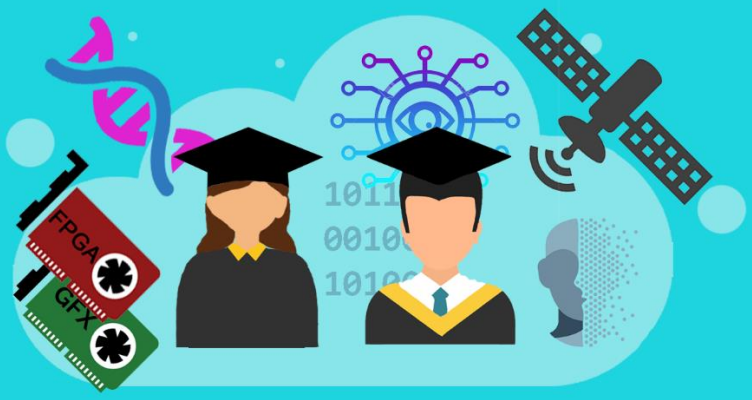
Diploma Thesis

Microproccessors and Digital Systems Laboratory

Academic year 2020-2021



Accelerating DNA sequence aligner using OpenCL programming model

The development of next-generation sequencing (NGS) technologies has dramatically changed the landscape of human genetics research. NGS platforms have now the capacity to generate billions of short fragments of DNA in a matter of hours. These small pieces of DNA, called reads, are the input to various types of genomic analysis such as variant calling and differential gene expression. The first step in any genomic analysis pipeline however is short read alignment, which entails finding a specific location on the reference human genome where a short read is best mapped. The vast amount of sequencing data and the excessive time requirements for this step to execute, have put considerable strain on the computing systems used for genome analysis. Since the throughput of NGS technologies does not cease its exponential growth, there is an ever-present need for identifying bottlenecks and proposing accelerated solutions for popular aligner tools.

Several aligners have been developed that rely on a seed-and-extend model for aligning the short reads. According to this model, in the seeding step, each short read is further fragmented in short pieces, called seeds, that align exactly on the reference genome. In the seed-extension step, each seed is extended so that the whole short read aligns with the reference, allowing mismatches. Most of the state-of-the-art aligners implement variations of the Smith-Waterman string matching algorithm to perform the seed-extension step. The valid alignments of the reads are reported in an output file following a standard format, and available for use in the next steps of the genomics pipeline.

The scope of this diploma thesis is to accelerate widely-used Bowtie2 aligner using the OpenCL programming model. The target hardware device can be either CPU, FPGA or GPU thanks to OpenCL versatility. OpenCL-based accelerators can be developed for various stages of the aligner, including the seeding step, the seed-extension step (e.g. SmithWaterman etc) and the final reporting step. All accelerators developed will be integrated in the aligner to allow for an end-to-end performance enhancement and evaluation. The final design will be measured in a cluster with access to both FPGA and GPU devices.

PREREQUISITES: C/C++ programming, OpenCL programming, Scripting Skills

Academic Advisor1: Konstantina Koliogeorgi, konstantina@microlab.ntua.gr

Academic Advisor2: Sotirios Xydis, sxydis@microlab.ntua.gr

Academic Advisor3: Dimitrios Soudris, dsoudris@microlab.ntua.gr