IN-[illegible]

[illegible]

Computing with Beowulf

Parallel computers built out of mass-market parts are cost-effectively performing data processing and simulation tasks.

The Supercomputing (now known as "SC") series of conferences celebrated its 10th anniversary last November. While vendors have come and gone, the dominant paradigm for tackling big problems still is a shared-resource, commercial supercomputer. Growing numbers of users needing a cheaper or dedicated-access alternative are building their own supercomputers out of mass-market parts. Such machines are generally called Beowulf-class systems after the 11th century epic.

This modern-day Beowulf story began in 1994 at NASA's Goddard Space Flight Center. A laboratory for the Earth and space sciences, computing managers there threw down a gauntlet to develop a $50,000 gigaFLOPS workstation for processing satellite data sets. Soon, Thomas Sterling and Don Becker were working on the Beowulf concept at the University Space Research Association (USRA)-run Center of Excellence in Space Data and Information Sciences (CESDIS). Beowulf clusters mix three primary ingredients: commodity personal computers or workstations, low-cost Ethernet networks, and the open-source Linux operating system.

Celebrated as a voluntary effort with no owner, Linux is a freely available Unix flavor that programmers can modify or extend at will. It currently runs on 386 and upward Intel-compatible, Compaq/Digital Alpha, Power PC, and, as of December, Sun UltraSPARC chips. Becker had been contributing to Linux by writing software to drive the different network adapters these chips require and subsequently added software to connect multiple networks. "Networking is a major component of turning an operating system into a cluster operating system," Becker said. Software updates continue as new chips and networks are released.

The first Beowulf linked 16 Intel 486 processors (100 MHz) with standard Ethernet (10 megabits per second) and performed 50 megaFLOPS on applications, about 1/20th of the initial goal. It took the release of the Pentium line, specifically the 200-MHz Pentium Pro, and 100 megabit-per-second Fast Ethernet to reach one gigaFLOPS for $50,000. That was achieved independently on two 16-processor machines running cosmology simulations at NASA Jet Propulsion Laboratory/Caltech and Los Alamos National Laboratory in 1996.

With performance improvements passed on by the mass market at constant or decreasing cost, institutions worldwide are turning to Beowulf systems to meet their data processing and simulation needs. Beowulfs probe human body molecular dynamics at the National Institutes of Health, fight computer crime for NASA, and teach California high school students about parallel computing. True to the original motivation, Earth and space science applications are proving among the most successful.

One of the larger Beowulfs is Goddard's Highly-parallel Integrated Virtual Environment, or HIVE for short. "Social insects are a good analogy for parallel computing," said its chief creator John Dorband. The HIVE corrals 128 Pentium Pro 200-MHz chips with Fast Ethernet. Most of the $210,000 to build it came from NASA's Regional Earth Science Applications Centers Program, which recently awarded nine grants to study problems ranging from water resource management to urban sprawl. The plan is for some Centers to have their own HIVEs.

An early step in using Earth science data is distinguishing among different kinds of land cover and water as well as their various uses. One way to accomplish that is image segmentation, searching through satellite observations for regions with common attributes. "The goal is to make a thematic map including identification of water, forest, agriculture, and human development, such as homes, roads, and industrial complexes," said Goddard scientist Jim Tilton. "Image segmentation is especially good for tracking changes in land cover, if you use it consistently across many data sets."

Most image labeling routines try to label each pixel separately, without reference to neighboring image pixels. Tilton said a more coherent approach is to divide the image into regions of similar pixels and label the regions instead. In his segmentation technique, Tilton first partially segments smaller sections of the image and then repeatedly recombines them until the whole image is assembled and segmented.

"The way this implementation divides up the problem is particularly effective on the HIVE," Tilton said. For example, the HIVE took 1.4 hours to segment a large (3,456-by-2,688-pixel) Landsat image. The same task required 1.8 hours on 512 processors of Goddard's CRAY T3E, a widely installed supercomputer. Tilton estimates his segmentation routine is 130 times more cost-effective on the HIVE as compared to the CRAY T3E.

Besides using processors with a much larger memory cache, another reason why the HIVE succeeds is that Tilton's algorithm requires "very little interprocessor communication." Minimizing communications is desirable when designing Beowulf applications because Ethernet networks are slower than the multiple-gigabit-per-second networks found in commercial supercomputers and thus delay passing data between processors.

So far, Tilton can label basic outlines of land and water and features like dense urban areas and major roads. He is now focusing on getting more specific in a "reliable and robust way," especially with higher-resolution and more complex satellite observations coming in the near future.

Observations also play a fundamental role in cosmology, providing a check against theories of the universe's evolution. Lately, cosmological models seem to spawn like rabbits. Michael Gross, a USRA physicist based at Goddard, has details on nearly 100 of them. He uses the HIVE to see how well the models match critical observational data.

A few hundred thousand years after the Big Bang, the universe cooled enough for neutral atoms to form. In the process, the sky emitted microwave background radiation that can be measured today. Gross examines measurements considered the "most constraining." NASA's Cosmic Background Explorer satellite looked at large scales. Various ground-based and balloon observations have captured parts of the sky around one degree across, "the distance light could travel when the microwave background was emitted," Gross said. Serving as a companion check is optical telescope readings of the three-dimensional distribution of galaxies in the local universe.

The models themselves vary in how much and what types of matter they include and in the speed at which the universe expands. Detector experiments carried out last year show that subatomic particles called neutrinos likely have mass, and some cosmological models make room for non-zero mass neutrinos in the universe's matter. Many models also favor a cosmological constant, an idea originating with Einstein for a repulsive force that accelerates the universe's expansion. With supernova observations pointing to the force's existence, the journal *Science* named the accelerating universe its "Breakthrough of the Year" for 1998.

Encompassing these variations, Gross studied 96 cosmological models on the HIVE. The results were presented in a workshop paper by University of California, Santa Cruz cosmologist Joel Primack and Gross. Their findings point to a role for neutrinos if the universe has more than 40 percent of the matter to eventually collapse itself (known as critical density). Gross explained that neutrinos make it slightly harder to form galaxy clusters, and compensating for this increases fluctuations in the degree-scale microwave background.

Overall, two models of those tested are the best fits. One is a cosmological constant universe with 50 percent of critical density and 10 percent of its matter as neutrinos. The other is a similar universe with 40 percent of critical density and only massless neutrinos.

Gross's computational scheme assigns one model per HIVE processor. "You take the same problem and run it on each processor with different parameters," he said. The 96-model study "would have taken three months on the DEC Alpha workstation I had been using," Gross said. "It ran in a little over a day on the HIVE. That makes it an enabling technology for these types of parameter studies."

Processing the cosmological observations themselves becomes more daunting as instruments increase in power and resolution. Starting operations in April is the European Southern Observatory's Very Large Telescope (VLT). It will combine four 8-meter mirrors, housed in separate structures in northern Chile, into an instrument that can gather far more light with greater resolution than the Hubble Space Telescope.

As the world's largest optical observatory, VLT will "be capable of data rates and data volumes in excess of any existing astronomical facility," said Peter Quinn, who heads the Observatory's Data Management and Operations Division. Each night the telescope could easily collect 100 gigabytes of observation data. To draw knowledge from this data pile, the Observatory is exploring Beowulf systems as dedicated data processing engines.

In most instances VLT data will enter an archive, where they will undergo a few levels of processing before being distributed to astronomers. First, "there are instrument effects you must remove to get the data from the sky that you want," Quinn said. "It is very intensive, a pixel-by-pixel operation." Current tests use frames from the 4-meter

New Technology Telescope (NTT); at 8,000 by 8,000 pixels, they are already large. In 2001, VLT will have a 32,000-by-32,000-pixel camera. Next, "you turn the raw information into physical numbers and start asking scientific questions of the images," he said. Quinn cited measuring the total energy given off by individual stars as one example.

NTT frame processing averages about 70 megaFLOPS per processor on Caltech's Naegling, a $300,000 Beowulf named for the hero's sword. This 160-processor system connects 200-MHz Pentium Pro and 300-MHz Pentium II chips with Fast Ethernet. Coordinated by Caltech's John Salmon, the Observatory has had steady access to 10 processors since January 1998. If additional performance scaling tests fare as well, plans call for installing a 16-Alpha-processor Beowulf at Observatory headquarters in Garching, Germany. The system would grow to 128 processors in 2000.

Beyond serving archival needs, Quinn envisions Beowulf as a platform for real-time processing on VLT. During an observing session, astronomers could get a quick reading and re-aim the telescope to get better results. Quinn is particularly excited about the potential of using such a capability with VLT's infrared camera. "To study objects very far away, we need to look at the infrared part of the spectrum," he said. "The problem is that everything around us is warm and glowing in the infrared, and you have to subtract this foreground signal from the sky."

"People haven't been able to do that in real-time because of the processing problem," Quinn lamented. "Between VLT and Beowulf we may have the opportunity to immediately see what the infrared data looks like and make decisions. We won't have to go home, study the data, and ready another experiment." Real-time decision-making would require a second Beowulf in Chile. "The advantage of Beowulf is that the machinery is relatively cheap, and you can afford to have multiple versions," Quinn said.

The Observatory's 128-processor blueprint is similar to the present-day Avalon system built by Mike Warren and colleagues at Los Alamos National Laboratory. This $313,000 Beowulf's 140 Alpha chips clock at 533 MHz and combine for 29.6 gigaFLOPS on a molecular dynamics simulation. Performing 48.6 gigaFLOPS on the Linpack benchmark makes Avalon the fastest Beowulf and ranks it 113th on the latest TOP500 list of world supercomputers. The Observatory machine would certainly be faster with a new generation of chips due out next year.

Those chips should surpass one gigaFLOPS performance, leading Sterling, now at Caltech, to the conclusion that "we are really on track for doing a $1 million teraFLOPS machine in three years." Such a computer would be possible with about 1,000 processors.

Systems of that size will require more innovative network topologies than those used in today's Beowulfs. Another challenge to building them is maintaining the reliability of system software. "Software coming from many different individuals may have bugs because it is not tested on everything," Dorband said. Since larger Beowulfs will accommodate more users, they also will need the administration tools that are routine with vendor-built supercomputers.

For instance, "in very large clusters, you have to expect a small number of processors to be down at any one time," said Phil Merkey, a CESDIS Beowulf-builder. "There really isn't support for monitoring that right now with commodity parts." Sterling listed other system software requirements: scheduling, check-pointing of codes, performance measurement and visualization, real-time debugging, accounting, and diagnostics.

Two new efforts address such software concerns. USRA formed Scyld Computing Services (Scyld is Beowulf's father) in January. This for-profit holding company provides commercial-level support for Beowulf installations, such as tracking down and fixing bugs and developing robust installation and monitoring tools. Becker, Scyld's chief technologist, wants to see Beowulfs become more consistent with each other to encourage third parties to develop applications software. Toward that end, the company will later contribute their software to the wider Beowulf community through such means as Red Hat Software's series of open-source "extreme.linux" CD-ROMs.

Developers worldwide are working on Beowulf administration software but without coordination. To pull these activities together and "make a common set of software that has been shaken out," Sterling and Daniel Savarese are spearheading Grendel (the name of the monster Beowulf slays in the epic). Caltech acts as a "surrogate user," running the software packages on Naegling and reporting findings back to developers. The modified tools will then be available on a Caltech World Wide Web site.

—Jarrett Cohen

More Info

More information about Beowulf and these applications is available from the following sources:

- *How to Build a Beowulf: A Guide to the Implementation and Application of PC Clusters,* Thomas L. Sterling, John Salmon, Donald J. Becker, and Daniel F. Savarese (MIT Press, March 1999)
- "Beowulf Project at CESDIS" — http://www.beowulf.org/
- "The Very Large Telescope Project" — http://www.eso.org/projects/vlt/
- "Grendel Software" — http://www.cacr.caltech.edu/beowulf/grendel/