GPU Resource Partitioning on SDumont II

NUMA Locality Impact on CPU-GPU Bandwidth

Pablo Alessandro Santos Hugen

Universidade Federal do Rio Grande do Sul - Instituto de Informática

December 2025

Outline

Motivation

SDumont II GH200 node allocation modes:

- **Exclusive**: Full node for one job
- **Shared**: SLURM GRES partitions resources

Questions:

1. How does bandwidth vary with NUMA affinity?
2. Does SLURM preserve locality in shared mode?

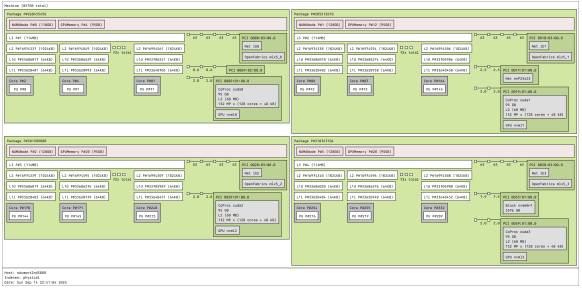
GH200 Node Architecture

Specifications:

- 4x NVIDIA GH200 GPUs (120GB HBM3)
- 288 ARM cores (72 per NUMA)
- NVLink-C2C: 900 GB/s (local)
- NVLink 4.0 between GPUs

NUMA affinity:

- GPU 0 ↔ NUMA 0
- GPU 1 ↔ NUMA 1
- GPU 2 ↔ NUMA 2
- GPU 3 ↔ NUMA 3



Experimental Setup

Tool: nvbandwidth v0.6

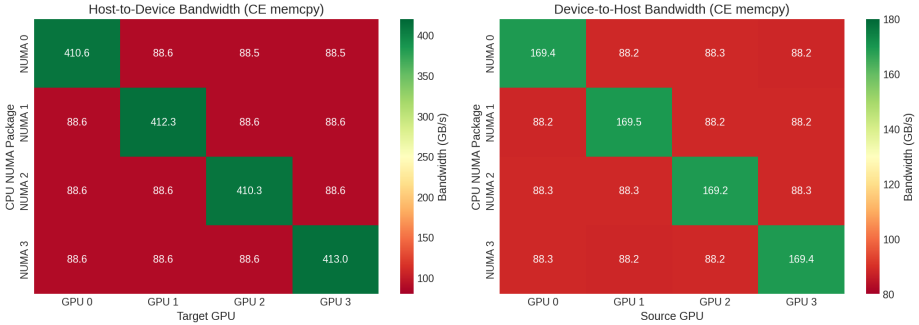
Exclusive queue:

- Pin process to each NUMA (0-3) with `numactl`
- Measure bandwidth to all 4 GPUs
- Result: 4×4 bandwidth matrix

Shared queue:

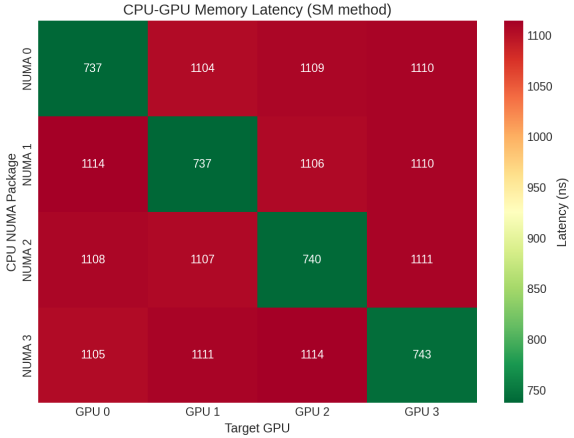
- Submit 4 concurrent jobs with 1 GPU each
- Observe SLURM's NUMA-GPU mapping

Host-to-Device Bandwidth



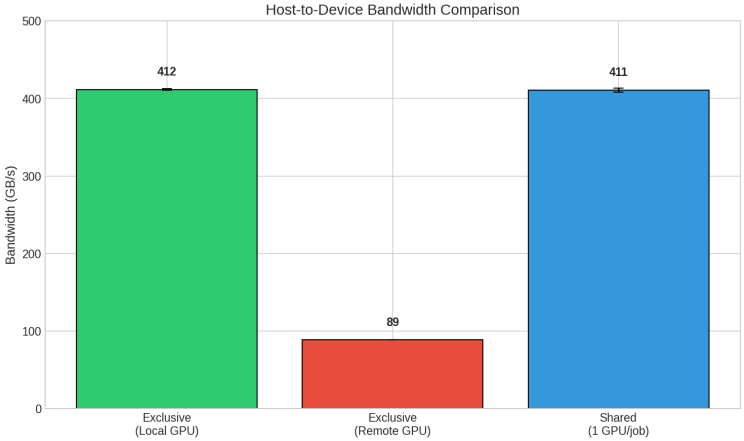
- **Local (diagonal):** 411.6 GB/s
- **Remote (off-diagonal):** 88.6 GB/s
- **Ratio:** 4.65×

CPU-GPU Latency



- Local: **739 ns**
- Remote: **1109 ns**
- Ratio: 1.5×

Exclusive vs Shared Queue



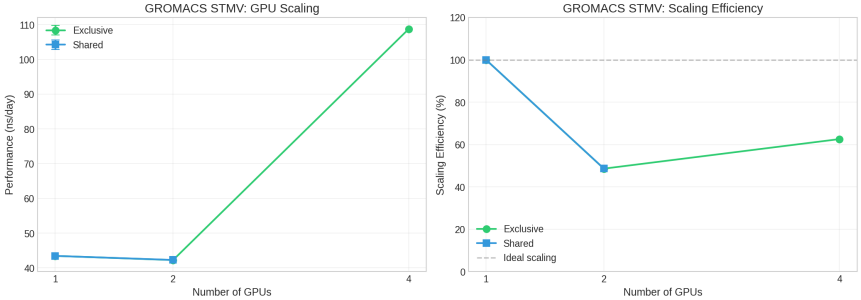
- Shared queue achieves ∼100% of local bandwidth
- SLURM preserves NUMA locality automatically

GROMACS Benchmark

Setup: GROMACS 2023.2, STMV (∼1M atoms), GPU offload: nb, bonded, pme

Queue	GPUs	ns/day	Scaling
Exclusive	1	43.45	1.00×
Exclusive	2	42.25	0.97×
Exclusive	4	108.75	**2.50**×
Shared	1	43.45	1.00×
Shared	2	42.28	0.97×

GROMACS GPU Scaling



- **1-GPU:** Same performance in both queues
- **2-GPU:** No scaling benefit
- **4-GPU:** 2.5× speedup

Summary

Metric	Value
H2D Bandwidth (Local)	411.6 GB/s
H2D Bandwidth (Remote)	88.6 GB/s
D2H Bandwidth (Local)	169.4 GB/s
D2H Bandwidth (Remote)	88.3 GB/s
Local/Remote Ratio	**4.65**×

Conclusions

Bandwidth:

- Local: 411 GB/s | Remote: 88 GB/s (4.6× difference)
- Shared queue maintained NUMA locality

GROMACS (STMV):

- Single-GPU: Same performance in both queues
- 4-GPU: 2.5× speedup | 2-GPU: No benefit

Guidance:

- Single-GPU jobs: Shared queue has no penalty
- Multi-GPU GROMACS (STMV): Use 1 or 4 GPUs

Thank you!

Questions?