DataDirect NETWORKS

A Scalable Sequencing Architecture

Selecting a sub-optimal file system can negatively impact the drug discovery pipeline by as much as 43 percent

Executive Summary

It is very difficult to identify the bottlenecks in a poorly architected sequencing system. Despite endemic system issues, the architecture can appear to be functioning at peak performance. This paper focuses on the storage component and the impact that a file system alone can have on the drug discovery pipeline. Data presented show that in an apples-to-apples comparison, between the legacy NFS protocol and DataDirect Networks GridScaler file system, NFS can negatively impact a discovery pipeline by as much as 43 percent over GridScaler.

By Mike May, PhD

Produced by *Bio-IT World* and the Cambridge Healthtech Media Custom Publishing Group

Over the past 10 years, the efficiency and accuracy of sequencing technology has significantly accelerated biological research and discovery. Today, the energy and resources focused on molecular and cellular biology, bioinformatics, proteomics and the emerging field of pharmacogenomics continue to accelerate at a break-neck pace. It is projected that in the second decade of the 21st century, we could gain a full understanding of the workings of our DNA. This newfound knowledge would empower us to improve our collective quality of life through a better understanding of how a specific genetic variation impacts a drug's efficacy or toxicity or possibly eradicate hundreds of genetically based disorders.

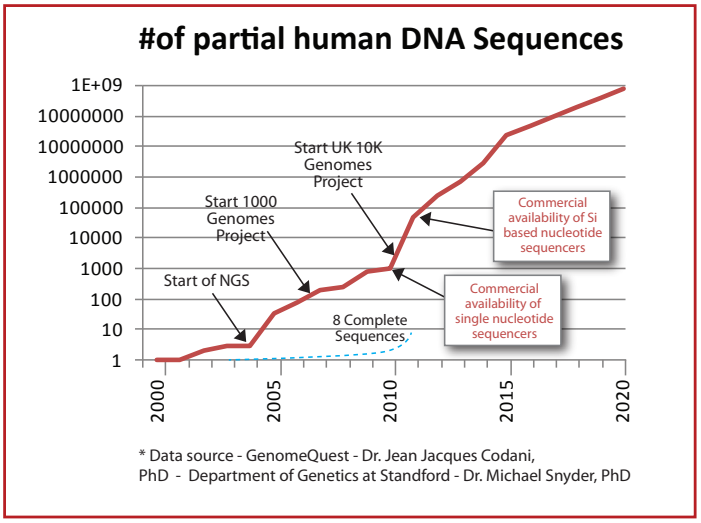

Figure 1. The number of human sequences is expected to increase exponentially.

In July of 2010, GenomeQuest's company blog discussed the implications of exponential growth of global whole-genome sequencing, including some comments about the total number of human genomes sequenced. A single human genome—23 pairs of chromosomes composed of a total of about six billion base pairs—requires about 60 gigabytes (GB) of storage in FASTA format. The GenomeQuest posting forecasts that the number of human genomes sequenced will explode from the single human genome sequenced in the early 2000s to 1,000 genomes in 2010, 50,000 genomes in 2011, 250,000 genomes in 2012 and 25 million by 2015 (See Figure 1).

The key technology for unraveling the secrets of DNA are the myriad of commercial sequencers available from various companies, including Illumina, Life Technologies, Pacific Biosciences, 454 Life Sciences and others. These sequencers interface to a computer network, which correlates and concatenates the billions of contigs—overlapping segments of DNA—that have been streamed to or stored on a network-attached storage system (NAS).

Accommodating the output rate of the sequencers requires a precisely designed and balanced system. The peak rate of data (base pairs) produced by these sequencers is already approaching 25 billion in 24 hours, or roughly an equivalent volume equal to 5.8 hours per human genome. In FASTA format, 25 billion base pairs create 500 GB of data. So one sequencer can generate 58 megabytes (MB) per second, and that number will probably double in the next 18 months.

To analyze a reference genome run, researchers require a computer platform with three key features: a fast and well-engineered network, a scalable and adaptive file system that supports a single name-space and a fast multi-core compute cluster, which runs under the API - Message- Passing Interface (MPI). For de novo align/assembly, the MPI/MPICH cluster is replaced or augmented with fat SMP nodes, which are commodity compute nodes with more than one terabyte (TB) of DRAM. Overall, a sequencing system consists of our key hardware elements:

1. Sequencer;
2. Computer network: comprised of 1 or 10 gigabit Ethernet (Gig/E), InfiniBand or other proprietary fabrics;
3. Compute nodes or cluster: a collection of heterogeneous compute nodes connected together via a network running under MPI/MPICH with or without fat SMP nodes and
4. A storage network: a scalable file system that supports a multi-petabyte (PB) single name-space (see Figure 2).

These four components make up the hierarchy of the gene-sequencing architecture. Each system depends on the other and must have the ability to adapt and scale to meet current and future sequencing needs. If one component creates a bottleneck, then the performance of the entire sequencing system suffers. This article focuses on the network-attached storage system and its relationship to the sequencer and the computer platform.

Faster Filing

To collect these data and make them available for others to analyze and process, bioinformatics experts rely on a network-attached file system. This file system allows a researcher to sit down at a terminal (client) and access files in some other location. From a researcher's point-of-view, the remote files and compute resource appear local or native. In 1984, Sun Microsystems developed a protocol for this process with the now seemingly redundant name of Network File System (NFS). This is a UNIX open-source (free) protocol that any lab can implement on top of their file system, and the NFS protocol is available in several versions, both open-source and commercial. Most, if not all, compute and storage protocols/filing systems deployed within sequencing centers worldwide are based on the ever popular and free Linux operating system and the x86 processor (not free). Smaller, cost-conscious sequencing centers, such as academic institutions, rely upon a Gig/E network, a heterogeneous MPI Linux cluster and the NFS protocol. In some cases, this works fine, but problems—mostly related to the protocol/file system—can and do emerge. The next tier, such as larger research universities, uses more and faster sequencers interfaced to multi-core x86 commodity servers, the network is 10 Gig/E or InfiniBand and a centralized NAS file system maintained by professional IT managers, augmented by computer science graduate students. In time, though, sequencing equipment is updated, and it soon spews out 58 MB per second per sequencer. As the alignment and assembly processes scale on a cluster system, the name-space scalability of traditional NFS is unable to scale past a few 100 processes per file system volume. At this point, the limitations of a legacy network and the distributed NFS name-space become a major bottleneck in the bioinformatics data center, spurring a need for change.

Most state-of-the-art disk storage protocols and file systems are parallel, delivering very high throughput. This performance is measured in terms of input/output (I/O) operations per second (IOPS) verses Read/Write throughput in GB per second. These sophisticated and full integrated parallel file systems allow data and programs to be simultaneously accessed over multiple computers by multiple users, all within a single namespace. NFS was designed for a point-to-point storage access. Hence, adding more file servers only adds capacity with new a name-space, not performance, whereas an integrated parallel file system's performance scales with added capacity.

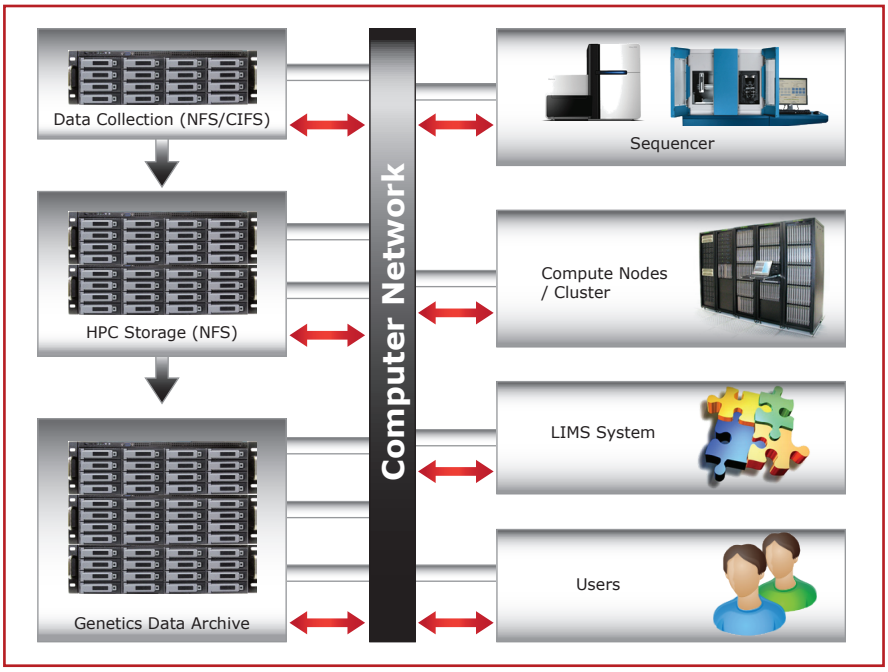

Figure 2. Example of a networked-attached sequencing system.

For front-end tasks, like in-place high-speed sequence-analysis, parallel file systems provide higher I/O performance by "striping" blocks of data from individual files over multiple disks, and reading and writing these blocks in parallel—hence the term "parallel file system." These systems are further optimized by their ability to seamlessly interoperate with native clients, also in parallel. This ability to support multiple clients is superior to traditional legacy protocols by supporting parallelize and aggregate I/O to the storage pool. This ability to read and write in parallel is ideal for bioinformatics and sequencing, where multiple clients or sequencers are reading and writing to the same file system. Some key support features provided by these systems include high availability, support for heterogeneous clusters, disaster recovery, security, the ability to dynamically move data (in the background) to a less vital storage resource and more. This class of features is important for the system administrator for "tuning" and future scaling of the storage system.

The DataDirect Networks (DDN) GridScaler parallel file storage system incorporates file services, data management and DDN Silicon Storage Architecture (S2A) and Storage Fusion Architecture (SFA) storage in a unified, scalable platform from 100 TBs to 10s of PBs all within a single name-space. DDN S2A and SFA storage platforms provide world-leading bandwidth ranging from 2–10 GB per second per storage platform, and serve as scalable building blocks to build out GridScaler systems. These storage solutions are ideal for bioinformatics and proteomics for both front-end and back-end processing. On the front-end, they are unsurpassed in dealing with a large number of files from multiple sources. They are also ideally suited for an in-place sequence process, where contigs must be randomly stored and recalled with minimal latency. On the back-end of the pipeline, large post-processed files are easily archived onto slower, higher density disks for later analysis. Rules or policies can be created that dictate that if a given dataset is not accessed in 30–45 days, it is moved to a disk array that literally spins hard drives down to conserve power with no changes to the file format and within the same single scalable file system - which can be accessed within seconds. Real-time performance analysis can enable system optimization; for this, DDN provides GridScaler clients (client software) with the ability to analyze the I/O used in an application. Monitoring metrics include a wide range of real-time data, including I/O request sizes, I/O completion latencies, open and close calls and more. With these numbers, the system administrator can assess an application's performance and then improve it. Such capabilities do not come with an NFS system.

In short, the DDN GridScaler platform with either SFA or S2A systems provide significantly better overall throughput and flexibility than legacy NFS-based systems. In addition, GridScaler can grow to multiple PBs in a single name-space and is equipped with a rich feature set to facilitate optimal tuning for a given process flow. Also, from the system administrators and users perspective, a DDN parallel storage system will have the same look and feel as the legacy file system (file view and hierarchy); it will just allow applications to use the storage resources more efficiently, effectively and in parallel, resulting in higher throughput, lower latency and more productive users.

Technology put to the test

To put data behind the above claims that GridScaler alone can accelerated an in-place high-speed sequence analysis job, DDN—with the help and cooperation of a key sequence OEM that can only be called "the sequencing company"—compared the performance of a legacy NFS storage platform against the S2A9900 storage system running the GridScaler file system. The platform was deployed at the sequencing company facility and interfaced to its workflow, which looks primarily for variations in gene sequences, often through re-sequencing. For example, it can search for single nucleotide polymorphisms (SNPs), insertions and deletions, and count genes to quantify copy numbers. As with the sequencing company name, the application software used is denoted as "sequencing software."

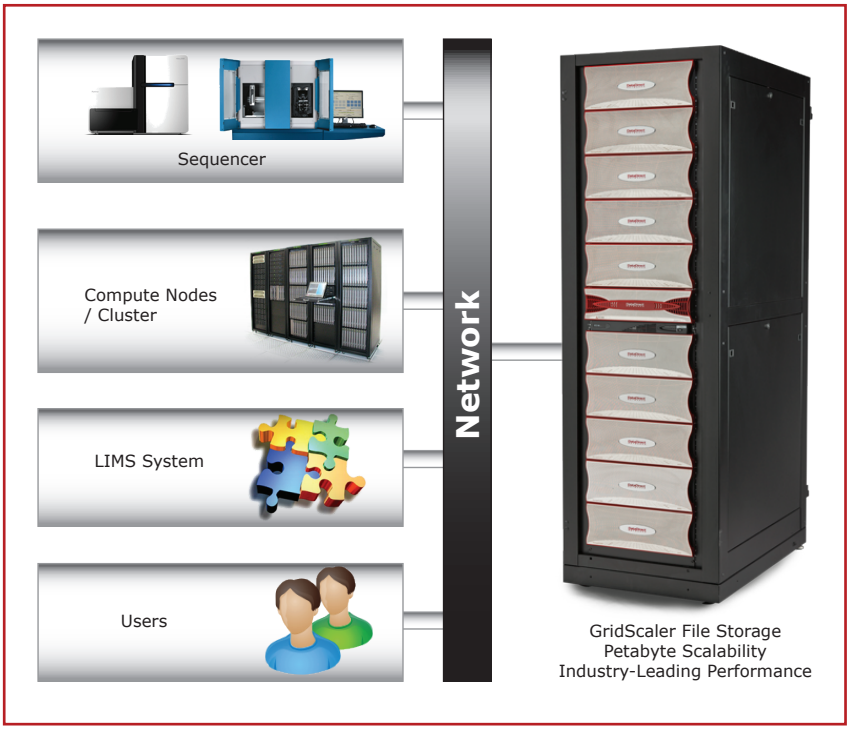

Figure 3. Example S2A9990 – GridScaler set-up.

For this test or benchmarking, DDN developed a proof-of-concept architecture. This consisted of six clients connected to a data network by way of a 1 Gig/E network fabric. The S2A9900 — GridScaler Gateway servers connected to the same data network with two 10 Gig/E lines: one for the GridScaler protocol and the other for the NFS protocol. Sequence file data were stored on the DDN S2A9900 storage system and a DDN EF3010 server stored the metadata from the sequencing runs (see Figure 3).

Reviewing the Results

The sequencing software used in the benchmarks includes several stages of computation. These steps start with the raw image or TIFF data from the sequencing company's gene-sequencing device. The first stage uses a dataset composed of more than one million files, which range in size from kilobytes (KB) to MB. These files get randomly read, reduced via an algorithm, and then are written into about 512 larger files—each of them being 1 GB with imbedded indexes for the next stage of the process. Although this sequence-data analysis involves several more stages, the first stage is the most I/O intensive. So the data-collect and the benchmarking results presented will examine only that stage.

NFS vs. GridScaler

	NFS/commodity Hardware	GridScaler/S2A -SFA
Capacity per Volume	Underlying File System Dependent, 16 -256TB	10PB ob object -based storage in a single volume, single name-space
Drive Varieties	SATA/SAS & SSD (generally not mixed)	SATA/SAS & SSD (in a tiered storage pool)
RAID Types	Underlying File System Dependent: RAID6 Overhead As Much as 40%	RAID 1,5 & 6 RAID6 Overhead Generally 22%
SATAssure (real time data protection engine)	No	Yes
Partial Rebuilds	No	Yes
Max Bandwidth	HW depend <2GB/sec.	2 to 12GB/second
IOPS	FS depend <40K (random)	100Ks of SPEC -SFS IOPS
Drives pre 4Uenclosure	Typically 15 to 18	60
Interconnect to host	Gb/E or 10Gb/E	Active/Active 10Gb/E or 40Gb InfiniBand
Read/Write at same speed?	No	Yes
D-MAID Power Savings support	No	Yes
Storage Pools	No	Yes Policy -Based Migration
HSM	No	Yes File System Integration to Support Tape Environments

NFS client vs. GridScaler Client

	NFS	GridScaler
Max. Network Transfer Size	32KB	4MB
Single Client File Access	Point:Point – One NFS Server per File Request	Federated – One Client Can See 1 -200+ Gateways all serving a single file
InfiniBand Support	No. 10GbE Performance < 800MB/s Latencies can be high	Yes. Full RDMA Support. 40Gb IB Performance > 3GB/s Latencies are very low
Native Client Support	Ubiquitous	Enterprise Linux Kernels Windows 2008
Client Scalability	<128 Concurrent Connections per Filer	4000+ Concurrent Client Connections per Volume, Directory or File

Figure 4. Features of NFS clients vs. GridScaler Native client.

The benchmarking tests measure the time (wall clock) required to complete the various stages of analysis, first exporting legacy NFS (via GridScaler) on the S2A9900 and then running GridScaler. Not surprisingly, the GridScaler approach ran the noted stages of sequencing analysis in 15 minutes and 5 seconds while NFS required more than 21 minutes and 30 seconds (same data sets and computations on the same nodes, with no tuning). In other words, removing the hardware from the equation, GridScaler alone ran 29.8 percent faster than NFS or put another way, NFS can negatively impact a drug discovery pipeline by as much as 43 percent over GridScaler.

This 29.8 percent gain in performance with GridScaler comes from many of its stated advantages over NFS (see Figure 4). However, the largest part of the difference comes from packet or payload size for input/output (Read/Write) operations. With NFS, packets are constrained to a maximum of just 32 KB. With GridScaler, I/O packet size can scale to 4 MB, or up to 125 times larger than NFS's maximum packet size. Moreover, I/O requests with an NFS approach are not always uniform, and that leads to a failure to use all of the available bandwidth between the client and storage server. By comparison, GridScaler clients perform optimal I/O exchanges with the server by using a payload that is a "best match" to the size of the genetic data, which leads to equally optimal I/O with the storage system.

Beyond the use of larger dynamic packets, GridScaler clients provide additional advantages in I/O speed. For example, an NFS client interacts with a single server, but a GridScaler client can simultaneously perform I/O with all of the servers in a system. So GridScaler's combination of large packet size and simultaneously communicating with multiple servers takes full advantage of the bandwidth available between the clients and servers. If a system includes a large numbers of clients, then the overall performance is even more dependent on the bandwidth of the fabric-interface—between the compute and storage servers.

Less traffic across the network also speeds up GridScaler. An NFS protocol, for instance, requires an extra layer of communication between the clients and the file system. With GridScaler, clients communicate directly with a clustered file system, which leads to less message overhead resulting in faster I/O for the clients.

Tomorrow's Tasks

Over the past five years, enhancements in throughput from sequencing technology have improved more than 100-fold. There is an ongoing and endless stream of new processes to improve accuracy and the rate of sequencing, including improved PCR techniques, Pyrosequencing of single nucleotides, sequential versus convergent ligation of multiple peptide fragments, DNA nano-ball, semiconductor micro machines and many more. These new and emerging processes continually push the speed at which proteins and DNA and RNA sequences are classified and enumerated, and the processes indirectly impact the compute and storage requirements needed to keep up with the flood of data. This trend will surely grow even steeper with new technologies and players, including Helicos BioScience, Complete Genomics, Oxford Nanopore Technologies and now "The Chip is the Machine" from Ion Torrent (soon to be acquired by Life Technologies). To maintain pace, IT teams need computational and storage capabilities that can grow and scale with the advances upstream in the data-generation pipeline.

It has been shown that in-place sequencing times can be reduced by as much as 29.8 percent with the Grid- Scaler parallel file system. Applying this savings to a pharma research pipeline, doing sequencing for pharmacogenomics correlation studies or DNA sequencing could translate to a reduction of the pipeline by weeks or possibly months.

To help sequencing centers ready for tomorrow, DDN has designed, tested and deployed their S2A and SFA storage controllers with native GridScaler in to numerous sequencing centers throughout the world. Today billions of base-pairs have been flawlessly stored and archived on DDN platforms. DDN customers can seamlessly add components at any level—clients, servers or storage nodes—and the GridScaler parallel file system will automatically make use of the added computing resources. Looking forward, we can expect a broader market take-up of these performance-enabling tools as traditional file systems evolve. Enhancements to the NFS protocol, such as NFS 4.1 (also known as parallel NFS (pNFS)) promise to provide many of the same parallel file semantics to the NFS client that are available today by the GridScaler platform.

As scientists gain the capabilities to explore genomes faster—looking at longer reads of DNA and pushing through more bases in a day—it will change not only our understanding of biology but also impact a wide range of applied sciences, from biotechnology to pharmaceuticals. Turning data into knowledge, however, demands the ability to file and analyze the raw sequence data. To make the most of the seeming tsunami of DNA data, IT teams need a fast and efficient file system—just what GridScaler provides.