

Sandia's Revolutionary Cplant: The World's First Virtual Supercomputer

“Another kind of revolution is going on. A major government laboratory like Sandia is willing to spend \$9.6 million plus a significant amount of in-house development to make a supercomputer out of a supply of off-the-shelf parts.”

—Rolf Riesen, Lead Cplant Software Developer

In late January 2001, Sandia National Laboratories in Albuquerque, New Mexico, received another 280 Compaq *AlphaServer*™ DS10L systems running Linux. These joined the 1,800 Compaq *AlphaServer* DS10L systems already at Sandia's Cplant Antarctica, a large-scale, massively parallel computing resource created from a cluster of commodity computing and networking components.

Formerly composed of only 600 Alpha workstations, Cplant (short for “Computational Plant”) has already been ranked 44th among the world's fastest supercomputers, as well as the largest “production” Linux cluster. Bill Camp, director of Sandia's Center for Computation, Computers, and Math, believes that Cplant has had the world's “most powerful Linux computer clusters for at least two years now.” He also predicts that “by 2005, we will be at the top of the list of the world's fastest supercomputers.”

Origins of Cplant

Sandia is a national security laboratory operated for the U.S. Department of Energy (DOE) by the Sandia Corporation, a Lockheed Martin company. Sandia designs all non-nuclear components for the nation's nuclear weapons, performs a wide variety of energy research and development projects, and works on assignments that respond to national security threats—both military and economic. Sandia also seeks partnerships with appropriate U.S. industry and government groups to collaborate on emerging technologies that can support its mission.

As a principal contractor to the DOE, Sandia is active in the DOE's Accelerated Strategic Computing Initiative (ASCI), which funds the design and operation of parallel computers that are among the world's most powerful systems. The fastest supercomputers in the world are critical to the DOE's science-based stockpile stewardship program, which requires extremely high computational speeds to simulate nuclear explosions and make sense out of the resulting flood of data. ASCI Red, Sandia's Intel-built supercomputer, was the fastest machine in the world for several years until early July 2000, when another DOE supercomputer—ASCI White—was built for the Lawrence Livermore National Laboratory by IBM. At Sandia, ASCI Red was quickly becoming over-subscribed, and backup was needed.

Sandia researchers set about to create a “poor man's” ASCI Red architecture by combining high-performance commodity parts with Sandia software. Because they had helped to develop the system software that made Red the fastest computer in the world, they believed they could succeed with an off-the-shelf version. So they took up the task of physically linking the highest performance commodity PCs in the world into a tightly knit cluster, and developing the software to make it work. They called this project “Cplant.”

A new concept in supercomputing

“Cplant, to put it simply, is a collaborative venture between Sandia and Compaq to build the world's first *virtual* supercomputer,” says Camp.

“Supercomputers for the past decade have traditionally been purchased as turnkey machines from the world's largest computer makers,” says Neil Pundit, manager of the Scalable Computing Systems department. “Such machines have cables, connection boxes, as well as monitors and testing equipment, already built in place. In Cplant we are following a new path, assembling a supercomputer out of parts, open-source software, and our own developments. We wanted to partner with a company that is as much committed to innovation as Sandia. Compaq is at the forefront of supercomputing technology—cluster technology, which is the basis of Cplant.”

Cplant, or “Computational Plant,” refers to physical computational hardware. But there is a second meaning, as in an organic plant that grows, evolves, and is pruned. A key concept in the project is that each year a new phase (or branch) will be added to the plant to increase its capability with the latest cost-effective hardware components. After three years, older and possibly obsolete hardware should be pruned away. For this strategy to work, Cplant had to be designed with scalable units—basic building blocks—that enable the machine to adapt to change.

Improving on Beowulf

The idea of combining off-the-shelf computers to create an inexpensive computational cluster with supercomputer functionality is generally credited to Thomas Sterling, who was one of the creators of the Beowulf system in the mid 1990s. Beowulf systems are typically loosely coupled, and devoted to doing no more than a few applications for small groups of researchers. As such, they achieve cost savings and simplicity, but sacrifice scalability, balance, and generality. Cplant differs from Beowulf clusters in that it is “a true multipurpose supercomputer,” says Camp. “Scientists can run any program in exactly the same fashion as though they were using ASCI Red.”

Cplant also has scalability. As Arthur Hale, Sandia’s deputy director of Scalability, puts it, “One of the biggest things that we want is to be able to scale up to an order of 10K of these servers in a cluster.” The software developed by Sandia is up to that challenge: “Most researchers have a hard time convincing their sponsors that this approach is feasible,” says Rolf Riesen. “Ordinarily, the software out there doesn’t scale to such numbers of nodes. Our software, on the other hand, has already run. So Sandia has jumped out ahead of the pack.”

Three Phases of Cplant, 1997-2000

Cplant has already been through several phases. The original Cplant was assembled in 1997 from 128 Digital Personal 433au Workstations based on Alpha 21164 processors running at 433 MHz. In 1998, 450 Personal 500au Workstations were added, also based on the 21164 processor. Four *AlphaServer* 1200 systems were delivered as well, to act as I/O servers for the cluster. In 1999 the Cplant facility was brought up to 1,400 processors, when 800 *AlphaStation*TM XP1000 systems running at 500 MHz were purchased. The *AlphaStation* XP1000s were based on the next-generation 21264 processor and chip set, with significantly higher sustained performance per processor. That same year, Cplant ordered 16 *AlphaServer* DS20 systems and a 3 TB Compaq *StorageWorks*TM solution. In each phase, Cplant was able to incorporate the latest technology and have it work with existing components.

In the course of the year 2000, some 1,800 *AlphaServer* DS10L systems were added, more than doubling the total size of Cplant. Because the *AlphaServer* DS10L is less than two inches tall, up to 42 DS10L systems can be packaged in a standard rack. The Sandia design packages 33 systems to a rack—more than four times the density of previous Cplant cabinets—leaving room for high-performance interconnections, networking, and system management components. The new racks are designed to require minimum external connections, allowing the major functional units of the systems to be integrated and tested during their manufacture at Compaq. The result is greatly simplified installation and maintenance.

Cplant’s *AlphaServer* systems run a modified version of Red Hat Linux (V5.1), plus the parallel systems software developed in the Cplant project. Internal communications among processors are carried out over the newest Myrinet links and switches developed by Myricom Corporation. The several internal communications networks in Cplant are critical to managing the computer as a single resource, and to carrying large parallel jobs.

Phase 4 begins

In January 2001, Bill Camp summed up the situation at Cplant as follows:

“We have *AlphaServer* DS10 systems with 1024 compute processors up and running with a parallel file system and the new Myrinet switches and NICs. This is .954 teraflops (1024 x 466 MHz x 2 flops/cycle). We have DS10s with another 256 compute processors in that cluster, but currently operating as an independent subcluster. This subcluster is currently being used for systems software development and

testing. It will be combined with the 1024 into a single cluster within a few weeks. This will bring us to 1.193 teraflops.

“We have 9 more DS10 cabinets with 32 compute processors each in receiving. This is 280 compute processors. They will be combined within the big cluster in a couple of months. That will bring us to 1.454 teraflops.

“This summer we will take the 512 *AlphaServer* XP1000 systems from Cplant Siberia (the California subcluster), put in the new Myrinet NICs and switches and combine them into Cplant Antarctica. This will bring us to 1.966 teraflops in the full cluster by summer.

“By the way, the DS10 cabinets actually have 33 processors. We only count the 32 that are used for computational (versus systems) tasks. If we counted all of them, the total would actually be 2.12 teraflops.”

The open source philosophy

“Sandia likes Linux for clusters because it’s open source,” says Arthur Hale. Following this philosophy, Cplant will shortly release its own software to the general public. As Rolf Riesen puts it, “Everyone in the world will help us improve it. Otherwise we would have this proprietary code that no one knows about, and something else that may not be as good could become the standard to be improved. And when we hired people, they wouldn’t have experience with the systems we’re running.”

Bill Blake, vice president of Compaq’s High Performance Technical Computing Group, applauds this notion. “Sandia is doing pioneering work in building truly large Linux systems, using a combination of open source software, their researcher’s own development, and hardware, tools, and compilers from Compaq.”

“We like Linux with Compaq,” says Hale, “because of their software offerings on the Linux platform, particularly the compiler technology. Getting performance out of the Alpha on Linux is very exciting.”

Business results

- More efficient operation of mission-critical applications
- Exceptional price/performance for floating-point, intensive calculations
- Increased computational power for complex 3D modeling

What makes it work

Systems

4 *AlphaServer* 1200s, 16 *AlphaServer* DS20s, 2,080 *AlphaServer* DS10s
128 Compaq Personal 433au Workstations, 450 Compaq Personal 500au Workstations, 800 Compaq *AlphaServer* XP1000 Professional Workstations

Software

Linux

Beta Linux compilers

Storage

StorageWorks systems with HSZ80 controllers

January 2001

Spokespersons

Bill Camp

Director of the Center for Computation, Computers, and Math

Sandia National Laboratories

Albuquerque, New Mexico

505-844-5678

<http://www.sandia.gov>

Neil Pundit
Manager of Scalable Computing Systems

Arthur Hale
Deputy Director of Scalability

Rolf Riesen
Lead Cplant Software Developer

© 2001 Compaq Computer Corporation. All rights reserved.

For more information about Compaq products and services, call 1-800-AT-COMPAQ or visit the Compaq Worldwide Web site at www.Compaq.com

This case study was prepared with the cooperation of the companies described. The material, however, does not constitute an endorsement of Compaq products by the company.

Product names mentioned herein may be trademarks and/or registered service marks of their respective companies. They do not represent an endorsement by Compaq Computer Corporation.

Trademarks

Compaq and the Compaq logo are registered in the U.S. Patent and Trademark Office. AlphaServer, AlphaStation, and StorageWorks are trademarks of Compaq Information Technologies Group, L.P. in the U.S. and/or other countries. All other product names mentioned herein may be trademarks or registered trademarks of their respective companies.