Delivering Sovereign AI

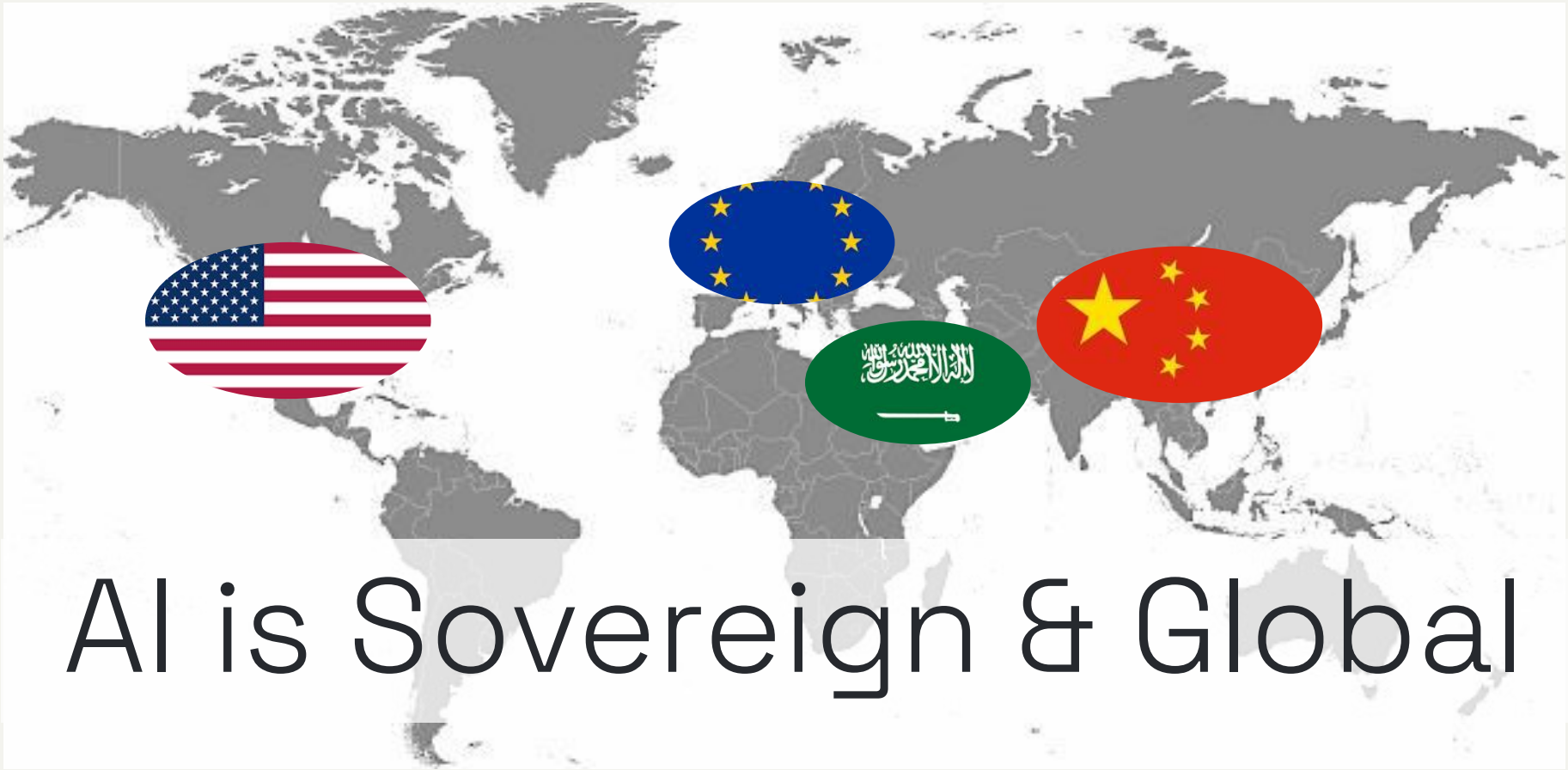
groq

Chris Stephens

VP, CTO @ Groq

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

groq



AI is Sovereign & Global

groq

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，请认真贯彻执行。

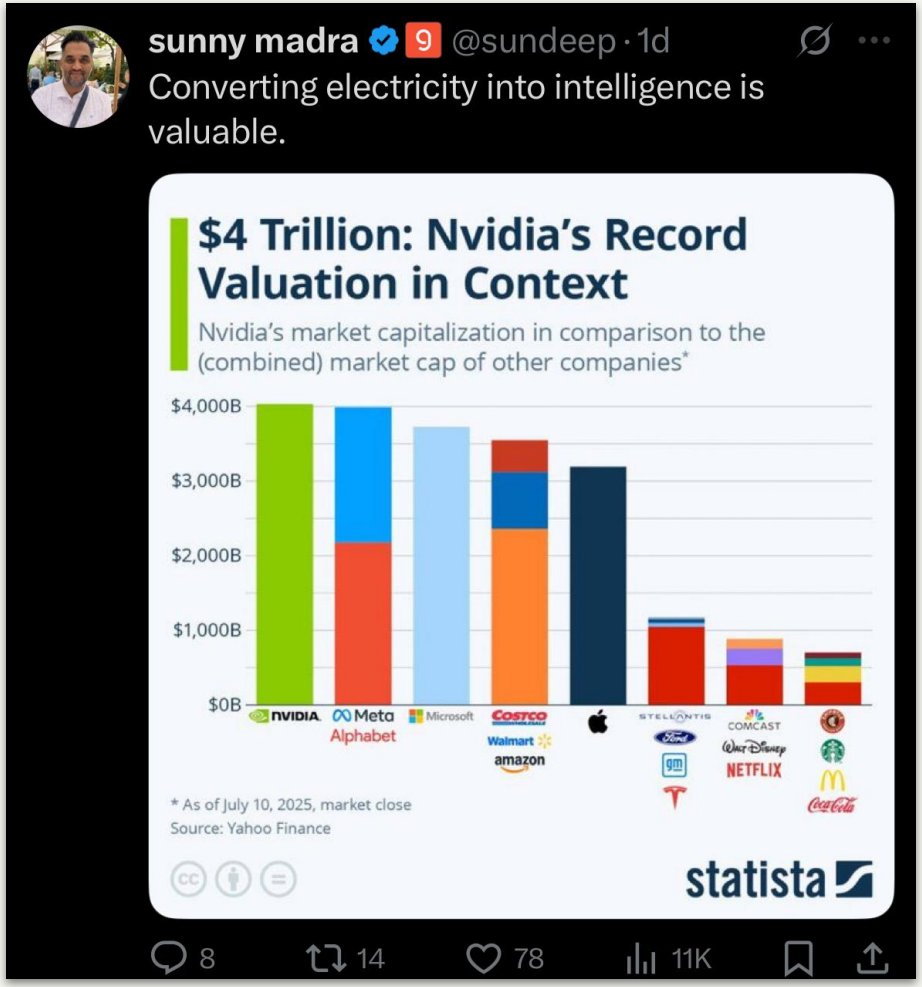
2017年7

（此件公开发布）

新一代人工智能发展规划

groq

Why?


sunny madra @sundeep · 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*
$4,000B
$3,000B
$2,000B
$1,000B
$0B
NVIDIA
Meta
Alphabet
Microsoft
Costco
Walmart
amazon
COMCAST
NETFLIX
* As of July 10, 2025, market close
Source: Yahoo Finance
statista
8
14
78
11K

groq

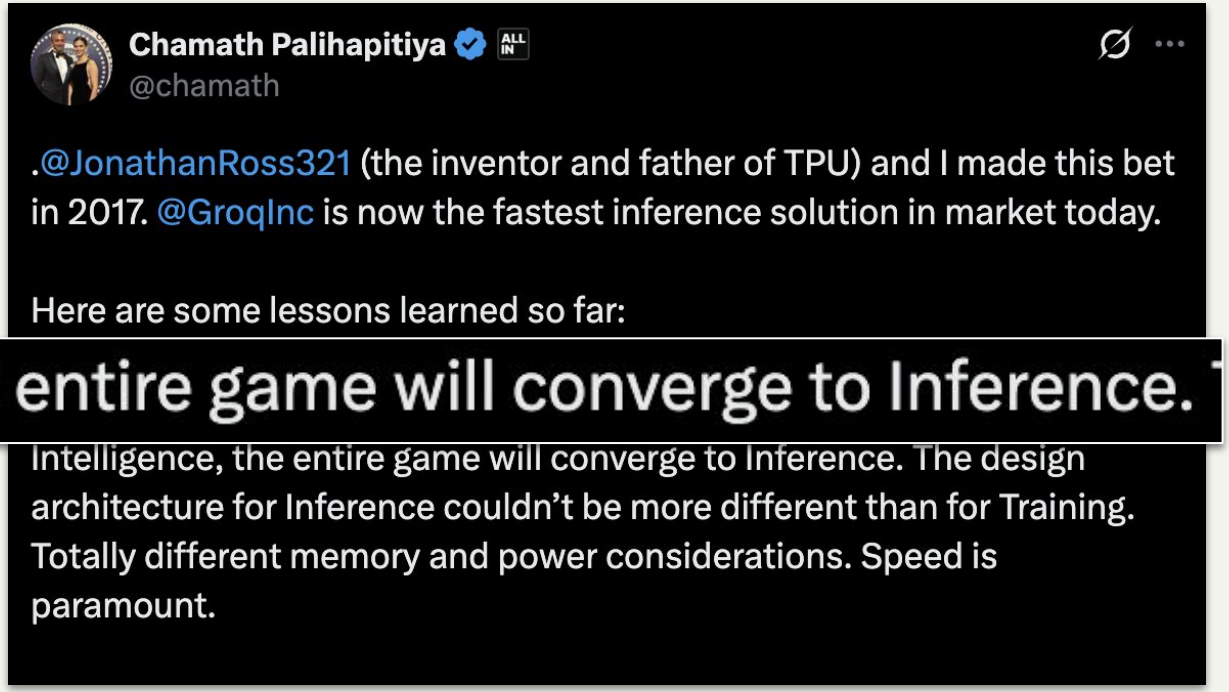
Global AI Trends

Why?

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

groq

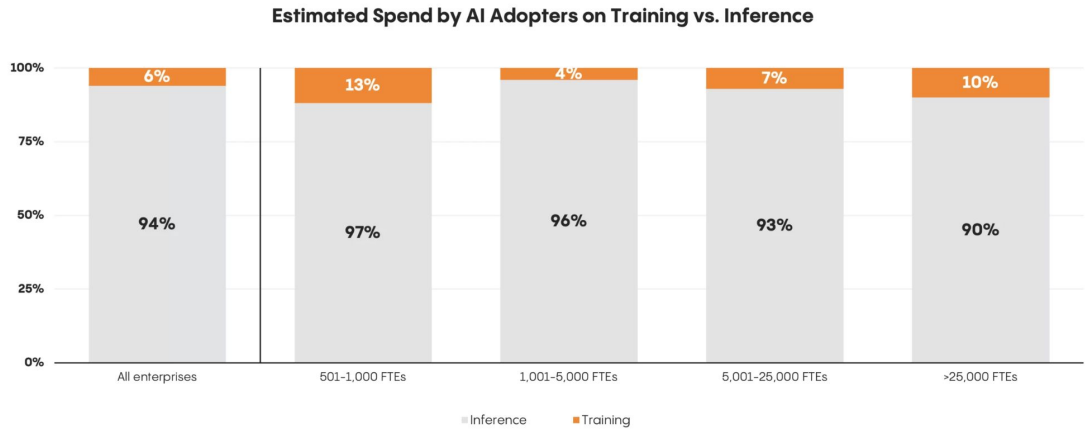
Why?


Chamath Palihapitiya
ALL IN
@chamath
.@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.

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



© 2024 Menlo Ventures

https://www.accenture.com/content/dam/accenture/final/industry/high-tech/document/Accenture-Unleashing-full-potential-of-AI.pdf
https://menlovc.com/perspective/the-modern-ai-stack-design-principles-for-the-future-of-enterprise-ai-architectures/

groq

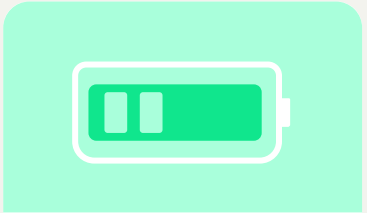
THE PROMISE OF AI IS

Massive and Unrealized

1950s–2000s

Algorithms

Enabled first mechanical automation of reasoning



2010s–PRESENT

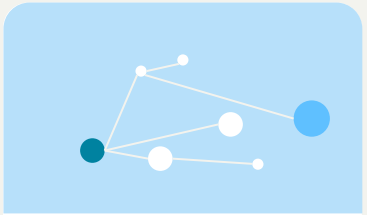
Training

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

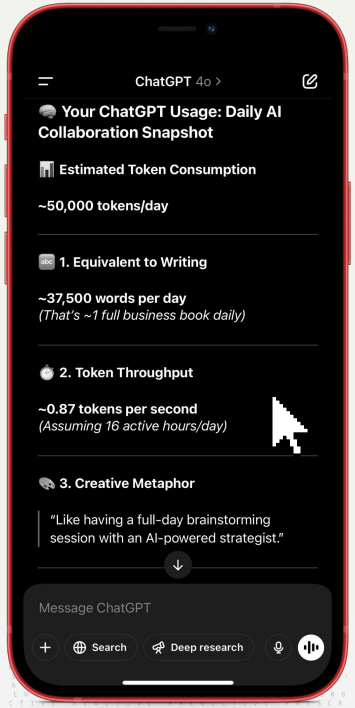
Instant, affordable, intelligent AI woven into daily life

Only the beginning

This will be huge.

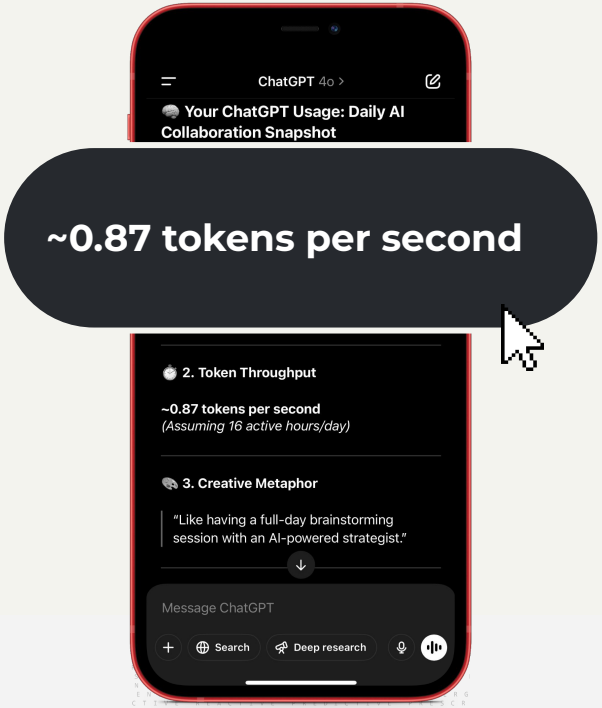
Public + private demand is persistent, concurrent, and growing

We all need Inference



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

how much inference will be you need?



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

Groq v GPU

50M People by the Numbers

12x

Economically Efficient

CapEx

75%

Minimized Running Cost

OpEx

10x

Optimized Ownership

TCO

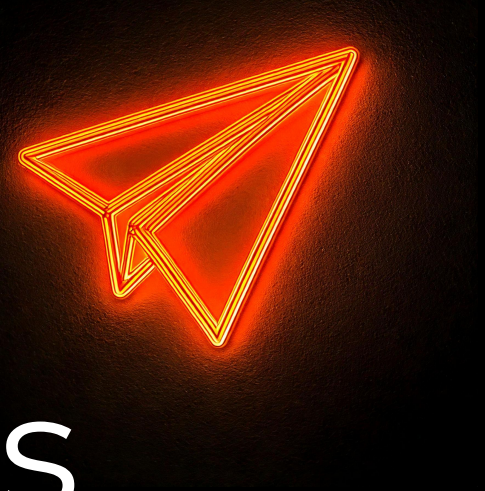
5x

Energy Efficiency

Consumption

groq

We've done this ~~twice~~ 3x!!



Groq Proprietary Under NDA

Feb 2025

Our First International **Expansion**

Groq + Humain

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

Since 1975
ARAB NEWS
The Voice of a Changing Region

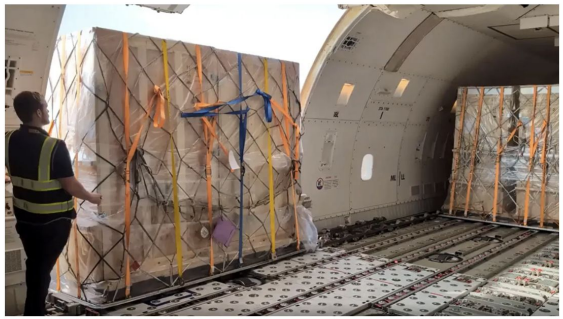
Middle East AI News

Groq LPUs ship to Saudi to build world’s largest AI infrastructure cluster

Groq LPUs and GroqRacks were shipped to Aramco Digital in December

CARRINGTON MALIN
JAN 03, 2025

4 Share



Groq LPUs and GroqRacks being loaded onto a National Air Cargo Boeing 747 cargo plane (Image credit: Jonathan Ross, CEO, Groq)

Groq Proprietary Under NDA

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.

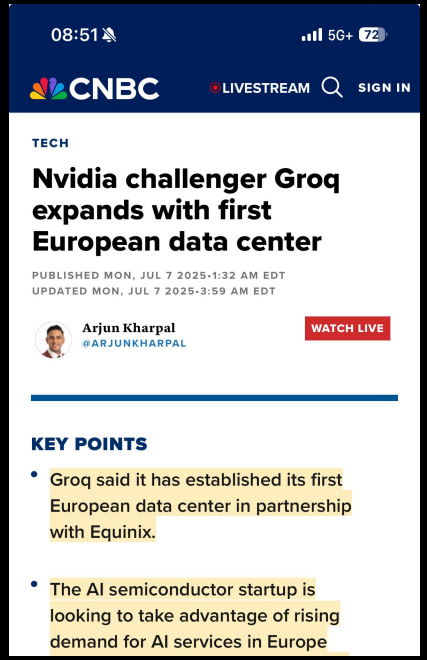
> “Groq’s technology delivers the speed and efficiency our customers need—now, not years from now.”
>
> Mirko Bibic, President & CEO, BCE and Bell Canada

Groq Proprietary Under NDA

May 2025

Our **First** European Expansion

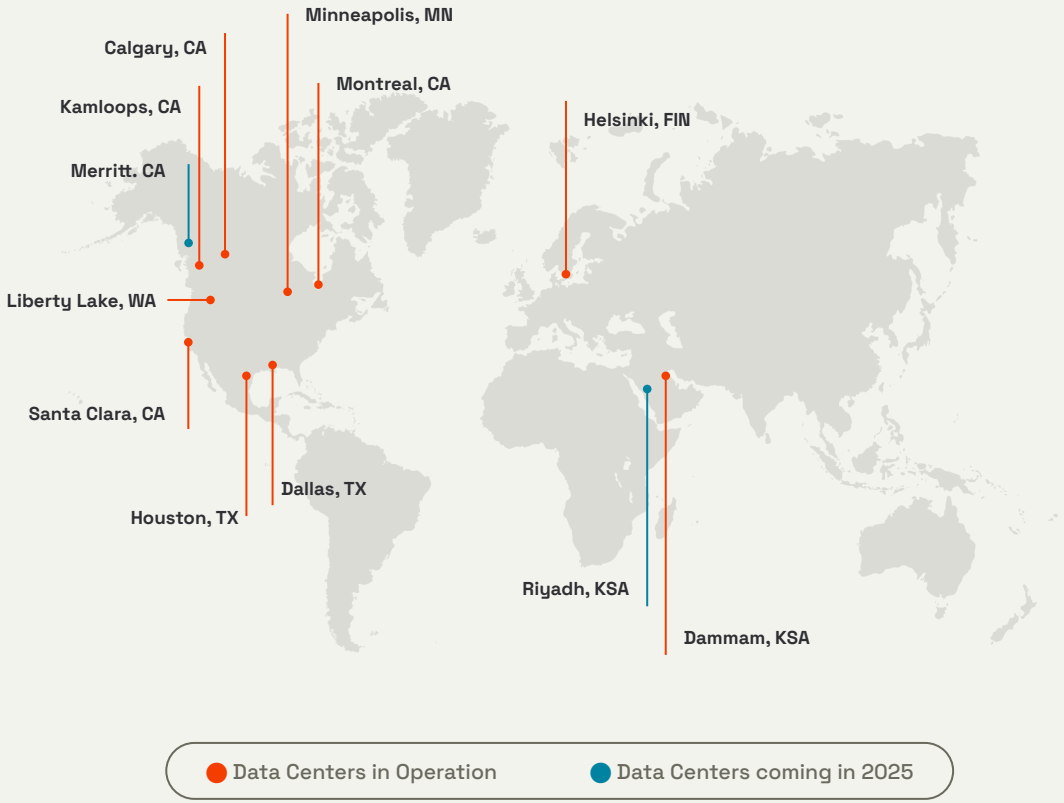
Groq + Equinix

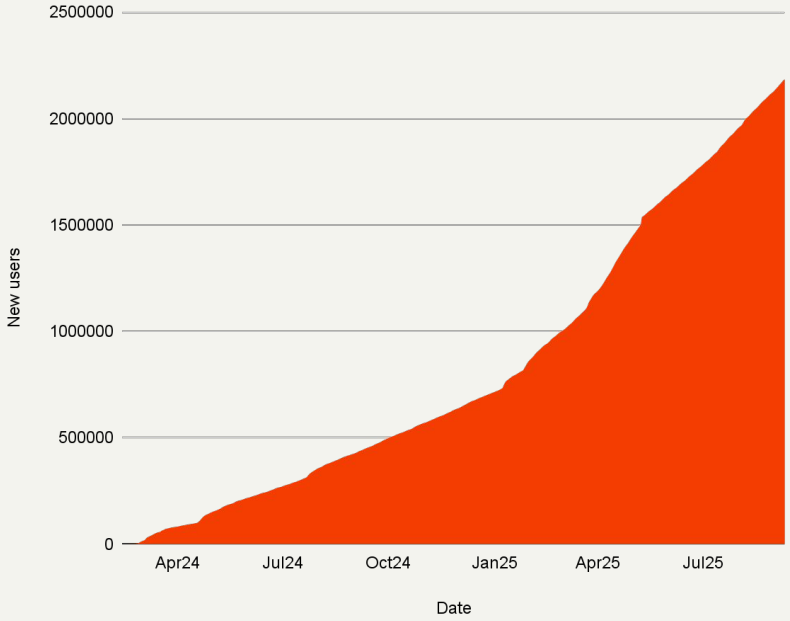


> “We’re not as supply limited, and that’s important for inference,” Ross told CNBC’s “Squawk Box Europe.”

Groq Proprietary Under NDA

Groq's Data Center Footprint is Global





DEVELOPERS CHOOSE GROQ

2.2M

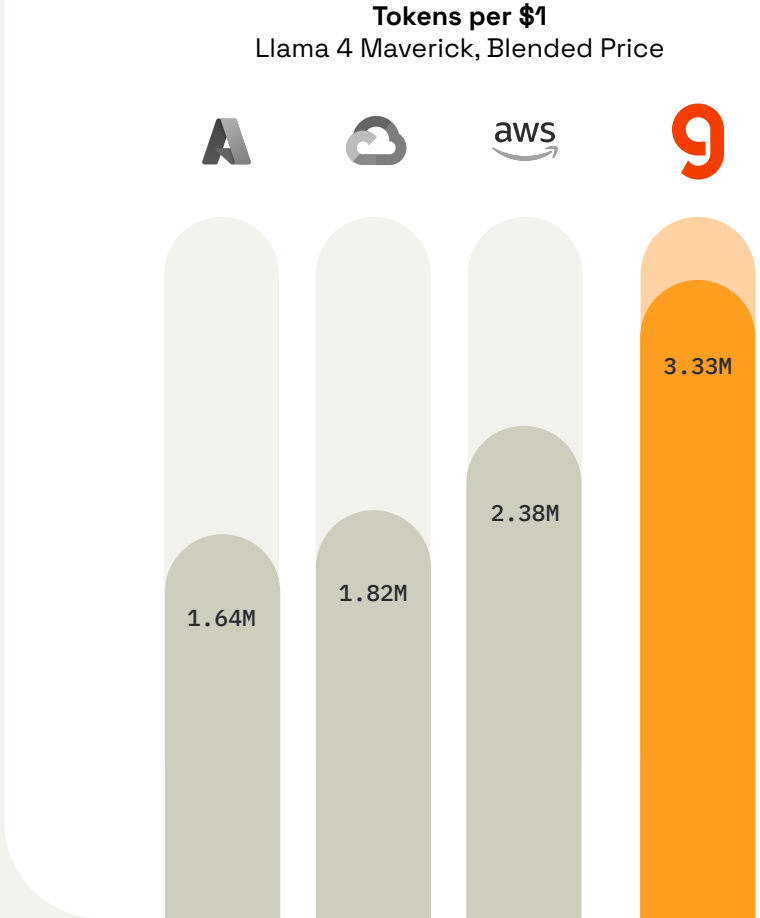
joined in 18 months

PRODUCTION SCALE X ECONOMICS

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.

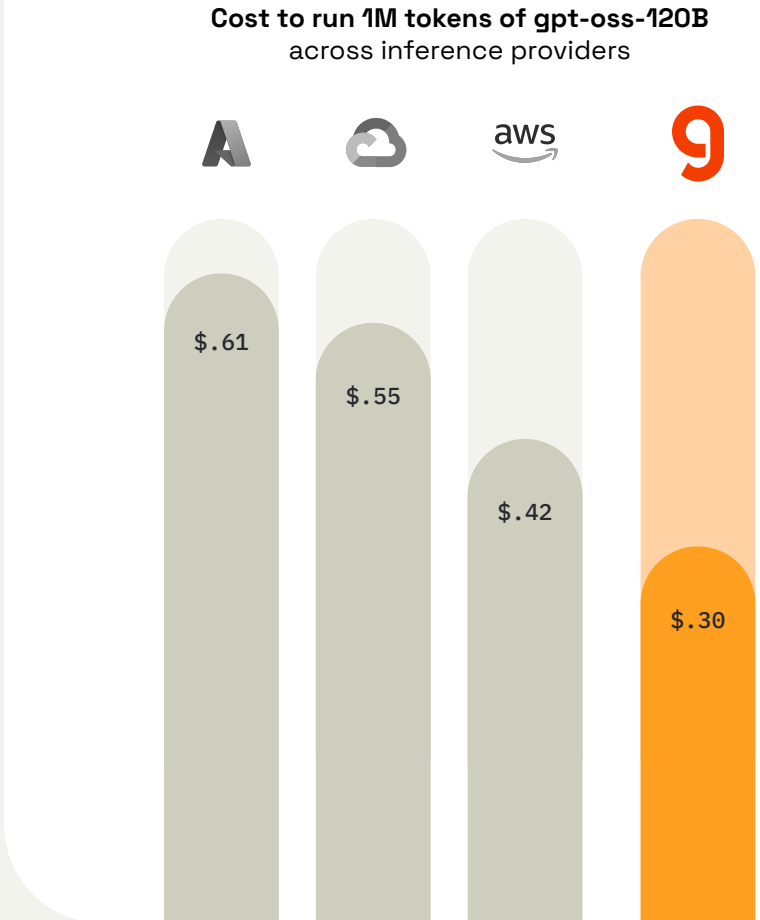


PRODUCTION SCALE X ECONOMICS

AI Success = How Well You **Scale**

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

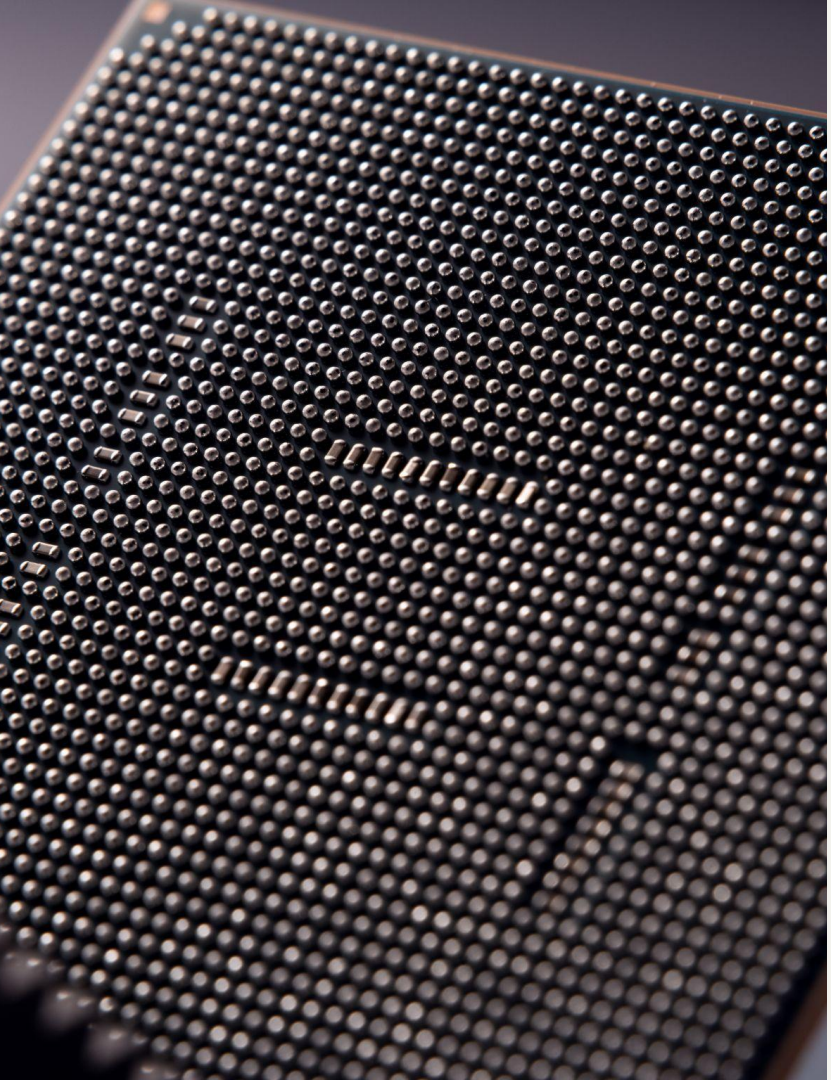
There is a better way.



groq

The LPU is
the **cartridge**.

GroqCloud is
the **console**.



groq

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

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.

Custom Compiler, Fully In Control

Groq's purpose-built compiler enables static scheduling and deterministic execution, ensuring predictable performance at every scale.

Power Efficient

Air-cooled by design, Groq's LPU and GroqRack require no complex cooling and power infrastructure, cutting operating costs and lowering environmental impact.

Direct Connectivity

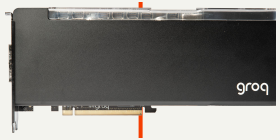
The LPU aligns hundreds of chips to act as a single core. The compiler predicts data arrival precisely, coordinating compute and network scheduling without caches or switches.

Full Stack Innovation

SILICON + SYSTEMS + SOFTWARE + NETWORKING + APIS

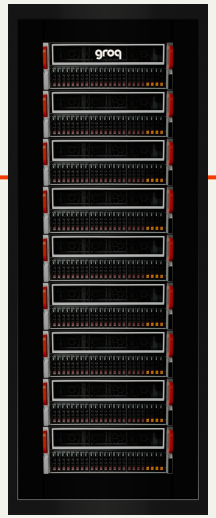
GroqChip

GroqCard

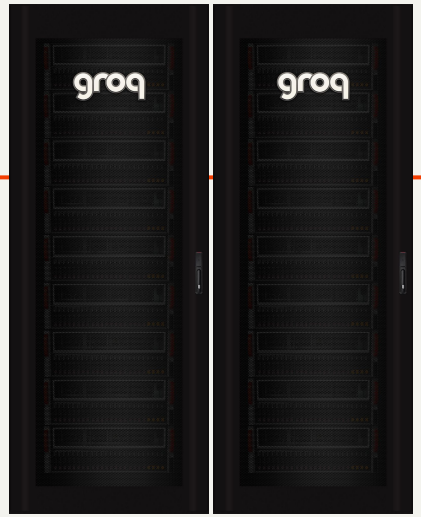


GroqNode

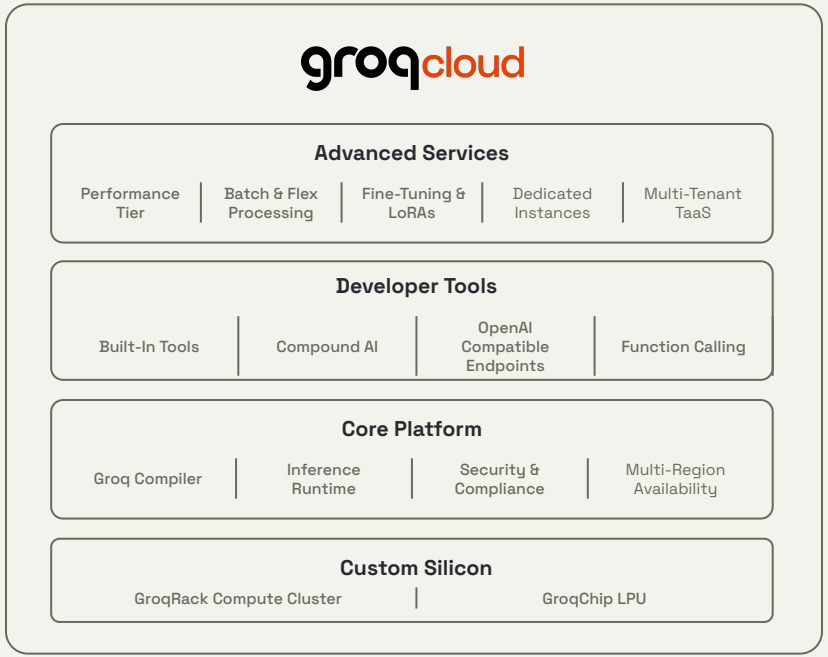
GroqRack



GroqCluster



GroqCloud



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

groq


Kent Beck on X: "I've b
x.com/KentBeck/status/1648413998025707520
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