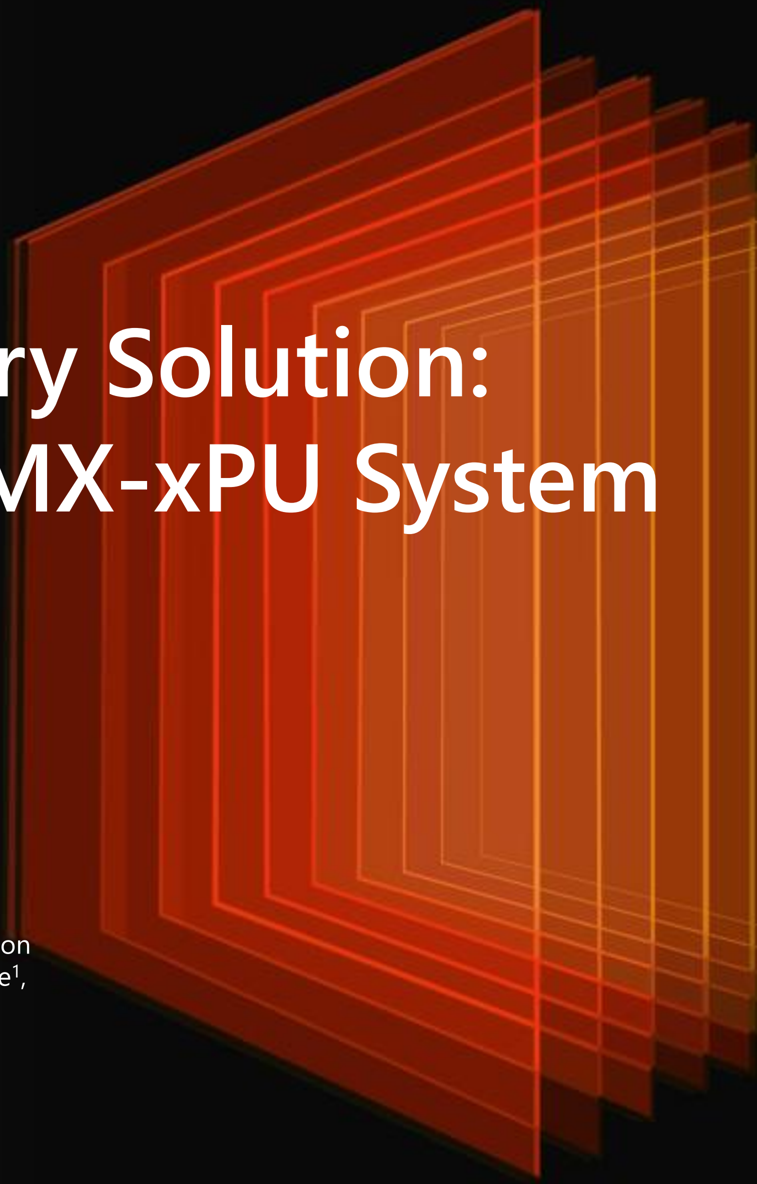
SK hynix AI-Specific Computing Memory Solution: From AiM device to Heterogeneous AiMX-xPU System for Comprehensive LLM Inference

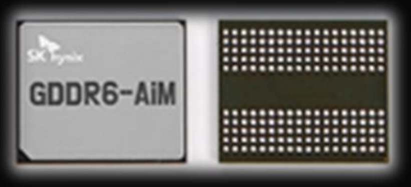


Guhyun Kim[1], Jinkwon Kim[1], Nahsung Kim[1], Woojae Shin[1], Jongsoon Won[1], Hyunha Joo[1], Haerang Choi[1], Byeongju An[1], Gyeongcheol Shin[1], Dayeon Yun[1], Jeongbin Kim[1], Changhyun Kim[1], Ilkon Kim[1], Jaehan Park[1], Yosub Song[1], Byeongsu Yang[1], Hyeongdeok Lee[1], Seungyeong Park[1], Wonjun Lee[1], Seonghun Kim[1], Yonghoon Park[1], Yousub Jung[1], Gi-Ho Park[2], and Euicheol Lim[1]

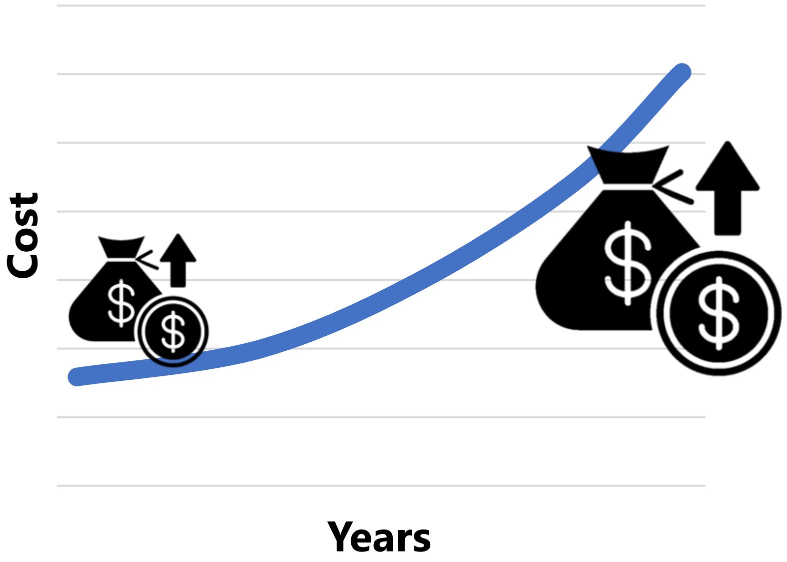
[1]SK hynix inc, [2]Sejong University

- Recap Accelerator-in-Memory (AiM) & AiMX
- System Extensions of AiMX Card for Datacenter
- AiM & AiMX for On-device AI
- Design Choices for Future AiM/AiMX
- Conclusion

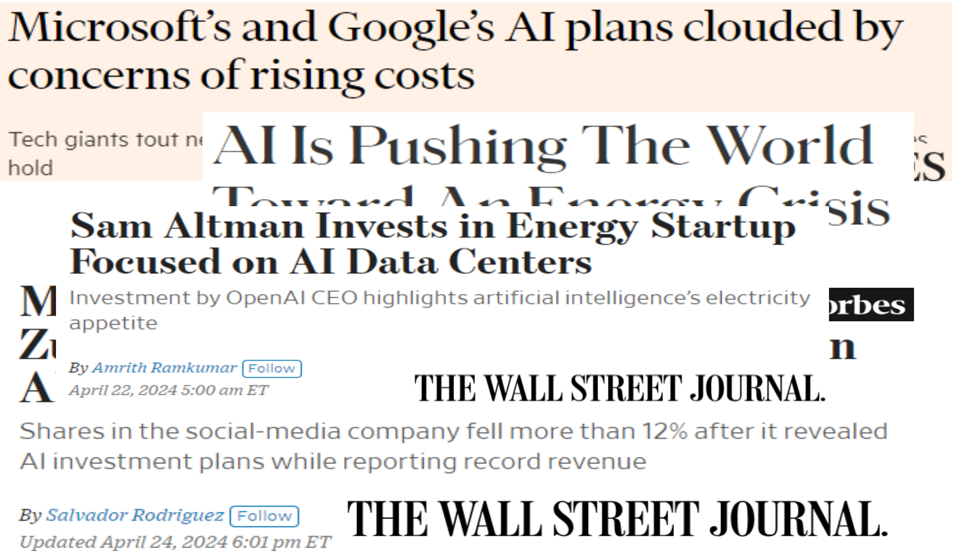
Recap Accelerator-in-Memory (AiM) and AiMX



Is LLM Sustainable?



Too Expensive Operating Expenditure



Microsoft's and Google's AI plans clouded by concerns of rising costs

Tech giants tout ne
hold

AI Is Pushing The World Toward An Energy Crisis

Sam Altman Invests in Energy Startup Focused on AI Data Centers

Investment by OpenAI CEO highlights artificial intelligence's electricity appetite

By *Amrith Ramkumar* Follow
April 22, 2024 5:00 am ET

M
Zu
A

orbes
n

THE WALL STREET JOURNAL.

Shares in the social-media company fell more than 12% after it revealed AI investment plans while reporting record revenue

By *Salvador Rodriguez* Follow
Updated April 24, 2024 6:01 pm ET

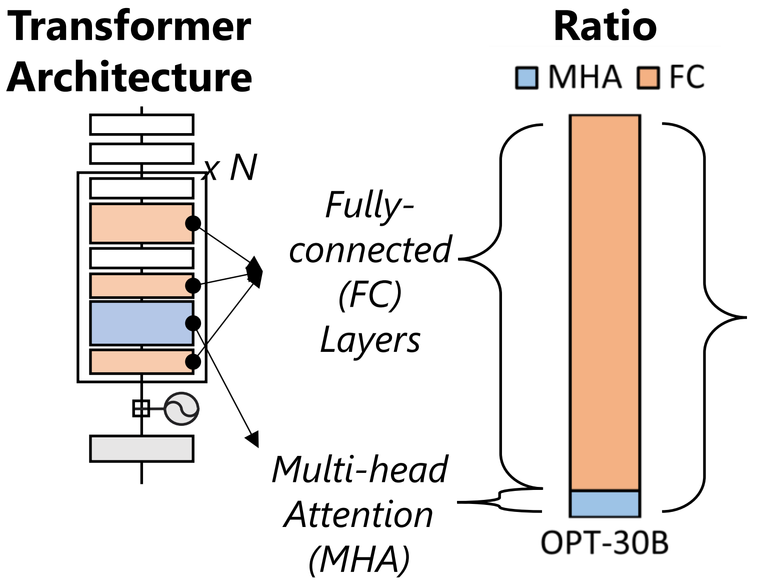
THE WALL STREET JOURNAL.

© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

Large Language Model – Memory Bound

LLM Architecture

- Mainly Consists of Matrix-Vector Multiplications (or GEMV)



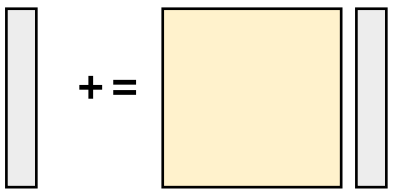
(*) Assumptions: batch1 inference during output token generation phase

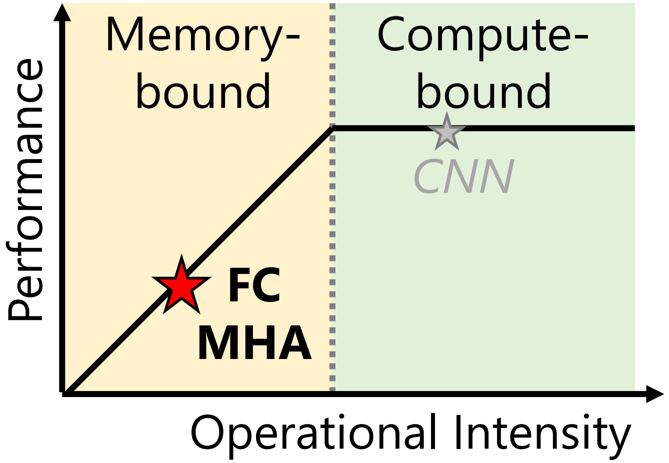
Matrix-Vector Multiplication

- GEMV: Memory BW-Bound with Low Arithmetic Intensity

Matrix-Vector Multiplication (GEMV)

$$y \leftarrow \alpha A x + \beta y$$

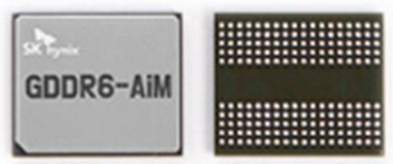




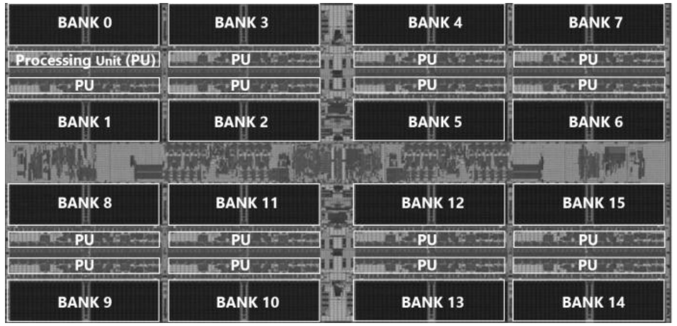
Accelerator-in-Memory: “True All-Bank Parallelism”

SK hynix’s First GDDR6-based Processing-in-Memory Product Sample

GDDR6-AiM Package



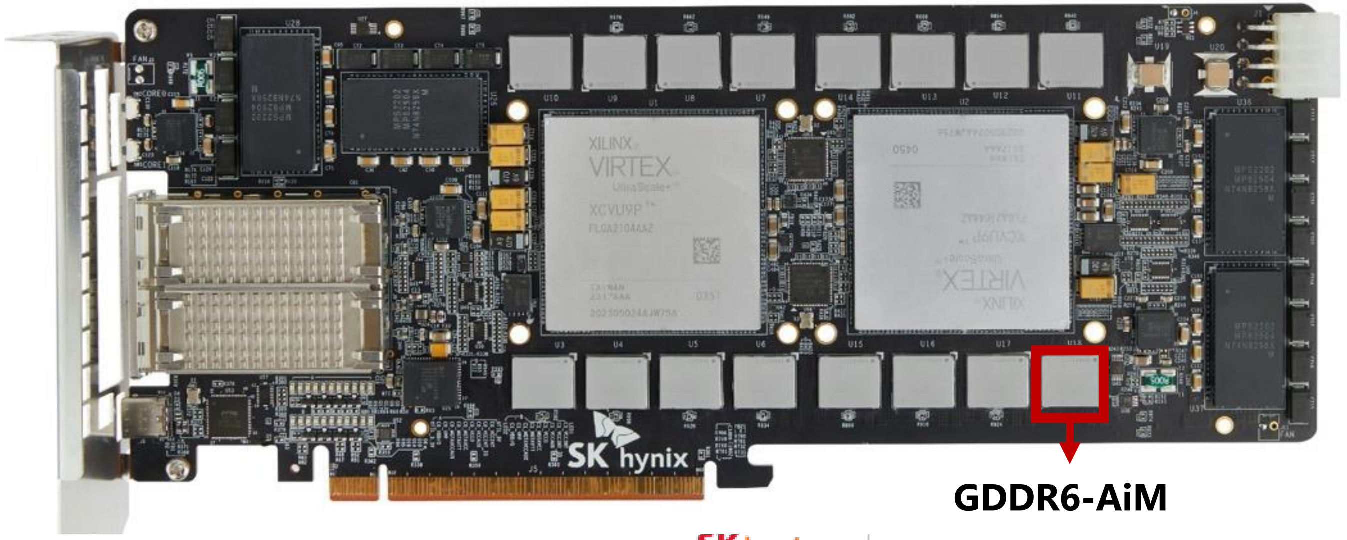
GDDR6-AiM Die Photograph



GDDR6-AiM Specification* (per die)

(External) Bandwidth**	32 GB/s
Operating Speed	1 GHz
Compute Throughput**	512 GFLOPS
Internal Bandwidth**	512 GB/s
Numeric Precision	BF16

AiMX Card Prototype



AiMX Card Prototype Specification

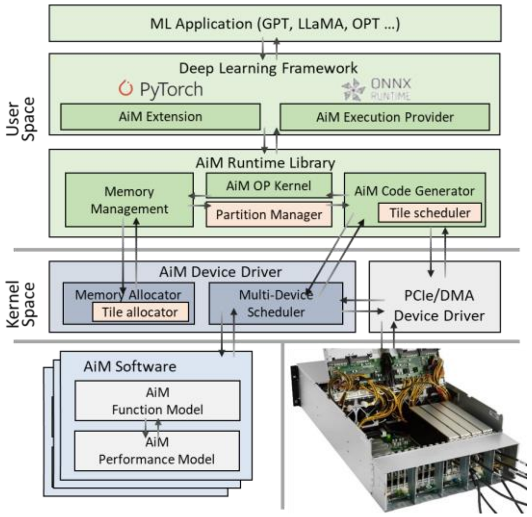
Host Interface		PCIe Gen3 x8x8 (bifurcated)
Form Factor		FHFL (A100/A30 compatible)
Configuration		2 FPGA*** x 16 AiM package
AiM	Capacity	16 GB
	Bandwidth	170 GB/s (@2.67Gbps****)

(*) S.Lee et al., ISSCC’22
(**) Defined as a peak during burst operations
(***) Xilinx Virtex Ultrascale+ (VU9P)
(****) 1/6 of peak data rate of GDDR6, 16Gbps (or 1TB/s)

SK hynix | © SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

Current Status of AiMX

AiM SDK Arch.



ML Application (GPT, LLaMA, OPT …)

User Space

Deep Learning Framework

PyTorch

ONNX RUNTIME

AiM Extension

AiM Execution Provider

AiM Runtime Library

Memory Management

AiM OP Kernel

Partition Manager

AiM Code Generator

Tile scheduler

Kernel Space

AiM Device Driver

Memory Allocator

Tile allocator

Multi-Device Scheduler

PCIe/DMA Device Driver

AiM Software

AiM Function Model

AiM Performance Model

Live DEMO in 2023

AI HARDWARE & EDGE AI SUMMIT

OCP GLOBAL SUMMIT

SC23 Denver, CO | i am hpc.

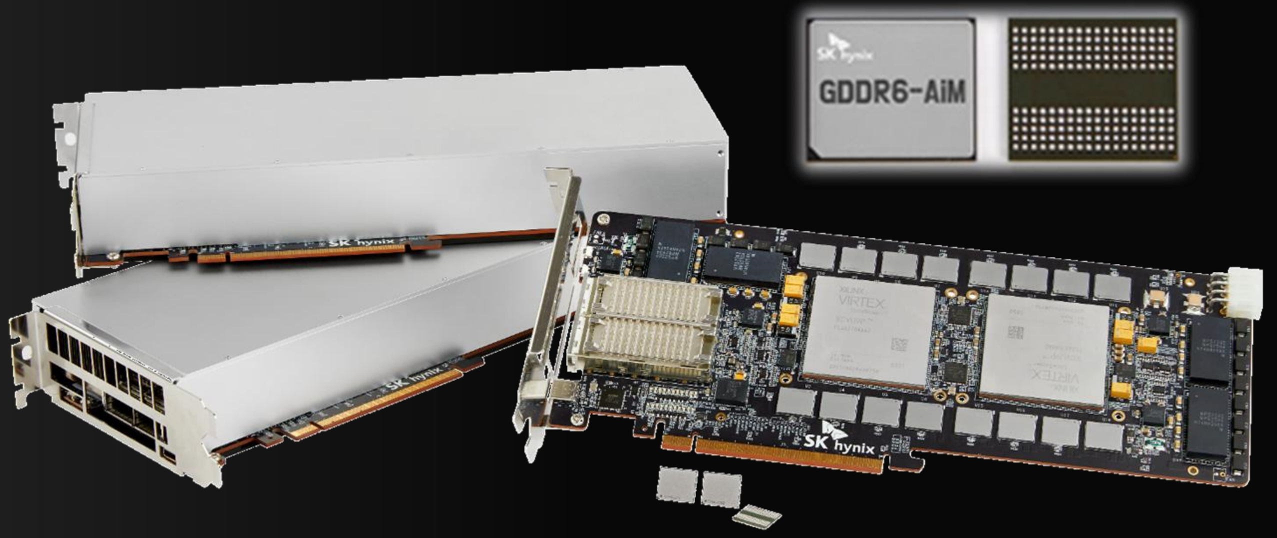
For Detailed Information..

- *[ISSCC'22/JSSC'22] 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"*
- ***[HC35] Memory-Centric Computing with SK Hynix's Domain-Specific Memory***
- *[SC23] Cost-Effective LLM Inference Solution Using SK hynix's AiM (Accelerator-in-Memory)*
- *White Papers: https://product.skhynix.com/support/downloads/kits.go*

© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

System Extensions of AiMX Card for Datacenter

- LLM Trend in Datacenter
- MHA in AiM
- Extended AiMX Card



Large Language Model Trend in Datacenter (1)

Fully-Connected Layer

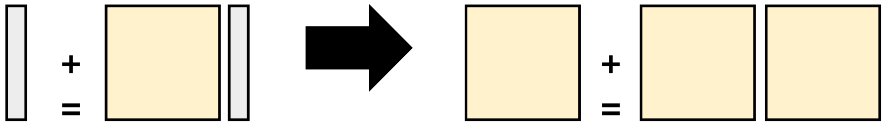
- Increase batch size
- Memory-bound GEMV → Compute-bound GEMM

Matrix-Vector Product (GEMV)

$y \leftarrow \alpha Ax + \beta y$

Matrix-Matrix Product (GEMM)

$C \leftarrow \alpha AB + \beta C$



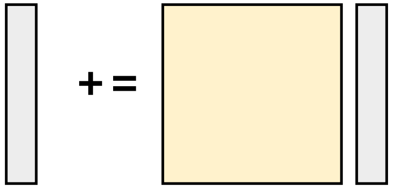
Batch size 1

Batch Size *N*

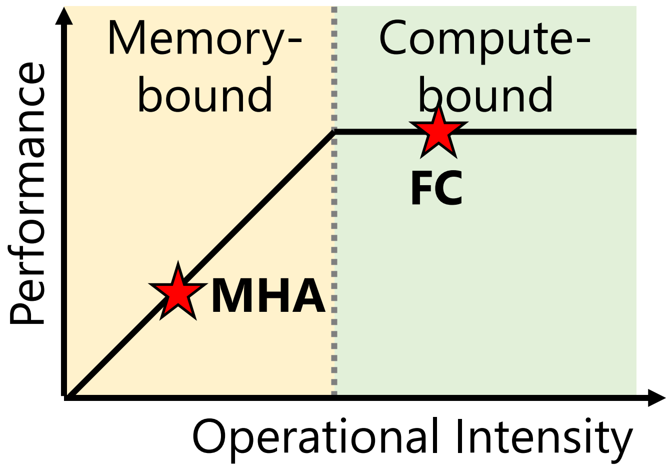
Multi-Head Attention

- Remain as GEMV
- MHA portion increase as batch size increase

$$y \leftarrow \alpha Ax + \beta y$$



Property of LLMs in Datacenter

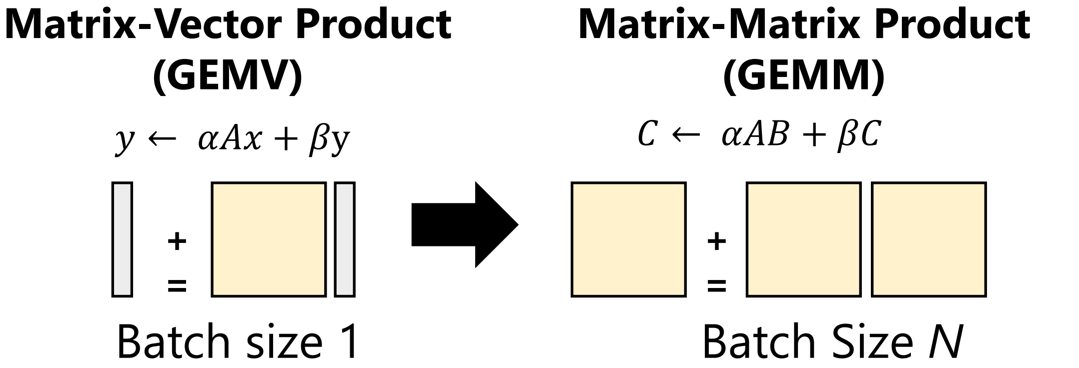


© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

Large Language Model Trend in Datacenter (2)

Fully-Connected Layer

- Increase batch size
- Memory-bound GEMV → Compute-bound GEMM



Matrix-Vector Product (GEMV)

$y \leftarrow \alpha Ax + \beta y$

Batch size 1

Matrix-Matrix Product (GEMM)

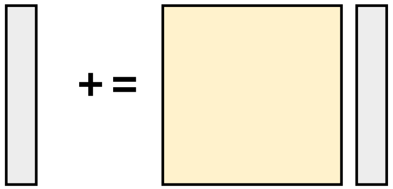
$C \leftarrow \alpha AB + \beta C$

Batch Size *N*

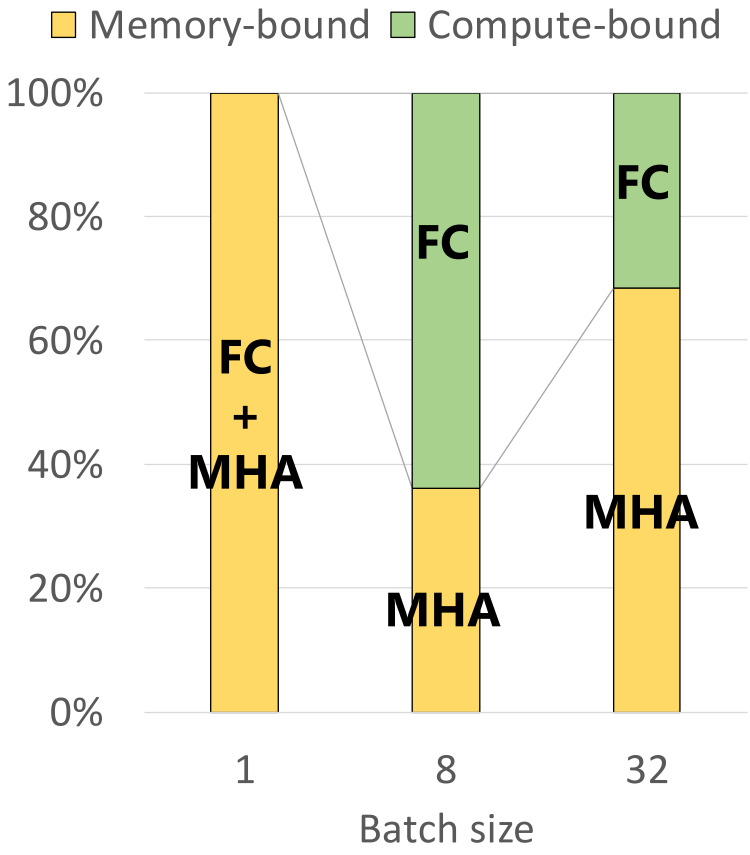
Multi-Head Attention

- Remain as GEMV
- MHA portion increase as batch size increase

$$y \leftarrow \alpha Ax + \beta y$$



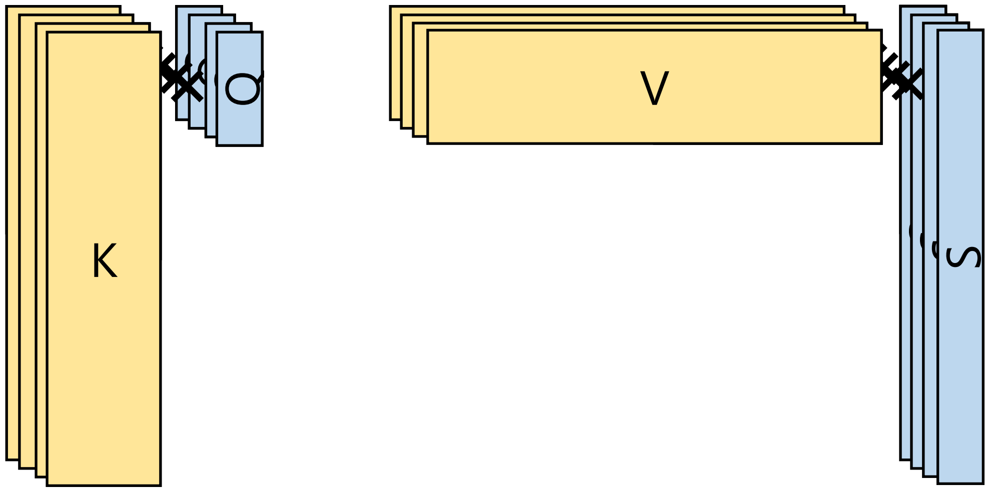
Property of LLMs in Datacenter



© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

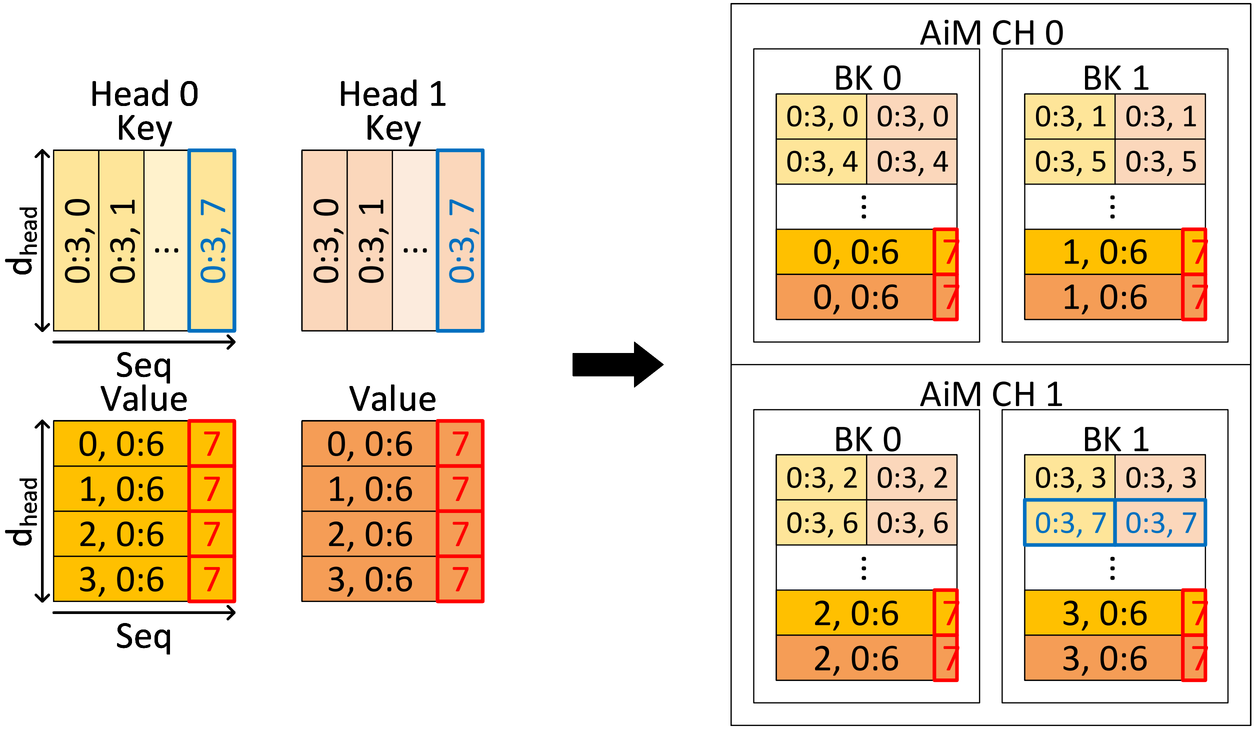
Multi-Head Attention in AiM

Multi-Head Attention



- **Asymmetric Matrix**
  - QK^T: Small Input Vector Size
  - SV: Small Output Vector Size
  - Many Heads
- **Consistency**
  - Key, Value Matrix Updated as Input Token Given

AiM Aware Key/Value Matrix Placement

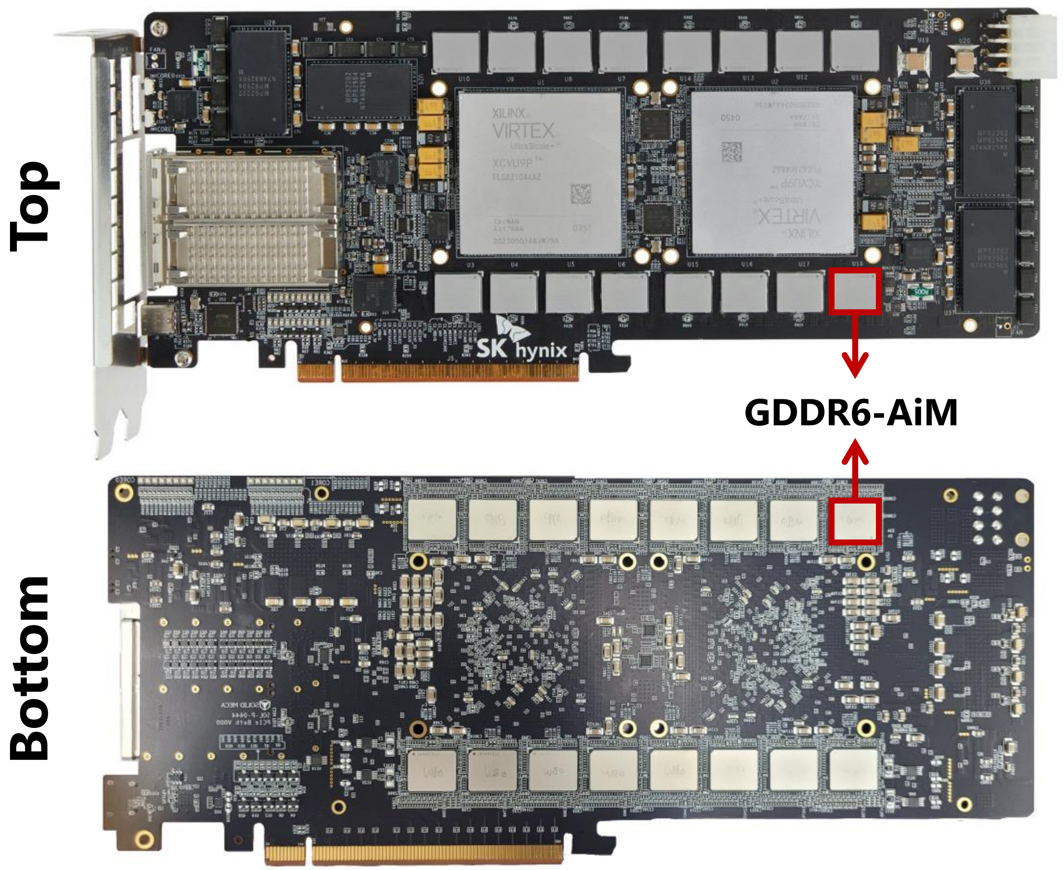


- **Key**
  - Gather Newly Generated Key Vectors into 1 Bank
- **Value**
  - Spread Newly Generated Value Vectors across All Banks

© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

Extended AiMX Card

Extended AiMX Card Prototype

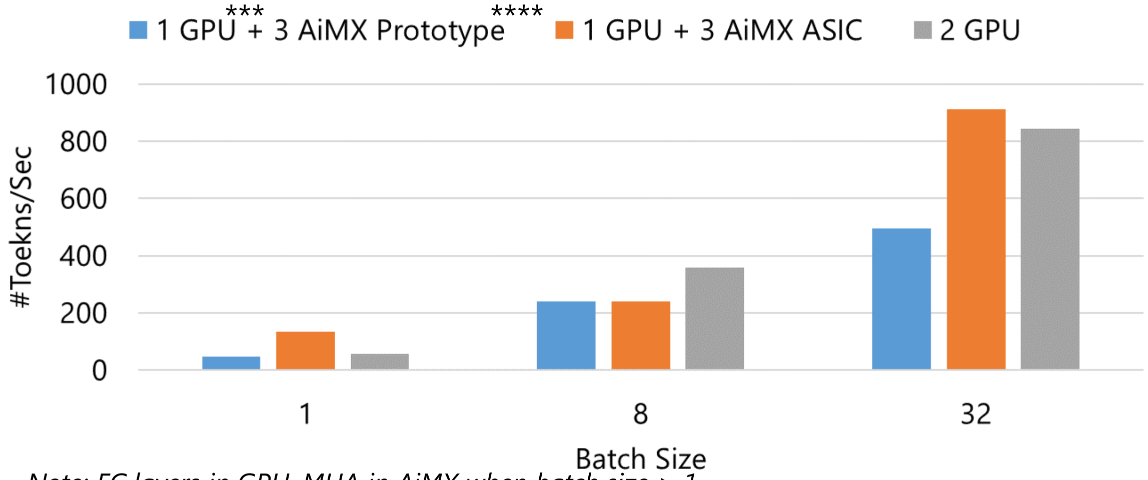


Specification

Form Factor		FHFL (H100/A100 compatible)
Configuration		2 FPGA* x 32 AiM package
AiM	Capacity	32 GB
	Bandwidth	170 GB/s (@2.67Gbps**)
Thermal Cooling		Passive

- Non-JEDEC Rank-like GDDR6 Configuration To Overcome IO Limit of FPGA

Performance Evaluation (OPT3-30B)



(*) Xilinx Virtex Ultrascale+ (VU9P)
(**) 1/6 of peak data rate of GDDR6, 16Gbps (or 1TB/s)

Note: FC layers in GPU, MHA in AiMX when batch size > 1
(***) H100 80GB GPU
(****) Estimated by in-house performance model because extended AiMX prototype cards are in bring up process. Number of AiMX is set to match capacity of H100 GPU.

© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

Plan of AiM/AiMX for Datacenter

Live Demo in 2024

AI HARDWARE & EDGE AI SUMMIT

San Jose,
SEP 09-12

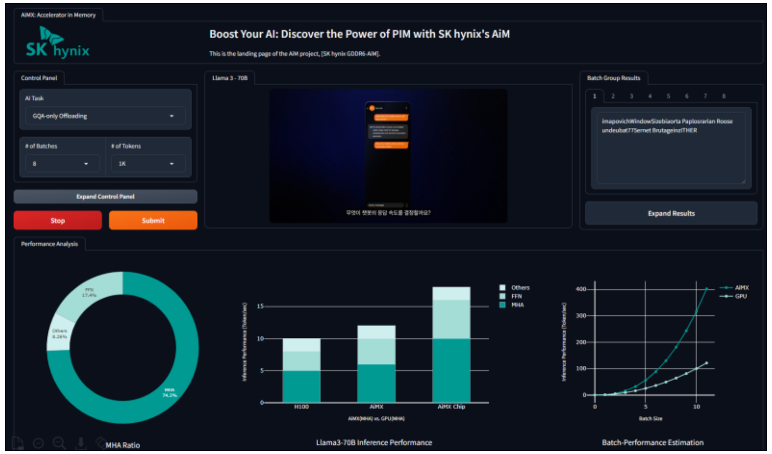
OCP GLOBAL SUMMIT

San Jose,
OCT 15-17

SC24 Atlanta, GA hpc creates.

Atlanta,
NOV 17-22

LLM Inference Demonstration


AiMX: Accelerator in Memory

SK hynix

Boost Your AI: Discover the Power of PIM with SK hynix's AiM

This is the landing page of the AiM project, [SK hynix GDDR6-AiM].

Control Panel

AI Task: GQA-only Offloading

# of Batches: 8

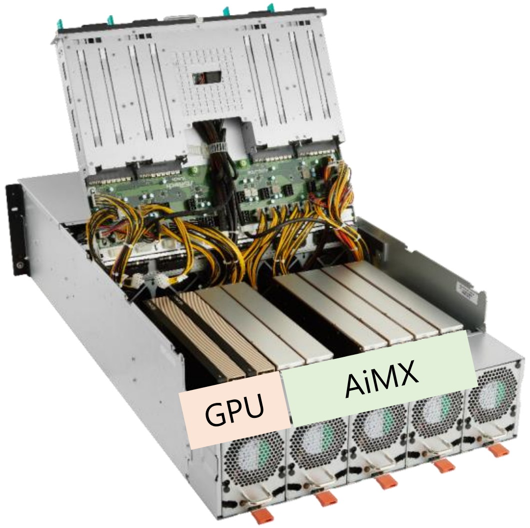
# of Tokens: 1K

무엇이 챗봇의 응답 속도를 결정할까요?

Batch Group Results

imapovichWindowSizebiaorta Paplosrarian Roose undeubat77Sernet BrutageinzITHER

Performance Analysis



- Larger Model
- Multi-head Attention w/ Multi-batch

Future Target

- High Capacity Solution
  - LPDDR-AiM
  - Up to 256GB/Card
- Super excellency in both System Performance and Energy Efficiency

SDK is Available Now!
Collaborate with SK hynix!

SKhynix_PIM@skhynix.com

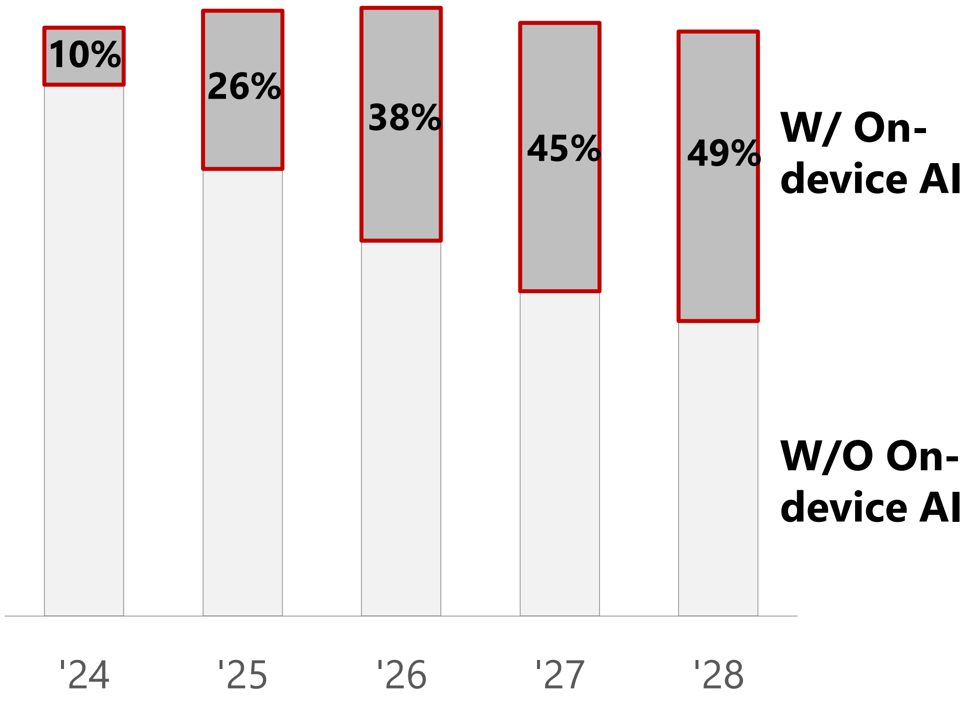
© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

AiM & AiMX for On-device AI

- On-device AI Trend
- LPDDR-AiM
- AiMX for On-device AI

On-device AI Prospect

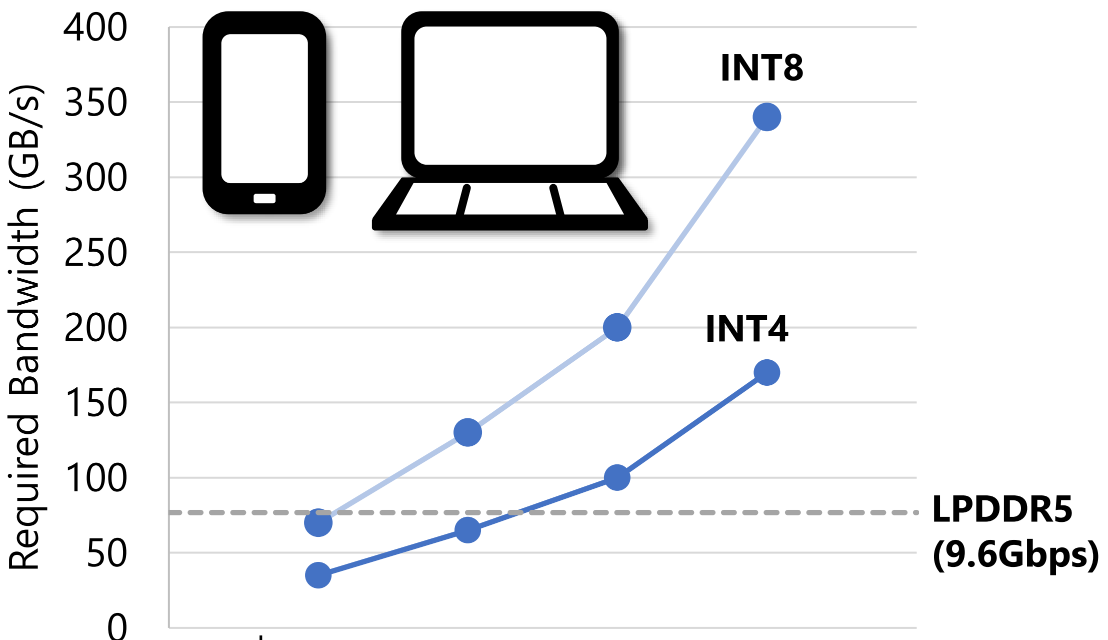
On-device AI Smartphone Prospect



Source: SK hynix Marketing Forecast

Bandwidth Requirement Prospect for On-device AI

Assumption:10 tokens/sec required

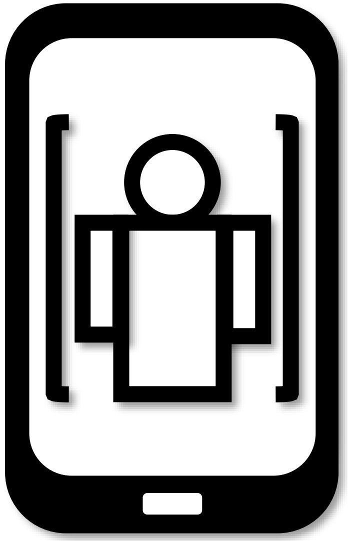


Year		2023	2024	2025	2026
Model Size		7B	13B	20B	34B
Capacity (GB)	INT8	7	13	20	34
	INT4	3.5	6.5	10	17

SK hynix | © SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

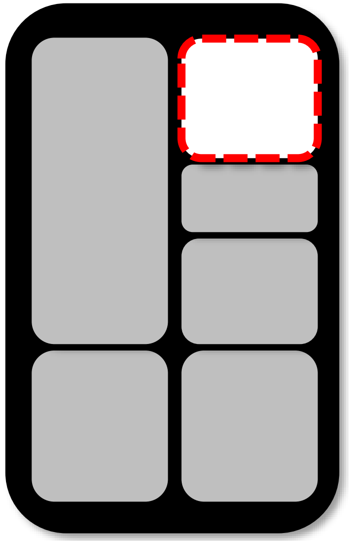
Why PIM for on-device AI?

Individualization



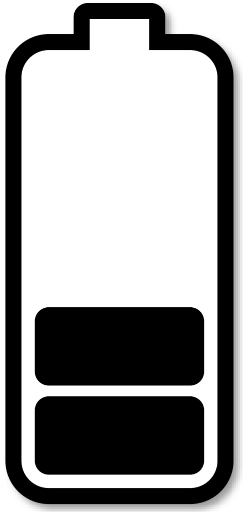
Low Batch Size
Memory Bound

Form Factor



Restricted Area

Battery



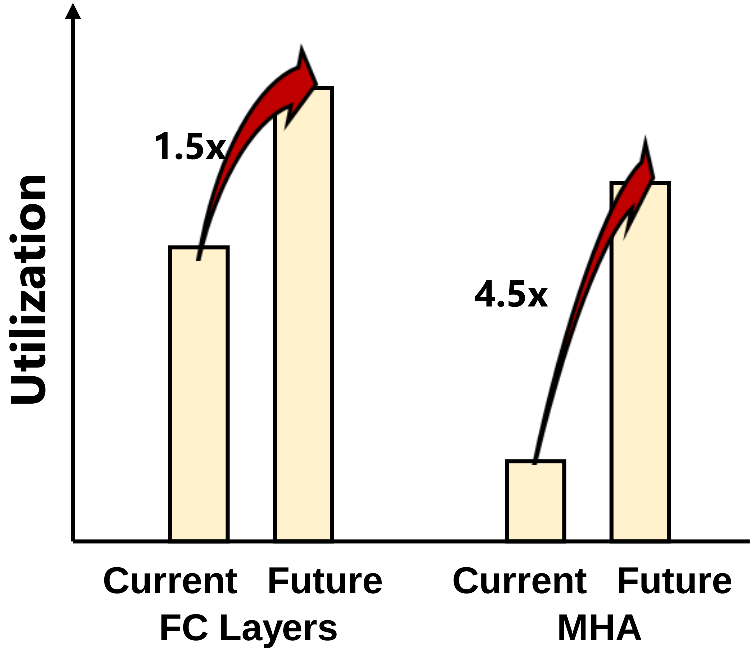
Energy Efficient

- Accelerate GEMV
- Replace Main Memory
- High Energy Efficiency

© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

LPDDR-AiM

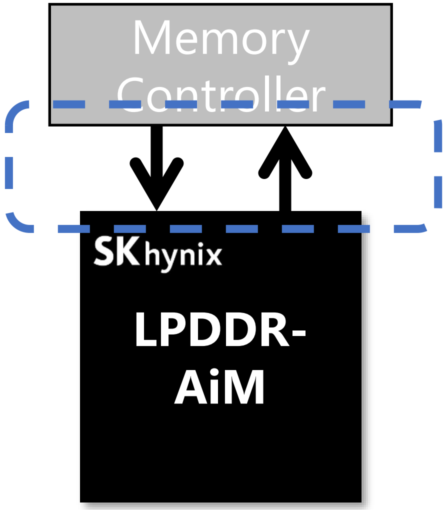
Higher Utilization



- Optimize Datapath
- Relieve State Diagram Restriction

Note: Assuming both LPDDR5-based with INT8 precision, but different architecture

Compatibility



- Not Change Existing LPDDR Commands/Performance
- Protocol to Minimize Overhead Between PIM <> RD/WR

Expected Specification

LPDDR-AiM (per die)	
Memory Density (GB)	1~2
Organization	X16
IO Data rate	9.6
(External) Bandwidth*	19.2 GB/s
Numeric Precision	INT4/8
Processing Unit (PU)	16 PU/die
Compute Throughput**	307.2 GOPS
Internal Bandwidth*	153.6 GB/s

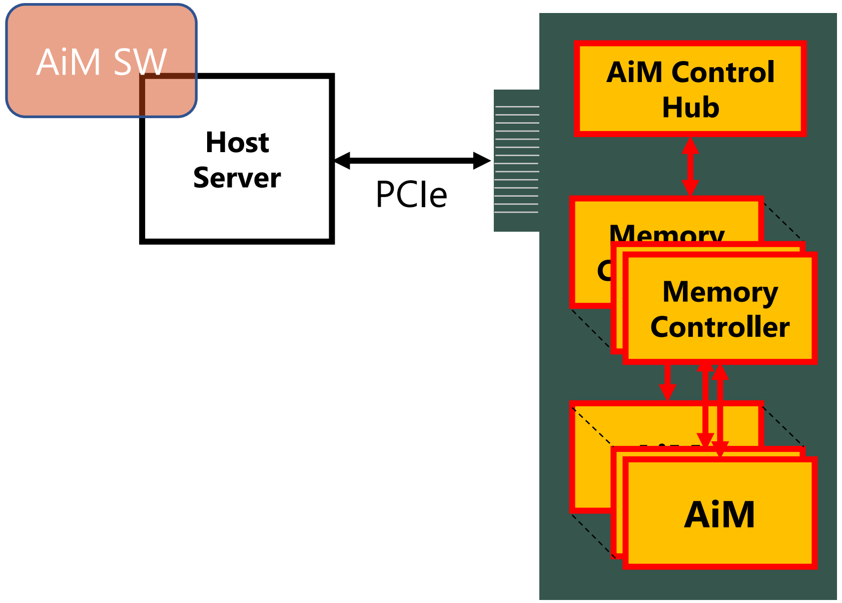
LPDDR-AiM (per package)	
Number of Dies	4~8
Memory Density (GB)	4~16
Organization	X64
(External) Bandwidth*	76.8 GB/s
Compute Throughput**	1228.8 GOPS
Internal Bandwidth*	614.4 GB/s

Note: Assuming AiM is based on LPDDR5, existing fastest LPDDR
(*) Defined as a peak during burst operations (tCCDL)
(**) INT8 based estimation

© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

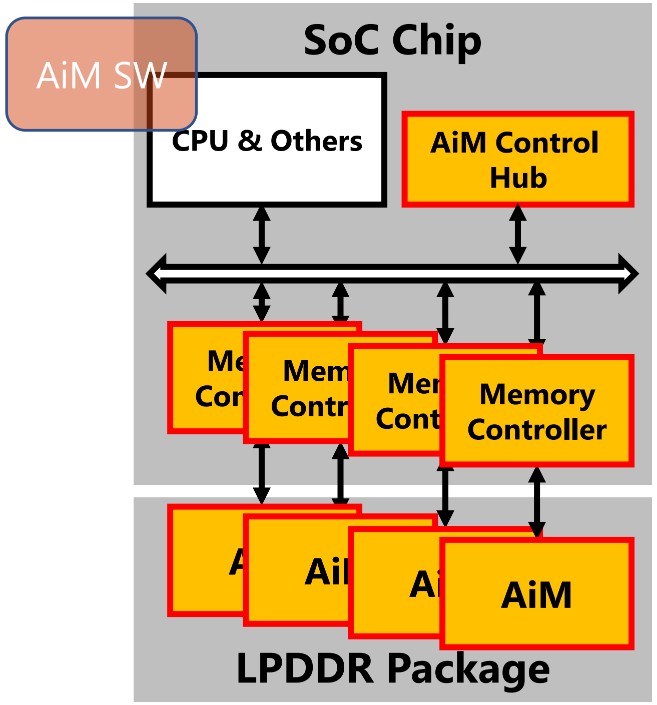
AiMX System for On-device AI

AiMX System for Datacenter



AiMX Card

AiMX System for On-device AI



- Similar Architecture as AiMX System for Datacenter
- Need to Modify Current Mobile AP or Client CPU

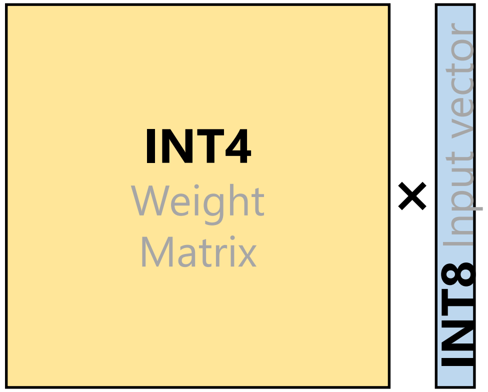
© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

Design Choices for Future AiM/AiMX

- AiM
- SoC
- Software

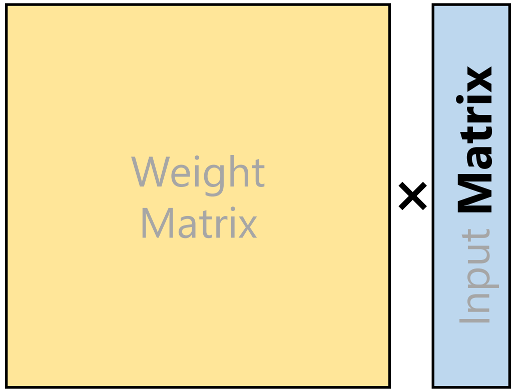
Design Choices - AiM

Precision



- Binary, INT, FP, BF, MX, …
- Scale Factor
- Heterogeneous Precision

Functionality

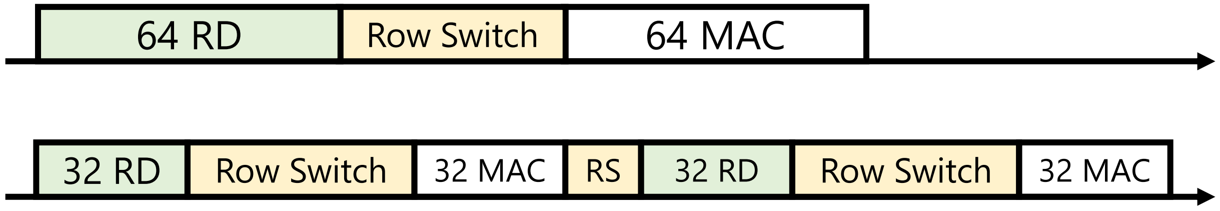


- GEMV
- GEMM
  - Batch, GQA, …
  - Hybrid Bonding
- Other Application?

SK hynix | © SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

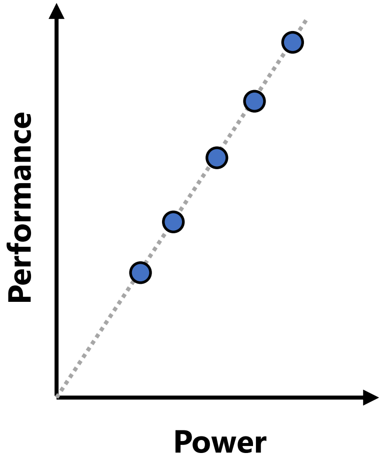
Design Choices - SoC

Arbitration



- Coarse-grained Normal & PIM Interleaving
  - Minimize Switch Overhead
  - *cf) Write Request Draining*

Power/ Thermal Balance

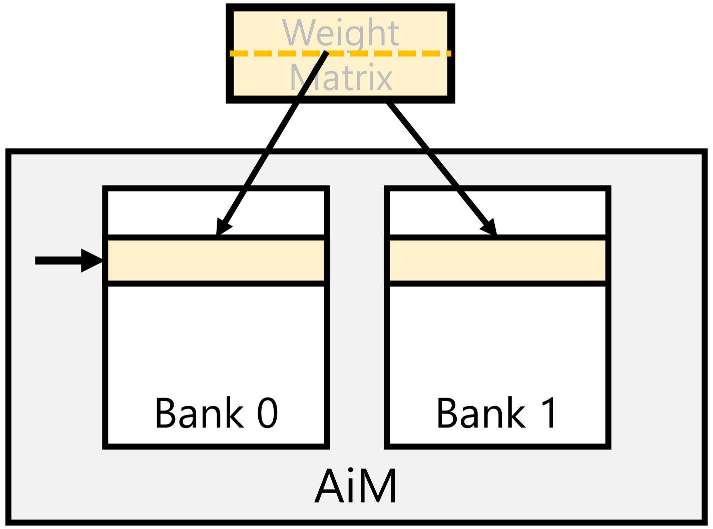


- Power Throttling
  - MAC-to-MAC Latency
  - #Banks to MAC
- Dynamic Power Supply

© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

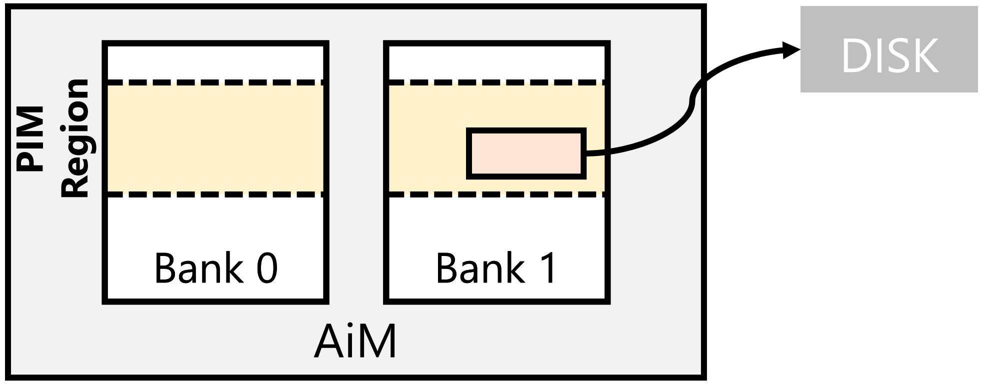
Design Choices - Software

Large Page Size



- To Place Weight Data into AiM-Aware Manner

Memory Management



- PIM-Aware Memory Swap Policy

© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

Conclusion

Beyond Memory..

Architecture

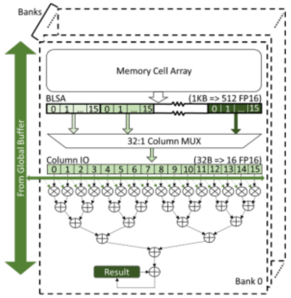
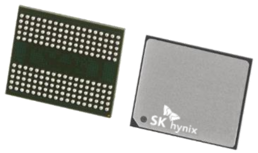


Fig. 4. Newton Datapath

Chip

GDDR6-AiM

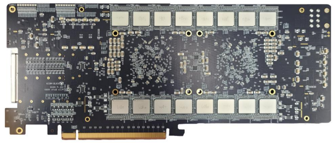


System

AiMX

Extension

Extended AiMX



2020 **2022** **2023** **2024**

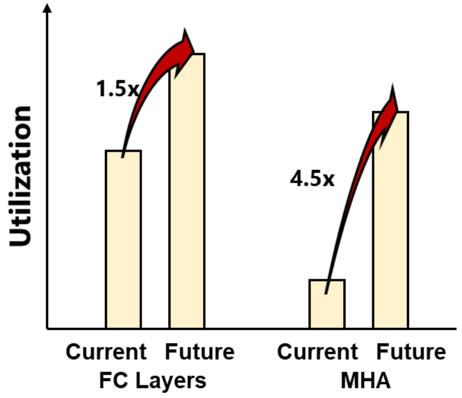
High Cap./Perf. Solution for Datacenter

Application, Hybrid Bonding, CXL-PIM, ...

AiM/AiMX for On-device AI

SDK is Available Now!
Collaborate with SK hynix!

SKhynix_PIM@skhynix.com

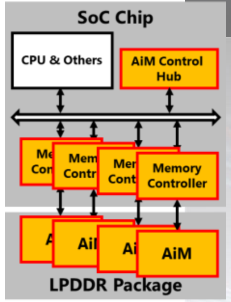


Architecture

SK hynix

LPDDR-AiM

Chip



System

Boost Your App. with SK hynix!

© SK hynix Inc. This material is proprietary of SK hynix Inc. and subject to change without notice.

Thank You

Q&A

SKhynix_PIM@skhynix.com