Article October 31, 2024 • 4 minute read

How much is an Nvidia A100?

Margaret Shen
Head of Business Operations

Nvidia A100 GPUs were launched in 2020, and while they are no longer the most bleeding edge GPUs on the market, they still offer great performance for training and deploying LLMs and diffusion models.

A100 Configurations and Pricing

The Nvidia A100 price varies based on configuration. The two primary factors influencing the price are GPU memory (40GB vs 80GB) and form factor (PCIe vs SXM).

40GB vs. 80GB

This refers to the amount of VRAM on the GPU, and determines how large of a model you can run/fine-tune. Please refer to our articles on how much VRAM you need to fine-tune or serve transformers models for an estimate on how much bang for your buck you would get.

PCIe vs SXM

Generally, the SXM version is more expensive than the PCIe version.

That's because SXM GPUs are directly socketed onto the motherboard, enabling more direct and high-bandwidth connections. If your primary goal is to effectively run larger transformer models, the SXM version is the better choice due to its enhanced performance capabilities.

Direct Purchase Price from Nvidia

The current market prices for the Nvidia A100 vary from $8,000 to $10,000 for the 40GB PCIe model to $18,000 to $20,000 for a 80GB SXM model.

Alternatives to Direct Purchase: A100s in the Cloud

Traditional Cloud Platforms - AWS, GCP, OCI

For most companies, the high upfront costs of setting up their own data centers lead them to seek cloud platforms instead. Traditionally, companies have made GPU reservations lasting 1-3 years, with Amazon, Google, and Oracle dominating this market.

AWS (Amazon) and GCP (Google) also provide flexible purchase models for A100s, including spot and on-demand options. The on-demand model allows users to pay for A100s by the second, but this flexibility comes at a higher per-hour price compared to reservations. Spot pricing enables users to utilize unused capacity and is also billed by the second, although it does not guarantee that workloads won't be pre-empted, making it generally cheaper than on-demand options.

Below are comparison tables of current A100 GPU per-hour list prices across these platforms.

GPU type	Purchase model	AWS*	GCP*	OCI
A100 40GB SXM	1 year reservation	$2.52	$2.31	$3.05
	3 year reservation	$1.56	$1.29	n/a
	On-demand	$4.10	$3.67	n/a
	Spot	$1.15	$1.17	n/a
A100 80GB SXM	1 year reservation	$3.15	n/a	$4.00
	3 year reservation	$1.95	n/a	n/a
	On-demand	$5.12	$5.12	n/a
	Spot	n/a	$1.57	n/a

**Prices based on us-east-1 for AWS and us-central1 for GCP. Note that prices will vary based on region.*

AWS also offers a Savings Plan that provides discounts from on-demand pricing in exchange for a 1 or 3-year commitment to a baseline $/hr usage. This plan is more flexible than a reservation but typically comes with a higher price.

Finding information on A100s for these cloud providers can be challenging due to differing naming conventions. Below is a guide on the instance type names associated with A100s for each provider:

GPU type	AWS	GCP	OCI
A100 40GB SXM	p4d.24xlarge	a2-highgpu-* a2-megagpu-*	bm-gpu4.8
A100 80GB SXM	p4de.24xlarge	a2-ultragpu-*	bm.gpu.a100-v2.8

Note that AWS and OCI only offer A100s in configurations of 8 GPUs, while GCP provides configurations ranging from 1 to 16.

Serverless Compute Startups

Many companies are exploring alternatives to the hyperscalers for several reasons:

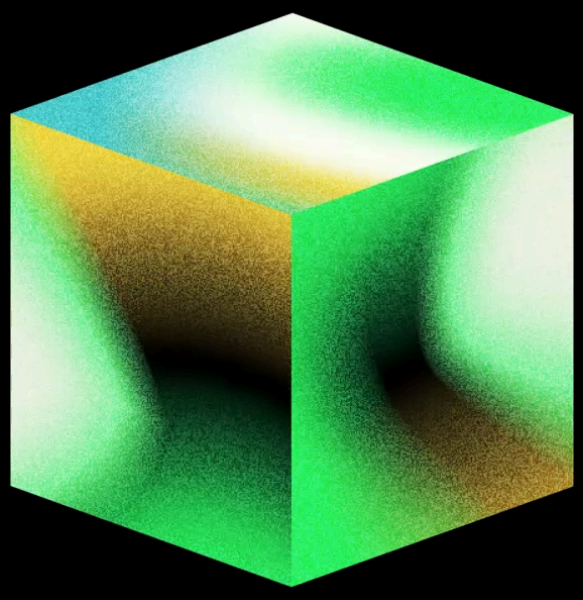
- Inflexibility of GPU configurations
- Lack of availability without making a big commitment
- Slow and manual provisioning of resources
- Overhead of configuring and managing infrastructure

Emerging GPU platforms are addressing these challenges by offering a more flexible model for accessing and scaling resources. By adopting a serverless approach, these platforms can spin up GPUs for users only when needed, simplifying the management of the underlying infrastructure. Below is a comparison of A100 per-hour prices for the most popular serverless GPU providers.

GPU Type	Modal	Lambda Labs	Runpod	Baseten
A100 40GB SXM	$2.78	$1.29	n/a	n/a
A100 80GB SXM	$3.40	$1.79	$2.72	$6.144

While these prices may be slightly higher than the spot or reservation prices of the hyperscalers, they do not reflect the full picture. Serverless options can quickly autoscale GPUs and charge based on usage, leading to significantly higher utilization and lower overall costs for workloads with unpredictable volume.

Interested in exploring the Nvidia A100 price options? On Modal, you can deploy a function with an A100 attached in just a few minutes. Try it out!



Ship your first app in minutes.

Get Started

$30 / month free compute

Use Cases

- Language Model Inference
- Image, Video & 3D
- Audio Processing
- Fine-Tuning
- Job Queues & Batch Processing
- Sandboxing Code
- Computational Biology

Popular Examples

- Serve LLM APIs with vLLM
- Create Custom Art of Your Pet
- Analyze Parquet files from S3 with DuckDB
- Run hundreds of LoRAs from one app
- Replace your CEO with an LLM

Resources

- Documentation
- Pricing
- Slack Community
- Articles
- GPU Glossary
- Model Library

Company

- About
- Blog
- Careers
- Privacy Policy
- Security & Privacy
- Terms