Towards in the field fast pathogens detection using FPGAs

S. Casale-Brunet, T. Schuepbach, N. Guex, C. Iseli, A. Bridge, D. Kuznetsov, C. Sigrist, P. Lemercier, I. Xenarios, E. Bezati
SIB, Swiss Institute of Bioinformatics, Switzerland

***Abstract*—The rapid detection and identification of infectious pathogens are a critical need for healthcare in both developed and developing countries. There is a need for portable diagnostics and epidemiological surveillance systems that can be deployed in remote field settings for infectious diseases testing, such as pandemic "hotspots" in Africa or southeast Asia. Together with the Vital-IT and Swiss-Prot groups, we developed a fast pathogen detection embedded hardware accelerator (no bigger than 75x50mm) capable of analyzing up to 250Mnt/s of DNA sequence. To deploy this solution, an embedded DNA Next Generation Sequencer (NGS) is used for retrieving the genetic material from patients sample at the point of care (e.g., the Oxford Nanopore Technologies MinION). Successively, the collected DNA sequences are analized in order to detect the presence of known pathogen patterns. Currently, up to 100 different pathogens can be analized at the same time using one single embedded accelerator. Due to the low-power consumption of this hardware accelerator, it is possible to give access to worldwide health organizations and local administrations to a complete real-time embedded and battery powered surveillance solution. Moreover, this solution can greatly scale to big computing clusters, where a larger pathogens number can be analized in parallel, outperforming current state of art computationally expensive solutions.**

***Keywords*-pathogens detection; FPGA; genomics**

I. Introduction

The in-development pathogen detection pipeline is composed of three different stages. As depicted in Fig. 1, the first stage consists of retrieving the metagenome from a patient's sample. Successively, a base-calling procedure is performed on a laptop: this consists on transforming the electric signals of the DNA sequencer machine onto a digital representation (e.g., Fasta or Fastq) of the DNA collected sequences. Successively, all the sequences are streamed through the HW portable accelerator for pathogens detection (HPAPD). A first version of the HPAPD has been implemented on a

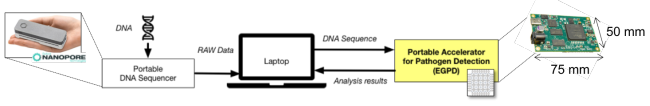

Figure 1: Portable accelerator for pathogens detection (HPAPD).

Xilinx Artix-7 XC7A200T FPGA running at 200MHz. The communication between the PC, used for streaming the DNA sequences and collecting the analysis results, is performed through a USB 3.0 integration module. The overall size of the HPAPD is no bigger than 75x50mm and it can be powered by a battery or through the USB power interface used for the data exchange. As illustrated in the following, the HW accelerator outperforms current state of art software (SW) implementation (up to 8 times on general purpose CPUs) and it can achieve a throughput of up to 250Mnt/s of analized DNA nucleotides.

II. Experimental results

The set of pathogens that need to be analized are encoded using a proprietary methodology. These are successively transformed to RTL using a specific compiler developed by authors. For each DNA sequence, the accelerator reports to the PC the number of occurrences of each pathogen through a single serial USB 3.0 interface. Hence, the aim of the following experiment has been not to evaluate the diagnostic accuracy of the pipeline (that is out of the scope of this work) but to measure the HPAPD performances (i.e., defined in terms of analized DNA nucleotides per second, nt/s) in its worst-case scenario. This can be summarized as when the maximum number of pathogens can be synthesized on the FPGA and streamed reads are short.

With this objective in mind, a set of 100 synthetic pathogens have been generated in order that a high number of pathogen occurrences can be found along the streamed sequences. Throughput evaluation has been done using two distinct set of DNA sequence reads. As illustrated in Table Ia, the first one is the assembled chromosomes of the GRCh37 reference human genome assembly and the second one is an example of short reads sequenced using an

Table I: DNA sequences and throughput performances.

(a) Input sequences and size distribution.

Input	Data size (Mnt)	Sequences	Length (nt)			
			total	average	min	max
GRCh37	2994.45	24	3095677412	128986558	48129895	249250621
W303	2339.44	410344	2392848698	5831	5	191145

(b) Throughput and execution time performances.

Pathogens	Sequence	Occurrences	Execution time (s)		Throughput (Mnt/s)		Speed-up
			SW	HW	SW	HW	
10	GRCh37	10233804	82.22	12.13	36.42	246.83	6.78
	W303	7194869	61.62	28.24	37.97	82.84	2.18
20	GRCh37	10235087	84.46	12.12	35.45	**247.06**	6.97
	W303	7195251	64.34	28.73	36.36	81.43	2.24
50	GRCh37	10628722	88.76	12.26	33.74	244.35	7.24
	W303	7687152	72.41	28.21	32.31	82.92	2.57
100	GRCh37	21257444	93.07	12.18	32.17	245.94	**7.64**
	W303	15374304	74.06	28.11	31.59	83.22	2.63

ONT MinION sequencer. Performances have been compared with the SW implementation running on an Intel Kaby Lake i7-8550U CPU. As it can be seen from Table Ib, the HW accelerator outperforms (up to 7.64 times) the multi-threaded SW implementation and it can achieve an analysis throughput of up to c.a. 250Mnt/s for long reads and up to c.a. 85Mnt/s for short reads sequences.

ACKNOWLEDGMENT

This project has been partially founded by the Fonds National Suisse (FNS) BRIDGE program.