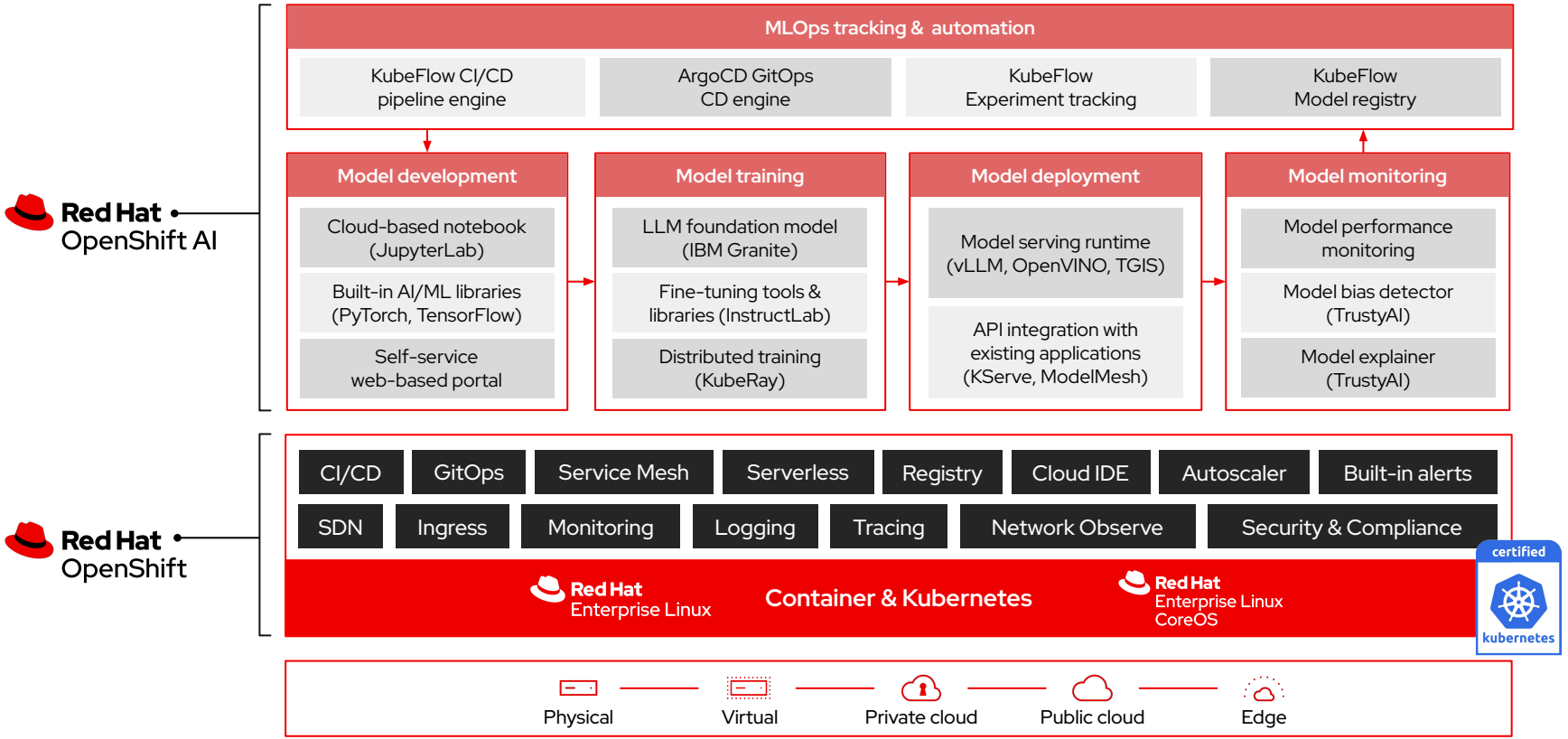
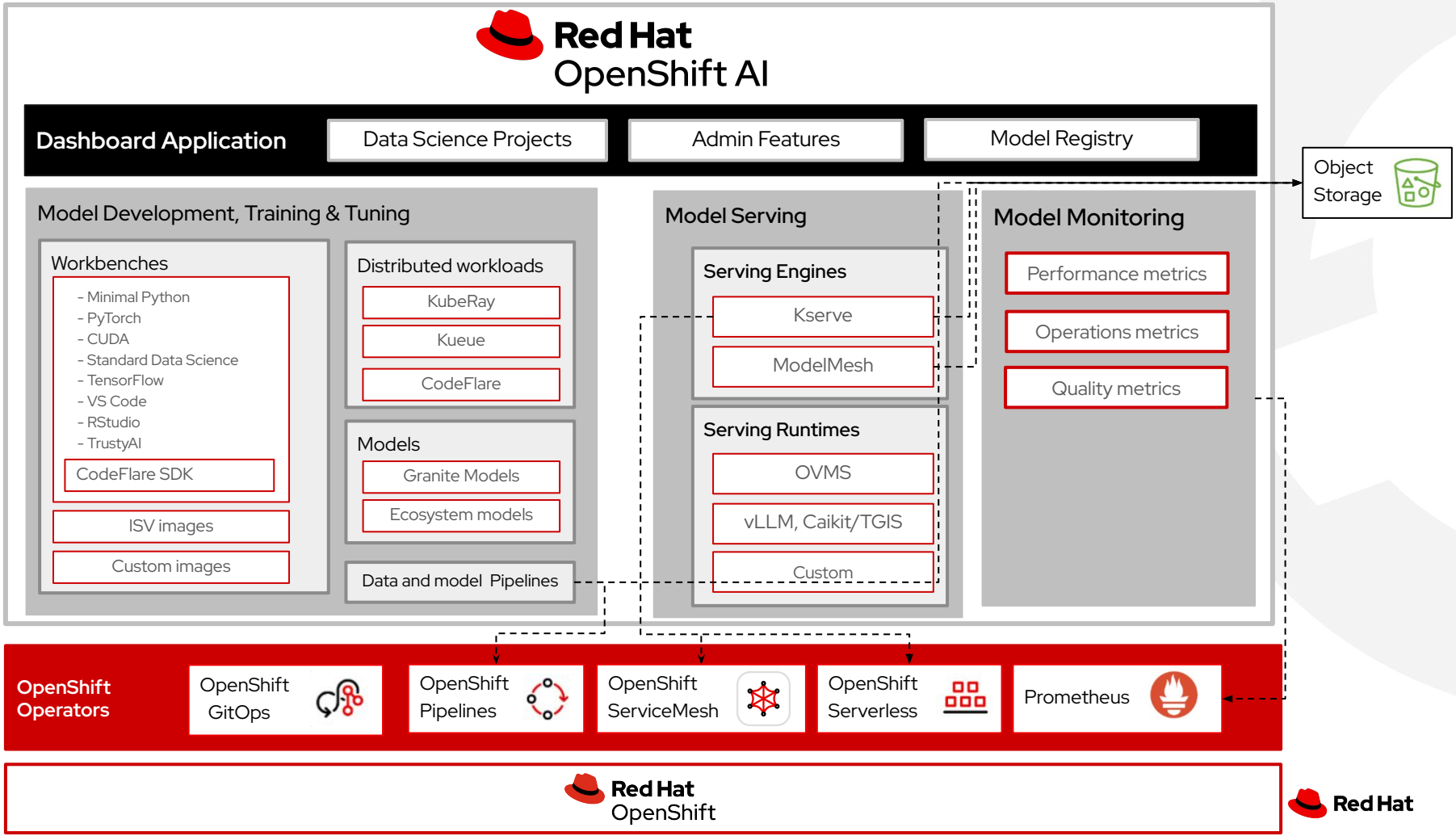
Red Hat OpenShift AI

Red Hat

OpenShift AI: the **MLOps** platform for **data scientists** and **operation teams** across **hybrid clouds**




Red Hat
OpenShift AI
Dashboard Application
Data Science Projects
Admin Features
Model Registry
Object Storage
Model Development, Training & Tuning
Workbenches
- Minimal Python
- PyTorch
- CUDA
- Standard Data Science
- TensorFlow
- VS Code
- RStudio
- TrustyAI
CodeFlare SDK
ISV images
Custom images
Distributed workloads
KubeRay
Kueue
CodeFlare
Models
Granite Models
Ecosystem models
Data and model Pipelines
Model Serving
Serving Engines
Kserve
ModelMesh
Serving Runtimes
OVMS
vLLM, Caikit/TGIS
Custom
Model Monitoring
Performance metrics
Operations metrics
Quality metrics
OpenShift Operators
OpenShift GitOps
OpenShift Pipelines
OpenShift ServiceMesh
OpenShift Serverless
Prometheus
Red Hat OpenShift

Red Hat

Red Hat OpenShift AI - Key features

Model development

Interactive, collaborative UI for **seamless access** AI/ML tooling, libraries, frameworks, etc.

Model serving

Model serving routing for **deploying models to production** environments

Model monitoring

Centralized monitoring for **tracking models performance and accuracy**

Data & model pipelines

Visual editor for **creating and automating** data science pipelines

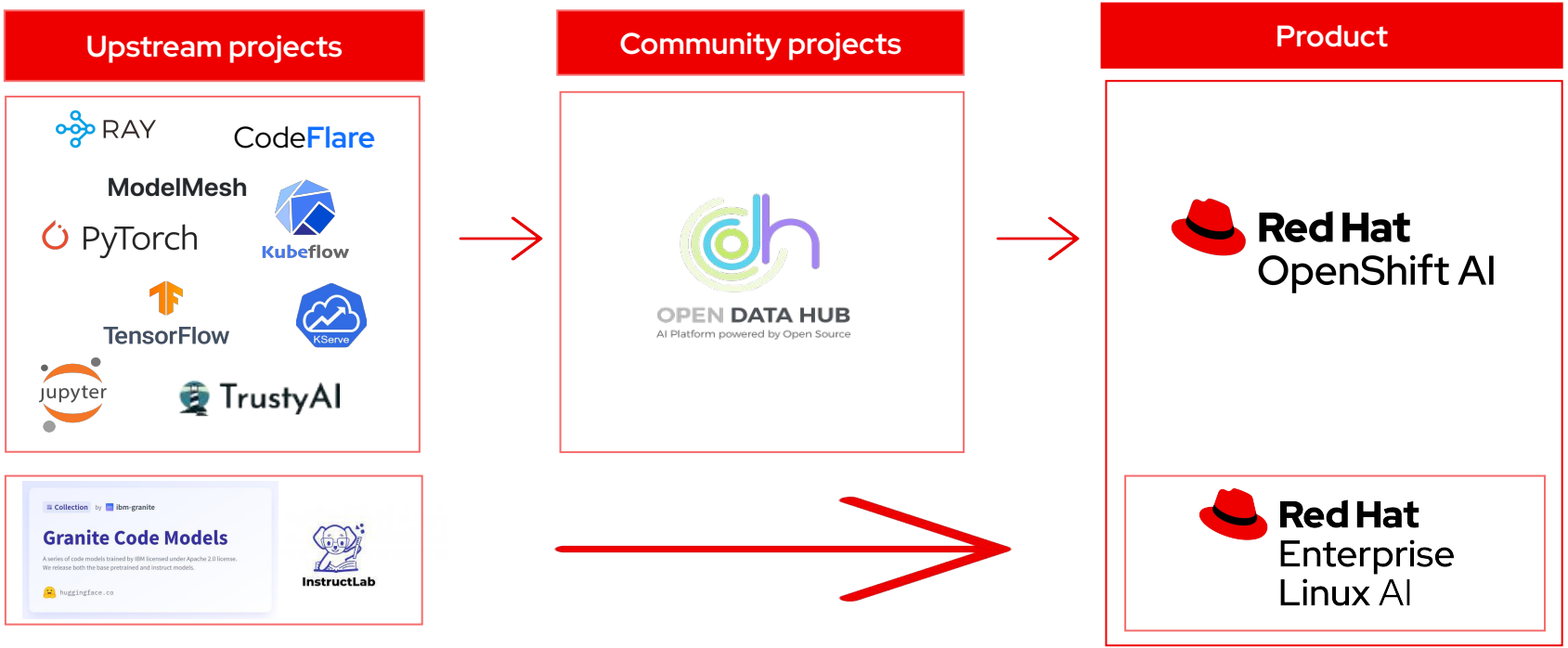
Distributed workloads

Seamless experience for **efficient data processing, model training, and tuning**

Trust & AI guardrails

Improve LLM accuracy, performance, latency and **transparency**

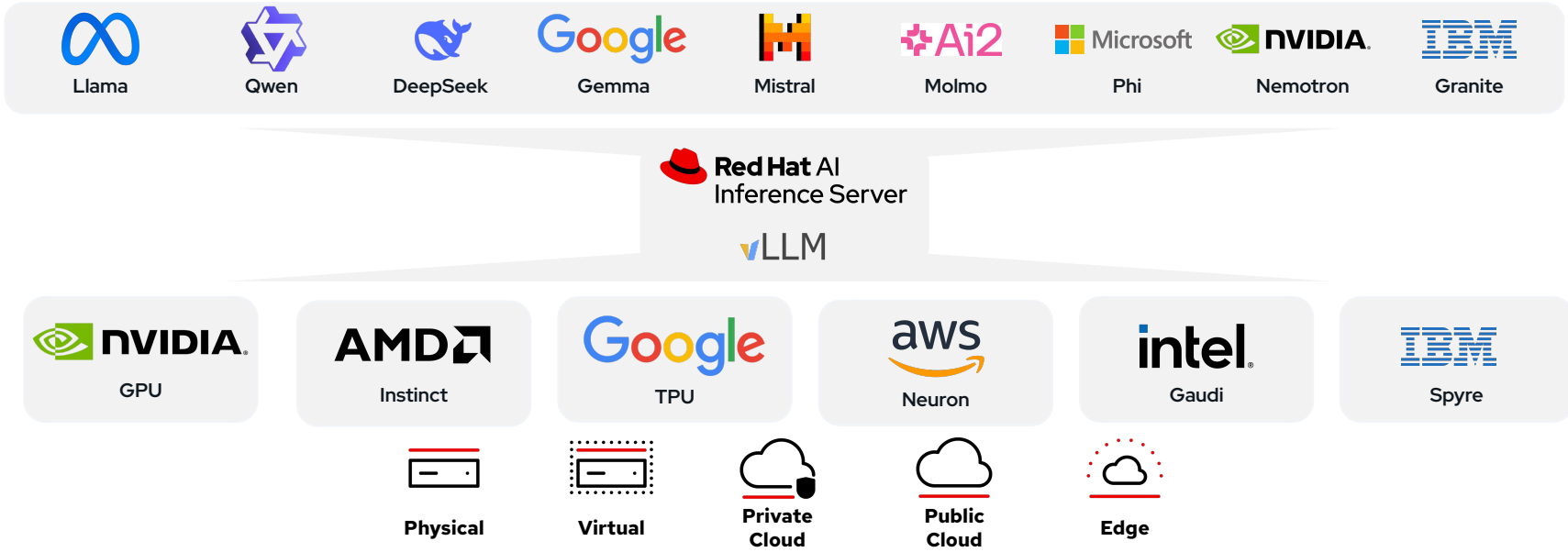
Red Hat's AI/ML engineering is 100% open source



Red Hat

Red Hat AI Inference Server

vLLM connects model creators to accelerated hardware providers



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

Red Hat

Use case #1: Cost reduction by model optimization

Red Hat

Model quantization by Neural Magic

Leader in LLM serving, now acquired by Red Hat

Quantized by
Neural Magic

NEURAL MAGIC

DeepSeek-R1-Distill-Llama-70B

DeepSeek-R1-Distill-Llama-8B

DeepSeek-R1-Distill-Qwen-32B

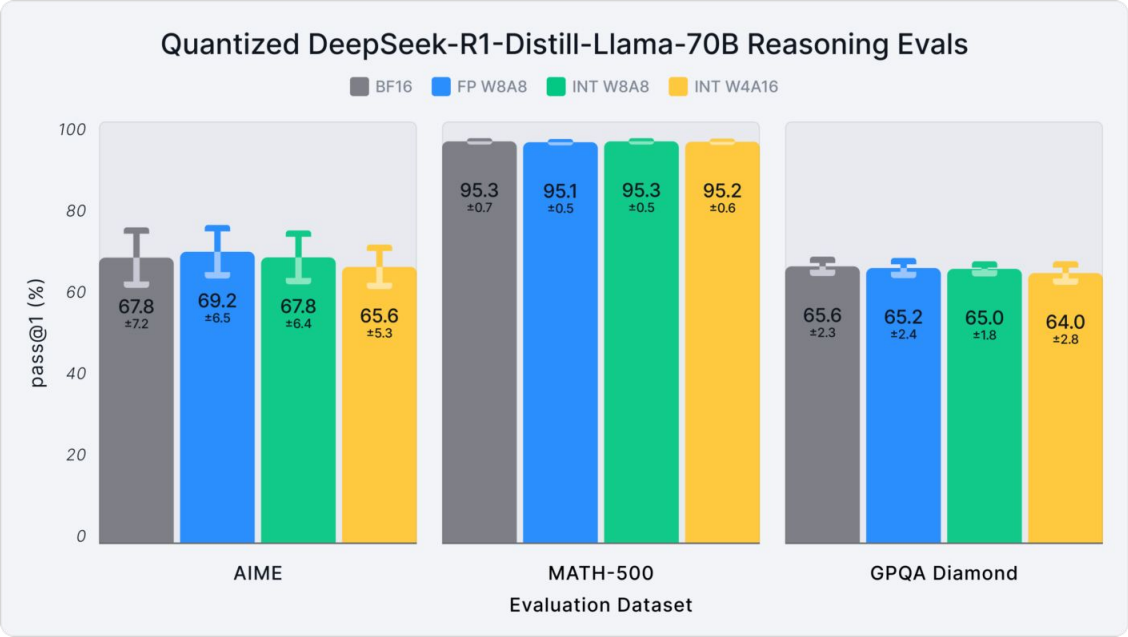
DeepSeek-R1-Distill-Qwen-14B

DeepSeek-R1-Distill-Qwen-7B

DeepSeek-R1-Distill-Qwen-1.5B

Red Hat

Fewer GPUs and $$$, yet **same level of model accuracy**



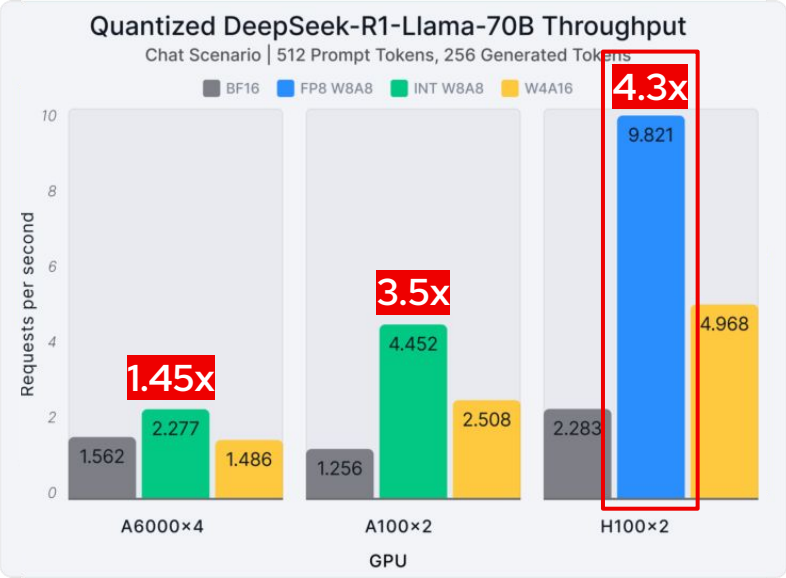
- **Nearly 100% accuracy recovery** in W8A8 compression (i.e. 50% of model compressed)
- Only around 2.67% accuracy loss in W4A16 compression (i.e. 75% of model weight compressed)

Reference: https://neuralmagic.com/blog/quantized-deepseek-r1-models-deployment-ready-reasoning-models/

Red Hat

Neural Magic **reduces DeepSeek deployment $$$**

Use fewer GPU cards and resources, yet achieve the same = Save cost!



LLM serving throughput

↑430%

- Powered by **Neural Magic**, company acquired by Red Hat
- Advanced **quantization** (i.e. model compression) techniques

Reference: https://neuralmagic.com/blog/quantized-deepseek-r1-models-deployment-ready-reasoning-models/

Red Hat

Use case #2: GPU sharing

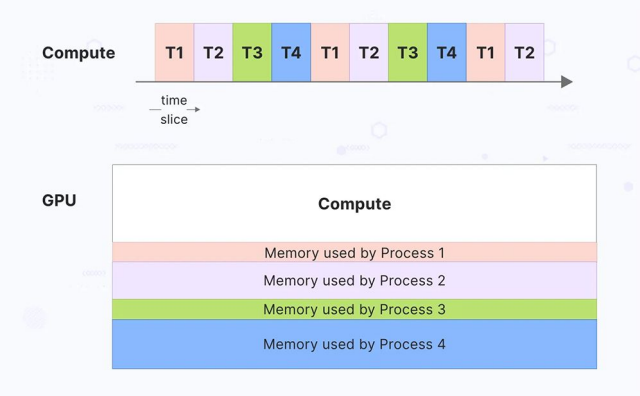
Red Hat

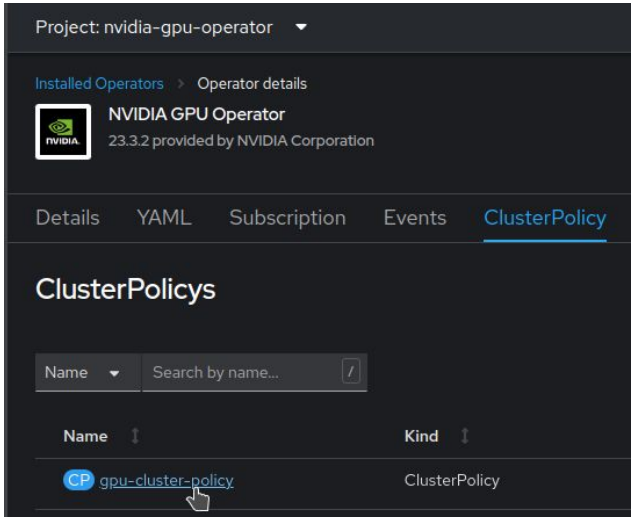
GPU sharings across teams on **OpenShift AI**

Leverage NVIDIA time-slicing and MIG using **NVIDIA GPU Operator**

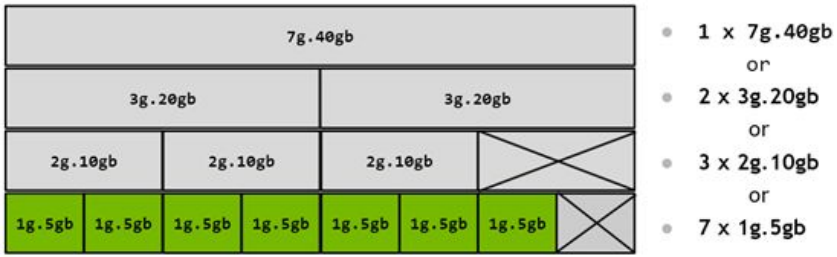
NVIDIA GPU OPERATOR

Time-slicing





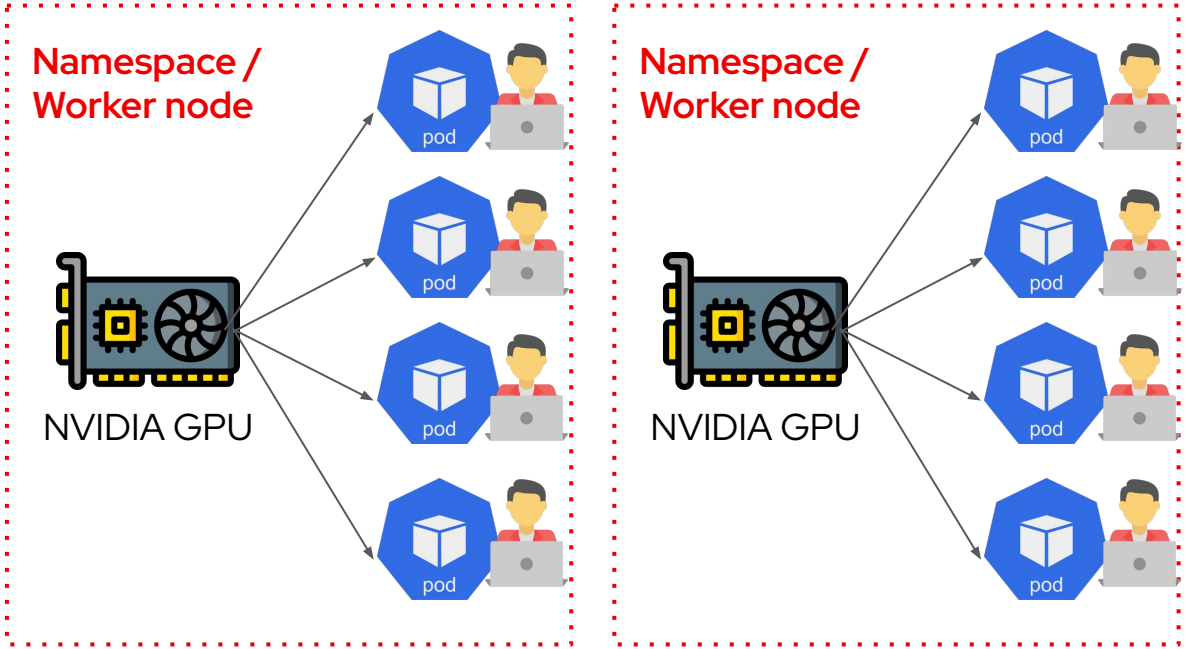
Multi-instance GPU (MIG)

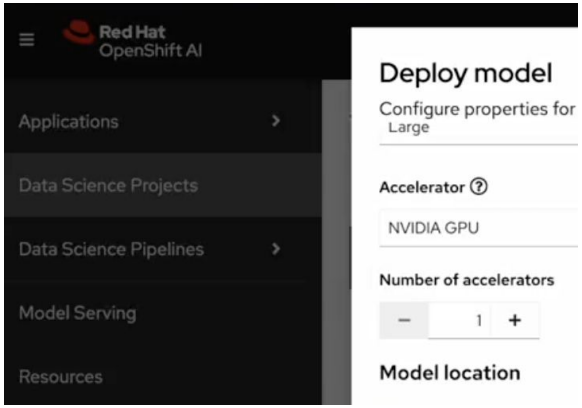


- 1 x 7g.40gb
  or
- 2 x 3g.20gb
  or
- 3 x 2g.10gb
  or
- 7 x 1g.5gb

Red Hat

Results of GPU sharing on OpenShift AI





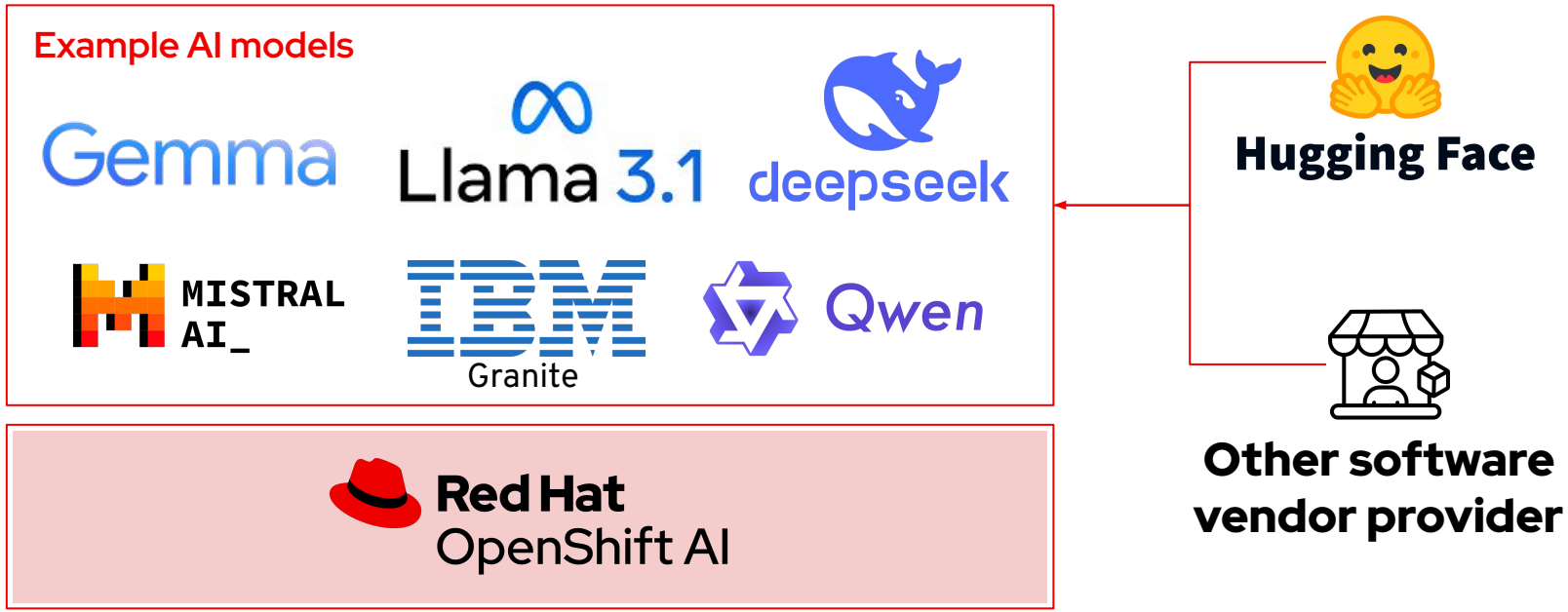
Red Hat

Use case #3: Diverse models support

Red Hat

Centralize AI model deployment on OpenShift AI

Bring whatever AI models to the platform



Note: Hugging Face and 3rd-party models are provided and supported by 3rd-party model providers

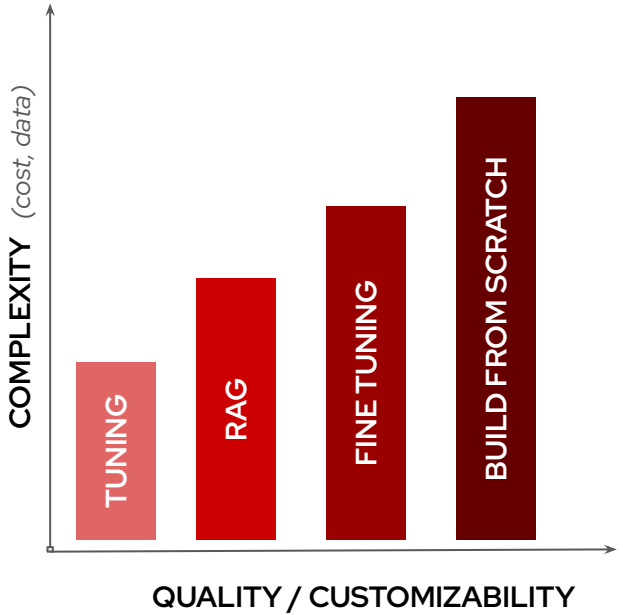
Red Hat

Use case #4: Integrate LLM with your knowledge

Red Hat

How to build and extend AI for your own business?

Pick the right way based on cost and customizability

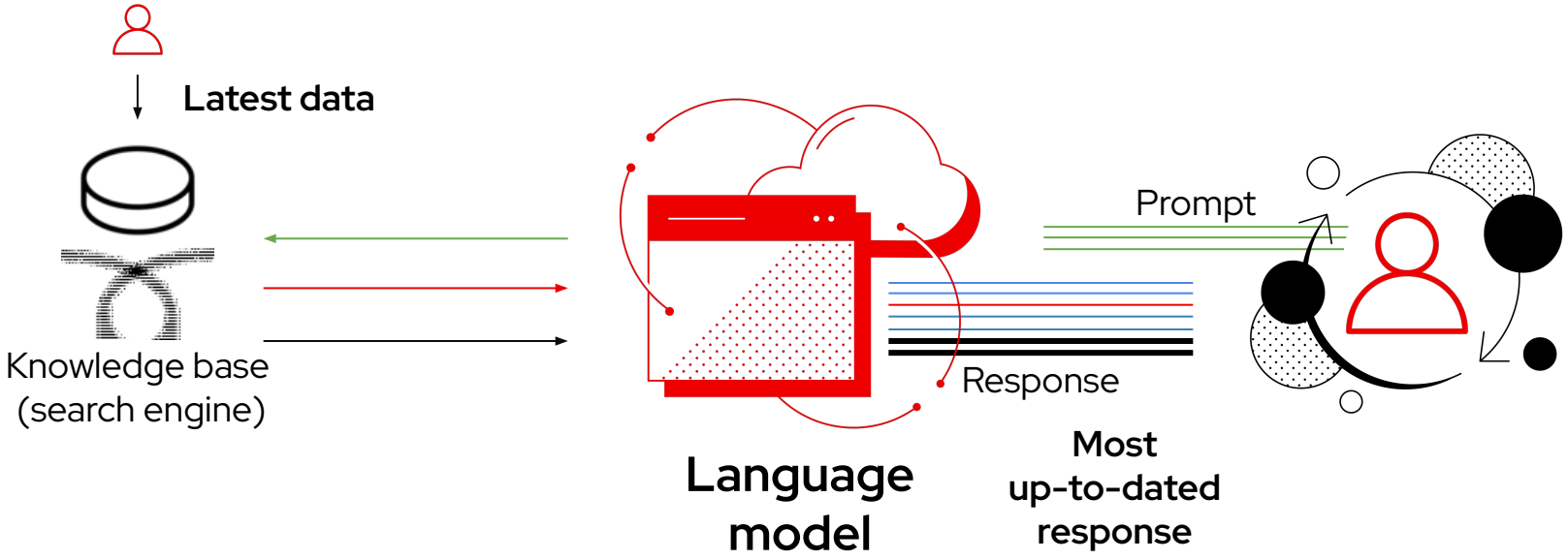


- **Prompt tuning:** Prepend prompt with data.
- **Retrieval augmented generation (RAG):** return facts from external sources.
- **Fine tuning:** Retrain small part of the model with your data.
- **Training from scratch:** only make sense for predictive AI.

Red Hat

Retrieval-augmented generation (RAG)

Augment knowledge base with new data



Use case #5: Fine-tuning small language models

Red Hat

IBM Granite

What is **IBM Granite 3.1**?

Pre-trained foundation model **provided and co-supported by IBM and Red Hat**

TRUE open sourced AI model

- **8B chat model, up to 32B code** SLM
- Apache 2.0 license
- Open weights on Hugging Face[1]
- **Open training data set**[2]

Multi-lingual support

- **English**, German, Spanish, French, Japanese, Portuguese, Arabic, Czech, Italian, Korean, Dutch and **Chinese**

Designed for Enterprise

- Up to **128K token context length**
- Language, code, agentic AI tasks……

IP indemnification assurance

- **Protect customers from legal liability** if Granite infringes on others' IP

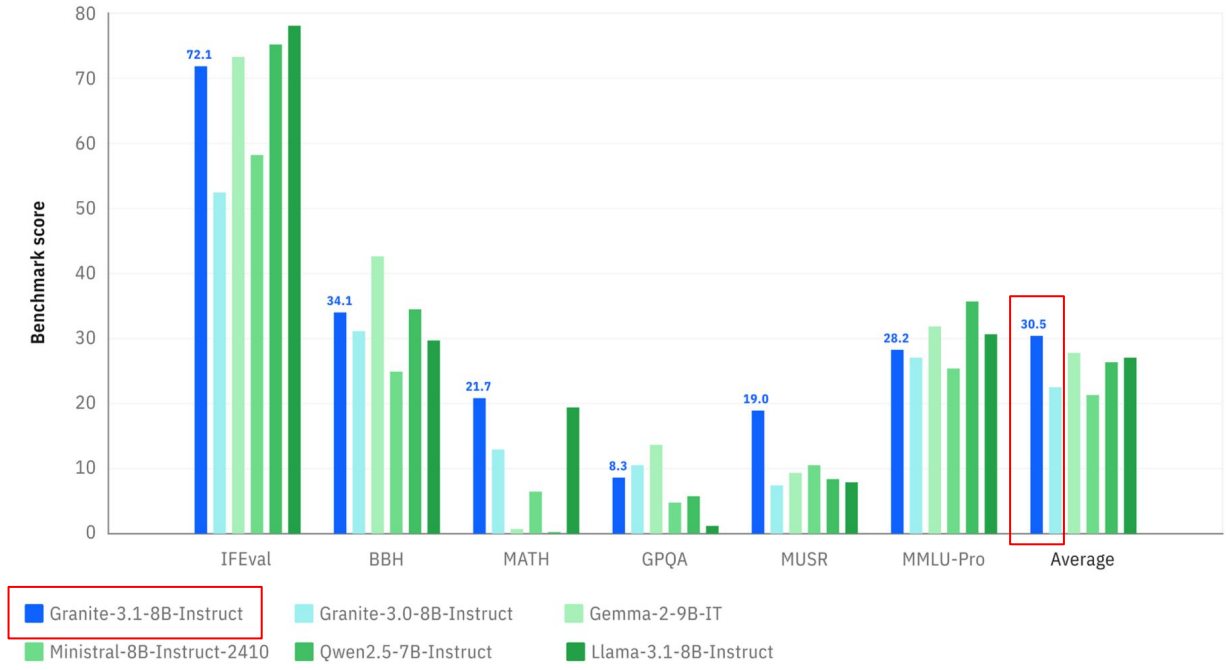
[1] https://huggingface.co/ibm-granite/granite-3.1-8b-instruct

[2] https://github.com/ibm-granite/granite-3.0-language-models/blob/main/paper.pdf

Red Hat

Granite can even **outperform** SOTA[1] language model

Hugging Face OpenLLM Leaderboard benchmarks within the same class of model size



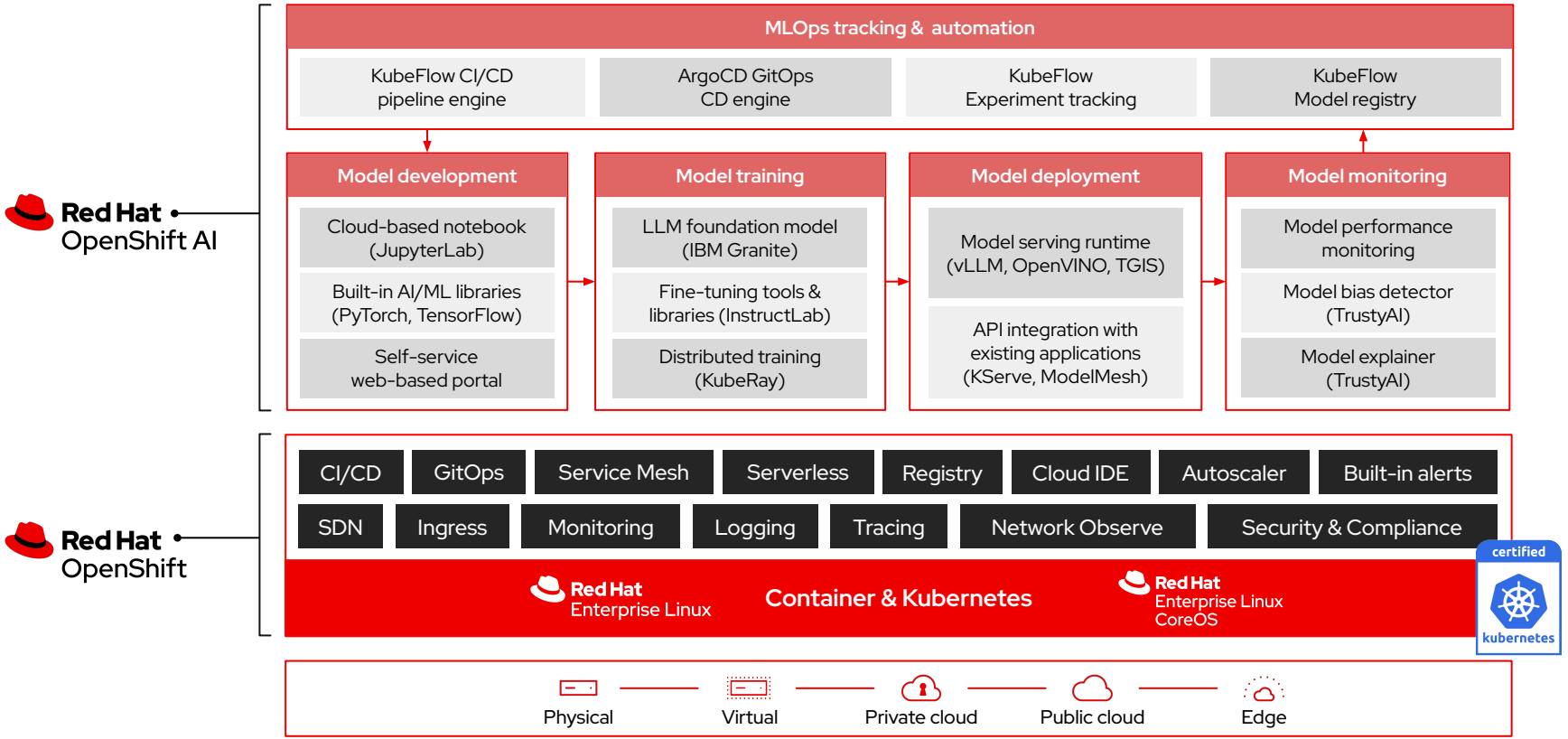
[1] State-of-the-Art

Red Hat

Use case #6: Streamline AI/ML with MLOps, across clouds

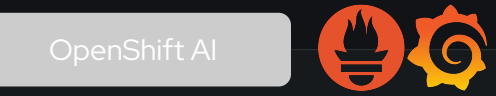
Red Hat

OpenShift AI: the **MLOps** platform for **data scientists** and **operation teams** across **hybrid clouds**

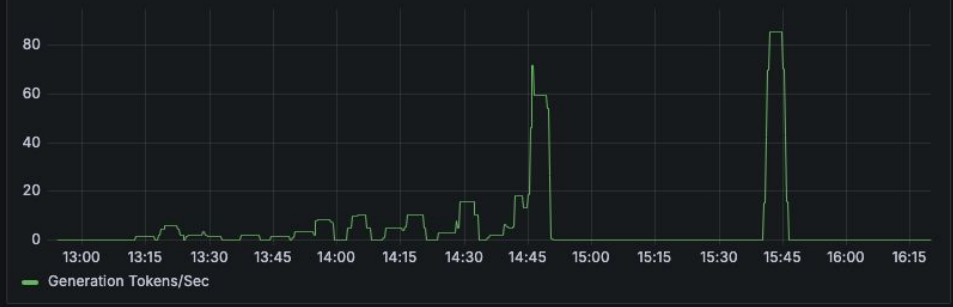


datasource Prometheus model_name granite-3b

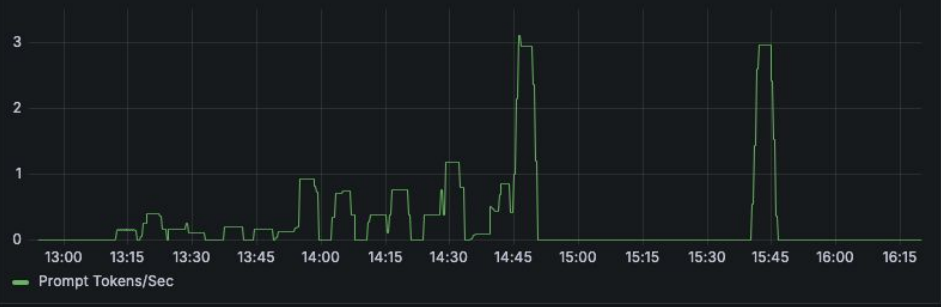
OpenShift AI



vLLM Generation Token Throughput

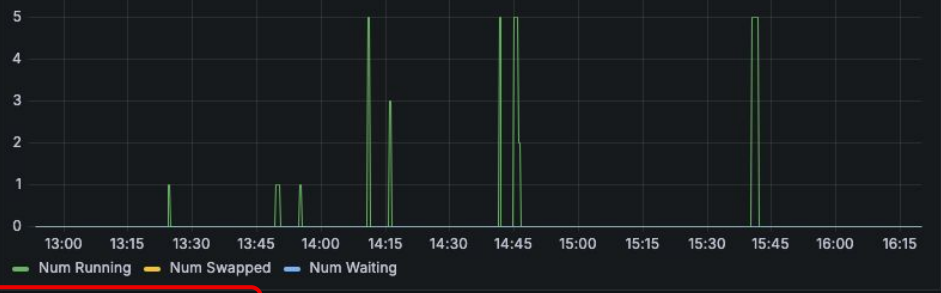


vLLM Prompt Token Throughput

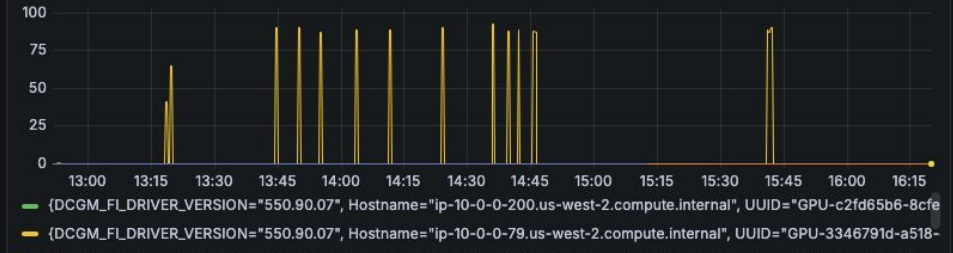


E2E Request Latency

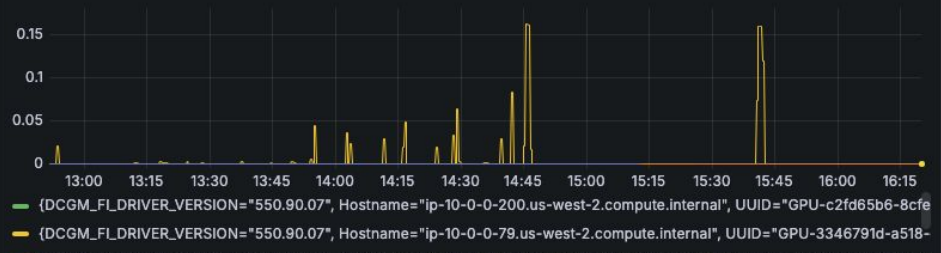
Scheduler State



GPU Utilization



GPU Tensor Core Utilization




Request Prompt Length
Prompt Length
Inf
2000
1000
500
200
100
50
20
10
5
2
1
0
13:00 13:15 13:30 13:45 14:00 14:15 14:30 14:45 15:00 15:15 15:30 15:45 16:00 16:15
0
60
Request Generation Length
Generation Length
Inf
2000
1000
500
200
100
50
20
10
5
2
1
0
13:00 13:15 13:30 13:45 14:00 14:15 14:30 14:45 15:00 15:15 15:30 15:45 16:00 16:15
0
24.4
OpenShift AI
Successfully Processed Requests
80
60
40
20
0
13:00 13:15 13:30 13:45 14:00 14:15 14:30 14:45 15:00 15:15 15:30 15:45 16:00 16:15
length
stop
Cache Utilization
8%
6%
4%
2%
0%
13:00 13:15 13:30 13:45 14:00 14:15 14:30 14:45 15:00 15:15 15:30 15:45 16:00 16:15
GPU Cache Usage
CPU Cache Usage
Time Per Output Token Latency
10 ms
9 ms
8 ms
7 ms
6 ms
5 ms
13:00 13:15 13:30 13:45 14:00 14:15 14:30 14:45 15:00 15:15 15:30 15:45 16:00 16:15
P99
P95
P90
P50
Time To First Token Latency
1.5 ms
1 ms
500 µs
0 s
13:00 13:15 13:30 13:45 14:00 14:15 14:30 14:45 15:00 15:15 15:30 15:45 16:00 16:15
P99

Values Red Hat can bring in AI

Red Hat

Provide governance

- **Central place** to deploy and use AI models
- **Standardize outputs** using approved AI models

Reduce privacy and legal risk

- Run AI models **privately**, **no data is ever leaked out**
- **IP indemnification** assurance from Red Hat

Increase flexibility and control

- **Tools to allow you shaping** your own AI for business
- **Free to choose** Red Hat-provided or 3rd party models

Reduce cost

- **Run SLM over LLM**, scale AI models horizontally
- Provide AI model **inference distribution and optimization**