SCIENGINES

WHITEPAPER
v2.1 (2012)
info@sciengines.com

Accelerated Smith-Waterman on RIVYERA Hardware

Implementation of Smith-Waterman algorithm for sequence alignment on the massively parallel FPGA-computing architecture RIVYERA S3-5000

Abstract

With rapidly increasing sizes of databases, scientists these days oftentimes face the problem of a lack of computational power. Applications and bioinformatics analysis of large datasets may require months of execution time on general purpose computers and scientific progress is stalled. This affects DNA as well as protein-centric research and becomes more and more a major obstacle to modern science as the scope of analysis grows.
Therefore, alternative computing-architectues, e.g. based on FPGAs, are considered for solving the computational problem: The RIVYERA massively parallel reconfigurable computer is such a computing-architecture and its characteristics are depicted in this whitepaper. Benefits of more than 500% computational performance at same cost are observed in comparison to ordinary high-performance computers.

SciEngines GmbH
24118 Kiel
Germany
www.sciengines.com

1. Introduction

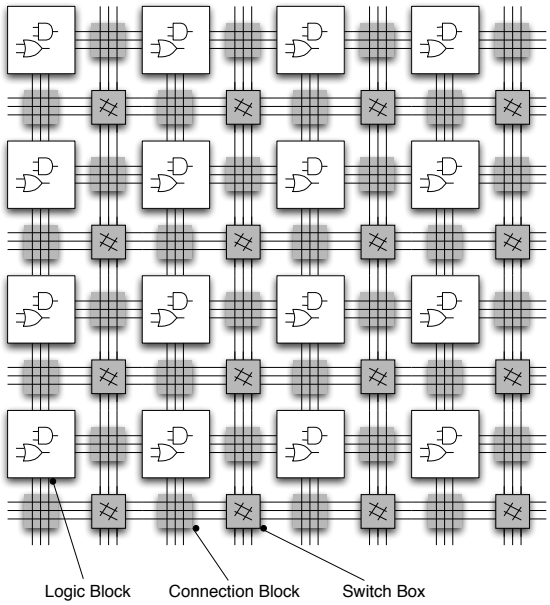
SciEngines' RIVYERA is a massively parallel reconfigurable supercomputer used for high performance analysis and processing. It is designed to perform specific applications at very high speeds. This speed-up furthermore enables users to utilize algorithms with higher quality - algorithms they usually would not consider due to their prohibitive execution time on ordinary computers, such as Smith-Waterman. Bioinformatics algorithms often are ideal candidates for implementation on FPGA (Field programmable gate array) computers because they are generally well parallelizable, which can be fully taken advantage off since each of chips can contain hundreds of processing cores, depending on the complexity of the algorithm. Such massive parallelization enables significant acceleration, even at a lower clock-frequency, which helps save energy and cost.
The RIVYERA S3-5000 is a 3HU, high-density supercomputer and contains 128 FPGA chips. 16 computing cards, each containing eight FPGAs with additional external memory per chip, are coupled via a high throughput backplane. The cards are hooked up through a broadband connection to an internal, preconfigured, off-the-shelf server that runs any regular OS and serves as the interface to the rest of the network. Furthermore, several RIVYERA computers can be interconnected with close to linear performance-scaling. The interconnected supercomputers act like they are just a single large one with an extended address space.

2. Field-Programmable Gate Arrays

A Field-Programmable Gate Array (FPGA) is an integrated circuit containing programmable logic components and programmable interconnects. Millions of configuration bits per FPGA can be programmed to perform whatever logical function is needed and may be reconfigured "in the field" by adjusting the function of logic blocks as well as routing in the connection blocks and switch boxes (figure 1).

During this configuration, the chip can be optimized for the currently required algorithm and can furthermore be structured to conduct several calculations within one chip at the same time.
SciEngines takes this to the next level by manufacturing computers that each contain up to 256 of such FPGAs – leading to massively parallel computing with the performance of small data-centers for specific applications. An example of such a massively parallel FPGA computer is the RIVYERA.

Figure 1 – FPGA structure



3. Applications and Algorithms

One of the most frequently used applications in Life Sciences is the sequence alignment. It searches for sequence similarities in proteins, DNA and/or RNA. It is used to find functional and structural similarities, evolutionary rela-tionships as well as for identification of mutations and Single Nucleotide Polymorphisms (SNPs). It is also part of the assembly of sequences / genomes / transcriptomes from Next-Generation-Sequencing data. In this context, the most accurate algorithm is Smith-Waterman [1], while the most common tool based on heuristics is BLAST [2]. With the increasing focus on short-read NGS data, also a variety of additional, fast sequence alignment tools are commonly used but are usually lacking in accuracy.

3.1 Smith-Waterman description

Smith-Waterman is a non heuristic algorithm to find the optimal local alignment. In comparison to the widely used heuristic approaches, Smith-Waterman reduces the number of false positives as well as raising the quality of the overall alignment by sacrificing speed.

The main differences are:
• The Smith-Waterman algorithm guarantees to find the optimal local alignment between the sequences, whereas heuristics like BLAST only approximates this.

• To guarantee the optimal alignments, the Smith-Waterman algorithm needs to perform many more computations than the BLAST algorithm, making it significantly slower on regular CPUs.
• Smith-Waterman is better in the late stages of the research due to its accuracy. It frees the user from analyzing data afterwards for false positives. Whenever absolutely correct results are needed over estimations, Smith-Waterman is the algorithm of choice.
The benefits of using Smith-Waterman compared to commonly used heuristics are also described in: ”Sensitivity and selectivity in protein similarity searches: a comparison of Smith-Waterman in hardware to BLAST and FASTA” [3].

3.2 Accelerated Smith-Waterman

SciEngines has designed the RIVYERA FPGA-computer that runs Smith-Waterman up to 1’600 times faster than a single 3 GHz core of a desktop computer while consuming very low amounts of electrical energy - 650 Watts. A search taking more than half a day on the desktop PC can be reduced to minutes on the FPGA computer. This enables researchers to run analysis with significantly higher quality but no longer execution time than before.
Our implementation of the Smith-Waterman algorithm allows full utilization of all reconfigurable chips and optimized hardware accelerated DNA-database searching. Providing the same command line as NCBI-based tools, the RIVYERA hardware accelerated Smith-Waterman can be integrated into any custom analysis tool-chain or up-to-date third party application using the standard BLAST tool-chain.

4. Performance Evaluation

We have evaluated the accelerated version of Smith-Waterman on a RIVYERA (see figure 2) against the best-performing implementations that were available from reliable sources. That said, we always recommend a test-run with real data on each architecture, to receive truly reliable performance estimates for the specific situation.

4.1 How to calculate GCUPS

The Smith-Waterman search algorithm compares all nucleotides in the query sequence with all nucleotides in a chosen database. Based on this concept, the speed of a Smith-Waterman database search can be calculated as the number of comparisons per second.
The speed is calculated using the following equation: query • database = CUPS (1) time
• query is the number of nucleotides in the query sequence
• database is the number of nucleotides in the database
• time is the execution time
• CUPS is the number of comparisons per second being performed. The unit for comparisons-per-second is cell updates per seconds or CUPS. GCUPS being “billion (10^9) cell updates per second”.

4.2 Benchmarked systems

We have used the following configurations for our benchmarks:

- RIVYERA S3-5000 (built-No. 0025-2010) with 16 Computing-Cards (Xilinx Spartan-3 5000, XC3S5000-5), 1 Interface-Card running RIVYERA implementation of Smith-

Figure 2 – RIVYERA S3-5000 front

Waterman.

- SSEARCH (FASTA package) running on a desktop computer with an 8 core 2.66 GHz Nehalem-EP CPU and 12 GB of RAM.
- Smith-Waterman was tested on a dataset consisting of 2.480.868 reads, each with 100 basepairs length plus the reverse complement. The data was retrieved from the 1000 genomes project NA12878 and aligned with hg19 chromosom 6 (171 mpb size).
- The data has been transferred onto the RIVYERA user hard disk drive as well as onto the desktop computer. Both hard disk drives are identical, same vendor and model.
- The unix time command has been used to calculate user-cpu-time for the overall process. The measurements are exclusively writing output files or printing to the screen.
- In addition to the RIVYERA and the desktop computer implementation, CUPS performance data has been retrieved from publicized research papers in order to compare alternative computing-architectures.

4.3 Results

Figure 3 clearly shows the significant performance advantage that such a FPGA-based architecture can realize. When factoring in the very low energy consumption and running costs, the cost-performance ratio shows significant advantages as well, with a factor of 3-5x when compared to ordinary hardware.

5. Future developments

Currently, the Smith-Waterman is also being implemented on a large model of the newest generation of the RIVYERA – the RIVYERA S6-LX150. With a maximum of 256 FPGAs of the model Spartan-6 LX150, this model is expected to deliver up to 12.0 TCUPS performance per computer. With additionally planned implementation-improvements, another 10 to 20% speed-increase are anticipated.
For high-speed NGS data-generation, even these speeds may not be sufficient for whole genome sequencing. For answering this challenge, a number of different approaches for pre-filtering NGS data and only running Smith-Waterman on the most relevant data are evaluated.

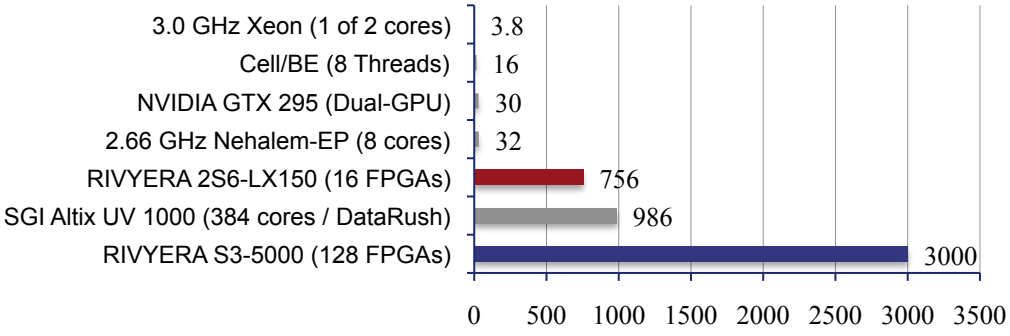
6. Conclusion

The RIVYERA S3-5000 runs at a speed of up to 25 Giga Cell Updates Per Second (GCUPS) per single FPGA, compared to speeds of up to 4 GCUPS peak on a 3 GHz CPU core. The overall performance of a RIVYERA equipped with 16 computing cards is 3.0 TCUPS. A comparable computer cluster consisting of 3 GHz CPU cores would require more than 750 cores if a linear scale-up can be achieved on such an ordinary architecture. Linking multiple RIVYERA computers via the uplink feature provides even greater speedup. Thus, the utilization of reconfigurable high-performance computers can enable new approaches to sequence alignment and improve results by avoiding heuristic approaches. Alternatively, bottle-necks in the analysis-speed for very large problem sets can be circumvented but a few pre-filtering steps may be required for coping with the scale of modern NGS data.

About SciEngines

Germany based SciEngines GmbH provides High Performance Reconfigurable Computing (HPRC) solutions – from smaller scale PCIe and

Figure 3 – Smith-Waterman Benchmark results in GCUPS [4], [5], [6]



USB accelerators to high-performance computers and massively parallel FPGA clusters. Services such as algorithm optimization, custom hardware-design and IT consulting provide additional value for SciEngines' clients.

SciEngines can be contacted at info@sciengines.com or by phone at +49 (0) 431 53 02 48 2.

References

[1] T. F. Smith, M. S. Waterman, Identification of Common Molecular Subsequences, Journal of Molecular Biology 147 (1981) 195–197 doi:10.1016/0022–2836(81)90087–5.
[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic Local Alignment Search Tool, Journal of Molecular Biology 215 (3) (1990) 403–410.
[3] M. D. J. D. R. T. Shpaer, E. G.; Robinson, Sensitivity and selectivity in protein similarity searches: a comparison of smith-waterman in hardware to blast and fasta., Genomics 38 (1996) 179–19.
[4] G. Pfeiffer, S. Baumgart, J. Schröder, M. Schimmler, A Massively Parallel Architecture for Bioinformatics, in: ICCS2009, Lecture Notes in Computer Science, Vol. 5544, Springer, 2009, pp. 994–1003.
[5] Y. Liu, B. Schmidt, D. Maskell, CUDASW++2.0: enhanced Smith-Waterman protein database search on CUDA-enabled GPUs based on SIMT and virtualized SIMD abstractions, BMC Res Notes 2010, 3:93.
[6] Press Release, Pervasive Software company website, Pervasive DataRush on SGI® Altix® UV 1000 Shatters Smith-Waterman Throughput Record by 43 Percent, September 27th 2010.