HOSTED·AI

GPU Mesh

Build a neocloud without GPU CAPEX.
Maximize GPU utilization and revenue.

Overview

- ↗ What is GPU Mesh?
- ↗ Hosted•ai + GPU Mesh economics
- ↗ GPU Mesh infrastructure types
- ↗ GPU Mesh infrastructure pricing

Requirements

- ↗ How to be a supplier
- ↗ How to be a buyer

Get started

- ↗ Next steps + more information

What is GPU Mesh?

GPU Mesh is a GPU capacity sharing network for service providers. It's an integrated part of the hosted·ai platform.

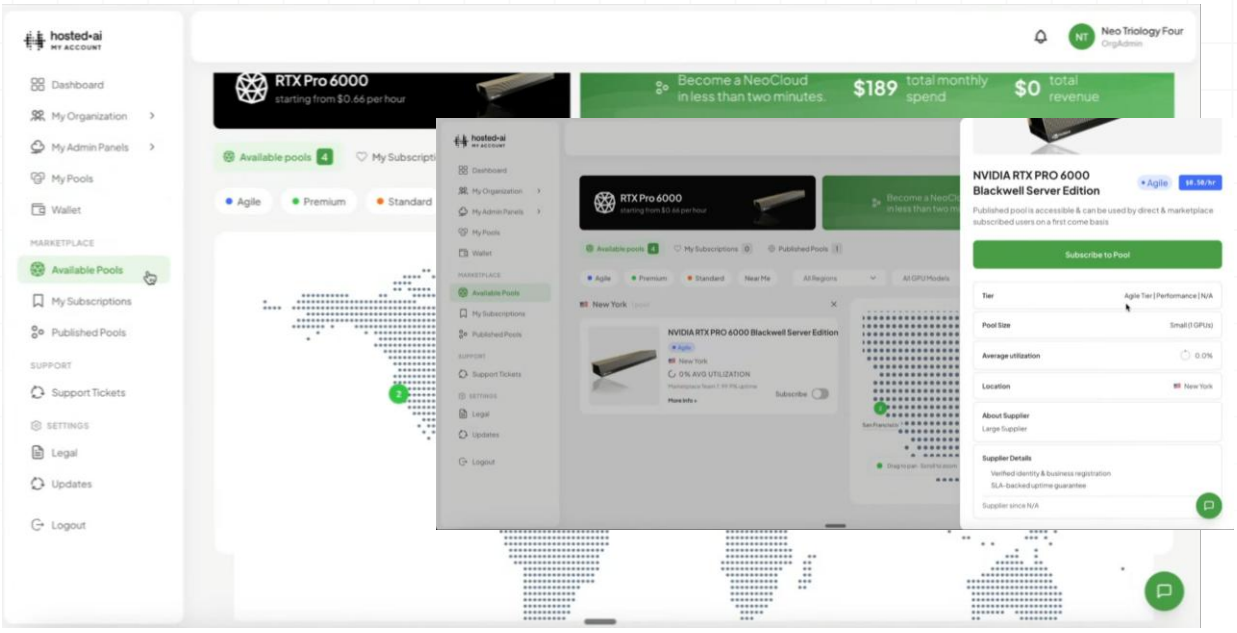
The hosted·ai platform automates orchestration, provisioning and billing of GPU infrastructure, either as bare metal GPU servers; GPU VMs; or true multi-tenant GPUaaS, based on hosted·ai's GPU abstraction, pooling and scheduling features.

GPU Mesh integrates with hosted·ai to enable service providers to sell true multi-tenant GPUaaS using GPU infrastructure resources from other providers.

It removes the CAPEX barrier for service providers to build and sell a GPU cloud - a 'neocloud' - and provides an easy channel for GPU infrastructure owners to monetize GPU in the most efficient and profitable way.

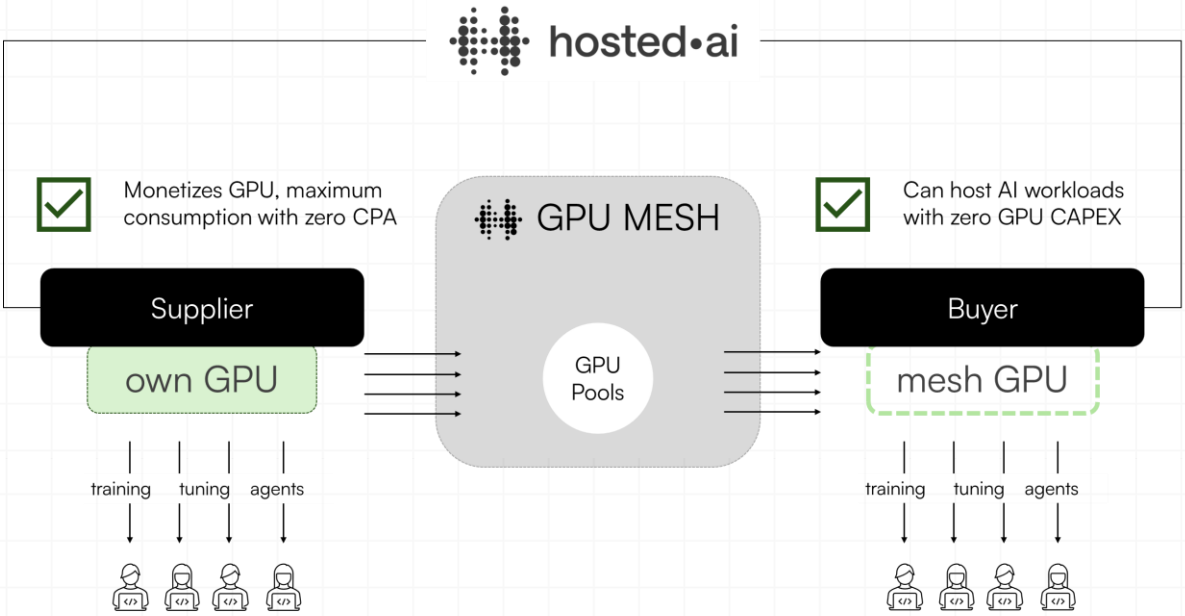
hosted·ai + GPU Mesh:

- Lowers entry barriers for AI cloud / GPU cloud hosting
- Provides easy access to global GPU resources for companies building and scaling neoclouds
- Enables more efficient, profitable GPU through resource pooling, optimized utilization and overcommit
- A complete GPU orchestration solution with a global demand channel built in



What is GPU Mesh?

You can use GPU Mesh to supply and sell GPU infrastructure; to buy and re-sell GPU infrastructure; or to do both at the same time.



Suppliers

- Suppliers are companies with GPU infrastructure
- Using hosted•ai, they create virtual GPU pools with optimized utilization and overcommit, and publish those pools to the Mesh
- Other service providers subscribe to those pools and re-sell them to customers; suppliers are paid for GPU resources that are consumed
- Publishing GPU pools to the Mesh does not make them exclusively available through the Mesh — you can sell those pools to your direct customers at the same time

Buyers

- Buyers are service providers that need GPU infrastructure.
- Using hosted•ai, they can subscribe to GPU pools on GPU Mesh, and re-sell that infrastructure as GPUaaS to their end customers
- Buyers are billed for GPU resources their customers consume
- Buyers can sell GPU Mesh resources to customers alongside any on-premises GPU infrastructure they have
- The source of GPU Mesh infrastructure is invisible to the buyer's end customer

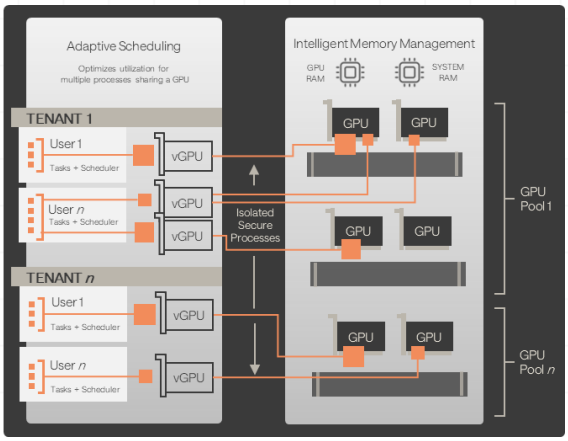
Hosted·ai + GPU Mesh economics

hosted·ai and GPU Mesh make GPUaaS accessible and profitable for service providers, by combining ultra-efficient GPU orchestration with a capacity sharing network.

The hosted·ai platform creates virtual GPU pools from physical GPU infrastructure. End user workloads get GPU resources from the virtual pool, rather than getting exclusive access to individual GPU cards.

This abstraction means resources that would otherwise be idle can be sold to other users. This is the foundation for profitable GPU infrastructure — maximizing utilization. It transforms GPU economics.

Amplifying this is hosted·ai's ability to overcommit GPU pools, further increasing utilization and multiplying revenue per GPU by the overcommit, or sharing, ratio.



GPU Mesh provides a sharing network for that ultra-efficient, virtual GPU pool infrastructure, connecting GPU supply with demand and delivering benefits for suppliers and buyers alike.

Supplier benefits: increased GPU utilization and revenue per GPU

Most GPU infrastructure today is sold as a static resource, with entire GPUs or fixed instances provisioned to a single customer. However, only about 40% of that GPU is used on average across training and inference, leaving 60% of your GPU investment effectively idle and unmonetized.

Using hosted·ai as the control plane for GPUs enables you to pool their resources, sell them to multiple tenants, and overcommit them to reach close to 100% utilization, and potentially 2-4x more revenue per physical GPU.

By publishing GPU pools to the Mesh, you also create new channels for GPU consumption without sales and marketing effort.

Buyer benefits: removing CAPEX from building and scaling a GPU cloud

The cost of GPU infrastructure is a significant barrier to market entry, for service providers trying to meet the explosive growth in demand for AI cloud.

At the same time, it can be difficult, as well as expensive, to source GPUs at sufficient scale to build a compelling GPU cloud / neocloud business.

Using hosted·ai with GPU Mesh, you can build a GPU cloud without GPU infrastructure CAPEX, and use ultra-efficient GPU resources to offer competitive GPUaaS pricing while still making healthy margins.

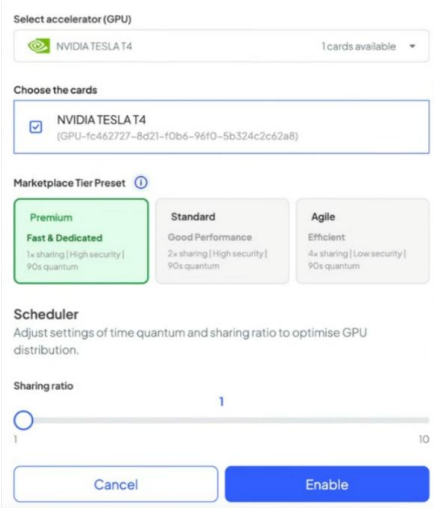
It also makes it easy to scale any existing GPU estate with new locations and GPU types, again without CAPEX.

GPU Mesh infrastructure types

GPU Mesh infrastructure is provided in three tiers: premium, standard, and agile.

Suppliers publish GPU pools configured for each tier. Each tier has different sharing (overcommit) and workload scheduling options to suit different buyer use cases, and to provide different monetization opportunities for infrastructure suppliers.

- **Premium tier**: 1x sharing, highest performance. Best for workloads that need minimum latency and highest security.
- **Standard tier**: 2x sharing, high performance. Best for workloads that need a balance of security, performance and price.
- **Agile tier**: 4x sharing, highest efficiency. Best for workloads that need optimal pricing.



Sharing and scheduling explained

Each tier uses a preset configuration of the hosted•ai workload scheduling and sharing settings for each pool. Those settings determine security and performance, and the potential revenue per GPU for suppliers, according to the level of GPU sharing/overcommit.

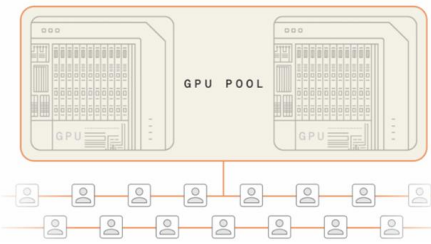
Sharing ratio

The sharing ratio controls how much overcommit is allowed - in other words, how many virtual GPUs are available for customers to buy from each physical GPU in the pool.

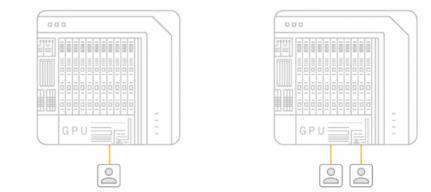
- The Premium tier has no sharing, delivering the highest performance. It's best suited to mission-critical and latency-sensitive workloads.
- The Standard tier has 2x sharing, and the Agile tier has 4x sharing - effectively doubling or quadrupling the number of customers that can simultaneously consume the GPU resources available.

  For example, with 8 GPUs in a pool, and 4x sharing, 32 virtual GPUs are available for customers to rent simultaneously.

hosted•ai — customers per GPU



Equivalent with passthrough / MIG

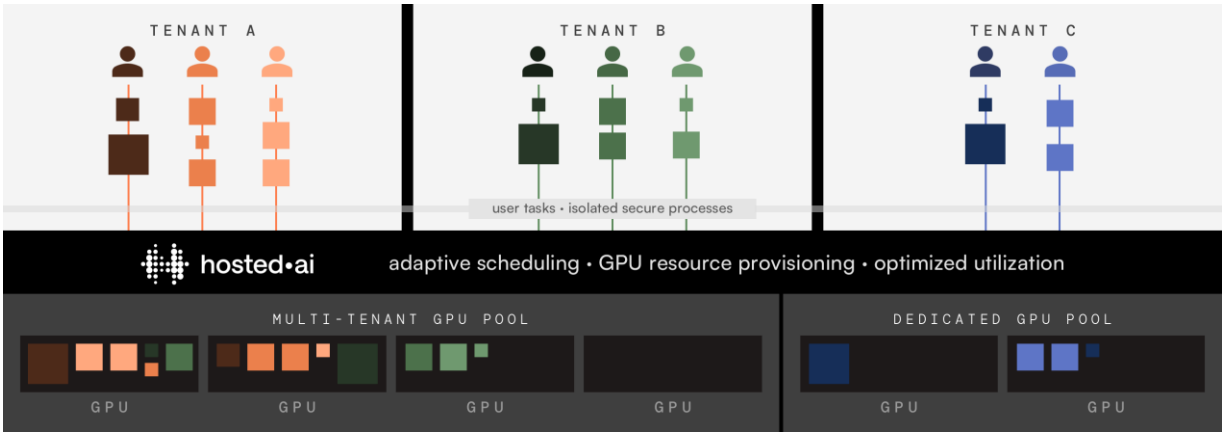


GPU Mesh infrastructure types

Scheduling type

The scheduling type controls how multi-tenant workloads are given access to GPUs in the pool, and how much utilization optimization can take place.

- **The Agile tier** uses spatial scheduling, which maximizes utilization: workloads can co-exist on GPUs. This offers excellent utilization efficiency, but lower security.
- **The Standard tier** uses temporal scheduling, which offers the highest security: workloads are swapped in and out of GPUs in the pool, and have full access to GPU resources while they have priority.
- **The Premium tier** has no GPU sharing, but is also based on temporal scheduling. This can become relevant depending on how the pool is provisioned to end users.
    - For example, a service provider provisions the entire pool to a single tenant, but that tenant is an enterprise with a team of users.
    - The team gets access to the entire pool of GPUs, and temporal scheduling comes into play when multiple team members access the pool simultaneously.



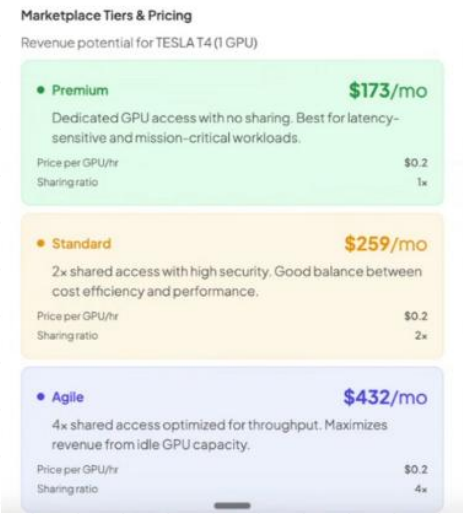
See this video for a detailed explanation of the hosted·ai scheduler.

GPU Mesh infrastructure pricing

GPU Mesh tiers: wholesale pricing

Each tier has a wholesale price per hour, per GPU type, based on a percentage of typical retail GPUaaS market rates. The wholesale Mesh price per GPU tier is set by hosted·ai to reflect those market rates — typically between 50 and 80 percent of retail.

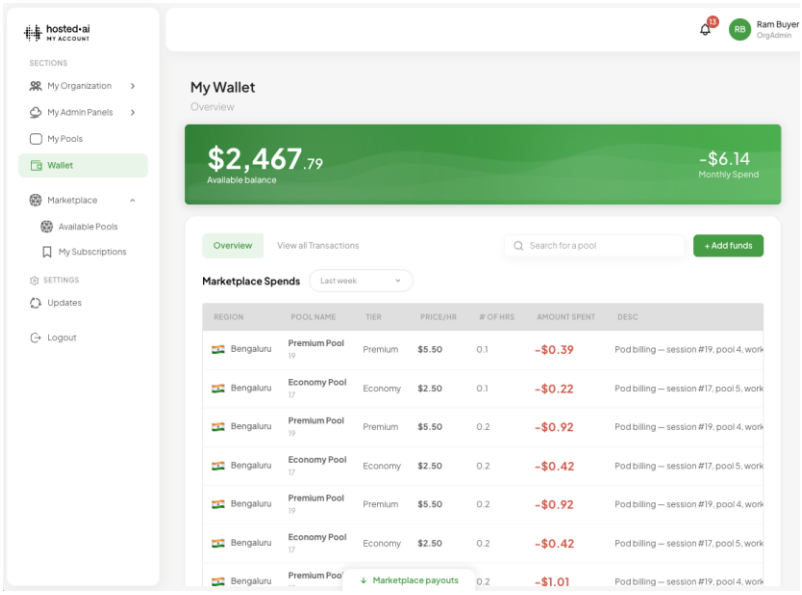
- **GPU Mesh buyers** set their own retail price for their GPUaaS offering, and pay the wholesale price only when their customers consume from the GPU pool.
- **GPU Mesh Suppliers** are paid at the wholesale price when their GPU resources are consumed.



GPU Mesh tiers: paying and getting paid

All GPU Mesh users have a wallet they use for invoices and payments for GPU Mesh transactions.

- **Buyers** use their wallet balance to fund GPU Mesh purchases. Subscribing to locations is free. Buyers are only charged when customers consume GPU resources from the Mesh. Buyers must have credit in their wallet to use Mesh resources. There are various payment and auto-top-up options available.
- **Suppliers** use their wallet to receive payments for GPU Mesh consumption. When the supplier's GPU is consumed via a buyer, the supplier's wallet is credited with the purchase. There are various options for receiving payment, with weekly as the most frequent option, and various options for transferring credits out of your wallet to your destination of choice.



GPU Mesh: how to be a supplier

(publishing and monetizing GPU infrastructure)

This is a high-level view of the process to get your infrastructure published to GPU Mesh, so you can start generating revenue through network.

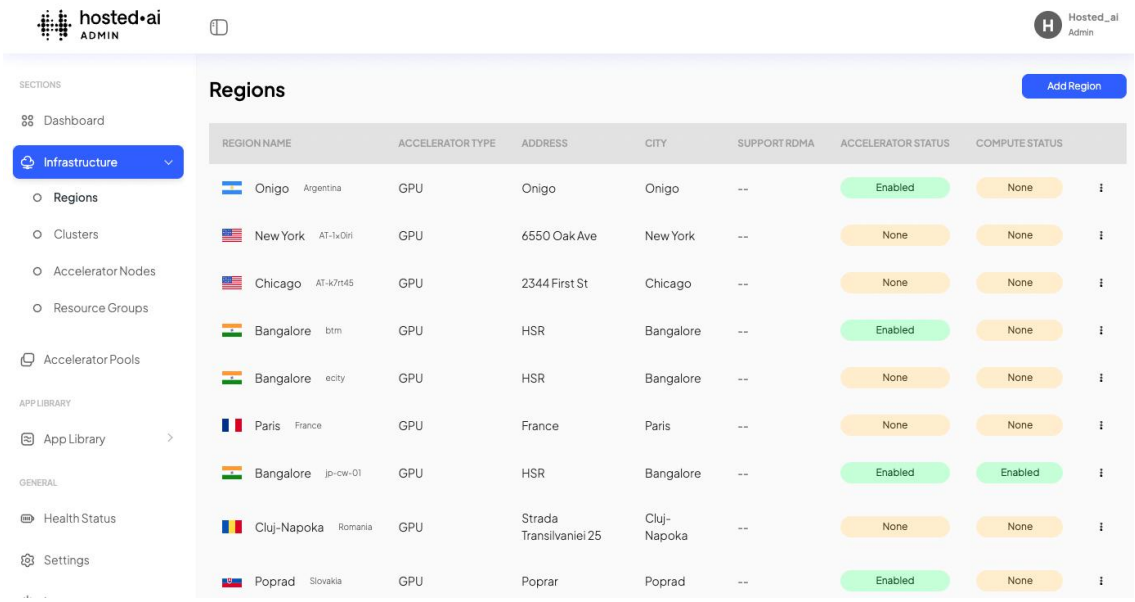
Check out the video links for a more detailed walkthrough, or contact the hosted•ai team for a demo. As part of your customer onboarding process, the hosted•ai team will take you through each of these steps and options to get you up and running as quickly as possible.

1. Deploy hosted•ai, then set up your infrastructure

- The hosted•ai stack can be deployed to a wide range of commodity server hardware under Ubuntu
- Use hosted•ai's infrastructure management tools to create regions for different infrastructure types, selecting the country, GPU type, network fabric type, and other characteristics
- Use hosted•ai's cluster management tools to discover and add physical nodes (GPU, CPU and storage) to each region
- See this video for a detailed walkthrough

2. Add GPUaaS nodes in each region

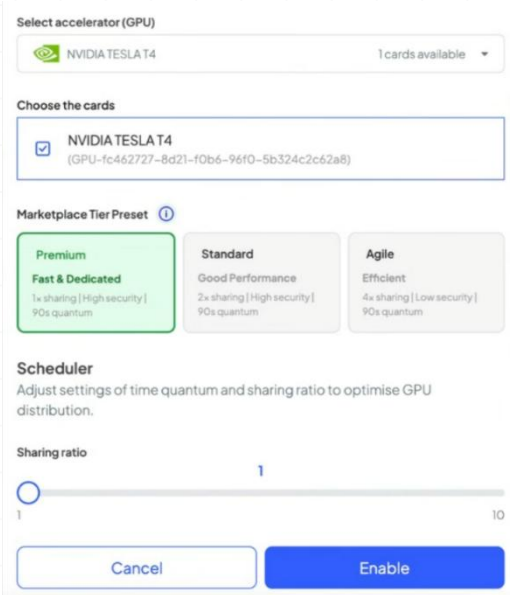
- In the hosted•ai admin panel, add GPUaaS nodes and choose their role
- Initialize the nodes to automatically install the required software and drivers, then enable GPUaaS for each node
- Auto-scan the initialized nodes to detect their GPU, CPU, RAM and storage resources. Collectively the nodes form a Kubernetes cluster used for GPUaaS
- See this video for a detailed walkthrough



GPU Mesh: how to be a supplier
(publishing and monetizing GPU infrastructure)

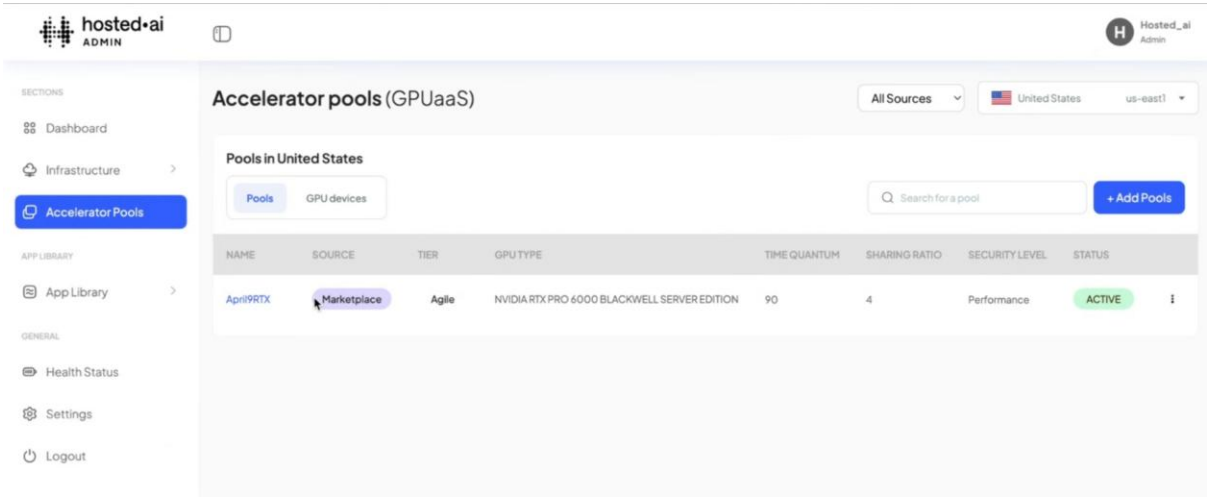
3. Connect to GPU Mesh and create GPU pools for multi-tenant GPUaaS

- Link your GPU Mesh account to your hosted•ai admin panel using a secure key
- In the hosted•ai admin panel, add physical GPUs to virtual GPU pools. A pool can contain any number of GPU cards of the same type
- Then, choose one of the Mesh-compatible presets for scheduling and overcommit: Premium, Standard or Agile for each pool
- Enable the pool to publish it



4. Get paid for consumption

- At a high level, that's it: once you have completed the GPU pool set-up for Mesh, the pools you publish become part of a global demand network, and you are monetizing that capacity with zero effort
- When workloads are deployed using your infrastructure, you are paid for consumption at the GPU Mesh wholesale rates
- There are GPU Mesh SLAs that infrastructure suppliers must adhere to, including notice requirements for removal of the pools you publish
- Any GPU pools you publish to Mesh can be viewed, managed and monitored alongside your own on-premises GPU pools through the hosted•ai admin panel



GPU Mesh: how to be a buyer

(buying and reselling GPU infrastructure)

This is a very high-level view of the set-up needed to sell GPU Mesh infrastructure to your end customers, so you can launch a neocloud without GPU CAPEX.

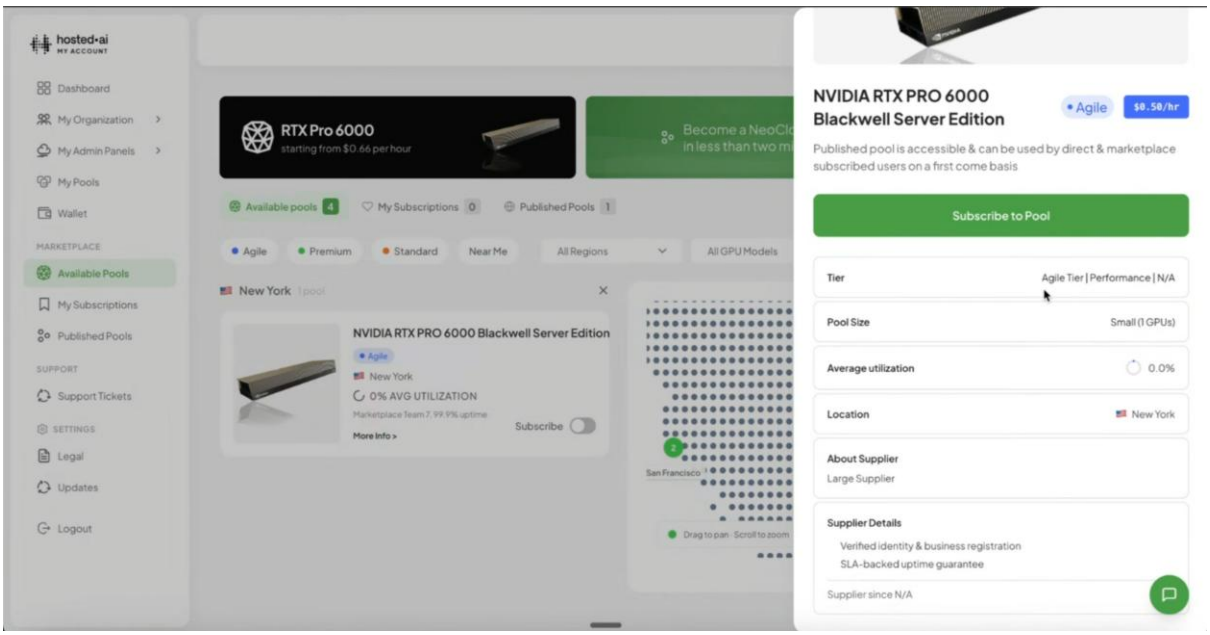
You'll use the same hosted·ai toolset as a provider with their own infrastructure, but without the need to configure physical infrastructure clusters. The physical infrastructure is sourced through GPU Mesh. Check out the video links for a more detailed walkthrough, or contact the hosted·ai team for a demo.

1. Deploy hosted·ai and set up your GPU Mesh account

- The hosted·ai stack can be deployed to a wide range of commodity server hardware under Ubuntu
- The hosted·ai team will create your GPU Mesh account during your customer onboarding process
- Once your admin account is enabled for GPU Mesh, you'll be able to invite other team members too

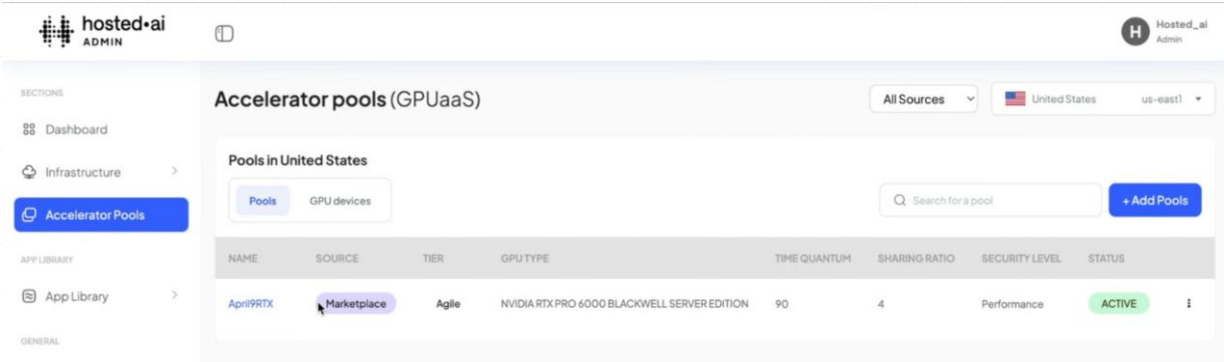
2. Connect hosted·ai to your GPU Mesh account, and subscribe to GPU pools

- You'll need to connect your GPU Mesh account to your hosted·ai admin panel using a secure key
- In your GPU Mesh account, browse the available GPU types and locations, and subscribe to as many as you want. Each location has helpful information about the available capacity, the size of the GPU pool available, and the price
- Any pools you subscribe to are shown in your GPU Mesh account, along with other tools and information to manage your account



3. Set up policies for the GPU Mesh pools you subscribed to

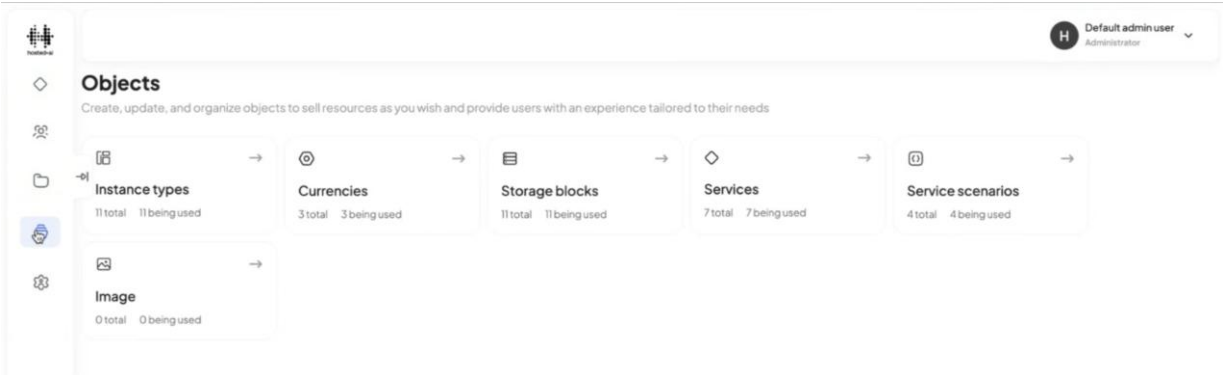
- At this point, you can use hosted•ai admin panel to set up your GPUaaS offering in just the same way you would set up GPUaaS based on your own on-premises GPU infrastructure
- The pools you have subscribed to are shown in your hosted•ai admin panel. This is where you configure how those pools are provisioned to your end users
- Use the tools in hosted•ai to set access policies, pricing policies and more



A full step-by-step is outside the scope of this guide, but this video gives a flavor of the process. For more information, contact the hosted•ai team for a demo.

4. Create users, resource types and policies

- The hosted•ai user panel is where you create profiles and accounts for your users, and configure the resources and policies that apply to their services



A full step-by-step is outside the scope of this guide, but this video gives a flavor of the process. For more information, contact the hosted•ai team for a demo.

Next steps

Using hosted·ai and GPU Mesh, you can launch a neocloud / GPU cloud in a matter of days.

Get in touch for a demo and more information, and start your journey to simple, profitable GPU cloud.

https://hosted.ai/demo

hello@hosted.ai

About hosted·ai

We're a software company with decades of experience in virtualization, cloud, hosting and AI.

- Founded in 2024
- Public launch and first customers onboarded in early 2025
- Our core platform technologies have been in development since 2018

Learn more and get in touch at https://hosted.ai.