BioArch: a reconfigurable hardware accelerator designed for bioinformatics workloads

S. Karen Khatamifard[1], Meisam Razaviyayn[2], Ulya R. Karpuzcu[1]
[1]University of Minnesota, Electrical Engineering, Minneapolis, MN, [2]University of Southern California, Industrial and Systems Engineering, Los Angeles, CA

Recent advances in sequencing technologies have revolutionized medicine and biology. Modern sequencing platforms can sequence tens of billions of bases per each run. Processing these massive datasets can take up to hours or days even in the presence of significant amount of computational resources. These computationally expensive tasks motivate the use of hardware-level acceleration with optimized computing architectures. In this talk, we discuss how widely-used computational tasks in bioinformatics can significantly benefit from the use of optimal hardware architectures and algorithms. In particular, we introduce BioArch, a reconfigurable hardware accelerator designed for bioinformatics workloads. BioArch aims to accelerate major reference-guided and de novo tasks in computational biology.

BioArch, capable of analyzing both short and long reads, has two central hardware components: one for pre-aligning reads to reference sequence(s), Filter Unit (FilterU); and the other for finding exact pairwise similarity score between two given short/long reads, Match Unit (MatchU). FilterU can efficiently prune the candidate hits with the possibility of using parallel FilterUs to prune even more aggressively. MatchUs, on the other hand, evaluates the similarity of two given sequences in constant time independent of the sequencing error rate or read lengths. MatchU's design is based on the clever use of processing in-memory (PIM) technologies, capable of handling simple (mostly integer) computations inside the memory where the data reside. PIMs are recently introduced as an energy efficient novel form of computing. PIMs can effectively remove all data transfers to and from memory, which is the main performance and energy bottleneck of today's data-intensive applications.

In addition to standard k-mer hashing strategies, BioArch benefits from a novel hash-based similarity evaluation which has been recently introduced in the image-processing community. This hashing function is developed with the training of deep convolutional neural networks on sequencing data. Our hash-based neural network leads to a linear time alignment of similar PacBio sequences with more than 98% accuracy.

In the last part of the talk, we share our numerical experiments on BioArch for two important case studies: 1) short-read alignment of Illumina reads and 2) De novo transcriptome sequencing with PabBio long reads. Our simulations show orders of magnitude higher throughput (7.5x) and energy efficiency (109.0x), when compared to representative, optimized state-of-the-art software-based algorithms.