Tutorial on Memory-Centric Computing: Conclusion Remarks

Geraldo F. Oliveira
Prof. Onur Mutlu

ISCA 2024
29 June 2024
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
- 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\textsuperscript{a,b}, Saugata Ghose\textsuperscript{b,c}, Juan Gómez-Luna\textsuperscript{a}, Rachata Ausavarungnirun\textsuperscript{d}

SAFARI Research Group

\textsuperscript{a}ETH Zürich  
\textsuperscript{b}Carnegie Mellon University  
\textsuperscript{c}University of Illinois at Urbana-Champaign  
\textsuperscript{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"

SAFARI

Special Research Sessions & Courses (I)

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

https://www.youtube.com/watch?v=qeukNs5XI3g
Special Session at ISVLSI 2022: 9 cutting-edge talks

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

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

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

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

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

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

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

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

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

https://www.youtube.com/playlist?list=PL5Q2soXY2Zi8KzG2CQYRNQOVD0G0BrnKy
Short weekly lectures

Hands-on projects

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 further our understanding of the PIM paradigm.

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](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](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=PL5Q2soXY2Zi8KzG2CQYRNQOVD0G0BrnK](https://www.youtube.com/watch?v=QLL0wQ9I4Dw&list=PL5Q2soXY2Zi8KzG2CQYRNQOVD0G0BrnK)
- Youtube Livestream (Spring 2022):
  - [https://www.youtube.com/watch?v=9e4Chnwdovo&list=PL5Q2soXY2Zi-841fUYYUK9EsXKhQKRPyX](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](https://www.youtube.com/onurmutlulectures)
Processing-in-Memory Course (Spring 2023)

- Short weekly lectures
- Hands-on projects

https://www.youtube.com/playlist?list=PL5Q2soXYZZi_EObuoAZVsQ_o6UySWQHz

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

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., UPEM, Neubio, and Alibaba) 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

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/

![Real PIM Tutorial](https://events.safari.ethz.ch/isca-pim-tutorial/)

**Tutorial Materials**

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

Watch the tutorial on YouTube: [https://www.youtube.com/live/GIb5EgSrWk0?feature=share](https://www.youtube.com/live/GIb5EgSrWk0?feature=share)

Watch the tutorial on Safari: [https://events.safari.ethz.ch/isca-pim-tutorial/](https://events.safari.ethz.ch/isca-pim-tutorial/)

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

International Symposium on Computer Architecture (ISCA)

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

[https://www.youtube.com/live/GIb5EgSrWk0?feature=share](https://www.youtube.com/live/GIb5EgSrWk0?feature=share)

[https://events.safari.ethz.ch/isca-pim-tutorial/](https://events.safari.ethz.ch/isca-pim-tutorial/)
Real PIM Tutorial [ASPLOS 2023]

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

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

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?

Directory:
- Graphs
- Media

https://www.youtube.com/watch?v=f5-nT1tbz5w

https://events.safari.ethz.ch/real-pim-tutorial/
Real PIM Tutorial [MICRO 2023]

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

2,560-DPU Processing-in-Memory System

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

Table of Contents
- Real-world Processing-in-Memory Systems for Modern Workloads
  - Tutorial Description
  - Agenda (Tentative, October 29, 2023)

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 1980s, PIM systems are becoming a reality with the advent of the first commercial products and prototypes.

A number of startups (e.g., UPMEM, Neuromatter) are already commercializing real PIM hardware, each with its own design philosophy 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).

PMI 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.

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 AvDIMM, 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://www.youtube.com/live/ohUooNS1xQI

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 & tools
Infrastructures for PIM Research
Research challenges & opportunities

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
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 University
- [https://safari.ethz.ch/](https://safari.ethz.ch/)
- omutlu@gmail.com

**Pinned**

- **ramulator** (Public)
  - A Fast and Extensible DRAM Simulator, with built-in support for modeling many different DRAM technologies including DDRx, LPDDRx, GDDRx, WI0x, HBMx, and various academic proposals. Described in the...
  - C++
  - 442 stars
  - 195 forks

- **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 publi...
  - C
  - 100 stars
  - 38 forks

- **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 stars
  - 120 forks

- **rowhammer** (Public)
  - C
  - 208 stars
  - 41 forks

- **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 limitation...
  - Verilog
  - 104 stars
  - 26 forks

- **Pythia** (Public)
  - C++
  - 85 stars
  - 25 forks

[https://github.com/CMU-SAFARI/](https://github.com/CMU-SAFARI/)
Tutorial on Memory-Centric Computing: Conclusion Remarks

Geraldo F. Oliveira
Prof. Onur Mutlu

ISCA 2024
29 June 2024