# Unified Designs of Multi-rail-aware MPI Allreduce and Alltoall Operations Across Diverse GPU and Interconnect Systems

Chen-Chun Chen, Jinghan Yao, Lang Xu, Hari Subramoni and Dhabaleswar K. (DK) Panda
The Ohio State University
{chen.10252, yao.877, xu.3304}@osu.edu, {subramon, panda}@cse.ohio-state.edu

## Research Motivation

- HPC systems are evolving rapidly with: diverse **GPUs**, **interconnects, multi-NIC** per node for high throughput
- Collective communication performance is critical. E.g.: Amber and heFFTe rely on **Allreduce** and **Alltoall**.
- Existing GPU-aware MPI collectives are limited:
  - Mostly **CPU-centric** designs for inter-node (no intra-node, such as **GPU kernel**, optimization)
  - Underutilize **GPU Direct RDMA** and GPU compute capabilities

## Research Challenges

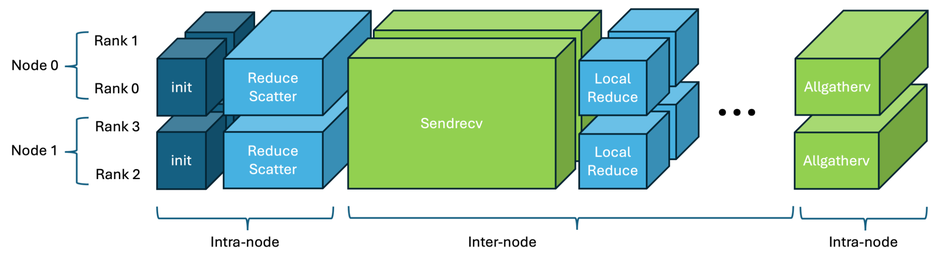
- What strategies are needed to design a high-performance GPU-aware MPI Allreduce and Alltoall inter-node operation?
- What approaches can further optimize throughput?
- Can we develop unified designs compatible with modern heterogeneous CPU, GPU, and interconnect systems?

## Overview of the Designs

- **Persistent GPU** buffer: avoid CPU staging
- **Two-level hierarchy** (intra + inter node): leader-based communication to isolate fast local ops from slower inter-node transfers
- **Multi-leader** strategy: **overlap** intra-node (ReduceScatter) with inter-node (Allreduce), and then intra-node (Allgather)
- Unified for Allreduce and Alltoall
  - Diverse GPU vendors: **NVIDIA**, **AMD**, Intel
  - Diverse networks: **Infiniband**, **Slingshot**, **Omni-Path**
  - Multiple algorithms: RD, RSA (Allreduce) and Pairwise, Scattered (Alltoall)

## Design of the Multi-leader Two-level Allreduce Algorithm

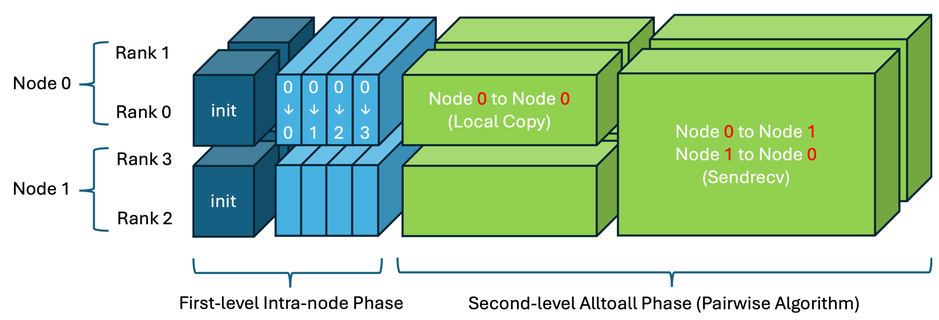
- **Two-Level Allreduce Structure:**
  - **Intra-node** phase: use GPU kernels for ReduceScatter and Allgather, employ IPC to access peer GPU buffers
  - **Inter-node** phase: ranks with same local rank form leader groups, each group runs a second-level Allreduce
- **Persistent** GPU buffer
- Early-triggered and **pipelined** optimization
  - Divide data into num_chunks



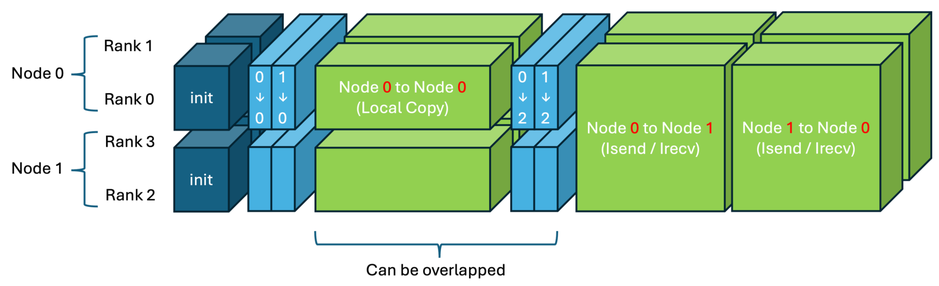

**Dataflow of the proposed multi-Leader two-Level Allreduce algorithm**

## Multi-leader Two-level Alltoall Algorithm

- **Push Algorithm:**
  - Steps:
    - **Each rank pushes data** directly to the GPU buffers of peers (via IPC)
    - After all intra-node pushes finish**, leader ranks** perform inter-node Alltoall
  - Limitation: No overlap possible due to **sync barrier** between phases
- **Pull Algorithm:**
  - Steps:
    - Inter-node **Irecv** launched early for each peer
    - Intra-node Pull: use IPC to **copy required segments** into a **persistent GPU buffer**
    - Once ready, launch **Isend**
  - Enables fine-grained **overlap** of intra- and inter-node stages

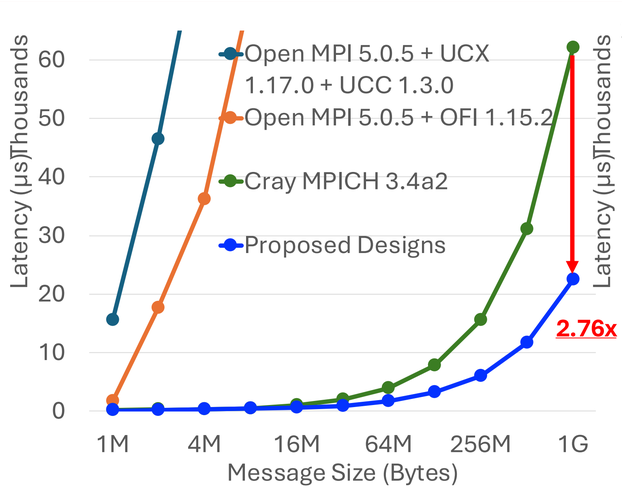



**Alltoall Push Algorithm**

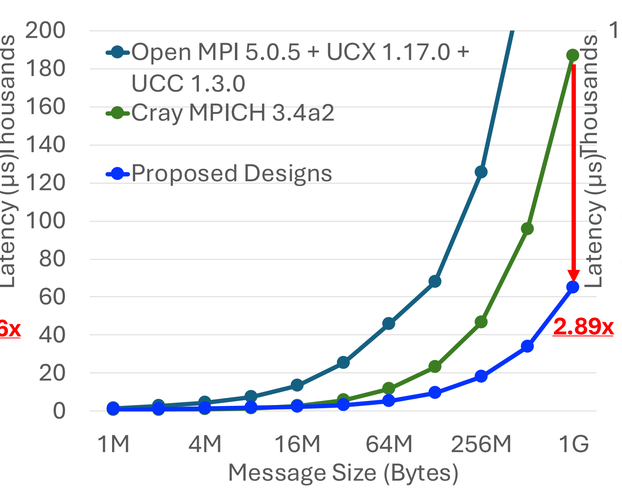



**Alltoall Pull Algorithm**

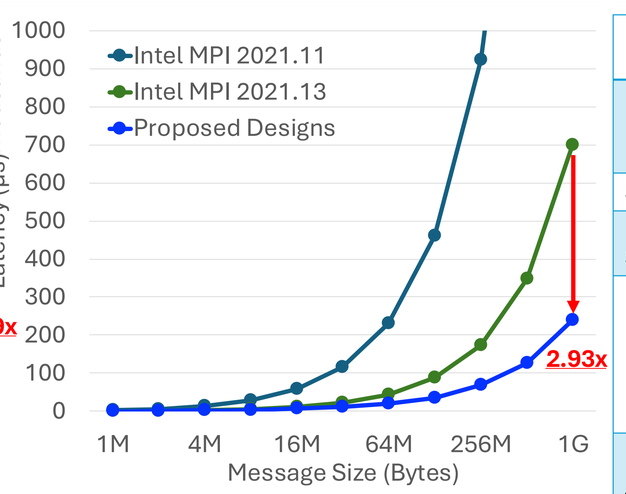
## Benchmark-level Performance Evaluations - Allreduce




**Isambard-AI – 8 nodes (32 GPUs)**




**Frontier – 8 nodes (64 GPUs)**

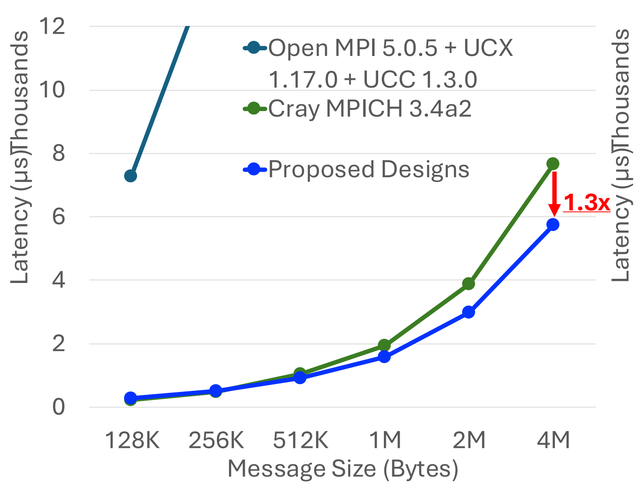



**Stampede3 – 8 nodes (64 GPUs)**

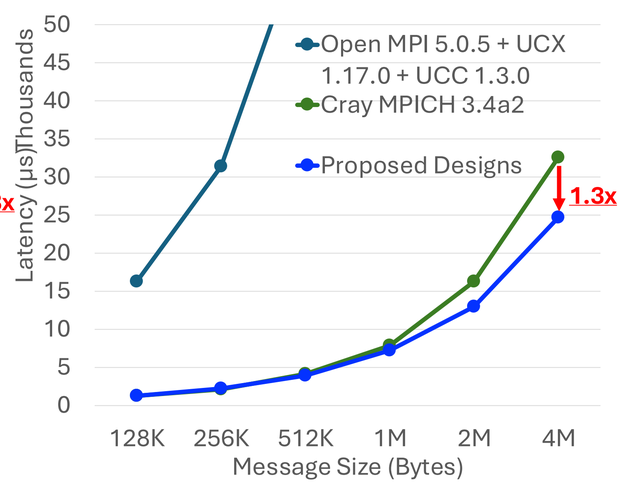
## Experimental Setup

| | Isambard-AI | Frontier | Cardinal | Stampede3 |
|---|---|---|---|---|
| CPU | NVIDIA Grace CPU | AMD EPYC 7A53 | Intel Xeon Platinum 8470 | Intel Xeon Platinum 8468 |
| Sockets | 4 | 1 | 2 | 2 |
| Core/ Sockets | 72 | 64 | 52 | 48 |
| GPU | 4 NVIDIA GH200 GPUs | 4 AMD MI250X (8 GCDs) | 4 NVIDIA H100 GPUs | 4 Intel Data Center GPU Max 1550s (8 Stacks) |
| Inter-connection | 4 HPE Slingshot 200 Gbps NICs | 4 HPE Slingshot 200 Gbps NICs | 4 Infiniband NDR400 HCAs | Cornelis CN5000 Omni-Path 100 Series |

## Benchmark-level Performance Evaluations - Alltoall

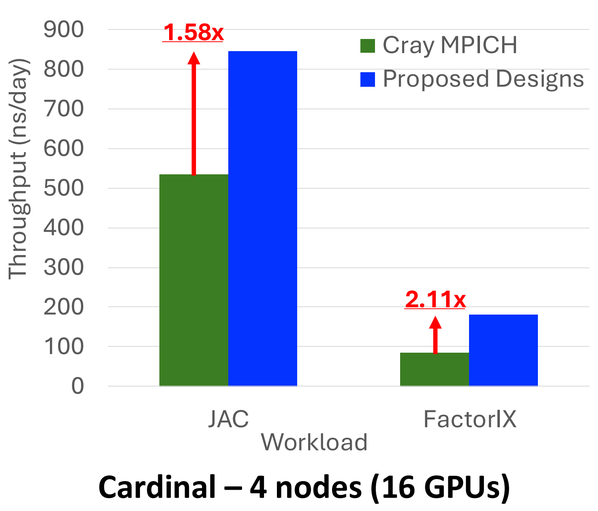



**Isambard-AI – 8 nodes (32 GPUs)**

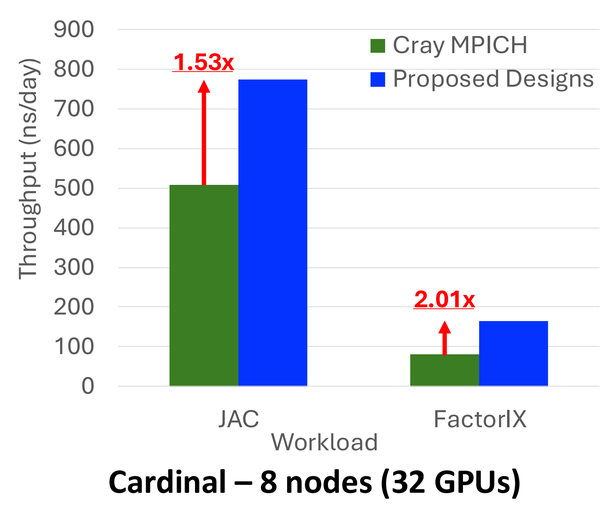



**Isambard-AI – 8 nodes (32 GPUs)**

## Application-level Performance Evaluations – Amber (Allreduce)

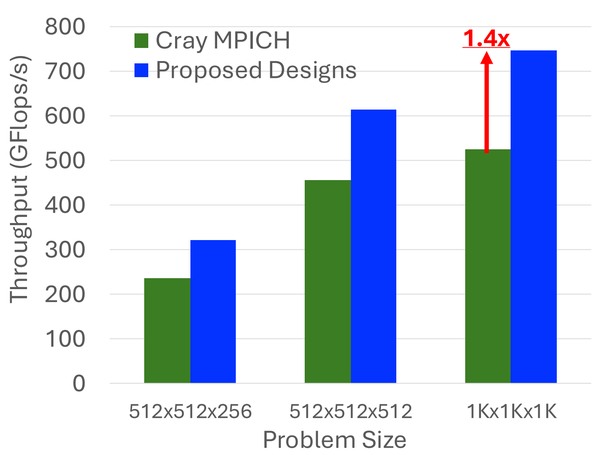



**Cardinal – 4 nodes (16 GPUs)**

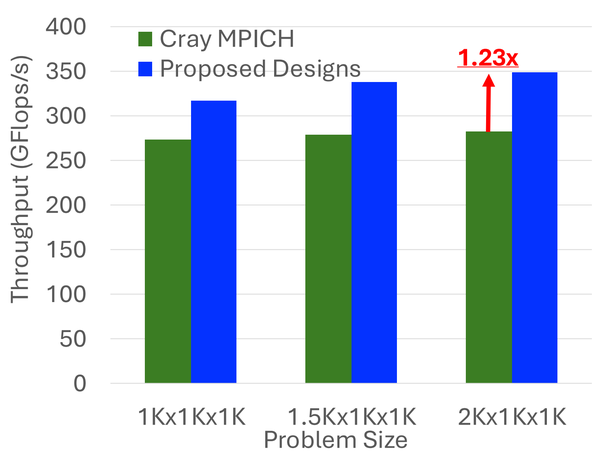



**Cardinal – 8 nodes (32 GPUs)**

## Application-level Performance Evaluations – heFFTe (Alltoall)




**Isambard-AI – 8 nodes (32 GPUs)**




**Frontier – 8 nodes (64 GPUs)**

## Conclusion

- We proposed unified, GPU-aware, and multi-rail-optimized MPI collectives.
  - Multi-leader, two-level **Allreduce** with **persistent GPU buffer**, optimize with **early-triggered pipelined overlap** to hide inter-node latency.
  - **Push** and **Pull** variants for GPU-resident **Alltoall** using IPC.
- Available in MVAPICH-Plus 4.0 (https://mvapich.cse.ohio-state.edu)
- **Benchmark results**:
  - Up to **2.7× - 2.9×** Allreduce speedup
  - Up to **1.3×** Alltoall speedup
- **Application-level impact**:
  - Up to **2×** performance gain in Amber
  - Up to **1.4×** performance gain in heFFTe

THE OHIO STATE UNIVERSITY

MVAPICH

References:
1. C. Chen, J. Yao, L. Xu, H. Subramoni, D. Panda: "Unified Designs of Multi-rail-aware MPI Allreduce and Alltoall Operations Across Diverse GPU and Interconnect Systems", 39th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2025)

Acknowledgements: This research is supported in part by NSF grants #1818253, #1854828, #2311830, #2312927, #2323116, #2415201, and XRAC grant #NCR-130002.