Google Skills

Using profiling to understand your model's inference performance

Deploying LLMs for real-time applications means every millisecond of latency counts, and every dollar of serving cost matters. Your model might be giving great answers, but if it's too slow or expensive, its real-world value plummets. How do you pinpoint what's slowing your model down or driving up costs?

This is where profiling comes in. To help you diagnose and fix these issues, we want to introduce you to a powerful tool from our internal toolbox: **XProf.**

What is XProf?

The X stands for XLA compiler, which helps you run ML operations on TPUs, GPUs, and even CPUs. XProf was previously known as the ‘TensorFlow profiler,’, but it has evolved.

Now part of the **OpenXLA** project, XProf is a powerful, framework-agnostic profiling tool that works seamlessly with **JAX, PyTorch/XLA**, and **Keras** on Google Cloud. It’s the same tool we use at Google to optimize inference for products like Gemini, Search, and YouTube.

It gives you a detailed view of your model's execution, helping you understand, debug, and optimize your inference code to get the most out of your hardware.

How does XProf work?

Getting started with XProf involves two main steps: capturing a profile from your inference server and then visualizing it.

You can **capture a profile** in two ways:

- **Programmatically:** You can add a few lines of code to your inference server to start and stop the profiler. This allows you to capture a trace of specific events, like the processing of a batch of requests.
- **On-demand:** For moments when you notice high latency or unexpected behavior in your live server, you can manually trigger the profiler to capture data for a specific duration and diagnose the issue in real-time without interrupting service.

To make this process even smoother, we’ve introduced the Cloud Diagnostics XProf library. This tool helps you host TensorBoard on a VM or GKE pod, and easily share links to your profiles for team collaboration.

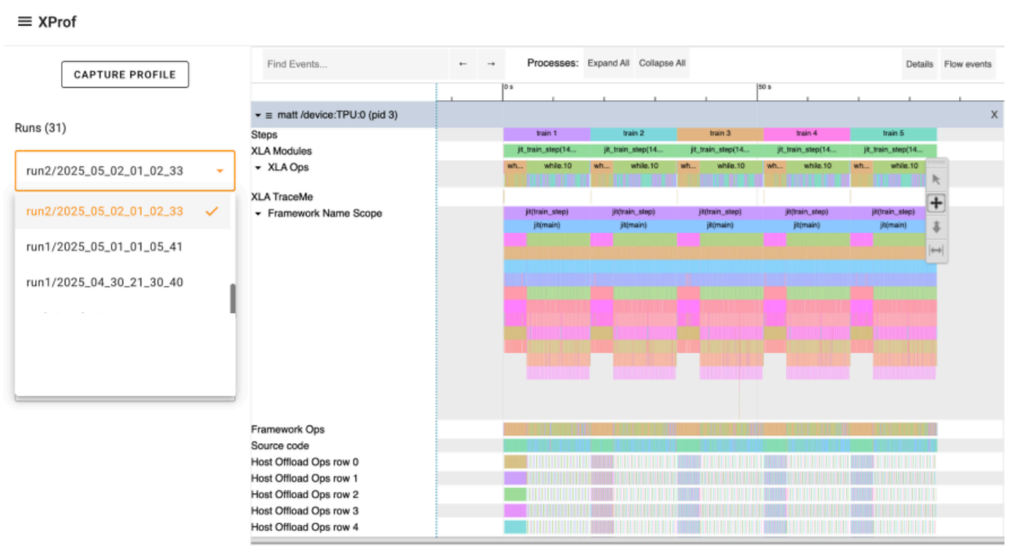
What can you see on XProf?

With XProf, you’re not just looking at raw numbers. A suite of powerful visualization tools tailored for debugging inference performance.

Trace Viewer
Get a detailed timeline of all the operations running on the host (CPU) and the accelerator (TPU/GPU) during an inference request. You can see precisely how long each part of the pipeline takes—from pre-processing to the model's forward pass, and post-processing—to identify latency bottlenecks.

Memory Viewer
Keep an eye on your memory usage over time. This tool helps you optimize memory allocation per request, allowing you to maximize batch sizes for higher throughput and prevent out-of-memory errors that can crash your server.

Graph Viewer
Visualize the computational graph of your model to understand how operations are connected and identify potential structural inefficiencies that can be optimized for faster inference.



These tools can help you pinpoint performance bottlenecks, understand hardware utilization, and ultimately, deliver a faster, more cost-efficient inference service.

Learn more: XProf profiler and Cloud Diagnostics XProf library.