Google Skills

Guided tutorial

GKE Inference Quickstart

These capabilities are exciting, but constantly being told you need to use a new tool or technique (or risk falling behind) can be overwhelming. Selecting the right combination of models, servers, and hardware is a complex process.

That brings us to GKE Inference Quickstart (GIQ). Think of this as a database of continuously benchmarked and tested inference stack configurations optimized for price and performance. You simply tell it your needs, and it can:

Analyze performance and cost

Explore available configurations and filter them based on your performance and cost requirements by using the gcloud container ai profiles list command. To view the complete set of benchmarking data for a specific configuration, use the gcloud container ai profiles benchmarks list command. This lets you identify the most cost-effective hardware for your specific performance requirements.

Run your own benchmarks:

The provided configurations and performance data are based on benchmarks that use the ShareGPT dataset. Performance for your workloads might vary from this baseline. To measure your model's performance under various conditions, you can use the experimental inference-benchmark tool.

Deploy manifests

After your analysis, generate and deploy an optimized Kubernetes manifest. You can optionally enable optimizations for storage and autoscaling. You can deploy from the Google Cloud console or by using the kubectl apply command. Before you deploy, ensure you have sufficient accelerator quota for the selected GPUs or TPUs in your Google Cloud project.