Hewlett Packard
Enterprise

E2000 Performance From Microbenchmarks to Applications

Michael Moore, Principal Engineer
Bill Loewe, Sakib Samar, Chris Walker

May 08, 2025

Agenda

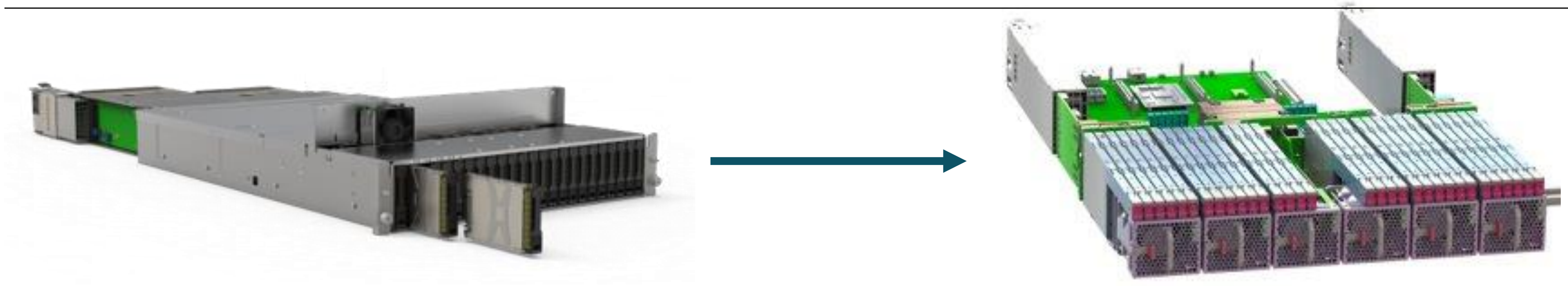
E2000 Description

Microbenchmarks

Applications on E2000

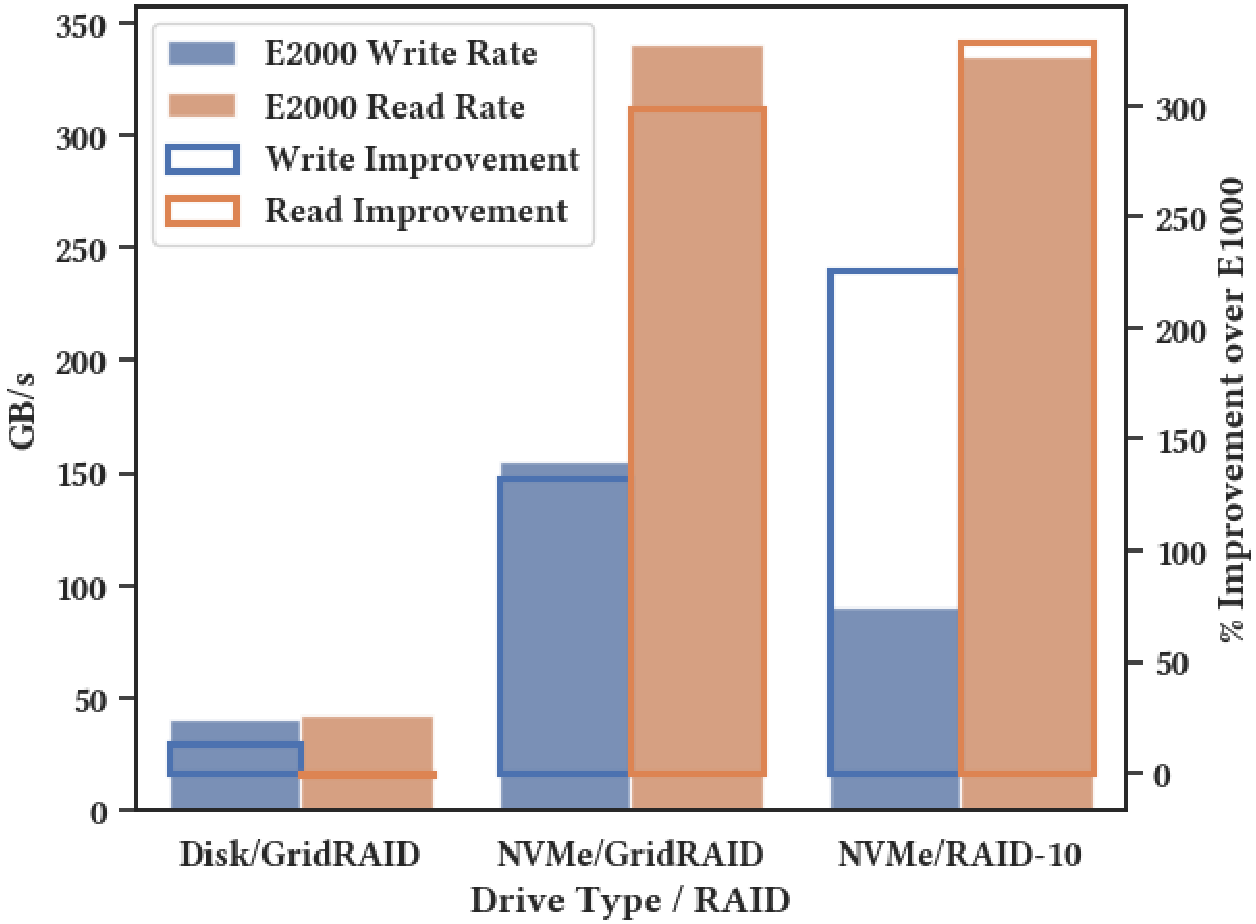
E1000 and E2000 Enclosure Description

Component	E1000	E2000
CPU	AMD 7502P, 2.5 GHz, 32 cores	AMD 9345P, 3.25 GHz, 32 cores
Memory	DDR4, 8 channels, 256 GB	DDR5, 12 channels, 384 GB
PCIe	Gen 4	Gen 5
NVMe devices	24 x U.2	32x ES.3
Primary NVMe Drive	Samsung 1733a, 7.5 GB/s read, 4.1 GB/s write	Samsung 1743, 14.0 GB/s read, 6.0 GB/s write
Network Connectivty	200 Gbps per Interface	Up to 400 Gbps per Interface
Targets	2 storage targets	4 storage targets



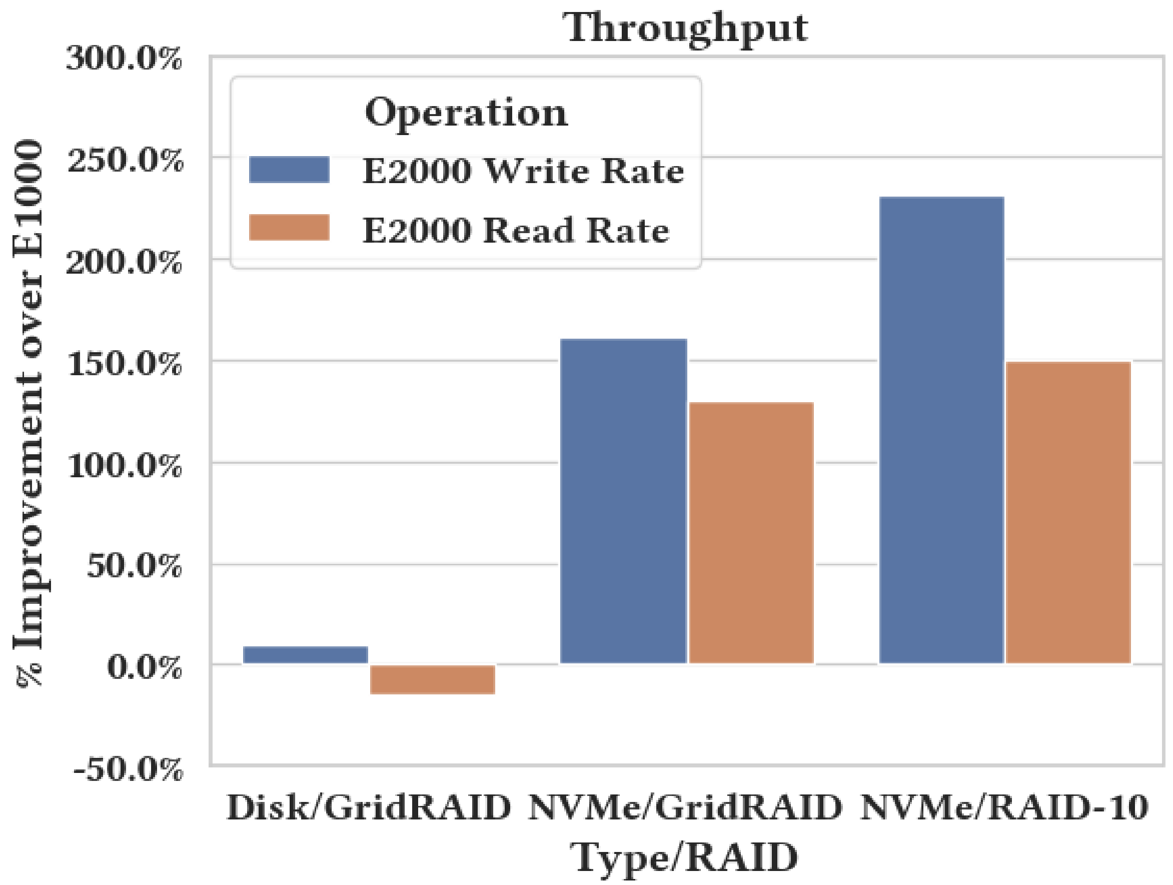
E2000 RAID Scalable Unit Performance

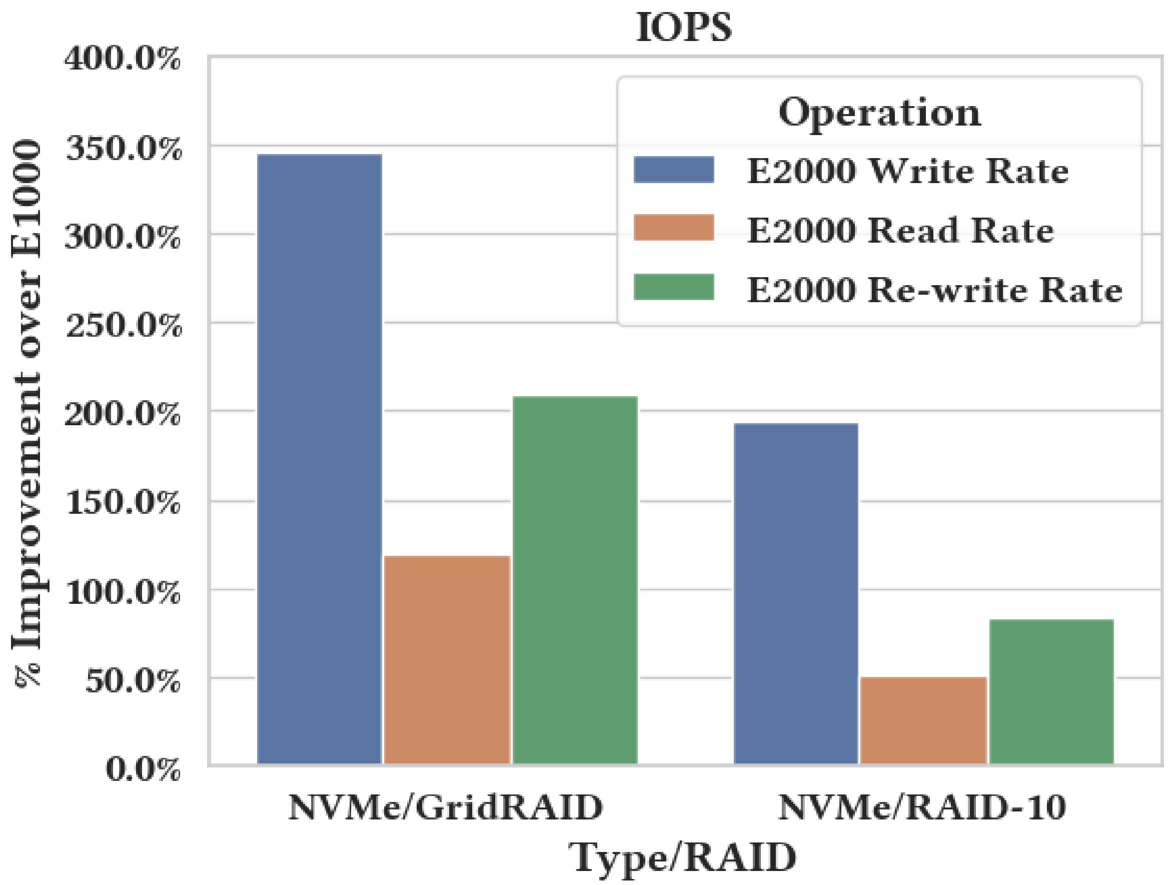
- NVMe based OSTs on the E2000 with ldiskfs show local (obdfilter-survey) performance
  - Greater than only drive quantity and per-drive performance
  - Write performance is drive limited
  - Read performance exceeds available network connectivity (>200 GB/s)



E2000 Throughput and IOPS

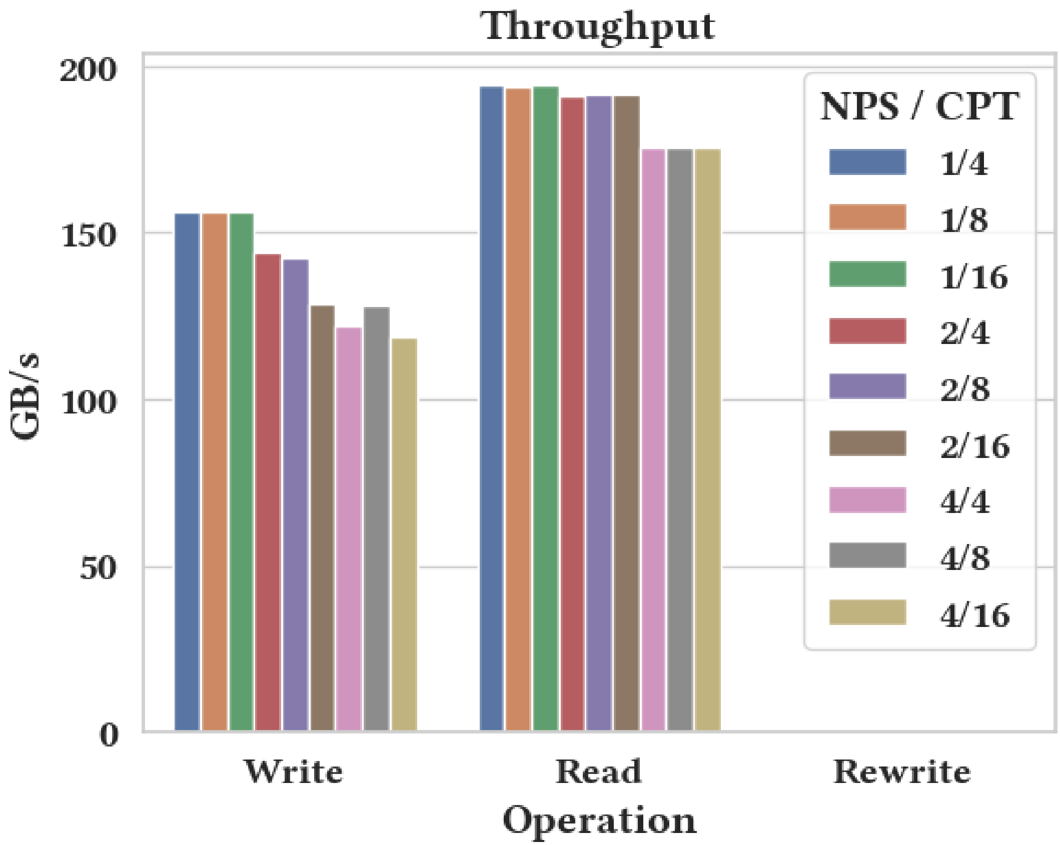
- IOR E2000 Improvement Relative to E1000 using ldiskfs

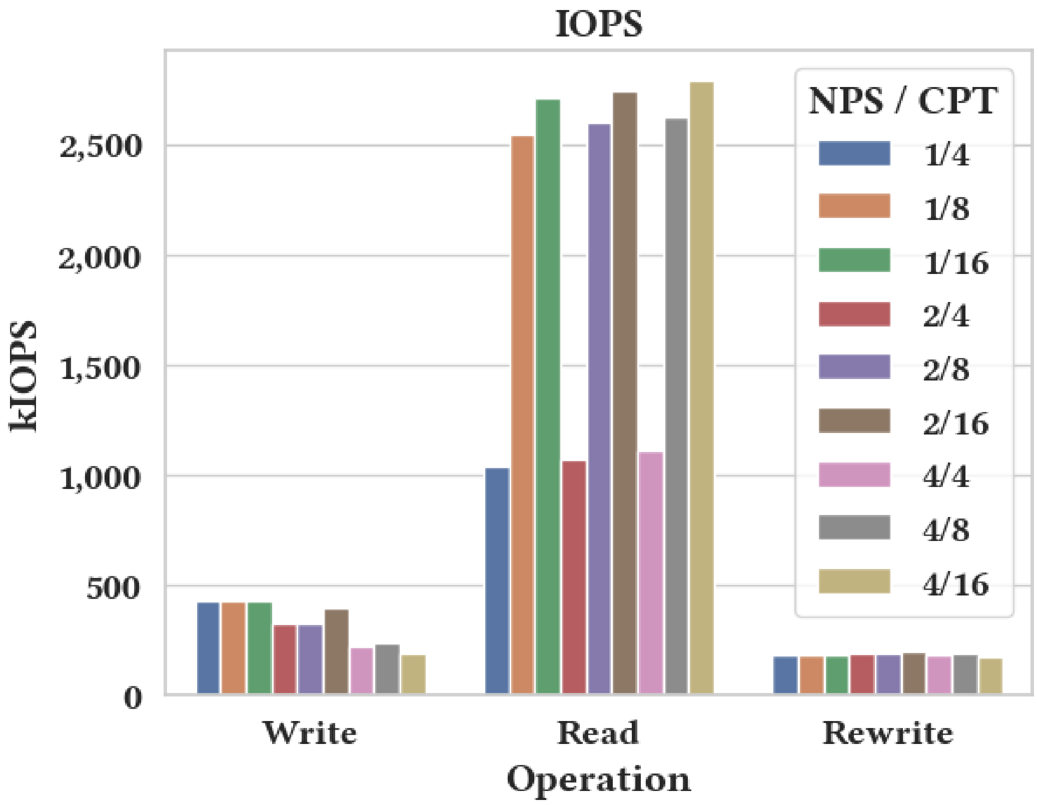




E2000 NPS and CPT Comparison, ldiskfs on GridRAID

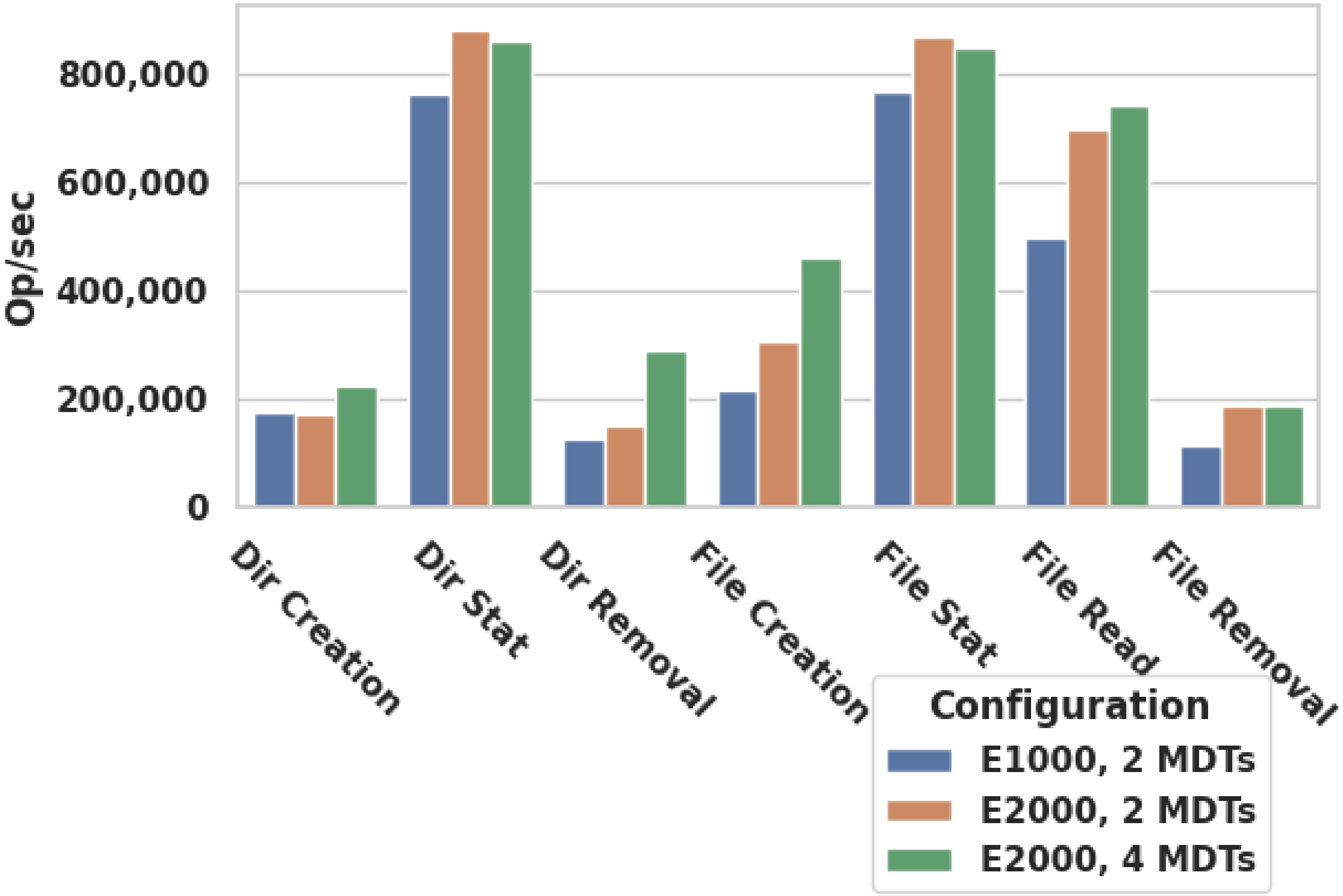
- NPS and CPT each influence performance
  - NPS=1 provides better throughput for write and read
  - CPT=8 or 16 provides dramatically better read IOPS





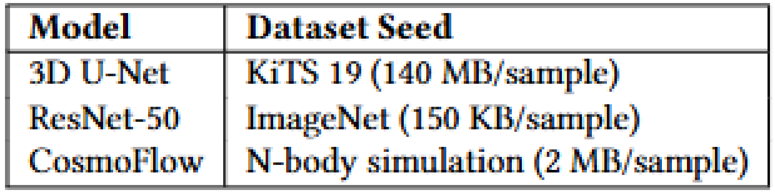
E2000 Metadata Performance

- Unique Working Directory mdtest, 0-byte files

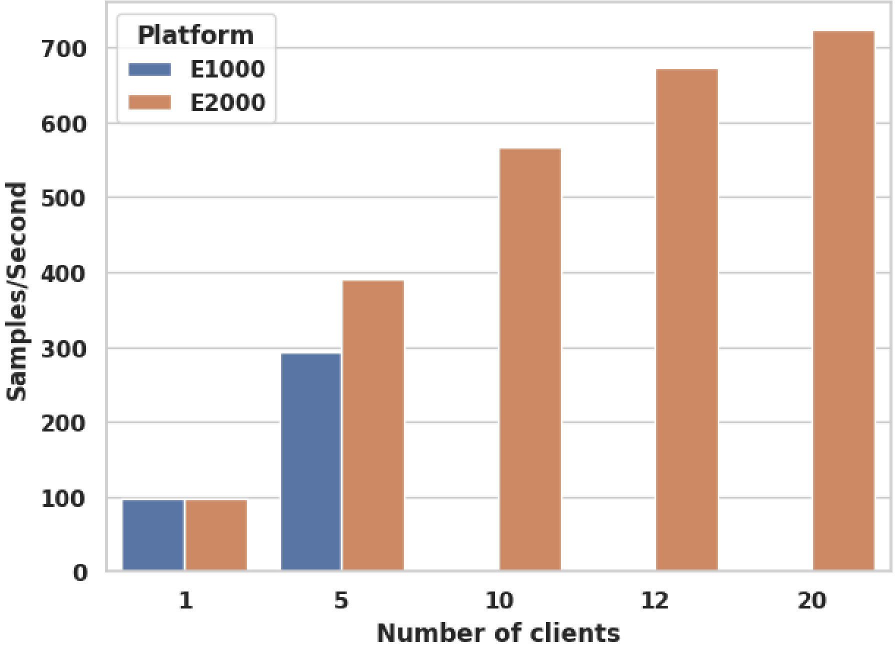


E2000 MLPerf Storage

- The MLPerf Storage Benchmark, provided by MLCommons, measures how efficiently a storage system can supply data to the AI training processes and achieve high throughput and utilization.
- The table on the right shows the details of each of the workloads, including the IO sizes.
- The figure shows the performance of the Unet3D workload using the E2000 and comparing it against the submitted results of the E1000.
- The E2000 obtained higher throughput and samples/second when running the Unet3D workload, and it scaled further with a higher client count.
- The details of each workload, the performance metric, and the passing criteria can be found in the MLCommons website.

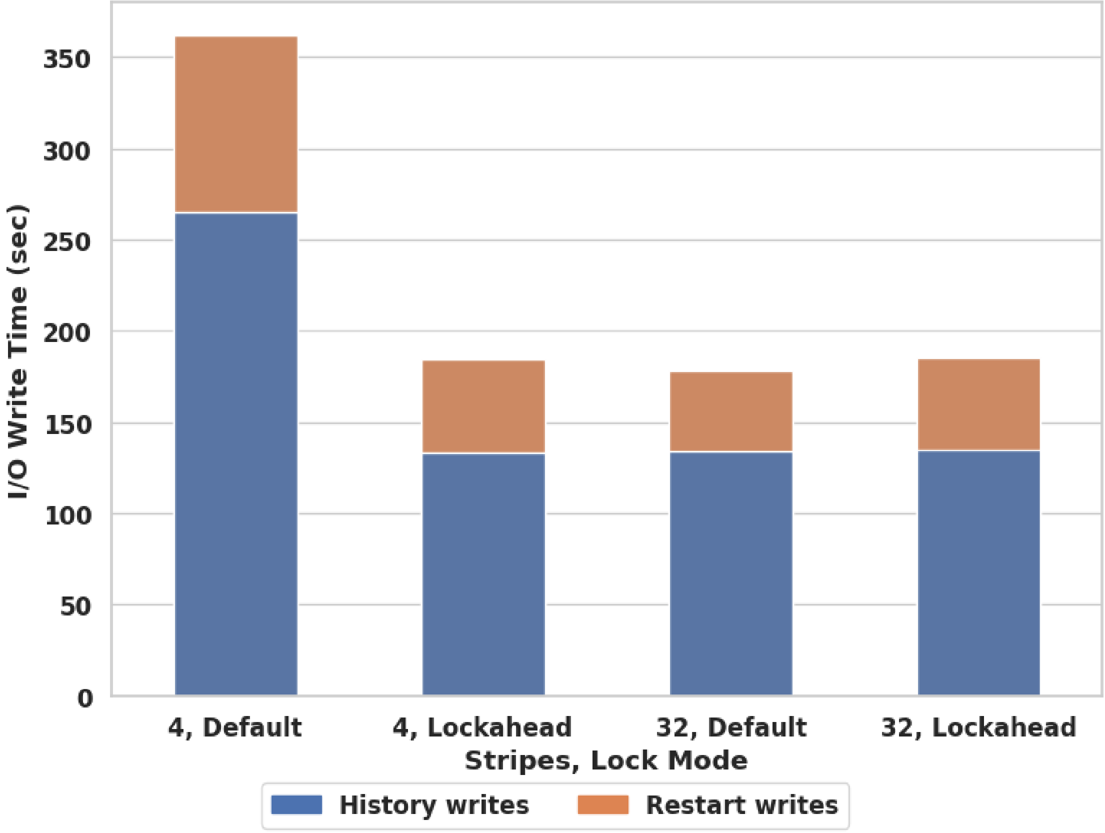


Model	Dataset Seed
3D U-Net	KiTS 19 (140 MB/sample)
ResNet-50	ImageNet (150 KB/sample)
CosmoFlow	N-body simulation (2 MB/sample)



Weather Research and Forecasting (WRF) on E2000

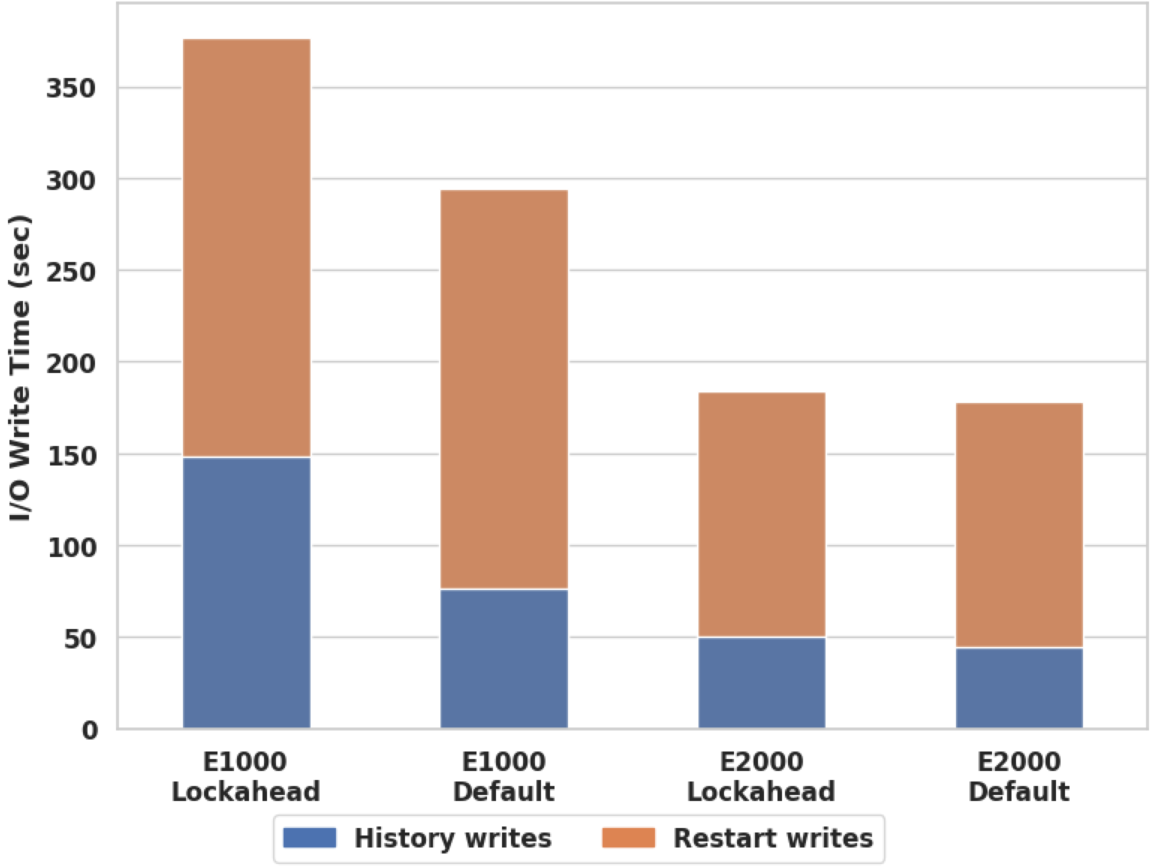
- I/O Write time is the cumulative sum of WRF reported history and restart file time across all files in the job
- Collective MPI-IO writes with Collective Buffering (CB)
  - WRF History and Restart writes:
    - Lustre striping:
      - Traditional, 1 stripe on each of 4 OSTs
      - Overstriping: 8 stripes on each of 4 OSTs
    - 32 nodes with one CB aggregator rank per node
      - For 4 stripes: `cray_cb_nodes_multiplier=8`
      - For 32 stripes: default MPI-IO hints
    - Lustre Locking
      - Default: default MPI-IO hints (lock mode 0)
      - Lustre Lockahead : `cray_cb_write_lock_mode=2`
- Lockahead 5-10% slower than optimal but scalable



WRF Comparison between E1000 and E2000

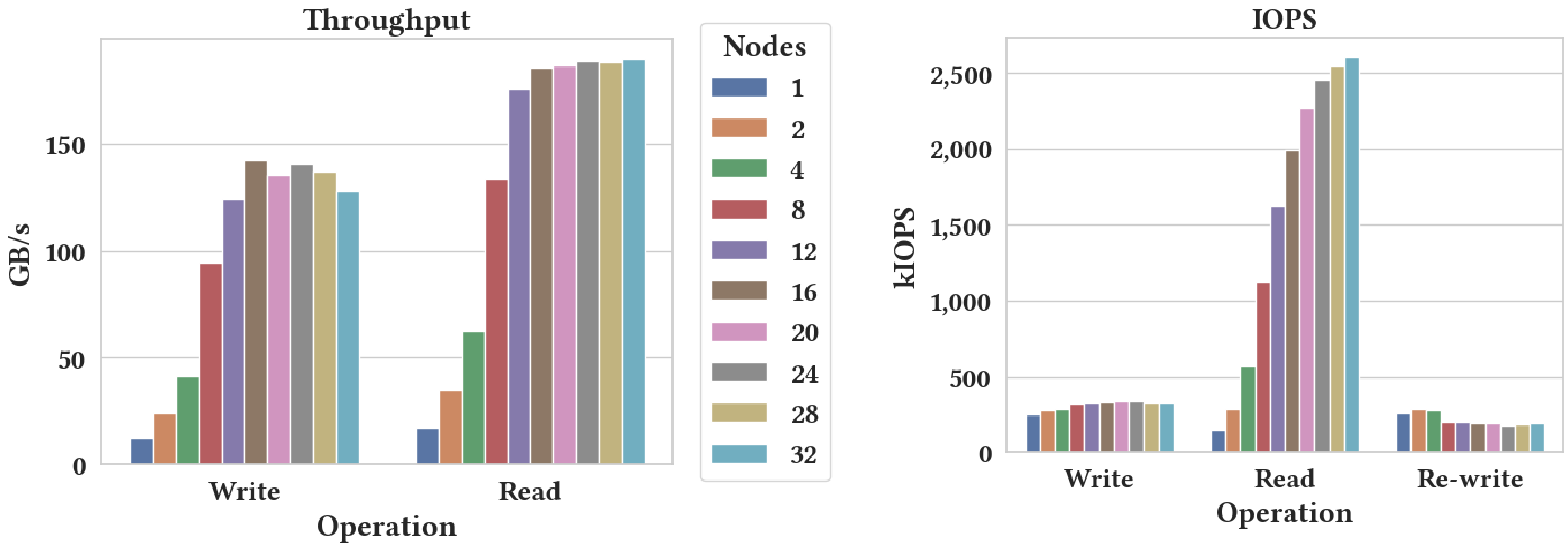
- Comparison of same WRF use case on two different systems
  - E1000, Slingshot 200, dual socket Milan CPU
  - E2000, IB NDR, single socket Genoa CPU
  - Only comparing file write time but CPU could influence performance
  - 1 x E1000-F (2 OSTs) to 1 x E2000-F (4 OSTs)
- Optimal 32 aggregator striping
  - 52% reduction in I/O write time
- Scalable 32 aggregator striping
  - 39% reduction in I/O write time

- Same % of application time spent on I/O
  - For ~50% reduction in runtime, still 12%-13% for I/O



Application Changes for E2000

- Metadata and Disk-based OSTs shouldn’t require changes
- E2000-F has similar per OST performance but twice the number of OSTs
- 12-16 nodes per E2000-F for throughput and write IOPS and 32 nodes for read IOPS*



**8 PPN for throughput, 24 PPN for IOPS*

Conclusion

- E2000 demonstrates significant improvement over the E1000 per scalable unit
  - E2000-F:
    - > 150% write and 100% read throughput improvement over E1000-F (GridRAID/ldiskfs)
    - > 300% write and 100% read IOPS improvement over E1000-F (GridRAID/ldiskfs)
  - E2000 MDU:
    - > 110% file create, 10% file stat, and 66% file removal improvement over E1000 MDU (unique)
- For applications migrating to E2000
  - Double the compute nodes for E2000-F scalable unit; similar node count per NVMe-based OST

Thank you

Bill Loewe, william.loewe@hpe.com
Michael Moore, michael.moore@hpe.com
Sakib Samar, sakib.samar@hpe.com
Chris Walker, chris.walker@hpe.com