Paint It Green!

How Open Source Helps Reduce the Environmental Impact of AI

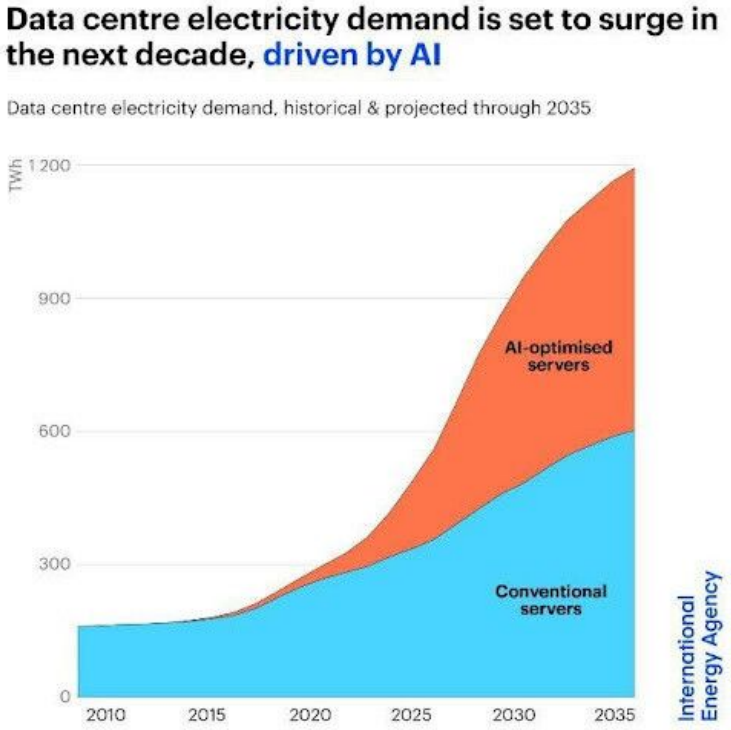
Danijel Soldo

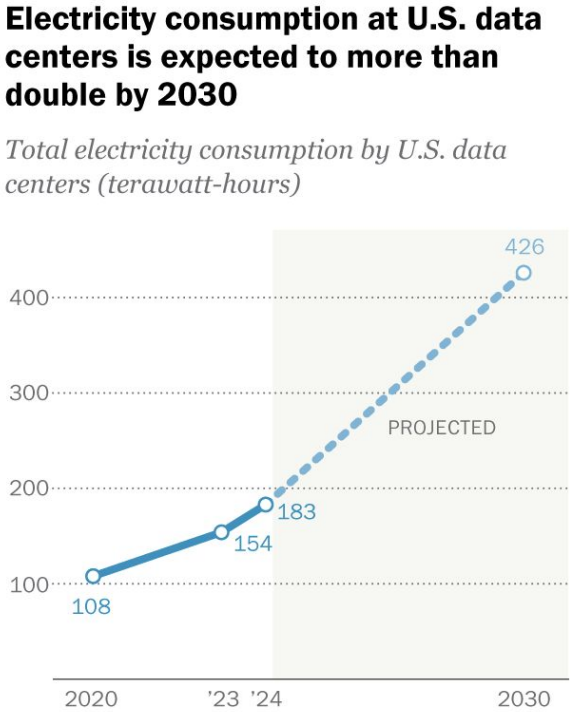
Solution Architect @ Red Hat

danijel@redhat.com

Red Hat

Impact #1: Electricity / Carbon emissions





Red Hat

Impact #2: Water

Infrastructure > Data Centres

Google data centre soaks up a third of Oregon city's water supply

News By Bobby Hellard published December 22, 2022

The tech giant has been labelled a "water vampire" after its facility increased water consumption every year since opening for the purposes of cooling

A single data center can consume up to 5 million gallons of drinking water a day, enough to supply thousands of households or farms.

https://utulsa.edu/news/data-centers-draining-resources-in-water-stressed-communities/

“A single 100-word email generated by ChatGPT costs 3 bottles of water.”

https://www.washingtonpost.com/technology/2024/09/18/energy-ai-use-electricity-water-data-centers/

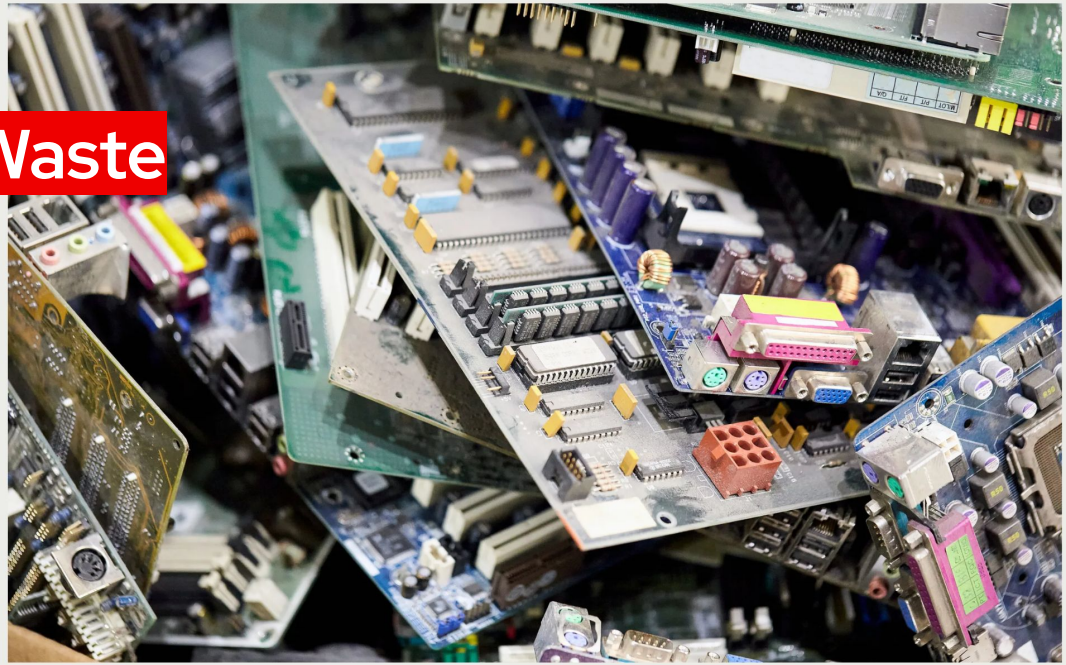
Red Hat

Impact #3: E-Waste

Generative AI Has a Massive E-Waste Problem › Rapid growth could result in an annual e-waste stream of 2.5 million tonnes by 2030

BY KATHERINE BOURZAC | 04 NOV 2024 | 3 MIN READ |

Katherine Bourzac is a freelance journalist based in San Francisco, Calif.



MINDFUL MEDIA/ISTOCK

Red Hat


Apocalypse Now?

Cut it ...

And let's get to action!

It’s now estimated that 80–90% of computing power for AI is used for **inference**.

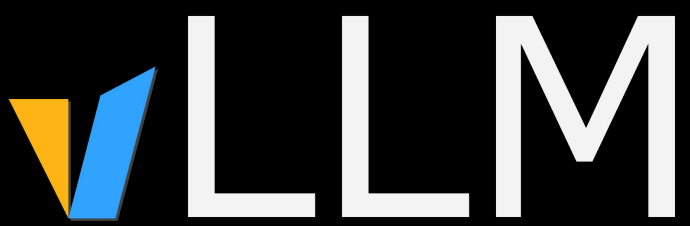
MIT Technology Review

Red Hat

Open Source Projects

for more efficient **AI Inferencing**

#1



Easy, fast, and cheap LLM serving for everyone

Octoverse 2025

Top Open Source Projects by contributors

Rank	Repository	Short description
1	vllm-project/vllm	High-throughput LLM inference engine
2	microsoft/vscode	Widely used open source code editor
3	openai/codex	Lightweight coding agent that runs in the terminal
4	huggingface/transformers	Core library for model loading & fine-tuning
5	godotengine/godot	Game engine for 2D/3D development

https://github.blog/news-insights/octoverse/octoverse-a-new-developer-joins-github-every-second-as-ai-leads-typescript-to-1/

About vLLM

The de-facto standard in open source model serving

PyTorch Foundation

Llama Granite Gemma Qwen DeepSeek Mistral

vLLM

Fast and easy to use open source inference server

NVIDIA GPU AMD Instinct intel Gaudi Google TPU aws Neuron IBM AIU

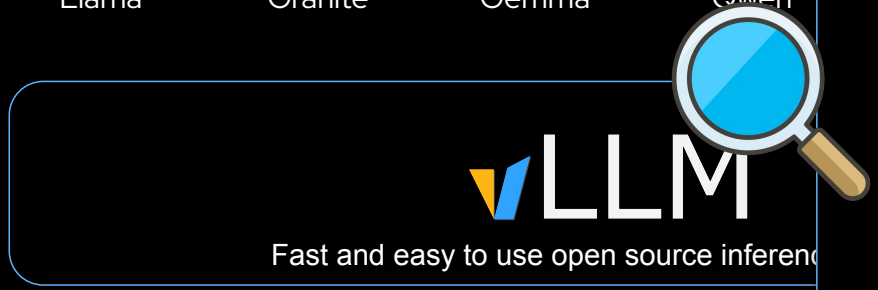
Red Hat

About vLLM

The de-facto standard in open source model serving

PyTorch Foundation

Llama Granite Gemma Qwen



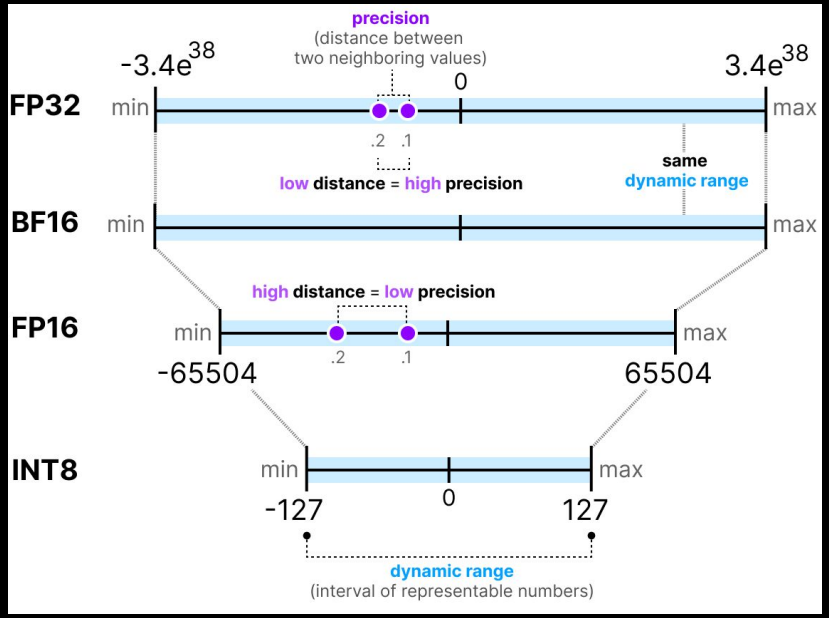
Optimizations!

- Paged Attention
- Advanced KV-Cache
- Prefix caching
- Spec Decode
- Chunked Pre-fill
- Multi-LoRA
- **Quantization**
- **Distributed Inference**

NVIDIA GPU AMD Instinct intel Gaudi Google TPU Neuron IBM AIU

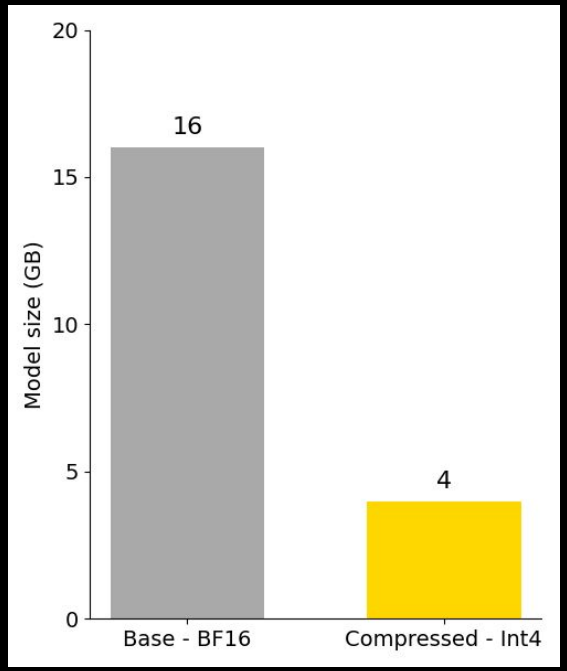
Red Hat

Quantization aims to reduce the precision of a model's weights from high to low precision formats (e.g. FP32 to INT8 / FP8) without dropping model quality.



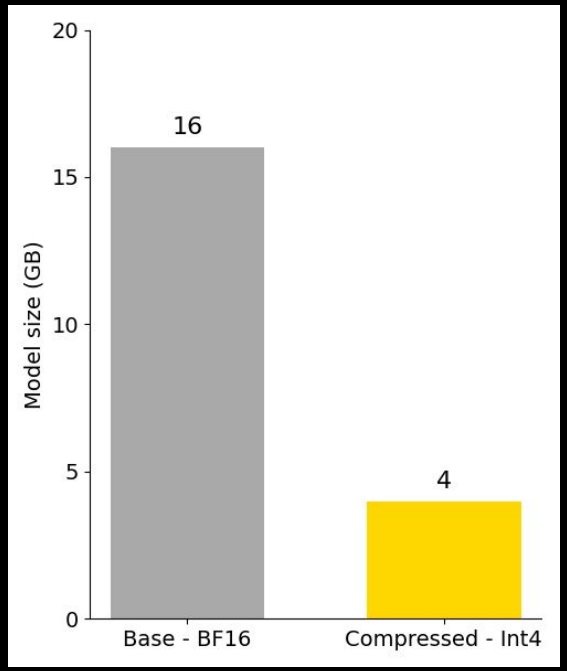
BF16 FP8 W8A8 INT W8A8 INT W4A16

LLM Compressor



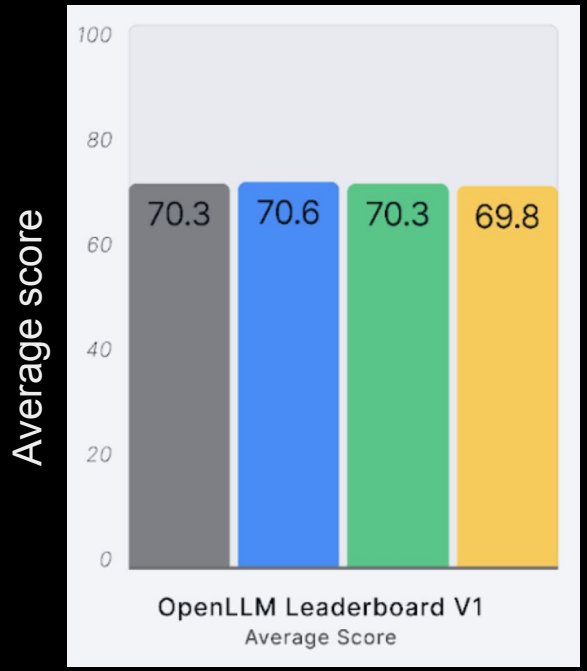
Model size

LLM Compressor



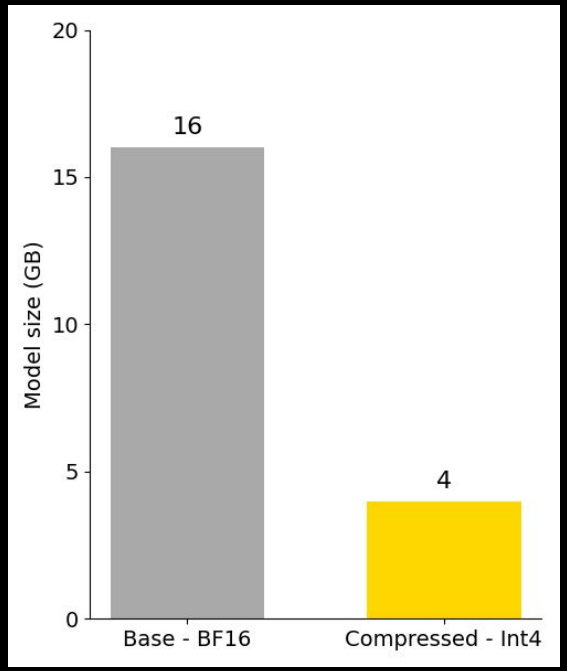
Model size

lm-evaluation-harness



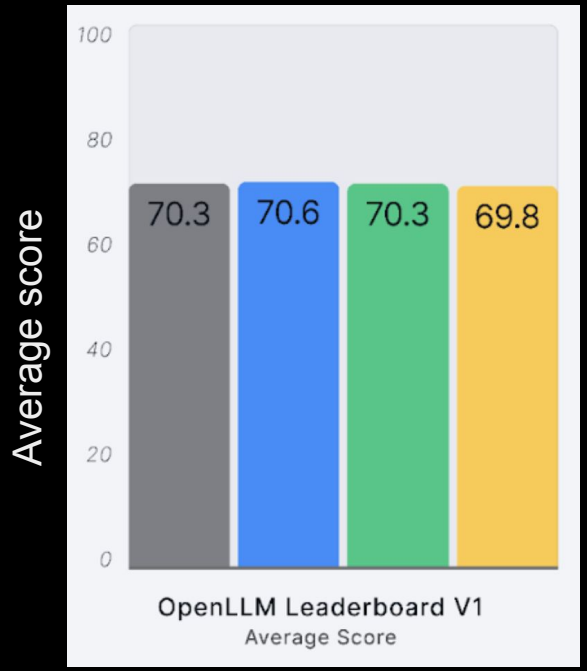
Accuracy

LLM Compressor



Model size

lm-evaluation-harness

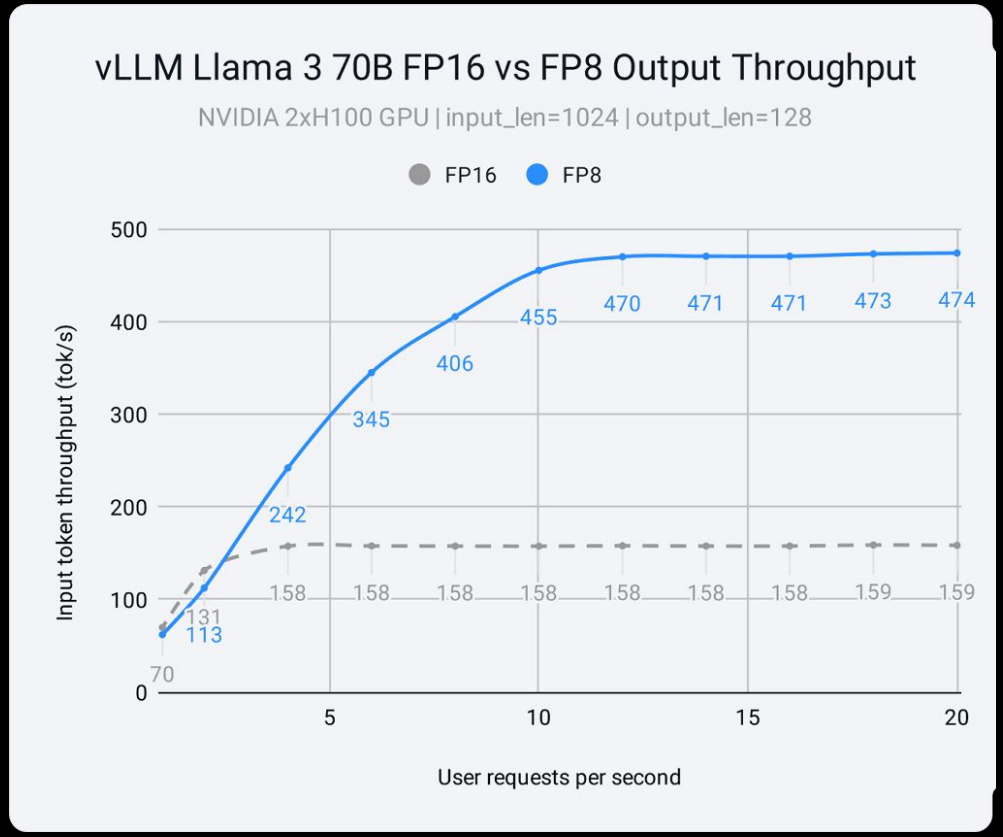


Accuracy

vLLM

Inference

Quantization enables more tokens for fixed hardware



Get Started with Quantization in vLLM

Validated models by Red Hat AI

→ *red.ht/optimized-models*

LLM Compressor

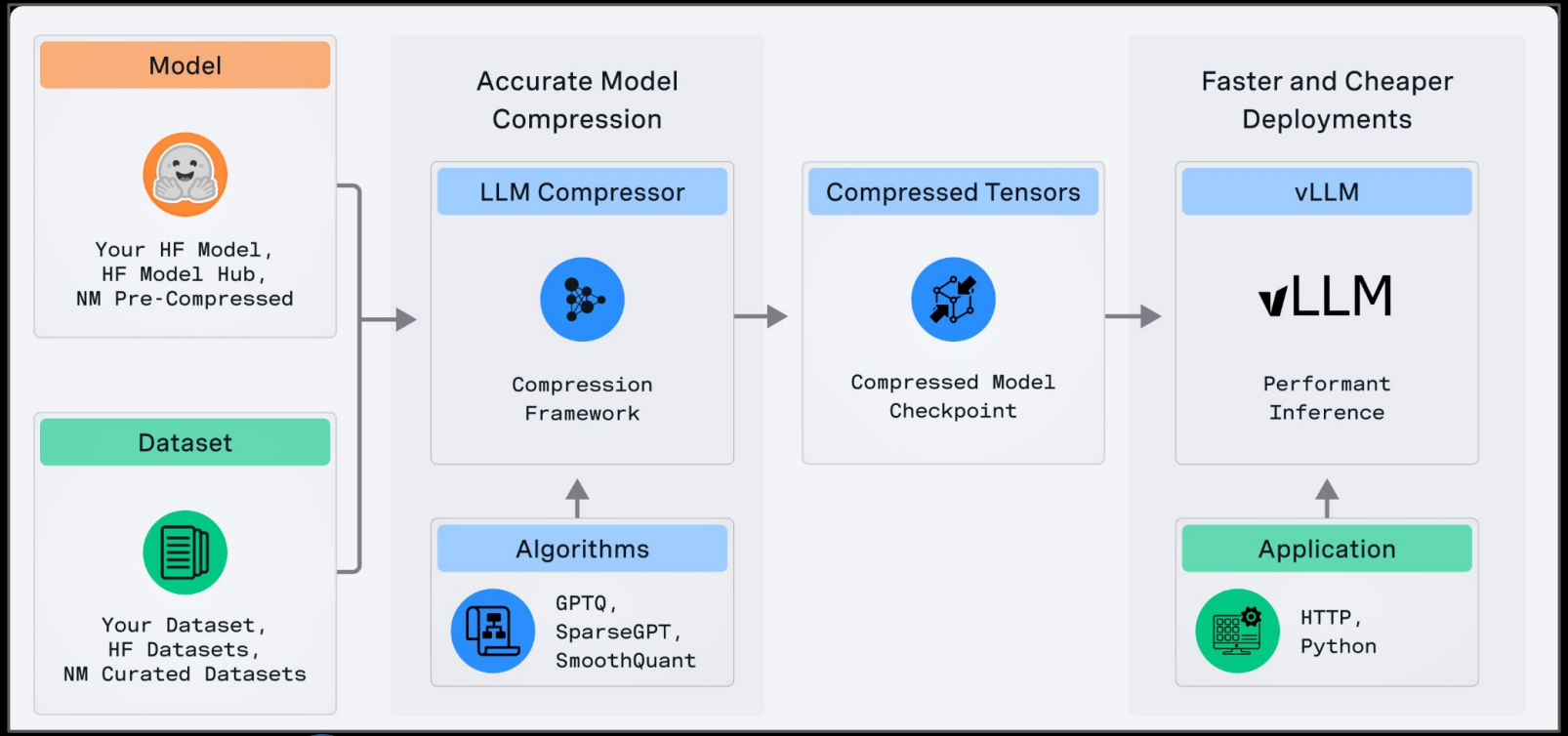
→ *red.ht/llm-compressor*

#2

LLM Compressor

easy-to-use library for optimizing models for deployment with vLLM

LLM Compressor

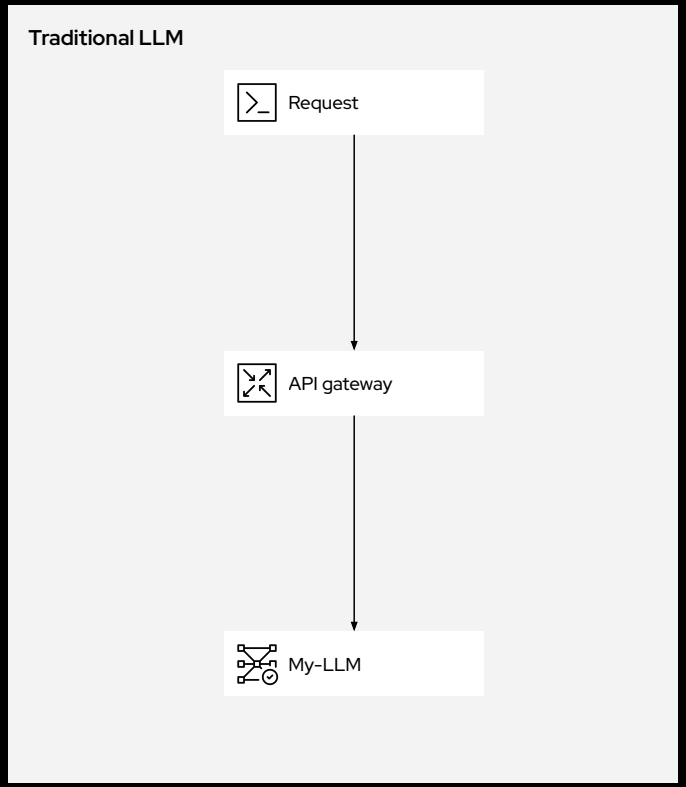


https://github.com/vllm-project/llm-compressor

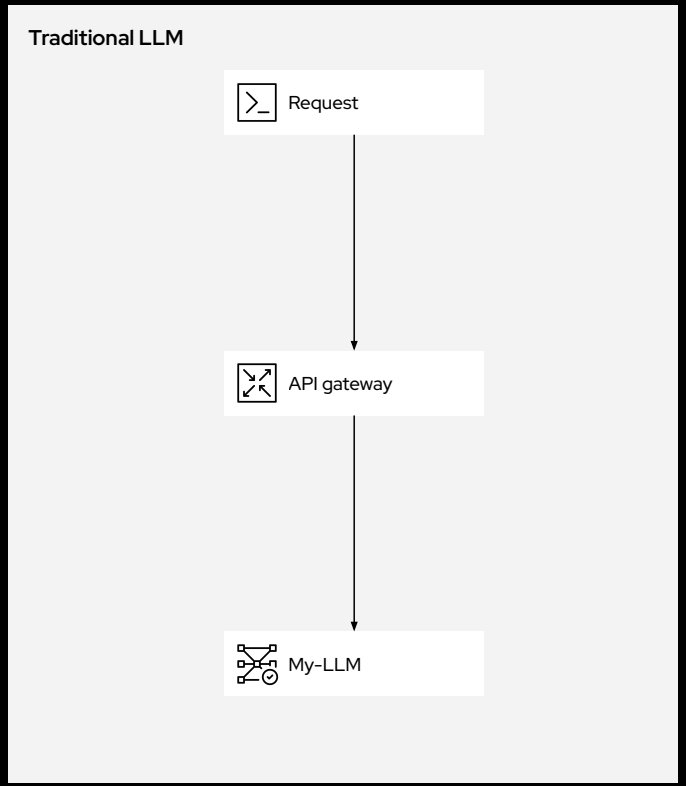
Red Hat

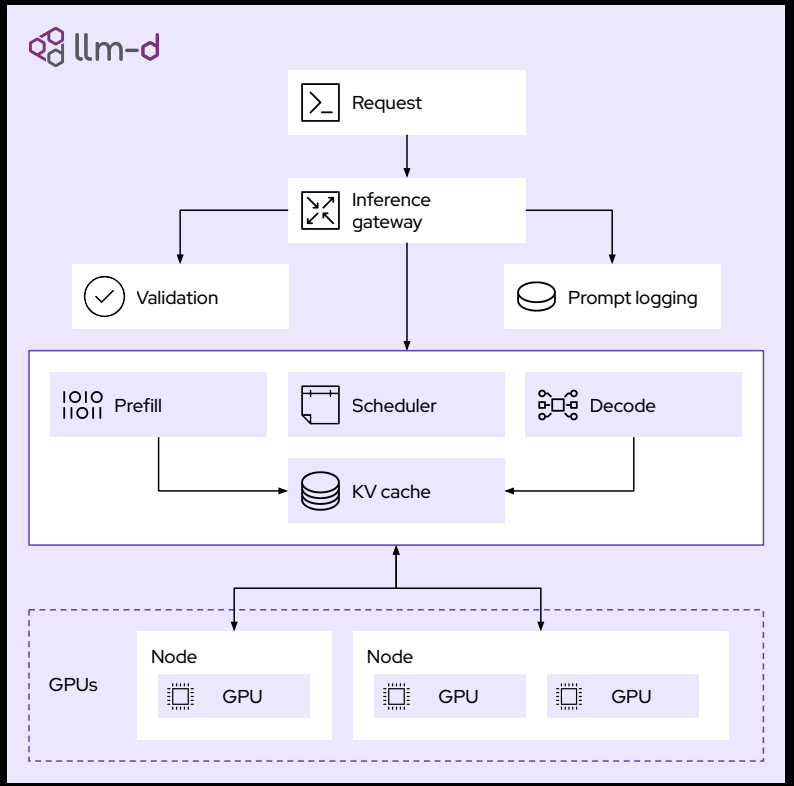
- **For practitioners:** off-the-shelf validated model for use with vLLM
- **For developers:** easy to use recipes with heavily optimized default configurations for all quantization formats supported in vLLM
- **For researchers:** fine-grained control over model quantization (per-channel, per-tensor, per-group), symmetric/asymmetric, with or without calibration data, activation reordering, dynamic/static activation quantization, etc.

#3 llm-d



Mostly Black Box

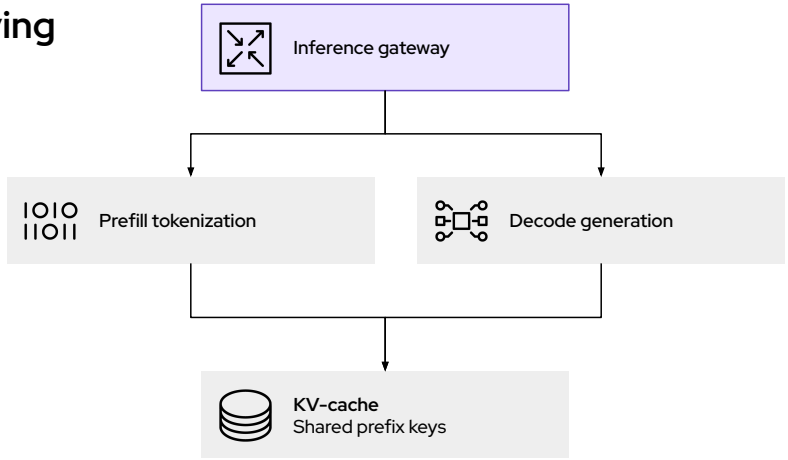




From Black Box to First-Class Citizen

Key Innovation: Distributed Model Serving

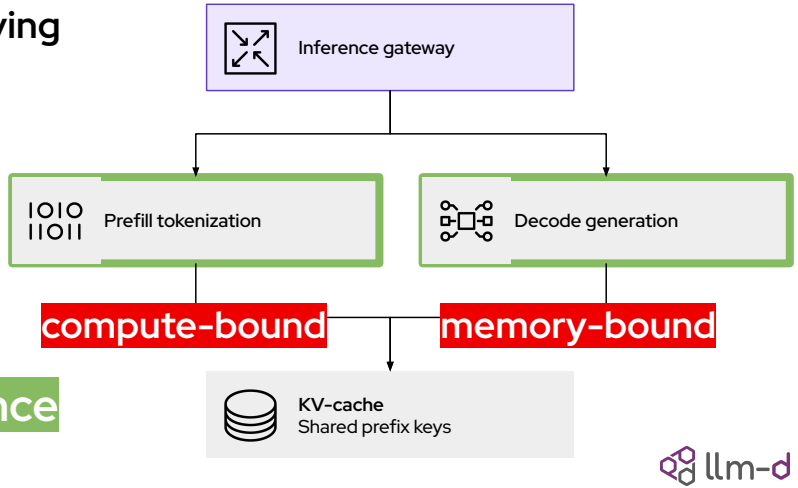
llm-d includes a special load scheduler that ensures each request is routed to the correct model server, built using Kubernetes' Gateway API inference extension



Key Innovation: Distributed Model Serving

llm-d includes a special load scheduler that ensures each request is routed to the correct model server, built using Kubernetes' Gateway API inference extension

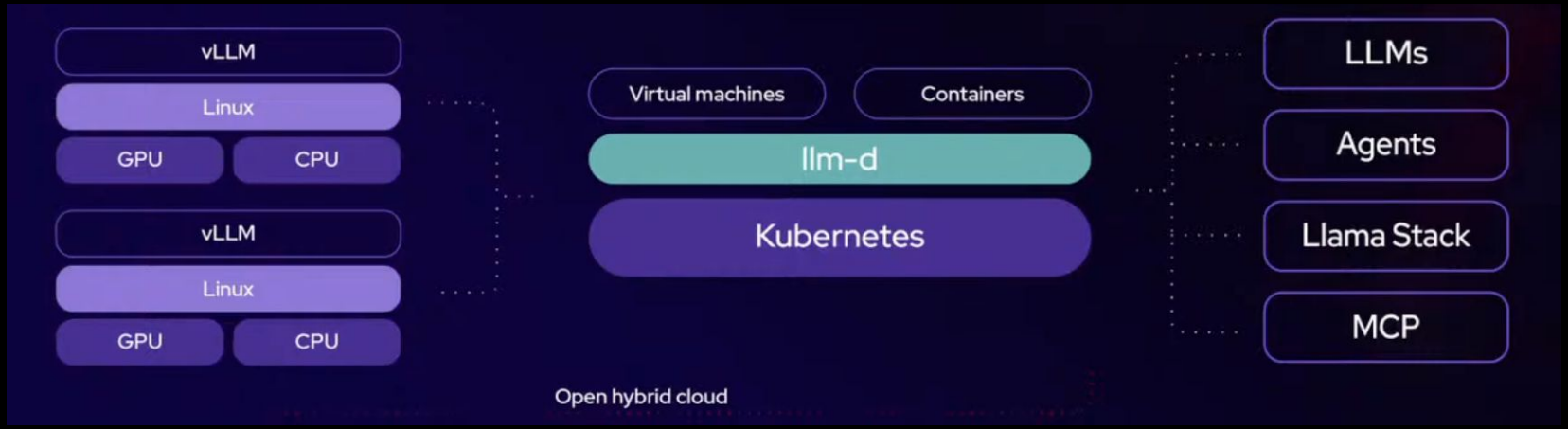
Disaggregation - a big chance for cost optimization!



Llm-d performance

llm-d

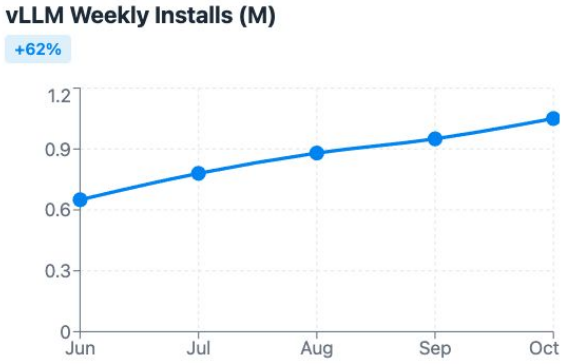
A Kubernetes-native, high-performance distributed LLM Inference framework

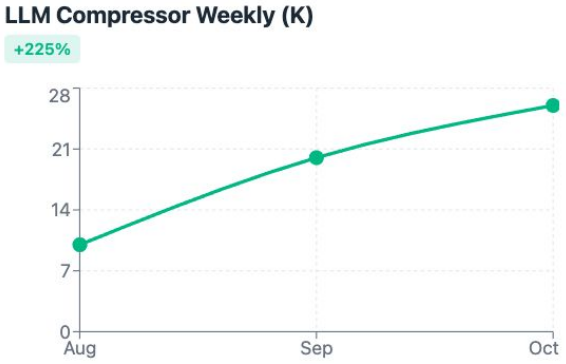


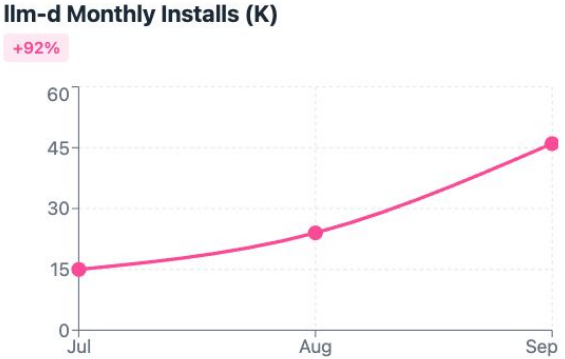
AMD cisco CoreWeave Google Hugging Face IBM intel Lambda Mistral AI NVIDIA

Red Hat

The momentum is real







#4 Kepler

Kubernetes Efficient Power Level Exporter

Project Kepler - in a nutshell

- CNCF Sandbox Project
- Prometheus exporter for energy metrics
- Multi-level attribution:

  Node → Pod → Container → VM → Process
- Hardware sensor-based (Intel RAPL, Redfish ...)
- Kepler Operator available

Kepler Major Rewrite in : v0.10.0

faster, safer, and cleaner

- Enhanced Performance & Accuracy
  - Dynamic RAPL zone detection:
  - Smarter power attribution
  - Better environment detection
  - Lean & efficient: Uses far fewer resources than the old Kepler version.

- Reduced Security Requirements
  - Read-only access only: Needs access to /proc and /sys, nothing else.
  - No privileged capabilities
  - Much safer footprint

- Modernized architecture & Service-oriented design

Current limitations:

- Only bare-metal, GPU & platform to come
- Only RAPL/powercap framework

Wrap Up: Resources

- https://github.com/vllm-project/vllm
- https://github.com/vllm-project/llm-compressor
- https://github.com/llm-d/llm-d
- https://github.com/sustainable-computing-io/kepler

Open Source

=

Open for

Community
Collaboration
Action!

Thank You.

Danijel Soldo
danijel@redhat.com