Red Hat Summit

Connect

Agentic AI in Action

Red Hat & Intel Shaping the Future of Enterprise AI

London

9 October

2025

Red Hat | intel

David Hellewell

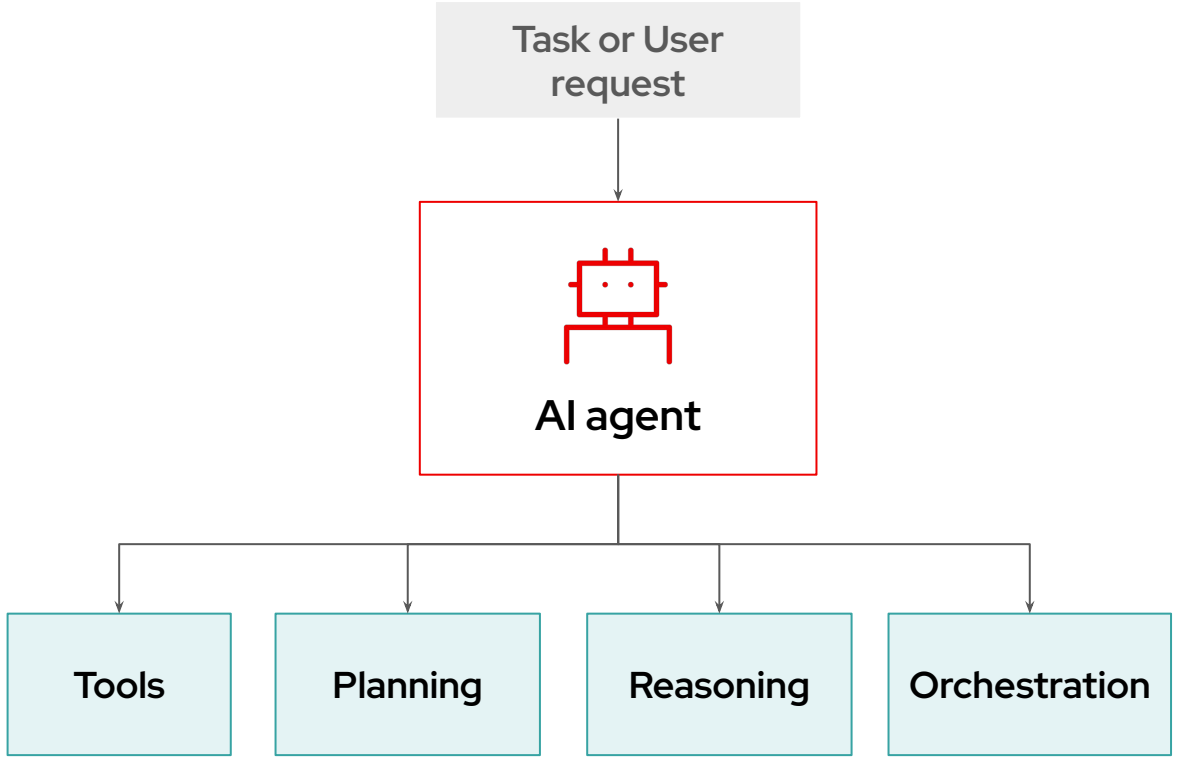
FSI Technical Leader
Intel

Jamie Hackett

AI SSA, EMEA
Red Hat

Intro to Agentic AI

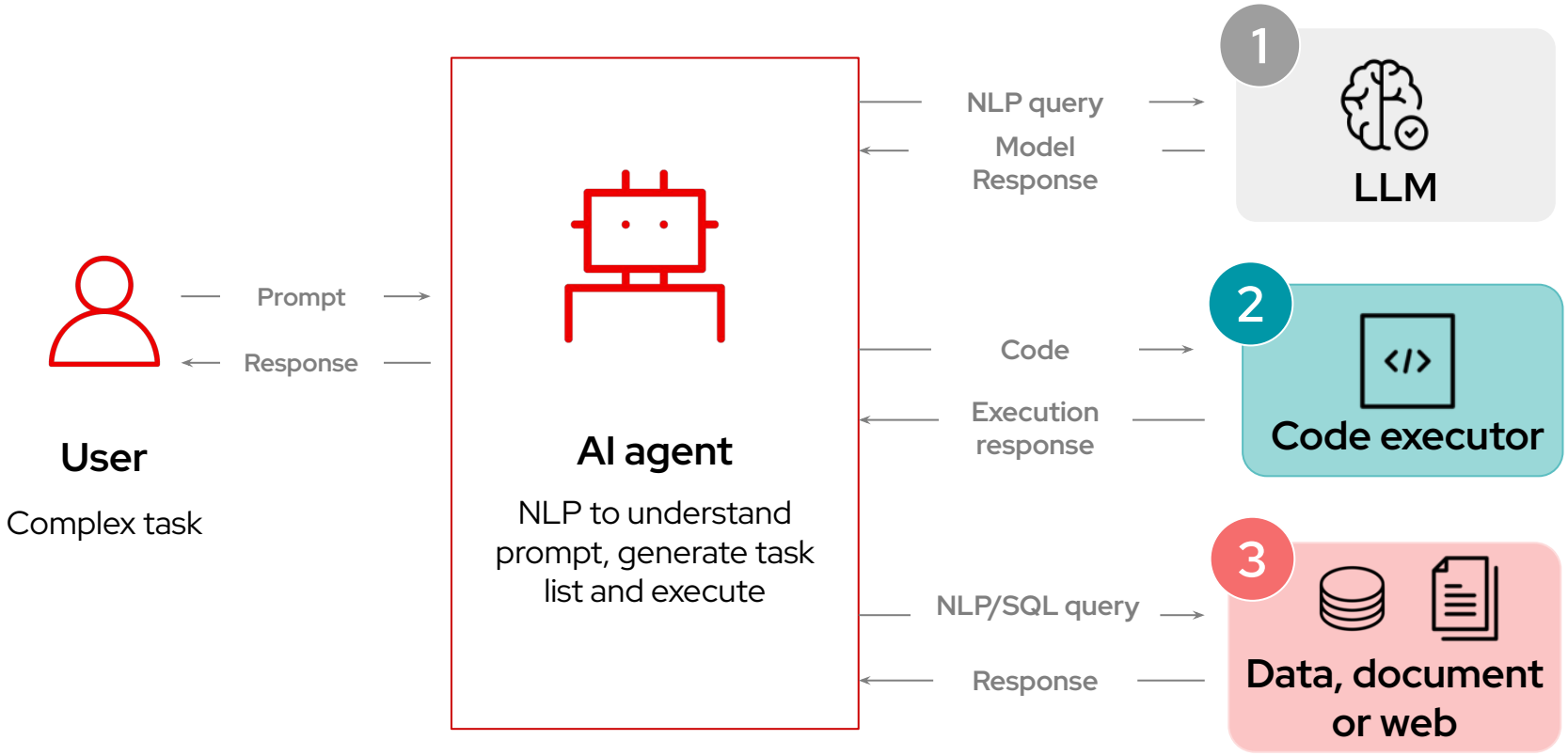
The components of an AI Agent system



- **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.

AI agents integrate models, functions & tools

Gen AI Models, Predictive AI Models, Code Functions, Search & more



Red Hat

Intel's AI Strategy and Capabilities

Red Hat | intel.

Intel® Xeon® 6 Processors for AI

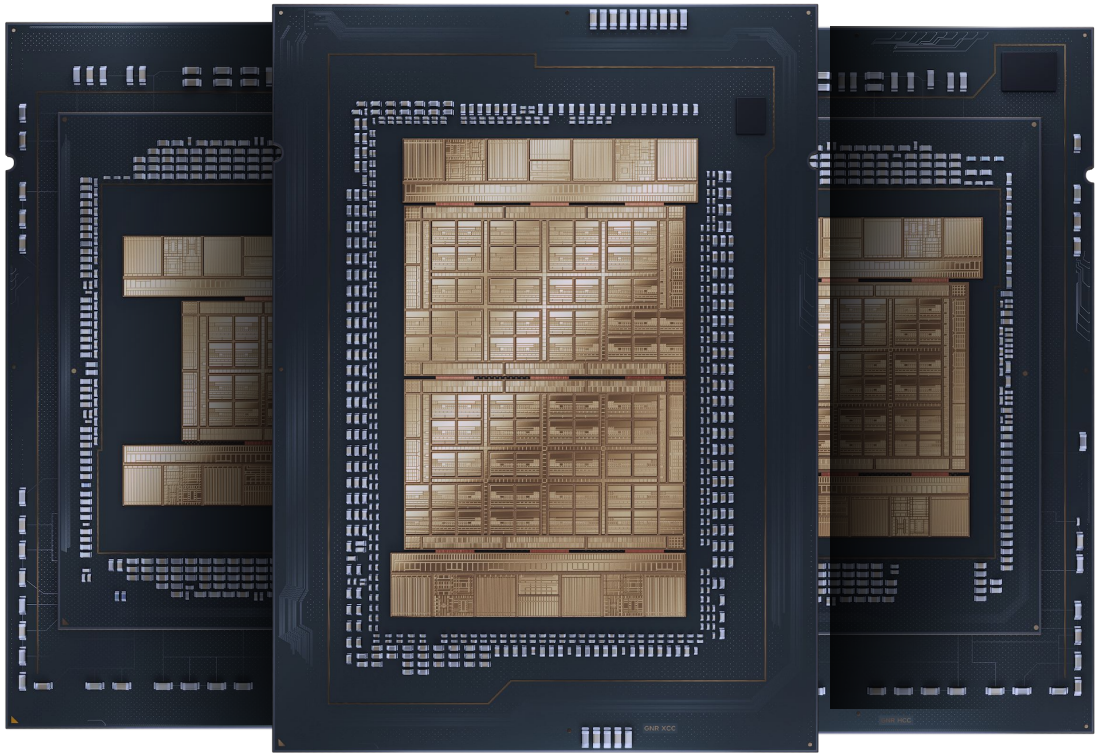
World's Best CPU for AI

Up to 128 P-cores
on 6900-series
up to 86 P-cores on 6500/6700-series

More bandwidth & cache
MRDIMM memory support
Up to 504MB low latency LLC

AI accelerators built-in
Intel® AMX, Intel® AVX-512,
and Intel® AVX-2

Comprehensive SW suite
AI development across classical ML
and small GenAI models



The Most Deployed Host CPU

Superior I/O performance
up to 192 PCIe 5.0 lanes

High Single Threaded Performance
With Intel's latest generation P-core

Top Tier Memory Support
30% higher memory B/W with MRDIMMs
Expandability with CXL 2.0

Ready for Deployment
DC-MHS & NVIDIA MGX™
form factors supported

Intel AMX Accelerates DEEP LEARNING Use Cases

Intel® Advanced Matrix Extensions (AMX)

BF16, INT8, and FP16 precision

- Recommender Systems
- Natural Language Processing
- Image Recognition Object Detection

Intel® Advanced Vector Extensions (AVX-512)

FP32 and FP64 precision

- Data Analytics
- Classical Machine Learning

Many DL workloads are “mixed precision” and
5th Gen Xeon can seamlessly transition between AMX and AVX-512 as needed

Intel Guadi 3 Accelerator for AI Inference

Delivering 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)

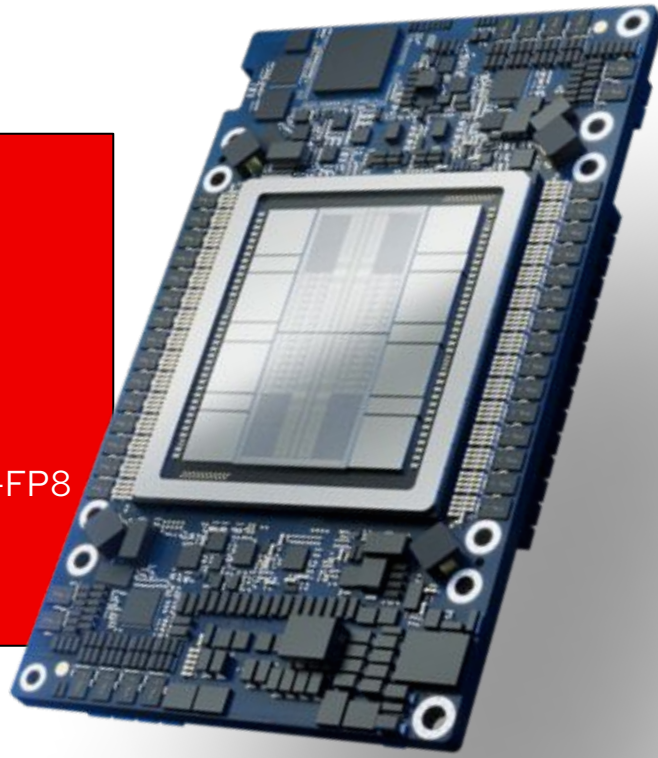
on Mistral-8x7B-Instruct-v01
Vs. leading GPU competitor

With long input and short output sizes

Up to
92%
More cost efficient
(tokens per dollar)

on Llama-3.1-405B-Instruct-FP8
vs. NVIDIA H200

with large context sizes

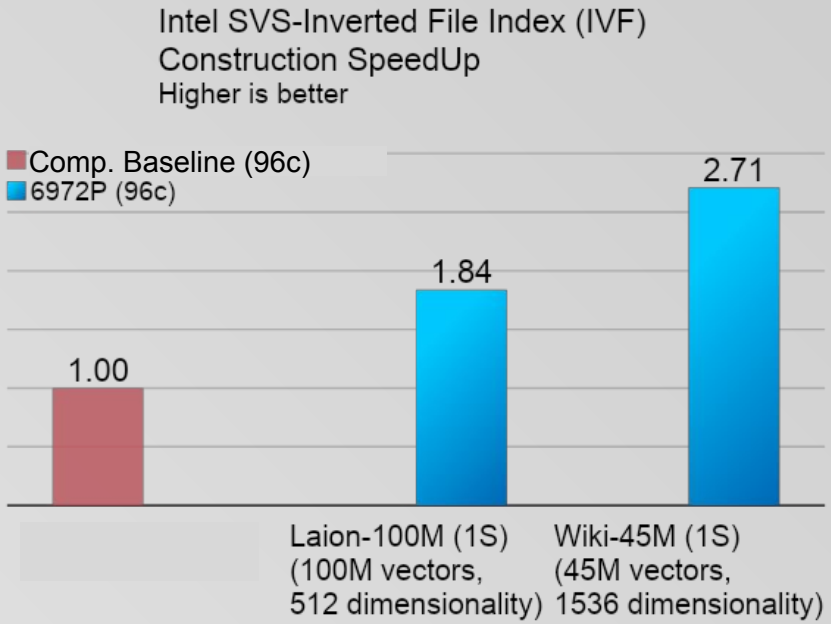


Source: Signal65 Lab Insights Report - Intel Gaudi 3 Accelerates AI at Scale on IBM Cloud, Intel-commissioned study by Signal65, published April 2025. See Signal65 report source for workloads and configurations. Results may vary.

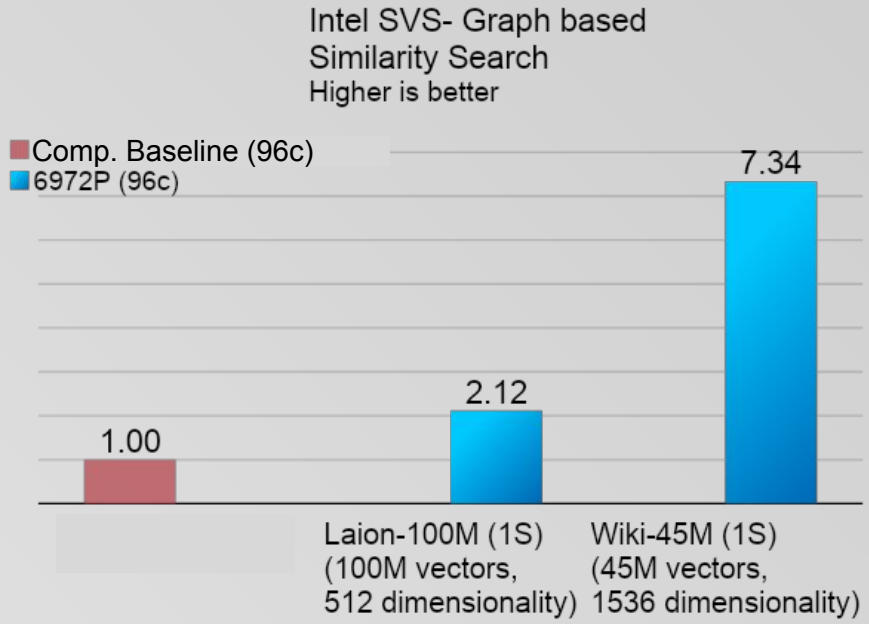
Intel® SVS Vector Optimizations for Vector Databases

Intel® Scalable Vector Search (SVS)

Improve Vector database indexing by leveraging Intel AMX instructions



Improve Vector database search with Intel vector optimizations using Intel SVS



See backup for workload and configurations. Results may vary.

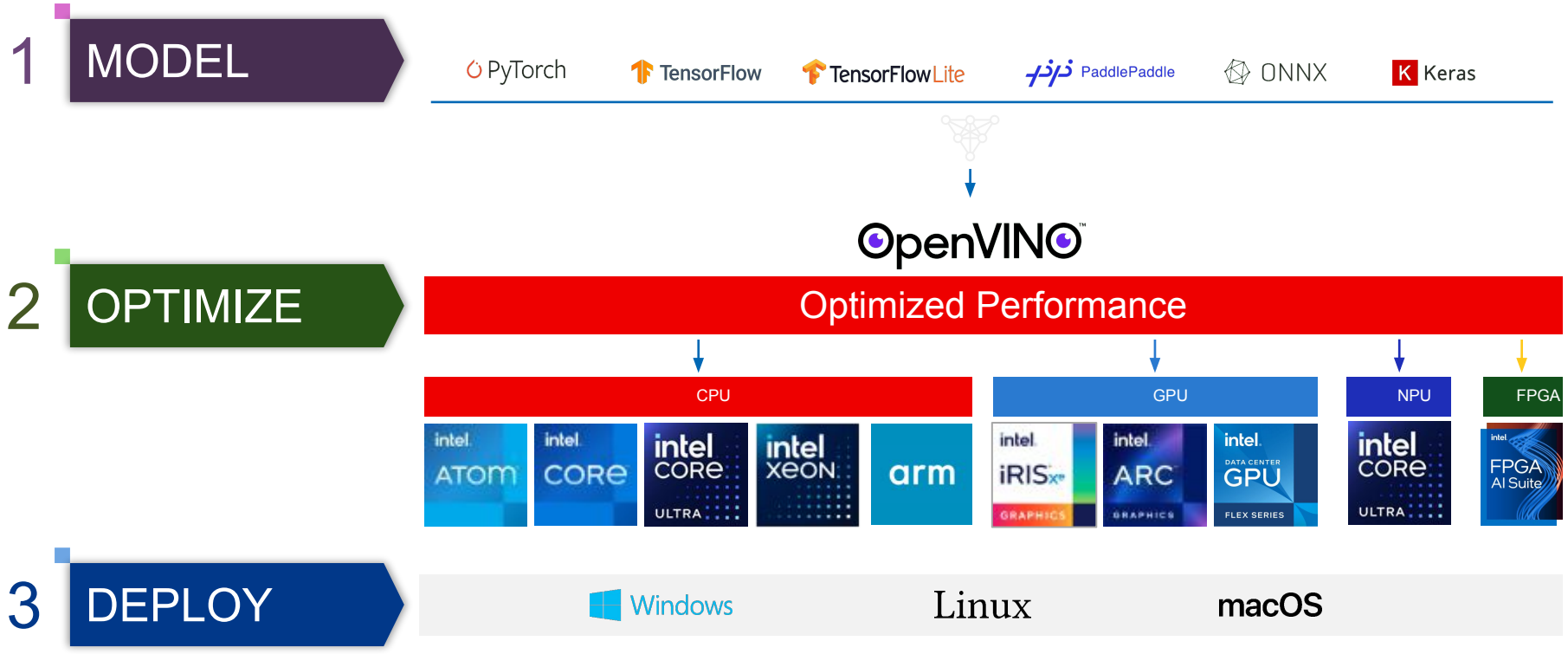
Intel Openvino Toolkit

Fast, Accurate Results with High-Performance

Convert and optimize models, and deploy across a mix of hardware

Hundreds of models supported across GenAI, Computer Vision, multi-modal

Use native APIs or employ
- Triton Server
- LangChain
- Torch.compile
- Hugging Face Optimum Intel
- ONNX Runtime with OpenVINO backend



Powered by oneAPI

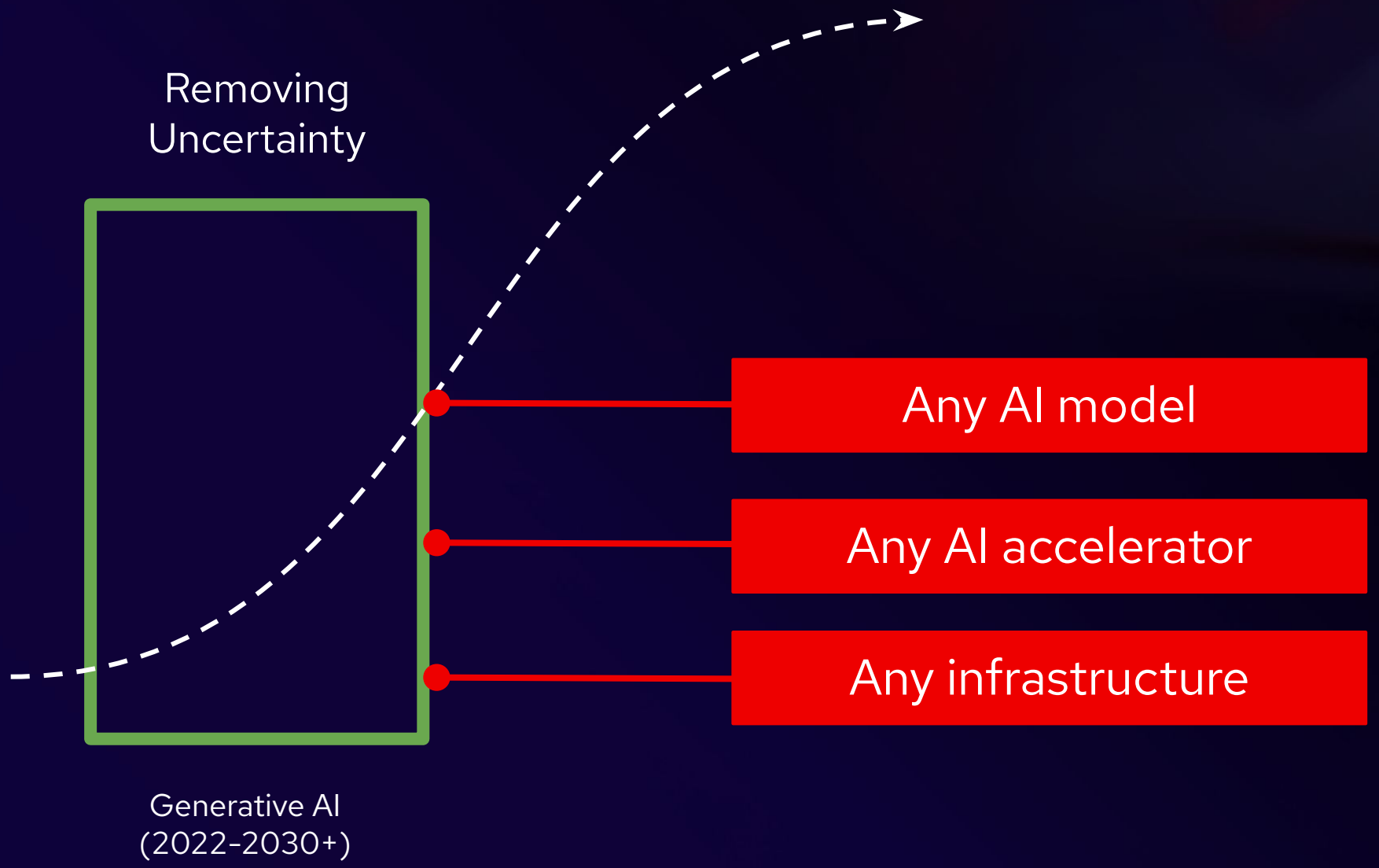
The productive, smart path to freedom for accelerated computing from the economic and technical burdens of proprietary alternatives.

Red Hat | intel.

Red Hat's AI Strategy and Capabilities

Red Hat | intel

Red Hat AI - Enabling AI Success



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

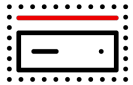
Red Hat
Enterprise Linux AI

Red Hat
OpenShift AI

Trusted, Consistent and Comprehensive foundation

NVIDIA AMD intel Hardware Acceleration Google aws IBM

Physical



Virtual

Private Cloud

Public Cloud

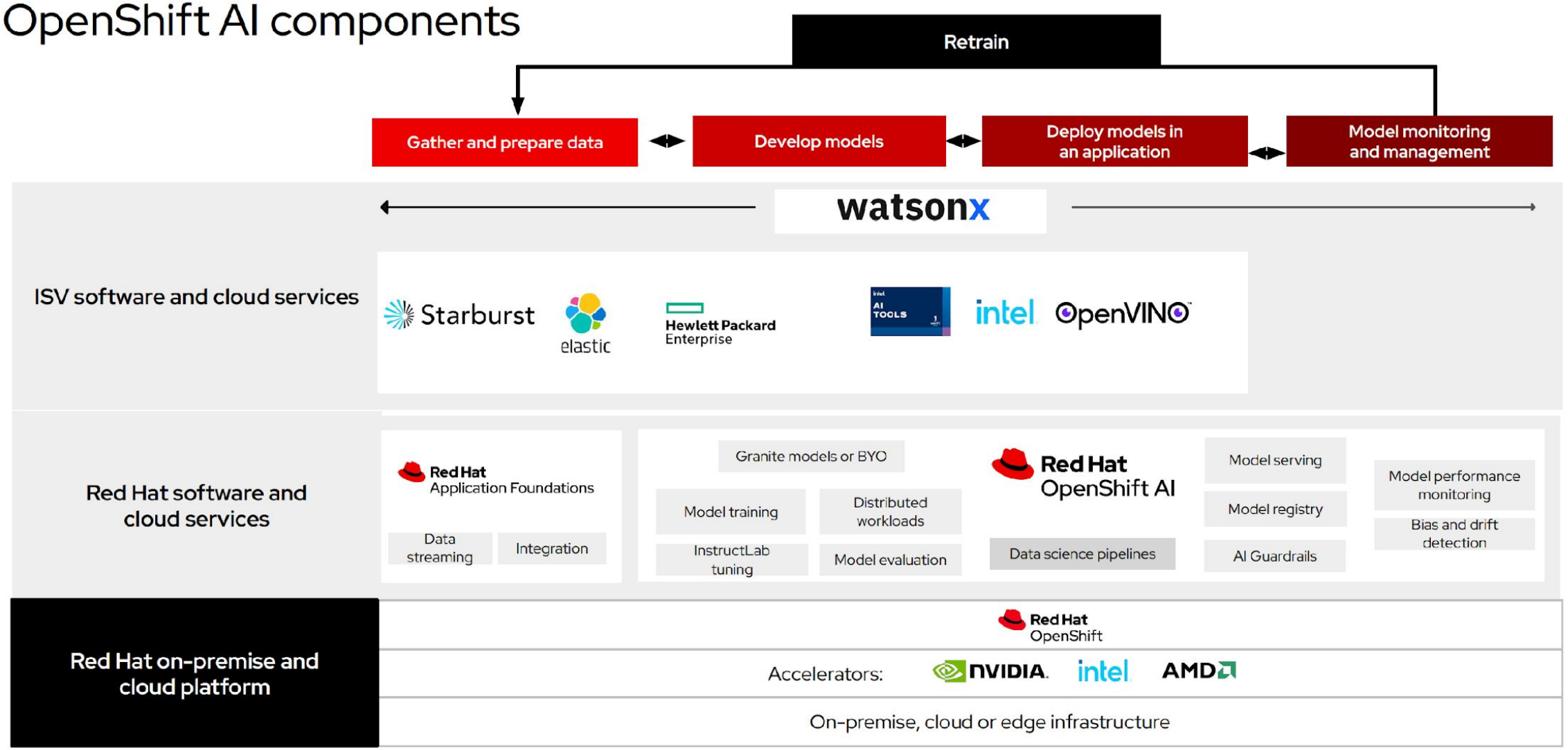
Edge

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

Red Hat AI Platform

Red Hat | intel

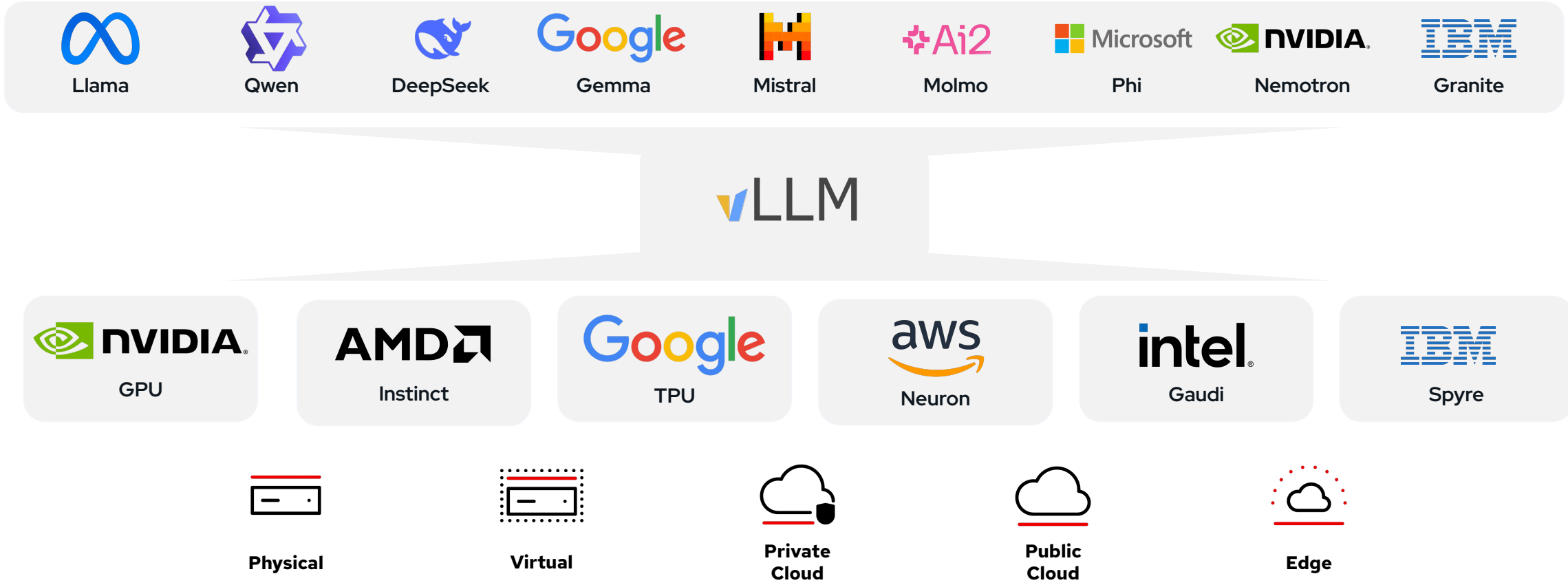
OpenShift AI components



Red Hat

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

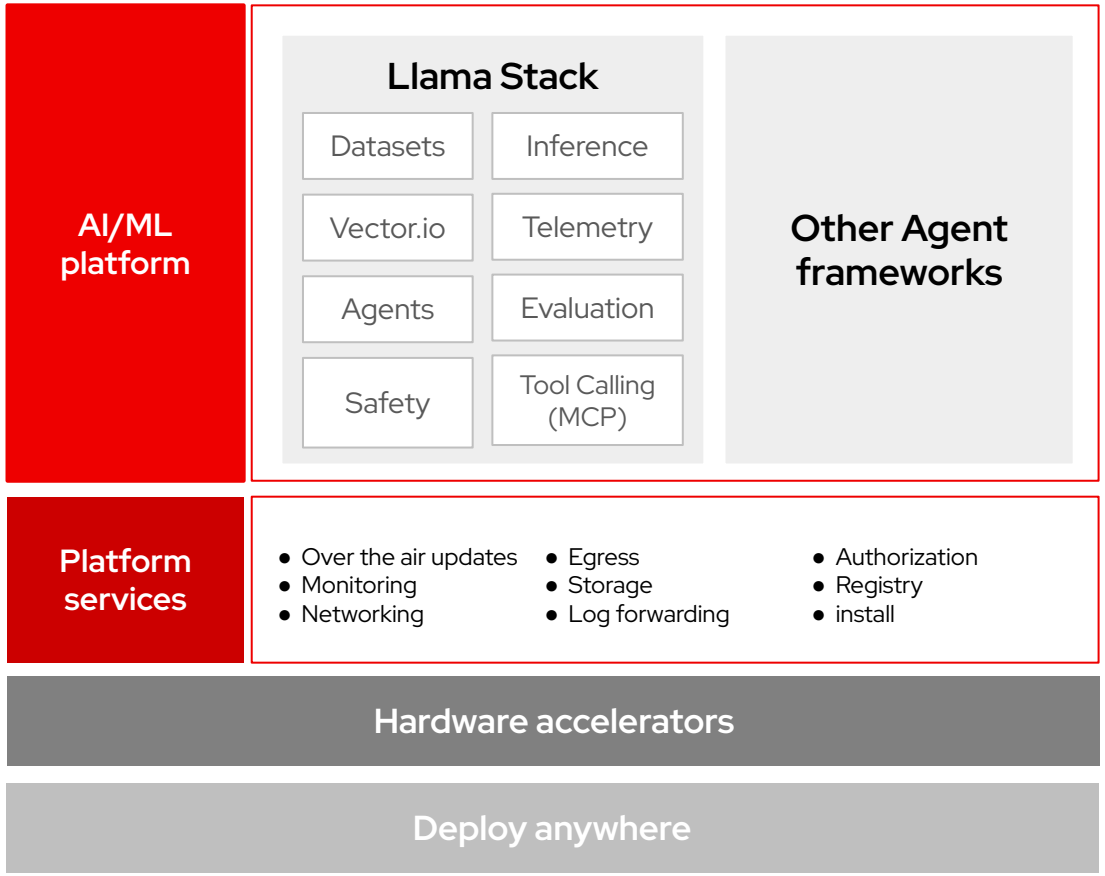
Validated models

- 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

A modular approach to building AI agents



Red Hat AI allows to:

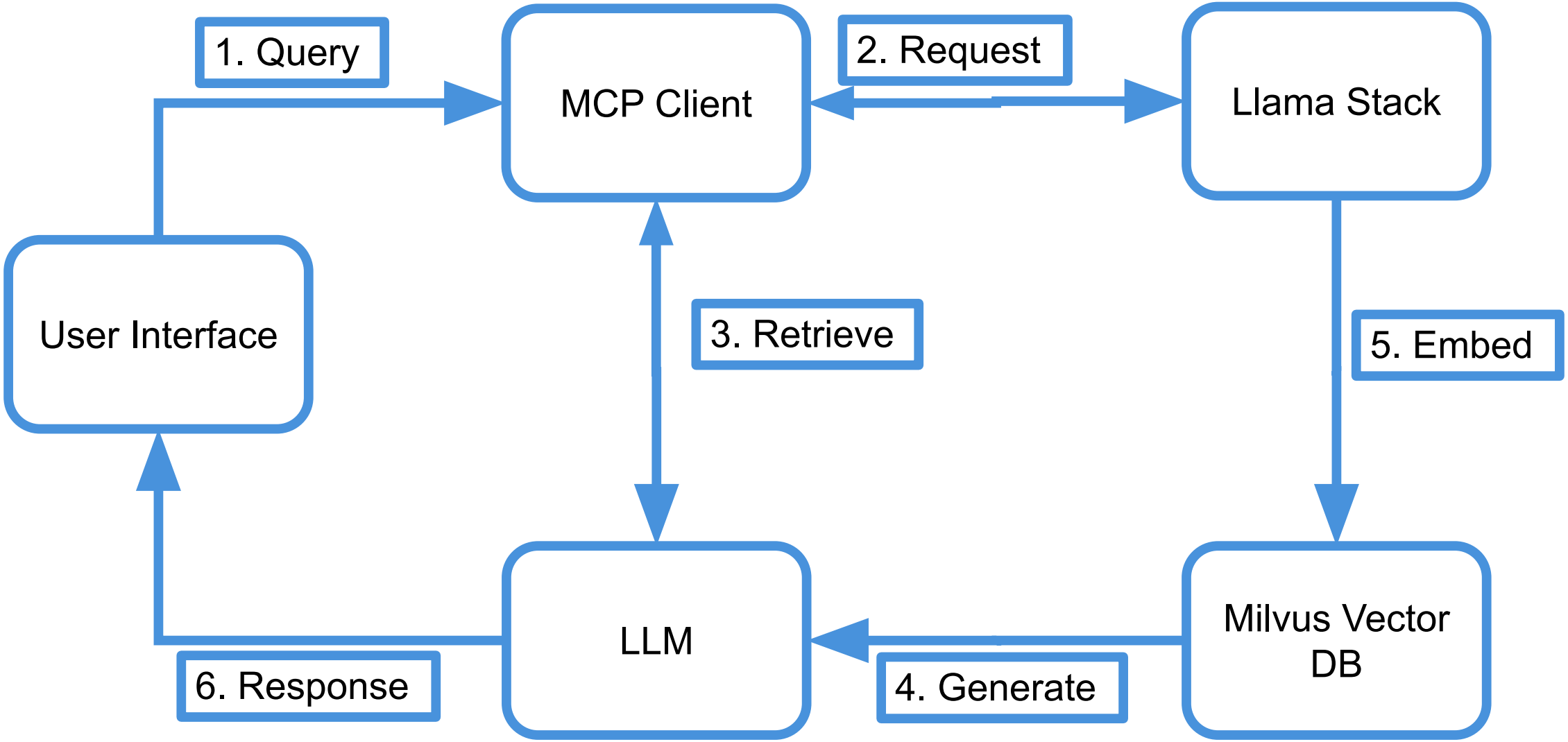
- Build agents using **Llama Stack's native capabilities and implementations**.
- **Bring compatible Llama Stack implementations** to OpenShift AI.
- **Use your own agent framework** and selectively incorporate Llama Stack APIs.
- **Build with Core Primitives** and manage your own agent framework as a standard workloads.

Red Hat

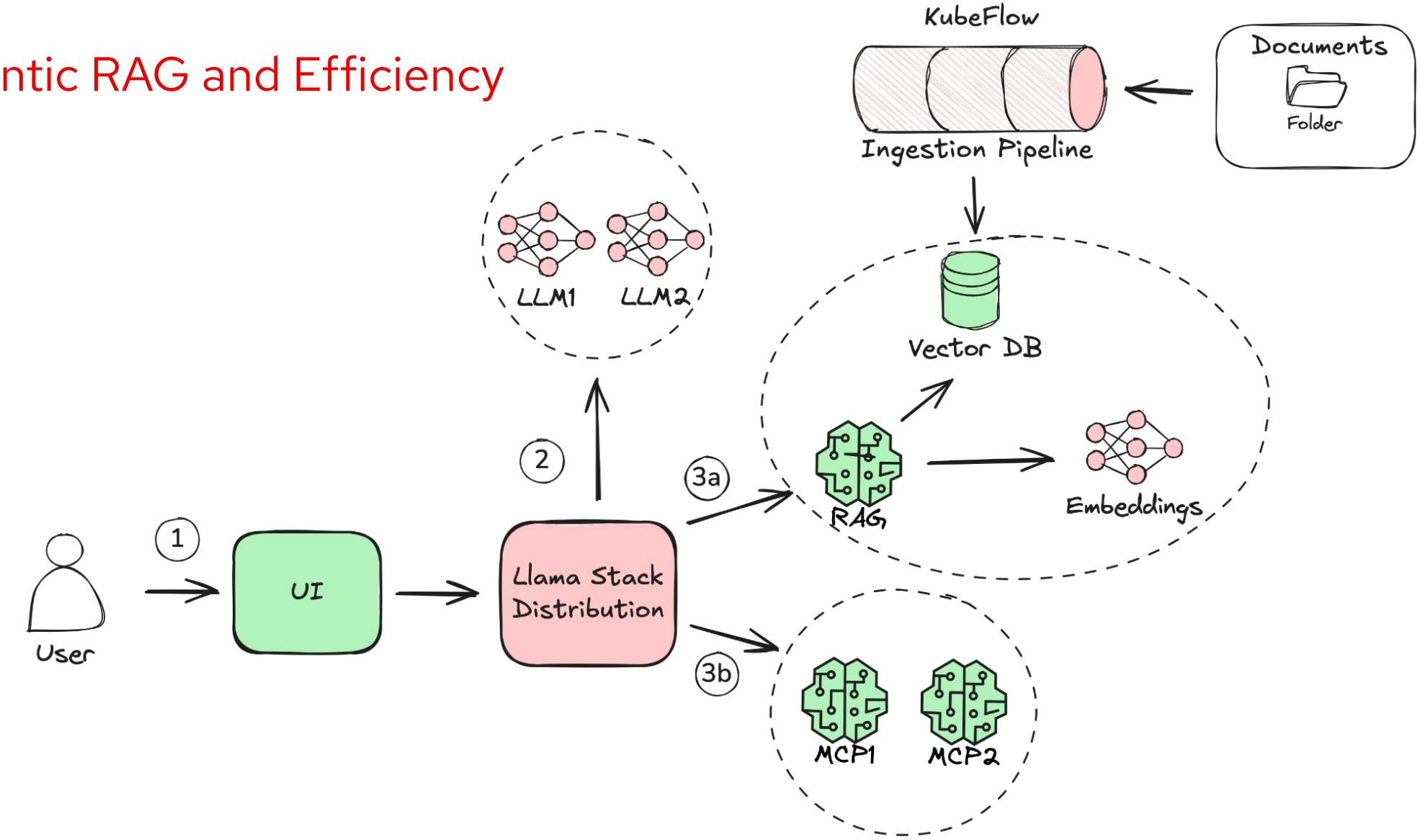
Agentic AI Demo

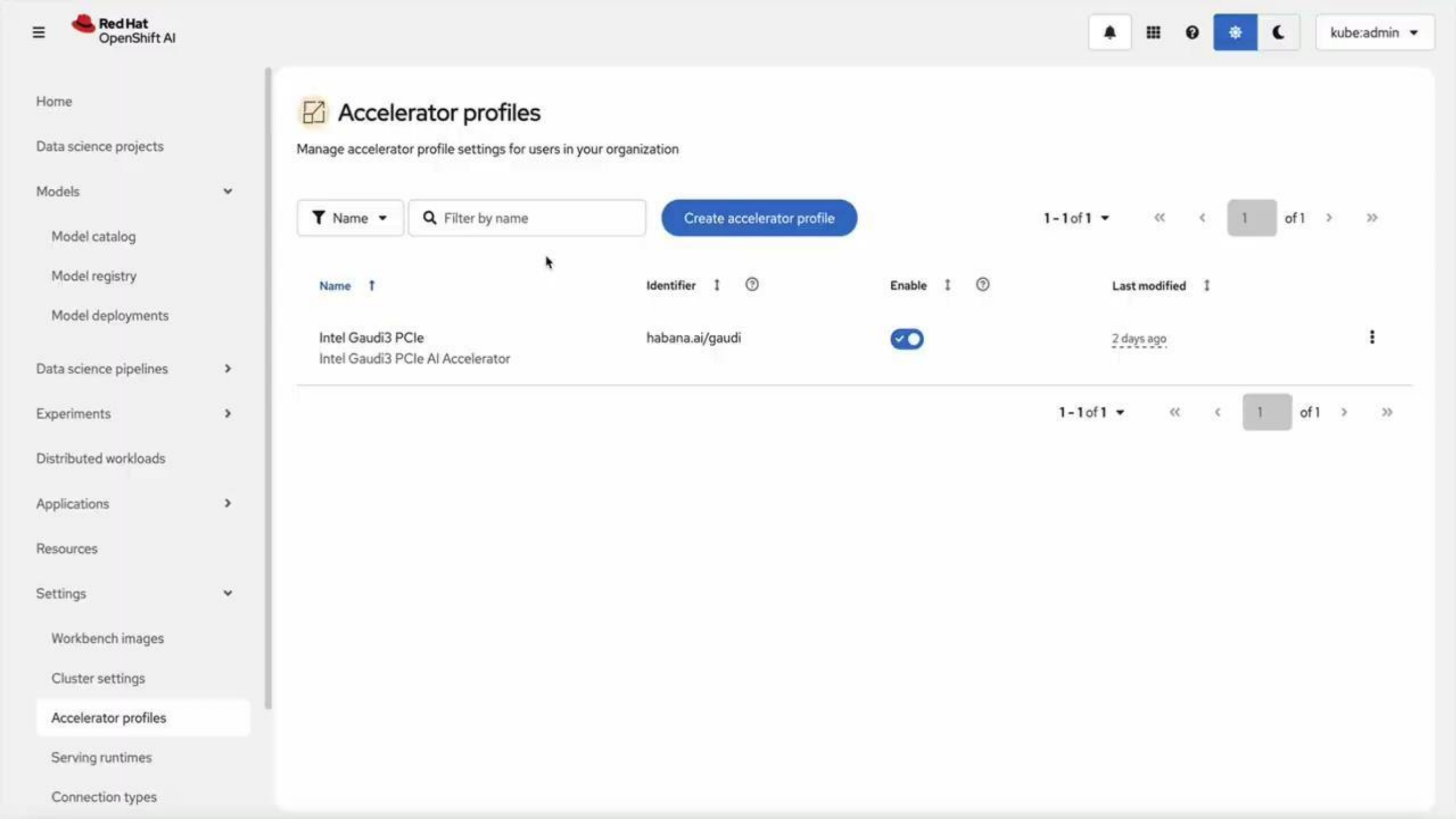
Red Hat | intel.

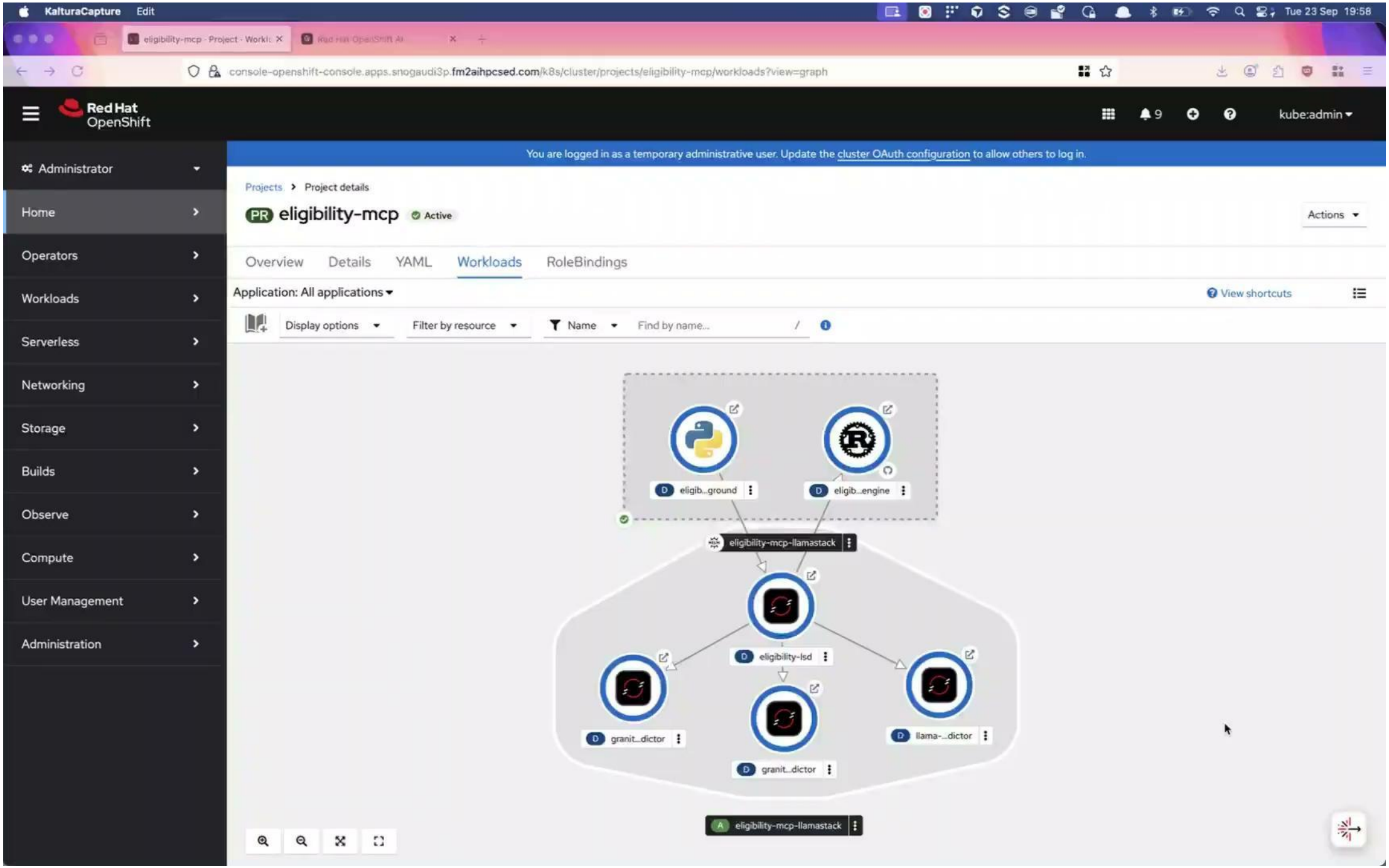
Agentic AI Demo Architecture



Agentic RAG and Efficiency




Red Hat OpenShift AI
kube:admin
Home
Data science projects
Models
Model catalog
Model registry
Model deployments
Data science pipelines
Experiments
Distributed workloads
Applications
Resources
Settings
Workbench images
Cluster settings
Accelerator profiles
Serving runtimes
Connection types
Accelerator profiles
Manage accelerator profile settings for users in your organization
Name
Filter by name
Create accelerator profile
1 - 1 of 1
1
of 1
Name
Identifier
Enable
Last modified
Intel Gaudi3 PCIe
Intel Gaudi3 PCIe AI Accelerator
habana.ai/gaudi
2 days ago
1 - 1 of 1
1
of 1


KalturaCapture Edit
Tue 23 Sep 19:58
console-openshift-console.apps.snogaudi3p.fm2aihpcsed.com/k8s/cluster/projects/eligibility-mcp/workloads?view=graph
Red Hat OpenShift
kube:admin
Administrator
Home
Operators
Workloads
Serverless
Networking
Storage
Builds
Observe
Compute
User Management
Administration
You are logged in as a temporary administrative user. Update the cluster OAuth configuration to allow others to log in.
Projects > Project details
PR eligibility-mcp
Active
Actions
Overview
Details
YAML
Workloads
RoleBindings
Application: All applications
View shortcuts
Display options
Filter by resource
Name
Find by name...
eligib...ground
eligib...engine
eligibility-mcp-llamastack
eligibility-lsd
granit...dictor
granit...dictor
llama-...dictor
eligibility-mcp-llamastack

Q & A

Apply for a **free** Gaudi 3 Proof of Concept in **30 seconds**

Choose your GenAI or Virtualization PoC:

- Building Inference, RAG, AgenticAI, Model-as-a-Service, and other AI Use Cases with Intel Gaudi and Xeon
- Optimize finetuning with intel Gaudi

Why work with Intel + Red Hat?:

- Benefit from access to free highly qualified experts from Red Hat and Intel and free access to the latest hardware to build your AI use case / application.

If selected, a Intel / Red Hat representative will contact you via email.

Come visit the Intel and Red Hat booths to learn more!

Red Hat Summit

Connect

Thank you


linkedin.com/company/red-hat


facebook.com/redhatinc


youtube.com/user/RedHatVideos


twitter.com/RedHat