Google Skills

Addressing common challenges

You've learned how to provision TPUs and the different architectural components. If you haven't taken our 'Architecting an inference stack' course, we encourage you to take it in parallel with this one, as it includes foundational information and best practices for frameworks, models, and orchestrators that will not be covered in this course.

The two TPU-specific challenges that we'll cover are:

- Interoperability between accelerators and frameworks
- Performance and utilization tracking

Interoperability between accelerators

Deploying LLMs often involves a trade-off between price, performance, and availability. Traditionally, achieving the best balance meant separate deployments, complex management, and higher costs. But what if we could simplify this for you?

This is where vLLM comes in. You're likely familiar with it by now, but if you didn't take our inference course, watch the next video.