International Symposium on Microarchitecture (MICRO)

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

https://ethz.zoom.us/j/68459763236?pwd=NEIzK1JGV0pHTjRaUzV5NmRna3JkZz09

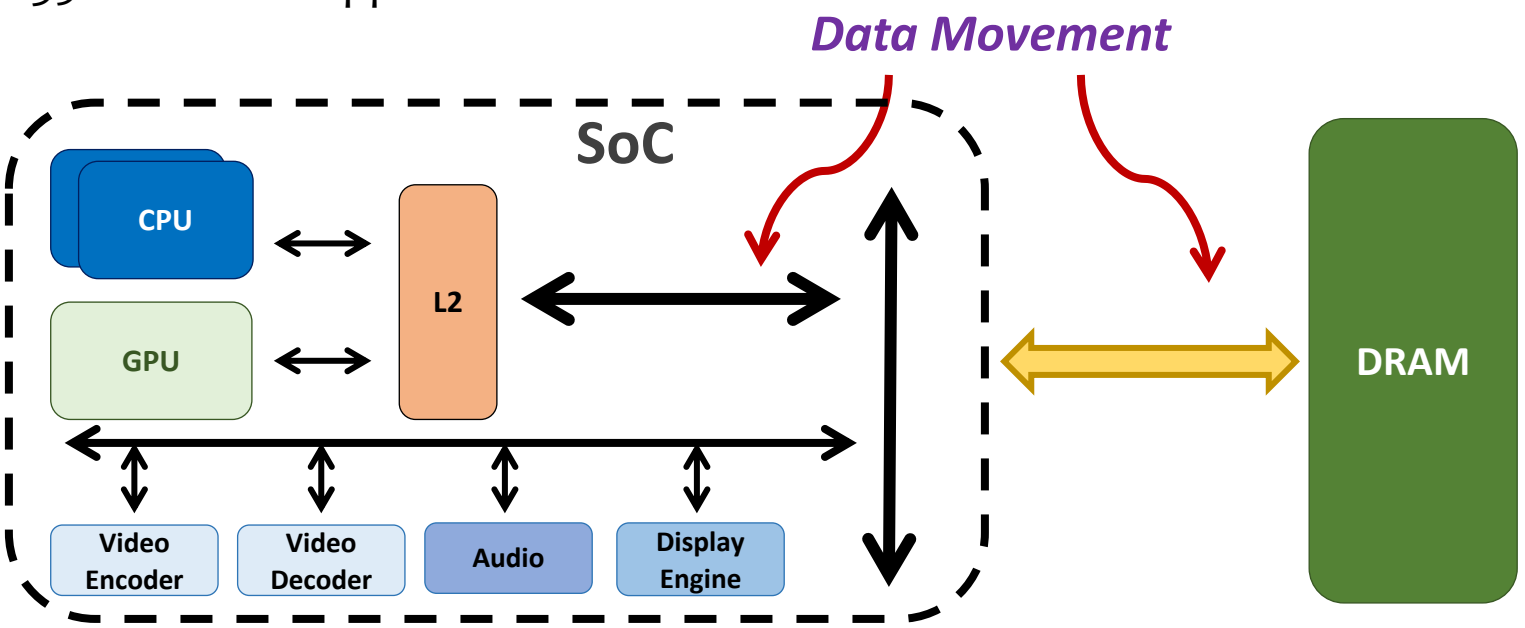
https://www.youtube.com/live/ohU00NSIxOI?si=h1I9eUx7KCScMHrg

October 29th, 2023

SAFARI

Data Movement in Computing Systems

- Data movement dominates performance and is a major system energy bottleneck
- Total system energy: data movement accounts for
  - 62% in consumer applications*,
  - 40% in scientific applications★,
  - 35% in mobile applications☆



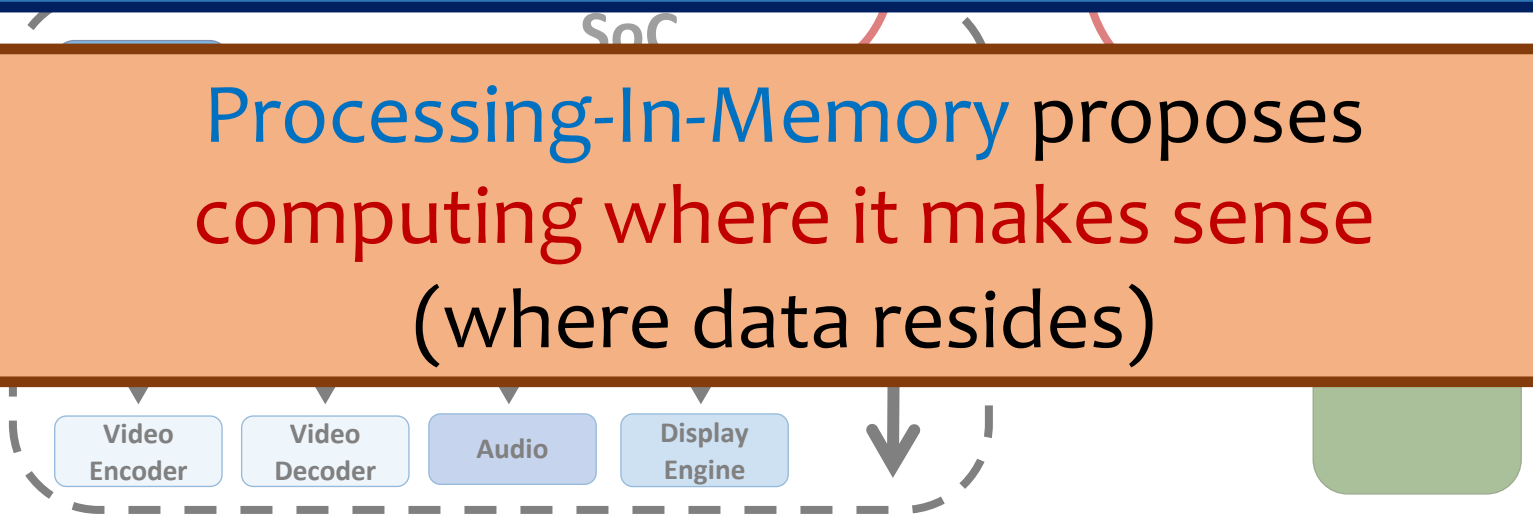
* Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS 2018
★ Kestor et al., "Quantifying the Energy Cost of Data Movement in Scientific Applications," IISWC 2013
☆ Pandiyan and Wu, "Quantifying the energy cost of data movement for emerging smart phone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

Data Movement in Computing Systems

- Data movement dominates performance and is a major system energy bottleneck
- Total system energy: data movement accounts for
  - 62% in consumer applications*,

Compute systems should be more data-centric

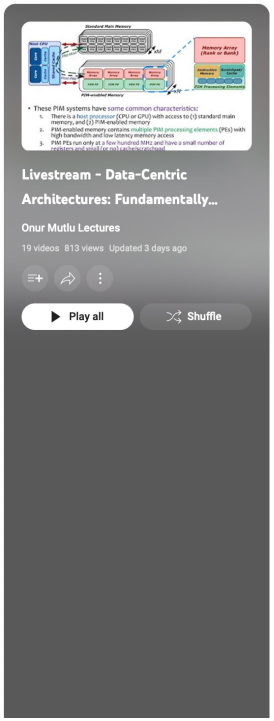
Processing-In-Memory proposes computing where it makes sense (where data resides)

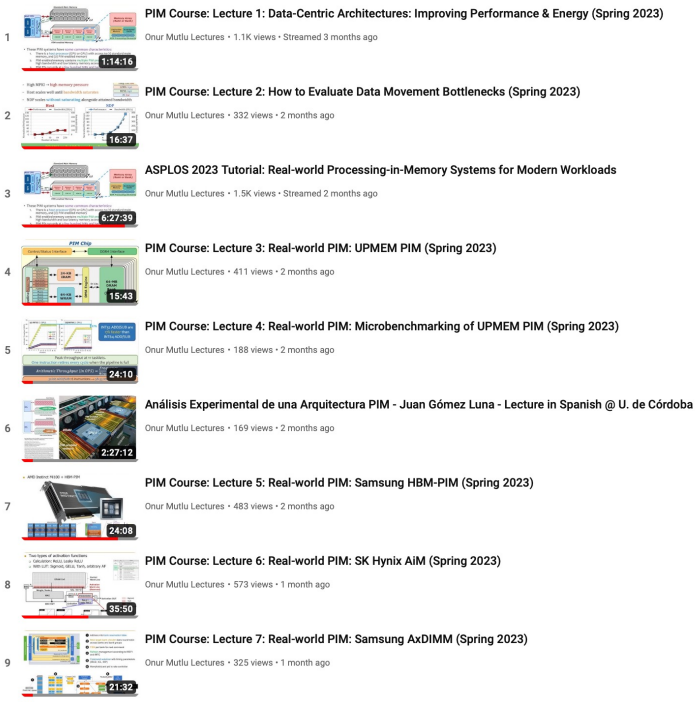


* Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS 2018
★ Kestor et al., "Quantifying the Energy Cost of Data Movement in Scientific Applications," IISWC 2013
☆ Pandiyan and Wu, "Quantifying the energy cost of data movement for emerging smart phone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

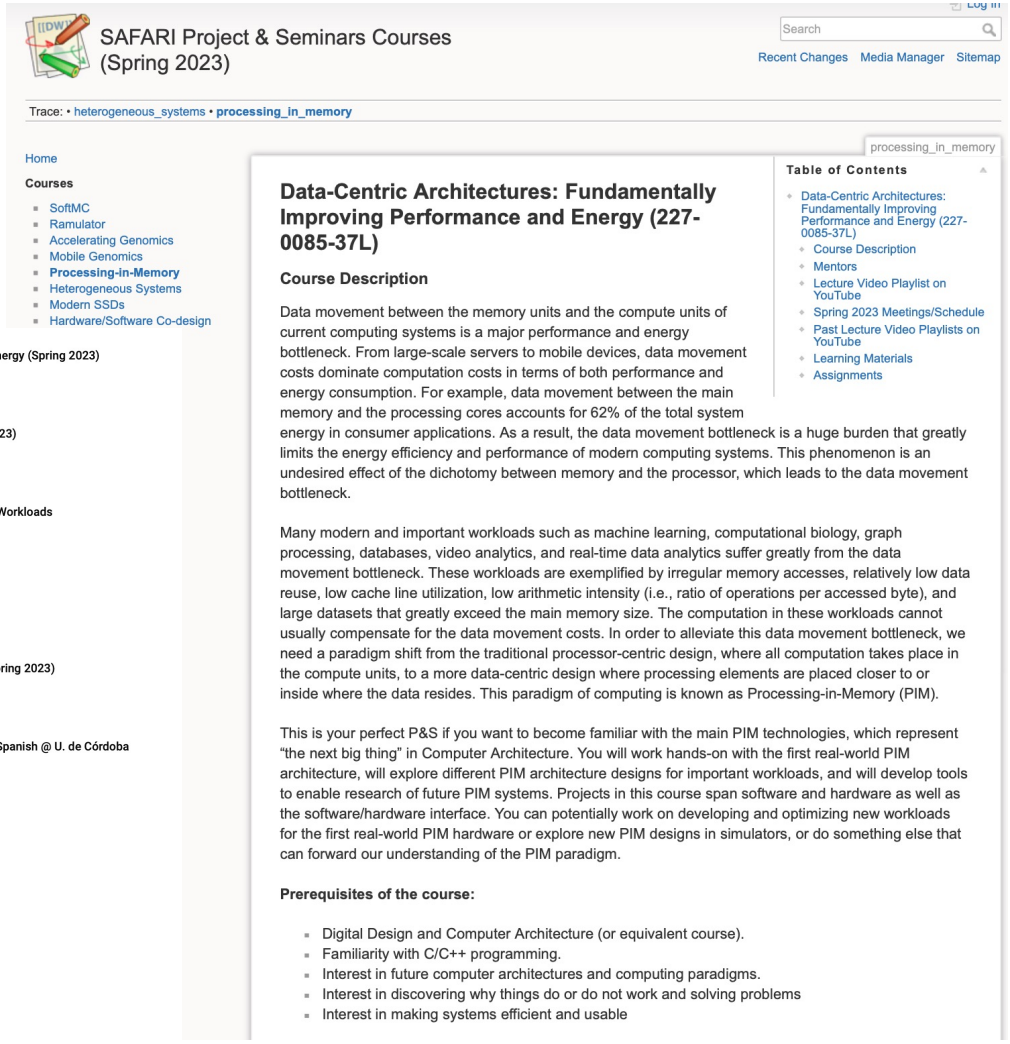
Processing-in-Memory Course (Spring 2023)

- Short weekly lectures
- Hands-on projects





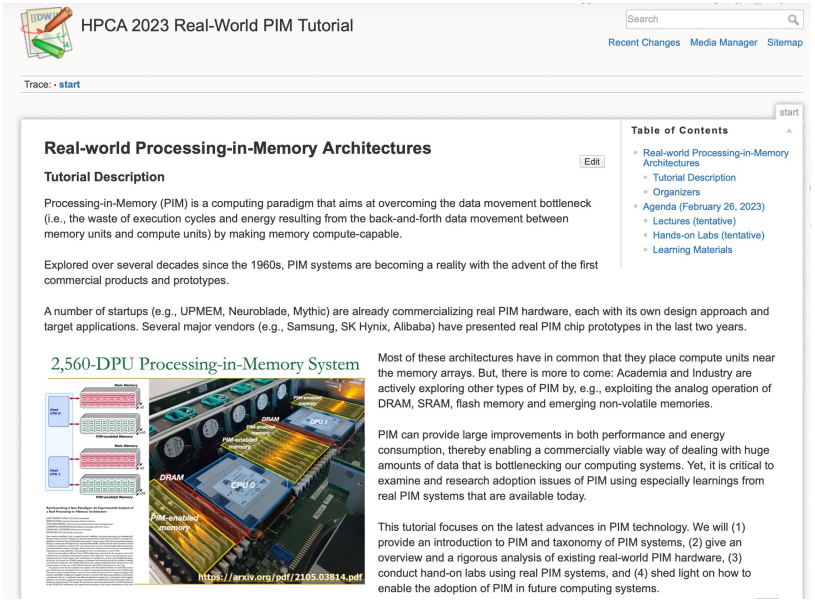
https://www.youtube.com/playlist?list=PL5Q2soXY2Zi_EObuoAZVSq_o6UySWQHvZ

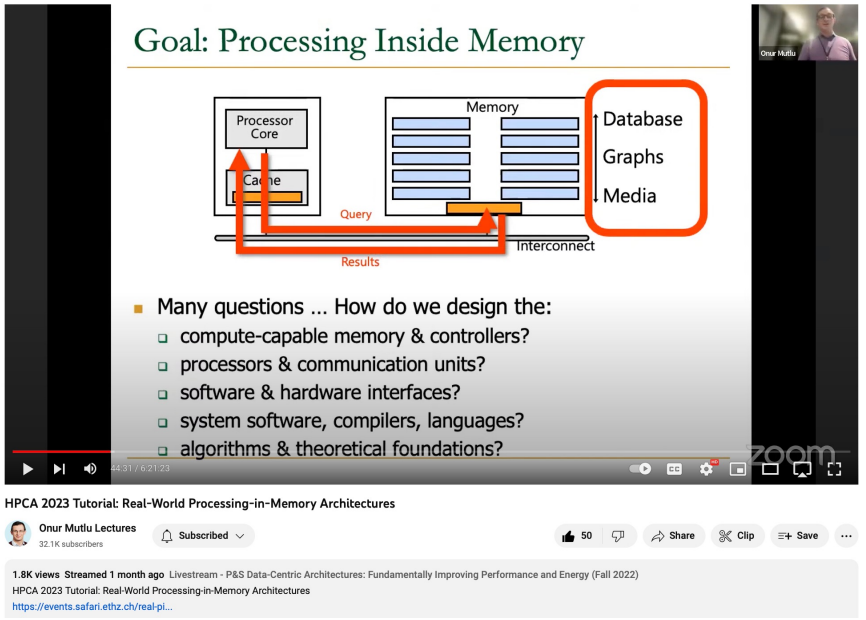


https://safari.ethz.ch/projects_and_seminars/spring2023/doku.php?id=processing_in_memory

Real PIM Tutorial (HPCA 2023)

- February 26th: Lectures + Hands-on labs + Invited lectures





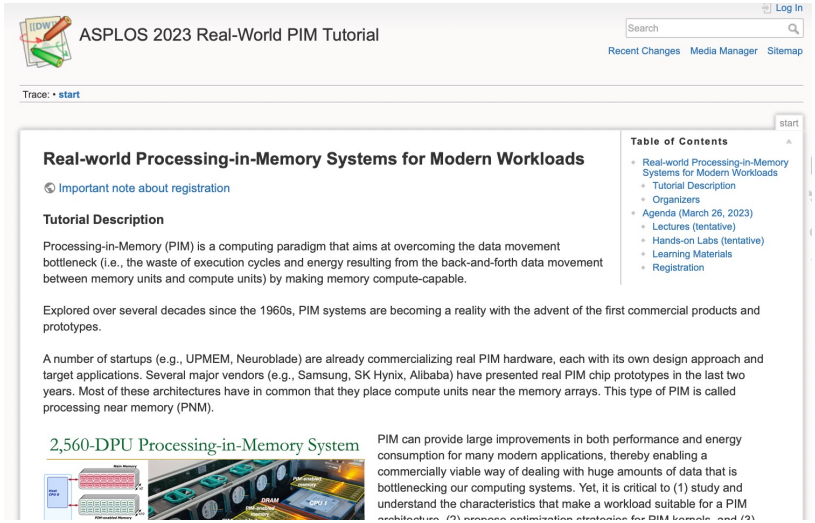
Time	Speaker	Title	Materials
8:00am-8:40am	Prof. Onur Mutlu	Memory-Centric Computing	(PDF) (PPT)
8:40am-10:00am	Dr. Juan Gómez Luna	Processing-Near-Memory: Real PNM Architectures Programming General-purpose PIM	(PDF) (PPT)
10:20am-11:00am	Dr. Dimin Niu	A 3D Logic-to-DRAM Hybrid Bonding Process-Near-Memory 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) (PPT)
1:30pm-2:10pm	Dr. Juan Gómez Luna	Processing-Using-Memory: Exploiting the Analog Operational Properties of Memory Components	(PDF) (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) (PPT)
3:40pm-5:40pm	Dr. Juan Gómez Luna	Hands-on Lab: Programming and Understanding a Real Processing-in-Memory Architecture	(Handout) (PDF) (PPT)

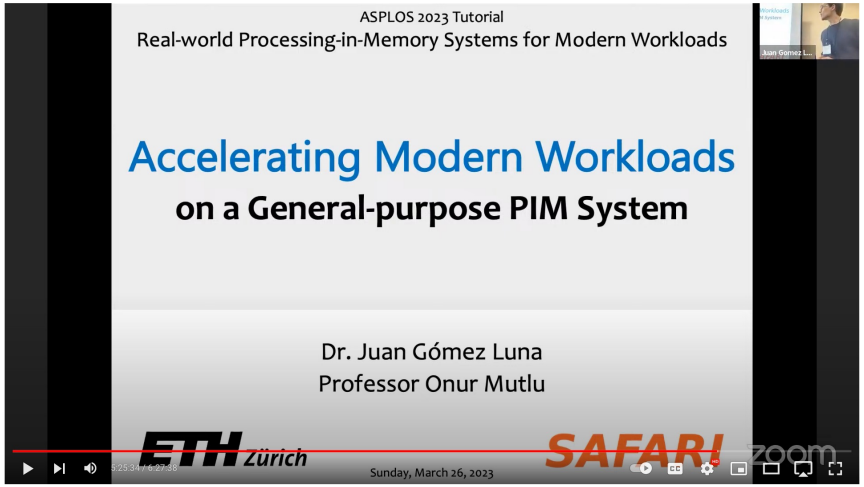
https://www.youtube.com/live/f5-nT1tbz5w?feature=share

https://events.safari.ethz.ch/real-pim-tutorial/doku.php?id=start

Real PIM Tutorial (ASPLOS 2023)

- March 26th: Lectures + Hands-on labs + Invited lectures





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

Onur Mutlu Lectures
32.1K subscribers
Subscribed
33
Share
Clip
Save
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
://events.safari.ethz.ch/asplos-...

https://www.youtube.com/live/oYCaLcT0Kmo?feature=share

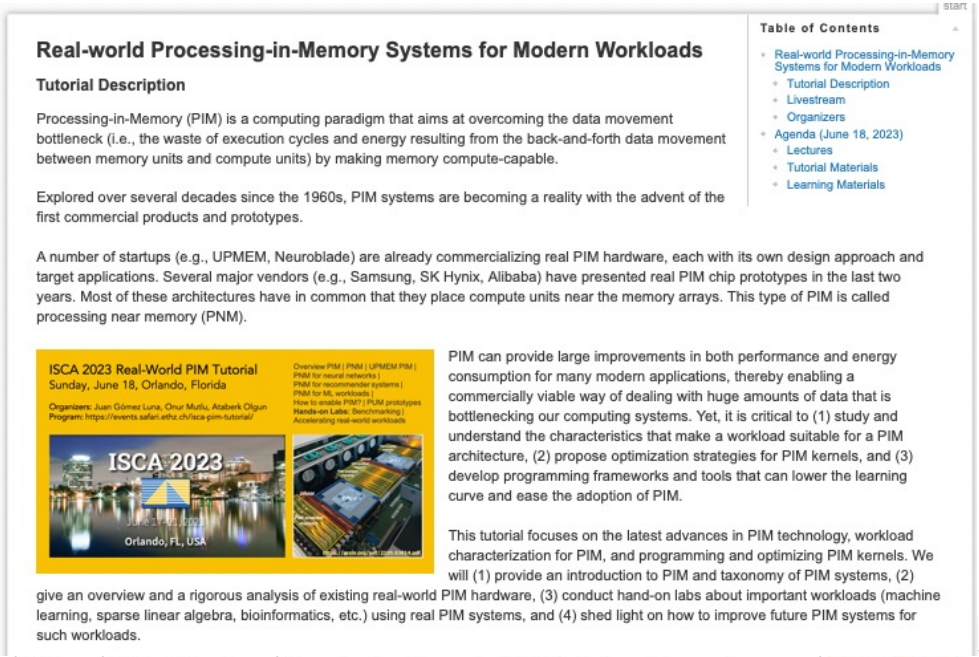
Tutorial Materials

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

https://events.safari.ethz.ch/asplos-pim-tutorial/doku.php?id=start

Real PIM Tutorial (ISCA 2023)

- June 18th: Lectures + Hands-on labs + Invited lectures



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).

Table of Contents

- Real-world Processing-in-Memory Systems for Modern Workloads
  - Tutorial Description
  - Livestream
  - Organizers
- Agenda (June 18, 2023)
  - Lectures
  - Tutorial Materials
  - Learning Materials

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.

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)
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) (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) (PDF) (PPT)

SAMSUNG

Introducing Real-world HBM-PIM Powered System for Memory-bound Applications

Samsung Electronics
DRAM Design Team
Sukhan Lee

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

Onur Mutlu Lectures
36.6K subscribers
Subscribed
61
Share
Clip
Save

1,976 views Streamed live on Jun 18, 2023 Livestream - Data-Centric Architectures: Fundamentally Improving Performance and Energy (Spring 2023)
ISCA 2023 Tutorial: Real-world Processing-in-Memory Systems for Modern Workloads
https://events.safari.ethz.ch/isca-pi...

https://www.youtube.com/live/GIb5EgSrWk0?si=Ws5bzE4UhHsj-mII

https://events.safari.ethz.ch/isca-pim-tutorial/doku.php?id=start

Real PIM Tutorial (MICRO 2023)

- October 29th: Lectures + Hands-on labs + Invited lectures

MICRO 2023 Real-World PIM Tutorial

Search

Recent Changes Media Manager Sitemap

Trace: • start

start

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

Edit

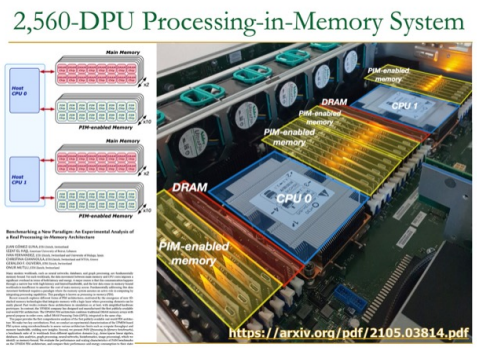
Tutorial Description

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.

Table of Contents

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.

Edit

Livestream

YouTube livestream

https://youtube.com/live/ohU00NSIxOI?feature=share

Agenda (EDT GMT-4)

- 7:55am-8:00am, Dr Juan Gómez Luna, "Welcome & Agenda."
- 8:00am-9:20am, Prof. Onur Mutlu / Geraldo F. Oliveira, "Memory-centric Computing: Introduction to PIM as a Paradigm to Overcome the Data Movement Bottleneck."
  - PIM taxonomy: PNM (processing near memory) and PUM (processing using memory).
  - DAMOV Workload Characterization Methodology.
- 9:20am-10:20am, Dr. Juan Gómez Luna, "Processing-Near-Memory: Real PNM."
  - PNM prototypes: Samsung HBM-PIM, SK Hynix AiM, Samsung AxDIMM, Alibaba HB-PNM.
  - UPMEM PIM: Architecture Characterization, Programming.
- Coffee break (10:20am-10:40am)
- 10:40am-11:20am, Prof. Youngsok Kim (Yonsei University), "PID-Join: A Fast In-Memory Join Algorithm for Commodity PIM-Enabled DIMMs."
- 11:20am-12:00pm, Dr. Abu Sebastian (IBM Research - Zürich), "PUM Based on Memristive Devices: The IBM HERMES Project Chip."
- Lunch break (12:00pm-1:00pm)
- 1:00pm-2:00pm, Geraldo F. Oliveira, "Processing-Using-DRAM: Ambit, SIMDRAM, pLUTo."
- 2:00pm-3:15pm, Dr. Juan Gómez Luna, "Accelerating Modern Workloads on a General-purpose PIM System: Machine leaning, Genomics..."
- 3:15pm-3:45pm, Dr. Juan Gómez Luna, "Adoption Issues: How to Enable PIM?"
- 3:45pm-4:15pm, Dr. Juan Gómez Luna, "SimplePIM: A Software Framework for High-level PIM Programming."
- 4:15pm-5:00pm, Ataberk Olgun, "Processing-Using-Memory Prototypes: PiDRAM."
- 5:00pm-5:10pm, Dr. Juan Gómez Luna, "Introduction/Preparation for Hands-on Labs."
  - Optional - Hands-on Lab: Programming and Understanding a Real PIM Architecture.

Join Algorithm for Commodity PIM DIMMs

- PID-Join: A Fast In-Memory Join Algorithm for Commodity PIM-Enabled DIMMs
  - Prof. Youngsok Kim

- **Abstract:** Modern dual in-line memory modules (DIMMs) now allow applications to offload their memory-intensive operations onto the in-DIMM processors (IDPs) located near the memory banks of a DIMM. The IDPs can greatly accelerate in-memory joins whose performance is frequently bottlenecked by main-memory accesses. In this talk, I will present PID-Join, a fast in-memory join algorithm which exploits the IDPs to further improve the performance of in-memory joins. PID-Join is optimized for UPMEM DIMMs, currently the only publicly-available DIMMs equipped with the IDPs. PID-Join achieves significant speedups over the existing CPU-based join implementations by prototyping and evaluating hash, sort-merge, and nested-loop algorithms optimized for the IDPs, enabling fast inter-IDP communication using host CPU cache streaming and vector instructions, and facilitating fast rank-wise data transfers between the IDPs and the main memory.

- **Bio:** Youngsok Kim is currently an assistant professor with the Department of Computer Science at Yonsei University. His research interests span computer architecture and system software with an emphasis on architecture-conscious database management systems, performance modeling, and architectural support for deep learning. He received his BSc and PhD in Computer Science and Engineering from POSTECH. He was a post-doc researcher at Seoul National University before joining Yonsei University, and was an intern at Consumer Hardware, Google Inc. and S.LSI Business, Samsung Electronics Co., Ltd. during his PhD studies.

PUM based on Memristive Devices

- PUM based on Memristive Devices: The IBM HERMES Project Chip
  - Dr. Abu Sebastian

- **Abstract:** I will introduce analog in-memory computing (or processing using memory) based on memristive technology. A multi-tile, mixed-signal AIMC prototype chip (The IBM HERMES Project Chip) aimed at deep learning inference application will be presented. This chip fabricated in 14nm CMOS technology comprises 64 AIMC cores/tiles based on phase-change memory technology. I will delve deeper into the device, circuits, architectural and algorithmic aspects of AIMC based on this prototype chip. Finally, I will conclude with a brief overview of on-going research activities.

- **Bio:** Dr. Abu Sebastian is one of the technical leaders of IBM's research efforts towards next generation AI Hardware and manages the in-memory computing group at IBM Research - Zurich. He is the author of over 200 publications in peer-reviewed journals/conference proceedings and holds over 90 US patents. In 2015 he was awarded the European Research Council (ERC) consolidator grant and in 2020, he was awarded an ERC Proof-of-concept grant. He was an IBM Master Inventor and was named Principal and Distinguished Research Staff Member in 2018 and 2020, respectively. In 2019, he received the Ovshinsky Lectureship Award for his contributions to "Phase-change materials for cognitive computing". In 2023, he was conferred the title of Visiting Professor in Materials by University of Oxford. He is a distinguished lecturer and fellow of the IEEE. Homepage: https://research.ibm.com/people/abu-sebastian

International Symposium on Microarchitecture (MICRO)

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

https://ethz.zoom.us/j/68459763236?pwd=NEIzK1JGV0pHTjRaUzV5NmRna3JkZz09

https://www.youtube.com/live/ohU00NSIxOI?si=h1I9eUx7KCScMHrg

October 29th, 2023

SAFARI