Shared computing grid cuts data mountains down to size

May 10 2005

Although University of Wisconsin-Madison professors Wesley Smith and David Schwartz operate in completely different scientific spheres - one seeking to explore the fundamental properties of matter and the other trying to wrest free the secrets of the human genome - both have the same dilemma: They are awash in a sea of data.

To make sense of the human genome, for example, Schwartz and the small army of scientists engaged in one of biology's grandest projects must sort through 20,000-25,000 genes and the hundreds of millions of base pairs - long, contiguous sequences of DNA that are the genes' biochemical memory.

Such tasks, Schwartz notes, are computationally intense. With a handful of computers, analysis of one small portion of the genome might take a year. But now, thanks to a visionary computing initiative called Grid Laboratory of Wisconsin (GLOW), Schwartz can whip through daunting sequences of DNA like nobody's business.

"The work we're doing wouldn't be possible without GLOW," says Schwartz, a UW-Madison professor of genetics and chemistry. "It's been catalytic for our research. What might take a year with a couple of computers can now be done in a day."

GLOW is a campus-wide distributed computing environment in which hundreds of individual personal computer-sized processors work in concert to sort through the massive data sets acquired by people such as

Schwartz; or to power the simulations that Smith, a UW-Madison professor of physics, uses to presage experiments planned for the high-energy particle accelerators that provide a deep understanding of matter.

"The roots of GLOW are very deep," says Miron Livny, a UW-Madison professor of computer science who, during the past 20 years, has devised a computing template known as Condor that, like the animal it is named for, is a scavenger. It gathers all available processing power from hundreds of pooled GLOW computers around campus and directs those unused cycles to the service of science being done by Smith, Schwartz and a half dozen other programs from chemical engineering to cancer research.

The power of GLOW is derived from harnessing the available cycles of many small computers, which almost never use their full computing capacity. Participating teams contribute processors and, with the help of $1.2 million from the National Science Foundation (NSF) and a $500,000 match from UW-Madison, the project has been able to assemble additional racks of processors that are managed by the individual projects, but which are available to the GLOW pool when not in local use.

By some estimates, most computers are used to only about 30 percent of their capacity. GLOW, using Condor, takes advantage of this spare capacity by dividing large research tasks into small ones and sending those tasks over a network to idle computers to work on. GLOW, in effect, has transformed the UW-Madison campus into a laboratory to study the power and issues associated with distributed computing in an intensive research environment, says Livny.

"Each group has full control over its own resources," says the UW-Madison computer scientist, "but when they are not being used locally, they must be available for use by the GLOW community."

GLOW, according to Livny, is built so that very different applications might be applied. For example, data collected by the IceCube Neutrino Telescope at the South Pole is processed very differently from data that might inform UW-Madison cancer researchers looking for a new drug.

"The applications are very different, so we try not to see the application," says Livny. "If we deal with the specifics of each of these fields, we're doomed. We try to be as generic as possible."

Despite that guiding philosophy, Livny and his group in computer science work closely with the various groups participating in GLOW to adapt their application so they can harness the power of GLOW and other Condor resources.

One of the challenges for GLOW is convincing faculty that participation can be of benefit to their projects. Even with hundreds of processors deployed in the collective interest, and with more being added all the time, there is still a competition for a finite resource.

It is not always easy, Livny says, to transform a project that is built around dedicated computing resources into one utilizing the power of a large pool of opportunistic resources.

"These things don't happen overnight," Livny acknowledges. "Part of the challenge is to demonstrate to the scientists that they will get value out of it."

But to Smith the advantages of obtaining access to a fire-breathing computational resource are obvious. "In my field, many researchers now think within a 'box of calculational limitations.' The ability to think outside that box means we can do things completely differently."

Source: University of Wisconsin-Madison