Red Hat Summit

Connect

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

Performance, Security, and Sovereignty in Action

Utrecht 11th November 2026

Red Hat | intel.

Legal Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for additional details.

No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

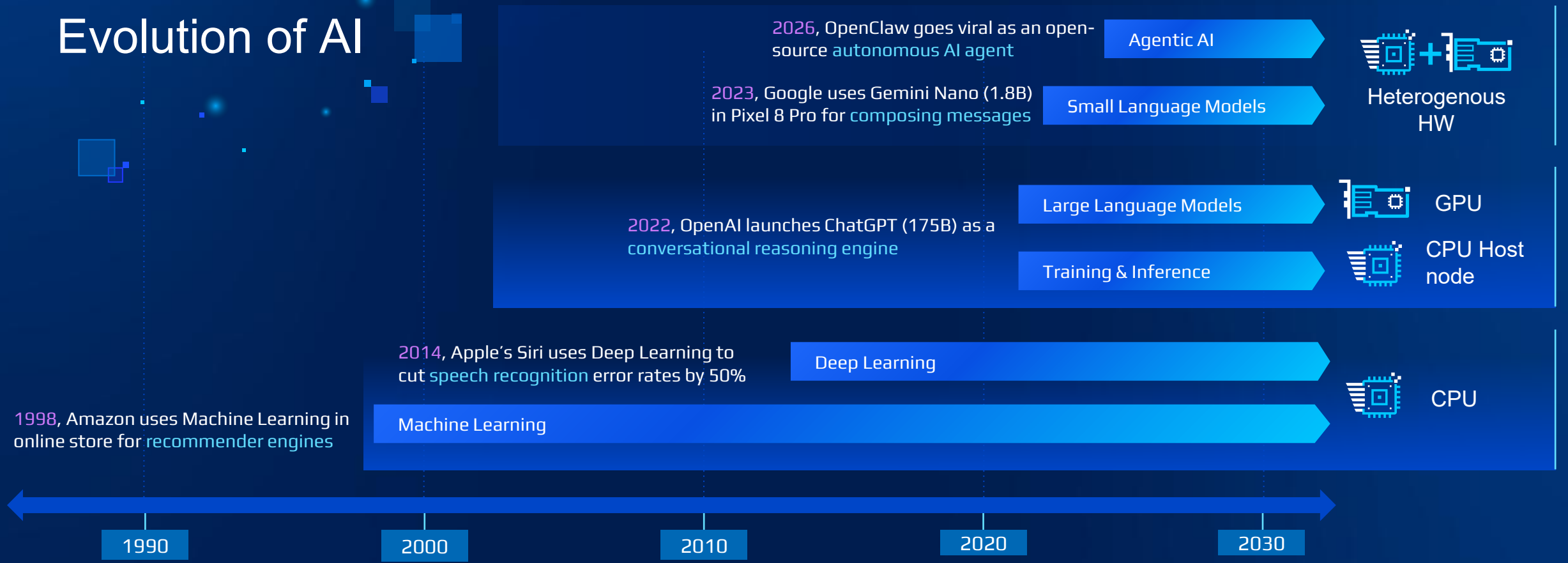
© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

AI Inflection Point

Red Hat | intel.

Advances in AI are evolving, accelerating and stacking

HW needs are changing and SW is responding with flexible, fully integrated solutions.



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



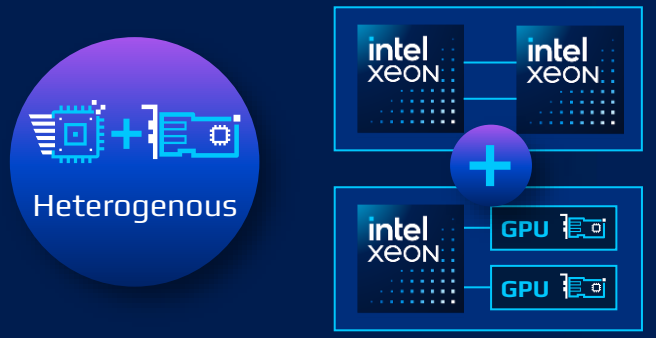
Run AI on Intel Xeon

- 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



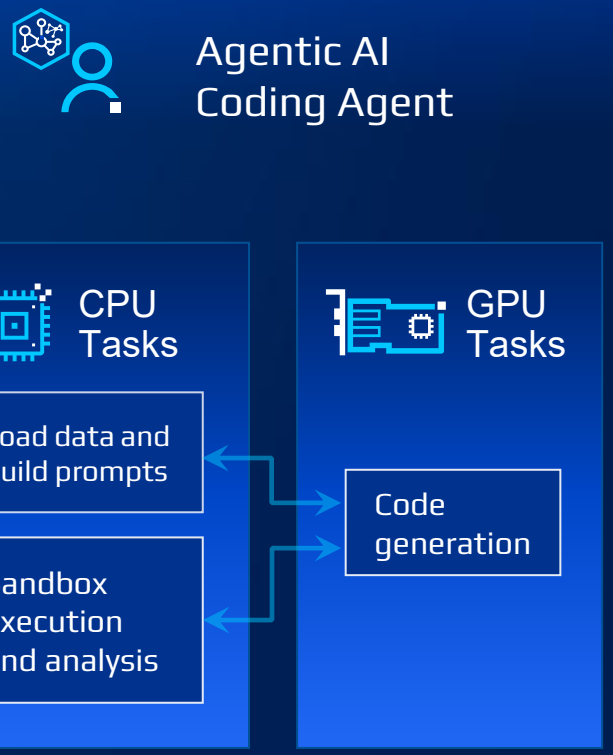
Deploy Intel Xeon with GPUs for Agentic AI workloads

- 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

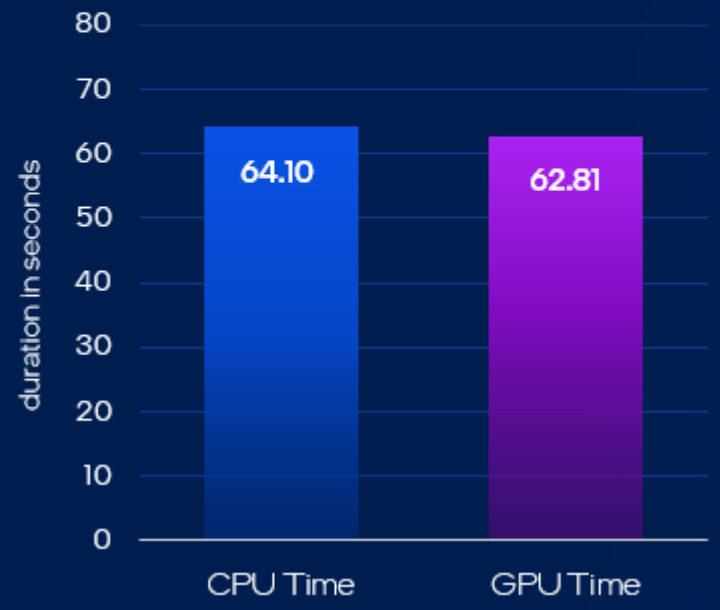
Red Hat × intel

Agentic AI Example: Software Code Generation Agent

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



Coding agent workflow



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

Source: Intel White Paper: Agentic AI Requires More CPUs



Red Hat® is the world's leading provider of open source enterprise software solutions, offering enterprise-grade Linux, cloud, container, and Kubernetes technologies to organizations. Their portfolio empowers businesses to build, deploy, and manage applications consistently across on-premise, hybrid, and multi-cloud environments. Intel and Red Hat are collaborating to provide optimized, open source AI platforms with Xeon that enable organizations to develop and scale applications.

The CPU has become increasingly important in AI that Red Hat OpenShift AI 3.4 became the first release to officially start supporting Xeon. This consists of vLLM model serving images that enable Xeon's Advanced Matrix Extensions (AMX) to accelerate inference on select popular Red Hat models. These models are used on AI Quickstarts, business use cases that can be rapidly deployed with minimal steps. Running managed applications on Xeon with Red Hat's ecosystem simplifies development and provides security across hybrid and multi-cloud environments.

Find out more about the Red Hat and Intel partnership.

Visit Red Hat Website

Visit Intel Software Catalogue

INTEL® GPU D-SERIES
CODENAMED

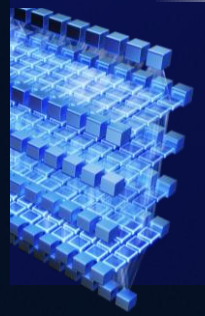
Crescent Island

Compute designed for inference

BUILT FOR AGENTIC AI

1311/2622
TFLOPs
FP8/MxFP4

32
Xᵉ Cores
Xᵉ 3P GPU IP



256
XMX engines

16
deep systolic

Xᵉ Matrix Extensions (XMX)

LARGE CONTEXT CAPACITY

1.5TB/s
Memory bandwidth

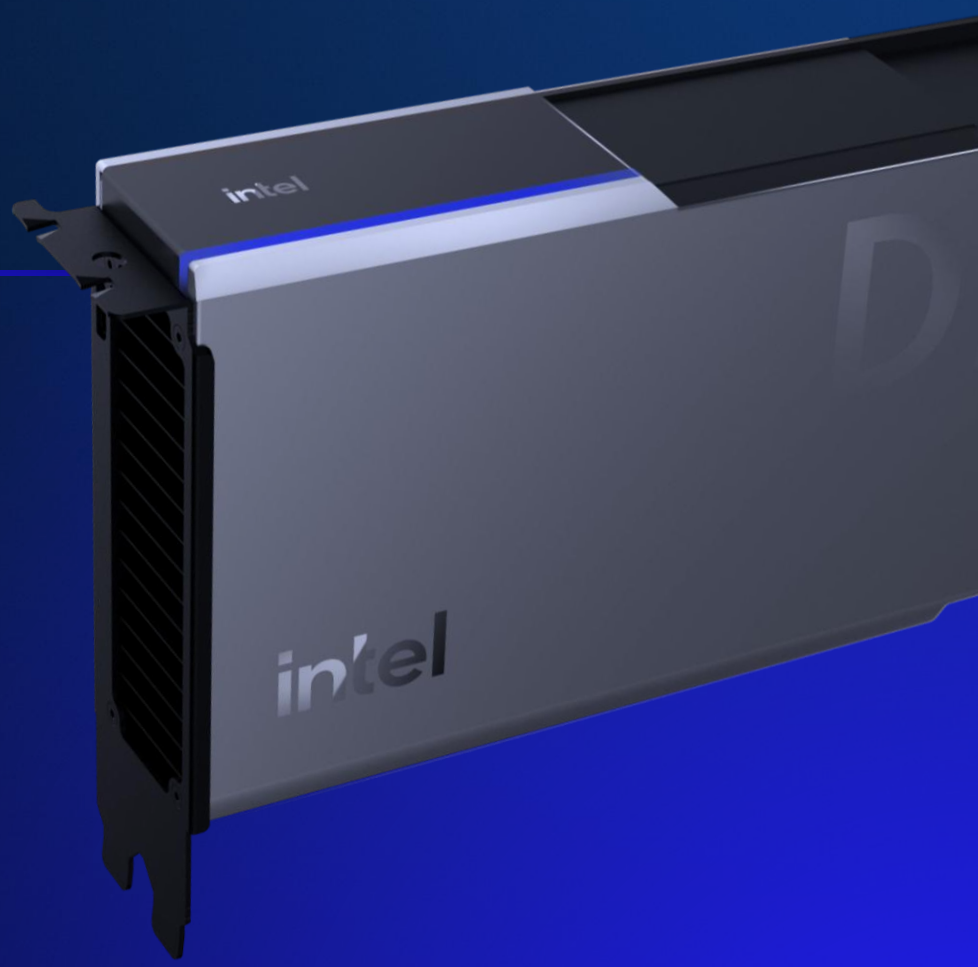
Up to
480GB*
LPDDR5x

LOW-POWER FAMILIAR FORM

350W
Air-cooled PCIe

OPEN SOFTWARE STACK

Day 0
Functionality



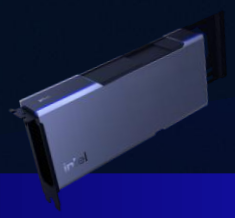
*Intel branded PCIe card has 160GB LPDDR5x; design enables partners to build ODM branded cards with flexible options up to 480GB memory.

intel

INTEL® GPU D-SERIES CODENAMED

Crescent Island Scalable Performance

Across the full AI model continuum



x1



x8



x80

Throughput

AI Model Continuum

GPT-oss	Kolibri-1	Llama 4	Qwen3	Deep Seek	GLM-4.5	Llama	Trinity	Deep Seek	GLM 5.3	Kimi K2	Mistral	Deep Seek	Grok 3	Kimi K3
20B	78B	Scout	235B	V4-Flash	-	Maverick	400B	V3-R1	753B	K2, K2.5 & K2.6	Le Chonk	V4 Pro	2.7T	

Fast Agents | Multimodal Doc Q&A | Agentic Coding | Knowledge Work | Agentic Workflows | Frontier Reasoning

Routing | High Vol. Chat | Multimodal Reasoning | Agent Swarms | Long Horizon Coding

intel

Sovereign Confidential AI

Red Hat | intel.

Sovereign AI infrastructure – Wherever you need it

Deploy AI with Verifiable Trust across any environment with Intel Confidential Computing Solutions

Public Cloud



Run Sovereign workloads on Hyperscale infrastructure

Private Cloud



Maintain control over infrastructure for highly regulated workloads

Hybrid & Multi Cloud



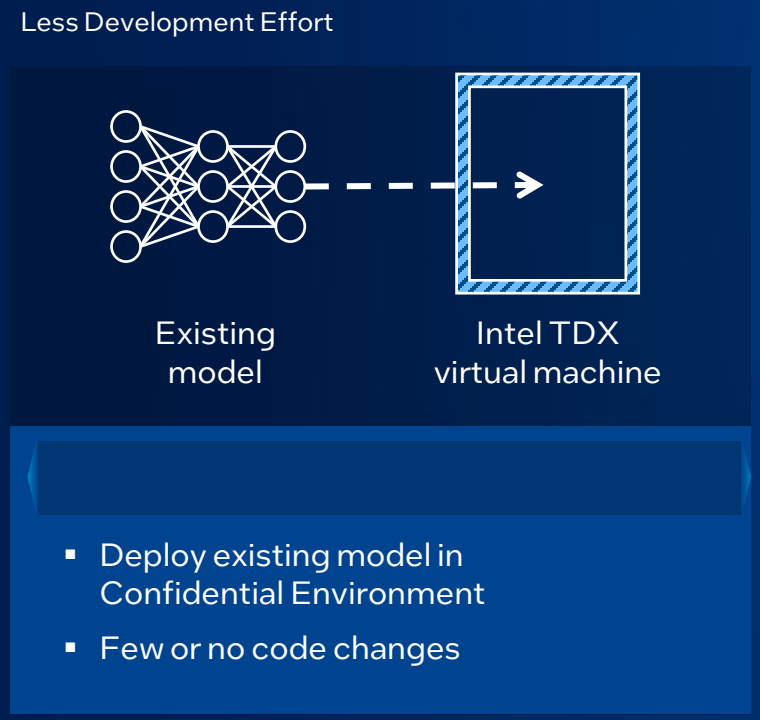
Move workloads without compromising security, governance or compliance

Confidential Computing with Intel® TDX

Sovereign AI with Intel® TDX

Sovereign AI isn't just about where data lives. It's about who owns the models, the pipelines, and the proof.

- Intel ® TDX protects container data while it's being processed
- Trust becomes verifiable
- Provides a trusted, scalable AI infrastructure foundation with security embedded from silicon to system.
- Extends governance, security, and AI processing consistently across distributed locations. The same architecture works across public, private, and hybrid cloud
- Agentic AI requires stronger protection for dynamic environment



Confidential Computing for Digital Sovereignty

Prevalent Protection Mechanisms



Data Protection at Rest



Data Protection in Flight

New Protection Frontier by Confidential Computing



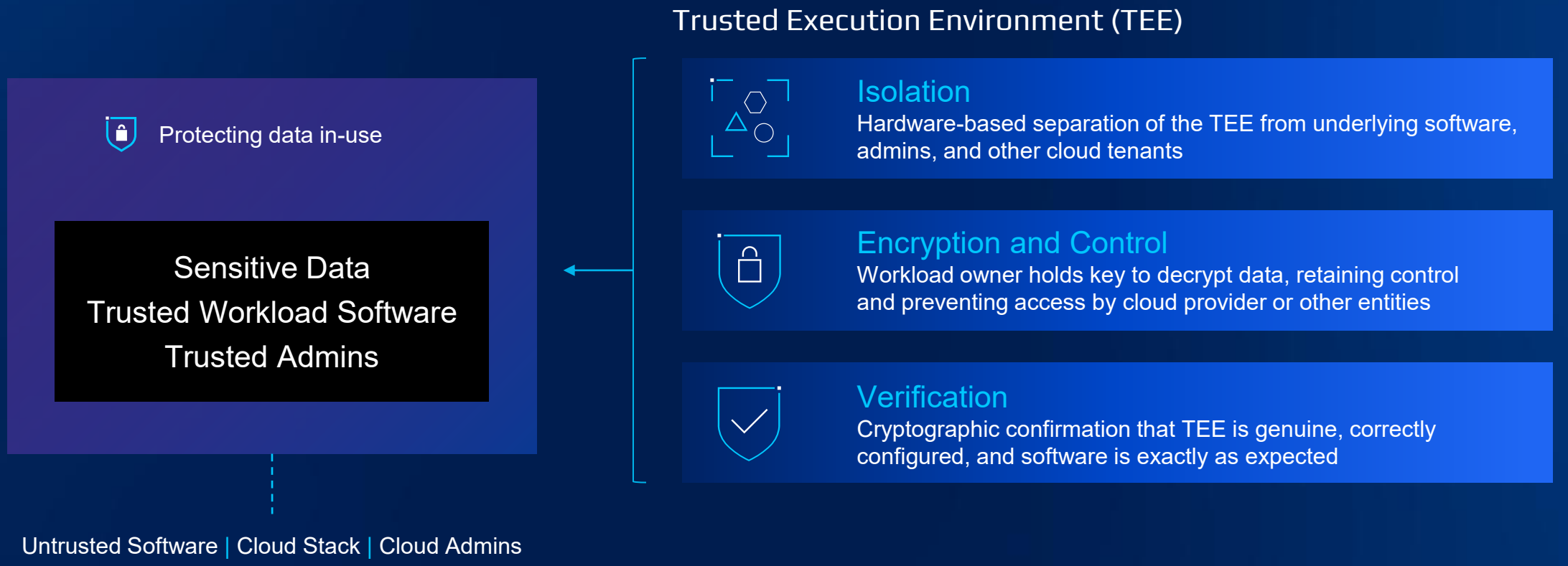
Data Protection in Use



Attestation

Confidential Computing on Red Hat and Intel

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



Agentic AI in Red Hat AI Platform

Red Hat | intel.

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**



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

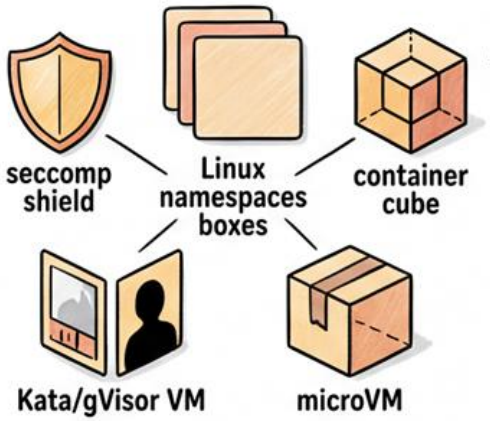
- The same **Cloud-native** principles **repurposed** for agentic workloads
- **Agent safety**, guardrails,and red teaming (Nemo, Garak, etc)
- **Deploy** on various footprints

Red Hat

MORE AUTONOMY REQUIRES MORE CONSTRAINTS

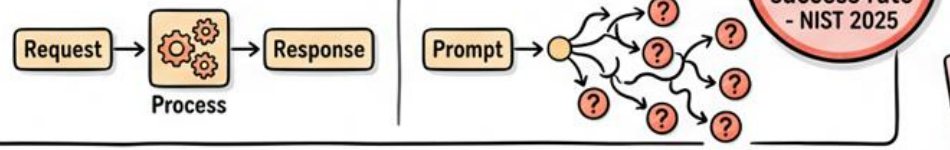
The case for behavioral agent security

SANDBOXING ISN'T NEW



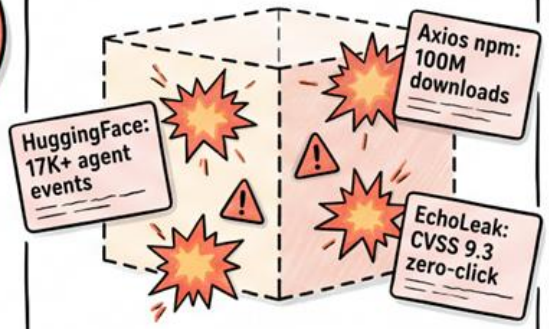
These control WHERE code runs

AGENTS CHANGE EVERYTHING



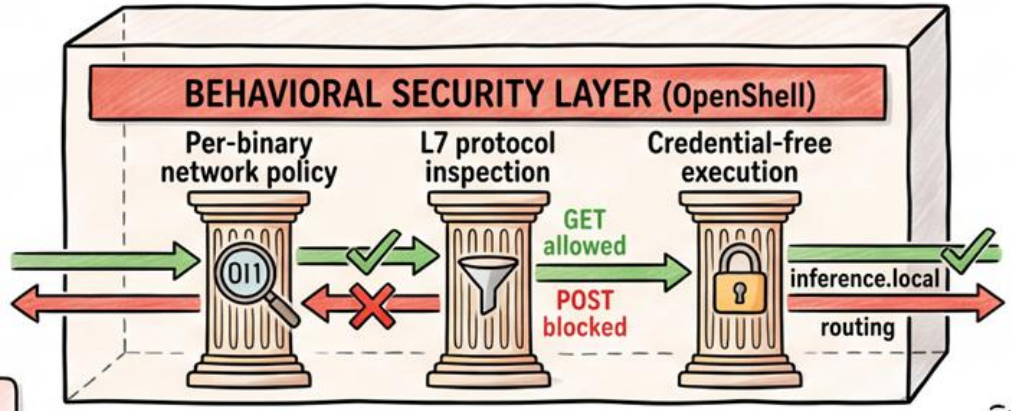
81% attack success rate - NIST 2025

ISOLATION ALONE FAILS

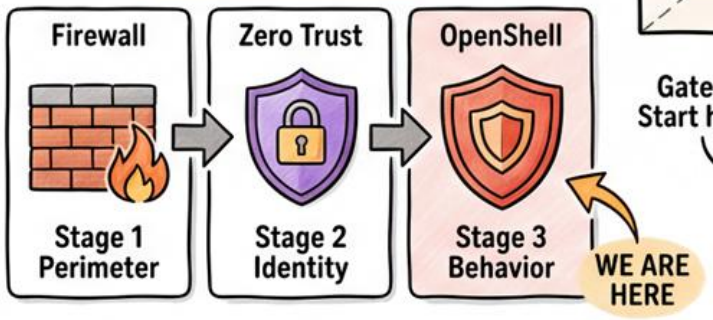


Container walls were never breached.
Attacks happened INSIDE.

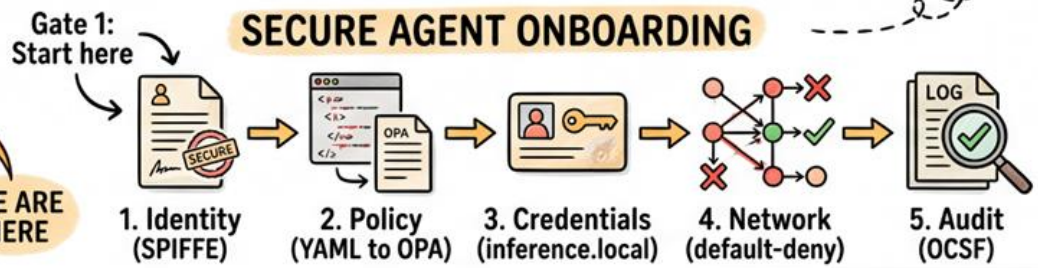
BEHAVIORAL SECURITY LAYER (OpenShell)



THE EVOLUTION



SECURE AGENT ONBOARDING



Stats corner

- **90%** over-permissioned
- **14.4%** security approval
- **88%** pilot failure
- **455K** malicious packages 2025

OWASP NIST

RED HAT AI: AGENTIC STRATEGY 2026

Your Agent. Our Platform. Production-Ready.

THE GAP

Frameworks keep changing

LangGraph · CrewAI · ADK · Strands · Custom

No framework ships production infrastructure

- Security
- Identity
- Observability
- Governance

Production needs don't

AGENTIC INFERENCE

- Agent Harness
- OpenShell Sandbox
- API Translation (OGX)
- Gateway + Guardrails
- Knowledge Retrieval (RAG, Vector Stores)
- Disaggregated Serving (llm-d)
- Model Server (vLLM)
- Hardware (GPU)

8 layers, all open, all debuggable

BYOA

(Bring Your Own Agent)

CUSTOMERS

Any API. Zero code changes.

THE ENTRYPOINT

THE DIFFERENTIATOR

THE PLATFORM STACK

- Customer chooses any framework — LangGraph · CrewAI · ADK · Strands · Custom
- Secure Agent Onboarding — Sandbox · SPIFFE/SPIRE · Identity
- Gateway (optional) — MaaS Gateway · MCP Gateway · Inference Gateway · Rate limiting, Auth, Routing, Quotas
- OGX (optional) — Responses API · Messages API (Planned) · Connectors · RAG · Multi-provider backend (Anthropic · Gemini · OpenAI x)
- llm-d (optional) — disaggregated routing, sits on top of vLLM
- vLLM (CORE) ★ — Chat Completions · Tool Calling
- AgentOps ☆ — Deep tracing · EvalHub · Cost tracking · Guardrails
- Infrastructure — OpenShift · GPUs · TPUs · RHEL AI

ONLY layer Red Hat doesn't own

Red Hat owns these

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)

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

Explore	Build	Operate
GenAI Studio Playground AI Hub	Starter Kits for 10 frameworks OGX APIs MCP Catalog	AgentOps EvalHub Cost tracking Guardrails

From idea to production

From Metal to Agents

Cloud · On-Prem · Disconnected · Air-Gapped

Red Hat

Agentic AI with Intel Xeon and vLLM CPU on RHOAI 3.5

Red Hat | intel.

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

Red Hat Summit

Connect

Thank you

linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

Red Hat | intel