HyperAccel Orion
Server Information
Accelerator
8 x A100
80 GB, 2,039 GB/s HBM
600 GB/s NVLink
8 x U55C
16 GB, 460 GB/s HBM
100 GB/s 10G QSFP28
Network
Expandable Synchronization Link (ESL)
Architecture
Streamlined Memory Access, Streamlined Execution Engine
Network
Expandable Synchronization Link (ESL)
Software
HyperDex Framework
General AI Service
Support Model
GPT, OPT, LLaMA, and their variants
Model Size
Up to 100 Billion parameters
API
OpenAI-based
Simultaneous Access
1-16 clients
Performance Results
Average Latency per Output Token
Millisecond (1.7ms) to generate an output token during actual inference

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

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

LPU™ Architecture
• Specialized to stream data transferred at high bandwidth
• Main computing unit that largely consists of multiple MAC trees for vector-matrix multiplication
• Vector operations such as Softmax, normalization
• 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)
• 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 synchronization to continuously run in tandem to hide all sync overhead except the tail-latency

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