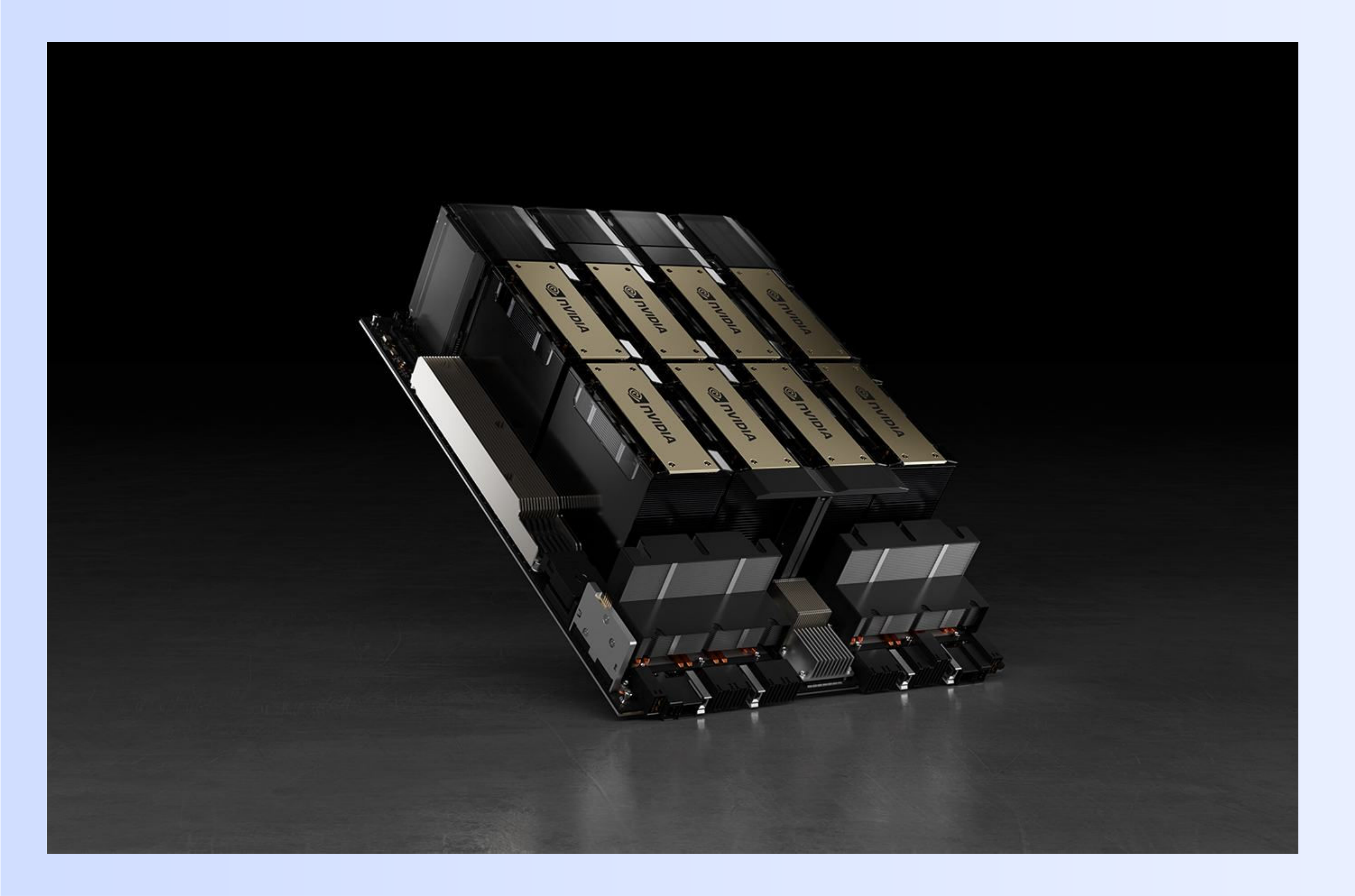
Unlocking Savings: Why Buying NVIDIA H100 GPUs Beat AWS Rental Costs



The speed at which AI, machine learning and high-performance computing are advancing is leading to difficult decisions for businesses in a huge range of different industries.

Some of the key things that companies are having to consider in light of these developments, are the benefits of cloud versus on-premises solutions.

Those using NVIDIA H100 GPUs often find themselves faced with a dilemma. Because choosing to rent from AWS or investing in colocation can have a significant impact on a company's bottom line. But how can companies make savings here, without impacting performance?

In this article, we'll discuss the cost implications and breakeven point of decisions like these, and we'll talk about one solution that's rapidly growing in popularity: the hybrid IT approach.

The Cost of Purchasing NVIDIA H100 GPUs

If a company opts to purchase an NVIDIA H100 GPU upfront, they can expect to pay in the region of $30,000. We'll also need to factor in costs associated with colocation here though, such as power, cooling, and space. These expenses set a typical business back $3,600 per year.

The Cost of Renting NVIDIA H100 GPUs on AWS

First things first, let’s tackle the math. If a company opts to rent NVIDIA H100 GPUs on AWS, we can calculate the cost at approximately $48,741.60 per year.

AWS offers the powerful EC2 p5.48xlarge instances, equipped with 8 NVIDIA H100 GPUs, at approximately $44.50 per hour (NVIDIA Blog) (Amazon Web Services, Inc.). We can break this cost down to around $5.56 per hour for a single H100 GPU.

So, for continuous, round-the-clock usage, the annual rental cost per H100 GPU would be $5.56 per hour × 24 hours/day × 365 days/year, which brings us to our figure of $48,741.60 per year.

Working Out The Total Annual Cost

When you purchase an NVIDIA H100 GPU, you spread the purchase cost over its useful life, which is typically five years. The depreciation per year is $3,000, with $15,000 recovery value. Add the annual colocation cost and you’re looking at $3,000 + $3,600 = $6,600 per year

Working Out The Total Annual Cost

When you purchase an NVIDIA H100 GPU, you spread the purchase cost over its useful life, which is typically five years. The depreciation per year is $3,000, with $15,000 recovery value. Add the annual colocation cost and you’re looking at $3,000 + $3,600 = $6,600 per year

The Breakeven Point

To determine how long it takes for the purchase to become more cost-effective than renting, we need to compare the annual costs:

- **Annual Rental Cost on AWS: $48,741.60**
- **Annual Cost of Purchased GPU: $6,600**

The breakeven period is calculated by dividing the purchase price by the difference in annual costs:
Breakeven Period = $30,000 / ($48,741.60 − $6,600) ≈ 0.71

So, it would take approximately 8.5 months to break even.

How to Get The Best of Both Worlds: Introducing the Hybrid IT Approach

While the financial benefits of purchasing and colocating NVIDIA H100 GPUs are clear, there is a way to access even greater advantages, and that’s through the hybrid IT approach.

A hybrid IT approach is a great strategy for so many organizations. And here’s why:

Flexibility and Scalability

A hybrid approach combines the strengths of both on-premises and cloud solutions. You can leverage on-premises GPUs for consistent, high-demand workloads while utilizing cloud resources to handle peak loads or temporary projects. This ensures you're not over-provisioning hardware that sits idle during off-peak times.

Cost Management

By purchasing GPUs for your baseline needs and renting additional capacity during spikes, you can optimize costs. This hybrid model prevents the need for significant capital expenditures on hardware that may only be needed intermittently.

Performance and Latency

On-premises GPUs can provide better performance and lower latency for certain applications, especially those requiring frequent data transfers or low-latency processing. For applications where latency is less critical, cloud GPUs can be an effective supplement.

Business Continuity and Disaster Recovery

A hybrid approach enhances business continuity and disaster recovery capabilities. On-premises GPUs can continue to operate even if there's a cloud service disruption, while cloud resources can be rapidly deployed if there's an issue with on-premises infrastructure.

Data Security and Compliance

Some workloads and data may need to remain on-premises due to security, compliance, or regulatory requirements. A hybrid model allows sensitive data to be processed locally while taking advantage of the cloud's scalability for less sensitive tasks.

Unlock savings for your business

For organizations with continuous, high-volume computational needs, purchasing NVIDIA H100 GPUs and colocating them in a data center can be significantly more cost-effective than renting from AWS.
Renting offers flexibility and eliminates upfront costs of course, but the long-term savings that can be made by purchasing and colocating start to become evident in less than a year.

A hybrid IT approach offers a compelling combination of cost-efficiency, scalability, performance and flexibility. By strategically blending on-premises and cloud resources, businesses can optimize their GPU utilization, manage costs effectively, and ensure robust performance and security.