

Any model, any accelerator, any cloud – Red Hat's AI vision uncovered



Andreas Bergqvist

AI Sales Specialist
Red Hat





Connect

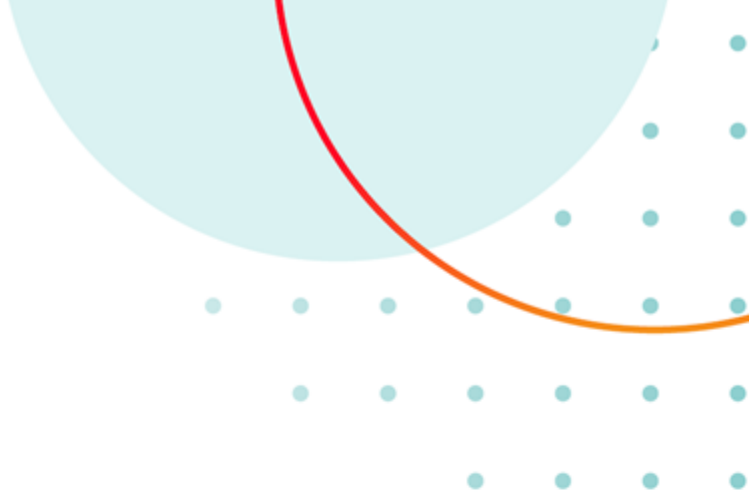
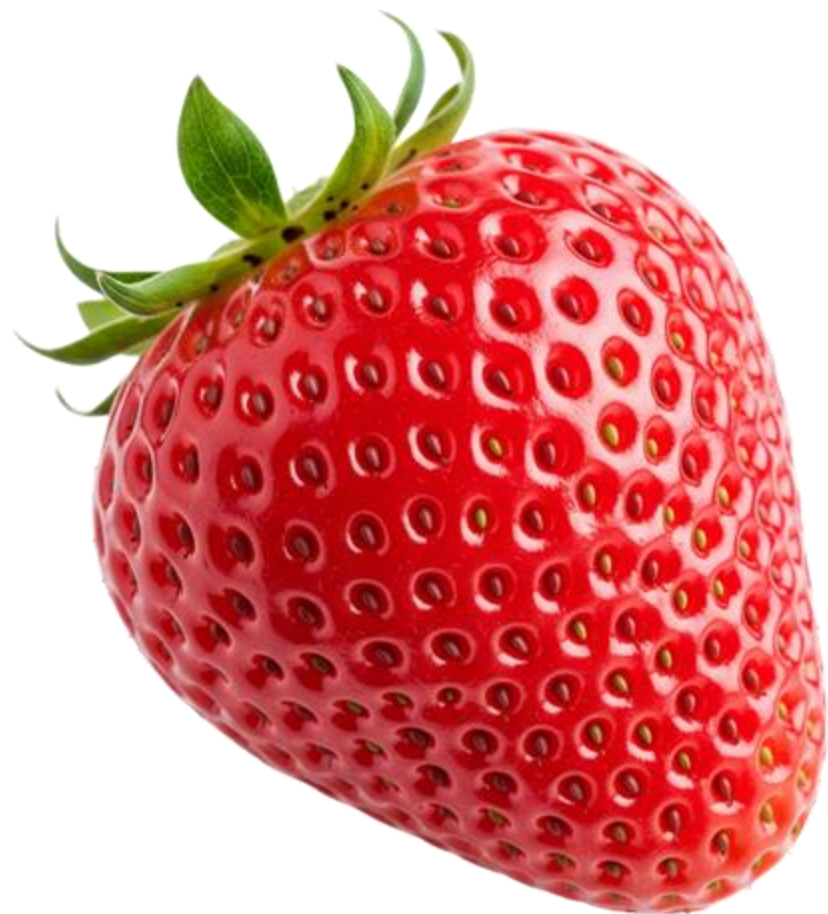
Red Hat AI

Any Cloud, Any Model, Any Accelerator

Andreas Bergqvist

Red Hat AI EMEA





What Red Hat is known for



An **enterprise software company** with an **open source** development model



Broad, innovative offerings fine-tuned for **hybrid** and **AI**



Trusted, comprehensive, and **consistent** portfolio



“Your choice of where to run AI will be
everywhere

And it's going to be based on **open source.**”

Matt Hicks, CEO Red Hat ^[1]

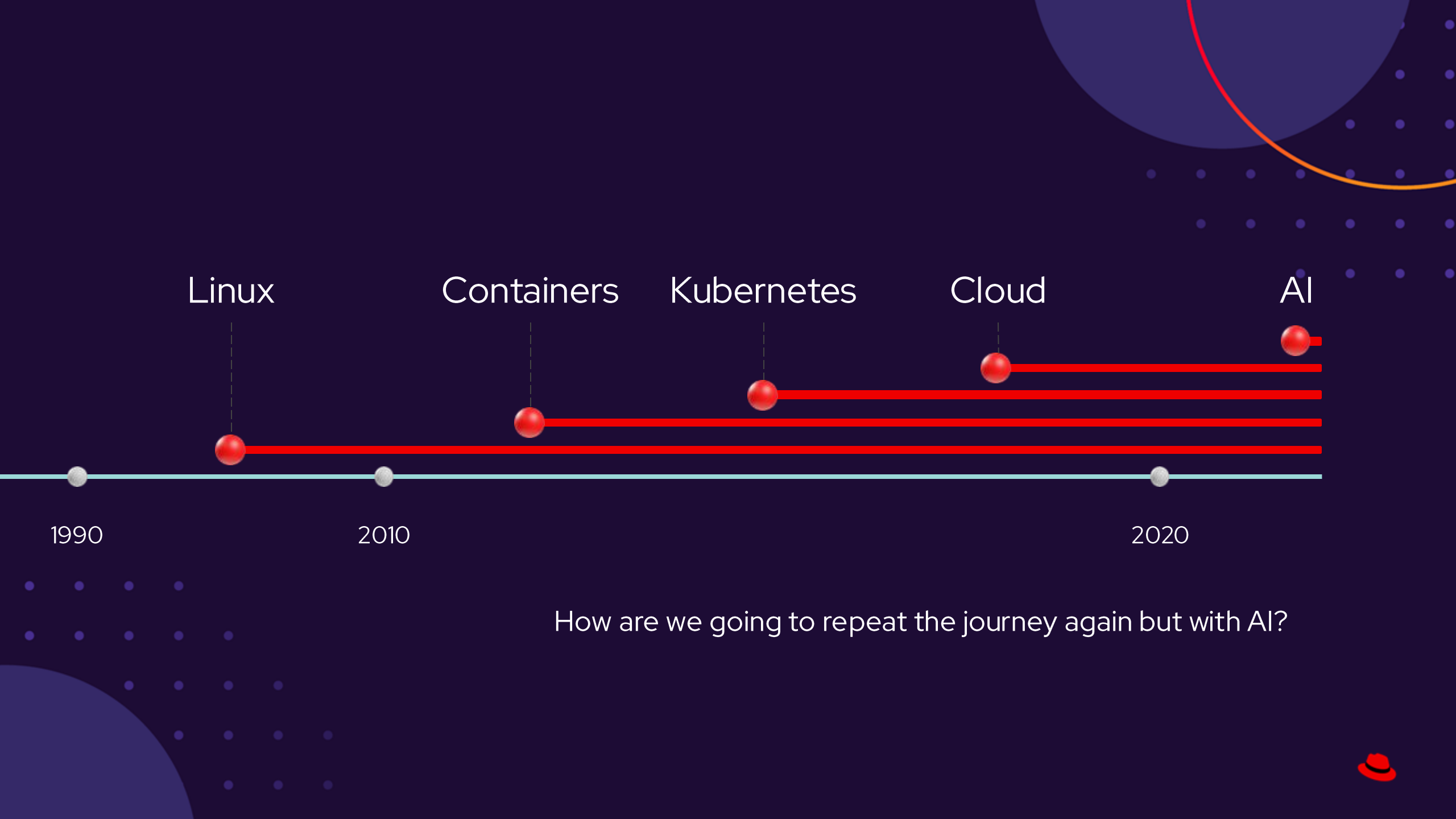
Our goal:

Help enterprise customers being able to leverage AI on Hybrid cloud
with platform flexibility, scalability, and robustness

Source:

[1] [Unleashing the Power of Hybrid Cloud and AI](#) (Matt Hicks, CEO Red Hat at Red Hat Summit





Linux

Containers

Kubernetes

Cloud

AI

1990

2010

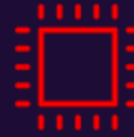
2020

How are we going to repeat the journey again but with AI?



Choices in accelerators
and infrastructure. It's
your call. **We support it all.**

AI Infrastructure



Processors/
Accelerators



Public
clouds



Private
clouds

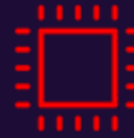


Edge
computing



Let's put your foundation
to work with a
**comprehensive AI
platform.**

AI Infrastructure



Processors/
Accelerators



Public
clouds



Private
clouds

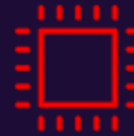


Edge
computing



Choosing models for
gen AI can be complex.
We made it simpler.

AI Infrastructure



Processors/
Accelerators



Public
clouds



Private
clouds

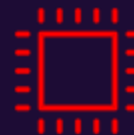


Edge
computing



A boost in productivity and
efficiency, from the jump.
Meet Lightspeed..

AI Infrastructure



Processors/
Accelerators



Public
clouds




Private
clouds



Edge
computing

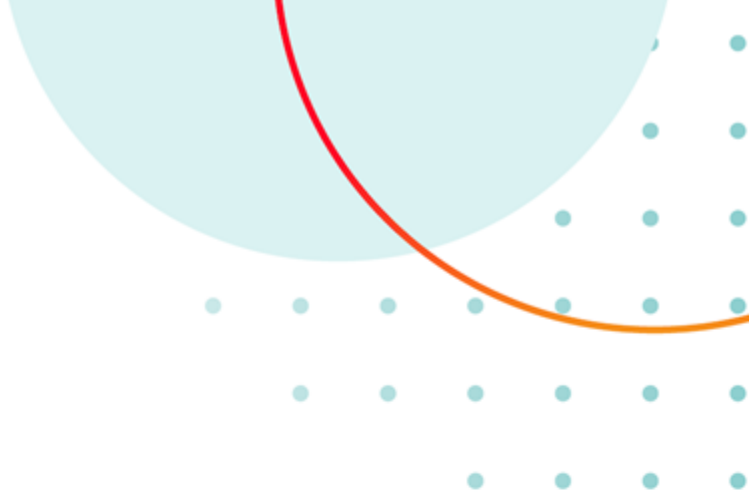
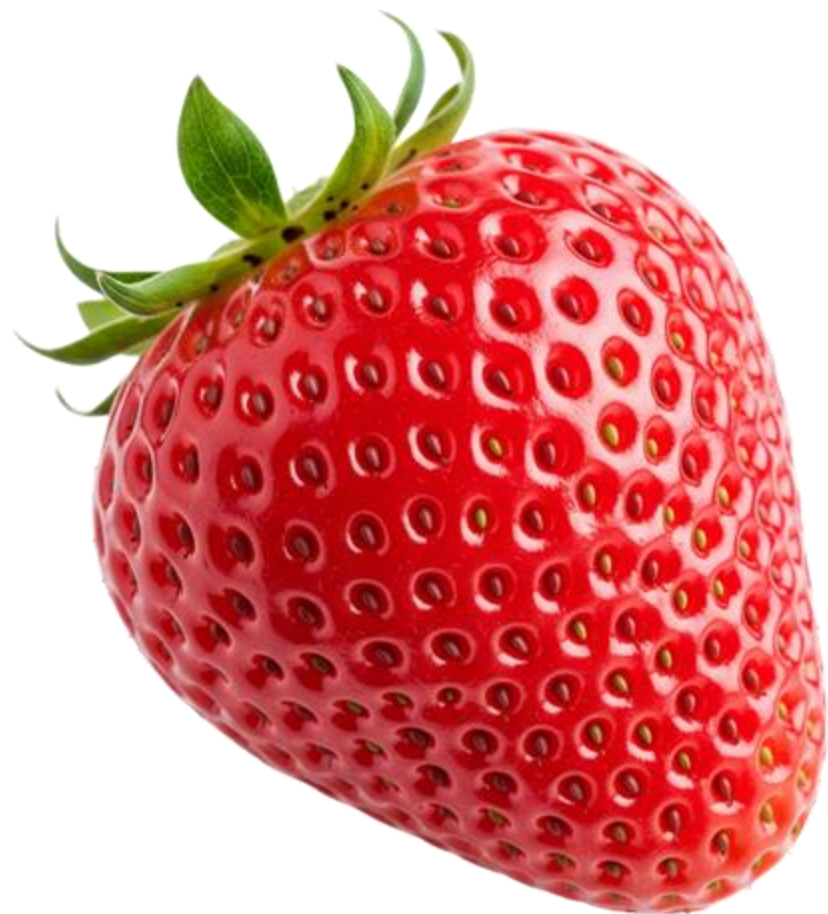


 **Red Hat**
Enterprise Linux

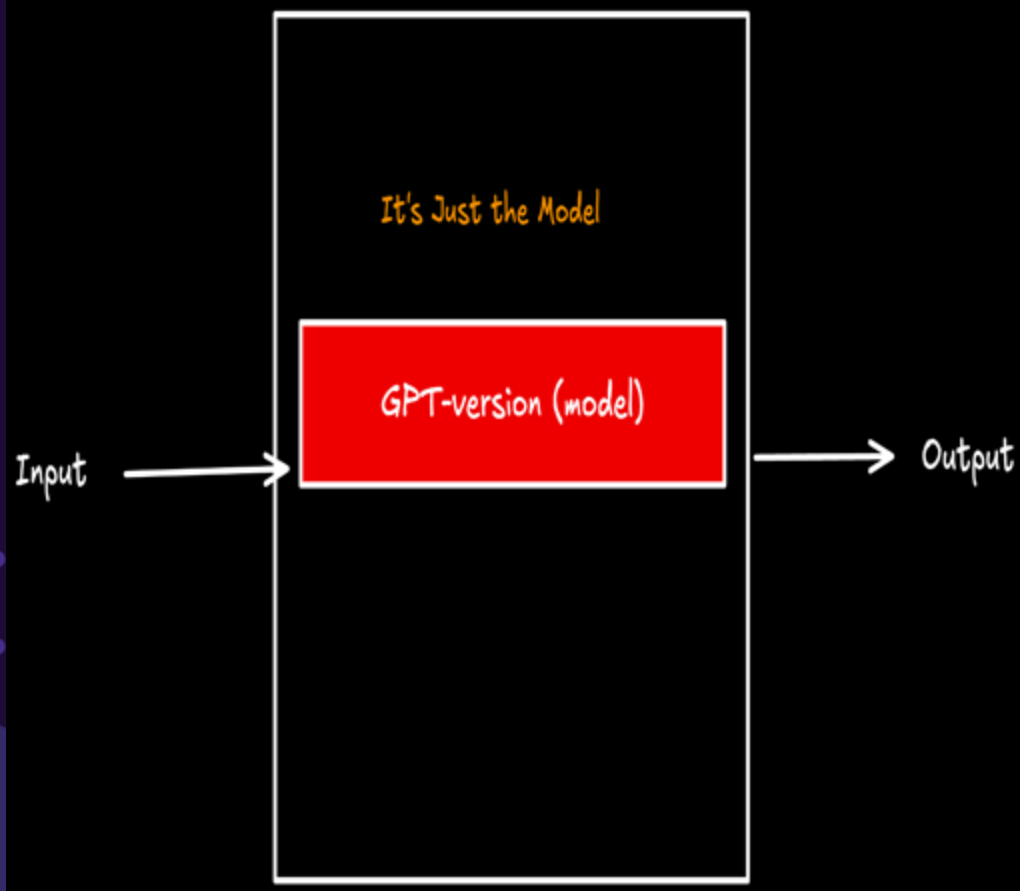
 **Red Hat**
Ansible Automation
Platform

 **Red Hat**
OpenShift

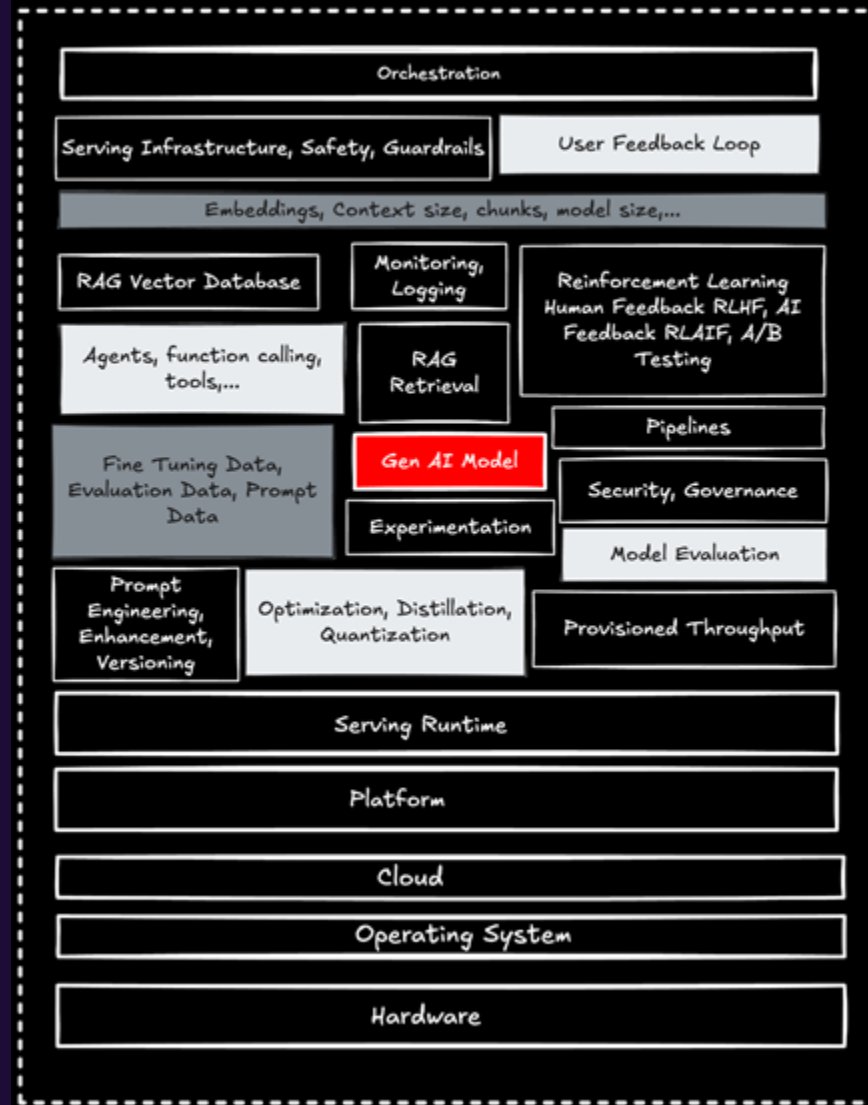




Perception



Reality





Where do Red Hat invest?

3 areas to keep an eye on



1

Llama Stack

An open source generative AI platform.



More info will come in the coming weeks
Tell your friends

More info will come in the coming weeks
Tell your friends

Why nVidia is the highest valued company in the world

LLM Model	Model Size (Parameters)	Inference VRAM (FP16 / INT8 / INT4)	Fine-Tuning VRAM (Full / PEFT)	Recommended NVIDIA GPU(s)	Cloud Cost (per hour, on-demand)	Purchase Cost (Approx. USD)
Llama 3 8B	8B	~16 GB / ~9 GB / ~5 GB	~70-80 GB / ~12-16 GB	RTX 4090, L40S	~\$0.20 - \$1.80	~\$1,800 - \$2,500
Mixtral 8x7B	47B (effective)	~94 GB / ~50 GB / ~28 GB	>300 GB / ~48 GB	2x L40S, A100 80GB, H100 80GB	~\$1.80 - \$4.00	~\$9,000 - \$17,000
Llama 3 70B	70B	~140 GB / ~75 GB / ~40 GB	>400 GB / ~60-80 GB	2x L40S, 2x A100 80GB, H100 80GB	~\$1.80 - \$4.50	~\$9,000 - \$35,000



2

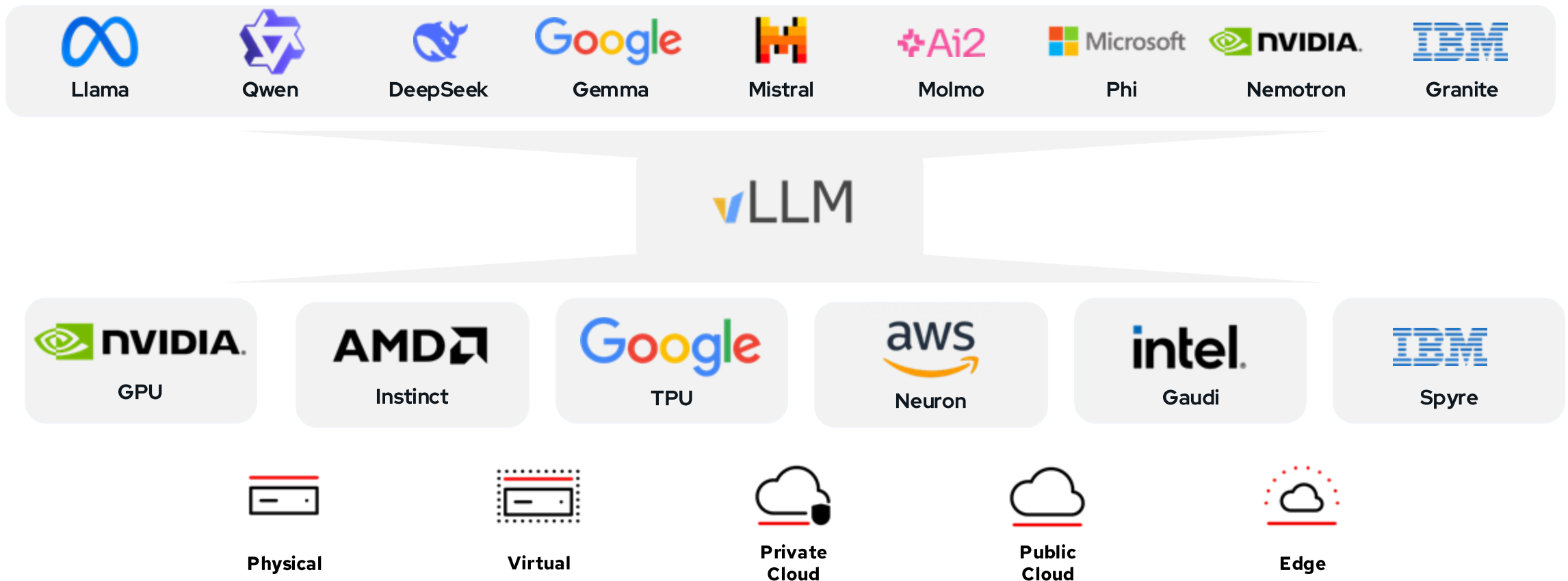
vLLM

Standardised and Efficient inference



Red Hat AI the inference engine for the hybrid cloud

vLLM supports the key models on the key hardware accelerators



Fast, flexible and scalable inference

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



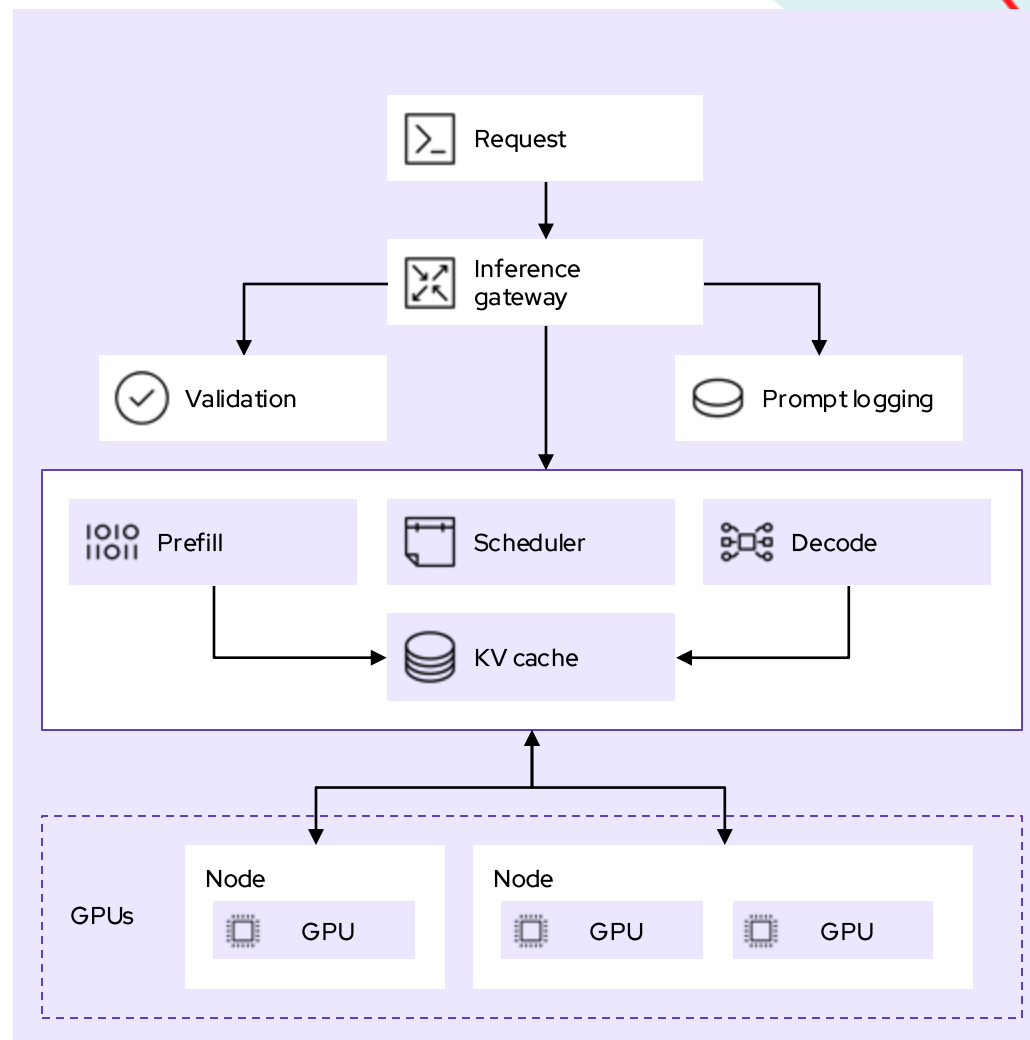
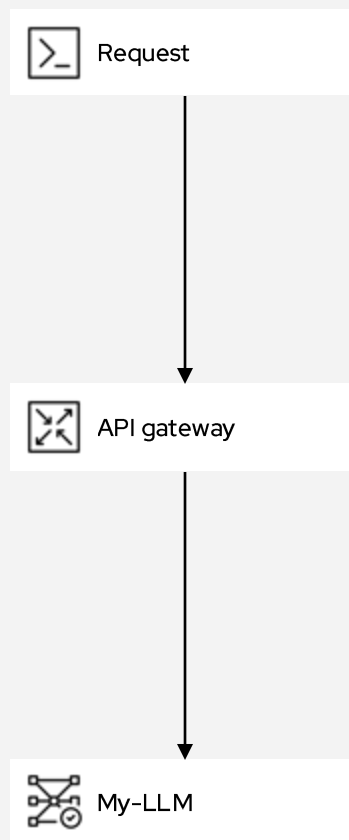
3

Llm-d

Distributed inference



Traditional LLM



What is llm-d?

An open-source, Kubernetes-native framework for distributed LLM inference



An open-source framework
for distributed large language
model (LLM) inference that
runs natively on Kubernetes.

- ▶ Joint open source initiative by Red Hat, Google, NVIDIA, Hugging Face, and many more organizations
- ▶ Optimized for Kubernetes based distributed platforms (OpenShift)
- ▶ Designed to optimize scale, latency, and flexibility for AI workloads
- ▶ Built to extend and interoperate with vLLM
- ▶ **Maximizes performance across multi-tenant and multi-model workloads with intelligent scheduling and resource management**





Red Hat AI



Red Hat AI
Inference Server



Red Hat
Enterprise Linux AI



Red Hat
OpenShift AI

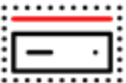
Trusted, Consistent and Comprehensive foundation



Hardware Acceleration



Physical



Virtual



Private
Cloud



Public
Cloud



Edge



Final words



AI is tasty but also complex

Red hat has 3 decades of making complex stuff easier.

AI is Red Hats biggest investment. Test us to see what we are up to in our open, no lock in world



Connect

Thank you



linkedin.com/company/red-hat



facebook.com/redhatinc



youtube.com/user/RedHatVideos



twitter.com/RedHat

