Important Disclosures

Independent Third-Party Research Report

Important Disclosure

This research report was prepared by an independent third-party provider and is made available by StartEngine Crowdfunding, Inc. (“StartEngine”) for informational purposes only. StartEngine and its affiliates, including StartEngine Primary, LLC, a registered broker-dealer, have not independently verified the information contained herein and do not guarantee its accuracy or completeness.

The views expressed in this report are those of the independent provider and do not necessarily reflect the views of StartEngine, its affiliates, or any of their employees.

This report does not constitute an offer to sell or a solicitation of an offer to buy any securities, nor does it constitute a recommendation to buy, sell, or hold any securities. Please carefully review the risks, conflicts of interest, and disclosures provided in the **Important Disclosures** section located at the end of this document.

Important Disclosures

Independent Third-Party Research

Important Disclosures

This research report was prepared by an independent third-party research provider without input, review, or influence by StartEngine Crowdfunding, Inc. ("StartEngine") or its affiliates and is distributed from StartEngine for informational purposes only. StartEngine has not independently verified the information contained herein and does not guarantee its accuracy or completeness. This report reflects the views of the independent provider and does not necessarily represent the views of StartEngine or its affiliates.

StartEngine does not endorse or make any representations regarding the conclusions or recommendations contained in this report. This report does not constitute a recommendation or solicitation by StartEngine to buy, sell, or hold any securities.

Investing in private companies involves a high degree of risk, including the loss of your entire investment. Past performance does not guarantee future returns.

This report is made available to all individuals on startengine.com and may be distributed to individuals who have expressed interest in specific companies or industries.

Please carefully review the company-specific risks, conflicts of interests, and disclosures below.

Company Specific Disclosures:

Conflict of Interest: StartEngine or its affiliates sell membership interests of:

- Series 21-1, a Series of StartEngine Private LLC which indirectly owns Groq, Inc. Series D Preferred Stock via a special purpose vehicle;
- Series 21-2, a Series of StartEngine Private LLC which indirectly owns Groq, Inc. Series B Preferred Stock via a special purpose vehicle;
- Series 2-1, a Series of StartEngine Private Funds LLC which indirectly owns Groq, Inc. Series D Preferred Stock via a special purpose vehicle;
- Series 3-1, a Series of StartEngine Private Funds LLC which indirectly owns Groq, Inc. Series B Preferred Stock via a special purpose vehicle;
- Series 6-1, a Series of StartEngine Private Funds LLC which indirectly owns Groq, Inc. Series E Preferred Stock via a special purpose vehicle; and
- Series 6-2, a Series of StartEngine Private Funds LLC which indirectly owns Groq, Inc. Series E Preferred Stock via a special purpose vehicle.

Security Ownership: StartEngine or its affiliates indirectly owns Series B, Series D, and Series E Preferred Stock of Groq, Inc.

For further questions regarding this report or its distribution, please contact StartEngine at contact@startengine.com representative.

SACRA

EQUITY RESEARCH

UPDATED
07/18/2025

Groq

TEAM

Jan-Erik Asplund
Co-Founder
jan@sacra.com

Marcelo Ballve
Head of Research
marcelo@sacra.com

DISCLAIMERS

Groq

Visit Website

AI inference chip and cloud tool for running large language models efficiently

#ai #ai-chips

REVENUE
$90,000,000
2024

VALUATION
$3,600,000,000
2025

FUNDING
$901,130,000
2024

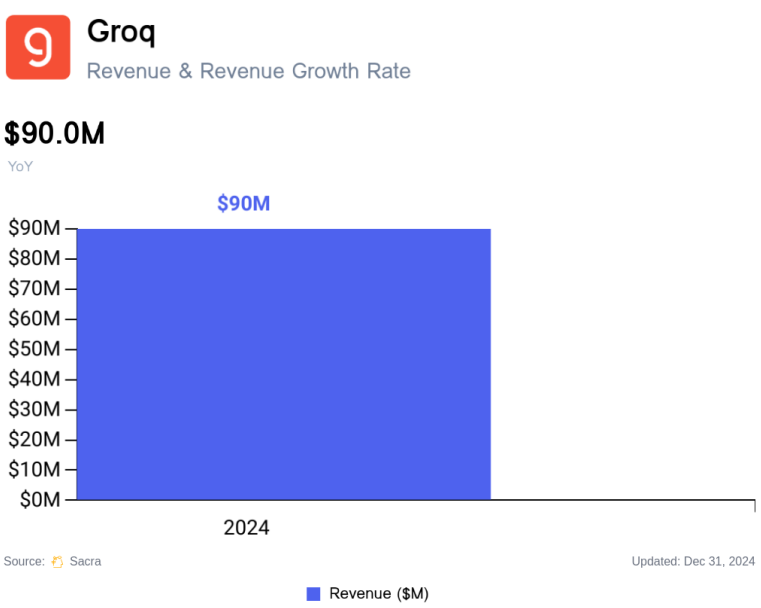
Details

HEADQUARTERS
Mountain View, CA

CEO
Jonathan Ross

SACRA

Revenue



Sacra estimates that Groq generated $90M in revenue in 2024. The company primarily generates revenue from selling cloud services for companies to run AI on its chips, similar to how companies purchase access to OpenAI models or AI models from Amazon Web Services. Groq has told investors that it is projecting $500M in revenue for the year 2025,.

Groq also sells its chip systems and data center operating services to other companies, including telecommunications company Bell Canada. The company reports that nearly 2 million developers and teams have used its services, indicating strong adoption of its GroqCloud platform.

The revenue growth reflects the company's transition from a pure hardware play to a cloud-first business model, capitalizing on the massive demand for high-speed AI inference capabilities across enterprise customers and the developer community.

Valuation

Groq raised $640 million in a Series D round in August 2024, led by BlackRock Private Equity Partners. The round brought Groq's total funding to over $1 billion since its founding in 2016.

Other investors in the Series D include Neuberger Berman, Type One Ventures, Cisco Investments, Global Brain, and Samsung Catalyst Fund. Previous major rounds included a $300 million Series C in April 2021 co-led by Tiger Global Management and D1 Capital.

Key strategic investors across the company's funding history include Founders Fund, Addition, and The Spruce House Partnership, reflecting strong institutional backing from both traditional venture capital and strategic corporate investors.

Product

Groq is a specialized AI inference company built around its proprietary Language Processing Unit (LPU) chip architecture. The LPU uses a statically-scheduled tensor streaming processor design that moves data through a deterministic pipeline, delivering ultra-low latency and up to 10x energy efficiency compared to traditional GPUs.

The core product is GroqCloud, a fully managed inference platform that developers can access through an OpenAI-compatible REST API. Developers can switch to Groq by changing just three lines of code—their API key, base URL, and model name—and immediately start streaming tokens at hundreds of tokens per second with sub-10 millisecond first-token latency.

GroqCloud supports curated open-source models like Llama-3, Qwen, and Mixtral, with performance benchmarks showing speeds like 1,345 tokens per second for Llama-3 8B and 662 tokens per second for Qwen-3 32B. The platform includes Python and JavaScript SDKs, streaming responses, and integrations with popular frameworks like LangChain, LlamaIndex, and Vercel AI SDK.

For enterprise customers, Groq offers GroqRack compute clusters—on-premises or colocation racks containing 64 to 576+ LPUs per rack. These systems target hyperscalers, sovereign clouds, and regulated industries requiring data residency, with customers like Aramco Digital ordering hundreds of racks for large-scale deployments.

Business Model

Groq operates a three-tier business model that ramps customers from cloud services to dedicated hardware. The company's B2B go-to-market approach targets both individual developers and enterprise customers through different engagement models.

GroqCloud uses a pay-per-token pricing model similar to OpenAI, where customers pay only for tokens generated. The platform offers both self-serve access with no minimum commitments and enterprise plans with annual volume commitments and volume discounts. This usage-based model allows customers to easily estimate costs before engaging with sales teams.

The hardware business involves selling GroqRack systems directly to enterprises, telecommunications companies, and cloud providers. These customers typically require on-premises deployment for data sovereignty, regulatory compliance, or performance requirements that cloud services cannot meet.

Groq's cost structure benefits from its vertically integrated approach, controlling both chip design and the software stack. The company manufactures chips through foundry partnerships, currently using GlobalFoundries' 14nm process with plans to move to Samsung's 4nm process for next-generation chips. This vertical integration allows Groq to optimize the entire inference pipeline while maintaining higher margins than pure software plays.

Competition

GPU incumbents

Nvidia continues to dominate the AI inference market with its H200, B200, and GH200 Grace-Hopper systems, maintaining over 80% of deployed inference GPUs. While Nvidia typically requires 8-16 GPUs to match Groq's token-per-second performance on large language models, the company's CUDA software ecosystem remains the stickiest competitive barrier. Nvidia is pushing its Blackwell B200 and GB200 systems to reset the performance-per-watt curve while bundling networking components to lock in customers.

AMD is positioning its MI300X, MI325X, and upcoming MI355X chips as cost-effective alternatives, with some benchmarks showing advantages on specific workloads. However, AMD's ROCm tooling still lags significantly behind CUDA in developer adoption, and the company faces supply chain challenges that have delayed competitive responses to Nvidia's latest generations.

Intel's Gaudi 3 chips claim 50% better performance than H100 at 40% lower total cost of ownership, but software fragmentation across oneAPI and SynapseAI has slowed enterprise adoption. Intel is targeting cloud partnerships by bundling Gaudi 3 with other data center components.

Specialized inference chips

Cerebras and SambaNova are building competing ASIC architectures optimized for AI inference, arguing that GPU economics break down when batch sizes are small for applications like chat, agent loops, and code assistance. These companies are positioning around extreme throughput-per-watt and deterministic latency, similar to Groq's approach.

Hyperscalers are developing internal silicon solutions including Google's TPU v5, Amazon's Inferentia 3, and Microsoft's Cobalt and Maia chips. While these primarily serve internal workloads, they represent potential competition for third-party inference demand as hyperscalers may offer these capabilities to external customers.

Cloud inference platforms

Traditional cloud providers like AWS, Google Cloud, and Microsoft Azure are expanding their AI inference offerings, leveraging existing customer relationships and integrated billing. These platforms can bundle inference with other cloud services, creating switching costs that pure-play inference companies like Groq must overcome through superior performance or pricing.

TAM Expansion

New products and technology

Groq is developing next-generation 4nm LPU chips through its partnership with Samsung Foundry, promising significant performance-per-watt improvements over current 14nm parts. This technology advancement opens opportunities in higher-value systems markets and enables support for larger context windows and real-time multimodal models.

The company's acquisition of Definitive Intelligence bundled data preparation, orchestration, and dashboard capabilities with Groq hardware, allowing enterprises to purchase turnkey inference systems rather than just chips. This vertical software stack expansion moves Groq up the value chain from hardware provider to complete solution vendor.

GroqCloud's tokens-as-a-service model transforms raw silicon into recurring software revenue streams. The Series D funding will add over 100,000 LPUs to the cloud by early 2025, enabling support for more sophisticated AI applications and larger-scale deployments.

Customer base expansion

Groq has grown from zero to over 360,000 developers in 18 months, with 75% of Fortune 100 companies maintaining accounts on the platform. By supporting open-source communities around models like Llama 3, Gemma, and Mixtral, Groq expands beyond Big Tech buyers into small and medium business developer tool budgets.

The company's reseller partnership with Carahsoft and FedRAMP roadmap provide pathways into US defense and civil agencies requiring secure, low-latency LLM inference. This public sector expansion represents a significant market opportunity given government AI adoption initiatives.

Enterprise customers are increasingly evaluating alternatives to Nvidia-based solutions due to cost and supply constraints, creating opportunities for Groq to capture market share in industries requiring high-performance inference at scale.

Geographic expansion

Groq secured a $1.5 billion commitment from Saudi Arabia in February 2025 to fund the largest non-hyperscaler inference cluster with over 19,000 LPUs. This positions Groq as the reference platform for Arabic language models and establishes a major presence in the Middle East market.

The company launched its first European data center region in Helsinki in July 2025, meeting data sovereignty requirements while leveraging green hydroelectric power for lower operating costs. This European expansion addresses regulatory compliance needs for multinational customers.

Groq is exploring partnerships to co-locate LPUs at remote renewable energy sites, allowing the company to monetize stranded energy while adding global capacity cost-effectively. This approach could enable expansion into emerging markets with abundant renewable resources.

Risks

Manufacturing concentration: Groq currently relies on GlobalFoundries' 14nm process for chip production, creating potential supply chain vulnerabilities. While the company has signed Samsung for next-generation 4nm chips, any manufacturing disruptions or geopolitical tensions affecting foundry access could significantly impact Groq's ability to scale production and meet growing demand.

Software ecosystem: Despite offering OpenAI-compatible APIs, Groq faces the challenge of competing against Nvidia's deeply entrenched CUDA ecosystem. Developers and enterprises have invested heavily in CUDA-based tooling, libraries, and workflows, creating switching costs that superior hardware performance alone may not overcome. The company must continue investing heavily in software development to match the breadth and maturity of established GPU software stacks.

Market timing: Groq's success depends on the continued growth of AI inference workloads and the shift from training-focused to inference-focused compute spending. If the AI market experiences a downturn or if new model architectures emerge that favor different hardware approaches, Groq's specialized LPU architecture could become less relevant, potentially stranding the company's significant R&D investments and manufacturing commitments.

Funding Rounds

Series D

Share Name	Issue Price	Issued At
Series D	$16.07811	Jan 2024
Series D-2	$14.4703	Jan 2024
Series D-1	$11.54007	Jan 2024

Series C

Share Name	Issue Price	Issued At
Series C	$11.54007	Apr 2021

Series B

Share Name	Issue Price	Issued At
Series B-2	$7.92373	Aug 2020
Series B-1	$7.45763	Aug 2020
Series B	$6.7554	Sep 2018

Series A

Share Name	Issue Price	Issued At
Series A-1	$4.8874	Sep 2018
Series A	$0.97244	Dec 2016

Figures sourced from the latest Certificate of Incorporation we have available.

DISCLAIMERS

This report is for information purposes only and is not to be used or considered as an offer or the solicitation of an offer to sell or to buy or subscribe for securities or other financial instruments. Nothing in this report constitutes investment, legal, accounting or tax advice or a representation that any investment or strategy is suitable or appropriate to your individual circumstances or otherwise constitutes a personal trade recommendation to you.

This research report has been prepared solely by Sacra and should not be considered a product of any person or entity that makes such report available, if any.

Information and opinions presented in the sections of the report were obtained or derived from sources Sacra believes are reliable, but Sacra makes no representation as to their accuracy or completeness. Past performance should not be taken as an indication or guarantee of future performance, and no representation or warranty, express or implied, is made regarding future performance. Information, opinions and estimates contained in this report reflect a determination at its original date of publication by Sacra and are subject to change without notice.

Sacra accepts no liability for loss arising from the use of the material presented in this report, except that this exclusion of liability does not apply to the extent that liability arises under specific statutes or regulations applicable to Sacra. Sacra may have issued, and may in the future issue, other reports that are inconsistent with, and reach different conclusions from, the information presented in this report. Those reports reflect different assumptions, views and analytical methods of the analysts who prepared them and Sacra is under no obligation to ensure that such other reports are brought to the attention of any recipient of this report.

Published on Jul 18th, 2025