NEBIUS

October 2025

Whitepaper

The economics of AI clusters

This whitepaper explores the cost dynamics of AI infrastructure for foundation model training and highlights how Nebius AI Cloud helps to maximize return on AI investments.

Introduction

Building a foundation AI model is expensive, and AI infrastructure takes the biggest part of the capital investment required. Regardless of the deployment approach, on-premises or cloud, this kind of infrastructure is extremely complex and sophisticated. The complexity and computational potential that was previously accessed by a limited number of academics is now at the disposal of a growing number of AI enthusiasts, ML practitioners and business leaders.

Yet growing availability doesn't mean affordability. These complex systems are expensive, and their complexity could become a black hole for your investments if you don't know exactly how your money will be spent.

In this whitepaper, we break down AI infrastructure costs and explore which parts could be substantially decreased. We also cover how Nebius, a purpose-built AI cloud, cuts training costs through continuous optimization of each piece of the stack.

AI workloads require more to run

Compared to traditional cloud computing tasks, AI workloads require more intensive calculations, more data and more time to deliver tangible results. To perform these calculations, you need a more sophisticated infrastructure with more compute power, larger storage volumes and higher data transfer throughput.

AI workloads are distinctive because they predominantly run on Graphics Processing Units (GPUs) or other accelerators such as Tensor Processing Units (TPUs) developed by Google and Trainium chips developed by AWS. Compared to CPUs, these chips demonstrate significantly higher performance, but consume more power. For effective operation, they need complementary systems like high-speed bandwidth, high-throughput storage and extra cooling.

AI workloads, including model training and model inference, rely heavily on parallel computing. This computational approach splits a machine learning job across multiple GPUs. Each compute node does its own part of the job, contributing to overall progress. By running jobs in parallel and scaling the GPU cluster size, ML engineers have an opportunity to accelerate the production pipeline and achieve better results, faster.

Time is money

AI infrastructure directly impacts your ability to scale and grow. With greater capacity, you are able to strengthen your competitive position efficiently and reach your goals faster. However, sophisticated hardware-software systems cost an enormous amount of money. This is the largest category of capital expenditure in every AI-focused startup. Since building, operating and owning a supercomputer does not fit on the balance sheet of most companies, there is an alternative approach. ML teams often leverage GPU-as-a-Service from multiple cloud service providers to obtain GPU compute at an hourly rate.

To assess the level of these investments, let's take a quick look at some calculations, based on existing industry pricing rates. If a GPU hour costs $2 and you order 8 GPUs for an hour, you will need to pay 2 * 8 * 1 = $16. If you order 3,000 GPUs, one hour will cost $6,000. **If you use 3,000 GPUs for a full day, it will cost you $144,000.** With this data, you can see that this level of investment requires steady financial backing.

For example, if you need this cluster to pre-train a foundation model, which can take weeks or months, you should reserve several million dollars for AI compute only. And if you manage to complete your job a couple of days earlier, **you will save hundreds of thousands of dollars for your next AI project.**

The cost of AI training

Let's walk through an example. Imagine you are training an LLM and you need to estimate how much time it will take to complete a distributed job on a 3000-GPU cluster. In theory, it could be accomplished in 3 days. To make this example more illustrative, we assume that we have ordered this cluster from Cloud X — a fictional GPU cloud service provider that delivers a baseline level of quality according to existing industry standards.

"Three days in theory" means 100% utilization of computational power of GPUs, no interruptions, code issues or dataset inconsistencies. This 3-day period means that everything goes smoothly and predictably. But in reality, interruptions occur. Infrastructure can fail. It needs maintenance and troubleshooting. And due to physical limitations, it's impossible to achieve 100% utilization of GPU chips.

According to NVIDIA DGX Cloud benchmarks, the GPU utilization (Model FLOPS Utilization, MFU) is about 45-55% of the theoretical hardware limit, depending on model you need to train[1]. So, if we take 50% of GPU utilization, the actual training time of the given job will take 6 days or 144 hours:

```
Training time (Cloud X) = Training time in theory * Hours a day * 100 / Cloud X MFU =
                          3 * 24 * 100 / 50% = 144 hours
```

While this calculation is a helpful data point, it doesn't include the risk of job interruptions. In order to make this time-to-completion estimate closer to reality, we need to factor in potential downtime from accidental job interruptions, as well as the checkpointing process, a reliability measure ML practitioners use to save the progress of training jobs.

According to studies, job interruptions can occur every 9.8 hours for a 3,000-GPU cluster[2]. And for every job interruption, the cluster needs an average of 1 hour to detect the failure and restore its state[3]. By adding potential job interruptions into the equation, we add an extra 14.7 hours to our initial estimate:

```
Number of interruptions (Cloud X) = Training time / Mean Time Between Failure (MTBF) =
                                    144 / 9.8 = 14.7
```

```
Interruption time (Cloud X) = Number of interruptions * Mean Time To Recovery (MTTR) =
                              14.7 * 1 = 14.7 hours
```

Checkpoint frequency can vary widely with the use case, model architecture, dataset size and training methodology. For reference, we assume a 3-hour checkpointing interval[4]. While checkpointing time also depends on multiple factors, including model size, network and storage conditions, we assume that with local storage it takes about 3 to 5 minutes.

```
Checkpointing time (Cloud X) = Training time / Checkpoint interval * Checkpoint duration =
                               144 / 3 * 5 / 60 = 4 hours
```

Additionally, it adds at least 4 hours of checkpointing to the initial time estimate.

(1) Average MFU numbers from NVIDIA DGX Cloud benchmarks. GitHub.
(2) Projected MTBF for 3,000 GPUs, based on data from Sivathanu, M., Zhao, Y., & Reddi, V. J. (2024). Revisiting Reliability in Large-Scale Machine Learning Research Clusters. Arxiv.
(3) Estimated AI cluster recovery time (MTTR) in a non-automated environment, based on reported customer experience in the industry.
(4) Following the More-efficient recovery from failures during large-ML-model training paper by Amazon Science.

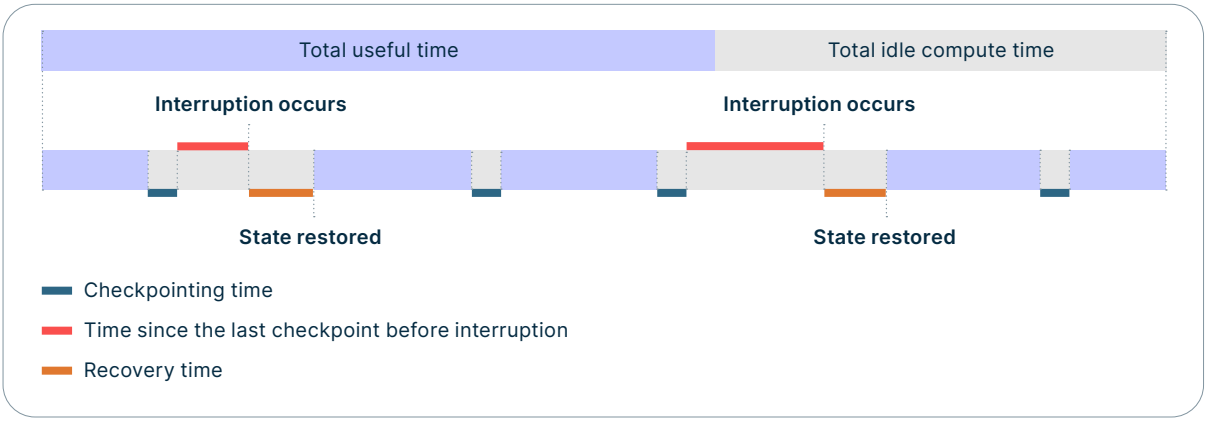

Figure 1. The structure of idle compute time that occurs during training, due to checkpoints and job interruptions.

Every time a failure occurs, training must roll back to the last saved checkpoint, which creates wasted time that does not contribute to model progress (Figure 1). Assuming job interruptions are uniformly distributed across the checkpoint interval, the expected rollback time is half of the checkpoint interval. Thus, the total wasted time can be estimated as:

```
Rollback time (Cloud X) = Checkpoint interval / 2 * Number of interruptions =
                          3 / 2 * 14.7 = 22.1 hours
```

Let's also estimate that, for cluster setup and maintenance, you need roughly 5 hours.

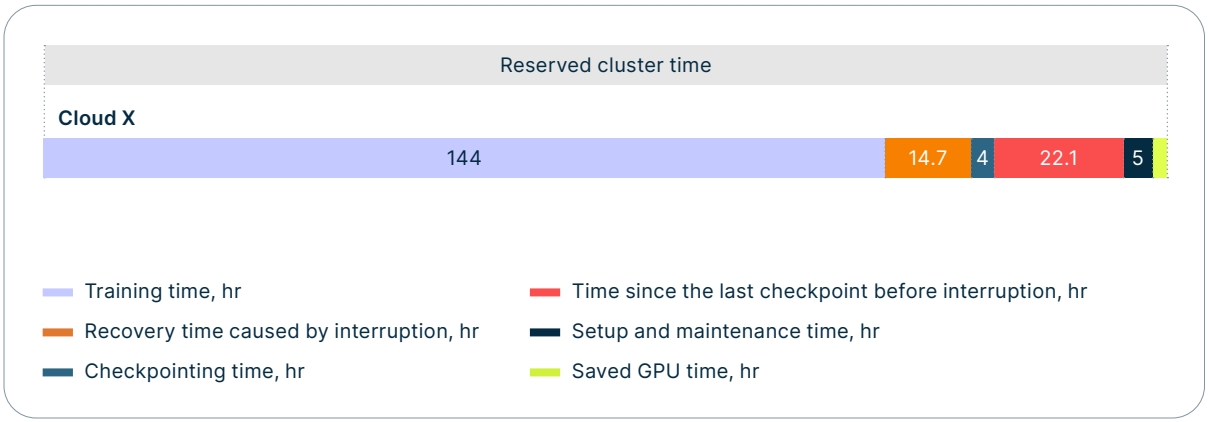

Figure 2. The structure of reserved compute time from Cloud X.

Let's sum up all of the considerations above, to estimate the total amount of time you will need to perform the LLM training job on a 3,000 GPU cluster and how much it will likely cost:

```
144 + 14.7 + 4 + 22.1 + 5 = 189.8 hours
```

If you reserved 8 days (192 hours) of this cluster, you would save 2.2 hours of compute time that you can reuse for other ML jobs or experiments.

How Nebius achieves cost-efficiency with reliable infrastructure

In this section, we examine how Nebius provides fast and reliable infrastructure for large-scale training, to maximize cost-efficiency by ensuring a stable and resilient compute environment.

According to DGX Cloud Benchmarks and MLCommons® results, Nebius delivers about 100% of the industry performance benchmark for GPU utilization. Average GPU cloud providers deliver 95-97% of the industry benchmark performance. Given the example above, if Nebius provides 100% of potential performance compared to 97% by Cloud X, it will result in 51.55% MFU and reduce the actual training time to 139.7 hours:

```
MFU (Nebius) = MFU of Cloud X * 100 / Benchmark performance of Cloud X =
                          50% * 100 / 97 = 51.55%
```

```
Training time (Nebius) = Training time in theory * Hours a day * 100 / Nebius MFU =
                        3 * 24 * 100 / 51.55% = 139.7 hours
```

Nebius delivers on cluster reliability, demonstrating 33 hours of stable operation on average for a 3000-GPU cluster. The peak value of this performance is 56.6 hours, based on a customer's recent production training cluster. Since we have an automated cluster recovery mechanism, it only takes about 12 minutes to detect an issue and restore the cluster training state. This results in 50 minutes or 0.8 hours of total idle time caused by job interruptions:

```
Number of interruptions (Nebius) = Training time / Mean Time Between Failure (MTBF) =
                                  139.7 / 33 = 4.2
```

```
Interruption time (Nebius) = Number of interruptions * Mean Time To Recovery (MTTR) =
                             4.2 * 12 / 60 = 0.8 hours
```

We follow the established recommendations for checkpointing every 3 hours. Our high-performance networking and shared storage infrastructure enables these checkpoints to complete in as little as 3 minutes, minimizing system idle time:

```
Checkpointing time (Nebius) = Training time / Checkpoint interval * Checkpoint duration
                            = 139.7 / 3 * 3 / 60 = 2.3 hours
```

The last thing to consider for reliability is the time lost between the interruption and the last checkpoint. Since Nebius shows 3.3x times fewer interruptions than Cloud X (4.2 vs. 14.7), we can estimate the rollback time for our cluster as follows:

```
Rollback time (Cloud X) = Checkpoint interval / 2 * Number of interruptions =
                          3 / 2 * 4.2 = 6.3 hours
```

Setup and maintenance time are also decreased on Nebius. We deliver clusters fully ready to work, with preinstalled drivers and libraries, so setup time for our cluster is faster than the industry standard. Additionally, we provision GPU capacity after rigorous multi-stage checks. We also enable our customers with end-to-end monitoring to catch errors at early stages and reduce the number of maintenance events. So, for this category, we can expect about 3 hours at this scale.

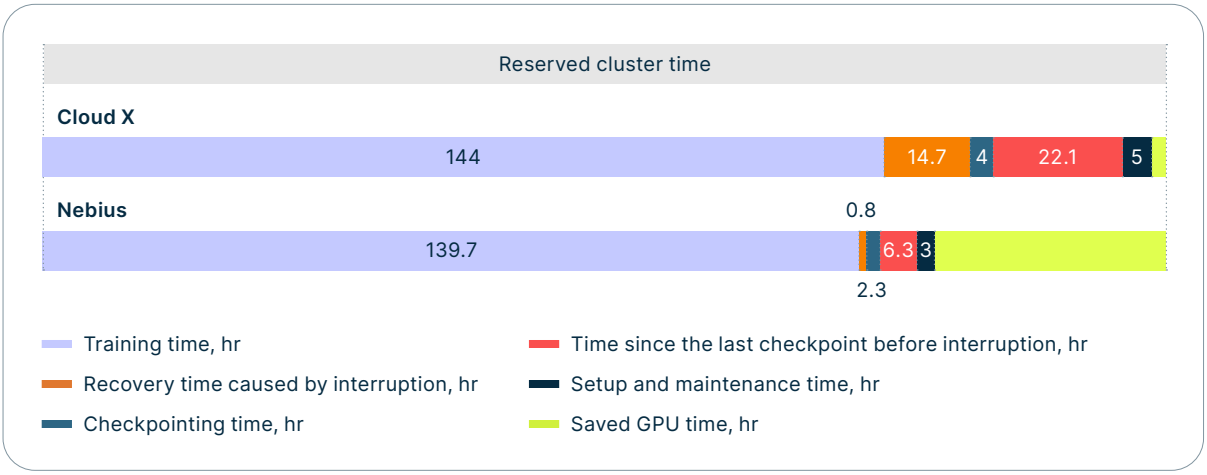

Figure 3. The structure of reserved compute time from Nebius

In total, completion of this training job on Nebius' cluster will take 152.4 hours:

```
139.7 + 0.8 + 2.3 + 6.3 + 3 = 152.1
```

With this hourly estimate, you save 39.9 hours compared to Cloud X, enabling you to get a head start on your next AI project.

From GPU pricing to Total Cost of Ownership

This example highlights the gap between a basic GPU cloud provider's infrastructure and the advanced infrastructure of Nebius. For a glimpse of how it impacts the TCO of AI infrastructure, we can augment this data with publicly available, on-demand pricing for NVIDIA HGX H200.

A GPU cloud provider with baseline features, like Cloud X, tends to have a lower price per GPU hour than hyperscalers or Nebius. This naturally leads to lower overall costs for reserved cluster capacity at Cloud X. However, if we look at the price you need to pay to get the job done, we'll see that these costs are relatively on par for both providers. At the same time, Nebius delivers a huge advantage in saved GPU time compared to the fictional competitor, Cloud X.

	Nebius	**Cloud X**
GPU hour, in USD	$3.50	$2.80
Reserved time, hours	192	192
Total training time, hours	152.1	189.8
Saved time, hours	39.9	2.2
Reserved cluster cost, in USD	$2,016,000	$1,612,800
Training job cost, in USD	**$1,597,050**	**$1,594,320**
Saved GPU cost, in USD	**$418,950**	**$18,480**

If we take NVIDIA HGX B200, which shows a smaller difference in on-demand pricing, we can see that with Nebius, you pay less for training jobs and save significantly by reducing compute time.

	Nebius	Cloud X
GPU hour, in USD	$5.50	$4.90
Reserved time, hours	192	192
Total training time, hours	152.1	189.8
Saved time, hours	39.9	2.2
Reserved cluster cost, in USD	$3,168,000	$2,736,000
Training job cost, in USD	**$2,509,650**	**$2,704,650**
Saved GPU cost, in USD	**$658,350**	**$31,350**

In this comparison, we leave storage costs out of the equation, implying that they are equal for both providers regardless of the quality and throughput parameters.

This illustrates how TCO works. While direct costs may be higher, the investment yields more training steps or model tokens compared to other cloud providers. To deliver this cost efficiency to Nebius customers, we continuously enhance our core reliability and performance features, introducing innovations across every layer of our vertically integrated AI cloud.

What else matters for cost-efficient AI at scale

Nebius was purpose-built as an AI infrastructure provider for modern generative AI workloads. Beyond delivering reliable, supercomputer-grade clusters for parallel model training and inference, we provide features that make AI adoption easier and more cost-effective.

Free expert support

Nebius offers white-glove customer support from world-class engineers, free of charge. When you are investing millions of dollars in model development or inference, direct access to AI subject matter experts is critical. This expertise not only reduces time-to-value, but also helps maximize the return from every hour of prepaid GPU compute.

Managed AI orchestration

Nebius delivers a fully managed orchestration environment that reduces the need for costly in-house DevOps expertise. For example, Managed Soperator, our solution for Slurm clusters, reduces manual labor by dozens of hours each week. Managed Soperator simplifies cluster configuration, maintenance and scaling, compared to a traditional Slurm-only approach.

No buffer capacity costs

Nebius does not charge customers for the buffer capacity required for rapid node recovery. By provisioning GPU compute through our cloud platform, we maintain a shared pool of spare nodes to cover automatic replacement. This can instantly reduce total cluster costs by 10–20%, compared to other vendors.

Let’s see how these benefits translate into direct savings for the customer:

	Nebius	Cloud X
White-glove customer support from high-class engineers	$0	Extra fee
On-site DevOps specialist to configure and maintain Slurm environment	$0	**$14,440** 192 hours of DevOps in the US, $75/hour
Pre-paid buffer capacity for quick node swap	$0	**$115,200** 10% of a 3,000-GPU cluster running 192 hours, $2 per GPU hour

At Nebius, our goal is to combine world-class infrastructure with economics that work in real life for our customers, ensuring every GPU hour delivers maximum value. If you are exploring how to scale your generative AI workloads more efficiently, we’d be glad to share our expertise and invite you to test the environment yourself. Get in touch with us to discuss how Nebius can optimize the performance and cost structure of your AI initiatives.