



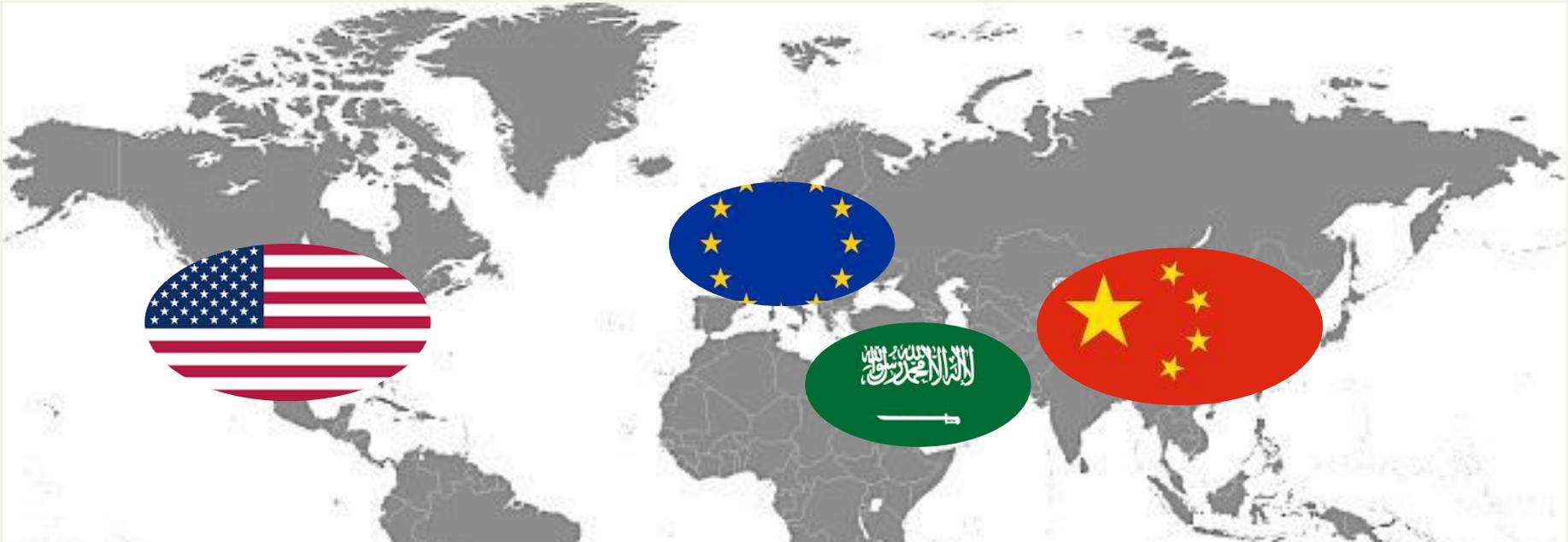
# Delivering Sovereign AI



## Chris Stephens

VP, CTO @ Groq

- + 3x Chief Data Officer
- + Faculty @ Carnegie Mellon
- + Father of 5



AI is Sovereign & Global

# Everyone is Making a Plan



"AI will improve our healthcare, spur our research and innovation and boost our competitiveness. We want AI to be a force for good and for growth. We are doing this through our own European approach – based on openness, cooperation and excellent talent. But our approach still needs to be supercharged. This is why, together with our Member States and with our partners, we will mobilise unprecedented capital through InvestAI for European AI gigafactories. This unique public-private partnership, akin to a CERN for AI, will enable all our scientists and companies – not just the biggest - to develop the most advanced very large models needed to make Europe an AI continent."

Ursula von der Leyen, President of the European Commission

索引号: 000014349/2017-00142 主题分类: 科技、教育科技  
发文机关: 国务院 成文日期: 2017年07月08日  
标题: 国务院关于印发新一代人工智能发展规划的通知  
发文字号: 国发〔2017〕35号 发布日期: 2017年07月20日

## 国务院关于印发 新一代人工智能发展规划的通知 国发〔2017〕35号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：  
现将《新一代人工智能发展规划》印发给你们，请认真贯彻落实。

(此件公开发布)

新一代人工智能发展规划

### Saudi Arabia

The Regional Hub for AI and Technology Investments + **\$14.9B**

<b>\$1.5B</b>	
Expanding Groq's investments in its project to launch the world's largest AI inference node in KSA through LPUs technology	Investing in digital infrastructure for AI by launching a global hub in Saudi Arabia to serve regional and global demand
<b>\$2B</b>	
Establishing an advanced manufacturing and technology center based on AI and robotics, and establishing a regional headquarters for Lenovo in Riyadh	Introduce ALLAM AI PC and provide a model of ALLAM on Qualcomm AI cloud

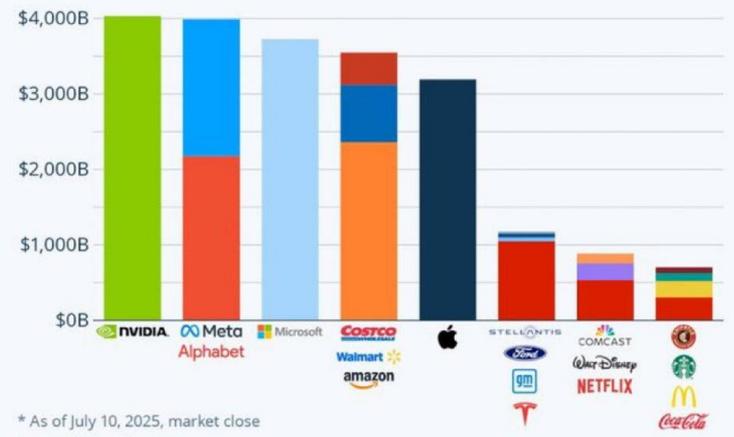
Why?



**sunny madra** @sundeeep · 1d  
Converting electricity into intelligence is valuable.

### \$4 Trillion: Nvidia's Record Valuation in Context

Nvidia's market capitalization in comparison to the (combined) market cap of other companies\*



\* As of July 10, 2025, market close  
Source: Yahoo Finance



8    14    78    11K

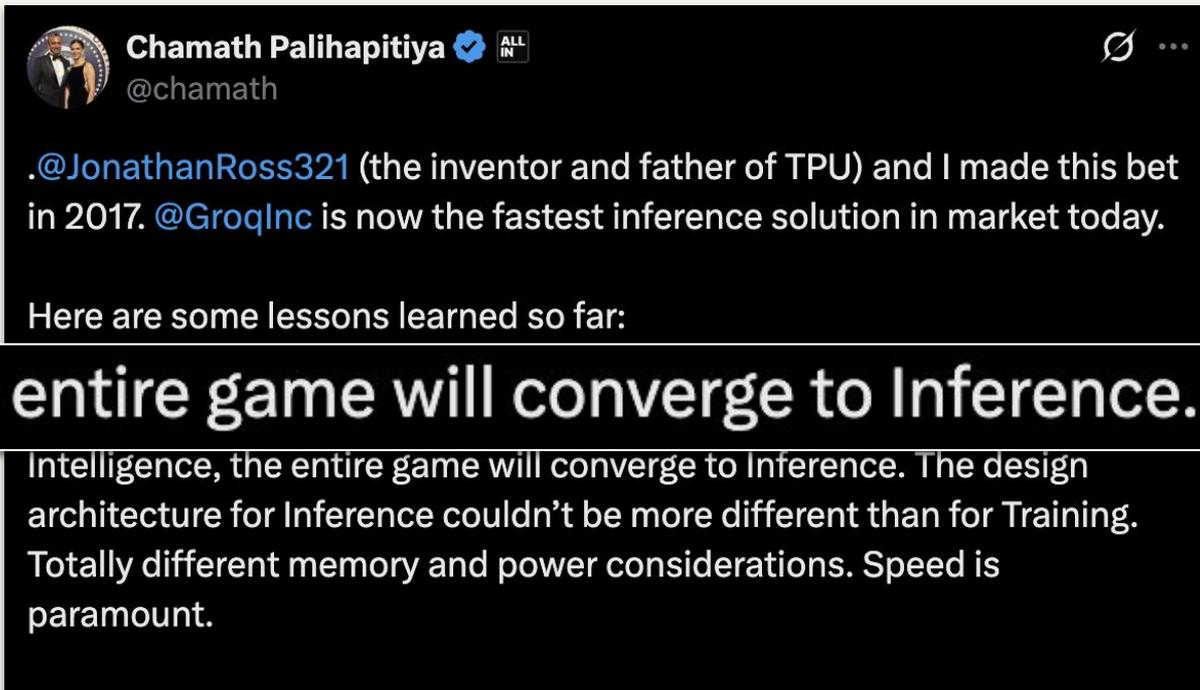
# Global AI Trends

---

Why?

- 1 Data & Compute Residency
- 2 AI security standards
- 3 Custom model ecosystems
- 4 Inference speed & cost explosion

Why?



A screenshot of a tweet from Chamath Palihapitiya (@chamath). The tweet text reads: ".@JonathanRoss321 (the inventor and father of TPU) and I made this bet in 2017. @GroqInc is now the fastest inference solution in market today. Here are some lessons learned so far: **the entire game will converge to Inference.** Intelligence, the entire game will converge to Inference. The design architecture for Inference couldn't be more different than for Training. Totally different memory and power considerations. Speed is paramount."

**the entire game will converge to Inference.**

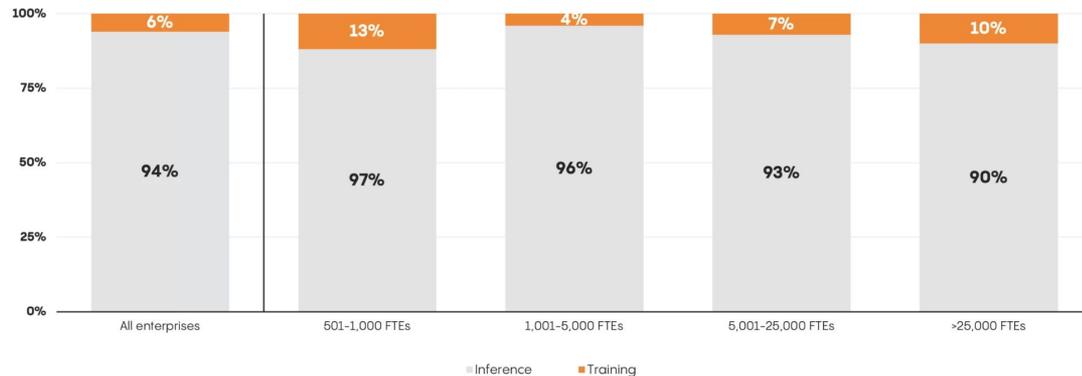
Intelligence, the entire game will converge to Inference. The design architecture for Inference couldn't be more different than for Training. Totally different memory and power considerations. Speed is paramount.

# Work is Inference

The first phase of AI is training where a model “learns”

The power comes from **inference**, where the learning is used to **solve problems**

Estimated Spend by AI Adopters on Training vs. Inference



© 2024 Menlo Ventures

<https://www.accenture.com/content/dam/accenture/financial/industry/high-tech/document/Accenture-Unlocking-Full-Potential-of-AI.pdf>  
<https://menlovc.com/perspective/the-modern-ai-stack-design-principles-for-the-future-of-enterprise-ai-architectures/>



THE PROMISE OF AI IS

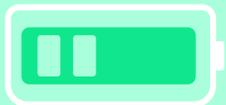
# Massive and Unrealized



1950s–2000s

## Algorithms

Enabled first mechanical automation of reasoning



2010s–PRESENT

## Training

Created today's foundation models, but required massive compute and power



TODAY

## Inference

Accessible and within reach for every business with predictable costs and clear ROI.



FUTURE

## AI-Native World

Instant, affordable, intelligent AI woven into daily life

Only the beginning

# This will be huge.

Public + private demand is persistent, concurrent, and growing

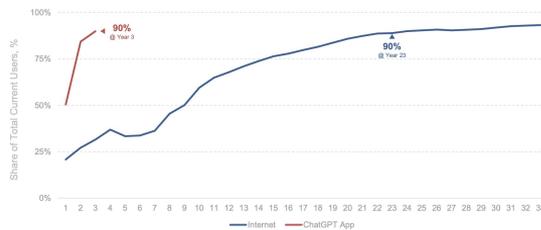
AI User Growth (ChatGPT as Foundational Indicator) = +8x to 800MM in Seventeen Months

ChatGPT User Growth (MM) – 10/22-4/25, per OpenAI



AI Global Adoption (ChatGPT as Foundational Indicator) = Have Not Seen Likes of This Around-the-World Spread Before

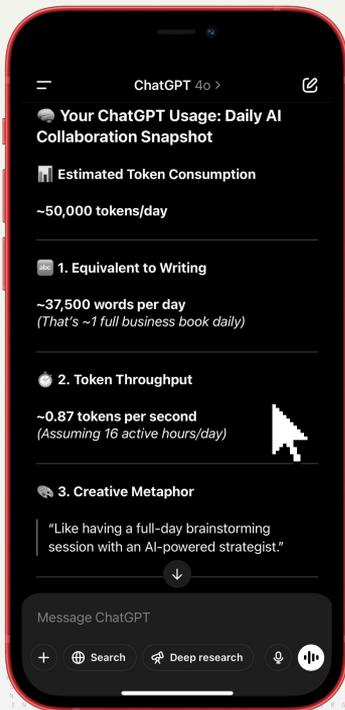
Internet vs. ChatGPT Users – Percent Outside North America (1990-2025), Per ITU & Sensor Tower



Note: Year 1 for Internet = 1990; year 33 = 2023. Year 1 for ChatGPT app = 5/23; year 3 for ChatGPT app = 5/25. ChatGPT app monthly active users (MAU) shown. Note that ChatGPT is not available in China, Russia and several other countries as of 5/25. China data may be subject to governmental limitations due to government restrictions. Includes only Android, iPhone & iPad users. Figures may underestimate from ChatGPT user base (e.g., desktop or mobile webpage users). Figures per United Nations definitions. Figures show % of total current users in that year – note that by year 3 for ChatGPT, we had per finished percentages could mean in coming months. Data for adoptive ChatGPT app only. Country-level data may be missing for select years, see per ITU. Source: United Nations / International Telecommunications Union (ITU); Sensor Tower (2023)



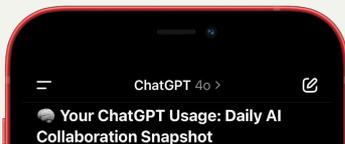
We all need  
Inference



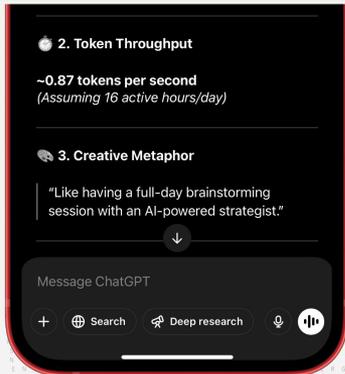
...most of the work will  
be inference, and

how much inference  
will be you need?

---



**~0.87 tokens per second**



If your country ran AI like electricity

How much inference compute would you need per citizen?

---



**~3T** tokens  
per day

---

50M People

**THIS IS A NATION-SCALE INFRASTRUCTURE NEED**

SENT INSIGHTS HIGH-SPEED LOW-POWER AUTONOMOUS PREDICTIVE ANALYT  
DATA-DRIVEN DECISION SUPPORT SYSTEM ARCHITECTURE DESIGN ENGINEER  
INNOVATION TECHNOLOGY ADVANCEMENT BREAKTHROUGH INNOVATION DISRUPT  
TRANSFORMATION OPTIMIZATION AUTOMATION INTEGRATION PRECISION QU  
MPUTE CORE FABRIC MATRIX VECTOR TENSOR GRAPH NODE EDGE FLOW S

Groq v GPU  
50M People by the Numbers



12x

Economically  
Efficient

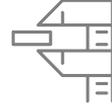
CapEx



75%

Minimized  
Running Cost

OpEx



10x

Optimized  
Ownership

TCO



5x

Energy  
Efficiency

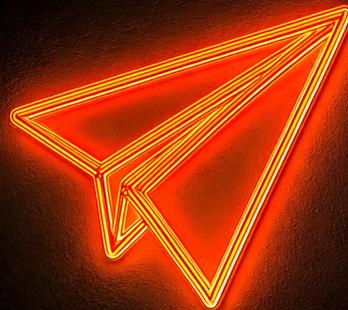
Consumption

groq

We've  
done this

~~twice.~~

**3x!!**



Feb 2025

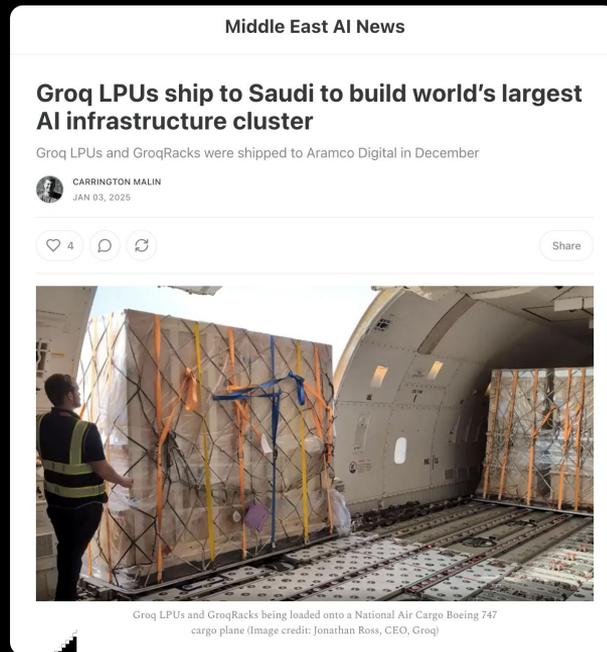
# Our First International Expansion

Groq + Humain



“Aramco’s tech arm takes the lead as global giants sign deals at LEAP.”

Since 1972  
**ARAB NEWS**  
*The Voice of a Changing Region*



May 2025

# Our **Next** International Expansion

Groq + Bell Canada



CISION

## Groq Becomes Exclusive Inference Provider for Bell Canada's Sovereign AI Network

PR Newswire

Wed, May 28, 2025 at 9:36 AM EDT • 2 min read



*New data centers across North America expand Groq's network, now serving over 20 million tokens per second*

MOUNTAIN VIEW, Calif., May 28, 2025 /PRNewswire/ -- Groq, the pioneer in fast AI inference, today announced an exclusive partnership with Bell Canada to power Bell AI Fabric, the country's largest sovereign AI infrastructure project.

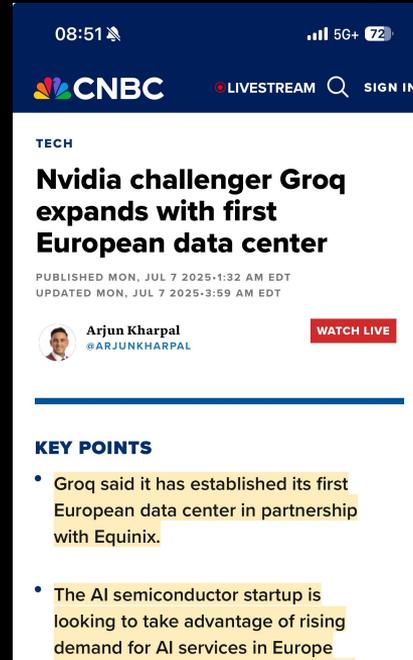
“Groq’s technology delivers the speed and efficiency our customers need—now, not years from now.”

Mirko Bibic, President & CEO, BCE and Bell Canada

May 2025

# Our **First** European Expansion

Groq + Equinix



08:51 5G+ 72

CNBC LIVESTREAM SIGN IN

TECH

## Nvidia challenger Groq expands with first European data center

PUBLISHED MON, JUL 7 2025-1:32 AM EDT  
UPDATED MON, JUL 7 2025-3:59 AM EDT

Arjun Kharpal @ARJUNKHARPAL WATCH LIVE

### KEY POINTS

- Groq said it has established its first European data center in partnership with Equinix.
- The AI semiconductor startup is looking to take advantage of rising demand for AI services in Europe

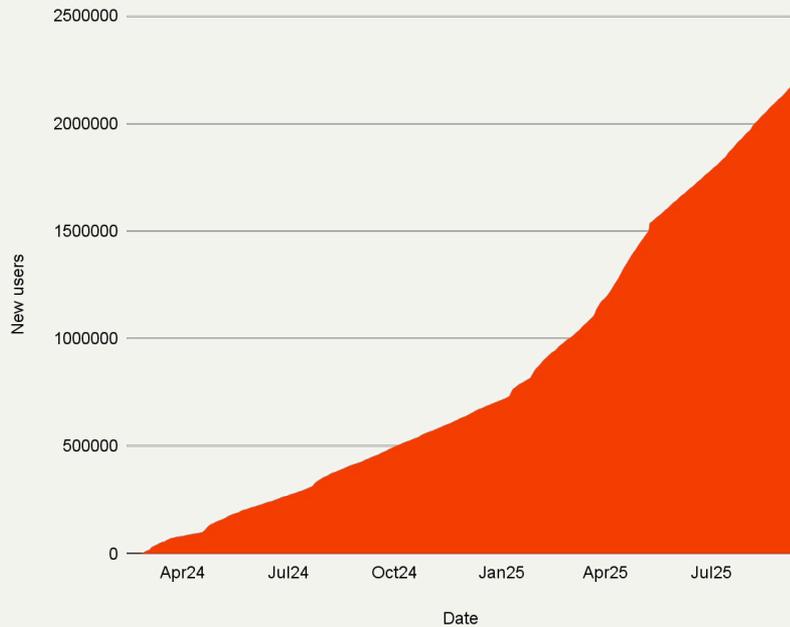
“We’re not as supply limited, and that’s important for inference,” Ross told CNBC’s “Squawk Box Europe.”

# Groq's Data Center Footprint is Global



● Data Centers in Operation

● Data Centers coming in 2025



DEVELOPERS CHOOSE GROQ

2.2M

joined in 18 months

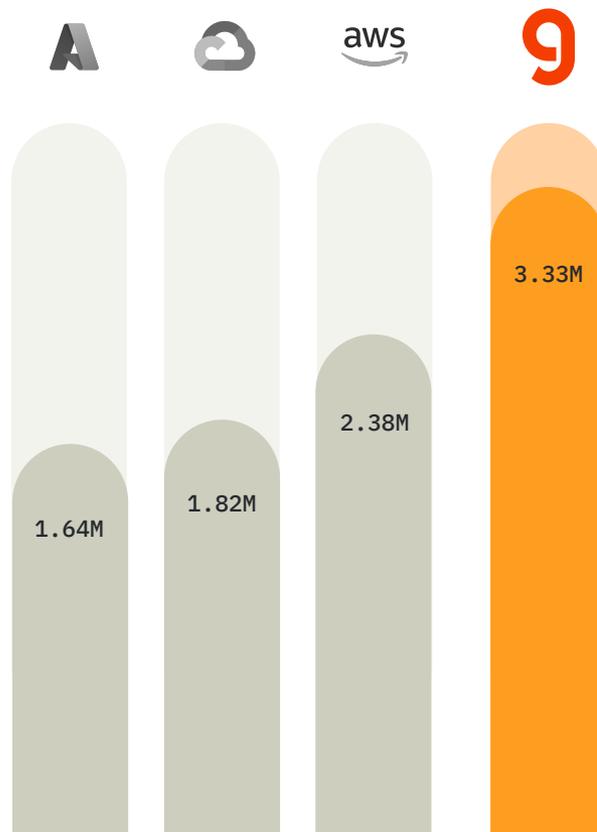
# And How Much **More** You Get For Your Dollar

Close the gap between idea and production, without trade-offs in performance, intelligence, or price.

Stop paying more for less.

Pay less for more.

**Tokens per \$1**  
Llama 4 Maverick, Blended Price



# AI Success = How Well You **Scale**

Defaulting to GPUs is the least good, most expensive choice. GPUs deliver overbuilt capacity at runaway cost.

There is a better way.

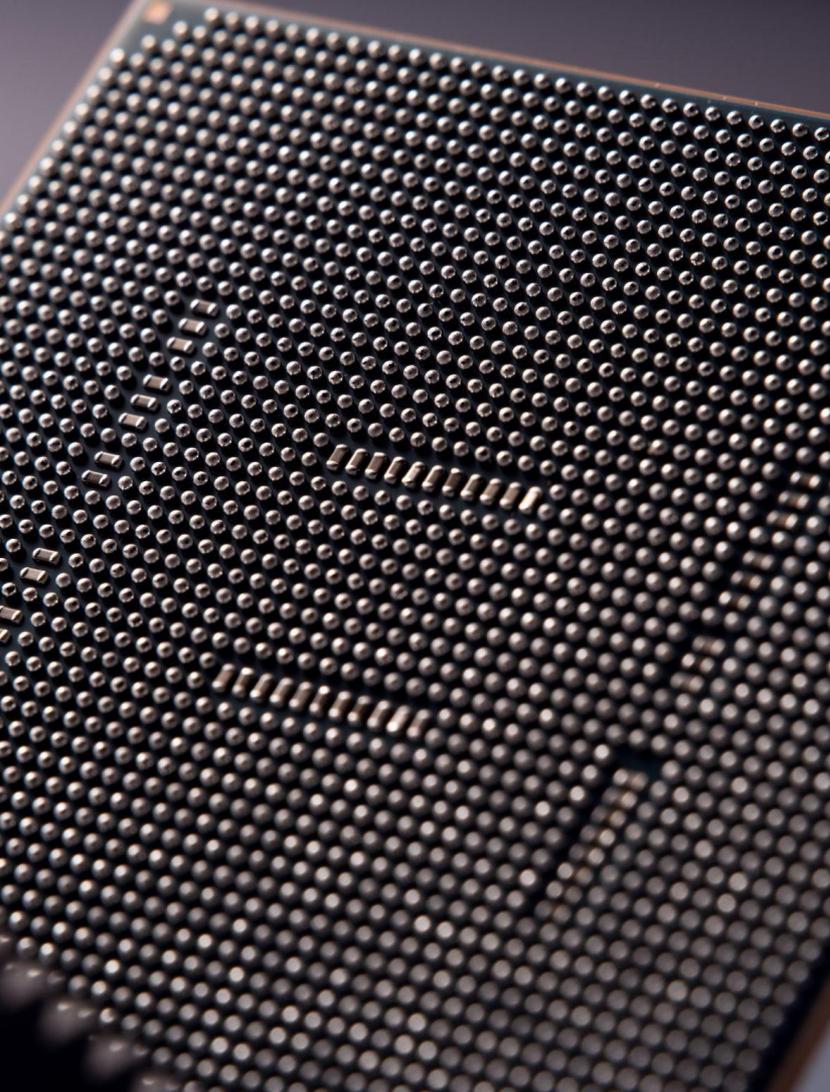
Cost to run 1M tokens of gpt-oss-120B  
across inference providers



The LPU is  
the **cartridge**.

GroqCloud is  
the **console**.





# Meet the LPU

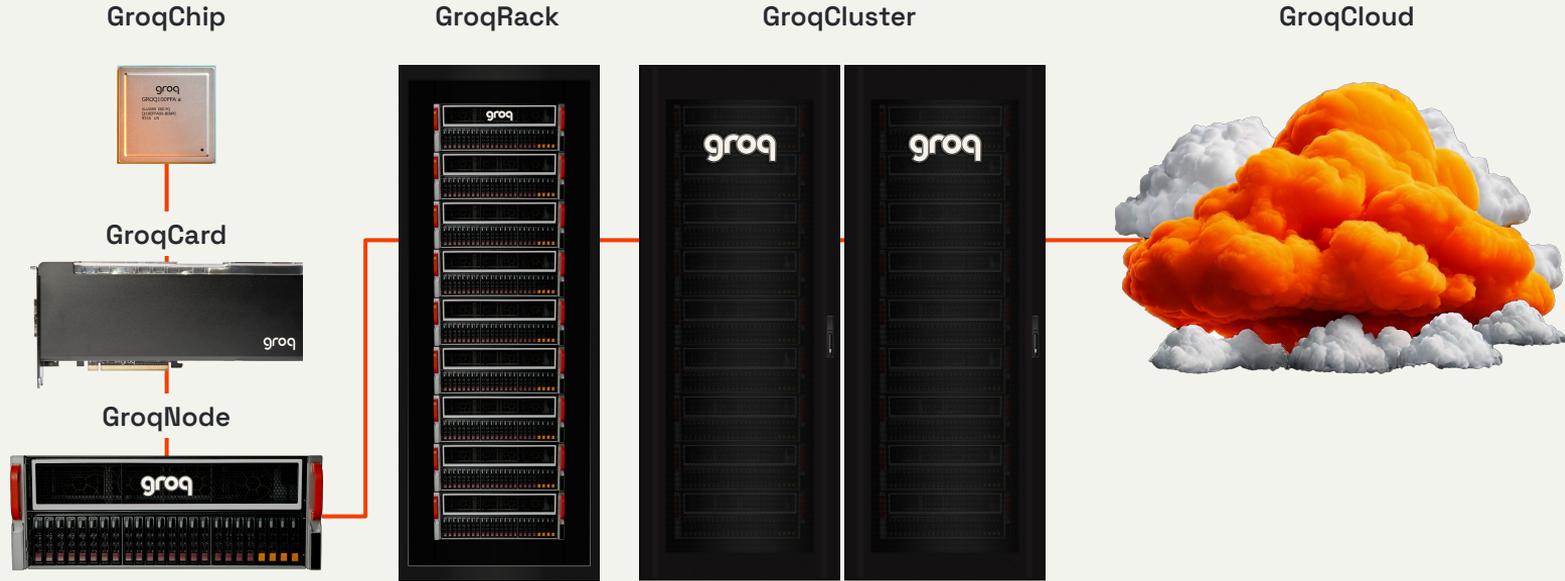
**BUILT FOR SPEED AND PRECISION**

Compiler and software-defined, single-core architecture removes traditional software complexity while continuous, token-based execution delivers consistent performance without tradeoffs. Every cycle is accounted for. No wasted operations, no unpredictable delays.

Single Core & On-Chip SRAM	Custom Compiler, Fully In Control	Power Efficient	Direct Connectivity
The LPU integrates hundreds of MB of SRAM as primary weight storage (not cache), cutting latency and feeding compute units at full speed. This enables efficient tensor parallelism across chips.	Groq's purpose-built compiler enables static scheduling and deterministic execution, ensuring predictable performance at every scale.	Air-cooled by design, Groq's LPU and GroqRack require no complex cooling and power infrastructure, cutting operating costs and lowering environmental impact.	The LPU aligns hundreds of chips to act as a single core. The compiler predicts data arrival precisely, coordinating compute and network scheduling without caches or switches.

# Full Stack Innovation

SILICON + SYSTEMS + SOFTWARE + NETWORKING + APIS





### Advanced Services

Performance Tier | Batch & Flex Processing | Fine-Tuning & LoRAs | Dedicated Instances | Multi-Tenant TaaS

### Developer Tools

Built-In Tools | Compound AI | OpenAI Compatible Endpoints | Function Calling

### Core Platform

Groq Compiler | Inference Runtime | Security & Compliance | Multi-Region Availability

### Custom Silicon

GroqRack Compute Cluster | GroqChip LPU

# Meet GroqCloud

THE WORLD'S FIRST CLOUD FOR INFERENCE,  
BUILT FROM SILICON UP.

Create more value from AI than you spend to run it.  
Fast, low cost inference that holds steady at scale.

#### Grow Without Surprise Bills

Other inference providers turn growth into a budget gamble with elastic pricing and surprise surcharges. Groq pricing is flat and linear, so every new user drives predictable revenue, not operational risk.

#### Speed that Never Slips

Our edge? Custom silicon. Groq delivers low latency that stays consistent across traffic, regions, and workloads. Every design choice focuses on keeping intelligence fast and affordable.

#### Intelligence You Can Trust

Groq's LPU architecture delivers precise outputs without accuracy loss. What you test is what you get in production. No batching tricks or hidden compromises. Just raw, consistent performance every time.

GLOBAL COVERAGE | SOVEREIGN AI | SECURITY





**Kent Beck** 🌻 🟩  
@KentBeck

I've been reluctant to try ChatGPT. Today I got over that reluctance. Now I understand why I was reluctant.

The value of 90% of my skills just dropped to \$0. The leverage for the remaining 10% went up 1000x. I need to recalibrate.

12:51 PM · Apr 18, 2023 · **1.4M** Views