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 18-1, a series of StartEngine Private LLC, which owns Common Stock of Lambda, Inc.
- Series 18-2, a series of StartEngine Private LLC, which owns Common Stock of Lambda, Inc.
- Series 18-3, a series of StartEngine Private LLC, which owns Series D-1 Preferred Stock of Lambda, Inc.
- Series 18-4, a series of StartEngine Private LLC, which owns Series Seed Preferred Stock of Lambda, Inc.
- Series 18-5, a series of StartEngine Private LLC, which owns Series Seed Preferred Stock of Lambda, Inc.

Security Ownership: StartEngine or its affiliates owns Common Stock, Series D-1 Preferred Stock, and Series Seed Preferred Stock of Lambda, Inc.

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

SACRA

EQUITY RESEARCH

UPDATED
05/10/2026

Lambda Labs

TEAM

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

Marcelo Ballve
Head of Research
marcelo@sacra.com

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.

Lambda Labs

Visit Website

Cloud GPU provider for training and fine-tuning AI models at scale

#ai #cloud-gpus

REVENUE
$760,000,000
2025

VALUATION
$5,900,000,000
2026

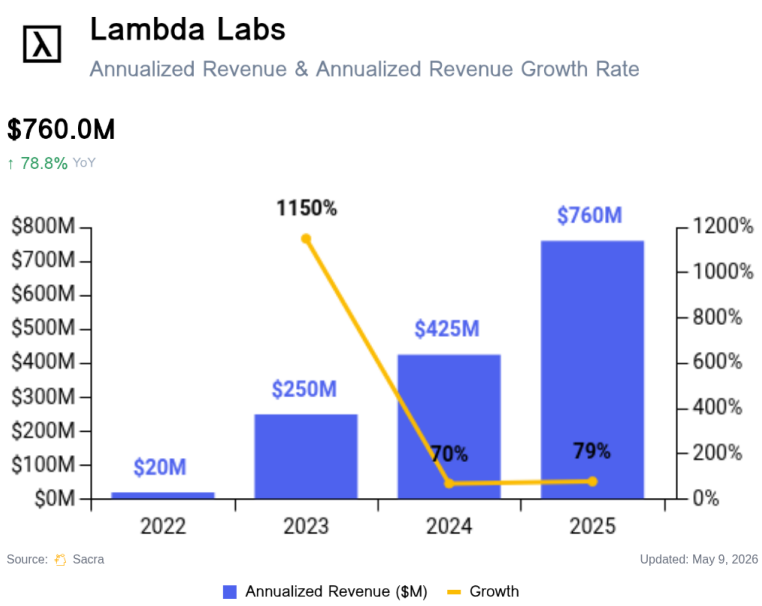
FUNDING
$2,300,000,000
2025

Details

HEADQUARTERS
San Jose, CA

CEO
Stephen Balaban

Revenue



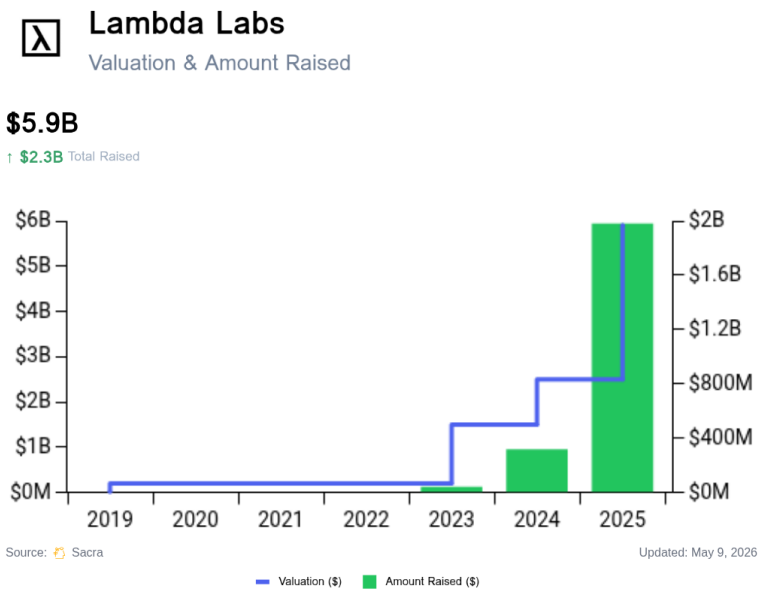
Sacra estimates that Lambda Labs hit $760M in annualized revenue in 2025, up 79% year-over-year from $425M at the end of 2024.

The company generated more than $520M in revenue over the trailing twelve months from October 2024 to September 2025, with Q3 sales up 80% year-over-year. In the first half of 2025 alone, revenue rose 33% to more than $250M, with Q2 contributing more than $140M—up nearly 60% year-over-year.

Lambda's cloud GPU rental business contributes the bulk of revenue, while legacy hardware sales make up a shrinking portion. Lambda's pricing emphasizes cost efficiency—Nvidia H100 PCIe instances at roughly $2.49 per hour compared to $4.25 at CoreWeave—helping drive utilization and expand its customer base of developers and enterprises.

Gross margin in the first half of 2025 was about 50%, or roughly 61% excluding non-cloud lines. The first-half net loss improved to about $24M, with the full trailing-twelve-month loss running approximately $175M. Lambda was in talks to raise $350M in pre-IPO financing, with Mubadala Capital in discussions to lead the round at roughly a 20% discount to the eventual IPO price, with an IPO targeted for the second half of 2026.

Valuation & Funding



Lambda raised over $1.5B in its Series E round led by TWG Global in November 2025, with participation from US Innovative Technology Fund and existing investors. Lambda declined to disclose the post-round valuation; prior to the round, Bloomberg reported discussions that could value the company at $4B–$5B.

This was up from $2.5B at its Series D in 2025. Earlier, Lambda was valued at $1.5B in its February 2024 Series C, $205.1M in its March 2023 Series B, and $87.5M in its July 2021 Series A.

Since its 2012 founding, Lambda has raised about $2.36B in equity funding, with notable investors including TWG Global, US Innovative Technology Fund, Nvidia, B Capital, and Gradient Ventures.

Product

In 2013, Lambda Labs launched its first product—a facial recognition API for Google Glass that let developers build AI-powered apps that could do things like remember faces and track down a specific face in a crowd, creating controversy and driving more than 5 million API calls every month.

In 2017, Lambda Labs pivoted to selling hardware built for AI and deep learning alongside their Face Recognition API, including a dedicated GPU laptop (the TensorBook), and physical workstations with 4 GPUs for more intensive training.

Lambda Labs sold their enterprise workstations and servers to customers like Amazon, Apple, Raytheon, and MIT, each one being pre-configured with TensorFlow, Pytorch, and Caffe to make it quicker to get started. The main customers of Lambda Labs were researchers working on the emerging fields of image recognition and speech generation, alongside early natural language processing.

Today, rebranded as Lambda (lambda.ai), the company is primarily a cloud platform that lets developers get access to on-demand, enterprise-grade cloud GPUs for AI development. The scarcity of GPU compute has been one of the key themes in the AI boom of 2023, with massive cloud providers like AWS struggling to meet demand due to chip shortages upstream at TSMC—creating an opportunity for new cloud providers like CoreWeave and Lambda that have beneficial relationships with Nvidia.

Lambda's cloud product has expanded well beyond single-node instances into a full-spectrum compute portfolio spanning multiple GPU generations. The company offers H100 SXM instances and production-scale "1-Click Clusters" on HGX H100 and HGX B200—including multi-node clusters offering on-demand access to 16–512 B200 GPUs connected via NVIDIA Quantum-2 InfiniBand. Lambda also operates the industry's first hydrogen-powered, production-grade NVIDIA GB300 NVL72 systems at ECL's Mountain View campus, where each system receives 142 kW of compute power; Lambda has since doubled its footprint at that facility to 100%.

Lambda holds NVIDIA Exemplar Cloud status—one of the first GPU cloud providers to receive NVIDIA's performance validation for mission-critical AI training workloads at scale, with H100 infrastructure benchmarked within 5% of NVIDIA's published baselines—and the company reports serving 200,000+ AI developers and teams.

Looking further ahead, Lambda has announced bare-metal instances on NVIDIA Vera Rubin NVL72, a production-scale GB300 NVL72 supercluster with Quantum-X Photonics networking (where Lambda is leading one of the largest deployments of NVIDIA Quantum-X InfiniBand co-packaged optics switches, covering 10,000+ GB300 GPUs), early adoption of NVIDIA Vera CPUs, and early NVIDIA BlueField-4 STX adoption—with Vera Rubin production availability and STX-based platform deployments planned for the second half of 2026.

Business Model

Lambda Labs, like CoreWeave in the GPUs-on-demand space and like other cloud providers such as AWS, operates on a model where it rents out computing resources (such as GPU power) to businesses and developers over the cloud.

One key differentiation with Lambda Labs is that where CoreWeave is cloud-only, Lambda Labs also has on-premises options. Companies that want to get more compute for less money, ensure data security, and work on big datasets can purchase Lambda workstations without using them as a cloud provider.

Lambda has raised over $1.5B in Series E funding led by TWG Global, with participation from Thomas Tull's US Innovative Technology Fund (USIT) and existing investors, with proceeds directed toward gigawatt-scale AI factories and supercomputing infrastructure for hyperscalers, enterprises, and frontier labs. To fund next-generation NVIDIA AI accelerator infrastructure and expanded data center capacity, Lambda secured a $1B syndicated senior secured credit facility—upsized from an original $275M facility established in August 2025—arranged by J.P. Morgan. The breadth of Lambda's customer base is underscored by the fact that Nvidia itself has agreed to rent back chips from Lambda—roughly 10,000 chips under a $1.3B, four-year agreement plus a separate $200M deal for 8,000 more chips.

Lambda has also shifted from leasing third-party data center space toward owning and building its own facilities, a move that reduces long-term per-unit costs and increases control over infrastructure quality and capacity planning. That owned footprint includes a leased 21MW presence at Prime Data Centers' AI-ready campus in Vernon, California.

To support this infrastructure scale-up, Lambda has materially strengthened its leadership bench. Co-founder Stephen Balaban serves as CTO and co-founder Michael Balaban as CPO, while Michel Combes—former CEO of Sprint and SoftBank Group International—leads the company as CEO, focused on capital formation, infrastructure expansion, and building operating systems to support growth. John Donovan, former AT&T Communications CEO, serves as chairman of the board. The broader leadership team includes Leonard Speiser as COO overseeing supercomputer deployment and operations, Jerry Hunter—former AWS infrastructure leader and Snap COO—as Vice Chairman, and Charles Fisher as CFO leading capital strategy and financial operations.

Expenses

Lambda Labs incurs a significant upfront cost when purchasing GPUs and setting up data centers. However, these GPUs have a useful life of several years, during which Lambda Labs can continually rent them out. The operational costs include electricity (GPUs are power-hungry), cooling (to prevent overheating), and staffing (for maintenance and customer support).

Improving the efficiency of data center operations (e.g., reducing electricity consumption, negotiating better rates for electricity, or improving cooling systems) can lower operational costs and thus improve margins.

Margin

The cost of a GPU for Lambda Labs includes the purchase price and the operational costs over its lifespan. The revenue from a GPU is the cumulative amount paid by customers to rent the GPU over time. Lambda Labs aims to maximize the utilization of each GPU to ensure that the revenue generated far exceeds the cost.

Margins are generally lowest on Lambda Labs's higher-end GPUs. For example, a high-end H100 PCIe card might cost Lambda Labs roughly $30,000. That GPU is then rented out at an average of $2.49 per hour. Assuming an 80% utilization rate, it would generate roughly $17,268 in revenue per year ($2.49/hour × 19 hours/day × 365 days/year).

However, cheaper GPUs like the A10 could generate much greater margins. At 80% utilization, an A10—which Lambda Labs could have bought for ~$3,500 a few years ago and is now rented out at $0.75 per hour—could generate $5,201 in revenue every year.

Competition

The market for GPU cloud services is highly competitive, with several key players, including major cloud providers like Amazon Web Services, Google Cloud and Azure as well as upstarts like Lambda Labs and Together AI, each offering unique advantages and targeting different segments of the AI and machine learning industry.

Big Cloud

The biggest long-term competition for Lambda Labs is likely to be the major three cloud providers: Google Cloud ($75B in revenue in 2023), Amazon Web Services ($80B in revenue in 2023) and Microsoft Azure ($26B in revenue in 2023).

With far greater revenue scale—vs. Lambda Labs's ~$250M in 2023—the big cloud platforms have the resources to invest both in acquiring GPUs and in developing their own silicon alternatives to Nvidia's GPUs.

However, between Lambda and the big cloud companies are also some "coopetitive" dynamics. In November 2025, Lambda announced a multibillion-dollar, multi-year agreement with Microsoft to deploy AI infrastructure powered by tens of thousands of NVIDIA GPUs, including GB300 NVL72 systems—positioning Lambda simultaneously as a competitor and a key supplier to one of its largest rivals.

CoreWeave

Like Lambda Labs, CoreWeave is a public cloud provider that purchases GPUs from Nvidia and rents them out to AI companies and companies building AI features.

Lambda Labs historically positioned itself as a better option for smaller companies and developers on less intensive workloads, offering Nvidia H100 PCIe GPUs at roughly $2.49 per hour compared to CoreWeave at $4.25 per hour. Lambda has since expanded its own portfolio to include HGX H100 and HGX B200 clusters—closing the product gap that previously left large-scale workloads to CoreWeave alone—and achieved NVIDIA Exemplar Cloud validation confirming performance within 5% of NVIDIA's published baselines, a credential that directly challenges CoreWeave's claim on enterprise training customers.

Together

Together is fundamentally a GPU reseller that rents GPUs from providers like Lambda Labs and CoreWeave, big cloud platforms like Google Cloud, and from other sources—academic institutions, crypto miners, other companies—and then rents those GPUs out to startups and AI companies, then bundling that in with software for training and fine-tuning open source AI models like Meta's Llama 2, Midjourney's Stable Diffusion, and its own RedPajama.

Sacra estimates that Together hit $10M in annual revenue run rate at the end of 2023, with 90% of that revenue coming from Forge, their bundled compute-and-training product that launched in June 2023. Forge promises A100 and H100 Nvidia server clusters at 20% of the cost of AWS.

TAM Expansion

Looking forward, the key dynamics in understanding Lambda Labs's durable advantage hinges on (1) the long-term state of the GPU industry, (2) the market for on-premises AI hardware, and (3) Lambda's own infrastructure build-out.

GPUs

At the root of the GPU shortage that has benefitted companies like Lambda Labs and CoreWeave to this point is a limitation at TSMC—Taiwan Semiconductor Manufacturing Company. The key shortage there is on chip-on-wafer-on-substrate (CoWoS) packaging capacity, which is used by all GPUs in the manufacturing process. TSMC expects the current shortage to last until around early 2026 and has a $2.9B packaging facility in the works, expected to be operational in 2027, that will further alleviate constraints.

The major cloud providers, as well as companies like Tesla, Meta and OpenAI, wanting to escape the dynamics of this shortage, have all begun or accelerated work on their own AI processors. That said, they're also dependent on TSMC to actually make their chips—and with Nvidia being one of TSMC's biggest and longest-term customers, Nvidia could still have an advantage on manufacturing, at least until shortages are completely alleviated.

On-premises

Lambda Labs, by offering on-premises workstations that can allow companies to train their AI models locally rather than in the cloud, is well-positioned to benefit from the rising tide of these kinds of workloads.

There are a few key reasons why companies would move their training on-premise vs. keeping it in the cloud, particularly as models improve and the need for compute ramps up further: cost, security, and big data.

Cost: Companies with stable, long-term compute needs might find that investing in on-premises infrastructure leads to lower total cost of ownership compared to continually renting cloud resources. This is particularly relevant for organizations that can efficiently manage and utilize their hardware over time.

Security: On-premises solutions can also provide businesses with control over their data, helping them meet stringent compliance requirements and data sovereignty laws by ensuring that sensitive data does not leave the company's premises.

Big data: When training AI models, it's often necessary to use large datasets of images, text, or audio files—so large that transferring them to a cloud-based environment for training can result in significant data transfer costs, also known as egress fees. These fees can add up quickly, especially when transferring large datasets, making it expensive to train AI models in the cloud. By keeping the data on-premises, organizations can avoid the costs associated with transferring data to and from the cloud.

Infrastructure Build-Out

Lambda is aggressively expanding its owned and long-term-leased data center footprint as a structural shift away from reliance on third-party space—a move that improves per-unit economics and gives Lambda greater control over capacity planning as it scales its frontier GPU roadmap. Lambda has articulated a goal of reaching 3 gigawatts of AI compute under management by 2030, targeting deployment of more than 1 million GPUs by the end of the decade.

Several major capacity additions are coming online in 2026. A 24MW AI Factory in Kansas City—expected to house more than 10,000 NVIDIA Blackwell Ultra GPUs with the potential to scale past 100MW—is slated to come online in early 2026, with the entire initial deployment dedicated to a single customer under a multi-year agreement. A partnership with EdgeConneX adds more than 30MW of additional capacity across a 23MW single-tenant Chicago site and an Atlanta facility, both targeted for 2026 readiness; the Chicago site uses hybrid direct-to-chip liquid cooling and air cooling. Lambda also holds a leased 21MW presence at Prime Data Centers' AI-ready LAX01 campus in Vernon, California, which spans 242,000 square feet across six data halls and is deployed for large-scale NVIDIA AI training and inference workloads. Together, these sites meaningfully increase Lambda's owned and long-term-leased capacity ahead of anticipated demand growth.

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 May 10th, 2026