Latency and Bandwidth Microbenchmarks of Six US Department of Energy Systems in the Top500

Carl Pearson*, Christopher M. Siefert*, Stephen L. Olivier*, Andrey Prokopenko†,
Timothy J. Fuller*, and Jonathan J. Hu*
*Sandia National Laboratories, Albuquerque, NM, USA
{cwpears, csiefer, slolivi, tjfulle, jhu}@sandia.gov
†Oak Ridge National Laboratory, Oak Ridge, TN, USA
prokopenkoav@ornl.gov

***Abstract*—Developers of portable high-performance computing applications are often concerned with basic performance properties across a variety of systems. Although supercomputing systems are comprehensively benchmarked during their acceptance testing process, results are not publicly disseminated and comparisons are typically restricted to immediate predecessor systems. This work presents selected single-node microbenchmarks of archetypal United States Department of Energy computers present in the June 2023 Top 500 list. These systems feature Intel or AMD CPUs, including Xeon Phi, and AMD or Nvidia GPUs, and provide a reasonable reference for what users can expect from these systems.**

***Index Terms*—Graphics processors, Computer performance, High performance computing, Software performance**

I. Introduction and Background

Many high-performance applications are concerned with achieving high performance on a variety of different high-performance computing (HPC) systems. Therefore, developers must understand key performance properties for a variety of systems. We present initial work on a reference for select single-node MPI latency and memory bandwidth on a representative set of systems. Specifically, we contribute

- MPI latency, CPU/GPU memory bandwidth, GPU copy latency, and GPU control latency benchmark results from six archetypal United States Department of Energy (DOE) systems in the June 2023 Top500 list.

This work augments related work gathering HPC performance data. First, performance measurements are usually made during "acceptance" testing when a computer is installed. Only a limit set of these results are made available to application developers.

Second, application developers make ad-hoc measurements during application development and performance tuning. These measurements range from simple measures like those we present, to highly application-specific scenarios. Unfortunately, they are scattered across a wide body of otherwise unrelated publications and it is unclear how comparable different sources are. This can lead to a fragmented and incomplete understanding of system performance in the HPC community.

Third, a variety of HPC microbenchmarks have been developed to measure specific latency, bandwidth, and compute scenarios. We leverage Comm|Scope [1], The Ohio State University's MPI microbenchmarks [2], and BabelStream [3], and evaluate them over a representative set of systems.

II. Experimental Setup

We present node-level latencies and bandwidths of six archetypal representatives of the fourteen DOE systems present in the June 2023 Top 500 [4]. Table I summarizes the six selected systems.

All benchmarks are selected from BabelStream 4.0 (BS), OSU Microbenchmarks 7.1 (OSU), and Comm|Scope 0.12.0 (CS). `pt2pt` (OSU) and `omp-stream` (BS) are used to measure CPU MPI latency and CPU STREAM bandwidth, respectively, for all systems. For systems with GPUs additional benchmarks are run. `MemcpyAsync`, `DeviceSynchronize`, and `kernel` from CS are used to measure intranode transfer bandwidths and GPU control latencies. `pt2pt` is further used to measure device-involved MPI latency. `{cuda,hip}-stream` from BS are used for device memory bandwidths.

The benchmarks are run in their default configuration to produce a sample (internally, the benchmarks may aggregate multiple measurements to produce the single reported sample, any such behavior is unmodified). The reported mean and standard deviation come from one hundred such samples. The

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. SAND2023-09018C. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

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

TABLE I
SUMMARY OF EVALUATED DOE COMPUTERS. [1] PRGENV-GNU.

System Name	Top500 Rank	Loc.	CPU	GPU	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	*none*	intel 2021.5.0	*none*	cray-mpich 7.7.20
Sawtooth	109	INL	Intel Cascade Lake	*none*	intel 19.0.5	*none*	intel-mpi 2019.0.117
Eagle	127	NREL	Intel Skylake	*none*	gcc 8.4.0	*none*	openmpi 4.1.0

TABLE II
STREAM COPY BANDWIDTHS [MEAN(SD) GB/s]

System	CPU	GPU
Frontier	111.97(0.24)	1368.69(0.11)
Summit	237.42(0.24)	805.30(0.11)
Perlmutter	112.91(0.26)	1396.47(0.24)
Trinity	256.64(2.11)	N/A
Sawtooth	238.70(8.39)	N/A
Eagle	208.24(0.92)	N/A

TABLE III
MPI LATENCY [MEAN(SD) µs]

	CPU to CPU		GPU to GPU	
System	Socket	Node	Socket	Node
Frontier	0.45(0.01)	N/A	N/A	0.44(0.00)
Summit	0.35(0.08)	0.86(0.00)	18.2(0.22)	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)		N/A	N/A
Eagle	0.17(0.00)	0.38(0.01)	N/A	N/A

TABLE IV
INTRANODE BANDWIDTH [MEAN(SD) GB/s]

	Host/GPU		GPU/GPU		
System	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	19.3(0.05)	N/A	N/A

TABLE V
GPU CONTROL AND MEMCOPY LATENCIES [MEAN(SD) µs]

				GPU/GPU			
System	Kernel	Sync	Host/GPU	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	4.70 (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

default environment on the systems is left unmodified, except to to enable GPU + MPI programming environments.

III. DISCUSSION

Tables II, III, IV, and V present the mean and (standard deviation) of STREAM COPY bandwidth, intranode bandwiths, MPI latencies, and GPU latencies.

Table II: there is an obvious distinction between single-socket and dual-socket systems in terms of aggregate CPU memory bandwidth. Perlmutter measurements used 40GB A100 GPUs, the majority in the system.

Table III: The column sub-headings indicate a shared regime for the latency measurement. CPU latencies are much lower than GPUs, except for Frontier. Frontier's GPU-to-GPU latency features a dicontinuous jump to 6.9 µs at 64KiB (not shown), suggesting a different implementation for larger buffers. The GPU latencies via MPI are faster than via the GPU programming APIs themselves, reflecting optimizations such as GPUDirect.

Table IV: For Frontier, "A", "B", and "C" refer to quad- dual- or single- Infinity Fabric Links. "D" refers to GPUs separated by two hops. "A" and "B" having the same bandwidth suggests the MI250 DMA engine can only generate about 50 GB/s of memory traffic. For Summit, "A" and "B" refer to same-socket or other-socket transfers.

Table V: "Kernel" is the wall-time consumed launching an empty kernel. "Sync" is the same for a device synchronize on an empty work queue. GPU-to-GPU subheadings have the same meaning as Tab. IV.

IV. CONCLUSION AND FUTURE WORK

This work demonstrates initial steps towards a simple reference for HPC node. An expanded work that includes a more thorough set of microbenchmarks across all DOE systems above rank 150 in the Top500 will be published at the 2023 Performance Modeling, Benchmarking, and Simulation workshop at The International Conference for High Performance Computing, Networking, Storage, and Analysis ("Supercomputing"). Followup work will attempt to identify and measure key inter-node communication performance properties.

ACKNOWLEDGMENT

The authors would also like to thank Christopher Knight of Argonne National Laboratory and James Elliott of Sandia National Laboratories for consulting on configuration and run options on various machines.

REFERENCES

[1] 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.

[2] OSU micro-benchmarks. [Online]. Available: http://mvapich.cse.ohio-state.edu/benchmarks/

[3] 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.

[4] TOP500 June 2023. [Online]. Available: https://www.top500.org/lists/top500/2023/06/