A Reconfigurable Computing Model for Biological Research Application of Smith-Waterman Analysis to Bacterial Genomes

High-throughput technologies in the field of biology have led to an exponential growth in the amount of data generated over the past several years. This is witnessed in the sequencing of the genomes of mouse, rat, rice, Arabidopsis, human and many bacteria and viruses. Ever increasing computational resources are required to annotate and explore the resulting sequences and to refine our current models of understanding. This data explosion is forcing computational biologists to search for innovative computational designs to meet the growing demands in the field of biology.

Computer power has also been growing at an exponential pace. Even with this growth, however, the requirements to process the biological data far outstrips the ability of traditional computing to meet the challenge of converting the data into information or knowledge. Simply annotating and updating current databases has become a task that occupies large fractions of many groups' computational resources in order to complete their task before the next release of the data. Therefore, computational analyses must be designed to exploit as much concurrency in their models and computational systems as possible. Fortunately, many problems in computational biology are inherently parallel, and benefit from concurrent computing models. Projects such as whole genome alignments or comparisons are perfectly suited to take advantage of massively scalable compute engines such as FPGAs.

Over the past several years, key computational biology algorithms such as the Smith-Waterman and Hidden Markov computations have been implemented on FPGAs, and have enabled many computational analyses that were previously impractical. However, the complexity of programming the FPGAs and the inability to scale a single task across multiple logic boards with multiple gigabytes of memory have severely hampered the wide-spread use of this technology.

Star Bridge Systems has developed a reconfigurable computing system using FPGAs that can deliver 10X to 100X or greater improvement in computational efficiency (compared to traditional RISC processor based machines) for many problems by tailoring hardware allocations to match the needs of applications. Built from commodity components, FPGAs, the Hypercomputer provides benefits usually associated with expensive high performance computer technology. A high level of programmability offers the end user the ability to dynamically change the make-up and organization of the entire compute substrate – computational elements, communications topology and memory – to optimize the system for the problem at hand through the use of Starbridge System's Viva® software. The enhanced programmability of dynamically reconfigurable supercomputers running with Viva can improve time-to-solution by significantly expediting the process of developing, modifying and validating software.

Starbridge has developed a patented FPGA architecture that combines any number from two to hundreds of FPGAs to achieve optimum hardware performance. It is fractal-like in design, i.e. it follows a pattern in which the structure of each level of hardware resources is repeated at the next higher level. Up to eleven FPGAs are arrayed on proprietary PCBs, which in turn may be expanded to multiple boards operating in a PCI-X (or other) communications environment.

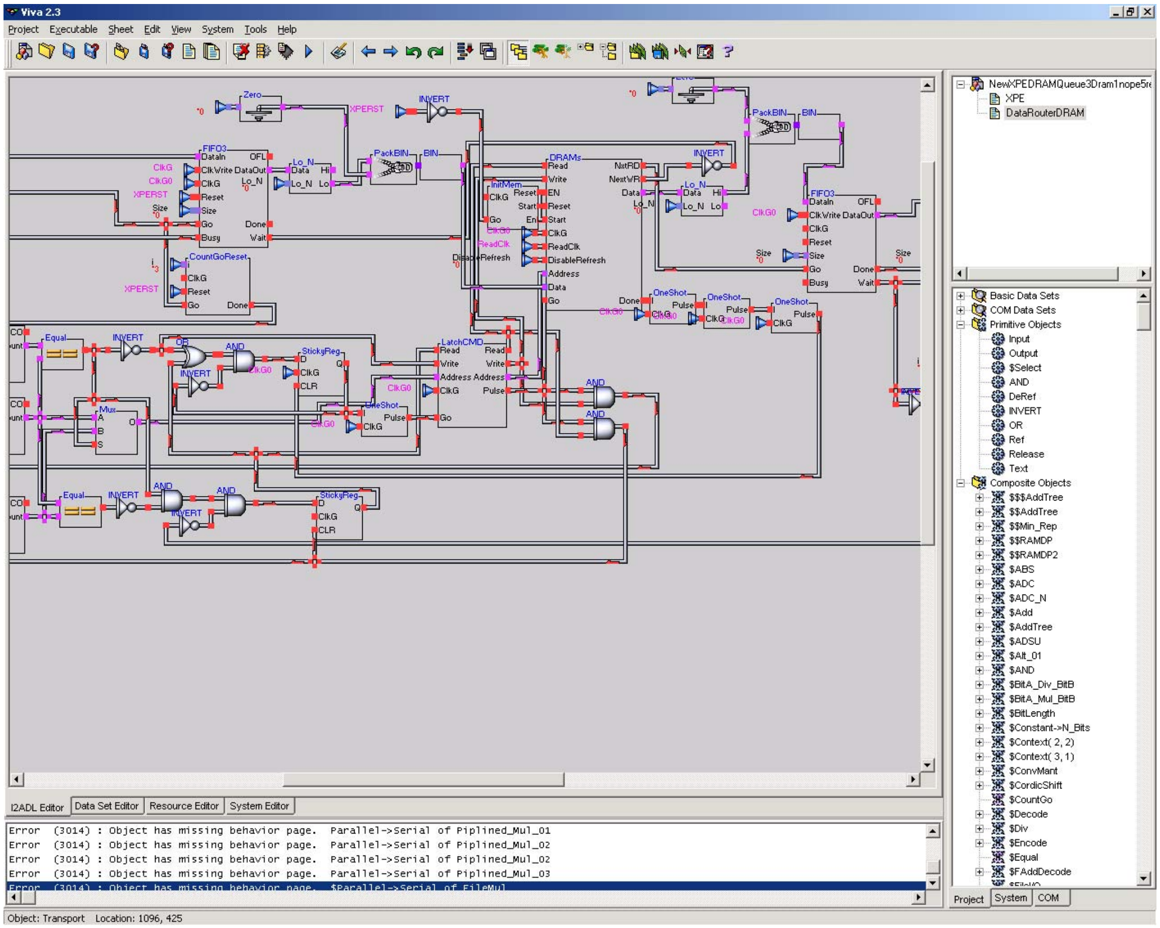
One system employs a dual processor motherboard and a single Hypercomputer board with nine Xilinx XC2V6000-BG1152 Virtex-II FPGAs and two XC2V4000-BG1152 Virtex-II FPGAs, yielding approximately 62 million gates per board. Three FPGAs function as the PCI-X bus interface, cross-point switch interface and router interface. Each of the remaining eight FPGAs, along with the cross-point switch, is equipped with four parallel memory channels interfaced to .5 gigabytes of double data rate DRAM per channel, i.e. 2 gigabytes of RAM per FPGA. This eleven-chip system features 18 gigabytes of RAM, configured with 36 64-bit parallel memory channels, yielding aggregate memory bandwidth of greater than 50 gigabytes per second per board. FPGAs are organized in groups of four with 50 I/O connections between each FPGA and the other three FPGAs in the quad. A total of 560 external I/O lines are available for communicating with other boards or digital systems. In a larger system boards are also organized in groups of four with one interconnect board in a similar fractal-like hierarchy.

With this architecture the following communication bandwidths are achievable:

Bus controller to quads:	20 GB/s	I/O board to board:	22.5 GB/s
Cross-point switch to quads:	10 GB/s	FPGA to FPGA:	7.5 GB/s
Router to quads:	30 GB/s	FPGA to FPGA inter-quad:	10.5 GB/s
Inter-quad:	12 GB/s	Aggregate bandwidth:	84 GB/s

Viva Description

Starbridge's Viva software runs on a traditional CPU and includes a high-level graphical algorithm description language. It emphasizes high-level control flow and reusable library function calls. The user selects library functions to perform calculations, then interconnects these into a control-flow graph via a "point-and-click" interface. (Figure 1)A synthesis tool develops net lists directly from the high level language, and placement and routing tools place the net lists directly into the reconfigurable Hypercomputer. In this manner, Viva instantiates algorithmic block diagram descriptions of desired behavior directly into system level hardware configurations. Computational, communications and memory objects may all be created by this basic process. The objects created may then be executed in whatever system was previously described as the implementation environment. The result is a set of algorithm design tools for users who lack the time or expertise to create their own application specific circuits using traditional design software.



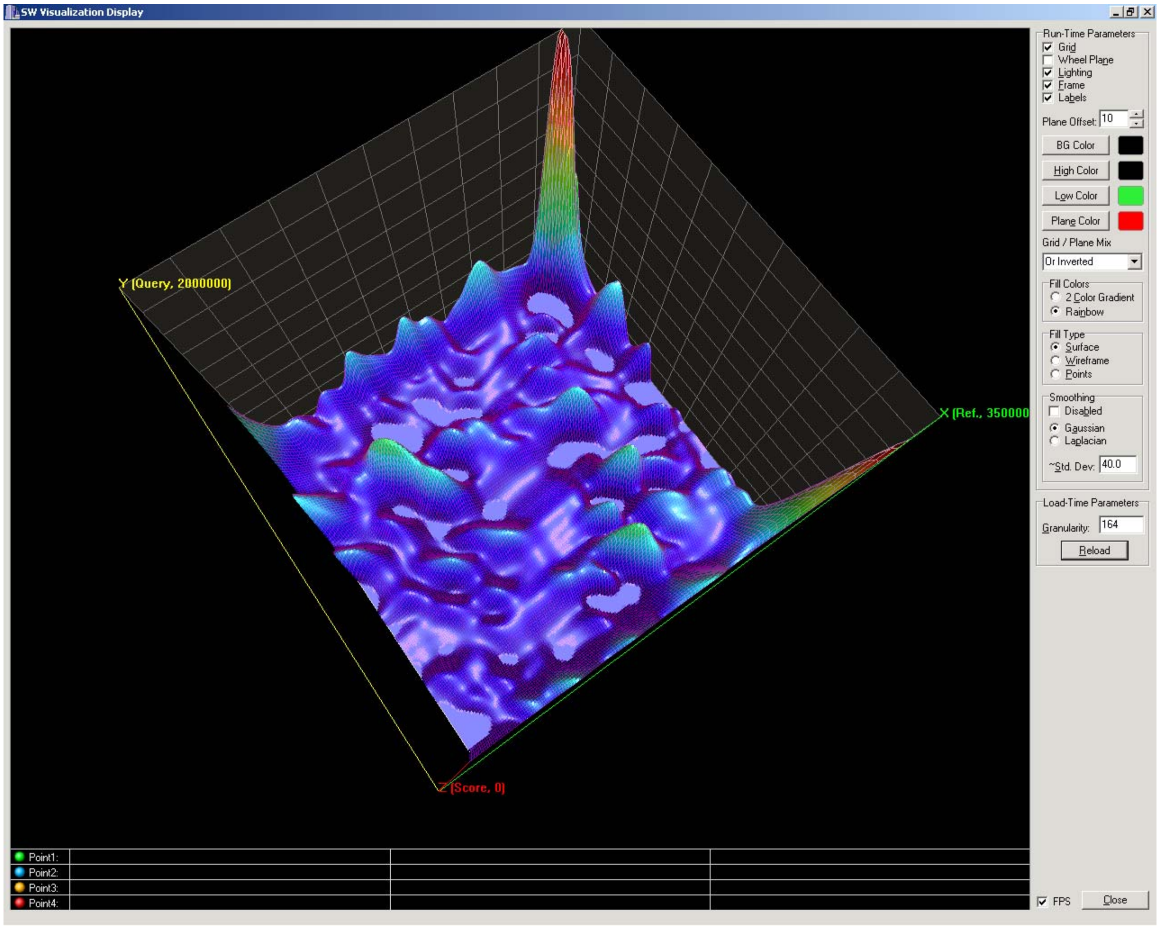
```
Error  (3014) : Object has missing behavior page.  Parallel->Serial of Piplined_Mul_01
Error  (3014) : Object has missing behavior page.  Parallel->Serial of Piplined_Mul_02
Error  (3014) : Object has missing behavior page.  Parallel->Serial of Piplined_Mul_02
Error  (3014) : Object has missing behavior page.  Parallel->Serial of Piplined_Mul_03
Error  (3014) : Object has missing behavior page.  $Parallel->Serial of FileMul
```

Object: Transport   Location: 1096, 425

Smith-Waterman Examples

The Smith Waterman implementation on the Star Bridge Hypercomputer takes advantage of all of the features of the hardware and Viva as the high level programming language. A basic Smith Waterman computational cell has been constructed in Viva with a pipeline structure so that data is passed from one computation cell to the next in a hierarchical order. In this way as many Smith Waterman computations are performed in parallel using as many FPGAs as available in the system. The high communications bandwidth, and parallel memory channels with high memory bandwidth from cell to cell and from FPGA to FPGA in the Star Bridge System enables a scalable structure for expansion. As new hardware is added, either by adding FPGAs, or by adding a new computer with the same system architecture, more SW cells can be computed in parallel. For example if one Hypercomputer board can perform 500 SW cell computations in parallel, two Hypercomputers connected together can perform 1000 SW cell computations in parallel. A thousand Hypercomputers can be chained together without degrading the efficiency of any of the processing cells. This is a sustained compute system, not a peak compute system that can only achieve peak performance on a temporary basis.

The Smith Waterman compute structure is built as a pipeline. The sequence is continuously read into the system as the SW matrix is progressing through the computational cycle. The size of the files can be any size. The X chromosome of 147 million bases versus the Y chromosome of 60 million bases is an early example of the capabilities of the system. To visualize the results, a graphical interface has been developed that interfaces to the output from the Hypercomputer that allows the user to interact with the data to find local similarities and search for global patterns. (Figure 2)



Run-Time Parameters
☑ Grid
☐ Wheel Plane
☑ Lighting
☑ Frame
☑ Labels

Plane Offset: 10

BG Color ■
High Color ■
Low Color 🟩
Plane Color 🟥

Grid / Plane Mix: Or Inverted

Fill Colors
○ 2 Color Gradient
● Rainbow

Fill Type
● Surface
○ Wireframe
○ Points

Smoothing
☐ Disabled
● Gaussian
○ Laplacian
~Std. Dev: 40.0

Load-Time Parameters
Granularity: 164
Reload

☑ FPS

Visualization of Smith-Waterman results from the bacteria Coxiella Burnetii (1,995,275 bases) and Synechocystis PCC6830 (3,573,470 bases). The computation of the Smith-Waterman matrix (7.13 X10^{12} cells) took approximately 10 minutes on a single Hypercomputer board.

Acknowledgements

This work was performed in collaboration with the National Cancer Institute – Frederick and SAIC Frederick.