



Network Appliance™ Filers in High-Throughput Life Science Research

Blackstone Computing and Network Appliance | April 2002 | TR-3157

TECHNICAL REPORT

Network Appliance, a pioneer and industry leader in data storage technology, helps organizations understand and meet complex technical challenges with advanced storage solutions and global data management strategies.

Table of Contents

1. Executive Summary	3
2. IT Challenges in the Life Sciences: Managing the Data Deluge	3
3. Network Appliance Filers: Meeting the Storage Needs of Life Sciences	4
4. Deployment Scenario: The Optimal BLAST Compute Farm	6
5. Conclusion	12
6. Appendix A: NetApp: Serving a Broad Range of Applications & Further Benefits Details.....	13
7. Appendix B: Configuring Filers: Key Considerations	14
8. Appendix C: Additional Information on NetApp Security and Data Protection Features.....	14
9. Footnotes	15

1. Executive Summary

The rapidly growing deluge of data in life sciences has produced strong demand for storage systems. Life science companies need systems that are easy to use, scale, and manage, as well as allow people and other systems shared access to data. For these reasons, network-attached storage systems, such as NetApp® filers, have become an increasingly popular solution. This paper is an overview of key computing and data storage challenges faced by life science organizations and how NetApp filers—as the central storage solution between enterprise and compute farm networks—help companies overcome these challenges. This paper also features a deployment scenario of a BLAST compute farm (including NetApp filers) that is optimized for near-linear scalability.

2. IT Challenges in the Life Sciences: Managing the Data Deluge

The announcements on the completion of first-draft sequencing efforts of the human genome in 2000 were just the most visible highlights of the significant scientific and technological milestones achieved recently within the life sciences. Many of these recent advances would have been impractical or extremely difficult to accomplish without simultaneous significant advancements in the automation of molecular biology laboratory techniques or without the harnessing of the massive amounts of compute power needed to analyze and manage the flood of resulting data.

The huge increase in scientific data made possible by the advent of robotics and high-speed laboratory automation technologies has enabled those who study biology to move from single-gene or single-experiment approaches toward faster and more comprehensive systematic research directions. These new large-scale research efforts, in turn, have allowed the previously rarified field of bioinformatics to explode onto center stage.

One of the single biggest infrastructure challenges facing life science companies today is the complexity of dealing with the increasing "tsunami" of biological data. Over the past few years, storage demands have grown by orders of magnitudes. No longer simply managing gigabytes of data, organizations are now struggling to handle terabyte volumes and to stay ahead of the increasing demand for capacity.

More specifically, organizations must deal with the accelerated growth in both public domain and internally generated data resources that must be kept available and online for analysis.

The international GenBank (see figure 2.1 [\(1\)](#)) sequence repository is just one of many Internet-accessible public data sets that organizations integrate into their internal discovery process. Another example is the growing list of finished and annotated genomes.

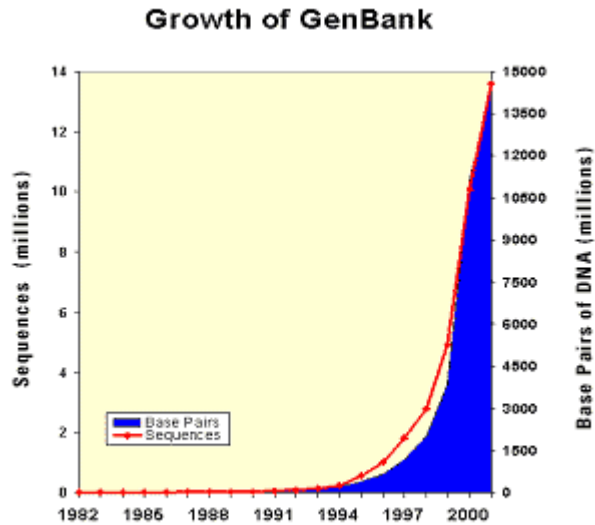


Figure 2.1

Compounding the issue, most data sources need to be parsed or manipulated into multiple formats to accommodate the wide range of analysis systems and algorithms in use today. Formats vary widely from structured data (e.g., relational databases) to unstructured data (e.g., binary and ASCII flat files).

Illustrating the situation, shortly after completing its own large-scale storage centralization project, the genomics department at one Boston area institution found itself having to expand its storage capacity from 3 to 14 terabytes in a single year. The additional infrastructure was necessary to support ongoing and emerging research systems and technologies that are continuously generating significant quantities of both structured and unstructured data.

3. Network Appliance Filers: Meeting the Storage Needs of Life Sciences

The sheer size and diversity of raw life science data that must be made available throughout the research organization makes network storage very compelling for storage needs.

Scientific data is useless if stored in isolation-invisible and unavailable to the systems and researchers that need it the most. Network Appliance's (NetApp) "filer" solutions address this challenge by making it easy for multiple systems and users—regardless of hardware or operating system—to simultaneously share access to the same files. In this way, essential research data can be stored once and made instantly available everywhere it is needed.

As an example of the utility that NetApp provides, consider a highly automated discovery research operation, such as large-scale sequencing or microarray-based gene expression studies. Each experiment produces electronic data that is collected regularly by the PC-based system responsible for controlling the instrument. With the PC connected to a network storage volume, the raw data can be pumped to a central repository where it is immediately available for open access. QA technicians, instrument operators, and scientists are then able to access the latest results immediately (regardless of their hardware platform or operating system).

Additionally, many systems and servers within the research-computing infrastructure may also need read/write access to the same data. These include archival and disaster recovery systems, as well as analysis pipelines running on large UNIX® clusters and scripts that manipulate the data for bulk loading into visualization applications or data warehouses.

Beyond these infrastructure needs, NetApp provides many key advantages for life science organizations. These benefits include:

- Rapid installation and configuration
- Very low administrative burden
- Centralized storage, which is easy to manage, deploy, and scale
- Ability to leverage existing IP networking infrastructure and staff skills
- Multiprotocol support that allows widest range of clients/client operating systems
- Multiple clients that can share read/write access to the same data

NetApp filers and caches provide low total cost of ownership (TCO)—a measurement of dollars spent on storage technology, as well as consideration of issues such as ease of administration, business interruption (minimization), and administrative costs and time* over the life cycle of these devices.

An independent study by INPUT demonstrates that NetApp offers significantly lower TCO than other vendors. Also, a case study at J.D. Edwards describes their ROI using caches. These and other NetApp comparative and value-based studies are available at http://www.netapp.com/tech_library/#reports.

* The industry analyst firm, Gartner Group, estimates 73 percent of total costs are related to administrative time. Minimizing administrative time provides for more time and budget for performance tuning and new projects.

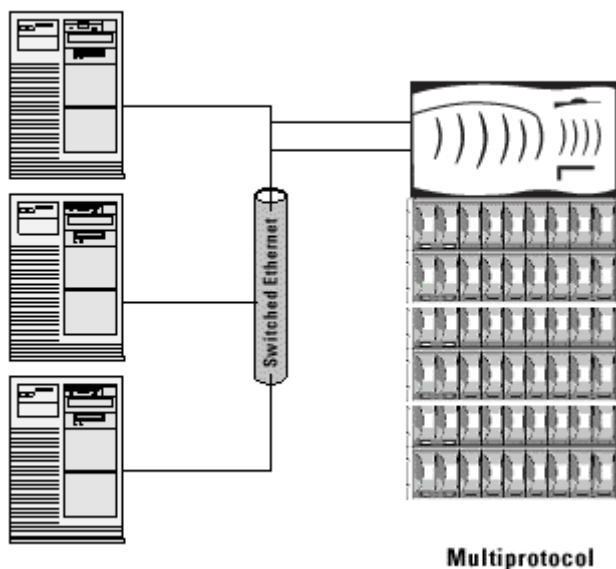


Figure 3.1 Sample network storage configuration.

4. Deployment Scenario: The Optimal BLAST Compute Farm

4.1. NetApp Filers in Compute Farm Environments—An Introduction

Compute farms are increasingly being brought into enterprise IT infrastructures as mission-critical production research computing platforms. Since a significant number of commonly used life science analytical tools are data and compute intensive, careful attention must be paid to the relationship between data locality, network topology, and file system performance within the compute farm environment.

It turns out that for modest-size compute farms (up to 20 CPUs), a simple configuration of the NetApp filers with NFSv3 [\(2\)](#) file services configured with consistent mount points and shared file systems for each node in the compute farm will provide for linear scaling of performance. Scaling beyond this, CPU resources are no longer the bottleneck and the rate of scientific analysis achievable may actually slow down with the addition of more CPU resources.

Blackstone Computing has worked with leading life sciences research organizations to develop software technology and best practices to enable scalable high-throughput computing. The remainder of this paper details a deployment scenario including the approach and technology required to deliver linear scaling of computing performance in proportion to the number of concurrent CPUs available, and discusses how the computing elements interoperate with the network topology and the NetApp filer to achieve this result.

4.2. First, What Is BLAST?

One of the most frequently asked questions in bioinformatics is "does this DNA or protein sequence have any significant similarity to that of any previously known genes or proteins?" This question is often asked for large numbers of query sequences, and often the target databases that must be searched are very large. To keep up with changes (i.e., new query sequences and constantly updated target databases), these searches typically have to be repeated at set intervals.

A researcher interested in a particular sequence will often wish to screen it against databases such as:

- Every unique sequence ever internally generated, handled, or acquired by the organization or lab
- All of the freely available sequences in GenBank and all other public repositories
- Every sequence that has ever been patented in any country around the world.

These databases can be very large and are typically stored as ASCII text files along with various associated binary index files used to speed performance for particular algorithms. These files can typically range from a few kilobytes to tens of gigabytes.

One of the most commonly used applications for searching sequence databases is a tool called BLAST (Basic Local Alignment Search Tool). The two primary implementations of BLAST are [WU-BLAST](#) and [NCBI-BLAST](#).

4.3. What Is a Compute Farm?

A compute farm refers to a set of individual servers ("compute nodes") interconnected via a scalable network infrastructure designed to suit the application mix, and configured with management (or "middleware") software, which creates a single image of the compute farm. The scientist sees only one large computing resource, not hundreds of discrete computers. The middleware software provides the scheduling functions. Blackstone Computing's PowerCloud™, an example of middleware software, also provides data balancing features, along with the ability to measure environments and allocate resources dynamically according to objectives.

Any applications—including the majority of popular informatics applications—that have the property of being "embarrassingly parallel" can be readily accelerated on a compute farm. Compute farms typically use standard TCP/IP networking and generally require no changes to the applications to provide value.

To accommodate future scaling and to simplify maintenance of the compute farm, Blackstone recommends a hierarchical networking topology consisting of the following major subsystems:

- A portal server—a dual-homed machine serves as the access point for scientists to a single resource image, which comprises, yet masks, the varied compute farm resources.
- An administration server—to manage configuration control of the entire compute farm and to host system management software, including PowerCloud.
- A compute subsystem—consisting of racked compute servers connected via a TCP/IP network.
- A storage subsystem—consisting of high-performance, highly available, scalable storage resources.

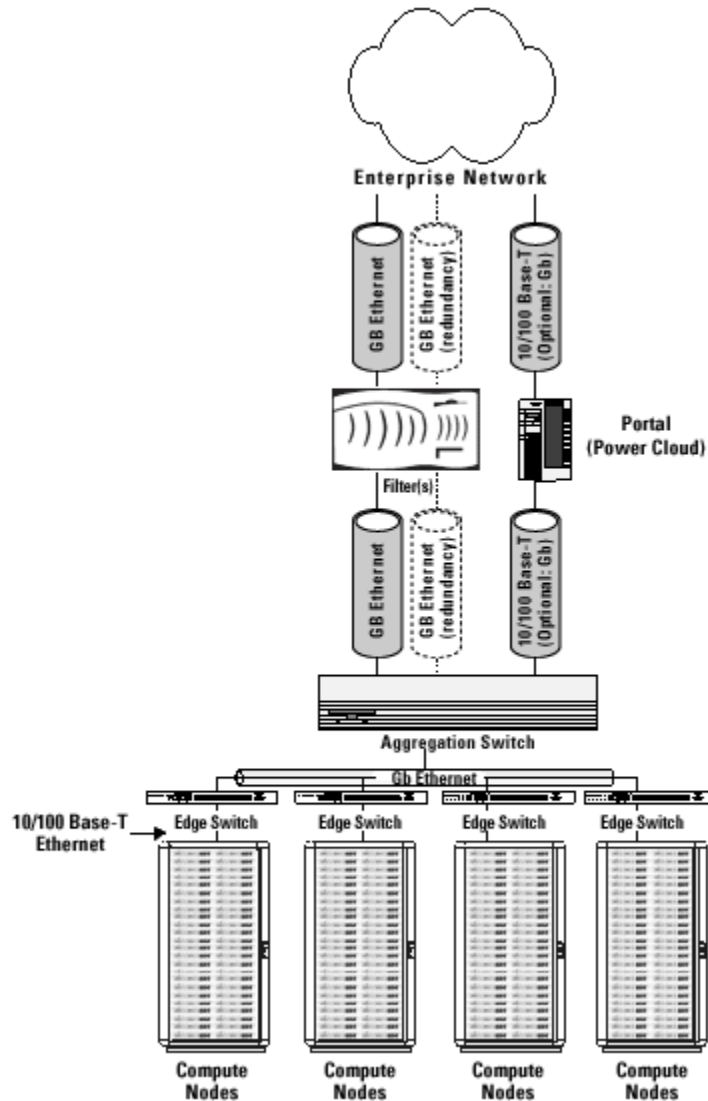


Figure 4.1. NetApp filer(s) serving both enterprise and compute farm needs.

"Edge" switches are used to provide 100Base-T connections between compute servers local to a rack and 1000Base-T uplinks to an "aggregation" switch. The aggregation switch provides for high bandwidth between the compute subsystem, the storage subsystem, and the administration and portal servers. It is also recommended that the switches be remotely manageable and SNMP aware for the following reasons:

- Network traffic monitoring and bottleneck identification
- Ability to map node MAC addresses to particular switch ports
- Rapid isolation and identification of network problems

The connection to the file server subsystem provides long-term online storage of end-user data and access to offline archival facilities. To maintain the anonymity of compute elements, no data other than temporary files should be stored locally on the compute nodes. Thus, a shared file server, such as a NetApp filer, is required.

This architecture's hierarchical network topology ensures scalability. The input/output (I/O) traffic generated by compute servers remains isolated to the servers contained in the physical rack, as is traffic associated with portal access and file server resources. This also avoids any physical cabling limitations associated with large scaling of compute farm resources.

The filer has two primary roles in the compute farm:

1. Store critical configuration, status, and log files on a common network file system. This allows the compute farm to gracefully recover from any individual server failure without losing jobs, results, or data.
2. Store the very large target data sets that BLAST queries require. Searches can be run directly against the network file system or the filer could be used as an intermediate staging point for data sets that are locally cached as needed to local disks contained within the individual server elements of the compute farm. The latter approach is highlighted in the remaining subsections.

A major benefit of the Network Appliance filer architecture in life sciences is the **ability to connect the NetApp filer to both the enterprise backbone network and the compute farm network.**

In this way, a single storage device can be used for all user data, laboratory data, application binaries, and scientific databases and repositories. Scientific productivity is greatly enhanced by eliminating the need for scientists to FTP their data from the enterprise data store to the local compute farm storage.

In addition, **centralized data storage has huge benefits to IT in terms of backup, recovery, scalability, and manageability.**

4.4. Considerations for Distributing BLAST on a Compute Farm

BLAST processing lends itself to processing on a compute farm because the data being queried can be segmented and separate instances of the BLAST application can be executed in parallel to achieve the results. This technique is called data decomposition.

Because BLAST is a so-called embarrassingly parallel application, organizations can shorten the turnaround time of their BLAST searches by increasing the number of CPUs in their compute farm (a.k.a. increasing the "*computing width*"), and thereby directly improve the productivity of their scientific research by allowing more sequence searches in a shorter time. (3)

While BLAST throughput is influenced by a number of factors (including CPU architecture and processor clock speed), since BLAST processing can be made parallel through data decomposition, the ability to increase the computing width introduces two additional, rate-limiting constraints:

1. Available physical memory

BLAST tries to load the target database directly into memory for optimal performance. BLAST databases that cannot be coresident in the computer's random access memory (RAM) during processing will take up to ten times longer to finish because the machine operating system must manage data flow between its RAM and its local hard disk drive. This situation, referred to as "paging," should be avoided at all costs.

Adding more RAM doesn't solve the problem. The most commonly used target sequence

databases (e.g., GenBank) are too large to fit into the 2- to 4GB RAM limitations of today's 32-bit computers. Sixty-four-bit systems built on Intel, Sun, Compaq, or IBM hardware eliminate these constraints, however, with the growth rate in biological data sources continuing to multiply, the best solution strikes a balance between number of processors and amount of physical memory per processor.

The key is to consider the physical RAM available to do the processing when performing the data decomposition prior to running BLAST. For example, if the compute farm contains 1GB of RAM per CPU, then the BLAST database segments need to be sized considerably less than 1GB to accommodate the application and data occupying RAM, and other memory-resident system software running local on the server.

This approach offers two benefits: a) higher performance by allowing increasing computing width; b) higher performance by ensuring that all applications and data will be coresident in RAM on the server, thus eliminating performance degradations due to paging.

2. Speed of the I/O subsystem

For this paper's purposes, the I/O subsystem comprises the network and the file server resources available to feed the BLAST applications and data to the compute farm. Consider a 10-CPU compute farm processing a BLAST request in which 500 query sequences are to be compared against three databases, of sizes 350MB, 600MB, and 1.8GB, respectively. A common strategy is to break the task into 25 jobs consisting of 20 sequences per job. As jobs are scheduled to each compute server, an NFS file transfer is initiated for each of the three databases. At 16MB per second throughput, this 27.5GB transfer will require approximately 29 minutes to complete. Since individual BLAST sequence searches usually complete in seconds, this overhead can be a high price to pay.

At larger scaling increments (100 CPUs or more), it is not uncommon to see a compute farm loaded with hundreds of jobs operating with mostly idle CPUs because each individual analytical process is waiting for pending I/O requests to complete. The problem is one of scale: No matter how large or over-engineered the network and the file server resources, a large (more than 10) number of I/O-hungry dual-processor machines can greatly tax the system with many large sequential read requests.

When the physical memory problem is solved with intelligent data decomposition of the sequence databases, the performance of the overall BLAST processing rests on the filer's ability to keep up with requests for application binaries and the database segments, and for the network resources to deliver these reliably to requesting CPUs. Therefore, the speed of the underlying file access and I/O subsystem has a dramatic effect on the pace at which discovery research can take place. (4)

4.5. High-Performance I/O Subsystem Considerations and Strategies

Without a scalable platform for discovery, scientists will have to limit the scope and expectation of their scientific inquiry. This is the worst possible outcome, since limits of computation are dictating the capacity to do science, not the other way around. To deal with this, life sciences IT teams have two principal objectives:

1. Invest in higher-performance storage and network resources (i.e., Gigabit Ethernet networks)

2. Deploy more clever data-balancing strategies to achieve scalability

Invest more in storage and networking hardware

The natural inclination is to increase the size of the data pipe to allow more data throughput from the storage to computing.

However, while network and filer performance are always important, the solution in this case is not about changing the storage architecture or fortifying the network-these simply shift the bottleneck. The ultimate solution is to reduce the load on the network and file server resources by bringing the data to be processed closer to the CPU that will process it only when needed. This is referred to as "data balancing."

Deploy NetApp network storage in concert with intelligent data balancing

The solution employs the local disk storage of the compute servers and automation to load balance the delivery of data files across the compute nodes themselves. Using this model, a setup job stages "chunks" of data from the NetApp filer to a subset of compute servers using NFS. This operation is very efficient due to NetApp's fast implementation of NFSv3. Once the staging transfers are complete, the client nodes themselves become both compute servers and file servers to the remaining client nodes. Jobs request data sets and the automation initiates server-to-server copies using sockets over IIOIP. In this way file serving is load-balanced, and application performance and predictability are greatly enhanced since all application I/O occurs local to the server node.

4.6. Blackstone Computing and NetApp Filers: Putting It All Together

Blackstone Computing has integrated the industry's NCBI-BLAST application with the company's PowerCloud Data Balancer software. The key features of Blackstone Computing's high-throughput BLAST solution include data decomposition of the target databases for the BLAST queries, accelerated database distribution within the compute farm, and automated updating for the frequently accessed databases on compute nodes. Combined with the key features of Network Appliance filer technology, a truly scalable and optimized BLAST compute farm solution is created with the following benefits:

- Linear scaling from 10 to 100 to 1,000 CPUs Very predictable application performance
- Commodity network, client nodes, and file server technology for the lowest cost;
- Improved reliability
- Easy administration, simple network diagram, easy backup, and easy scaling

This blend of compute farm architecture, NetApp file servers, and NCBI-BLAST enhanced with Blackstone's PowerCloud software yields *near-linear performance scalability*.

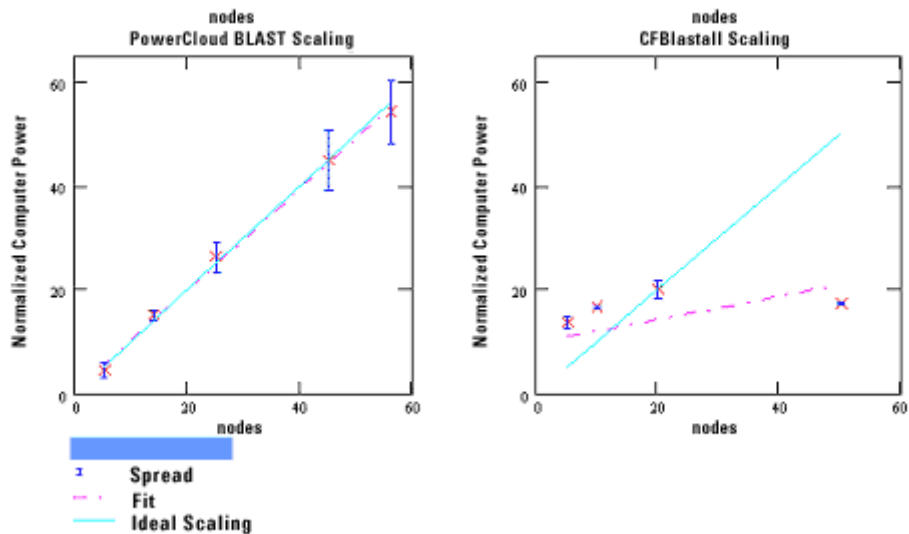


Figure 4.2. Blackstone Computing and NetApp benchmark on linear scaling.

The compute farm without Blackstone's data balancing and BLAST acceleration technologies (graph on right) suffered from I/O bottlenecks such that 50 nodes only delivered the power of 20. In contrast, the enhanced solution scaled with 98% efficiency.

As an example, a Boston-area biotech compared NCBI BLAST with and without Blackstone Computing's data balancing and BLAST-acceleration technologies. The company ran 100 jobs (10 sequences/job) against a 4GB EST database on a 130-CPU Linux® compute farm and compared processing time and CPU utilization. The traditional compute farm took more than two hours to process the jobs and utilized only five to seven percent of available CPU power. Blackstone's software, which alleviated network bottlenecks for the NetApp filer, delivered the same jobs in about ten minutes using 50 to 70% of available CPU power.

Additionally, a large pharmaceutical increased its compute farm size about five-fold to just over 50 (dual-CPU) servers. The new nodes were slightly faster than the old ones. Using Blackstone's data-balancing and BLAST-acceleration technologies, they ran a benchmark containing thousands of protein sequences against a popular genome database. The project results ran approximately five times faster than on the smaller system. The linear scalability of the compute farm reached nearly 100%.

5. Conclusion

A major benefit of the Network Appliance filer architecture in life sciences is the ability to connect the device to both the enterprise backbone network and the compute farm network. From these networks, users and systems have shared concurrent access to data. Furthermore, such centralized data storage has tremendous benefits to IT in terms of backup, recovery, scalability, and manageability.

NetApp filers are an excellent choice for life science organizations requiring a single, scalable solution—featuring a fast implementation of NFSv3—for both centralized management of user data and delivery of data to high-throughput compute farm applications, such as BLAST.

Network Appliance, Inc., and Blackstone Computing provide the life sciences industry with best-

Network Appliance Inc.

of-breed, complementary solutions that meet growing processing and data storage needs. More information on these companies can be found at <http://www.netapp.com> and <http://www.BlackstoneComputing.com>

6. Appendix A: NetApp: Serving a Broad Range of Applications & Further Benefits Details

In addition to compute farm storage, NetApp filers serve a broad range of applications in life sciences today, as detailed below.

Network Appliance Filers Serve a Broad Range of Applications in Life Sciences	
Portable personal files and user home directories	Centralizing the storage of user home directories and personal files on a filer appliance allows researchers to maintain a single personal repository that follows them regardless of the desktop, notebook, or server system they log in to. From a management standpoint, backups and capacity demands are far easier to handle if staff files are stored in a central location.
Canonical data repository	Organizations often have hundreds of gigabytes or even terabytes of flat file sequence, structure, or annotation data that must be maintained, curated, and updated. Often different groups within the same organizations may be downloading, formatting, and publishing the same data. Filers can be used to host the definitive "canonical" data sources within an enterprise or organization.
Instrument data collection & experimental result repository	High-throughput PC-assisted laboratory instruments that are not already tightly integrated with laboratory information management systems (LIMS) back-end databases can save result data directly to the filer rather than local disk. Raw result data is then immediately available to be processed or picked up by knowledge management platforms, automated analysis pipelines, data warehouse loading applications, and visualization/browsing tools.
Oracle usage	NetApp filers are qualified and regularly used in conjunction with Oracle® databases. For more information refer to TR3023, available online at http://www.netapp.com/tech_library/3023.html
Application and algorithm hosting	Organizations making widespread use of internally developed software applications or algorithms that are constantly in flux due to the software development may have difficulties with version control and software update issues. Hosting rapidly changing applications via network-attached filers enforces uniform version control and instantaneous distribution of new software revisions.

Further benefits of NetApp technology:

NetApp Makes Network Storage:	
Scalable and Manageable	Powerful and Highly Available
A filer product line that scales from 50GB to 18TB	Native (not emulated) multiprotocol support within Data ONTAP microkernel OS

Storage expansion without downtime or access disruption	Excellent published NFSv3 performance benchmarks
Fewer than 50 administrative commands (all available via GUI, serial, telnet, or rsh)	WAFL file system and RAID4 that protects data without parity-related performance hits
Snapshot™ technology that simplifies backup as users can easily browse and recover their own files	Battery-backed NVRAM, ensuring that all write operations complete even after power or hardware failures
	A cluster architecture that provides a framework for a highly available and load-balanced storage backbone

7. Appendix B: Configuring Filers: Key Considerations

Configuring Filers: Key Considerations	
Multiple network interfaces	A single Gigabit Ethernet connection should directly connect the NetApp filer to the aggregation switch per four Gigabit Ethernet uplinks from the edge switches to the aggregation switch. Additional Gigabit Ethernet connections can be used to provide redundancy and improved availability or storage services.
Maximize Fibre Channel loops	A single Network Appliance 800-series Fibre Channel adapter can support up to 56 devices contained within four shelf enclosures. Expansion past 56 drives requires an additional adapter. Performance gains can be realized by adding Fibre Channel loops before the 56-device limit is met.
Network Time Protocol (NTP)	Due to the importance of file creation, access, and modification time stamps in networked environments, the NetApp filer as well as all cluster server and compute elements should be configured to synchronize their clocks to a common standard reference. In absence of a centralized time server, the cluster portal node can be configured as an NTP master.
Hot spares	At least one hot spare disk should be configured. The disk needs to be at least as large as the largest available online disk.
Network Appliance clustering	Clustered filers are generally not strictly necessary in the BLAST-optimized (as example) compute farm due primarily to the reliability of the standalone filers and the short recovery time in situations where a forced reboot occurs.

8. Appendix C: Additional Information on NetApp Security and Data Protection Features

Time-to-Market, Data Protection, and Security
Using NetApp's Snap technologies, operational efficiency is increased dramatically. Bulk backups, restores, and copies can be done within minutes. This is not only beneficial from an administration standpoint, but also from a development standpoint. Consider a situation when a development team spends many hours on a simulation process and, at a particular point of time, a restoration of the database back to the starting point is required. This can

Network Appliance Inc.

be rendered easily and quickly using the SnapRestore® feature. Also, data can be protected using the SnapMirror® feature to copy data to a remote filer/location asynchronously.

Concerning security, NetApp filers' Data ONTAP software supports access control list (ACL) and UNIX file permissions. NetApp supports both Network File System (NFS) for UNIX and Common Internet File System (CIFS) for Windows® protocols as well as a mixed protocol environment, which is flexible yet secure. The CIFS model resembles the Windows NT®/Windows 2000 security model, supports Active Directory and Kerberos, and contains antivirus protection on file access. The NFS model resembles UNIX style permissions. The mixed model allows all data to be accessible by all protocols, UNIX permissions (user ID, group ID, user permissions, group permissions, and other permissions) and NT access control lists (user security identification [SID], primary group SID, file ACL) within the same file system while giving the administrator control over which file-level security model will be used in each area of the file server.

The SecureShare® technology ensures data integrity by providing file locking across multiple file-sharing protocols. File- and directory-level security can be configured to use Windows NT, UNIX, or "mixed" permissions in single- and multiprotocol applications.

Further, NetApp filers uses 'qtrees,' which are the basis for security configuration. On a NetApp filer, qtrees allow a flexible segmentation of a file system for quota management, opportunistic lock (olock) configuration, backup management, and security styles. Qtree security styles can be changed dynamically, deleted, or created, providing flexibility for the administrator to implement the security style that matches the security needs. Conversion of a UNIX operating system-style qtree into an NTFS-style qtree, for example, can be accomplished quickly and easily as storage needs change.

Additionally, NetApp filers also support the Network Data Management Protocol (NDMP), which is an open standard for centralized control of enterprise-wide data management. The NDMP architecture allows backup application vendors to control native backup-and-restore facilities in NetApp filers and other network-attached servers. Initially coauthored by Network Appliance and Legato Systems, NDMP is now being actively enhanced and adopted by the storage industry.

9. Footnotes

1. Source: <http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html>
2. It is important to note that compute farm management software is much easier to install and maintain if used with a common NFS mount for configuration files, fault fault-tolerance features, and logging operations.
3. For this reason, compute farm technology is a compelling alternative to symmetric multiprocessor (SMP) computer technology. The price/performance advantage enjoyed by compute farm technology over SMP is usually about 5:1.
4. NetApp filers support gigabit connections between the filer and storage device directly (NIC to NIC) or through gigabit switches. NetApp filers use an operating system called Data ONTAP™, which is based on the trademarked "Write Anywhere File Layout" (WAFL®) technology. One of the most common latency issues related to writing to disk is

the speed of the operation, which involves not only writing data to disk, but also maintaining a parity bit. With the WAFL implementation of the RAID4 configuration, the head is not restricted to a specific spot on the disk to write parity, instead it can write anywhere on the disk that is closest to the data spot. This eliminates seek time on the head, thereby improving performance at every read and write operation. This is the most basic way that a filer improves performance and guarantees performance gains for all users who read or write data.