Designs for accelerating Bioinformatic problem solving using FPGAs based HPC system

Gaurav Kumar Singh
Paderborn University
gauravks@mail.uni-paderborn.de

Abstract

The Human Genome project was marked as completed in the year 2003 which opened vast avenues for research towards developing and enhancing Genomic analysis techniques. With such vast sequence database available, the Genomic research greatly depends on bioinformatics capabilities and improvements to the computational speed would help in analysis the data to discover causes and treatments for various diseases. This has been a major driving factor to develop techniques to increase the processing capabilities of the existing algorithms, tools and techniques by utilizing advancements in computing power. FPGA based acceleration for existing algorithms presents very promising advantages, reducing processing times by huge factor compared to other CPU and GPU based techniques. Similarly, the introduction of high performance clusters to distribute the processing has already been used and shown to be effective for large sequence analysis.

Combining these technologies together possess great benefits for even speeding up the analysis of the huge databases. This paper will present some of the techniques and heterogenous system which have been developed and utilized to speed the the genome analysis from years to days. Initially we look at bioinformatics application areas where FPGA and HPC system are beneficial. Then the paper describes some of the tools and algorthmic accelerators using FPGA and HPC in a heterogenous system. The last part presents some systems and evaluation results in terms of speedup compared to existing systems and tools.

1 Introduction

Humans quest to understand the basic biological processes lead to development of research areas such as biochemistry and biotechnology. Multiple decades of research in the biological molecules helped us in understanding the existence DNA and genome which defines how a living organism behaves and exist. On the other hand the advancement in computer technologies and increased use of them in healthcare, biomedical and computational biological research has helped find cure and medical treatments for many complex health issues and save many lives over the years.

In efforts to increase the knowledge of genomes, the field of Bioinformatic was created. Bioinformatic majorly involves the study of biological molecules (biomolecules) which build up the cells of the living organisms. As with the other biological fields, bioinformatic aims at utilizing the capabilities of the computer science to build and analyse molecular sequences (genes) of genomes. In this direction, The *Human Genome Project*[1] was started in late 1990 and was completed in 2003 successfully. "A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method"[1]. This was a huge step but also presented the problem of huge processing times for analysis of such a large database of genome for extracting any useful information. The existing algorithms for database searches such as Smith-Waterman [2] based on dynamic programming for local similarity estimation and heuristics based BLAST [3] were limited by high computation times. Schmidt, Schröder, and Schimmler [4] have demonstrated a parallel system which helps to speed up the molecular sequence analyse. The main limiting factors at this point were the processing capabilities of the computing units on which the algorithms were running.

Add pic of human genome from somewhere

Various methods in the past decades have evolved to provide higher computing and data processing capabilities for various application domain. The earliest method being hardware acceleration provided by symmetric multiprocessing which allows distribution of computing to different processor sharing a common memory. The next major acceleration achieved has been the development of high performance computing clusters (HPC). HPC system work by splitting and distributing the problem over multiple similar processing units popularly known as nodes. Each of the node, consists of a high performance processor with multiple cores and sharing the same common memory. The nodes in the clusters are connected to each other with high speed Ethernet connections for exchange of data and control information as shown in fig<>. Each node can be used to process a sub-set of the data parallely decreasing the overall computing time for the problem. Due to such benefits, these techniques have being used to speed up the Bioinformatic algorithms by modification to work on these HPC clusters and utilize the benefits. [5], [6] gives implementation of the famous Smith-Waterman algorithm on HPC systems. A various number of parallel implementation for BLAST such as mpiBLAST [7] are available as well which prove to be more time efficient.

Add pic of cluster

Another step in increasing the processing capabilities of the clusters is use of GPU. GPUs allows offloading the vector based arithmetic operations for large datasets. The GPUs prove to be excellent accelerators for reducing the processing time of complex calculation on large amount of data which are common in many of the application domain utilizing the clusters. Liu, Schmidt, and Muller-Wittig [8] have presented such a system which is capable of performing 10 times faster compared to serial versions.

Though the parallel implementation with CPU and GPU help in achieving faster processing time, its heavily dependant on the size of the cluster. Also the speedup is de-

[1] `http://https://www.genome.gov/`

pendant on the size of the problem as well. These reasons made researchers to look for areas for improving the execution times of the algorithm by using hardware based accelerators for the algorithms to reduce processing time for each operation. This is where the FPGA has helped a lot by providing opportunities to implement the algorithms directly in the hardware. The flexibility of FPGA based accelerators makes them very useful to design application specific acceleration hardware and also re-use them for different kinds of problems. Currently a lot of accelerators are available which the bioinformatic community is benefiting from. This paper would discuss some of these implementation and give an overview of how such accelerators are integrated with the HPC clusters to build heterogenous systems which are used to achieve very high processing speeds to reduce the time from days to hours required for some bioinformatic application.

The rest of the paper is divided into 3 sections. Section 2 introduces the bioinformatic application domain giving details of algorithms and tools popularly used. Section 3 will present the optimization techniques for genome comparison by FPGA and heterogenous systems and the last section presents results achieved by such optimization for some of the current systems.

2 Bioinformatic

Introduction to the concept of Bioinformatic

2.1 Application areas

2.1.1 Genome Assembly

2.1.2 Contenct-based Search

2.1.3 Genome comparison

2.1.4 Pattern Matching

2.1.5 Genome Databases

2.2 Algorithms

2.2.1 Dynamic Programming

References [9], [2]

2.2.2 Seed-based Heuristics

Reference [10], [3]

2.2.3 Languages Models and Profiles

References [11] [12]

3 Optimization of Genome comparison algorithms

This section would give specific example for Techniques and related system designs which use FPGA based HPC systems to accelerate the solution finding.

Add subsection to describe the techniques along with references.

3.1 FPGA accelerators for Smith-Waterman

3.2 Heterogenous system designs

4 Evaluation and comparison

This section should highlight the possible acceleration which is possible with the FPGA based system using the Evaluation data of different techniques and present a comparative study of how this techniques vary to each other and to traditional HPC and serial computing.

This should be able to highlight the advantages of using FPGAs for certain problem to reduce cost and time for for the problems.

5 Conclusion

Timelogic accelerators

References

[1] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The Sequence of the Human Genome", en, *Science*, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001. DOI: 10.1126/science.1058040.

[2] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences", *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, Mar. 1981. DOI: 10.1016/0022-2836(81)90087-5.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool", *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct. 1990. DOI: 10.1016/S0022-2836(05)80360-2.

[4] B. Schmidt, H. Schröder, and M. Schimmler, "Massively Parallel Solutions for Molecular Sequence Analysis", in *Proceedings of the 16th International Parallel and Distributed Processing Symposium*, ser. IPDPS '02, Washington, DC, USA: IEEE Computer Society, 2002, pp. 201–.

[5] A. Boukerche, A. C. M. A. d. Melo, M. Ayala-Rincon, and T. M. Santana, “Parallel Smith-Waterman Algorithm for Local DNA Comparison in a Cluster of Workstations”, en, in *Experimental and Efficient Algorithms*, ser. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, May 2005, pp. 464–475. DOI: 10.1007/11427186_40.

[6] W. S. Martins, J. B. Del Cuvillo, F. J. Useche, K. B. Theobald, and G. R. Gao, “A MULTITHREADED PARALLEL IMPLEMENTATION OF A DYNAMIC PROGRAMMING ALGORITHM FOR SEQUENCE COMPARISON”, en, WORLD SCIENTIFIC, Dec. 2000, pp. 311–322. DOI: 10.1142/9789814447362_0031.

[7] A. E. Darling, L. Carey, and W.-c. Feng, “The Design, Implementation, and Evaluation of mpiBLAST”, in *In Proceedings of ClusterWorld 2003*, 2003.

[8] W. Liu, B. Schmidt, and W. Muller-Wittig, “CUDA-BLASTP: Accelerating BLASTP on CUDA-Enabled Graphics Hardware”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1678–1684, Nov. 2011. DOI: 10.1109/TCBB.2011.33.

[9] E. Rucci, A. D. Giusti, M. Naiouf, G. Botella, C. García, and M. Prieto-Matias, “Smith-Waterman algorithm on heterogeneous systems: A case study”, in *2014 IEEE International Conference on Cluster Computing (CLUSTER)*, Sep. 2014, pp. 323–330. DOI: 10.1109/CLUSTER.2014.6968784.

[10] A. Mahram and M. C. Herbordt, “Fast and Accurate NCBI BLASTP: Acceleration with Multiphase FPGA-based Prefiltering”, in *Proceedings of the 24th ACM International Conference on Supercomputing*, ser. ICS ’10, New York, NY, USA: ACM, 2010, pp. 73–82. DOI: 10.1145/1810085.1810099.

[11] T. Oliver, L. Y. Yeow, and B. Schmidt, “Integrating FPGA acceleration into HMMer”, *Parallel Computing*, High-Performance Computational Biology, vol. 34, no. 11, pp. 681–691, Nov. 2008. DOI: 10.1016/j.parco.2008.08.003.

[12] N. Abbas, S. Derrien, S. Rajopadhye, P. Quinton, A. Cornu, and D. Lavenier, “Combining execution pipelines to improve parallel implementation of HMMER on FPGA”, *Microprocessors and Microsystems*, vol. 39, no. 7, pp. 457–470, Oct. 2015. DOI: 10.1016/j.micpro.2015.06.006.