Secure AI Native Architecture (SAINA): Building trust for the AI era

A Spectro Cloud white paper

v1.0, March 2026

Summary

We are witnessing a fundamental shift in how organizations operate. Artificial Intelligence has moved beyond experimental pilots and research labs. It is now the engine driving critical decision-making, product innovation, research, and national defense. To support this, organizations are building "AI Factories" — massive, high-performance data centers designed solely to create and run intelligence.

These environments are powerful, but they are also fragile. The security tools that protected applications for the last decade are simply not designed for AI. They cannot protect complex training, inferencing, or analytical workloads running across shared GPUs. They cannot detect a subtle "poisoning" attack buried deep inside a training dataset.

And they cannot secure data that must remain unencrypted in memory while it is being processed.

The result is a dangerous dilemma: engineers are forced to choose between performance and security. To get the speed they need, they may turn off the firewalls and disable the security agents. They leave the doors to the "crown jewels" wide open.

We believe you shouldn't have to choose between security and performance. This paper introduces the Secure AI Native Architecture (SAINA), a blueprint developed by Spectro Cloud to rebuild trust from the bottom up. By anchoring security in silicon rather than the software, SAINA establishes a trusted, high-performance operating environment.

The SAINA architecture delivers:

- **Hardware-enforced isolation:** Using dedicated NVIDIA BlueField data processing units (DPUs) to physically separate data and traffic from different teams and tenants. This allows them to share expensive hardware resources without risk.
- **Sovereign storage security:** Using external hardware security modules (HSMs) like Fortanix to generate and manage unique strong encryption keys for high-performance storage systems such as WEKA, DDN, and VAST Data. This greatly reduces the risk of data theft.
- **Active defense:** Using NVIDIA DOCA Argus to monitor for threats directly from the system memory, spotting advanced attacks and malware that traditional software-based security agents miss.
- **Confidential computing:** Protecting your data and models even while they are being used by the GPU, ensuring that no one — not even the infrastructure administrator — can see your raw information. This applies to training, inference, and any other sensitive compute workload.

From SENA to SAINA

At Spectro Cloud, we're passionate about spreading security best practices, especially in emerging and therefore risky computing paradigms. Before SAINA, we released SENA: the Secure Edge Native Architecture, built in partnership with Intel to tackle the unique challenges of edge computing environments. **Check it out here.**

Contents

1. The new security reality of AI infrastructure

To understand why a new architecture is needed, we have to look honestly at why the old one is poorly suited to the needs of today's AI factories. Essentially, it's about the shapes and speeds of AI traffic, and how enterprise security appliances slow them down. Let's unpack why.

The physics of the "AI Factory" and performance cost

Even after the world of enterprise IT moved on from a pure perimeter-based 'castle and moat' approach and into an era of Zero Trust, the landscape of IT security essentially involved monitoring and controlling the flow of traffic through control points. SIEM and DLP tools monitor infiltration and exfiltration; internal segmentation strategies recreate the perimeter at smaller scale, augmented by extended user access and identity management controls.

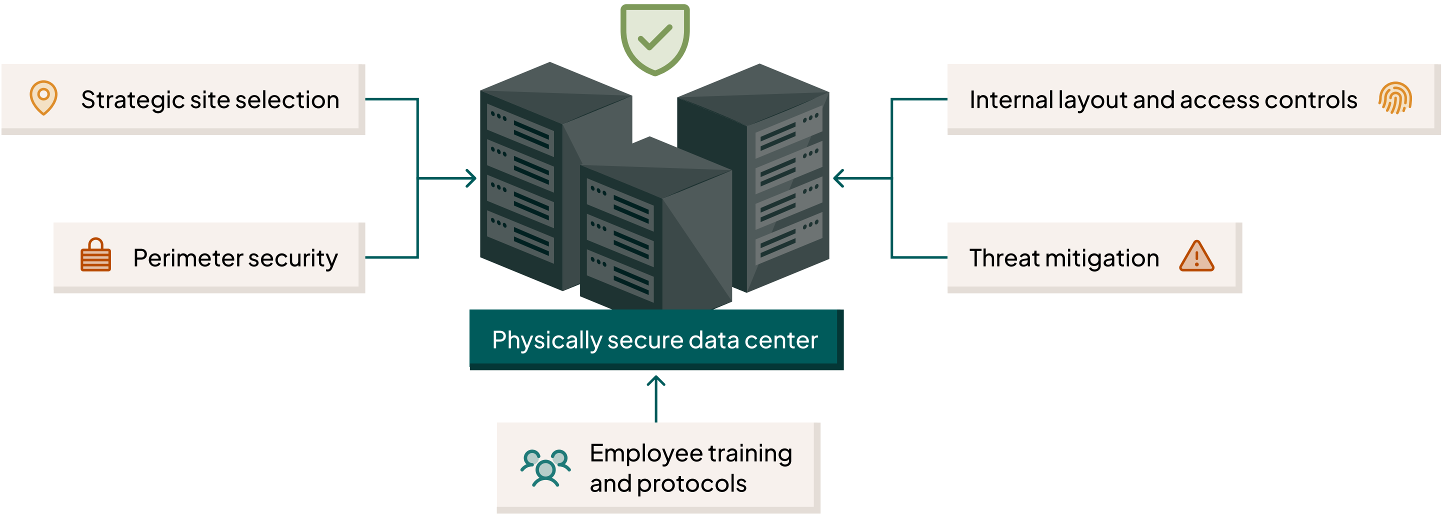

Strategic site selection
Perimeter security
Internal layout and access controls
Threat mitigation
Physically secure data center
Employee training and protocols

Figure 1: Security elements in the data center

The AI Factory turns every security checkpoint into a bottleneck.

The primary workload is distributed compute — training, inferencing, and large-scale data processing — which requires constant communication between different GPUs, servers and storage systems. This "East-West" traffic is colossal. Modern GPU interconnects operate at speeds of 400 Gbps to 800 Gbps per port.

Resource-intensive security or network tasks, like deep packet inspection, placed on the host CPU sap critical cycles away from accelerating AI workloads. This prevents the system overall from reaching its optimal performance.

Faced with this compromise, infrastructure teams are often tempted to adopt a flat and open network design, removing internal firewalls and traffic inspection points.

If an attacker breaches a single node — perhaps through a compromised Python library — they find themselves in a fully trusted zone where they can move laterally to any other node with impunity.

The “Crown Jewels” vulnerability

Any compromise to security architecture is dangerous in the era of the AI factory, because the assets at risk have changed. Your proprietary models, inferencing pipelines, and the training data itself are the most valuable intellectual property on the planet.

This data, whether it’s patient records or classified sensor data, must be decrypted and loaded into the GPU’s memory to be processed. This creates a vulnerability window where the most sensitive secrets sit as plain text in memory, accessible to anyone with “root” access to the host server.

At the same time, many teams may expect access to this data, which is also in constant change: new data being collected and added to the dataset, models retrained and fine-tuned. And GPUs themselves, where this data is being used, may be shared widely across different tenants, even simultaneously.

The security agent dilemma

The standard answer to internal threats is endpoint detection and response (EDR) agents. But in the high-performance world of AI there are two problems:

- **Resource contention:** An agent that wakes up to scan a file causes “jitter,” disrupting the precise data access timing required for distributed training or inference.
- **Stability risk:** Agents often hook into the kernel, causing conflicts with the specialized drivers and libraries required by the GPU stack.

The industry compromise is to disable these security agents on GPU nodes, leaving the most critical infrastructure effectively unmonitored.

Feature	Traditional enterprise architecture	Modern AI Factory architecture
Traffic flow	North-South (User to App)	East-West (GPU to GPU)
Bandwidth	10 Gbps – 25 Gbps per server	400 Gbps – 800 Gbps per GPU
Inspection	Perimeter firewall / Deep Packet Inspection	Impossible with legacy appliances (too slow)
Endpoint security	Heavy Agents (EDR/AV)	Disabled (agents cause performance jitter)
Trust model	Trust the internal network, rely on perimeter security	Zero Trust (but often unimplemented)
Most sensitive assets at risk	Database records	Models, inference pipelines, training data

Table 1: The security gap: traditional vs. AI native

2. The SAINA philosophy: trust the silicon

Security needs to live and run somewhere.

We've established that the AI factory is too performance-intensive to use agents, and standalone security appliances cause bottlenecks..

We also can't automatically trust the operating system stack — like all conventional software, it's vulnerable to tampering and exploits.

So where do we put security? We put it in the hardware.

The Secure AI Native Architecture (SAINA) is built on a "Silicon-to-Software" philosophy. We assume the host operating system might be compromised.

Instead, we use NVIDIA BlueField Data Processing Units (DPUs). The DPU is a "computer in front of the computer." It runs its own secure operating system and is physically isolated from the host CPU.

The DPU runs all the security controls, and because it's electronically airgapped from the host OS, these controls cannot be disabled, even with server root access.

While strong security is the main benefit of using the DPU in this way, the DPU also offloads all the intensive computation of security, networking, and storage I/O processing from the CPU. Every host CPU cycle is then dedicated to accelerating AI training or inference, achieving optimum performance without resource contention.

To manage the DPU, Spectro Cloud PaletteAI orchestrates this silicon layer, translating high-level governance policies into low-level hardware rules.

3. Starting with trust: the hardware foundation

Almost every modern AI factory is made up of many, many servers, together working as nodes in one or more large Kubernetes clusters. Our security efforts therefore need to start with each of these host servers, from the moment they boot and load their operating systems. This is because attackers can compromise even the OS kernel, giving them access to everything that happens afterwards.

Hardware root of trust: measured boot

Before we run any workload, we must establish that each host server booted the correct, uncompromised OS. To do this, we use a cryptographic process anchored in hardware.

The process begins with the server's firmware and a Trusted Platform Module (TPM 2.0), which almost all modern servers have. The TPM is a separate physical chip on the server motherboard, designed specifically for security.

- **Secure boot:** The system's immutable boot ROM verifies the digital signature of the software booting up, each time the system boots — including after patches and upgrades or power cycles.
- **Measured boot:** As the system loads the operating system kernel and core drivers, the TPM records a unique cryptographic hash ("digital fingerprint") of each component. The TPM stores these records in shielded memory registers. This audit trail is append-only, ensuring that malware cannot erase the boot record and evidence of its own loading.

Remote attestation: the gatekeeper

Once the host system boots, Spectro Cloud PaletteAI performs remote attestation before allowing the node to join the cluster, following the flow shown in the image below.

1. **Challenge:** PaletteAI challenges the host system, which responds with a "Quote."
2. **Verification:** The Quote contains the digital fingerprints (measurements) taken by the TPM, cryptographically signed by the TPM's internal, secret key.
3. **Appraisal:** PaletteAI verifies the signature and compares the reported measurements against a trusted "Golden Record" of approved hashes.

This process ensures, with mathematical certainty, that the host operating system has not been tampered with. If the host kernel, firmware, or drivers are modified, PaletteAI rejects the node, preventing it from accessing the network or sensitive workloads.

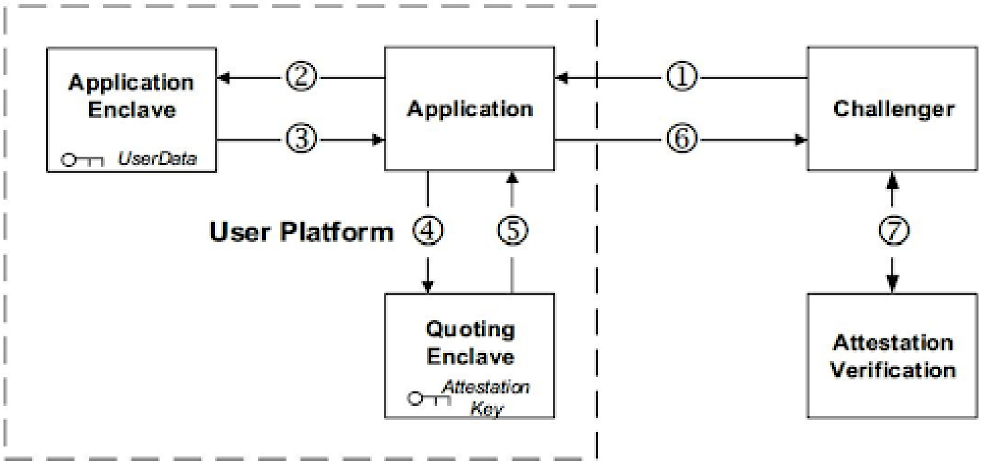


Figure 2: The attestation process

4. Keeping tenants separate: true multi-tenancy

In an AI Factory, running AI workloads at scale, different departments or even different companies are likely to be using the same infrastructure at the same time, even down to the level of an individual GPU.

In this scenario, strict multi-tenancy is not just an IT policy: it's a security necessity. SAINA enforces isolation across three crucial planes: compute, networking, and storage.

Compute isolation: slicing the GPU with MIG

Sharing a powerful GPU is risky due to side-channel attacks, in which one user infers data about another by monitoring shared resources like cache timing or power usage.

SAINA uses NVIDIA Multi-Instance GPU (MIG) technology on Hopper and Blackwell architectures to achieve GPU slicing safely.

MIG physically partitions the GPU hardware, dividing a single accelerator into up to seven fully isolated instances. Each MIG instance receives dedicated compute units (SMs), its own High Bandwidth Memory (HBM) slice, and, crucially, its own physically isolated L2 Cache banks, as shown in the image below.

This full-scope isolation eliminates the shared resources that side-channel attacks exploit. PaletteAI orchestrates this slicing, ensuring that each tenant receives a resource-guaranteed, securely isolated partition for their training, inference, or data processing workloads.

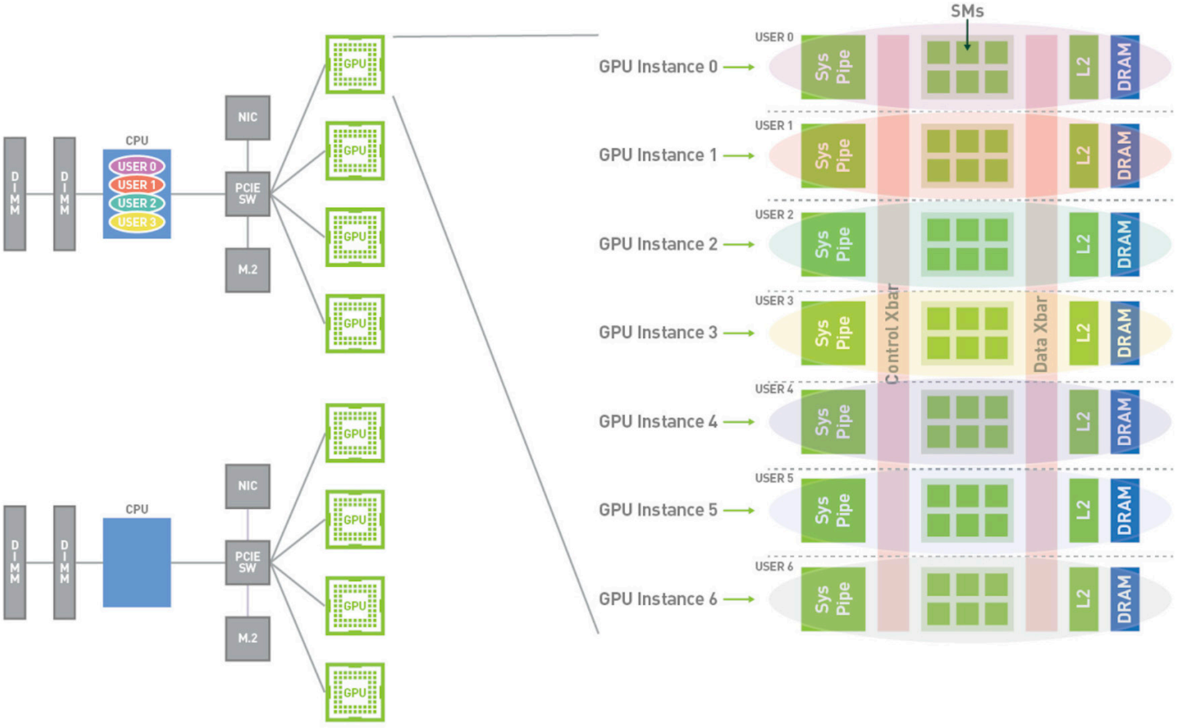


Figure 3: Slicing a GPU with MIG

Network isolation: the electronic air gap

SAINA achieves network segmentation without sacrificing performance by offloading network enforcement to the DPU.

We use the NVIDIA DOCA Platform Framework (DPF) and OVN-Kubernetes to achieve this. The Open Virtual Network (OVN) control plane runs on the dedicated Arm cores of the BlueField-3 DPU.

When a packet leaves a GPU, it goes straight to the DPU. The DPU's hardware accelerators check the firewall rules (ACLs) and route the traffic at wire speed (up to 400 Gbps), completely bypassing the host CPU.

The result is an electronic air gap. Even if one tenant gains root access to its host server, it cannot see or interfere with the network traffic of another tenant.

Storage isolation: the crypto-shredding workflow

Data at rest requires protection that is mathematically provable. We use the concept of crypto-shredding by integrating PaletteAI with hardware security modules (HSMs) like Fortanix for sovereign storage security. This integration targets high-performance, AI-native storage systems such as Weka, DDN, and VAST Data.

The workflow guarantees secure, cryptographic data destruction.

When a tenant workspace is first provisioned, PaletteAI requests a unique, high-entropy encryption key from the Fortanix HSM. This key never leaves the HSM boundary, unless access is given.

PaletteAI passes this key securely to the Container Storage Interface (CSI) driver of the storage system, which creates an encrypted volume using that tenant-specific key.

When the project ends, PaletteAI sends a command to the HSM to delete the key. Deleting the encryption key instantly renders the encrypted data blocks on the disk into random noise.

This provides immediate and provable data destruction that meets compliance and governance requirements.

5. Active defense: the invisible security guard

Isolation and encryption are necessary, but it's always good practice to assume that a breach will occur — and so we need to plan for how to detect a breach, without consuming host resources.

SAINA employs active defense using NVIDIA's DOCA Argus. This is an agentless security tool running entirely on the BlueField-3 DPU.

DMA: the inspection superpower

Argus uses **Direct Memory Access (DMA)**, a feature of the PCIe bus that allows the DPU to read the host server's system RAM directly, without asking the host CPU for permission.

Because Argus operates autonomously outside the host operating system's control plane, the host OS (and any malware running on it) has no idea it is being watched. There is no performance impact, and the attacker cannot kill the monitoring process.

To simplify incident forensics, Argus takes real-time snapshots of the volatile memory and analyzes them for evidence of malicious activity.

What Argus detects

Argus provides real-time situational awareness by inspecting memory structures for threats. It checks the SHA256 hashes of running executables and loaded libraries against a baseline to detect code injection or file size mismatches. It monitors I/O and process memory access patterns associated with unauthorized bulk data exfiltration, often indicators of model theft. And to deal with malware, it inspects kernel structures directly to find hidden processes or reverse shells that the compromised OS tries to conceal.

From detection to response

When Argus detects a high-severity alert, it doesn't just log it. Because it controls the network interface via the DPU, Argus can instantly and autonomously apply a firewall rule to sever the network connection of the compromised node. This containment happens in milliseconds, trapping the attacker before exfiltration or lateral movement can occur.

Threat type	Detection mechanism	Response action
Code injection	Scans memory against known-good binary hashes (Process Attestation)	Quarantine node; alert SIEM
Reverse shells	Detects unauthorized network connections and associated processes	Sever network connection immediately
Model exfiltration	Monitors I/O patterns for massive, unauthorized data reads from memory	Block network egress; freeze node
Hidden processes	Inspects kernel memory structures directly, bypassing OS camouflage	Report concealed activity to security platform

Table 2: DOCA Argus threat detection and response

6. Protecting what matters: data and workloads

We have secured the infrastructure. Now we must protect the data and code while they are being processed.

Confidential computing (data-in-use)

Encryption has traditionally covered data at rest and data in transit. But data always has to be decrypted to be processed. This "data-in-use" vulnerability is where the most sensitive data — including proprietary models and sensitive training sets — is exposed.

SAINA uses confidential computing technology available in modern host systems and GPUs to close this gap.

The Trusted Execution Environment (TEE) creates a hardware-protected enclave — a "black box" — within the secure processing unit (CPU or GPU). Data remains encrypted as it moves across the PCIe bus and is only decrypted inside the silicon of the GPU or CPU core, processed, and immediately re-encrypted. This process prevents unauthorized users from accessing or modifying the data or the code running within the enclave.

This means that even if a rogue infrastructure administrator gains access to the physical server and dumps the memory, they will only see encrypted noise.

Supply chain security

Practitioners are always keen to try the latest and greatest available technology, from new AI tools to a host of community models. But the proliferation of open-source models and containers creates a toxic supply chain problem, making "model poisoning" a major concern.

PaletteAI acts as a strict border control agent for all incoming software, enabling practitioners to have timely access to the ecosystem, but also putting the necessary guardrails in place to avoid risks.

The platform engineering team responsible for the environment first set a policy guardrail: "Only allow binaries, models, or containers signed by our approved corporate key."

When a data scientist attempts to deploy a workload, PaletteAI's admission controller uses tools like Cosign and Sigstore to verify digital signatures on the container images and models, including large language models (LLMs). If the artifact is unsigned or invalid, deployment is blocked, preventing the introduction of compromised software.

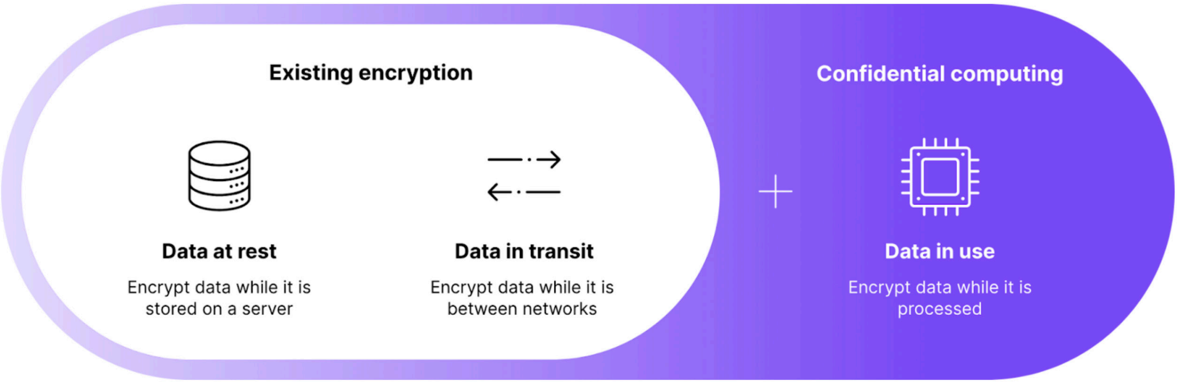

Existing encryption
Data at rest
Encrypt data while it is stored on a server
Data in transit
Encrypt data while it is between networks
+
Confidential computing
Data in use
Encrypt data while it is processed

Figure 4: Confidential computing extends traditional encryption to cover 'data in use'.

7. The management plane: Spectro Cloud PaletteAI

SAINA's many powerful functions depend on sophisticated hardware — DPUs, TPMs, HSMs and more. That power is useless if it is too complex to manage. Spectro Cloud PaletteAI is the unifying platform that makes SAINA operational at scale. It abstracts the complexity of the silicon into usable workflows for humans.

Two personas, one platform

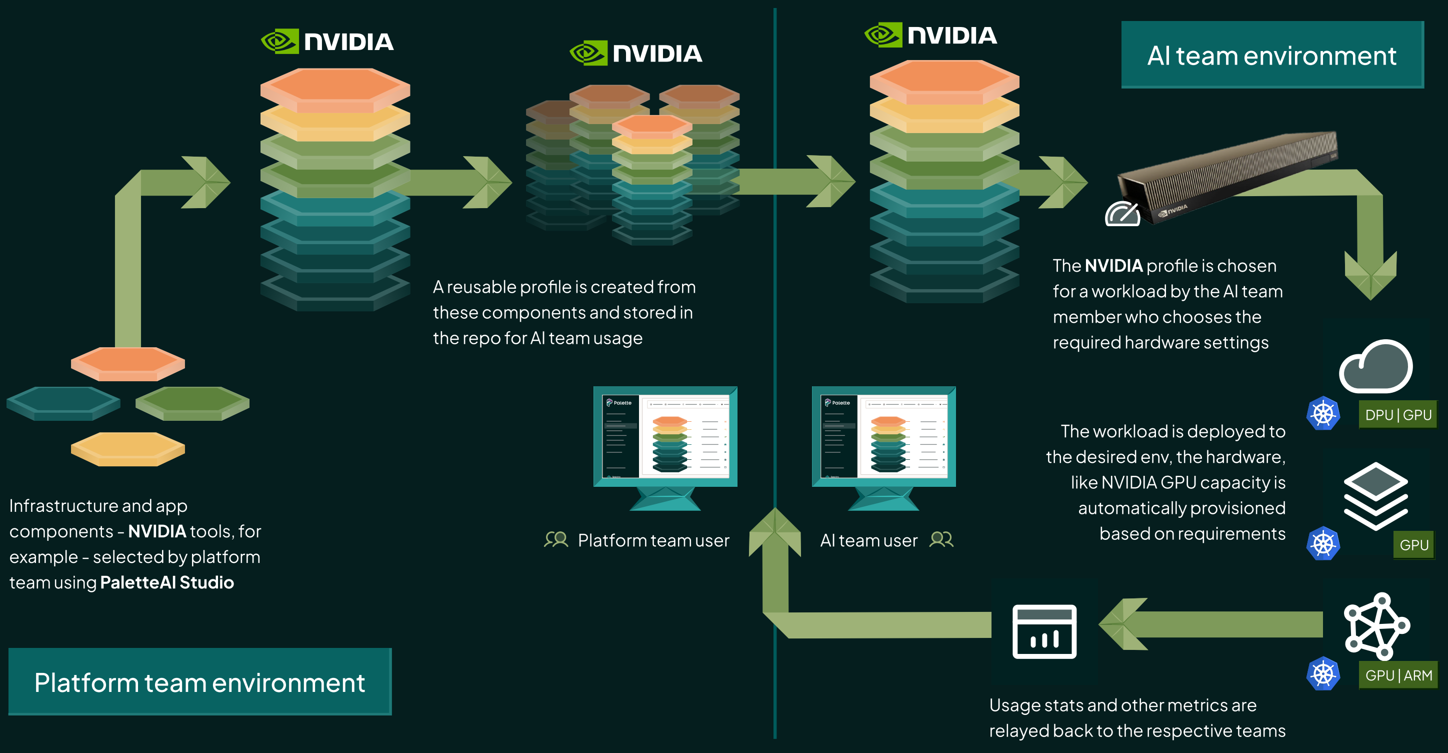
PaletteAI is designed to eliminate the conflict between security and development velocity by separating roles:

PaletteAI Studio (for the platform engineer): This is the command center. Engineers define the "Golden Profiles," that power the AI factory, customizing and building on prebuilt profiles provided by Spectro Cloud and its partners.

Profiles cover the true full stack of the AI environment, from the OS, storage and networking up to popular AI tooling and models.

Engineers then set the OVN firewall rules, configure the Fortanix integration, and manage the infrastructure layer. Their goal is control, compliance, and setting policy guardrails.

Practitioner Workspaces (for the data scientist): This is the self-service portal. Data scientists don't need to see the complexity. They just select a secure template, launch their notebook, and get a GPU-accelerated workspace that is pre-hardened, network-isolated, and actively monitored. Their goal is speed and agility.



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

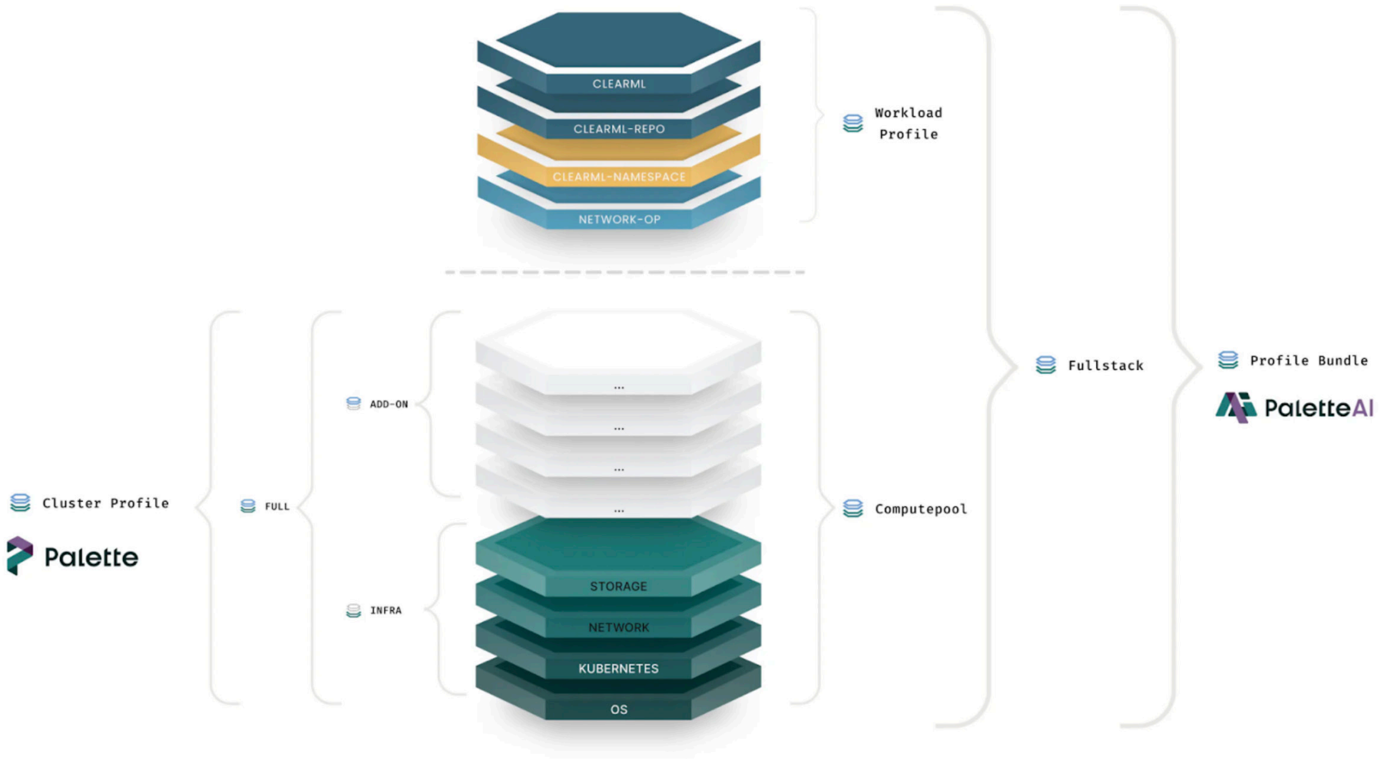


Figure 5: PaletteAI's various profile types enable declarative modeling of the full infrastructure stack, literally from OS to model

PaletteAI VerteX for regulated industries

For government, defense, and highly regulated industries, Spectro Cloud offers an edition of PaletteAI where the entire cryptographic stack is validated to FIPS 140–3 standards.

This adherence is mandatory for US Federal systems and ensures every layer uses validated encryption modules.

FIPS 140-3™

Many AI Factories must operate disconnected from the public internet (often referred to as airgapped).

To suit these scenarios, PaletteAI supports the packaging of all cluster updates, applications, container images, and Kubernetes manifests into single, verifiable airgap deployment artifacts (content bundles). These bundles can be scanned, transferred securely, and loaded into the disconnected cluster.

CONCLUSION:

Building a new architecture for trust in the AI era

AI factories have fundamentally unique security challenges that demand a new security architecture. That's what we've set out to achieve with SAINA. It:

- Ensures performance by offloading security to the BlueField-3 DPU.
- Guarantees isolation using MIG and the electronic airgap.
- Delivers provable data security via HSM-backed crypto-shredding.
- Enables invisible, real-time defense with DOCA Argus.
- Is all managed simply and consistently by Spectro Cloud PaletteAI.

With SAINA, we're not just building infrastructure. We're building trust: with our users, customers, partners and other stakeholders. And in the AI era, trust is the most valuable currency of all.

Get in touch

If you'd like to explore more of the security controls described in this paper, or see first hand how PaletteAI simplifies their deployment and management in your AI factory, get in touch at **spectrocloud.com/get-started**.

Appendix: SAINA components

Component	Function	Technology used	Security benefit
Management plane	Orchestration & governance	**Spectro Cloud PaletteAI**	Unified policy enforcement, lifecycle management, user persona separation.
Root of trust	Boot integrity	**TPM 2.0 / Host System**	Verifies the integrity of the host system's kernel and firmware before cluster join.
Network security	Segmentation & firewall	**BlueField-3 DPU + OVN**	Offloads firewall to hardware, freeing host CPU cycles; creates electronic airgap.
Compute security	Workload isolation	**NVIDIA MIG (Multi-Instance GPU)**	Physically partitions GPU cache and compute units to prevent side-channel attacks.
Storage security	Data-at-rest protection	**Fortanix HSM + PaletteAI**	Crypto-shredding for instant, provable data destruction on Weka, DDN, VAST Data.
Active defense	Threat detection	**DOCA Argus**	Agentless, invisible memory introspection via DMA to catch run-time malware.
Data privacy	Data-in-use protection	**Confidential Computing (TEE)**	Keeps data encrypted inside the secure processing hardware during compute.
Supply chain	Software integrity	Cosign + Sigstore	Blocks unsigned or unverified containers, models, and binaries from deployment.

Table 3: SAINA component summary

Spectro Cloud