Latency Processing Unit (LPU™) for Acceleration of Hyperscale AI Models

Seungjae Moon, Seongmin Hong, Junseo Cha, Gyubin Choi,
Jung-Hoon Kim, Junsoo Kim, Sekyong Song, and Joo-Young Kim

HyperAccel Co., Ltd.

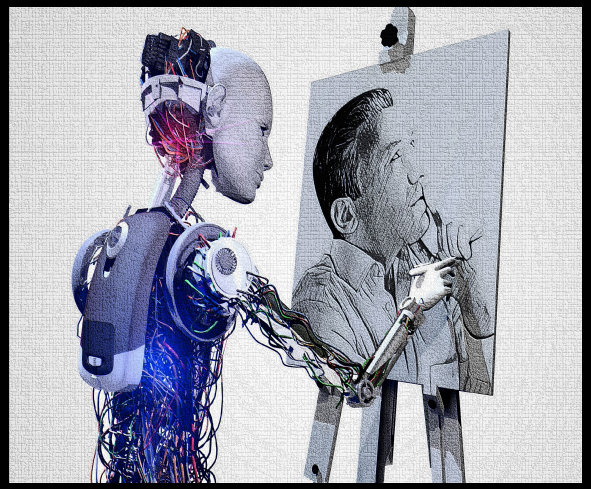
HYPER ACCEL

HYPER-ACCELERATED SOLUTIONS FOR MISSION-CRITICAL WORKLOADS

Outline

- Hyperscale AI Models
- LPU™ Architecture
- Performance Results
- Summary

Generative AI



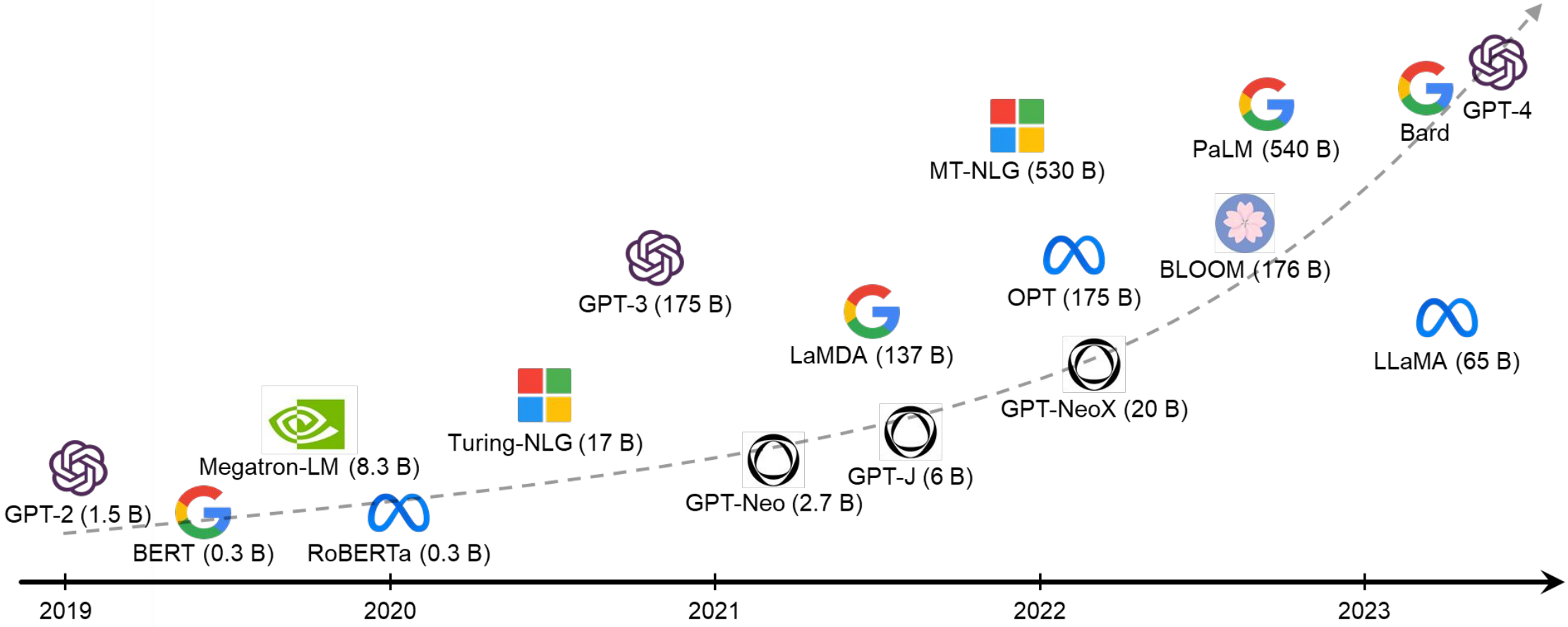
- Figure 1: “Image of artificial intelligence drawing a man on a canvas”
- Figure 2: “What if Picasso painted Elon Musk?”

Generative AI

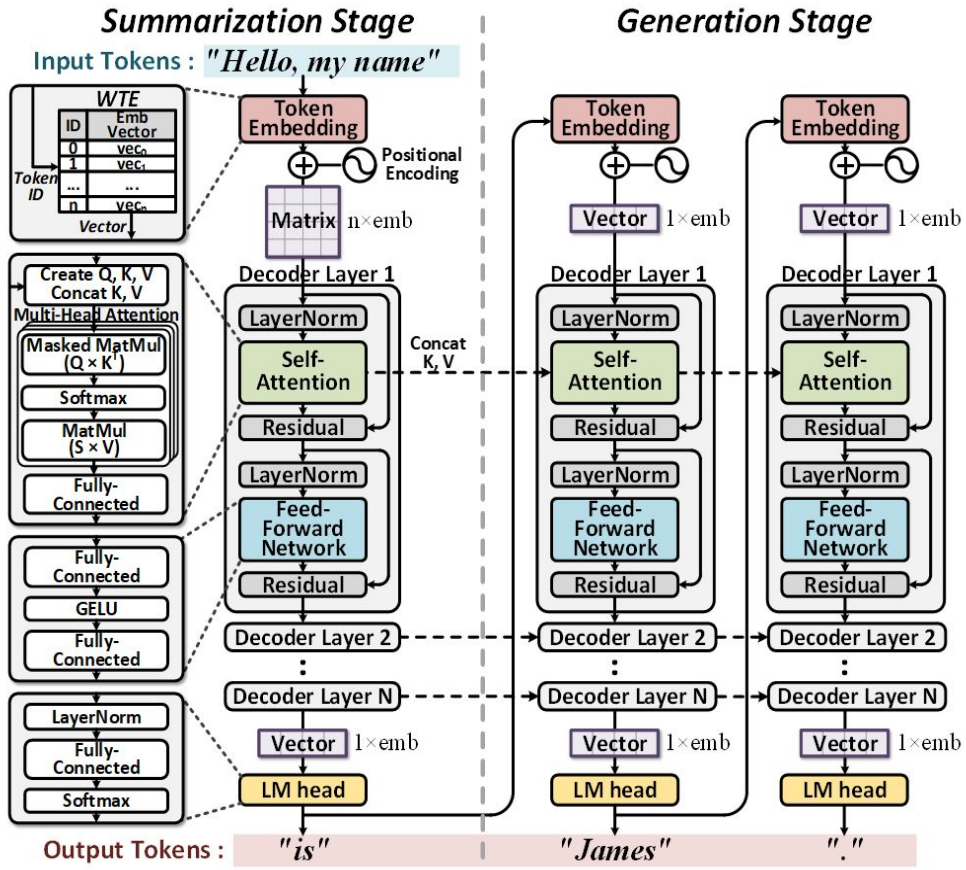
CREATIVITY

- Figure 1: “Image of artificial intelligence drawing a man on a canvas”
- Figure 2: “What if Picasso painted Elon Musk?”

Hyperscale AI and Large Language Model (LLM)



LLM Training vs. Inference



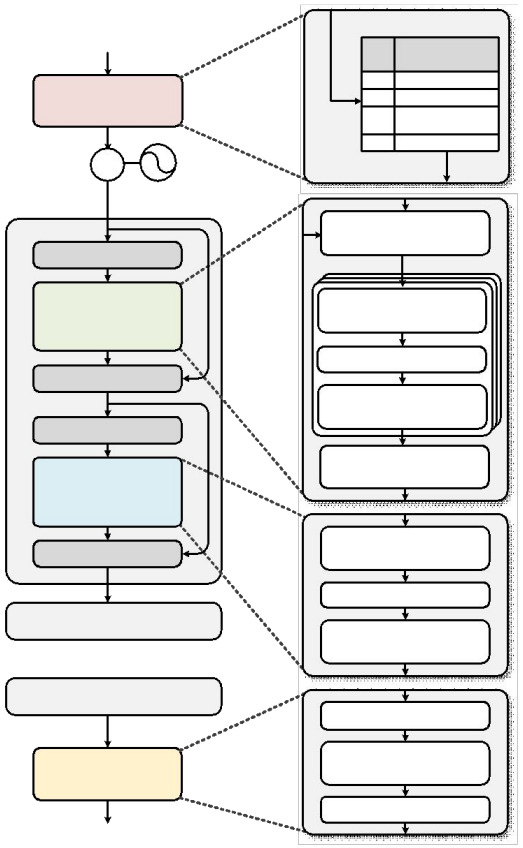
Training

- Large batch
- Compute-intensive
- Throughput-oriented hardware

Inference

- Small batch
- Memory-intensive
- Latency-oriented hardware

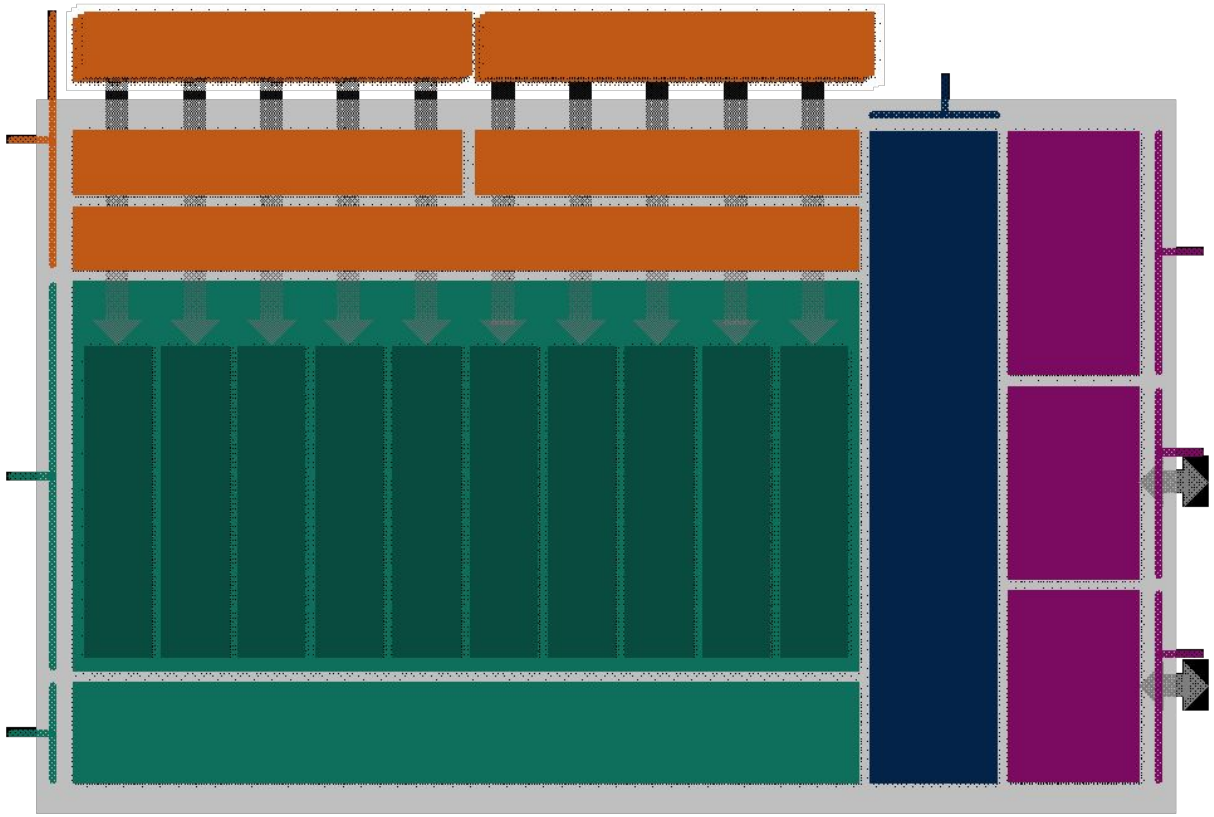
Requirements for LLM Inference Accelerator



- **Execution of vector and matrix operations**
  due to distinct model structure
- **Optimization for small batch size**
  due to different request conditions
- **Maximum usage of memory bandwidth**
  due to memory bottleneck
- **Parallelization and scalable network**
  due to LLM's growing computational and physical memory requirements

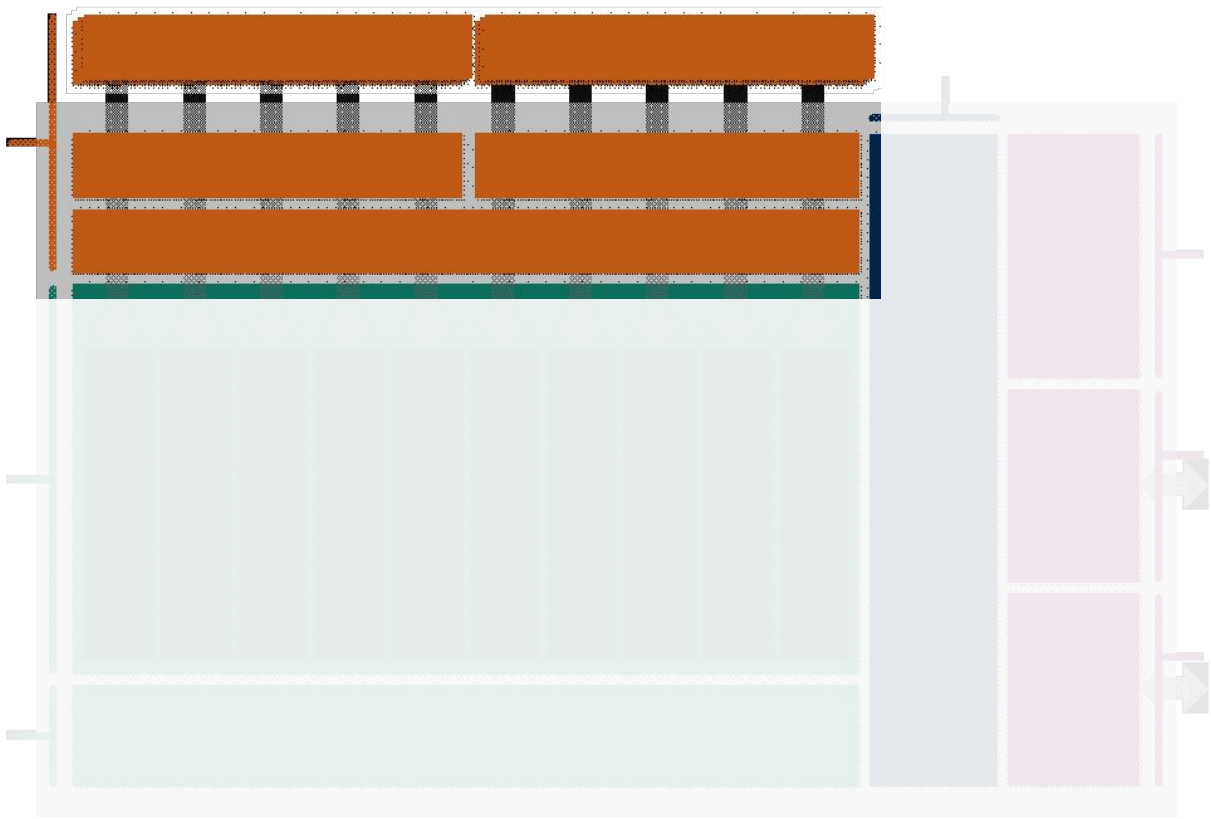
Latency Processing Unit (LPU™)

- World-first hardware accelerator for large language model inference



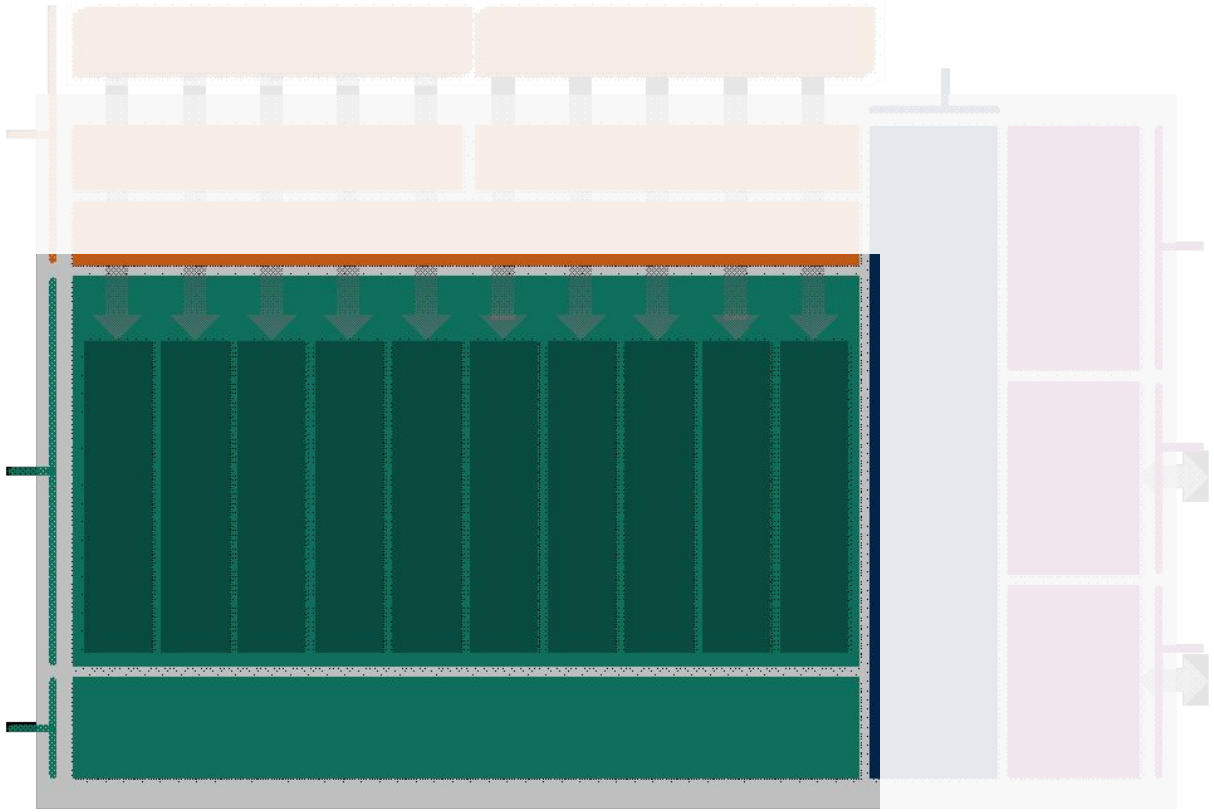
Latency Processing Unit (LPU™)

- World-first hardware accelerator for large language model inference



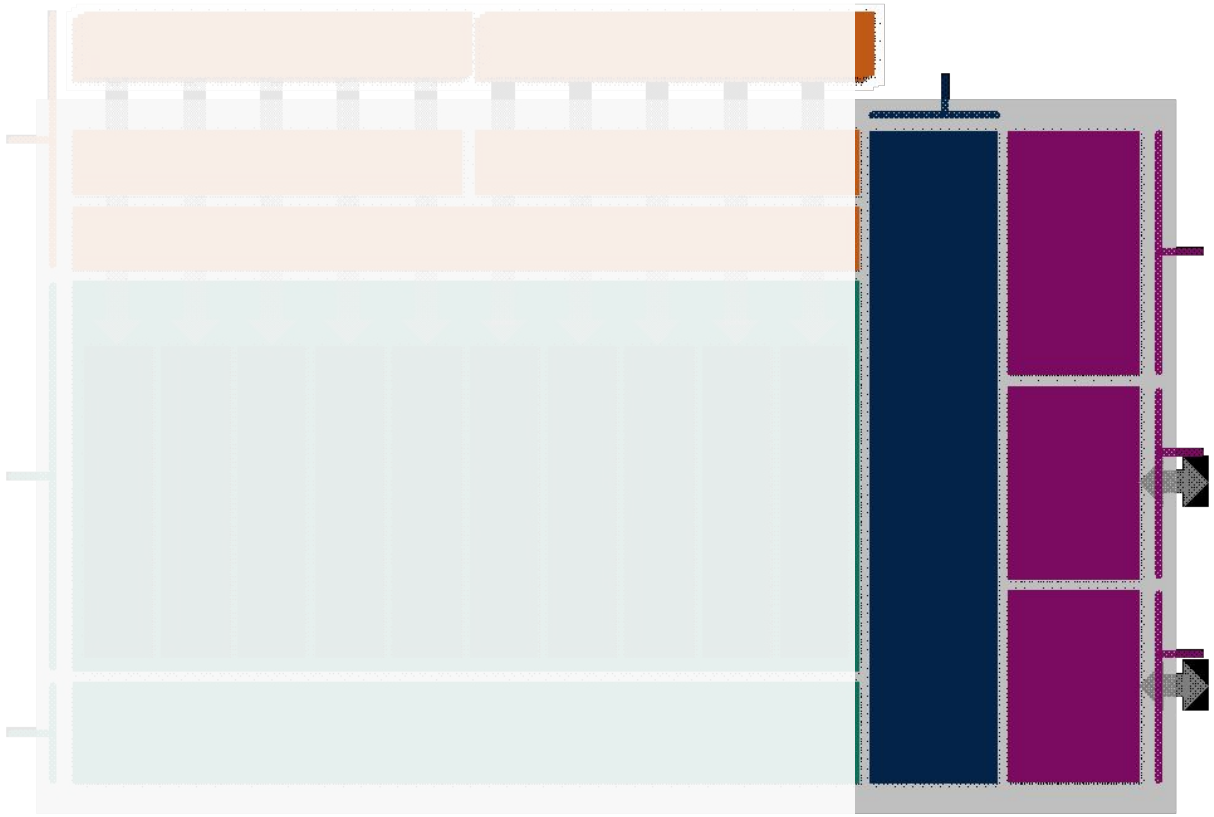
Latency Processing Unit (LPU™)

- World-first hardware accelerator for large language model inference



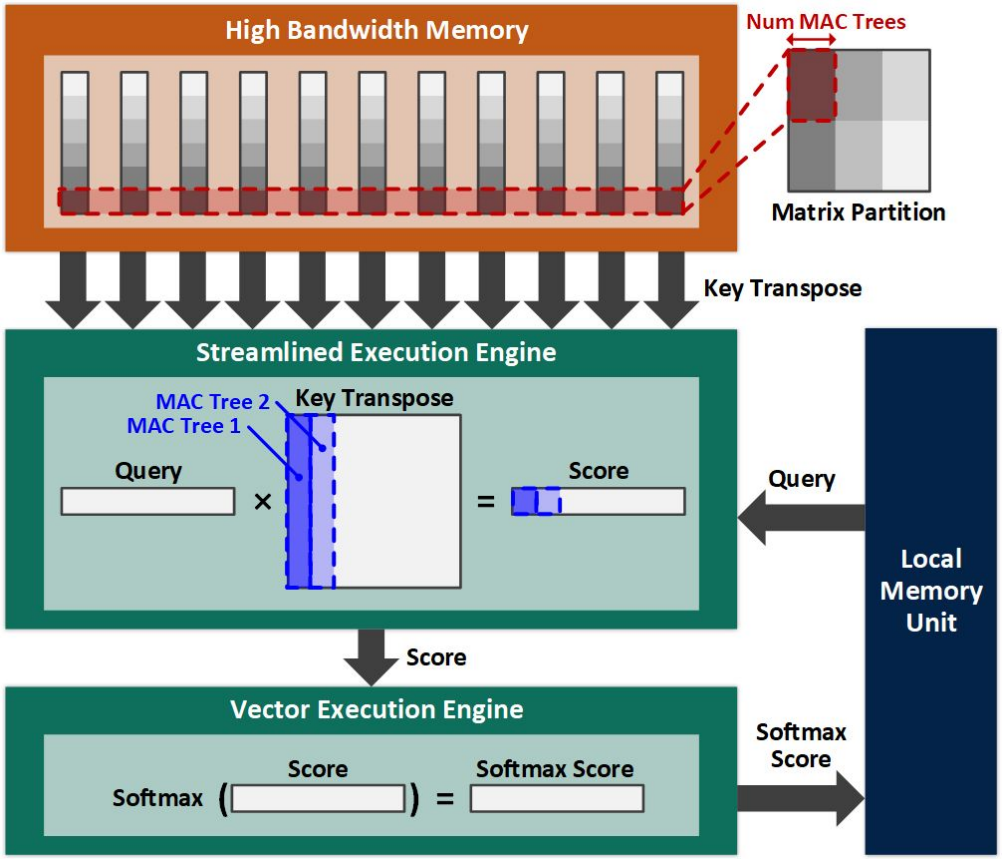
Latency Processing Unit (LPU™)

- World-first hardware accelerator for large language model inference



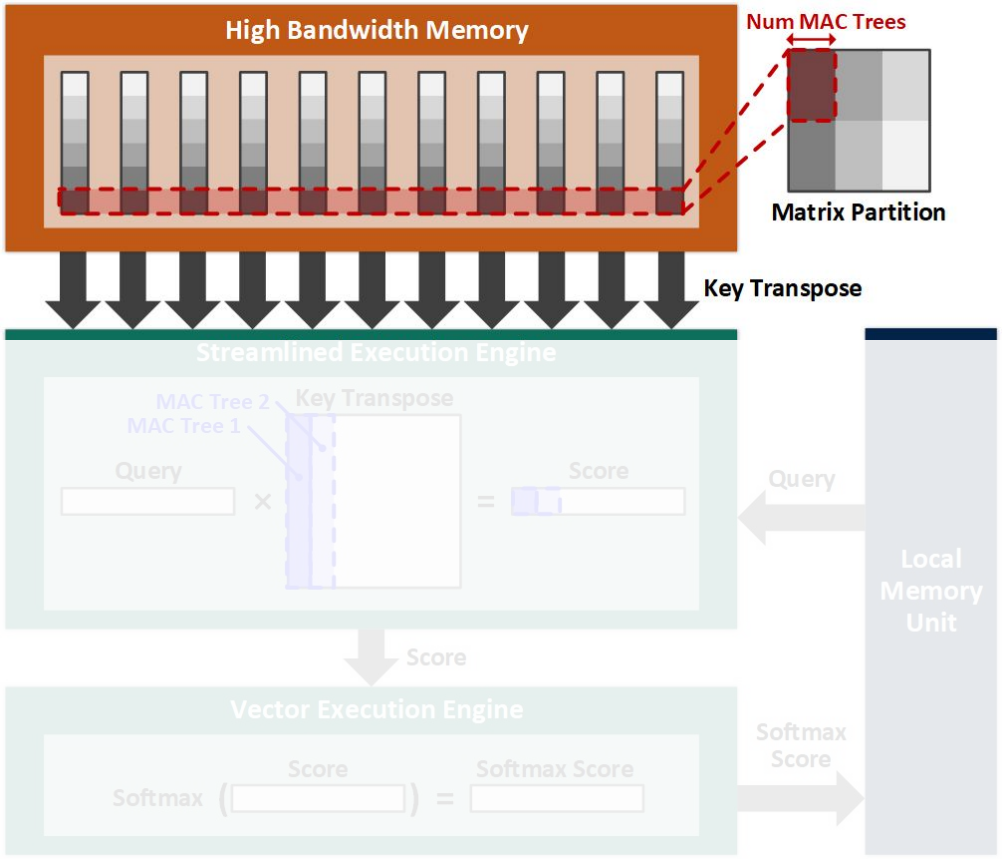
LPU™ Microarchitecture

- Maximize the memory bandwidth usage with optimized tiling scheme (~90%)
- Transpose operation using the strobe signal when writing data to memory
- Low-latency and high-throughput execution engine with number of MAC trees that exactly matches the incoming bandwidth
- Out of order scheduling to allow both engines to work on independent operations when required



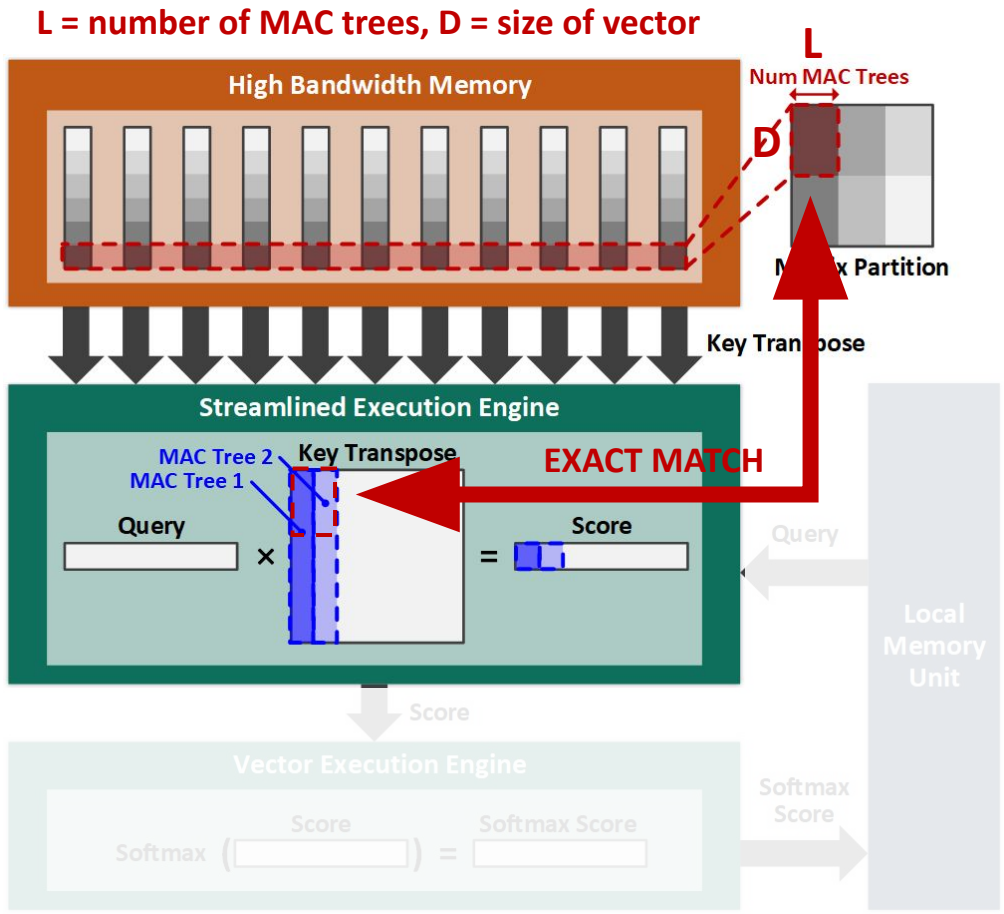
LPU™ Microarchitecture

- Maximize the memory bandwidth usage with optimized tiling scheme (~90%)
- Transpose operation using the strobe signal when writing data to memory
- Low-latency and high-throughput execution engine with number of MAC trees that exactly matches the incoming bandwidth
- Out of order scheduling to allow both engines to work on independent operations when required



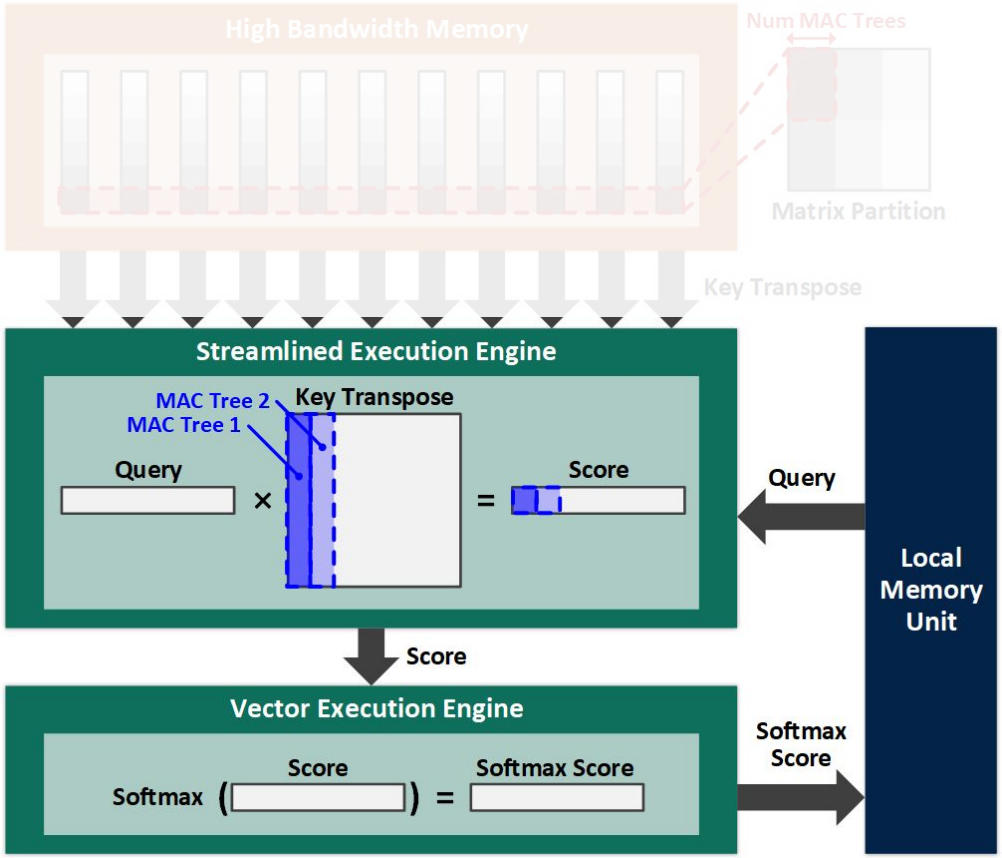
LPU™ Microarchitecture

- Maximize the memory bandwidth usage with optimized tiling scheme (~90%)
- Transpose operation using the strobe signal when writing data to memory
- Low-latency and high-throughput execution engine with number of MAC trees that exactly matches the incoming bandwidth
- Out of order scheduling to allow both engines to work on independent operations when required



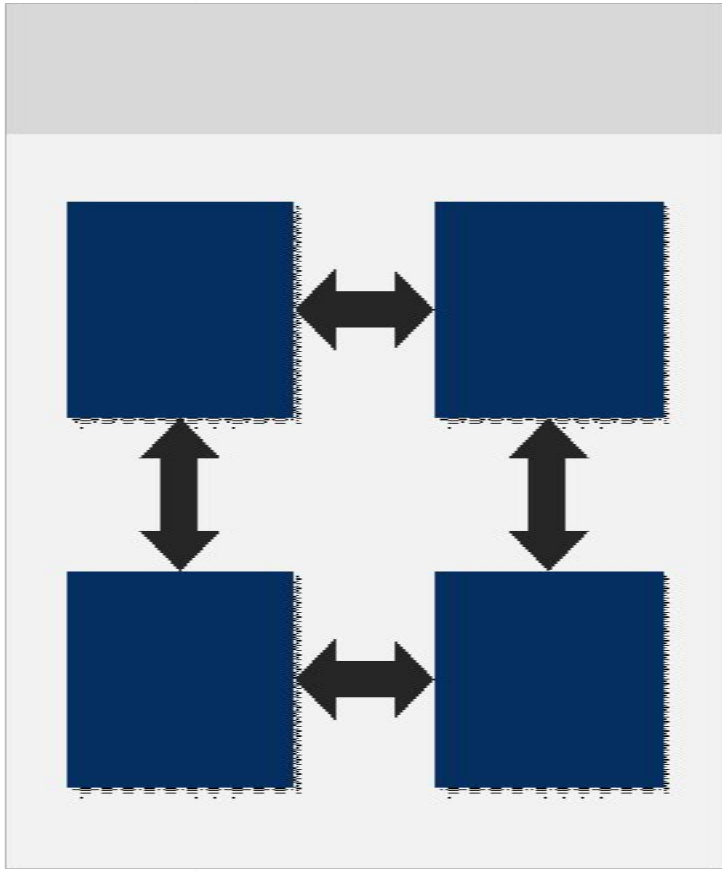
LPU™ Microarchitecture

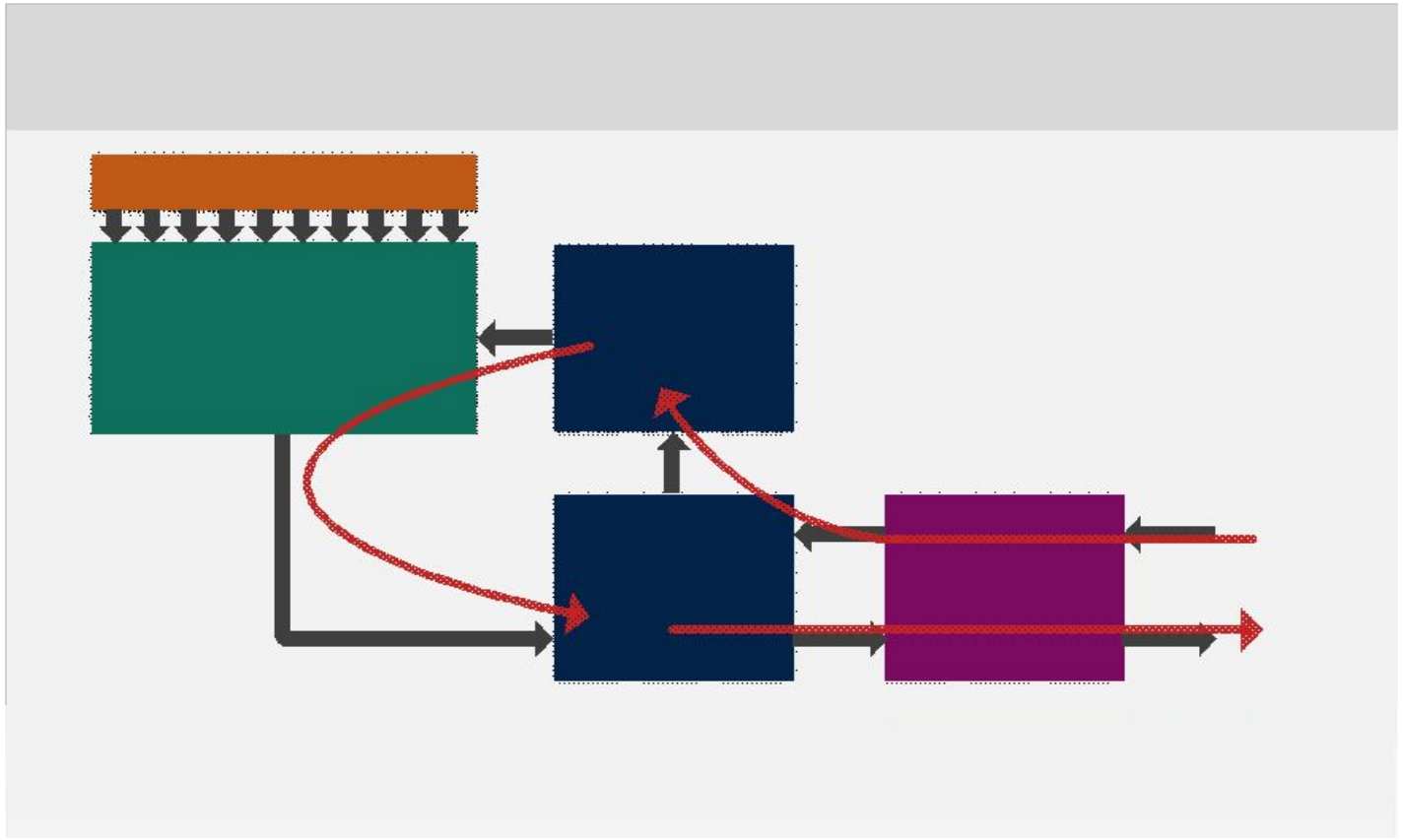
- Maximize the memory bandwidth usage with optimized tiling scheme (~90%)
- Transpose operation using the strobe signal when writing data to memory
- Low-latency and high-throughput execution engine with number of MAC trees that exactly matches the incoming bandwidth
- Out of order scheduling to allow both engines to work on independent operations when required



Expandable Synchronization Link (ESL)

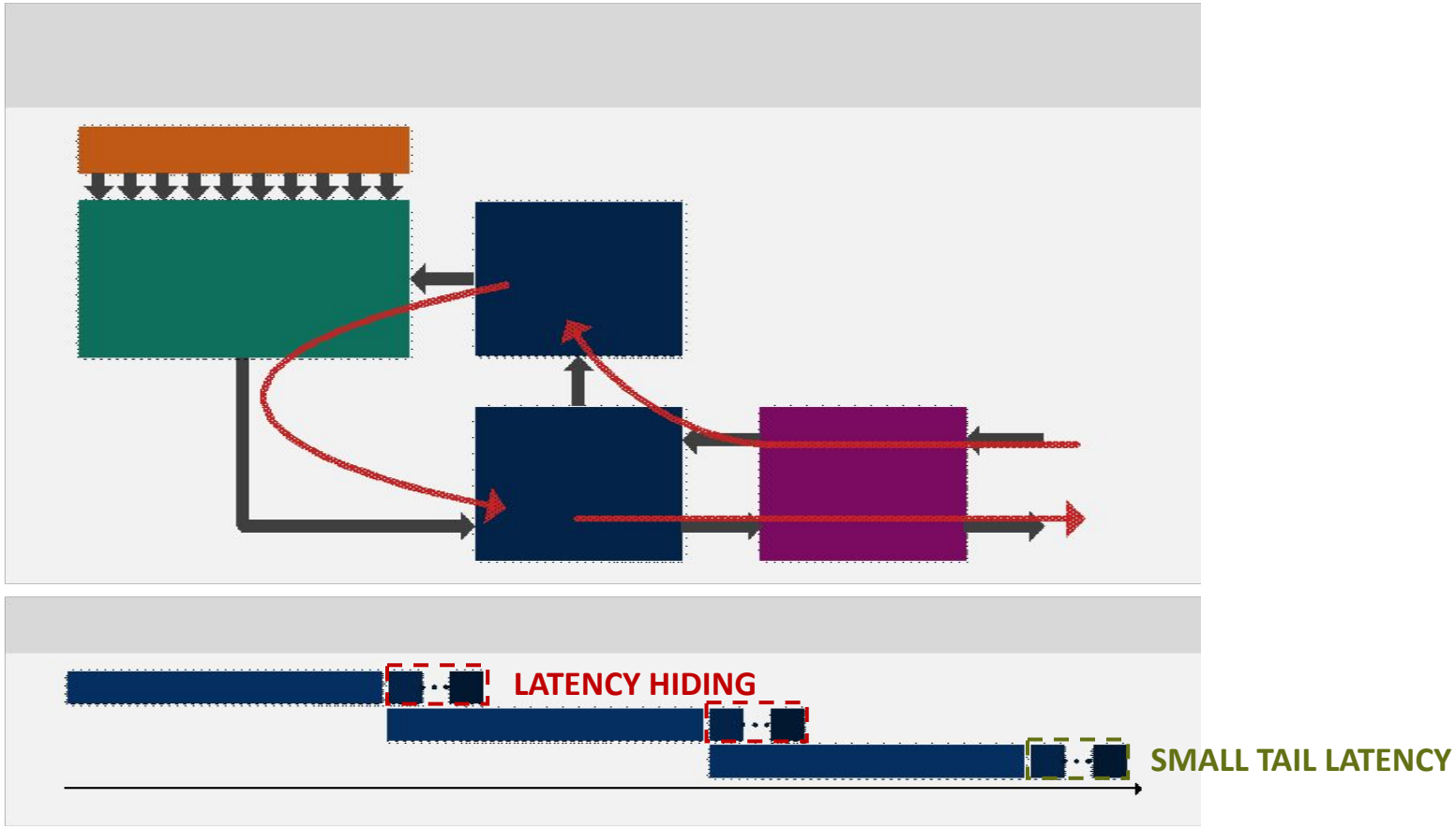
- Highly expandable network that performs data synchronization with latency hiding





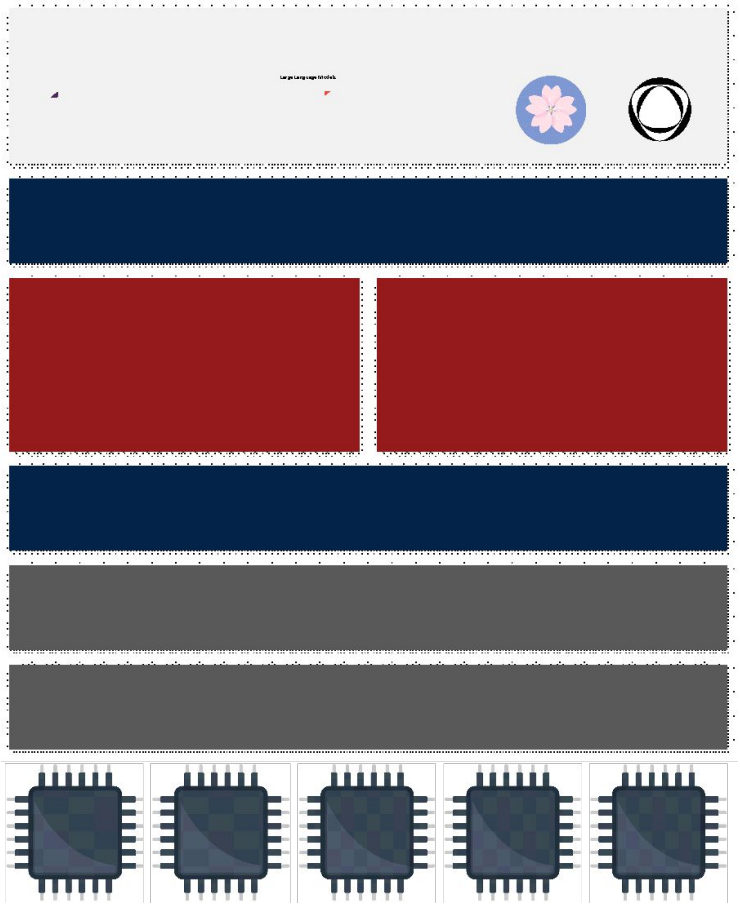
Expandable Synchronization Link (ESL)

- Highly expandable network that performs data synchronization with latency hiding

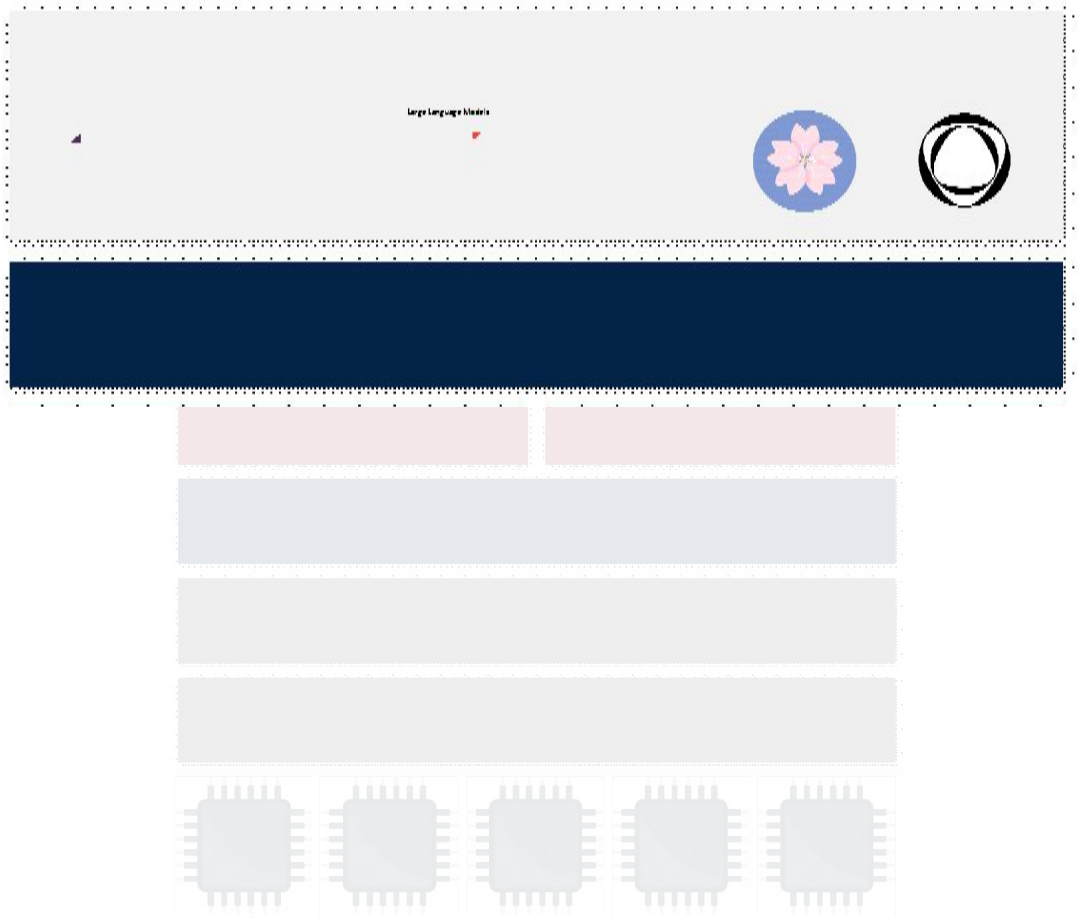


HyperDex Software Stack

- Bridging large language models to LPU-based systems

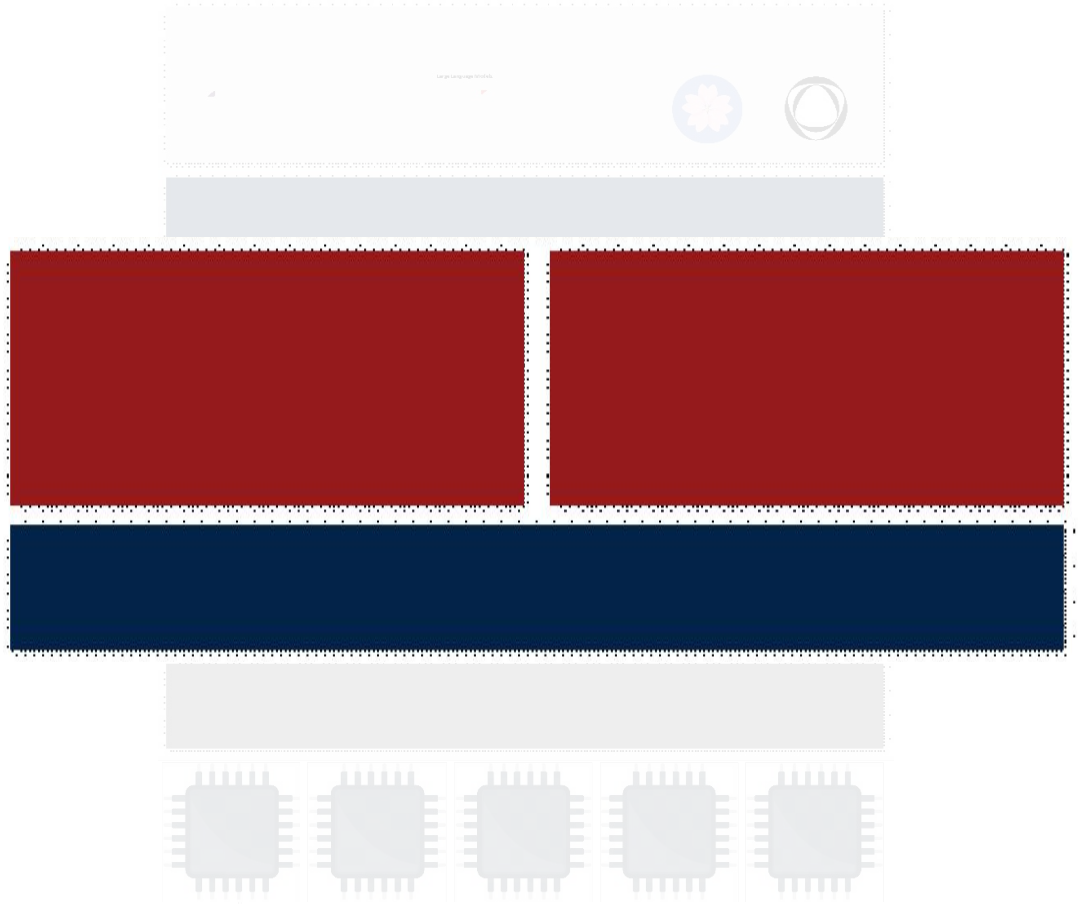


HyperDex Software Stack



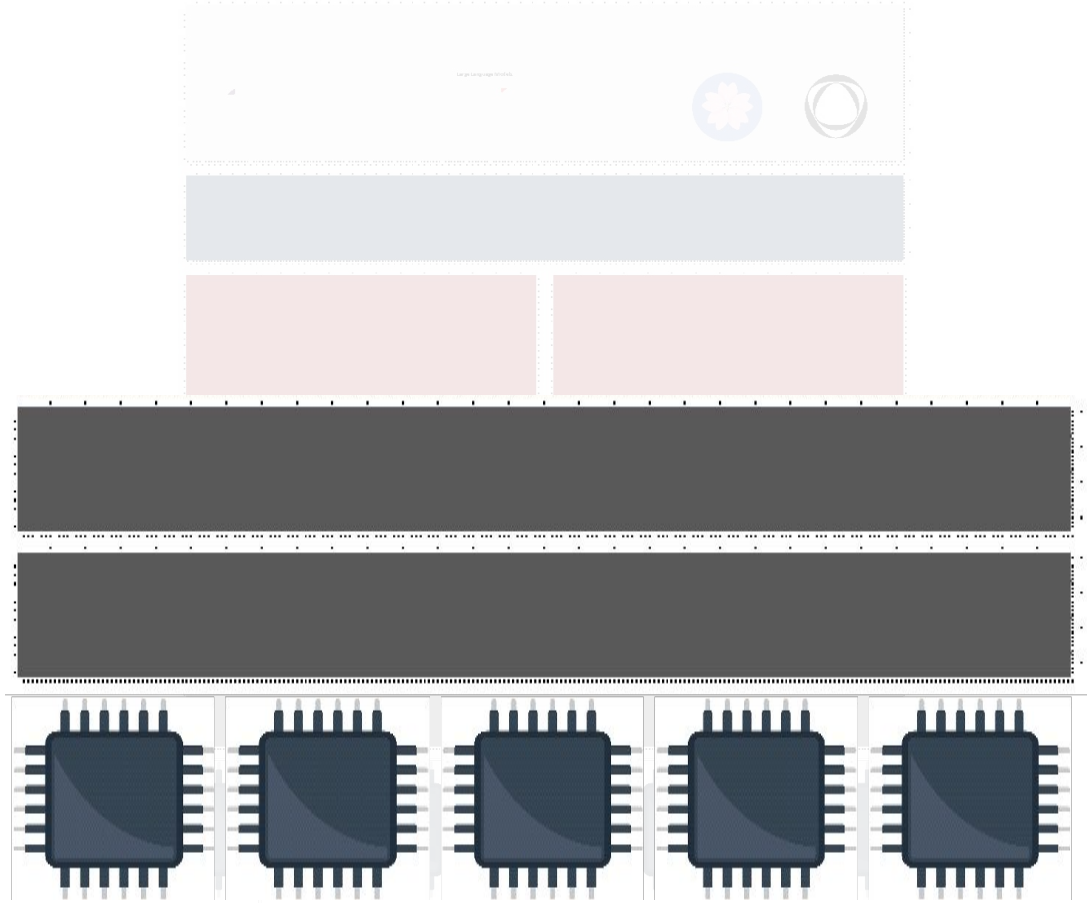
- Support various large language models
- Intra-layer parallelism for self-attention and feed-forward network
- Partitions the models parameters across multiple devices
- Optimal memory allocation and alignment of model parameters
- Parallel instruction chaining for maximum latency saving

HyperDex Software Stack

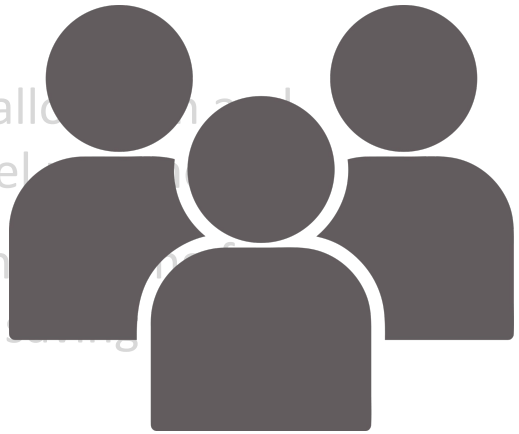


- Support various large language models
- Intra-layer parallelism for self-attention and feed-forward network
- Partitions the models parameters across multiple devices
- Optimal memory allocation and alignment of model parameters
- Parallel instruction chaining for maximum latency saving

HyperDex Software Stack

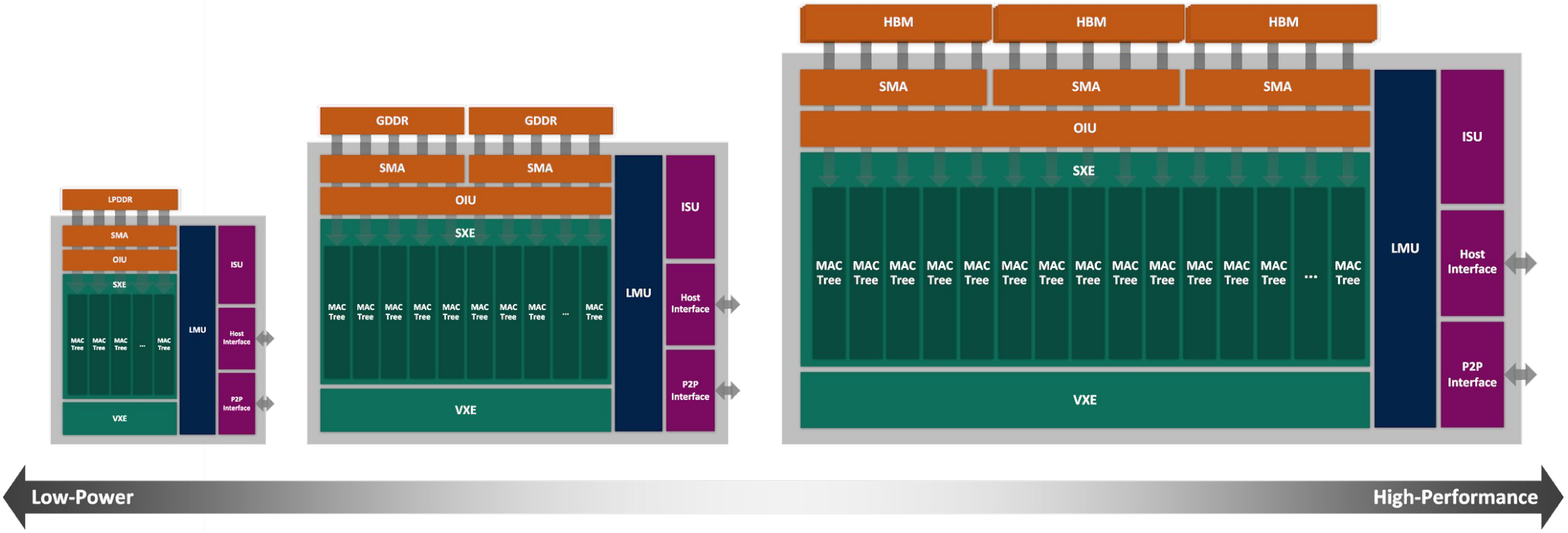


- Support various large language models
- Intra-layer parallelism for self-attention and feed-forward network
- Partitions the models parameters across multiple devices
- Optimal memory allo[illegible] and alignment [illegible] odel
- Parallel ins[illegible]action [illegible] imu[illegible]atency [illegible]



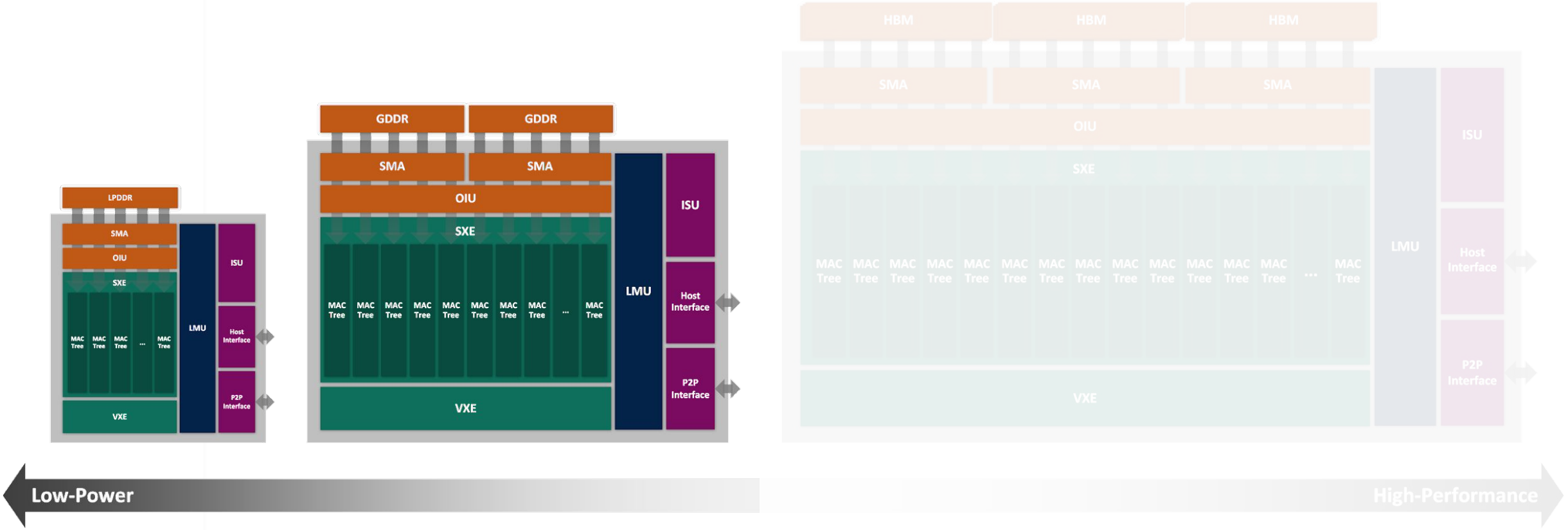
LPU™ IP Products

- Highly scalable hardware solution from low-power to high-performance



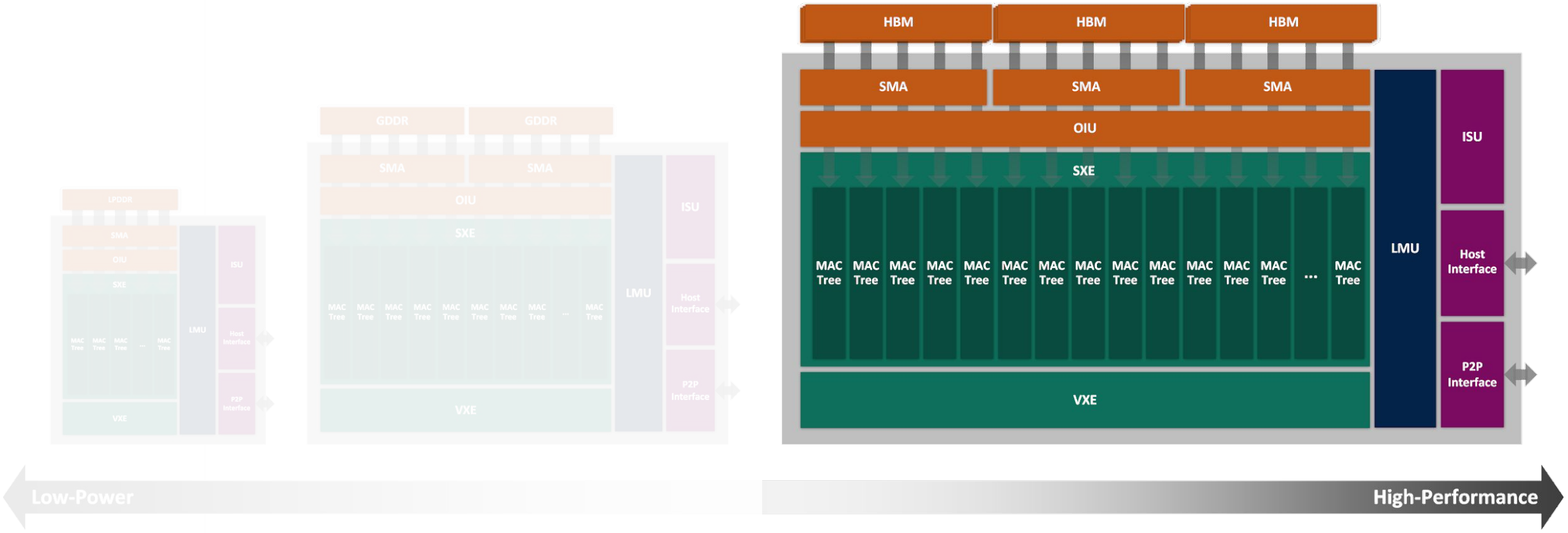
LPU™ IP Products

- Highly scalable hardware solution from low-power to high-performance

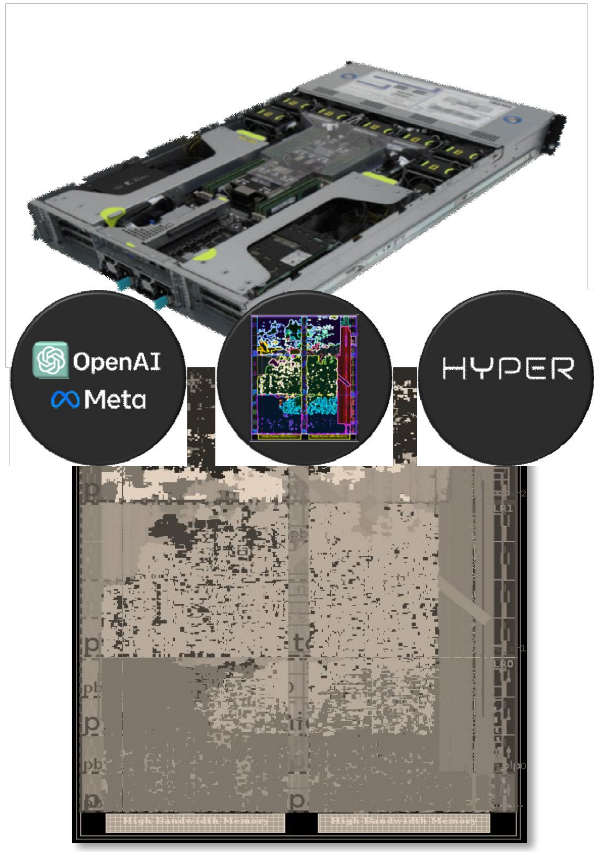


LPU™ IP Products

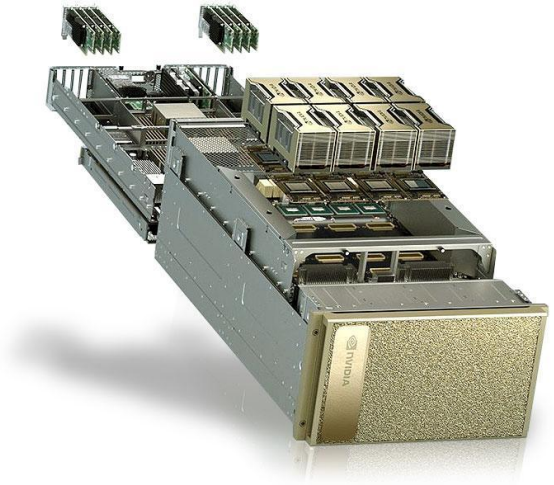
- Highly scalable hardware solution from low-power to high-performance



Performance Results



16x LPU™ Implementation on AMD Alveo U55C



NVIDIA/
FasterTransformer

NVIDIA

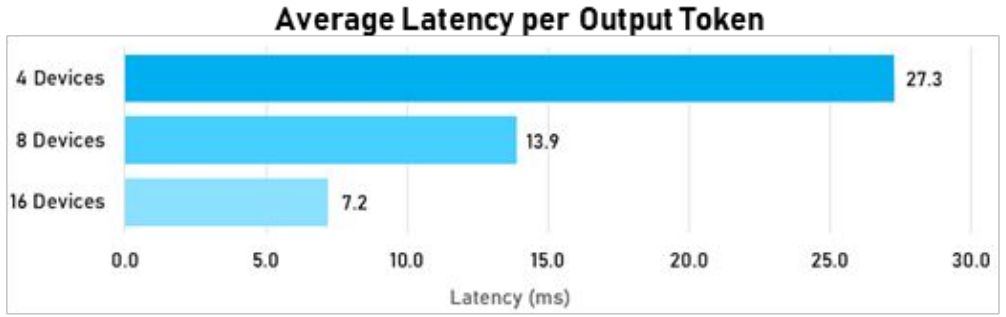
Transformer related optimization, including BERT, GPT

35 Contributors · 192 Issues · 4k Stars · 659 Forks

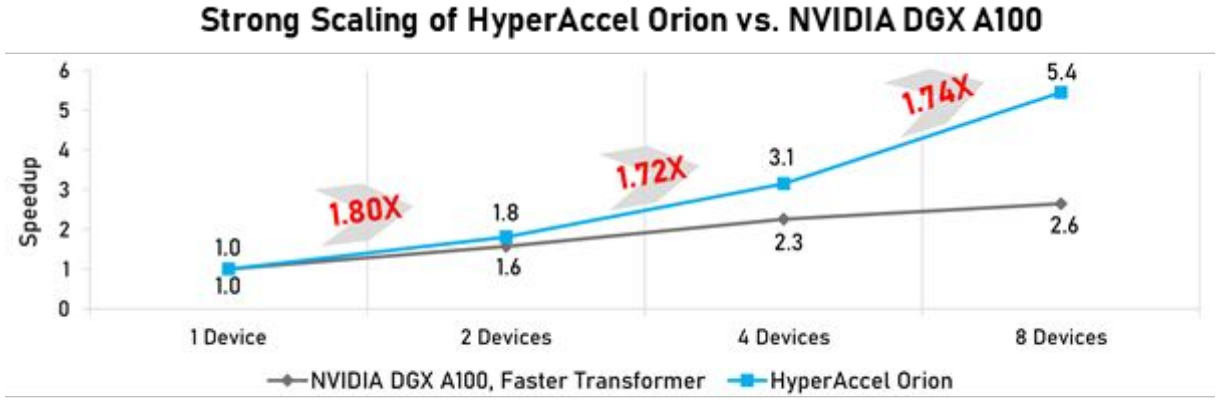
NVIDIA DGX A100 running FasterTransformer Library

Performance Results

- **Millisecond** (7.2ms) latency per output token and **1.76X** scalability



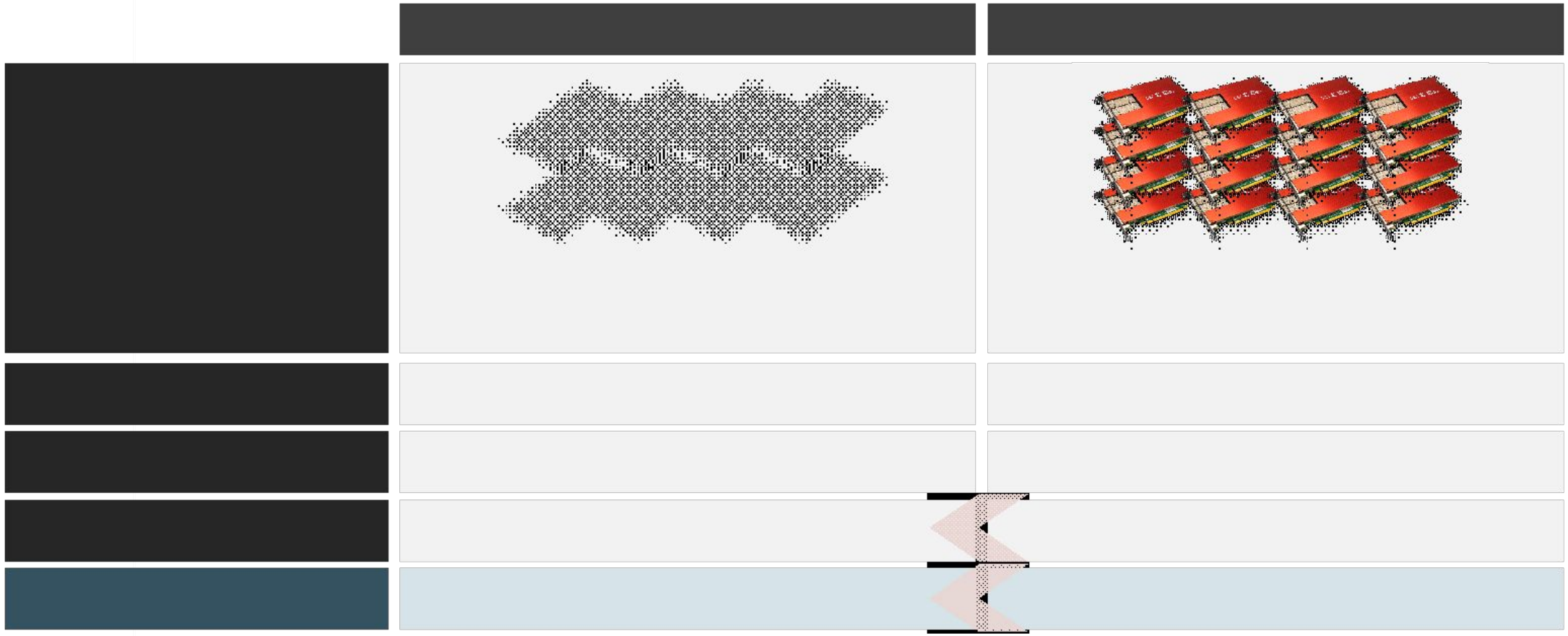
Millisecond (7.2ms) to generate an output token during GenAI inference



1.76x speedup on average for doubling the number of devices

Performance Results

- **1.49X** speedup and **2.35X** cost-effectiveness



Summary

- **We present LPU™, a latency-oriented architecture for hyperscale AI models**
  - ✔ Compute core supporting end-to-end transformer operations including various matrix, vector, and non-linear functions
  - ✔ Maximizing HBM bandwidth with an optimized tiling scheme and dataflow
  - ✔ Parallel computation using Intra-layer model parallelism and peer-to-peer communication

- **LPU™ provides the highest efficiency for hyperscale AI models**
  - ✔ Multi-LPU server achieves **1.49X speedup and 2.35X cost-effectiveness** compared to the GPU counterpart DGX A100 with state-of-the-art transformer library (FasterTransformer)
  - ✔ Achieving **high scalability of 1.76X** speedup for doubling the number of devices

- **We are building LPU silicon IPs, SW stack, and utilities**
  - ✔ Various configuration of hardware accelerator, model types and sizes, quantization scheme
  - ✔ Proof of concepts (POCs) and deployments in companies, large and small

Thank You

Questions? Feel Free to contact us!

- Email: sj.moon@hyperaccel.ai
- Website: www.hyperaccel.ai

Poster Session 6:00-7:00 PM

This will appear at HotChips 2023

- *"HyperAccel LPU: Accelerating Hyperscale Models for Generative AI"*