`Accelerating Genome Sequencing 100X with FPGAs

Olaf O. Storaasli
Oak Ridge National Laboratory
Olaf@ornl.gov

Dave Strenski
Cray Inc.
Stren@cray.com

Abstract

The performance of two Cray XD1 systems with Virtex-II Pro 50 and Virtex-4 LX160 FPGAs was evaluated using the FASTA computational biology program for human genome (DNA and protein) sequence comparisons. FPGA speedups of 50X (Virtex-II Pro 50) and 100X (Virtex-4 LX160) over a 2.2 GHz Opteron were obtained. FPGA coding issues for human genome data are described.

FASTA Algorithm

FASTA [**1**] is used for protein: protein, DNA:DNA, protein: translated DNA and ordered or unordered peptide searches. It calculates similarity statistics for biologists to determine if alignments are random or homotopic. FASTA, who's input format is widely used by other search tools (i.e. BLAST [**2**]) relies on ssearch34 [**3**]. 98.6% of ssearch34 time is spent in the FLOCAL_ALIGN function, a Smith-Waterman FPGA pipeline algorithm [**3-4**] (**Fig. 1**) to calculate the maximum alignment score for two sequences.

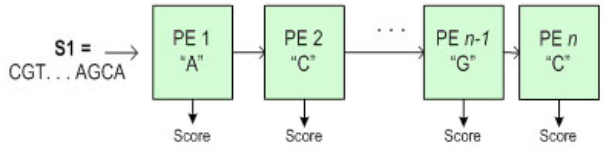

Figure 1. Smith-Waterman FPGA Pipeline

One query character is preloaded into each processing element which then calculate scores in the column of that query character. The database string (S1) is shifted through the pipeline so each database character is compared to each query character in parallel, resulting in a table of scores.

OpenFPGA Benchmark Results:

Single FPGA and Opteron results were obtained for the 4GB human genome sequence openfpga.org benchmark.

Bacillus_anthracis DNA comparison: Genome matching was performed on Virtex2, Virtex4 and Opetron Cray XD1 configurations for 18 DNA query sequences: AE017024-AE017041 on a large database, AE016879 for two outputs:

Detailed: -Q –H –f -10 –g -3 **–d 10** –b 10 –s
Minimal: -Q –H –f -10 –g -3 **–d 0** –b 10 –s

Each query sequence (~300 thousand characters) was compared with the 5 million character database, runs which took over 3 days on the 2.2 GHz Opteron. As the FPGA Smith-Waterman code was limited to a maximum query size of 16k characters (and maximum database size of 512k characters), code was written to split the input query and database into smaller sequences. Ssearch34 results were then obtained for 16k and 8k query sizes for two output options on two Cray XD1 systems (ORNL's *Tiger*-Virtex-II Pro 50 and Cray's *Pacific*-Virtex4 LX160) and compared with 1 Opteron to determine FPGA speedups (**Figs. 2-3**).

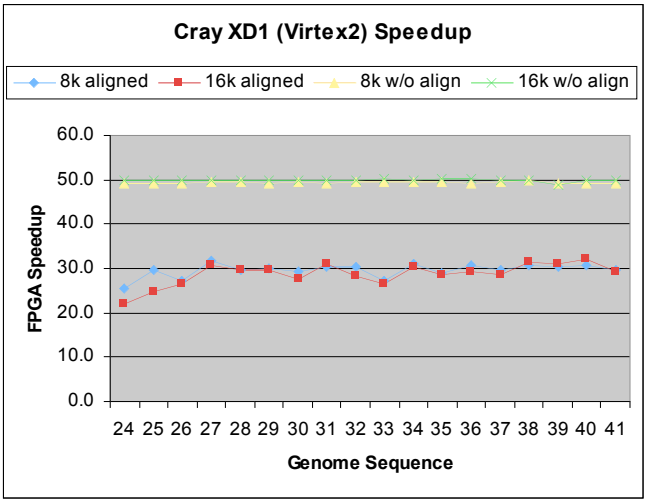

Figure 2. Virtex-II Pro 50 FPGA speedup

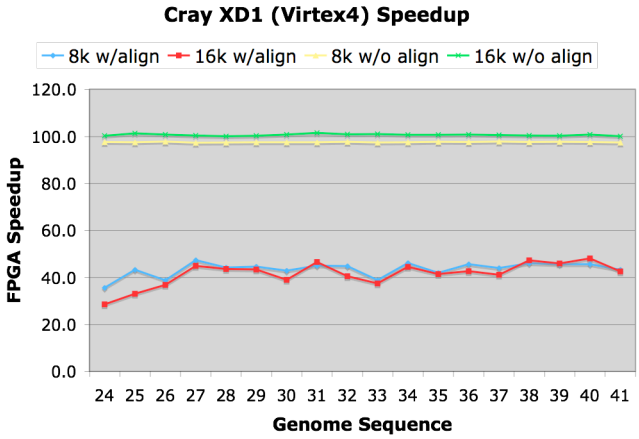

Figure 3. Virtex-4 LX160 speedup

Figsures 2-3 (blue/red) for detailed alignment sequence output shows Virtex-II/Virtex-4 speedups of 29X/43X (1.8/3.9 standard deviation). Reduced output (yellow/green) increased speedups to 50X/100X (0.16/0.13 standard deviation) with 16k queries slightly faster than 8k. Output IO (performed by the Opteron) was small for the Opteron's 75-hour solution time, so it was not optimized. However, reducing the additional output gave significant speedups up to 100X, but only minor reductions on the Opteron.

Clearly, the Virtex4 FPGA's 100X speedup outperformed Virtex2's 50X speedup, so that searches formerly taking 100 days (ie: 14 weeks) can now be completed in one day.

Analyis: With 3X more gates, the Virtex-4 LX160 can process 3x more copies of the algorithm in parallel than the Virtex-II Pro 50. This added logic space allows it to run faster despite its 125MHz clock (less than 140MHz for Virtex-II Pro 50. The slower speed pemits signals to travel across the FPGA. The Virtex-4 design has 128 SWPEs, compared to 48 for the Virtex-II, however, more code optimization could double the FPGA performance to 200X. The original 100 MHz Virtex II algorithm was increased to 140 MHz and similar Virtex4 optimization is also possible .

Query and Database Sizes: Speedup similarity for 8k and 16k queries, prompted additional study on the impact of query and database size on FPGA speedup. The same query sequence and database set was run 30 more times splitting the query sizes into sequences of length 1k, 2k, 4k, 8k, and 16k. The database was then split into sequences sizes: 16k, 32k, 64k, 128k, 256k, and 512k characters. Virtex II Pro FPGA speedups varied from 37-50X the Opteron (**Fig. 4**) for 1k-16k queries with minor variations in data size.

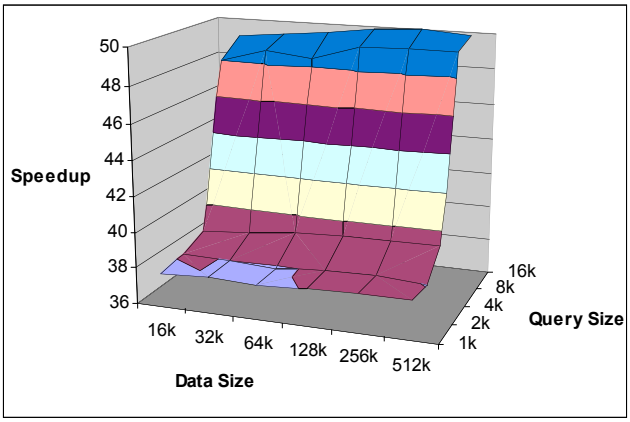

Figure 4. Speedup for Virtex-II Pro 50 FPGA

Larger query sizes (8k and 16k) gave better (~50X) speedups as in **Fig. 13** [2](100X expected for Virtex4).

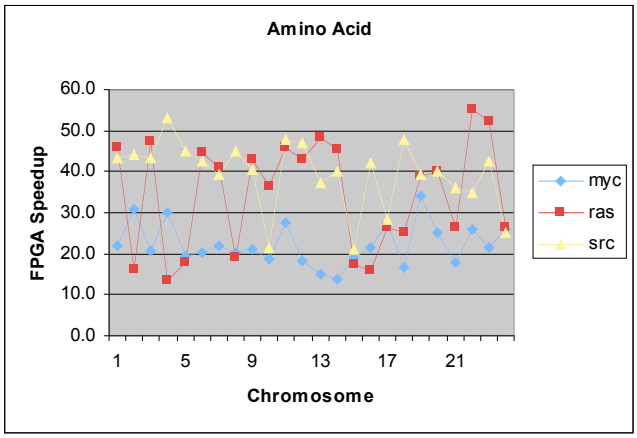

Figure 5. Virtex-II speedup for myc, ras and src sequences

Amino Acid Search: FPGA speedups (Fig. 5) for Amino acid queries (myc, ras, src) in the openfpga.org benchmark showed similar results but with a wider variation among chromosomes (particularly src) which is,attributed to longer sequence lengths.

Conclusions

FPGA performance was evaluated using the FASTA code for comprehensive biological DNA and amino acid sequencing on Cray XD1 computers with both Virtex-II Pro 50 and Virtex-4 LX160 FPGAs. Significant speedups of up to 100x over a 2.2 GHz Opteron processor were achieved. These results indicate similar speedups are likely for acceleration modules (DRC's and Xtreme data) that fit in Opteron sockets, both in small embedded systems and Cray XT supercomputers.

References

[1] FASTA Sequence Comparison Code: fasta.bioch.virginia.edu

[2] MitrionC/BLAST: http://www.hpcwire.com/hpc/1274236.html

[3] Margerm, Steve, and Maltby, Jim; Accelerating the Smith-Waterman Algorithm on the Cray XD1, Cray WP-0060406 2006.

[4] Storaasli, Olaf, Yu, Weikuan, Strenski, Dave, and Malby, Jim; Perfomance Evaluation of FPGA-Based Biological Applications, Cray Users Group Proceedings, Seattle WA, May 2007.

Acknowledgment

This research was sponsored by the Laboratory Directed Research & Development Program of ORNL managed by UT-Battelle for the U. S. Department of Energy on Contract.DEAC0500OR22725. The U.S. Government retains a non-exclusive, royalty-free license to publish or copy the published form of this contribution, or allow others to do so, for U.S. Government purposes.