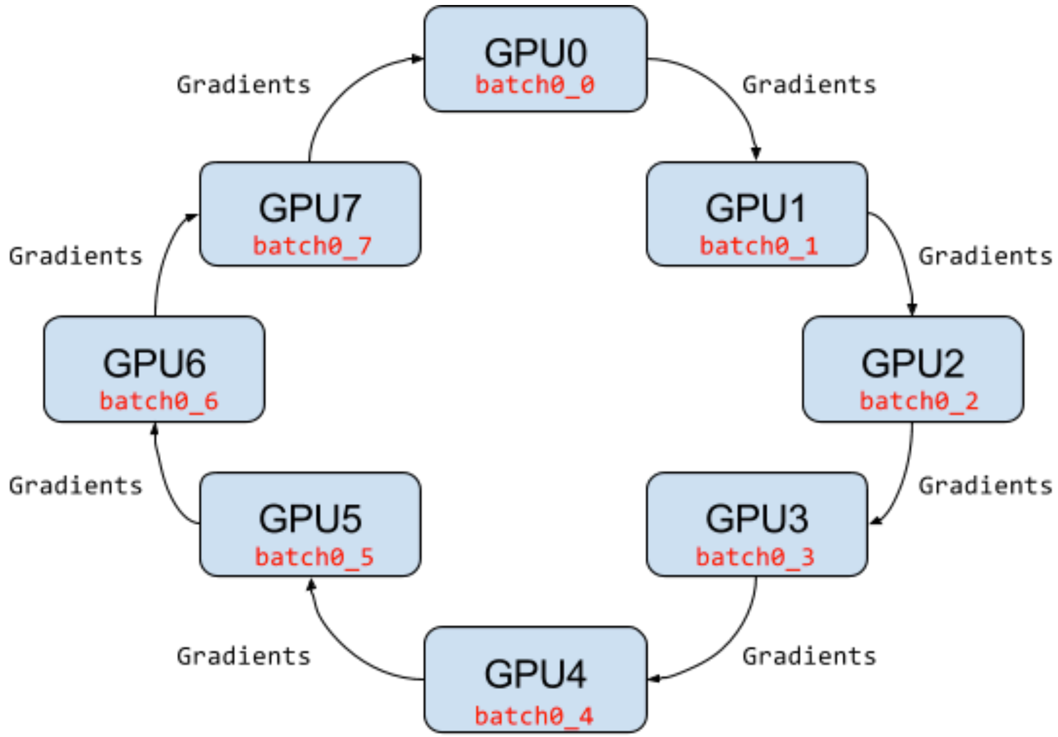
- Introduction
- Problem Statement
- **Detailed Description and Results**
  - Scalable Streaming Broadcast for InfiniBand Networks
  - Efficient Scheduling of Non-contiguous Data Transfer
  - GPU-enabled Zero-copy Transfer for Non-contiguous Data
  - **GPU-enabled Reduction operations**
- Ongoing Work and Future Research Directions
- Broader Impact on the HPC Community
- Expected Contributions

Motivated Example #3 – Reduction Op. for DL Training

- Can GPU resources help improving compute-intensive communications?
  - E.g., MPI_Reduce, MPI_Allreduce, MPI_Scan
  - **Emerging distributed deep learning training**
    - Exchange and update weights
  - Requires fast and high-bandwidth solutions



Ben-Nun T, Hoefler T. Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. arXiv preprint arXiv:1802.09941. 2018 Feb 26.



https://www.oreilly.com/ideas/distributed-tensorflow

How to leverage GPUs for MPI Reduction Operations?

- **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!

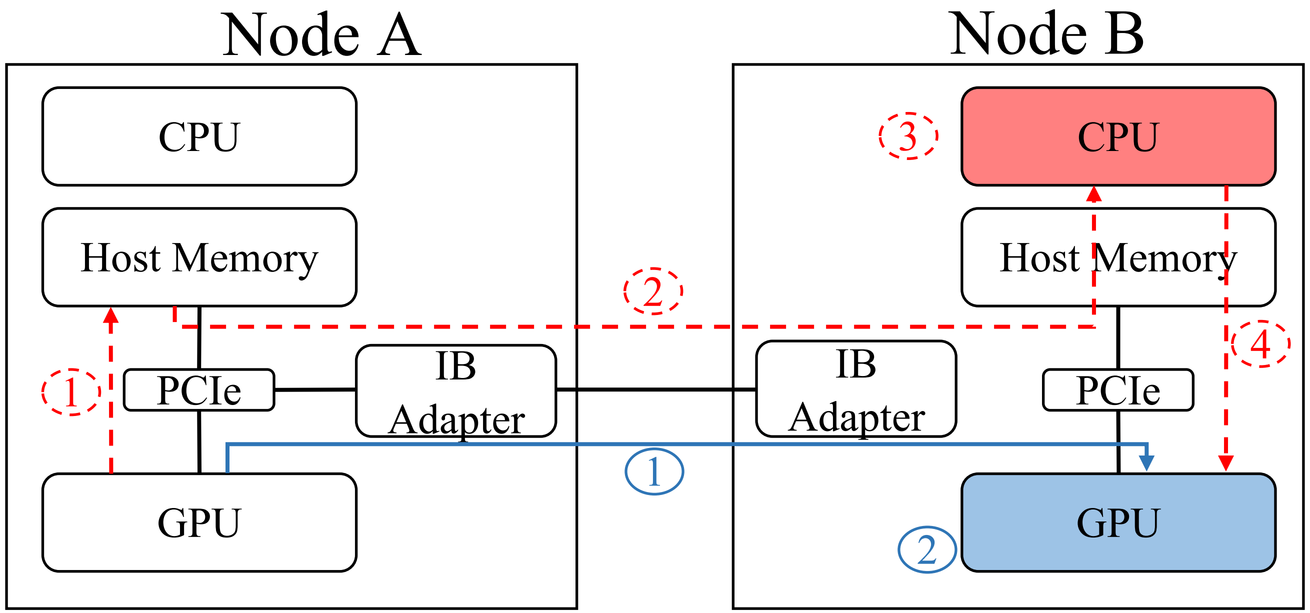
Fast

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



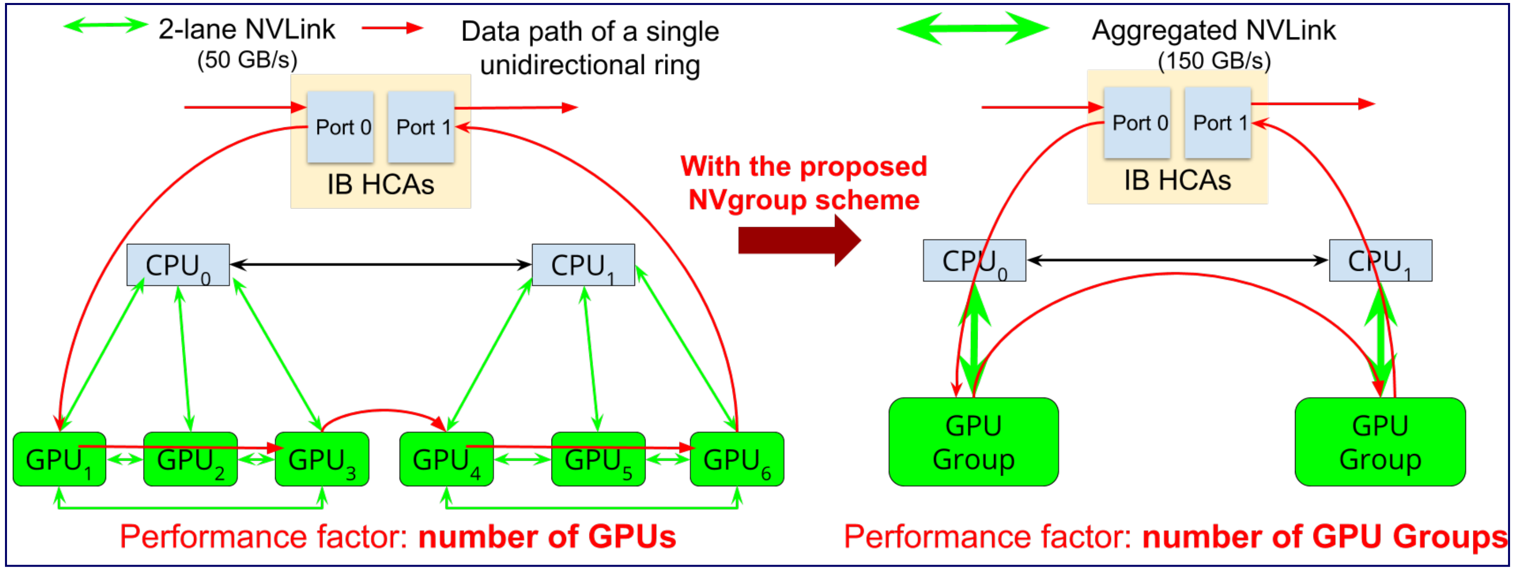
Ching-Hsiang Chu et al., "CUDA Kernel based Collective Reduction Operations on Large-scale GPU Clusters, " IEEE/ACM CCGrid 2016

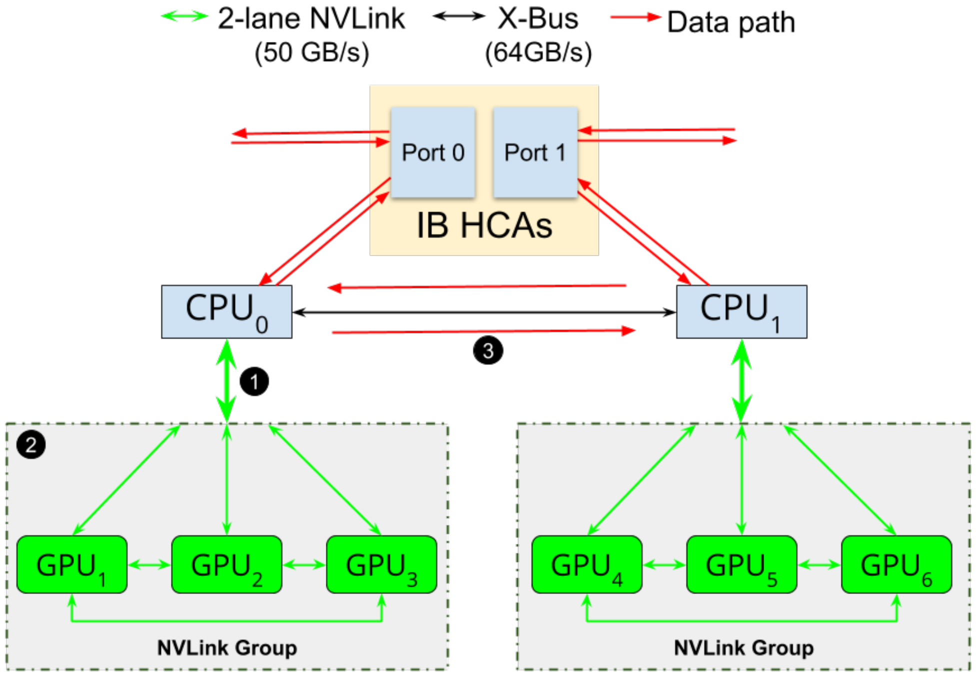
Alternative and Extended Designs

<table>
<tr><th>Communication</th><th>Computation</th><th>Design</th><th>Algorithm</th><th>Benefit</th></tr>
<tr><td rowspan="4">Host<->Host</td><td rowspan="3">CPU</td><td>BR-H-HH (Default)</td><td>Binomial-Reduce</td><td rowspan="2">Large scale, small messages</td></tr>
<tr><td>RD-H-HH (Default)</td><td>Recursive doubling</td></tr>
<tr><td>GR-H-HH</td><td rowspan="4">Gather-Reduce</td><td rowspan="4">Small scale, small messages</td></tr>
<tr><td rowspan="7">GPU</td><td>GR-HH</td></tr>
<tr><td>Host<->Device (GDR)</td><td>GR-HD / GR-DH</td></tr>
<tr><td rowspan="4">Device<->Device (GDR)</td><td>GR-DD</td></tr>
<tr><td>BR-DD</td><td>Binomial-Reduce</td><td rowspan="4">Large messages for any scale</td></tr>
<tr><td>BRB-DD</td><td>Binomial-Reduce-Bcast</td></tr>
<tr><td>RD-DD</td><td rowspan="2">Recursive doubling</td></tr>
<tr><td>Host<->Device (GDR)</td><td>RD-HD/RD-DH</td></tr>
</table>

Proposed NVGroup Allreduce

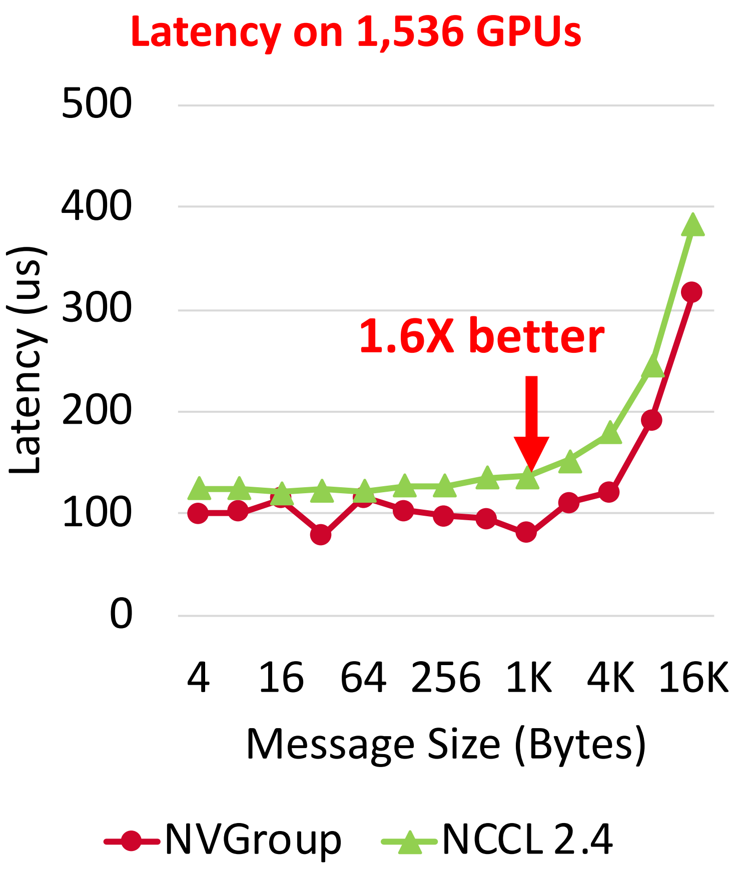
- Grouping GPUs which are fully connected by NVLinks
  - Contention-free communication within the group
- Cooperative Reduction Kernels to exploit load-compute-store primitives over NVLinks

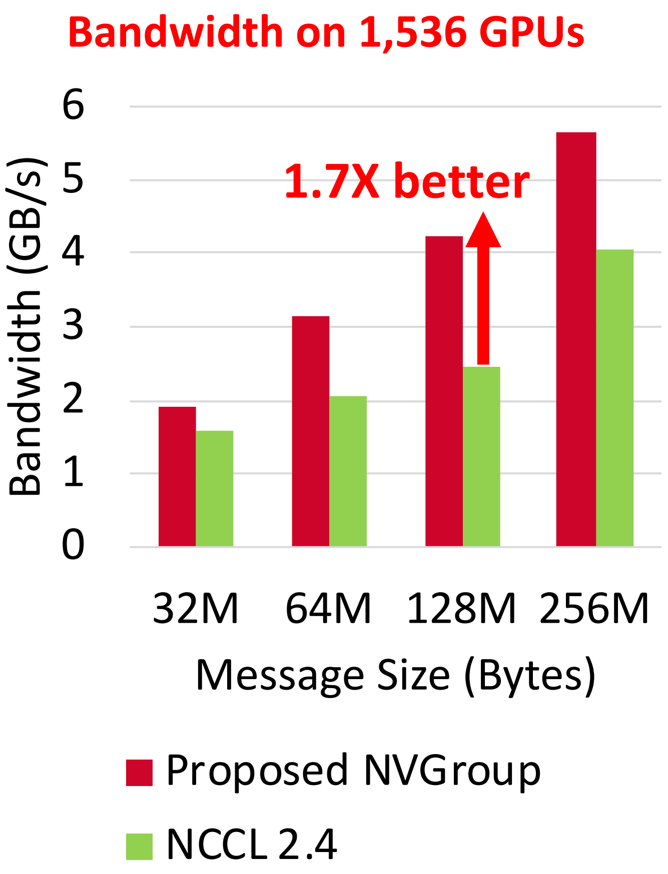


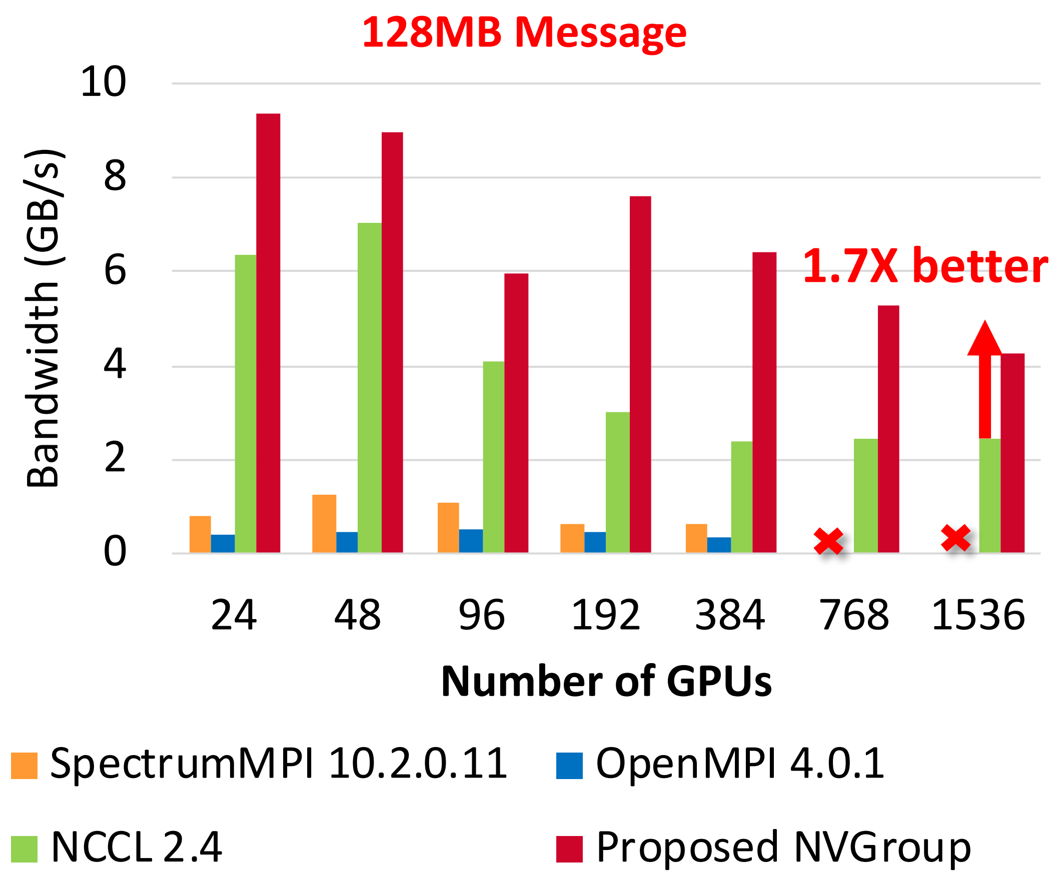


Ching-Hsiang Chu et al., "NV-Group: Cooperative and Link-Efficient Reductions for Deep Learning on NVLink-enabled Dense GPU Systems, " (to be submitted)

Preliminary Results – Allreduce Benchmark





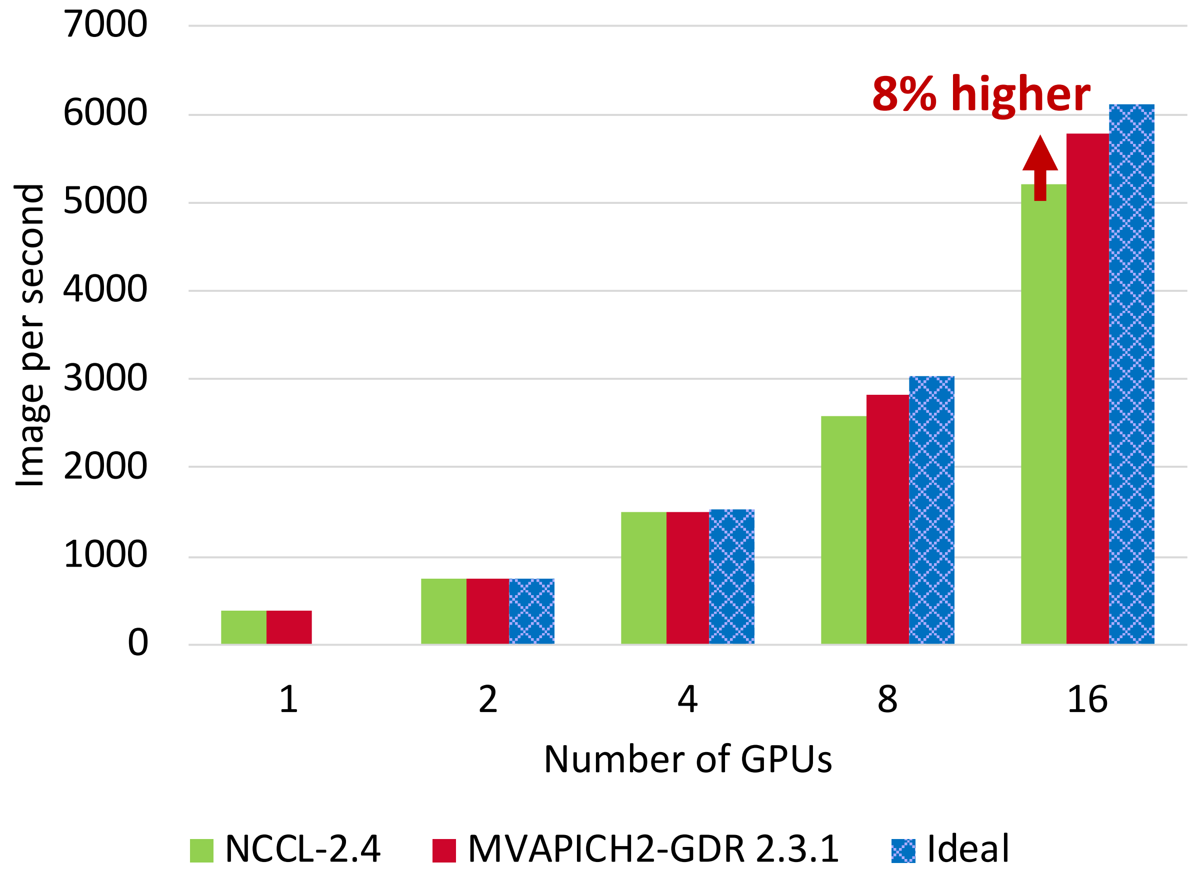


#1 Summit Platform: Dual-socket IBM POWER9 CPU, 6 NVIDIA Volta V100 GPUs, and 2-port InfiniBand EDR Interconnect

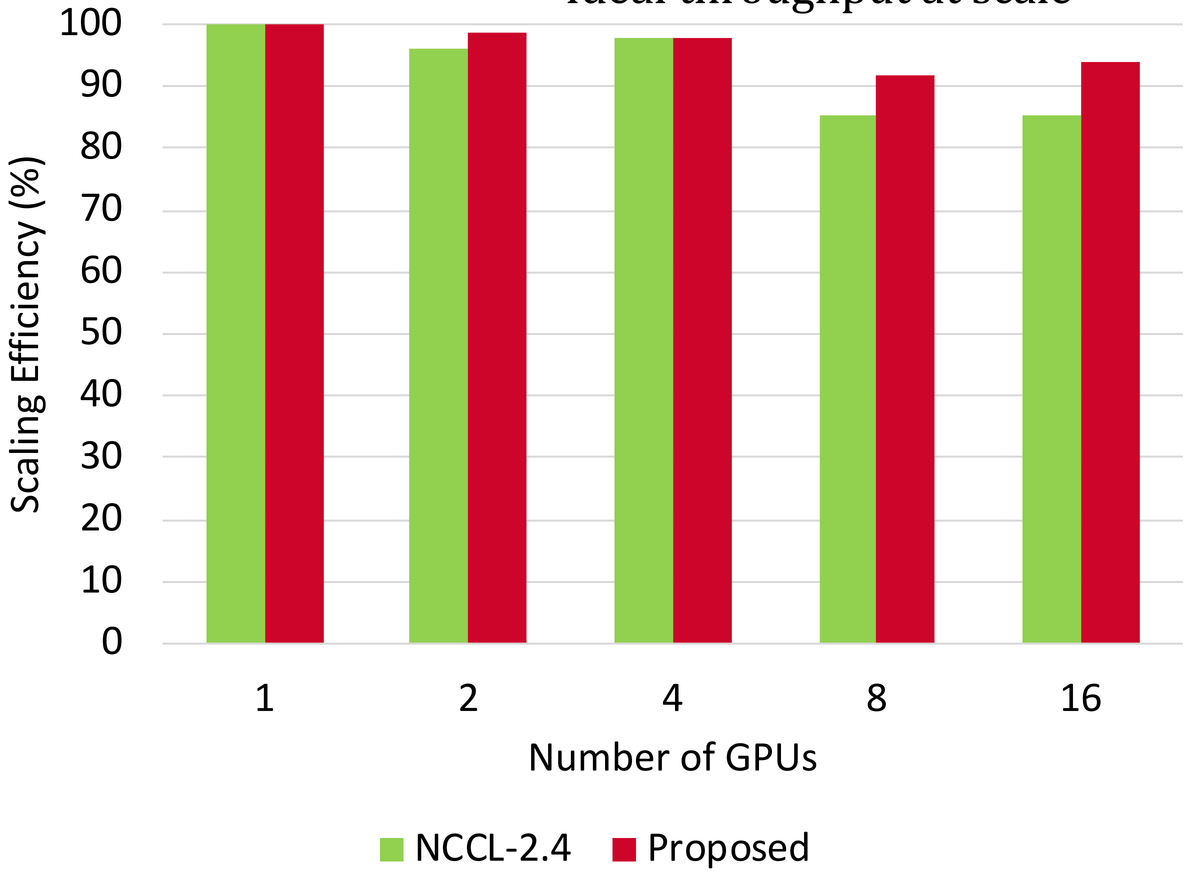
Ching-Hsiang Chu et al., "NV-Group: Cooperative and Link-Efficient Reductions for Deep Learning on NVLink-enabled Dense GPU Systems, " (to be submitted)

Preliminary Results – Distributed Deep Learning Training

- ResNet-50 Training using TensorFlow benchmark on a DGX-2 machine (16 Volta GPUs)



$$\text{Scaling Efficiency} = \frac{\text{Actual throughput}}{\text{Ideal throughput at scale}} \times 100\%$$



Ching-Hsiang Chu et al., "NV-Group: Cooperative and Link-Efficient Reductions for Deep Learning on NVLink-enabled Dense GPU Systems, " (to be submitted)

Outline

- Introduction
- Problem Statement
- Detailed Description and Results
- Ongoing Work and Future Research Directions
- **Broader Impact on the HPC Community**
- Expected Contributions

MVAPICH2 Project

- High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.1), Started in 2001, First version available in 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2011
  - Support for GPGPUs (MVAPICH2-GDR) and MIC (MVAPICH2-MIC), Available since 2014
  - Support for Virtualization (MVAPICH2-Virt), Available since 2015
  - Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  - Support for InfiniBand Network Analysis and Monitoring (OSU INAM) since 2015
  - **Used by more than 3,000 organizations in 89 countries**
  - **More than 553,000 (> 0.5 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (June ‘19 ranking)
    - 3rd ranked 10,649,640-core cluster (Sunway TaihuLight) at NSC, Wuxi, China
    - 16th, 556,104 cores (Oakforest-PACS) in Japan
    - 19th, 367,024 cores (Stampede2) at TACC
    - 31st, 241,108-core (Pleiades) at NASA and many others
  - Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, and OpenHPC)
  - http://mvapich.cse.ohio-state.edu

Empowering Top500 systems for over a decade

Partner in the 5th ranked TACC Frontera System

Impact to the Community

- **Accelerating GPU-enabled HPC applications worldwide**
  - MVAPICH2-GDR is widely used on many top-ranked GPU clusters worldwide including Summit (#1), Sierra (#2), ABCI (#8), Lassen (#10) and more
- **Enabling fast weather forecasting**
  - MeteoSwiss uses the COSMO numerical weather forecasting model for the production of regional and local forecast products
  - Use MVAPICH2-GDR to accelerate GPU Communication
- **Supporting scalable and reliable data dissemination**
  - DoD streaming applications are using MVAPICH2-GDR to accelerate GPU-based broadcast operations
  - Tutorial sessions: PETTT ’17, PETTT ‘18

Outline

- Introduction
- Problem Statement
- Detailed Description and Results
- Ongoing Work and Future Research Directions
- Broader Impact on the HPC Community
- **Expected Contributions**

Expected Contributions

- Improving Scale-out and Scale-up performance by exploiting features in modern Interconnects
  - Enabling Scalable Broadcast by using InfiniBand hardware multicast and GPUDirect RDMA
  - Link-efficient schemes by exploiting load-store primitives over GPU interconnects
- Efficient GPU-enabled Communication Middleware
  - GPU-enabled MPI derived datatype processing
  - GPU-enabled reduction operations
- Significant impact on the community
  - The abstraction for accelerator-enabled communication middleware
  - Benefit HPC & ML/DL workloads
  - Broader outreach through MVAPICH2-GDR public releases

Thank You!

Questions?

chu.368@osu.edu

http://web.cse.ohio-state.edu/~chu.368