*via*bandwidth

MARKET REPORT

State of GPU Cloud

An independent snapshot of the global GPU compute market, built from a directory where operators are verified against network evidence rather than self-reported claims. This is what the market looks like right now and how buyers should read it.

VIABANDWIDTH RESEARCH · Q3 2026 · FIGURES AS OF JULY 2026

588	157	45+	48
GPU OPERATORS INDEXED	NETWORK-VERIFIED DIRECT OPERATORS	ACCELERATOR MODELS TRACKED	COUNTRIES WITH INDEXED CAPACITY

EXECUTIVE SUMMARY

The GPU cloud market runs far wider than the dozen names that dominate the headlines. Right now viabandwidth indexes *588 operators offering GPU compute*, roughly nine times the count carried by the price-comparison sites that track only the best-known providers. The market is also younger and more fragmented than it looks from the outside. Of those 588 operators, *157 are network-verified direct operators* that run their own metal on their own networks, while a large tail resell or broker capacity that physically sits somewhere else. The newest accelerators have moved from launch to broad availability quickly, with H200 and B200 already listed by hundreds of operators, yet the specialist neocloud segment remains light on network footprint compared with the established cloud and carrier operators that also sell GPU. For a buyer, the practical consequence is simple. The shortlist worth building is not the fifteen famous names, it

is the far larger set of verified operators, where knowing which of them actually run the hardware is the difference between a real quote and a reseller markup.

01 · Methodology

How we measured this

Every operator in this report is drawn from the viabandwidth directory, where a listing is verified against independent network and infrastructure signals before it is trusted, not taken from operator marketing. Operators are classified as direct, marketplace or reseller from that evidence rather than from how they describe themselves. The figures below are a dated snapshot from July 2026. The live directory updates continuously, so a filtered view on the site may differ slightly from a fixed report, which is expected. We count operators, not chips or revenue, because chip inventory and financials are not something any directory can verify. We would rather report what is real than estimate what is not. This report makes no growth or trend claims, because it is a single point in time. The quarterly baseline it establishes is what makes future growth measurable.

> **What this report is not.** It is not a ranking by GPU count, price or performance, nor an endorsement of any operator. Where an operator lists an accelerator, that reflects what the operator publishes, verified for operator identity and network presence, not a physical audit of installed hardware.

02 · Composition

Who actually runs the hardware

The single most useful cut of the market is the sourcing split, because it separates operators who run their own infrastructure from those who resell or broker it. It is also the one thing buyers can almost never see on the open web, the place where a directory verified against network evidence earns its keep.

Share of the 588 indexed GPU operators by how they source capacity. Direct operators run their own infrastructure; resellers and marketplaces sell or aggregate capacity that sits elsewhere.

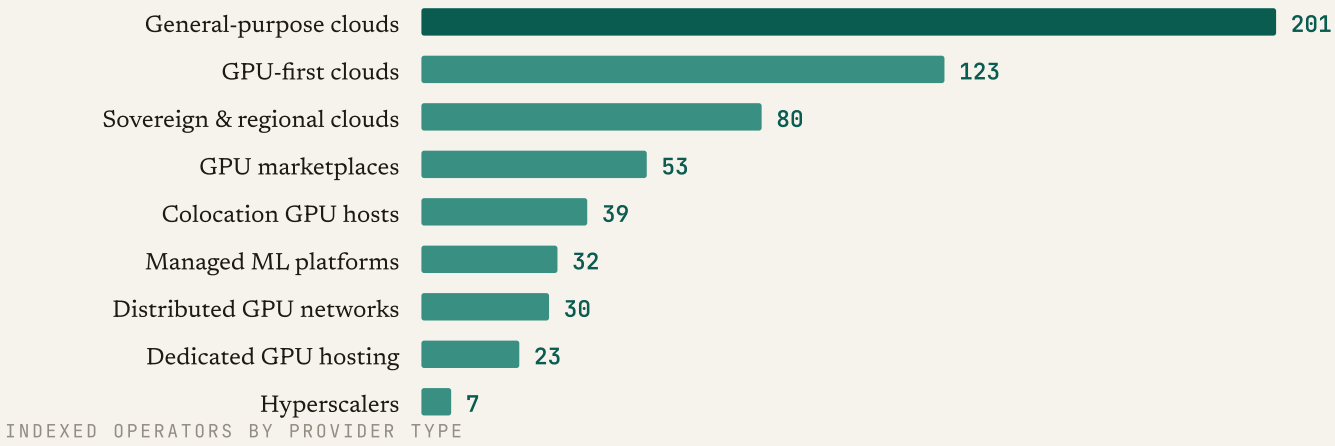
SOURCING CLASSIFICATION	OPERATORS	SHARE
Direct operator (own infrastructure)	375	64%
Reseller (capacity sourced elsewhere)	76	13%
Marketplace (aggregates third-party supply)	32	5%
Not yet classified	105	18%

Of the 375 direct operators, 157 are network-verified, meaning they announce and run the address space their compute sits on. That verified-direct set is the sharpest shortlist a GPU buyer can start from, because it screens out the resellers and brokers whose price carries a markup and whose location is somebody else's.

Only 157 of the 588 indexed operators announce and run their own network. The remaining 431 sell compute that sits on capacity somewhere else, which is exactly what a network-verified directory is built to tell apart.

What kinds of providers these are

The sourcing split says who owns the hardware. It does not say what sort of business each operator runs, which is a separate question. Sorted by the kind of provider they are, the market is led by general-purpose clouds that rent GPUs alongside everything else, followed by a deep field of GPU-first clouds built specifically for accelerated compute, with a substantial sovereign and regional layer serving buyers who need capacity inside a particular jurisdiction.



INDEXED OPERATORS BY PROVIDER TYPE

The 588 operators sorted by the kind of provider each one is. General-purpose clouds are the single largest group, yet the GPU-first, sovereign and marketplace layers together outnumber them.

READING THE PROVIDER TYPES

General-purpose clouds

Full-service clouds that rent GPUs alongside storage, compute and the rest of their menu.

GPU-first clouds

Providers built specifically for accelerated compute, where GPUs are the whole business.

Sovereign & regional clouds

Operators serving a single country or region, often for data residency.

GPU marketplaces

Platforms that aggregate GPU supply from many independent hosts.

Colocation GPU hosts

Datacenter and colocation operators that also rent GPU capacity.

Managed ML platforms

Providers that run the training and inference stack for you, not just raw servers.

Distributed GPU networks

Decentralized networks that pool GPUs across many locations.

Dedicated GPU hosting

Single-tenant GPU servers rented bare or fully managed.

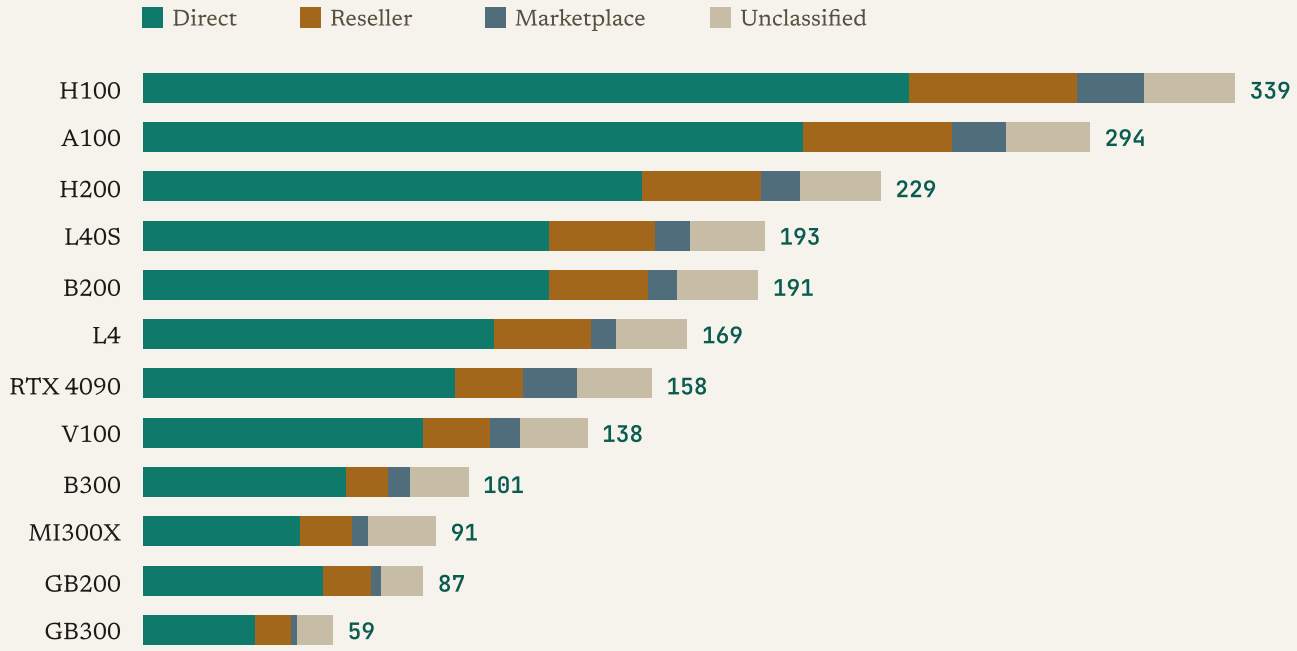
Hyperscalers

The largest global clouds, the AWS, Azure, Google and Oracle class.

That spread is where the choice lives. The names a buyer reaches for first are a small set of general-purpose clouds, while the GPU-first, sovereign and colocation operators hold the capacity that fits a specific region or budget, for anyone who looks past the front page.

Which chips the market is actually offering

Across all 588 indexed operators, the accelerator lineup shows both the long install base of prior-generation cards and how fast the newest silicon has reached broad availability. Counts below are the number of indexed operators that list each accelerator.



Number of indexed operators listing each accelerator, split by how the operator sources capacity. Direct operators account for the bulk of every model.

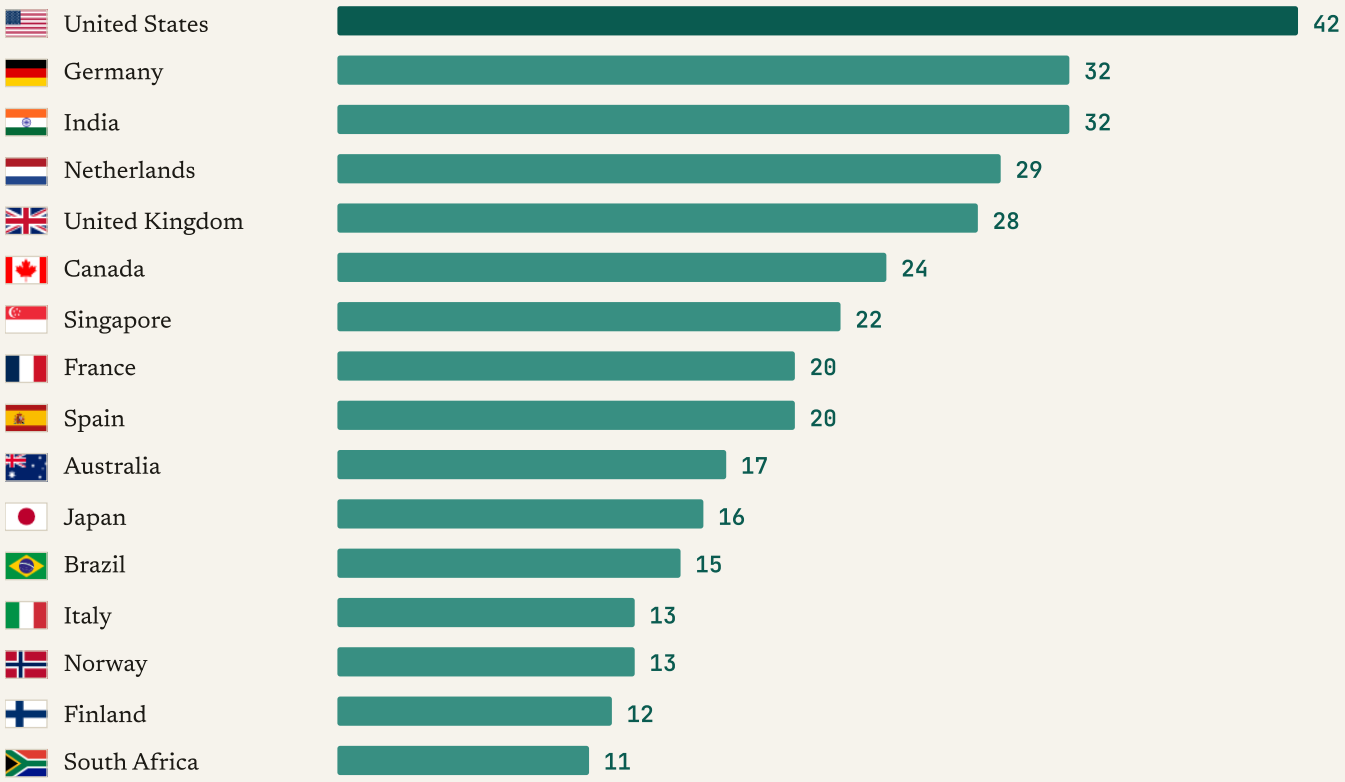
H100 and A100 remain the workhorses of the market by a wide margin, so buyers chasing only the newest silicon are competing for a much thinner supply. The current generation has spread fast underneath them, with **H200 at 229 operators and B200 at 191**, while the rack-scale GB200 and GB300 systems stay concentrated among a smaller set of operators, which is where availability rather than price becomes the binding constraint.

AMD as the credible second source

AMD's MI300X now appears with 91 operators and the MI325X with 31, which makes AMD a real second-source option rather than a rounding error. For a buyer whose workload is not locked to CUDA, that is the difference between one supply queue and two.

Where the capacity sits

Among operators that publish their deployment locations, the map is broader than the usual United States story. The following counts operators that state GPU capacity in each country, so an operator present in several markets is counted in each.



OPERATORS STATING CAPACITY, BY COUNTRY

Operators that state GPU capacity in each country. Depth outside the United States, including the Nordic sovereign markets, is real and gives residency-constrained buyers somewhere to go.

COUNTRY	OPERATORS WITH STATED CAPACITY
United States	42
Germany	32
India	32
Netherlands	29
United Kingdom	28
Canada	24

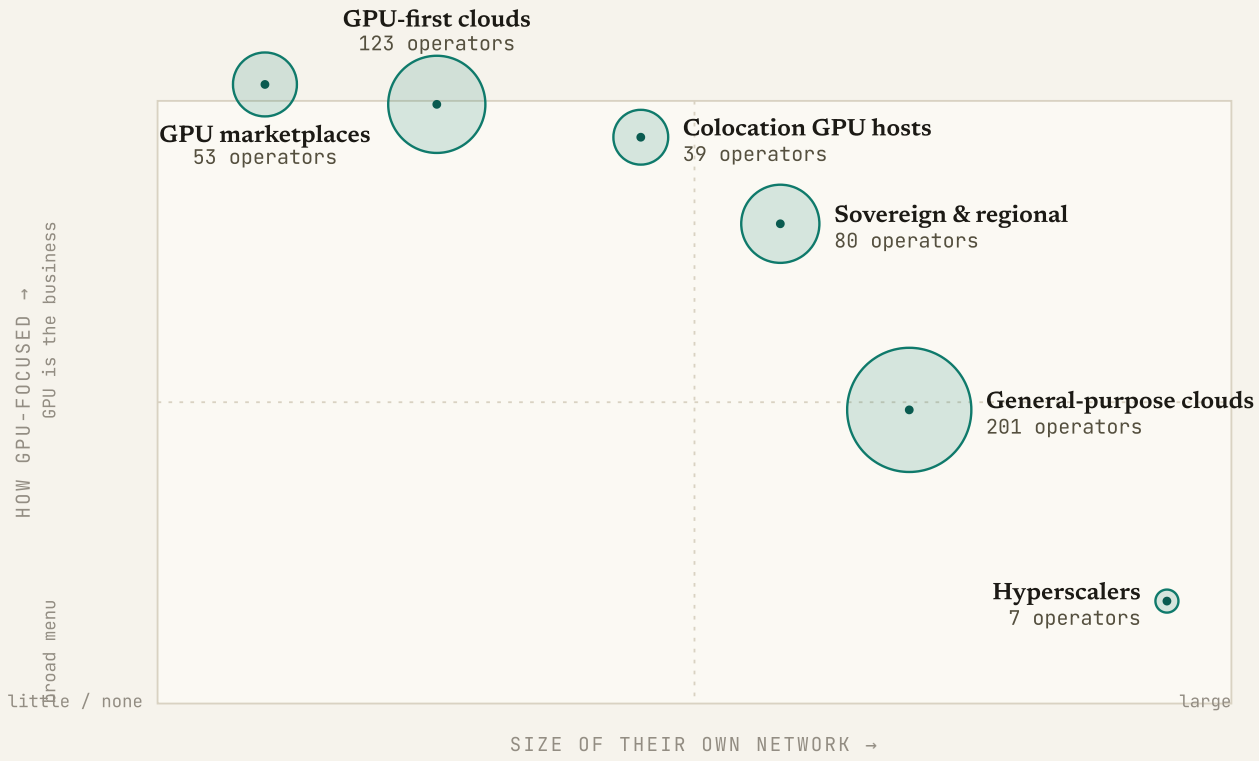
COUNTRY	OPERATORS WITH STATED CAPACITY
Singapore	22
France	20
Spain	20
Australia	17
Japan	16
Brazil	15

The United States leads but does not dominate. **Germany and India sit level as the next markets**, with the depth of the Netherlands and Singapore pointing to how much sovereign and regional demand now shapes where operators place GPU capacity. A buyer with a data-residency or latency requirement has real choice outside the usual hyperscaler regions.

05 · Network footprint

Scale is not the same as specialism

Ranking GPU-listed operators by the size of their own network footprint surfaces a useful truth: the largest networks in the GPU market belong to established cloud and carrier operators, not the specialist neoclouds. Footprint here means the operator's own routed IPv4 address space, which reflects overall network scale, not installed GPU inventory.



Provider types placed by the size of their own network and how much of the business is GPU. Bubble size shows the number of indexed operators. Among direct operators the sixteen with the largest networks list two to three accelerators on average, while the leaner GPU-first operators average more than five, so the biggest networks are rarely the deepest GPU specialists. Positioning is illustrative by provider type, not a ranking of any operator.

OPERATOR	OWN IPV4 FOOTPRINT
OVHcloud	4.59M
Tata Communications	4.46M
IBM	3.89M
Leaseweb	2.13M
Cisco	1.73M
INAP / Constant	1.47M
Samsung SDS	1.37M

The read for buyers is that **network scale and GPU specialism are different things**. A large carrier footprint signals reach and stability but not necessarily the newest accelerators or the best price per

GPU-hour, while a lean neocloud may run current silicon on leased capacity. Both are legitimate. Knowing which is which, before outreach, is the point of a verified directory.

06 · For buyers

What this means for procurement

01 **Start from the verified-direct set, not the famous names.** The 157 network-verified direct operators are the shortlist least likely to carry a reseller markup or a location that is not their own.

02 **Separate the accelerator you want from the one that is available.** H100 and A100 supply runs deep while the current generation spreads fast, with rack-scale GB200 and GB300 still supply-constrained rather than price-constrained.

03 **Treat AMD as a real second source.** MI300X availability across 91 operators means a genuine alternative supply queue for workloads not locked to CUDA.

04 **Use geography as leverage.** With capacity indexed across 48 countries and real depth outside the United States, a residency or latency requirement is a filter, not a dead end.

- A credit in the Related research module shown on the relevant GPU operator profiles
- Exclusive to a single sponsor per edition, never shared
- Independent by design, so the credit is labelled and never changes the figures or the verification behind them

To present the next edition, write to **advertise@viabandwidth.com**.

---

viabandwidth is the verified directory of GPU compute, colocation and carrier networks.

Operators are verified using independent network and infrastructure signals. Figures are a snapshot as of July 2026 and the live directory updates continuously. This report counts indexed operators and the accelerators they publish, verified for operator identity and network presence. It does not audit installed hardware or make any claim about price, performance or revenue. Paid placement in the directory is always labelled and never affects verification or the figures in this report.

Cite as: viabandwidth, State of GPU Cloud, Q3 2026, viabandwidth.com/reports/state-of-gpu-cloud.