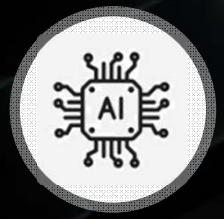
PIM use case: Cost-effective LLM accelerator using AiM (Sk hynix's PIM)

Euicheol Lim

SK hynix

AI chatbot : Game changer of AI Market

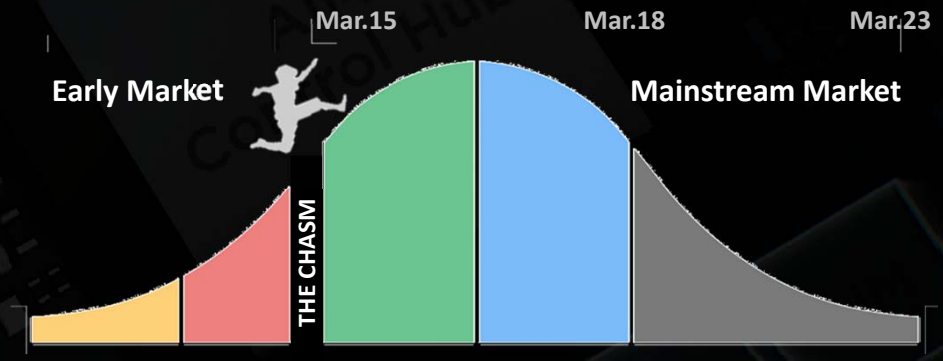
- AI chatbot is opening a new mainstream market for AI Services but OpEx issues have to be solved



Innovative Technology

Investment

WOW Factor



"AI chatbot is crossing the CHASM"

Too Expensive
Operating Expenditure

AI chatbot inference – Prompt & Response

- AI chatbot Inference consists of input processing (prompt) and answer generating (response)
- Especially, response stage is significantly memory intensive

Prompt : Comprehension

Input: 9 Words / Model Parameter Read: 1

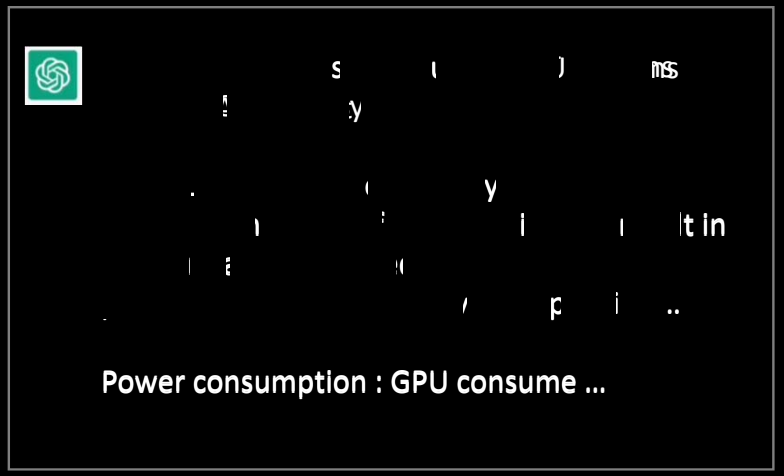
What are the issues of running Chatbot with GPUs?

What	are	the
issues	of	running
Chatbot	with	GPUs

Computing-Intensive

Response: Generating Answer

Output: 196 Words / Model Parameter Read : 261



Memory-Intensive

1) For easy understanding, we use 'Words' as a unit of sequence length as opposed to 'Token'. Since 1 Token is the same as 0.75 Word on average, the number of 'Read' is 261.

Is GPU sufficient for AI chatbot service?

- Need higher performance and more cost-effective computing infra than current system

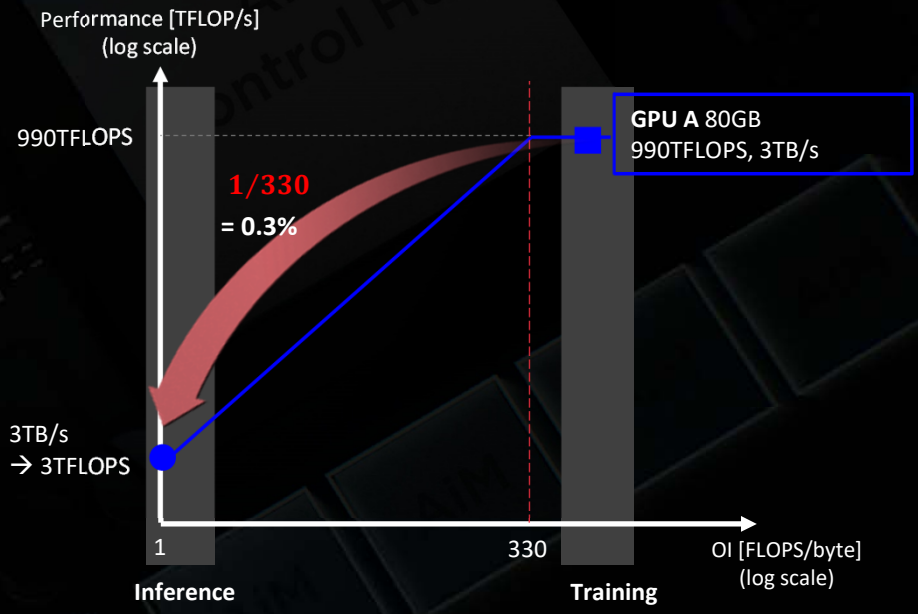
Inference Time with GPU System

175B LLM model case

GPU A (80GB, 3TB/s) x 5

Model size	**350GB**
Bandwidth	**15TB/s**
Processing Time (1 token)	**23 ms** (350GB/15TB/s)
Processing Time (261 token)	**6.0 sec**

Why so slow?

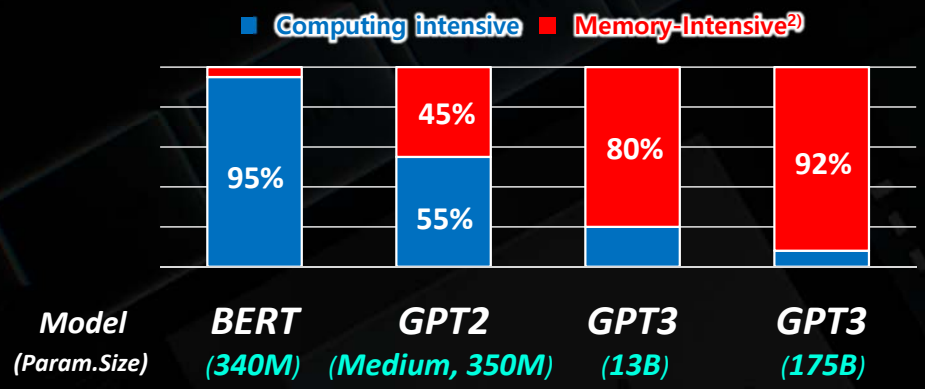


- **Extremely low performance per cost**
  - Using only 0.3% of the GPU performance, Wasting GPU power consumption

Why PIM as a LLM service Accelerator?

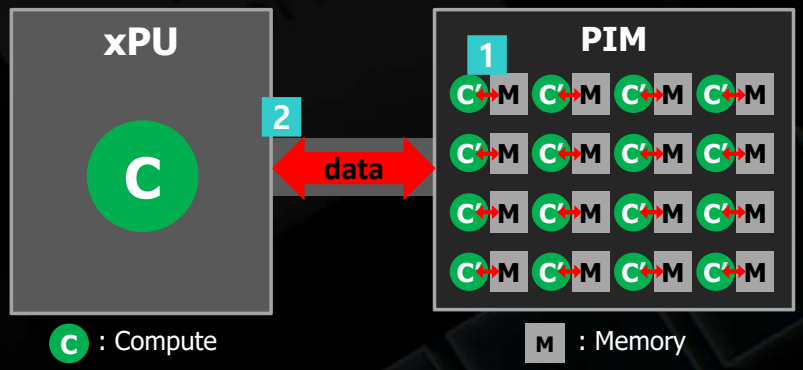
- PIM is the best option for LLM service which is highly memory-Intensive.

LLM service with Model & Size



- The larger the model, the **more memory intensive function (specifically, "GEMV"),** so **Memory Bandwidth for GEMV operation** has a greater impact on system performance than the processor

Feature of PIM

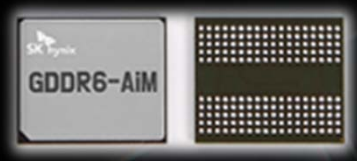


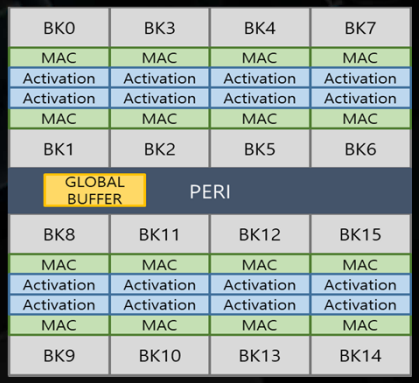
1	**Performance Improvement** By utilizing the higher Bandwidth inside the memory
2	**Energy Efficiency Improvement** By minimizing data movement between host and memory

PIM is suitable for accelerating Memory-Intensive Application like LLM inference

AiM introduction

- SK hynix's very first GDDR6-based processing-in-memory (PIM) product called AiM(Accelerator-in-Memory) is ready



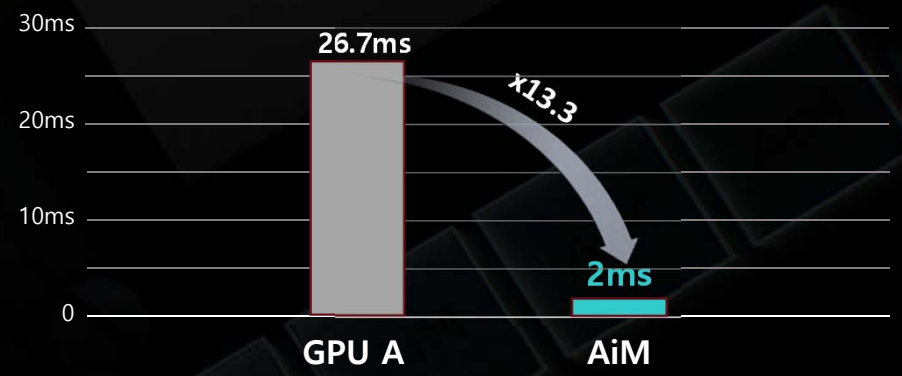


AiM	
Memory Density	1GB
Bandwidth-external	64 GB/s
Function Support	GEMV, Activation
GEMV Bandwidth	0.5 TB/s (x8 of external BW)
GEMV Performance	0.5 TFLOPS
Numeric Precision	Brain Floating Point 16 (BF16)
Targets	Memory-intensive AI applications

Critical Metric

Performance per Memory Capacity [TFLOPS/GB]

→ Memory BW per Memory Capacity [TB/s / GB] (if OI = 1)

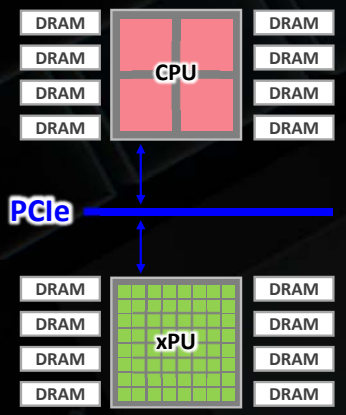


Token processing time = $\frac{1}{\text{Memory BW per Memory Capacity}}$

AiMX card - How to deploy AiM into existing system

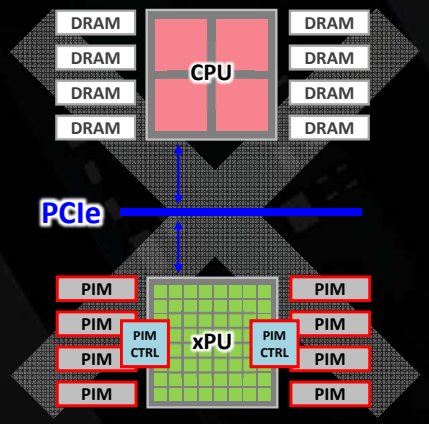
- Can be easily deployed into the existing system by adding the AiM based Accelerator (AiMX) Card, rather than by replacing DRAM with PIM

Conventional Computing System



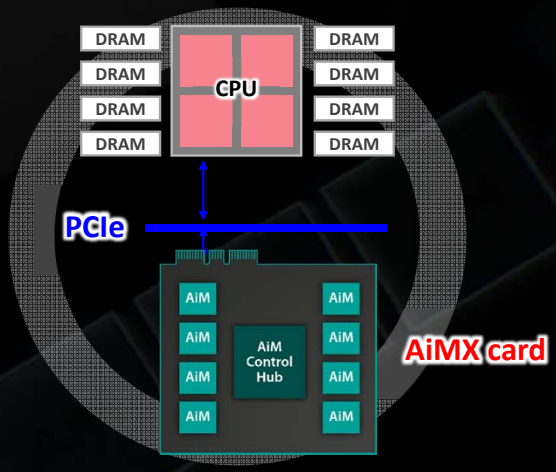
- Memory bottleneck

DRAM Replacement with PIM



- Host SoC (xPU) must be modified[1)]
- SW burden for memory mgmt.

AiM Accelerator (AiMX) Card

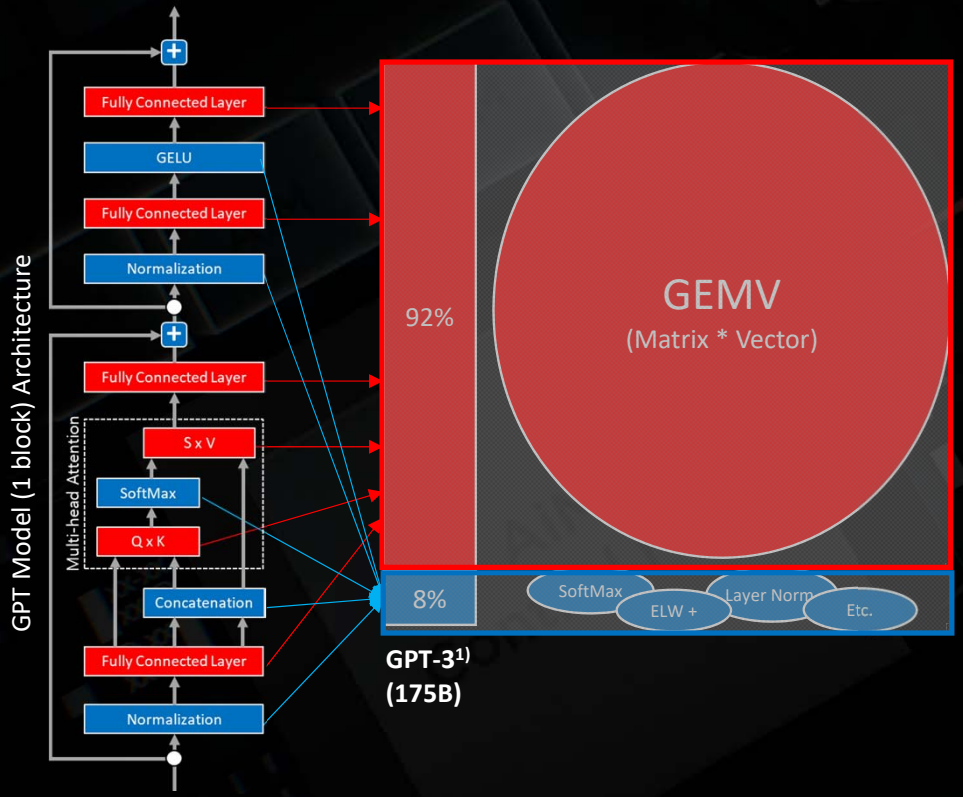


- No need to modify any existing xPUs
- Need an AiM Control Hub chip
- SW modification can be minimized

1) Conventional memory controller + additional command for PIM operation + in order scheduling

AiMX architecture – Efficiency & Flexibility

- Efficiency: AiM chip processes large amount fixed memory-intensive function (GEMV) efficiently
- Flexibility: AiM-Control-Hub processes small amount various functions flexibility

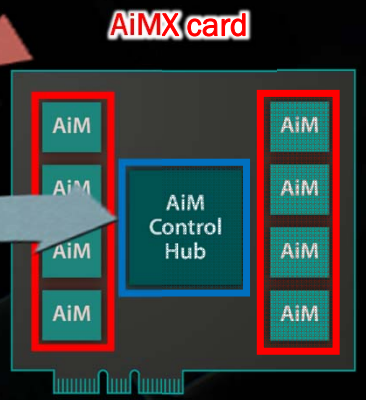


AiM

- Fixed Function (GEMV)
- Memory Intensive
- Large work
- Dedicated unit

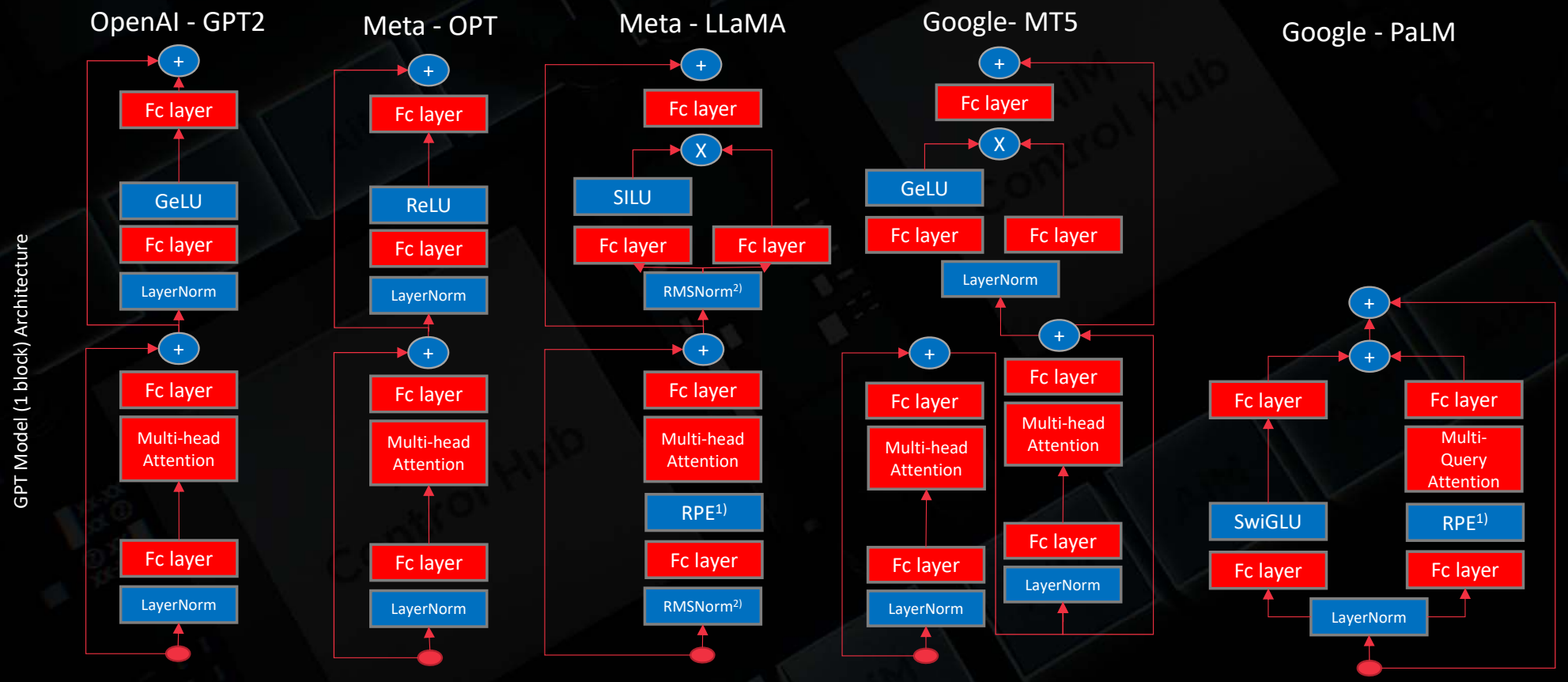
AiM Control Hub (ACH)

- Various Functions
- Compute Intensive
- Small work
- Flexible unit



1) Measured data using 1x V100 GPU with PyTorch (v2.0)

LLM Architectures Similarity



1) Rotary Positional Embedding 2) Root Mean Square Layer Normalization

AiMX benefit – Performance & Energy consumption

- Provide different level of customer experience by 13x shorter service latency
- Reduce operating cost significantly with 17% energy consumption

13x Shorter Service latency

175B model case

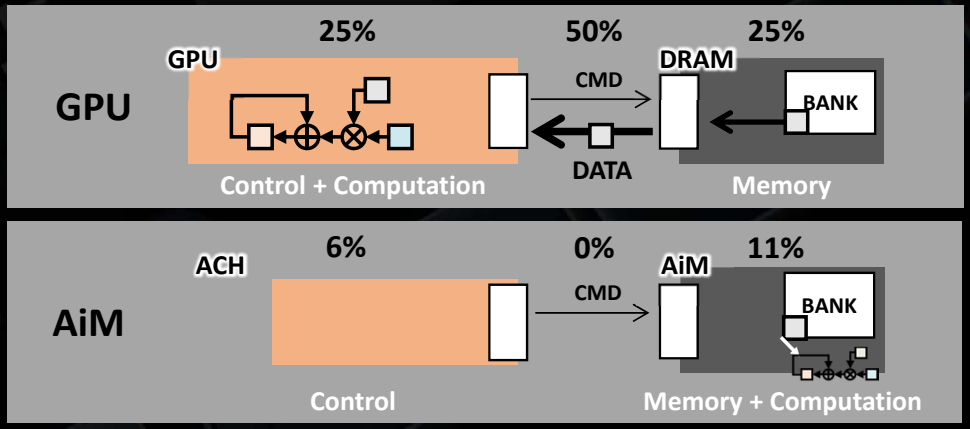
	GPU A (80GB, 3TB/s) x 5	AiM (16GB, 8TB/s) [1] x 25
Model size	350GB	350GB
Bandwidth	15TB/s	200TB/s
Processing Time (1 token)	23 msec (350GB/15TB/s)	1.8 msec (350GB/200TB/s)
Processing Time (261 token)	6.0 sec →	0.46 sec

13x Shorter Service latency

1) Prototype configuration (16 AiM chips per card)

GEMV energy consumption : 17% of GPU

1. Small controller and dedicated MAC unit in AiM
2. Remove off-chip data movement and reduce internal data movement
3. Reduce the static energy by short processing time



Example System configuration

- Most optimal computing infra for LLM service will be combination of GPU system and AiMX system
- GPU system for Prompt Stage, AiMX system for Response Stage

Prompt Stage
(Question Understanding)

- Input tokens are processed in parallel
- Needs just 1 time model data read for all token
- *Computing intensive*

GPU system

Response Stage
(Answer Generating)

- Output tokens are processed in serial
- Needs model data read in every token
- Extremely memory intensive

AiMX system

AiMX cards

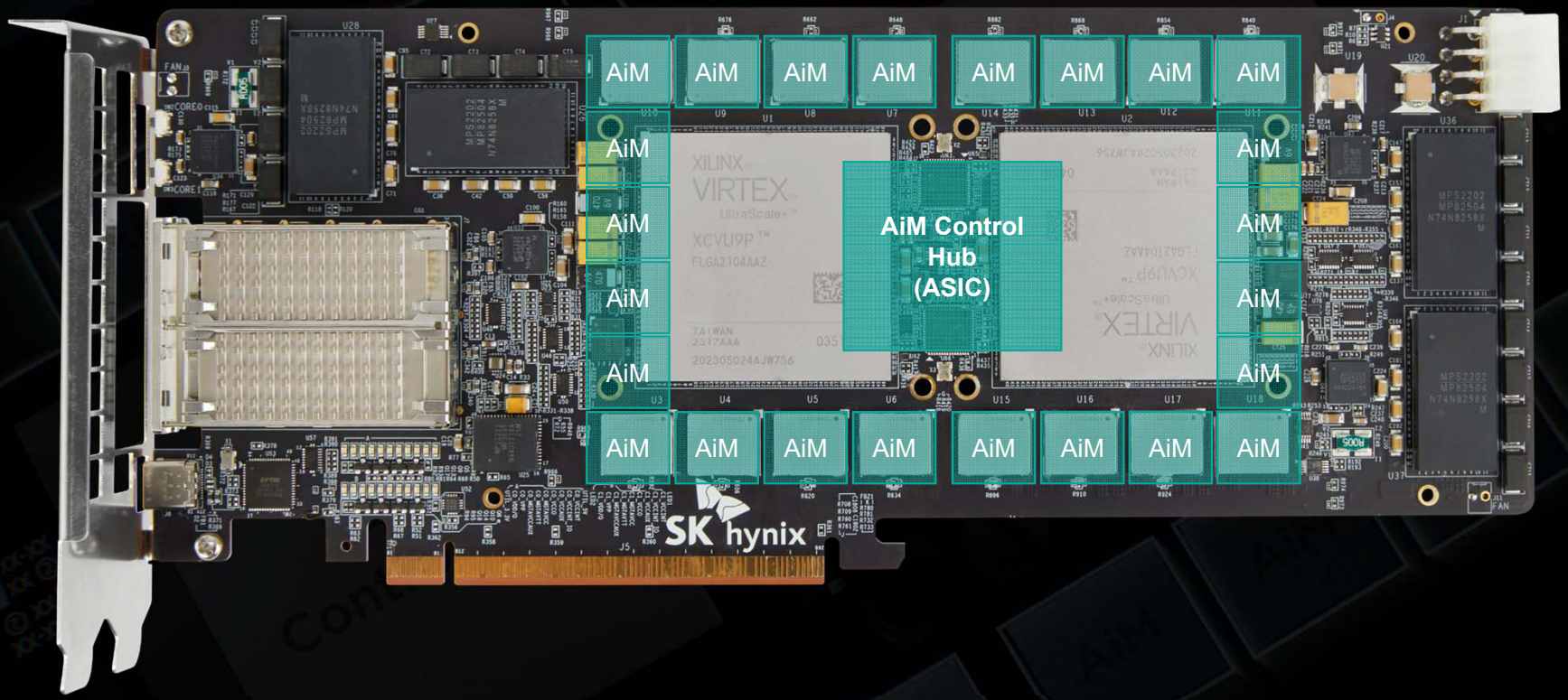
Showcase : Q&A AI app. on AiMX proto. system

- We developed an AiMX proto. card using FPGA chip and built an AiMX reference system optimized for LLM with GPU cards for Prompt Stage and AiMX cards for Response Stage
- Q&A AI application showcase with AiMX prototype system at SK hynix booth

AiMX Card prototype

AiMX reference system

AiMX prototype card



- When the AiM controller is implemented as an ASIC, more area can be used to mount more AiM chips to provide more memory capacity and more memory BW.

Direction of Next Generation AiM

- To catch up moving target of LLM service & infra trend

1. Higher memory capacity per card & server
2. Higher processing performance for multi-batching & beam search
3. Lower precision support for reduced model

Not only Datacenter market, there are client and mobile LLM computing demands


SK hynix
SK hynix Accelerator in Memory
AiMX