Data and Computational Science driven by Persistent Memory Supercomputer Pegasus

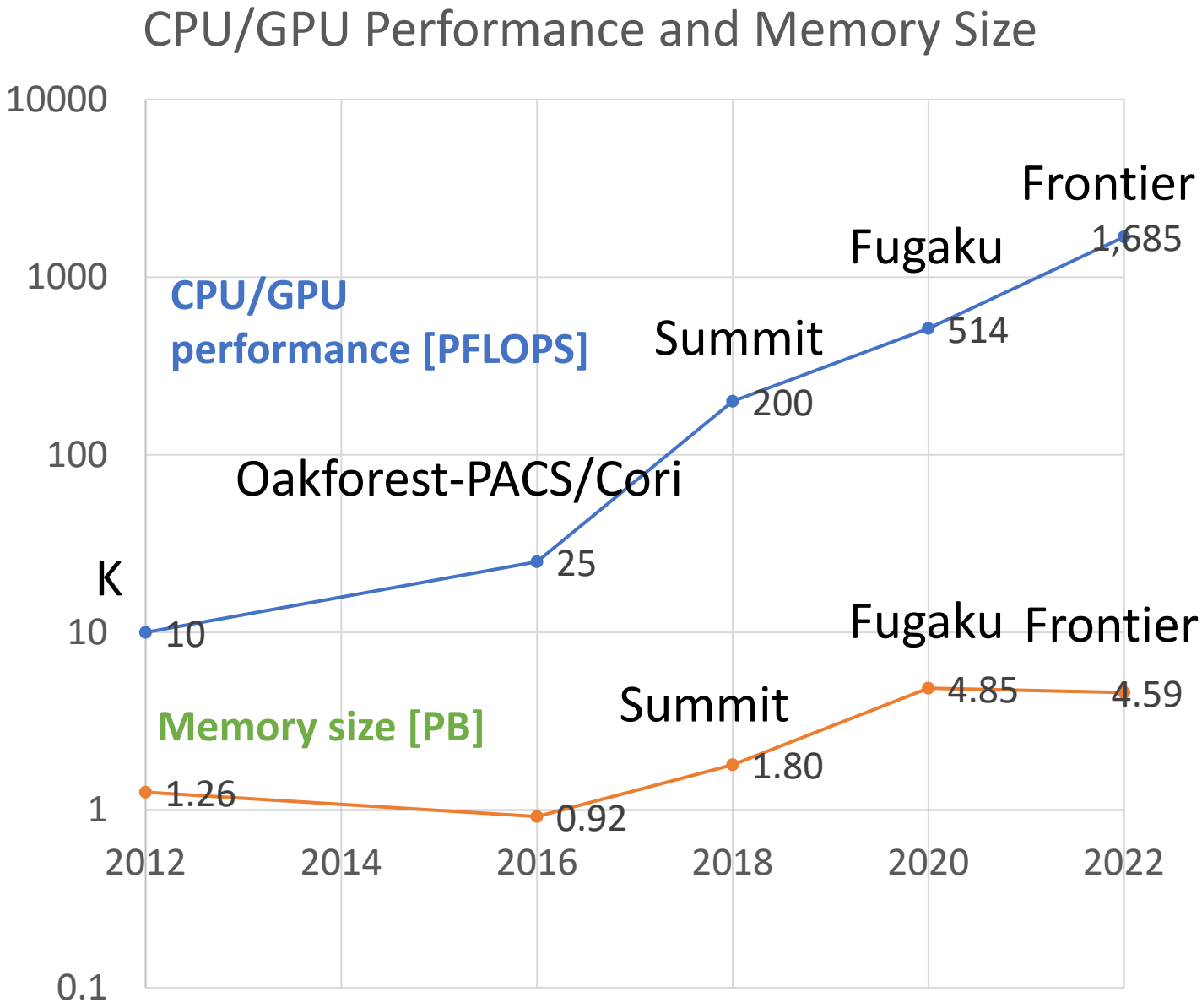
Osamu Tatebe

Center for Computational Sciences, University of Tsukuba

Pegasus background

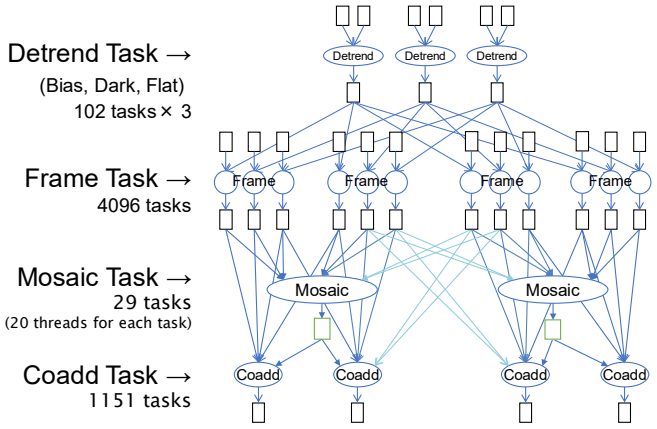
- CPU performance 170x, but memory size 3.6x in 10 years
- It matters for Data-driven and AI-driven Science
  - Memory size and Storage performance are important

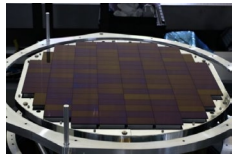
- Introduce Persistent Memory
  - Low power and cost effective
  - Big memory space and high-performance storage



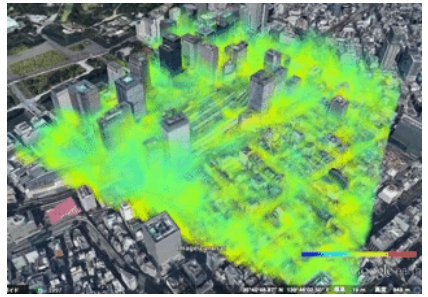
Design Goal of Pegasus

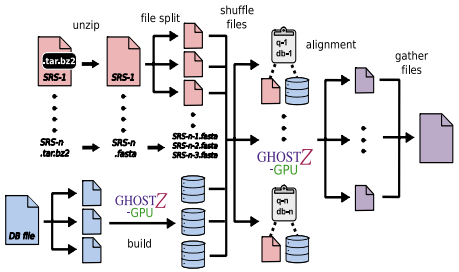
- Accelerates large-scale data analysis and big data AI by utilizing persistent memory for large memory space and high-performance storage
- Fosters new fields of large-scale data analysis, new applications of big data AI, and system software research

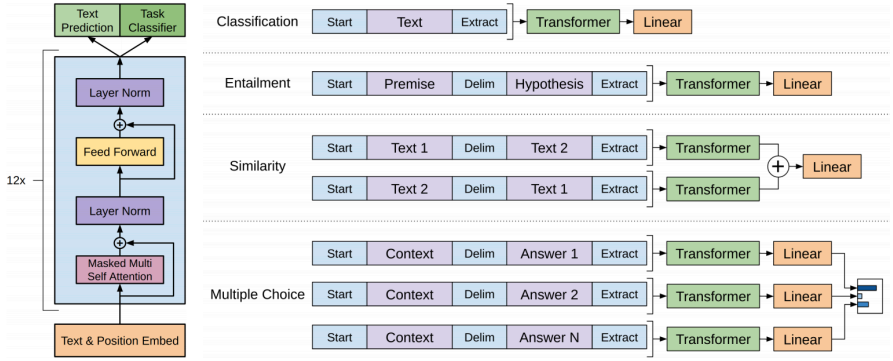




© NAOJ



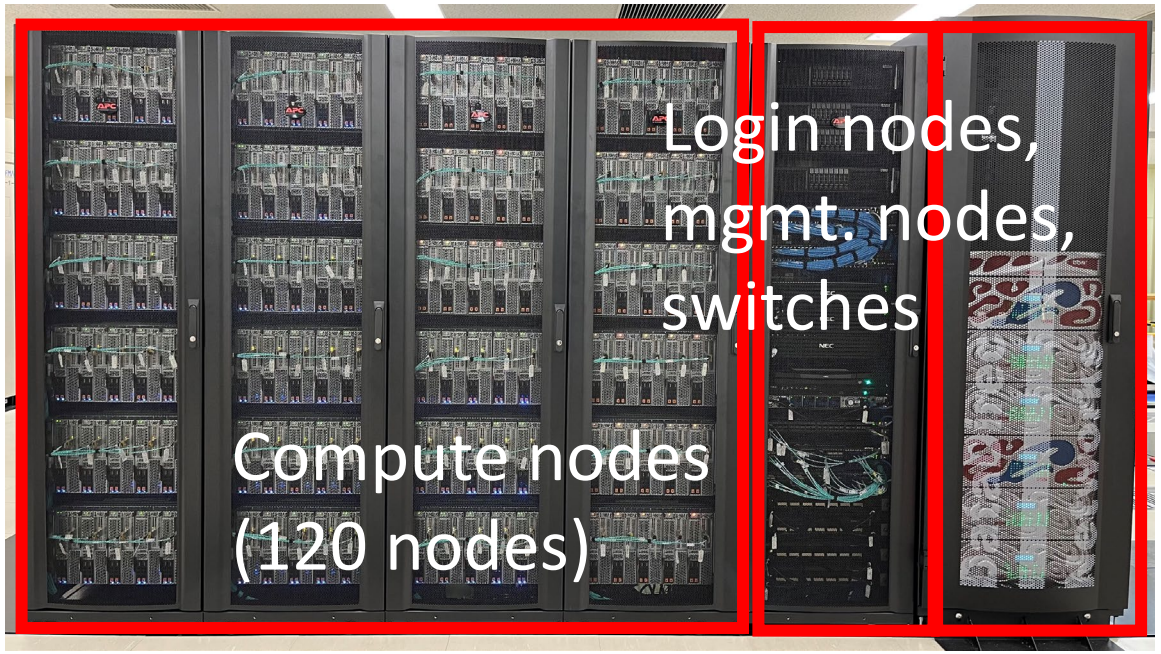




https://paperswithcode.com/method/gpt

Pegasus Highlights

- Build with 4th Gen Intel Xeon (SPR), NVIDIA H100 Tensor Core PCIe GPU, and Intel Optane persistent memory 300, which will strongly drive Big Data and AI



Parallel File System

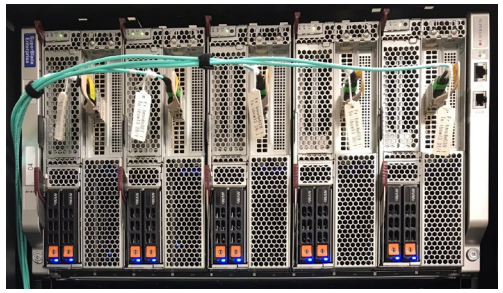
Design of Pegasus

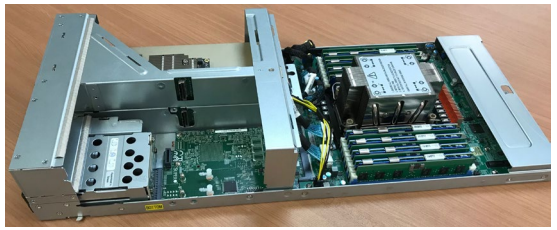
- Improve computational performance
  - Intel 4th Gen Xeon (SPR), NVIDIA H100 PCIe GPU
- Improve bandwidth of GPU memory, CPU memory, and I/O
  - HBM2E, DDR5, PCIe Gen5
  - Persistent Memory is installed in DDR5 DIMM slots
- Increase memory size
  - 2TiB + 128GiB available on all nodes
- Improve local storage performance using Persistent Memory and Node-local SSD
  - R&D of parallel caching file system

Pegasus Specification

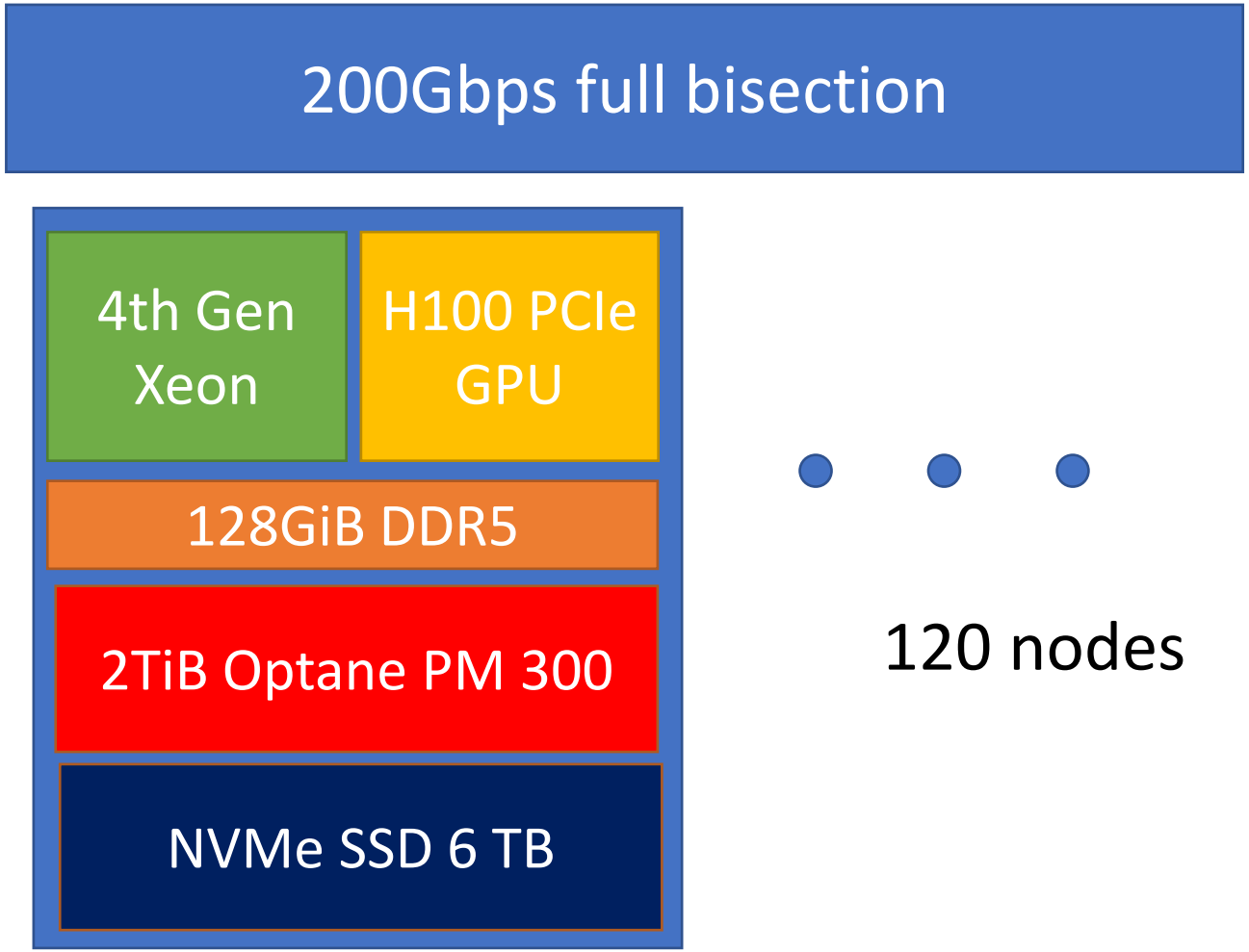
- Installed in Q4 2022
- Total Performance
  - 120 nodes, 6.5 PFlops, 240 TiB Pmem
- Node specification
  - 3.2 TFlops Intel Platinum 8468 (Sapphire Rapids)
  - 51 TFlops NVIDIA H100 PCIe GPU
  - 128 GiB DDR5 DRAM (282 GB/s)
  - 2 TiB Optane PM 300 series (Crow Pass)
  - 6 TB NVMe SSD (7 GB/s)
- Interconnection Network
  - NVIDIA Quantum-2 InfiniBand platform (200 Gbps) full bisection (InfinBand NDR200)
- Parallel File System
  - 7.1 PByte DDN EXAScaler (40 GB/s)

NEC LX B1000E Blade Enclosure

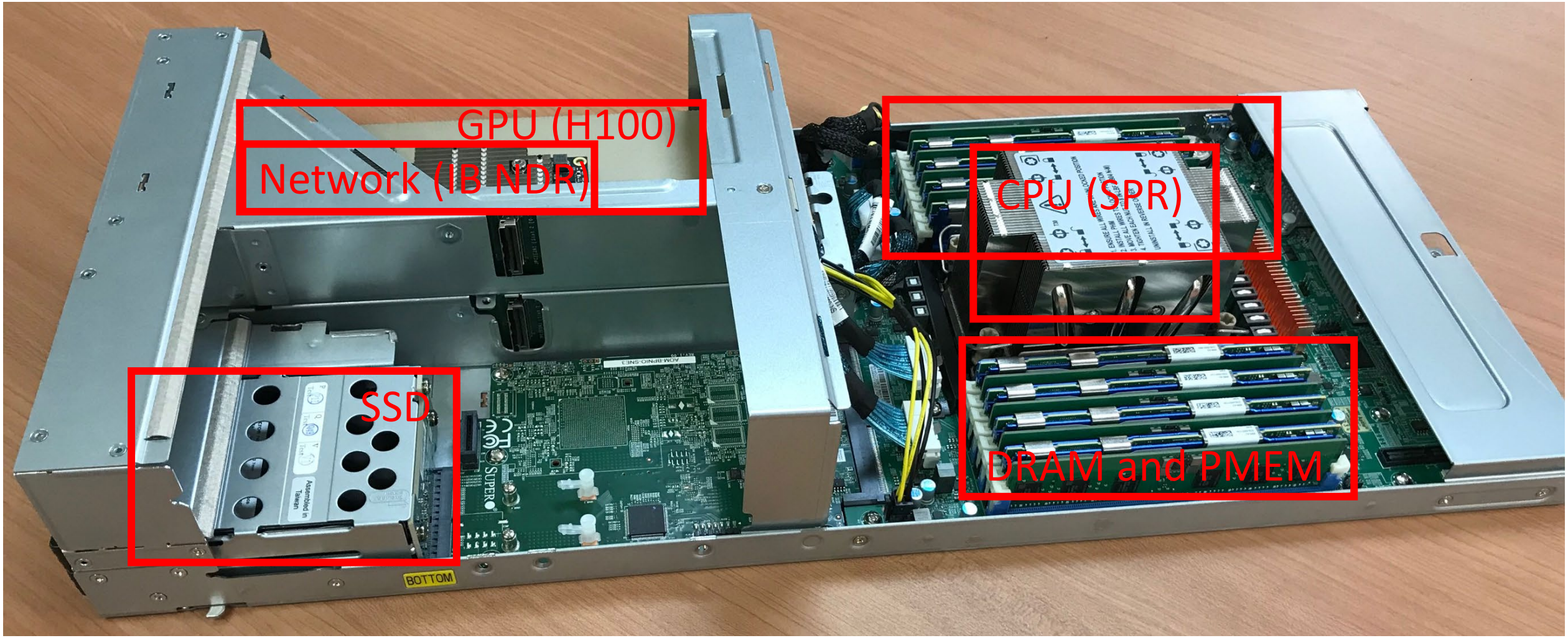




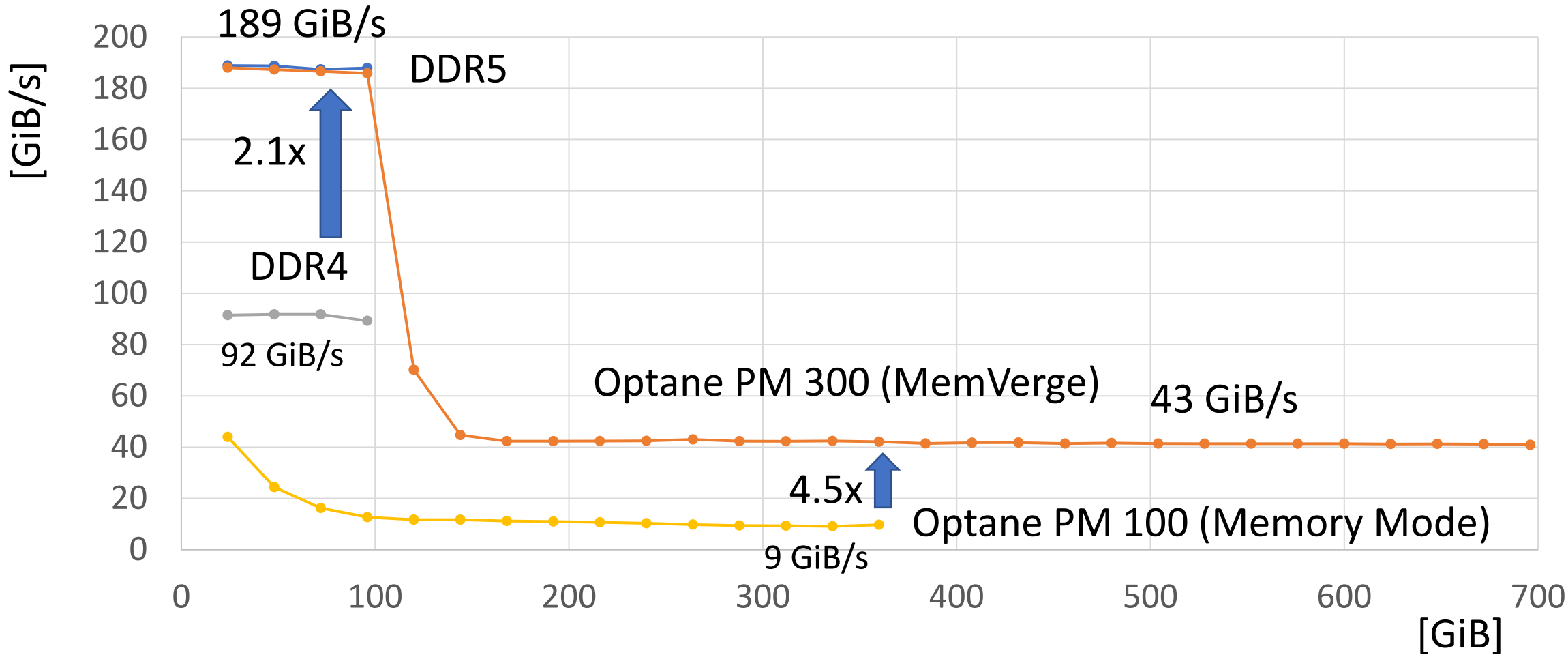
NEC LX 102Bk-6



Compute node (Blade)

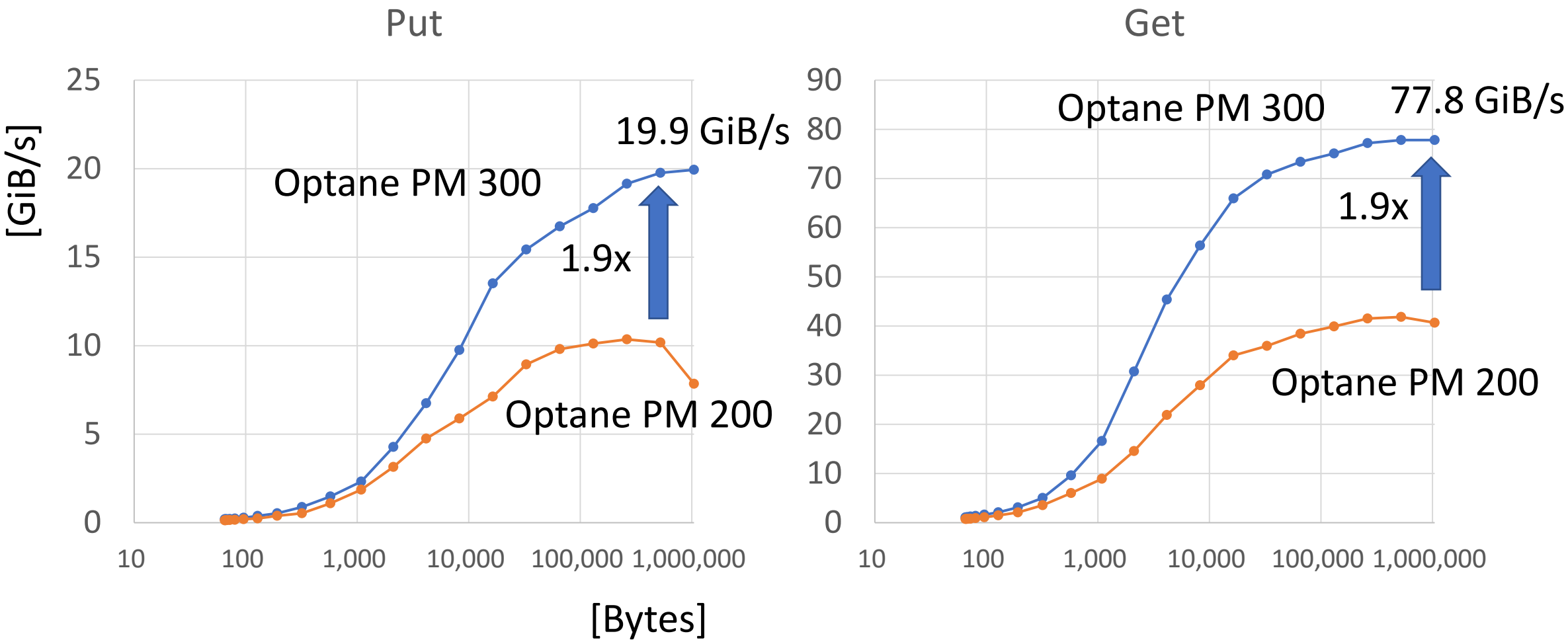


Stream Benchmark (Triad; a[] = b[] + s * c[])



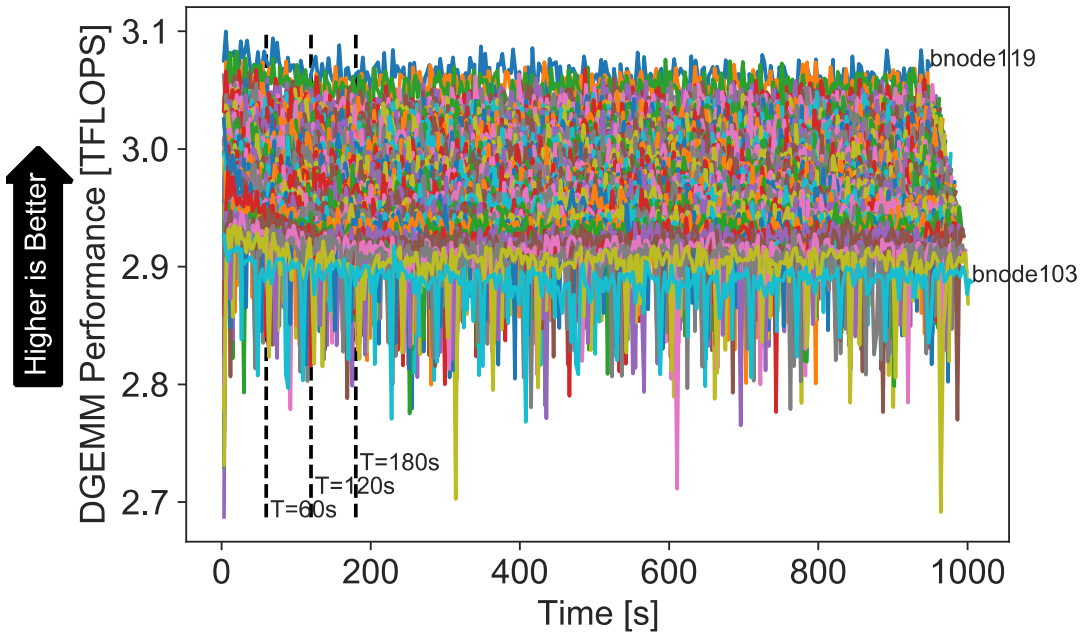
Special thanks to Akira Nukada

Persistent in-memory KVS Bandwidth (pmemkv/devdax)



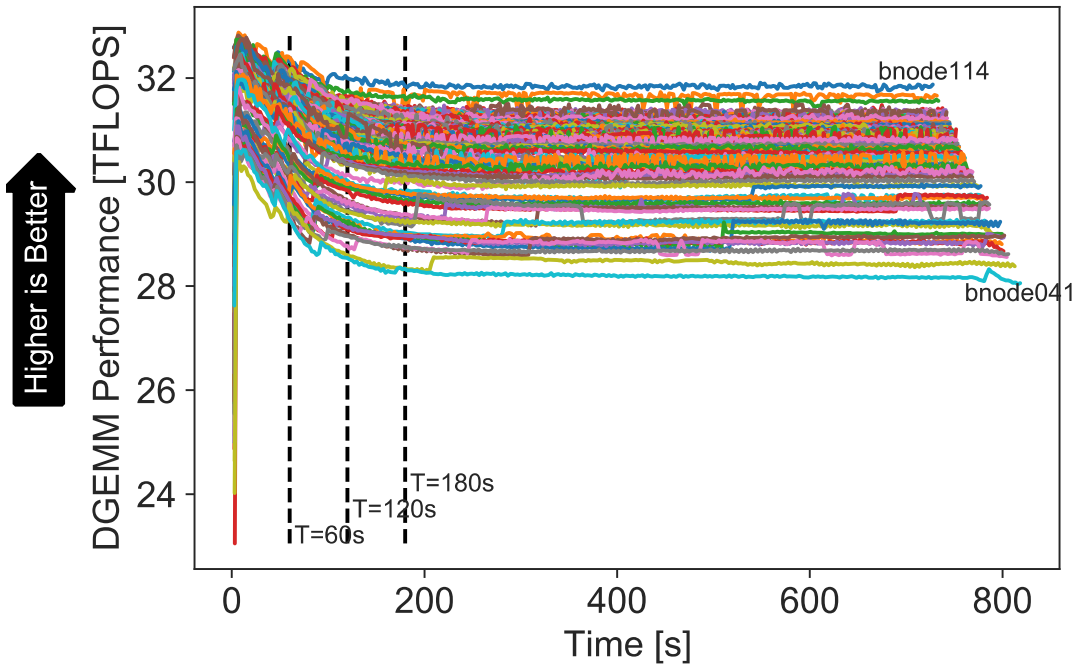
DGEMM (1) (C = a * A * B + b * C)

CPU (Sapphire Rapids), oneMKL



There is not so much turbo boost effect

GPU (H100), cuBLAS

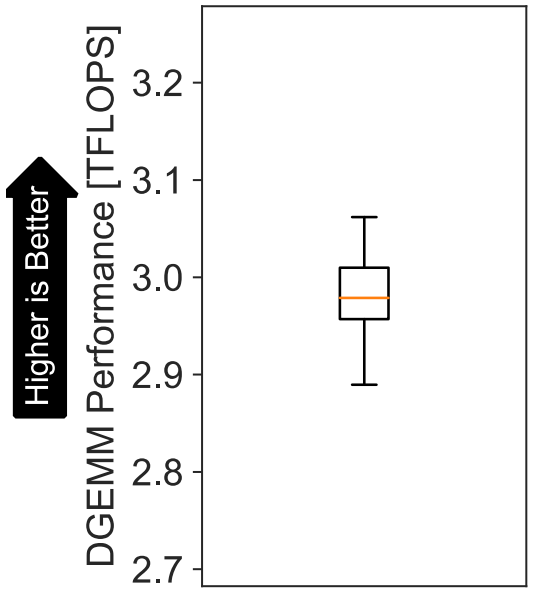


Early stage performs better due to turbo boost
After 180 seconds, the performance would be stable

Special thanks to Norihisa Fujita

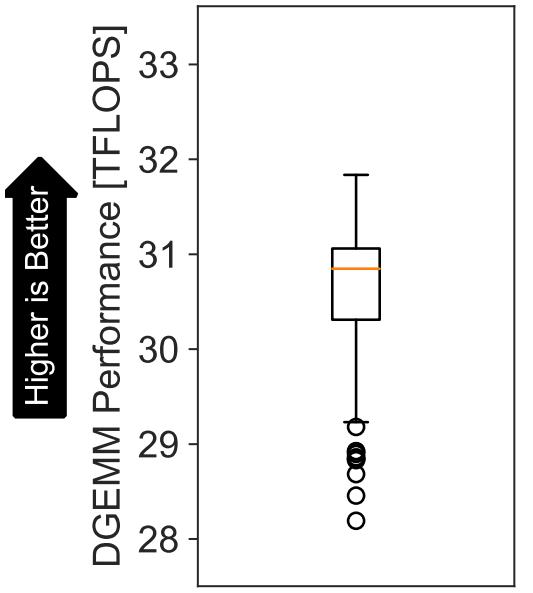
DGEMM (2)

CPU (Sapphire Rapids)



Median 2.98
Ave 2.98
93% of peak

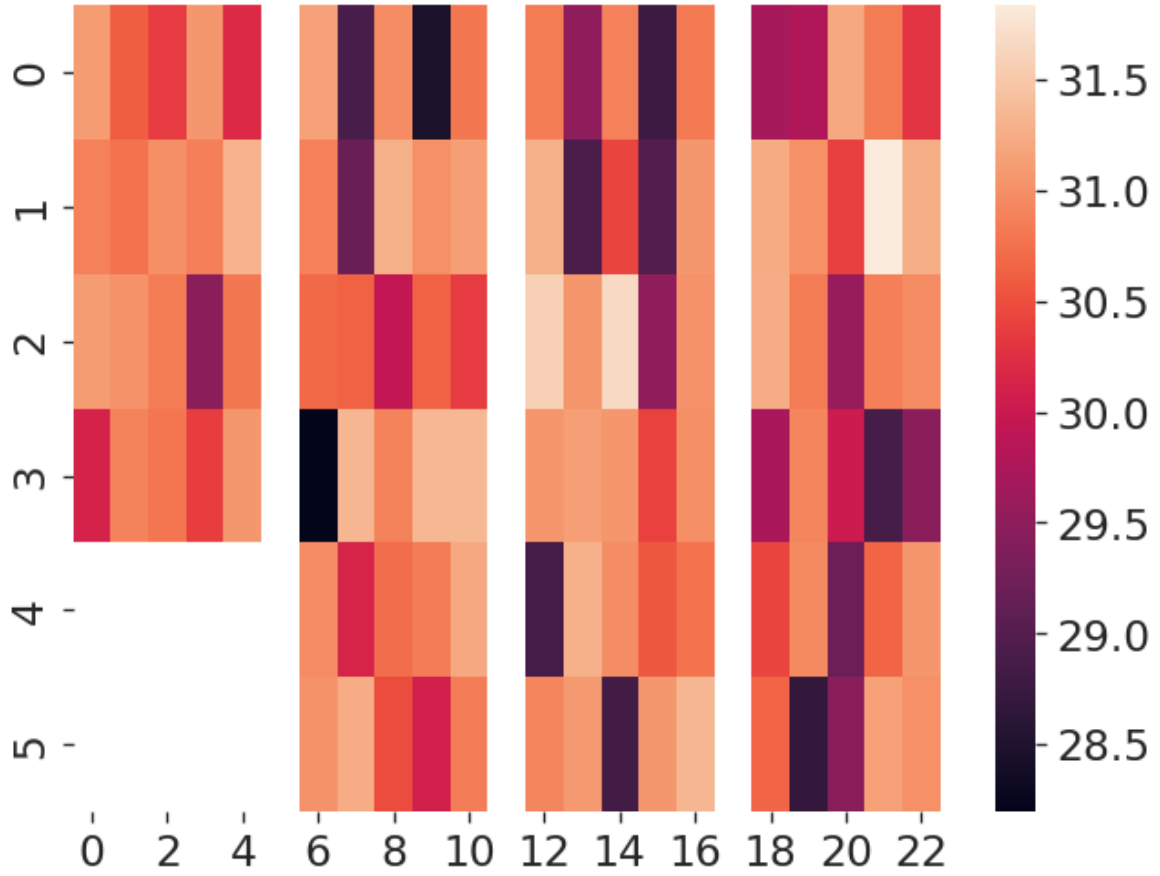
GPU (H100)



Median 30.85
Ave 30.56
60% of peak
15% performance difference

Special thanks to Norihisa Fujita

DGEMM GPU performance per server location



No significant relationship between GPU performance and server location

Special thanks to Kohei Hiraga

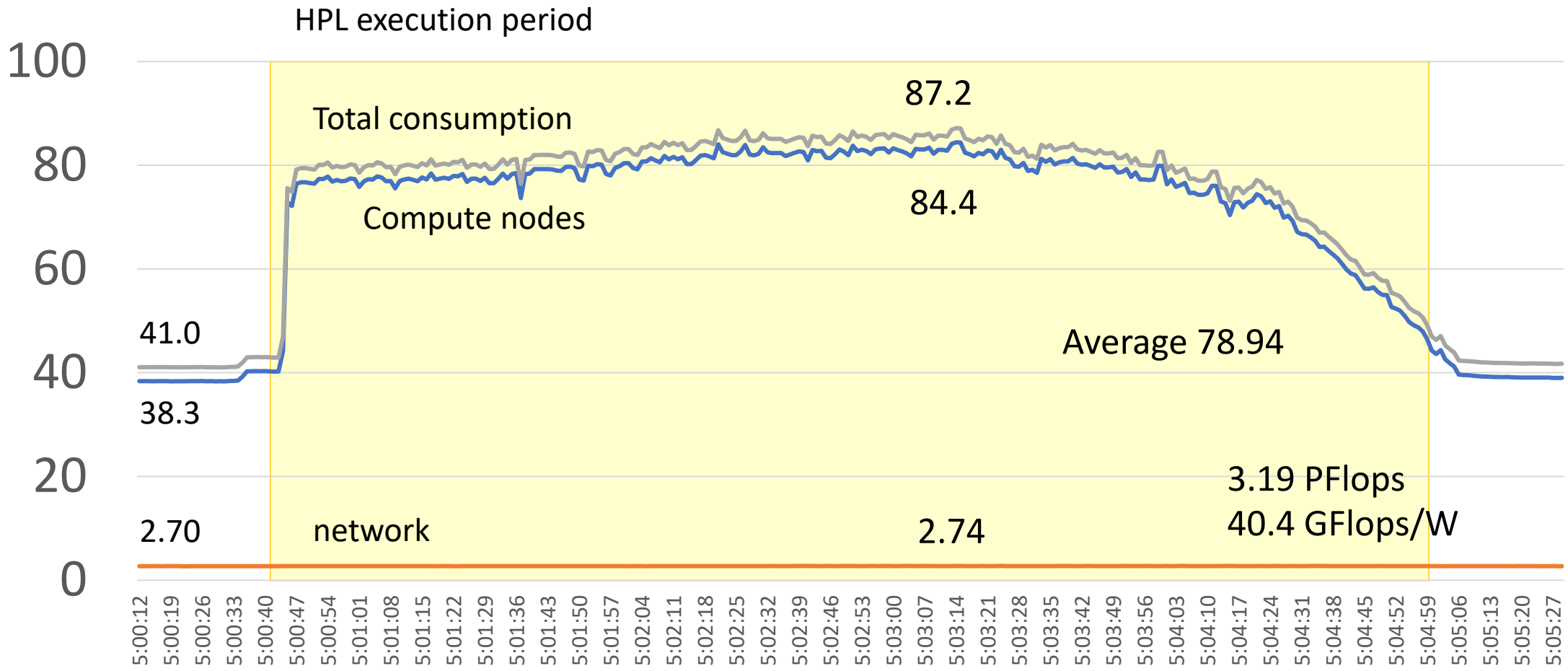
Green500

- #12 in ISC23 list
- Would be better in SC23

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
11	400	**MN-3** - MN-Core Server, Xeon Platinum 8260M 24C 2.4GHz, Preferred Networks MN-Core, MN-Core DirectConnect, Preferred Networks Preferred Networks Japan	1,664	2.18	53	40.901
12	190	**Pegasus** - NEC LX 102Bk-6, Xeon Platinum 8468 48C 2.1GHz, NVIDIA H100 80GB PCIe, Infiniband NDR, NEC Center for Computational Sciences, University of Tsukuba Japan	13,680	3.48	79	40.448
13	370	**Champollion** - Apollo 6500, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 80 GB, Mellanox HDR Infiniband, HPE Hewlett Packard Enterprise France	19,840	2.32	60	38.555
14	387	**SSC-21 Scalable Module** -	16,704	2.27	103	33.983

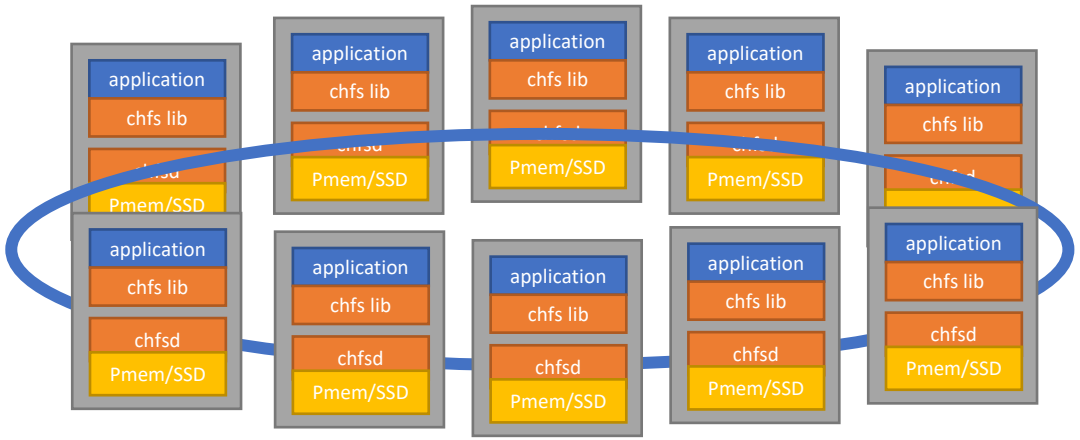
3.19 PFLOPS / 78.94 kW

Green500 Power Consumption [kW]

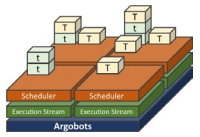


CHFS/Cache research and development

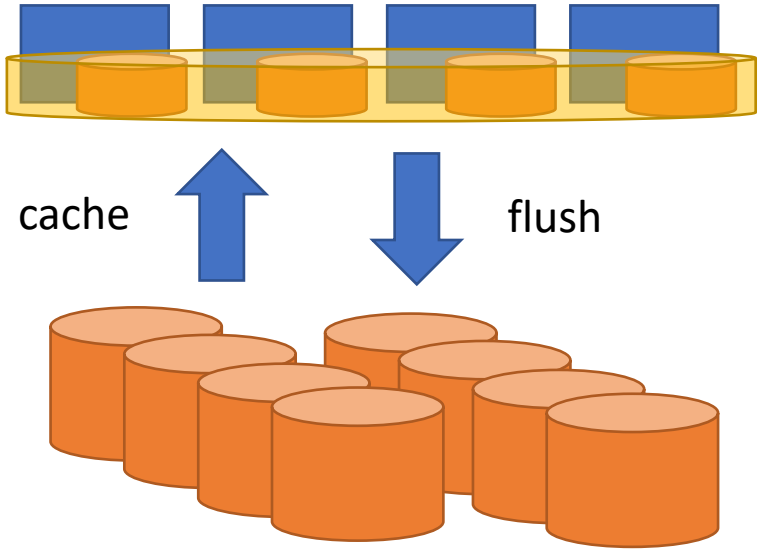
- Caching file system utilizing node-local persistent memory and storage
- Improve metadata performance by relaxing consistency with backend PFS in easy-to-understand semantics
- https://github.com/otatebe/chfs



Mochi

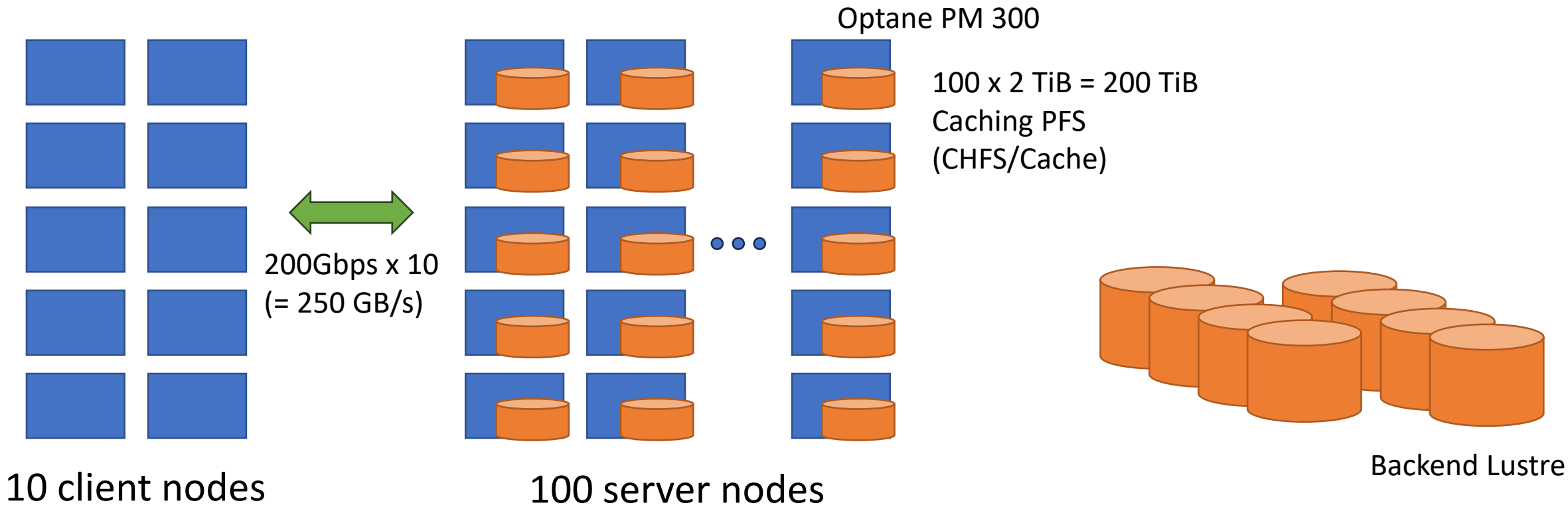


pmem.io

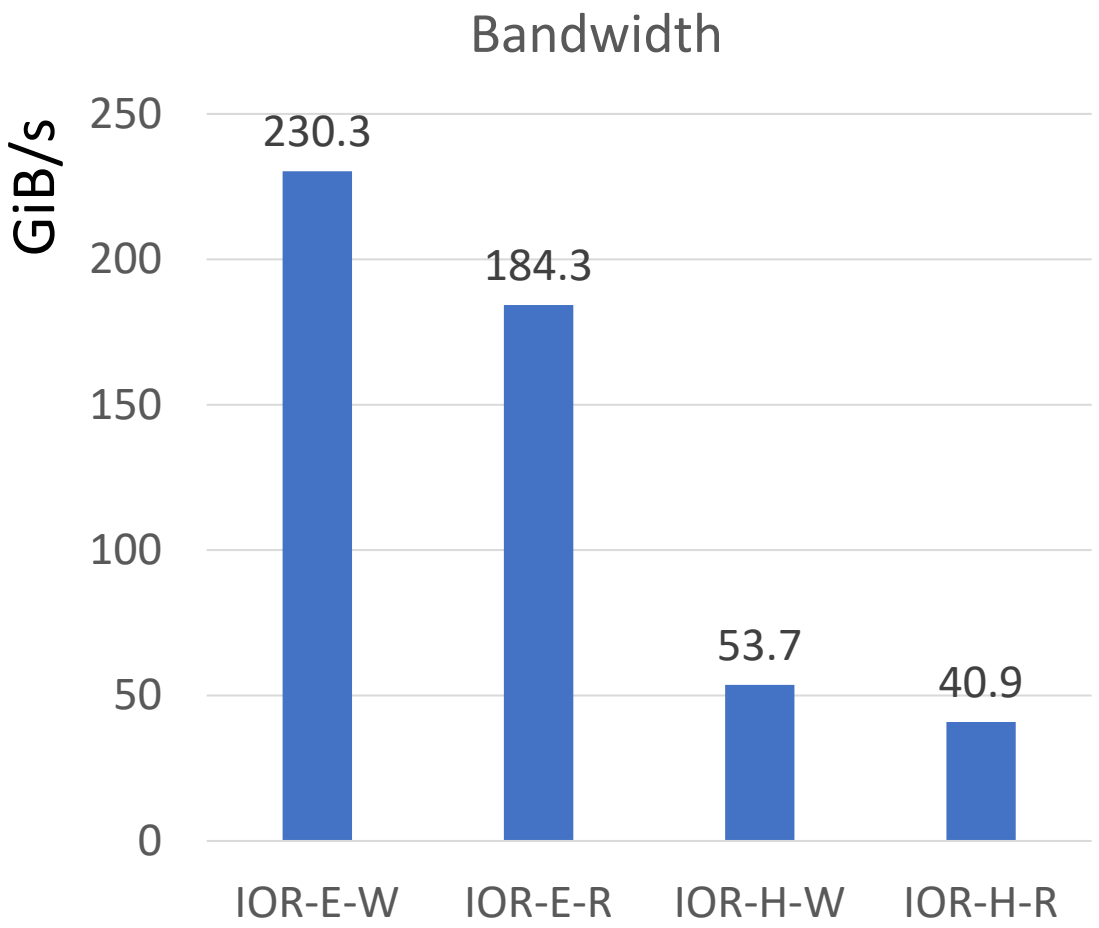


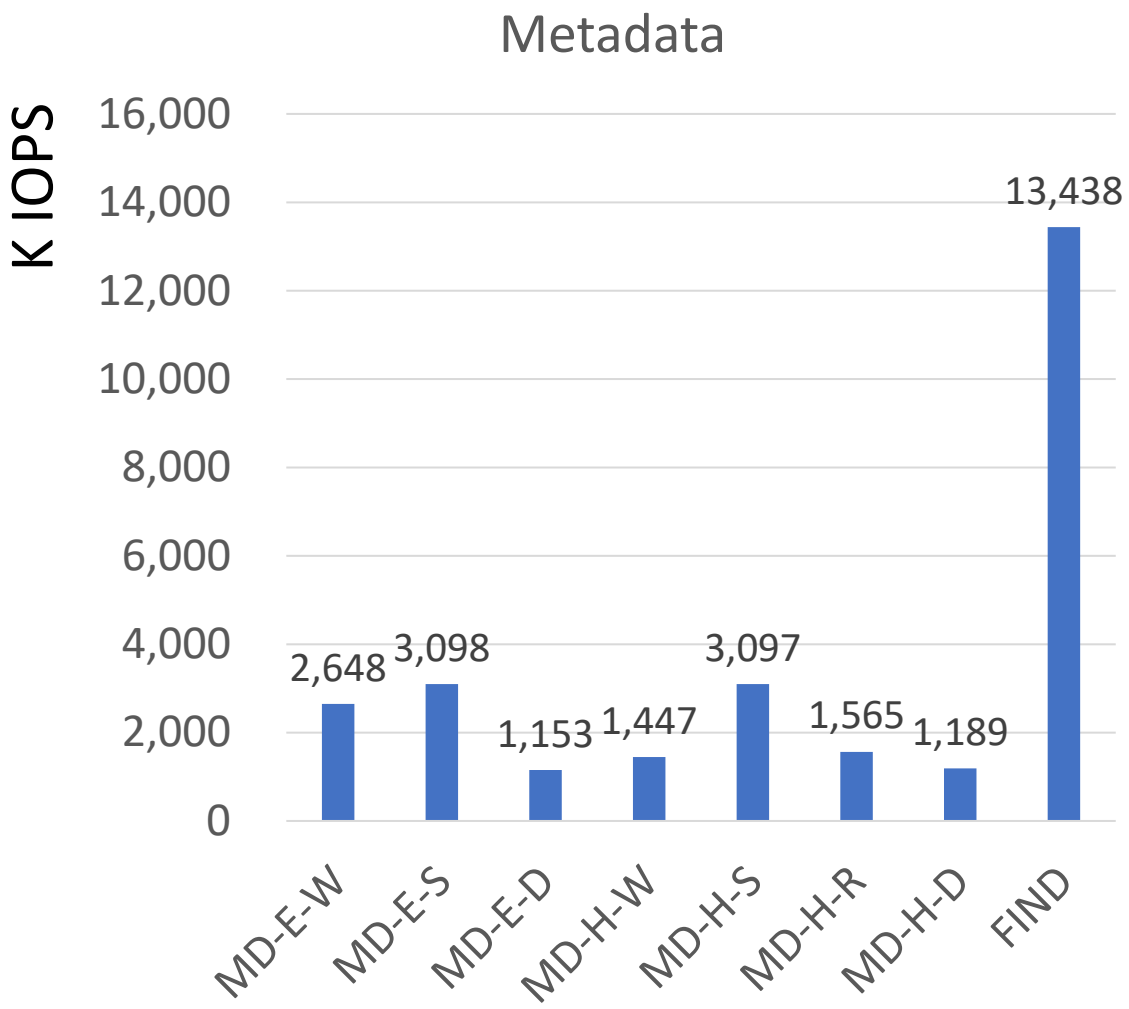
IO500 Challenge

- Top storage ranking by IO500 benchmark
- I/O Bandwidth, Metadata performance, Find performance



IO500 ISC23 Score





10	ISC23	LRZ	SuperMUC-NG-Phase2-10	Lenovo	DAOS	10	1,120	1,533.28	266.73	8,813.96
11	SC22	Cloudam HPC on OCI	HPC-OCI	Cloudam	BurstFS	10	720	1,285.21	95.90	17,224.05
12	ISC22	LRZ	SuperMUC-NG Phase2	Lenovo	DAOS	10	640	1,034.55	209.48	5,109.23
13	SC22	Meadowgate Technologies	Meadowgate	INTEL HPE	DAOS	10	1,280	1,014.24	213.15	4,826.12
14	SC21	BPFS Lab	Kongming		BPFS	10	800	972.60	96.26	9,827.09
15	ISC22	University of Cambridge	Cumulus	Dell/Intel	DAOS	10	2,000	963.00	216.78	4,277.86
16	ISC20	Intel	Wolf	Intel	DAOS	10	420	758.71	164.77	3,493.56
17	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960	612.87	105.28	3,567.85
18	ISC22	Lenovo	Lenovo-Lenox3	Lenovo	DAOS	10	720	544.18	115.94	2,554.14
19	SC21	National Research Center of Tranlational Medicine at Shanghai Ruijin Hospital	ASTRA	NRCTM	DAOS	10	360	511.02	87.50	2,984.61
20	ISC20	TACC	Frontera	Intel	DAOS	10	420	508.88	79.16	3,271.49
21	ISC23	University of Tsukuba	Pegasus	OSS	CHFS	10	480	484.41	98.24	2,388.60
22	ISC21	National Supercomputer Center in GuangZhou	Venus2	National Supercomputer Center in GuangZhou	kapok	10	480	474.10	91.64	2,452.87
23	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380	440.64	95.80	2,026.80

21	ISC23	University of Tsukuba	Pegasus	OSS	CHFS	10	480	484.41	98.24	2,388.60
22	ISC21	National Supercomputer Center in GuangZhou	Venus2	National Supercomputer Center in GuangZhou	kapok	10	480	474.10	91.64	2,452.87
23	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380	440.64	95.80	2,026.80

#21 in 10 node list

#28 in full list

Summary

- Pegasus was introduced in Q4 2022
  - Big memory and high-performance storage for data-driven and AI-driven science
- Stream Triad performance is 189 GB/s in DDR5 (2.1x better than DDR4) and 41 GB/s in Pmem (4.5x better than Optane 100)
- Pmemkv performance is 20 GiB/s for PUT and 78 GiB/s for GET (2.0x better than Optane 200)
- DGEMM performance is 31 TFlops in H100 (60% of peak), and 3 TFlops in SPR (93% of peak)
- HPL performance is 3.48 PFlops for performance, and 3.19 PFlops and 40.4 GFlops/watts for energy efficiency
  - #12 in ISC23 Green500 list, #190 in TOP500
- #21 in ISC23 IO500 10-node list, #28 in full list