communication and accelerator call overheads. In this paper, we propose a dataflow architecture for Smith-Waterman Matrix-fill and Traceback alignment stages, to perform short-read alignment on NGS data. The architectural decision of moving both stages on chip extinguishes the communication overhead, and coupled with radical software restructuring, allows for efficient integration into widely-used Bowtie2 aligner. This approach delivers ×18 speedup over the respective Bowtie2 standalone components, while our co-designed Bowtie2 demonstrates a 35% boost in performance.

***Index Terms*—Next Generation Sequencing, Reconfigurable Acceleration, Dataflow Computing, Bowtie2, Smith Waterman, Traceback**

I. Introduction

The development of next-generation sequencing (NGS) technologies has dramatically changed the landscape of human genetics research [1]. Advances in the field of DNA and RNA sequencing have led to effective genome mapping and have paved the way to personalized genomic medicine [2]. 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 [3] and differential gene expression [4]. 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 [5], there is an ever-present need for identifying bottlenecks and proposing accelerated solutions for popular aligner tools.

Several aligners [6], [7] 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

This work has been partially funded by EU Horizon 2020 program under grant agreement No 825061 EVOLVE (https://www.evolve-h2020.eu/)

1946-1488/19/$31.00 ©2019 IEEE
DOI 10.1109/FPL.2019.00021

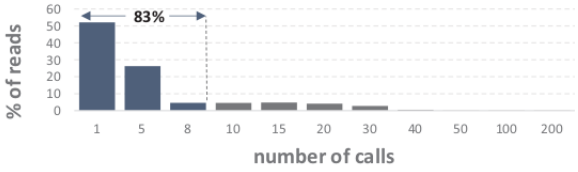

Fig. 1: Distribution of Matrix-Fill Function Calls Across Reads in Bowtie2.

provided by existing solutions. Co-design of NGS alignment exposes several challenges that can be only highlighted by holistically and carefully profiling a sequencer and modeling the behavior of a co-designed version.

In this paper, we design, implement and explore a novel high performance reconfigurable accelerator for Smith-Waterman backwards until it constructs a valid alignment path.

A profiling based on a 10 million short-read input dataset of length 100bp, that was collected as part of EU healthcare project AEGLE [9], indicates that Smith-Waterman dominates the execution time of Bowtie2 aligner [6] by a percentage of 60%. However, the histogram in Fig.1 showcases that this time is actually shared among independent Smith-Waterman tasks distributed across all reads. Each read alignment can invoke from one up to 270 Smith-Waterman *matrix-fill* tasks, each one followed by a *traceback* task. All individual *matrix-fill* tasks add up to the 56% of total execution time and *traceback* 4%. An initial naive approach would target this stage to employ hardware acceleration and tackle the alignment bottleneck. A straightforward integration of an accelerated *matrix-fill* phase of Smith-Waterman though [10], [11], [12] would introduce a huge overhead, due to both the immense amount of the accelerator calls and the transferring of the matrices to the CPU for the traceback stage. In fact, taking into account the time overhead provisioned by each call to the accelerator and the accelerator-CPU transfer time for each matrix, the overall execution time of the aligner can actually be increased. A challenge as such has also been noted by [13] regarding JVM-FPGA communication overhead.

Most existing works either propose standalone matrix-fill Smith-Waterman acceleration ignoring the traceback stage [10]–[12] and thus the communication overhead in a real system, or provide an end-to-end hardware implementation of both seed and extend phases [14], [15]. The latter architectural decision introduces immense memory requirements, to support storing the human genome on chip. Moreover, such systems constitute new tools that introduce a learning curve for biologists and come at the expense of safe-to-use results and advanced visualization analyses provided by well-known and defacto sequencing frameworks such as Bowtie2 [6], BWA [7]. There are only a few software/hardware co-design acceleration works for short-read alignment [16], [17]. Therefore, effective co-design of the NGS short-read alignment still remains an open issue, mainly due to narrow view on real integration of a matrix anti-diagonal per time step. The authors in [12] provide a very detailed architecture as such, that implements a multistage-PE design, and optimize each stage in terms of resources utilization and delay. Similarly in [20], the authors propose a reconfigurable accelerator that implements a modified equation to improve mapping efficiency of a single PE, and a special floor plan to cut down the interface components routing delay. In this paper, we aim to extend these designs by implementing the complete Smith-Waterman algorithm along with the Traceback procedure. In our final high-throughput real system, on-chip traceback diminishes matrix transfer overhead cost and thus enables efficient integration, without adding extra latency thanks to the pipelined scheduling of consecutive tasks.

There are only a few works of accelerated sequence alignment based on Smith-Waterman with Traceback. The authors in [21] propose a space efficient, global sequence alignment architecture that accelerates both the forward scan and traceback for variable reference lengths. The traceback procedure is