



POWERWULF CLUSTERS: BIOINFORMATICS

CASE STUDY

PROFILE

Dartmouth College's DISCOVERY Supercomputing Cluster

CHALLENGE

Processing large, multi-gigabyte data files for human genetics research Scaling processor cores to expand the compute cluster without increasing the amount of space needed.

SOLUTION

Dartmouth selected system integrator PSSC Labs to supply an initial PowerWulf Cluster consisting of 25 dual-core nodes in 2005, and now has nearly 90 nodes with 328 processor cores, 600 gigabytes of memory, and 11 terabytes of disk space.

IMPACT

Schmitt and his team have built the largest computing cluster at Dartmouth and one of the largest educational computing clusters in New England. This facility enables world-renowned research into the genetic causes of cancer and other diseases and also provides high-performance computing resources for engineering, physics, and other programs.



POWERWULF CLUSTERS: BIOINFORMATICS



CASE STUDY CONTINUED

ORGANIZATIONAL PROFILE

Founded in 1769, Dartmouth College is an Ivy League school that offers an outstanding undergraduate education along with world-famous graduate institutions, including the Tuck School of Business, Dartmouth Medical School, and The Thayer School of Engineering.

To enhance its medical research capabilities, the College's Norris-Cotton Cancer Center hired Jason Moore, a renowned genetics research scientist, to build a large computing cluster at the College in 2004. The cluster became known as the Dartmouth Initiative for SuperComputing Ventures in Education and Research, or DISCOVERY. Moore had overseen development of a similar cluster at Vanderbilt University, and Dartmouth wanted to provide similar or better facilities for its Computational Genetics Computing Laboratory.

CHALLENGE

Increases in computing power over the past two decades have driven far more sophisticated data analyses in the field of genetics. Many of these compute sessions involve massive files – as large as 20 gigabytes or more. A leading graduate educational institution, Dartmouth College wanted to provide its genetics students with up-to-date computing resources that would not only speed execution of their projects but enable new and highly sophisticated analyses.

In 2005, the College's Computational Genetics Lab began an effort to build a supercomputing server cluster and hired Peter Schmitt as the Lab's technical director. A former programmer with no prior experience in building clusters, Schmitt had a fast and steep learning curve. He interviewed server manufacturers and system integrators while evaluating cluster management software, educating himself about how these large systems were built and run.



POWERWULF CLUSTERS: BIOINFORMATICS



CASE STUDY CONTINUED

SOLUTION

After evaluating major suppliers of clustering hardware and software such as HP, IBM, and Sun as well as third-party system integrators, he selected PSSC Labs, a Southern California-based systems integrator focusing on high-performance server clusters for corporate and government clients. "PSSC Labs had the best combination of service and price," says Schmitt. "Some vendors had lower prices with no service, while others had great service with very high prices. PSSC Labs had just the right combination." In stating his requirements, Schmitt had one firm request. "We requested AMD Opteron™ processors in the servers because we believed their memory management was superior," he says. "Our cluster is 100 percent AMD Opteron processor-based." This even includes the processors in legacy servers the lab owned before PSSC brought in its equipment. PSSC Labs supplied servers with 64-bit Dual-Core AMD Opteron processors, along with 8 gigabytes of RAM, high-speed, low-latency InfiniBand interconnects, and one 80-gigabyte hard drive.

Although the servers arrived with nearly perfect configurations, selecting cluster management software involved a longer period of trial and error for Schmitt. "We started off with Maui and Torque as the cluster software," he says, "and we have now settled on Moab, which has been a great product."

In addition to the PSSC Labs PowerWulf cluster servers, Schmitt added some existing server nodes with single-core AMD Opteron processors to create a free pool of computing resources for the engineering, physics, and chemistry students. "We share the cluster's resources with the rest of the community," he says. "We have a buy-in process where these other departments actually purchase hardware nodes and get four years of access to the cluster. But there's always enough performance left over for the genetics jobs."

Students and professors in the Computational Genetics Lab develop and run their own applications using standard tools such as C++, FORTRAN, Perl, Python, and Java. Students from the engineering, physics and chemistry departments use applications such as Fluent (a computational fluid dynamics tool), EMAN (a set of image/volume processing tools that perform single particle reconstructions to determine the 3-dimensional structures of molecules), and MatLab (a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numerical computation).



POWERWULF CLUSTERS: BIOINFORMATICS



CASE STUDY CONTINUED

SOLUTION CONT.

As manager of the cluster, Schmitt also developed a web site that shows current utilization and usage information and facilitates scheduling to minimize request calls from potential users. The web site (<http://discovery.dartmouth.edu>) is open to anyone who wants to know more about the DISCOVERY cluster. As manager of the cluster, Schmitt also developed a web site that shows current utilization and usage information and facilitates scheduling to minimize request calls from potential users. The web site (<http://discovery.dartmouth.edu>) is open to anyone who wants to know more about the DISCOVERY cluster.

IMPACT

The performance and reliability of the PSSC Labs PowerWulf cluster servers have been outstanding, and the cluster now supports major projects in cancer research as well as other disciplines. "Students can now use hundreds of processors to handle a computationally-intensive problem or to process 26 gigabytes of data in a few days when it would have taken a year or more on a single system," says Schmitt.

In addition, InfiniBand's low latency server interconnects are also delivering fast results. "We have a user who saw a 70 percent improvement in processing speed on a job due to the InfiniBand interconnect," says Schmitt.

Although the PowerWulf cluster started with 25 dual-core nodes in 2005, Schmitt added dozens of nodes during 2006 to bring the cluster to its current total of 328 processor cores. For the future, he wants to increase the number of processing cores. "When Quad-Core AMD Opteron processors come out, we will replace our dual-core units as they reach their end-of-life cycles," says Schmitt. "We want to get to 500 CPUs within the next two years, and the only way we're going to get there in the space available is to use quad-core CPUs." Fortunately, the sophisticated power management capabilities of the AMD Opteron processor will allow Schmitt and his team to upgrade to Quad-Core AMD Opteron processors without linear increases in power requirements. Throughout the cluster's development, PSSC Labs has provided highly responsive support. "We get very excellent turnaround on our service requests, and we can always ship a system back to them if it's a big problem," says Schmitt.

By relying on AMD processor performance and PSSC Labs' deep knowledge of educational and scientific computing clusters, Peter Schmitt and the Dartmouth Computational Genetics Laboratory have built a truly scalable community resource that is helping speed the advance of medical science.