UCRL-TR-217288

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Certification of Completion of Level-2 Milestone 461: Deploy First Phase of I/O Infrastructure for Purple

M. Gary, D. Wiltzius

November 23, 2005

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Introduction

This report describes the deployment and demonstration of the first phase of the I/O infrastructure for Purple. The report and the references herein are intended to certify the completion of the following Level 2 Milestone from the ASC FY04–05 Implementation Plan, due at the end of Quarter 4 in FY05:

Milestone: 461

Title: Deploy First Phase of I/O Infrastructure for Purple

Category: Campaign 11—NA113, Advanced Simulation and Computing

ASC Program Element: Physical Infrastructure & Platforms, Ongoing Computing, PSE, Views, DisCom

The milestone is defined as follows:

"External networking infrastructure installation and performance analysis will be completed for the initial delivery of Purple. The external networking infrastructure includes incorporation of a new 10 Gigabit Ethernet fabric linking the platform to the LLNL High Performance Storage System (HPSS) and other center equipment. The LLNL archive will be upgraded to HPSS Release 5.1 to support the requirements of the machine and performance analysis will be completed using the newly deployed I/O infrastructure. Demonstrated throughput to the archive for this infrastructure will be a minimum of 1.5GB/s with a target of 3GB/s. Since Purple delivery is not scheduled until late Q3, demonstration of these performance goals will use parts of Purple and/or an aggregate of other existing resources."

Milestone integration/interfaces we defined as:

"Deployment of this phase of the I/O infrastructure for Purple requires integration and interface between Ongoing Computing, Physical Infrastructure & Platforms, and S&CS program elements at LLNL. Success will depend on the combined effort of these ASC program elements along with collaboration with IBM platform and archive personnel as well as 10 Gigabit Ethernet vendors."

The milestone was completed August 10, 2005, when the system known as pURPURA demonstrated an aggregate of 2.5GB/s from its four front-end nodes to the Secure Computing Facility (SCF) HPSS archive using the I/O infrastructure deployed as part of this milestone effort. This rate exceeded the milestone minimum of 1.5GB/s by 67%.

The following sections will briefly describe the Platform, Network and Archive architectures, efforts, and accomplishments which together represent the deployment and demonstration of the first phase of the I/O architecture for the Purple machine. These sections are followed by a description of the production demonstration that achieved 2.5 GB/s data rate from pURPURA to HPSS using this infrastructure.

Platform

The 100 TFLOP Purple system was delivered in two phases. The first phase consisted of the delivery of one-sixth of the machine, or 252 nodes, which arrived at LLNL in April 21st, 2005. This system, currently named pURPURA, was assembled and then moved to the Secure Computing Facility (SCF) in June of 2005. The remaining nodes of Purple were delivered in September. After acceptance completes, both Purple and pURPURA will be merged together. Following this, the system will be opened to limited availability planned for early CY06.

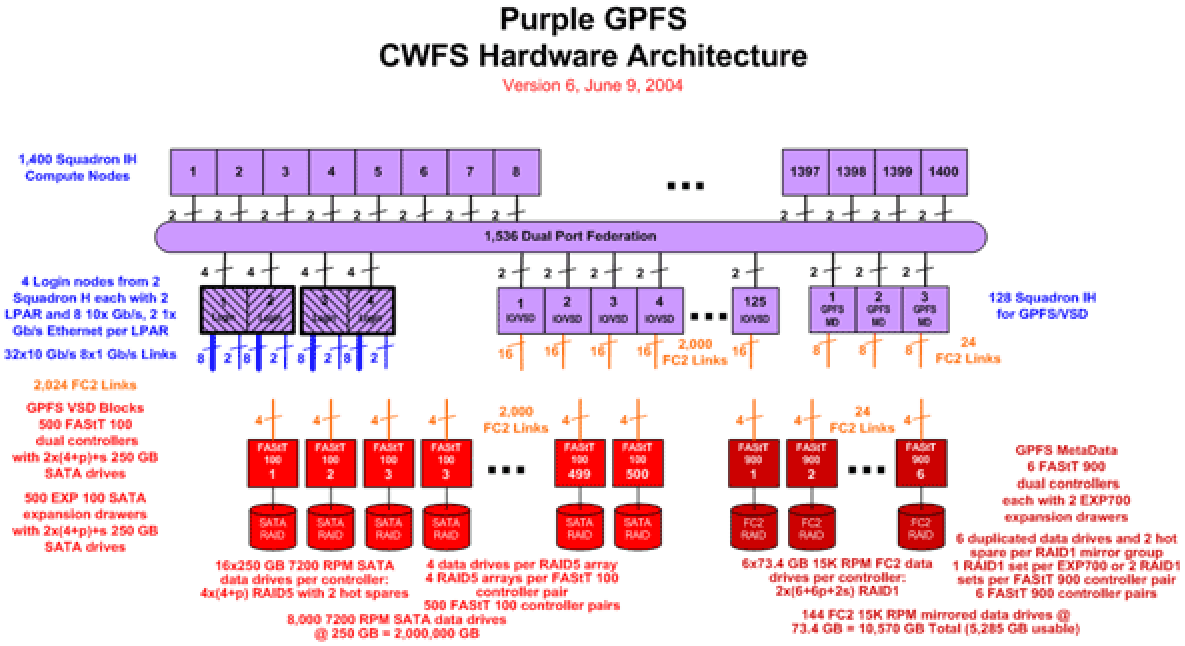
Purple GPFS
CWFS Hardware Architecture
Version 6, June 9, 2004
1,400 Squadron IH Compute Nodes
1
2
3
4
5
6
7
8
1397
1398
1399
1400
1,536 Dual Port Federation
4 Login nodes from 2 Squadron H each with 2 LPAR and 8 10x Gb/s, 2 1x Gb/s Ethernet per LPAR
32x10 Gb/s 8x1 Gb/s Links
1 IO/VSD
2 IO/VSD
3 IO/VSD
4 IO/VSD
125 IO/VSD
1 GPFS MD
2 GPFS MD
3 GPFS MD
128 Squadron IH for GPFS/VSD
2,000 FC2 Links
24 FC2 Links
2,024 FC2 Links
GPFS VSD Blocks
500 FAStT 100
dual controllers
with 2x(4+p)+s 250 GB SATA drives
500 EXP 100 SATA expansion drawers
with 2x(4+p)+s 250 GB SATA drives
FAStT 100 1
FAStT 100 2
FAStT 100 3
FAStT 100 3
FAStT 100 499
FAStT 100 500
SATA RAID
FAStT 900 1
FAStT 900 2
FAStT 900 6
FC2 RAID
GPFS MetaData
6 FAStT 900
dual controllers
each with 2 EXP700
expansion drawers
16x250 GB 7200 RPM SATA data drives per controller:
4x(4+p) RAID5 with 2 hot spares
4 data drives per RAID5 array
4 RAID5 arrays per FAStT 100 controller pair
500 FAStT 100 controller pairs
8,000 7200 RPM SATA data drives
@ 250 GB = 2,000,000 GB
6x73.4 GB 15K RPM FC2 data drives per controller:
2x(6+6p+2s) RAID1
6 duplicated data drives and 2 hot spare per RAID1 mirror group
1 RAID1 set per EXP700 or 2 RAID1 sets per FAStT 900 controller pair
6 FAStT 900 controller pairs
144 FC2 15K RPM mirrored data drives @
73.4 GB = 10,570 GB Total (5,285 GB usable)

Figure 1: Purple Hardware Architecture

Preparing and demonstrating the necessary network and archive infrastructure that will serve Purple was the goal of this milestone. Phase One Purple hardware, the pURPURA machine, was used to demonstrate that the necessary I/O infrastructure was designed, procured, deployed and was ready to satisfy the demands of the Purple machine itself. pURPURA consists of 252 nodes combining 248 Squadron IH Compute nodes (8-way Power5 @ 2.0GHz with 32GB of memory) with 4 Squadron H login nodes(8-way Power5 @ 2.0GHz with 32GB of memory) each with a total of eight 10Gigabit Ethernet connections.

The pURPURA machine was delivered, assembled and tested in LLNL's Open Computing Facility (OCF) and then swung over to the classified SCF facility. Once in the SCF, pURPURA was linked to the newly installed 10GBit Ethernet infrastructure and had the FTP/PFTP, NFT and HTAR file transfer interfaces installed on the login nodes. pURPURA entered Limited Availability on June 27th of 2005.

Networking

At the heart of the off-machine communication for Purple is the network infrastructure linking the platform with external resources including the archive, visualization, and the wide-area network for the tri-Lab remote user community. A system the scale of Purple required a significant I/O infrastructure build-out. It was determined that the final Purple machine would require 32 external 10GigE connections with its Phase 1 pURPURA instance requiring 16 10GigE drops.

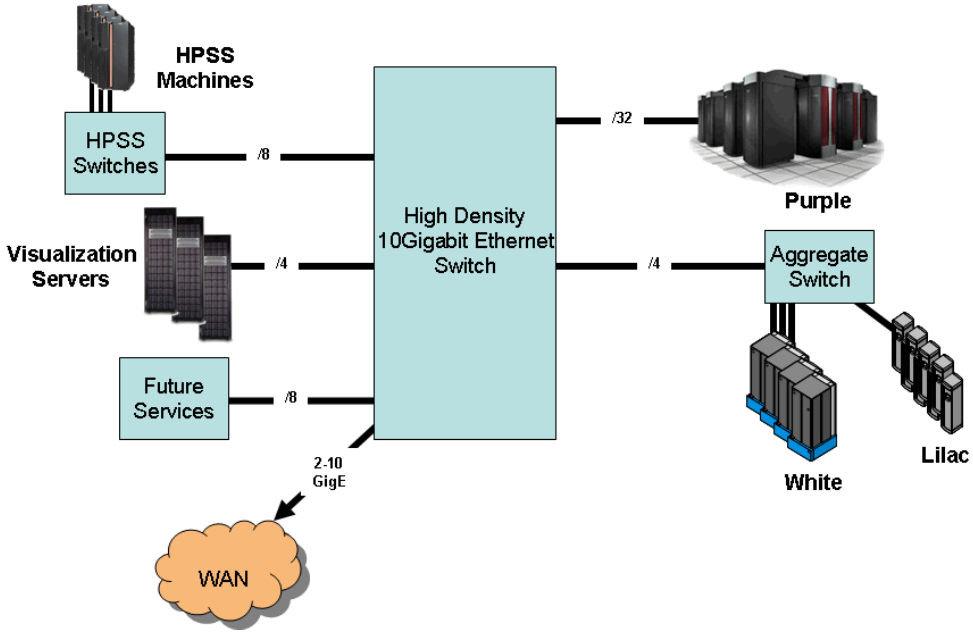

Figure 2: 10Gigabit Ethernet for Purple

The major challenges we faced to meet this requirement were

1) Finding a network switch that would provide a large number of 10Gigabit Ethernet interfaces providing adequate throughput for this immediate need. In CY05 the requirement was for a switch with about 60 10Gigabit Ethernet ports with about 300Gbps throughput (see Figure 2). As the Purple machine and environment grows, this switch will need to scale to provide 600Gbps in CY06. In CY07 we'd expect to provide over 100 10Gigabit Ethernet ports at about 1Tbps.
2) Purple would be the first machine with 10Gigabit Ethernet network interface cards (NICs) that would be able to generate >1Gbps in a data stream. This required some redesign of the I/O architecture of other resources, particularly the HPSS archive and our DisCom WAN connection accessing Purple through this 10Gigabit Ethernet switch.
3) This 10Gigabit Ethernet core and the larger data streams established a need for 10Gigabit Ethernet NICs. Hence we were required to seek cost effective yet high performance 10Gigabit Ethernet NICs.

We have successfully met these challenges. For 1), after meeting with many network vendors to learn about their near and longer term product roadmaps, we identified a product that is not only a good technical solution for the CY05 requirements but for the years beyond. Additionally, the price was substantially less than competing products. Addressing the second requirement required minor additional consideration of the network design for the HPSS archive. However, network modeling and analysis showed that the larger data flows presented by the 10Gigabit Ethernet NICs on Purple could reduce efficiency of the DisCom WAN in some cases. To avoid this potential problem, we purchased and installed new linecards for our edge router for the DisCom WAN to provide additional buffering. Finally, the last requirement was met after testing 10Gigabit Ethernet NICs from several vendors. The testing demonstrated performance of 3-6Gbps, which is adequate. Recently, we found another vendor that recently made available 10Gigabit Ethernet NICs at a fraction of the cost of the ones we've tested, so this activity will be ongoing.

Because of the significant up-front evaluation efforts, the deployment of the switch infrastructure was highly successful and immediately began meeting its performance expectations for the pURPURA platform.

Archive

One of the primary sources and destinations for off-platform data is archival storage. At LLNL both OCF and SCF centers provide state-of-the-art archival storage of data to ASC customers at all three laboratories in archives running the High Performance Storage System (HPSS). HPSS is developed in collaboration with five DOE laboratories and IBM Global Services. HPSS provides a unique blend of scalable, parallel archival storage interfaces and services to customers running at all three ASC centers. Rather than constraining data transfer to the speed of a single storage device, HPSS is designed to distribute data across a configurable amount of storage units and to remove other limits to scaling including number of files, directories, concurrent users, etc.

In order to supply the performance necessary to offload ASC platforms, and not hinder computation, a world-class array of storage hardware is deployed and supported underneath HPSS. This includes high performance disk arrays, tape subsystems, mover nodes, SANs, networks, robotics and petabytes of media. Together, this hardware and software supports unlimited storage for an unlimited amount of time at speeds capable of supporting the needs of ASC platforms.

Purple presented three major challenges for the HPSS system at LLNL.

- Scaling Archive Hardware
- Enhancing Software Infrastructure
- Archive Deployment

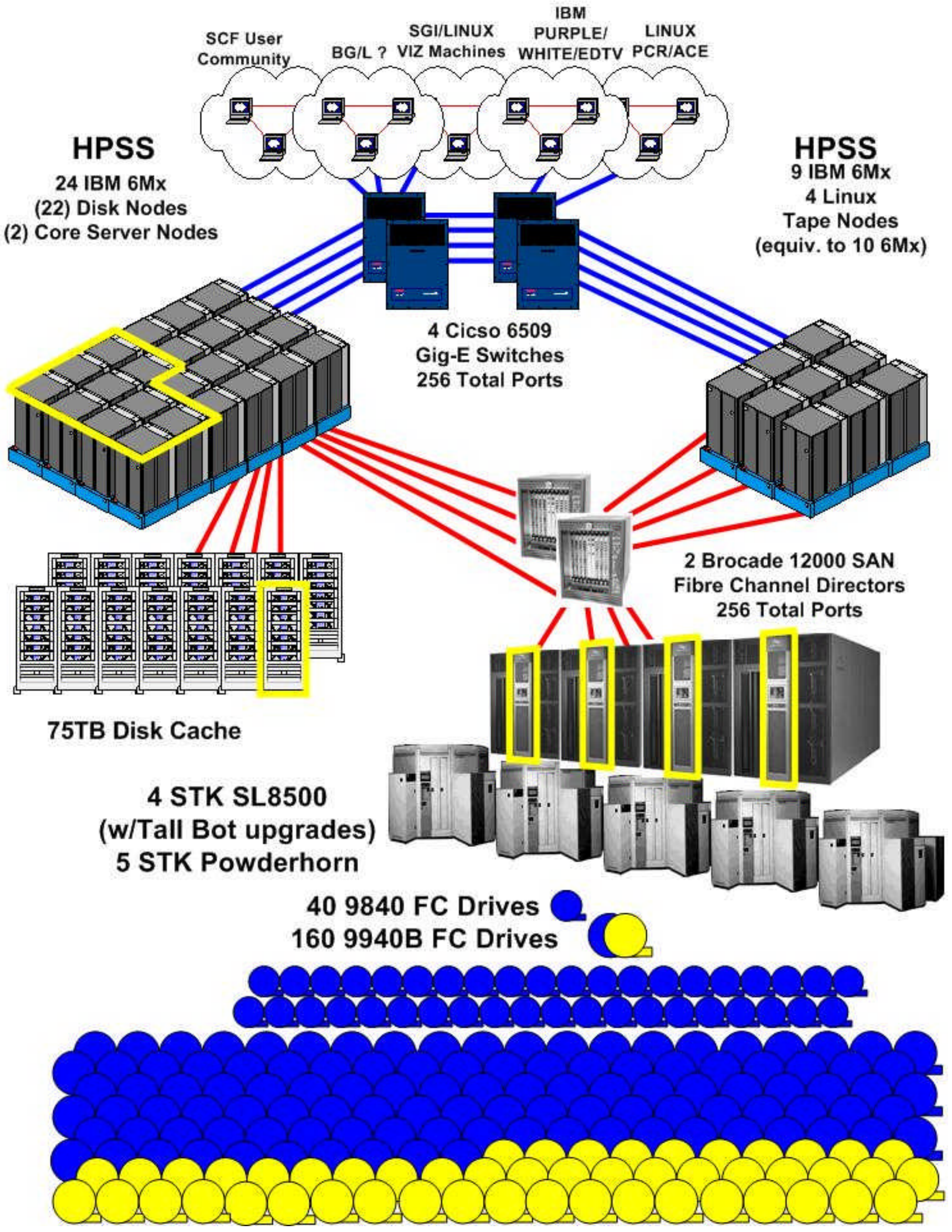

Figure 3: The SCF HPSS Archive

Scaling Archive Hardware

In order to handle the influx of data from Purple, the performance and capacity capabilities of the SCF archive needed to be enhanced. During both years of the lifetime of this milestone, members of the Data Storage Group (DSG) researched, tested, and procured the hardware infrastructure required to support the Purple machine. This work encompassed: data movers, disk cache, core server platforms, robotics, tape drives and media.

Data Movers

Data movers are computers that interface between a network and either a tape or disk device. They are typically machines designed and configured for high I/O throughput. A single data mover may handle one or more of the parallel streams involved in a transfer from a compute platform to archival storage.

In preparation for the Purple, DSG selected IBM p-Series 6M2 processors as Disk Movers. For Tape Movers, Xeon-based Linux nodes (configured as a cluster without interconnect) were chosen along with a number of 6M2 processors. The Linux movers were selected after substantial testing in a testbed environment.

Disk Cache

High speed disk is the first stop for user data entering the LLNL archives. The DSG has considerable experience in selecting, fielding and configuring high performance RAID disk arrays. For the HPSS disk cache we chose, tested and configured Data Direct Networks (DDN) Corporation disk arrays using S2A8500 controllers. These high speed FibreChannel RAID arrays are capable of sustaining over 600MB/s per controller and contain the fastest and most robust internal architecture of any of the arrays researched.

Core Server Platform

A core server platform is home to the "brains" of an HPSS system. The core services of the archive (name service, bitfile service, metadata service, physical volume library, etc) are housed on a single high-speed compute platform. This platform must be fast, reliable and be paired with very robust disks for metadata storage. For this job, DSG selected 8-way IBM p-Series 6M2 processors linked with IBM FastT disk storage. This combination was chosen because it provided a platform capable of hosting AIX, DB2, HPSS and was a single vendor solution between both processor and vitally important metadata disk.

Tape Drives and Media

Half-inch tape media remains the most cost effective archival storage medium for large-scale computer centers. The DSG performed a study of archive costs (disk versus tape) in January of 2005 to verify that this was still true. The study found that tape was:

- 6.7 times cheaper to purchase than disk
- 57 times cheaper than disk for yearly maintenance
- 342 times cheaper than disk for yearly power and cooling
- net 72 times cheaper to maintain on a per-year basis than disk

There is considerable competition in the half-inch tape market, but LLNL also has a considerable history fielding tape devices. In support of Purple we will be fielding three tape species:

1) *StorageTek 9840B Drives*. These enterprise class drives are the current workhorses of the LLNL archive and have proven themselves as excellent devices.

2) *StorageTek T10K Drives*. These drives are the next generation of enterprise drive being produced by StorageTek.

3) *IBM LTO-3 Drives*. These are mid-range tape drives which are more economical than the enterprise class, but at a sacrifice of speed and some robustness. DSG selected LTO-3 drives as a "second-copy" technology, housing one of two copies of select important files. If they prove to be robust enough, we may well be able to leverage them to drive down archive costs in the future.

Robotics

Because of manpower costs, tape automation now dominates all large-scale archives in computer centers around the world. Since 1987, LLNL has fielded the highly successful StorageTek 9310 silos. While we plan to continue using our 9310s, they are no longer being manufactured. We chose a new robotic technology, the StorageTek SL8500, during the move to our new computer center. This redundant, high speed robotics platform offers flexible expansion in both capacity and performance. It is also capable of housing all of the tape drive species planned at LLNL. LLNL procured seven of these robots and housed them in the new locations for each of our data archives. These robots will form the foundation for our archives for years to come.

Enhancing Software Infrastructure

In order to ensure continued viability of the archive for ASC Purple and other center users, the LLNL HPSS development team completed development, testing and packaging of HPSS Release 5.1. This release included an important system restructuring that eliminated the previous reliance upon Transarc's Encina transactional software, replacing that with a standard DB2 database for metadata operations. To further improve transactional performance, the HPSS Bitfile Server, Name Server and Storage Servers were merged into a single "Core Server". In addition, HPSS developers replaced the aging Sammi system administration infrastructure with an extensible and scalable Java-based solution.

These HPSS R5.1 upgrades were essential to ensure continued viability of HPSS in our center, and for improving the reliability and scalability of the archive. The changes yielded transactional performance rates of 3x-9x previously seen. Protocol efficiencies yielded improvements in per-file and aggregate transfer rates as well.

Archive Deployment

Once hardware technologies were chosen and HPSS software developed, they need to be deployed to end customers in an as non-disruptive manner as possible. Doing so required careful planning, many dozens of procurements, phased installation of hardware, extensive software testing, and conversion of the existing archives to HPSS R5.1. These efforts were accomplished over the two year period of this milestone and represent an incredible amount of work.

Hardware deployments of the chosen mover, disk, tape, robotic and SAN technologies were performed in concert with the moves of our data centers at LLNL. All installations were performed in a manner to minimize downtimes. Often the ASC user community would simply see enhanced performance, not realizing that equipment was being moved, configured, enhanced and installed behind the scenes.

The most involved part of the archive deployment effort for this milestone was the conversion of the LC HPSS archives to HPSS R5.1. Because of the decades of data investment involved, archive conversions are always risky propositions and can take up to a year to plan and carry out. At LLNL, the archive Operations Team works hand-in-hand with local HPSS developers on these efforts. After months of testing in local development environments, the OCF archive was successfully converted on July 4th, 2004 and the SCF was converted on September 20th 2004.

Demonstration

The milestone was written such that a delay in the delivery and deployment of the Phase 1 Purple machine would still allow the I/O infrastructure to be proven. Fortunately, pURPURA was delivered and integrated quickly allowing our milestone demonstration to use the very same infrastructure deployed for Purple itself.

On June 15th pURPURA was moved (swung) to the SCF environment and was hooked to the 10GigE infrastructure deployed as part of this milestone. DSG personnel installed three archival storage interfaces on the machine:

- FTP/PFTP version 1.0.10_pending – standard and parallel versions of FTP.
- NFT – a persistent archive interface.
- HTAR – a file aggregation interface

Configuration files were enhanced to include pURPURA. These files ensure that optimal networking configuration is used for transfers off of a new machine. pURPURA was built with four login nodes (pu241, pu242, pu243 and pu244). Each login node was installed with AIX v5.1 ML6, HPS 1. The HPSS system used was the production HPSS archive running on the "raven" complex. Each raven node was running IBM's AIX 5.1 operating system and HPSS R5.1 (tarfile 2 plus local mods).

On pURPURA, twelve Parallel FTP (PFTP) sessions were executed via a script hpss.test.6.run which launched other scripts executing the PFTP sessions (hpss.test.6.slave) on each login node.

Each of the twelve PFTP sessions moved a 30GB file on pURPURA to archival storage across the I/O infrastructure. In order to capture instantaneous data rates, the iostat command was run via script on the HPSS mover machines. The output from these commands was parsed via a C program called iosdif.c. The maximum write rate measured was measured at 14:13:35 with a rate of 2.55GB/s:

```
    From the file "050810.iorates.log.2":
```

```
14:13:35                 Megabytes Per Second
Disk Activity             Read          Write
System disks             0.106000    0.472000
Home file system         0.000000    0.050400
DB2 database             0.005600    0.083200
Disk cache              24.536400 2546.688000
```

Summary

LLNL platform, networking and archive teams worked closely together over two years to deploy a high performance I/O infrastructure for Phase 1 of the Purple machine. All milestone tasks were performed successfully and the milestone metric minimum of 1.5GB/s data transfers to the archive were exceeded by over 1GB/s.