What is llm-d

llm-d is a high-performance distributed inference serving stack optimized for production deployments on Kubernetes. Its key feature includes：

- Intelligent Routing
- Advanced KV-Cache Management
- Serving Large Models
- Operational Excellence
- Batch Processing

Why Alternative AI Accelearators using llm-d

llm-d slogan: Achieve SOTA Inference Performance On Any Accelerator

It already supports:

- AMD ROCm
- CPU x86_64
- Google TPU
- Intel XPU
- NVIDIA GPU
- Rebellions NPU

For new vendor it provides:

- Good abstraction
- Comprehensive adaptation guidelines
- Clear compatibility hierarchy

And it requests:

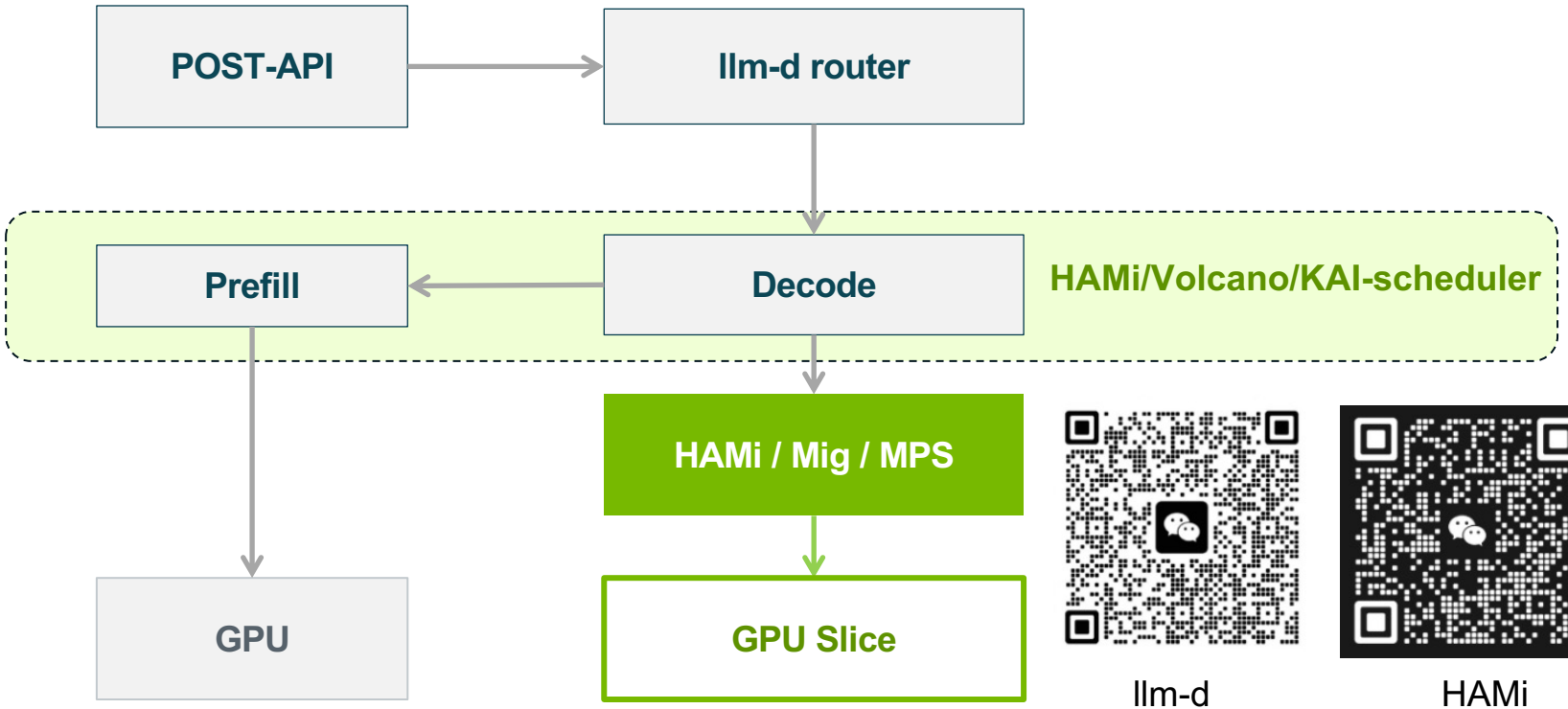
- Public inference engine image
- Open Dockerfile
- Vendor native inference optimization solution
- Contact information for vendor maintainers
- Nightly e2e environment for all realease

It provides diversity, standrad, and reliability for both vendors and users.

vLLM PD Disaggreation Demo

1. Check the device available
2. Check the llm-d deployment
3. Deploy inference instances
4. Post infrence requests
5. Check the vllm log

Optimize LLM-D inference



llm-d

HAMi