DAOS beyond Persistent Memory

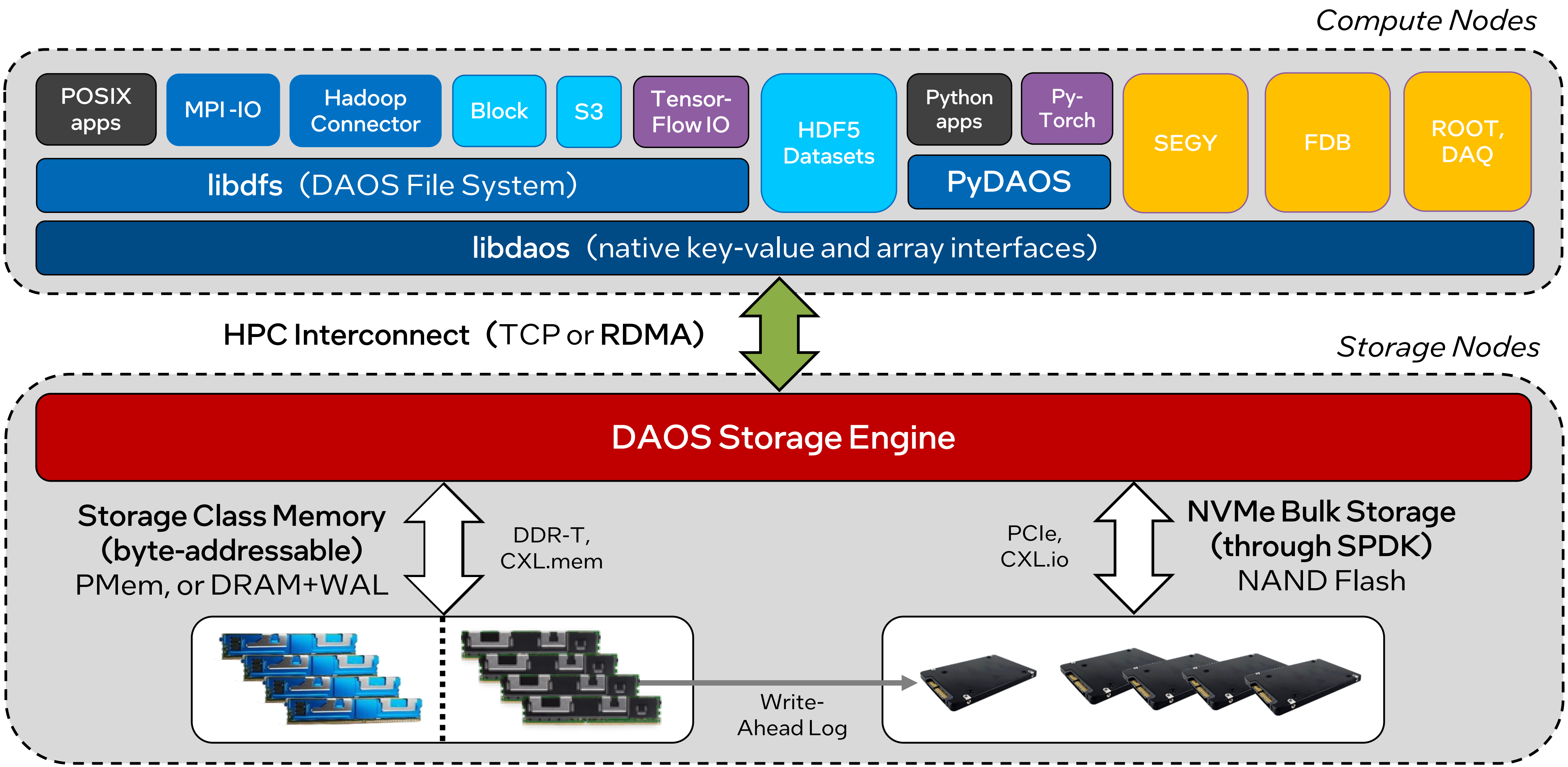
4th Workshop on Heterogeneous Memory Systems (HMEM)

Full ISC23 IXPUG Workshop Paper: https://doi.org/10.1007/978-3-031-40843-4_26

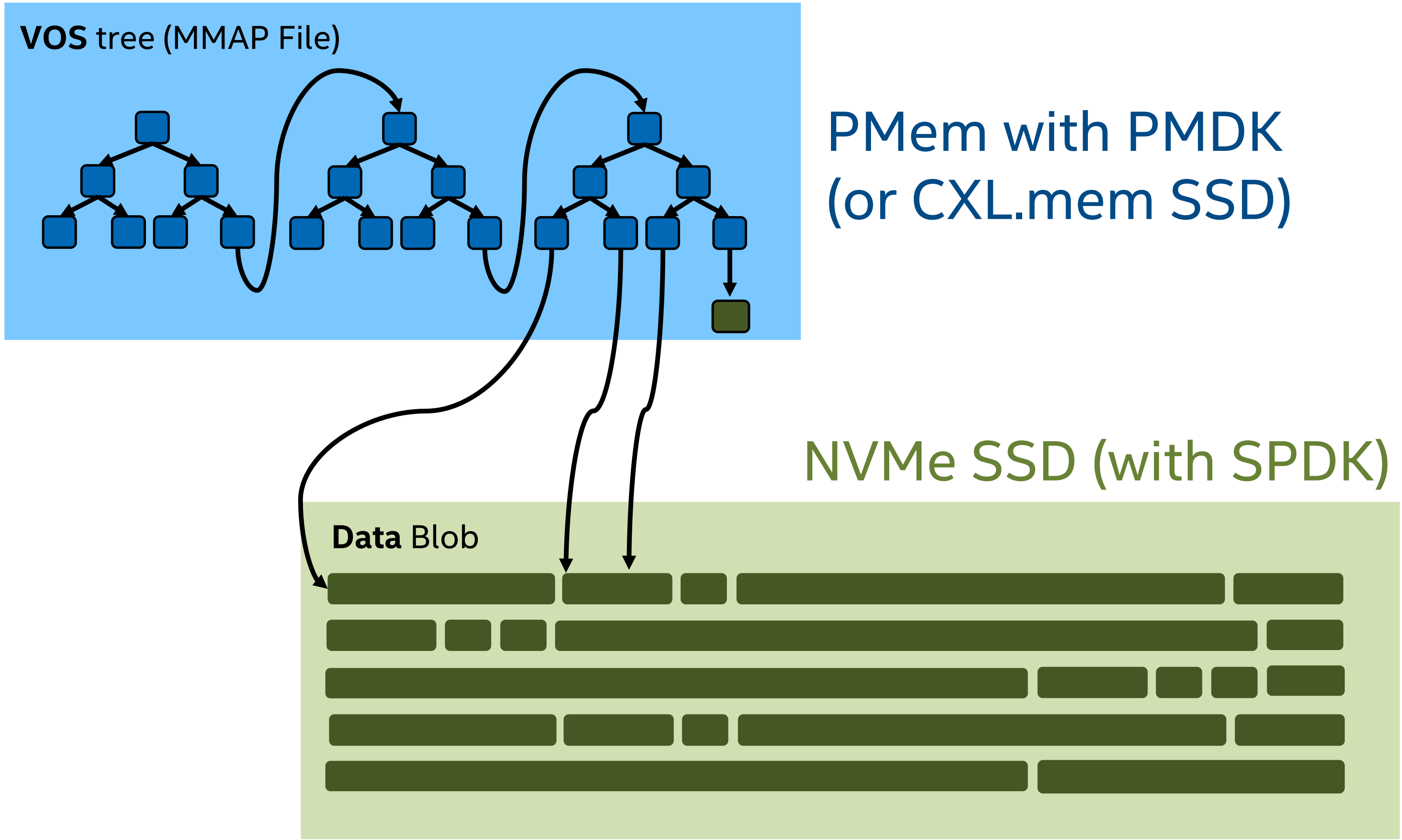
Michael Hennecke – Intel Corporation

17-Nov-2023

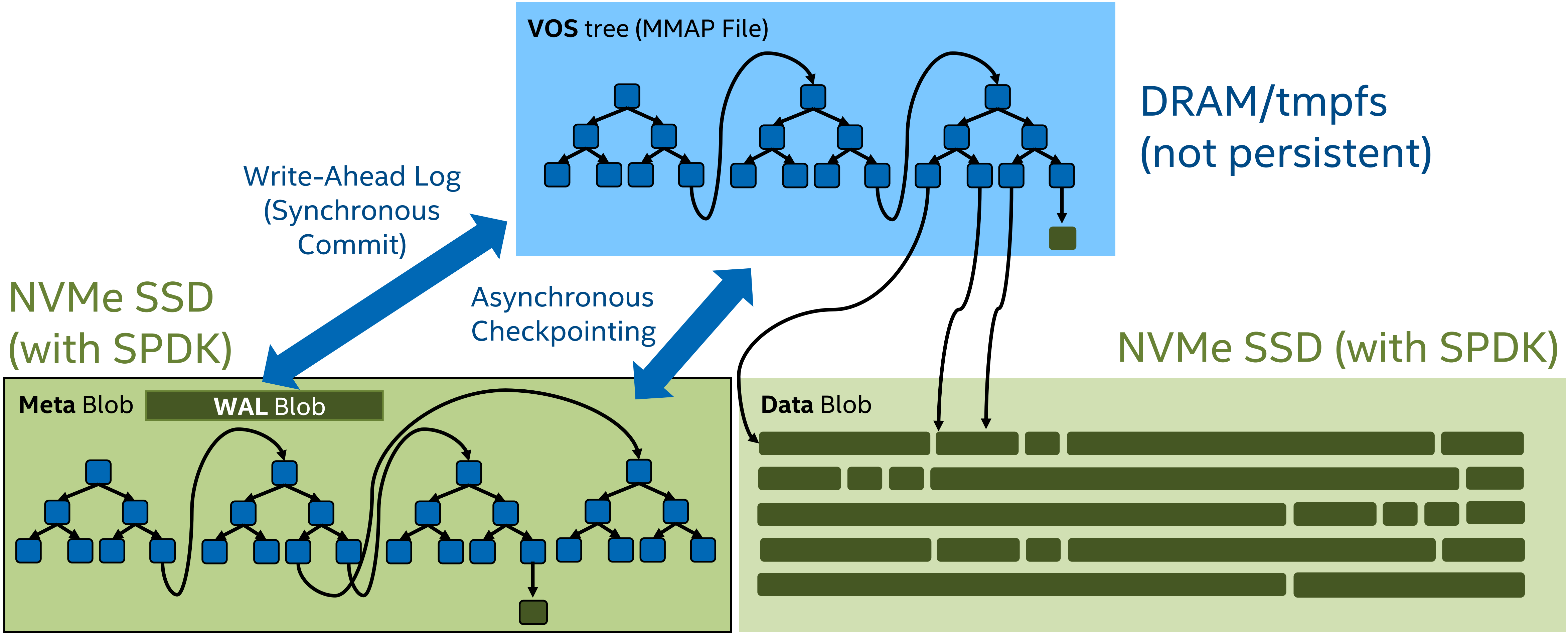
DAOS beyond Persistent Memory



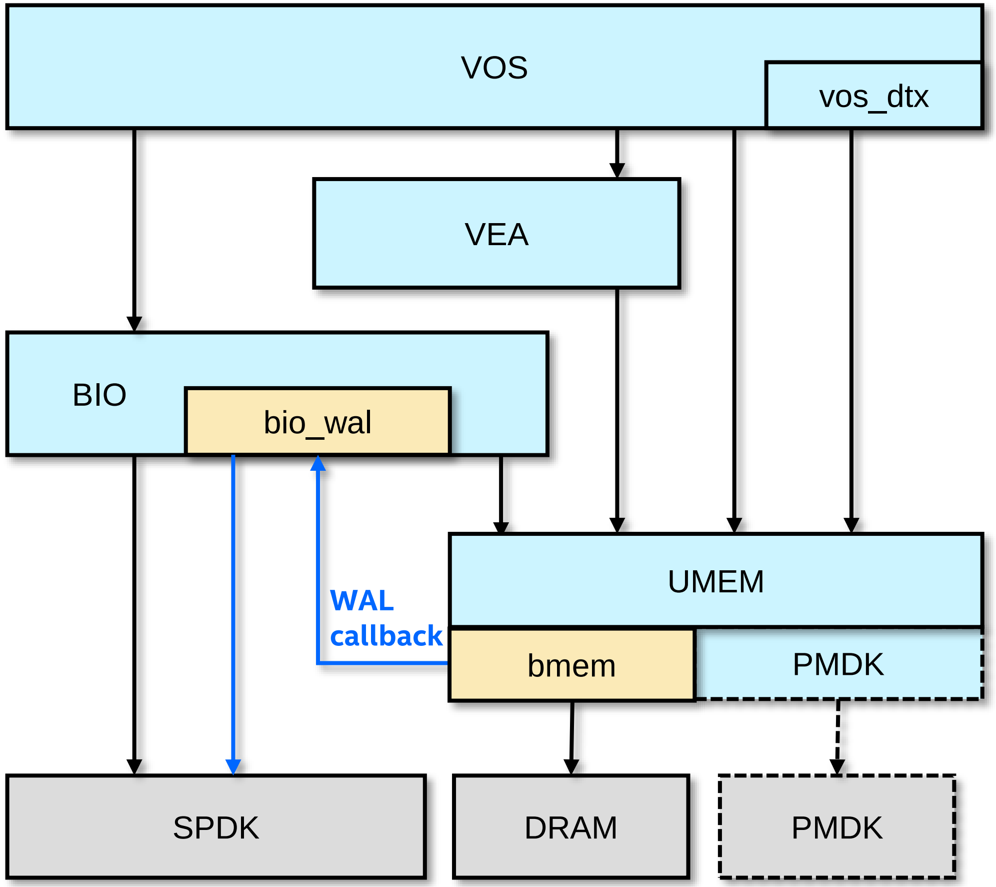
DAOS Backend using Persistent Memory



DAOS Backend using Volatile Memory



New DAOS Backend Stack Layering

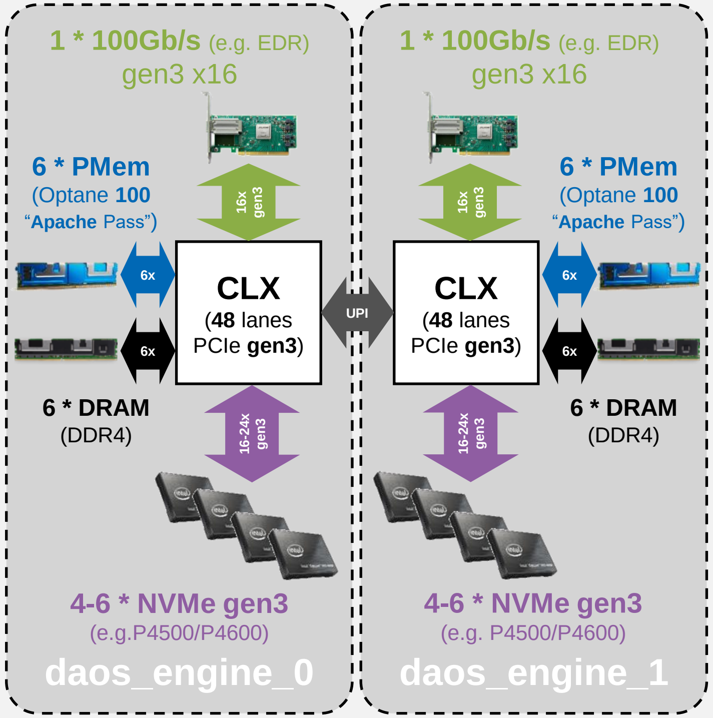


VOS = Versioning Object Store
VEA = Versioned Extent Allocation
BIO = Blob I/O
DTX = DAOS Transaction
UMEM = Unified Memory
PMDK = Persistent Memory Dev Kit
SPDK = Storage Performance Dev Kit
WAL = Write Ahead Log
bmem = Blob Memory allocator

Changes isolated to a few layers

DAOS Servers on 2nd and 3rd Gen Intel Xeon SP

2nd Gen Intel Xeon Scalable Processors



1 * 100Gb/s (e.g. EDR)
gen3 x16

6 * PMem
(Optane 100
“Apache Pass”)

CLX
(48 lanes
PCIe gen3)

6 * DRAM
(DDR4)

4-6 * NVMe gen3
(e.g.P4500/P4600)

daos_engine_0

1 * 100Gb/s (e.g. EDR)
gen3 x16

6 * PMem
(Optane 100
“Apache Pass”)

CLX
(48 lanes
PCIe gen3)

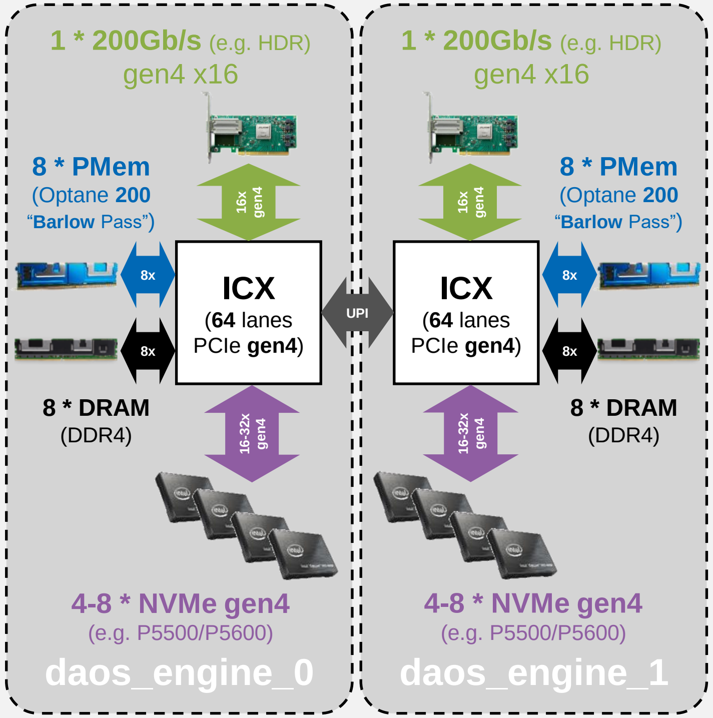
6 * DRAM
(DDR4)

4-6 * NVMe gen3
(e.g. P4500/P4600)

daos_engine_1

Bandwidth: **≤23 GiB/s** per server

3rd Gen Intel Xeon Scalable Processors



1 * 200Gb/s (e.g. HDR)
gen4 x16

8 * PMem
(Optane 200
“Barlow Pass”)

ICX
(64 lanes
PCIe gen4)

8 * DRAM
(DDR4)

4-8 * NVMe gen4
(e.g. P5500/P5600)

daos_engine_0

1 * 200Gb/s (e.g. HDR)
gen4 x16

8 * PMem
(Optane 200
“Barlow Pass”)

ICX
(64 lanes
PCIe gen4)

8 * DRAM
(DDR4)

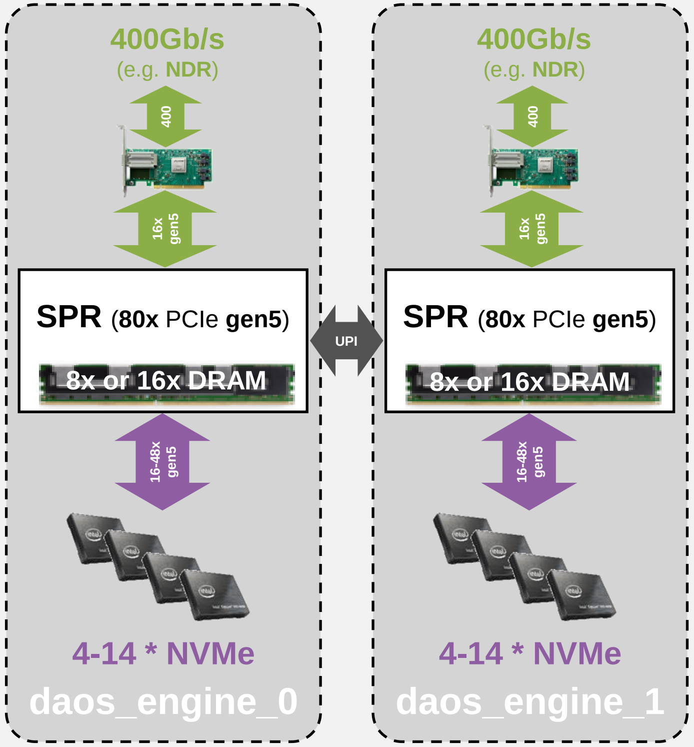
4-8 * NVMe gen4
(e.g. P5500/P5600)

daos_engine_1

Bandwidth: **≤46 GiB/s** per server

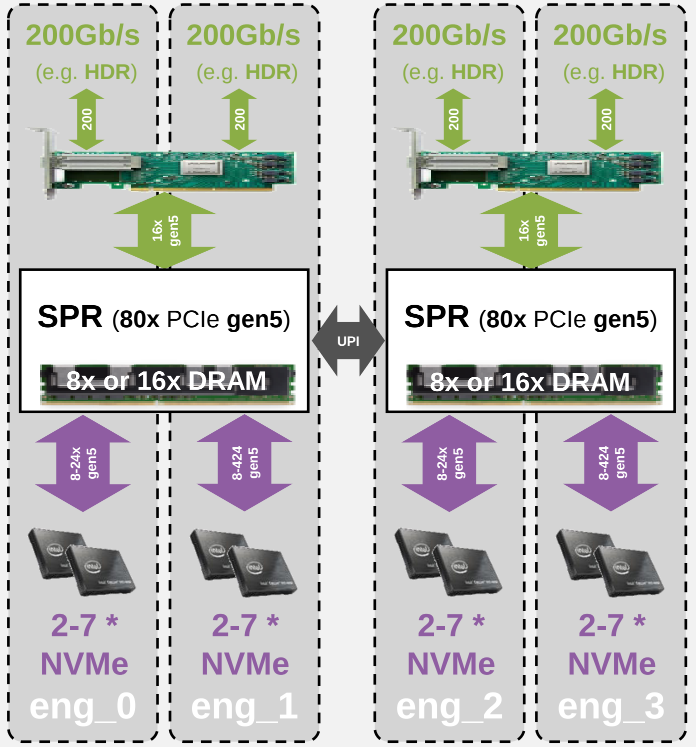
DAOS Server Design Options for 4th Gen Xeon SP

4th Gen Intel Xeon SP 2 * 400 Gbps



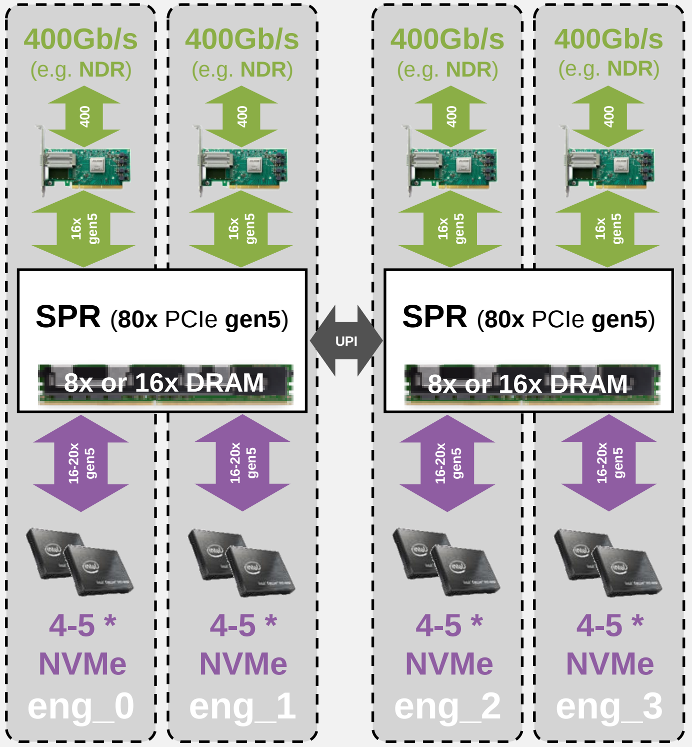
Bandwidth: **≤92 GiB/s** per server

4th Gen Intel Xeon SP 4 * 200 Gbps



Bandwidth: **≤92 GiB/s** per server

4th Gen Intel Xeon SP 4 * 400 Gbps

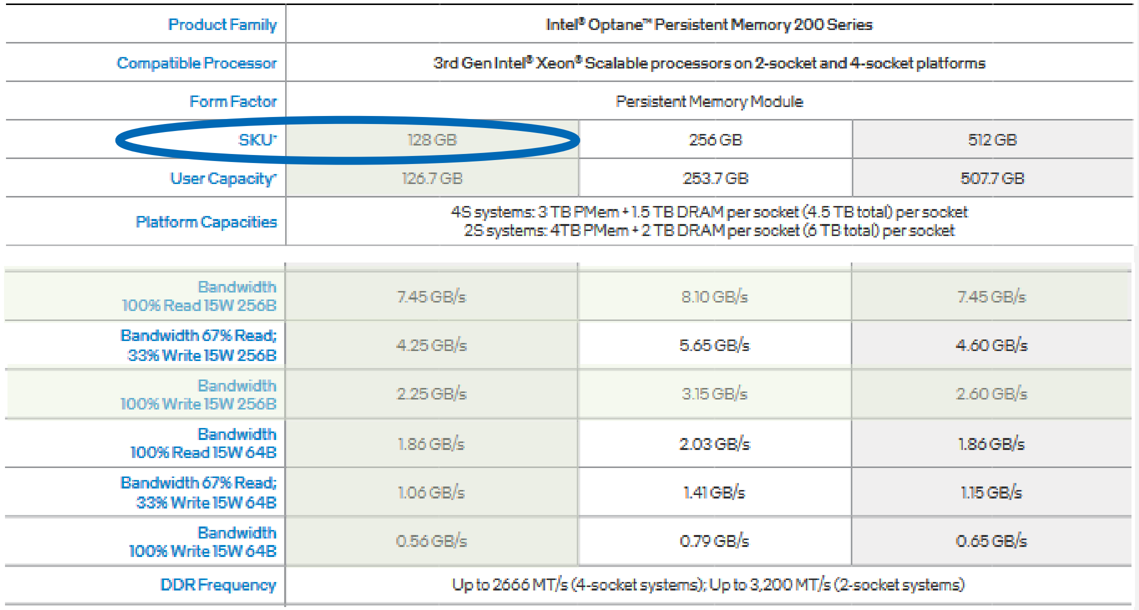


Bandwidth: **≤184 GiB/s** per server

Performance Expectations for Optane 200 PMem

- 1x PMem device bandwidth:
  - Read: 7.45 GB/s = 6.93 GiB/s (256B xfers)
  - Write: 2.25 GB/s = 2.06 GiB/s (256B xfers)
  - Read: 1.86 GB/s = 1.73 GiB/s (64B xfers)
  - Write: 0.56 GB/s = 0.52 GiB/s (64B xfers)
- With 8x PMem per CPU/DAOS engine:
  - Read: 8x 6.93 GiB/s = 55.4 GiB/s (256B xfers)
  - Write: 8x 2.06 GiB/s = 16.5 GiB/s (256B xfers)
  - Read: 8x 1.73 GiB/s = 13.8 GiB/s (64B xfers)
  - Write: 8x 0.52 GiB/s = 4.2 GiB/s (64B xfers)

https://www.intel.com/content/www/us/en/products/docs/memory-storage/optane-persistent-memory/optane-persistent-memory-200-series-brief.html

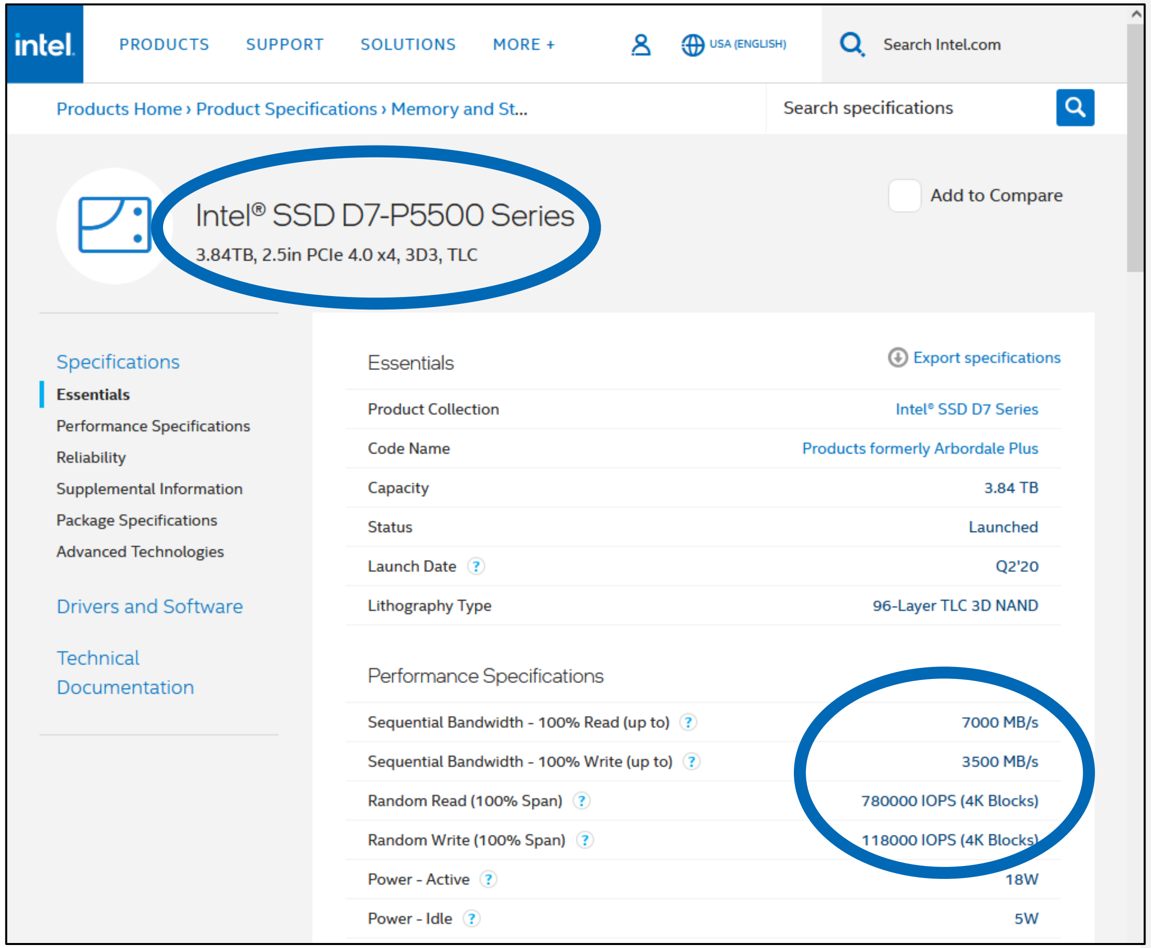


Product Family	Intel® Optane™ Persistent Memory 200 Series		
Compatible Processor	3rd Gen Intel® Xeon® Scalable processors on 2-socket and 4-socket platforms		
Form Factor	Persistent Memory Module		
SKU*	128 GB	256 GB	512 GB
User Capacity*	126.7 GB	253.7 GB	507.7 GB
Platform Capacities	4S systems: 3 TB PMem + 1.5 TB DRAM per socket (4.5 TB total) per socket 2S systems: 4TB PMem + 2 TB DRAM per socket (6 TB total) per socket		
Bandwidth 100% Read 15W 256B	7.45 GB/s	8.10 GB/s	7.45 GB/s
Bandwidth 67% Read; 33% Write 15W 256B	4.25 GB/s	5.65 GB/s	4.60 GB/s
Bandwidth 100% Write 15W 256B	2.25 GB/s	3.15 GB/s	2.60 GB/s
Bandwidth 100% Read 15W 64B	1.86 GB/s	2.03 GB/s	1.86 GB/s
Bandwidth 67% Read; 33% Write 15W 64B	1.06 GB/s	1.41 GB/s	1.15 GB/s
Bandwidth 100% Write 15W 64B	0.56 GB/s	0.79 GB/s	0.65 GB/s
DDR Frequency	Up to 2666 MT/s (4-socket systems); Up to 3,200 MT/s (2-socket systems)		

Performance Expectations for a gen4 NVMe Disk

- 1x NVMe disk 4kiB random IOPS:
  - Read: 780000/s * 4kiB = 2.98 GiB/s
  - Write: 118000/s * 4kiB = 0.45 GiB/s
- 4x NVMe disks 4kiB random IOPS:
  - Read: 4x 2.98 GiB/s = 11.9 GiB/s
  - Write: 4x 0.45 GiB/s = 1.8 GiB/s
- Latency of a *single* I/O operation has an impact on required queue depth to achieve good bandwidth
  - **Optane** SSDs perform better than NAND for low qdepth = low level of parallelism…

https://ark.intel.com/content/www/us/en/ark/products/202705/intel-ssd-d7p5500-series-3-84tb-2-5in-pcie-4-0-x4-3d3-tlc.html



Intel® SSD D7-P5500 Series

3.84TB, 2.5in PCIe 4.0 x4, 3D3, TLC

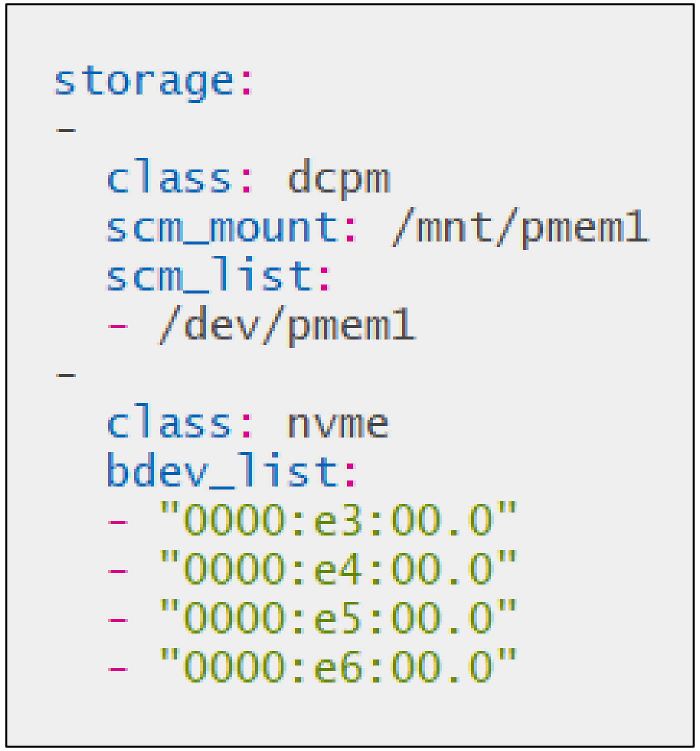
Essentials

Product Collection	Intel® SSD D7 Series
Code Name	Products formerly Arbordale Plus
Capacity	3.84 TB
Status	Launched
Launch Date ?	Q2'20
Lithography Type	96-Layer TLC 3D NAND

Performance Specifications

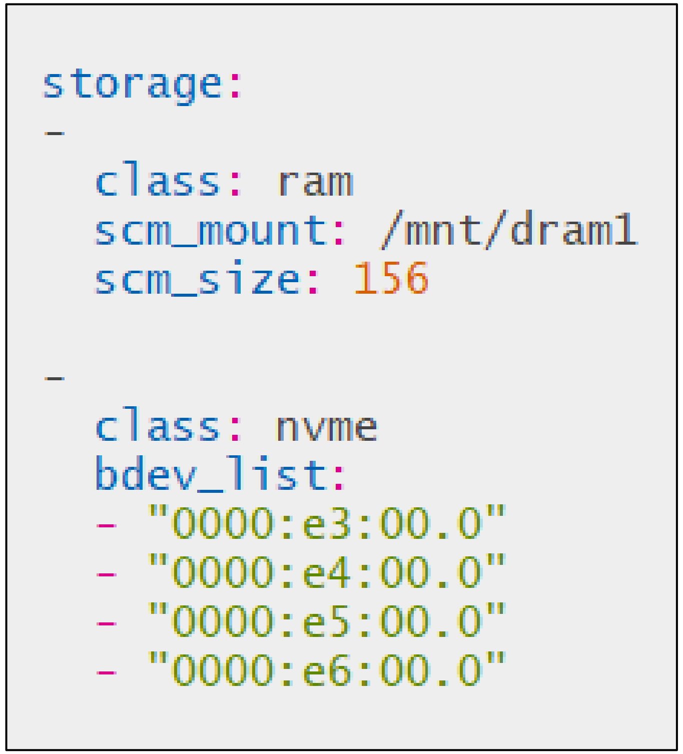
Sequential Bandwidth - 100% Read (up to) ?	7000 MB/s
Sequential Bandwidth - 100% Write (up to) ?	3500 MB/s
Random Read (100% Span) ?	780000 IOPS (4K Blocks)
Random Write (100% Span) ?	118000 IOPS (4K Blocks)
Power - Active ?	18W
Power - Idle ?	5W

“Traditional” Configuration Options in daos_server.yml



```
storage:
-
  class: dcpm
  scm_mount: /mnt/pmem1
  scm_list:
  - /dev/pmem1
-
  class: nvme
  bdev_list:
  - "0000:e3:00.0"
  - "0000:e4:00.0"
  - "0000:e5:00.0"
  - "0000:e6:00.0"
```

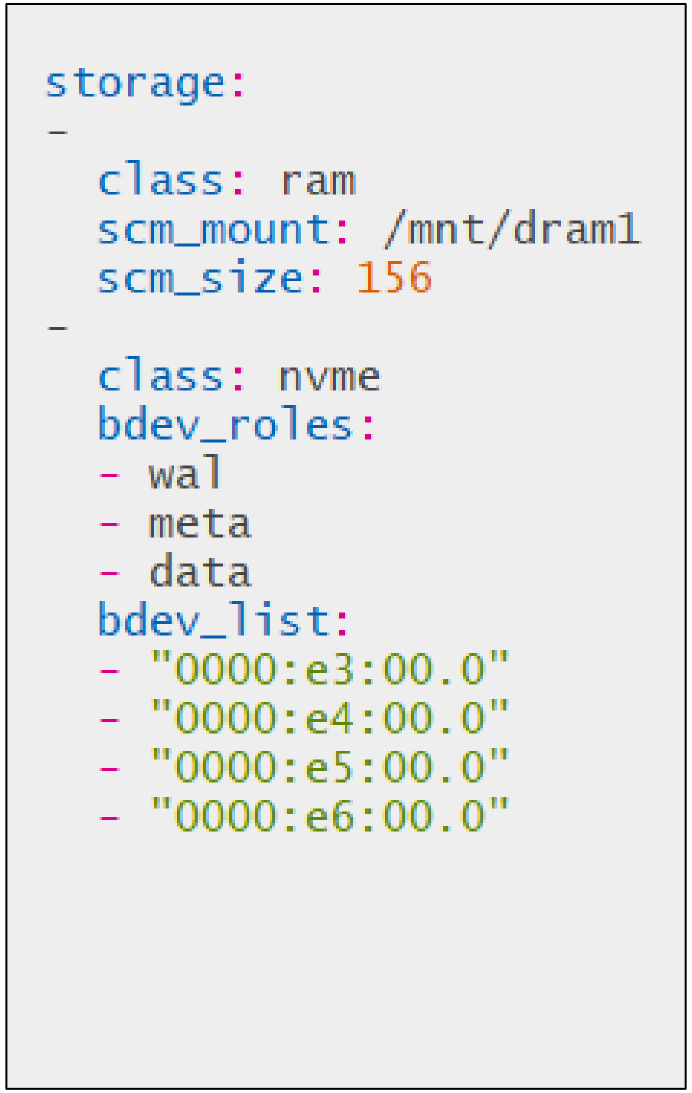
PMem-based DAOS



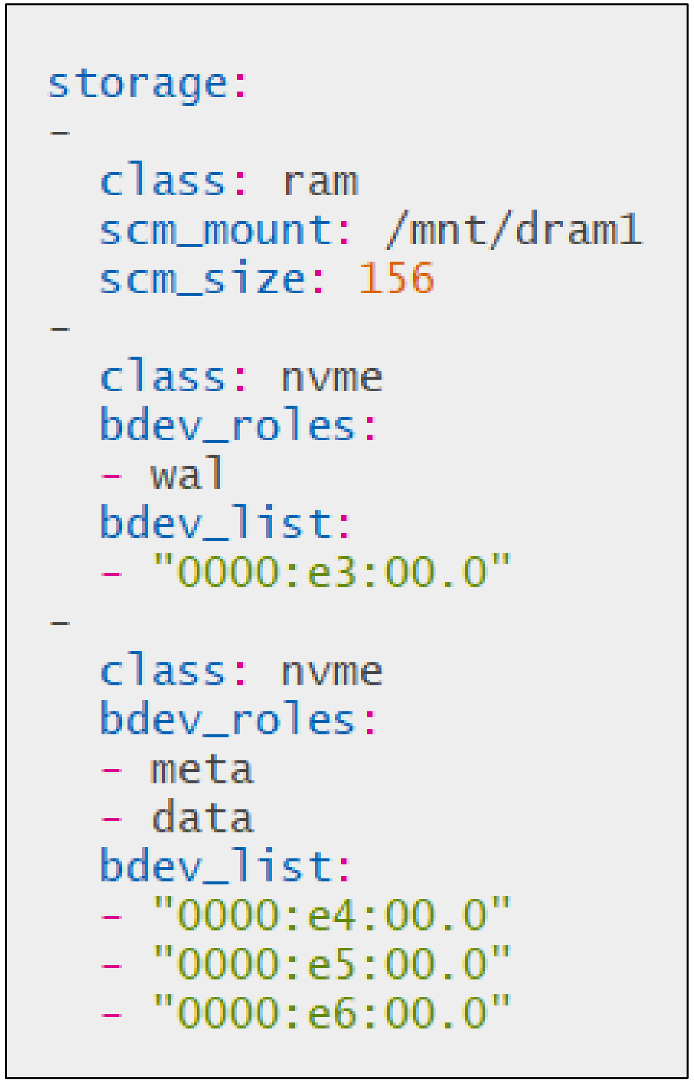
```
storage:
-
  class: ram
  scm_mount: /mnt/dram1
  scm_size: 156

-
  class: nvme
  bdev_list:
  - "0000:e3:00.0"
  - "0000:e4:00.0"
  - "0000:e5:00.0"
  - "0000:e6:00.0"
```

“Ephemeral” DAOS

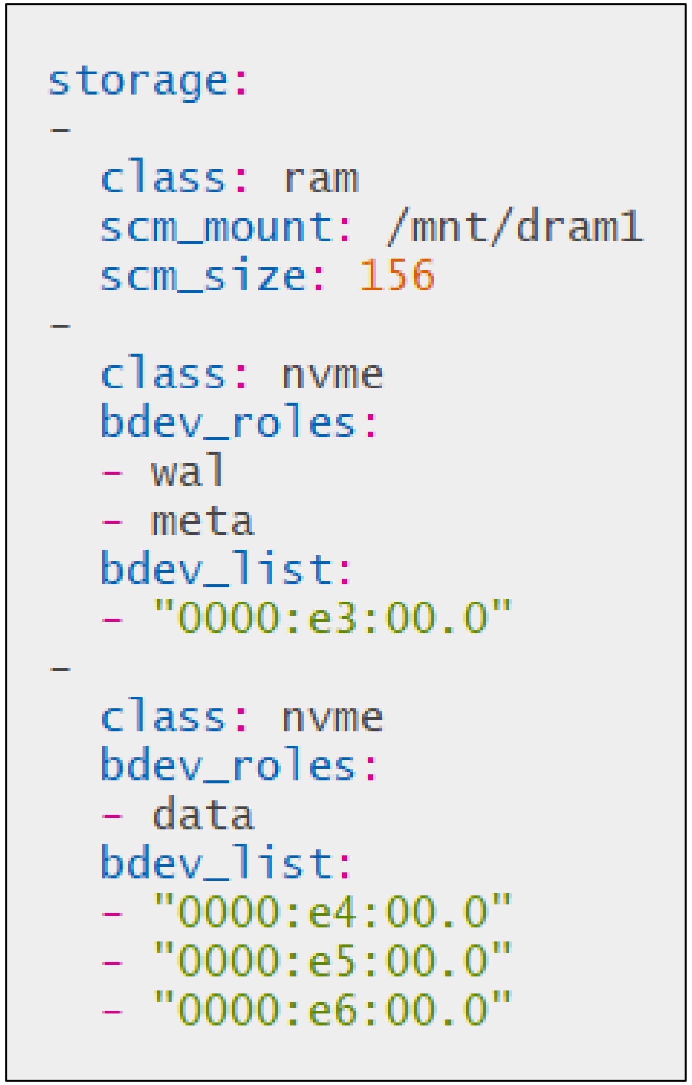
“MD-on-SSD” Configuration Options in daos_server.yml



```
storage:
-
  class: ram
  scm_mount: /mnt/dram1
  scm_size: 156
-
  class: nvme
  bdev_roles:
  - wal
  - meta
  - data
  bdev_list:
  - "0000:e3:00.0"
  - "0000:e4:00.0"
  - "0000:e5:00.0"
  - "0000:e6:00.0"
```



```
storage:
-
  class: ram
  scm_mount: /mnt/dram1
  scm_size: 156
-
  class: nvme
  bdev_roles:
  - wal
  bdev_list:
  - "0000:e3:00.0"
-
  class: nvme
  bdev_roles:
  - meta
  - data
  bdev_list:
  - "0000:e4:00.0"
  - "0000:e5:00.0"
  - "0000:e6:00.0"
```



```
storage:
-
  class: ram
  scm_mount: /mnt/dram1
  scm_size: 156
-
  class: nvme
  bdev_roles:
  - wal
  - meta
  bdev_list:
  - "0000:e3:00.0"
-
  class: nvme
  bdev_roles:
  - data
  bdev_list:
  - "0000:e4:00.0"
  - "0000:e5:00.0"
  - "0000:e6:00.0"
```

Metadata Performance

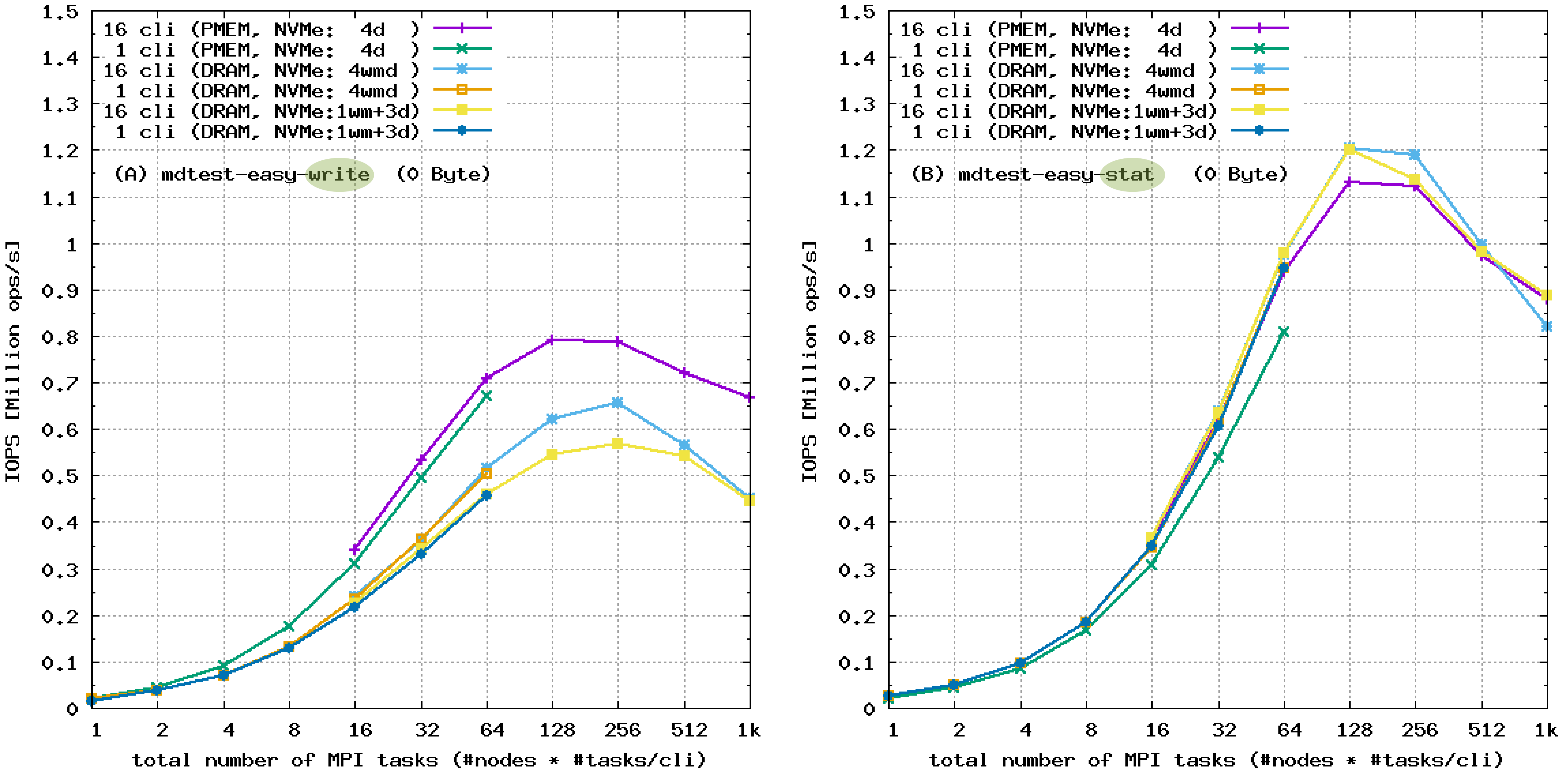
(1 engine @ 24 targets; HDR IB; 8TB pool; 30sec stonewall)

mdtest-easy (0-Byte files; dir-per-process)

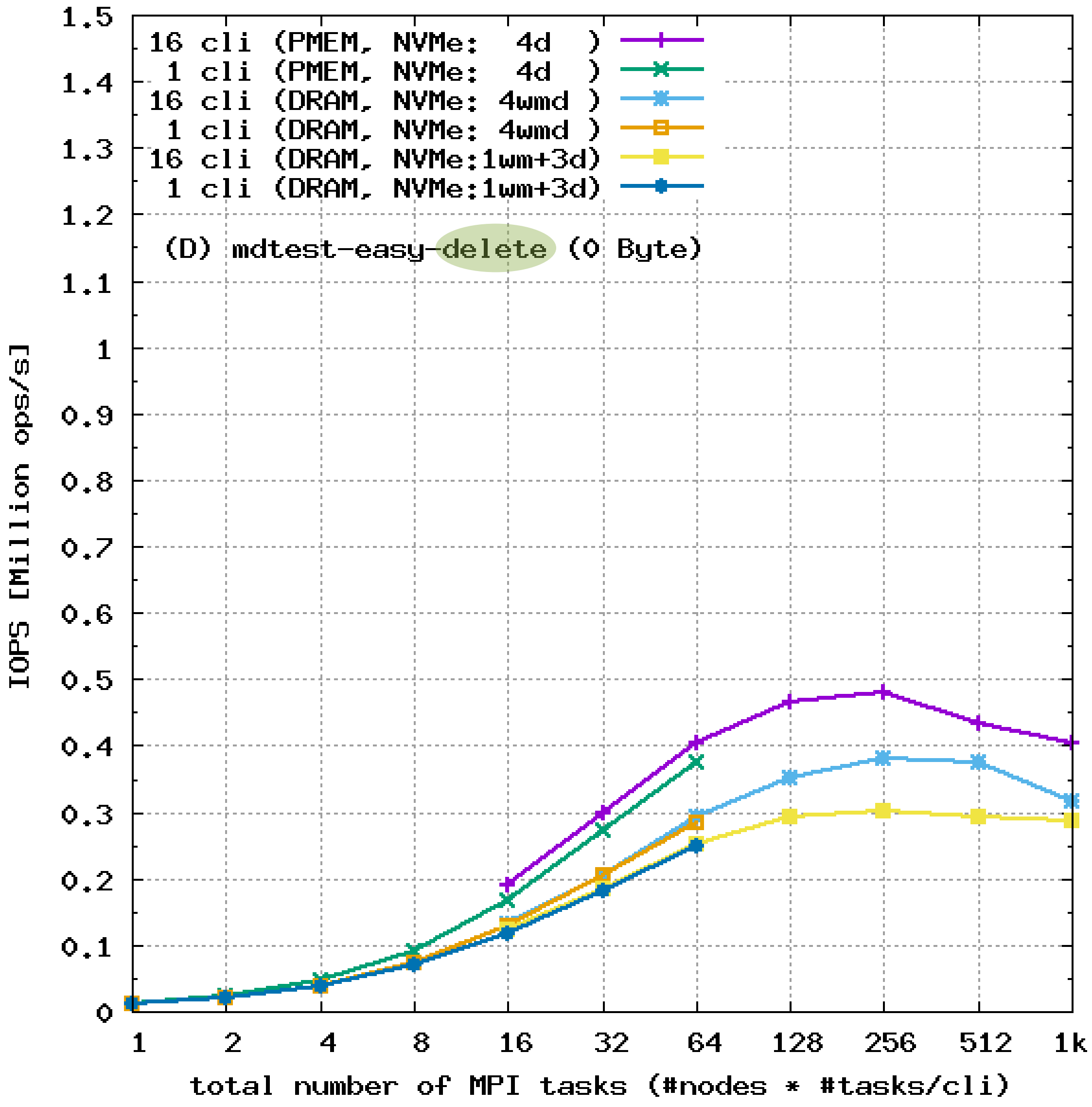
mdtest-hard (3901-Byte files; shared-dir)

mdtest-hard2 (7802-Byte files; shared-dir)

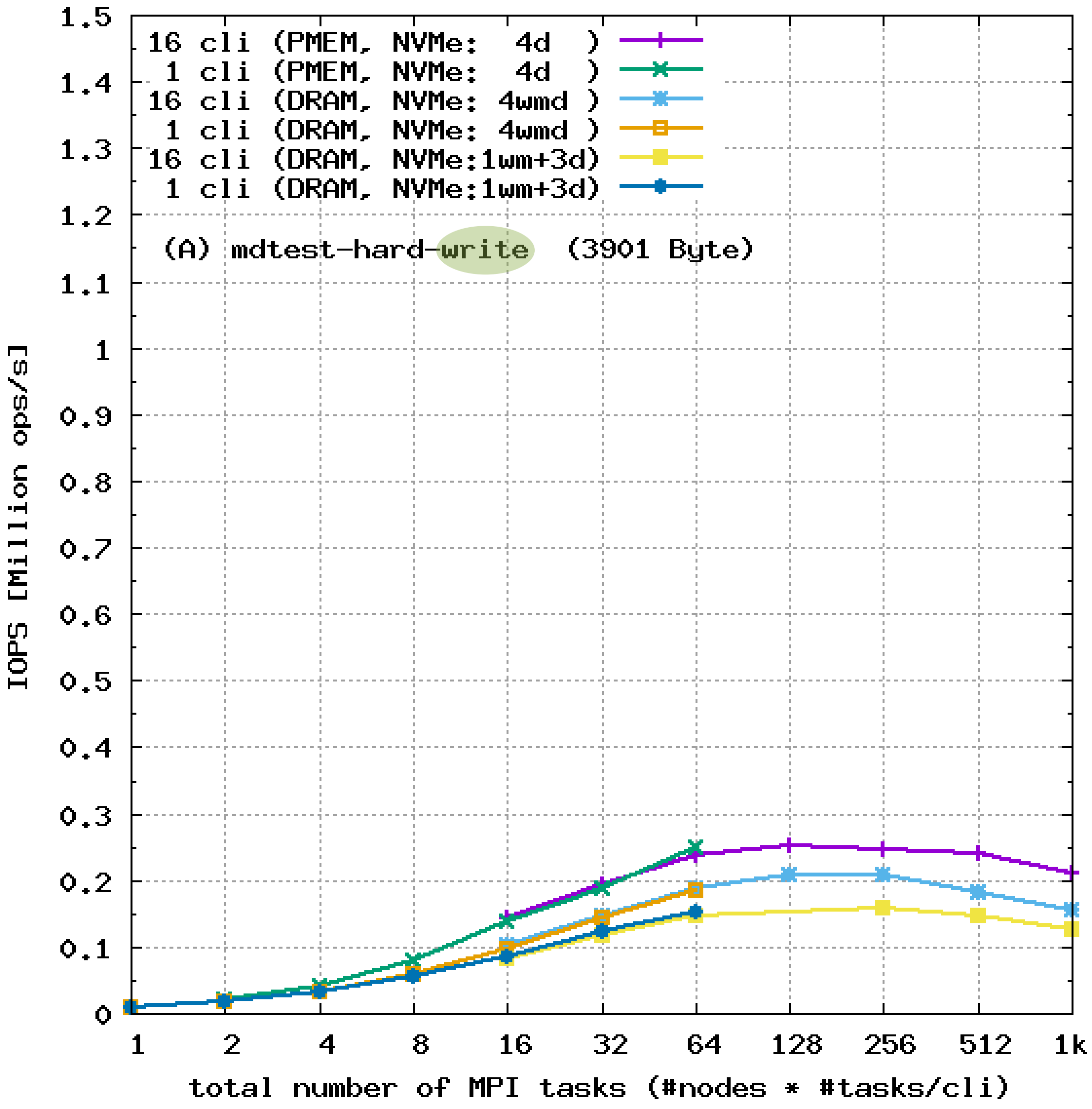
mdtest-easy (0-Byte files): (A) write (B) stat

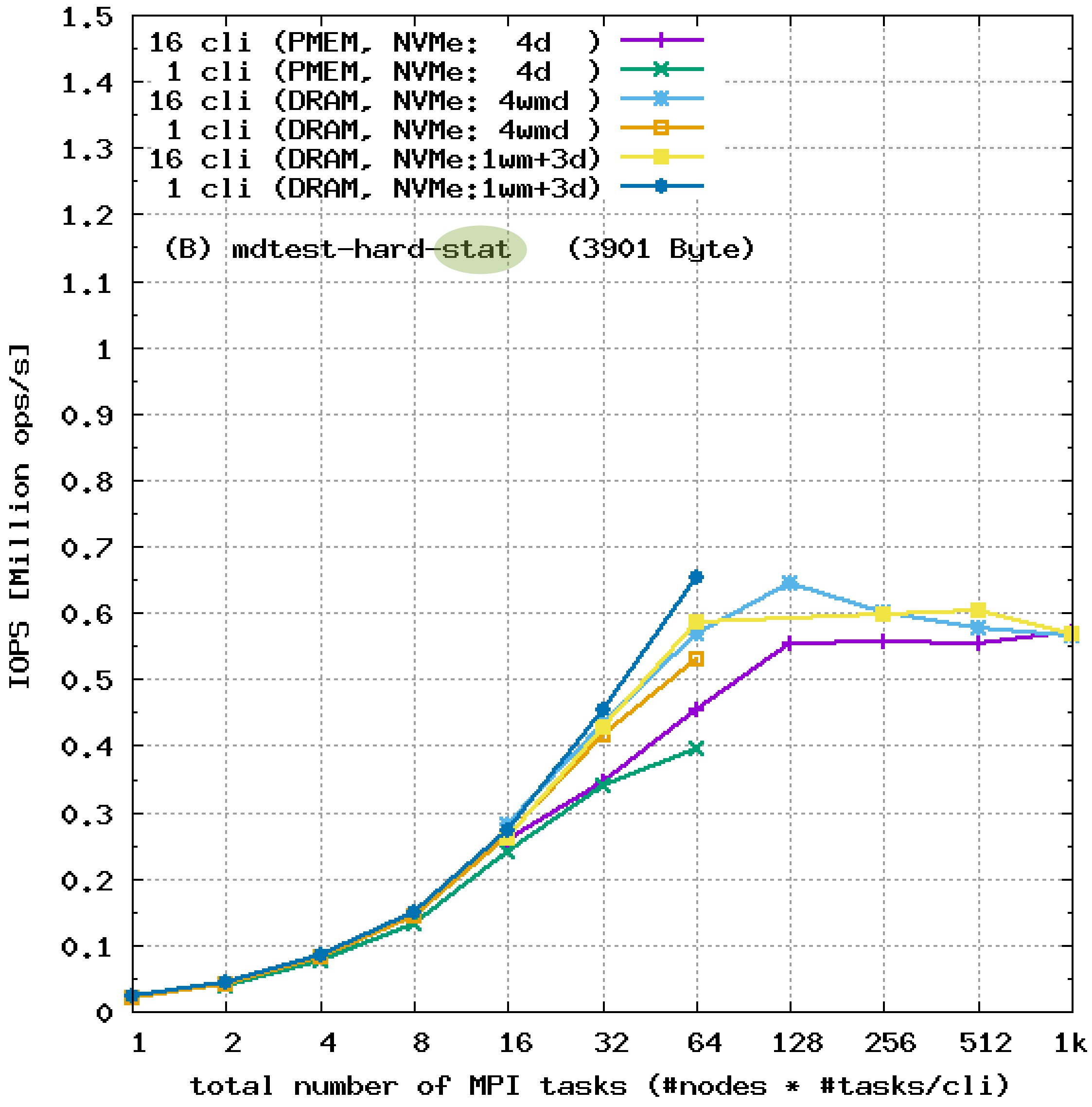


mdtest-easy (0-Byte files): (D) delete

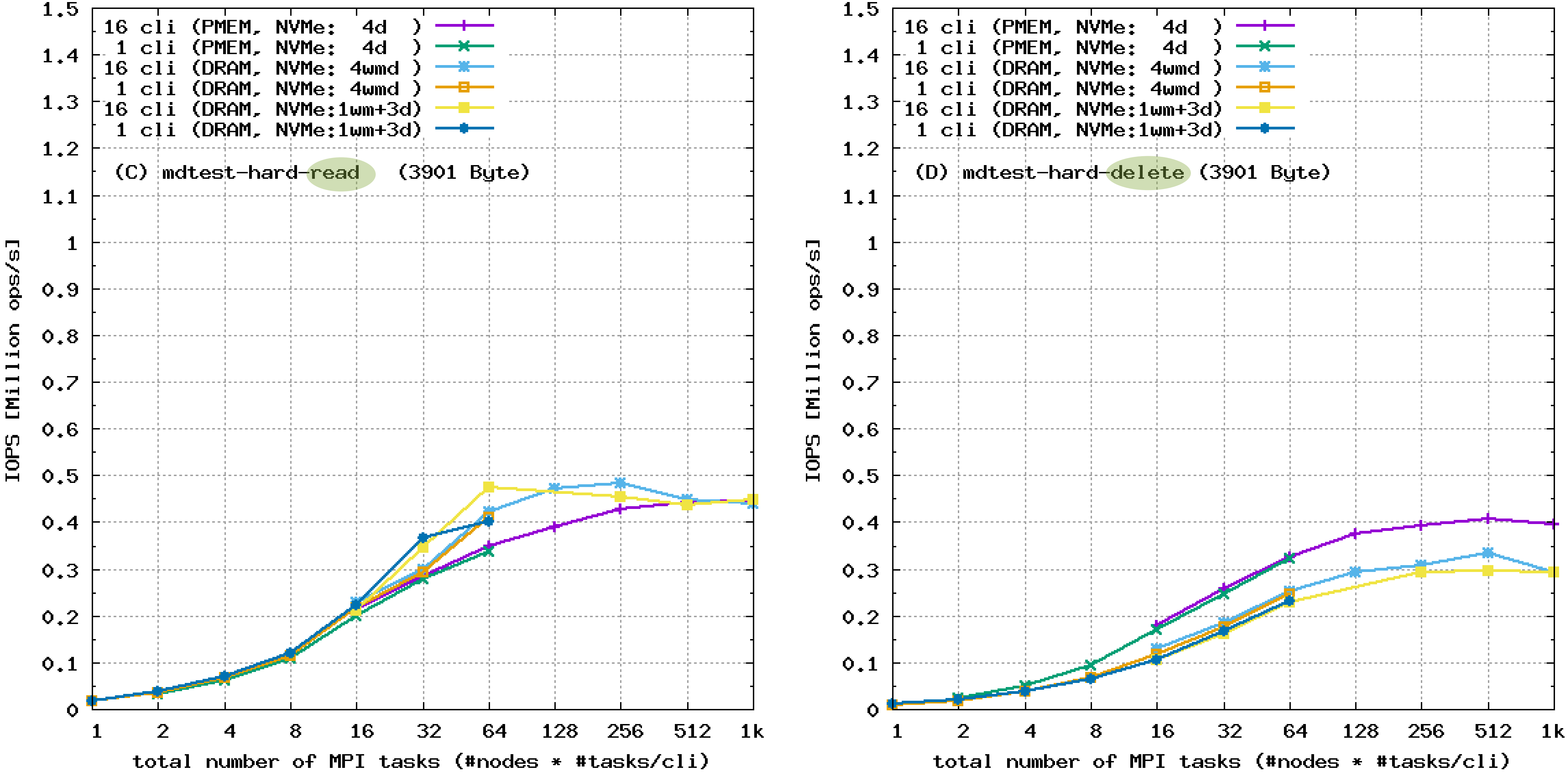


mdtest-hard (3901-Byte files): (A) write (B) stat

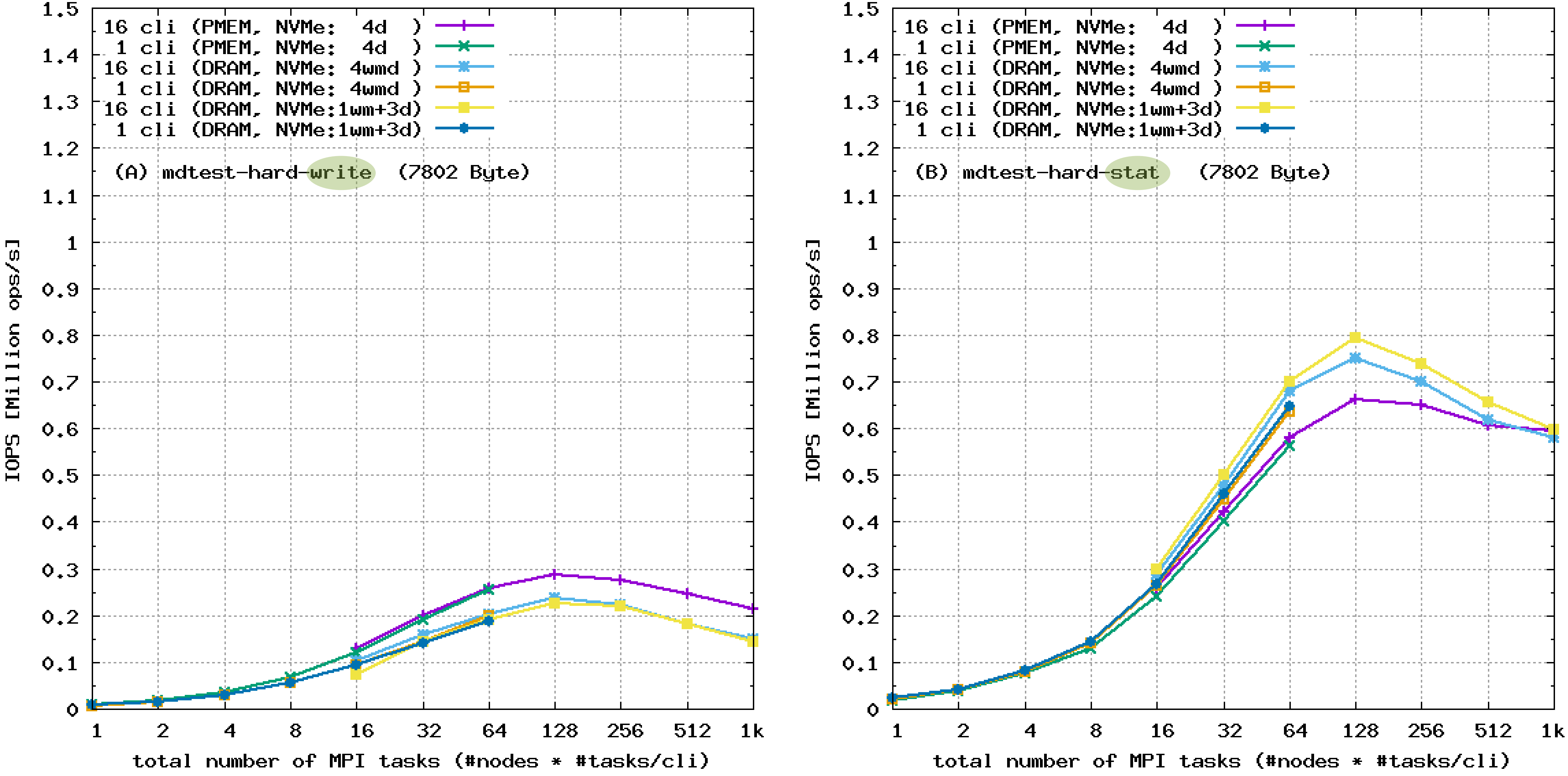




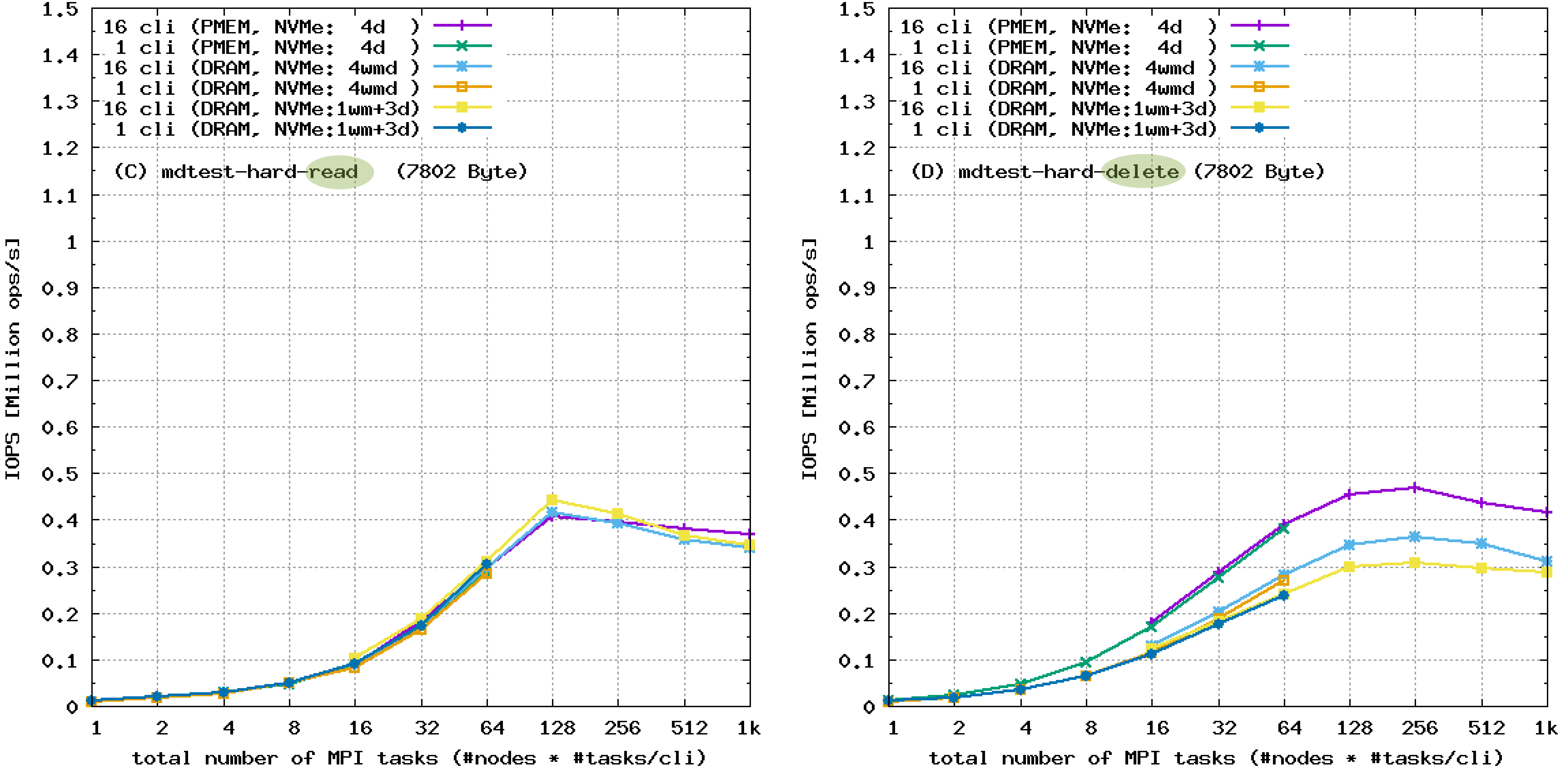
mdtest-hard (3901-Byte files): (C) read (D) delete



mdtest-hard2 (7802-Byte files): (A) write (B) stat



mdtest-hard2 (7802-Byte files): (C) read (D) delete



Summary and DAOS Resources

- DAOS Metadata-on-SSD (Phase 1) is implemented (DAOS 2.4 tech preview)
  - Comparable performance to DAOS on Optane PMem for mdtest-stat, mdtest-read.
  - Some (up to 20%) degradation for mdtest-write, mdtest-delete (synchronous WAL)
  - Full ISC23 workshop paper: https://doi.org/10.1007/978-3-031-40843-4_26
- Future Phase 2 of MD-on-SSD: Enable migration of "cold" metadata to data blobs
  - Will reduce **DRAM capacity** requirements (as a percentage of NVMe capacity)
- DAOS Community Resources:
  - Github: https://github.com/daos-stack/daos
  - Online doc: https://docs.daos.io/
  - Mailing list & slack: https://daos.groups.io/
  - Recordings from 7th DAOS User Group at SC23: https://dug.daos.io/
  - Intel landing page for DAOS: https://www.intel.com/content/www/us/en/high-performance-computing/daos.html

Thank you for attending – Questions?