Thesis Overview:

Performance and energy efficiency evaluation of heterogeneous systems for bioinformatics

Enzo Rucci
School of Computer Science, National University of La Plata, Argentina
PhD Thesis in Computer Science[1]
Advisors: Armando De Giusti, Marcelo Naiouf, Carlos García Sánchez, Guillermo Botella Juan
{erucci,degiusti,mnaiouf}@lidi.info.unlp.edu.ar, {garsanca,gbotella}@ucm.es

The power consumption problem represents one of the major obstacles for exascale systems design. As a consequence, the scientific community is searching for different ways to improve power efficiency of High-Performance Computing (HPC) systems [1]. One recent trend to increase compute power and, at the same time, limit power consumption of these systems lies in adding accelerators and co-processors, like NVIDIA/AMD graphic processing units (GPU) or Intel Xeon Phi co-processors. On the other hand, field programmable gate arrays (FPGA) stands as a promissory alternative due to their increasing computational capability, low power consumption and the development of new tools that reduce programming cost. These hybrid systems that use different processing resources are called heterogeneous systems and area capable of achieving better FLOPS/Watt ratios [2].

Bioinformatics is one of the areas affected by current HPC problems due to the exponential growth of biological data in the last years and the increasing number of bioinformatics applications demanding HPC to meet performance requirements. One of these applications is sequence alignment, which is considered to be fundamental procedure in biological sciences [3]. The alignment process compares two or more biological sequences and its purpose is to identify regions of similarity among them. The Smith-Waterman (SW) algorithm [4] is a popular method for local sequence alignment that has been used as the basis for many subsequent algorithms, and is often employed as a benchmark when comparing different alignment techniques. However, due to the quadratic computational complexity of Smith-Waterman algorithm, several heuristics are used in practice that reduce the execution time but at the expense of not guaranteeing to discover the optimal local alignments.

In order to process the ever increasing quantity of biological data with acceptable response times, it is necessary to develop new computational tools that are capable of accelerating key primitives and fundamental algorithms in an efficient manner from performance and energy consumption points of view. For that reason, this thesis considered, as general objective, evaluating performance and energy efficiency of HPC systems for accelerating Smith-Waterman biological sequence alignment.

First, the SW parallelization ways were studied and the available implementations over different hardware platforms (CPU, GPU, FPGA and Xeon Phi) were described. The analysis for each platform included temporal evolution, contributions, limitations and experimental work, and the results of each implementation. This analysis required the study of different hardware architectures and programming models as wells as the different sequence alignment methods.

Next, new algorithmic solutions for heterogeneous systems were developed. As GPUs are the dominant accelerator in HPC community today, and there is a wide availability of scientific research to use this kind of accelerator for sequence alignment, the development of new implementations for heterogeneous systems based on Xeon Phi's and FPGAs was prioritized. At the beginning of this thesis, no Xeon Phi-based implementation for sequence alignment was developed. On the contrary, there existed previous works for sequence alignment computing using FPGAs, although these implementations were developed using traditional hardware description languages (HDL) and most of them had one or more limitation that restricted its usage.

The experimental work started with Xeon Phi-based heterogeneous systems. As starting point, CPU and Xeon Phi implementations were individually developed and optimized before combining them into a hybrid solution. These implementations were compiled into a single tool called SWIMM. The tool SWIPE [5] was identified as the fastest implementation for CPU similarity search from the previous state-of-the-art study. Through experimental work carried out, it was demonstrated that the SSE-version of SWIMM is comparable to SWIPE, while the corresponding AVX2-version was able to outperform SWIPE by up to 1.4×. It is important to mention that, to the best of the author's knowledge, SWIMM is the first implementation using AVX2 instruction set. In

[1]Full text available at: http://sedici.unlp.edu.ar/handle/10915/53045

relation to Xeon Phi, there were no solutions for these co-processors when this research began. However, SWAPHI [6] and XSW 2.0 [7] were introduced in 2014. SWAPHI only exploits compute power of Xeon Phi while XSW 2.0 is able to take advantage of both, CPU and co-processor. It was demonstrated that the KNC-version of SWIMM is competitive with SWAPHI, is able to outperform the latter for medium and long-sized sequences. Additionally, the hybrid version of SWIMM with dynamic work distribution (SSE+KNC) outperformed its alternative XSW 2.0, achieving up to 3.9-fold speedups.

The next step in experimental work was the development of new algorithmic solutions for FPGA-based heterogeneous systems. Unlike state-of-the-art implementations, this thesis examined the benefits of a highly innovative technology in the form of supporting the OpenCL for FPGAs, instead of traditional HDLs like VHDL or Verilog. In the first place, performance and resource usage of different kernels were explored in order to find the most beneficial configuration. Subsequently, a hybrid version was developed using SWIMM code. These implementations were also compiled in a single tool called OSWALD. As far as the author knows, this is the first high-level programming implementation on FPGAs using OpenCL. Besides, this is one of few implementations that is portable, completely functional and general for these kind of systems. Unfortunately, the absence of source code prevented a comparison with other FPGA implementations and, while a theoretical analysis is still possible, the different devices, technologies and methodologies used complicate a direct and fair comparison.

The last step in experimental work consisted in relating performance achieved by the different systems used previously with their corresponding power consumption. According to the analysis carried out, CPUs offer a good balance between performance and power consumption from the exploitation of multithreading and vector instructions, especially those with AVX2 instruction set. On the other hand, the incorporation of accelerators to improved global system performance in all cases, although the improvement factor varied depending on the accelerator selected. Unfortunately, results were different from an energy efficiency point of view. Xeon Phi-based systems are not a good choice for this problem from an energy efficiency perspective principally due to the absence of low-range vector capabilities on this coprocessor. The performance gain is smaller than the increase in power consumption, and this translates to less GCUPS/Watt. Both, FPGAs and GPUs, can effectively exploit this kind of data type and that is why these accelerators are able to improve energy efficiency. GPU accelerated systems offer higher performance rates but at the expense of higher power consumption rates too. CPU-FPGA systems offer less GCUPS than GPU-based platforms. However, because its power consumptions is lower, the energy efficiency rates are higher. Unlike other performance and energy efficiency evaluations in SW context, this thesis used real and representative amino acids datasets, powerful hardware platforms and the best implementations available in literature so far for each device. For that reason, this evaluation can be more useful in real world.

According to the obtained results and the contributions, it is expected that this thesis contributes to a wider SW adoption from the bioinformatics community and to a more efficient sequence alignment computing from both, performance and power consumption perspectives.

References

[1] M. Giles and I. Reguly, “Trends in high-performance computing for engineering calculations”, Philosophical Transactions of the Royal Society A, vol. 372, no. 2022, p. 20130319, 2014.

[2] M. Vestias and H. Neto, “Trends of CPU, GPU and FPGA for high-performance computing,” in 24th International Conference on Field Programmable Logic and Applications (FPL), 2014, Sept 2014, pp. 1–6.

[3] B. Schmidt, “Bioinformatics: High Performance Parallel Computer Architectures”, B. Schmidt, Ed. CRC Press, 2010.

[4] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” Journal of Molecular Biology, vol. 147, no. 1, pp. 195–197, March 1981.

[5] T. Rognes, “Faster Smith-Waterman database searches with inter-sequence SIMD parallelization”, BMC Bioinformatics, vol. 12:221, 2011.

[6] Y. Liu and B. Schmidt, “SWAPHI: Smith-Waterman protein database search on Xeon Phi coprocessors,” in 25th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP 2014), 2014.

[7] L. Wang, Y. Chan, X. Duan, H. Lan, X. Meng, and W. Liu. (2014) “XSW 2.0: A fast Smith-Waterman Algorithm Implementation on Intel Xeon Phi Coprocessors”. Available at: http://sdu-hpcl.github.io/XSW/