

Lambda Labs Company Profile



Contents

[Cloud 3](#_Toc192268438)

[Lambda 1-Click clusters 3](#_Toc192268439)

[Features 3](#_Toc192268440)

[Pricing Structure 3](#_Toc192268441)

[Use Cases 3](#_Toc192268442)

[Summary 3](#_Toc192268443)

[On-Demand Cloud 3](#_Toc192268444)

[Features 3](#_Toc192268445)

[Pricing Structure 4](#_Toc192268446)

[Use Cases 4](#_Toc192268447)

[Summary 4](#_Toc192268448)

[Private Cloud 4](#_Toc192268449)

[Features 4](#_Toc192268450)

[Pricing Structure 5](#_Toc192268451)

[Use Cases 5](#_Toc192268452)

[Summary 5](#_Toc192268453)

[Lambda Chat 5](#_Toc192268454)

[Features 5](#_Toc192268455)

[Use Cases 6](#_Toc192268456)

[Summary 6](#_Toc192268457)

[Data Center 6](#_Toc192268458)

[NVIDIA DGX Systems 6](#_Toc192268459)

[Features 6](#_Toc192268460)

[Deployment Options 7](#_Toc192268461)

[Pricing Structure 7](#_Toc192268462)

[Use Cases 8](#_Toc192268463)

[Summary 8](#_Toc192268464)

[Scalar Servers 8](#_Toc192268465)

[Features 8](#_Toc192268466)

[Pricing Structure 8](#_Toc192268467)

[Use Cases 9](#_Toc192268468)

[Summary 9](#_Toc192268469)

[Desktop Workstations 9](#_Toc192268470)

[Lambda Vector Pro GPU Workstation 9](#_Toc192268471)

[Features 9](#_Toc192268472)

[Performance 10](#_Toc192268473)

[Pricing Structure 10](#_Toc192268474)

[Use Cases 10](#_Toc192268475)

[Summary 10](#_Toc192268476)

[Vector GPU Desktop 10](#_Toc192268477)

[Features 10](#_Toc192268478)

[Pricing Structure 11](#_Toc192268479)

[Use Cases 11](#_Toc192268480)

[Summary 11](#_Toc192268481)

[Lambda Vector One 11](#_Toc192268482)

[Features 11](#_Toc192268483)

[Pricing Structure 12](#_Toc192268484)

[Use Cases 12](#_Toc192268485)

[Summary 12](#_Toc192268486)

[Partnerships 12](#_Toc192268487)

[Company Value 12](#_Toc192268488)

[Pricing Strategy 13](#_Toc192268489)

# Cloud

## Lambda 1-Click clusters

**1-Click Clusters**, is a groundbreaking service that provides on-demand access to GPU clusters for AI model training.

### Features

* On-Demand Access - This service allows users to spin up multi-node GPU clusters featuring 16 to 512 NVIDIA H100 Tensor Core GPUs, all interconnected with NVIDIA Quantum-2 InfiniBand. This setup is designed for rapid training and fine-tuning of large AI models, enabling teams to leverage substantial computational power without lengthy procurement processes.
* Flexible Reservations - Unlike traditional GPU cloud services that often require long-term contracts, 1-Click Clusters can be rented for a minimum of just two weeks. This flexibility is particularly beneficial for smaller companies and startups that need temporary access to high-performance computing resources.
* Pre-Installed Software - Each cluster comes with essential AI frameworks pre-installed, including PyTorch, TensorFlow, NVIDIA CUDA, and NVIDIA cuDNN, allowing users to start their projects immediately without additional setup time.

### Pricing Structure

* The pricing for 1-Click Clusters is set at $4.49 per GPU per hour. There are no additional egress or ingress fees, and storage costs are $0.20 per GB per month24. This straightforward pricing model eliminates hidden costs, making it easier for teams to budget their projects.

### Use Cases

1-Click Clusters are particularly suited for -

* AI Startups: Companies looking to validate their models quickly without committing to expensive long-term contracts.
* Research Institutions: Teams needing substantial computational resources for short-term projects, such as drug discovery or multimodal AI development.

### Summary

Overall, Lambda's 1-Click Clusters represent a significant advancement in making powerful GPU resources more accessible and flexible for a wide range of users in the AI field.

## On-Demand Cloud

Lambda Labs offers an On-Demand Cloud service specifically designed for AI developers, providing access to powerful GPU resources for training and inference

### Features

* GPU Options - Users can access various NVIDIA GPUs, including the latest H100, A100, and A10 Tensor Core GPUs. This variety allows users to choose instances that best fit their workload and budget needs.
* Flexible Instances - Lambda provides multi-GPU instances, enabling configurations of 1x, 2x, 4x, and 8x NVIDIA GPUs. This flexibility is crucial for scaling workloads according to specific project requirements.
* User-Friendly API - The Lambda Cloud API allows developers to easily launch, terminate, and manage instances programmatically. This feature is particularly beneficial for automating workflows and integrating with existing development processes.
* Pre-Configured Environments - Instances come pre-installed with essential machine learning frameworks such as TensorFlow, PyTorch, and NVIDIA's CUDA and cuDNN. This setup minimizes the time needed for configuration, allowing users to focus on their projects immediately.
* One-Click Jupyter Access - Users can quickly connect to their GPU instances via Jupyter notebooks directly from their browsers, streamlining the development process.

### Pricing Structure

* On-demand instances are billed by the hour with no long-term commitments required. For example, NVIDIA H100 instances start at approximately $2.49 per hour, making it accessible for various budgets.

### Use Cases

Lambda's on-demand cloud is ideal for –

* AI Research Teams: Researchers needing temporary access to high-performance GPUs for model training and experimentation.
* Startups: Companies looking to develop AI solutions without the financial burden of long-term contracts or infrastructure investments.

### Summary

Overall, Lambda's On-Demand Cloud provides a robust platform tailored for AI developers, combining flexibility, powerful hardware options, and ease of use in a competitive pricing model.

## Private Cloud

Lambda Labs provides a Private Cloud solution that is tailored for organizations needing dedicated, high-performance GPU resources for AI and deep learning applications.

### Features

* Dedicated GPU Clusters - The Private Cloud allows users to reserve clusters of NVIDIA GPUs, including the latest H100 and H200 models. These clusters can be scaled from 64 to 32,000 GPUs, providing flexibility based on computational needs.
* High-Speed Networking - Clusters are equipped with NVIDIA Quantum-2 InfiniBand, which enhances data transfer speeds and reduces latency, essential for large-scale AI training tasks.
* Custom Configurations - Organizations can customize their cloud setup to fit specific workloads, whether they require a few GPUs for development or thousands for extensive model training.
* Integrated Software Stack - The private cloud comes with a fully integrated software stack, including popular machine learning frameworks such as TensorFlow, PyTorch, and NVIDIA's CUDA and cuDNN. This setup ensures that users can start their projects without extensive configuration.
* Storage Options - Lambda has introduced persistent storage solutions that allow users to maintain data across sessions. This feature is particularly useful for long-term projects where data continuity is crucial.

### Pricing Structure

Lambda offers competitive pricing for its Private Cloud services:

* The cost per GPU hour varies based on the type of GPU and the length of the contract. For instance, pricing for H100 GPUs can start as low as $1.84 per hour with longer-term contracts.

|  |  |  |
| --- | --- | --- |
| Contract Length (years) | Down Payment | Price Per GPU Hour |
| 3 | 100% | $1.84 |
| 2 | 100% | $2.09 |
| 1 | 100% | $2.29 |

### Use Cases

Lambda's Private Cloud is ideal for -

* Large Enterprises - Companies requiring substantial computational power for AI research and development without the overhead of managing physical hardware.
* Research Institutions - Academic and research organizations needing dedicated resources for experimental AI projects.

### Summary

Lambda Labs' Private Cloud offers a robust solution for organizations looking to leverage powerful GPU resources in a flexible, scalable environment tailored to AI workloads.

## Lambda Chat

Lambda Labs has developed a product known as Lambda Chat, which is designed to facilitate real-time collaboration and communication among users within the Lambda Cloud ecosystem.

### Features

* Real-Time Collaboration - Lambda Chat enables users to communicate instantly while working on projects, allowing for seamless collaboration among team members.
* Integration with Lambda Cloud - The chat feature is integrated into the Lambda Cloud platform, making it easy for users to discuss projects, share insights, and troubleshoot issues directly within their working environment.
* Team Management - Users can invite team members to join their Lambda Cloud accounts, allowing for centralized management of resources and billing. This feature streamlines administrative tasks and enhances team productivity.
* Multiplayer Jupyter Notebooks - The chat functionality supports collaborative features in Jupyter Notebooks, enabling multiple users to work on the same notebook simultaneously while discussing their work in real time.
* Resource Sharing - Teams can share resources such as instances, storage, and files directly through the chat interface, enhancing collaborative efforts on AI projects.

### Use Cases

* AI Development Teams - Ideal for teams working on AI models who need to communicate effectively while managing complex projects.
* Educational Settings - Useful for academic groups or classes where students can collaborate on projects and share resources.

### Summary

Overall, Lambda Chat enhances the collaborative experience within the Lambda Cloud environment, making it easier for teams to work together efficiently on AI development tasks.

# Data Center

## NVIDIA DGX Systems

Lambda Labs offers a range of NVIDIA DGX Systems designed for high-performance AI and deep learning tasks. These systems provide robust hardware capabilities and seamless integration for AI workflows.

### Features

* High-Performance Hardware -
  + DGX B200: Features 8 NVIDIA B200 Tensor Core GPUs, optimized for deep learning tasks.
  + DGX H100: Equipped with 8 NVIDIA H100 Tensor Core GPUs, providing exceptional performance for training and inference.
  + Both systems utilize NVLink and NVSwitch technology for high-speed GPU-to-GPU communication, enhancing overall throughput.
* Scalability -
  + The DGX systems can scale from single units to large clusters interconnected via NVIDIA Quantum-2 InfiniBand, allowing organizations to expand their computing resources as needed.
  + The DGX SuperPOD architecture supports scaling from two to thousands of interconnected DGX systems, making it suitable for extensive AI applications.
* Integrated Software Stack -
  + Each DGX system comes with the NVIDIA AI Enterprise software suite, which includes tools for managing AI workflows, cluster management, and optimized libraries for accelerated computing.
  + The systems are designed to run complex AI models efficiently, significantly reducing time to insight.
* Lifecycle Management Services -
  + Lambda provides lifecycle management services to ensure that customers can upgrade their hardware seamlessly as new NVIDIA architectures are released, maintaining the cutting edge of AI technology.

### Deployment Options

* Lambda-Hosted Clusters - Organizations can choose to have their DGX systems hosted in Lambda's secure data centres, which offer high-bandwidth internet access and VPN connectivity.
* On-Premises Deployment - Alternatively, customers can deploy DGX systems in their own data centres, with installation support from Lambda engineers.

### Pricing Structure

* NVIDIA DGX B200 –
  + Configuration - 10U system with 8x NVIDIA B200 Tensor Core GPUs.
  + Pricing - Specific pricing details for the B200 are generally available upon request from Lambda Labs, as it may vary based on customization and deployment options.
* NVIDIA DGX H100 -
  + Configuration - 8U system with 8x NVIDIA H100 Tensor Core GPUs.
  + Pricing - Detailed pricing for the H100 configuration is typically provided on inquiry, reflecting the advanced capabilities of the H100 GPUs.
* NVIDIA DGX SuperPOD -
  + Scaling Options - This system allows scaling from two to thousands of interconnected DGX systems, optimized for extensive AI workloads.
  + Pricing - Pricing for SuperPOD configurations also depends on specific requirements and is generally discussed directly with Lambda's sales team.
* Hourly Rental Options -
  + For cloud-based access to NVIDIA H100 GPUs, Lambda offers pricing starting at approximately $2.49 per hour for on-demand use. This flexible pricing model allows users to access high-performance computing resources without long-term commitments.
* Reserved Instances -
  + For longer commitments, Lambda provides reserved instance pricing that can significantly reduce costs. For example, prices can start as low as $1.84 per GPU per hour for a three-year contract with a full upfront payment.

| **System Type** | **Configuration** | **Price (Approx.)** |
| --- | --- | --- |
| NVIDIA DGX B200 | 10U system with 8x B200 GPUs | Contact for pricing |
| NVIDIA DGX H100 | 8U system with 8x H100 GPUs | Contact for pricing |
| NVIDIA SuperPOD | Scalable from two to thousands of units | Contact for pricing |
| On-Demand Access | NVIDIA H100 | Starting at $2.49/hr |
| Reserved Pricing | Long-term contracts (3 years) | Starting at $1.84/hr |

### Use Cases

NVIDIA DGX Systems are widely used across various industries:

* Healthcare - For applications like medical imaging and drug discovery.
* Finance - Supporting algorithmic trading and risk management.
* Automotive - Enhancing capabilities in autonomous driving and predictive maintenance.

### Summary

Lambda Labs' NVIDIA DGX Systems provide a powerful solution for organizations looking to leverage advanced AI capabilities. With high-performance hardware, scalable architecture, integrated software solutions, and flexible deployment options, these systems are well-suited for both research and production environments in AI development.

## Scalar Servers

Lambda Labs offers Scalar Servers, which are high-performance GPU servers designed for deep learning and AI workloads.

### Features

* GPU Options - Scalar Servers can be configured with up to 8 NVIDIA H100 NVL Tensor Core GPUs, providing substantial computational power for intensive AI tasks.
* High Performance - Each server can support 192 CPU cores and 384 threads. Up to 8192 GB of memory is available, along with 253 TB of NVMe SSD storage, ensuring fast data access and processing capabilities.
* Scalability - These servers can easily scale from a single unit to larger clusters, allowing organizations to expand their compute resources as needed. This is particularly useful for distributed training and hyperparameter optimization.
* Integrated Software Stack - Scalar Servers come pre-installed with the Lambda Stack, which includes popular machine learning frameworks such as PyTorch, TensorFlow, CUDA, and cuDNN. This setup minimizes configuration time and ensures compatibility.
* Networking and Storage - The servers feature high-performance parallel file systems optimized for machine learning, along with networking capabilities that support GPUDirect RDMA and GPUDirect Storage for efficient data transfer.
* Premium Support - Lambda provides extensive support options, including a warranty with advanced parts replacement, live technical assistance from machine learning engineers, and ongoing software updates.

### Pricing Structure

* On-Demand Pricing -
  + NVIDIA H100 GPUs: Pricing starts around $1.89 per hour for the use of H100 GPUs in the private cloud setup. This is competitive compared to other cloud providers, making it an attractive option for organizations needing high-performance computing for AI workloads.
  + NVIDIA A100 GPUs: Available for approximately $1.10 per hour, which is significantly lower than similar offerings from other cloud providers.
* Reserved Pricing -
  + For longer-term commitments, Lambda offers reserved pricing, which can further reduce costs. Specific pricing for reserved instances typically requires contacting Lambda sales for a quote based on the desired configuration and commitment length.
* Additional Costs -
  + Depending on the configuration, there may be additional costs associated with storage, memory, and networking capabilities.

### Use Cases

Lambda Scalar Servers are ideal for:

* AI Research Teams - Organizations conducting extensive research requiring powerful GPU resources.
* Enterprise Applications - Businesses needing scalable solutions for AI model training and inference.

### Summary

Overall, Lambda Scalar Servers are engineered to deliver exceptional performance and flexibility for deep learning applications, making them suitable for both research institutions and commercial enterprises looking to leverage advanced AI technologies.

# Desktop Workstations

## Lambda Vector Pro GPU Workstation

The Lambda Vector Pro GPU Workstation is designed for AI developers and machine learning engineers, offering powerful performance and advanced features tailored for deep learning tasks.

### Features

* GPU Options –
  + Supports up to 2 dual-slot NVIDIA GPUs, including:
  + NVIDIA RTX 4500 Ada Generation: 24 GB of GDDR6 memory.
  + NVIDIA RTX 5000 Ada Generation: 32 GB of GDDR6 memory.
  + Higher-end options like the NVIDIA RTX A6000 with 48 GB of memory are also available for more demanding applications.
* Processor Options -
  + Equipped with powerful AMD Ryzen Threadripper CPUs. Options include the Threadripper PRO 5995WX (64 cores) and other models up to 96 cores, providing substantial multi-threaded performance.
* Memory and Storage -
  + Can be configured with up to 1 TB of RAM, ensuring ample memory for large datasets and complex models.
  + Supports high-speed storage options, including NVMe SSDs, for quick data access.
* Pre-Installed Software -
  + Comes with the Lambda Stack, which includes popular machine learning frameworks like PyTorch and TensorFlow, along with NVIDIA drivers. This setup minimizes configuration time and ensures compatibility with the latest software updates.
* Cooling and Design -
  + The workstation is designed for quiet operation while maintaining effective cooling, essential for long training sessions.

### Performance

* The Lambda Vector Pro is optimized for tasks such as image generation, model training, and handling complex AI workloads. Users have reported significant performance improvements compared to previous setups, making it suitable for both research and production environments.

### Pricing Structure

* While specific pricing details can vary based on configuration, a typical Lambda Vector Pro workstation starts around $7,000 to $12,000, depending on the selected GPU and additional options.

### Use Cases

The Lambda Vector Pro is ideal for -

* AI Research Teams: Those needing a robust workstation for developing and training machine learning models.
* Small to Medium Enterprises: Companies looking to leverage AI capabilities without investing in large server infrastructures.

### Summary

Overall, the Lambda Vector Pro GPU Workstation provides a powerful, flexible solution for developers and researchers focused on advancing their AI projects.

## Vector GPU Desktop

The Lambda Vector GPU Desktop is a powerful workstation designed specifically for AI and machine learning tasks.

### Features

* GPU Options:
  + Supports up to 2 dual-slot PCIe GPUs.
  + NVIDIA RTX 4500 Ada Generation -24 GB of GDDR6 memory, 7,680 CUDA Cores, 240 Tensor Cores
  + NVIDIA RTX 5000 Ada Generation - 32 GB of GDDR6 memory, 12,800 CUDA Cores, 400 Tensor Cores
* Processor Options:
  + Equipped with high-performance AMD Ryzen Threadripper CPUs, including Threadripper PRO 7975WX (32-Core), Threadripper PRO 7985WX (64-Core), Threadripper PRO 7995WX (96-Core).
* Memory and Storage - Configurable with up to 1 TB of RAM, allowing for extensive data handling and model training. Supports NVMe SSDs for fast storage solutions.
* Zero Configuration Required - The workstation comes pre-installed with the Lambda Stack, which includes popular machine learning frameworks like TensorFlow and PyTorch. This setup eliminates the need for manual driver installations and configurations, allowing users to focus on their research.
* Optimized for AI Workloads - The Vector Desktop is designed for speed, value, and quiet operation, making it suitable for both development and production environments.

### Pricing Structure

* While specific pricing may vary based on configuration, the Lambda Vector GPU Desktop typically starts at around $4,000 to $8,000, depending on the selected components and GPUs. Higher configurations with more powerful GPUs or additional memory can increase the price significantly.

### Use Cases

The Lambda Vector GPU Desktop is ideal for:

* AI Developers: Those who require a robust workstation for developing and training machine learning models.
* Research Institutions: Labs needing powerful computing resources without the complexity of server setups.

### Summary

Overall, the Lambda Vector GPU Desktop provides a flexible and powerful solution tailored for AI and machine learning tasks, making it an excellent choice for professionals in the field.

## Lambda Vector One

The Lambda Vector One is a new single-GPU desktop PC designed for demanding AI and machine learning tasks.

### Features

* GPU - Equipped with 1x NVIDIA GeForce RTX 4090, which has 24 GB of liquid-cooled GDDR6 memory. This powerful GPU is capable of handling complex computations required for deep learning tasks.
* Processor - Powered by the AMD Ryzen™ 9 7950X, featuring 16 cores and 32 threads. This CPU provides significant processing power, making it suitable for multitasking and running heavy applications.
* Memory - Configurable with either 64 GB or 128 GB of DDR5 RAM, allowing for efficient handling of large datasets and models.
* Storage - The system can support an operating system drive of up to 3.84 TB M.2 (NVMe), along with additional storage options for data, accommodating up to three additional 3.84 TB M.2 NVMe drives.
* Networking - Includes a 10Gb Ethernet interface, providing high-speed network connectivity essential for data-intensive applications.
* Design and Cooling - The Vector One features a sleek design with an advanced liquid cooling system for both the CPU and GPU, ensuring optimal performance while maintaining a quiet operation (only emits about 39 dB SPL under load).
* Pre-installed Software - Comes with the Lambda Stack, which includes necessary frameworks like TensorFlow and PyTorch, enabling users to start their AI projects immediately without additional setup.

### Pricing Structure

The Lambda Vector One is competitively priced at around $4,400 to $5,500, making it an attractive option for those looking for a powerful yet compact desktop solution tailored for AI and machine learning tasks.

### Use Cases

The Lambda Vector One is ideal for -

* AI Researchers and Developers - Those who need a robust workstation to train neural networks and run complex models.
* Educational Institutions - Suitable for labs and classrooms focused on AI education and research.

### Summary

Overall, the Lambda Vector One provides a high-performance solution in a compact form factor, specifically designed to meet the needs of AI and machine learning professionals.

# Partnerships

Lambda Labs has established several strategic partnerships to enhance its AI and deep learning offerings.

Quantiphi - This collaboration aims to provide tailored AI solutions across various industries, including banking, healthcare, and telecommunications. Quantiphi's AI capabilities combined with Lambda's high-speed computing infrastructure facilitate the training of Large Language Models (LLMs) efficiently.

Weights & Biases - This strategic partnership integrates Weights & Biases' MLOps platform with Lambda's NVIDIA-accelerated systems. It aims to provide enterprises with the necessary infrastructure and tools for developing machine learning algorithms, particularly in computer vision and natural language processing.

NVIDIA - Lambda has been recognized as an NVIDIA Partner Network Solution Integration Partner of the Year for multiple years. This partnership allows Lambda to leverage NVIDIA’s technology to deliver advanced GPU cloud services and infrastructure for AI workloads37.

VAST Data - Lambda selected VAST Data’s technology as the storage backbone for its offerings, enhancing data management capabilities essential for AI applications4.

SK Telecom - This recent partnership aims to expand Lambda's AI Cloud services in South Korea, enabling local enterprises and research labs to access GPU cloud services through SK Telecom's infrastructure

# Company Value

Lambda Labs has achieved a valuation of $1.5 billion following its recent funding round, where it raised $320 million in Series C financing. This investment was led by the US Innovative Technology Fund, with participation from various notable investors, including B Capital and T. Rowe Price Associates, among others.

The company has experienced significant growth, with revenues increasing from $30 million in 2022 to over $500 million in 2023, and projections of nearly $600 million for 2024.

# Pricing Strategy

Tiered Pricing Models - Lambda offers tiered pricing based on the number of GPUs and the level of service required. This allows customers to choose a plan that fits their budget and needs, encouraging upgrades as their requirements grow.

Usage-Based Pricing - The company has adopted a usage-based pricing model for its cloud services, where customers pay based on actual GPU usage. This model is attractive to businesses with fluctuating workloads, as it allows them to scale costs with their resource consumption.

Subscription Discounts - Lambda provides discounts for long-term commitments and subscriptions, incentivizing customers to commit to longer contracts. This approach helps stabilize revenue streams and improves customer retention.

Bundled Services - By offering bundled services that combine hardware and software solutions (like workstations with pre-installed AI frameworks), Lambda encourages customers to purchase more comprehensive packages at a discounted rate compared to buying components separately.

Freemium Options - Implementing a freemium model for certain services can attract new users who may later convert to paid plans as they require more features or higher performance.

Custom Configurations - Lambda allows for custom configurations in its workstations and servers, enabling customers to tailor their purchases according to specific needs. This flexibility can lead to higher customer satisfaction and increased sales.

Enhanced Support Packages - Offering premium support packages for an additional fee can generate extra revenue while providing customers with peace of mind regarding technical assistance.

Value-Based Pricing - Lambda may employ value-based pricing strategies, where prices reflect the perceived value delivered to customers rather than just the cost of production. This approach can maximize revenue from high-value clients who benefit significantly from Lambda's offerings.