A Peak Performance Model for All-to-all on Hierarchical Systems and Its Applications

Rohini Uma-Vaideswaran[1], Joshua Romero[2], Daniel Dotson[1], David Appelhans[2], P. K. Yeung[1]

[1]Georgia Institute of Technology, [2]NVIDIA Corporation

Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS) Workshop, SC25

Supported by NSF via subaward TACC at UT Austin as a CSA Project

Introduction

- All-to-all (A2A) communication: crucial component of many scientific computing applications
  - ✓ Fluid dynamics: Direct Numerical Simulations (3D FFTs with distributed transposes)
  - ✓ Machine learning
  - ✓ ...

- Increasingly massive parallelism $\Longrightarrow$ higher fidelity, but A2A renders the application network bandwidth bound.

Scalability challenge $\Longleftrightarrow$ A2A performance

"Peak" performance model

A theoretical upper bound for all-to-all performance on a given system

Basis: **hierarchical structure** of emergent heterogeneous HPC platforms.

Key feature: model definition and parameters based on easily accessible **system/network specifications**.

Hierarchical Communication Pathways

"self"

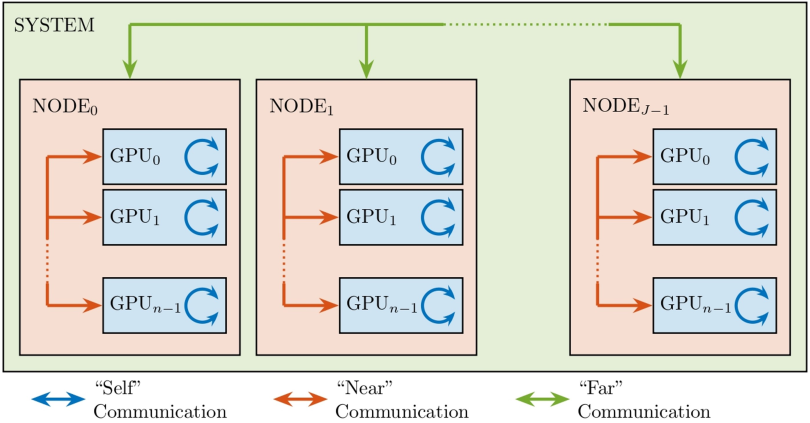
High-bandwidth memory

"near"

NVLink, Infinity Fabric, etc.

"far"

Infiniband, Slingshot, etc.



All-to-All Performance Model

- P processes globally, tightly connected groups of n processes, A2A message buffer of size m, and p2p messages of size m/P.

	"self"	"near"	"far"
# p2p Messages	1	$(n-1)$	$(P-n)$
Bandwidth	B_{self}	B_{near}	B_{far}
A2A Time	$t_{\text{self}} = \frac{m}{P}\frac{1}{B_{\text{self}}}$	$t_{\text{near}} = (n-1)\frac{m}{P}\frac{1}{B_{\text{near}}}$	$t_{\text{far}} = (P-n)\frac{m}{P}\frac{1}{B_{\text{far}}}$

- Assumption: concurrent p2p communication along all 3 pathways.

Peak All-to-All Performance Model

Time: $$t_{\text{a2a}} = \max\left\{ (P-n)\frac{m}{P}\frac{1}{B_{\text{far}}},\ (n-1)\frac{m}{P}\frac{1}{B_{\text{near}}},\ \frac{m}{P}\frac{1}{B_{\text{self}}} \right\}$$

Peak Bandwidth: $B_{\text{a2a}} = m/t_{\text{a2a}}$

Considerations while Applying the Model

Peak All-to-All Performance Model

$$\text{Time: } t_{\text{a2a}} = \max\left\{ (P-n)\frac{m}{P}\frac{1}{B_{\text{far}}},\ (n-1)\frac{m}{P}\frac{1}{B_{\text{near}}},\ \frac{m}{P}\frac{1}{B_{\text{self}}} \right\}$$

$$\text{Peak Bandwidth: } B_{\text{a2a}} = m/t_{\text{a2a}}$$

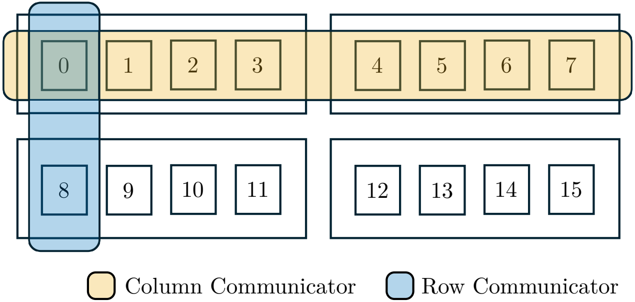
1. **"near" group** definition **dependent on application and node topology**.
2. Model parameters: **number of p2p messages and bandwidths** of each communication pathway based on **system specifications**.

Application: 2D Distributed Transposes

- Data distributed over P processes, decomposed into P_r rows and P_c columns.
- A2A over row and column communicators of different sizes and configurations.
- **Key caveat**: communicators may be split between different nodes, affects definition of "near" term.

E.g., $P = P_r \times P_c = 2 \times 8$, on a system with 4 GPUs per node, A2A time:

$$t_{\text{a2a}}^{\text{2D}} = t_{\text{col}} + t_{\text{row}} = t_{\text{a2a}}(P = 8,\ n = 4) + t_{\text{a2a}}(P = 2,\ n = 1)$$



Test Systems: *Vista*, *Alps* & NVL72

	Vista	***Alps***	**NVL72**
GPUs/node	1 NVIDIA GH200	4 NVIDIA GH200	4 NVIDIA GB200
"Far" BW (per GPU)	InfiniBand 50 GB/s	Slingshot 25 GB/s	InfiniBand 50 GB/s
"Near" BW (GPU-GPU)	–	NVLink 150 GB/s	NVLink and NVSwitch 900 GB/s
GPUs/"Near" group	–	4	72
"Self" BW per GPU	HBM3 4 TB/s		HBM3e 8 TB/s

"near" group size on *Vista* $n = 1 \implies t_{\text{near}} = 0$.

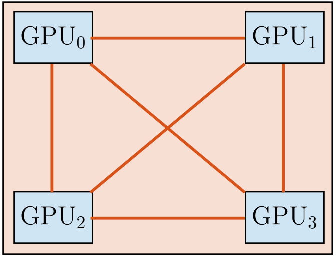
Applying the A2A Model on *Alps*

	GPUs/node	"Far" BW (per GPU)	"Near" BW (GPU-GPU)	GPUs/"Near" group	"Self" BW per GPU
Alps	4 NVIDIA GH200	Slingshot 25 GB/s	NVLink 150 GB/s	4	HBM3 4 TB/s

- On a node, each NVLink connected GPU-GPU pair can communicate over a 150 GB/s link.
- But the GPUs split their NVLink connections across peers: max. NVLink BW achieved only when communicating with all 3 peers.

$\implies B_{\text{near}}$ is a function of the number of "near" processes participating in A2A:

$$B_{\text{near}}(n) = (n-1) \times 150 \text{ GB/s}$$



Node topology on *Alps*

Key takeaway: Choice of B_{near} must account for the node topology.

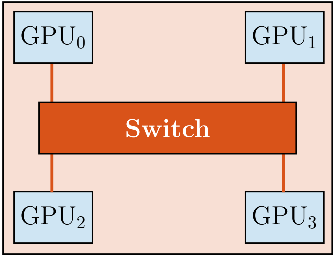
Applying the A2A Model on the NVL72 System

	GPUs/node	"Far" BW (per GPU)	"Near" BW (GPU-GPU)	GPUs/"Near" group	"Self" BW per GPU
NVL72	4 NVIDIA GB200	InfiniBand 50 GB/s	NVLink and NVSwitch 900 GB/s	72	HBM3e 8 TB/s

- Interconnected NVL72 domains, each consisting of 18 nodes with 4 GPUs each.
- Each domain $\equiv$ single, large, 72 GPU node.

$\Longrightarrow$ "near" pathway is *intra-domain*,
"far" pathway is *inter-domain*.

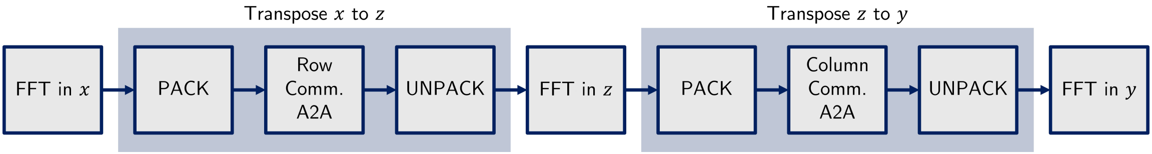
$$B_{\text{near}} = 900\ \text{GB/s}, \quad B_{\text{far}} = 50\ \text{GB/s}$$



Node topology on NVL72

Benchmark Code: GPU-enabled Direct Numerical Simulations

- Pseudospectral algorithm for direct numerical simulations of turbulent fluid flows.
- GPU algorithm: Yeung et al., *Computer Physics Communications, 2025* (Fortran, OpenMP offloading, GPU-aware MPI).
- A2A communication required for **distributed transposes** (1D or 2D) between 1D FFTs in three coordinate directions.
- 3D solution domain with N^3 grid points distributed among P processes, A2A message size $m = 4N^3/P$.
- A2A communication dominates time/step (≥ 80 %).



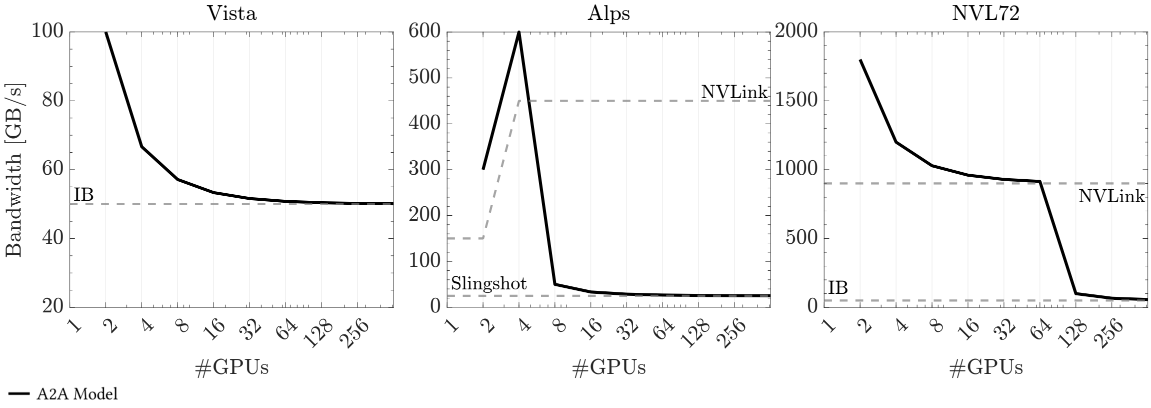
Testing Multiple A2A Backends with cuDecomp

- Adaptive 2D Decomposition library (Romero *et al.* PASC 2023).
- Multiple A2A backends for global transposition, to compare with model prediction.
- Key performance data collected: average runtime and bandwidth of A2A communication and local pack/unpack in distributed transpose.

Backend	Communication APIs
MPI_P2P MPI_P2P (pipelined)	`MPI_Isend, MPI_Irecv`
MPI_A2A	`MPI_Alltoall, MPI_Alltoallv`
NCCL NCCL (pipelined)	`ncclSend, ncclRecv`
NVSHMEM NVSHMEM (pipelined)	`nvshmemx_putmem_nbi_on_stream, nvshmemx_putmem_nbi`

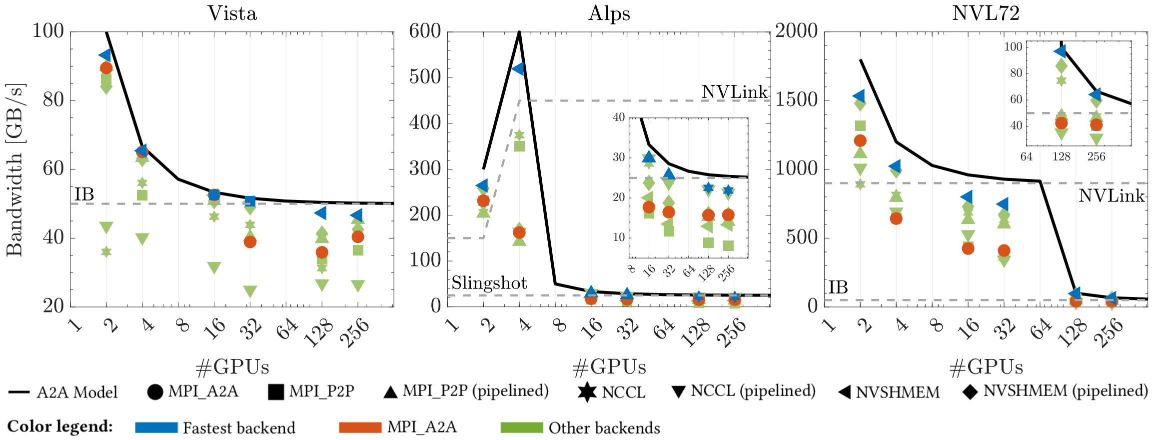
All-to-all Model Prediction

- Gradual transitions between “self”, “near” and “far” communication dominated regions.
- Performance boosts from “self” and “near” communication at all scales.
- Transition regions pushed to larger scales as “near” group size increases.



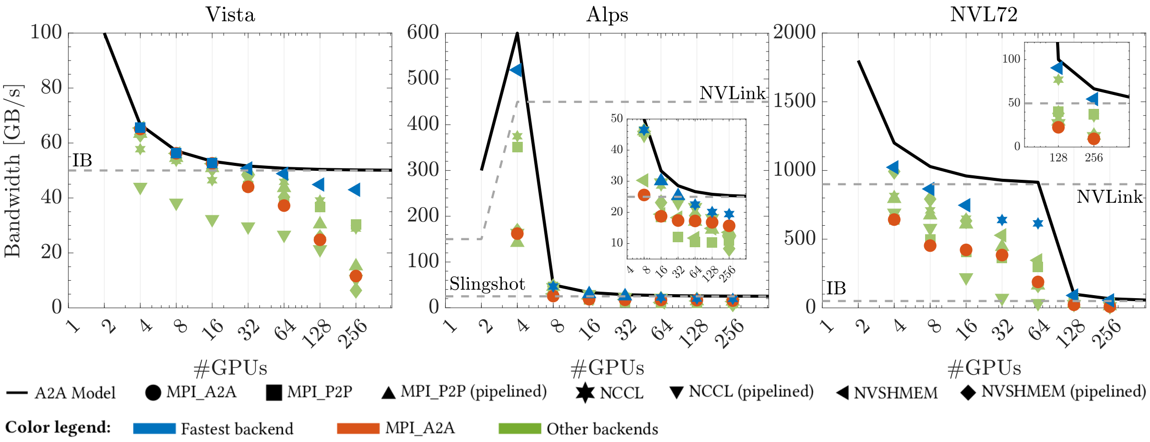
All-to-all Model Validation: Benchmarking Tests

- Fixed p2p message size for every 8x increase in #GPUs.
- At least one A2A backend achieves near peak model performance.
- Model acts as a reasonable upper bound for achievable A2A bandwidth.



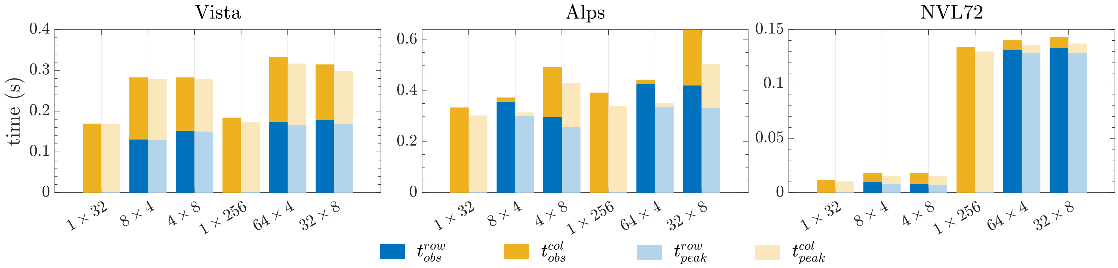
All-to-all Model Validation: Strong Scaling

- Decreasing p2p message size as #GPUs increases: 8 GiB to 125 MiB.
- Some impact from unmodeled latency terms at the smallest message sizes.



2D Distributed Transpose Analysis

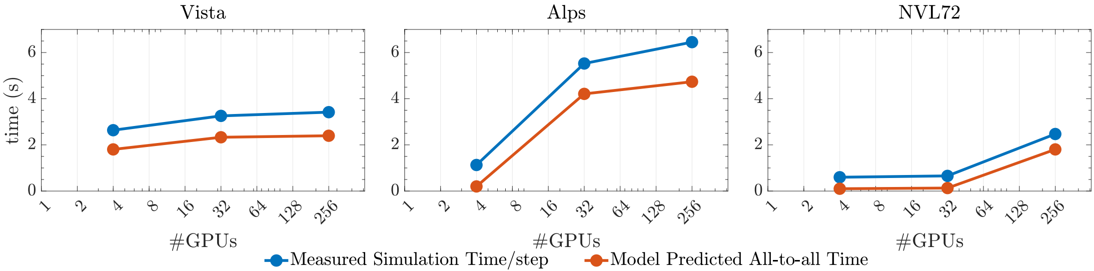
- Measured row (blue) and column (yellow) communicator A2A time (solid color) versus model predictions (faded color) for different $P_r \times P_c$.
- Fixed message size m over #GPUs $= 32$ and 256.
- $P_r = 1$ most performant, although significant impact of "near" group size.
- Systems with large "near" groups beneficial for 2D decompositions.



Weak Scaling of Benchmark Code

What can the model tell us about weak scaling in a communication bandwidth bound application?

- Measured simulation time/step closely mirrors the predicted theoretical best t_{a2a}.



Weak scaling is not flat: the near and self communication pathways contribute time improvements far beyond the “near” group size n.

Summary & Conclusions

Peak Performance Model

- ✓ Modeling paradigm to obtain an accurate quantitative measure of all-to-all performance given message and system parameters.
- ✓ Accounts for hierarchical structure of modern multi-GPU per node systems with global "far", tightly connected "near" and local "self" communication pathways.

Application and Validation

- ✓ Benchmarking tests on three systems validate theoretical upper bound from model as well as model predictions for 2D distributed transpose.
- ✓ "near" and "self" pathways boost communication well beyond "near" group size.
- ✓ No single communication library hits peak all-to-all performance across all process counts and systems.
- ✓ Weak scaling is not flat despite near-peak all-to-all communication performance.