Hardware-Accelerated DNA Sequencing

Zhongpan Wu, Karim Hammad, Yunus Dawji,
Ebrahim Gafar-Zadeh and Sebastian Magierowski

I. Short Summary

DNA sequencing denotes the physical and computational pipeline responsible for extracting the genetic code stored within the DNA molecules of living organisms. This process is fundamental to life science and is enabling progress towards personalized medicine. Basecalling is a critical computational task within the DNA sequencing pipeline. Its function is to predict the unique order of the four chemical DNA building blocks from physical measurements, and thus, produce a text-equivalent of DNA. By employing state-of-the-art nanosensors, today's sequencers have achieved accelerated measurement rates within hand-sized footprints. These devices place sequencing at the doorstep of real-time mobile applications, but the poor quality of their signals impose an extreme computational burden on basecalling, presently requiring large and power-hungry compute resources. To ease that burden, we demonstrate an FPGA-based hardware acceleration engine for the DNA basecalling task in collaboration with the Canadian Food Inspection Agency (CFIA) suitable for field-based biohazard detection applications.

II. General Description

Basecalling is an essential stage in the DNA sequencing process that consist of many physical and computational steps. As illustrated in Fig. 1, step 5, the basecalling step, of this pipeline converts physical measurements of DNA input snippets into a text representation of the equivalent succession of base molecules, adenine (A), cytosine (C), guanine (G), and thymine (T). The extracted text sequence identifies the so-called *primary structure* of the measured molecule.

DNA sequencing has made staggering leaps over the last four decades. Since Frederick Sanger's breakthrough in the area [1] throughputs have gone from $\sim 10^2$ base pairs (bp) per day; to

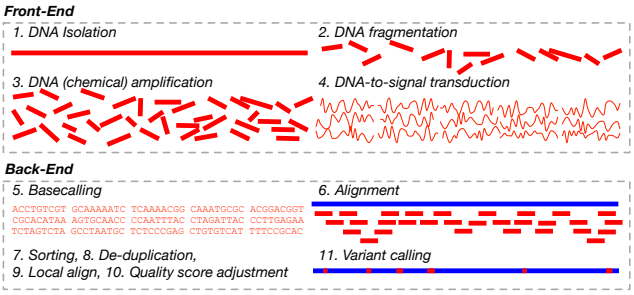

Fig. 1. DNA Sequencing Pipeline

10^6 bp/s using so-called next-generation sequencing machines (NGS). For instance, the $1M HiSeq X machine developed by market leader Illumina is able to reach 6×10^6 bp/s the equivalent of 7.5 human genomes per hour within a 200-kg, 1,500-W footprint.

Recent advances in semiconductor and sensing technologies [2] have lead to stunning miniaturization such as the hand-sized MinION sequencer [3]. This small device can ideally process (step 4 in Fig. 1) over 500 separate DNA samples in parallel and thus measure equivalent of 0.13 human genome per hour in a real-time streaming fashion while consuming 5 W. Such miniaturizations make the prospect of widely available on-field rapid genome analysis much more realistic. However, the acceleration achieved by this technology need to be matched by high performance computing (HPC) on the back-end of the sequencing system (*e.g.,* steps 5-11 in Fig. 1) to fully realize accelerated genomic analysis and interpretation.

In this work, we devoted our efforts to improve the computational efficiency of the basecalling task, in terms of speed and power, for a Hidden Markov Model (HMM)-based DNA sequencer by proposing a real-time FPGA-based hardware-accelerated engine.

A. System Overview

As illustrated in Fig. 2, the proposed FPGA-accelerated engine is a PC-based implementation

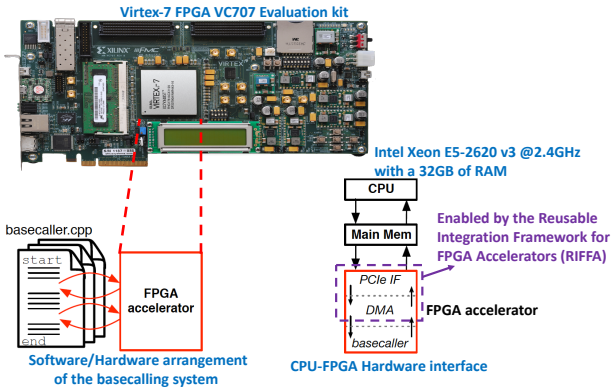

Fig. 2. Proposed FPGA-Accelerated Basecalling Engine

of the basecalling system. The basecaller processes multiple channel streams (*i.e.,* DNA fragments) of the measured data in a serial fashion (*i.e.,* one DNA fragment at a time).

The basecaller consists of a C++ program running on a PC (*i.e.,* Intel Xeon E5-2620 v3 @2.4GHz with a 32GB of RAM) and works cooperatively with an external fixed-point FPGA accelerator device (*i.e.,* Xilinx Virtex-7) over the PCIe bus. The C++ program encapsulates a simple API which offloads the HMM's intensive computations to the FPGA accelerator. The C++ program receives back the FPGA-computed results to finalize the DNA basecalled sequence on the CPU side. The PCIe link between the CPU and FPGA is enabled by the Reusable Integration Framework for FPGA accelerators (RIFFA) [4]. RIFFA provides the FPGA IP core with coherent access to the CPU's main memory via a direct-memory-access (DMA) bus master and PCIe endpoint.

B. Results

The results depicted in Fig. 3 compare the speed performance of our proposed FPGA-accelerated basecaller to that of the conventional CPU-based basecaller. It can be noticed from the upper plot that the proposed basecaller is outperforming the CPU-based basecaller by a factor of 1.46× at 60 MHz. In addition, the FPGA clock frequency demonstrates a marginal effect on the proposed basecaller's speed. This is due to our present use of a basic credit-base flow control scheme with small DMA buffer sizes (for development purposes) which achieves speeds of only ∼30-Mb/s. By expanding our DMA buffer size we shall realize the FPGA basecaller's inherent speed,

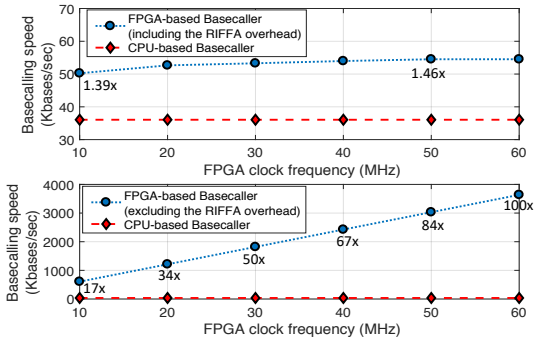

Fig. 3. Speed Performance for the Proposed Basecaller

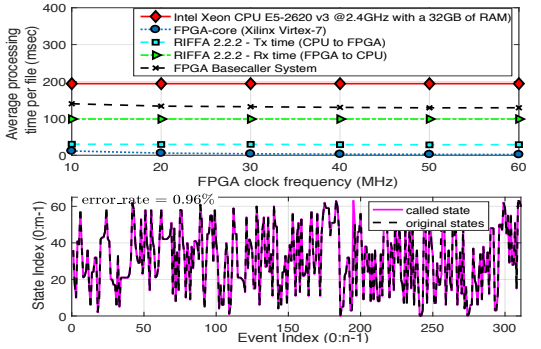

Fig. 4. RIFFA Transfer Speeds and Basecaller's Accuracy

which experimentally achieves 100× acceleration at a 60-MHz clock as shown in the bottom plot of Fig. 3. Further insights on the slow RIFFA communication speeds are portrayed in the top plot of Fig. 4 which reports the send and receive times over the RIFFA channel compared to the FPGA core processing time.

To achieve high utilization of the PCIe link bandwidth and attain faster speeds (as high as 3.64GB/sec), our streaming mechanism is in the process of being modified to convey more events in each transfer by increasing the aforementioned DMA buffers and multi-threading TX and RX channels. The FPGA IP core consumes 5.32 W (cf. 85-W CPU-based basecaller power). From another perspective, the bottom plot of Fig. 4 shows that our basecaller attains high level of accuracy with a 0.96% error rate.

REFERENCES

[1] F. Sanger et al., *Journal of Molecular Biology*, vol. 94, no. 3, pp. 441 – 448, May 1975.

[2] C. W. Fuller et al., *Proceedings of the National Academy of Science*, vol. 113, no. 19, pp. 5233 – 5238, May 2016.

[3] C. G. Brown et al., *Nature Biotechnology*, vol. 34, no. 8, pp. 810 – 811, Aug. 2016.

[4] M. Jacobsen et al., *ACM Trans. Reconfigurable Technol. Syst.*, vol. 8, no. 4, pp. 22:1–22:23, Sep. 2015.