Bridging Performance and Affordability in AI Silicon

Juntaek Oh

HYPER ACCEL

Source: 조선일보


엔비디아 GPU 26만장의 의미 ※B200 칩 기준
가격
최대 104억달러
(약 15조원)
연산 능력
중대형 AI
데이터센터 5곳
HBM
HBM3E 208만개 탑재
(6000억~9000억원)
필요 전력
500~600MW
(중형 화력발전소급)
국가 경쟁력
미국, 중국에 이은
3위 수준
엔비디아가
국내에 공급하는
GPU
과학기술
정보통신부
5만장
국가 AI 컴퓨팅센터 구축·독자 AI 파운데이션 모델 개발
삼성전자
5만장
세계 최대 반도체 AI 팩토리 구축
SK 그룹
5만장
제조 AI 클라우드 및 AI 팩토리 구축
현대차 그룹
5만장
AI 팩토리 통해 자율차, 스마트 팩토리, 로보틱스 강화
네이버
6만장
피지컬 AI 플랫폼 공동 개발
자료=업계 그래픽=이진영

Source: 조선일보

Cassandra Unchained @michaeljburry · Nov 10

Understating depreciation by extending useful life of assets artificially boosts earnings -one of the more common frauds of the modern era.

Massively ramping capex through purchase of Nvidia chips/servers on a 2-3 yr product cycle should not result in the extension of useful lives of compute equipment.

Yet this is exactly what all the hyperscalers have done. By my estimates they will understate depreciation by $176 billion 2026-2028.

By 2028, ORCL will overstate earnings 26.9%, META by 20.8%, etc. But it gets worse. More detail coming November 25th. Stay tuned.

Network/Compute Depreciation Useful Life (Years)

Company	2020	2021	2022	2023	2024	2025
META	3	4	4½	4½	4½	5½
GOOG	3	4	4	6	6	6
ORCL	5	5	5	5	6	6
MSFT	3	4	6	6	6	6
AMZN	4	4	5	5	6	5

Source: Company SEC Filings

한국경제 구독 PICK

"265조 사기극"...공매도 나선 '빅쇼트' 마이클 버리의 경고 [김인엽의 퓨처 디스패치]

김인엽 기자 TALK

입력 2025.11.11. 오전 10:26 · 수정 2025.11.11. 오전 11:02 기사원문

57 199 가가

버리 "AI데이터센터 감가상각 축소"
알파벳·MS 내용연수 3→6년 연장
이코노미스트 "시총 1100조 증발 우려"
엔비디아 GPU 개발 주기는 2→1년으로
오픈AI "주기 짧아지면 자금조달 어려워"
"AI 학습 아닌 추론 등 사용 가능" 주장도

사진=연합뉴스

Source: X, 한국경제

AMD CEO Lisa Su says AI data center market will be worth $1 trillion by 2030

AMD
팔로워 2,003,747명
5시간 •

At AMD Financial Analyst Day, our CEO Lisa Su projected that the data center market will reach $1 trillion by 2030, and AMD is positioned to lead that transformation.

“We have now all of the pieces to deliver full AI factories, and that is really our goal throughout this entire stack, across CPUs, GPUs, software, networking, and our cluster-level systems design.”

With leadership across AI infrastructure and multi-gigawatt deployments, AMD is powering the next era of high-performance computing.

Read more from Yahoo Finance

번역 표시

AMD CEO Lisa Su says AI data center market will be worth $1 trillion by 2030
finance.yahoo.com

Meta plans $600 billion US spend as AI data centers expand

By **Reuters**

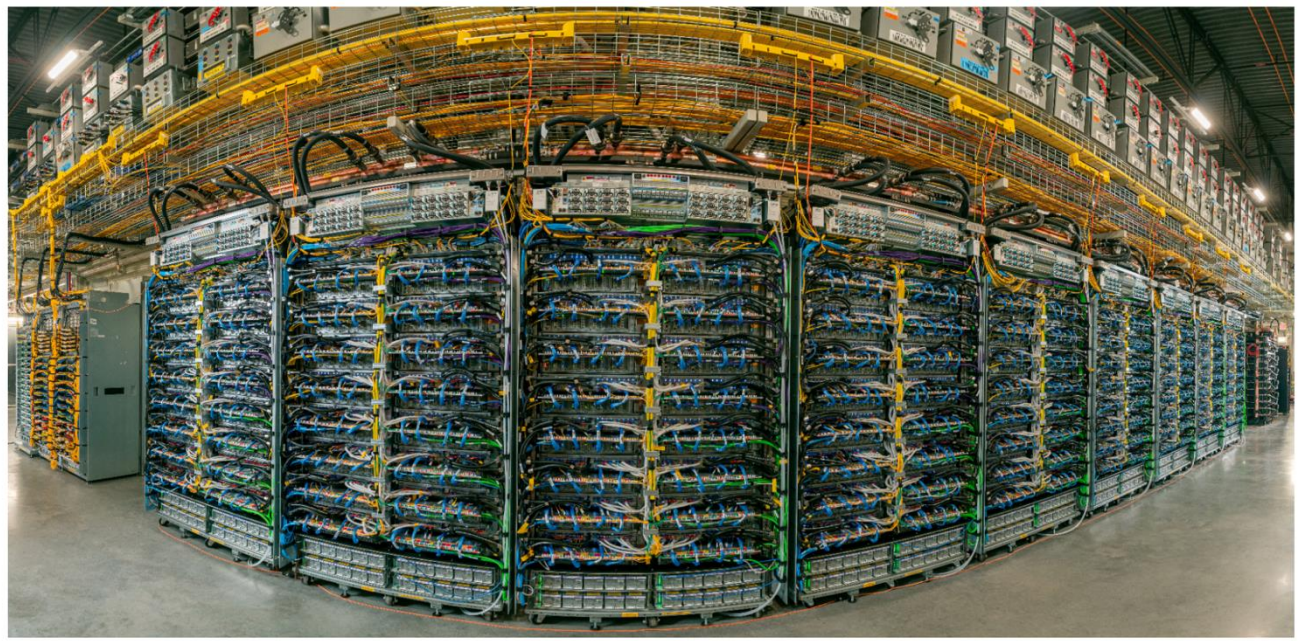
November 7, 2025 10:23 AM PST · Updated November 7, 2025

A teenager poses for a photo while holding a smartphone in front of a Meta logo in this illustration taken September 11, 2025.
REUTERS/Dado Ruvic/Illustration/File Photo Purchase Licensing Rights

Source: Linkedin, Reuters

Google Cloud Launches Ironwood TPUs, New Axion VMs for AI Inference

November 10, 2025
BY QUANTUM NEWS



Source: Quantum zeitgeist


cerebras raises
$1.1B
$8.1B VALUATION
Series G Funding
LED BY
Fidelity INVESTMENTS
ATREIDES MANAGEMENT
WITH PARTICIPATION FROM
TIGERGLOBAL
VALOR
1789 CAPITAL
ALTIMETER
ALPHA WAVE
BENCHMARK

Groq Raises $750 Million as Inference Demand Surges

The investment strengthens Groq's role in the American AI Stack, delivering fast, affordable compute worldwide.

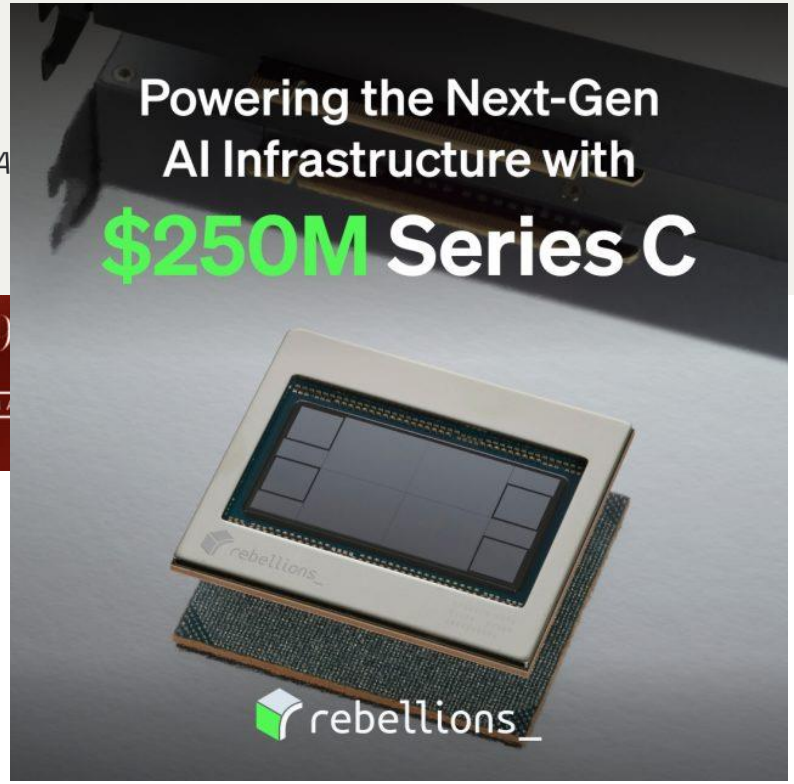
cerebras raises

Groq Raises $750 Million as Inference Demand Surges

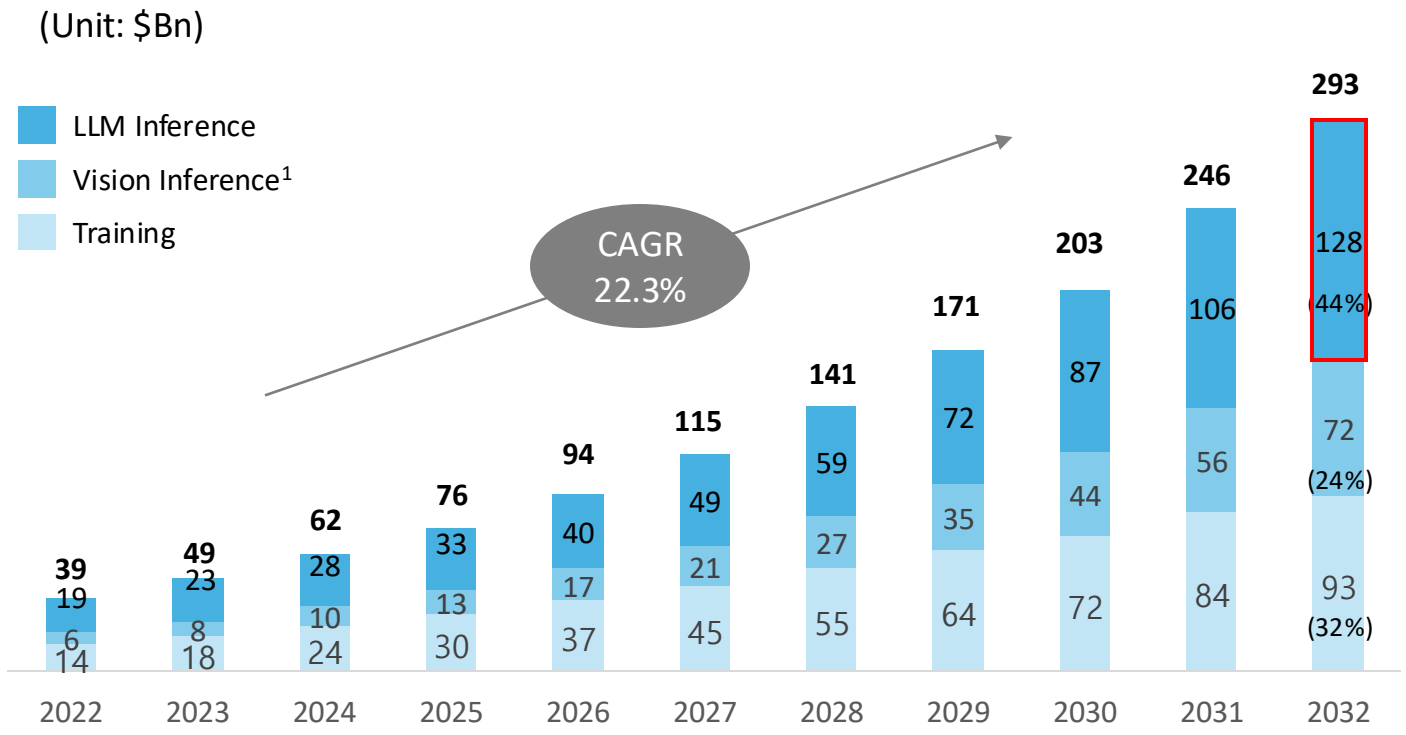
The investment strengthens Groq's role in the A
compute worldwide.

LED BY

Fidelity INVESTMENTS · ATREIDES MANAGEMENT · TIGERGLOBAL · VALOR · 1789 · ALTIMETER · ALPHA WAVE · BENCHMA



AI Chips for LLM Inference are Poised to Lead the Market



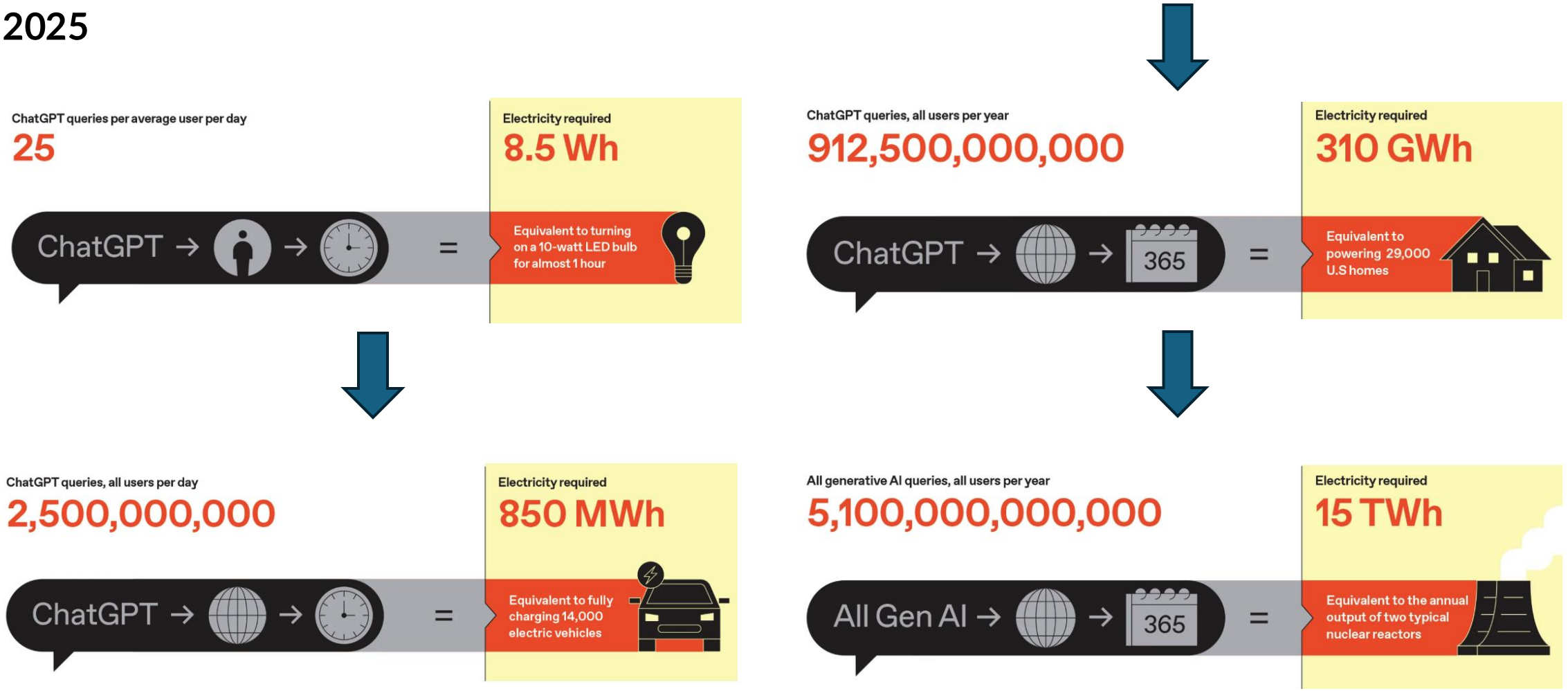
LLM Inference Chip to Comprise 44% of the Total Market by 2032

- Increasing demand for NLP applications across industries
- Growing complexity of language models
- Performance and efficiency benefits

Source: IDTechEx, Bloomberg, HyperAccel Analysis

[1] CNN Inference

2025

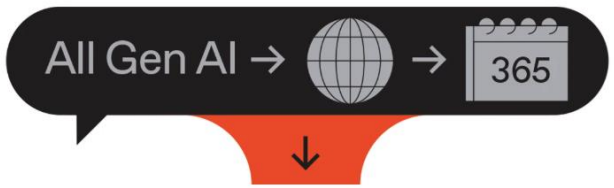


Source: IEEE Spectrum

2030

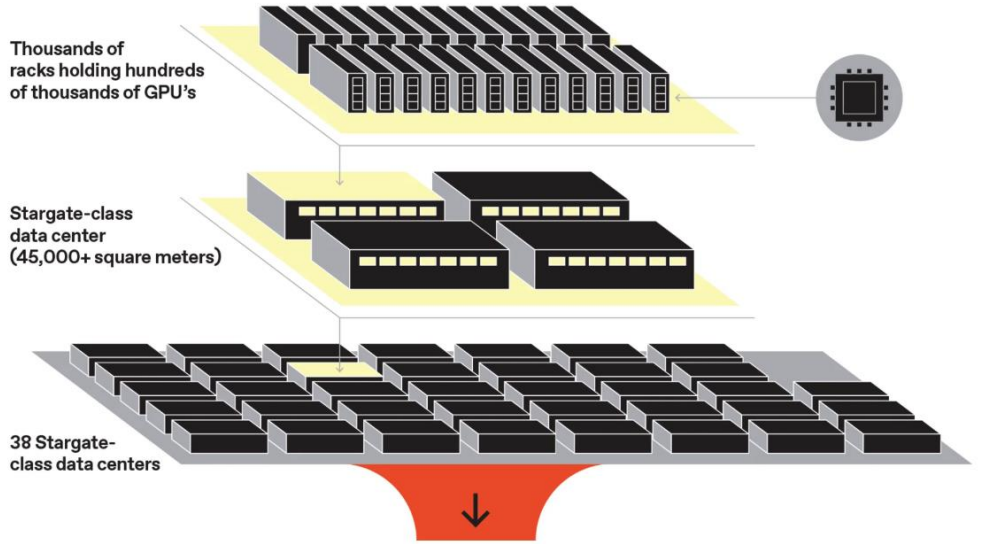
All generative AI queries, all users per year

120,000,000,000,000



Construction required

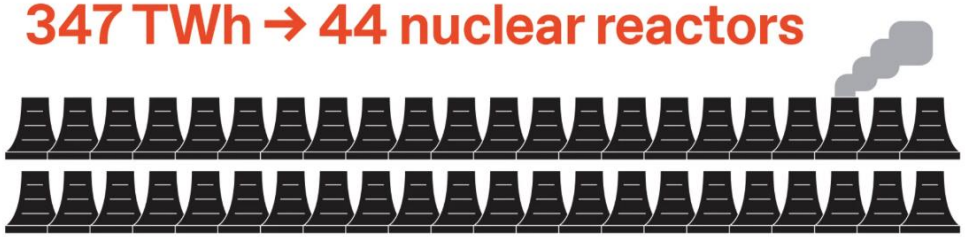
38 Stargate-class data centers



The Schneider Electric report estimates that all generative AI queries consume 15 TWh in 2025 and will use 347 TWh by 2030; that leaves 332 TWh of energy—and compute power—that will need to come online to support AI growth. That implies the construction of dozens of data centers along the lines of the Stargate Project, which plans to build the first ever 1-gigawatt facilities. Each of these facilities will theoretically consume 8.76 TWh per year—so 38 of these new campuses will account for the 332 TWh of new energy required.

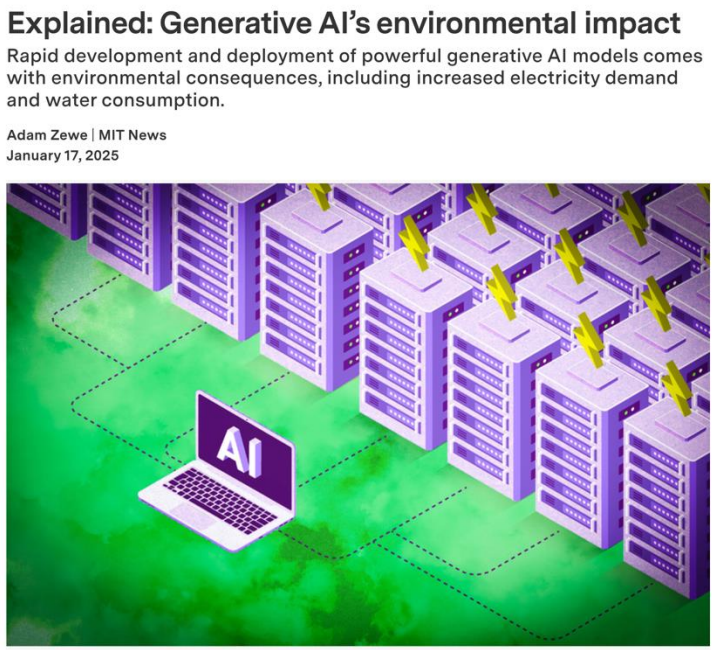
Electricity required

347 TWh → 44 nuclear reactors



Source: IEEE Spectrum

Why Microsoft's AI Chip Supply Chain Awaits Power



100MW data center capacity may remain unused for years due to power shortage

Source: MIT News, TECHZINE, SupplyChain

HyperAccel Steps Forward With Unique GenAI Chip Solutions to Win the AI Opportunity

Fast, Efficient, and Affordable for Generative AI Inference

HyperAccel is an ***GenAI Chip Startup*** hyperfocused on developing disruptive technologies and solutions specifically for ***Generative AI Inferencing***. Designed from foundational understanding of transformer architecture Large Language Models, our **LPU(LLM Processing Unit)**aims for higher throughput performance with order of magnitude gains in cost and energy efficiency as compared to GPU's in market today. We dedicate ourselves to help our customers scale ***Fast, Efficient, and Affordable GenAI Services for Everyone!***

HyperAccel Vision : GenAI is for Everyone

Engineered for GenAI Leadership

Specialized in LLM

We offer unique GenAI chips designed from first principles for LLM Inference – the HyperAccel LPU

Our ASIC models achieve **2×** higher throughput, **20×** better cost efficiency, and **5×** better energy efficiency compared to NVIDIA H100[1]

Deep-Tech Founders & World-Class Team

Launched in January 2023, HyperAccel has grown to Series A and curated a team of over 70 world-class semiconductor HW, SW, and Systems professionals from leading companies

Disruptive Technologies

Patented Hardware Designs

Proprietary Full-Stack Software Platform

Distributed/Runtime-Dynamic Execution

Scalable/Extensible Architecture

Lighthouse Customers

NAVER Cloud for Datacenter AI ASIC

Global Consumer Electronics Company (TBA) for Edge AI ASIC

[1] Llama 3.1 70B; 4096 Input/Output Token Length; Batch 32

Company Profile

Jan 2023

KAIST

Spin-off

Aug 2023

$4.5M

(Seed)

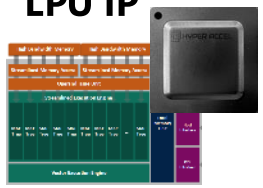
Oct 2023

AMD

Orion AI Server (FPGA)

Sep 2024

LPU IP



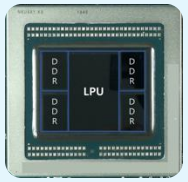
Dec 2024

$40M

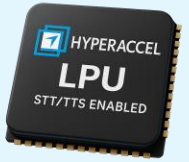
(Series A)

2026

Two ASIC Products



NAVER Cloud



70
Employees

50
Patents

$3.5M
Revenue

The Leadership Team

AI Computing & Infra
18 yrs Semiconductor Expert, 9 yrs @ Microsoft

Founder & CEO

Professor EE, KAIST
Head of AI Semiconductor Systems Research Lab

Engineering Leader, Microsoft Azure
2017 - 2019

Senior Researcher, Microsoft Research
2014 - 2017

Researcher, Microsoft Research
2012 - 2014

Joo-Young Kim

Microsoft

Education

Ph.D., KAIST (2010)
M.S., KAIST (2007)
B.S., KAIST (2005)

World-Class Engineering Team

- HW Engineering (24)
- SW Engineering (25)
- Systems Engineering (7)

NPU Computing
13 yrs Semiconductor Expert, 10 yrs @ Samsung

CTO

Neubla CTO
2021 - 2023

Samsung C-Lab Leader
2021 - 2021

Staff Engineer, Samsung S. LSI
2017 - 2020

SoC Engineer, Samsung S. LSI
2011 - 2015

Jinwon Lee

neubla

SAMSUNG

Education

M.S., SNU (2017)
B.S., SNU (2002)

Strategy & Business Development
20 yrs Management Expert, 13 yrs Tier-1 Consulting

CSO

OKIT Inc. Founder & CEO
2020 - 2024

Loplat COO
2017 - 2020

BCG Principal
2012 - 2017

Oliver Wyman Manager
2005 - 2011

Yongwoong Jung

BCG

Education

INSEAD MBA (Class of 2012)
B.S., Korea University (2004)

Business & Operations Team

- BD, Sales, Mktg, Finance, HR, Admin (8)

We Saw the Need to Specialize Having Deep Knowledge of LLM's the Effects of Large Models, Sequential Processing, and Memory Intensive for GenAI Inference

Input

Hello, my name

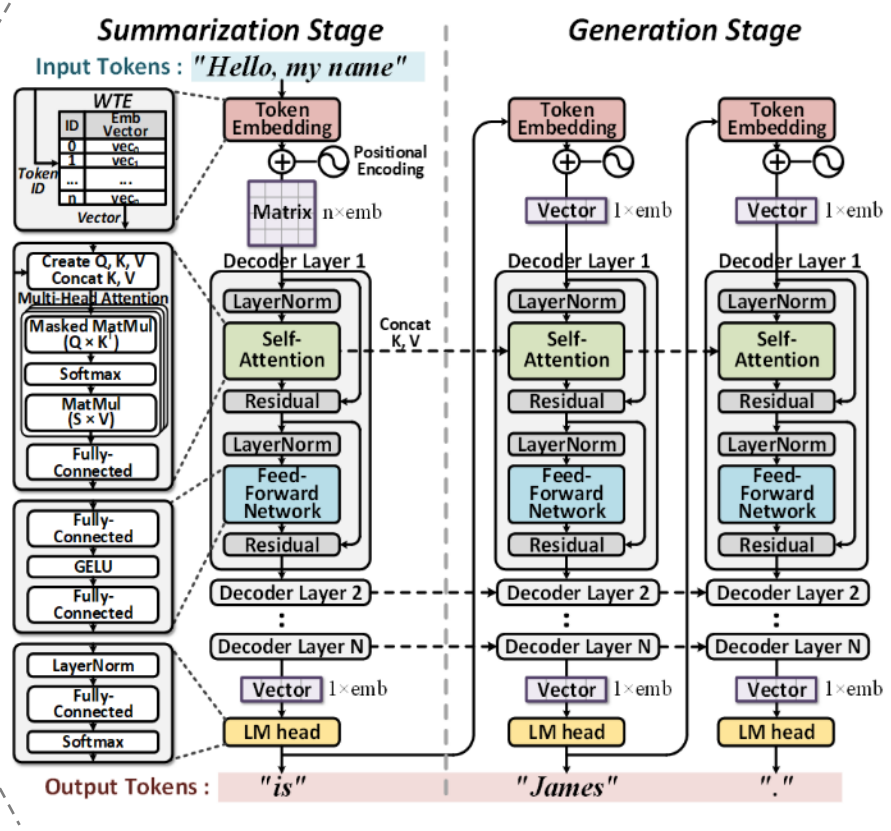
Large Language Model

ChatGPT LLAMA AI Gemini

Output

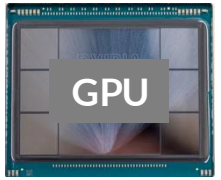
is James Smith and…

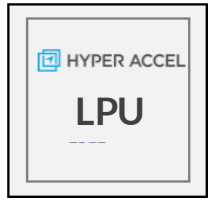
Transformer Architecture



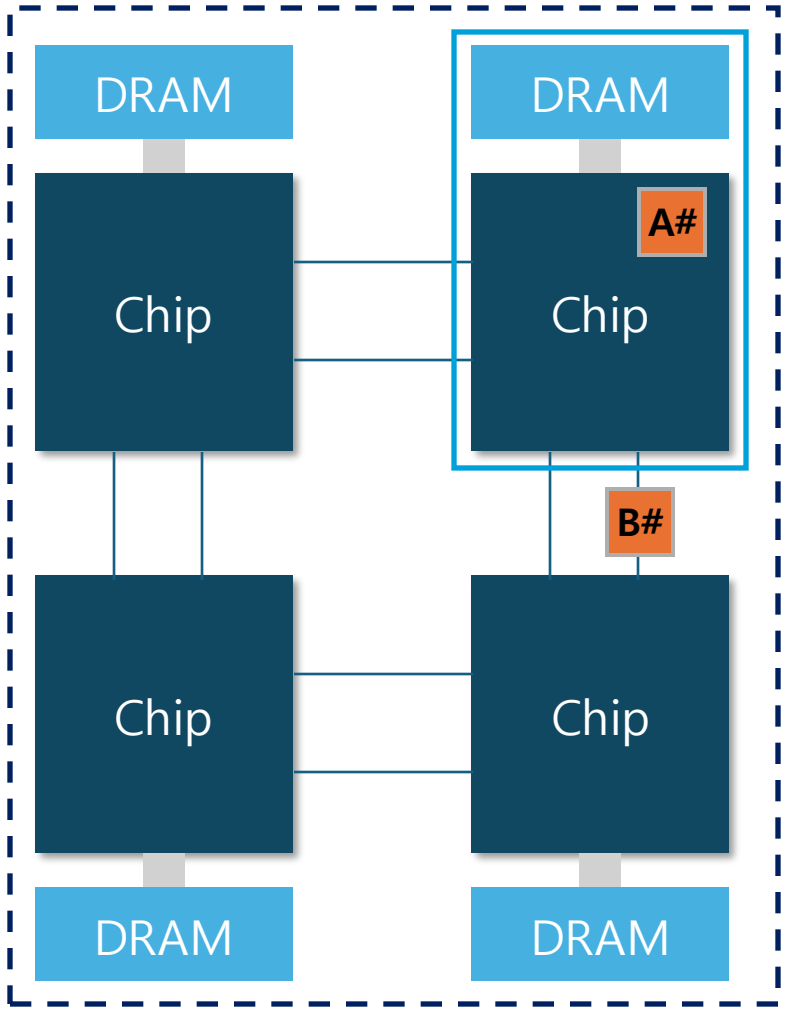
Large Models + Sequential Processing

Training (Model Making)	VS	Inference (Service)
Large Batch		Small Batch
Compute Intensive		***Memory Intensive***
Throughput Oriented		Service/Latency Oriented





We Built Our LPU with Unique HW, SW, and System Designs to Deliver on Our Promise of High Performance and Affordability/Sustainability



Single LPU

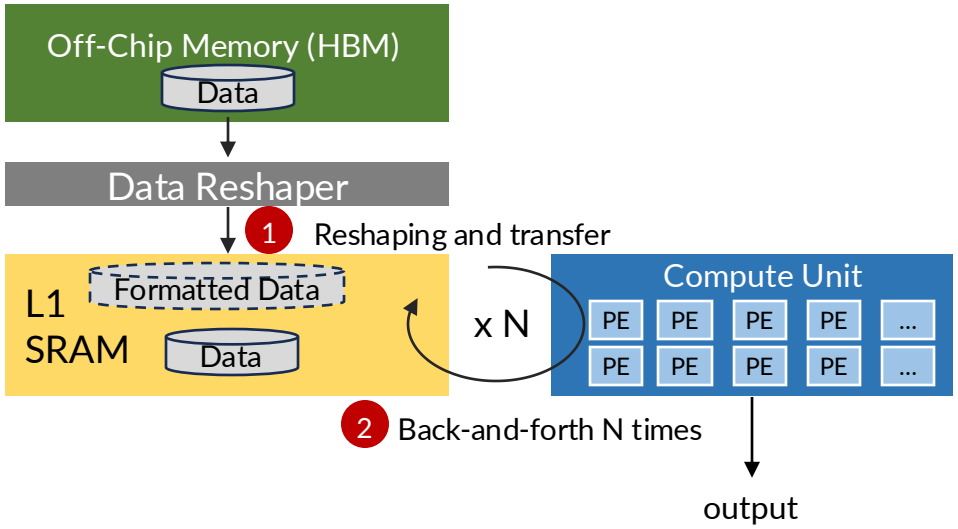
Multi LPU

	High Performance	Affordability/Sustainability
A1	**Maximized Memory Bandwidth Utilization** - Streamlined Memory Access	**Standard LPDDR Memory Replaces HBM** - Widely Available and Low Cost - 2x the Max Capacity of HBM - ½ the Price of HBM - 60% less power consumption - No HBM Advanced Packaging Costs - Comparatively Lower IP Costs
A2	**Specialized Compute Engines for End-to-End LLM Operations** - Matrix and Vector Execution Engines	
A3	**Adaptable HW Architecture Design for Chip-Level Scale Up/Down**	
B1	**System-Level Extensibility via Peer-to-Peer Expandable Synchronization Link (ESL)**	**Standard PCIe and Ethernet for NVLink-type Capabilities without Proprietary Charges and High Cost**
B2	**Full-Stack HyperDex Software Platform to Optimize the Performance of Multi-LPU System**	

Streamlined Dataflow Architecture

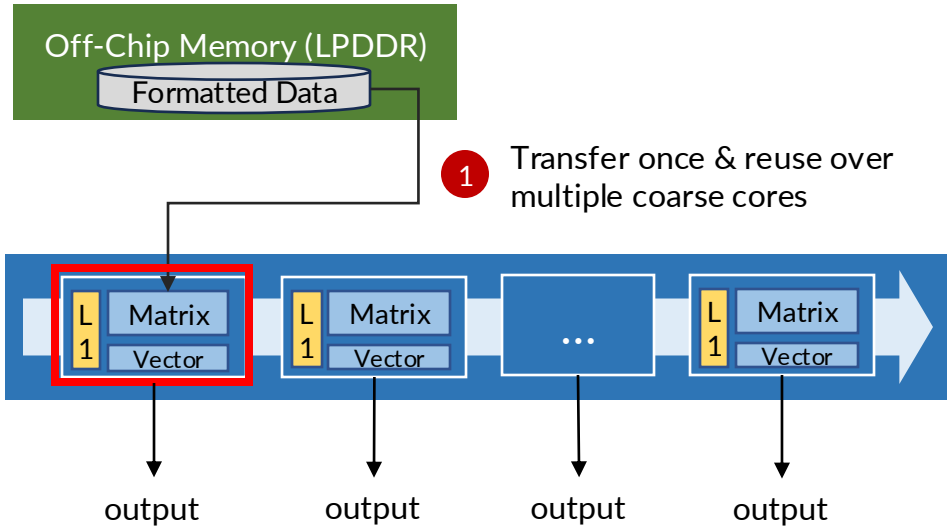
Conventional (GPU)

- 1000s of Small Cores
- Hierarchical access w/ lots of data movements



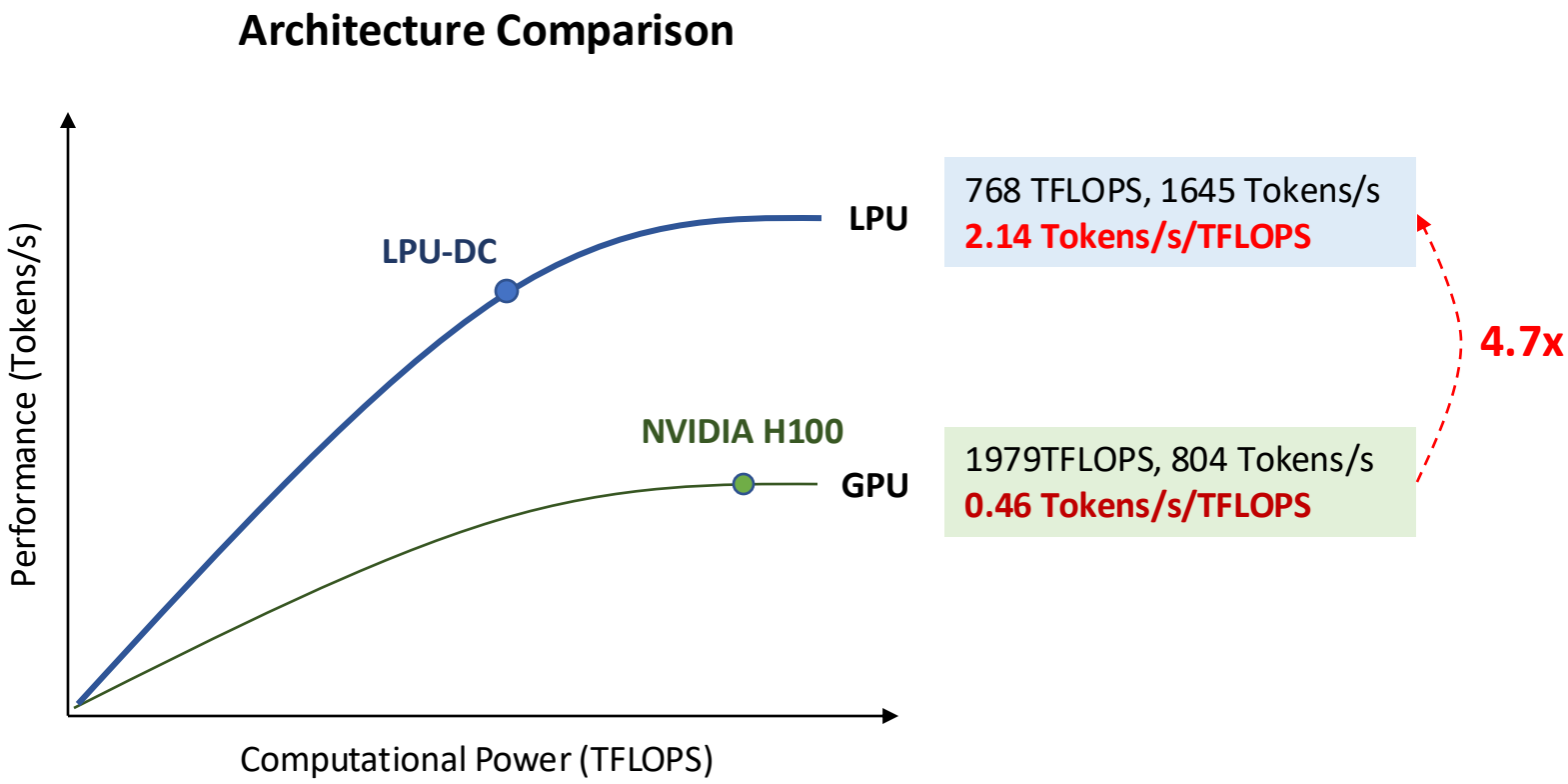
Ours (LPU)

- 10s of Big Cores
- Streamlined dataflow w/ maximum data reuse



Maintain ~90% Utilization of Given Peak BW

Achieving 5x LLM Performance per TFLOPS

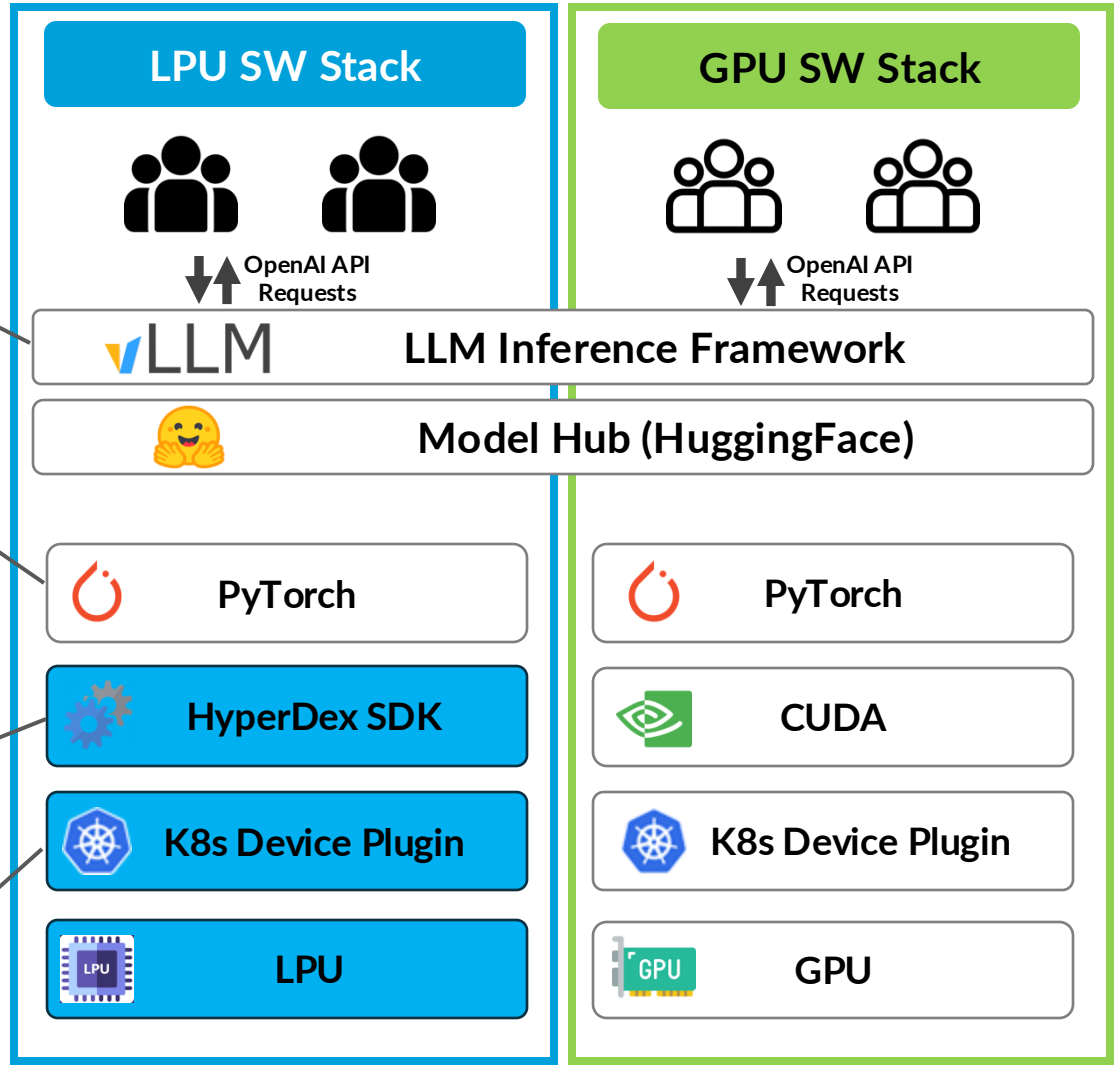


Prioritizing Software, We Built From the Start Our Full-Stack HyperDex Software Platform to Optimize HW Performance and Engage GenAI Developers

vLLM Plugin: Enable the use of the LLM inference framework vLLM on LPU

eDSL: Domain-Specific Language for writing high-performance, efficient LPU kernels

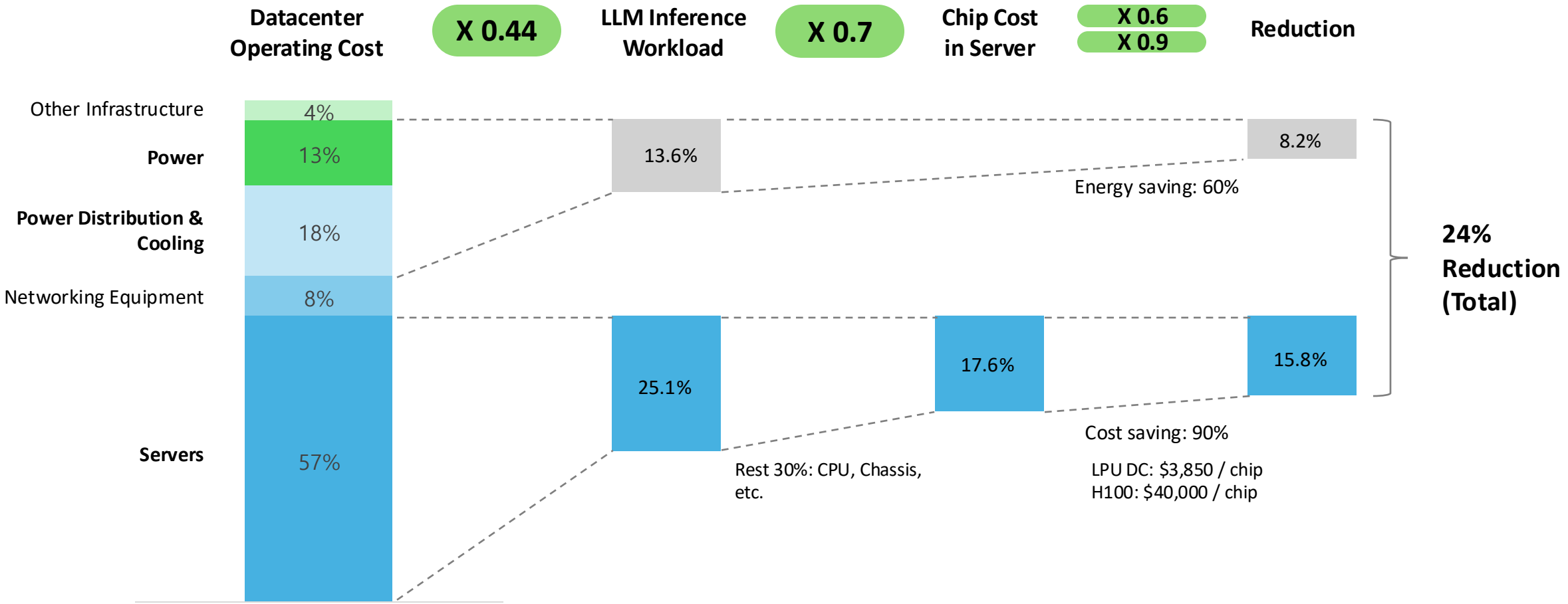
Compile & run the binary on LPU with tensor parallelism for running large models
Device Plugin for LPU deployment in a cloud environment



Seamless LLM inference experience for developers familiar with GPUs

- Use same OpenAI API requests to GPU setup
- Run vLLM OpenAI-compatible server
- Get LLM models from the HuggingFace Hub

Datacenter Cost Reduction



Source: AWS, HyperAccel Analysis

HyperAccel Product Roadmap

4Q 2023

On-Premise

3Q 2025

Cloud

Q1 2026: MXFP4 Quantization (4x Performance Boost)
Q2 2026: Compatibility Extension (Leading FPGA's)

LPU FPGA Server

Orion™

Latency Optimized

Orion™ Gen2

+ Quantization
+ Compatibility Extension FPGA's

Amazon EC2 F2 Instances

LPU ASIC-DC Chip

Customer Inputs

Silicon Prototype (Samsung 4nm MPW)

+ Latency/Throughput
+ Batching Support

4Q 2024

Production ASIC (Samsung 4nm Single Run)

+ Multi-Modality, Input/Output Batching
+ Quantization, Full SW Stack

Datacenter

HYPER ACCEL
Bertha
SZVB3I
2512

1Q 2026

LPU ASIC-Edge Chip

On-Device

Production ASIC (TSMC 6nm Single Run)

+ Multi-Modality (STT/TTS)
+ Power Optimization

4Q 2026

2023 2024 2025 2026 2027

HyperAccel Product Portfolio

Product	**LPU ASIC-DC Chip** Available 1Q 2026	**Orion™ LPU FPGA Server** Commercially Available	**LPU ASIC-Edge Chip** Available 4Q 2026
Target Customer	Hyperscalers Cloud Service Providers AI Cloud Providers Datacenters	Telecom Research Labs Universities Startups	Automotive Consumer Electronics/Smart Home AI Devices Robotics IoT
Value Proposition	High Performance Cost & Energy Efficient Attractive Pricing	Affordable Performance Cost & Energy Efficient Realtime Throughput	Edge Performance Cost & Energy Efficient Mass-Market Pricing
Support & Maintenance	HW: Server Partner ASIC: HyperAccel SW: HyperAccel	HW (incl. FPGA): Server Partner FPGA (Image): HyperAccel SW: HyperAccel	HW: OEM Partner ASIC: HyperAccel SW: HyperAccel

HYPER ACCEL

Hyper-Accelerated Solutions For Mission-Critical Workloads