Spectro Cloud | PaletteAI VerteX

GPU optimization for mission-ready AI infrastructure

Getting more from every GPU dollar: from loading dock to live, and every day after

The GPU utilization problem

GPU infrastructure is the most expensive line item in most AI budgets. A single NVIDIA DGX system costs hundreds of thousands of dollars. For defense agencies and government programs running dozens or hundreds of GPU nodes across classified and unclassified environments, the investment is enormous — and the pressure to demonstrate ROI is real.

Most of that capacity goes underused. Industry benchmarks consistently show average GPU utilization rates of 30–50% across enterprise environments. Government agencies often fare worse: static allocation policies, lengthy procurement cycles, and manual provisioning workflows mean GPUs sit idle for weeks or months between workloads.

GPU optimization is really two problems. The first is time to value — how quickly new hardware goes from loading dock to running production workloads. The second is sustained utilization — how effectively GPU resources are shared, scheduled, and monitored once they're live. Most organizations struggle with both.

70%

GPU utilization improvement with PaletteAI VerteX

5 clicks

To deploy a pretrained or custom model

30 days

Classified AI factory: loading dock to production

Source: Spectro Cloud customer deployments

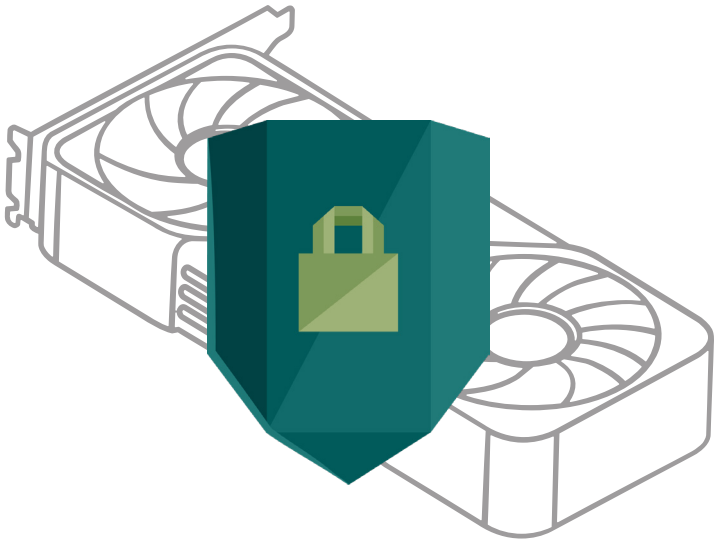
Scenario 1: New deployments — loading dock to live

Standing up a production-ready AI environment in a classified or air-gapped network is a notoriously slow process. Integrating GPU drivers, CUDA runtimes, Kubernetes orchestration, AI frameworks, high-speed networking, and security controls requires deep expertise and careful sequencing. When each host has to be configured individually (which is common in defense environments) a single deployment can consume months of engineering effort.

PaletteAI VerteX takes a different approach. Complete GPU environments — OS, Kubernetes, GPU operator, drivers, AI frameworks, networking, security policies — are defined as declarative full-stack blueprints. Once a blueprint is validated, it can be deployed consistently to bare-metal servers, edge nodes, or cloud instances. No hand-configuration. No node-by-node variation.

The platform auto-discovers GPU hardware attributes (model, family, memory, count, MIG capability) before cluster formation, enabling smarter scheduling decisions from the start. One-click GPU operator deployment handles driver installation, monitoring, and lifecycle automatically.

For defense programs operating under tight timelines, this is the difference between months of manual integration and a repeatable, auditable workflow that gets hardware operational in days.



In the field: classified AI factory in 30 days

A federal technology provider had spent ten months hand-configuring an AI environment built on NVIDIA HGX systems and Spectrum-X networking for a classified government program. Configuration varied from node to node. The team couldn't trust the environment enough to declare it production-ready.

With 30 days left before go-live, they switched to PaletteAI. The platform captured the full environment — bare-metal provisioning, networking, storage, and the Run:ai workload layer — in a single declarative blueprint. The day before go-live, when a full cluster rebuild was needed, the team completed it without breaking a sweat. GPU capacity was then doubled in days, with immediate full utilization.

Read the full story: **bit.ly/classified-story**

Spectro Cloud

Scenario 2: Optimizing live GPU resources

Once GPU hardware is deployed, the challenge shifts to sustained utilization. In most government environments, GPU nodes are statically assigned to individual teams or programs. Reprovisioning takes too long, so teams reserve capacity they don't always need. The result: expensive hardware sits idle while other programs wait in queue.

PaletteAI VerteX addresses this through several concrete mechanisms.

Shared GPU pools and GPU-as-a-Service

Rather than locking GPUs to specific teams, PaletteAI VerteX turns GPU clusters into shared pools that multiple teams access on demand. AI practitioners request GPU resources through pre-approved, self-service templates. Platform teams define the guardrails — hardware profiles, security policies, approved software stacks — and the platform handles provisioning automatically. This shared model is what drives the 70% utilization improvement: GPUs become elastic resources instead of fixed allocations.

Intelligent autoscaling

Autoscaling policies dynamically adjust GPU, CPU, and memory allocation in real time based on utilization thresholds. When a training workload spikes demand, nodes scale up. When a job completes, they scale back down. A distance-minimizing algorithm selects which machines to add or remove, optimizing utilization while minimizing churn and workload disruption. Cooldown periods prevent thrashing between scaling events.

GPU-aware node selection

When provisioning clusters, PaletteAI VerteX avoids wasting GPU resources on non-GPU roles. The platform prioritizes allocating GPU nodes to GPU workloads first, then fills remaining capacity with CPU-only workers. If no CPU-only workers are available, it selects GPU machines with the lowest GPU count — keeping high-value accelerators available for the workloads that need them.

Multi-tenant quota management

A layered quota system prevents any single team or program from consuming all available GPU capacity — a real concern in environments with shared superpods or multi-tenant clusters. Quotas can be set at the namespace, project, or tenant level, scoped per GPU family (for example: 48 NVIDIA H100s for one project, 12 A100s for another). Lower-level resources can't exceed the limits set above them.

Utilization monitoring and energy visibility

GPU utilization metrics are sourced from DCGM and node exporter, exposed through Grafana dashboards for real-time and historical visibility. Teams can see exactly where GPUs are working hard, where they're idle, and how energy consumption tracks over time. Bare-metal power management capabilities let organizations physically power down unused GPU servers and restore them when demand returns — cutting energy costs without sacrificing readiness.

A partner you can trust

Meeting the standards you expect

PaletteAI VerteX is Spectro Cloud's FIPS 140-3 validated, self-hosted edition — purpose-built for environments that demand the highest levels of security and compliance. It operates fully in air-gapped environments with no dependency on external connectivity, and enforces zero-trust architecture with policy-driven network segmentation and workload isolation by default.

We're a Silver member of the Agentic AI Foundation and hold Certified AI Platform status from the CNCF. Spectro Cloud holds SOC 2 Type II certification and ISO 27001:2022 accreditation. For defense and intelligence programs, this means AI infrastructure that meets security requirements from day one — without forcing teams into rigid, locked-down platforms that limit what they can build.

That's why we're trusted by the US Army, Navy and Air Force, Airbus Defence and Space, and many other public sector and commercial organizations.

Integrated with the full NVIDIA AI stack

GPU optimization depends on tight integration across the full hardware and software stack. Misconfigured networking, incompatible driver versions, or poorly tuned storage can bottleneck GPU throughput just as effectively as underallocation.

Spectro Cloud is an NVIDIA preferred partner, validated to the NVIDIA AI Factory for Government reference architecture. Through our PaletteAI VerteX platform, we you can manage the complete NVIDIA accelerated computing stack as a single, version-controlled blueprint: GPU operator, NVIDIA NIM for inference, Run:ai for workload scheduling, DOCA for DPU-accelerated networking and security, BlueField DPU support, and Spectrum-X high-speed networking. When these components are deployed and lifecycle-managed together — rather than hand-integrated one by one — GPU resources can actually reach the utilization levels the hardware is designed for.

Ready to talk GPU optimization?

Whether you're standing up new AI infrastructure or looking to get more from what you already have, we'd welcome the conversation. Spectro Cloud offers guided demos, proof-of-concept engagements, and architecture workshops tailored to government and defense requirements.

Request a demo: **spectrocloud.com/get-started**

Contact our government team: **gov@spectrocloud.com**

Spectro Cloud