

UC BERKELEY CENTER FOR INTEGRATIVE GENOMICS

THE CUSTOMER

The Center for Integrative Genomics brings together researchers from traditionally separated fields of study to analyze and compare the genome sequences of a broad spectrum of organisms in order to determine the mechanisms responsible for evolutionary diversity among animals, plants and microbes. Faculty at the Center are drawn from distinct academic departments at UC Berkeley and Lawrence Berkeley National Laboratories, including molecular and cellular biology, integrative biology, computer science, bioengineering, plant and microbial biology, mathematics, and public health. One of the Center's primary objectives is to decode regulatory DNA - the regions of the genome that control gene expression - and the protein factors that bind to them. By understanding how a common set of genes is deployed differently in various organisms, they hope to reveal the mechanisms of evolution.

THE CHALLENGE

When the Center opened in 2002, project leaders Michael Levine PhD, Gene Myers PhD, and Lior Patcher PhD realized that to conduct comparative analysis at the scale necessary to be successful, they needed to build an IT environment that could process, store and retrieve multi-gigabyte files containing the DNA sequences of various organisms. Most comparative analyses involve two such large data sets, for example, the genome of a human vs. the genome of a mouse. The undistilled results of such comparisons deliver datasets typically on the same order as the input. But a given multi-way analysis involves performing tens and hundreds of such computations. The Center wanted a solution that delivered the performance and flexibility to conduct many such analyses on a weekly basis. While compute power was an issue, the more important consideration was to have a single conceptual data system that contained all the data and results, and that could deliver the inputs and receive the results from the CPU's in a highly efficient and thus necessarily, parallel manner. "The ability to quickly and easily conduct comparative analyses from a single, easily maintained file system is primary to our



Industry: Life Sciences

The Challenge:

To quickly and easily conduct comparative analyses of hundreds of computations required to accomplish their mission to research and understand gene regulation. They wanted a storage solution with exceptional I/O performance, seamless scalability and one that was easy to manage. It also had to meet their price/performance ratio for an academic budget.

Panasas Solution:

The fully integrated software/hardware solution included the Panasas® ActiveScale® Operating Environment and the PanFS™ parallel file system with the Panasas DirectFLOW® protocol.

Key Results:

- Parallel access to compute cluster for exceptional I/O performance
- Maximized cluster utilization
- Flexibility in running data set comparisons
- Maximized ROI from clustered computing environment

mission," said Gene Myers, from the Department of Computer Science, University of California, Berkeley. "It is only through extensive testing that we will be able to understand gene regulation."

On the compute side of the solution, the Center deployed a Linux cluster with 36 CPUs to meet their processing needs. But key to the overall solution was the deployment of a high performance storage system. Specifically, the

“After conversations with several storage vendors, it was clear that Panasas was the only company that really understood what we needed”

- Gene Myers

Department of Computer Science,
University of California, Berkeley



Center needed a storage solution that could deliver exception I/O performance, scale seamlessly to handle a growing number of data sets, and be flexible enough to handle a quick change of one, or both, of the data sets being compared. The team knew that a distributed file system could offer many of the things they were looking for, but they needed to find a solution that was easy to manage and one that delivered an exceptional price/performance ratio.

THE SOLUTION

The team at the Center looked at several different solutions. Most standard storage products had the architectural limitation of delivering data through one, or possibly two, ethernet ports; a choke point that immediately eliminated them from consideration. Several solutions involved purchasing commodity components and then adding a software layer that provided a distributed file system, but the price/performance ratio on these approaches was poor. Only the Panasas® Parallel Storage Cluster, an integrated software/hardware distributed file system, provided the desired price/performance ratio for an academic budget and delivered the integration needed to simplify installation and management as well as reduce recurring administration costs. “After conversations with several storage vendors, it was clear that Panasas was the only company that really understood what we needed,” said Myers.

A three shelf, multi-TB Panasas Storage Cluster was deployed to support the 36 CPU cluster configured in a 6 x 6 matrix. The head node of each row in the cluster matrix has two Gigabit Ethernet connections to the Panasas system. Data is pipelined through the configuration and all CPUs in the system are able to receive and push data simultaneously at the peak capacity of the network.

THE VALUE

Once moved into production, the Panasas Storage Cluster enabled the Center to realize the benefits of a comprehensive cluster architecture. “Every processor in our architecture is now running at peak performance,” said Myers. “The resulting parallelism is delivering a 6X performance improvement in our environment.” The Panasas system is designed from the ground up to deliver exceptional random I/O and data access throughput, breaking the traditional storage bottleneck and allowing direct disk-to-cluster node access.

Further benefit for the center was the flexibility and ease of management offered by a single, global namespace. “With other solutions, we were forced to trade off manageability and flexibility for greater performance,” commented Myers. “With Panasas, we can get it all.” Finally, the lab is confident knowing that as the system grows in size it will still be flexible and easy to manage, plus the performance will scale with capacity in linear fashion.

“Every processor in our architecture is now running at peak performance. The resulting parallelism is delivering a 6X performance improvement in our environment.”

- Gene Myers

Department of Computer Science,
University of California, Berkeley



Accelerating Time to Results™

6520 Kaiser Drive Fremont, California 94555 Phone: 1-888-PANASAS Fax: 510-608-4798 www.panasas.com
1-888-PANASAS (US & Canada) 00 (800) PANASAS2 (UK & France) 00 (800) 787-702 (Italy) +001 (510) 608-7790 (All Other Countries)

©2007 Panasas Incorporated. All rights reserved. Panasas, the Panasas logo, Accelerating Time to Results and ActiveScale are trademarks or registered trademarks of Panasas, Inc. in the United States and other countries. All other trademarks are the property of their respective owners.