SAMSUNG

Real-world Implementation and Future of AI Acceleration Systems based on Processing-in-Memory

Byeongho Kim, Ph.D.
Samsung Electronics

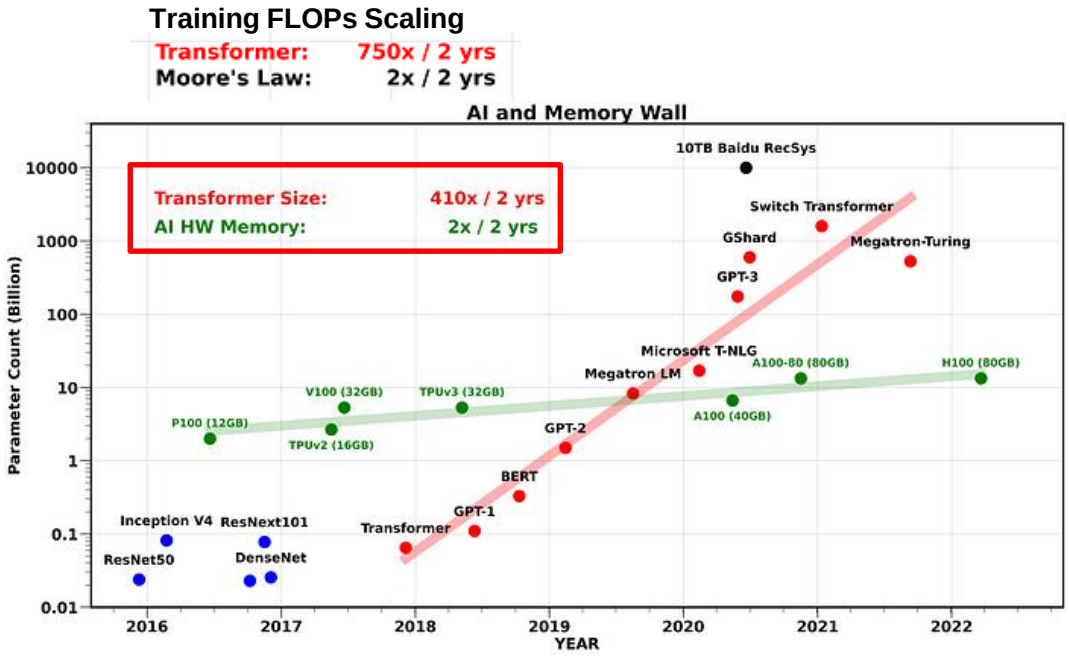
SAMSUNG

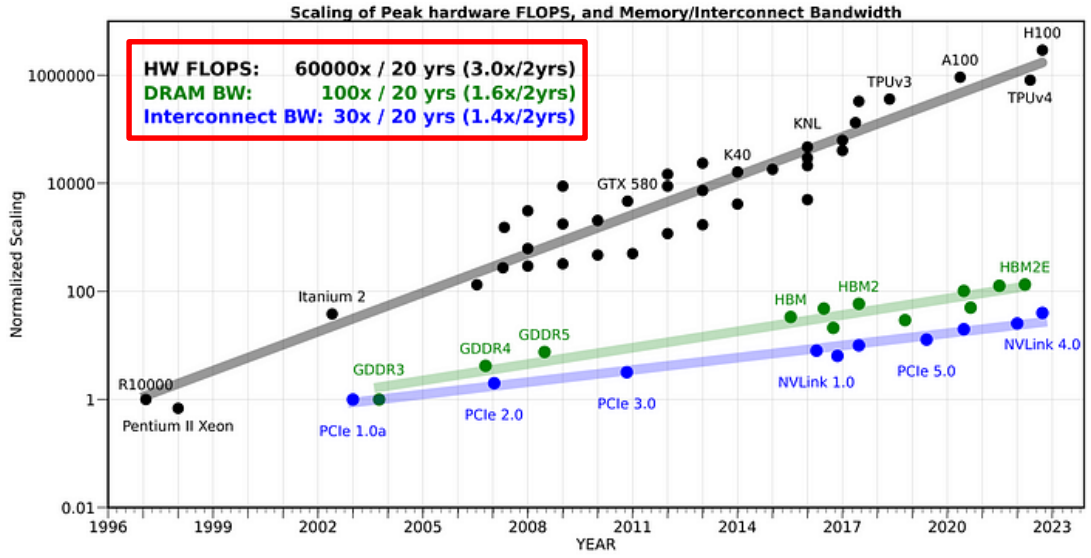
Outline

- Generative AI and Memory Requirement
- Memory Solutions for AI
  - New Memory Hierarchy
- Processing-in-Memory
  - PIM and PNM
  - HBM-PIM and other commercial PIM technology
  - Challenges on commercializing PIM
  - Next: LPDDR-PIM
- Summary

Large-scale AI and Memory Wall

- New applications have already reached **memory wall**
  - AI applications are bottlenecked by communication overhead rather than compute.
  - Scaling rate of AI model (FLOPS & Parameters) far exceeds that of memory bandwidth/capacity.

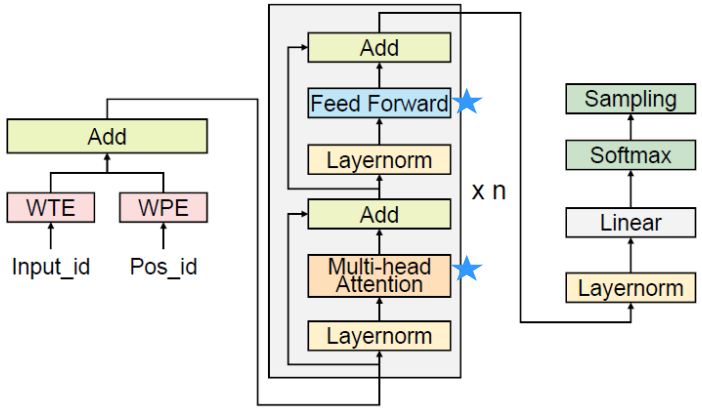


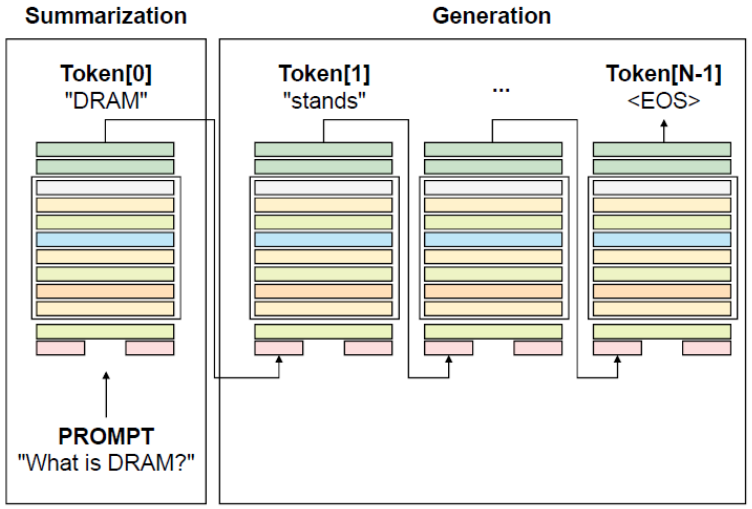


*https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

GPT Characteristics

- Transformer decoder-based structure and two phases
  - Generation stage dominates the execution time.





★: Layer which includes GEMV in Generation stage

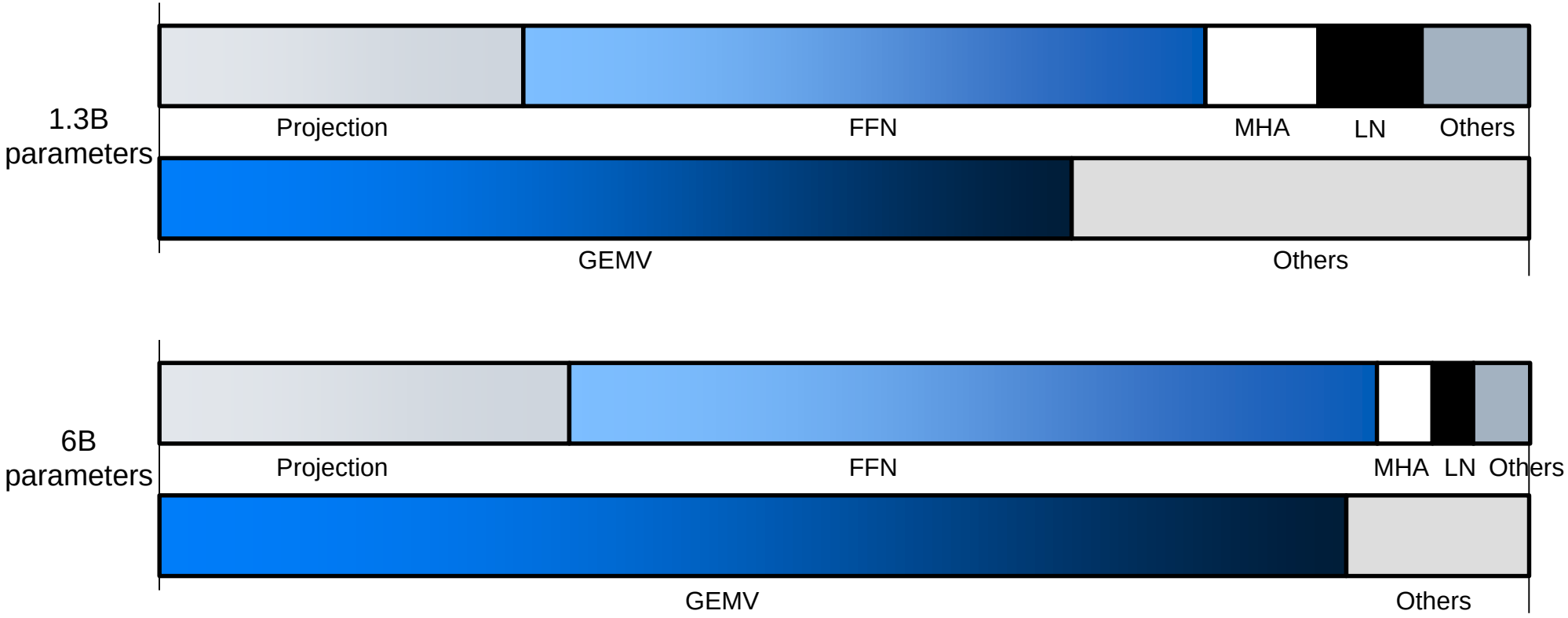
# Output tokens (L_{out}) \ # Input tokens (L_{in})	2	8	32	128	512	2,048
2,048	99.9	99.9	99.9	99.9	99.8	99.2
512	99.8	99.8	99.8	99.8	99.2	96.8
128	99.2	99.2	99.2	99.1	96.7	88.2
32	96.9	96.8	96.7	96.3	87.9	64.5
8	87.5	87.4	87.0	85.6	62.1	29.1
2	50.0	49.8	48.9	45.8	19.0	5.5

Generation stage time ratio in GPT-3 execution time*

*J. Choi, Unleashing the Potential of PIM: Accelerating Large Batched Inference of Transformer-Based Generative Models, IEEE CAL, 2023

GPT Characteristics

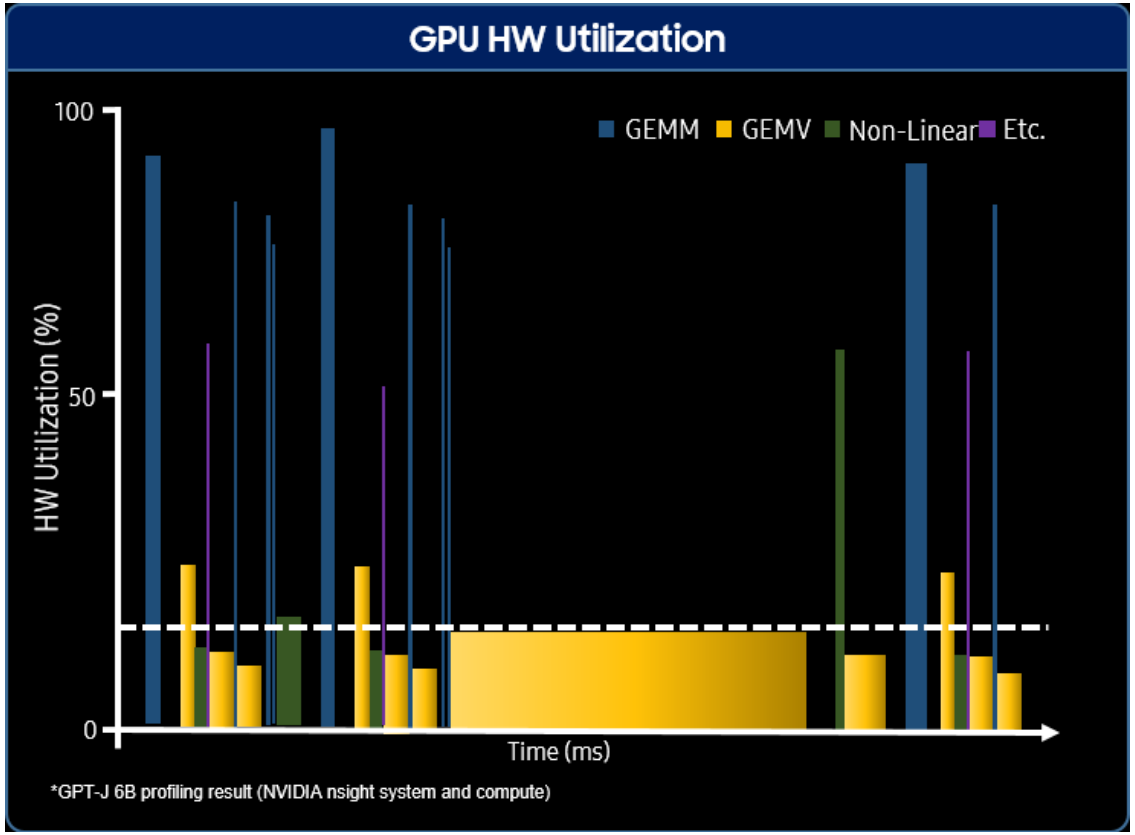
- GEMV portion can be 60–80% of total generation latency



*Profiling result is measured in A100 System (Fastertransformer + GPT-J, FP16, Input/Output token:32/32)
GPT-j: Google JAX framework

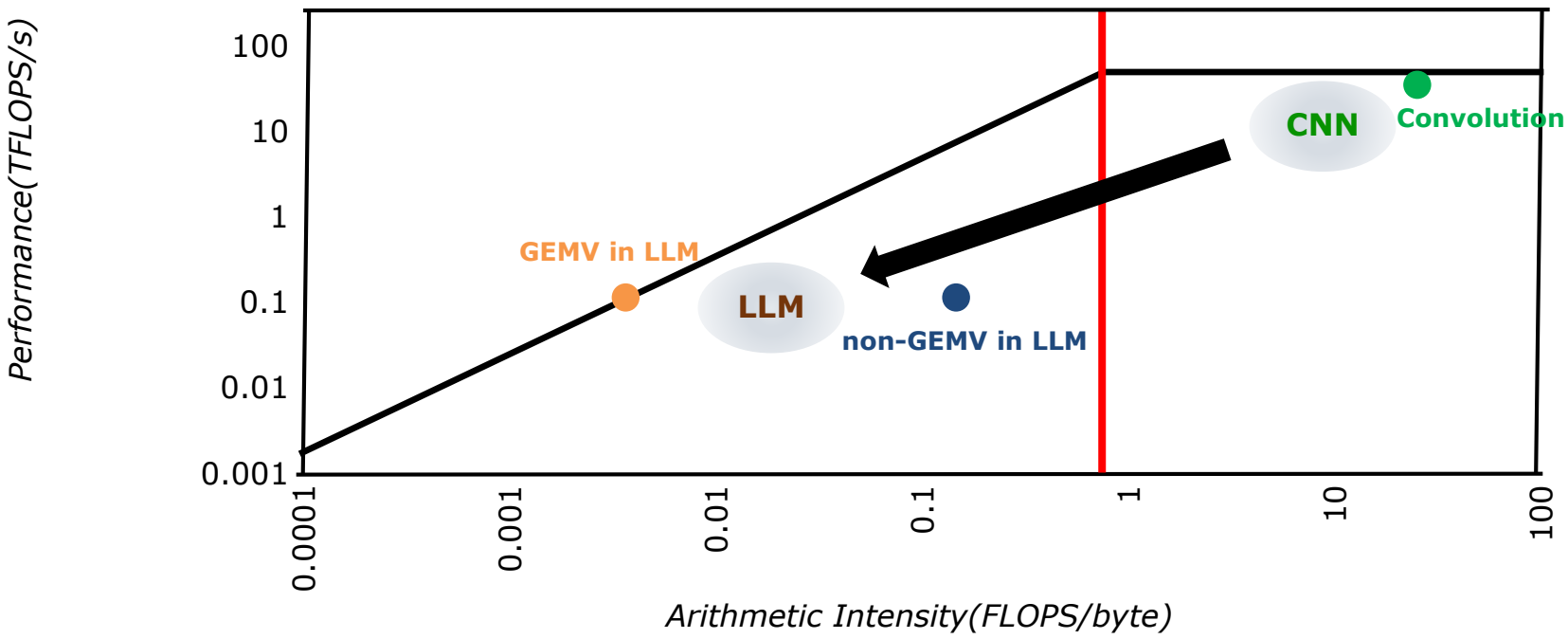
GPT H/W Utilization Breakdown

- As # of tokens increases, GEMV dominate inference time
- Max. utilization is limited by memory bandwidth on GEMV



Memory Solution for GPT

- AI needs Higher Bandwidth Memory
  - Paradigm shift in AI: CNN → LLM(GPT)



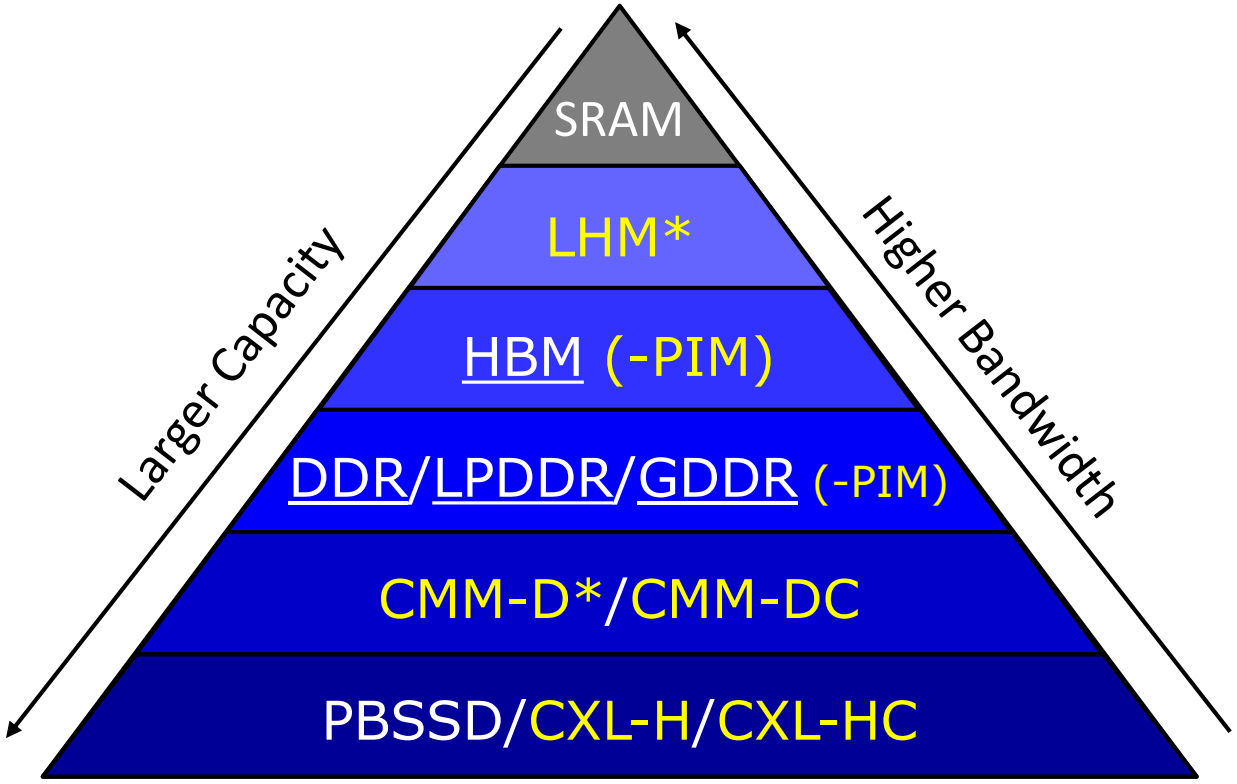
SAMSUNG

Outline

- Generative AI and Memory Requirement
- Memory Solution for AI
  - Memory Hierarchy
- Processing-in-Memory
  - PIM and PNM
  - HBM-PIM and other commercial PIM technology
  - Challenges on commercializing PIM
  - Next: LPDDR-PIM
- Summary

High Bandwidth Memory Solutions for AI

- HBM provides the Highest Bandwidth in the Market
  - Various memory solution can be used depending on System



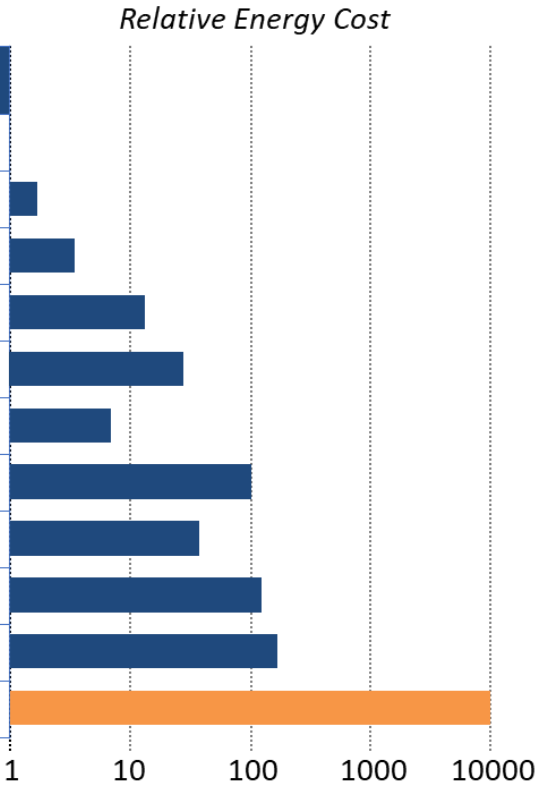
- Yellow-color memories are newly proposed.
* LHM: Low-power High-bandwidth Memory (Ex. LLW)
* CMM: CXL Memory Module

- The products underlined are currently commercially available.

Besides Bandwidth, Energy also Matters

- Another Limitation of Von Neumann Architecture
  - DRAM consumes large energy to transfer data

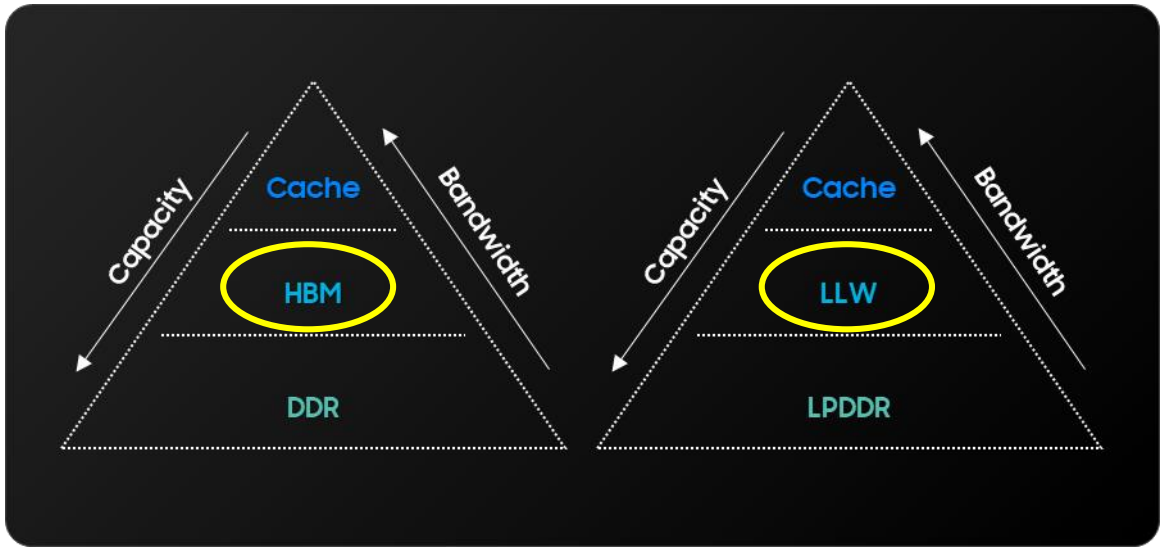
Operation	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

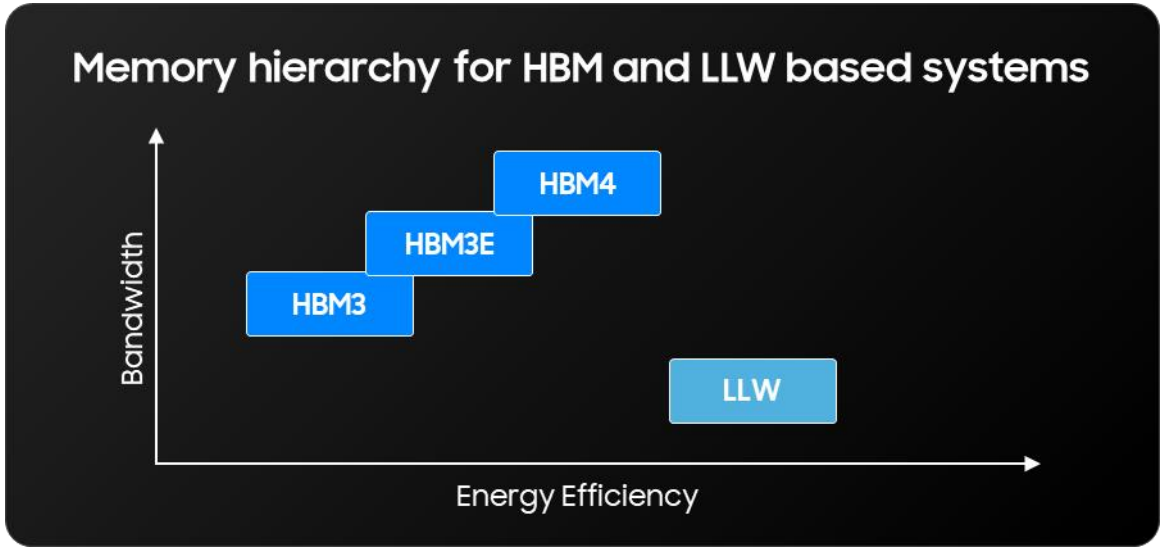


Source : Computing's Energy Problem (and what we can do about it) (ISSCC'14)

Near Memory Solution

- Each near memory solution has its role in the memory hierarchy
  - High Bandwidth Memory (HBM) **for HPC**
  - Low-Latency Wide-IO (LLW) **for Mobile**





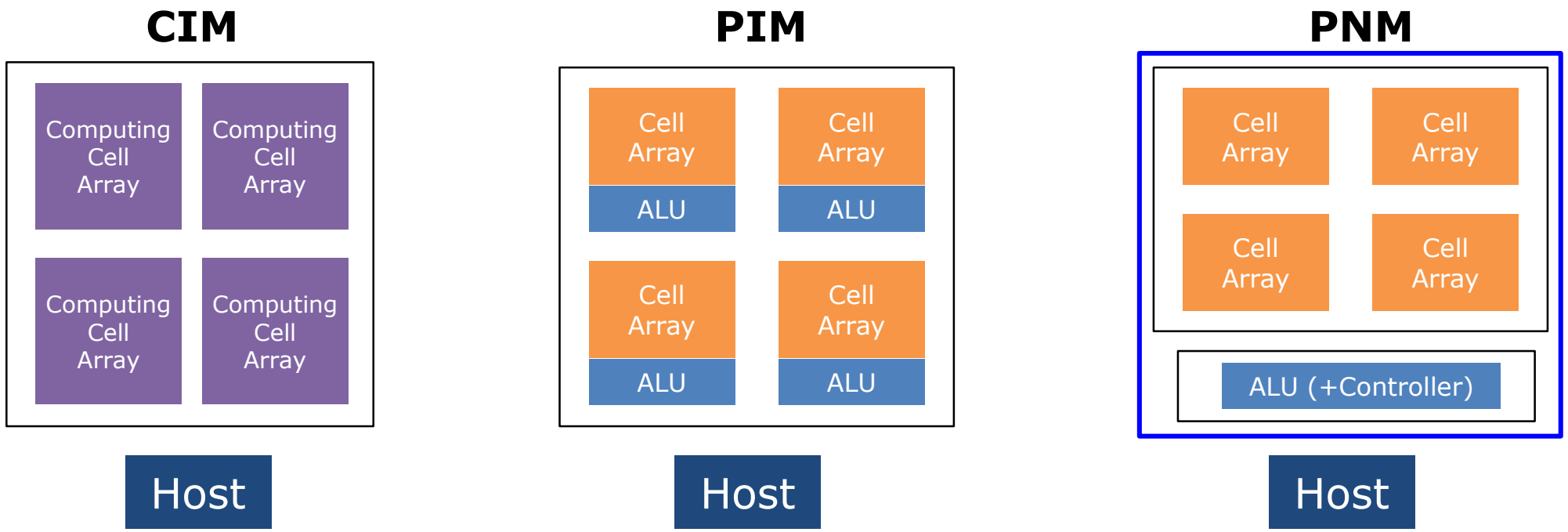
SAMSUNG

Outline

- Generative AI and Memory Requirement
- Memory Solution for AI
  - Memory Hierarchy
- Processing-in-Memory
  - PIM and PNM
  - HBM-PIM and other commercial PIM technology
  - Challenges on commercializing PIM
  - Next: LPDDR-PIM
- Summary

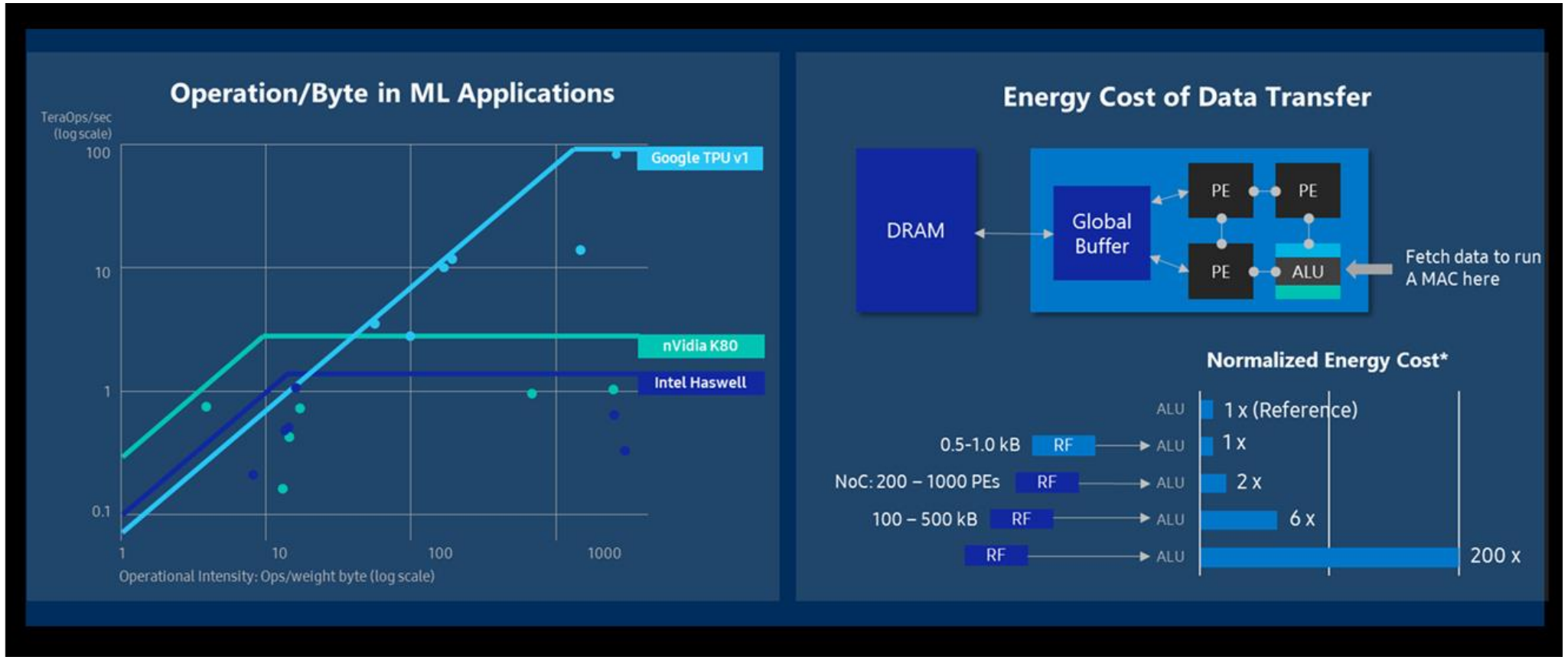
Intelligent Memory and Types

- Three distinct categories in this talk
  - CIM: use memory array as a processing unit
  - PIM: use embedded logic near memory array as a processing unit
  - PNM: use an additional chip for processing inside a memory package or a set



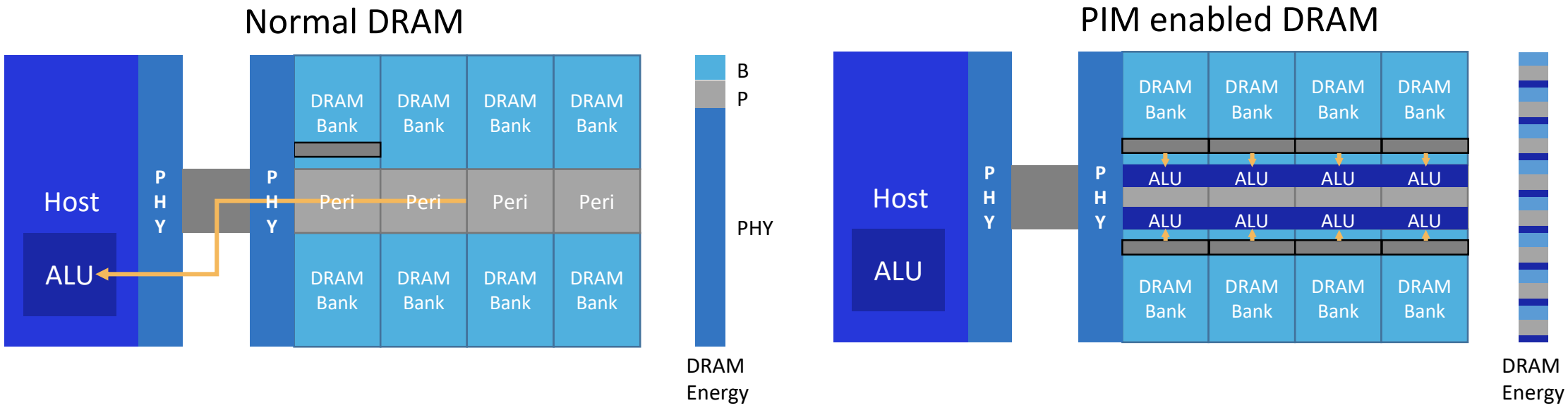
PIM: Renewed Interest

- ML workloads w/ growing model size need more frequent DRAM accesses, limiting performance and dominating energy consumption



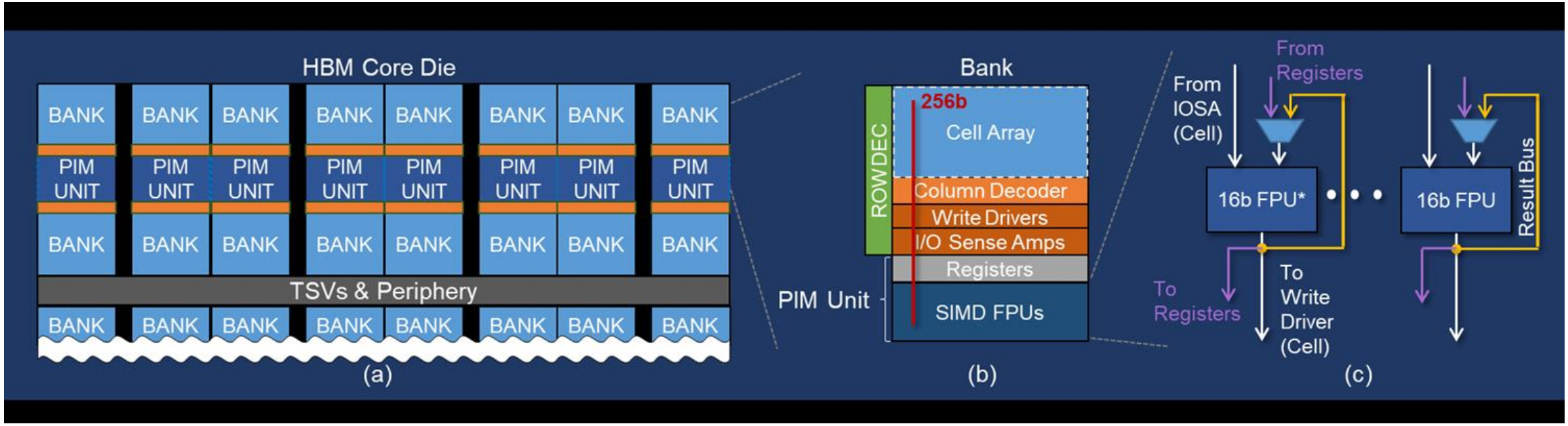
Processing-in-Memory (PIM)

- Utilize internal memory bandwidth by bank-level parallelism
  - Proposed by major DRAM vendors (Samsung & SK-Hynix)



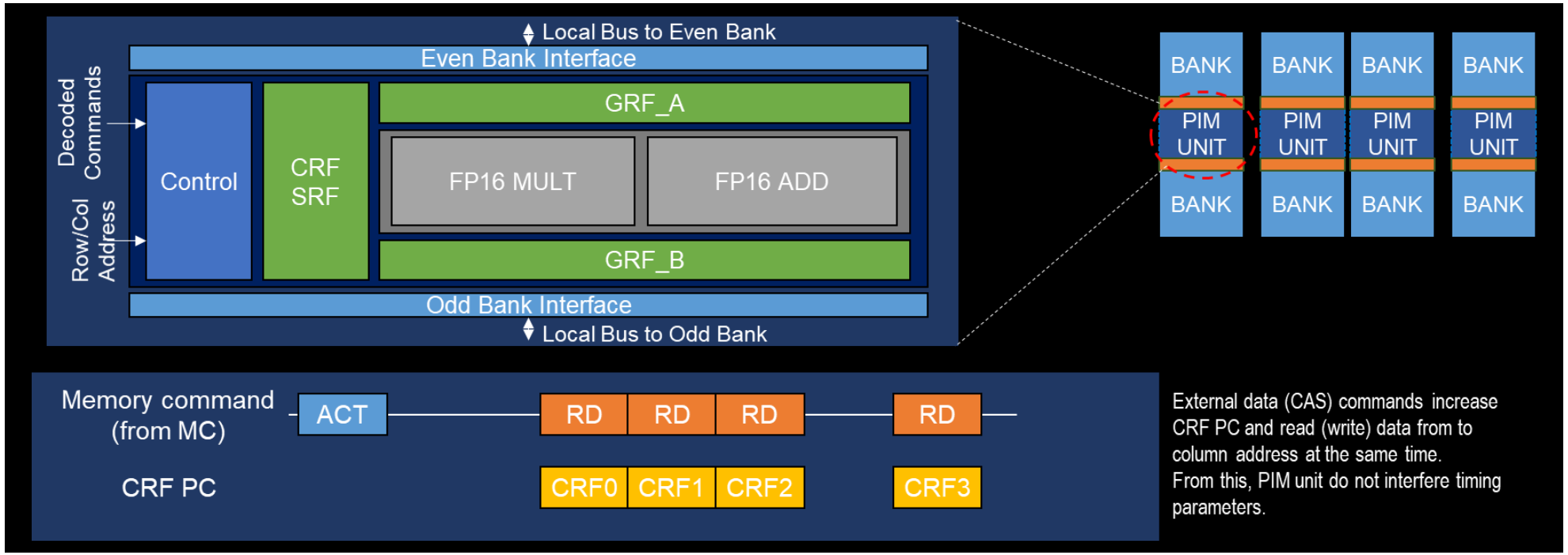
Overview of PIM Architecture

- High on-chip compute bandwidth w/o changing DRAM core circuitry
  - Place SIMD FPU at bank IO boundary
  - Exploit bank-level parallelism: access multiple banks/FPUs in a lockstep manner
- Expose high on-chip bandwidth of standard DRAM to processors
  - Build on industry standard DRAM interfaces and preserve deterministic DRAM timing
  - i.e., a DRAM RD/WR command triggers execution of a PIM instruction



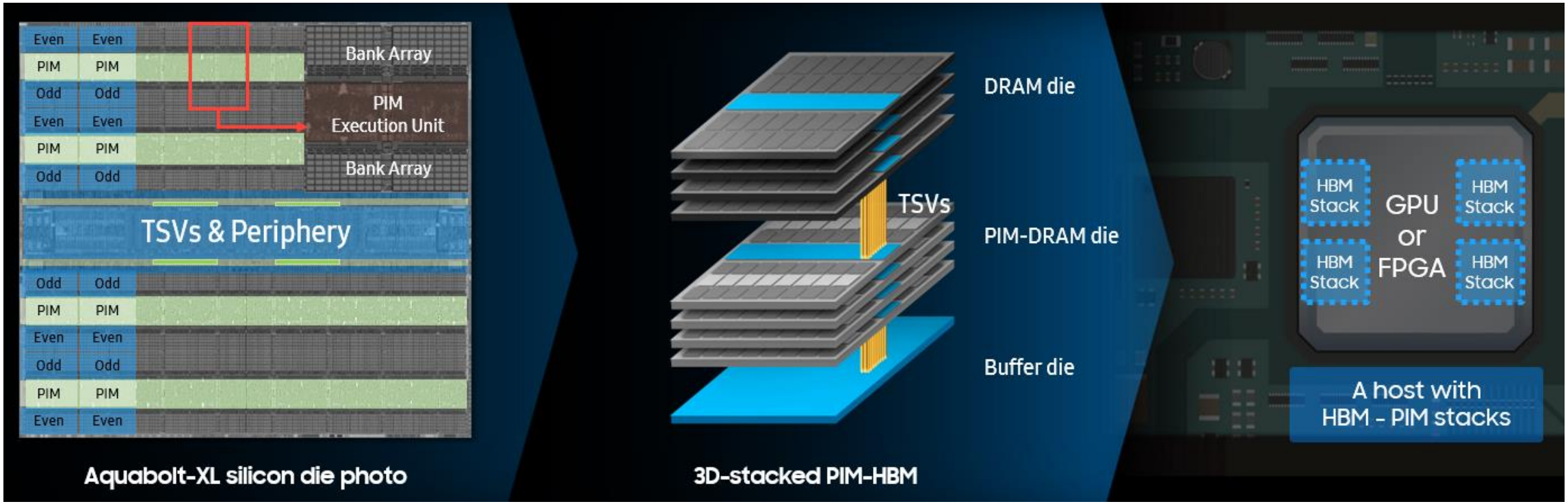
PIM-DRAM: Microarchitecture

- Consist of three major components with DRAM local bus interface:
  - A 16-lane FP16 SIMD FPU array: a pair of 16 FP16 multipliers and adders
  - Register files: Command, General, and Scalar register files (CRF, GRF, and SRF)
  - A PIM unit controller (fetch and decode, controls pipeline signals, forward)



HBM-PIM Implementation

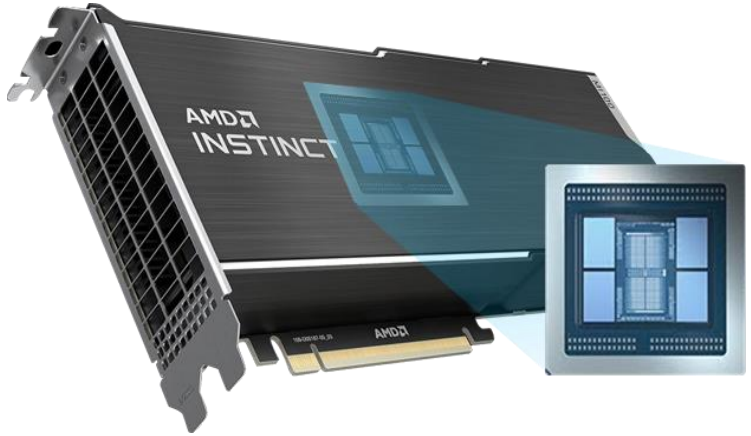
- Based on a commercial HBM2 design
  - Off-chip and on-chip bandwidth: 1.23 TB/s and 4.92 TB/s



HBM-PIM Powered Systems

- Collaborated with two system-board companies

AMD MI50/100-PIM GPU



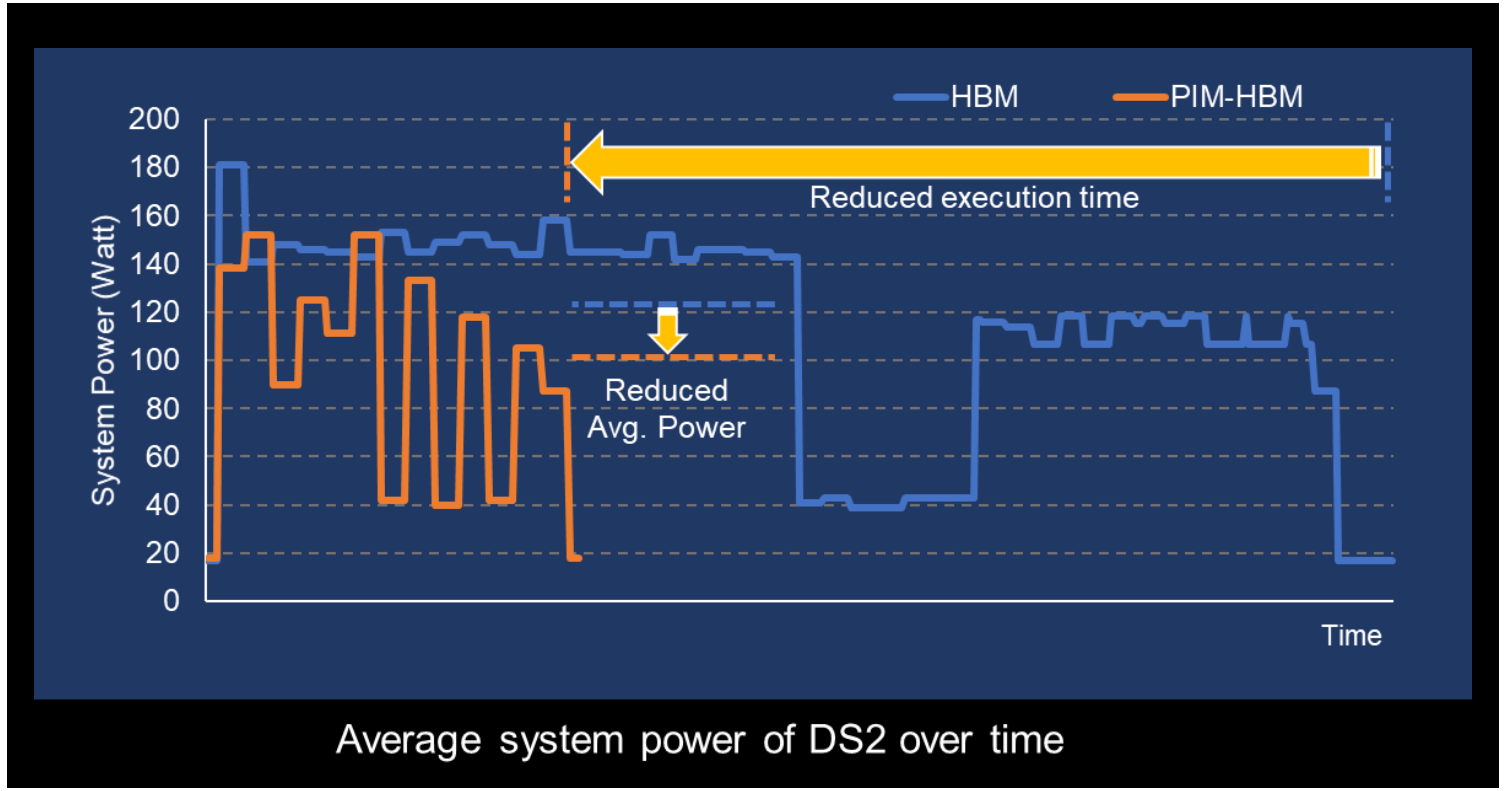
· Capacity, 24GB (4 cubes)
· PIM performance, 4.9 TFLOPS

Xilinx Alveo U280 FPGA

· Capacity, 12GB (2 cubes)
· PIM performance, 1.9 TFLOPS

HBM-PIM Evaluation – Performance/Power/Energy

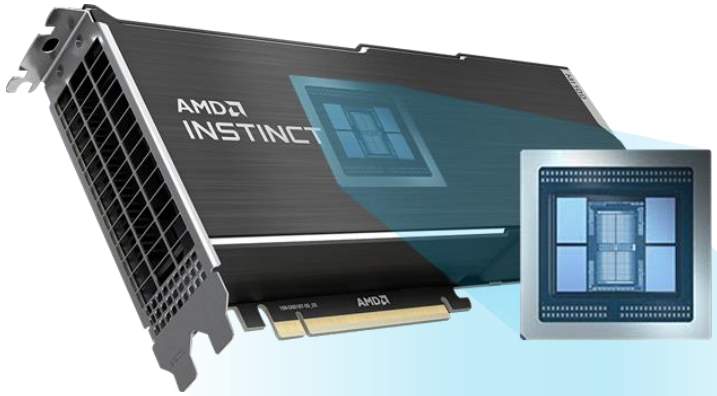
- PIM-HBM improves energy efficiency by
  - both shorter execution time and lower average power consumption.



HBM-PIM Cluster

- Installed 96 AMD MI100 GPUs fabricated with HBM-PIM

AMD MI100-PIM GPU



· Capacity, 24GB (4 cubes)
· PIM performance, 4.9 TFLOPS

Server node

· 8 MI100-PIM GPUs per node

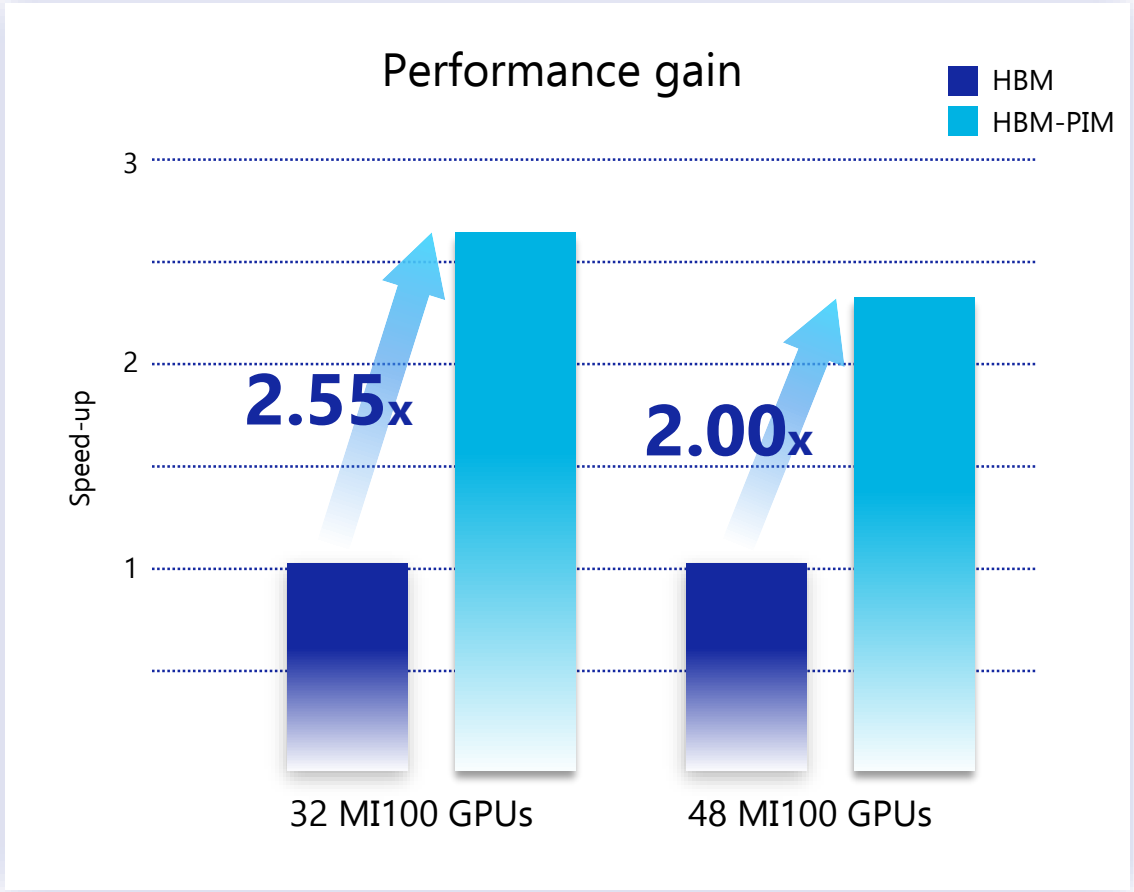
HBM-PIM cluster

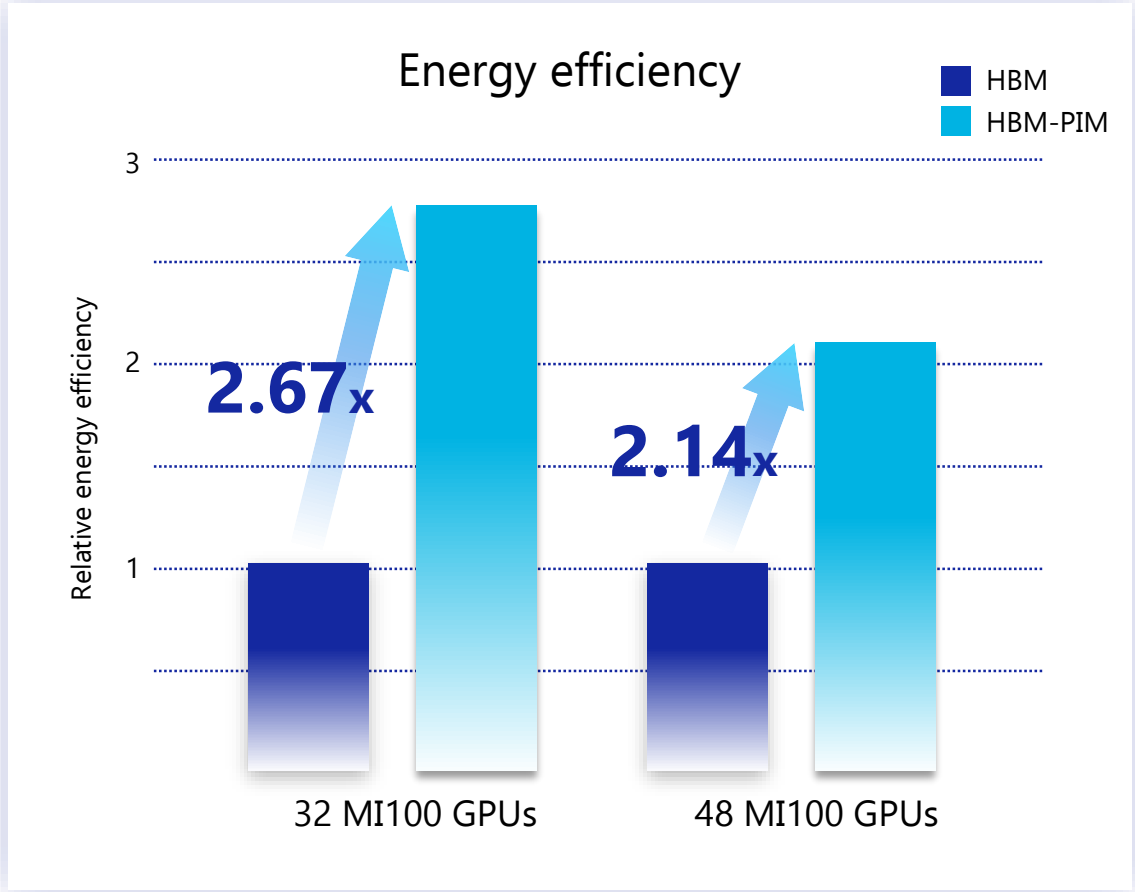
· 12 nodes interconnected
through 200G InfiniBand network
· Total 96 MI100-PIM GPUs in a cluster

Real-world Implementation and Future of AI Acceleration Systems
based on Processing-in-Memory

Evaluation Results on MoE Model

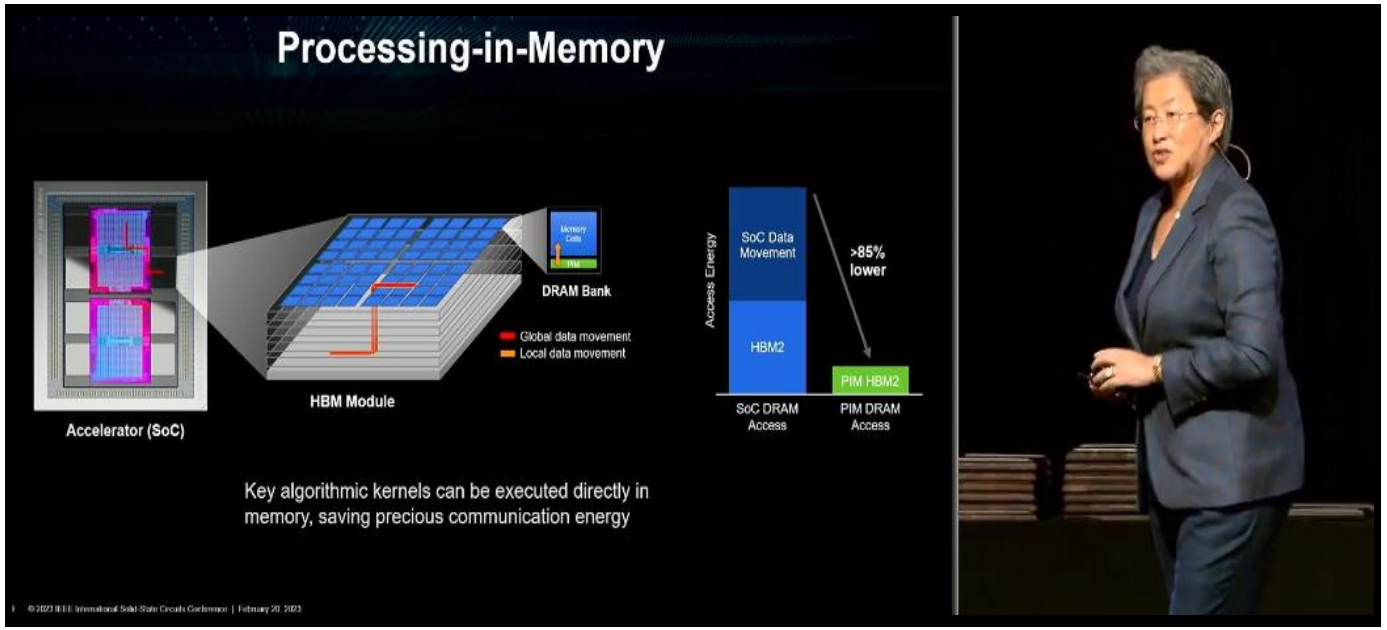
- Performance 2x and Energy efficiency 3x compared to normal GPU





PIM Value on GPT

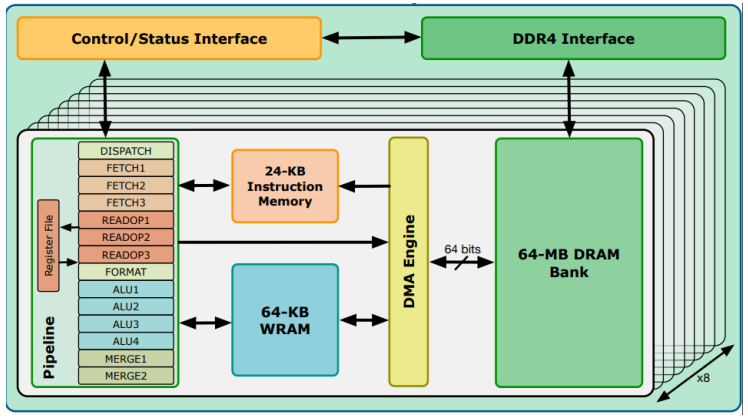
- OpenAI's focus is on developing new AI technologies and pushing the boundaries of what is possible with AI, so it's possible that they will explore the use of PIM technology at some point in the future.
- AMD, access energy can be improved by execute the main algorithm kernel directly in memory
  - Dr. Lisa Su, Energy reduction by 85% compared to using conventional HBM

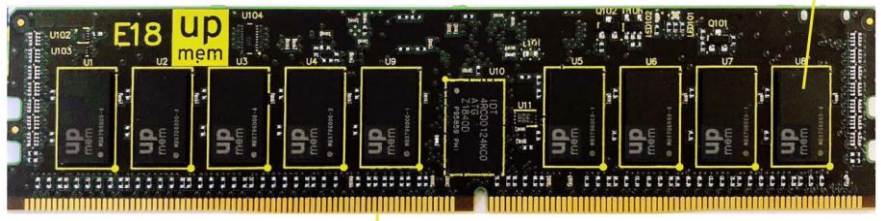


Real-world Implementation and Future of AI Acceleration Systems
based on Processing-in-Memory

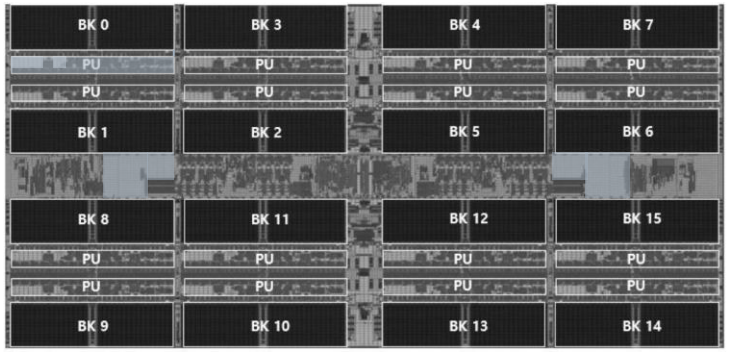
Case Study on Commercial DRAM Technology

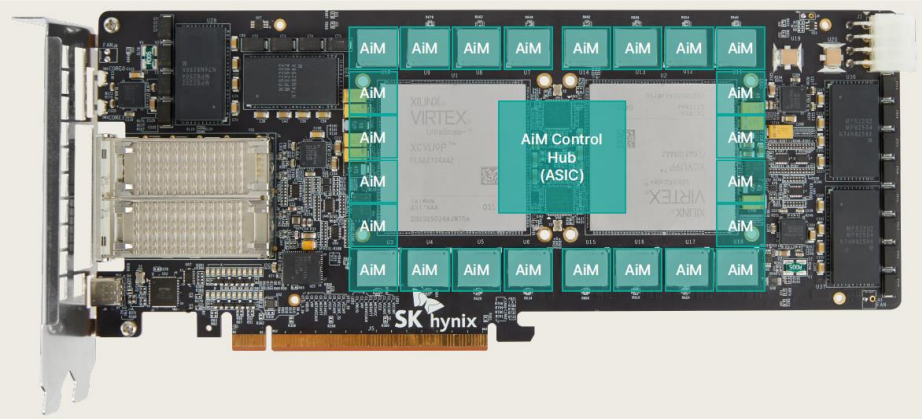
- UPMEM
- Hynix GDDR6
  - AiMX (PIM cluster) based on Hynix GDDR6





Reference: The true Processing In Memory accelerator, Hotchips '19





On-Device AI & LPDDR-PIM

- Growing Importance of On-Device AI
  - Data center costs and power consumption are increasing
  - Privacy concerns are rising as sensitive data is transmitted to the cloud for processing
  - Network connectivity is not always reliable or available, particularly in remote areas
- LPDDR-PIM improves battery life and prevents memory over-provisioning just for bandwidth

Benefits of On-Device AI

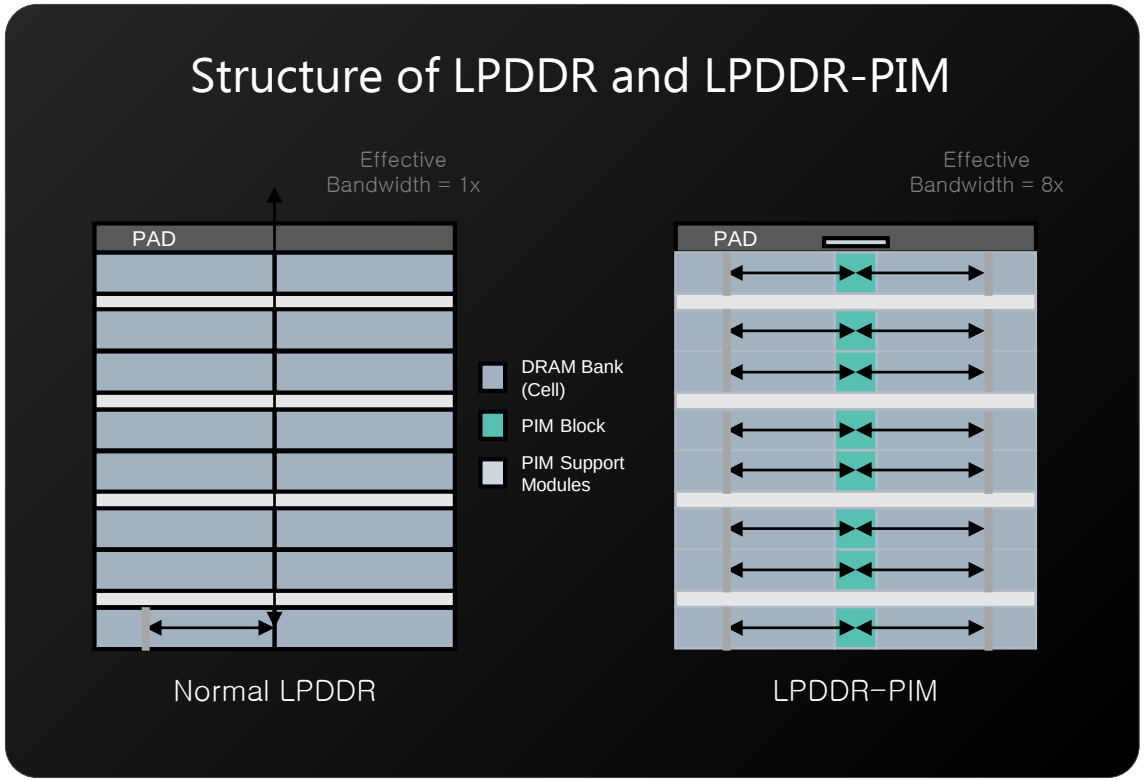
Privacy Protection

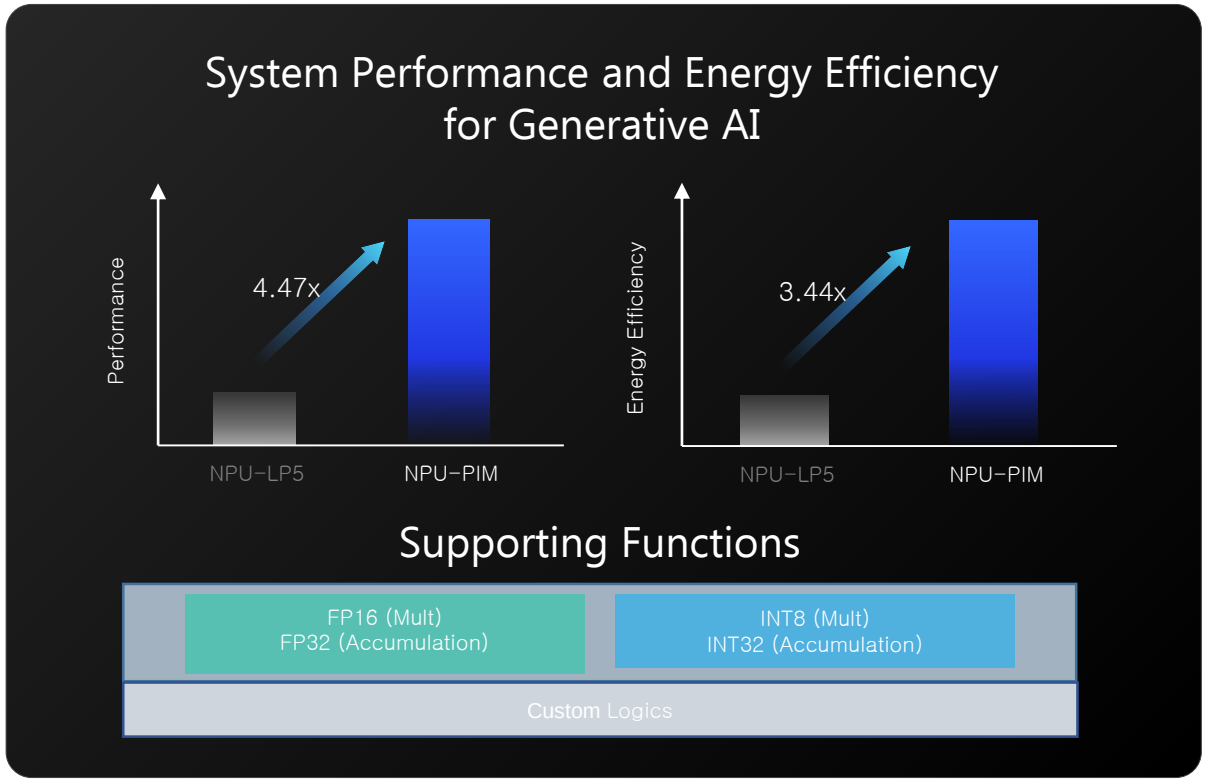
Low Latency

Low Power & Low Cost

LPDDR-PIM Introduction

- LPDDR-PIM improves performance and energy efficiency of the system with in-DRAM processing
  - Performance: Utilizes up to 8x higher in-DRAM bandwidth by bank parallel operation
  - Energy Efficiency: Reduces data movement energy by utilizing in-DRAM processing unit

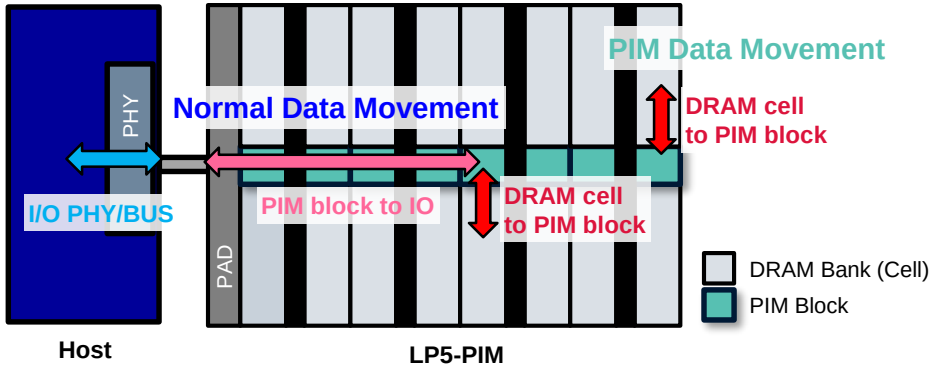


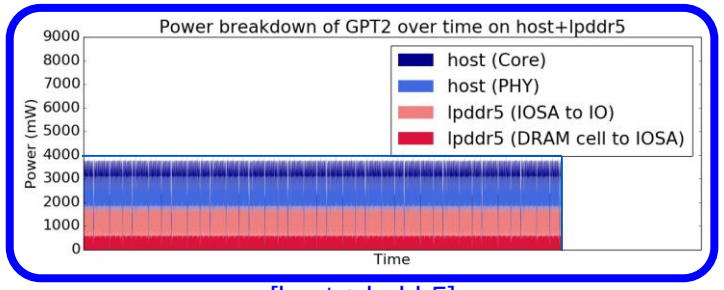


Real-world Implementation and Future of AI Acceleration Systems based on Processing-in-Memory

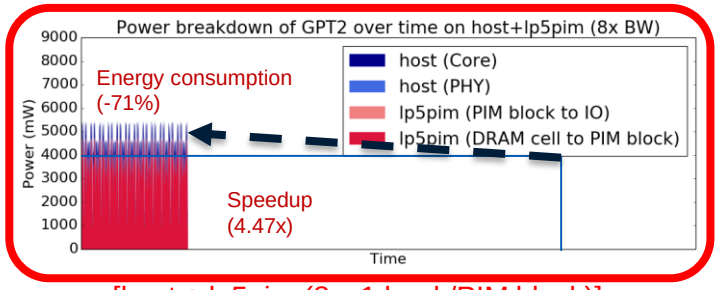
LPDDR-PIM System Perf./Power Analysis

- LPDDR-PIM improves energy efficiency by shorter execution time.
  - Power consumption of DRAM internal component (red) increases proportionally
  - Power consumption of global I/O bus (light red) and I/O PHYs (light blue) considerably decreases

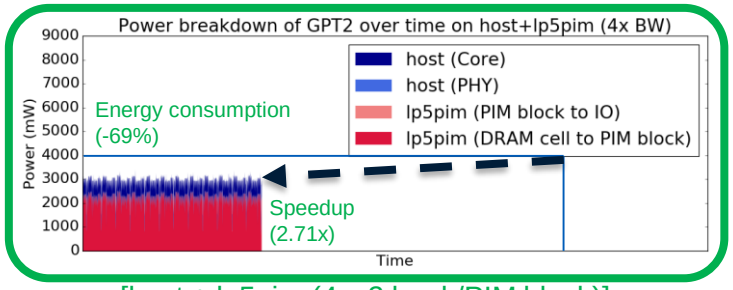




[host + lpddr5]



[host + lp5pim (8x: 1 bank/PIM block)]



[host + lp5pim (4x: 2 bank/PIM block)]

Summary

- Generative AI requires **High-Bandwidth** and **High-Capacity** Memories.
- Memory vendors provide Various Memory Solutions to meet requirements.
  - HBM for server and LLW for edge and mobile
- Processing capability in Memory enables higher bandwidth and energy efficiency.
  - CIM, PIM and PNM are meaningful and heavily studied in school and industry.
  - **PIM**: Still lots of challenges to be solved for commercializing.
  - Specially, **LPDDR-PIM** is being prepared for on-device AI.
  - **Need strong collaboration** between system, processor, memory and software.