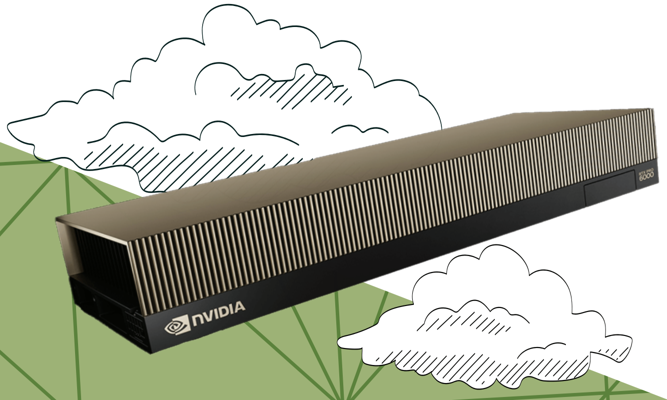
GPU as a Service with Spectro Cloud and NVIDIA

GPUaaS is key to unlocking AI adoption

Most enterprises are looking to leverage NVIDIA GPU-accelerated computing to power today's explosion of AI training and inferencing workloads — at scale and across environments, including the edge.

The end goal for many is delivering GPU-as-a-Service (GPUaaS): the ability to provision and manage GPU resources on Kubernetes on demand, without manual configuration or deep infrastructure knowledge.

With GPUaaS, you can abstract the complexity of hardware and software setup, allowing teams to focus on building and running AI workloads at scale, unlocking a host of use cases:



- **Enterprise AI platforms:** Run and scale AI/ML workloads across cloud, edge, and on-prem with consistent infrastructure and automation.
- **Model development and training:** Give data science teams fast, self-service access to GPU-ready environments for building and training models.
- **Edge and mission-critical AI:** Deploy GPU-powered AI in constrained or disconnected environments with full lifecycle control and security.
- **High-performance AI pipelines:** Optimize throughput and latency with BlueField-3 DPU acceleration for networking, storage, and security operations.
- **AI in regulated environments:** Ensure compliance and security for sensitive workloads in public sector, healthcare, and finance.
- **Retrieval-augmented generative AI:** Connect AI applications to internal data sources for accurate, domain-specific outputs while ensuring security and compliance.
- **Agentic AI with enterprise data:** Deploy autonomous AI agents that reason and act using proprietary enterprise data to automate and optimize business operations.

Making GPUaaS more than a pipe dream

Out of the box, Kubernetes lacks built-in GPU scheduling, monitoring, and lifecycle coordination. To unlock a GPUaaS experience for your users, you need the ability to deploy and configure a broad and complex set of NVIDIA software into each cluster — a task that until today required significant operational effort. The challenges are significant:

- **Complex GPU stack integration:** While NVIDIA provides a powerful and flexible GPU stack, operationalizing it consistently across environments can be time-consuming without automation.
- **Disjointed environments:** Running AI workloads across cloud, edge, and on-prem often leads to inconsistent configurations and fragmented management.
- **Scaling infrastructure reliably:** Expanding and maintaining GPU-enabled clusters with performance, versioning, and governance in mind can be challenging at enterprise scale.
- **Mixed workload environments:** Supporting both modern container-based and legacy VM-based AI tools in the same stack increases operational complexity.
- **Adopting DPUs efficiently:** NVIDIA BlueField offers transformative acceleration, but integrating it into Kubernetes workflows still requires operational tooling and lifecycle integration.

Spectro Cloud: automation for real-world GPUaaS

Now, Spectro Cloud has worked closely with NVIDIA to bring the GPUaaS experience to reality. Our Palette Kubernetes management platform repeatably deploys key NVIDIA software stacks into Kubernetes clusters.

The result is on-demand GPU capability, backed by Palette's automation of the full cluster lifecycle. Palette also supports VM-based AI tooling alongside containers in the same clusters, unifying operations in a single platform.

Palette's GPUaaS capabilities are built for teams running complex, performance-sensitive workloads — particularly in industries with strict security, latency, or operational constraints.

About Spectro Cloud

Spectro Cloud delivers simplicity and control to organizations running Kubernetes at any scale.

With its Palette platform, Spectro Cloud empowers businesses to deploy, manage, and scale Kubernetes clusters effortlessly — from edge to data center to cloud — while maintaining the freedom to build their perfect stack.

Trusted by leading organizations worldwide, Spectro Cloud transforms Kubernetes complexity into elegant, scalable solutions, enabling customers to master their cloud-native journey with confidence.

How it works

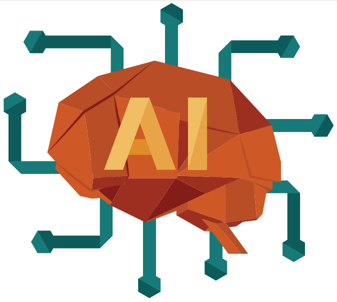
When an application team needs a new GPU-enabled cluster to deploy an AI workload, the platform team needs only to turn to Spectro Cloud Palette.

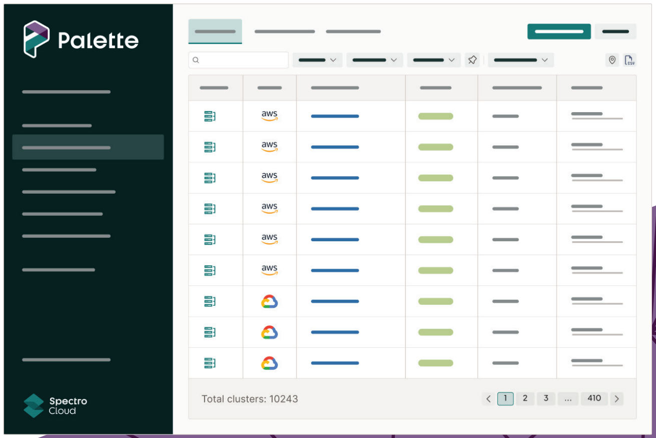
With a few clicks — or via API or Terraform — they can provision one or more clusters on cloud, edge, or data center infrastructure , complete with the full stack: OS, Kubernetes, platform tooling, NVIDIA GPU Operator, GPU Schedulers (such as Kai, Kueue, Armada), NVIDIA NIM microservices and NVIDIA NeMo, and NVIDIA BlueField with NVIDIA DOCA support.

Palette handles the entire process automatically, turning days of work into minutes and delivering a true, on-demand GPU-as-a-Service experience.

Here's how GPUaaS works with Palette:

- **Easily create a GPU stack:** Palette orchestrates full-stack GPU environments using the NVIDIA GPU Operator, including drivers, container runtimes, and device plugins, fully deployed via declarative Cluster Profiles.
- **Run it anywhere, seamlessly:** GPU-enabled clusters can be spun up across clouds, data centers, or edge locations with consistent configs and policies — no manual tuning required.
- **Automate lifecycle management:** GPU components are managed alongside Kubernetes and the OS, with built-in support for upgrades, monitoring, and policy enforcement.
- **Add DPU acceleration (Tech Preview):** When enabled, Palette can also provide BlueField components like DOCA Platform Framework (DPF), and OVN-Kubernetes with hardware acceleration — offloading networking, storage, and security from the CPU and GPU enabling your GPUaaS setups to run efficiently.
- **Benefit from unified, scalable AI infrastructure:** Teams can manage GPU and DPU stacks from a single pane of glass — provisioning, scaling, securing, and tearing down AI clusters as needed to maximize utilization and performance.





Total clusters: 10243

Key benefits of GPUaaS with Spectro Cloud and NVIDIA

With full-stack automation and unified management, platform teams can move faster, reduce complexity, and get the most out of their GPU and DPU investments.

- **Less manual work:** Skip tedious setup with prebuilt automations.
- **Faster time to value:** Launch AI workloads in minutes instead of days.
- **Better hardware utilization:** Make the most of your GPU and DPU investments across cloud, edge, and on-prem.
- **Reduced risk:** Automated lifecycle management cuts human error and closes vulnerabilities faster.

Your first choice for GPUaaS

Spectro Cloud is a proud NVIDIA partner with a proven track record of powering AI workloads for enterprises and public sector organizations.

- **Trusted NVIDIA partner:** We're an NVIDIA Preferred Partner, a graduate of the NVIDIA Inception startup program, and a member of the Jetson™ Ecosystem.
- **Award-winning AI innovation:** We're helping organizations like Dentsply Sirona, GE HealthCare, RapidAI, the US Air Force and many others leverage AI in production, from edge to cloud.
- **First choice for Kubernetes:** Rated Leader in GigaOm's Radar for Managed and Edge Kubernetes two years running, nobody knows K8s like we do.
- **Safe and secure:** From ISO 27001 and SOC 2 to FIPS cryptography, we take your trust seriously.

MANAGED KUBERNETES V.5
GIGAOM RADAR REPORT
LEADER
2025

THE CHANNEL CO.
CRN
AI 100
2025

Get started

To learn more about Spectro Cloud Palette, and to arrange a 1:1 demo about our GPUaaS capabilities with NVIDIA, visit **spectrocloud.com/get-started.**