MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

THE OHIO STATE UNIVERSITY

Efficient Non-contiguous Data Transfer using MVAPICH2-GDR for GPU-enabled HPC Applications

Ching-Hsiang Chu

chu.368@osu.edu

Ph.D. Candidate
Department of Computer Science and Engineering
The Ohio State University

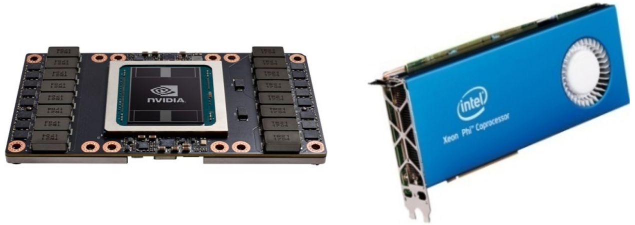
Outline

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

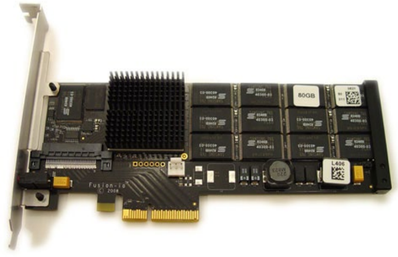
Trends in Modern HPC Architecture: Heterogeneous

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)

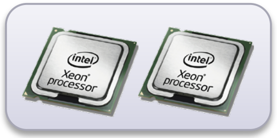
#2 Sierra (17,280 GPUs)
#10 Lassen (2,664 GPUs)

#8 ABCI
(4,352 GPUs)

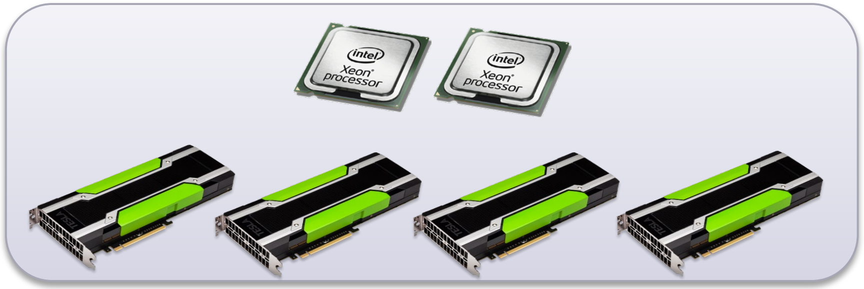
#22 DGX SuperPOD
(1,536 GPUs)

Architectures: Past, Current, and Future

Multi-core CPUs within a node

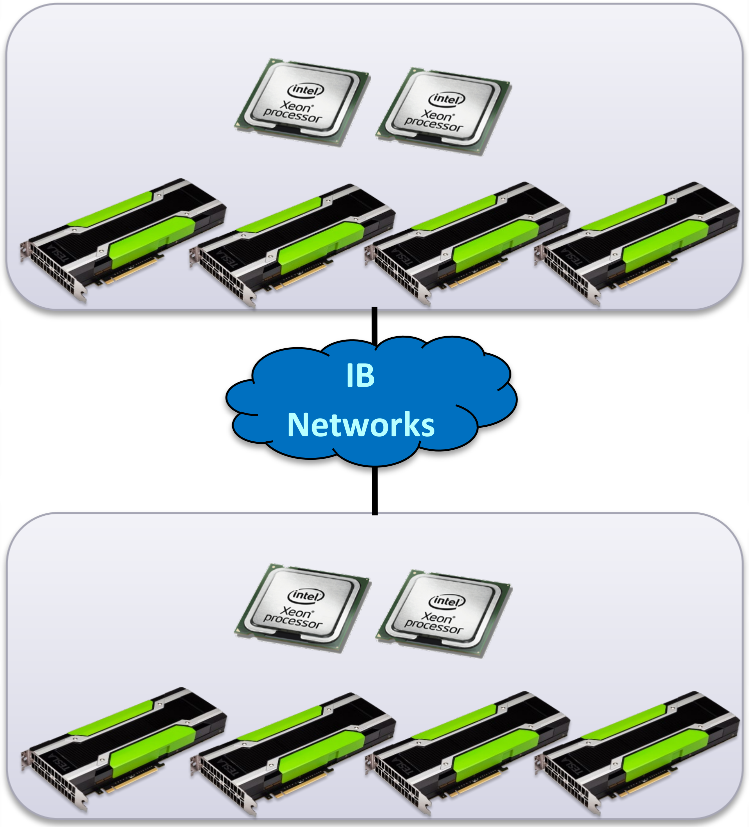


Multi-core CPUs + Multi-GPU within a node

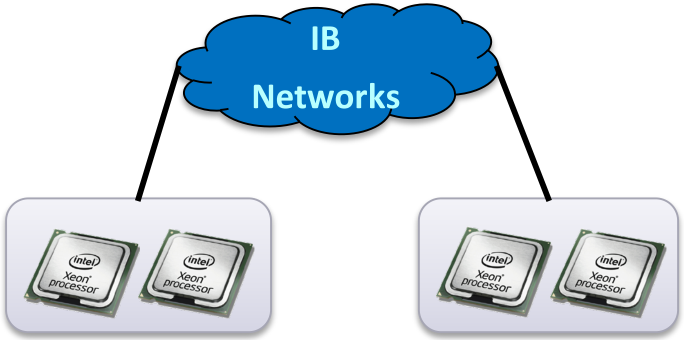


Multi-core CPUs + Multi-GPU across nodes

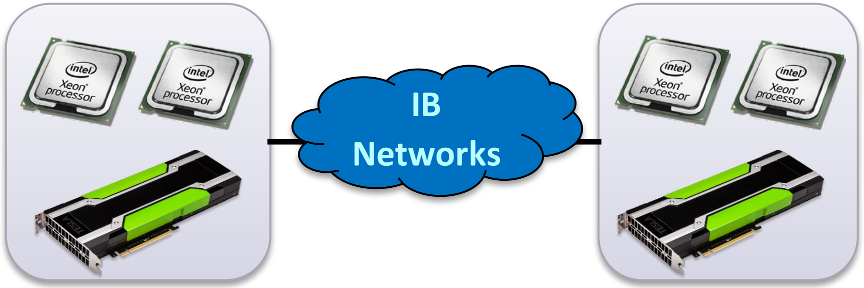
(E.g., Sierra/Summit, Frontier)



Multi-core CPUs across nodes



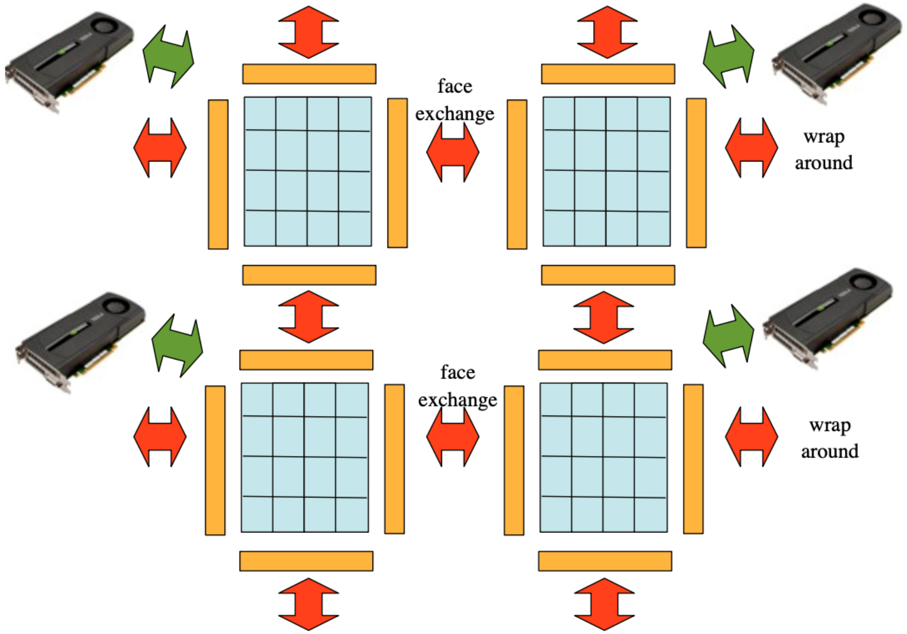
Multi-core CPUs + Single GPU across nodes



Motivated Example – Non-contiguous Data Transfer

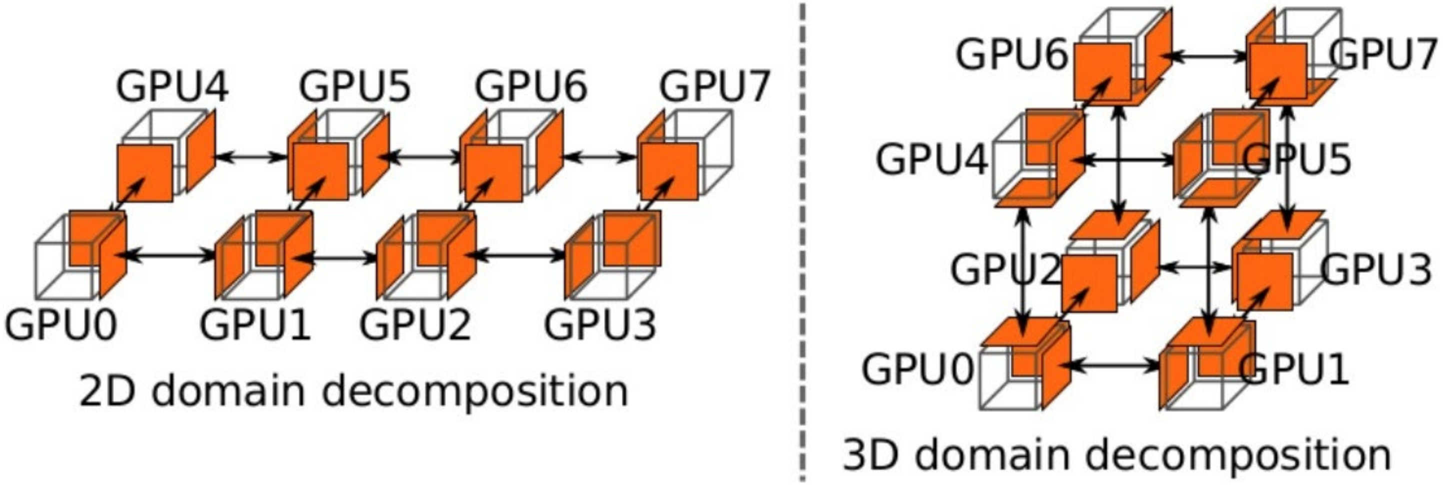
- Wide usages of MPI derived datatype for Non-contiguous Data Transfer
  - Requires Low-latency and high overlap processing

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

Outline

- **Introduction**
- **Advanced Designs in MVAPICH2-GDR**
  - Asynchronous designs for Maximizing Overlap
  - Zero-copy (Pack-free) on Dense-GPU systems
- **Concluding Remarks**

Existing GPU-enabled MPI Datatype Processing

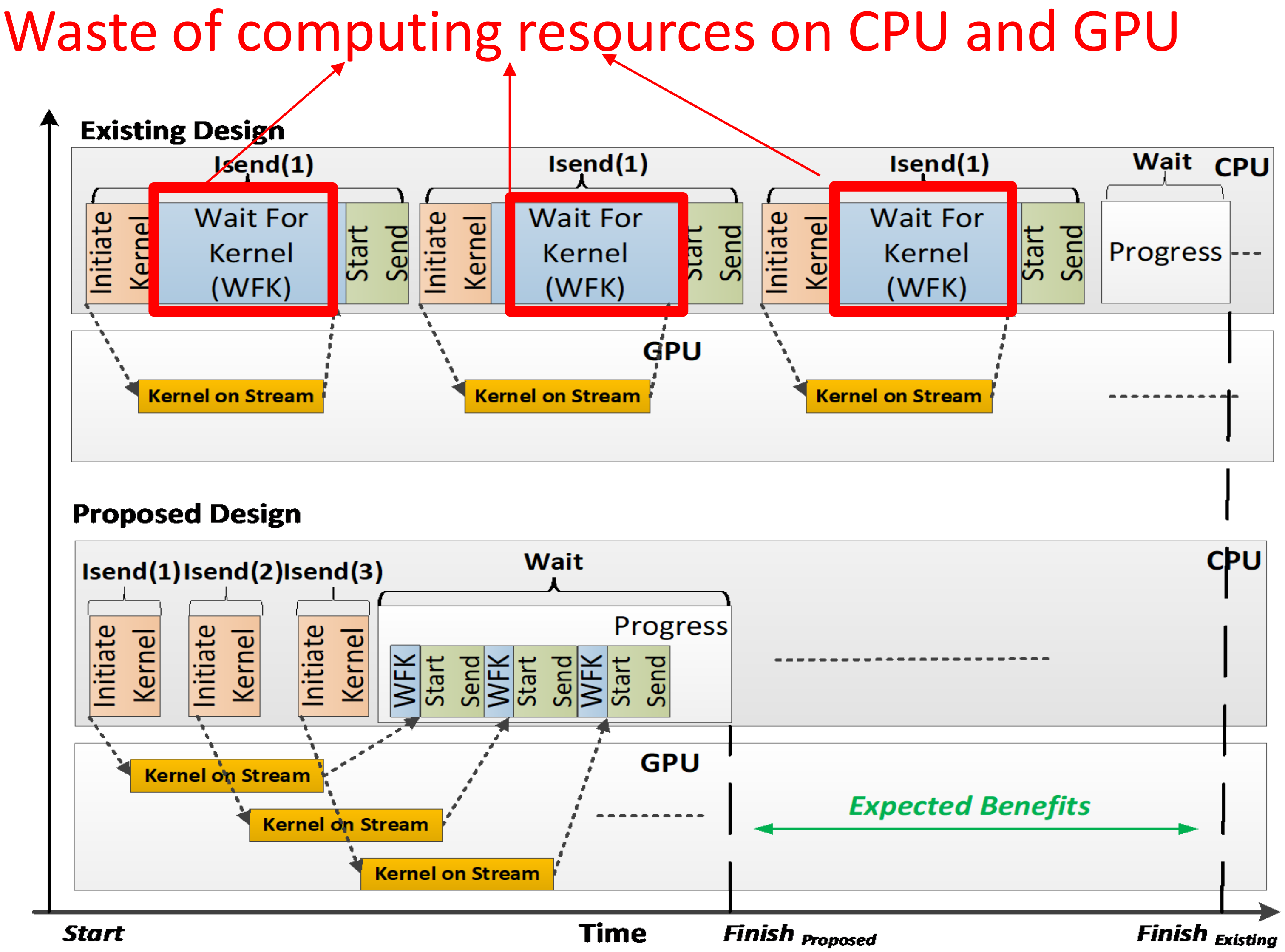
Common Scenario

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

MPI_Waitall (…);

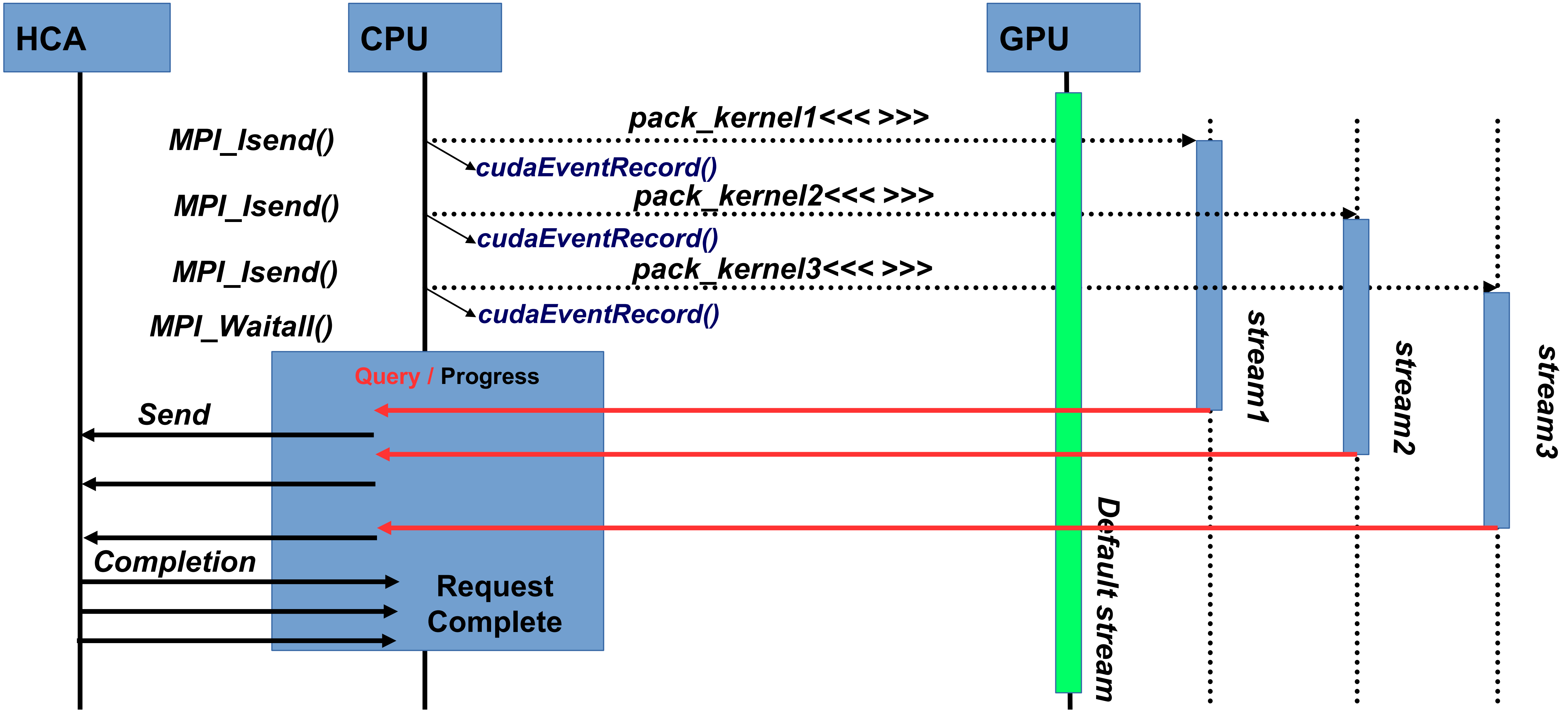
*A, B…contain non-contiguous MPI Datatype

Waste of computing resources on CPU and GPU

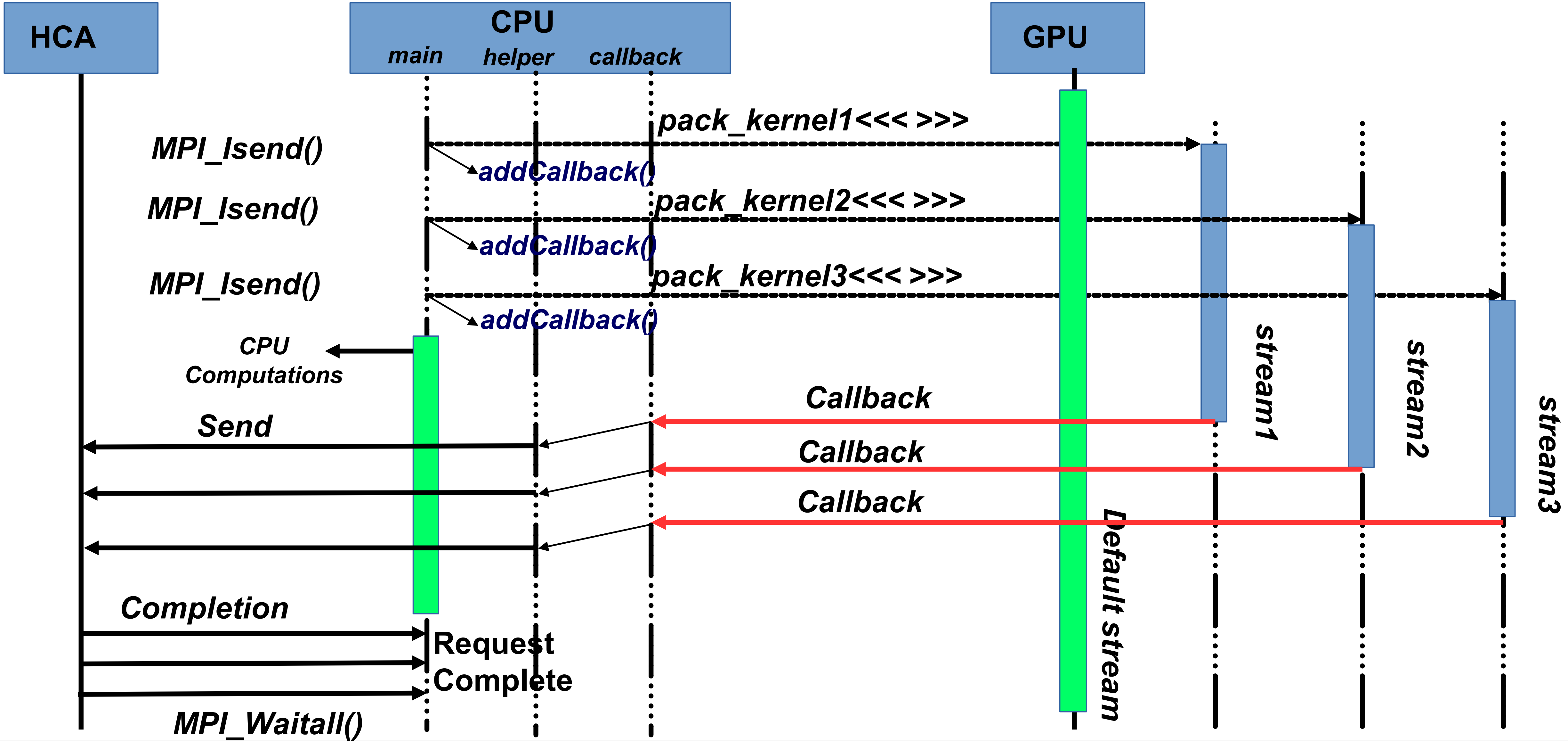


Ching-Hsiang Chu et al., "Exploiting Maximal Overlap for Non-Contiguous Data Movement Processing on Modern GPU-enabled Systems, " IEEE IPDPS 2016.

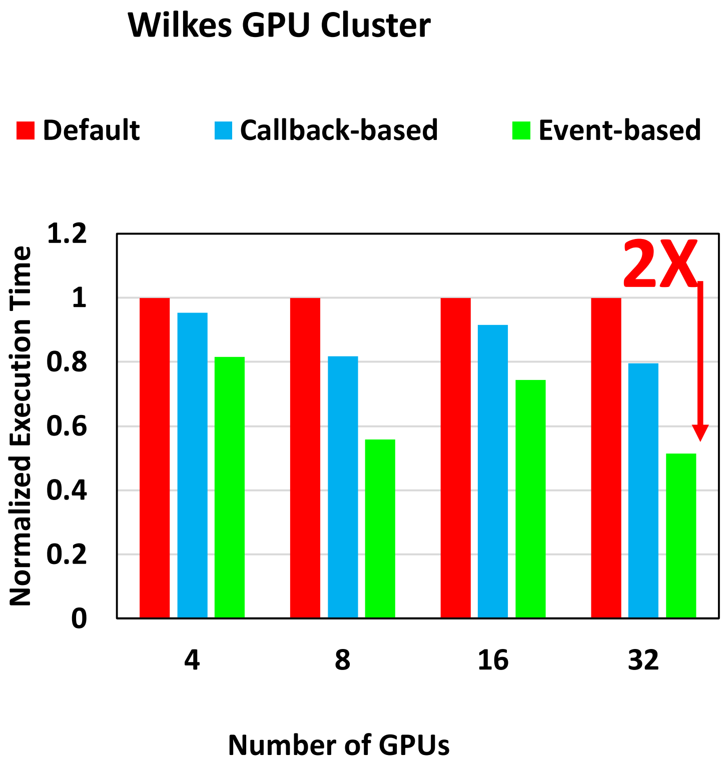
Proposed Event-based Design – Low Latency

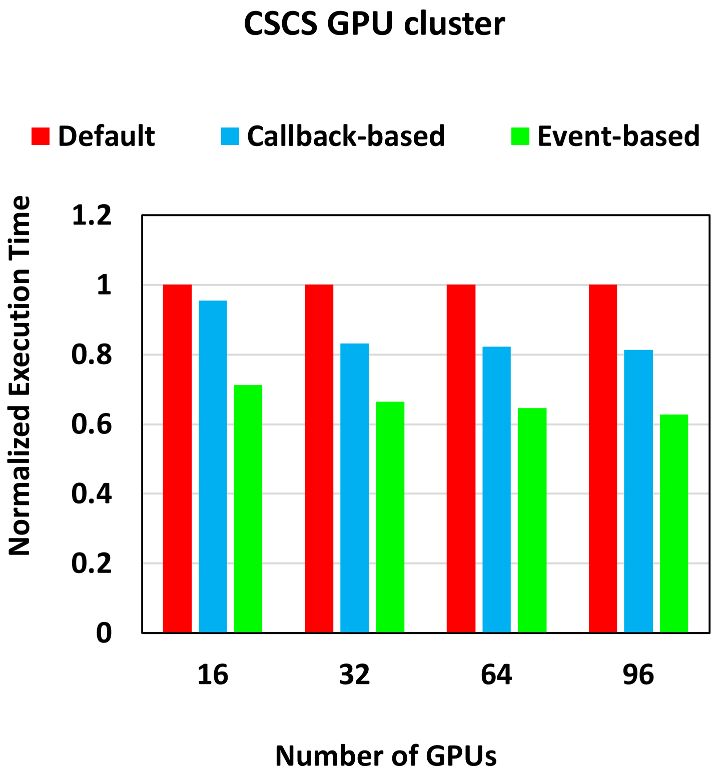


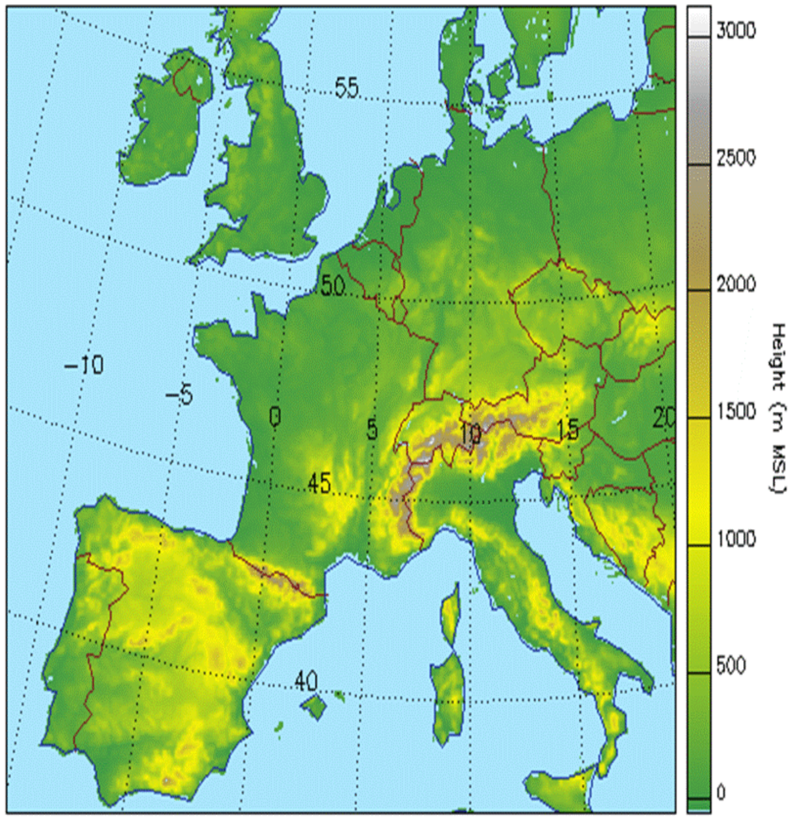
Proposed Callback-based Design – High Overlap



Application-Level Evaluation (Cosmo) and Weather Forecasting in Switzerland







Cosmo model: http://www2.cosmo-model.org/content/tasks/operational/meteoSwiss/

- 2X improvement on 32 GPUs nodes
- 30% improvement on 96 GPU nodes (8 GPUs/node)

On-going collaboration with CSCS and MeteoSwiss (Switzerland) in co-designing MV2-GDR and Cosmo Application

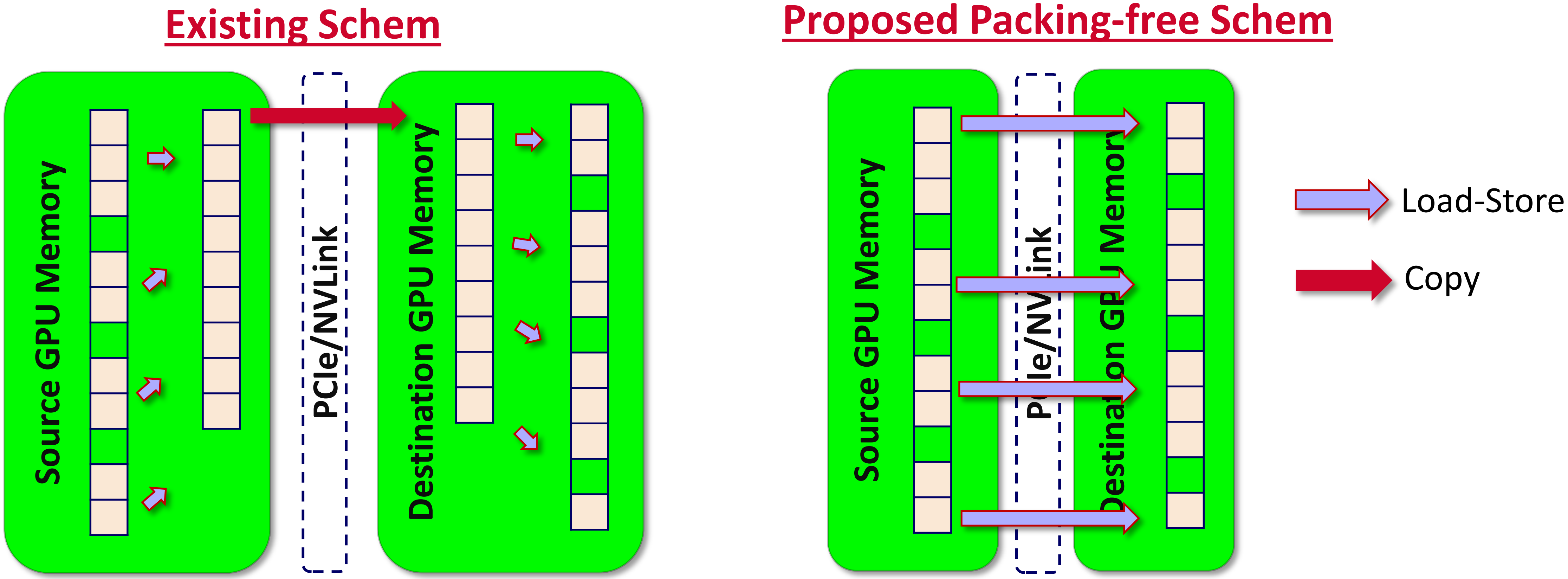
C. Chu, K. Hamidouche, A. Venkatesh, D. Banerjee , H. Subramoni, and D. K. Panda, Exploiting Maximal Overlap for Non-Contiguous Data Movement Processing on Modern GPU-enabled Systems, IPDPS'16

Outline

- **Introduction**
- **Advanced Designs in MVAPICH2-GDR**
  - Asynchronous designs for Maximizing Overlap
  - Zero-copy (Pack-free) on Dense-GPU systems
- **Concluding Remarks**

Proposed Zero-copy (packing-free) Datatype Transfer

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

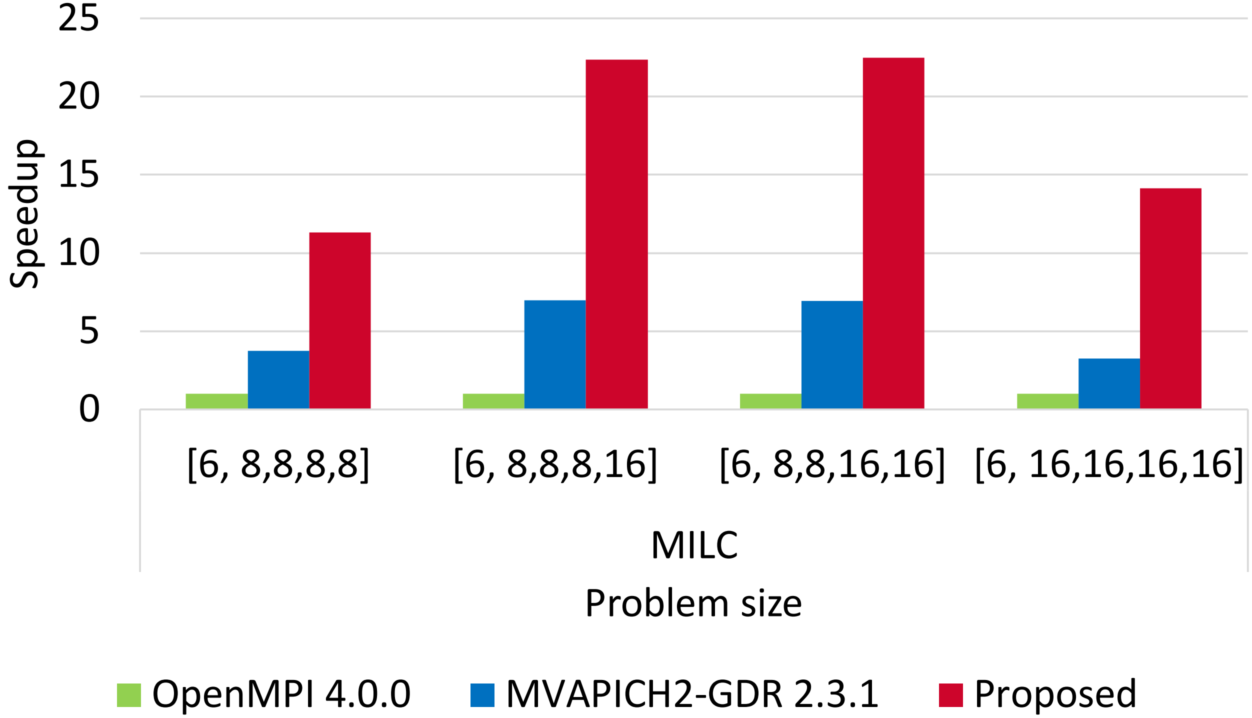


Ching-Hsiang Chu et al., “High-Performance Adaptive MPI Derived Datatype Communication for Modern Multi-GPU Systems”, to appear in HiPC 2019.

Performance Evaluation

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

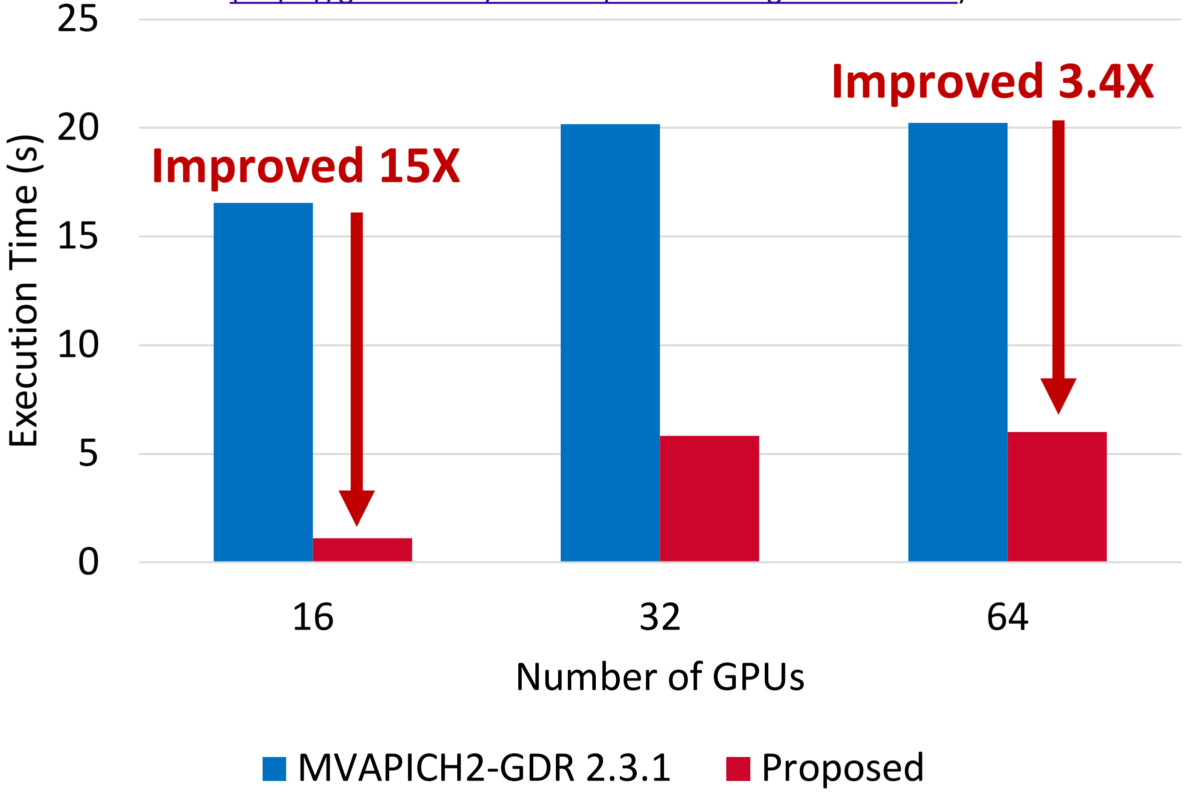
GPU-based DDTBench mimics MILC communication kernel



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", to appear in HiPC 2019.

Outline

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

Concluding Remarks

- **Efficient MPI derived datatype processing for GPU-resident data**
  - Asynchronous GPU kernels to achieve high overlap between communication and computation
  - Zero-copy schemes for Dense-GPU with high-speed interconnects like PCIe and NVLink
- **These features are included since MVAPICH2-GDR 2.3.2**
  - http://mvapich.cse.ohio-state.edu/
  - http://mvapich.cse.ohio-state.edu/userguide/gdr/

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/