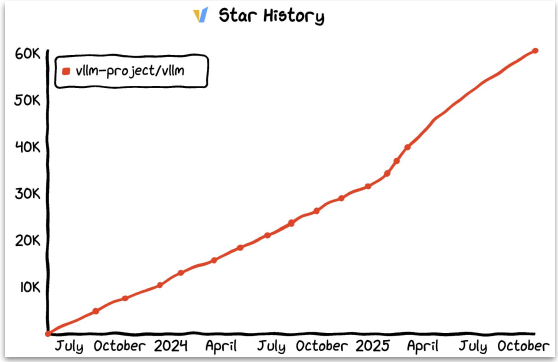
vLLM Project Update

Tyler Michael Smith
Core Maintainer, vLLM &
Chief Architect, Inference Engineering @ Red Hat

What is vLLM?

The **High-Throughput** and **Memory-Efficient** open-source serving engine for LLMs



Most Popular LLM Serving Engine

- 70K+ GitHub stars, 800+ PRs/month
- 500K++ GPUs deployed 24/7
- 10K+ members in slack.vllm.ai

>2000 Contributors from >50 major companies

AMD, IBM, Google, amazon, ORACLE, NVIDIA, CoreWeave, THINKING MACHINES, ROBLOX, intel., AI21, Anyscale, embedded, arm, Tencent

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

```
$ uv pip install vllm --torch-backend=auto
$ vllm serve deepseek-ai/DeepSeek-V3.1 -tp 8
```

Broad Model Support (>100 arches)

Llama, DeepSeek, Qwen, GPT-OSS, Gemma, Mistral, Phi, Granite, Cohere

Flexible Device Parallelism

Tensor, Pipeline, Expert, Data, Context Parallel
Disagg Prefill/Decode, Disagg Encoder

Wide Hardware Support

NVIDIA CUDA, AMD ROCm, intel. Gaudi/XPU, Google TPU, aws Neuron, Ascend, CPU, Metal, METAX MACA, rebellions_ RBLN, IBM Spyre, Cambricon MLU, Baidu Kunlun

Diverse Project Ecosystem

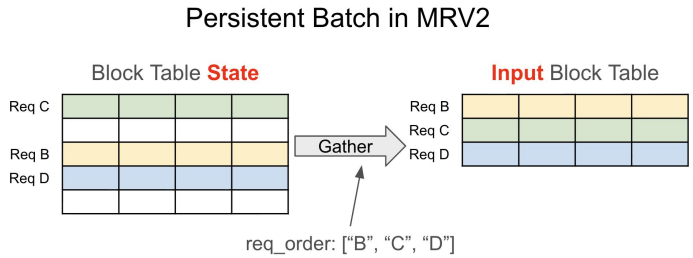
llm-d, vLLM-Omni, GuideLLM, vLLM Semantic Router, LLM Compressor, PECULATORS, Production Stack

vLLM

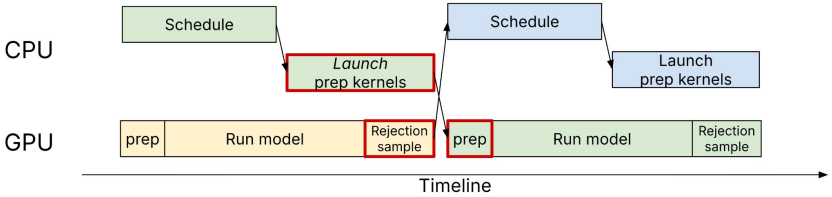
What's next: Model Runner V2

Ground-up Reimplemented Execution Core: Cleaner. More Efficient. More Modular.

Decoupled Persistent Batching Design



Async-First Design - No CPU↔GPU Syncs



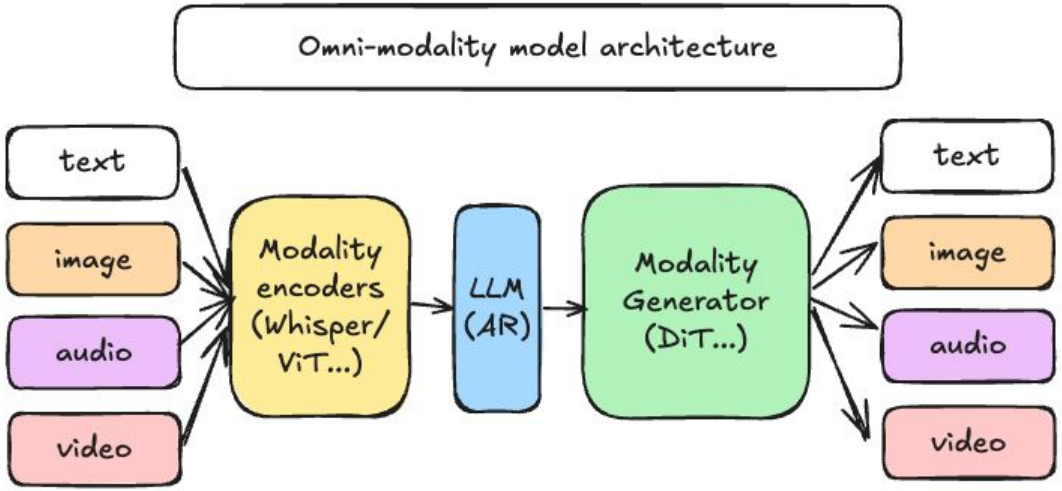
Better Performance For High Concurrency

- MRV1 vs MRV2 on Qwen3-0.6B with 1×GB200

vLLM-Omni

Easy, Fast, and Cheap Omni-Modality Model Serving

- **Omni-modal Input**
  - Embed audio, image, video input into latent vectors
- **Omni-modal Output**
  - Diffusion models (DiT) convert latent vectors into multimodal output
- **Autoregressive Inference**
  - Leverage vLLM for autoregressive transformer inference




```
$ vllm serve Qwen/Qwen-Image-2512 --omni --port 8091
```

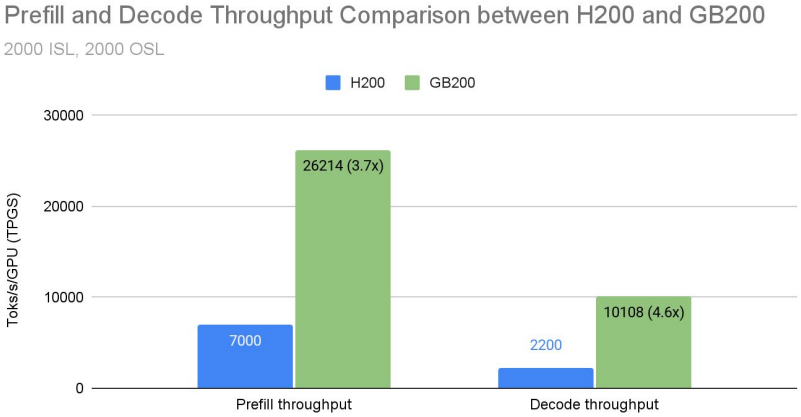
```
$ curl -X POST http://localhost:8091/v1/images/generations \
  -H "Content-Type: application/json" \
  -d '{"prompt": "a dragon laying over the spine of the Green Mountains of Vermont", "size": "1024x1024",}' | jq -r '.data[0].b64_json' | base64 -d > dragon.png
```

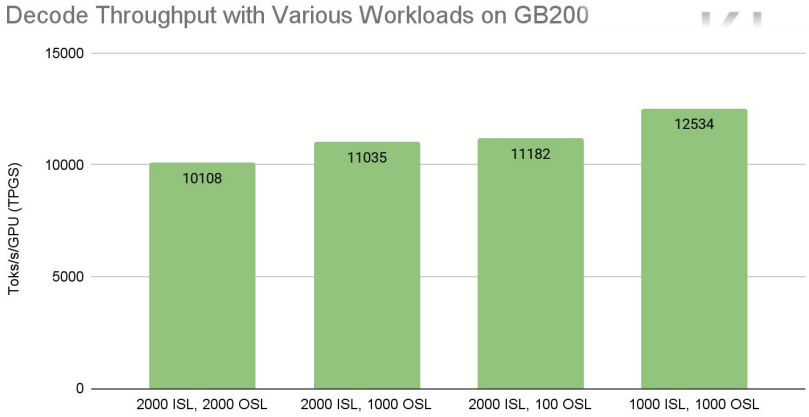

vLLM

Wide Expert Parallelism on H200 and GB200

Reached 26.2k tokens/s per GPU on P and 10.1K tokens/s per GPU on D

deepseek R1





NVFP4 GEMM & MoE Dispatch, FP8 MLA GEMM, kernel fusions, MoE DP Chunk and the addition of fast EPLB + DBO for decode deliver compound gains.

PyTorch CONFERENCE EUROPE 2026

vLLM

vLLM Q2 Roadmap

See roadmap.vllm.ai for more information

- #sig-core
  - MRV2 hardening toward default
  - KV cache manager rethink
  - Address scheduler preemption & HoL blocking
- #sig-ci
  - Continue regular release cadence!
  - Time-to-signal → 30 min; model eval × hardware matrix; automatic test target determination
- #sig-model-performance
  - +GB200, B300, MI355
  - Data-driven kernel SoL tracking
  - Revamped accuracy CI
- #sig-omni
  - Disaggregated large-scale serving for omni & diffusion models
- #sig-quantization
  - Online quantization refactor
  - Deterministic dispatch oracle
  - W1–8 low-bit kernel expansion
- #sig-large-scale-serving
  - Production-ready Elastic EP
  - Zero-cost async EPLB
  - Experimental fault-tolerant EP
- #sig-torch-compile
  - torch.compile integration hardening
  - Improvements to vLLM CustomOps
- #sig-multimodality
  - ViT CUDA graph + torch.compile on by default
  - More flexible encoder API

Get Involved in vLLM

2000+ contributors • 50+ companies • 500K+ GPUs deployed 24/7

Try it now

```
$ uv pip install vllm --torch-backend=auto
```

Connect

- slack.vllm.ai — 10K+ members, find your SIG
- github.com/vllm-project/vllm — 800+ PRs/mo
- vllm.ai/events — Weekly SIG meetings
- roadmap.vllm.ai — Full Q2 roadmap will be posted