P&S Processing-in-Memory

Real-World Processing-in-Memory Architectures: Alibaba Hybrid Bonding PNM Engine

Dr. Juan Gómez Luna

Prof. Onur Mutlu

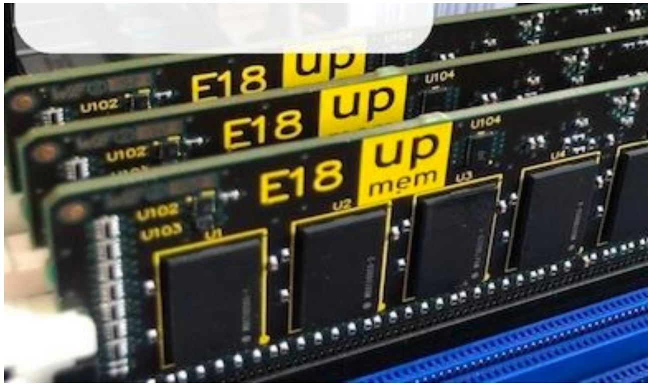
ETH Zürich

Spring 2022

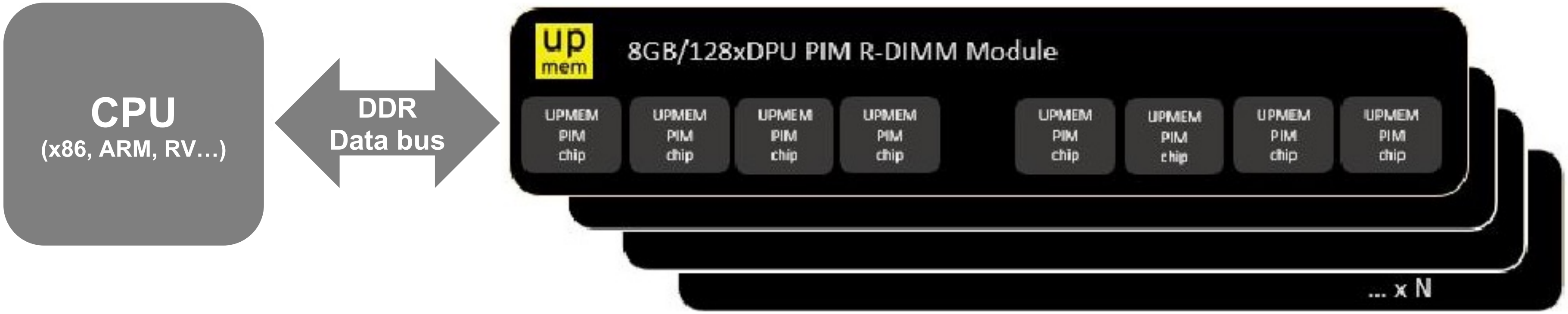
12 May 2022

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.



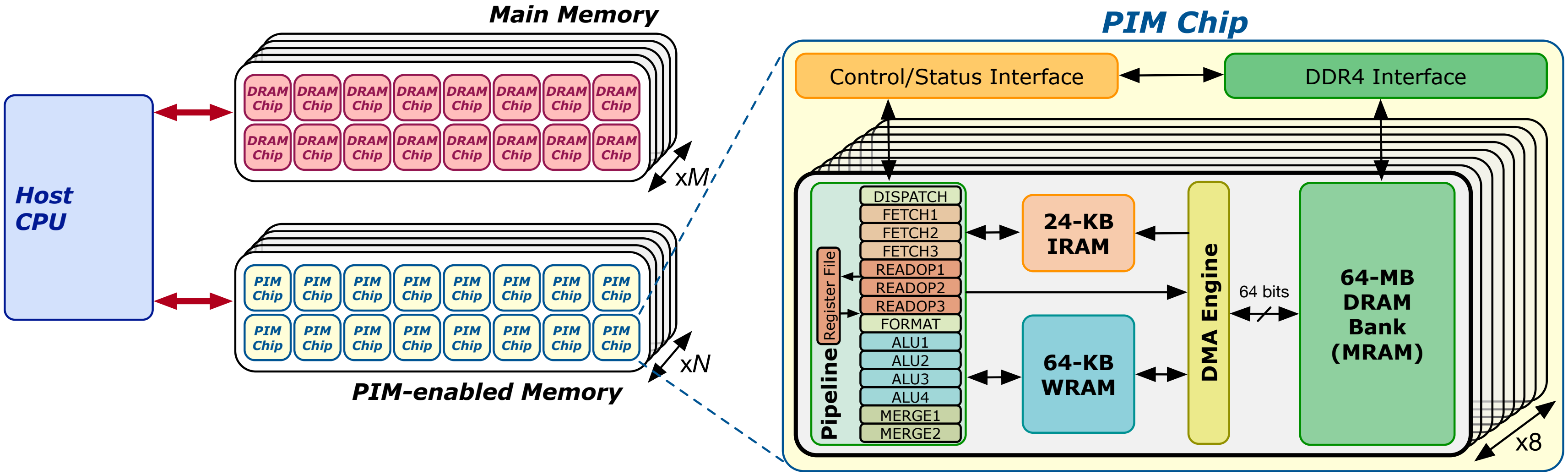
- Replaces **standard** DIMMs
  - DDR4 R-DIMM modules
    - 8GB+128 DPUs (16 PIM chips)
    - Standard 2x-nm DRAM process
  - **Large amounts of** compute & memory bandwidth



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/

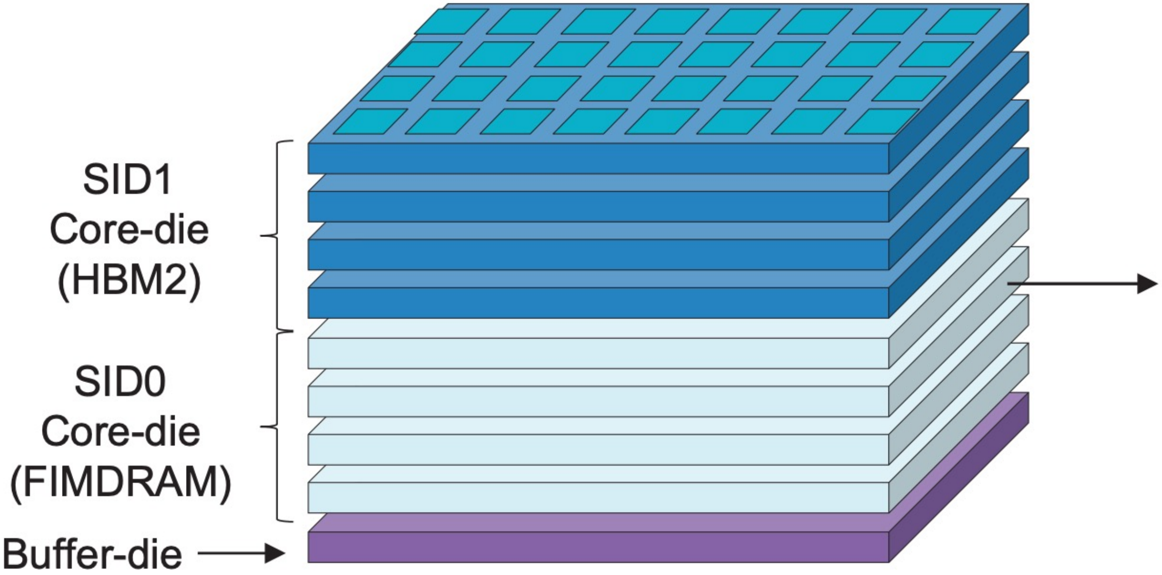
UPMEM PIM System Organization

- 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



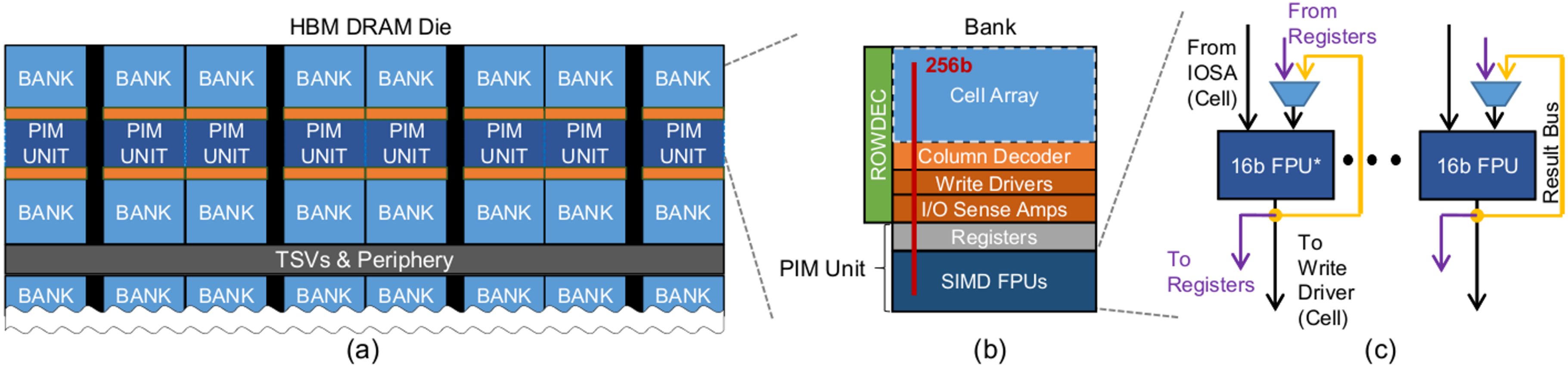
[3D Chip Structure of HBM with FIMDRAM]

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)

Kwon et al., A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications, ISSCC 2021

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

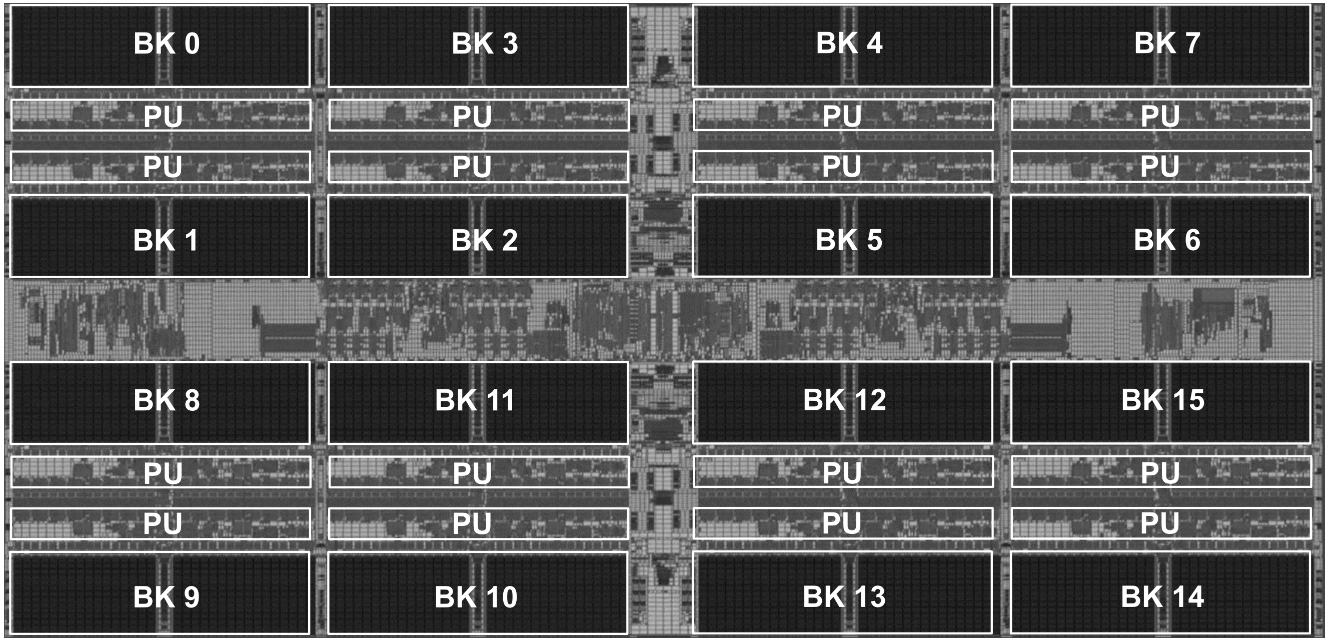


Lee et al., Hardware Architecture and Software Stack for PIM Based on Commercial DRAM Technology, ISCA 2021

AiM: Chip Implementation

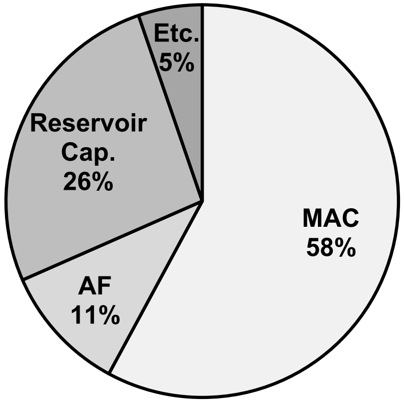
- 4 Gb AiM die with 16 processing units (PUs)

AiM Die Photograph



1 Process Unit (PU) Area

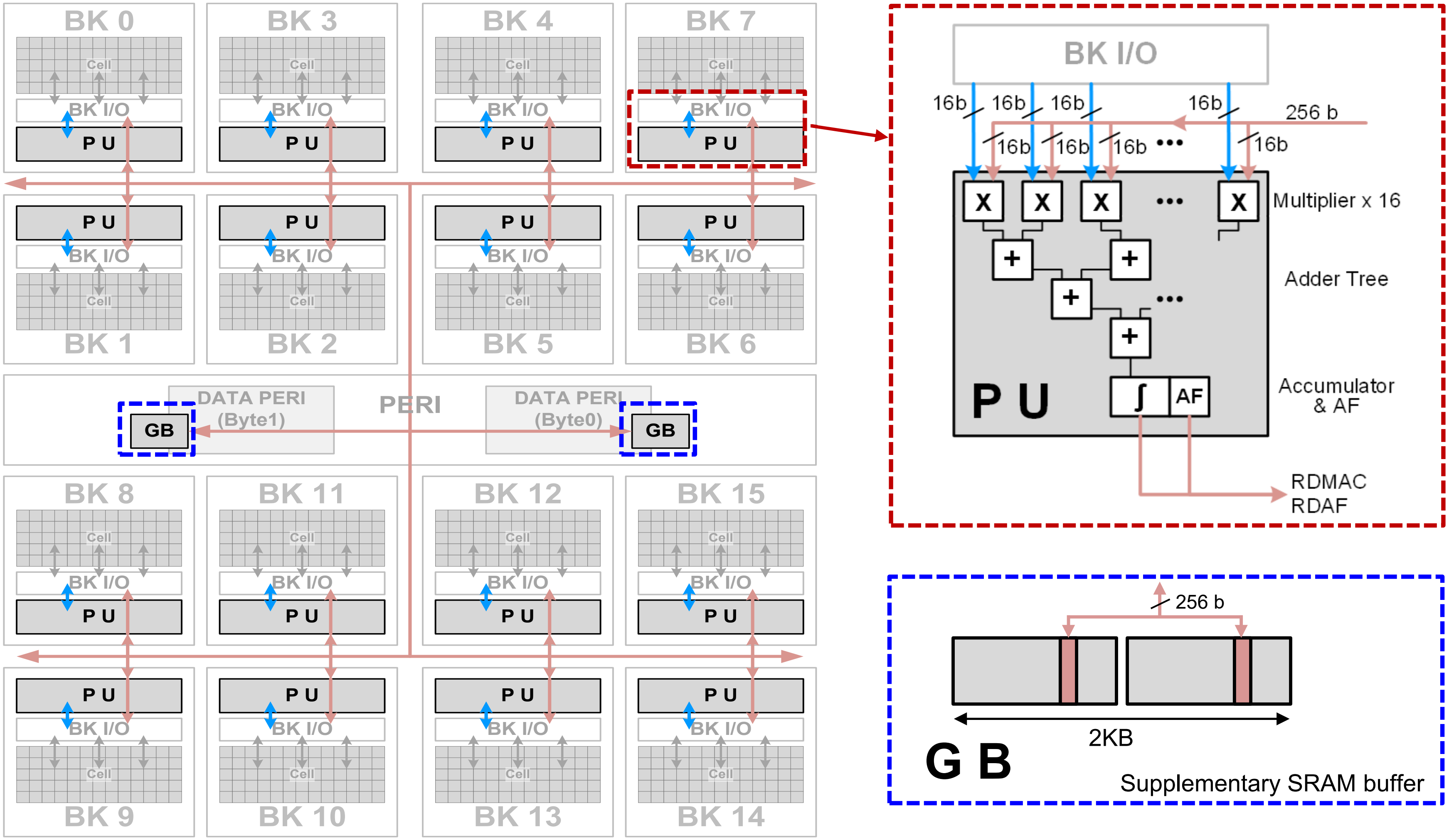
Total	0.19mm^2
MAC	0.11mm^2
Activation Function (AF)	0.02mm^2
Reservoir Cap.	0.05mm^2
Etc.	0.01mm^2



Lee et al., 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, ISSCC 2022

AiM: System Organization

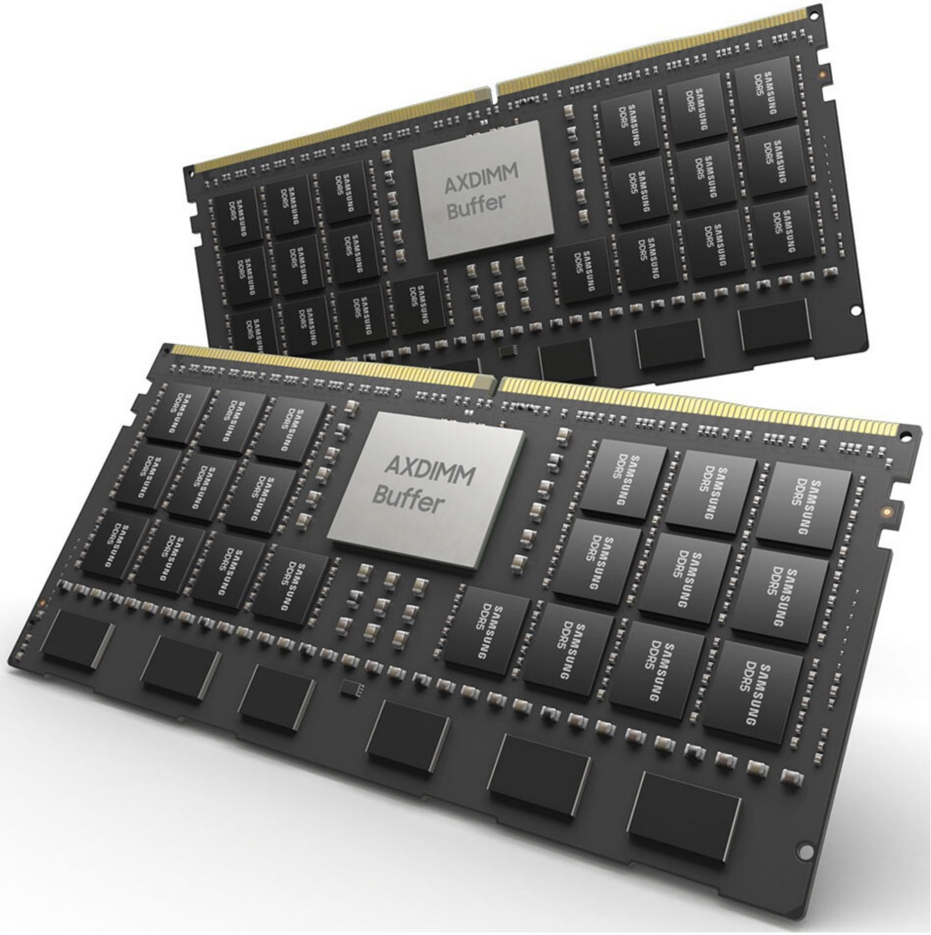
- GDDR6-based AiM architecture



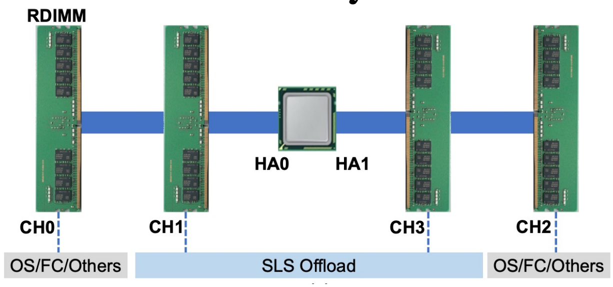
Lee et al., A 1ynm 1.25V 8Gb, 16Gb/s/pin GDDR6-based Accelerator-in-Memory supporting 1TFLOPS MAC Operation and Various Activation Functions for Deep-Learning Applications, ISSCC 2022

Samsung AxDIMM (2021)

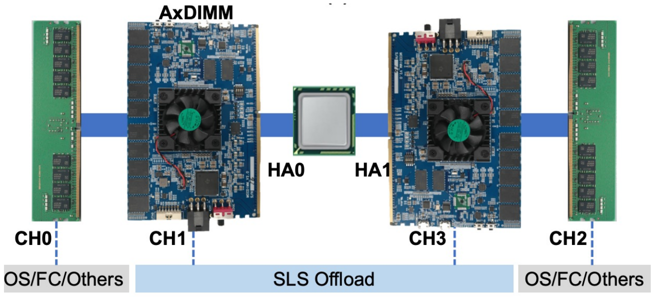
- DIMM-based PIM
  - DLRM recommendation system

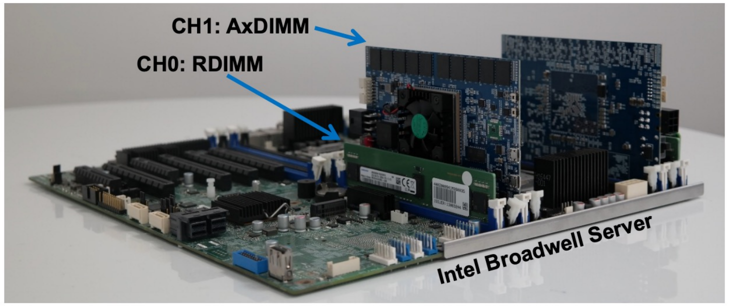


Baseline System



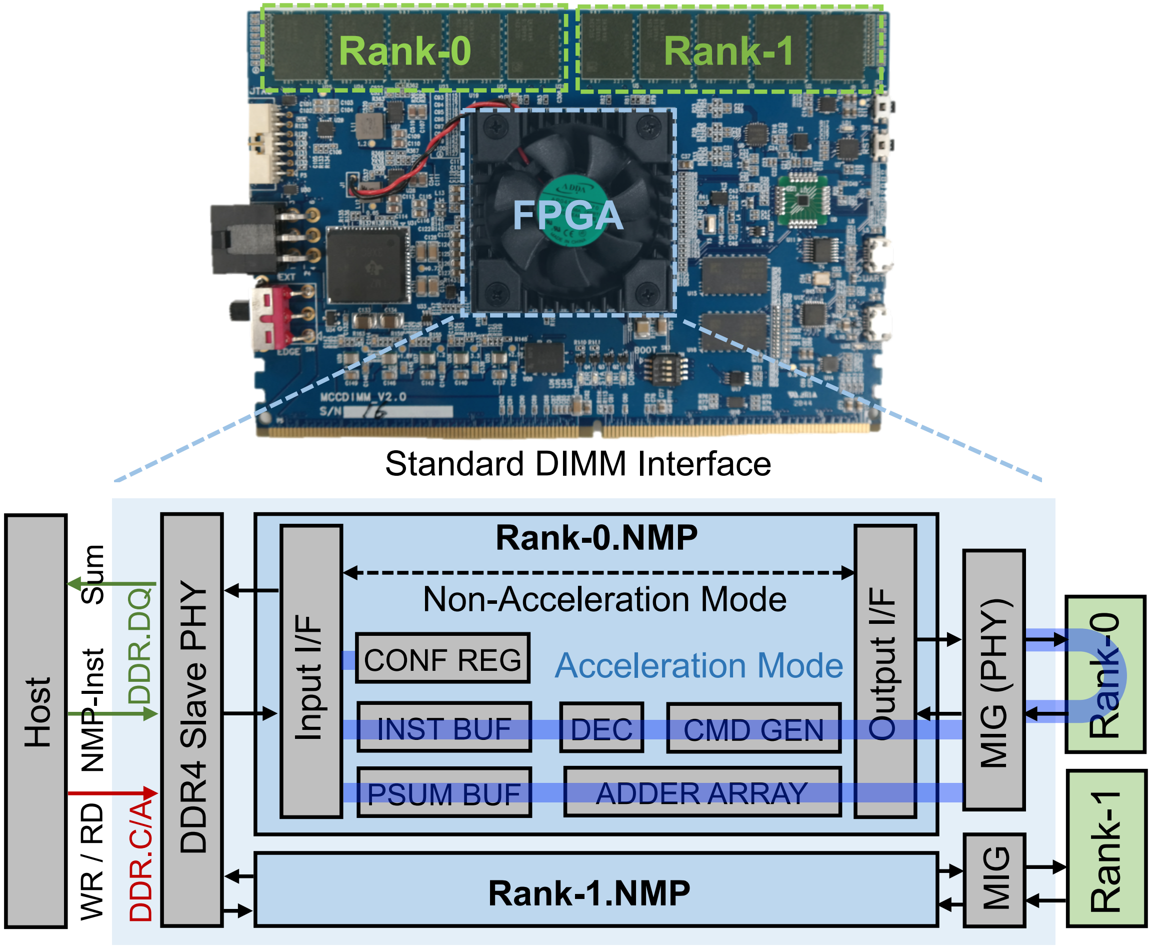
AxDIMM System





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

AxDIMM Design: Hardware Architecture



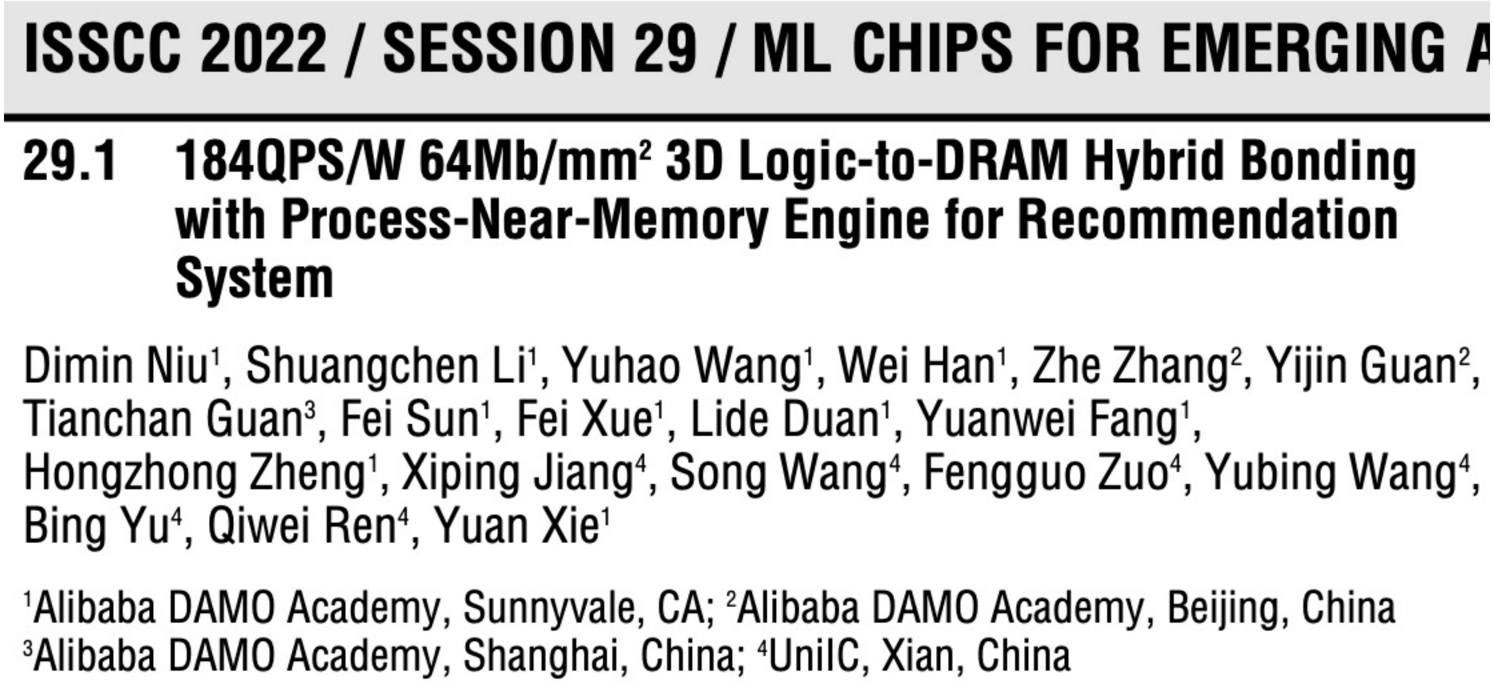
NMP-Inst (64 bits)

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

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

Alibaba 3D Logic-to-DRAM Hybrid Bonding with Processing-near-Memory Engine

Hybrid Bonding with PnM Engine (ISSCC 2022)



ISSCC 2022 / SESSION 29 / ML CHIPS FOR EMERGING A

29.1 184QPS/W 64Mb/mm^2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System

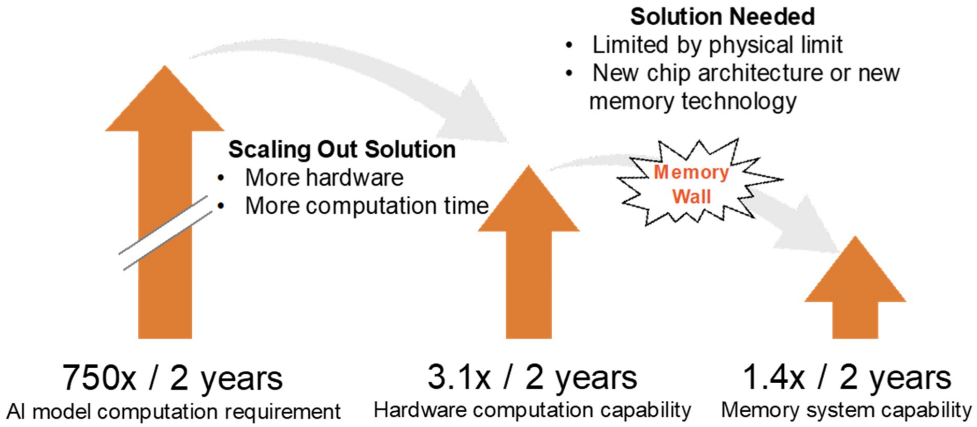
Dimin Niu[1], Shuangchen Li[1], Yuhao Wang[1], Wei Han[1], Zhe Zhang[2], Yijin Guan[2], Tianchan Guan[3], Fei Sun[1], Fei Xue[1], Lide Duan[1], Yuanwei Fang[1], Hongzhong Zheng[1], Xiping Jiang[4], Song Wang[4], Fengguo Zuo[4], Yubing Wang[4], Bing Yu[4], Qiwei Ren[4], Yuan Xie[1]

[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

Processing-in-Memory for Machine Learning

- Memory bandwidth is not enough for many ML workloads



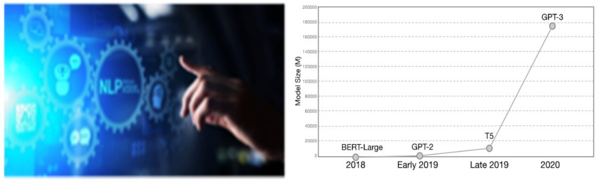
Solution Needed
- Limited by physical limit
- New chip architecture or new memory technology

Scaling Out Solution
- More hardware
- More computation time

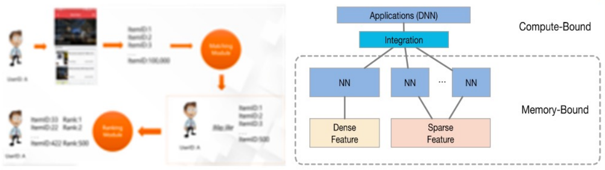
750x / 2 years
AI model computation requirement

3.1x / 2 years
Hardware computation capability

1.4x / 2 years
Memory system capability



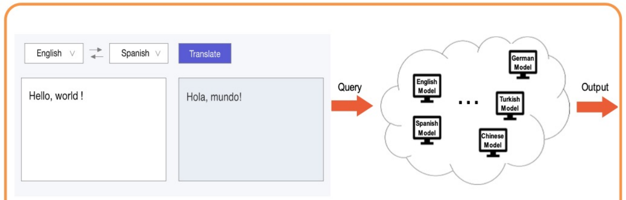
Natural Language Processing



Recommendation Systems

A 10TB Customer-Item Graph
Item
Customer
GNN
Features
Features
Vector ops Approximation
Graph Analytics - Fraud Detection
Scalar ops Accurate comp.

Graph Neural Network



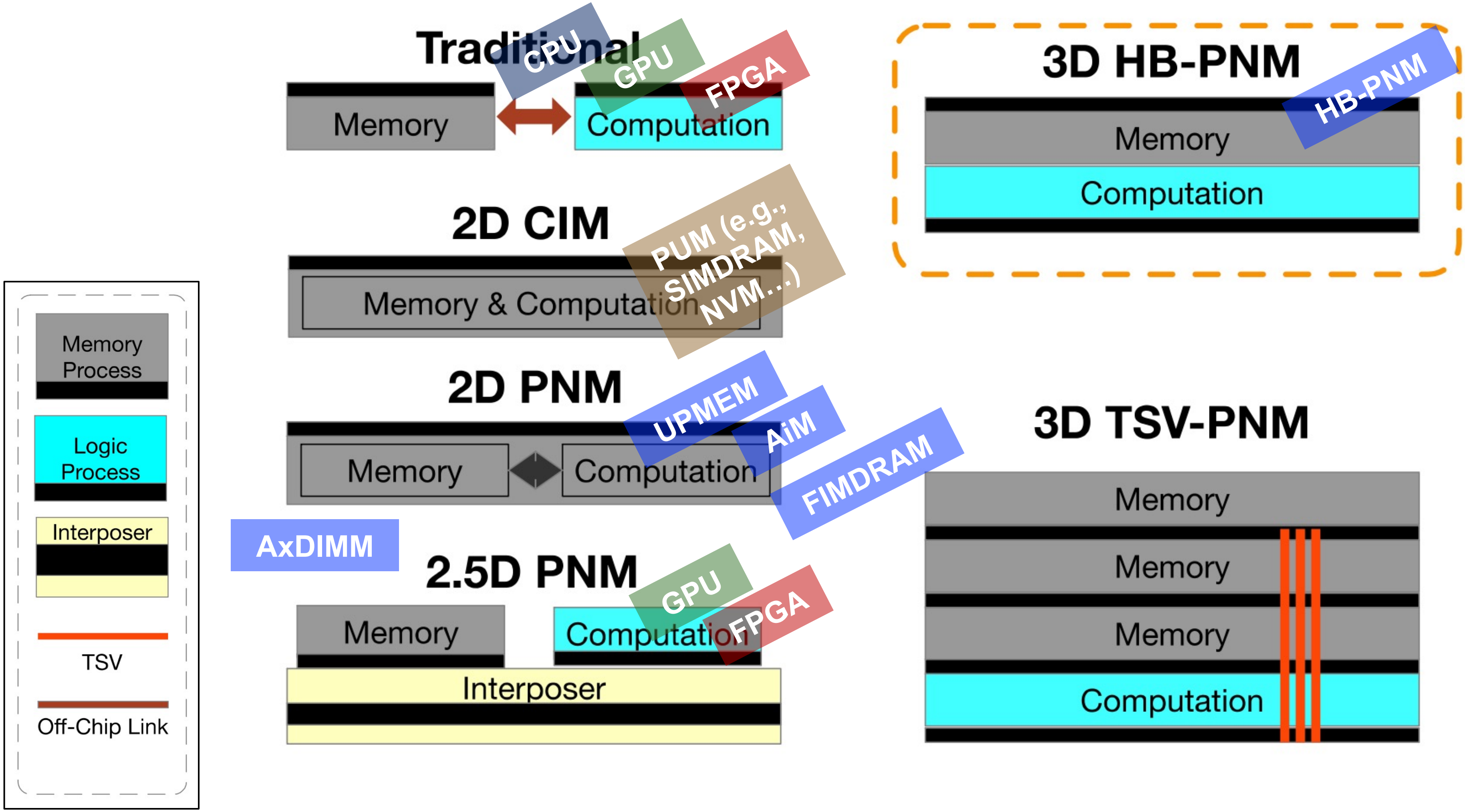
Hello, world !

Hola, mundo!

Multi-Task Online Inference

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Processing-in-Memory Classification

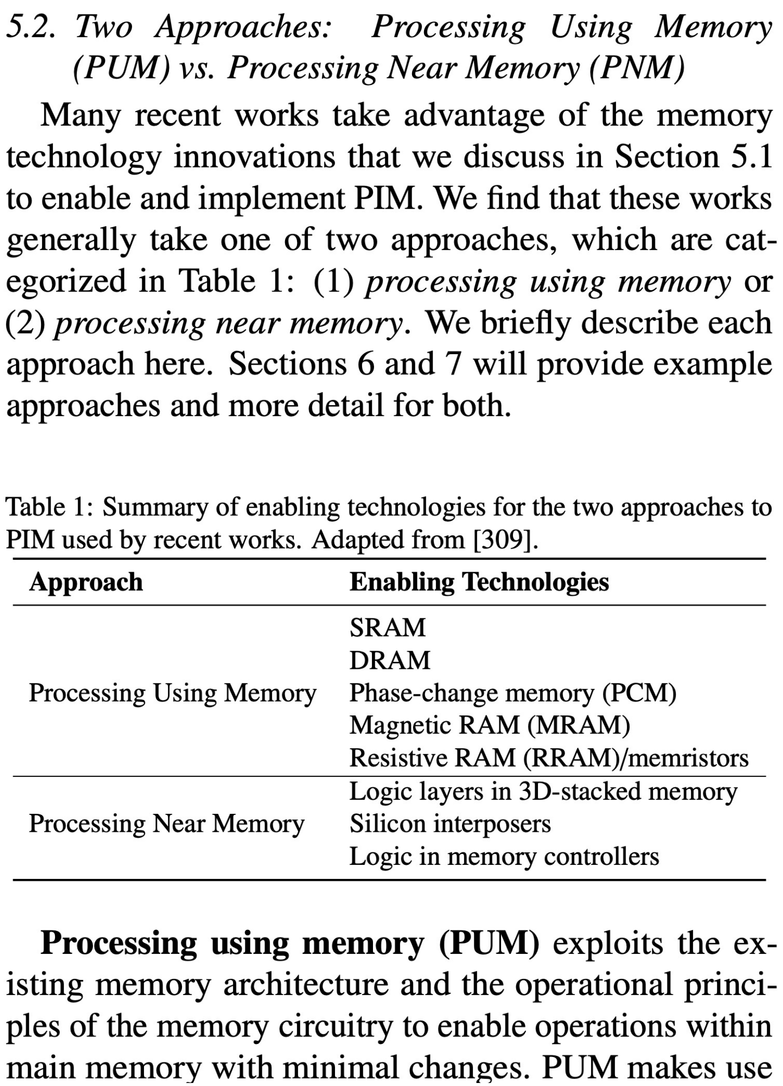


PNM : Process Near Memory
CIM: Compute In Memory

HB : Hybrid Bonding
TSV: Through-silicon Via

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

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].

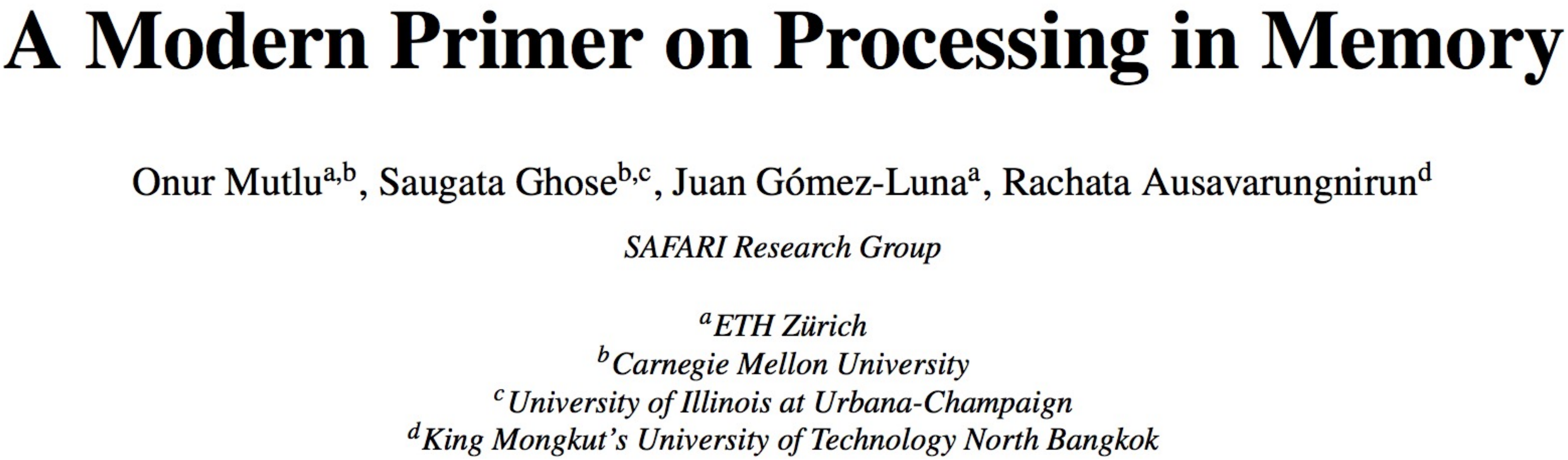
Approach	Enabling Technologies
Processing Using Memory	SRAM DRAM Phase-change memory (PCM) Magnetic RAM (MRAM) Resistive RAM (RRAM)/memristors
Processing Near Memory	Logic layers in 3D-stacked memory Silicon interposers Logic in memory controllers

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)]

https://people.inf.ethz.ch/omutlu/pub/ModernPrimerOnPIM_springer-emerging-computing-bookchapter21.pdf

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

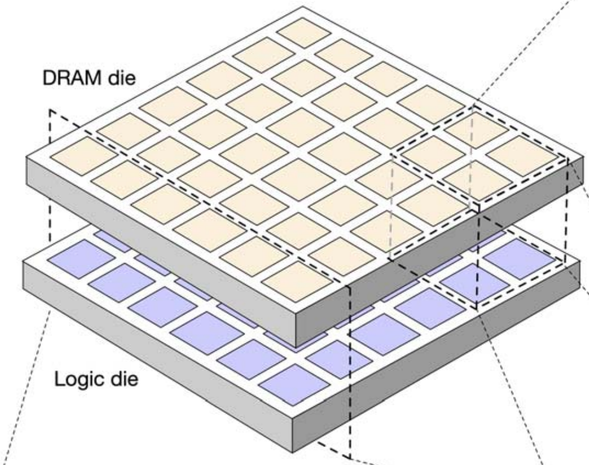
Onur Mutlu, Saugata Ghose, Juan Gomez-Luna, and Rachata Ausavarungnirun,
"A Modern Primer on Processing in Memory"
Invited Book Chapter in ***Emerging Computing: From Devices to Systems - Looking Beyond Moore and Von Neumann***, Springer, to be published in 2021.

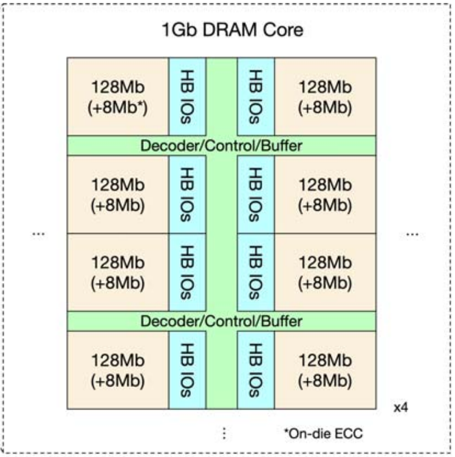
https://arxiv.org/pdf/1903.03988.pdf

HB-PNM: Overall Architecture (I)

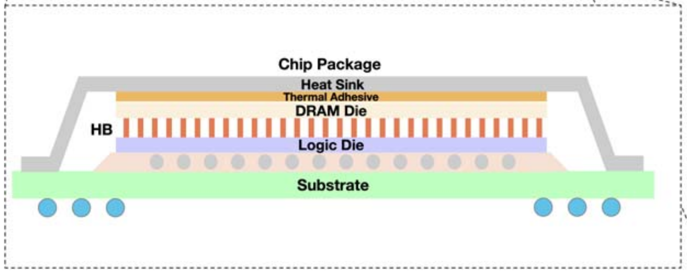
- 3D-stacked logic die and DRAM die vertically bonded by hybrid bonding (HB)

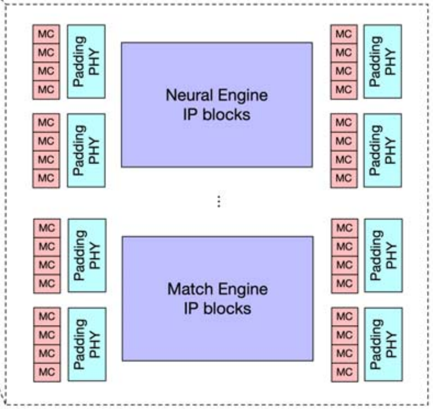
3D-stacked illustration of the DRAM die and logic die





DRAM array layout illustration and its imposed design constraints on logic die





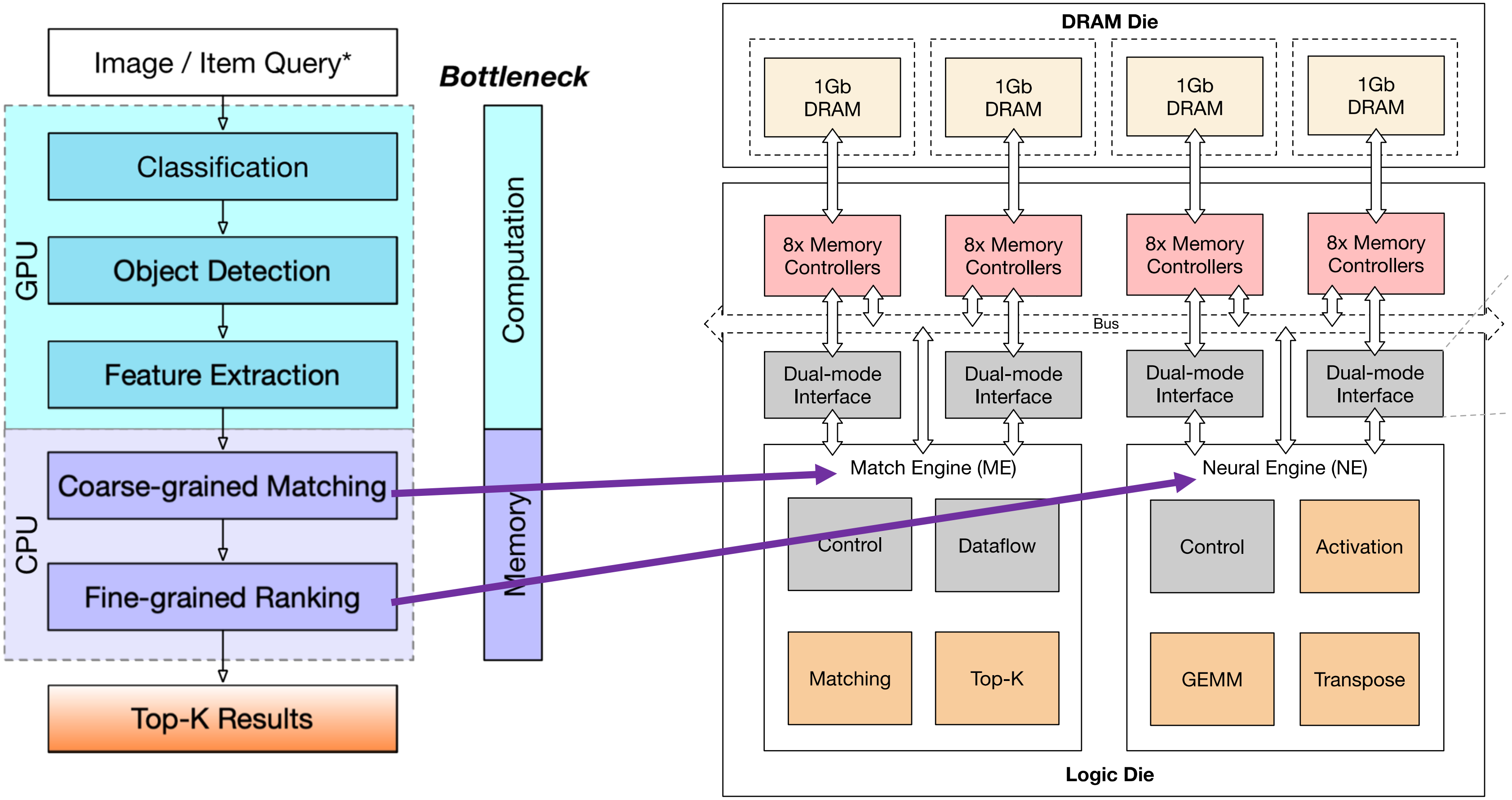
Cross-section illustration of the logic die and DRAM die vertically bonded by HB in a chip package

Logic die physical constraints due to hybrid bonding PHY and MC

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM: Overall Architecture (II)

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

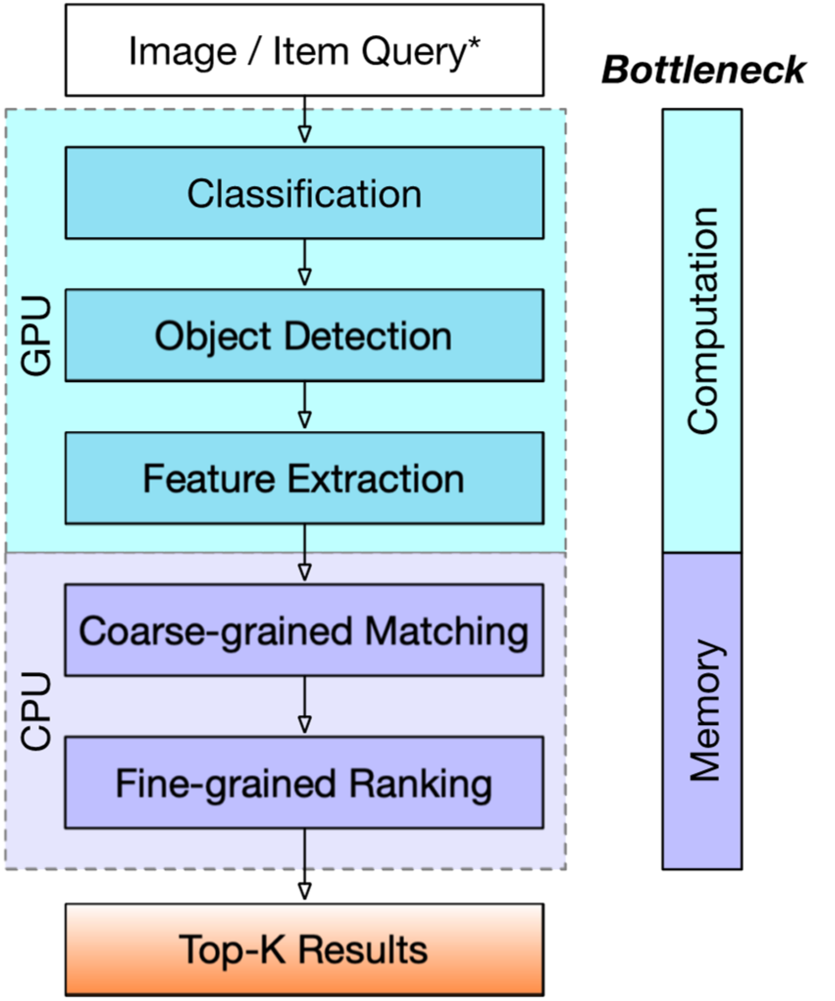


Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Recommendation Systems

Feature Generation + Matching & Ranking

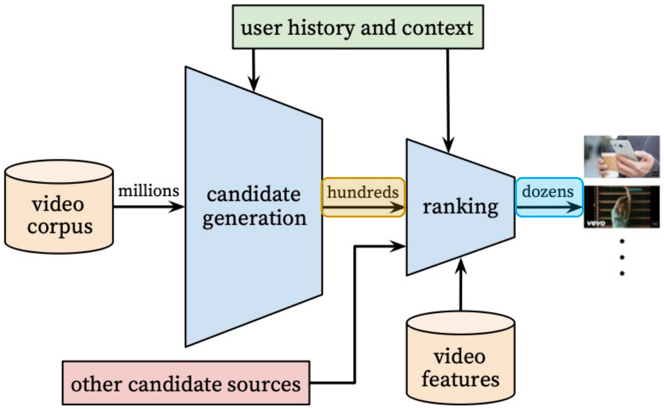
- Recommendation system
  - Feature generation
    - Classification, object detection, feature extraction
    - Compute-bound
    - Good fit for GPU
  - Matching and ranking
    - Coarse-grained matching, fine-grained ranking
    - Memory-bound
      - Most latency (89.87%) and energy (82.97%)
      - Typically run on CPU



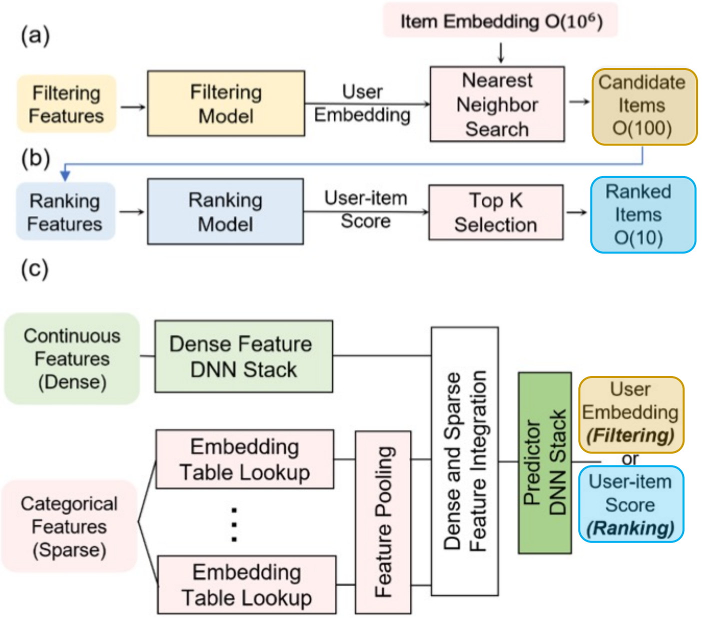
Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Recommendation Systems

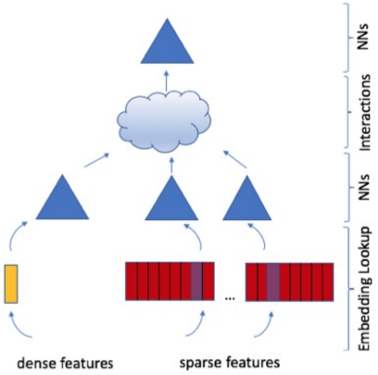
- Candidate recommendations are retrieved and then ranked



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



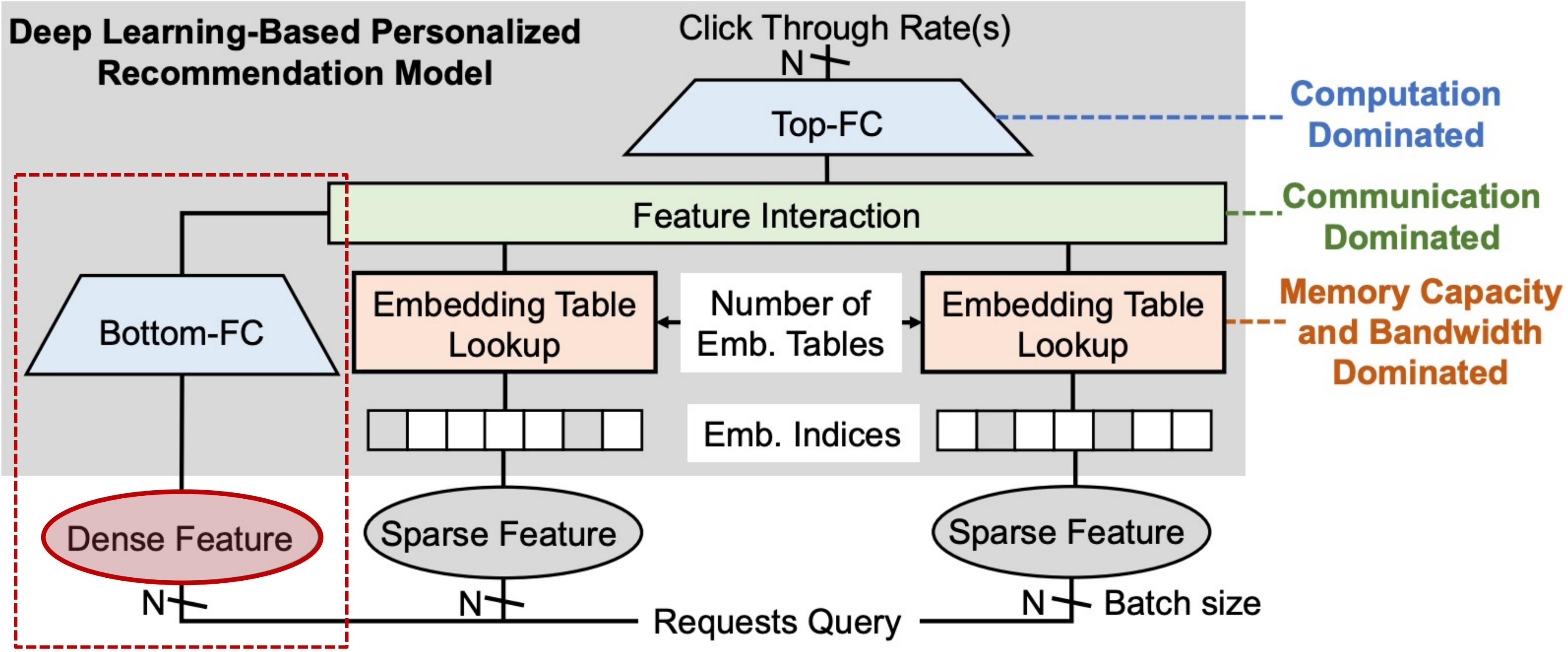
Li et al., iMARS: An In-Memory-Computing Architecture for Recommendation Systems, arXiv:2202.09433, 2022



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

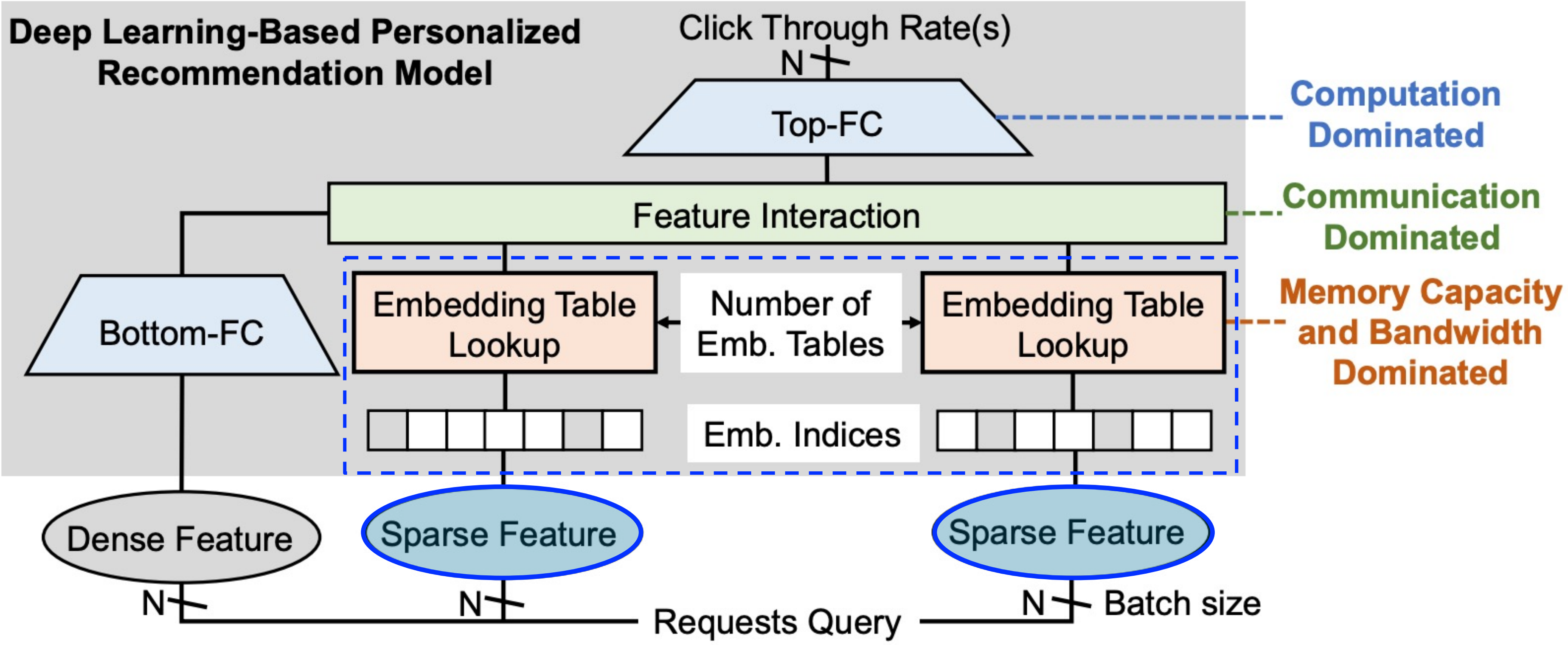


Dense features: continuous inputs in vectors and matrices are processed by typical DNN layers (e.g., fully connected layers)

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

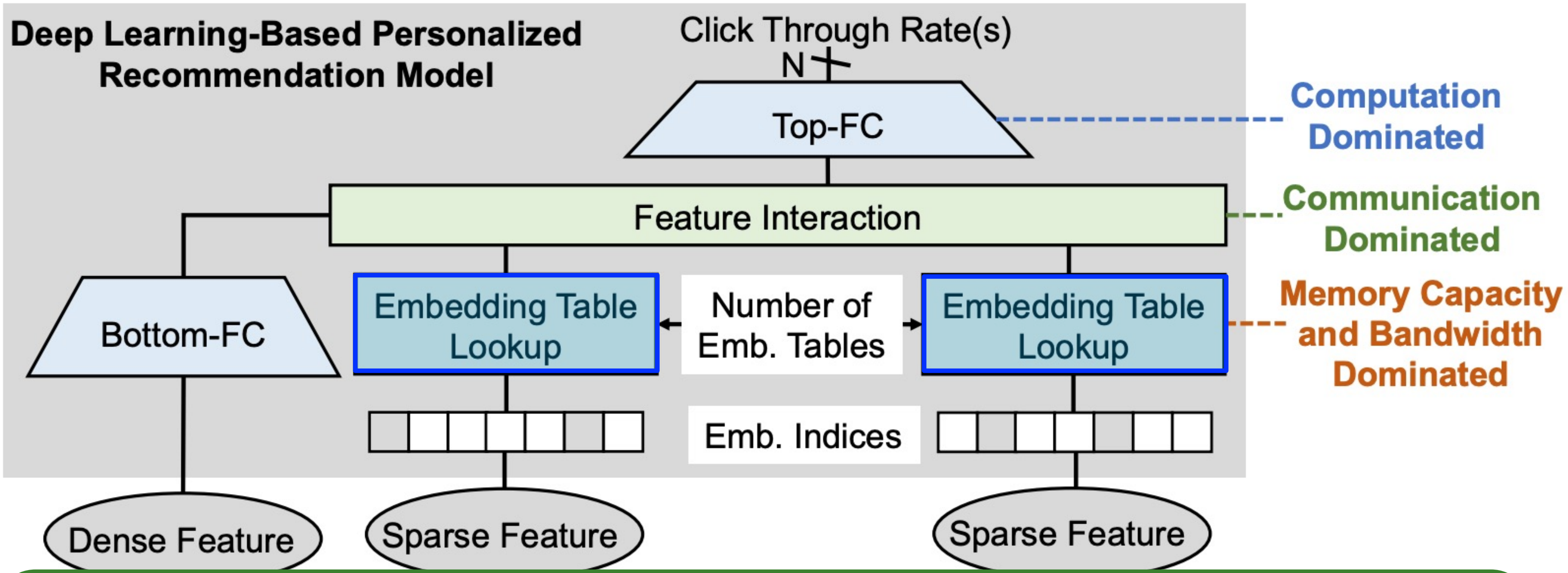


Sparse features: for categorical inputs;
processed by indexing large embedding tables

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

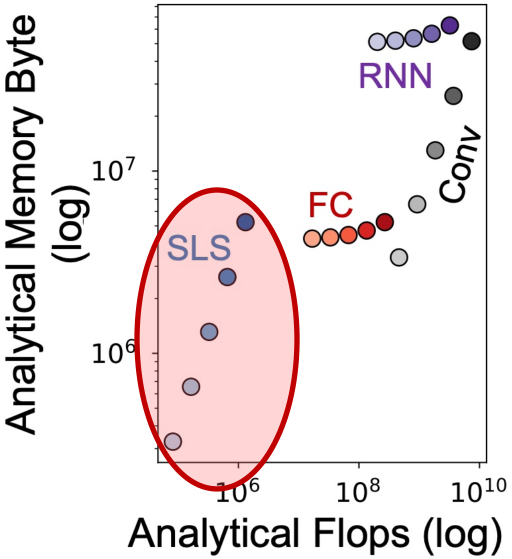


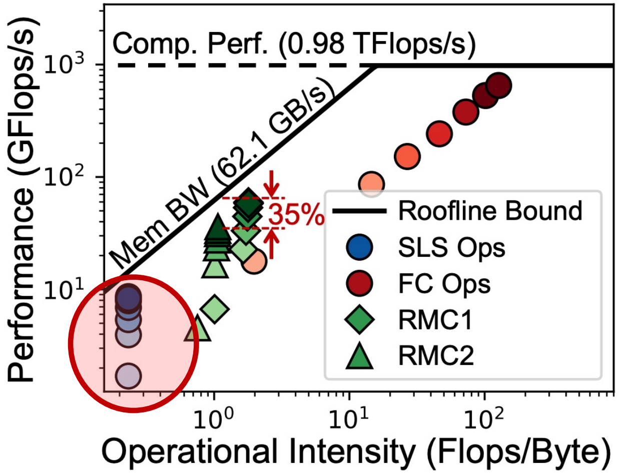
Embedding tables are organized as a set of potentially millions of vectors: lookup and pooling operations represent sparse features learned during training and generally exhibit Gather-Reduce pattern, via Caffe2's SparseLengths (SLS) operators

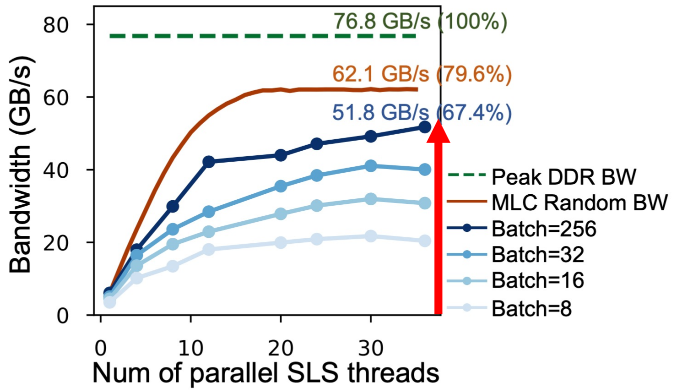
Ke et al. "RecNMP: Accelerating personalized recommendation with near-memory processing," ISCA 2020

DLRM Performance Characterization

- Identifying key performance bottlenecks for the DLRM system







SparseLengths (SLS) operators:

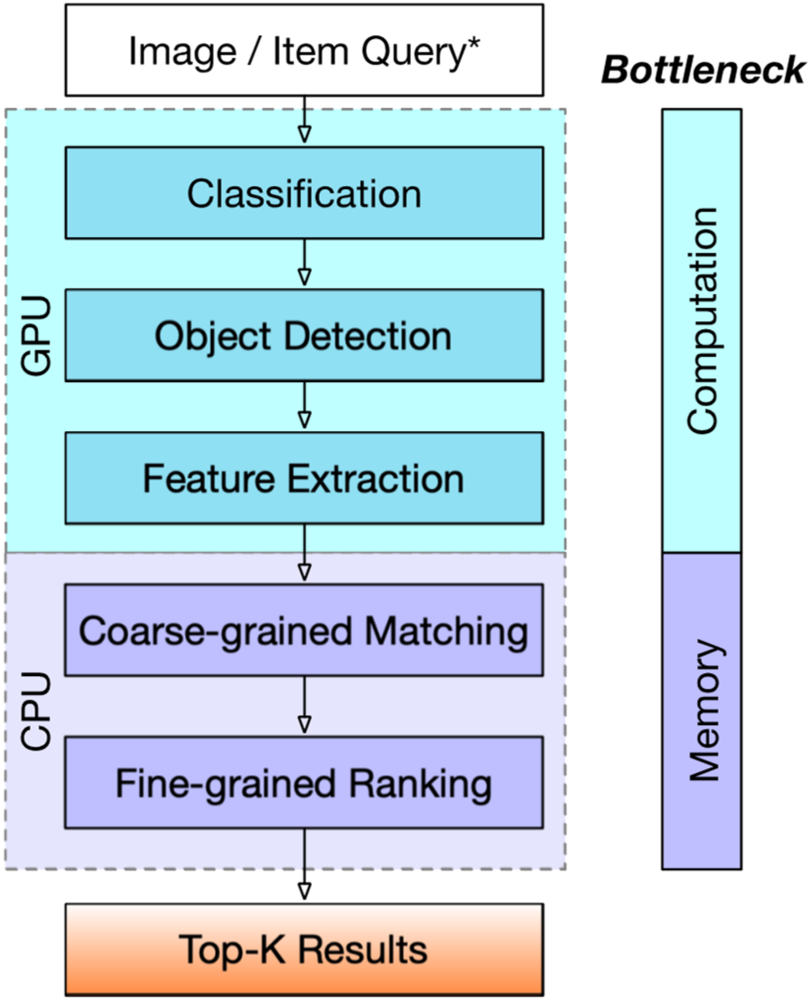
- Low FP intensity
- Larger batch size:
  - Higher memory footprint
  - Higher memory intensity

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

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

Feature Generation + Matching & Ranking

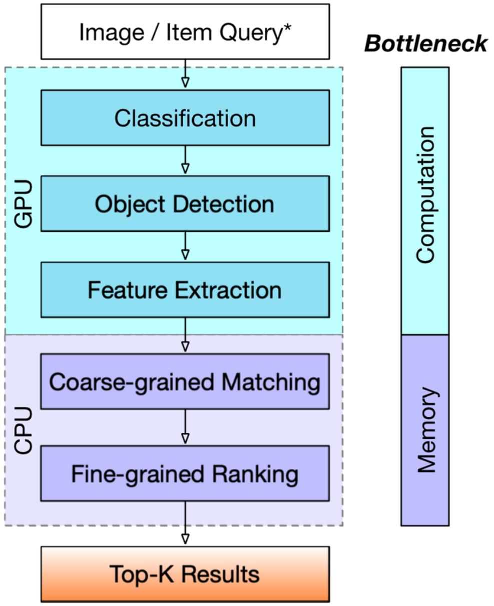
- Recommendation system
  - Feature generation
    - Classification, object detection, feature extraction
    - Compute-bound
    - Good fit for GPU
  - Matching and ranking
    - Coarse-grained matching, fine-grained ranking
    - Memory-bound
      - Most latency (89.87%) and energy (82.97%)
      - Typically run on CPU

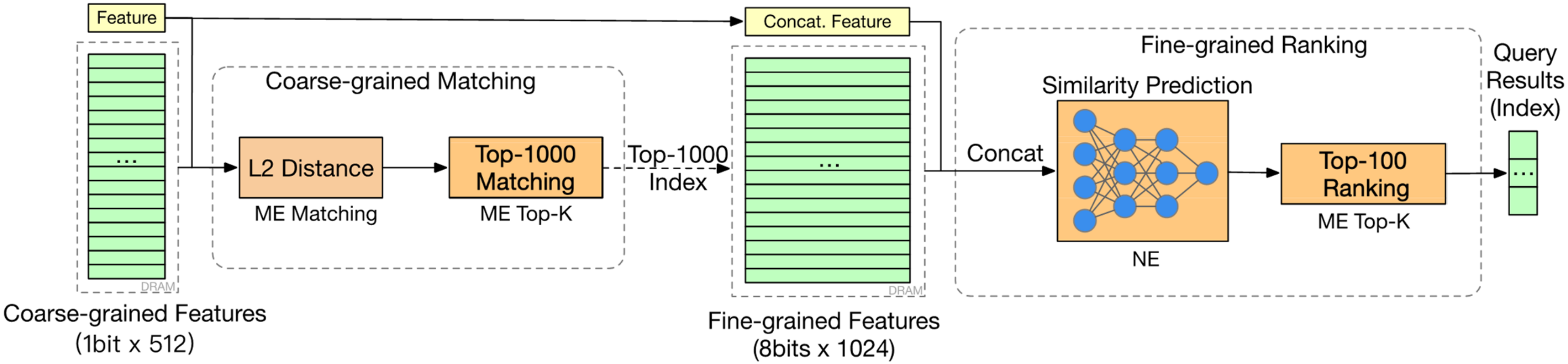


Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Matching & Ranking

- Coarse-grained matching
  - Binary feature vectors with 512 dimensions
  - Distance calculation
  - Top-1000 items selected from 40K items
- Fine-grained ranking
  - Features with 8 bits x 1024 dimensions
  - 3-layer MLP (2048-256-64-1) for similarity prediction
  - Top-100 ranking results from 1K items



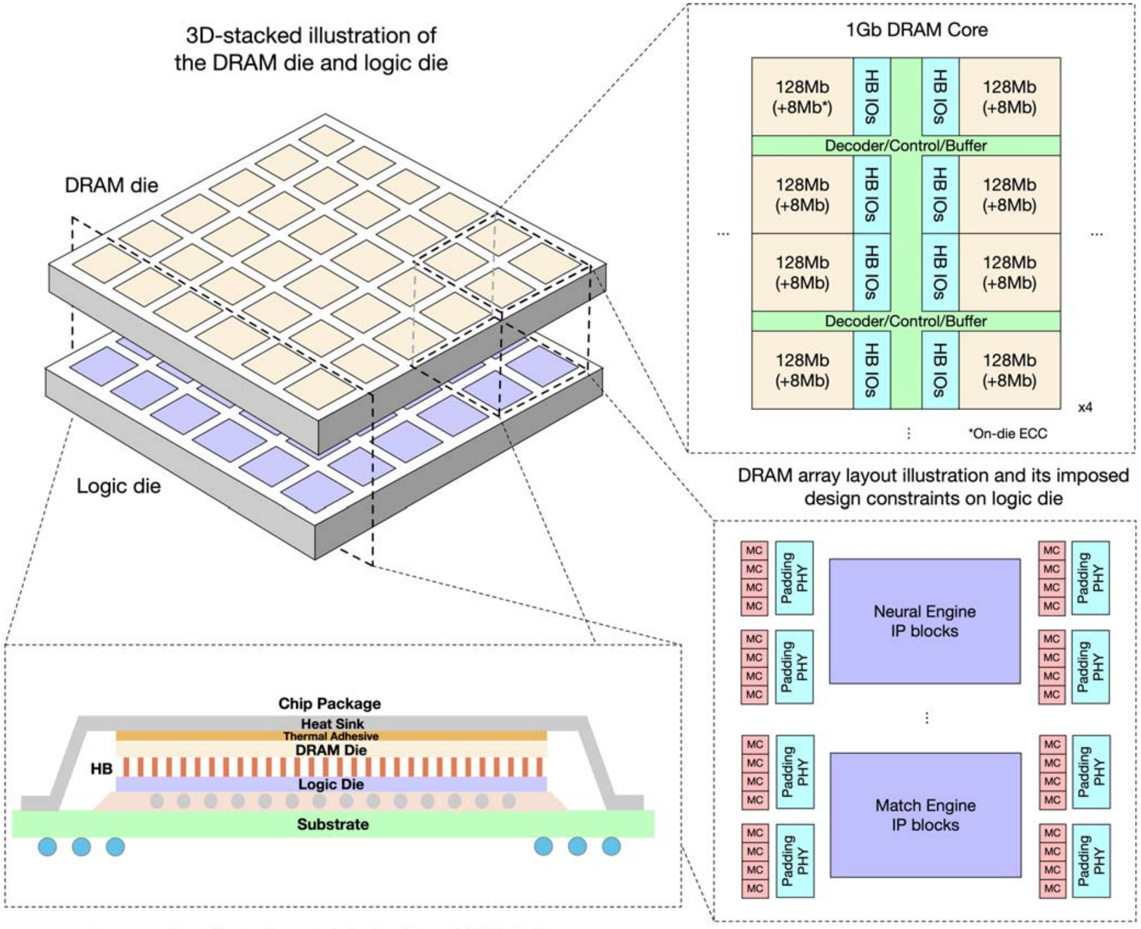


Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

3D Logic-to-DRAM Hybrid Bonding

HB-PNM: Overall Architecture (I)

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



3D-stacked illustration of the DRAM die and logic die

DRAM array layout illustration and its imposed design constraints on logic die

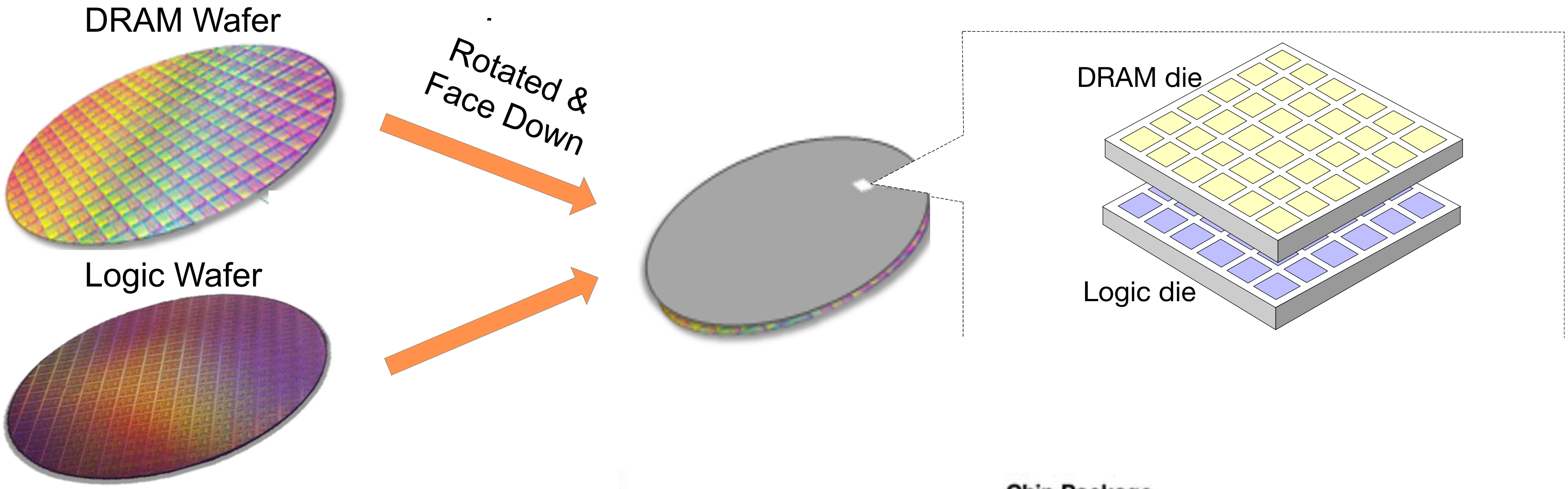
Cross-section illustration of the logic die and DRAM die vertically bonded by HB in a chip package

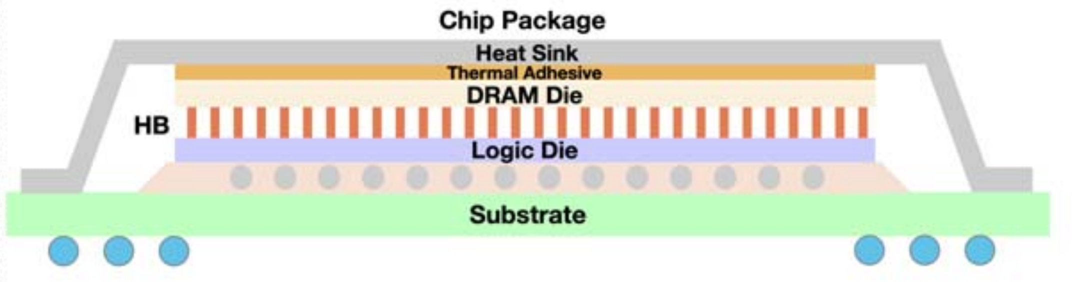
Logic die physical constraints due to hybrid bonding PHY and MC

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM: Chip Implementation

- 3D Logic-to-DRAM Hybrid Bonding
  - Face-to-face hybrid wafer bonding



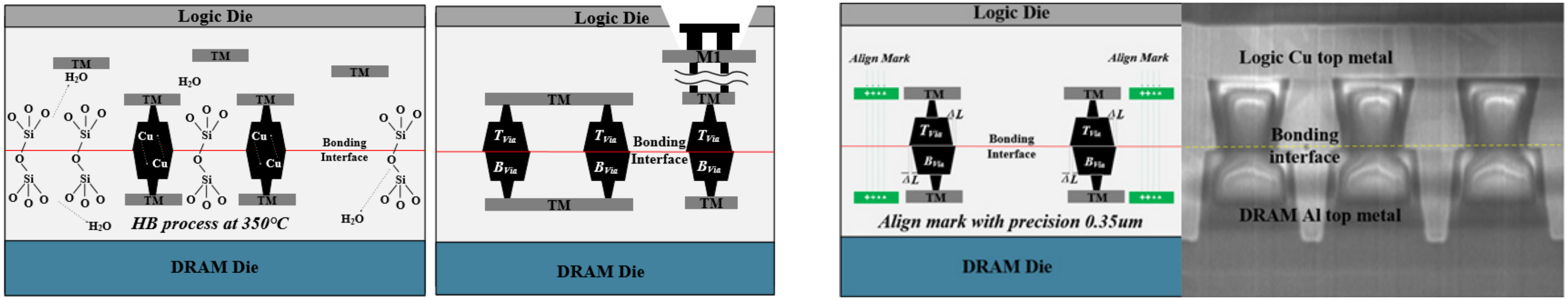


Cross-section illustration

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM: Hybrid-Bonding Interconnection

- 3D Logic-to-DRAM Hybrid Bonding
  - Face-to-face hybrid wafer bonding
    - Logic and memory manufactured independently: this avoids the challenges of integrating logic into memory chips
  - Cu-Cu direct fusion with low bonding temperature (<350ºC)
  - Much denser vias than other 3D-stacking technologies
    - Low pitch size (3 um) vs. HBM microbumps (35 um[1])
    - High inter-layer bandwidth (1.38 TB/s) vs. HBM2E (460 GB/s[2])

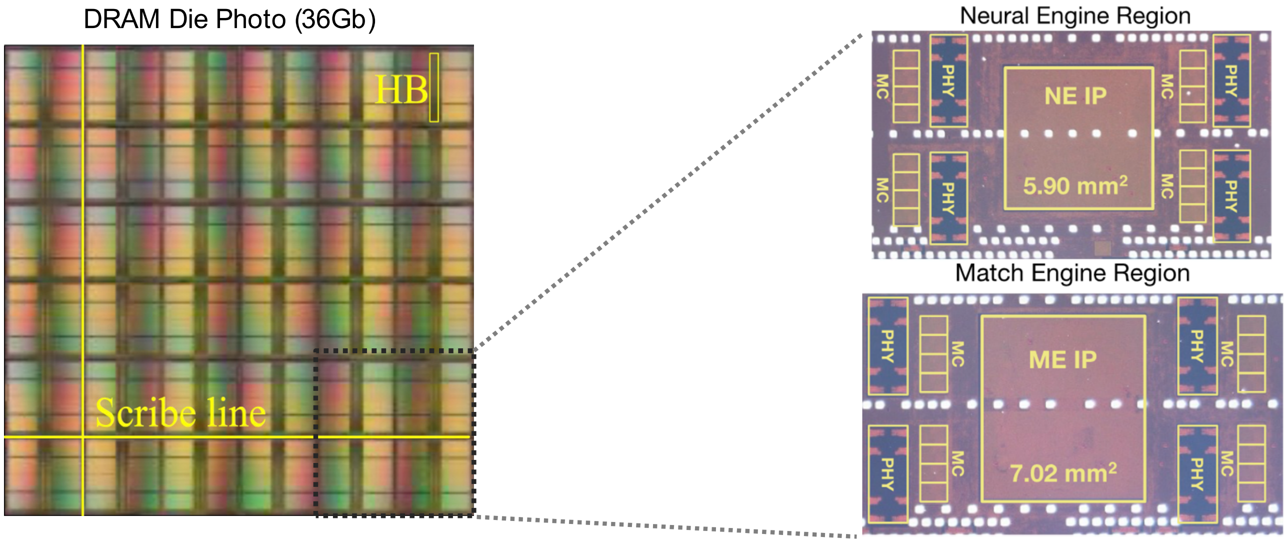


[1] Kim et al. Signal Integrity and Computing Performance Analysis of a Processing-In-Memory of High Bandwidth Memory (PIM-HBM) Scheme, IEEE TCPMT, 2021
[2] https://product.skhynix.com/products/dram/hbm/hbm2e.go

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM: DRAM Die and Logic Die

- DRAM die and logic die



DRAM Die		
Technology	25nm	
Area	Total*	602.22 mm^2
	Neural Engine	32 mm^2
	Match Engine	32 mm^2
Voltage	1.1 V	
Frequency (max)	150 MHz	
Power	300 mW per 1Gb	
Bandwidth**	153.60 GB/s / 1.38 TB/s	

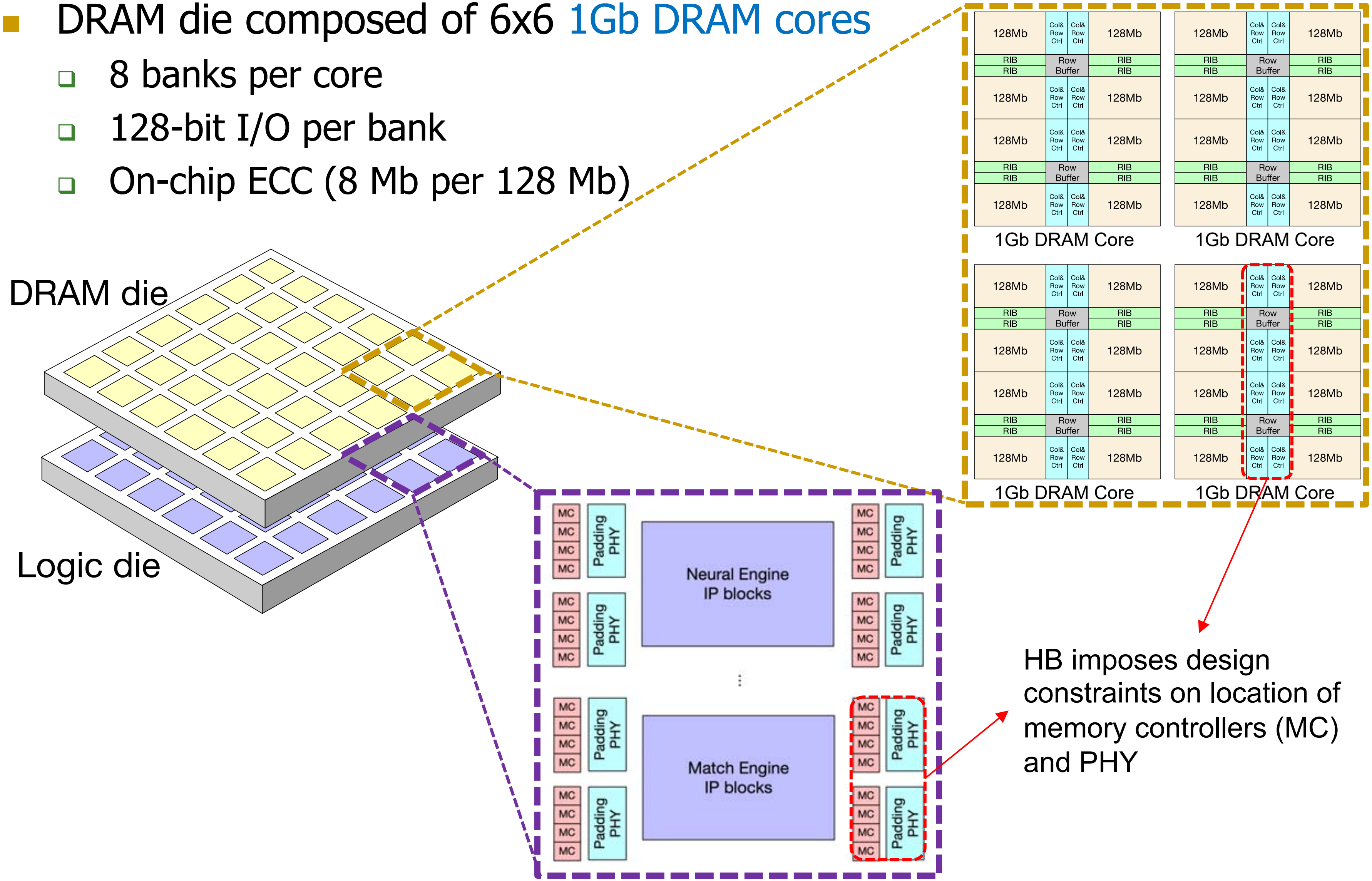
Logic Die		
Technology	55nm	
Area	Total*	602.22 mm^2
	Neural Engine	5.90 mm^2
	Match Engine	7.02 mm^2
# of MC	16 per IP	
Voltage	1.2 V	
Frequency	300 MHz	
Power	977.70 mW	
Precision	INT8	

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM Architecture

HB-PNM Architecture

- DRAM die composed of 6x6 1Gb DRAM cores
  - 8 banks per core
  - 128-bit I/O per bank
  - On-chip ECC (8 Mb per 128 Mb)

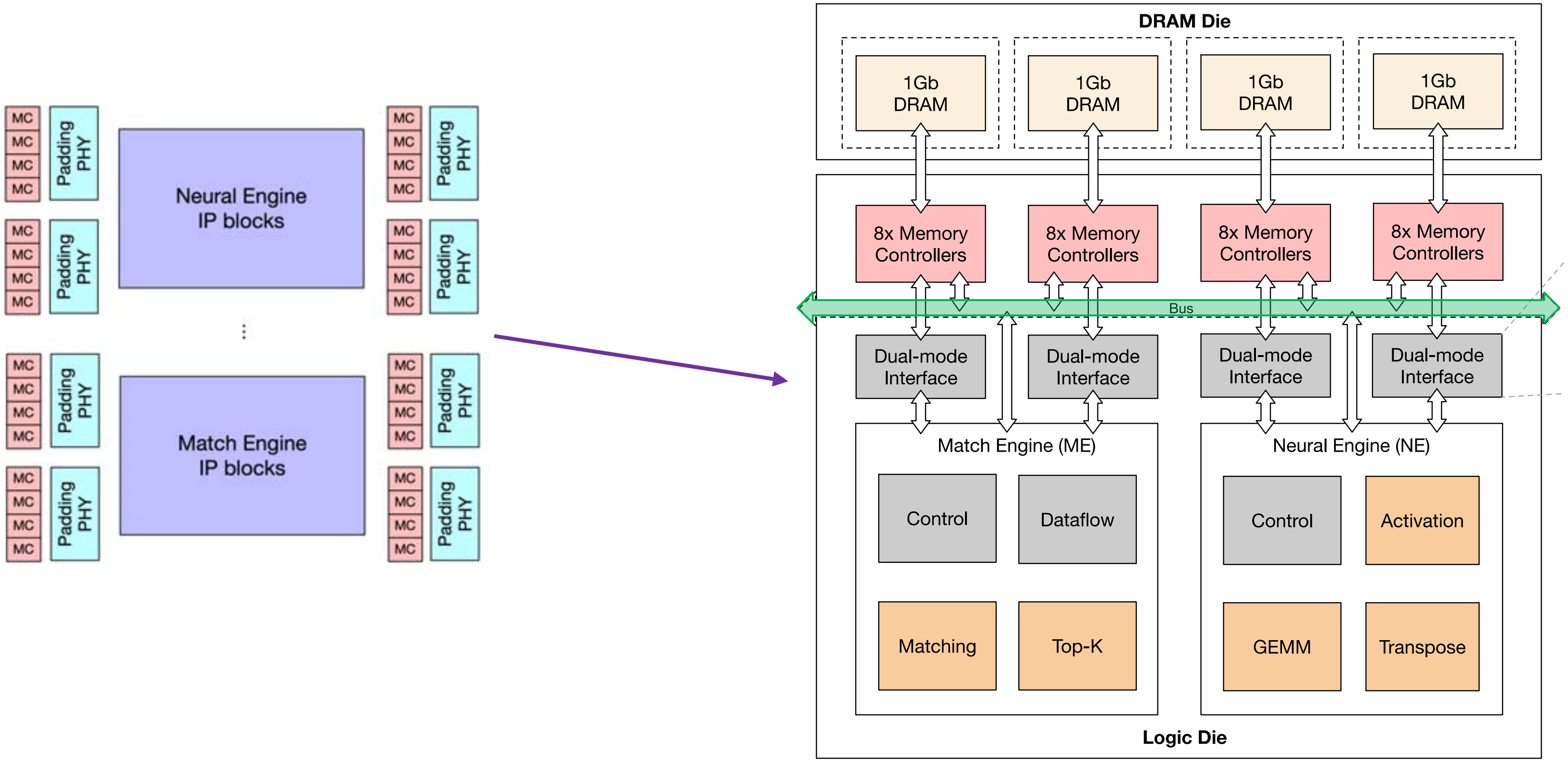


HB imposes design constraints on location of memory controllers (MC) and PHY

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM: Logic Die

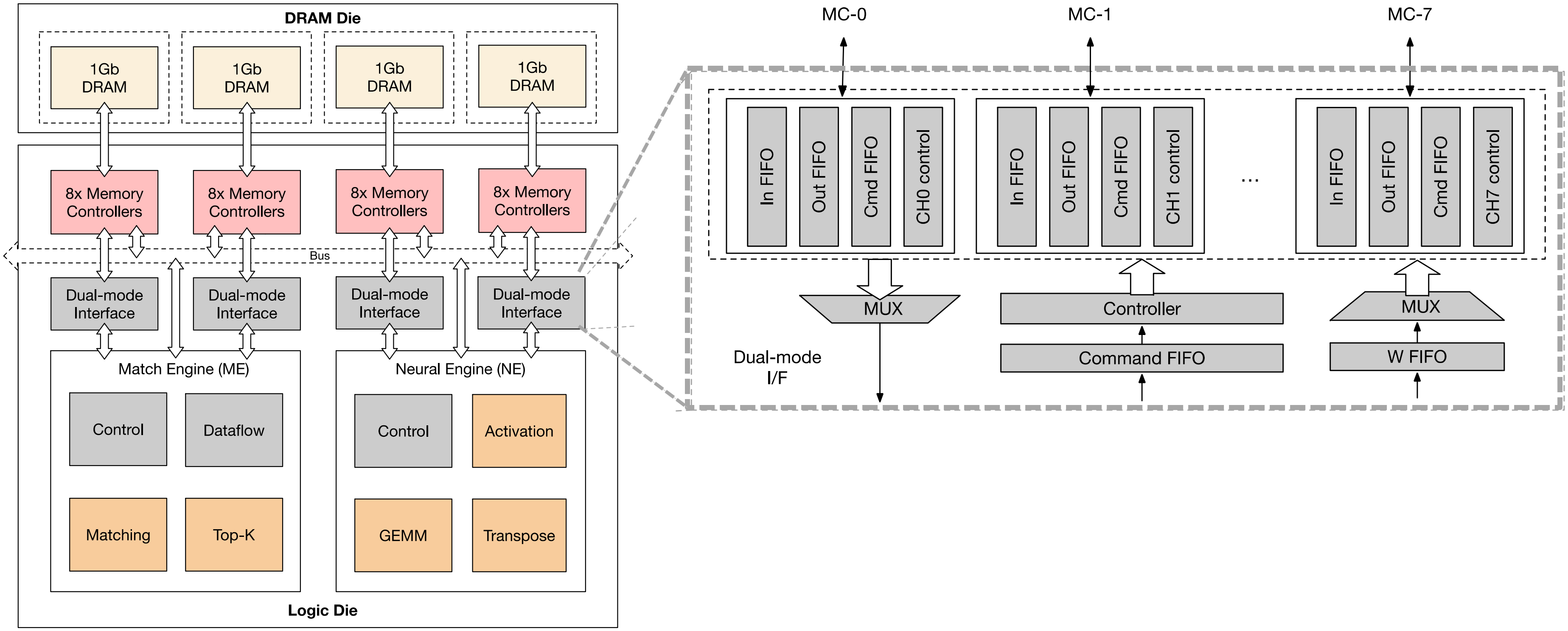
- Match engine and neural engine for matching and ranking in a recommendation system
  - Direct access to their counterpart DRAM blocks
  - Access to other DRAM blocks via on-chip bus



Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM Logic Die: Dual-mode Interface

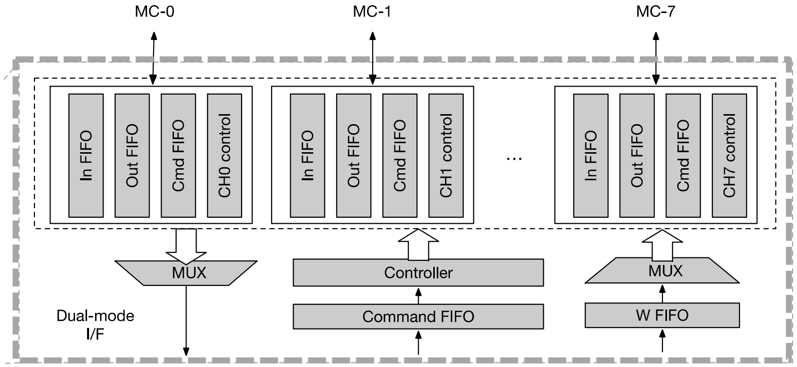
- Dual-mode interface can switch between
  - All 8 banks in lock-step for full bandwidth
  - Single channel (1 of 8 banks)

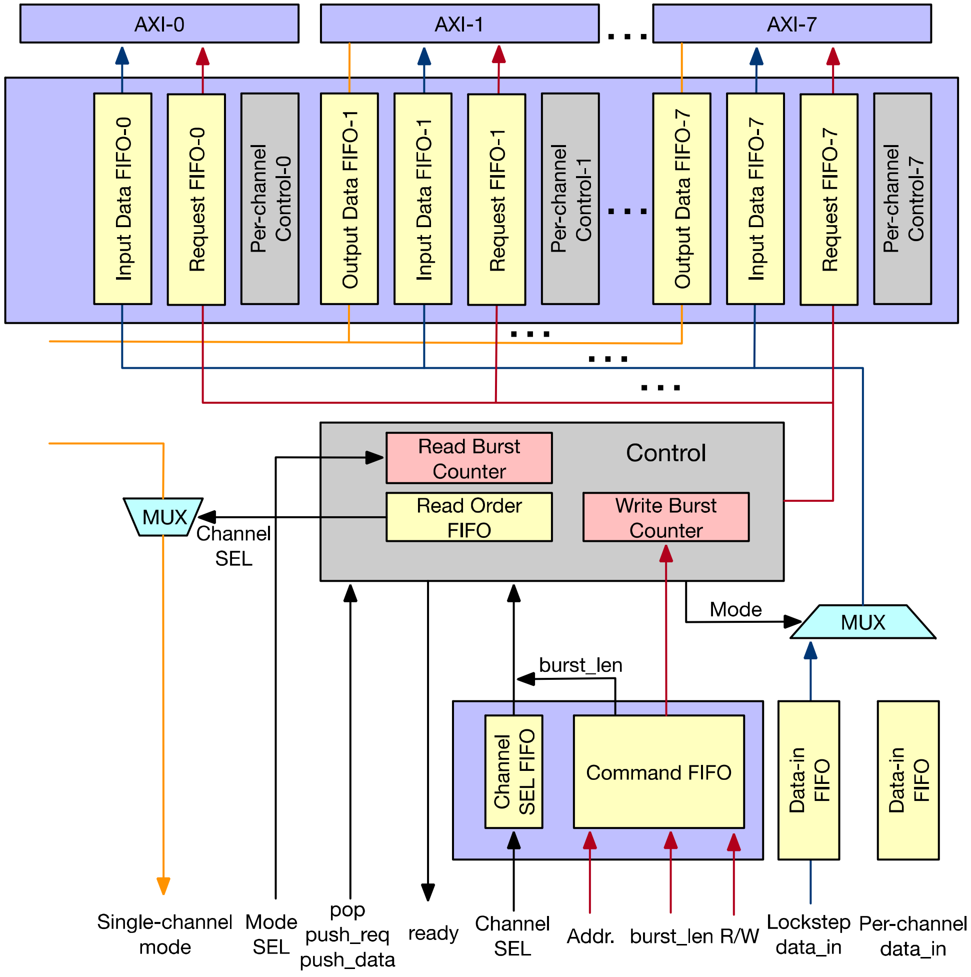


Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Neural Engine: Interface Bridge

- Support for single-channel mode and lockstep mode
- Read/write counters to support burst requests

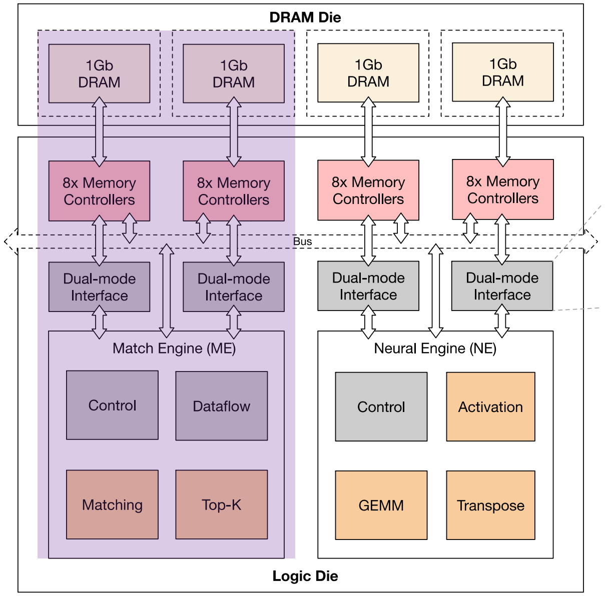


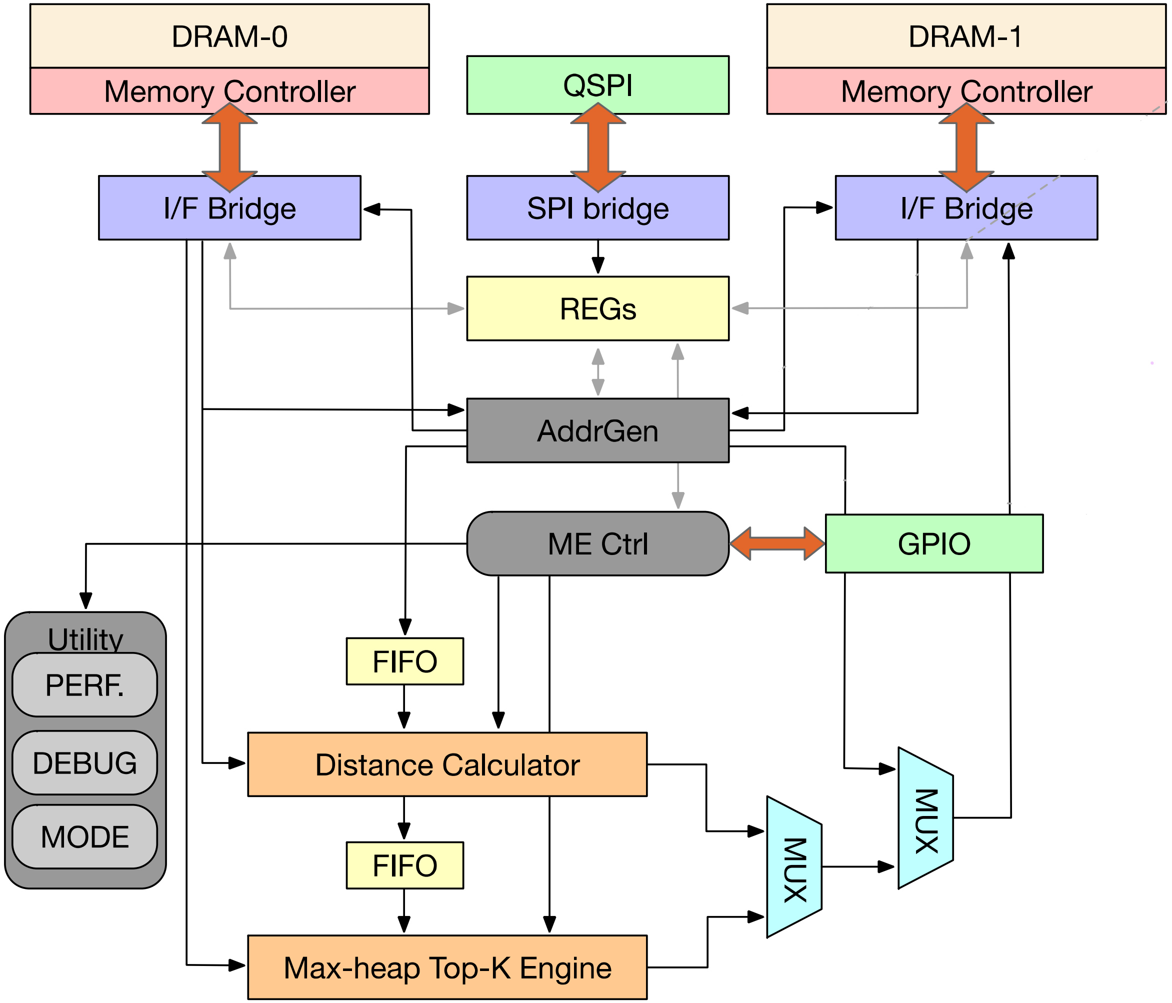


Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM: Match Engine

- Responsible for coarse-grained matching

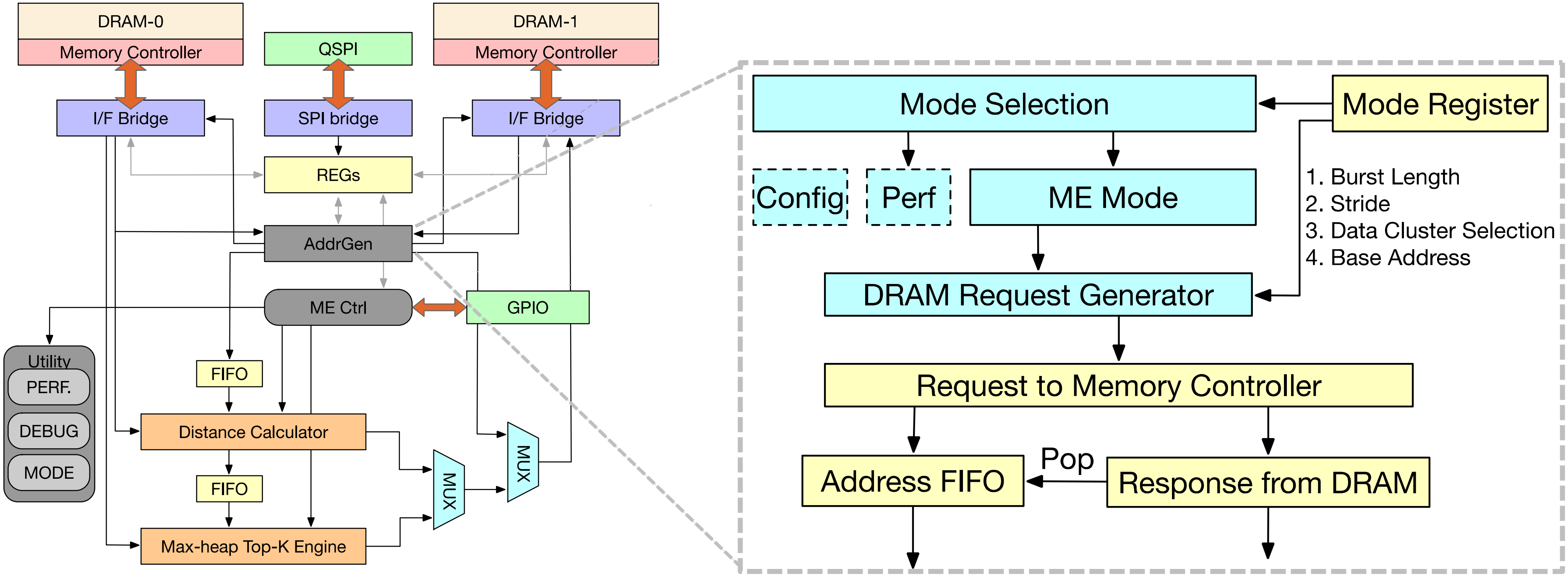




Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Match Engine: Address Generator

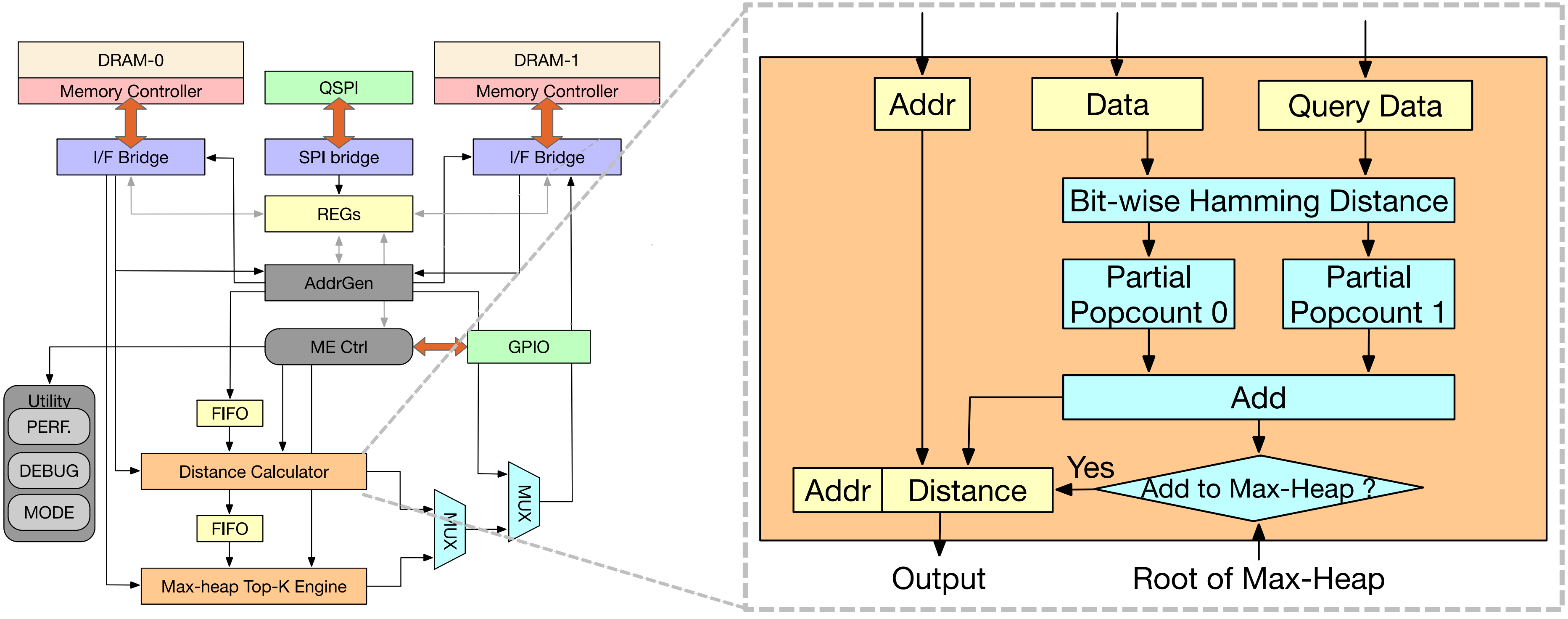
- AddGen generates the address of the input query
  - Mode selection for different access patterns
  - Configurable via registers



Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Match Engine: Distance Calculator

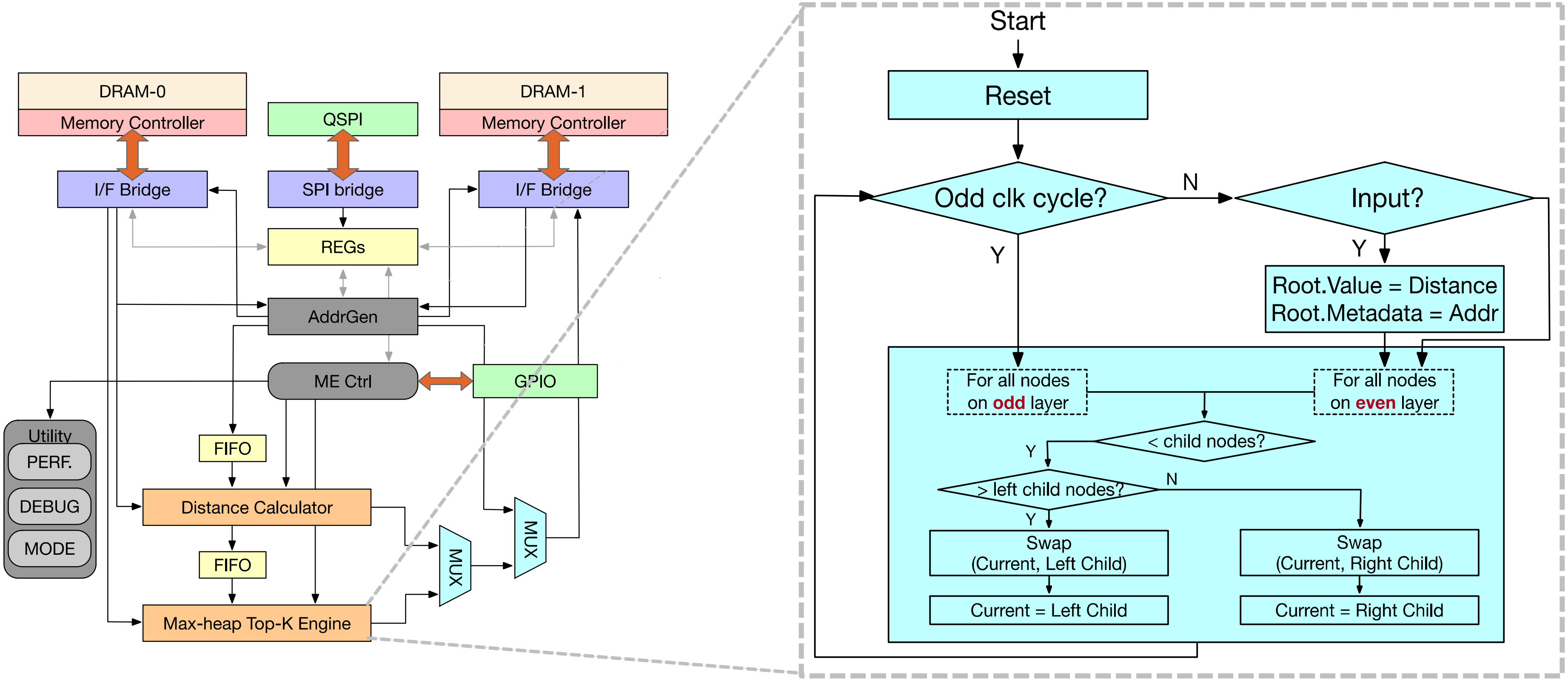
- Distance calculator obtains similarity between input query and feature vectors
  - It computes Hamming distance of two 512-bit vectors
  - Distance is filtered by root of max-heap



Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Match Engine: Top-K Engine (I)

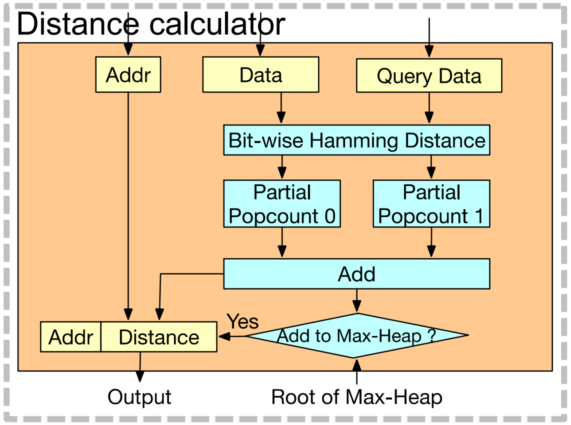
- Max-heap hardware block and data structure with 1000 nodes <address, distance> for the 1000 shortest distances
  - New input every two cycles

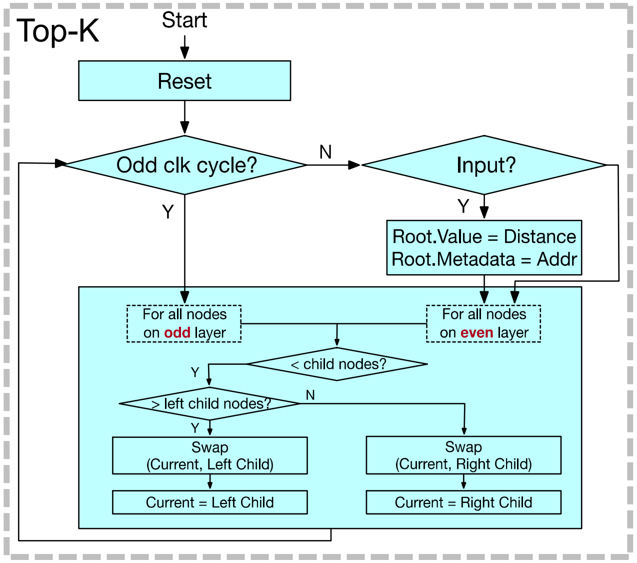


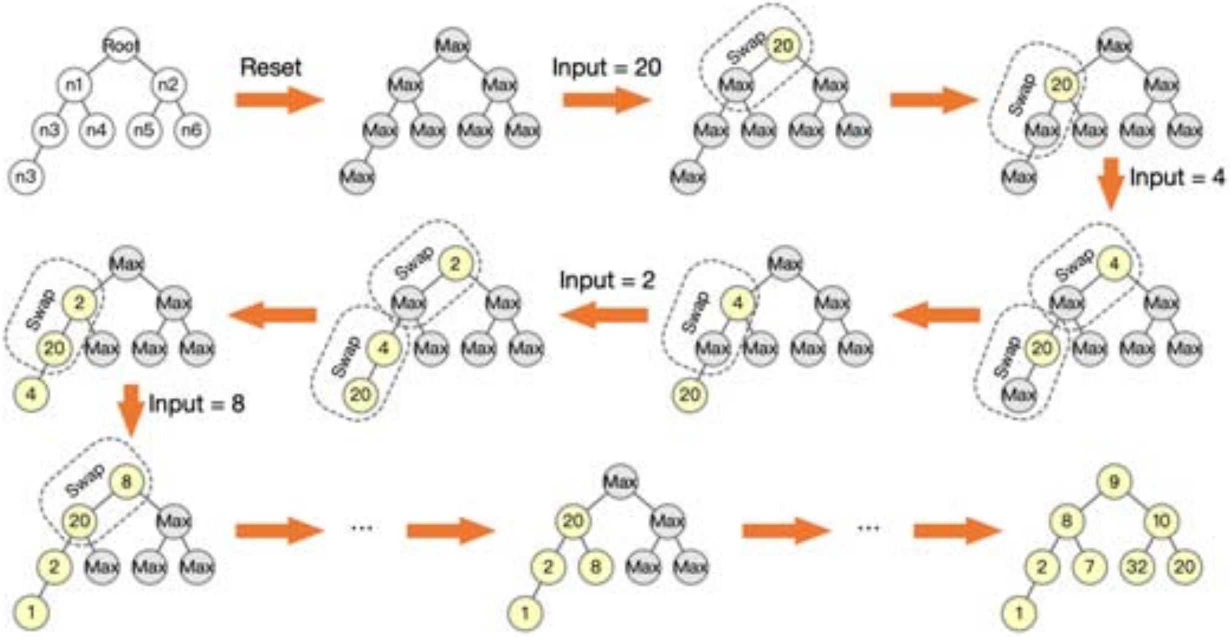
Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Match Engine: Top-K Engine (II)

- Max-heap hardware block and data structure with 1000 nodes <address, distance> for the 1000 shortest distances





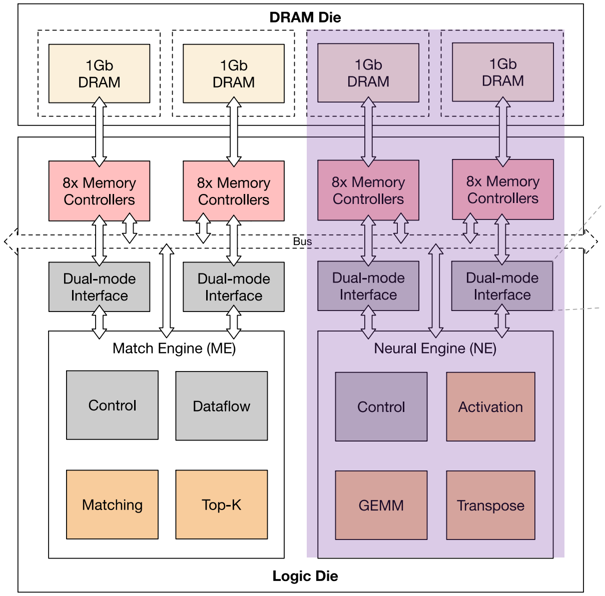


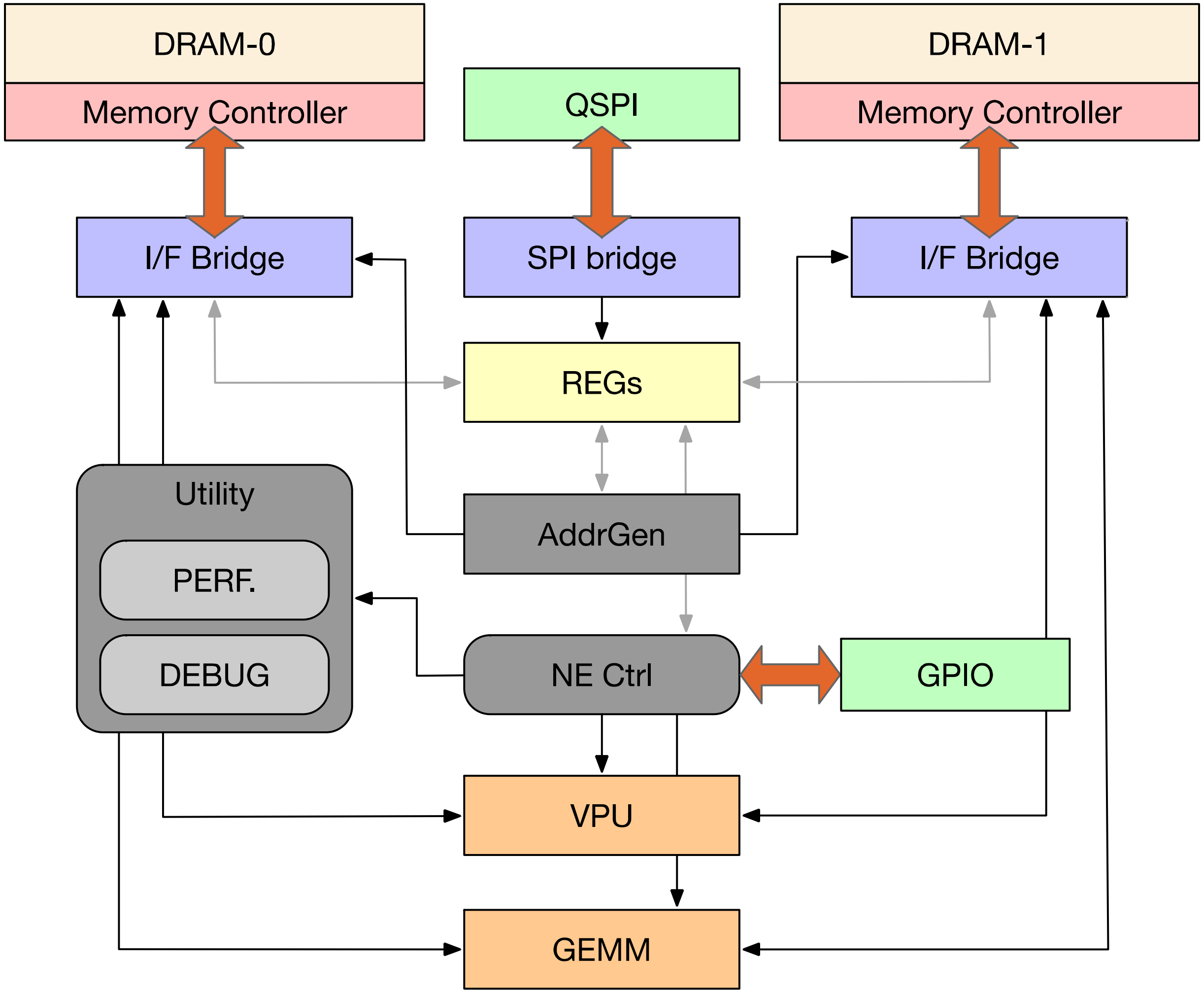
If Distance < left child, swap left child;
Else, swap right child

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM: Neural Engine

- Responsible for similarity prediction for fine-grained ranking

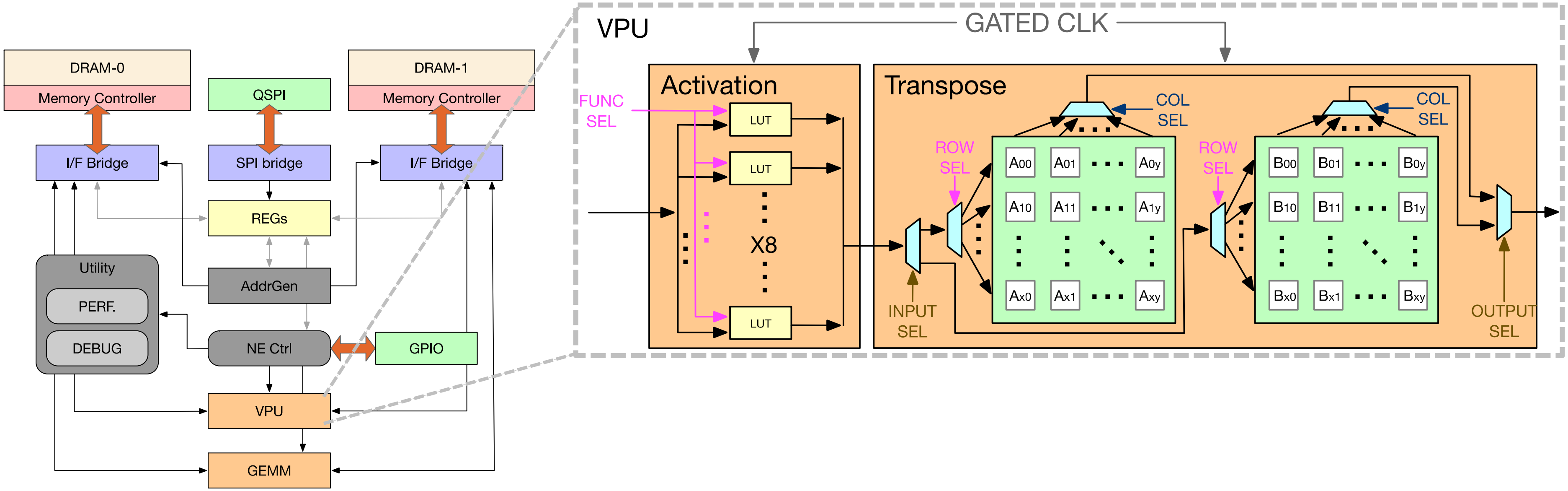




Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Neural Engine: Vector Processing Unit

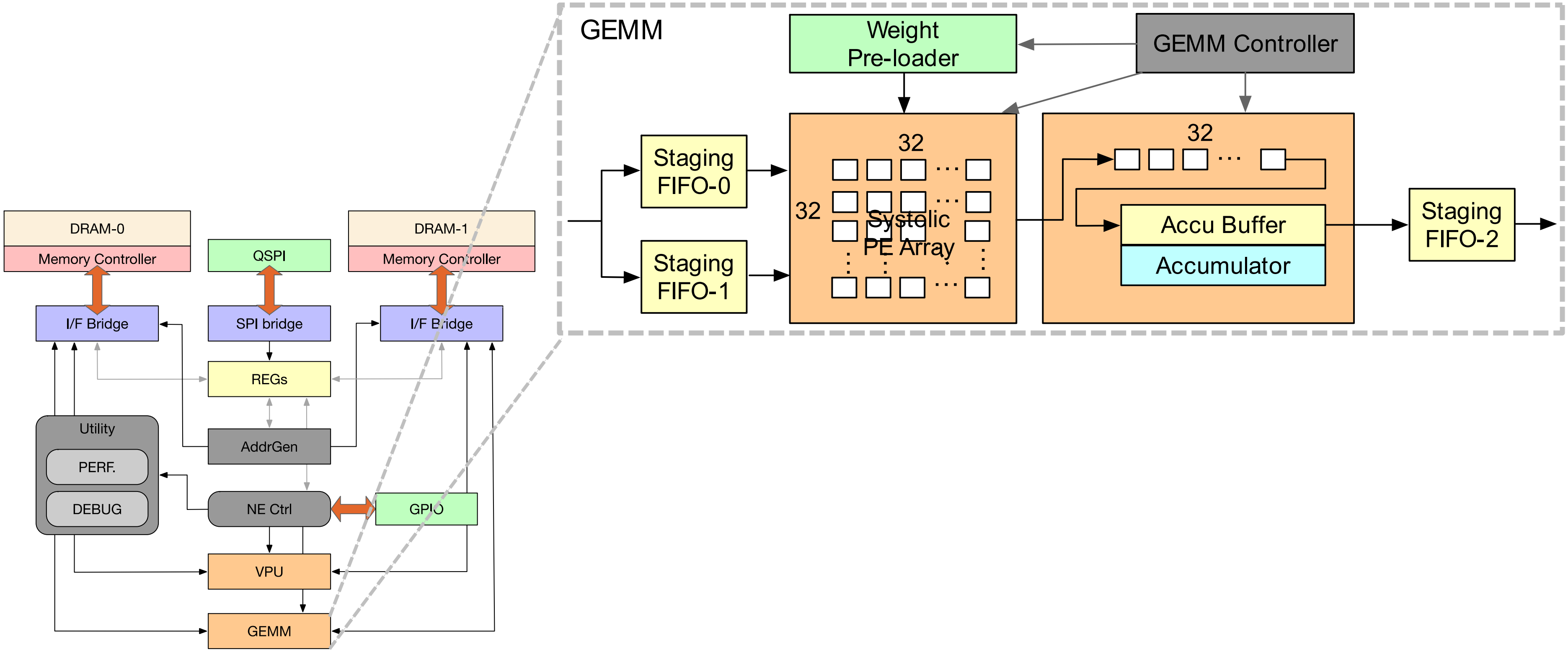
- Activations based on LUTs
  - Support for GeLU and Exp
- Transpose
  - Transpose 16x16 matrix with ping-pong array
  - 2D register file array
  - Row-based writes and column-based reads



Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Neural Engine: GEMM

- 32x32 INT8 fully-pipelined systolic array
  - Partial sums accumulated in INT32 accumulator



Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Lecture on Systolic Arrays

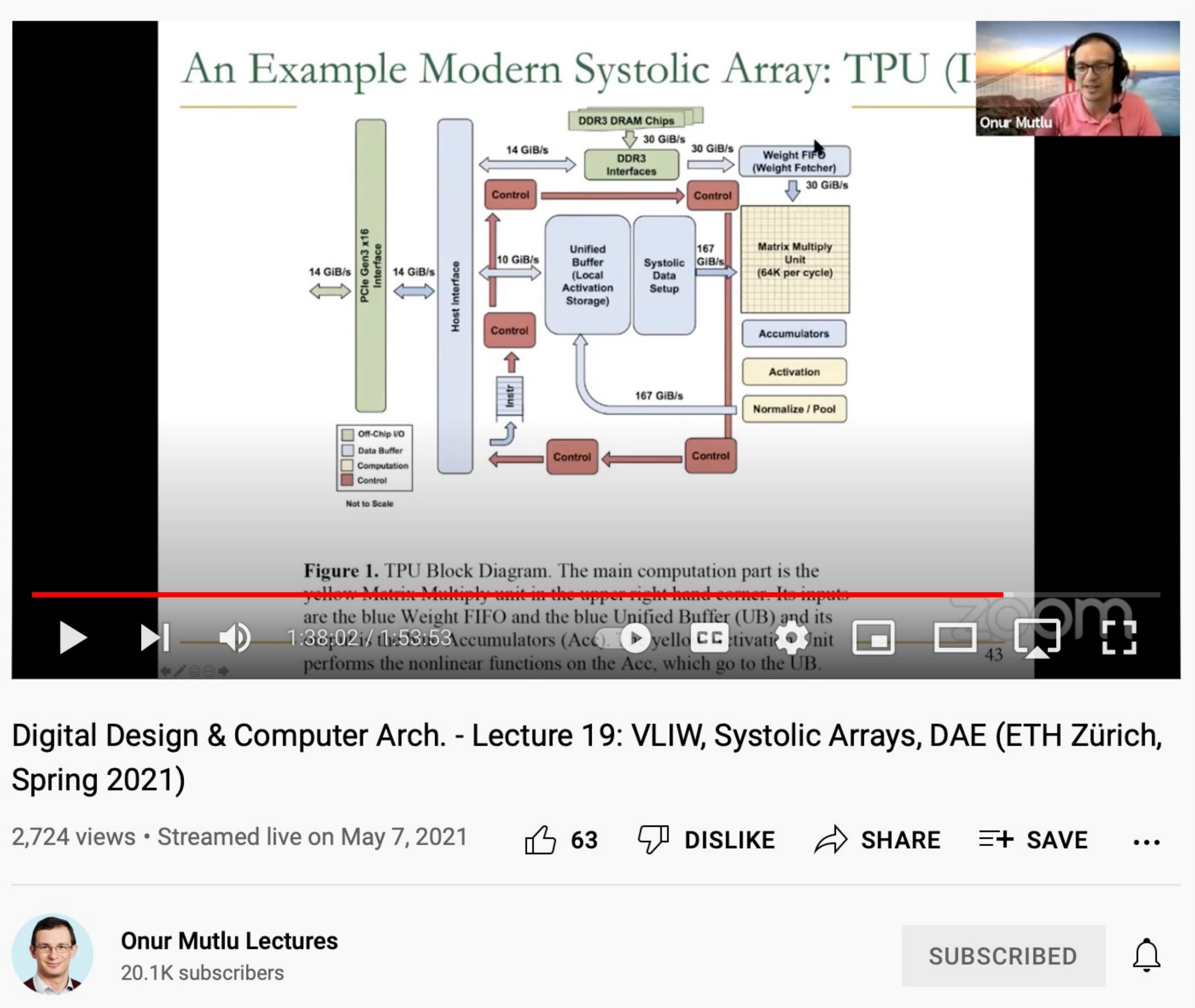


Figure 1. TPU Block Diagram. The main computation part is the yellow Matrix Multiply unit in the upper right hand corner. Its inputs are the blue Weight FIFO and the blue Unified Buffer (UB) and its output is the blue Accumulators (Acc). The yellow Activation Unit performs the nonlinear functions on the Acc, which go to the UB.

Digital Design & Computer Arch. - Lecture 19: VLIW, Systolic Arrays, DAE (ETH Zürich, Spring 2021)

2,724 views • Streamed live on May 7, 2021

63 DISLIKE SHARE SAVE ...

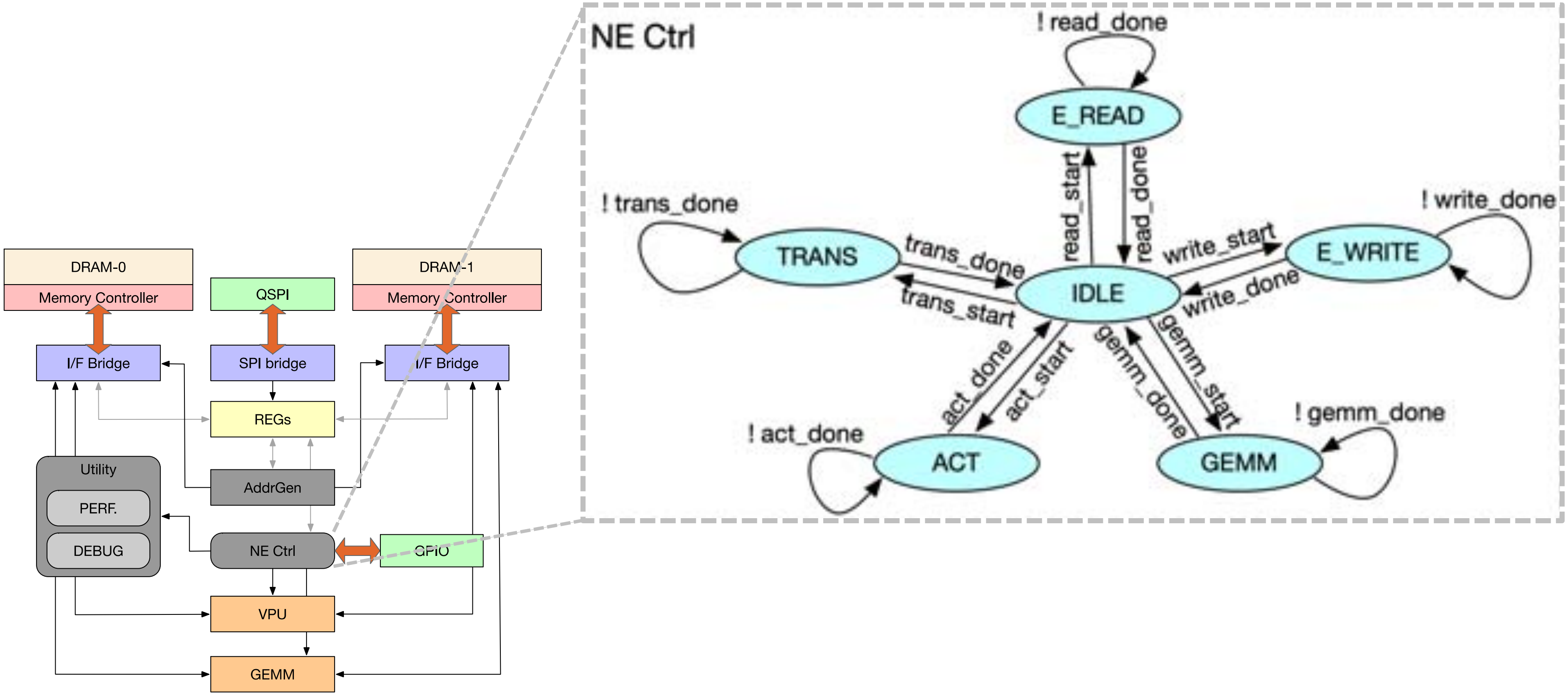
Onur Mutlu Lectures
20.1K subscribers

SUBSCRIBED

https://youtu.be/UtLy4Yagdys?t=2948

Neural Engine: Finite State Machine

- Five working states and one idle state
  - Each working state is for one instruction



Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

HB-PNM: Key Feature Summary

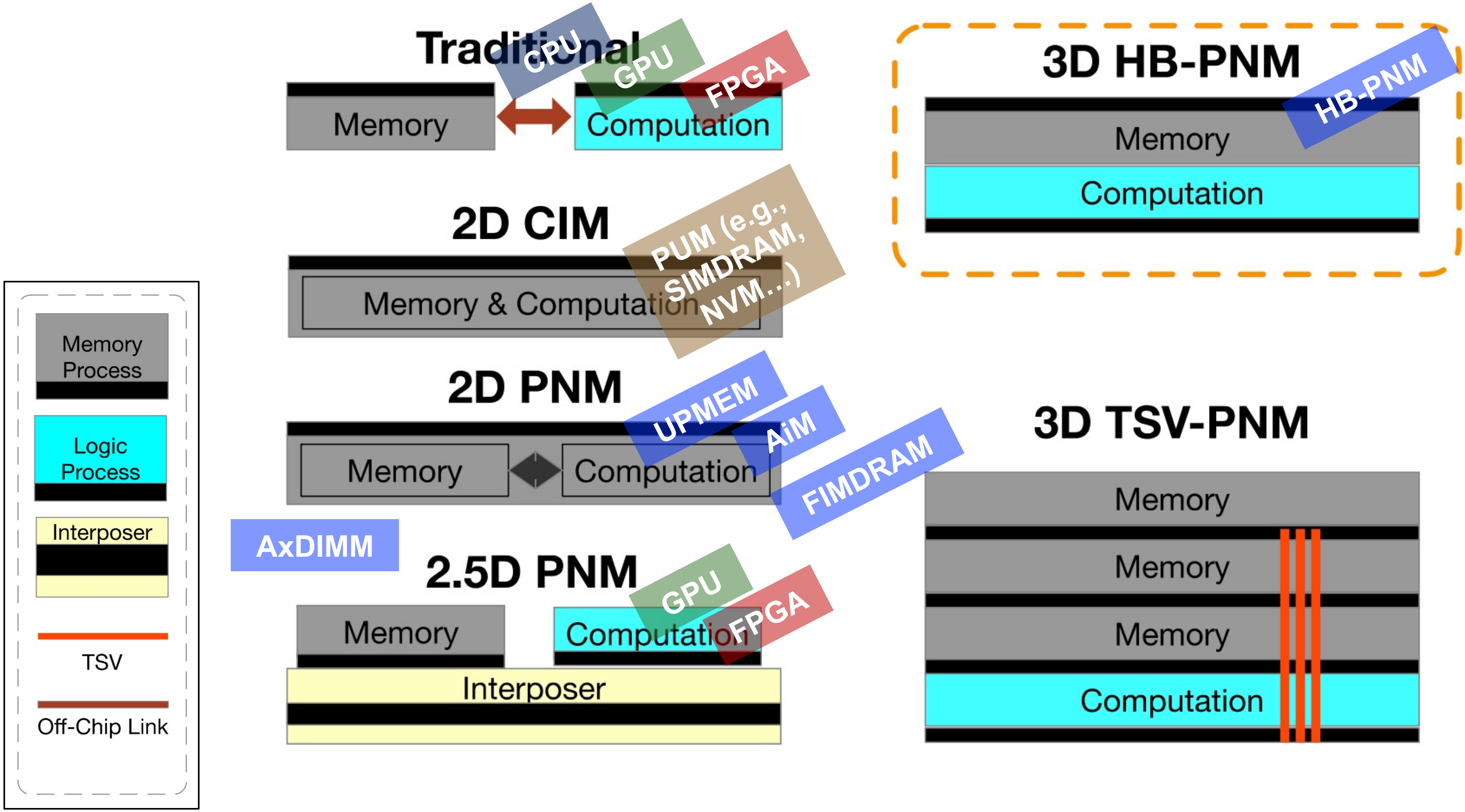
- Comparison table

	2D CIM *	UPMEM PIM **	A100 GPU ***	FIMDRAM ****	This work
Type of Memory	SRAM	DDR4	HBM2	HBM2	LPDDR4
Technology (Memory/Logic)	16nm	2xnm / 2xnm	1y$^{\#}$ / 7nm	20nm / 20nm	25nm / 55nm
Capacity	4.5 Mb	8GB / DIMM	80GB	6GB / cube	4.5GB
Bandwidth	-	128GB/s / DIMM	1935GB/s	1200GB/s / cube$^{\#}$	38.4GB/s / 1Gb
Frequency (Logic)	200MHz	500MHz	1410MHz	300MHz	300MHz
Bandwidth/Capacity (a.u.)	-	16	24.2	200	**307**
Energy	-	~25pJ/bit	4.47pJ/bit	2.75pJ/bit	**0.88pJ/bit**

* H. Jia et al, ISSCC 2021.
** F. Devaux et al, Hotchip 2019
*** J. Choquette et al, Hotchip 2020
**** Y. C. Kwon et al, ISSCC 2021

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Processing-in-Memory Classification

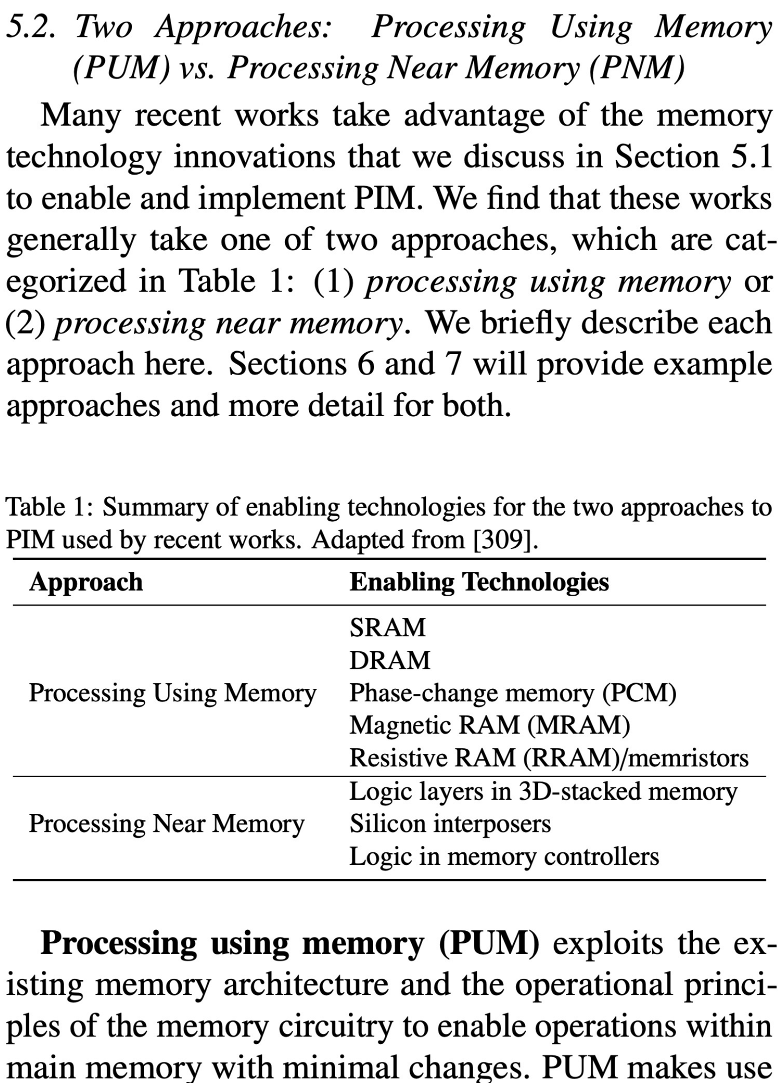


PNM : Process Near Memory
CIM: Compute In Memory

HB : Hybrid Bonding
TSV: Through-silicon Via

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

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

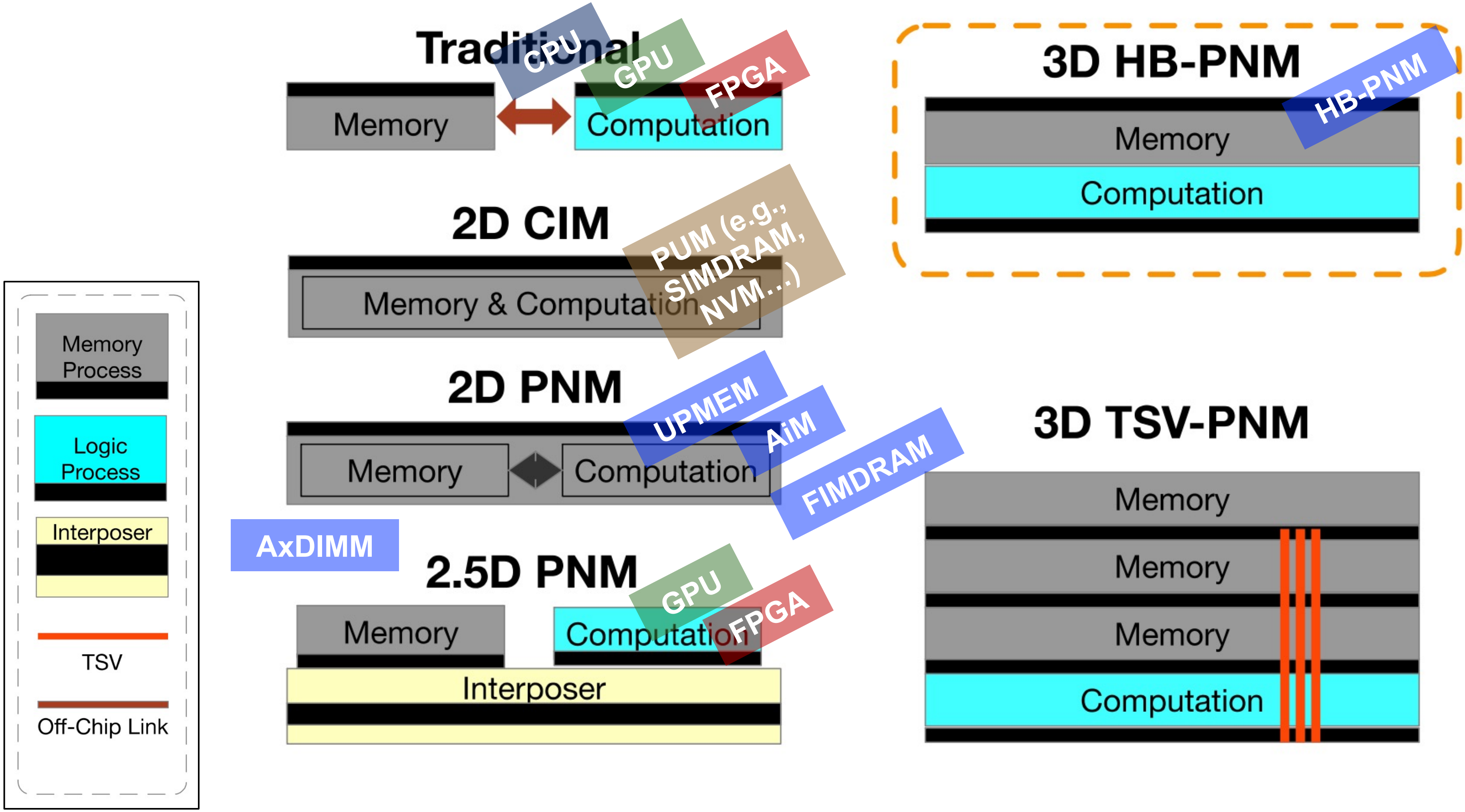
Onur Mutlu, Saugata Ghose, Juan Gomez-Luna, and Rachata Ausavarungnirun,
"A Modern Primer on Processing in Memory"
Invited Book Chapter in ***Emerging Computing: From Devices to Systems - Looking Beyond Moore and Von Neumann***, Springer, to be published in 2021.
[Tutorial Video on "Memory-Centric Computing Systems" (1 hour 51 minutes)]

https://people.inf.ethz.ch/omutlu/pub/ModernPrimerOnPIM_springer-emerging-computing-bookchapter21.pdf

Similarities and Differences among Current PIM Systems

- Similarities
  - Current real-world processing-in-memory architectures follow a processing-near-memory approach
  - All based on DRAM memory
- Differences
  - Near-bank (UPMEM, FIMDRAM, AiM, HB-PNM) vs. near-chip (AxDIMM)
  - General-purpose (UPMEM) vs. special-function (FIMDRAM, AiM, HB-PNM)
  - FGMT (UPMEM) vs. SIMD (FIMDRAM, AiM, AxDIMM) vs. systolic array (HB-PNM)
  - Natively integer (UPMEM, HB-PNM) vs. floating point (FIMDRAM)
    - FP16 (FIMDRAM) vs. BF16 (AiM) vs. FP32 (AxDIMM)
  - DDR4 (UPMEM, AxDIMM) vs. LPDDR4 (HB-PNM) vs. HBM2 (FIMDRAM) vs. GDDR6 (AiM)

Processing-in-Memory Classification

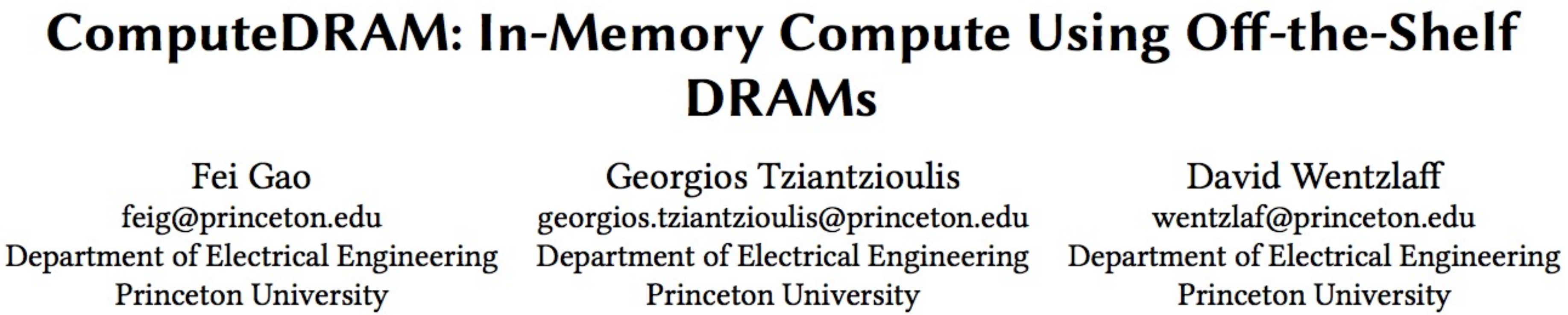


PNM : Process Near Memory
CIM: Compute In Memory

HB : Hybrid Bonding
TSV: Through-silicon Via

Niu et al., 184QPS/W 64Mb/mm2 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System, ISSCC 2022

Processing-using-Memory in Real DRAM Chips



ComputeDRAM: In-Memory Compute Using Off-the-Shelf DRAMs

Fei Gao
feig@princeton.edu
Department of Electrical Engineering
Princeton University

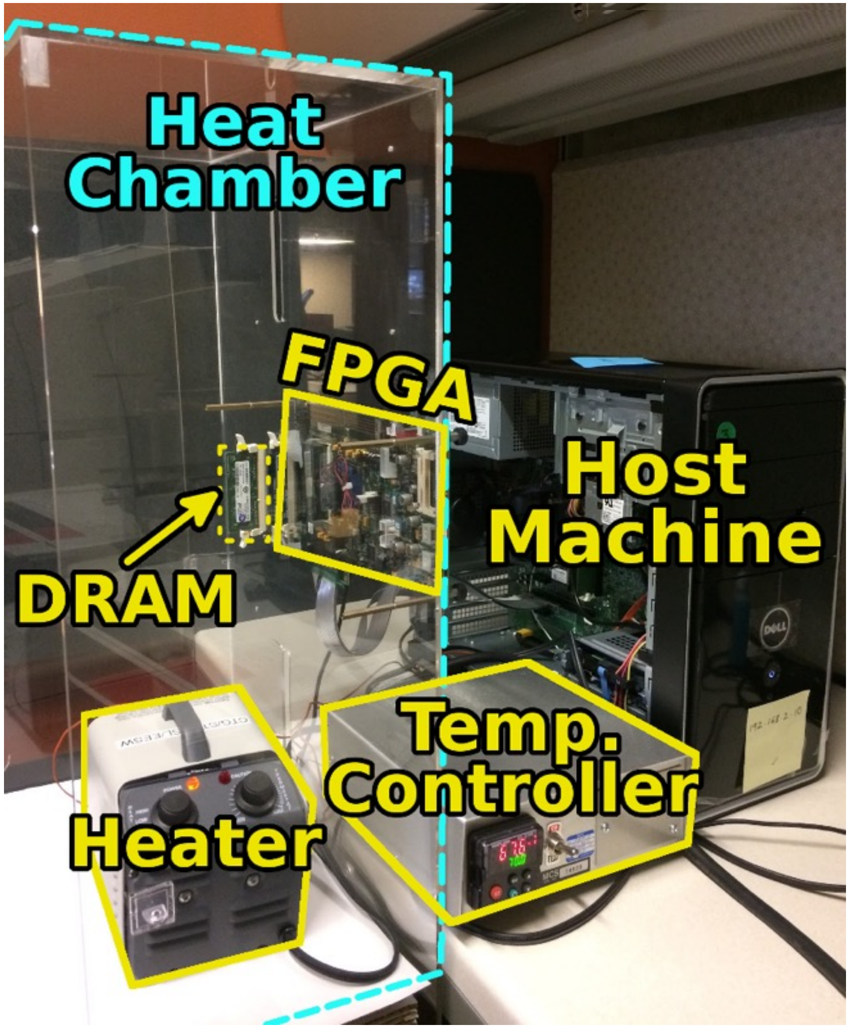
Georgios Tziantzioulis
georgios.tziantzioulis@princeton.edu
Department of Electrical Engineering
Princeton University

David Wentzlaff
wentzlaf@princeton.edu
Department of Electrical Engineering
Princeton University

https://parallel.princeton.edu/papers/micro19-gao.pdf

SoftMC: Open Source DRAM Infrastructure

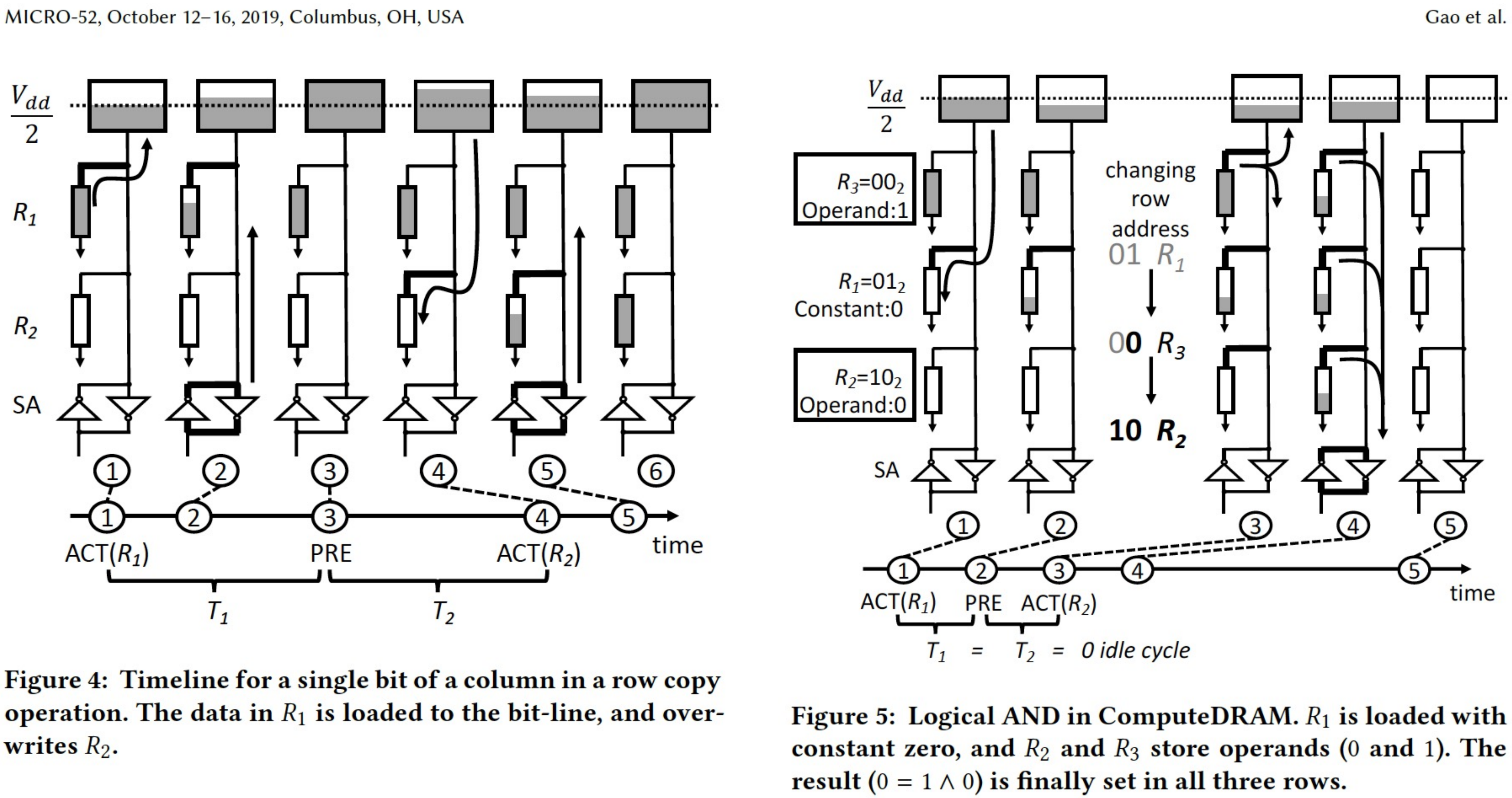
- Hasan Hassan et al., "**SoftMC: A Flexible and Practical Open-Source Infrastructure for Enabling Experimental DRAM Studies**," HPCA 2017



- Flexible
- Easy to Use (C++ API)
- Open-source

github.com/CMU-SAFARI/SoftMC

RowClone & Bitwise Ops in Real DRAM Chips



MICRO-52, October 12–16, 2019, Columbus, OH, USA

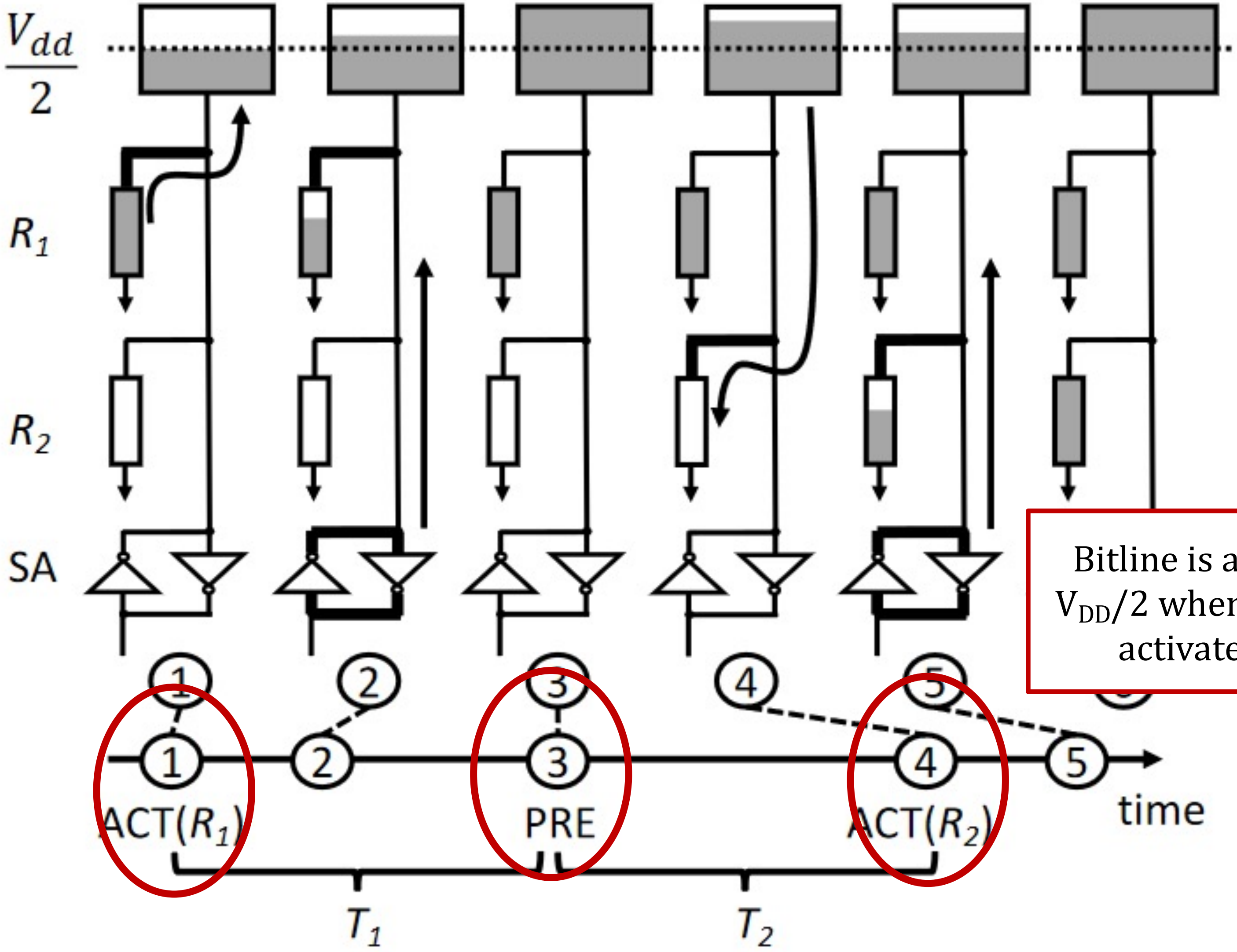
Gao et al.

Figure 4: Timeline for a single bit of a column in a row copy operation. The data in R_1 is loaded to the bit-line, and overwrites R_2.

Figure 5: Logical AND in ComputeDRAM. R_1 is loaded with constant zero, and R_2 and R_3 store operands (0 and 1). The result ($0 = 1 \land 0$) is finally set in all three rows.

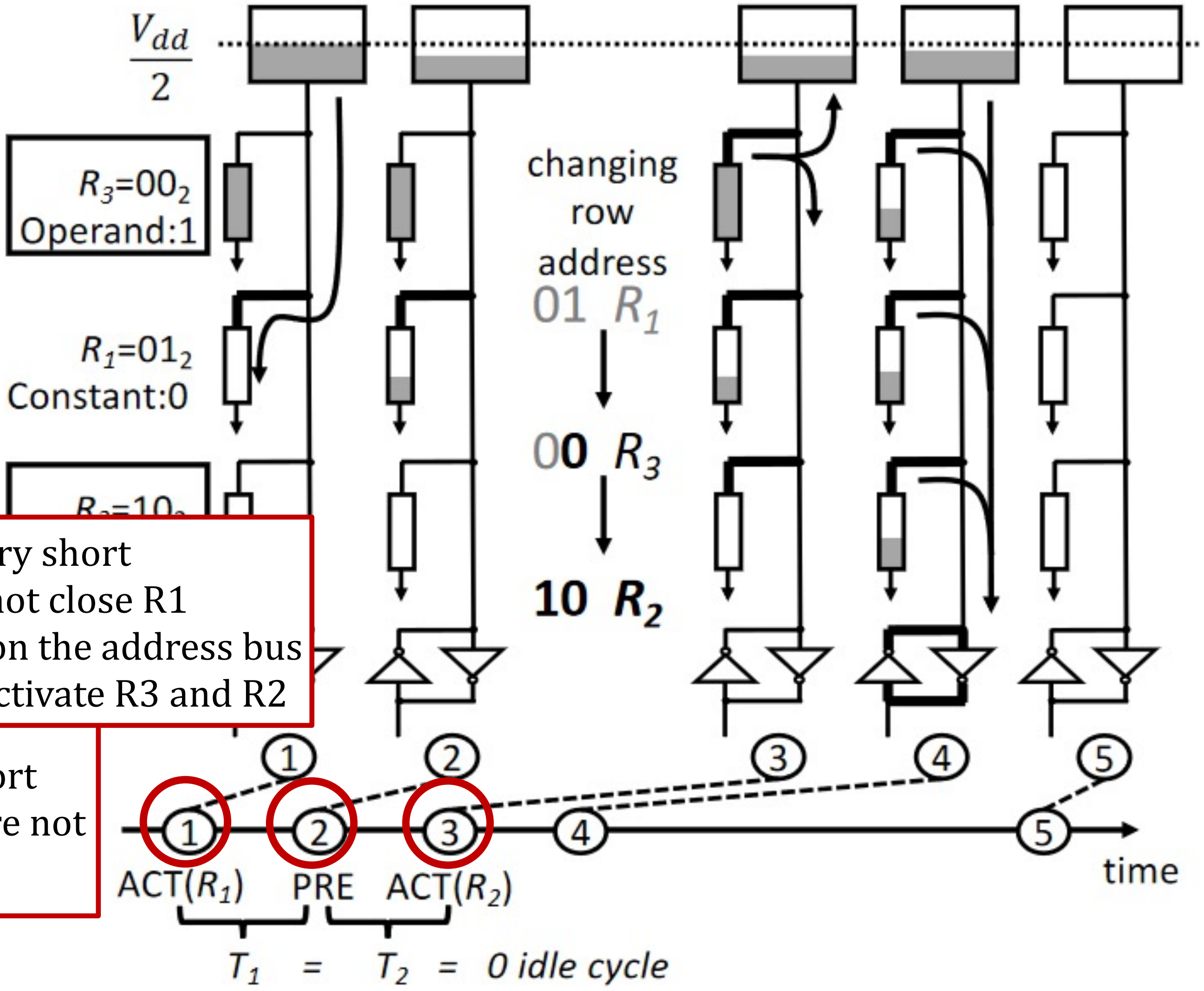
https://parallel.princeton.edu/papers/micro19-gao.pdf

Row Copy in ComputeDRAM



Bitline is above $V_{DD}/2$ when R2 is activated.

Bitwise AND in ComputeDRAM



T2 very short
PRE cannot close R1
R3 will appear on the address bus
ACT(R2) will activate R3 and R2

T1 very short
Sense amps are not activated

Experimental Methodology (I)

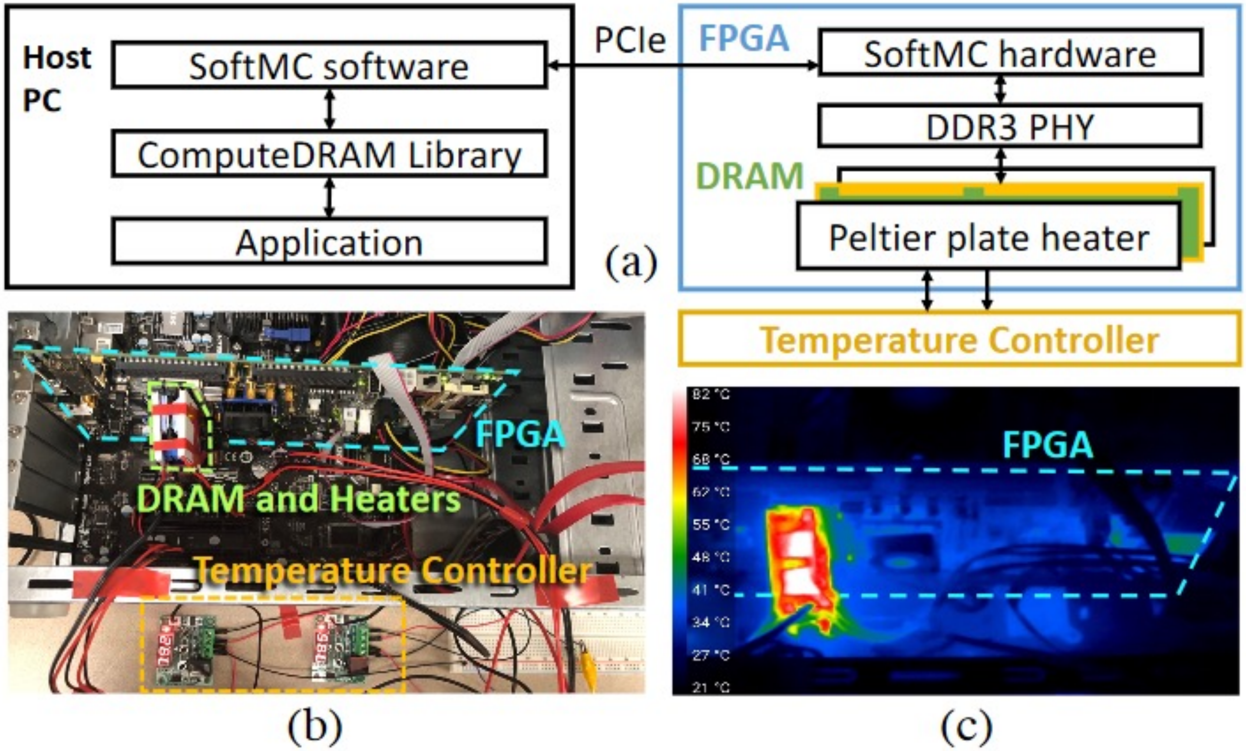


Figure 9: (a) Schematic diagram of our testing framework. (b) Picture of our testbed. (c) Thermal picture when the DRAM is heated to 80 °C.

Experimental Methodology (II)

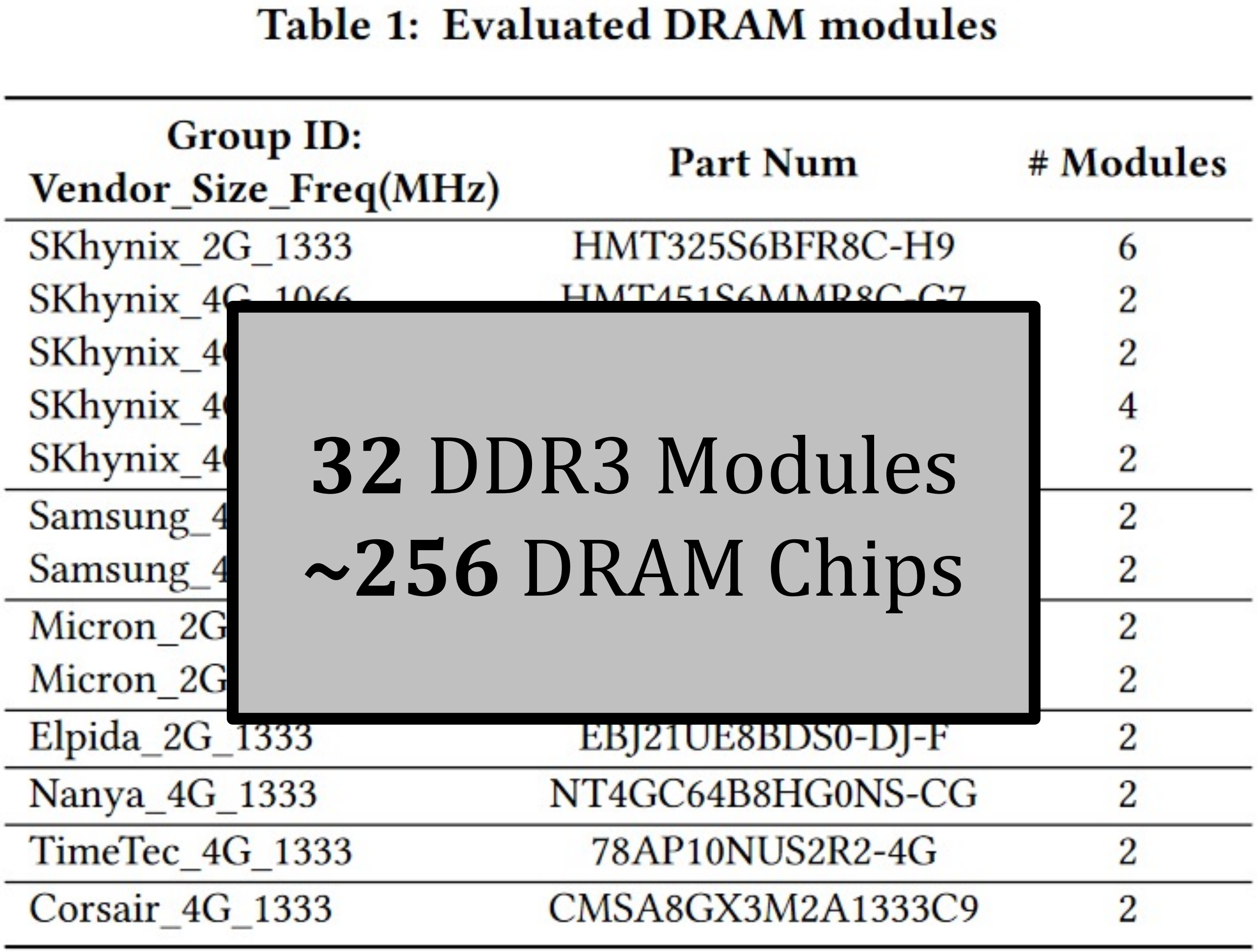


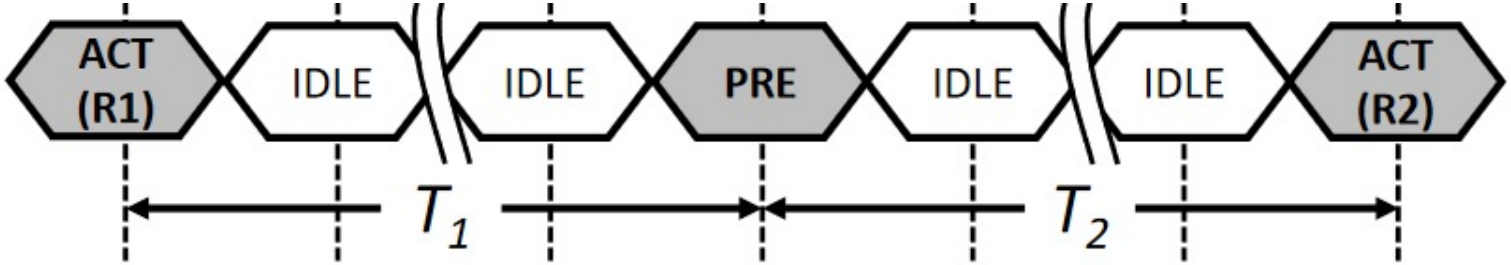
Table 1: Evaluated DRAM modules

Group ID: Vendor_Size_Freq(MHz)	Part Num	# Modules
SKhynix_2G_1333	HMT325S6BFR8C-H9	6
SKhynix_4G_1066	HMT451S6MMP8C-G7	2
SKhynix_4	[illegible]	2
SKhynix_4	[illegible]	4
SKhynix_4	[illegible]	2
Samsung_4	[illegible]	2
Samsung_4	[illegible]	2
Micron_2G	[illegible]	2
Micron_2G	[illegible]	2
Elpida_2G_1333	EBJ21UE8BDS0-DJ-F	2
Nanya_4G_1333	NT4GC64B8HG0NS-CG	2
TimeTec_4G_1333	78AP10NUS2R2-4G	2
Corsair_4G_1333	CMSA8GX3M2A1333C9	2

32 DDR3 Modules
~**256** DRAM Chips

Proof of Concept (I)

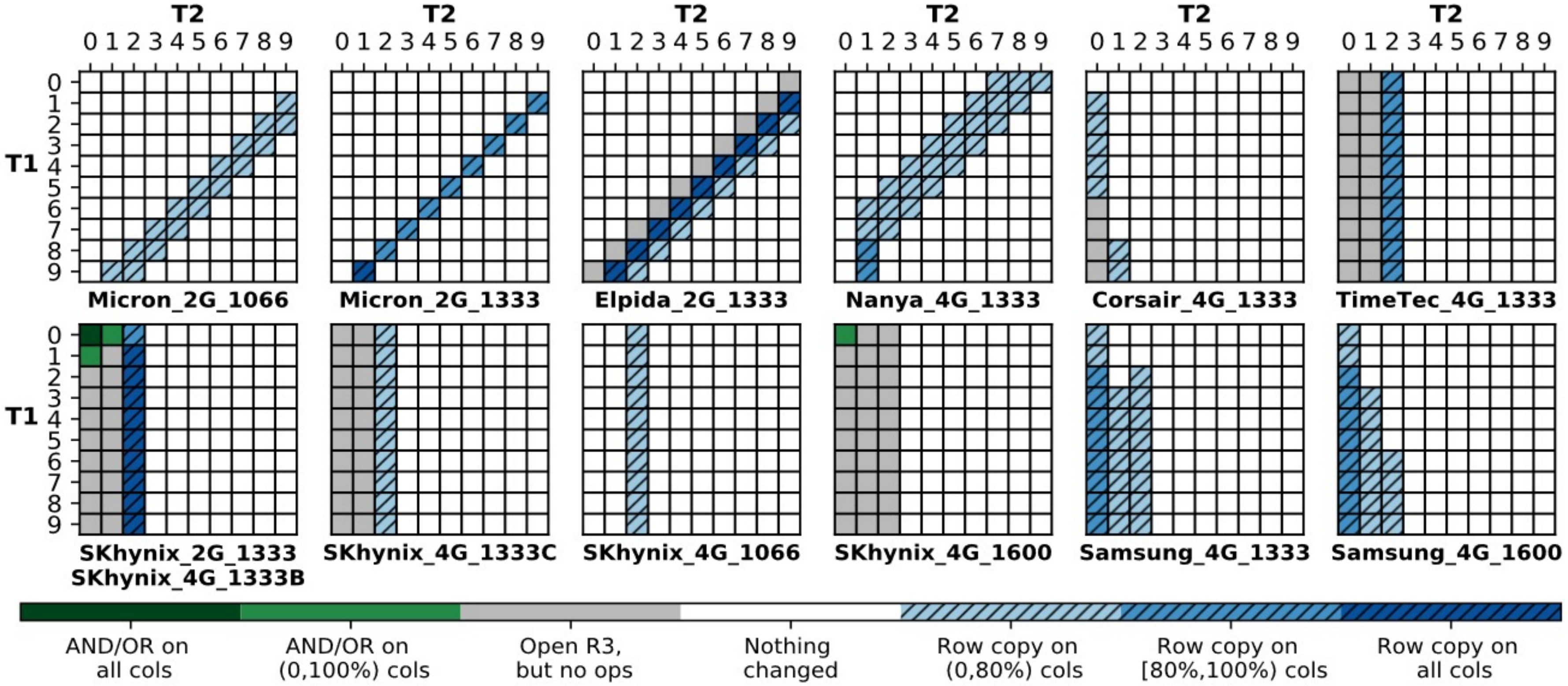
- How they test these memory modules:
  - Vary T_1 and T_2, observe what happens.



SoftMC Experiment

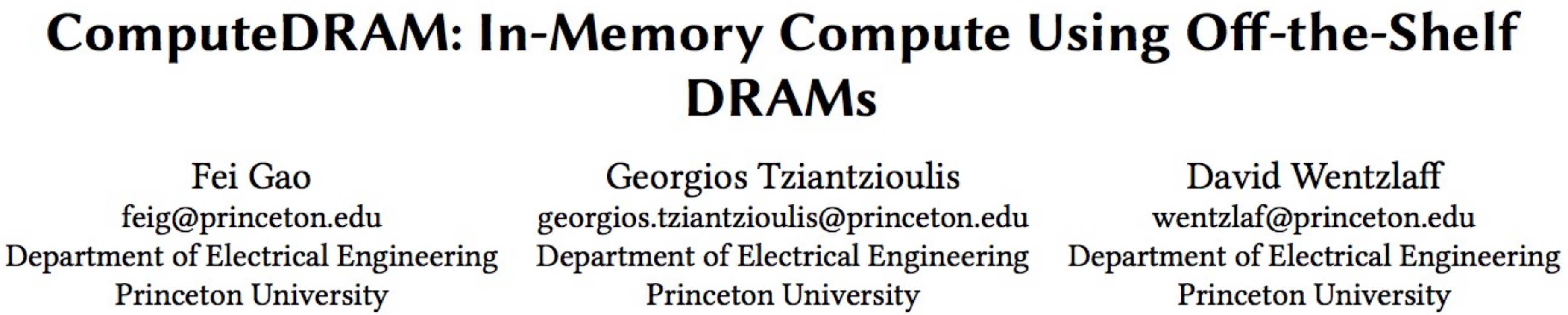
1. Select a random subarray
2. Fill subarray with random data
3. Issue ACT-PRE-ACTs with given T_1 & T_2
4. Read out subarray
5. Find out how many columns in a row support either operation
   - Row-wise success ratio

Proof of Concept (II)



- Each grid represents the success ratio of operations for a specific DDR3 module.

Processing-using-Memory in Real DRAM Chips



ComputeDRAM: In-Memory Compute Using Off-the-Shelf DRAMs

Fei Gao
feig@princeton.edu
Department of Electrical Engineering
Princeton University

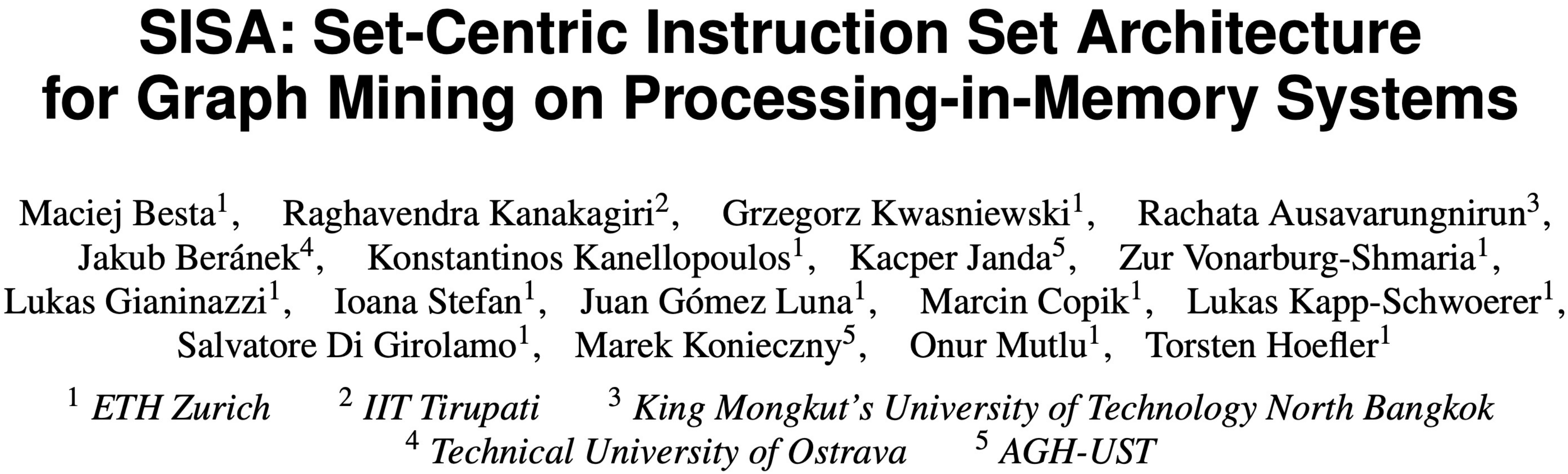
Georgios Tziantzioulis
georgios.tziantzioulis@princeton.edu
Department of Electrical Engineering
Princeton University

David Wentzlaff
wentzlaf@princeton.edu
Department of Electrical Engineering
Princeton University

https://parallel.princeton.edu/papers/micro19-gao.pdf

PnM and PuM Working Synergistically

- Maciej Besta, Raghavendra Kanakagiri, Grzegorz Kwasniewski, Rachata Ausavarungnirun, Jakub Beránek, Konstantinos Kanellopoulos, Kacper Janda, Zur Vonarburg-Shmaria, Lukas Gianinazzi, Ioana Stefan, Juan Gómez-Luna, Marcin Copik, Lukas Kapp-Schwoerer, Salvatore Di Girolamo, Nils Blach, Marek Konieczny, Onur Mutlu, and Torsten Hoefler, **"SISA: Set-Centric Instruction Set Architecture for Graph Mining on Processing-in-Memory Systems"** *Proceedings of the 54th International Symposium on Microarchitecture (**MICRO**)*, Virtual, October 2021. [Older arXiv version]



SISA: Set-Centric Instruction Set Architecture for Graph Mining on Processing-in-Memory Systems

Maciej Besta[1], Raghavendra Kanakagiri[2], Grzegorz Kwasniewski[1], Rachata Ausavarungnirun[3], Jakub Beránek[4], Konstantinos Kanellopoulos[1], Kacper Janda[5], Zur Vonarburg-Shmaria[1], Lukas Gianinazzi[1], Ioana Stefan[1], Juan Gómez Luna[1], Marcin Copik[1], Lukas Kapp-Schwoerer[1], Salvatore Di Girolamo[1], Marek Konieczny[5], Onur Mutlu[1], Torsten Hoefler[1]

[1] *ETH Zurich* [2] *IIT Tirupati* [3] *King Mongkut's University of Technology North Bangkok*
[4] *Technical University of Ostrava* [5] *AGH-UST*

Upcoming Lectures

- PUM architectures and prototypes

- Case studies
  - SpMV on UPMEM PIM architecture
  - Neural network accelerators for the edge
  - Hybrid transactional and analytical processing (HTAP) databases

- Enabling the adoption of PIM

P&S Processing-in-Memory

Real-World Processing-in-Memory Architectures: Alibaba Hybrid Bonding PNM Engine

Dr. Juan Gómez Luna

Prof. Onur Mutlu

ETH Zürich

Spring 2022

12 May 2022