Dataflow acceleration of Smith-Waterman with Traceback for high throughput Next Generation Sequencing

Konstantina Koliogeorgi*, Nils Voss†, Sotiria Fytraki †,

Sotirios Xydis*, Georgi Gaydadjiev †, Dimitrios Soudris*

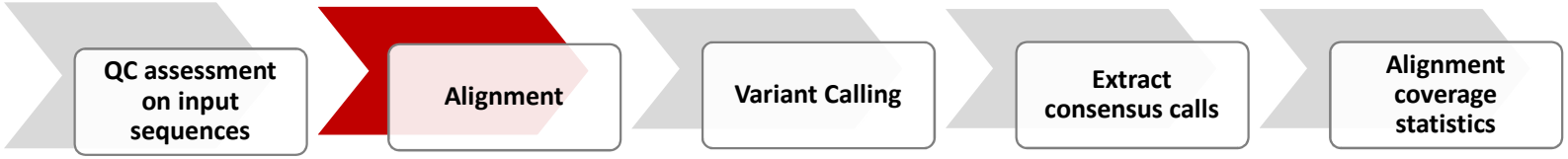
** National Technical University of Athens, Greece, {konstantina, sxydis, dsoudris}@microlab.ntua.gr*

Maxeler Technologies UK {nvoss, sfytraki, georgi}@maxeler.com

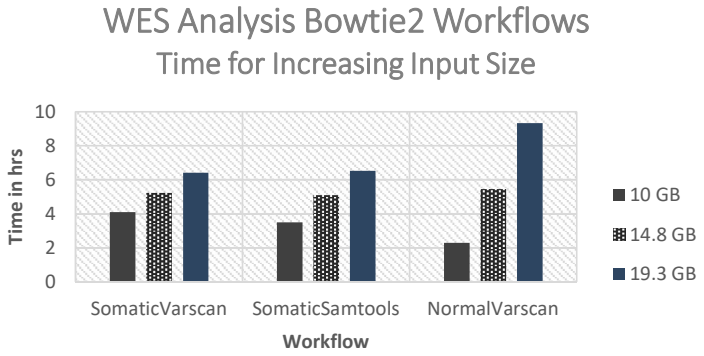
FPL 2019 Conference

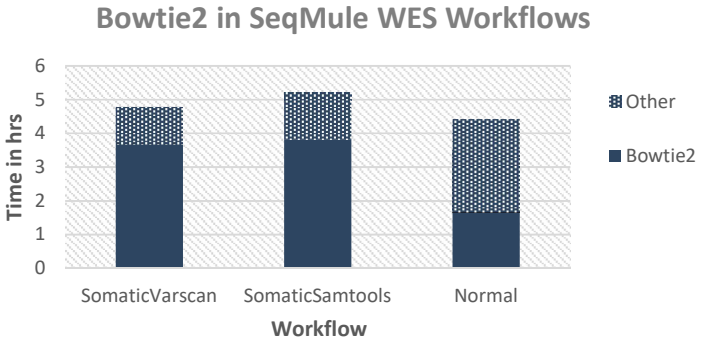
Genome Sequencing

- Genome represents entire genetic information of an organism
- Next-Generation Sequencing technologies allow to compare individual to reference genome
- Typical genomic workflow e.g. SeqMule
  - short read alignment: reads ~100 bases long



- Operate on huge amount of data



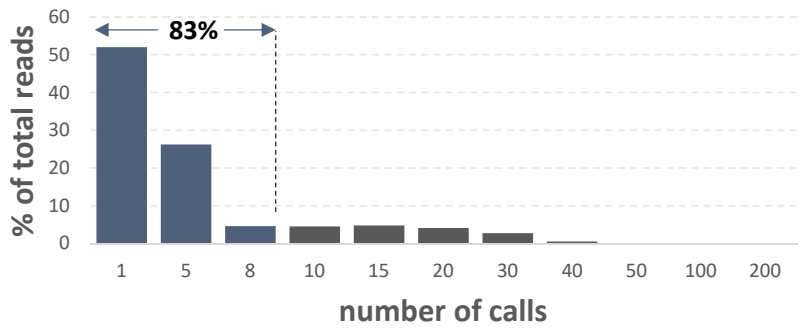


- Aligners Bottleneck in Workflow => in need of acceleration!

Problem Statement

- Most Aligners utilize Seed & Extend Model
  - Fragment reads into short pieces (seeds) that align exactly to genome
  - Extend seeds to full alignment with SmithWaterman

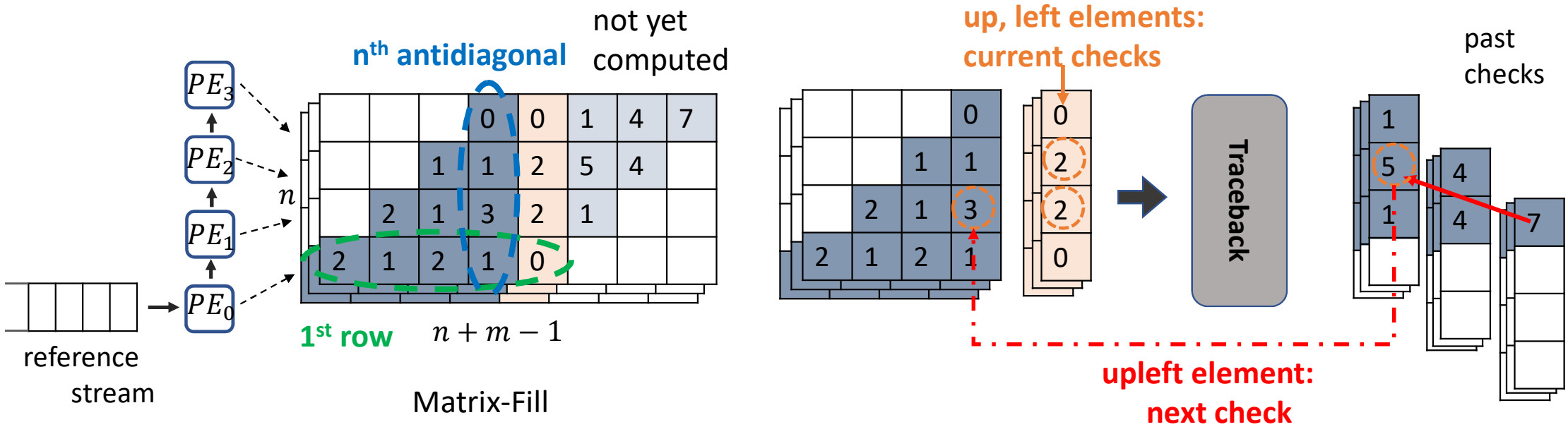
- SmithWaterman
  - Matrix Fill Stage followed by Traceback
  - Takes up 60% (55% + 5% respectively) of total time
  - Distributed over hundreds of tasks per read
  - **calling & data transfer** overhead



- Challenge
  - Co-designed Solution to avoid overhead
  - Extract parallelism to further boost performance

Standalone Optimized Dataflow Implementation

- Matrix Fill Calculates Matrices E,H,F
- Traceback traverses matrices in reverse order to construct alignment path

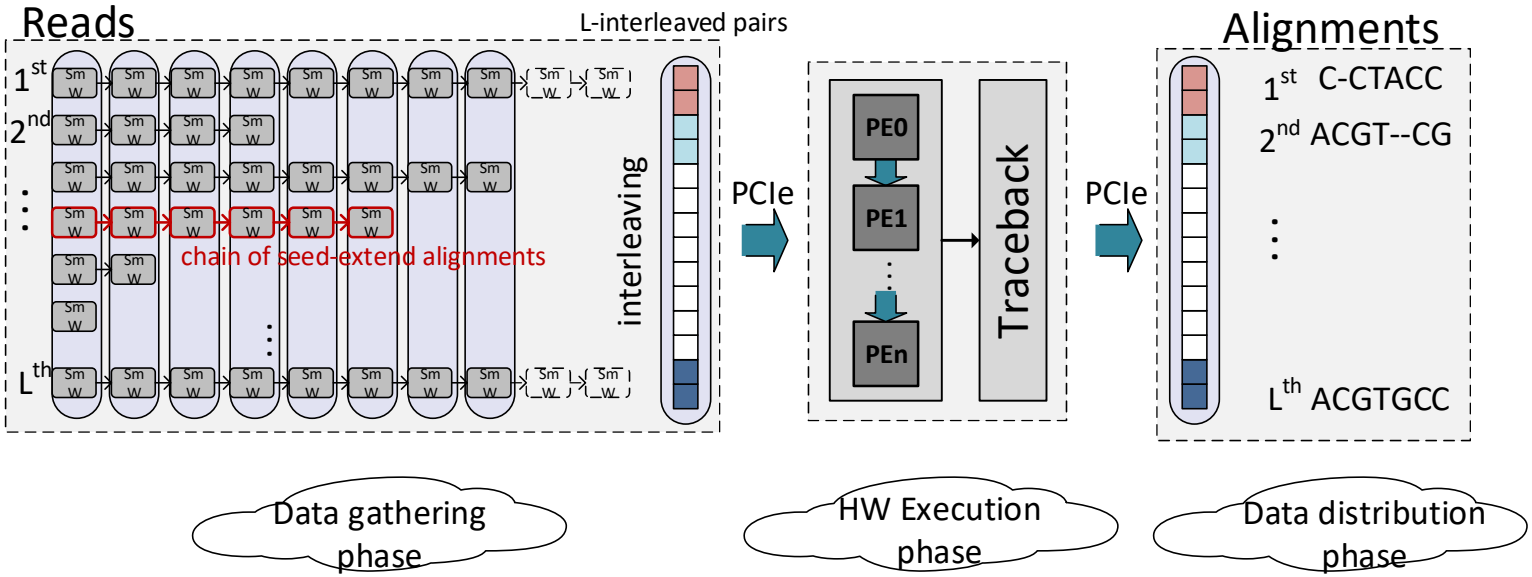


- Interleaving Data Scheme
  - Interlace data from subsequent read-reference pairs
- Double Buffering
  - operate in pipeline fashion

Proposed Integration Architecture

Key Architectural Decisions

- Move Traceback on Hardware to alleviate transfer cost
- Major Software Restructure to constraint number of accelerator calls



Results

- x18 speedup standalone
- x1,55 speedup end to end

Thank you for your attention!