MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

Network Assisted Non-Contiguous Transfers for GPU-Aware MPI Libraries

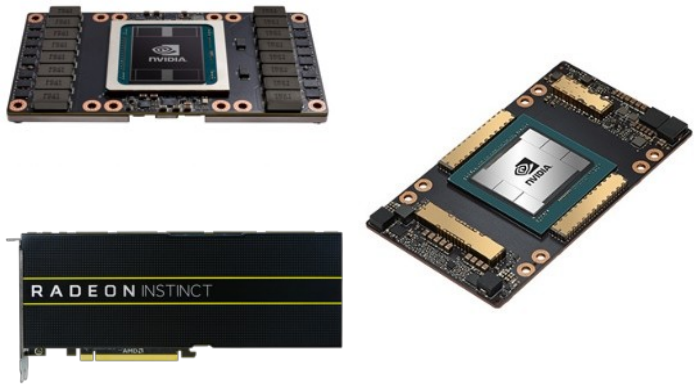
Presentation at the 10th Annual MVAPICH User Group (MUG) Meeting (MUG '22)

Kaushik Kandadi Suresh, Kawthar Shafie Khorassani, Chen-Chun Chen,
Bharath Ramesh, Mustafa Abduljabbar, Aamir Shafi, Hari Subramoni and
Dhabaleswar K. Panda

E-mail: {kandadisuresh.1, shafiekhorassani.1, chen.10252, ramesh.113, abduljabbar.1, shafi.16, subramoni.1, panda.2} @osu.edu

The Ohio State University

Trends in Modern HPC Clusters

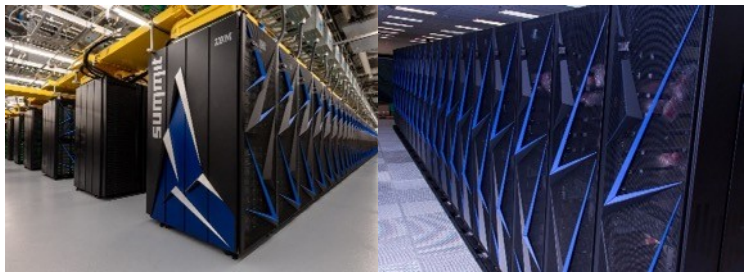


Accelerators (NVIDIA, AMD GPUs)
High compute power
High peak memory bandwidth (A100: 12440 Gb/s memory

- Increased use of GPUs on modern clusters due to high compute capability and power efficiency.
- MPI is widely used in these systems for large scale parallel applications
- Non-Contiguous data exchanges are prevalent in many of these MPI based applications

#1 Frontier
(37,000 GPUs)

#2 Fugaku
(158,976 nodes with A64FX ARM CPU, a GPU-like processor)



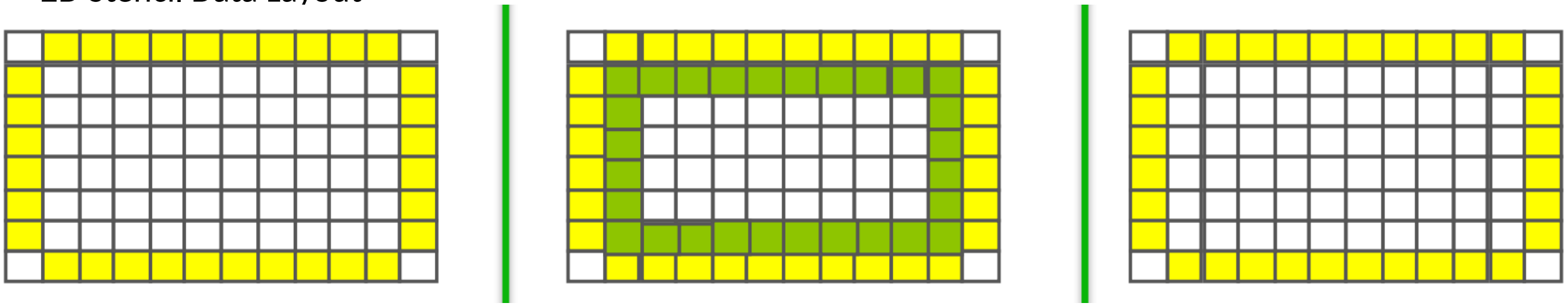
#4 Summit (27,648 GPUs)
#5 Sierra (17,280 GPUs)

#8 Selene
NVIDIA DGX A100 SuperPOD
(2,240 GPUs)

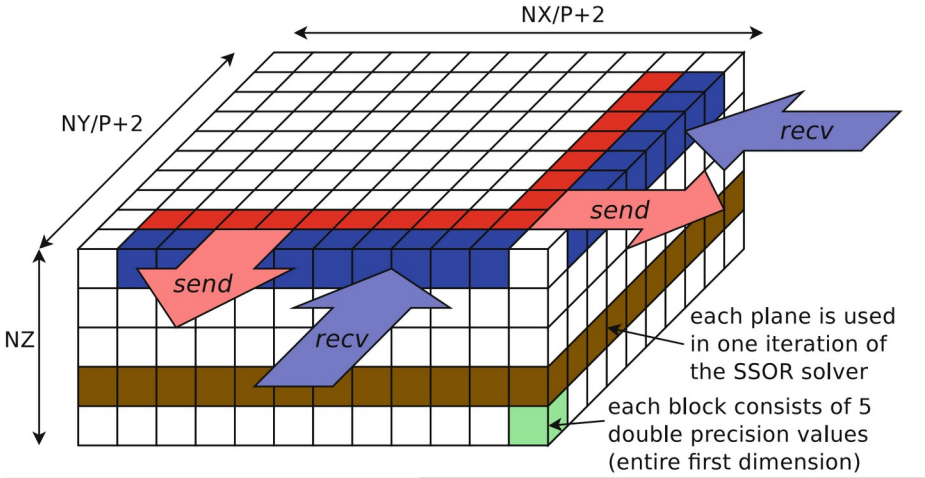
https://www.top500.org/

Existence of Non-Contiguous Exchanges in HPC

- 2D Stencil Data Layout



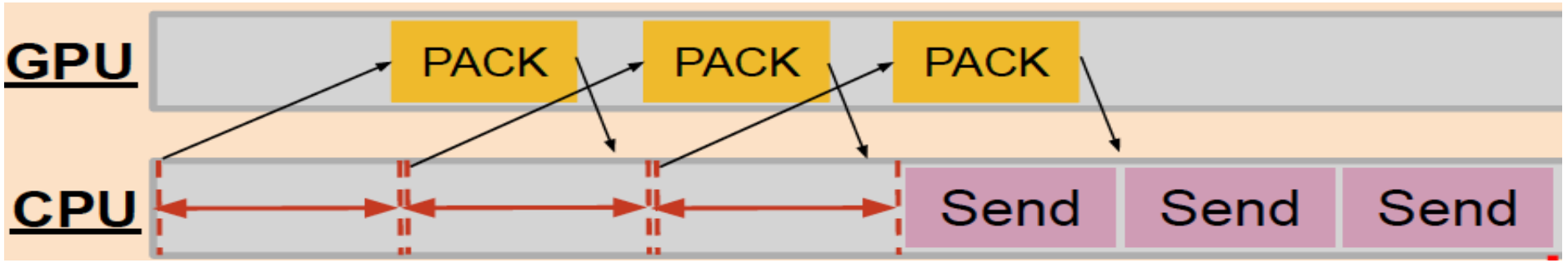
- Data Layout in NAS LU



1) https://www.mcs.anl.gov/~thakur/sc16-mpi-tutorial/slides.pdf
2) Schneider T., Gerstenberger R., Hoefler T. (2012) Micro-applications for Communication Data Access Patterns and MPI Datatypes. In: Träff J.L., Benkner S., Dongarra J.J. (eds) Recent Advances in the Message Passing Interface. EuroMPI

State-of-the-Art GPU Non-Contiguous schemes

- Optimized pack-unpack kernels [1,3]
- Efficient overlap of pack-unpack kernel with data exchange[2]

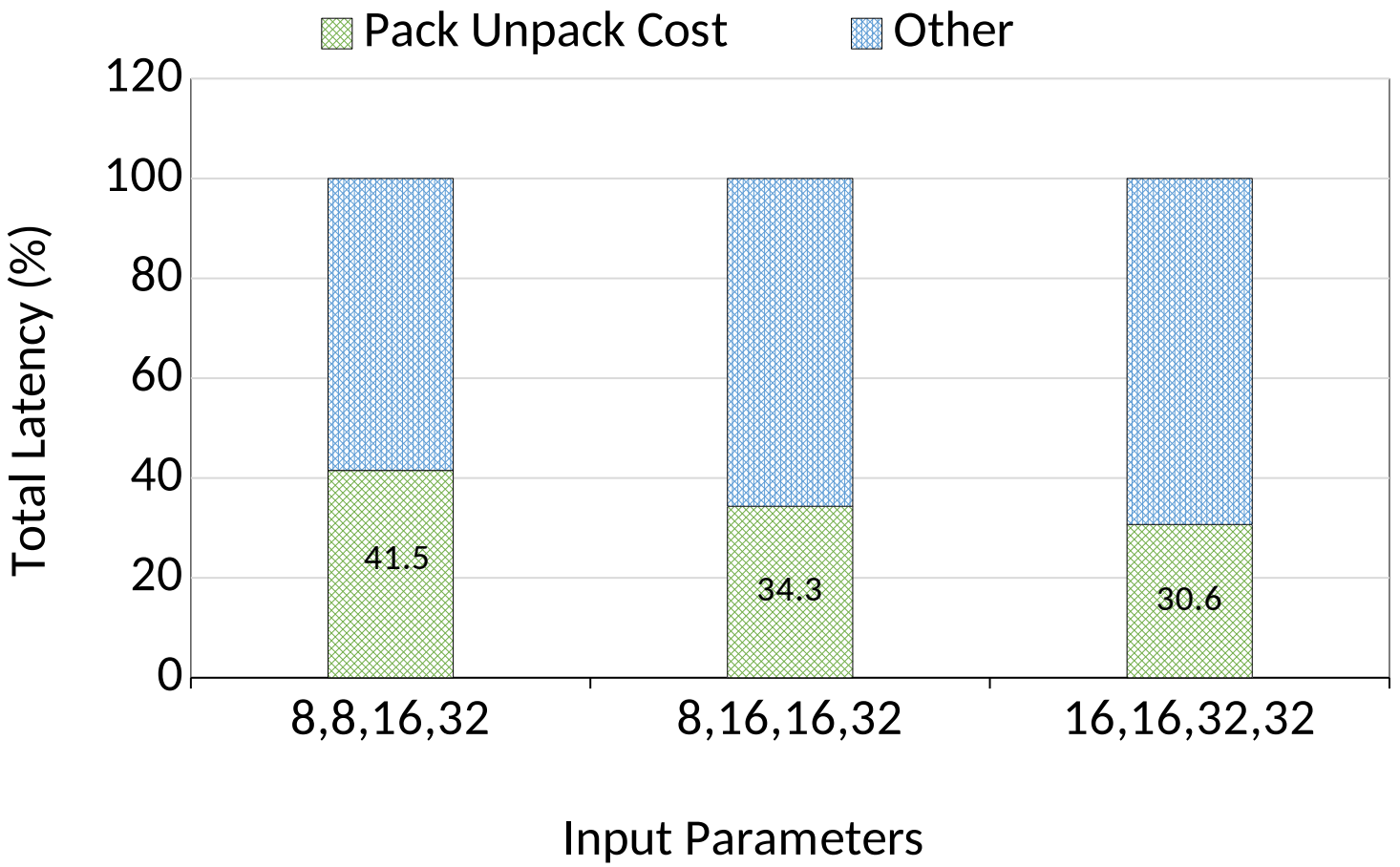


- Fusing multiple kernels to amortize launch overhead[4]

[1] R. Shi, X. Lu, S. Potluri, K. Hamidouche, J. Zhang and D. K. Panda, "HAND: A Hybrid Approach to Accelerate Non-contiguous Data Movement Using MPI Datatypes on GPU Clusters," ICPP 2014.
[2] 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 2016.
[3] Wei Wu, George Bosilca, Rolf vandeVaart, Sylvain Jeaugey, and Jack Dongarra. “GPU-Aware Non-contiguous Data Movement In Open MPI, “ HPDC 2016.
[4] C. -H. Chu, K. S. Khorassani, Q. Zhou, H. Subramoni and D. K. Panda, "Dynamic Kernel Fusion for Bulk Non-contiguous Data Transfer on GPU Clusters,” CLUSTER 2020

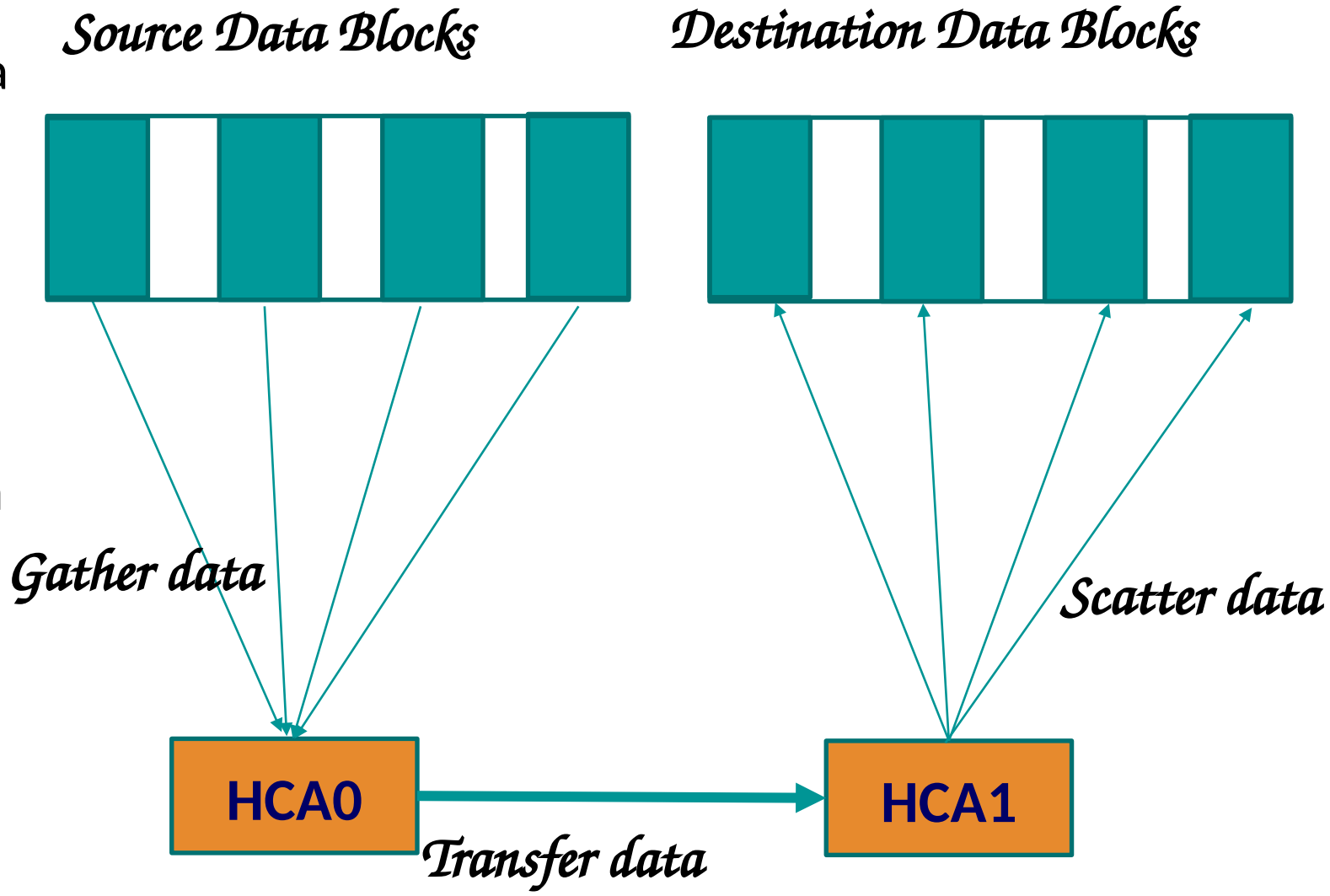
Pack Overhead

- Modified DDT bench that exchanges non-contiguous GPU buffers.
- MIMD Lattice Computation (MILC) application layouts from DDT bench are used.
- Pack cost up-to 40% of total transfer time



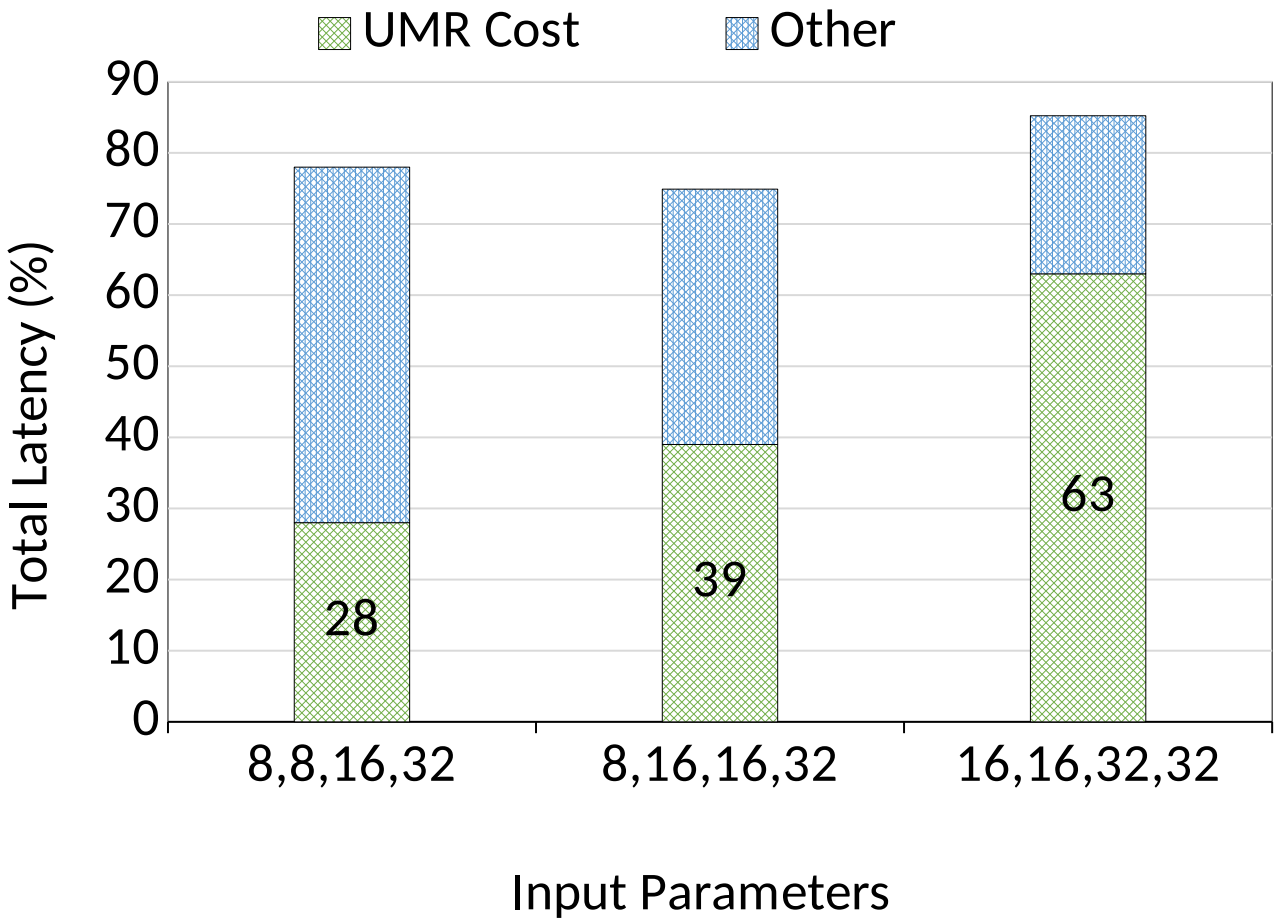
Pack free Non-Contiguous Exchange in modern HCAs

- Modern HCAs support direct exchange of non-contiguous data
- Source HCA gathers data from source GPU buffer
- Destination HCA scatter data to destination GPU buffer
- User-Mode Memory Registration (UMR) is a registration mode introduced by Mellanox. This allows us to map and exchange non-contiguous memory regions without host/GPU level packing.



UMR Overheads

- UMR requires mkey creation and mapping to non-contiguous buffers
- mkey creation could cost up-to 200 us
- mkey mapping of can consume up-to 60% of the total time.
- Sender and Receiver may need to exchange several mkeys depending on layouts which could degrade the performance



Contributions

- Propose UMR-based design for exchanging non-contiguous data
- Enhance the proposed design to amortize the overheads associated with UMR
- Demonstrate the usefulness of the proposed schemes by comparing the performance of the proposed designs on real application layouts in GPU-based HPC clusters

Overview of the MVAPICH2 Project

- **High Performance open-source MPI Library**
- **Support for multiple interconnects**
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), and AWS EFA, Rockport Networks, and Slingshot
- **Support for multiple platforms**
  - x86, OpenPOWER, ARM, Xeon-Phi, GPGPUs (NVIDIA and AMD)
- **Started in 2001, first open-source version demonstrated at SC '02**
- Supports the latest MPI-3.1 standard
- http://mvapich.cse.ohio-state.edu
- **Additional optimized versions for different systems/environments:**
  - MVAPICH2-X (Advanced MPI + PGAS), since 2011
  - MVAPICH2-GDR with support for NVIDIA (since 2014) and AMD (since 2020) GPUs
  - MVAPICH2-MIC with support for Intel Xeon-Phi, since 2014
  - MVAPICH2-Virt with virtualization support, since 2015
  - MVAPICH2-EA with support for Energy-Awareness, since 2015
  - MVAPICH2-Azure for Azure HPC IB instances, since 2019
  - MVAPICH2-X-AWS for AWS HPC+EFA instances, since 2019
- **Tools:**
  - OSU MPI Micro-Benchmarks (OMB), since 2003
  - OSU InfiniBand Network Analysis and Monitoring (INAM), since 2015

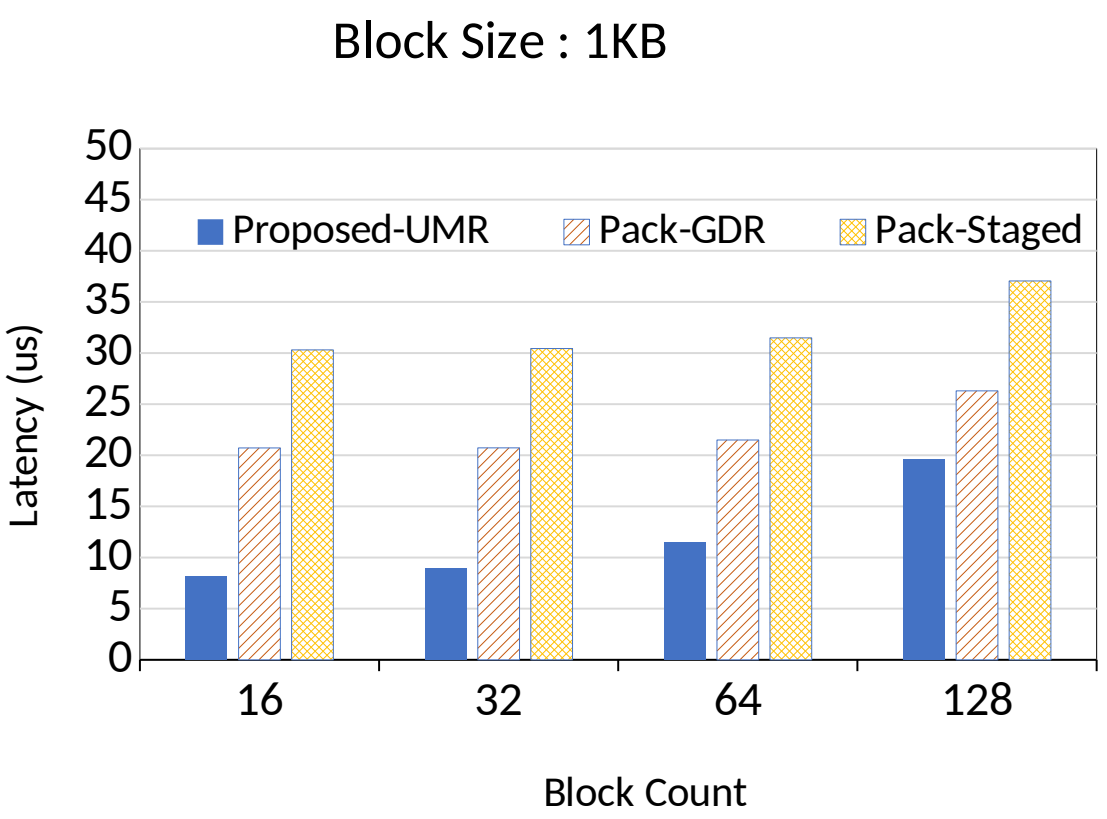
- **Used by more than 3,279 organizations in 90 countries**
- **More than 1.61 Million downloads from the OSU site directly**
- Empowering many TOP500 clusters (Jun '22 ranking)
  - 6th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 16th, 448, 448 cores (Frontera) at TACC
  - 26th, 391,680 cores (ABCI) in Japan
  - 42th, 570,020 cores (Nurion) in South Korea and many others
- Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, OpenHPC, and Spack)
- Partner in the 16th ranked TACC Frontera system
- **Empowering Top500 systems for more than 16 years**

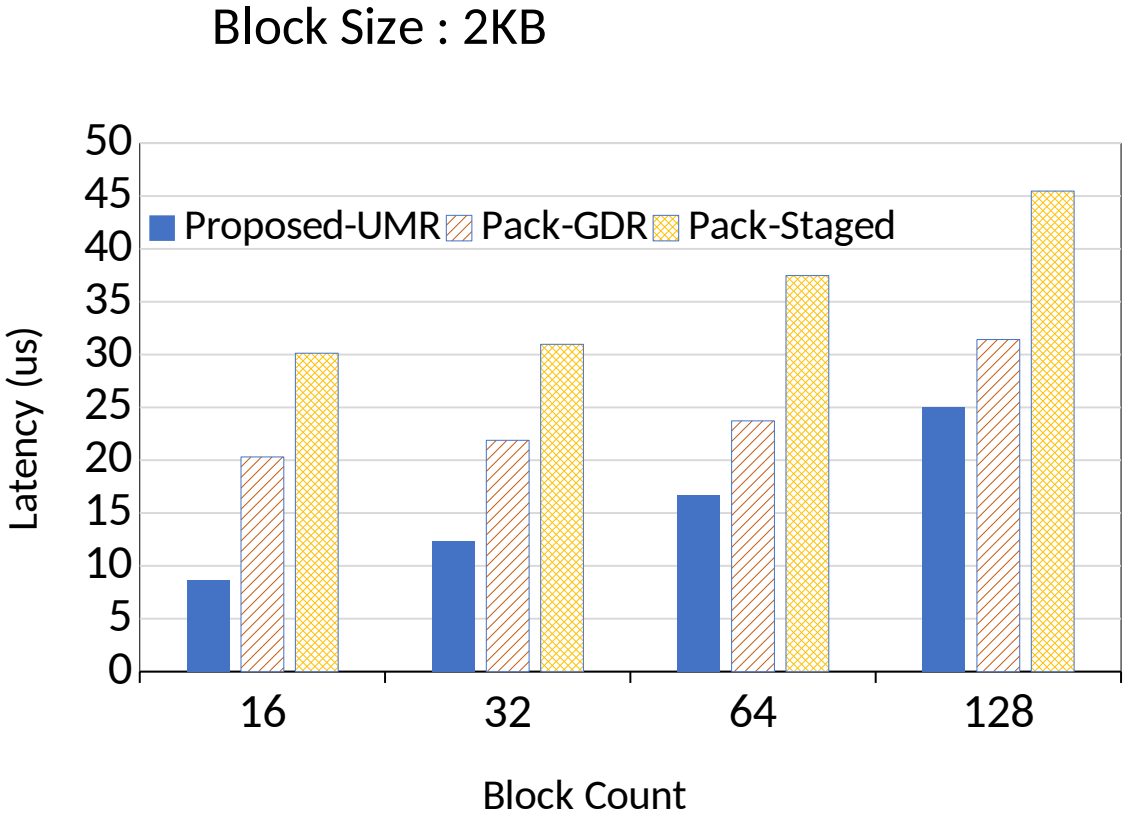
Experimental Setup

	MRI (OSU)	ThetaGPU
CPU Model	AMD EPYC 7713	AMD EPYC 7742
System memory	256 GB	1 TB
GPUs	4 NVIDIA A100	8 NVIDIA A100
Interconnects	PCIe Gen3 GPU<->HOST<->HCA	NVLink/NVSwitch GPU<->PCIe Switch<->HCA
NVIDIA driver version	410.79	410.48

- MPI libraries :
  - MVAPICH2-GDR-2.3.6, OpenMPI-4.1.3 + UCX-1.12.1
- Benchmarks and Applications Kernels:
  - OMB with Vector DDT
  - Modified DDT Bench with GPU support

Vector Benchmark Results



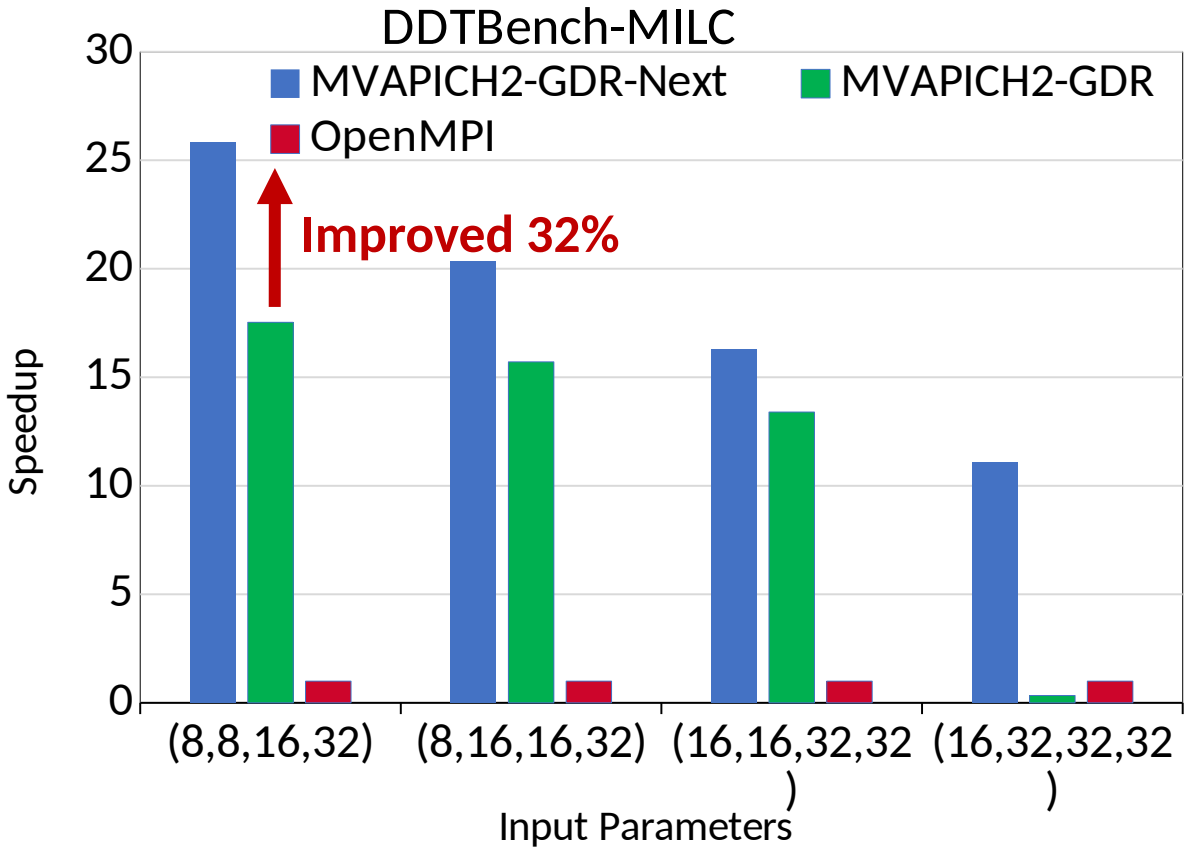


- Modified OMB osu_latency with block lengths of 1KB, 2KB
- Proposed UMR performs up-to 2X better than pack-GDR and 3X better than pack-staged

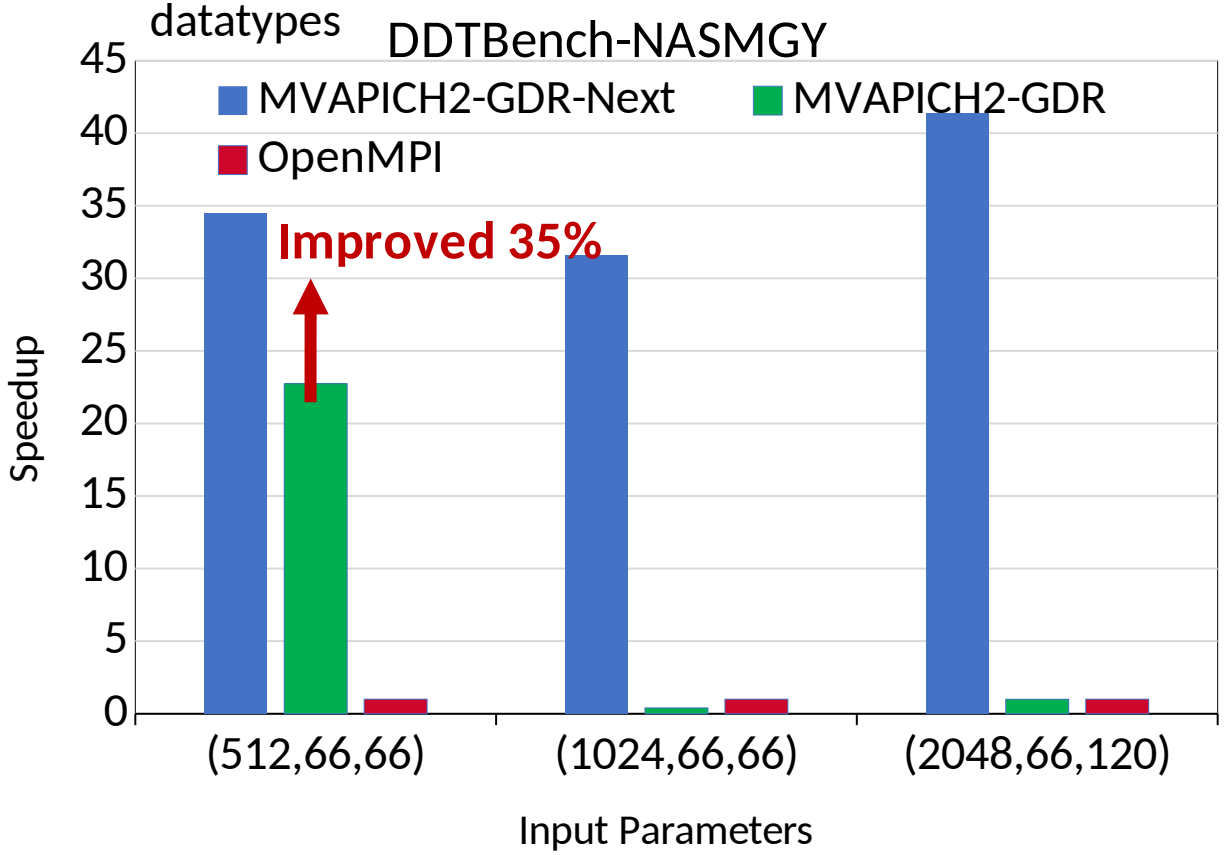
DDTBench Results

- **1 GPU per Node, 2 Node experiment. Speed-up relative to OpenMPI**

- Uses nested vector datatype for 4D face exchanges.



- 3D face exchanges with vector and nested vector datatypes

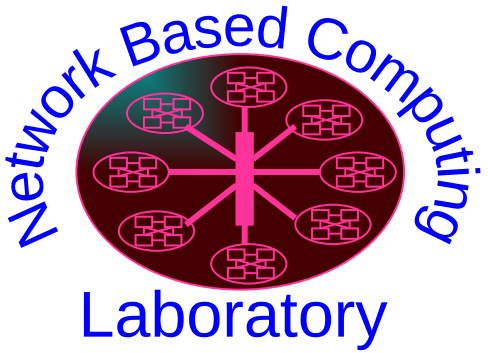


Conclusion & Future Work

- Conclusion
  - Past work on non-contiguous data movement focused on optimizing packing and unpacking kernels
  - Proposed a basic UMR based scheme to handle non-contiguous data on GPUs being communicated across the network.
  - Enhanced the proposed scheme
  - Proposed design achieves up to 2X improvement in performance over state-of-the-art schemes at the micro-benchmark level and up-to 35% improvement on application layouts.
- Future work
  - Provide support for AMD GPUs
  - Evaluate these designs on large scale HPC applications

Thank You!

kandadisuresh.1@osu.edu



Follow us on

https://twitter.com/mvapich

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

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

HiBD
High-Performance
Big Data

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/

HiDL
High-Performance
Deep Learning

The High-Performance Deep Learning Project
http://hidl.cse.ohio-state.edu/