

ABOUT HYPERACCEL

Beyond Acceleration. Built for Inference.

HyperAccel is an AI semiconductor startup pioneering LPU (LLM Processing Unit) technology purpose-built for LLM inference.

 **\$45M**
Currently raising
Series B

 **90+**
Employees
More than 80% are engineers.

 **2023**
Founded
Gangnam, Seoul, Korea

KEY CUSTOMERS & PARTNERS

NAVER Cloud

 **LG Electronics**

SAMSUNG

 **SEMI FIVE**

ADVANTECH

Inventec

HPE

AWARDS & RECOGNITION

-  **IEEE/ACM DAC 2023** – Best Presentation Award
-  **IEEE VLSI Design 2024** – Best Exhibitor Award
-  **KPAS 2024** – Promising AI Semiconductor Startup (Korea)
-  **IEEE Micro 2025** – Best Paper Award
-  **ICTGC 2025** – AI Core Technology & Chip Award (Taiwan)
-  **ICT Merit Award 2026** – Order of Merit, Prime Minister (Korea)
-  **WIS 2026 Innovation Award** – Minister of Science and ICT (Korea)

WE'RE HIRING!

Apply via our website or scan the QR code



 hyperaccel.ai

   HyperAccel

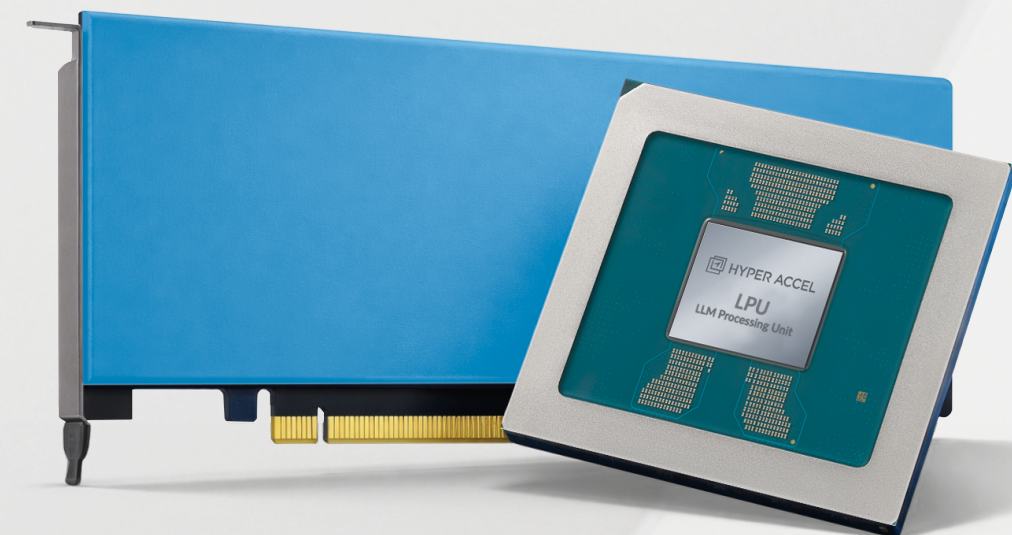
 contact@hyperaccel.ai



HYPER ACCEL

Hyper-Accelerated Solutions For Mission-Critical Workloads

Built for LLM Inference. More Tokens, Less Budget.



1/5

TCO (\$/hr) vs. H100

6x

Per-Watt & Per-Dollar Efficiency
up to 6.2x, Qwen3-235B FP8

5x

TPS Efficiency per TFLOPS
2.14 vs. 0.46 tok/s/TFLOPS (H100)

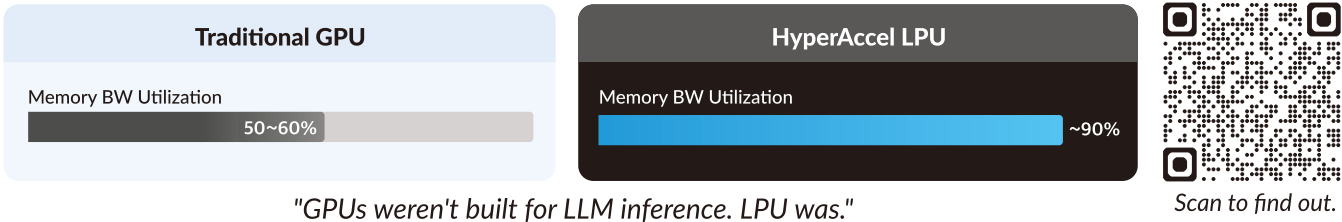
Powered by HyperAccel's LPU SoC, purpose-built for LLM inference at data center scale.

WHY HyperAccel LPU

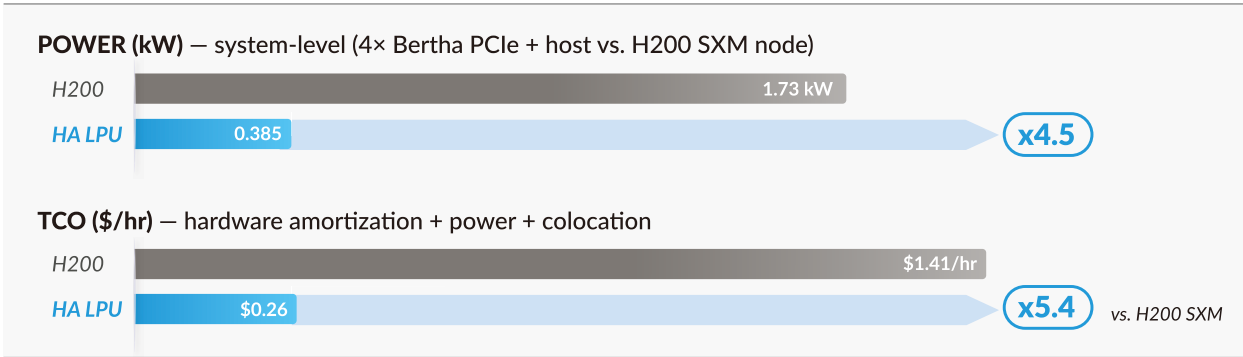
More Tokens. Less Power. Lower Cost.

World's first LLM inference silicon with LPDDR5X

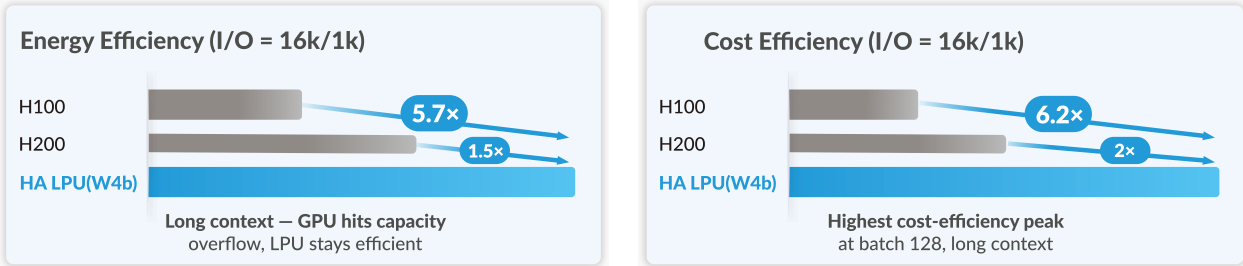
Architecture Advantages: Streamlined Dataflow



System-Level Cost & Power vs. H200 SXM



Benchmark: Up to 6× Per-Watt & Per-Dollar (Qwen3-235B, FP8, TP=4)



Full Spec Comparison

	H100 SXM	H200 SXM	HyperAccel LPU
Process Node	TSMC 4N	TSMC 4N	Samsung 4nm
Peak Compute (BF16)	989.5 TFLOPS	989.5 TFLOPS	384 TFLOPS
Peak Compute (INT8)	1,979 TOPS	1,979 TOPS	768 TOPS
Peak Compute (MXFP4)	N/A	N/A	1.5 PFLOPS
TDP	700W	700W	250W
On-Chip SRAM	83MB	83MB	256MB
Memory Capacity	HBM3 80GB	HBM3e 141GB	LPDDR5X 192GB
Memory BW	3.5 TB/s	4.8 TB/s	546 GB/s
Form Factor	SXM / PCIe card	SXM / PCIe card	Dual Slot PCIe card

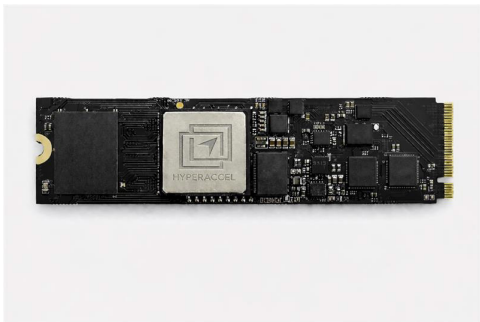
Full Product Portfolio

From data center scale to edge intelligence

LPU M.2 Modul: The Core of Edge AI

Coming Soon

The LPU M.2 is a next-generation AI accelerator designed to run LLMs and Transformer models directly on embedded devices. No server required—just pure, on-device intelligence.



for On-device AI

Key Specifications

- 16 TFLOPs (BF16) and 32 TOPs (INT8) compute performance
- 16GB LPDDR5X memory
- MXM form factor
- Optimized architecture for LLM and Transformer-based inference on edge devices

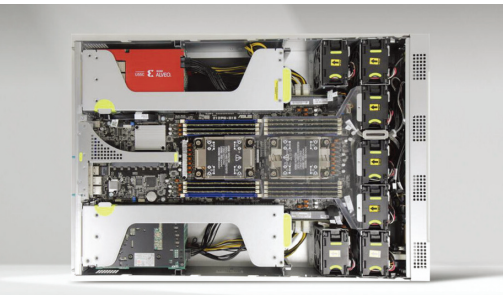
Applications

Enabling Physical AI across industries — from industrial automation robots and humanoid robots to drones, smart retail, and AI kiosks.

Orion Server (HX-F55X)

Available Now

Orion is powered by AMD Alveo™ U55C and AMD Virtex™ UltraScale+™ FPGAs, engineered to accelerate transformer-based large language models at data center scale.



for Data center-scale LLM serving

- Accelerator8 x Flux 55X
- DRAM Bandwidth3.68 TB/s
- DRAM Capacity128 GB
- TPS175 (Llama2 7b)
- TDP576W

Forte-55X Accelerator

Available Now

Built on AMD Xilinx U55C FPGA, F55X delivers low-latency, privacy-focused AI inference with 460 GB/s HBM bandwidth — optimized for single inference workloads.



for Ultra-low latency single inference

- DRAM BandwidthHBM2, 460 GB/s
- DRAM Capacity16 GB
- Power Consumption75W
- Batch Size1
- Form FactorSingle Slot
- Supported ModelsGPT, OPT, Llama, Claude, Phi