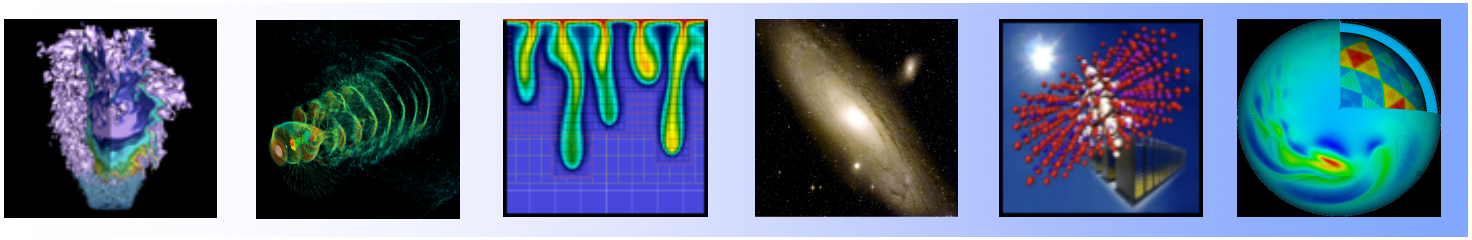


DVS, GPFS and External Lustre at NERSC – How It's Working on Hopper

Tina Butler, Rei Chi Lee, Gregory Butler

05/25/11

CUG 2011

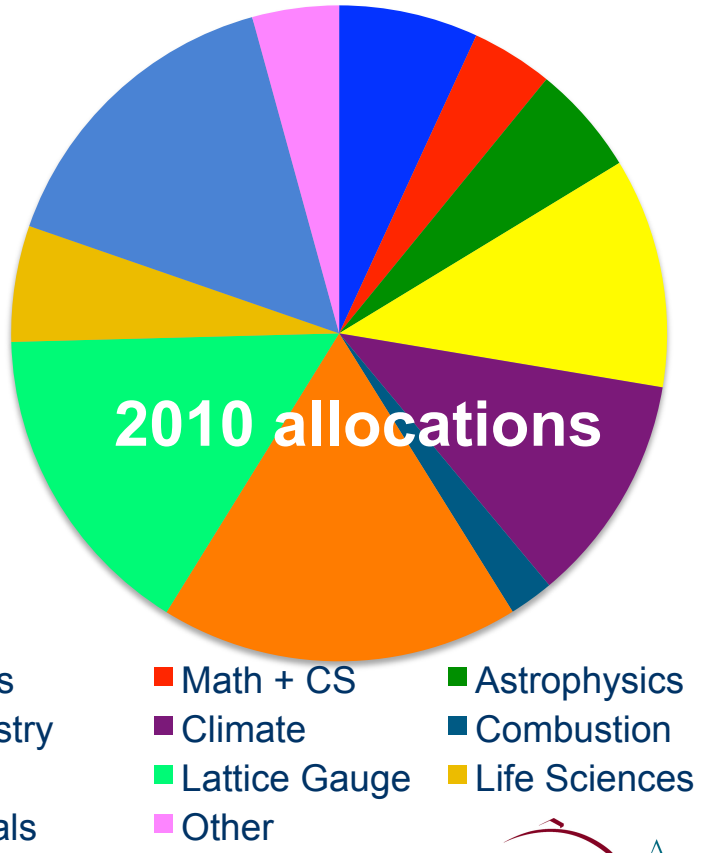
U.S. DEPARTMENT OF ENERGY | Office of Science

National Energy Research Scientific Computing Center

BERKELEY LAB Lawrence Berkeley National Laboratory

NERSC is the Primary Computing Center for DOE Office of Science

- **NERSC serves a large population**
  - Approximately 3000 users, 400 projects, 500 codes
- **Focus on "unique" resources**
  - Expert consulting and other services
  - High end computing & storage systems
- **NERSC is known for:**
  - Excellent services & diverse workload



NERSC Systems

Large-Scale Computing Systems

Franklin (NERSC-5): Cray XT4
- 9,532 compute nodes; 38,128 cores
- ~25 Tflop/s on applications; 356 Tflop/s peak

Hopper (NERSC-6): Cray XE6
- Phase 1: Cray XT5, 668 nodes, 5344 cores
- Phase 2: Cray XE6, 6384 nodes, 153216 cores 1.28 Pflop/s peak

Clusters

140 Tflops total

Carver
- IBM iDataplex cluster

PDSF (HEP/NP)
- ~1K core throughput cluster

Magellan Cloud testbed
- IBM iDataplex cluster

GenePool (JGI)
- ~5K core throughput cluster

NERSC Global Filesystem (NGF)

Uses IBM's GPFS
- 1.5 PB capacity
- 10 GB/s of bandwidth

HPSS Archival Storage

- 40 PB capacity
- 4 Tape libraries
- 150 TB disk cache

Analytics

Euclid
(512 GB shared memory)

Dirac GPU testbed (48 nodes)

Lots of users, multiple systems, lots of data

- At the end of the 90's it was becoming increasingly clear that data management was a huge issue.
- Users were generating larger and larger data sets and copying their data to multiple systems for pre- and post-processing.
- Wasted time and wasted space
- Needed to help users be more productive

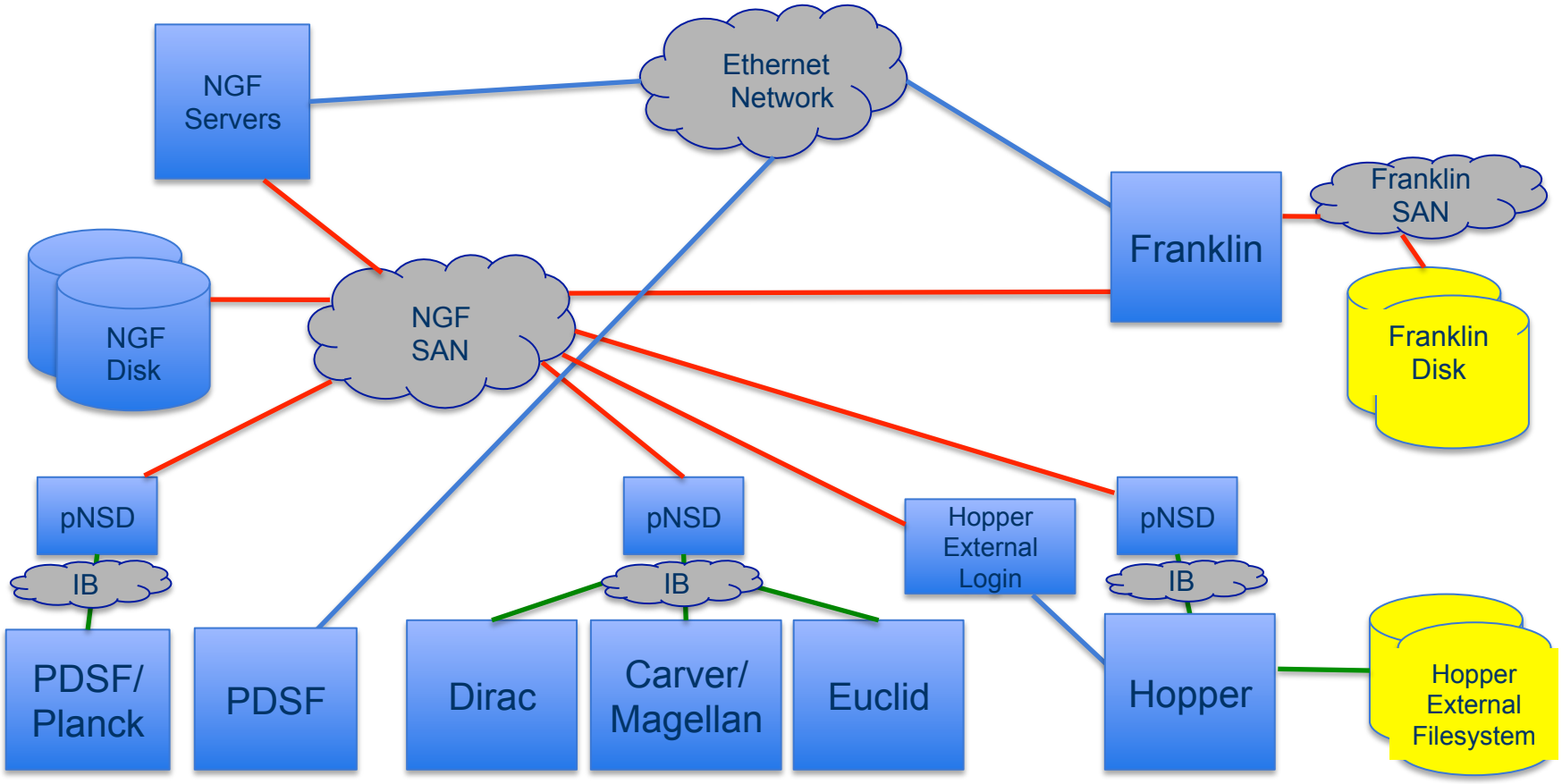
Global Unified Parallel Filesystem

- In 2001 NERSC began the GUPFS project.
  - High performance
  - High reliability
  - Highly scalable
  - Center-wide shared namespace
- Assess emerging storage, fabric and filesystem technology
- Deploy across all production systems

NERSC Global Filesystem (NGF)

- First production in 2005 using GPFS
  - Multi-cluster support
  - Shared namespace
  - Separate data and metadata partitions
  - Shared lock manager
  - Filesystems served over Fibre Channel and Ethernet
  - Partitioned server space through private NSDs

NERSC Global Filesystem (NGF)



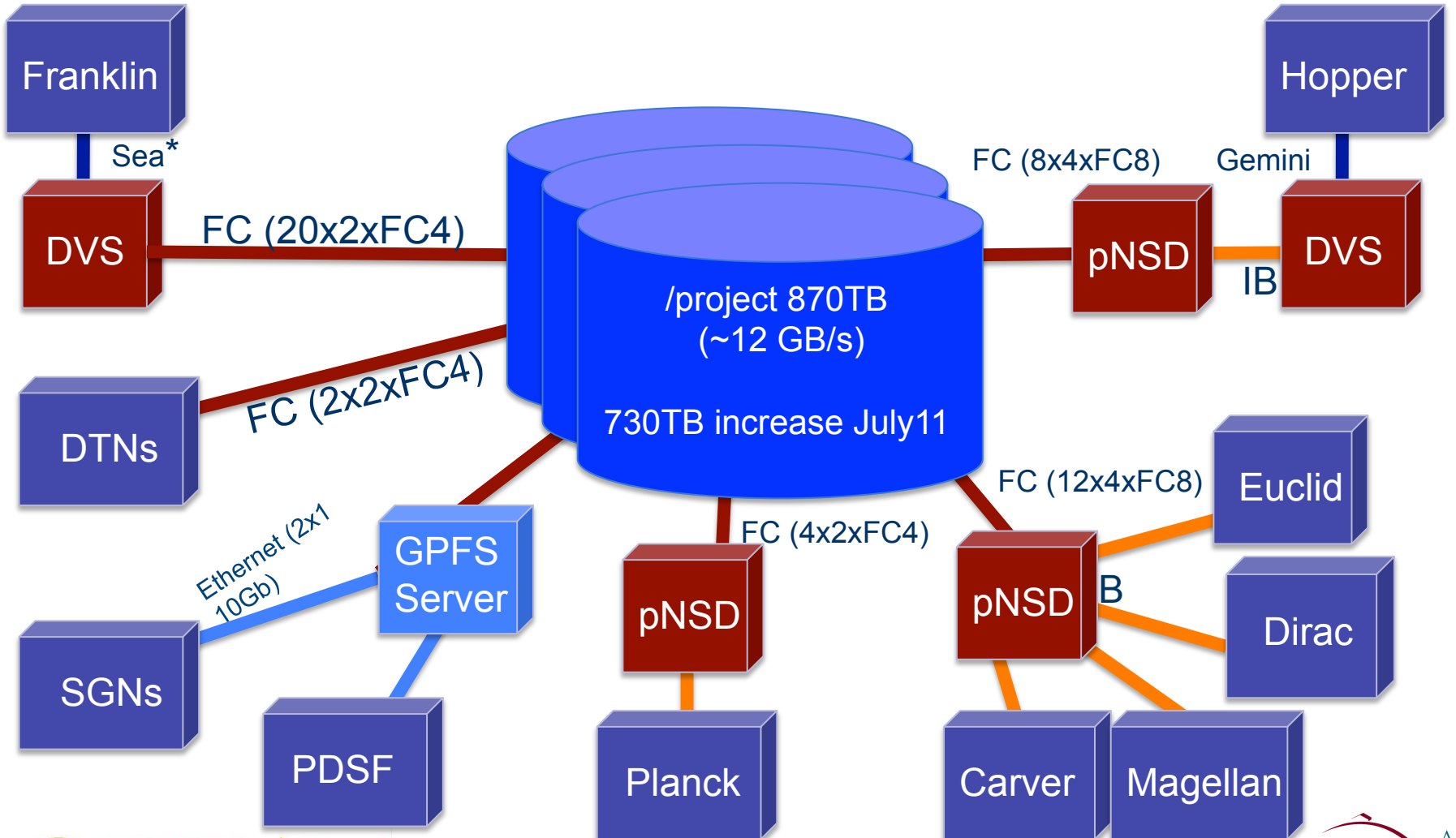
NGF Configuration

- NSD servers are commodity
  - 28 core servers
  - 26 private NSD servers
    - 8 for hopper; 14 for carver; 8 for planck (PDSF)
- Storage is heterogeneous
  - DDN 9900 for data LUNs
  - HDS 2300 for data and metadata LUNs
  - Have also used Engenio and Sun
- Fabric is heterogeneous
  - FC-8 and 10 GbE for data transport
  - Ethernet for control/metadata traffic

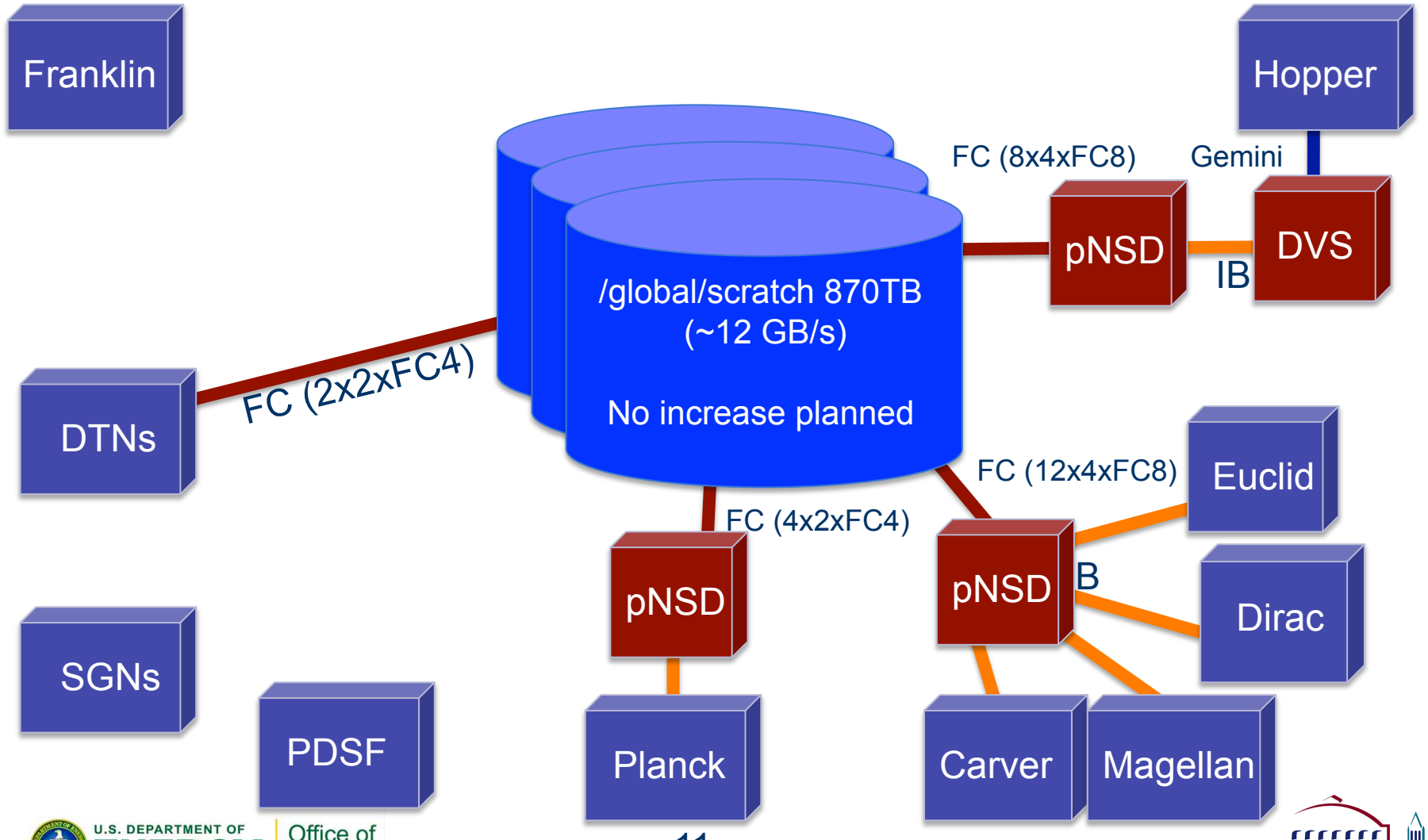
NGF Filesystems

- Collaborative - /project
  - 873 TB, ~12 GB/s, served over FC-8
  - 4 DDN 9900
- Scratch - /global/scratch
  - 873 TB, ~12 GB/s, served over FC-8
  - 4 DDN 9900s
- User homes – /global/u1, /global/u2
  - 40 TB, ~3-5 GB/s, served over Ethernet
  - HDS 2300
- Common area - /global/common, syscommon
  - ~5 TB, ~3-5 GB/s, served over Ethernet
  - HDS 2300

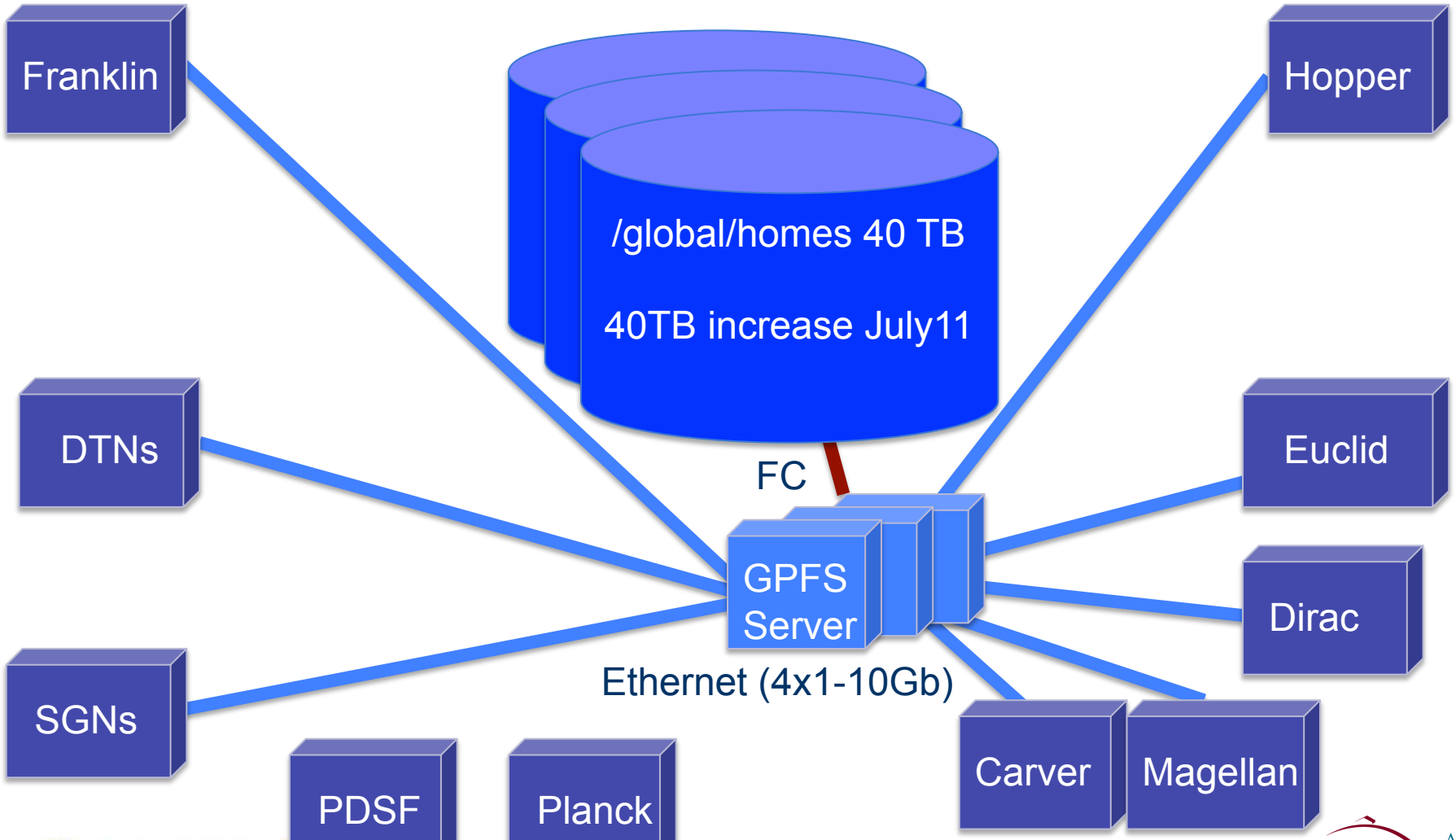
NGF /project


Franklin
Sea*
DVS
FC (20x2xFC4)
DTNs
FC (2x2xFC4)
Ethernet (2x1 10Gb)
SGNs
GPFS Server
PDSF
/project 870TB
(~12 GB/s)
730TB increase July11
FC (4x2xFC4)
pNSD
Planck
FC (8x4xFC8)
pNSD
Hopper
Gemini
DVS
IB
FC (12x4xFC8)
pNSD
Euclid
B
Dirac
Carver
Magellan

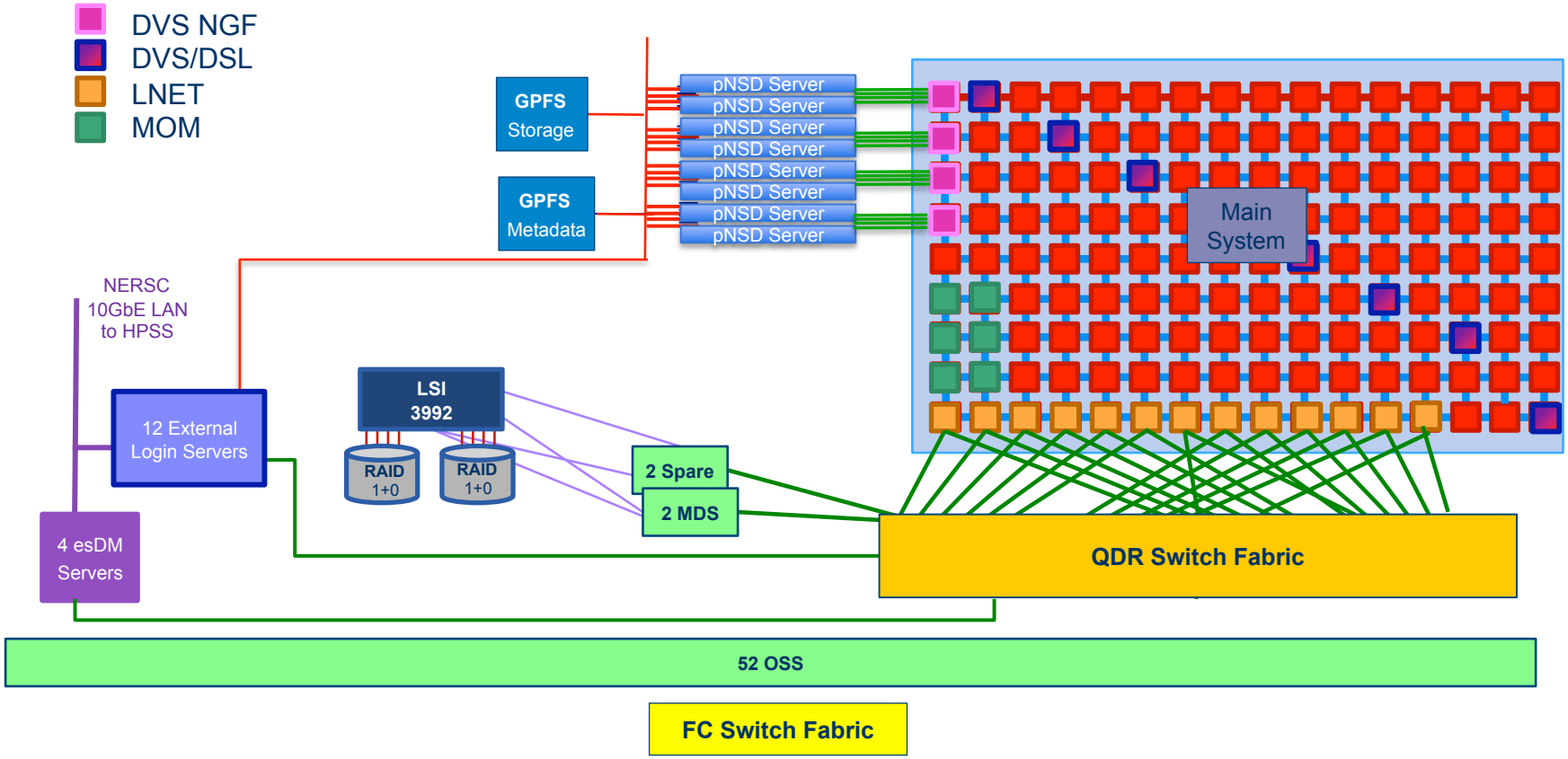
NGF global scratch



NGF global homes



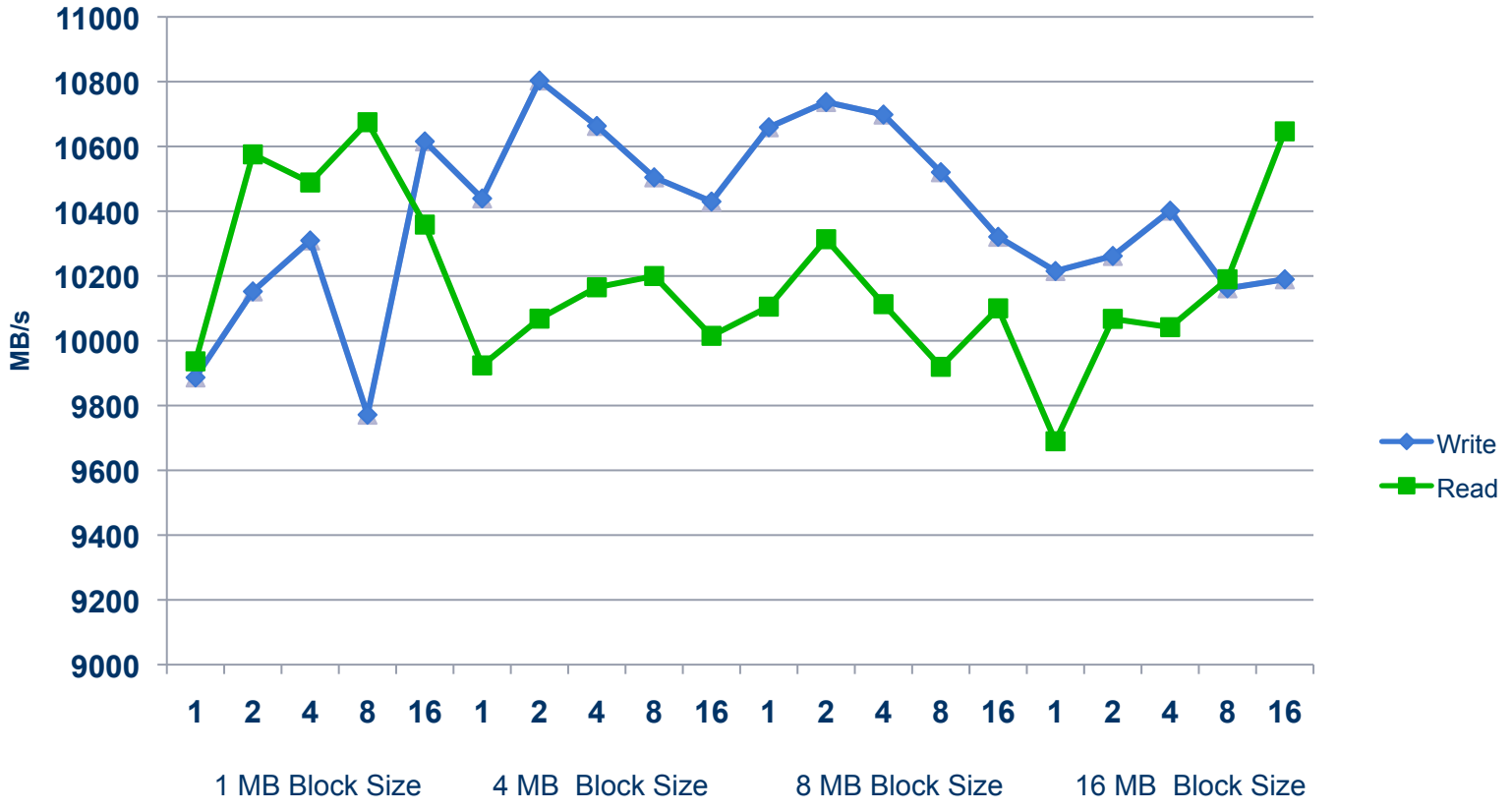
Hopper Configuration



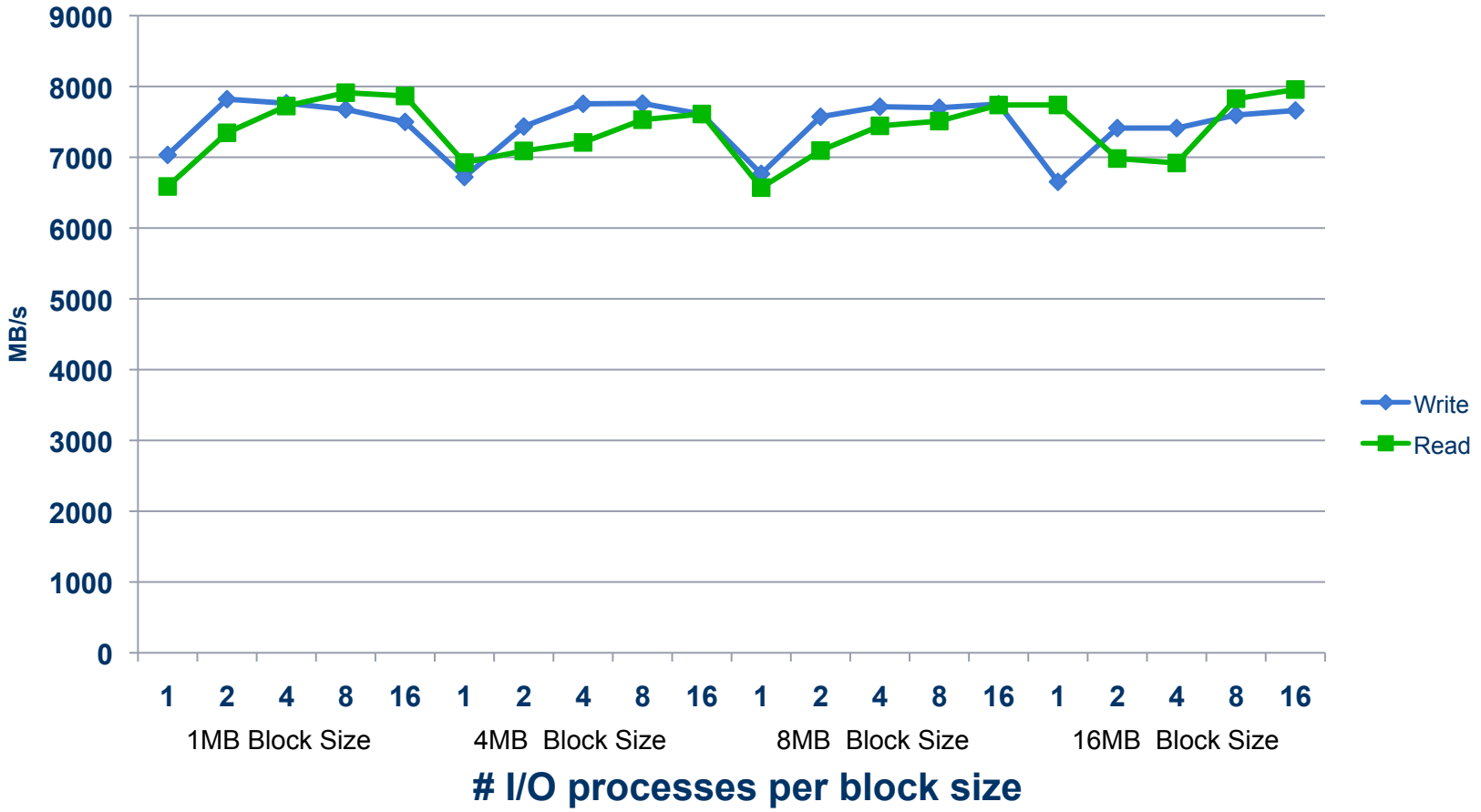
DVS on Hopper

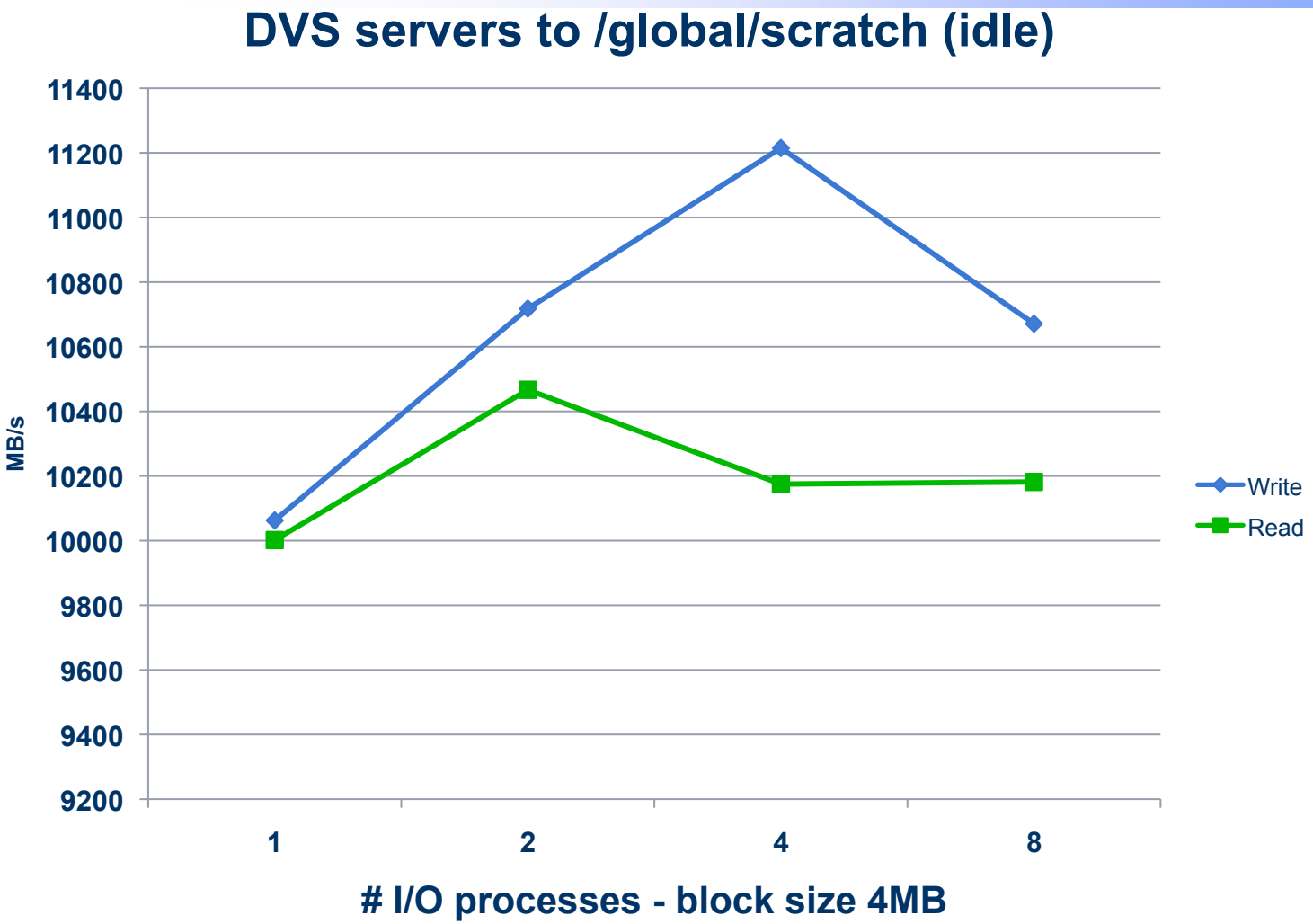
- 16 DVS servers for NGF filesystems
  - IB connected to private NSD servers
  - GPFS remote cluster serving compute and MOM nodes
  - 2 DVS nodes dedicated to MOMs
  - Cluster parallel
- 32 DVS DSL servers on repurposed compute nodes
  - Loadbalanced for shared root

pNSD servers to /global/scratch (idle)

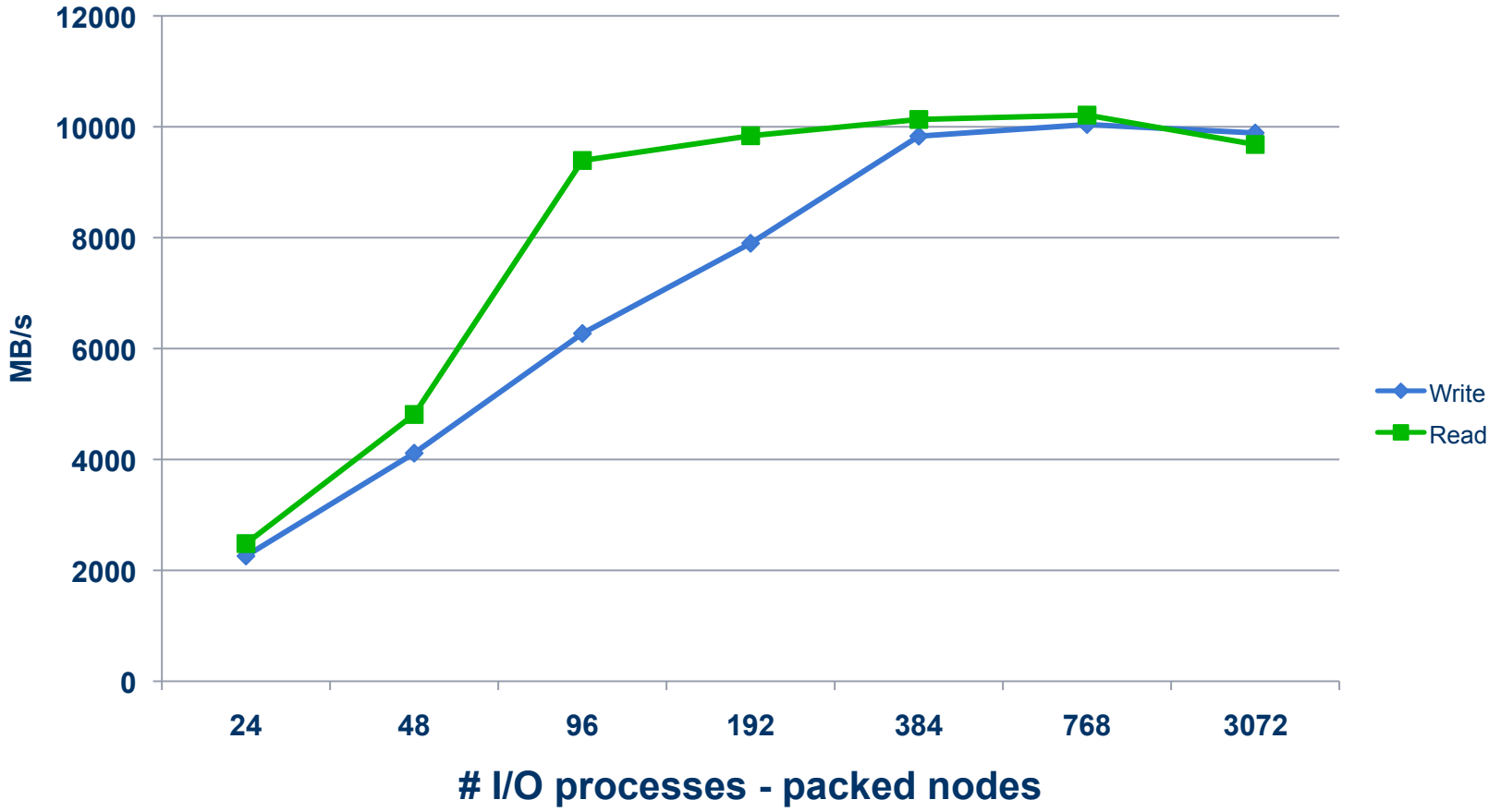


#I/O processes per block size

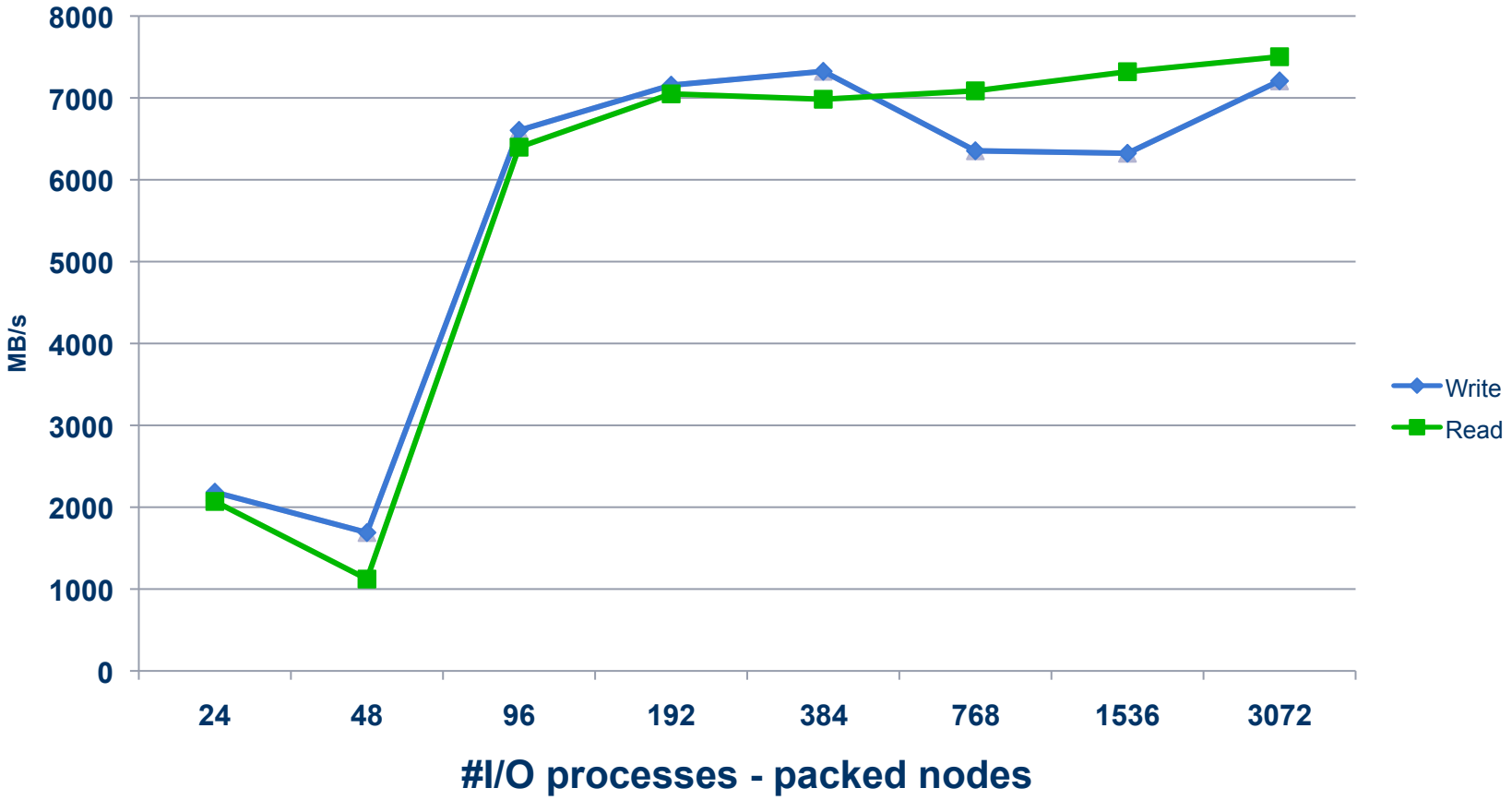

pNSD servers to /global/scratch (busy)
MB/s
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1 2 4 8 16 1 2 4 8 16 1 2 4 8 16 1 2 4 8 16
1MB Block Size
4MB Block Size
8MB Block Size
16MB Block Size
I/O processes per block size
Write
Read


DVS servers to /global/scratch (idle)
11400
11200
11000
10800
10600
10400
10200
10000
9800
9600
9400
9200
MB/s
1
2
4
8
I/O processes - block size 4MB
Write
Read

Hopper compute nodes to /global/scratch (idle)



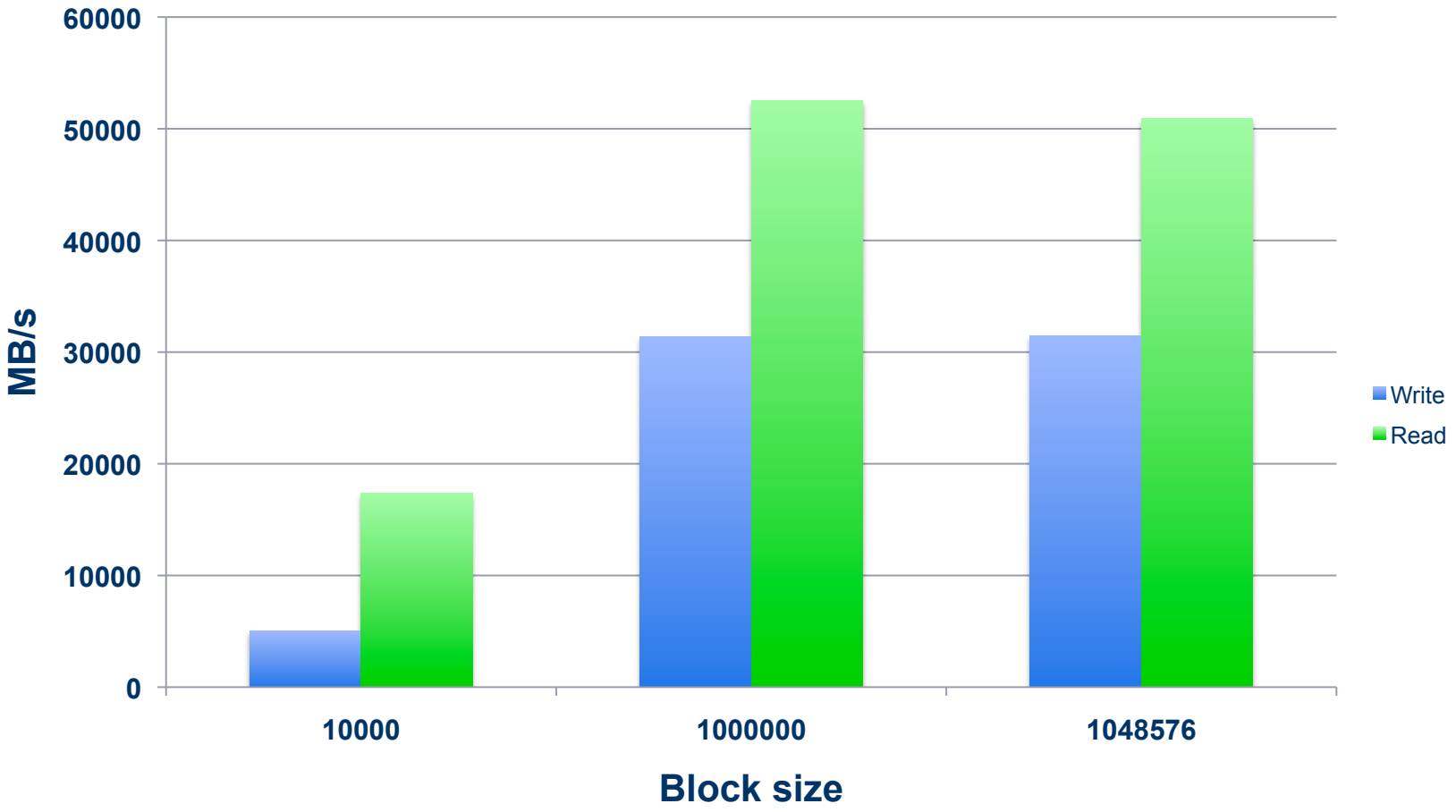
Hopper compute nodes to /global/scratch (busy)



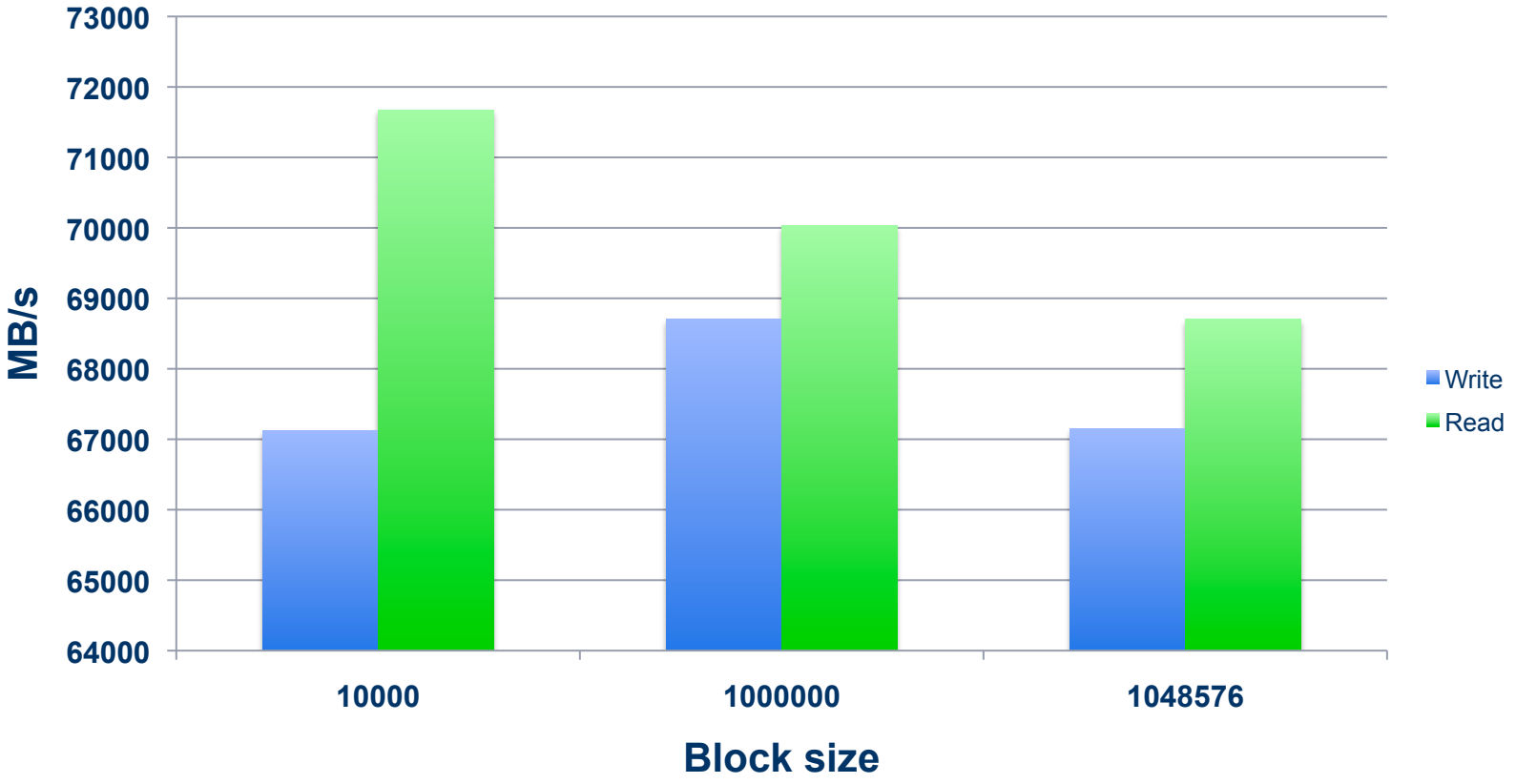
Hopper Filesystems

- External Lustre
  - 2 local scratch filesystems
  - 2+ PBs user storage
  - Aggregate 70 GB/s
- External nodes
  - 26 LSI 7900
  - 52 OSSes with 6 OSTs per OSS
  - 4 MDS with failover
- 56 LNET routers

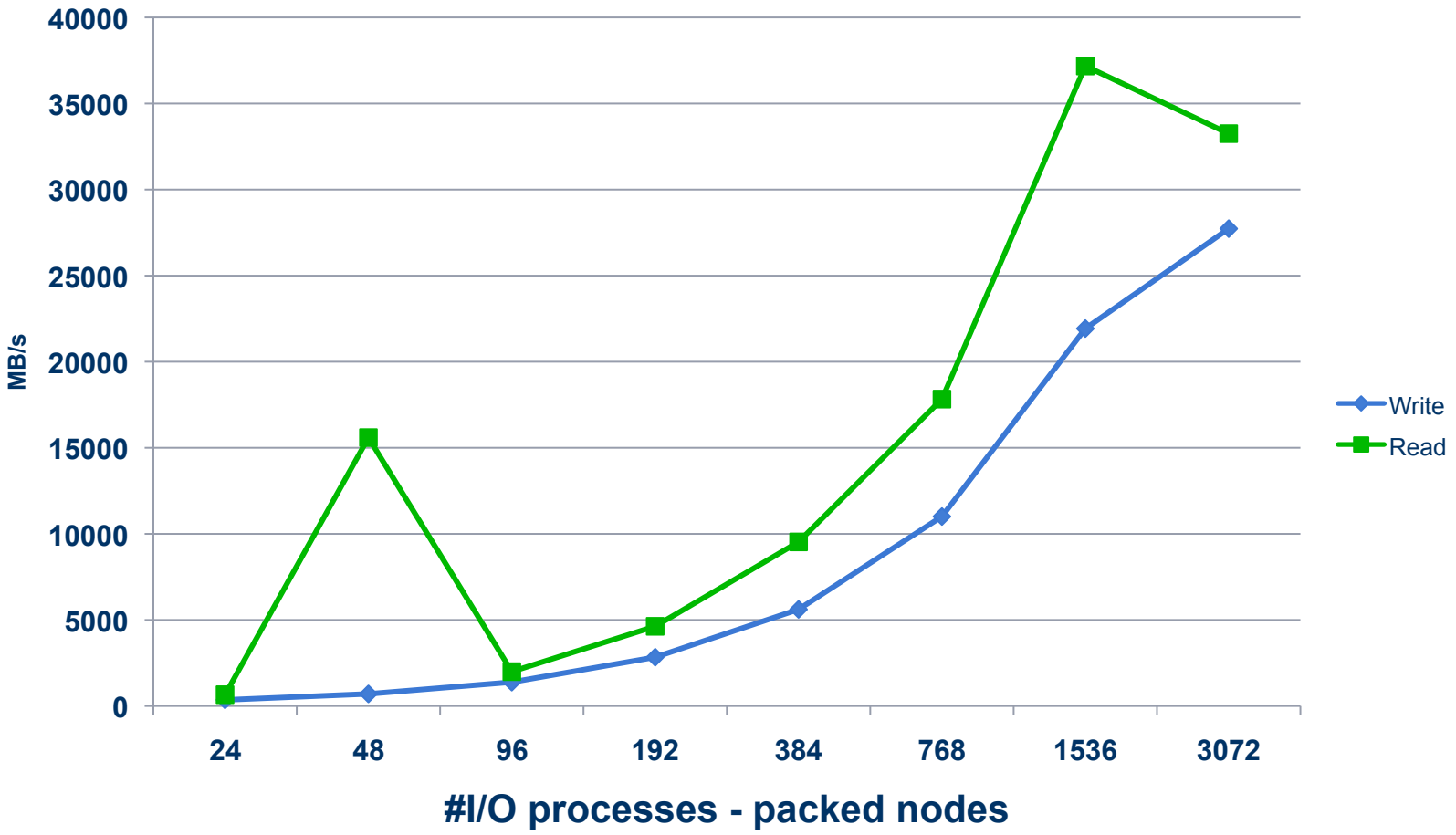
IOR 2880 MPI Tasks MPI-IO Aggregate



IOR 2880 MPI Tasks File Per Processor -- Aggregate



Hopper compute nodes to /scratch (lustre)



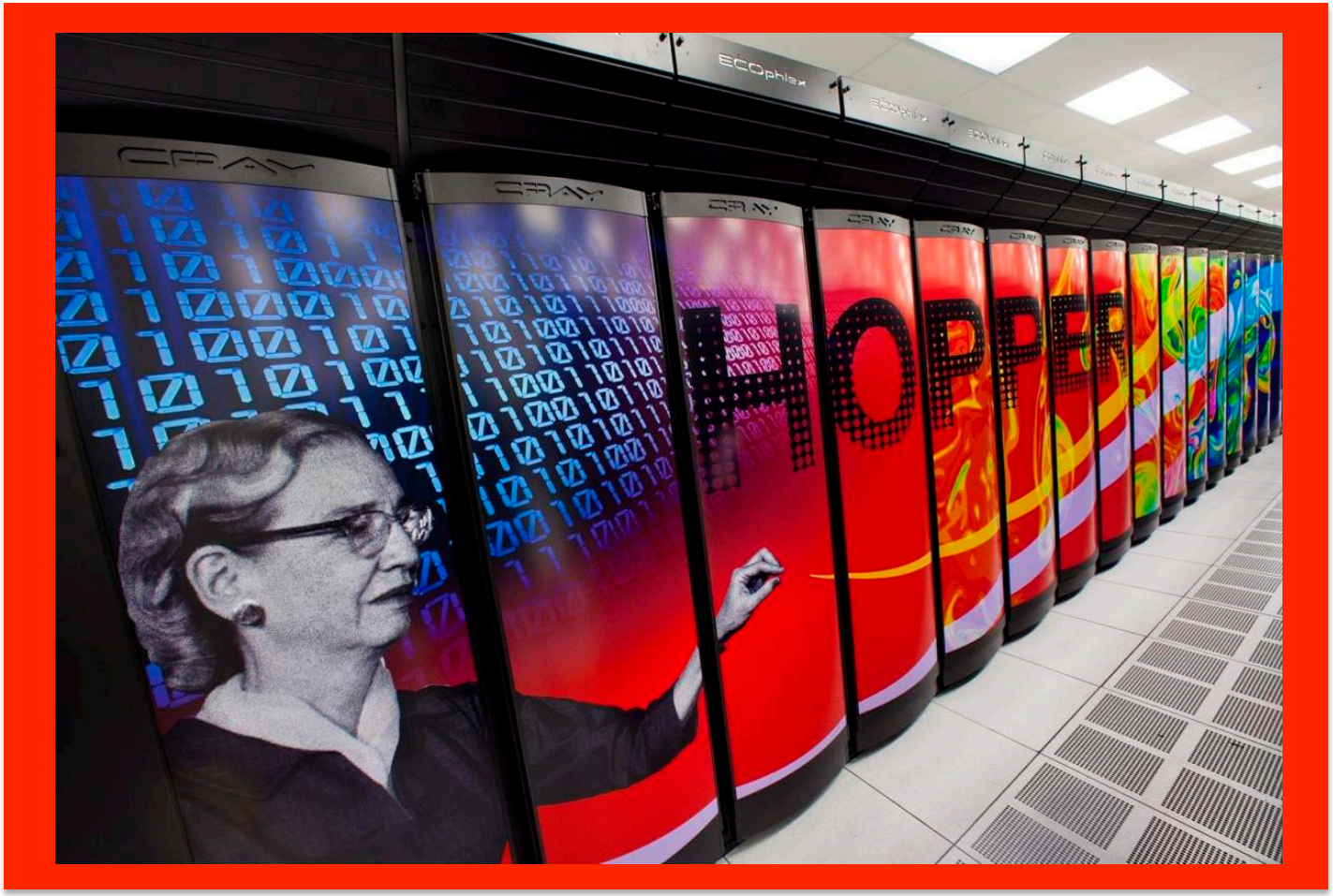
Conclusions

- The mix of dedicated external Lustre and shared NGF filesystems works well for user workflows with mostly good performance.
- Shared file I/O is an issue for both Lustre and DVS-served filesystems.
- Cray and NERSC working together on DVS and shared file I/O issues through Center of Excellence.

Acknowledgments

This work was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC02-05CH11231.

This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy.


ECOphlex
CRAY
HOPPER