

Ph.D. Thesis Defense



Accelerating Genome Sequence Analysis via Efficient Hardware/Algorithm Co-Design

Damla Senol Cali

July 15, 2021 | 10:00 a.m. EDT | Virtual public defense (Zoom)

Abstract

Genome sequence analysis plays a pivotal role in enabling many medical and scientific advancements in personalized medicine, outbreak tracing, the understanding of evolution, and forensics. Modern genome sequencing machines can rapidly generate massive amounts of genomics data at low cost. However, the analysis of genome sequencing data is currently bottlenecked by the computational power and memory bandwidth limitations of existing systems, as many of the steps in genome sequence analysis must process a large amount of data. Our goals in this dissertation are to (1) characterize the real-system behavior of the genome sequence analysis pipeline and its associated tools, (2) expose the bottlenecks and tradeoffs of the pipeline and tools, and (3) co-design fast and efficient algorithms along with scalable and energy-efficient customized hardware accelerators for the key pipeline bottlenecks to enable faster genome sequence analysis.

First, we comprehensively analyze the tools in the genome assembly pipeline for long reads in multiple dimensions (i.e., accuracy, performance, memory usage, and scalability), uncovering bottlenecks and tradeoffs that different combinations of tools and different underlying systems lead to. We show that we need high-performance, memory-efficient, low-power, and scalable designs for genome sequence analysis in order to exploit the advantages that genome sequencing provides. Second, we propose GenASM, an acceleration framework that builds upon bitvector-based approximate string matching (ASM) to accelerate multiple steps of the genome sequence analysis pipeline. We co-design our highly-parallel, scalable and memory-efficient algorithms with low-power and area-efficient hardware accelerators. We evaluate GenASM for three different use cases of ASM in genome sequence analysis and show that GenASM is significantly faster and more power- and area-efficient than state-of-the-art software and hardware tools for each of these use cases. Third, we implement an FPGA-based prototype for GenASM, where state-of-the-art 3D-stacked memory (HBM2) offers high memory bandwidth and FPGA resources offer high parallelism by instantiating multiple copies of the GenASM accelerators. Fourth, we propose GenGraph, the first hardware acceleration framework for sequence-to-graph mapping. Instead of representing the reference genome as a single linear DNA sequence, genome graphs provide a better representation of the diversity among populations by encoding variations across individuals in a graph data structure, avoiding a bias towards any one reference. GenGraph enables the efficient mapping of a sequenced genome to a graph-based reference, providing more comprehensive and accurate genome sequence analysis.

Overall, we demonstrate that genome sequence analysis can be accelerated by co-designing scalable and energy-efficient customized accelerators along with efficient algorithms for the key steps of genome sequence analysis.

Zoom public defense

Meeting ID: 980 8964 2738 (<https://cmu.zoom.us/j/98089642738>)

Please contact **Damla Senol Cali** (dsenol@andrew.cmu.edu) or **Tunca Cali** (calit@duq.edu) for the meeting password.

Committee

Onur Mutlu

CMU-ECE, ETH Zurich

Co-advisor

Saugata Ghose

CMU-ECE, University of
Illinois Urbana-Champaign

Co-advisor

James C. Hoe

CMU-ECE

Can Alkan

Bilkent University