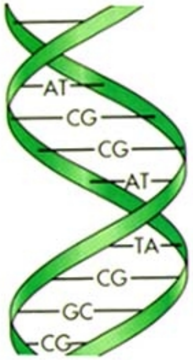
Dhruv Jain

Prof. M. Balakrishnan

FPGAS
As accelerators for
NEXT-GENERATION SEQUENCING (NGS)
Applications

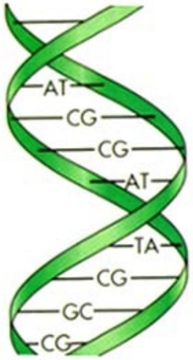


Brief Introduction

- Sequence Alignment
  - Discipline of Bioinformatics which concerns itself with arranging sequences of DNA, RNA or protein to identify regions of similarity.
  - Usually a small query seqeunce aligned against large reference gnome

- Next Generation Sequencing
  - Reads small pieces between 20 and 1000 bases, depending on the technology used.
  - So, there is a need to speed up alignment to save processing time

- Use of FPGAs to achieve parallelization of sequence alignment algorithms



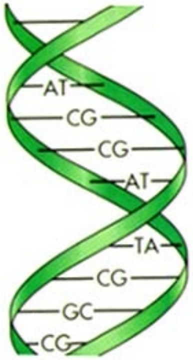
Methodology

- **Profiling**:
  - Code of the open source software tools obtained from web
  - Profiled on runnning on real data. Compute intensive kernels have been identified

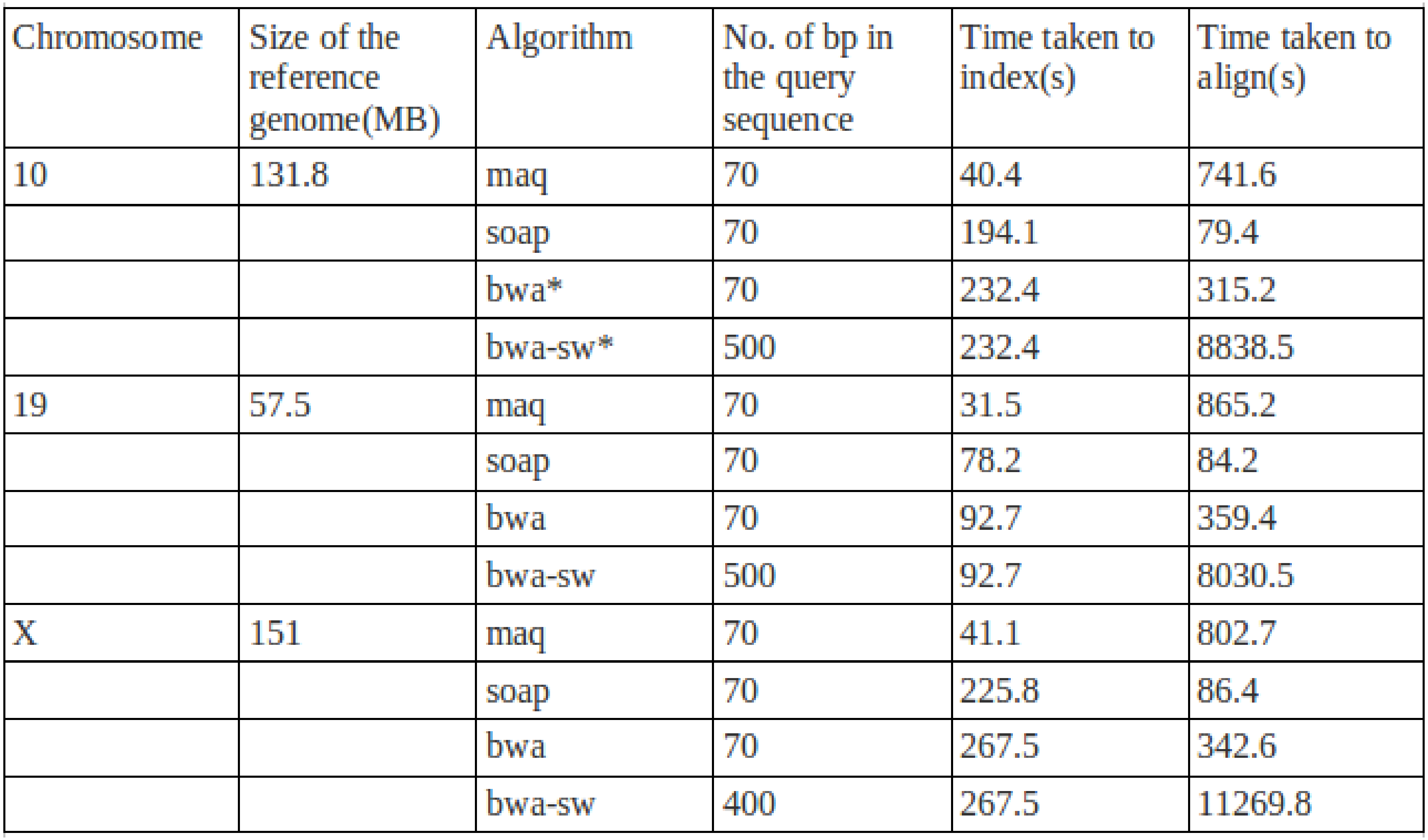
- **High Level Performance Estimation**:
  - Rough estimate of the resources consumed by the kernel on the FPGA
    - Estimation of the number of functional units to exploit parallelism.
    - Estimation of communication interface and memory hierarchy

- **Selection of desired algorithm**
  - Based on the performance estimation

- **Hardware-Software Codesign**
  - Code divided into specfic parts to execute on FPGA; Processor

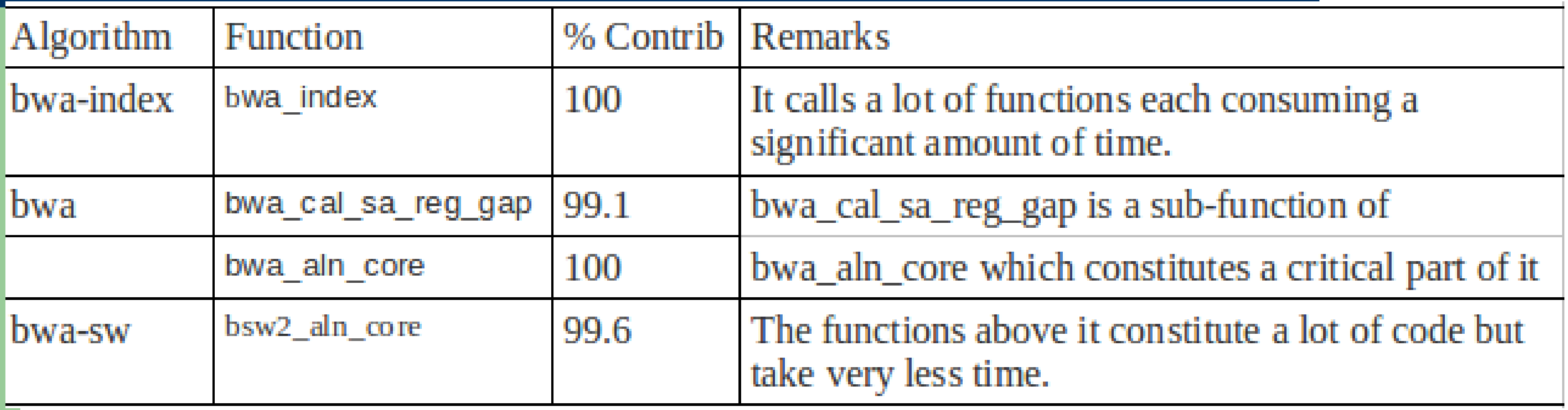


Profiling Results



Chromosome	Size of the reference genome(MB)	Algorithm	No. of bp in the query sequence	Time taken to index(s)	Time taken to align(s)
10	131.8	maq	70	40.4	741.6
		soap	70	194.1	79.4
		bwa*	70	232.4	315.2
		bwa-sw*	500	232.4	8838.5
19	57.5	maq	70	31.5	865.2
		soap	70	78.2	84.2
		bwa	70	92.7	359.4
		bwa-sw	500	92.7	8030.5
X	151	maq	70	41.1	802.7
		soap	70	225.8	86.4
		bwa	70	267.5	342.6
		bwa-sw	400	267.5	11269.8

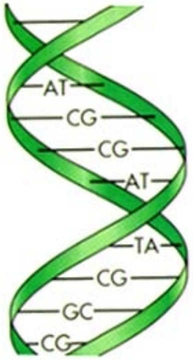
Profiling Results of BWA



Algorithm	Function	% Contrib	Remarks
bwa-index	bwa_index	100	It calls a lot of functions each consuming a significant amount of time.
bwa	bwa_cal_sa_reg_gap	99.1	bwa_cal_sa_reg_gap is a sub-function of
	bwa_aln_core	100	bwa_aln_core which constitutes a critical part of it
bwa-sw	bsw2_aln_core	99.6	The functions above it constitute a lot of code but take very less time.

Chose BWA because:

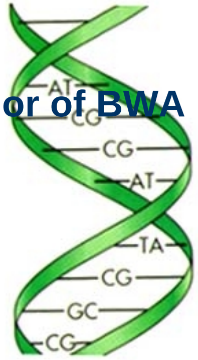
- Likely to show an increase on FPGA
  - BWT already implemented and showed 52% improvement in time (Martinez et. al)
  - Evident from the above results (Amdahl's law)
- Popular (highly cited and widely used)



Why BWA over SOAP?

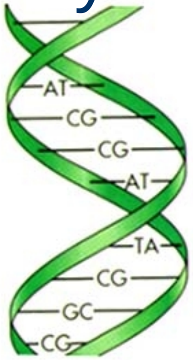
“People choose bowtie and bwa more often probably because both ***natively support the SAM output****, while soap2 not. Bowtie is often seen in RNA-seq/ChIP-seq because it is extremely fast for single-end reads and because the whole tophat/cufflink package is very useful. BWA is often seen for SNP/indel calling because it does gapped alignment and produces fewer false alignment. BWA/stampy/novoalign* ***estimate mapping quality*** *which is at times useful. Bowtie/soap2 do not, which is why they are faster”*

- Heng Li, Author of BWA



BWA Algorithm in brief

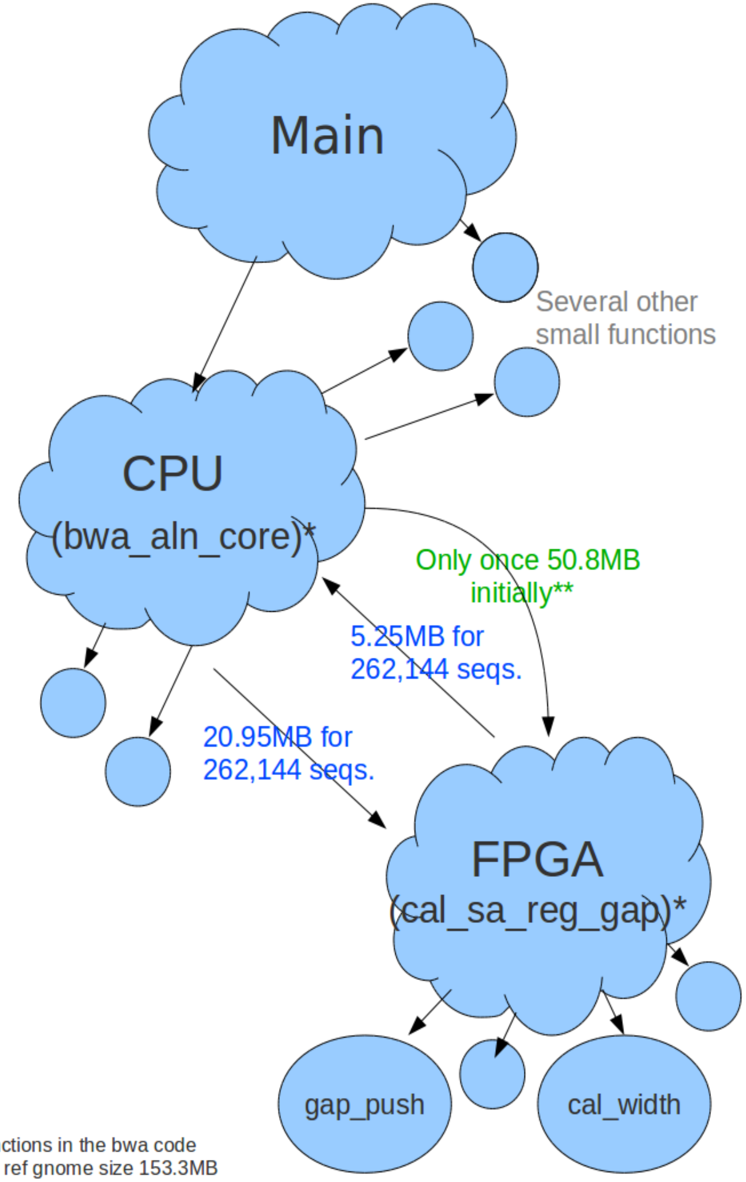
- Constructs a **suffix array interval** (BWT) from ref gnome and aligns short query sequence to it (Li, Durbin, Sanger-UK, 2009)
- Initial construction of suffix array takes large time linear in length of query sequence
- Alignment time linear w.r.t the length of query sequence using **backward search**



Data Flow Graph

For 150MB gnome and 50bp query sequences:

- Time taken in each iteration = 0.32ms
- Mem required **α** size of ref. seq.

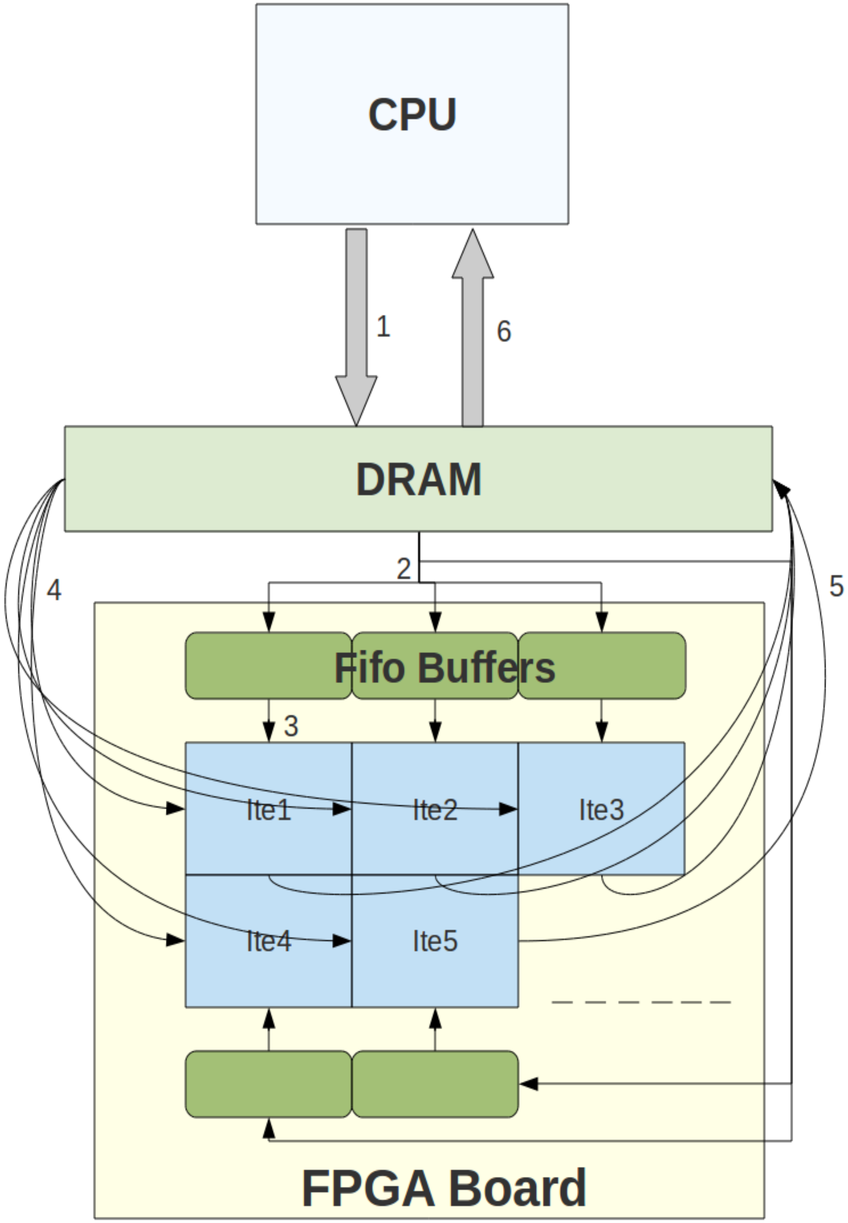


* functions in the bwa code
**for ref gnome size 153.3MB

Figure1: Data flow graph of bwa-software tool

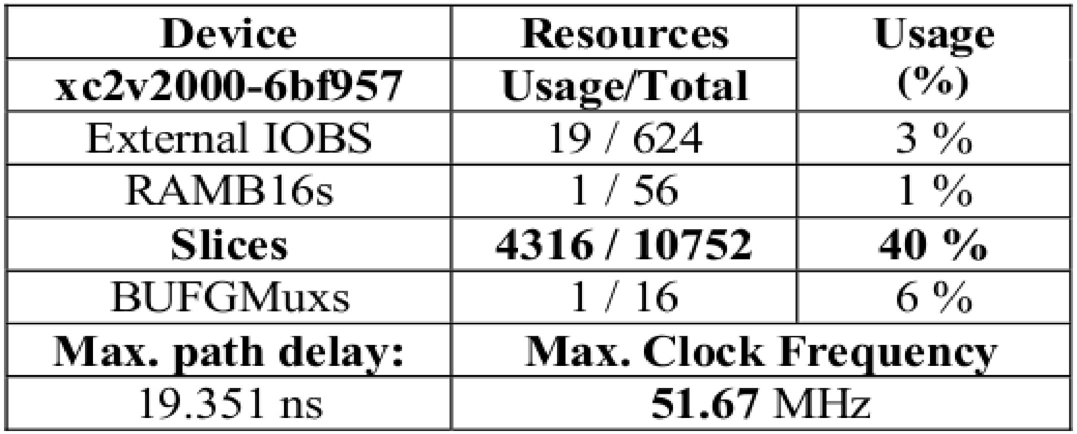
Hardware Software Co-design

1. CPU sends the ref. gnome and the query seq to the DRAM memory.
2. Ref. gnome is transfered from DRAM to all the FIFO buffers simultaneously
3. Ite blocks read the ref. gnome data part by part from the FIFO Buffers
4. Ite blocks read query sequence from DRAM via BRAM
5. The Ite block process data and send result back to the DRAM via BRAM
6. CPU receives the output data from DRAM



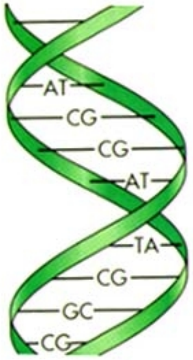
BWT Algorithm (for indexing)

- Already implemented on FPGA
- Enchaned by converting serial data input to parallel



Device	Resources	Usage (%)
xc2v2000-6bf957	**Usage/Total**	
External IOBS	19 / 624	3 %
RAMB16s	1 / 56	1 %
Slices	**4316 / 10752**	**40 %**
BUFGMuxs	1 / 16	6 %
Max. path delay:	**Max. Clock Frequency**	
19.351 ns	**51.67** MHz	

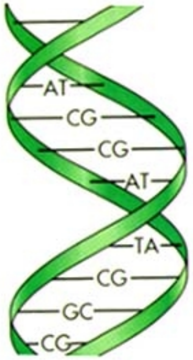
PLACE AND ROUTE REPORT



Parallel Alignment

- Currently in Implementation
  - According to Figure 2

- Projected Speedup
  - Depends on no. of bp in ref and query seq to a great extent
  - For a million bp ref Seq, 50bp query seq, **expected ~25**
  - Dominated by **Mem Bandwith** (DRAM) – 12.8Gbps
  - Dominated by **Size of Mem** on board and not by area
    - BRAM: 38.3MB (XC6V5X576T - largest available)
    - Distr. RAM: 7.6MB



Conclusion and Future Work

- ✓ Provided insight into Hardware-Software codesign of sequence alignment software tools which will speedup entire chain of its applications like biological research, diagnostic, biotechnology, forensic biology and biological systematics

- ✓ Work can be **compared with other possible hardware implementations** like GPUs for performance enhancement