AI Data Center Security Architecture Blueprint by Check Point

How to secure corporate AI infrastructure and Private LLMs with Check Point's AI security technologies stack. Secure access to the AI Data Center and Public Cloud, and protect AI workloads, applications, data security and management.

1. Introduction

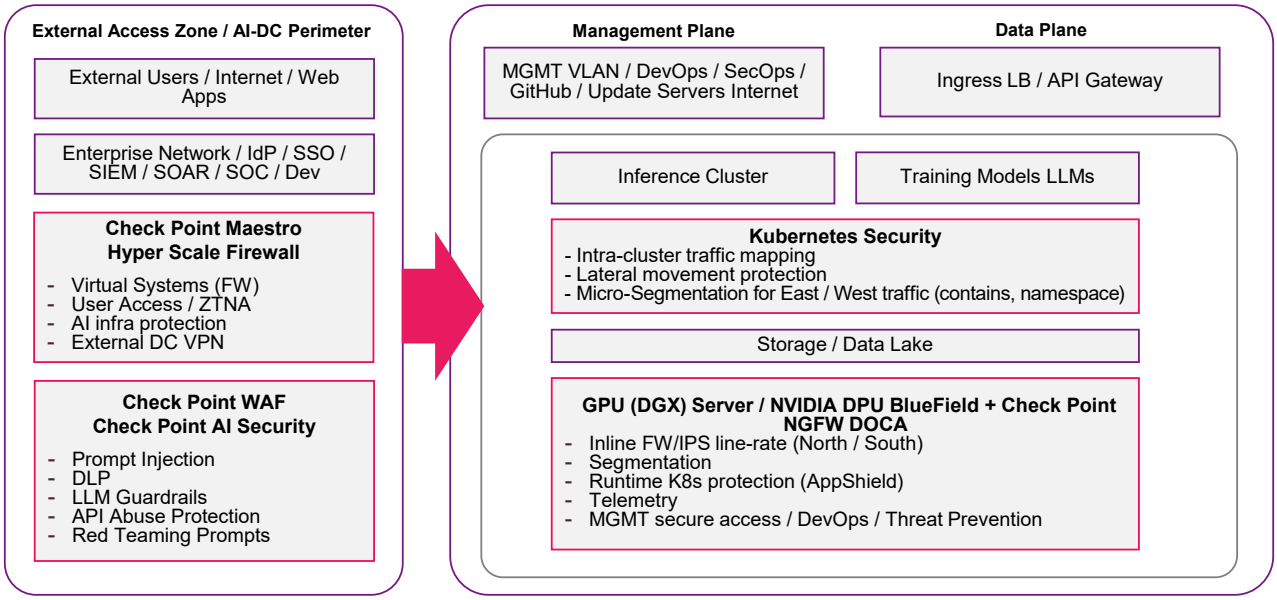
The adoption of private AI and LLM (Large Language Model) infrastructures by enterprises introduces a new class of risks. Unlike traditional IT workloads, AI data centers manage sensitive training data, powerful GPU clusters, distributed inference services, and high-throughput pipelines that can easily become attack vectors.

Organizations face threats to data, intellectual property, AI models, and end-users. Building AI capabilities without embedding security increases exposure to poisoning, data leakage, and governance failures.

To ensure resilience, AI data centers must be secured end-to-end — from the fabric and GPU clusters to Kubernetes workloads, and API-driven inference workloads and services.

2. Value Check Point provides for AI Data Centers Security

Check Point enables organizations to confidently adopt and scale AI by addressing both traditional IT threats and the new, unique risks of AI-driven environments. Check Point solutions based on the modern security technologies and integration with advanced 3rd party products, part of general open-garden strategy; - provides embedded cyber security by design to cover all sensitive blocks of AI-Data Center and Private LLMs.



The solution provides layered protection approach from the access towards AI workloads as shown on the figure above:

- **Perimeter Layer Protection** – external access control zone, entry point to the AI Data Center fabric, secured by Zero-Trust (ZTNA) enabled Hyper Scale NGFW (Maestro)
- **Application Layer Protection** – API based security of AI application and Agentic / LLM layer protection which is different from "traditional" api security
- **AI-Server Layer Protection** – HW embedded NGFW to take a security close to the AI workloads, segment servers (DGX) and protect Management traffic.
- **Kubernetes Layer Protection** – ensure East-West traffic inside K8s cluster visibility and containers micro-segmentation policy enforcement

The value delivered goes far beyond technology, translating directly into measurable business outcomes:

Business Value	Description
Protection of Intellectual Property and Data Assets	Safeguards proprietary AI models, training datasets, and inference results — preventing theft or manipulation of high-value R&D assets.
Business Continuity and Service Reliability	Ensures AI services remain resilient and available even under attack, minimizing downtime and avoiding costly disruptions.
Regulatory Compliance and Trust	Provides governance, traceability, and auditability to meet emerging AI regulations and maintain customer and regulator trust.
Risk Reduction and Cost Avoidance	Reduces the likelihood of breaches, data leakage, and compliance fines, protecting both finances and brand reputation.
Operational Efficiency	Simplifies security operations across training and inference with centralized policy management and automation, lowering overhead for DevOps and SecOps.
Secure Innovation and Faster AI Adoption	Embeds protection from the ground up, enabling organizations to adopt and scale AI with confidence while safeguarding users and customers.
Customer and Partner Confidence	Demonstrates strong security commitment, becoming a differentiator in markets where trust and reliability drive adoption and partnerships.

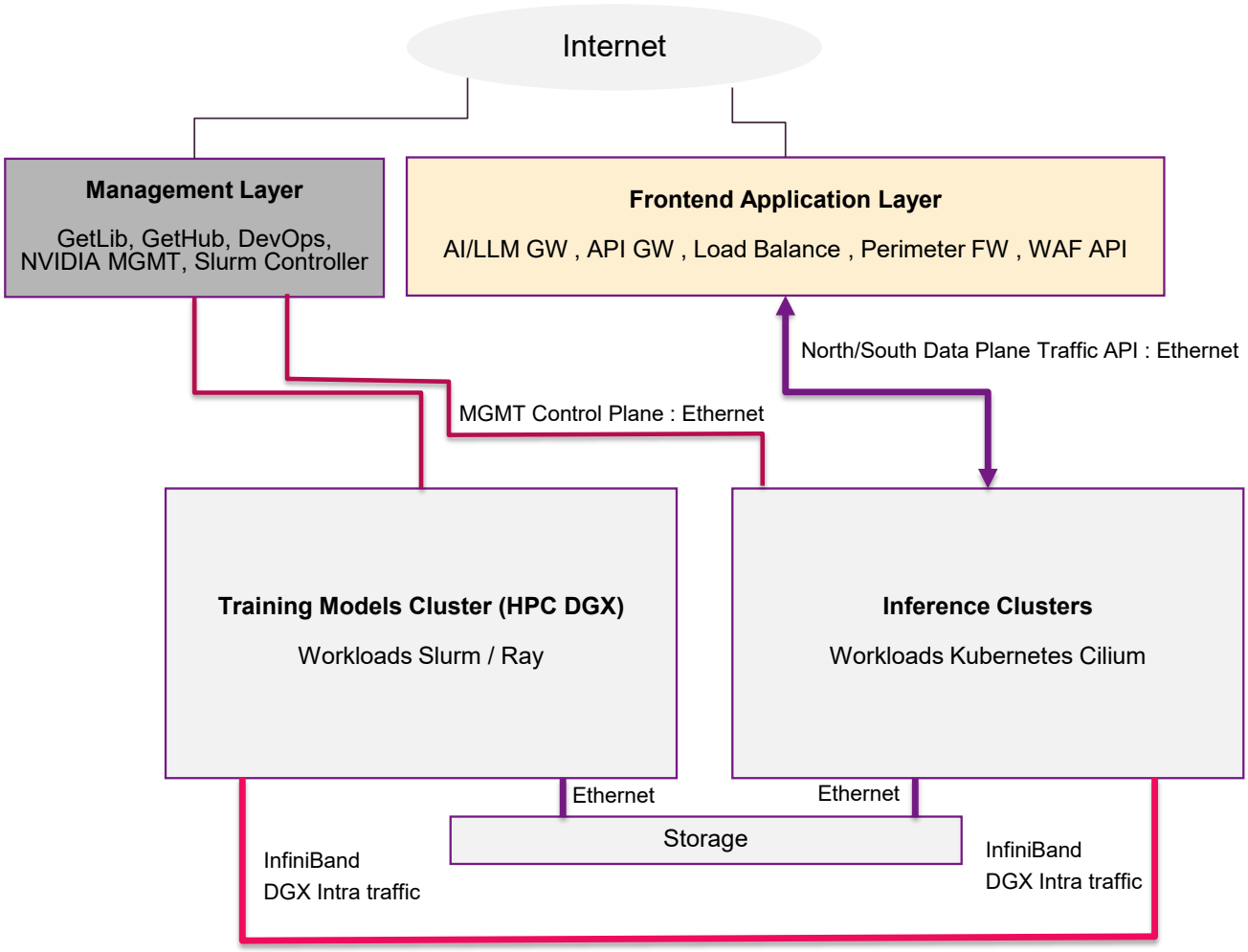
3. High-Level Overview of AI Data Center Architecture

An AI data center is based on model training and inference domains at scale, combining high-performance GPU clusters (e.g NVIDIA), secure connectivity, and orchestration layers.

At the edge, a frontend application layer with API gateways, load balancers, firewalls, and WAFs manages and protects user and application traffic, while a dedicated management layer hosts DevOps, SecOps, and control functions over isolated VLANs.

Inference Cluster hosts deployed AI models that process user queries and application requests in real time. It consists typically of GPU-powered DGX servers orchestrated by Kubernetes with Cilium or other K8s CNI technologies. This cluster handles model inference requests, ensuring high-performance execution and efficient resource utilization across distributed nodes.

Training clusters, typically based on DGX servers, use InfiniBand interconnects for ultra-fast GPU communication and frameworks like Slurm or Ray to coordinate distributed workloads, whereas Inference clusters rely on Kubernetes with Cilium for real-time model serving and policy enforcement.



** Slurm = Simple Linux Utility for Resource Management*

4. AI Infrastructure Security Risks

AI infrastructure introduces unique risks that extend beyond traditional IT systems. These risks directly affect the confidentiality, integrity, and availability of sensitive data, models, and applications. Unlike standard data centers, AI environments combine high-performance computing, large-scale data pipelines, and distributed training clusters — all of which create new attack surfaces, regulatory challenges, and risks of misuse.

Category	Key Risks
Infrastructure & Platform Risks	• Compromise of servers, workloads, or system integrity • Lateral movement through East-West traffic within AI clusters • Compromise of containers or workloads via malicious libraries from GitHub • Exploitation of misconfigured FWs, API GW, or exposed MGMT interfaces / DevOps misuse • Container escape or privilege escalation attacks targeting Kubernetes runtimes • API gateway bypass or misuse exposing model endpoints directly • Resource exhaustion or GPU/memory DoS attacks impacting AI availability
AI Supply Chain Risks	• Tampering in third-party frameworks, APIs, or pre-trained models • Dependency poisoning through open-source libraries • Unverified model weights or firmware updates introducing hidden risks
Data & Training Risks	• Training data poisoning creating hidden backdoors • PII leakage from prompts, responses, or logs • Cross-tenant data bleeding in multi-tenant GPU or inference environments • Unauthorized access or exfiltration of proprietary datasets
Model Risks	• Model poisoning attacks compromising training integrity • Model theft or extraction via API abuse or query enumeration • Adversarial attacks causing targeted misclassification • Model drift and performance degradation over time
Application Risks	• Prompt injection and jailbreak attempts bypassing model controls • Output manipulation or generation of harmful/unethical content • RAG poisoning affecting context accuracy • Agent hijacking or manipulation of autonomous behavior • Vulnerabilities in third-party integrations or plug-ins
AI Governance & Operational Risks	• Lack of AI system accountability or auditability • Insufficient access control and monitoring across AI pipelines • Shadow AI deployments bypassing enterprise governance • Misalignment between security, compliance, and DevOps ownership
Compliance & Regulatory Risks	• Violations of AI-specific regulations (EU AI Act, U.S. Executive Order 14110) • Failure to meet model explainability and "right to explanation" (GDPR) • Breach of data residency and cross-border AI processing laws • Non-compliance with industry frameworks (HIPAA, PCI-DSS, ISO 42001)

5. AI Security by Design

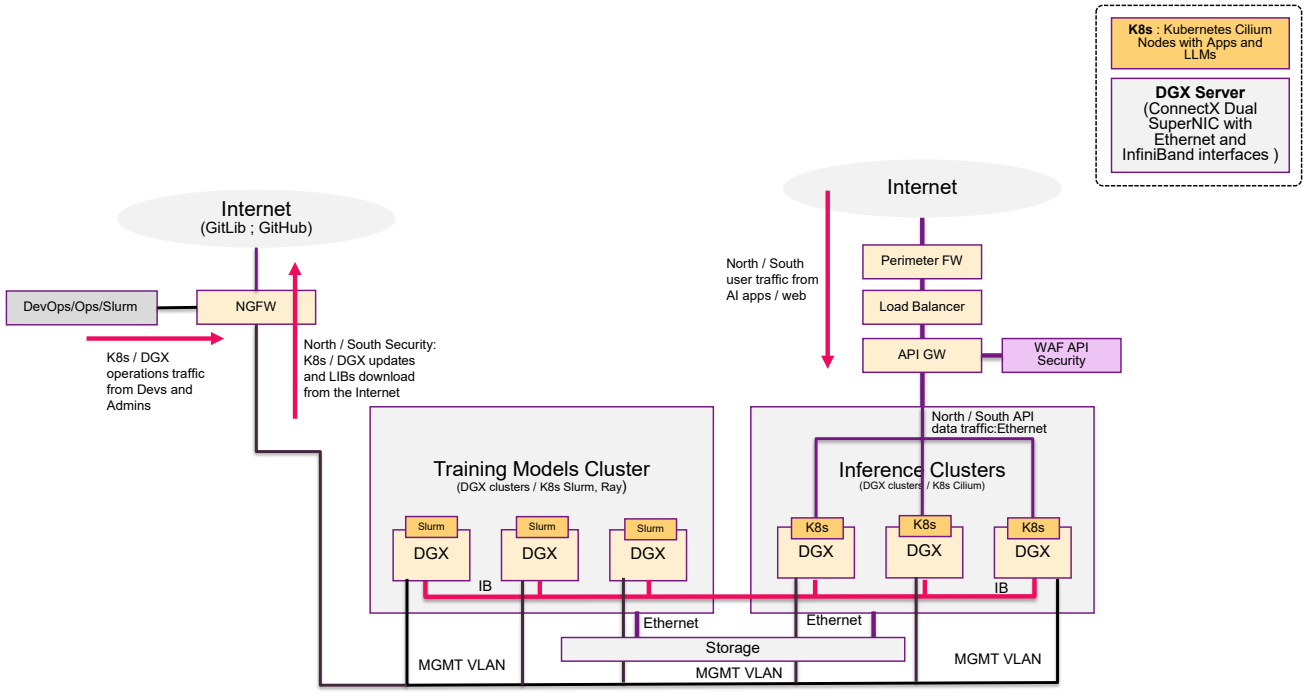
"AI must be Secure by Design. This means that manufacturers of AI systems must consider the security of the customers as a core business requirement, not just a technical feature, and prioritize security throughout the whole lifecycle of the product, from inception of the idea to planning for the system's end-of-life. It also means that AI systems must be secure to use out of the box, with little to no configuration changes or additional cost."
- CISA, Software Must Be Secure by Design, and Artificial Intelligence Is No Exception

AI must be **Secure by Design**, not an afterthought bolted on top of existing systems.

- **Zero Trust everywhere:** every user, API, and service interaction must be authenticated, authorized, and continuously validated.
  - Microsegmented service mesh with mTLS between training and inference zones.
  - Strong NHI (Non-Human Identity) security: service accounts, API key management, and secret rotation.
  - Context-aware access control based on model risk and data sensitivity.
- **AI-Native Security Controls:**
  - LLM gateways with prompt injection detection and rate limiting.
  - Input sanitization and output filtering using policy-based guardrails.
  - Model access control via RBAC/ABAC and runtime protection.
  - Agent authorization boundaries to isolate agentic AI actions.
- **Data & model integrity:** enforce signed models, encrypted data pipelines, and isolated training zones.
- **Continuous validation:** runtime monitoring, adversarial red-teaming, and anomaly detection to expose blind spots.
- **Governance & accountability:** embed policy controls that regulate what AI can access, generate, or share.
  - Maintain audit trails across training, deployment, and inference workflows.
  - Ensure alignment with EU AI Act, GDPR, and sector mandates (HIPAA, PCI-DSS, ISO 42001).

Value: Secure by Design transforms AI from "functional but fragile" into resilient, governed, and business-ready.

6. AI Data Center Generic Security



** Slurm = Simple Linux Utility for Resource Management*

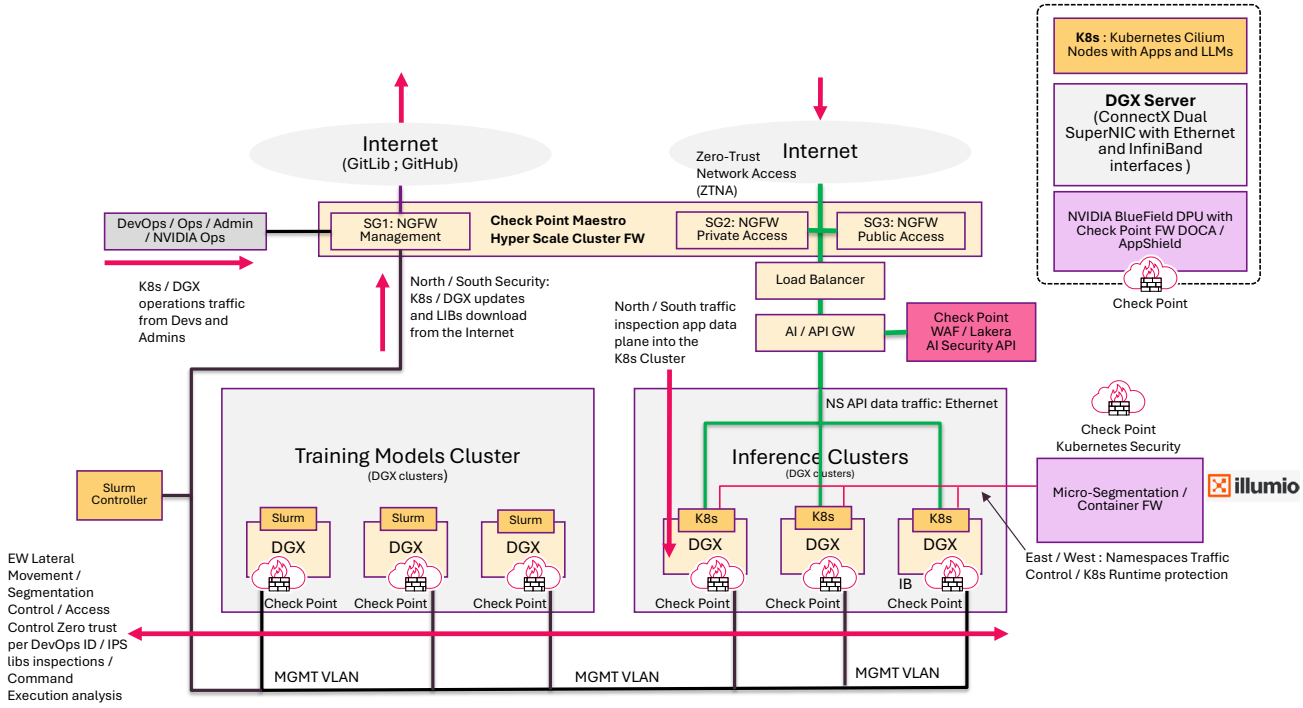
This design illustrates a typical secure AI data center architecture that separates training and inference clusters while enforcing strict controls accross dedicated network zones (Training, Inference, Storage and Management VLANs and segments).

The Training clusters in an isolated environment with no direct Internet access, leverages high-speed interconnects for model development. The Inference clusters handle real-time workloads through Kubernetes orchestration, with users connecting to AI workloads by generating prompts via API gateway located at the front end of the AI fabric.

Security is layered across the stack: the management plane is isolated and protected, ensuring DevOps and administrative access is tightly controlled; north-south traffic from the internet is secured with perimeter firewalls, WAF API / AI-aware gateways that enforce prompt injection detection, rate limiting, and guardrails. The access control is enforced for east-west traffic between clusters and storage, for segmentation and monitoring.

Key principles such as AI-aware security, Zero Trust access, and continuous inspection of both API and DevOps traffic ensure that sensitive models, data, and workloads remain protected against tampering, leakage, and misuse.

7. Check Point AI Data Center Security Architecture



** Slurm = Simple Linux Utility for Resource Management*

The Check Point AI Data Center Security architecture delivers end-to-end protection for both training and inference domains by tightly integrating network, application, and AI-specific security controls.

- At the core, **Check Point Maestro Hyperscale Firewall** provides elastic security with the ability to carve out **separate Security Groups (SGs)**, functioning as independent virtual firewalls. Each SG can enforce distinct security policies—for example, one segment dedicated to DevOps/ administrative access, another to private API traffic, and another for public-facing inference services. This segmentation not only simplifies policy management but also ensures operational resilience by isolating policy domains, updates, and maintenance windows.
- On the application side, **Check Point WAF integrates with AI Security**, adding advanced protection tailored for AI environments.
- Check Point secures AI APIs and LLM endpoints by detecting **prompt injection attacks, data exfiltration attempts, adversarial queries, and misuse of generative AI outputs**, enabling safe consumption of AI services without exposing sensitive models or data.
- The design also clearly separates the **management plane (MGMT VLAN)**, which handles Kubernetes, DevOps, and DGX operational traffic, from the **data plane (Ethernet/InfiniBand)** that carries training and inference workloads. This segregation enforces strict **Zero Trust principles**, ensuring that user/API traffic cannot bypass management protections.
- At the node level, **NVIDIA BlueField DPUs integrated with Check Point DOCA FW and AppShield** provide inline inspection and micro-segmentation for ingress/egress Ethernet traffic before it reaches the InfiniBand fabric.

- In parallel, **Check Point integrates with Illumio** solution and enforces east-west traffic segmentation within Kubernetes clusters, controlling traffic between namespaces, pods, and services, as well as quarantining malicious containers.
- Together, this architecture unifies traditional firewalling, AI-specific application security, and in-fabric DPU enforcement into a **Zero Trust AI Data Center blueprint**, capable of protecting sensitive training data, inference APIs, and operational environments against both traditional and emerging AI-driven threats.

8. Use Cases and Security Components

The following use cases highlight how Check Point secures AI data centers across training, inference, management, and application layers, while directly delivering a value:

Use Case	Check Point Component	Value
Segregating Training and Inference Domains and Servers Protection	Check Point NGFW running on DGX BlueField, enforcing inter-zone traffic policies and segregation of data and control planes.	Reduces risk of lateral movement, ensuring sensitive training data and production inference remain isolated and protected.
Zero Trust User and DevOps Access	Check Point Maestro NGFW Security Groups (SGs) create dedicated access policies for DevOps, admins, and external users.	Ensures least-privilege access, minimizes insider threats, and improves compliance posture.
AI Application Security	Check Point WAF integrated with AI security validates API calls, protects against prompt injection, and secures AI-specific threats, DLP and more.	Protects AI applications from novel attacks, preserving customer trust and regulatory compliance.
Container and Namespace Segmentation and Protection	Check Point integration with Illumio enforces micro-segmentation across K8s namespaces and services. As well as integrating with Maestro NGFW via Logs API to block and quarantine infected workloads.	Prevents unauthorized east–west traffic inside clusters, reducing attack surface, protect from lateral movement inside K8s nodes.
Out-of-Band Threat Detection	BlueField DPUs mirror traffic to Check Point IDS/ Maestro for anomaly and behavior analysis without impacting performance.	Identifies threats early with minimal impact on GPU workloads, maintaining high system throughput.
Training Data Protection and Poisoning Detection	Check Point AI Agent Security + Check Point WAF (API validation) + Data FLow DLP Controls	Detects and blocks poisoned or manipulated training data, prevents ingestion of harmful or corrupted datasets, and ensures clean, trusted data flows into model training pipelines.
Adversarial Attack and Inference Abuse Detection	Check Point AI Agent Security + Check Point WAF behavioral analysis + Sig-based & ML-based anomaly detection	Identifies adversarial prompts, inference-time model manipulation attempts, and abnormal query patterns, protecting AI outputs from hijacking, jailbreaks, and harmful content generation.
Model Theft and Exfiltration Prevention	Check Point NGFW egress filtering + DLP + Bluefield DPU + Model Protection Policies	Prevents unauthorized export of model parameters, embeddings, or sensitive weights, blocks covert exfiltration channels, and detects abnormal

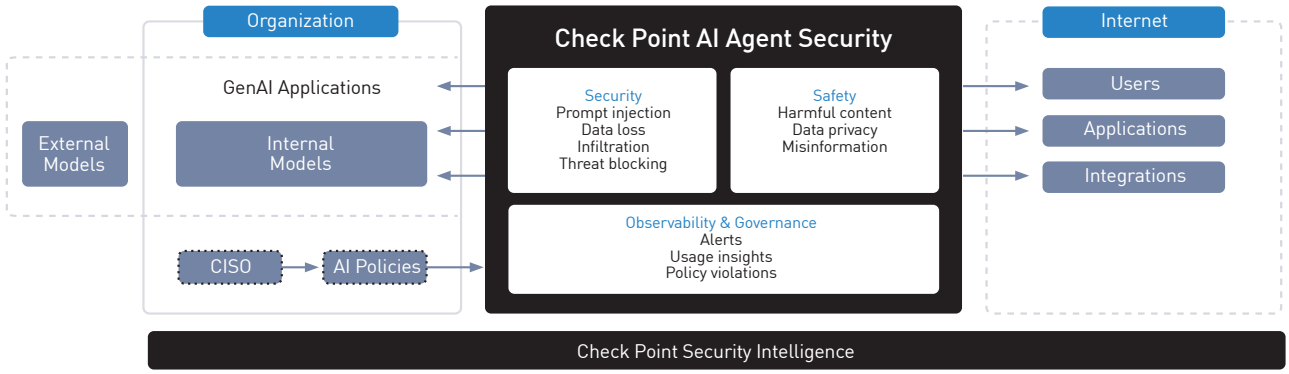
Use Case	Check Point Component	Value
Secure API Traffic Management	API traffic from AI apps is inspected by perimeter firewalls, WAF, and DOCA-based controls before entering inference clusters.	Protects against API abuse, supply chain vulnerabilities, and DDoS attacks.
Data Security and Compliance Monitoring	Check it Centralized telemetry from DGX, K8s, BlueField, integrated into Check Point management	Enables visibility, auditing, and reporting for compliance frameworks (AI Act, GDPR, HIPAA).

9. Runtime security for your GenAI

Check Point AI Agent Security: AI-Native Runtime Guard for LLM Security

Capabilities

- Check Point AI Agent Security provides real-time visibility and control over GenAI applications by intercepting both inputs (prompts) and outputs (generated content).
- It supports customizable "guardrails" (predefined or custom) including prompt defenses, data leakage prevention, content moderation, malicious link detection, and policy-driven filters.
- It continuously evolves threat models via Check Point's intelligence platform, updating detections (e.g. adversarial queries, injections) with low latency and minimal false positives.
- Centralized policy and control: You can define policies across multiple GenAI apps without touching application code, enabling consistent enforcement and monitoring.
- Low overhead, high performance: The solution is designed to add minimal latency to inference pipelines, making it usable in production-scale GenAI.



Use Cases in AI / Private LLM Environments

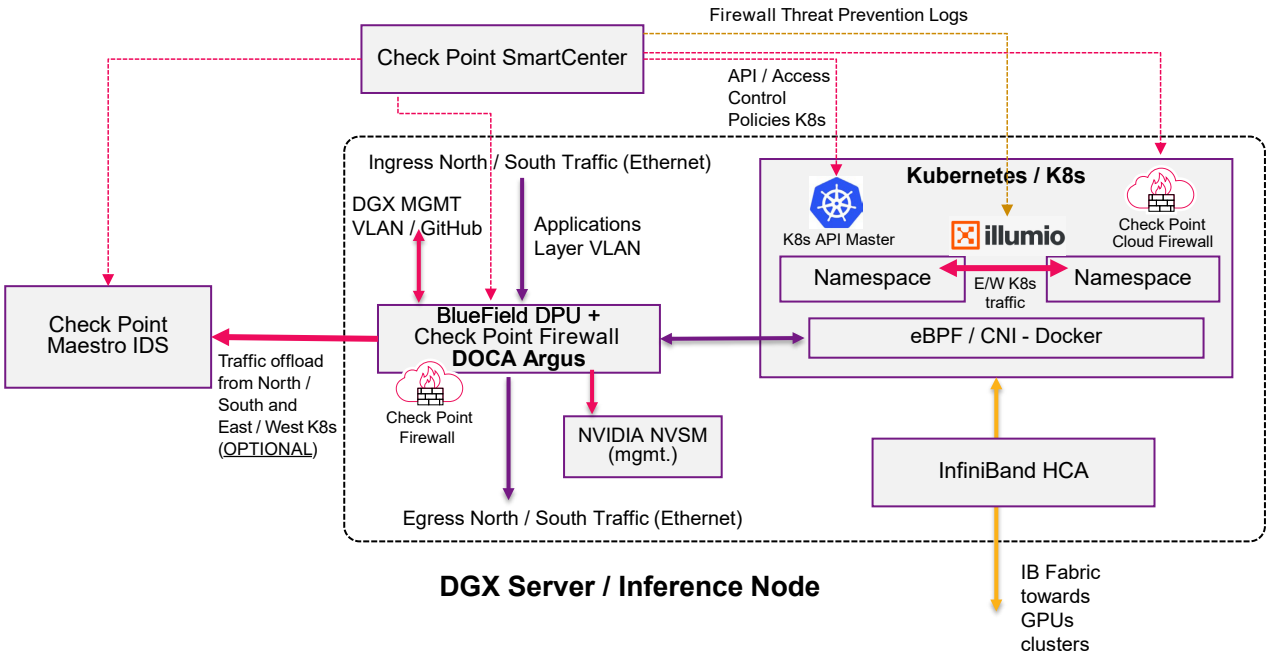
- Conversational Agents / Chatbots: Protects against prompt injection, exposure of system prompts or internal instructions, and out-of-bound responses.
- RAG / Document Agents: Secures retrieval-augmented generation pipelines, guarding against poisoned documents, tampered references, and unauthorized data leakages.

- GenAI Gateways / API Frontends: Used at the boundary of AI systems to centralize protection of API traffic, apply policies across services, and monitor AI service behavior.
- Connected Agents / Integrations: When AI agents call external services (APIs, third-party tools), Check Point ensures security over those agent workflows (blocking malicious API calls, ensuring flow integrity).
- Model Context Protocol (MCP) Architectures: Check Point can wrap MCP tool calls (prompts, resources, tools) to secure interactions and prevent injection or misuse in emerging AI frameworks.

Value & Impact for an AI Data Center / Private LLM Security

- AI-specific defense layer: Check Point fills a gap that traditional security tools can't cover — attacks crafted in natural language (prompt injections, subtle manipulations) that could lead to data leaks or model abuse.
- Separation of concerns / clean interface: Because Check Point functions through APIs at the application layer, it doesn't require deep instrumentation in GPU nodes or container runtimes — simplifies deployment.
- Consistent policy across environments: You can apply the same security governing logic across multiple inference clusters, AI apps, and gateways, ensuring uniform protection and lowering policy fragmentation.
- Support for compliance & auditing: Logging, threat detection, and response features provide traceability for AI interactions. For audits, AI regulation compliance, or investigation.
- Low friction adoption: Minimal latency impact and code-agnostic integration make it viable in production AI environments — security without sacrificing performance.
- Adaptive security posture: Since threat models evolve, continuous intelligence updates help maintain protection against novel AI threats (jailbreaks, chained attacks) even after deployment.

10. Protecting NVIDIA DGX Servers and Kubernetes



Use Cases

- Secure DevOps access to AI fabric DGX nodes
- Protection for MGMT VLAN
- Block malicious code to be uploaded / downloaded (GitHub / GitLab)
- Secure Application Data Plane (API) user access (NW)
- Runtime Threat Detection with DOCA
- EW K8s access control and micro-segmentation (Ilumio)

Check Point security is embedded directly into NVIDIA BlueField DPUs through the DOCA framework, allowing enforcement of inline firewall and microsegmentation policies at the hardware-accelerated NIC level. This provides high-performance inspection of north–south Ethernet traffic (ingress/egress) before it reaches the GPU/CPU workloads inside the DGX server. The integration extends Check Point's advanced firewall and IPS protections without consuming host resources, preserving GPU cycles for AI workloads.

Key Features

- Hardware-accelerated FW/IPS: Runs Check Point's next-gen firewall engine on the DPU, enabling low-latency policy enforcement. Zero negative impact on AI system performance.
- Microsegmentation: Isolates applications, VLANs, and tenants at the NIC level, limiting blast radius of compromise.
- Traffic Offload: Optionally mirrors traffic to Check Point Maestro IDS for anomaly detection and advanced behavioral analytics.
- Policy Orchestration: Managed centrally via Check Point SmartCenter, ensuring consistent rules across DPUs, DGX servers, and Kubernetes clusters.

AppShield on BlueField

DOCA Argus is a DOCA service running on NVIDIA® BlueField® networking platforms, designed to immediately detect and enable response to attacks, minimizing their potential impact and risk.

The DOCA Argus framework provides real-time situational awareness and runtime threat detection by inspecting host memory using advanced memory forensics. Live machine introspection is performed at the hardware level, analyzing specific snippets of volatile host memory to monitor threats in real time without impacting system performance. DOCA Argus does not violate privacy, as information is extracted only from kernel structures.

Unlike conventional tools, Argus runs independently of the host, requiring no agents, integration, or reliance on host-based resources. This agentless, zero-overhead design enhances system efficiency and ensures resilient security in any compute environment, including bare-metal, virtualized, containerized, and multi-tenant infrastructures. By operating outside the host, isolated in its own trust domain, DOCA Argus remains invisible to attackers—even if the system is compromised.

Use Cases:

- **Threat Intelligence Correlation (Hash Reputation)**
  Extracts MD5/SHA hashes of executed files and checks them against Check Point ThreatCloud for malware detection.
- **SBOM Integrity Enforcement**
  Detects deviations from expected SBOM baselines, identifying tampered libraries, supply-chain attacks, or unauthorized components.
- **Reverse Shell Detection**
  Identifies suspicious process and syscall patterns that indicate reverse shells or post-exploitation C2 activity.
- **Automated Remediation via Playblocks / SOAR**
  When malicious behavior is detected, Check Point invokes and triggers playblocks or any SOAR system for remediation e.g. —blocking traffic and quarantining the compromised node / container, alerting SOC, enforcing security policies accordingly.

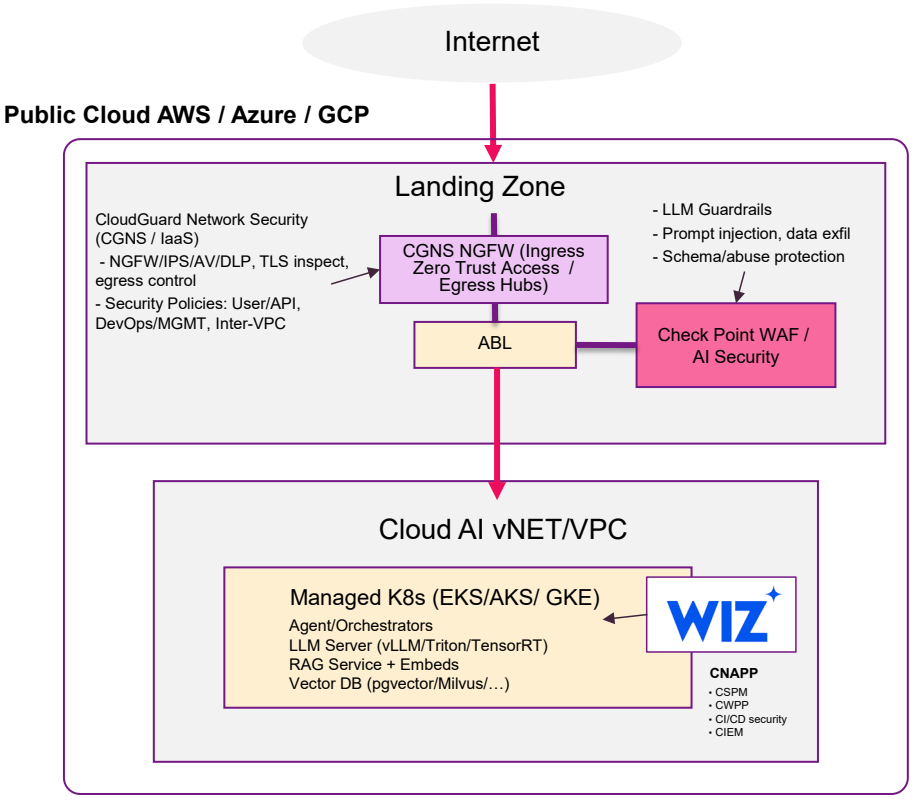
Illumio Containers Security

Illumio delivers deep visibility and micro-segmentation within Kubernetes clusters, operating via its C-VEN or agentless architecture to block lateral movement and isolate compromised workloads.

On detection of threats at the perimeter (for example, when the Check Point NGFW identifies bot traffic or access to a compromised website), NGFW logs can be forwarded to Illumio, enabling it to instantly quarantine affected containers and enforce enforcement policies at the node or pod level.

This combined approach ensures that an intrusion detected at the edge triggers immediate containment inside the cluster, preventing attacker propagation across containers or services. By linking perimeter threat prevention with runtime workload isolation, enterprises gain a robust, layered defense in containerized environments.

11. Protecting Public Cloud AI

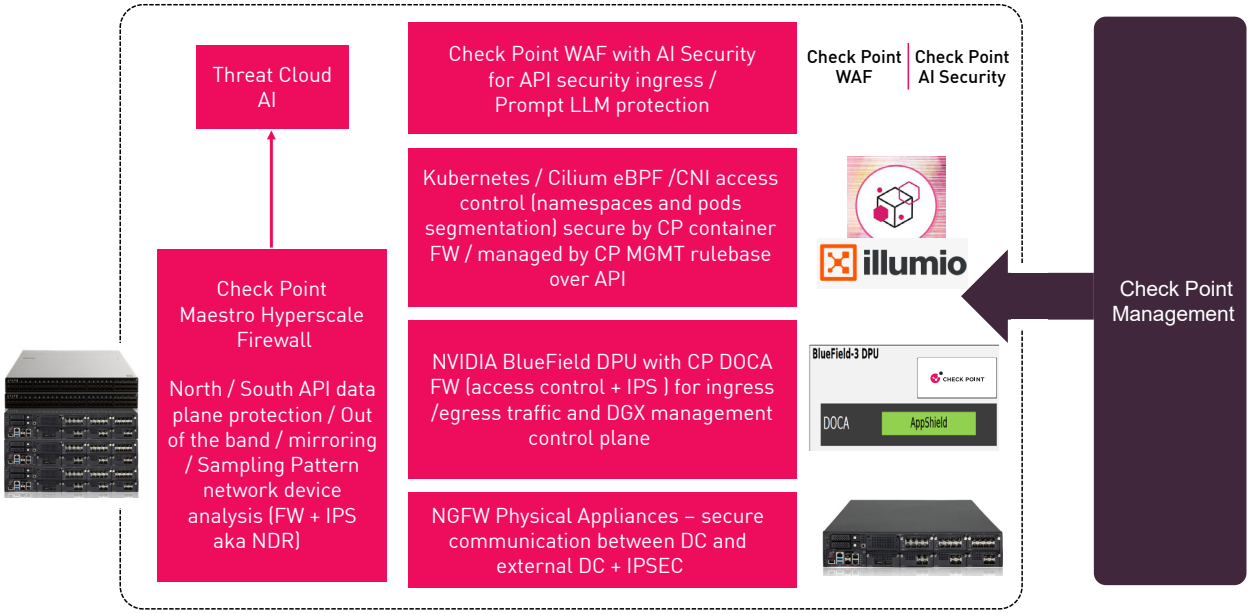


In public cloud environments, AI workloads and private LLMs are deployed across managed Kubernetes platforms such as Amazon EKS, Azure AKS, and Google GKE. The same layered security approach used in AI data centers is applied here, ensuring protection across network, application, and workload levels.

Check Point Cloud Firewall delivers ingress/egress firewalling, segmentation, and Zero Trust access control, while Check Point AI Security provides AI-native WAF protection to defend against prompt injection, data exfiltration, and schema abuse.

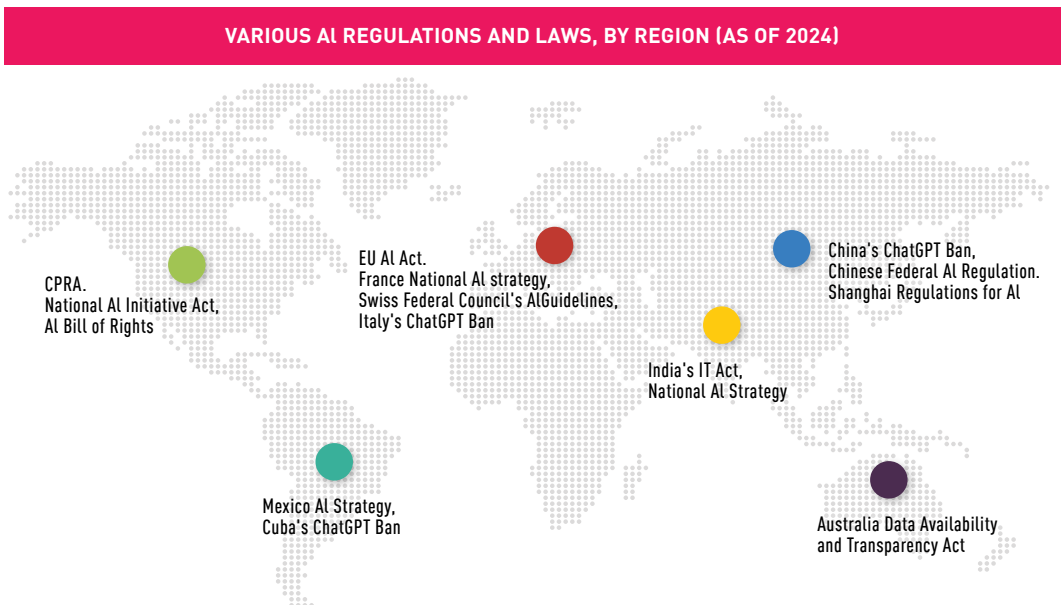
To secure cloud-native workloads, Wiz CNAPP integration strengthens posture management, vulnerability detection, and runtime protection across containers and Kubernetes clusters, creating a unified and compliant security fabric for public cloud AI infrastructures.

12. Security Technologies Full Stack for AI security



13. Alignment of Check Point AI DC Security with AI Governance Frameworks

*Through 2026, at least 80% of unauthorized AI transactions will be caused by **internal violations of enterprise / governance** policies concerning information oversharing, unacceptable use or misguided AI behavior **rather than malicious attacks (Gartner)***



Check Point's AI Data Center security blueprint protects infrastructure and workloads and also enables governance, trust, and compliance per **NIST AI RMF** and **Gartner AI TRiSM**:

Framework	Key Requirement	Check Point AI DC Alignment
NIST AI RMF	Governance & Transparency	SmartCenter delivers centralized policy management, audit trails, and full visibility across training and inference domains.
	Risk Mitigation	BlueField DPU integration with Check Point Firewall and Kubernetes segmentation prevents oversharing, misuse, and lateral movement.
	Lifecycle Security	End-to-end protection from data ingestion and model training to inference APIs exposed to end users (Check Point AI Agent Security)
Gartner AI TRiSM	Trustworthiness & Integrity	Check Point and BF AppShield provide runtime protection, detect anomalies, and stop model/output manipulation.
	Explainability & Accountability	Unified logging and monitoring allow traceability of API calls, user activity, and container behavior to meet audit/compliance needs.
	Policy Enforcement & Zero Trust	NGFWs, container security - enforce Zero Trust at management and data planes, reducing insider and external risks.

Value: By embedding AI-aware controls into the network, container, and application layers, Check Point extends beyond perimeter security to address policy-driven AI risks, ensuring organizations can meet governance standards, pass audits, and deploy AI responsibly.

14. Summary

Check Point's AI Data Center Security Blueprint enables organizations to adopt AI with confidence by protecting sensitive models, data, and applications from misuse and regulatory risk. Leveraging Maestro, integrated AI Security, and NVIDIA DPU integration, the blueprint delivers business continuity, regulatory compliance, and customer trust while reducing the risk of data leakage, insider threats, and unauthorized AI use. This approach safeguards multi-million-dollar AI investments, accelerates secure innovation, and ensures AI services remain reliable and resilient at scale.

About Check Point

Check Point Software Technologies Ltd. www.checkpoint.com) is a leading protector of digital trust, utilizing AI-powered cyber security solutions to safeguard over 100,000 organizations globally. Through its Infinity Platform and an open garden ecosystem, Check Point's prevention-first approach delivers industry-leading security efficacy while reducing risk. Employing a hybrid mesh network architecture with SASE at its core, the Infinity Platform unifies the management of on-premises, cloud, and workspace environments to offer flexibility, simplicity and scale for enterprises and service providers.

Worldwide Headquarters
5 Shlomo Kaplan Street, Tel Aviv 6789159, Israel | Tel: +972-3-753-4599

U.S. Headquarters
100 Oracle Parkway, Suite 800, Redwood City, CA 94065 | Tel: 1-800-429-4391
www.checkpoint.com