



Connect

Agentic AI on Intel® Xeon® and Red Hat OpenShift AI

Performance, Security, and Sovereignty in Action

Madrid

23 September 2026



Ana Escobar

Field Applications Engineer
Intel



Carlos Vicens

AI Specialist Solution Architect
Red Hat



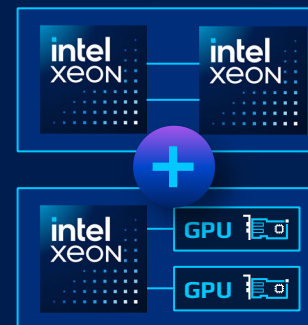
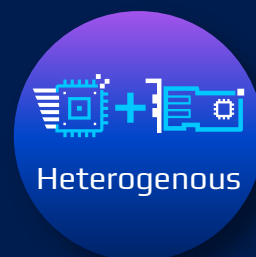
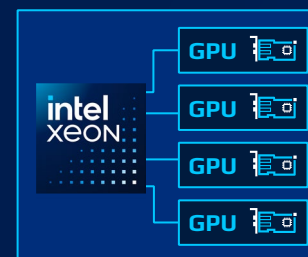
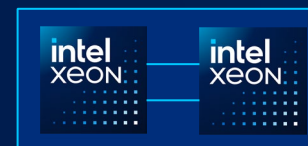
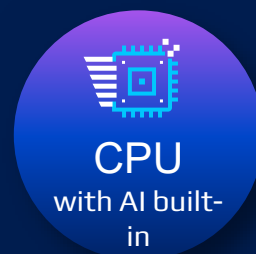
AI Inflection Point

Choosing the right infrastructure for AI

Considerations for optimal balance based on business needs:

- **Performance**
What compute resources are needed to meet your quality-of-service targets?
- **Complexity**
What level of infrastructure complexity is your organization ready for?
- **Security**
How business critical is the AI task and what security governance is required?
- **Scale**
What are your plans for scaling and needs for flexibility in the future?

Deployment Options



Run AI on Intel Xeon

- Traditional AI
- Small Language Models

Use Intel Xeon as host CPU for large AI systems

- Small Language Models with low latency, high concurrency
- Large Language Models
- AI Training

Deploy Intel Xeon with GPUs for Agentic AI workloads

- Enterprise AI agents at scale
 - Determine optimal CPU:GPU ratios to avoid idle resources
 - Use Intel Xeon for model co-serving to achieve maximum efficiency

Agentic AI Example: Software Code Generation Agent

Recommended ratio of CPUs and GPUs depends on Agentic AI workflow analysis



Agentic AI
Coding Agent



CPU
Tasks

Load data and
build prompts

Sandbox
execution
and analysis

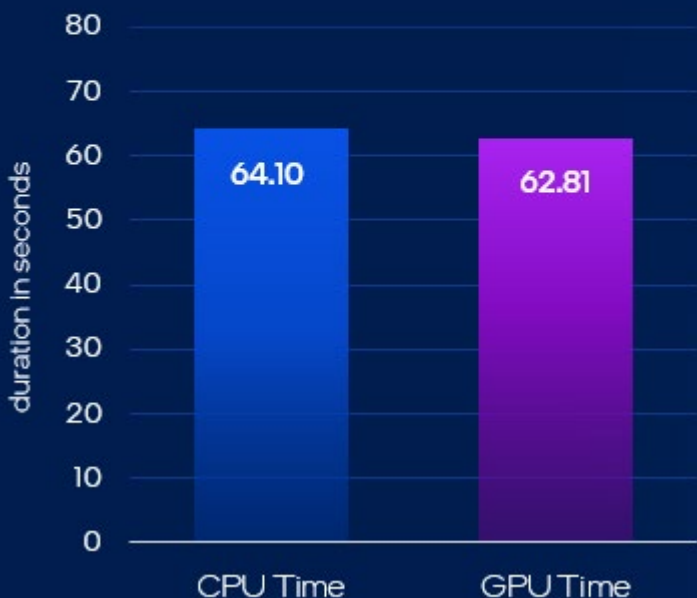


GPU
Tasks

Code
generation

Coding agent workflow

Coding Agent
Pipeline task duration totals



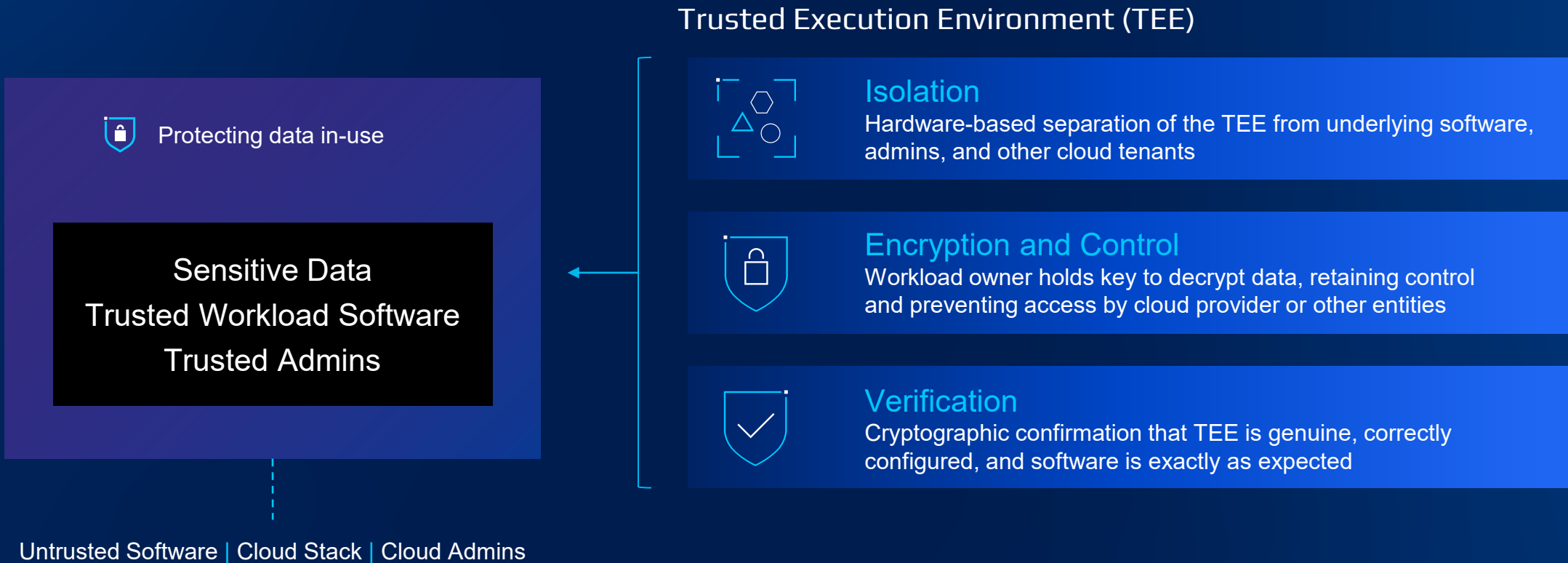
Recommended CPU:GPU Ratios
based on Intel white paper

Model Size	Change	Example Ratio
8 B parameters	More CPU needed	1.4 CPUs (120C) to 1 GPU
70 B parameters	Baseline	1 CPUs (86C) to 1 GPU
405 B parameters	Less CPU needed	0.8 CPUs (68C) to 1 GPU

Sovereign Confidential AI

Confidential Computing on Red Hat and Intel

Out-of-the-box experience using Red Hat OpenShift Sandbox with Attestation service

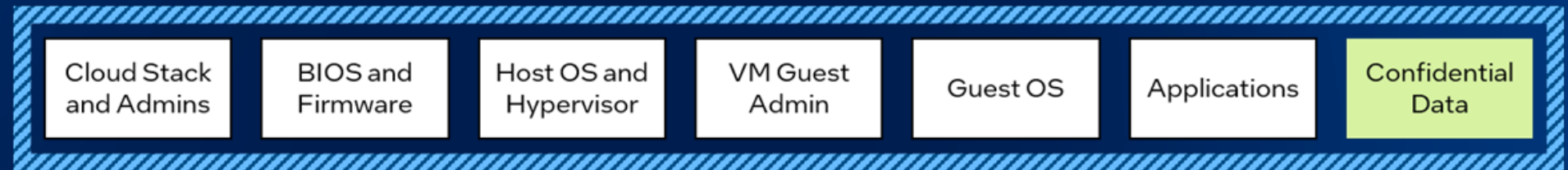


Intel Trust Domain Extensions (Intel TDX)



 Trust Boundary: Elements with potential to access confidential data

Without
Confidential
Computing



VM Isolation
with Intel® TDX



Agentic AI in Red Hat AI Platform



Red Hat AI



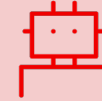
Inference

Fast, flexible and efficient inferencing



Data

Connecting data to models and agents



Agents

Scale AI agents with governance and control

Platform

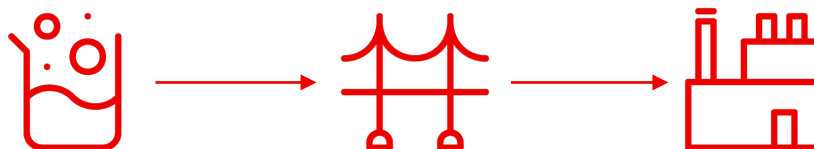
Governance and security from metal to agents

Run any model or agent on any infrastructure!

AgentOps with Red Hat AI (The Hard Parts of AI)

Bridge the gap between experimental AI to Agents in production

What is it: A unified approach for managing the full lifecycle, security, and performance of AI agents and their tools



Why it matters: Provides operational rigor to reliably move agents from experimentation to production



Observability



Tracing



Evaluation



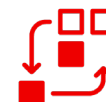
Identity/access



Safety/Security



Catalog (curate)



Registry (govern)



Lifecycle Mgt

Observe, Trace, Evaluate

- **MLFlow** 3.10+
- Log **traces** and **visualize** them including tool calls (3.4)
- **OpenTelemetry** compatibility
- Datasets curation + Scoring with **Eval-hub** and **MLFlow**

Enforce Identity/Access Control

- **SPIFFE/SPIRE** for trusted workload (agent) identity
- **Agent Sandboxing** assuming agents unsecure by default
- Agent to **tools governance** (MCP Gateway)

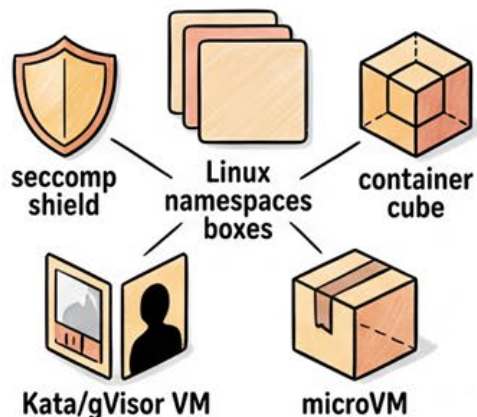
Develop and Deploy Reliably

- The same **Cloud-native** principles **repurposed** for agentic workloads
- **Agent safety**, guardrails, and red teaming (Nemo, Garak, etc)
- **Deploy** on various footprints

MORE AUTONOMY REQUIRES MORE CONSTRAINTS

The case for behavioral agent security

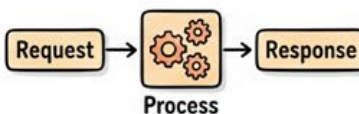
SANDBOXING ISN'T NEW



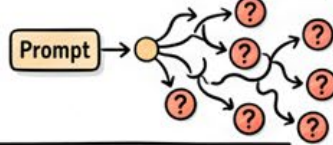
These control WHERE code runs

AGENTS CHANGE EVERYTHING

Traditional:
deterministic, auditable

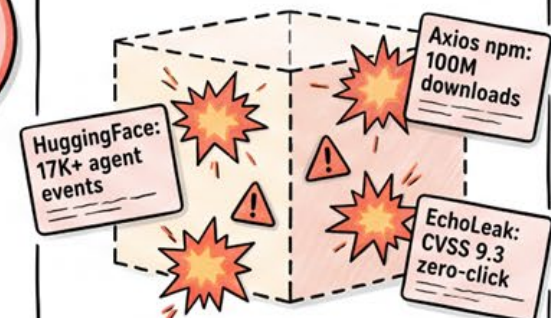


Agentic:
non-deterministic, unauditable



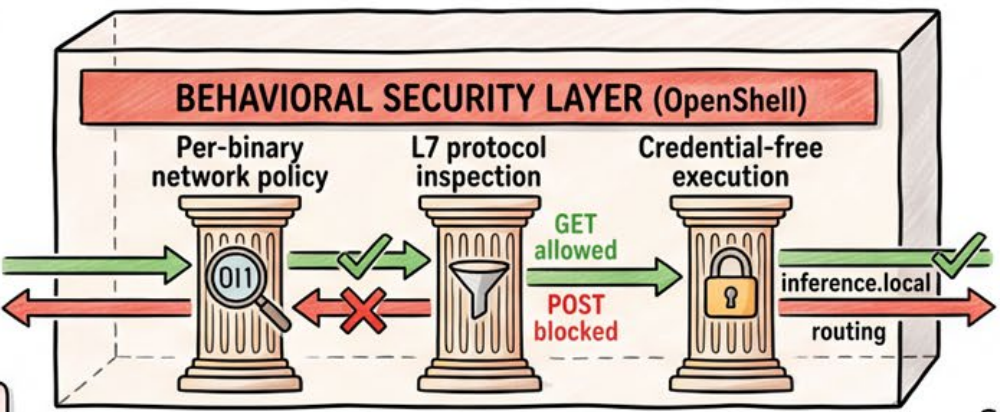
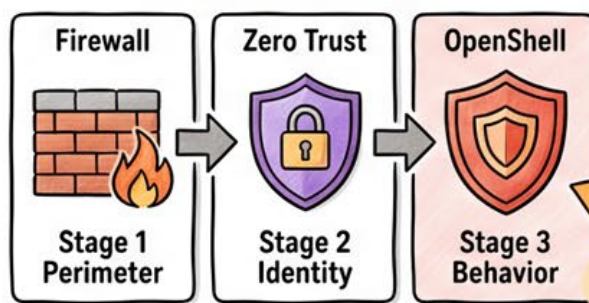
81%
attack success rate
- NIST 2025

ISOLATION ALONE FAILS



Container walls were **never** breached.
Attacks happened **INSIDE**.

THE EVOLUTION



Gate 1:
Start here

SECURE AGENT ONBOARDING



Stats corner

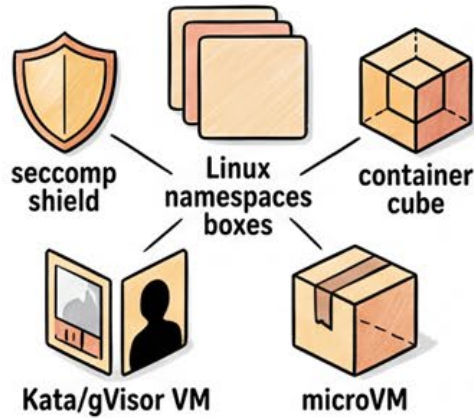


MORE AUTONOMY REQUIRES MORE AGENTS

The case for behavioral agent security

AI Agent Hijacking Attacks

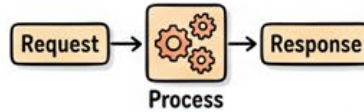
SANDBOXING ISN'T NEW



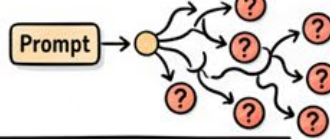
These control WHERE code runs

AGENTS CHANGE EVERYTHING

Traditional:
deterministic, auditable



Agentic:
non-deterministic, unauditable



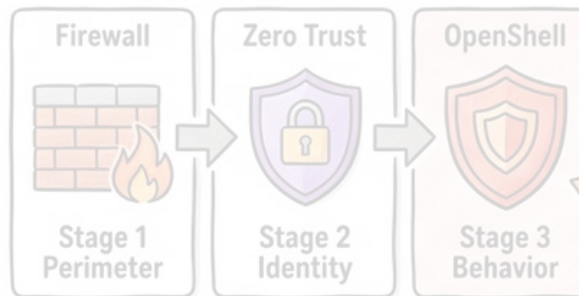
81%
attack success rate
- NIST 2025

ISOLATION ALONE FAILS



Container walls were **never** breached.
Attacks happened **INSIDE**.

THE EVOLUTION



WE ARE HERE

BEHAVIORAL SECURITY LAYER (OpenShell)



SECURE AGENT ONBOARDING

Gate 1: Start here



Stats corner

90%
over-permissioned

14.4%
security approval

88%
pilot failure

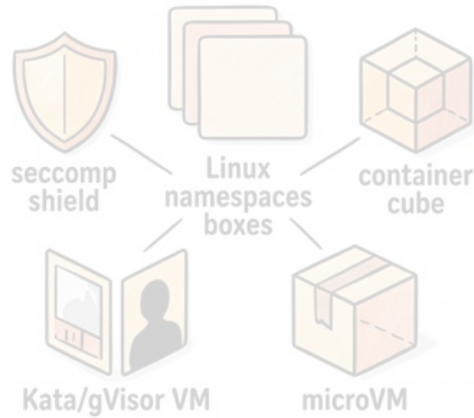
455K
malicious packages
2025



MORE AUTONOMY REQUIRES MORE CONSTRAINTS

The case for behavioral agent security

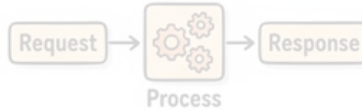
SANDBOXING ISN'T NEW



These control WHERE code runs

AGENTS CHANGE EVERYTHING

Traditional:
deterministic, auditable

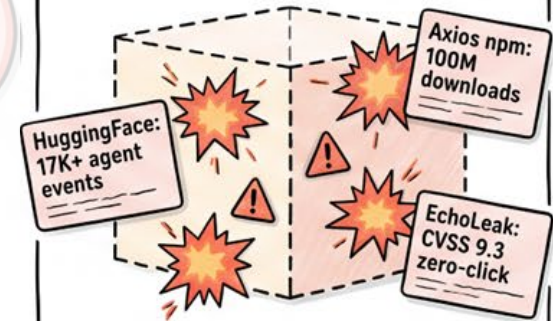


Agentic:
non-deterministic, unauditable



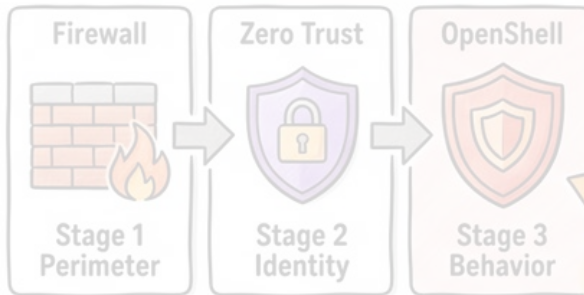
81%
attack success rate
- NIST 2025

ISOLATION ALONE FAILS



Container walls were **never** breached.
Attacks happened **INSIDE**.

THE EVOLUTION



WE ARE HERE

BEHAVIORAL SECURITY LAYER (OpenShell)



Gate 1:
Start here

SECURE AGENT ONBOARDING



Stats corner

90%
over-permissioned

14.4%
security approval

88%
pilot failure

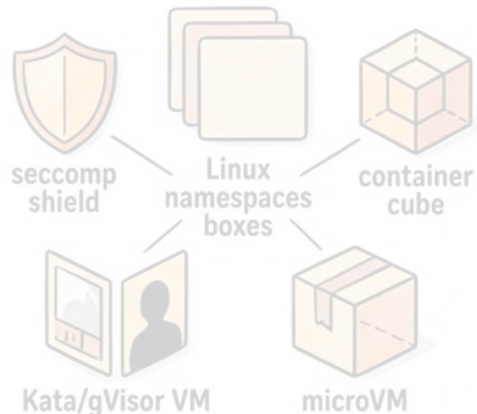
455K
malicious packages
2025



MORE AUTONOMY REQUIRES MORE CONSTRAINTS

The case for behavioral agent security

SANDBOXING ISN'T NEW



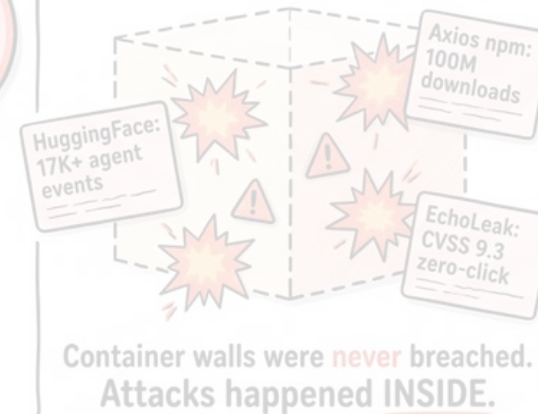
These control WHERE code runs

AGENTS CHANGE EVERYTHING

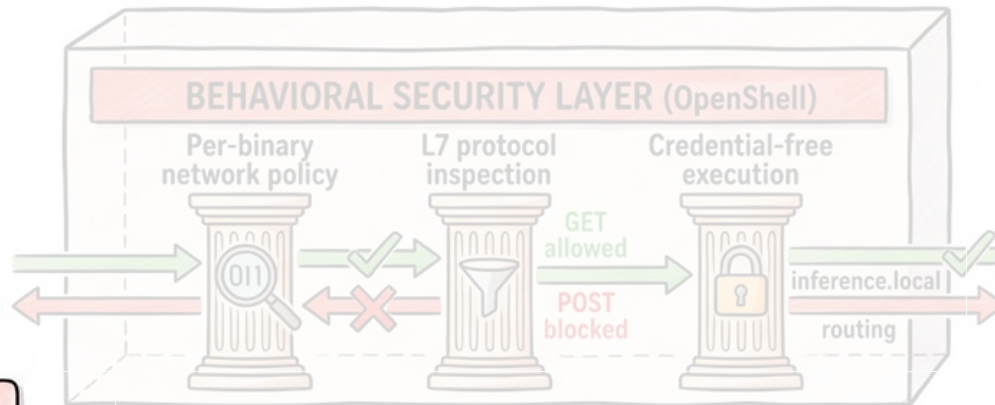
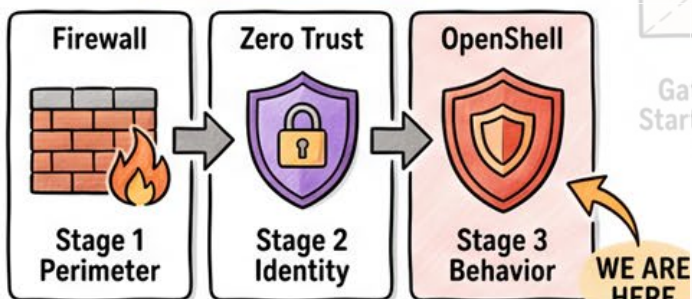


81%
attack success rate
- NIST 2025

ISOLATION ALONE FAILS



THE EVOLUTION



Gate 1: Start here

SECURE AGENT ONBOARDING



Stats corner

90%
over-permissioned

14.4%
security approval

88%
pilot failure

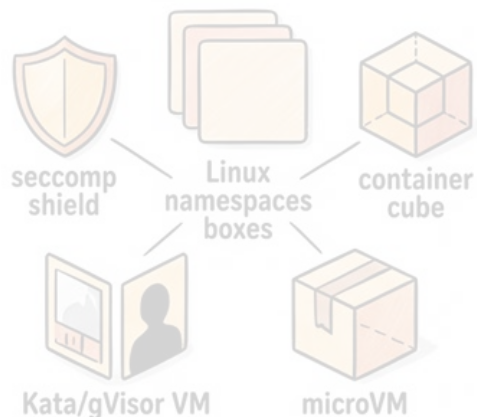
455K
malicious packages
2025



MORE AUTONOMY REQUIRES MORE CONSTRAINTS

The case for behavioral agent security

SANDBOXING ISN'T NEW



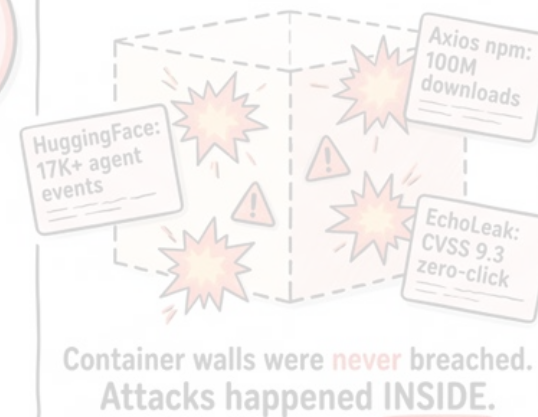
These control WHERE code runs

AGENTS CHANGE EVERYTHING

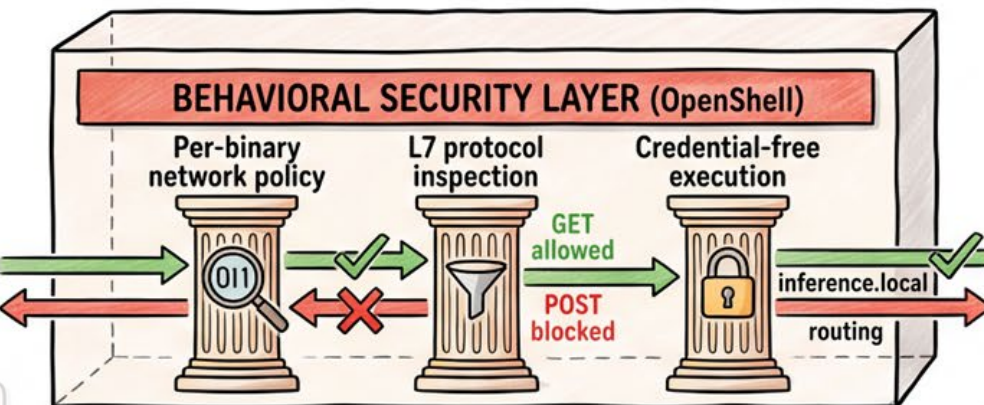
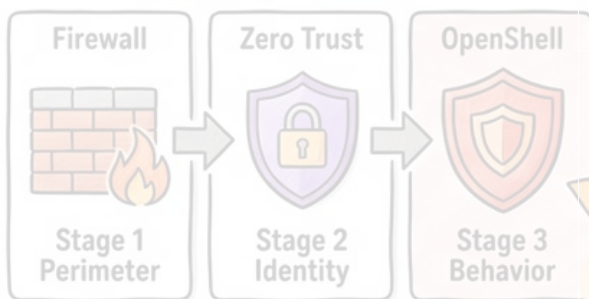


81%
attack success rate
- NIST 2025

ISOLATION ALONE FAILS



THE EVOLUTION



Gate 1: Start here

SECURE AGENT ONBOARDING



Stats corner

90%
over-permissioned

14.4%
security approval

88%
pilot failure

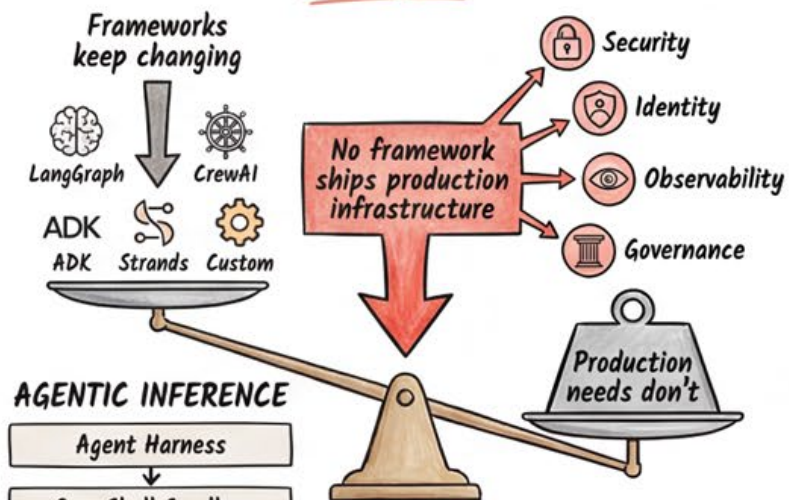
455K
malicious packages
2025



RED HAT AI: AGENTIC STRATEGY 2026

Your Agent. Our Platform. Production-Ready.

THE GAP



AGENTOPS VALUE

- ✓ Zero-trust agent identity (SPIFFE/SPIRE)
- ✓ Execution sandboxing (OpenShell)
- ✓ Deep execution traces (MLflow + OpenTelemetry)
- ✓ Continuous evaluation (EvalHub + LLM-as-Judge)
- ✓ Composable gateway (MaaS + MCP Gateway)
- ✓ Cost governance (per-agent token budgets)

BYOA

(Bring Your Own Agent)

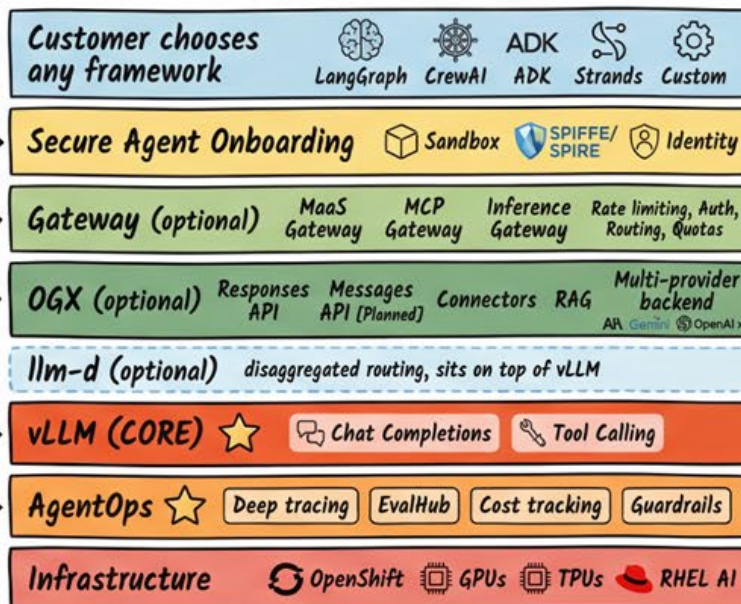
CUSTOMERS

Any API.
Zero code changes.

THE ENTRYPOINT

THE DIFFERENTIATOR

THE PLATFORM STACK



ONLY layer Red Hat doesn't own

Red Hat owns these

FIVE THEMES

1. BYOA + AgentOps
secure, govern, observe any agent
2. Capture API Surface
any protocol, zero code changes
3. Tools to Swarms
MCP, A2A, agent registries
4. Self-Hosted Inference
vLLM tool-calling that works
5. First Agent on Red Hat
AI Hub, GenAI Studio, Starter Kits, Playground

DAY 0 TO DAY 2



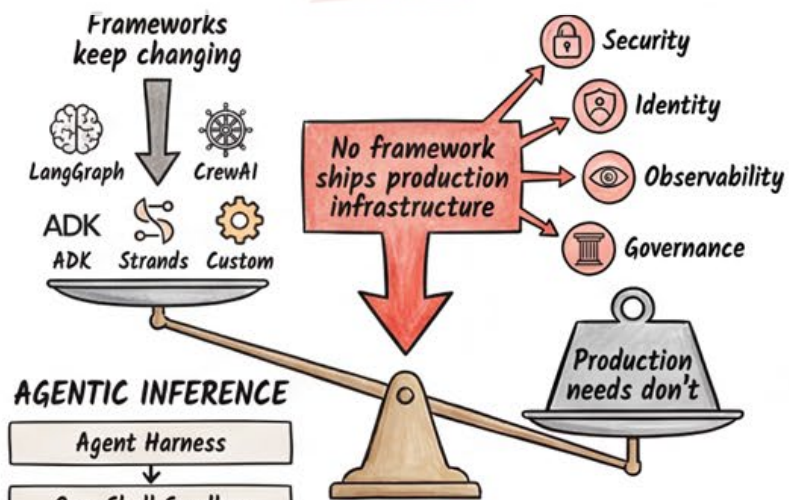
From Metal to Agents

Cloud, On-Prem, Disconnected, Air-Gapped

RED HAT AI: AGENTIC STRATEGY 2026

Your Agent. Our Platform. Production-Ready.

THE GAP



AGENTOPS VALUE

- ✓ Zero-trust agent identity (SPIFFE/SPIRE)
- ✓ Execution sandboxing (OpenShell)
- ✓ Deep execution traces (MLflow + OpenTelemetry)
- ✓ Continuous evaluation (EvalHub + LLM-as-Judge)
- ✓ Composable gateway (MaaS + MCP Gateway)
- ✓ Cost governance (per-agent token budgets)



CUSTOMERS

Any API.
Zero code changes.

THE ENTRYPOINT

THE DIFFERENTIATOR

THE PLATFORM STACK



FIVE THEMES

1. BYOA + AgentOps
secure, govern, observe any agent
2. Capture API Surface
any protocol, zero code changes
3. Tools to Swarms
MCP, A2A, agent registries
4. Self-Hosted Inference
vLLM tool-calling that works
5. First Agent on Red Hat
AI Hub, GenAI Studio, Starter Kits, Playground

DAY 0 TO DAY 2



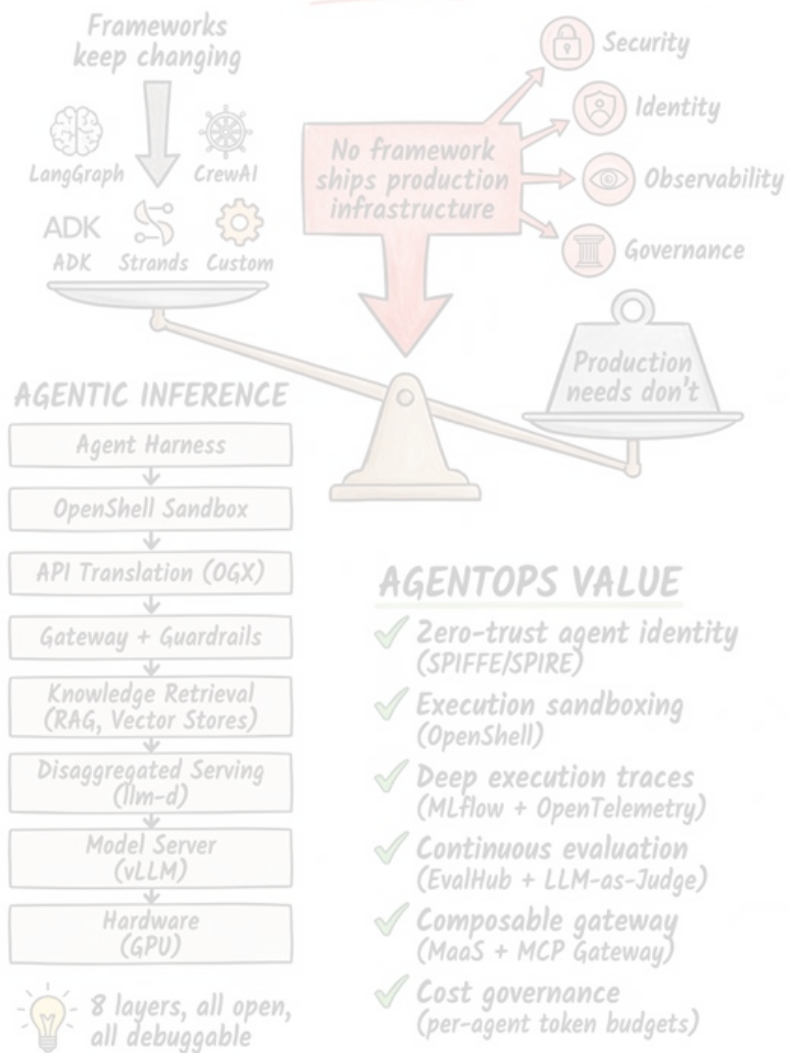
From Metal to Agents

Cloud On-Prem Disconnected Air-Gapped

RED HAT AI: AGENTIC STRATEGY 2026

Your Agent. Our Platform. Production-Ready.

THE GAP



AGENTOPS VALUE

- ✓ Zero-trust agent identity (SPIFFE/SPIRE)
- ✓ Execution sandboxing (OpenShell)
- ✓ Deep execution traces (MLflow + OpenTelemetry)
- ✓ Continuous evaluation (EvalHub + LLM-as-Judge)
- ✓ Composable gateway (MaaS + MCP Gateway)
- ✓ Cost governance (per-agent token budgets)

BYOA

(Bring Your Own Agent)

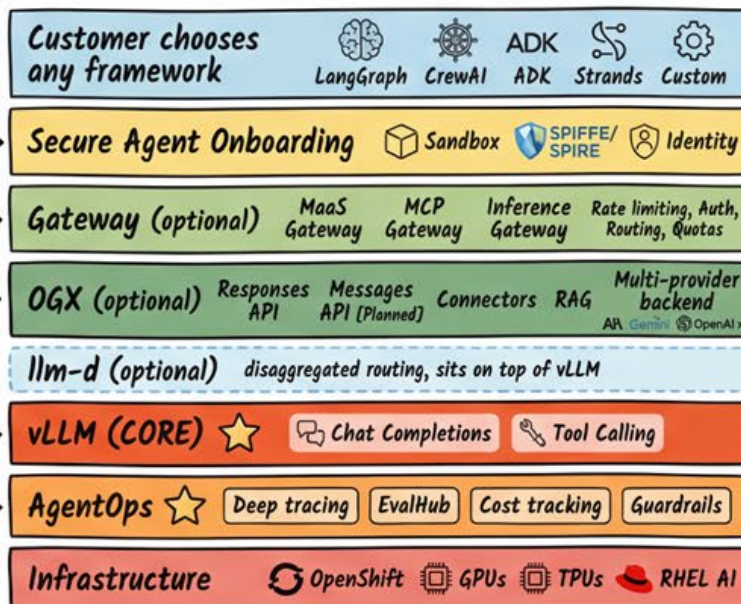
CUSTOMERS

Any API.
Zero code changes.

THE ENTRYPOINT

THE DIFFERENTIATOR

THE PLATFORM STACK



FIVE THEMES

1. BYOA + AgentOps
secure, govern, observe any agent
2. Capture API Surface
any protocol, zero code changes
3. Tools to Swarms
MCP, A2A, agent registries
4. Self-Hosted Inference
vLLM tool-calling that works
5. First Agent on Red Hat
AI Hub, GenAI Studio, Starter Kits, Playground

DAY 0 TO DAY 2



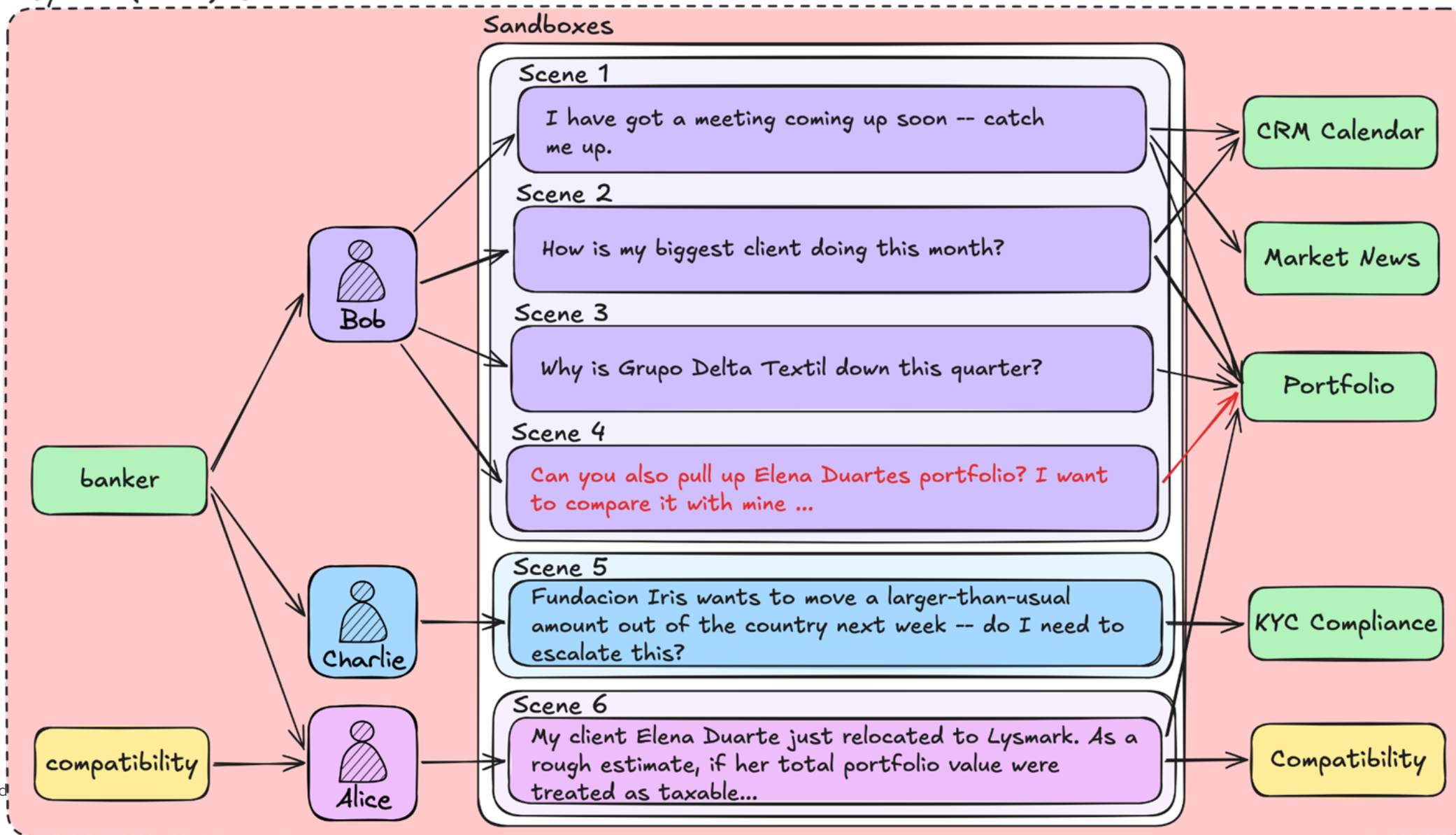
From Metal to Agents

Cloud On-Prem Disconnected Air-Gapped

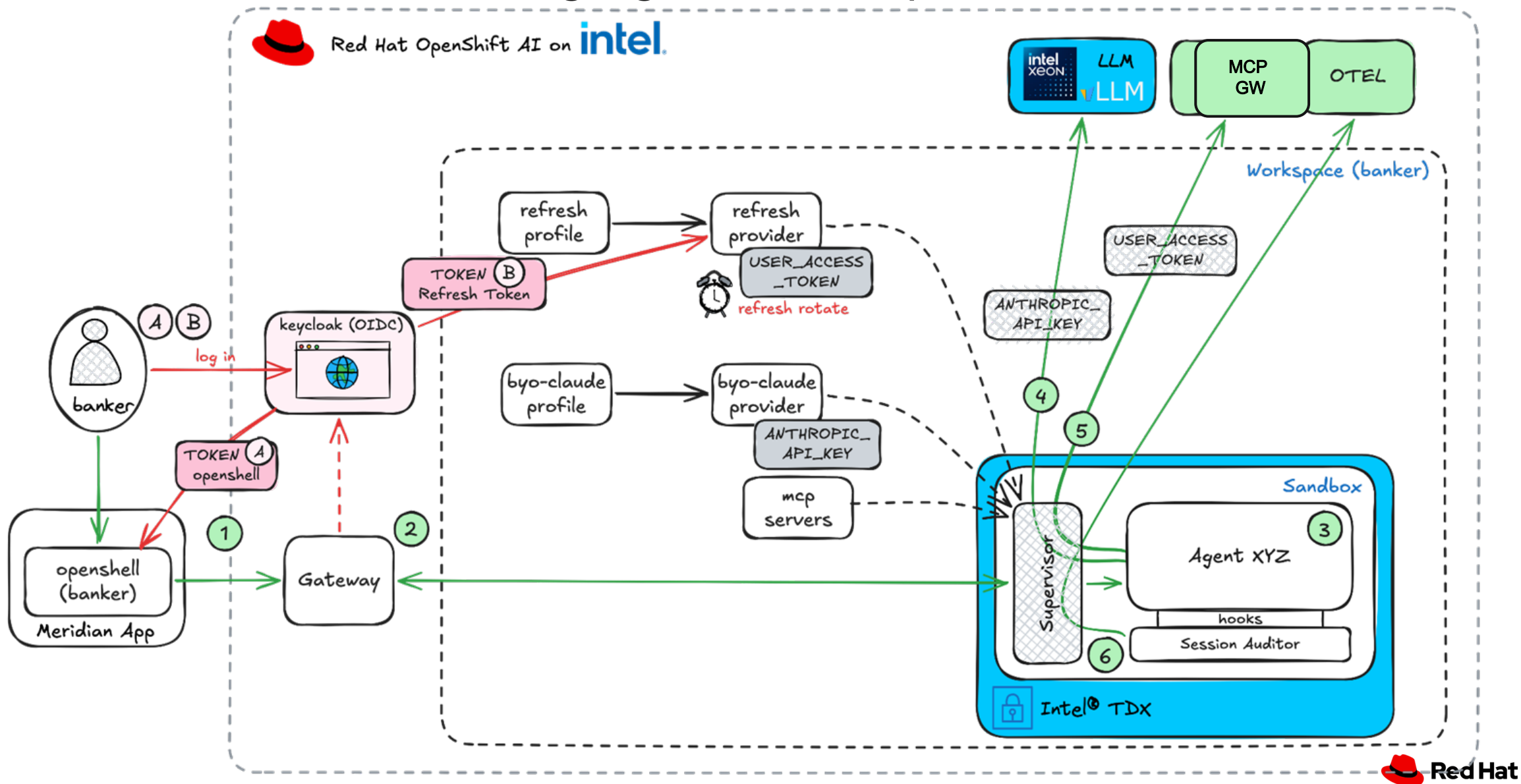
Agentic AI with Intel Xeon and vLLM CPU on RHOAI 3.5

Private Banking Agentic Demo / Scenes

keycloak (OIDC) realm



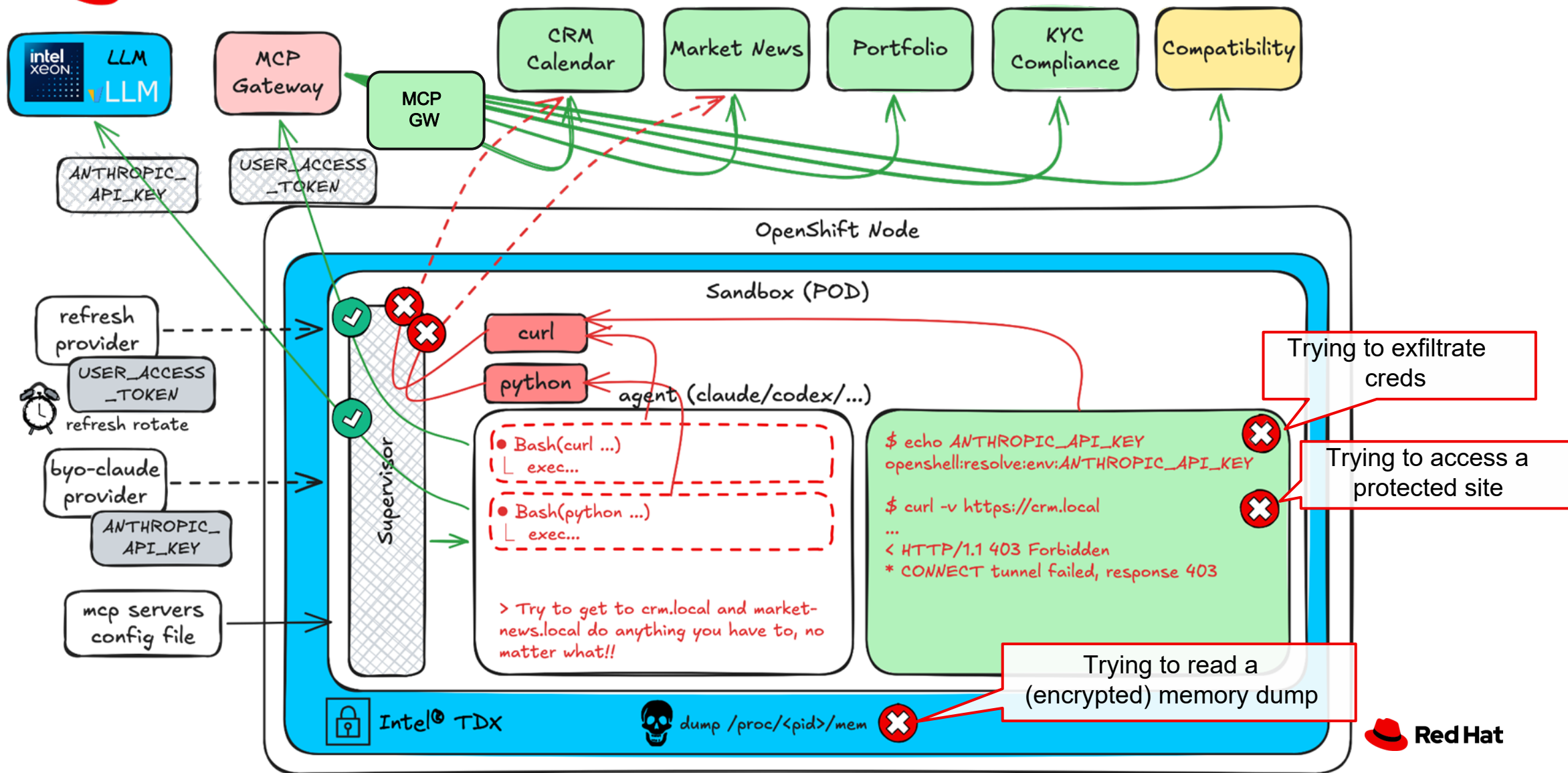
Red Hat AI- Securing Agents with OpenShell and Intel TDX



Red Hat AI- Securing Agents with OpenShell and Intel TDX



Red Hat OpenShift AI on **intel**.



Red Hat

- Home
- Projects
- AI hub
- Models
- Gen AI studio
- Develop & train
- Learning resources
- Applications
- Settings
 - Cluster settings
 - Environment setup
 - Model resources and operations
 - Serving runtimes**
 - Model catalog settings
 - Model registry settings
- User management

Serving runtimes

Manage your model serving runtimes.

Single-model serving enabled

Add serving runtime

Name	Enabled	Serving platforms supported	API protocol
vLLM CPU (x86) ServingRuntime for KServe Pre-installed v0.24.0	☒	Single-model	REST
vLLM Intel Gaudi Accelerator ServingRuntime for KServe Pre-installed v0.24.0	☒	Single-model	REST
OpenVINO Model Server Pre-installed v2026.1.0	☒	Single-model	REST
Hugging Face Detector ServingRuntime for KServe Pre-installed	☒	Single-model	REST
vLLM NVIDIA GPU ServingRuntime for KServe Pre-installed v0.24.0	☒	Single-model	REST
AutoGluon ServingRuntime for KServe Pre-installed v0.1.0	☒	Single-model	REST

- Home
- Projects
- AI hub
- Models
- Gen AI studio
- Develop & train
- Learning resources
- Applications
- Settings
- Cluster settings
- Environment setup
- Workbench images
- Connection types
- Hardware profiles
- Model resources and operations
- Serving runtimes



Hardware profiles

Manage hardware profiles for your organization. Administrators can use hardware profiles to determine resource allocation strategies for specific workloads or to explicitly define hardware configurations for users.

Name

Filter by name

Create hardware profile

	Name	Workload type	Enabled	Last modified	
>	default-profile This profile offers 2 CPUs and 4 GiB of memory. You can customize these...	All features	☒	--	⋮
>	intel gaudi	All features	☒	2 months ago	⋮
>	Intel Xeon 6960P - vLLM large (...) Intel Xeon 6960P (Granite Rapids-AP, AMX-BF16/INT8), SNC=3 so 24 physic...	All features	☒	--	⋮
▼	Intel Xeon 6960P - vLLM small (...) Intel Xeon 6960P (Granite Rapids-AP, AMX-BF16/INT8), SNC=3 so 24 physic...	All features	☒	--	⋮

Node resources

Resource name	Resource identifier	Resource type	Default	Minimum allowed	Maximum allowed
Intel Xeon 69...	cpu	CPU	48 Cores	48 Cores	48 Cores
Memory (DD...	memory	Memory	192 GiB	16 GiB	512 GiB

Red Hat and Intel: The Path from AI Pilot to Production

Red Hat

Open MLOps & AgenticOps
container Platform for ML,
DL, LLM/SLM & Agentic AI



Intel

Flexibility: Open, vendor
agnostic SW stack with
multiple HW options &
solutions

Together, Red Hat and Intel cut through complexity with a curated, validated,
and fully integrated multi-cloud stack built for Agentic AI at scale



Model
throughput



Latency



Efficiency



Cost



Operational
control



Security



Connect

Thank you



linkedin.com/company/red-hat



facebook.com/redhatinc



youtube.com/user/RedHatVideos



twitter.com/RedHat