New Challenges of Benchmarking All-Flash Storage for HPC

National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory
Berkeley, California, USA

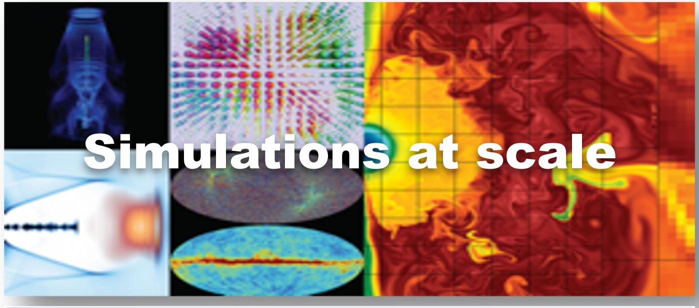
Glenn K. Lockwood
Alberto Chiusole
Nicholas J. Wright

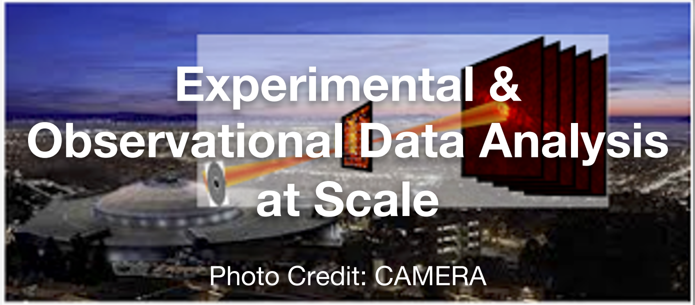
November 15, 2021

NERSC is the mission HPC facility

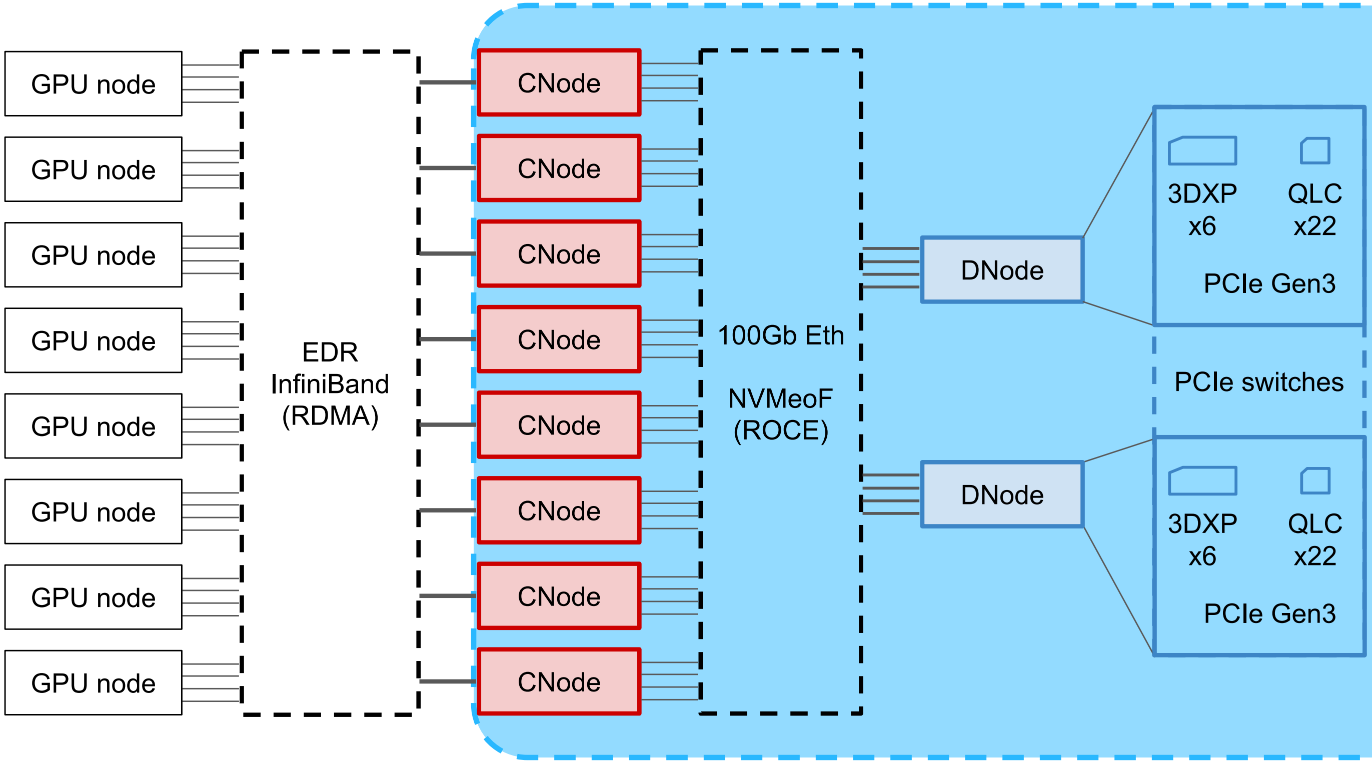
for DOE Office of Science

- Diverse workload type and size:
  - *Many* science domains
  - Experimental/AI-driven workloads
  - 7,000 active users, 700 apps
- Checkpoint/restart only part of the picture
  - New all-flash file systems make big promises
  - What does evaluating these technologies look like in practice?



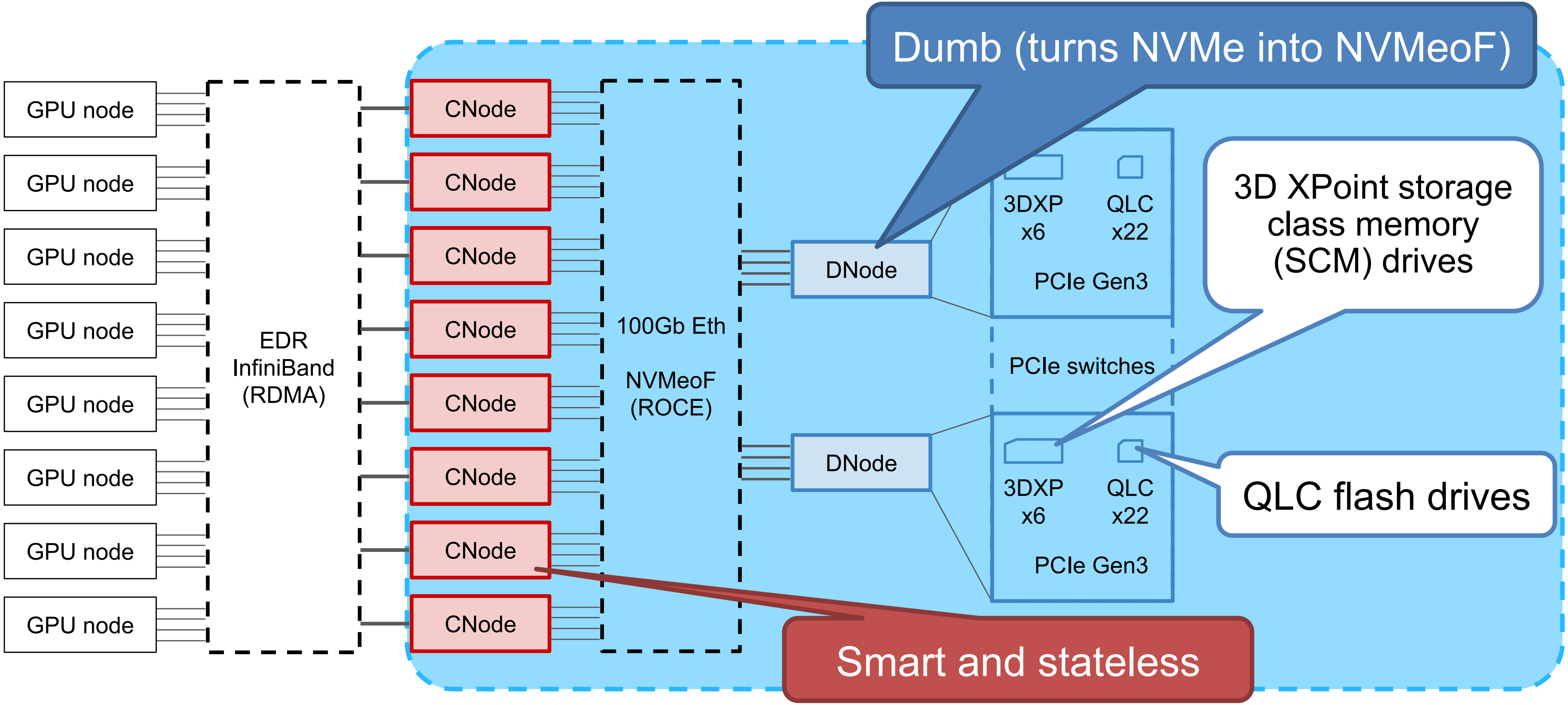


Test system – VAST Universal Storage



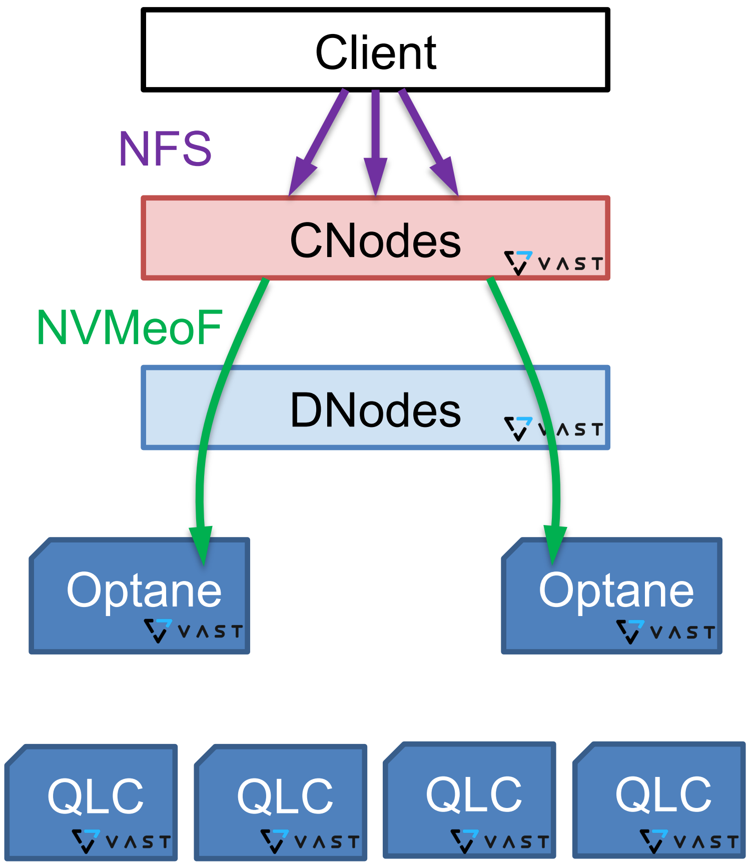
- 620 TB
- 40 GB/s read
- 5 GB/s write
- 168 KIOPS write
- 336 KIOPS read

Test system – VAST Universal Storage



VAST write path in a nutshell

1. Client write goes to any "CNode" (no locality)
2. CNode replicates write to two Optane drives and bucketed based on LSH to build multi-GB stripes



VAST write path in a nutshell

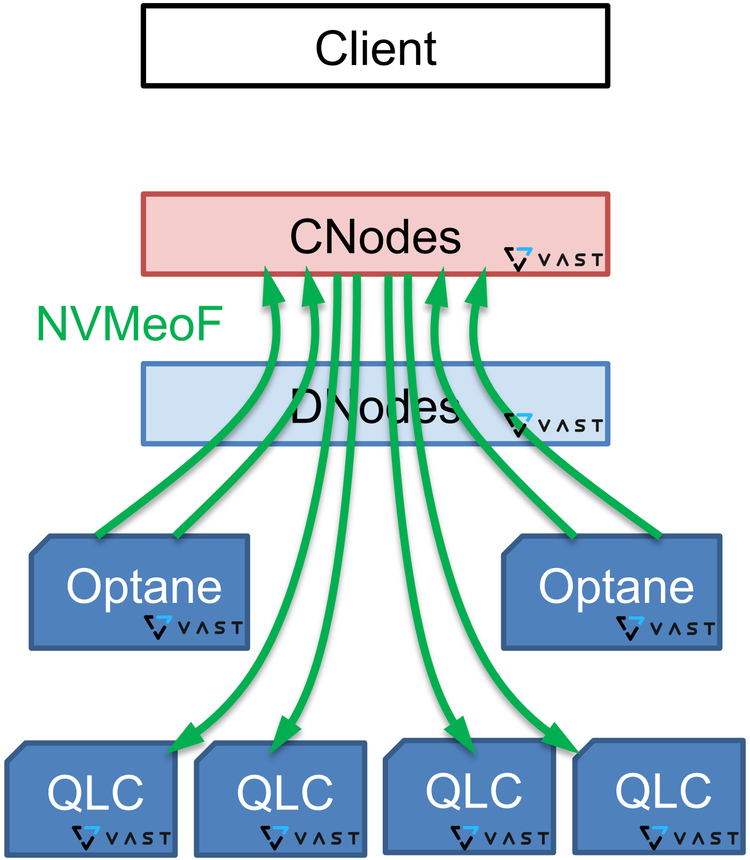
1. Client write goes to any "CNode" (no locality)
2. CNode replicates write to two Optane drives and bucketed based on LSH to build multi-GB stripes

sync

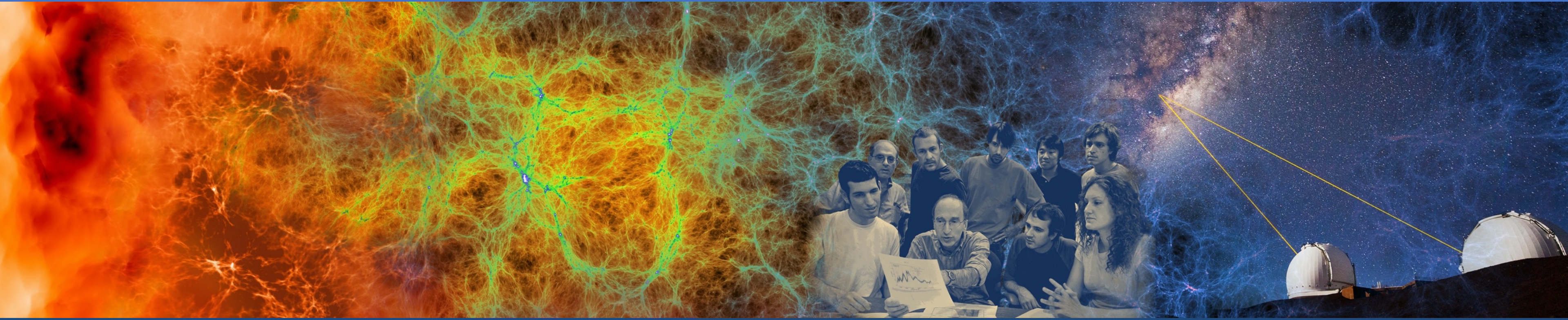
---

async

3. Full stripes are compressed, EC'ed and written to QLC

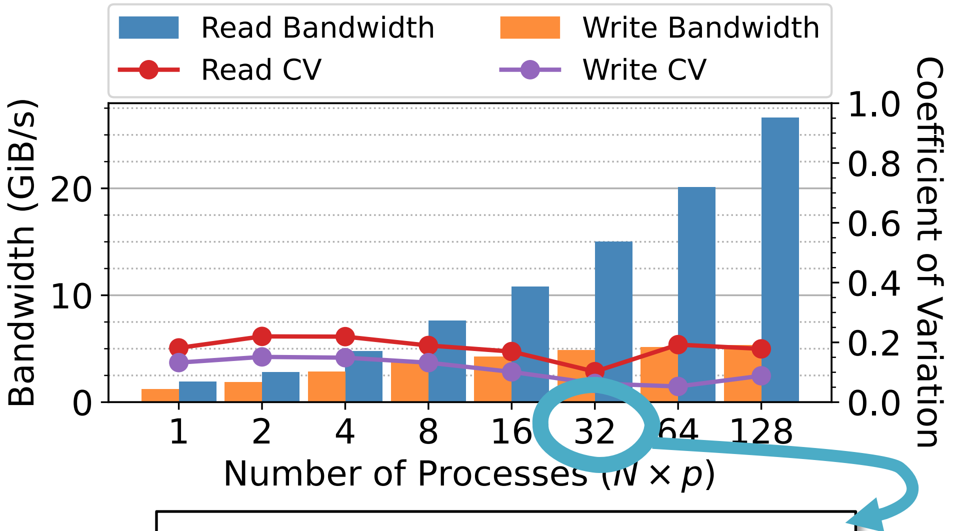


Bandwidth beyond the hero number



Testing performance versatility

- Run file-per-proc IOR write tests followed by read tests
- Test at many scales and I/O sizes
  - Node count $N = \{1, 2, 4, 8\}$
  - Procs/node $p = \{1, 2, 4, 8, 16\}$
  - I/O size $t = \{4\ Ki, 512\ Ki, 1\ Mi, 4\ Mi, 8\ Mi, 32\ Mi\}$
- Express performance as averages

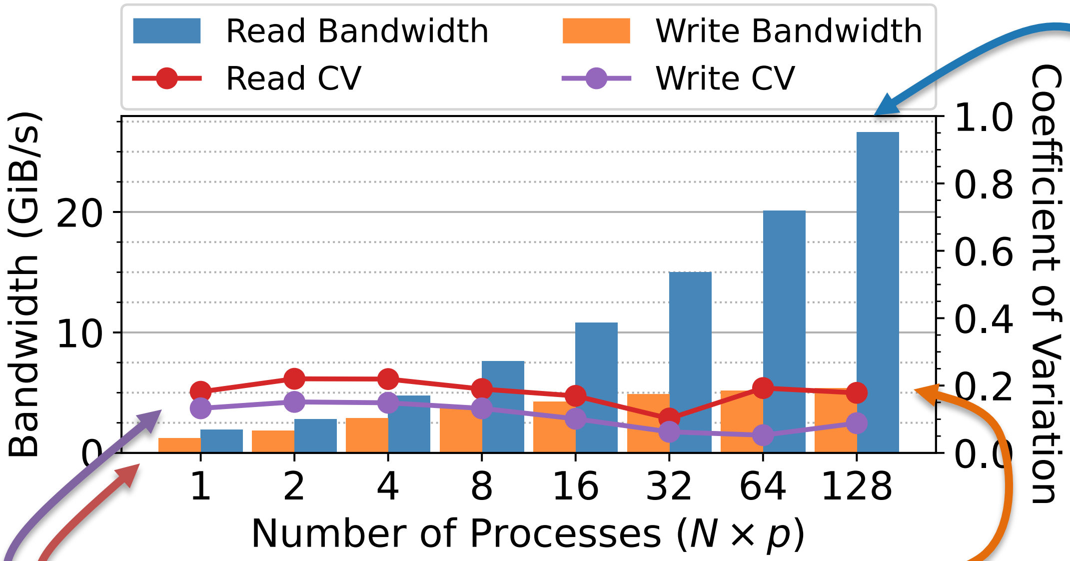


$x = 32$ includes mean of 5 tests *each*:

- $(N = 8, p = 4)$
- $(N = 4, p = 8)$
- $(N = 2, p = 16)$

for I/Os of sizes $t = \{4\ KiB, \ldots, 32\ MiB\}$

Sequential I/O performance tested naïvely



Average read bandwidth scales well for all N, p, t;

- $\mu > 25$ GB/s (with NFS!)
- Good for data analytics

Could add more, smaller Optane drives to shift write:read performance

Reliable bandwidth independent of

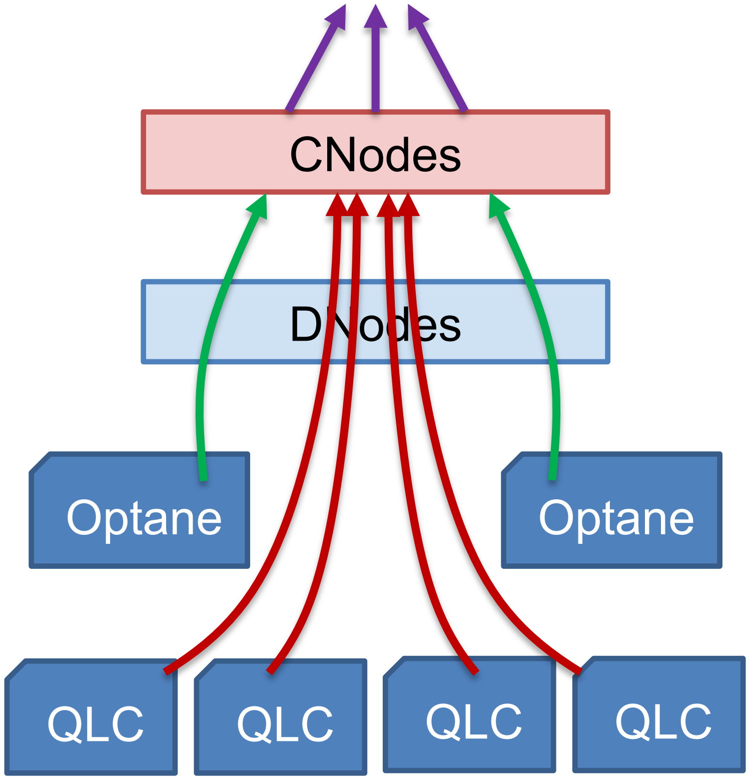
- I/O size (t = 4 KiB to 32 MiB)
- N vs. p

Average write bandwidth saturates at 5 GB/s:

- synchronous replication
- writes must land on SCM (12 of 56 drives)

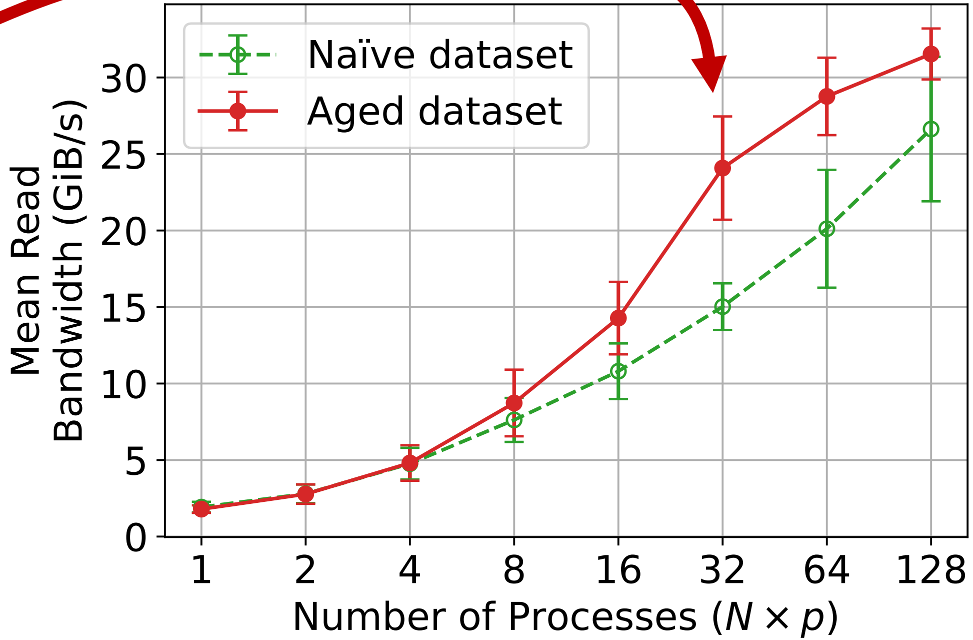
Sequential reads in hybrid SCM/QLC

- Reads can come from either
  1. **12× SCM (30 GB/s theoretical)**
  2. **44× QLC (140 GB/s theoretical)**
- What happens if we don't read-after-write?
  1. Step 1: Write data
  2. Step 2: Artificially age data (flush SCM with throwaway data)
  3. Step 3: Read data from step 1

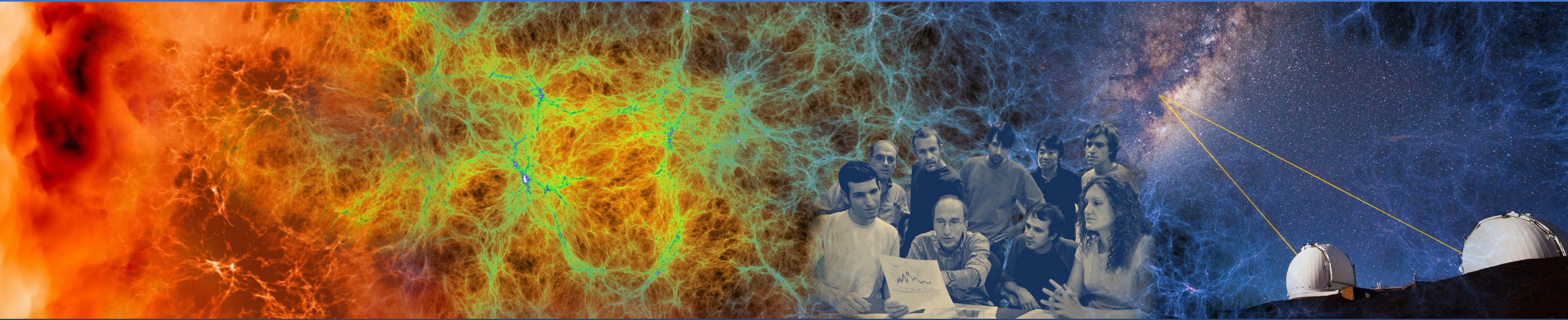


Reading aged data is faster in hybrid SCM/QLC

- Data gets faster as it ages!
  - > 50% higher read bandwidth
  - 44 QLC vs. 12 SCM SSDs
- Most user data is "aged"
  - NERSC: 2.2 PB/day for 35 PB file system (write 6% per day)
  - VAST uses ~0.5% capacity to receive new writes
  - Data is "old" after ~2 hours

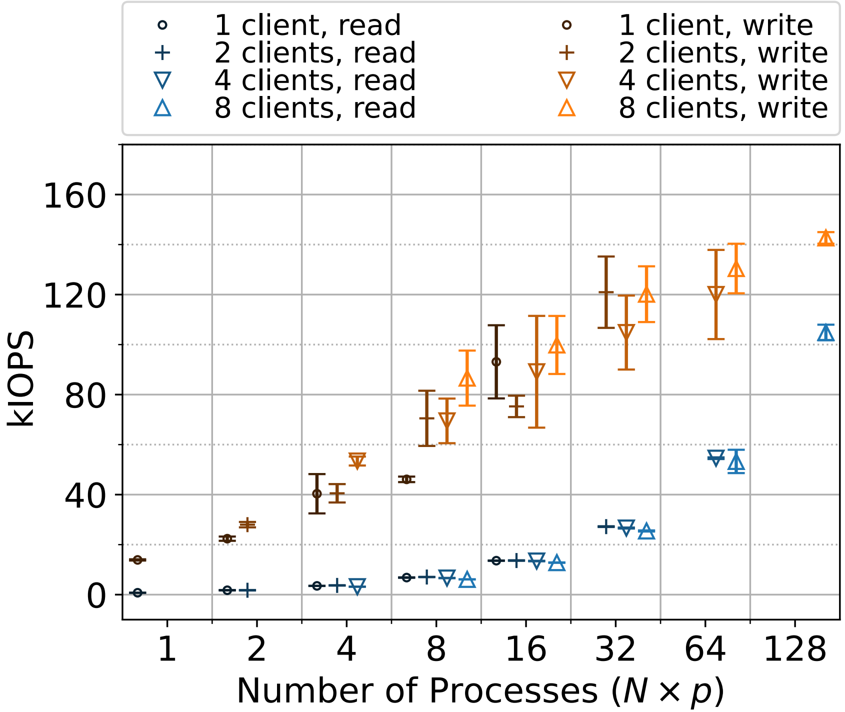


Measuring IOPS in a meaningful way



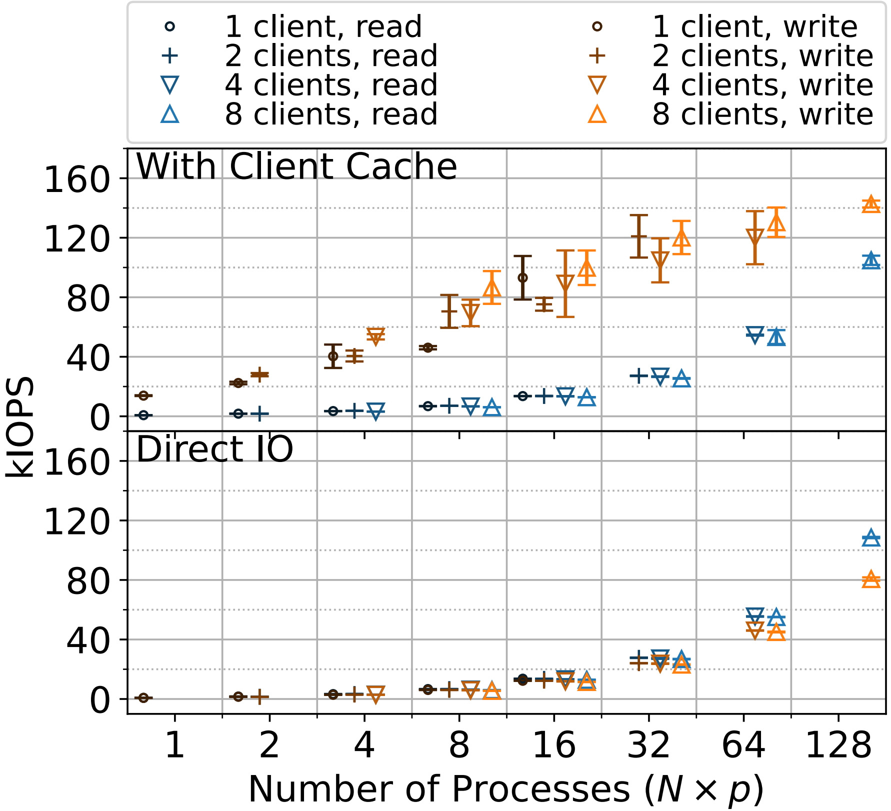
Measuring random I/O performance the normal way

- Run file-per-proc IOR
  - write 4 KiB at random offsets
  - read 4 KiB from aged dataset
- IOPS are insensitive to N vs p
- Read IOPS not close to saturation
- Write IOPS show
  - high peak performance
  - wide variation run-to-run
  - actually measuring write-back reordering performance



What is a random write anyway?

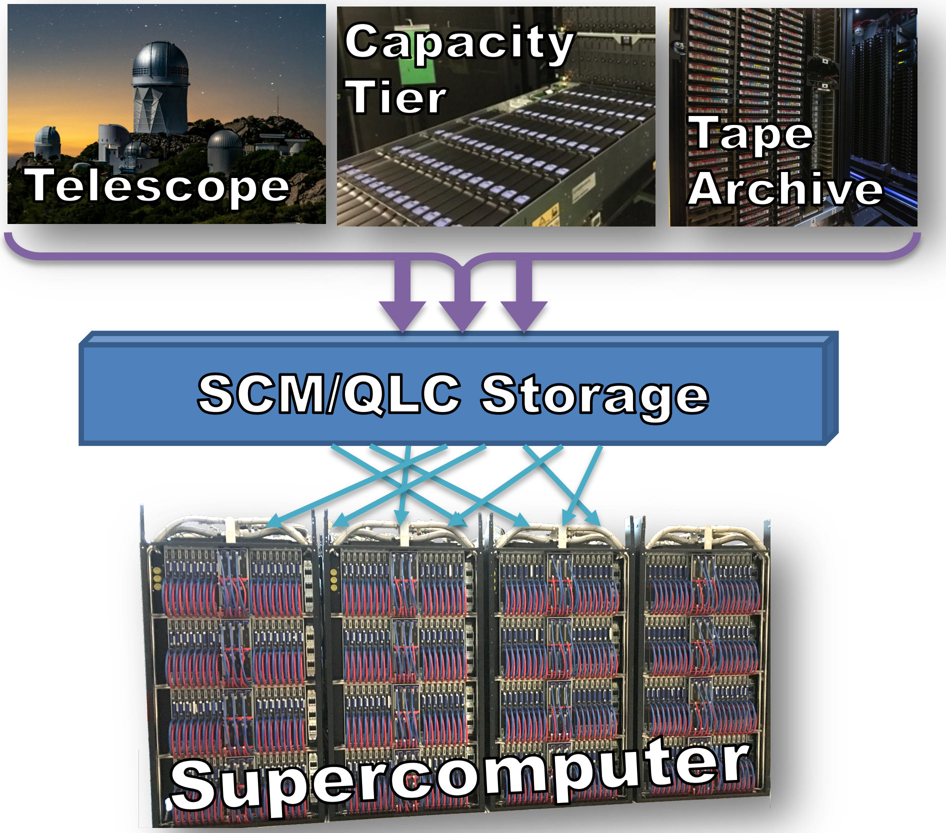
- O_DIRECT reduces apparent write IOPS
- Which is "true performance?"
  - True random writes are rare
  - Random, direct I/O is rarer
- **Application** performance should include write-back
- **System** performance is better measured with O_DIRECT



SCM/QLC + AI training workflow: it's complicated

Archetypal AI training workflow

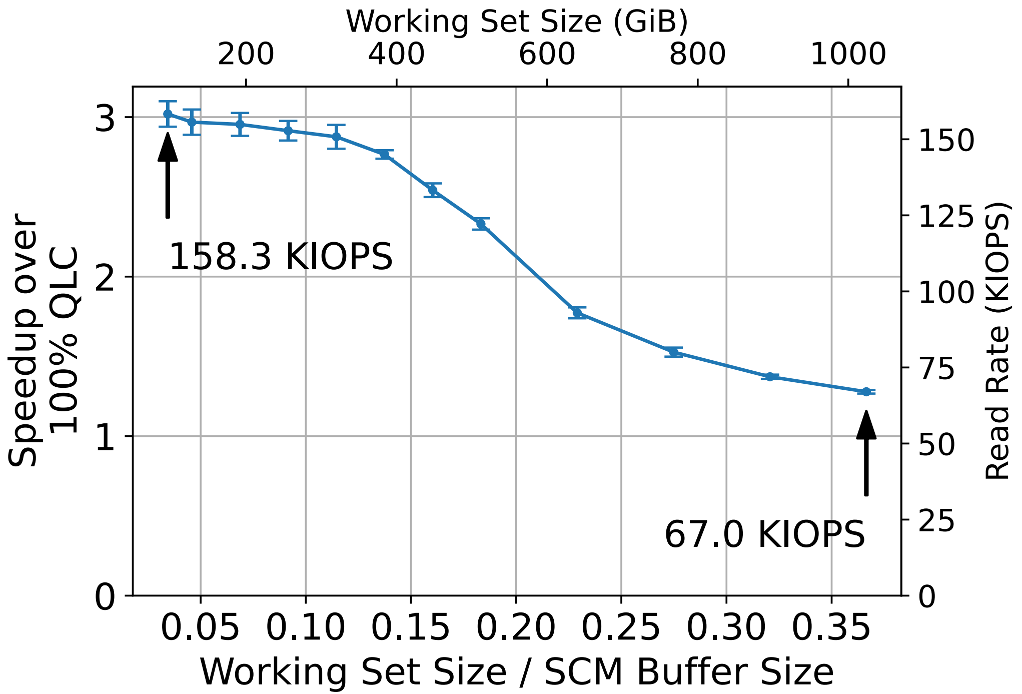
1. Data streamed into SCM/QLC storage
   - Origin: inside or outside of data center
   - I/O: large, sequential writes
2. Data randomly read
   - Begin immediately after step 1
   - I/O: intense random reads



Read IOPS depend on dataset size

- Datasets partly overflow from SCM to QLC
  - SCM: bandwidth ↓
  - QLC: bandwidth ↑
- Random read rate varies with dataset size

Could make IOPS more predictable with stage/punch/pin commands



Conclusions

We shouldn't benchmark all-flash with methods developed for HDD!

<u>Performance versatility</u>

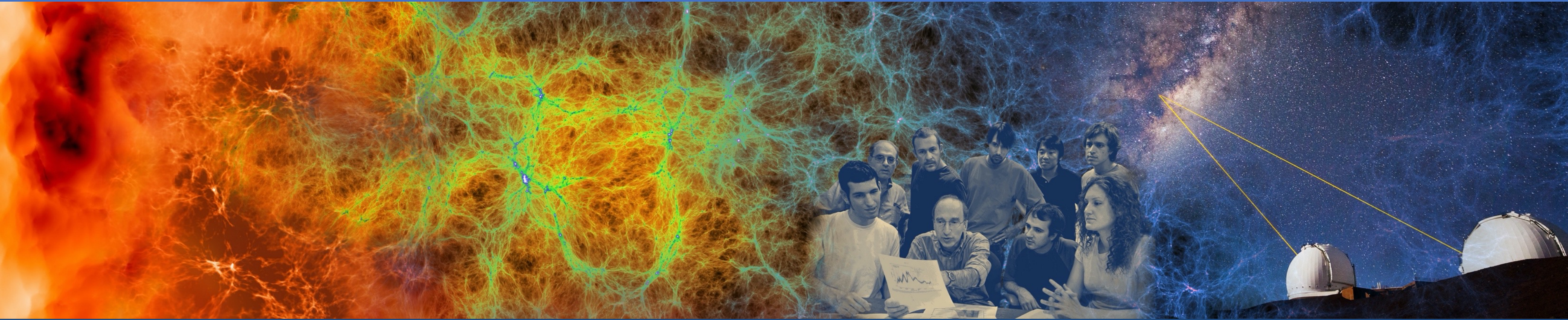
- All-flash can give consistent bandwidth at all I/O sizes – so measure them
- No one "true" value for IOPS – consider: app or system?

<u>SCM/QLC complicates performance analysis</u>

- Reading "new" data can be misleading!
- "New" data has lower sequential but higher random performance
- "Aged" data has higher sequential but lower random performance

Thank you!

This material is based upon work supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-05CH11231. This research used resources and data generated from resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.



NERSC BERKELEY LAB U.S. DEPARTMENT OF ENERGY Office of Science