

March 1, 2008 | Bio-IT World > Add-On Speed for Bioinformatics

Add-On Speed for Bioinformatics

By Mike May

March 1, 2008 | Today's pharmaceutical simulations eat up computer run time faster than CPU-makers can churn out more-powerful chips. "Some calculations performed in molecular dynamics, for example, can easily take a month, and many problems are not even feasible," says Nathan Woods, chief scientist at XtremeData in Schaumburg, Illinois. As a result, researchers and hardware designers often opt to speed up simulations with hardware accelerators instead of adding CPUs.

Even without taking on new territory, such as molecule-level simulations, researchers in bioinformatics still face a data-CPU divide. "NIH's GenBank is approximately doubling in size year after year," says Teck Hiong Chua, vice president of business development at Progeniq in Singapore. Consequently, researchers need more computing power to search through those data. "The 400 next-generation sequencing instruments already deployed have the capacity to generate 41 terabases in a year," says Martin Gollery, senior bioinformatics scientist for Active Motif in Carlsbad, California. "That's over 550 times more data than is currently stored in GenBank." So even if Moore's Law stayed on pace, the growth in data would easily overwhelm hardware advances. This imbalance, though, gives add-on accelerators an edge.

Instead of running algorithms through programs that get interpreted and carried out by a CPU, developers essentially hard wire algorithms into the add-on accelerators. These devices work like a chip that was designed to do a specific task. As a result, they run faster than a CPU for many algorithms. Progeniq, for example, developed its BioBoost 4.0 around USB 2.0. "It can fit into an existing PCI slot or sit on the desk plugged into a workstation or laptop," says Chua. When running Progeniq's recent Hidden Markov Model (HMM) sequence analysis on the BioBoost 4.0, says Chua, "this gives an average speed up of 25 to 85 times — depending on the sequence length and database — versus a normal workstation." He adds that Progeniq will soon release applications for Smith-Waterman and ClustalW algorithms. The base model of BioBoost 4.0, to just run say the HMM, costs about \$7,500. To get the BioBoost 4.0 with a complete suite of applications, a researcher will pay an additional \$5,000–10,000

Adding Algorithms

CLC bio in Aarhus, Denmark, makes a variety of accelerators for bioinformatics, including its Bioinformatics Cube. This FPGA (field-programmable gate array)-based device can be connected to a computer via a USB port, and works with various operating systems, including Linux, Mac OS X, and Windows. "It can be programmed with any of our workbenches or command lines," says Bjarne Knudsen, CLC bio's chief scientific officer. He adds that the Cube can speed up Smith-Waterman algorithms by as much as 100 times. The Cube also runs a variety of BLAST algorithms. "We are continually developing new algorithms," Knudsen says, such as the forthcoming HMM algorithm. Depending on the algorithms desired, the Cube costs about \$10,000. In addition, CLC bio will build custom Cube solutions when requested.

Meanwhile, other companies keep extending their add-on options. In March 2008, for example, Active Motif's TimeLogic division releases its NextEngine accelerator for Linux-based servers. Each of these FPGA-based PCIe cards can process HMM comparisons at the speed of 550 2.6-gigahertz Xeon CPU cores, according to Christopher Hoover, TimeLogic marketing manager. Even previous Active Motif accelerators attracted attention. For example, Terry Gaasterland, director of the Scripps Genome Center at the Scripps Institution of Oceanography of the University of California, San Diego, says, "I have four TimeLogic boards and the associated software to perform DNA- and protein-sequence alignment as well as software for protein-pattern searching by HMM." Gaasterland describes their performance as simply, "Excellent."

"Many researchers new to sequence analysis need to analyze Solexa, 454, or SOLiD data," Hoover says. "These groups are finding out that mapping shorter reads accurately to reference genomes can pose significant computational challenges." Consequently, some research groups might want more than just an accelerator. They might prefer a complete computing package, such as the Code-Quest Biocomputing Workstation, which accelerates BLAST, Smith-Waterman, HMM analysis, and gene modeling. The CodeQuest can also simplify the development of analyses. It includes a drag-and-drop interface called PipeWorks that lets users stitch together search tools and filters. "Researchers that wouldn't attempt to code workflows in Perl can annotate genomic data without interfacing with a bioinformatics expert," says Hoover. Even at a cost of \$25,000, he says that CodeQuest can pay for itself in 1–2 years by offsetting the power, cooling, and server-room rack-space costs required by clusters.

Researchers who prefer to crank up the capabilities of their current computers might choose XtremeData's in-socket accelerators. These plug into extra-CPU slots on a computer's motherboard. "By plugging into a socket, our device is a peer with a CPU," says Woods, "and that opens up a host of acceleration benefits." This FPGA-based device serves a very specific need in a highly parallel way. For example, XtremeData's most recent XD2000i is designed as an accelerator for Intel's Bensley Enterprise Platform that runs on the Xeon processor. "Our device sits on the same bus as the processor," says Woods, "and that gives us a large bandwidth to main



memory.” He adds that this is crucial for speeding up calculations on large data sets, such as those used in molecular dynamics or to sequence genes.

The acceleration provided by the XD2000i depends on the application. If a program includes a loop of integer calculations — common in genomics — then “the XD2000i could speed up those operations by 10 to 100 times,” says Woods. “We have direct experience on computer tomography, and there one FPGA replaces about 10 CPUs on integer calculations.”

Hardware accelerators already speed up research in many fields, including agricultural genomics. “Handling comparative assembly techniques for large genomes — plus the downstream annotation required to understand all the genes identified — places a huge demand on IT infrastructure,” Gollery says. “Forty percent of the TimeLogic customers are engaged in plant genomics because it addresses these burdens.”