CoreWeave

Visit Website

CoreWeave provides public cloud infrastructure for GPU compute.

#ai

Details

HEADQUARTERS

Roseland, NJ

CEO

Michael Intrator

REVENUE

$465,000,000

2023

VALUATION

$7,000,000,000

2023

GROWTH RATE (Y/Y)

1,760%

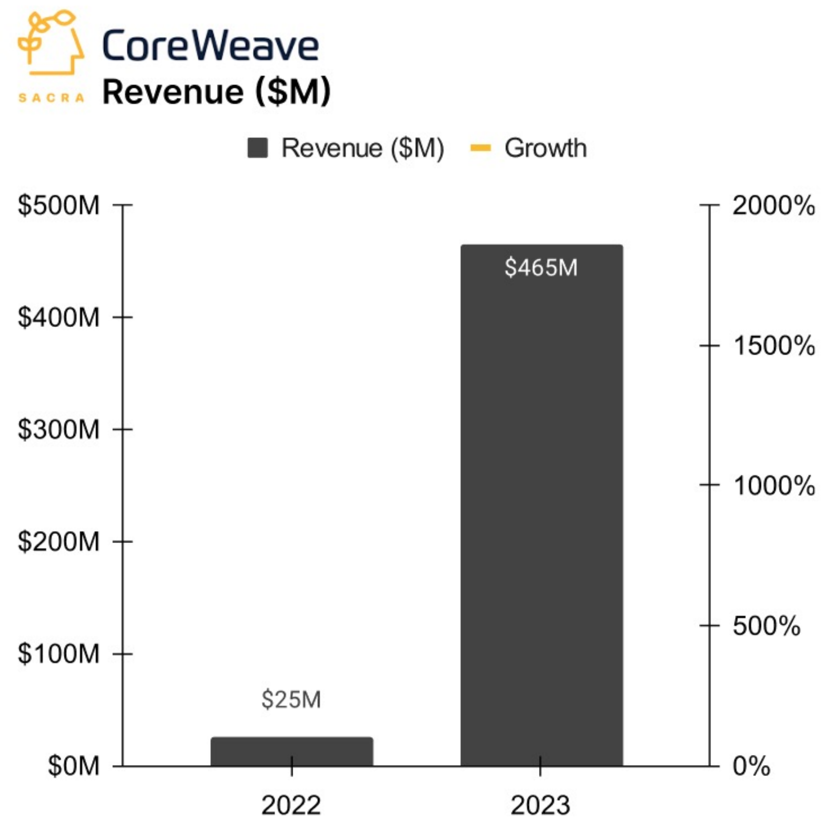
2023

FUNDING

$3,500,000,000

2023

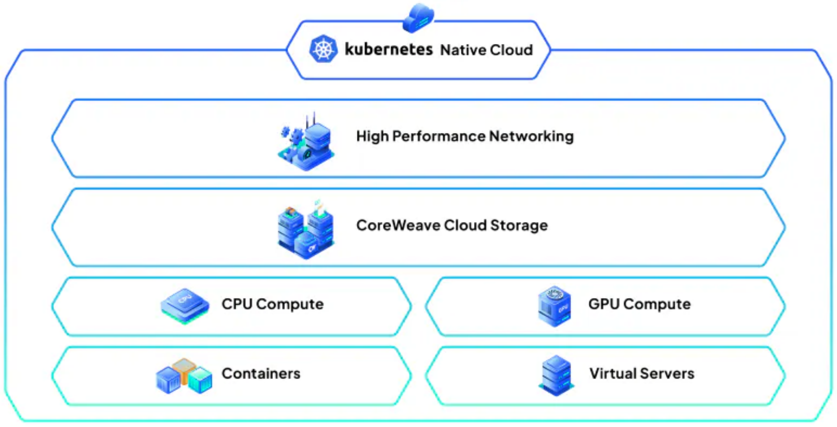
Revenue



Sacra estimates that CoreWeave hit $465M in revenue in 2023, up 1760% or about 19x from $25M in 2022, off the rapid acceleration of demand for GPU compute from cloud providers, LLM companies, and every app looking to integrate generative AI features. CoreWeave projects $2.3B of revenue in 2024, with signed cloud contracts in excess of $7B through 2026, up from $5B in early 2023.

A significant portion of that $7B in contracts came from Microsoft, which agreed on a multi-year deal for CoreWeave to supply it with GPU compute amid rising demand from Azure cloud customers that Microsoft has not been able to meet.

Product



CoreWeave was founded in 2017 as Atlantic Crypto, an Ethereum mining company that bought Nvidia graphics processing units (GPUs) both to mine its own crypto and rent out GPU servers to other crypto miners. In early 2019, Atlantic Crypto changed its name to CoreWeave and pivoted to providing GPUs-on-demand for generalized computing purposes rather than focusing on crypto.

Through this period, CoreWeave built infrastructure for delivering that GPU compute across seven global facilities—positioning them well for the flood of demand for GPU compute that arrived in 2022 with the generative AI boom.

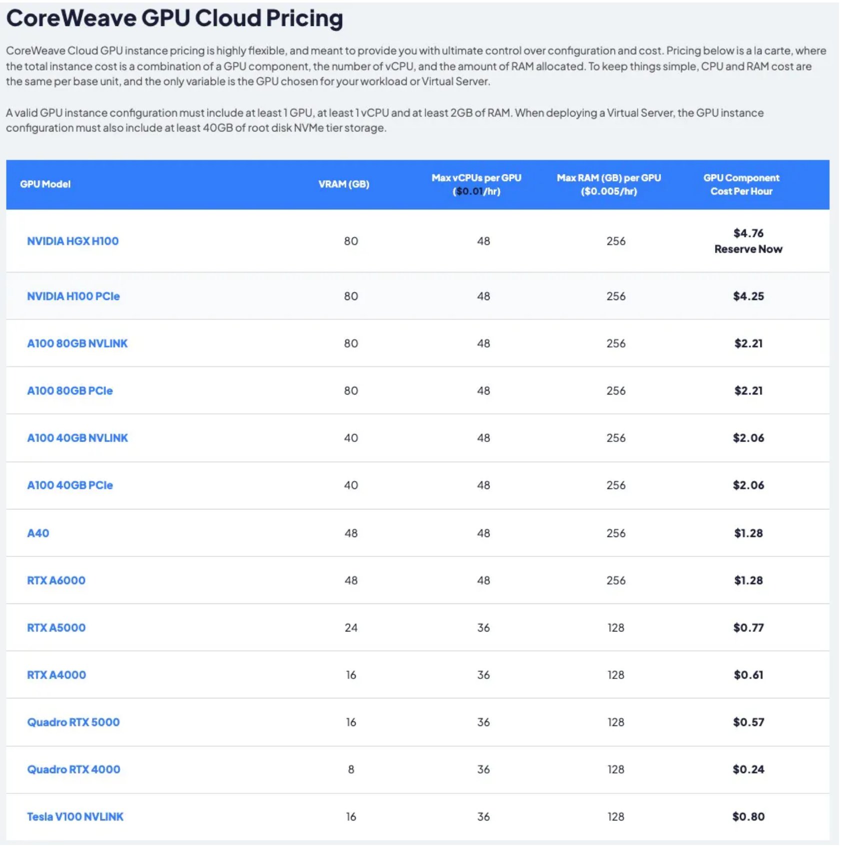
Today, CoreWeave is fundamentally a GPU-first cloud platform that lets developers and business access compute remotely the same way they would with Amazon Web Services or Azure.

What differentiates CoreWeave is their far greater availability of the high-end GPUs that are designed for training and running large, complex, AI workloads. With 45,000 GPUS, CoreWeave is the largest private provider of GPUs in North America. In early 2023, CoreWeave was one of the first cloud providers to offer access to the new Nvidia H100 Tensor Core GPUs on its platform.

While AWS and Amazon customers report resources shortages on their cloud platforms, CoreWeave's most favored nation relationship with Nvidia has allowed them to both scale faster to meet demand and offer higher-powered GPUs.

For the AI text adventure game AI Dungeon, which is based on GPT-2, serving 1.6M users of their game drove response times up and cost too much to continue running the product on AWS's Cortex GPU compute platform. Switching to Tesla V100 GPUs delivered via CoreWeave's cloud cut AI Dungeon's response time down by 50%.

Business Model



CoreWeave GPU Cloud Pricing

CoreWeave Cloud GPU instance pricing is highly flexible, and meant to provide you with ultimate control over configuration and cost. Pricing below is a la carte, where the total instance cost is a combination of a GPU component, the number of vCPU, and the amount of RAM allocated. To keep things simple, CPU and RAM cost are the same per base unit, and the only variable is the GPU chosen for your workload or Virtual Server.

A valid GPU instance configuration must include at least 1 GPU, at least 1 vCPU and at least 2GB of RAM. When deploying a Virtual Server, the GPU instance configuration must also include at least 40GB of root disk NVMe tier storage.

GPU Model	VRAM (GB)	Max vCPUs per GPU ($0.01/hr)	Max RAM (GB) per GPU ($0.005/hr)	GPU Component Cost Per Hour
NVIDIA HGX H100	80	48	256	**$4.76** **Reserve Now**
NVIDIA H100 PCIe	80	48	256	**$4.25**
A100 80GB NVLINK	80	48	256	**$2.21**
A100 80GB PCIe	80	48	256	**$2.21**
A100 40GB NVLINK	40	48	256	**$2.06**
A100 40GB PCIe	40	48	256	**$2.06**
A40	48	48	256	**$1.28**
RTX A6000	48	48	256	**$1.28**
RTX A5000	24	36	128	**$0.77**
RTX A4000	16	36	128	**$0.61**
Quadro RTX 5000	16	36	128	**$0.57**
Quadro RTX 4000	8	36	128	**$0.24**
Tesla V100 NVLINK	16	36	128	**$0.80**

CoreWeave, like other cloud providers, operates on a model where it rents out computing resources (such as GPU power) to businesses and developers.

CoreWeave's ~85% gross margins come from the difference between the cost of maintaining these resources (including the initial investment in hardware, ongoing electricity, cooling, maintenance, and support staff costs) and the revenue generated from customers paying to use these resources.

Customers pay CoreWeave for the computing power they use, typically on a per-hour basis. This payment model is attractive to customers because it allows for flexible scaling of resources based on demand, and they only pay for what they use. CoreWeave sets the rental price based on market demand, the specific GPU model (newer models with better performance command higher prices), and the operational costs to ensure a profitable margin.

CoreWeave, like AWS, has an expansion motion in offering different kind of services on top of the basic product of GPU compute. So far, CoreWeave has added on specialized solutions for data storage, networking, CPU compute, each priced on a similar pay-as-you-go basis.

Expenses

CoreWeave 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 CoreWeave 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.

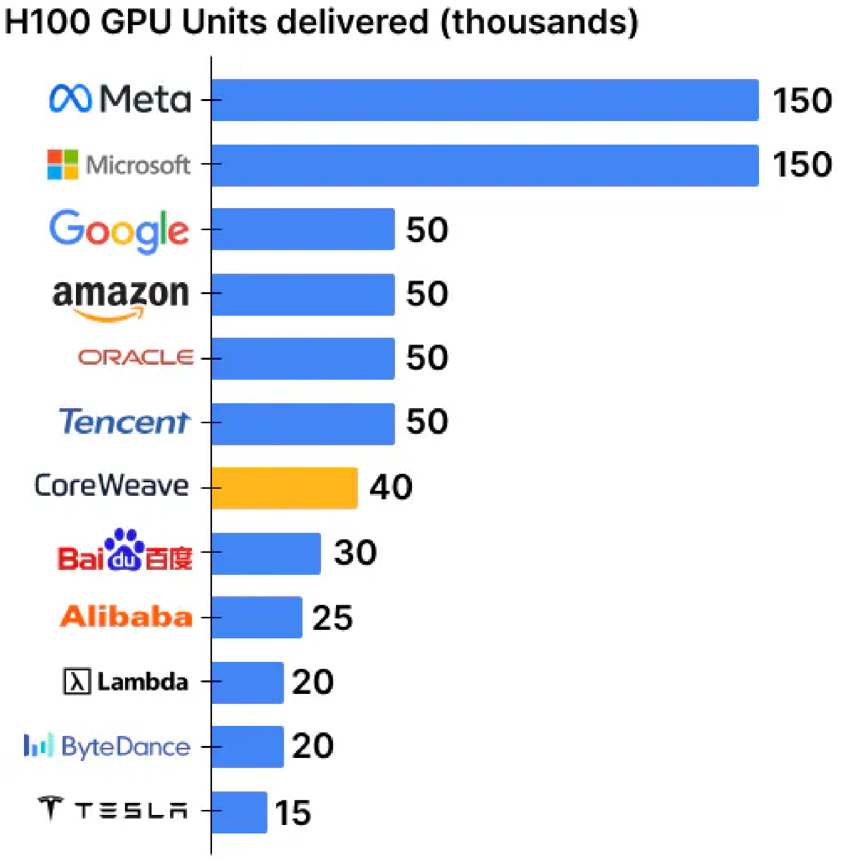
Margin

The cost of a GPU for CoreWeave 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. CoreWeave aims to maximize the utilization of each GPU to ensure that the revenue generated far exceeds the cost.

Margins are generally lowest on CoreWeave's higher-end GPUs. For example, A high-end H100 PCIe card might cost CoreWeave roughly $30,000. That GPU is then rented out at an average of $4.25 per hour. Assuming an 80% utilization rate, it would generate roughly $29,473 in revenue per year ($1/hour * 12 hours/day * 365 days/year)—roughly break-even assuming that they don't hit 100% utilization.

However, cheaper GPUs like the A40, which CoreWeave could have bought in bulk in 2021 before the generative AI boom, could generate much greater margins. At 80% utilization, an A40—which had a sticker price of $4,500 three years ago and is now rented out by CoreWeave at $1.278 per hour—could generate $8,877 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 CoreWeave 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. CoreWeave's ~$465M 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.

So far, CoreWeave has been able to outmatch the biggest cloud providers on access to GPUs because they've enjoyed preferential treatment from Nvidia, which has allocated GPUs away from Amazon, Google and Microsoft and towards CoreWeave. Notably, CoreWeave is the only major cloud provider customer of Nvidia's that is not developing its own AI chips to try to compete with Nvidia, making it a good customer for Nvidia to support.

Lambda Labs

Like CoreWeave, Lambda Labs is a public cloud provider that purchases GPUs from Nvidia and rents them out to AI companies and companies building AI features. Also like CoreWeave, Lambda Labs has received generous allocations of Nvidia GPUs and was in talks with Nvidia for investment in 2023, but as of February 2024, that deal hasn't happened.

Lambda Labs is generally positioning itself as a better option for smaller companies and developers working on less intensive computational tasks, offering Nvidia H100 PCIe GPUs at a price of roughly $2.49 per hour, compared to CoreWeave at $4.25 per hour. On the other hand, Lambda Labs does not offer access to the more powerful HGX H100—$27.92 per hour for a group of 8 at CoreWeave—which is designed for maximum efficiency in large-scale AI workloads.

Lambda Labs generated about $20M in revenue in 2020 and was projecting $250M in 2023 and $600M in 2024 as of July last year. Lambda Labs is backed by Thomas Tull's US Innovative Technology fund, B Capital, SK Telecom, Crescent Cove, Mercato Partners, 1517 Fund, Bloomberg Beta, and Gradient Ventures.

Together

Together is fundamentally a GPU reseller that rents GPUs from 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

To date, CoreWeave's rapidly accelerating growth has been driven by high demand for GPUs and compute, combined with low supply. CoreWeave's favored partner status with Nvidia has allowed them to offer better availability than the major cloud platforms while also undercutting them on price.

Looking forward, the key dynamics in understanding CoreWeave's durable advantage hinges on **(1)** the long-term state of the GPU industry, and **(2)** CoreWeave's ability to build a differentiated AI compute platform.

GPUs

At the root of Nvidia's GPU shortage 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. Currently, TSMC expects that the current shortage will last until about March 2026. TSMC recently announced a plan to build a $2.9B packaging facility that will be operational in 2027, further alleviating shortages.

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.

Tech

Last year, CoreWeave reported record-breaking LLM benchmark results using Nvidia HGX H100 instances—CoreWeave's platform, composed of one of the biggest clusters of HGX GPUs in the world, came in 29x faster than the next-fastest competitor.

That's one sign that CoreWeave could compete with the major cloud providers even if the present GPU shortages come to an end. CoreWeave's infrastructure has been designed from the ground-up to serve GPU compute at scale—since 2017 when the company was working on Ethereum mining as Atlantic Crypto.

Over the last few years, AI workloads have generally been increasing in size and complexity, creating a scaling revenue opportunity for companies like CoreWeave that specialize in serving the customers with the biggest compute needs.

If CoreWeave can continue to outmatch cloud rivals on performance, then they could protect their moat even in the absence of the supply-and-demand dynamics that have powered their growth to $465M in revenue so far.

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.

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 Feb 16th, 2024