Red Hat Summit

Connect

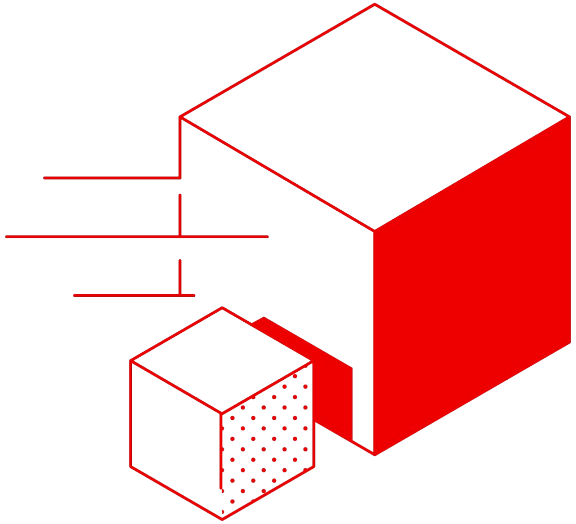
Becoming a private AI Provider with Red Hat AI

Model-as-a-Service while maximizing GPU cost efficiency

Sander Snel

Sr. Specialist Solution Architect AI Platform
Red Hat

Inference is where the real world value happens



- **Powers the AI experience** where users interact with models
- **Can happen anywhere** across hardware, models, and the hybrid cloud
- **Creates value for AI initiatives** by delivering on desired business outcomes

Source:
What is AI Inference?, Jan 2025, Red Hat
AI Inference Market Size, Share & Trends, Feb 2025, Markets and Markets
Amazon Found Every 100ms of Latency Cost them 1% in Sales, July 2023, Gigaspaces

The Operational Challenges in the Inference Era

Infrastructure cost

Requires substantial compute power to deliver expected experience

Operational complexities

Non- standardized approach creates inefficiencies

Deployment constraints

Inference across hybrid environments can lack flexibility

vLLM

Build the fastest and easiest-to-use
open-source LLM inference & serving engine

The De Facto OSS Inference Platform

vLLM is emerging as the Linux of GenAI Inference

High Performance

- Advanced algorithms for high QPS serving
- Single server/GPU to distributed/multi GPU
- Competitive with Proprietary NVIDIA stack
- The “comparison point” for alternative methods

Cross Platform

Enable all accelerators

NVIDIA GPU

AMD Instinct

intel Gaudi

Google TPU

aws Neuron

IBM AIU

Enable all OEMs

DELL

Lenovo

cisco

Hewlett Packard Enterprise

Easy To Use

- Native Hugging Face integration
- Simple APIs for online and offline inference
- Broadest feature set and model support
- Developer and IT productivity

Bringing robust enterprise inference to the Red Hat hybrid cloud.

The value of vLLM

Deliver fast, flexible and scalable inference

Faster response time

vLLM can achieve higher throughput, this translates to processing more tasks or requests within a given amount of time.

Efficient memory management

vLLM organizes virtual memory, this translates to handling larger models and longer sequences more effectively within a given hardware setup.

Reduce hardware costs

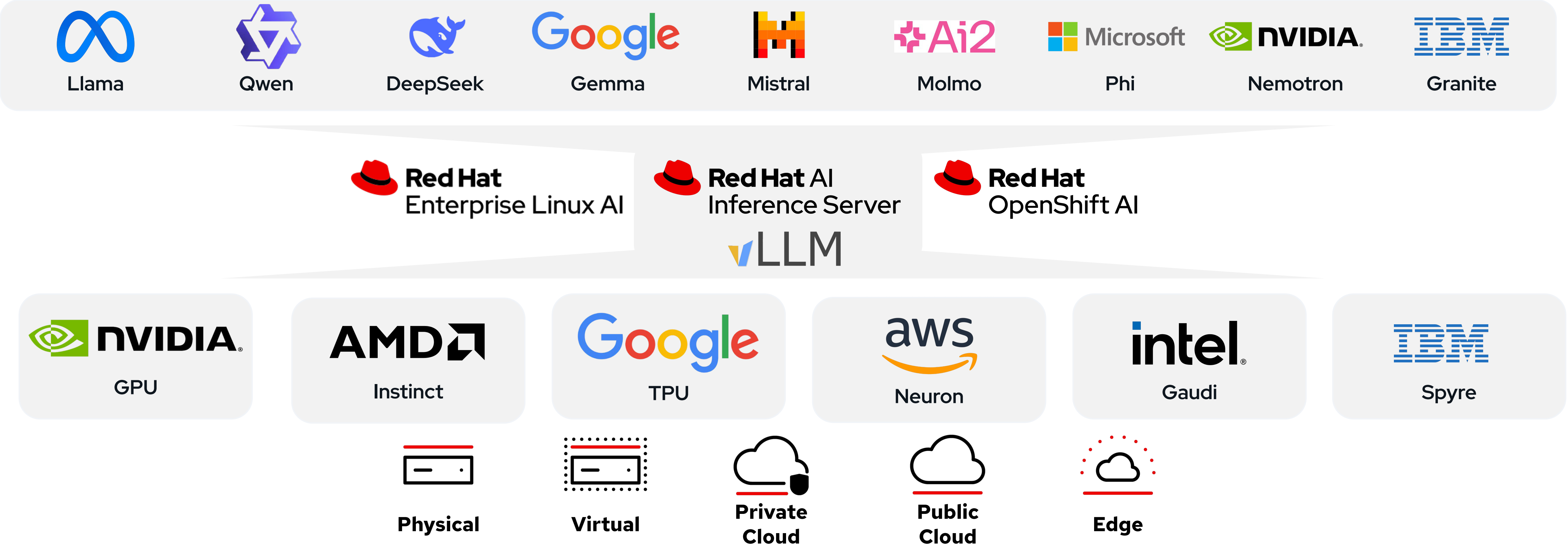
vLLM offers a more efficient use of resources, which is equivalent to fewer GPUs needed to handle the processing of LLMs.

Designed for security and scale

Self-hosting an LLM with vLLM provides you with more control over data privacy and usage, as well as an ability to handle growing demand.

Red Hat AI Inference Server

vLLM connects model creators to accelerated hardware providers



Llama
Qwen
DeepSeek
Gemma
Mistral
Molmo
Phi
Nemotron
Granite

Red Hat Enterprise Linux AI
Red Hat AI Inference Server
Red Hat OpenShift AI

vLLM

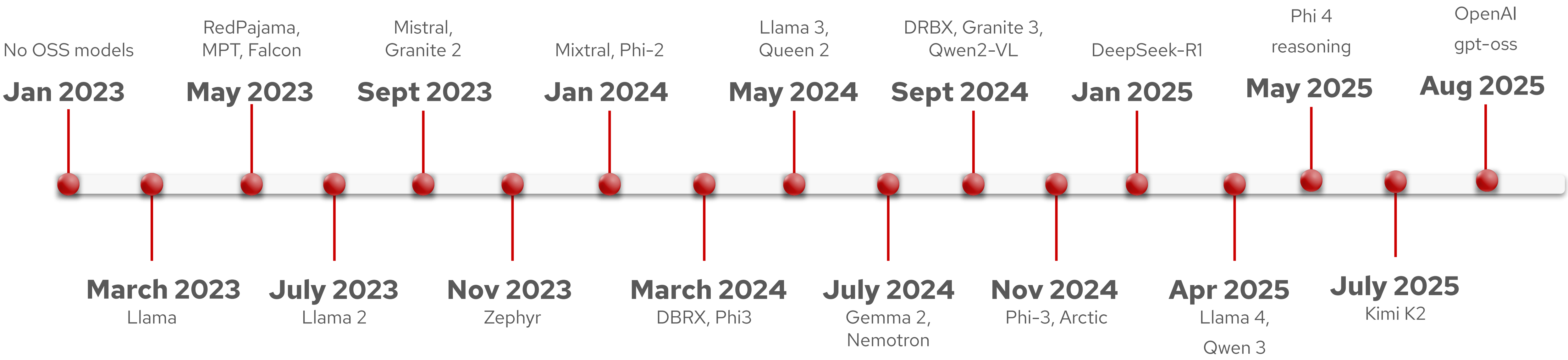
GPU
Instinct
TPU
Neuron
Gaudi
Spyre

Physical
Virtual
Private Cloud
Public Cloud
Edge

Single platform to run any model, on any accelerator, on any cloud

Expanding choice of models

There has been an explosion of capability from open-source over the last 2 years



Jan 2023
No OSS models

March 2023
Llama

May 2023
RedPajama, MPT, Falcon

July 2023
Llama 2

Sept 2023
Mistral, Granite 2

Nov 2023
Zephyr

Jan 2024
Mixtral, Phi-2

March 2024
DBRX, Phi3

May 2024
Llama 3, Queen 2

July 2024
Gemma 2, Nemotron

Sept 2024
DRBX, Granite 3, Qwen2-VL

Nov 2024
Phi-3, Arctic

Jan 2025
DeepSeek-R1

Apr 2025
Llama 4, Qwen 3

May 2025
Phi 4 reasoning

July 2025
Kimi K2

Aug 2025
OpenAI gpt-oss

Meta · Ai2 · NVIDIA · MISTRAL AI_ · Hugging Face · databricks · snowflake · IBM · Google · Microsoft · deepseek

Red Hat AI repository on Hugging Face

Collection of third-party models

Llama

Qwen

Gemma

Mistral, Voxtral

DeepSeek

Phi

Molmo

Granite

Nemotron

OpenAI
GPT-oss

K2

SMOLI M3 3B

Choice of Models

- Transformers (Dense, MOE), Multi-modal LLMs, Embeddings Models, Hybrid / Novel Attention, Vision
- Hugging Face compatible (safe tensors), OCI-compatible containers

Validated models

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

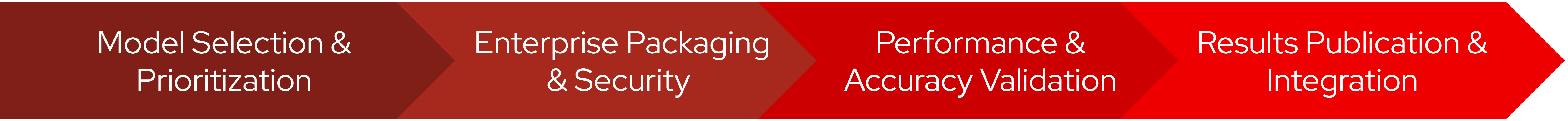
Optimized models

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

Cut GPU costs with inference optimized models.

Red Hat AI Model Validation Methodology

A rigorous and transparent process to deliver trusted, enterprise-ready AI models



Model Selection & Prioritization

We prioritize models for validation based on a continuous analysis of the AI ecosystem, driven by several key inputs:

- **Customer Demand**
- **OSS Market Leadership**

Enterprise Packaging & Security

Selected models are packaged as OCI artifacts and ModelCars. This enterprise-grade packaging is a critical step that enables:

- **Security**
- **Lifecycle Management**

Performance & Accuracy Validation

Each model is rigorously validated for performance across diverse hardware (NVIDIA, AMD, etc.) and various use cases.

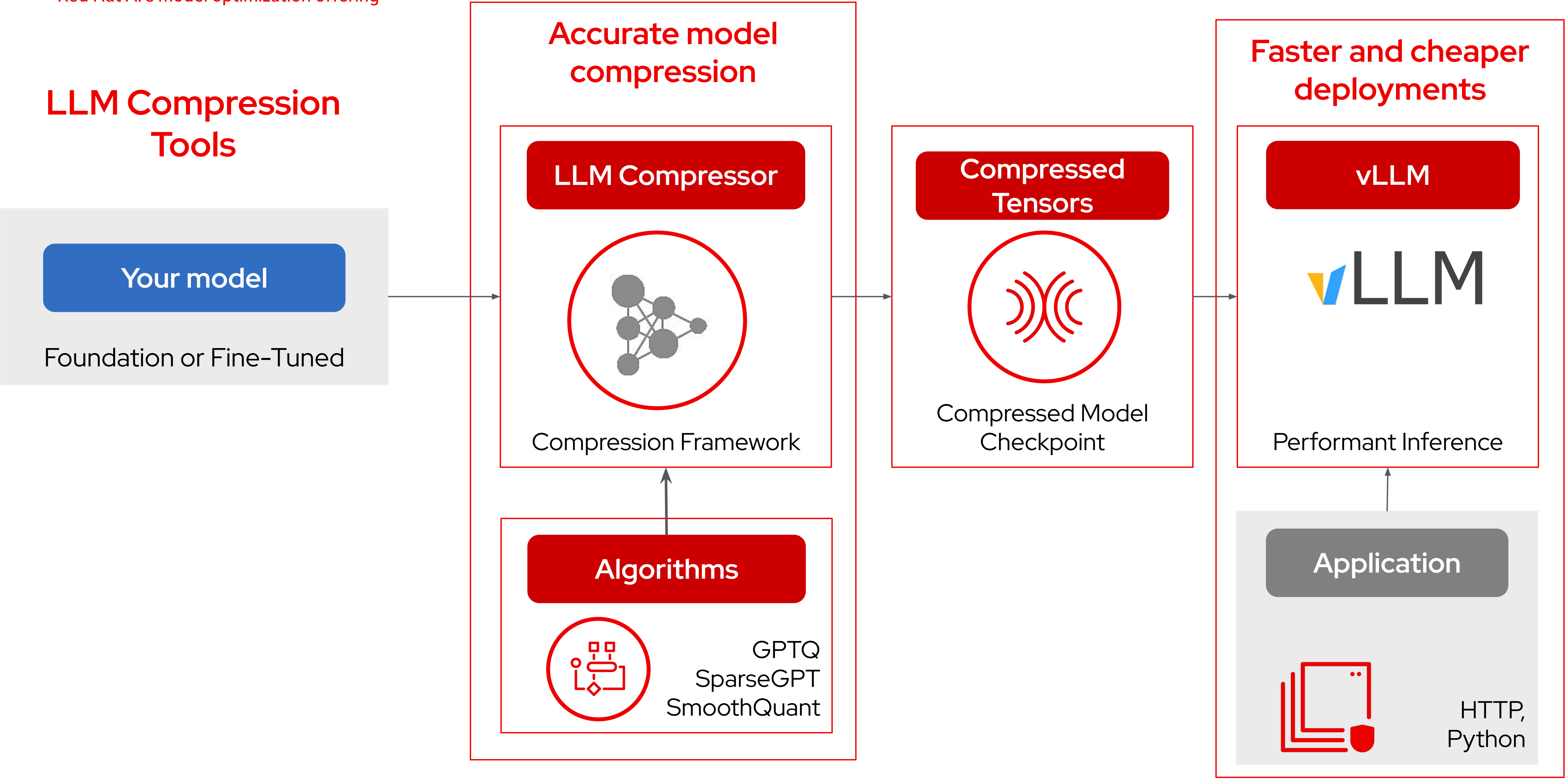
- **Performance Benchmarking**
- **Accuracy Evaluations**

Results Publication & Integration

All generated data is aggregated and published to empower our teams and customers:

- **Customer-Facing Model Catalog** (Coming to RHOAI in 3.0)

LLM Compression Tools



Your model

Foundation or Fine-Tuned

Accurate model compression

LLM Compressor

Compression Framework

Algorithms

GPTQ
SparseGPT
SmoothQuant

Compressed Tensors

Compressed Model Checkpoint

Faster and cheaper deployments

vLLM

Performant Inference

Application

HTTP,
Python

Red Hat: Leaders in Open Source GenAI Inference

Red Hat has built a comprehensive set of model optimization capabilities to drive operational efficiencies

Third-party validated and optimized models

Llama

Qwen

Gemma

Mistral

DeepSeek

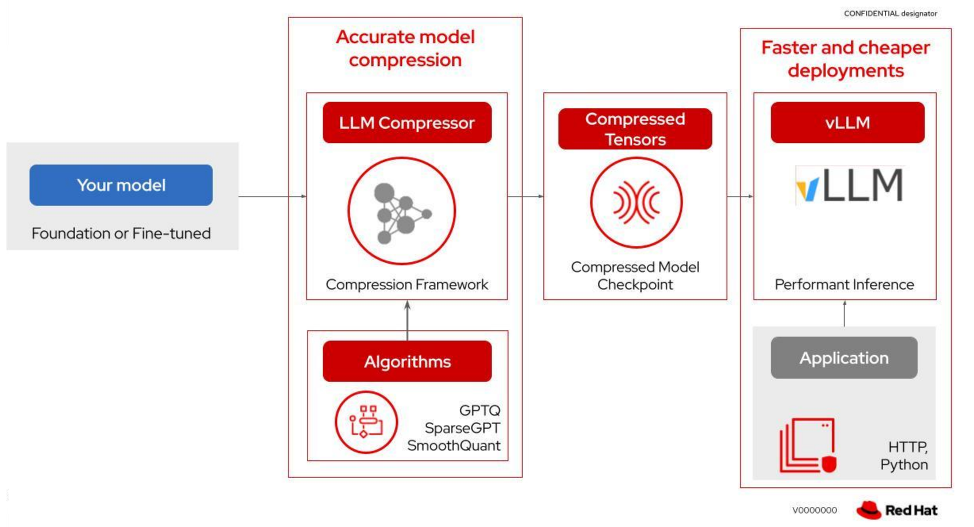
Phi

Molmo

Granite

Nemotron

LLM Compression Tools



Hosted on the Red Hat AI repository on Hugging Face

Connecting Models to the Hardware through vLLM

Expanding Hardware Support

NVIDIA	GB200 and RTX PRO 6000
AMD	Instinct MI35x series
intel	Gaudi new hardware plugin
Google	TPU architecture
aws	Neuron enablement
IBM	Spyre Hardware plugin
rebellions_	Hardware plugin for ATOM
METAX 沐曦	Hardware plugin with MACA

Day-Zero Model Support

gpt-oss

Qwen3-Next

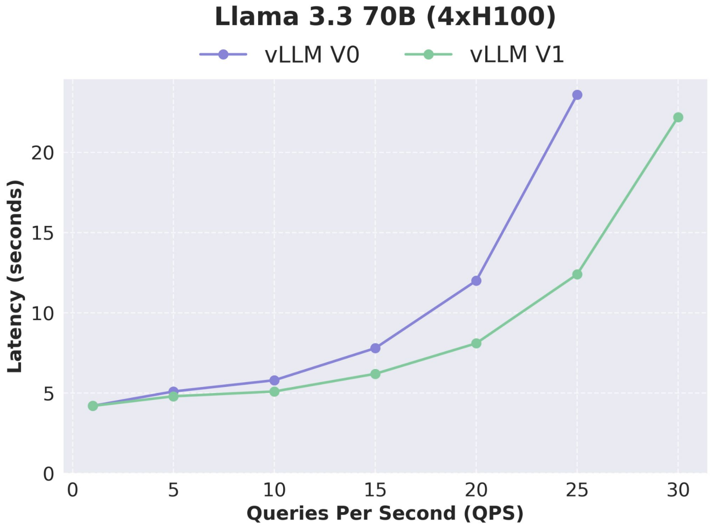
Gemma 3n

Multimodal in vLLM 0.10.1

Qwen2.5-Omni

PRITHVI GEOSPATIAL

V1 Unified Architecture



Re-architecture of vLLM core components, including the scheduler, KV cache manager, worker, sampler, and API server.

Red Hat AI delivers consistent, fast and cost-effective inference

Select a large language model

A catalog of ready-to-use, third party validated and optimized models

Choose an inference runtime

vLLM

An optimized engine to deliver fast, cost-effective, and consistent inference

Choose the hardware that works best for you

AMD NVIDIA Google intel aws IBM

vLLM connects model creators to accelerated hardware providers

Scale AI inference when ready

llm-d

Llm-d provides consistent, distributed, inference at scale

Inference at scale everywhere

Distributed, scalable gen AI inference for Enterprise AI

Red Hat AI

Now includes llm-d

- Lower infrastructure cost & increased efficiency
- Faster response times for multi-turn & agent workloads
- Simplified management for platform administrators

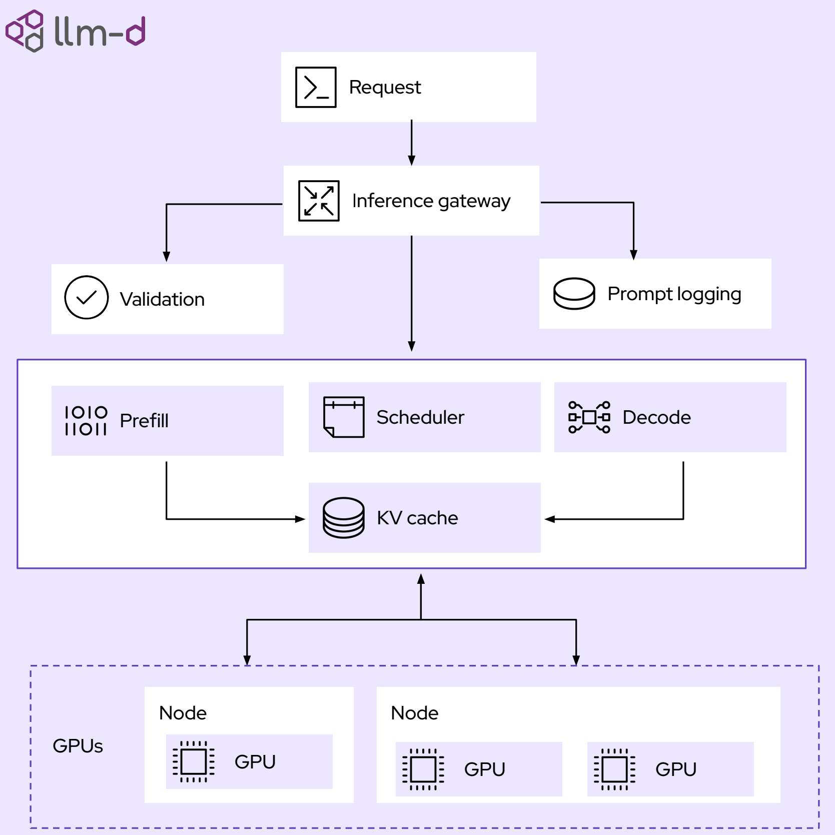
Deliver faster, cheaper, and more manageable AI systems for enterprise production

llm-d reimagines how LLMs run on Kubernetes

Distributed Inference with llm-d

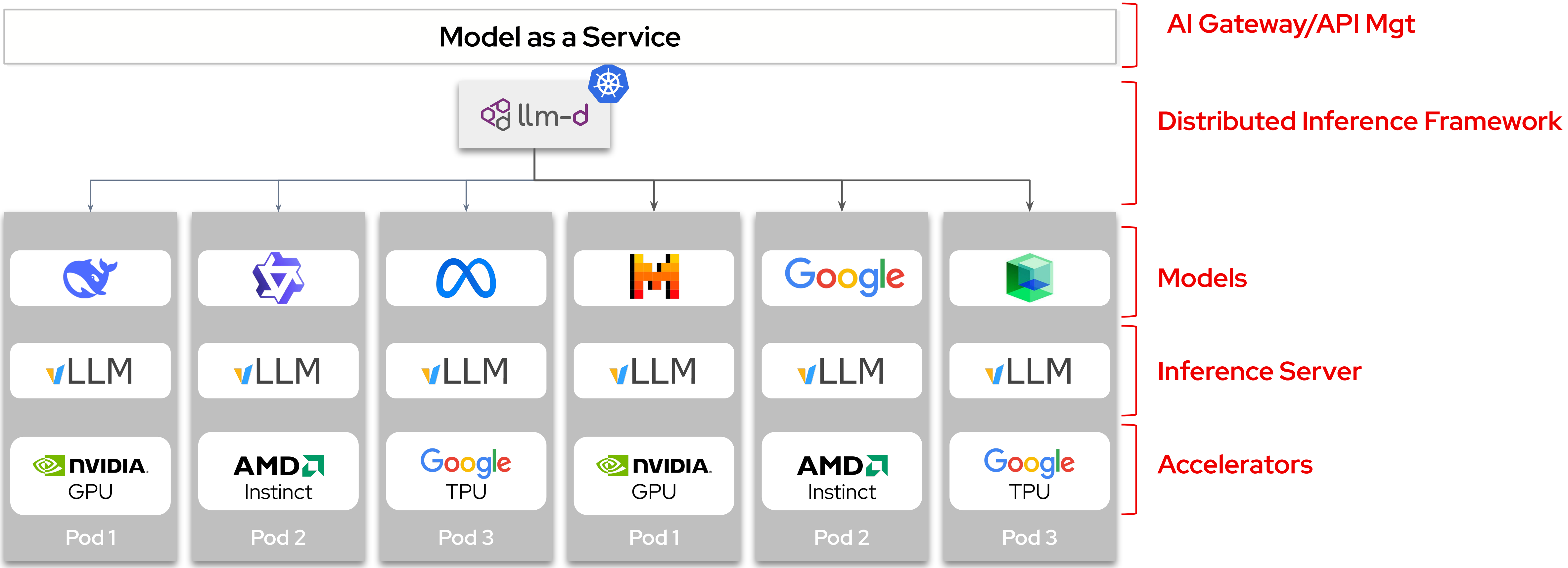
Maximize GPU utilization and deliver on your SLOs with distributed inference

- Joint open source project by Red Hat, Google, NVIDIA, AMD, Hugging Face, and many more
- Kubernetes-Native Architecture for simple deployment and management of GenAI models
- Optimized GenAI Inference to accelerate LLM's and MoE
- Intelligent Resource Utilization to reduce inference costs
- High Performance and Scalability to meet demanding Service Level Objectives (SLOs).
- Supported on Heterogeneous Hardware like NVIDIA and AMD GPUs (and many more to come in the future)



Enterprise GenAI inference platform

Holistic approach to optimize and operationalize deployment and scaling of open-source LLMs



Become the **Private AI Provider** for your organization

What is Models as a Service

- Strategy delivering central AI services privately
- Model service consumed by large audience
- Accessible to Developers and Associates
- GPUs invisible to user, critical for cost optimization

Why IT should become the Private AI Provider

- Compliant with existing security, data & privacy policies
- Predictable costs & increased utilization
- Reduce time to market with AI applications
- Unified & impactful service delivery

How value is created

- AI managed like any other workload
- Innovation across entire organization

Use cases

Private Assistant (RAG)

Embedding Application

Code Assistant

Image Generation

Agentic

Your next app

Red Hat can help

Red Hat OpenShift AI

Red Hat OpenShift

Flexible and Efficient Inference

- GA distributed inference (llm-d)
- New validated and optimized models
- vLLM enhancements
- LLM Compressor GA

Connecting Models to Data

- Modular and extensible approach for: data ingestion, synthetic data generation, tuning, evaluations.
- RAG enhancements & partner integrations
- Continual Post Training Algorithm
- Feature Store GA

Red Hat AI

Agentic AI

- AI experiences: AI hub and gen AI studio
- Model Context Protocol support & MCP Server access in gen AI studio
- Llama Stack API integration

AI Platform

- Model catalog and registry GA
- Model as a Service provider enhancements and API Mgt integration
- GPU as a Service enhancements

Single platform to run any model, on any accelerator, on any cloud

Red Hat Summit

Connect

Thank you

in linkedin.com/company/red-hat

f facebook.com/redhatinc

▶ youtube.com/user/RedHatVideos

twitter.com/RedHat