copyright 2002 Alan Radding

Independent Assessment

by Alan Radding

April 26, 2002

Bioinformatics Jumps on the Grid

Which Industry Goes Next?

The grid computing bandwagon is starting to roll, and bioinformatics and life sciences companies are among the first to jump aboard. Although grid computing appears to be an excellent fit for bioinformatics companies--small players with little more than desktop computer systems but huge computer processing needs--grid computing should attract companies outside of bioinformatics.

Grid computing, which describes an approach to computing that aggregates the processing power of hundreds, thousands, even millions of individual systems to handle compute-intensive tasks, has been kicking around for a decade or more under various names. Initially it was described as the meta-computer--thousands of discrete systems loosely coupled through a network contributing spare cycles to big processing tasks. More recently, it has been embedded in the concept of the compute utility, which enables companies to tap into seemingly infinite amounts of processing on demand. The utility provider establishes a massive processing grid and makes processing cycles available to clients as needed at a metered rate.

SETI (Search for Extraterrestrial Intelligence) is the best-known example of grid computing. SETI involves sifting through massive volumes of radio data collected from space for evidence of intelligence life. Such a task, however, requires unfathomable amounts of processing power on a continuous basis. To get that processing power at a low cost (essentially free), SETI researchers borrow unused processing cycles from millions of home computers. In SETI@Home, about three million users to date have voluntarily loaded a piece of

copyright 2002 Alan Radding

code in the form of a screen saver to their Internet-connected PC. Instead of spinning images of flying toasters when the PC is idle, the SETI software downloads a small chunk of raw data, processes it, and reports the results back to the central organization. SETI, in effect, is grid computing on a cosmic scale.

Compute grids today come in three flavors: public like SETI, shared like a compute utility, and private. The bioinformatics industry is generally opting for private computing grids to harness the power of a company's installed base of systems and avoid the need to buy a supercomputer.

Infinity Pharmaceuticals, for example, uses a local grid of 100 PCs to handle a small set of compute-intensive applications. "These applications are very spiky and might suddenly need thousands of hours of processing to solve a problem," CIO Andy Palmer told me recently. By the end of the year, he expects his grid to reach 1000 PCs.

For Infinity, grid computing makes sense. "We're only a small company, but we do have all these systems sitting around. Grid computing gives us a way to utilize this power more efficiently," Palmer continued.

Structural Bioinformatics already had large-scale supercomputer power through an SGI system and a 120-processor IBM Linux cluster when it turned to grid computing. The supercomputer and cluster is used to perform the massive calculations required to analyze proteins. One calculation alone can take 14 hours. That leaves the company few processing cycles for other types of analysis and modeling CFO Ralph Barry told me. So, the company turned to grid computing to harness the spare processing cycles sitting on desktops throughout the company. "Grid computing gives us a way to get more processing without buying more hardware," Barry noted.

The Achilles heel, the weak spot, of grid computing has always been the software. Conventional monolithic applications with their massive amounts of code riddled with critical interdependencies didn't lend themselves to being broken up and run in small chunks in parallel on many machines. In addition, setting up and managing the sprawling grids proved to be a formidable task.

copyright 2002 Alan Radding

Two developments, however, have come together to make grid computing practical for the bioinformatics industry and, I believe, for many other industries: 1) the advent of highly modular applications consisting of loosely coupled components along with the emergence of Web services and 2) the arrival of vendors like Avaki Corp. and Platform Computing to provide ready-to-run software and systems that simplify the task of setting up a grid and managing the operation. Both Infinity Pharmaceuticals and Structural Bioinformatics adopted the Avaki grid solution.

The European Bioinformatics Institute (EBI), a non-profit research organization, selected Platform Computing Inc.'s LSF MultiCluster for its grid solution. LSF MultiCluster is an intelligent workload management solution that EBI uses to manage a workload across a 140-processor multi-platform server farm, which includes Compaq AlphaServers, SGI servers and 40 Linux PCs running EBI's 200-gigabyte database of flat files on protein and genomic sequencing

Other vendors too are rushing into grid computing. Sun Microsystems provides Sun Grid Engine 5.2.3 software that manages and submits jobs to specified machines within a cluster. Sun Grid Engine is intended specifically for the processing of compute intensive tasks and maximizing the available power of the local network. A recent alliance with Avaki should help Sun extend the scope of the Sun Grid Engine.

IBM has set up a Grid Innovation Centre and is partnering with Avaki to develop both compute and data grids. IBM also is teaming with Platform, allowing it to demonstrate its grid systems at the Grid Innovation Centre.

Microsoft has identified grids as an ideal platform for its .NET initiative. It reportedly has joined the Globus Project, a joint grid computing research effort by the U.S. Department of Energy's Argonne National Laboratory and the University of Southern California's Information Sciences Institute.

copyright 2002 Alan Radding

Although the bioinformatics and life sciences industry are early adopters of grid computing, one can easily envision grids being used throughout corporate computing for business intelligence, data mining, risk management, and more. The financial services industry, for instance, represents an ideal candidate for compute grids. Banks and investment houses already deploy thousands or tens of thousands of overpowered Pentium-based PCs, most used for little more than email and word processing. By tying these systems together in a grid, investment firms could tap their spare processing power to run, say, Monte Carlo simulations, a compute-intensive risk management analytical application.

You can expect the big systems platform vendors to add support for grid computing one way or another. Software vendors should begin thinking about how they can architect, modularize, componentize, or otherwise optimize their software to run on grids. Before long, grids will be popping up everywhere.