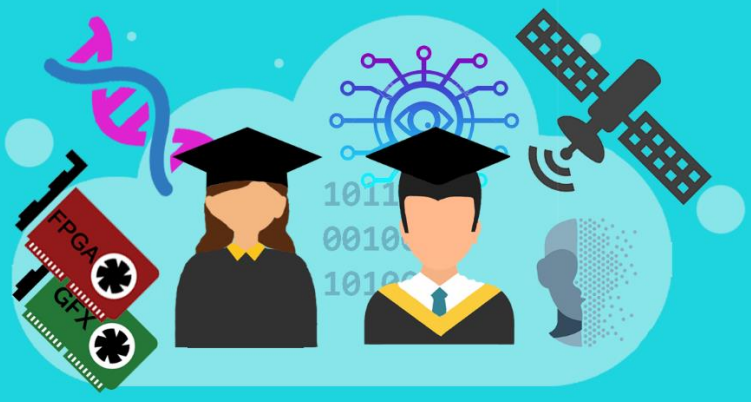
Diploma Thesis

Microproccessors and Digital Systems Laboratory

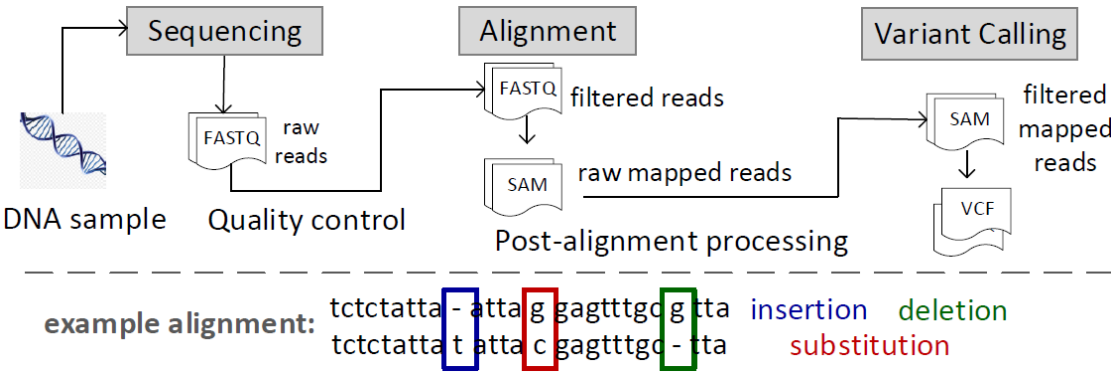
Academic year 2020-2021



Accelerating NGS genomic pipelines leveraging Heterogeneous Computing and Serverless Computing Architectures

A genome contains an organism's complete set of genetic instructions and information. Genome study is the focus of genomic workflows (e.g. variant calling, differential gene expression, phylogeny creation) and is essential to discover the mechanisms that lead to genetic diversity. The first step of a typical genomics workflow converts raw signals from individuals into short fragments of bases (i.e. nucleotide bases A,C,G,T), called short reads. Short read alignment then performs the mapping of the short reads to a location in the reference genome that is most likely its origin. Read alignments are read by the variant calling step, which identifies differences between the aligned reads and the reference genome. The exponential growth of sequencing data and the excessive time requirements of these workflows, have put considerable strain on the computing systems used for genome analysis and have stressed the need for accelerating the individual stages or the entire workflow.

At the same time, heterogeneous clusters and cloud infrastructures have lately been a popular platform for running genomic pipelines. Therefore, we can leverage technologies and features of these platforms to boost genomics performance. Serverless computing can be employed to achieve that goal. In this new computing and scheduling model, large applications are transformed into more structured ones with smaller execution units. Each of these smaller tasks is scheduled to servers and assigned resources depending on the availability and requirements. Deployment of tasks is decided by the framework upon a user request for a workflow execution. Once deployed, the tasks communicate with each other in an event-driven manner, without needless interactions with the framework.



In this diploma thesis, the goal is to accelerate a genomic pipeline using serverless computing. The genomic pipeline will include various stages of aligners and variant callers and it should change dynamically according to user configurations. The scheduler will then break the pipeline into independent stages and schedule the flow of data between stages with the help of specialized frameworks. Depending on the stage, the framework can also exploit heterogeneous recourses and accelerate the stage with pre-built GPU and FPGA accelerators. As a next level of optimizations, the same tactic can be applied within each stage and therefore extend the depth of the pipeline.

PREREQUISITES: C/C++ programming, Scripting Skills, FPGA/GPU programming

Related Frameworks: SeqMule, Bowtie2, GATK, Kubernetes, OpenWhisk

Contact:

Konstantina Koliogeorgi, konstantina@microlab.ntua.gr
Achilleas Tzenetopoulos Ph.D.student: (atzenetopoulos@microlab.ntua.gr)
Dimosthenis MasourosPh.D. student: (dmasouros@microlab.ntua.gr)
Sotirios Xydis Ass.Prof.: (sxydis@microlab.ntua.gr)
Dimitrios Soudris Prof.: (dsoudris@microlab.ntua.gr)