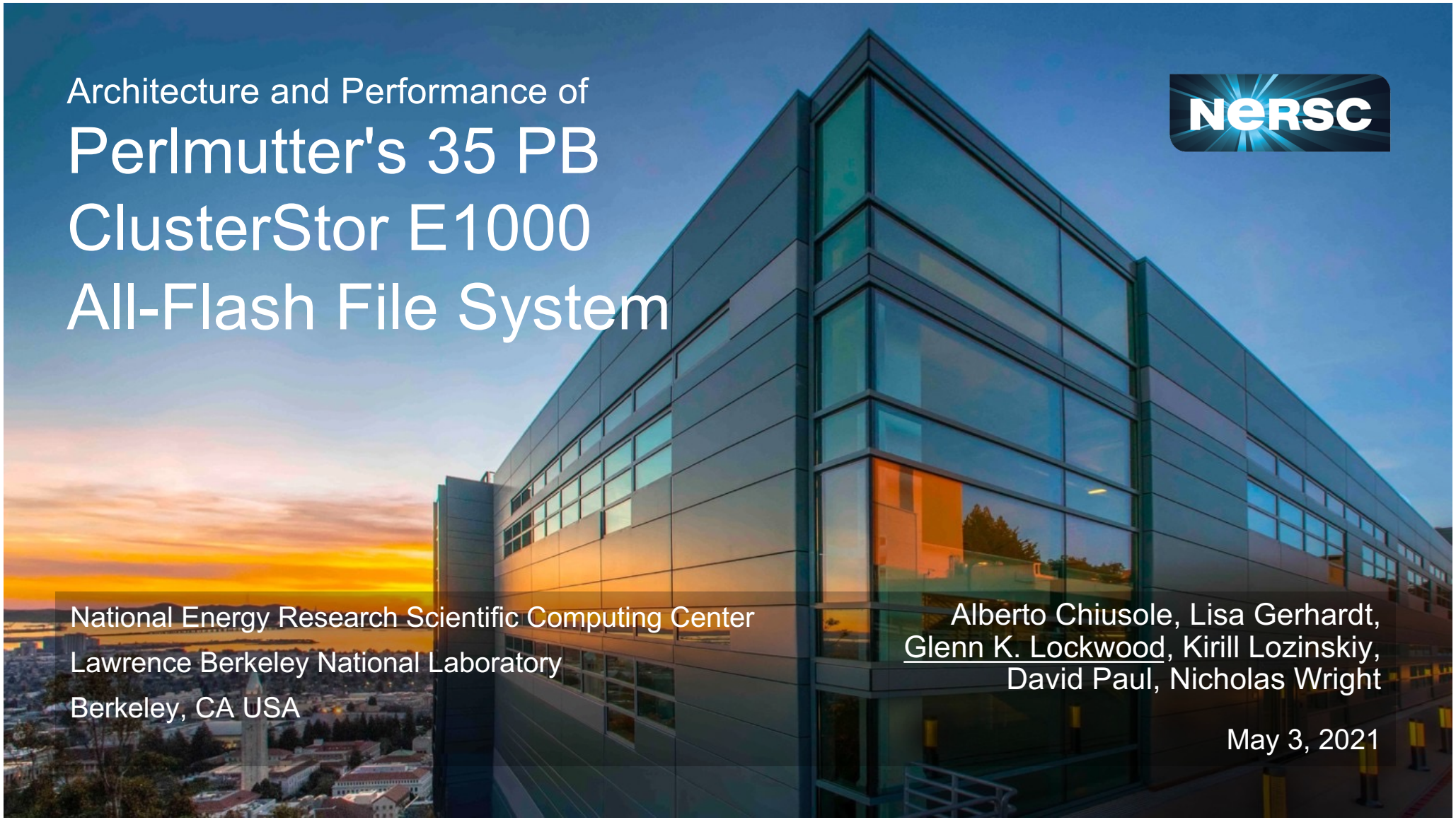


Architecture and Performance of

Perlmutter's 35 PB ClusterStor E1000 All-Flash File System

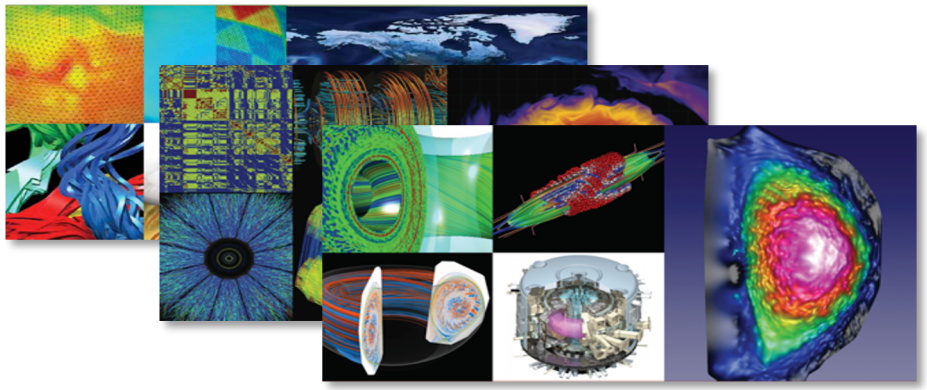
National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory
Berkeley, CA USA

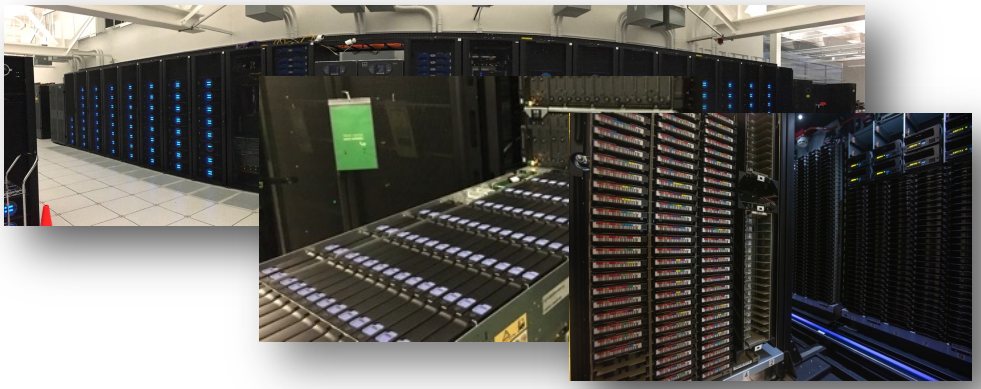
Alberto Chiusole, Lisa Gerhardt,
Glenn K. Lockwood, Kirill Lozinskiy,
David Paul, Nicholas Wright

May 3, 2021

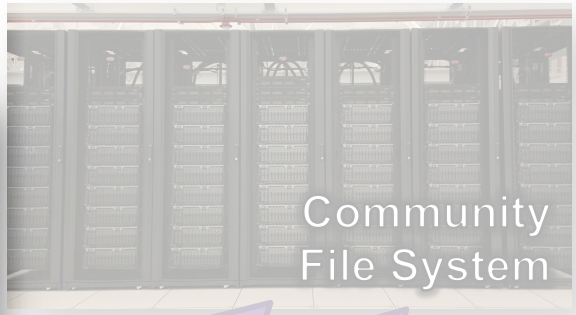
NERSC is the mission computing facility

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





- **NERSC serves a diverse community**
  - 8,000 active users
  - 900 projects
  - 700 applications (sim, data, AI)
  - 1,800 publications
- **NERSC systems designed accordingly**
  - Versatile performance – not just peak
  - Multitenant computing
  - Many jobs running at many scales (40% of hours go to capability jobs)
- **Perlmutter designed with this philosophy**



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

24x Gateway nodes
2x Slingshot NICs
2x 200G HCAs

SAN

SAN

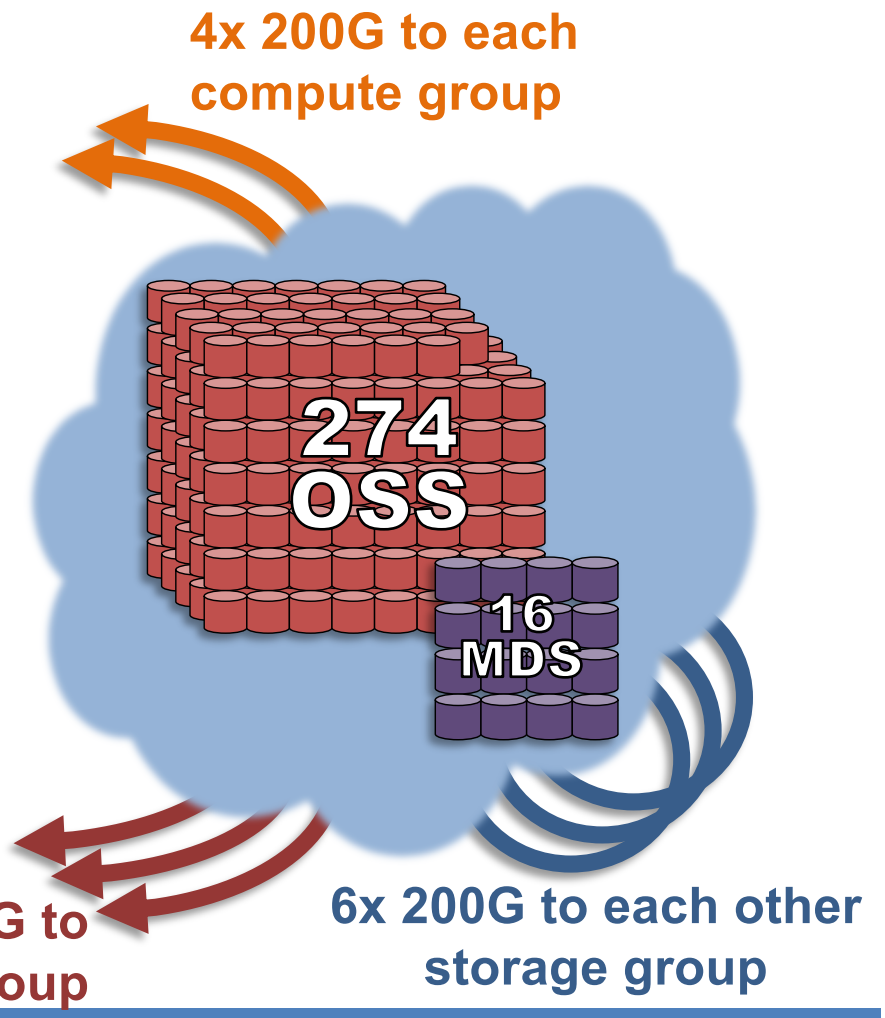
2x Arista 7804 routers
400 Gb/s/port
> 10 Tb/s routing

WAN

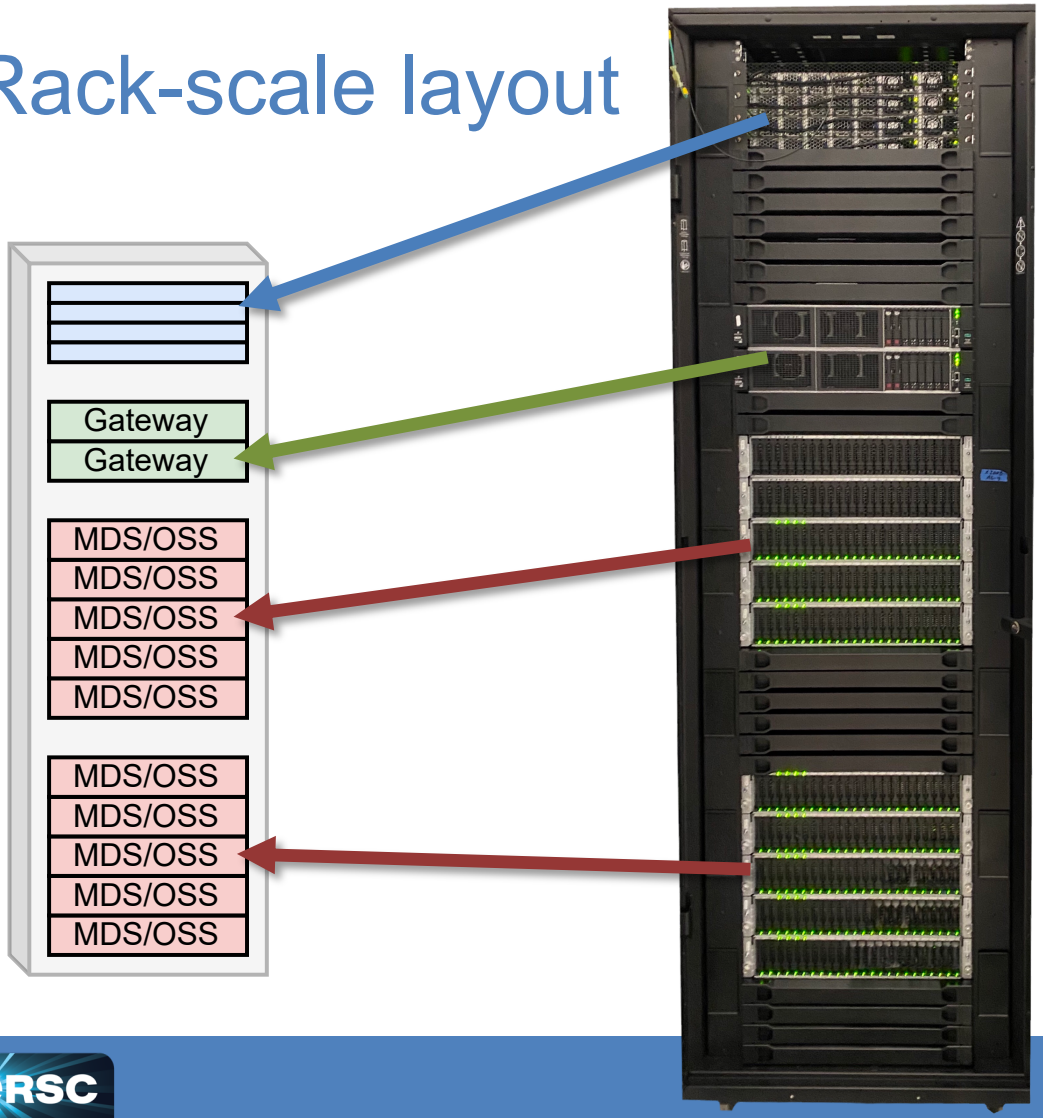
3,072 CPU nodes
2x AMD Epyc 7763
1x Slingshot NIC

Perlmutter's I/O Subsystem

- 35 PB usable, all-NVMe Lustre
- 274x OSSes
- 16x MDSes
- Directly integrated on dragonfly
  - No LNet routers
  - Four dragonfly groups for file system
  - File system remains available even if compute cabinets are down



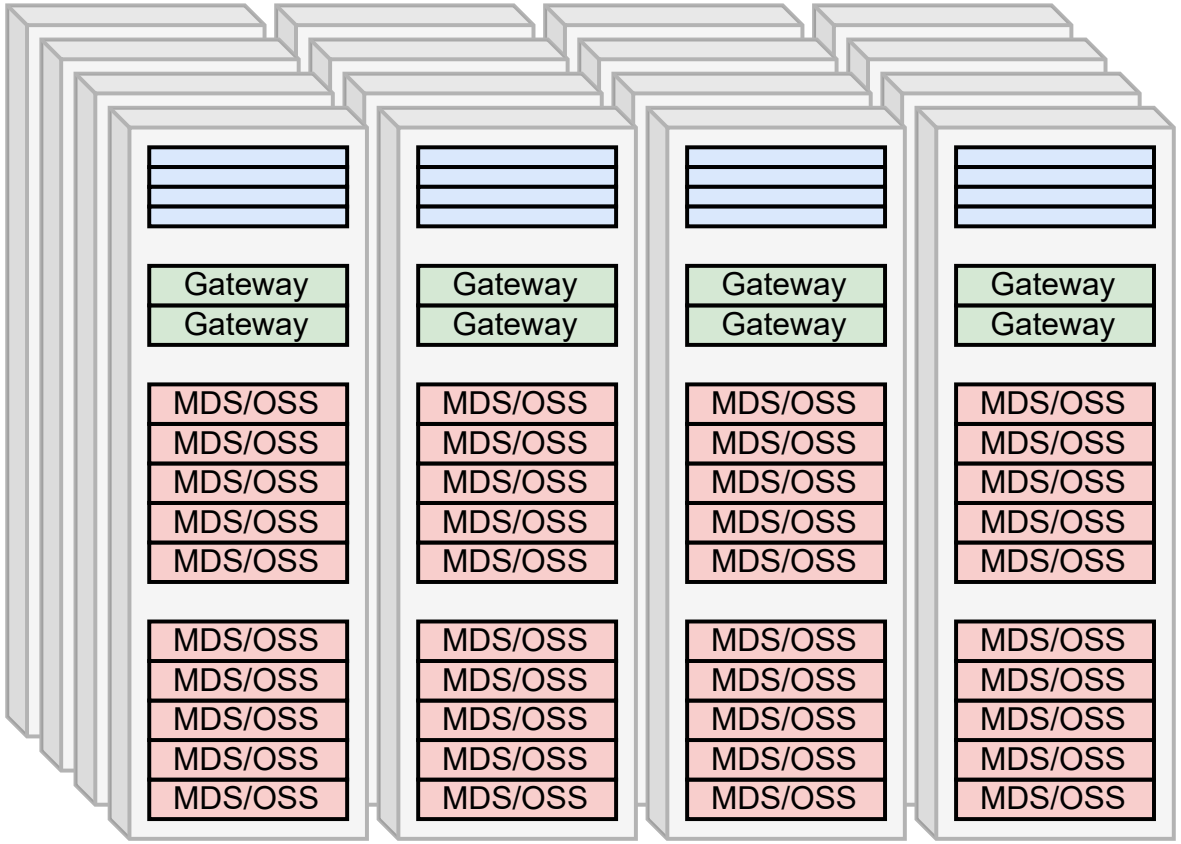
Rack-scale layout



Each rack =

- ~10 ClusterStor E1000 enclosures
- ~2 gateway nodes – Slingshot to InfiniBand
- Slingshot switch complex

Rack-scale layout



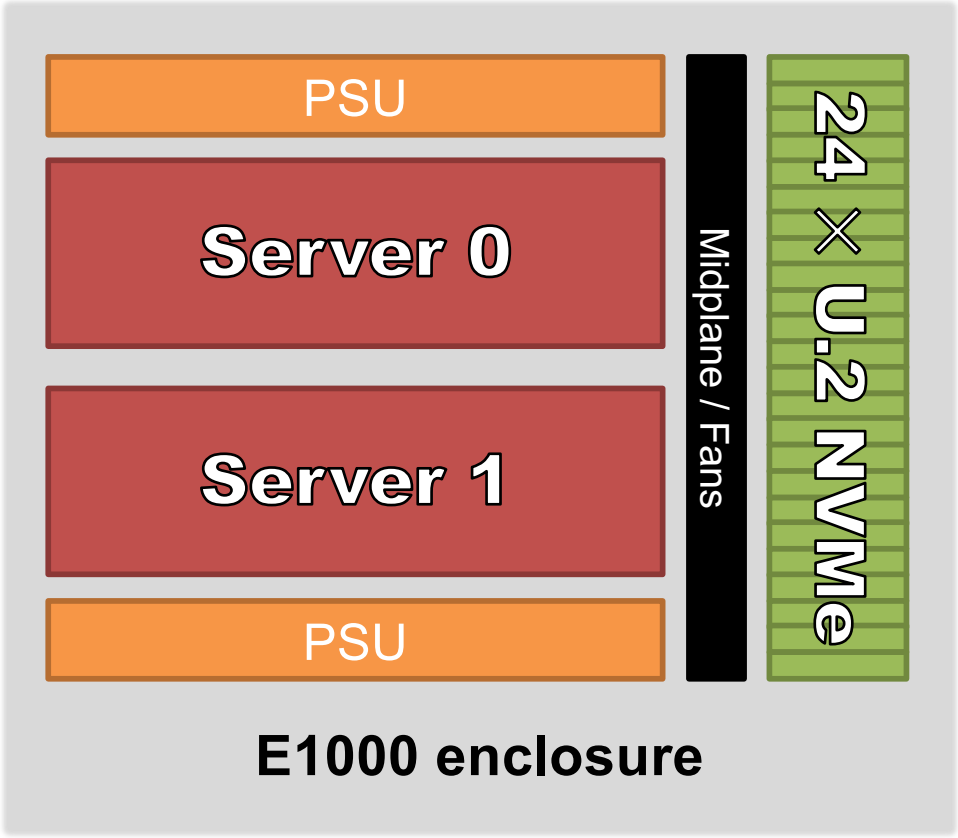
Each rack =

- ~10 ClusterStor E1000 enclosures
- ~2 gateway nodes – Slingshot to InfiniBand
- Slingshot switch complex

4 racks = 1 group

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

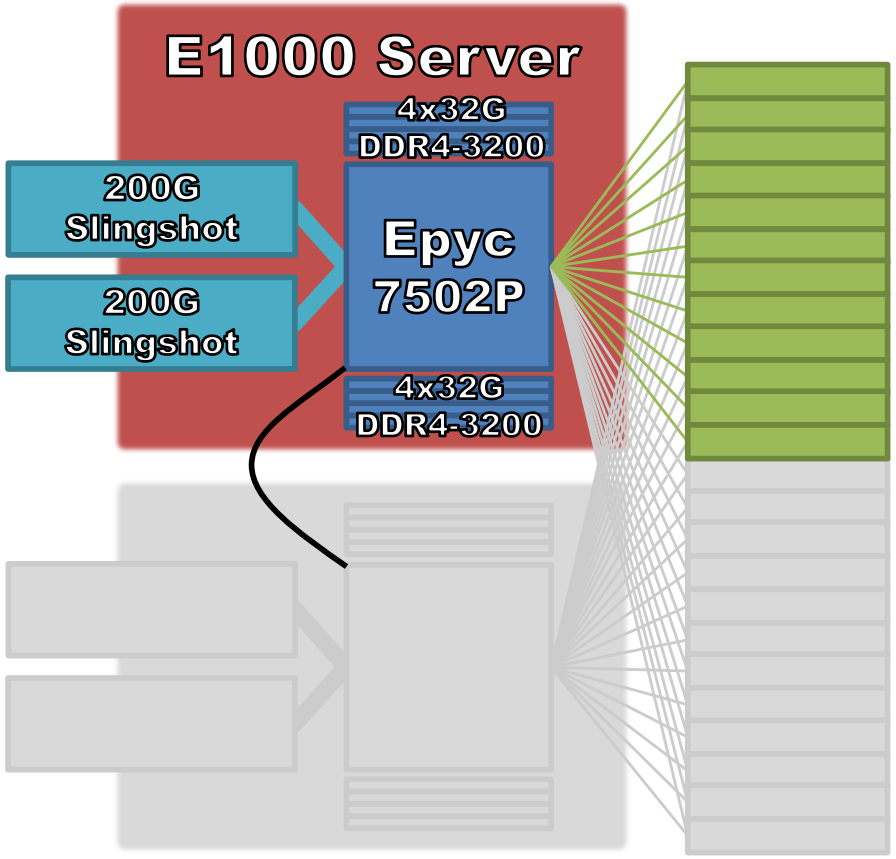
Inside a single 2U E1000 enclosure



Designed to be reliable – no single points of failure

- 2x servers (OSS or MDS)
- Redundant PSUs, fans, fan controllers
- 24x U.2 15.36 TB NVMe drives
  - Samsung PM1733
  - Dual-ported PCIe Gen4 (2x2)
  - Each server sees 24x drives
- Infrastructure for heartbeating and failover

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 + ldiskfs to maximize performance
  - OST = 8 + 2 + 1 RAID6
  - MDT = 11-way RAID10

Small-scale end-to-end performance

Configuration

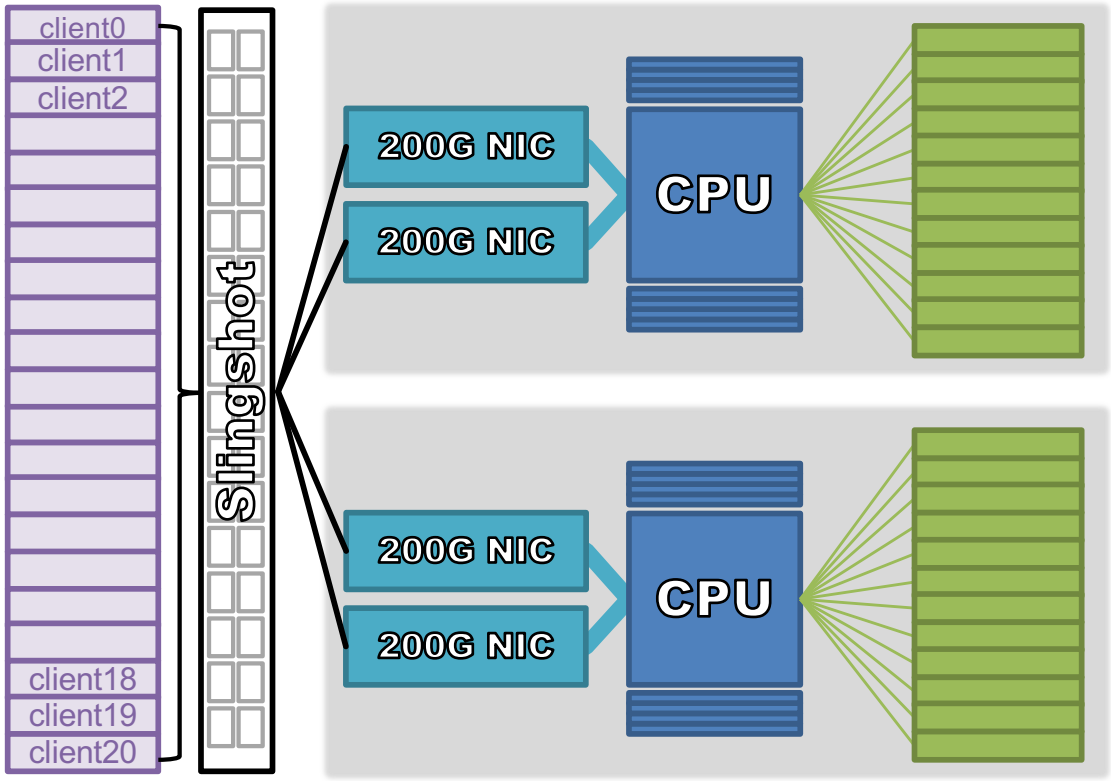
- 21x clients
- 2x OSSes (1x E1000)
- Slingshot interconnect

Bandwidth

- Writes: 28 GB/s/OSS
- Reads: 41 GB/s/OSS

IOPS

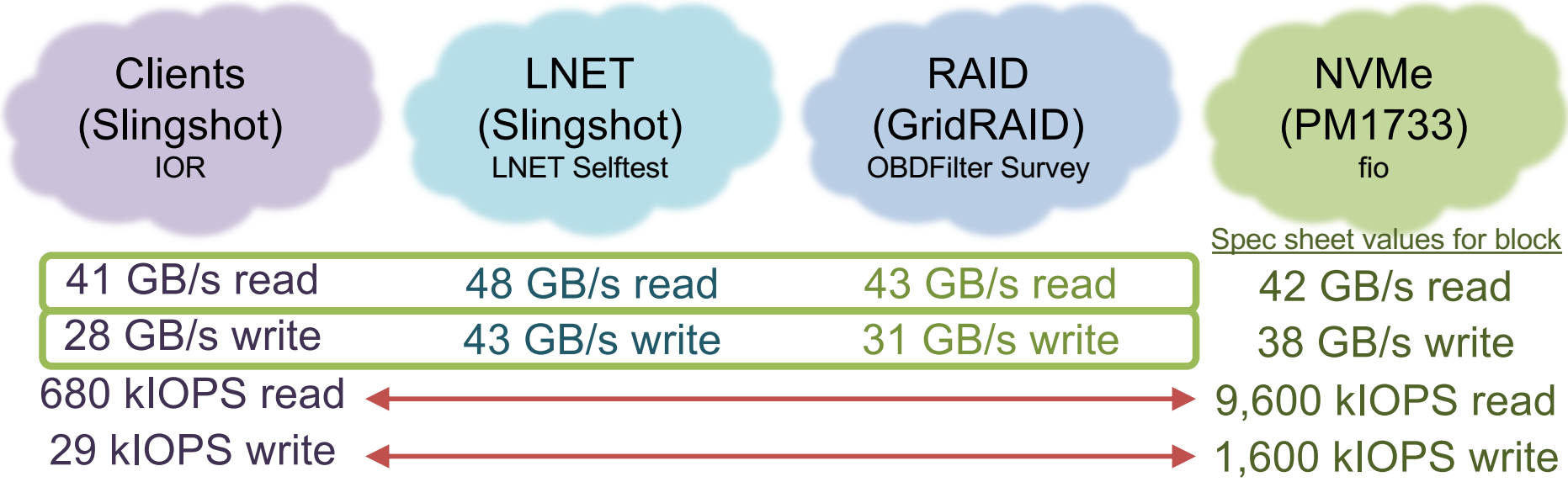
- Writes: 29 kIOPS/OSS
- Reads: 680 kIOPS/OSS



Note:

- Data provided by HPE from smaller-scale network than Perlmutter
- These data reflect newer Linux kernel than obdfilter/lnet-selftest tests

E1000 efficiently delivers NVMe performance to clients

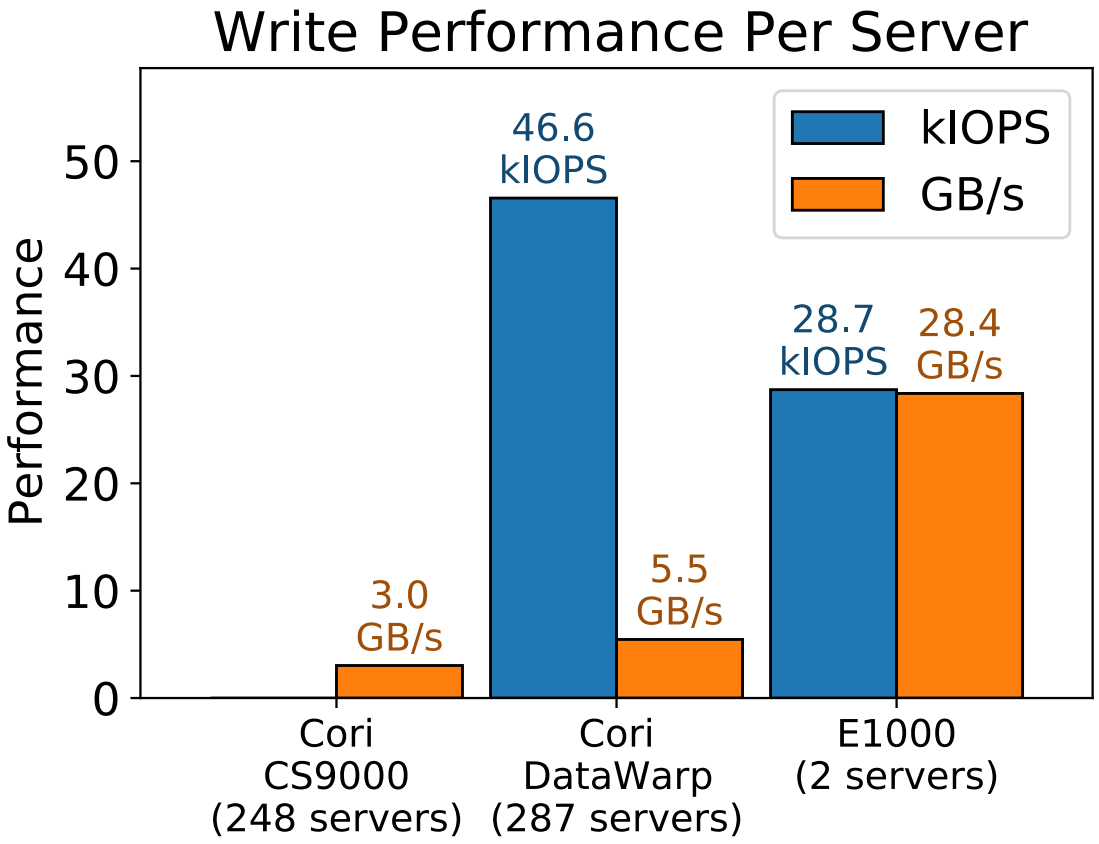


Lustre, Slingshot, and GridRAID deliver excellent performance efficiency

- 90% NVMe block bandwidth (remember: 8+2 on writes)
- 10% NVMe block IOPS (after read-modify-write penalty)

Note:
- Data provided by HPE from smaller-scale network than Perlmutter
- These data reflect newer Linux kernel than obdfilter/lnet-selftest tests

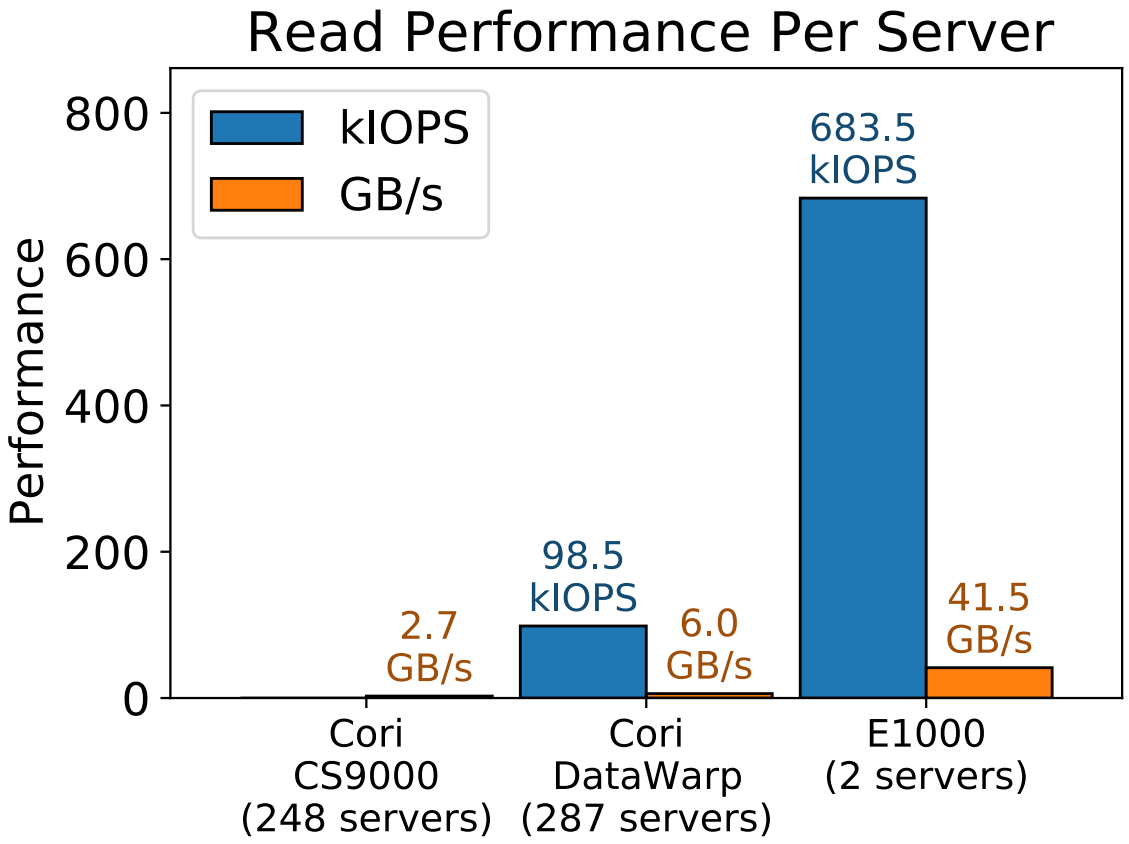
How does this compare to Cori for *writes*?



- *NOTE: E1000 data does not reflect network scaling*
- **Write bandwidth** up to 9x previous generation
- **Write IOPS** falls short of DataWarp
  - RAID6 (E1000) vs RAID0 (DataWarp)
  - Perlmutter traded IOPS for resilience (DataWarp failed if one drive failed)

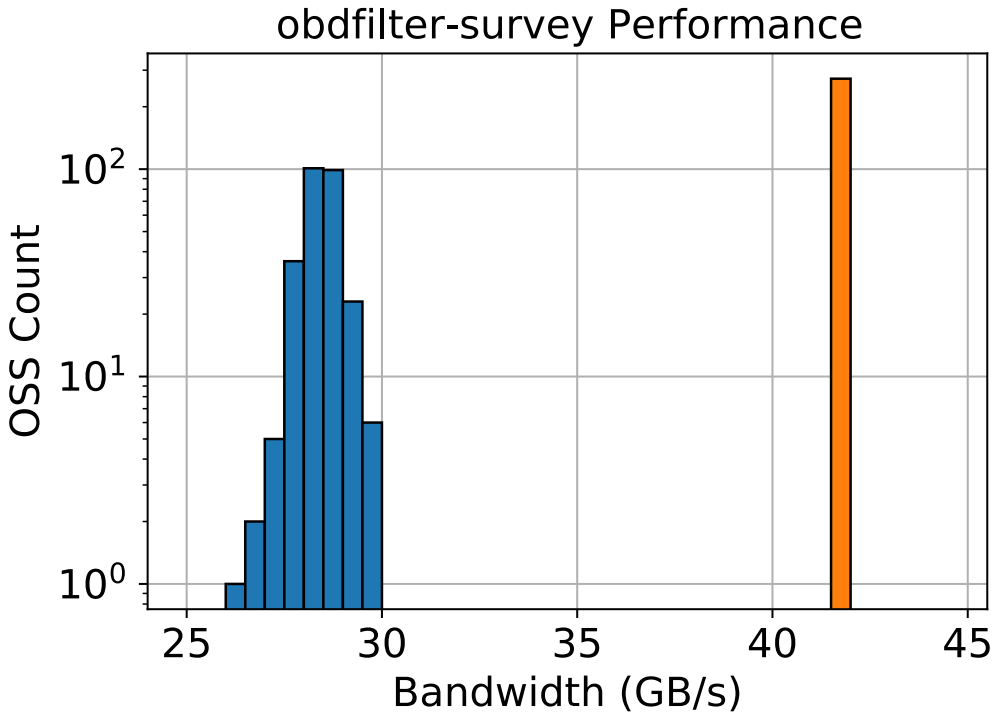
How does this compare to Cori for *reads*?

- *NOTE: E1000 data does not reflect network scaling*
- **Read bandwidth** up to 15x previous generation
- **Read IOPS** may be most exciting part
  - NERSC is read-heavy
  - Expecting much better interactive responsiveness
  - Expecting less variation from contention



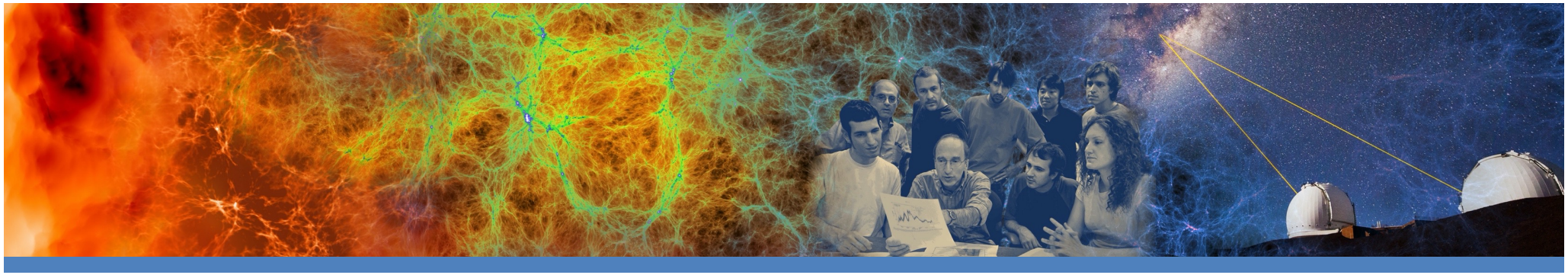
A few observations from going all-NVMe so far

- **Performance of writes and reads differ**
  - Reads > Writes
  - Writes *vary* more
- **HPE saw SSD OSTs slow with age (> 5 drive writes)**
  - …but fstrim completely restores write performance!
  - NERSC anticipates monthly trim (expect ~130 PB/month written)



Take-aways and next steps

- Perlmutter's 35 PB all-NVMe file system is built on HPE Cray E1000
- Lustre, GridRAID, and ldiskfs efficiently deliver bandwidth *and* IOPS from NVMe to clients
- At-scale testing and tuning to follow
  - Scaling up to O(1,000) compute nodes and 274 OSSes this month
  - Metadata/DNE testing kicking off



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.