Collaborative Bandwidth-Efficient Intra-Node Allreduce

AmirHossein Sojoodi, Ali Farazdaghi,
Hamed Sharifian, Ryan E. Grant, and Ahmad Afsahi

Parallel Processing Research Laboratory (PPRL),
Department of Electrical and Computer Engineering

Smith Engineering
Queen’s University, Kingston, Canada

AsHES 2025,
Milan, Italy, June 3rd , 2025

1 Introduction

2 Background

3 Design and Evaluation

4 Conclusion

1. HPC and Cluster Computing
2. Multi-GPU Clusters
3. Multi-Path Communication
4. Heterogeneous Computing
5. Let's Beat NCCL!

1. Allreduce Operation in HPC
2. GPU-Aware Communication Libraries
3. Heterogeneous Computing

Target Topologies

- Systems with NVLink and PCIe interconnects
  - Any number of NVLinks for Peer-to-Peer communication
  - PCIe paths between GPUs and the Host

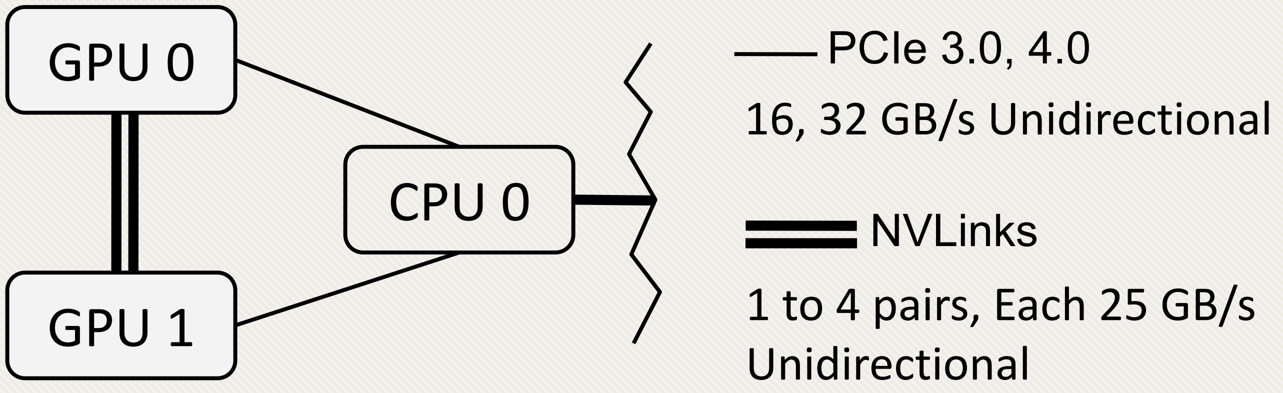


Figure 1: Target topologies for multi-path Allreduce framework

1. Design Objectives
2. Framework Architecture
3. Data Flow Profile

Design Objectives

1. Multi-Path Communication
   – Utilizing all available band width to decrease data transfer overhead between the GPUs
2. Leveraging Idle CPUs in the Data Path
   – Concurrently utilizing the computational capabilities of both CPUs and GPUs
3. Low Overhead
   – A low-overhead pipelined communication scheme to overlap computation and communication
4. Data Integrity
   – Ensuring data integrity and consistency during data transfers and computations.
5. Reducing CPU Involvement in the Control Path
   – Utilizing asynchronous communication primitives to offload the tasks asynchronously

Framework

- Multi-path Heterogeneous Pairwise Exchange Allreduce on two GPUs using NVLink and PCIe.

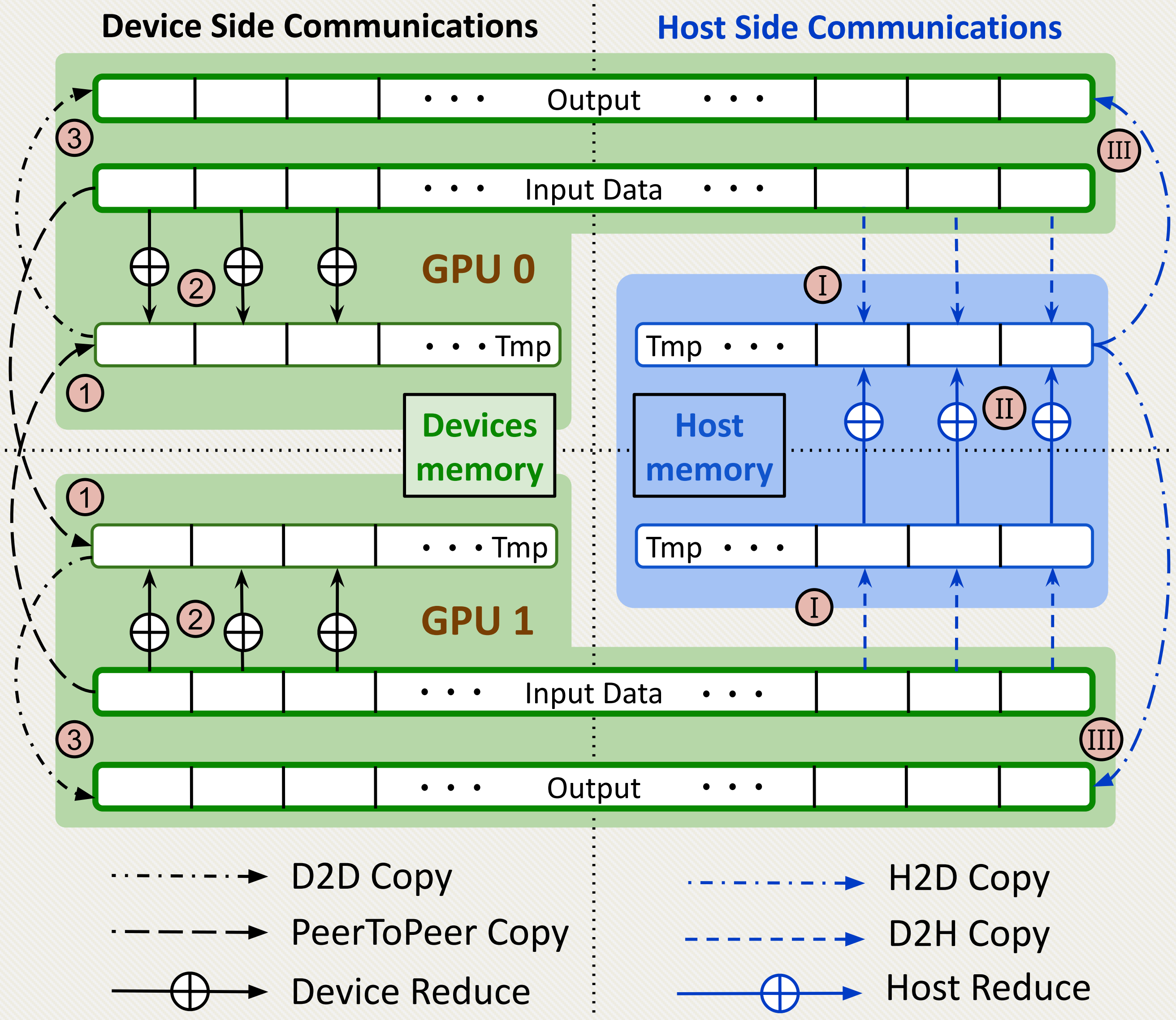


Figure 2: Multi-Path Heterogeneous Allreduce on Two GPUs

Data Flow Timeline

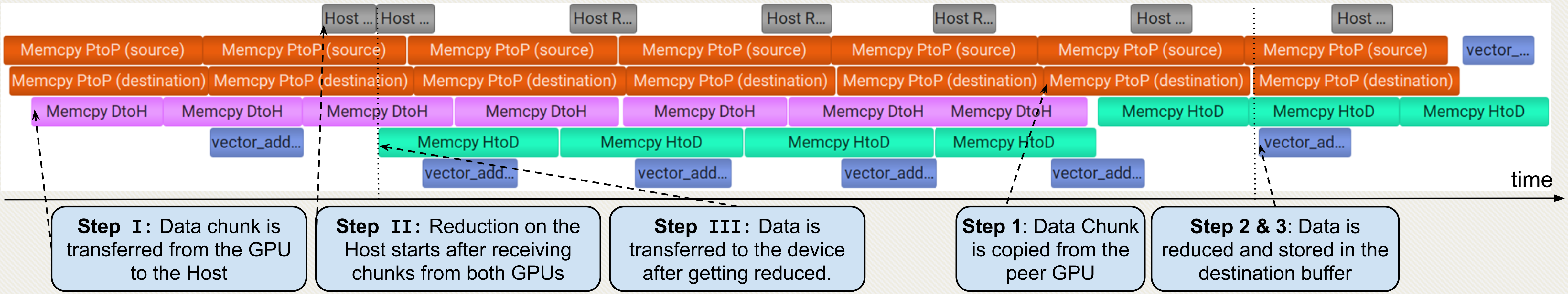


Figure 3: A summarized NVIDIA Nsight Systems timeline profile of the multi-path heterogeneous Allreduce (128 MB) on one of the GPUs on Daisy cluster.

A Experimental Setup

B Allreduce Algorithms Evaluations on Mist Cluster

C Allreduce Algorithms Evaluations on Daisy Cluster

D Allreduce Algorithms Evaluations on Narval Cluster

Experimental Setup

- Experimentally developed in user-level
- On topologies with at least 2 GPUs connected via NVLink and PCIe
- Compared with:
  - NCCL (v2.24.3-1)
  - MPI (v5.0.2) configured with:
    - UCC (v1.3)
    - UCX (v1.17)
    - CUDA (v12.6)

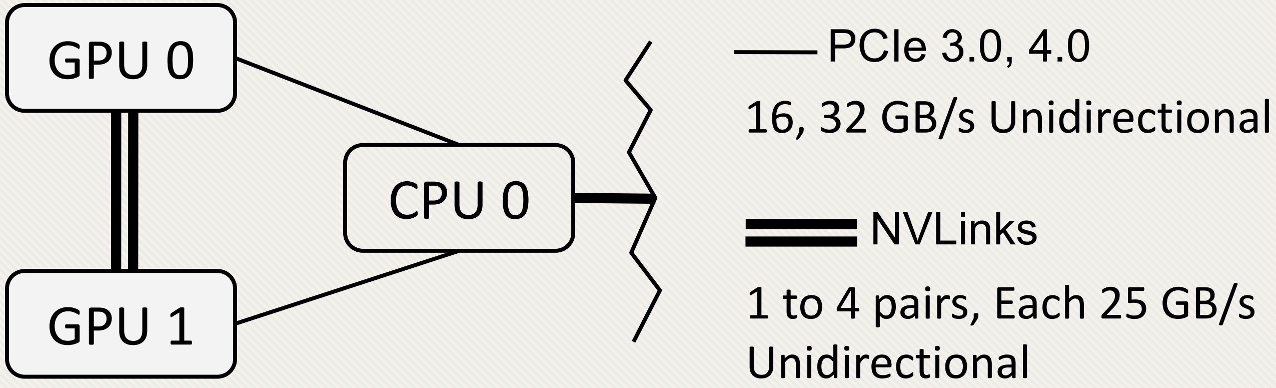


Figure 1: Target topologies for multi-path Allreduce framework

Allreduce Evaluations - Mist Cluster

- Mist of SciNet - Canada
  - 4 x NVIDIA V100 per node (2 socket)
  - Each GPU pair on each socket have three NVLinks
  - Power9 8335-GTH with 16 cores (64 HWT)
  - One CPU and two of the GPUs are configured as a NUMA node

A) MPI_UCC_UCP
B) MPI_UCC_NCCL
C) Pairwise Exchange
D) Kernel-Based Pairwise Exchange
E) Pipelined Pairwise Exchange
F) Segmented Ring
G) Kernel-Based Segmented Ring
H) Pipelined Segmented Ring
I) Multi-path Heterogenous Pairwise Exchange

Figure 4: Speedup Figures Legend

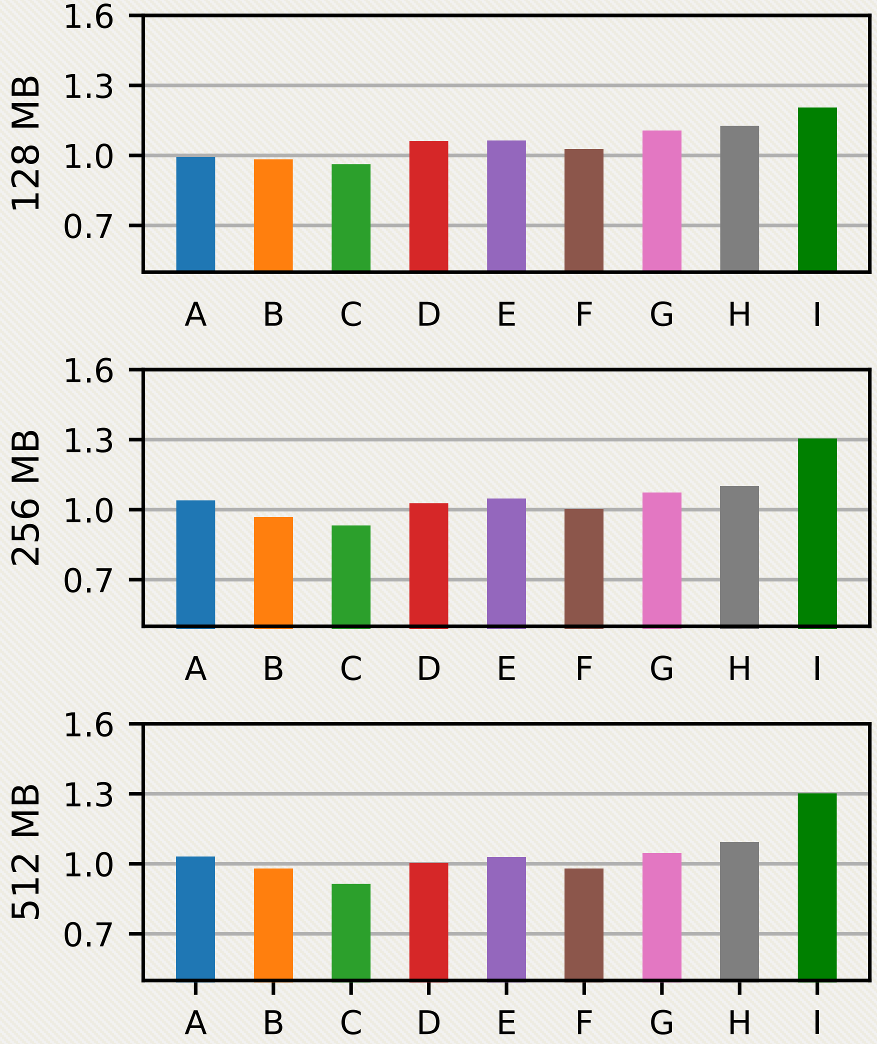


Figure 5: Allreduce algorithms comparison against standalone NCCL on Mist

Allreduce Evaluations - Daisy Cluster

- Daisy – Local Cluster at Queen’s University
  - 4 x NVIDIA A100 per node (2 socket)
  - Each GPU pair on each socket have four NVLinks
  - Intel Xeon Gold 6338, with 32 cores (64 HWT)
  - One CPU and two of the GPUs are configured as a NUMA node

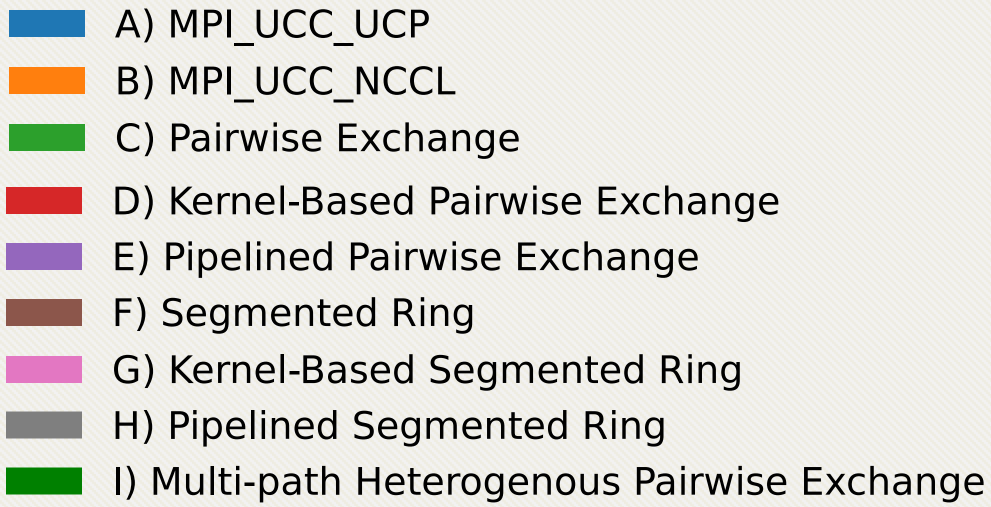


Figure 4: Speedup Figures Legend

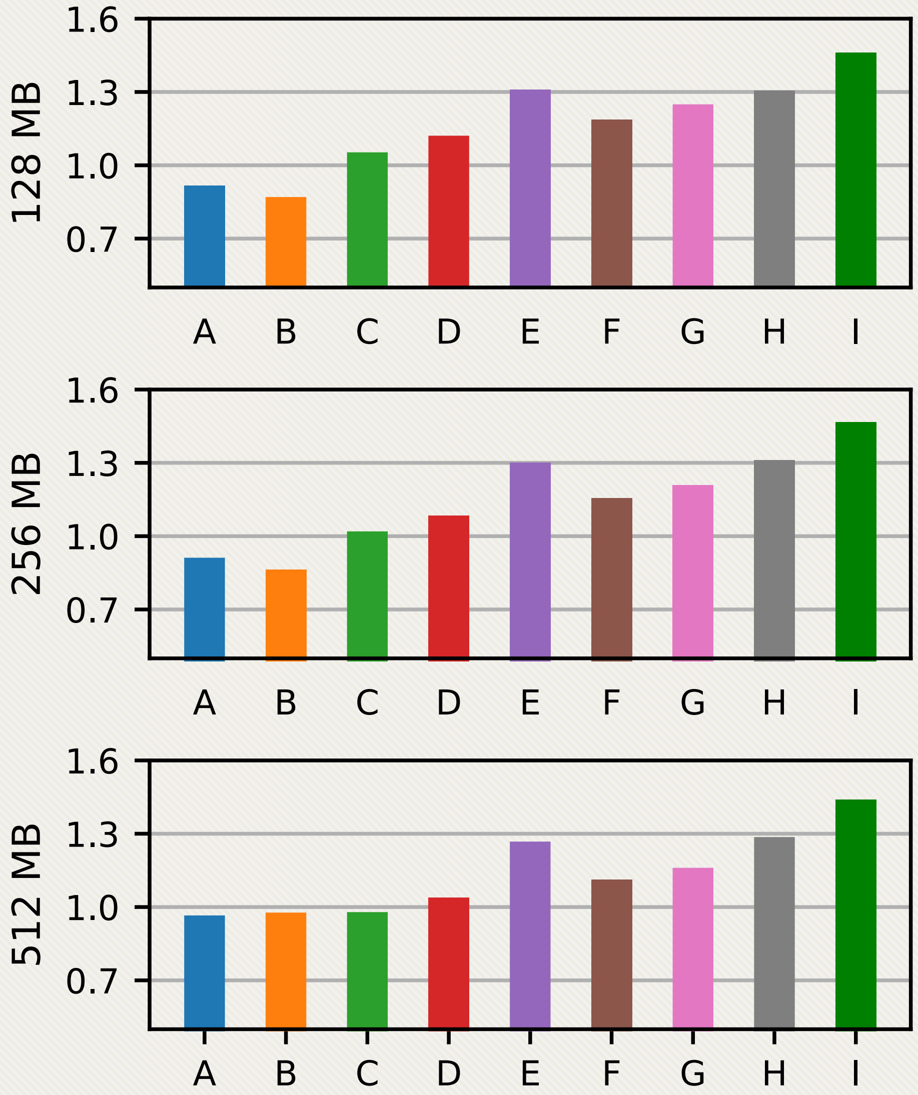


Figure 6: Allreduce algorithms comparison against standalone NCCL on Daisy

Allreduce Evaluations on Narval Cluster

- Narval of Digital Research Alliance of Canada
  - An eight-NUMA node system with 4 x NVIDIA A100
  - GPUs are connected by four NVLinks (full mesh)
  - AMD EPYC 7413 with total of 24 cores (no HTW)
  - Each GPU with a specific memory channel is connected and configured as a NUMA node

A) MPI_UCC_UCP
B) MPI_UCC_NCCL
C) Pairwise Exchange
D) Kernel-Based Pairwise Exchange
E) Pipelined Pairwise Exchange
F) Segmented Ring
G) Kernel-Based Segmented Ring
H) Pipelined Segmented Ring
I) Multi-path Heterogenous Pairwise Exchange

Figure 4: Speedup Figures Legend

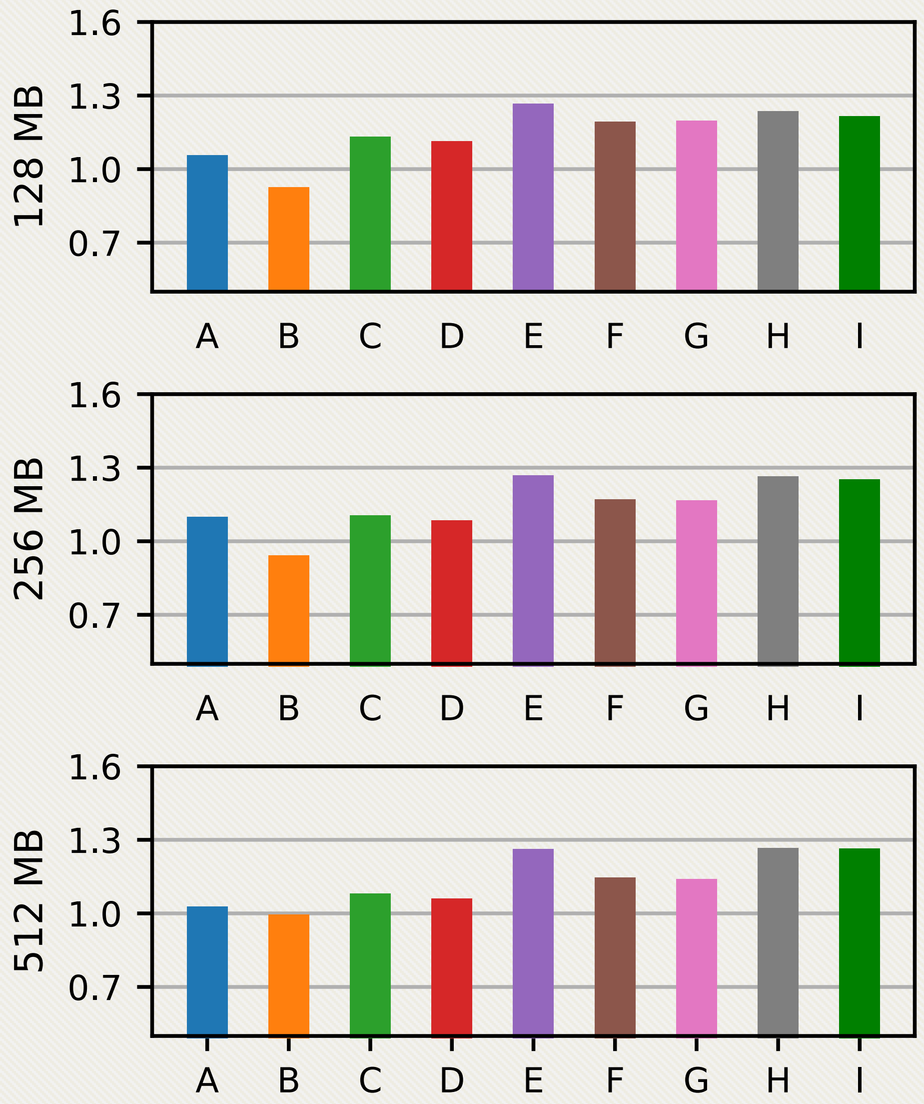


Figure 7: Allreduce algorithms comparison against standalone NCCL on Narval

1. Conclusions
2. Future Directions

Conclusions

- A preliminary feasibility study on collaborative Allreduce
- Considerable acceleration opportunity for:
  - Large message sizes
  - High number of available CPU cores
  - No/limited simultaneous traffic on PCIe channels
- Future Directions
  - Integration to the communication libraries
  - Dynamic configuration (with performance modeling)
    - Topology awareness
    - Automatically adapt to ongoing communication
  - Explore tightly coupled topologies, like superchips

Acknowledgments

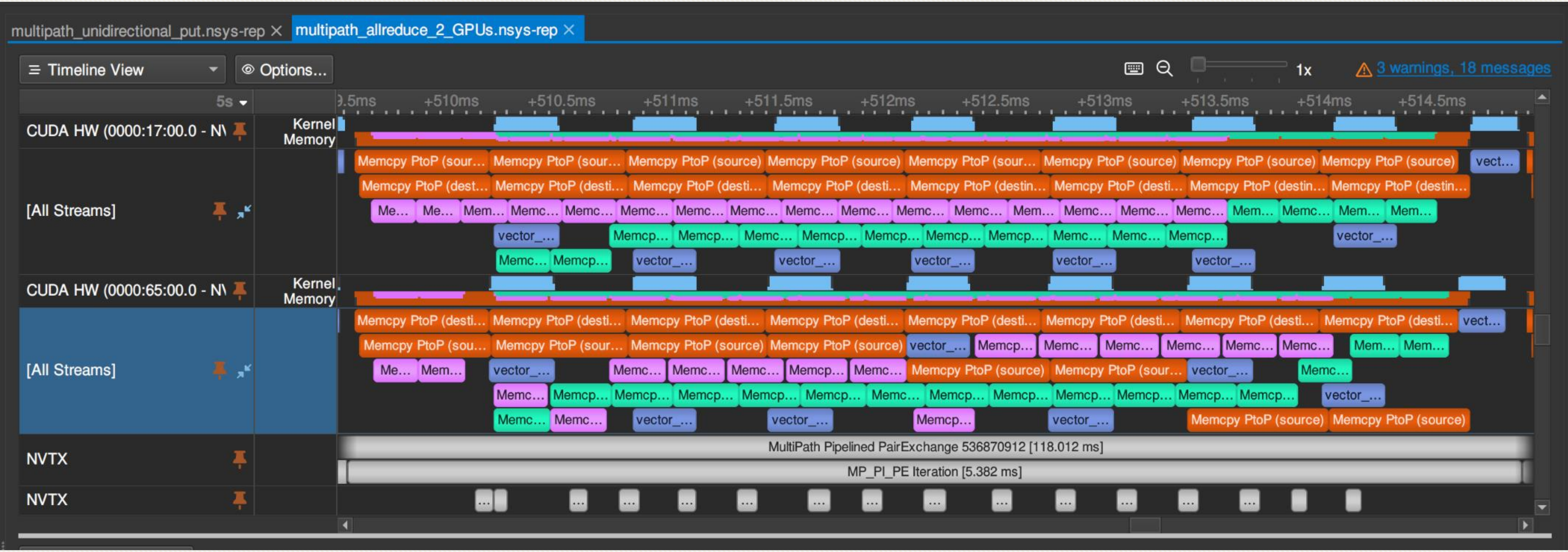
- Natural Sciences and Engineering Research Council of Canada
- Digital Research Alliance of Canada

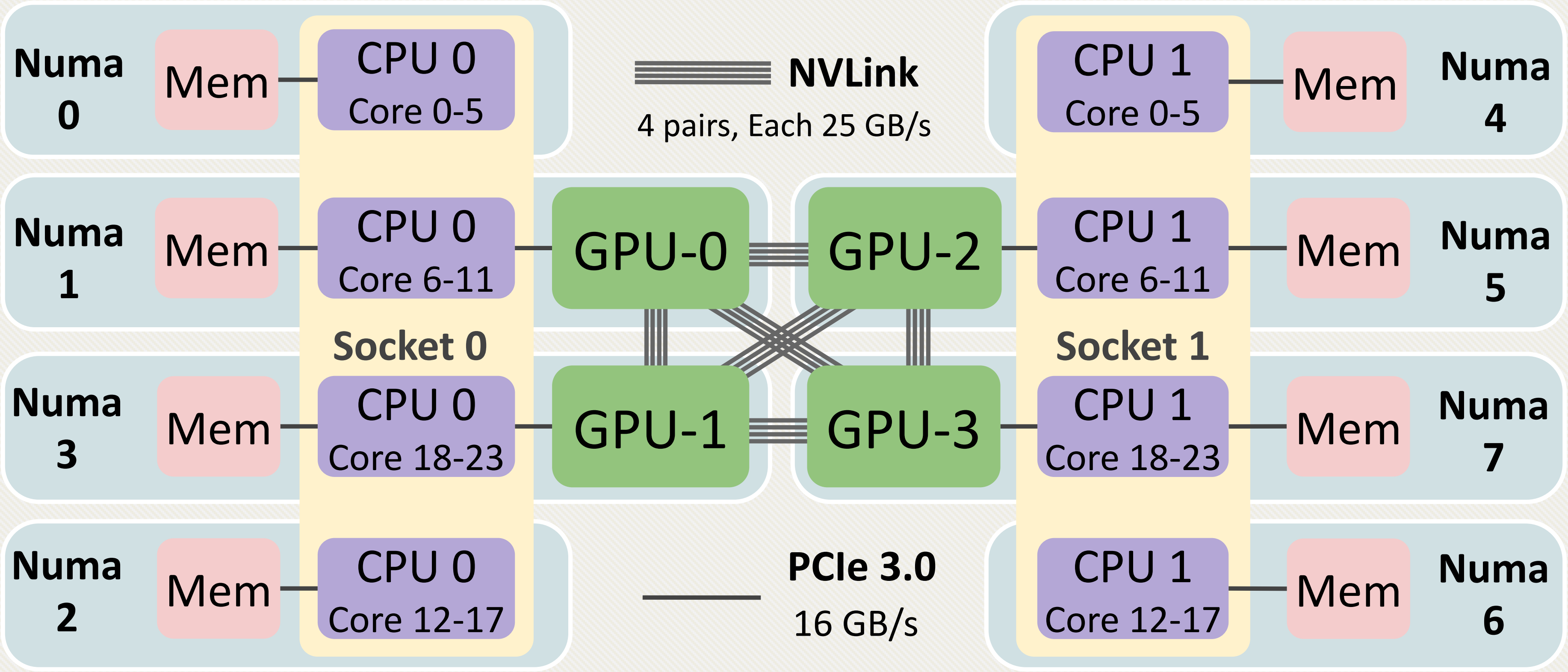
Thank You!

Instead of cursing the darkness, better light a candle!

Questions, Comments, and Ideas are Welcome!

amirsojoodi.github.io


multipath_unidirectional_put.nsys-rep
multipath_allreduce_2_GPUs.nsys-rep
Timeline View
Options...
1x
3 warnings, 18 messages
+510ms
+510.5ms
+511ms
+511.5ms
+512ms
+512.5ms
+513ms
+513.5ms
+514ms
+514.5ms
CUDA HW (0000:17:00.0 - NV
Kernel
Memory
[All Streams]
Memcpy PtoP (source)
Memcpy PtoP (destin...
CUDA HW (0000:65:00.0 - NV
NVTX
MultiPath Pipelined PairExchange 536870912 [118.012 ms]
MP_PI_PE Iteration [5.382 ms]


Numa 0
Mem
CPU 0
Core 0-5
NVLink
4 pairs, Each 25 GB/s
CPU 1
Core 0-5
Mem
Numa 4
Numa 1
Mem
CPU 0
Core 6-11
GPU-0
GPU-2
CPU 1
Core 6-11
Mem
Numa 5
Socket 0
Socket 1
Numa 3
Mem
CPU 0
Core 18-23
GPU-1
GPU-3
CPU 1
Core 18-23
Mem
Numa 7
Numa 2
Mem
CPU 0
Core 12-17
PCIe 3.0
16 GB/s
CPU 1
Core 12-17
Mem
Numa 6