

PIM Architectures for Bioinformatics

Dr. Konstantina Koliogeorgi

kkoliogeorgi@ethz.ch

<https://people.inf.ethz.ch/>

ICS 2025

08 June 2025

SAFARI

ETH zürich

Brief Self Introduction

- Konstantina Koliogeorgi

- Senior Researcher and Lecturer @ SAFARI
- PhD, National Technical University of Athens, 2023
- kkoliogeorgi@safari.ethz.ch



- Research & Teaching Areas

- Hardware/Software Co-Design
- Heterogeneous System Architecture
- Reconfigurable Computing and Architectures
- Hardware Acceleration
- Optimized Architectures for Genome analysis
- High Level Synthesis Tools
- Design Space Exploration

Agenda

- Brief Introduction to Genomics
- Data Movement Bottlenecks during analysis
- Designing algorithms and architectures that tackle data movement overhead
 - Target Multiple Steps of Pipeline
 - Leverage Processing-In-Memory
 - Leverage In-Storage Processing

Faster, Scalable & Accurate Genome Analysis



Uncovering and treating diseases
linked to genomic variations



Altering genomes to solve
fundamental challenges of life

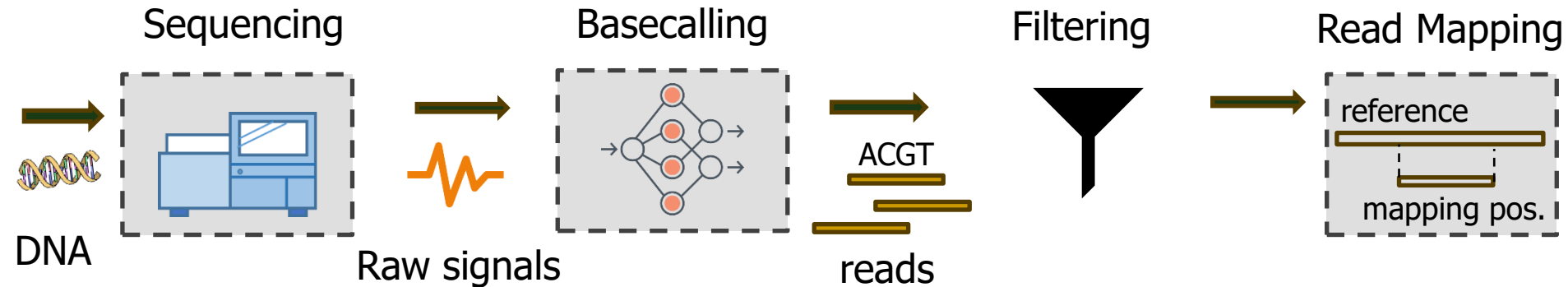


Detecting **pathogens**
in the environment



Rapid surveillance of
disease outbreaks

Typical Genome Sequence Analysis



Sequencing

extract small fragments of the original DNA sequence

Basecalling

convert raw signals to DNA bases
Deep Neural Networks

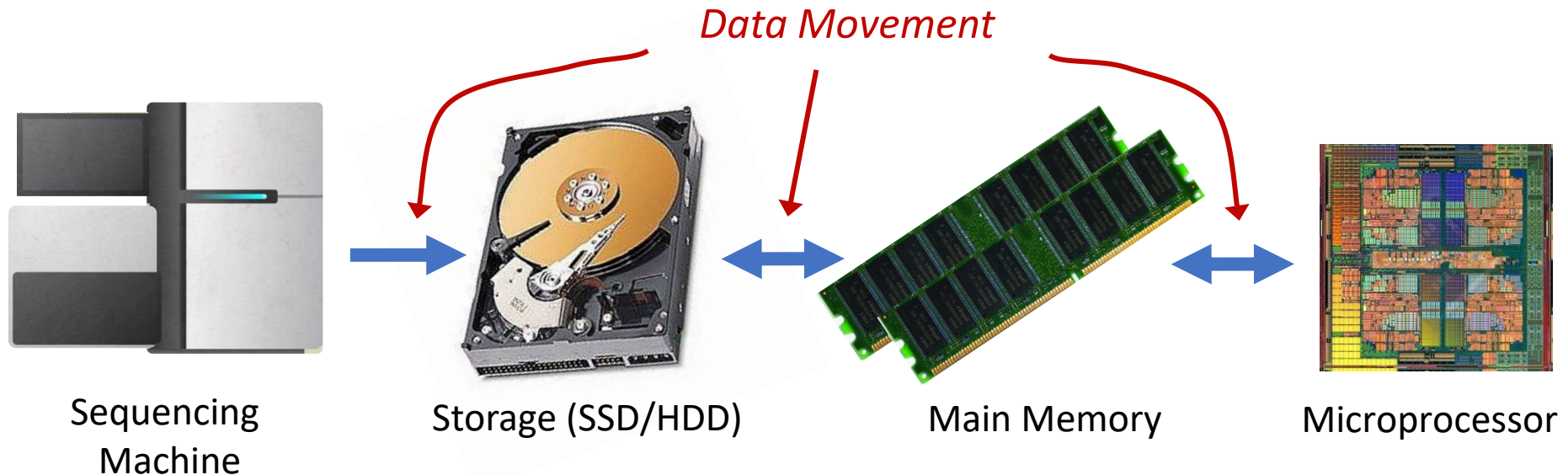
Read mapping

aligns reads to potential matching locations in the reference genome

Significant **barrier**
to **genome analyses**

Data Movement Dominates Performance

- **Data movement** dominates performance and is a **major** system **energy bottleneck** (accounting for 40%-62%)



Single **memory** request **consumes >160× - 800× more energy** compared to performing an **addition operation**

- Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS 2018
- Kestor et al., "Quantifying the Energy Cost of Data Movement in Scientific Applications," IISWC 2013
- Pandiyan and Wu, "Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms," IISWC 2014

Data analysis
is performed
far away from the data

We need to orchestrate
algorithms and architectures
to handle data well

Genomic Analysis Steps in Memory

We need to design algorithms
that fit processing-in-memory

Processing Using Memory



The image shows a YouTube video player interface. The video content displays a title slide for 'Computer Architecture' with the subtitle 'Lecture 6a: RowHammer & Secure and Reliable Memory II'. The speaker is identified as Prof. Onur Mutlu from ETH Zürich, dated Fall 2021, with a specific date of 15 October 2021. The video player includes standard controls like play, volume, and a progress bar showing 1:51 / 2:47:55. Below the video, the channel name 'Onur Mutlu Lectures' is shown with 19.6K subscribers and a 'SUBSCRIBED' button. The video title is 'Computer Architecture - Lecture 6: Processing using Memory (Fall 2021)' and it has 721 views, streamed live on Oct 15, 2021. Engagement icons for likes (26), comments (0), and share are visible.

Computer Architecture
Lecture 6a: RowHammer &
Secure and Reliable Memory II

Prof. Onur Mutlu
ETH Zürich
Fall 2021
15 October 2021

1:51 / 2:47:55

ETH ZÜRICH D-ITET

Computer Architecture - Lecture 6: Processing using Memory (Fall 2021)

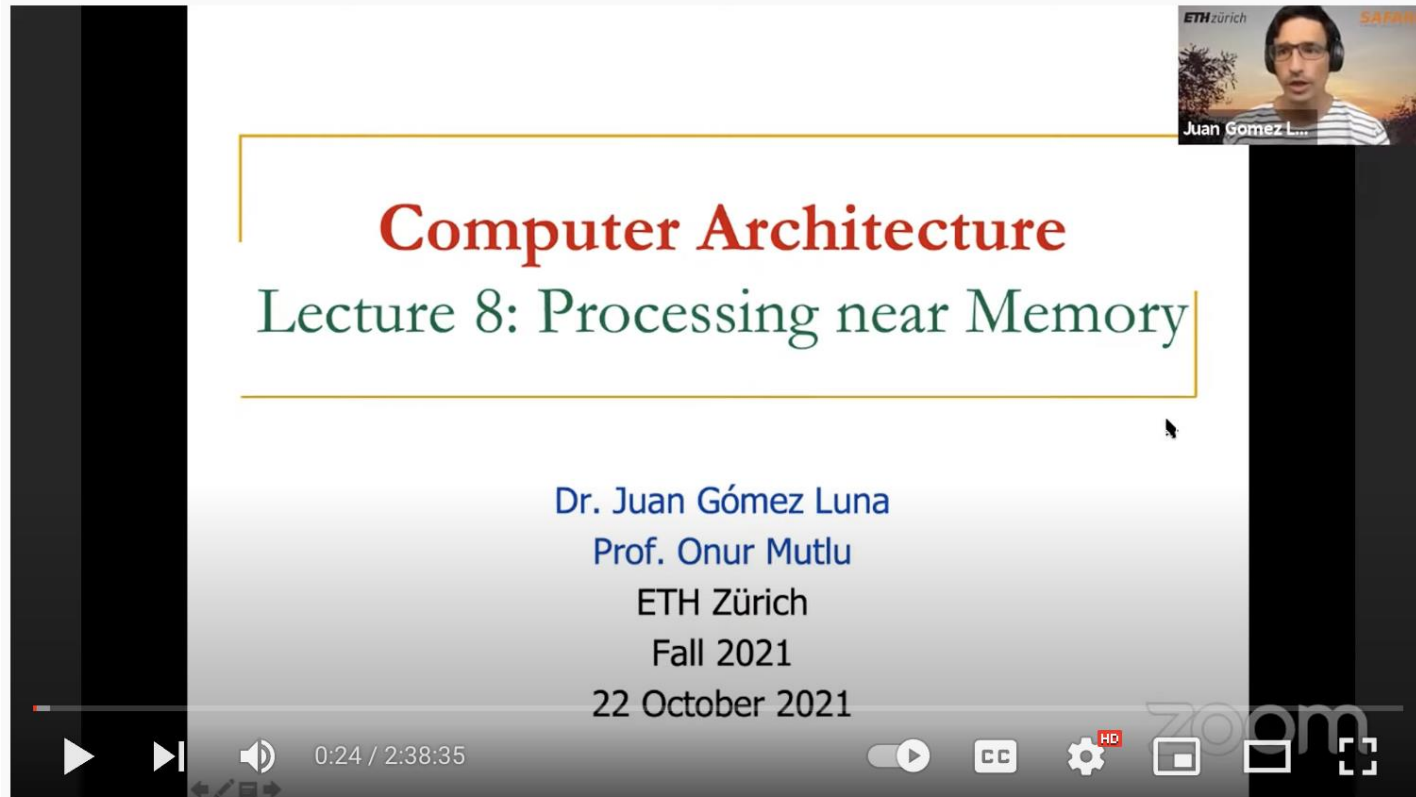
721 views • Streamed live on Oct 15, 2021

26 0 SHARE SAVE ...

Onur Mutlu Lectures
19.6K subscribers

SUBSCRIBED

Processing Near Memory



Computer Architecture
Lecture 8: Processing near Memory

Dr. Juan Gómez Luna
Prof. Onur Mutlu
ETH Zürich
Fall 2021
22 October 2021

0:24 / 2:38:35

zoom

Computer Architecture - Lecture 8: Processing near Memory (Fall 2021)

759 views • Streamed live on Oct 22, 2021

👍 33 💬 0 ➦ SHARE ⚙️ SAVE ...

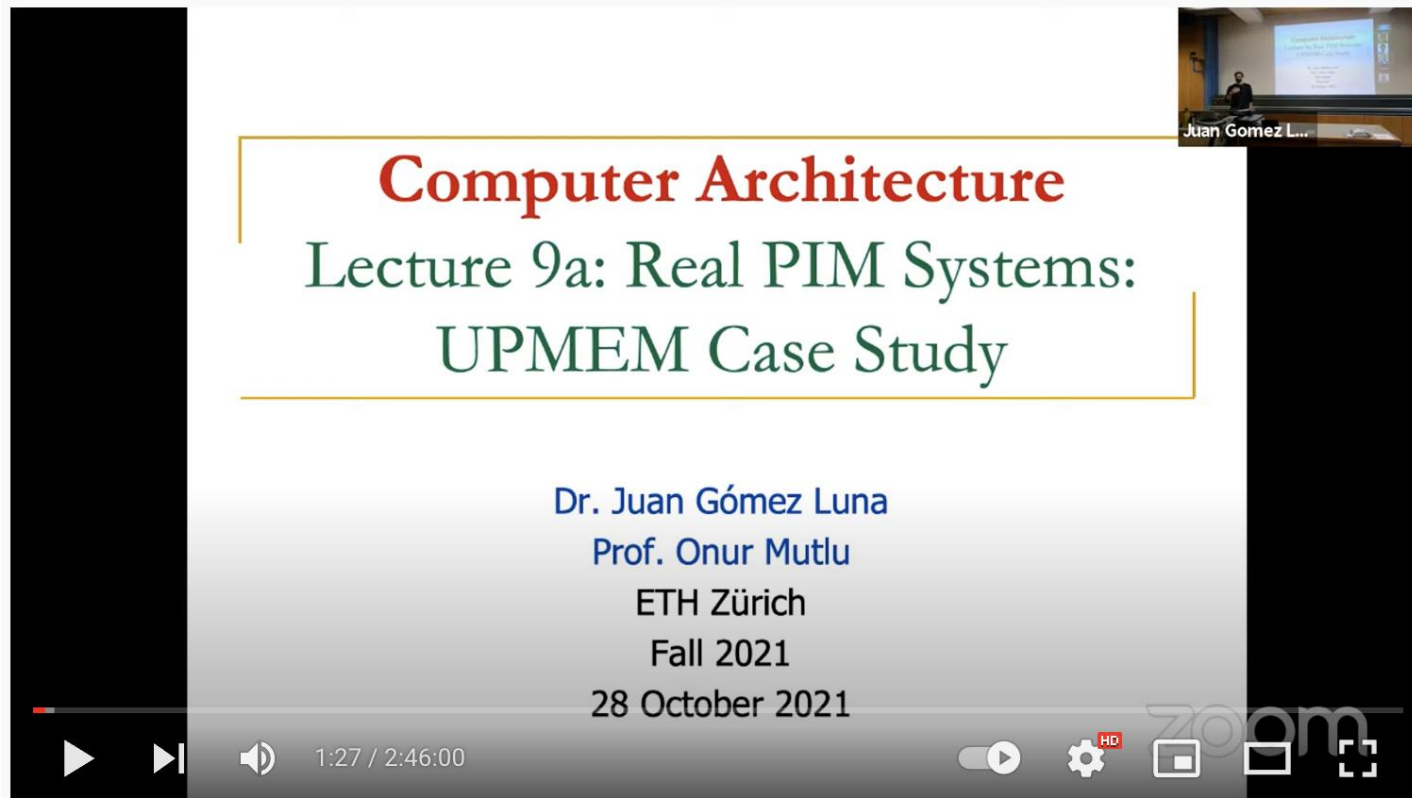


Onur Mutlu Lectures
19.6K subscribers

SUBSCRIBED



Using Real PIM System



The video player shows a lecture slide with the following text:

Computer Architecture
Lecture 9a: Real PIM Systems:
UPMEM Case Study

Dr. Juan Gómez Luna
Prof. Onur Mutlu
ETH Zürich
Fall 2021
28 October 2021

The video player controls at the bottom show a progress bar at 1:27 / 2:46:00, a play button, a volume icon, and a 'zoom' watermark. A small inset in the top right corner shows a person in a lecture hall.

Computer Architecture - Lecture 9: Real PIM Systems: UPMEM Case Study (Fall 2021)

137 views • Streamed live 5 hours ago

👍 11 💬 0 ➦ SHARE ⚙️ SAVE ...



Onur Mutlu Lectures
19.6K subscribers

SUBSCRIBED



Raw Signal Translation using PIM [MICRO '23]

- Taha Shahroodi, Gagandeep Singh, Mahdi Zahedi, Haiyu Mao, Joel Lindegger, Can Firtina, Stephan Wong, Onur Mutlu, and Said Hamdioui,
"Swordfish: A Framework for Evaluating Deep Neural Network-based Basecalling using Computation-In-Memory with Non-Ideal Memristors"
Proceedings of the 56th International Symposium on Microarchitecture (MICRO), Toronto, ON, Canada, November 2023.
[\[Slides \(pptx\) \(pdf\)\]](#)
[\[arXiv version\]](#)

Swordfish: A Framework for Evaluating Deep Neural Network-based Basecalling using Computation-In-Memory with Non-Ideal Memristors

Taha Shahroodi¹ Gagandeep Singh^{2,3} Mahdi Zahedi¹ Haiyu Mao³ Joel Lindegger³ Can Firtina³
Stephan Wong¹ Onur Mutlu³ Said Hamdioui¹

¹TU Delft ²AMD Research ³ETH Zürich

Using PIM for filtering

- 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.
[arxiv.org Version \(pdf\)](#)

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

Jeremie S. Kim^{1,6*}, Damla Senol Cali¹, Hongyi Xin², Donghyuk Lee³, Saugata Ghose¹, Mohammed Alser⁴, Hasan Hassan⁶, Oguz Ergin⁵, Can Alkan^{4*} and Onur Mutlu^{6,1*}

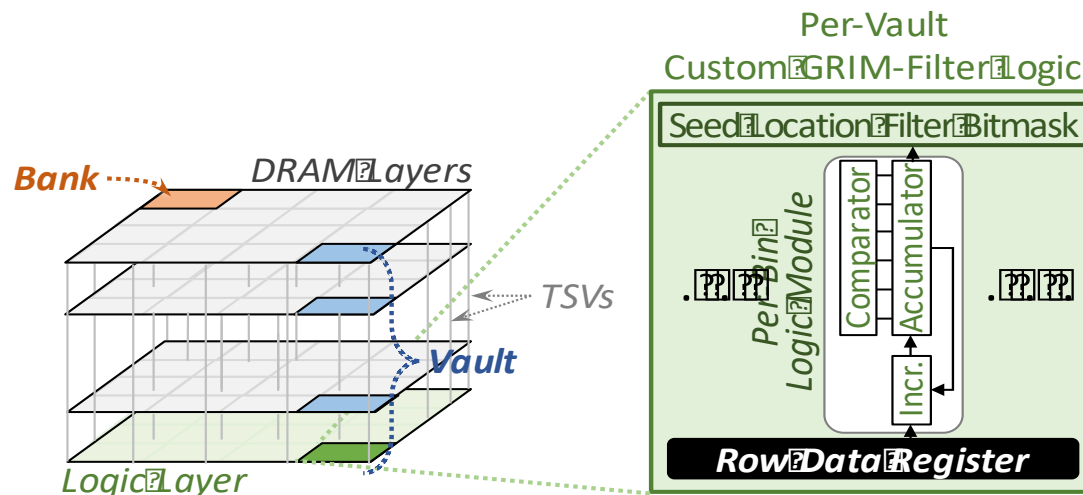
From The Sixteenth Asia Pacific Bioinformatics Conference 2018
Yokohama, Japan. 15-17 January 2018

GRIM-Filter in 3D-Stacked DRAM

1. Highly Parallel mechanism

2. Memory Bound: Given the frequent accesses to memory, we find that GRIM-Filter is memory bound

These properties together make GRIM-Filter a good algorithm to be run in 3D-Stacked DRAM



AIM (PIM Sequence Alignment Framework)

Safaa Diab, Amir Nassereldine, Mohammed Alser, Juan Gómez-Luna,
Onur Mutlu, Izzat El Hajj

[“A Framework for High-throughput Sequence Alignment using Real Processing-in-Memory Systems”](#)

arXiv, 2022

[\[Source code\]](#)

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

Safaa Diab¹, Amir Nassereldine¹, Mohammed Alser², Juan Gómez Luna², Onur Mutlu², Izzat El Hajj¹

¹*American University of Beirut, Lebanon* ²*ETH Zürich, Switzerland*

Accelerating Sequence-to-Graph Mapping

- Damla Senol Cali, Konstantinos Kanellopoulos, Joel Lindegger, Zülal Bingöl, Gurpreet S. Kalsi, Ziyi Zuo, Can Firtina, Meryem Banu Cavlak, Jeremie Kim, Nika Mansouri Ghiasi, 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 Cali¹ Konstantinos Kanellopoulos² Joël Lindegger² Zülal Bingöl³
Gurpreet S. Kalsi⁴ Ziyi Zuo⁵ Can Firtina² Meryem Banu Cavlak² Jeremie Kim²
Nika Mansouri Ghiasi² Gagandeep Singh² Juan Gómez-Luna² Nour Almadhoun Alserr²
Mohammed Alser² Sreenivas Subramoney⁴ Can Alkan³ Saugata Ghose⁶ Onur Mutlu²

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

Genome Sequence Analysis

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

Linear Reference: ACG**T**ACGT

Read: ACGG

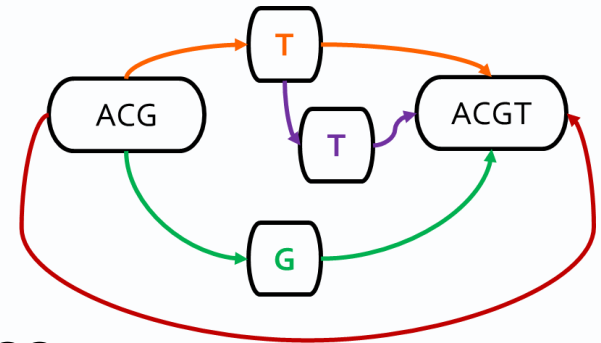
Alternative Sequence: ACG**G**ACGT

Alternative Sequence: ACG**TT**ACGT

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**.

Analysis of State-of-the-Art Tools

Based on our analysis with **GraphAligner** and **vg:**

SW

Observation 1: Alignment step is the bottleneck

Observation 2: Alignment suffers from high cache miss rates

Observation 3: Seeding suffers from the DRAM latency bottleneck

Observation 4: Baseline tools scale sublinearly

Observation 5: Existing S2S mapping accelerators are unsuitable for the S2G mapping problem

HW

Observation 6: Existing graph accelerators are unable to handle S2G alignment

SeGraM: First Graph Mapping Accelerator

Our Goal:

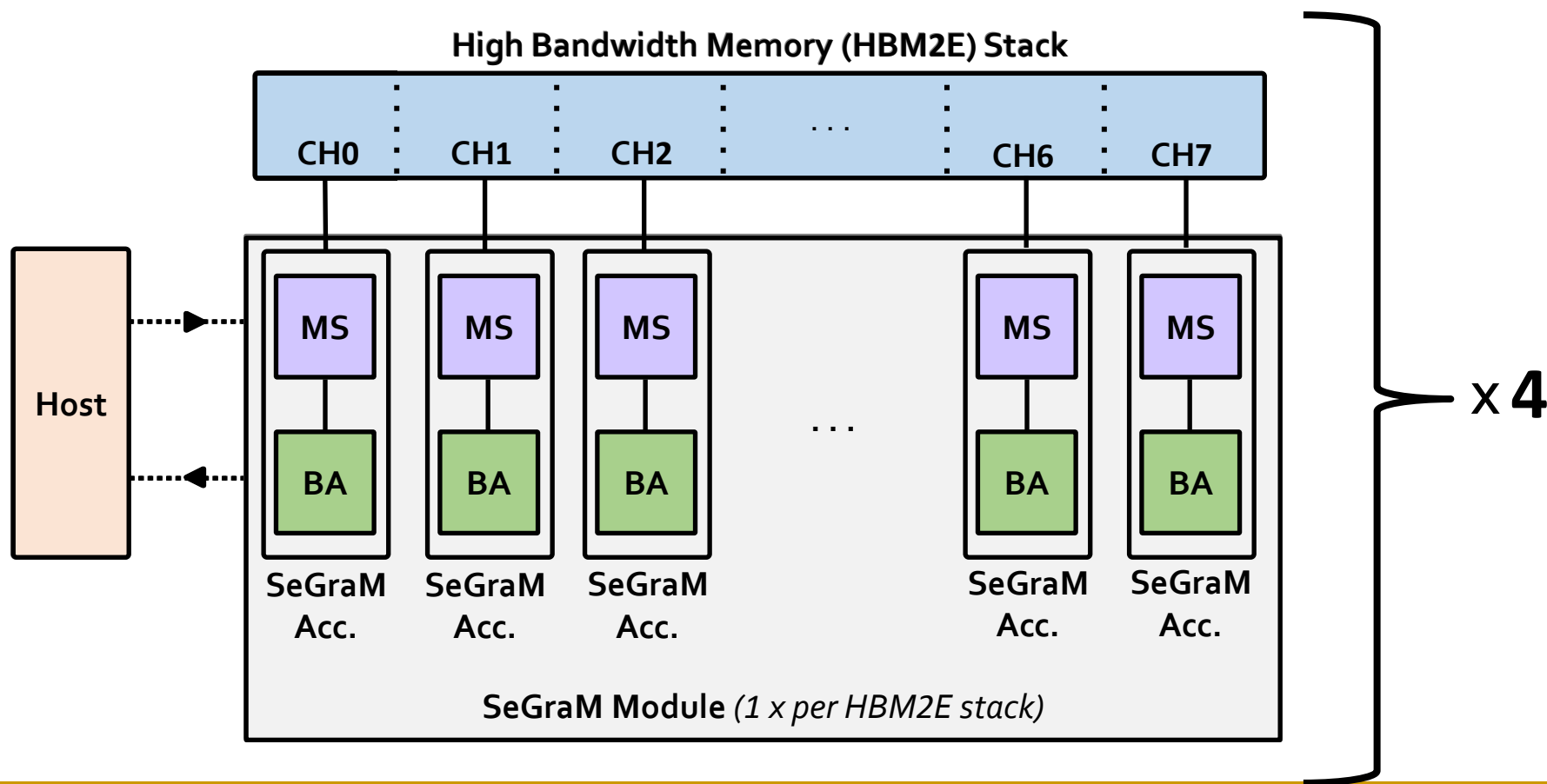
Specialized, high-performance, scalable, and low-cost algorithm/hardware co-design that alleviates bottlenecks in **multiple steps** of sequence-to-graph mapping

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

Overall System Design of SeGraM

High-Bandwidth Memory (HBM): Enables low-latency and highly-parallel memory access



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 Mao¹ Mohammed Alser¹ Mohammad Sadrosadati¹ Can Firtina¹ Akanksha Baranwal¹
Damla Senol Cali² Aditya Manglik¹ Nour Almadhoun Alserr¹ Onur Mutlu¹
¹*ETH Zürich* ²*Bionano Genomics*

Overview: Two Limitations

Multiple steps in genome analysis



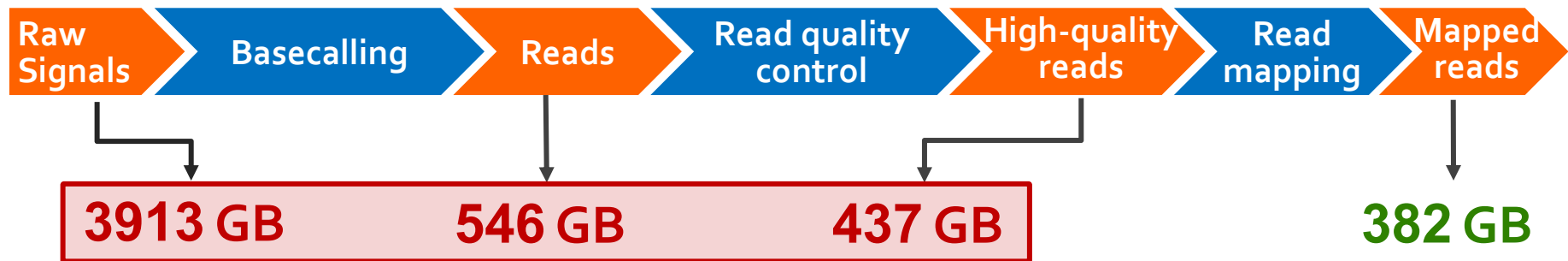
Large data movement
between multiple steps



A lot of
wasted computation
done on data that is
later discovered to be
useless

Limitation 1: Large Data Movement

- Using a human dataset in [NC'19] as an example:

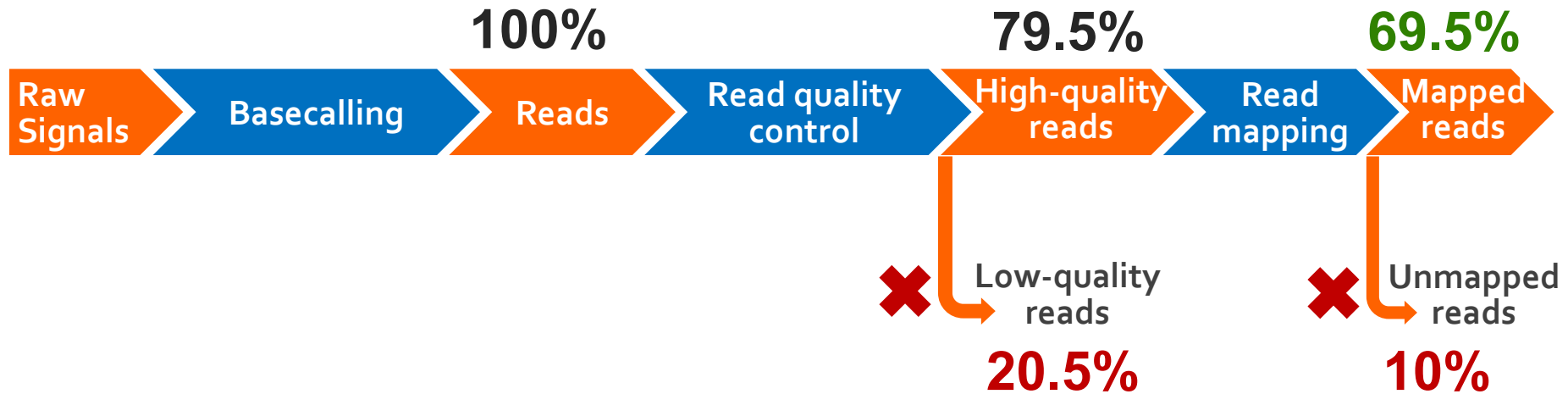


Large data movement between genome analysis steps

[NC'19] Rory Bowden, Robert W Davies, Andreas Heger, Alistair T Pagnamenta, Mariateresa de Cesare, Laura E Oikkonen, Duncan Parkes, Colin Freeman, Fatima Dhalla, Smita Y Patel, et al. Sequencing of human genomes with nanopore technology. Nature Communications, 2019.

Limitation 2: Wasted Computation

- Using a human dataset in [NC'19] as an example:



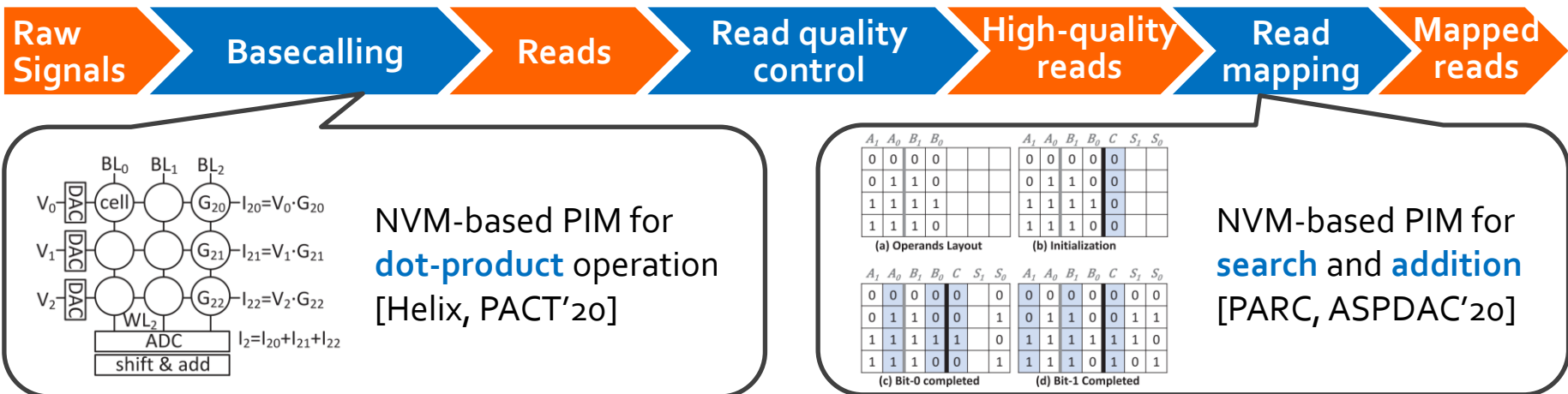
A considerable amount of computation on **useless data** due to

- Low-quality reads
- Unmapped reads

[NC'19] Rory Bowden, Robert W Davies, Andreas Heger, Alistair T Pagnamenta, Mariateresa de Cesare, Laura E Oikkonen, Duncan Parkes, Colin Freeman, Fatima Dhalla, Smita Y Patel, et al. Sequencing of human genomes with nanopore technology. Nature Communications, 2019.

State-of-the-art Works

- ❑ NVM-based PIM is an efficient technique to reduce data movement by processing data using or near memory



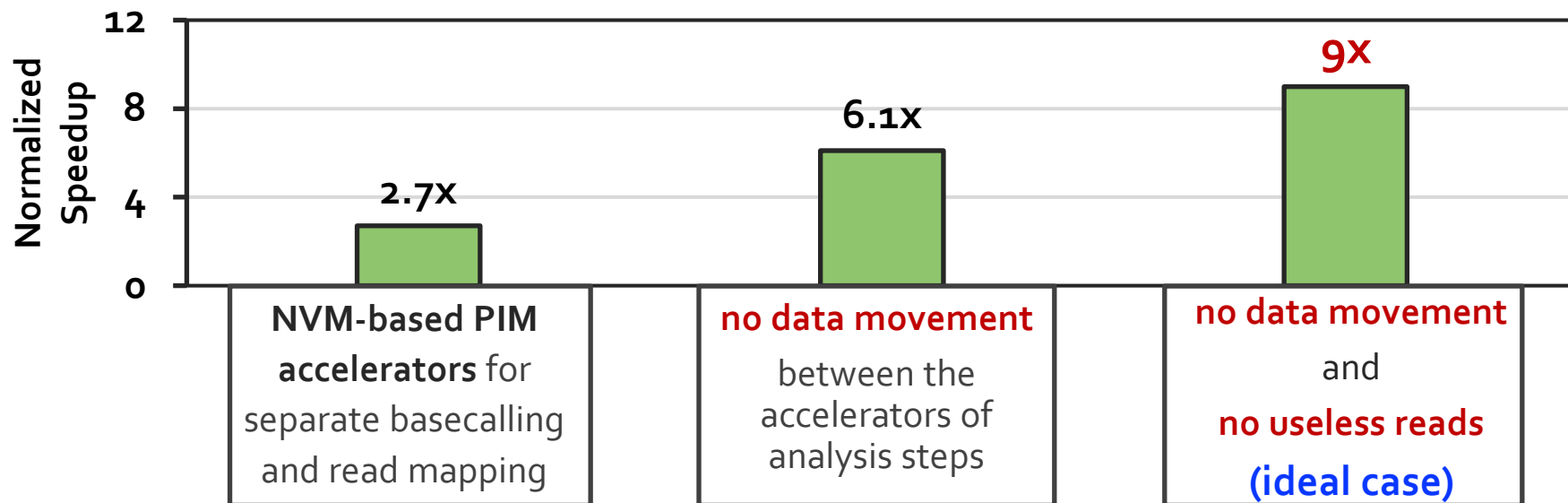
- Reduce the data movement in a single genome analysis step
- Exacerbate the data movement overhead between analysis steps

No prior work tackles data movement between analysis steps and reduces useless computation

Goal and Opportunities

Goal: Efficiently accelerate the entire genome analysis pipeline while **minimizing data movement and useless computation**

- We perform a study to quantify potential performance benefits
 - Results are normalized to the performance of GPU



Overview: GenPIP

❑ **GenPIP:** A fast and energy-efficient **in-memory** acceleration system for the Genome analysis PIpline via **tight integration of genome analysis steps**

❑ **GenPIP** has two key techniques

- **Chunk-based pipeline (CP)**

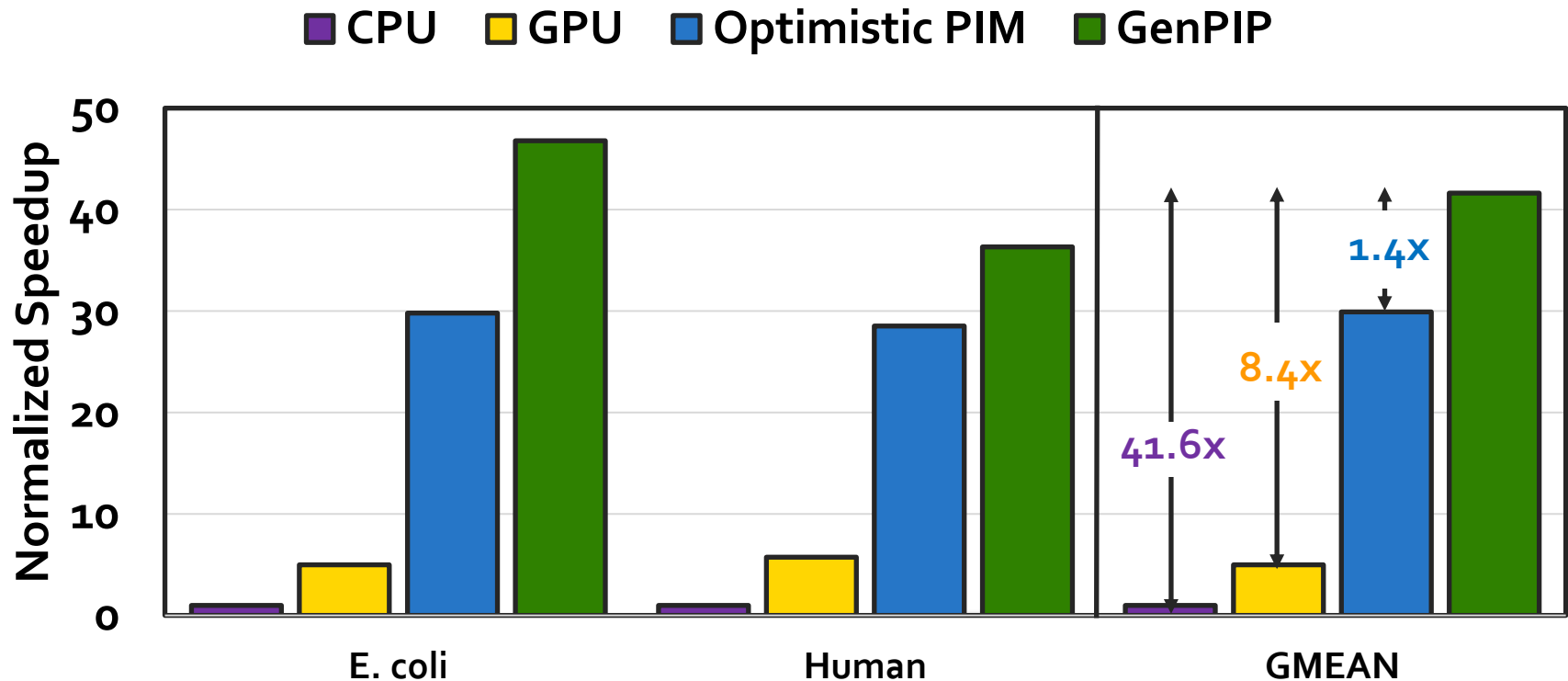
- **Provides fine-grained collaboration** of genome analysis steps

- **Early rejection (ER)**

- **Timely stops the execution on useless data** by predicting which reads will not be useful

❑ **GenPIP** outperforms state-of-the-art software & hardware solutions using **CPU**, **GPU**, and **optimistic PIM** by **41.6x**, **8.4x**, and **1.4x**, respectively.

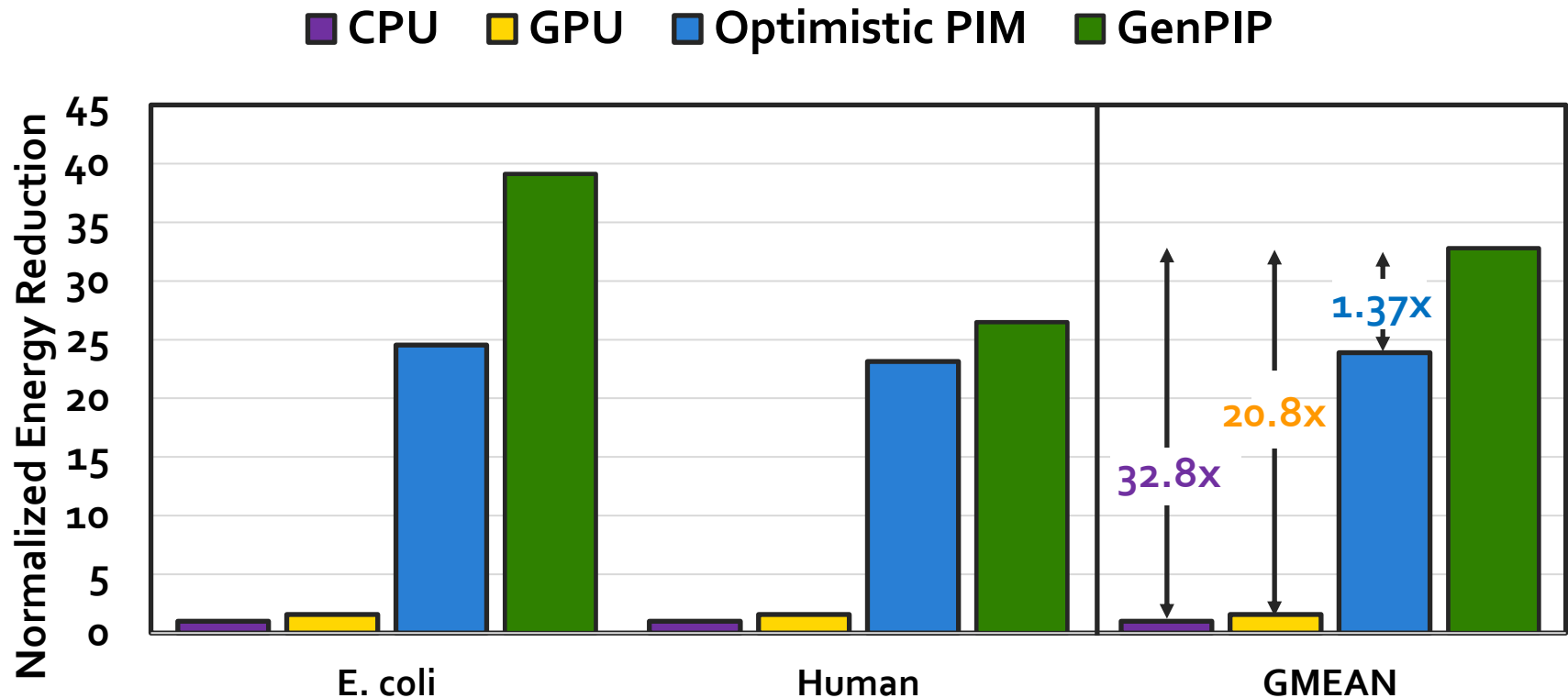
Key Results – Performance



GenPIP provides 41.6x, 8.4x, and 1.4x speedup over CPU, GPU, and optimistic PIM

Both CP and ER are critical to the speedup

Key Results – Energy Efficiency



GenPIP provides **32.8x**, **20.8x**, and **1.37x** energy savings over CPU, GPU, and optimistic PIM

ER is especially critical to the energy efficiency

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 Mao¹ Mohammed Alser¹ Mohammad Sadrosadati¹ Can Firtina¹ Akanksha Baranwal¹
Damla Senol Cali² Aditya Manglik¹ Nour Almadhoun Alserr¹ Onur Mutlu¹
¹ETH Zürich ²Bionano Genomics

Can we **process data closer**
to where it is **stored?**

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)]

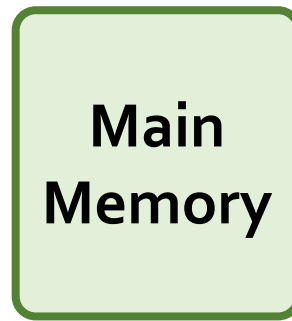
GenStore: A High-Performance In-Storage Processing System for Genome Sequence Analysis

Nika Mansouri Ghiasi¹ Jisung Park¹ Harun Mustafa¹ Jeremie Kim¹ Ataberk Olgun¹
Arvid Gollwitzer¹ Damla Senol Cali² Can Firtina¹ Haiyu Mao¹ Nour Almadhoun Alserr¹
Rachata Ausavarungnirun³ Nandita Vijaykumar⁴ Mohammed Alser¹ Onur Mutlu¹

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

Genome Sequence Analysis

Data Movement from Storage



Alignment

Computation Unit
(CPU or Accelerator)

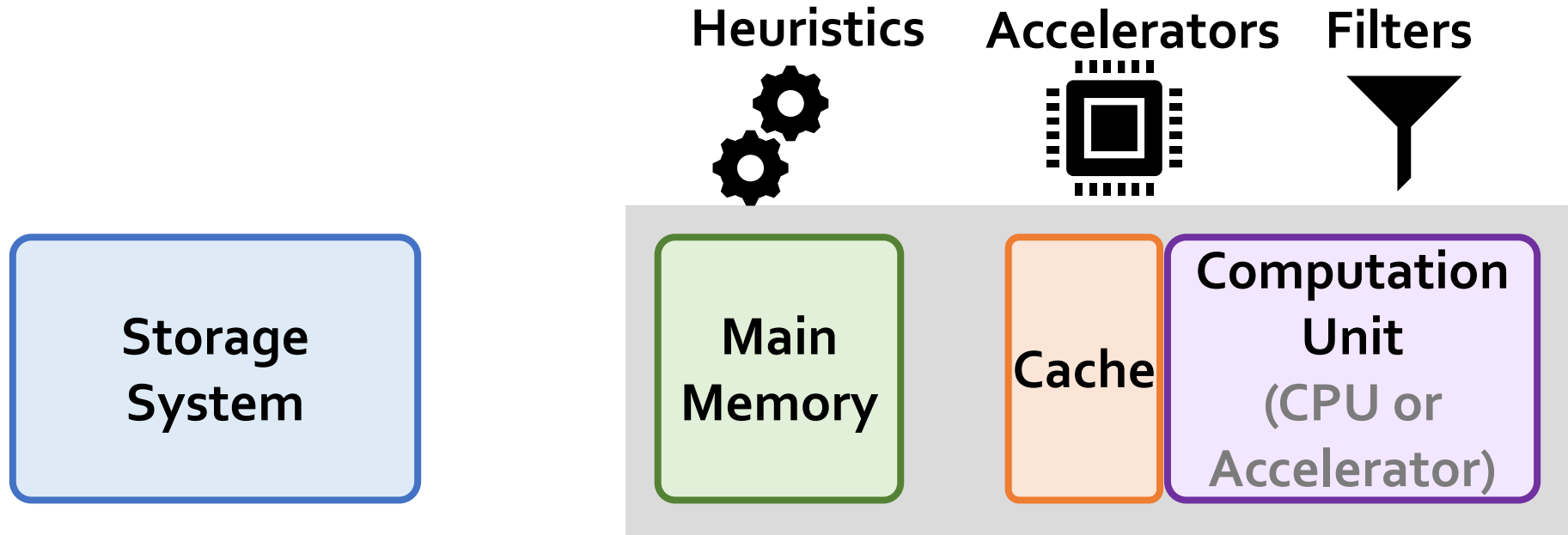


Computation overhead



Data movement overhead

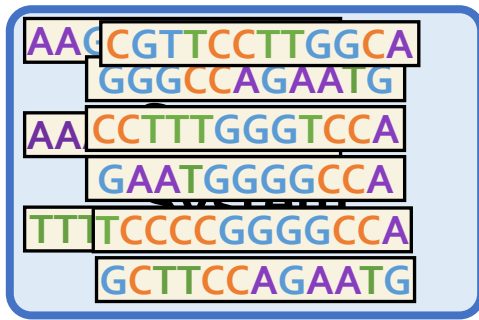
Accelerating Genome Sequence Analysis



Key Idea



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



Filtered Reads

**Main
Memory**

Cache

**Computation
Unit**
(CPU or
Accelerator)

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

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 Ghiasi¹ Jisung Park¹ Harun Mustafa¹ Jeremie Kim¹ Ataberk Olgun¹
Arvid Gollwitzer¹ Damla Senol Cali² Can Firtina¹ Haiyu Mao¹ Nour Almadhoun Alserr¹
Rachata Ausavarungnirun³ Nandita Vijaykumar⁴ Mohammed Alser¹ Onur Mutlu¹

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

In-Storage Metagenomics [ISCA 2024]

- Nika Mansouri Ghiasi, Mohammad Sadrosadati, Harun Mustafa, Arvid Gollwitzer, Can Firtina, Julien Eudine, Haiyu Mao, Joel Lindegger, Meryem Banu Cavlak, Mohammed Alser, Jisung Park, and Onur Mutlu,

"MegIS: High-Performance and Low-Cost Metagenomic Analysis with In-Storage Processing"

Proceedings of the 51st Annual International Symposium on Computer Architecture (ISCA), Buenos Aires, Argentina, July 2024.

[[Slides \(pptx\)](#)] [[pdf](#)]

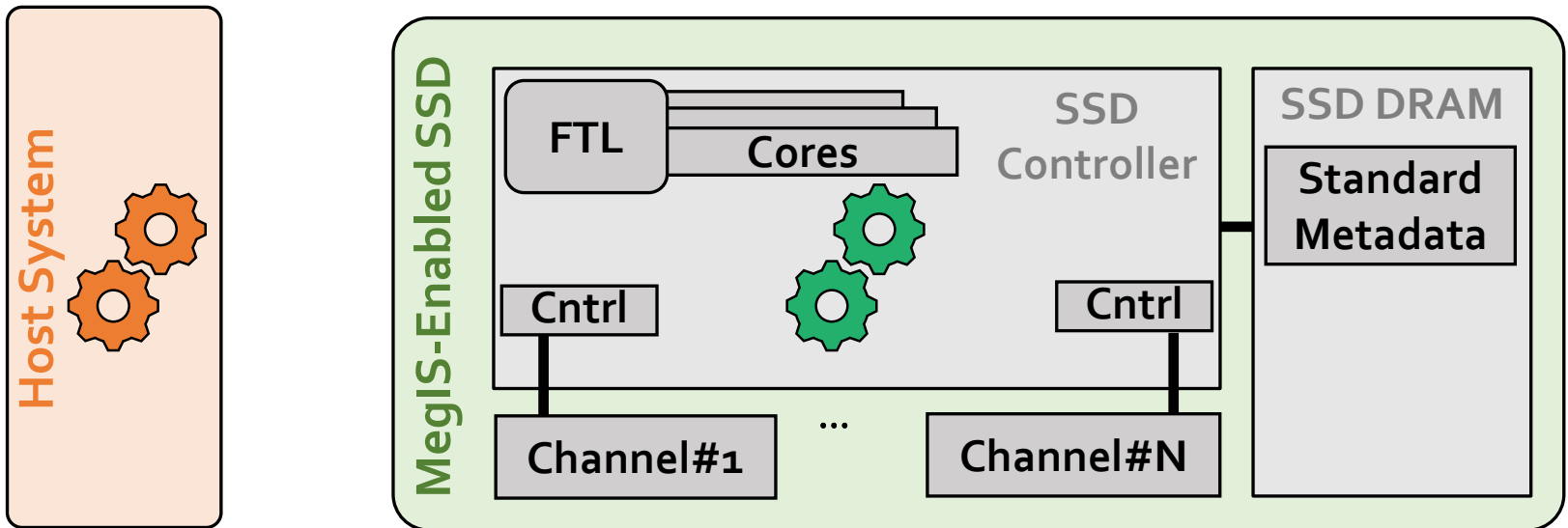
[[arXiv version](#)]

MegIS: High-Performance, Energy-Efficient, and Low-Cost Metagenomic Analysis with In-Storage Processing

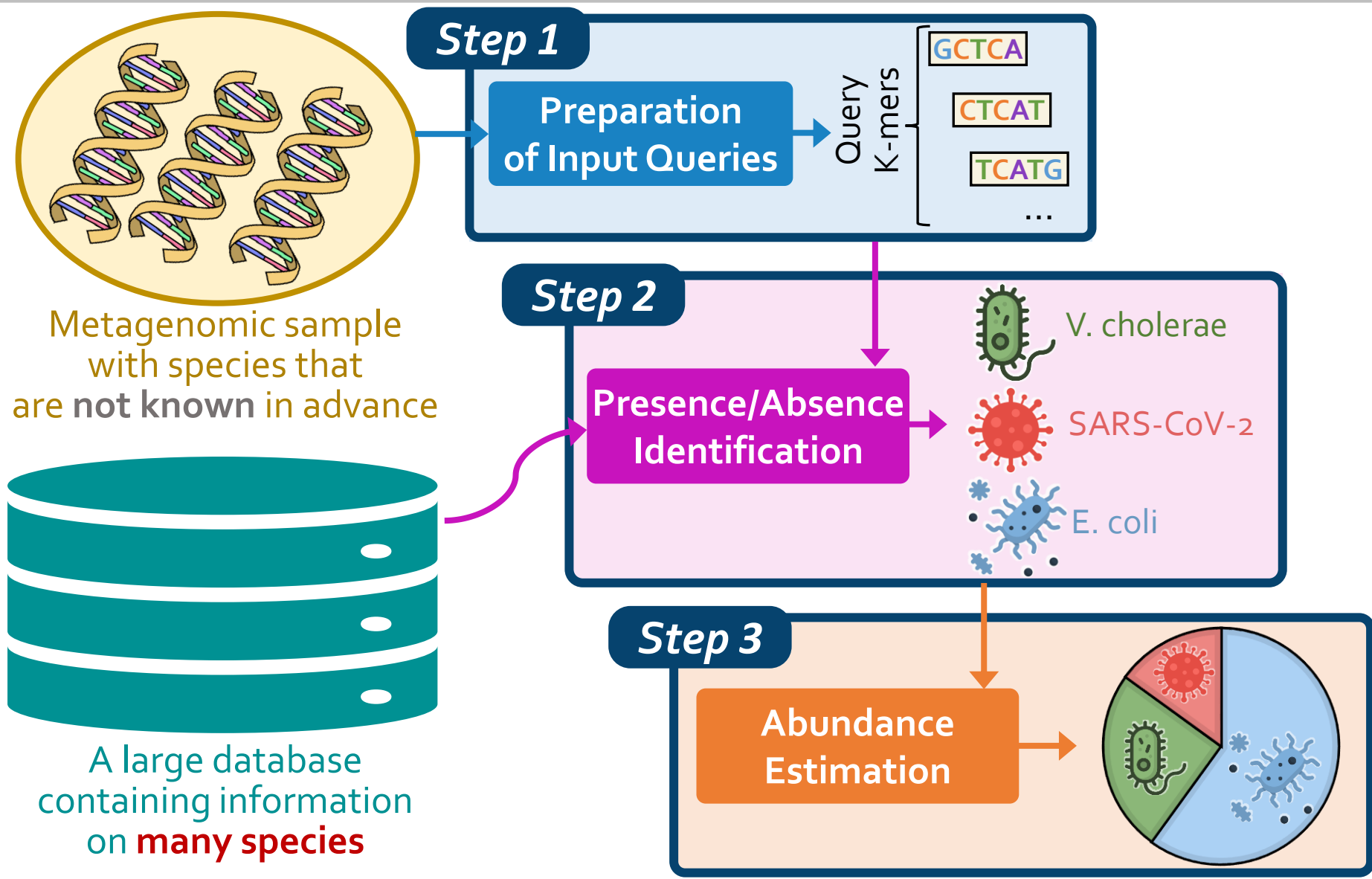
Nika Mansouri Ghiasi¹ Mohammad Sadrosadati¹ Harun Mustafa¹ Arvid Gollwitzer¹
Can Firtina¹ Julien Eudine¹ Haiyu Mao¹ Joël Lindegger¹ Meryem Banu Cavlak¹
Mohammed Alser¹ Jisung Park² Onur Mutlu¹
¹ETH Zürich ²POSTECH

MegIS: Metagenomics In-Storage

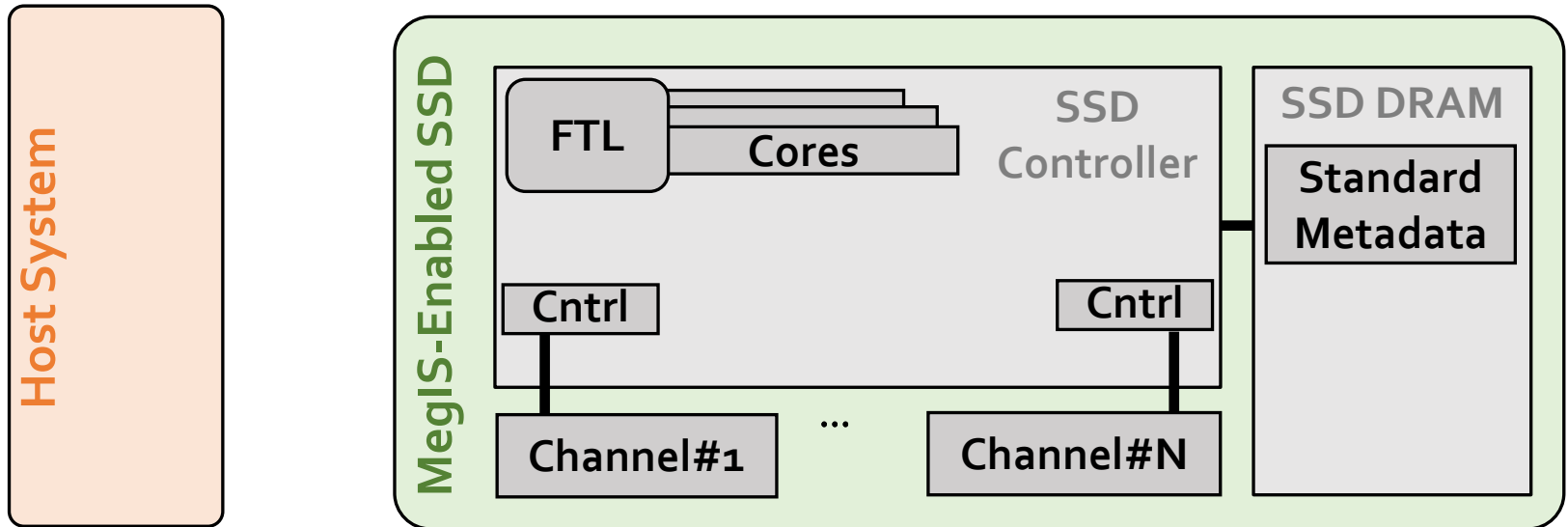
- First in-storage system for *end-to-end* metagenomic analysis
- **Idea:** Cooperative in-storage processing for metagenomic analysis
 - Hardware/software co-design between the **storage system** and **host system**



MegIS's Steps



MegIS Hardware-Software Co-Design



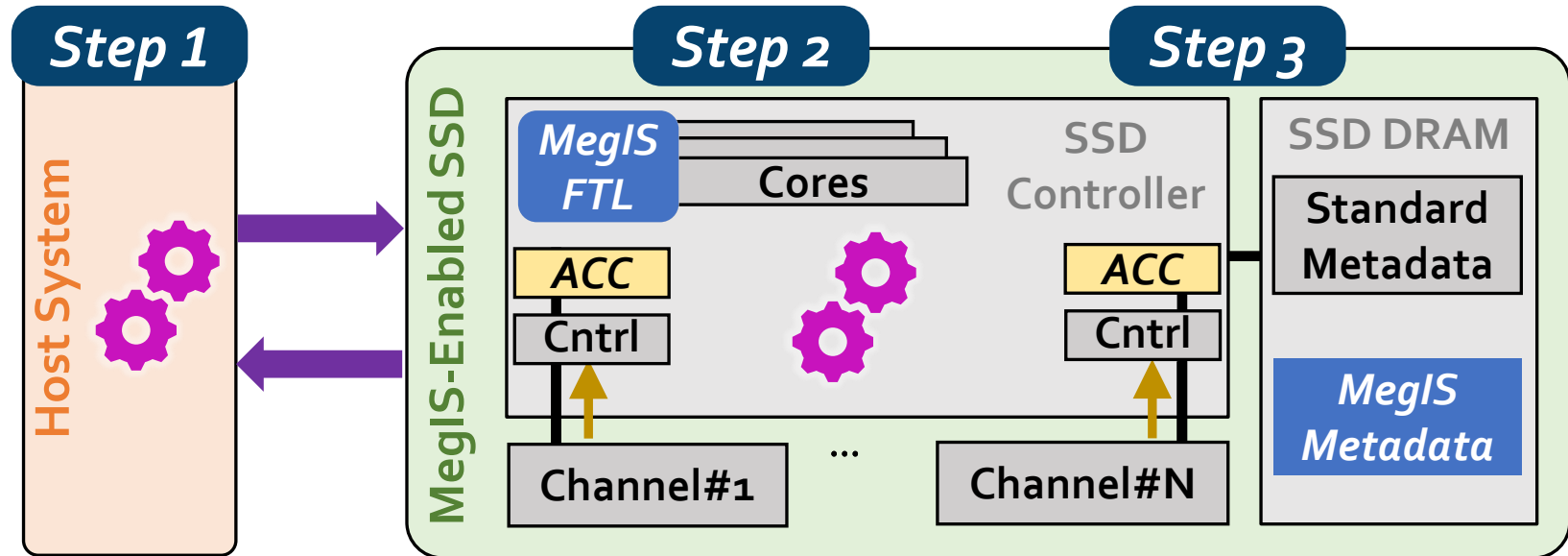
MegIS Hardware-Software Co-Design

Task partitioning and mapping

- Each step executes in its most suitable system

Data/computation flow coordination

- Reduce communication overhead
- Reduce #writes to flash chips



Storage-aware algorithms

- Enable efficient access patterns to the SSD

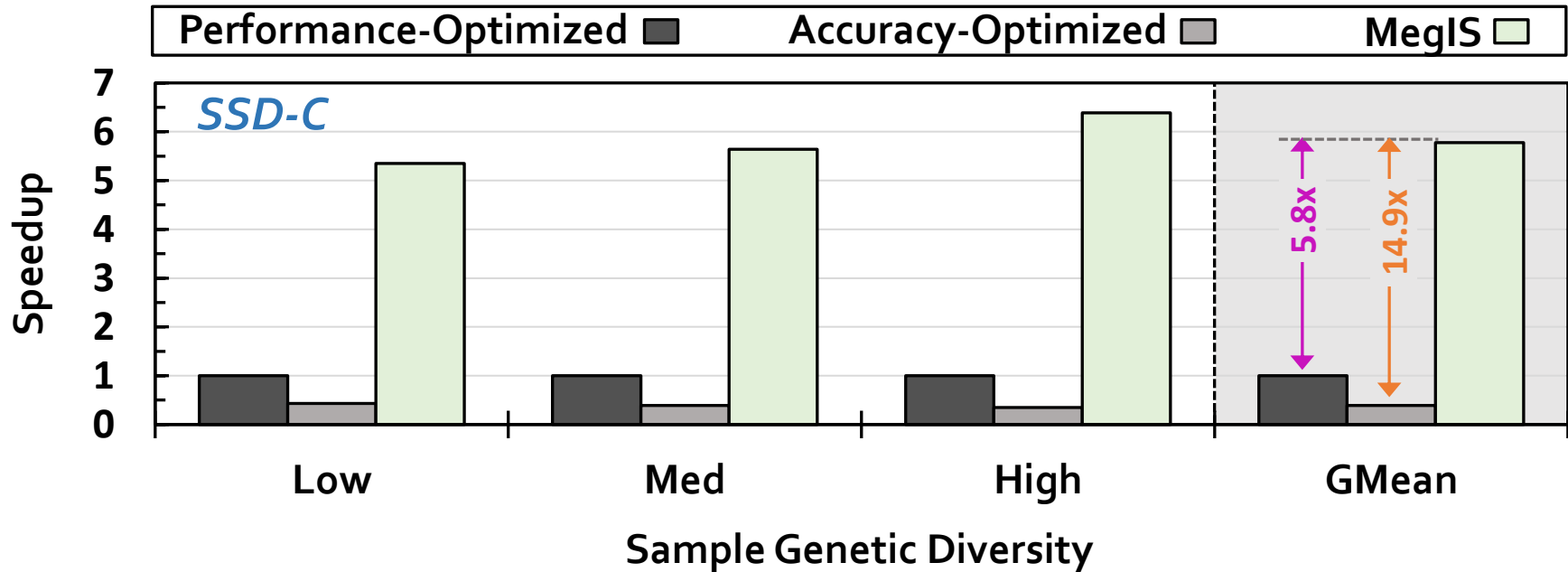
Lightweight in-storage accelerators

- Minimize SRAM/DRAM buffer spaces needed inside the SSD

Data mapping scheme and Flash Translation Layer (FTL)

- Specialize to the characteristics of metagenomic analysis
- Leverage the SSD's full internal bandwidth

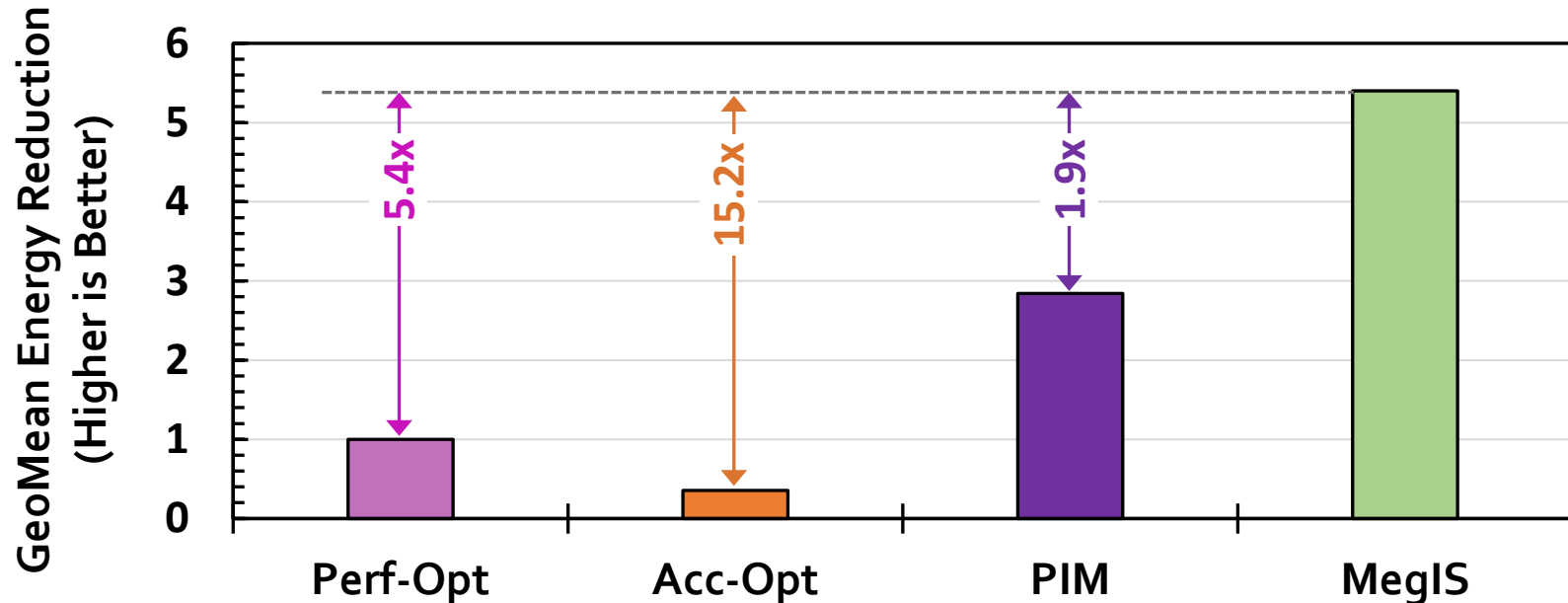
Evaluation: Speedup over the Software Baselines



MegIS provides significant speedup over both
Performance-Optimized and **Accuracy-Optimized** baselines

Evaluation: Reduction in Energy Consumption

- On average across different input sets and SSDs



MegIS provides significant energy reduction over the **Performance-Optimized**, **Accuracy-Optimized**, and **PIM** baselines

In-Storage Metagenomics [ISCA 2024]

- Nika Mansouri Ghiasi, Mohammad Sadrosadati, Harun Mustafa, Arvid Gollwitzer, Can Firtina, Julien Eudine, Haiyu Mao, Joel Lindegger, Meryem Banu Cavlak, Mohammed Alser, Jisung Park, and Onur Mutlu,

"MegIS: High-Performance and Low-Cost Metagenomic Analysis with In-Storage Processing"

Proceedings of the 51st Annual International Symposium on Computer Architecture (ISCA), Buenos Aires, Argentina, July 2024.

[[Slides \(pptx\)](#) ([pdf](#))]

[[arXiv version](#)]

MegIS: High-Performance, Energy-Efficient, and Low-Cost Metagenomic Analysis with In-Storage Processing

Nika Mansouri Ghiasi¹ Mohammad Sadrosadati¹ Harun Mustafa¹ Arvid Gollwitzer¹
Can Firtina¹ Julien Eudine¹ Haiyu Mao¹ Joël Lindegger¹ Meryem Banu Cavlak¹
Mohammed Alser¹ Jisung Park² Onur Mutlu¹
¹ETH Zürich ²POSTECH

Conclusion

- **System design for bioinformatics** is a critical problem
 - It has large scientific, medical, societal, personal implications
- This talk is about **accelerating** genomics by **alleviating data movement bottleneck**
- We covered various **recent works** on individual algorithms and pipelines
 - **PnM, PuM, ISP**
- **Many future opportunities exist**
 - **Especially with new sequencing technologies**
 - **Especially with new applications and use cases**

PIM Architectures for Bioinformatics

Dr. Konstantina Koliogeorgi

kkoliogeorgi@ethz.ch

<https://people.inf.ethz.ch/>

ICS 2025

08 June 2025

SAFARI

ETH zürich