GPUaaS revenue will reach USD130 billion by 2030, with USD9.6 billion projected for telecoms operators

October 2025

Joseph Attwood

AI-based applications are at the heart of many enterprises' plans for growth and development. However, most enterprises do not own the infrastructure on which their AI workloads run because this hardware is expensive and complex. Instead, they rely partly or entirely on infrastructure sourced through GPU-as-a-service (GPUaaS) offerings to run their AI training, fine-tuning and inferencing workloads.

In Analysys Mason's recently published *GPU-as-a-service (GPUaaS): worldwide forecast 2025–2030* report, we forecast that total GPUaaS revenue will grow from USD21 billion in 2024 to USD134 billion in 2030. This article expands on some of the trends that we expect to see in this market during the forecast period.

Figure 1: GPUaaS revenue, worldwide

Source: Analysys Mason
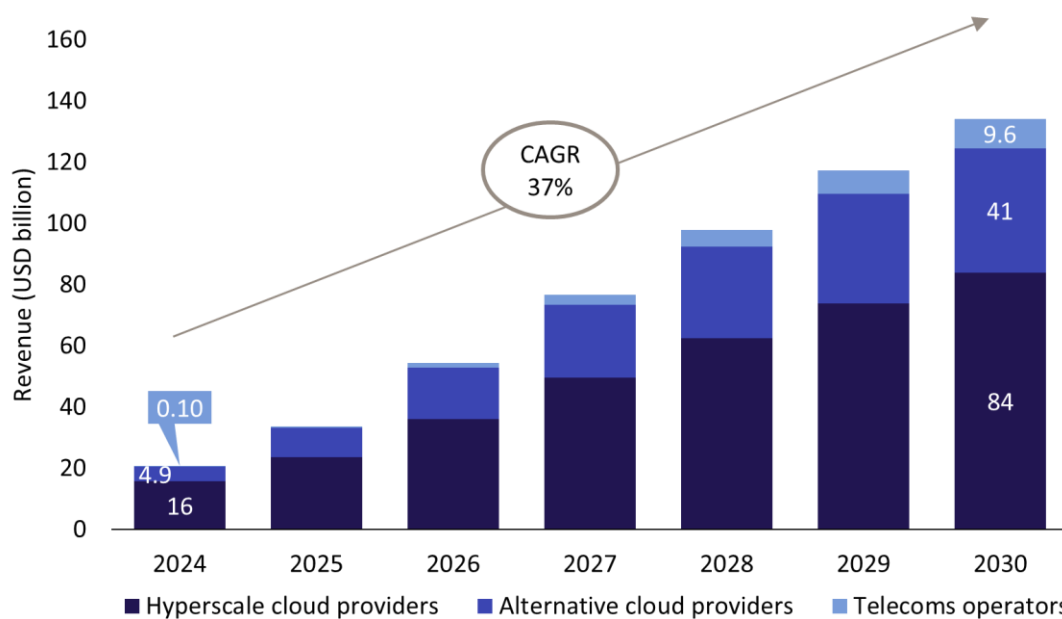
Hyperscale cloud providers account for the majority of revenue in the GPUaaS market, but their market share will gradually fall

Hyperscale cloud providers' (hyperscalers') dominance in the traditional cloud computing market has translated into an early lead in the GPUaaS market. In 2024, they accounted for 76% of all GPUaaS-related revenue. This was around 6% of their total cloud service revenue in 2024; we expect this figure to grow to around 8% in 2025. Hyperscalers often charge higher GPU rental prices than their competitors, but their solutions represent a compelling choice for enterprises for a number of reasons: hyperscalers have large amounts of GPUaaS

infrastructure; they have strong portfolios of AI applications and platforms; and they benefit from data gravity[1] because many enterprises already store data on their cloud platforms. However, we forecast that hyperscalers' share of total GPUaaS revenue will fall to 63% by 2030.

Hyperscale cloud providers' main competitors are alternative cloud providers, a category that consists of Tier 2 cloud providers and neoclouds, with neoclouds (such as CoreWeave, Lambda and Nebius) being the primary threat for hyperscalers. Neoclouds are companies that have launched GPUaaS solutions without first having 'traditional' cloud computing offerings. They have entered the market and scaled rapidly over the last 3 years and can offer competitive pricing because they have launched efficient GPU clouds and their businesses are not burdened by having to maintain legacy infrastructure and services. However, it is uncertain how viable some of these neoclouds will be in the long term because they often have high levels of debt and either have limited profits or are making a net loss, and must contend with falling GPU rental prices.

We have also seen around 30 telecoms operators launch GPUaaS offerings. Most of these offerings are still very new, so we put GPUaaS revenue for telecoms operators for 2024 at just under USD100 million, growing to reach USD9.6 billion by 2030.[2] However, there is still uncertainty around whether telecoms operators will be able to successfully compete against hyperscalers and neoclouds in this market.

Many GPUaaS providers hope sovereign cloud GPUaaS will be a monetisable differentiator for them, but the reality is uncertain

Data privacy and sovereignty regulations can impose restrictions on the countries in which AI infrastructure may be housed when hosting certain AI workloads. Evolving sovereignty regulations and increasing geopolitical tensions are expected to drive enterprises' demands for domestic AI infrastructure over the forecast period. A subset of the GPUaaS demand from enterprises with sovereignty concerns will be reserved for providers that are also headquartered in the country/region in which they are deploying AI infrastructure and can meet further sovereign cloud requirements. Telecoms operators and some neoclouds are hoping to capitalise on the opportunity to provide GPUaaS infrastructure that meets these requirements.

However, it is unclear how the demand for 'full' sovereign cloud GPUaaS will evolve. It is possible that most enterprises only care about getting access to infrastructure that is deployed in a particular country, but not about further sovereign cloud requirements (such as the location of a GPUaaS provider's headquarters). If this is the case, then specialist sovereign cloud GPUaaS providers will inevitably face fierce competition in the future as hyperscalers and neoclouds expand their GPUaaS footprints to a growing number of countries.

It is not clear to what extent providers will be able to charge a premium for offering sovereign cloud GPUaaS, which may limit their ability to monetise their sovereign cloud credentials.

AI inferencing will account for a growing share of GPUaaS workloads, reaching 78% of GPUaaS revenue in 2030

We estimate that enterprise spending on solutions for AI training and inferencing workloads accounted for roughly equivalent shares of GPUaaS revenue in 2024. Over the forecast period, GPUaaS compute used for AI

---

[1] Data gravity is the phenomenon where large stores of data tend to attract applications, workloads and more data to exist on the same platform as themselves.

[2] Our *GPU-as-a-service (GPUaaS): worldwide forecast 2025–2030* report also predicts the share of GPUaaS revenue that is attributed to AI-related professional services and AI applications/platforms provided by telecoms operators.

analysys mason

training workloads will increase as model developers create more and bigger models (in terms of the number of parameters used). Similarly, AI inferencing demands will increase as more enterprises and consumers adopt AI solutions and make requests to these solutions more frequently. The growing size of generative AI (GenAI) models, and a move towards reasoning models that do more 'thinking' when producing outputs, will also drive up the amount of compute used for inferencing. We expect growth in inferencing workloads to outpace the growth for training workloads. Inferencing workloads will therefore account for a growing share of GPUaaS revenue over the forecast period.

GPUaaS revenue from enterprises running fine-tuning workloads will also increase over the forecast period. This is because enterprises are increasingly integrating AI into their critical business processes and therefore require AI models that can be tuned on their data. However, many organisations are not currently well-equipped to fine-tune AI models, for example due to a lack of skillsets and/or available, well-governed data. As these challenges are resolved, enterprises will be more able to fine-tune AI models.

GPUaaS is also still being used for non-AI use cases such a cryptocurrency mining, gaming and visual effects rendering. However, growth in revenue from AI GPUaaS workloads will far outpace that for non-AI GPUaaS workloads.

Enterprise spending on GPUaaS will continue to increase over the next 5 years and beyond, but the market is being contested by a growing number of providers with different backgrounds. Capturing market share will be dependent on providers' abilities to carve out niches in terms of price, location of infrastructure, sovereign cloud capabilities, AI applications/platforms, AI-related professional services and other differentiators.

Analysys Mason's Cloud and AI Infrastructure *research programme provides data and strategic analysis on the GPUaaS market. Our research includes detailed forecasts and reports that provide insights on the strategies of GPUaaS providers (hyperscalers, neoclouds and telecoms operators) and AI hardware and software infrastructure vendors that serve these GPUaaS providers. To find out more about this content and our expertise in this area, contact* Joseph Attwood*.*

analysys mason