FOUNDATIONAL | 1,852 words | 7 min read | Updated August 2026

What Is a NeoCloud?

A NeoCloud is a cloud provider built specifically for AI workloads. This guide covers what the term means, how the architecture works from the network fabric down to the power, how NeoClouds compare with hyperscalers, traditional cloud, and colocation, and what to check before you sign.

1. Executive Summary

A NeoCloud is a cloud provider built for one job: running AI workloads on large fleets of GPUs. Instead of a general purpose cloud's broad catalog, it sells dense accelerator capacity, the dedicated fabric that binds those accelerators into one cluster, and the storage and orchestration to keep them working. The term is also written as neocloud, one word. What customers buy is predictability: training runs that finish when the schedule says they will. This article covers what a NeoCloud is, how one is built from the fabric down to the power, how it compares with hyperscalers, traditional cloud, and colocation, and what to check before signing, starting with whether a provider owns its infrastructure or resells it.

2. Key Takeaways

- A NeoCloud is a cloud provider whose primary product is GPU as a Service, built for AI training, fine tuning, and inference rather than general IT.
- The defining feature is not the GPU. It is the dedicated back end fabric between GPUs and the removal of virtualization layers above them.
- What customers buy is predictability. Job completion time matters more than peak benchmark performance.
- A NeoCloud is a business model. A NeoCloud data center is a building. Once GPUs run inside it, the building is an AI factory. Some providers own neither.
- Judge providers on useful output per dollar and on energized capacity, not the GPU hour price. Power gates deployment more often than chip allocation.

3. What Is a NeoCloud?

A NeoCloud, also written neocloud and sometimes called an AI cloud or an AI NeoCloud, is a specialized cloud infrastructure provider dedicated to AI workloads. It gives customers access to clusters of accelerators, almost always GPUs, together with the networking, storage, and scheduling needed to run large scale training, fine tuning, and production inference. The commercial model is usually described as GPU as a Service, or GPUaaS. The terms carry no technical difference.

What separates a NeoCloud from a general purpose cloud is the order the design decisions were made in. Traditional clouds were built around CPUs, virtualization, and a wide service catalog, then adapted for AI. A NeoCloud starts at the workload and works backward through the fabric, the rack, the cooling, and the power.

Why did NeoClouds emerge?

Two forces created the category: worldwide scarcity of top end accelerators, and revenue diversification by the largest chip producers, who had reason to route supply beyond the small group of hyperscalers that dominated early demand.

NeoClouds began as stopgaps, and that origin still shapes the economics. Bare metal as a service is a thin business, and McKinsey's assessment is that providers who stay there are the most exposed, with durable positions more likely in sovereign compute and specialized workloads than in direct competition with hyperscalers.

What problem does a NeoCloud solve?

General purpose cloud infrastructure carries overhead that AI workloads feel acutely. Virtualization costs throughput. Shared networking introduces variance between nodes. Neither matters for a web application. Both matter when a training job synchronizes thousands of GPUs many times an hour and one slow node stalls the run.

NeoClouds strip out layers of abstraction so job completion time becomes predictable rather than probabilistic. Predictability, not peak benchmark performance, is what customers are buying.

Is a NeoCloud the same as an AI data center?

No. An AI data center is the physical facility: power, cooling, high density racks, and the fabric connecting them. A NeoCloud is the business that operates one, or leases space in one, and sells access to the compute inside. One is a building, the other is a service model. Once GPUs are running, that facility is commonly called an AI factory. Companies owning both layers control more of their cost structure.

4. How Does a NeoCloud Work?

A NeoCloud is usually described in three technical layers. In practice there is a fourth underneath, and it decides whether the other three ever get built.

The full stack, end to end:

Power > Data center and cooling > Compute > Networking > Storage > Orchestration > AI workload

1. AI optimized compute

Servers run bare metal or with minimal virtualization, so little sits between the job and the silicon. Nodes are dense, typically four to eight GPUs each, and providers expose hardware topology so frameworks can be tuned to the actual machine rather than an abstraction.

2. Two networks, not one

NeoClouds run a front end network of standard Ethernet for management and user access, and a separate back end fabric carrying only GPU to GPU traffic: lossless Ethernet or InfiniBand, commonly 400G or 800G, in non blocking spine and leaf topologies, using RDMA or RoCE so accelerators reach each other's memory without routing through the CPU. Training depends on constant collective communication, so that dual network design,

not the GPU model, is the clearest marker of a genuine AI cluster.

3. Disaggregated and isolated infrastructure

Compute, storage, and networking scale independently. Tenants are separated at the network and identity layer, because models and training data are among the most valuable assets a customer owns.

4. The layer underneath: power

Everything above assumes energized megawatts. A single rack scale system such as NVIDIA's GB300 NVL72 draws on the order of 120 kW to 140 kW and is fully liquid cooled, well beyond what most colocation halls were built to deliver. Land, power, and shell gate deployment more often than chip allocation does.

5. NeoCloud vs. Hyperscaler vs. Traditional Cloud

The three models coexist, and most serious AI programs use more than one. The question is not which is best overall but which is best at the layer you need.

Factor	Traditional cloud	Hyperscaler	NeoCloud
Primary design goal	Abstraction and multi tenancy, CPU centric	Global scale and service breadth	Deterministic performance, accelerator centric
Compute model	Virtualized instances	Virtualized instances plus reserved AI stacks	Bare metal or minimally virtualized dense GPU nodes
Networking	Standard Ethernet	Standard Ethernet plus AI fabrics in select regions	Dedicated lossless back end fabric, 400G to 800G, RDMA or RoCE
Consumption models	On demand and reserved	On demand, reserved, and serverless by token or request	On demand, reserved, and multi year dedicated clusters
Service breadth and reach	Broad IT catalog, regional to global	Widest catalog, global footprint	Narrow, infrastructure focused, fewer regions
Relationship to power	Abstracted from the customer	Abstracted from the customer	Often core to the economics
Best fit	General enterprise applications	Breadth, compliance coverage, global footprint	Sustained training, fine tuning, high volume inference

A NeoCloud is also compared with colocation. Colocation sells space, power, and cooling; the customer brings the hardware. A NeoCloud sells the compute itself, racked, networked, and operated. The difference is who carries the capital cost of the accelerators and who is accountable when a node fails mid job.

6. Why It Matters

For a buyer, the NeoCloud decision comes down to two numbers that rarely appear on a price sheet. The first is useful output per dollar rather than dollars per GPU hour. A cheaper hour on a cluster with an unstable fabric or noisy neighbors costs more per finished training run than a more expensive hour on a cluster that holds its throughput. The second is time. Capacity contracted but not energized is not capacity, and an interconnection delay is indistinguishable from a product delay.

For an operator, the pressures run the other way. Accelerators depreciate quickly, energy is a large and volatile line item, and multi tenancy makes consistent job completion times hard to guarantee. Global data center electricity consumption was about 415 TWh in 2024, and the International Energy Agency projects it will more than double to around 945 TWh by 2030. Those pressures separate operators who own their infrastructure from those who resell it, and they surface in contract terms long before marketing.

7. The Bentaus Perspective

Bentaus is an energy first NeoCloud. The dividing line it draws is between providers that rent compute and providers that own the layers beneath it.

The internal shorthand is land, power, and shell, plus compute, plus management. Secure the land, the power, and the building. Put GPUs inside it and it becomes an AI factory. Add orchestration and it becomes operable.

Bentaus owns and operates its power assets and its data center infrastructure. Compute runs on AMD and NVIDIA platforms. Bentaus is a certified AMD NeoCloud partner, as well as a Supermicro partner. Ziani Systems' Power Asset Orchestrator manages both layers, energy and compute.

That orchestration is what makes ownership useful. GPU power draw can be modulated in response to grid signals, so an AI data center operates as a flexible load and takes part in demand response rather than sitting on the grid as a fixed, non interruptible draw.

Most NeoClouds treat power as a procurement problem. Bentaus treats it as part of the product.

Explore Bentaus NeoCloud | Talk with the Bentaus team

8. Frequently Asked Questions

What are examples of NeoCloud companies?

Lists of top NeoCloud providers mix four kinds of company: publicly traded GPU cloud specialists, venture backed AI cloud startups, energy and infrastructure operators that moved into AI compute, and aggregators reselling capacity they do not own. Identify that last group before you sign: a reseller can quote a price but cannot commit hardware, a power contract, or a remediation timeline. Ask whether the provider owns the GPUs, the facility, and the power.

What are the key differences between a hyperscaler and a NeoCloud?

Breadth against depth. A hyperscaler offers hundreds of managed services across dozens of regions, with AI compute one line in a large catalog. A NeoCloud offers a narrow set of services built around dense GPU clusters and the fabric connecting them, in fewer locations. Hyperscalers win on ecosystem and reach. NeoClouds win on price performance, hardware transparency, and access to current generation accelerators.

What is the difference between cloud and NeoCloud?

A traditional cloud is built on general purpose, CPU centric architecture and prioritizes abstraction and multi tenancy. A NeoCloud is accelerator centric and prioritizes raw, predictable performance. A NeoCloud is a type of cloud, not a replacement for one. Most organizations run both.

Is it spelled neocloud or neo cloud?

One word, neocloud, is standard in industry usage. Bentaus writes it NeoCloud with a capital C. Two word and hyphenated forms appear in search queries but are rare in technical writing. No standards body governs the term, which is why the spelling is still settling.

What is a NeoCloud data center?

A facility engineered around AI workloads rather than general enterprise IT: high density racks, liquid cooling, a dedicated back end fabric, and a power design built for sustained high draw with large transients. It is the building. The NeoCloud is the business.

9. Continue Learning

- NeoCloud vs. Hyperscaler: What's the Difference? (bentaus.com/ai-knowledge-center/neocloud-vs-hyperscaler)
- What Is an AI Factory? (bentaus.com/ai-knowledge-center/what-is-an-ai-factory)
- What Is GPUaaS (GPU as a Service)? (bentaus.com/ai-knowledge-center/what-is-gpuaas)
- Why Power Is Becoming the Biggest Constraint on AI Infrastructure (bentaus.com/ai-knowledge-center/power-constraint-ai-infrastructure)

10. Author, Review and Sources

Written by: Niv Calderon, Head of Partnerships and Marketing, Bentaus

Published: August 2026

Last updated: August 2026

Sources

- Cisco Systems, What Is Neocloud? cisco.com
- DriveNets, Neocloud Providers: The Future of GPUaaS for AI Workloads, drivenets.com
- McKinsey & Company, The evolution of neoclouds and their next moves, 19 November 2025
- International Energy Agency, Energy and AI, 2025
- NVIDIA, GB300 NVL72 platform documentation