JCAHPC

University of Tsukuba
Center for Computational Sciences
筑波大学　計算科学研究センター

Storage System of the Oakforest-PACS Supercomputer

Osamu Tatebe

Center for Computational Sciences

University of Tsukuba

Oakforest-PACS (OFP) [Since 2016.12]

- Compute nodes
  - 25 PFLOPS (**Xeon Phi KNL**)
  - 8,208 nodes
  - 897 TiByte memory
  - 100 Gbps **Omni-Path**, fat tree, full bisection bandwidth
- Parallel File System (Lustre)
  - 26.2 PByte
  - 500 GByte/sec (physical peak)
  - 4 MDS x 3 sets, 10 ES14KX
- File Cache System (aka Burst Buffer)
  - 940 TByte NVMe SSD
  - 1.56 TByte/sec (physical peak)
    - All memory dump takes 10 minutes
  - 25 **IME14K with Omni-Path**

Parallel File System for OFP

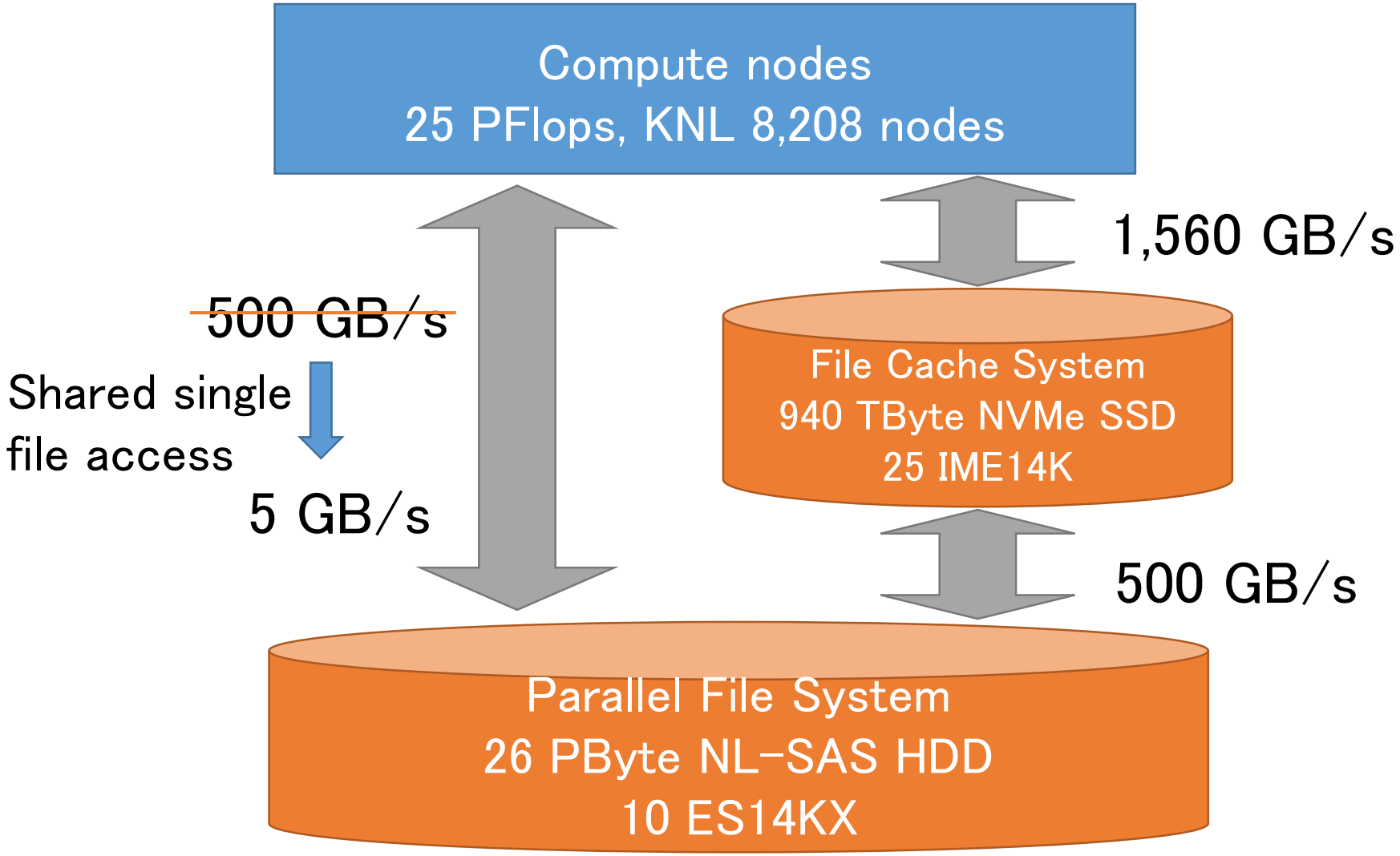
- Lustre File System
  - 26.2 PByte
  - 500 GByte/sec (physical peak)
    - 100 Gbps x 4 ports x 10 nodes

- 3 set of 4 MDS
- 10 ES14KX

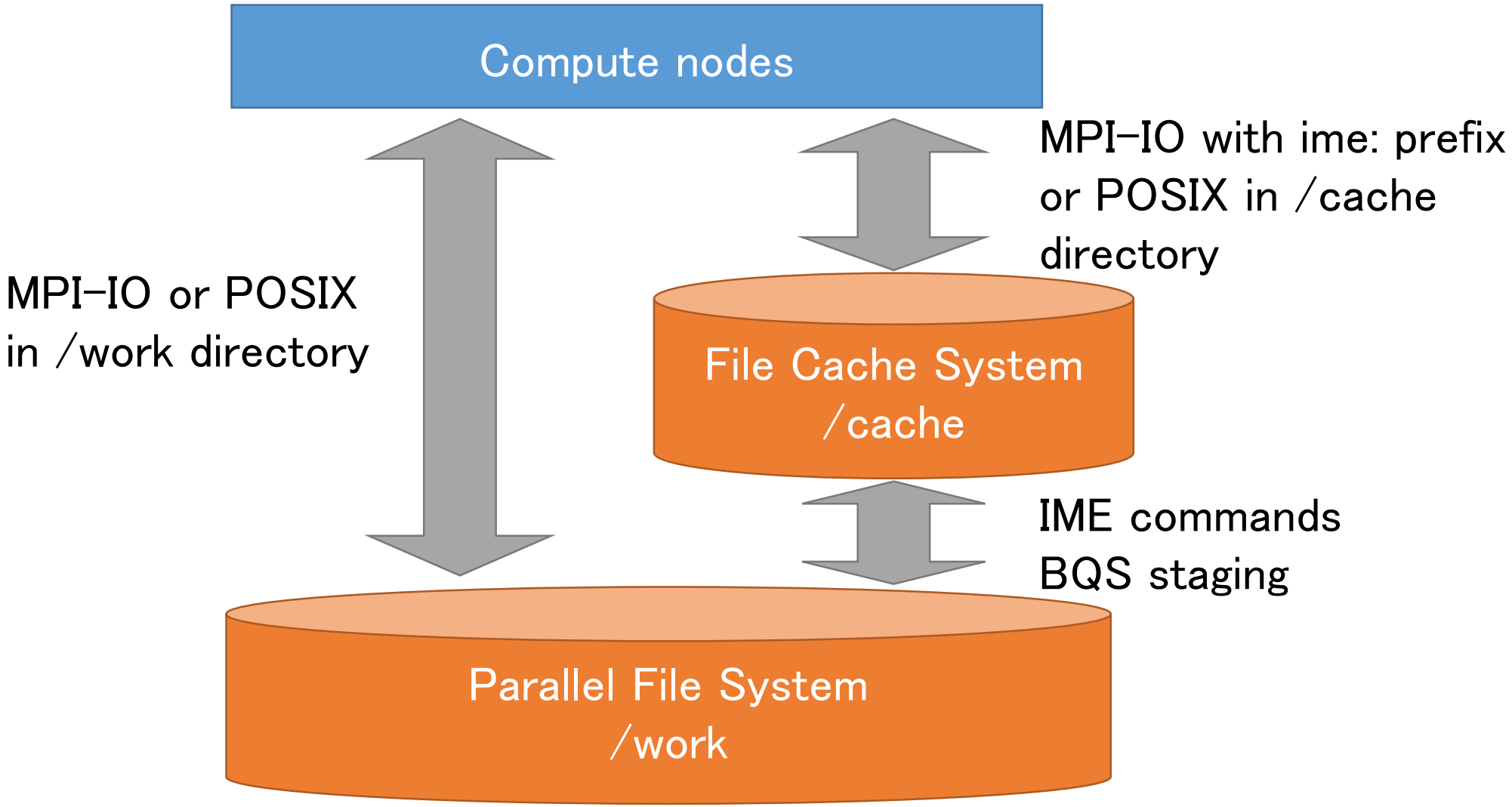
File Cache System (Burst Buffer)

- Infinite Memory Engine (IME)
  - 940 TByte NVMe SSD
  - 1.56 TByte/sec (physical peak)
    - 1.3 GB/s x 48 SSD x 25 nodes
  - 25 IME14K

Storage System for OFP



Access to File System Cache



/cache and /work (PFS) has the same namespace

IO-500 Benchmark (1)

- IO Issues in Supercomputers
  - Not peak storage performance, but real performance by applications
  - Lower priority than CPU and network in supercomputers
  - Encourage R&D for hard access patterns
- IO Benchmarks and TOP 500 list
  - BoF at SC16, ISC17
  - First list announced at SC17
  - Julian Kunkel (DKRZ), Jay Lofstead (Sandia), John Bent (Seagate)
- Mixed benchmarks for bandwidth and metadata performance in HPC typical access patterns

IO500

November 2017
io500.org

#	information				io500		
	system	institution	filesystem	client nodes	score	bw	md
						GiB/s	kIOP/s
1	Oakforest-PACS	JCAHPC	IME	2048	101.48	471.25	21.85
2	Shaheen	Kaust	DataWarp	300	70.90	151.53	33.17
3	Shaheen	Kaust	Lustre	1000	41.00	54.17	31.03
4	JURON	JSC	BeeGFS	8	35.77	14.24	89.83
5	Mistral	DKRZ	Lustre	100	32.15	22.77	45.39
6	Sonasad	IBM	Spectrum Scale	10	21.63	4.57	102.38
7	Seislab	Fraunhofer	BeeGFS	24	18.75	5.13	68.58
8	EMSL Cascade	PNNL	Lustre	126	11.17	4.88	25.57
9	Serrano	SNL	Spectrum Scale	16	4.25	0.65	27.98

IO-500 Benchmark (2)

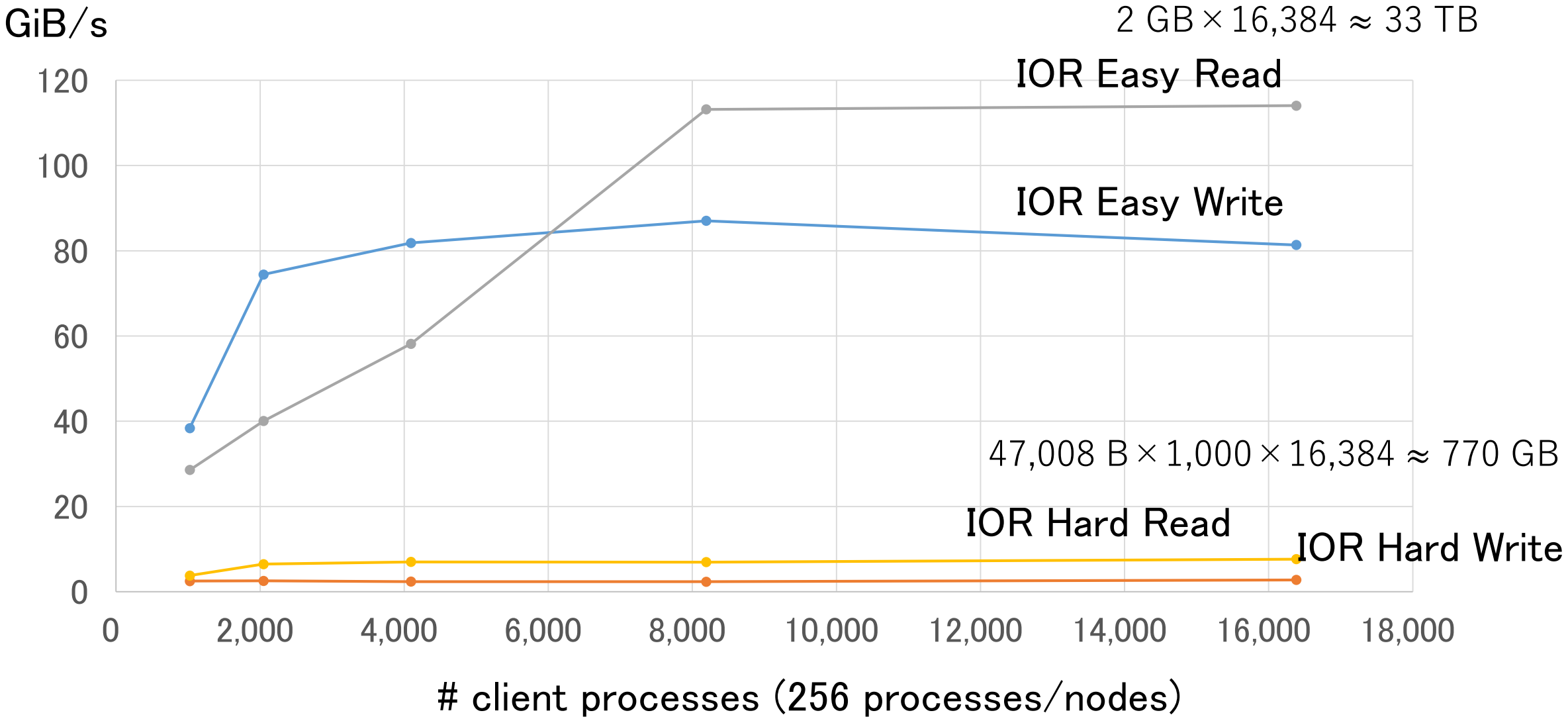
- Bandwidth – 4 test cases
  - File per process write/read [IOR Easy, BW1/3]
  - Single shared file write/read of 47,008B block size [IOR Hard, BW2/4]
- Metadata performance – 8 test cases
  - 0byte file creations/stats/deletions in unique directories [MDT Easy, IOPS1/4/6]
  - 3,901B file creations/stats/reads/deletions in a shared directory [MDT Hard, IOPS2/5/7/8]
  - Find [IOPS3]
- File size and # files are decided so that file writes and creations takes more than 5 minutes
  - Reduce the buffer cache effect
  - (Rule of thumb shows all memory dump will take 5 to 10 minutes)

IO-500 Benchmark Score

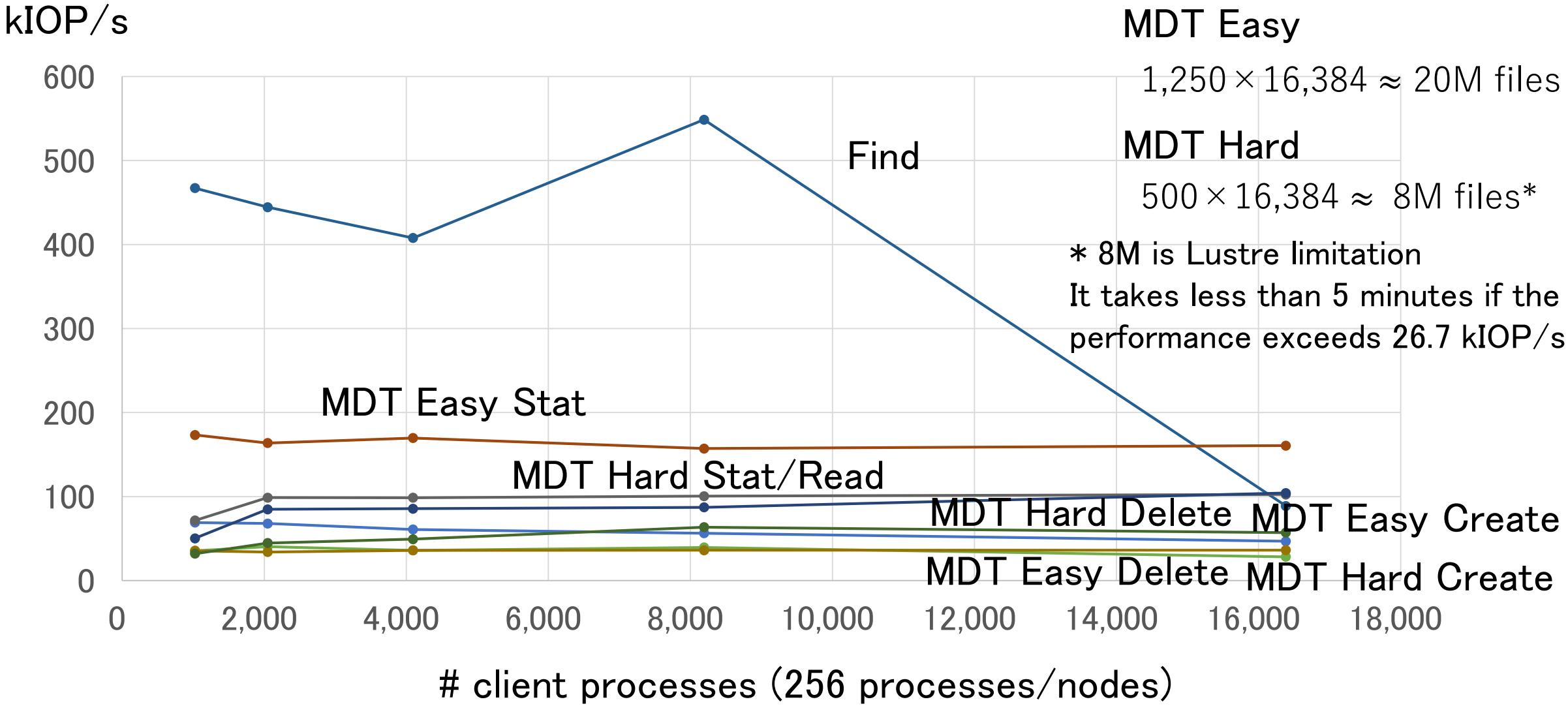
- Geometric mean of geometric mean of bandwidths and geometric mean of metadata performances

$$\sqrt{\sqrt[4]{BW1 \times BW2 \times BW3 \times BW4} \times \sqrt[8]{IOPS1 \times IOPS2 \times \cdots \times IOPS8}}$$

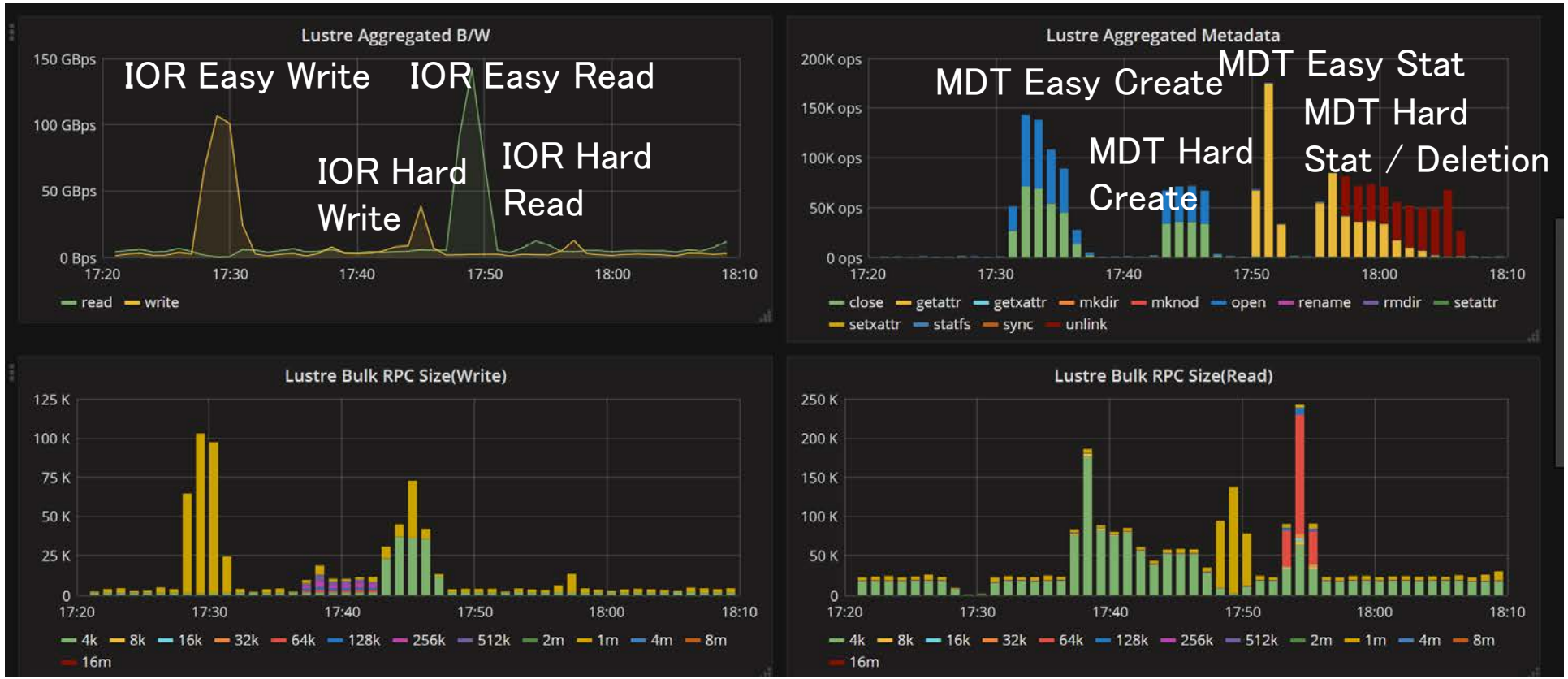
Lustre IO-500 Bandwidth



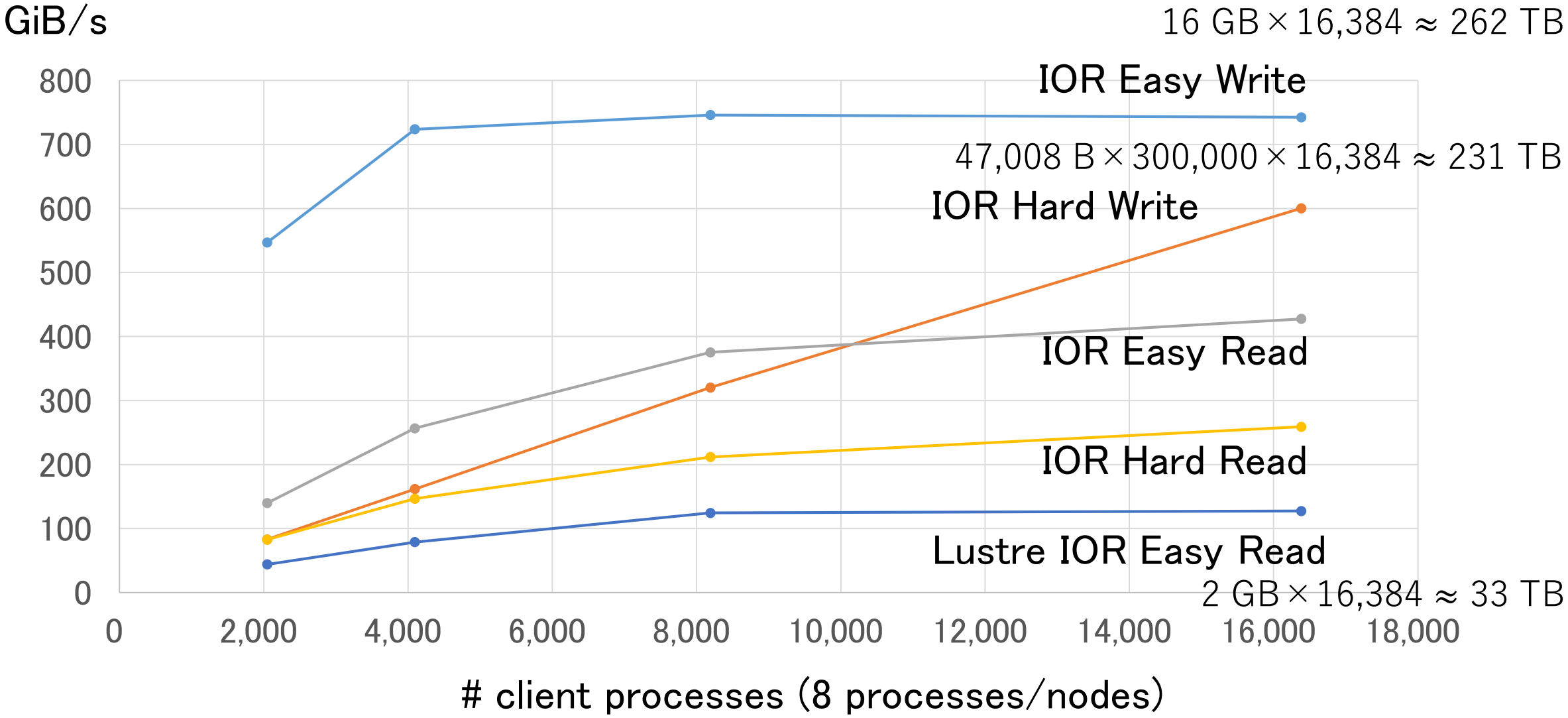
Lustre IO-500 Metadata



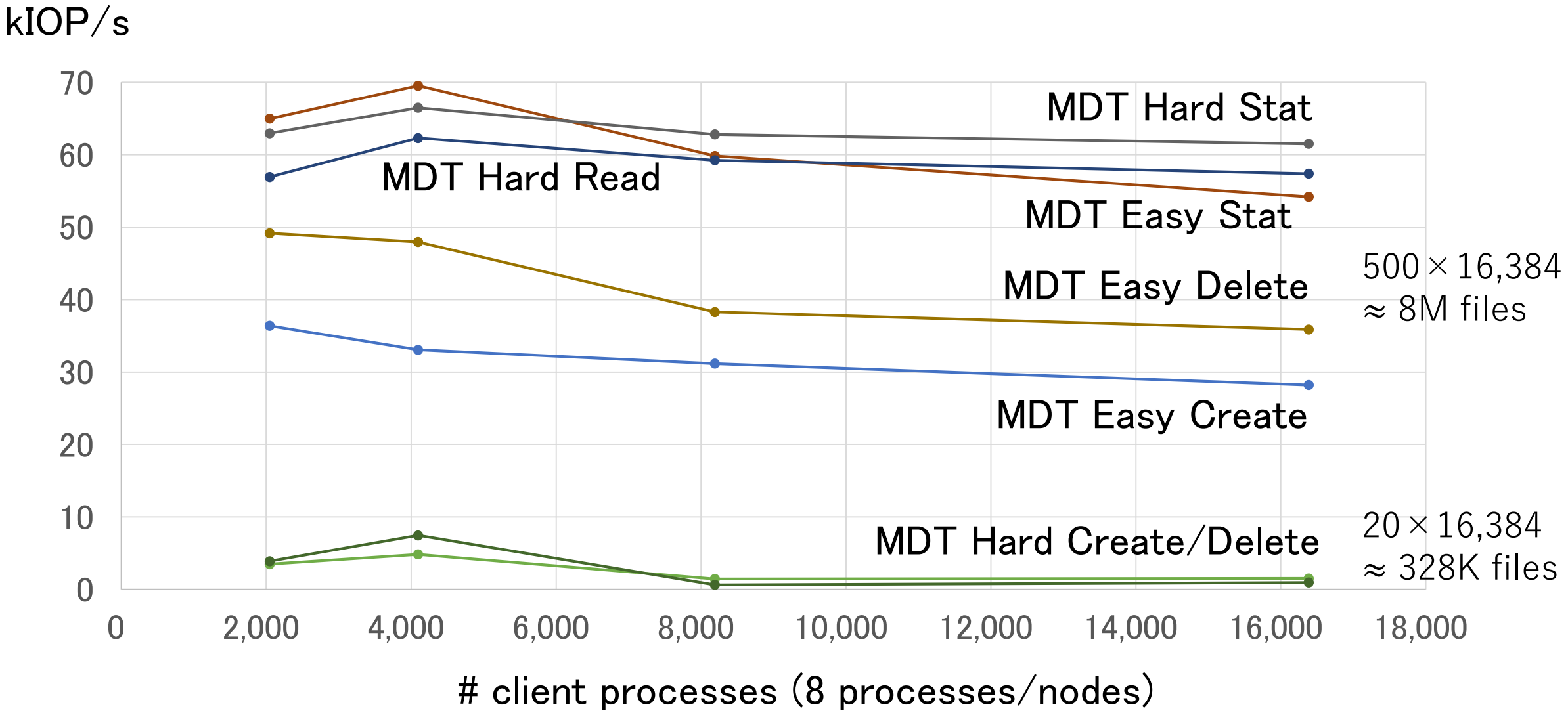
Lustre Performance Monitor



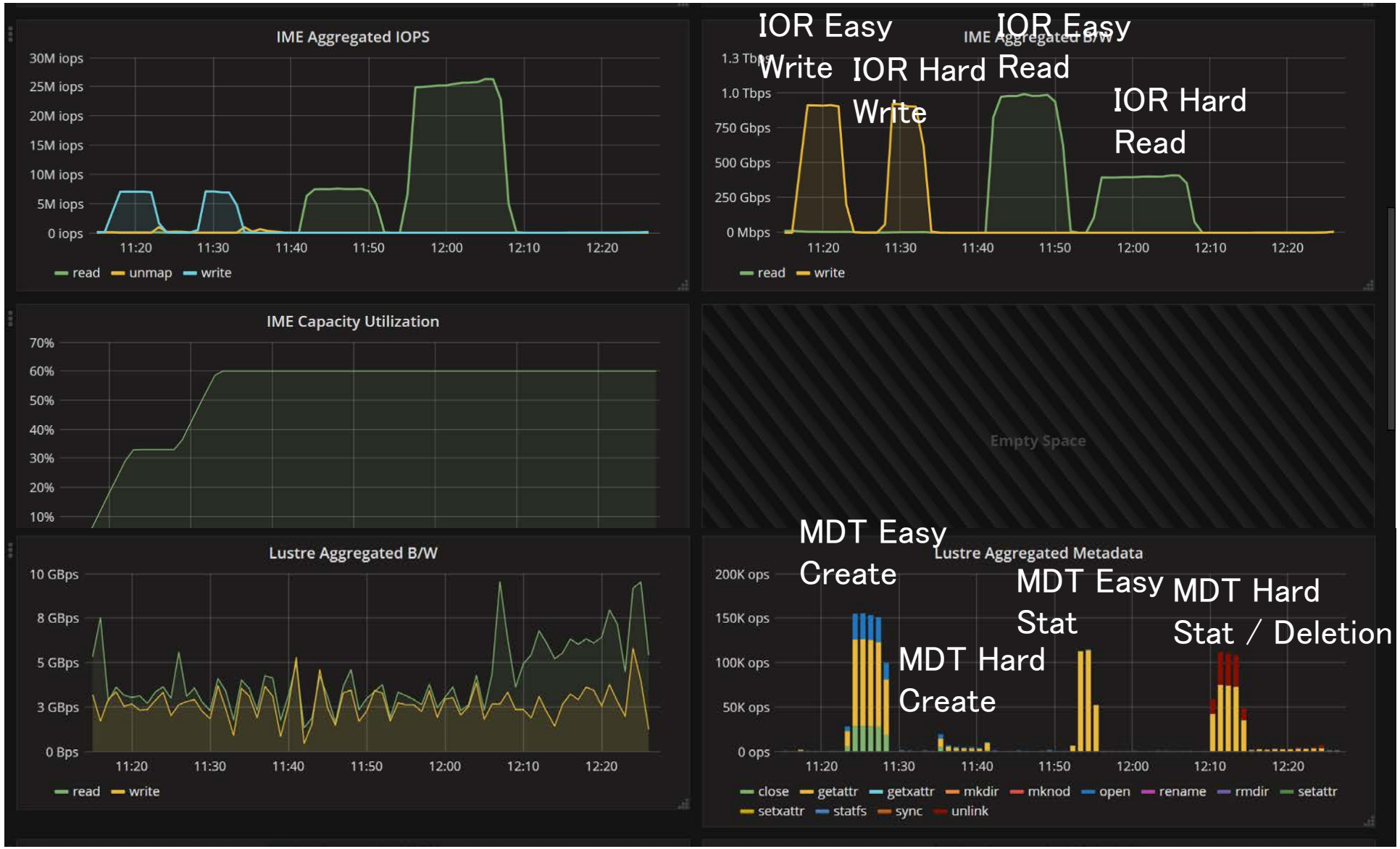
IME IO-500 Bandwidth



IME IO-500 Metadata



IME Performance Monitor



Summary

- IO-500 Benchmark expects
  - Storage system R&D for both bandwidth and metadata performance
  - More attention for storage in supercomputer procurement
- IO-500 benchmark result

	Bandwidth [GiB/s]	Metadata [kIOP/s]	Score
Lustre	21.4	88.78*	42.18*
IME	471.25	21.85	101.48

* MDT Hard takes less than 5 minutes

- IME bandwidth and metadata performance can be improved more