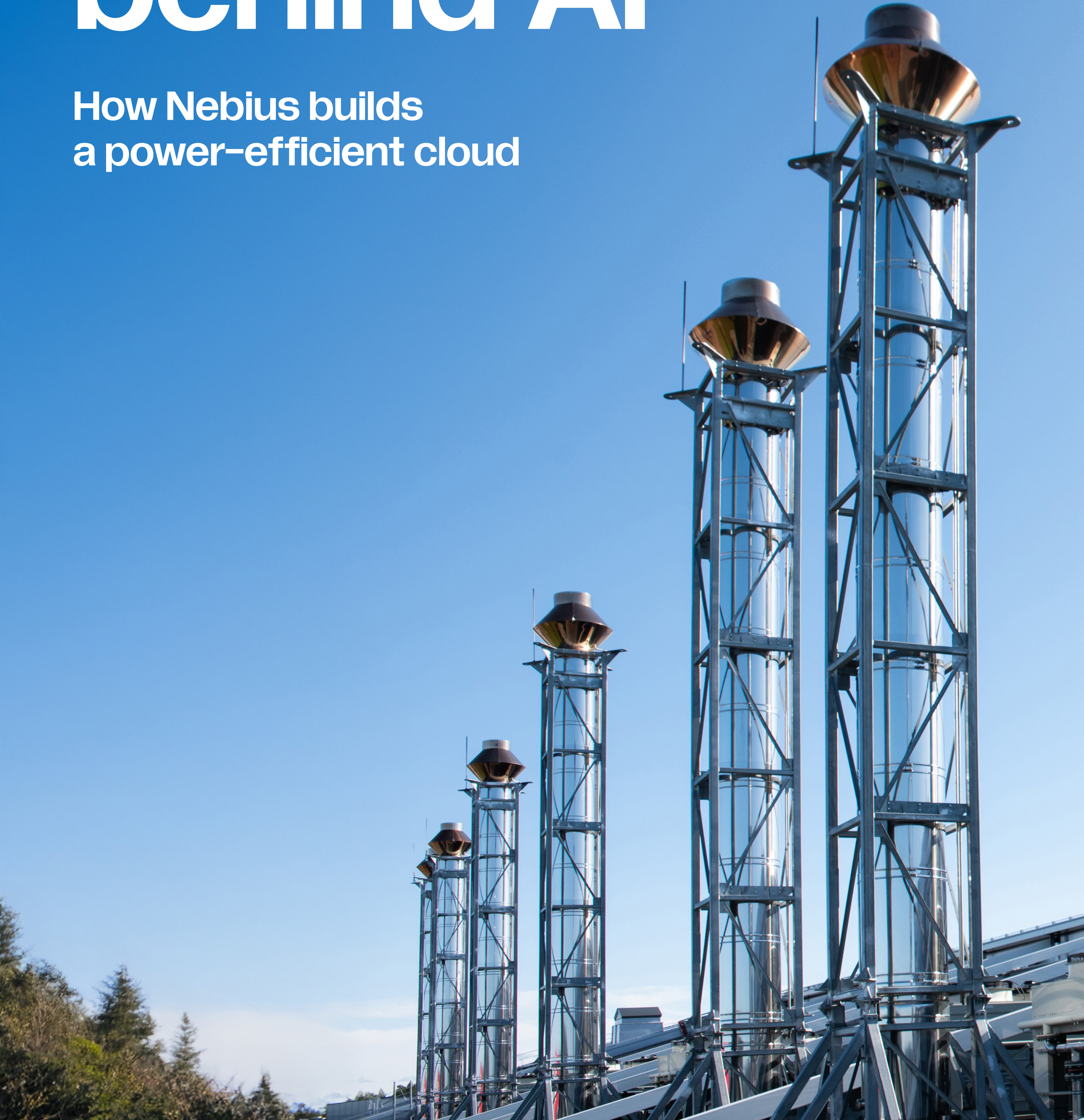
Whitepaper | January 2026

The energy behind AI

How Nebius builds a power-efficient cloud



The energy behind AI

AI workloads require a lot of energy, and this makes sustainability a key topic for any cloud provider. To put it simply: power goes in, and AI results — tokens trained or served — come out. Behind this process are four layers that define how efficient an AI cloud can be: the model layer, the cluster layer, the fabric layer and the data center layer. In this whitepaper, we explain how each layer works and how Nebius improves efficiency across the stack, from software engineering to hardware design and datacenter operations.

AI needs power. A lot of it.

When we type a question into a chatbot, the reply feels instant, but behind the scenes, a model runs thousands of mathematical operations across GPUs and other specialized hardware. All of this consumes electricity, which is why leaders in the field often remind us that even a simple “please” sent to a chatbot has a real energy cost. And this is only inference. Long before a model answers a single query, a significant amount of energy has already been spent on training, which can require thousands of servers running continuously for weeks or months.

The link between digital technology and energy use is not new. Data centers have been major power consumers for more than a decade. In 2010, they already used over 1% of global electricity[1], a number that grew to around 1.5% in 2024[2]. The absolute amount of energy use, however, kept growing, driven by continuous technological innovations, from smartphone apps to video streaming.

Now, AI represents the next major wave of demand. Electricity consumption from data centers has grown 12% per year over the last five years[3], in large part due to the rise of AI. Some projections show that in the US alone, electricity use could reach 325–580 TWh per year by 2028[4], which is roughly comparable to the annual electricity consumption of the UK[5] or Germany today.

The message is clear: AI is becoming a major new consumer of power, and how that power is managed will determine its sustainability.

[1] Jonathan G. Koomey, Growth in Data Center Electricity Use 2005 to 2010 (Oakland, CA: Analytics Press, Aug. 1, 2011).

[2] International Energy Agency, Energy and AI (Paris: IEA, April 10, 2025).

[3] International Energy Agency, Energy and AI: Energy Demand from AI (Paris: IEA, April 10, 2025).

[4] Berkeley Lab, Energy Analysis & Environment Impacts Division, 2024 United States Data Center Energy Usage Report (Berkeley, CA: Lawrence Berkeley National Laboratory, Dec. 20, 2024).

[5] Considering the United Kingdom’s annual electricity demand is approximately 320 terawatt-hours, while Germany’s annual electricity consumption is approximately 460 terawatt-hours. Sources: U.K. Department for Energy Security and Net Zero; Bundesnetzagentur (Germany).

With great power comes great responsibility

The Spider-Man quote captures a reality in AI: powerful technologies bring major opportunities but also tangible risks and a significant environmental impact. And while the long-term AI implications are still a part of a broader debate, its growing energy footprint is already measurable and well understood.

This creates responsibility for everyone involved in the process: from end users generating AI requests to engineers designing hardware for data centers where AI workloads are hosted. In this context, we treat "responsibility" as an opportunity to influence the reduction of the energy footprint of AI across its complex production chain.

Seen through this lens, the rapid expansion of AI and the growing share of global electricity consumed by data centers mean that AI infrastructure providers will have an increasingly meaningful impact on the energy landscape in the coming decades.

Energy in, tokens out: Understanding the black box

From the outside, an AI infrastructure provider can look like a black box: energy goes in, and AI output, that is, the tokens[6] generated during training or inference, come out (Figure 1). But the path between those two points is anything but simple. Turning electricity into useful AI work depends on a stack of interconnected systems and components, each with its own efficiency trade-offs.

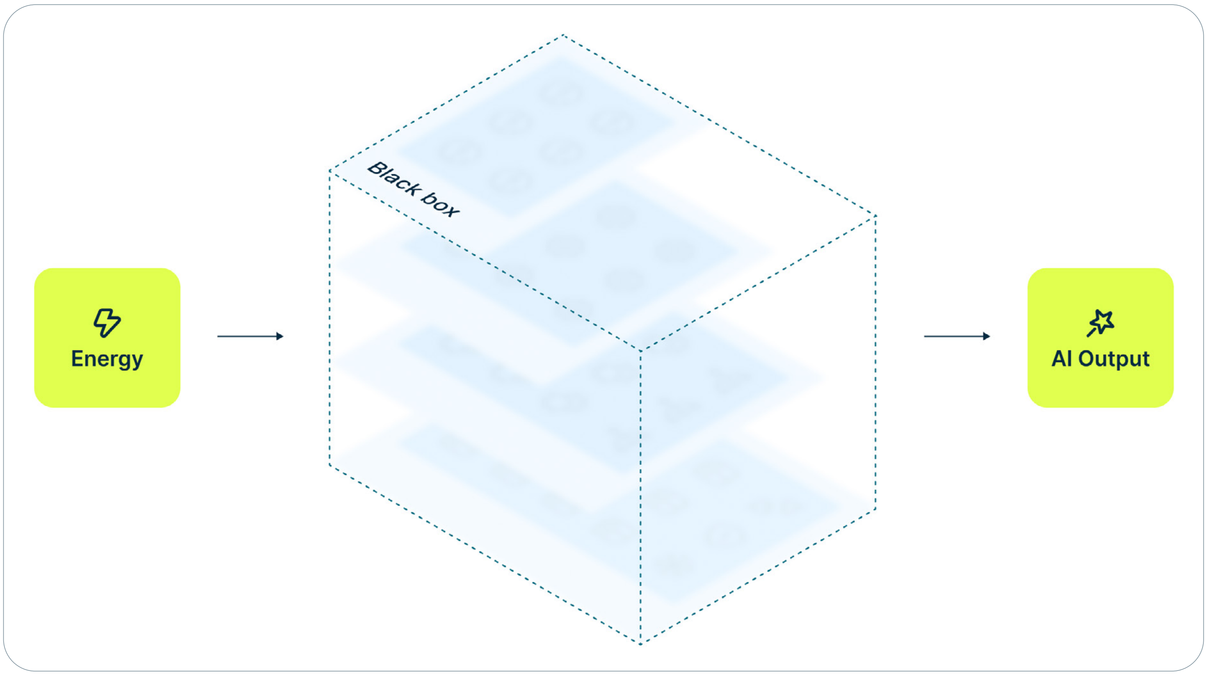


Figure 1. AI infrastructure providers produce AI output from energy.

Providers build and operate these systems in very different ways. Hardware choices, cluster designs, computing fleet management and data center architecture can vary dramatically. As a result, two providers can draw the same amount of power yet produce very different amounts of AI output.

To understand why and to see where sustainability gains, and losses, can come from, we need to look inside this black box. This way we can examine how energy travels through the system and identify parts with the strongest impact on overall energy use.

[6] Pieces of text, like a short word, part of a word, a symbol generated by an AI model.

Four layers of efficiency

Modern cloud providers serve many customers at once. Each customer's workload is shaped by its own mix of factors, from intensity and timing to software set-ups. As a result, each customer leaves a different sustainability footprint and requires their own optimization strategy.

All of these workloads, however, run on the same underlying data center resources. That means that the cloud provider is still accountable for the total power used in a facility and can leverage its infrastructure to turn this power into AI output more efficiently.

To make this easier to understand, we break the energy-to-AI output process into four layers of efficiency (Figure 2). This framework shows where we have direct influence, where efficiency depends on customer choices, and where it is shaped by upstream hardware and software vendors. It also helps tie engineering decisions to measurable sustainability outcomes. By tracking these outcomes, we can better control and further optimize the environmental impact of AI workloads running in our infrastructure.

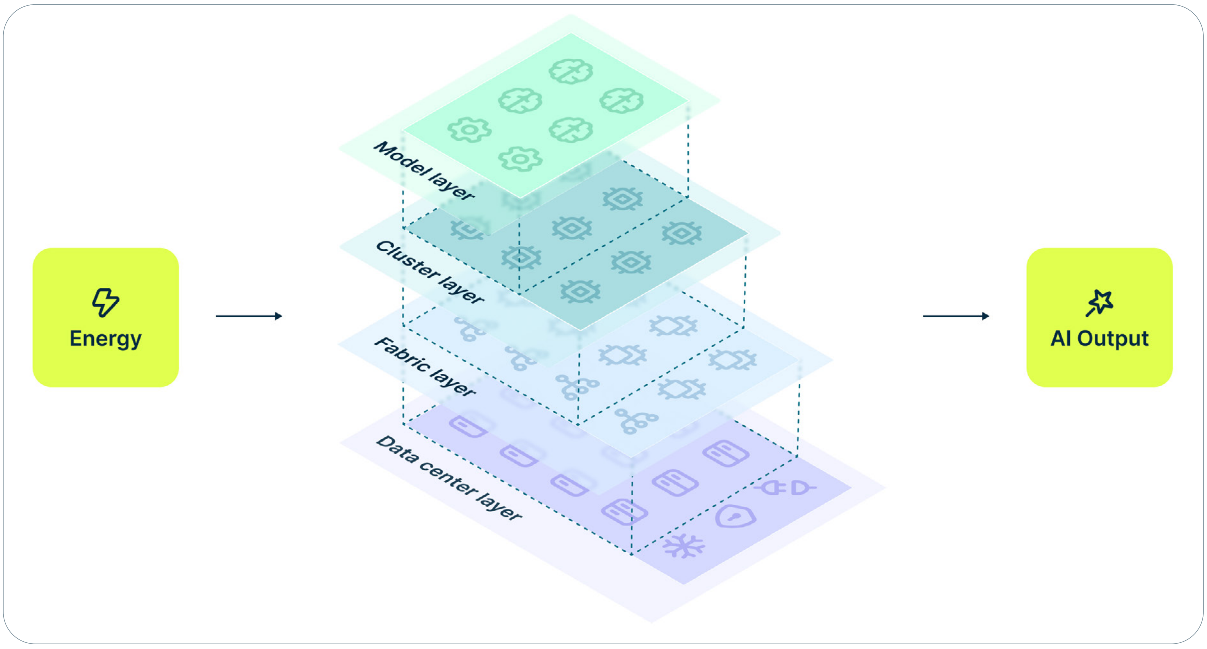


Figure 2. Four layers of efficiency we factor in on the journey from energy to tokens.

The energy-to-AI output process is complex, shaped by different groups of stakeholders, each influencing the final energy footprint to a varying degree at different layers of efficiency. The table below provides examples of such stakeholders and describes their roles within the value chain.

Stakeholders in the Energy-to-AI Output Process

Stakeholder group	Users	Hardware and software vendors	AI infrastructure provider
Who's in	AI labs and model developers Builders of AI-native products Enterprises adopting AI features in operations or products Public sector	Vendors of AI hardware Developers of software for AI stack AI labs and model developers	Cloud providers
Role	Use AI infrastructure to train or run ML models; define workload patterns and model choices	Supply the GPUs, systems, components and software that enable AI; define performance limits	Builds and manages the compute environment and provides it to the customer

Model-layer efficiency

Efficiency at the model level refers to how a model runs and interacts with the underlying AI hardware. Model parameters, hardware characteristics and runtime optimizations together shape this model, and their interactions affect how much energy is required to generate each token. For example, on the same hardware, Llama 3 405B consumes about 2.3x more energy per token than GPT-3 175B, which is roughly 8.8 Joules per token versus 3.6 Joules per token[7]. The reverse is also true: the same model can draw different amounts of energy depending on the hardware it is optimized for.

At this layer, these groups play the key roles:

- **User**, who selects the model and its configuration.
- **Hardware and software vendors**, who define specifications and capabilities of hardware and software used to run the model.
- **AI infrastructure provider**, who provides the compute environment and can optimize the runtime for the model.

Cluster-layer efficiency

Cluster-layer efficiency comes from the capabilities of provisioned GPU clusters as well as how they are put to use. A well-built cluster provides higher throughput, fewer interruptions, and more predictable performance. This reduces idle time, minimizes retries and shortens the total runtime of a job.

The efficiency logic is simple: If a training run is complete 30% faster because the cluster is more stable and performant, its energy use drops by roughly the same amount as fewer GPU-hours are needed to produce the same AI output.

At this layer, two groups have the most influence:

- **User**, whose workload structure and configuration choices affect workload behavior and how effectively the cluster is utilized.
- **AI infrastructure provider**, who ensures the cluster's reliability and consistent performance and provides tools for improving workload behavior.

Fabric-layer efficiency

In Nebius terminology, a compute fabric is a pool of physically interconnected servers that operate as a single compute resource. A data center can host several such fabrics, and each one is typically shared by multiple users, isolated from one another as separate tenants.

Tenant activity is constantly shifting. Users request different numbers of GPUs, their jobs start and finish on different schedules, and their workloads vary in runtime. As these patterns change, some servers or racks may remain powered on but idle — in some cases still drawing as much as 12-25% (depending on the GPU model) of their active power consumption, as our internal observations indicate.

These idle GPU-hours — we refer to them as GPU × hour slices — represent lost potential: capacity (and power draw) is there, but it is not producing tokens or training progress (Figure 3). This lowers fabric-level efficiency, calculated as the amount of AI work generated per unit (for instance, MWh) of available fabric capacity.

Two groups play the key role at this layer:

- **User**, who influences utilization through workload patterns and job timing.
- **AI infrastructure provider**, who manages resource allocation, scheduling and the ability to leverage small unused slices.

(7) Semianalysis. H100 vs GB200 NVL72 Training Benchmarks - Power, TCO, and Reliability Analysis, Software Improvement Over Time (2025).

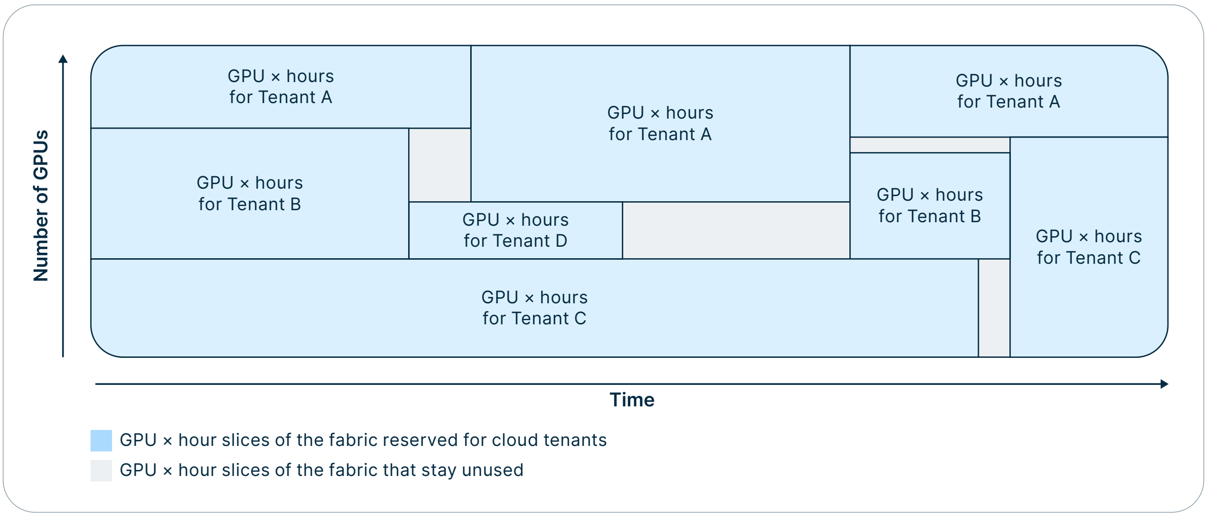


Figure 3. GPU × hours representation of a cloud fabric with some underutilized slices (gray)

Data center-layer efficiency

Data center-layer efficiency reflects how effectively a facility turns its total energy intake into power delivered to IT equipment. The standard metric for this is PUE (Power Usage Effectiveness):

```
PUE = total data center energy consumption ÷ energy consumed by IT equipment
```

Total energy includes both the power used by servers and GPUs as well as various overheads, including cooling, power conversion and distribution losses, and the energy required for physical infrastructure such as lighting and security systems. In 2025, the global average annual PUE was at 1.54, meaning that for every watt the IT equipment used, data centers needed an extra 54% just to keep the environment running[8].

Two groups shape efficiency at this level:

- **Hardware and software vendors**, who influence hardware requirements that data centers must support, such as heat density and cooling needs for advanced AI systems.
- **AI infrastructure provider**, who designs and operates the data center, including cooling systems, connectivity and networking.

[8] Uptime Institute, Uptime Institute Global Data Center Survey Results 2025 (New York, NY: Uptime Institute, 2025).

Our shared responsibility

Across the energy-to-AI output chain, responsibility is shared. Each stakeholder influences efficiency in different ways, and their decisions often set the boundaries for everyone else. To make these relationships clear, and to show where Nebius's influence is strongest, we use a simple influence matrix.

Layer-Role Influence Matrix

Layer	User influence	Hardware and software influence	AI infrastructure provider influence	
			Typical provider	Nebius
Model	High	High	Low	High
Cluster	Medium	Low	High	High
Fabric	Medium	Low	High	High
Data center	No impact	Medium	High*	High

*This value applies to AI infrastructure providers that operate data centers they own.

No impact — The stakeholder cannot affect efficiency at this layer. For example, a user cannot change the physical design of the data center and therefore has no control over its PUE.

Low — The stakeholder has only limited, indirect influence on efficiency at this layer. For example, hardware and software vendors define the quality of the components for AI clusters, but their contribution is significantly limited by the implementation on the AI infrastructure provider's side.

Medium — The stakeholder's decisions meaningfully affect efficiency, but do not fully determine it. For example, a user can improve training code and scheduling, but the gains still depend on the stability and performance of the underlying cluster.

High — The stakeholder's core design and operational decisions define efficiency at this layer. For example, the AI infrastructure provider decides on the data center's cooling architecture, which has a major impact on the facility's energy use.

In this framework, Nebius acts as the AI infrastructure provider operating its own data centers, and holds high influence on the cluster, fabric and data center layers. Unlike a typical provider, Nebius also runs an inference platform and develops tools that improve the performance of open-source models. This gives Nebius influence at the model layer as well.

As a vertically-integrated AI cloud provider, Nebius has multiple opportunities to optimize energy utilization across all four layers of energy-to-AI process, from the moment energy enters our facilities to how it's converted into AI output.

The next sections outline the measures we apply at each of these layers to maximize AI output from a given unit of energy.

How we combine software and hardware efficiency to drive sustainability progress

Nebius offers a vertically-integrated AI cloud. We design the hardware, develop the software stack and build the data centers where everything runs. Because these layers sit under one roof, we can make engineering decisions that cut across them.

We choose where to innovate, how components should interact and which trade-offs produce the best efficiency at each layer of the energy-to-AI chain. This structure gives us direct control over both the individual optimizations and the way they fit together, which is a meaningful opportunity to maximize sustainability gains.

Software engineering

Model-layer improvements

Instead of running inference on a standard out-of-the-box deployment, Nebius Token Factory, a production inference platform, applies runtime optimizations that enhance model-hardware interaction by increasing model throughput and reducing compute needs, without changing the model or the underlying hardware.

Token Factory also offers post-training optimization tools that add a set of improvements on top of an open-source model's core capabilities. These methods allow models to produce more useful output per unit of compute on a given hardware than their default configurations typically allow.

Cluster-layer improvements

Nebius applies a broad set of reliability mechanisms to keep production AI clusters stable and resilient. Active and passive health checks, along with automated node isolation and replacement, reduce failures and cut down the time needed to recover a training job after an interruption.

This all has a measurable effect. One of our customers recorded 56.6 hours of stable operation on a 3,000-GPU production cluster, while studies show that clusters of similar size typically experience job interruptions roughly every 9.8 hours[9].

[9] 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

Performance optimizations such as topology-aware scheduling (placing jobs on GPUs that are physically close to each other) further shorten total job time. Faster completion improves customer ROI and reduces energy use, since fewer GPU-hours, and therefore less energy, are needed to finish the same workload (Figure 4).

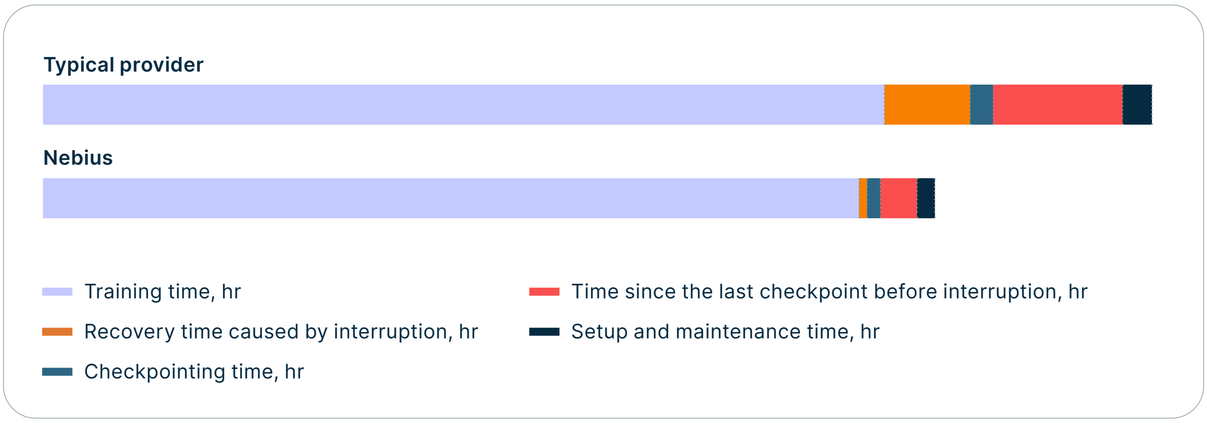


Figure 4. Simplified representation of how Nebius outperforms competitors by providing stable and resilient AI clusters.

We provide end-to-end observability of how a workload actually behaves on the cluster (GPU activity, idle patterns, power consumption and more) via Grafana dashboards and the Nebius console. With this data, customers can detect underperforming jobs and tune workloads in ways that reduce idle time and shorten total job duration.

Fabric-layer improvements

A virtualized cloud fabric lets us make far better use of GPUs we operate. On-demand and preemptible instances (short-lived GPU allocations that can be reclaimed when capacity is needed elsewhere) give customers elasticity while allowing us to harvest small unused slices in the GPU × hours fabric. This minimizes fragmentation and increases the amount of AI work produced per watt delivered to the fabric.

This is a key difference from bare-metal providers, whose resources are allocated in large, fixed chunks, leading to gaps between customer reservations and, therefore, idling hardware. Virtualization lets us pack workloads more tightly and use the entire fabric more efficiently (Figure 5).

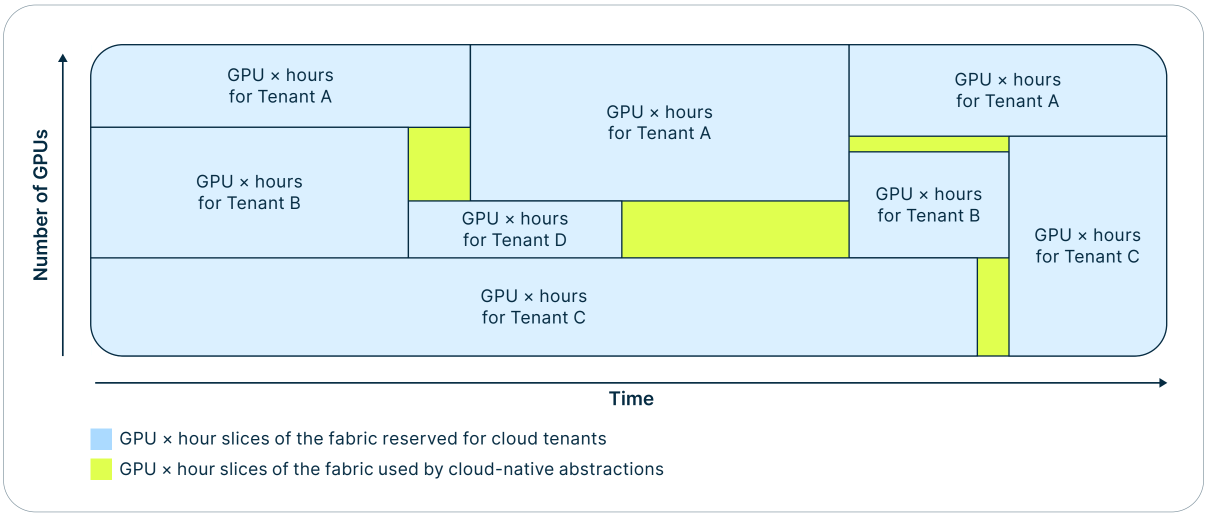


Figure 5. Simplified representation of how Nebius uses virtualization to capture small unused capacity that bare-metal setups leave idle.

Virtualization also reduces the amount of hardware that must be held back as spare capacity. Bare-metal environments typically reserve 10–20% of each tenant's cluster for node replacement, leaving those GPUs idle. Nebius maintains a single shared buffer across many tenants and allocates it dynamically whenever auto-healing is required. As a result, we need only about 4–5% of cluster capacity as buffer, significantly cutting idle power draw.

Storage contributes as well. Each fabric hosting active tenants requires block, file and object storage, and Nebius provides all of these through a unified, software-defined storage architecture. This allows diverse storage options to run on the same hardware instead of separate, duplicated systems. With fewer dedicated hardware nodes per workload, total facility power consumption is lower.

Hardware engineering

Cluster-layer Improvements

We design and refine our server and rack hardware in-house, testing them extensively to reach performance targets that support durability, resilience and predictable behavior under load. Unlike providers that rely on off-the-shelf solutions, we avoid inherited design issues that would otherwise drive failure rates and troubleshooting overhead. We also develop our own proprietary firmware. Controlling this low-level software gives us tighter reliability guarantees, reduces operational variance and allows faster troubleshooting when issues arise.

Together, this helps ensure clusters stay stable, workloads run without unnecessary interruptions, and maintenance (when required) is fast and predictable. Fewer failures mean less rework and shorter time-to-result, which ultimately lowers the total energy required to produce AI output.

Fabric-layer improvements

Nebius's servers are designed for high packing density and predictable power behavior, even as workloads fluctuate. We engineer them to operate reliably across wider temperature ranges than off-the-shelf hardware would typically tolerate, making GPUs less sensitive to thermal swings while also reducing cooling needs in the data center.

Coupled with an innovative cooling layout, these designs cut server power draw by about 20% compared with standard OEM servers[10] in our tests, effectively giving 20% more AI output per watt. Stable thermals and consistent node performance let the cloud platform pack workloads more tightly, reducing fragmentation and safely reclaiming small GPU slices.

[10] Custom Nebius server recorded an average power draw of 8.2 kilowatts, compared with 10.0 kilowatts for a comparable Gigabyte G593-SD0 GPU server, representing an approximate 20% reduction in server power consumption. Measurements were taken under controlled test conditions across inlet temperatures of 25°C, 30°C, 35°C, and 40°C.

Higher per-server efficiency and lower fragmentation mean each fabric we operate can deliver more compute from the same energy budget, improving the overall energy-to-AI ratio for active tenants.

Data center-layer improvements

To achieve top-tier PUE levels, our data centers combine innovative cooling technologies with an optimized facility layout. We have been using free-cooling systems and are now deploying closed-loop liquid cooling with no chillers or water loops, using air as the primary coolant. This design reduces cooling overhead, typically one of the major contributors to data center overheads, and enables efficient recovery of heat as a byproduct of energy consumption.

With these engineering decisions, our own data centers have been recently operating at PUE levels as low as 1.15, among the most efficient across the peer group[11].

Measuring sustainability outcomes

Improving efficiency across different layers has a measurable effect on Nebius' actual energy footprint. In 2024, we avoided roughly 20 GWh of electricity consumption[12]. About 10 GWh came from more energy-efficient hardware designs, and another 10 GWh from lower data-center overheads reflected by the data centers' annual PUE outperforming global averages. Without these engineering choices, total consumption could have been 20–30% higher.

Nebius also recovered a significant share of the energy it consumed. Between 2022 and 2024, over 50 GWh was captured as heat and supplied to the local municipal heating network in Mäntsälä, Finland[13]. This is roughly equivalent to the heating needs of 2,500 Finnish households over the same period. In 2024 alone, this recovered heat was enough to cover 65% of the local network's needs. Because this replaces heat that would otherwise be produced from conventional fuels, the associated emissions were cut by 54%, or by an estimated 3,220 tons of CO_2[12]. This equates to taking roughly 770 gasoline-powered passenger vehicles off the road for a full year[14].

> **"Working with Nebius, we've found a way to turn innovation and the presence of a state-of-the-art data center into something that directly benefits all of our residents. The heat recovered from the servers warms homes and public buildings — reducing emissions and strengthening our local energy system."**
>
> Hannu Laurila, Mayor of Mäntsälä, Finland

Scaling the sector responsibly

As demand for AI grows, considering the infrastructure powering it will only become more essential in the years ahead. Each layer of the AI production chain presents an opportunity to scale the sector responsibly and with greater resource efficiency in mind. By taking an engineering-led approach, we can ensure that every watt is used with purpose.

(11). Benchmarking used latest PUE figures reported by hyperscalers and other clouds.

(12) Nebius Group N.V., 2024 Sustainability Report (Amsterdam: Nebius Group, July 10, 2025).

(13) Madeleine North, "Here's How Data Centre Heat Can Warm Your Home," World Economic Forum (June 18, 2025).

(14) Passenger vehicle equivalency calculated using U.S. EPA standard emissions factors: one gasoline-powered passenger vehicle emits approximately 4.19 metric tons of CO_2 per year. Dividing 3,220 metric tons of CO_2 by 4.19 metric tons per vehicle per year yields approximately 770 vehicles removed from the road for one year.

NEBIUS

Nebius is the ultimate AI cloud. We combine custom hardware, proprietary software and energy-efficient data centers to deliver unmatched speed, scale and lower costs — on your terms. Whether you're building foundation models, fine-tuning or scaling inference globally, Nebius gives you the performance of a supercomputer with the flexibility of a hyperscaler.

Learn more at nebius.com.