Sandia National Laboratories

Exceptional service in the national interest

LATENCY AND BANDWIDTH MICROBENCHMARKS OF US DEPARTMENT OF ENERGY SYSTEMS IN THE JUNE 2023 TOP500 LIST

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

Sandia National Laboratories [1]Oak Ridge National Laboratory

14th IEEE Intl. Workshop on PMBS

November 13th, 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.

U.S. DEPARTMENT OF ENERGY

NNSA

SAND2023-13518C

MOTIVATION

- These slides: **carlpearson.net/talk/20231113_pmbs/**

- Portable application codes
  - Performance on a variety of machines

- "Acceptance Testing" inadequate for community knowledge base
  - Part of the supercomputer procurement process
  - Performance (and stability) of **one machine**
  - Varying degrees of public release

- Our goal: provide a one-stop shop for key intra-node performance measures on representative machines

Any subjective views or opinions that might be expressed do not necessarily represent the views of the U.S. Government.

SYSTEM SELECTION

- Top500[1] is a biannual list of the 500 computers with the fastest LINPACK performance
  - June 2023 was the most recent at the time of the work
- US Department of Energy has a wide variety of machines
- Some machines have been decommissioned since June 2023
  - Summit will stop accepting batch jobs in December
- Arbitrary cutoff at #150

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

SELECTED SYSTEMS (HARDWARE)

Name	Top500 Rank	Loc.	CPU	GPU
Frontier	1	ORNL	AMD EPYC	AMD MI250X
Summit	5	ORNL	IBM POWER9	NVIDIA V100
Sierra	6	LLNL	IBM POWER9	NVIDIA V100
Perlmutter	8	NERSC	AMD EPYC 7763	NVIDIA A100[1]
Polaris	19	ANL	AMD EPYC 7532	NVIDIA A100
Trinity	27	LANL	Intel Xeon Phi 7250	--
Lassen	36	LLNL	IBM Power9	NVIDIA V100
Theta	94	ANL	Intel Xeon Phi 7230	--
Sawtooth	109	INL	Intel Xeon Platinum 8268	--
RZVernal	116	LLNL	AMD EPYC	AMD MI250X
Eagle	127	NREL	Intel Xeon Gold 6154	--
Tioga	132	LLNL	AMD EPYC	AMD MI250X
Manzano	141	SNL	Intel Xeon Platinum 8268	--

[1]40GB

SELECTED SYSTEMS (HARDWARE)

Name	Top500 Rank	Loc.	CPU	GPU
Frontier	1	ORNL	AMD EPYC	AMD MI250X
Summit	5	ORNL	IBM POWER9	NVIDIA V100
Sierra	6	LLNL	IBM POWER9	NVIDIA V100
Perlmutter	8	NERSC	AMD EPYC 7763	NVIDIA A100[1]
Polaris	19	ANL	AMD EPYC 7532	NVIDIA A100
Trinity	27	LANL	Intel Xeon Phi 7250	--
Lassen	36	LLNL	IBM Power9	NVIDIA V100
Theta	94	ANL	Intel Xeon Phi 7230	--
Sawtooth	109	INL	Intel Xeon Platinum 8268	--
RZVernal	116	LLNL	AMD EPYC	AMD MI250X
Eagle	127	NREL	Intel Xeon Gold 6154	--
Tioga	132	LLNL	AMD EPYC	AMD MI250X
Manzano	141	SNL	Intel Xeon Platinum 8268	--

[1]40GB

IBM, Intel, and AMD CPUs

SELECTED SYSTEMS (HARDWARE)

Name	Top500 Rank	Loc.	CPU	GPU
Frontier	1	ORNL	AMD EPYC	AMD MI250X
Summit	5	ORNL	IBM POWER9	NVIDIA V100
Sierra	6	LLNL	IBM POWER9	NVIDIA V100
Perlmutter	8	NERSC	AMD EPYC 7763	NVIDIA A100[1]
Polaris	19	ANL	AMD EPYC 7532	NVIDIA A100
Trinity	27	LANL	Intel Xeon Phi 7250	--
Lassen	36	LLNL	IBM Power9	NVIDIA V100
Theta	94	ANL	Intel Xeon Phi 7230	--
Sawtooth	109	INL	Intel Xeon Platinum 8268	--
RZVernal	116	LLNL	AMD EPYC	AMD MI250X
Eagle	127	NREL	Intel Xeon Gold 6154	--
Tioga	132	LLNL	AMD EPYC	AMD MI250X
Manzano	141	SNL	Intel Xeon Platinum 8268	--

[1] 40GB

multi-core CPUs

Intel Knight's Landing

SELECTED SYSTEMS (HARDWARE)

Name	Top500 Rank	Loc.	CPU	GPU
Frontier	1	ORNL	AMD EPYC	AMD MI250X
Summit	5	ORNL	IBM POWER9	NVIDIA V100
Sierra	6	LLNL	IBM POWER9	NVIDIA V100
Perlmutter	8	NERSC	AMD EPYC 7763	NVIDIA A100[1]
Polaris	19	ANL	AMD EPYC 7532	NVIDIA A100
Trinity	27	LANL	Intel Xeon Phi 7250	--
Lassen	36	LLNL	IBM Power9	NVIDIA V100
Theta	94	ANL	Intel Xeon Phi 7230	--
Sawtooth	109	INL	Intel Xeon Platinum 8268	--
RZVernal	116	LLNL	AMD EPYC	AMD MI250X
Eagle	127	NREL	Intel Xeon Gold 6154	--
Tioga	132	LLNL	AMD EPYC	AMD MI250X
Manzano	141	SNL	Intel Xeon Platinum 8268	--

[1] 40GB

With and without GPUS

SELECTED SYSTEMS (HARDWARE)

Name	Top500 Rank	Loc.	CPU	GPU
Frontier	1	ORNL	AMD EPYC	AMD MI250X
Summit	5	ORNL	IBM POWER9	NVIDIA V100
Sierra	6	LLNL	IBM POWER9	NVIDIA V100
Perlmutter	8	NERSC	AMD EPYC 7763	NVIDIA A100[1]
Polaris	19	ANL	AMD EPYC 7532	NVIDIA A100
Trinity	27	LANL	Intel Xeon Phi 7250	--
Lassen	36	LLNL	IBM Power9	NVIDIA V100
Theta	94	ANL	Intel Xeon Phi 7230	--
Sawtooth	109	INL	Intel Xeon Platinum 8268	--
RZVernal	116	LLNL	AMD EPYC	AMD MI250X
Eagle	127	NREL	Intel Xeon Gold 6154	--
Tioga	132	LLNL	AMD EPYC	AMD MI250X
Manzano	141	SNL	Intel Xeon Platinum 8268	--

[1] 40GB

AMD and Nvidia GPUs

SELECTED SYSTEMS (HARDWARE)

Name	Top500 Rank	Loc.	CPU	GPU
Frontier	1	ORNL	AMD EPYC	AMD MI250X
Summit	5	ORNL	IBM POWER9	NVIDIA V100
Sierra	6	LLNL	IBM POWER9	NVIDIA V100
Perlmutter	8	NERSC	AMD EPYC 7763	NVIDIA A100[1]
Polaris	19	ANL	AMD EPYC 7532	NVIDIA A100
Trinity	27	LANL	Intel Xeon Phi 7250	--
Lassen	36	LLNL	IBM Power9	NVIDIA V100
Theta	94	ANL	Intel Xeon Phi 7230	--
Sawtooth	109	INL	Intel Xeon Platinum 8268	--
RZVernal	116	LLNL	AMD EPYC	AMD MI250X
Eagle	127	NREL	Intel Xeon Gold 6154	--
Tioga	132	LLNL	AMD EPYC	AMD MI250X
Manzano	141	SNL	Intel Xeon Platinum 8268	--

[1]40GB

Similar nodes

Frontier / RZVernal / Tioga

Summit / Sierra / Lassen

Perlmutter / Polaris

Trinity / Theta

SELECTED SYSTEMS (SOFTWARE)

- Multiple versions and vendors for compilers and MPI available on each system
- We stick with as default an experience as possible
  - ”log in and compile”
  - Our expectation is that defaults should be well-behaved
  - Sometimes, defaults are too old to compile benchmarks (e.g. gcc 4.9.3 on Lassen)

SELECTED SYSTEMS (SOFTWARE)

Name	Top500 Rank	Compiler	Acc. Toolchain	MPI
Frontier	1	amd-mixed/5.3.0		cray-mpich/8.1.23
Summit	5	xl/16.1.1-10	cuda/11.0.3	spectrum-mpi/10.4.0.3-20210112
Sierra	6	gcc/8.3.1	cuda/10.1.243	spectrum-mpi/rolling-release
Perlmutter	8	gcc/11.2.0	cuda/11.7	cray-mpich/8.1.25
Polaris	19	nvhpc/21.9	cuda/11.4	cray-mpich/8.1.16
Trinity	27	intel/2022.0.2	--	cray-mpich/7.7.20
Lassen	36	gcc/8.3.1	cuda/10.1.243	spectrum-mpi/rolling-release
Theta	94	intel/19.1.0.166	--	cray-mpich/7.7.14
Sawtooth	109	intel/19.0.5	--	intel-mpi/2019.0.117
RZVernal	116	amd/5.6.0		cray-mpich/8.1.26
Eagle	127	gcc/8.4.0	--	openmpi/4.1.0
Tioga	132	amd/5.6.0		cray-mpich/8.1.26
Manzano	141	intel/16.0	--	openmpi/1.10

SELECTED SYSTEMS (SOFTWARE)

Name	Top500 Rank	Compiler	Acc. Toolchain	MPI
Frontier	1	amd-mixed/5.3.0		cray-mpich/8.1.23
Summit	5	xl/16.1.1-10	cuda/11.0.3	spectrum-mpi/10.4.0.3-20210112
Sierra	6	gcc/8.3.1	cuda/10.1.243	spectrum-mpi/rolling-release
Perlmutter	8	gcc/11.2.0	cuda/11.7	cray-mpich/8.1.25
Polaris	19	nvhpc/21.9	cuda/11.4	cray-mpich/8.1.16
Trinity	27	intel/2022.0.2	--	cray-mpich/7.7.20
Lassen	36	gcc/8.3.1	cuda/10.1.243	spectrum-mpi/rolling-release
Theta	94	intel/19.1.0.166	--	cray-mpich/7.7.14
Sawtooth	109	intel/19.0.5	--	intel-mpi/2019.0.117
RZVernal	116	amd/5.6.0		cray-mpich/8.1.26
Eagle	127	gcc/8.4.0	--	openmpi/4.1.0
Tioga	132	amd/5.6.0		cray-mpich/8.1.26
Manzano	141	intel/16.0	--	openmpi/1.10

GNU, LLVM, and vendor compilers

SELECTED SYSTEMS (SOFTWARE)

Name	Top500 Rank	Compiler	Acc. Toolchain	MPI
Frontier	1	amd-mixed/5.3.0		cray-mpich/8.1.23
Summit	5	xl/16.1.1-10	cuda/11.0.3	spectrum-mpi/10.4.0.3-20210112
Sierra	6	gcc/8.3.1	cuda/10.1.243	spectrum-mpi/rolling-release
Perlmutter	8	gcc/11.2.0	cuda/11.7	cray-mpich/8.1.25
Polaris	19	nvhpc/21.9	cuda/11.4	cray-mpich/8.1.16
Trinity	27	intel/2022.0.2	--	cray-mpich/7.7.20
Lassen	36	gcc/8.3.1	cuda/10.1.243	spectrum-mpi/rolling-release
Theta	94	intel/19.1.0.166	--	cray-mpich/7.7.14
Sawtooth	109	intel/19.0.5	--	intel-mpi/2019.0.117
RZVernal	116	amd/5.6.0		cray-mpich/8.1.26
Eagle	127	gcc/8.4.0	--	openmpi/4.1.0
Tioga	132	amd/5.6.0		cray-mpich/8.1.26
Manzano	141	intel/16.0	--	openmpi/1.10

OpenMPI, Cray/MPICH, IBM, and Intel MPIs

BENCHMARK SELECTION

- Used pre-existing open-source microbenchmarks
  - Chosen based on portability, familiarity, and established use in the community

- BabelStream [1]
  - serial, OpenMP, CUDA, and HIP (among others)
  - STREAM bandwidth for serial, OpenMP, CUDA, and HIP
- OSU MPI Microbenchmarks [2]
  - MPI
  - Point-to-point latency
- Comm|Scope [3]
  - CUDA and HIP
  - CPU/GPU and GPU/GPU memcpy bandwidth, GPU control latencies

We only use MPI intra-node

- Measures mirror how many DOE applications are written
  - MPI to communicate between processes
  - GPU/GPU communication handled by GPU-aware MPI or by staging through host

[1] 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.
[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

CPU MEMORY BANDWIDTH AND MPI LATENCY

	Memory Bandwidth (GB/s)			MPI Latency (us)	
Rank. Name	One Core	All Cores	Peak (all cores)	On-Socket	On-Node
29. Trinity	12.36 ± 0.16	347.28 ± 5.76	> 450	0.67 ± 0.01	0.99 ± 0.01
94. Theta	18.76 ± 0.58	119.72 ± 0.54	> 450	5.95 ± 0.01	6.25 ± 0.05
109. Sawtooth	13.06 ± 0.35	238.70 ± 8.39	281.50	0.48 ± 0.01	0.48 ± 0.01
127. Eagle	13.45 ± 0.03	208.24 ± 0.92	255.91	0.17 ± 0.00	0.38 ± 0.01
141. Manzano	15.27 ± 0.05	234.86 ± 0.12	281.50	0.32 ± 0.00	0.56 ± 0.01

theoretical

Ranks on same socket

Ranks on different sockets

CPU MEMORY BANDWIDTH AND MPI LATENCY

	Memory Bandwidth (GB/s)			MPI Latency (us)	
Rank. Name	One Core	All Cores	Peak (all cores)	On-Socket	On-Node
29. Trinity	12.36 ± 0.16	347.28 ± 5.76	> 450	0.67 ± 0.01	0.99 ± 0.01
94. Theta	18.76 ± 0.58	119.72 ± 0.54	> 450	5.95 ± 0.01	6.25 ± 0.05
109. Sawtooth	13.06 ± 0.35	238.70 ± 8.39	281.50	0.48 ± 0.01	0.48 ± 0.01
127. Eagle	13.45 ± 0.03	208.24 ± 0.92	255.91	0.17 ± 0.00	0.38 ± 0.01
141. Manzano	15.27 ± 0.05	234.86 ± 0.12	281.50	0.32 ± 0.00	0.56 ± 0.01

- Peak number is MCDRAM bandwidth
  - Our working set fits in MCDRAM
- MCDRAM configured as a transparent cache

- These two systems are very similar on paper

CPU MEMORY BANDWIDTH AND MPI LATENCY

	Memory Bandwidth (GB/s)			MPI Latency (us)	
Rank. Name	One Core	All Cores	Peak (all cores)	On-Socket	On-Node
29. Trinity	12.36 ± 0.16	347.28 ± 5.76	> 450	0.67 ± 0.01	0.99 ± 0.01
94. Theta	18.76 ± 0.58	119.72 ± 0.54	> 450	5.95 ± 0.01	6.25 ± 0.05
109. Sawtooth	13.06 ± 0.35	238.70 ± 8.39	281.50	0.48 ± 0.01	0.48 ± 0.01
127. Eagle	13.45 ± 0.03	208.24 ± 0.92	255.91	0.17 ± 0.00	0.38 ± 0.01
141. Manzano	15.27 ± 0.05	234.86 ± 0.12	281.50	0.32 ± 0.00	0.56 ± 0.01

- Whether same-socket or cross-socket impacts performance varies

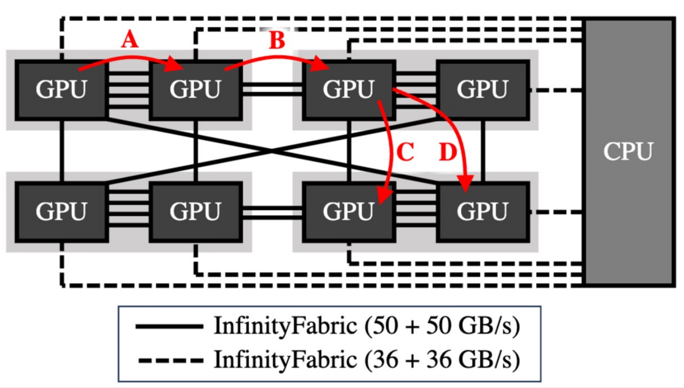
MEASUREMENTS FOR ACCELERATOR SYSTEMS

	Acc. Memory Bandwidth (GB/s)		MPI Latency (us)	MPI Latency (us) Acc.-to-Acc.			
Rank. Name	Measured	Peak	Host-to-Host	A	B	C	D
1. Frontier	1336.35 ± 1.11	1600	0.45 ± 0.01	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00
5. Summit	786.43 ± 0.11	900	0.34 ± 0.07	18.10 ± 0.22	19.30 ± 0.15		
6. Sierra	861.40 ± 0.65	900	0.38 ± 0.01	18.72 ± 0.12	19.76 ± 0.37		
8. Perlmutter	1363.74 ± 0.23	1555.2	0.46 ± 0.06	13.50 ± 0.13			
19. Polaris	1362.75 ± 0.17	1555.2	0.21 ± 0.00	10.42 ± 0.03			
36. Lassen	861.03 ± 0.53	900	0.37 ± 0.00	18.68 ± 0.20	19.72 ± 0.13		
116. RZVernal	1291.38 ± 0.77	1600	0.49 ± 0.00	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.00	0.49 ± 0.01
132. Tioga	1336.81 ± 0.97	1600	0.49 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.49 ± 0.01

theoretical

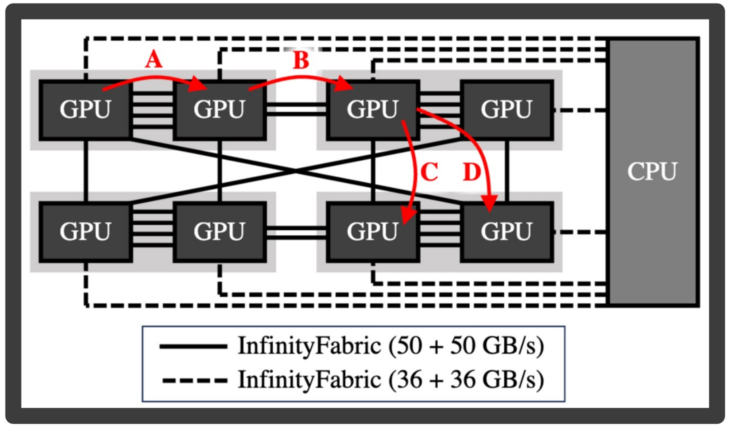
Two ranks on same CPU socket

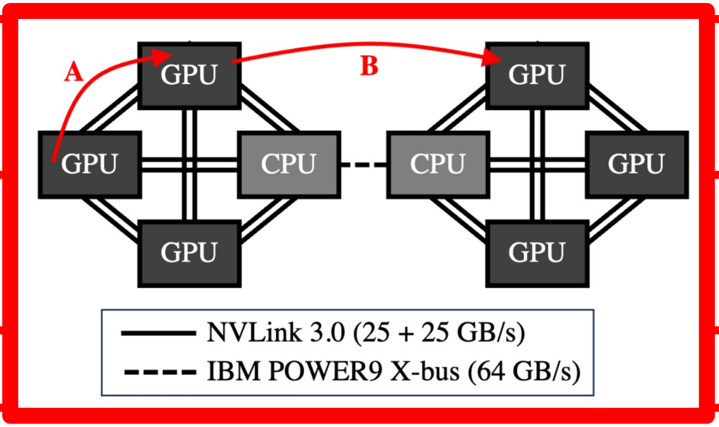
Two ranks on different accelerators near/fast (A) … to … far/slow (D) (more details on coming slides)



	MPI Latency (us) Acc.-to-Acc.			
Rank. Name	A	B	C	D
1. Frontier	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00
5. Summit	18.10 ± 0.22	19.30 ± 0.15		
6. Sierra	18.72 ± 0.12	19.76 ± 0.37		
8. Perlmutter	13.50 ± 0.13			
19. Polaris	10.42 ± 0.03			
36. Lassen	18.68 ± 0.20	19.72 ± 0.13		
116. RZVernal	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.00	0.49 ± 0.01
132. Tioga	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.49 ± 0.01

No measurable MPI latency difference on this node type

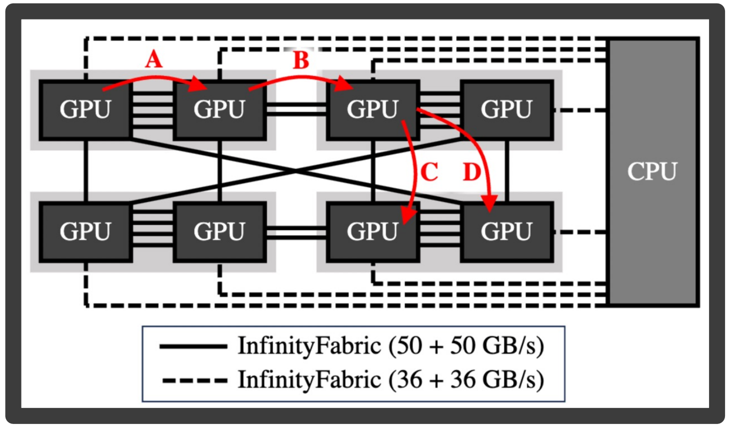


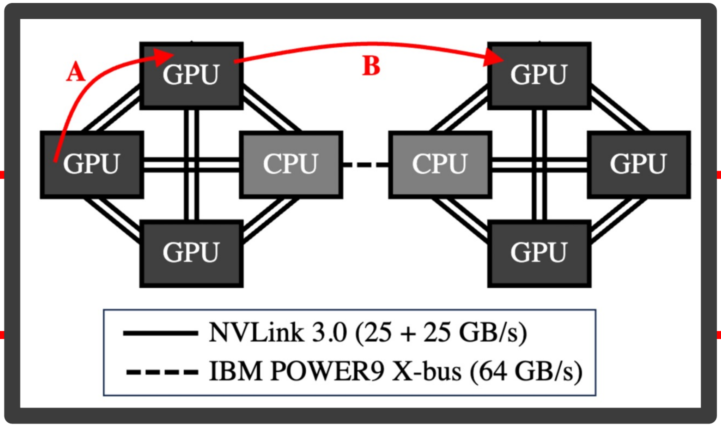


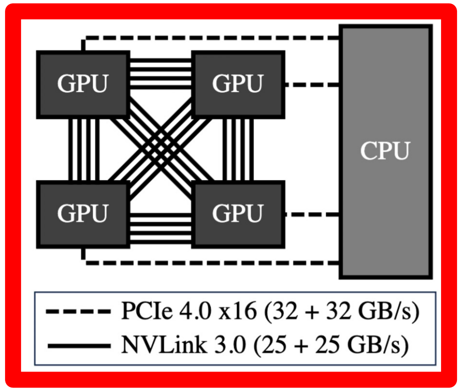
	MPI Latency (us) Acc.-to-Acc.			
Rank. Name	A	B	C	D
1. Frontier	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00
5. Summit	18.10 ± 0.22	19.30 ± 0.15		
6. Sierra	18.72 ± 0.12	19.76 ± 0.37		
8. Perlmutter	13.50 ± 0.13			
19. Polaris	10.42 ± 0.03			
36. Lassen	18.68 ± 0.20	19.72 ± 0.13		
116. RZVernal	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.00	0.49 ± 0.01
132. Tioga	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.49 ± 0.01

Although Summit / Sierra / Lassen have different numbers of GPUs connected by different lanes, A/B have the same meaning

Going across X-bus has a measurable latency penalty

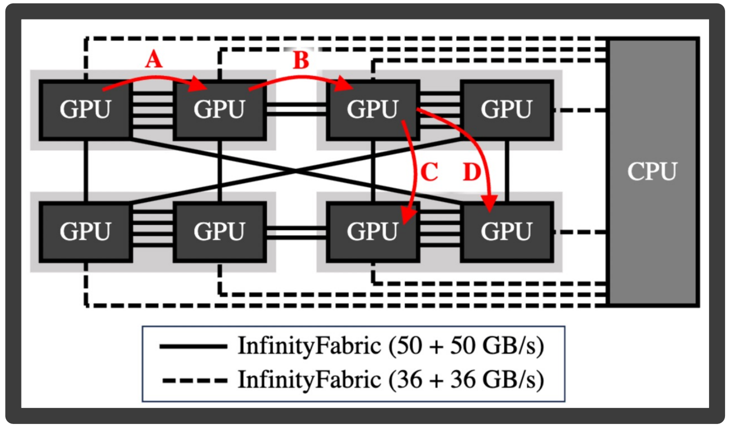


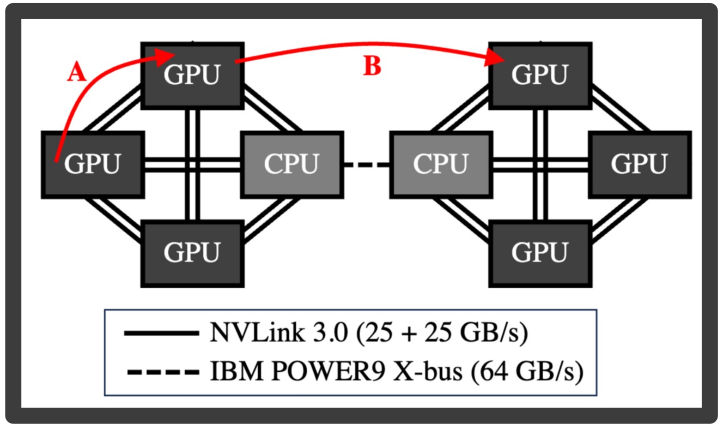


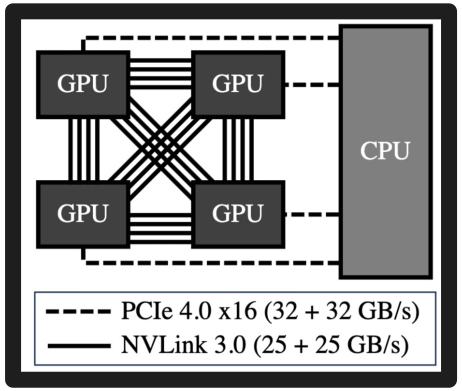


	MPI Latency (us) Acc.-to-Acc.			
Rank. Name	A	B	C	D
1. Frontier	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00
5. Summit	18.10 ± 0.22	19.30 ± 0.15		
6. Sierra	18.72 ± 0.12	19.76 ± 0.37		
8. Perlmutter	13.50 ± 0.13			
19. Polaris	10.42 ± 0.03			
36. Lassen	18.68 ± 0.20	19.72 ± 0.13		
116. RZVernal	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.00	0.49 ± 0.01
132. Tioga	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.49 ± 0.01

Identical all-to-all GPU connections







	MPI Latency (us) Acc.-to-Acc.			
Rank. Name	A	B	C	D
1. Frontier	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00	0.44 ± 0.00
5. Summit	18.10 ± 0.22	19.30 ± 0.15		
6. Sierra	18.72 ± 0.12	19.76 ± 0.37		
8. Perlmutter	13.50 ± 0.13			
19. Polaris	10.42 ± 0.03			
36. Lassen	18.68 ± 0.20	19.72 ± 0.13		
116. RZVernal	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.00	0.49 ± 0.01
132. Tioga	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.49 ± 0.01

Extremely low latency for small ($\leq 2^8$) messages.

ACCELERATOR CONTROL LATENCIES

	Kernel (us)		(H -> D + D -> H) / 2		D -> D Latency (us)			
Rank. Name	Launch	Wait	Latency (us)	Bandwidth (GB/s)	A	B	C	D
1. Frontier	1.51 ± 0.00	0.14 ± 0.00	12.91 ± 0.02	24.87 ± 0.01	12.02 ± 0.05	12.56 ± 0.03	12.68 ± 0.02	12.02 ± 0.10
5. Summit	4.84 ± 0.01	4.31 ± 0.01	7.82 ± 0.07	44.88 ± 0.00	24.97 ± 0.16	27.44 ± 0.14		
6. Sierra	4.13 ± 0.01	5.59 ± 0.02	7.27 ± 0.23	63.40 ± 0.01	23.91 ± 0.16	27.70 ± 0.12		
8. Perlmutter	1.77 ± 0.01	0.98 ± 0.00	4.24 ± 0.01	24.74 ± 0.00	14.74 ± 0.41			
19. Polaris	1.83 ± 0.00	1.32 ± 0.01	5.33 ± 0.02	23.71 ± 0.00	32.84 ± 0.30			
36. Lassen	4.56 ± 0.00	5.52 ± 0.01	7.76 ± 0.32	63.34 ± 0.02	24.56 ± 0.28	27.69 ± 0.10		
116. RZVernal	2.16 ± 0.01	0.12 ± 0.00	12.20 ± 0.07	24.88 ± 0.00	9.85 ± 0.01	12.58 ± 0.00	12.45 ± 0.02	10.21 ± 0.01
132. Tioga	2.15 ± 0.01	0.12 ± 0.00	12.19 ± 0.04	24.88 ± 0.00	9.85 ± 0.02	12.59 ± 0.01	12.46 ± 0.01	10.12 ± 0.02

launch empty kernel

sync empty queue

async memcpy

A/B/C/D same as previous slides

ACCELERATOR CONTROL LATENCIES

	Kernel (us)		(H -> D + D -> H) / 2		D -> D Latency (us)			
Rank. Name	Launch	Wait	Latency (us)	Bandwidth (GB/s)	A	B	C	D
1. Frontier	1.51 ± 0.00	0.14 ± 0.00	12.91 ± 0.02	24.87 ± 0.01	12.02 ± 0.05	12.56 ± 0.03	12.68 ± 0.02	12.02 ± 0.10
5. Summit	4.84 ± 0.01	4.31 ± 0.01	7.82 ± 0.07	44.88 ± 0.00	24.97 ± 0.16	27.44 ± 0.14		
6. Sierra	4.13 ± 0.01	5.59 ± 0.02	7.27 ± 0.23	63.40 ± 0.01	23.91 ± 0.16	27.70 ± 0.12		
8. Perlmutter	1.77 ± 0.01	0.98 ± 0.00	4.24 ± 0.01	24.74 ± 0.00	14.74 ± 0.41			
19. Polaris	1.83 ± 0.00	1.32 ± 0.01	5.33 ± 0.02	23.71 ± 0.00	32.84 ± 0.30			
36. Lassen	4.56 ± 0.00	5.52 ± 0.01	7.76 ± 0.32	63.34 ± 0.02	24.56 ± 0.28	27.69 ± 0.10		
116. RZVernal	2.16 ± 0.01	0.12 ± 0.00	12.20 ± 0.07	24.88 ± 0.00	9.85 ± 0.01	12.58 ± 0.00	12.45 ± 0.02	10.21 ± 0.01
132. Tioga	2.15 ± 0.01	0.12 ± 0.00	12.19 ± 0.04	24.88 ± 0.00	9.85 ± 0.02	12.59 ± 0.01	12.46 ± 0.01	10.12 ± 0.02

AMD kernel latencies

ACCELERATOR CONTROL LATENCIES

	Kernel (us)		(H -> D + D -> H) / 2		D -> D Latency (us)			
Rank. Name	Launch	Wait	Latency (us)	Bandwidth (GB/s)	A	B	C	D
1. Frontier	1.51 ± 0.00	0.14 ± 0.00	12.91 ± 0.02	24.87 ± 0.01	12.02 ± 0.05	12.56 ± 0.03	12.68 ± 0.02	12.02 ± 0.10
5. Summit	4.84 ± 0.01	4.31 ± 0.01	7.82 ± 0.07	44.88 ± 0.00	24.97 ± 0.16	27.44 ± 0.14		
6. Sierra	4.13 ± 0.01	5.59 ± 0.02	7.27 ± 0.23	63.40 ± 0.01	23.91 ± 0.16	27.70 ± 0.12		
8. Perlmutter	1.77 ± 0.01	0.98 ± 0.00	4.24 ± 0.01	24.74 ± 0.00	14.74 ± 0.41			
19. Polaris	1.83 ± 0.00	1.32 ± 0.01	5.33 ± 0.02	23.71 ± 0.00	32.84 ± 0.30			
36. Lassen	4.56 ± 0.00	5.52 ± 0.01	7.76 ± 0.32	63.34 ± 0.02	24.56 ± 0.28	27.69 ± 0.10		
116. RZVernal	2.16 ± 0.01	0.12 ± 0.00	12.20 ± 0.07	24.88 ± 0.00	9.85 ± 0.01	12.58 ± 0.00	12.45 ± 0.02	10.21 ± 0.01
132. Tioga	2.15 ± 0.01	0.12 ± 0.00	12.19 ± 0.04	24.88 ± 0.00	9.85 ± 0.02	12.59 ± 0.01	12.46 ± 0.01	10.12 ± 0.02

AMD X86 + NVIDIA A100

IBM POWER9 + NVIDIA V100

ACCELERATOR CONTROL LATENCIES

	Kernel (us)		(H -> D + D -> H) / 2		D -> D Latency (us)			
Rank. Name	Launch	Wait	Latency (us)	Bandwidth (GB/s)	A	B	C	D
1. Frontier	1.51 ± 0.00	0.14 ± 0.00	12.91 ± 0.02	24.87 ± 0.01	12.02 ± 0.05	12.56 ± 0.03	12.68 ± 0.02	12.02 ± 0.10
5. Summit	4.84 ± 0.01	4.31 ± 0.01	7.82 ± 0.07	44.88 ± 0.00	24.97 ± 0.16	27.44 ± 0.14		
6. Sierra	4.13 ± 0.01	5.59 ± 0.02	7.27 ± 0.23	63.40 ± 0.01	23.91 ± 0.16	27.70 ± 0.12		
8. Perlmutter	1.77 ± 0.01	0.98 ± 0.00	4.24 ± 0.01	24.74 ± 0.00	14.74 ± 0.41			
19. Polaris	1.83 ± 0.00	1.32 ± 0.01	5.33 ± 0.02	23.71 ± 0.00	32.84 ± 0.30			
36. Lassen	4.56 ± 0.00	5.52 ± 0.01	7.76 ± 0.32	63.34 ± 0.02	24.56 ± 0.28	27.69 ± 0.10		
116. RZVernal	2.16 ± 0.01	0.12 ± 0.00	12.20 ± 0.07	24.88 ± 0.00	9.85 ± 0.01	12.58 ± 0.00	12.45 ± 0.02	10.21 ± 0.01
132. Tioga	2.15 ± 0.01	0.12 ± 0.00	12.19 ± 0.04	24.88 ± 0.00	9.85 ± 0.02	12.59 ± 0.01	12.46 ± 0.01	10.12 ± 0.02

PCIE 4.0 between CPU and GPU

NVLink between CPU and GPU

ACCELERATOR CONTROL LATENCIES

	Kernel (us)		(H -> D + D -> H) / 2		D -> D Latency (us)			
Rank. Name	Launch	Wait	Latency (us)	Bandwidth (GB/s)	A	B	C	D
1. Frontier	1.51 ± 0.00	0.14 ± 0.00	12.91 ± 0.02	24.87 ± 0.01	12.02 ± 0.05	12.56 ± 0.03	12.68 ± 0.02	12.02 ± 0.10
5. Summit	4.84 ± 0.01	4.31 ± 0.01	7.82 ± 0.07	44.88 ± 0.00	24.97 ± 0.16	27.44 ± 0.14		
6. Sierra	4.13 ± 0.01	5.59 ± 0.02	7.27 ± 0.23	63.40 ± 0.01	23.91 ± 0.16	27.70 ± 0.12		
8. Perlmutter	1.77 ± 0.01	0.98 ± 0.00	4.24 ± 0.01	24.74 ± 0.00	14.74 ± 0.41			
19. Polaris	1.83 ± 0.00	1.32 ± 0.01	5.33 ± 0.02	23.71 ± 0.00	32.84 ± 0.30			
36. Lassen	4.56 ± 0.00	5.52 ± 0.01	7.76 ± 0.32	63.34 ± 0.02	24.56 ± 0.28	27.69 ± 0.10		
116. RZVernal	2.16 ± 0.01	0.12 ± 0.00	12.20 ± 0.07	24.88 ± 0.00	9.85 ± 0.01	12.58 ± 0.00	12.45 ± 0.02	10.21 ± 0.01
132. Tioga	2.15 ± 0.01	0.12 ± 0.00	12.19 ± 0.04	24.88 ± 0.00	9.85 ± 0.02	12.59 ± 0.01	12.46 ± 0.01	10.12 ± 0.02

Similar systems – configuration differences?

ACCELERATOR CONTROL LATENCIES

	Kernel (us)		(H -> D + D -> H) / 2		D -> D Latency (us)			
Rank. Name	Launch	Wait	Latency (us)	Bandwidth (GB/s)	A	B	C	D
1. Frontier	1.51 ± 0.00	0.14 ± 0.00	12.91 ± 0.02	24.87 ± 0.01	12.02 ± 0.05	12.56 ± 0.03	12.68 ± 0.02	12.02 ± 0.10
5. Summit	4.84 ± 0.01	4.31 ± 0.01	7.82 ± 0.07	44.88 ± 0.00	24.97 ± 0.16	27.44 ± 0.14		
6. Sierra	4.13 ± 0.01	5.59 ± 0.02	7.27 ± 0.23	63.40 ± 0.01	23.91 ± 0.16	27.70 ± 0.12		
8. Perlmutter	1.77 ± 0.01	0.98 ± 0.00	4.24 ± 0.01	24.74 ± 0.00	14.74 ± 0.41			
19. Polaris	1.83 ± 0.00	1.32 ± 0.01	5.33 ± 0.02	23.71 ± 0.00	32.84 ± 0.30			
36. Lassen	4.56 ± 0.00	5.52 ± 0.01	7.76 ± 0.32	63.34 ± 0.02	24.56 ± 0.28	27.69 ± 0.10		
116. RZVernal	2.16 ± 0.01	0.12 ± 0.00	12.20 ± 0.07	24.88 ± 0.00	9.85 ± 0.01	12.58 ± 0.00	12.45 ± 0.02	10.21 ± 0.01
132. Tioga	2.15 ± 0.01	0.12 ± 0.00	12.19 ± 0.04	24.88 ± 0.00	9.85 ± 0.02	12.59 ± 0.01	12.46 ± 0.01	10.12 ± 0.02

Similar systems – configuration differences?

ACCELERATOR SUMMARY

Acc.	Memory BW (GB/s)	MPI Latency (us)	Kernel (us)		H2D / D2H		D2D
			Launch	Wait	Latency (us)	BW (GB/s)	Latency (us)
V100	786.43–861.40	18.10–18.72	4.13–4.84	4.31–5.59	7.27–7.82	44.88–63.40	23.91–24.97
A100	1362.75–1363.74	10.42–13.50	1.77–1.83	0.98–1.32	4.24–5.33	23.71–24.74	14.74–32.84
MI250X	1291.38–1336.81	0.44–0.50	1.51–2.16	0.12–0.14	12.19–12.91	24.87–24.88	9.85–12.02

CONCLUSION

- 2x-10x difference in key measures across contemporaries
  - event 50x across active systems (0.12us vs 5.59us kernel wait latency MI250X vs V100)
- Newer systems tend to be faster
  - Not across the board
- Intra-node topologies are complicated
  - Physical and/or manufacturing reasons
  - Does not always impact simple measurements
- Software and/or system configuration introduces significant differences in similar hardware
  - e.g. Memory Bandwidth on Theta and Trinity
  - Difficult to predict what your software will actually achieve

CURRENT GAPS AND FUTURE WORK

- Working on open-sourcing the benchmark scripts
- Not a *completely* representative sample
  - Focused on machines used by DOE
  - Missing interesting machines, e.g. Fugaku (Top500 #2, RIKEN)
- No inter-node measurements
  - Very important to application performance
  - Has all the complexity of intra-node measures, combined with network performance
    - Actually increases intra-node complexity too – e.g. GPU-to-NIC mapping
  - Thinking about a digestible set of key inter-node measures
- Cloud
  - Cloud-native HPC clusters, or spillover
  - Intra-node measurement approach could be similar
  - Maybe additional challenges for inter-node measures
- Enough interest to regenerate results and iterate on the approach every year?

ACKNOWLEDGEMENTS

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-NA-0003525. 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 **Argonne Leadership Computing Facility**, a DOE Office of Science User Facility supported under Contract No. DE-AC02-06CH11357. This research used resources of the **Los Alamos National Laboratory**, supported by the US Department of Energy under contract No 89233218CNA000001. This research used resources of the **Lawrence Livermore National Laboratory**, supported by the US Department of Energy under contract DE-AC52-07NA27344. 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.

Sandia National Laboratories

Exceptional service in the national interest

LATENCY AND BANDWIDTH MICROBENCHMARKS OF US DEPARTMENT OF ENERGY SYSTEMS IN THE JUNE 2023 TOP500 LIST

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

Sandia National Laboratories [1]Oak Ridge National Laboratory

14th IEEE Intl. Workshop on PMBS

November 13th, 2023

Questions?

cwpears@sandia.gov

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.

SAND2023-13518C