Neoclouds: GPU-First Clouds for the AI Boom

Introduction talk at GroupBrain

Jap Papreja

IIT Mandi

1st November 2026

Roadmap

1. Warm-up: Where does AI run?
2. Why neoclouds exist
3. Neocloud vs. hyperscaler
4. Players and their stories
5. Beyond just GPUs
6. The money
7. Risks and the bubble debate
8. What's next

A simple question

You type a prompt into ChatGPT...

Where does it actually run?

- Not on your phone
- In a data center, on thousands of GPUs
- So... who owns those GPUs?

Quick ques

Who here has ever used a cloud GPU?

The cloud you already know

- **AWS, Microsoft Azure, Google Cloud**, the "hyperscalers"
- Rent computing by the hour instead of buying servers
- Hundreds of services: storage, databases, networking, AI tools
- Built to serve everyone: websites, banks, games, startups

Think of it as. . .

A giant **department store**: something for every customer.

Why did neoclouds appear?

1. **Demand exploded:** training and serving large models needs GPUs by the thousand
2. **Supply was tight:** GPUs were scarce, and the big clouds had waitlists
3. **Focus wins:** a company built only around GPUs can build and ship capacity faster

Neocloud

A cloud provider built specifically around GPU computing for AI workloads.

Neocloud vs. hyperscaler

	Hyperscaler	**Neocloud**
Focus	Everything: hundreds of services	GPUs for AI, little else
Hardware	Mixed: CPUs, storage, some GPUs	Dense GPU clusters, fast networks
Access	Virtualised, managed services	Often bare metal or Kubernetes
Contracts	Pay-as-you-go plus enterprise deals	Often long-term reserved capacity
Strength	Ecosystem, compliance, trust	Speed, GPU expertise, flexibility

Department store vs. specialist shop

Not pure rivals: Microsoft is also a major *customer* of neoclouds.

The landscape

Company	Known for
CoreWeave	The largest; started in crypto mining; listed on Nasdaq in 2025
Nebius	Spun out of Yandex; "full-stack" AI cloud; Nasdaq: NBIS
Lambda	Developer-friendly GPU cloud; raising money ahead of a planned IPO
Crusoe	Energy-first: builds data centers around power (e.g. the Abilene campus)
Nscale	UK-based, Nvidia-backed; raised $3.36B ahead of its planned IPO
Together AI	Focused on running (inference) open models

Two kinds of neoclouds

GPU clouds

- **Rent the GPUs**
- CoreWeave, Nebius, Lambda
- You run your own model
- Billed per GPU-hour

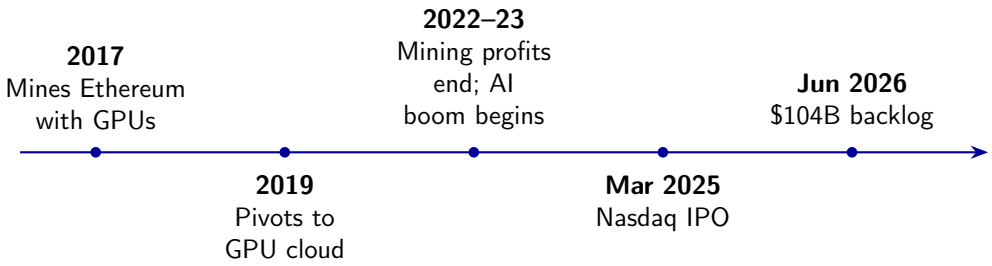
Inference clouds

- **Rent the model**
- Together AI, Fireworks, Baseten
- You call an API
- Billed per token

Renting a kitchen vs. ordering a meal.

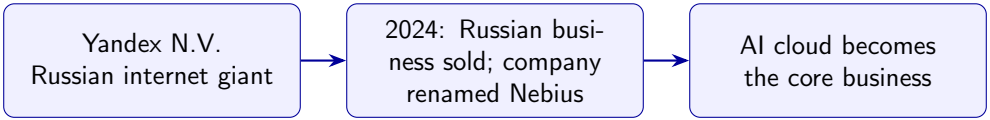
The lines are blurring: many now offer both.

CoreWeave: from crypto mining to AI cloud



- **Why it worked:** it already had GPUs and data-center know-how
- Q2 2026 revenue: $2.58B; 2026 guidance: $12.4–13.2B

Nebius: from Yandex to AI cloud



- Annualised revenue: \$1.25B (end 2025) $\rightarrow$ \$3.0B (June 2026)
- Target: \$7–9B by end of 2026; Q2 revenue \$582M (+454% year on year)
- A reported \$27B multi-year contract with Meta

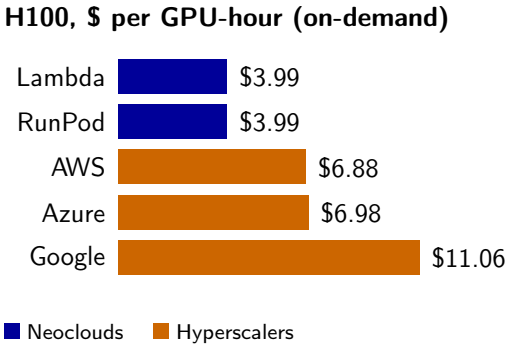
You can rent a GPU today

- Rent a GPU by the hour from a laptop
- Providers such as Lambda, RunPod, Together AI
- The giants (CoreWeave, Nebius) mostly serve big, long-term customers

Try it

Serve a small open model on a rented GPU and measure how fast it responds.

Which GPUs for our lab?



Our pick: a neocloud

- About 40–65% cheaper per GPU-hour
- Hourly billing, no long contract
- No data-transfer fees (Lambda, RunPod)

Choose a hyperscaler if

you have free cloud credits, or data rules require it.

List prices, Oct 2026 (IntuitionLabs). AWS and Google are 8-GPU nodes, Azure 1 GPU. Prices change often.

Running an open-source model

Pick a model from Hugging Face, then choose where it runs:

1. Hosted API

Together AI, Fireworks
Pay per token, no setup.
Fastest start.

2. Rent a GPU

RunPod or Lambda, with vLLM
You control the setup.
Best for many users.

3. Own machine

Ollama on a laptop or lab PC
Free, and data stays with you.
Best for testing.

Rule of thumb

An 8B model needs about 16 GB of GPU memory (less if quantised).

It's not just "lots of GPUs"

Talk fast

Thousands of GPUs must act like *one* computer. This needs specialised networks (InfiniBand, NVLink).

Power and cooling

AI racks draw far more power than normal servers. Liquid cooling, and finding enough electricity, is the real bottleneck.

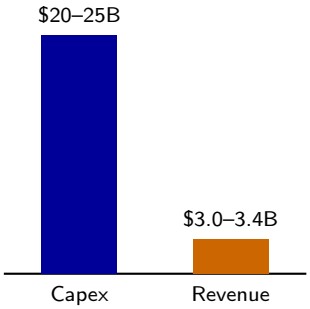
Reliability

At this scale something is always failing. Health checks and fast replacement *are* the product.

Example: Crusoe's contracted capacity is measured in gigawatts, not servers.

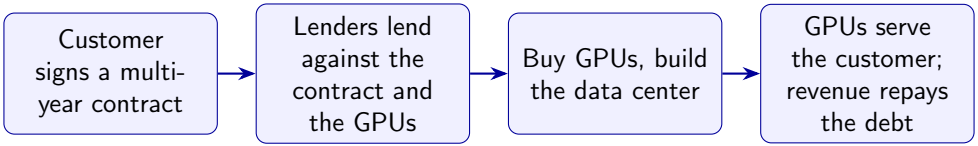
Why it is so costly

- GPUs, buildings, power and networking are paid for before the customer pays
- Nebius, 2026 guidance: capex $20–25B vs. revenue $3.0–3.4B
- One estimate: about $60B of neocloud capex in 2025



Nebius 2026 guidance, bars drawn to scale (midpoints)

How it gets financed



- The contract backlog is what makes lenders comfortable
- Nvidia sells the chips *and* sometimes invests in the customer (e.g. $1B in Nscale's latest round), a pattern critics call "circular financing"

The scale in numbers

CoreWeave backlog

$104B

Contracted revenue at 30 June 2026, about 8 years of this year's revenue

Nebius run-rate

$3.0B

Annualised revenue, up from $1.25B six months earlier

CoreWeave debt

$30B+

Long-term debt; Q3 interest guided at $0.86–0.94B for one quarter

Market forecast

$35B → $237B

One analyst projection, 2026 to 2031 (estimates vary widely)

Few customers, big contracts

- The biggest buyers: Microsoft, OpenAI, Meta
- CoreWeave: no customer above about 35% of backlog in late 2025, down from about 50% a quarter earlier
- Nebius: growth driven by large Meta and Microsoft deals

Why it matters

Long contracts make revenue predictable. A few customers also means a few points of failure.

Buildout or bubble?

The buildout case

- Contracted backlog worth years of revenue
- Capacity is near sold out
- Customers prepay to lock in scarce GPUs
- Inference demand is growing

The bubble case

- Heavy debt and rising interest bills
- A few customers carry the whole story
- How long is a GPU actually useful?
- Circular financing; custom chips (TPU, Trainium) from big clouds; more efficient models

Where the field is heading

- **Moving up the stack:** from renting GPUs to managed inference and tooling
- **Inference takes over:** one forecast puts it at about 80% of neocloud demand by 2030
- **IPO wave:** Nscale has filed and Lambda is raising ahead of a planned listing
- **Consolidation:** not every neocloud will survive a downturn
- **Sovereign and regional clouds:** countries want compute at home

Takeaways

1. Neoclouds are GPU-first clouds built for the AI boom
2. They win on speed and focus; hyperscalers win on breadth
3. The business is capital heavy: debt, long contracts, few big customers
4. Lasting infrastructure or bubble? Inference is where the next phase is decided

Sovereign AI and India

- Governments want AI compute (and data) on home soil
- India: the IndiaAI Mission provides subsidised GPU access through local providers
- Local players such as Yotta and E2E Networks
- Open question: can regional clouds compete on scale and cost?

Inside an AI cluster

- **NVLink:** GPU-to-GPU links inside a server or rack
- **InfiniBand / Ethernet (RoCE):** the network between servers
- **Slurm vs. Kubernetes:** how jobs get scheduled on the cluster
- **Liquid cooling:** dense racks produce too much heat for air

The GPU depreciation debate

- A GPU is bought for cash, then "used up" on paper over its assumed life
- Assume 3 years vs. 6 years and the yearly cost, and the profit, change a lot
- Loans are sized on what the GPUs are expected to be worth later
- New chip generations arrive every year or so, which keeps the debate alive

Sources

- CoreWeave Q2 2026: TIKR, 24/7 Wall St, Insider Monkey, Simply Wall St, Crypto Briefing (Aug 2026)
- CoreWeave Q3 2025 customer concentration: Data Center Dynamics
- Nebius Q2 2026: Quartr summary; Benzinga; Q1 2026 coverage via Pluang
- Nscale $3.36B financing: Fierce Network; Storage Newsletter (Oct 2026)
- Lambda: TechCrunch (6 Oct 2026); Crusoe: reports of June and July 2026
- Market size: Mordor Intelligence via SDxCentral; inference share: ABI Research; 2025 capex: J.P. Morgan estimate

Questions?

Thank you