International Symposium on Computer Architecture (ISCA)

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

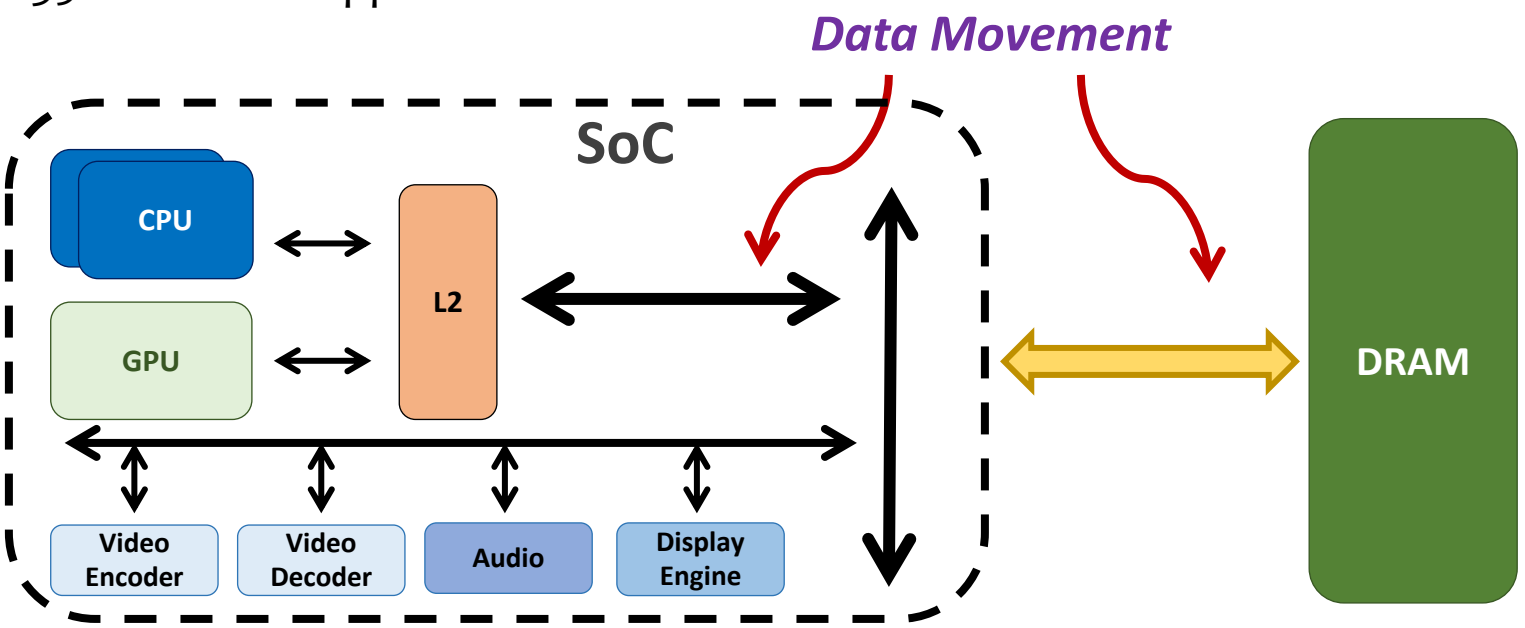
https://www.youtube.com/live/GIb5EgSrWk0?feature=share

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



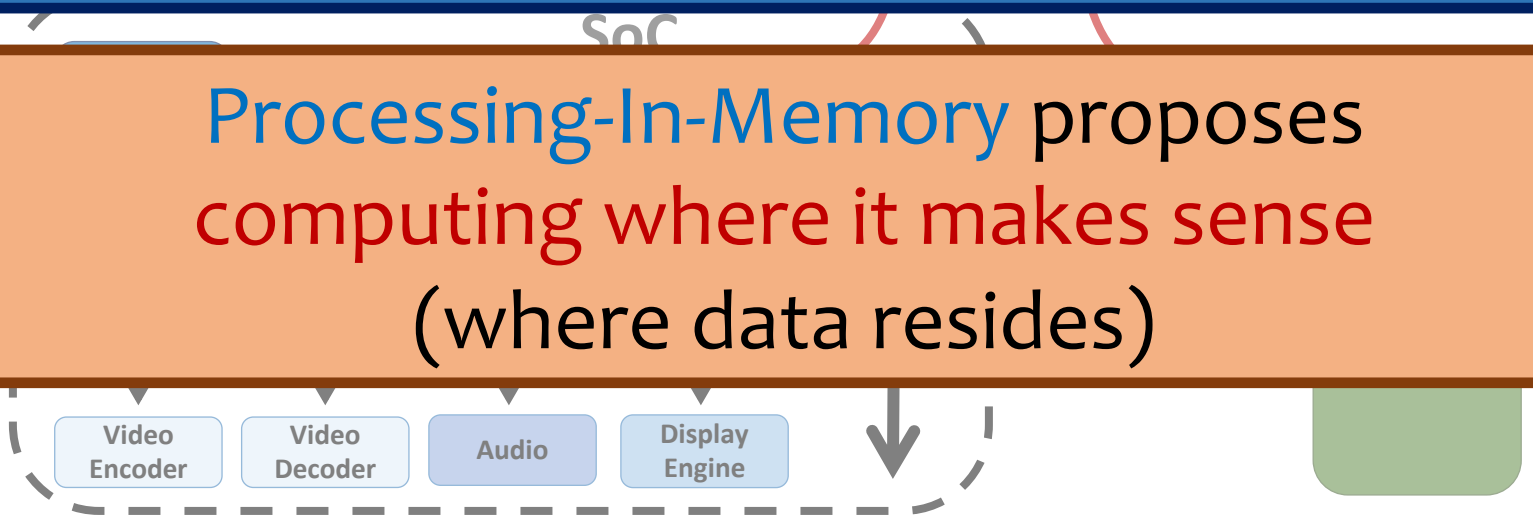
* Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS 2018
★ Kestor et al., "Quantifying the Energy Cost of Data Movement in Scientific Applications," IISWC 2013
☆ Pandiyan and Wu, "Quantifying the energy cost of data movement for emerging smart phone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

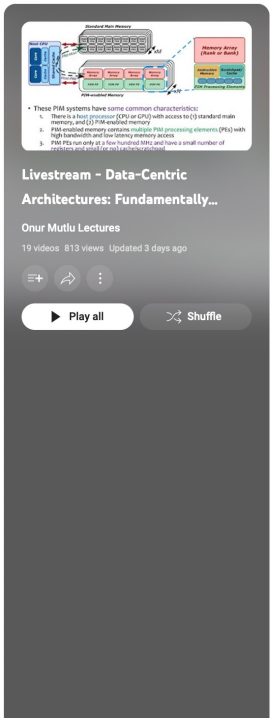
Processing-In-Memory proposes computing where it makes sense (where data resides)

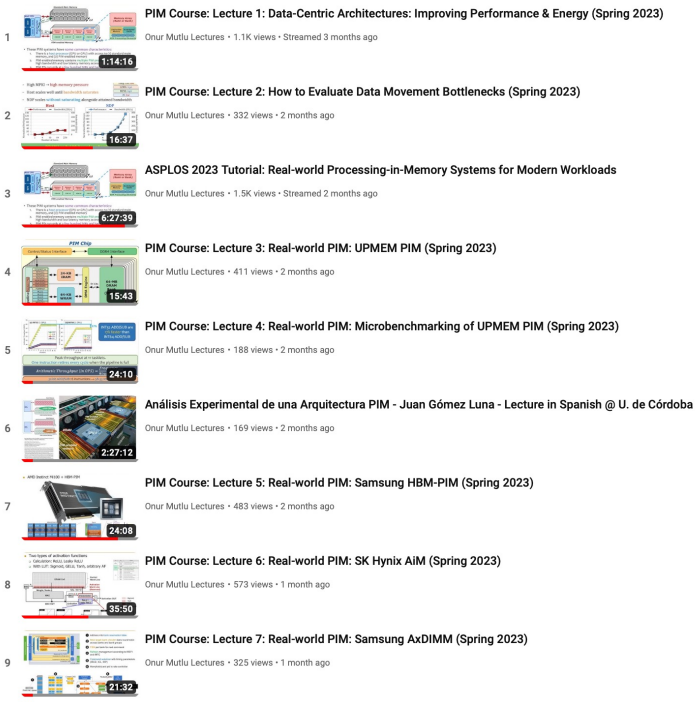


* Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS 2018
★ Kestor et al., "Quantifying the Energy Cost of Data Movement in Scientific Applications," IISWC 2013
☆ Pandiyan and Wu, "Quantifying the energy cost of data movement for emerging smart phone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

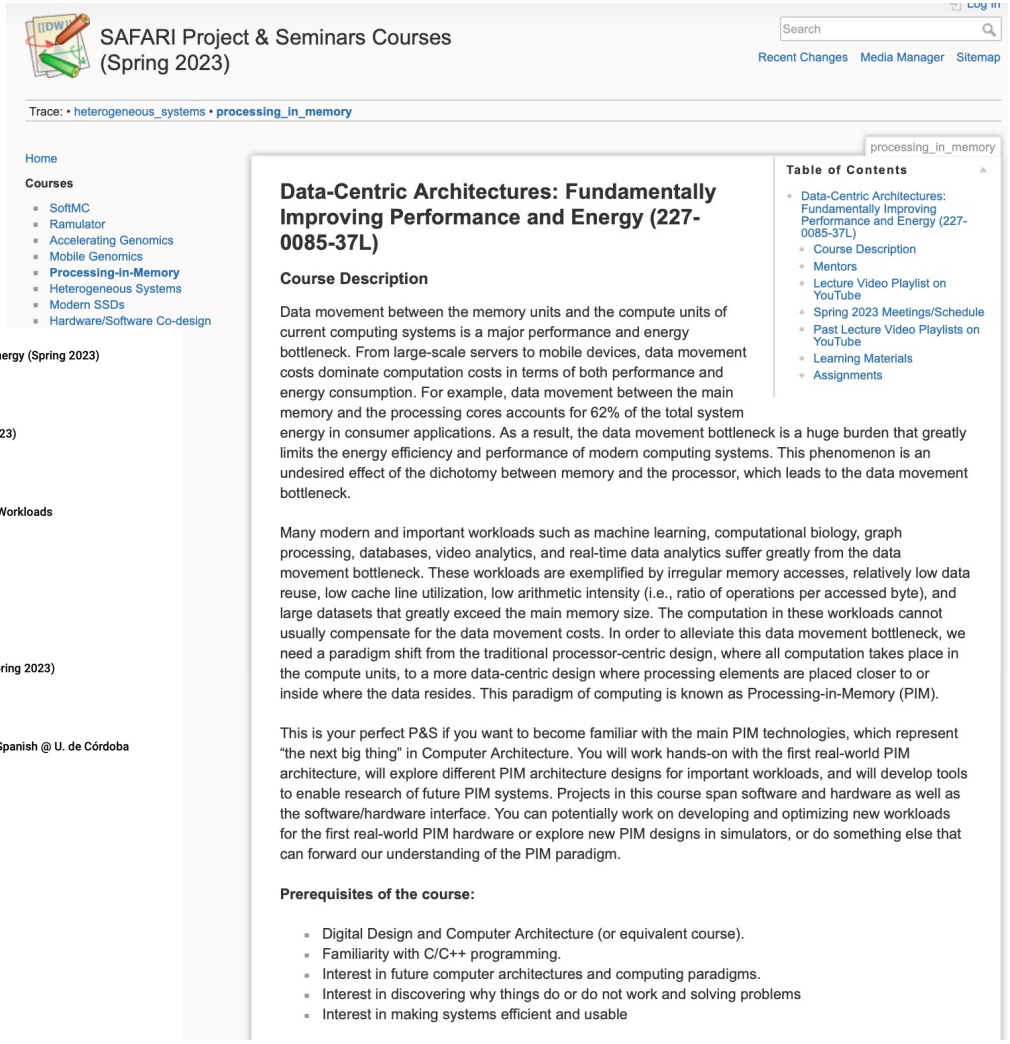
Processing-in-Memory Course (Spring 2023)

- Short weekly lectures
- Hands-on projects





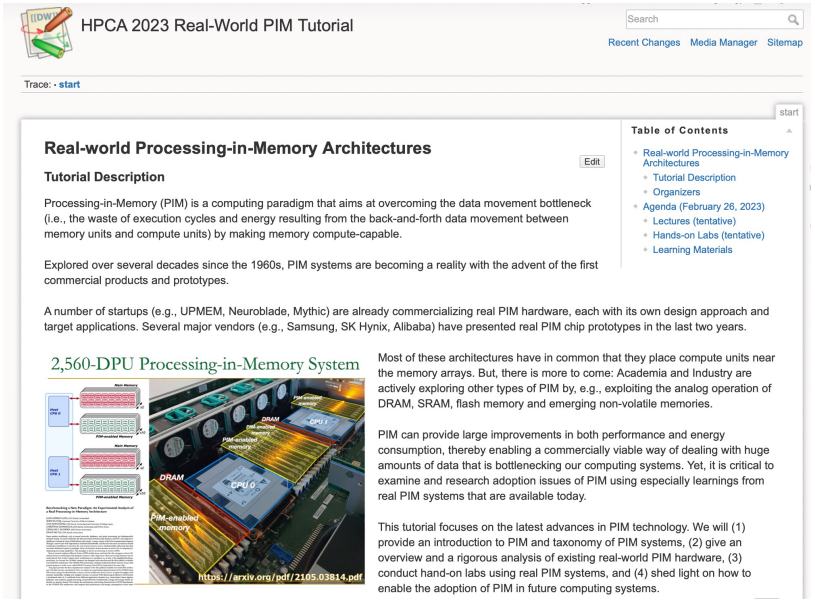
https://www.youtube.com/playlist?list=PL5Q2soXY2Zi_EObuoAZVSq_o6UySWQHvZ

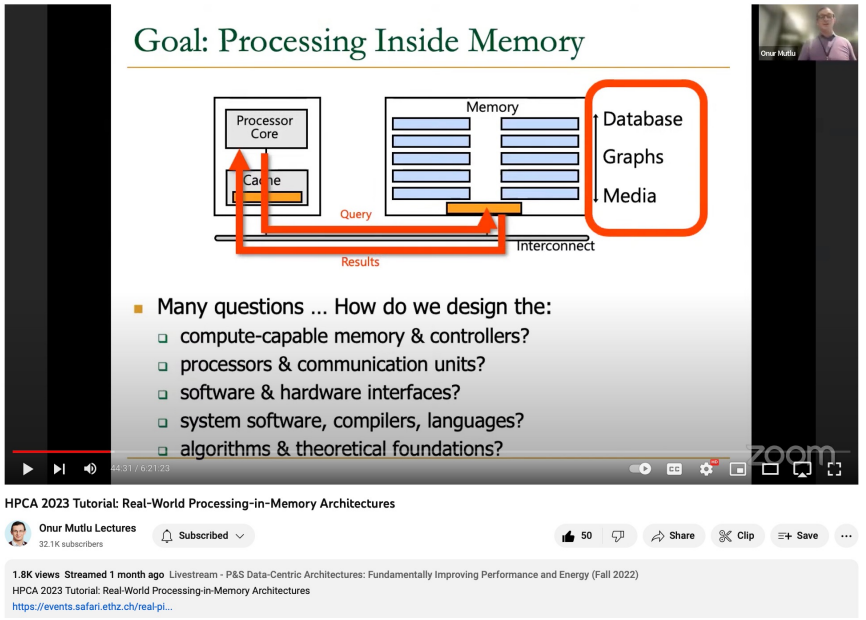


https://safari.ethz.ch/projects_and_seminars/spring2023/doku.php?id=processing_in_memory

Real PIM Tutorial (HPCA 2023)

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





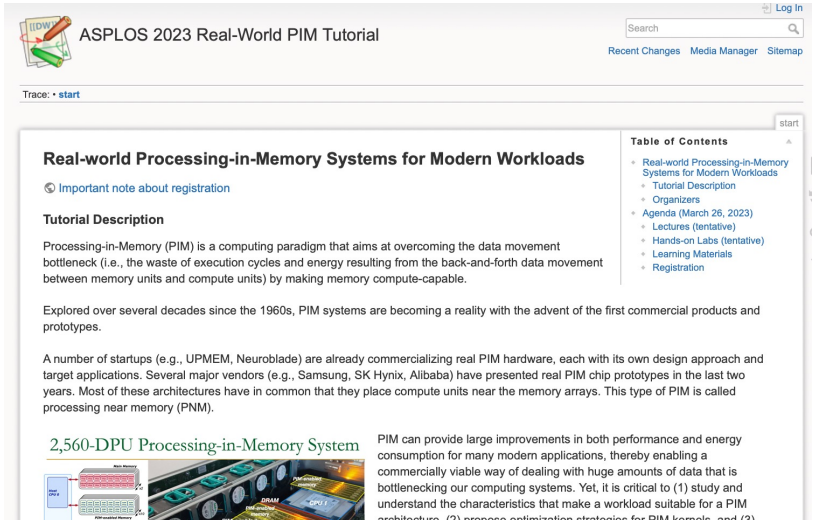
Time	Speaker	Title	Materials
8:00am-8:40am	Prof. Onur Mutlu	Memory-Centric Computing	(PDF) (PPT)
8:40am-10:00am	Dr. Juan Gómez Luna	Processing-Near-Memory: Real PNM Architectures Programming General-purpose PIM	(PDF) (PPT)
10:20am-11:00am	Dr. Dimin Niu	A 3D Logic-to-DRAM Hybrid Bonding Process-Near-Memory 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)

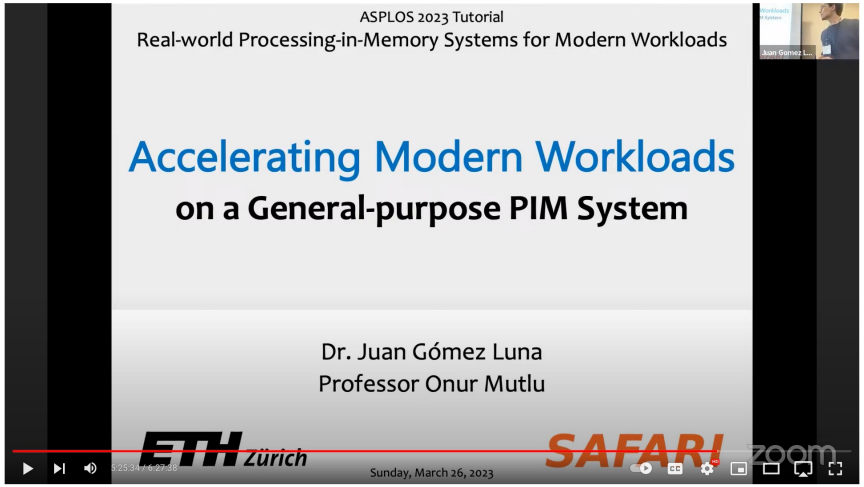
https://www.youtube.com/live/f5-nT1tbz5w?feature=share

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

ISCA 2023 Real-World PIM Tutorial

Search

Recent Changes Media Manager Sitemap

Trace: • start

start

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

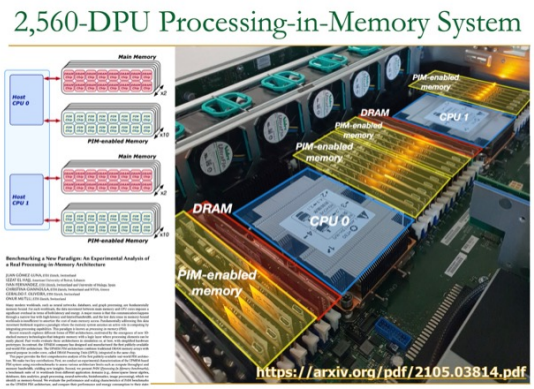
Table of Contents

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.

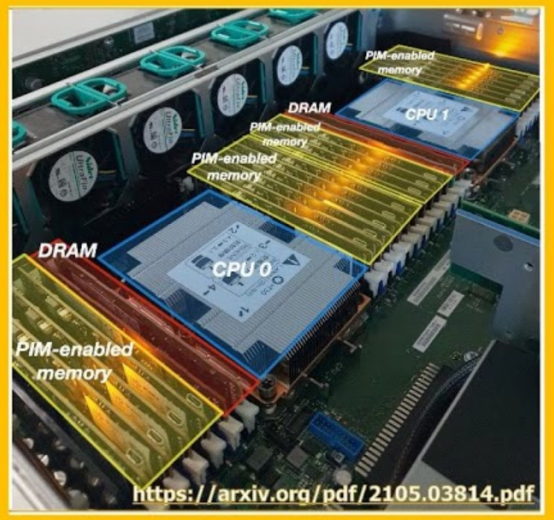
This tutorial focuses on the latest advances in PIM technology, workload characterization for PIM, and programming and optimizing PIM kernels. We will (1) provide an introduction to PIM and taxonomy of PIM systems, (2) give an overview and a rigorous analysis of existing real-world PIM hardware, (3) 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.

Real PIM Tutorial (ISCA 2023)

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



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

Onur Mutlu Lectures
33.4K subscribers
Subscribed
8
Share
Save

1 waiting Scheduled for 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://www.youtube.com/live/GIb5EgSrWk0?feature=share

Agenda

- 9:00am-10:20am, Prof. Onur Mutlu/Dr. Juan Gómez Luna, "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).
- 10:20am-11:00pm, 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 and Programming.
- Coffee break (11:00am-11:20am)
- 11:20am-11:50am, Prof. Izzat El Hajj (AUB), "High-throughput Sequence Alignment using Real Processing-in-Memory Systems".
- 11:50am-12:30pm, Dr. Christina Giannoula (UofT), "SparseP: Towards Efficient Sparse Matrix Vector Multiplication on Real Processing-In-Memory Systems".
- Lunch break (12:30pm-2:00pm)
- 2:00pm-2:45pm, Dr. Sukhan Lee (Samsung), "Introducing Real-world HBM-PIM Powered System for Memory-bound Applications".
- 2:45pm-3:30pm, Dr. Juan Gómez Luna/Ataberk Olgun, "Processing-Using-Memory and PUM Prototypes: Ambit/SIMDRAM, PiDRAM".
- Coffee break (3:30pm-4pm)
- 4:00pm-4:40pm, Dr. Juan Gómez Luna, "Accelerating Modern Workloads on a General-purpose PIM System".
- 4:40pm-5:20pm, Dr. Juan Gómez Luna, "Adoption Issues: How to Enable PIM?"
- 5:20pm-5:30pm, Dr. Juan Gómez Luna, "Introduction/Preparation for Hands-on labs".
- Optional - Hands-on Lab: Programming and Understanding a Real PIM Architecture.

Sequence Alignment on Real PIM

- High-throughput Sequence Alignment using Real Processing-in-Memory Systems
  - Prof. Izzat El Hajj

- **Abstract:** Sequence alignment is a memory bound computation whose performance in modern systems is limited by the memory bandwidth bottleneck. Processing-in-memory architectures alleviate this bottleneck by providing the memory with computing competencies. We present Alignment-in-Memory (AIM), a framework for high-throughput sequence alignment using processing-in-memory which we have implemented and evaluated it on UPMEM, the first publicly-available general-purpose programmable processing-in-memory system. Our evaluation shows that a real processing-in-memory system can substantially outperform server-grade multi-threaded CPU systems running at full-scale when performing sequence alignment for a variety of algorithms, read lengths, and edit distance thresholds. We hope that our findings inspire more work on creating and accelerating bioinformatics algorithms for such real processing-in-memory systems. Our code is available at: https://github.com/safaad/aim.

- **Bio:** Izzat El Hajj is an Assistant Professor in the Department of Computer Science at the American University of Beirut. His research interests are in application acceleration and programming support for emerging parallel processors and memory technologies, with a particular interest in GPUs and processing-in-memory. Izzat received his M.S. and Ph.D. in Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign and his B.E. in Electrical and Computer Engineering at the American University of Beirut.

SpMV on a Real PIM System

- SparseP: Towards Efficient Sparse Matrix Vector Multiplication for Real Processing-In-Memory Systems
  - Dr. Christina Giannoula

- **Abstract**: Several manufacturers have already started to commercialize near-bank Processing-In-Memory (PIM) architectures. Near-bank PIM architectures place simple cores close to DRAM banks and can yield significant performance and energy improvements in parallel applications by alleviating data access costs. Real PIM systems can provide high levels of parallelism, large aggregate memory bandwidth and low memory access latency, thereby being a good fit to accelerate the widely-used, memory-bound Sparse Matrix Vector Multiplication (SpMV) kernel.

  This talk provides the first comprehensive analysis of SpMV on a real-world PIM architecture, and presents SparseP, the first SpMV library for real PIM architectures. We will discuss three key contributions that we make in our Sigmetrics 2022 paper. First, we implement a wide variety of software strategies on SpMV for a multithreaded PIM core and characterize the computational limits of a single multithreaded PIM core. Second, we design various load balancing schemes across multiple PIM cores, and two types of data partitioning techniques to execute SpMV on thousands of PIM cores: (1) 1D-partitioned kernels to perform the complete SpMV computation only using PIM cores, and (2) 2D-partitioned kernels to strive a balance between computation and data transfer costs to PIM-enabled memory. Third, we compare SpMV execution on a real-world PIM system with 2528 PIM cores to state-of-the-art CPU and GPU systems to study the performance and energy efficiency of various devices. SparseP software package provides 25 SpMV kernels for real PIM systems supporting the four most widely used compressed matrix formats, and a wide range of data types. Our extensive evaluation provides new insights and recommendations for software designers and hardware architects to efficiently accelerate SpMV on real PIM systems.

- **Bio**: Christina Giannoula is a Postdoctoral Researcher at the University of Toronto working with Prof. Gennady Pekhimenko and the EcoSystem research group. She is also working with the SAFARI research group, which is led by Prof. Onur Mutlu. She received her Ph.D. in October 2022 from School of Electrical and Computer Engineering (ECE) at the National Technical University of Athens (NTUA) advised by Prof. Georgios Goumas, Prof. Nectarios Koziris and Prof. Onur Mutlu. Her research interests lie in the intersection of computer architecture, computer systems and high-performance computing. Specifically, her research focuses on the hardware/software co-design of emerging applications, including graph processing, pointer-chasing data structures, machine learning workloads, and sparse linear algebra, with modern computing paradigms, such as large-scale multicore systems, disaggregated memory systems and near-data processing architectures. She has several publications and awards for her research on these topics.

Samsung HBM-PIM

- Introducing Real-world HBM-PIM Powered System for Memory-bound Applications
  - Dr. Sukhan Lee
- **Abstract:** Since the introduction of Samsung's groundbreaking high bandwidth memory with in-memory processing (HBM-PIM) in 2021, a number of HBM-PIM enabled systems, including a GPU cluster and FPGA, have been developed. These advancements have been presented and showcased at various conferences and journals, spanning from ISSCC 2021 to Memcon 2023. In this tutorial, we provide a comprehensive overview of HBM-PIM, covering its architectural aspects, the associated software ecosystem, and the structure of PIM-powered systems. We also present energy-efficient performance results for memory-bound applications achieved by these systems.
- **Bio:** Sukhan Lee received a Ph.D. degree in intelligent convergence systems from Seoul National University. In 2018, he joined the Memory Division, Samsung Electronics, Hwaseong, Korea, where he has been involved in DRAM circuit design. His research interests include memory microarchitecture and neural network system hardware architecture and design.

International Symposium on Computer Architecture (ISCA)

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

https://www.youtube.com/live/GIb5EgSrWk0?feature=share

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

SAFARI