Architecture and Performance of the Perlmutter 35 PB All-NVMe Lustre File System at NERSC

National Energy Research Scientific Computing Center

Lawrence Berkeley National Laboratory

Berkeley, CA USA

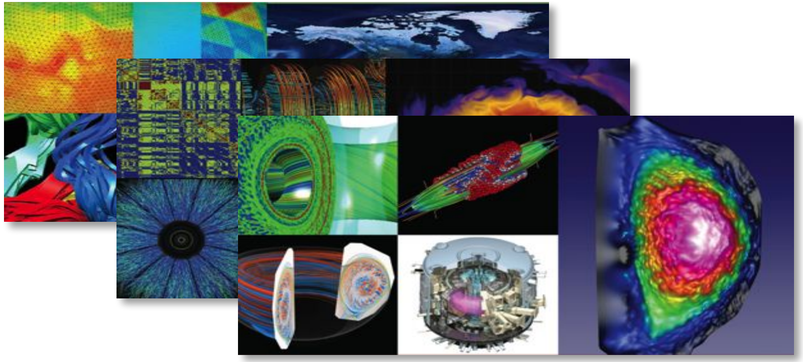
Alberto Chiusole, Lisa Gerhardt, Glenn K. Lockwood, Kirill Lozinskiy, David Paul, Nicholas Wright

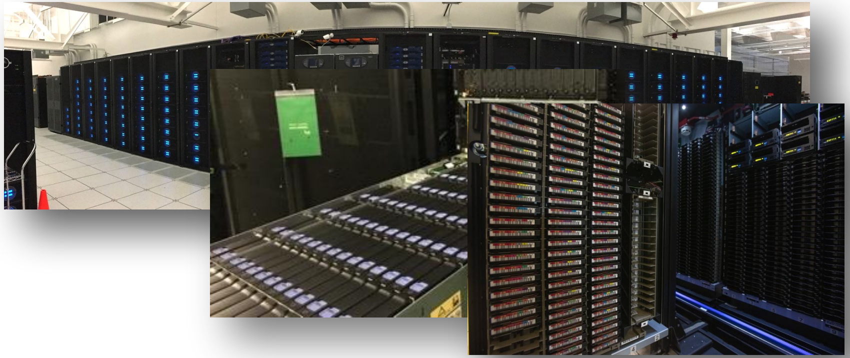
SC21: Analyzing Parallel I/O

November 16, 2021

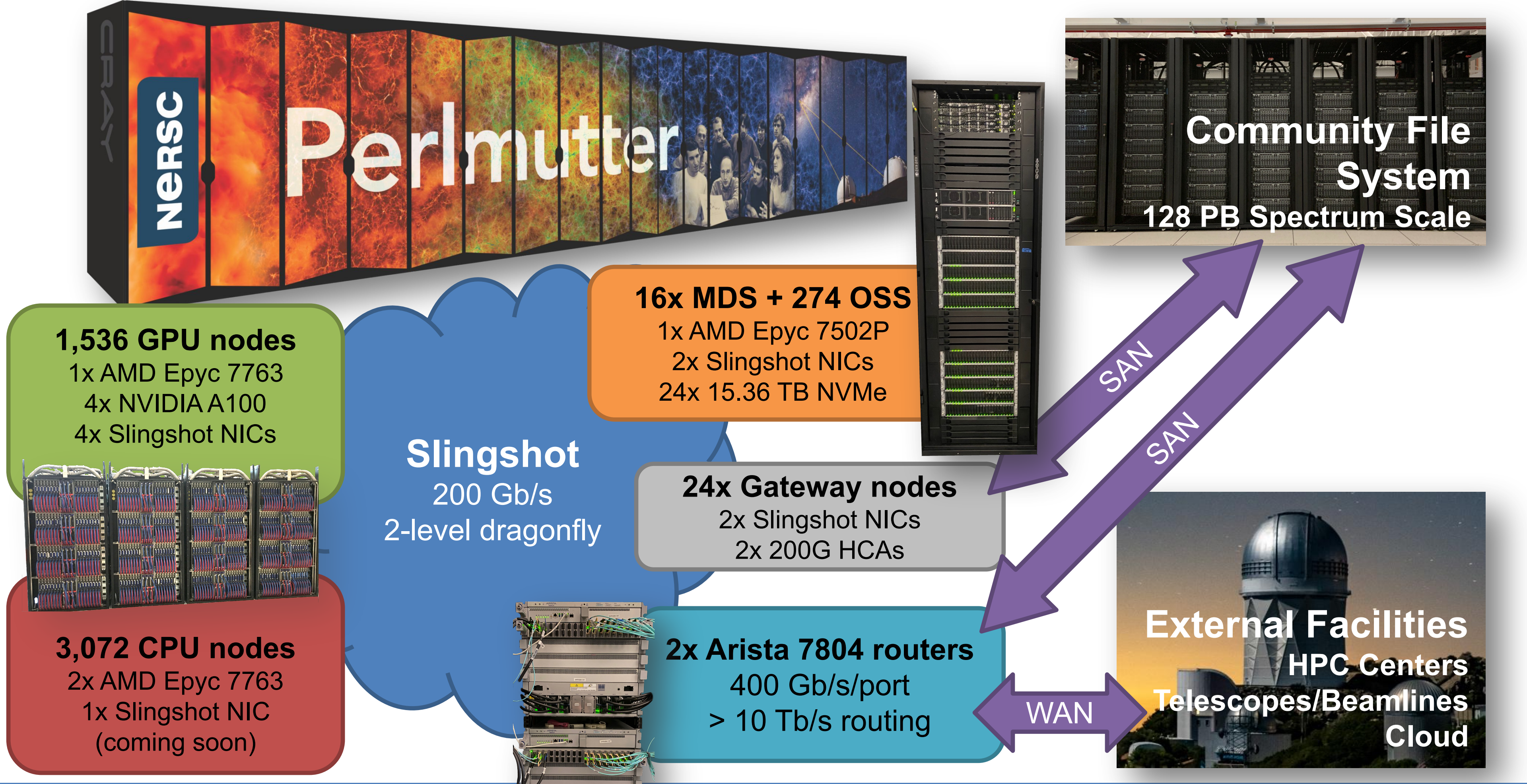
NERSC is the mission computing facility

for the U.S. Department of Energy Office of Science





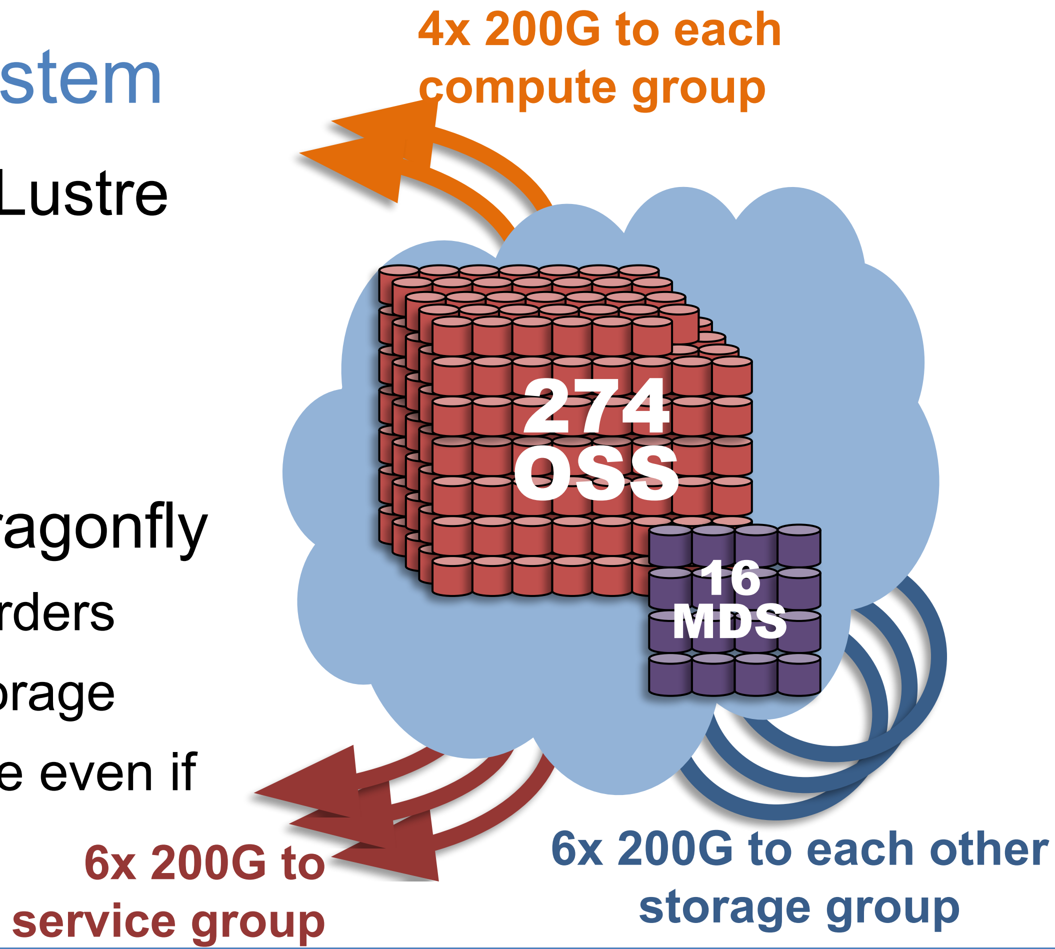
- **Diverse user community**
  - 8,000 active users, 900 projects
  - 700 applications (simul., data, AI)
- **We design for our workload**
  - Many jobs at many scales (40% of hours go to *capability jobs*)
  - Small, incoherent I/O
  - Not just checkpoint/restart!
- **Flash is ideal for versatile performance**



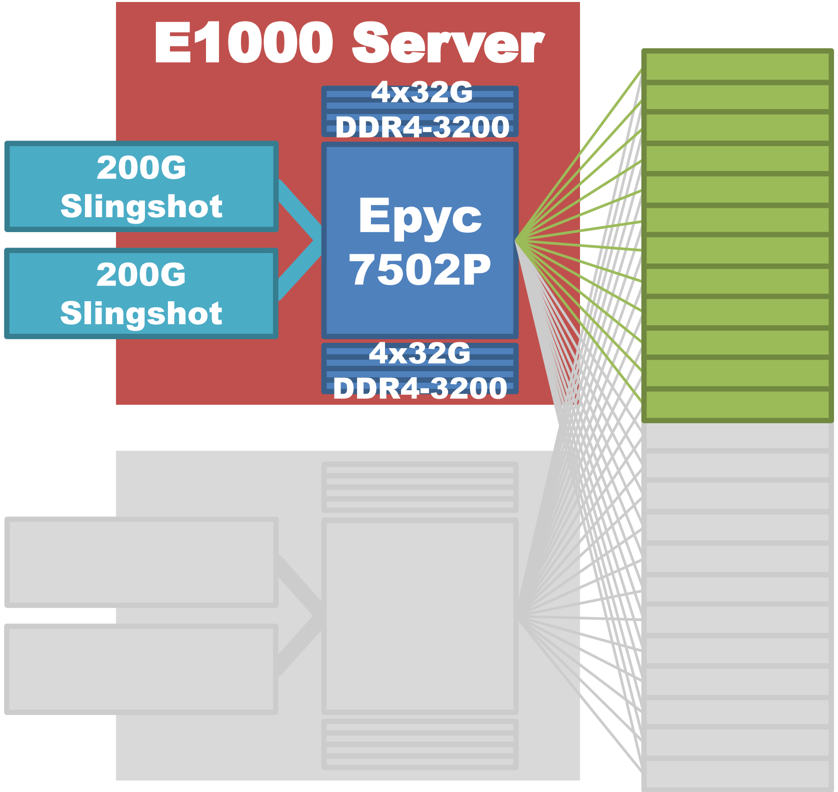
CRAY
NERSC
Perlmutter
Community File System
128 PB Spectrum Scale
16x MDS + 274 OSS
1x AMD Epyc 7502P
2x Slingshot NICs
24x 15.36 TB NVMe
1,536 GPU nodes
1x AMD Epyc 7763
4x NVIDIA A100
4x Slingshot NICs
Slingshot
200 Gb/s
2-level dragonfly
SAN
SAN
24x Gateway nodes
2x Slingshot NICs
2x 200G HCAs
3,072 CPU nodes
2x AMD Epyc 7763
1x Slingshot NIC
(coming soon)
2x Arista 7804 routers
400 Gb/s/port
> 10 Tb/s routing
WAN
External Facilities
HPC Centers
Telescopes/Beamlines
Cloud

Perlmutter's I/O Subsystem

- **35 PB usable**, all-NVMe Lustre
  - 274x OSSes
  - 16x MDSes
  - 3,480x SSDs total
- **Directly integrated** on dragonfly
  - No LNet routers or I/O forwarders
  - Four dragonfly groups for storage
  - File system remains available even if compute cabinets are down



Servers architected to maximize performance



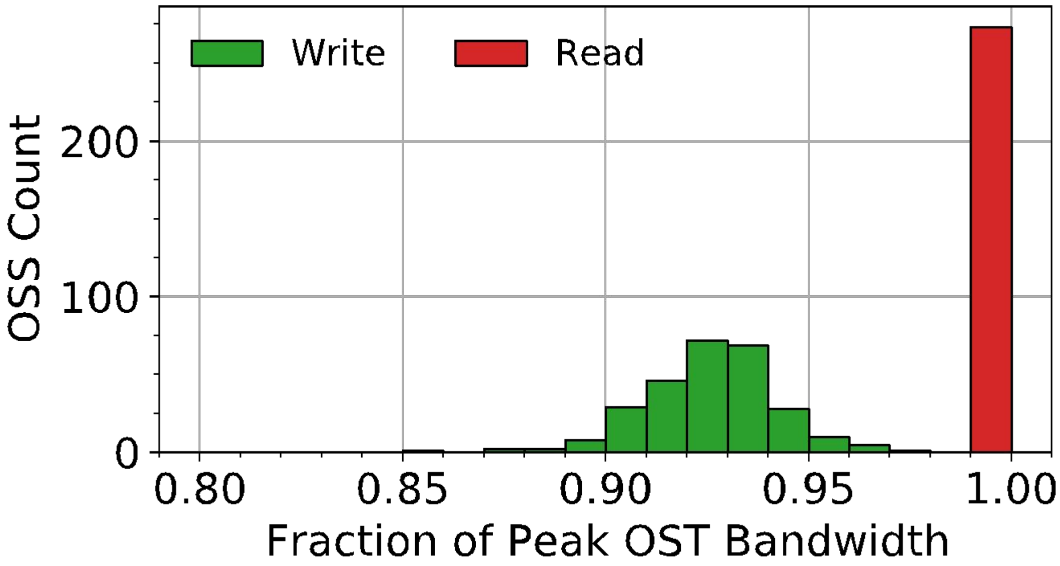
- Single-socket AMD Rome (128x PCIe Gen4 lanes)
  - Allows switchless design
  - 48 lanes for 24x NVMes
  - 32 lanes for 2x NICs
- **One** server = **one** OST/MDT
- **One** OST/MDT = **12x** NVMe
- GridRAID (HPE) + ldiskfs to maximize performance
  - OST = 8+2+1 RAID6 (GridRAID)
  - MDT = 11-way RAID10 (mdraid)

Performance efficiency – GridRAID and ldiskfs

SSD spec sheet
- 3.2 GB/s write
- 3.5 GB/s read

obdfilter-survey
- Writes: 92.6% of peak
  ~3.0 GB/s/SSD
- Reads: 99.9% of peak
  ~3.5 GB/s/SSD

obdfilter-survey results



GridRAID + ldiskfs efficiently **delivers NVMe bandwidth**

Performance efficiency – Slingshot and LNet

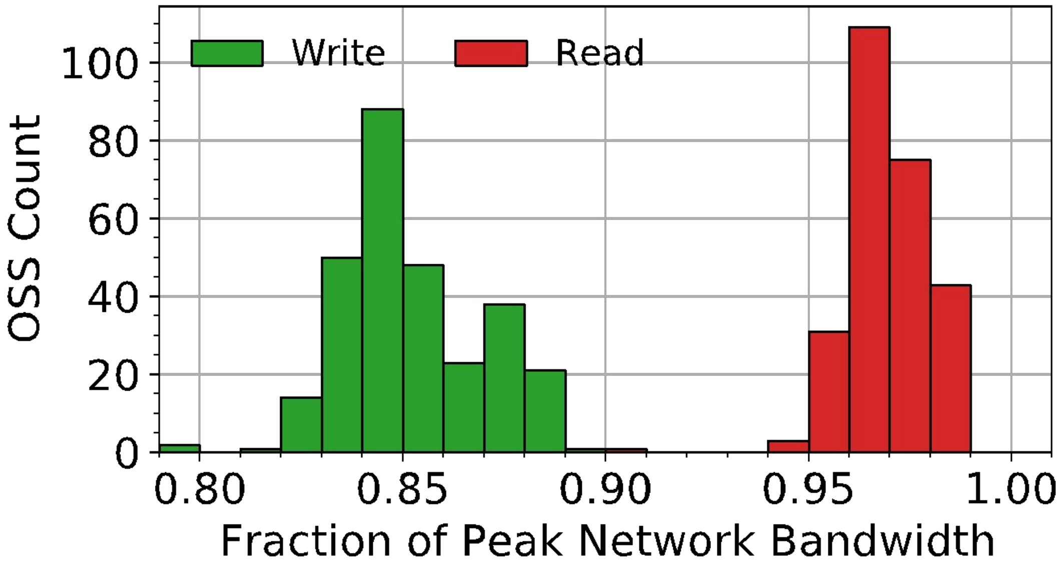
Line rate for NICs

- 2x 200 Gb Slingshot
- 50 GB/s line rate

LNet self-test

- Writes: 84.8% of peak
  ~42 GB/s/OSS
- Reads: 97.0% of peak
  ~48 GB/s/OSS

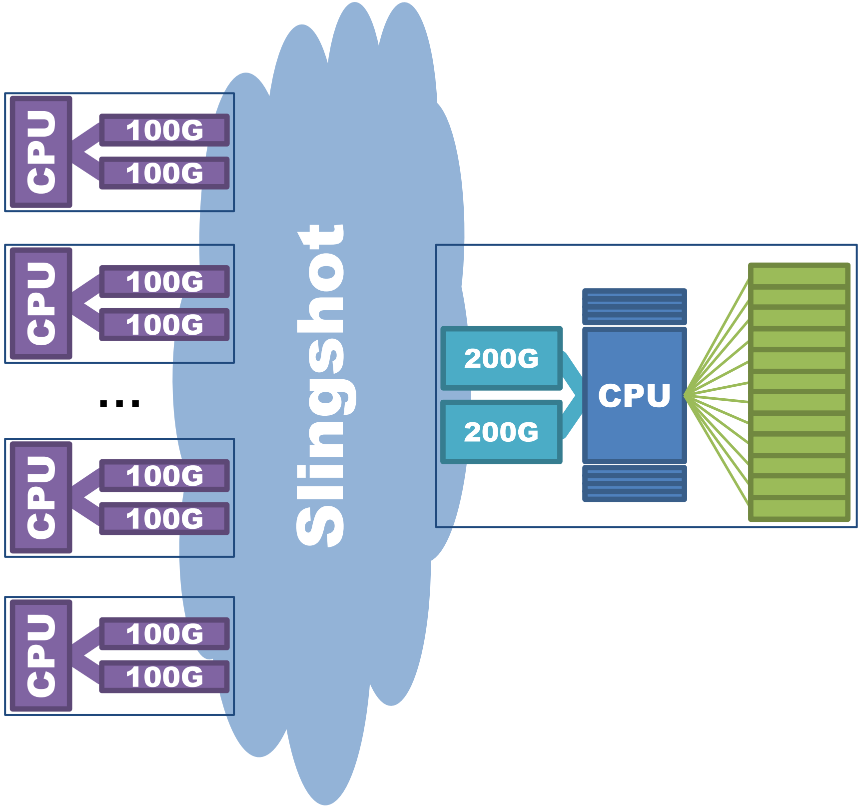
LNet self-test results



Slingshot and LNet multi-rail also efficiently **delivers bandwidth**

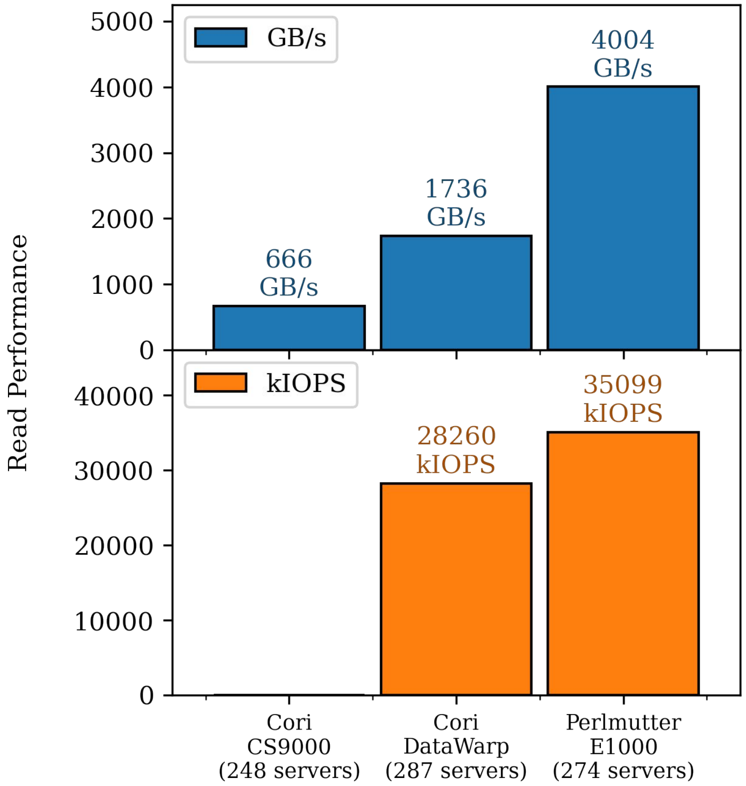
Performance capability of *one* NVMe OSS

- **Bandwidth**
  - Writes: 27 GB/s/OSS
  - Reads: 41 GB/s/OSS
- **IOPS**
  - Writes: 29 kIOPS/OSS
  - Reads: 1,400 kIOPS/OSS
- **Configuration**
  - IOR w/ 20 to 250 clients
  - 1 OSS, 1 OST (12 NVMes in 8+2)
  - Slingshot interconnect
  - Lustre version 2.12.4+cray
  - Perlmutter has **274 OSSes**

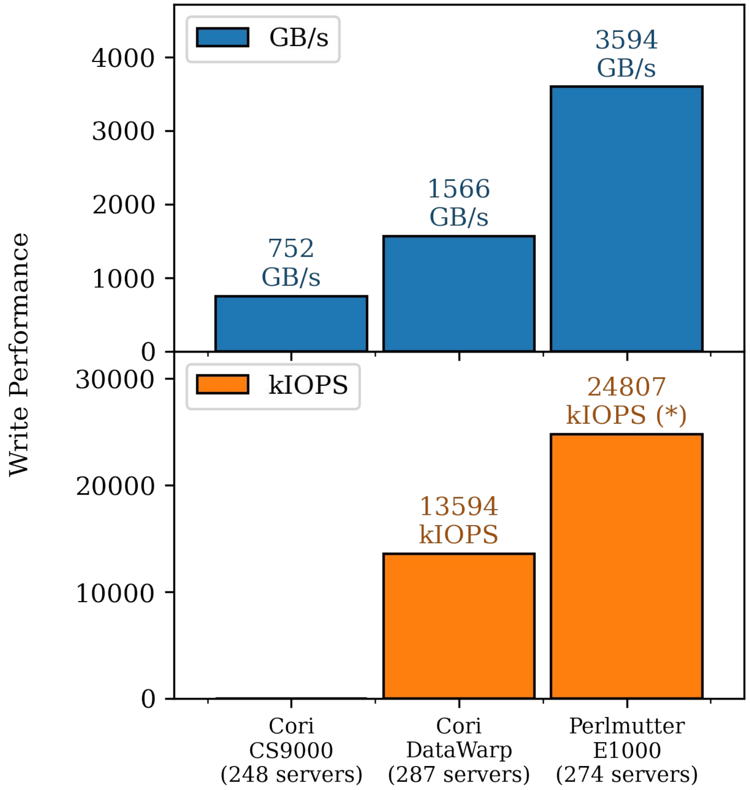


How does this compare to Cori for *reads*?

- **Read bandwidth** up to 2.3x - 6x higher than Cori
- **Read IOPS** are promising for user experience
  - NERSC is read-heavy
  - Expecting much better interactive responsiveness
  - Expecting **less variation** from contention (more predictable performance)



How does this compare to Cori for *writes*?



- **Write bandwidth** is 2x - 4.7x higher than Cori
- **Write IOPS** seems better than Datawarp
  - *most of this actually merit of OS cache
  - Far better than Cscratch anyway
  - RAID6 (8+2+1) vs RAID0 (DW)
  - Perlmutter traded IOPS for **resilience**

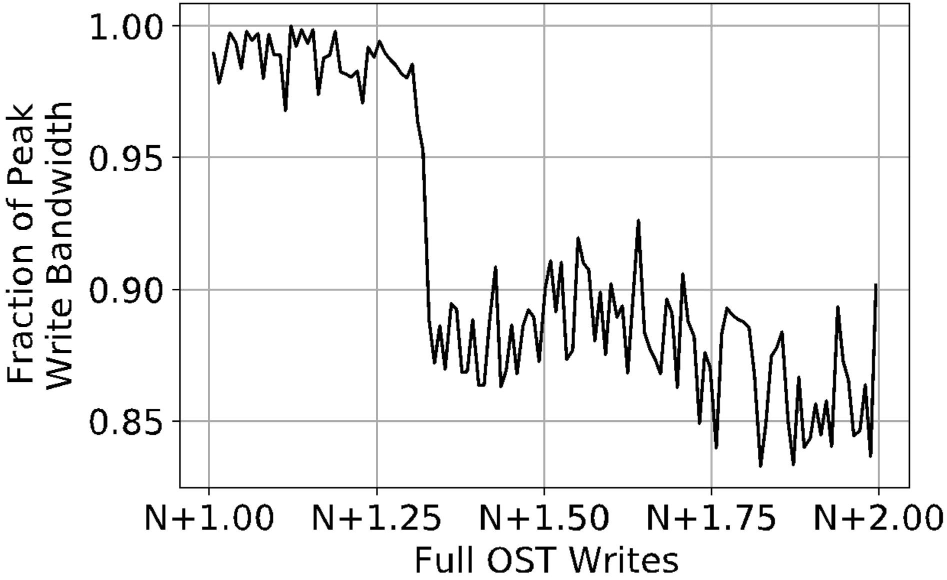
What about *metadata*?

- mdtest 3.3
- “production” run
  - 230 clients x 6 procs/client = 1380 procs
  - **1.6 M** file/s created
- “full-scale” run
  - 1382 clients x 2 procs/client = 2764 procs
  - **1.3 M** file/s deleted
- Great improvement for User Experience on the system
- Not comparable to Cori scratch

A few surprises so far…

SSD OSTs slow with age

- analogous to HDD fragmentation
- ~10% write bandwidth lost after ~5 full OST writes
- *fstrim* completely restores write performance!
- We anticipated monthly trim
  - 5x OSTs = 665 TB
  - expect: 2.2 – 2.9 PB/day
  - 5x OST writes = 60 - 80 days
- Currently performing it **nightly**



Note: N was not carefully measured

Bugs found (so far)

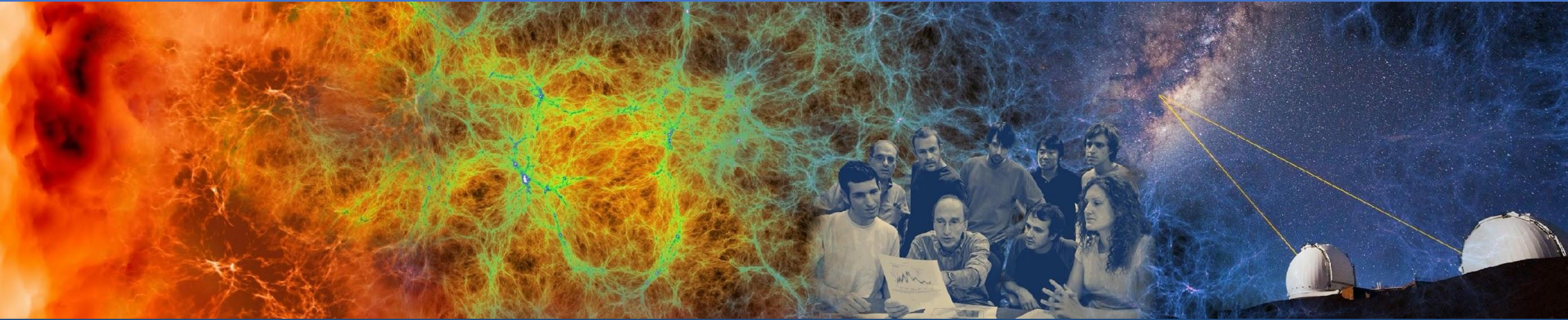
- Progressive File Layout (PFL) striping
  - IOR w/ Cray-MPICH was crashing/freezing when doing I/O against a PFL-striped directory
  - Quickly patched
  - Is NERSC the first to use PFL in prod?

- (unrelated to Perlmutter's Lustre scratch) MPICH's *MPI_File_write_all* with *romio_no_indep_rw=true* hint freezes against a GPFS file system
  - Lots of clients using our GPFS file systems
  - Patched in upstream MPICH
  - Will land soon on Perlmutter

Thank you!

Special thanks to Peter Bojanic, John Fragalla, Jeff Hudson, Cory Spitz, and the HPE Cray Storage R&D team for their insights. Work should be credited to Ershaad Basheer, Alberto Chiusole, David Fox, Lisa Gerhardt, Kirill Lozinskiy, David Paul, Chris Samuel, Nicholas Wright, and other members of the NERSC-Cray Lustre Center of Excellence.

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

NERSC's first foray into NVMe at scale: Cori (2015)

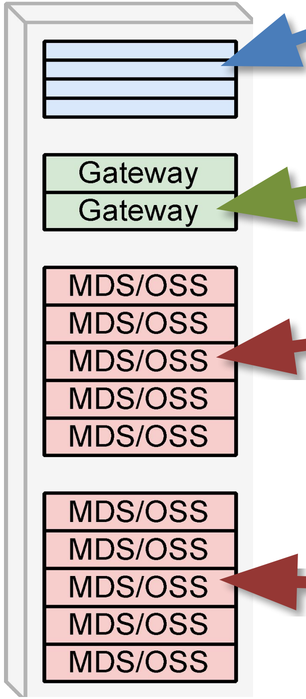
Cori – Cray XC-40

- 2,388 Intel Haswell nodes
- 9,688 Intel KNL nodes

I/O Subsystem

- 1.8 PB, 1.5 TB/s burst buffer
  - DataWarp File System
  - 1,152 NVMe SSDs
  - RAID0
- 30 PB, 700 GB/s scratch
  - Lustre File System
  - 10,168 HDDs
  - 8+2 RAID6

Rack-scale layout



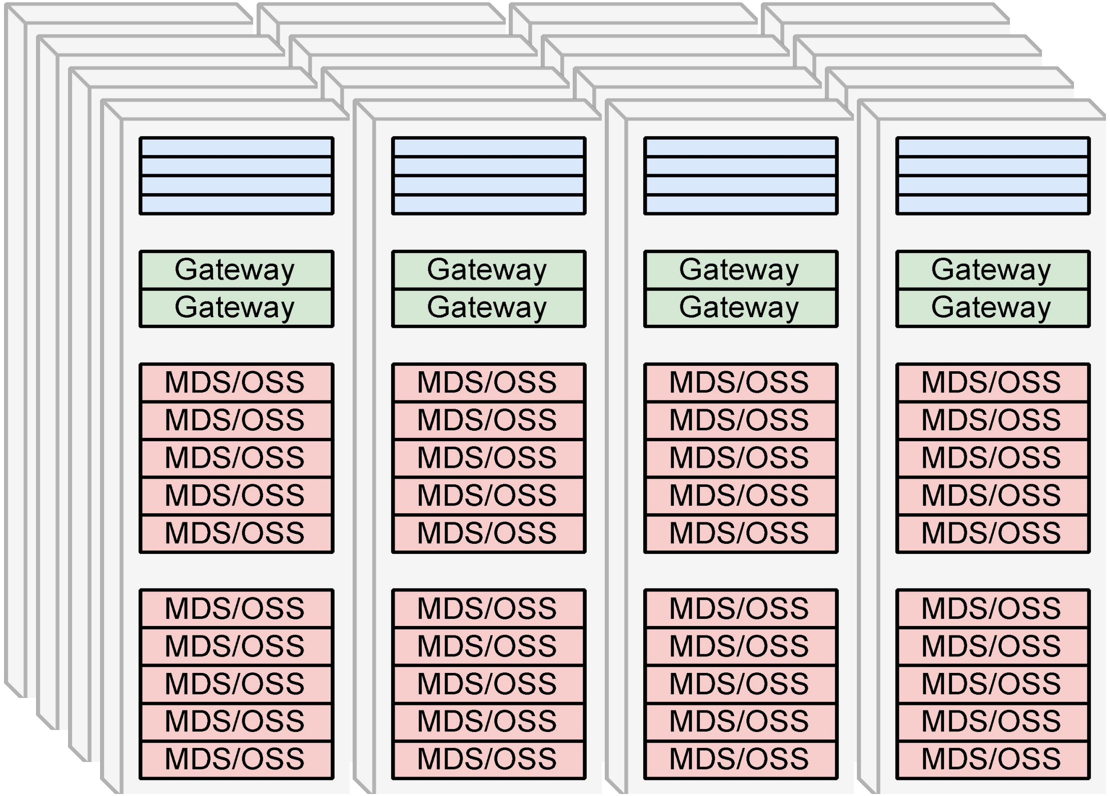
Physically (approx.)

- 10x 2U24 enclosures
- 240x 15.36T NVMe SSD
- 2x gateway nodes – Slingshot to InfiniBand
- Slingshot switch complex

Logically (approx.)

- 20 OSS and/or MDS
- 3.6 PB raw
- 1.6 PB (if all MDSes) - 2.7 PB (if all OSSes) usable

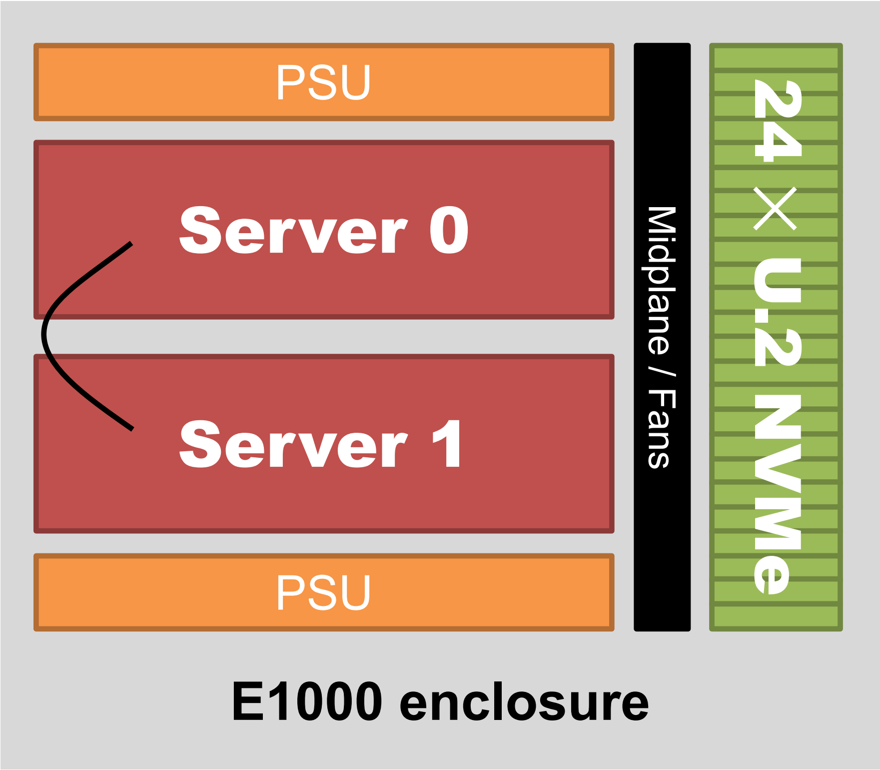
Rack-scale layout



4 racks = 1 group

- Four groups total
- Each connected to every other group in the system
- Compute I/O can get dedicated global links
- 14.4 TB/s/dir to computes

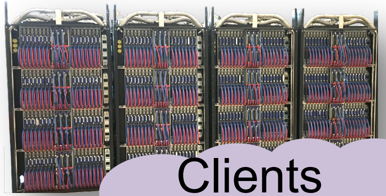
Inside a 2U24 enclosure

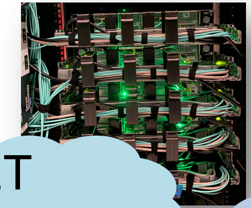


Designed for <u>reliability</u>

- No single points of failure
  - 2x servers (OSS or MDS)
  - Redundant PSUs, fans, etc
- 24x U.2 15.36 TB NVMe drives
  - Samsung PM1733
  - Dual-ported PCIe Gen4 (2x2)
  - Each server sees 24x drives
- Heartbeating and failover

Perlmutter file system: excellent performance efficiency





Clients (Slingshot)
IOR

41 GB/s read
27 GB/s write
1400 kIOPS read
29 kIOPS write

LNET (Slingshot)
LNET Self-test

48 GB/s read
42 GB/s write

RAID (GridRAID)
OBDFilter Survey

43 GB/s read
31 GB/s write

NVMe (PM1733)
fio

42 GB/s read
38 GB/s write
9,600 kIOPS read
1,600 kIOPS write

88.4%(w) / 97.2%(r) NVMe block bandwidth (remember: 8+2 on writes)

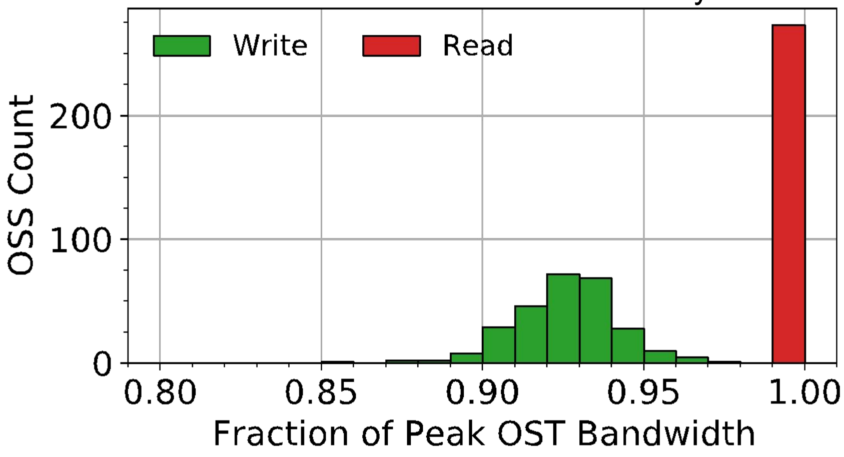
5.33%(w) / 15.1%(r) NVMe block IOPS (after read-modify-write penalty)

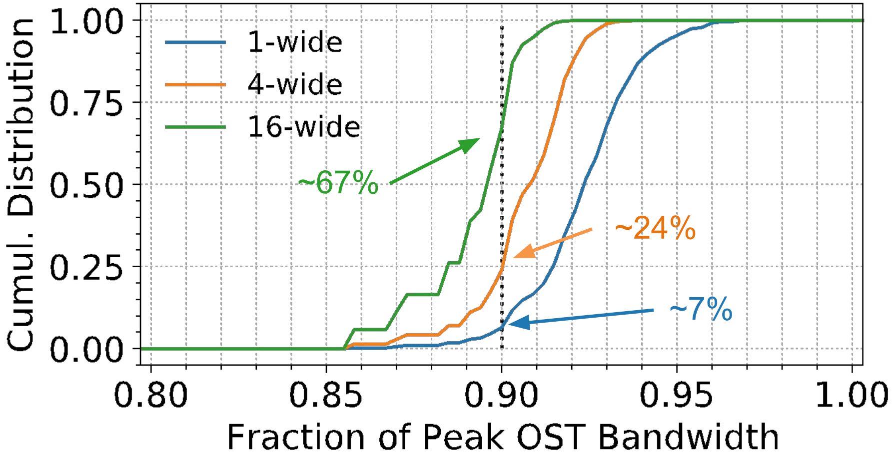
A few surprises so far…

Read and write bandwidths differ

- Reads faster than writes
  - write parity overhead
  - NVMe is faster on reads
- Writes *vary* more
- Must balance
  - stripe width (high bandwidth)
  - write variability (straggling OSTs)

obdfilter-survey results





Take-aways and next steps

- Perlmutter's 35 PB all-NVMe file system is built on HPE Cray E1000
- Lustre, GridRAID, ldiskfs, Slingshot, and LNet multi-rail efficiently deliver bandwidth *and* IOPS from NVMe to clients
- More scale results to follow!
  - Scaled up to O(1,000) compute nodes and 274 OSSes already!
  - Metadata/DNE testing kicking off
  - Progressive File Layout (PFL): automatic file striping enabled