Spectro Cloud | **NVIDIA**

Manage AI at scale with Spectro Cloud and NVIDIA

One powerful platform to deploy and manage the full AI infrastructure stack, from bare metal to model.

Your partners for the next era of AI

Deploying AI applications at scale is complex. Precious GPUs and other infrastructure often sits underutilized due to complex operations and challenges in maintaining it in production.

Through our PaletteAI platform, we help enterprises and public sector organizations accelerate every stage of AI operations: from hardware provisioning to model management.

We collaborate with NVIDIA to provide a tightly integrated solution across the stack, to support enterprise AI at scale.

Solving the challenges of AI at scale

Working with NVIDIA and bringing together other leading partners across the AI ecosystem, we help you solve the six key challenges common to AI initiatives.

Infrastructure underutilization

With infrastructure guardrails and dynamic node allocation, you can optimize which workloads get which resources.

Production journey

With automated day 2 operations and security, we smooth the transition from experimentation to production.

Architecture complexity

Instead of manually stitching together tools, we provide full end-to-end stacks, from bare metal to ML models, to deploy with just a few clicks.

Hybrid and multicloud scenarios

Because we're environment agnostic, you get the same experience whether your stacks are on prem or in the cloud.

Security and compliance requirements

We tick the boxes with support for FIPS 140–3, access management and networking isolation.

Cost predictability

AI investments already make CFOs wince. We price simply per GPU so you can forecast AI budgets and predict ROI.

What's your use case?

Together with NVIDIA, we help enterprises, neoclouds, service providers and public sector agencies apply AI in four key scenarios. Whatever your focus, we have a solution to fit.

Sovereign AI

Sovereign AI projects are often greenfield, large-scale, and highly security conscious. Our single private platform is ready to help you keep tenants isolated and workloads humming.
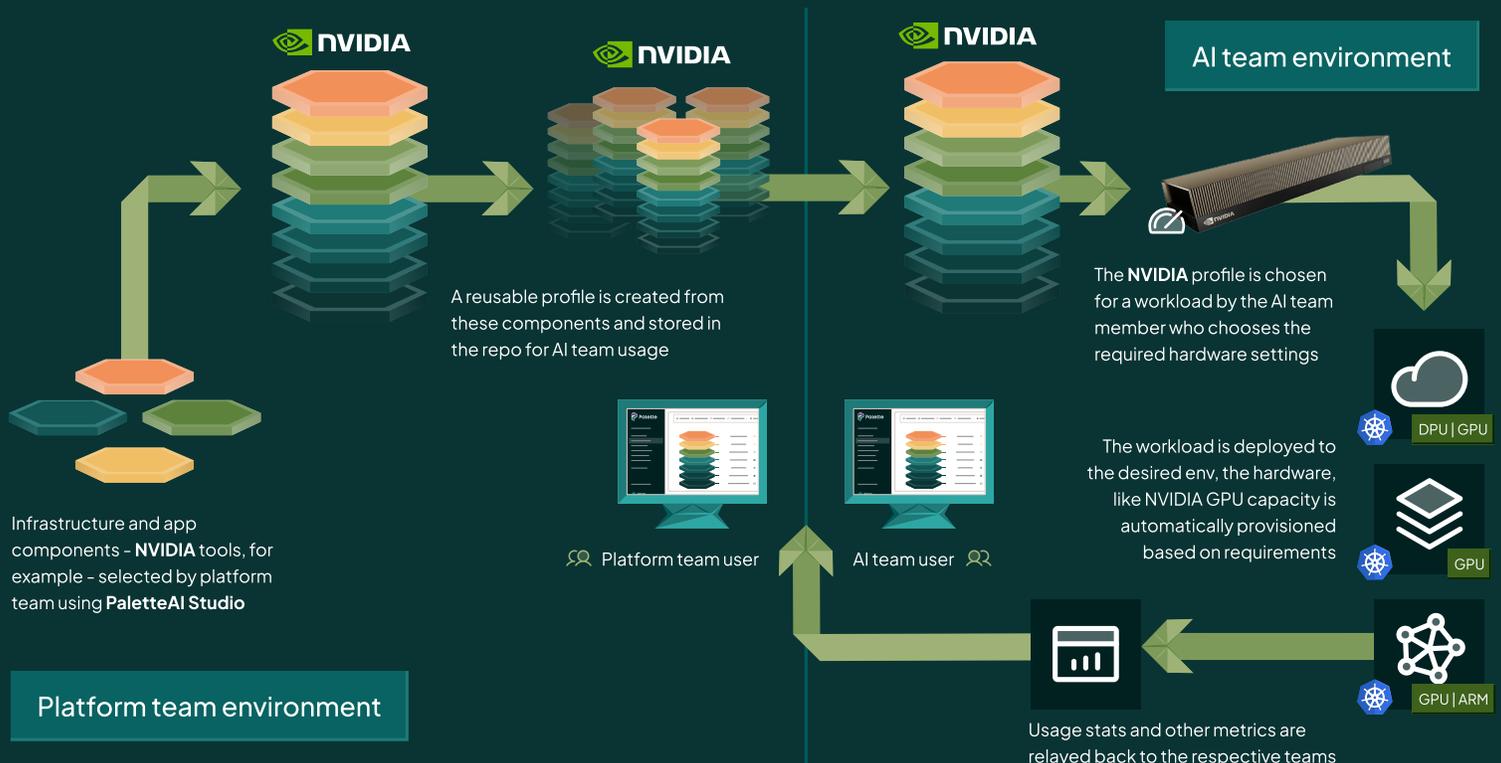
AI Factories

Accelerating AI workflows at scale — that's the purpose of an AI factory. We have a blueprint to help you build and run one efficiently.

AI as a Service

If you're looking to deliver AI capabilities — training, inference, or genAI models — on demand to your users, we give you the simple full-stack control you need.

Edge AI

From computer vision to interactive agents, AI can make a huge difference at the edge where data and users live. We help you deploy it efficiently and securely.

Infrastructure and app components - **NVIDIA** tools, for example - selected by platform team using **PaletteAI Studio**

A reusable profile is created from these components and stored in the repo for AI team usage

Platform team user

AI team user

AI team environment

The **NVIDIA** profile is chosen for a workload by the AI team member who chooses the required hardware settings

The workload is deployed to the desired env, the hardware, like NVIDIA GPU capacity is automatically provisioned based on requirements

DPU | GPU

GPU

GPU | ARM

Usage stats and other metrics are relayed back to the respective teams

Platform team environment

Spectro Cloud

Spectro Cloud and NVIDIA: better together for AI use cases

Our product portfolio integrates at different layers of the stack to ensure a seamless experience to deploy and then maintain AI solutions.

Security

BlueField DPUs & Zero Trust

FIPS 140-3

Security patches within SLAs

Multi-tenancy

Spectro Cloud + **NVIDIA**

AI platform
- Use NVIDIA NIMs with our Model-as-a-Service capability
- Deploy NVIDIA Run.ai within a few clicks

Kubernetes
Get the NVIDIA GPU Operator up and running with a few clicks

Networking
Accelerate AI workloads with Spectrum-X & PaletteAI

Infrastructure
- Follow NVIDIA AIDP blueprint within PaletteAI
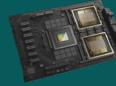- Use PaletteAI with NVIDIA Base Command Manager

Platform
- Seamlessly enable the use of NVIDIA CUDA
- Ensure immutability through Kairos

NVIDIA GB100   NVIDIA H100   NVIDIA B100   Kairos

A trusted force in AI infrastructure

At Spectro Cloud, we power application infrastructure for some of the world's largest enterprises: from GE HealthCare and Airbus, to T-Mobile and Yum! Brands.

They trust us because we take seriously the things that matter to them: security and compliance. Openness. Partnerships. Learn more about our commitments at spectrocloud.com

Next steps

To find out more about our work with NVIDIA and our solutions for AI, and to book a 1:1 meeting to discuss your projects, visit **spectrocloud.com/nvidia**.