Sandia National Laboratories

Exceptional service in the national interest

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

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

Sandia National Laboratories, [1]Oak Ridge National Laboratory

Tue Nov 2, 2023

Poster Lightning Talks, Cluster 2023

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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.

SAND2023-11711C

INTRO

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

BabelStream [2]

- CPU memory bandwidth
- GPU memory bandwidth

OSU MPI Microbenchmarks [3]

- CPU/CPU MPI Latency
- GPU/GPU MPI Latency

Comm|Scope [4]

- GPU control latencies
- CPU/GPU copy bandwidth
- GPU/GPU copy bandwidth

- 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

Upcoming paper in PMBS'23 (SC workshop)
More systems
More benchmarks
More discussion

[1] TOP500 June 2023. [Online]. Available: https://www.top500.org/lists/top500/2023/06/
[2] 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.
[3] OSU micro-benchmarks. [Online]. Available: http://mvapich.cse.ohio-state.edu/benchmarks/
[4] 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

SYSTEMS

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.

- Variety of CPUs and GPUS of interest to the U.S. DOE
- Maintain system defaults wherever possible
  - modules, compilers, drivers, etc…
- Mean & standard deviation of 100 samples

STREAM COPY BANDWIDTH

dual-socket CPUs

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).

MPI LATENCY

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.

CPU/GPU TRANSFER BANDWIDTHS

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

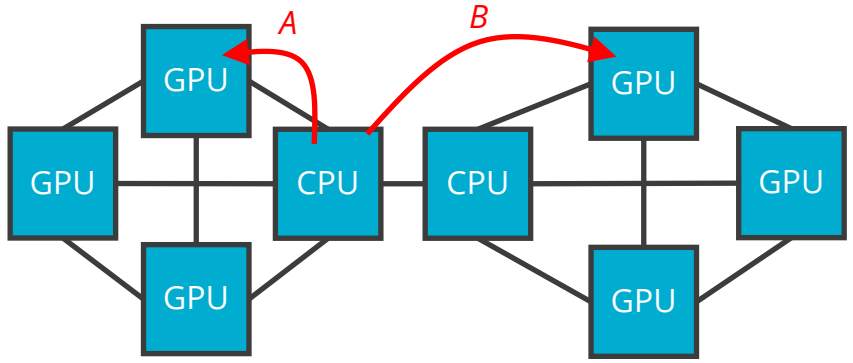

Fig. 2: Summit node showing transfer kinds (Tab. 3).

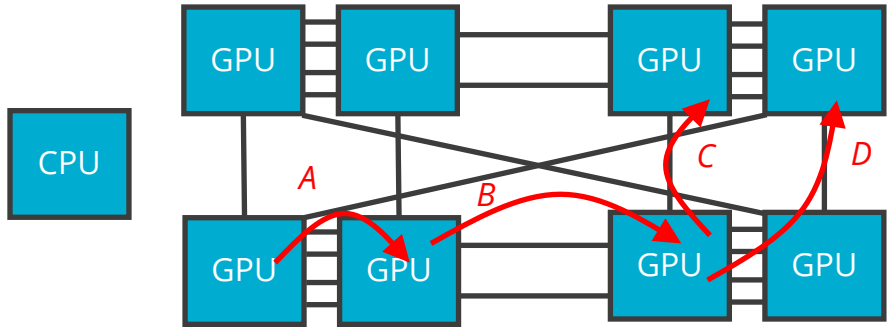

Fig. 3: Frontier node showing transfer kinds (Tabs. 3, 4).

GPU CONTROL AND TRANSFER LATENCIES

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.

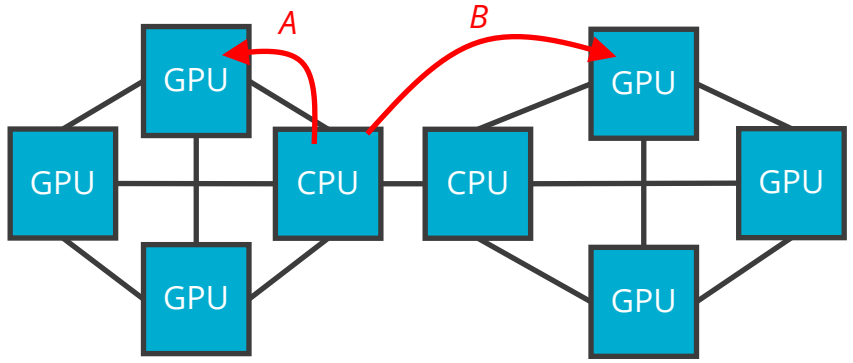

Fig. 2: Summit node showing transfer kinds (Tab. 3).

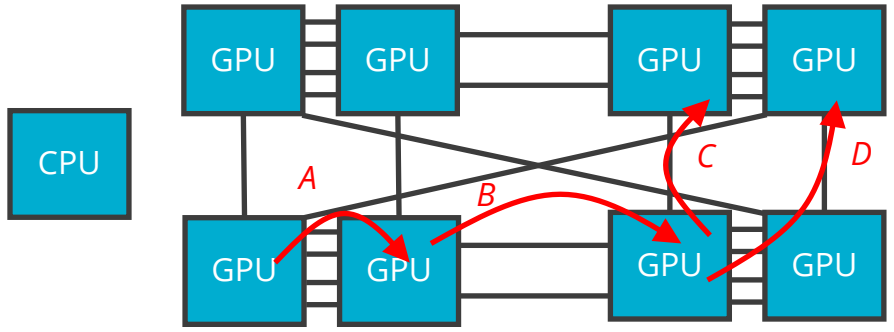

Fig. 3: Frontier node showing transfer kinds (Tabs. 3, 4).

THANK YOU