Accelerating Genome Analysis

A Primer on an Ongoing Journey

Onur Mutlu
omutlu@gmail.com
https://people.inf.ethz.ch/omutlu

14 April 2023
BIO-Arch Workshop @ RECOMB

SAFARI ETH zürich Carnegie Mellon

Overview

- System design for bioinformatics is a critical problem
  - It has large scientific, medical, societal, personal implications

- This talk is about accelerating a key step in bioinformatics: genome sequence analysis
  - In particular, read mapping

- Many bottlenecks exist in accessing and manipulating huge amounts of genomic data during analysis

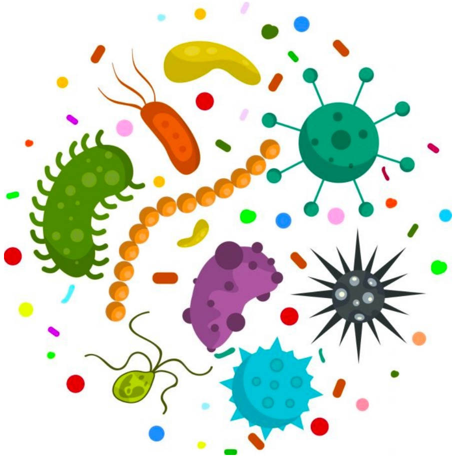
- Many recent ideas to accelerate read mapping
  - My personal journey since September 2006

Our Dream (circa 2007)

- An embedded device that can perform comprehensive genome analysis in real time (within a minute)
    - Which of these DNAs does this DNA segment match with?
    - What is the likely genetic disposition of this patient to this drug?
    - What disease/condition might this particular DNA/RNA piece associated with?
    - . . .

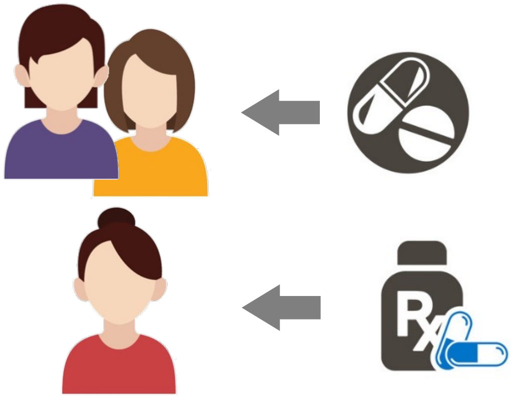
We Need Faster & Scalable Genome Analysis

Understanding **genetic variations, species, evolution, …**



Predicting the presence and relative abundance of **microbes** in a sample

Rapid surveillance of **disease outbreaks**



Developing **personalized medicine**

And, many, many other applications …

A Bright Future for Intelligent Genome Analysis

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu
"Accelerating Genome Analysis: A Primer on an Ongoing Journey" IEEE Micro, August 2020.

MinION from ONT

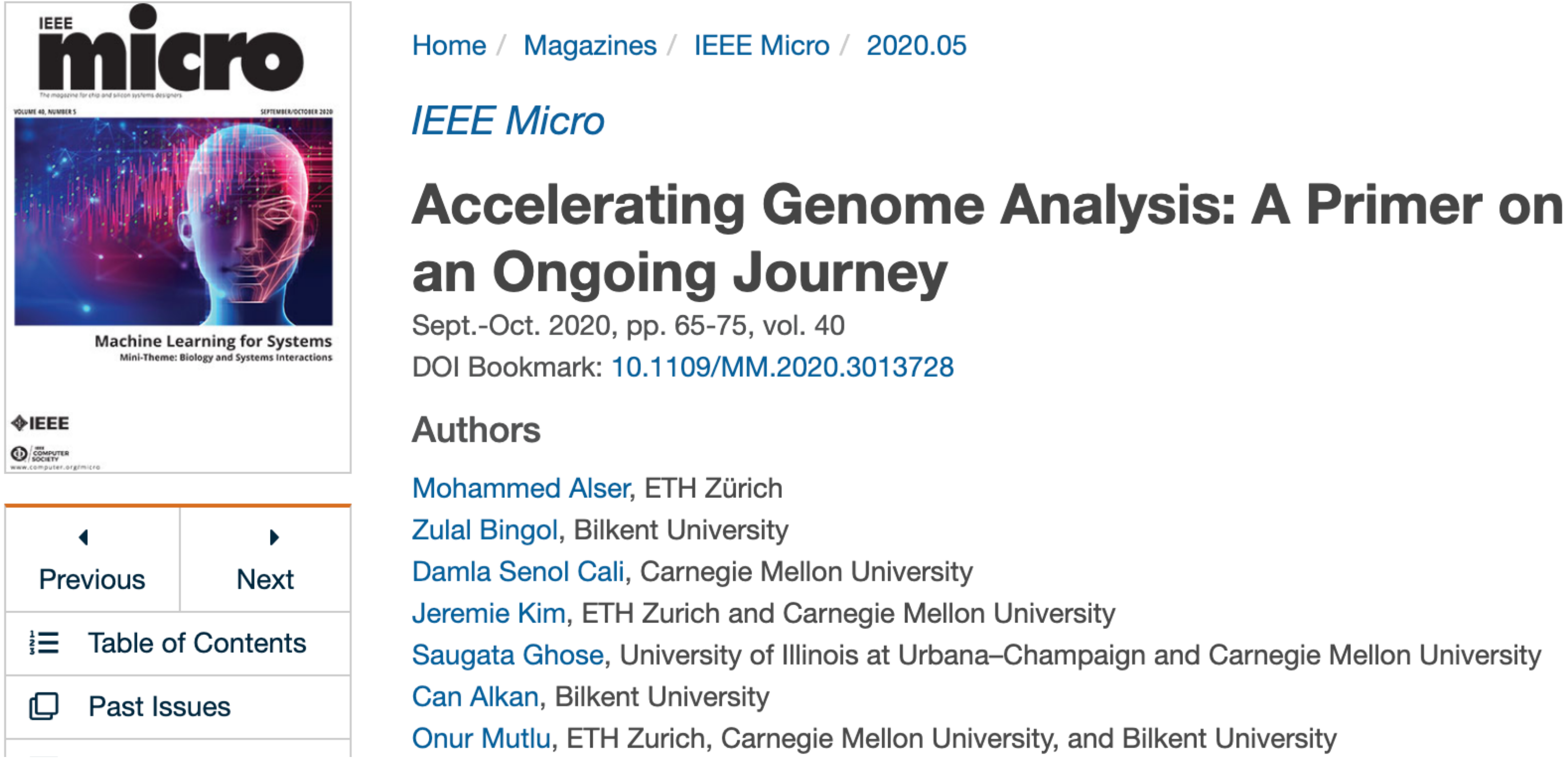
Accelerating Genome Analysis: A Primer on an Ongoing Journey
Sept.-Oct. 2020, pp. 65-75, vol. 40
DOI Bookmark: 10.1109/MM.2020.3013728

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications
July-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

SmidgION from ONT

A Few Overview Readings (I)

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu
"Accelerating Genome Analysis: A Primer on an Ongoing Journey"
IEEE Micro, August 2020.



IEEE Micro

Accelerating Genome Analysis: A Primer on an Ongoing Journey

Sept.-Oct. 2020, pp. 65-75, vol. 40
DOI Bookmark: 10.1109/MM.2020.3013728

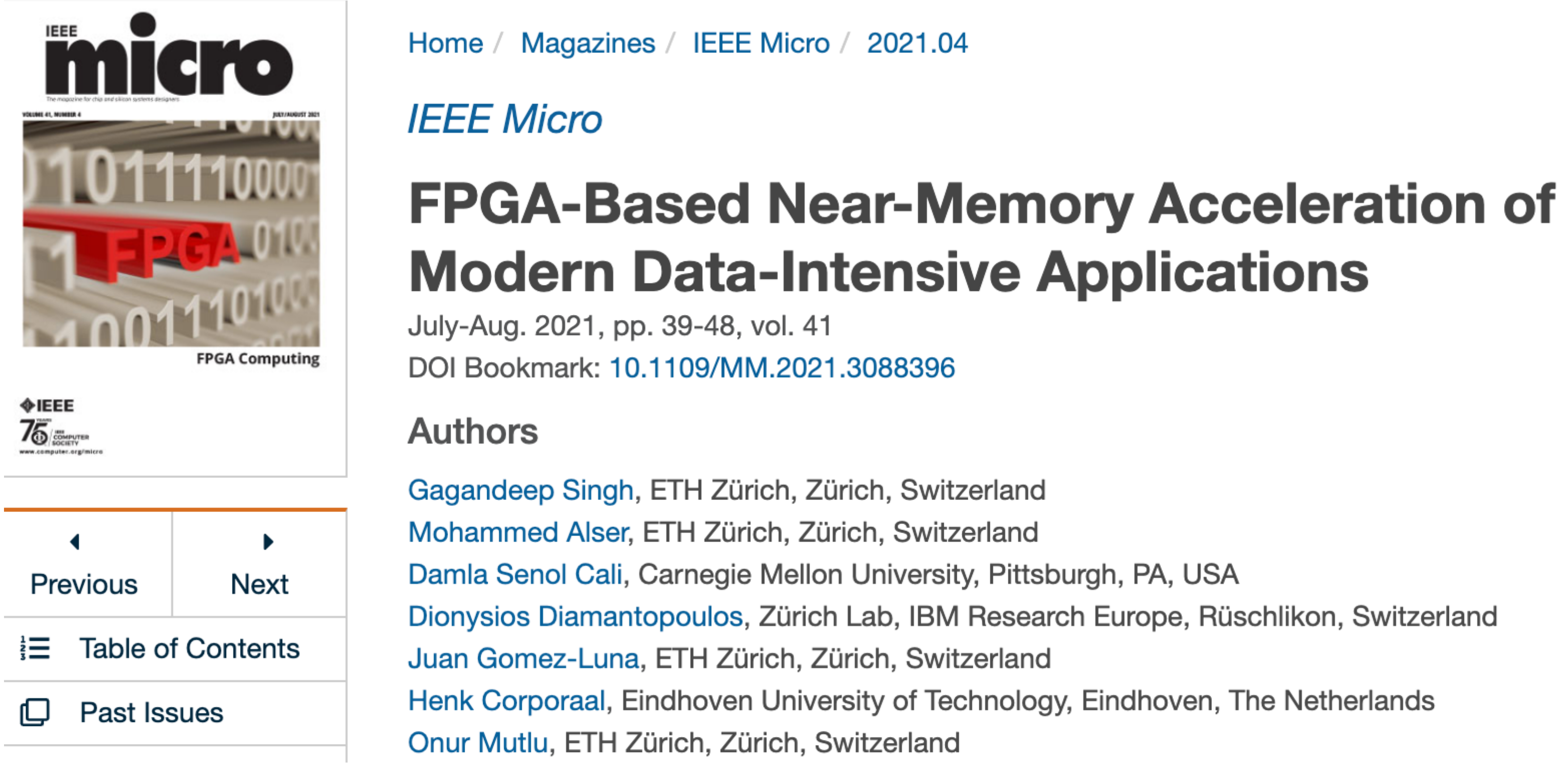
Authors

Mohammed Alser, ETH Zürich
Zulal Bingol, Bilkent University
Damla Senol Cali, Carnegie Mellon University
Jeremie Kim, ETH Zurich and Carnegie Mellon University
Saugata Ghose, University of Illinois at Urbana–Champaign and Carnegie Mellon University
Can Alkan, Bilkent University
Onur Mutlu, ETH Zurich, Carnegie Mellon University, and Bilkent University

https://arxiv.org/pdf/2008.00961.pdf

A Few Overview Readings (II)

Gagandeep Singh, Mohammed Alser, Damla Senol Cali, Dionysios Diamantopoulos, Juan Gomez-Luna, Henk Corporaal, Onur Mutlu,
[“FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications“]()
IEEE Micro, 2021.
[[Source Code]()]



IEEE Micro

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications

July-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

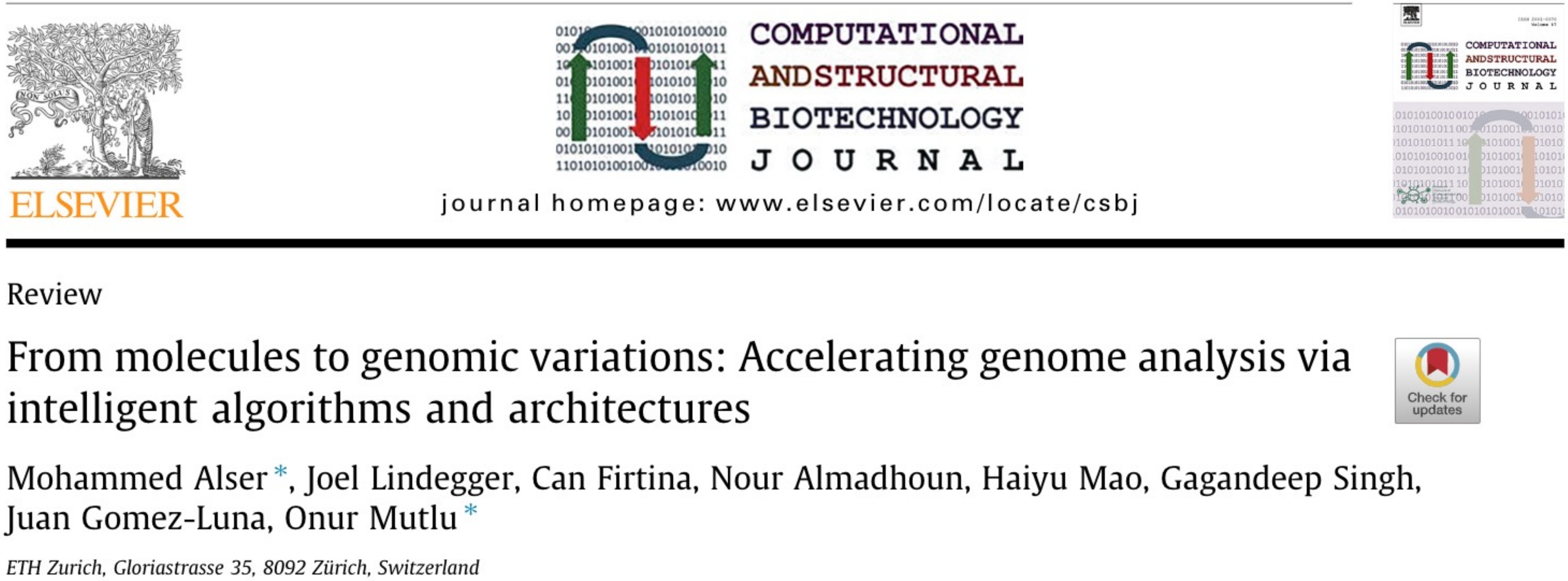
Authors

Gagandeep Singh, ETH Zürich, Zürich, Switzerland
Mohammed Alser, ETH Zürich, Zürich, Switzerland
Damla Senol Cali, Carnegie Mellon University, Pittsburgh, PA, USA
Dionysios Diamantopoulos, Zürich Lab, IBM Research Europe, Rüschlikon, Switzerland
Juan Gomez-Luna, ETH Zürich, Zürich, Switzerland
Henk Corporaal, Eindhoven University of Technology, Eindhoven, The Netherlands
Onur Mutlu, ETH Zürich, Zürich, Switzerland

[https://arxiv.org/pdf/2106.06433.pdf]()

A Few Overview Readings (III)

Mohammed Alser, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez-Luna, Onur Mutlu
"**From Molecules to Genomic Variations: Intelligent Algorithms and Architectures for Intelligent Genome Analysis**"
Computational and Structural Biotechnology Journal, 2022
[Source code]



ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

From molecules to genomic variations: Accelerating genome analysis via intelligent algorithms and architectures

Check for updates

Mohammed Alser *, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez-Luna, Onur Mutlu *

ETH Zurich, Gloriastrasse 35, 8092 Zürich, Switzerland

https://arxiv.org/pdf/2205.07957.pdf

A Few Overview Readings (IV)

- **To appear at DAC 2023**

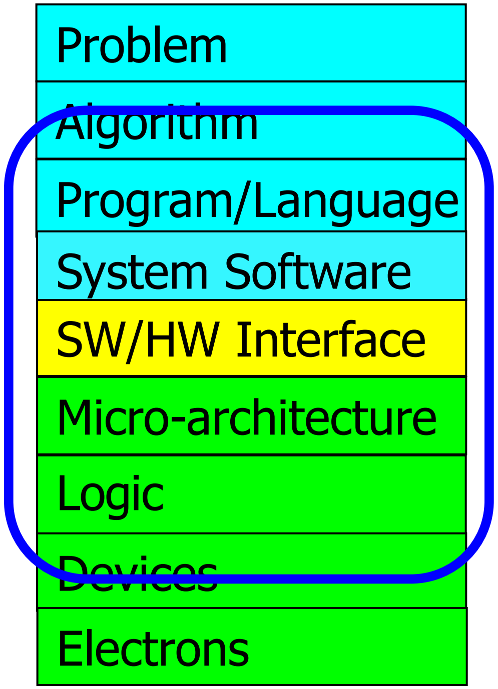
Invited: Accelerating Genome Analysis via Algorithm-Architecture Co-Design

Onur Mutlu Can Firtina

ETH Zürich

Algorithm-Arch-Device Co-Design is Critical

Computer Architecture (expanded view)



Problem

Algorithm

Program/Language

System Software

SW/HW Interface

Micro-architecture

Logic

Devices

Electrons

Agenda

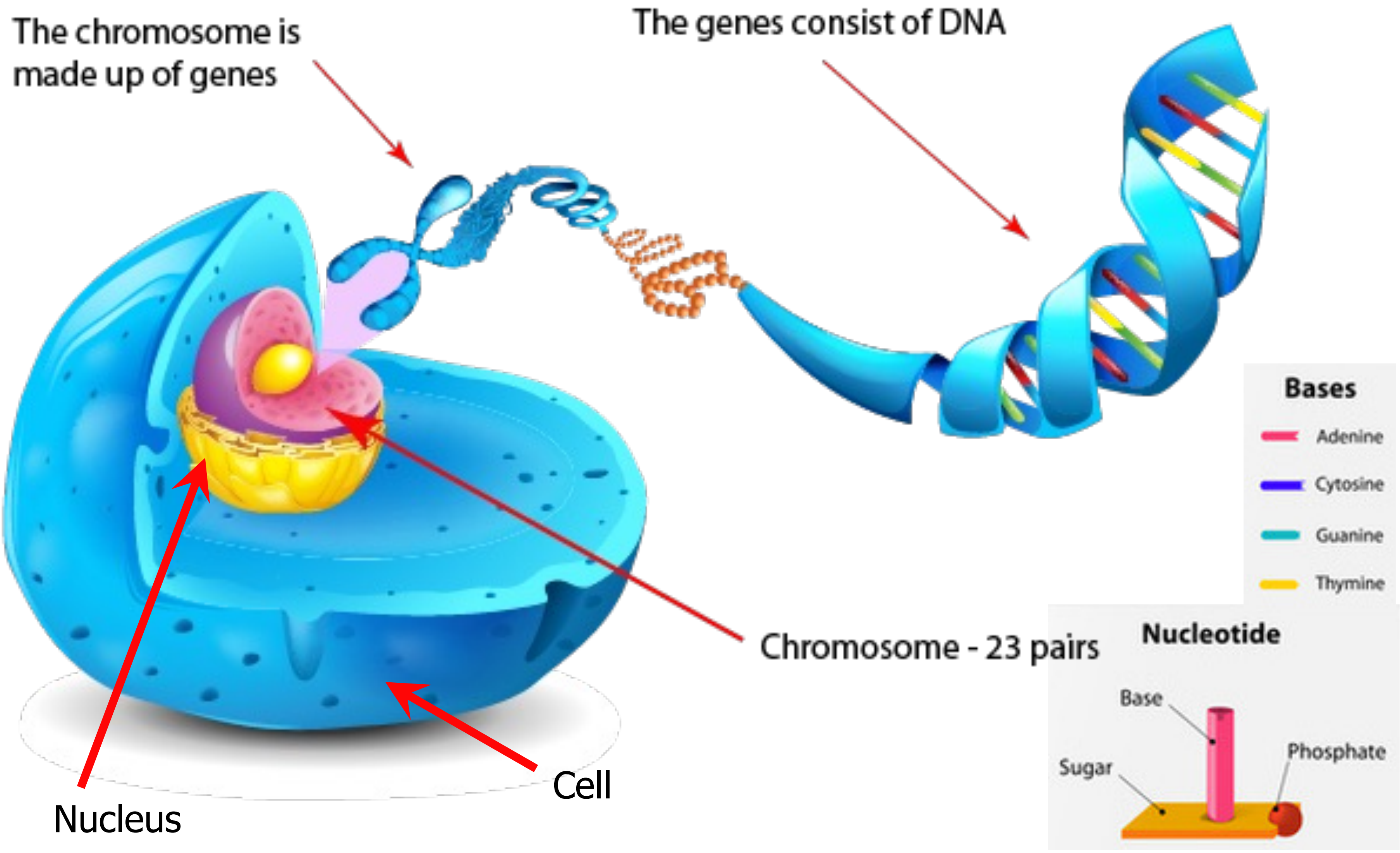
- The Problem: DNA Read Mapping
    - State-of-the-art Read Mapper Design

- Algorithmic Acceleration
    - Exploiting Structure of the Genome
    - Exploiting SIMD Instructions

- Hardware Acceleration
    - Specialized Architectures
    - Processing in Memory & Storage

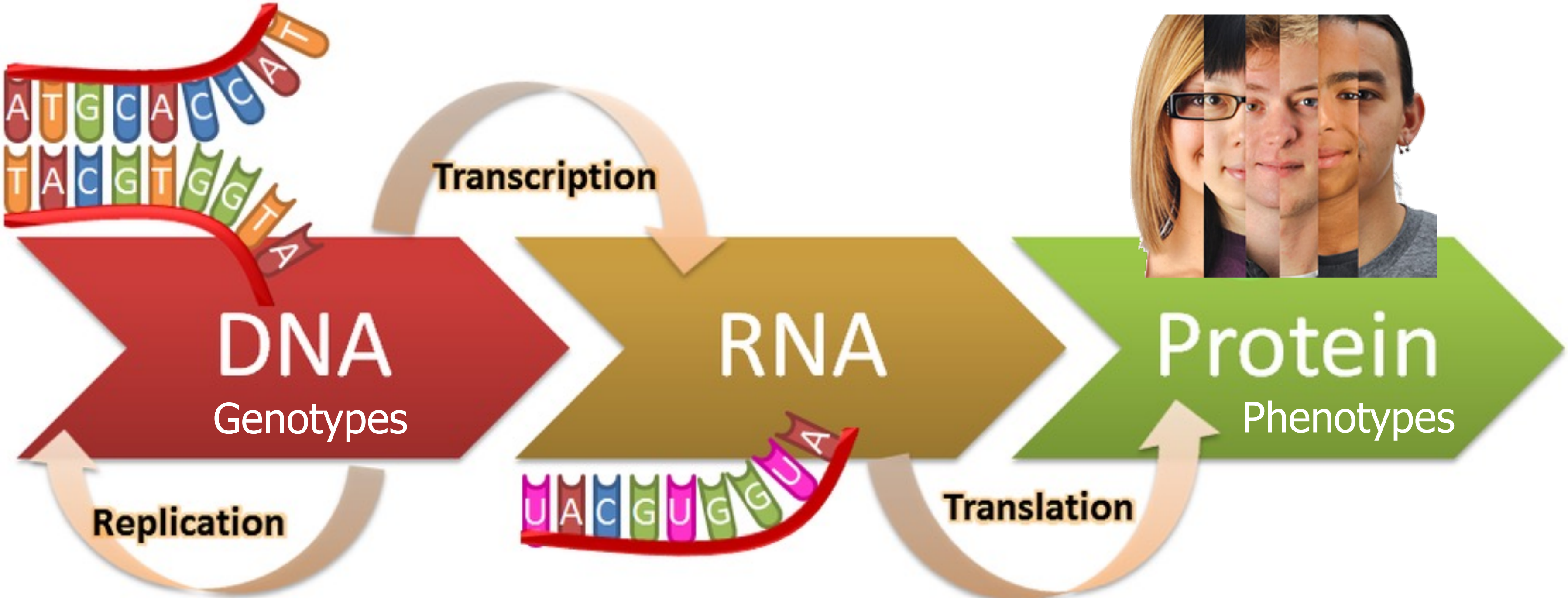
- Future Opportunities: New Technologies & Applications

What Is a Genome Made Of?



The discovery of DNA’s double-helical structure (Watson+, 1953)

The Central Dogma of Molecular Biology



Transcription

DNA
Genotypes

RNA

Protein
Phenotypes

Replication

Translation

ALL-TIME FAVORITES

Whenever you think of pie—what kind do you think of first? Of course, the old favorites that you and your family have enjoyed over and over again. Here are the best recipes there are for the old stand-bys.

McCallum Vintage Recipe Diner

Old-Fashioned Apple Pie

BAKE at 450° for 10 minutes, then at 375° for 40 to 50 minutes. MAKES 9-inch pie.

Prepare...... pastry for two-crust pie as directed on **Pillsbury Golden Pie Crust Mix** or **Sticks** package. Fit bottom crust loosely into 9-inch piepan.

Combine...... in small bowl
¾ to 1 cup **sugar** (depending on sweetness of apples)
2 tablespoons **flour**
1 teaspoon grated **lemon rind**
½ teaspoon **cinnamon** and
¼ teaspoon **salt**. Set aside.

Arrange...... 6 cups pared and sliced **apples** (6 medium) in closely packed layers in pastry-lined pan, sprinkling cinnamon-sugar mixture over each layer.

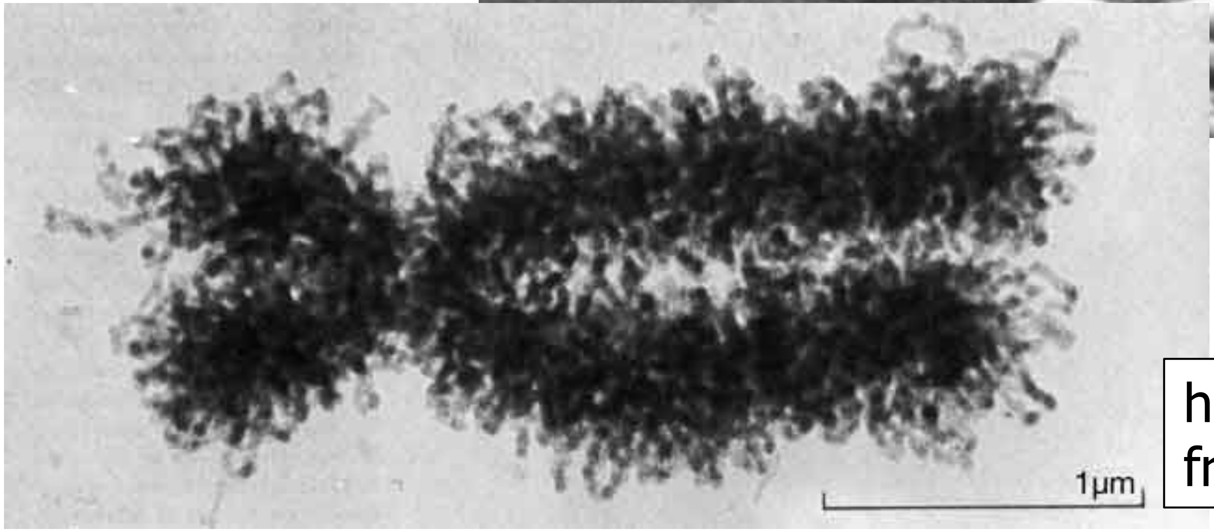
Sprinkle...... 1 tablespoon **lemon** juice over apples. Dot with 2 tablespoons **butter**.

Roll out...... remaining dough. Cut slits for escape of steam. Moisten rim of bottom crust. Place top crust over filling. Fold edge under bottom crust, pressing to seal. Flute.

Bake........ at 450° for 10 minutes, then at 375° for 40 to 50 minutes until golden brown.

Bake cheese into your next apple pie by making a cheese-flavored crust. Follow the recipe on the Pillsbury Golden Pie Crust Mix or Sticks package.

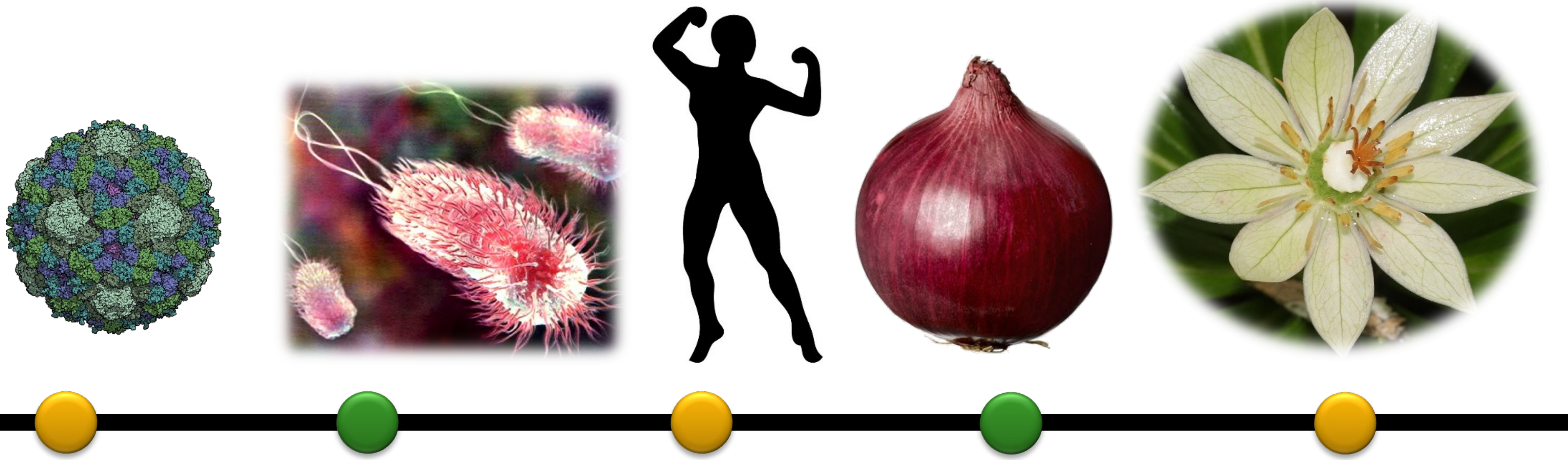
DNA Under Electron Microscope



human chromosome #12
from HeLa’s cell

CCTCCTCAGTGCCACCCAGCCCACTGGCAGCTCCCAAACA
GGCTCTTATTAAAACACCCTGTTCCCTGCCCCTTGGAGTG
AGGTGTCAAGGACCTAAACTAAAAAAAAAAAAAGAAAA
AGAAAAGAAAAAAGAATTTAAAATTTAAGTAATTCTTTGAA
AAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATG
TGCTAAACAGCACTTTTT**TTGACCATTAT**TTTGGATCTGAAA
GAAATCAAGAATAAATGAAGGACTTGATACATTGGAAGA
GGAGAGTCAAGGACCTACAGAAAAAAAAAAAAAAGAAA
AAGAAAAGAAAAAAGA**A**TTTAAAATTTAAGTAATTCTTTGA
AAAAAACTAATTTCTAAGCTTCTT**C**ATGTCAAGGACCTAAT
GTCTGTGTTGCAGGTCTTCTTGCATTTCCCTGTCAAAAGA
AAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTA
ATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCAGGCC
GGCTCTTATTAAAACACCCTGTTCCCTGCCCCTTGGAGTG

How Large is a Genome?



Phi X174 virus
5.386 Kilo bp

E. coli O157:H7
5.44 Million bp

Homo Sapiens
3.2 Billion bp

Onion, Allium Cepa
16 Billion bp

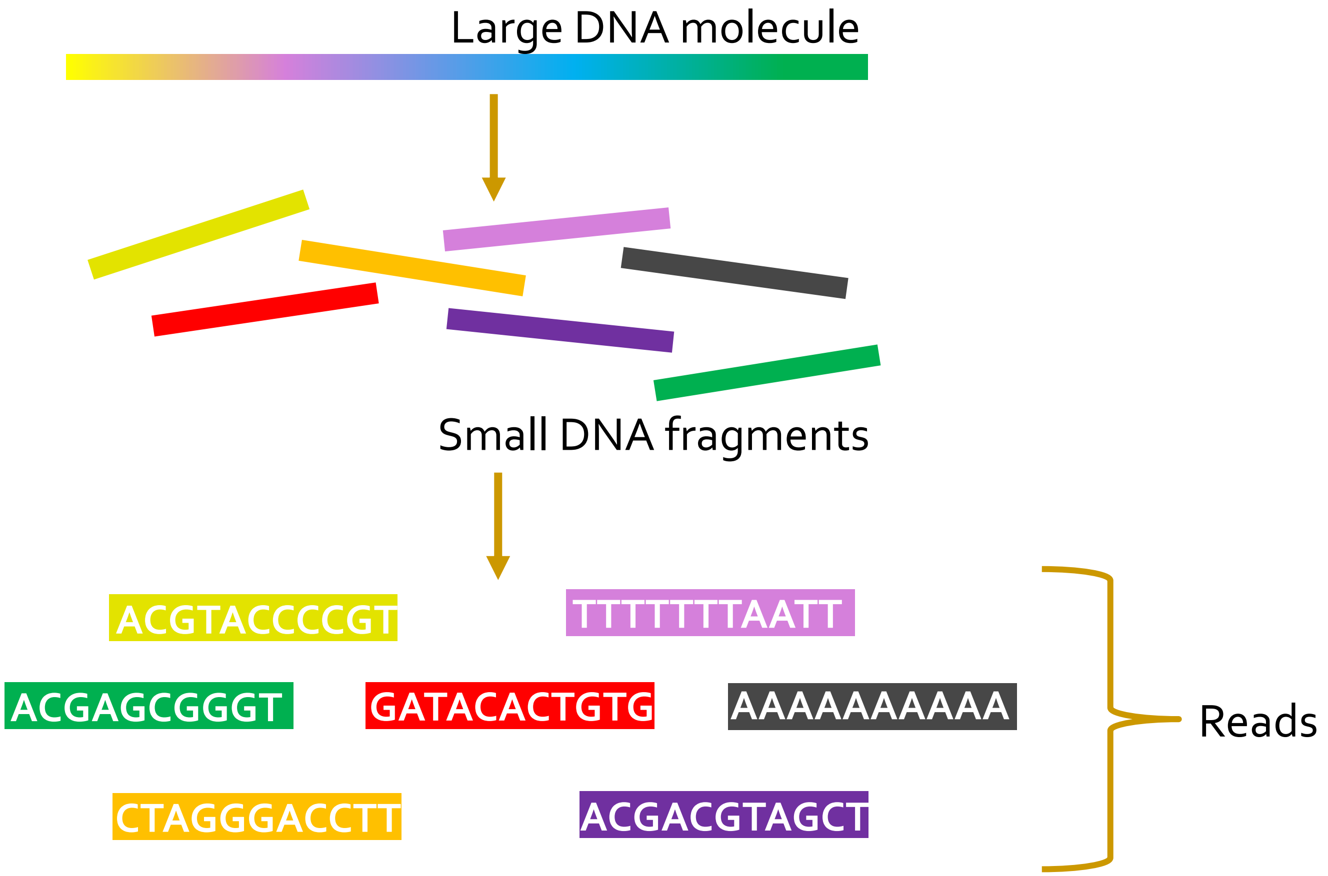
Paris Japonica
149 Billion bp

DNA Sequencing

- Goal:
  - Find the complete sequence of A, C, G, T's in an organism's DNA

- Challenge:
  - There is no machine that takes long DNA as an input, and gives the complete sequence as output
  - All sequencing machines chop DNA into pieces and identify relatively small pieces (but not how they fit together)

Genome Sequencing



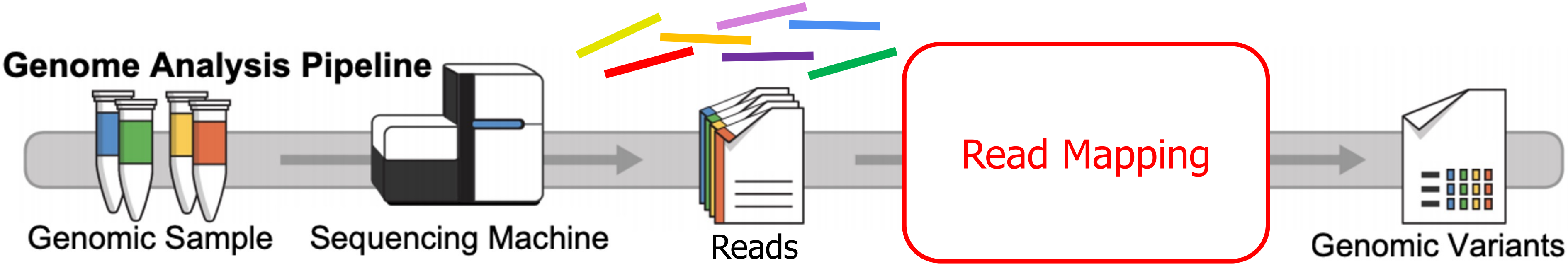
Large DNA molecule

Small DNA fragments

ACGTACCCCGT
TTTTTTTAATT
ACGAGCGGGT
GATACACTGTG
AAAAAAAAAA
CTAGGGACCTT
ACGACGTAGCT

Reads

Genome Sequencing and Analysis



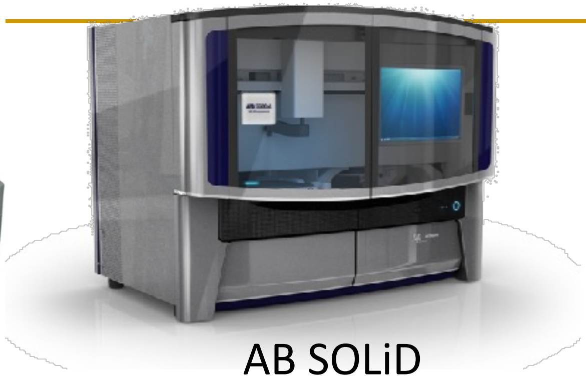
Current sequencing machines provide **small randomized fragments** of the original DNA sequence

Alser+, "[Technology dictates algorithms: Recent developments in read alignment]", Genome Biology, 2021

Untangling Yarn Balls & DNA Sequencing

Genome Sequencers


Roche/454


AB SOLiD


Illumina MiSeq


Complete Genomics


Oxford Nanopore MinION


Illumina NovaSeq 6000


Illumina HiSeq2000


Pacific Biosciences RS


Oxford Nanopore GridION


Ion Torrent PGM


Ion Torrent Proton

… and more! All produce data with different properties.

SAFARI

High-Throughput Sequencers

Illumina MiSeq

Pacific
Biosciences
Sequel II

Oxford
Nanopore
PromethION

Oxford Nanopore MinION

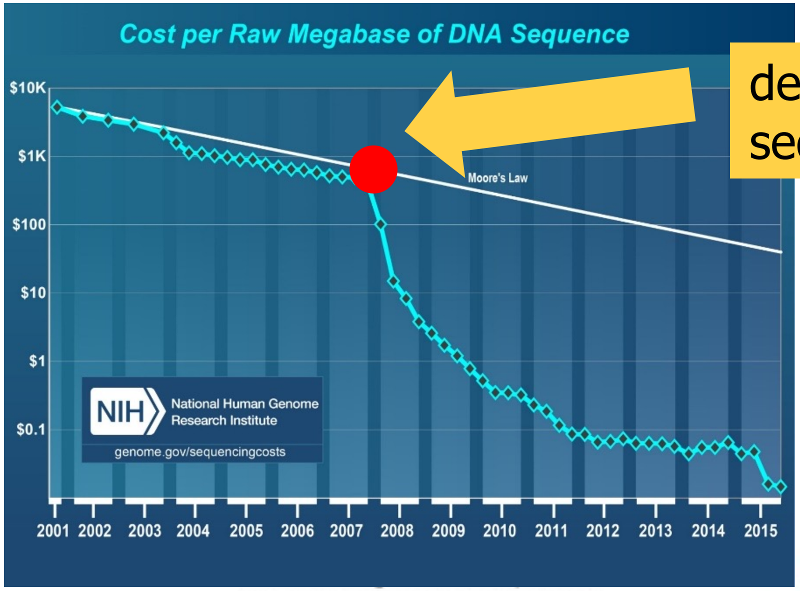
Illumina NovaSeq 6000

Pacific Biosciences RS II

Oxford
Nanopore
SmidgION

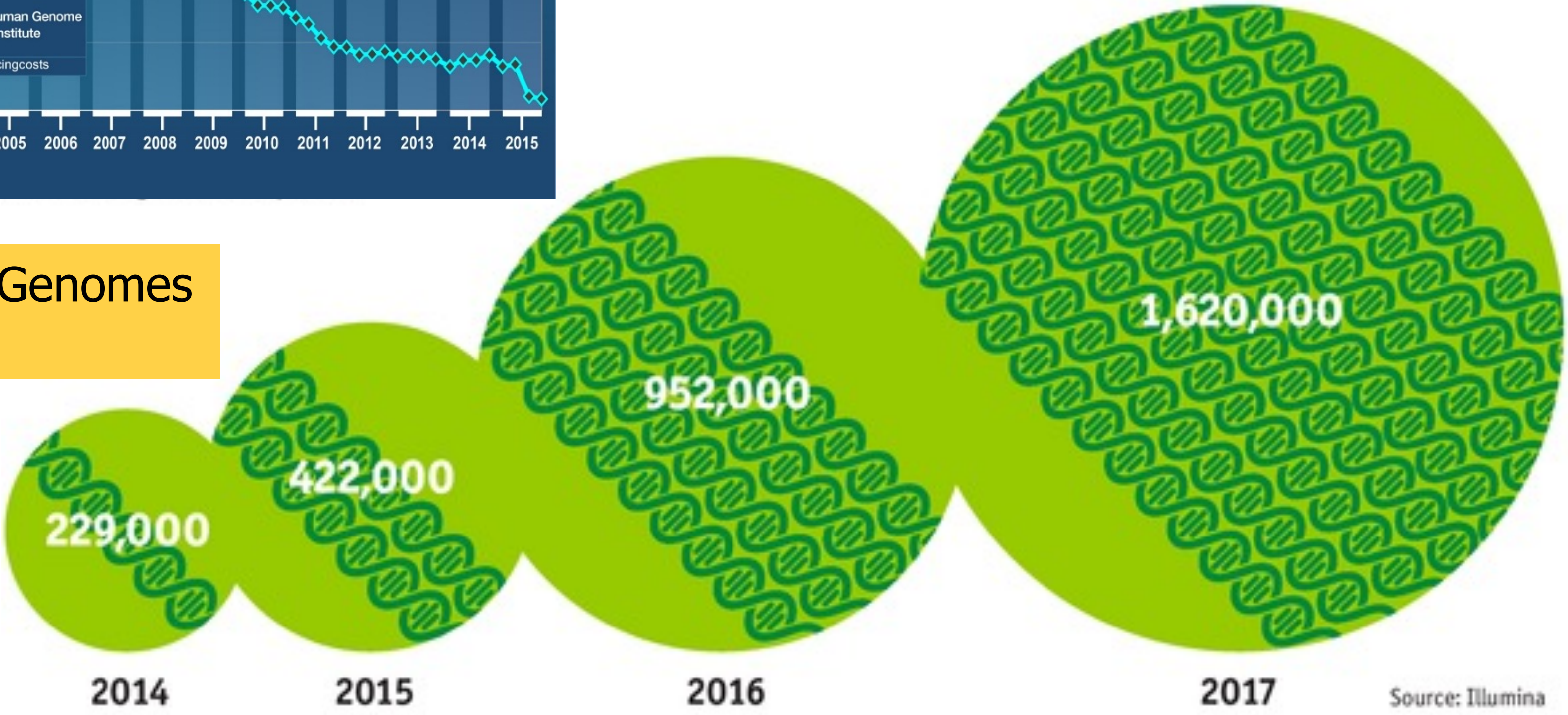
… and more! All produce data with different properties.

The Genomic Era



development of high-throughput sequencing (HTS) technologies

Number of Genomes Sequenced



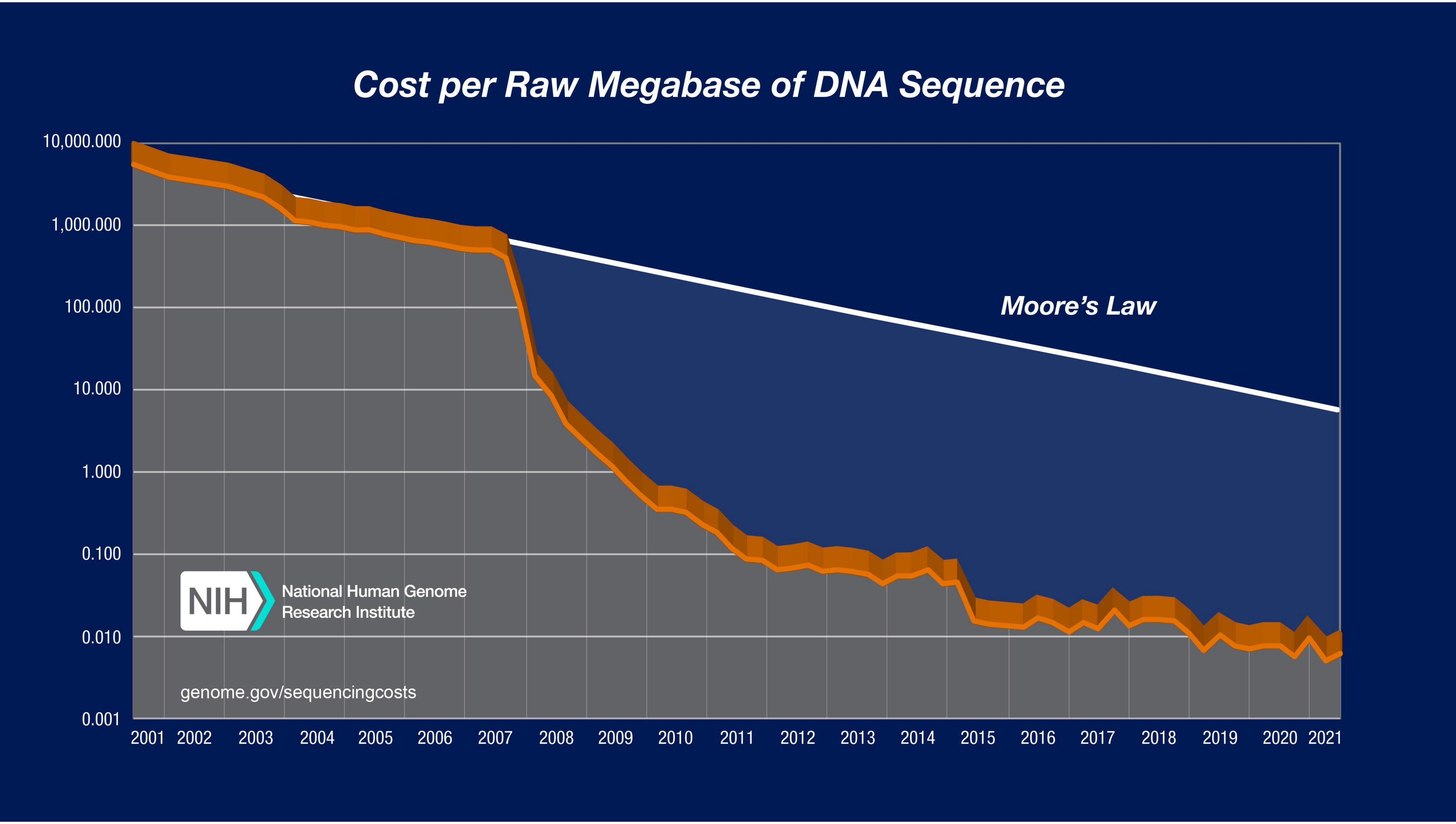
Source: Illumina

The Economist

SAFARI

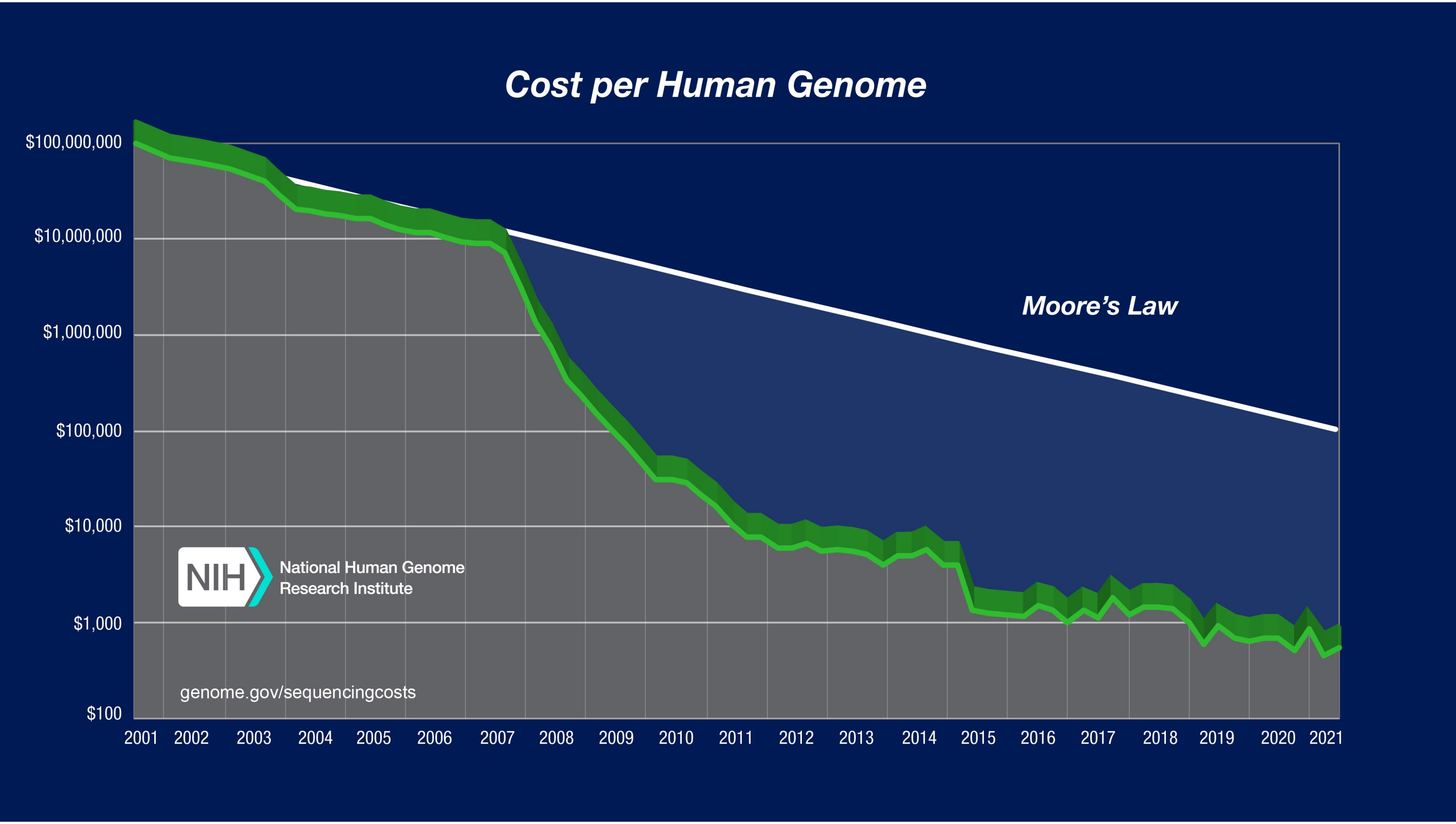
http://www.economist.com/news/21631808-so-much-genetic-data-so-many-uses-genes-unzipped

Genome Sequencing Cost Is Reducing



*From NIH (https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data)

Genome Sequencing Cost Is Reducing

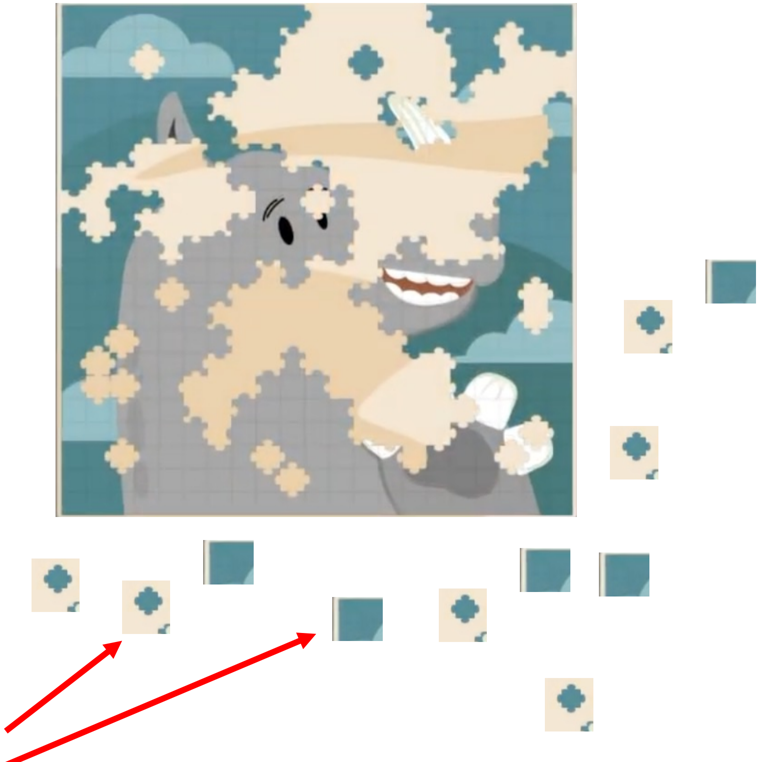


*From NIH (https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data)

Solving the Puzzle

.FASTA file

.FASTQ file



Reference genome

Sequenced Reads

https://www.pacb.com/smrt-science/smrt-sequencing/hifi-reads-for-highly-accurate-long-read-sequencing/

Newer Genome Sequencing Technologies

Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions

Damla Senol Cali ✉, Jeremie S Kim, Saugata Ghose, Can Alkan, Onur Mutlu

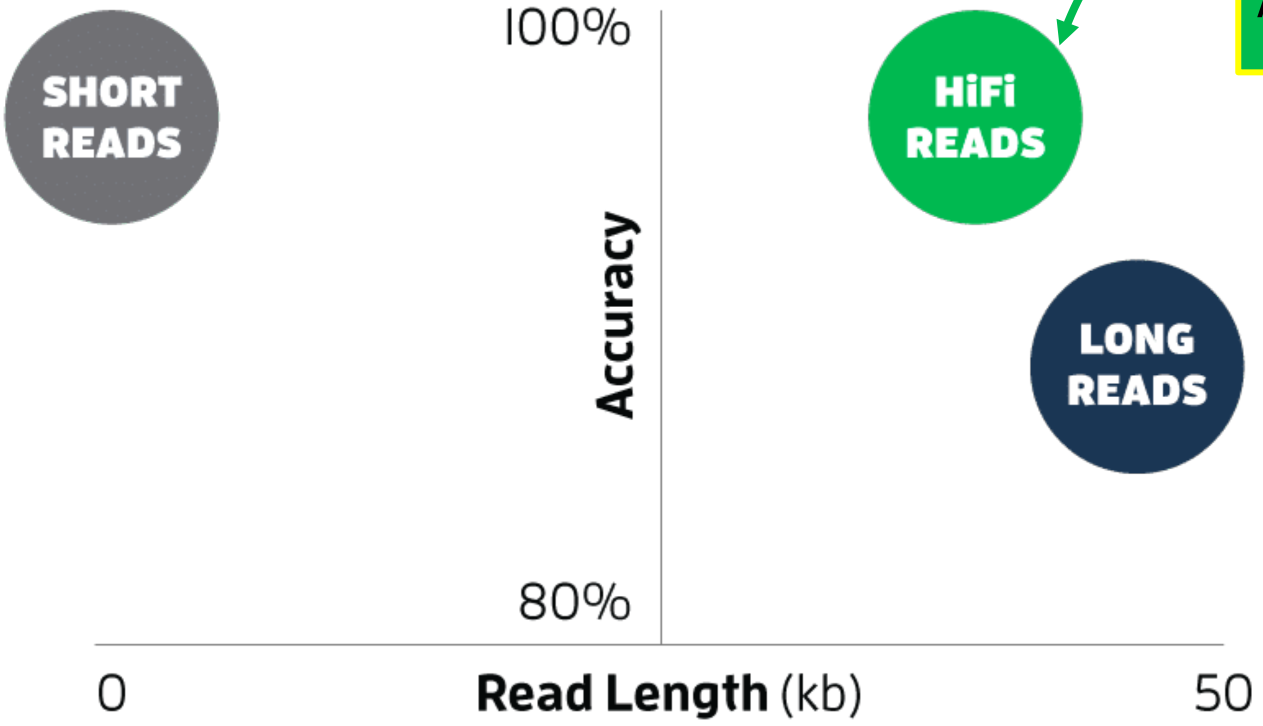
Briefings in Bioinformatics, bby017, https://doi.org/10.1093/bib/bby017

Published: 02 April 2018 **Article history** ▾

Oxford Nanopore MinION

Senol Cali+, "**Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions**," Briefings in Bioinformatics, 2018.
[Open arxiv.org version] [Slides (pptx) (pdf)] [Talk Video at AACBB 2019]

Types of Genomic Reads



Long: 10-20 kb
Accurate: 99.8%

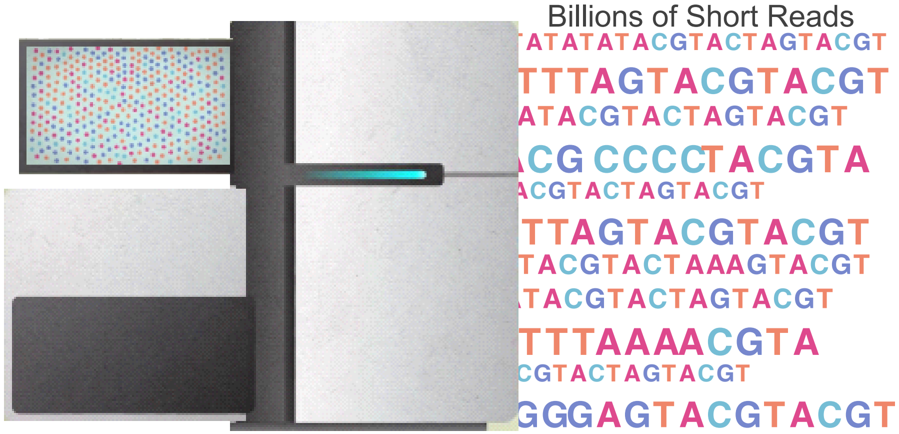
But still very expensive!

Wenger+, "[Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome]()", *Nature Biotechnology*, 2019

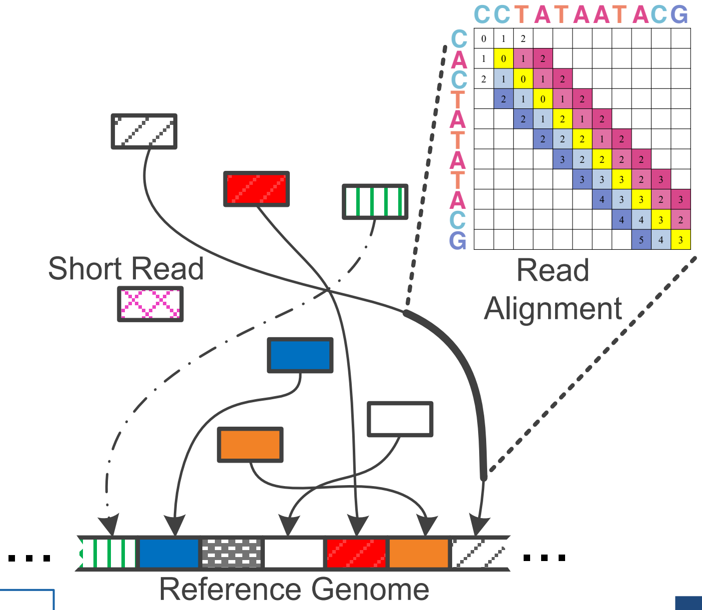
https://labs.wsu.edu/genomicscore/illumina-sequencing/
https://pacbio.gs.washington.edu/

Genome Analysis

1 Sequencing



2 Read Mapping



3 Variant Calling

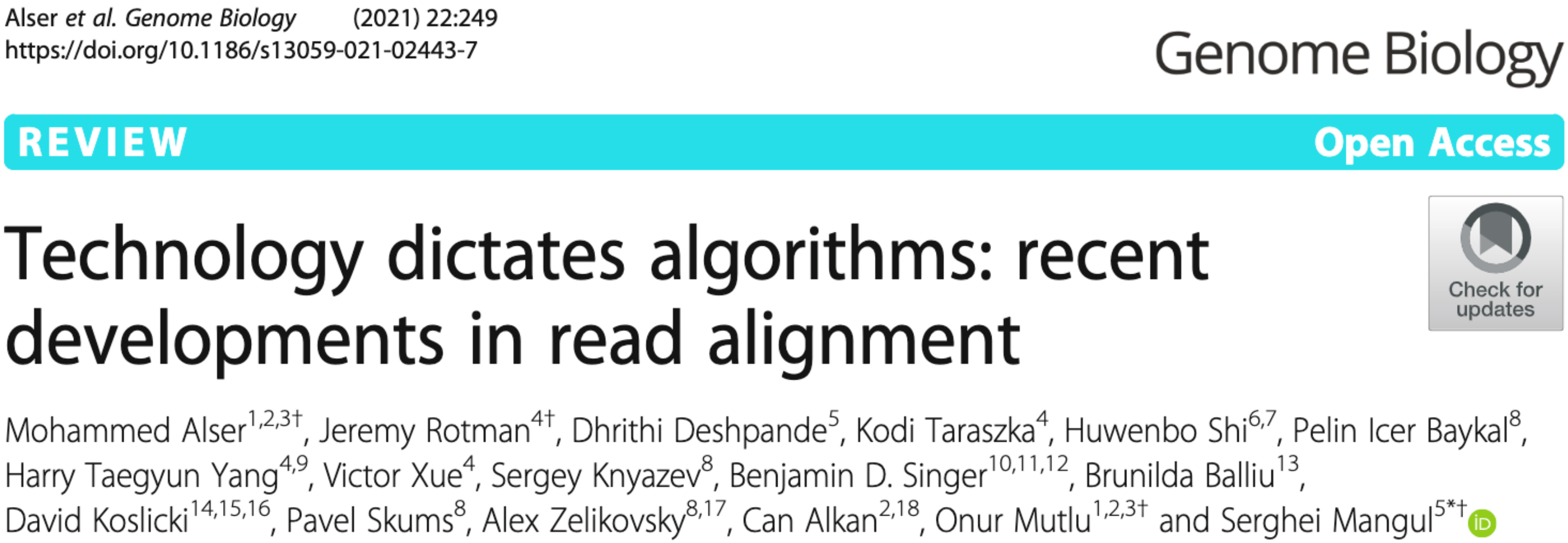
```
reference: TTTATCGCTTCCATGACGCAG
read1:        ATCGCATCC
read2:       TATCGCATC
read3:            CATCCATGA
read4:          CGCTTCCAT
read5:               CCATGACGC
read6:             TTCCATGAC
```

4 Scientific Discovery

Read Mapping Techniques in 111 Pages

In-depth analysis of 107 read mappers (1988-2020)

Mohammed Alser, Jeremy Rotman, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, Victor Xue, Sergey Knyazev, Benjamin D. Singer, Brunilda Balliu, David Koslicki, Pavel Skums, Alex Zelikovsky,
Can Alkan, Onur Mutlu, Serghei Mangul
"Technology dictates algorithms: Recent developments in read alignment"
Genome Biology, 2021
[Source code]



Alser *et al. Genome Biology* (2021) 22:249
https://doi.org/10.1186/s13059-021-02443-7

Genome Biology

REVIEW

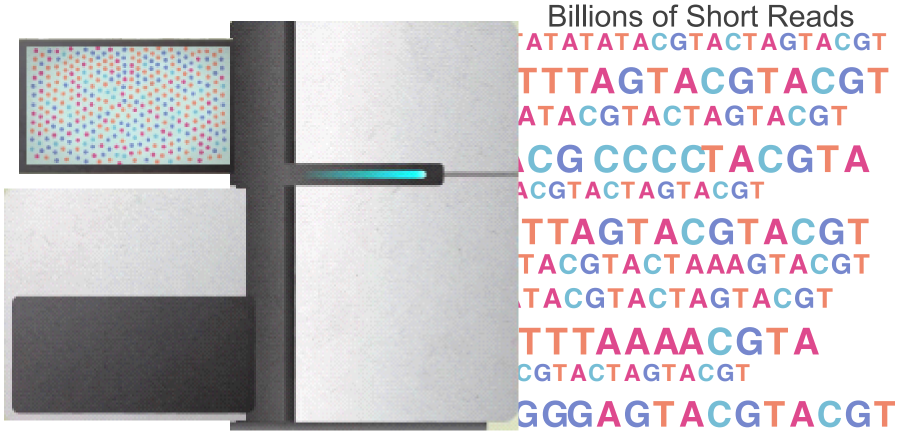
Open Access

Technology dictates algorithms: recent developments in read alignment

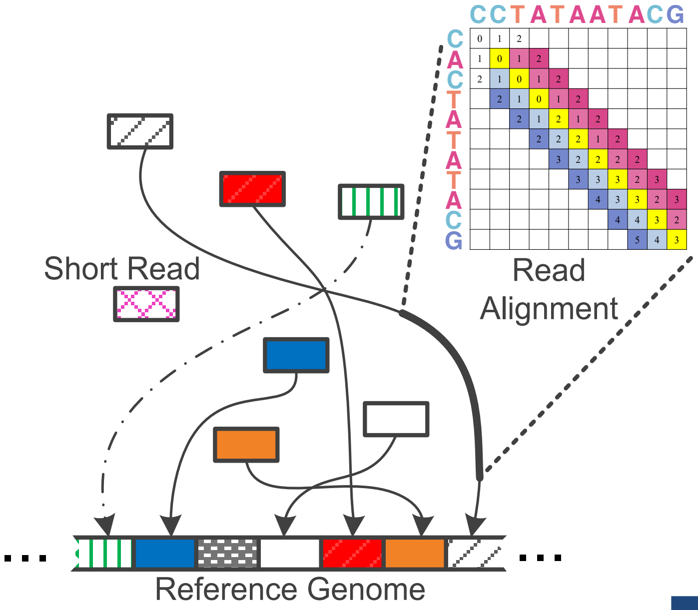
Check for updates

Mohammed Alser[1,2,3†], Jeremy Rotman[4†], Dhrithi Deshpande[5], Kodi Taraszka[4], Huwenbo Shi[6,7], Pelin Icer Baykal[8], Harry Taegyun Yang[4,9], Victor Xue[4], Sergey Knyazev[8], Benjamin D. Singer[10,11,12], Brunilda Balliu[13], David Koslicki[14,15,16], Pavel Skums[8], Alex Zelikovsky[8,17], Can Alkan[2,18], Onur Mutlu[1,2,3†] and Serghei Mangul[5*†]

One Problem

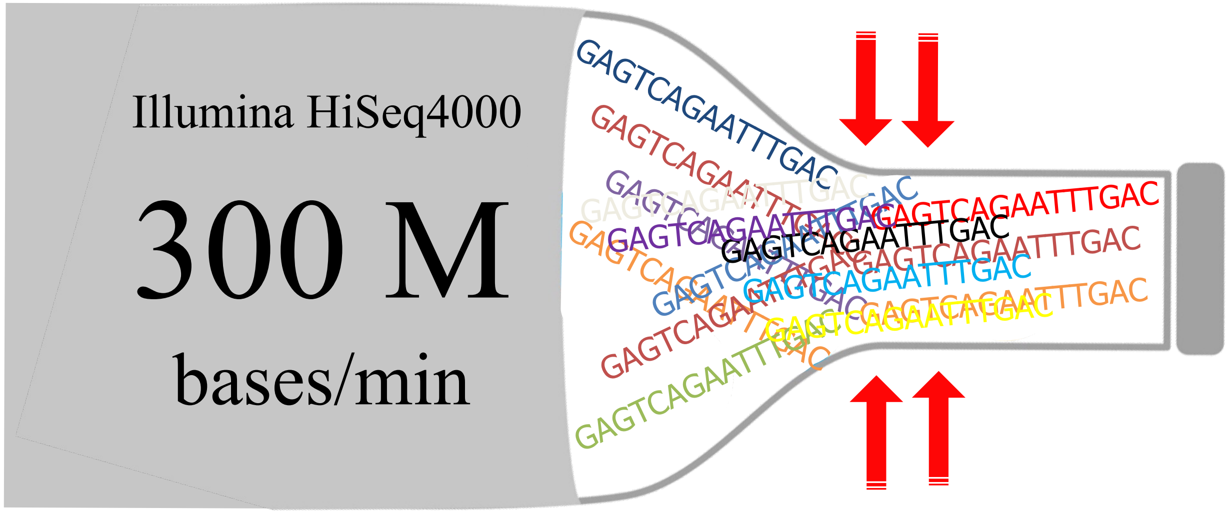


1 Sequencing



Read Mapping 2

We Are Bottlenecked in Read Mapping



Illumina HiSeq4000

300 M

bases/min

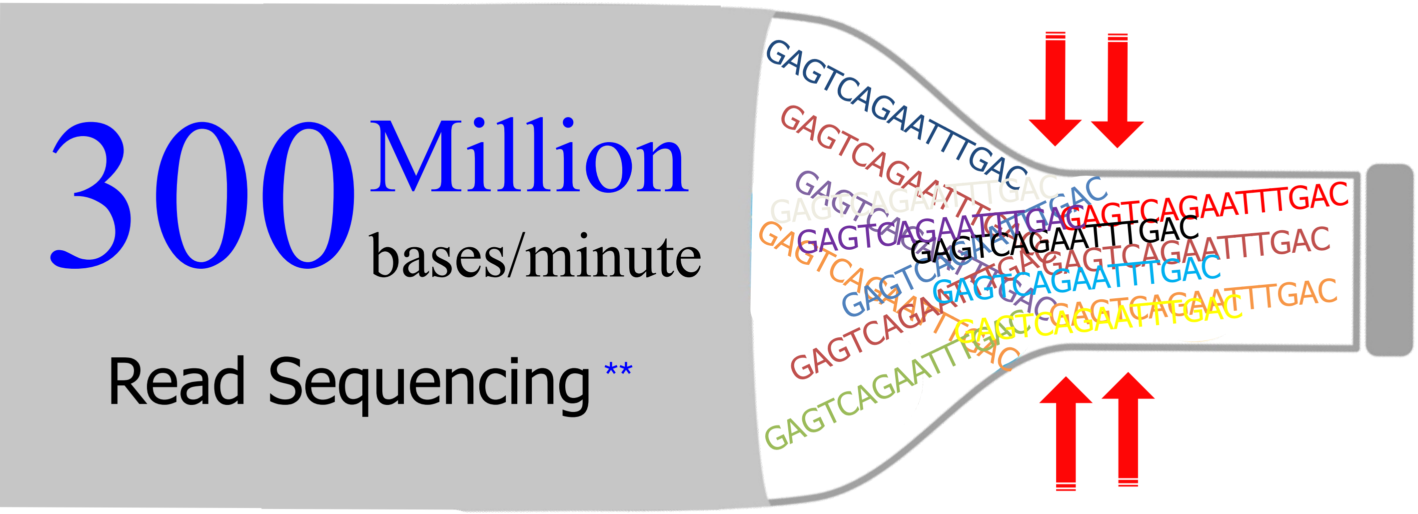
on average

2 M

bases/min

(0.6%)

The Read Mapping Bottleneck



300 Million bases/minute

Read Sequencing **

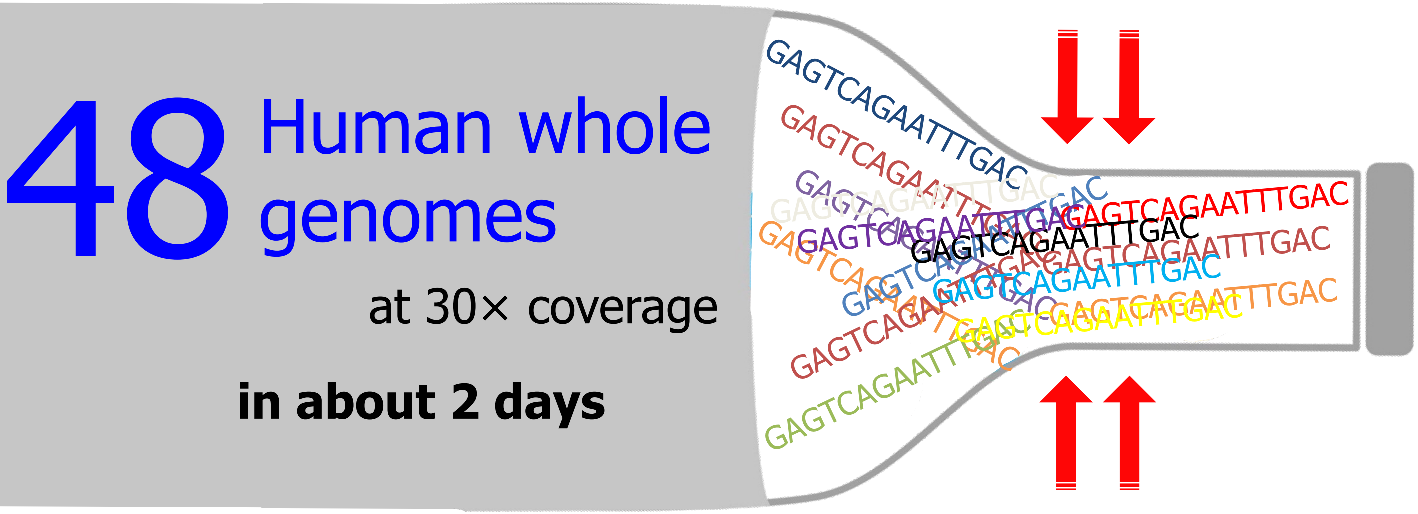
2 Million bases/minute

Read Mapping *

150x slower

* BWA-MEM
** HiSeqX10, MinION

The Read Mapping Bottleneck



48 Human whole genomes

at 30× coverage

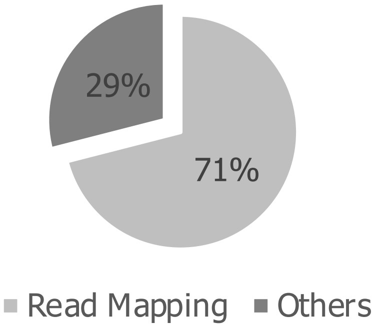
in about 2 days

Illumina NovaSeq 6000

1 Human genome

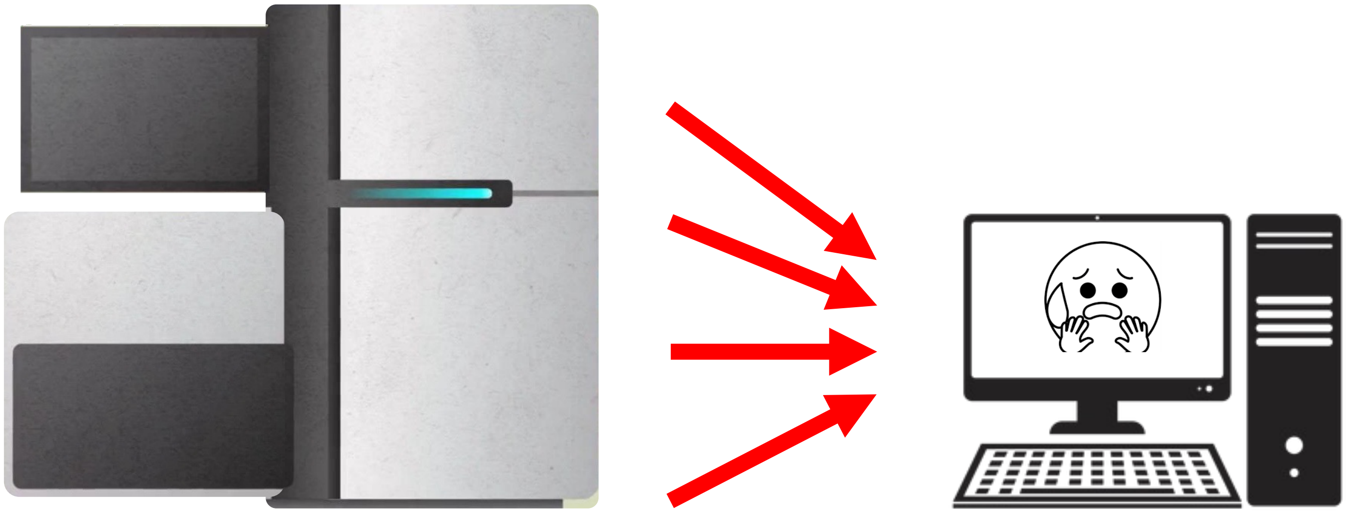
32 CPU hours

on a 48-core processor



Goyal+, "Ultra-fast next generation human genome sequencing data processing using DRAGENTM bio-IT processor for precision medicine", *Open Journal of Genetics,* 2017.

Problems with (Genome) Analysis Today



Special-Purpose Machine
for **Data Generation**

General-Purpose Machine
for **Data Analysis**

FAST

SLOW

Slow and inefficient processing capability
Large amounts of data movement

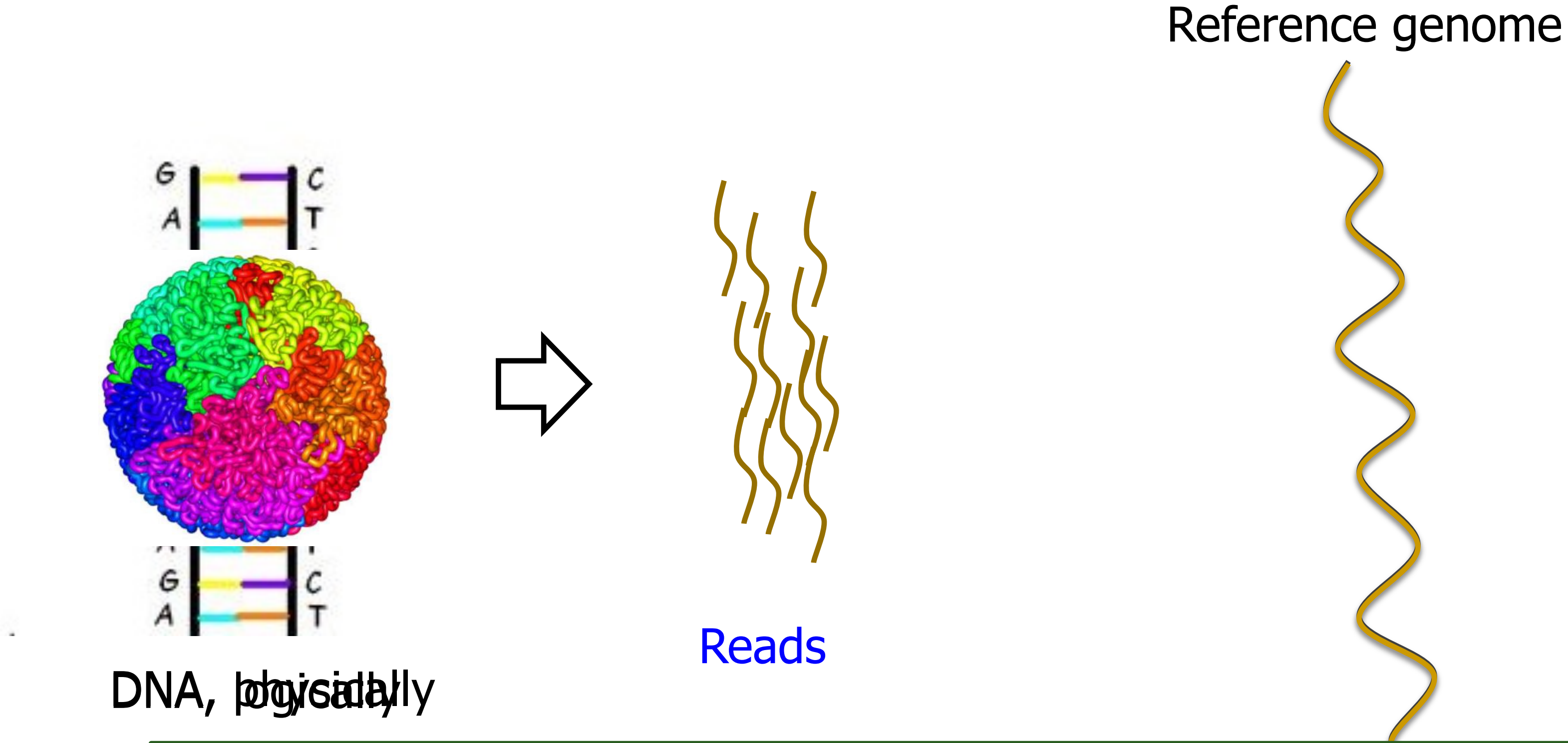
SAFARI This picture is similar for many "data generators & analyzers" today

One Problem

Need to construct
the entire genome
from many sequenced reads

Read Mapping

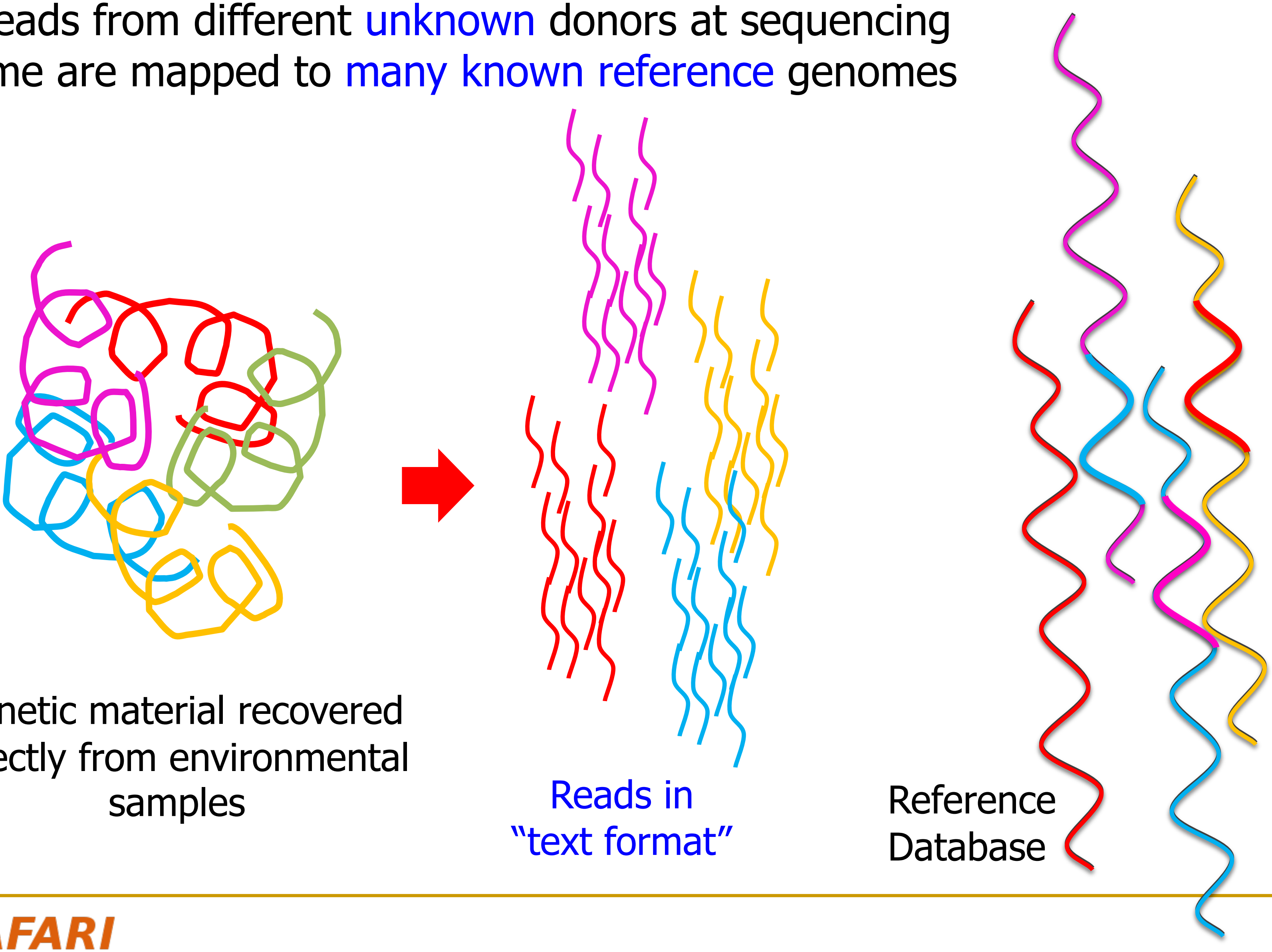
- Map many short DNA fragments (reads) to a known reference genome with some differences allowed



Mapping short reads to reference genome is challenging (billions of 50-300 base pair reads)

Read Mapping for Metagenomic Analysis

Reads from different unknown donors at sequencing time are mapped to many known reference genomes

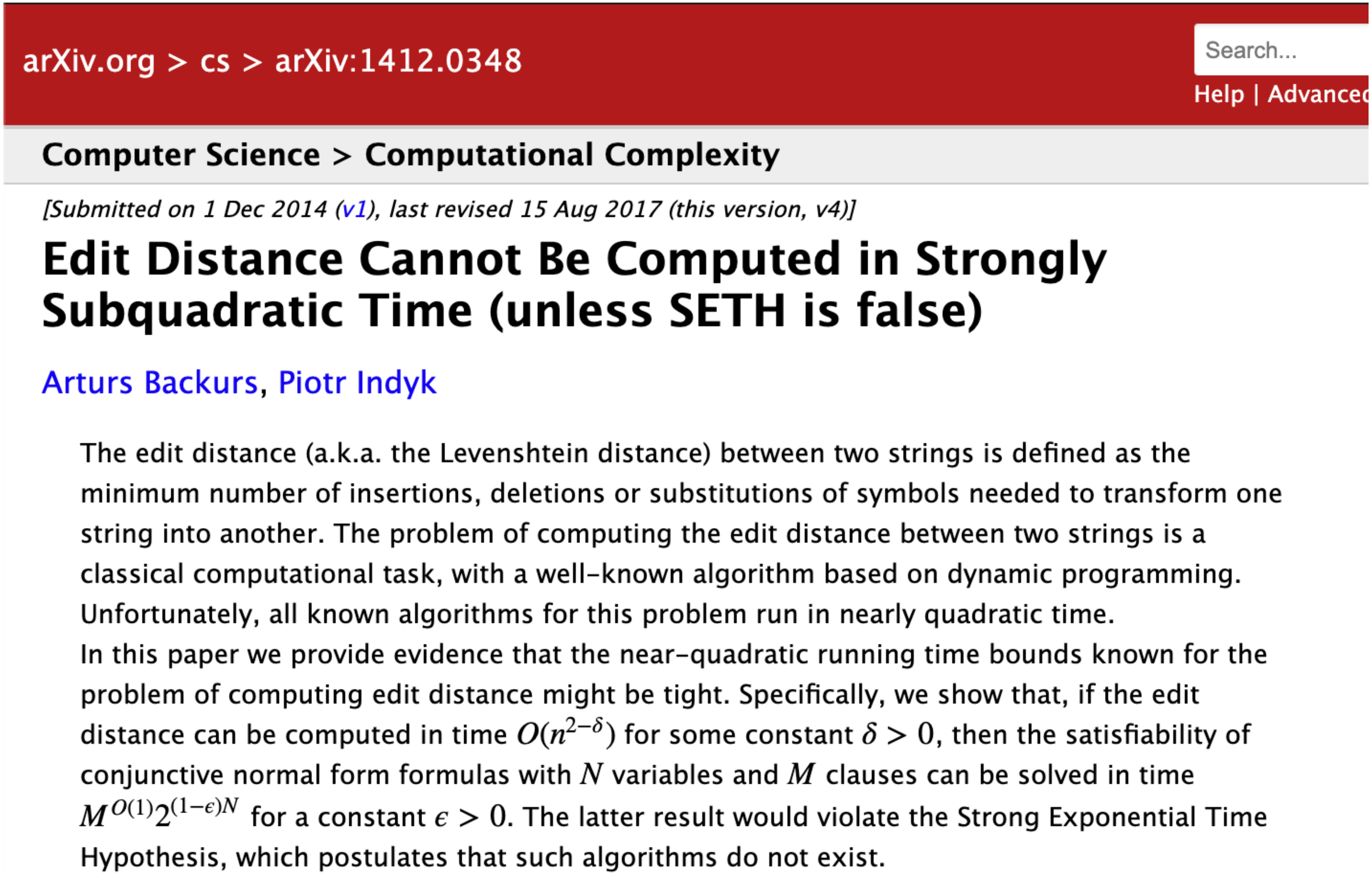


Genetic material recovered directly from environmental samples

Reads in “text format”

Reference Database

Computational Cost is Mathematically Proven



Computer Science > Computational Complexity

[Submitted on 1 Dec 2014 (v1), last revised 15 Aug 2017 (this version, v4)]

Edit Distance Cannot Be Computed in Strongly Subquadratic Time (unless SETH is false)

Arturs Backurs, Piotr Indyk

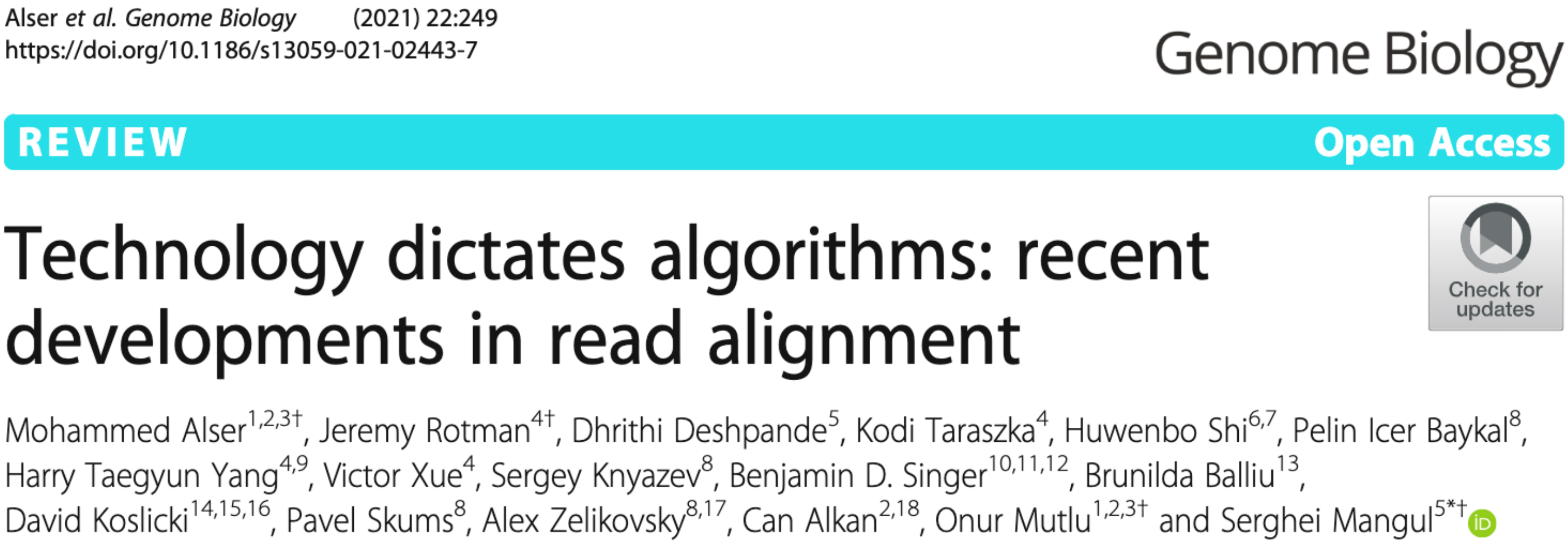
The edit distance (a.k.a. the Levenshtein distance) between two strings is defined as the minimum number of insertions, deletions or substitutions of symbols needed to transform one string into another. The problem of computing the edit distance between two strings is a classical computational task, with a well-known algorithm based on dynamic programming. Unfortunately, all known algorithms for this problem run in nearly quadratic time.
In this paper we provide evidence that the near-quadratic running time bounds known for the problem of computing edit distance might be tight. Specifically, we show that, if the edit distance can be computed in time $O(n^{2-\delta})$ for some constant $\delta > 0$, then the satisfiability of conjunctive normal form formulas with N variables and M clauses can be solved in time $M^{O(1)} 2^{(1-\epsilon)N}$ for a constant $\epsilon > 0$. The latter result would violate the Strong Exponential Time Hypothesis, which postulates that such algorithms do not exist.

https://arxiv.org/abs/1412.0348

Read Mapping Techniques in 111 Pages

In-depth analysis of 107 read mappers (1988-2020)

Mohammed Alser, Jeremy Rotman, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, Victor Xue, Sergey Knyazev, Benjamin D. Singer, Brunilda Balliu, David Koslicki, Pavel Skums, Alex Zelikovsky,
Can Alkan, Onur Mutlu, Serghei Mangul
"Technology dictates algorithms: Recent developments in read alignment"
Genome Biology, 2021
[Source code]



Alser *et al. Genome Biology* (2021) 22:249
https://doi.org/10.1186/s13059-021-02443-7

Genome Biology

REVIEW **Open Access**

Technology dictates algorithms: recent developments in read alignment

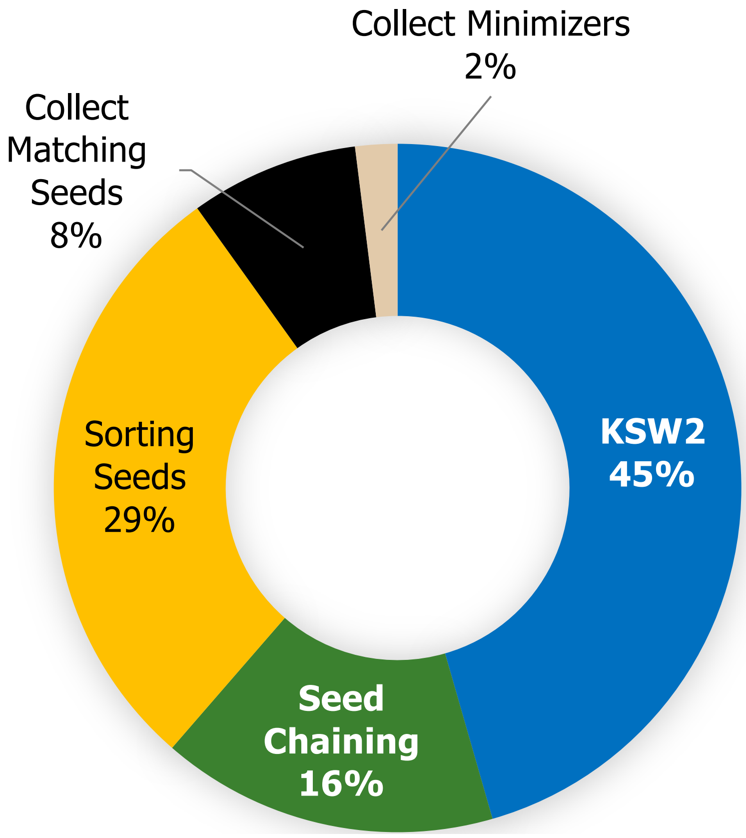
Check for updates

Mohammed Alser[1,2,3†], Jeremy Rotman[4†], Dhrithi Deshpande[5], Kodi Taraszka[4], Huwenbo Shi[6,7], Pelin Icer Baykal[8], Harry Taegyun Yang[4,9], Victor Xue[4], Sergey Knyazev[8], Benjamin D. Singer[10,11,12], Brunilda Balliu[13], David Koslicki[14,15,16], Pavel Skums[8], Alex Zelikovsky[8,17], Can Alkan[2,18], Onur Mutlu[1,2,3†] and Serghei Mangul[5*†]

Read Mapping Execution Time (Modern)

>60%

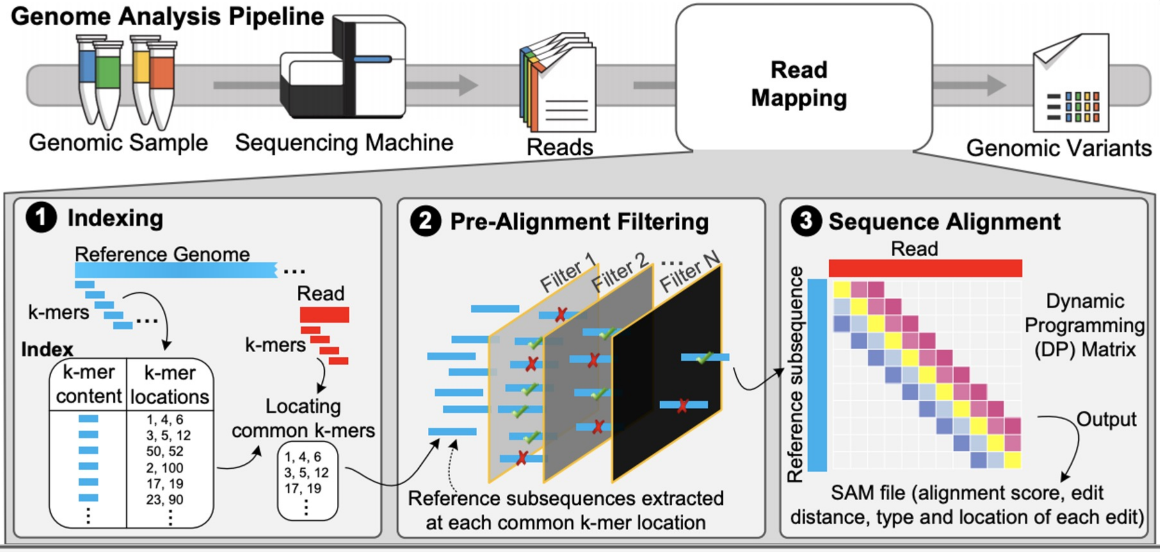
of the read mapper's execution time is spent in sequence alignment



minimap2

ONT FASTQ size: 103MB (151 reads), Mean length: 356,403 bp, std: 173,168 bp, longest length: 817,917 bp

Accelerating Read Mapping

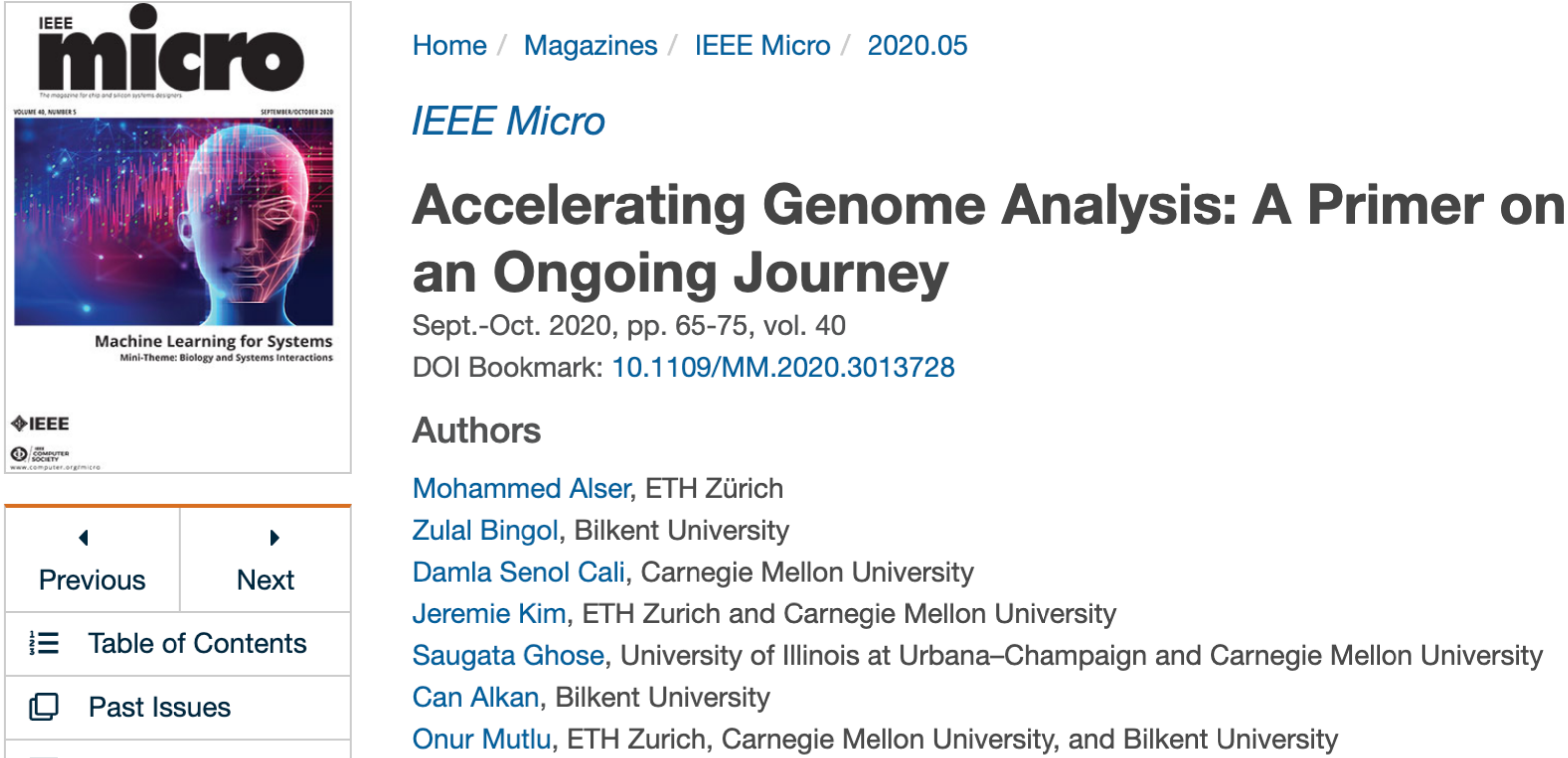


Accelerating Indexing	Accelerating Pre-Alignment Filtering	Accelerating Alignment
Reducing the number of seeds	q-gram filtering	Accurate alignment accelerators
Reducing data movement during indexing	Pigeonhole principle	Heuristic-based alignment accelerators
	Base counting	
	Sparse DP	

Alser+, "Accelerating Genome Analysis: A Primer on an Ongoing Journey", IEEE Micro, 2020.

Detailed Analysis of Tackling the Bottleneck

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu
"Accelerating Genome Analysis: A Primer on an Ongoing Journey"
IEEE Micro, August 2020.



Home / Magazines / IEEE Micro / 2020.05

IEEE Micro

Accelerating Genome Analysis: A Primer on an Ongoing Journey

Sept.-Oct. 2020, pp. 65-75, vol. 40
DOI Bookmark: 10.1109/MM.2020.3013728

Authors

Mohammed Alser, ETH Zürich
Zulal Bingol, Bilkent University
Damla Senol Cali, Carnegie Mellon University
Jeremie Kim, ETH Zurich and Carnegie Mellon University
Saugata Ghose, University of Illinois at Urbana–Champaign and Carnegie Mellon University
Can Alkan, Bilkent University
Onur Mutlu, ETH Zurich, Carnegie Mellon University, and Bilkent University

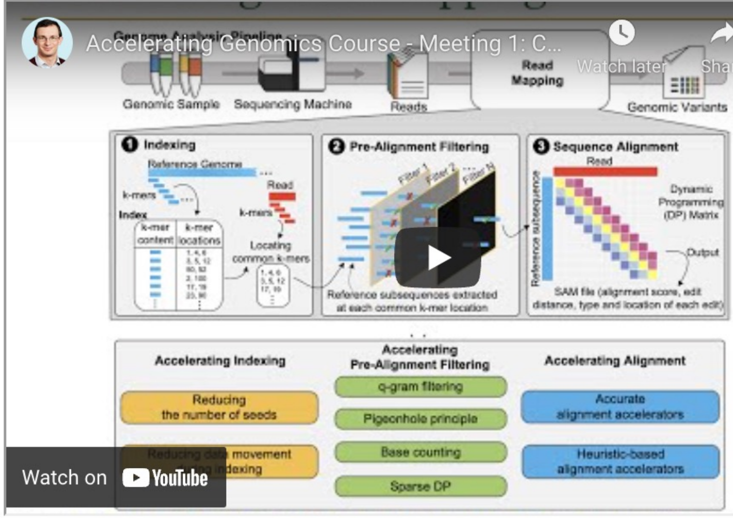
Genomics Course (Fall 2022)

- **Fall 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/fall2022/doku.php?id=bioinformatics
- **Spring 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=bioinformatics

- **Youtube Livestream (Fall 2022):**
  - https://www.youtube.com/watch?v=nA41964-9r8&list=PL5Q2soXY2Zi8tFlQvdxOdizD_EhVAMVQV
- **Youtube Livestream (Spring 2022):**
  - https://www.youtube.com/watch?v=DEL_5A_Y3TI&list=PL5Q2soXY2Zi8NrPDgOR1yRU_Cxxjw-u18

- Project course
  - Taken by Bachelor's/Master's students
  - Genomics lectures
  - Hands-on research exploration
  - Many research readings

https://www.youtube.com/onurmutlulectures



Spring 2022 Meetings/Schedule

Week	Date	Livestream	Meeting	Learning Materials
W1	11.3 Fri.	YouTube Live	M1: P&S Accelerating Genomics Course Introduction & Project Proposals (PDF) (PPT)	Required Materials Recommended Materials
W2	18.3 Fri.	YouTube Live	M2: Introduction to Sequencing (PDF) (PPT)	
W3	25.3 Fri.	YouTube Premiere	M3: Read Mapping (PDF) (PPT)	
W4	01.04 Fri.	YouTube Premiere	M4: GateKeeper (PDF) (PPT)	
W5	08.04 Fri.	YouTube Premiere	M5: MAGNET & Shouji (PDF) (PPT)	
W6	15.4 Fri.	YouTube Premiere	M6: SneakySnake (PDF) (PPT)	
W7	29.4 Fri.	YouTube Premiere	M7: GenStore (PDF) (PPT)	
W8	06.05 Fri.	YouTube Premiere	M8: GRIM-Filter (PDF) (PPT)	
W9	13.05 Fri.	YouTube Premiere	M9: Genome Assembly (PDF) (PPT)	
W10	20.05 Fri.	YouTube Live	M10: Genomic Data Sharing Under Differential Privacy (PDF) (PPT)	
W11	10.06 Fri.	YouTube Premiere	M11: Accelerating Genome Sequence Analysis (PDF) (PPT)	

Agenda

- The Problem: DNA Read Mapping
  - State-of-the-art Read Mapper Design

- Algorithmic Acceleration
  - Exploiting Structure of the Genome
  - Exploiting SIMD Instructions

- Hardware Acceleration
  - Specialized Architectures
  - Processing in Memory & Storage

- Future Opportunities: New Technologies & Applications

GateKeeper: FPGA-Based Alignment Filtering

- Mohammed Alser, Hasan Hassan, Hongyi Xin, Oguz Ergin, Onur Mutlu, and Can Alkan
  "GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping"
  Bioinformatics, [published online, May 31], 2017.
  [Source Code]
  [Online link at Bioinformatics Journal]

GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping

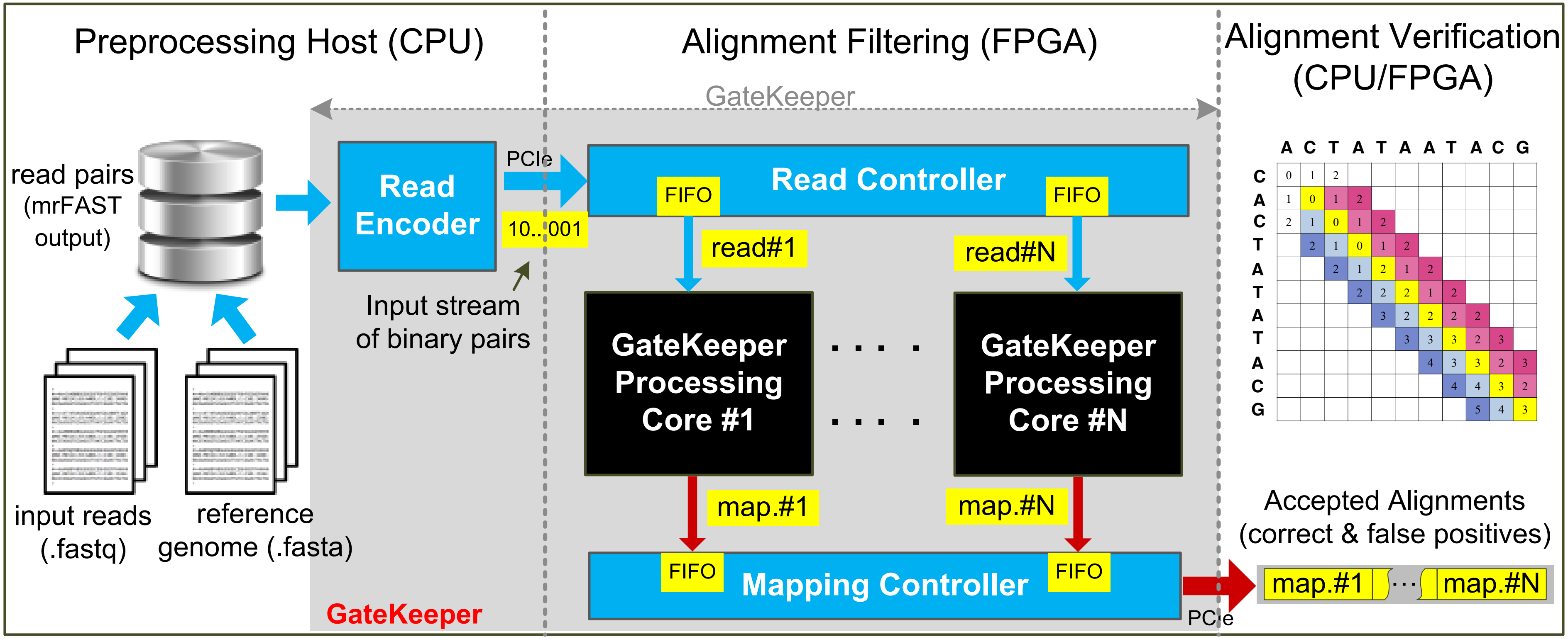
Mohammed Alser ✉, Hasan Hassan, Hongyi Xin, Oğuz Ergin, Onur Mutlu ✉, Can Alkan ✉

Bioinformatics, Volume 33, Issue 21, 1 November 2017, Pages 3355–3363,
https://doi.org/10.1093/bioinformatics/btx342

Published: 31 May 2017 **Article history** ▾

GateKeeper Accelerator Architecture

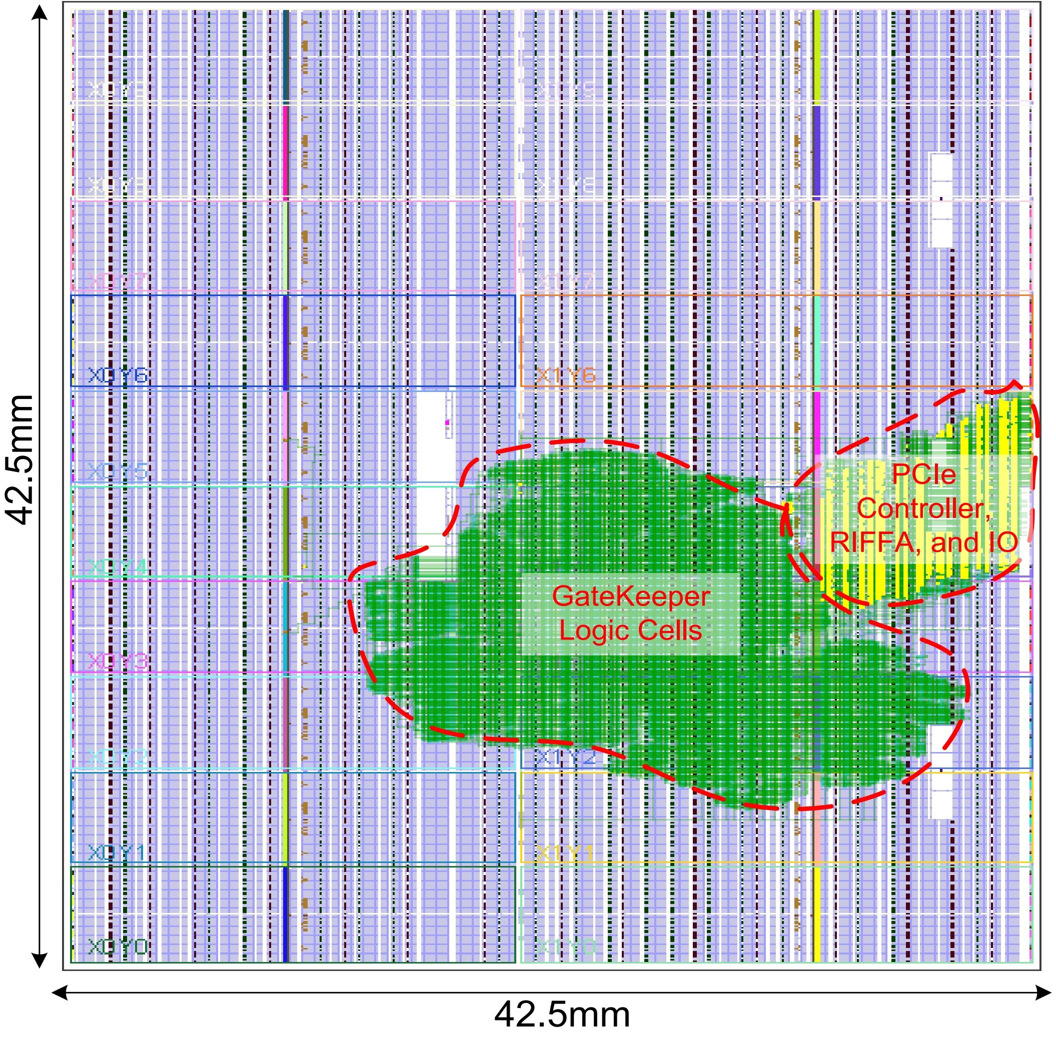
- **Maximum data throughput** =~13.3 billion bases/sec
- Can examine **8 (300 bp) or 16 (100 bp) mappings concurrently** at 250 MHz
- **Occupies 50%** (100 bp) to **91%** (300 bp) of the FPGA slice LUTs and registers



	A	C	T	A	T	A	A	T	A	C	G
C	0	1	2								
A	1	0	1	2							
C	2	1	0	1	2						
T		2	1	0	1	2					
A			2	1	2	1	2				
T				2	2	2	1	2			
A					3	2	2	2	2		
T						3	3	3	2	3	
A							4	3	3	2	3
C								4	4	3	2
G									5	4	3

FPGA Chip Layout

GateKeeper: 17.6%, PCIe Controller, RIFFA, and IO: 5%



Read length:

300 bp

Error threshold:

E=15

GateKeeper: Speed & Accuracy Results

90x-130x faster filter

than SHD (Xin et al., 2015) and the Adjacency Filter (Xin et al., 2013)

4x lower false accept rate

than the Adjacency Filter (Xin et al., 2013)

10x speedup in read mapping

with the addition of GateKeeper to the mrFAST mapper (Alkan et al., 2009)

Freely available online

github.com/BilkentCompGen/GateKeeper

GateKeeper Conclusions

- FPGA-based pre-alignment greatly speeds up read mapping
  - 10x speedup of a state-of-the-art mapper (mrFAST)

- FPGA-based pre-alignment can be integrated with the sequencer
  - It can help to hide the complexity and details of the FPGA
  - Enables **real-time filtering** while sequencing
  - Paves the way to **on-device genome analysis**

More on GateKeeper

- Mohammed Alser, Hasan Hassan, Hongyi Xin, Oguz Ergin, Onur Mutlu, and Can Alkan
  "GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping"
  Bioinformatics, [published online, May 31], 2017.
  [Source Code]
  [Online link at Bioinformatics Journal]

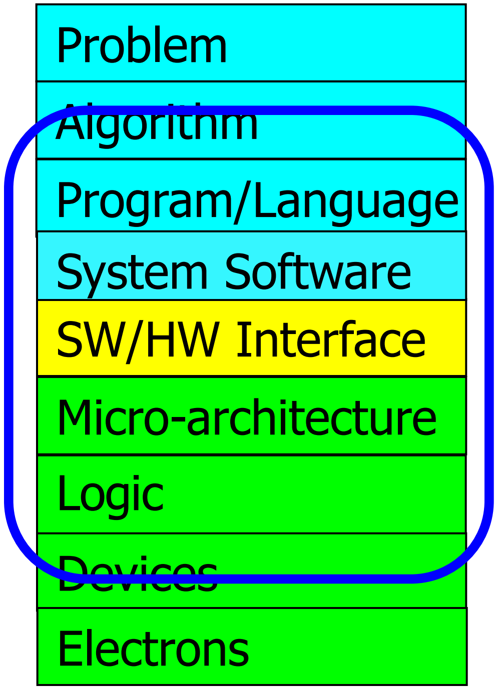
GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping

Mohammed Alser ✉, Hasan Hassan, Hongyi Xin, Oğuz Ergin, Onur Mutlu ✉, Can Alkan ✉

Bioinformatics, Volume 33, Issue 21, 1 November 2017, Pages 3355–3363, https://doi.org/10.1093/bioinformatics/btx342

Published: 31 May 2017 **Article history** ▾

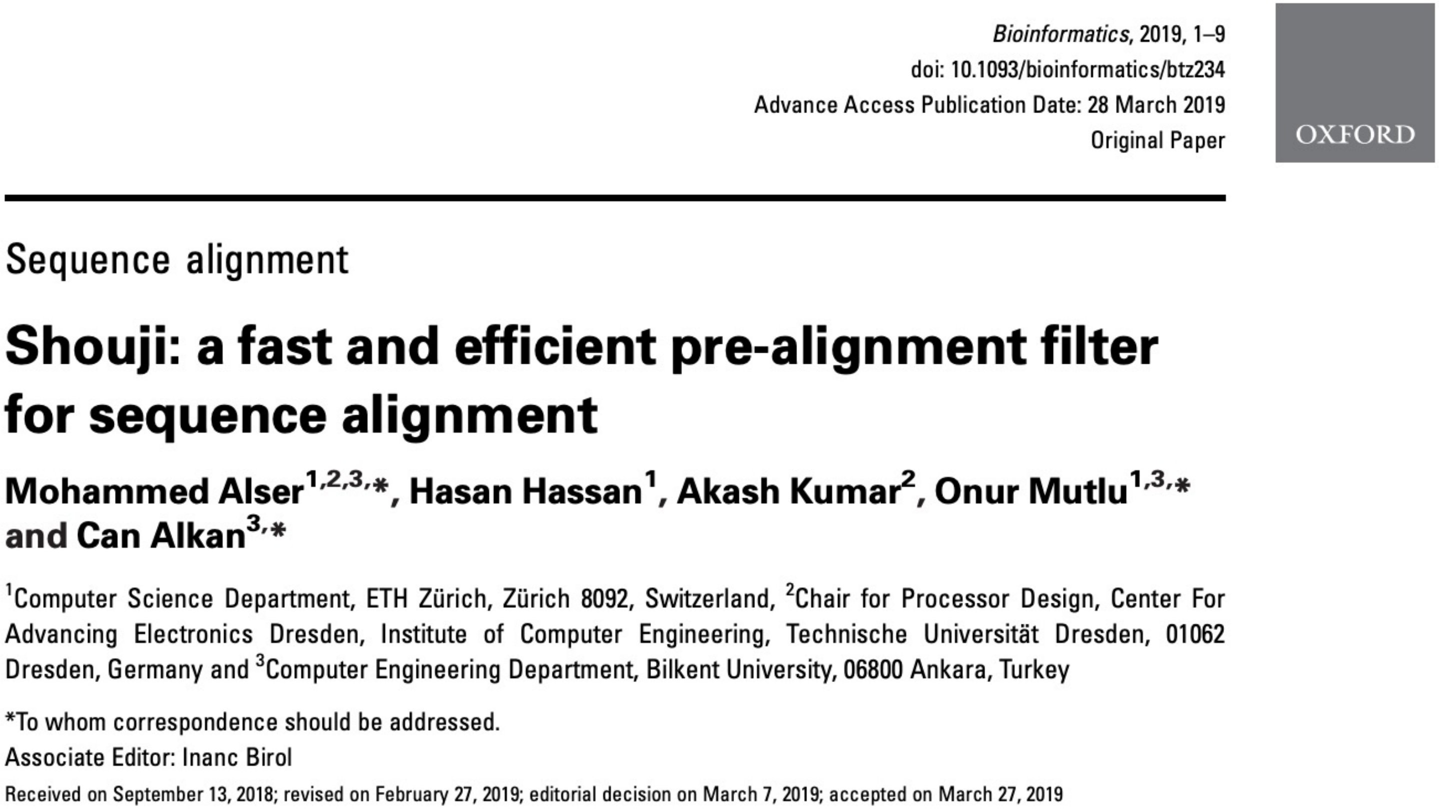
Algorithm-Arch-Device Co-Design is Critical



Computer Architecture (expanded view)

Shouji (障子) [Alser+, Bioinformatics 2019]

Mohammed Alser, Hasan Hassan, Akash Kumar, Onur Mutlu, and Can Alkan,
"Shouji: A Fast and Efficient Pre-Alignment Filter for Sequence Alignment"
Bioinformatics, [published online, March 28], 2019.
[Source Code]
[Online link at Bioinformatics Journal]



Bioinformatics, 2019, 1–9
doi: 10.1093/bioinformatics/btz234
Advance Access Publication Date: 28 March 2019
Original Paper

OXFORD

Sequence alignment

Shouji: a fast and efficient pre-alignment filter for sequence alignment

Mohammed Alser[1,2,3,*], Hasan Hassan[1], Akash Kumar[2], Onur Mutlu[1,3,*] and Can Alkan[3,*]

[1]Computer Science Department, ETH Zürich, Zürich 8092, Switzerland, [2]Chair for Processor Design, Center For Advancing Electronics Dresden, Institute of Computer Engineering, Technische Universität Dresden, 01062 Dresden, Germany and [3]Computer Engineering Department, Bilkent University, 06800 Ankara, Turkey

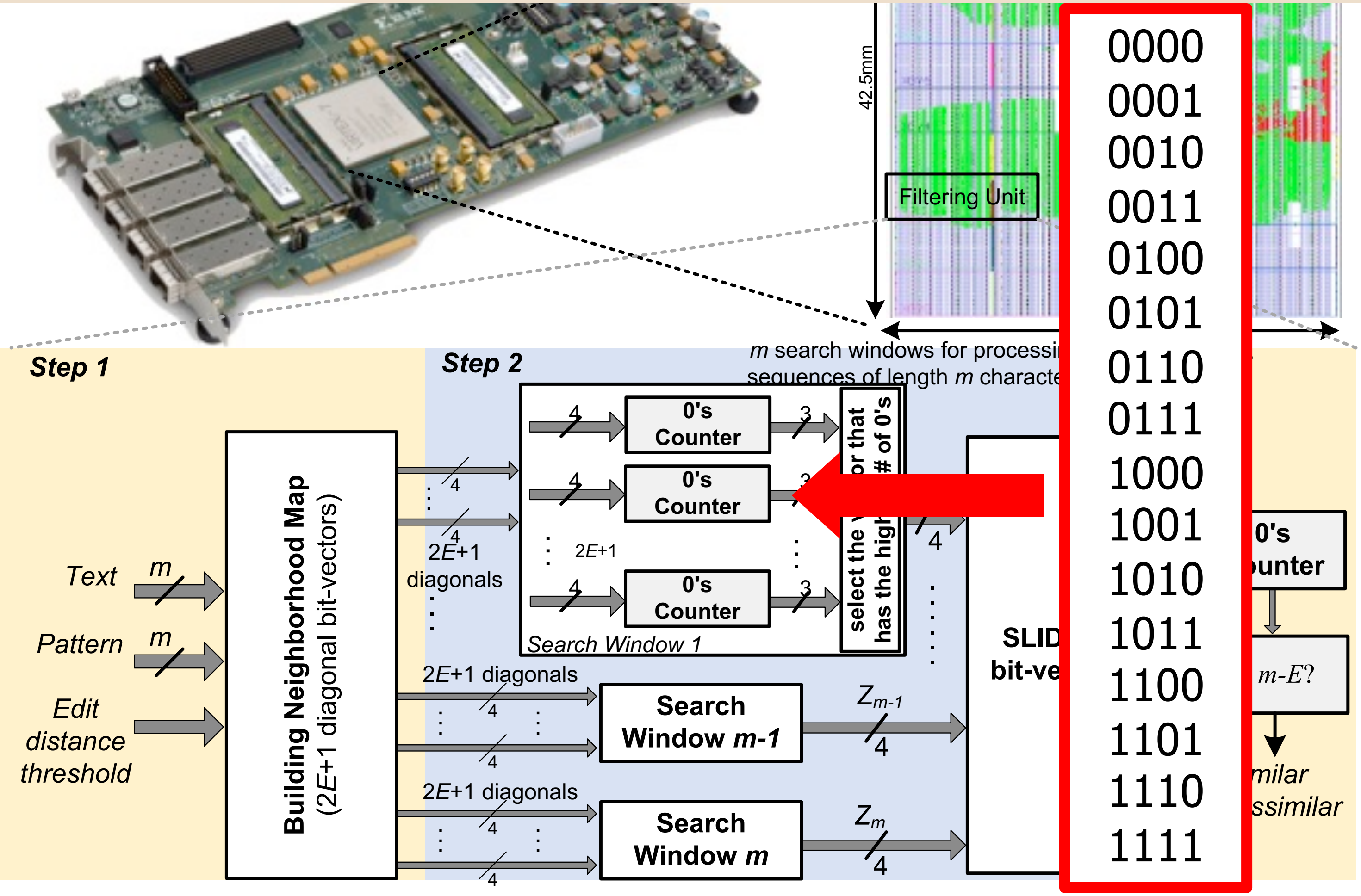
*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on September 13, 2018; revised on February 27, 2019; editorial decision on March 7, 2019; accepted on March 27, 2019

Hardware Implementation

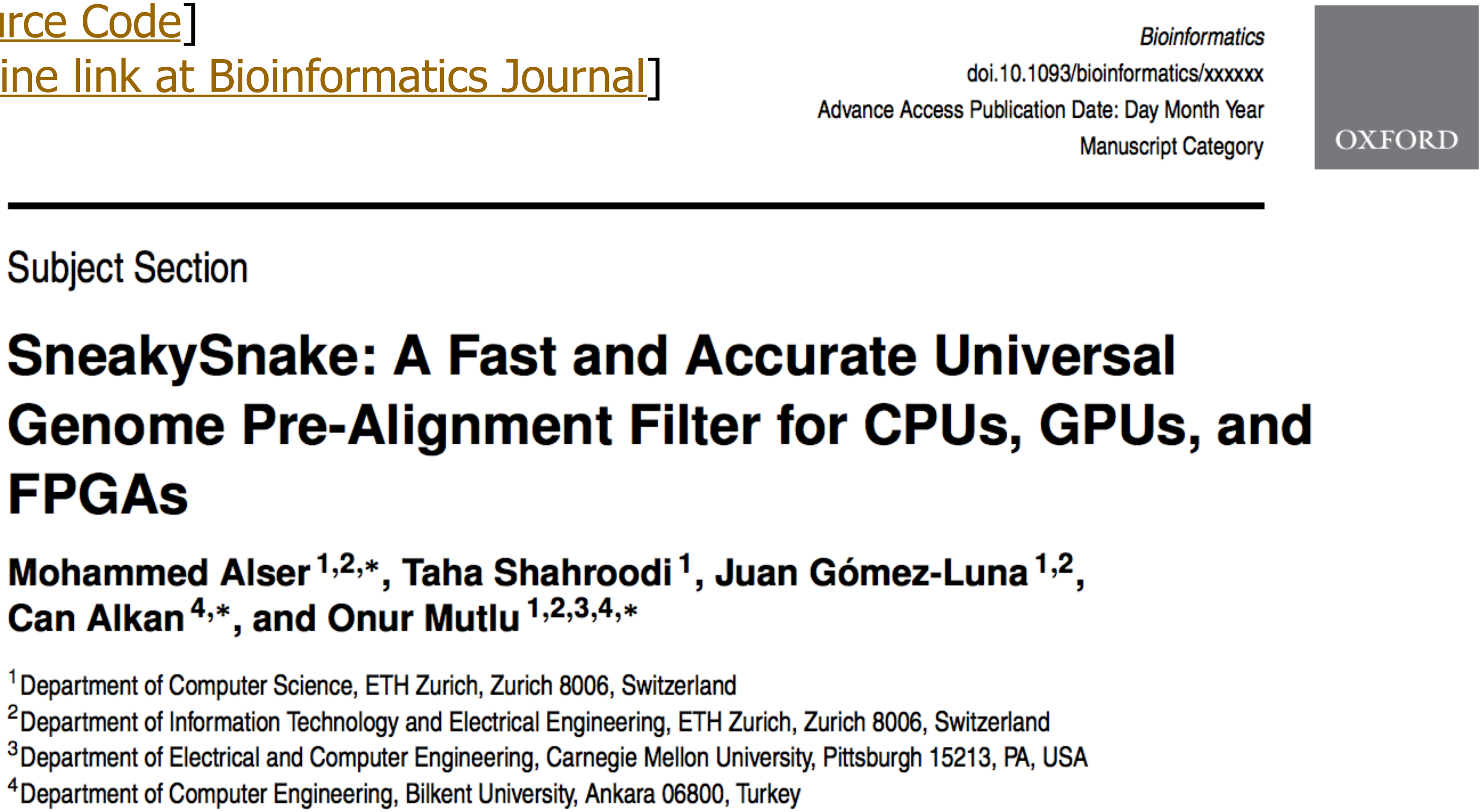
Counting is performed concurrently for all bit-vectors and all sliding windows in a single clock cycle using multiple 4-input LUTs



0000
0001
0010
0011
0100
0101
0110
0111
1000
1001
1010
1011
1100
1101
1110
1111

SneakySnake [Alser+, Bioinformatics 2020]

Mohammed Alser, Taha Shahroodi, Juan-Gomez Luna, Can Alkan, and Onur Mutlu,
"SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs"
Bioinformatics, to appear in 2020.
[Source Code]
[Online link at Bioinformatics Journal]



Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category
OXFORD

Subject Section

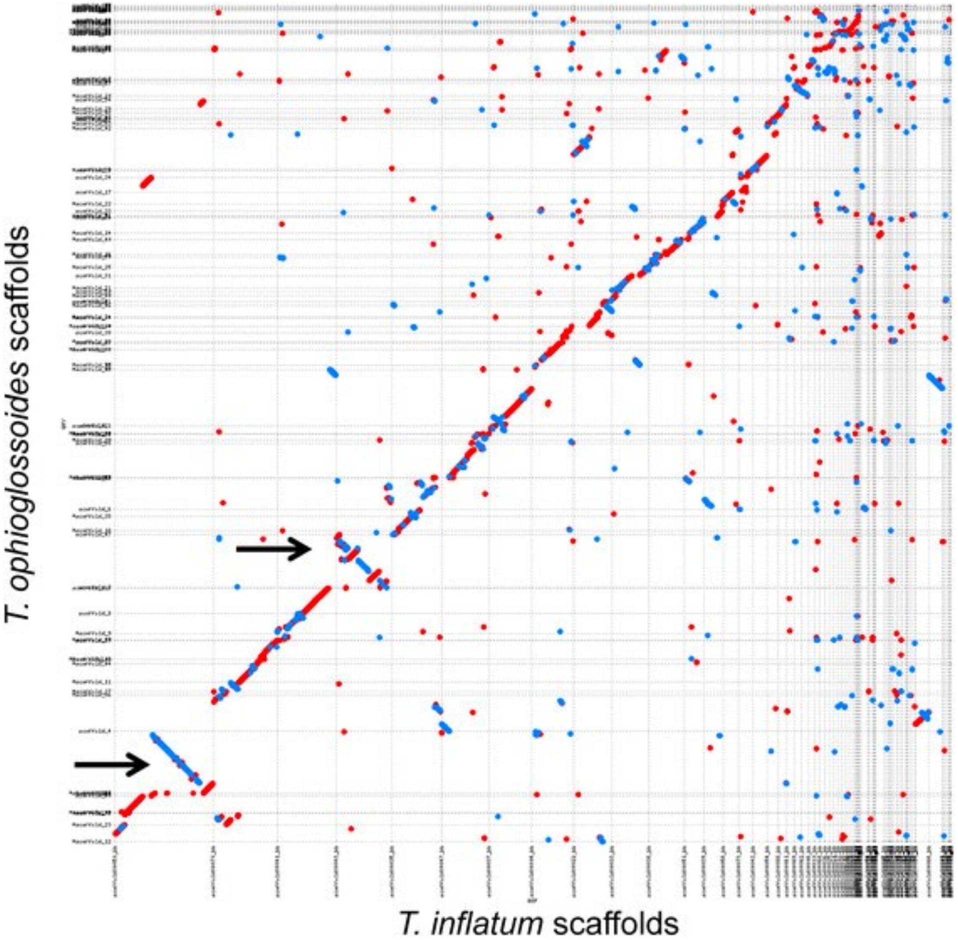
SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs

Mohammed Alser [1,2,*], Taha Shahroodi [1], Juan Gómez-Luna [1,2], Can Alkan [4,*], and Onur Mutlu [1,2,3,4,*]

[1] Department of Computer Science, ETH Zurich, Zurich 8006, Switzerland
[2] Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8006, Switzerland
[3] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh 15213, PA, USA
[4] Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey

SneakySnake

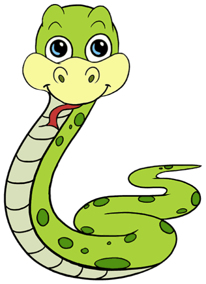
- **Key observation:**
  - Correct alignment is a sequence of non-overlapping long matches

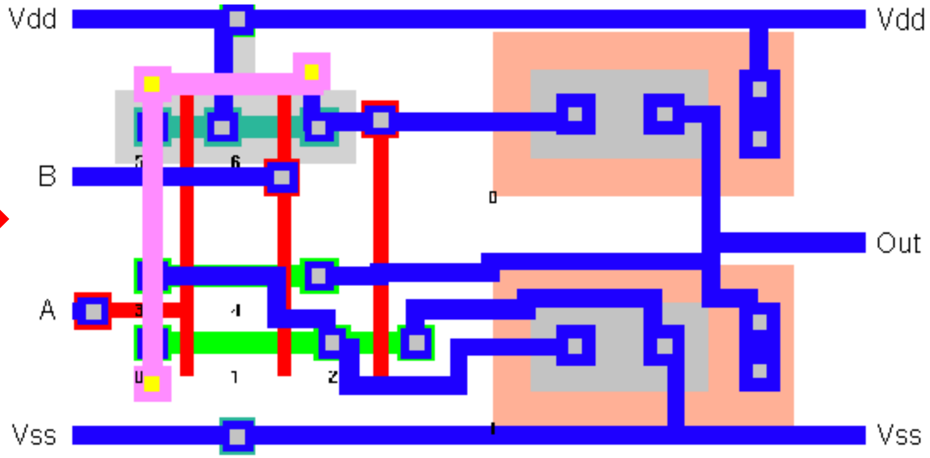


Dot plot, dot matrix
(Lipman and Pearson, 1985)

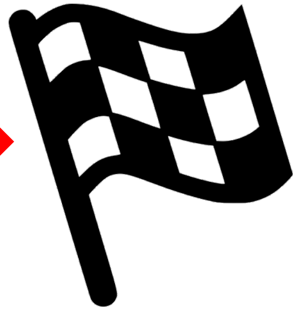
SneakySnake

- **Key observation:**
  - Correct alignment is a sequence of non-overlapping long matches
- **Key idea:**
  - Reduce the approximate string matching problem to the Single Net Routing problem in VLSI chip layout





VLSI chip layout



SneakySnake

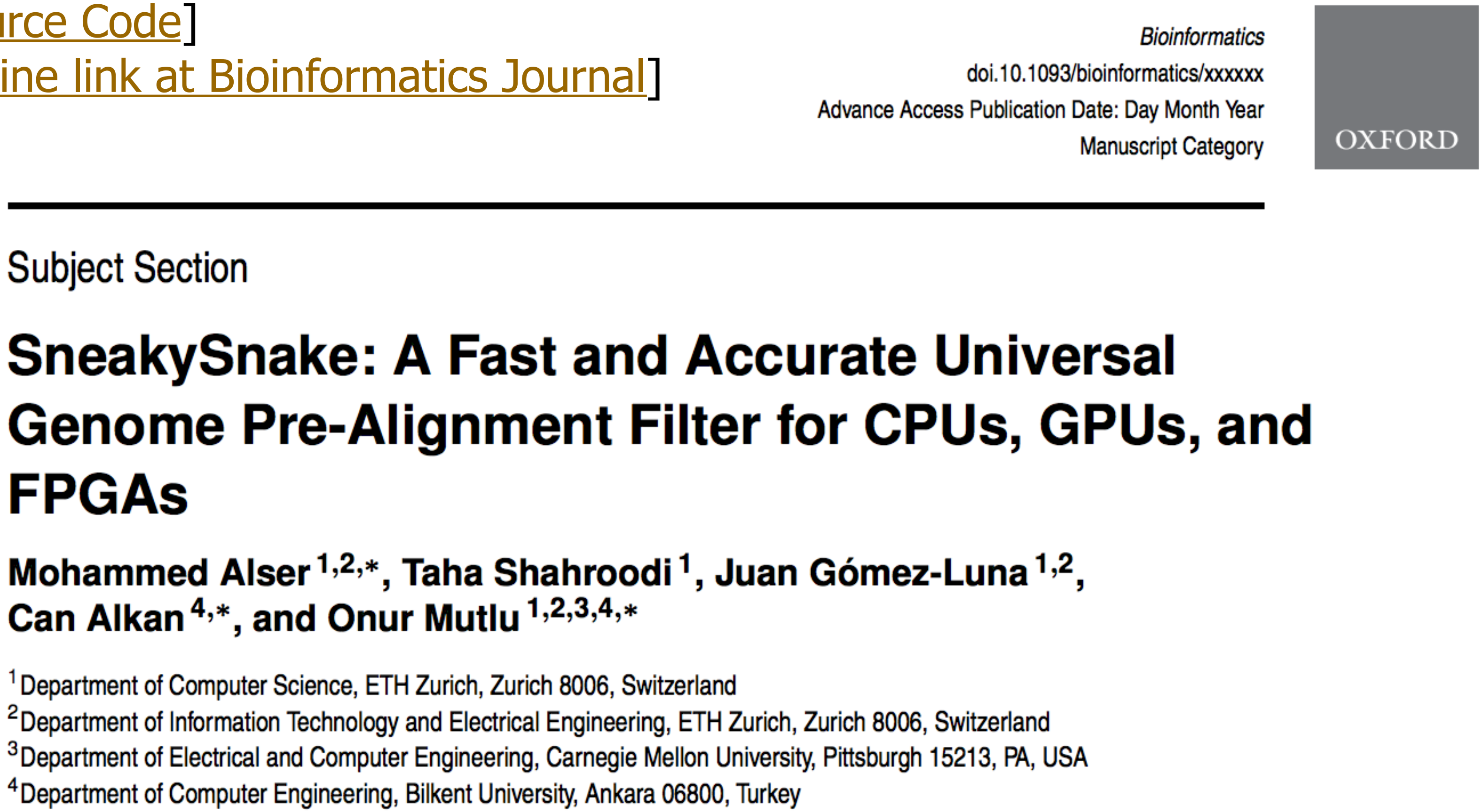
- **Key observation:**
  - Correct alignment is a sequence of non-overlapping long matches

- **Key idea:**
  - Reduce the approximate string matching problem to the Single Net Routing problem in VLSI chip layout

- **Key result:**
  - SneakySnake is up to four orders of magnitude more accurate than Shouji (Bioinformatics'19) and GateKeeper (Bioinformatics'17)
  - SneakySnake greatly accelerates state-of-the-art CPU sequence aligners, Edlib (Bioinformatics'17) and Parasail (BMC Bioinformatics'16)
    - by up to 37.7× and 43.9× (>12× on average), on CPUs
    - by up to 413× and 689× (>400× on average) *with FPGAs/GPUs*

More on SneakySnake [Alser+, Bioinformatics 2020]

Mohammed Alser, Taha Shahroodi, Juan-Gomez Luna, Can Alkan, and Onur Mutlu, **"SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs"**
Bioinformatics, to appear in 2020.
[Source Code]
[Online link at Bioinformatics Journal]



Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

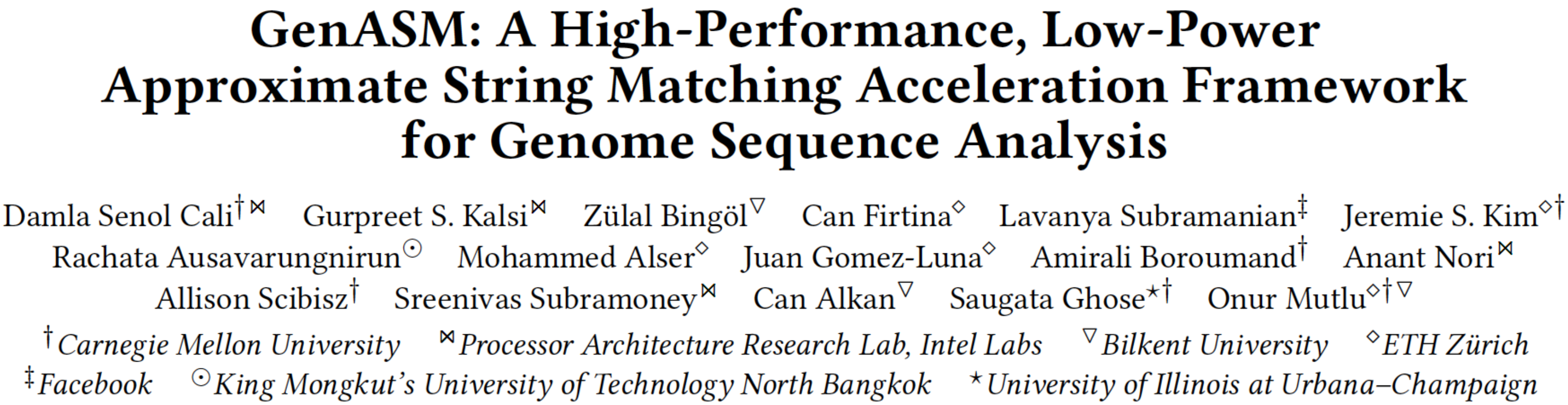
SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs

Mohammed Alser [1,2,*], Taha Shahroodi [1], Juan Gómez-Luna [1,2], Can Alkan [4,*], and Onur Mutlu [1,2,3,4,*]

[1] Department of Computer Science, ETH Zurich, Zurich 8006, Switzerland
[2] Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8006, Switzerland
[3] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh 15213, PA, USA
[4] Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey

GenASM Framework [MICRO 2020]

- Damla Senol Cali, Gurpreet S. Kalsi, Zulal Bingol, Can Firtina, Lavanya Subramanian, Jeremie S. Kim, Rachata Ausavarungnirun, Mohammed Alser, Juan Gomez-Luna, Amirali Boroumand, Anant Nori, Allison Scibisz, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu, **"GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis"**
  Proceedings of the 53rd International Symposium on Microarchitecture (***MICRO***), Virtual, October 2020.
  [Lighting Talk Video (1.5 minutes)]
  [Lightning Talk Slides (pptx) (pdf)]
  [Talk Video (18 minutes)]
  [Slides (pptx) (pdf)]



GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis

Damla Senol Cali[†⋈] Gurpreet S. Kalsi[⋈] Zülal Bingöl[▽] Can Firtina[◇] Lavanya Subramanian[‡] Jeremie S. Kim[◇†]
Rachata Ausavarungnirun[⊙] Mohammed Alser[◇] Juan Gomez-Luna[◇] Amirali Boroumand[†] Anant Nori[⋈]
Allison Scibisz[†] Sreenivas Subramoney[⋈] Can Alkan[▽] Saugata Ghose[⋆†] Onur Mutlu[◇†▽]

[†] *Carnegie Mellon University* [⋈] *Processor Architecture Research Lab, Intel Labs* [▽] *Bilkent University* [◇] *ETH Zürich*
[‡] *Facebook* [⊙] *King Mongkut's University of Technology North Bangkok* [⋆] *University of Illinois at Urbana–Champaign*

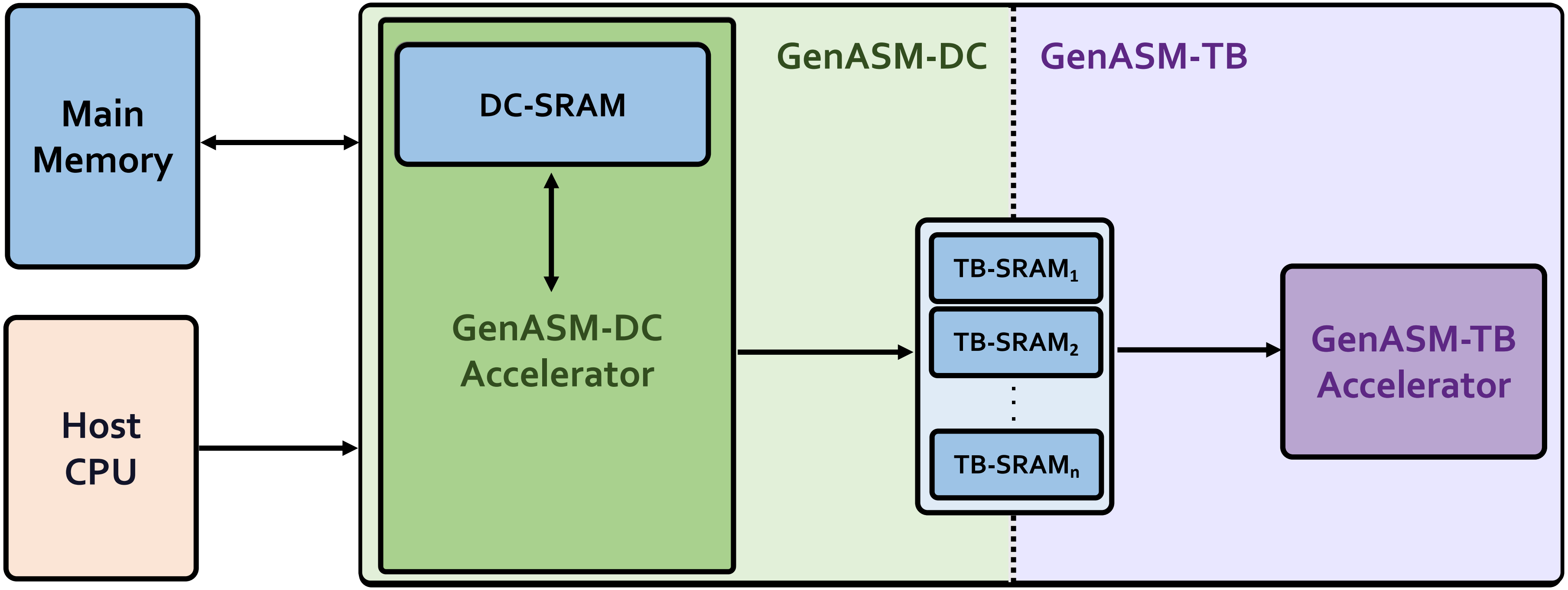
GenASM: ASM Framework for GSA

Our Goal:

Accelerate approximate string matching
by designing a fast and flexible framework,
which can accelerate *multiple steps* of genome sequence analysis

- **GenASM:** *First* ASM acceleration framework for GSA
  - Based on the *Bitap* algorithm
    - Uses fast and simple bitwise operations to perform ASM
  - Modified and extended ASM algorithm
    - Highly-parallel Bitap with long read support
    - Bitvector-based novel algorithm to perform *traceback*
  - Co-design of our modified scalable and memory-efficient algorithms with low-power and area-efficient hardware accelerators

GenASM: Hardware Design

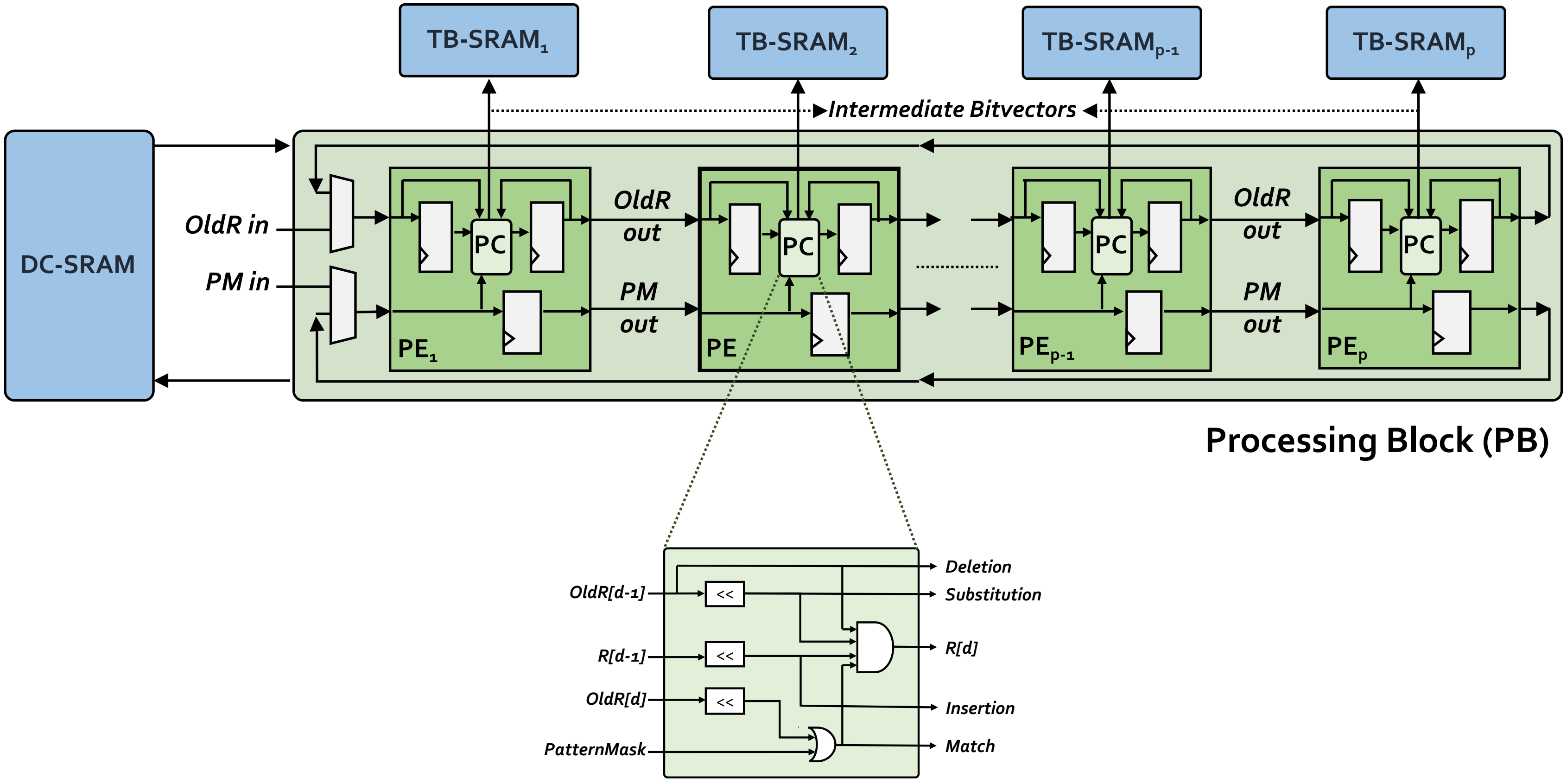


GenASM-DC:
generates bitvectors
and performs edit
Distance **C**alculation

GenASM-TB:
performs **T**race**B**ack
and assembles the
optimal alignment

GenASM-DC: Hardware Design

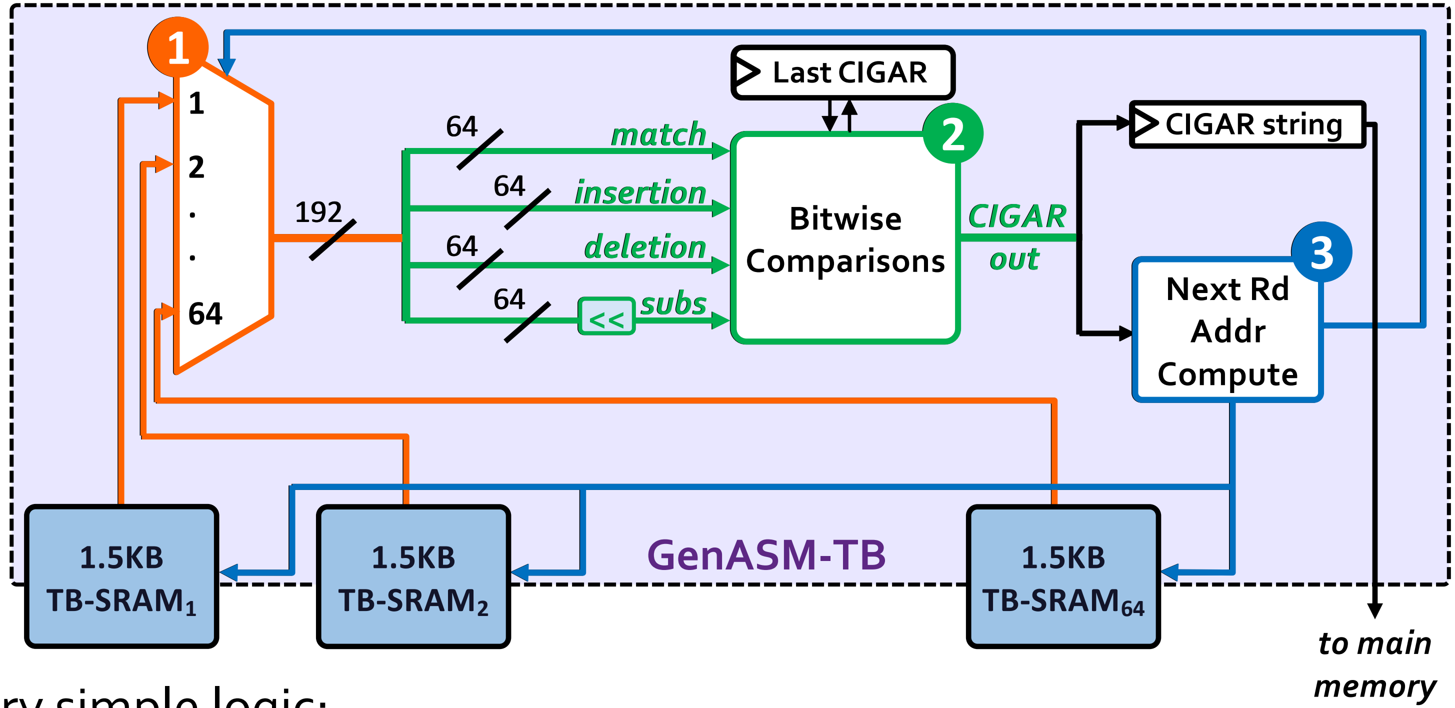
- **Linear cyclic systolic array** based accelerator
  - Designed to maximize parallelism and minimize memory bandwidth and memory footprint



Processing Block (PB)

Processing Core (PC)

GenASM-TB: Hardware Design



- Very simple logic:

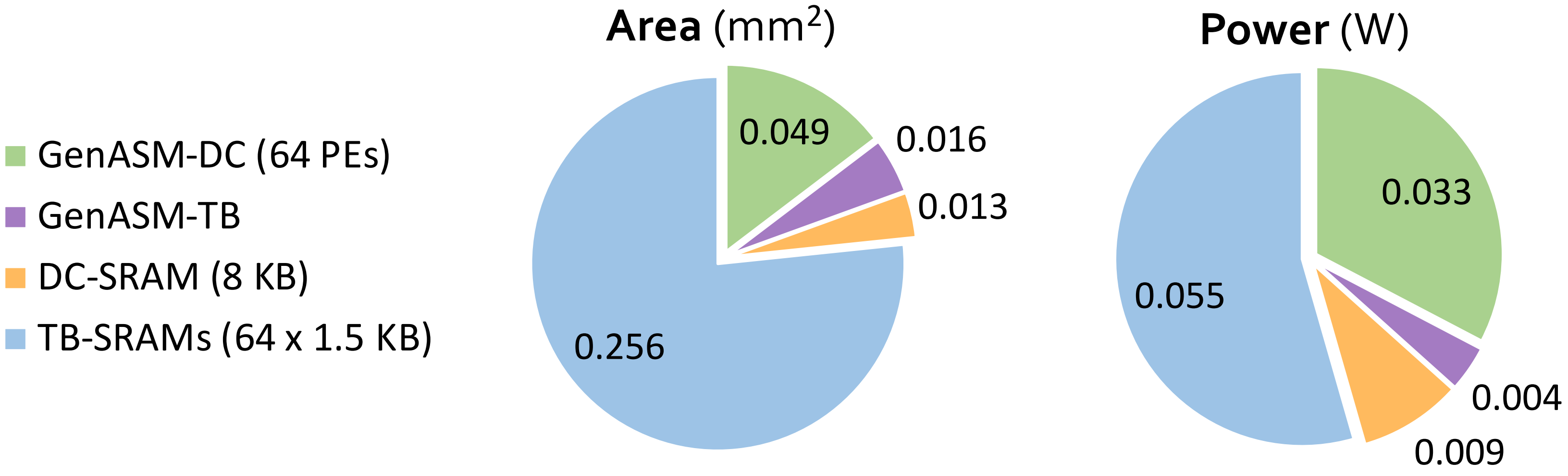
  ① Reads the bitvectors from one of the TB-SRAMs using the computed address

  ② Performs the required bitwise comparisons to find the traceback output for the current position

  ③ Computes the next TB-SRAM address to read the new set of bitvectors

Key Results – Area and Power

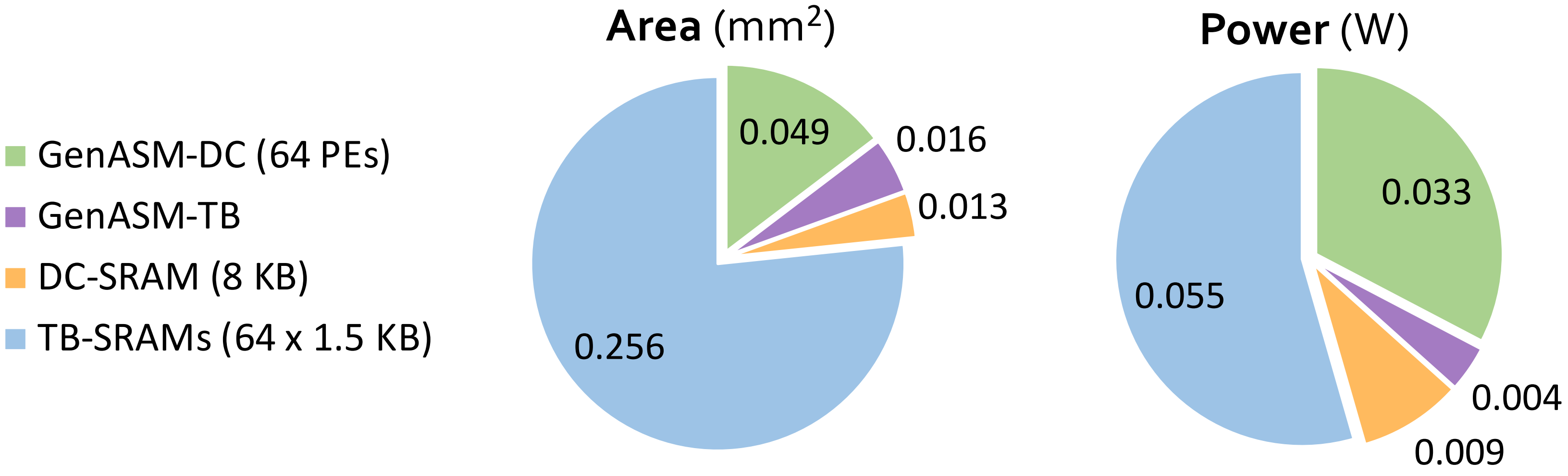
- Based on our **synthesis** of **GenASM-DC** and **GenASM-TB** accelerator datapaths using the Synopsys Design Compiler with a **28nm** LP process:
  - Both GenASM-DC and GenASM-TB operate **@ 1GHz**



Total (1 vault):	0.334 mm^2	0.101 W
Total (32 vaults):	10.69 mm^2	3.23 W
% of a Xeon CPU core:	**1%**	**1%**

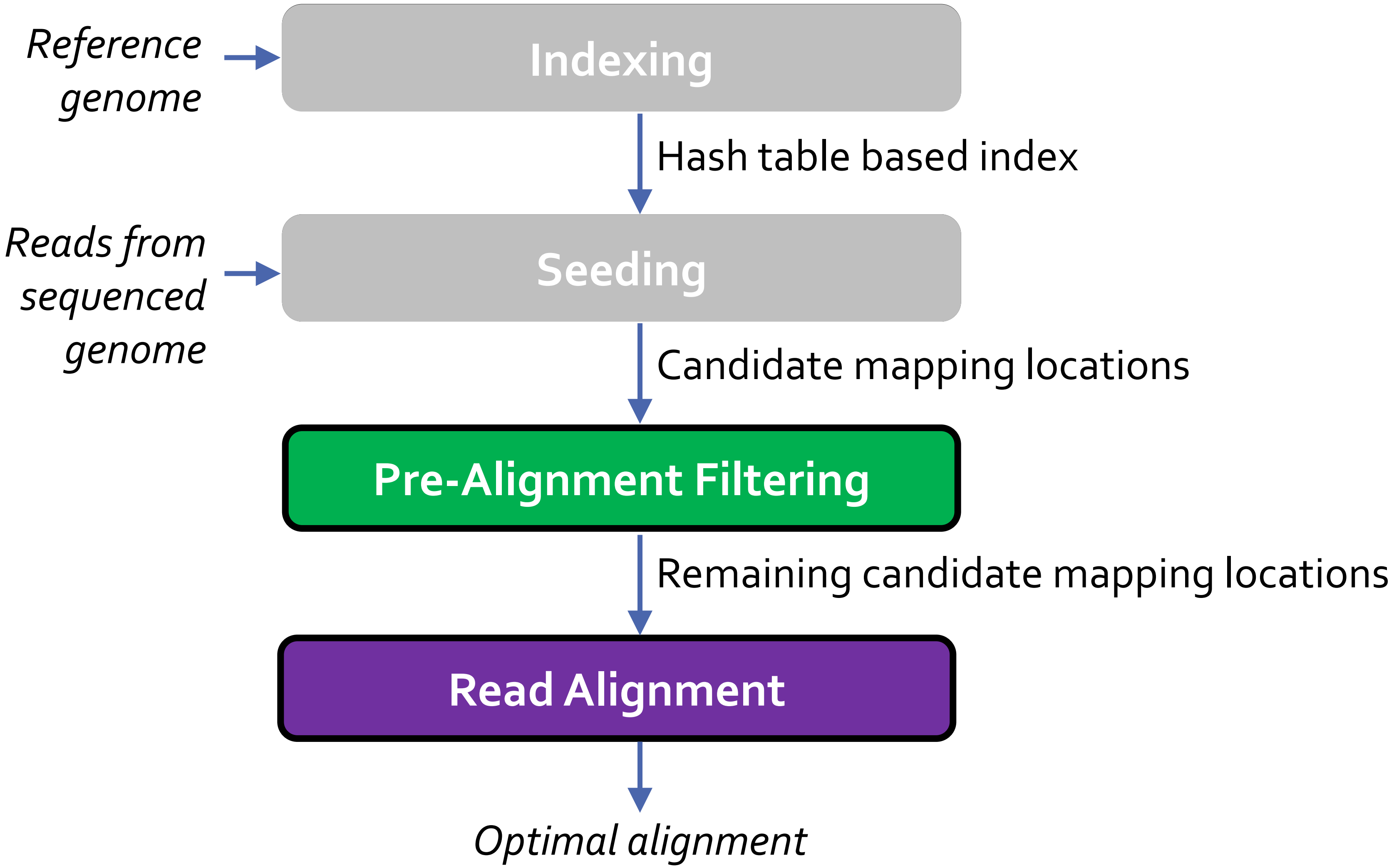
Key Results – Area and Power

- Based on our **synthesis** of **GenASM-DC** and **GenASM-TB** accelerator datapaths using the Synopsys Design Compiler with a **28nm** LP process:
  - Both GenASM-DC and GenASM-TB operate **@ 1GHz**



GenASM has low area and power overheads

Use Cases of GenASM



Reference genome

Indexing

Hash table based index

Reads from sequenced genome

Seeding

Candidate mapping locations

Pre-Alignment Filtering

Remaining candidate mapping locations

Read Alignment

Optimal alignment

Use Cases of GenASM (cont'd.)

(1) Read Alignment Step of Read Mapping

- ○ Find the optimal alignment of how reads map to candidate reference regions

(2) Pre-Alignment Filtering for Short Reads

- ○ Quickly identify and filter out the unlikely candidate reference regions for each read

(3) Edit Distance Calculation

- ○ Measure the similarity or distance between two sequences

- ❑ We also discuss other possible use cases of GenASM in our paper:
  - ○ Read-to-read overlap finding, hash-table based indexing, whole genome alignment, generic text search

Key Results

(1) Read Alignment

- ☐ **116×** speedup, **37×** less power than **Minimap2** (state-of-the-art **SW**)
- ☐ **111×** speedup, **33×** less power than **BWA-MEM** (state-of-the-art **SW**)
- ☐ **3.9×** better throughput, **2.7×** less power than **Darwin** (state-of-the-art **HW**)
- ☐ **1.9×** better throughput, **82%** less logic power than **GenAx** (state-of-the-art **HW**)

(2) Pre-Alignment Filtering

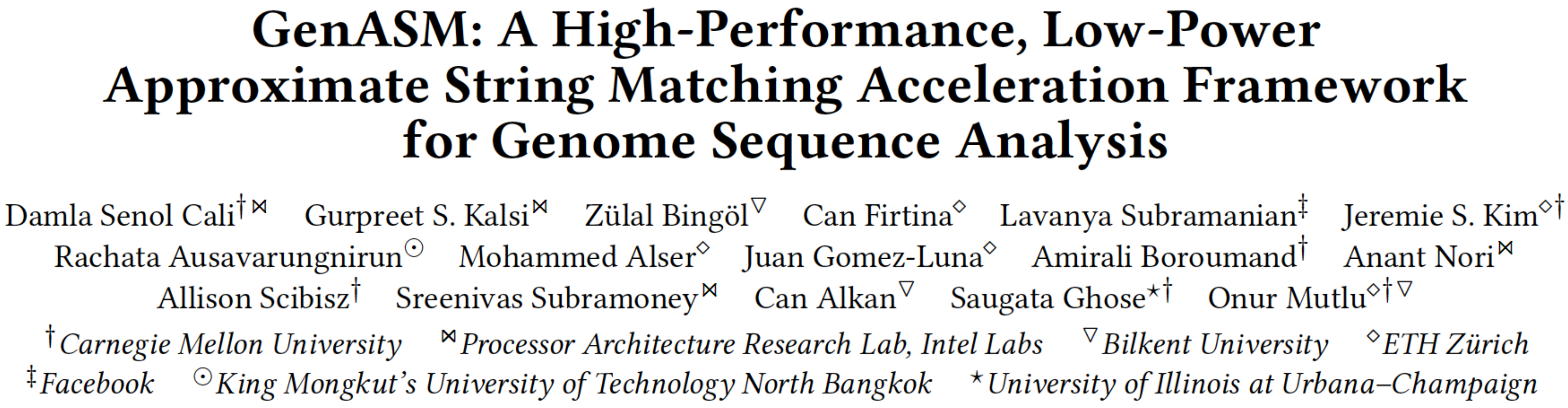
- ☐ **3.7×** speedup, **1.7×** less power than **Shouji** (state-of-the-art **HW**)

(3) Edit Distance Calculation

- ☐ **22–12501×** speedup, **548–582×** less power than **Edlib** (state-of-the-art **SW**)
- ☐ **9.3–400×** speedup, **67×** less power than **ASAP** (state-of-the-art **HW**)

More on GenASM Framework **[MICRO 2020]**

- Damla Senol Cali, Gurpreet S. Kalsi, Zulal Bingol, Can Firtina, Lavanya Subramanian, Jeremie S. Kim, Rachata Ausavarungnirun, Mohammed Alser, Juan Gomez-Luna, Amirali Boroumand, Anant Nori, Allison Scibisz, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu,
  "GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis"
  Proceedings of the 53rd International Symposium on Microarchitecture (***MICRO***), Virtual, October 2020.
  [Lighting Talk Video (1.5 minutes)]
  [Lightning Talk Slides (pptx) (pdf)]
  [Talk Video (18 minutes)]
  [Slides (pptx) (pdf)]



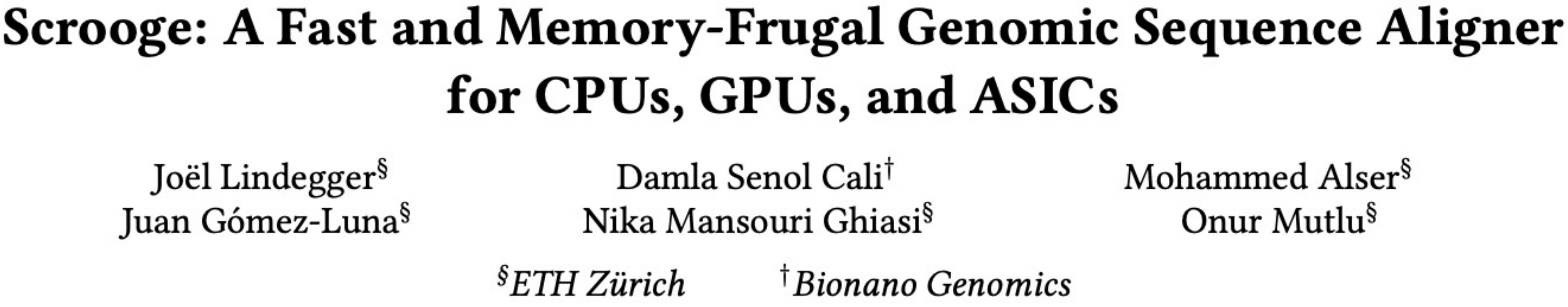
GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis

Damla Senol Cali[†⋈] Gurpreet S. Kalsi[⋈] Zülal Bingöl[▽] Can Firtina[◇] Lavanya Subramanian[‡] Jeremie S. Kim[◇†]
Rachata Ausavarungnirun[⊙] Mohammed Alser[◇] Juan Gomez-Luna[◇] Amirali Boroumand[†] Anant Nori[⋈]
Allison Scibisz[†] Sreenivas Subramoney[⋈] Can Alkan[▽] Saugata Ghose[⋆†] Onur Mutlu[◇†▽]

[†] *Carnegie Mellon University* [⋈] *Processor Architecture Research Lab, Intel Labs* [▽] *Bilkent University* [◇] *ETH Zürich*
[‡] *Facebook* [⊙] *King Mongkut's University of Technology North Bangkok* [⋆] *University of Illinois at Urbana–Champaign*

Scrooge: Overcoming GenASM Limitations

- Joël Lindegger, Damla Senol Cali, Mohammed Alser, Juan Gómez-Luna, Nika Mansouri Ghiasi, and Onur Mutlu,
  "Scrooge: A Fast and Memory-Frugal Genomic Sequence Aligner for CPUs, GPUs, and ASICs"
  Bioinformatics, [published online on] 24 March 2023.
  [Online link at Bioinformatics Journal]
  [arXiv preprint]
  [Scrooge Source Code]



Scrooge: A Fast and Memory-Frugal Genomic Sequence Aligner for CPUs, GPUs, and ASICs

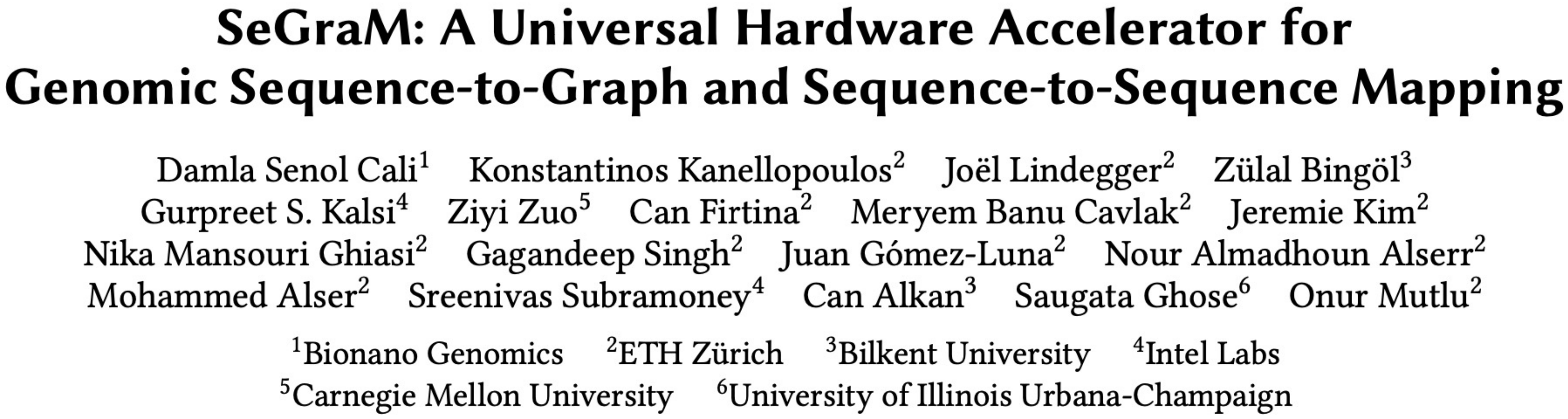
Joël Lindegger[§] Damla Senol Cali[†] Mohammed Alser[§]
Juan Gómez-Luna[§] Nika Mansouri Ghiasi[§] Onur Mutlu[§]

[§]*ETH Zürich* [†]*Bionano Genomics*

https://arxiv.org/pdf/2208.09985.pdf

Accelerating Sequence-to-Graph Mapping

- Damla Senol Cali, Konstantinos Kanellopoulos, Joel Lindegger, Zulal Bingol, Gurpreet S. Kalsi, Ziyi Zuo, Can Firtina, Meryem Banu Cavlak, Jeremie Kim, Nika MansouriGhiasi, Gagandeep Singh, Juan Gomez-Luna, Nour Almadhoun Alserr, Mohammed Alser, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu,
  "SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping"
  *Proceedings of the 49th International Symposium on Computer Architecture (**ISCA**)*, New York, June 2022.
  [arXiv version]



SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping

Damla Senol Cali[1] Konstantinos Kanellopoulos[2] Joël Lindegger[2] Zülal Bingöl[3]
Gurpreet S. Kalsi[4] Ziyi Zuo[5] Can Firtina[2] Meryem Banu Cavlak[2] Jeremie Kim[2]
Nika Mansouri Ghiasi[2] Gagandeep Singh[2] Juan Gómez-Luna[2] Nour Almadhoun Alserr[2]
Mohammed Alser[2] Sreenivas Subramoney[4] Can Alkan[3] Saugata Ghose[6] Onur Mutlu[2]

[1]Bionano Genomics [2]ETH Zürich [3]Bilkent University [4]Intel Labs
[5]Carnegie Mellon University [6]University of Illinois Urbana-Champaign

https://arxiv.org/pdf/2205.05883.pdf

Genome Sequence Analysis

- Mapping the reads to a reference genome (i.e., ***read mapping***) is a *critical step* in genome sequence analysis

Linear Reference: ACGTACGT

Read: ACGG

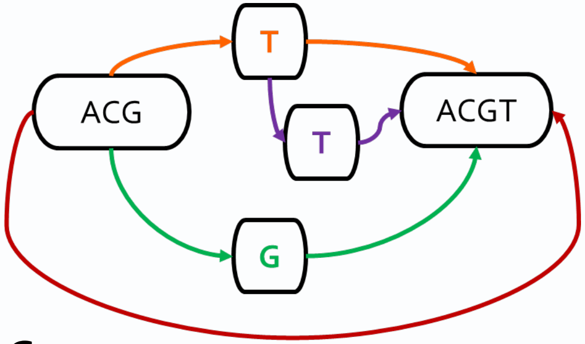
Alternative Sequence: ACGGACGT

Alternative Sequence: ACGTTACGT

Alternative Sequence: ACG–ACGT

Sequence-to-Sequence (S2S) Mapping

Graph-based Reference:



Read: ACGG

Sequence-to-Graph (S2G) Mapping

Sequence-to-graph mapping results in **notable quality improvements.**

However, it is a **more difficult** computational problem,

with **no prior hardware design.**

SeGraM: First Graph Mapping Accelerator

Our Goal:

Specialized, high-performance, scalable, and low-cost algorithm/hardware co-design that alleviates bottlenecks in **multiple steps** of sequence-to-graph mapping

SeGraM: *First universal algorithm/hardware co-designed genomic mapping accelerator* that can effectively and efficiently support:

☐ Sequence-to-graph mapping

☐ Sequence-to-sequence mapping

☐ Both short and long reads

Sequence-to-Graph Mapping Pipeline

Pre-Processing Steps (Offline)

Linear reference genome

Known genetic variations

0.1

Genome Graph Construction

(construct the graph using a linear reference genome and variations)

Genome graph

0.2

Indexing

(index the nodes of the graph)

Hash-table-based index (of graph nodes)

Reads from sequenced genome

1

Seeding

(query the index & find the seed matches)

Candidate mapping locations (subgraphs)

2

Filtering/Chaining/Clustering

(filter out dissimilar query read and subgraph pairs)

Remaining candidate mapping locations (subgraphs)

3

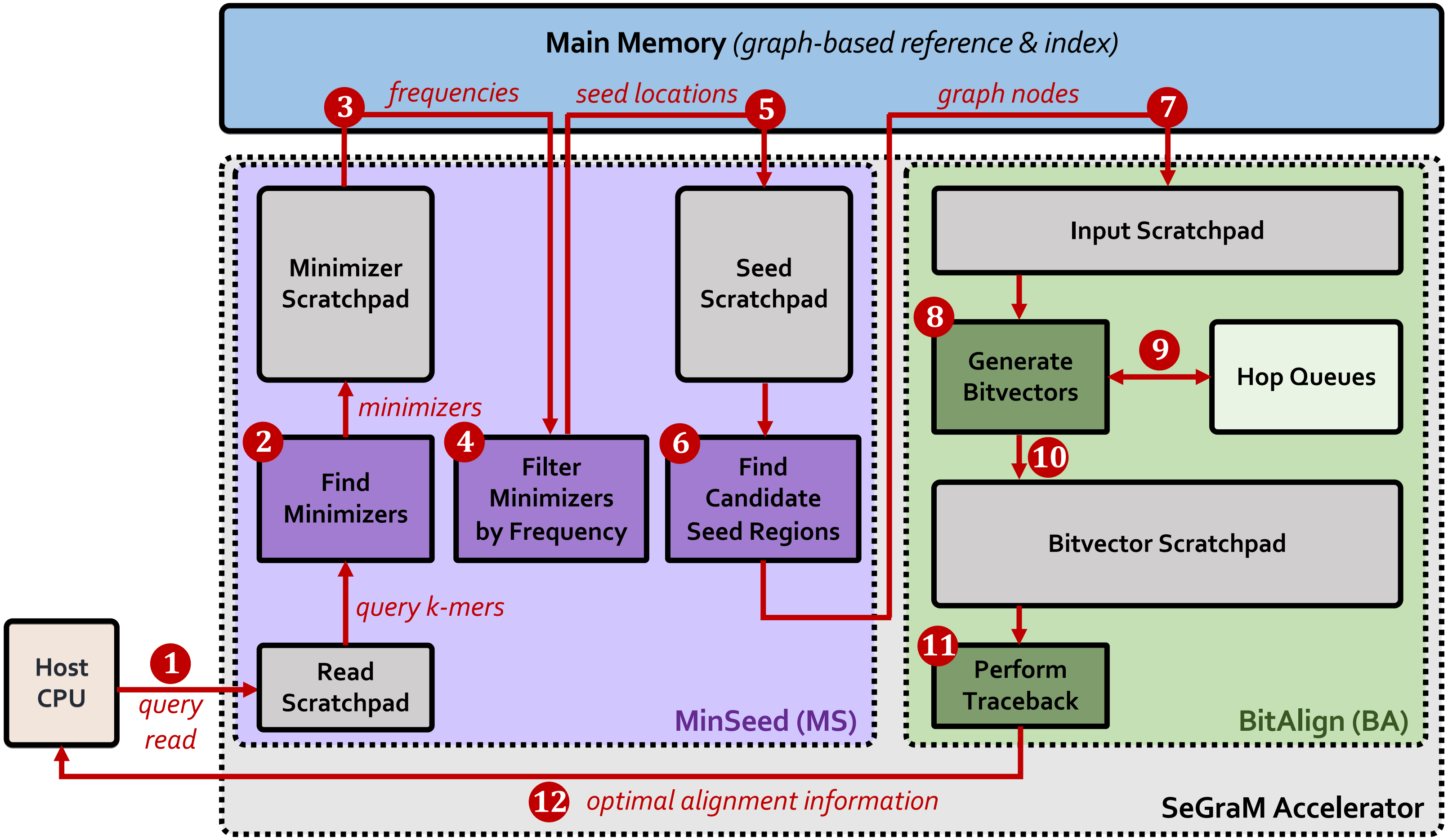
S2G Alignment

(perform distance/score calculation & traceback)

Optimal alignment between read & subgraph

Seed-and-Extend Steps (Online)

SeGraM Hardware Design



MinSeed: *first* hardware accelerator for **Min**imizer-based **Seed**ing

BitAlign: *first* hardware accelerator for (**Bit**vector-based) sequence-to-graph **Align**ment

Use Cases & Key Results

(1) Sequence-to-Graph (S2G) Mapping

- ☐ **5.9×/106×** speedup, **4.1×/3.0×** less power than **GraphAligner** for long and short reads, respectively (state-of-the-art **SW**)
- ☐ **3.9×/742×** speedup, **4.4×/3.2×** less power than **vg** for long and short reads, respectively (state-of-the-art **SW**)

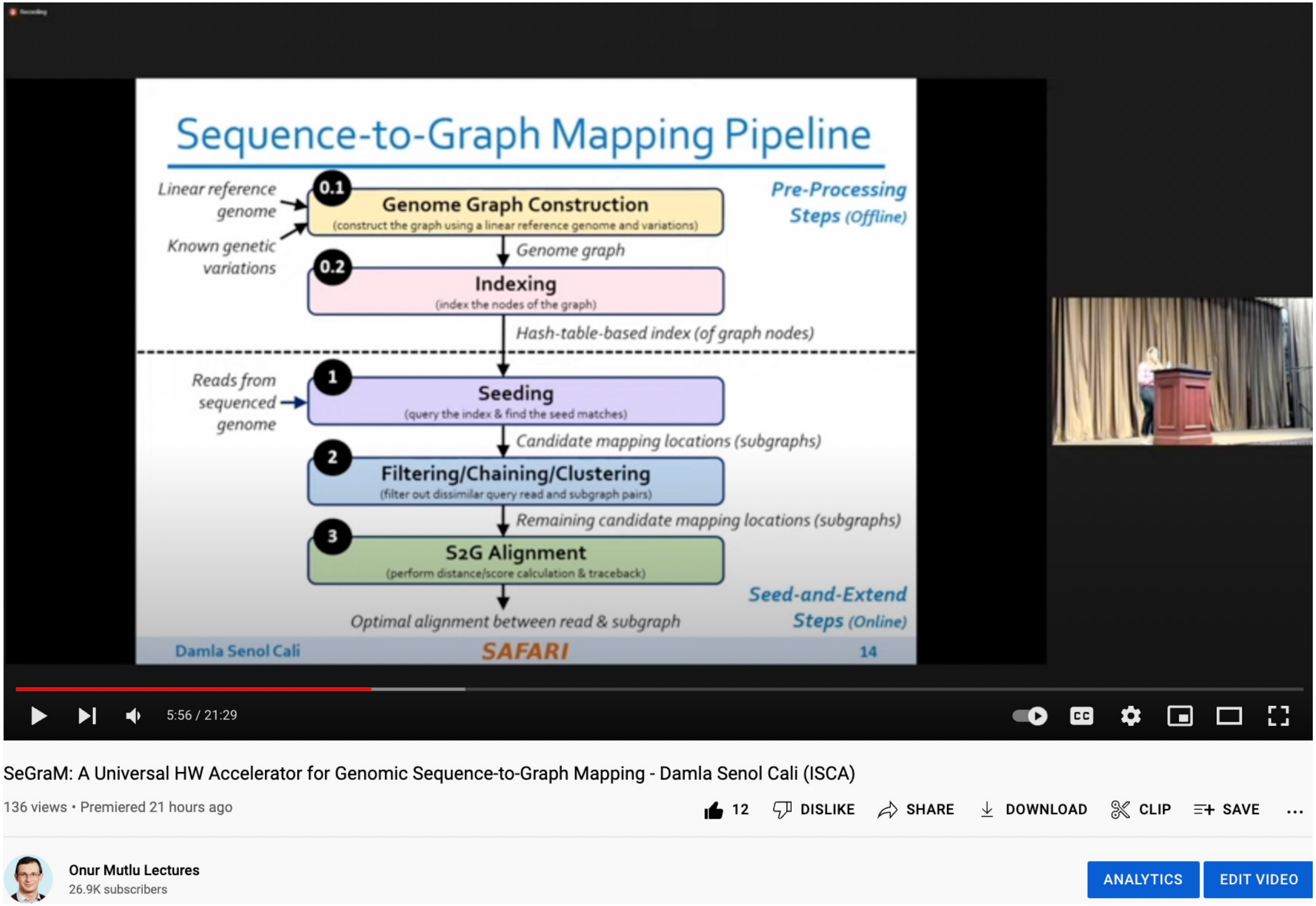
(2) Sequence-to-Graph (S2G) Alignment

- ☐ **41×–539×** speedup over **PaSGAL** with AVX-512 support (state-of-the-art **SW**)

(3) Sequence-to-Sequence (S2S) Alignment

- ☐ **1.2×/4.8×** higher throughput than **GenASM** and **GACT of Darwin** for long reads (state-of-the-art **HW**)
- ☐ **1.3×/2.4×** higher throughput than **GenASM** and **SillaX of GenAX** for short reads (state-of-the-art **HW**)

SeGraM Talk Video



Sequence-to-Graph Mapping Pipeline

Linear reference genome → **0.1** **Genome Graph Construction** (construct the graph using a linear reference genome and variations)

Known genetic variations → **0.1**

Pre-Processing Steps (Offline)

↓ *Genome graph*

0.2 **Indexing** (index the nodes of the graph)

↓ *Hash-table-based index (of graph nodes)*

Reads from sequenced genome → **1** **Seeding** (query the index & find the seed matches)

↓ *Candidate mapping locations (subgraphs)*

2 **Filtering/Chaining/Clustering** (filter out dissimilar query read and subgraph pairs)

↓ *Remaining candidate mapping locations (subgraphs)*

3 **S2G Alignment** (perform distance/score calculation & traceback)

↓ *Optimal alignment between read & subgraph*

Seed-and-Extend Steps (Online)

Damla Senol Cali SAFARI 14

SeGraM: A Universal HW Accelerator for Genomic Sequence-to-Graph Mapping - Damla Senol Cali (ISCA)

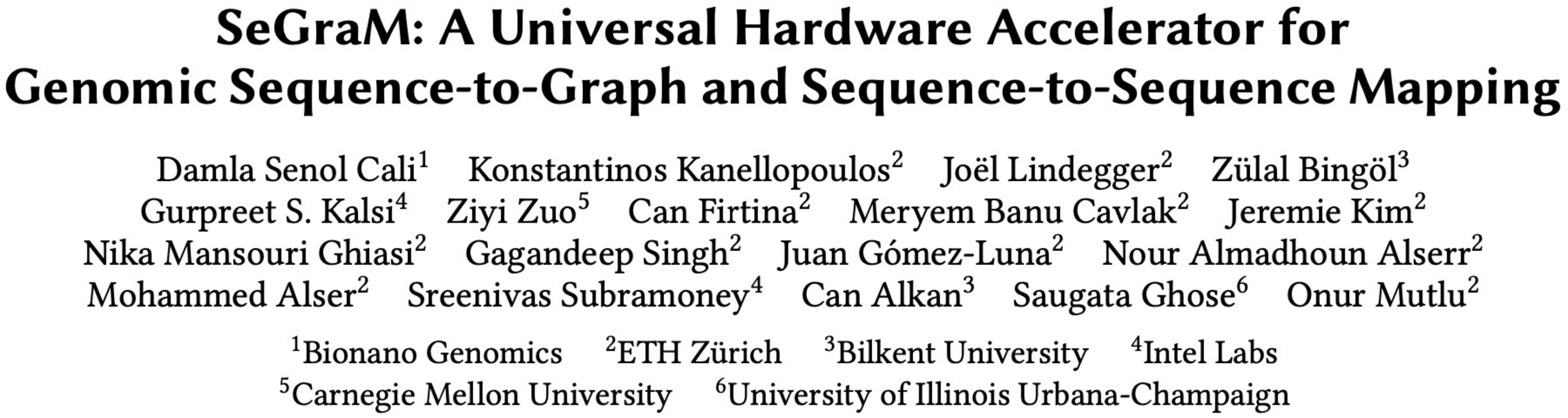
136 views • Premiered 21 hours ago

Onur Mutlu Lectures
26.9K subscribers

https://www.youtube.com/watch?v=gyjqYoyDP9s

Accelerating Sequence-to-Graph Mapping

- Damla Senol Cali, Konstantinos Kanellopoulos, Joel Lindegger, Zulal Bingol, Gurpreet S. Kalsi, Ziyi Zuo, Can Firtina, Meryem Banu Cavlak, Jeremie Kim, Nika MansouriGhiasi, Gagandeep Singh, Juan Gomez-Luna, Nour Almadhoun Alserr, Mohammed Alser, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu,
  "SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping"
  Proceedings of the 49th International Symposium on Computer Architecture (***ISCA***), New York, June 2022.
  [arXiv version]



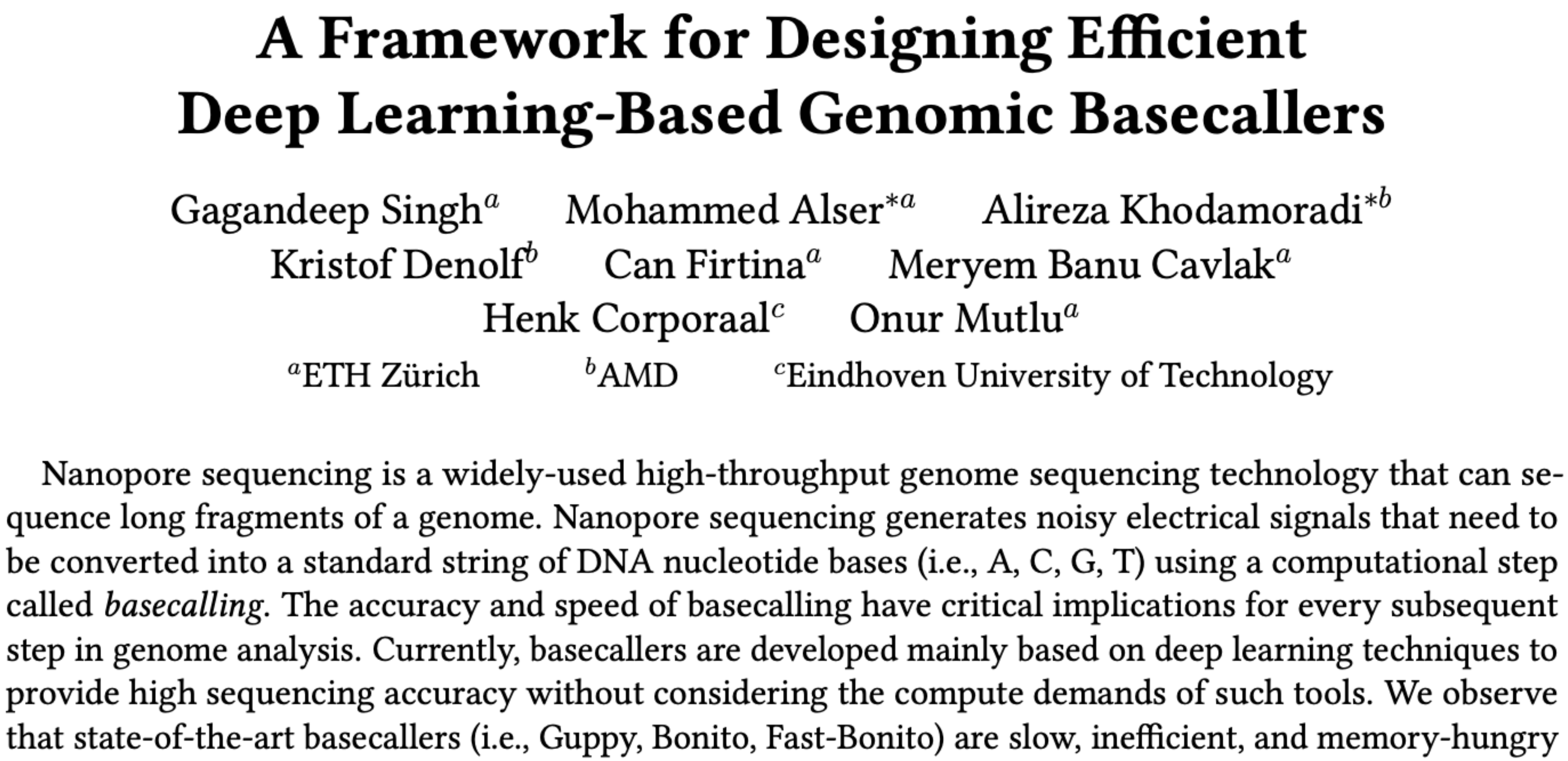
SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping

Damla Senol Cali[1] Konstantinos Kanellopoulos[2] Joël Lindegger[2] Zülal Bingöl[3]
Gurpreet S. Kalsi[4] Ziyi Zuo[5] Can Firtina[2] Meryem Banu Cavlak[2] Jeremie Kim[2]
Nika Mansouri Ghiasi[2] Gagandeep Singh[2] Juan Gómez-Luna[2] Nour Almadhoun Alserr[2]
Mohammed Alser[2] Sreenivas Subramoney[4] Can Alkan[3] Saugata Ghose[6] Onur Mutlu[2]

[1]Bionano Genomics [2]ETH Zürich [3]Bilkent University [4]Intel Labs
[5]Carnegie Mellon University [6]University of Illinois Urbana-Champaign

https://arxiv.org/pdf/2205.05883.pdf

Designing & Accelerating Basecallers



A Framework for Designing Efficient Deep Learning-Based Genomic Basecallers

Gagandeep Singh[a] Mohammed Alser*[a] Alireza Khodamoradi*[b]
Kristof Denolf[b] Can Firtina[a] Meryem Banu Cavlak[a]
Henk Corporaal[c] Onur Mutlu[a]
[a]ETH Zürich [b]AMD [c]Eindhoven University of Technology

Nanopore sequencing is a widely-used high-throughput genome sequencing technology that can sequence long fragments of a genome. Nanopore sequencing generates noisy electrical signals that need to be converted into a standard string of DNA nucleotide bases (i.e., A, C, G, T) using a computational step called *basecalling*. The accuracy and speed of basecalling have critical implications for every subsequent step in genome analysis. Currently, basecallers are developed mainly based on deep learning techniques to provide high sequencing accuracy without considering the compute demands of such tools. We observe that state-of-the-art basecallers (i.e., Guppy, Bonito, Fast-Bonito) are slow, inefficient, and memory-hungry

https://arxiv.org/pdf/2211.03079.pdf

Agenda

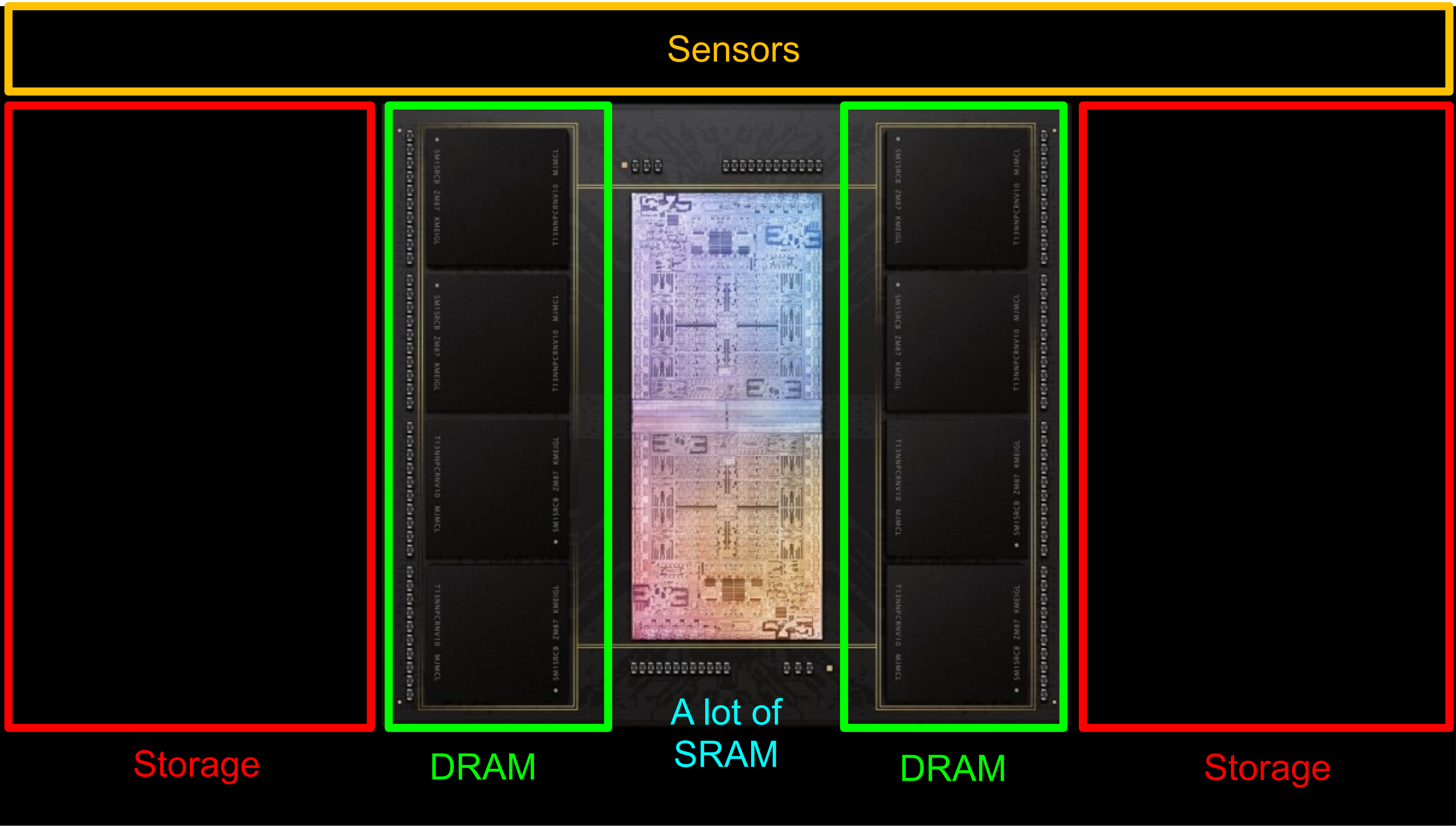
- The Problem: DNA Read Mapping
    - State-of-the-art Read Mapper Design

- Algorithmic Acceleration
    - Exploiting Structure of the Genome
    - Exploiting SIMD Instructions

- Hardware Acceleration
    - Specialized Architectures
    - Processing in Memory & Storage

- Future Opportunities: New Technologies & Applications

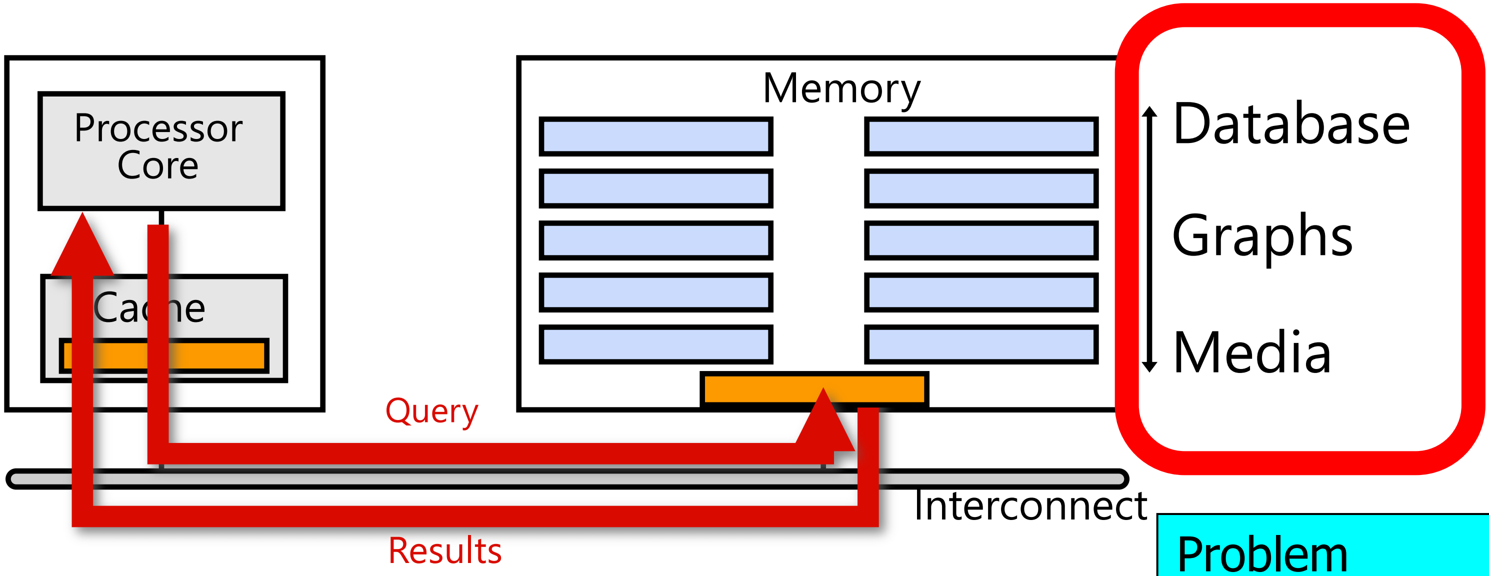
Process Data Where It Makes Sense



Apple M1 Ultra System (2022)

https://www.gsmarena.com/apple_announces_m1_ultra_with_20core_cpu_and_64core_gpu-news-53481.php

Goal: Processing Inside Memory



- Many questions … How do we design the:
  - compute-capable memory & controllers?
  - processors & communication units?
  - software & hardware interfaces?
  - system software, compilers, languages?
  - algorithms & theoretical foundations?

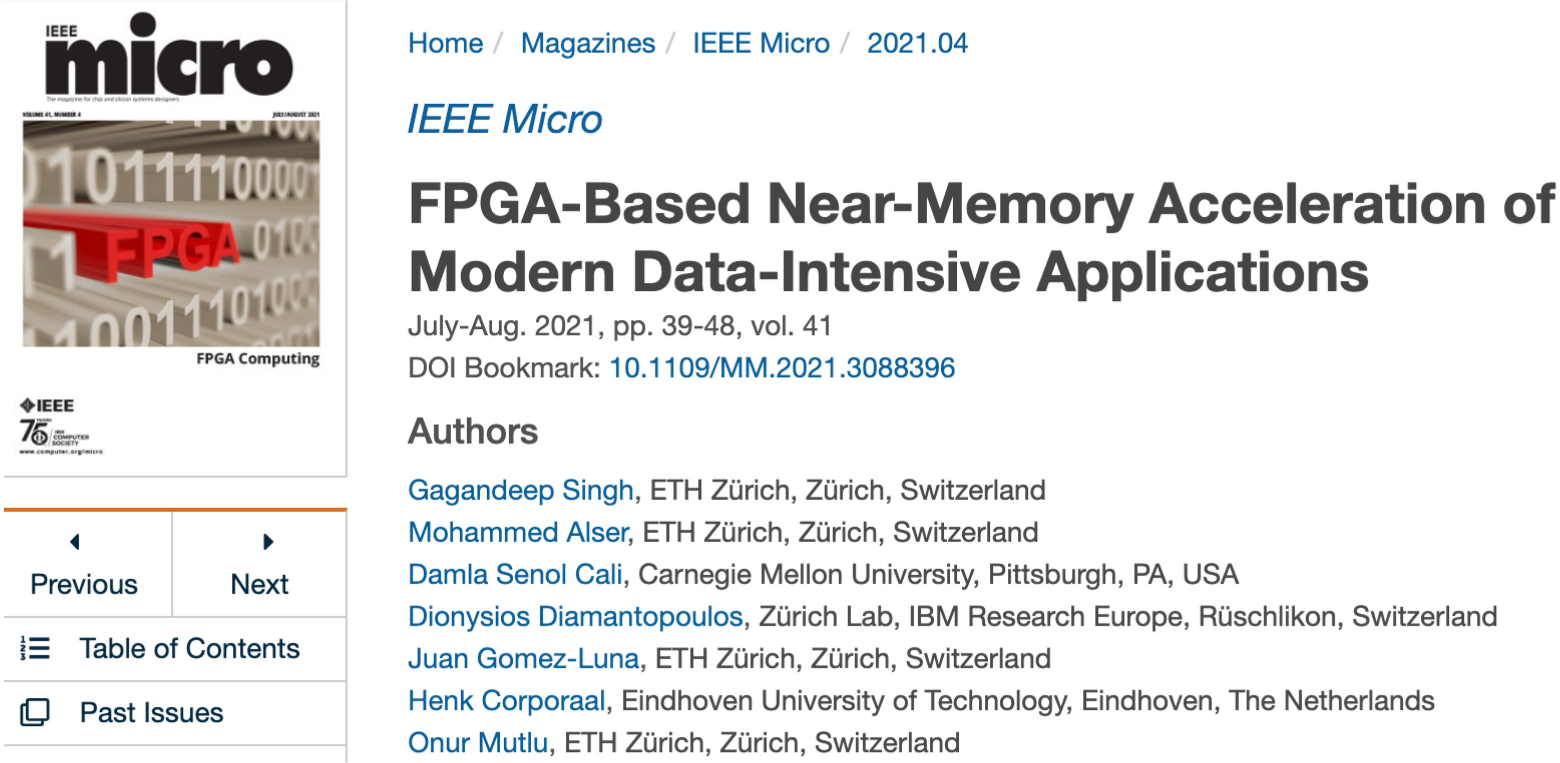
Problem
Algorithm
Program/Language
System Software
SW/HW Interface
Micro-architecture
Logic
Devices
Electrons

Read Mapping & Filtering in Memory

We need to design
mapping & filtering algorithms
that fit processing-in-memory

Near-Memory Pre-Alignment Filtering

Gagandeep Singh, Mohammed Alser, Damla Senol Cali, Dionysios Diamantopoulos, Juan Gomez-Luna, Henk Corporaal, Onur Mutlu,
"FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications"
IEEE Micro, 2021.
[Source Code]



IEEE Micro

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications

July-Aug. 2021, pp. 39-48, vol. 41

DOI Bookmark: 10.1109/MM.2021.3088396

Authors

Gagandeep Singh, ETH Zürich, Zürich, Switzerland
Mohammed Alser, ETH Zürich, Zürich, Switzerland
Damla Senol Cali, Carnegie Mellon University, Pittsburgh, PA, USA
Dionysios Diamantopoulos, Zürich Lab, IBM Research Europe, Rüschlikon, Switzerland
Juan Gomez-Luna, ETH Zürich, Zürich, Switzerland
Henk Corporaal, Eindhoven University of Technology, Eindhoven, The Netherlands
Onur Mutlu, ETH Zürich, Zürich, Switzerland

Near-Memory SneakySnake

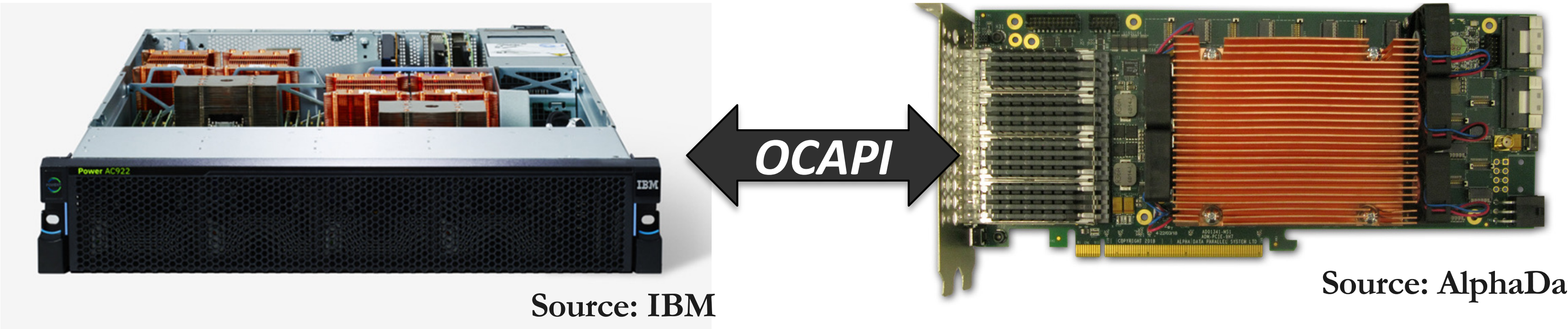
- Problem: Read mapping is heavily bottlenecked by data movement from main memory

- Solution: Perform read mapping near where data resides using specialized logic

- We carefully redesign the accelerator logic of SneakySnake to exploit near-memory computation capability on real FPGA boards that use HBM (high-bandwidth memory)

- **Near-memory SneakySnake** improves **performance** and **energy efficiency** by 27.4× and 133×, respectively, over a 16-core (64-thread) IBM POWER9 CPU

Near-Memory Acceleration using FPGAs



Source: IBM

Source: AlphaData

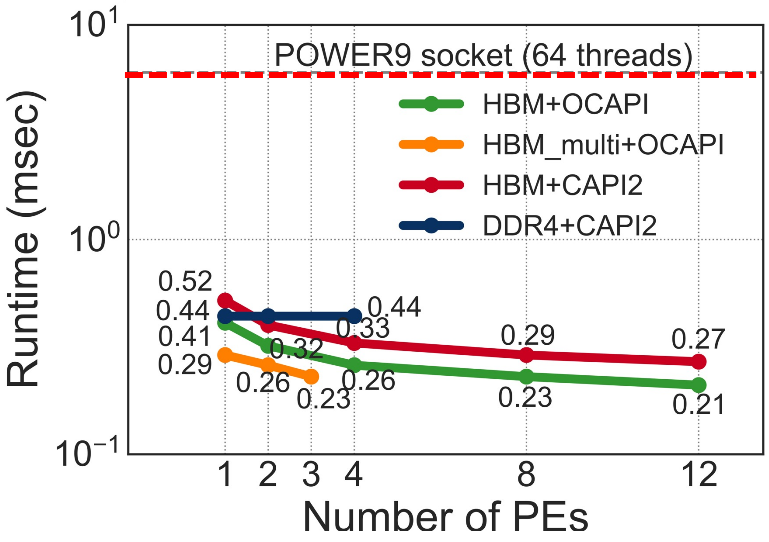
IBM POWER9 CPU

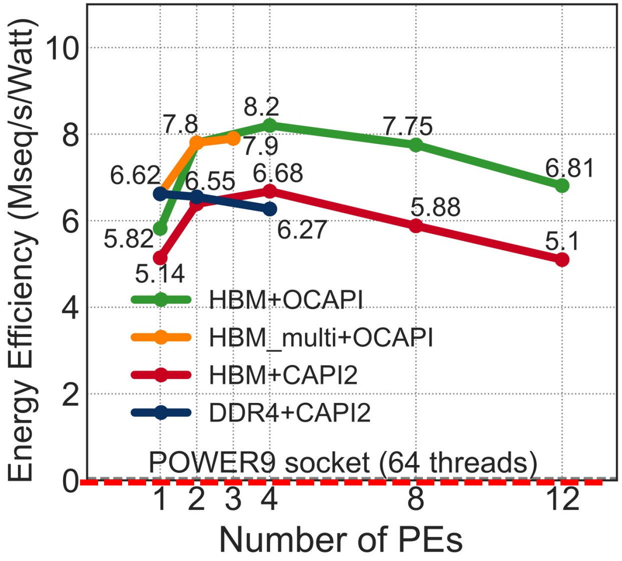
HBM-based FPGA board

Near-HBM FPGA-based accelerator

Two communication technologies: **CAPI2** and **OCAPI**
Two memory technologies: **DDR4** and **HBM**
Two workloads: **Weather Modeling** and **Genome Analysis**

Performance & Energy Greatly Improve





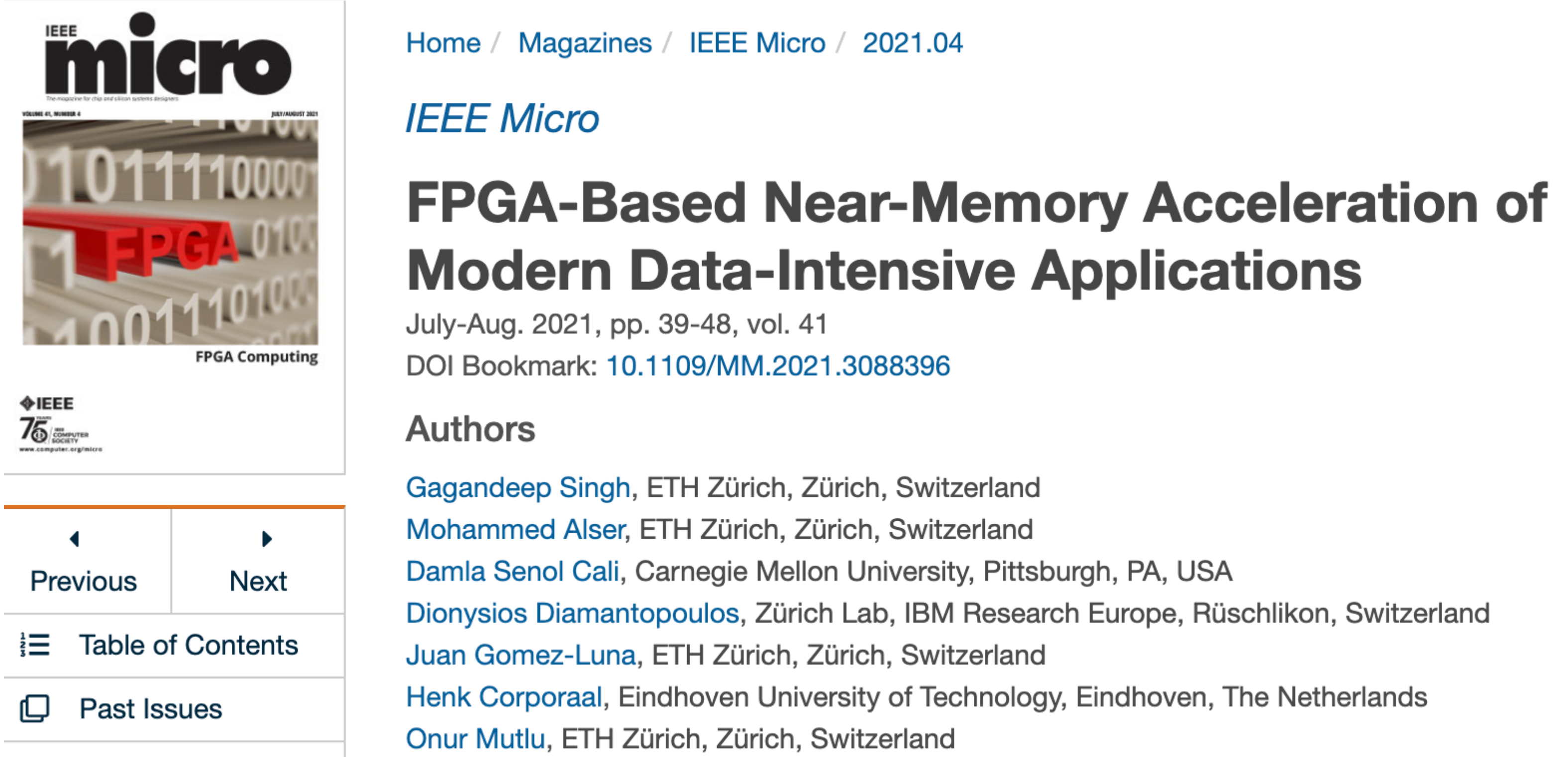
5-27× performance vs. a 16-core (64-thread) IBM POWER9 CPU

12-133× energy efficiency vs. a 16-core (64-thread) IBM POWER9 CPU

HBM alleviates memory bandwidth contention vs. DDR4

More On Near-Memory SneakySnake

Gagandeep Singh, Mohammed Alser, Damla Senol Cali, Dionysios Diamantopoulos, Juan Gomez-Luna, Henk Corporaal, Onur Mutlu,
"FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications"
IEEE Micro, 2021.
[Source Code]



IEEE Micro

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications

July-Aug. 2021, pp. 39-48, vol. 41

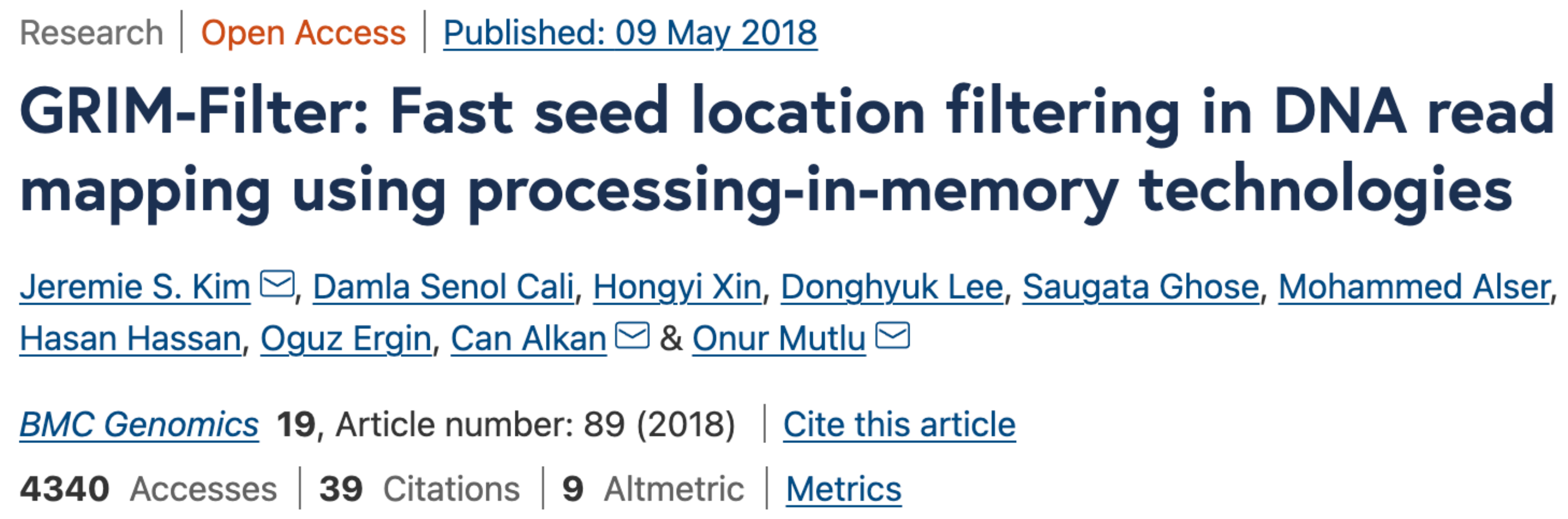
DOI Bookmark: 10.1109/MM.2021.3088396

Authors

Gagandeep Singh, ETH Zürich, Zürich, Switzerland
Mohammed Alser, ETH Zürich, Zürich, Switzerland
Damla Senol Cali, Carnegie Mellon University, Pittsburgh, PA, USA
Dionysios Diamantopoulos, Zürich Lab, IBM Research Europe, Rüschlikon, Switzerland
Juan Gomez-Luna, ETH Zürich, Zürich, Switzerland
Henk Corporaal, Eindhoven University of Technology, Eindhoven, The Netherlands
Onur Mutlu, ETH Zürich, Zürich, Switzerland

Location Filtering in 3D-Stacked PIM

- Jeremie S. Kim, Damla Senol Cali, Hongyi Xin, Donghyuk Lee, Saugata Ghose, Mohammed Alser, Hasan Hassan, Oguz Ergin, Can Alkan, and Onur Mutlu,
  "GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies"
  BMC Genomics, 2018.
  *Proceedings of the 16th Asia Pacific Bioinformatics Conference (**APBC**)*, Yokohama, Japan, January 2018.
  [Slides (pptx) (pdf)]
  [Source Code]
  [arxiv.org Version (pdf)]
  [Talk Video at AACBB 2019]



Research | Open Access | Published: 09 May 2018

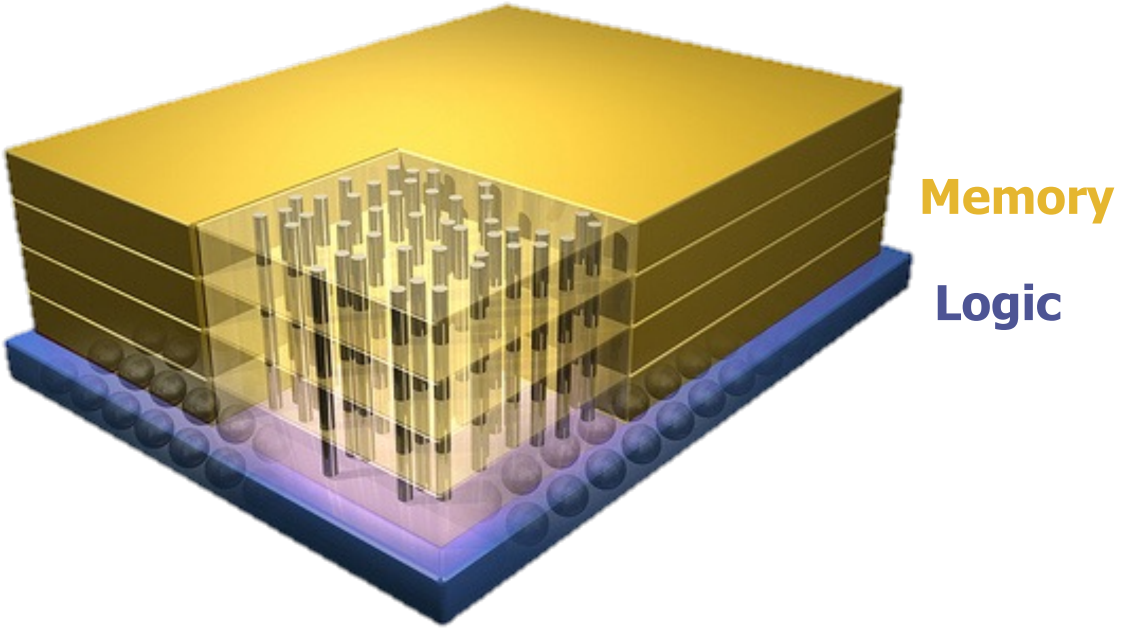
GRIM-Filter: Fast seed location filtering in DNA read mapping using processing-in-memory technologies

Jeremie S. Kim, Damla Senol Cali, Hongyi Xin, Donghyuk Lee, Saugata Ghose, Mohammed Alser, Hasan Hassan, Oguz Ergin, Can Alkan & Onur Mutlu

BMC Genomics **19**, Article number: 89 (2018) | Cite this article

4340 Accesses | **39** Citations | **9** Altmetric | Metrics

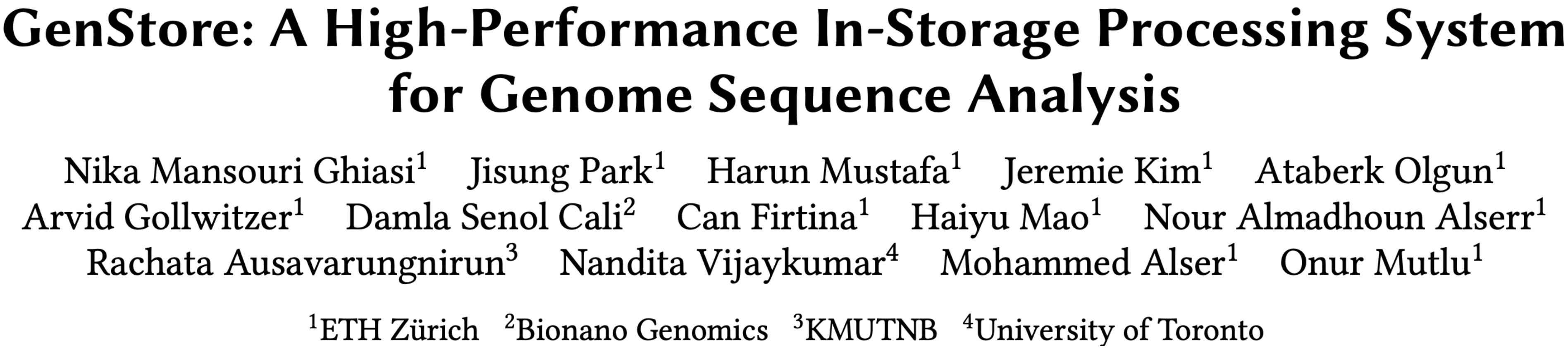
Opportunity: 3D-Stacked Logic+Memory



Other "True 3D" technologies under development

In-Storage Genome Filtering **[ASPLOS 2022]**

- Nika Mansouri Ghiasi, Jisung Park, Harun Mustafa, Jeremie Kim, Ataberk Olgun, Arvid Gollwitzer, Damla Senol Cali, Can Firtina, Haiyu Mao, Nour Almadhoun Alserr, Rachata Ausavarungnirun, Nandita Vijaykumar, Mohammed Alser, and Onur Mutlu,
  "GenStore: A High-Performance and Energy-Efficient In-Storage Computing System for Genome Sequence Analysis"
  *Proceedings of the 27th International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*, Virtual, February-March 2022.
  [Lightning Talk Slides (pptx) (pdf)]
  [Lightning Talk Video (90 seconds)]

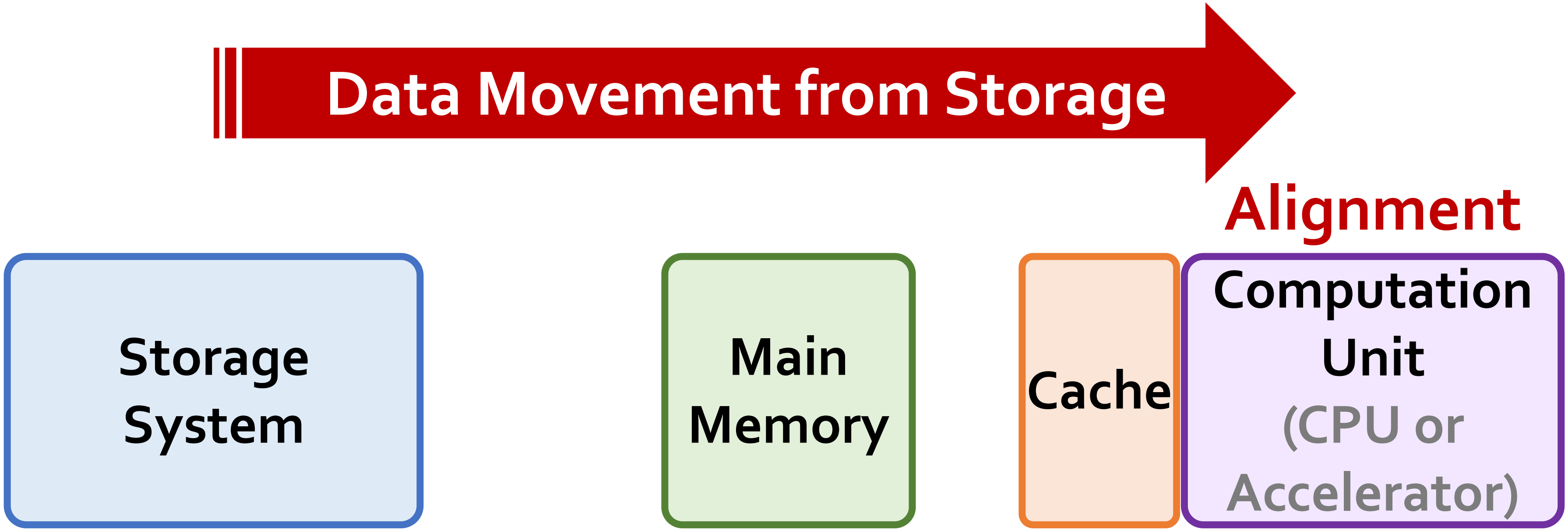


GenStore: A High-Performance In-Storage Processing System for Genome Sequence Analysis

Nika Mansouri Ghiasi[1] Jisung Park[1] Harun Mustafa[1] Jeremie Kim[1] Ataberk Olgun[1]
Arvid Gollwitzer[1] Damla Senol Cali[2] Can Firtina[1] Haiyu Mao[1] Nour Almadhoun Alserr[1]
Rachata Ausavarungnirun[3] Nandita Vijaykumar[4] Mohammed Alser[1] Onur Mutlu[1]

[1]ETH Zürich [2]Bionano Genomics [3]KMUTNB [4]University of Toronto

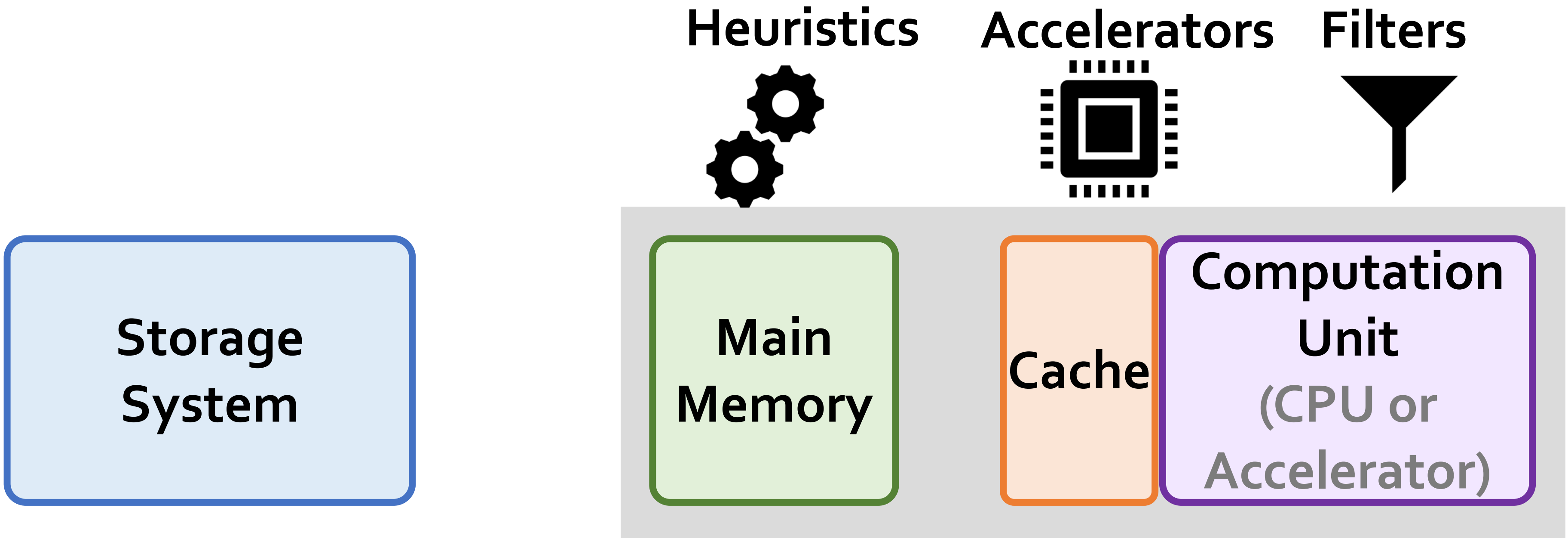
Genome Sequence Analysis



✗ Computation overhead

✗ Data movement overhead

Accelerating Genome Sequence Analysis



Heuristics

Accelerators

Filters

Storage System

Main Memory

Cache

Computation Unit

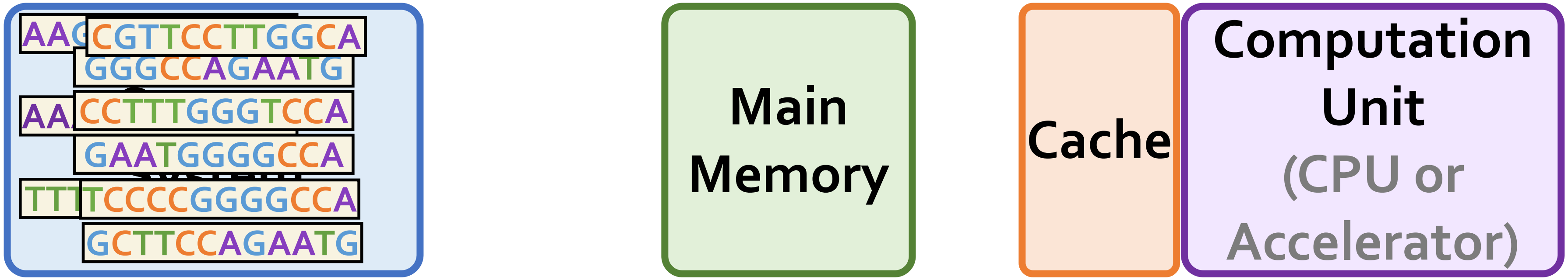
(CPU or Accelerator)

✓ Computation overhead

✗ Data movement overhead

Key Idea

Filter reads that do *not* require alignment *inside the storage system*



Filtered Reads

Exactly-matching reads

Do not need expensive approximate string matching during alignment

Non-matching reads

Do not have potential matching locations and can skip alignment

Filtering Opportunities

- Sequencing machines produce one of two kinds of reads
  - **Short reads:** highly accurate and short
  - **Long reads:** less accurate and long

Reads that do not require the expensive alignment step:

Exactly-matching reads

Do not need expensive approximate string matching during alignment

- Low sequencing error rates (short reads) combined with
- Low genetic variation

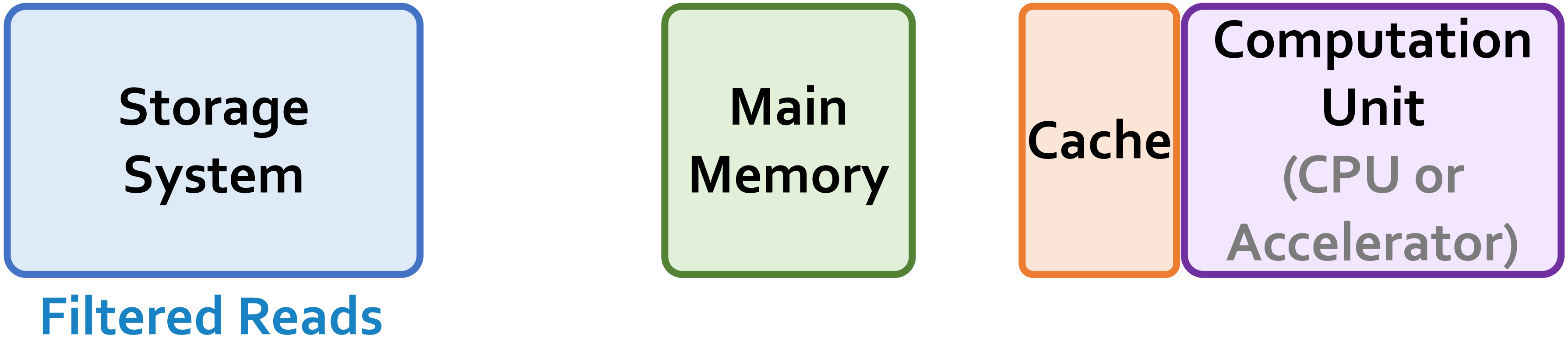
Non-matching reads

Do not have potential matching locations, so they skip alignment

- High sequencing error rates (long reads) or
- High genetic variation (short or long reads)

Challenges

Filter reads that do *not* require alignment *inside the storage system*

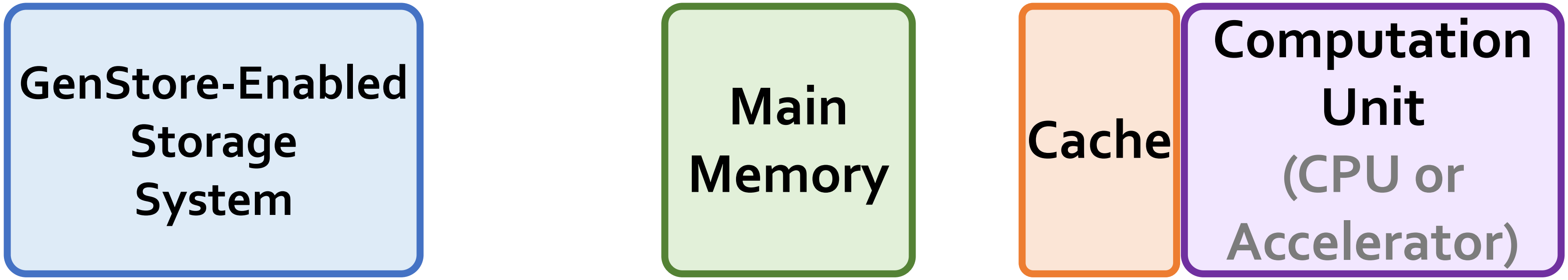


Read mapping workloads can exhibit different behavior

There are limited hardware resources in the storage system

GenStore

Filter reads that do *not* require alignment *inside the storage system*



GenStore-Enabled Storage System

Main Memory

Cache

Computation Unit (CPU or Accelerator)

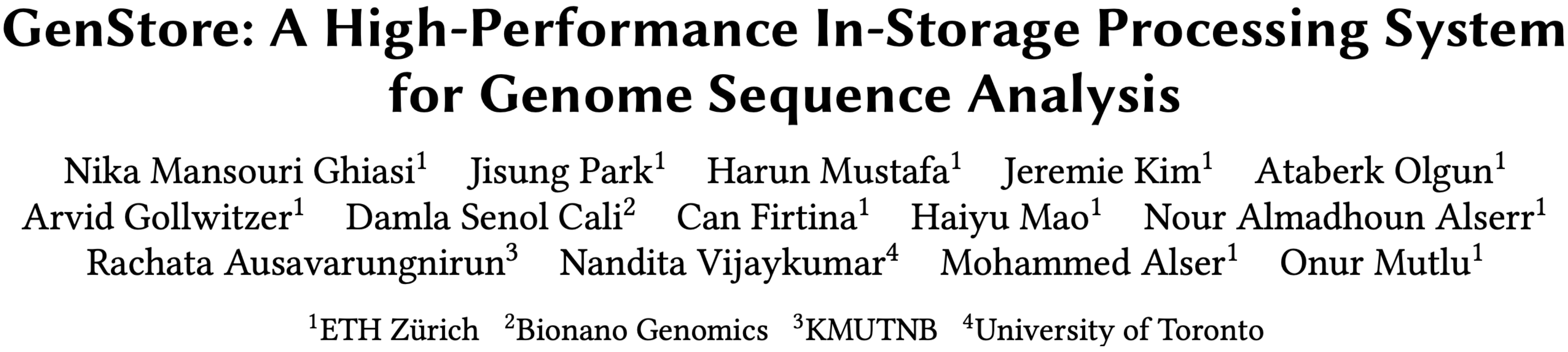
✓ Computation overhead

✓ Data movement overhead

GenStore provides significant speedup (1.4x - 33.6x) and energy reduction (3.9x – 29.2x) at low cost

In-Storage Genome Filtering **[ASPLOS 2022]**

- Nika Mansouri Ghiasi, Jisung Park, Harun Mustafa, Jeremie Kim, Ataberk Olgun, Arvid Gollwitzer, Damla Senol Cali, Can Firtina, Haiyu Mao, Nour Almadhoun Alserr, Rachata Ausavarungnirun, Nandita Vijaykumar, Mohammed Alser, and Onur Mutlu,
  "GenStore: A High-Performance and Energy-Efficient In-Storage Computing System for Genome Sequence Analysis"
  *Proceedings of the 27th International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*, Virtual, February-March 2022.
  [Lightning Talk Slides (pptx) (pdf)]
  [Lightning Talk Video (90 seconds)]

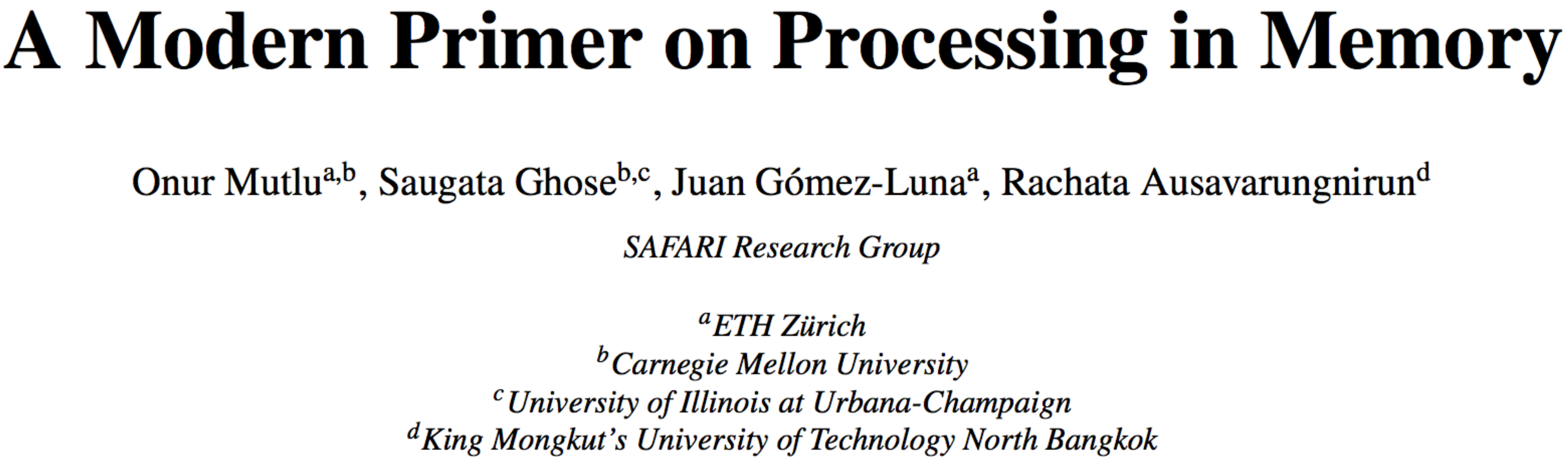


GenStore: A High-Performance In-Storage Processing System for Genome Sequence Analysis

Nika Mansouri Ghiasi[1] Jisung Park[1] Harun Mustafa[1] Jeremie Kim[1] Ataberk Olgun[1]
Arvid Gollwitzer[1] Damla Senol Cali[2] Can Firtina[1] Haiyu Mao[1] Nour Almadhoun Alserr[1]
Rachata Ausavarungnirun[3] Nandita Vijaykumar[4] Mohammed Alser[1] Onur Mutlu[1]

[1]ETH Zürich [2]Bionano Genomics [3]KMUTNB [4]University of Toronto

PIM Review and Open Problems



A Modern Primer on Processing in Memory

Onur Mutlu[a,b], Saugata Ghose[b,c], Juan Gómez-Luna[a], Rachata Ausavarungnirun[d]

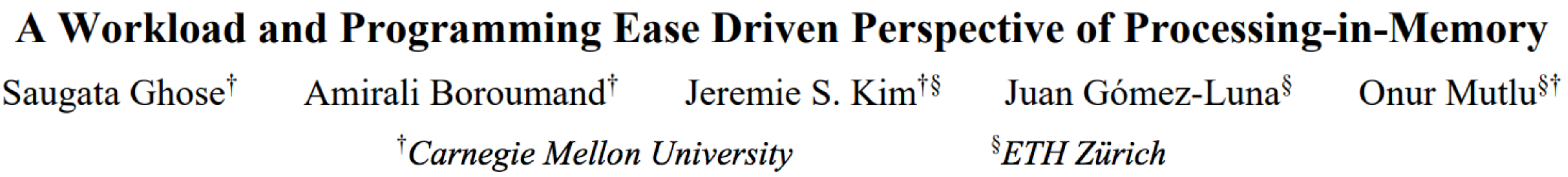
SAFARI Research Group

[a] *ETH Zürich*
[b] *Carnegie Mellon University*
[c] *University of Illinois at Urbana-Champaign*
[d] *King Mongkut's University of Technology North Bangkok*

Onur Mutlu, Saugata Ghose, Juan Gomez-Luna, and Rachata Ausavarungnirun,
"A Modern Primer on Processing in Memory"
Invited Book Chapter in ***Emerging Computing: From Devices to Systems - Looking Beyond Moore and Von Neumann***, Springer, to be published in 2021.

https://arxiv.org/pdf/1903.03988.pdf

PIM Review and Open Problems (II)



A Workload and Programming Ease Driven Perspective of Processing-in-Memory

Saugata Ghose[†] Amirali Boroumand[†] Jeremie S. Kim[†§] Juan Gómez-Luna[§] Onur Mutlu[§†]

[†]*Carnegie Mellon University* [§]*ETH Zürich*

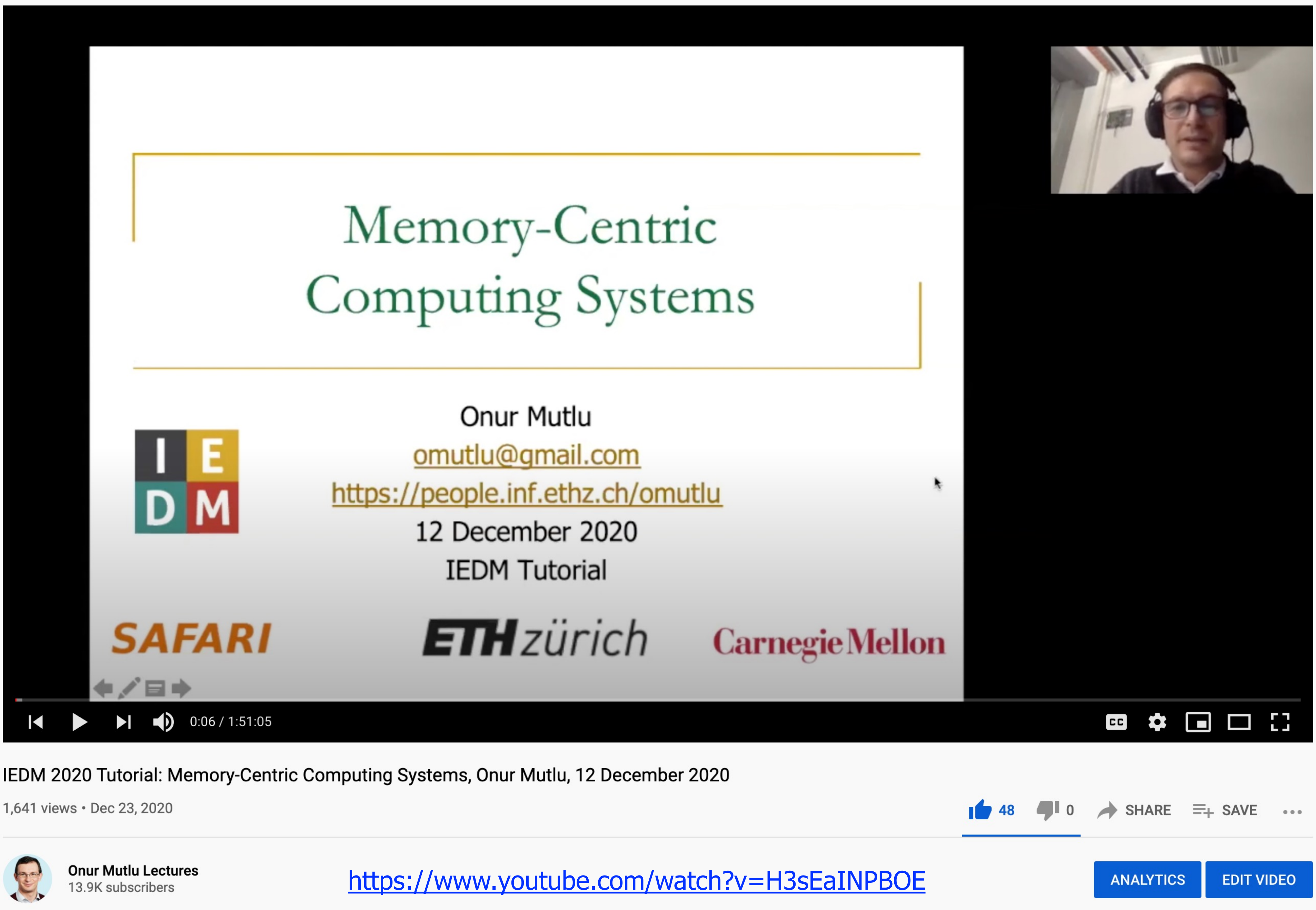
Saugata Ghose, Amirali Boroumand, Jeremie S. Kim, Juan Gomez-Luna, and Onur Mutlu,
"Processing-in-Memory: A Workload-Driven Perspective"
Invited Article in IBM Journal of Research & Development, Special Issue on Hardware for Artificial Intelligence, to appear in November 2019.
[Preliminary arXiv version]

https://arxiv.org/pdf/1907.12947.pdf

More on Processing-in-Memory

- Onur Mutlu,
  "Memory-Centric Computing Systems"
  Invited Tutorial at *66th International Electron Devices Meeting (**IEDM**)*, Virtual, 12 December 2020.
  [Slides (pptx) (pdf)]
  [Executive Summary Slides (pptx) (pdf)]
  [Tutorial Video (1 hour 51 minutes)]
  [Executive Summary Video (2 minutes)]
  [Abstract and Bio]
  [Related Keynote Paper from VLSI-DAT 2020]
  [Related Review Paper on Processing in Memory]

https://www.youtube.com/watch?v=H3sEaINPBOE



Memory-Centric Computing Systems

Onur Mutlu
omutlu@gmail.com
https://people.inf.ethz.ch/omutlu
12 December 2020
IEDM Tutorial

IEDM 2020 Tutorial: Memory-Centric Computing Systems, Onur Mutlu, 12 December 2020

1,641 views • Dec 23, 2020

https://www.youtube.com/watch?v=H3sEaINPBOE

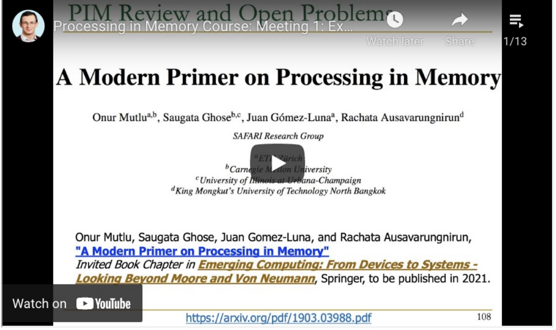
https://www.youtube.com/onurmutlulectures

PIM Course (Fall 2022)

- **[Fall 2022 Edition:](https://safari.ethz.ch/projects_and_seminars/fall2022/doku.php?id=processing_in_memory)**
  - https://safari.ethz.ch/projects_and_seminars/fall2022/doku.php?id=processing_in_memory
- **Spring 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=processing_in_memory

- **Youtube Livestream (Fall 2022):**
  - https://www.youtube.com/watch?v=QLL0wQ9I4Dw&list=PL5Q2soXY2Zi8KzG2CQYRNQOVD0GOBrnKy
- **Youtube Livestream (Spring 2022):**
  - https://www.youtube.com/watch?v=9e4Chnwdovo&list=PL5Q2soXY2Zi-841fUYYUK9EsXKhQKRPyX

- Project course
  - Taken by Bachelor's/Master's students
  - Processing-in-Memory lectures
  - Hands-on research exploration
  - Many research readings



A Modern Primer on Processing in Memory

Onur Mutlu[a,b], Saugata Ghose[b,c], Juan Gómez-Luna[a], Rachata Ausavarungnirun[d]

SAFARI Research Group

[a]ETH Zürich
[b]Carnegie Mellon University
[c]University of Illinois at Urbana-Champaign
[d]King Mongkut's University of Technology North Bangkok

Onur Mutlu, Saugata Ghose, Juan Gomez-Luna, and Rachata Ausavarungnirun,
"A Modern Primer on Processing in Memory"
Invited Book Chapter in ***Emerging Computing: From Devices to Systems - Looking Beyond Moore and Von Neumann***, Springer, to be published in 2021.

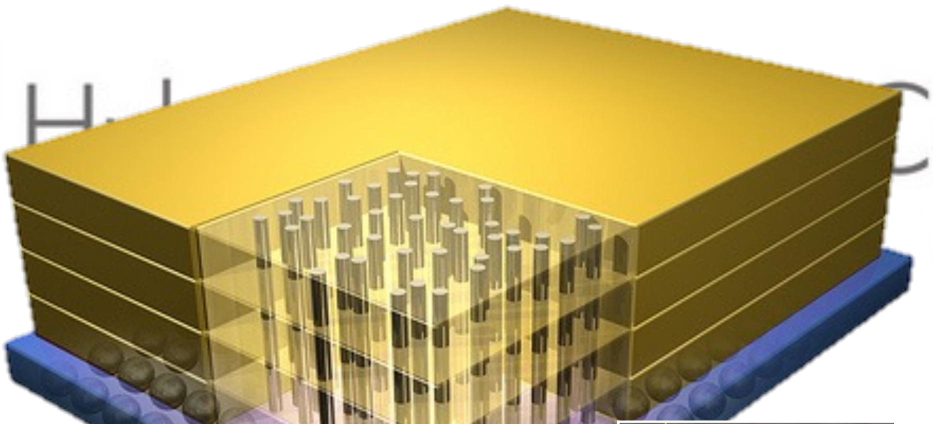
https://arxiv.org/pdf/1903.03988.pdf 108

Spring 2022 Meetings/Schedule

Week	Date	Livestream	Meeting	Learning Materials	Assignments
W1	10.03 Thu.	YouTube Live	M1: P&S PIM Course Presentation (PDF) (PPT)	Required Materials Recommended Materials	HW 0 Out
W2	15.03 Tue.		Hands-on Project Proposals		
	17.03 Thu.	YouTube Premiere	M2: Real-world PIM: UPMEM PIM (PDF) (PPT)		
W3	24.03 Thu.	YouTube Live	M3: Real-world PIM: Microbenchmarking of UPMEM PIM (PDF) (PPT)		
W4	31.03 Thu.	YouTube Live	M4: Real-world PIM: Samsung HBM-PIM (PDF) (PPT)		
W5	07.04 Thu.	YouTube Live	M5: How to Evaluate Data Movement Bottlenecks (PDF) (PPT)		
W6	14.04 Thu.	YouTube Live	M6: Real-world PIM: SK Hynix AiM (PDF) (PPT)		
W7	21.04 Thu.	YouTube Premiere	M7: Programming PIM Architectures (PDF) (PPT)		
W8	28.04 Thu.	YouTube Premiere	M8: Benchmarking and Workload Suitability on PIM (PDF) (PPT)		
W9	05.05 Thu.	YouTube Premiere	M9: Real-world PIM: Samsung AxDIMM (PDF) (PPT)		
W10	12.05 Thu.	YouTube Premiere	M10: Real-world PIM: Alibaba HB-PNM (PDF) (PPT)		
W11	19.05 Thu.	YouTube Live	M11: SpMV on a Real PIM Architecture (PDF) (PPT)		
W12	26.05 Thu.	YouTube Live	M12: End-to-End Framework for Processing-using-Memory (PDF) (PPT)		
W13	02.06 Thu.	YouTube Live	M13: Bit-Serial SIMD Processing using DRAM (PDF) (PPT)		
W14	09.06 Thu.	YouTube Live	M14: Analyzing and Mitigating ML Inference Bottlenecks (PDF) (PPT)		
W15	15.06 Thu.	YouTube Live	M15: In-Memory HTAP Databases with HW/SW Co-design (PDF) (PPT)		
W16	23.06 Thu.	YouTube Live	M16: In-Storage Processing for Genome Analysis (PDF) (PPT)		
W17	18.07 Mon.	YouTube Premiere	M17: How to Enable the Adoption of PIM? (PDF) (PPT)		
W18	09.08 Tue.	YouTube Premiere	SS1: ISVLSI 2022 Special Session on PIM (PDF & PPT)		

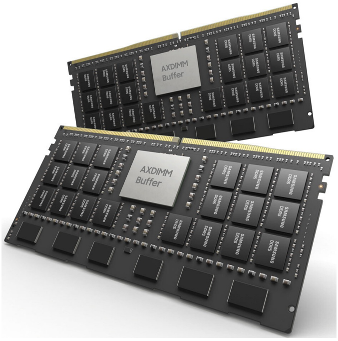
https://www.youtube.com/onurmutlulectures

Processing-in-Memory Landscape Today

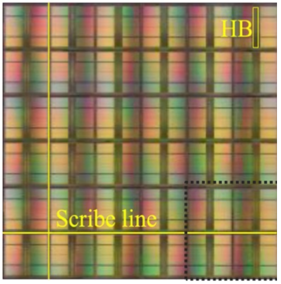


Micron

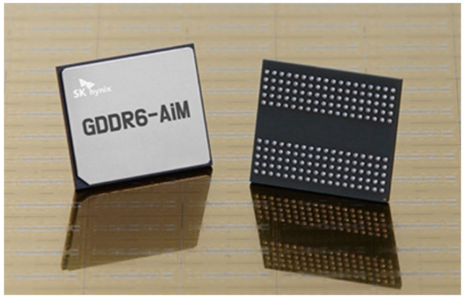
AUTOMATA PROCESSING



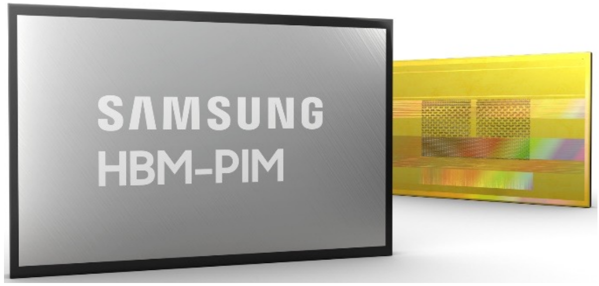
[Samsung 2021]



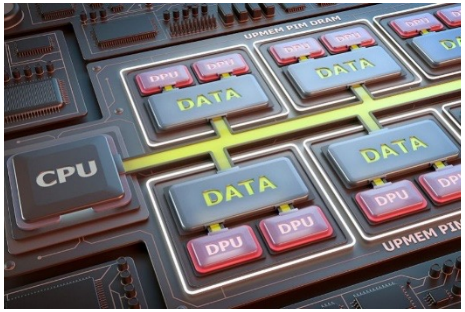
[Alibaba 2022]



[SK Hynix 2022]



[Samsung 2021]

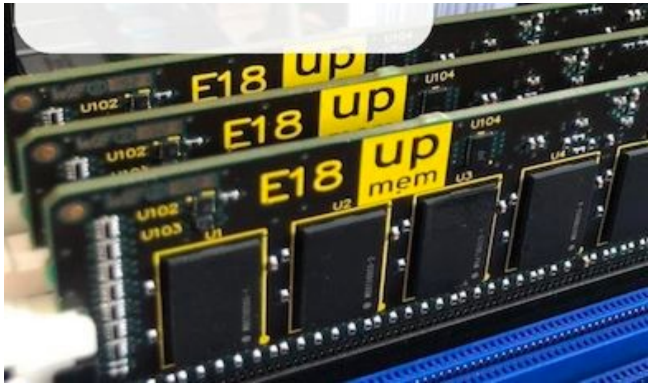


[UPMEM 2019]

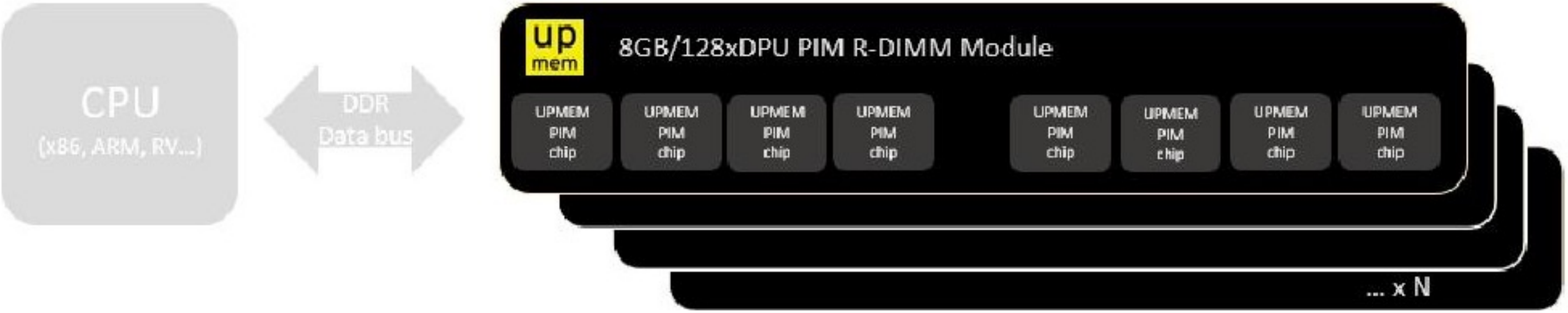
This does not include many experimental chips and startups

UPMEM Processing-in-DRAM Engine (2019)

- Processing in DRAM Engine
- Includes **standard DIMM modules**, with a **large number of DPU processors** combined with DRAM chips.



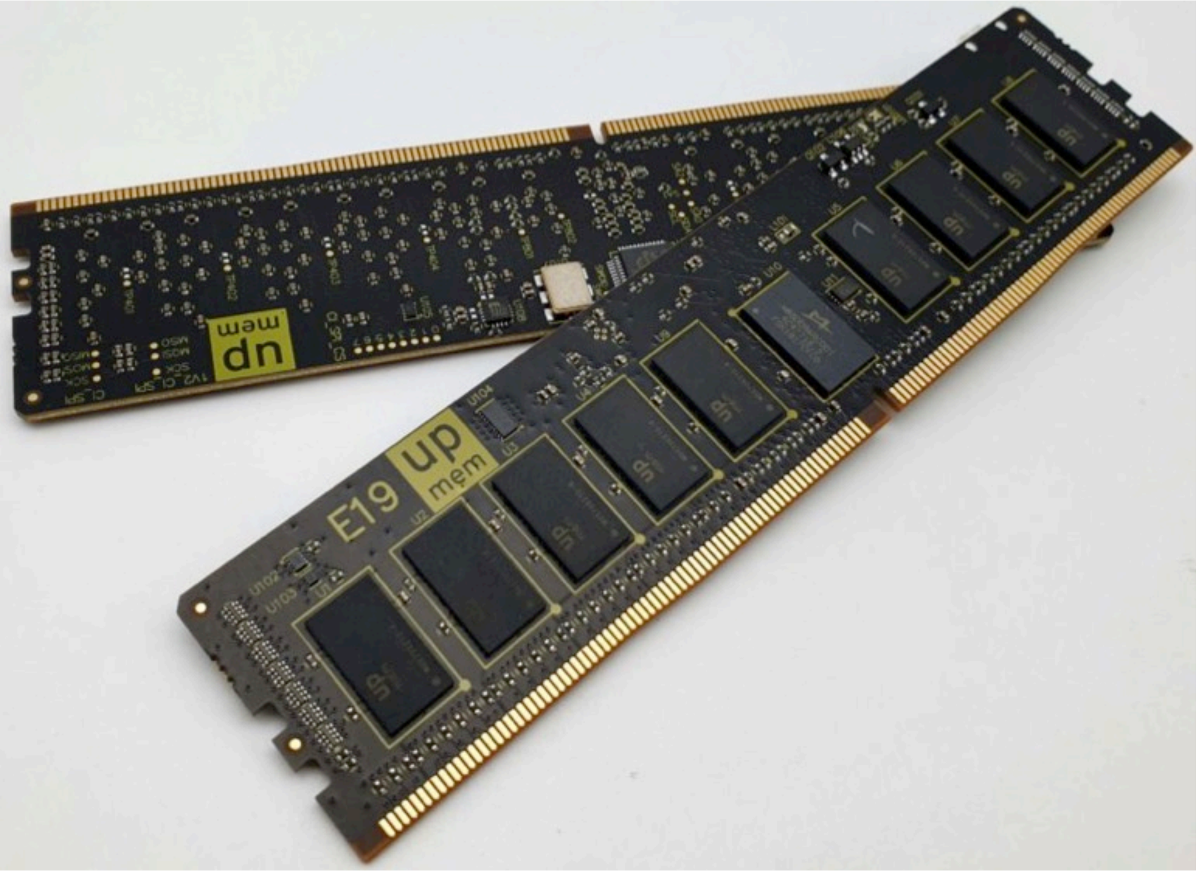
- Replaces **standard** DIMMs
  - DDR4 R-DIMM modules
    - 8GB+128 DPUs (16 PIM chips)
    - Standard 2x-nm DRAM process
  - **Large amounts of** compute & memory bandwidth



https://www.anandtech.com/show/14750/hot-chips-31-analysis-inmemory-processing-by-upmem
https://www.upmem.com/video-upmem-presenting-its-true-processing-in-memory-solution-hot-chips-2019/

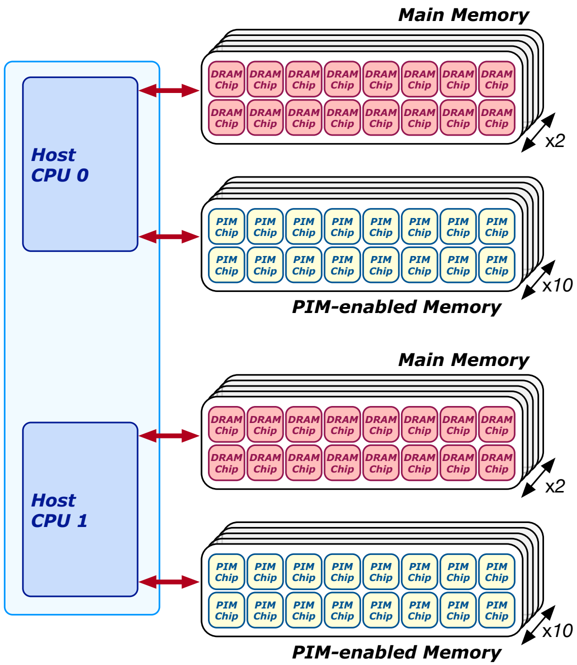
UPMEM Memory Modules

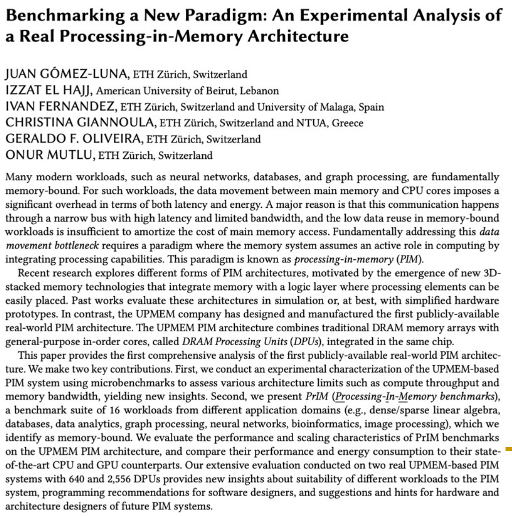
- E19: 8 chips DIMM (1 rank). DPUs @ 267 MHz
- P21: 16 chips DIMM (2 ranks). DPUs @ 350 MHz



www.upmem.com

2,560-DPU Processing-in-Memory System





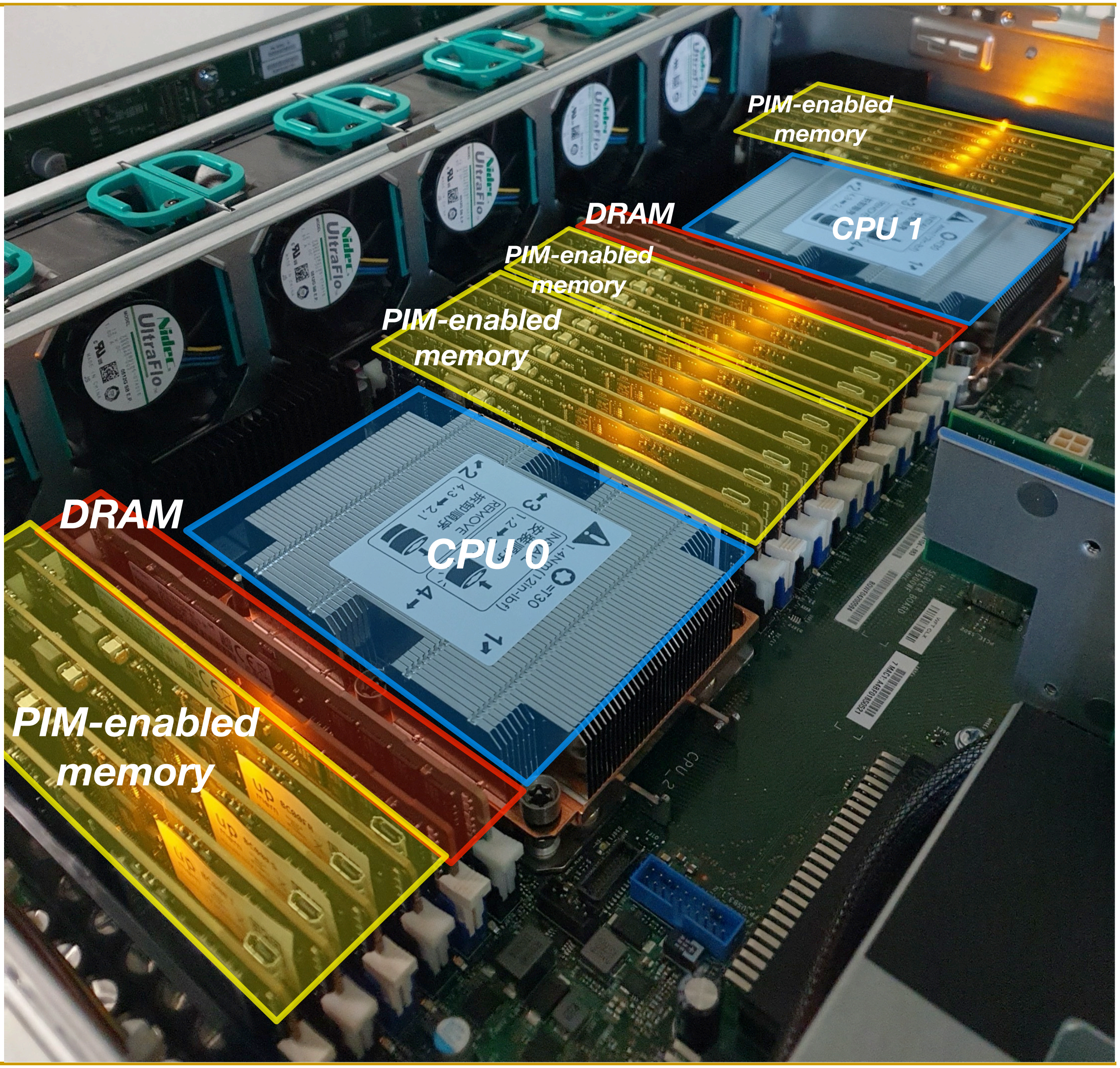
Benchmarking a New Paradigm: An Experimental Analysis of a Real Processing-in-Memory Architecture

JUAN GÓMEZ-LUNA, ETH Zürich, Switzerland
IZZAT EL HAJJ, American University of Beirut, Lebanon
IVAN FERNANDEZ, ETH Zürich, Switzerland and University of Malaga, Spain
CHRISTINA GIANNOULA, ETH Zürich, Switzerland and NTUA, Greece
GERALDO F. OLIVEIRA, ETH Zürich, Switzerland
ONUR MUTLU, ETH Zürich, Switzerland

Many modern workloads, such as neural networks, databases, and graph processing, are fundamentally memory-bound. For such workloads, the data movement between main memory and CPU cores imposes a significant overhead in terms of both latency and energy. A major reason is that this communication happens through a narrow bus with high latency and limited bandwidth, and the low data reuse in memory-bound workloads is insufficient to amortize the cost of main memory access. Fundamentally addressing this *data movement bottleneck* requires a paradigm where the memory system assumes an active role in computing by integrating processing capabilities. This paradigm is known as *processing-in-memory* (*PIM*).

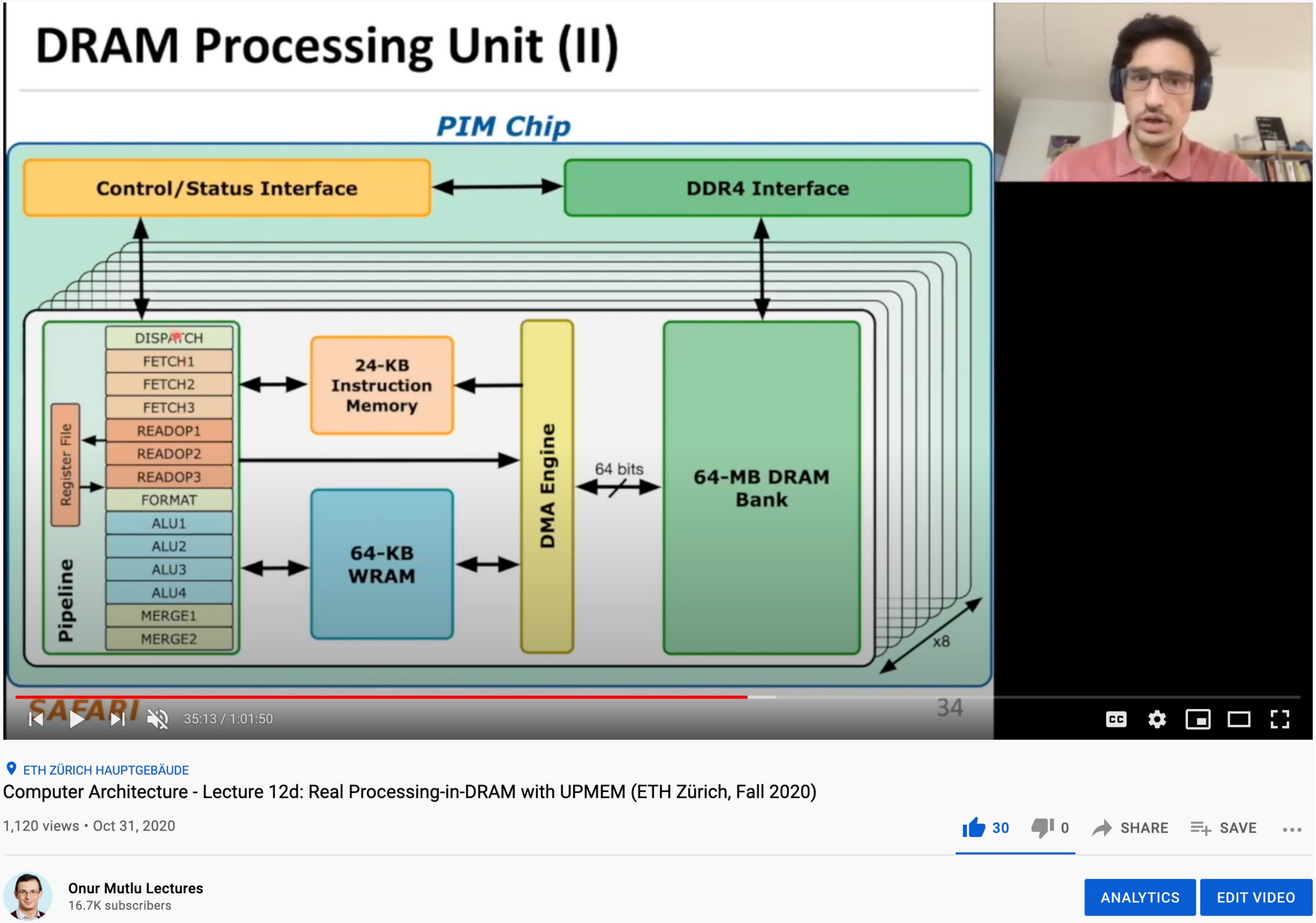
Recent research explores different forms of PIM architectures, motivated by the emergence of new 3D-stacked memory technologies that integrate memory with a logic layer where processing elements can be easily placed. Past works evaluate these architectures in simulation or, at best, with simplified hardware prototypes. In contrast, the UPMEM company has designed and manufactured the first publicly-available real-world PIM architecture. The UPMEM PIM architecture combines traditional DRAM memory arrays with general-purpose in-order cores, called *DRAM Processing Units* (*DPUs*), integrated in the same chip.

This paper provides the first comprehensive analysis of the first publicly-available real-world PIM architecture. We make two key contributions. First, we conduct an experimental characterization of the UPMEM-based PIM system using microbenchmarks to assess various architecture limits such as compute throughput and memory bandwidth, yielding new insights. Second, we present *PrIM* (*Processing-In-Memory benchmarks*), a benchmark suite of 16 workloads from different application domains (e.g., dense/sparse linear algebra, databases, data analytics, graph processing, neural networks, bioinformatics, image processing), which we identify as memory-bound. We evaluate the performance and scaling characteristics of PrIM benchmarks on the UPMEM PIM architecture, and compare their performance and energy consumption to their state-of-the-art CPU and GPU counterparts. Our extensive evaluation conducted on two real UPMEM-based PIM systems with 640 and 2,556 DPUs provides new insights about suitability of different workloads to the PIM system, programming recommendations for software designers, and suggestions and hints for hardware and architecture designers of future PIM systems.



https://arxiv.org/pdf/2105.03814.pdf

More on the UPMEM PIM System



DRAM Processing Unit (II)

ETH ZÜRICH HAUPTGEBÄUDE

Computer Architecture - Lecture 12d: Real Processing-in-DRAM with UPMEM (ETH Zürich, Fall 2020)

1,120 views • Oct 31, 2020

Onur Mutlu Lectures

16.7K subscribers

https://www.youtube.com/watch?v=Sscy1Wrr22A&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=26

SRC TECHCON Presentation

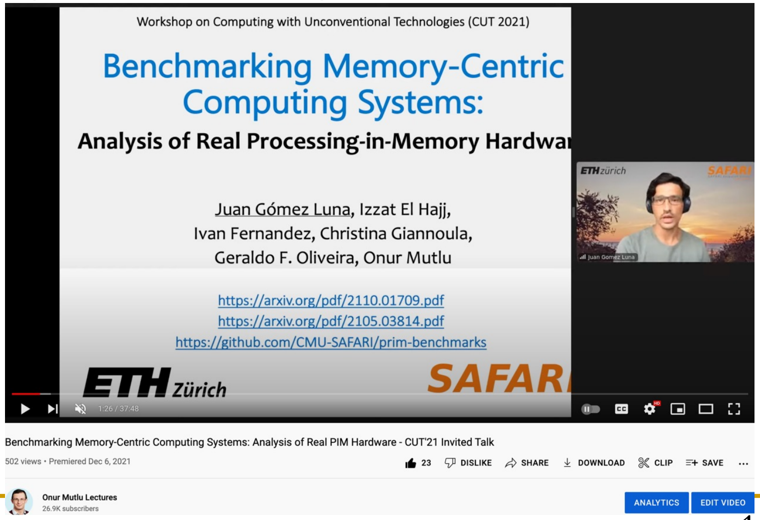
- Dr. Juan Gomez-Luna
  - Benchmarking Memory-Centric Computing Systems: Analysis of Real Processing-in-Memory Hardware
  - Based on two major works
    - https://arxiv.org/pdf/2105.03814.pdf
    - https://arxiv.org/pdf/2207.07886.pdf

Benchmarking Memory-Centric Computing Systems: Analysis of Real Processing-In-Memory Hardware

Year: 2021, Pages: 1-7
DOI Bookmark: 10.1109/IGSC54211.2021.9651614

Authors

Juan Gómez-Luna, ETH Zürich
Izzat El Hajj, American University of Beirut
Ivan Fernandez, University of Malaga
Christina Giannoula, National Technical University of Athens
Geraldo F. Oliveira, ETH Zürich
Onur Mutlu, ETH Zürich



Workshop on Computing with Unconventional Technologies (CUT 2021)

Benchmarking Memory-Centric Computing Systems:
Analysis of Real Processing-in-Memory Hardwa

Juan Gómez Luna, Izzat El Hajj,
Ivan Fernandez, Christina Giannoula,
Geraldo F. Oliveira, Onur Mutlu

https://arxiv.org/pdf/2110.01709.pdf
https://arxiv.org/pdf/2105.03814.pdf
https://github.com/CMU-SAFARI/prim-benchmarks

Benchmarking Memory-Centric Computing Systems: Analysis of Real PIM Hardware - CUT'21 Invited Talk

502 views • Premiered Dec 6, 2021

Onur Mutlu Lectures
26.9K subscribers

https://www.youtube.com/watch?v=nphV36SrysA

UPMEM PIM System Summary & Analysis

- Juan Gomez-Luna, Izzat El Hajj, Ivan Fernandez, Christina Giannoula, Geraldo F. Oliveira, and Onur Mutlu,
  "Benchmarking Memory-Centric Computing Systems: Analysis of Real Processing-in-Memory Hardware"
  *Invited Paper at Workshop on Computing with Unconventional Technologies (**CUT**)*, Virtual, October 2021.
  [arXiv version]
  [PrIM Benchmarks Source Code]
  [Slides (pptx) (pdf)]
  [Talk Video (37 minutes)]
  [Lightning Talk Video (3 minutes)]

Benchmarking Memory-Centric Computing Systems: Analysis of Real Processing-in-Memory Hardware

Juan Gómez-Luna	Izzat El Hajj	Ivan Fernandez	Christina Giannoula	Geraldo F. Oliveira	Onur Mutlu
ETH Zürich	*American University of Beirut*	*University of Malaga*	*National Technical University of Athens*	*ETH Zürich*	*ETH Zürich*

AIM (PIM Sequence Alignment Framework)

- Safaa Diab, Amir Nassereldine, Mohammed Alser, Juan Gómez Luna, Onur Mutlu, and Izzat El Hajj,
  "A Framework for High-throughput Sequence Alignment using Real Processing-in-Memory Systems"
  Bioinformatics, [published online on] 27 March 2023.
  [Online link at Bioinformatics Journal]
  [arXiv preprint]
  [AiM Source Code]

A Framework for High-throughput Sequence Alignment using Real Processing-in-Memory Systems

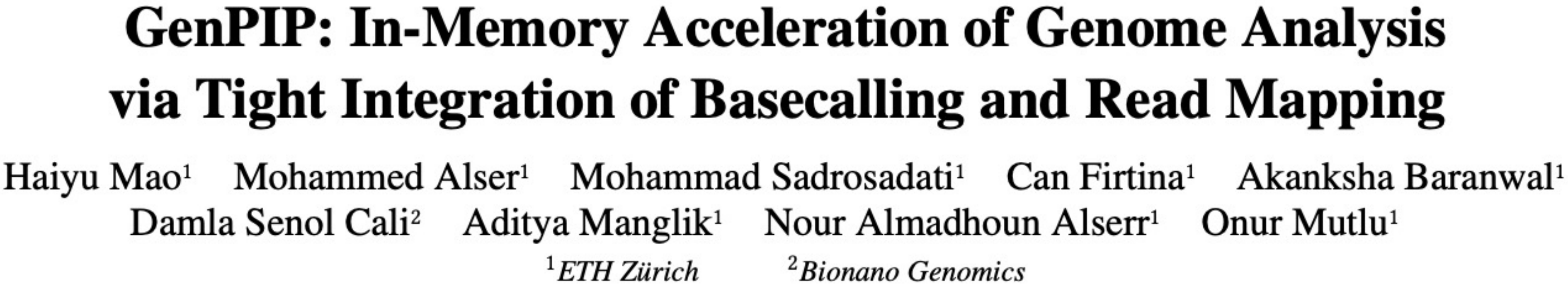
Safaa Diab [1] Amir Nassereldine [1] Mohammed Alser [2] Juan Gómez Luna [2]
Onur Mutlu [2] Izzat El Hajj [1]

[1] American University of Beirut [2] ETH Zürich

https://arxiv.org/pdf/2208.01243.pdf

Accelerating Basecalling + Read Mapping via PIM

- Haiyu Mao, Mohammed Alser, Mohammad Sadrosadati, Can Firtina, Akanksha Baranwal, Damla Senol Cali, Aditya Manglik, Nour Almadhoun Alserr, and Onur Mutlu,
  "GenPIP: In-Memory Acceleration of Genome Analysis via Tight Integration of Basecalling and Read Mapping"
  *Proceedings of the 55th International Symposium on Microarchitecture (**MICRO**)*, Chicago, IL, USA, October 2022.
  [Slides (pptx) (pdf)]
  [Longer Lecture Slides (pptx) (pdf)]
  [Lecture Video (25 minutes)]
  [arXiv version]



GenPIP: In-Memory Acceleration of Genome Analysis via Tight Integration of Basecalling and Read Mapping

Haiyu Mao[1] Mohammed Alser[1] Mohammad Sadrosadati[1] Can Firtina[1] Akanksha Baranwal[1]
Damla Senol Cali[2] Aditya Manglik[1] Nour Almadhoun Alserr[1] Onur Mutlu[1]
[1] *ETH Zürich* [2] *Bionano Genomics*

https://arxiv.org/pdf/2209.08600.pdf

Agenda

- The Problem: DNA Read Mapping
  - State-of-the-art Read Mapper Design

- Algorithmic Acceleration
  - Exploiting Structure of the Genome
  - Exploiting SIMD Instructions

- Hardware Acceleration
  - Specialized Architectures
  - Processing in Memory & Storage

- Future Opportunities: New Technologies & Applications

Newer Genome Sequencing Technologies

Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions

Damla Senol Cali ✉, Jeremie S Kim, Saugata Ghose, Can Alkan, Onur Mutlu

Briefings in Bioinformatics, bby017, https://doi.org/10.1093/bib/bby017

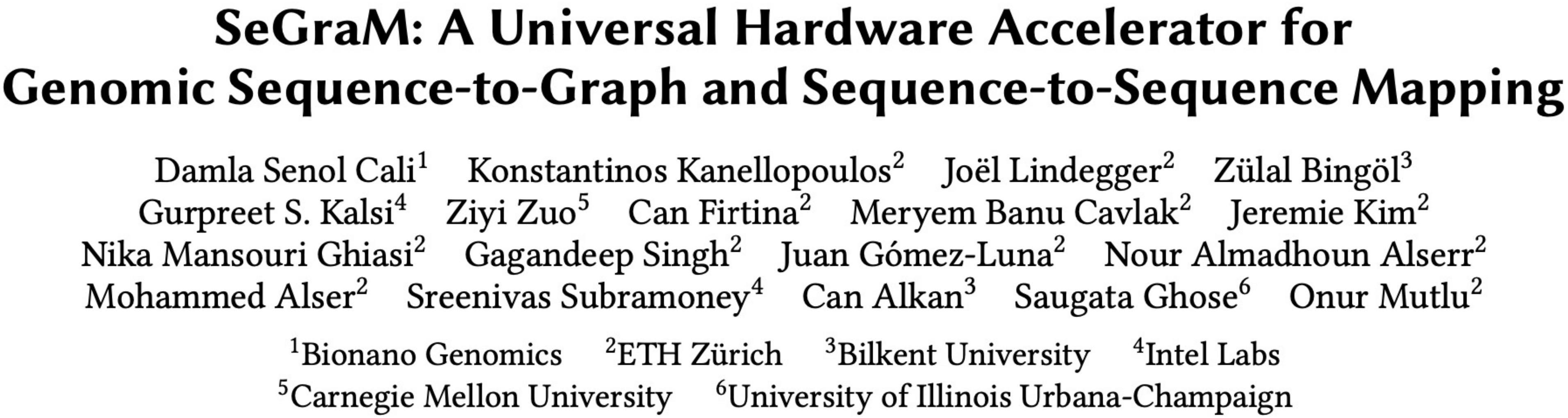
Published: 02 April 2018 **Article history** ▾

Oxford Nanopore MinION

Senol Cali+, "**Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions**," Briefings in Bioinformatics, 2018.
[Open arxiv.org version] [Slides (pptx) (pdf)] [Talk Video at AACBB 2019]

New Applications: Graph Genomes

- Damla Senol Cali, Konstantinos Kanellopoulos, Joel Lindegger, Zulal Bingol, Gurpreet S. Kalsi, Ziyi Zuo, Can Firtina, Meryem Banu Cavlak, Jeremie Kim, Nika MansouriGhiasi, Gagandeep Singh, Juan Gomez-Luna, Nour Almadhoun Alserr, Mohammed Alser, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu,
"SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping"
Proceedings of the 49th International Symposium on Computer Architecture (***ISCA***), New York, June 2022.
[arXiv version]



SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping

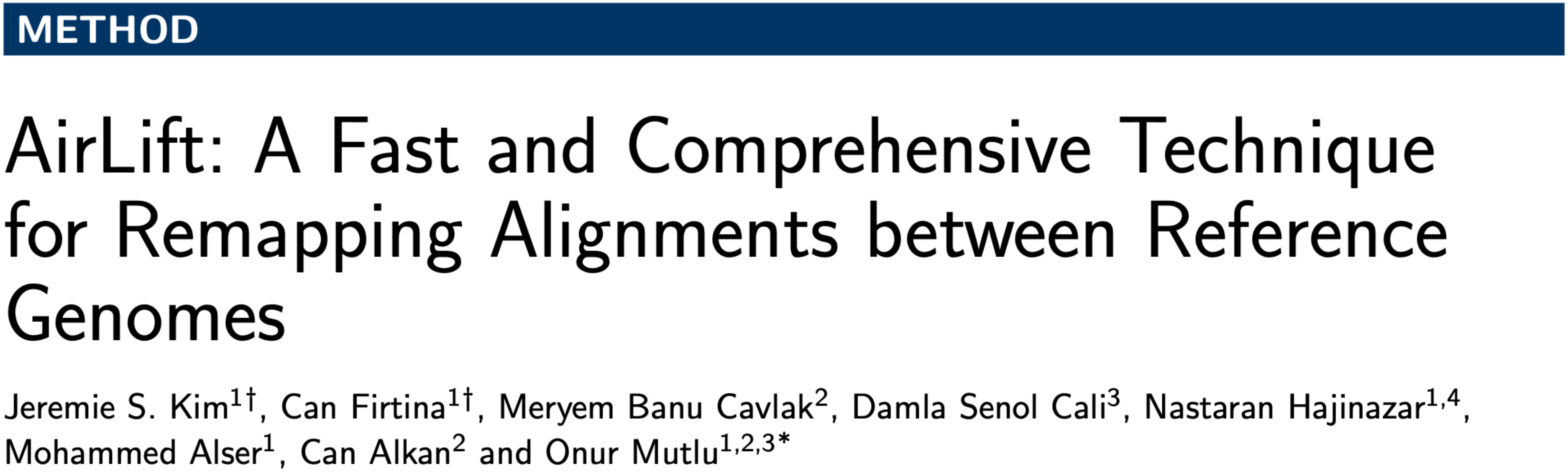
Damla Senol Cali[1] Konstantinos Kanellopoulos[2] Joël Lindegger[2] Zülal Bingöl[3]
Gurpreet S. Kalsi[4] Ziyi Zuo[5] Can Firtina[2] Meryem Banu Cavlak[2] Jeremie Kim[2]
Nika Mansouri Ghiasi[2] Gagandeep Singh[2] Juan Gómez-Luna[2] Nour Almadhoun Alserr[2]
Mohammed Alser[2] Sreenivas Subramoney[4] Can Alkan[3] Saugata Ghose[6] Onur Mutlu[2]

[1]Bionano Genomics [2]ETH Zürich [3]Bilkent University [4]Intel Labs
[5]Carnegie Mellon University [6]University of Illinois Urbana-Champaign

https://arxiv.org/pdf/2205.05883.pdf

New Applications: Frequent Reference Updates

- Jeremie S. Kim, Can Firtina, Meryem Banu Cavlak, Damla Senol Cali, Nastaran Hajinazar, Mohammed Alser, Can Alkan, and Onur Mutlu,
  "AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes"
  Preprint in arXiv and bioRxiv, 2021.
  [bioRxiv preprint]
  [arXiv preprint]
  [AirLift Source Code and Data]



METHOD

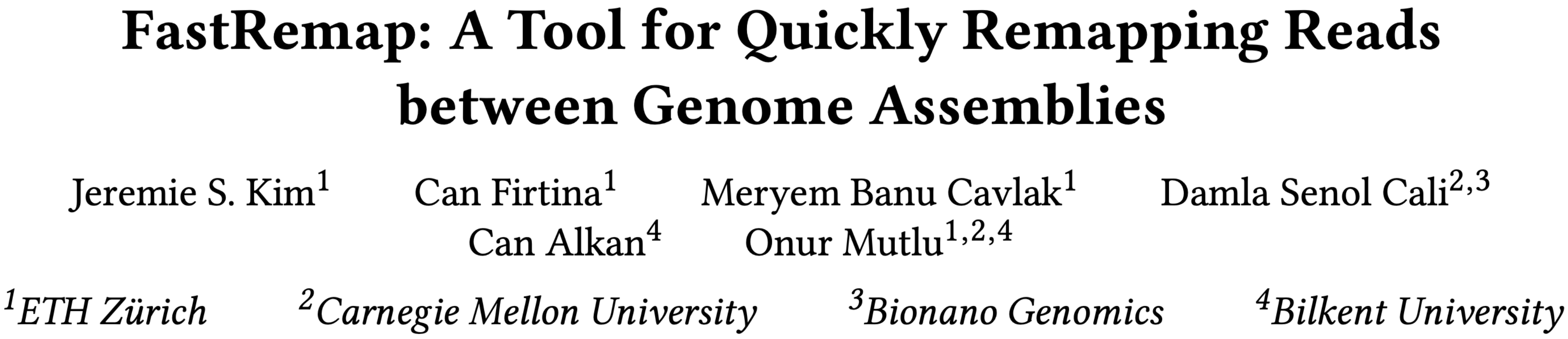
AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes

Jeremie S. Kim[1†], Can Firtina[1†], Meryem Banu Cavlak[2], Damla Senol Cali[3], Nastaran Hajinazar[1,4], Mohammed Alser[1], Can Alkan[2] and Onur Mutlu[1,2,3*]

https://arxiv.org/pdf/1912.08735.pdf

Mapping Constant Regions Between References

- Jeremie S. Kim, Can Firtina, Meryem Banu Cavlak, Damla Senol Cali, Can Alkan, and Onur Mutlu,
  "FastRemap: A Tool for Quickly Remapping Reads between Genome Assemblies"
  Bioinformatics, btac554.
  [FastRemap Source Code]



FastRemap: A Tool for Quickly Remapping Reads between Genome Assemblies

Jeremie S. Kim[1] Can Firtina[1] Meryem Banu Cavlak[1] Damla Senol Cali[2,3]
Can Alkan[4] Onur Mutlu[1,2,4]

[1]*ETH Zürich* [2]*Carnegie Mellon University* [3]*Bionano Genomics* [4]*Bilkent University*

https://arxiv.org/pdf/2201.06255.pdf

New Frontiers: Raw Signal Analysis

- **To appear at ISMB/ECCB 2023**

RawHash: Enabling Fast and Accurate Real-Time Analysis of Raw Nanopore Signals for Large Genomes

Can Firtina[1] Nika Mansouri Ghiasi[1] Joel Lindegger[1] Gagandeep Singh[1]
Meryem Banu Cavlak[1] Haiyu Mao[1] Onur Mutlu[1]
[1] *ETH Zurich*

https://arxiv.org/pdf/2301.09200.pdf

New Frontiers: Raw Signal Analysis

- **To appear at APBC 2023**

TargetCall: Eliminating the Wasted Computation in Basecalling via Pre-Basecalling Filtering

Meryem Banu Cavlak[1] Gagandeep Singh[1] Mohammed Alser[1] Can Firtina[1] Joël Lindegger[1]
Mohammad Sadrosadati[1] Nika Mansouri Ghiasi[1] Can Alkan[2] Onur Mutlu[1]
[1] *ETH Zürich* [2] *Bilkent University*

https://arxiv.org/pdf/2301.09200.pdf

A Bright Future for Intelligent Genome Analysis

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu
"Accelerating Genome Analysis: A Primer on an Ongoing Journey" IEEE Micro, August 2020.

MinION from ONT

Accelerating Genome Analysis: A Primer on an Ongoing Journey
Sept.-Oct. 2020, pp. 65-75, vol. 40
DOI Bookmark: 10.1109/MM.2020.3013728

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications
July-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

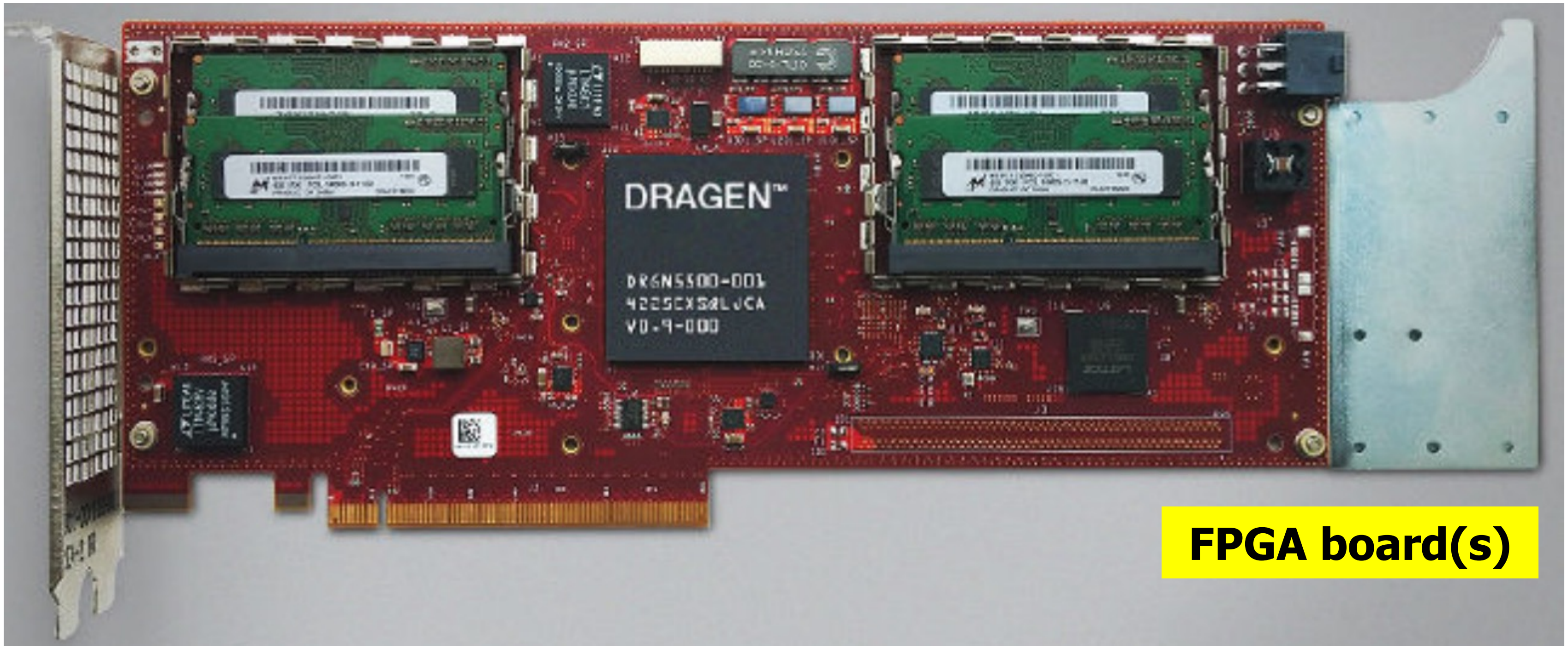
SmidgION from ONT

Conclusion

Things Are Happening In Industry

Illumina DRAGEN Bio-IT Platform (2018)

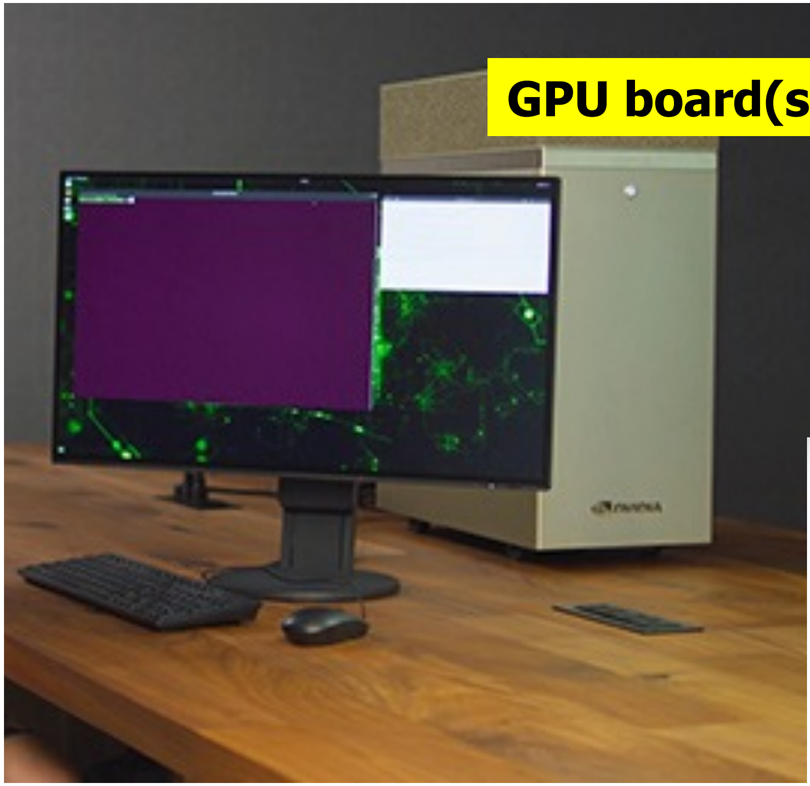
- Processes whole genome at 30x coverage in ~25 minutes with hardware support for data compression



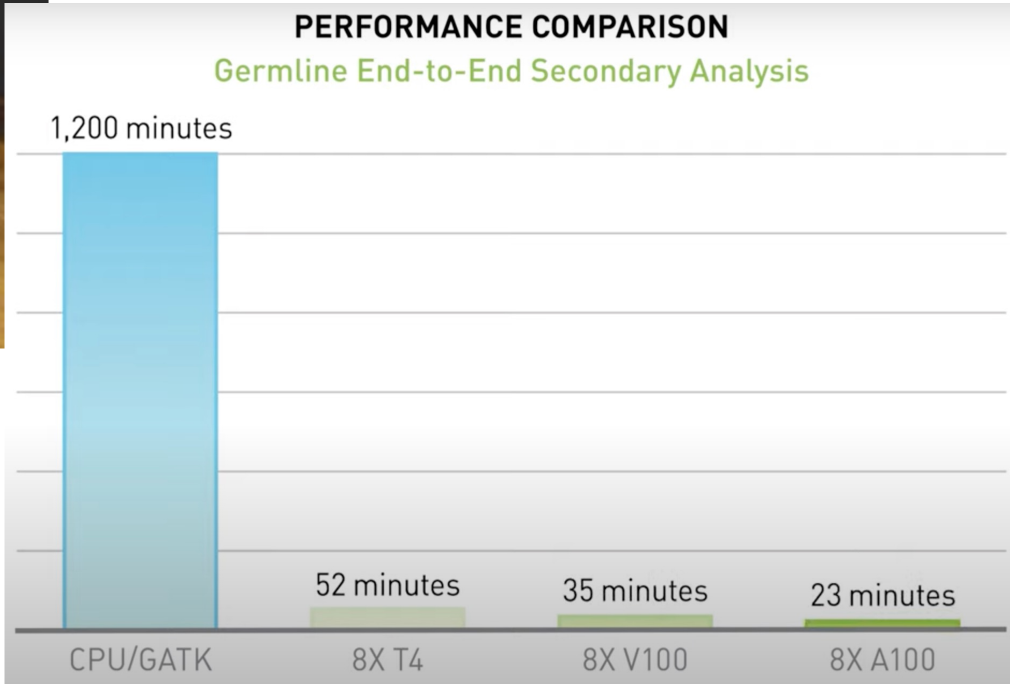
FPGA board(s)

emea.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html
emea.illumina.com/company/news-center/press-releases/2018/2349147.html

NVIDIA Clara Parabricks (2020)

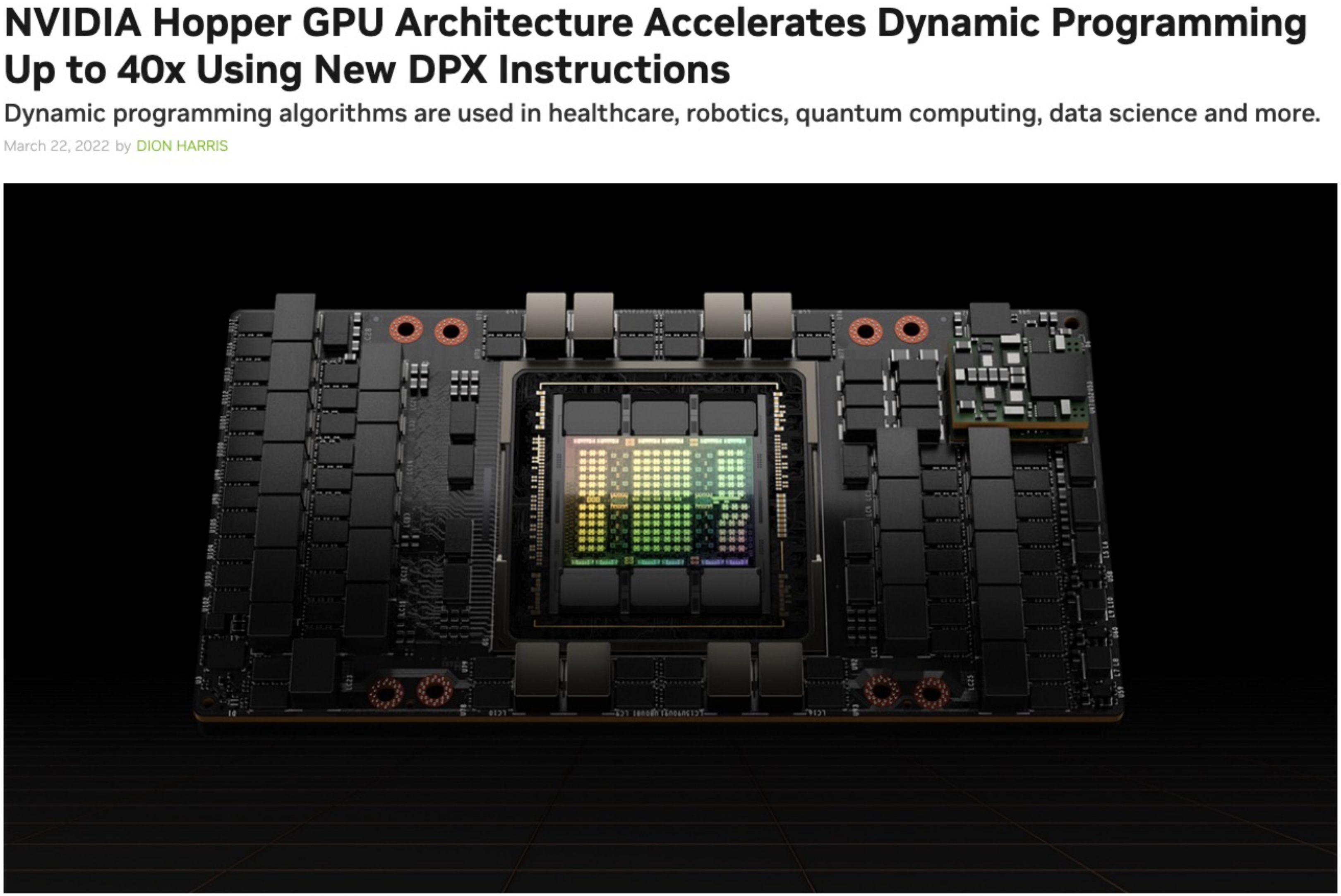


A University of Michigan startup in 2018 joined NVIDIA in 2020



https://developer.nvidia.com/clara-parabricks

NVIDIA Hopper DPX Instructions (2022)



NVIDIA Hopper GPU Architecture Accelerates Dynamic Programming Up to 40x Using New DPX Instructions

Dynamic programming algorithms are used in healthcare, robotics, quantum computing, data science and more.

March 22, 2022 by DION HARRIS

https://blogs.nvidia.com/blog/2022/03/22/nvidia-hopper-accelerates-dynamic-programming-using-dpx-instructions/

We are accelerating the transformation in how we analyze the human genome!

Bionano & NVIDIA:

Accelerating Analysis for Fast Time to Results

- Technological solution to **support higher throughput**
- **New high-performance algorithms** from Bionano
- **Powered by NVIDIA RTX™ 6000 Ada Generation GPUs**
- Analysis of highly complex cancer whole genomes in **less than 2 hours**
- Workflow tailored for a **small lab and IT footprint**

Recall Our Dream (from 2007)

- An embedded device that can perform comprehensive genome analysis in real time (within a minute)

- Still a long ways to go
  - Energy efficiency
  - Performance (latency)
  - Security & privacy
  - **Huge memory bottleneck**

Conclusion

- System design for bioinformatics is a critical problem
    - It has large scientific, medical, societal, personal implications

- This talk is about accelerating a key step in bioinformatics: genome sequence analysis
    - In particular, read mapping

- We covered various recent ideas to accelerate read mapping
    - My personal journey since September 2006

- **Many future opportunities exist**
    - **Especially with new sequencing technologies**
    - **Especially with new applications and use cases**

A Bright Future for Intelligent Genome Analysis

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu
“Accelerating Genome Analysis: A Primer on an Ongoing Journey” IEEE Micro, August 2020.

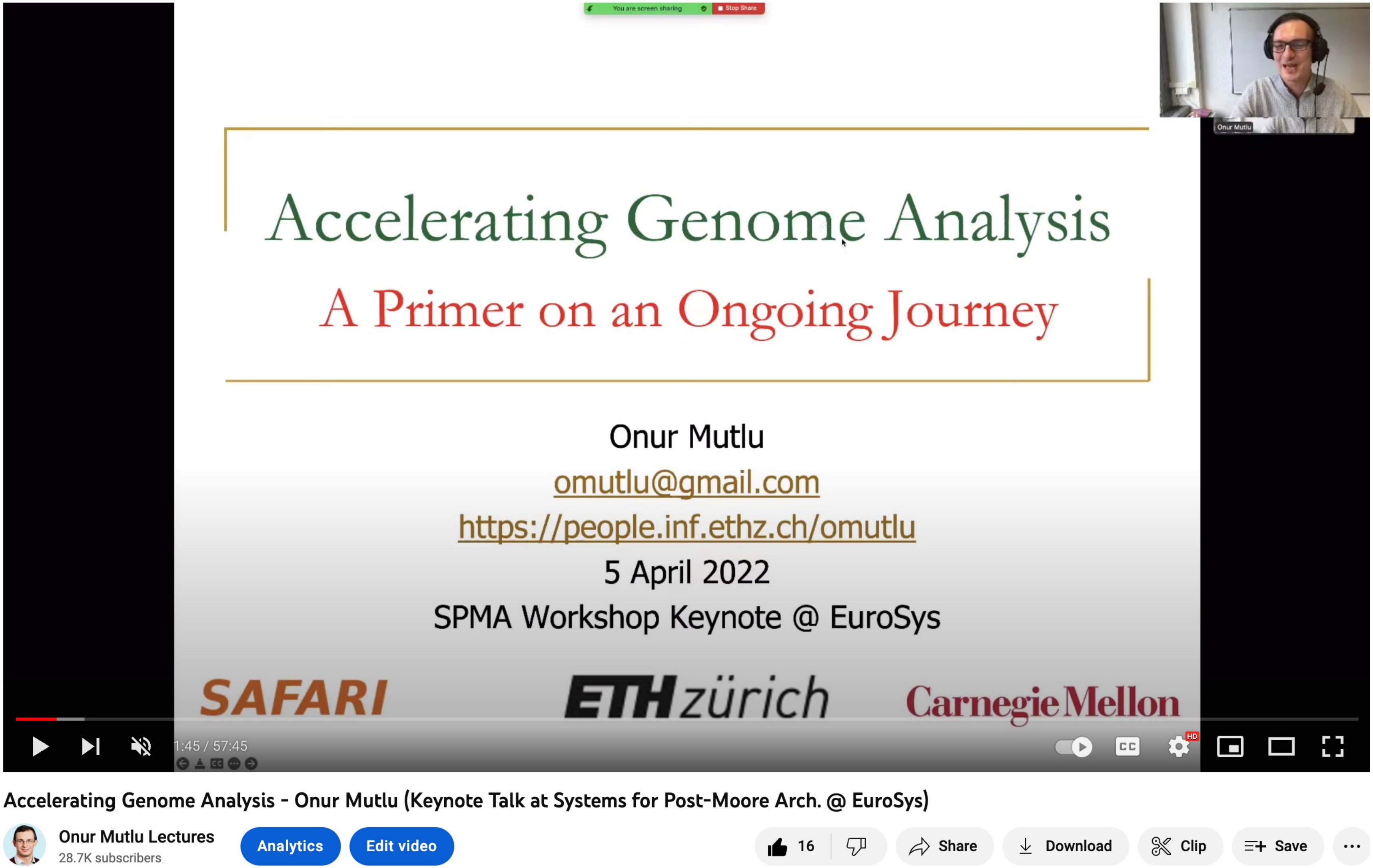
MinION from ONT

Accelerating Genome Analysis: A Primer on an Ongoing Journey
Sept.-Oct. 2020, pp. 65-75, vol. 40
DOI Bookmark: 10.1109/MM.2020.3013728

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications
July-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

SmidgION from ONT

A Longer Version of This Talk (I)



Accelerating Genome Analysis

A Primer on an Ongoing Journey

Onur Mutlu

omutlu@gmail.com

https://people.inf.ethz.ch/omutlu

5 April 2022

SPMA Workshop Keynote @ EuroSys

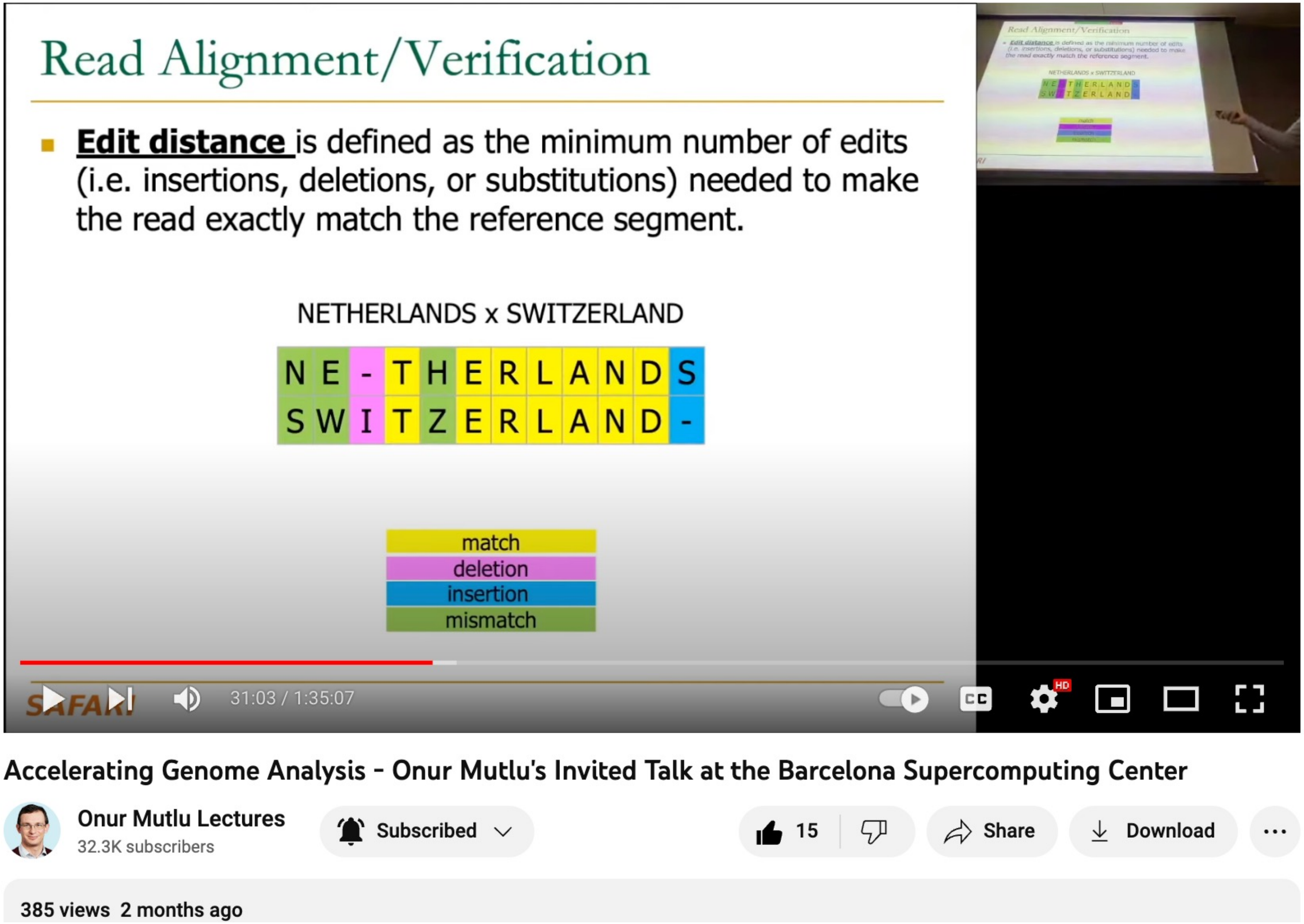
Accelerating Genome Analysis - Onur Mutlu (Keynote Talk at Systems for Post-Moore Arch. @ EuroSys)

Onur Mutlu Lectures

28.7K subscribers

https://www.youtube.com/watch?v=NCagwf0ivT0

A Longer Version of This Talk (II)



Read Alignment/Verification

- **Edit distance** is defined as the minimum number of edits (i.e. insertions, deletions, or substitutions) needed to make the read exactly match the reference segment.

NETHERLANDS x SWITZERLAND

N	E	-	T	H	E	R	L	A	N	D	S
S	W	I	T	Z	E	R	L	A	N	D	-

- 🟨 match
- 🟪 deletion
- 🟦 insertion
- 🟩 mismatch

Accelerating Genome Analysis - Onur Mutlu's Invited Talk at the Barcelona Supercomputing Center

https://www.youtube.com/watch?v=tVpg0XqU_c4

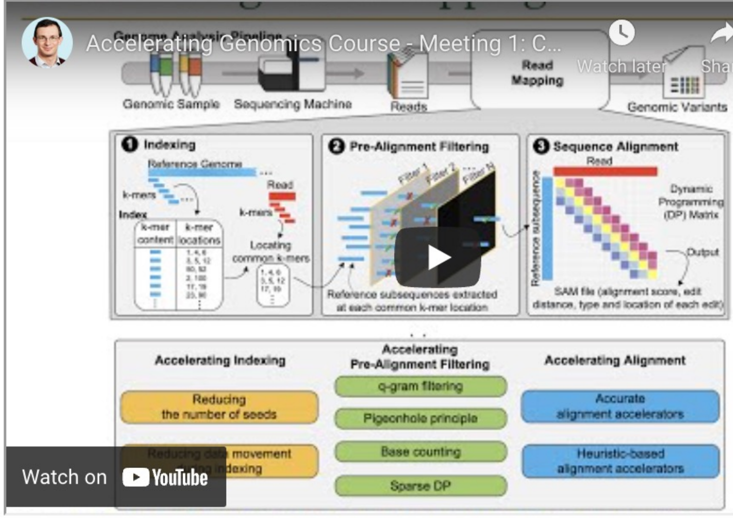
Genomics Course (Fall 2022)

- **Fall 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/fall2022/doku.php?id=bioinformatics
- **Spring 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=bioinformatics

- **Youtube Livestream (Fall 2022):**
  - https://www.youtube.com/watch?v=nA41964-9r8&list=PL5Q2soXY2Zi8tFlQvdxOdizD_EhVAMVQV
- **Youtube Livestream (Spring 2022):**
  - https://www.youtube.com/watch?v=DEL_5A_Y3TI&list=PL5Q2soXY2Zi8NrPDgOR1yRU_Cxxjw-u18

- Project course
  - Taken by Bachelor's/Master's students
  - Genomics lectures
  - Hands-on research exploration
  - Many research readings

https://www.youtube.com/onurmutlulectures



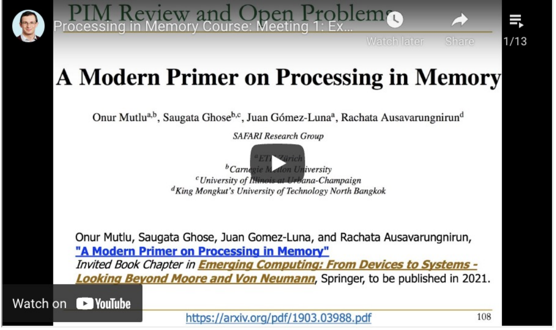
Accelerating Genomics Course - Meeting 1: C...

Spring 2022 Meetings/Schedule

Week	Date	Livestream	Meeting	Learning Materials
W1	11.3 Fri.	YouTube Live	M1: P&S Accelerating Genomics Course Introduction & Project Proposals (PDF) (PPT)	Required Materials Recommended Materials
W2	18.3 Fri.	YouTube Live	M2: Introduction to Sequencing (PDF) (PPT)	
W3	25.3 Fri.	YouTube Premiere	M3: Read Mapping (PDF) (PPT)	
W4	01.04 Fri.	YouTube Premiere	M4: GateKeeper (PDF) (PPT)	
W5	08.04 Fri.	YouTube Premiere	M5: MAGNET & Shouji (PDF) (PPT)	
W6	15.4 Fri.	YouTube Premiere	M6: SneakySnake (PDF) (PPT)	
W7	29.4 Fri.	YouTube Premiere	M7: GenStore (PDF) (PPT)	
W8	06.05 Fri.	YouTube Premiere	M8: GRIM-Filter (PDF) (PPT)	
W9	13.05 Fri.	YouTube Premiere	M9: Genome Assembly (PDF) (PPT)	
W10	20.05 Fri.	YouTube Live	M10: Genomic Data Sharing Under Differential Privacy (PDF) (PPT)	
W11	10.06 Fri.	YouTube Premiere	M11: Accelerating Genome Sequence Analysis (PDF) (PPT)	

PIM Course (Fall 2022)

- **Fall 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/fall2022/doku.php?id=processing_in_memory
- **Spring 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=processing_in_memory
- **Youtube Livestream (Fall 2022):**
  - https://www.youtube.com/watch?v=QLL0wQ9I4Dw&list=PL5Q2soXY2Zi8KzG2CQYRNQOVD0GOBrnKy
- **Youtube Livestream (Spring 2022):**
  - https://www.youtube.com/watch?v=9e4Chnwdovo&list=PL5Q2soXY2Zi-841fUYYUK9EsXKhQKRPyX
- Project course
  - Taken by Bachelor’s/Master’s students
  - Processing-in-Memory lectures
  - Hands-on research exploration
  - Many research readings



A Modern Primer on Processing in Memory

Onur Mutlu[a,b], Saugata Ghose[b,c], Juan Gómez-Luna[a], Rachata Ausavarungnirun[d]

SAFARI Research Group

[a]ETH Zürich
[b]Carnegie Mellon University
[c]University of Illinois at Urbana-Champaign
[d]King Mongkut’s University of Technology North Bangkok

Onur Mutlu, Saugata Ghose, Juan Gomez-Luna, and Rachata Ausavarungnirun, **"A Modern Primer on Processing in Memory"**
Invited Book Chapter in ***Emerging Computing: From Devices to Systems - Looking Beyond Moore and Von Neumann***, Springer, to be published in 2021.

https://arxiv.org/pdf/1903.03988.pdf 108

Spring 2022 Meetings/Schedule

Week	Date	Livestream	Meeting	Learning Materials	Assignments
W1	10.03 Thu.	YouTube Live	M1: P&S PIM Course Presentation (PDF) (PPT)	Required Materials Recommended Materials	HW 0 Out
W2	15.03 Tue.		Hands-on Project Proposals		
	17.03 Thu.	YouTube Premiere	M2: Real-world PIM: UPMEM PIM (PDF) (PPT)		
W3	24.03 Thu.	YouTube Live	M3: Real-world PIM: Microbenchmarking of UPMEM PIM (PDF) (PPT)		
W4	31.03 Thu.	YouTube Live	M4: Real-world PIM: Samsung HBM-PIM (PDF) (PPT)		
W5	07.04 Thu.	YouTube Live	M5: How to Evaluate Data Movement Bottlenecks (PDF) (PPT)		
W6	14.04 Thu.	YouTube Live	M6: Real-world PIM: SK Hynix AiM (PDF) (PPT)		
W7	21.04 Thu.	YouTube Premiere	M7: Programming PIM Architectures (PDF) (PPT)		
W8	28.04 Thu.	YouTube Premiere	M8: Benchmarking and Workload Suitability on PIM (PDF) (PPT)		
W9	05.05 Thu.	YouTube Premiere	M9: Real-world PIM: Samsung AxDIMM (PDF) (PPT)		
W10	12.05 Thu.	YouTube Premiere	M10: Real-world PIM: Alibaba HB-PNM (PDF) (PPT)		
W11	19.05 Thu.	YouTube Live	M11: SpMV on a Real PIM Architecture (PDF) (PPT)		
W12	26.05 Thu.	YouTube Live	M12: End-to-End Framework for Processing-using-Memory (PDF) (PPT)		
W13	02.06 Thu.	YouTube Live	M13: Bit-Serial SIMD Processing using DRAM (PDF) (PPT)		
W14	09.06 Thu.	YouTube Live	M14: Analyzing and Mitigating ML Inference Bottlenecks (PDF) (PPT)		
W15	15.06 Thu.	YouTube Live	M15: In-Memory HTAP Databases with HW/SW Co-design (PDF) (PPT)		
W16	23.06 Thu.	YouTube Live	M16: In-Storage Processing for Genome Analysis (PDF) (PPT)		
W17	18.07 Mon.	YouTube Premiere	M17: How to Enable the Adoption of PIM? (PDF) (PPT)		
W18	09.08 Tue.	YouTube Premiere	SS1: ISVLSI 2022 Special Session on PIM (PDF & PPT)		

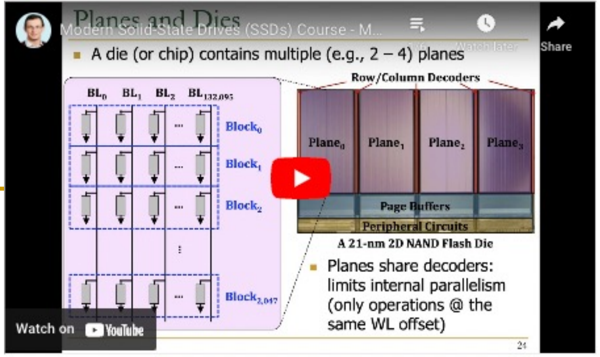
https://www.youtube.com/onurmutlulectures

SSD Course (Spring 2023)

- **Spring 2023 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/spring2023/doku.php?id=modern_ssds
- **Fall 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/fall2022/doku.php?id=modern_ssds

- **Youtube Livestream (Spring 2023):**
  - https://www.youtube.com/watch?v=4VTwOMmsnJY&list=PL5Q2soXY2Zi_8qOM5Icpp8hB2SHtm4z57&pp=iAQB
- **Youtube Livestream (Fall 2022):**
  - https://www.youtube.com/watch?v=hqLrd-Uj0aU&list=PL5Q2soXY2Zi9BJhenUq4JI5bwhAMpAp13&pp=iAQB

- Project course
  - Taken by Bachelor's/Master's students
  - SSD Basics and Advanced Topics
  - Hands-on research exploration
  - Many research readings



Planes and Dies

Modern Solid-State Drives (SSDs) Course - M...

- A die (or chip) contains multiple (e.g., 2 – 4) planes
- Planes share decoders: limits internal parallelism (only operations @ the same WL offset)

A 21-nm 2D NAND Flash Die

24

Fall 2022 Meetings/Schedule

Week	Date	Livestream	Meeting	Learning Materials	Assignments
W1	06.10		M1: P&S Course Presentation PDF PPT	Required Recommended	
W2	12.10	YouTube Live	M2: Basics of NAND Flash-Based SSDs PDF PPT	Required Recommended	
W3	19.10	YouTube Live	M3: NAND Flash Read/Write Operations PDF PPT	Required Recommended	
W4	26.10	YouTube Live	M4: Processing inside NAND Flash PDF PPT	Required Recommended	
W5	02.11	YouTube Live	M5: Advanced NAND Flash Commands & Mapping PDF PPT	Required Recommended	
W6	09.11	YouTube Live	M6: Processing inside Storage PDF PPT	Required Recommended	
W7	23.11	YouTube Live	M7: Address Mapping & Garbage Collection PDF PPT	Required Recommended	
W8	30.11	YouTube Live	M8: Introduction to MQSim PDF PPT	Required Recommended	
W9	14.12	YouTube Live	M9: Fine-Grained Mapping and Multi-Plane Operation-Aware Block Management PDF PPT	Required Recommended	
W10	04.01.2023	YouTube Premiere	M10a: NAND Flash Basics PDF PPT	Required Recommended	
			M10b: Reducing Solid-State Drive Read Latency by Optimizing Read-Retry PDF PPT Paper	Required Recommended	
			M10c: Evanesco: Architectural Support for Efficient Data Sanitization in Modern Flash-Based Storage Systems PDF PPT Paper	Required Recommended	
			M10d: DeepSketch: A New Machine Learning-Based Reference Search Technique for Post-Deduplication Delta Compression PDF PPT Paper	Required Recommended	
W11	11.01	YouTube Live	M11: FLIN: Enabling Fairness and Enhancing Performance in Modern NVMe Solid State Drives PDF PPT	Required	
W12	25.01	YouTube Premiere	M12: Flash Memory and Solid-State Drives PDF PPT	Recommended	

https://www.youtube.com/onurmutlulectures

Accelerating Genome Analysis

A Primer on an Ongoing Journey

Onur Mutlu
omutlu@gmail.com
https://people.inf.ethz.ch/omutlu

14 April 2023
BIO-Arch Workshop @ RECOMB

SAFARI ETH zürich Carnegie Mellon

Backup Slides for Further Info

Resources & Acknowledgments

Special Research Sessions & Courses

- Special Session at ISVLSI 2022: 9 cutting-edge talks

In-Memory Processing

ISVLSI 2022 Special Session

IEEE Computer Society Annual Symposium on VLSI

Adonis room
Ailathon resort, Paphos, Cyprus
July 4th, 2022

0:04 / 3:36:35 • Dr. Juan Gómez-Luna, "Introduction to the ISVLSI 2022 Special Session on Processing-in-Memory"

ISVLSI 2022 Special Session on Processing-in-Memory

1,286 views • Premiered Aug 9, 2022

Onur Mutlu Lectures
26.9K subscribers

https://www.youtube.com/watch?v=qeukNs5XI3g

Detailed Lectures on Genome Analysis

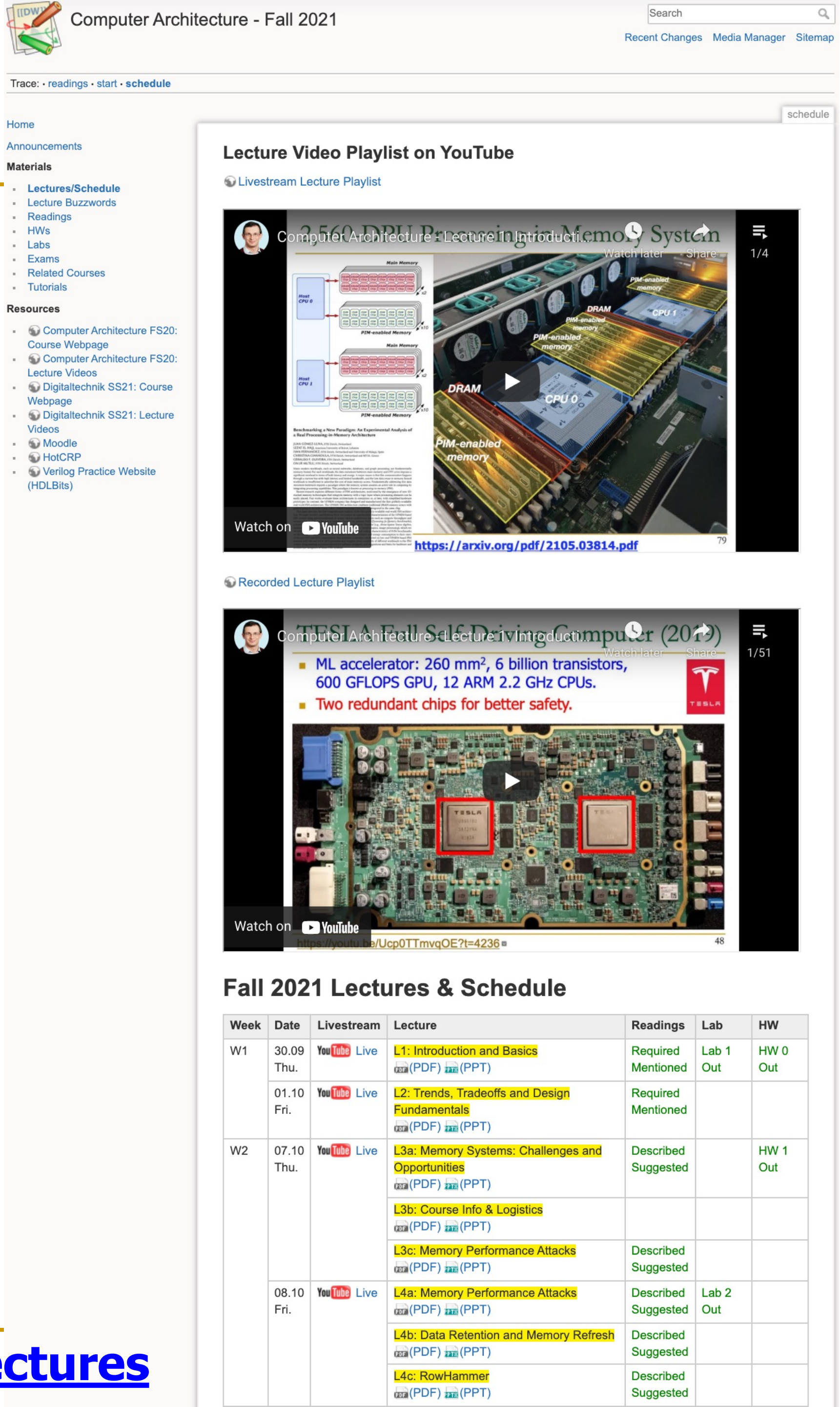
- Computer Architecture, Fall 2020, Lecture 3a
  - **Introduction to Genome Sequence Analysis** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=CrRb32v7SJc&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=5
- Computer Architecture, Fall 2020, Lecture 8
  - **Intelligent Genome Analysis** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=ygmQpdDTL7o&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=14
- Computer Architecture, Fall 2020, Lecture 9a
  - **GenASM: Approx. String Matching Accelerator** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=XoLpzmN-Pas&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=15
- Accelerating Genomics Project Course, Fall 2020, Lecture 1
  - **Accelerating Genomics** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=rgjl8ZyLsAg&list=PL5Q2soXY2Zi9E2bBVAgCqLgwiDRQDTyId

Comp Arch (Fall'21)

- **Fall 2021 Edition:**
  - https://safari.ethz.ch/architecture/fall2021/doku.php?id=schedule
- **Fall 2020 Edition:**
  - https://safari.ethz.ch/architecture/fall2020/doku.php?id=schedule

- **Youtube Livestream (2021):**
  - https://www.youtube.com/watch?v=4yfkM_5EFgo&list=PL5Q2soXY2Zi-Mnk1PxjEIG32HAGILkTOF
- **Youtube Livestream (2020):**
  - https://www.youtube.com/watch?v=c3mPdZA-Fmc&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN

- Master's level course
  - Taken by Bachelor's/Masters/PhD students
  - Cutting-edge research topics + fundamentals in Computer Architecture
  - 5 Simulator-based Lab Assignments
  - Potential research exploration
  - Many research readings



Computer Architecture - Fall 2021

Trace: · readings · start · **schedule**

Home

Announcements

Materials

- **Lectures/Schedule**
- Lecture Buzzwords
- Readings
- HWs
- Labs
- Exams
- Related Courses
- Tutorials

Resources

- Computer Architecture FS20: Course Webpage
- Computer Architecture FS20: Lecture Videos
- Digitaltechnik SS21: Course Webpage
- Digitaltechnik SS21: Lecture Videos
- Moodle
- HotCRP
- Verilog Practice Website (HDLBits)

Lecture Video Playlist on YouTube

Livestream Lecture Playlist

Recorded Lecture Playlist

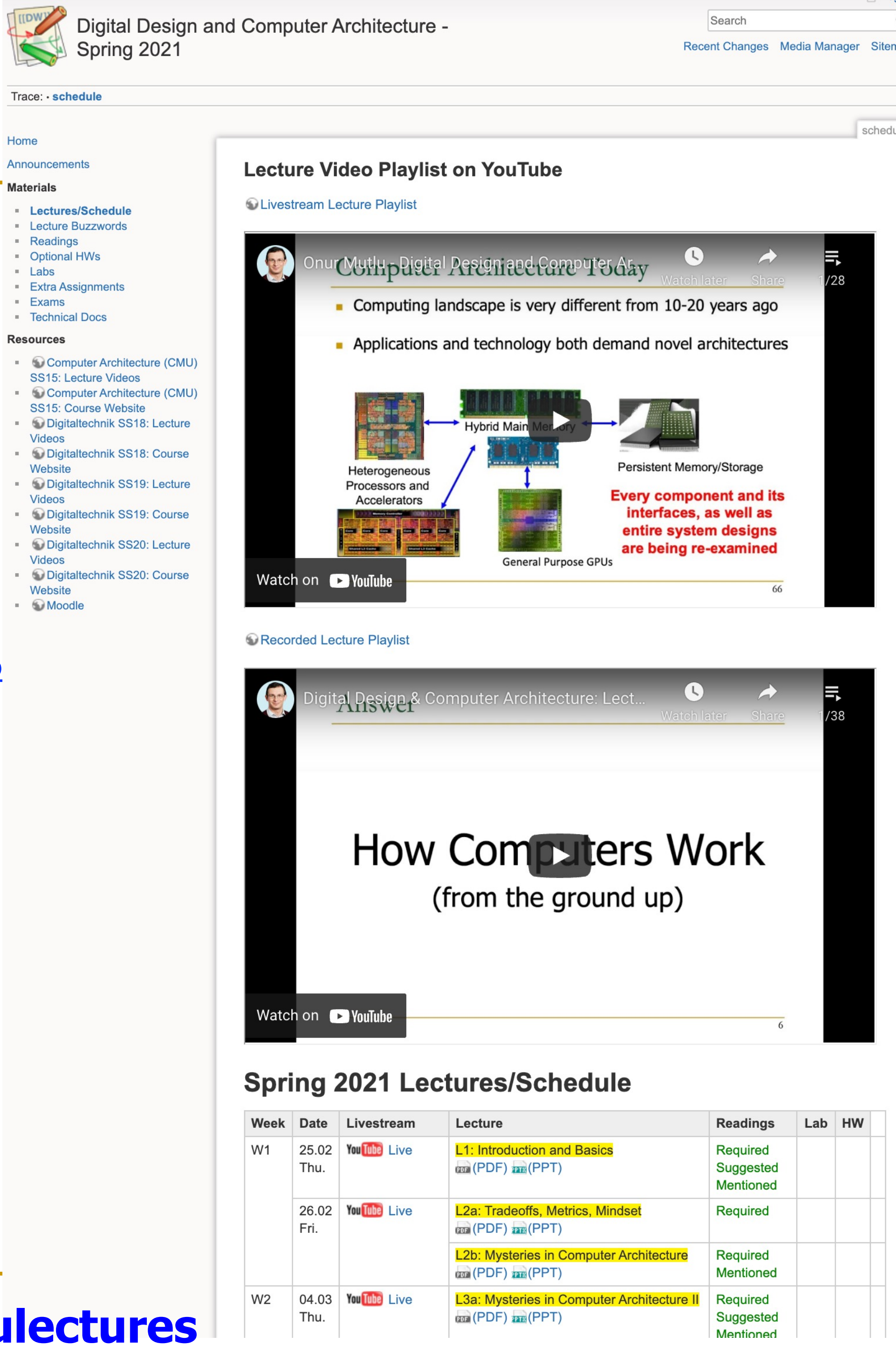
Fall 2021 Lectures & Schedule

Week	Date	Livestream	Lecture	Readings	Lab	HW
W1	30.09 Thu.	YouTube Live	L1: Introduction and Basics (PDF) (PPT)	Required Mentioned	Lab 1 Out	HW 0 Out
	01.10 Fri.	YouTube Live	L2: Trends, Tradeoffs and Design Fundamentals (PDF) (PPT)	Required Mentioned		
W2	07.10 Thu.	YouTube Live	L3a: Memory Systems: Challenges and Opportunities (PDF) (PPT)	Described Suggested		HW 1 Out
			L3b: Course Info & Logistics (PDF) (PPT)			
			L3c: Memory Performance Attacks (PDF) (PPT)	Described Suggested		
	08.10 Fri.	YouTube Live	L4a: Memory Performance Attacks (PDF) (PPT)	Described Suggested	Lab 2 Out	
			L4b: Data Retention and Memory Refresh (PDF) (PPT)	Described Suggested		
			L4c: RowHammer (PDF) (PPT)	Described Suggested		

https://www.youtube.com/onurmutlulectures

DDCA (Spring 2022)

- [Spring 2022 Edition:](https://safari.ethz.ch/digitaltechnik/spring2022/doku.php?id=schedule)
  - https://safari.ethz.ch/digitaltechnik/spring2022/doku.php?id=schedule
- Spring 2021 Edition:
  - https://safari.ethz.ch/digitaltechnik/spring2021/doku.php?id=schedule
- Youtube Livestream (Spring 2022):
  - https://www.youtube.com/watch?v=cpXdE3HwvK0&list=PL5Q2soXY2Zi97Ya5DEUpMpO2bbAoaG7c6
- Youtube Livestream (Spring 2021):
  - https://www.youtube.com/watch?v=LbC0EZY8yw4&list=PL5Q2soXY2Zi_uej3aY39YB5pfW4SJ7LlN
- Bachelor's course
  - 2nd semester at ETH Zurich
  - Rigorous introduction into "How Computers Work"
  - Digital Design/Logic
  - Computer Architecture
  - 10 FPGA Lab Assignments



Digital Design and Computer Architecture - Spring 2021

Search
Recent Changes Media Manager Siten

Trace: · schedule

Home
Announcements
Materials
- Lectures/Schedule
- Lecture Buzzwords
- Readings
- Optional HWs
- Labs
- Extra Assignments
- Exams
- Technical Docs

Resources
- Computer Architecture (CMU) SS15: Lecture Videos
- Computer Architecture (CMU) SS15: Course Website
- Digitaltechnik SS18: Lecture Videos
- Digitaltechnik SS18: Course Website
- Digitaltechnik SS19: Lecture Videos
- Digitaltechnik SS19: Course Website
- Digitaltechnik SS20: Lecture Videos
- Digitaltechnik SS20: Course Website
- Moodle

Lecture Video Playlist on YouTube

Livestream Lecture Playlist

Recorded Lecture Playlist

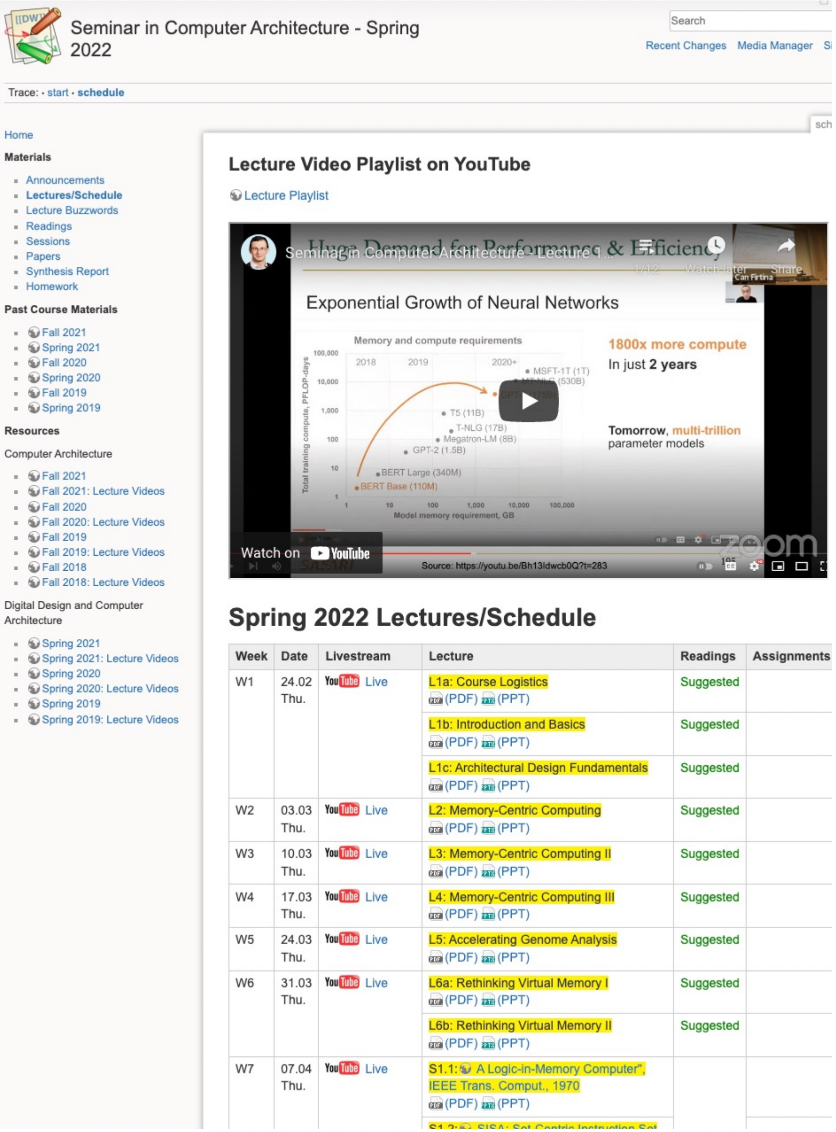
Spring 2021 Lectures/Schedule

Week	Date	Livestream	Lecture	Readings	Lab	HW
W1	25.02 Thu.	YouTube Live	L1: Introduction and Basics (PDF) (PPT)	Required Suggested Mentioned		
	26.02 Fri.	YouTube Live	L2a: Tradeoffs, Metrics, Mindset (PDF) (PPT)	Required		
			L2b: Mysteries in Computer Architecture (PDF) (PPT)	Required Mentioned		
W2	04.03 Thu.	YouTube Live	L3a: Mysteries in Computer Architecture II (PDF) (PPT)	Required Suggested Mentioned		

SAFARI https://www.youtube.com/onurmutlulectures

Seminar in Comp Arch (Spring & Fall)

- **[Spring 2022 Edition:](https://safari.ethz.ch/architecture_seminar/spring2022/doku.php?id=schedule)**
  - https://safari.ethz.ch/architecture_seminar/spring2022/doku.php?id=schedule
- **Fall 2021 Edition:**
  - https://safari.ethz.ch/architecture_seminar/fall2021/doku.php?id=schedule
- **Youtube Livestream (Spring 2022):**
  - https://www.youtube.com/watch?v=rS9UPk509AQ&list=PL5Q2soXY2Zi_hxizriwKmFHgcoe2Q8-m0
- **Youtube Livestream (Fall 2021):**
  - https://www.youtube.com/watch?v=4TcP297mdsI&list=PL5Q2soXY2Zi_7UBNmC9B8Yr5JSwTG9yH4
- Critical analysis course
  - Taken by Bachelor's/Masters/PhD students
  - Cutting-edge research topics + fundamentals in Computer Architecture
  - 20+ research papers, presentations, analyses



Seminar in Computer Architecture - Spring 2022

Trace: · start · schedule

Home

Materials

- Announcements
- **Lectures/Schedule**
- Lecture Buzzwords
- Readings
- Sessions
- Papers
- Synthesis Report
- Homework

Past Course Materials

- Fall 2021
- Spring 2021
- Fall 2020
- Spring 2020
- Fall 2019
- Spring 2019

Resources

Computer Architecture

- Fall 2021
- Fall 2021: Lecture Videos
- Fall 2020
- Fall 2020: Lecture Videos
- Fall 2019
- Fall 2019: Lecture Videos
- Fall 2018
- Fall 2018: Lecture Videos

Digital Design and Computer Architecture

- Spring 2021
- Spring 2021: Lecture Videos
- Spring 2020
- Spring 2020: Lecture Videos
- Spring 2019
- Spring 2019: Lecture Videos

Lecture Video Playlist on YouTube

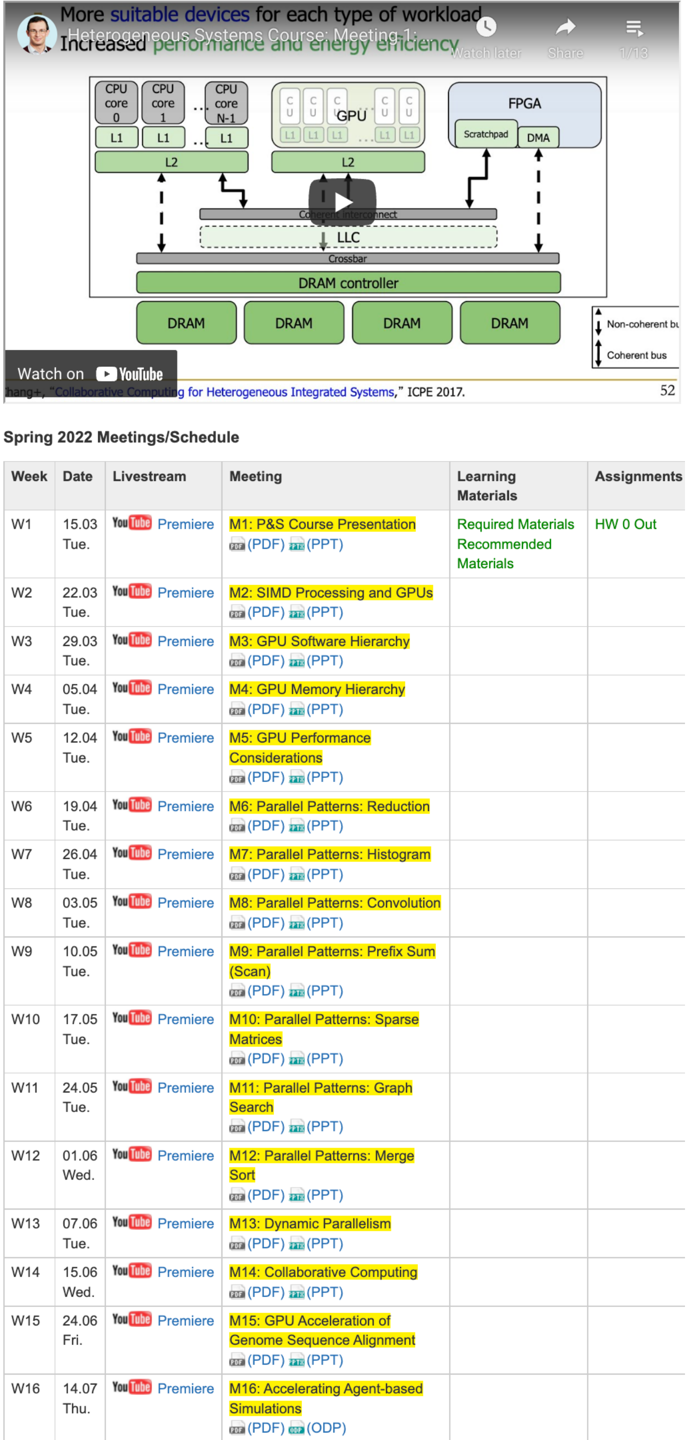
Lecture Playlist

Spring 2022 Lectures/Schedule

Week	Date	Livestream	Lecture	Readings	Assignments
W1	24.02 Thu.	YouTube Live	L1a: Course Logistics (PDF) (PPT)	Suggested	
			L1b: Introduction and Basics (PDF) (PPT)	Suggested	
			L1c: Architectural Design Fundamentals (PDF) (PPT)	Suggested	
W2	03.03 Thu.	YouTube Live	L2: Memory-Centric Computing (PDF) (PPT)	Suggested	
W3	10.03 Thu.	YouTube Live	L3: Memory-Centric Computing II (PDF) (PPT)	Suggested	
W4	17.03 Thu.	YouTube Live	L4: Memory-Centric Computing III (PDF) (PPT)	Suggested	
W5	24.03 Thu.	YouTube Live	L5: Accelerating Genome Analysis (PDF) (PPT)	Suggested	
W6	31.03 Thu.	YouTube Live	L6a: Rethinking Virtual Memory I (PDF) (PPT)	Suggested	
			L6b: Rethinking Virtual Memory II (PDF) (PPT)	Suggested	
W7	07.04 Thu.	YouTube Live	S1.1: A Logic-in-Memory Computer", IEEE Trans. Comput., 1970 (PDF) (PPT)		
			S1.2: SISA: Set-Centric Instruction Set		

Hetero. Systems (Spring'22)

- **[Spring 2022 Edition:](https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=heterogeneous_systems)**
  - https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=heterogeneous_systems
- **[Youtube Livestream:](https://www.youtube.com/watch?v=oFO5fTrgFIY&list=PL5Q2soXY2Zi9XrgXR38IM_FTjmY6h7Gzm)**
  - https://www.youtube.com/watch?v=oFO5fTrgFIY&list=PL5Q2soXY2Zi9XrgXR38IM_FTjmY6h7Gzm
- Project course
  - Taken by Bachelor's/Master's students
  - GPU and Parallelism lectures
  - Hands-on research exploration
  - Many research readings

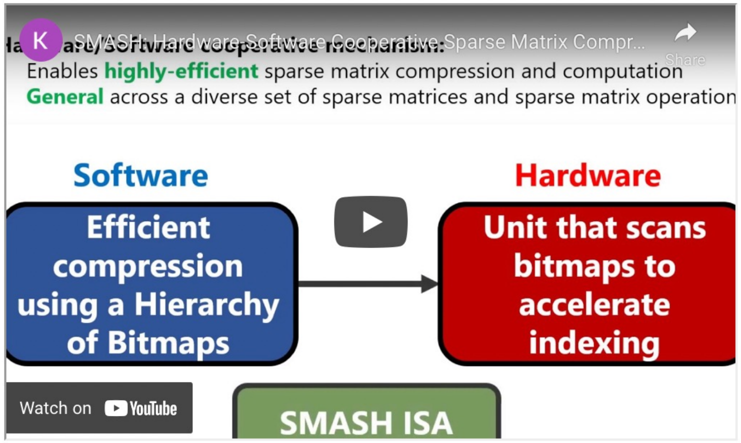


Spring 2022 Meetings/Schedule

Week	Date	Livestream	Meeting	Learning Materials	Assignments
W1	15.03 Tue.	YouTube Premiere	M1: P&S Course Presentation (PDF) (PPT)	Required Materials Recommended Materials	HW 0 Out
W2	22.03 Tue.	YouTube Premiere	M2: SIMD Processing and GPUs (PDF) (PPT)		
W3	29.03 Tue.	YouTube Premiere	M3: GPU Software Hierarchy (PDF) (PPT)		
W4	05.04 Tue.	YouTube Premiere	M4: GPU Memory Hierarchy (PDF) (PPT)		
W5	12.04 Tue.	YouTube Premiere	M5: GPU Performance Considerations (PDF) (PPT)		
W6	19.04 Tue.	YouTube Premiere	M6: Parallel Patterns: Reduction (PDF) (PPT)		
W7	26.04 Tue.	YouTube Premiere	M7: Parallel Patterns: Histogram (PDF) (PPT)		
W8	03.05 Tue.	YouTube Premiere	M8: Parallel Patterns: Convolution (PDF) (PPT)		
W9	10.05 Tue.	YouTube Premiere	M9: Parallel Patterns: Prefix Sum (Scan) (PDF) (PPT)		
W10	17.05 Tue.	YouTube Premiere	M10: Parallel Patterns: Sparse Matrices (PDF) (PPT)		
W11	24.05 Tue.	YouTube Premiere	M11: Parallel Patterns: Graph Search (PDF) (PPT)		
W12	01.06 Wed.	YouTube Premiere	M12: Parallel Patterns: Merge Sort (PDF) (PPT)		
W13	07.06 Tue.	YouTube Premiere	M13: Dynamic Parallelism (PDF) (PPT)		
W14	15.06 Wed.	YouTube Premiere	M14: Collaborative Computing (PDF) (PPT)		
W15	24.06 Fri.	YouTube Premiere	M15: GPU Acceleration of Genome Sequence Alignment (PDF) (PPT)		
W16	14.07 Thu.	YouTube Premiere	M16: Accelerating Agent-based Simulations (PDF) (ODP)		

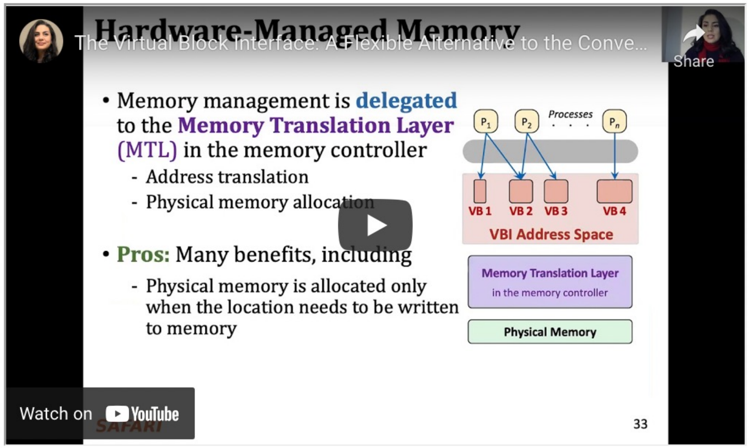
HW/SW Co-Design (Spring 2022)

- **Spring 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=hw_sw_codesign
- **Youtube Livestream:**
  - https://youtube.com/playlist?list=PL5Q2soXY2Zi8nH7un3ghD2nutKWWDk-NK
- Project course
  - Taken by Bachelor's/Master's students
  - HW/SW co-design lectures
  - Hands-on research exploration
  - Many research readings


SMASH: Hardware-Software Cooperative Sparse Matrix Compr...

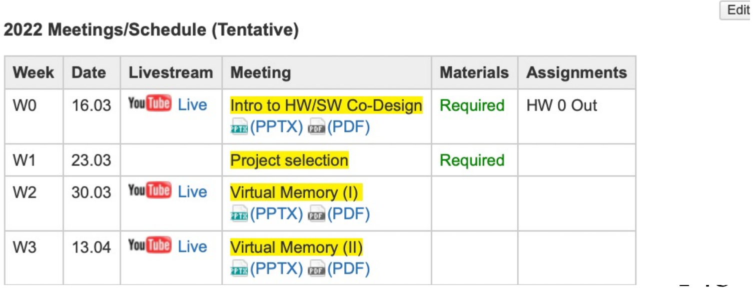
Enables **highly-efficient** sparse matrix compression and computation

General across a diverse set of sparse matrices and sparse matrix operation


The Virtual Block Interface: A Flexible Alternative to the Conve...

Hardware-Managed Memory

- Memory management is **delegated** to the **Memory Translation Layer** (MTL) in the memory controller
  - Address translation
  - Physical memory allocation
- **Pros:** Many benefits, including
  - Physical memory is allocated only when the location needs to be written to memory



2022 Meetings/Schedule (Tentative)

Week	Date	Livestream	Meeting	Materials	Assignments
W0	16.03	YouTube Live	Intro to HW/SW Co-Design (PPTX) (PDF)	Required	HW 0 Out
W1	23.03		Project selection	Required	
W2	30.03	YouTube Live	Virtual Memory (I) (PPTX) (PDF)		
W3	13.04	YouTube Live	Virtual Memory (II) (PPTX) (PDF)		

Funding Acknowledgments

- Alibaba, AMD, ASML, Google, Facebook, Hi-Silicon, HP Labs, Huawei, IBM, Intel, Microsoft, Nvidia, Oracle, Qualcomm, Rambus, Samsung, Seagate, VMware, Xilinx
- NSF
- NIH
- GSRC
- SRC
- CyLab
- EFCL

Acknowledgments

SAFARI

SAFARI Research Group

safari.ethz.ch

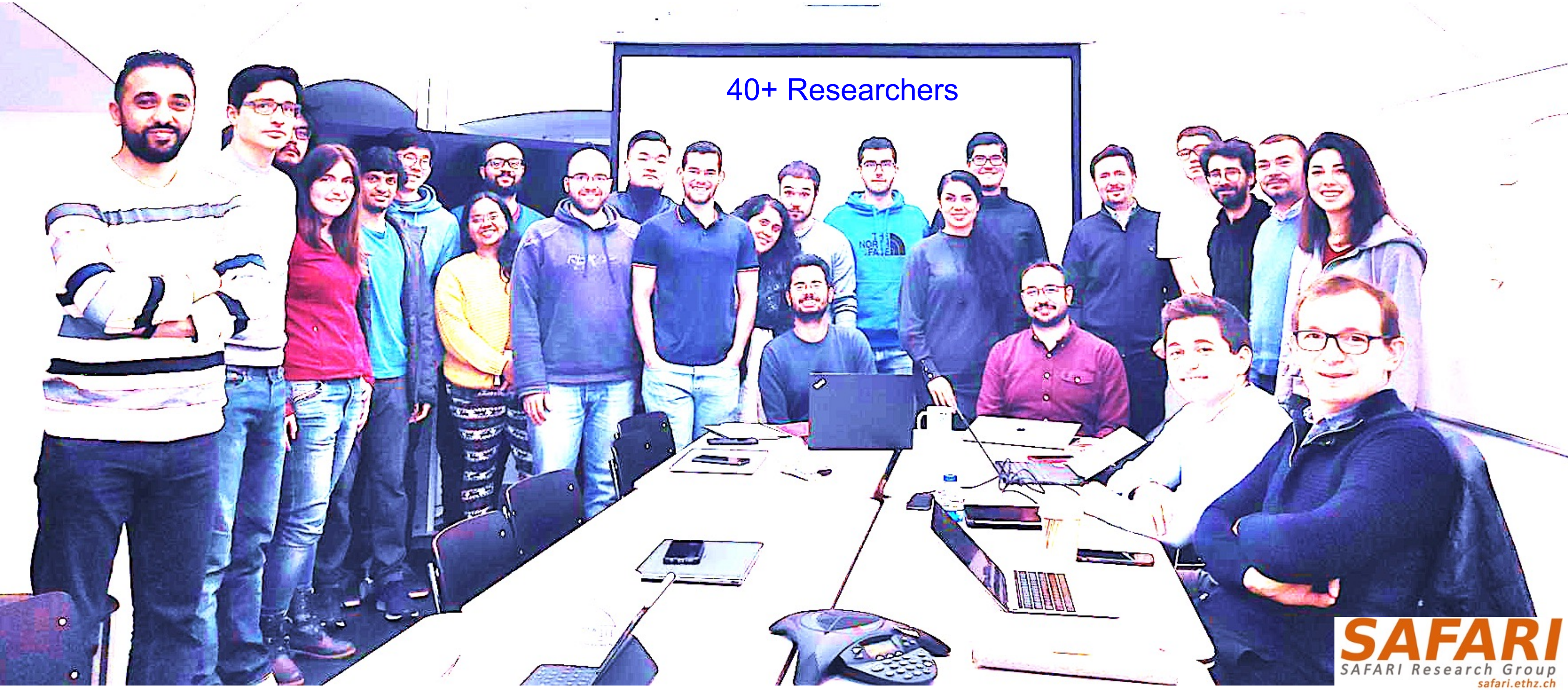
Think BIG, Aim HIGH!

https://safari.ethz.ch

Onur Mutlu's SAFARI Research Group

Computer architecture, HW/SW, systems, bioinformatics, security, memory

https://safari.ethz.ch/safari-newsletter-january-2021/

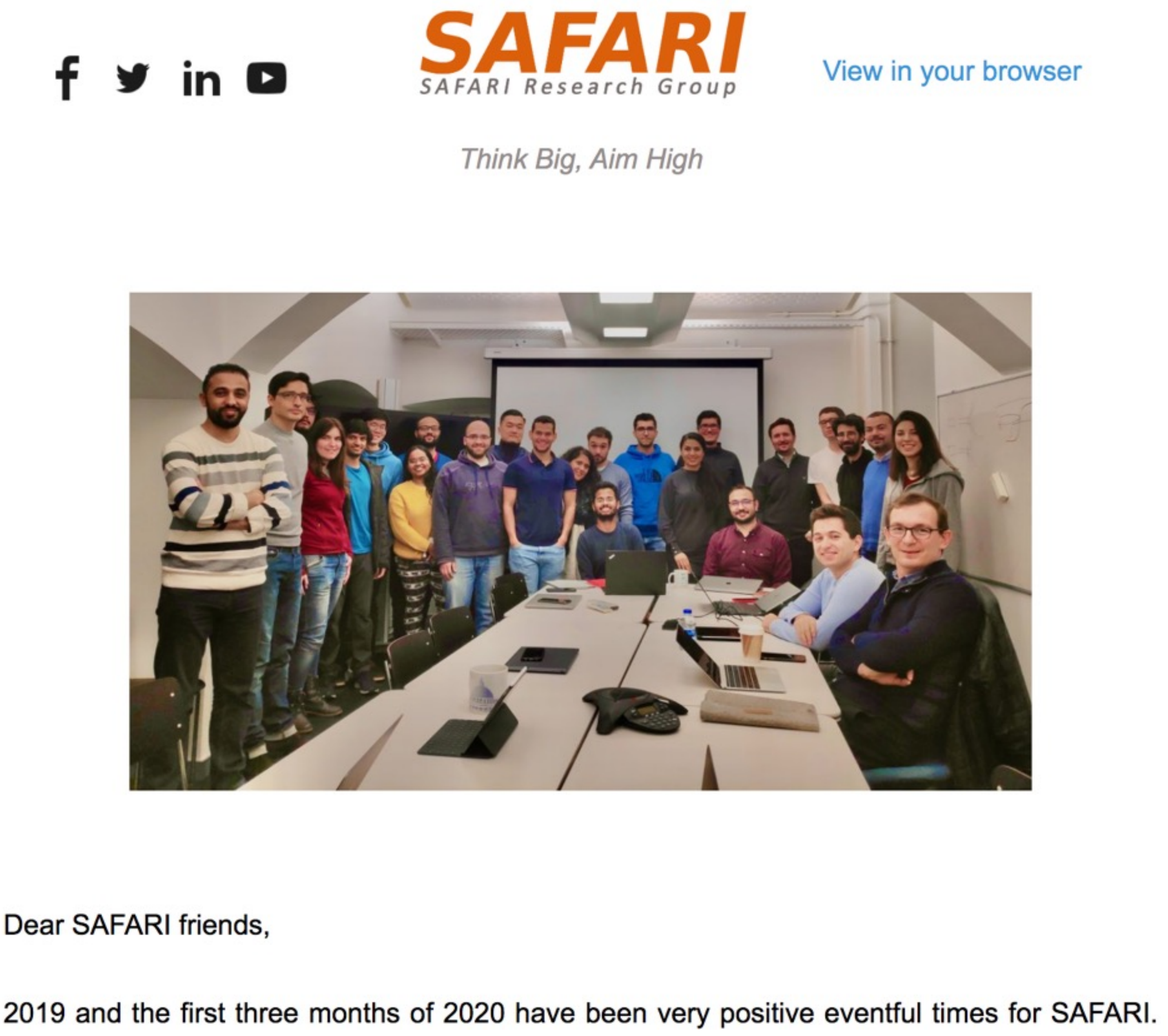


Think BIG, Aim HIGH!

https://safari.ethz.ch

SAFARI Newsletter April 2020 Edition

- https://safari.ethz.ch/safari-newsletter-april-2020/



SAFARI
SAFARI Research Group

View in your browser

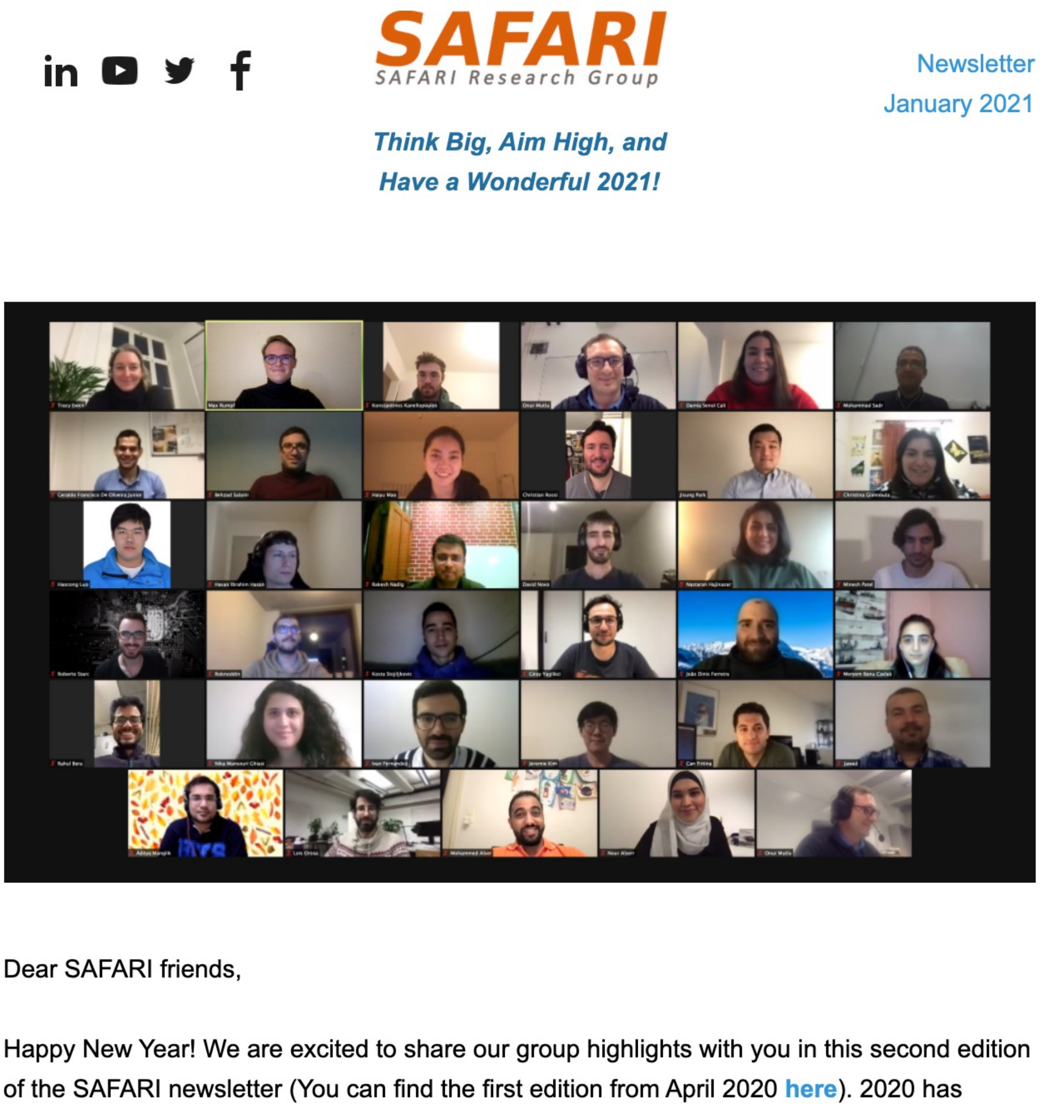
Think Big, Aim High

Dear SAFARI friends,

2019 and the first three months of 2020 have been very positive eventful times for SAFARI.

SAFARI Newsletter January 2021 Edition

- https://safari.ethz.ch/safari-newsletter-january-2021/



SAFARI

SAFARI Research Group

Newsletter
January 2021

Think Big, Aim High, and
Have a Wonderful 2021!

Dear SAFARI friends,

Happy New Year! We are excited to share our group highlights with you in this second edition of the SAFARI newsletter (You can find the first edition from April 2020 **here**). 2020 has

SAFARI Newsletter December 2021 Edition

- https://safari.ethz.ch/safari-newsletter-december-2021/

Think Big, Aim High

View in your browser

December 2021

Referenced Papers, Talks, Artifacts

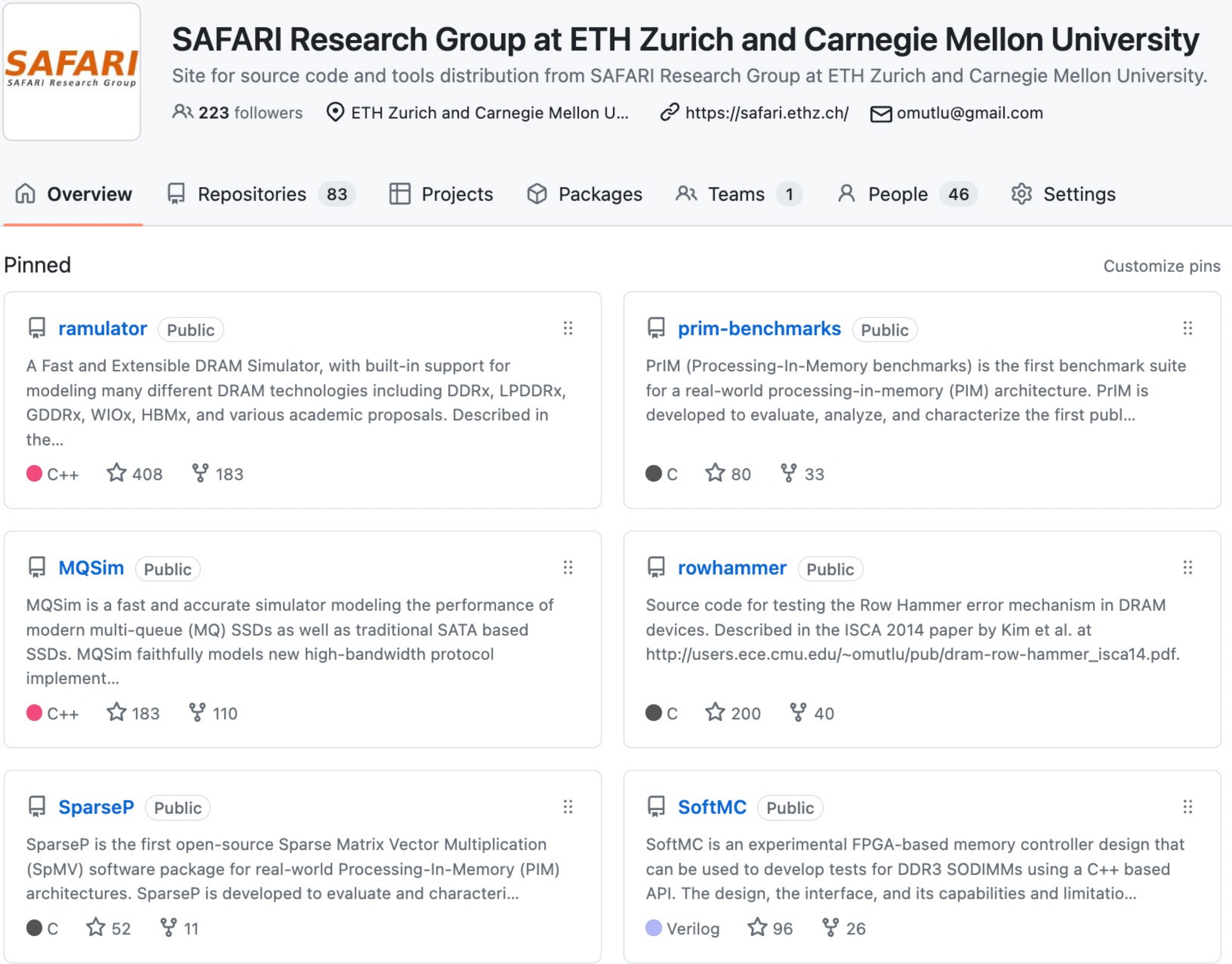
- All are available at

  https://people.inf.ethz.ch/omutlu/projects.htm

  https://www.youtube.com/onurmutlulectures

  https://github.com/CMU-SAFARI/

Open Source Tools: SAFARI GitHub



SAFARI Research Group at ETH Zurich and Carnegie Mellon University

Site for source code and tools distribution from SAFARI Research Group at ETH Zurich and Carnegie Mellon University.

223 followers · ETH Zurich and Carnegie Mellon U... · https://safari.ethz.ch/ · omutlu@gmail.com

ramulator Public

A Fast and Extensible DRAM Simulator, with built-in support for modeling many different DRAM technologies including DDRx, LPDDRx, GDDRx, WIOx, HBMx, and various academic proposals. Described in the...

C++ ☆ 408 183

prim-benchmarks Public

PrIM (Processing-In-Memory benchmarks) is the first benchmark suite for a real-world processing-in-memory (PIM) architecture. PrIM is developed to evaluate, analyze, and characterize the first publ...

C ☆ 80 33

MQSim Public

MQSim is a fast and accurate simulator modeling the performance of modern multi-queue (MQ) SSDs as well as traditional SATA based SSDs. MQSim faithfully models new high-bandwidth protocol implement...

C++ ☆ 183 110

rowhammer Public

Source code for testing the Row Hammer error mechanism in DRAM devices. Described in the ISCA 2014 paper by Kim et al. at http://users.ece.cmu.edu/~omutlu/pub/dram-row-hammer_isca14.pdf.

C ☆ 200 40

SparseP Public

SparseP is the first open-source Sparse Matrix Vector Multiplication (SpMV) software package for real-world Processing-In-Memory (PIM) architectures. SparseP is developed to evaluate and characteri...

C ☆ 52 11

SoftMC Public

SoftMC is an experimental FPGA-based memory controller design that can be used to develop tests for DDR3 SODIMMs using a C++ based API. The design, the interface, and its capabilities and limitatio...

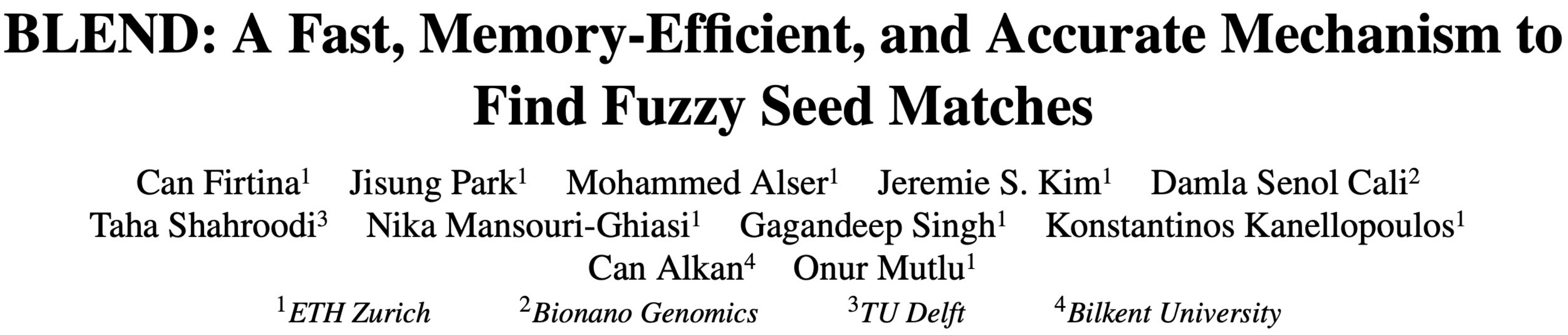
Verilog ☆ 96 26

https://github.com/CMU-SAFARI/

Some Other Recent Papers

Finding Approximate Seed Matches

- Can Firtina, Jisung Park, Mohammed Alser, Jeremie S. Kim, Damla Senol Cali, Taha Shahroodi, Nika Mansouri-Ghiasi, Gagandeep Singh, Konstantinos Kanellopoulos, Can Alkan, and Onur Mutlu,
  "BLEND: A Fast, Memory-Efficient, and Accurate Mechanism to Find Fuzzy Seed Matches"
  Preprint in *arXiv*, 2021.
  [arXiv preprint]
  [BLEND Source Code and Data]

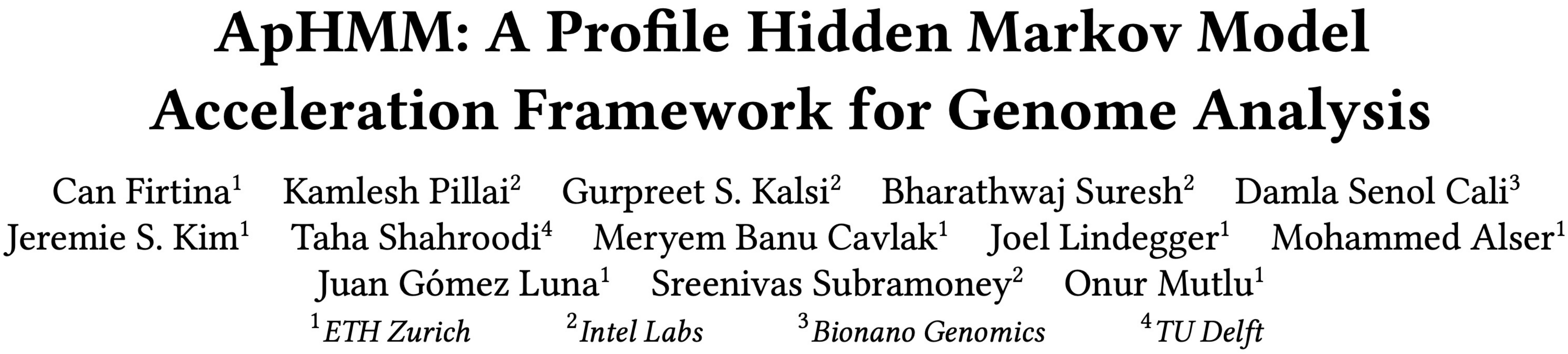


BLEND: A Fast, Memory-Efficient, and Accurate Mechanism to Find Fuzzy Seed Matches

Can Firtina[1] Jisung Park[1] Mohammed Alser[1] Jeremie S. Kim[1] Damla Senol Cali[2]
Taha Shahroodi[3] Nika Mansouri-Ghiasi[1] Gagandeep Singh[1] Konstantinos Kanellopoulos[1]
Can Alkan[4] Onur Mutlu[1]
[1]*ETH Zurich* [2]*Bionano Genomics* [3]*TU Delft* [4]*Bilkent University*

Hardware Acceleration for pHMMs

- Can Firtina, Kamlesh Pillai, Gurpreet S. Kalsi, Bharathwaj Suresh, Damla Senol Cali, Jeremie S. Kim, Taha Shahroodi, Meryem Banu Cavlak, Joel Lindegger, Mohammed Alser, Juan Gómez-Luna, Sreenivas Subramoney, and Onur Mutlu,
  "ApHMM: A Profile Hidden Markov Model Acceleration Framework for Genome Analysis"
  Preprint in *arXiv*, 2022.
  [Source Code]

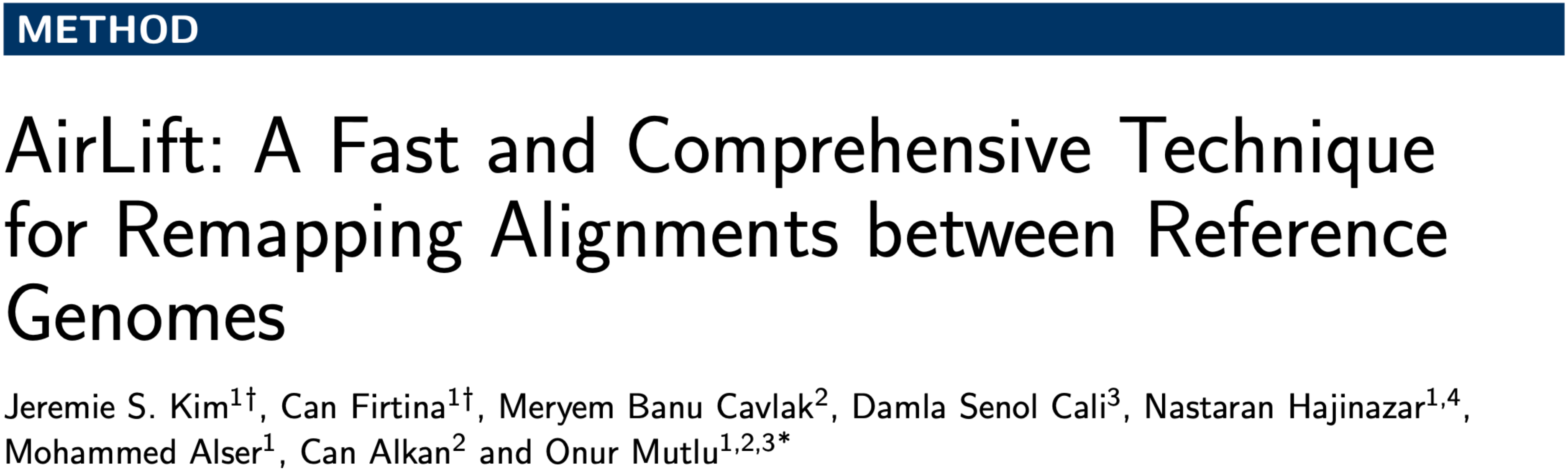


ApHMM: A Profile Hidden Markov Model Acceleration Framework for Genome Analysis

Can Firtina[1] Kamlesh Pillai[2] Gurpreet S. Kalsi[2] Bharathwaj Suresh[2] Damla Senol Cali[3]
Jeremie S. Kim[1] Taha Shahroodi[4] Meryem Banu Cavlak[1] Joel Lindegger[1] Mohammed Alser[1]
Juan Gómez Luna[1] Sreenivas Subramoney[2] Onur Mutlu[1]
[1] *ETH Zurich* [2] *Intel Labs* [3] *Bionano Genomics* [4] *TU Delft*

Remapping Reads Between References

- Jeremie S. Kim, Can Firtina, Meryem Banu Cavlak, Damla Senol Cali, Nastaran Hajinazar, Mohammed Alser, Can Alkan, and Onur Mutlu,
  "AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes"
  Preprint in arXiv and bioRxiv, 2021.
  [bioRxiv preprint]
  [arXiv preprint]
  [AirLift Source Code and Data]



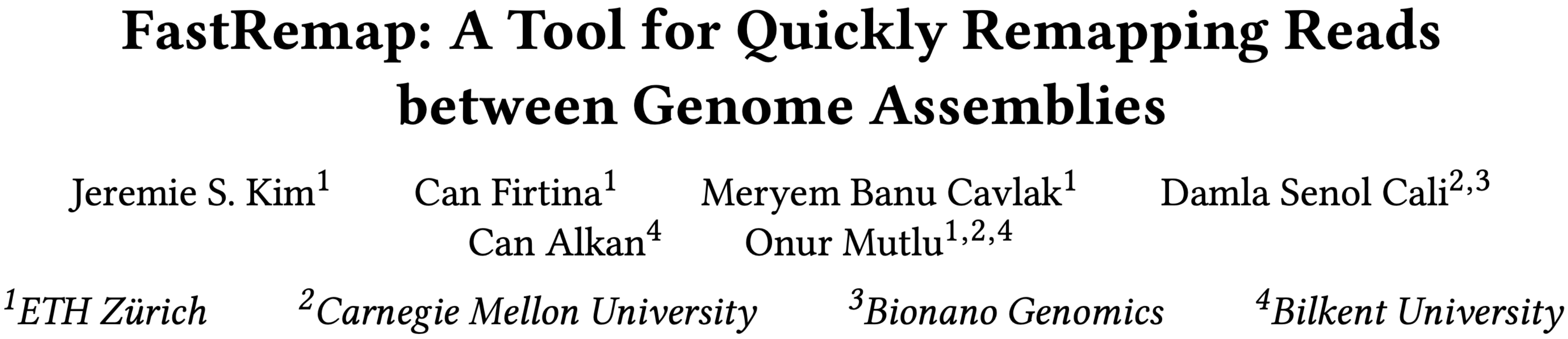
METHOD

AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes

Jeremie S. Kim[1†], Can Firtina[1†], Meryem Banu Cavlak[2], Damla Senol Cali[3], Nastaran Hajinazar[1,4], Mohammed Alser[1], Can Alkan[2] and Onur Mutlu[1,2,3*]

Mapping Constant Regions Between References

- Jeremie S. Kim, Can Firtina, Meryem Banu Cavlak, Damla Senol Cali, Can Alkan, and Onur Mutlu,
  "FastRemap: A Tool for Quickly Remapping Reads between Genome Assemblies"
  Bioinformatics, btac554.
  [FastRemap Source Code]



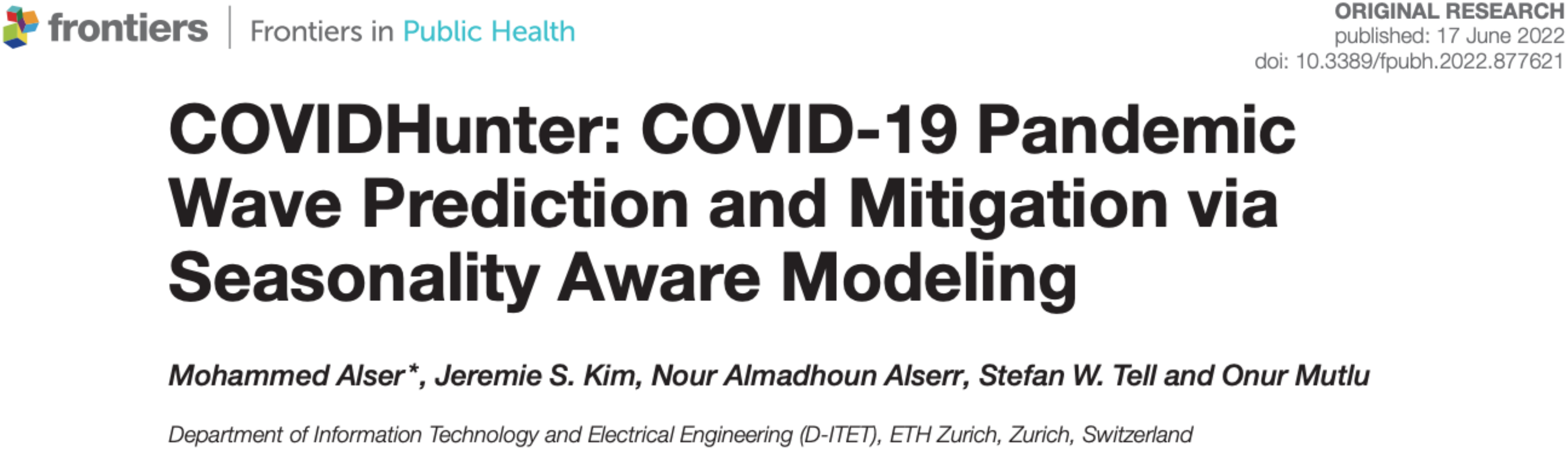
FastRemap: A Tool for Quickly Remapping Reads between Genome Assemblies

Jeremie S. Kim[1] Can Firtina[1] Meryem Banu Cavlak[1] Damla Senol Cali[2,3]
Can Alkan[4] Onur Mutlu[1,2,4]

[1] *ETH Zürich* [2] *Carnegie Mellon University* [3] *Bionano Genomics* [4] *Bilkent University*

COVIDHunter

Mohammed Alser, Jeremie S. Kim, Nour Almadhoun Alserr, Stefan W. Tell, Onur Mutlu
"COVIDHunter: COVID-19 Pandemic Wave Prediction and Mitigation via Seasonality Aware Modeling"
Frontiers in Public Health 2022
[Source Code]



frontiers | Frontiers in Public Health

ORIGINAL RESEARCH
published: 17 June 2022
doi: 10.3389/fpubh.2022.877621

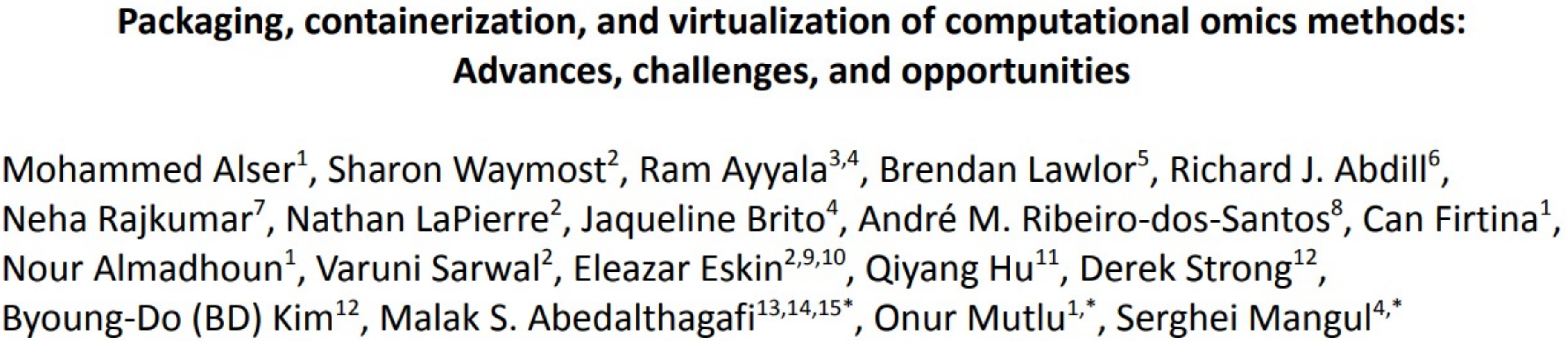
COVIDHunter: COVID-19 Pandemic Wave Prediction and Mitigation via Seasonality Aware Modeling

Mohammed Alser, Jeremie S. Kim, Nour Almadhoun Alserr, Stefan W. Tell and Onur Mutlu*

Department of Information Technology and Electrical Engineering (D-ITET), ETH Zurich, Zurich, Switzerland

Packaging Omics Methods

Mohammed Alser, Sharon Waymost, Ram Ayyala, Brendan Lawlor, Richard J. Abdill, Neha Rajkumar, Nathan LaPierre, Jaqueline Brito, Andre M. Ribeiro-dos-Santos, Can Firtina, Nour Almadhoun, Varuni Sarwal, Eleazar Eskin, Qiyang Hu, Derek Strong, Byoung-Do (BD)Kim, Malak S. Abedalthagafi, Onur Mutlu, Serghei Mangul
"Packaging, containerization, and virtualization of computational omics methods: Advances, challenges, and opportunities"
arrXiv 2022

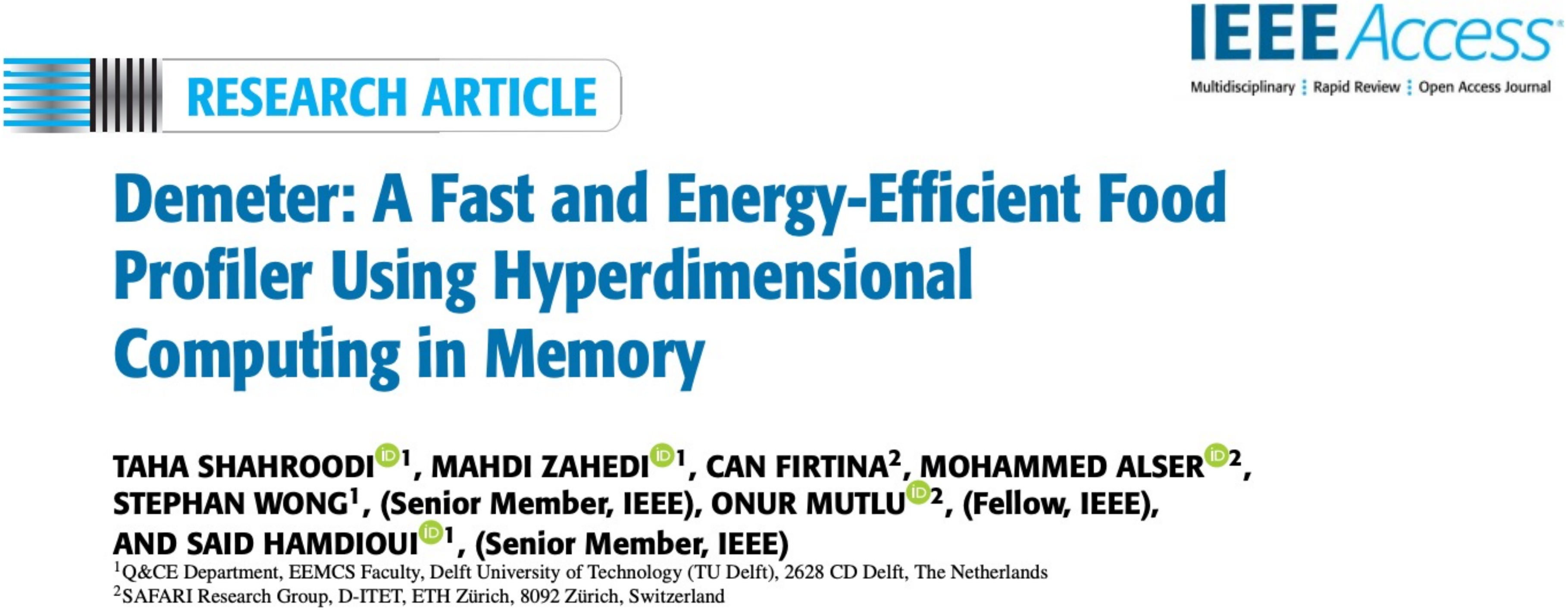


Packaging, containerization, and virtualization of computational omics methods: Advances, challenges, and opportunities

Mohammed Alser[1], Sharon Waymost[2], Ram Ayyala[3,4], Brendan Lawlor[5], Richard J. Abdill[6], Neha Rajkumar[7], Nathan LaPierre[2], Jaqueline Brito[4], André M. Ribeiro-dos-Santos[8], Can Firtina[1], Nour Almadhoun[1], Varuni Sarwal[2], Eleazar Eskin[2,9,10], Qiyang Hu[11], Derek Strong[12], Byoung-Do (BD) Kim[12], Malak S. Abedalthagafi[13,14,15*], Onur Mutlu[1,*], Serghei Mangul[4,*]

Demeter (HD Food Microbiome Profiling)

Taha Shahroodi, Mahdi Zahedi, Can Firtina, Mohammed Alser, Stephan Wong, Onur Mutlu, Said Hamdioui
"Demeter: A Fast and Energy-Efficient Food Profiler using Hyperdimensional Computing in Memory"
IEEE Access, 2022



IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

RESEARCH ARTICLE

Demeter: A Fast and Energy-Efficient Food Profiler Using Hyperdimensional Computing in Memory

TAHA SHAHROODI[1], MAHDI ZAHEDI[1], CAN FIRTINA[2], MOHAMMED ALSER[2], STEPHAN WONG[1], (Senior Member, IEEE), ONUR MUTLU[2], (Fellow, IEEE), AND SAID HAMDIOUI[1], (Senior Member, IEEE)
[1]Q&CE Department, EEMCS Faculty, Delft University of Technology (TU Delft), 2628 CD Delft, The Netherlands
[2]SAFARI Research Group, D-ITET, ETH Zürich, 8092 Zürich, Switzerland

End of Backup Slides