Google Skills

vLLm for LLM inference

In the previous video, Don mentions that vLLM is an easy-to-use library for LLM inference and serving. It allows you to switch between GPUs and TPUs easily with minimal coding changes, using a dual-container approach within a single pod.

Because vLLM uses different base images for GPUs and TPUs, we run two separate containers:

- vLLM OpenAI image for GPUs
- vLLM TPU image for TPUs

If the appropriate accelerator (GPU or TPU) is present, the container's vLLM server starts. Otherwise, it sleeps. This ensures only the correct vLLM server is active based on the underlying hardware.