In-Memory Processing

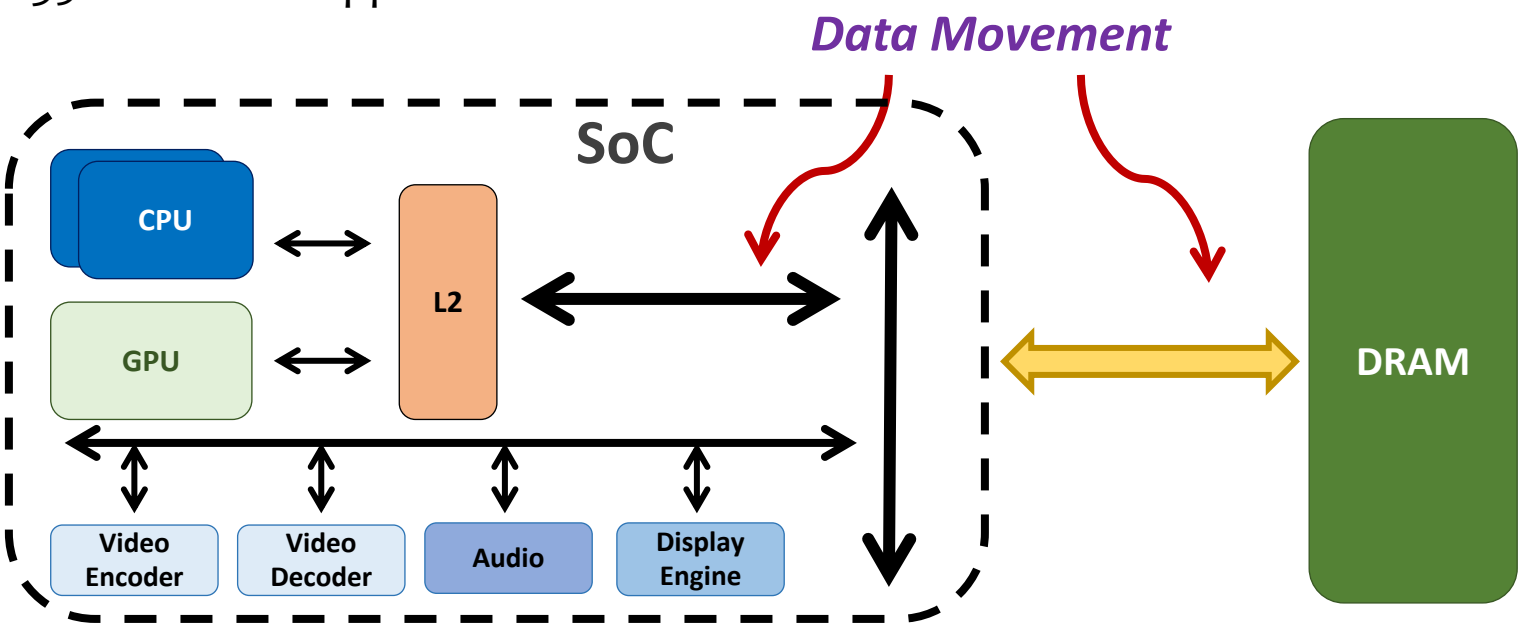
ISVLSI 2022 Special Session

IEEE Computer Society Annual Symposium on VLSI

Adonis room
Ailathon resort, Paphos, Cyprus
July 4th, 2022

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☆



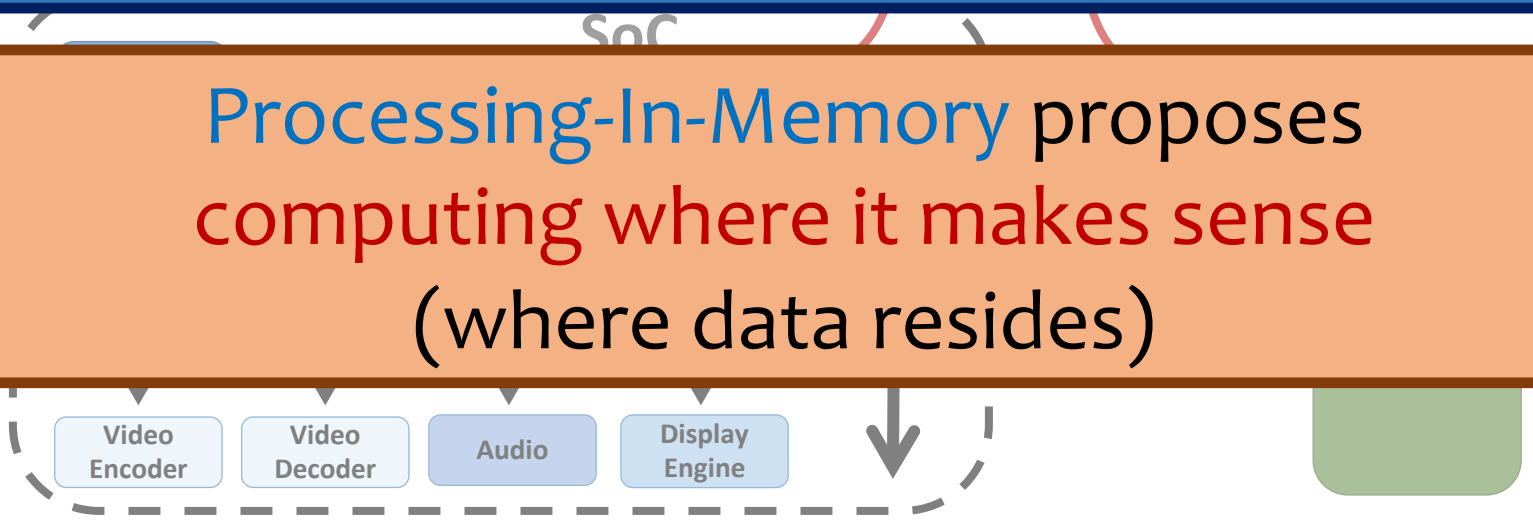
* Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS 2018
★ Kestor et al., "Quantifying the Energy Cost of Data Movement in Scientific Applications," IISWC 2013
☆ Pandiyan and Wu, "Quantifying the energy cost of data movement for emerging smart phone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

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

In-Memory Processing

- 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, PIM is nowadays becoming a reality with the advent of the UPMEM PIM architecture, the first commercially-available PIM architecture, and the recent announcements of new PIM architectures by major DRAM vendors (Samsung, SK Hynix). These architectures have in common that they place compute units near the memory arrays. But there is more to come: Academia and Industry keep actively exploring other types of PIM by, e.g., exploiting the analog operation of SRAM, DRAM, or non-volatile memories. As a result, PIM may become ubiquitous in the next few years by providing orders-of-magnitude improvements in performance and energy efficiency.

- This special session focuses on the latest advances in PIM technology. Either software for & benchmarking of real-world PIM architectures, or hardware & architecture proposals for future PIM systems are welcome to this special session.

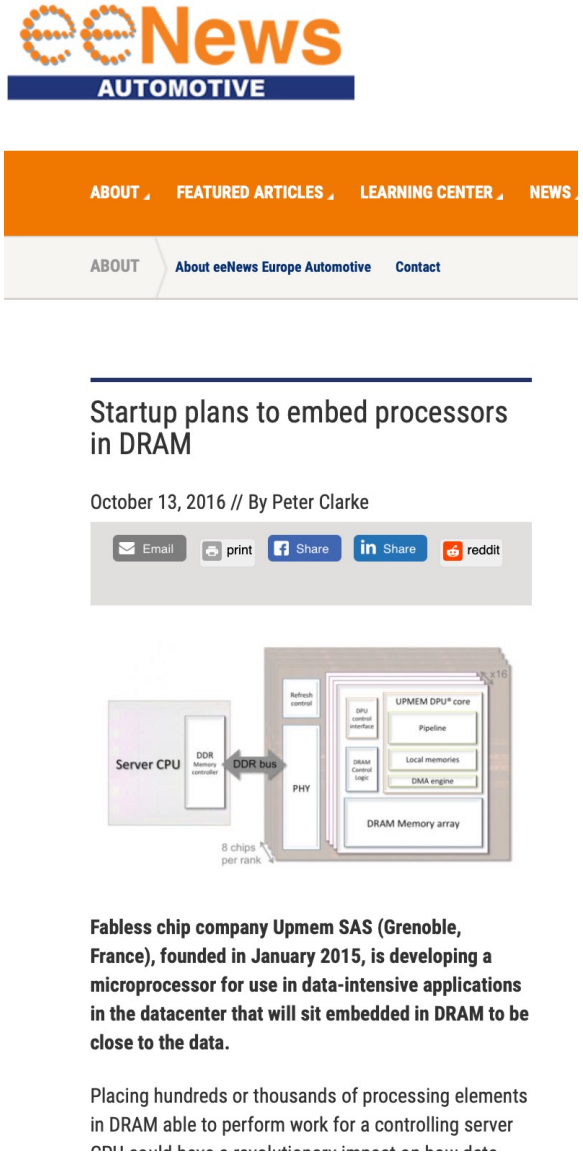
- Organizers: Dr. Juan Gómez Luna, Professor Onur Mutlu (ETH Zürich)

PIM Becomes Real

- UPMEM, founded in January 2015, announces the first real-world PIM architecture in 2016
- UPMEM's PIM-enabled DIMMs start getting commercialized in 2019

- In early 2021, Samsung announces FIMDRAM at ISSCC conference
- Samsung's LP-DDR5 and DIMM-based PIM announced a few months later

- In early 2022, SK Hynix announces AiM and Alibaba announces HB-PNM at ISSCC conference



eeNews AUTOMOTIVE

ABOUT | FEATURED ARTICLES | LEARNING CENTER | NEWS

ABOUT | About eeNews Europe Automotive | Contact

Startup plans to embed processors in DRAM

October 13, 2016 // By Peter Clarke

Email | print | Share | Share | reddit

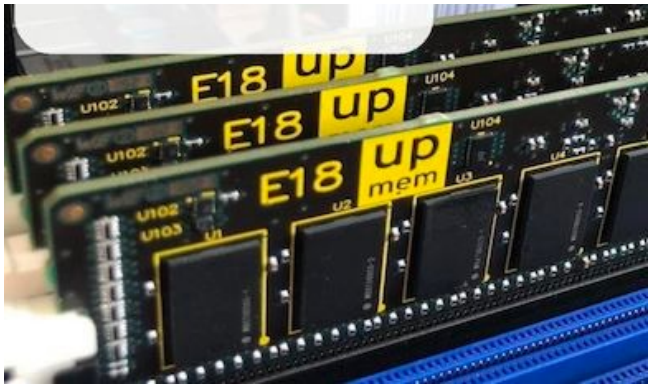
Fabless chip company Upmem SAS (Grenoble, France), founded in January 2015, is developing a microprocessor for use in data-intensive applications in the datacenter that will sit embedded in DRAM to be close to the data.

Placing hundreds or thousands of processing elements in DRAM able to perform work for a controlling server

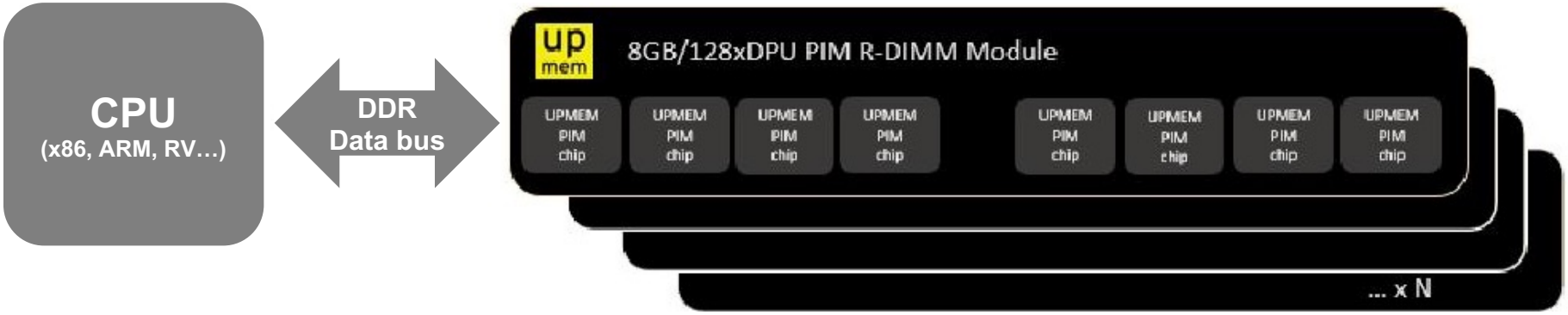
https://www.eenewsautomotive.com/news/startup-plans-embed-processors-dram-0#

UPMEM Processing-in-DRAM Engine (2019)

- Processing in DRAM Engine
- Includes **standard DIMM modules**, with a **large number of DPU processors** combined with DRAM chips.



- Replaces **standard** DIMMs
  - DDR4 R-DIMM modules
    - 8GB+128 DPUs (16 PIM chips)
    - Standard 2x-nm DRAM process
  - **Large amounts of** compute & memory bandwidth



https://www.anandtech.com/show/14750/hot-chips-31-analysis-inmemory-processing-by-upmem
https://www.upmem.com/video-upmem-presenting-its-true-processing-in-memory-solution-hot-chips-2019/

Samsung Function-in-Memory DRAM (2021)

Samsung Develops Industry's First High Bandwidth Memory with AI Processing Power

Korea on February 17, 2021 Audio Share

The new architecture will deliver over twice the system performance and reduce energy consumption by more than 70%

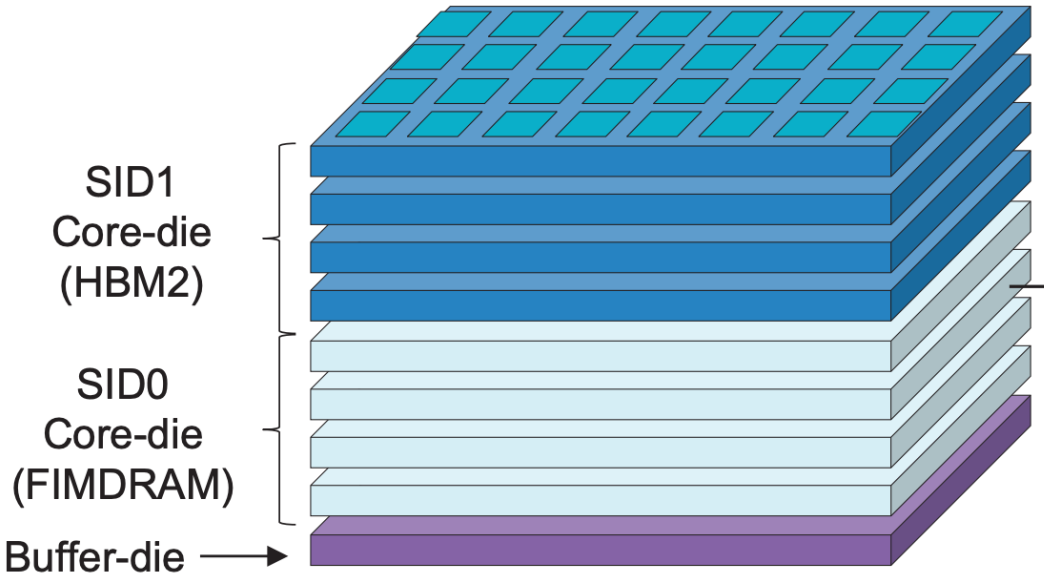
Samsung Electronics, the world leader in advanced memory technology, today announced that it has developed the industry's first High Bandwidth Memory (HBM) integrated with artificial intelligence (AI) processing power — the HBM-PIM. The new processing-in-memory (PIM) architecture brings powerful AI computing capabilities inside high-performance memory, to accelerate large-scale processing in data centers, high performance computing (HPC) systems and AI-enabled mobile applications.

Kwangil Park, senior vice president of Memory Product Planning at Samsung Electronics stated, "Our groundbreaking HBM-PIM is the industry's first programmable PIM solution tailored for diverse AI-driven workloads such as HPC, training and inference. We plan to build upon this breakthrough by further collaborating with AI solution providers for even more advanced PIM-powered applications."

https://news.samsung.com/global/samsung-develops-industrys-first-high-bandwidth-memory-with-ai-processing-power

FIMDRAM: Chip Structure

- FIMDRAM based on HBM2

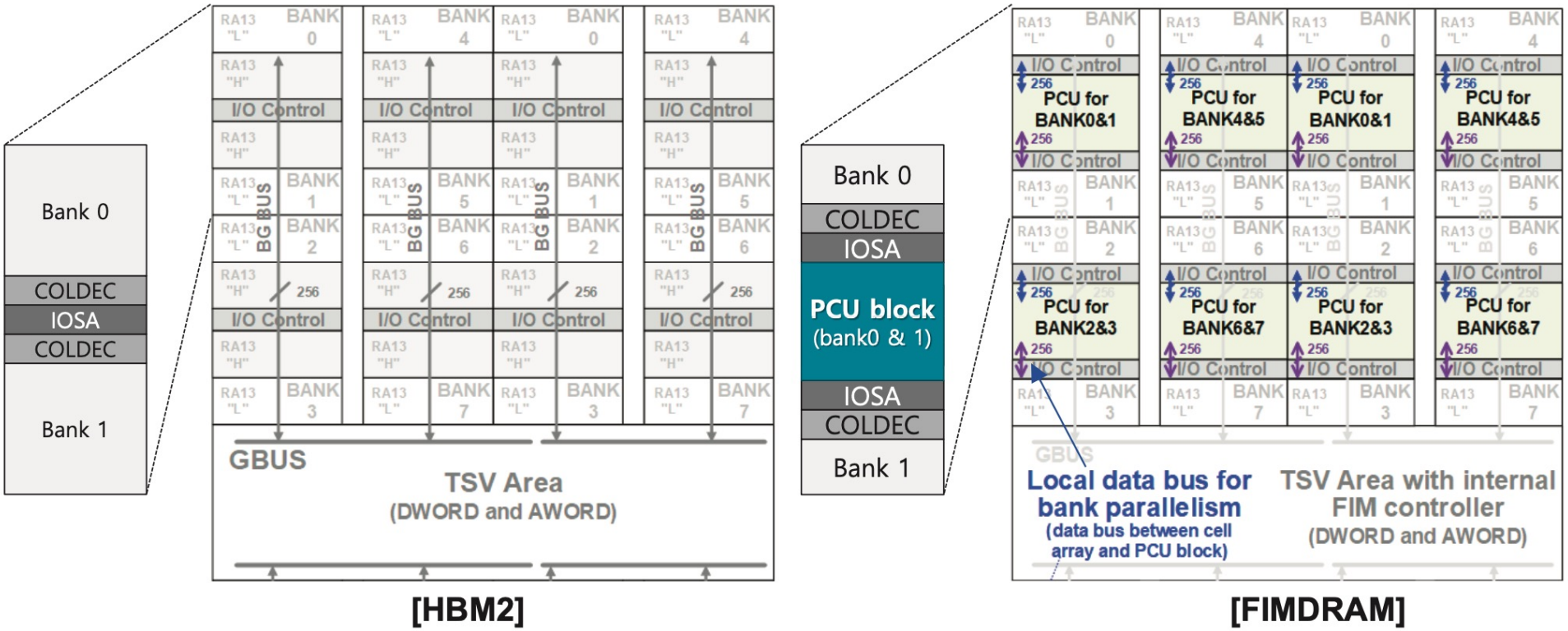


[3D Chip Structure of HBM with FIMDRAM]

Chip Specification
128DQ / 8CH / 16 banks / BL4
32 PCU blocks (1 FIM block/2 banks)
1.2 TFLOPS (4H)
FP16 ADD / Multiply (MUL) / Multiply-Accumulate (MAC) / Multiply-and- Add (MAD)

FIMDRAM: System Organization

- HBM2 vs. FIMDRAM



Samsung AxDIMM (2021)

Samsung Newsroom

CORPORATE | PRODUCTS | INSIGHTS | PRESS RESOURCES

Samsung Brings In-Memory Processing Power to Wider Range of Applications

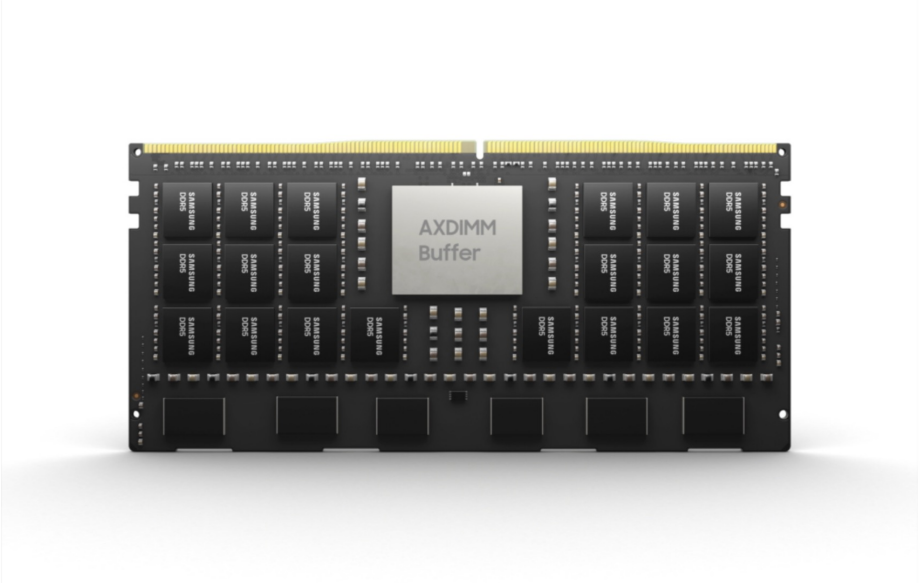
Korea on August 24, 2021 Audio Share

Integration of HBM-PIM with the Xilinx Alveo AI accelerator system will boost overall system performance by 2.5X while reducing energy consumption by more than 60%

PIM architecture will be broadly deployed beyond HBM, to include mainstream DRAM modules and mobile memory

Samsung Electronics, the world leader in advanced memory technology, today showcased its latest advancements with processing-in-memory (PIM) technology at Hot Chips 33—a leading semiconductor conference where the most notable microprocessor and IC innovations are unveiled each year. Samsung's revelations include the first successful integration of its PIM-enabled High Bandwidth Memory (HBM-PIM) into a commercialized accelerator system, and broadened PIM applications to embrace DRAM modules and mobile memory, in accelerating the move toward the convergence of memory and logic.

DRAM Modules Powered by PIM

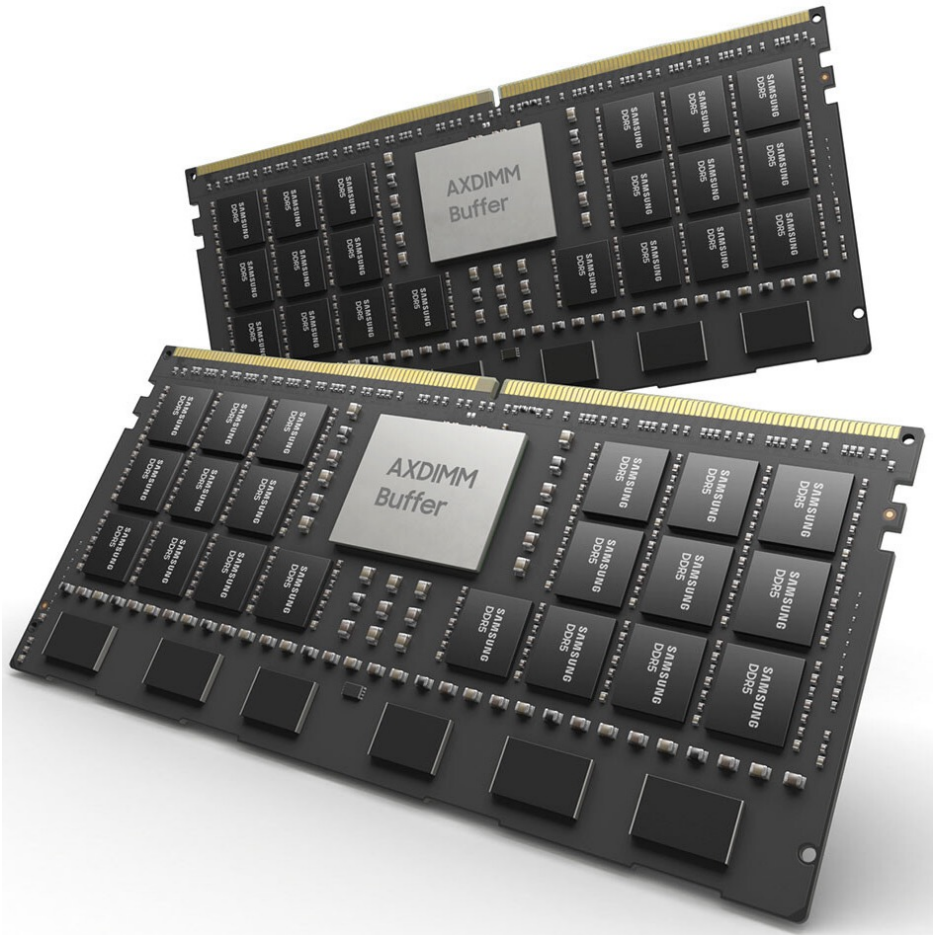


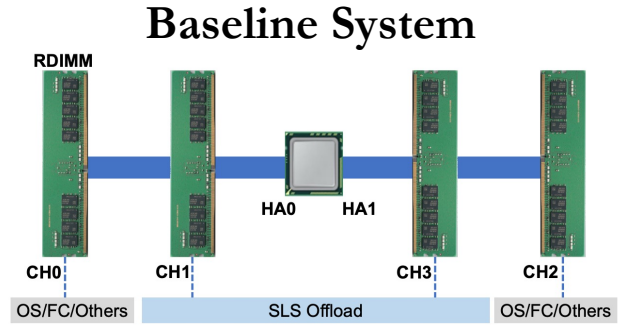
The Acceleration DIMM (AXDIMM) brings processing to the DRAM module itself, minimizing large data movement between the CPU and DRAM to boost the energy efficiency of AI accelerator systems. With an AI engine built inside the buffer chip, the AXDIMM can perform parallel processing of multiple memory ranks (sets of DRAM chips) instead of accessing just one rank at a time, greatly enhancing system performance and efficiency. Since the module can retain its traditional DIMM form factor, the AXDIMM facilitates drop-in replacement without requiring system modifications. Currently being tested on customer servers, the AXDIMM can offer approximately twice the performance in AI-based recommendation applications and a 40% decrease in system-wide energy usage.

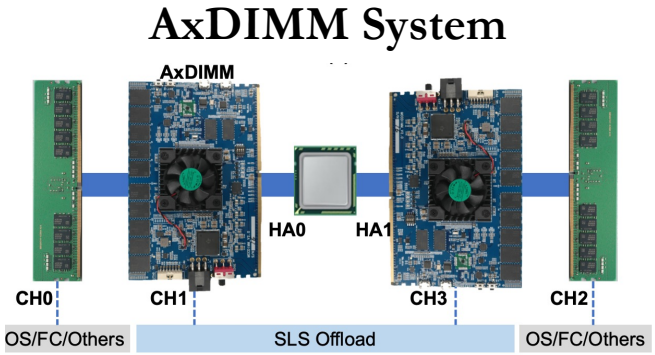
https://news.samsung.com/global/samsung-brings-in-memory-processing-power-to-wider-range-of-applications

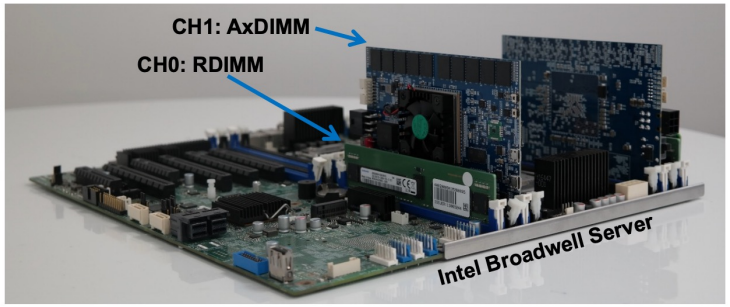
Samsung AxDIMM (2021)

- DIMM-based PIM
  - DLRM recommendation system









Ke et al. "Near-Memory Processing in Action: Accelerating Personalized Recommendation with AxDIMM", IEEE Micro (2021)

SK Hynix Accelerator-in-Memory (2022)

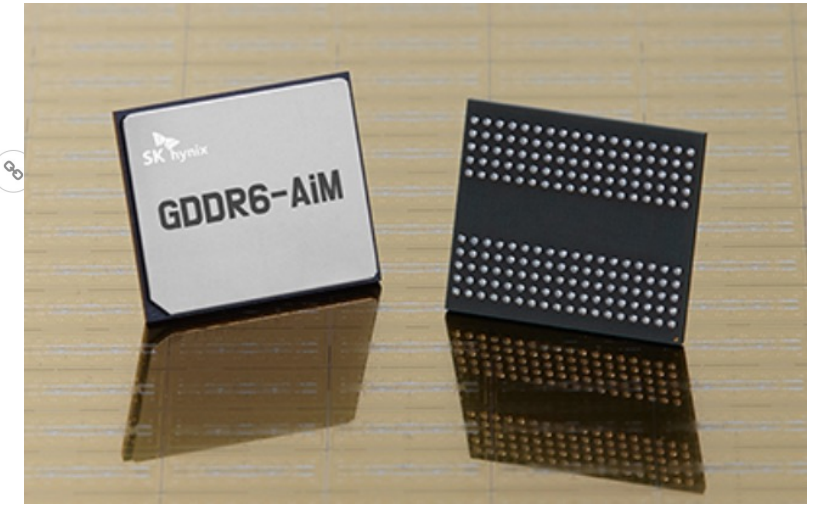
SKhynix NEWSROOM

ENG

INSIGHT | SK hynix STORY | PRESS CENTER | MULTIMEDIA | Search

SK hynix Develops PIM, Next-Generation AI Accelerator

February 16, 2022



Seoul, February 16, 2022

SK hynix (or "the Company", www.skhynix.com) announced on February 16 that it has developed PIM*, a next-generation memory chip with computing capabilities.

*PIM(Processing In Memory): A next-generation technology that provides a solution for data congestion issues for AI and big data by adding computational functions to semiconductor memory

It has been generally accepted that memory chips store data and CPU or GPU, like human brain, process data. SK hynix, following its challenge to such notion and efforts to pursue innovation in the next-generation smart memory, has found a breakthrough solution with the development of the latest technology.

SK hynix plans to showcase its PIM development at the world's most prestigious semiconductor conference, 2022 ISSCC*, in San Francisco at the end of this month. The company expects continued efforts for innovation of this technology to bring the memory-centric computing, in which semiconductor memory plays a central role, a step closer to the reality in devices such as smartphones.

*ISSCC: The International Solid-State Circuits Conference will be held virtually from Feb. 20 to Feb. 24 this year with a theme of "Intelligent Silicon for a Sustainable World"

For the first product that adopts the PIM technology, SK hynix has developed a sample of GDDR6-AiM (Accelerator* in memory). The GDDR6-AiM adds computational functions to GDDR6* memory chips, which process data at 16Gbps. A combination of GDDR6-AiM with CPU or GPU instead of a typical DRAM makes certain computation speed 16 times faster. GDDR6-AiM is widely expected to be adopted for machine learning, high-performance computing, and big data computation and storage.

11.1 A 1ynm 1.25V 8Gb, 16Gb/s/pin GDDR6-based Accelerator-in-Memory supporting 1TFLOPS MAC Operation and Various Activation Functions for Deep-Learning Applications

Seongju Lee, SK hynix, Icheon, Korea

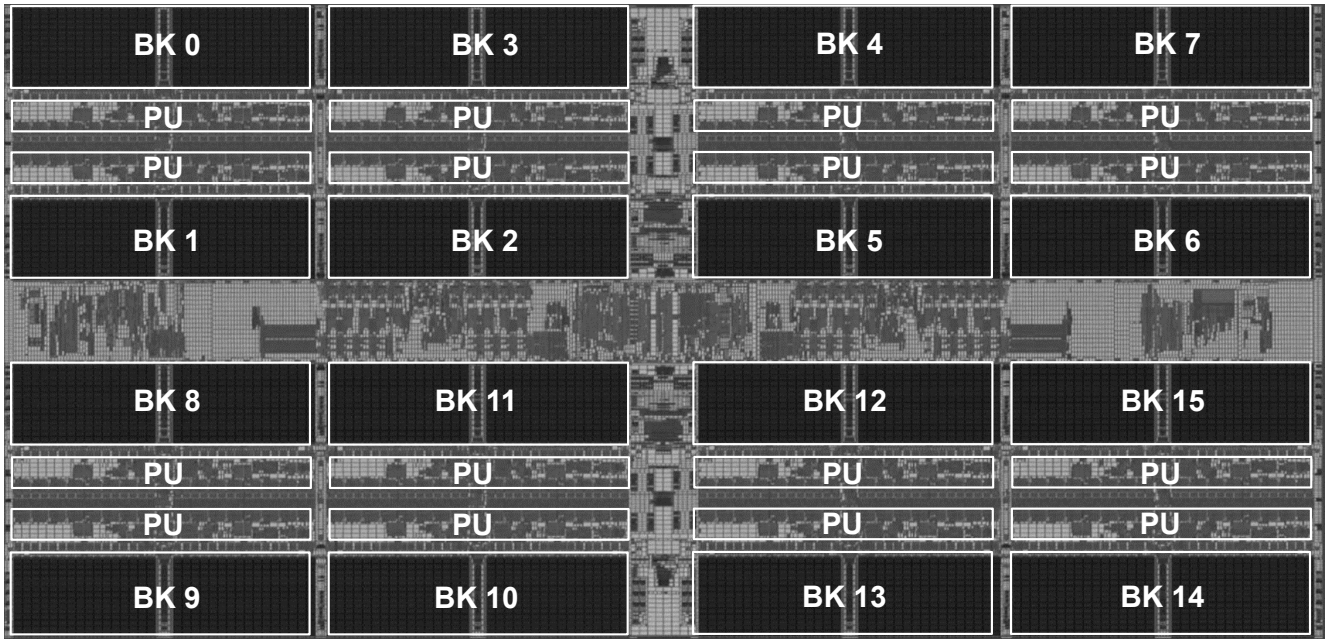
In Paper 11.1, SK Hynix describes an 1ynm, GDDR6-based accelerator-in-memory with a command set for deep-learning operation. The 8Gb design achieves a peak throughput of 1TFLOPS with 1GHz MAC operations and supports major activation functions to improve accuracy.

https://news.skhynix.com/sk-hynix-develops-pim-next-generation-ai-accelerator/

AiM: Chip Implementation

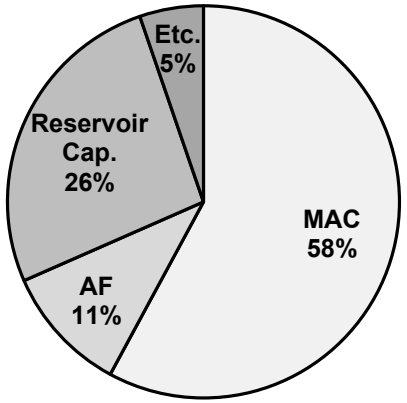
- 4 Gb AiM die with 16 processing units (PUs)

AiM Die Photograph



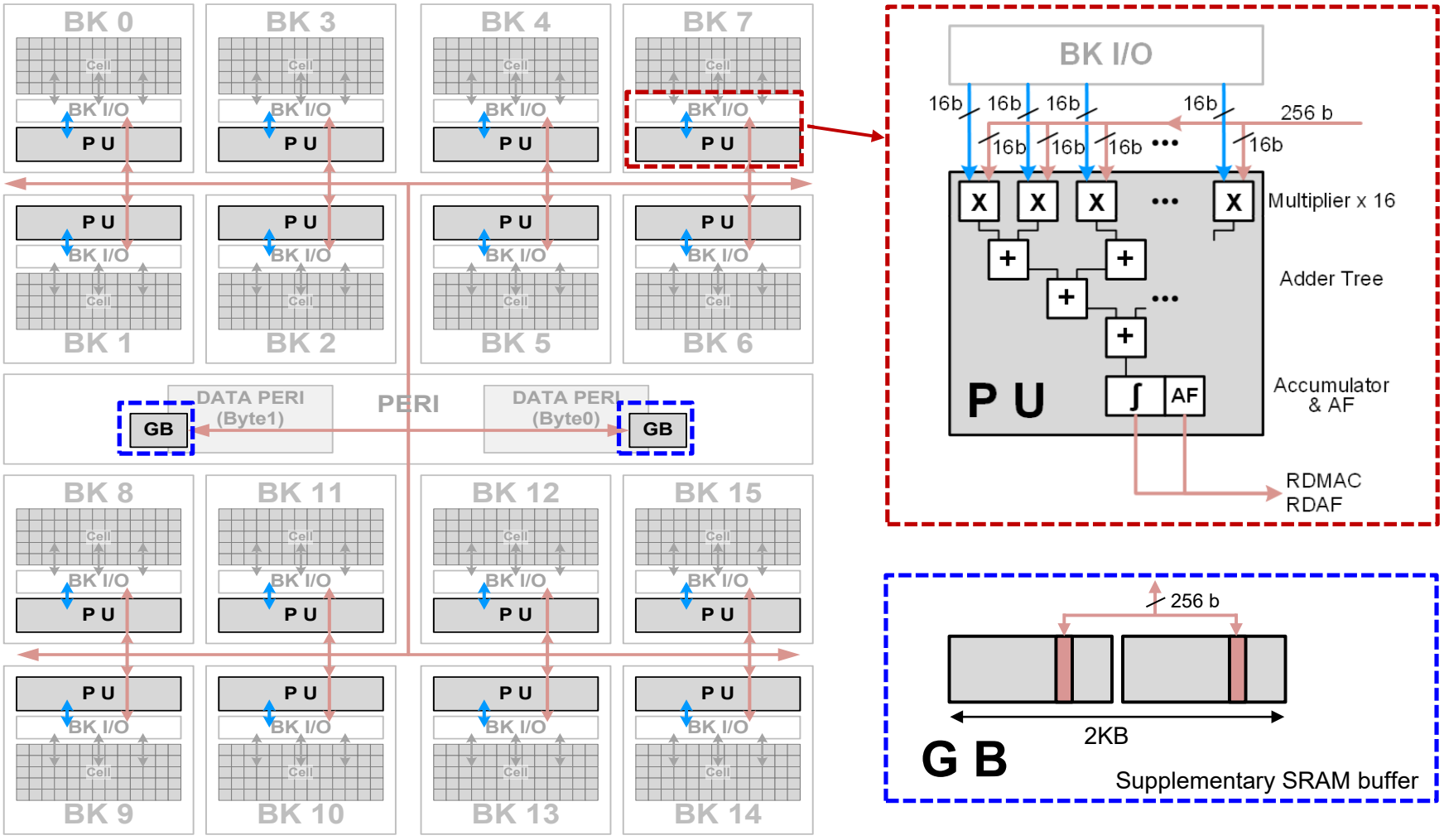
1 Process Unit (PU) Area

Total	0.19mm^2
MAC	0.11mm^2
Activation Function (AF)	0.02mm^2
Reservoir Cap.	0.05mm^2
Etc.	0.01mm^2



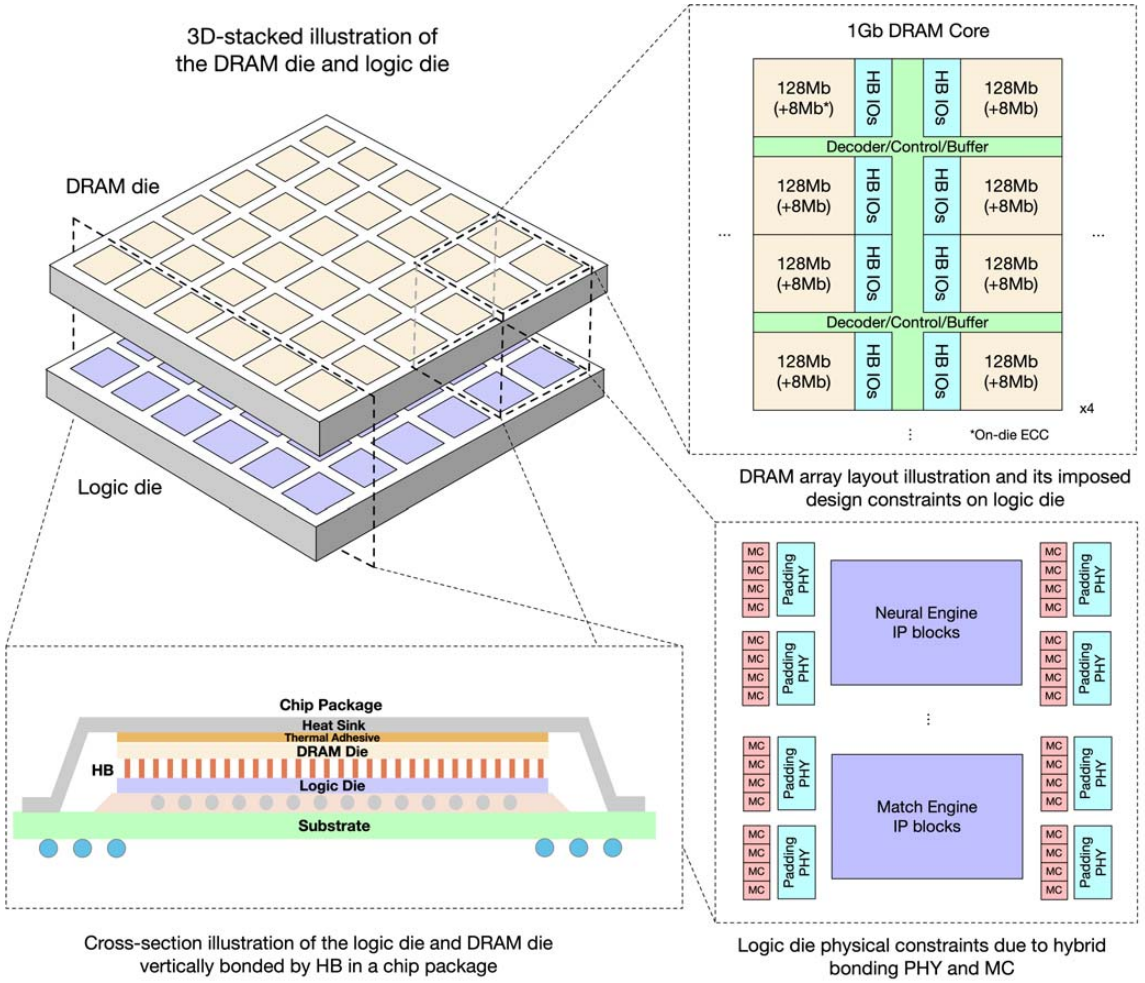
AiM: System Organization

- GDDR6-based AiM architecture



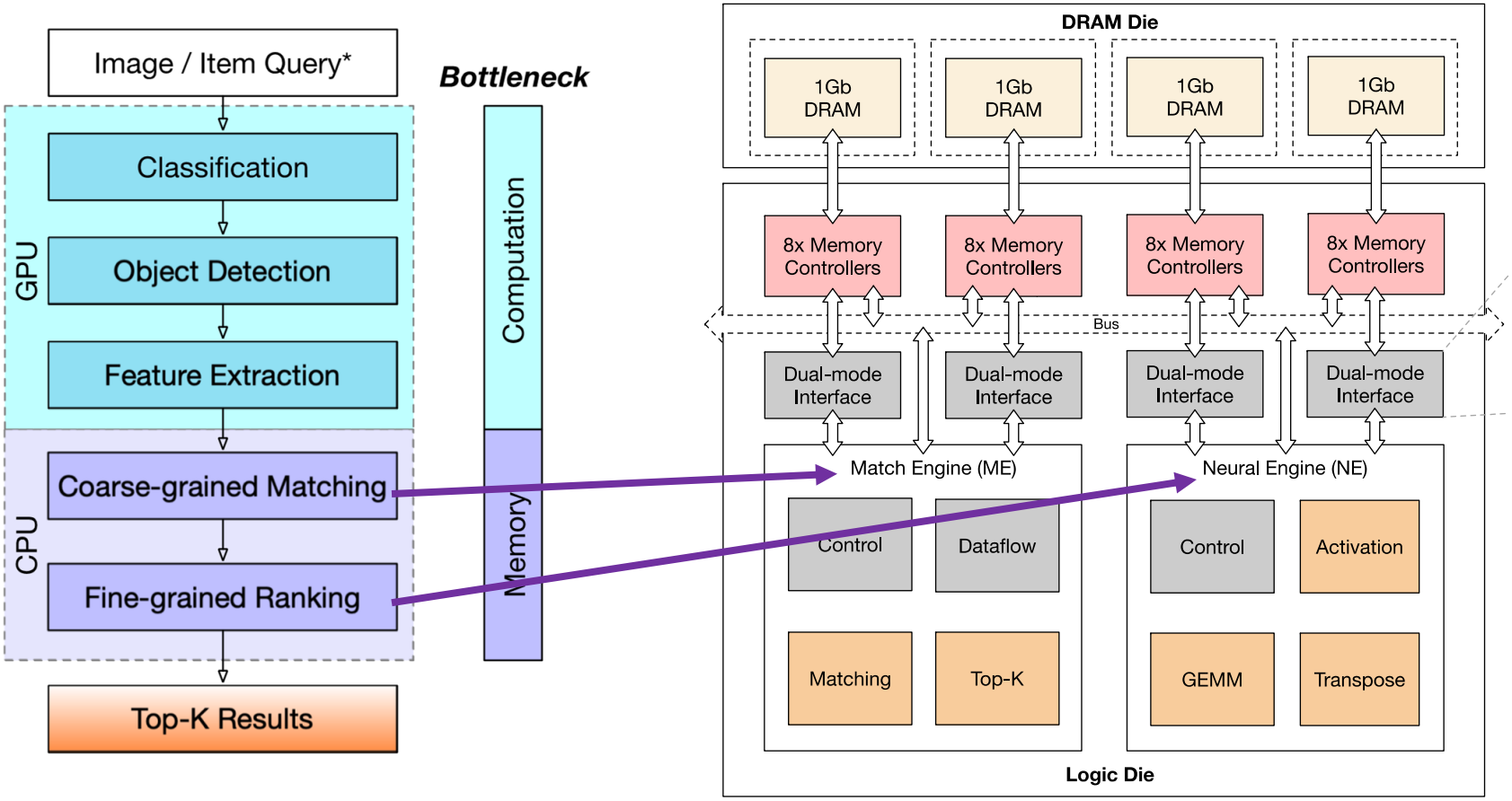
HB-PNM: Overall Architecture (I)

- 3D-stacked logic die and DRAM die vertically bonded by hybrid bonding (HB)



HB-PNM: Overall Architecture (II)

- Match engine and neural engine for matching and ranking in a recommendation system



Two PIM Approaches

5.2. Two Approaches: Processing Using Memory (PUM) vs. Processing Near Memory (PNM)

Many recent works take advantage of the memory technology innovations that we discuss in Section 5.1 to enable and implement PIM. We find that these works generally take one of two approaches, which are categorized in Table 1: (1) *processing using memory* or (2) *processing near memory*. We briefly describe each approach here. Sections 6 and 7 will provide example approaches and more detail for both.

Table 1: Summary of enabling technologies for the two approaches to PIM used by recent works. Adapted from [309].

Approach	Enabling Technologies
Processing Using Memory	SRAM DRAM Phase-change memory (PCM) Magnetic RAM (MRAM) Resistive RAM (RRAM)/memristors
Processing Near Memory	Logic layers in 3D-stacked memory Silicon interposers Logic in memory controllers

Processing using memory (PUM) exploits the existing memory architecture and the operational principles of the memory circuitry to enable operations within main memory with minimal changes. PUM makes use

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.
[Tutorial Video on "Memory-Centric Computing Systems" (1 hour 51 minutes)]

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

Barriers to Adoption of PIM

1. Functionality of and applications & software for PIM

2. Ease of programming (interfaces and compiler/HW support)

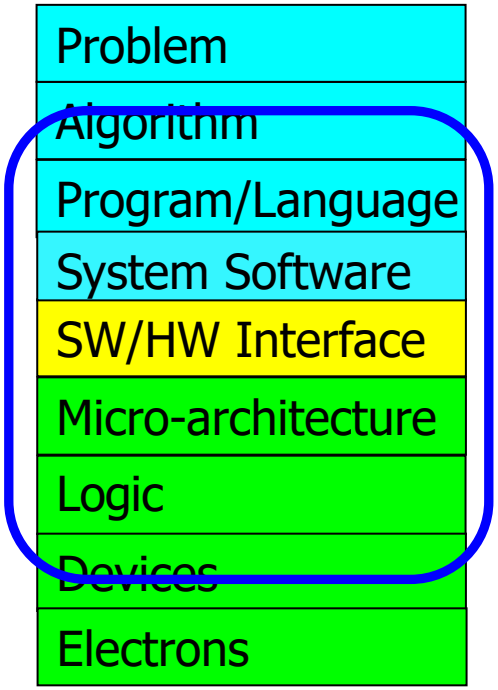
3. System support: coherence & virtual memory

4. Runtime and compilation systems for adaptive scheduling, data mapping, access/sharing control

5. Infrastructures to assess benefits and feasibility

All can be solved with change of mindset

We Need to Revisit the Entire Stack



We can get there step by step

Agenda

- Plenary talk I: In-Memory Processing
  - Professor Onur Mutlu (ETH Zürich, Switzerland)
- 16:00-17:15 - In-Memory Processing I
  - The Road to Widely Deploying Processing-In-Memory: Challenges and Opportunities
    - Saugata Ghose (University of Illinois Urbana-Champaign, USA)
  - Methodologies, Workloads, and Tools for Processing-In-Memory: Enabling the Adoption of Data-Centric Architectures
    - Geraldo Francisco De Oliveira Junior (ETH Zurich, Switzerland); Juan Gomez Luna (ETH, Switzerland); Saugata Ghose (University of Illinois Urbana-Champaign, USA); Onur Mutlu (ETH Zurich, Switzerland)
  - PiDRAM: An FPGA-Based Framework for End-To-End Evaluation of Processing-In-DRAM Techniques
    - Ataberk Olgun (ETH Zurich, Switzerland); Juan Gomez Luna (ETH, Switzerland); Konstantinos Kanellopoulos and Behzad Salami (ETH Zurich, Switzerland); Hasan Hassan (ETH Zurich); Oguz Ergin (TOBB University of Economics and Technology, Turkey); Onur Mutlu (ETH Zurich, Switzerland)
- 17:30-19:00 - In-Memory Processing II
  - Heterogeneous Data-Centric Architectures for Modern Data-Intensive Applications: Case Studies in Machine Learning and Databases
    - Geraldo Francisco De Oliveira Junior (ETH Zurich, Switzerland); Saugata Ghose (University of Illinois Urbana-Champaign, USA); Juan Gomez Luna (ETH, Switzerland); Onur Mutlu (ETH Zurich, Switzerland)
  - Exploiting Near-Data Processing to Accelerate Time Series Analysis
    - Ivan Fernandez (University of Malaga & ETH Zurich, Spain); Ricardo Quislant (University of Malaga, Spain); Christina Giannoula (National Technical University of Athens & ETH Zurich, Greece); Mohammed Alser and Juan Gomez Luna (ETH, Switzerland); Eladio D Gutierrez and Oscar Plata (University of Malaga, Spain); Onur Mutlu (ETH Zurich, Switzerland)
  - GenStore: In-Storage Filtering of Genomic Data for High-Performance and Energy-Efficient Genome Analysis
    - Nika Mansouri Ghiasi, Jisung Park, Harun Mustafa, Jeremie Kim, Ataberk Olgun and Arvid Gollwitzer (ETH Zurich, Switzerland); Damla Senol Cali (Bionano Genomics, USA); Can Firtina, Haiyu Mao and Nour Almadhoun Alserr (ETH Zurich, Switzerland); Rachata Ausavarungnirun (KMUTNB, Thailand); Nandita Vijaykumar (University of Toronto, Canada); Mohammed Alser and Onur Mutlu (ETH Zurich, Switzerland)
  - SparseP: Efficient Sparse Matrix Vector Multiplication on Real Processing-In-Memory Architectures
    - Christina Giannoula (National Technical University of Athens & ETH Zurich, Greece); Ivan Fernandez (University of Malaga & ETH Zurich, Spain); Juan Gomez-Luna (ETH, Switzerland); Nectarios Koziris and Georgios Goumas (National Technical University of Athens, Greece); Onur Mutlu (ETH Zurich, Switzerland)
  - Machine Learning Training on a Real Processing-In-Memory System
    - Juan Gomez Luna and Yuxin Guo (ETH, Switzerland); Sylvan Brocard, Julien Legriel and Remy Cimadomo (UPMEM, France); Geraldo Oliveira and Gagandeep Singh (ETH, Switzerland); Onur Mutlu (ETH Zurich, Switzerland)

In-Memory Processing

ISVLSI 2022 Special Session

IEEE Computer Society Annual Symposium on VLSI

Adonis room
Ailathon resort, Paphos, Cyprus
July 4th, 2022