P&S Processing-in-Memory

Real-World Processing-in-Memory Architectures: Samsung AxDIMM

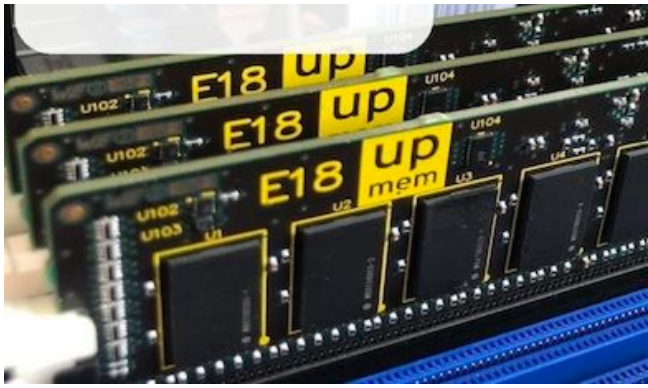
Dr. Juan Gómez Luna
Prof. Onur Mutlu
ETH Zürich

Fall 2022

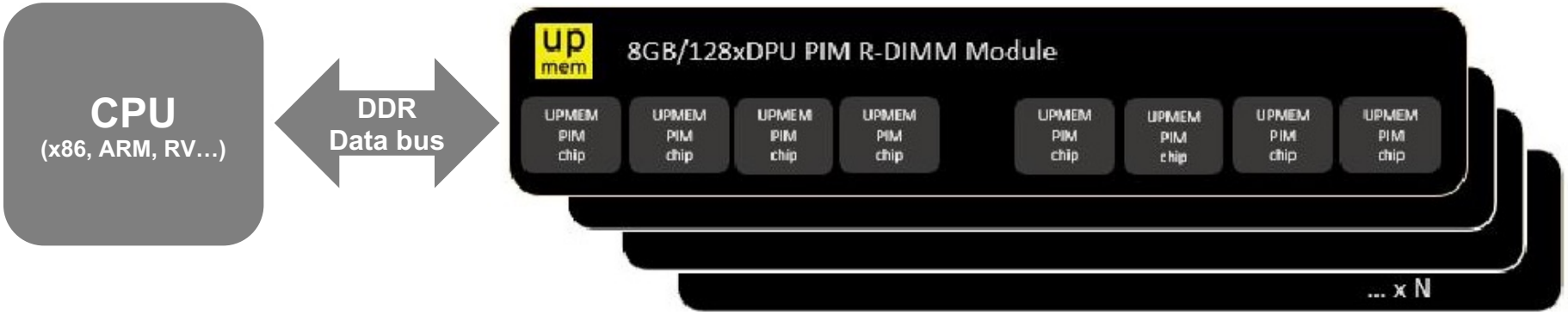
22 November 2022

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.



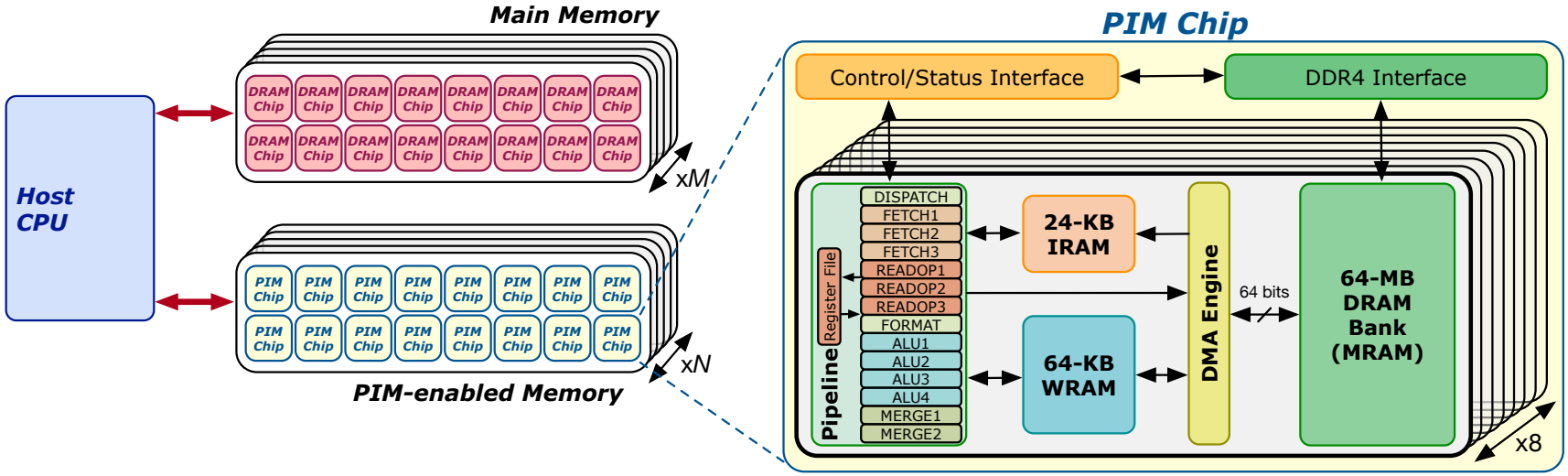
- Replaces **standard** DIMMs
  - DDR4 R-DIMM modules
    - 8GB+128 DPUs (16 PIM chips)
    - Standard 2x-nm DRAM process
  - **Large amounts of** compute & memory bandwidth



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/

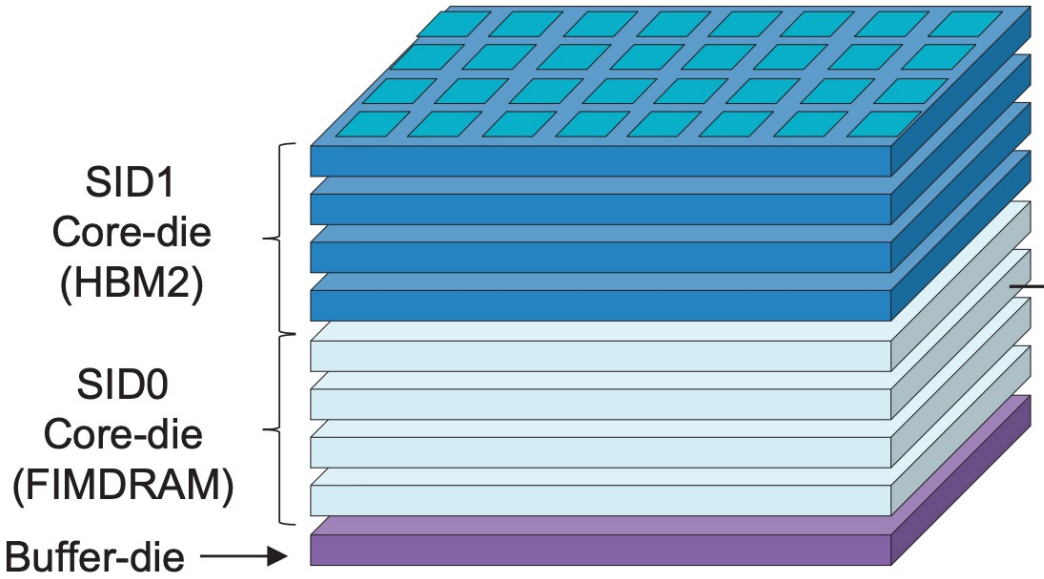
UPMEM PIM System Organization

- A UPMEM DIMM contains 8 or 16 chips
  - Thus, 1 or 2 ranks of 8 chips each
- Inside each PIM chip there are:
  - 8 64MB banks per chip: Main RAM (MRAM) banks
  - 8 DRAM Processing Units (DPUs) in each chip, 64 DPUs per rank



FIMDRAM: Chip Structure

- FIMDRAM based on HBM2

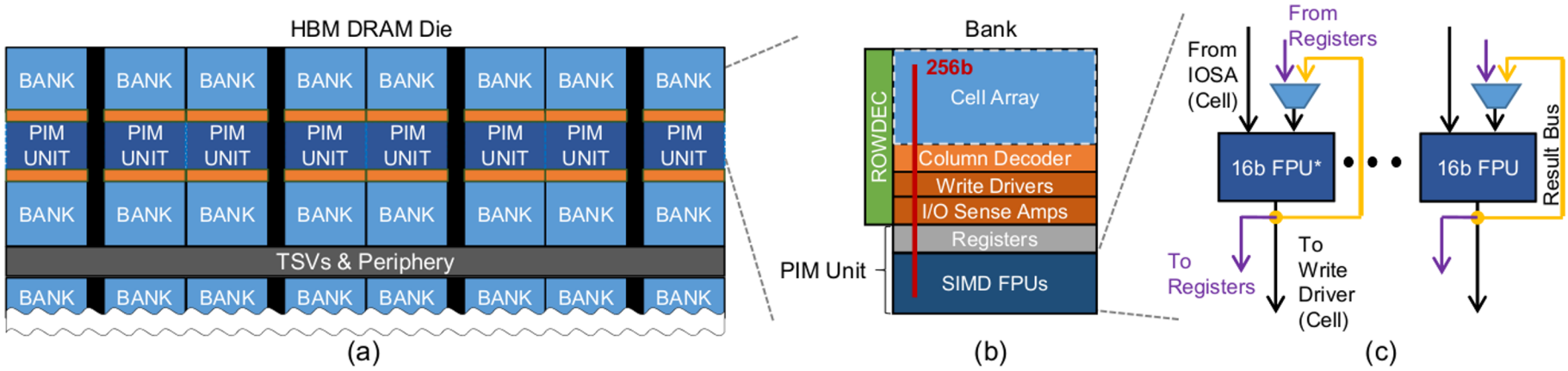


[3D Chip Structure of HBM with FIMDRAM]

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 (III)

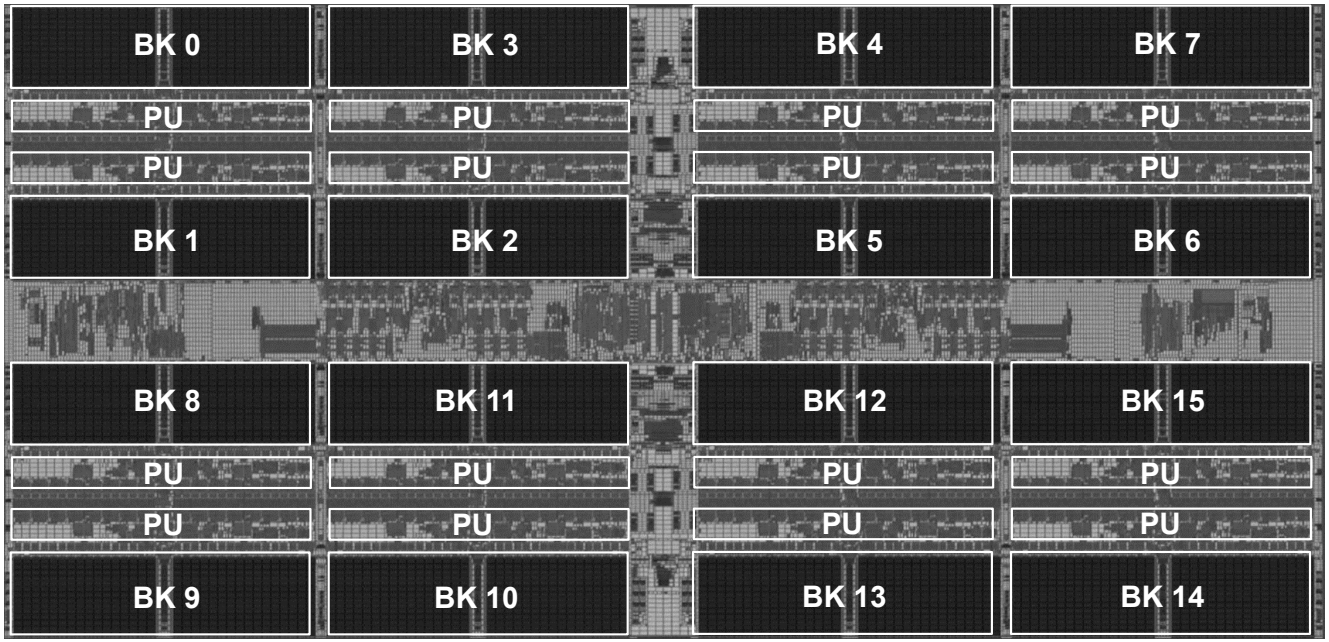
- PIM units respond to standard DRAM column commands (RD or WR)
  - Compliant with unmodified JEDEC controllers
- They execute one wide-SIMD operation commanded by a PIM instruction with deterministic latency in a lock-step manner
- A PIM unit can get 16 16-bit operands from IOSAs, a register, and/or the result bus



AiM: Chip Implementation

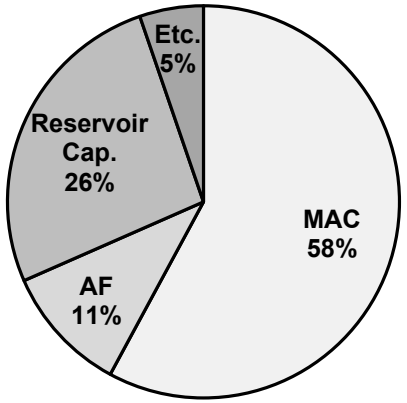
- 4 Gb AiM die with 16 processing units (PUs)

AiM Die Photograph



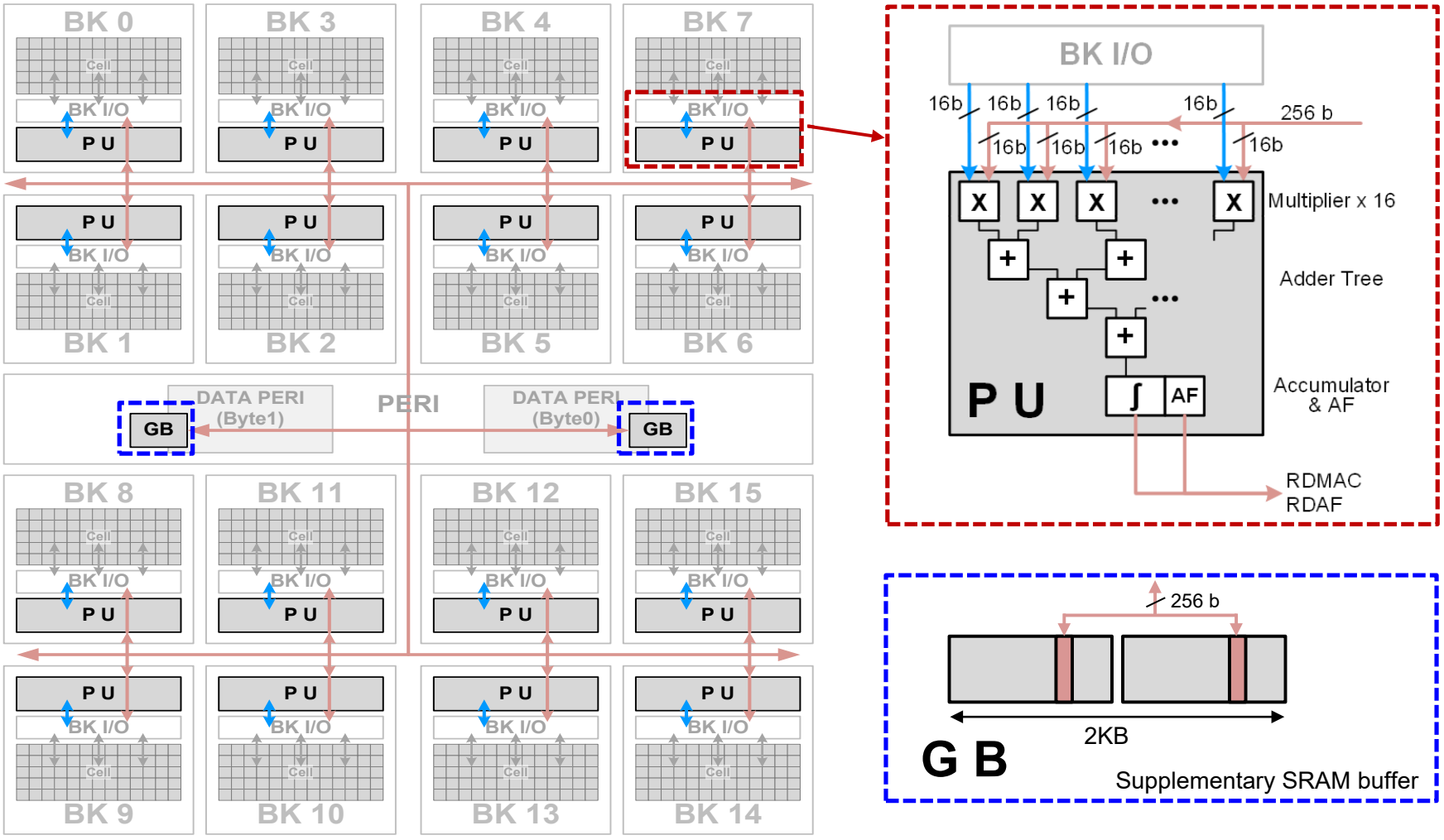
1 Process Unit (PU) Area

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



Samsung AxDIMM

Samsung AxDIMM (2021)

Samsung Newsroom

CORPORATE | PRODUCTS | INSIGHTS | PRESS RESOURCES

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

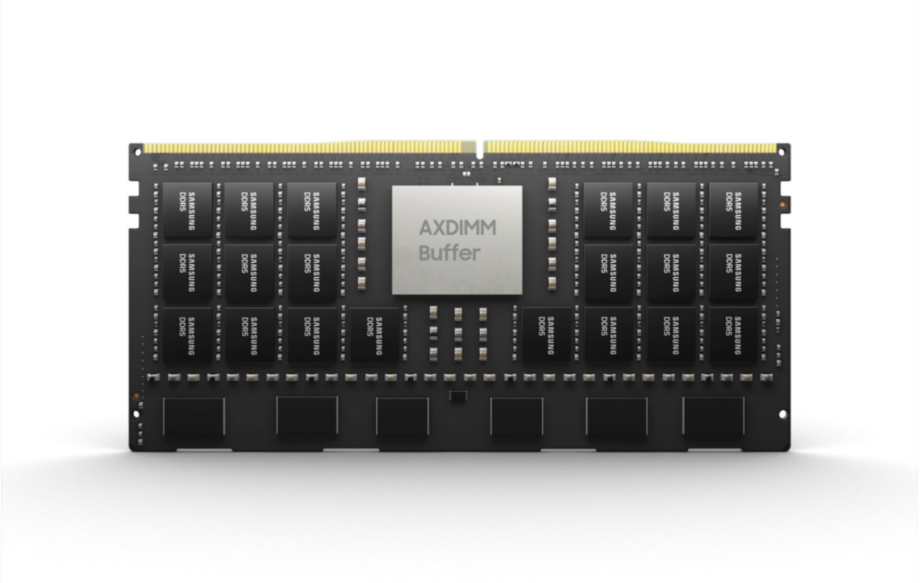
Korea on August 24, 2021 Audio Share

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

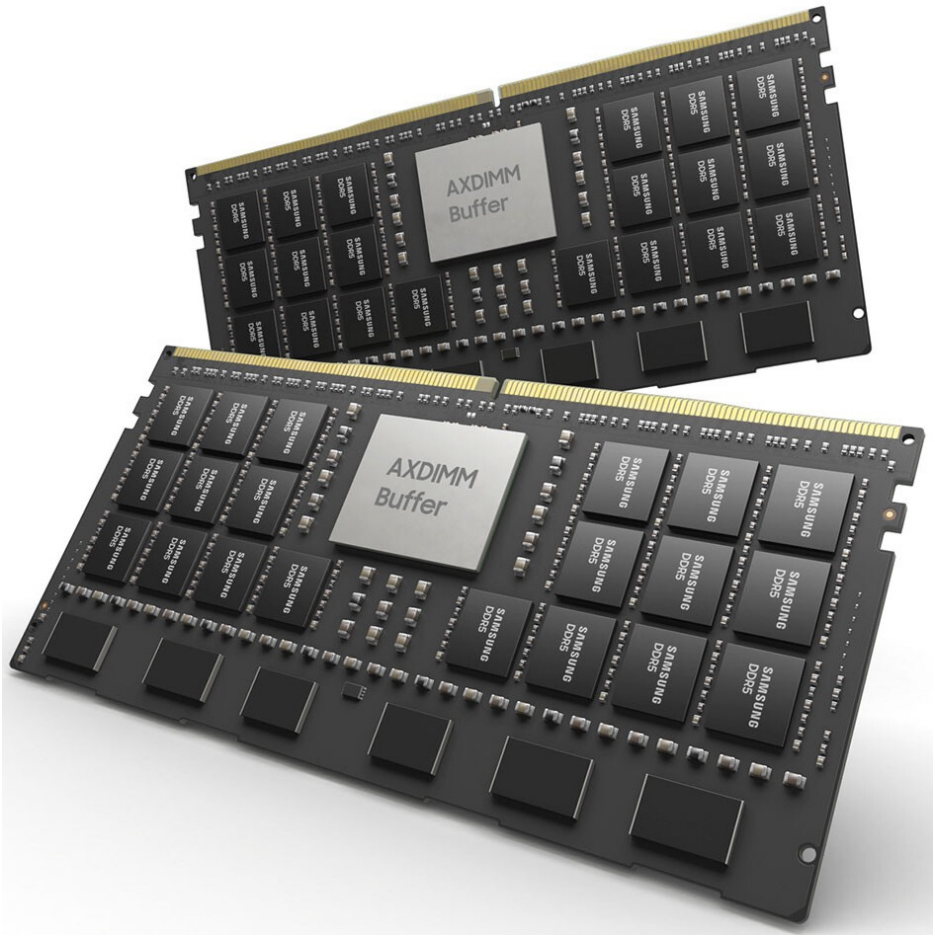


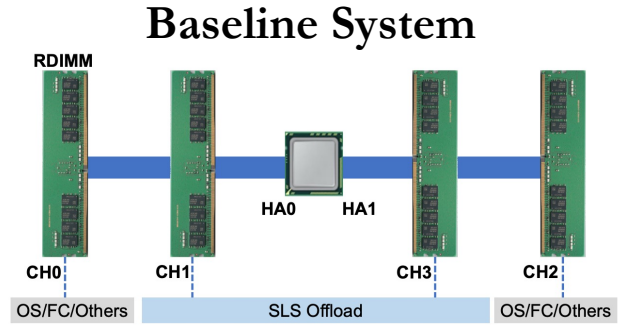
The Acceleration DIMM (AXDIMM) brings processing to the DRAM module itself, minimizing large data movement between the CPU and DRAM to boost the energy efficiency of AI accelerator systems. With an AI engine built inside the buffer chip, the AXDIMM can perform parallel processing of multiple memory ranks (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.

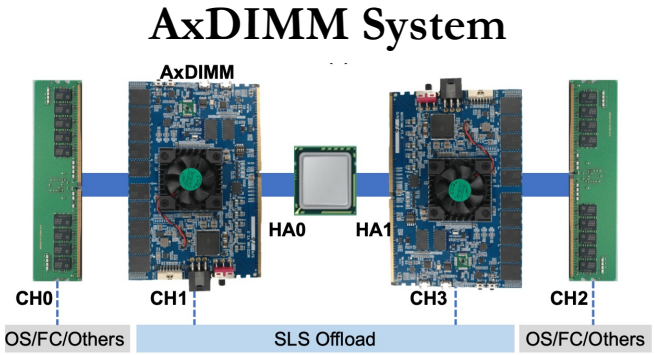
https://news.samsung.com/global/samsung-brings-in-memory-processing-power-to-wider-range-of-applications

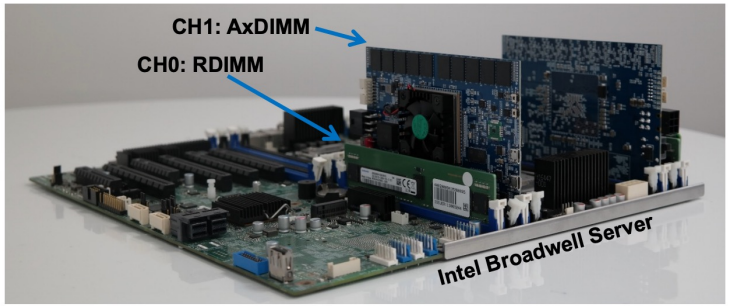
Samsung AxDIMM (2021)

- DIMM-based PIM
  - DLRM recommendation system





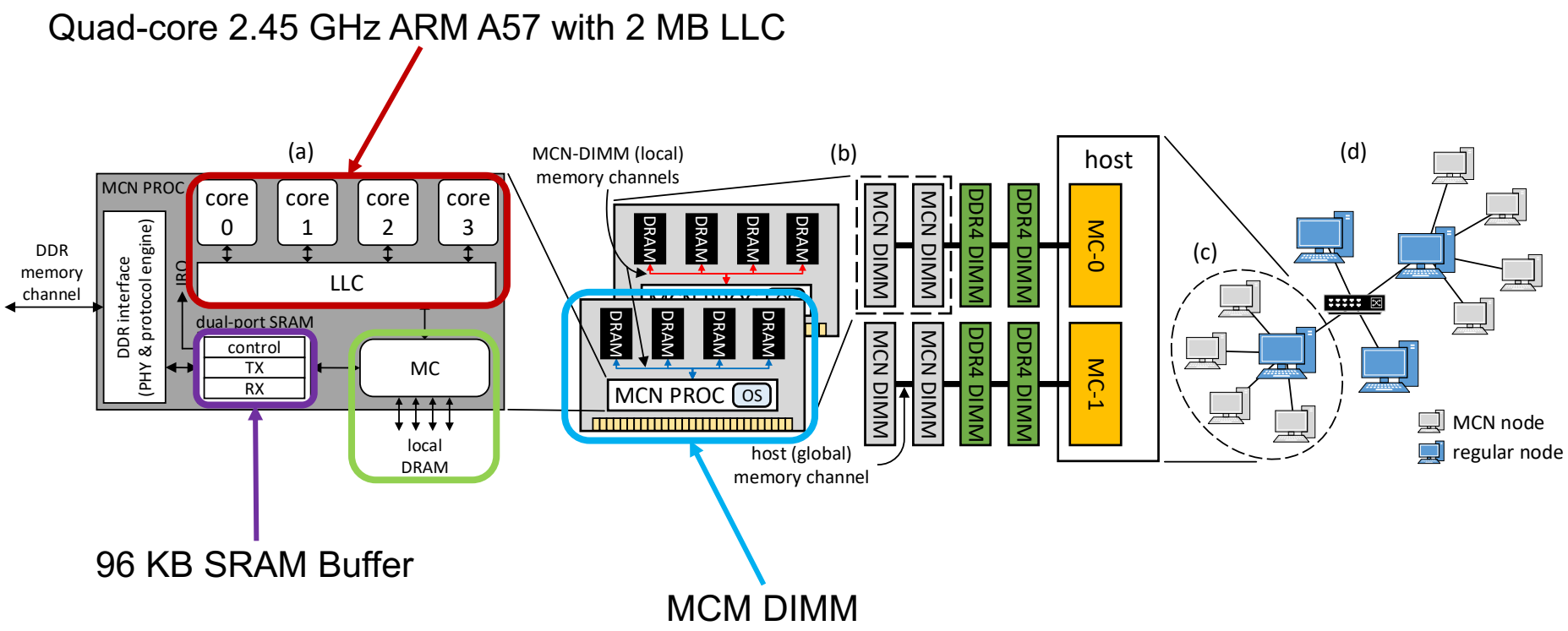




Ke et al. "Near-Memory Processing in Action: Accelerating Personalized Recommendation with AxDIMM", IEEE Micro (2021)

General-Purpose Near-Rank Approach

- Memory Channel Network (MCN) DIMMs



PnM with AxDIMM (IEEE Micro 2021)

Near-Memory Processing in Action: Accelerating Personalized Recommendation with AxDIMM

Liu Ke*†, Xuan Zhang†, Jinin So‡, Jong-Geon Lee‡, Shin-Haeng Kang‡, Sukhan Lee‡, Songyi Han‡, YeonGon Cho‡, JIN Hyun Kim‡, Yongsuk Kwon‡, KyungSoo Kim‡, Jin Jung‡, Ilkwon Yun‡, Sung Joo Park‡, Hyunsun Park‡, Joonho Song‡, Jeonghyeon Cho‡, Kyomin Sohn‡, Nam Sung Kim‡, Hsien-Hsin S. Lee*

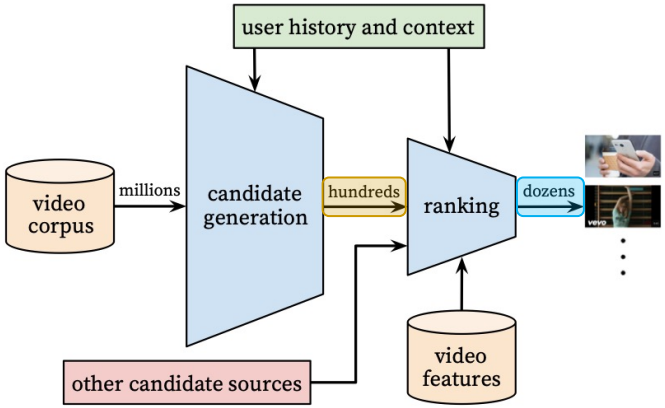
*Facebook, †Washington University in St. Louis, ‡Samsung

https://doi.org/10.1109/MM.2021.3097700

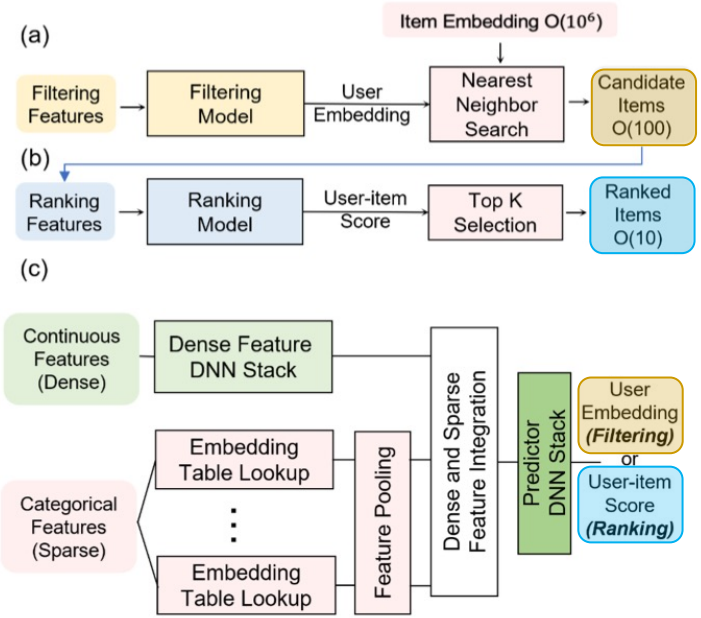
Recommendation Systems

Recommendation Systems

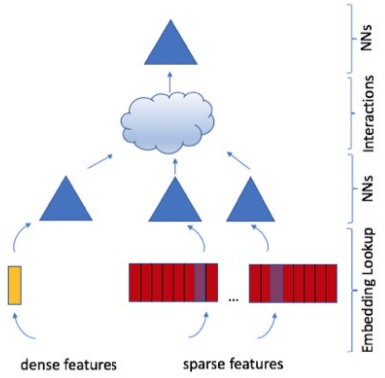
- Candidate recommendations are retrieved and then ranked



Covington et al., Deep Neural Networks for YouTube Recommendations, RecSys 2016



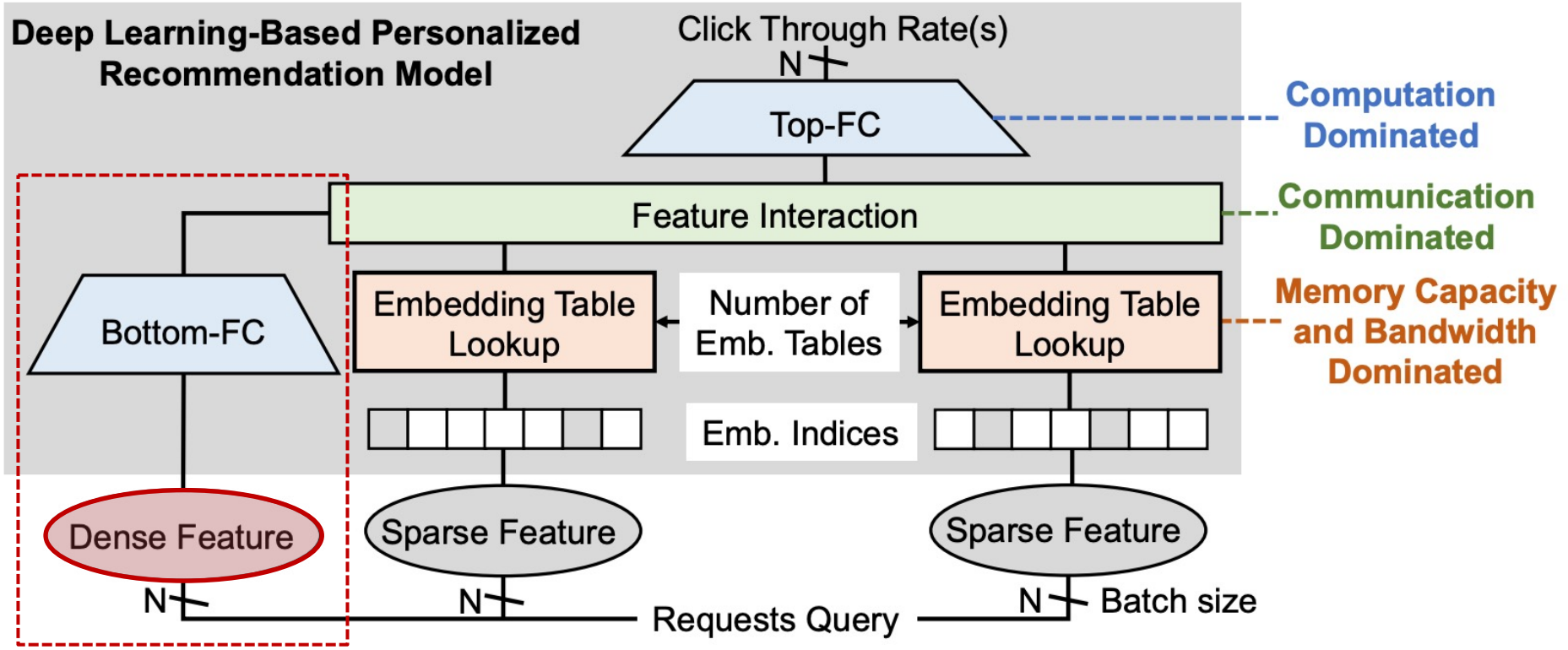
Li et al., iMARS: An In-Memory-Computing Architecture for Recommendation Systems, arXiv:2202.09433, 2022



Naumov et al., Deep Learning Recommendation Model for Personalization and Recommendation Systems, arXiv:1906.00091, 2019

Overview of Recommendation Models

- Personalized recommendation: recommend content to users, e.g., Facebook's DLRM recommendation system

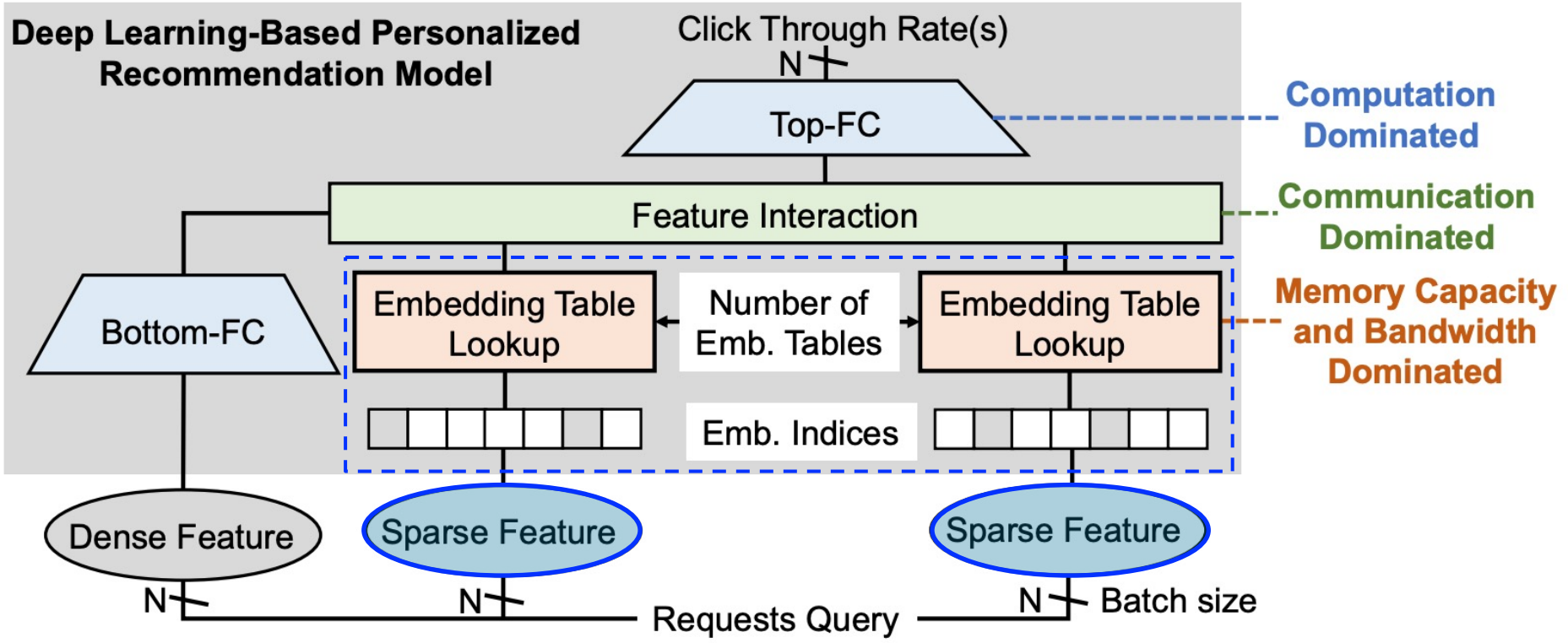


Dense features: continuous inputs in vectors and matrices are processed by typical DNN layers (e.g., fully connected layers)

Ke et al. "RecNMP: Accelerating personalized recommendation with near-memory processing," ISCA 2020

Overview of Recommendation Models

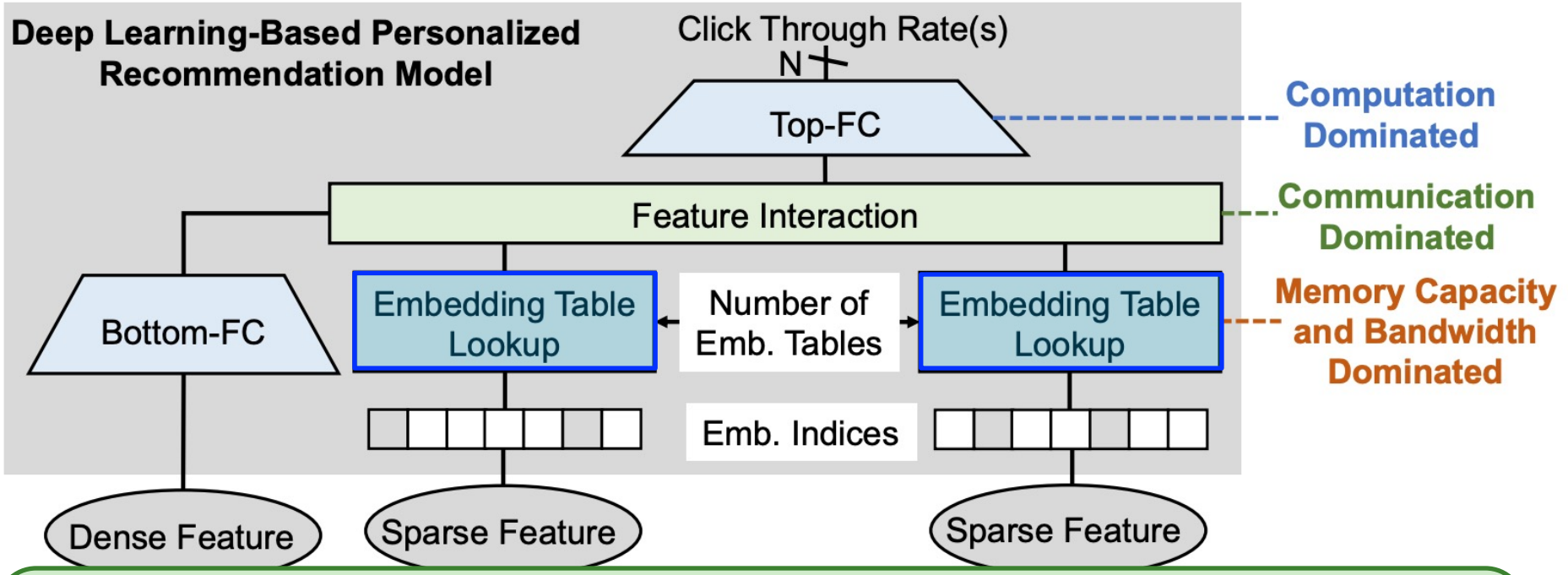
- Personalized recommendation: recommend content to users, e.g., Facebook's DLRM recommendation system



Sparse features: for categorical inputs;
processed by indexing large embedding tables

Overview of Recommendation Models

- Personalized recommendation: recommend content to users, e.g., Facebook's DLRM recommendation system

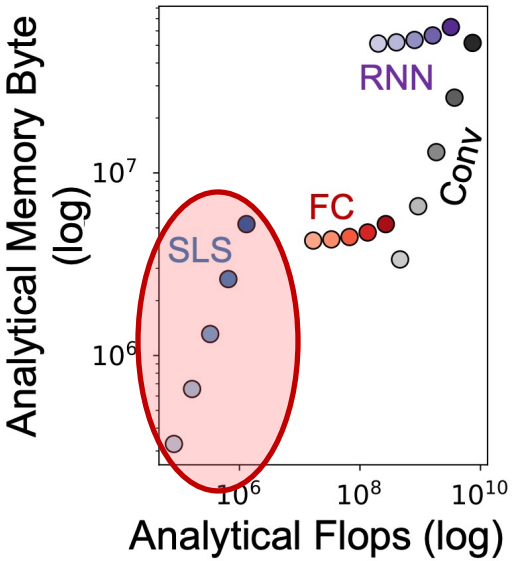


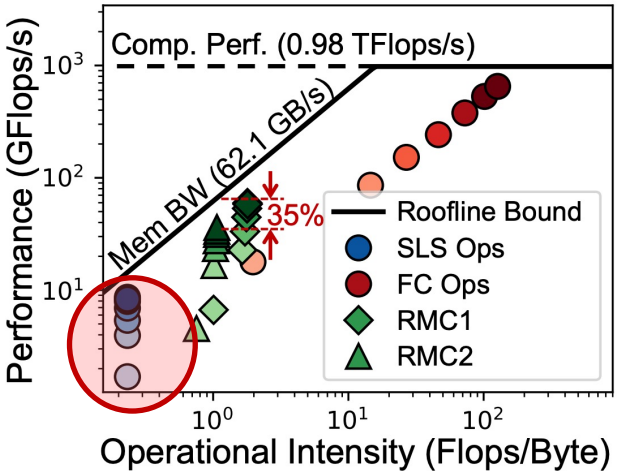
Embedding tables are organized as a set of potentially millions of vectors: lookup and pooling operations represent sparse features learned during training and generally exhibit Gather-Reduce pattern, via Caffe2's SparseLengths (SLS) operators

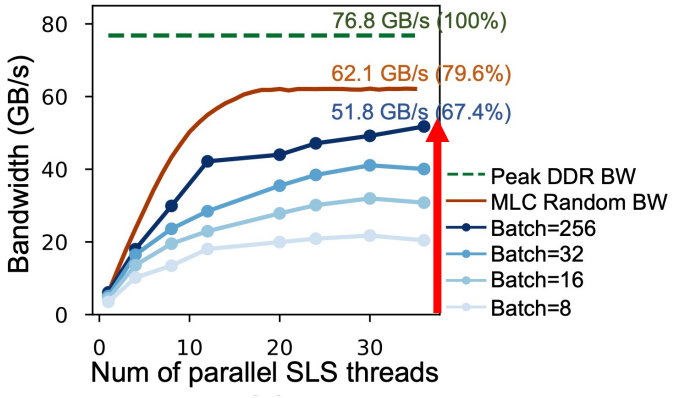
Ke et al. "RecNMP: Accelerating personalized recommendation with near-memory processing," ISCA 2020

DLRM Performance Characterization

- Identifying key performance bottlenecks for the DLRM system





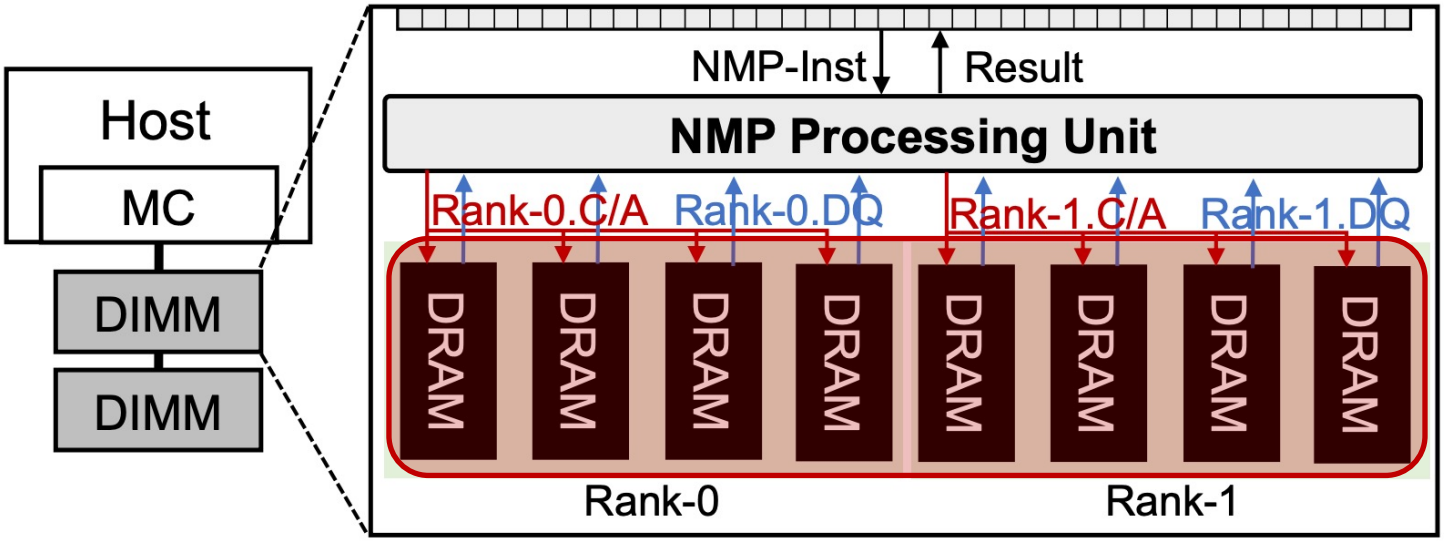


SparseLengths (SLS) operators:
- Low FP intensity
- Larger batch size:
  - Higher memory footprint
  - Higher memory intensity

The memory bandwidth can easily be saturated by embedding operations especially as both the batch size and the number of threads increase

RecNMP Architecture

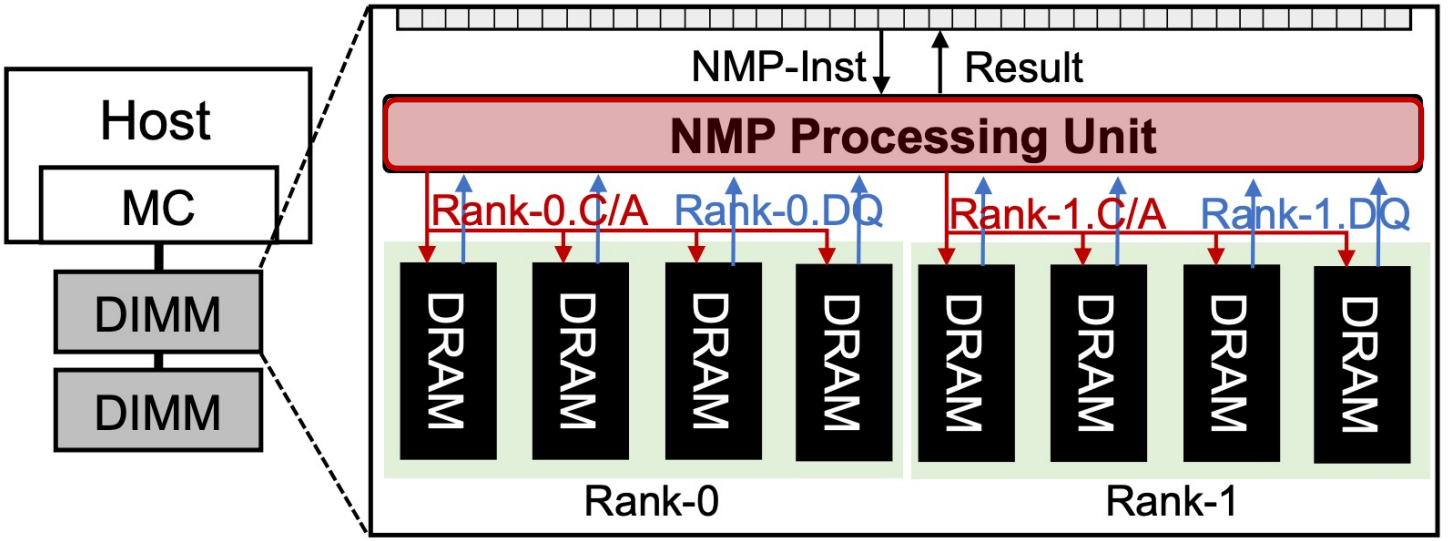
- DIMM-based NMP architecture for recommendation systems
  - Multiply the bandwidth by exploiting rank-level parallelism



Embedding entries are fetched from the concurrently activated ranks

RecNMP Architecture

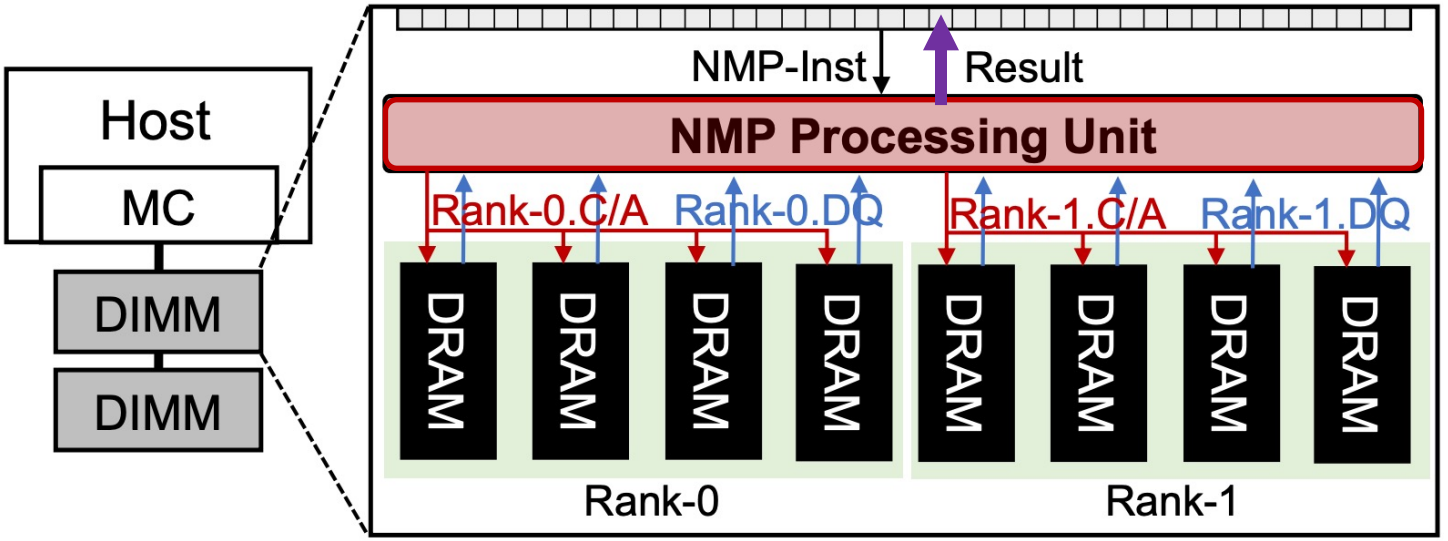
- DIMM-based NMP architecture for recommendation systems
  - Multiply the bandwidth by exploiting rank-level parallelism



The NMP PU performs the local embedding lookup and pooling functions at the memory side, producing the general Gather-Reduce execution pattern

RecNMP Architecture

- DIMM-based NMP architecture for recommendation systems
  - Multiply the bandwidth by exploiting rank-level parallelism

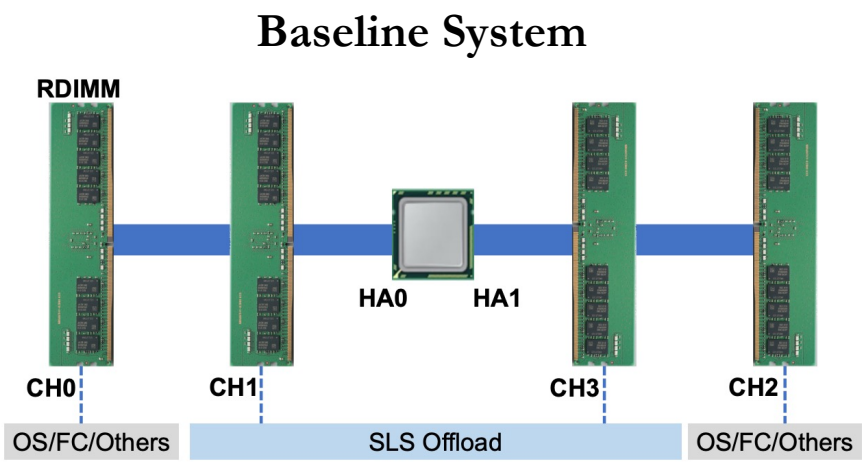


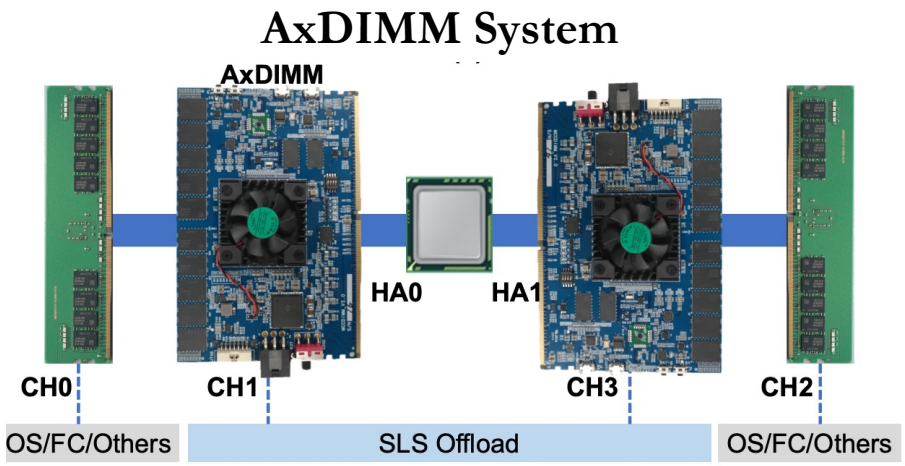
Element-wise summation of the embedding entries is performed inside the NMP PU, and the final pooling result is transferred back to host

AxDIMM Design: Overview

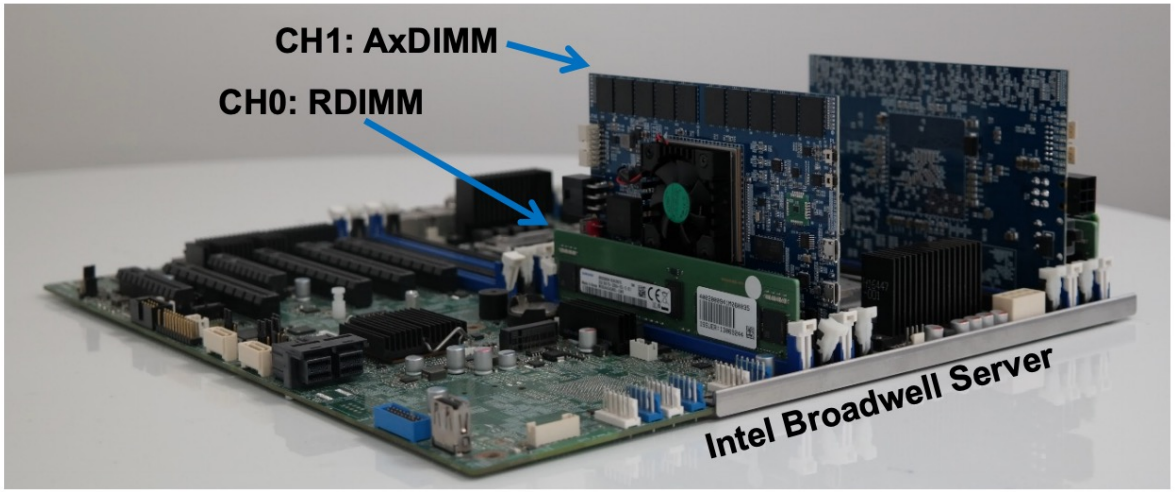
- Accelerator DIMM (AxDIMM)
  - DDR4-compatible FPGA-based platform with standard memory interfaces
- AxDIMM can potentially
  - support both in-order general-purpose processor and specialized accelerator modules
  - be an interesting prototyping platform for near-memory processing
- Personalized recommendation case study, including:
  - hardware implementation
  - software-stack support

AxDIMM System





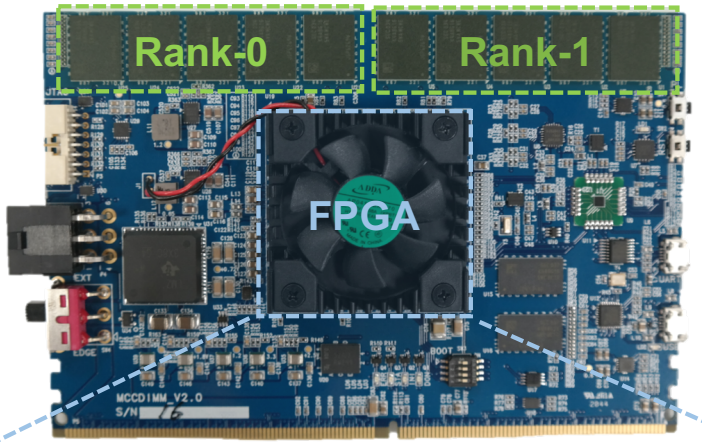
FPGA: Xilinx XCZU19EG FPGA



System was slowed down (1/3 of normal DDR4 memory channel speedup; CPU went from 3.2 GHz to 1.2 GHz) to keep up with the FPGA IO speed

AxDIMM Hardware & Architecture

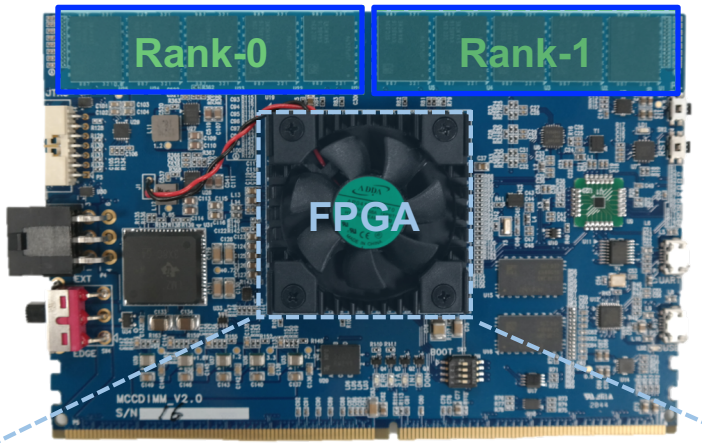
AxDIMM Design: Hardware Architecture



Standard DIMM Interface

FPGA board with standard DIMM interface:
It serves as a real-system
near-memory processing implementation

AxDIMM Design: Hardware Architecture

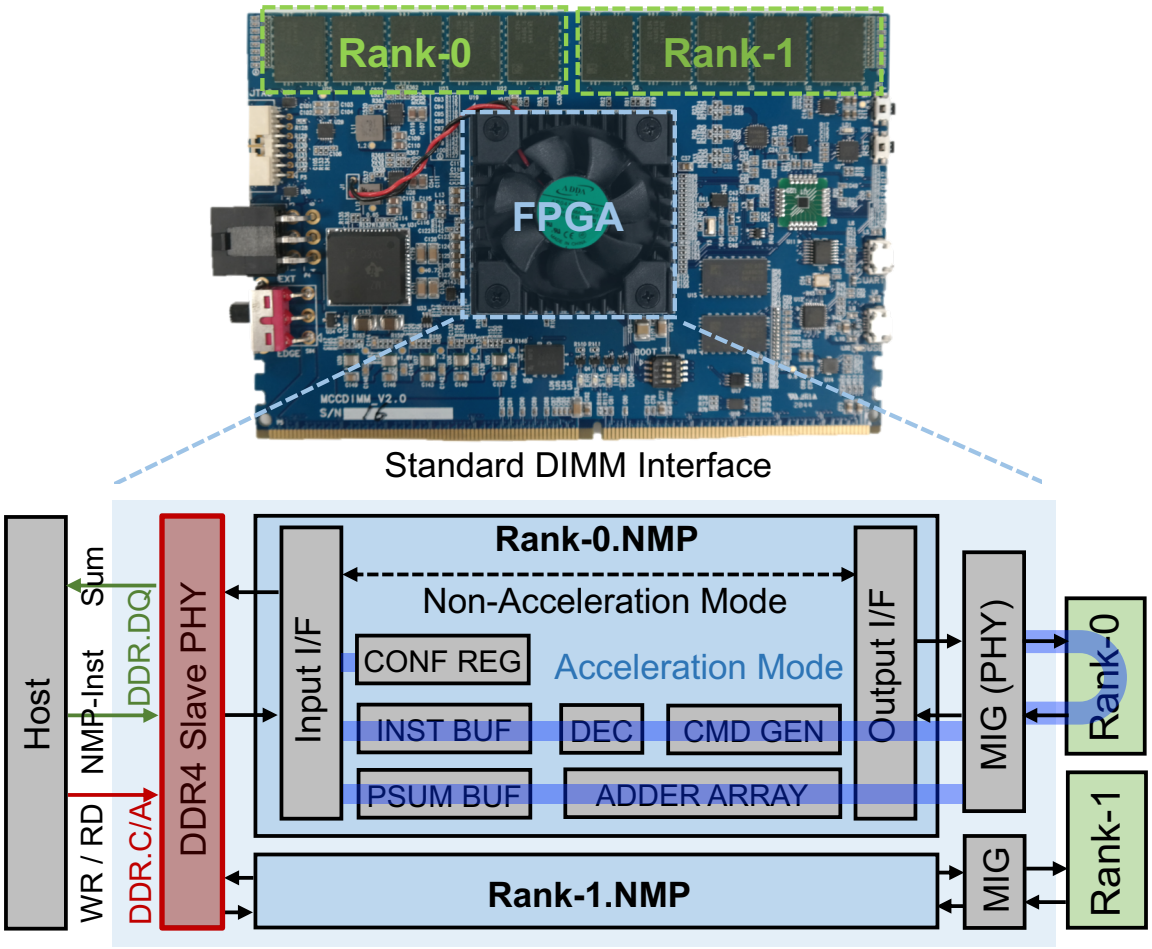


Standard DIMM Interface

Rank-level parallelism:
Two DRAM ranks are activated in parallel
to load embedding entries from memory

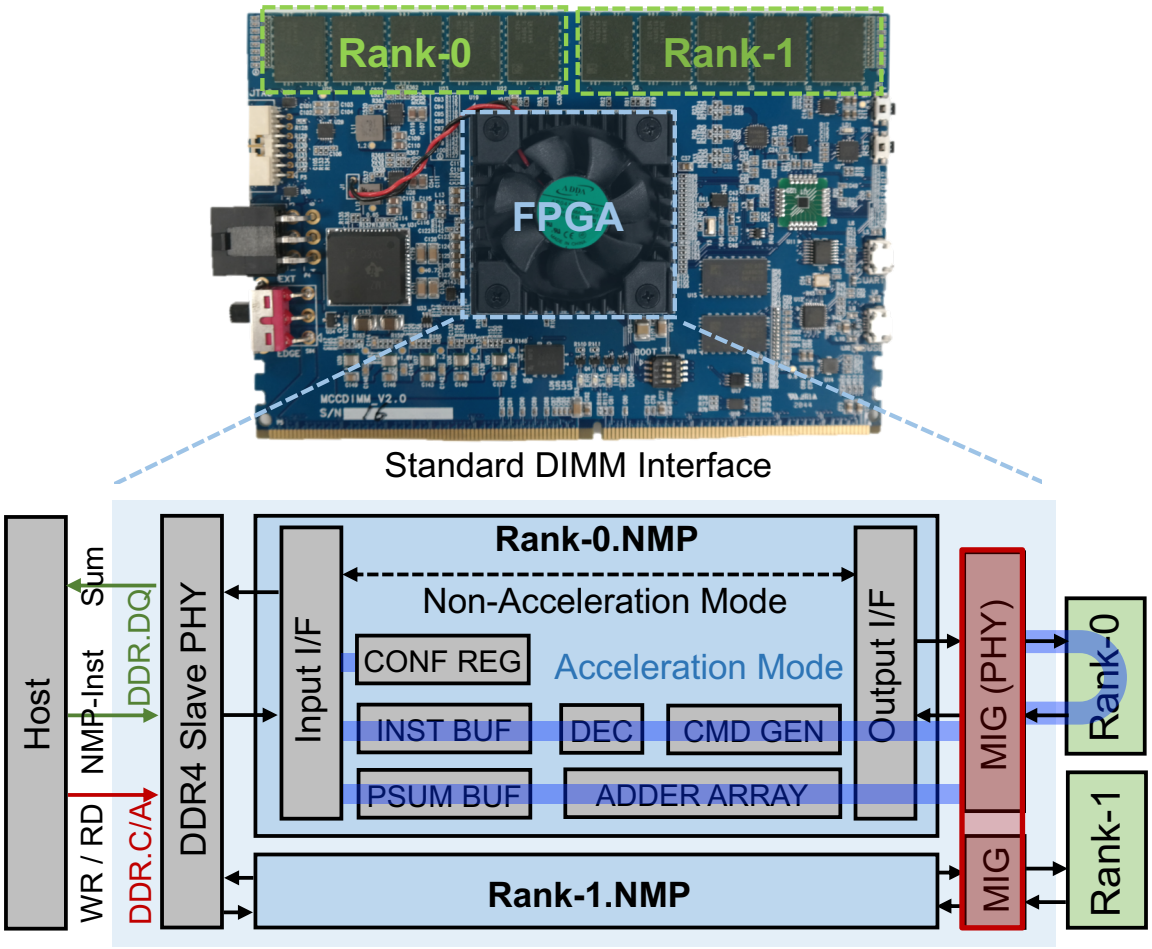
Element-wise summation
is performed inside the FPGA module

AxDIMM Design: Hardware Architecture



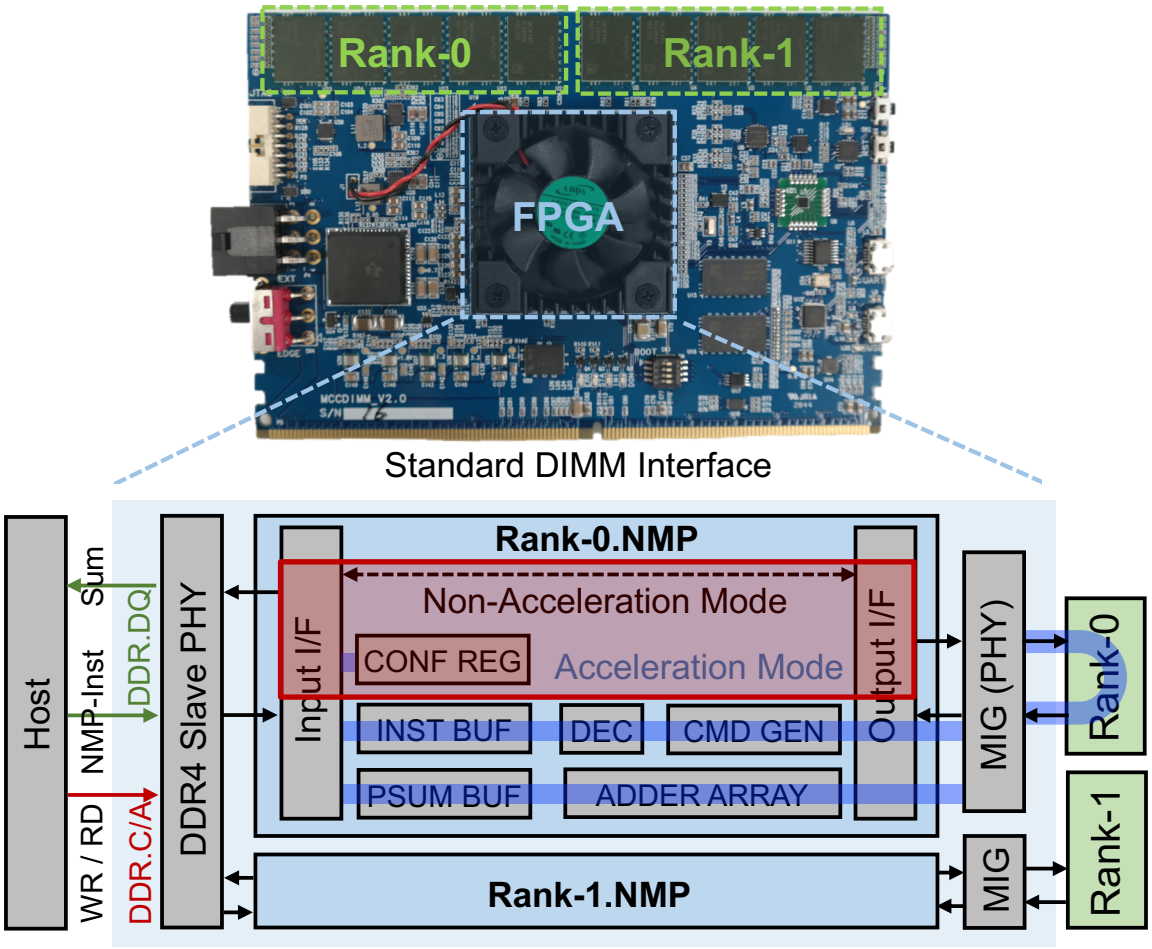
DDR4 slave PHY receives DRAM commands and NMP instructions (via DQ pins) from the host side

AxDIMM Design: Hardware Architecture



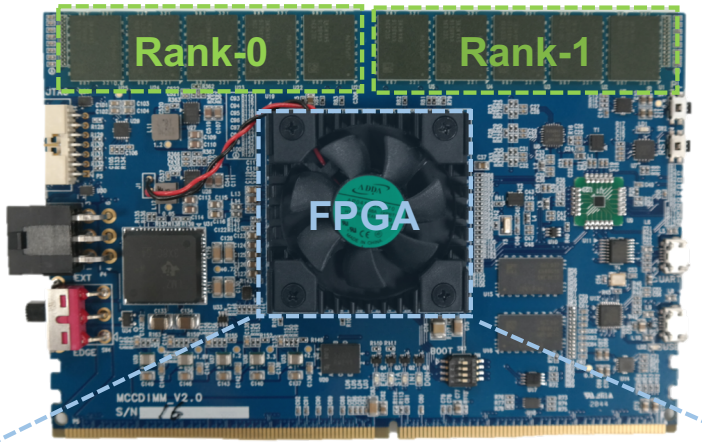
The memory interface generator (MIG) supports the internal rank accesses between Rank-NMP and the DRAM device

AxDIMM Design: Hardware Architecture

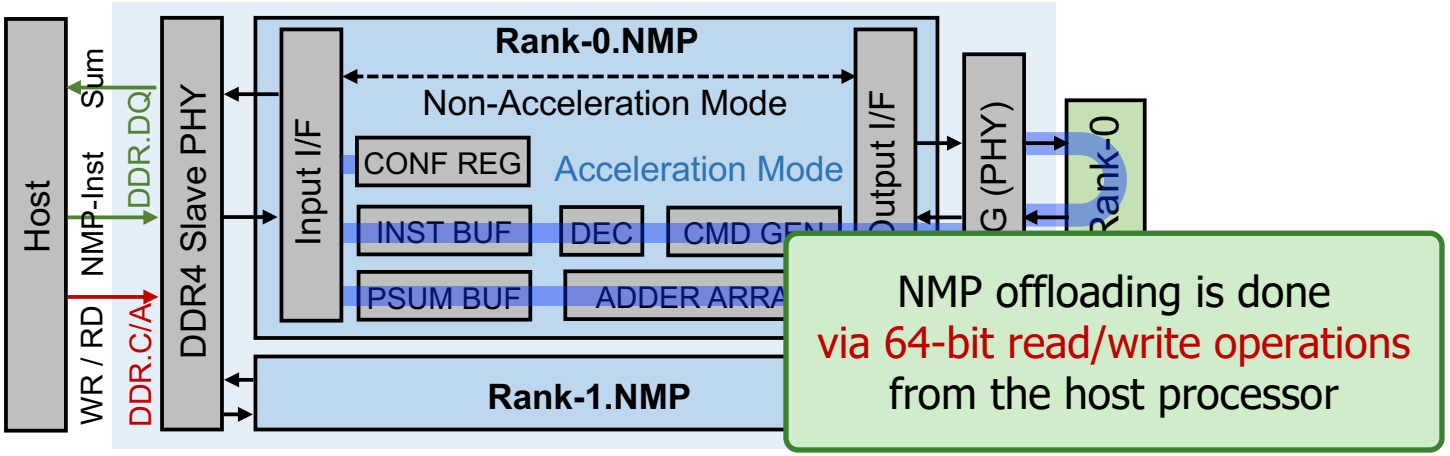


Two execution modes:
(1) non-acceleration mode
(2) acceleration mode (blocking)

AxDIMM Design: Hardware Architecture



Standard DIMM Interface

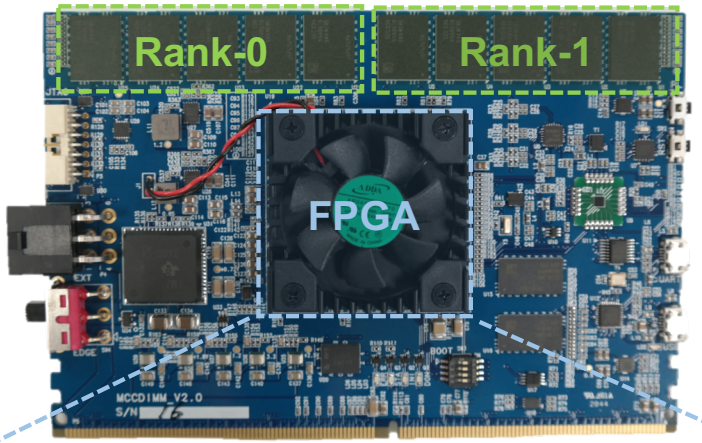


NMP offloading is done
via 64-bit read/write operations
from the host processor

NMP-Inst (64 bits)

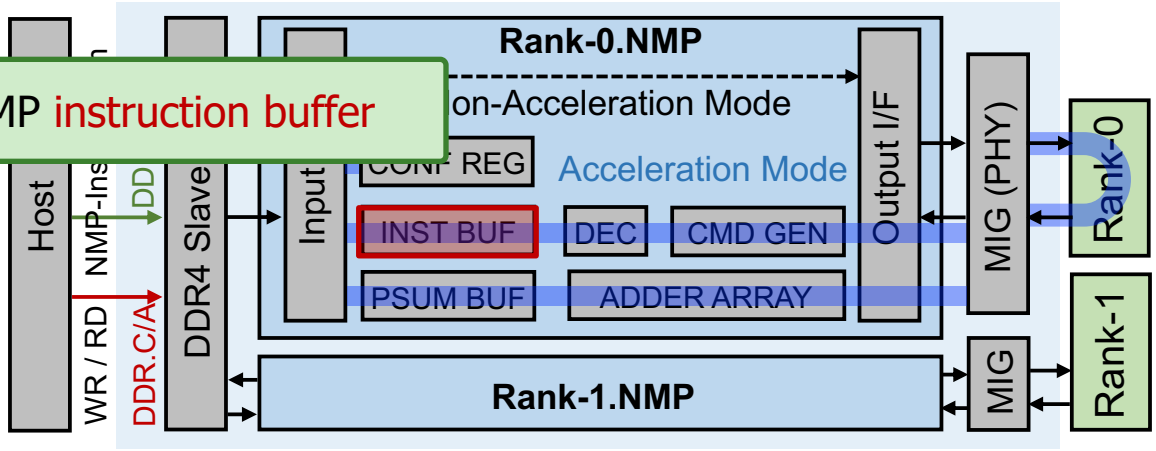
OpCode	Locality	PSUM Tag	Trace End	Reserved	Row Addr	BG	BA	Col Addr
2 bit	1 bit	12 bit	1 bit	17 bit	17 bit	2 bit	2 bit	10 bit

AxDIMM Design: Hardware Architecture



Standard DIMM Interface

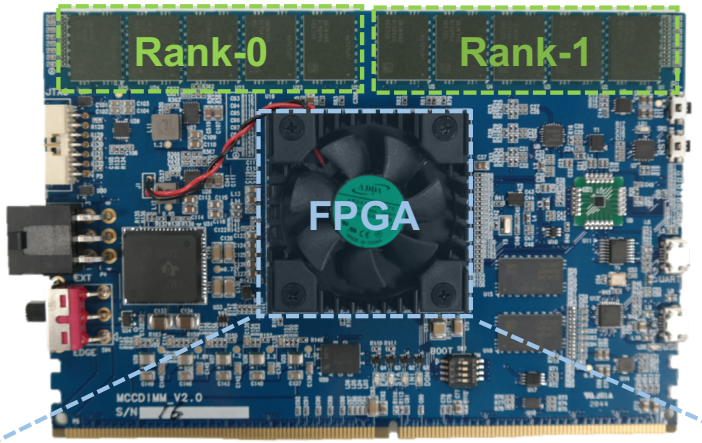
256-KB NMP instruction buffer



NMP-Inst (64 bits)

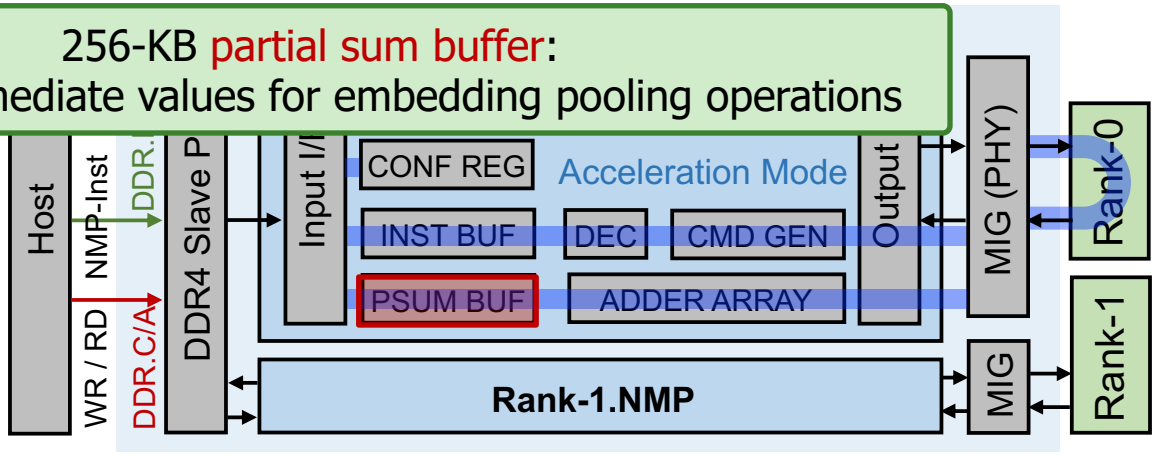
OpCode	Locality	PSUM Tag	Trace End	Reserved	Row Addr	BG	BA	Col Addr
2 bit	1 bit	12 bit	1 bit	17 bit	17 bit	2 bit	2 bit	10 bit

AxDIMM Design: Hardware Architecture



Standard DIMM Interface

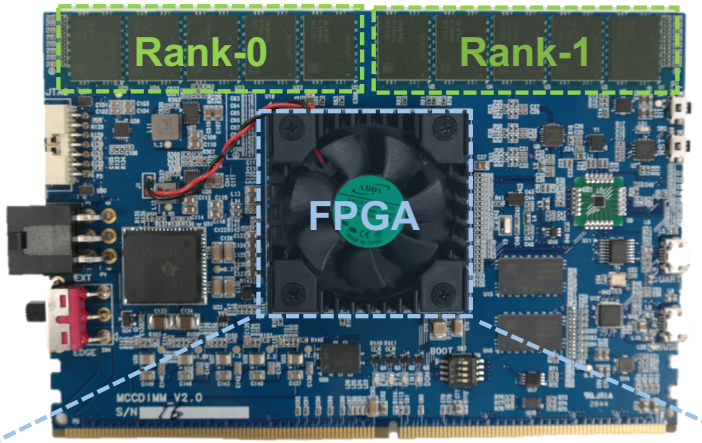
256-KB partial sum buffer:
It stores intermediate values for embedding pooling operations



NMP-Inst (64 bits)

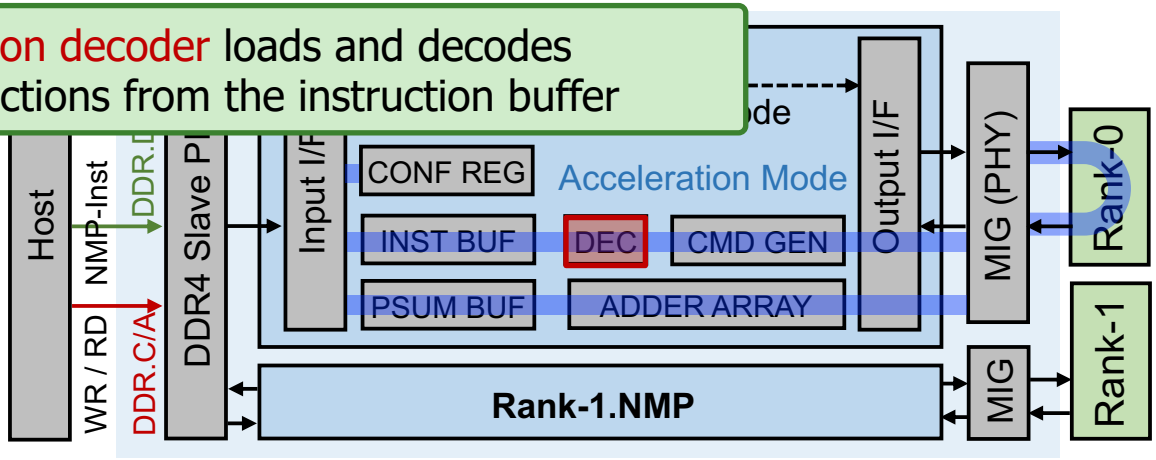
OpCode	Locality	PSUM Tag	Trace End	Reserved	Row Addr	BG	BA	Col Addr
2 bit	1 bit	12 bit	1 bit	17 bit	17 bit	2 bit	2 bit	10 bit

AxDIMM Design: Hardware Architecture



Standard DIMM Interface

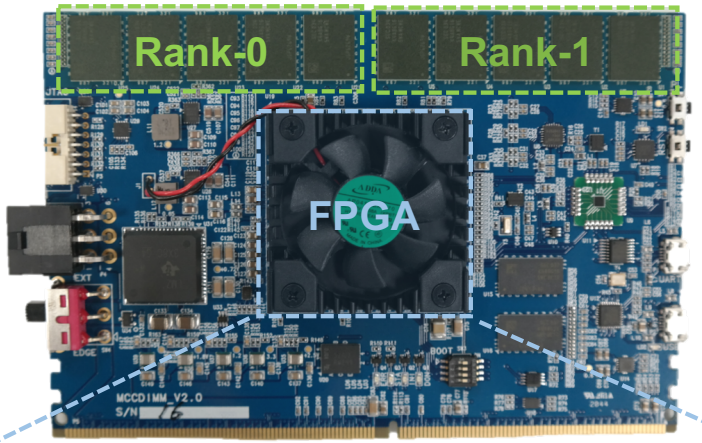
Instruction decoder loads and decodes
NMP instructions from the instruction buffer

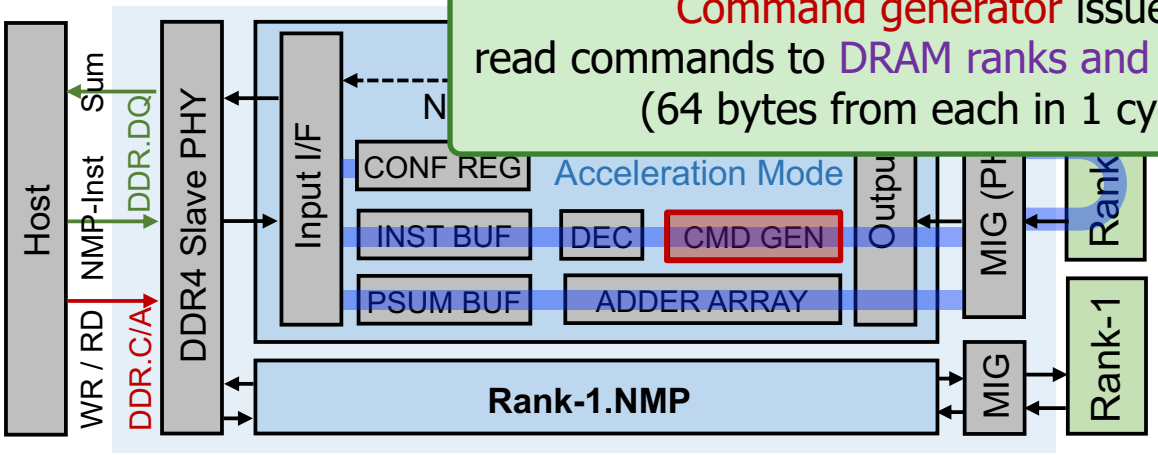


NMP-Inst
(64 bits)

OpCode	Locality	PSUM Tag	Trace End	Reserved	Row Addr	BG	BA	Col Addr
2 bit	1 bit	12 bit	1 bit	17 bit	17 bit	2 bit	2 bit	10 bit

AxDIMM Design: Hardware Architecture

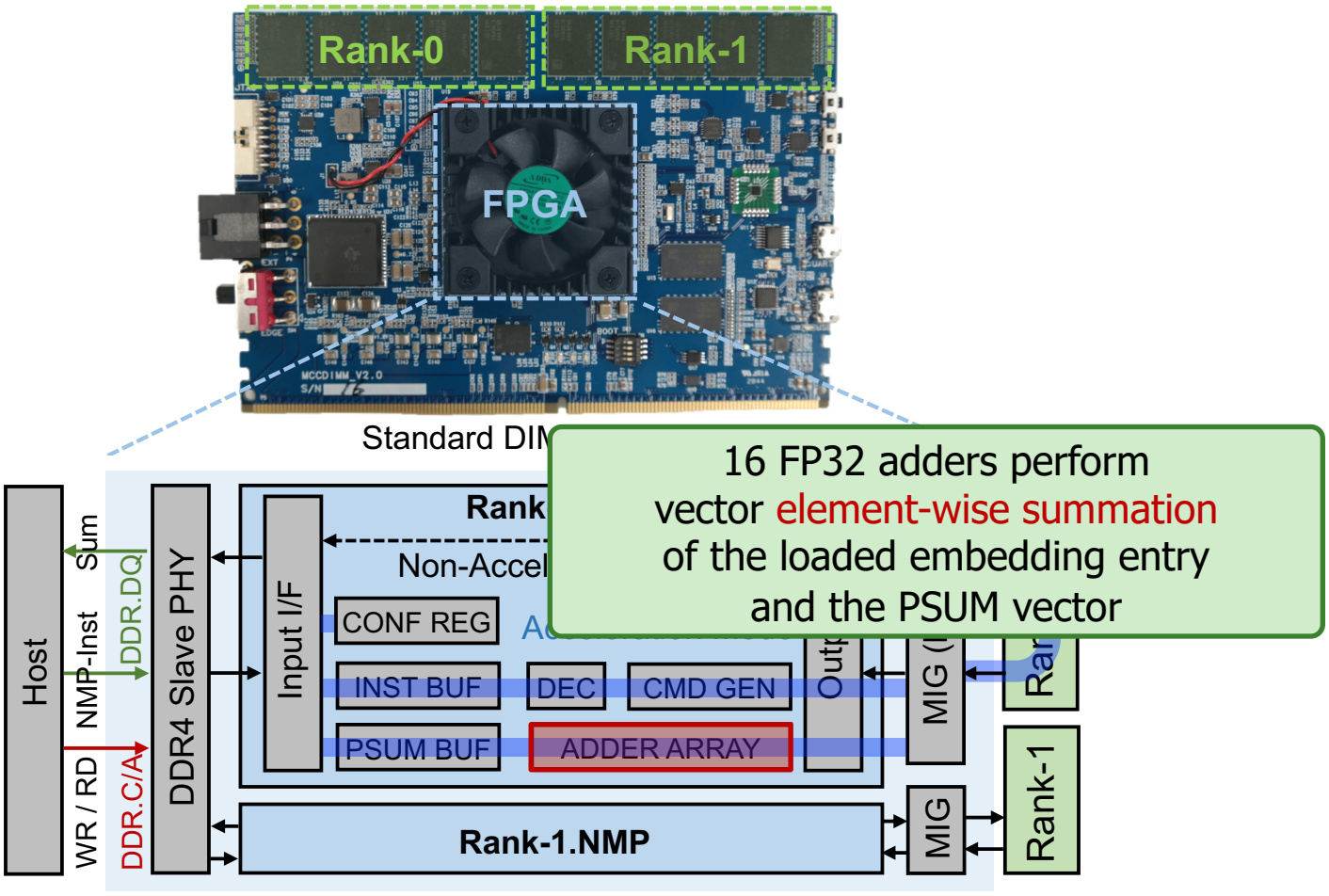




NMP-Inst
(64 bits)

OpCode	Locality	PSUM Tag	Trace End	Reserved	Row Addr	BG	BA	Col Addr
2 bit	1 bit	12 bit	1 bit	17 bit	17 bit	2 bit	2 bit	10 bit

AxDIMM Design: Hardware Architecture

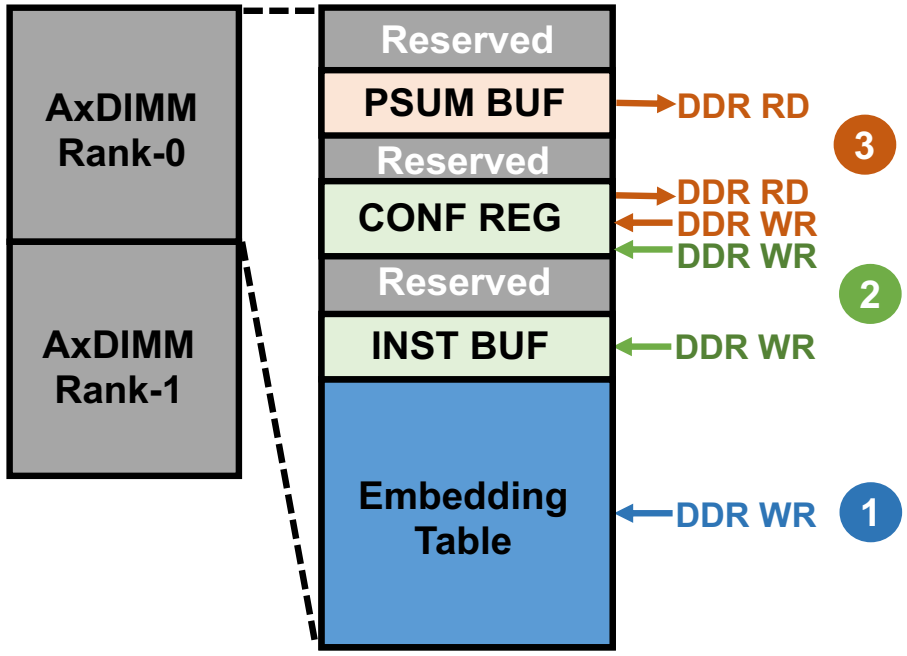


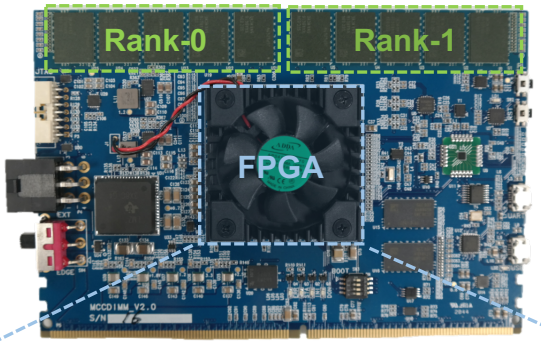
NMP-Inst (64 bits)

OpCode	Locality	PSUM Tag	Trace End	Reserved	Row Addr	BG	BA	Col Addr
2 bit	1 bit	12 bit	1 bit	17 bit	17 bit	2 bit	2 bit	10 bit

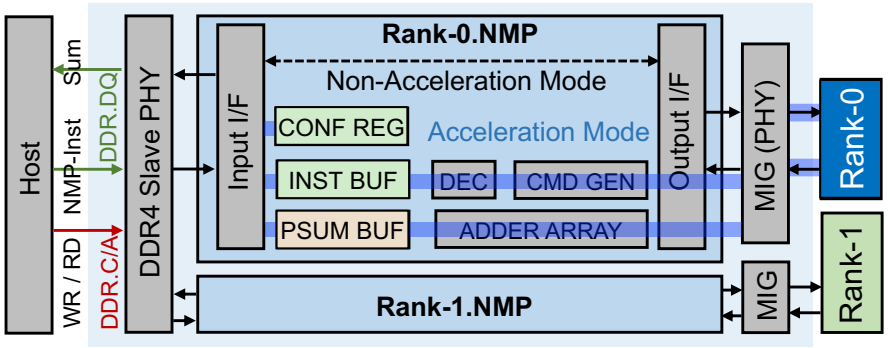
AxDIMM Design: Address Map

- Memory map of AxDIMM



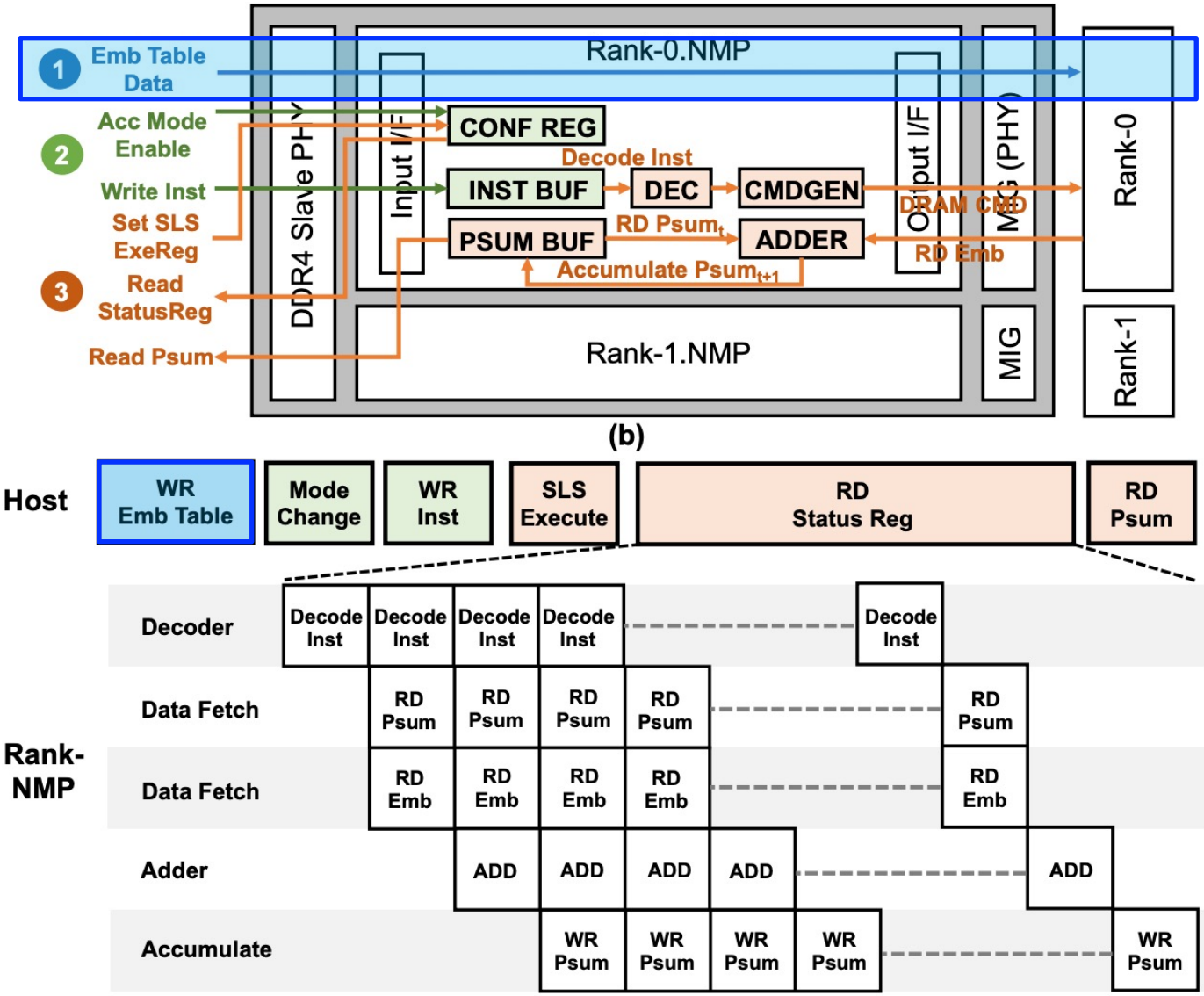


Standard DIMM Interface

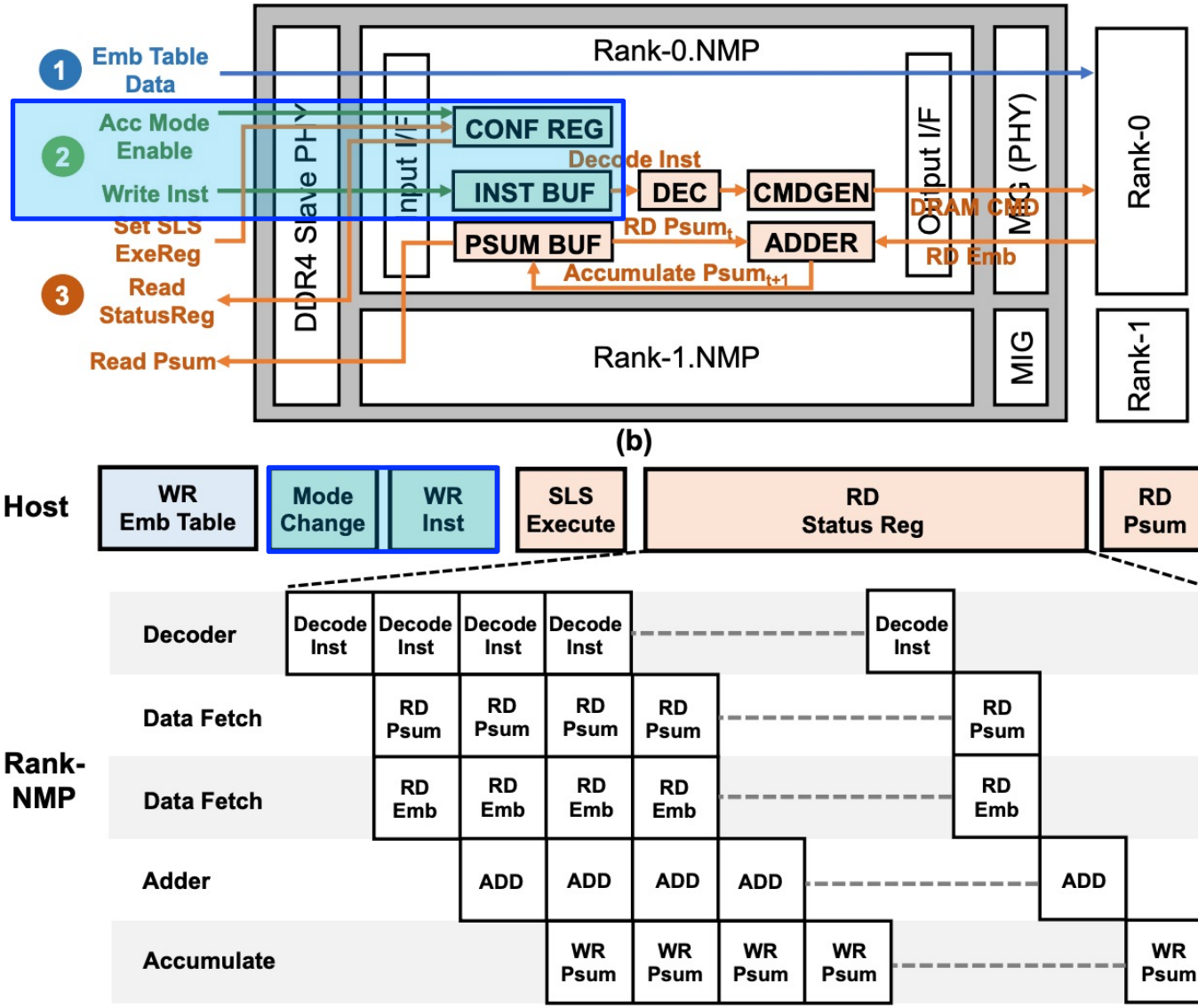


AxDIMM Execution Flow

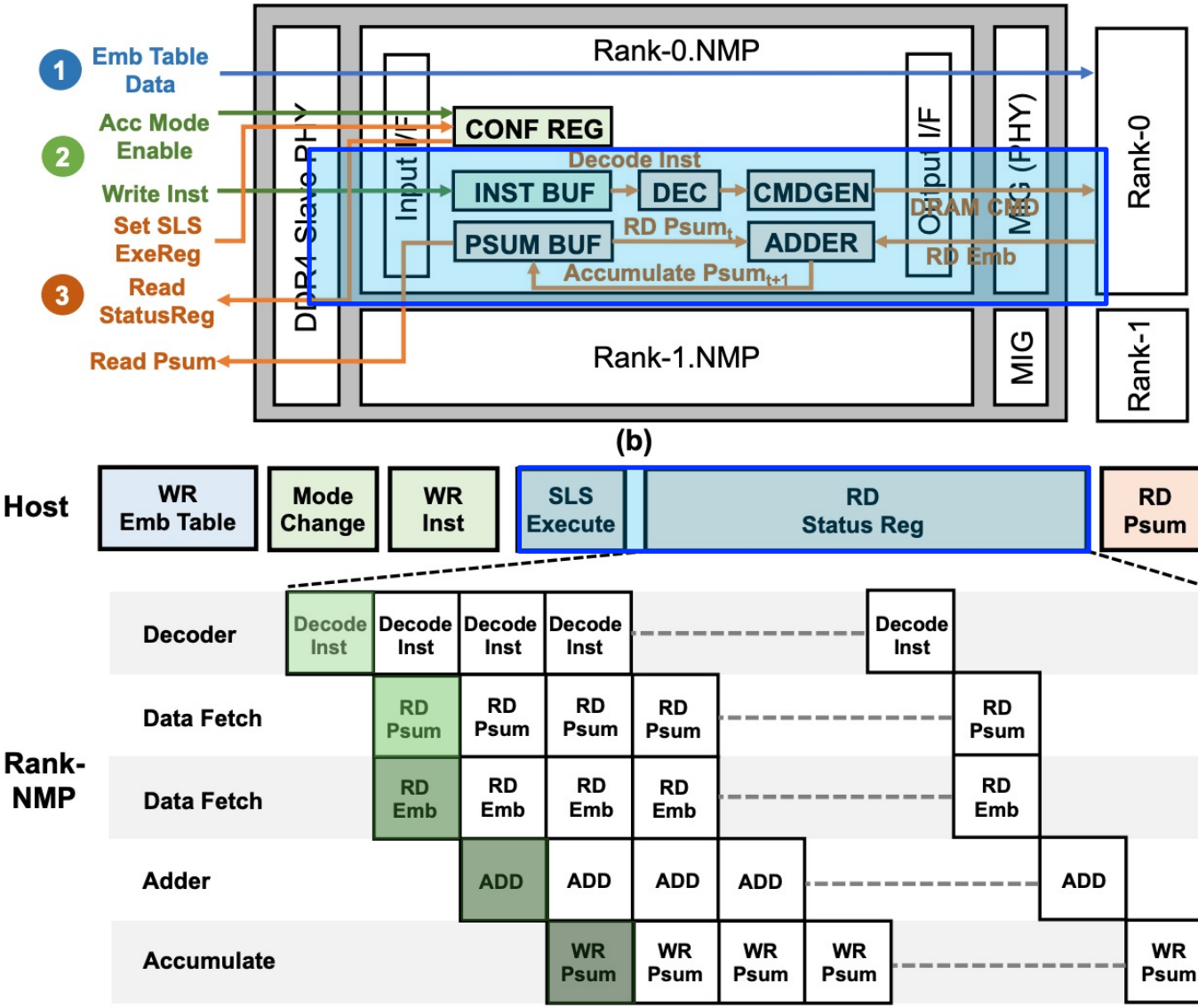
AxDIMM Design: Execution Flow



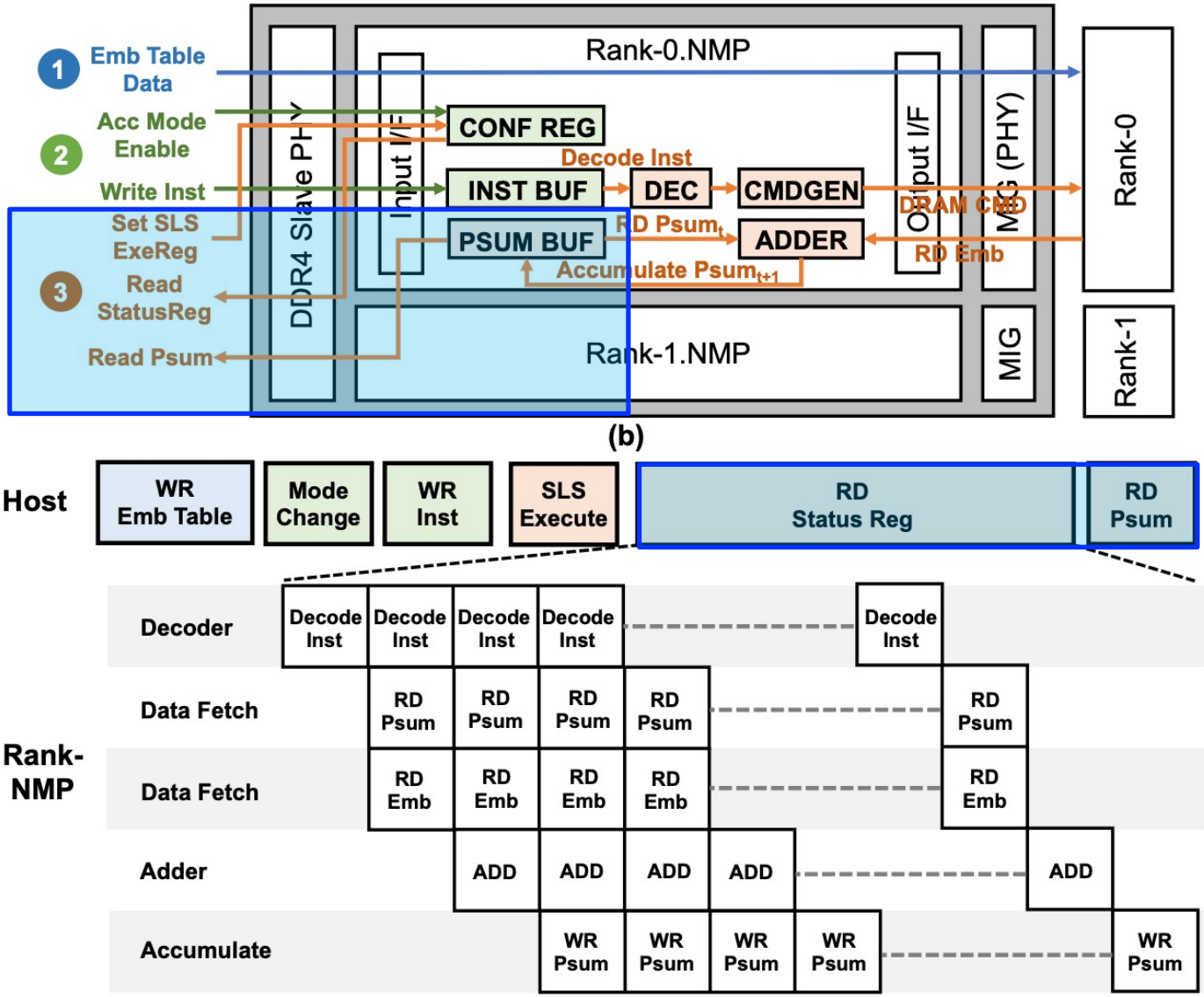
AxDIMM Design: Execution Flow



AxDIMM Design: Execution Flow



AxDIMM Design: Execution Flow



PnM with AxDIMM (IEEE Micro 2021)

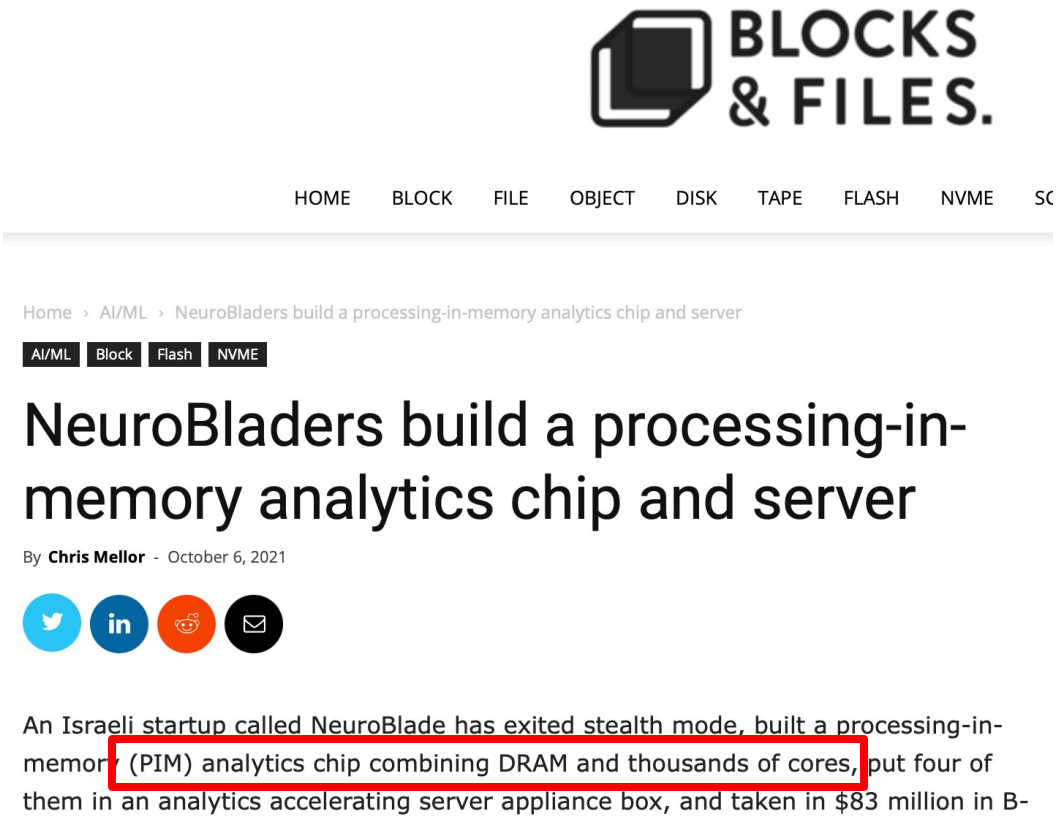
Near-Memory Processing in Action: Accelerating Personalized Recommendation with AxDIMM

Liu Ke*†, Xuan Zhang†, Jinin So‡, Jong-Geon Lee‡, Shin-Haeng Kang‡, Sukhan Lee‡, Songyi Han‡, YeonGon Cho‡, JIN Hyun Kim‡, Yongsuk Kwon‡, KyungSoo Kim‡, Jin Jung‡, Ilkwon Yun‡, Sung Joo Park‡, Hyunsun Park‡, Joonho Song‡, Jeonghyeon Cho‡, Kyomin Sohn‡, Nam Sung Kim‡, Hsien-Hsin S. Lee*

*Facebook, †Washington University in St. Louis, ‡Samsung

https://doi.org/10.1109/MM.2021.3097700

More Real-World PIM to Come



BLOCKS & FILES.

HOME BLOCK FILE OBJECT DISK TAPE FLASH NVME SC

Home › AI/ML › NeuroBladers build a processing-in-memory analytics chip and server

AI/ML Block Flash NVME

NeuroBladers build a processing-in-memory analytics chip and server

By **Chris Mellor** - October 6, 2021

An Israeli startup called NeuroBlade has exited stealth mode, built a processing-in-memory (PIM) analytics chip combining DRAM and thousands of cores, put four of them in an analytics accelerating server appliance box, and taken in $83 million in B-round funding.

The idea is to take a GPU approach to big data-style analytics and AI software by employing a massively parallel core design, but take it further by layering the cores on DRAM with a wide I/O bus architecture design linking the cores and memory to speed processing even more. This design vastly reduces data movement between storage and memory and also accelerates data transfer between memory and processing cores.

https://blocksandfiles.com/2021/10/06/neurobladers-build-a-processing-in-memory-analytics-chip-and-server/

NeuroBlade Patent (I)

(12) **United States Patent**
Sity et al.

(10) **Patent No.: US 10,762,034 B2**
(45) **Date of Patent: Sep. 1, 2020**

(54) **MEMORY-BASED DISTRIBUTED PROCESSOR ARCHITECTURE**

(71) Applicant: **NeuroBlade, Ltd.**, Hod-Hashron (IL)

(72) Inventors: **Elad Sity**, Kfar Saba (IL); **Eliad Hillel**, Kfar Saba (IL)

(73) Assignee: **NeuroBlade, Ltd.**, Hod-Hashron (IL)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **16/512,590**

(22) Filed: **Jul. 16, 2019**

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,837,747 A * 6/1989 Dosaka G11C 8/12 365/189.05

5,155,729 A 10/1992 Rysko et al.

(Continued)

FOREIGN PATENT DOCUMENTS

CA 2 149 479 C 5/2001

OTHER PUBLICATIONS

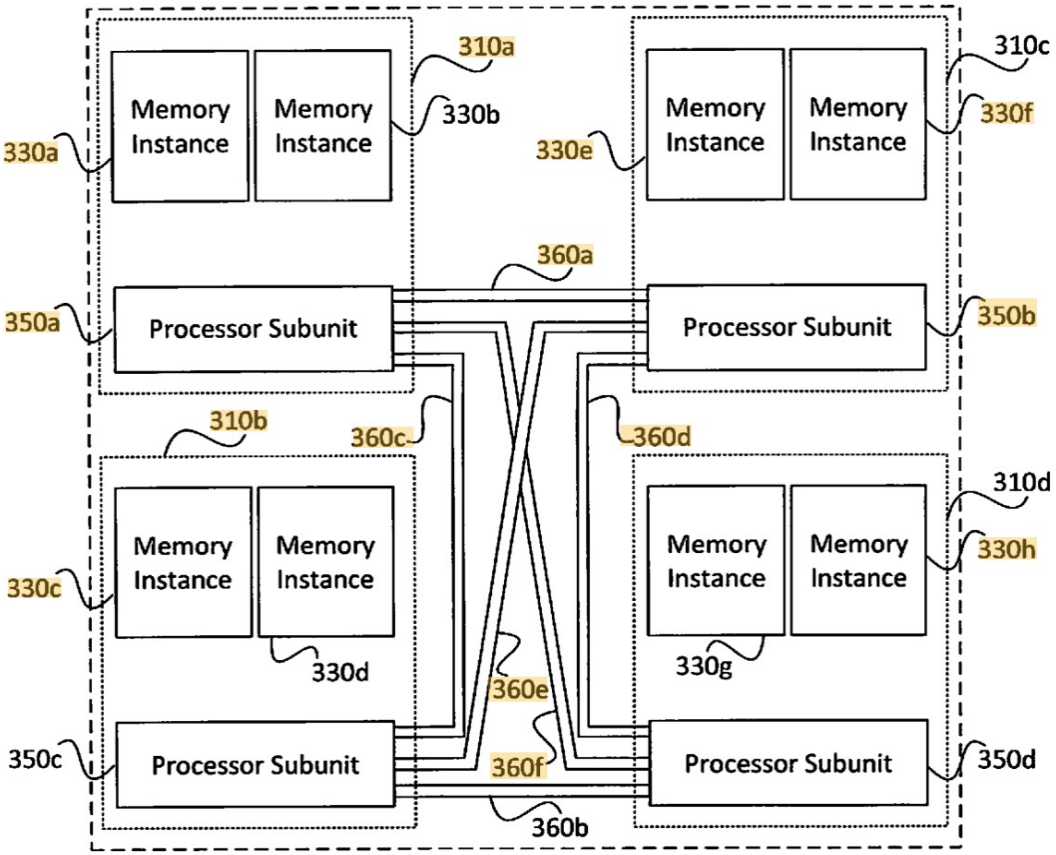
Ahn et al., "A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing," ISCA '15 (Jun. 13-17, 2015), pp. 105-117.

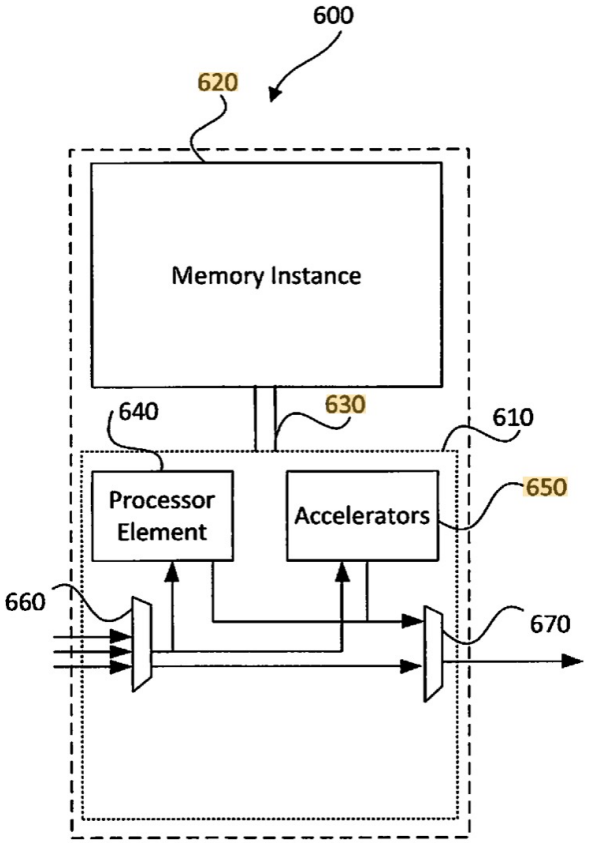
(57) **ABSTRACT**

Distributed processors and methods for compiling code for execution by distributed processors are disclosed. In one implementation, a distributed processor may include a substrate; a memory array disposed on the substrate; and a processing array disposed on the substrate. The memory array may include a plurality of discrete memory banks, and the processing array may include a plurality of processor subunits, each one of the processor subunits being associated with a corresponding, dedicated one of the plurality of discrete memory banks. The distributed processor may further include a first plurality of buses, each connecting one of the plurality of processor subunits to its corresponding, dedicated memory bank, and a second plurality of buses, each connecting one of the plurality of processor subunits to another of the plurality of processor subunits.

Sity et al., "Memory-based Distributed Processor Architecture," US 10,762,034 B2

NeuroBlade Patent (II)





Sity et al., “Memory-based Distributed Processor Architecture,” US 10,762,034 B2

NeuroBlade: Xiphos

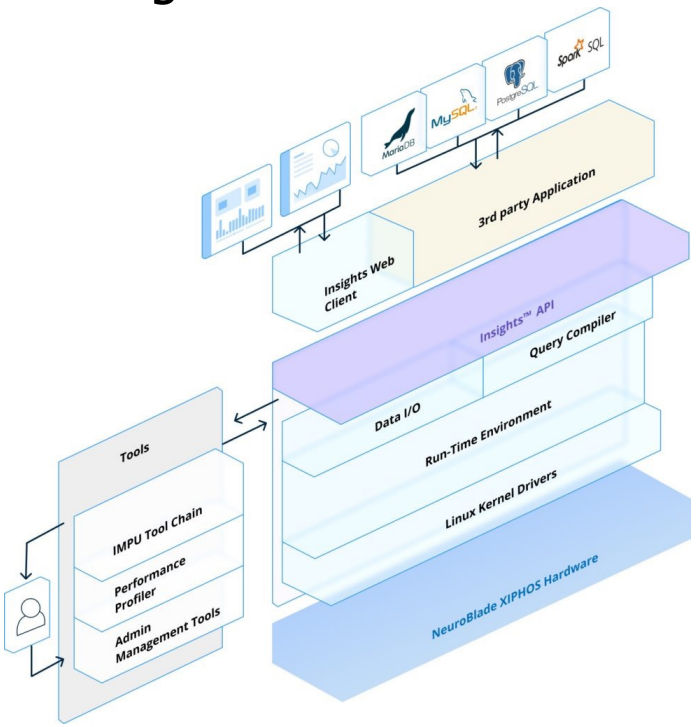
- PIM XRAM chip
  - IMPU (Intensive Memory Processing Unit)
- x86 CPU, 32 NVMe SSDs
- PCIe fabric: "Everything is connected on top of PCIe fabric."
- Wide I/O bus: multiple x16 PCIe buses

Xiphos appliance.

https://www.neuroblade.com

NeuroBlade: Sofware Suite

- Xiphos SW suite: Insights API
  - APIs for 3rd party applications and web client
- Data I/O
  - ETL process populates and updates local storage
- Query Compiler
  - Generates query execution plans
- Tools
  - E.g., visual profiler
- TPC benchmarks and queries



Hybrid Bonding with PnM Engine (ISSCC 2022)

ISSCC 2022 / SESSION 29 / ML CHIPS FOR EMERGING A

29.1 184QPS/W 64Mb/mm² 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System

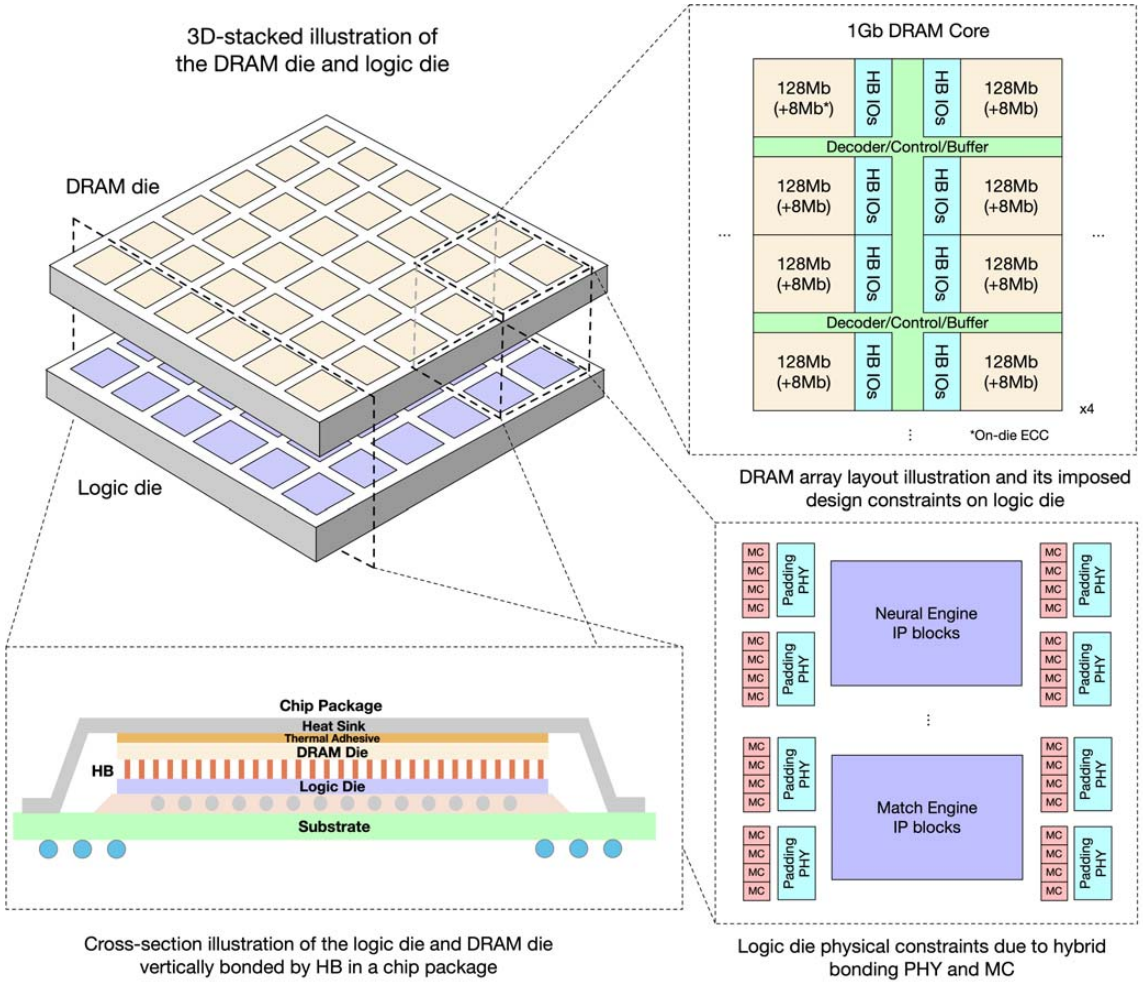
Dimin Niu[1], Shuangchen Li[1], Yuhao Wang[1], Wei Han[1], Zhe Zhang[2], Yijin Guan[2], Tianchan Guan[3], Fei Sun[1], Fei Xue[1], Lide Duan[1], Yuanwei Fang[1], Hongzhong Zheng[1], Xiping Jiang[4], Song Wang[4], Fengguo Zuo[4], Yubing Wang[4], Bing Yu[4], Qiwei Ren[4], Yuan Xie[1]

[1]Alibaba DAMO Academy, Sunnyvale, CA; [2]Alibaba DAMO Academy, Beijing, China
[3]Alibaba DAMO Academy, Shanghai, China; [4]UniIC, Xian, China

https://doi.org/10.1109/ISSCC42614.2022.9731694

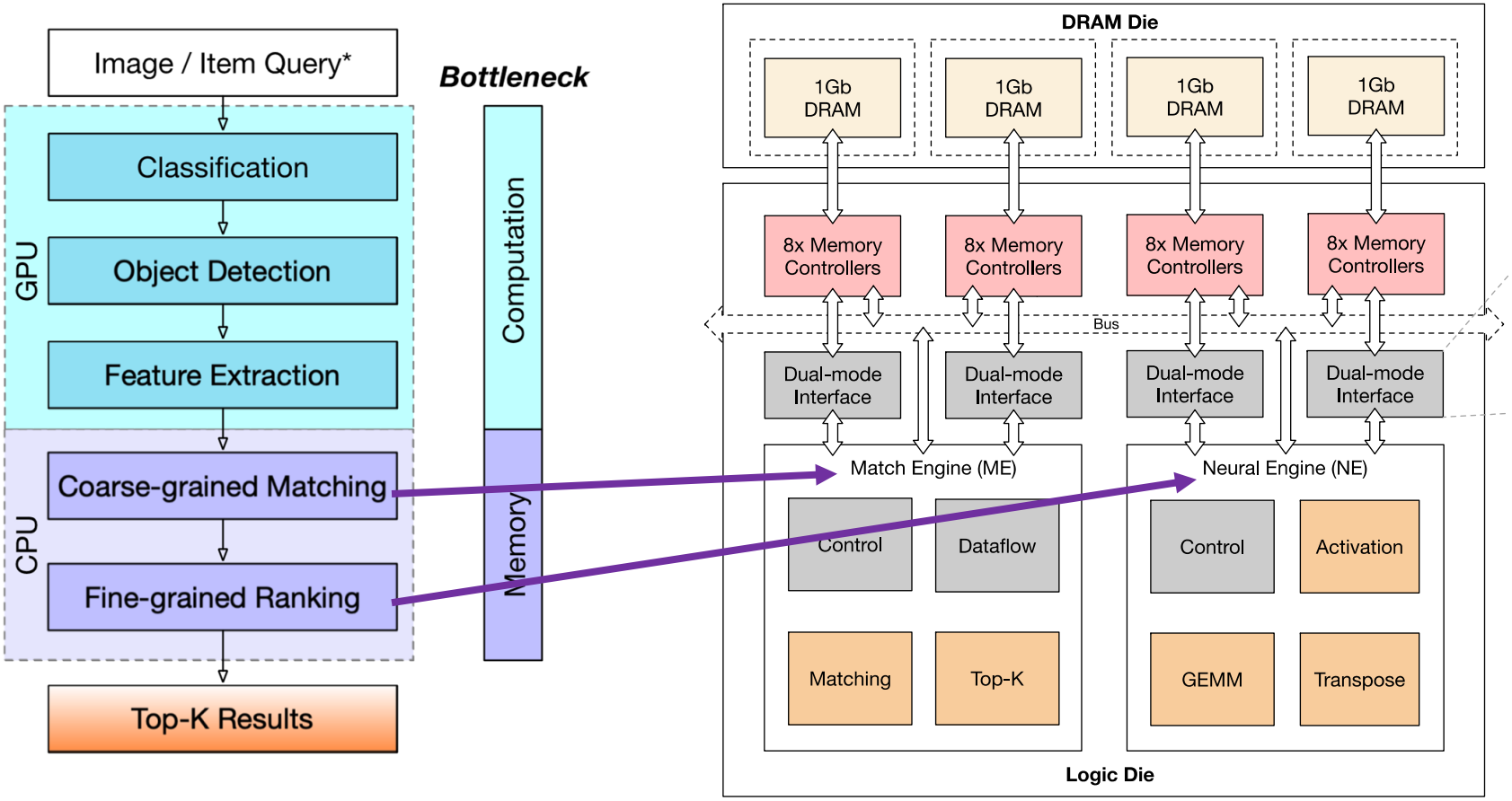
HB-PNM: Overall Architecture (I)

- 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



Upcoming Lectures

- More real-world PIM architectures
- PUM architectures and prototypes
- Enabling the adoption of PIM

P&S Processing-in-Memory

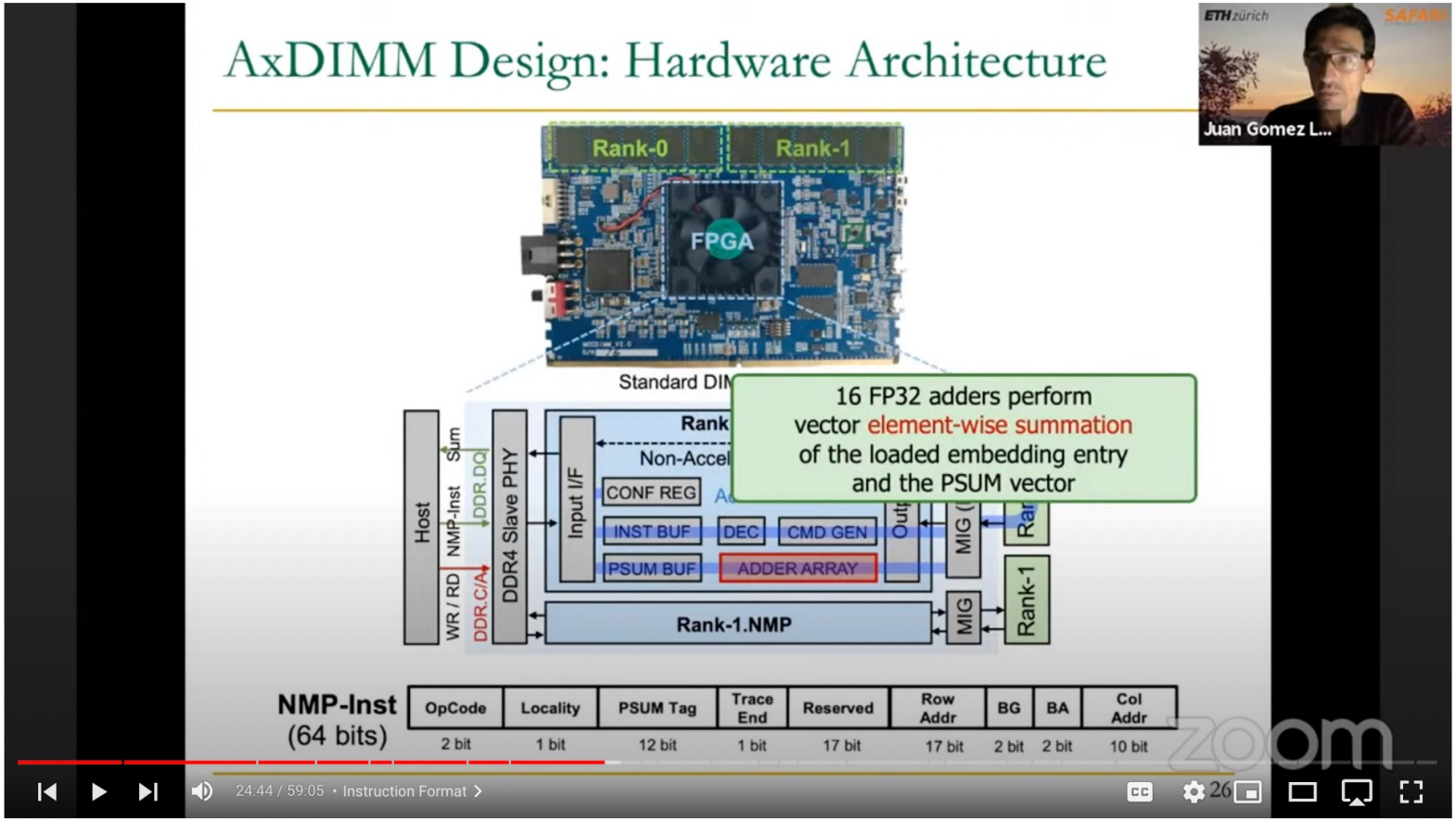
Real-World Processing-in-Memory Architectures: Samsung AxDIMM

Dr. Juan Gómez Luna
Prof. Onur Mutlu
ETH Zürich

Fall 2022

22 November 2022

Another Lecture on AxDIMM



Livestream - P&S Exploring the Processing-in-Memory Paradigm for Future Computing Systems (Fall 2021)

Processing in Memory Course: Meeting 5: Real-world PIM architectures IV - Fall'21

Onur Mutlu Lectures
29.4K subscribers
Subscribed
18
Share
Clip
Save

661 views Streamed live on Nov 2, 2021
Project & Seminar, ETH Zürich, Fall 2021
Exploring the Processing-in-Memory Paradigm for Future Computing Systems (https://safari.ethz.ch/projects_and_s...)

https://youtu.be/2FMQg786GKs