Sandia National Laboratories

LATENCY AND BANDWIDTH MICROBENCHMARKS OF SIX US DEPARTMENT OF ENERGY SYSTEMS IN THE TOP500

Carl Pearson[1], Christopher M. Siefert[1], Stephen L. Olivier[1], Andrey Prokopenko[2], Timothy J. Fuller[1], Jonathan J. Hu[1]

[1] Sandia National Laboratories [2] Oak Ridge National Laboratory

Problem

- Many applications are becoming performance-portable
- Acceptance testing results are not generally public
- Existing benchmark publications compare few systems
- Ad-hoc measurements fragmented through literature

Contribution

- MPI latency, CPU/accelerator memory bandwidth, accelerator copy latency, and accelerator control latency benchmark results from six archetypal systems in the June 2023 Top500 [1] list

System Name	Top500 Rank	Loc.	CPU	Accelerator	CPU Compiler	GPU Compiler	MPI
Frontier	1	ORNL	AMD Zen 3	AMD MI250X	hipcc 5.3.0		cray-mpich/8.1.23
Summit	5	ORNL	IBM POWER9	NVIDIA V100	xl/16.1.1-10	nvcc 11.0.3	spectrum-mpi/10.4.0.3-20210112
Perlmutter[1]	8	NERSC	AMD Zen 3	NVIDIA A100	gcc/11.2.0	nvcc 11.7.64	cray-mpich/8.1.25
Trinity	29	LANL	Intel KNL	--	intel/2021.5.0	--	cray-mpich/7.7.20
Sawtooth	109	INL	Intel Cascade Lake	--	intel/19.0.5	--	intel-mpi/2019.0.117
Eagle	127	NREL	Intel Skylake	--	gcc/8.4.0	--	openmpi/4.1.0

Table 1: Summary of representative DOE systems in the June 2023 Top500. [1]PrgEnv-gnu.

Measurement Strategy

- OSU MPI Microbenchmarks 7.1 [2]
- Comm|Scope 0.12.0 [3]
- BabelStream 4.0 [4]
- Default system environment + GPU/MPI enablement
- Mean and standard deviation of 100 samples

System Name	CPU	GPU
Frontier	111.97 ± 0.24	1,368.69 ± 0.11
Summit	237.42 ± 0.24	805.30 ± 0.11
Perlmutter	112.91 ± 0.26	1,396.47 ± 0.24
Trinity	256.64 ± 2.11	N/A
Sawtooth	238.70 ± 8.39	N/A
Eagle	208.24 ± 0.92	N/A

Table 2: BabelStream COPY bandwidths (GB/s).

STREAM COPY Bandwidth

- BabelStream's `omp-stream`, `hip-stream`, and `cuda-stream` benchmarks
- Single-socket systems feature lower aggregate CPU bandwidth
- Trinity, Sawtooth, and Eagle do not have accelerators

MPI Latency

- OSU benchmarks `pt2pt`
- Point-to-point MPI latency
- Hardware locality typically visible in latency measurements

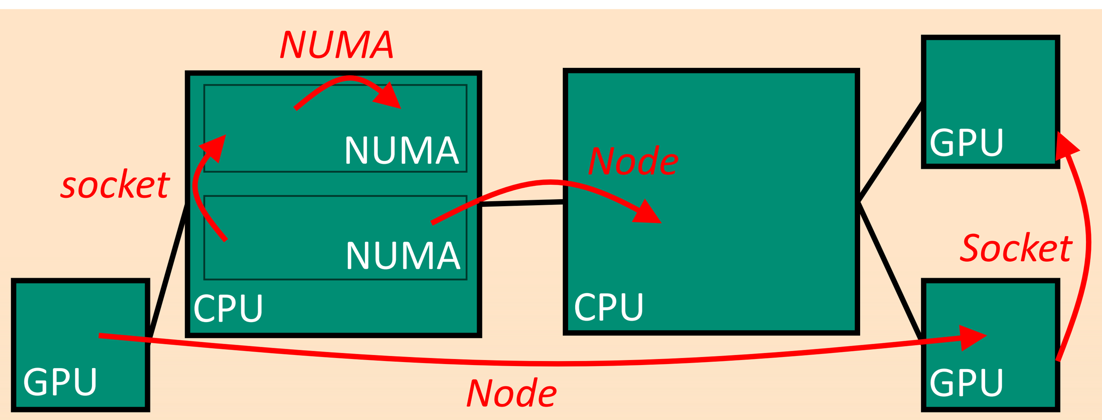

Fig. 1: Example of communication domains (Tab. 2).

System Name	On-Socket (μs)		GPU → GPU (μs)	
	Socket	Node	Socket	Node
Frontier	0.45 ± 0.01	N/A	N/A	0.43 ± 0.00
Summit	0.35 ± 0.08	0.86 ± 0.00	18.2 ± 0.22[2]	19.40 ± 0.20
Perlmutter	0.46 ± 0.06	1.11 ± 0.04	N/A	13.50 ± 0.13
Trinity	0.67 ± 0.01	0.99 ± 0.01	N/A	N/A
Sawtooth	0.48 ± 0.01[1]		N/A	N/A
Eagle	0.17 ± 0.00	0.38 ± 0.01	N/A	N/A

Table 3: MPI latencies. Column subheadings indicate the communication domain. [1] These two measurements are the same. [2] Refers to GPUs attached to the same POWER9 CPU.

Accelerator Intranode Bandwidth and Latencies

- Comm|Scope's `MemcpyAsync`, `DeviceSynchronize`, and `kernel` benchmarks
- Interconnect heterogeneity on Frontier and Summit (Figs. 2, 3) have a significant impact in measured transfer bandwidths. Latencies are not affected.

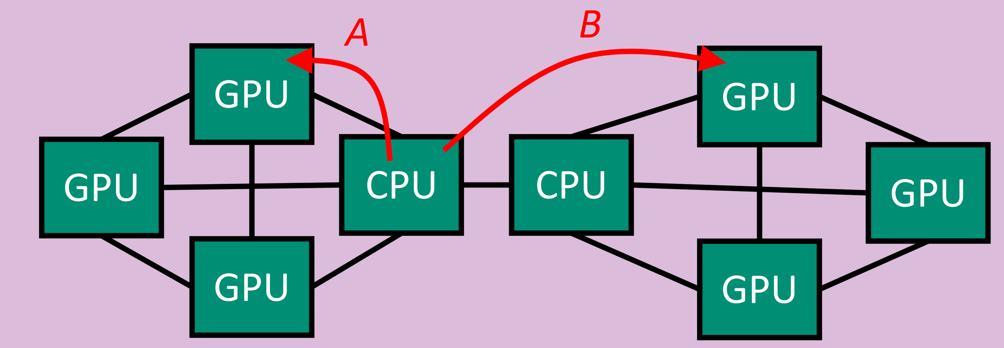

Fig. 2: Summit node showing transfer kinds (Tab. 3).

System Name	Host/GPU (GB/s)		GPU/GPU (GB/s)		
	A	B	A	B	C,D
Frontier	26.70 ± 0.00	N/A	50.90 ± 0.00	50.95 ± 0.00	36.95 ± 0.00
Summit	47.91 ± 0.00	37.61 ± 0.03	34.17 ± 0.01	30.29 ± 0.21	N/A
Perlmutter	26.50 ± 0.00	N/A	19.30 ± 0.05	N/A	N/A

Table 4: Intranode transfer bandwidths (GB/s). Host/GPU is mean of host-to-device and device-to-host

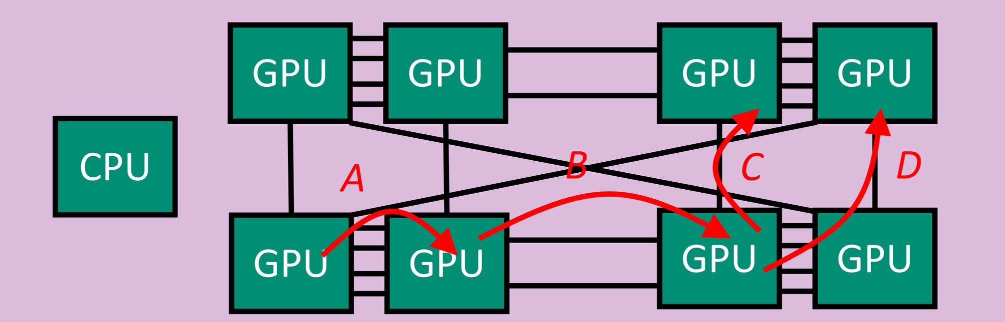

Fig. 3: Frontier node showing transfer kinds (Tabs. 3, 4).

System Name	Kernel (μs)	Sync (μs)	Host/GPU (μs)	GPU → GPU (μs)			
				A	B	C	D
Frontier	1.50 ± 0.00	0.14 ± 0.00	13.03 ± 0.05	12.02 ± 0.05	12.56 ± 0.03	12.68 ± 0.02	12.02 ± 0.10
Summit	4.7 ± 0.00	4.54 ± 0.00	7.70 ± 0.03	24.97 ± 0.15	27.44 ± 0.14	N/A	N/A
Perlmutter	1.77 ± 0.01	4.24 ± 0.01	4.24 ± 0.01	14.74 ± 0.41	N/A	N/A	N/A

Table 5: GPU control and transfer latencies.

References

[1] TOP500 June 2023. [Online]. Available: https://www.top500.org/lists/top500/2023/06/
[2] OSU micro-benchmarks. [Online]. Available: http://mvapich.cse.ohio-state.edu/benchmarks/
[3] C. Pearson, A. Dakkak, S. Hashash, C. Li, I.-H. Chung, J. Xiong, and W.-M. Hwu, "Evaluating characteristics of CUDA communication primitives on high-bandwidth interconnects," in Proceedings of the 2019 ACM/SPEC International Conference on Performance Engineering, 2019, pp. 209–218
[4] T. Deakin, J. Price, M. Martineau, and S. McIntosh-Smith, "Evaluating attainable memory bandwidth of parallel programming models via BabelStream," International Journal of Computational Science and Engineering, vol. 17, no. 3, pp. 247–262, 2018.

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2023-09020D

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

This research made use of Idaho National Laboratory computing resources which are supported by the Office of Nuclear Energy of the U.S. Department of Energy and the Nuclear Science User Facilities under Contract No. DE-AC07-05ID14517.

This poster has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy.

This research used resources of the Los Alamos National Laboratory, supported by the U.S. Department of Energy under contract No. 89233218CNA000001.

Sandia National Laboratories