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

Tina Butler, **Rei Chi Lee** *and* **Gregory F. Butler**,
National Energy Research Scientific Computing Center

ABSTRACT: Providing flexible, reliable and high performance access to user data is an ongoing problem for HPC centers. This paper will discuss how NERSC has partitioned its storage resources between local and global filesystems, and the configuration, functionality, operation and performance of these several different parallel filesystems in use on NERSC's Cray XE6, Hopper.

KEYWORDS: Filesystems, GPFS, Lustre, DVS

NERSC

The National Energy Research Scientific Computing Center (NERSC) is the primary high performance computing (HPC) facility for scientific research sponsored by the Office of Science in the US Department of Energy. NERSC is operated by Lawrence Berkeley National Laboratory (LBNL) and provides computing resources for more than 4000 researchers worldwide annually. NERSC has a highly diverse workload and supports applications at a wide range of scales and computational intensity.

NERSC systems

To support that range of users and applications, NERSC fields a new large system about every three years, and keeps each large system for 6-7 years. Smaller mid-range or special purpose systems are installed periodically.

Current NERSC systems include:

Hopper (NERSC-6), a Cray XE6, with 6384 24-core nodes, 1.28 PF peak.

Franklin (NERSC-5) a Cray XT4 with 9532 4-core nodes, 356 TF peak,

Carver, an IBM iDataplex cluster with 400 8-core nodes

Magellan, an IBM iDataplex cluster; cloud computing testbed

PDSF, throughput commodity cluster

Euclid, a Sunfire analytics system

Dirac, a GPU testbed

HPSS, an archival HSM storage system

NGF, the center-wide shared filesystem

NERSC and Global Filesystems

At the end of the 1990's NERSC was procuring and integrating a large IBM SP3 cluster, Seaborg. As part of that procurement, NERSC chose to provide all user storage on Seaborg using IBM's Global Parallel File System (GPFS). GPFS was quite new at that time, and there were some concerns about using a filesystem originally designed for multi-media streaming for user home directories. However, after some growing pains, and a number of problem reports and fixes, GPFS proved to be a reliable and high performance filesystem.

Around the same time, it became clear that user data was an increasingly significant issue from both management and productivity perspectives. Users run codes and do pre- and post-processing on multiple machines, and therefore need access to their data in multiple places. Copying files around and keeping them synchronized is a waste of researchers valuable time and a waste of significant amounts of storage.

In 2001, NERSC started the Global Unified Parallel File System (GUPFS) project. The goal of the project was to provide a scalable, high performance, high bandwidth shared filesystem for all NERSC production computing and support systems. The primary purpose of the GUPFS project was to make it easier to conduct advanced scientific research using NERSC systems. This was to be accomplished through the use of a shared filesystem providing a unified file namespace, operating on consolidated shared storage that could be directly accessed by all NERSC production systems. The first phase of the project was to evaluate emerging technologies in shared/clustered filesystem software, high performance storage area network (SAN) fabrics and high performance storage devices to determine which combination of hardware and software best fulfilled the requirements for a center-wide shared filesystem. The second phase of the project was a staged deployment of the chosen technologies in a production environment. A significant number of candidates for the various technologies were assessed on the GUPFS testbed system.

In October 2005 the first center-wide shared filesystem instance was deployed at NERSC based on DDN and Engenio storage, Fibre Channel (FC) fabric and GPFS. GPFS's resilience, remote clustering features and data management policies made it the choice for the NERSC's center-wide filesystem. This first instance of the NERSC Global Filesystem (NGF), /project, was designed to support collaborative projects that used multiple systems in their workflow. Initially /project provided 70 TB of high performance, permanent storage served over Fibre Channel and Ethernet; it has been extremely popular and has grown to 873 TB, with expansion to 1.5 PB coming later this year.

After the demonstrated success of /project, NERSC deployed NGF instances for user home directories (/global/u1 and /global/u2) and system common areas for shared software (/global/common). These smaller filesystems are served over 10 Gb Ethernet and are tuned for smaller block sizes and moderate performance. The latest NGF filesystem put in production is a global scratch area with size and performance similar to /project. All NERSC production systems are using NGF filesystems; Carver, Magellan, Euclid and Dirac are using NGF exclusively – they have no local user-writable storage.

NGF and Cray systems

At the time that NGF was entering production, NERSC's largest system was Franklin, a Cray XT4. Cray's filesystem of choice for XT systems has been Lustre. The standard Lustre installation on a Cray system includes directly attached storage arrays and metadata and object storage servers on service nodes. Access to the Lustre filesystem(s) from compute nodes running a lightweight OS is provided by a lightweight Lustre client.

Franklin was delivered with local Lustre filesystems for user home directories and scratch space. Making NGF filesystems accessible on Cray systems required a new software layer. Directly mounting GPFS filesystems on Franklin's 9,532 compute nodes was not technically or financially feasible. Fortunately, Cray had acquired the rights to the Data Virtualization Service (DVS) software originally developed by Extreme Scale (later Cassatt). DVS essentially provides I/O forwarding services for an underlying filesystem client. While XT login nodes were able to directly mount the NGF filesystems, using DVS on XT service nodes allowed the GPFS-based filesystems to be accessed by XT compute nodes. This capability was a major productivity gain for NERSC users.

Hopper's externalized architecture

Hopper, NERSC's newest system, is a Cray XE6. Hopper is the first large system that NERSC has procured since NGF was deployed as a production resource. The existence of NGF was a major factor in the architecting of the Hopper system. First, Hopper has two external scratch filesystems; the filesystems are Lustre-based, but were designed to be easily detached and incorporated into NGF at a later time if NERSC so chose. In the external filesystems, the metadata and object storage servers are commodity Linux servers – the connection to the interior of the XE6 is via Lustre Routers (LNET) on XIO service nodes.

NERSC also chose to externalize the login nodes for Hopper. At the time that Hopper was being procured, Franklin was having issues with overloaded internal login nodes, and interconnect instability. Externalizing the login functionality allowed the incorporation of commodity servers with more memory, processors and swap space than internal service nodes. The external login nodes and external filesystems were seen as a way to provide an environment where users can access their files, do pre- and post-processing, develop codes, and submit jobs even if the large compute resource was unavailable. The external environment was designed to have as few dependencies on the internal XE6 resources as possible. The external login nodes also have all NGF filesystems mounted. DVS is used on Hopper to project the NGF filesystems to compute nodes. Hopper external nodes are shown in Table 1.

Another benefit of the externalized server design became apparent when NERSC and Cray agreed to a phased delivery for the Hopper system. The first phase of Hopper was a modestly sized XT5, but included all the externalized login and filesystem servers and storage. The intent was to integrate the external storage and login nodes into the NERSC infrastructure in Phase 1 and resolve any issues prior to Phase 2 delivery. NERSC ended up being able to keep the Phase 1 system available for users while the Phase 2 XE6 was installed and integrated by splitting the login nodes and external scratch filesystems between the two systems. When the time came to move the second external Lustre filesystem to the XE6, the lifetime of the XT5 system was extended by transitioning user files to the NGF-based /global/scratch.

Quantity	Node Type
12	External Login Nodes
4	External Data Mover Nodes
52	Storage server nodes (Lustre OSS)
4	Metadata server nodes (Lustre MDS)
26	LSI Engenio 7900 storage system
208	16-slot drive enclosures
3120	1 TB 7.2K RPM SATA drives
2	LSI Engenio 3992 for metadata
24	450 GB 15K RPM FC drives for metadata
8	External pNSD Nodes
2	External Management Nodes

Table 1: Hopper External Nodes

For the Phase 2 delivery, another type of external server was introduced into the Hopper configuration. On the Phase 1 system, the DVS nodes for NGF had both Fibre Channel and 10 GbE interfaces and were directly attached to the NGF SAN, just as was done on Franklin. The new XIO blades for Phase 2 have only 1 slot per node. Also, as NGF has grown, it has become apparent that the FC SAN is stressed and that it would be beneficial to reduce the number of initiators on the fabric. Fortunately, GPFS supports a feature called private network shared disk servers or pNSDs. pNSDs are members of the GPFS owning cluster that are dedicated to serving a particular remote cluster. pNSDs provide dedicated performance as well as insulating the remote cluster from events that would otherwise require disruption of client nodes. Hopper's 8 pNSDs are connected to the Phase 2 DVS servers via QDR InfiniBand with metadata traffic routed through the network nodes.

Hopper's configuration

Hopper's internal node types and counts are shown in Table 2; overall system configuration is shown in Figure 1. The compute partition of the system has 6384 dual socket 2.1 GHz Magny Cours 12 core-based nodes providing 153,216 cores. Hopper can produce 3,677,184 CPU hours per day.

To support the compute partition, Hopper has several types of service nodes. Services that require an external interface like Lustre routers (IB), network nodes (10 GbE), or DVS (FC or IB) are sited on the latest version of Cray's service node, the XIO. Services that do not require direct outside access like Torque or PBS MOM nodes can be put on repurposed compute nodes. This allows more flexible configuration of these services and provides the greater resources of the compute node to the serial portion of the running application. On Franklin, there have been frequent problems with oversubscription of resources on MOM nodes. This has been mitigated on Hopper through the use of repurposed compute nodes. The other type of service node that can reside easily on a repurposed compute node is a shared root DVS server. These servers are used to project a shared root to compute nodes for applications using Dynamic Shared Libraries (DSL) or Cluster Compatibility Mode (CCM). CCM is not yet configured on Hopper, but is being investigated as an option for users whose workflows do not translate easily to the standard CLE model.

Quantity	Node Type
6384	Dual-socket 2.1 GHz 12 core Magny Cours Compute Nodes
24	MOM Nodes. (on Compute blades)
56	Lustre Router Nodes
32	Shared-root DVS Server Nodes (on Compute blades)
16	DVS Server Nodes
2	Network Nodes
4	RSIP Server Nodes
2	Boot Nodes
2	Syslog and System Database Nodes

Table 2: Hopper Internal Nodes

Hopper's filesystems

It is useful to think about the Hopper system as having an inside and an outside – the inside is the XE6 compute partition with its set of compute nodes and service nodes, all connected by the high performance Gemini interconnect. The outside is the cluster of external servers for logins, filesystems and data movement. The inside and the outside are linked by sets of filesystems. The first set of filesystems is the high performance external Lustre filesystems that are used for scratch space on Hopper. Each of the 2 scratch filesystems consists of 13 LSI 7900 storage subsystems each with 8 trays of 15 1 TB disk drives configured as 8+2 RAID 6. Each 7900 has 12 LUNs and 2 OSS nodes configured as failover pairs serving them. Each filesystem has two MDS nodes also configured for failover. The OSS nodes connect to the 56 LNET nodes on the inside of Hopper through a QDR InfiniBand fabric. The scratch filesystems are mounted on all Hopper internal and external nodes that are user-accessible through either direct login or batch submission – login, MOM, compute, and data mover nodes.

The other set of filesystems is the group of GPFS-based globally mounted filesystems provided by NGF. These filesystems are designed to present a consistent environment for users across platforms and to facilitate data sharing and reuse. NGF provides permanent space for user home directories (u1 and u2), a collaborative area (project), and space for applications and utilities maintained by NERSC for the user community (common). High performance temporary storage is provided by global scratch. Figure 2 shows the general layout of NGF.

Filesystem Performance

The external Lustre filesystems on Hopper were specified to provide at least 70 GB/s of aggregate bandwidth for user applications. The Hopper external scratch filesystems are at Lustre 1.8.4. NERSC uses IOR as one benchmark for filesystem performance. IOR aggregate results are shown in Figure 3 and 4. Figure 5 shows a set of IOR runs against /scratch, one of the external Lustre filesystems. These runs were made before final tuning was done on the filesystem, but they show the filesystem achieving close to design performance of 35 GB/s aggregate (per filesystem).

IOR results for Posix file per process tests have shown the best performance. Using Cray's iobuf library to buffer mismatched block sizes made a considerable improvement at the 10,000 and 1,000,000 byte block size tests. Shared file/MPI-IO performance is significantly less. NERSC is working with Cray to diagnose and improve issues with shared file I/O performance. Lustre at scale is also showing increased variability in performance. This has been dubbed the "slow LUN" problem, and appears to be related to poor placements of extents on some LUNs. IOR is particularly sensitive to this type of single thread slowdown.

Performance data for the NGF filesystem /global/scratch on Hopper is shown in Figures 6-10. /global/scratch is a shared resource; multiple systems have it mounted. Figures 6 & 7 show performance from the Hopper pNSD servers to NGF, as measured by lmdd, for both busy and relatively idle times on the filesystem. The idle case (Fig. 6) shows that the Hopper private NSDs are not getting the measured maximum bandwidth to /global/scratch of about 12 GB/s; the theoretical maximum is more like 20 GB/s. Work is in hand to reconfigure the Fibre Channel fabric to recover that lost bandwidth. The busy case shows lower (about 6.5 – 8 GB/s) but quite consistent performance to and from /global/scratch for the private NSDs. Figure 8 shows the results of running lmdd on Hopper DVS servers to /global/scratch. This is another mostly idle case, and the DVS servers are getting most of the available bandwidth to the filesystem. Figures 9 and 10 show the results of running IOR on Hopper compute nodes against /global/scratch - again there are busy and idle cases. The idle case shows a maximum performance of around 10 GB/s; the busy case a maximum of around 7.5 GB/s. These results compare well with the maximum available bandwidth achieved by the private NSDs.

Again, performance is not as good when applications are using shared file I/O. On NGF filesystems /global/scratch and /project, this has been attributed to the current DVS configuration where the number of DVS servers per file has been limited to prevent GPFS coherence thrashing. Improving DVS/GPFS performance is the goal of a NERSC/Cray Center of Excellence.

Conclusions

Hopper has been in production at NERSC since May 1, 2011. In the acceptance period when all user time was free, Hopper delivered over 320 million hours to scientific users. One of the reasons this rapid ramp up was possible was because the NGF filesystems provided easy transition from Franklin and the XT5-based Hopper Phase 1. The externalized filesystems allowed the flexibility to move storage resources incrementally from one system to another, keeping users productive through the installation and integration of the full Hopper system. DVS, both in serving NGF filesystems and enabling essentially a full Linux user environment through DSL, has been invaluable.

About the Authors

Tina Butler is a member of the Computational Systems Group at NERSC. Rei Chi Lee and Gregory Butler are members of Storage Systems Group at NERSC and are the principal architects and developers of the NERSC Global Filesystem.

Acknowledgment:

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.

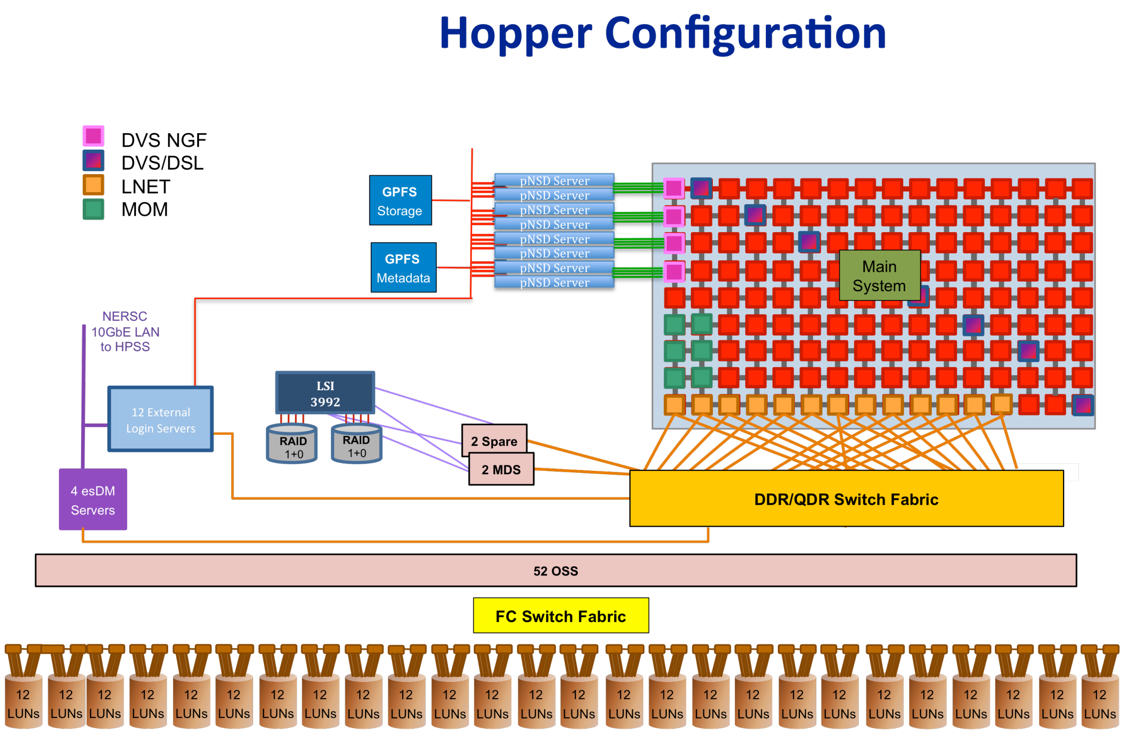

Figure 1. Hopper Configuration

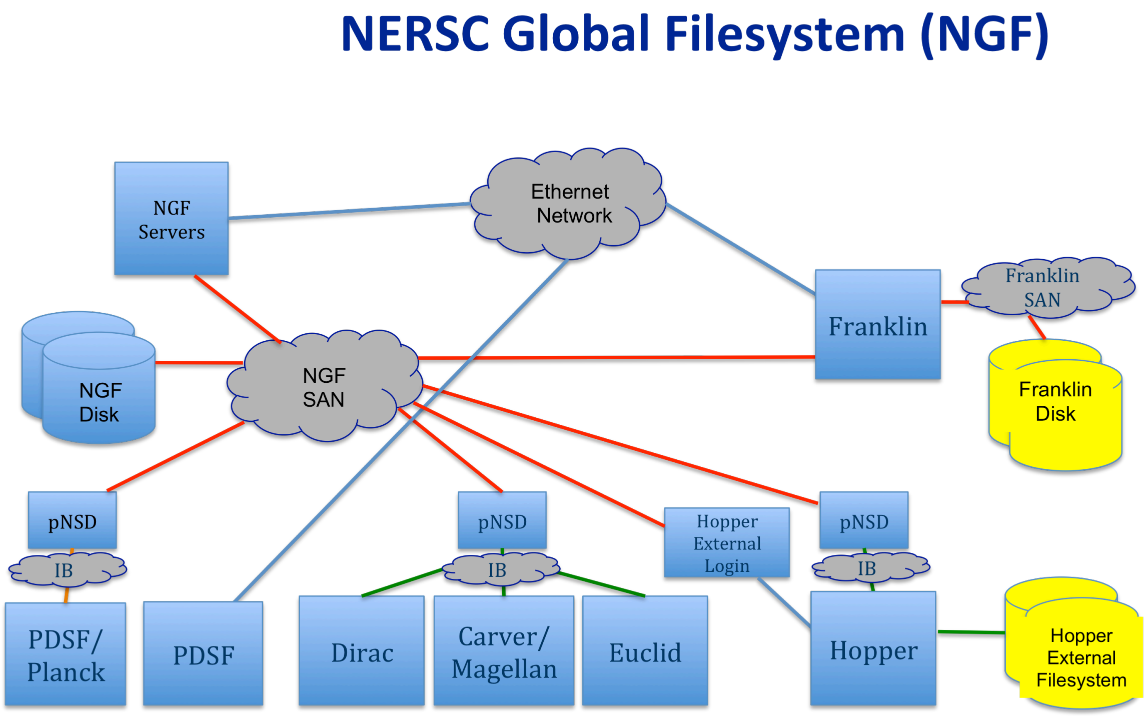

Figure 2. Nersc Global Filesystem

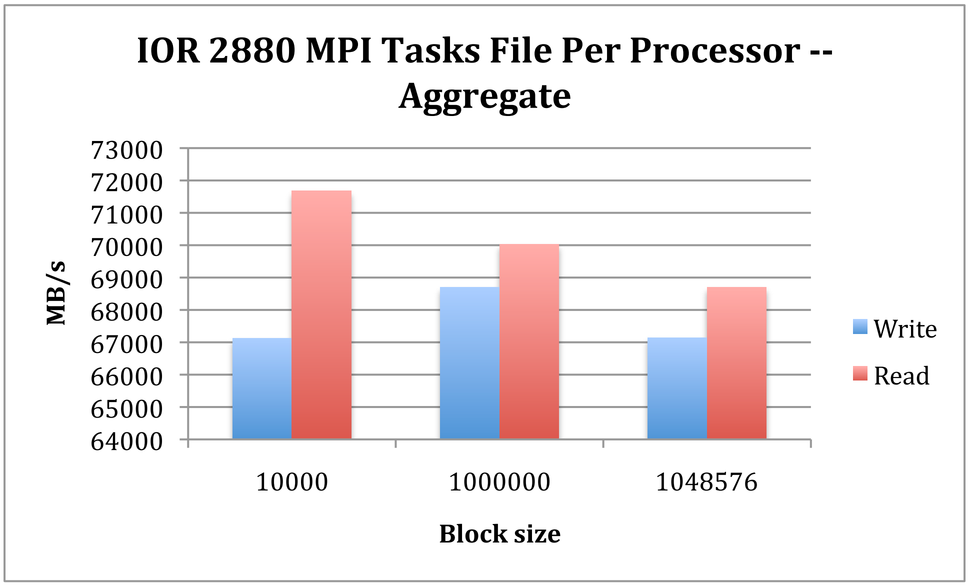

Figure 3. IOR File per Process

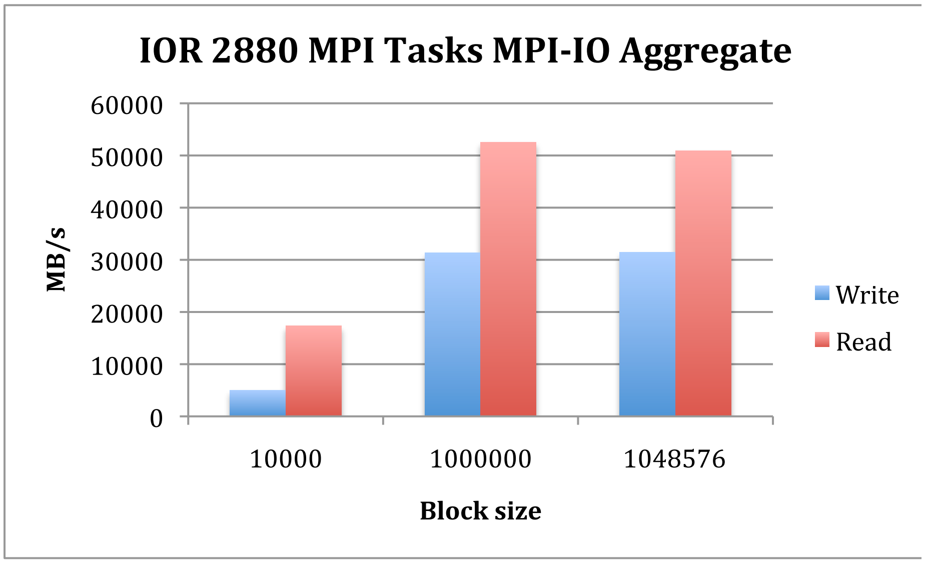

Figure 4. IOR MPI-IO

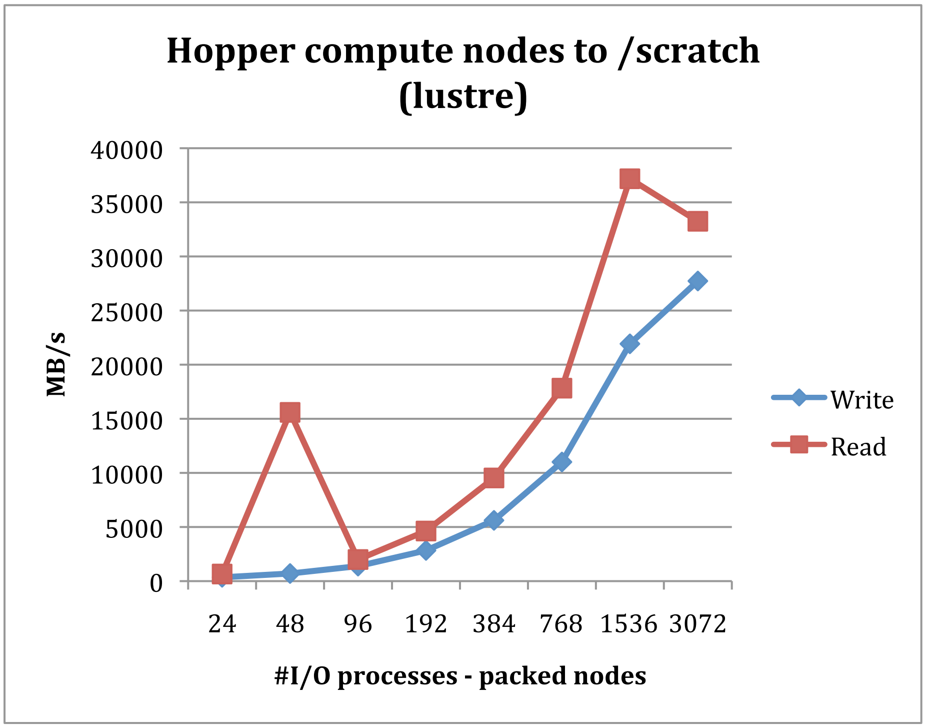

Figure 5. IOR File per process

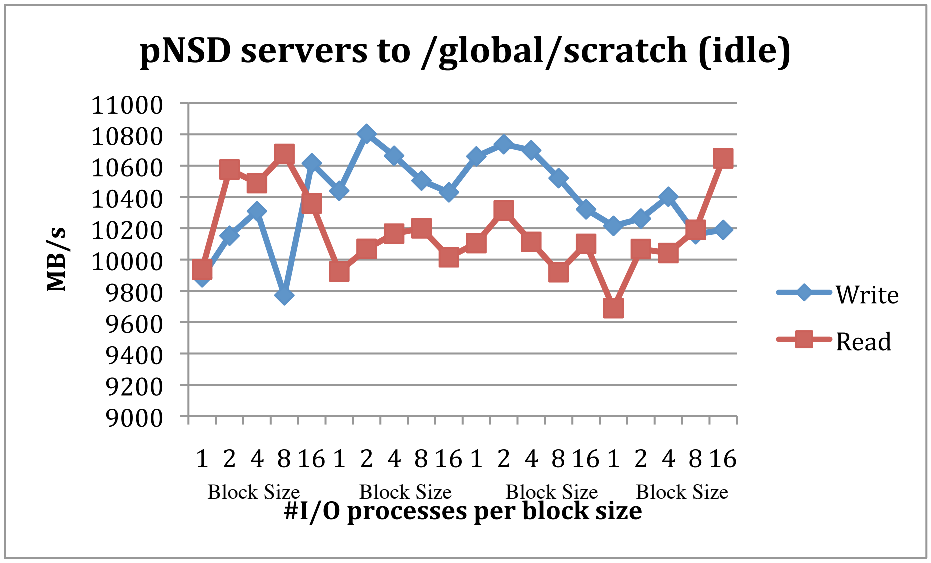

Figure 6. pNSD to /global/scratch (idle)

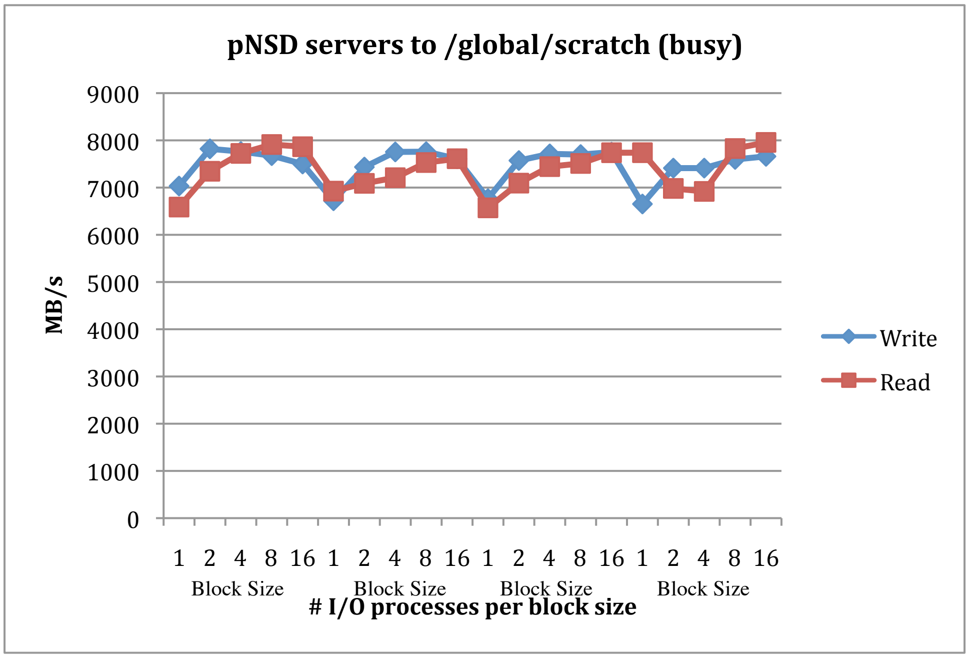

Figure 7. pNSD to /global/scratch (busy)

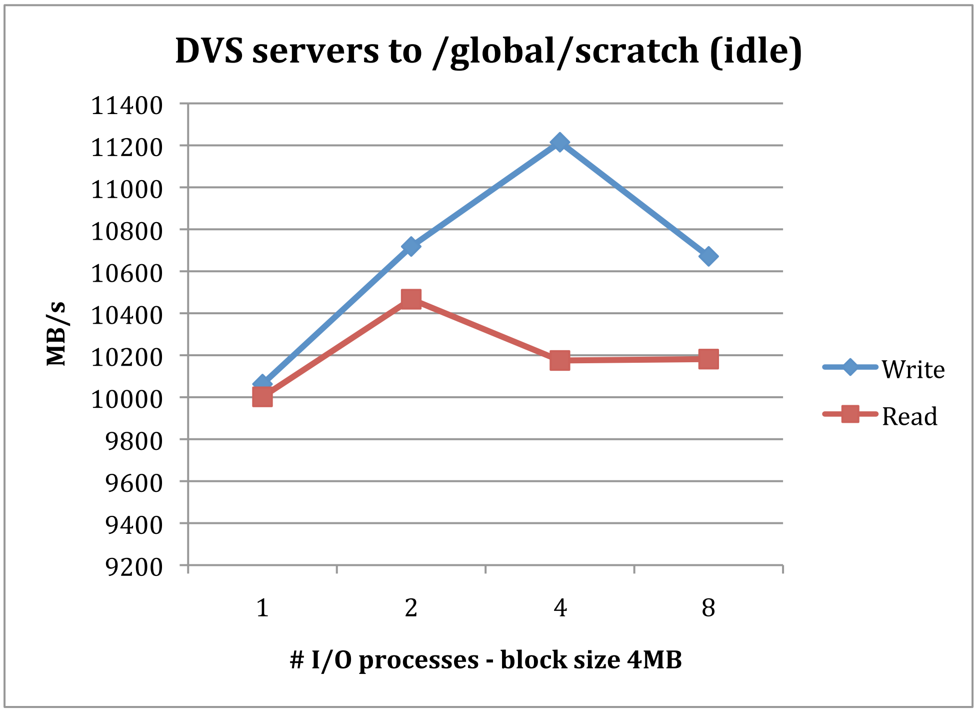

Figure 8. DVS server to /global/scratch (idle)

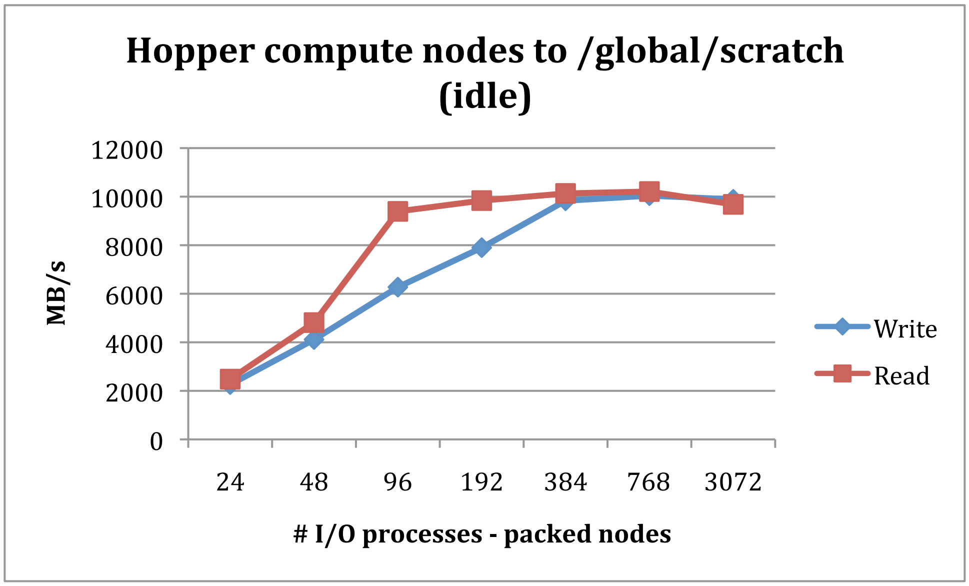

Figure 9. Compute nodes to /global/scratch (idle)

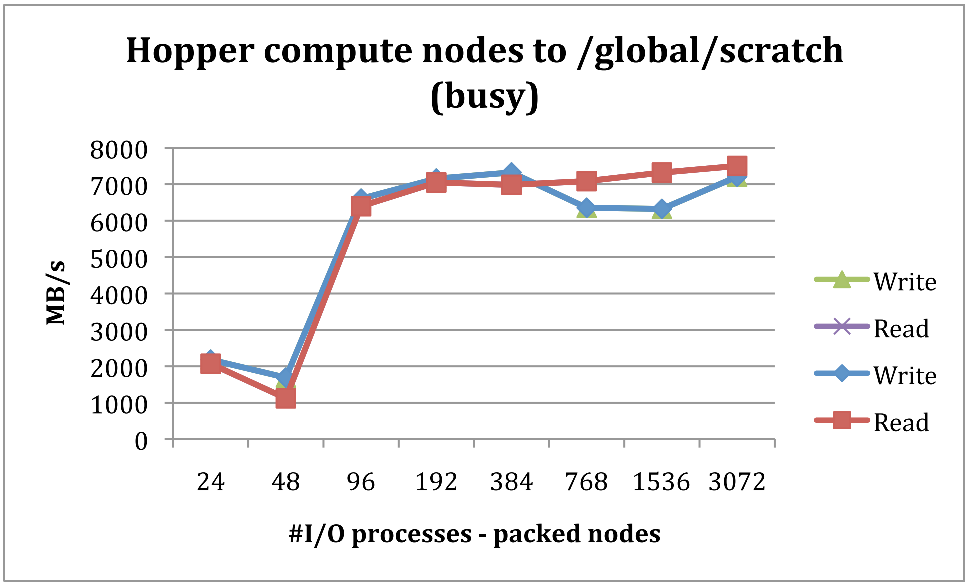

Figure 10. Compute nodes to /global/scratch (busy)