# Tutorial on Memory-Centric Computing: Conclusion Remarks

Geraldo F. Oliveira

Prof. Onur Mutlu

ISCA 2024

29 June 2024

SAFARI

ETH zürich

# Agenda

- Introduction to Memory-Centric Computing Systems
- Invited Talk by Prof. Minsoo Rhu:
  "*Memory-Centric Computing Systems – For AI and Beyond*"
- Coffee Break
- Real-World Processing-Near-Memory Systems
- Processing-Using-Memory Architectures for Bulk Bitwise Op.
- Invited Talk by Prof. Saugata Ghose:
  "*RACER and ReRAM PUM*"
- PIM Programming & Infrastructure for PIM Research
- Closing Remarks

# Challenge and Opportunity for Future

Fundamentally

Energy-Efficient

**(Data-Centric)**

Computing Architectures

# Challenge and Opportunity for Future

Fundamentally

High-Performance

**(Data-Centric)**

Computing Architectures

# Challenge and Opportunity for Future

Computing Architectures

with

Minimal Data Movement

# Concluding Remarks

- We must design systems to be **balanced**, **high-performance**, **energy-efficient** (all at the same time) → intelligent systems
  - **Data-centric, data-driven, data-aware**

- Enable computation capability inside and close to memory

- This can
  - Lead to **orders-of-magnitude** improvements
  - **Enable new applications & computing platforms**
  - **Enable better understanding of nature**
  - **…**

- Future of **truly memory-centric computing** is bright
  - We need to do research & design across the computing stack

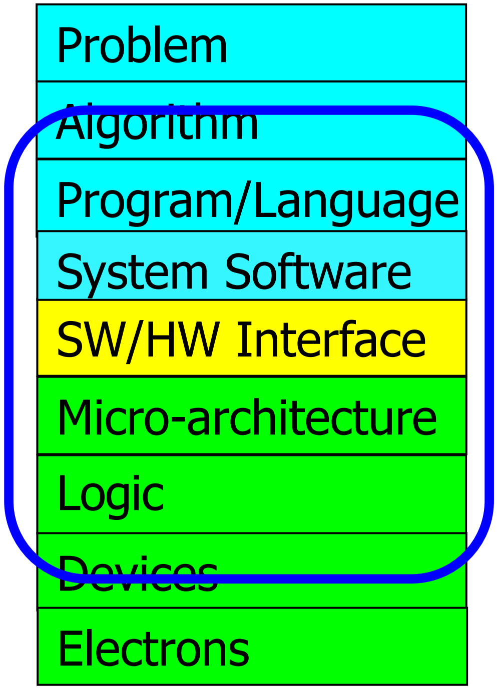
# Fundamentally Better Architectures

**Data-centric**

**Data-driven**

**Data-aware**

# We Need to Revisit the Entire Stack



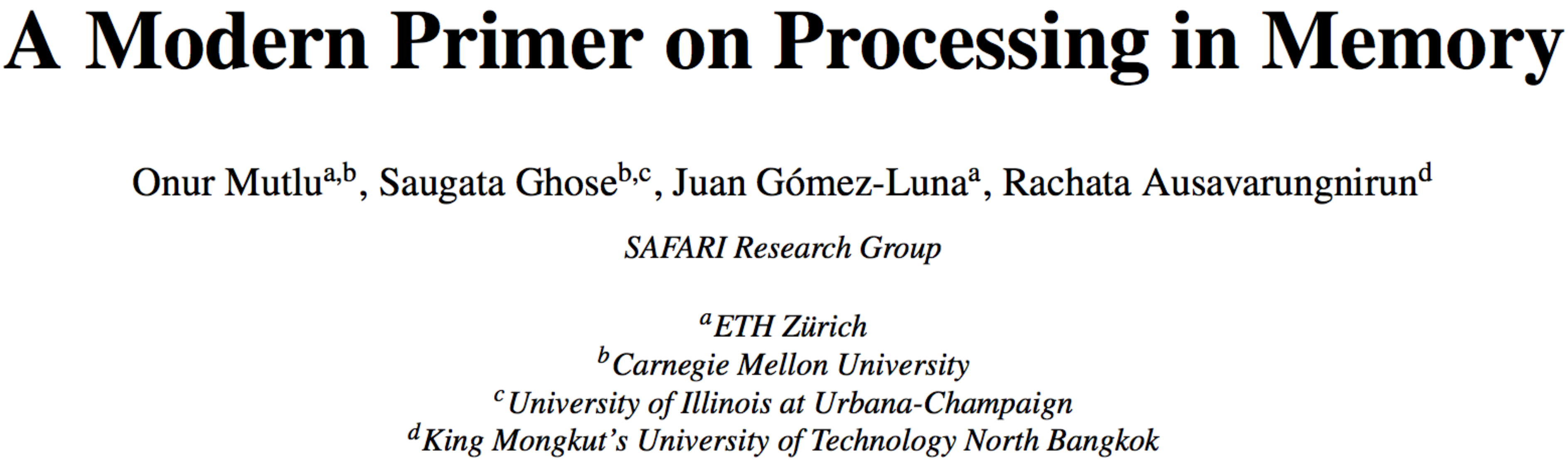
**We can get there step by step**

# We Need to Exploit Good Principles

- Data-centric system design
- All components intelligent
- Better (cross-layer) communication, better interfaces
- Better-than-worst-case design
- Heterogeneity
- Flexibility, adaptability

**Open minds**

# 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*

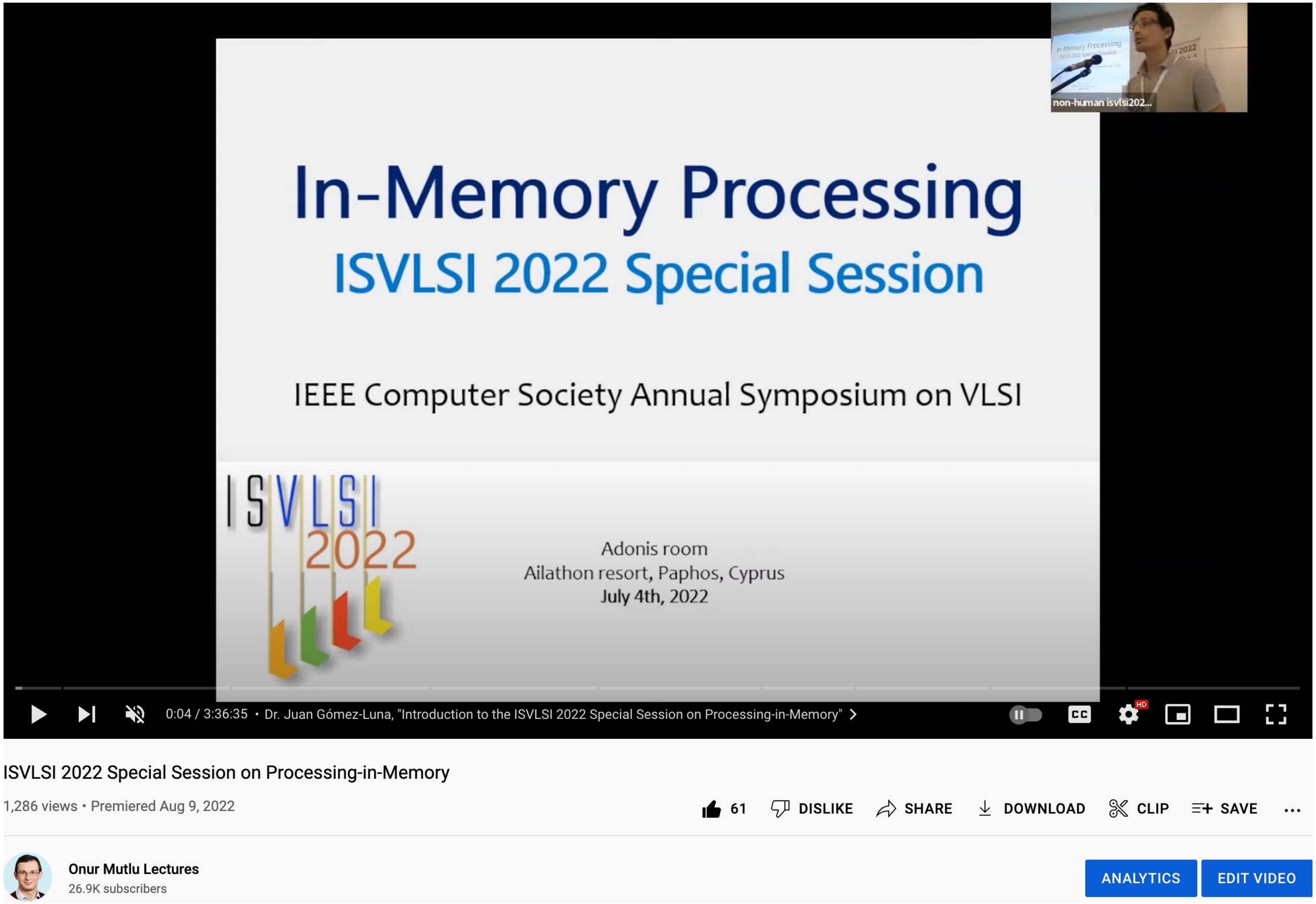
[a] *ETH Zürich*
[b] *Carnegie Mellon University*
[c] *University of Illinois at Urbana-Champaign*
[d] *King Mongkut's University of Technology North Bangkok*

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**

# Special Research Sessions & Courses (I)

- 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

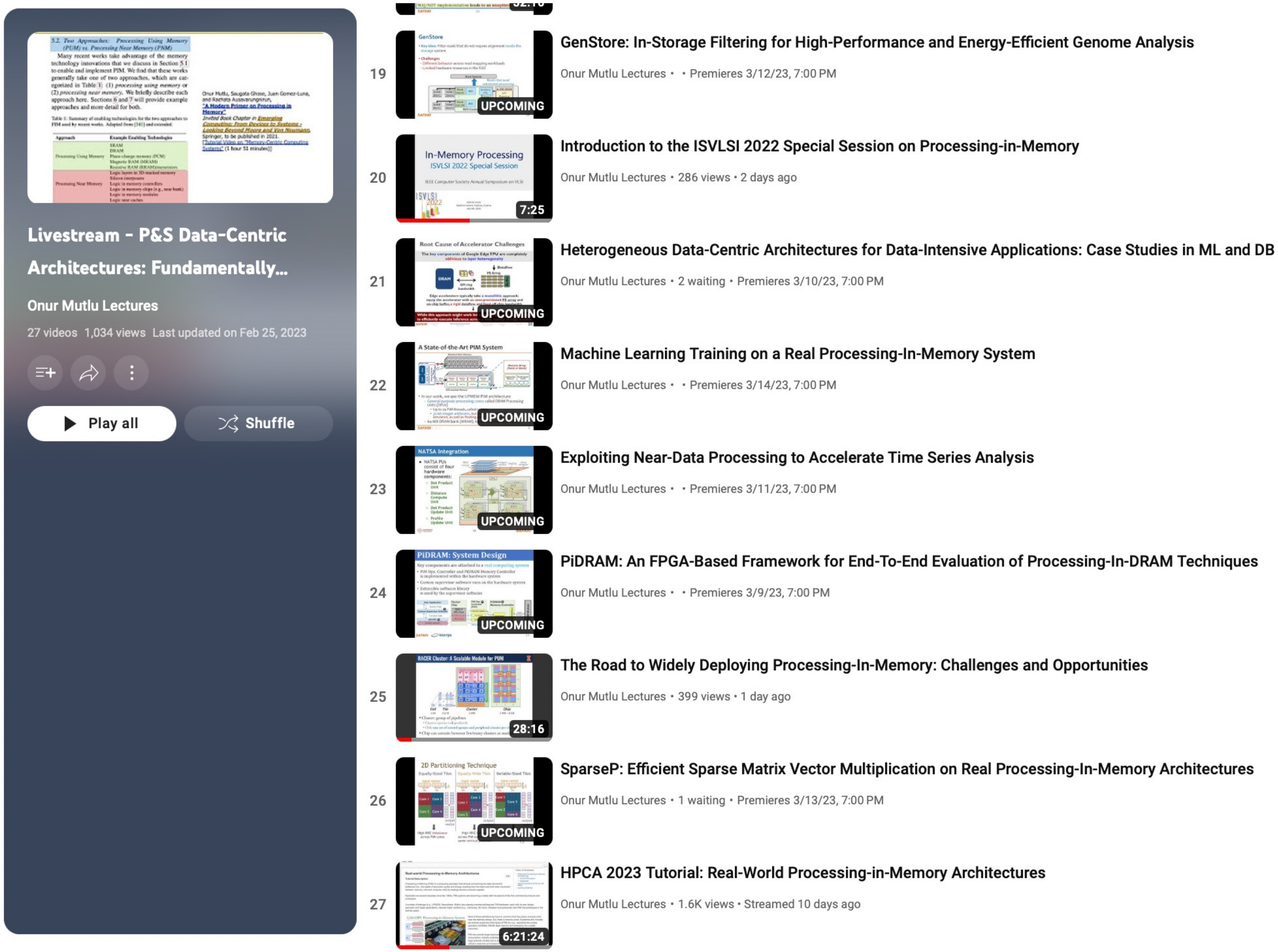
1,286 views • Premiered Aug 9, 2022

Onur Mutlu Lectures
26.9K subscribers

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

# Special Research Sessions & Courses (II)

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



**Livestream - P&S Data-Centric Architectures: Fundamentally...**

Onur Mutlu Lectures

27 videos 1,034 views Last updated on Feb 25, 2023

19 GenStore: In-Storage Filtering for High-Performance and Energy-Efficient Genome Analysis
Onur Mutlu Lectures • • Premieres 3/12/23, 7:00 PM
UPCOMING

20 Introduction to the ISVLSI 2022 Special Session on Processing-in-Memory
Onur Mutlu Lectures • 286 views • 2 days ago
7:25

21 Heterogeneous Data-Centric Architectures for Data-Intensive Applications: Case Studies in ML and DB
Onur Mutlu Lectures • 2 waiting • Premieres 3/10/23, 7:00 PM
UPCOMING

22 Machine Learning Training on a Real Processing-In-Memory System
Onur Mutlu Lectures • • Premieres 3/14/23, 7:00 PM
UPCOMING

23 Exploiting Near-Data Processing to Accelerate Time Series Analysis
Onur Mutlu Lectures • • Premieres 3/11/23, 7:00 PM
UPCOMING

24 PiDRAM: An FPGA-Based Framework for End-To-End Evaluation of Processing-In-DRAM Techniques
Onur Mutlu Lectures • • Premieres 3/9/23, 7:00 PM
UPCOMING

25 The Road to Widely Deploying Processing-In-Memory: Challenges and Opportunities
Onur Mutlu Lectures • 399 views • 1 day ago
28:16

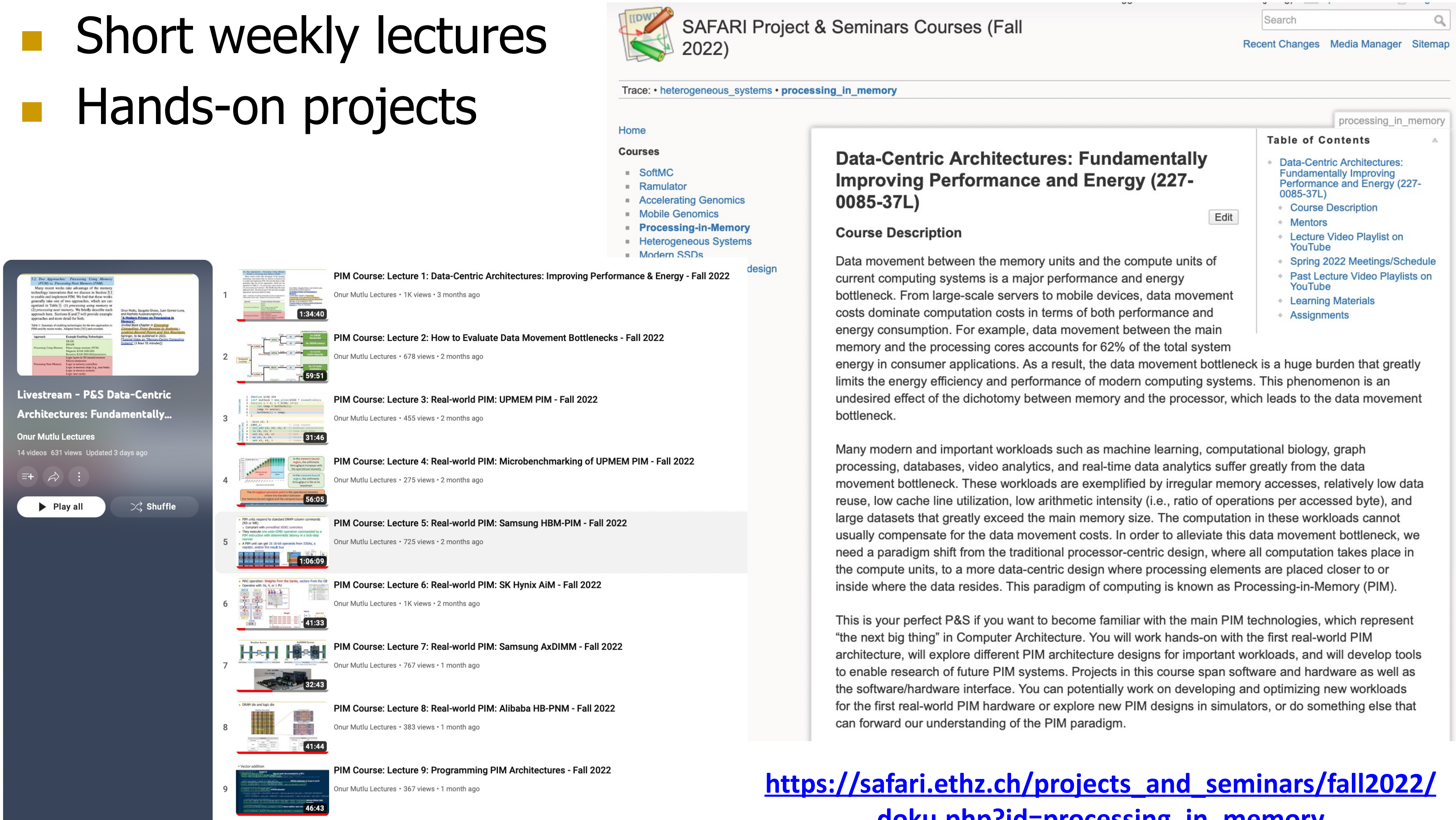
26 SparseP: Efficient Sparse Matrix Vector Multiplication on Real Processing-In-Memory Architectures
Onur Mutlu Lectures • 1 waiting • Premieres 3/13/23, 7:00 PM
UPCOMING

27 HPCA 2023 Tutorial: Real-World Processing-in-Memory Architectures
Onur Mutlu Lectures • 1.6K views • Streamed 10 days ago
6:21:24

https://www.youtube.com/playlist?list=PL5Q2soXY2Zi8KzG2CQYRNQOVD0GOBrnKy

# Processing-in-Memory Course (Fall 2022)

- Short weekly lectures
- Hands-on projects



SAFARI Project & Seminars Courses (Fall 2022)

Trace: • heterogeneous_systems • processing_in_memory

Courses

- SoftMC
- Ramulator
- Accelerating Genomics
- Mobile Genomics
- Processing-in-Memory
- Heterogeneous Systems
- Modern SSDs

## Data-Centric Architectures: Fundamentally Improving Performance and Energy (227-0085-37L)

### Course Description

Data movement between the memory units and the compute units of current computing systems is a major performance and energy bottleneck. From large-scale servers to mobile devices, data movement costs dominate computation costs in terms of both performance and energy consumption. For example, data movement between the main memory and the processing cores accounts for 62% of the total system energy in consumer applications. As a result, the data movement bottleneck is a huge burden that greatly limits the energy efficiency and performance of modern computing systems. This phenomenon is an undesired effect of the dichotomy between memory and the processor, which leads to the data movement bottleneck.

Many modern and important workloads such as machine learning, computational biology, graph processing, databases, video analytics, and real-time data analytics suffer greatly from the data movement bottleneck. These workloads are exemplified by irregular memory accesses, relatively low data reuse, low cache line utilization, low arithmetic intensity (i.e., ratio of operations per accessed byte), and large datasets that greatly exceed the main memory size. The computation in these workloads cannot usually compensate for the data movement costs. In order to alleviate this data movement bottleneck, we need a paradigm shift from the traditional processor-centric design, where all computation takes place in the compute units, to a more data-centric design where processing elements are placed closer to or inside where the data resides. This paradigm of computing is known as Processing-in-Memory (PIM).

This is your perfect P&S if you want to become familiar with the main PIM technologies, which represent “the next big thing” in Computer Architecture. You will work hands-on with the first real-world PIM architecture, will explore different PIM architecture designs for important workloads, and will develop tools to enable research of future PIM systems. Projects in this course span software and hardware as well as the software/hardware interface. You can potentially work on developing and optimizing new workloads for the first real-world PIM hardware or explore new PIM designs in simulators, or do something else that can forward our understanding of the PIM paradigm.

Livestream - P&S Data-Centric Architectures: Fundamentally...

Onur Mutlu Lectures

14 videos 631 views Updated 3 days ago

1 PIM Course: Lecture 1: Data-Centric Architectures: Improving Performance & Energy - Fall 2022 — Onur Mutlu Lectures • 1K views • 3 months ago — 1:34:40

2 PIM Course: Lecture 2: How to Evaluate Data Movement Bottlenecks - Fall 2022 — Onur Mutlu Lectures • 678 views • 2 months ago — 59:51

3 PIM Course: Lecture 3: Real-world PIM: UPMEM PIM - Fall 2022 — Onur Mutlu Lectures • 455 views • 2 months ago — 31:46

4 PIM Course: Lecture 4: Real-world PIM: Microbenchmarking of UPMEM PIM - Fall 2022 — Onur Mutlu Lectures • 275 views • 2 months ago — 56:05

5 PIM Course: Lecture 5: Real-world PIM: Samsung HBM-PIM - Fall 2022 — Onur Mutlu Lectures • 725 views • 2 months ago — 1:06:09

6 PIM Course: Lecture 6: Real-world PIM: SK Hynix AiM - Fall 2022 — Onur Mutlu Lectures • 1K views • 2 months ago — 41:33

7 PIM Course: Lecture 7: Real-world PIM: Samsung AxDIMM - Fall 2022 — Onur Mutlu Lectures • 767 views • 1 month ago — 32:43

8 PIM Course: Lecture 8: Real-world PIM: Alibaba HB-PNM - Fall 2022 — Onur Mutlu Lectures • 383 views • 1 month ago — 41:44

9 PIM Course: Lecture 9: Programming PIM Architectures - Fall 2022 — Onur Mutlu Lectures • 367 views • 1 month ago — 46:43

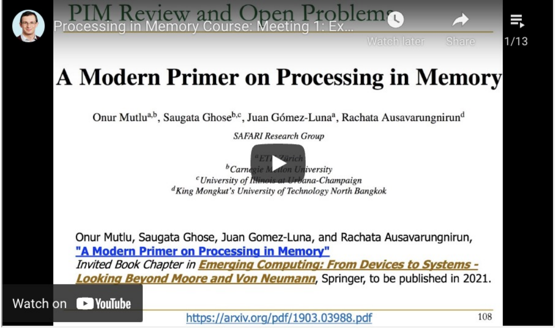
https://safari.ethz.ch/projects_and_seminars/fall2022/doku.php?id=processing_in_memory

https://youtube.com/playlist?list=PL5Q2soXY2Zi8KzG2CQYRNQOVD0GOBrnKy

# 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

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


PIM Review and Open Problems
Processing in Memory Course: Meeting 1: Ex...
Watch later
Share
1/13
Watch on YouTube

A Modern Primer on Processing in Memory

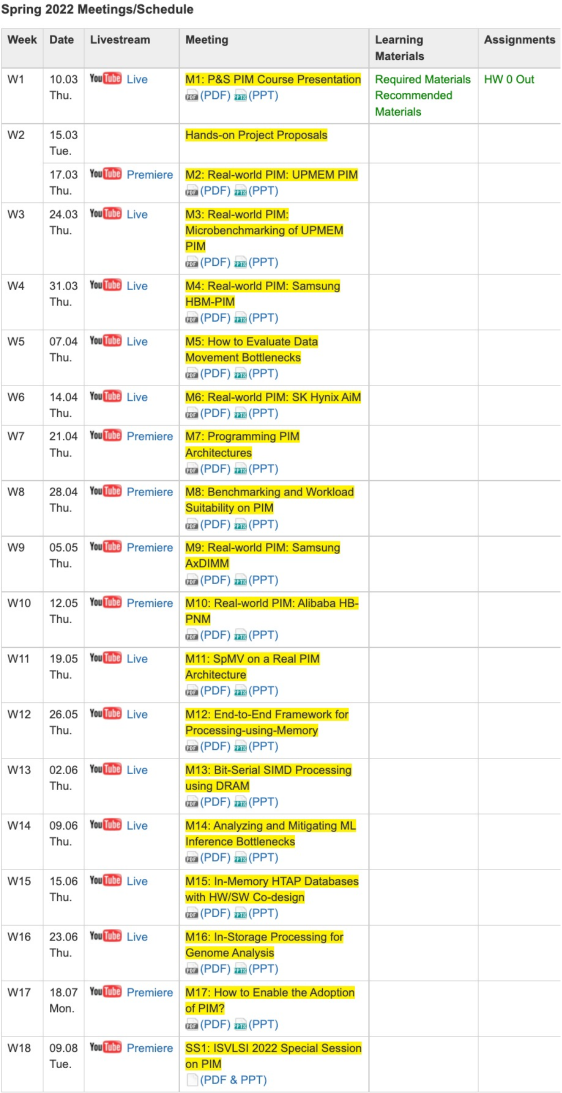
Onur Mutlu[a,b], Saugata Ghose[b,c], Juan Gómez-Luna[a], Rachata Ausavarungnirun[d]

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

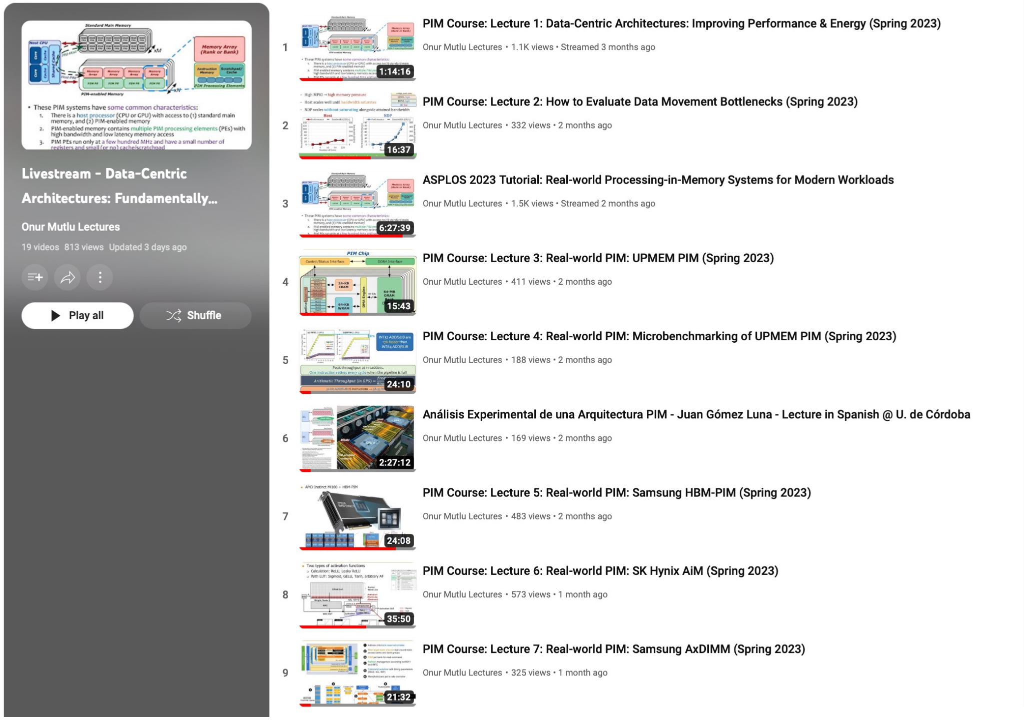


**Spring 2022 Meetings/Schedule**

| Week | Date | Livestream | Meeting | Learning Materials | Assignments |
|---|---|---|---|---|---|
| W1 | 10.03 Thu. | YouTube Live | M1: P&S PIM Course Presentation (PDF) (PPT) | Required Materials<br>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) | | |

# Processing-in-Memory Course (Spring 2023)

- Short weekly lectures
- Hands-on projects



Livestream - Data-Centric Architectures: Fundamentally...

Onur Mutlu Lectures

19 videos 813 views Updated 3 days ago

1 PIM Course: Lecture 1: Data-Centric Architectures: Improving Performance & Energy (Spring 2023)
Onur Mutlu Lectures • 1.1K views • Streamed 3 months ago
1:14:16

2 PIM Course: Lecture 2: How to Evaluate Data Movement Bottlenecks (Spring 2023)
Onur Mutlu Lectures • 332 views • 2 months ago
16:37

3 ASPLOS 2023 Tutorial: Real-world Processing-in-Memory Systems for Modern Workloads
Onur Mutlu Lectures • 1.5K views • Streamed 2 months ago
6:27:39

4 PIM Course: Lecture 3: Real-world PIM: UPMEM PIM (Spring 2023)
Onur Mutlu Lectures • 411 views • 2 months ago
15:43

5 PIM Course: Lecture 4: Real-world PIM: Microbenchmarking of UPMEM PIM (Spring 2023)
Onur Mutlu Lectures • 188 views • 2 months ago
24:10

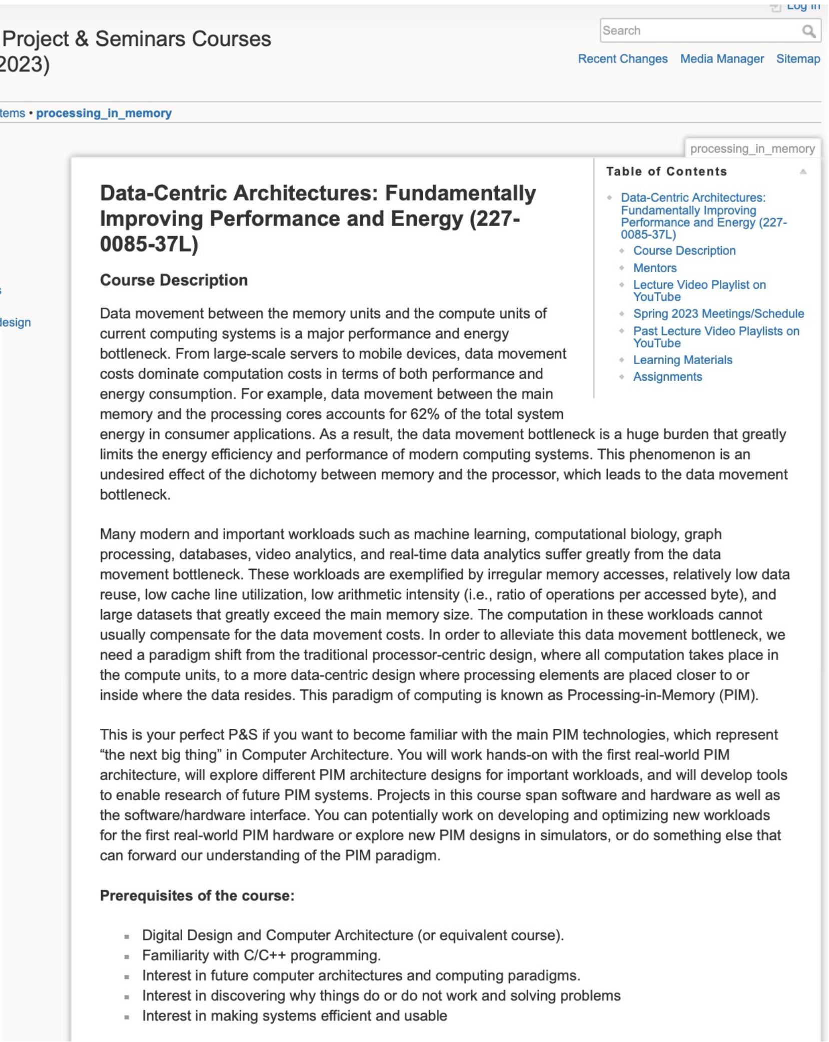
6 Análisis Experimental de una Arquitectura PIM - Juan Gómez Luna - Lecture in Spanish @ U. de Córdoba
Onur Mutlu Lectures • 169 views • 2 months ago
2:27:12

7 PIM Course: Lecture 5: Real-world PIM: Samsung HBM-PIM (Spring 2023)
Onur Mutlu Lectures • 483 views • 2 months ago
24:08

8 PIM Course: Lecture 6: Real-world PIM: SK Hynix AiM (Spring 2023)
Onur Mutlu Lectures • 573 views • 1 month ago
35:50

9 PIM Course: Lecture 7: Real-world PIM: Samsung AxDIMM (Spring 2023)
Onur Mutlu Lectures • 325 views • 1 month ago
21:32

https://www.youtube.com/playlist?list=PL5Q2soXY2Zi_EObuoAZVSq_o6UySWQHvZ



SAFARI Project & Seminars Courses (Spring 2023)

Trace: • heterogeneous_systems • processing_in_memory

Home

Courses

- SoftMC
- Ramulator
- Accelerating Genomics
- Mobile Genomics
- Processing-in-Memory
- Heterogeneous Systems
- Modern SSDs
- Hardware/Software Co-design

processing_in_memory

Table of Contents

- Data-Centric Architectures: Fundamentally Improving Performance and Energy (227-0085-37L)
  - Course Description
  - Mentors
  - Lecture Video Playlist on YouTube
  - Spring 2023 Meetings/Schedule
  - Past Lecture Video Playlists on YouTube
  - Learning Materials
  - Assignments

## Data-Centric Architectures: Fundamentally Improving Performance and Energy (227-0085-37L)

### Course Description

Data movement between the memory units and the compute units of current computing systems is a major performance and energy bottleneck. From large-scale servers to mobile devices, data movement costs dominate computation costs in terms of both performance and energy consumption. For example, data movement between the main memory and the processing cores accounts for 62% of the total system energy in consumer applications. As a result, the data movement bottleneck is a huge burden that greatly limits the energy efficiency and performance of modern computing systems. This phenomenon is an undesired effect of the dichotomy between memory and the processor, which leads to the data movement bottleneck.

Many modern and important workloads such as machine learning, computational biology, graph processing, databases, video analytics, and real-time data analytics suffer greatly from the data movement bottleneck. These workloads are exemplified by irregular memory accesses, relatively low data reuse, low cache line utilization, low arithmetic intensity (i.e., ratio of operations per accessed byte), and large datasets that greatly exceed the main memory size. The computation in these workloads cannot usually compensate for the data movement costs. In order to alleviate this data movement bottleneck, we need a paradigm shift from the traditional processor-centric design, where all computation takes place in the compute units, to a more data-centric design where processing elements are placed closer to or inside where the data resides. This paradigm of computing is known as Processing-in-Memory (PIM).

This is your perfect P&S if you want to become familiar with the main PIM technologies, which represent "the next big thing" in Computer Architecture. You will work hands-on with the first real-world PIM architecture, will explore different PIM architecture designs for important workloads, and will develop tools to enable research of future PIM systems. Projects in this course span software and hardware as well as the software/hardware interface. You can potentially work on developing and optimizing new workloads for the first real-world PIM hardware or explore new PIM designs in simulators, or do something else that can forward our understanding of the PIM paradigm.

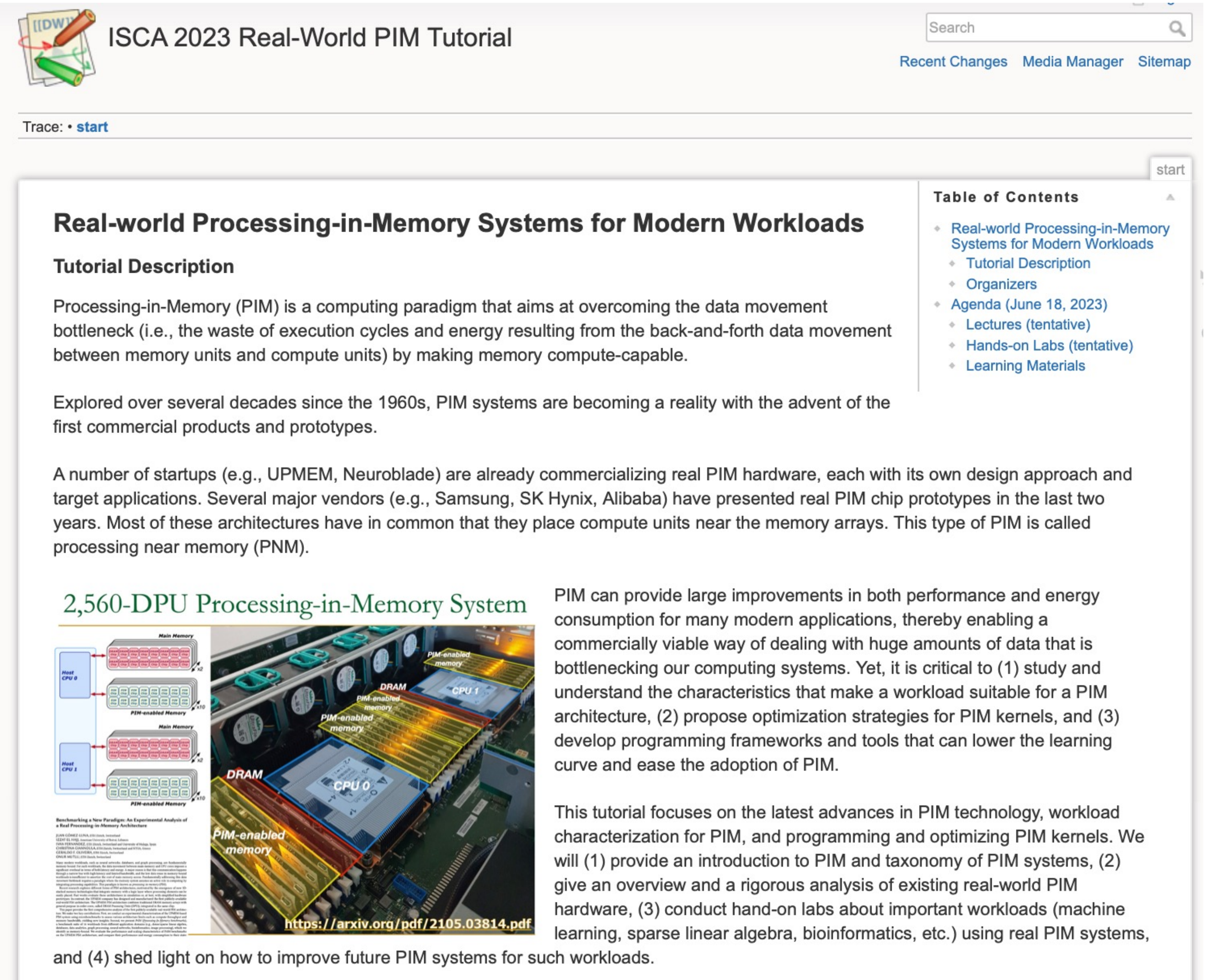
**Prerequisites of the course:**

- Digital Design and Computer Architecture (or equivalent course).
- Familiarity with C/C++ programming.
- Interest in future computer architectures and computing paradigms.
- Interest in discovering why things do or do not work and solving problems
- Interest in making systems efficient and usable

https://safari.ethz.ch/projects_and_seminars/spring2023/doku.php?id=processing_in_memory

# Real PIM Tutorials [ISCA'23, ASPLOS'23, HPCA'23]

- June, March, Feb : Lectures + Hands-on labs + Invited talks



ISCA 2023 Real-World PIM Tutorial

Trace: • start

## Real-world Processing-in-Memory Systems for Modern Workloads

### Tutorial Description

Processing-in-Memory (PIM) is a computing paradigm that aims at overcoming the data movement bottleneck (i.e., the waste of execution cycles and energy resulting from the back-and-forth data movement between memory units and compute units) by making memory compute-capable.

Explored over several decades since the 1960s, PIM systems are becoming a reality with the advent of the first commercial products and prototypes.

A number of startups (e.g., UPMEM, Neuroblade) are already commercializing real PIM hardware, each with its own design approach and target applications. Several major vendors (e.g., Samsung, SK Hynix, Alibaba) have presented real PIM chip prototypes in the last two years. Most of these architectures have in common that they place compute units near the memory arrays. This type of PIM is called processing near memory (PNM).

PIM can provide large improvements in both performance and energy consumption for many modern applications, thereby enabling a commercially viable way of dealing with huge amounts of data that is bottlenecking our computing systems. Yet, it is critical to (1) study and understand the characteristics that make a workload suitable for a PIM architecture, (2) propose optimization strategies for PIM kernels, and (3) develop programming frameworks and tools that can lower the learning curve and ease the adoption of PIM.

This tutorial focuses on the latest advances in PIM technology, workload characterization for PIM, and programming and optimizing PIM kernels. We will (1) provide an introduction to PIM and taxonomy of PIM systems, (2) give an overview and a rigorous analysis of existing real-world PIM hardware, (3) conduct hand-on labs about important workloads (machine learning, sparse linear algebra, bioinformatics, etc.) using real PIM systems, and (4) shed light on how to improve future PIM systems for such workloads.

https://events.safari.ethz.ch/isca-pim-tutorial/

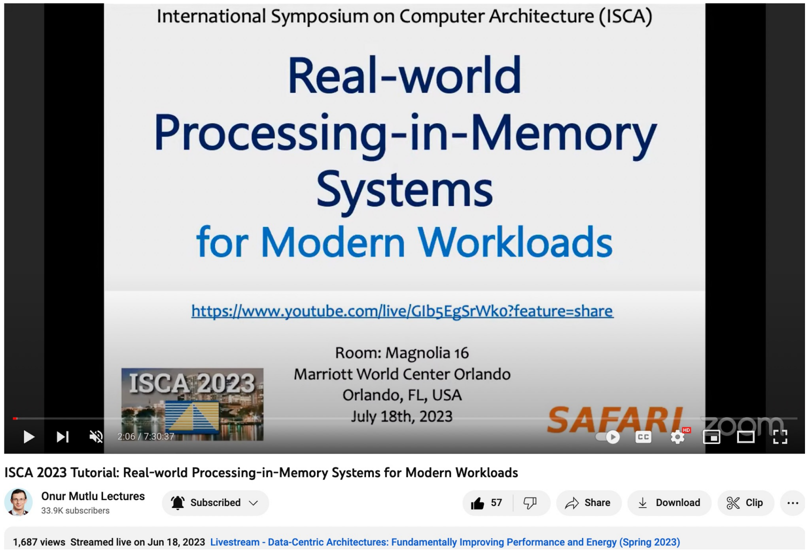
# Real PIM Tutorial [ISCA 2023]

- June 18: Lectures + Hands-on labs + Invited talks

ISCA 2023 Real-World PIM Tutorial
Sunday, June 18, Orlando, Florida

**Organizers:** Juan Gómez Luna, Onur Mutlu, Ataberk Olgun
**Program:** https://events.safari.ethz.ch/isca-pim-tutorial/

Overview PIM | PNM | UPMEM PIM | PNM for neural networks | PNM for recommender systems | PNM for ML workloads | How to enable PIM? | PUM prototypes
**Hands-on Labs:** Benchmarking | Accelerating real-world workloads



International Symposium on Computer Architecture (ISCA)

Real-world Processing-in-Memory Systems for Modern Workloads

https://www.youtube.com/live/GIb5EgSrWko?feature=share

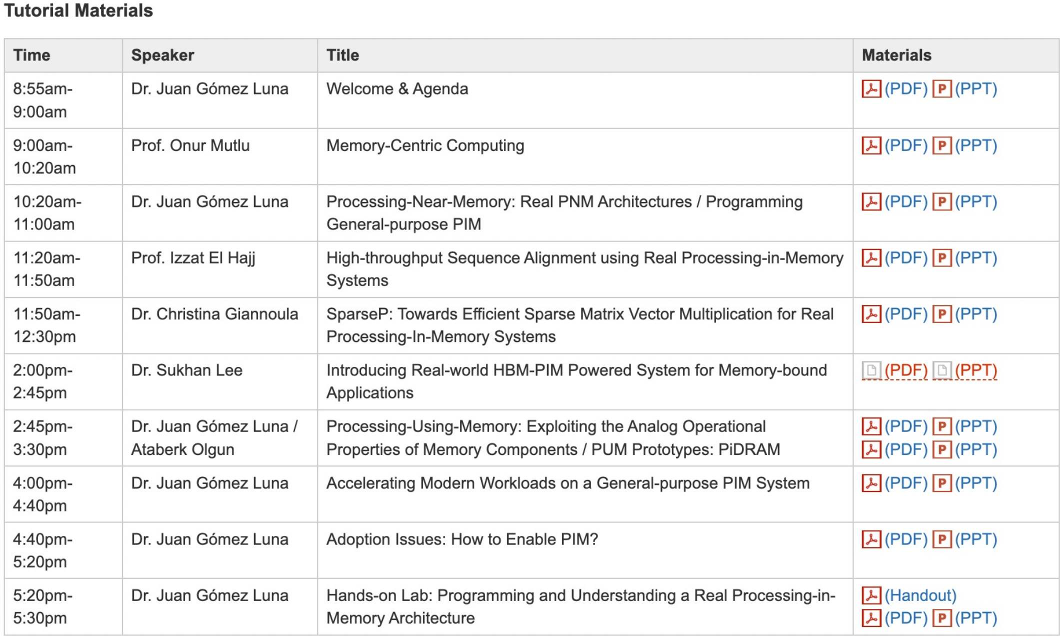
Room: Magnolia 16
Marriott World Center Orlando
Orlando, FL, USA
July 18th, 2023

ISCA 2023 Tutorial: Real-world Processing-in-Memory Systems for Modern Workloads

Onur Mutlu Lectures
33.9K subscribers

1,687 views Streamed live on Jun 18, 2023 Livestream - Data-Centric Architectures: Fundamentally Improving Performance and Energy (Spring 2023)

ISCA 2023
SAFARI
zoom
2:06 / 7:30:37
Subscribed
57
Share
Download
Clip



**Tutorial Materials**

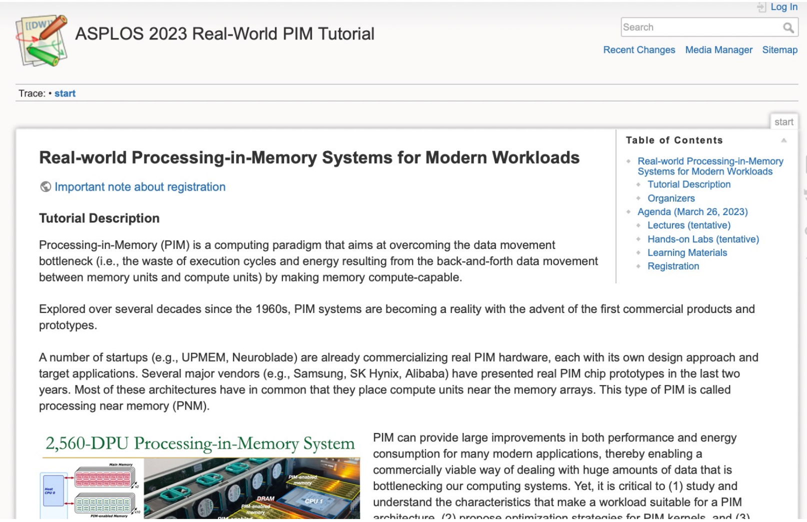
| Time | Speaker | Title | Materials |
|---|---|---|---|
| 8:55am-9:00am | Dr. Juan Gómez Luna | Welcome & Agenda | (PDF) (PPT) |
| 9:00am-10:20am | Prof. Onur Mutlu | Memory-Centric Computing | (PDF) (PPT) |
| 10:20am-11:00am | Dr. Juan Gómez Luna | Processing-Near-Memory: Real PNM Architectures / Programming General-purpose PIM | (PDF) (PPT) |
| 11:20am-11:50am | Prof. Izzat El Hajj | High-throughput Sequence Alignment using Real Processing-in-Memory Systems | (PDF) (PPT) |
| 11:50am-12:30pm | Dr. Christina Giannoula | SparseP: Towards Efficient Sparse Matrix Vector Multiplication for Real Processing-In-Memory Systems | (PDF) (PPT) |
| 2:00pm-2:45pm | Dr. Sukhan Lee | Introducing Real-world HBM-PIM Powered System for Memory-bound Applications | (PDF) (PPT) |
| 2:45pm-3:30pm | Dr. Juan Gómez Luna / Ataberk Olgun | Processing-Using-Memory: Exploiting the Analog Operational Properties of Memory Components / PUM Prototypes: PiDRAM | (PDF) (PPT)<br>(PDF) (PPT) |
| 4:00pm-4:40pm | Dr. Juan Gómez Luna | Accelerating Modern Workloads on a General-purpose PIM System | (PDF) (PPT) |
| 4:40pm-5:20pm | Dr. Juan Gómez Luna | Adoption Issues: How to Enable PIM? | (PDF) (PPT) |
| 5:20pm-5:30pm | Dr. Juan Gómez Luna | Hands-on Lab: Programming and Understanding a Real Processing-in-Memory Architecture | (Handout)<br>(PDF) (PPT) |

https://www.youtube.com/live/GIb5EgSrWk0

https://events.safari.ethz.ch/isca-pim-tutorial/

# Real PIM Tutorial [ASPLOS 2023]

- March 26: Lectures + Hands-on labs + Invited talks



ASPLOS 2023 Real-World PIM Tutorial

Trace: • start

## Real-world Processing-in-Memory Systems for Modern Workloads

Important note about registration

Table of Contents

- Real-world Processing-in-Memory Systems for Modern Workloads
  - Tutorial Description
  - Organizers
- Agenda (March 26, 2023)
  - Lectures (tentative)
  - Hands-on Labs (tentative)
  - Learning Materials
  - Registration

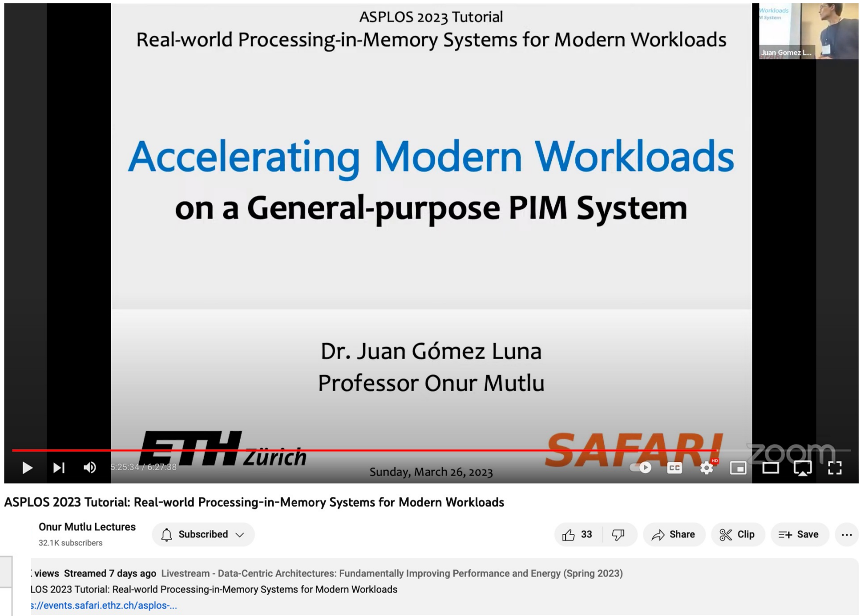
### Tutorial Description

Processing-in-Memory (PIM) is a computing paradigm that aims at overcoming the data movement bottleneck (i.e., the waste of execution cycles and energy resulting from the back-and-forth data movement between memory units and compute units) by making memory compute-capable.

Explored over several decades since the 1960s, PIM systems are becoming a reality with the advent of the first commercial products and prototypes.

A number of startups (e.g., UPMEM, Neuroblade) are already commercializing real PIM hardware, each with its own design approach and target applications. Several major vendors (e.g., Samsung, SK Hynix, Alibaba) have presented real PIM chip prototypes in the last two years. Most of these architectures have in common that they place compute units near the memory arrays. This type of PIM is called processing near memory (PNM).

PIM can provide large improvements in both performance and energy consumption for many modern applications, thereby enabling a commercially viable way of dealing with huge amounts of data that is bottlenecking our computing systems. Yet, it is critical to (1) study and understand the characteristics that make a workload suitable for a PIM architecture, (2) propose optimization strategies for PIM kernels, and (3)



ASPLOS 2023 Tutorial
Real-world Processing-in-Memory Systems for Modern Workloads

Accelerating Modern Workloads
on a General-purpose PIM System

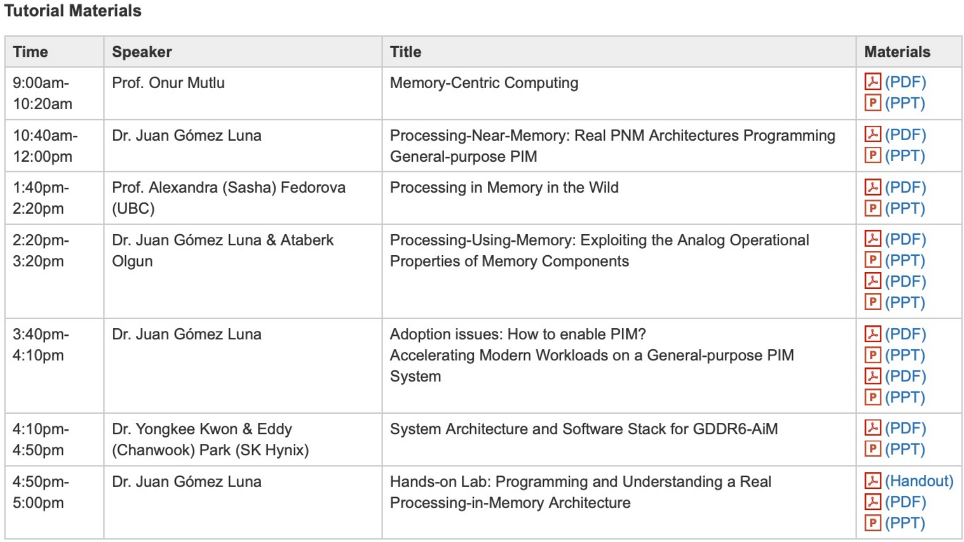
Dr. Juan Gómez Luna
Professor Onur Mutlu

Sunday, March 26, 2023

ASPLOS 2023 Tutorial: Real-world Processing-in-Memory Systems for Modern Workloads

Onur Mutlu Lectures
32.1K subscribers

views Streamed 7 days ago Livestream - Data-Centric Architectures: Fundamentally Improving Performance and Energy (Spring 2023)
LOS 2023 Tutorial: Real-world Processing-in-Memory Systems for Modern Workloads
s://events.safari.ethz.ch/asplos-...



### Tutorial Materials

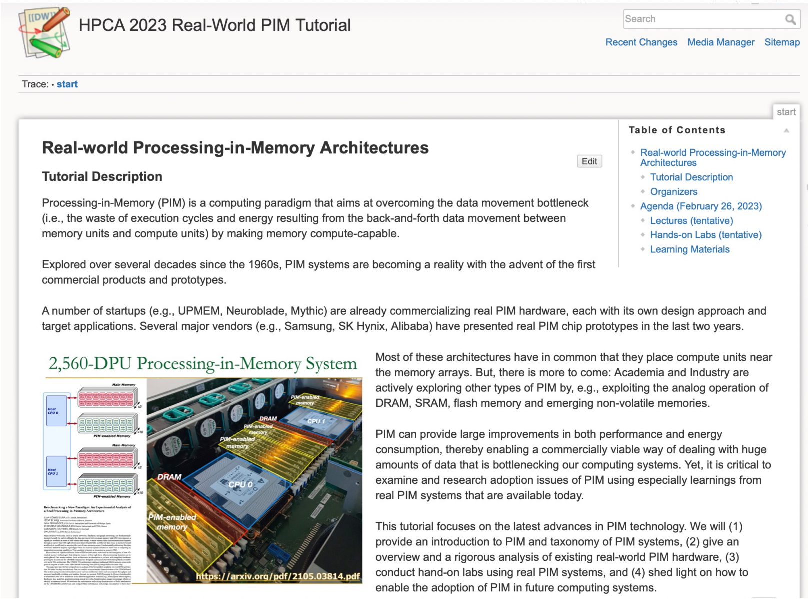
| Time | Speaker | Title | Materials |
|---|---|---|---|
| 9:00am-10:20am | Prof. Onur Mutlu | Memory-Centric Computing | (PDF)<br>(PPT) |
| 10:40am-12:00pm | Dr. Juan Gómez Luna | Processing-Near-Memory: Real PNM Architectures Programming General-purpose PIM | (PDF)<br>(PPT) |
| 1:40pm-2:20pm | Prof. Alexandra (Sasha) Fedorova (UBC) | Processing in Memory in the Wild | (PDF)<br>(PPT) |
| 2:20pm-3:20pm | Dr. Juan Gómez Luna & Ataberk Olgun | Processing-Using-Memory: Exploiting the Analog Operational Properties of Memory Components | (PDF)<br>(PPT)<br>(PDF)<br>(PPT) |
| 3:40pm-4:10pm | Dr. Juan Gómez Luna | Adoption issues: How to enable PIM?<br>Accelerating Modern Workloads on a General-purpose PIM System | (PDF)<br>(PPT)<br>(PDF)<br>(PPT) |
| 4:10pm-4:50pm | Dr. Yongkee Kwon & Eddy (Chanwook) Park (SK Hynix) | System Architecture and Software Stack for GDDR6-AiM | (PDF)<br>(PPT) |
| 4:50pm-5:00pm | Dr. Juan Gómez Luna | Hands-on Lab: Programming and Understanding a Real Processing-in-Memory Architecture | (Handout)<br>(PDF)<br>(PPT) |

**https://www.youtube.com/watch?v=oYCaLcT0Kmo**

**https://events.safari.ethz.ch/asplos-pim-tutorial/**

# Real PIM Tutorial [HPCA 2023]

- February 26: Lectures + Hands-on labs + Invited Talks



HPCA 2023 Real-World PIM Tutorial

Trace: · start

## Real-world Processing-in-Memory Architectures

### Tutorial Description

Processing-in-Memory (PIM) is a computing paradigm that aims at overcoming the data movement bottleneck (i.e., the waste of execution cycles and energy resulting from the back-and-forth data movement between memory units and compute units) by making memory compute-capable.

Explored over several decades since the 1960s, PIM systems are becoming a reality with the advent of the first commercial products and prototypes.

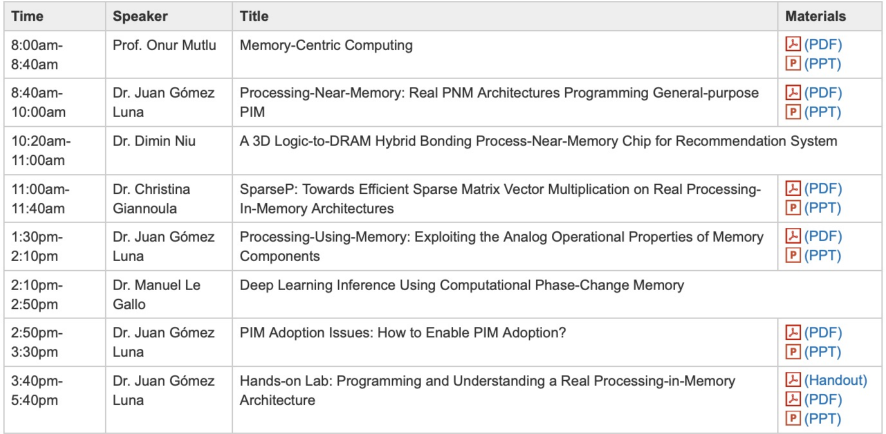
A number of startups (e.g., UPMEM, Neuroblade, Mythic) are already commercializing real PIM hardware, each with its own design approach and target applications. Several major vendors (e.g., Samsung, SK Hynix, Alibaba) have presented real PIM chip prototypes in the last two years.

2,560-DPU Processing-in-Memory System

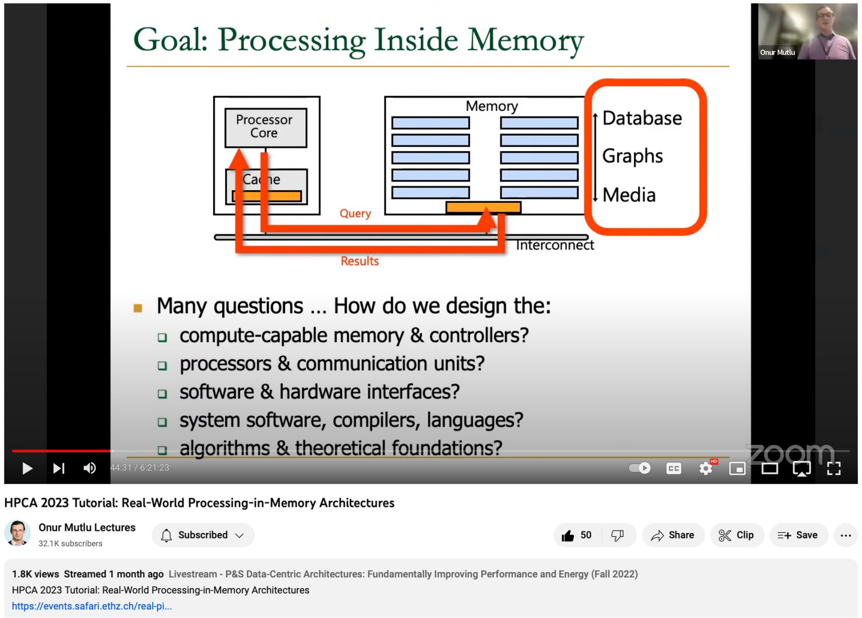
Most of these architectures have in common that they place compute units near the memory arrays. But, there is more to come: Academia and Industry are actively exploring other types of PIM by, e.g., exploiting the analog operation of DRAM, SRAM, flash memory and emerging non-volatile memories.

PIM can provide large improvements in both performance and energy consumption, thereby enabling a commercially viable way of dealing with huge amounts of data that is bottlenecking our computing systems. Yet, it is critical to examine and research adoption issues of PIM using especially learnings from real PIM systems that are available today.

This tutorial focuses on the latest advances in PIM technology. We will (1) provide an introduction to PIM and taxonomy of PIM systems, (2) give an overview and a rigorous analysis of existing real-world PIM hardware, (3) conduct hands-on labs using real PIM systems, and (4) shed light on how to enable the adoption of PIM in future computing systems.



| Time | Speaker | Title | Materials |
|---|---|---|---|
| 8:00am-8:40am | Prof. Onur Mutlu | Memory-Centric Computing | (PDF)<br>(PPT) |
| 8:40am-10:00am | Dr. Juan Gómez Luna | Processing-Near-Memory: Real PNM Architectures Programming General-purpose PIM | (PDF)<br>(PPT) |
| 10:20am-11:00am | Dr. Dimin Niu | A 3D Logic-to-DRAM Hybrid Bonding Process-Near-Memory Chip for Recommendation System | |
| 11:00am-11:40am | Dr. Christina Giannoula | SparseP: Towards Efficient Sparse Matrix Vector Multiplication on Real Processing-In-Memory Architectures | (PDF)<br>(PPT) |
| 1:30pm-2:10pm | Dr. Juan Gómez Luna | Processing-Using-Memory: Exploiting the Analog Operational Properties of Memory Components | (PDF)<br>(PPT) |
| 2:10pm-2:50pm | Dr. Manuel Le Gallo | Deep Learning Inference Using Computational Phase-Change Memory | |
| 2:50pm-3:30pm | Dr. Juan Gómez Luna | PIM Adoption Issues: How to Enable PIM Adoption? | (PDF)<br>(PPT) |
| 3:40pm-5:40pm | Dr. Juan Gómez Luna | Hands-on Lab: Programming and Understanding a Real Processing-in-Memory Architecture | (Handout)<br>(PDF)<br>(PPT) |



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?

HPCA 2023 Tutorial: Real-World Processing-in-Memory Architectures

Onur Mutlu Lectures
32.1K subscribers

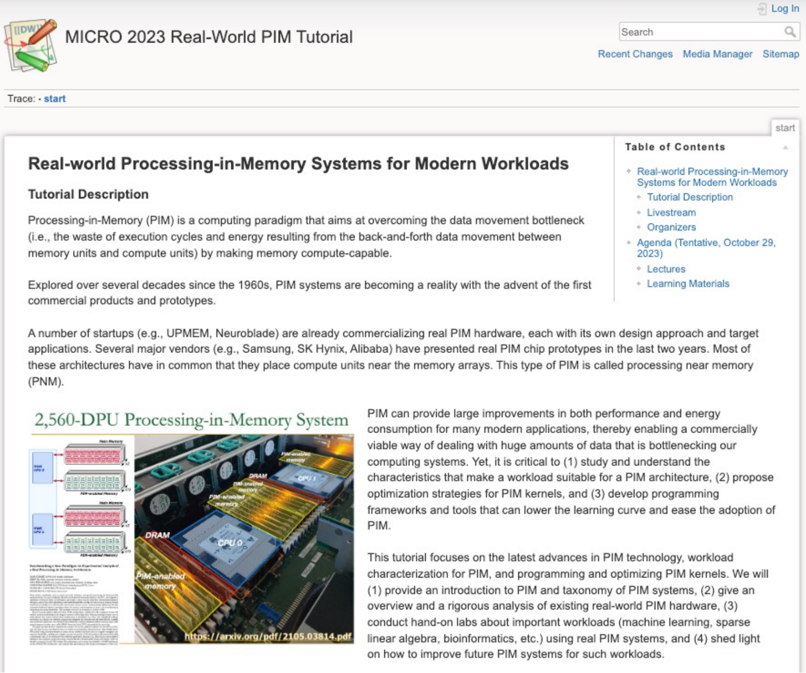
1.8K views Streamed 1 month ago Livestream - P&S Data-Centric Architectures: Fundamentally Improving Performance and Energy (Fall 2022)
HPCA 2023 Tutorial: Real-World Processing-in-Memory Architectures
https://events.safari.ethz.ch/real-pi...

[https://www.youtube.com/watch?v=f5-nT1tbz5w](https://www.youtube.com/watch?v=f5-nT1tbz5w)

[https://events.safari.ethz.ch/real-pim-tutorial/](https://events.safari.ethz.ch/real-pim-tutorial/)

# Real PIM Tutorial [MICRO 2023]

- October 29: Lectures + Hands-on labs + Invited talks



Log In
MICRO 2023 Real-World PIM Tutorial
Search
Recent Changes Media Manager Sitemap
Trace: · start
start

Table of Contents

- Real-world Processing-in-Memory Systems for Modern Workloads
  - Tutorial Description
  - Livestream
  - Organizers
- Agenda (Tentative, October 29, 2023)
  - Lectures
  - Learning Materials

## Real-world Processing-in-Memory Systems for Modern Workloads

### Tutorial Description

Processing-in-Memory (PIM) is a computing paradigm that aims at overcoming the data movement bottleneck (i.e., the waste of execution cycles and energy resulting from the back-and-forth data movement between memory units and compute units) by making memory compute-capable.

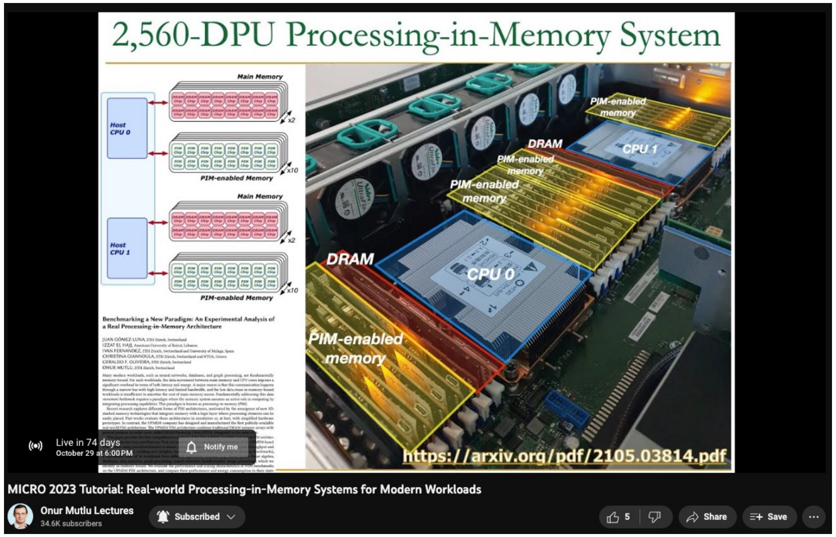
Explored over several decades since the 1960s, PIM systems are becoming a reality with the advent of the first commercial products and prototypes.

A number of startups (e.g., UPMEM, Neuroblade) are already commercializing real PIM hardware, each with its own design approach and target applications. Several major vendors (e.g., Samsung, SK Hynix, Alibaba) have presented real PIM chip prototypes in the last two years. Most of these architectures have in common that they place compute units near the memory arrays. This type of PIM is called processing near memory (PNM).

2,560-DPU Processing-in-Memory System
DRAM
PIM-enabled memory
CPU 0
CPU 1
https://arxiv.org/pdf/2105.03814.pdf

PIM can provide large improvements in both performance and energy consumption for many modern applications, thereby enabling a commercially viable way of dealing with huge amounts of data that is bottlenecking our computing systems. Yet, it is critical to (1) study and understand the characteristics that make a workload suitable for a PIM architecture, (2) propose optimization strategies for PIM kernels, and (3) develop programming frameworks and tools that can lower the learning curve and ease the adoption of PIM.

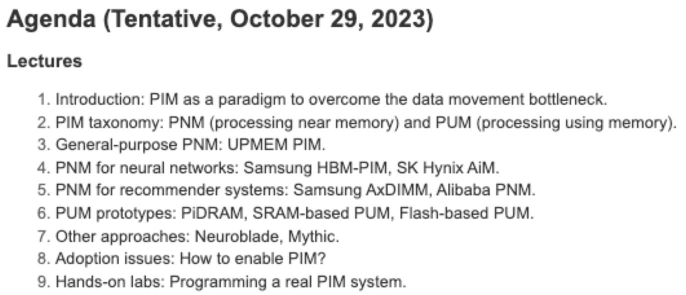
This tutorial focuses on the latest advances in PIM technology, workload characterization for PIM, and programming and optimizing PIM kernels. We will (1) provide an introduction to PIM and taxonomy of PIM systems, (2) give an overview and a rigorous analysis of existing real-world PIM hardware, (3) conduct hands-on labs about important workloads (machine learning, sparse linear algebra, bioinformatics, etc.) using real PIM systems, and (4) shed light on how to improve future PIM systems for such workloads.



2,560-DPU Processing-in-Memory System
Main Memory
Host CPU 0
PIM-enabled Memory
Main Memory
Host CPU 1
PIM-enabled Memory
x2
x10
Benchmarking a New Paradigm: An Experimental Analysis of a Real Processing-in-Memory Architecture
DRAM
PIM-enabled memory
CPU 0
CPU 1
https://arxiv.org/pdf/2105.03814.pdf
Live in 74 days
October 29 at 6:00 PM
Notify me

MICRO 2023 Tutorial: Real-world Processing-in-Memory Systems for Modern Workloads

[https://www.youtube.com/live/ohUooNSIxOI](https://www.youtube.com/live/ohUooNSIxOI)



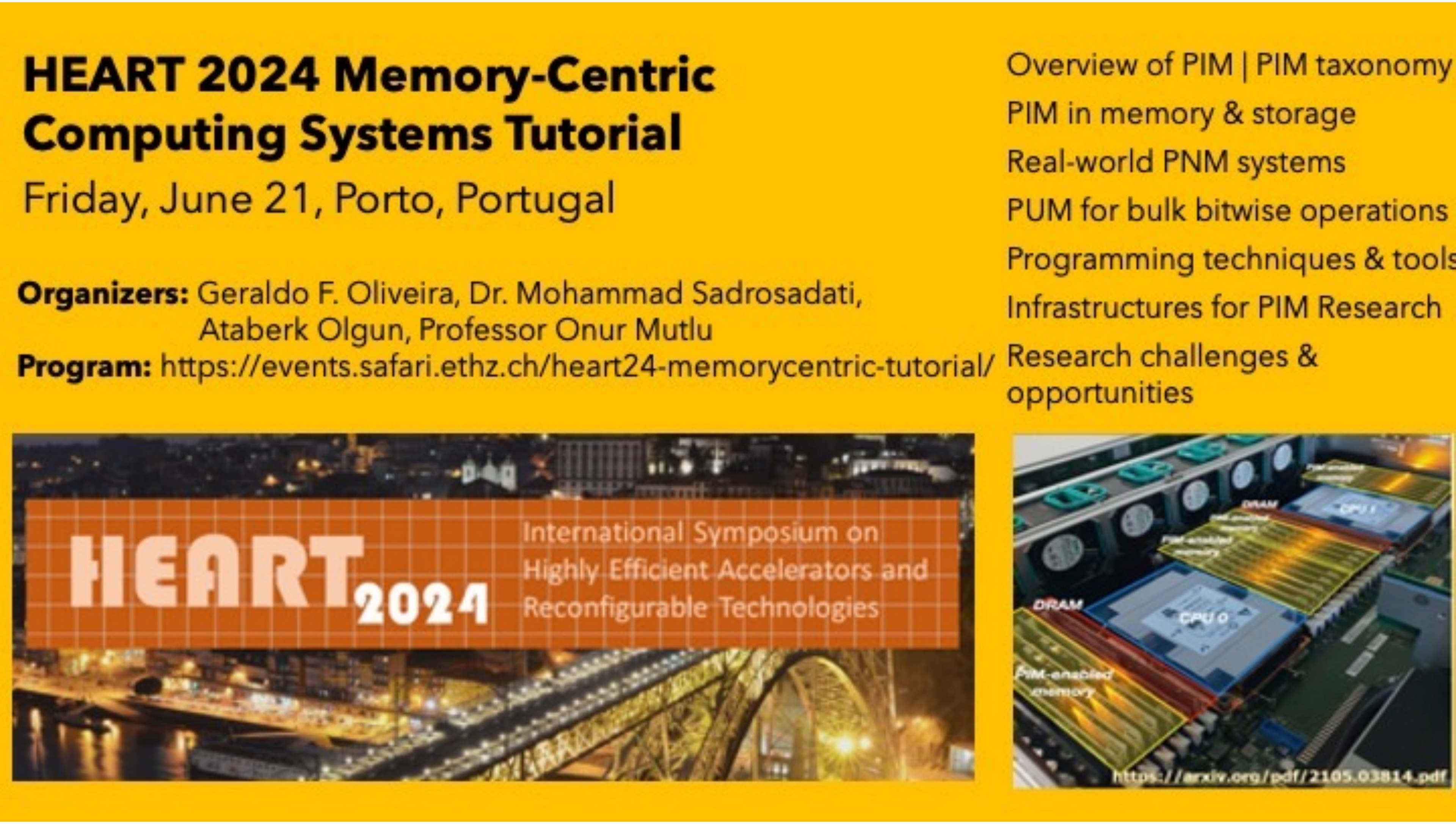
## Agenda (Tentative, October 29, 2023)

### Lectures

1. Introduction: PIM as a paradigm to overcome the data movement bottleneck.
2. PIM taxonomy: PNM (processing near memory) and PUM (processing using memory).
3. General-purpose PNM: UPMEM PIM.
4. PNM for neural networks: Samsung HBM-PIM, SK Hynix AiM.
5. PNM for recommender systems: Samsung AxDIMM, Alibaba PNM.
6. PUM prototypes: PiDRAM, SRAM-based PUM, Flash-based PUM.
7. Other approaches: Neuroblade, Mythic.
8. Adoption issues: How to enable PIM?
9. Hands-on labs: Programming a real PIM system.

[https://events.safari.ethz.ch/micro-pim-tutorial](https://events.safari.ethz.ch/micro-pim-tutorial)

# PIM Tutorial at HEART 2024



## HEART 2024 Memory-Centric Computing Systems Tutorial

Friday, June 21, Porto, Portugal

**Organizers:** Geraldo F. Oliveira, Dr. Mohammad Sadrosadati, Ataberk Olgun, Professor Onur Mutlu

**Program:** https://events.safari.ethz.ch/heart24-memorycentric-tutorial/

Overview of PIM | PIM taxonomy

PIM in memory & storage

Real-world PNM systems

PUM for bulk bitwise operations

Programming techniques & tool

Infrastructures for PIM Research

Research challenges & opportunities

HEART 2024
International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies
DRAM
CPU 0
CPU 1
PIM-enabled memory
https://arxiv.org/pdf/2105.03814.pdf

https://events.safari.ethz.ch/heart24-memorycentric-tutorial

# This PIM Tutorial at ISCA 2024

**ISCA 2024 Memory-Centric Computing Systems Tutorial**

Saturday, June 29, Buenos Aires, Argentina

**Organizers:** Geraldo F. Oliveira, Dr. Mohammad Sadrosadati,
Ataberk Olgun, Professor Onur Mutlu

**Program:** https://events.safari.ethz.ch/isca24-memorycentric-tutorial/

Overview of PIM | PIM taxonomy

PIM in memory & storage

Real-world PNM systems

PUM for bulk bitwise operations

Programming techniques & tools

Infrastructures for PIM Research

Research challenges & opportunities

[https://events.safari.ethz.ch/isca24-memorycentric-tutorial](https://events.safari.ethz.ch/isca24-memorycentric-tutorial)

# Referenced Papers, Talks, Artifacts

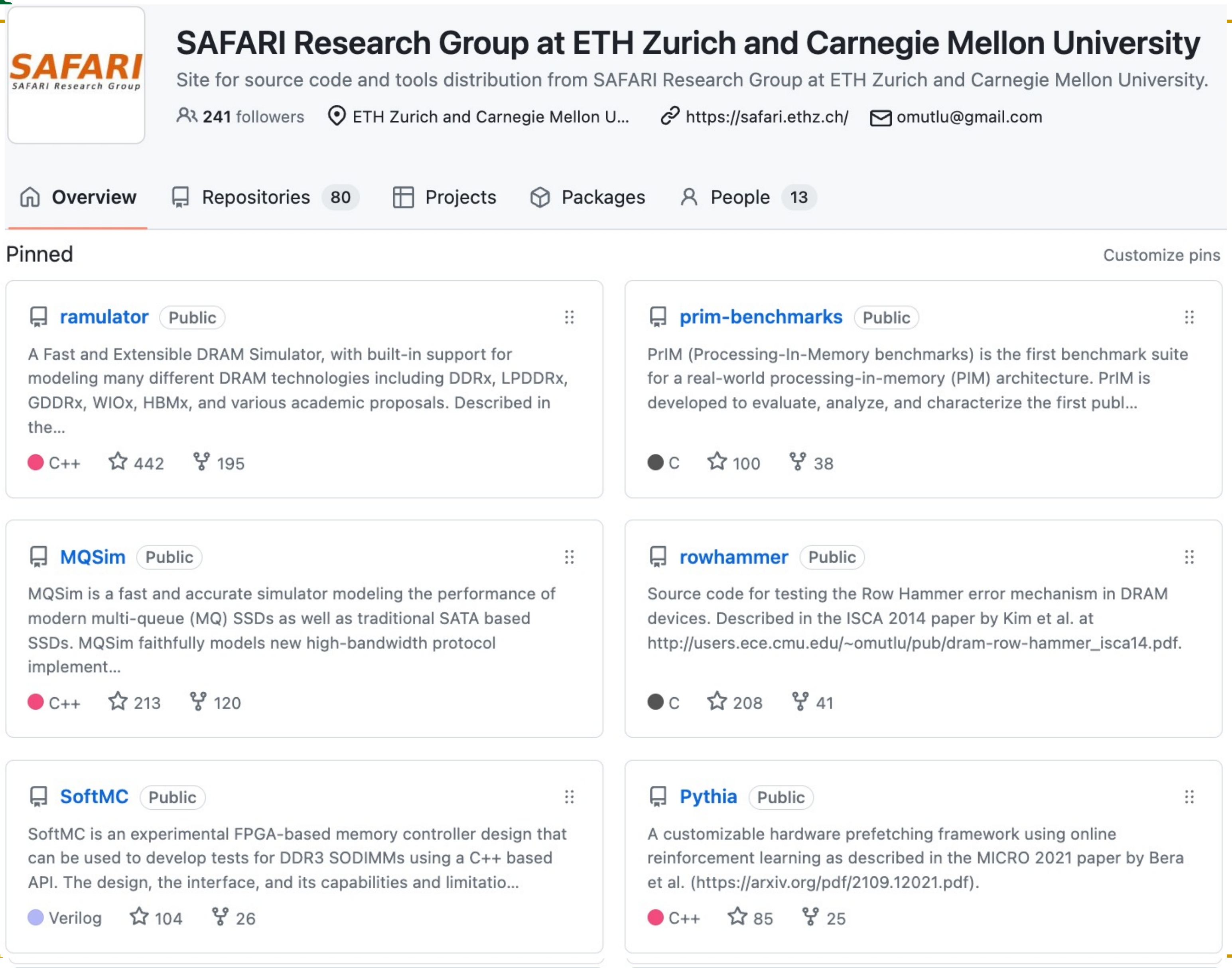
- All are available at

  **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.

241 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++ ☆ 442 195

**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 ☆ 100 38

**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++ ☆ 213 120

**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 ☆ 208 41

**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 ☆ 104 26

**Pythia** Public

A customizable hardware prefetching framework using online reinforcement learning as described in the MICRO 2021 paper by Bera et al. (https://arxiv.org/pdf/2109.12021.pdf).

● C++ ☆ 85 25

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

# Tutorial on Memory-Centric Computing: Conclusion Remarks

Geraldo F. Oliveira

Prof. Onur Mutlu

ISCA 2024

29 June 2024

SAFARI

ETH zürich