INTELLIGENCE BRIEFING
Security Command Center
TLP:CLEAR
2026-06-14 04:24 UTC

NVIDIA and CrowdStrike Push Security Into AI Infrastructure Silicon, But Real-World Validation Is Still Ahead

SECURITY ANALYSIS | MEDIUM | CVSS 5.0

SCC Item ID	SCC-STY-2026-0195
Type	Security Analysis
Severity	MEDIUM
CVSS Base Score	5.0
Affected Products	NVIDIA Vera BlueField-4 STX, NVIDIA DOCA (Argus, Vault, Flow), CrowdStrike Falcon Next-Gen SIEM, CrowdStrike Charlotte Agentic SOAR, VAST Data Zero Trust Framework
Discovery Source	Rss:T1 Threatintel

Executive Summary

NVIDIA and CrowdStrike have announced a joint initiative to embed security controls directly into AI infrastructure silicon, using NVIDIA's Vera BlueField-4 STX data processing unit and its DOCA software stack to generate security telemetry from agentic AI workloads that previously produced no usable signal for SOC tooling. CrowdStrike plans to ingest that telemetry into Falcon Next-Gen SIEM for correlated detection across what the industry is calling the 'AI factory' layer. All claims remain pre-production, STX-based platforms are not expected in customer environments until H2 2026, making this a significant architectural bet rather than a validated defense, and organizations building AI infrastructure today should treat the visibility gap as a present-day risk, not a future one.

Technical Analysis

The announcement addresses a structural problem that has grown alongside the rapid deployment of agentic AI workloads: autonomous agents operating at the infrastructure layer generate no native security telemetry, leaving SOC teams blind to lateral movement, credential abuse, and data access patterns that occur entirely below the host OS visibility horizon. NVIDIA's BlueField-4 STX DPU is designed to close that gap by offloading security functions into silicon, with three DOCA software components carrying the load. DOCA Argus provides telemetry and observability, capturing network flows and process behavior at the DPU layer independent of the workload OS. DOCA Vault handles secrets and key management for the AI workload environment. DOCA Flow enforces network policy at line rate. CrowdStrike's announced integration would ingest Argus telemetry into Falcon Next-Gen SIEM, enabling correlation between infrastructure-layer signals and traditional host and

identity telemetry. VAST Data's complementary positioning adds a storage-layer Zero Trust claim, addressing data access governance for AI pipelines that move large volumes of model weights, embeddings, and context memory across high-density compute clusters. The relevant MITRE ATT&CK surface is broad: T1078 (Valid Accounts) covers agent credential abuse in multi-model orchestration; T1083 (File and Directory Discovery) maps to agent reconnaissance of context memory stores; T1530 (Data from Cloud Storage) covers unauthorized agent data access; T1210 (Exploitation of Remote Services) addresses lateral movement across AI infrastructure nodes; T1040 and T1557 address network-layer interception risks in agent-to-agent communication. The corresponding CWE classes, CWE-284 (Improper Access Control), CWE-200 (Sensitive Information Exposure), and CWE-693 (Protection Mechanism Failure), describe the structural weaknesses this initiative targets, not vulnerabilities in any shipping product. The critical caveat is timing: production availability is projected for H2 2026, and no adversarial validation data exists yet. Organizations deploying agentic AI infrastructure today are operating in the gap this architecture is designed to fill, with no equivalent in-silicon telemetry source currently available at scale. The practical implication for security architects is that existing perimeter and host-based controls are structurally mismatched to the AI workload threat model, and compensating controls need to be designed now rather than deferred until the NVIDIA-CrowdStrike stack ships.

Action Checklist

1. Step 1: Assess exposure, audit your current and planned AI infrastructure deployments for agentic workloads, multi-model orchestration systems, and high-density GPU compute clusters where infrastructure-layer telemetry is absent; specifically identify whether NVIDIA BlueField DPUs, DOCA components, or VAST Data storage are in scope
2. Step 2: Review controls, evaluate whether existing SIEM and EDR coverage extends to the DPU and storage layers of AI infrastructure, or whether a visibility gap exists; verify NIST AC-3 (Access Enforcement) and AC-6 (Least Privilege) are applied to agent service accounts and orchestration identities, not just human accounts
3. Step 3: Map agent identity and access, apply CIS 5.1 (Establish and Maintain an Inventory of Accounts) and CIS 5.4 (Restrict Administrator Privileges to Dedicated Administrator Accounts) to AI agent identities; agentic workloads frequently run with over-provisioned permissions that map directly to T1078 abuse scenarios
4. Step 4: Address network policy gaps, enforce segmentation between AI compute, storage, and orchestration layers aligned with NIST AC-4 (Information Flow Enforcement); review NIST IA-2 (Authentication) for agent service account authentication and NIST AC-5 (Separation of Duties) applicability to multi-factor enforcement in your orchestration framework
5. Step 5: Build a compensating control plan for the 2026 gap, document what telemetry sources currently cover agentic AI workloads; engage NVIDIA, CrowdStrike, and VAST Data on roadmap timelines; establish logging baselines per NIST AU-2 (Event Logging) and CIS 8.2 (Collect Audit Logs) for AI pipeline activity so anomaly baselines exist before the new stack arrives
6. Step 6: Update threat model, incorporate T1083 (agent context memory reconnaissance), T1530 (unauthorized agent data access), and T1078 (agent credential abuse) into your AI workload threat register; brief leadership on the structural nature of the visibility gap rather than framing it as a vendor announcement

IR / Forensic Enrichment

Triage Priority	STANDARD
Escalation Criteria	Escalate to urgent if active agentic workloads are confirmed running on BlueField DPUs or VAST Data storage without any existing telemetry coverage, or if an agent service account is found with cluster-admin or root-equivalent privileges and no compensating monitoring — conditions that represent a live, undetected lateral movement surface requiring immediate containment action.
Recovery Notes	Once segmentation controls and agent identity restrictions are in place, validate that agentic workload pipelines (inference serving, multi-model orchestration, RAG data retrieval from VAST Data) continue to function correctly under the new least-privilege and network policy constraints before declaring recovery. Monitor orchestration controller logs and DPU management interface access logs for a minimum of 30 days post-remediation for anomalous agent authentication attempts, unexpected cross-segment flows, or service account usage outside established behavioral baselines. Establish a formal review gate tied to NVIDIA DOCA Argus and CrowdStrike Falcon Next-Gen SIEM integration availability so that compensating controls are formally retired and replaced — not simply abandoned — when the vendor telemetry pipeline reaches production validation.
Forensic Artifacts	BlueField DPU ARM subsystem process list and DOCA service state ('systemctl status doca-*', 'ps aux' via out-of-band DPU SSH) — captures which DOCA components (Argus, Vault, Flow) were active at time of assessment and whether any unexpected processes were running in the DPU's isolated execution environment | Kubernetes RBAC role bindings for all agentic service accounts ('kubectl get clusterrolebindings -o yaml') — documents over-provisioned agent identities that could be abused for lateral movement across the AI compute cluster in a T1078-analog attack against orchestration infrastructure | VAST Data S3-compatible access logs or NFS audit trails showing which agent identities accessed which storage namespaces and at what volume — establishes baseline for detecting unauthorized agent data access (T1530 analog) across the AI training and inference data fabric | Network flow records or tcpdump captures on trunk interfaces between GPU compute, storage fabric, and orchestration control plane subnets — surfaces any existing unauthorized east-west traffic that the current telemetry-absent environment would not otherwise detect prior to DOCA Flow integration | Sysmon Event ID 1 (Process Creation) logs on orchestration controller hosts showing parent-child process trees for agentic runtimes (python, ray, triton) — establishes the pre-hardening behavioral baseline needed to detect anomalous agent process spawning after DOCA Argus telemetry integration becomes available

Per-Action IR Details

Step 1: Assess exposure — audit your current and planned AI infrastructure deployments for agentic workloads, multi-model orchestration systems, and high-density GPU compute clusters where infrastructure-layer telemetry is absent; specifically identify whether NVIDIA BlueField DPUs, DOCA components, or VAST Data storage are in scope

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: establishing IR capability and identifying gaps before an incident occurs

Controls: CIS 1.1 (Establish and Maintain Detailed Enterprise Asset Inventory), CIS 2.1 (Establish and Maintain a Software Inventory)

Compensating: Run 'lspci | grep -i mellanox' or 'lspci | grep -i bluefield' on GPU cluster hosts to enumerate installed BlueField DPUs without enterprise asset tooling. Cross-reference against rack manifests or IPMI/BMC inventory exports. Maintain findings in a shared spreadsheet tagged by workload type (agentic orchestrator, inference node, storage fabric). Two-person team can complete a 50-node audit in a single shift using parallel SSH via pdsh or Ansible ad-hoc commands.

Evidence: This step does not alter live state. However, document the current asset snapshot — BlueField firmware version ('mlxfwmanager --query'), DOCA service status ('systemctl list-units doca*'), and VAST Data cluster node inventory — before any changes are made, so a pre-remediation baseline exists for later comparison.

Step 2: Review controls — evaluate whether existing SIEM and EDR coverage extends to the DPU and storage layers of AI infrastructure, or whether a visibility gap exists; verify NIST AC-3 (Access Enforcement) and AC-6 (Least Privilege) are applied to agent service accounts and orchestration identities, not just human accounts

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: identifying detection and logging gaps prior to incident; establishing monitoring baselines

Controls: NIST AC-3 (Access Enforcement), NIST AC-6 (Least Privilege), CIS 5.1 (Establish and Maintain an Inventory of Accounts)

Compensating: Without a SIEM, use osquery to enumerate service accounts on orchestration hosts: 'SELECT username, description, shell FROM users WHERE shell != "/sbin/nologin";'. Cross-reference against expected agentic service accounts (LangChain runners, Ray cluster workers, Triton Inference Server service identities). For BlueField DPU management plane, query 'mlnx_bf_configure' and review embedded ARM core process lists via the DPU's out-of-band SSH interface to identify unexpected privileged processes.

Evidence: Before modifying any service account permissions or access policies, capture a point-in-time snapshot of current account privileges: 'getent passwd' and 'getent group' on orchestration hosts, plus any Kubernetes RBAC role bindings ('kubectl get clusterrolebindings -o yaml') governing agent service accounts. This volatile configuration state will be overwritten if accounts are modified and must precede any access enforcement changes.

Step 3: Map agent identity and access — apply CIS 5.1 (Establish and Maintain an Inventory of Accounts) and CIS 5.4 (Restrict Administrator Privileges to Dedicated Administrator Accounts) to AI agent identities; agentic workloads frequently run with over-provisioned permissions that map directly to T1078 abuse scenarios

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: establishing access management baselines and identity governance for non-human actors before adversarial exploitation

Controls: CIS 5.1 (Establish and Maintain an Inventory of Accounts), CIS 5.4 (Restrict Administrator Privileges to Dedicated Administrator Accounts), NIST AC-6 (Least Privilege), NIST AC-2 (Account Management)

Compensating: Export Kubernetes service account tokens and RBAC bindings ('kubectl get serviceaccounts --all-namespaces -o yaml' and 'kubectl get rolebindings,clusterrolebindings --all-namespaces -o yaml') and flag any agent identity bound to cluster-admin or system:masters. For bare-metal orchestration (Ray, Slurm), run 'sudo -l -U ' on each node to surface over-provisioned sudo rights granted to LLM runner or inference service accounts. Document findings before restricting — restriction is the next action and alters live state.

Evidence: Before restricting any agent identity, capture active session data for those accounts: 'who', 'last', 'lastlog', and 'ps aux | grep ' on each orchestration host. In Kubernetes environments, capture live pod security context ('kubectl get pods --all-namespaces -o jsonpath="{..securityContext}"'). These represent volatile runtime state that will be lost once account permissions are modified or pods are restarted.

Step 4: Address network policy gaps — enforce segmentation between AI compute, storage, and orchestration layers aligned with NIST AC-4 (Information Flow Enforcement); review D3-UAP (User Account Permissions) and D3-MFA (Multi-factor Authentication) applicability to agent authentication in your orchestration framework

NIST Phase: Containment

Reference: NIST 800-61r3 §3.3 — Containment Strategy: isolating affected system segments and restricting lateral movement paths before confirmed exploitation occurs in an environment with known visibility gaps

Controls: NIST AC-4 (Information Flow Enforcement), CIS 4.4 (Implement and Manage a Firewall on Servers), CIS 6.3 (Require MFA for Externally-Exposed Applications), CIS 6.5 (Require MFA for Administrative Access)

Compensating: Use Linux network namespaces or nftables rules to enforce micro-segmentation between GPU compute nodes, VAST Data storage fabric, and the orchestration control plane without a commercial NDR platform. A Sigma rule monitoring for unexpected east-west connections from DPU management IPs to production orchestration subnets can be deployed on any syslog aggregator. Wireshark/tcpdump captures on trunk interfaces between AI compute VLANs will surface unauthorized flows. Example: 'tcpdump -i -n "src net and dst net and not port 443"'.

Evidence: Before implementing any firewall rule changes or network policy enforcement that would drop or redirect live traffic, capture 'netstat -ano' or 'ss -tulnp' on all BlueField DPU management interfaces and orchestration controller nodes. Also capture 'ip route show table all' and active iptables/nftables rulesets ('iptables-save', 'nft list ruleset'). These reflect the live network state — enforcement actions will permanently alter observable traffic flows and any evidence of unauthorized lateral movement paths will be obscured.

Step 5: Build a compensating control plan for the 2026 gap — document what telemetry sources currently cover agentic AI workloads; engage NVIDIA, CrowdStrike, and VAST Data on roadmap timelines; establish logging baselines per NIST AU-2 (Event Logging) and CIS 8.2 (Collect Audit Logs) for AI pipeline activity so anomaly baselines exist before the new stack arrives

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: establishing logging infrastructure, detection baselines, and tool readiness to support future detection and analysis when the NVIDIA DOCA/Argus telemetry pipeline becomes operationally available

Controls: NIST AU-2 (Event Logging), NIST AU-3 (Content Of Audit Records), NIST AU-6 (Audit Record Review, Analysis, And Reporting), NIST AU-12 (Audit Record Generation), CIS 8.2 (Collect Audit Logs)

Compensating: Deploy Sysmon on orchestration controller hosts with a configuration that captures process creation (Event ID 1), network connections (Event ID 3), and file creation (Event ID 11) for known agentic runtime executables (python, ray, triton, langchain). Forward via Winlogbeat or rsyslog to an ELK stack or Graylog instance. For VAST Data, enable S3-compatible access logging if the API interface is active and ship logs to the same aggregator. Establish a 30-day rolling baseline of normal agent process trees and API call rates before the DOCA Argus integration arrives so anomaly thresholds are data-driven rather than guessed.

Evidence: This step does not alter live state directly, but the act of enabling new logging on previously unmonitored DPU management interfaces or VAST Data storage APIs may affect agent performance or trigger access policy evaluations. Before enabling audit logging on the BlueField DPU ARM subsystem, capture the current DOCA service configuration ('systemctl status doca-*') and note which DOCA components (Argus, Vault, Flow) are installed versus active, as this establishes the pre-logging baseline against which new telemetry will be compared.

Step 6: Update threat model — incorporate T1083 (agent context memory reconnaissance), T1530 (unauthorized agent data access), and T1078 (agent credential abuse) into your AI workload threat register; brief leadership on the structural nature of the visibility gap rather than framing it as a vendor announcement

NIST Phase: Post Incident

Reference: NIST 800-61r3 §4 — Post-Incident Activity: lessons learned, threat model updates, and policy improvements driven by identified gaps; updating detection and IR playbooks before the next incident

Controls: NIST AU-6 (Audit Record Review, Analysis, And Reporting)

Compensating: Facilitate a tabletop exercise scenario in which an agentic LLM orchestrator running on NVIDIA BlueField-equipped infrastructure is compromised via an over-provisioned service account (T1078 analog) and begins enumerating VAST Data storage buckets (T1530 analog). Use the MITRE ATLAS matrix (which extends ATT&CK to AI/ML systems) as a free reference to construct realistic attack chains specific to agentic workloads — it covers adversarial ML tactics not present in the core ATT&CK enterprise matrix. Document findings as updated threat register entries with no-cost detection hypotheses (e.g., Sigma rules for anomalous agent API call volumes) that can be implemented while the DOCA/Falcon integration is pending.

Evidence: This step does not alter live state. Archive all current threat register documents, risk assessment outputs, and existing SIEM detection rule sets before updating them, so the pre-update state is preserved for audit trail purposes (NIST AU-11 Audit Record Retention). This is particularly important if the organization is subject to compliance frameworks requiring documented evidence of threat model evolution over time.

Detection Guidance

Current detection opportunities are constrained by the visibility gap this story describes, but meaningful hunting is possible with existing tooling. Review AU-2 (Event Logging) coverage against these specific sources: orchestration platform logs for agent-to-agent API calls that deviate from expected workflow patterns (T1071); storage access logs for bulk reads of embedding stores, model weight directories, or context memory paths outside normal pipeline schedules (T1530, T1083); network flow data for east-west traffic between AI compute nodes that lacks expected application-layer signatures (T1040, T1557); identity provider logs for service account authentications originating from unexpected source IPs or at unexpected times (T1078). For organizations running NVIDIA BlueField DPUs in current deployments (BlueField-2 or BlueField-3), DOCA Argus telemetry may already be available in limited form; confirm whether that data is being forwarded to SIEM. NIST SI-4 (System Monitoring) should be scoped explicitly to include AI pipeline processes, not just traditional workloads. Audit service account permissions for orchestration identities against AC-6 (Least Privilege), over-provisioned agent accounts are a primary risk surface for T1078 abuse. NIST AU-2 (Audit Event Logging) applies to agent identity monitoring on compute nodes. NIST SI-7 (Software, Firmware, and Information Integrity) can surface unauthorized modifications to DOCA configuration files or secrets management stores. Organizations relying solely on perimeter controls for AI infrastructure should flag this as a CWE-693 (Protection Mechanism Failure) finding in their risk register and track it against the H2 2026 in-silicon telemetry availability timeline.

Framework Mappings

MITRE-ATTACK

- **T1210** — Exploitation of Remote Services
- **T1078** — Valid Accounts
- **T1190** — Exploit Public-Facing Application
- **T1040** — Network Sniffing
- **T1557** — Adversary-in-the-Middle
- **T1083** — File and Directory Discovery
- **T1530** — Data from Cloud Storage
- **T1071** — Application Layer Protocol

NIST-800-53R5

- **AC-6** — Least Privilege
- **SC-7** — Boundary Protection
- **SI-2** — Flaw Remediation
- **AC-2** — Account Management
- **IA-2** — Identification and Authentication (Organizational Users)
- **IA-5** — Authenticator Management
- **CA-8** — Penetration Testing
- **RA-5** — Vulnerability Monitoring and Scanning
- **SI-7** — Software, Firmware, and Information Integrity

- **CA-7** — Continuous Monitoring
- **SI-4** — System Monitoring
- **AC-3** — Access Enforcement
- **SC-28** — Protection of Information at Rest

OWASP-TOP10-2021

- **A01:2021** — Broken Access Control

CIS-V8

- **6.1** — Establish an Access Granting Process
- **6.2** — Establish an Access Revoking Process
- **8.2** — Collect Audit Logs

SOC2-TSC

- **CC6.1** — The entity implements logical access security software, infrastructure, and architectures over protected information assets

HIPAA-SECURITY

- **164.312(a)(1)** — Access Control

NIST-CSF-2

- **DE.CM-01** — Networks and network services are monitored

MITRE ATT&CK Mapping

Technique ID	Technique Name	Tactic
T1210	Exploitation of Remote Services	Lateral-Movement
T1078	Valid Accounts	Defense-Evasion
T1190	Exploit Public-Facing Application	Initial-Access
T1040	Network Sniffing	Credential-Access
T1557	Adversary-in-the-Middle	Credential-Access
T1083	File and Directory Discovery	Discovery
T1530	Data from Cloud Storage	Collection
T1071	Application Layer Protocol	Command-And-Control

Sources

Source	URL	Tier
Blog	https://www.crowdstrike.com/en-us/blog/crowdstrike-nvidia-bring-ent...	T3
	https://nvidianews.nvidia.com/news/nvidia-vera-bluefield-4-stx-brin...	T3
	https://futurumgroup.com/insights/crowdstrike-falcon-aims-to-see-in...	T3
	https://www.stocktitan.net/news/NVDA/nvidia-vera-blue-field-4-stx-b...	T3
VAST Data's Zero Trust Framework for Agentic AI	https://www.vastdata.com/blog/vast-zero-trust-agentic-ai-nvidia-blu...	T3

Generated 2026-06-14 04:24 UTC by TJS Security Command Center