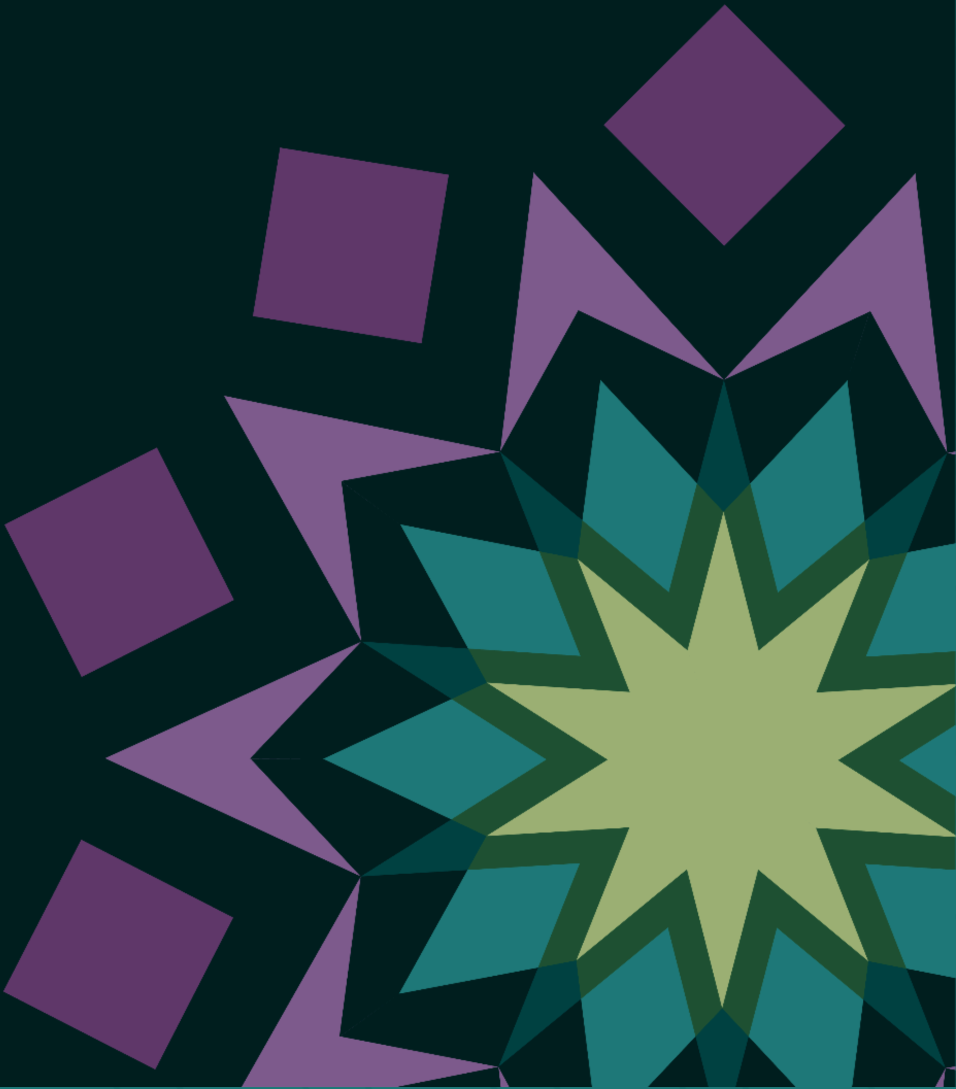
Spectro Cloud

PaletteAI

Building Enterprise AI Factories with PaletteAI

The factory assembly line

V1.1, March 13, 2026



Version history

Version	Date	Changes
V 1.0	October 28, 2025	Initial release
V1.1	March 13, 2026	Minor updates

Contents

Introduction

AI is changing the world around us — fast. Generative AI is a key technological driver for innovation. According to Gartner[1], 80% of existing applications will be refactored to include embedded AI.

For those in the infrastructure and platform space, it means applying their expertise in new and exciting ways, designing and building the cornerstone hardware and software into platforms our users – developers and data science teams – need to drive that innovation.

These platforms need to be ready to roll *yesterday*, be highly secure, be fully utilized, while offering a superb experience to their users. But platform teams face pressure and technical challenges to deliver – a hefty challenge for sure.

We've learned from internal developer platforms – applying those lessons learned to internal AI platforms. Rigid standardization doesn't work; platforms should enable data science team flexibility, maximize their productivity and maximize hardware utilization, while balancing those needs with guardrails for security, compliance and cost.

Like physical factories that powered the industrial revolution, **AI Factories** drive the AI revolution, but instead of producing physical goods from raw materials, AI Factories transform data and electricity into intelligence, insights and tokens with great scale and efficiency. Every enterprise needs an AI Factory to deliver fast, repeatable, flexible, and efficient AI outcomes.

In this white paper, we'll look at what these challenges platform teams face, why they are the way they are, and how to address them, turning bespoke and brittle AI infrastructure into smoothly-running AI Factories.

Let's dive in.

[1] 2024 Gartner AI Survey

Section 1: How did we get here?

The rate of change of the AI space is incredible. New models drop weekly, new technologies are announced daily, and the ecosystem of tools is large and sprawling, growing at an incredible pace.

AI ecosystem is now more than 2.5× cloud native's size
CB Insights , Stanford AI Index Report, other supporting research

Enterprises face significant challenges in designing, scaling, and optimizing these systems, often requiring specialized expertise and substantial time and financial investment building systems from scratch.

Underneath the huge diversity of impressive AI models and innovative AI apps lies a complex cloud-native infrastructure stack, built on containers and Kubernetes. The ecosystem that powers the AI boom is truly mind-boggingly large and complex.

The rise of AI-optimized data centers

Organizations of all kinds and sizes are taking advantage of AI across industries, including government and defense, and have given rise to a new breed of sovereign clouds, neoclouds and GPU-as-a-service and AI service providers.

Many enterprises are not just turning to hyperscalers or neoclouds for their compute resources. They are building entirely new data center facilities, with millions of dollars in hardware and facilities; one of the biggest infrastructure building booms in history.

Compounding complexity

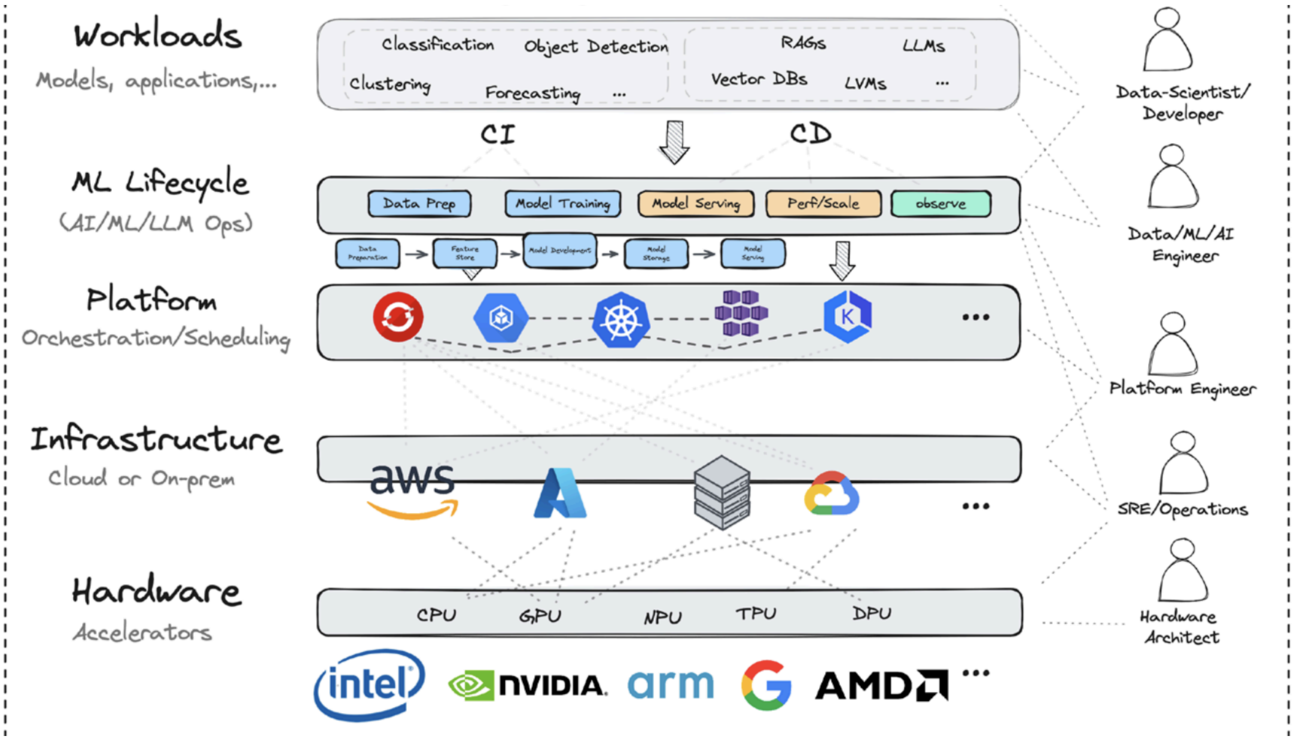
Within those facilities, a radically different data center architecture has emerged to optimize for AI: from GPU and DPU hardware acceleration, physical networking and storage, operating systems, Kubernetes, overlay and multi-tenant networking, to petabyte-scale storage and many packages and frameworks used by AI/ML practitioners across the ML lifecycle.

The road to AI factories is full of engineering challenges, which can be complex, time-consuming, and resource-intensive to solve.

Time to market

Building AI factories from scratch can take months, especially when coupled with the interoperability challenges inherent in integrating hundreds of components that all need to be precisely configured, integrated and tuned.

And when platforms are too slow to deliver, or don't deliver at all, data science teams resort to 'shadow AI', deployed without the blessing of IT and without proper controls — sacrificing governance, security, and ROI.



And while cloud-native approaches and Kubernetes enable repeatable and scalable deployments of AI workloads by decomposing AI infrastructure and workflows into modular components, the pain of 'integration hell' is felt deeply by platform teams building AI platforms.

"When you put everything together, it never works. Or as I like to call it, integration hell."
Shane Wighton (Stuff Made Here):

Talent shortage

Operating in the complex AI landscape requires deep systems expertise across a wide range of domains, from networking and storage to GPU drivers and Kubernetes.

Integrating so many complex components often leads to technical shortcuts, stability issues, duct-taped quick fixes that break at every update and configuration change, usability problems, and even glaring security issues. This all makes it hard to maintain the platform's security, usability and user experience, performance and resource utilization.

Accelerated infrastructure for AI is a precious and in-demand resource, especially GPUs. However, the sheer complexity of these environments is a major cause of deployments taking weeks or months, all while the capital investment in AI hardware isn't returning a value.

> *AI adoption gap: 95% of Generative AI pilots fail to show ROI because enterprises struggle with infrastructure complexity to operationalize models.*
> *MIT Sloan Management Review*

With this much complexity and so many technologies involved, AI practitioners cannot be expected to have deep, specific infrastructure skills needed to deploy their AI functions and workloads, leading to the need for guardrails and help by platform teams.

> *Kubernetes skills gap: 78% of IT roles now demand AI skills — yet enterprises report critical shortages in LLM, gen-AI, ethics and security expertise. Platform teams understand K8s; AI teams don't — leading to friction and inefficiency.*
> *Source: AI Workforce Consortium*

Cost efficiency

Even when deployment succeeds and AI platforms are in use, making sure the hardware resources are efficiently utilized is critical to the success of AI projects and investments.

> On average, 30–70% of expensive GPU and DPU accelerators resources sit idle.
> Source: SemiAnalysis

Often caused by lack of flexibility to deploy workloads in shared clusters, poor resource sharing and scheduling and lack of true multi-tenancy solutions, expensive GPUs are often underutilized, leading to lower ROI than predicted.

It’s the business outcome, stupid

The cocktail of a fast-changing landscape, complex hardware and software stacks and the highly-specialized skills required often lead to sub-optimal results. With platforms taking longer than expected to be deployed, or having usability, stability, performance, utilization or security issues, many AI projects struggle to deliver on their capital-intensive ROI.

Because many projects are so strategic to the business, the pressure is on platform teams to overcome these challenges. They need to deliver these platforms yesterday, with top-notch performance, utilization, user experience, and security.

As we’ve seen, getting there is no small feat, and platform teams can use a bit of help getting them there – so that they don’t have to deliver a brittle platform held together by duct-taped components, but deploy a smoothly-running AI Factory instead.

Constituents of good

But what are the constituents that make AI factories so good?

Enterprises simply look for guidance to get more value from their AI investments (across hardware, software and engineers) faster to decrease the time-to-market of their innovations. By adopting an AI factory approach, businesses streamline AI infrastructure, reduce complexity, close skills gaps and drive engineering productivity.

To use a factory-related comparison: AI stacks are deployed through an ‘assembly line’ of sorts: highly standardized declarative profiles of pre-validated designs containing the desired states of the entire stack, including pre-integration of all hardware and software components. They remove the variability and guesswork of manually deploying and integrating components.

This approach results in a repeatable, consistent and quick deployment, massively reducing deployment time, giving AI practitioners earlier and full access to environments, massively improving the time-to-value for these platforms.

By using these validated design blueprints, organizations don’t have to re-invent the wheel, reducing the complexity of design and depth of skills required to make the deployment and integration of components happen. Automation and declarative desired-state deployments mean they’re repeatable – ensuring every cluster is provisioned consistently and in line with security, governance and operational best practices, further increasing usability, stability, performance and utilization.

Following these designs frees up data science teams – letting them focus exclusively on their own domain, removing the need for them to spend time on Kubernetes and other infrastructure toil, increasing their productivity and agility, freeing up time to spend on keeping up with new technologies and ecosystem innovation.

NVIDIA is the market leader in Enterprise AI, and its Enterprise AI Factory validated designs help organizations build AI factories that are cost effective, scalable and high-performing.

Together with its ecosystem partners, they deliver a full, validated stack of integrated accelerated computing, high-performance networking and AI software to eliminate the burden of designing and building these systems from scratch, mitigating deployment risks and enhancing time-to-value.

It integrates with existing enterprise systems, data sources, and operational infrastructure, to create a seamless experience for AI practitioners, while following enterprise guardrails for security, operational control and corporate governance.

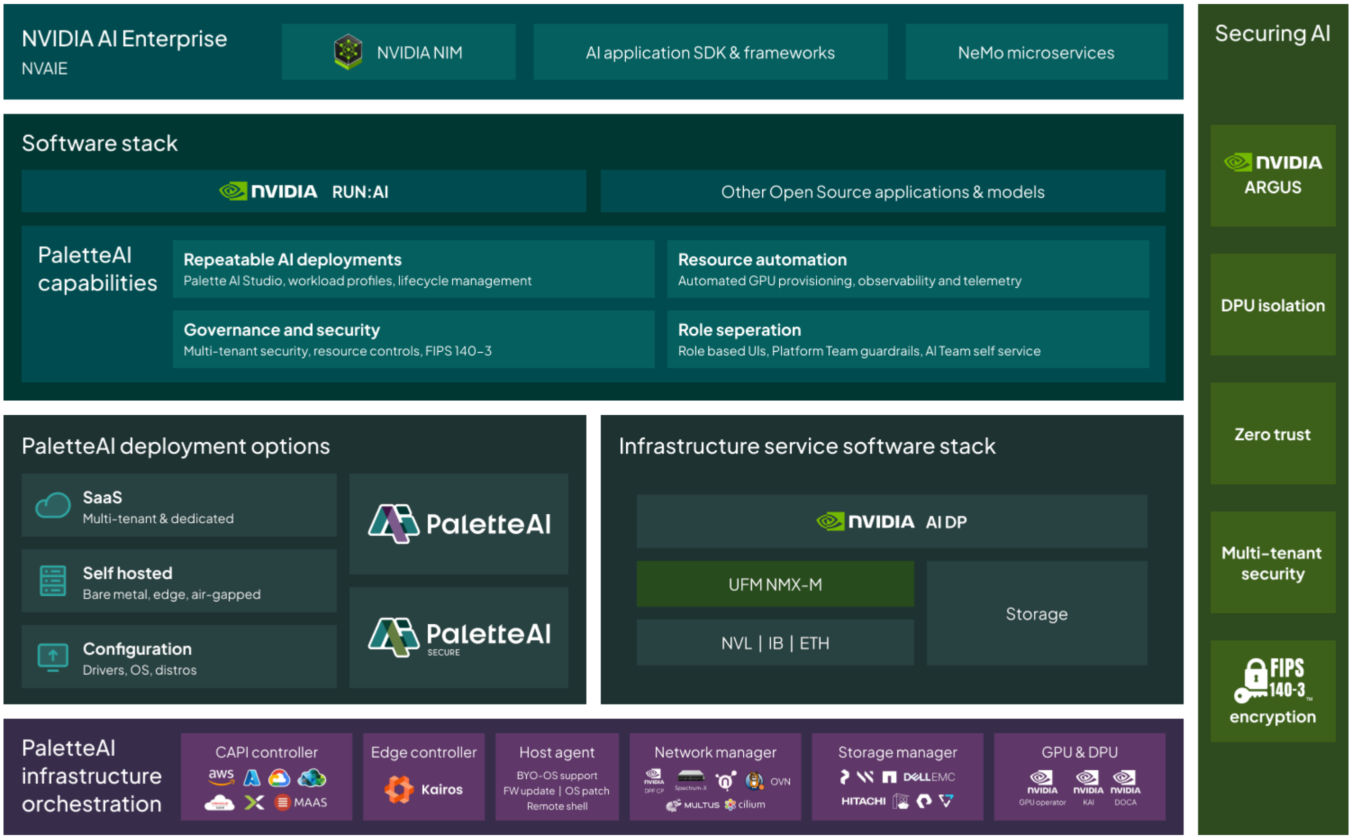
In the following section, we'll look at how PaletteAI, the assembly line for Enterprise AI factories, helps bring this vision to life.

Section 2: Introducing PaletteAI

With PaletteAI, enterprises can build repeatable, consistent and secure AI infrastructure while maintaining operational control, while deploying, managing, and scaling AI workloads. PaletteAI is a unified platform that bridges the gap between Kubernetes infrastructure and AI workflows. It abstracts away cluster provisioning, GPU guardrails and other infrastructural toil from data science teams.

Catering to both platform teams and AI practitioners, PaletteAI deploys and manages both AI infrastructure and the workloads running on top, spanning data science and AI engineering.

It combines automated infrastructure orchestration with workload scheduling and deployment, allowing platform teams to provide governance guardrails and codify technical best practices, while giving AI practitioners benefits from repeatable, consistent and toil-free deployments of their applications and workloads, seamlessly integrating with existing MLOps stacks and lifecycles.



PaletteAI seamlessly integrates NVIDIA's validated designs for AI Enterprise Factories, enabling enterprises to bring NVIDIA's vision to life with full support for the entire NVIDIA enterprise suite across Kubernetes, storage, networking and GPUs while slotting into corporate IT best practices, including zero-trust security powered by NVIDIA BlueField DPUs and the DOCA framework.

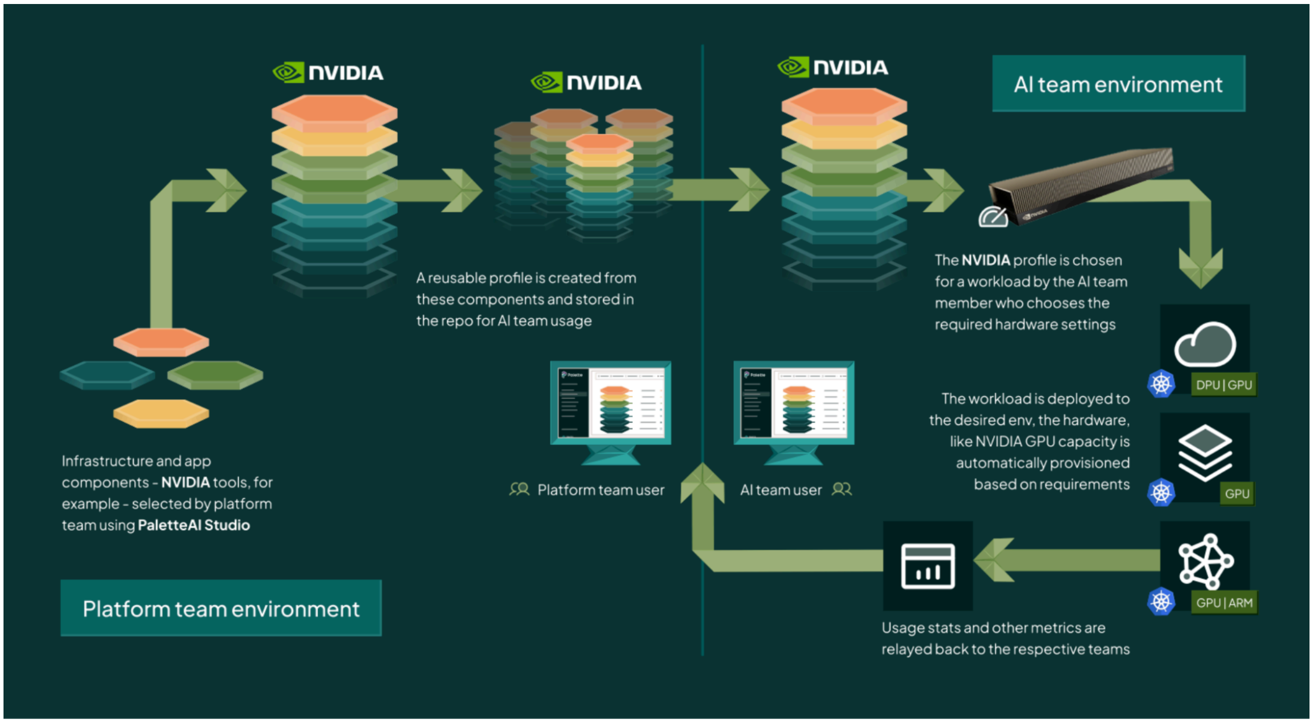
Platform teams have complete operational control over AI infrastructure with multi-tenancy support (SSO, RBAC, etc), cost and resource management, policy-based cluster support (making clusters available for multi-tenant, single tenant or even app-exclusive workload deployments), infrastructure and workload observability, as well as declarative lifecycle management cluster updates and configuration changes.

With repeatable deployments for AI workloads and applications including NVIDIA NIM and Triton for inferencing and NVIDIA NeMo for Agentic AI, as well as industry standard tools like Run:AI,ClearML and KubeFlow, PaletteAI provides AI practitioners with a seamless experience across all phases of the ML lifecycle with repeatable, consistent provisioning of applications for AI practitioners across the enterprise, from AI factory to AI edge.

In short, platform teams can design guardrailed templates and policies, while AI teams deploy freely within approved boundaries.

Section 3: How PaletteAI works

PaletteAI brings together platform and practitioner workflows into a unified system for building, deploying, and managing AI environments with consistency, security, and scale through specific workflows for each, focused on the work that needs to be done.



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

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

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

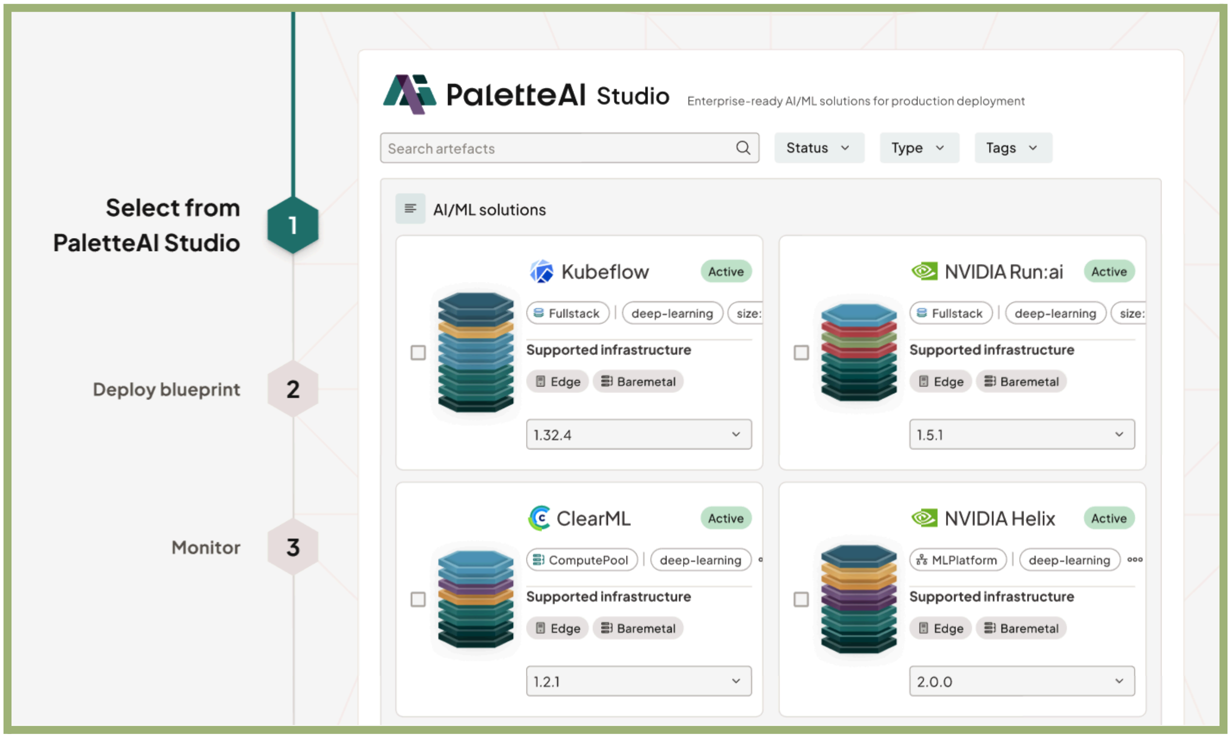
Platform teams define complete cluster configurations that span the full stack from the operating system and Kubernetes distribution to storage and networking, as well as GPU and DPU drivers, NVIDIA Operators and more, combining these layers into a validated, ready-to-go deployment for AI-accelerated infrastructure. Platform teams then use these profiles to deploy hardware stacks with the correct software suite, making them available for consumption.

Additionally, platform teams make available AI and ML tools in PaletteAI Studio, enabling AI practitioners to deploy these tools quickly in a repeatable, consistent manner to Kubernetes clusters of their choice. These deployments include enterprise guardrails for security, operational control and corporate governance put in place by platform teams, so AI practitioners do not need to busy themselves with infrastructural toil, but instead can focus on the customizations they need, deploying the right AI/ML workloads and models to the right cluster.

PaletteAI Studio

The AI stack is not only split up along the ML lifecycle axis of data prep, experimentation, model training, model serving and inference, but also fragmented across an immense number of possible products and open source projects for each. Organizations require a catalog of out-of-the-box workloads, need customizability to fit workloads to the organization's specific model pipeline needs, and need the ability to roll your own workload definitions.

PaletteAI caters to all of the vastly varying needs of different data science teams across enterprises through the PaletteAI Studio, a catalog and customization engine to make sure organizations can tailor workloads to their needs.



Instead of forcing AI practitioners to build container images or write Kubernetes YAML, they can now solely focus on the ML Ops lifecycle.

By exposing only those infrastructural choices that matter to them (such as deploying a model to a shared or dedicated cluster), data science teams no longer need to know the nitty-gritty details of Kubernetes and infrastructure.

This, in turn, unlocks self-service capabilities for data science teams so that they're no longer dependent on platform teams to deploy and operate AI and ML application stacks while delegating the underlying infrastructure to platform engineering teams asynchronously.

PaletteAI plays into a wide variety of use cases that benefit from decoupling infrastructure provisioning from consumption, including consistent deployment and AI model serving across environments in cloud and edge locations and MLOps automation and continuous training and deployment workflows for faster experiment iteration.

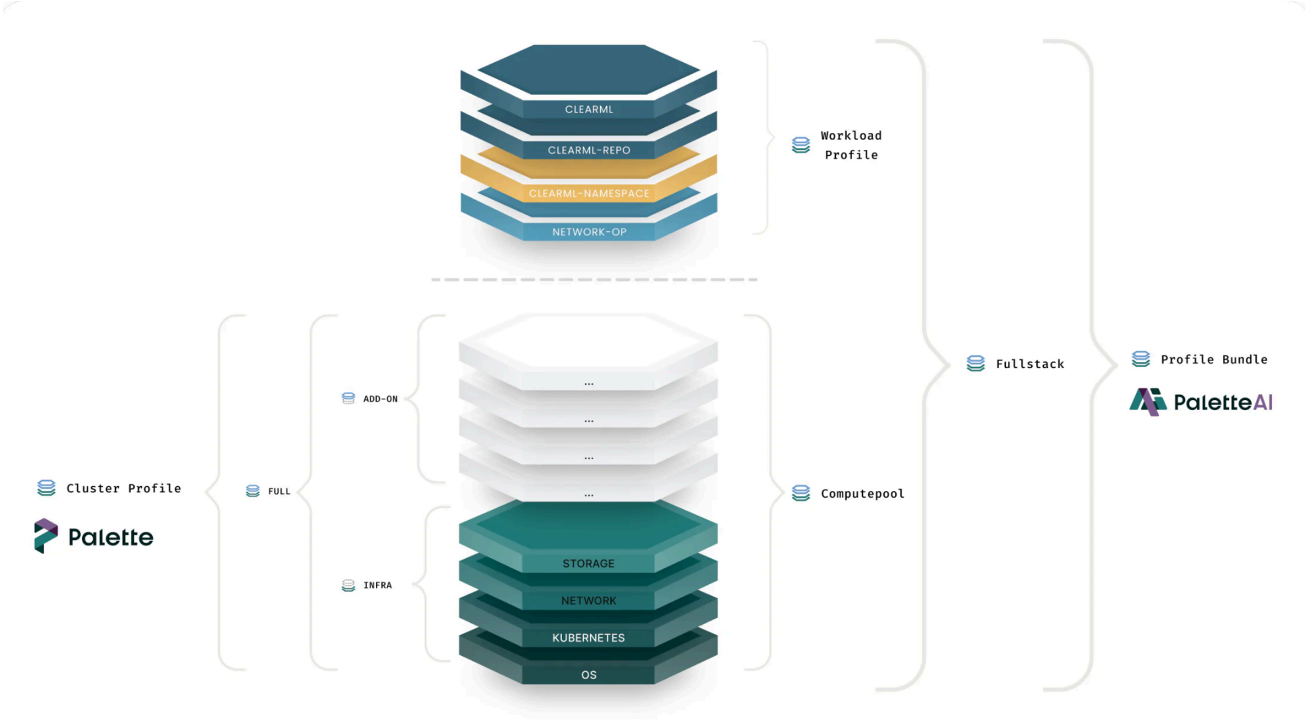
Security and multitenancy

PaletteAI uses a hierarchical organizational structure with Tenants and Projects to manage access control and resource organization for ML/AI deployments. Platform admins can use these to isolate clusters to users, create shared clusters, and apply role-based access control to separate tenants, teams or even individual users.

On the hardware side, PaletteAI supports enforcing zero-trust security powered by NVIDIA BlueField DPUs and the DOCA framework. To obtain GPU multi-tenancy, PaletteAI integrates with leading solutions such as Netris or Aviz.

Profile bundles

PaletteAI manages provisioning and lifecycle of AI infrastructure, from operating system, Kubernetes, networking, storage, NVIDIA drivers to Operators through Profile Bundles. These are declarative configurations, consisting of many layers combined into a single 'stack'.

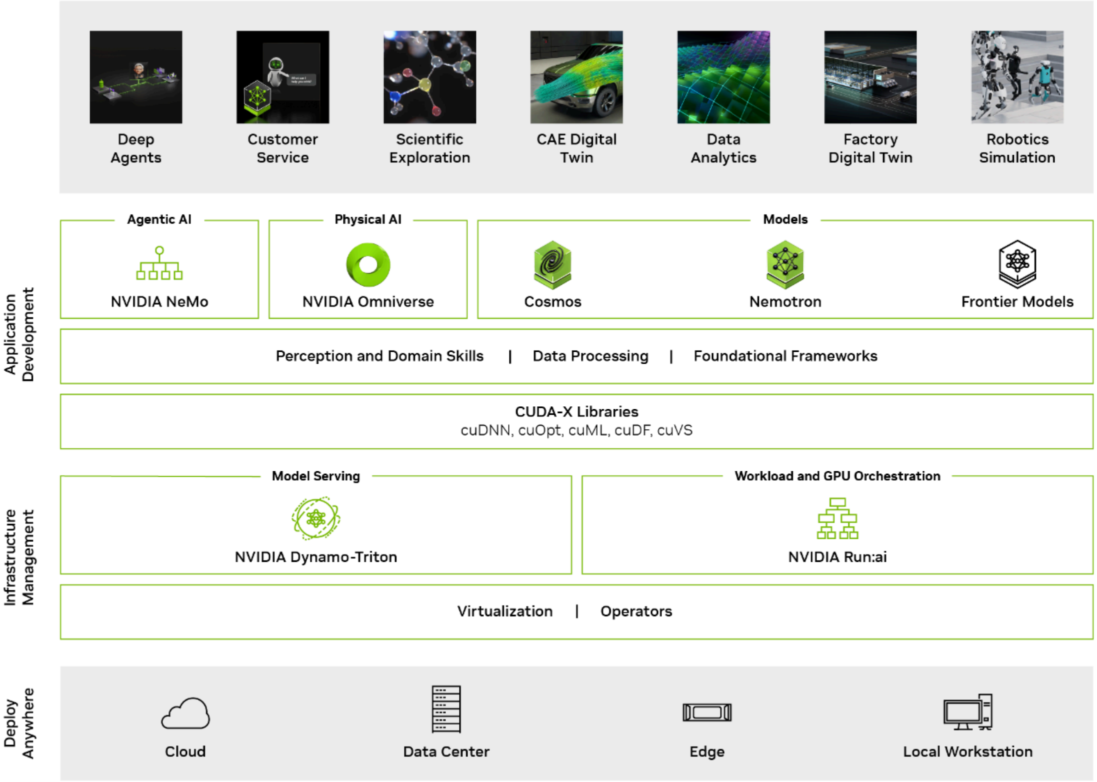


Additionally, clusters are linked to their original profile. Any changes made to the original profile are reconciled, allowing one-to-many configuration management using a single cluster profile to manage many clusters. Platform engineers can configure different clusters for different teams to use, and each cluster operates in shared or dedicated mode for optimal control over utilization and security.

PaletteAI ships with various ready-to-go stacks to choose from, including the NVIDIA Enterprise AI Factory validated design to get started quickly with validated hardware stacks. Alternatively, platform teams can mix and match layers into a full stack, and customize existing stacks.

NVIDIA AI Enterprise components

The product's reconciliation engine applies these full configuration stacks to a hardware environment, provisioning the entire software layer, configuring the hardware (including DPUs), forming a Kubernetes cluster and installing the correct AI plumbing, such as the NVIDIA Network and GPU Operators, NIM microservices, NeMo and other components for NVIDIA-based AI factories.



Workload profiles

Similarly, PaletteAI manages provisioning and the lifecycle of the AI workloads through WorkloadProfiles. These are declarative configurations that model how applications are deployed on a Kubernetes cluster and combine aspects important to platform engineers relating to security, resources, performance best practices with aspects important to AI engineers.

By abstracting away the infrastructural details from data science teams, they no longer have to care about cluster specifics or Kubernetes in general; they simply can deploy applications with a minimal set of deployment variables.

Which deployment variables are exposed to data science teams is determined by platform teams, who have the flexibility to specify defaults and options. This powerful separation of responsibilities helps data scientists deploy the apps they need quickly and without dependence on platform teams, while platform teams can make sure the right enterprise guardrails are present.

With only a minimal set of deployment variables, for instance which version of the application, and which cluster to deploy to, deploying complex AI workloads has never been easier.

By fully adopting a GitOps approach, PaletteAI workload configurations are versioned and available for deployment via both the kubectl command and the web interface. PaletteAI exclusively uses Kubernetes Custom Resource Definitions (CRDs) to power the overall functionality of PaletteAI, meaning all PaletteAI state is stored in Kubernetes — not in a separate database. Each workload is a deployed instantiation of a particular workload profile and deployed applications are tied to a versioned WorkloadProfile, enabling PaletteAI to reconcile revisions of the profile with the deployed application.

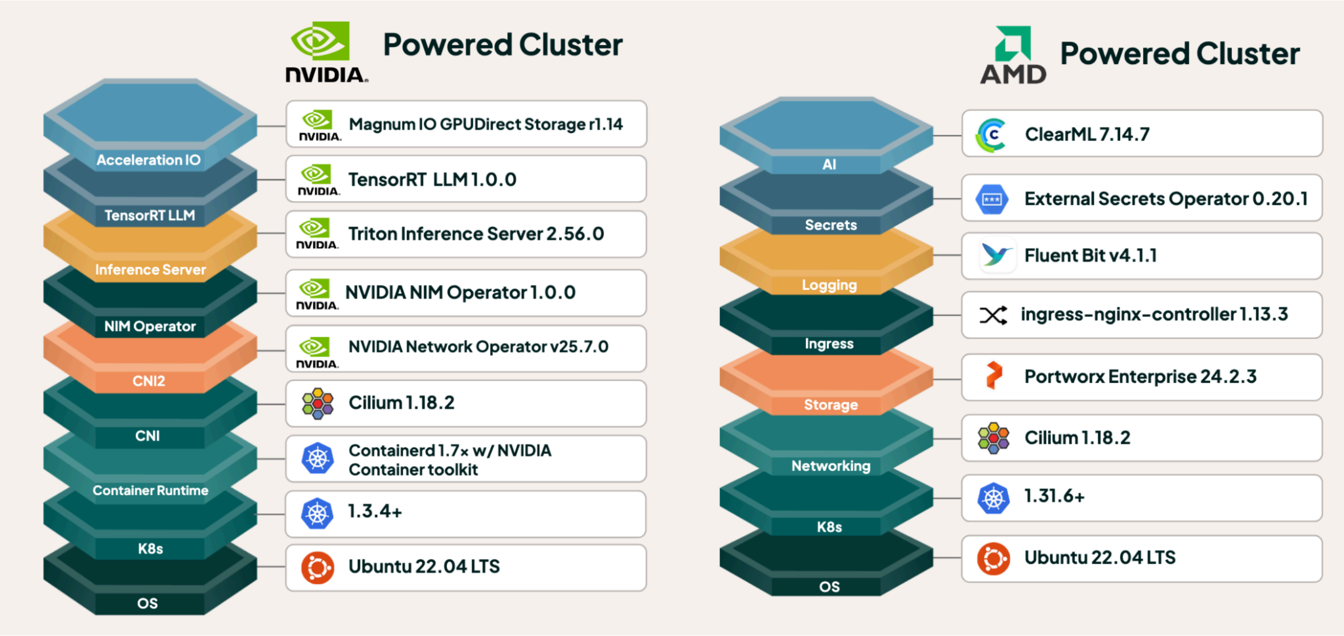
Multi-cluster workload placement

PaletteAI comes with multi-cluster scheduling support that allows AI practitioners to select where their workload must run based on hardware characteristics, the type of workload and the physical location clusters are located in.

Putting it all together: an example

When we put it all together, we end up with the ability to deploy an AI/ML workload end-to-end, from hardware and core infrastructure software all the way up to the platform AI engineers use in their day-to-day work.

To illustrate what that would look like, let's look at a couple of examples.



In this example, PaletteAI is used to deploy a full AI stack to a set of physical servers and includes the hardware configuration, operating system, Kubernetes cluster configuration with storage and networking, security tooling and AI-specific layers. It's the ability to model and deploy a stack end-to-end that makes PaletteAI so powerful.

Section 4: Next steps

In this white paper, we've seen that AI fuels the fastest transformation in decades in enterprise infrastructure, reshaping infrastructure, software ecosystems, development workflows and applications — all at once.

Kubernetes, with all of its complexity and potential, is the heart of the technical foundation — but not the solution by itself. The ecosystem of cloud-native tooling surrounding Kubernetes is large and mature. With PaletteAI, you can shape that foundation based on validated designs and single click deployments (cutting time-to-deploy from weeks to hours), so that platform teams can safeguard security, performance, utilization (increase GPU utilization up to 70%) and other corporate IT governance, while data sciences teams can deploy and manage AI workloads faster, more consistently and without infrastructure friction.

This white paper was only an introduction to PaletteAI and we understand reading is a poor substitute for experiencing a product first-hand. Visit spectrocloud.com/get-started to arrange a live, hands-on demo to see how your data science team can ship models faster without toil.

About Spectro Cloud

With our Palette and PaletteAI platforms, Spectro Cloud solves how enterprises and public sector organizations manage full-stack application and AI infrastructure in any environment: from edge to cloud, and from metal to model.

Using the power of cloud-native technologies like Kubernetes, we give platform engineers and operations teams flexibility to choose their perfect stack, while benefiting from complete repeatable consistency. We automate the full lifecycle of complex infrastructure at scale, for massive cost savings and better business outcomes. Learn more at spectrocloud.com

About NVIDIA

NVIDIA (NASDAQ: NVDA) is the world leader in AI and accelerated computing.

About NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud-native software platform that accelerates the development and deployment of production-ready AI. It includes a full suite of enterprise-grade AI and data science tools, optimized frameworks, pretrained models, and infrastructure software, all certified to run on NVIDIA-accelerated systems. With NVIDIA AI Enterprise, enterprises gain consistent, supported operations across data center, cloud, and edge, enabling them to build, scale, and manage AI workloads with performance, security, and efficiency.

Spectro
Cloud