

Accelerating AI-Driven Genome Analysis

via Algorithm-Architecture Co-Design

Can Firtina | canfirtina@gmail.com

Senior Researcher and Lecturer at ETH Zurich

Incoming Assistant Professor of Computer Science at UMD

June 16, 2025

PASC25, Challenges in Systems Design for Omics, Switzerland

ETH zürich

SAFARI

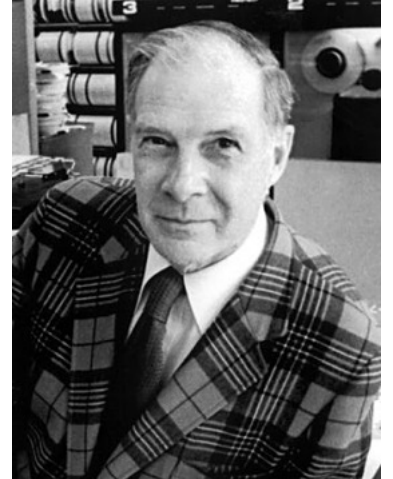


UNIVERSITY OF
MARYLAND

The Goal of Computing: Beyond Numbers

“The purpose of **computing** is [to gain] **insight**, not numbers”

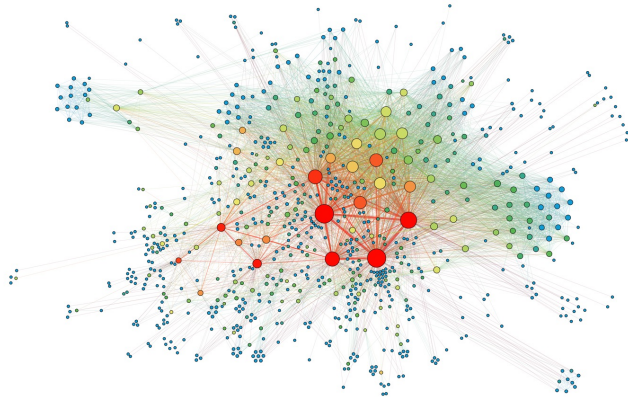
Richard Hamming



Big Data is Everywhere



Large recommender systems



Graph/Tree Processing



Large Language Models



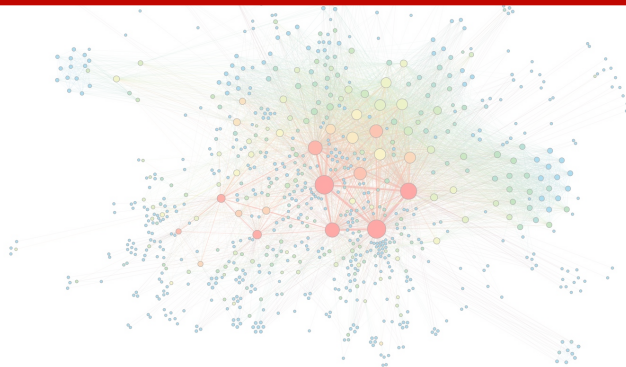
Genomics

Computing is Bottlenecked by Data

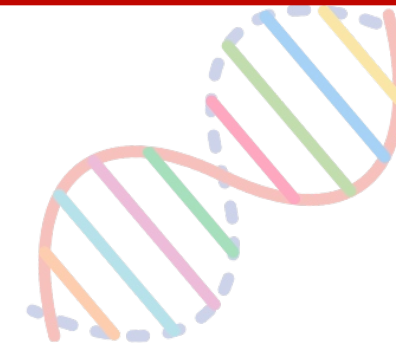


 Claude
3.5 Sonnet

**Data movement →
Performance & Energy Bottleneck**

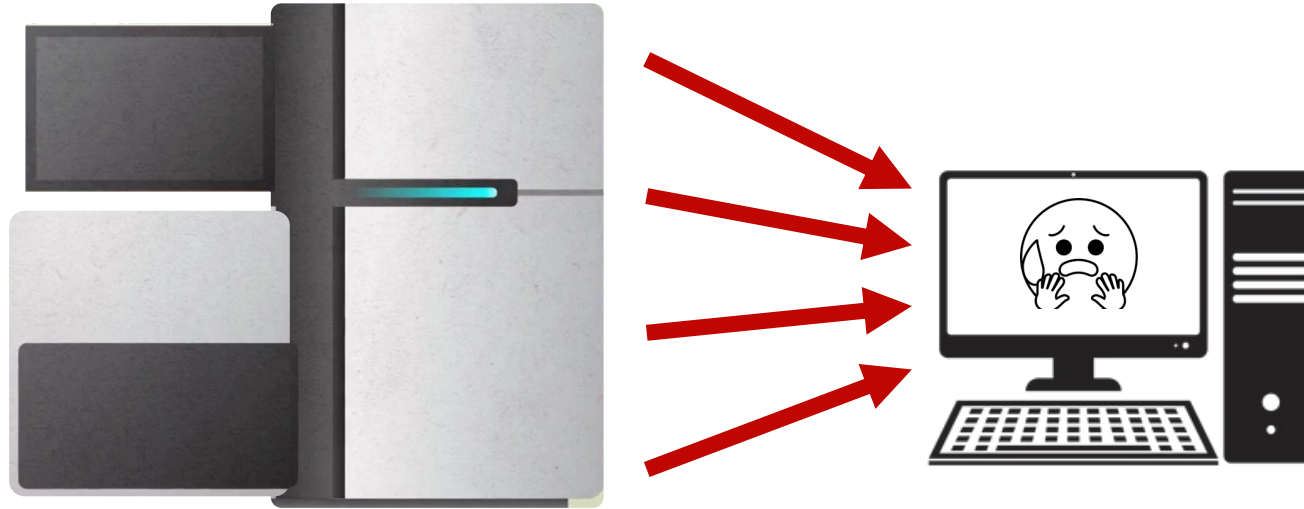


Graph/Tree Processing



Genomics

Problems with Data Analysis Today



Special-Purpose Machine
for **Data Generation**

FAST

General-Purpose Machine
for **Data Analysis**

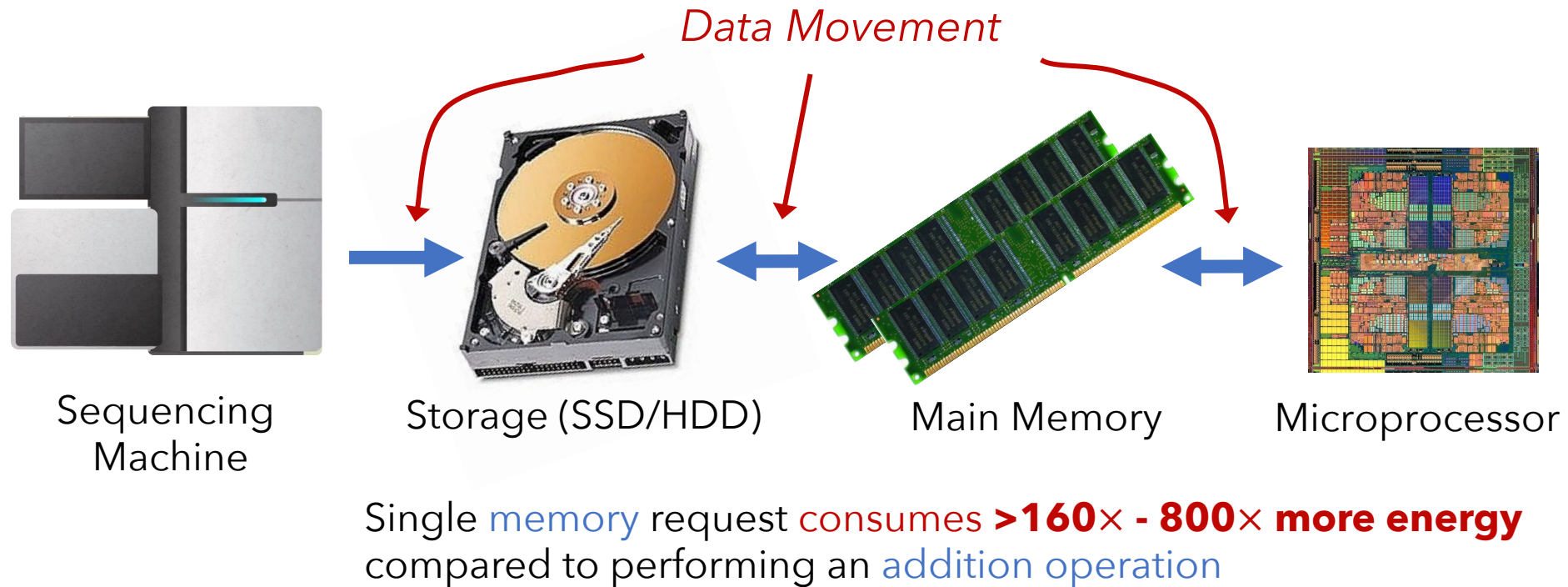
SLOW

Processing capabilities oblivious to data analysis requirements

Large amounts of data movement

Data Movement Dominates Performance

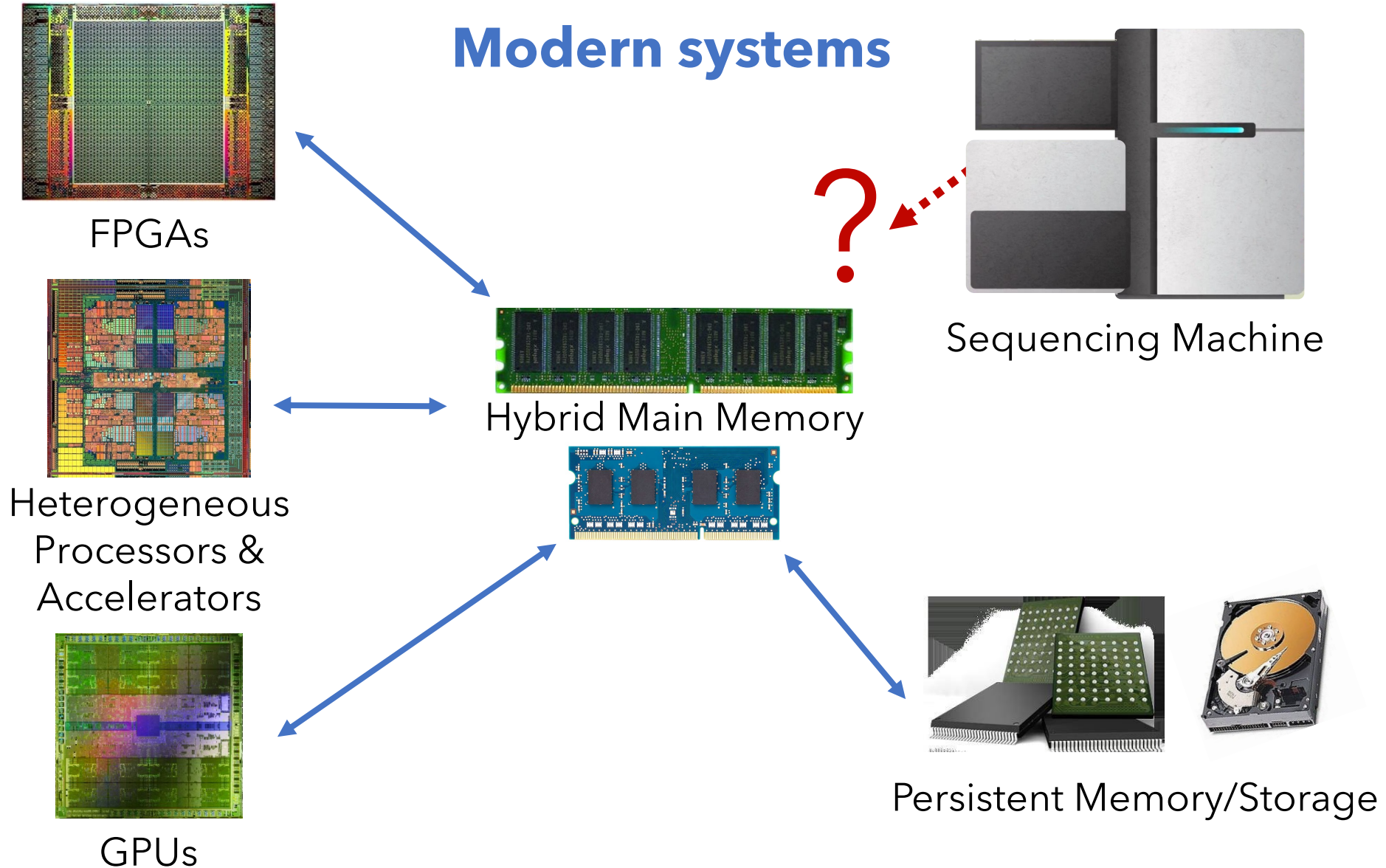
- **Data movement** dominates performance and is a **major system energy bottleneck** (accounting for 40% - 62%)



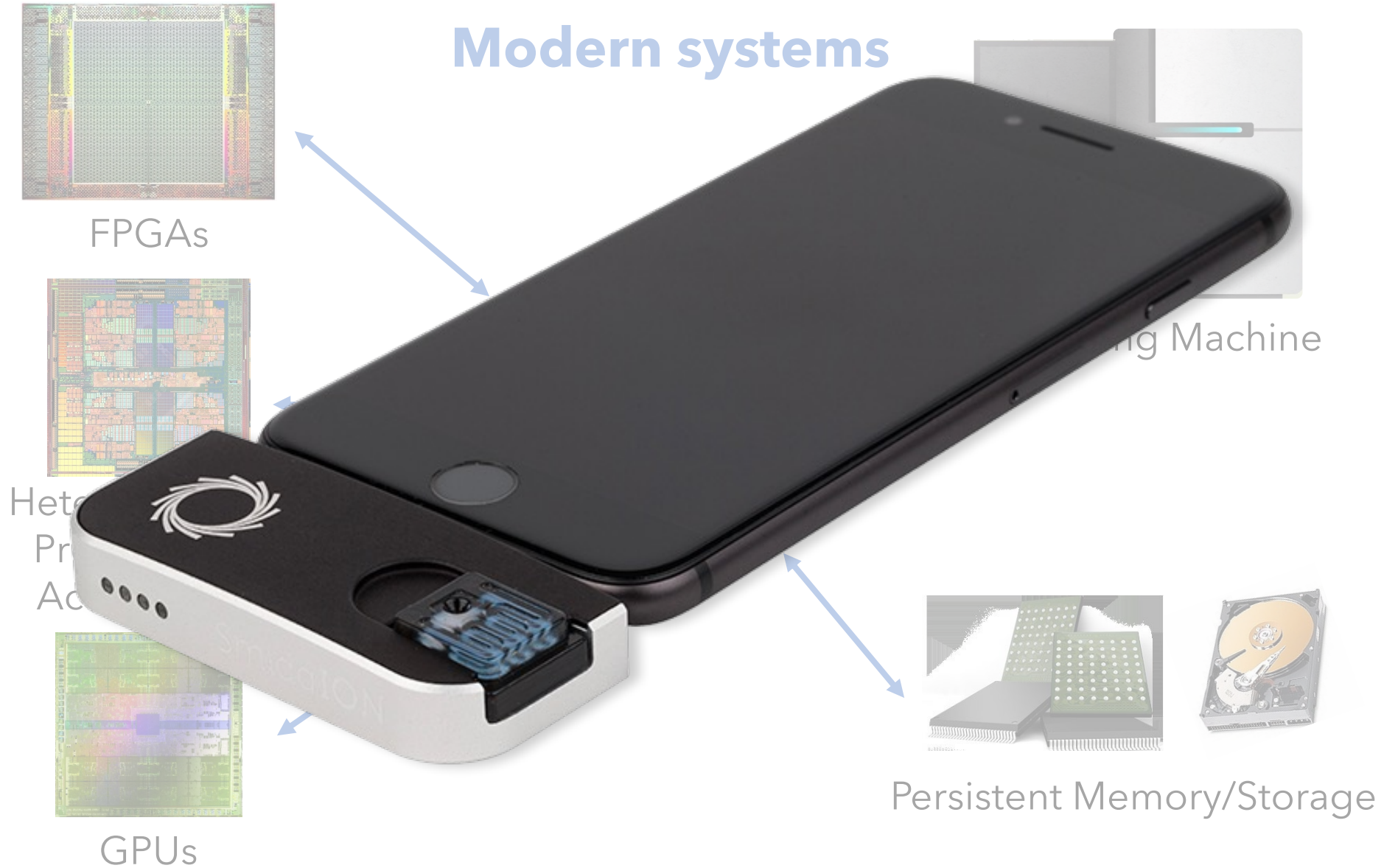
- Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS 2018
- Kestor et al., "Quantifying the Energy Cost of Data Movement in Scientific Applications," IISWC 2013
- Pandiyan and Wu, "Quantifying the energy cost of data movement for emerging smart phone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

**We need to co-design
algorithms and architecture
to handle data well**

Pushing Towards New Architectures

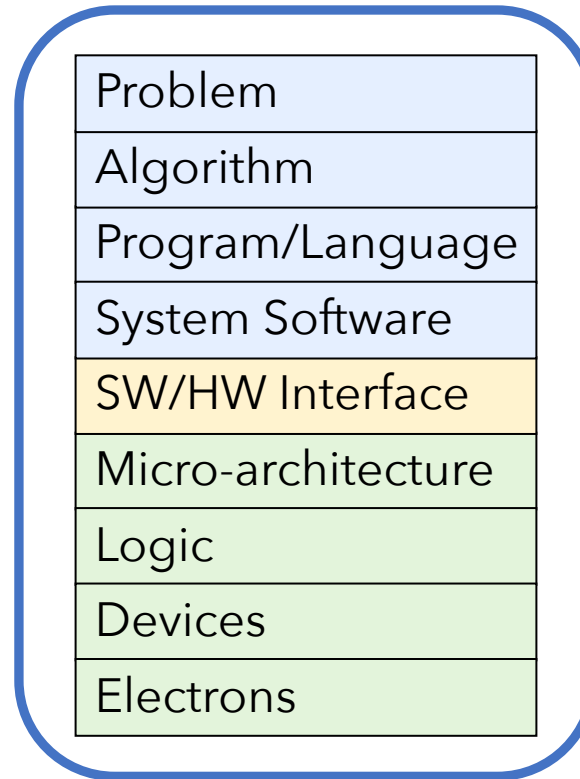


Pushing Towards New Architectures



Algorithm-Architecture Co-design is Critical

Computer Architecture (expanded view)



Accelerating Genome Analysis [DAC '23]

- Onur Mutlu and Can Firtina,
"Accelerating Genome Analysis via Algorithm-Architecture Co-Design"
*Invited Special Session Paper in Proceedings of the 60th Design Automation Conference (**DAC**),
San Francisco, CA, USA, July 2023.*
[\[Related Invited Paper\]](#)
[\[arXiv version\]](#)
[\[Slides \(pptx\) \(pdf\)\]](#)
[\[Talk Video at DAC 2023\]](#) (38 minutes, including Q&A)

Accelerating Genome Analysis via Algorithm-Architecture Co-Design

Onur Mutlu Can Firtina
ETH Zürich

Analyzing Genomes Reveals Key Insights



Personalized Medicine



Altering genomes to solve fundamental challenges of life

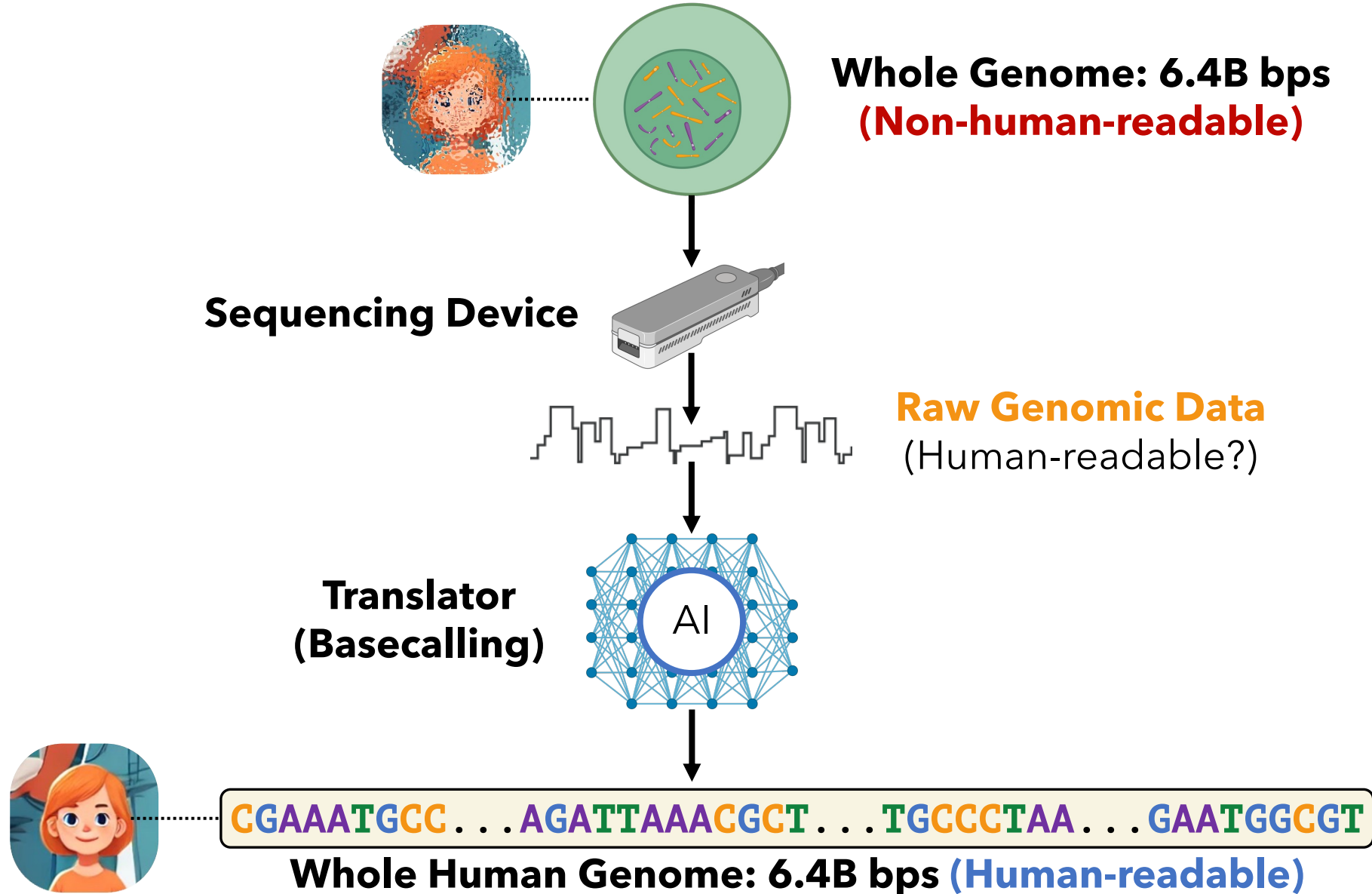


Detecting **pathogens** in the environment

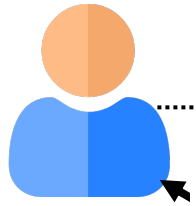


Rapid surveillance of **disease outbreaks**

Generating the Entire Genomic Sequence

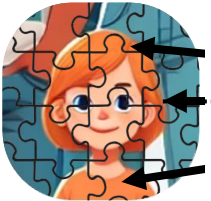


Genomic Data: Giant Jigsaw Puzzle to Solve



Human (Reference) Genome: 3.2 Billion bps (haploid)

CGAAATGCC...AGATTAAACGGT...TGCCCTAA...GAATGGCGT



?

CGAAATGCC...AGATTAAATGGT...TGCCCTAA...GAATGGCGT

Entire Genomic Sequence of an Organism



Small pieces (reads)



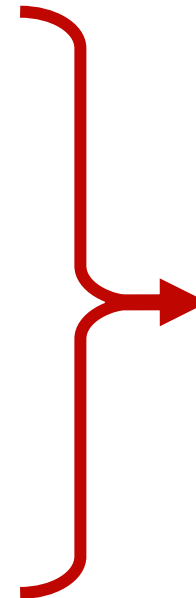
Unknown origins



Non-identical reference genome

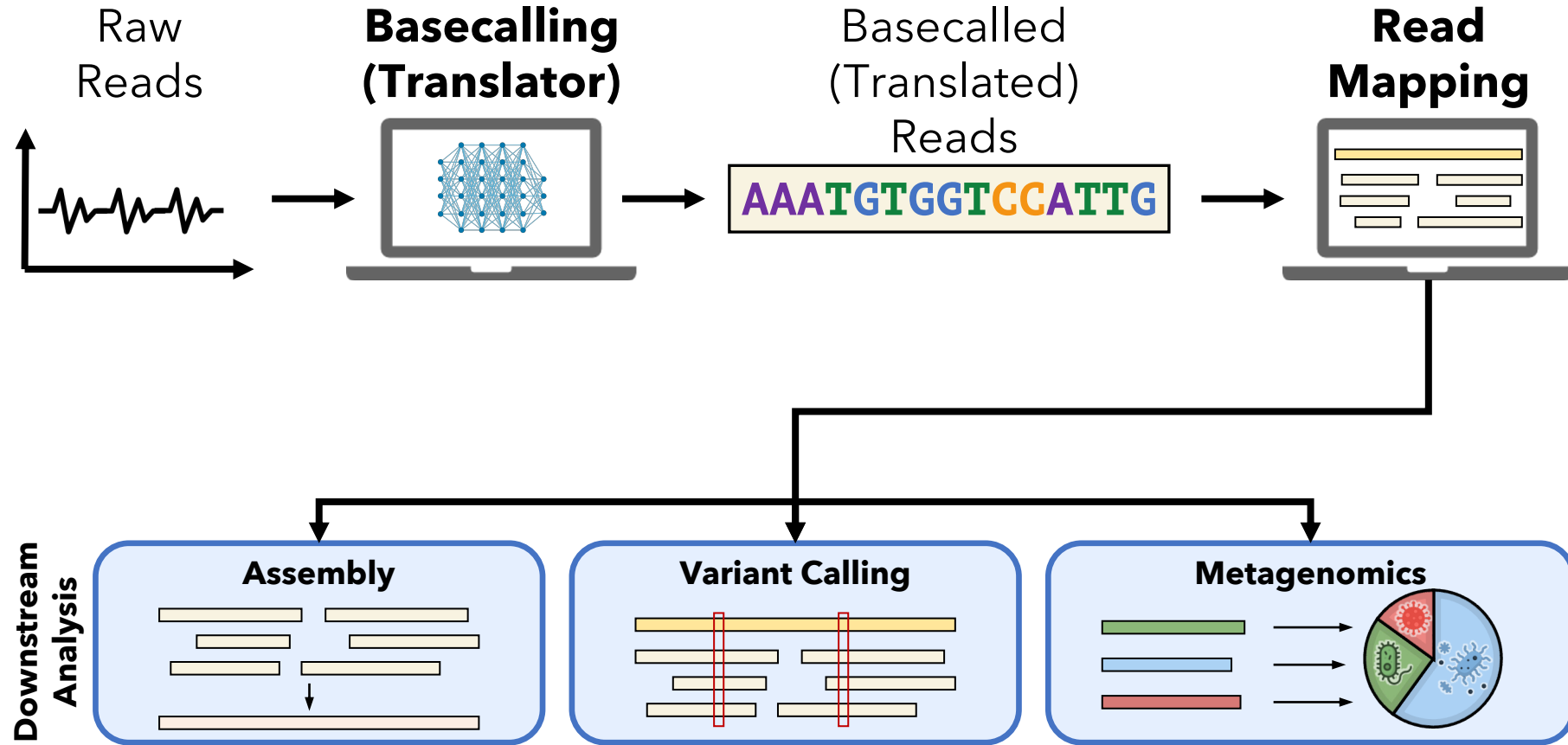


Large volume of data to analyze

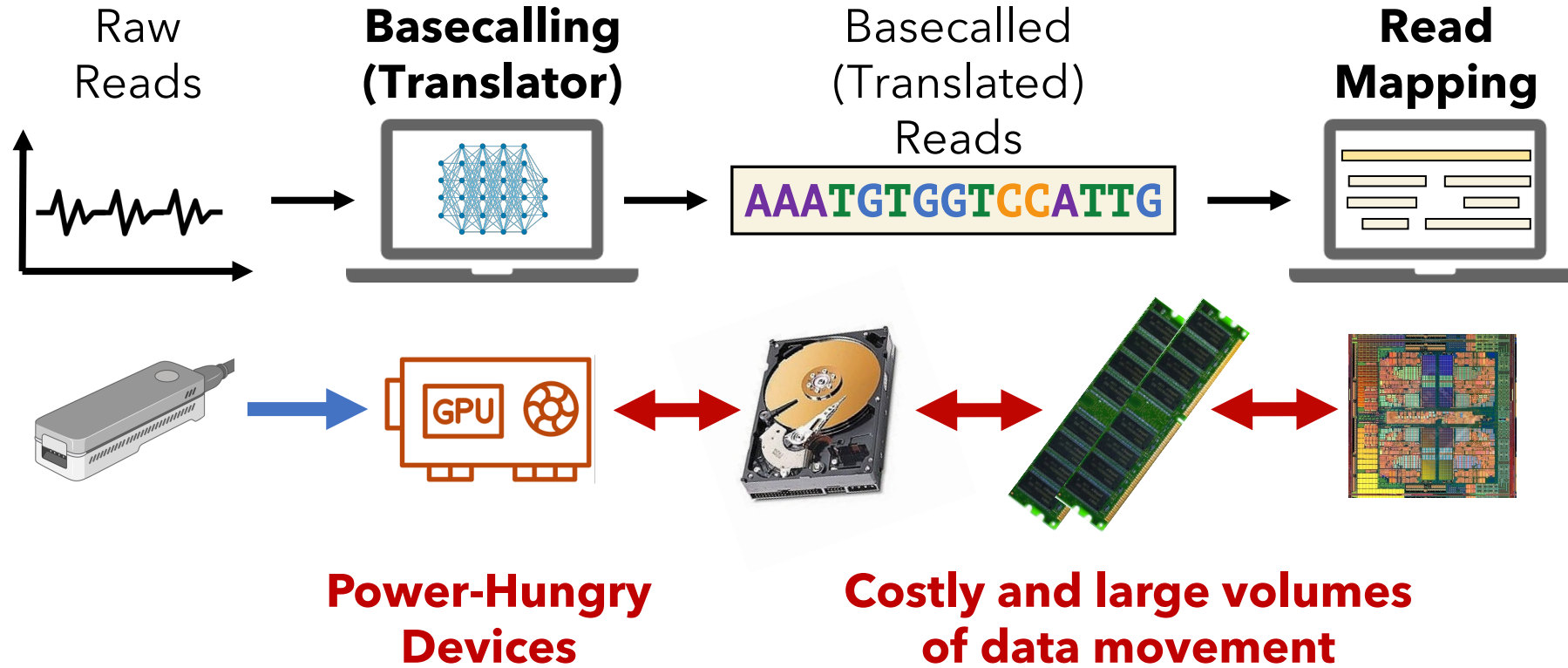


**Data Science for
Massive Datasets
(TBs to PBs)**

A Common Genome Analysis Pipeline



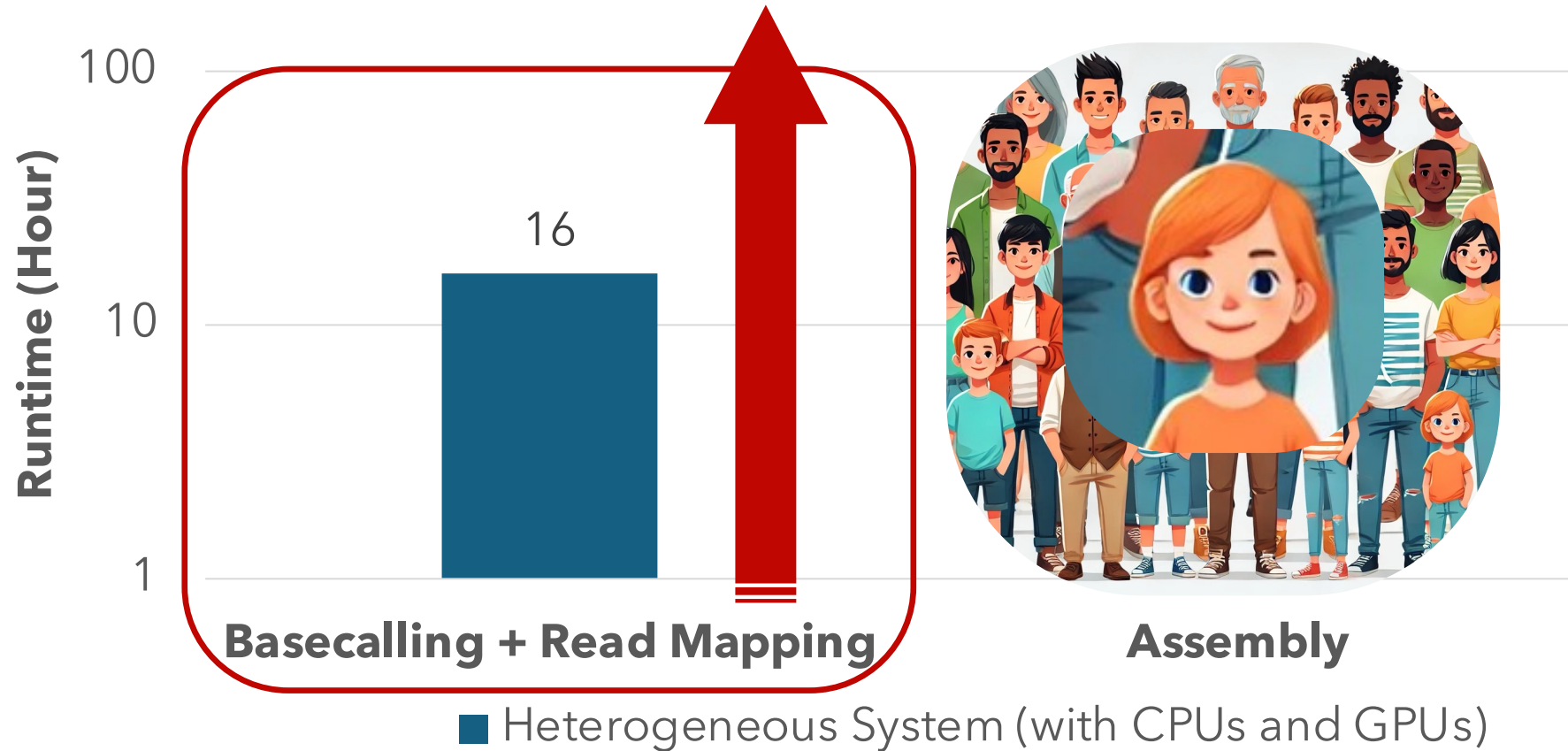
Genome Analysis is Energy Inefficient



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

Genome Analysis is Computationally Costly



Significant computation and energy overhead

Limited Application Scope and Accuracy



**Energy-efficient analysis for
Resource-constrained devices**

**Urgent analysis
within minutes**



**Large-scale analysis
without sacrificing accuracy
and performance**



Enabling New Directions in Genome Analysis **via New Algorithms**

New Frontiers: Raw Signal Analysis [ISMB '23]

- [Can Firtina](#), Nika Mansouri Ghiasi, Joel Lindegger, Gagandeep Singh, Meryem Banu Cavlak, Haiyu Mao, and Onur Mutlu,

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

Proceedings of the [31st Annual Conference on Intelligent Systems for Molecular Biology \(ISMB\)](#) and the [22nd European Conference on Computational Biology \(ECCB\)](#), July 2023

[\[Online link at Bioinformatics Journal\]](#)

[\[arXiv version\]](#)

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

[\[Talk Video at ISMB/ECCB 2023\]](#) (19 minutes)

[\[RawHash Source Code\]](#)

Bioinformatics, 2023, **39**, i297–i307
<https://doi.org/10.1093/bioinformatics/btad272>
ISMB/ECCB 2023

OXFORD

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,*}

¹Department of Information Technology and Electrical Engineering, ETH Zurich, 8092 Zurich, Switzerland

*Corresponding author. Department of Information Technology and Electrical Engineering, ETH Zurich, Gloriastrasse 35, 8092 Zurich, Switzerland.
E-mail: firtinac@ethz.ch (C.F.), omutlu@ethz.ch (O.M.)

Real-Time Raw Signal Analysis [Bioinformatics '24]

- [Can Firtina](#), Melina Soysal, Joël Lindegger, and Onur Mutlu, **"RawHash2: Mapping Raw Nanopore Signals Using Hash-Based Seeding and Adaptive Quantization"** *Bioinformatics*, July 2024.
[[Online link at Bioinformatics Journal](#)]
[[arXiv version](#)]
[[RawHash Talk Video](#)] (19 minutes)
[[RawHash2 Source Code](#)]

Bioinformatics, 2024, **40(8)**, btae478
<https://doi.org/10.1093/bioinformatics/btae478>
Advance Access Publication Date: 30 July 2024
Applications Note



Sequence analysis

RawHash2: mapping raw nanopore signals using hash-based seeding and adaptive quantization

Can Firtina ^{1,*}, Melina Soysal ¹, Joël Lindegger ¹, Onur Mutlu ^{1,*}

¹Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8092, Switzerland

*Corresponding authors. Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8092, Switzerland.
E-mail: firtinac@ethz.ch (C.F.); omutlu@ethz.ch (O.M.)

New Directions: Assembling Raw Signals

- [Can Firtina](#), Maximilian Mordig, Harun Mustafa, Sayan Goswami, Nika Mansouri Ghiasi, Stefano Mercogliano, Furkan Eris, Joël Lindegger, Andre Kahles, and Onur Mutlu

"Rawsample: Overlapping and Assembling Raw Nanopore Signals using a Hash-based Seeding Mechanism"

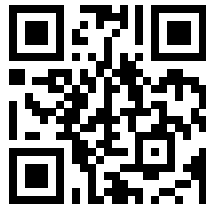
[arXiv](#), October 2024

[\[arXiv version\]](#)

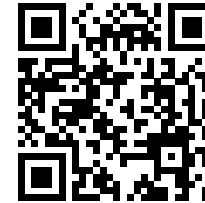
[\[Talk Video at CSHL Data Science 2024\]](#) (16 minutes)

[\[Source Code\]](#)

arXiv



Source code



Rawsample: Overlapping and Assembling Raw Nanopore Signals using a Hash-based Seeding Mechanism

Can Firtina¹ Maximilian Mordig^{1,2} Harun Mustafa^{1,3,4} Sayan Goswami¹ Nika Mansouri Ghiasi¹
Stefano Mercogliano¹ Furkan Eris¹ Joël Lindegger¹ Andre Kahles^{1,3,4} Onur Mutlu¹

¹*ETH Zurich*

²*Max Planck Institute for Intelligent Systems*

³*University Hospital Zurich*

⁴*Swiss Institute of Bioinformatics*

Quickly Aligning Raw Signals [IEEE Access '24]

- Joel Lindegger, Can Firtina, Nika Mansouri Ghiasi, Mohammad Sadrosadati, Mohammed Alser, and Onur Mutlu,

"RawAlign: Accurate, Fast, and Scalable Raw Nanopore Signal Mapping via Combining Seeding and Alignment"

IEEE Access, December 2024.

[[Online link at IEEE Access Journal](#)]

[[arXiv version](#)]

[[RawAlign Source Code](#)]

RawAlign: Accurate, Fast, and Scalable Raw Nanopore Signal Mapping via Combining Seeding and Alignment

**JOËL LINDEGGER^{ID}, CAN FIRTINA^{ID}, NIKA MANSOURI GHIASI, MOHAMMAD SADROSADATI^{ID},
MOHAMMED ALSER^{ID}, AND ONUR MUTLU^{ID}, (Fellow, IEEE)**

Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland

Corresponding authors: Joël Lindegger (jmlindegger@gmail.com), Can Firtina (firtinac@ethz.ch), and Onur Mutlu (omutlu@ethz.ch)

This work was supported in part by the Semiconductor Research Corporation (SRC), in part by the ETH Future Computing Laboratory, in part by European Union's Horizon Programme for Research and Innovation under Grant 101047160-BioPIM, and in part by the Swiss National Science Foundation (SNSF) under Grant 200021_213084.

In-Storage Raw Signal Analysis [ICS '25]

- Melina Soysal, Konstantina Koliogeorgi, Can Firtina, Nika Mansouri Ghiasi, Rakesh Nadig, Haiyu Mao, Geraldo Francisco Oliveira Junior, Yu Liang, Klea Zambaku, Mohammad Sadrosadati, and Onur Mutlu

"MARS: Processing-In-Memory Acceleration of Raw Signal Genome Analysis Inside the Storage Subsystem"

In 39th International Conference on Supercomputing (ICS), June 2025.

[\[arXiv version\]](#)

MARS: Processing-In-Memory Acceleration of Raw Signal Genome Analysis Inside the Storage Subsystem

Melina Soysal[†] Konstantina Koliogeorgi[†] Can Firtina[†] Nika Mansouri Ghiasi[†]
Rakesh Nadig[†] Haiyu Mayo^{*} Geraldo F. Oliveira[†]
Yu Liang[†] Klea Zambaku[†] Mohammad Sadrosadati[†] Onur Mutlu[†]

[†] *ETH Zürich*

^{*} *King's College London*

Tolerating Noise in String Matching [NARGAB '23]

- [Can Firtina](#), Jisung Park, Mohammed Alser, Jeremie S. Kim, Damla Senol Cali, Taha Shahroodi, Nika Mansouri Ghiasi, Gagandeep Singh, Konstantinos Kanellopoulos, Can Alkan, Onur Mutlu, **"BLEND: A Fast, Memory-Efficient, and Accurate Mechanism to Find Fuzzy Seed Matches in Genome Analysis"**
[NAR Genomics and Bioinformatics](#), March 2023.
[[Online link at NAR Genomics and Bioinformatics Journal](#)]
[[arXiv version](#)]
[[bioRxiv version](#)]
[[Talk Video at RECOMB 2023](#)] (23 minutes)
[[BLEND Source Code](#)]



Volume 5, Issue 1
March 2023

JOURNAL ARTICLE

BLEND: a fast, memory-efficient and accurate mechanism to find fuzzy seed matches in genome analysis

[Can Firtina](#) ✉, [Jisung Park](#), [Mohammed Alser](#), [Jeremie S Kim](#), [Damla Senol Cali](#), [Taha Shahroodi](#), [Nika Mansouri Ghiasi](#), [Gagandeep Singh](#), [Konstantinos Kanellopoulos](#), [Can Alkan](#), [Onur Mutlu](#) ✉

NAR Genomics and Bioinformatics, Volume 5, Issue 1, March 2023, lqad004,

Mitigating Useless Computations [IEEE/ACM TCBB '24]

- 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"

IEEE/ACM TCBB, August 2024.

[[Online link at IEEE/ACM TCBB Journal](#)]

[[arXiv version](#)]

Presented at the 21st Asia Pacific Bioinformatics Conference (APBC),
Changsha, China, April 2023.

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

[[Talk Video at BIO-Arch 2023 Workshop](#)] (22 minutes)

[[AirLift Source Code](#)]

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

Jeremie S. Kim^{1,†} Can Firtina^{1,†} Meryem Banu Cavlak¹ Damla Senol Cali²

Nastaran Hajinazar^{1,3} Mohammed Alser¹ Can Alkan⁴ Onur Mutlu^{1,2,4}

¹*ETH Zurich*

²*Carnegie Mellon University*

³*Simon Fraser University*

⁴*Bilkent University*

*Equal contribution

Mitigating Useless Computations [Bioinformatics '22]

- 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*, October 2022.
[[Online link at Bioinformatics Journal](#)]
[[arXiv preprint](#)]
[[FastRemap Source Code](#)]

Bioinformatics, 38(19), 2022, 4633–4635

<https://doi.org/10.1093/bioinformatics/btac554>

Advance Access Publication Date: 17 August 2022

Applications Note



Genome analysis

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,
Can Alkan ³ and **Onur Mutlu** ^{1,3}

¹Department of Computer Engineering, ETH Zurich, D-ITET, Zurich 8006, Switzerland, ²Department of Computer Engineering, Bionano Genomics, San Diego, CA 92121, USA and ³Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey

Utilizing AI/ML in
Genome Analysis
**via New Algorithms and
Architectures**

Error Correction using ML [Bioinformatics '20]

- [Can Firtina](#), Jeremie S. Kim, Mohammed Alser, Damla Senol Cali, A. Ercument Cicek, Can Alkan, and Onur Mutlu,
"Apollo: A Sequencing-Technology-Independent, Scalable, and Accurate Assembly Polishing Algorithm"
[Bioinformatics](#), June 2020.
[[Online link at Bioinformatics Journal](#)]
[[arXiv version](#)]
[[Slides \(pptx\)](#)][[\(pdf\)](#)]
[[Source Code](#)]

Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm

FREE

[Can Firtina](#), [Jeremie S Kim](#), [Mohammed Alser](#), [Damla Senol Cali](#), [A Ercument Cicek](#),
[Can Alkan](#) ✉, [Onur Mutlu](#) ✉

Bioinformatics, Volume 36, Issue 12, 15 June 2020, Pages 3669–3679,
<https://doi.org/10.1093/bioinformatics/btaa179>

Published: 13 March 2020 **Article history** ▼

Accelerating ML & Genome Graphs [ACM TACO '24]

- [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 Gomez Luna](#), [Sreenivas Subramoney](#) and [Onur Mutlu](#),

"ApHMM: Accelerating Profile Hidden Markov Models for Fast and Energy-efficient Genome Analysis"

[ACM Transactions on Architecture and Code Optimization](#) (**TACO**), Feb 2024.

[[ACM Digital Library version](#)]

[[arXiv version](#)]

Presented at the [19th HiPEAC Conference](#), Munich, Germany, January 2024.

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

[[Talk Video HiPEAC 2024](#)] (35 minutes)

[[ApHMM Source Code](#)]

RESEARCH-ARTICLE | **OPEN ACCESS** |  

ApHMM: Accelerating Profile Hidden Markov Models for Fast and Energy-efficient Genome Analysis

Authors:  [Can Firtina](#),  [Kamlesh Pillai](#),  [Gurpreet S. Kalsi](#),  [Bharathwaj Suresh](#),  [Damla Senol Cali](#),  [Jeremie S. Kim](#),  [Taha Shahroodi](#),  [Meryem Banu Cavlak](#),  [Joël Lindegger](#),  [Mohammed Alser](#),  [Juan Gómez Luna](#),  [Sreenivas Subramoney](#),  [Onur Mutlu](#) ([Less](#)) | [Authors Info & Claims](#)

ACM Transactions on Architecture and Code Optimization, Volume 21, Issue 1 • Article No.: 19, Pages 1 - 29

<https://doi.org/10.1145/3632950>

Translating Raw Signals using ML [Genome Biology '24]

- Gagandeep Singh, Mohammed Alser, Kristof Denolf, [Can Firtina](#), Alireza Khodamoradi, Meryem Banu Cavlak, Henk Corporaal and Onur Mutlu, [**"RUBICON: A Framework for Designing Efficient Deep Learning-Based Genomic Basecallers"**](#) *Genome Biology*, February 2024.
[[arXiv version](#)]
[[Online link at Genome Biology Journal](#)]
[[RUBICON Source Code](#)]

Method | [Open access](#) | Published: 16 February 2024

RUBICON: a framework for designing efficient deep learning-based genomic basecallers

[Gagandeep Singh](#), [Mohammed Alser](#), [Kristof Denolf](#), [Can Firtina](#) ✉, [Alireza Khodamoradi](#), [Meryem Banu Cavlak](#), [Henk Corporaal](#) & [Onur Mutlu](#) ✉

[Genome Biology](#) **25**, Article number: 49 (2024) | [Cite this article](#)

Eliminating Useless Basecalling [Frontiers In Genetics '24]

- M. Banu Cavlak, Gagandeep Singh, Mohammed Alser, Can Firtina, Joel Lindegger, Mohammad Sadrosadati, Nika Mansouri Ghiasi, Can Alkan, and Onur Mutlu,
"TargetCall: Eliminating the Wasted Computation in Basecalling via Pre-Basecalling Filtering"
Frontiers in Genetics, October 2024.
[[Online link at Frontiers in Genetics Journal](#)]
[[arXiv Version](#)]
[[Talk Video at BIO-Arch 2023 Workshop](#)]
[[TargetCall Source Code](#)]

TargetCall: eliminating the wasted computation in basecalling via pre-basecalling filtering

Meryem Banu Cavlak^{1*}, Gagandeep Singh¹, Mohammed Alser¹,
Can Firtina^{1*}, Joël Lindegger¹, Mohammad Sadrosadati¹,
Nika Mansouri Ghiasi¹, Can Alkan² and Onur Mutlu^{1*}

¹SAFARI Research Group, Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland, ²Department of Computer Engineering, Bilkent University, Ankara, Türkiye

Basecalling 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

DNN Hardware Acceleration

DNN execution is dominated by:

Vector-Matrix
Multiplication (VMM)

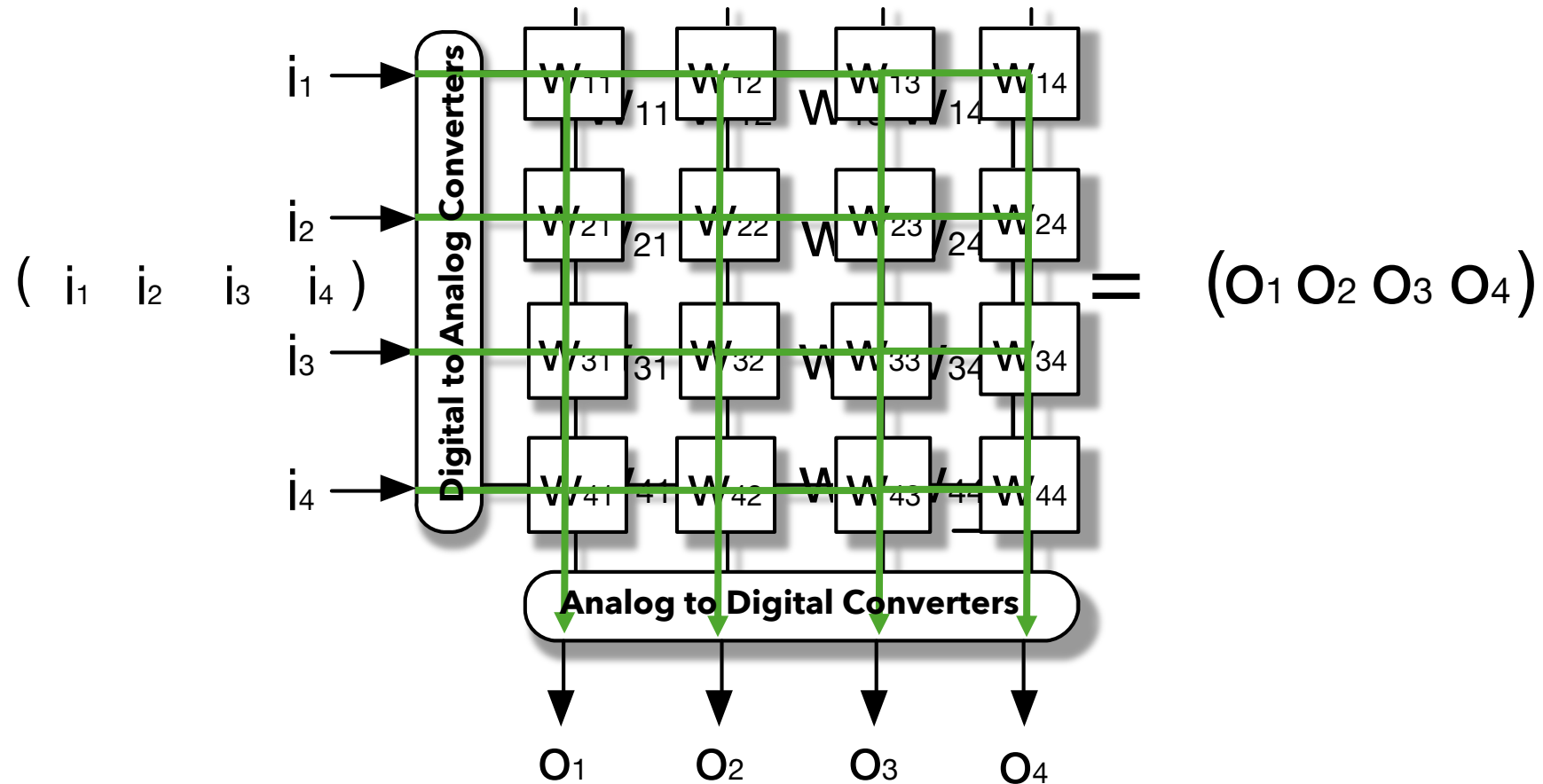
Data movement
between memory and
accelerator

Memristor-based PIM for DNN Acceleration

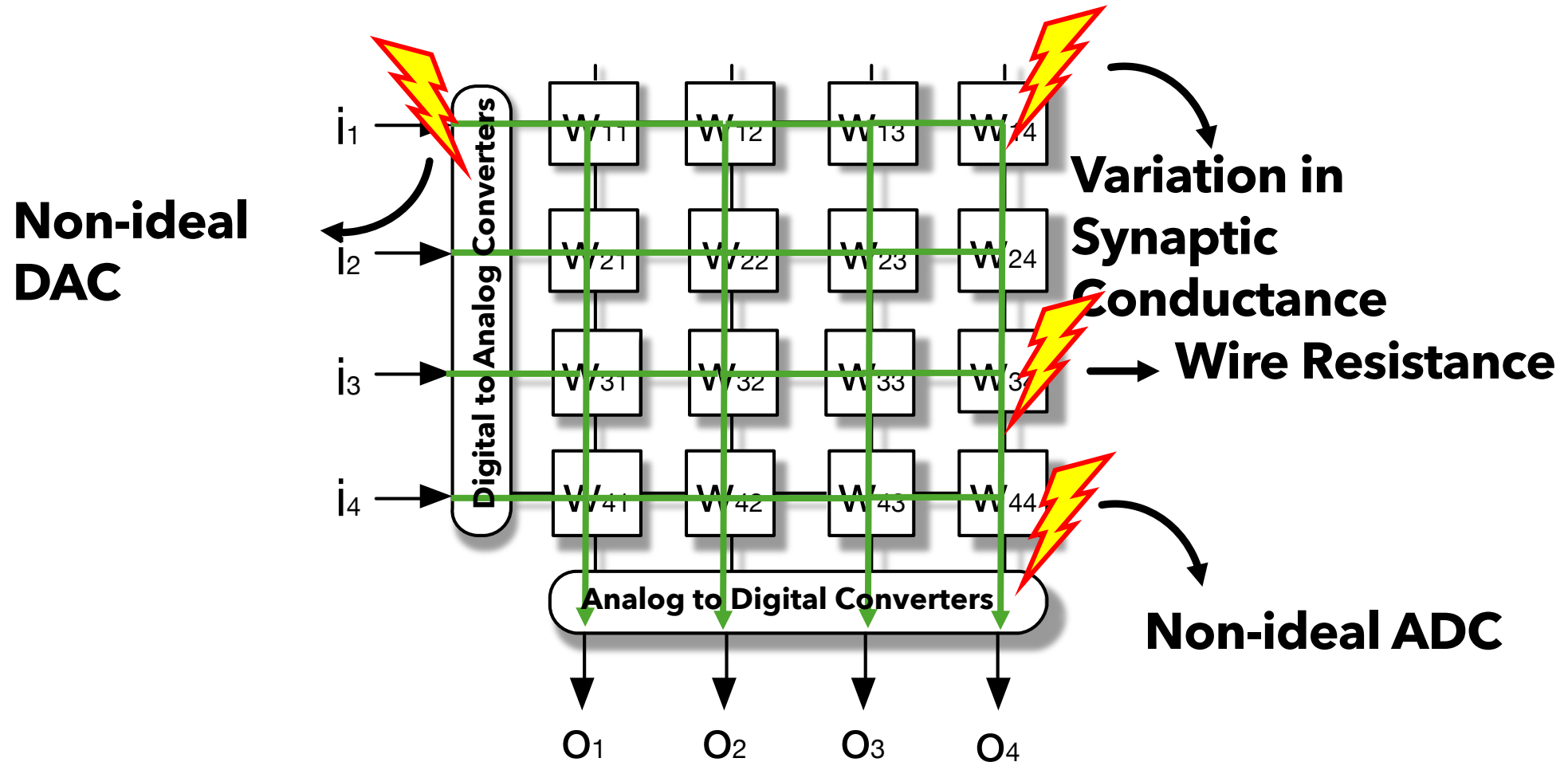
Memristor-based
crossbars support
VMM

Processing in Memory
(PIM) minimizes data
movement

VMM in Memristor-based Crossbars

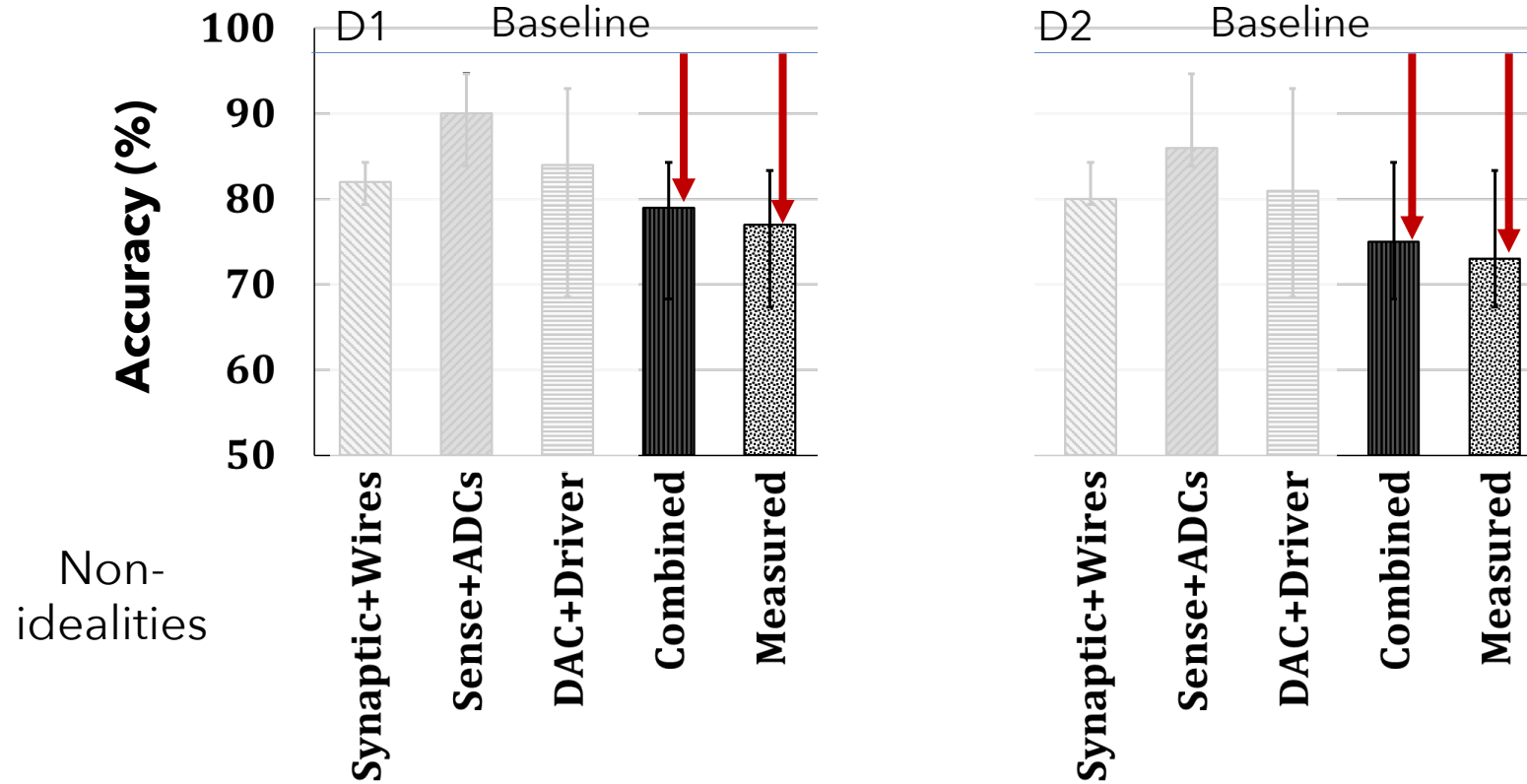


Non-idealities in Memristor-based Crossbars



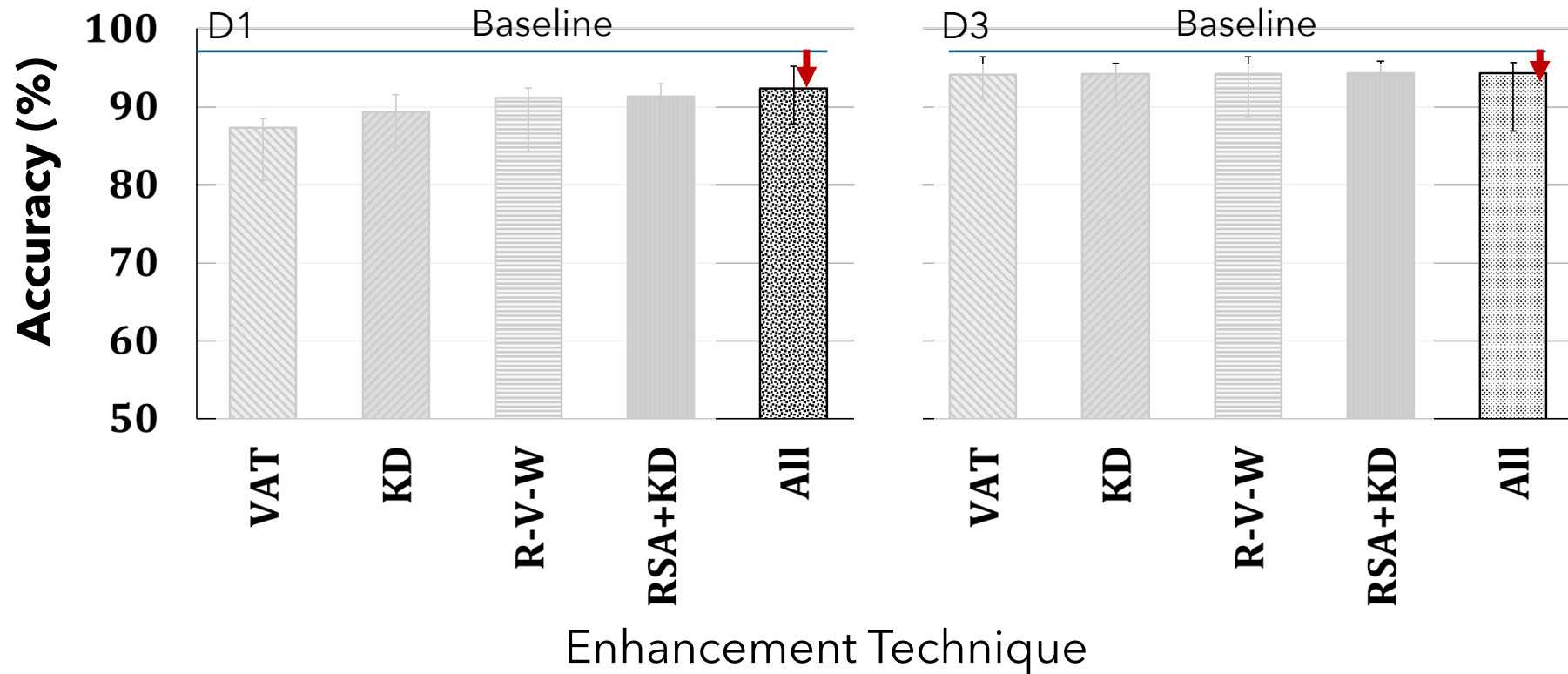
Subpar utilization of trained models
due to non-idealities

Accuracy: All Non-idealities without Mitigation



Combined non-idealities leads to significant accuracy loss (**>18%**)

Accuracy: Enhancement Techniques on All Non-idealities



Accuracy loss drops to 6% from 18% after handling all non-idealities

ML-based Genome Analysis via PIM [MICRO '22]

- 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](#)]

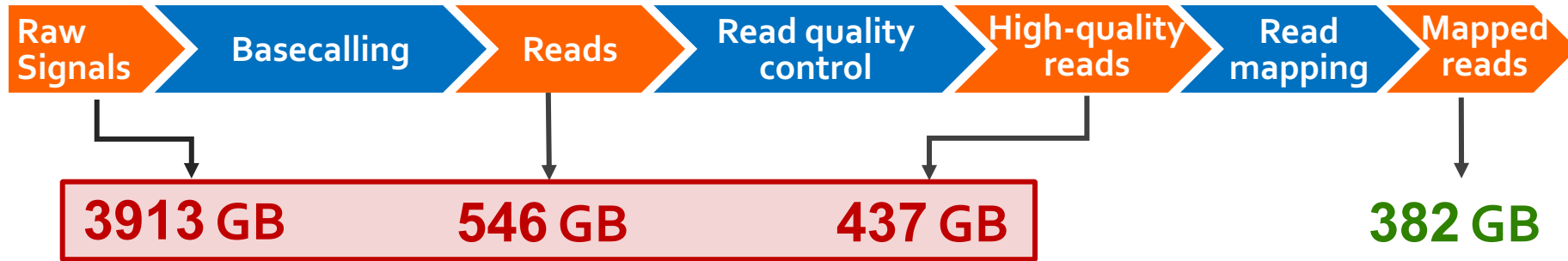
GenPIP: In-Memory Acceleration of Genome Analysis via Tight Integration of Basecalling and Read Mapping

Haiyu Mao¹ Mohammed Alser¹ Mohammad Sadrosadati¹ Can Firtina¹ Akanksha Baranwal¹
Damla Senol Cali² Aditya Manglik¹ Nour Almadhoun Alserr¹ Onur Mutlu¹

¹*ETH Zürich*

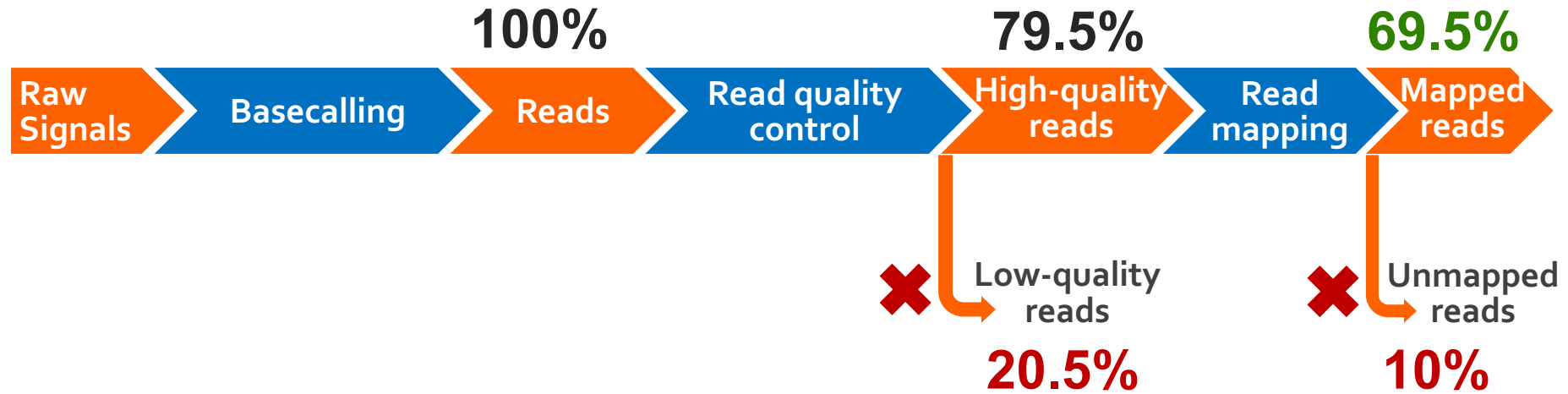
²*Bionano Genomics*

Problem 1: Large Data Movement



Large data movement between genome analysis steps

Problem 2: Wasted Computation



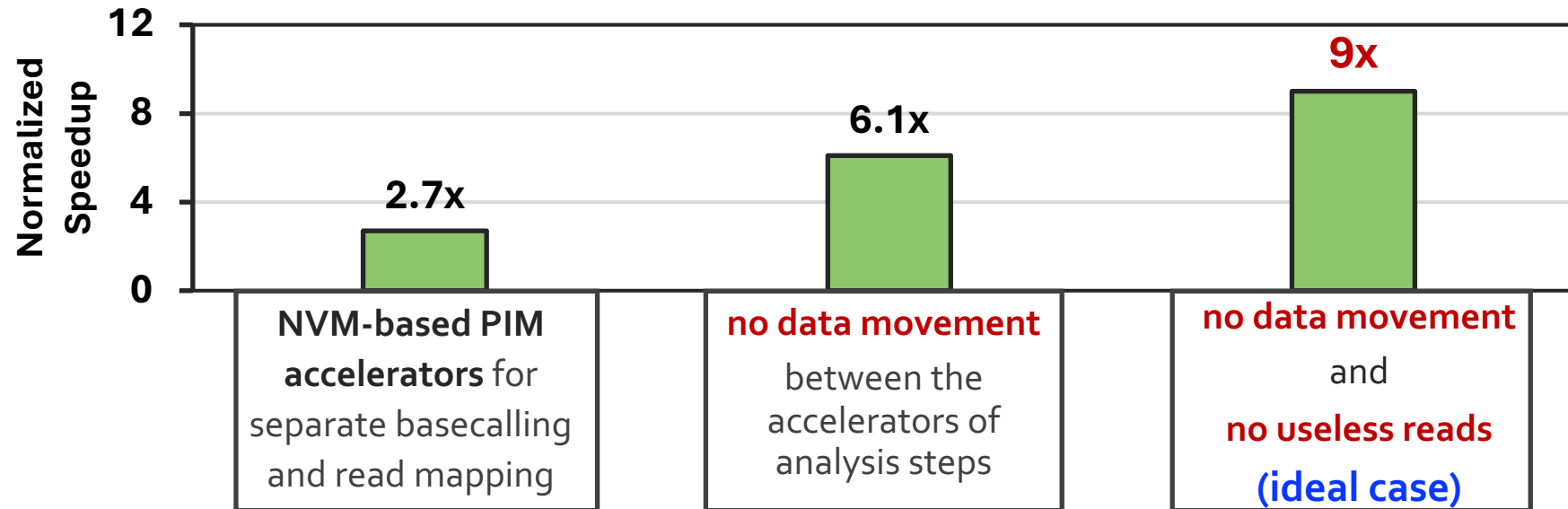
A considerable amount of computation on **useless data** due to

- Low-quality reads
- Unmapped reads

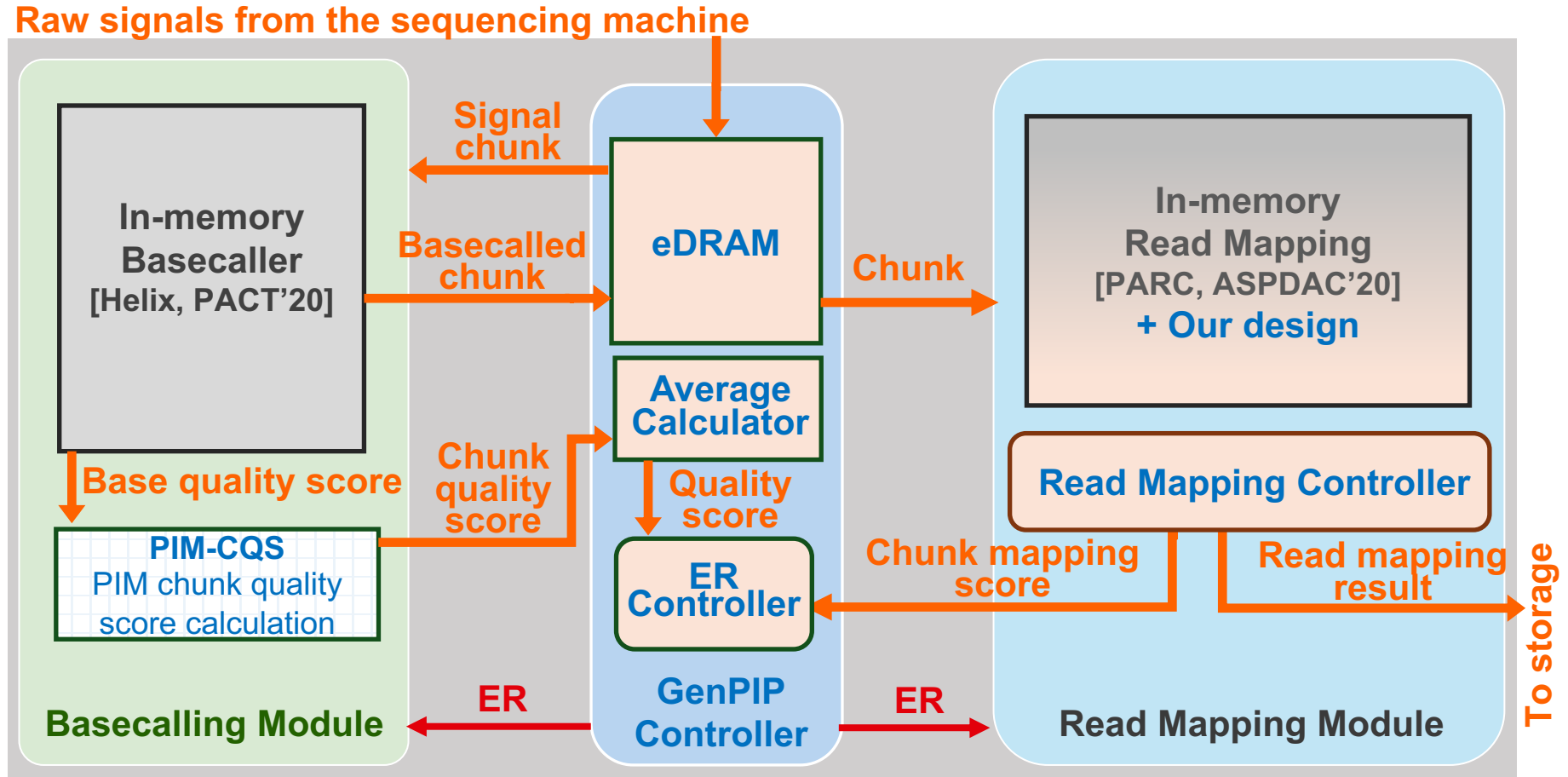
Goal and Opportunities

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

We perform a study to quantify potential performance benefits



GenPIP Overview



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

Fast, Accurate, and Efficient
Genome Analysis
**via Algorithm-Architecture
co-design**

Accelerating String Matching [MICRO '20]

- 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, 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.

[[Lightning 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

Accelerating Genome Graphs [ISCA '22]

- Damla Senol Cali, Konstantinos Kanellopoulos, Joel Lindegger, Zulal Bingol, 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.

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

[[arXiv version](#)]

[[SeGraM Source Code and Datasets](#)]

[[Talk Video](#) (22 minutes)]

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

In-Storage Filtering for Genomics [ASPLOS '22]

- 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, Feb 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 '24]

- 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

Accelerating Genome Analysis [DAC '23]

- Onur Mutlu and Can Firtina,
[**"Accelerating Genome Analysis via Algorithm-Architecture Co-Design"**](#)
*Invited Special Session Paper in Proceedings of the [60th Design Automation Conference \(DAC\)](#),
San Francisco, CA, USA, July 2023.*
[\[Related Invited Paper\]](#)
[\[arXiv version\]](#)
[\[Slides \(pptx\) \(pdf\)\]](#)
[\[Talk Video at DAC 2023\]](#) (38 minutes, including Q&A)

Accelerating Genome Analysis via Algorithm-Architecture Co-Design

Onur Mutlu Can Firtina
ETH Zürich

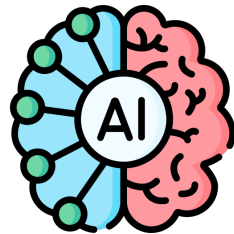
Overview of Our Research Interests

Goal: Enable **Fast, Energy-Efficient, Accurate, and Scalable Genome Analysis**
by Designing Algorithms and Architectures

Today's Talk

Real-Time and Raw Sequencing Data Analysis

[ISMB/ECCB'23]
[Bioinformatics'24]
[IEEE Access'24]
[arXiv'24 & HiTSeq'24]

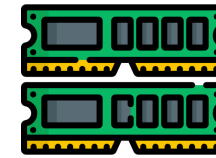
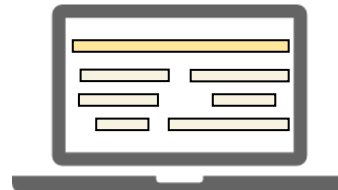


Algorithm & Hardware Design for AI in Genome Analysis

[Bioinformatics'20]
[ACM TACO'24]
[Genome Biology'24]
[Front. in Genetics'24]

Accurate and Fast Algorithm Design for Read Mapping

[Bioinformatics'22]
[NARGAB'23 & RECOMB'23]
[arXiv'23]
[arXiv'23]
[IEEE/ACM TCBB'24]



Data-Centric Architectures for Genome Analysis

[MICRO'20]
[MICRO'22]
[ISCA'22]
[ASPLOS'22]
[IEEE Access'22]
[MICRO'23]
[ISCA'24]



Reviews

[CSBJ'22]
[DAC'23]
[Nature Protocols'24]

Today's Talk



Significant computation and energy overhead



Limited accuracy and application scope

**Algorithm & Hardware Design
for AI in Genome Analysis**



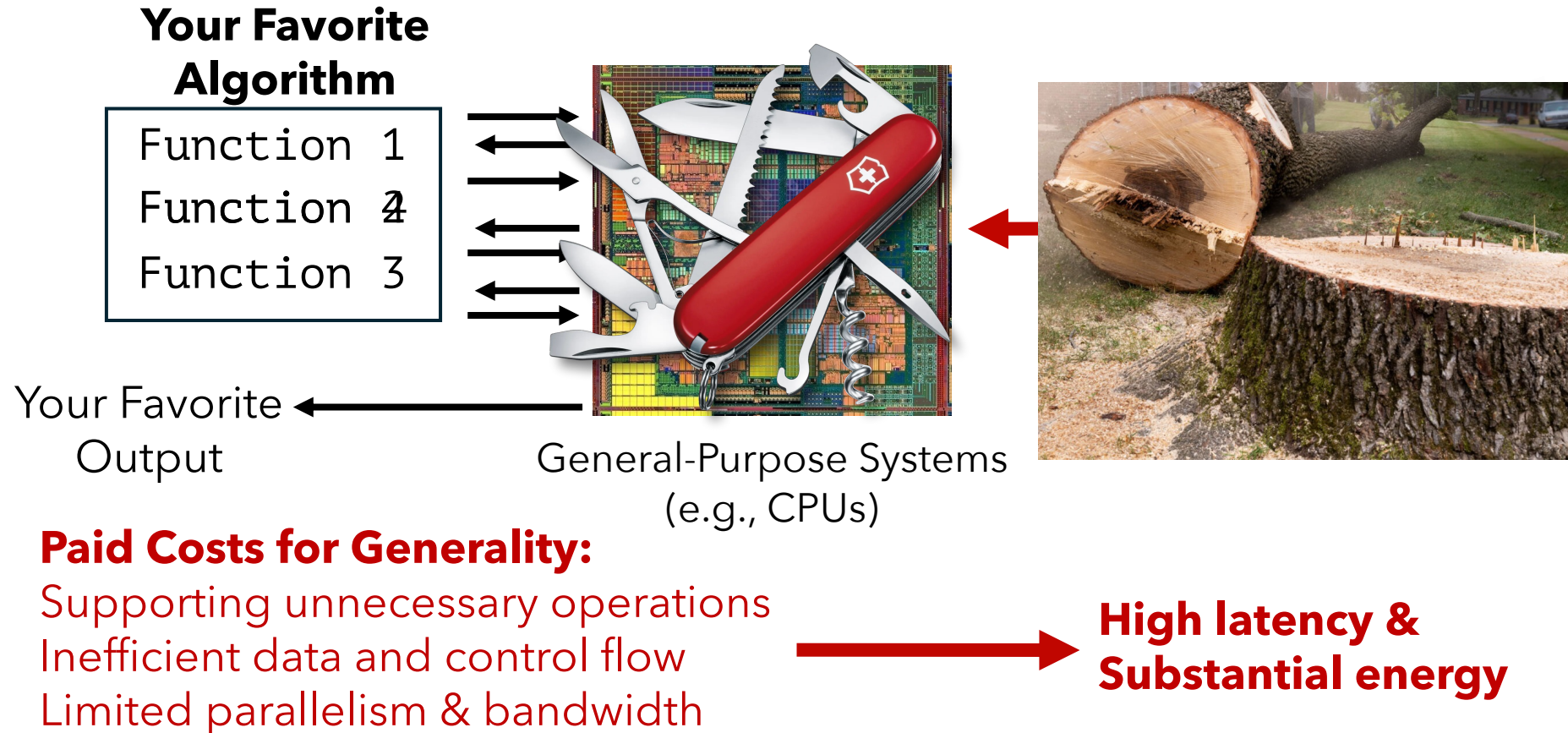
**ApHMM
(TACO'24 &
HiPEAC'24)**

***Accelerating ML operations
in Genome Graphs***

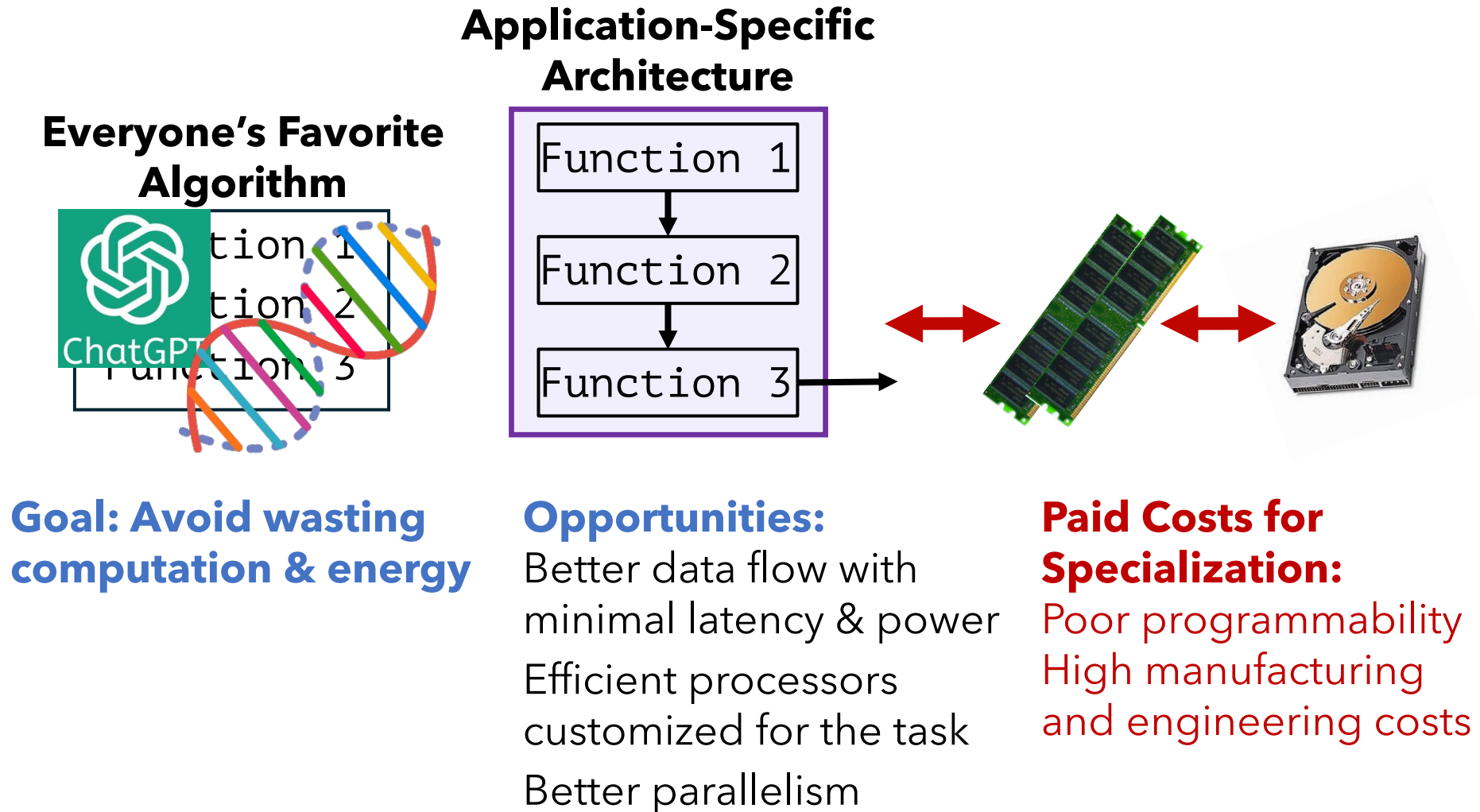
General-Purpose "Things"



Algorithm Design Demands General-Purpose Computing



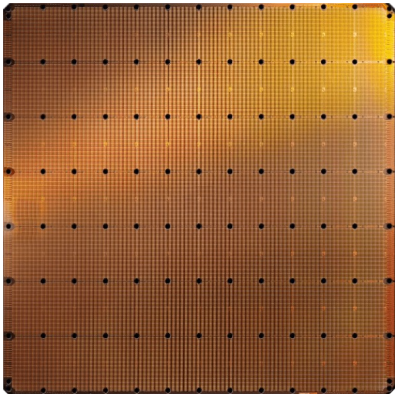
Great Algorithms Demand Specialization



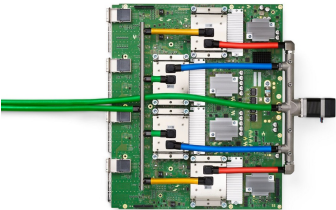
Acceleration Efforts for AI and Genomics

Cerebras WSE-3

The largest ML accelerator chip



Google's TPUs:



NVIDIA H100



Scale your studies with ease

Process high-throughput data quickly with hardware acceleration

Process high-throughput data quickly with hardware acceleration. With four field-programmable gate arrays (FPGAs) onboard you have the most powerful DRAGEN analysis ever enabling you to process NovaSeq X System data easily. Perform up to four simultaneous applications per flow cell in a single run.

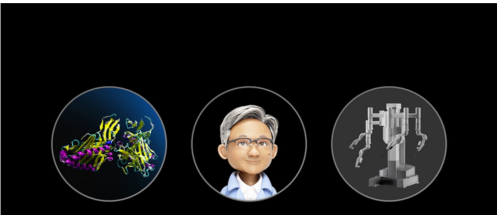
Reduce data footprint, manage and store data easily with lower costs and lower energy consumption, with built-in compression that reduces FASTQ file sizes by up to 80%.

Stream data directly to Illumina Connected Analytics or BaseSpace Sequence Hub on the cloud for scalable data management, analysis, and aggregation.

NVIDIA Partners With Industry Leaders to Advance Genomics, Drug Discovery and Healthcare

IQVIA, Illumina, Mayo Clinic and Arc Institute Harness NVIDIA AI and Accelerated Computing to Transform \$10 Trillion Healthcare and Life Sciences Industry

January 13, 2025



Absci and AMD Announce Collaboration and Strategic Investment to Accelerate the Future of AI Drug Discovery

01/08/2025

PDF Version

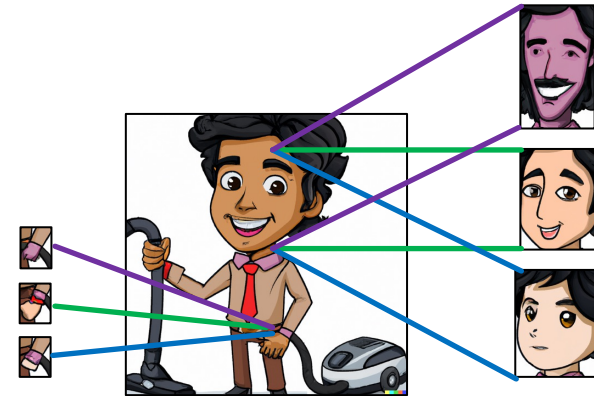
Important and Commonly Used Applications Demand Specialization

Probabilistic Graphs in Genomics

Probabilistic graphs are commonly used **to identify variations** between sequences

To avoid **redundant** comparisons and storage

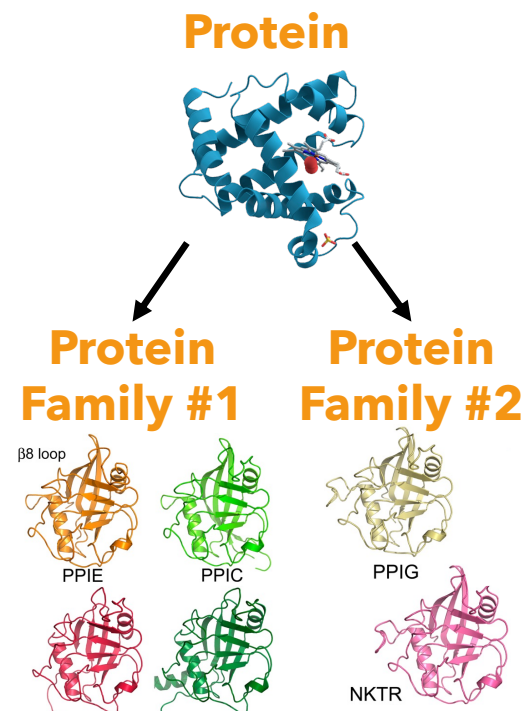
To provide **rich information** on expected variations



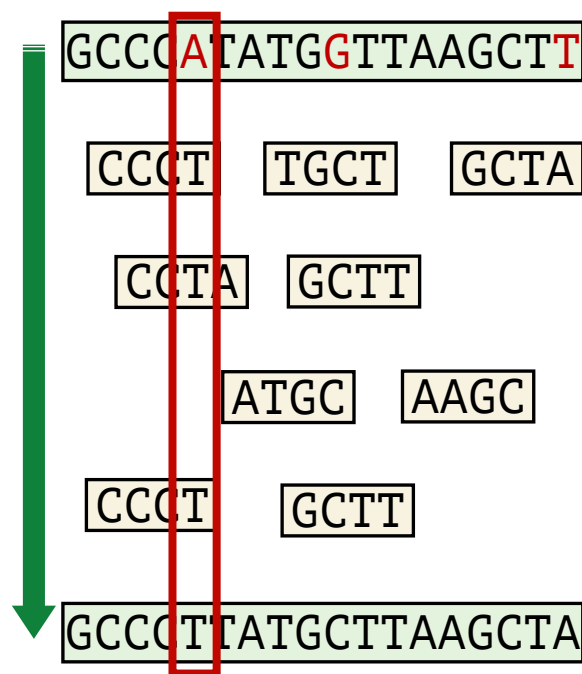
Reference Genome
Graph

Graph-Based Applications in Genomics

Protein Family Search with Graph Alignment



Error Correction with Probabilistic Graphs

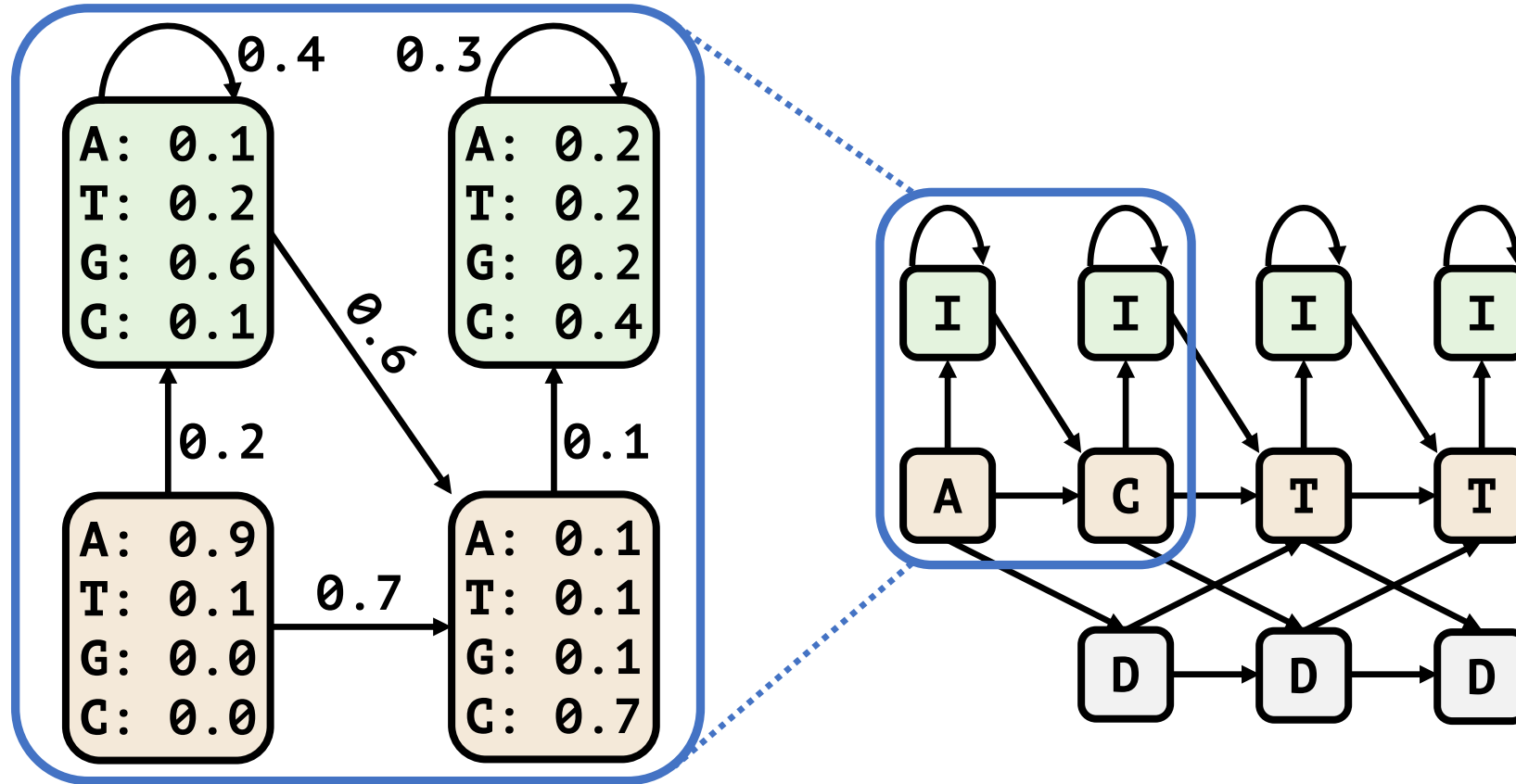


Multiple Sequence Alignment

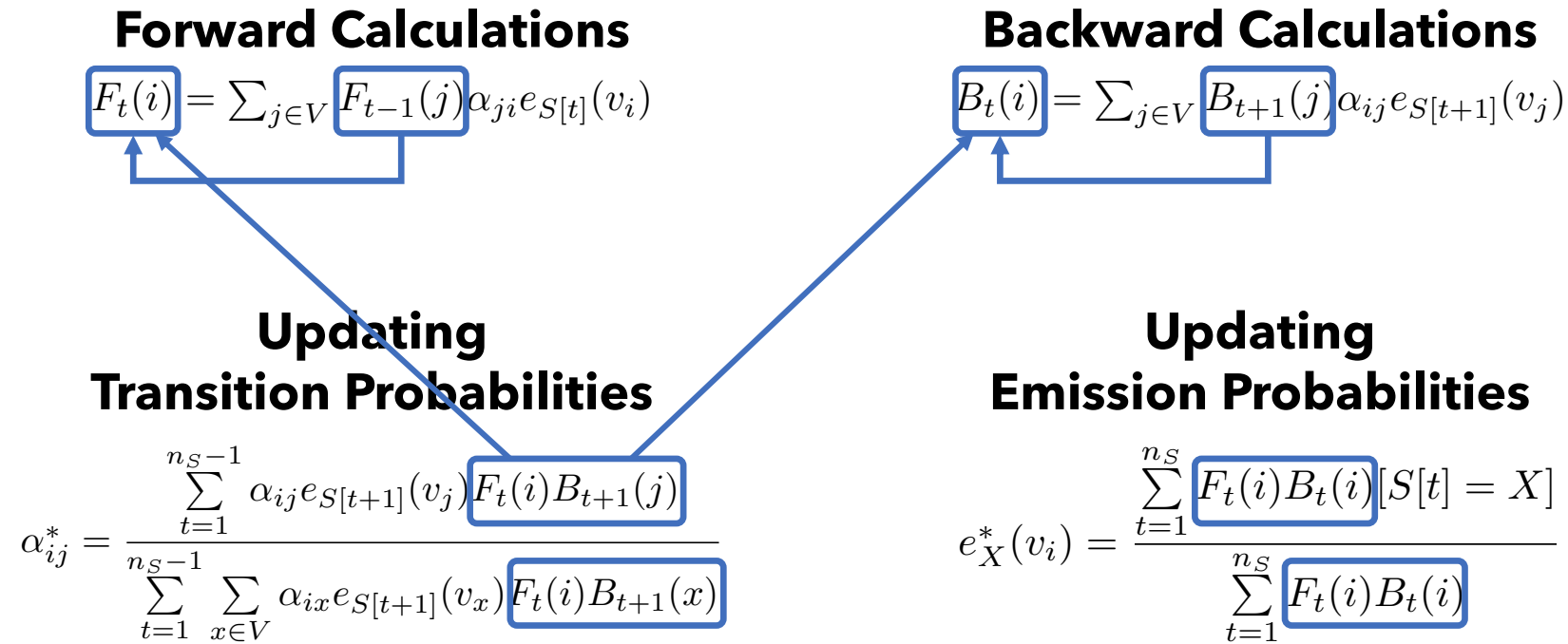


A Common Graph Structure in Genomics

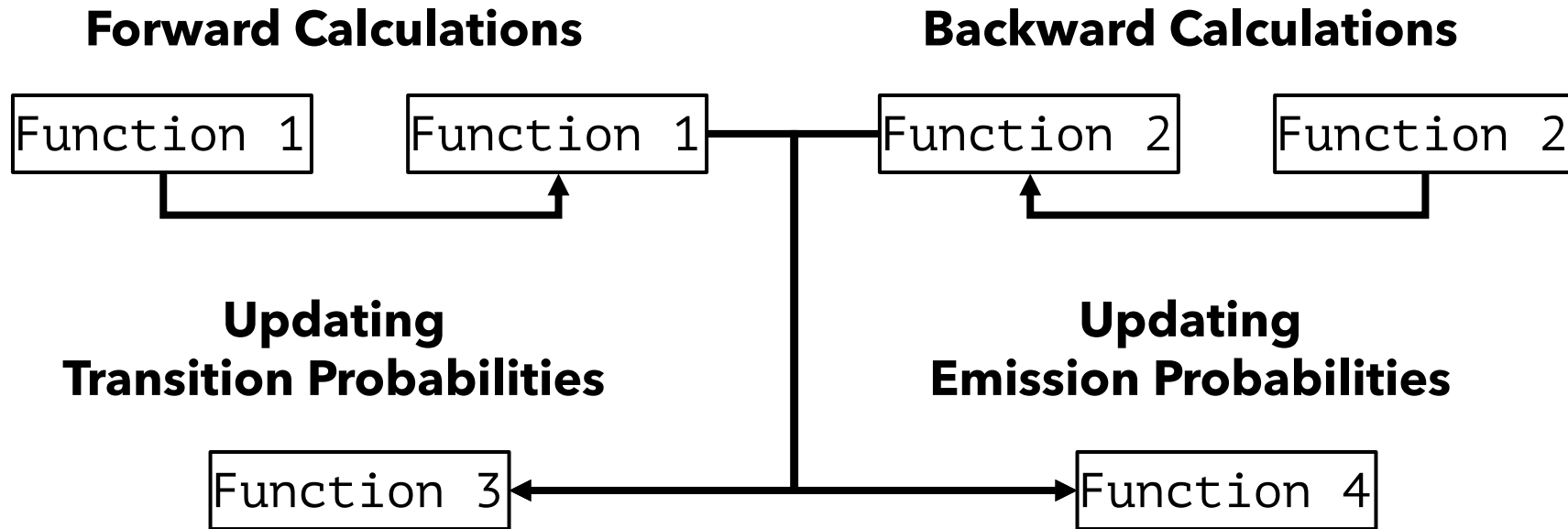
Profile Hidden Markov Models (PHMMs):
A probabilistic graph structure



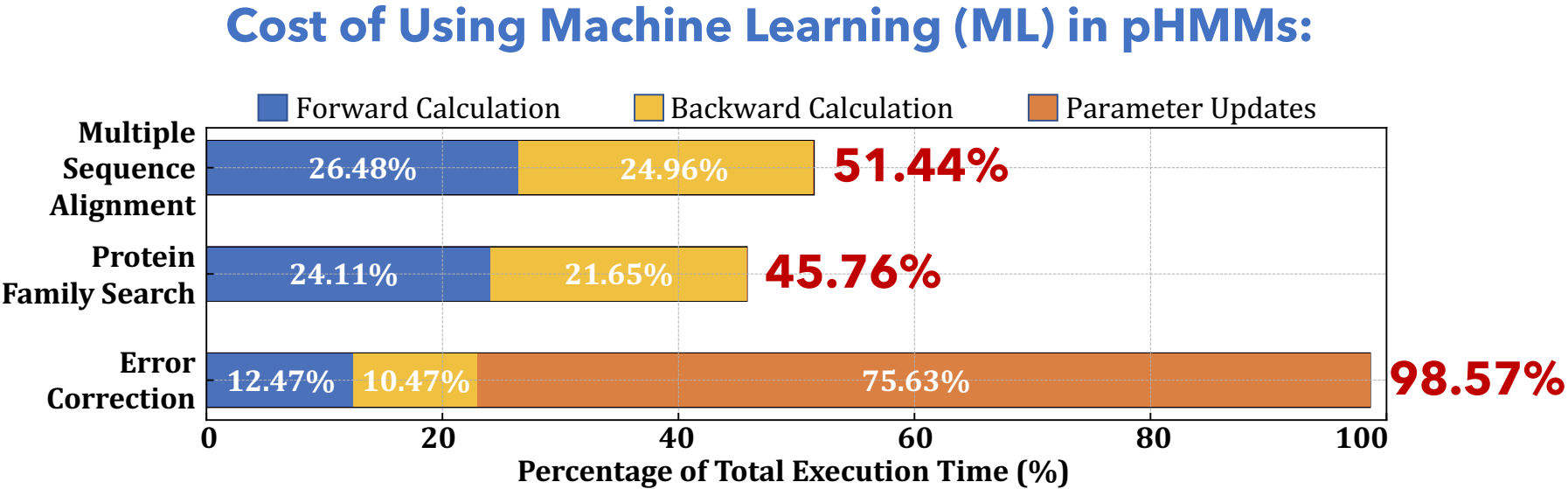
Underlying ML Technique in Genomic Graphs



Underlying ML Technique in Genomic Graphs



Costly Use of ML in Genomic Graphs



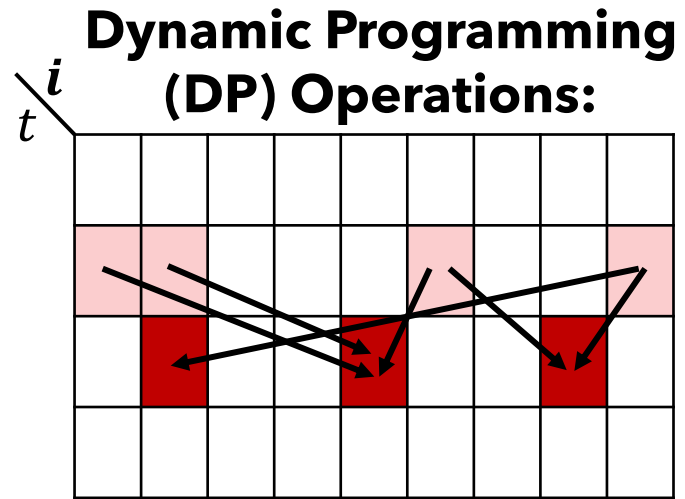
Goal

Enable **rapid and power-efficient**
use of the underlying ML technique in genome graphs

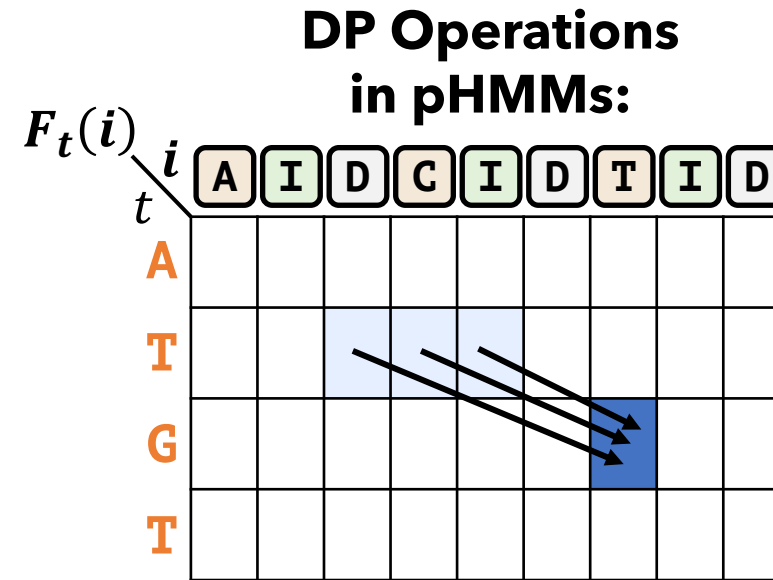
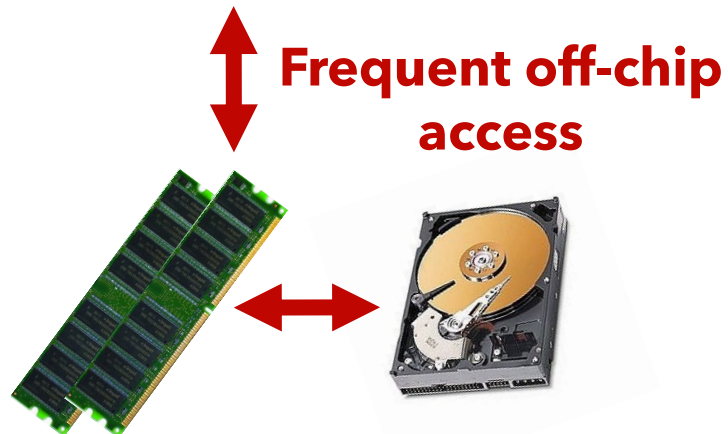
Key Approach: HW-SW Co-Design

- **Reduce redundant data storage** by utilizing the fixed data pattern
 - **Reduce unnecessary computations** with quick filtering **SW**
 - **Avoid repeated operations** by utilizing lookup tables
-
- **Reduce data movement** by processing data where it makes sense **HW**
 - **Flexible and efficient** hardware design

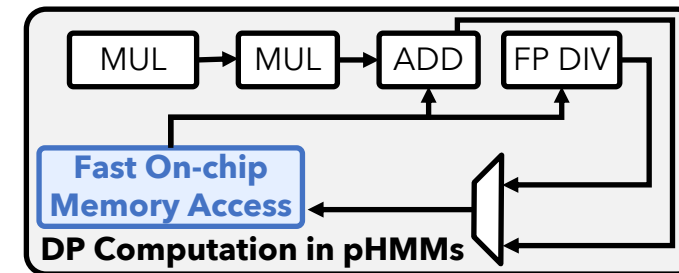
HW: Simple Data Pattern in DP Calculations



Many Different Patterns



Simple Pattern



Store and access data
where it makes sense

SW: Reducing Unnecessary Computations

Observation:
Calculating the entire DP
is not needed

$F_t(i)$

$t \backslash i$	A	I	D	C	I	D	T	I	D
A									
T									
G									
T									

Dense DP Calculations

**Filter by
sorting**



$F_t(i)$

$t \backslash i$	A	I	D	C	I	D	T	I	D
A									
T									
G									
T									

Sparse (Filtered) DP Calculations

Challenge:
Sorting is costly in hardware

SW: Reducing Unnecessary Computations

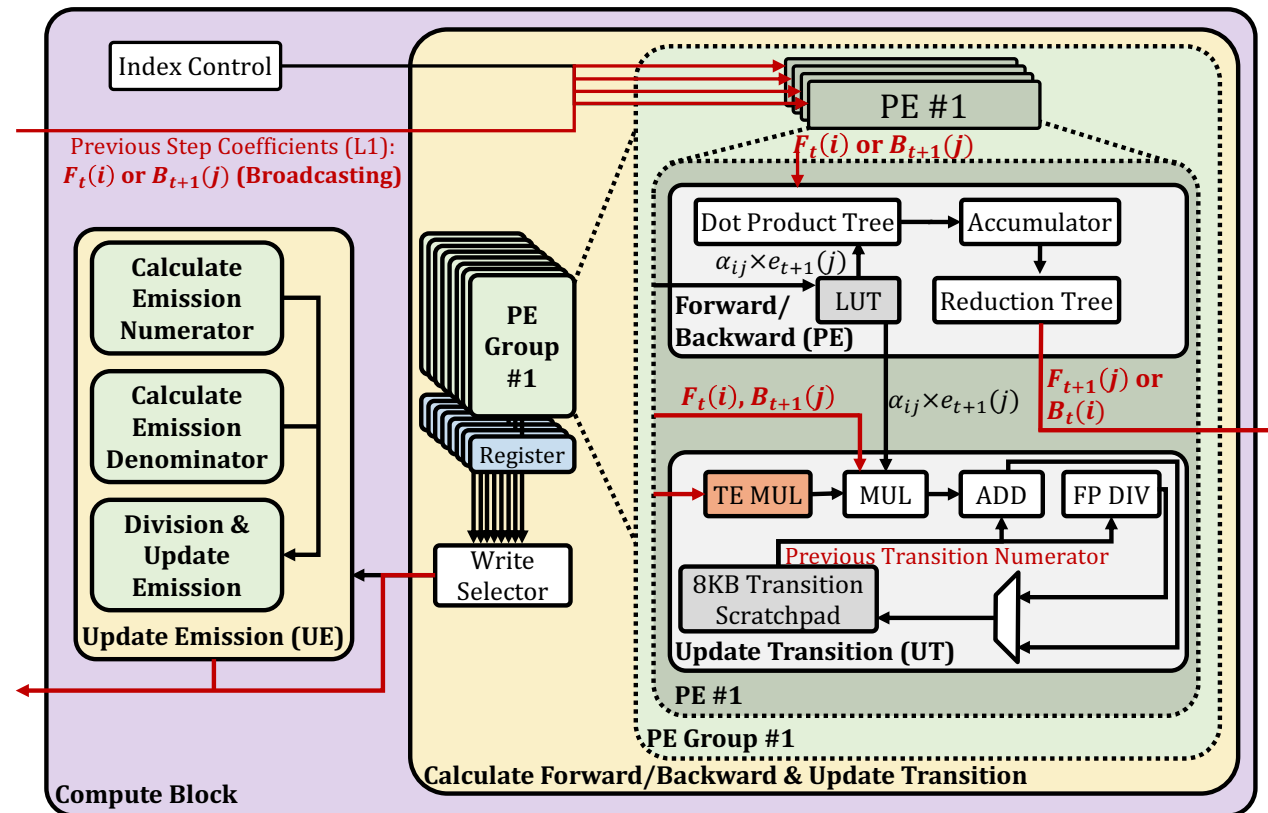
Software co-design:

Replace sorting with a cheaper filtering

State IDs	Range
8, 9	1.00 - 0.94
10, 14	0.94 - 0.88
15, 16, 18	0.88 - 0.82
11, 20, 21, ...	0.82 - 0.76
13, 17, 19, ...	0.76 - 0.70
...	...
...	...
...	...
...	0.06 - 0.00
Histogram Filter	

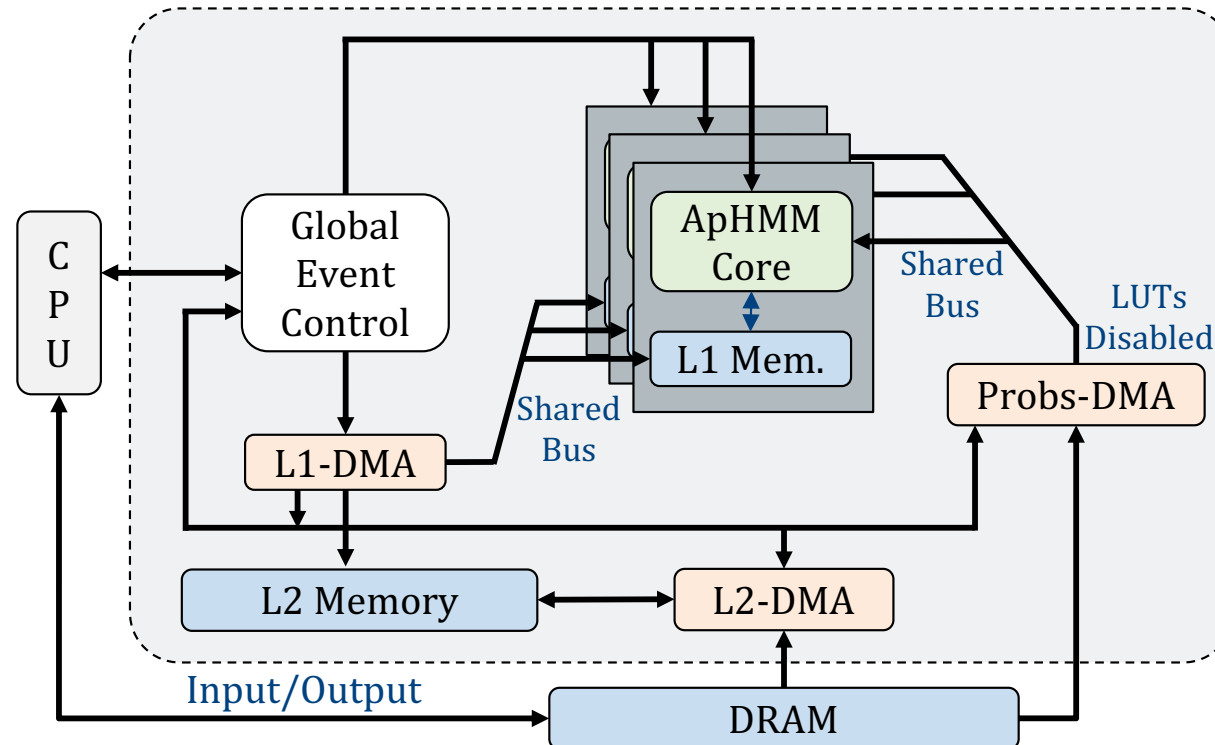
Putting the Optimizations Together

Effectively implementing the hardware and software optimizations
to improve performance and reduce power requirements

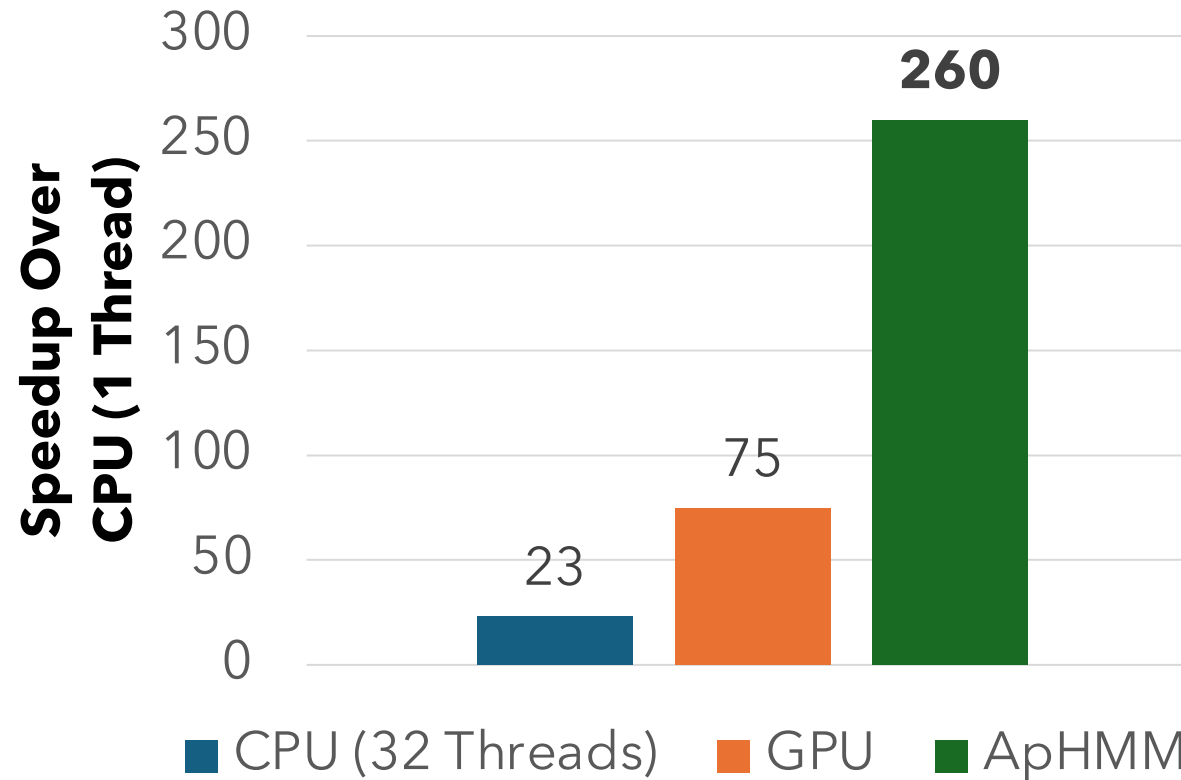


Integrating ApHMM in a Complete System

The specialized task can be **offloaded** to an accelerator **whenever needed**

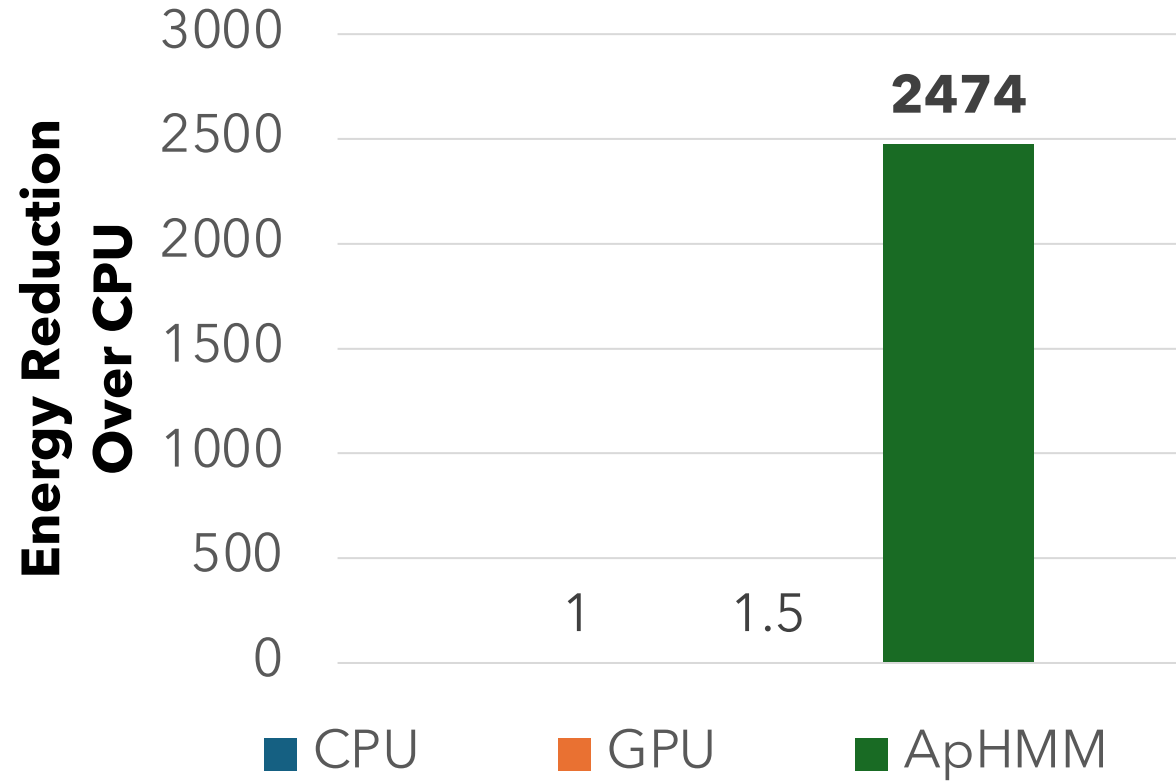


Key Results: Performance



Specialized design enables computation
up to two orders of magnitude faster:
Better data flow with reduced latency and improved parallelism

Key Results: Energy Consumption



...and a significant reduction in energy consumption
by up to three orders of magnitude:
simpler hardware with a minimized off-chip memory accesses
and data movement

Specialization Provides
Significant Performance
and Energy Benefits

Today's Talk



Significant computation and energy overhead



Limited accuracy and application scope

**Algorithm & Hardware Design
for AI in Genome Analysis**



**ApHMM
(TACO'24 &
HiPEAC'24)**

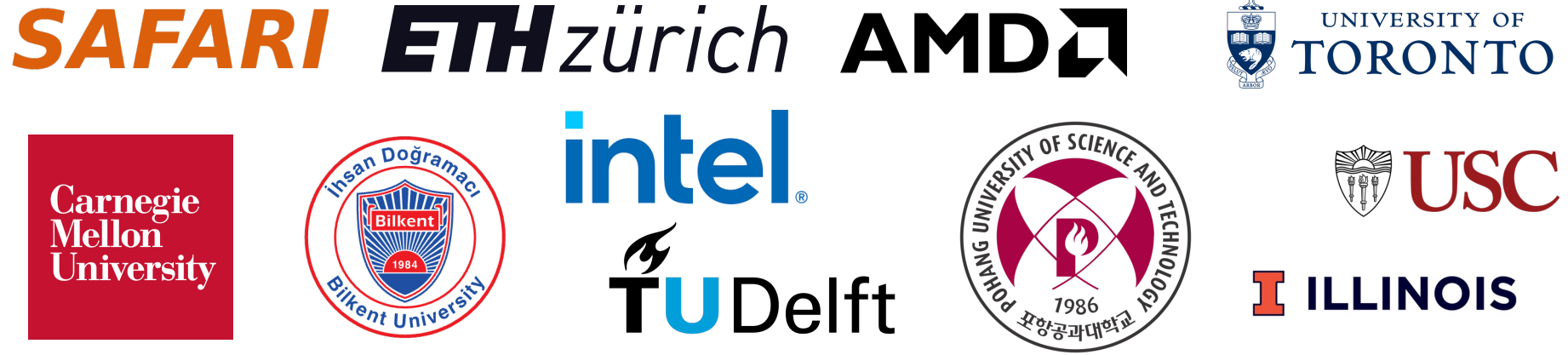
***Accelerating ML operations
in Genome Graphs***

Summary of Key Takeaways

- There are **many ways to utilize AI in genomics** (and life sciences more broadly)
 - We only scratched the surface of those applications today
- It is critical to **consider the requirements of your application**
 - Portable? Low latency? Highest accuracy?
- It is critical to **consider the limitations of your system**
 - And integrate these limitations in the analysis
- **Data movement** will continue to be a major system bottleneck
 - Generating biological data is getting faster and cheaper
- **Many future opportunities exists with emerging technologies**

Acknowledgements

- **Collaborators:**



- **Funding agencies and industry sponsors:**

- BioPIM, SNSF, Intel, Google, Huawei, Microsoft, VMware, and SRC

Accelerating AI-Driven Genome Analysis

via Algorithm-Architecture Co-Design

Can Firtina | canfirtina@gmail.com

Senior Researcher and Lecturer at ETH Zurich

Incoming Assistant Professor of Computer Science at UMD

June 16, 2025

PASC25, Challenges in Systems Design for Omics, Switzerland

ETH zürich

SAFARI



UNIVERSITY OF
MARYLAND