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]UnilC, Xi'an, China

Self Introduction

Education Background

- B.S and M.S degree in electronic engineering from Tsinghua University
- Ph.D. degree in computer engineering from Pennsylvania State University

Work Experience

- Computing Technology Lab, DAMO Academy, Alibaba since 2019
- Memory Solutions Lab, Samsung Semiconductor 2014 - 2019

Research Interests

- Computer Architecture, Computing in/with Memory, Non-volatile memory

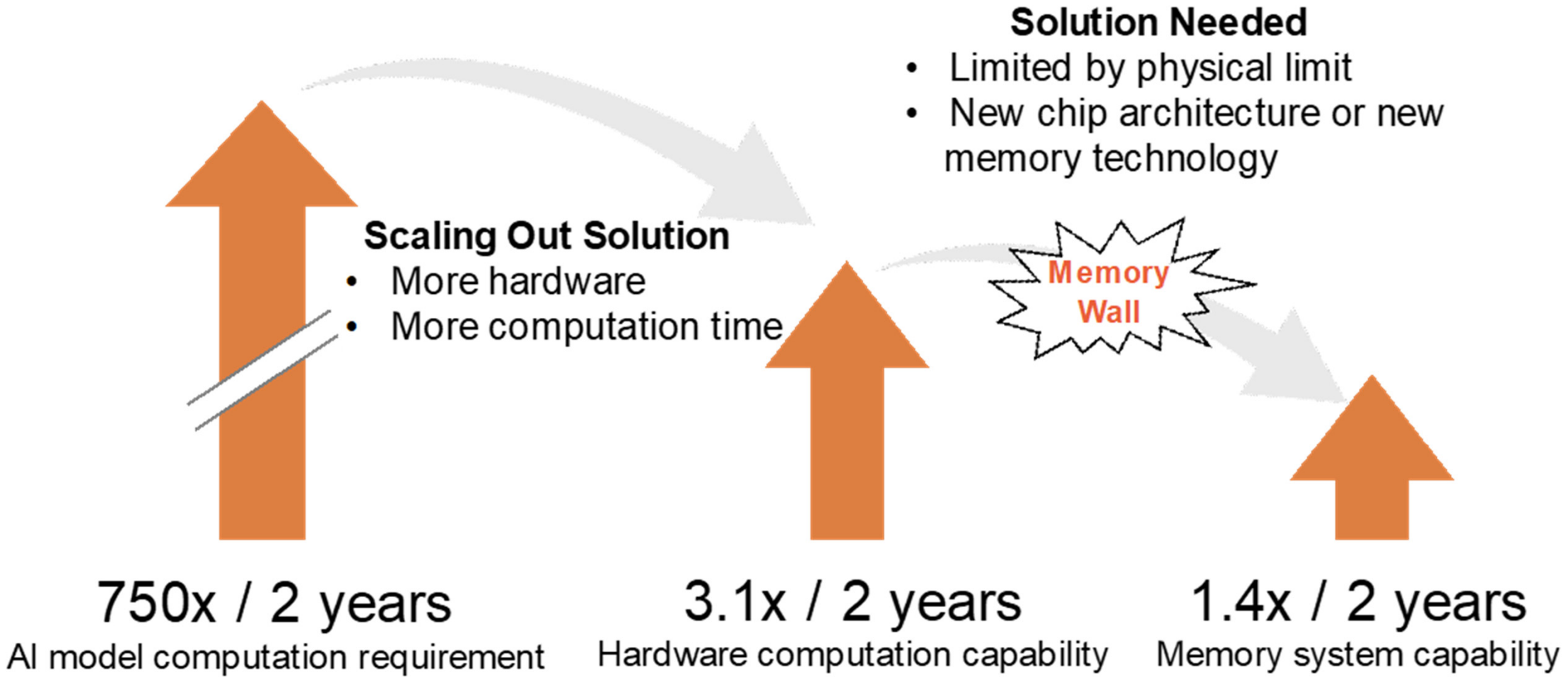
Outline

- **Motivation**
- **System and Chip Architecture**
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- **Measurement Results**
- **Conclusion**

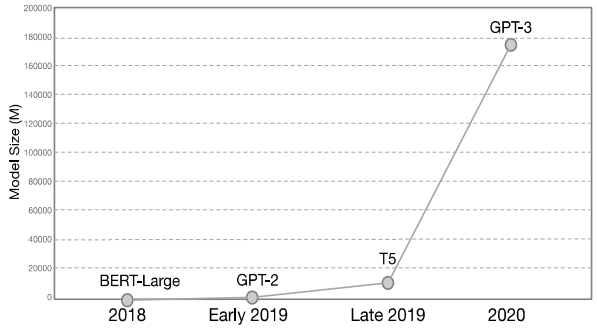
Outline

- **Motivation**
- System and Chip Architecture
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- Measurement Results
- Conclusion

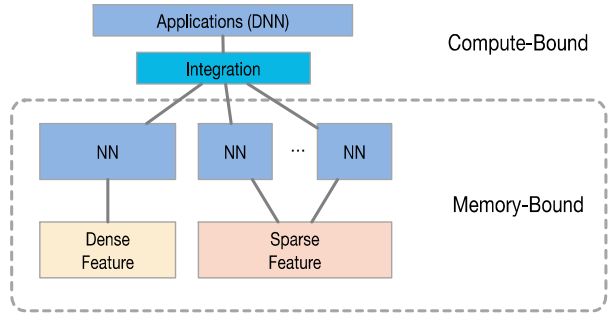
Memory Wall in AI Era



Memory-Bound Applications

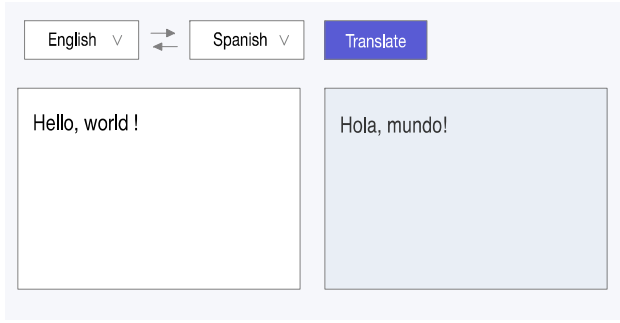


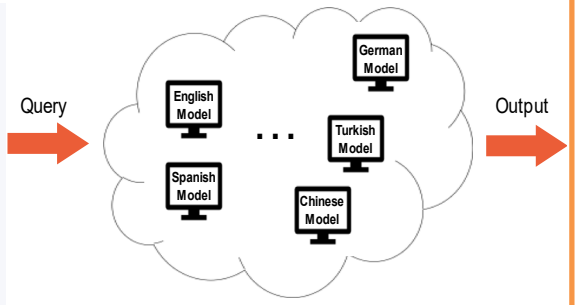
Natural Language Processing



Recommendation Systems

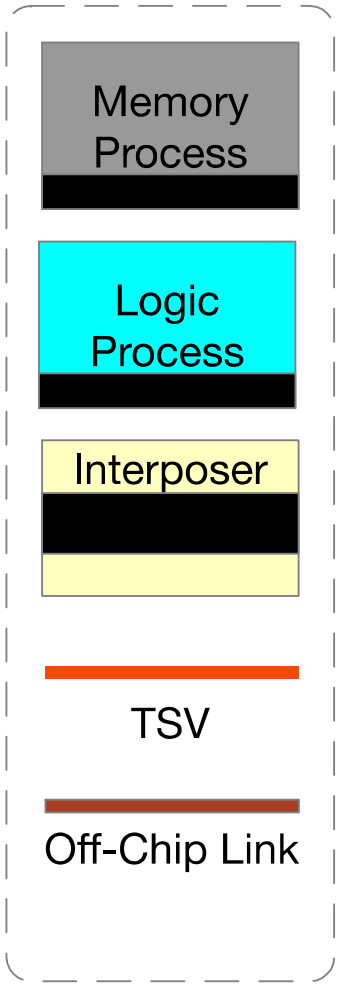
Graph Neural Network

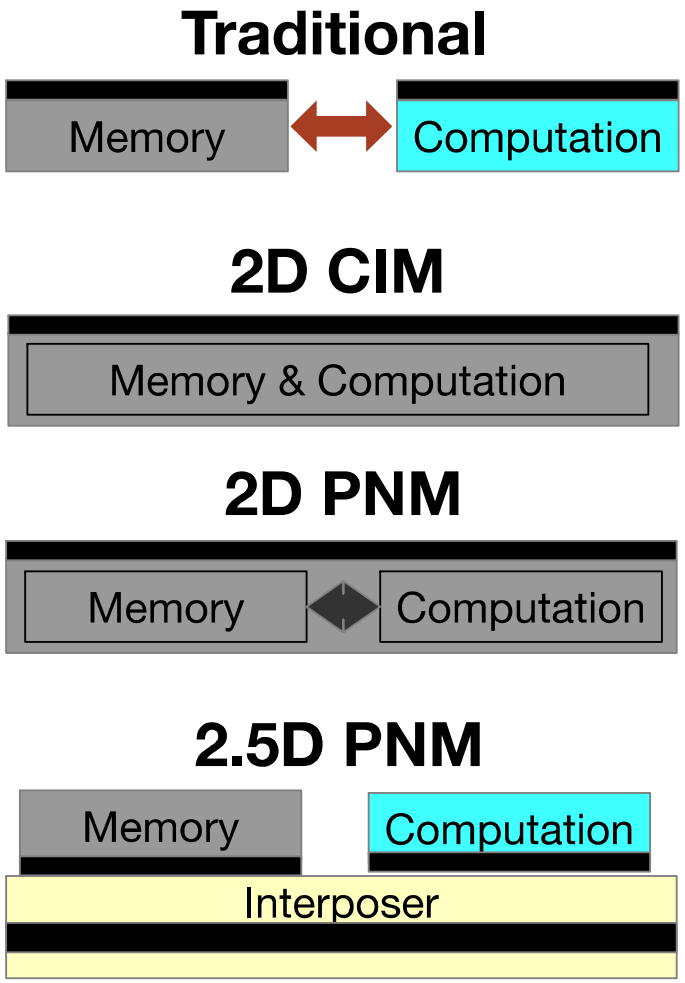


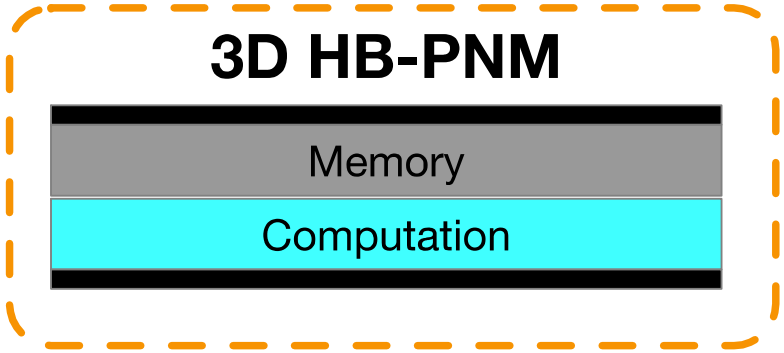


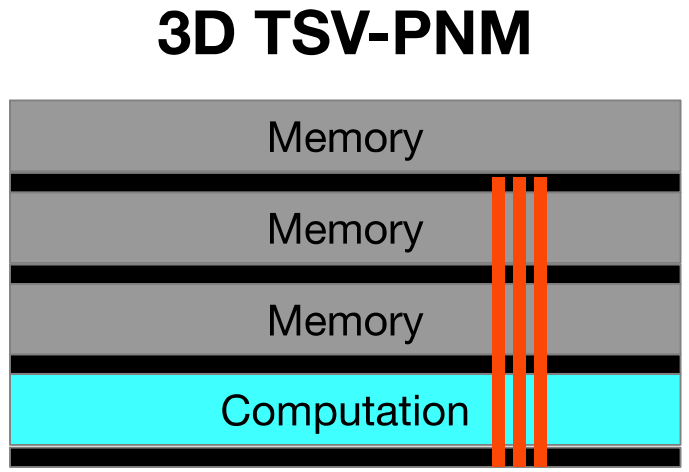
Multi-Task Online Inference

State-of-the-art PNM/CIM Solutions







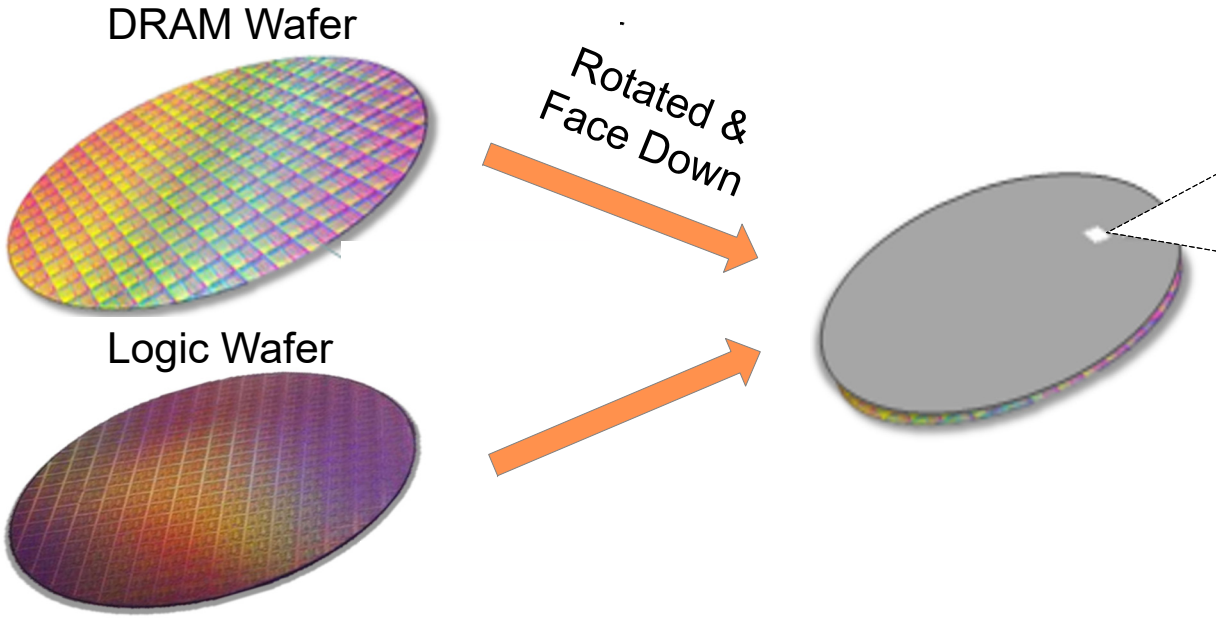


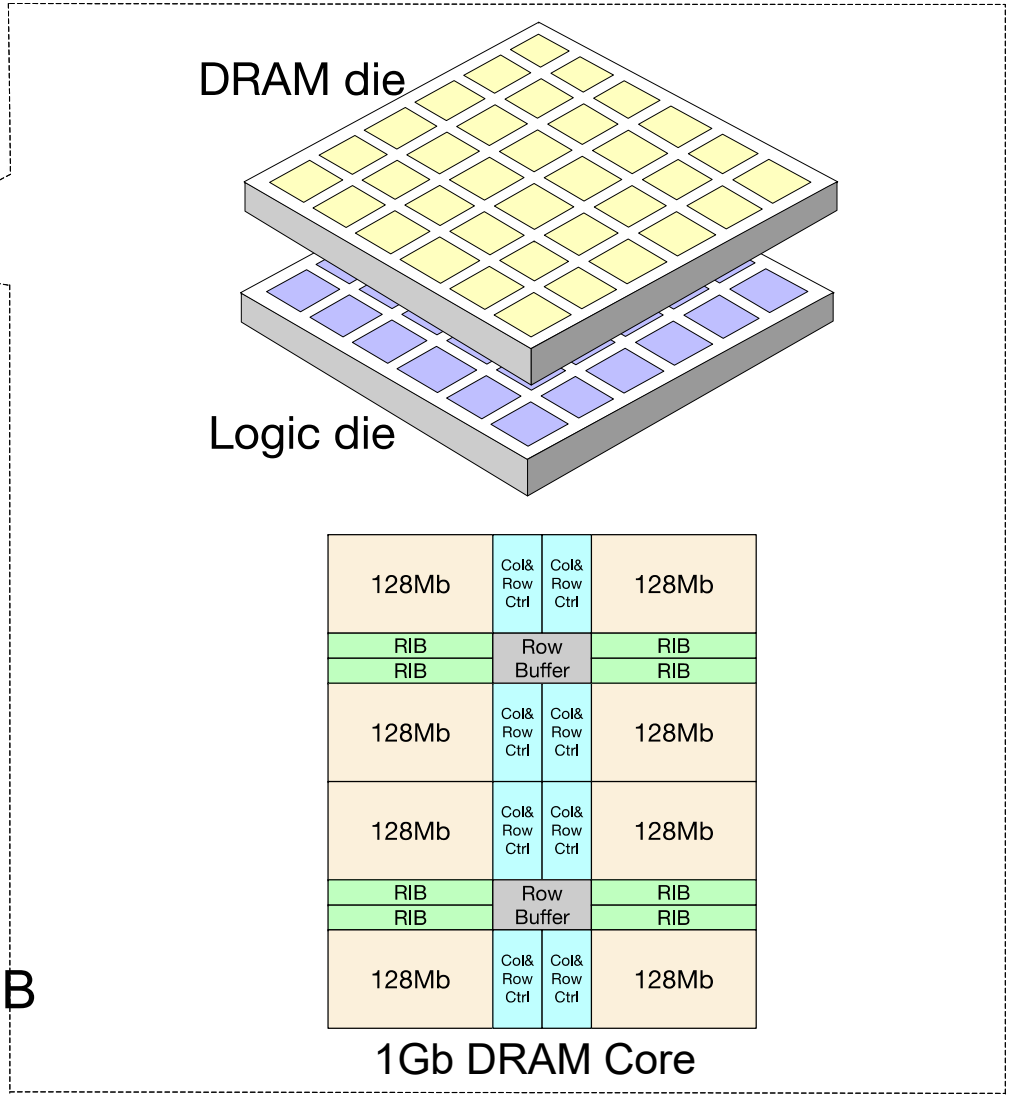
PNM : Process Near Memory
CIM: Compute In Memory
HB : Hybrid Bonding
TSV: Through-silicon Via

Outline

- Motivation
- **System and Chip Architecture**
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- Measurement Results
- Conclusion

3D Logic-to-DRAM Hybrid Bonding

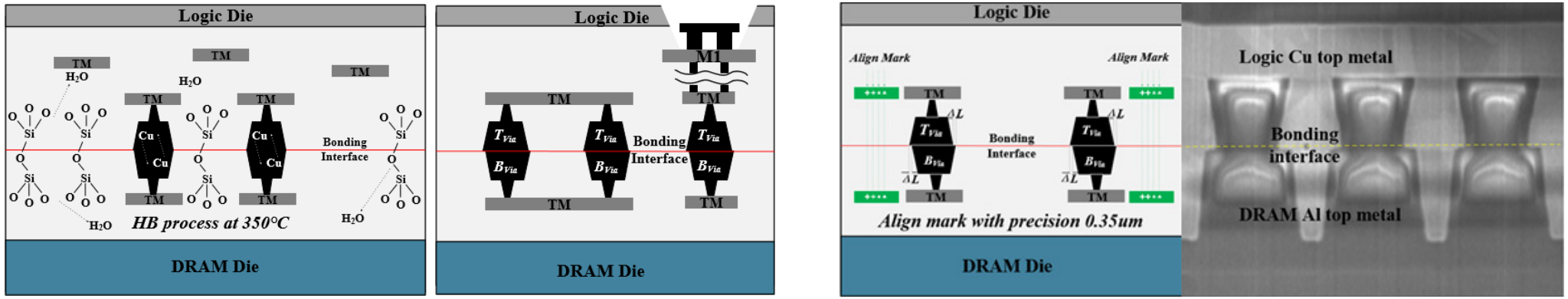




- Logic-to-DRAM face-to-face **H**ybrid wafer **B**onding
- 25nm DRAM technology with 36 x 1Gbits array
- 1Gbits DRAM core with 8 banks and on-chip ECC
- Each bank with 128 bits I/O, and implemented with HB

Hybrid-bonding Interconnection

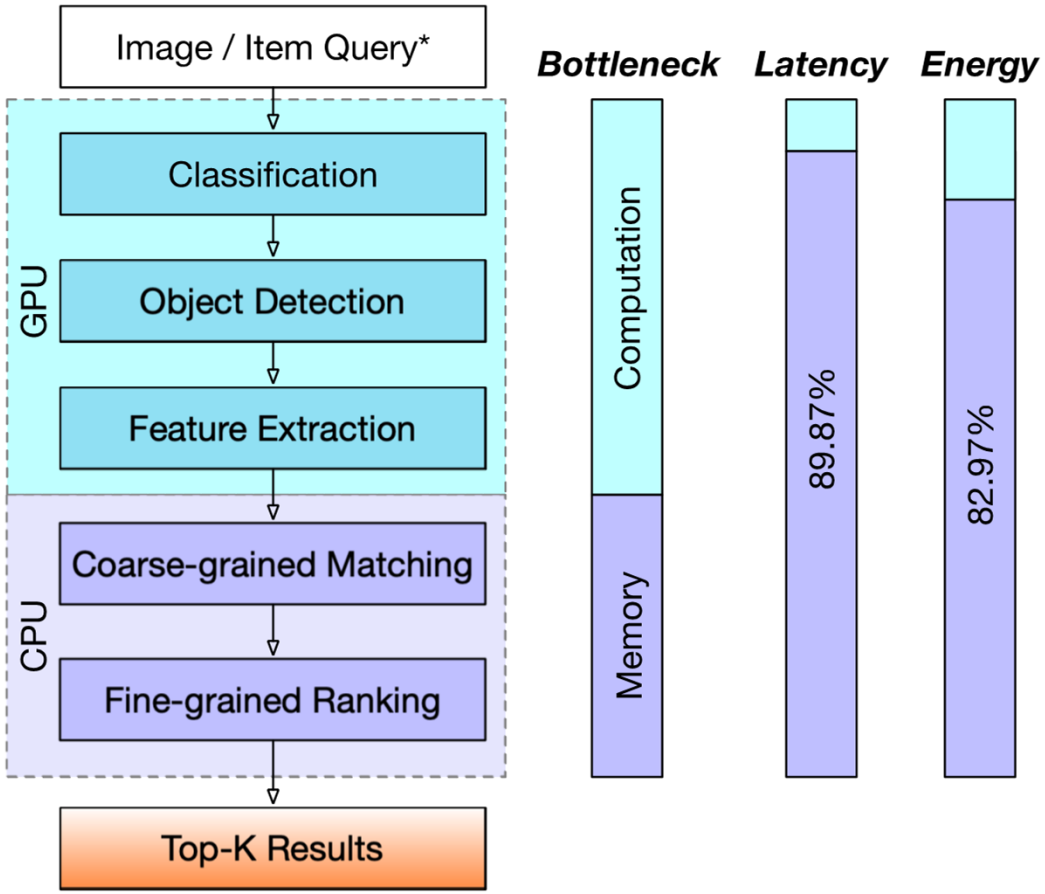
- Cu-Cu direct fusion with low bonding temperature (< 350°C)
- Up to 110,000/mm² integration density
- Small pitch size of 3μm
- Align marker with high precision of 0.35μm



Outline

- Motivation
- **System and Chip Architecture**
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- Measurement Results
- Conclusion

Typical Recommendation System

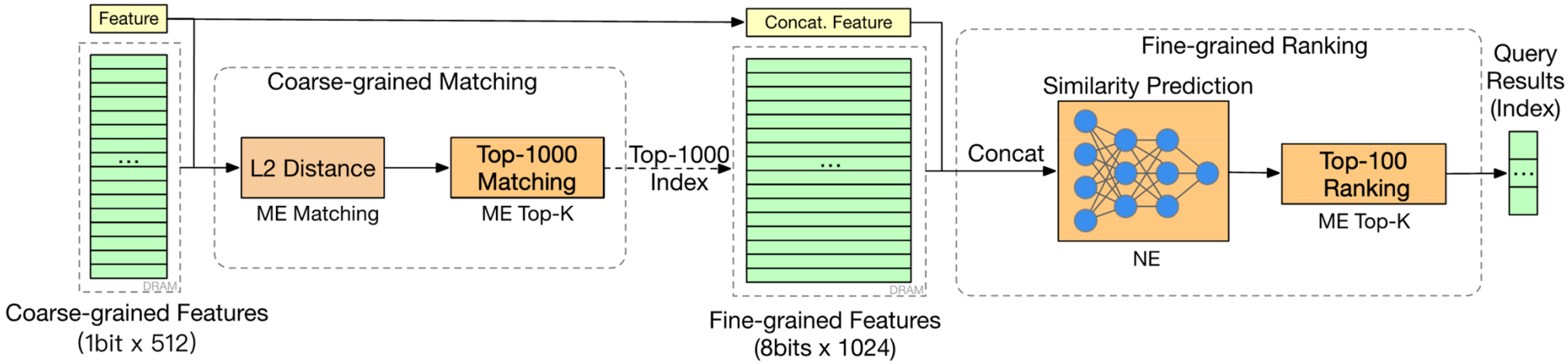


*Item feature can be extracted from different methods. Here is a typical case for image queries.

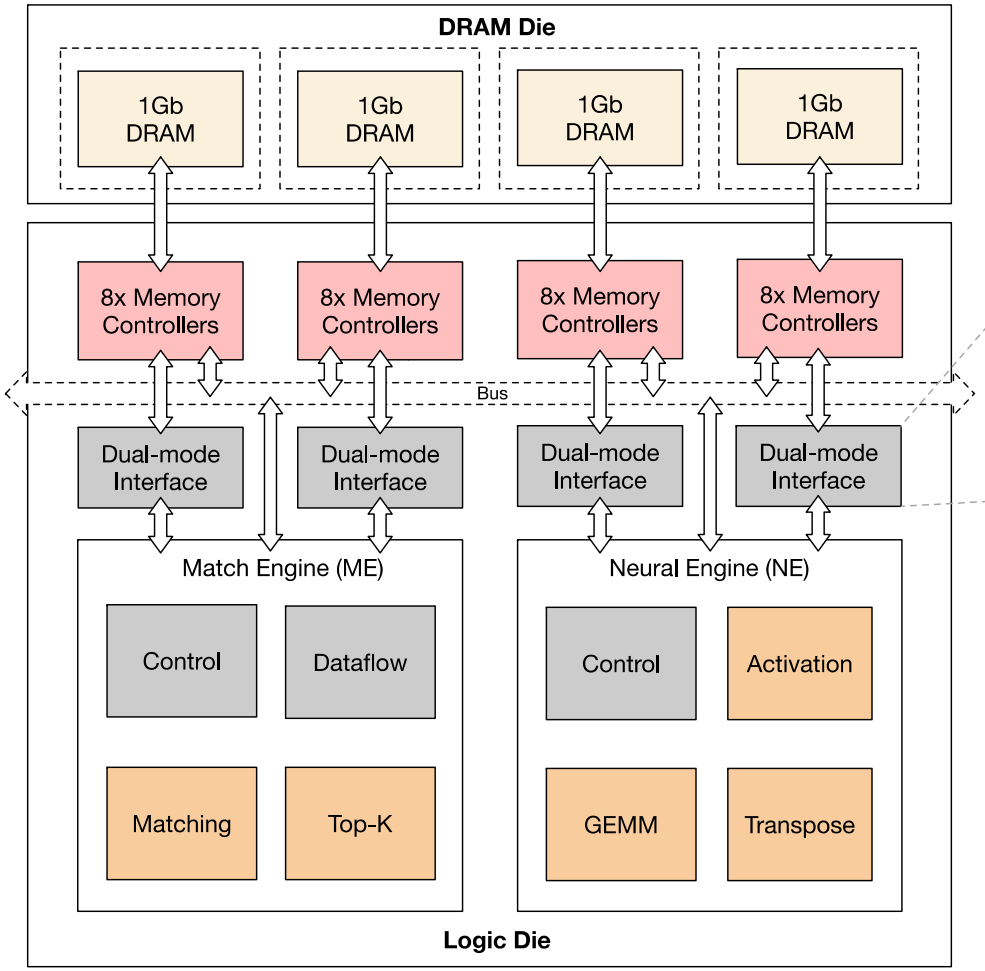
- A two-step Recommendation System
  - Feature Generation
    - Classification, object detection and feature extraction
    - Computation-bound
    - Typically executed on GPU
  - Matching & Ranking
    - Coarse-grained matching and fine-grained ranking
    - Memory-bound
    - Typically executed on CPU and commercial DRAM as external memory
    - Consumes most latency (**89.87%**) and energy (**82.97%**)
    - Requires **high-bandwidth, large-capacity and energy-efficient** memory

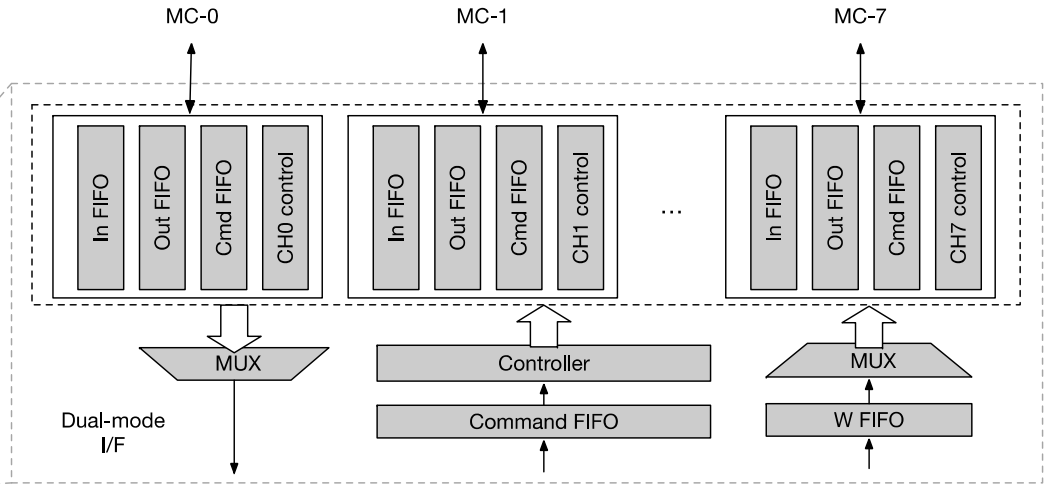
Ranking & Matching

- Coarse-grained Matching
  - Coarse-grained features with 1bit x 512 dimensions
  - Matching: L2 distance calculation
  - Top-1000 items selected from 40K items
- Fine-grained Ranking
  - Fine-grained features with 8bits x 1024 dimensions
  - Similarity prediction: three-layer MLP (2048-256-64-1)
  - Top-100 ranking results selected from 1K items



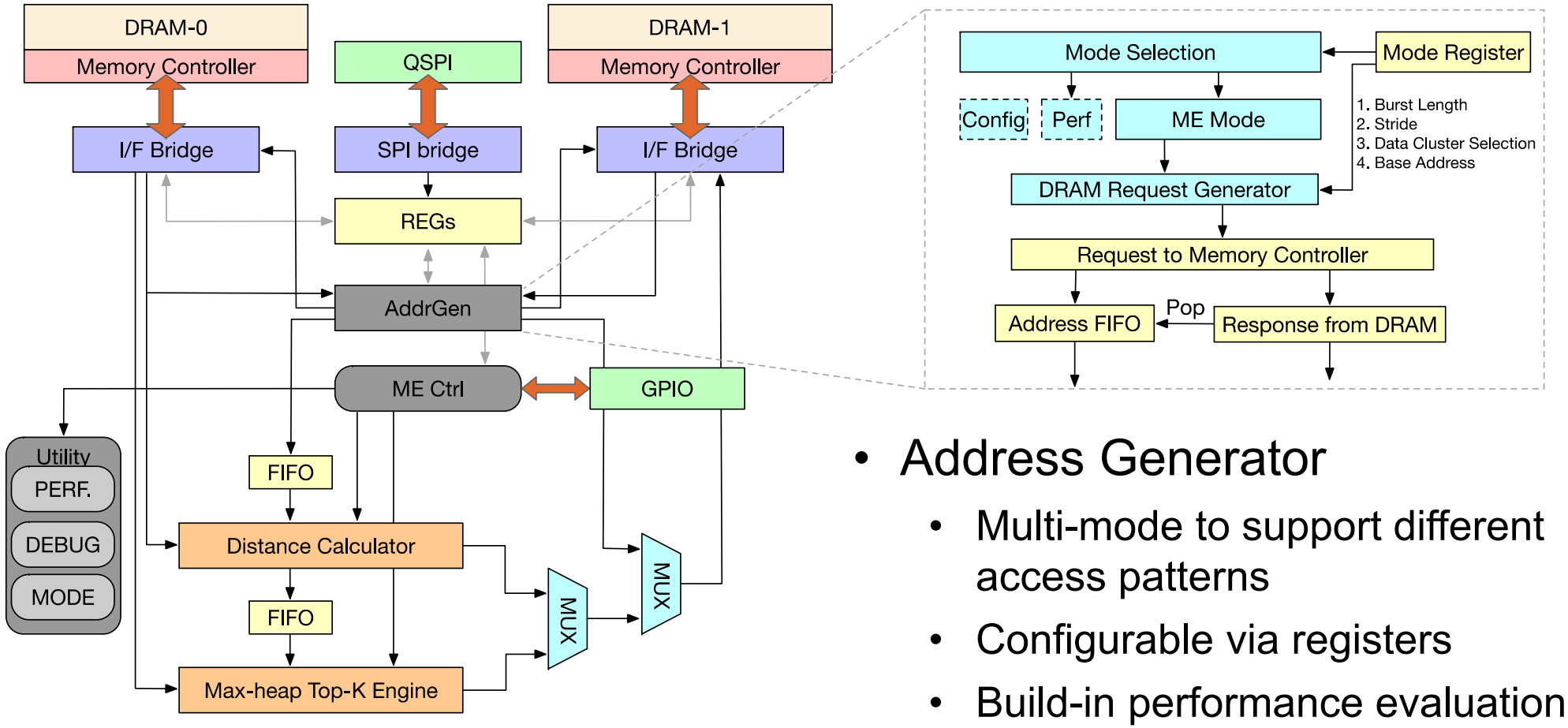
Overall Architecture





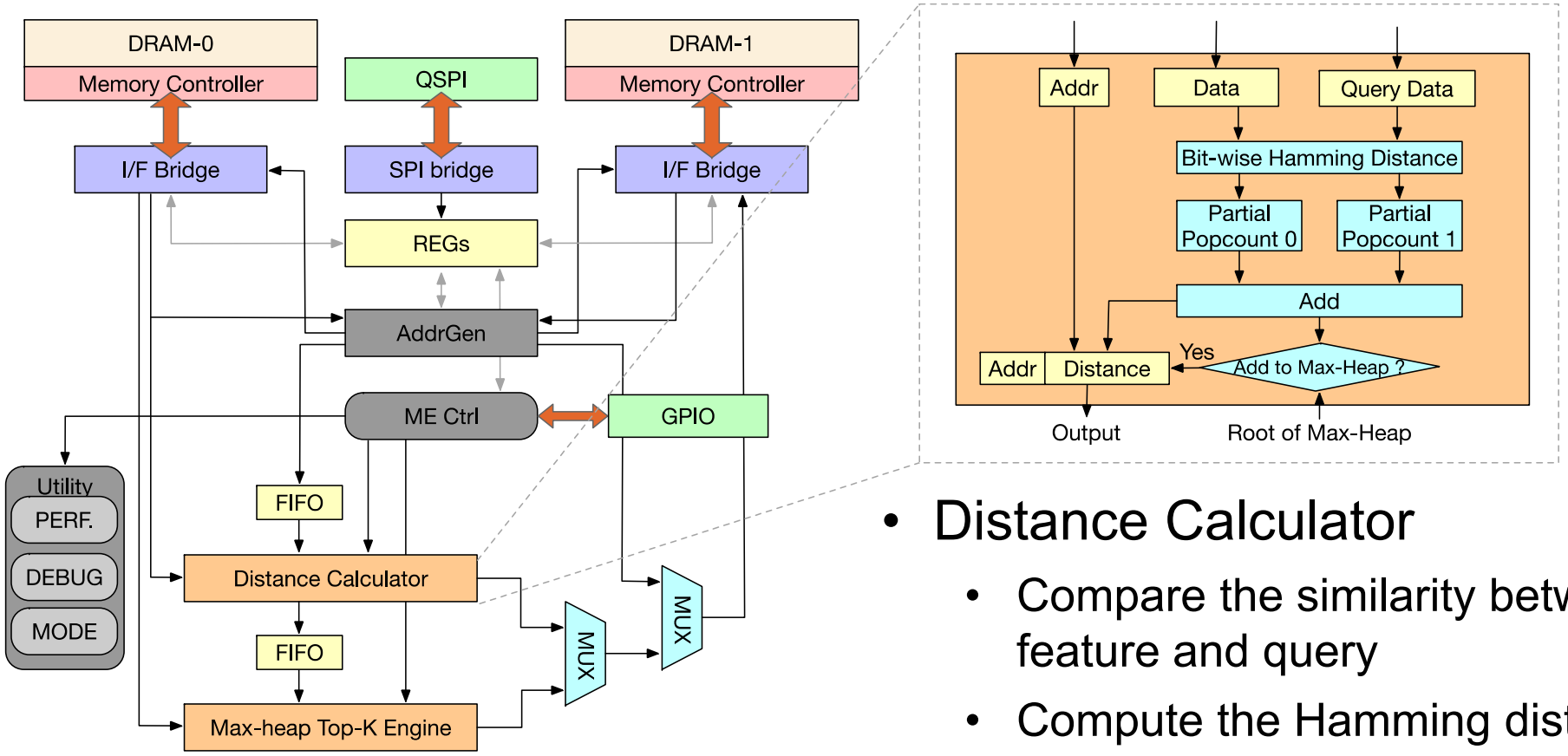
- Memory
  - 4 x 1Gb blocks with 4096 bits I/O
  - 38.4GB/s on-chip bandwidth per block
- Compute
  - Match Engine: Coarse-grained Matching
  - Neural Engine: Fine-grained Ranking
- Dual-mode Interface

Match Engine Architecture (1)



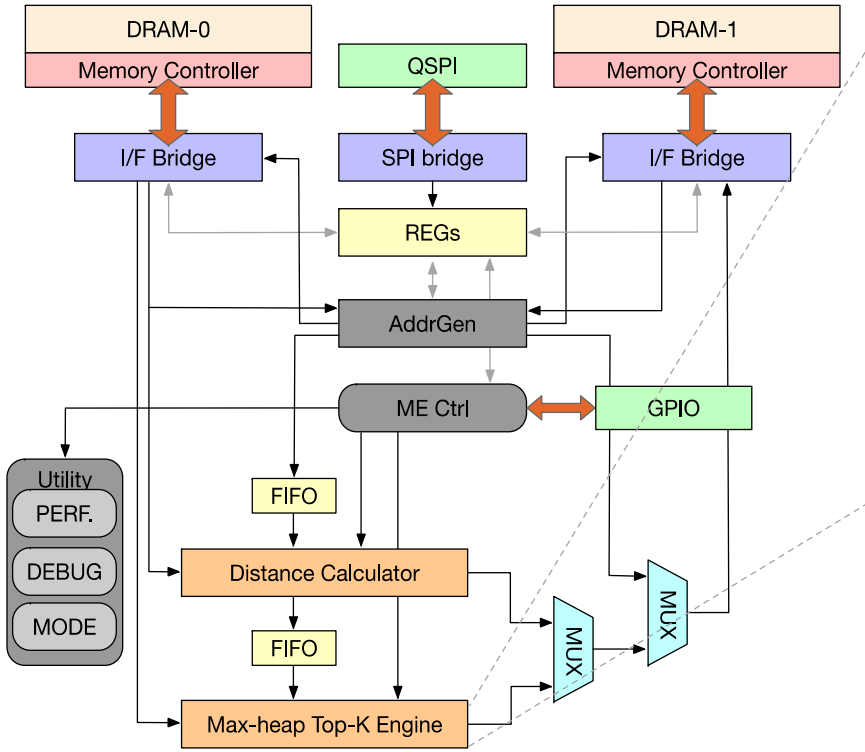
- Address Generator
  - Multi-mode to support different access patterns
  - Configurable via registers
  - Build-in performance evaluation mode & performance counter

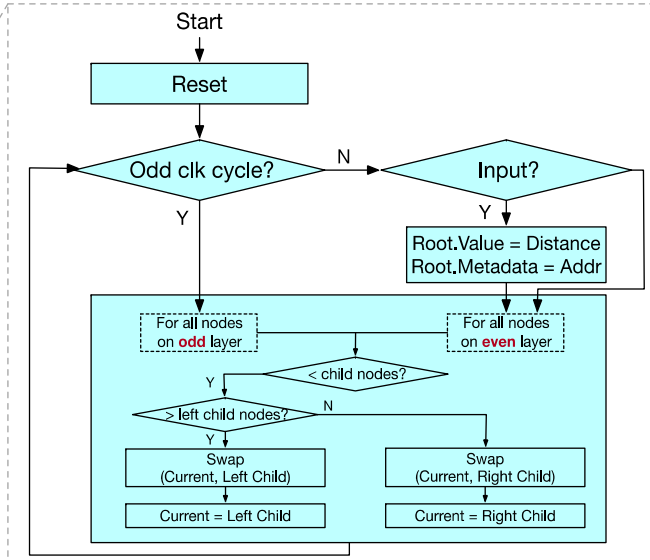
Match Engine Architecture (2)

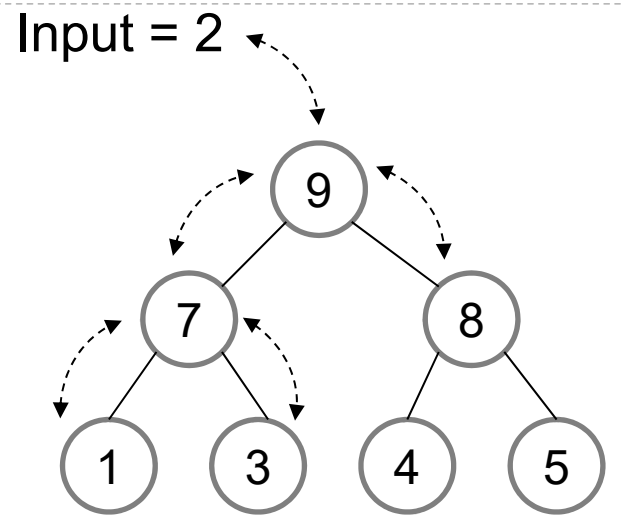


- Distance Calculator
  - Compare the similarity between input feature and query
  - Compute the Hamming distance of two 512-bit features
  - Filtered by rot of max-heap

Match Engine Architecture (3)

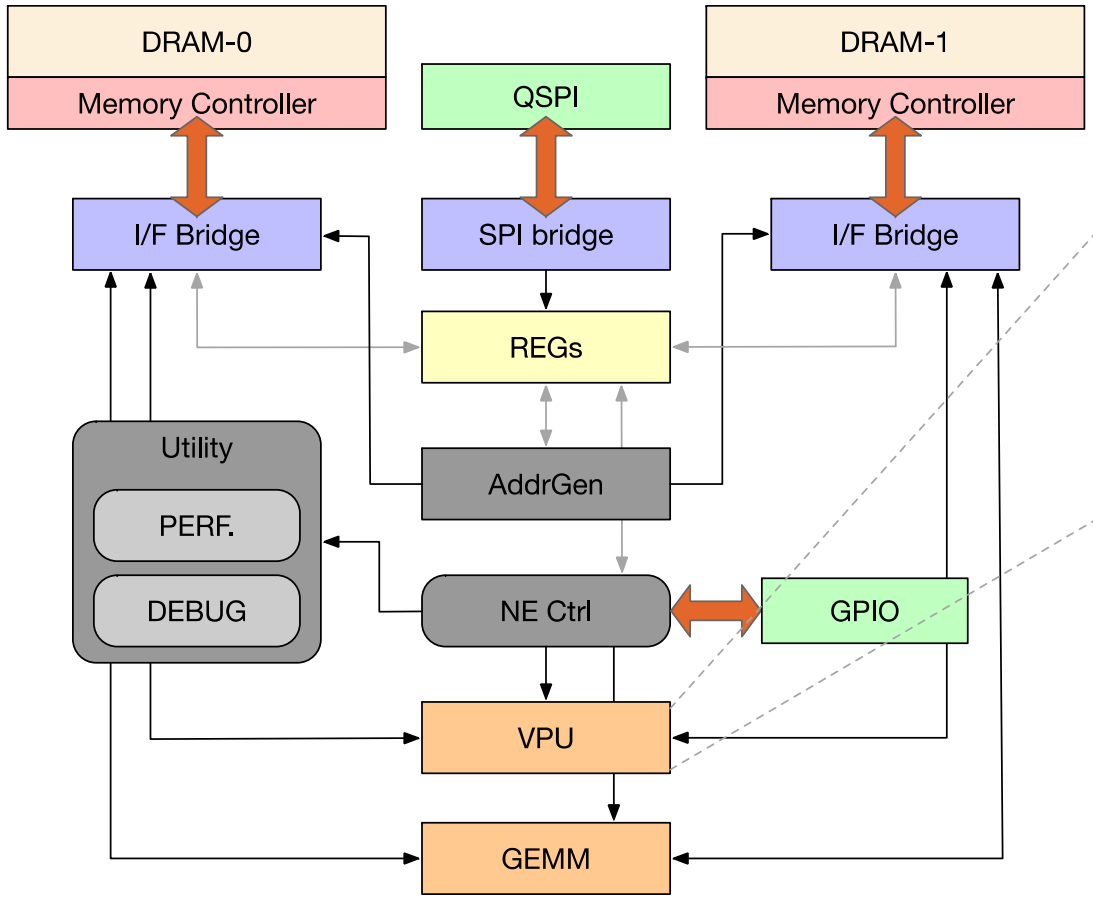


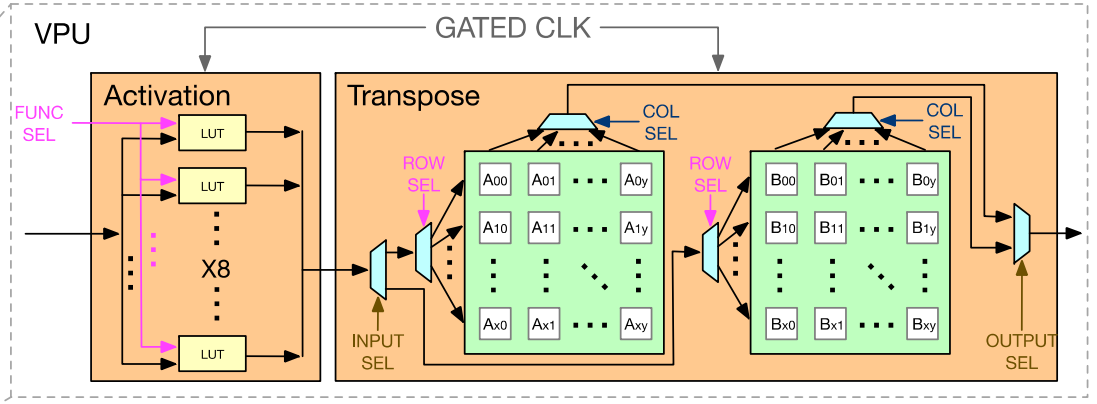




- Top-k Engine
  - Maintain a max-heap hardware block
  - Receives input every two cycles
  - Alternately heapifies nodes in odd layers and even layers
  - Stores the top-1000 matching results

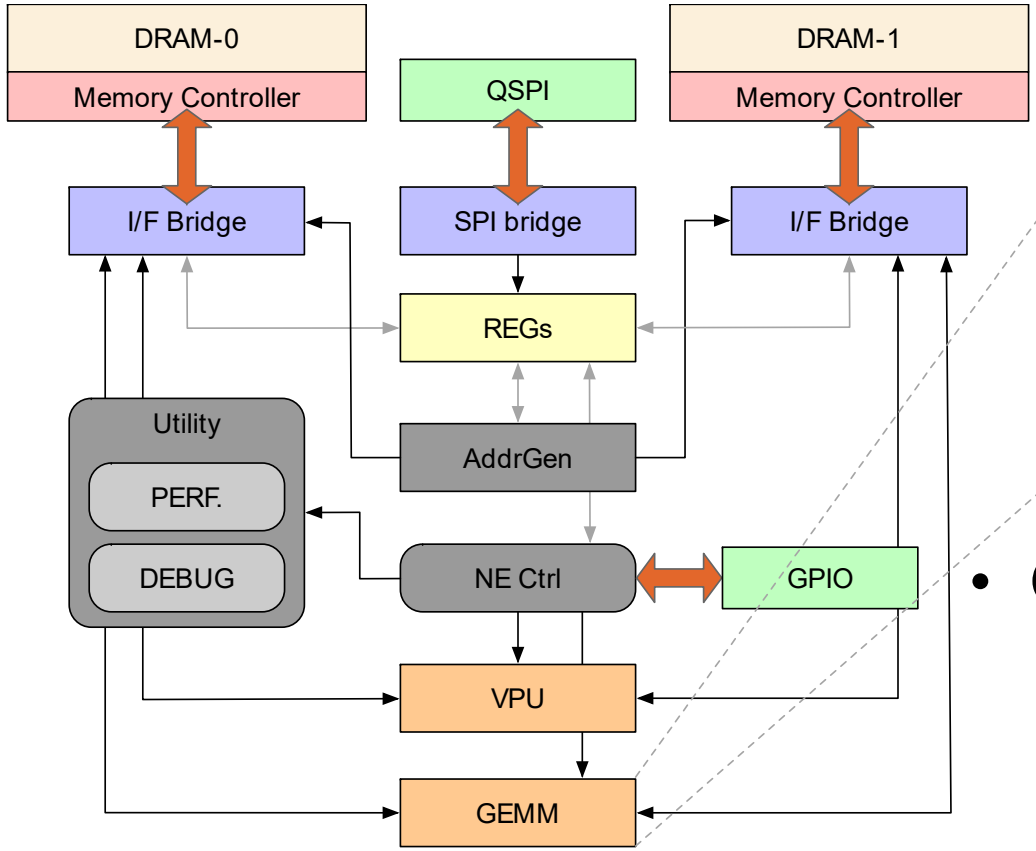
Neural Engine Architecture (1)

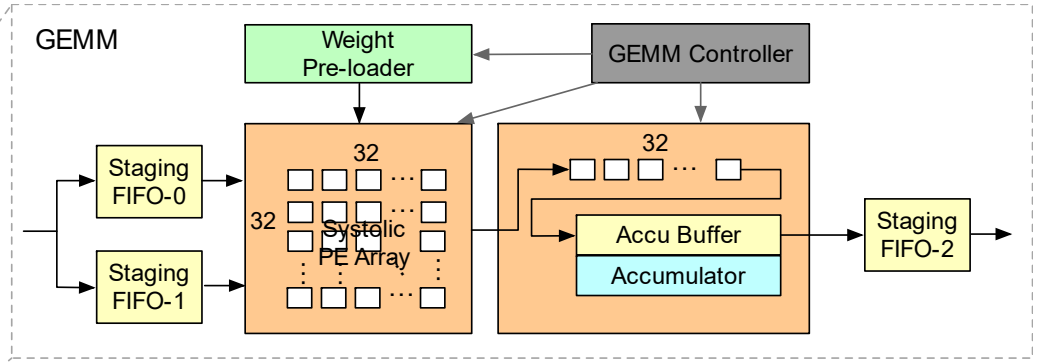




- Vector Process Unit
  - Activations
    - LUT based design
    - Supports GeLU & Exp
  - Transpose
    - Transpose 16x16 matrix with ping-pong array
    - Implemented with 2D register file array
    - Supports row-based writes and column-based reads

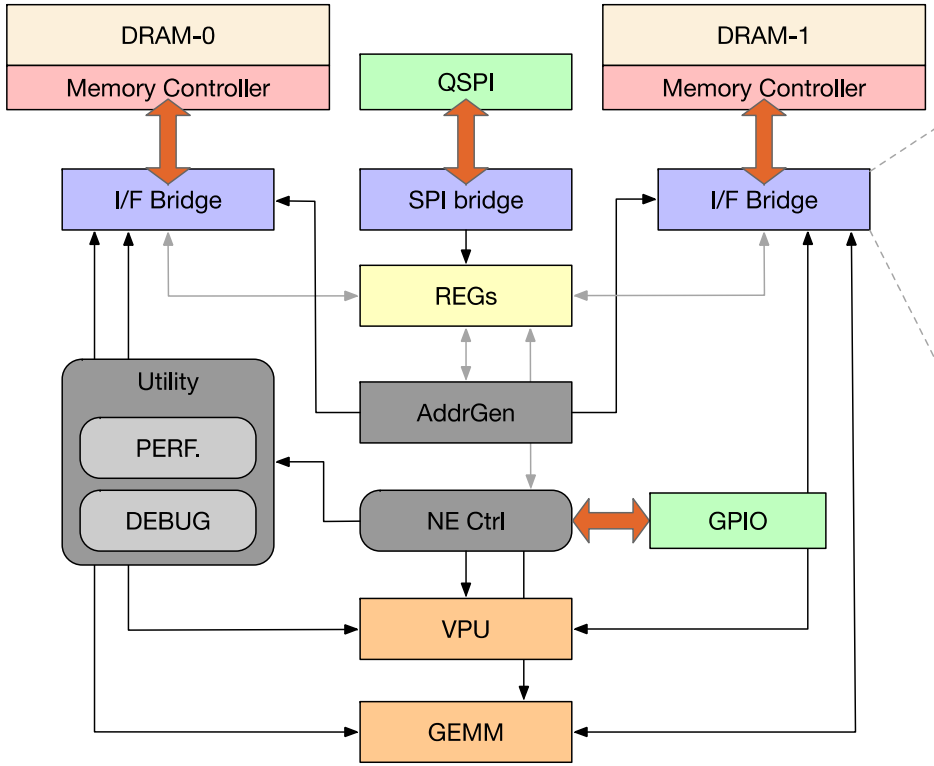
Neural Engine Architecture (2)

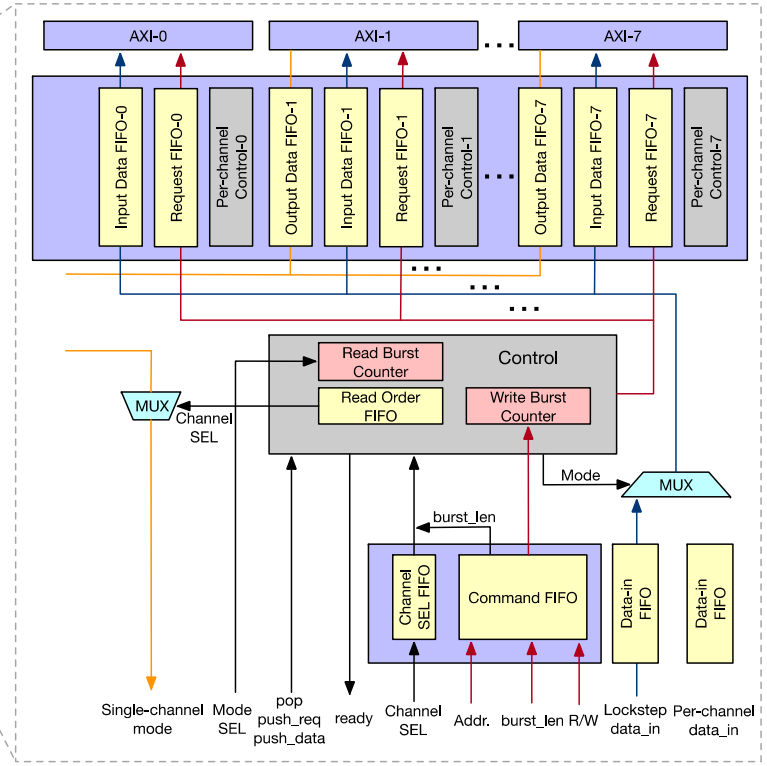




- GEMM
  - 32 by 32 systolic PE array (INT8)
  - Partial sum accumulated by the accumulator (INT32)
  - 600GOPS (@300MHz)

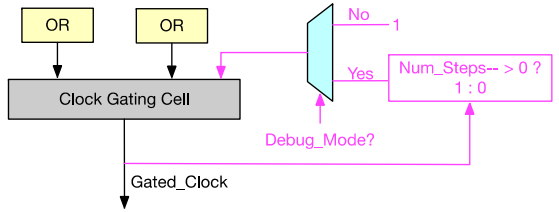
Interface Bridge & Debug





- Support both single-channel mode and lockstep-mode
- Read/write counter to support burst requests
- Support cycle-wise debug with clock gating

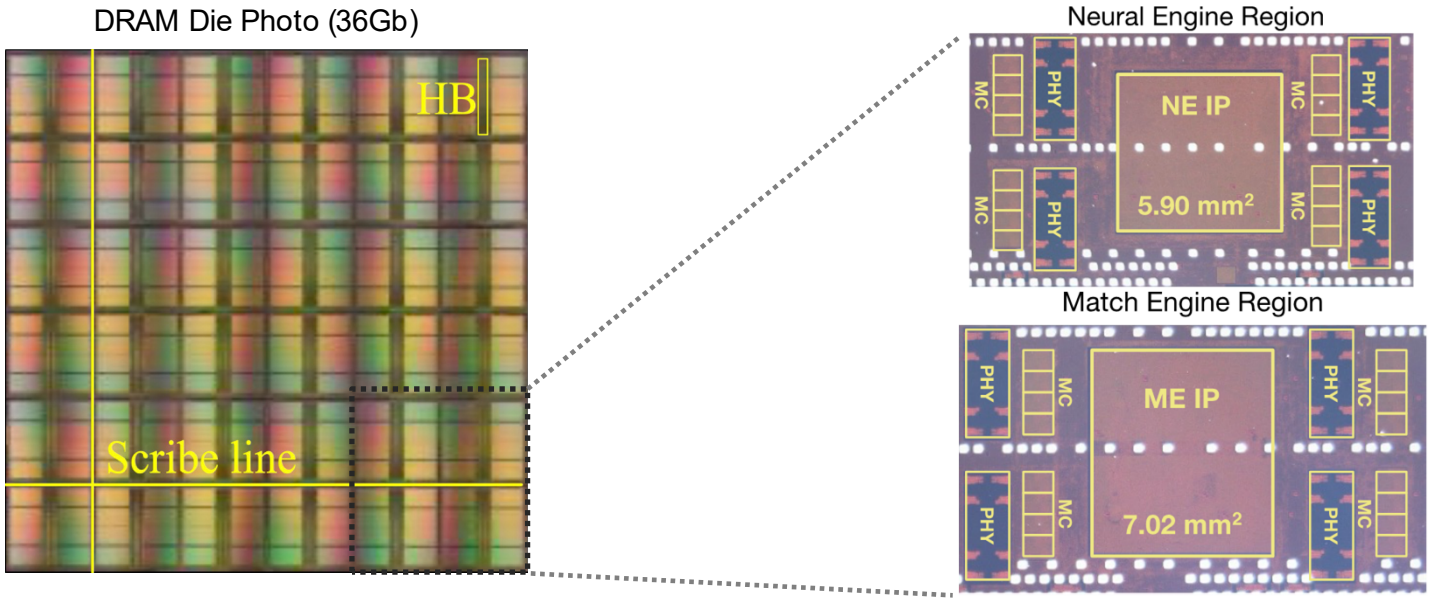
Lock-step DEBUG



Outline

- Motivation
- System and Chip Architecture
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- **Measurement Results**
- Conclusion

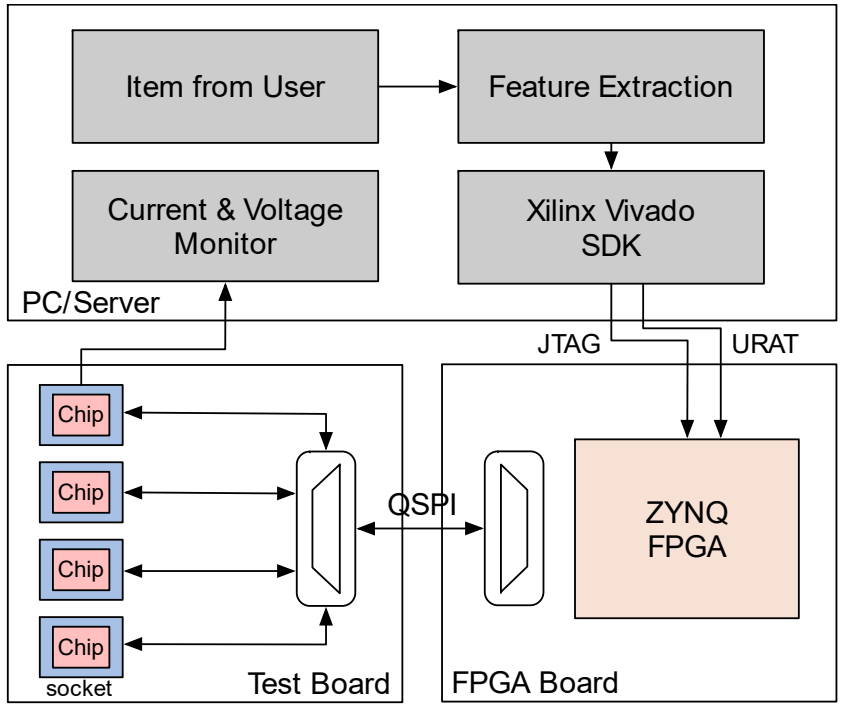
Die Photo and Summary

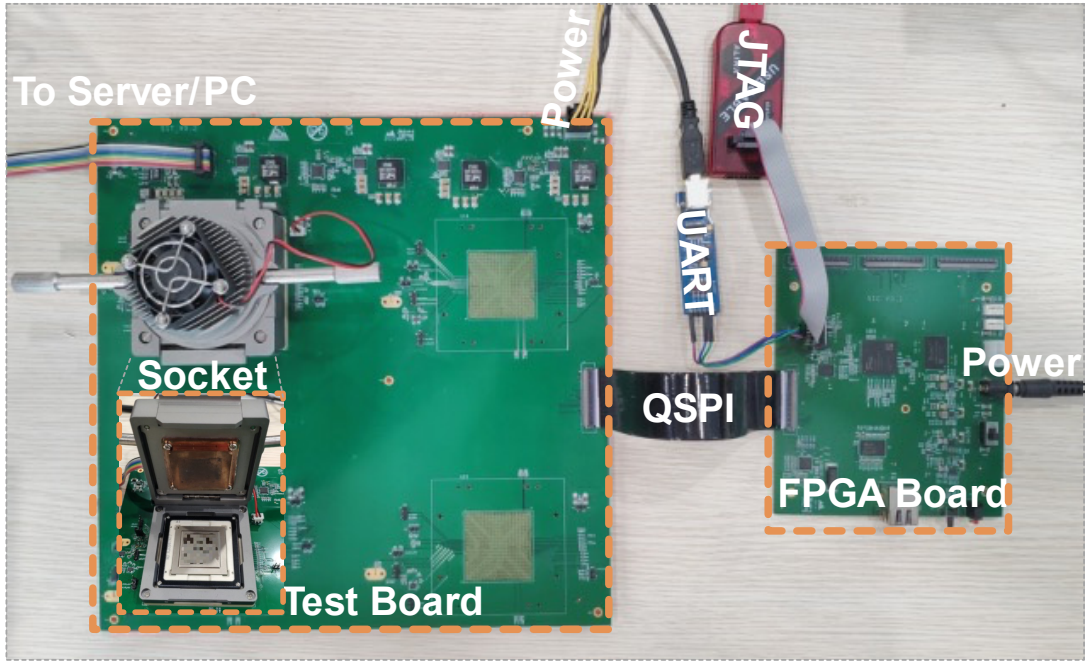


DRAM Die		
Technology	25nm	
Area	Total*	602.22 mm²
	Neural Engine	32 mm²
	Match Engine	32 mm²
Voltage	1.1 V	
Frequency (max)	150 MHz	
Power	300 mW per 1Gb	
Bandwidth**	153.60 GB/s / 1.38 TB/s	

Logic Die		
Technology	55nm	
Area	Total*	602.22 mm²
	Neural Engine	5.90 mm²
	Match Engine	7.02 mm²
# of MC	16 per IP	
Voltage	1.2 V	
Frequency	300 MHz	
Power	977.70 mW	
Precision	INT8	

Evaluation Platform





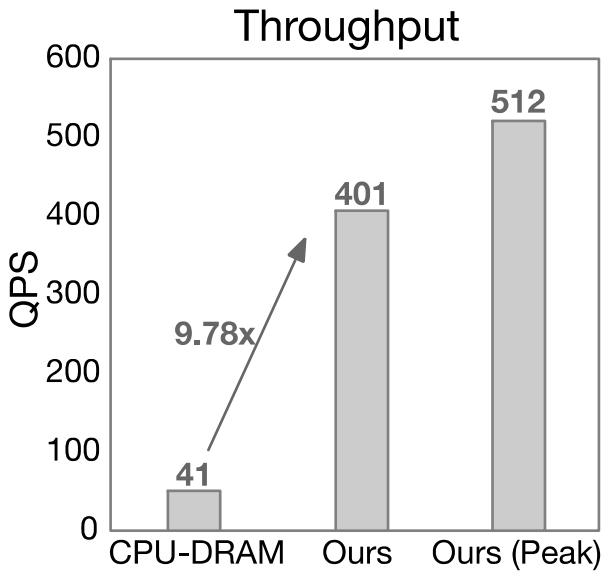
- Test board capable to mount up to 4 HB
- FPGA board responsible to write/read data and generate configuration to the chip register

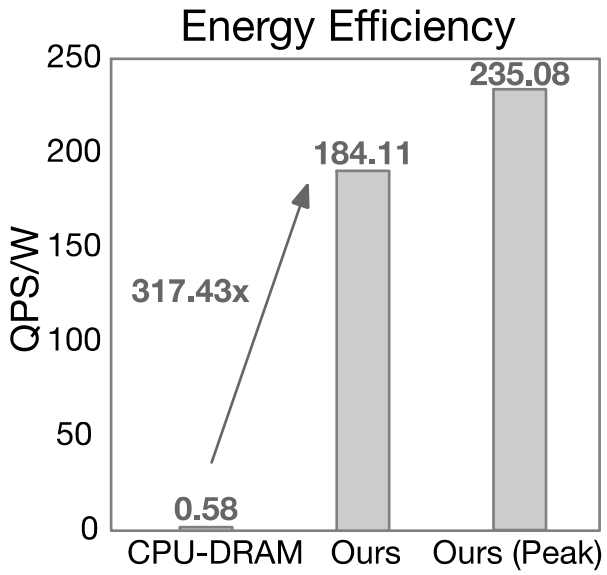
Performance

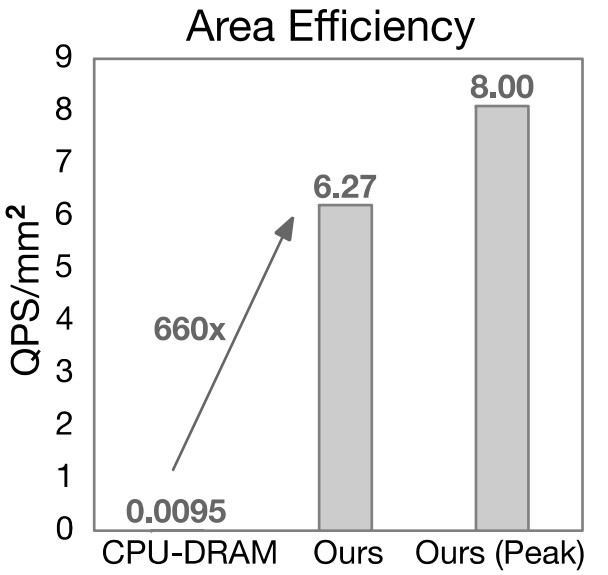
	CPU - DRAM*	This work
Logic Technology	14 nm	55 nm
Frequency	2.20 GHz	300 MHz
Area	4294 mm^2	64 mm^2 (4Gb)
Precision	INT8	INT8
Power**	70.17 W (TDP: 125 W)	2.178 W

** CPU: Intel Xeon Gold 5220@2.20GHz, tested on Pytorch*
*** CPU power measured by PyRAPL*

- Measurement vs. Peak: **~20%** initialization and memory subsystem overhead







QPS: Queries per second

Comparison

	2D CIM *	2D PNM **	2.5D PNM ***	3D TSV (Hybrid) ****	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**

#Estimated

- High off-chip bandwidth
- High bandwidth per capacity
- Low energy per 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

Outline

- Motivation
- System and Chip Architecture
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- Measurement Results
- Comparison
- **Conclusion**

Conclusion

- Memory-bound application can significantly benefit from process-near-memory and computing-in-memory
- A 3D Logic-to-DRAM Hybrid Bonding Chip with Process-Near-Memory Engine for Recommendation System is demonstrated featuring with:
  - High-bandwidth and energy-efficient memory with hybrid bonding
  - High-throughput streaming processing units for matching and ranking
  - **2.4GB/s/mm^2** bandwidth density and **0.88pJ/bit** energy consumption
  - **~10x** performance improvement and over **300x** energy-efficiency improvement over conventional CPU+DRAM system