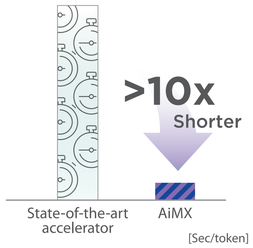
Boost Your AI: Discover the Power of PIM with SK hynix's AiM!

AiMX as Generative AI Accelerator System

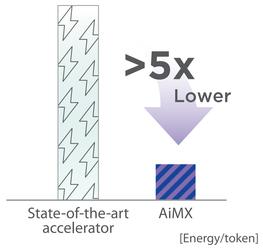
AiMX is an AiM-centric reference system targeting inference of generative AI applications.

Shorter Service Latency



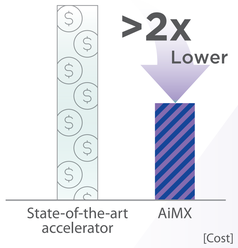
by providing much higher effective memory bandwidth

Lower Energy Consumption

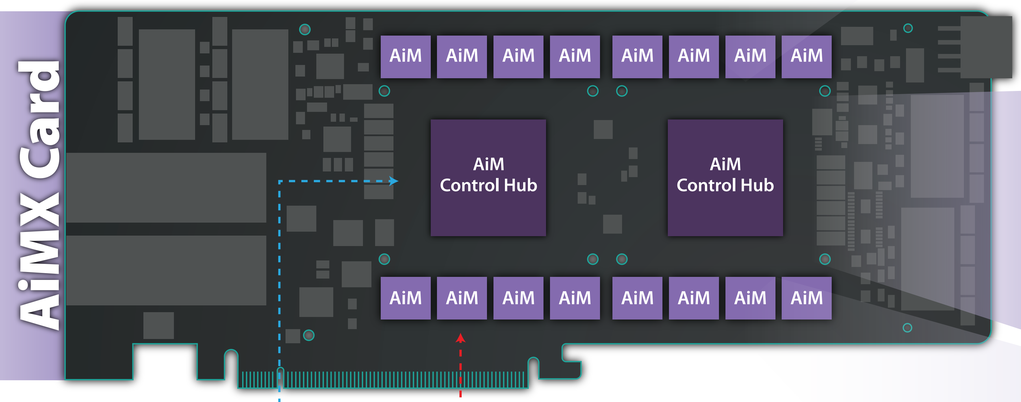


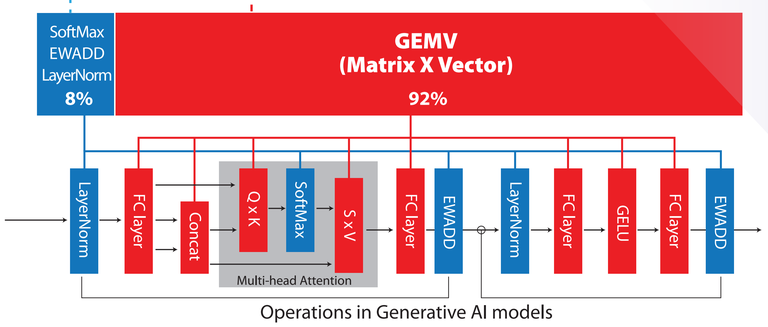
by reducing most off-chip data transfer

Lower Cost



by using lightweight controller instead of an expensive xPU

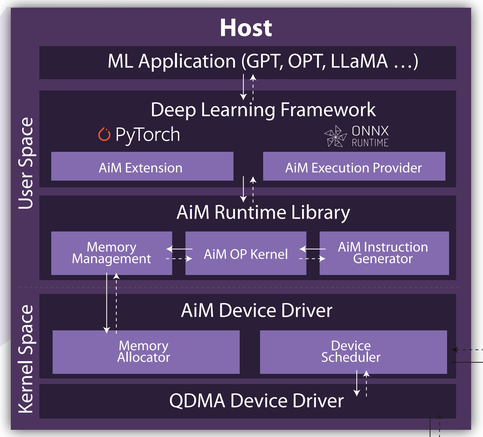




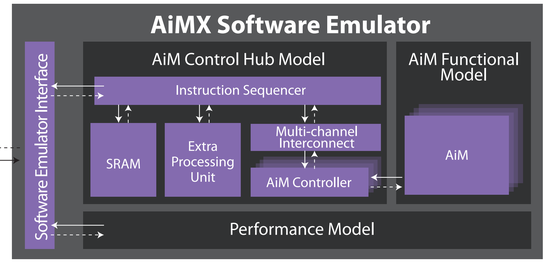
Operations in Generative AI models

Measured on GPT-3 (175B, single decoder block) with PyTorch (v2.0) using 1 x V100 GPU

AiMX Software Stack



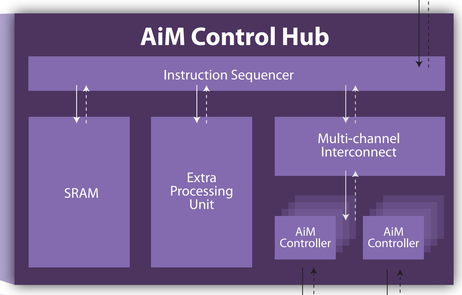
- Provides various AiM operation kernels
- Generates AiM instructions and dispatches them to the AiM control hub
- Supports AiMX software emulator for developing AI applications even without a physical hardware board



· Emulates the functionalities of the AiM control hub and AiMs
· Supports the flexible hardware model
· Supports the AiM performance analytical model

Macro Instruction

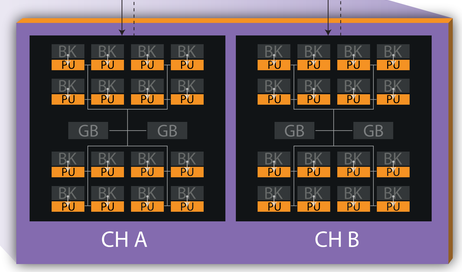
AiM Control Hub



- Controls multiple AiM chips efficiently
- Executes various small operations such as Layer Normalization and SoftMax through the extra processing unit

AiM Command

AiM



- SK hynix's first Processing-In-Memory
- Accelerates large fixed GEMV operations by exploiting true bank parallelism
- 0.5 TB/s bandwidth and 1GB capacity per chip
- Large non-reused weight matrices located in each bank
- Small reused input vectors located in Global Buffer (GB)