Red Hat Summit

Connect

Agentic AI in Action

Red Hat & Intel Shaping the Future of Enterprise AI

intel | Red Hat

Albertano Caruso

Sales Application Engineer
Intel

Natale Vinto

Evangelism Director
Red Hat

Agenda

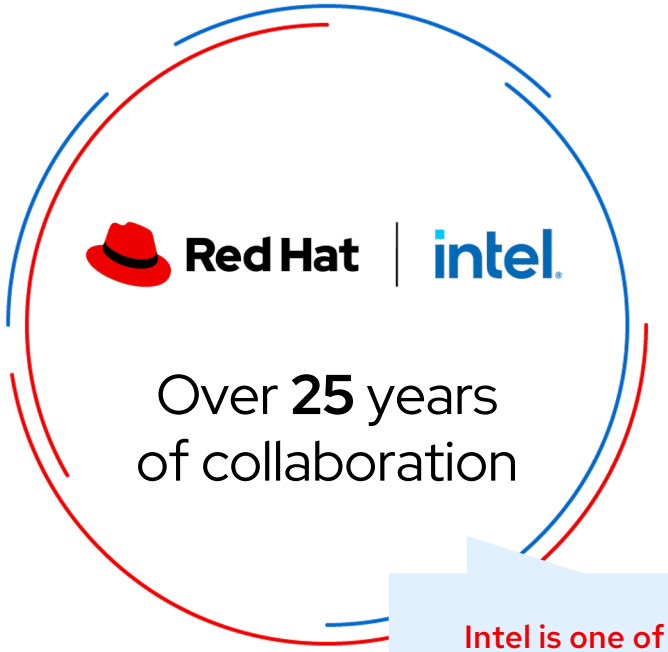
- Intel – Red Hat partnership
- Intel AI Strategy
- Red Hat AI Strategy
- Intel AI Accelerators
- Intel Confidential Computing
- Intel AI Software
- Red Hat AI Platform
- Intro to Agentic AI
- Agentic AI Demo

Intel - RH Partnership

Open source software: Intel is committed

Intel® has a long history with Linux®, actively participating in open source development and collaboration with the Linux community, to ensure hardware is well-supported and delivers optimal performance on Linux-based systems.

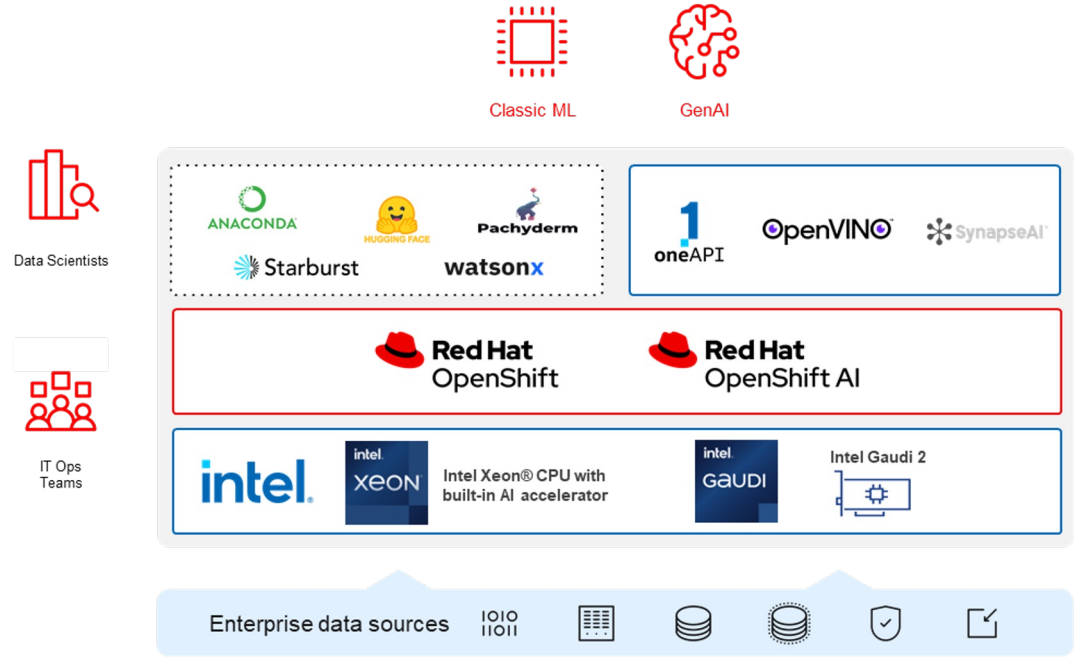
Intel contributes to more than 100 different open source projects, from the Linux kernel to cloud orchestration and plugins for Kubernetes.



Intel is one of the key contributors
Linux Kernel Corporate Contributor since 2007[1]

Source:
[1] Open source software: Intel engineers are engaged. Intel Innovation Digital 2021.

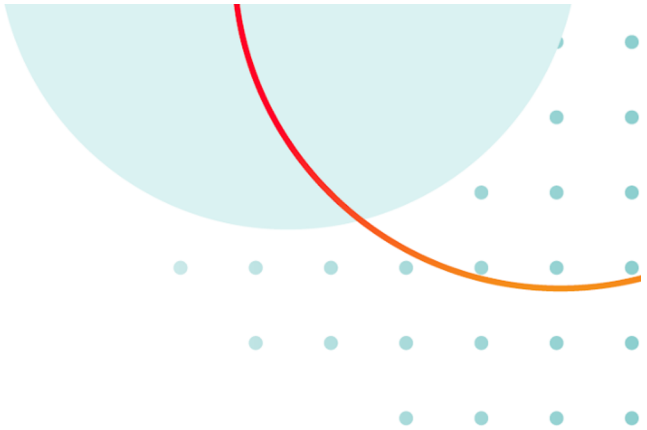
Real Customer Example: AI Sweden



- Collaborating to deliver AI solutions
- Deeper, product collaboration focused on customer enablement with OpenShift AI, Intel Xeon, Gaudi 2 and the Intel AI Suite
- Testing, validation, and proof of concepts
- Receive support for building AI applications

Red Hat | intel

Intel's AI Strategy and Capabilities

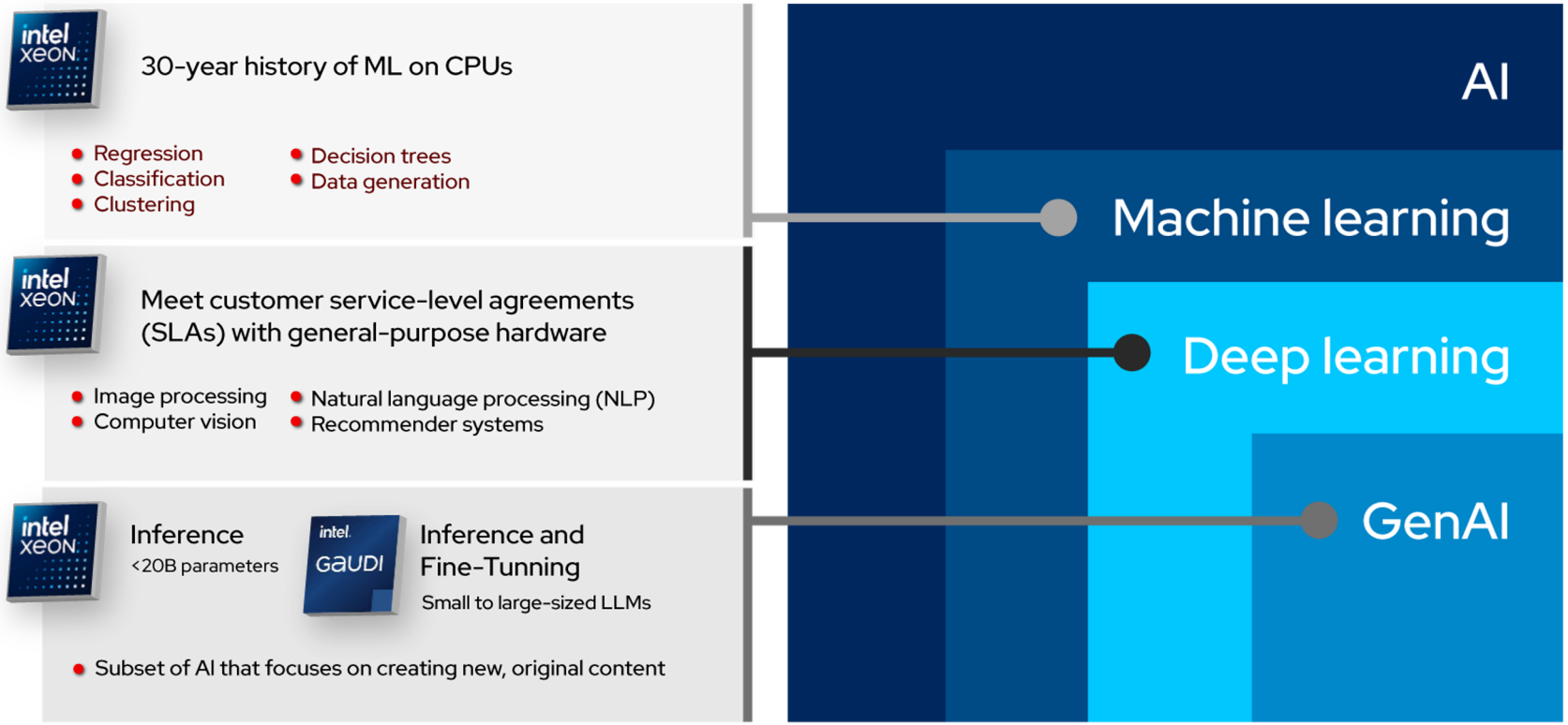


Intel's AI Strategy

Open	Less cost, No lock in
Innovation	AIPC to Edge to Datacenter & Cloud
Efficient	Performance per $ & per W leadership
Secure	Data as your IP & Models as your IP

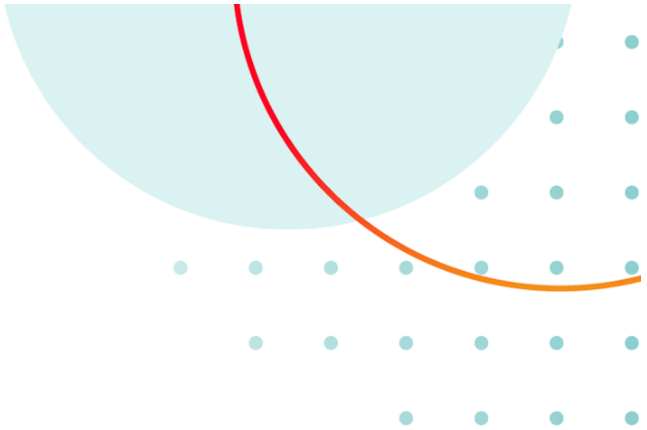
The AI Hierarchy: Mapping ML, DL, and GenAI with Intel

Discover how Intel® processors fuel AI workloads across inference, training, and next-generation GenAI applications



Red Hat | intel

Red Hat's AI Strategy and Capabilities



Red Hat AI

Accelerate the development and delivery of AI solutions across hybrid-cloud environments

Increase efficiency with **fast, flexible and efficient inferencing**

Simplified and consistent experience for **connecting models to data**

Flexibility and consistency when **scaling AI across the hybrid cloud**

Accelerate Agentic AI delivery and stay at the forefront of innovation

Red Hat AI

Red Hat AI Inference Server

Red Hat Enterprise Linux AI

Red Hat OpenShift AI

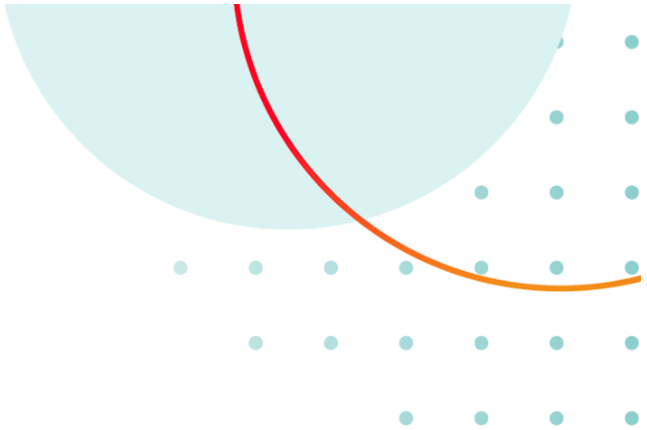
Trusted, Consistent and Comprehensive foundation

NVIDIA AMD intel Hardware Acceleration Google aws IBM

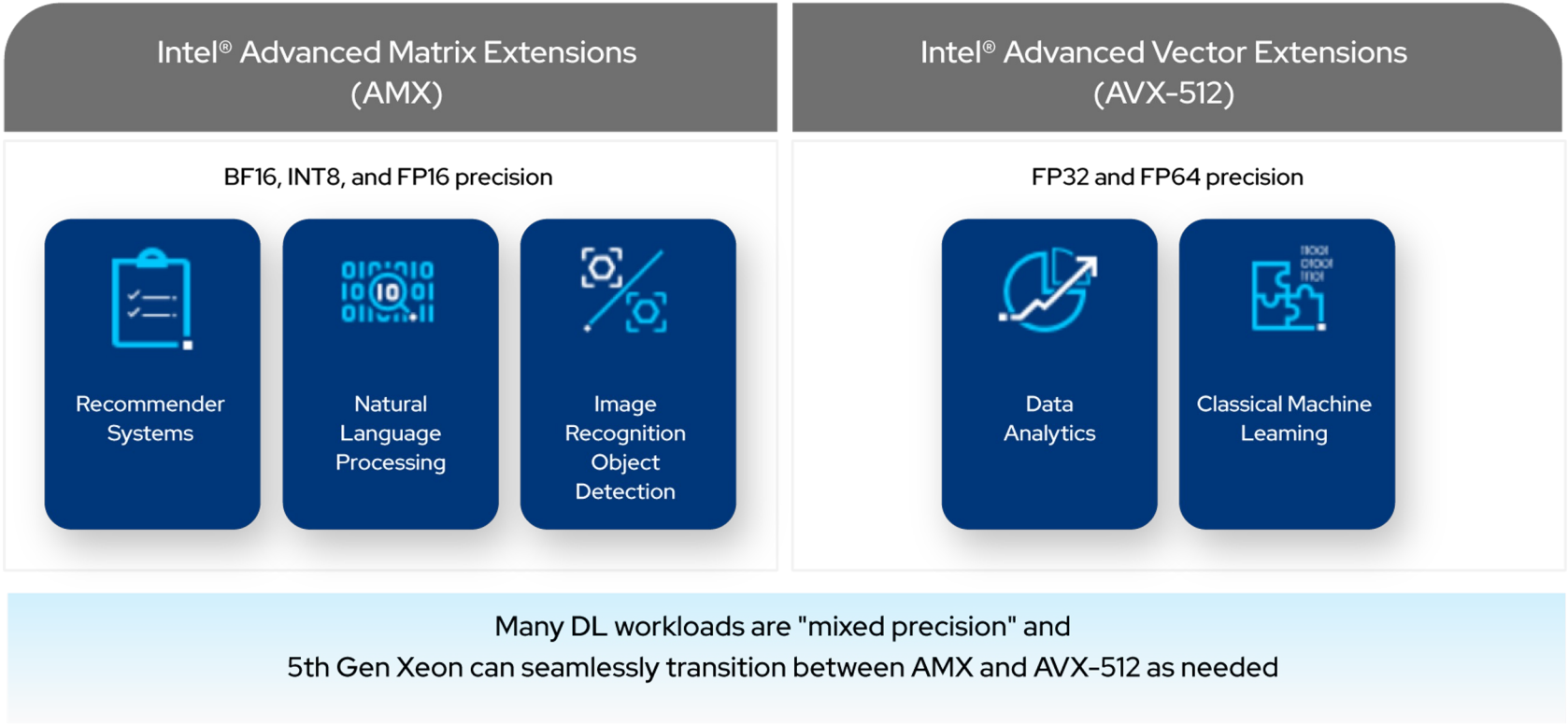
* NVIDIA, AMD, Intel, Google TPU supported in Red Hat AI. AWS Inferentia/Neuron IBM AIU are on our roadmap

Red Hat

Intel AI Accelerators



Intel® AMX Accelerates **DEEP LEARNING** Use Cases



Resolve Customer Queries Faster with More Concurrent Users in Your LLMs and Agents

- Get superior performance for batch, real-time inference, and training for small and medium language models with Intel® Xeon® processors.
- Use your CPU for cost-effective model updates.

Large language models (LLMs)

Intel Xeon 6 vs. AMD EPYC Turin

Llama2-7B

Up to

1.38x

higher throughput

with Intel Xeon 6980P
vs. AMD EPYC 9965'

Intel Xeon 6 vs. 5th Gen Intel Xeon

GPTJ-6B

Up to

2x

Higher performance

Intel Xeon 6980P
vs. Intel Xeon 8592+2

Llama-13B

Up to

2x

Higher performance

Intel Xeon 6980P
vs. Intel Xeon 8592+2

Llama2-7B

Up to

2.3x

Higher training performance

Intel Xeon 6980P
vs. Intel Xeon 8592+3'

5th Gen Intel Xeon vs. 3rd Gen Intel Xeon

Llama2-13B

Up to

2.1x

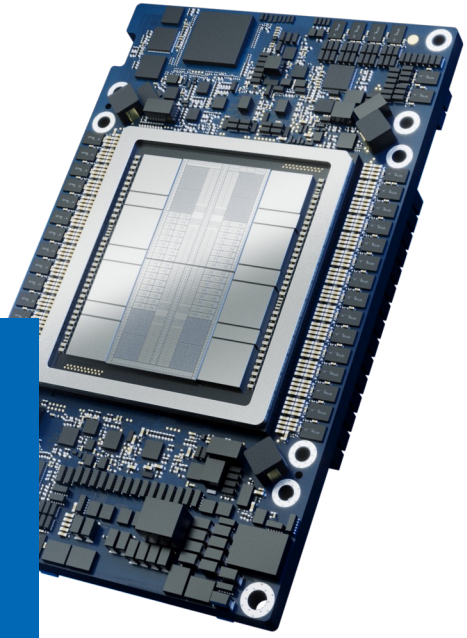
real-time inference performance speedup

5th Gen Intel Xeon vs.
3rd Gen Intel Xeon4

Red Hat | intel.

Intel® Gaudi® 3 AI Accelerator: AI Inferencing

Price Performance Advantage



Up to
43%

Higher throughput
(tokens per second)

on IBM Granite-3.1-8B-Instruct
vs. leading GPU competitor
with small context sizes

Up to
120%

More cost efficient
(tokens per dollar)

on Mixtral-8x7B-Instruct-v0.1
vs. leading GPU competitor
with long input and short output sizes

Up to
92%

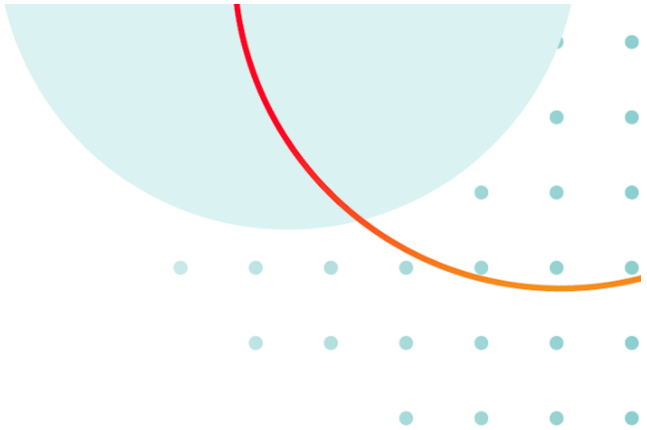
More cost efficient
(tokens per dollar)

on Llama-3.1-405B-Instruct-FP8
vs. leading GPU competitor
with large context sizes

Source: Signal65 Lab Insight Whitepaper - Intel Gaudi 3 AI Accelerator at Scale on IBM Cloud, Intel-commissioned study by Signal65, published April 22, 2025. Reported numbers are inferencing results on Intel Gaudi 3 vs. NVIDIA H200. See source for workloads and configurations. Results may vary.

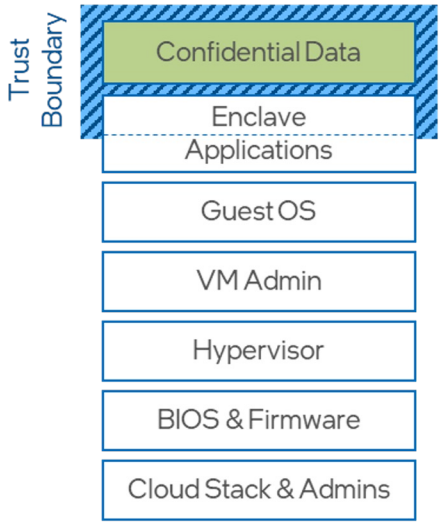
Red Hat | intel

Intel Confidential Computing



App Isolation

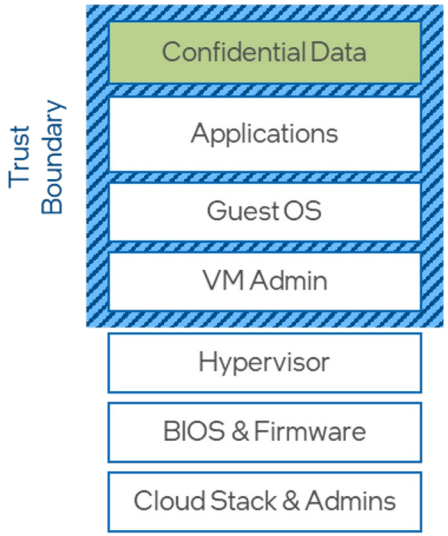
Intel® SGX



Smallest trust boundary for greatest data protection & code integrity

VM Isolation

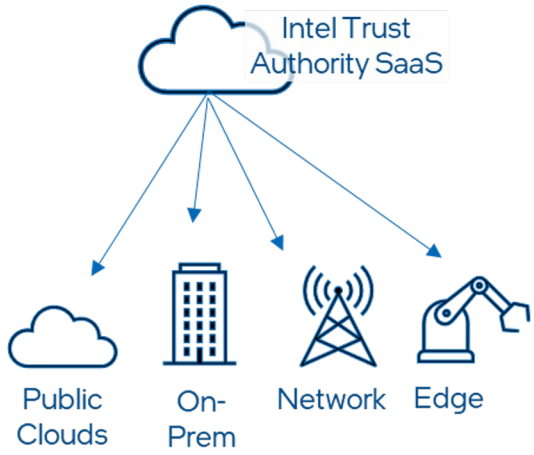
Intel® TDX



Most straightforward path to greater security for legacy apps

Trust Services

Intel® Tiber™ Trust Authority



Uniform, independent attestation of trustworthy environments

Founded on Intel's Security-First Development & Lifecycle Support

Red Hat | intel.

Confidential AI Helps Protect Data & Models In-Use

Utilizing Confidential Computing for Containers with Intel TDX

Hardware-Based Protection of Data In-Use
With Intel Trusted Domain Extensions (TDX)

+

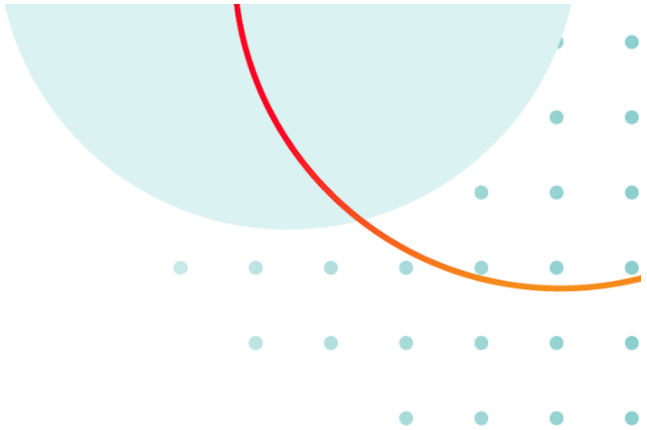
OpenShift / OpenShift AI sandboxed containers

The Confidential Containers project

Confidential Computing is about protecting data in-use.
You do not have to trust the system admins of the providers any longer.

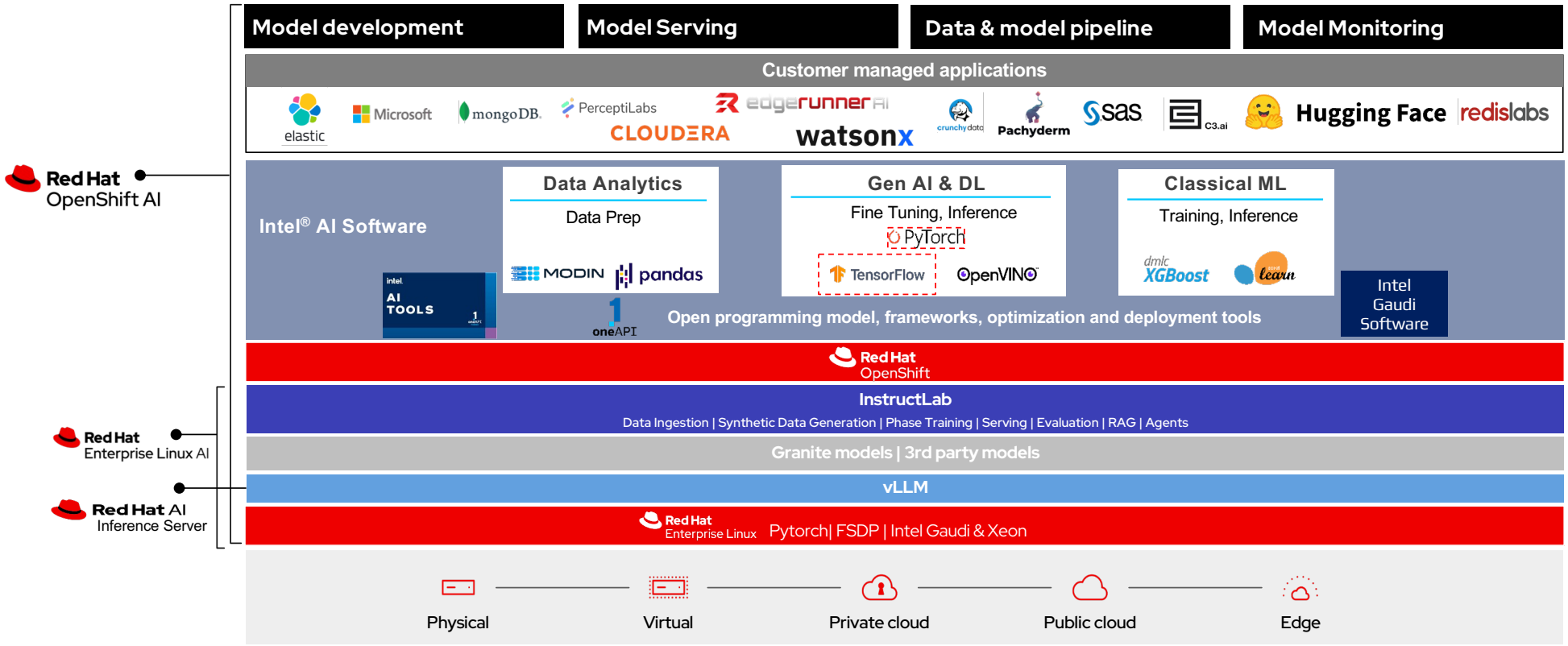
Red Hat | intel.

Intel AI Software

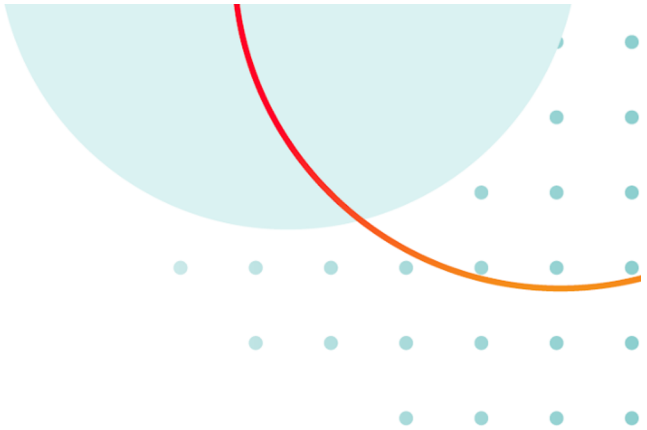


Red Hat AI with Intel AI platform

Generative AI and MLOps capabilities for building flexible, trusted AI solutions at scale

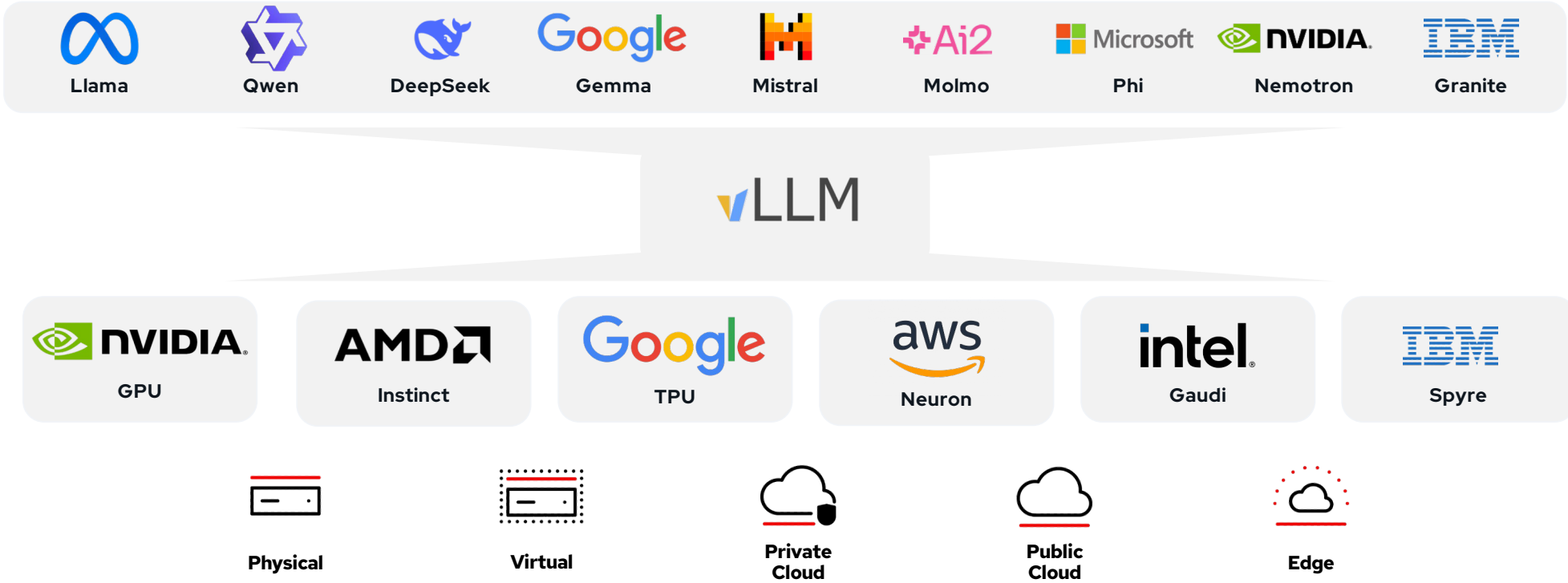


Red Hat AI Platform



Red Hat AI the inference engine for the hybrid cloud

vLLM supports the key models on the key hardware accelerators



Red Hat AI repository on Hugging Face

A collection of third-party validated and optimized large language models

Broad Collection of models

Llama
Qwen
Gemma
Mistral
DeepSeek
Phi
Molmo
Granite
Nemotron

Validated models

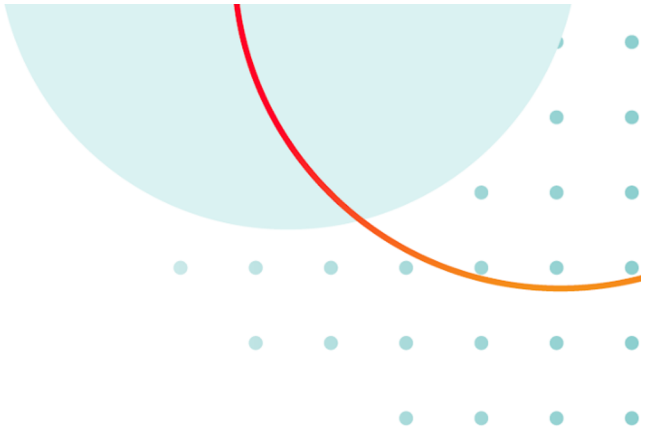
- Tested using realistic scenarios
- Assessed for performance across a range of hardware
- Done using GuideLLM benchmarking and LM Eval Harness

Optimized models

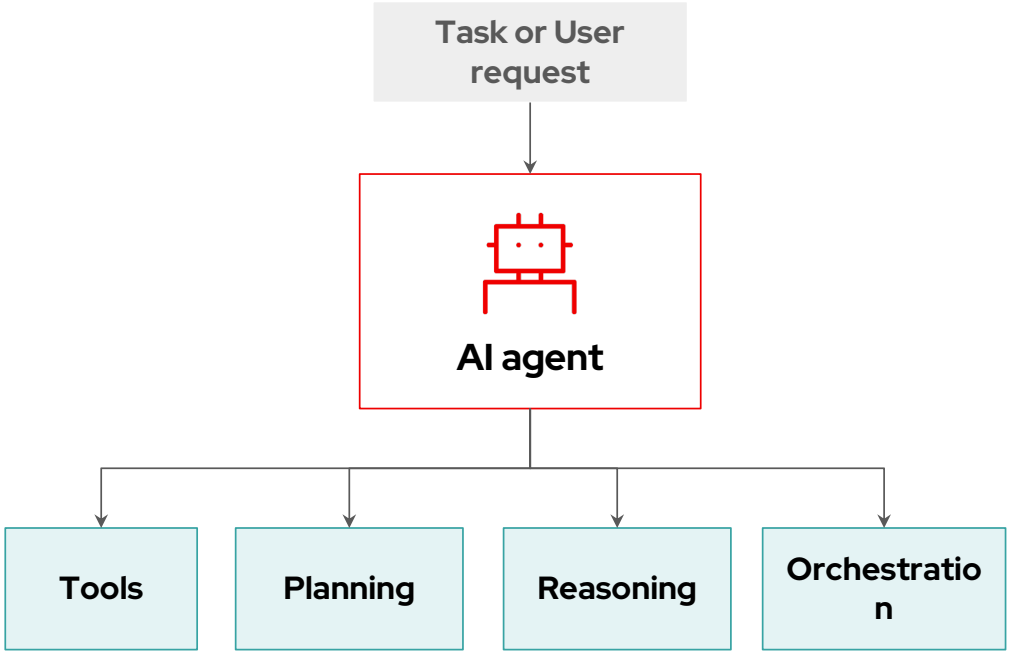
- Compressed for speed and efficiency
- Designed to run faster, use fewer resources, maintain accuracy
- Done using LLM Compressor with latest algorithms

Red Hat

Intro to Agentic AI

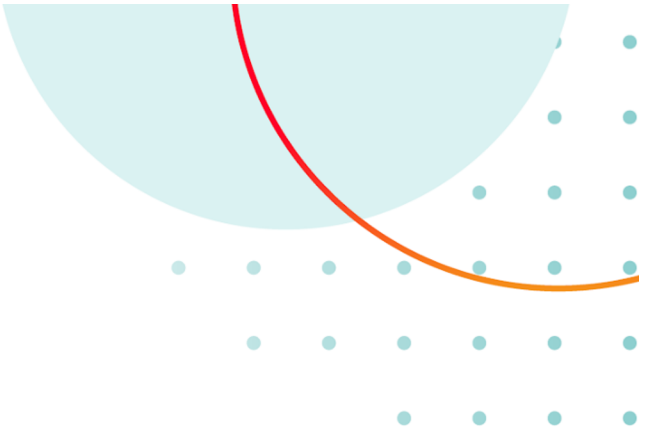


The components of an AI Agent system

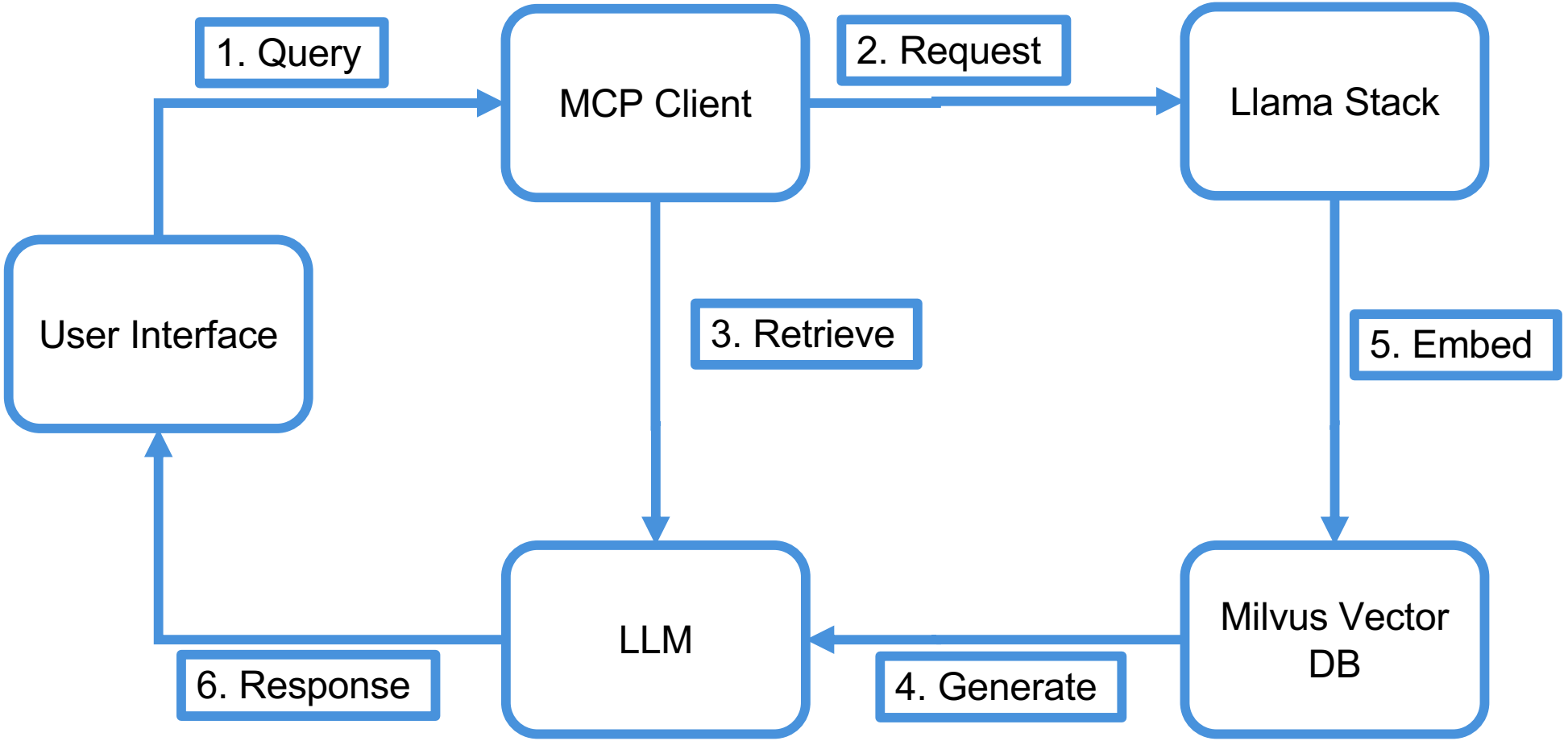


- **Tool Utilization:** Leverages external tools to gather data and perform tasks.
- **Planning and Execution:** Develops and executes multistep plans to achieve goals autonomously.
- **Reasoning:** Applies logic and contextual understanding to make informed decisions.
- **Orchestration:** Coordinates actions, tools, and agents to dynamically adjust and complete tasks.
- **Communication protocols:** enables the connections between the components.

Agentic AI Demo



Agentic AI Demo Architecture



Q & A

Red Hat Summit

Connect

Grazie


linkedin.com/company/red-hat


facebook.com/redhatinc


youtube.com/user/RedHatVideos


twitter.com/RedHat