Leveraging Succinct Data Structures for DNA Sequence Mapping on FPGA

27th IEEE Reconfigurable Architecture Workshop, May 18th 2020

Guido Walter Di Donato *guidowalter.didonato@mail.polimi.it*

Alberto Zeni *alberto.zeni@mail.polimi.it*

Lorenzo Di Tucci *lorenzo.ditucci@polimi.it*

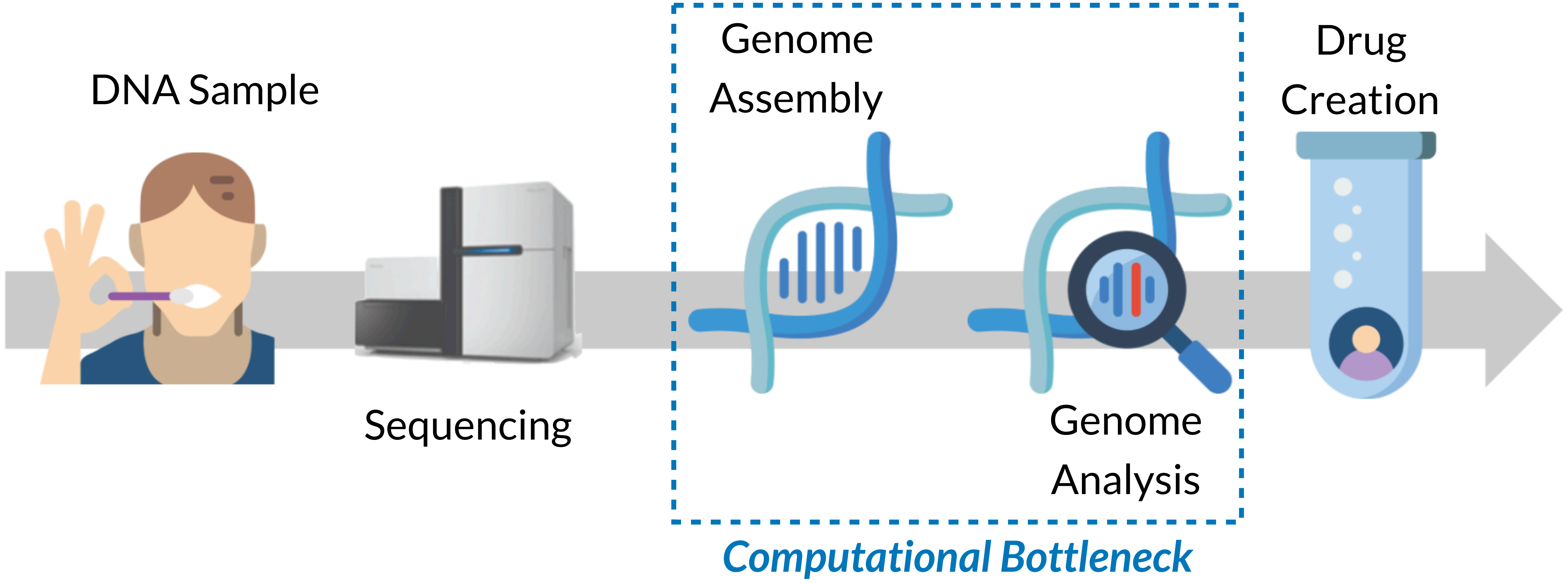
Marco D. Santambrogio *marco.santambrogio@polimi.it*

POLITECNICO MILANO 1863

POLITECNICO MILANO 1863 NECST laboratory

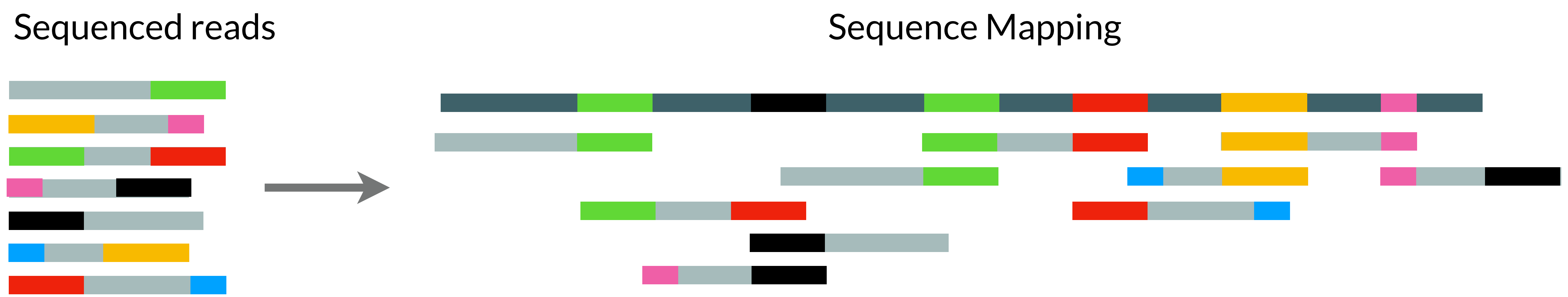
Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB) - Politecnico di Milano

DNA Analysis for Precision Medicine



Sequence Mapping

Sequence mapping is a computational intensive step involved in **genome assembly** and **genomic analysis** pipelines, leading to **long execution times** and **high computational costs.**[1]



[1] Beretta, S.. Algorithms for strings and sequences: Pairwise alignment. *Encyclopaedia of Bioinformatics and Computational Biology.* Oxford: Academic Press, 2019.

POLITECNICO MILANO 1863

POLITECNICO MILANO 1863 NECST laboratory

RICHARD AND LOAN HILL DEPARTMENT OF BIOENGINEERING COLLEGES OF ENGINEERING AND MEDICINE

Sequence Mapping

Sequence mapping is the **computational bottleneck** of many genome assembly and genome analysis pipelines:

- **Long execution time**
- **High computational cost**

[1] Beretta, S.. Algorithms for strings and sequences: Pairwise alignment. *Encyclopaedia of Bioinformatics and Computational Biology.* Oxford: Academic Press, 2019.

Burrows-Wheeler Transform

Reversible string permutation that can be searched directly and has long strings of repeated characters (**great for compression**).

TACGACGTCGACT$ →

1	**TACGACGTCGACT$**
2	ACGACGTCGACT$T
3	CGACGTCGACT$TA
…	…
13	T$TACGACGTCGAC
14	$TACGACGTCGACT

Sort each row by their prefixes in lexicographic order

14	$TACGACGTCGACT
2	ACGACGTCGACT$T
5	ACGTCGACT$TACG
…	…
1	**TACGACGTCGACT$**
8	TCGACT$TACGACG

←

BWT output

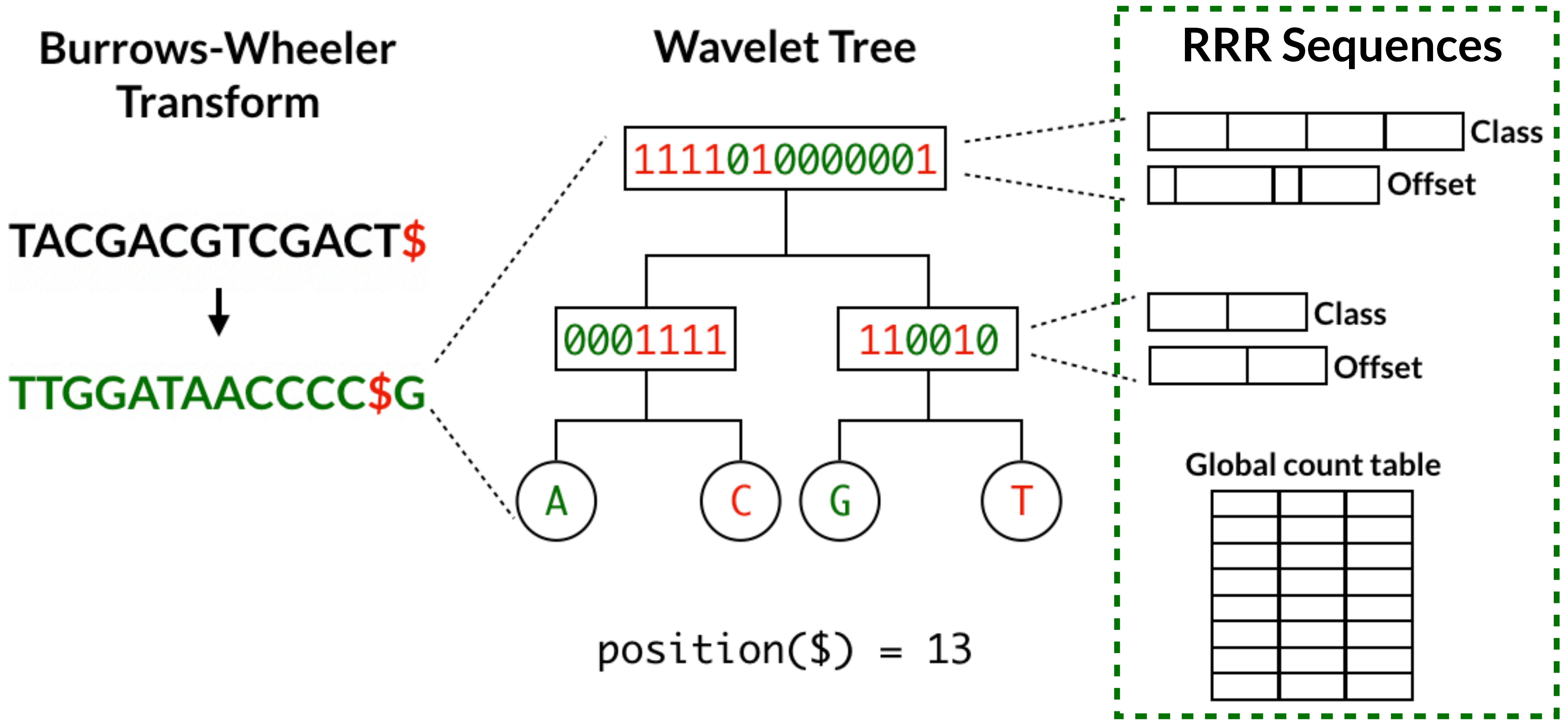
TTGGATAACCCC$G

14-2-5-11-3-9-6-12-4-10-7-13-1-8

Suffix array

BWaveR Data Structure

A flexible combination of **succinct data structures**, namely ***Wavelet Tree*** and ***RRR Sequences***, able to store long sequences in an **amount of space "close" to the theoretic lower bound**, while still allowing for **efficient rank query operations**.



Parameters: ***block size*** & ***superblock factor***

S., Beller, T., Moffat, A., and Petri, M.: From theory to practice: Plug and play with succinct data structures, 2014

Backward Search on BWaveR Data Structure

INITIALIZATION

$start = C(X) + 1$

$end = C(X+1)$

TTGGATAACCCC$G

C(X)	$	A	C	G	T
	0	1	4	8	11

ITERATIONS

end < pos($)

$start = C(X) + rank_{WT}(X, start-1) + 1$

$end = C(X) + rank_{WT}(X, end)$

start < pos($) ≤ end

$start = C(X) + rank_{WT}(X, start-1) + 1$

$end = C(X) + rank_{WT}(X, end-1)$

start ≥ pos($)

$start = C(X) + rank_{WT}(X, start-2) + 1$

$end = C(X) + rank_{WT}(X, end-1)$

Backward Search on BWaveR Data Structure

INITIALIZATION

$start = C(X) + 1$

$end = C(X+1)$

ITERATIONS

end < pos($)

$start = C(X) + rank_{WT}(X, start-1) + 1$

$end = C(X) + rank_{WT}(X, end)$

A pattern of **length *p*** is found in **O(*p*) time, independently of the reference size!**

start < pos($) ≤ end

$start = C(X) + rank_{WT}(X, start-1) + 1$

$end = C(X) + rank_{WT}(X, end-1)$

TTGGATAACCCC$G

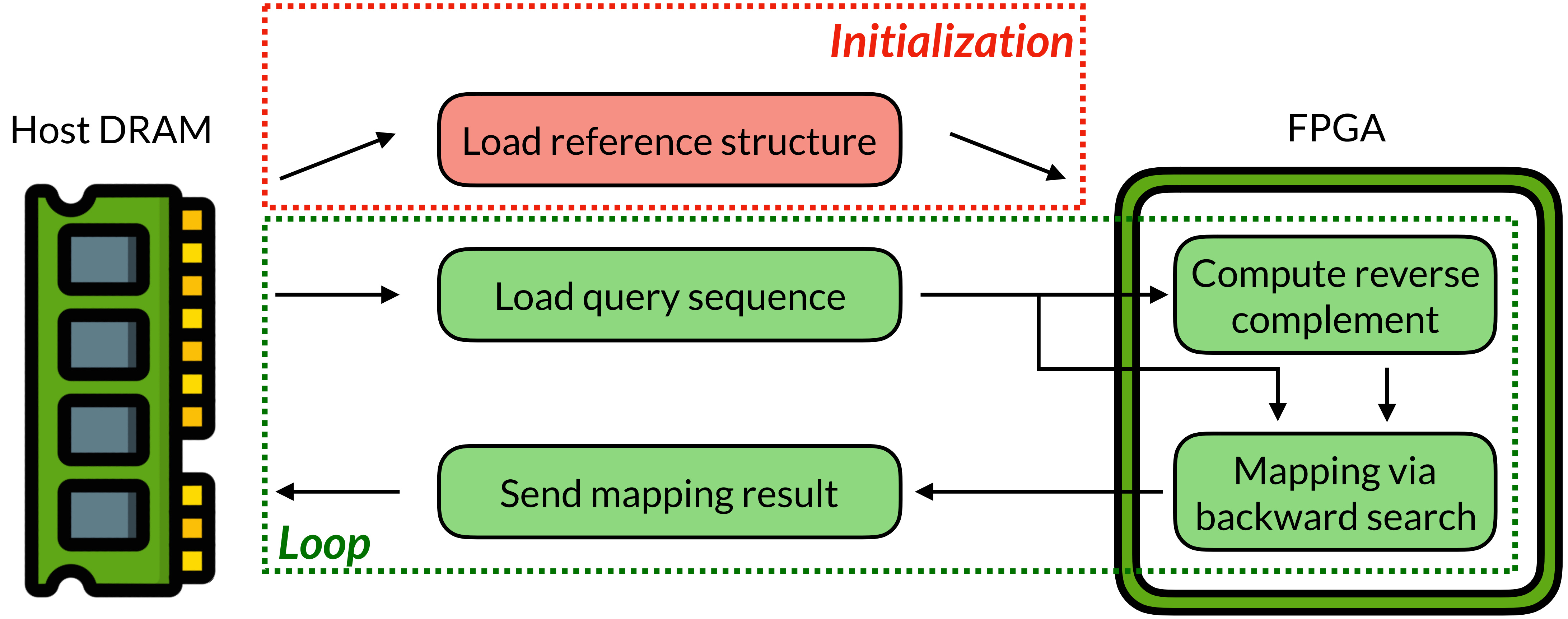
C(X)	$	A	C	G	T
	0	1	4	8	11

start ≥ pos($)

$start = C(X) + rank_{WT}(X, start-2) + 1$

$end = C(X) + rank_{WT}(X, end-1)$

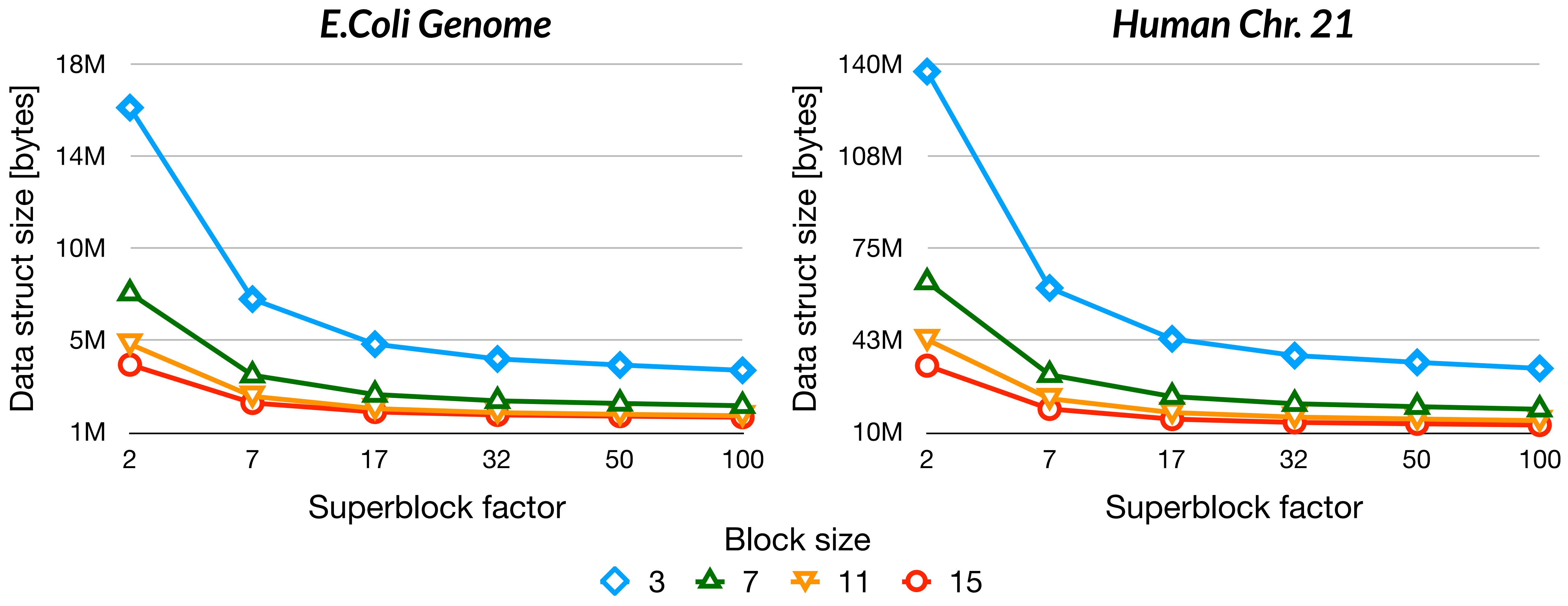
Custom Sequence Mapping Architecture



FPGA-tailored optimizations:

1. Reference structure stored in **BRAM** to **reduce memory access time**
2. Dimensioned query sequences structures to **exploit memory burst**
3. **Parallel mapping** of each query sequence and its reverse complement
4. Implemented rank queries using a ***reduction*** strategy

Results - Memory Footprint



Memory reduction up to

63.7% w.r.t BWT sequence

98.1% w.r.t naive Occ matrix

Memory reduction up to

68.3% w.r.t BWT sequence

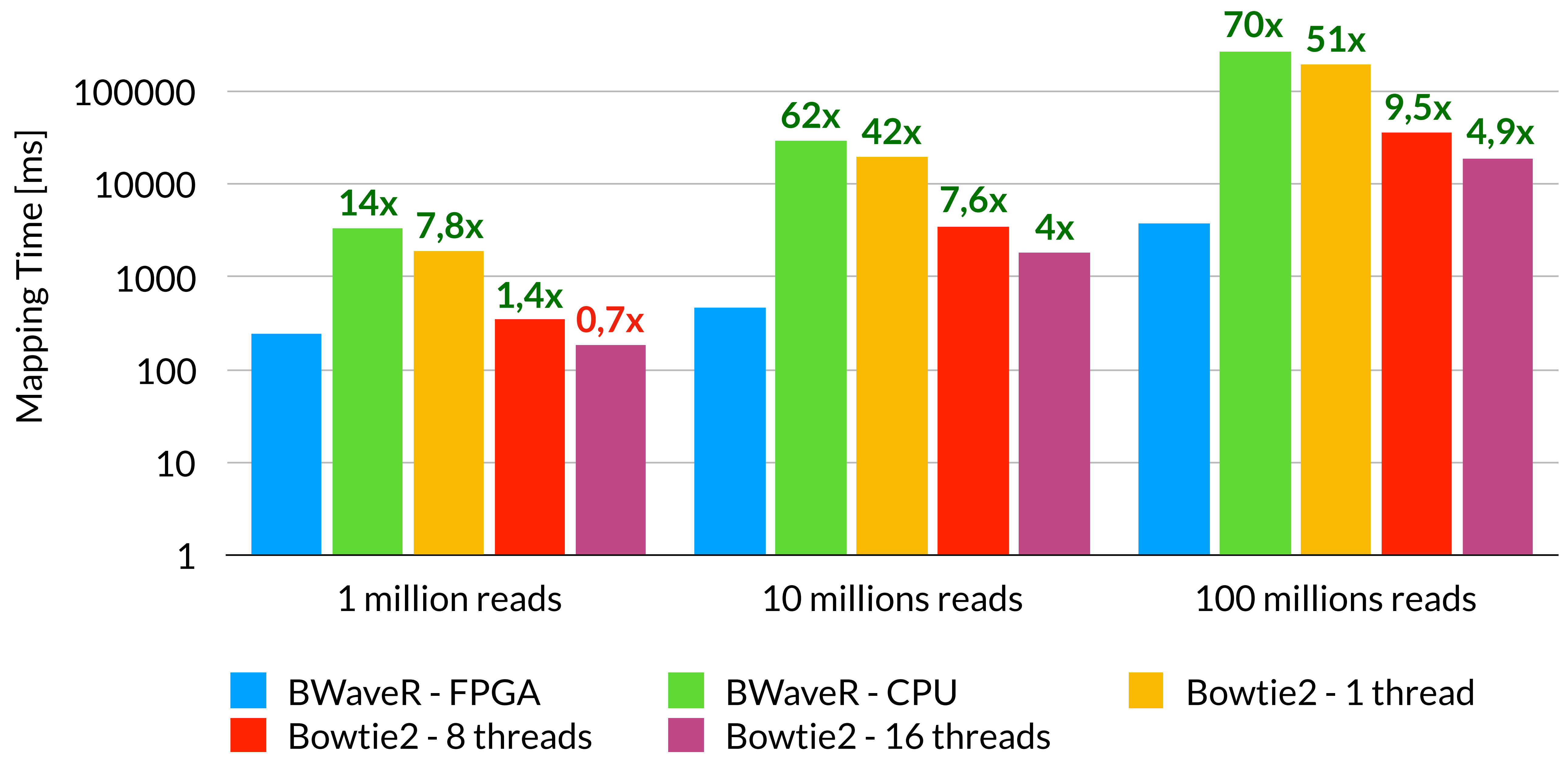
98.4% w.r.t naive Occ matrix

POLITECNICO
MILANO 1863

POLITECNICO MILANO 1863
NECST laboratory

Results - Mapping Time (HW)

Time for mapping 100 bp reads against Human Chr. 21 (*b* = 15, *sf* =50)

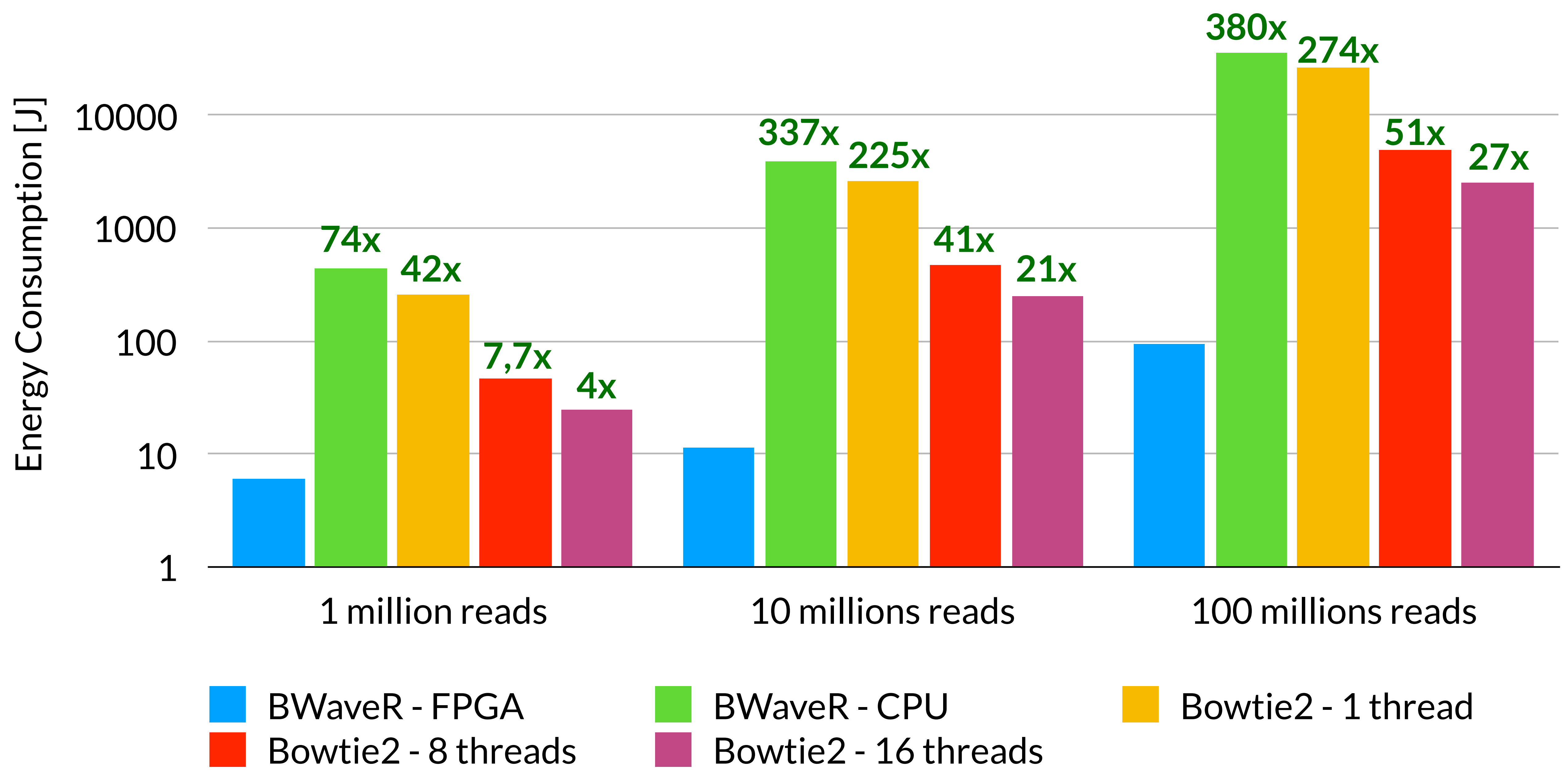


POLITECNICO MILANO 1863

POLITECNICO MILANO 1863 NECST laboratory

Results - Energy Consumption (HW)

Energy for mapping 100 bp reads against Human Chr. 21 (*b* = 15, *sf* =50)



POLITECNICO MILANO 1863

POLITECNICO MILANO 1863 NECST laboratory

Conclusions

This paper presents **BWaveR**, a **fast** and **memory efficient** sequence mapper leveraging **succinct data structures** and **HW acceleration on FPGA**:

- ➡ **Flexibility of usage**
- ➡ **Great compression capabilities**
- ➡ **Time independence from reference size**
- ➡ **Low energy consumption**
- ➡ **Low execution time**

BWaveR

For questions regarding this work, email:

Guido Walter Di Donato: ***guidowalter.didonato@mail.polimi.it***