Supercharging Proteomics Discovery

By Jim King, Contributing Editor, *Bio·IT World*

Produced by Cambridge Healthtech
Media Group Custom Publishing

sgi®

www.sgi.com

Supercharging Proteomics Discovery

The explosion in genomics research has been a boon to biomedical research and drug discovery, but genetic sequencing and expression data is only the first step in the biological cascade. It is translation of the genome into proteins that drives the true nature of biological systems, and that falls under the purview of proteomics – the study of the proteins and metabolites, and their interactions.

Proteomics is a burgeoning field that has become central to drug discovery, diagnostics, systems biology, and synthetic biology, among many other applications. It has tremendous potential as a vehicle for the discovery of biomarkers that could provide greater accuracy in a patient's prognosis, or guide clinical trial investigators as they attempt to determine a drug's safety and efficacy. Such studies could ultimately usher in the long-awaited 'personalized medicine,' which would use diagnostics to maximize the curative power of medications by allowing physicians to prescribe specific drugs for subsets of patients with a corresponding genetic profile. Protein-protein interaction studies can assist in the design of new drugs and the identification of novel drug targets. Proteomics also has potential as a diagnostic tool to assist in the diagnosis and prognosis of diseases, such as distinguishing early-stage liver cancer in patients with hepatitis C liver cirrhosis.

The study of proteins and their interactions can reveal the nature of biological pathways. Once elucidated, these mechanisms could be used, for example, in the creation of genetically modified bacteria capable of cleaning up oil spills, or producing ethanol from cellulosic plant feed stocks.

But to fulfill its promise, proteomics requires intense computational power. Industry, government, and academic laboratories are generating massive data sets that promise to unlock many biological secrets, but this boon comes at a price: mountains of that data can swiftly overwhelm any computational environment, slowing analysis to a standstill.

Many of today's research problems rely on huge volumes of data, the analysis of which can lead to computing bottlenecks. Systems biology approaches are even more complicated because researchers want to drill down to the minutest level of details, even as they maintain cross-referencing to investigate relationships between proteins. These computational problems become large enough that the traditional approach — assembling more CPU power in an ever-expanding cluster — begins to generate limiting returns. Numerous CPUs sending and receiving large amounts of data create an I/O bottleneck. Those processors also eat up power and space, and management of such clusters become unwieldy.

Computational limits can also prevent researchers from asking really big questions. In an ideal world, researchers would like to incorporate many lines of inquiry into a single series of experiments. For example, an experiment to investigate the genetic basis of mental retardation in fruit flies — pursued as a model of Fragile X syndrome and other forms of human mental retardation — might also include an analysis of the fly metabolites (its metabolome) as well as its proteins. Determining associations and performing cross-correlations, perhaps across multiple species of fly, would quickly overwhelm any traditional computing cluster.

Typically, researchers respond to such limitations by simplifying the experiment so that computational resources can handle the burden, but that limits the power of the experiment and analysis. The kind of in-depth analysis that can really push forward the boundaries of science requires a totally different approach, such as marrying traditional clusters with Field Programmable Gate Array (FPGA) devices that can eliminate computational bottlenecks.

THE PROTEOMICS APPLIANCE

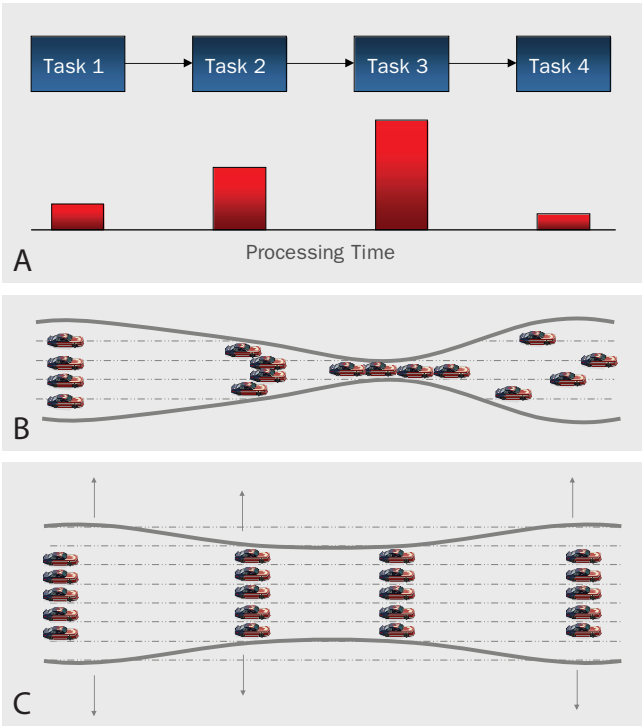
To respond to these needs, SGI has developed the Proteomics Appliance. It is based on the SGI® RASC™ (Reconfigurable Application Specific Computing) system using Intel® Itanium® and Quad-Core Intel ® Xeon ® Processors. SGI's technology partner, Singapore-based Progeniq, has already loaded the Proteomics Appliance with the proteomics applications Smith-Waterman and ClustalW, and it is expected that by the end of March it will also be loaded with HMMer. BLASTp will follow soon after.

Preliminary test runs with the SGI Proteomics Appliance show a 20 to 30-fold speedup over a 2.4GHz Intel Core 2 Quad Q6600 when using Smith-Waterman to search the SWISS-PROT database with query sequences ranging from 500 to 1,000 amino acids. The performance of Clustal W shows an approximate 7-fold speedup over the same 2.4GHz Intel Core 2 Quad Q6600 when aligning a set of 124 sequences.

These already impressive boosts will be improved further as the researchers at SGI and Progeniq further optimize the system. Combined with the SGI® Altix® system, the Proteomics Appliance can run analyses as much as hundreds of times faster than traditional solutions, with only a minimal increase in power consumption. The Appliance's reconfigurable hardware processor can 'rewire' itself on the fly according to whichever application needs to be accelerated.

That capability makes the Proteomics Appliance well-suited to a common bottleneck that afflicts proteomics analyses. Such bottlenecks occur because proteomics analyses often incorporate several applications, and frequently one of them demands the lion's share — 90-95% — of processing time. The Proteomics Appliance is adept at easing the strain of computational bottlenecks, thus dramatically accelerating the entire analysis. A typical computational job consists of a series of tasks in a workflow. This workflow can be likened to a multi-lane road, where the number of lanes at the position of an individual task represents the amount of computational time it demands. The most computationally intensive tasks are like a multi-lane highway that narrows temporarily to a single lane, causing a bottleneck in the road and resulting in a traffic jam.

The Proteomics Appliance can be assigned to the most intensive task ('Task 3' in Fig. A), accelerating it by a factor of 50, for example. The effect of this is to change the shape of the road. The bottleneck at Task 3 (Fig. B) is now alleviated, and the proportion of processing time redistributed amongst Tasks 1, 2, and 4 (Fig. C). Parallelizing Tasks 1, 2, and 4 over multiple processors or cores



can widen the road at these points, giving the maximum throughput for the entire workflow and dramatically reducing the overall analysis time.

This improved performance means a significant decrease in the time required for analysis. An experiment that might take months to complete on a traditional system could be performed in days or even hours on the SGI Proteomics Appliance.

Increased overall analysis speed isn't the only advantage that the Proteomics Appliance can offer. It can also give a researcher greater flexibility. Often, the optimal choice for a research application is not the fastest, and limited processing capacity forces researchers to choose another application, sacrificing optimal performance for speed. The Proteomics Appliance's dramatic increase in performance eliminates the need to make this sort of compromise, freeing the researcher to choose the right software package for the project with no worries about overwhelming computational resources.

TIME TO FAST FORWARD: BRINGING NEW TECHNOLOGY TO THE FOREFRONT

Over the years, there have been a number of attempts based on a variety of technologies to speed up the most commonly used life science applications. Notable among the earlier efforts were Paracel's special high performance systems that used Application-Specific Integrated Circuits (ASICs) to accelerate a number of programs and Timelogic's FPGA accelerator boards that helped off-load computationally-intensive tasks from a CPU.

SGI has developed a next generation FPGA-based solution for the life sciences that overcomes many of the limitations of earlier systems. The SGI Proteomics Appliance provides the ideal mix of high performance, power efficiency, and simplified system management. Unlike traditional processors, which are serial in nature, FPGAs are inherently parallel, allowing multiple functions to be performed simultaneously.

The Proteomics Appliance is built on the SGI RASC platform, which combines the high-performance SGI Altix architecture with leading-edge FPGA technology to deliver a complete platform for accelerating workloads. This solution can be used for a mix of CPU-based and FPGA-accelerated applications in a number of domains, or it can be optimally configured to run one set of applications within a specific domain.

From a systems perspective, the SGI RASC solution leverages numerous SGI Altix features like its scalable, high-bandwidth shared memory system architecture. This design enables systems to be configured with virtually any mixture of CPUs, FPGAs, memory and I/O so the configuration can be adapted to the needs of a specific customer. Single systems with over 5 TFLOPS of performance, 100 TB of memory and 10's of GBytes/sec of I/O could be created.

The SGI RASC platform is also optimized for high-performance. First, it integrates two of today's largest FPGAs (the Xilinx Virtex 4) onto a single board. And because a single RASC blade with two FPGAs has two 6.4GByte/sec direct connections to the shared memory infrastructure, applications using the FPGAs can access enormous amounts of memory much faster than on other system designs. The combination of state-of-the-art FPGA performance and scalable high-performance system architecture means that developers and users are able to achieve sizable real-life performance improvements.

POWER AND SPACE EFFICIENCY

Increasingly, the power needed to run HPC cluster and to cool data centers is becoming a real issue. Over the course of a year, the power consumption can run several hundred thousand dollars — more than the servers themselves. That represents a real financial burden. It is also an environmental concern in an era when green computing initiatives have gained increased attention.

In addition to pure acceleration of research results, the SGI RASC solution offers a significant power savings over a high-throughput cluster. These savings come in three ways.

First, FPGA devices consume about ten times less power than a typical system. The SGI RASC architecture delivers over 75% of input power to the computational system, compared to about 50%

for 'pizza box' servers, and it means that from the start, the Proteomics Appliance consumes about 1/3 less energy.

Second, FPGA-based solutions are significantly faster than standard CPUs running proteomics applications, so a system need only run at peak performance for a shorter period of time.

Third, by consuming significantly less power to begin with, the system generates less heat than standard computing systems and as a result, less energy needs to be expended to cool the data center.

If you combine these savings together and factor in the added processing power of the FPGA + CPU as compared to just a CPU, you end up with an overall energy usage for an FPGA solution that is much less than an equivalent CPU-based only solution.

Naturally, using less power cuts electric bills and is good for the environment, but the benefits don't stop there. Added to an existing setup, the Proteomics Appliance is much denser than a standard high-throughput cluster, taking up significantly less rack space.

When you consider that industry surveys indicate 96% of computer rooms will run out of capacity within the next 5 years, denser systems that consume less power and are therefore easier to cool can clearly have a big impact on the bottom line.

SGI EXTENDS A RICH HISTORY OF ACCELERATING SCIENTIFIC RESEARCH

SGI has long been an enabler of scientific research. Early Challenge and Power Challenge systems brought high performance technical computing to the mass market. For example, in 1996, a 64-CPU Power Challenge Array was used by SGI and EMBL to analyze more than 6000 protein sequences from the genome of yeast (*Saccharomyces cerevisiae*).

SGI extended that mid-range leadership to the high-end by developing NUMAflex® technology which enabled it to grow shared memory systems to up to 1,024 processors with 100s of GBytes/second of memory bandwidth.

As Linux® became more commonly used in the scientific community, and particularly in the life sciences community, SGI became the first major vendor dedicated to providing open source solutions. SGI is currently shipping SGI Altix, its fourth generation of scalable shared memory system based on NUMAflex . Early life science adopters of the SGI Altix included the National Cancer Institute, the University of Arizona the Memorial Sloan-Kettering Cancer Center, the Center for Biological Sequence Analysis (CBS) of the Technical University of Denmark, Beijing-Normal University, Merck, and the Wellcome Trust Sanger Institute.

In addition to its work on the systems side, SGI has a rich history of partnerships with mathematical analysis and database vendors working with them in the early days of Linux adoption to ensure scientists and researchers could get optimized performance when running their applications. Within the life sciences community, SGI worked with a wide variety of partners such as Gaussian, Schrodinger, SCM, and a number of open-source application providers to maximize performance on SGI systems. ■