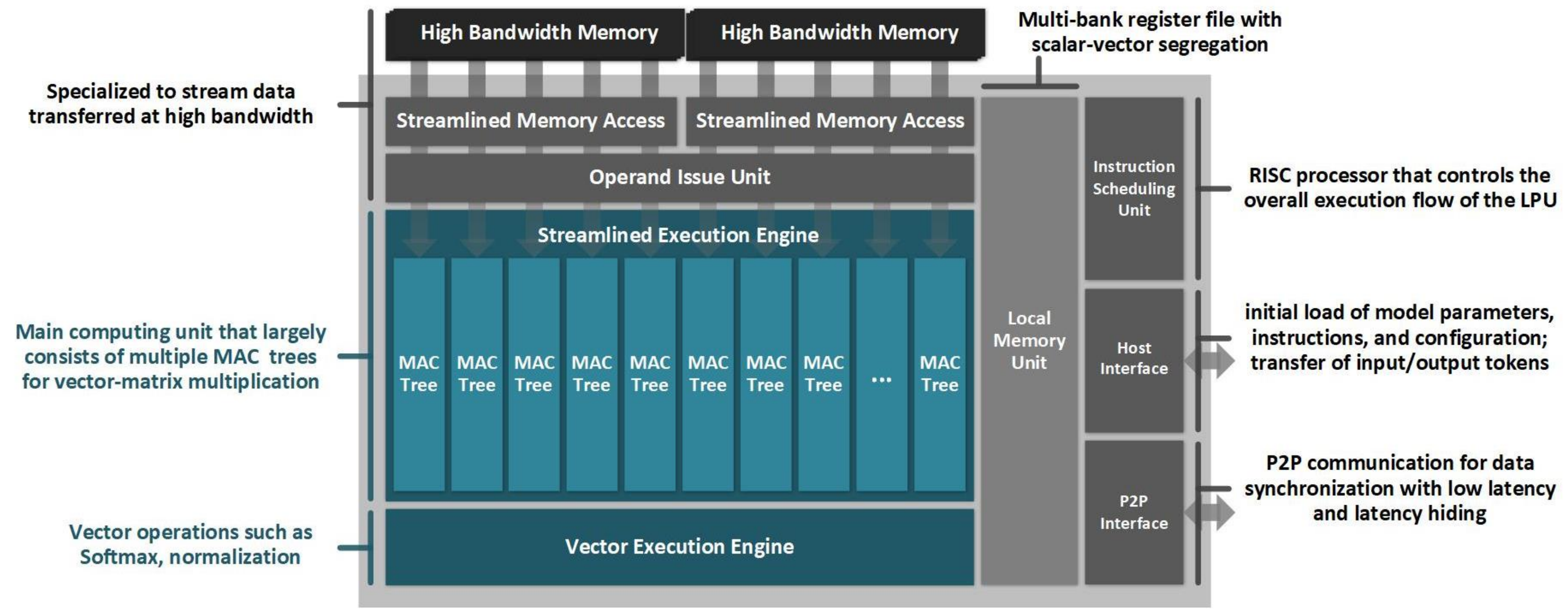
HyperAccel Latency Processing Unit (LPU^{TM}) Accelerating Hyperscale Models for Generative AI

Seungjae Moon, Junsoo Kim, Jung-Hoon Kim, Junseo Cha, Gyubin Choi, Seongmin Hong, and Joo-Young Kim
HyperAccel, Hwaseong-si, Republic of Korea

Introduction

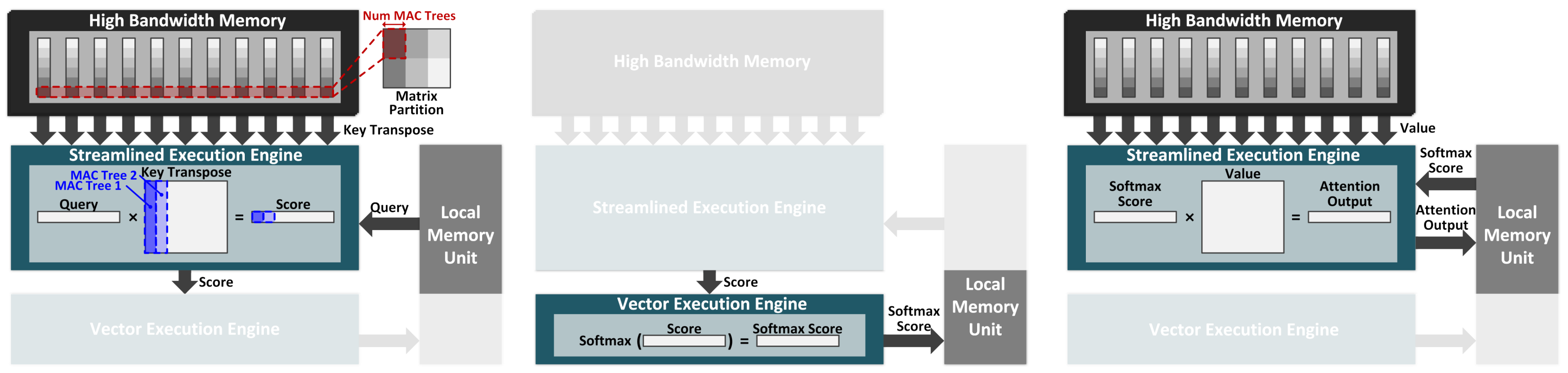
- The fundamental goal of AI is to create human-like intelligence. Generative AI has enabled AI to do what we thought was innate to only humans: show creativity.
- Transformer-based large language models (LLM) with multi-billion parameters, such as OpenAI GPT, Meta LLaMA, can create original texts and visual contents.
- For efficient model Inference, a latency-oriented and scalable hardware for small-batch memory-intensive workloads is required to meet the needs of different users
- **Latency Processing Unit**, the world-first hardware accelerator dedicated for the end-to-end inference of LLM.

LPU^{TM} Architecture


High Bandwidth Memory
High Bandwidth Memory
Multi-bank register file with scalar-vector segregation
Specialized to stream data transferred at high bandwidth
Streamlined Memory Access
Streamlined Memory Access
Operand Issue Unit
Streamlined Execution Engine
Main computing unit that largely consists of multiple MAC trees for vector-matrix multiplication
MAC Tree
...
Local Memory Unit
Instruction Scheduling Unit
RISC processor that controls the overall execution flow of the LPU
Host Interface
initial load of model parameters, instructions, and configuration; transfer of input/output tokens
P2P Interface
P2P communication for data synchronization with low latency and latency hiding
Vector operations such as Softmax, normalization
Vector Execution Engine

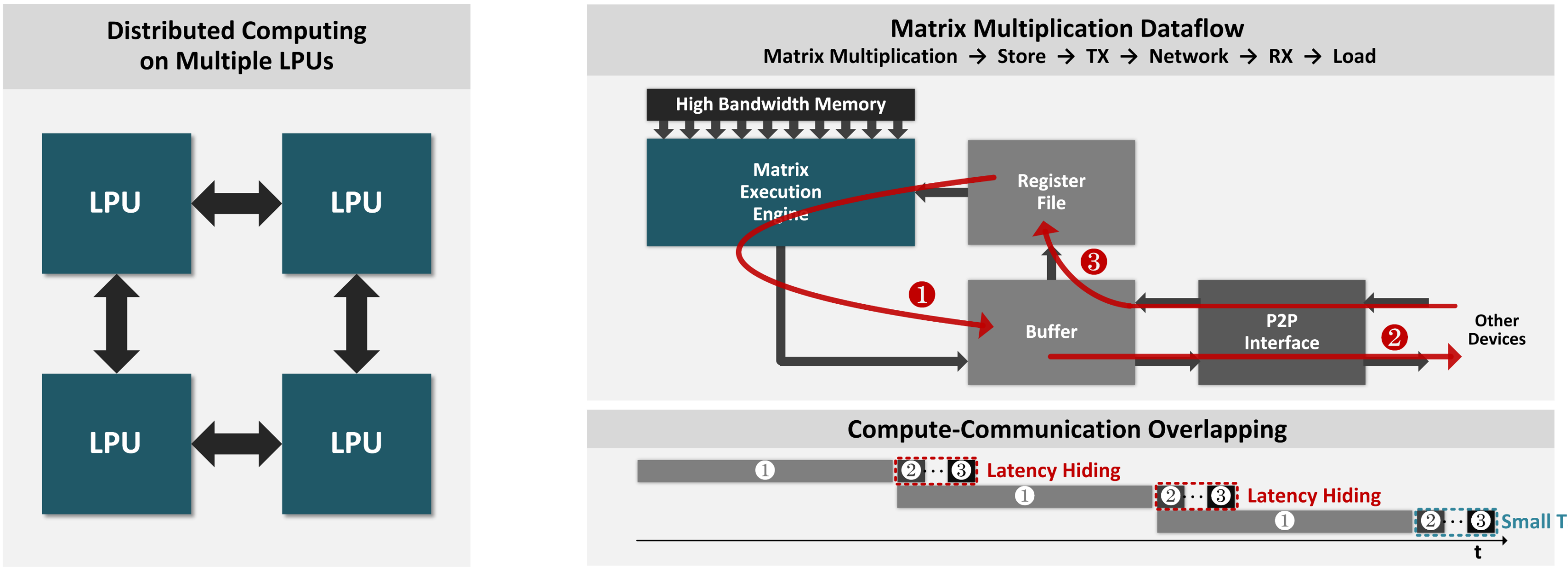
- Connects all channels of high bandwidth memory to the execution engines with datapath that exactly matches the incoming bandwidth
- Utilizes hardware-aware memory mapping and tiling to remove the need for any data reshaping and switching
- Consists of low-latency and high-throughput custom multiply-accumulate (MAC) trees, multi-precision arithmetic function unit, and special function units
- Out-of-order scheduling to allow simultaneous execution of independent matrix and vector operations for maximum hardware utilization
- Achieves effective bandwidth usage of 90% during end-to-end LLM inference

Illustration of Attention Operation



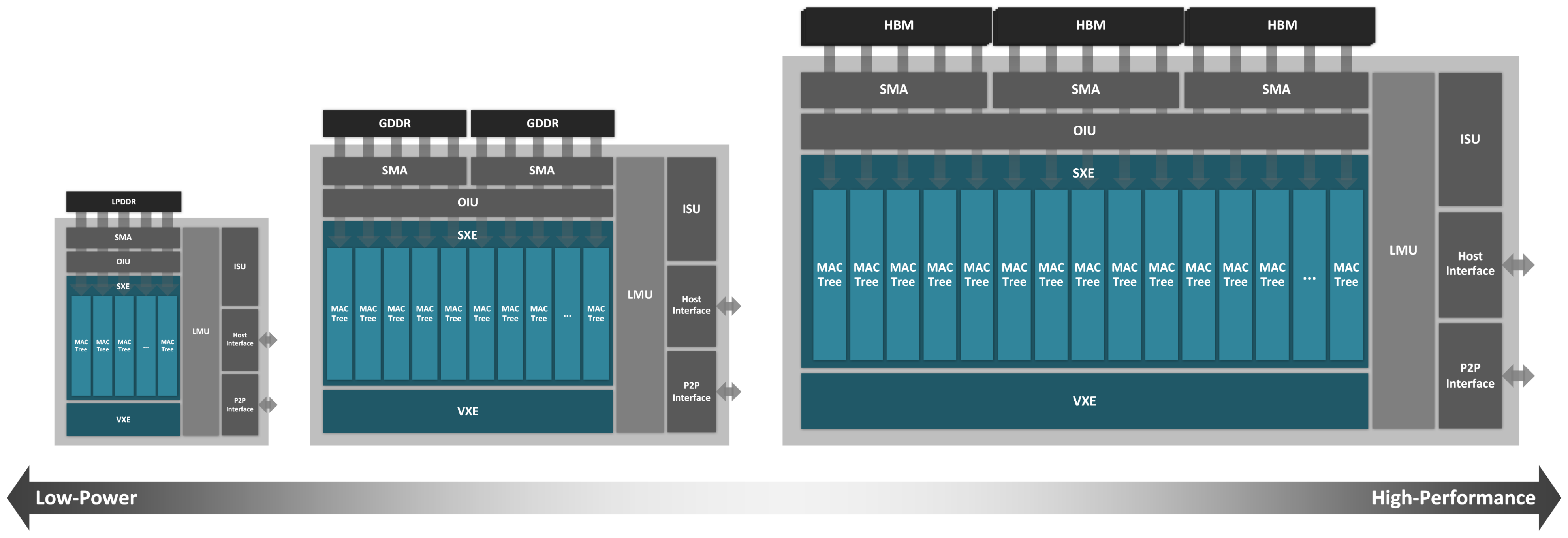
(1) Query * Key Transpose (2) Softmax (3) Softmax Score * Value

Expandable Synchronization Link (ESL)


Distributed Computing on Multiple LPUs
LPU
LPU
LPU
LPU
Matrix Multiplication Dataflow
Matrix Multiplication → Store → TX → Network → RX → Load
High Bandwidth Memory
Matrix Execution Engine
Register File
Buffer
P2P Interface
Other Devices
1
2
3
Compute-Communication Overlapping
Latency Hiding
Latency Hiding
Small Tail Latency
t

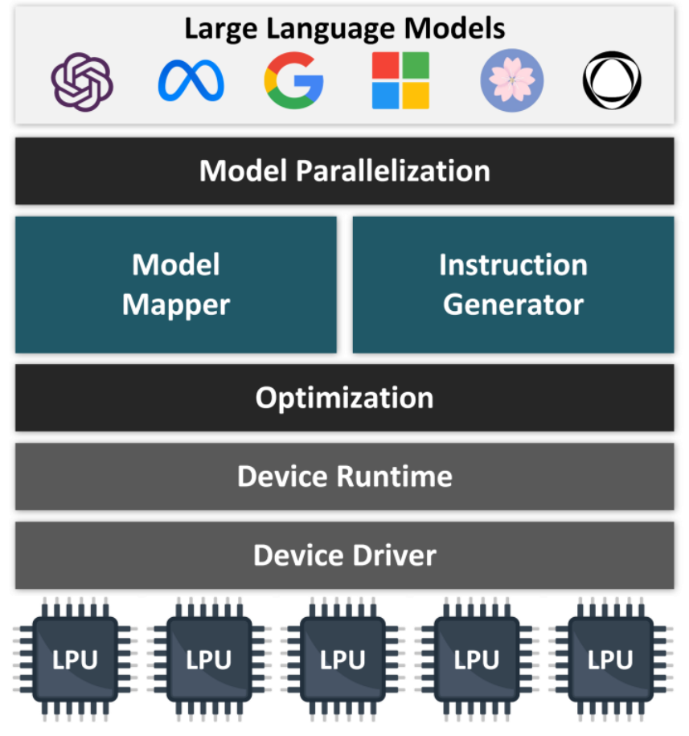
- Lightweight full-duplex peer-to-peer (P2P) communication technology that performs data synchronization with low latency and latency hiding
- Low-latency by minimal packet overhead, direct path I/O, and short dataflow
- Latency-hiding by custom protocol that enables execution and data synchronization to continuously run in tandem to hide all sync overhead except the tail-latency

IP Products



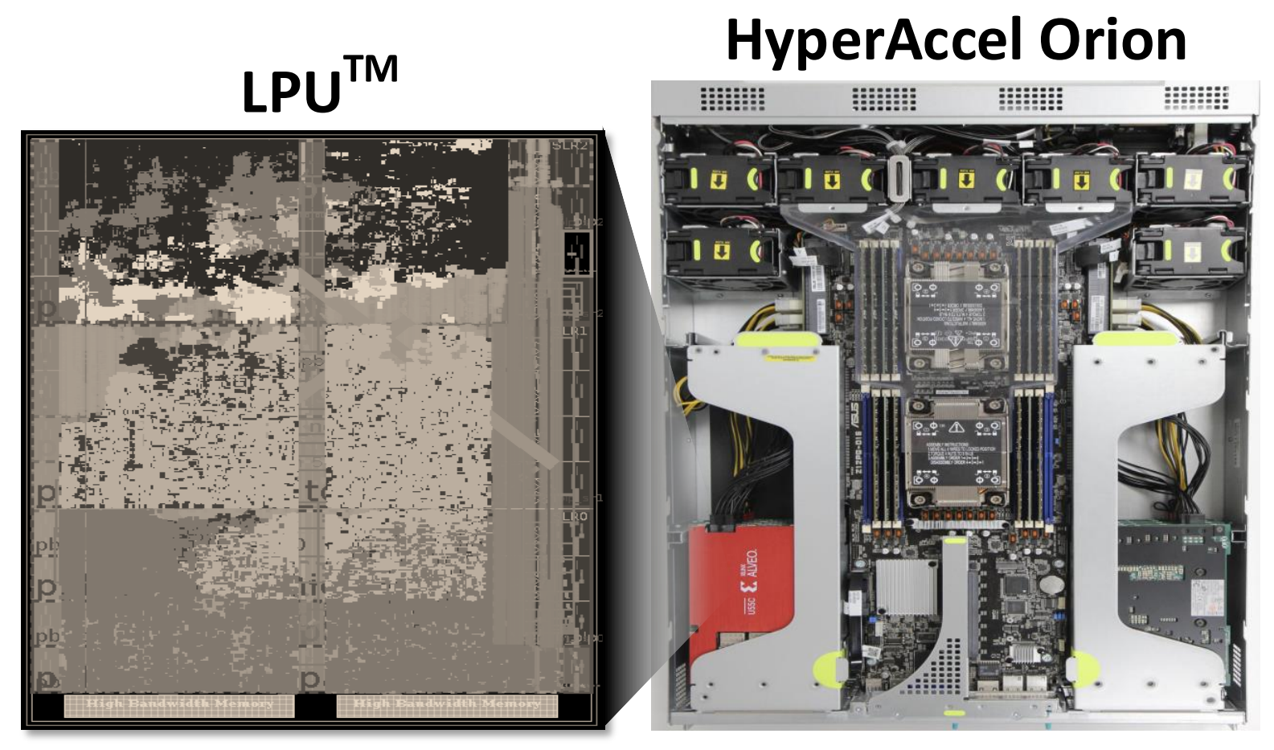
- Highly flexible to reconfigure both memory types and compute resources for low-power and high-performance (baseline: GDDR with 16 lanes x 64 vector dimension MAC trees in SXE)
- **Low-power:** scale down memory bandwidth to that of LPDDR with fewer MAC trees in SXE
- **High-performance:** scale up memory bandwidth to that of HBM with more MAC trees in SXE

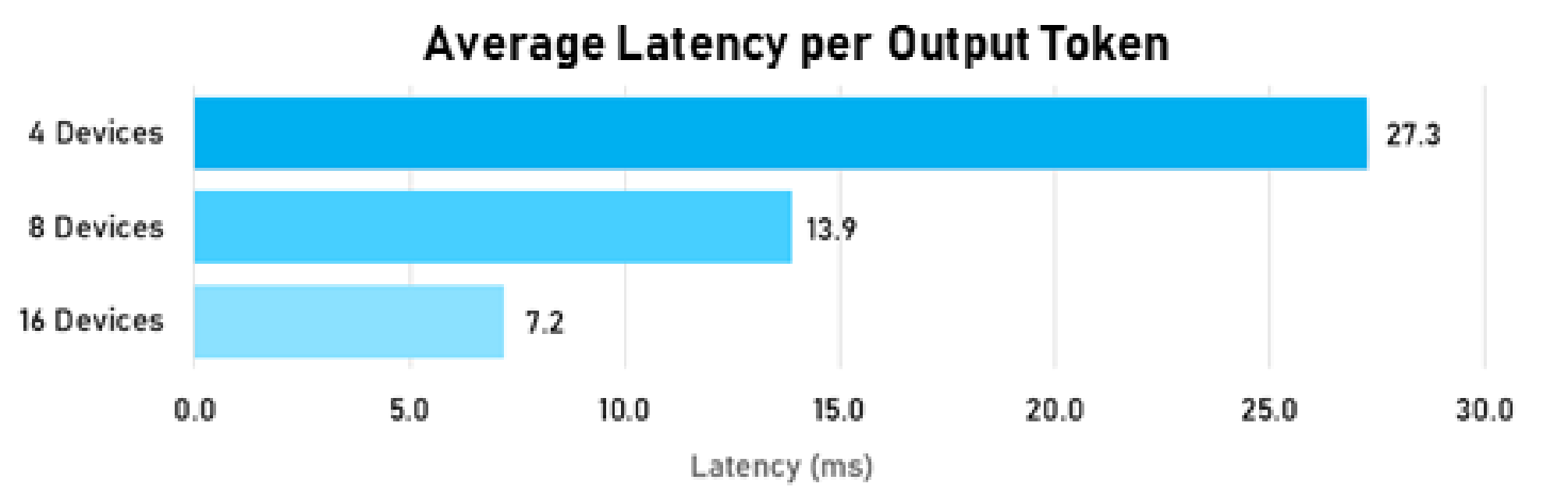
HyperDex Software Stack



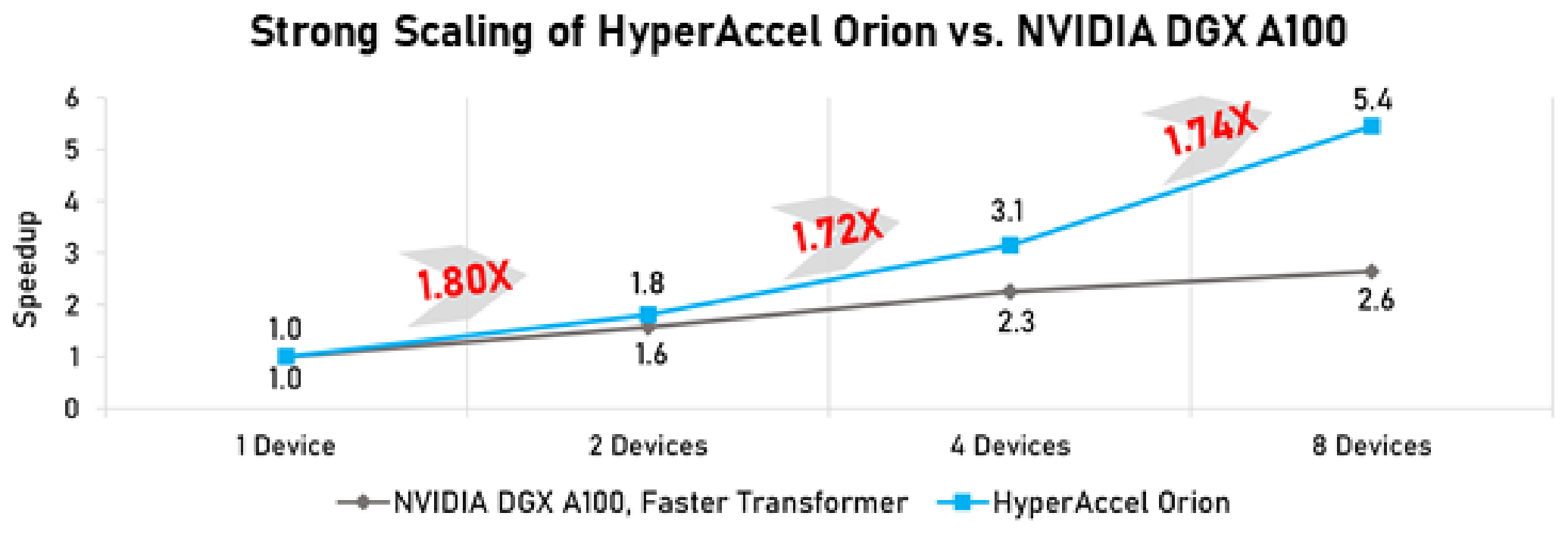
- Bridges LPU platform at the application-level through standard API
- Supports various LLMs, such as GPT, OPT, LLaMA, and their variants
- Intra-layer parallelism of model parameters for parallelizable operations
- Optimal memory allocation and alignment of model parameters
- Parallel instruction chaining for minimum control overhead

Performance Results



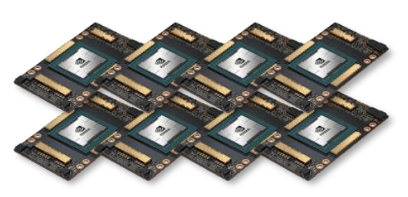


Millisecond (7.2ms) to generate an output token during GenAI inference



1.76x speedup on average for doubling the number of devices

SYSTEM SPECIFICATIONS		
	HyperAccel Orion	
Server Information		
Accelerator	8x Latency Processing Unit	16x Latency Processing Unit
HBM Total Bandwidth	3.68 TB/s	7.36 TB/s
HBM Memory Capacity	128 GB	256 GB
DSP Slice	72K	144K
Performance	13.9 ms per output token	7.2 ms per output token
System Power Usage	1.4 kW max	2.9 kW max
Form Factor	PCIe single slot	
Architecture	Streamlined Memory Access, Streamlined Execution Engine	
Network	Expandable Synchronization Link 2x QSFP28, 2x100Gb/s	
Software	HyperDex framework	
Generative AI Service		
Service	Transformer-based natural language generation	
Supported Model	GPT, OPT, LLaMA, and their variants	
Model Size	Up to 100 Billion parameters	
API	OpenAI-based	
Simultaneous Access	1-16 clients	

	NVIDIA DGX A100	HyperAccel Orion
Accelerator	 **8 x A100** 80 GB, 2,039 GB/s HBM 600 GB/s NVLink	 **16 x U55C** 16 GB, 460 GB/s HBM 100 Gbit/s QSFP28
Maximum Power	**3,200 W**	**2,400 W**
Cost	**$119,992** (1 GPU = $14,999)	**$75,952** (1 FPGA = $4,747)
Performance	**93.9 tokens/s** ×1.49	**139.8 tokens/s**
Performance /cost	**782.5 tokens/s/milion$** ×2.35	**1,840.6 tokens/s/milion$**

- **HyperAccel Orion** (16 LPUs), HyperDex vs. **NVIDIA DGX A100** (8 GPUs), FasterTransformer
- GPT3-20B, 32-input/128-output text generation
- Orion achieves **7.2ms per output token** (~140 tokens per second)
- Orion achieves **1.76x** scalability for doubling the number of devices (vs. 1.39x of DGX A100)
- Orion achieves **1.49x** speedup and **2.35x** cost-effectiveness compared to DGX A100

Feel free to contact us! **Email:** contact@hyperaccel.ai **Website:** hyperaccel.ai **Linkedin:** linkedin.com/company/hyperaccel