

INSIDER

Newsletter for the Employees of Ames Laboratory ■ Volume 11, Number 7 ■ July/August 2000



The Octopus Finds Needles in the Haystack

Cluster computer sorts through mounds of genetics data

It's not impossible to find a needle in a haystack. If you have enough time — or enough arms — it can be done. That's more or less the reasoning behind a “cluster computer” designed by Ames Lab's Scalable Computing Laboratory and built by Iowa State University's zoology and genetics department. Called the Octopus, this machine has 18 computers that function in concert to do the work of a much costlier supercomputer.

In this case, the “haystack” is the huge amount of genetic data being generated by various genome discovery projects. The high-profile Human Genome Project, for instance, will ultimately result in data on the three billion chemical base pairs that make up human DNA. Researchers at Iowa State University, working with various colleges around the country, are dealing with similar amounts of data with their maize gene discovery project.

The “needles” in these haystacks are sequences of DNA, or genes, that living cells follow to make proteins. Those proteins, in turn, perform the elemental work of life.

Determining gene sequences and how they express themselves in genetic traits is difficult at best. One challenge is the sheer amount of data involved; just scanning through the human genome data at a rate of 100 bits per second would take close to a year.

Another complicating factor is *continued on page 2*



Brad Powers (left), an ISU senior in biology, and David Halstead, associate scientist in the SCL, are shown with the Octopus, an 18-node cluster computer they helped build for ISU's zoology and genetics department.



that perhaps only the Human Genome Project can afford the brute force method of direct sequence gathering. Even with the importance of maize — the class of crops responsible for more than 70 percent of the caloric value of the human diet worldwide — researchers must instead find more elegant, less costly ways to crack its genetic code. These methods involve the use of sophisticated computer algorithms acting on huge databases.

Bioinformatics to the rescue

The computational demands of genomics is what led ISU's Volker Brendel, an associate professor in zoology and genetics, to instigate the Octopus project. Brendel is a member of the maize gene discovery team and an expert in the nascent field of "bioinformatics." This relatively new discipline taps tools and techniques from biology as well as mathematics and computer science. It relies on sophisticated statistical techniques and efficient software algorithms to interpret genetic data, leading to a better understanding of biological systems.

Like most anything related to information technology, bioinformatics research can be accelerated by the use of faster computers. To be sure, high-performance supercomputers have the computational horsepower and memory needed to handle gene data with ease. But the cost of these systems starts somewhere north of several hundred thousand dollars. So instead, Brendel considered a cluster computer to be built using run-of-the-mill — and therefore low-cost — personal computer components. Soon, Brendel was in touch with the experts at the Scalable Computing Laboratory.

Cluster computing is one of

SCL's specialties. SCL scientists have built or designed some eight clusters for Ames Laboratory and ISU researchers. The SCL also runs two of its own cluster computers, including a \$665,000 cluster of dual processor IBM Power3 workstations made possible via a Shared University Research grant from IBM. The SCL team has even written a cluster "cookbook" that provides a how-to on constructing PC clusters (available on the Web at www.scl.ameslab.gov/Projects/ClusterCookbook/).

Still, the design of a cluster is a shade more difficult than cooking

*"The true research aspect that we help other groups with is how to get their code to run in parallel."
—David Halstead*

dinner. Cluster computers achieve high speeds by dividing the work of the software and running various sections simultaneously on individual processors called "nodes." As such, various components — including processor speed, memory and hard disk space — need to be "balanced" for a particular application to achieve optimal speeds. For the Octopus, SCL researchers ran benchmarks of their own design to determine the most appropriate characteristics for the system.

The lab also provides ongoing assistance to Brendel's group to help write software that can take advantage of the cluster. "Sticking these machines in a room, turning them on, and so on, is arduous but doable. The true research aspect that we help other groups with is how to get their code to run in parallel," says David Halstead, SCL associate scientist who is sometimes called

"Clusterman" around the Lab for his expertise in the field.

Plain vanilla, but fast

Each node of the Octopus consists of a 450 MHz Intel Pentium II microprocessor, 512 megabytes of memory and a nine-gigabyte hard disk. By itself, each would make a fairly powerful desktop computer. Yet, the parts are "plain vanilla," says Halstead. The use of these standard parts translates into low cost, he explains, adding that each node can also have a life after the cluster as a desktop system for word processing and other routine computing tasks.

As for the software, "Brendel's algorithms are ideal applications for cluster computing," explains Halstead. "The problem is fairly linear," which means that adding more processors leads directly to faster processing. The \$45,000 Octopus has a theoretical capacity of roughly 3.2 billion floating point operations per second, which makes it about half as powerful as the \$300,000 ALICE cluster operated by SCL in the Ames Lab. Brendel is also planning to follow SCL's suggestion to add a "terminal room" with desktop systems that can be used either by students as regular computers or as additional nodes for Octopus.

The whole situation "couldn't have worked out better," says Brendel of the design and building of the Octopus, which essentially took about three months. The machine is now up and running in ISU's Molecular Biology Building. "We are still in the infancy of using the machine. Basically, we are using it to speed up sequential code for gene finding and annotation," he says.

Thanks to their low cost, cluster computers are becoming more common in universities and research institutes; in fact, Ames Lab and ISU are home to at least

nine such systems. But most clusters are made for physics, engineering, weather forecasting or just to learn more about cluster computing. The Octopus, in contrast, is probably one of the few built specifically for bioinformatics, according to Brendel. What's more, it's located in the same building where computational biologists do their research. "It lets me stay close to my data," he says. ■

~ Robert Mills