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 offers two allocation modes for GH200 nodes:

- **Exclusive queue**: Full node reserved for one job
- **Shared queue**: SLURM GRES scheduling among jobs

Research questions:

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

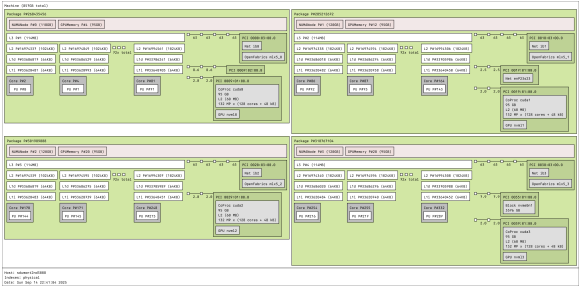
GH200 Node Architecture

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

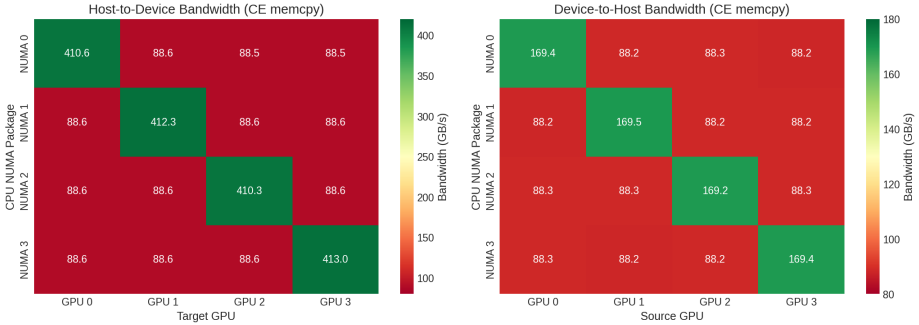
Exclusive queue:

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

Shared queue:

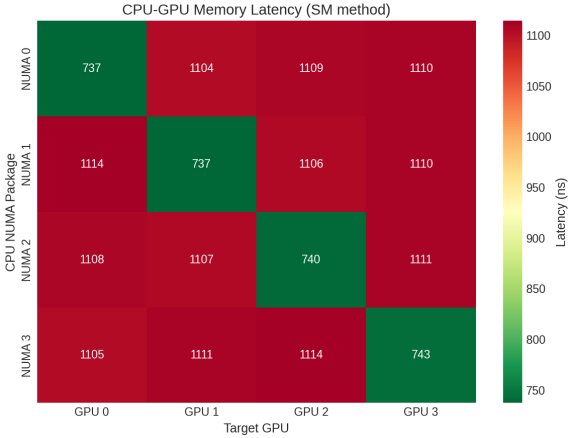
- Submit 4 concurrent jobs with 1 GPU each
- Observe SLURM's NUMA-GPU mapping

Host-to-Device Bandwidth



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

CPU-GPU Latency

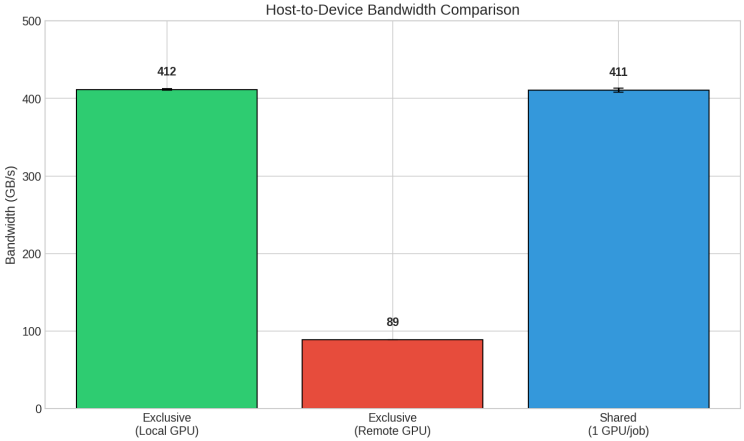


- Local: **739 ns**
- Remote: **1109 ns**
- Ratio: 1.5×

Observation

Remote access is bandwidth-limited, not latency-limited.

Exclusive vs Shared Queue



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

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

1. **NUMA locality is critical:** 4.65× bandwidth difference
2. **SLURM preserves locality:** Shared queue maintains CPU-GPU affinity
3. **GPU-GPU is NUMA-independent:** NVLink unaffected by CPU placement

Workload	Recommendation
Memory-bound	Exclusive queue
Compute-bound	Shared queue OK
Multi-GPU	Either queue

Thank you!

Questions?