



Storage Performance for Life Science Applications

An Isilon Systems® Whitepaper

August 2009

Prepared by:



Contents

Abstract	2
Performs Real Work the Way We Do It	2
Comparing Four Shared Storage Devices at RosettaInpharmatics	2
Performance Scales Linearly with Capacity.....	3
HPC /Storage for Science Explained	4
Cluster Computing with Shared Storage	4
High Performance Computing.....	5
High Throughput Computing	5
CPU-bound Analyses	6
IO-bound Analyses	7
Clustered Storage	7
Benchmarking Isilon for Science.....	8
Neuro-Imaging.....	8
Neuro-Imaging Data Growth Requirements	8
Neuro-Imaging Data Analysis using FSL.....	9
Neuro-Imaging FSL Performance Requirements	9
Next-Generation Sequencing	9
Next-Generation Sequencing Analysis using Illumina Pipeline.....	10
Illumina Pipeline is IO-bound	10
Benchmarking with bonnie++	10
IO Performance of Isilon under Load.....	11
IO Performance of Isilon with Scale.....	12
Conclusion	12

Abstract

Many branches of Life Science research involve the generation, accumulation, analysis, and distribution of “large” amounts of data. The analysis of these data is often compute and IO intensive and can require hours, days, and months to complete. The wall-clock analysis time can be reduced by spreading the computational load over a number of computers working simultaneously. However the rate of time reduction is often limited by the IO capacity of a shared storage system serving data to an increasing number of client computers.

In this paper, we share the experiences of industry and academia, including RosettaInpharmatics and UCLA conducting various cross-vendor storage benchmarking experiments and analysis when performing “real” Life Science analyses. In all these tests, Isilon storage was found to offer the greatest IO performance that scales linearly with storage capacity.

We’ll also explain the differences between achieving high-performance and high-throughput computing when executing many discrete processes that are CPU or IO-bound and how this relates to scaling clusters of computers or clusters of storage.

And finally, we perform our own benchmarking experiments to determine how well Isilon’s symmetric clustered storage performs as the shared file system for the data-intensive scientific research areas of Neuro-imaging and Next-Generation DNA sequencing, looking more closely at how Isilon’s IO performance scales with storage capacity using an open source IO benchmarking tool.

Performs Real Work the Way We Do It

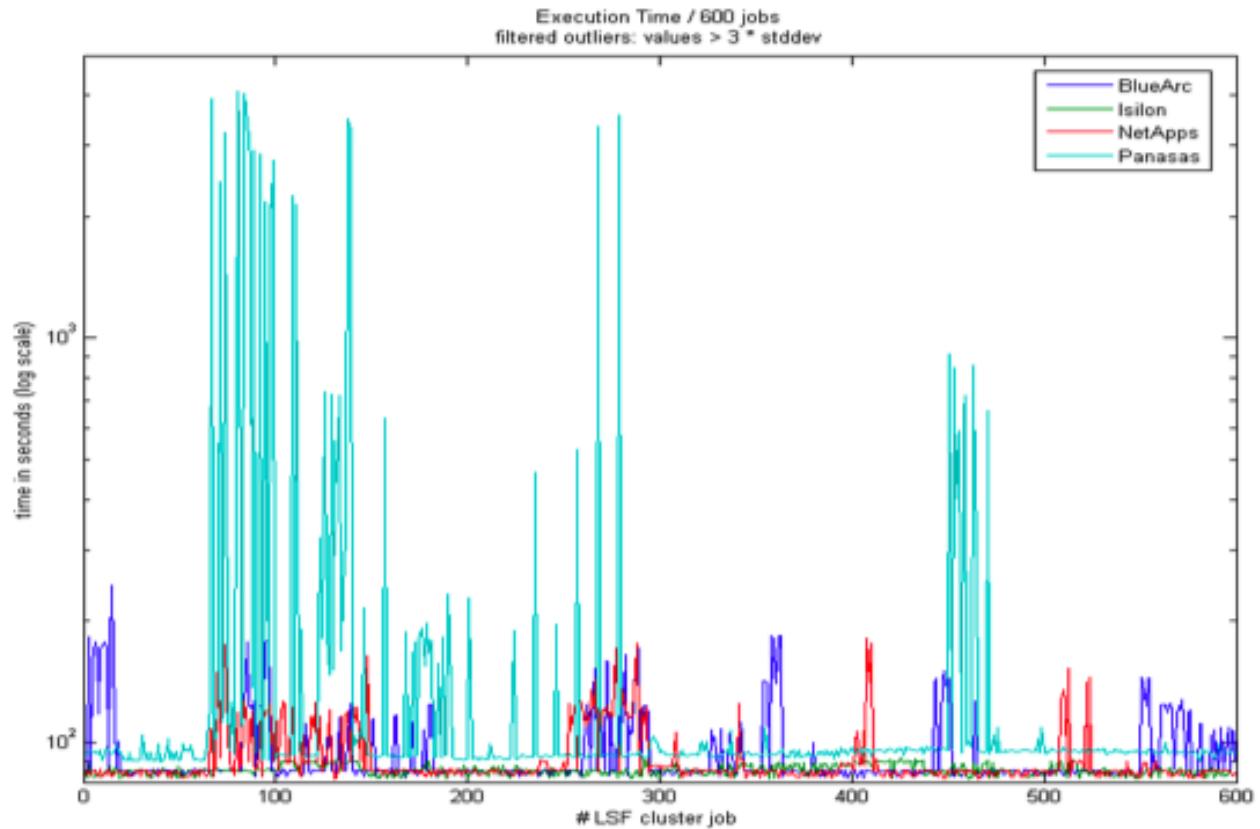
Comparing Four Shared Storage Devices at Rosetta Inpharmatics

At this year’s Bio-IT World Expo (April, 2009) in Boston, John Dey the UNIX Operations Manager, Rosetta Inpharmatics gave a presentation, “High Throughput Storage for Clusters”, where he compared the performance of four shared storage devices; BlueArc, Isilon, NetApp, and Panasas. In the testing, these systems served as the shared file system for a cluster of computers, performing as he said, “Real work, the way we do it”. In this experiment, the wall-clock execution time (log scale, 0 to 10,000 seconds) was reported for each of 600 informatics jobs, scheduled by a distributed resource management system (Platform LSF), operating on a cluster of computers, served by the shared storage device in question. While we do not know the exact nature of the jobs being executed (what application and whether they are CPU or IO bound), we can safely assume they are applications typically used by informaticists at a pharmaceutical company.

In plotting the results, Dey shows that while all of the storage devices perform comparably for some of the jobs (25-75, 300-450, etc), there is a broad distribution of wall-clock execution time (10s of seconds to 10,000s of seconds) across storage devices for more than half of the jobs (0-25, 75-300, 450-475, etc).

The Panasas storage device shows the greatest deviation from the mean performance and the BlueArc and NetApp storage devices show comparable performance.

The Isilon storage device shows the best performance, consistently across all 600 jobs (dark green line along the bottom of the chart).



Performance Scales Linearly with Capacity

The Laboratory of Neuro-Imaging (LONI) at UCLA is an institution leading the development of advanced computational algorithms and scientific approaches for the comprehensive and quantitative mapping of brain structure and function. Researchers at LONI study the development and function of the brain using a variety of imaging modalities: PET (Positron Emission Tomography), SPECT (Single Photon Emission Computed Tomography), fMRI (functional Magnetic Resonance Imaging), and DTI (Diffusion Tensor Imaging).

LONI permits its researchers, students and industry collaborators to gain access to a vast database of neuro-imaging data and run analysis workflows using “The LONI Pipeline Execution Environment” on a 1200 core HPC cluster. Discrete workflow jobs have an execution time ranging from 10 minutes to weeks accessing data from single scans that range in size from 20MB to several hundred gigabytes per sample.

The LONI Imaging-genetics modeling tool analyzes the 3D imaging data and genotypic information for several groups/populations of 100's of subjects. Each 3D neuro-imaging volume is comprised of voxel intensities defined on a 256^3 lattice. This tool fits a multi-linear regression model on 6,000 genes (regressors) at each voxel of the data, where the imaging data is the dependent variable. The result of this analysis is a collection of 0.5MB files storing the regression model for each voxel. The entire analysis stores terabytes of results, which imposes significant stress on their computational and NFS infrastructure.

LONI's framework for representation, analysis and visualization of multidimensional multi-resolution data requires intense computational and/or parallel processing power. The core of this computational complexity is modeling standard 3D volumes as multi-scale hierarchies of 3D volumes, which represent the data at different resolutions. This representation improves the access intensities at specific anatomical regions and resolution levels, however, it requires a significant data preprocessing overhead. The calculations for a single volume could range from 0.5-12 hrs, depending on the data volume and the chosen multi-resolution scaling.

To manage these intensive processing needs, in 2006, LONI transitioned from a sole reliance on SMP compute systems with attached SAN storage to utilizing a 600 core HPC cluster, however they found the SAN storage could not meet the IO requirements of their research. In an effort to find a suitable replacement for the SAN, LONI compared the performance of storage offerings from leading storage providers by executing their standard mix of analysis workflows on each vendor's products. LONI analysis found Isilon offered the greatest performance and the lowest cost. As a result, LONI deployed 13 node Isilon storage system (7 storage nodes with 6 accelerators) to support their 600 core HPC cluster.

Two years later, LONI doubled their compute cluster from 600 to 1200 cores to support an increased computational demand and scaled their Isilon storage system from 13 to 27 nodes to provide increased storage and IO capacity.

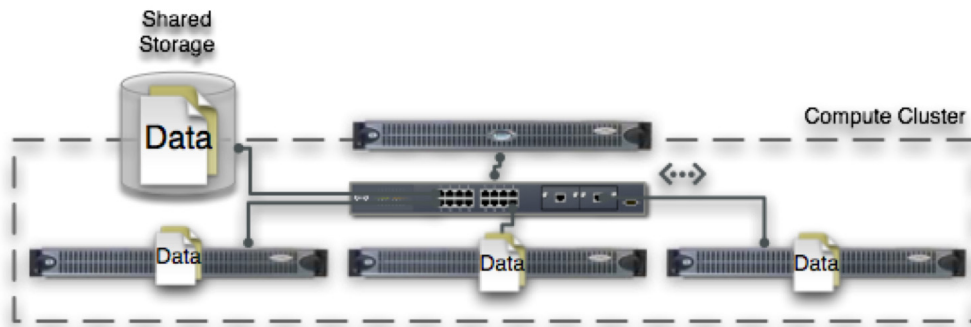
LONI reports observing a linear IO performance increase as they added additional Isilon storage nodes, without altering the overall storage architecture.

HPC /Storage for Science Explained

Cluster Computing with Shared Storage

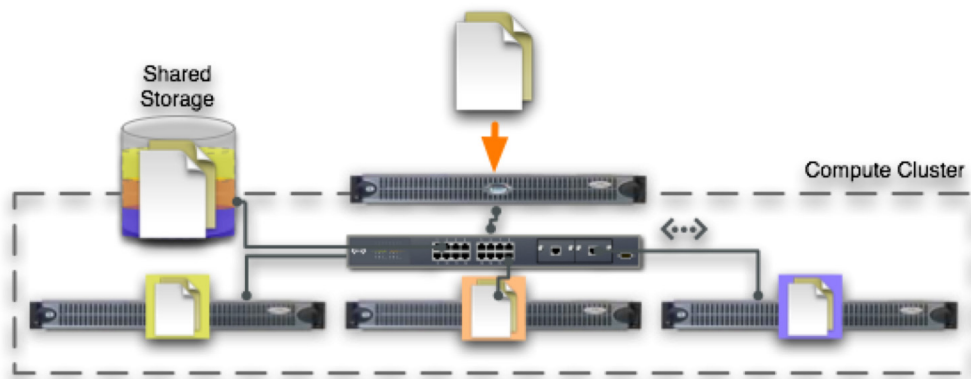
In scientific research settings, clusters of computers have become a common tool for research much like a microscope or centrifuge. A compute cluster is composed of a number of similarly provisioned computers attached to a shared storage system. Each computer in the cluster has common access to input data and a common repository for generated results. In addition, a distributed resource management system is available that enables users to submit their analyses for execution upon one or more anonymous computers to achieve either high performance (i.e. a fixed amount of work in the shortest time) or high throughput (i.e. the most work accomplished using a compute resource of a fixed size). As demand for compute capacity increases, additional computers can be readily added in small or large increments.

A compute cluster typically executes many analysis tasks simultaneously, either high performance computing or high throughput computing, where the performance of a single analysis task is typically limited by either the speed of the computer's processors (CPU-bound) or by the rate of access to the shared file system (IO-bound).



High Performance Computing

High performance computing aims to accomplish a given amount of computational work in less time, often at the expense of efficiency. Many scientific analyses are “embarrassingly parallel” in nature, where the wall-clock execution time of a single, long-running analysis can be reduced by partitioning the input data into many smaller pieces, permitting many discrete computers to operate simultaneously on their assigned portion. Ideally the performance through this parallelism scales linearly with the number of computers tasked to the analysis. However in practice, the performance is often limited by the IO capacity of the shared storage system. Image analysis associated with Next-Generation sequencing is often distributed over a cluster of computers for shorter wall-clock analysis time; however the parallelism can be constrained by file access to the image data.

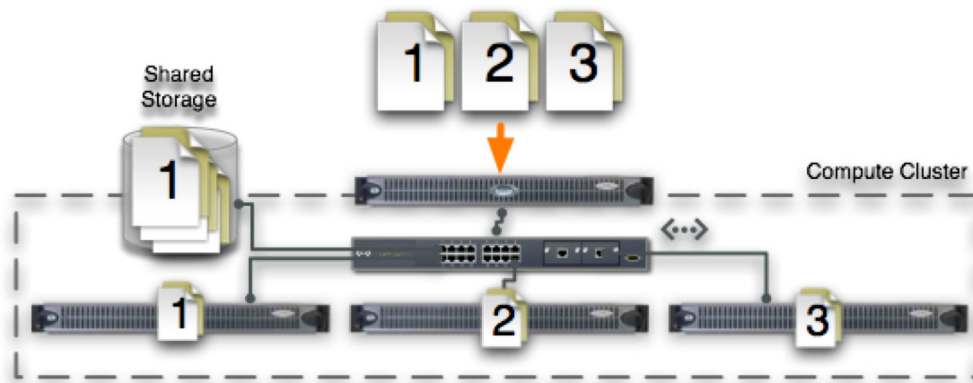


High Throughput Computing

Some scientific research involves a large number of short-running analyses, or analyses that are not easily parallelized. In this case clusters are used for high throughput computing. In contrast to high perform-

ance computing, high throughput computing aims to complete more computational work, more efficiently, given a fixed set of compute resources, sometimes at the expense of performance. That is to say, a single analysis might not be any faster, but the given compute resource is maintained at full activity to complete the greatest total number of analyses. An individual analysis is not partitioned into many smaller pieces, instead many independent analyses are performed at the same time with greater efficiency. Clusters of computers that provide a common, shared resource to many users executing many independent analysis tasks aim to achieve the greatest possible efficiency with high-throughput computing.

As with high performance computing, the efficiency of high throughput computing is often limited by the IO capacity of the shared storage system.



CPU-bound Analyses

Some analyses are CPU-bound. These analyses generally have input files that are “small” in size and number, where the direct computation is the slowest step within the overall analysis. The wall-clock execution time of the analysis can be reduced by using faster processors. Additionally, if the structure of the analysis algorithm permits parallelization, these analyses can be made faster by distributing the analysis on a cluster of computers. Phylogenetic analysis are commonly CPU-bound algorithms, where the input and resulting output data files are few in number and small in size, but require many CPU hours, days, or months of computation to complete.

Since performance is rate-limited by processing power, increasing the IO capacity of the shared storage system will have little to no effect on overall performance.



IO-bound Analyses

Alternatively, some analyses are IO-bound. These analyses generally have input files that are “large” in size or number, where accessing data on disk and bringing it up into physical memory is the slowest step. In this case, providing more or faster processors will have little to no effect on overall performance. The use of NCBI BLAST for sequence homology searching against large sequence databases is a familiar IO-bound use case, particularly when the database does not fit entirely within physical memory and must be iteratively swapped out to disk.

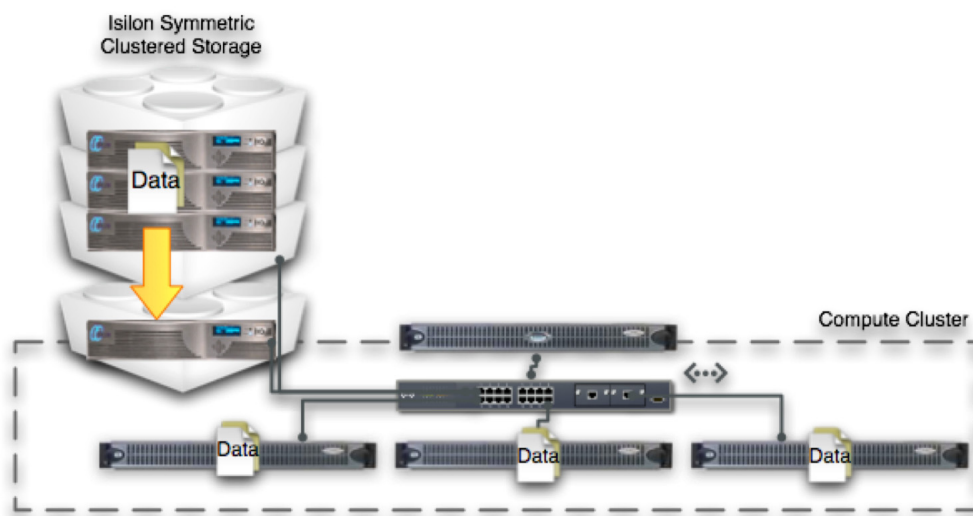
The only way to increase overall performance is to provide greater IO capacity to the shared storage system.



Clustered Storage

In a manner comparable to clustered computing, some storage systems have a clustered architecture. Isilon’s symmetric clustered storage system is composed of anonymous and identical components that each provides a unit of storage capacity and a unit of IO capacity, thereby permitting the simple incre-

mental scaling of both storage and IO without re-architecting the overall storage system. As demand for storage and IO capacity increases, additional units can be readily added in small or large increments.



Benchmarking Isilon for Science

In the following set of experiments, we aim to determine how well Isilon's symmetric clustered storage performs as the storage system for data-intensive scientific research. We could have used an IO performance benchmarking tool alone to determine rates of reading and writing data to disk, but we find these IO-only benchmarks to be somewhat misleading in that they report IO characteristics and not "real work, real research the way we do it". A more direct method of measuring the storage performance is to choose some real-world data-intensive experiments and measure the time of analysis. Presently, neuro-imaging and Next-Generation DNA sequencing are common experiments that involve compute and IO intensive analysis of massive amounts of data. Following are two examples of typical life science analysis applications and we will explore their characteristics in terms of CPU and IO, as well as a performance analysis of standard benchmarking tools.

Neuro-Imaging

Neuro-Imaging Data Growth Requirements

As in the earlier example at the Laboratory of Neuro Imaging (LONI) at UCLA, fMRI (Functional Magnetic Resonance Imaging), is a common neuro-imaging experiment to determine activated regions of the brain in response to a stimulus. This "brain mapping" is achieved by observing increased blood flow to the activated areas of the brain using an fMRI scanner. The scanning of a single human test subject might occur over a 60 to 90 minute period, with hundreds of discrete scans performed every few seconds, generating as much as 1GB of data per subject. A single instrument operating at only 50% capacity can produce many terabytes (1,000s of GBs) of data per year. The neuro-imaging centers interviewed for this paper utilize up to ten instruments, supporting dozens of scientists, each allocated a baseline of 2TB of disk

space for their ongoing experiments. While this rapid scaling is a significant challenge for many labs, data growth of 10 to 20 TB per year is not unusual in these environments.

Neuro-Imaging Data Analysis using FSL

FSL is a commonly used open source library of analysis tools for fMRI (and other) imaging data produced by members of the Analysis Group, at FMRIB, at the University of Oxford, in Oxford, UK. Some of the tools provide real-time analysis through a point-and-click interface, and others are used for batch processing of image data. Several of the more compute-intensive batch processing tools will take advantage of a compute cluster operating a distributed resource management system (e.g. LSF or SGE) for high performance analysis.

Neuro-Imaging FSL Performance Requirements

The FSL software comes with a set of example data called FEEDS that permits the user to perform an example usage of each of the tools that come with the FSL suite, and a validating script called RUN that will serially execute each of the tools against an appropriate set of data. Executing the RUN script against the FEEDS data set under different run-time conditions is a simple way of determining the performance of neuro-imaging analysis under various conditions.

Executing RUN on a dual-quad core system against the entire FEEDS data set on local disk took about 30 minutes. During this run we observed 100% CPU utilization on one of the cores, suggesting that neuro-imaging analysis is CPU-bound. Repeating the benchmark test with 4 simultaneous instances of RUN also completed in about 30 minutes, with 4 cores showing 100% CPU utilization, each accessing data from the same local disk. Repeating this experiment with the FEEDS data located on a local disk RAID or a shared Isilon storage system also executed in 30 minutes with 100% CPU utilization, confirming the execution of the FSL suite is CPU-bound.

We phoned the FSL development team and asked if there were conditions under which fMRI analyses could be performed that might be IO-bound. They indicated that some uses of FEAT (an FSL tool for fMRI image data pre-processing) with irregularly large files could produce an IO-bound condition. We repeated the benchmark tests, limiting our execution to only FEAT and found no difference in execution time with data located on different types of storage.

Next-Generation Sequencing

DNA sequencing has undergone a revolution in recent years. Driven by novel sequencing chemistries, micro-fluidic systems, and reaction detection methods, "Next-Generation" sequencing instruments from 454, Illumina, ABI, and Helicos offer 100 to 1000-fold increased throughput combined with an additional 100 to 1000-fold decreased cost per nucleotide when compared with conventional Sanger sequencing. The result for such labs is a dramatic increase in storage requirements from gigabytes to petabytes (1 million GB) in a couple years. Each Next-Generation sequencing platform is unique in terms of the nature

and volume of the data it generates. Typically, anywhere from 600GB (gigabytes) to 6TB (terabytes) of primary image data is written over a period of one to three days.

Next-Generation Sequencing Analysis using Illumina Pipeline

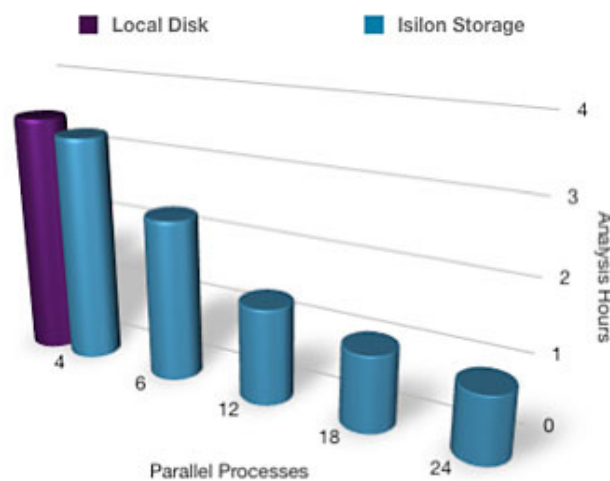
Illumina is one of four major vendors of Next-Generation sequencing instruments. The Illumina data analysis pipeline operates in 3 distinct phases (image analysis, base-calling, and alignment) to generate resulting DNA sequence from several hundred thousand image files of primary data. The alignment phase is CPU-bound with an execution time that is highly dependent upon the sample being sequenced. However the image analysis and base-calling phases tend to be IO-bound with execution times that are independent of the sample being sequenced. For this paper we benchmarked the execution time of the image analysis and base calling phases (one of eight tiles) using a 170GB validation data set.

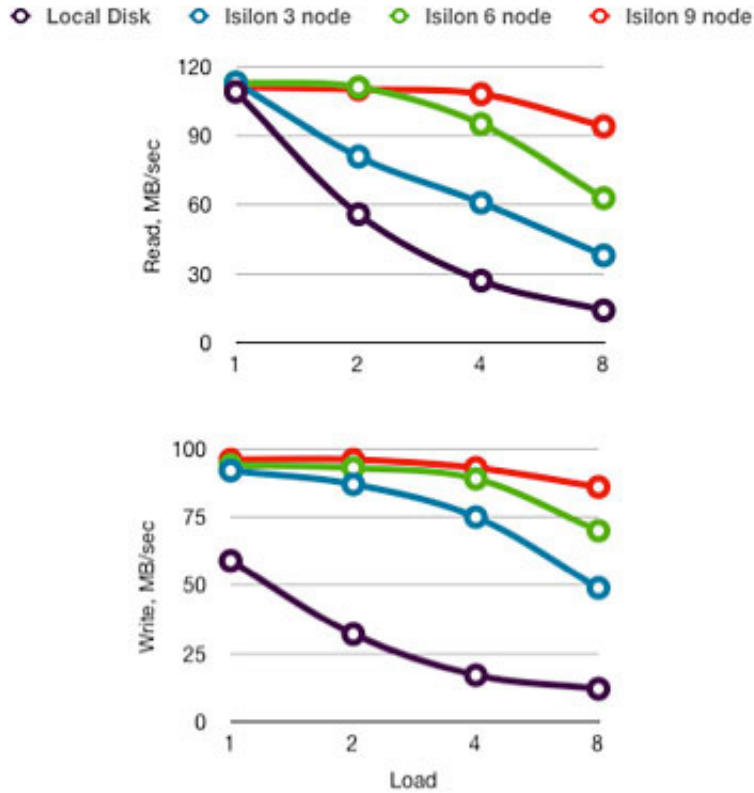
Illumina Pipeline is IO-bound

In this benchmarking experiment, we performed an Illumina analysis using 4 parallel processes on one system with data located on local disk as compared to execution on an increasing number of parallel processing spanning 1 to 6 cluster nodes with data located on a shared Isilon storage system. Performance was marginally faster on 4 processes using the Isilon storage as compared to local disk. However the analysis is highly parallelizable, with maximum performance (under an hour) when distributed over 18-24 processes increasing performance four-fold. Beyond 24 parallel processes little performance benefit was observed. Under the extreme condition of parallelizing over 140 parallel processes operating on 35 cluster nodes, performance remained nearly the same without significantly taxing the storage system, indicating that performance might be limited by the overhead of scheduling the many processes.

Benchmarking with bonnie++

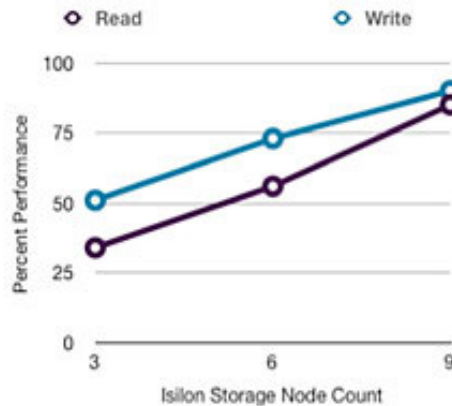
In these tests, we used bonnie++, an open-source benchmarking suite that performs a number of simple tests measuring the performance of either local disk or a shared file system. Among other things, it provides a measure of read and write performance.





IO Performance of Isilon under Load

Both local disk and the Isilon storage system show comparable read performance under light load (110 MB/sec). However the Isilon storage system shows nearly 60% greater write performance (95 MB/sec vs. 59 MB/sec). As the IO load increases from a factor of 1 to 8, the read and write performance of the local disk degrades rapidly. Whereas the read and write performance of the Isilon storage system is less affected by increased load and can be maintained through the incremental addition of storage nodes from 3 to 9.



IO Performance of Isilon with Scale

The plots above show the increased IO performance of the Isilon storage system when using an increasing number of storage nodes. Plotting the data above as percent performance as a function of node count shows a near linear performance IO capacity with increasing node count.

Conclusion

In this paper, we aimed to determine how well Isilon's symmetric clustered storage performs as the storage system for data-intensive scientific research. In our tests we found that our use of neuro-imaging analysis software to be CPU-bound and therefore the performance of these analyses were not affected by the IO performance capabilities of the storage system.

However, we also found that real-world neuro-imaging research performed simultaneously by many users, such as that being performed by LONI [is](#) reported to be dependent on the IO performance of the storage system.

We also found that the performance of next-generation sequencing analysis could be executed four-fold faster when distributed over a cluster of computers when using the Isilon storage system. We observed that the Isilon storage system could maintain a more consistent level of IO performance under increasing load, and that its IO performance increases linearly with node count.