Distributed Asynchronous Object Storage (DAOS)

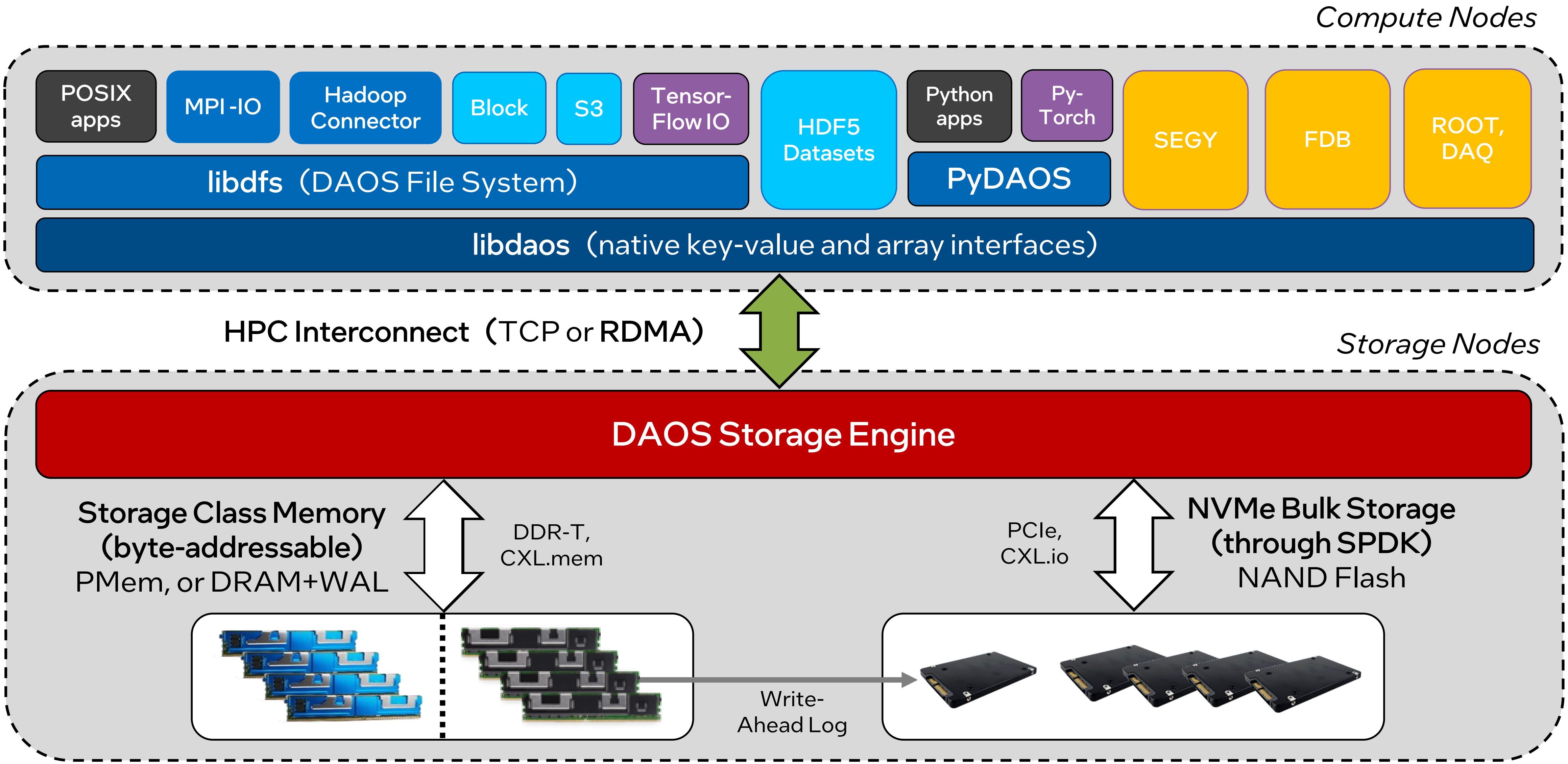
DAOS beyond Persistent Memory

Michael Hennecke et al.

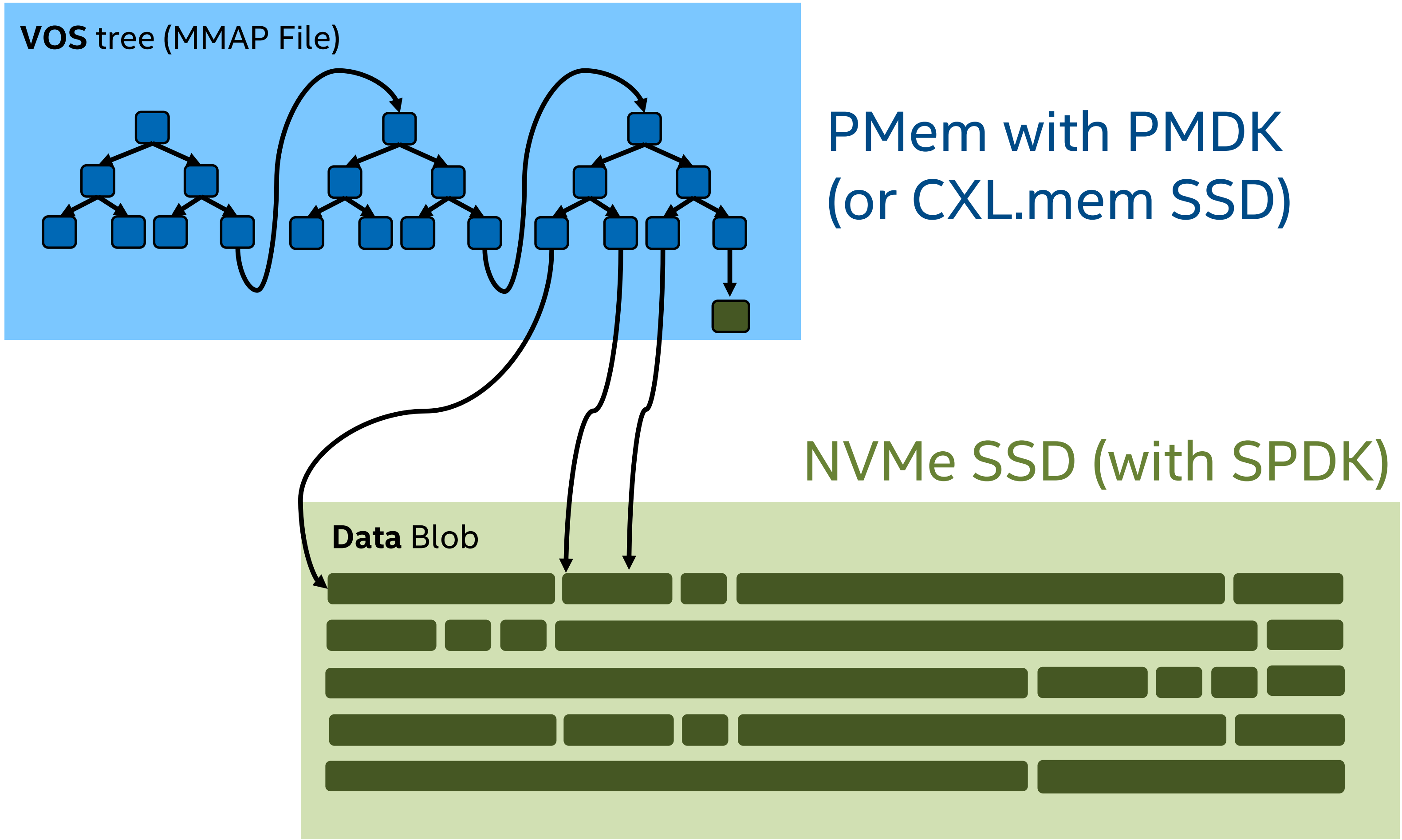
DUG'23; Monday 13-Nov-2023; 9:00am – 12:30pm MST

ISC23 workshop paper:
https://doi.org/10.1007/978-3-031-40843-4_26

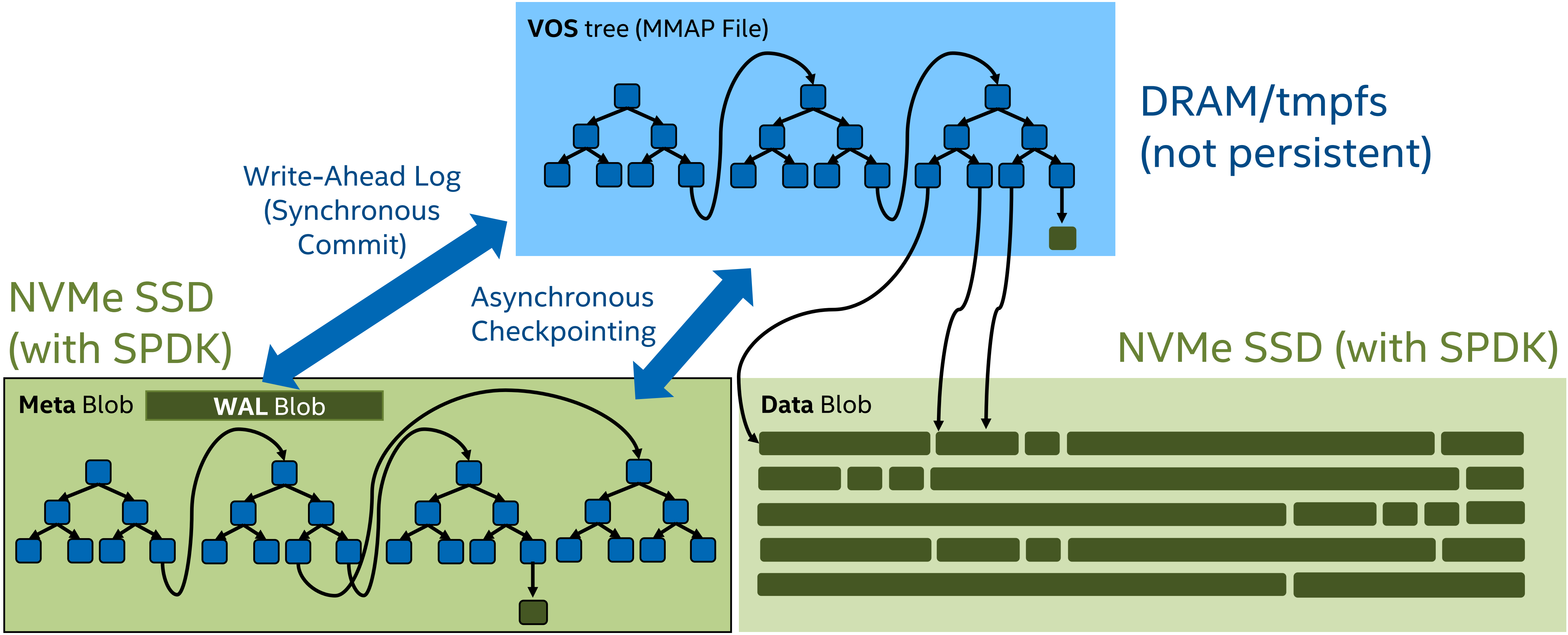
DAOS Software Architecture



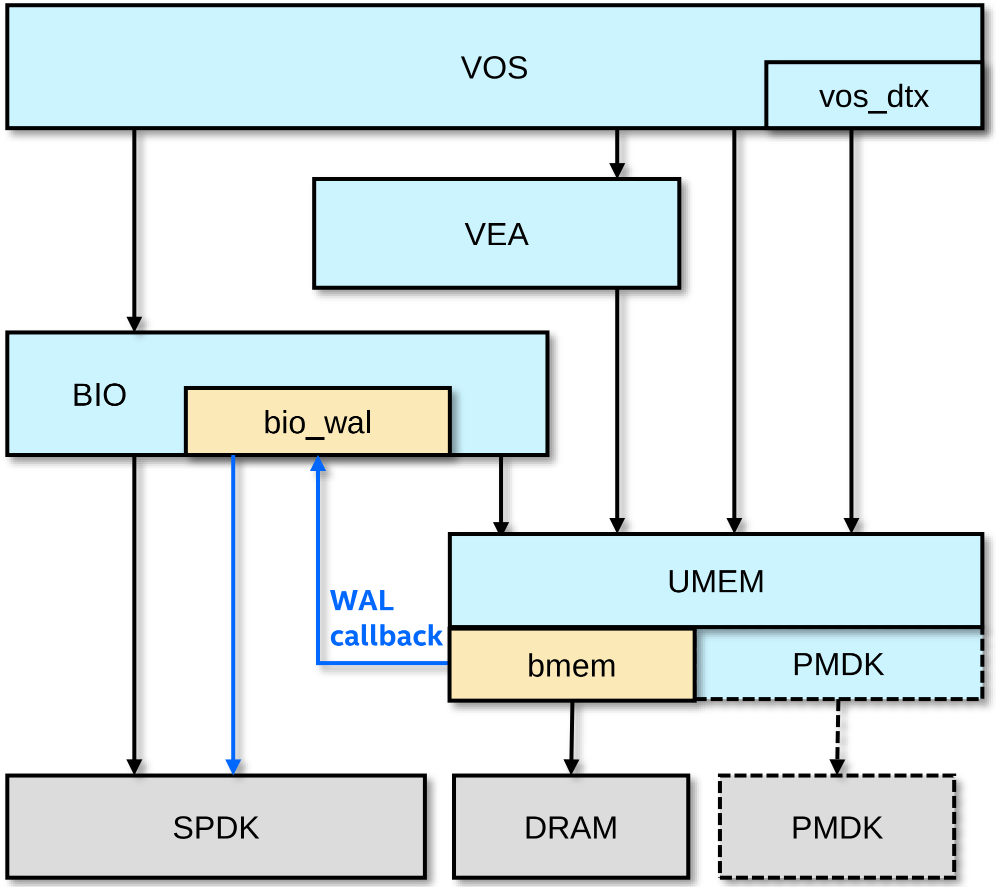
DAOS Backend using Persistent Memory



DAOS Backend using Volatile Memory



New DAOS Backend Stack Layering

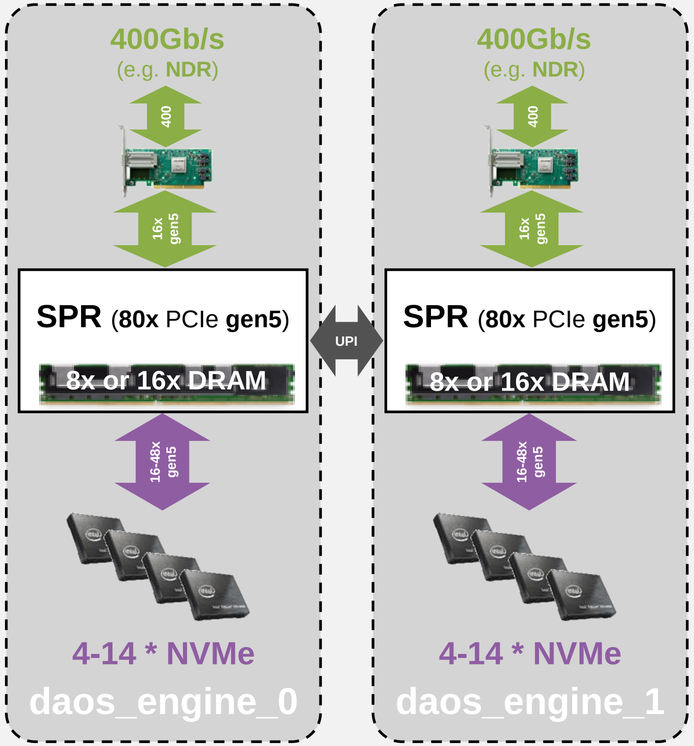


VOS = Versioning Object Store
VEA = Versioned Extent Allocation
BIO = Blob I/O
DTX = DAOS Transaction
UMEM = Unified Memory
PMDK = Persistent Memory Dev Kit
SPDK = Storage Performance Dev Kit
WAL = Write Ahead Log
bmem = Blob Memory allocator

Changes isolated to a few layers

DAOS Server Design Options for 4th Gen Xeon SP

4th Gen Intel Xeon SP
2 * 400 Gbps



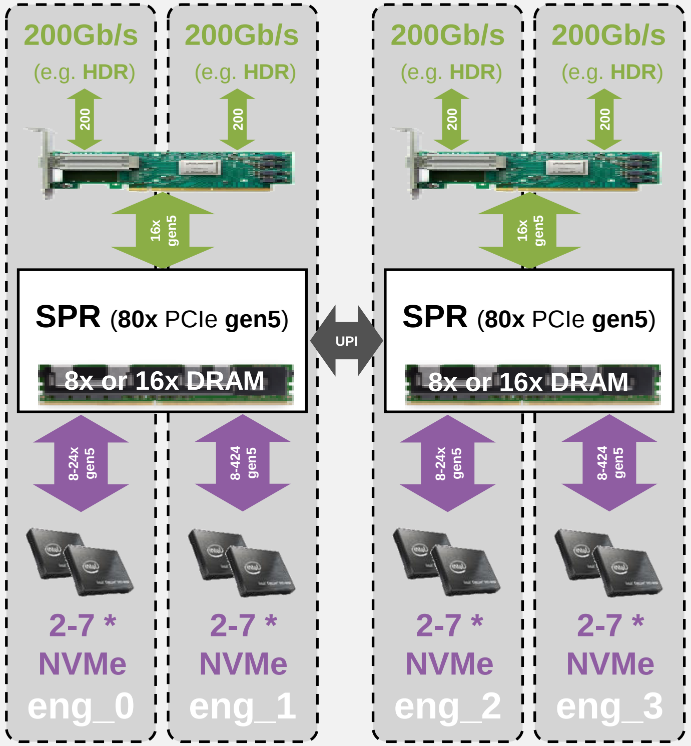
400Gb/s
(e.g. NDR)
400
16x gen5
SPR (80x PCIe gen5)
8x or 16x DRAM
16-48x gen5
4-14 * NVMe
daos_engine_0

UPI

400Gb/s
(e.g. NDR)
400
16x gen5
SPR (80x PCIe gen5)
8x or 16x DRAM
16-48x gen5
4-14 * NVMe
daos_engine_1

Bandwidth: **≤92 GiB/s** per server

4th Gen Intel Xeon SP
4 * 200 Gbps



200Gb/s
(e.g. HDR)
200

200Gb/s
(e.g. HDR)
200

16x gen5
SPR (80x PCIe gen5)
8x or 16x DRAM
8-24x gen5
2-7 * NVMe
eng_0

8-424 gen5
2-7 * NVMe
eng_1

UPI

200Gb/s
(e.g. HDR)
200

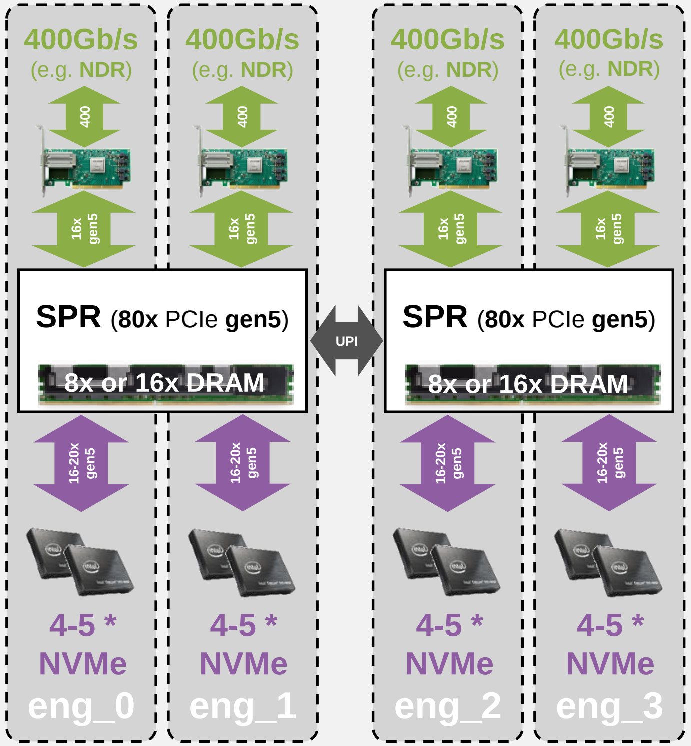
200Gb/s
(e.g. HDR)
200

16x gen5
SPR (80x PCIe gen5)
8x or 16x DRAM
8-24x gen5
2-7 * NVMe
eng_2

8-424 gen5
2-7 * NVMe
eng_3

Bandwidth: **≤92 GiB/s** per server

4th Gen Intel Xeon SP
4 * 400 Gbps



400Gb/s
(e.g. NDR)
400
16x gen5

400Gb/s
(e.g. NDR)
400
16x gen5

SPR (80x PCIe gen5)
8x or 16x DRAM
16-20x gen5
4-5 * NVMe
eng_0

16-20x gen5
4-5 * NVMe
eng_1

UPI

400Gb/s
(e.g. NDR)
400
16x gen5

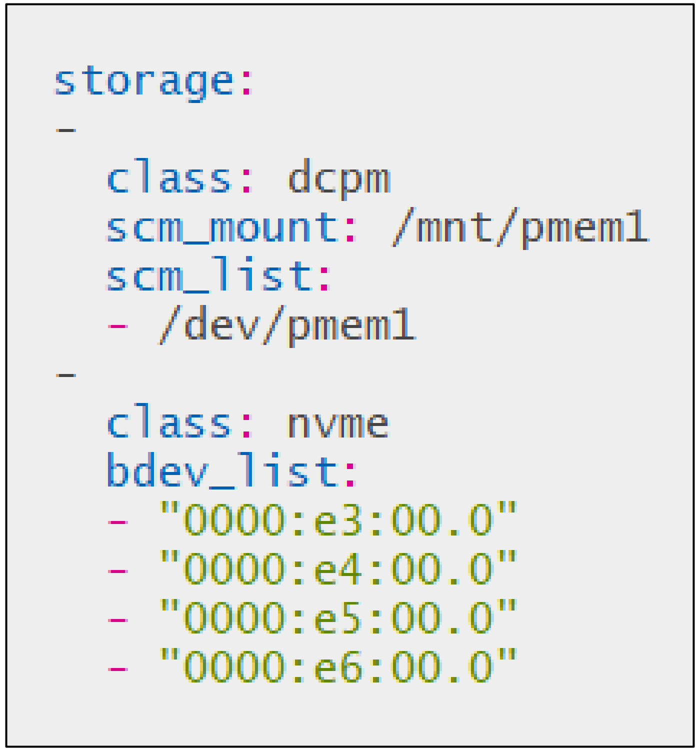
400Gb/s
(e.g. NDR)
400
16x gen5

SPR (80x PCIe gen5)
8x or 16x DRAM
16-20x gen5
4-5 * NVMe
eng_2

16-20x gen5
4-5 * NVMe
eng_3

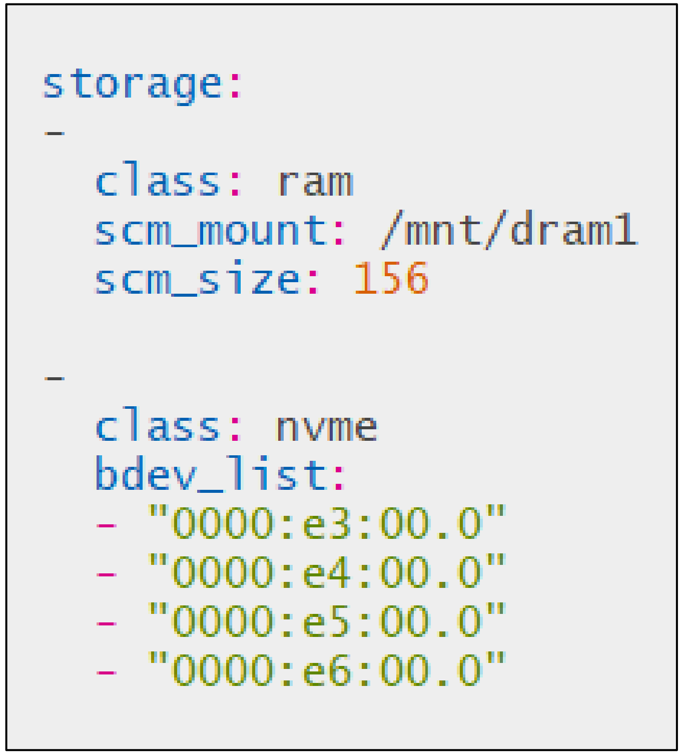
Bandwidth: **≤184 GiB/s** per server

“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"
```

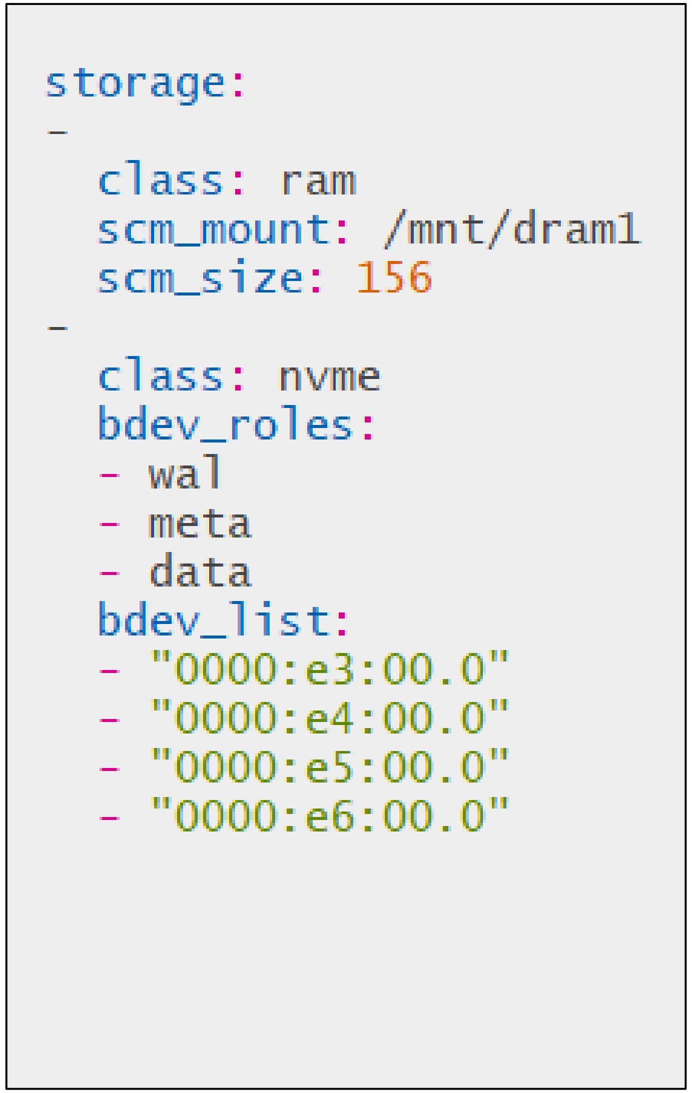
PMem-based DAOS



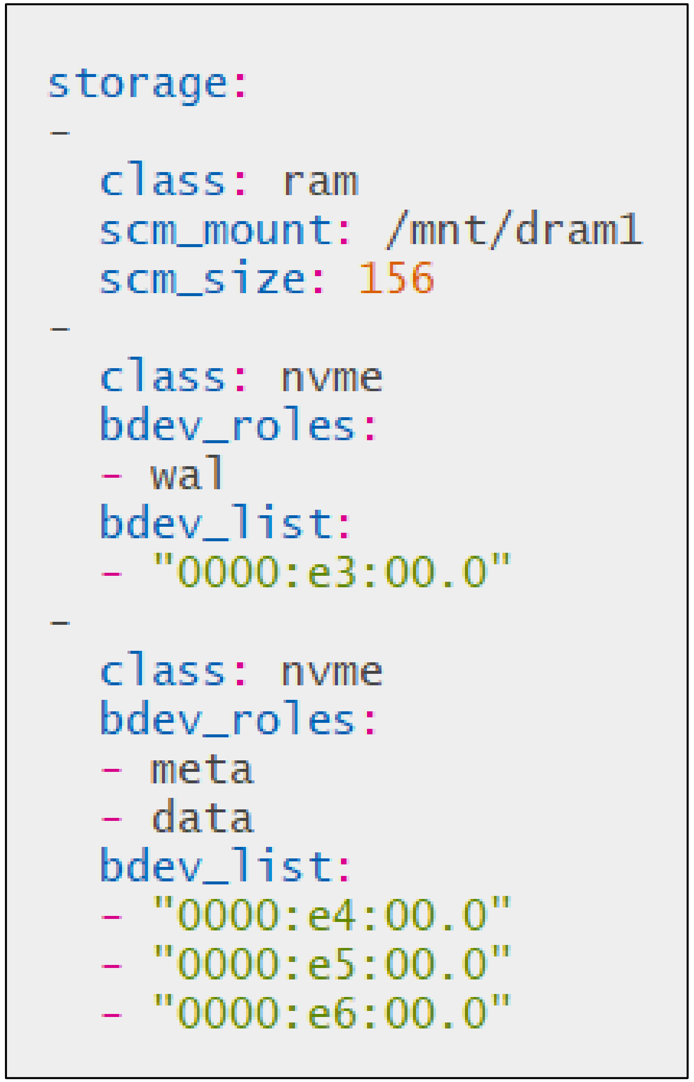
```
storage:
-
  class: ram
  scm_mount: /mnt/dram1
  scm_size: 156

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

“Ephemeral” DAOS

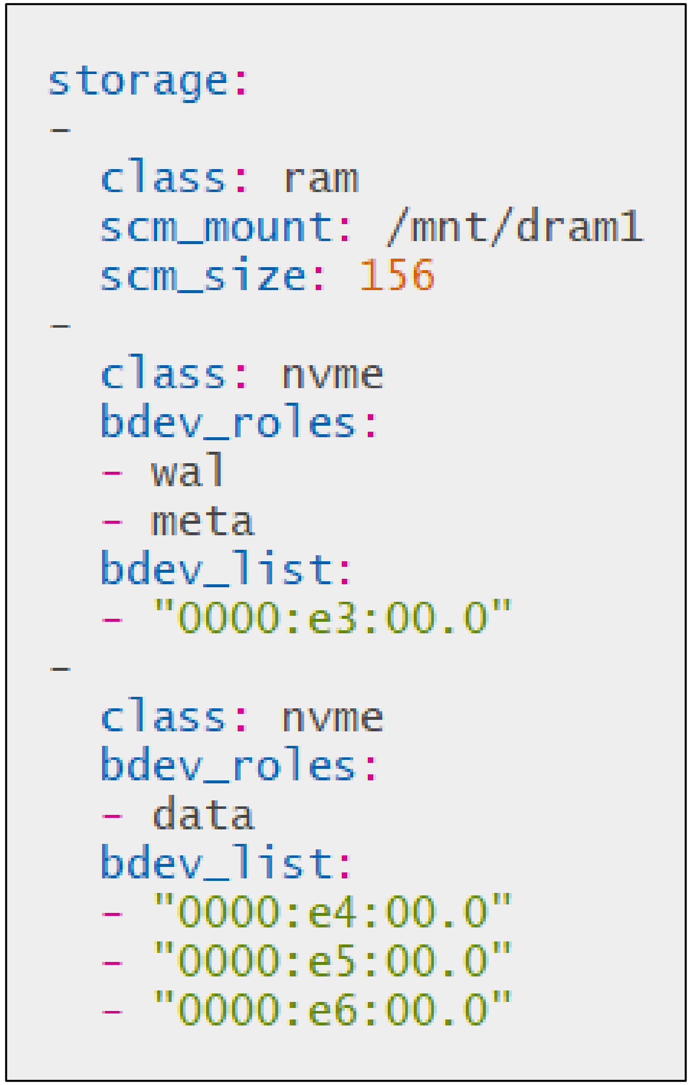
“MD-on-SSD” Configuration Options in daos_server.yml



```yaml
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"
```



```yaml
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"
```



```yaml
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

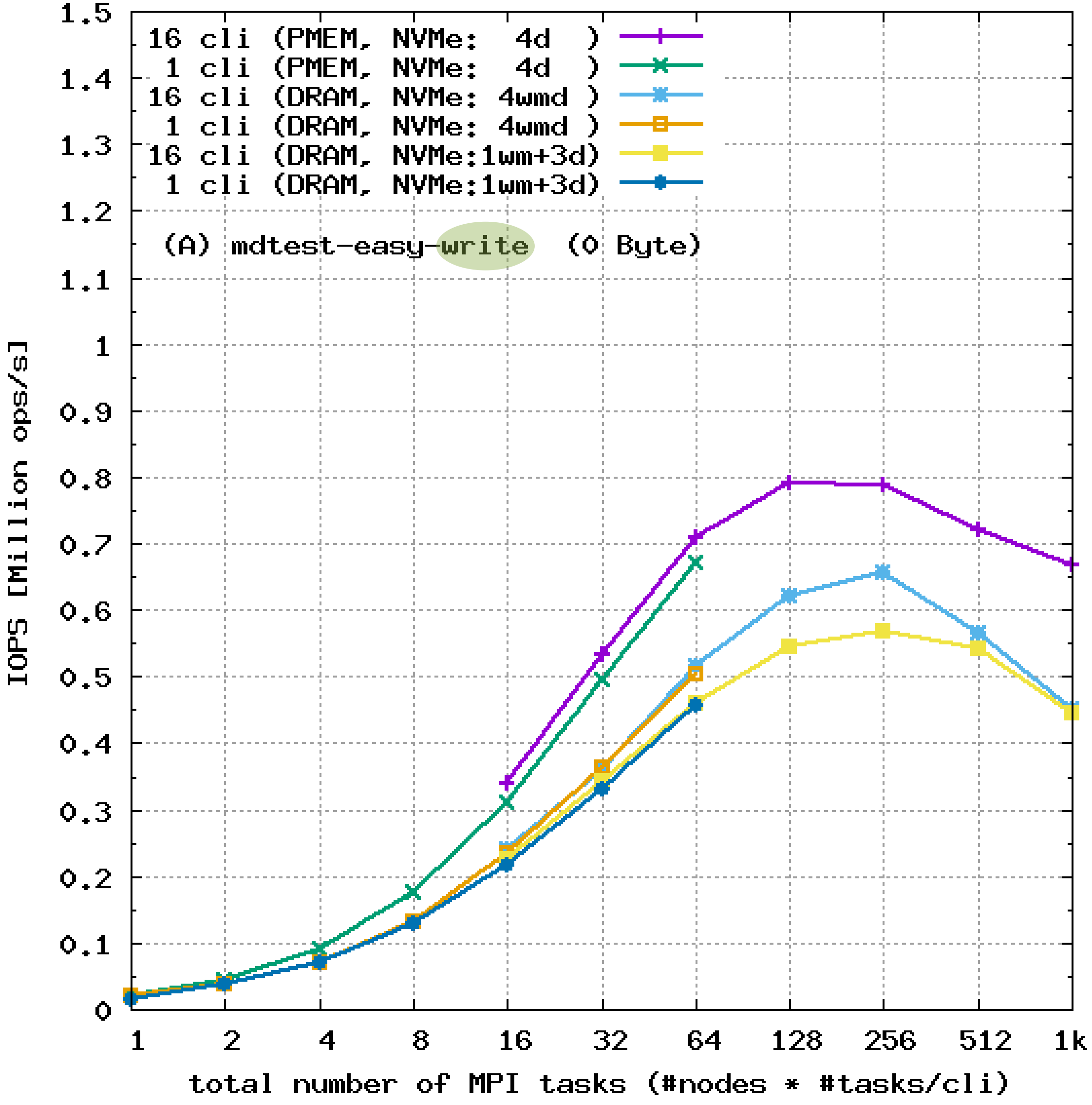
(1 engine @ 24 targets; HDR IB; 8TB pool; 30sec stonewall)

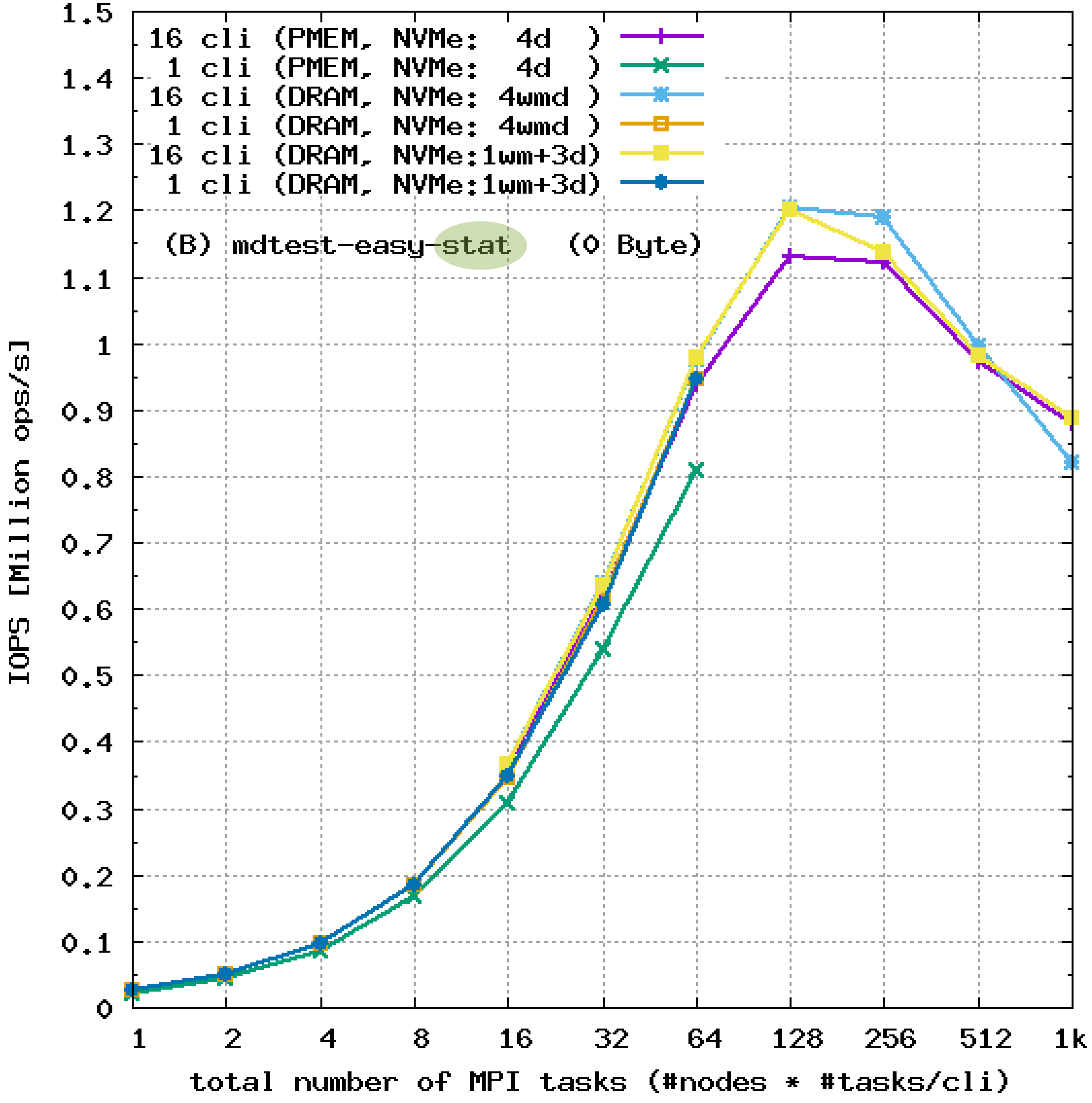
mdtest-easy (0-Byte files; dir-per-process)

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

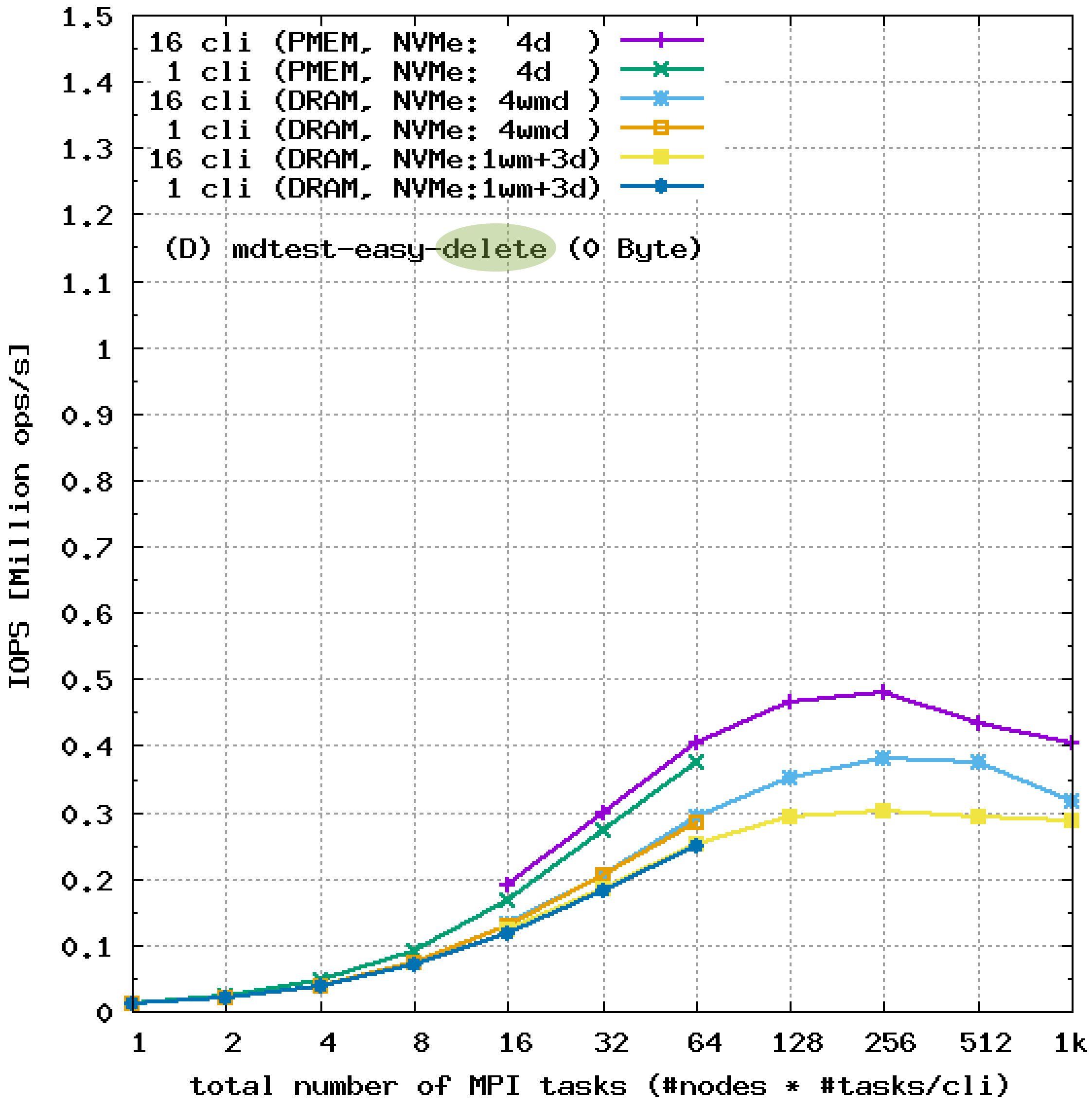
mdtest-hard2 (7802-Byte files; shared-dir)

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





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



Summary

- DAOS Metadata-on-SSD (Phase 1) is implemented
  - DAOS 2.4 tech preview; DAOS 2.6 generally available
  - Comparable performance to DAOS on PMem for mdtest-stat, mdtest-read. Some (≤20%) degradation for mdtest-write, mdtest-delete (synchronous WAL)
  - Usage/designation of NVMe devices depends on server config (perf. vs capacity, …)
- Future Phase 2 of MD-on-SSD: Enable migration of “cold” metadata from DRAM to “meta” blobs on NVMe
  - DAOS 2.8 tech preview; DAOS 3.0 generally available
  - Will reduce DRAM capacity requirements (as a percentage of NVMe capacity)
- ISC23 workshop paper: https://doi.org/10.1007/978-3-031-40843-4_26