# Exact Alignment with FM-Index on the Intel Xeon Phi Knights Landing Processor

Jose M. Herruzo[1] Sonia González-Navarro[1] Pablo Ibáñez[2]
Víctor Viñals[2] Jesús Alastruey-Benedé[2] Oscar Plata[1]

[1]Departamento de Arquitectura de Computadores
Universidad de Málaga

[2]Grupo de Arquitectura de Computadores
Universidad de Zaragoza

Accelerator Architecture in Computational Biology and Bioinformatics, 2018

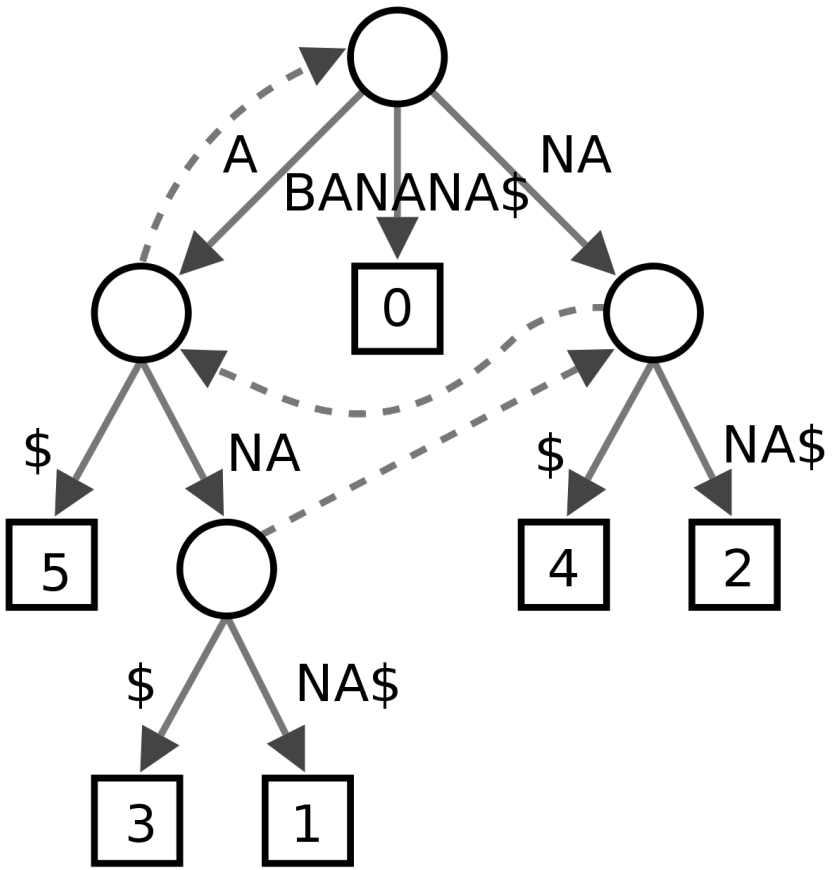
# Motivation

- Genomic Sequencing
  - New Sequencing Technologies

- Sequence alignment

# Motivation

## Indices




- Suffix Tree
- Hash Tables
- FM-Index

# Motivation
## BWT & FM-Index

- Burrows Wheeler Transform (BWT)

- FM-Index
  - Bowtie
  - BWA
  - SOAP2

banana$

| | |
|---|---|
| 0 | banana$ |
| 1 | anana$b |
| 2 | nana$ba |
| 3 | ana$ban |
| 4 | na$bana |
| 5 | a$banan |
| 6 | $banana |

# Outline

# Outline

# FM-Index

- Data structures
  - C array
  - Occ table
- Random memory accesses
- Memory bound algorithm
- Backward Search algorithm

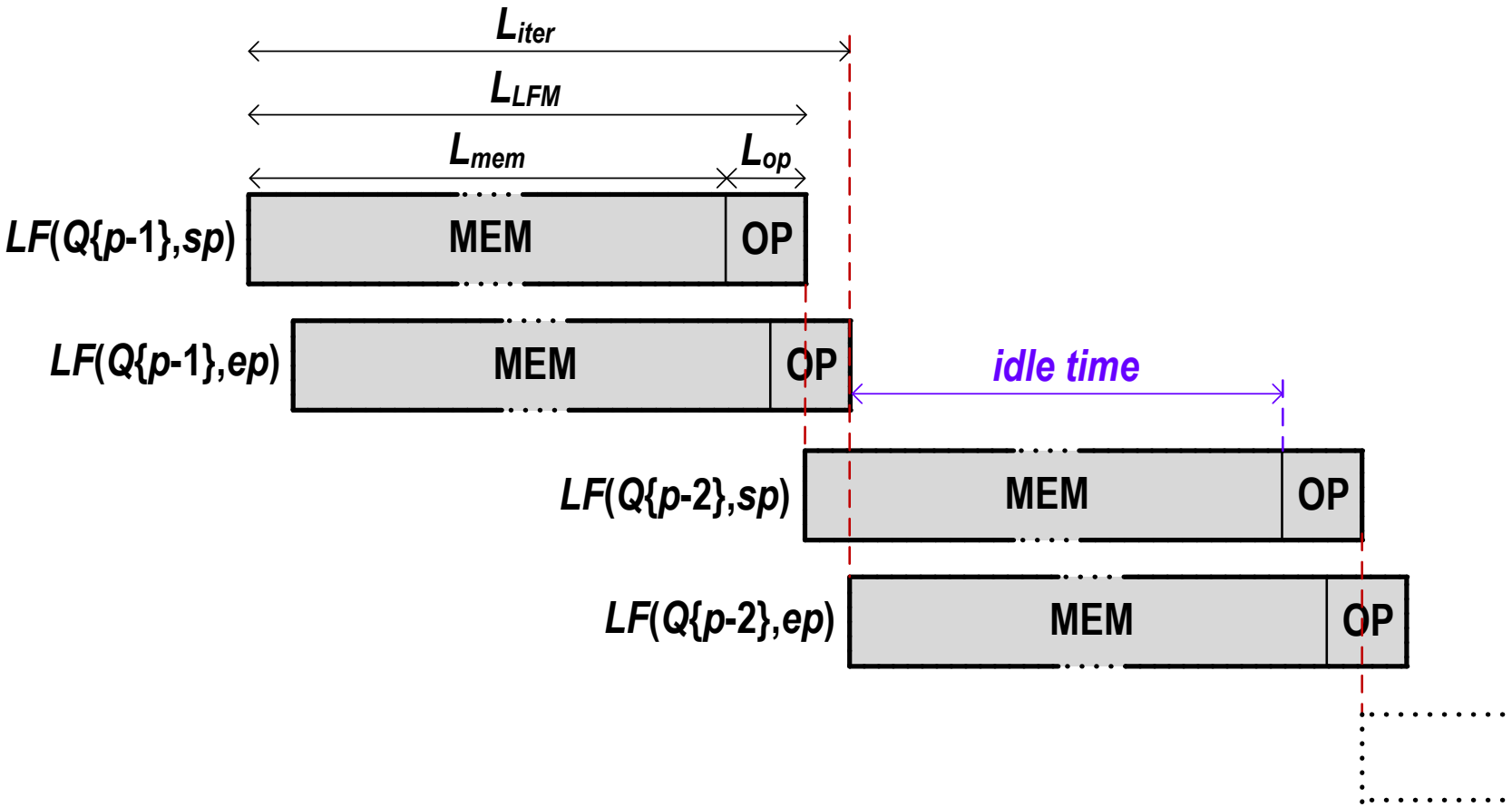
**Algorithm:** Backward Search Based on FM-index

**Input:** FM-index of $T$ text ($C$ & $Occ$), $Q$ query, $n$:$|T|$, $p$:$|Q|$
**Ouput:** ($sp$,$ep$): Interval pointers of $Q$ in $T$

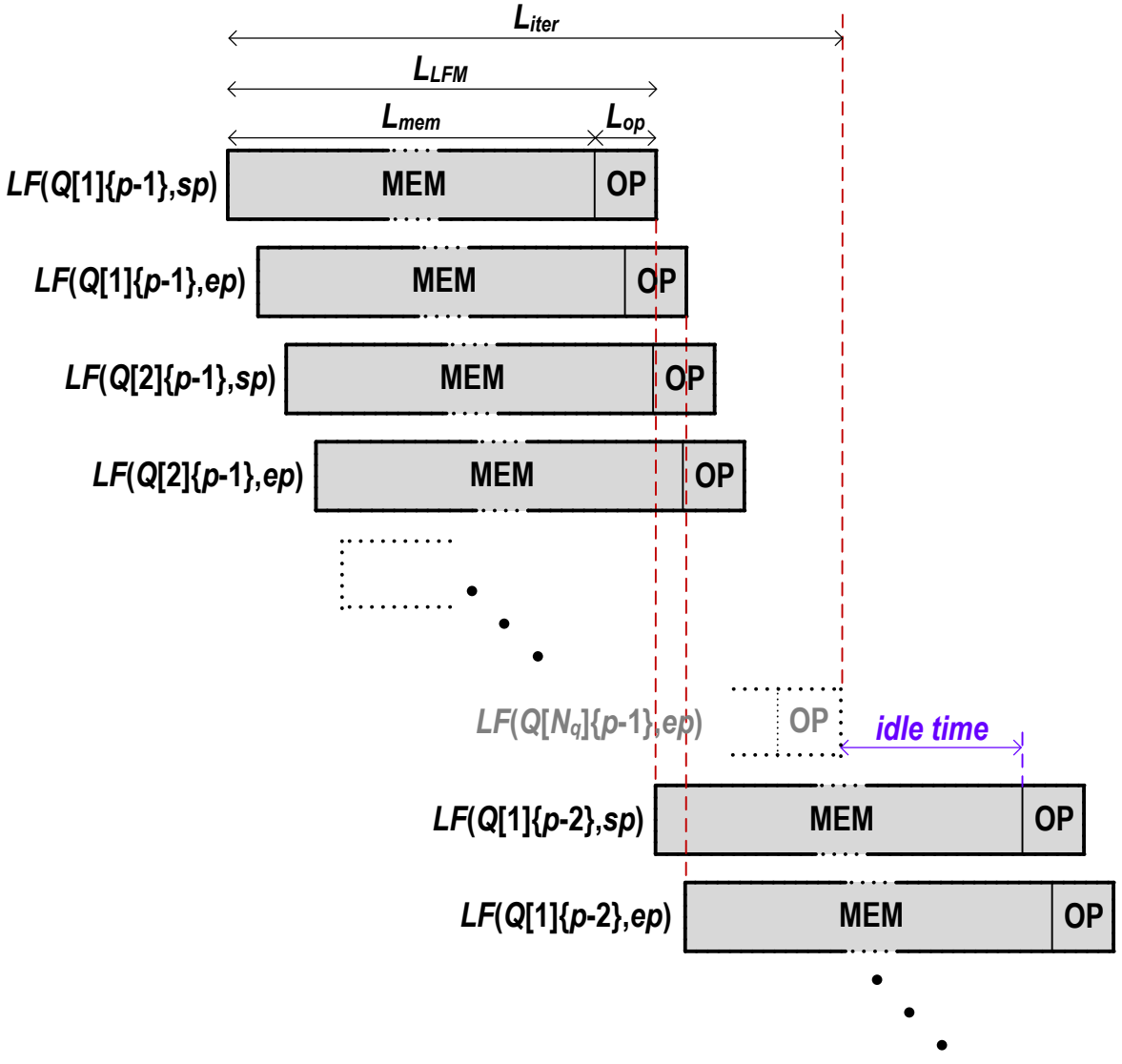
```
    begin
1:  sp = C[Q[p]]
2:  ep = C[Q[p]+1]
3:  for i from p-1 to 1 step -1
4:     sp = LF(Q[i],sp)
5:     ep = LF(Q[i],ep)
6:  end for
7:  return (sp+1,ep)
    end
```

2 LFop-chains

# FM-Index

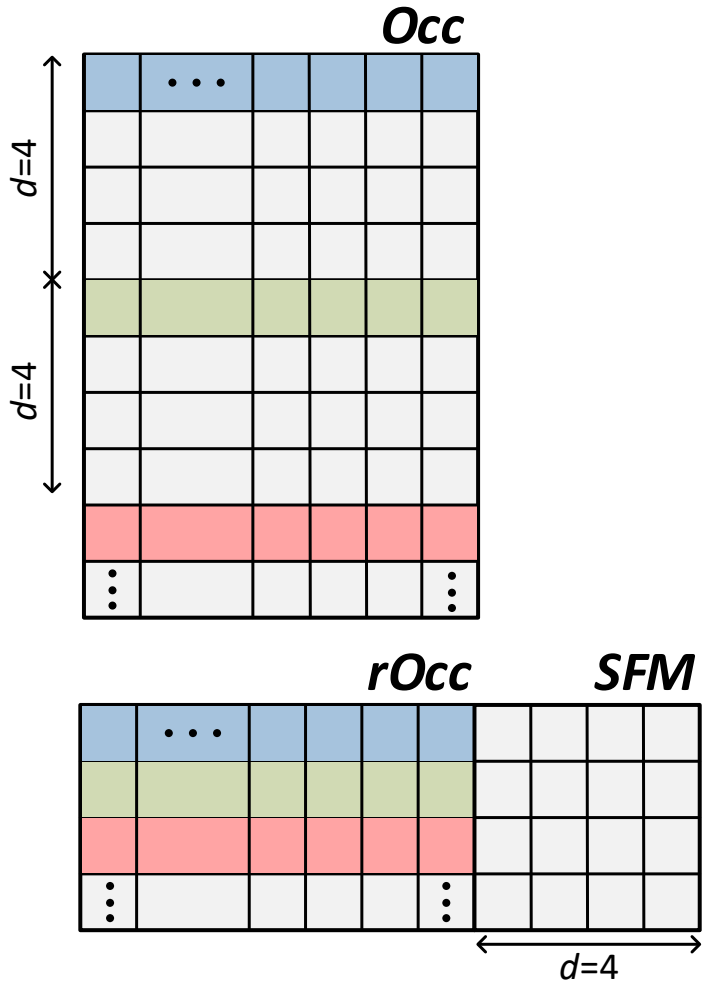

$L_{iter}$
$L_{LFM}$
$L_{mem}$
$L_{op}$
LF(Q{p-1},sp)
MEM
OP
LF(Q{p-1},ep)
MEM
OP
idle time
LF(Q{p-2},sp)
MEM
OP
LF(Q{p-2},ep)
MEM
OP

# FM-Index


$L_{iter}$
$L_{LFM}$
$L_{mem}$
$L_{op}$
LF(Q[1]{p-1},sp)
LF(Q[1]{p-1},ep)
LF(Q[2]{p-1},sp)
LF(Q[2]{p-1},ep)
LF(Q[Nq]{p-1},ep)
LF(Q[1]{p-2},sp)
LF(Q[1]{p-2},ep)
MEM
OP
idle time

# Outline

# Sampled FM-Index




Reduce Memory Footprint
↓
Increase computing requirements

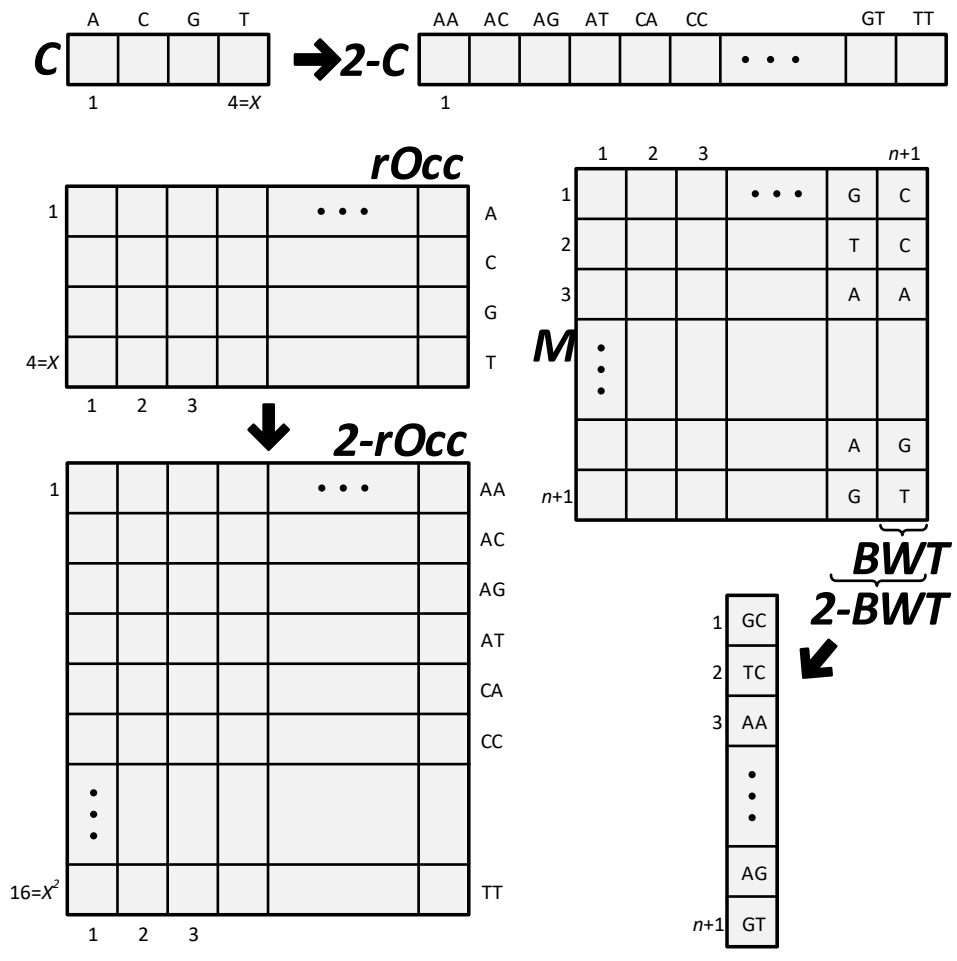
From Ferragina, P. and Manzini, G.: "Opportunistic Data Structures with Applications" (2000)

# Outline

# K-Step FM-Index

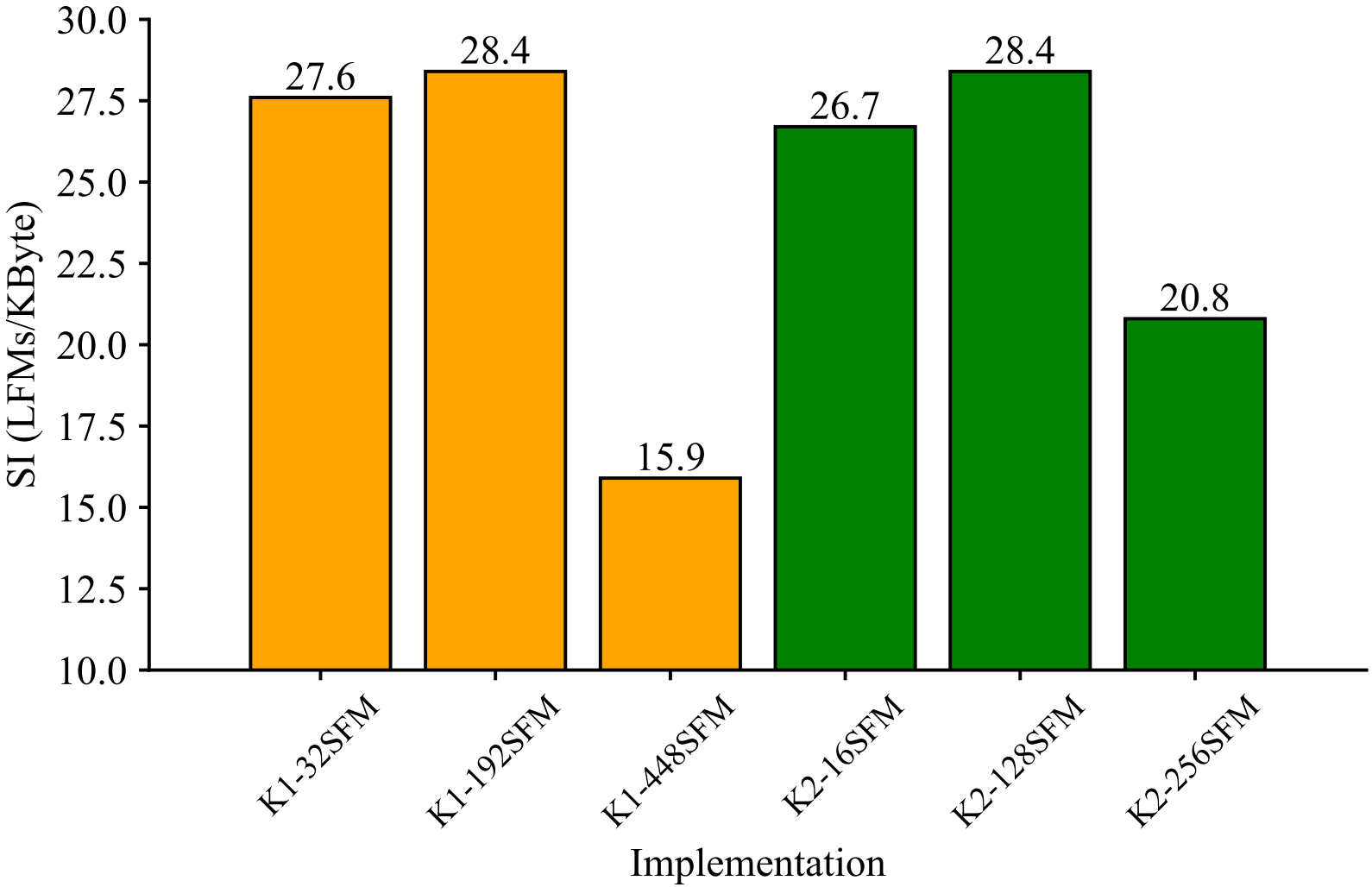
Searching several symbols per iteration

- Increased memory footprint
- Reduced number of LF operations



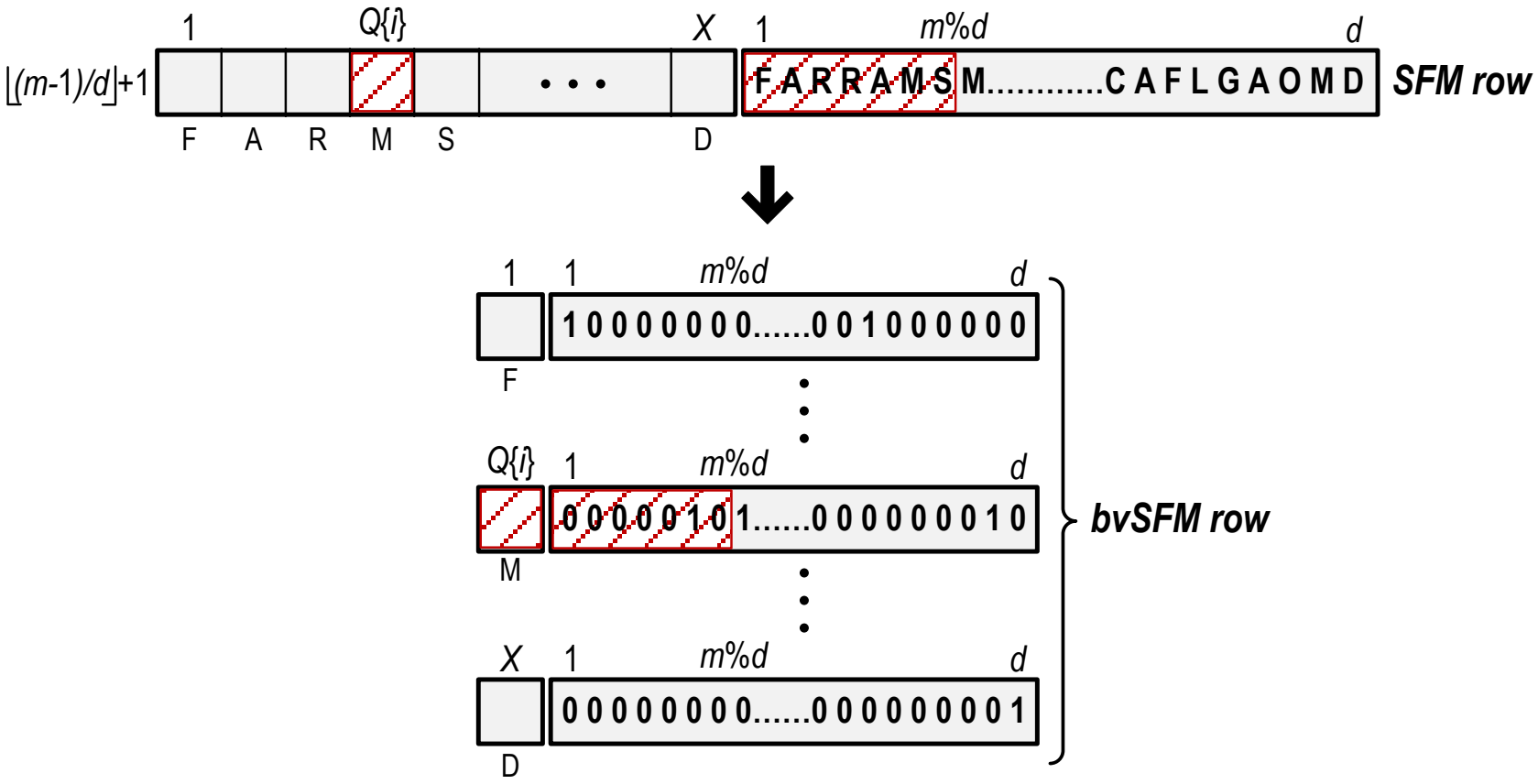

From Chacon et al. (2015)

# FM-Index Comparison



# Outline

# Bit-Vector FM-Index



# Bit-Vector FM-Index

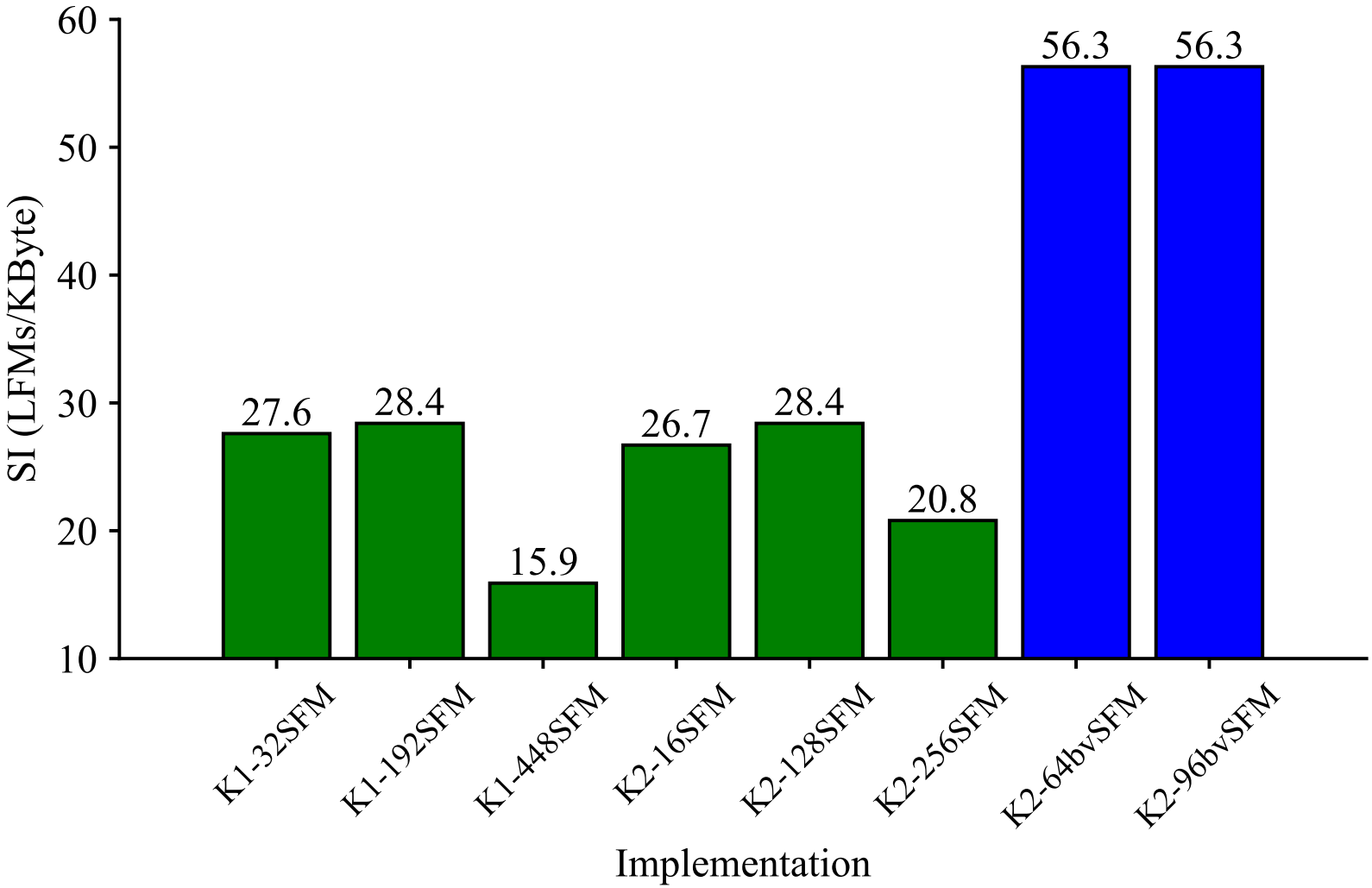
## Advantages

- Reduced data movement
- Reduced computing requirements

## Disadvantages

- Increased memory footprint

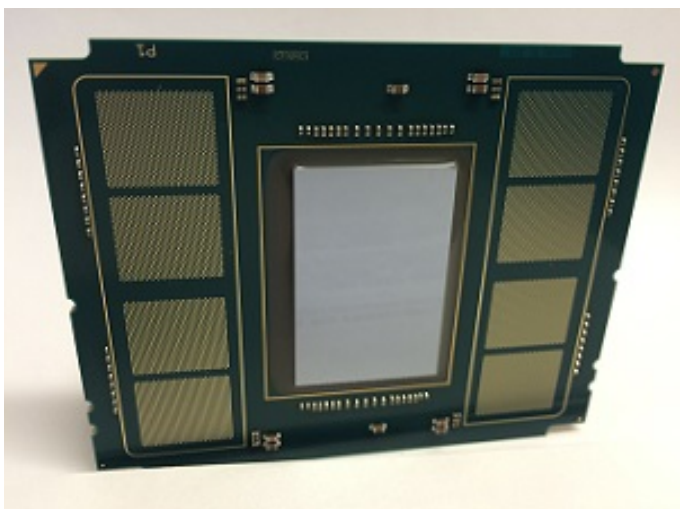
# FM-Index Comparison



# Outline

# Hardware Resources

| **Xeon Phi 7210** (KNL) | **Xeon E5-2630V4** (Broadwell) |
|---|---|
| 64 cores @ 1.3 GHz | 10 cores @ 2.2 GHz |
| 4 threads per core | 2 threads per core |
| 400 GB/s (MCDRAM) | 68 GB/s (DDR4) |



## Intel Xeon Phi

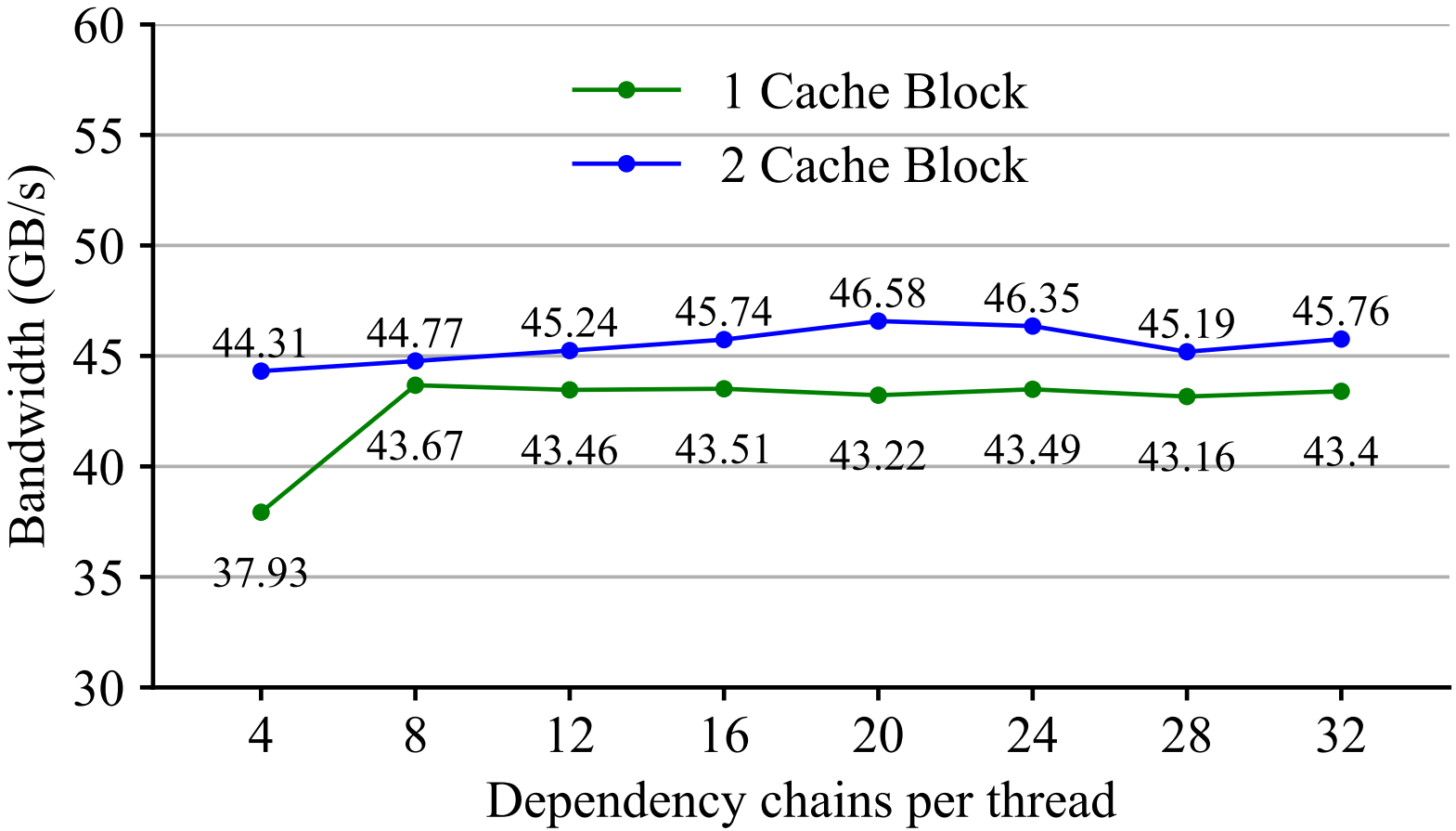
- AVX 512 Vectorial Processing Units
- High Bandwidth Memory (HBM)
  - Cache / Hybrid / Flat

# Outline

# Outline

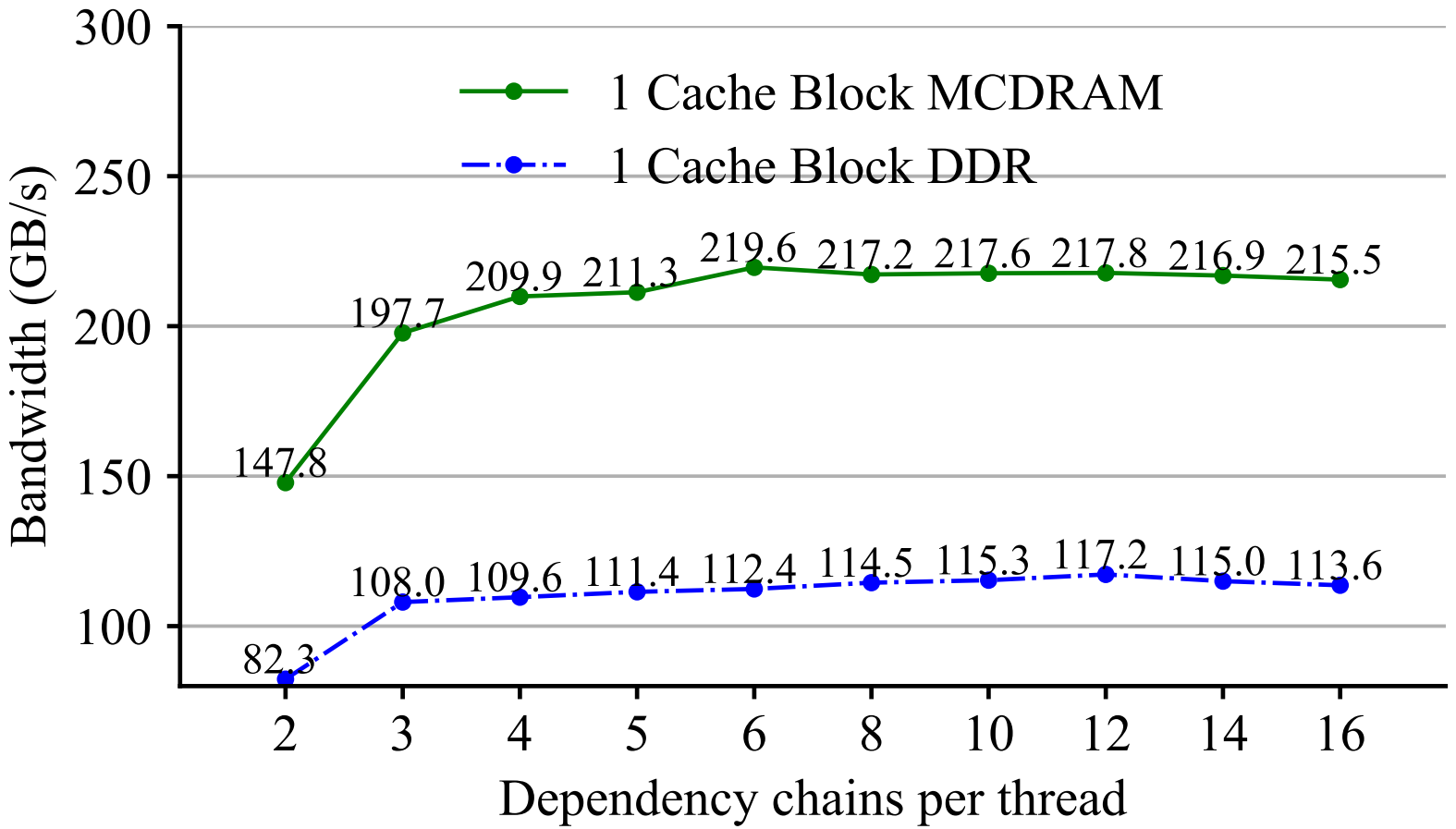
# Results

## Xeon Broadwell RANDOM Benchmark



# Results

## Xeon Phi KNL RANDOM Benchmark

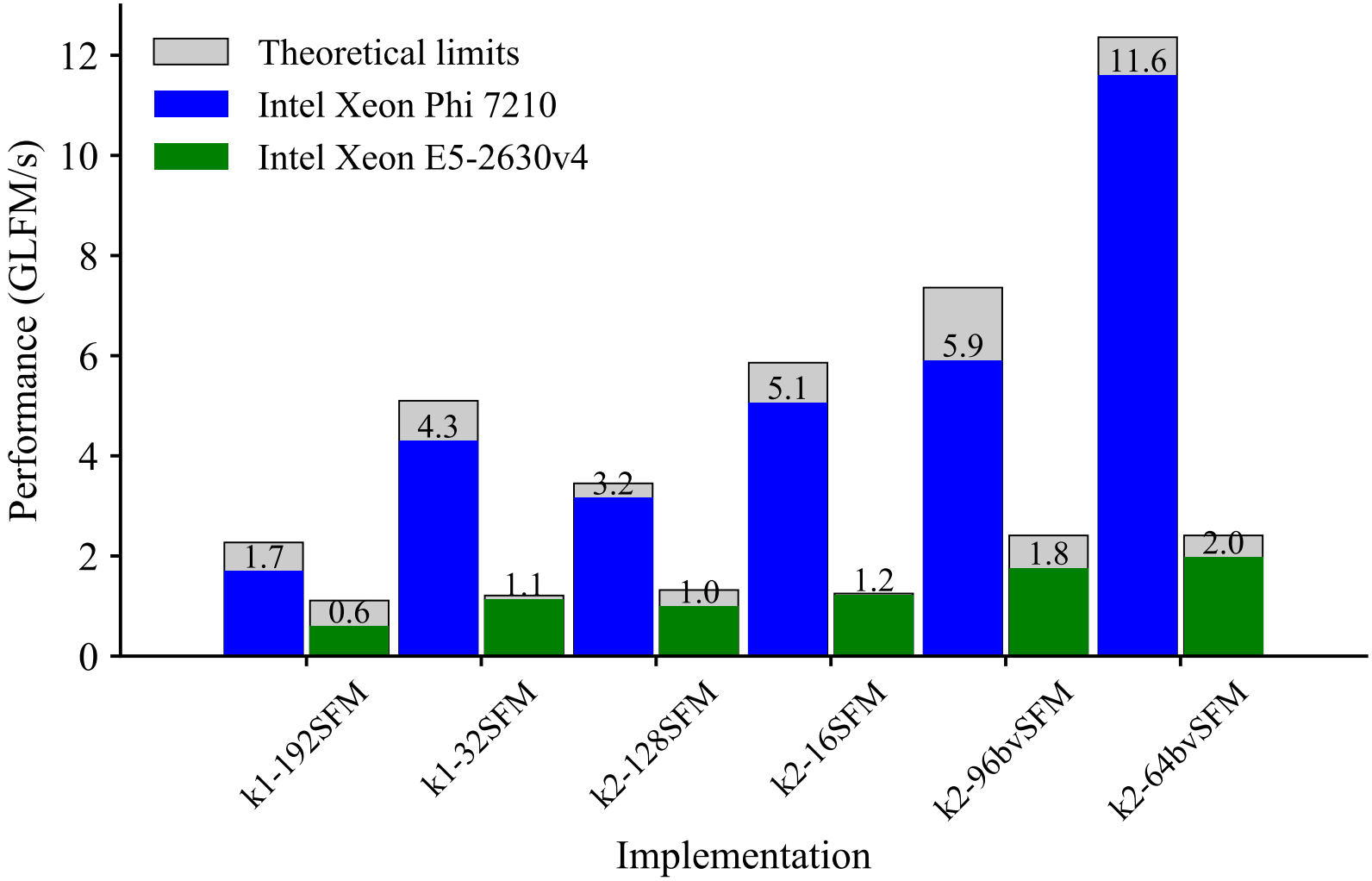


# Outline

# Results
## Experimental Setup

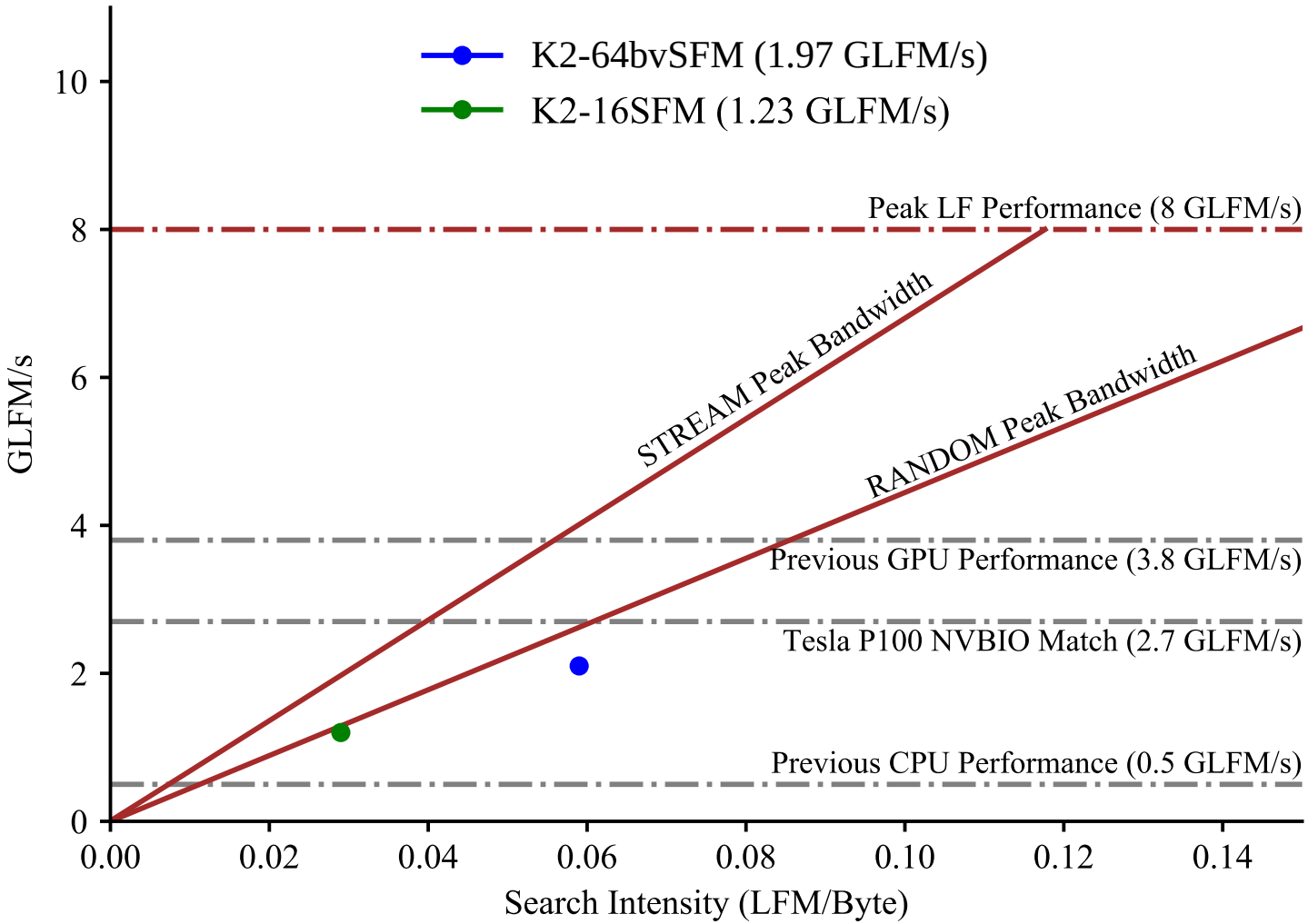
- Human Genome (Around 3GBases).
- 20 million input queries generated by Mason.
- 200 symbols per Sequence.
- Measurements started after loading the sequences into main memory.
- LFM/s as performance evaluation metric

# Results


Theoretical limits
Intel Xeon Phi 7210
Intel Xeon E5-2630v4
Performance (GLFM/s)
0
2
4
6
8
10
12
1.7
0.6
4.3
1.1
3.2
1.0
5.1
1.2
5.9
1.8
11.6
2.0
k1-192SFM
k1-32SFM
k2-128SFM
k2-16SFM
k2-96bvSFM
k2-64bvSFM
Implementation

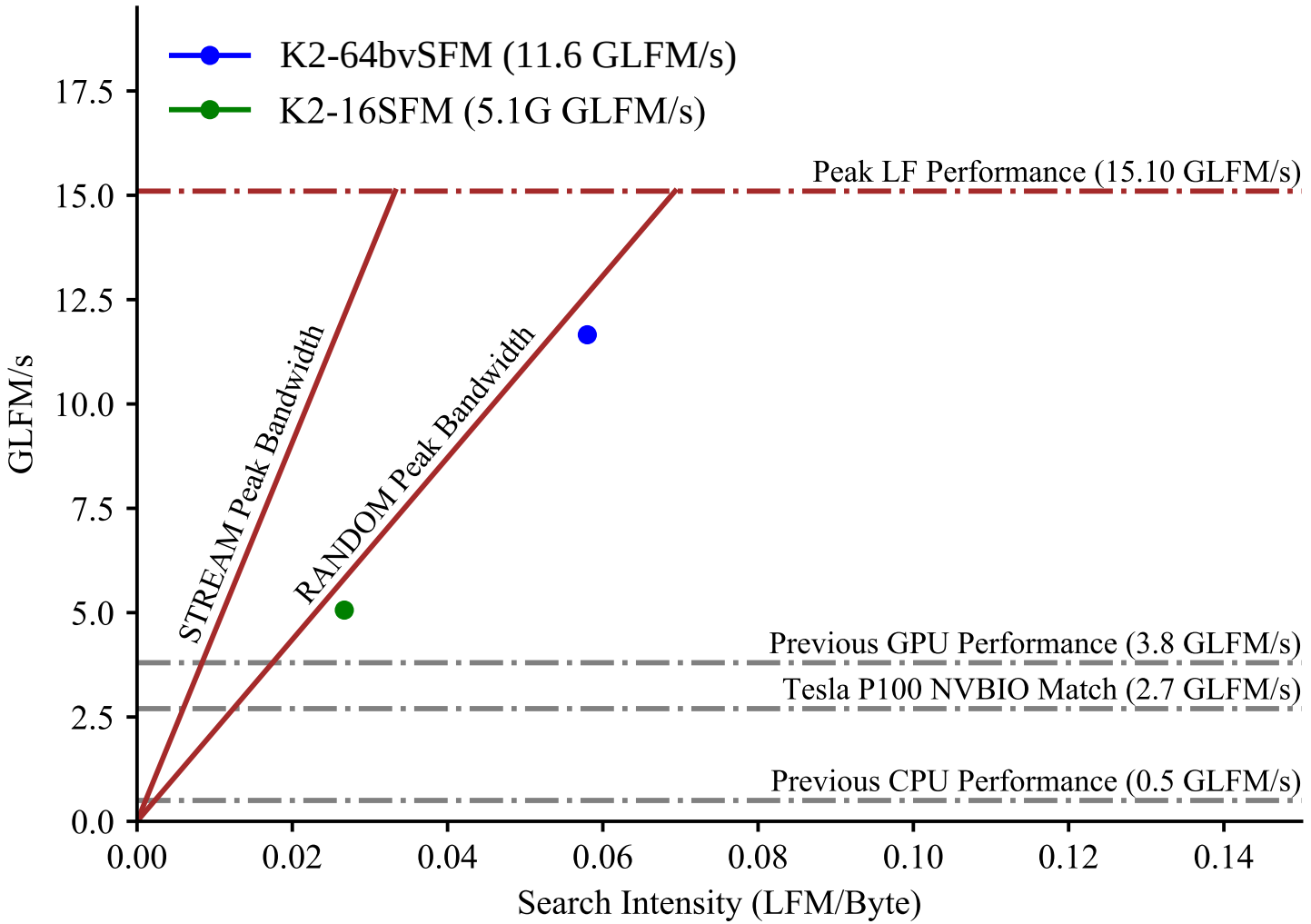
# Results

## Roofline (Broadwell)



# Results

## Roofline (Knights Landing)



# Outline

# Conclusions

- Reduced data movement
- Efficiently used memory bandwidth
- Great performance improvement
  - Up to 11.7 G LFM/s
  - 3x faster than previous GPU versions

# Exact Alignment with FM-Index on the Intel Xeon Phi Knights Landing Processor

Jose M. Herruzo[1] Sonia González-Navarro[1] Pablo Ibáñez[2]
Víctor Viñals[2] Jesús Alastruey-Benedé[2] Oscar Plata[1]

[1]Departamento de Arquitectura de Computadores
Universidad de Málaga

[2]Grupo de Arquitectura de Computadores
Universidad de Zaragoza

Accelerator Architecture in Computational Biology and Bioinformatics, 2018