The Center of Excellence in Bioinformatics at the University at Buffalo

CHALLENGE	SOLUTION	BENEFIT
Design and install a large-scale, high-performance computing (HPC) cluster for bioinformatics research that requires trillions of complex calculations per second and at least 10 TB of storage	A 2,000-node HPC cluster comprising dual-processor Dell™ PowerEdge™ 1650 and PowerEdge 2650 servers using Intel® Xeon™ and Intel Pentium® processors and running the Red Hat® Linux® operating system; a storage area network (SAN) with a Dell | EMC FC4700 storage array and two Dell PowerVault™ 136T tape libraries	High levels of processing power at a better price/performance compared to supercomputers; stable backup solution

The quest for a cure

Dell and the Center of Excellence in Bioinformatics at the University at Buffalo create Linux-based clusters running on Intel processors and dedicated to cutting-edge research in bioinformatics

The mystery of the human body has eluded scientists for centuries. Through the years, scientists have performed research and developed theories—with the help of supercomputers—about the human body and the deadly diseases that attack our very existence. In some cases, their research has resulted in effective treatments and powerful drugs that have all but annihilated specific diseases. But the quest for a cure—or at least some relief—for threats such as cancer, AIDS, or Alzheimer's disease continues.

Today, researchers and scientists in the field of bioinformatics focus on topics such as identifying, sequencing, and understanding the human genome—and developing molecular models of even the tiniest proteins in biological agents. This analysis requires high-end computing and visualization technology to help process resource-intensive research. Traditionally, the discipline's computational needs were met by supercomputers, which are very expensive. In recent years, however, high-performance computing (HPC) clusters built with commodity components have become a viable alternative, offering a cost-effective way to obtain the required processing power.

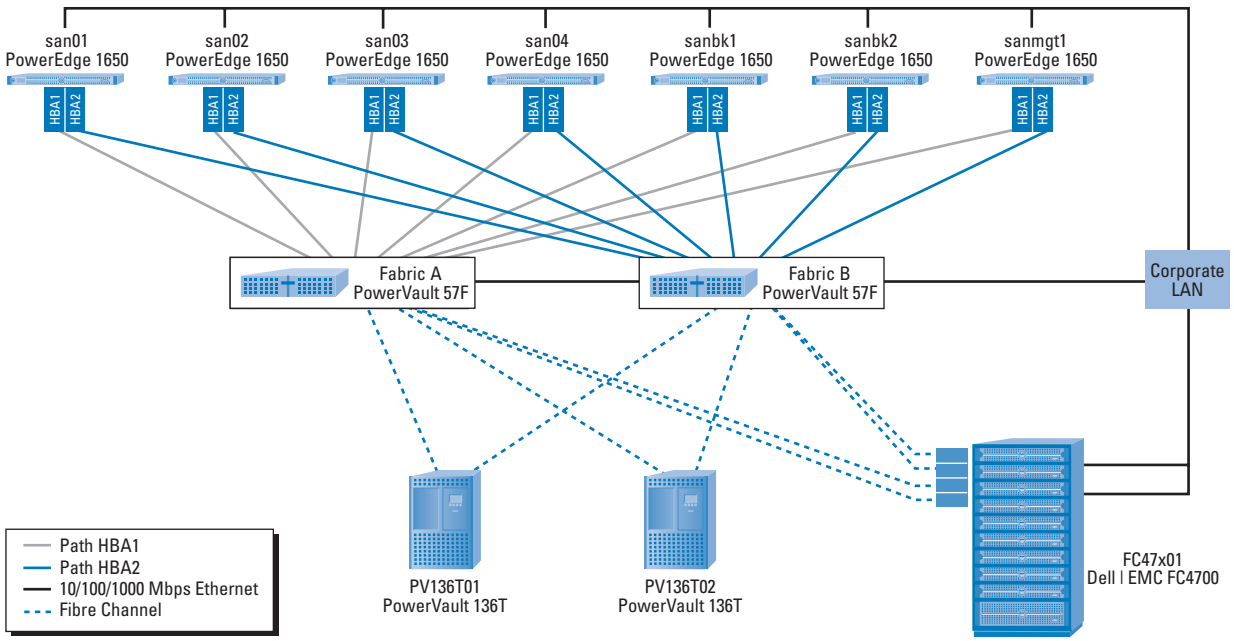
The newly formed Center of Excellence in Bioinformatics at the University at Buffalo (UB), a campus of The State University of New York, combines such high-end technologies as supercomputing and visualization with scientific expertise in disciplines such as genomics, proteomics, and bioimaging. Because of the high costs of supercomputers, the center wanted a cost-effective alternative for its new facility. As director of the Buffalo Center of Excellence in Bioinformatics, Dr. Jeffrey Skolnick wanted a fast, reliable, and scalable HPC cluster to support his research in computational biology. This cluster would be devoted to running proprietary software that performs protein-folding simulations and calculations. Because of the large amount of data generated by the calculations, the cluster needed to accommodate at least 10 TB of storage and have a stable backup solution.

Strategic staging and efficient teamwork

The importance of this research combined with the need for cost-effective processing set the stage for a partnership that included corporate, government, and non-profit organizations—all to serve the bioscience community. Dell provided the computing power as the corporate relationship in this equation with a high-performance computing (HPC) cluster.

Because the UB center was new, the site that would house the cluster—which was under construction—and the cluster itself were on a similar schedule for becoming operational. To build the cluster within the project schedule, Dell needed a temporary facility that could handle the power and HVAC demands of a 2,000-node cluster. At a maximum of 2 million BTUs per hour, this cluster required approximately 170 tons of cooling capacity. The team chose a facility on Long Island, New York, which had adequate power and cooling facilities so they could build the cluster in parallel with the construction of the new data center at UB. This nearby facility enabled quick delivery and installation of the configured cluster once the data center was completed.

Dell also faced a tight deadline: It had a little more than five weeks to build the entire cluster and a storage area network (SAN),



The Fibre Channel SAN environment at the Buffalo Center of Excellence in Bioinformatics

implement a backup solution, and perform acceptance testing. Dell created two teams that worked on three shifts. At the Long Island facility, one team racked and stacked equipment, configured software, and tested the configurations. This team then disassembled each rack as it was completed and shipped the hardware to the UB location, where the second team performed final configuration, testing, and verification of the cluster. Once installed at UB, the cluster passed all preset acceptance tests and goals set forth at the project's inception.

Dell servers give cluster the power to perform

To meet Dr. Skolnick's computational needs, 1,900 Dell™ PowerEdge™ 1650 servers became compute nodes and four served as master nodes that provided a centralized management platform for managing applications on the compute nodes.

A system of this magnitude requires maximum computing power in the smallest form factor possible to conserve space, power, and ultimately cost. PowerEdge 1650 servers had the density required to support this type of system, allowing 41 servers and an Ethernet switch to fit in each 42U rack. Each PowerEdge 1650 contained an 80 GB hard drive, dual Intel® Pentium® III processors at 1.26 GHz, 1 GB of RAM, and dual gigabit[1] network interface cards (NICs).

An HPC cluster with centralized storage has demanding I/O and bandwidth requirements. Therefore, four PowerEdge 1650 servers were dedicated to providing network file services to the cluster nodes. In addition, two PowerEdge 1650 servers served as dedicated backup and recovery management machines.

Based on Dr. Skolnick's requirements for a cost-effective, scalable, and reliable development environment, Dell installed 100 PowerEdge 2650 servers, each with a 73 GB Ultra SCSI hard drive, dual Intel Xeon™ processors at 2.4 GHz, 1 GB of RAM, and dual gigabit NICs. This configuration allows cluster users to execute jobs on the 1,900 production nodes, the 100 development nodes, or all 2,000 compute nodes.

Cisco® switches provided 100BaseT connections for the servers. Each switch also had dual gigabit uplinks that were connected to two Extreme Networks® switches, which created a highly redundant network infrastructure.

Designed for performance and reliability, the SAN incorporated ninety 181 GB Fibre Channel disk drives in a Dell|EMC FC4700 storage array and eight disk array enclosures (DAEs). Two Dell PowerVault™ 136T tape libraries provided tape backup facilities. All SAN devices connected to two PowerVault 57F[2] Fibre Channel switches.

Software enhances cluster manageability and operations

The PowerEdge server nodes run the Red Hat® Linux® operating system. To install and configure each node, Dell employed the MPI Software Technology Felix™ management system for Linux-based

[1] Gigabit Ethernet indicates compliance with IEEE® 802.3ab and does not connote speeds of 1 Gbps.

[2] Newer models are available at http://www.dell.com.

> "Dell's exceptional price/performance allowed us to acquire low-cost servers that will give us extremely high levels of computing power. Deploying industry-standard technology in the form of a server cluster enables us to process the massive amount of data that is critical when doing this type of research."
>
> – **Dr. Jeffrey Skolnick**
> Director, Center of Excellence in Bioinformatics
> University at Buffalo

clusters. Felix enables the master nodes to support the configuration and management of the compute nodes. It lets the cluster system administrator use a master node to monitor, manage, and maintain all or subsets of the cluster nodes, which can be grouped according to user-defined attributes.

Scheduling is the key to cluster operation, so Dell implemented the Platform LSF® 5 scheduling software for this important task. Because of its almost unlimited scalability—Platform LSF 5 can support up to 200,000 processors and more than 500,000 jobs in a cluster—the team felt confident that it would meet the high demands placed on a job scheduler by a 2,000-node cluster.

With more than 2,000 pieces of hardware in the cluster, the bioinformatics center needed a simple, scalable solution for monitoring vital statistics and maintaining the overall health of its investment. To provide this functionality, the team installed Dell OpenManage™ Server Administrator agents on each node and Dell OpenManage IT Assistant on four management nodes. Dell OpenManage is based on the industry-standard Simple Network Management Protocol (SNMP) for seamless integration into existing enterprise management platforms.

For backup and recovery, Dell used a combination of EMC® SnapView™ and VERITAS NetBackup DataCenter™ software. SnapView, a storage-array-based tool, captures snapshot images of a file system for nondisruptive, consistent backups of production data. It integrates seamlessly with NetBackup for data and backup media management. The NetBackup media servers comprised two PowerEdge 1650 servers running Red Hat Linux.

Massive processing at maximized price/performance

The bioinformatics research performed by Dr. Skolnick and his team requires massive computing power, historically the domain of multimillion-dollar supercomputers. Using the Dell cluster, the researchers can conduct their work at a fraction of the cost.

"Dell's exceptional price/performance allowed us to acquire low-cost servers that will give us extremely high levels of computing power," says Dr. Skolnick. "Deploying industry-standard technology in the form of a server cluster enables us to process the massive amount of data that is critical when doing this type of research."

According to Dr. Skolnick, the amount of data to be analyzed would take approximately 2,000 years to analyze on a single computer with one processor. By using the cluster, he expects to complete his initial data analysis in just six months. The Dell cluster, capable of performing more than 5 trillion calculations per second, is an important tool that allows Dr. Skolnick to further his center's vision: to enable the development of new medical treatments for cancer, Alzheimer's disease, AIDS, and other diseases by creating state-of-the-art algorithms for data acquisition, storage, management, and transmission.

Success is contagious

Based on the success of this cluster, the Center for Computational Research (CCR) at UB decided to deploy a 300-node Dell HPC cluster to assist general UB scientific research efforts, such as tracking pollution in the Great Lakes. This cluster—comprising 300 Dell PowerEdge 2650 servers, each with dual Intel Xeon processors at 2.4 GHz—has become the highest ranking Dell system on the TOP500 Supercomputer Sites list[3]. These UB clusters illustrate how standards-based systems are expanding into the territory once exclusively inhabited by proprietary supercomputers.

FOR MORE INFORMATION

http://www.dell.com/hpcc
http://www.dell.com/servers
http://www.dell.com/storage
http://www.intel.com

[3] TOP500 Supercomputer Sites, http://www.top500.org.