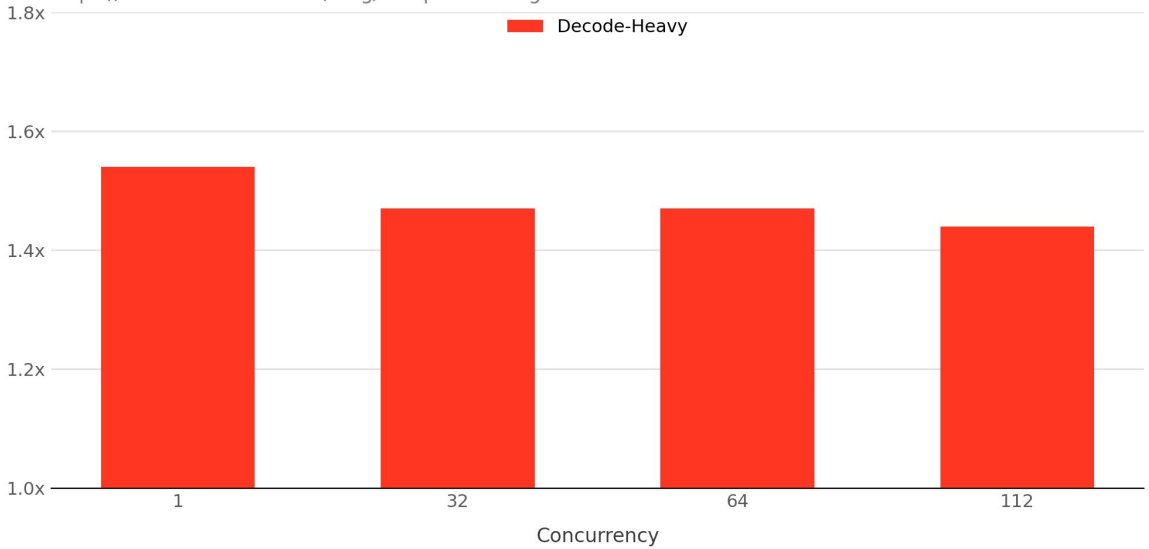
PEFT Serving Throughput Speedup

https://www.databricks.com/blog/fast-peft-serving-scale



Techniques:

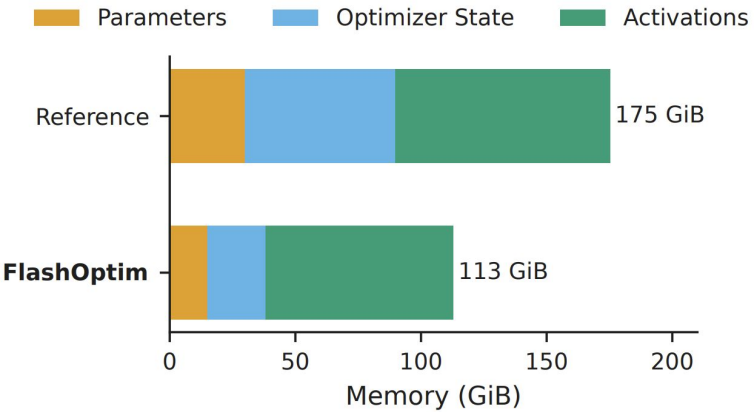
- Parallelized work
- Single cuda stream overlap
- Hybrid dtype attention
- Fused Hadamard Kernels
- Better CPU-GPU overlap

Daya Khudia: I work on making enterprise AI faster and efficient at databricks

The Data and AI company

From faster training to inference at scale

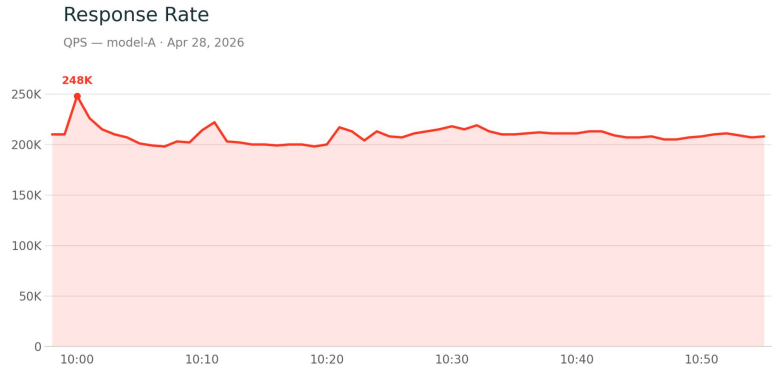
FlashOptim: Optimizers for Memory-Efficient Training (Spotlight @ICML26)



- Improved Weight splitting
- Companding to reduce quantization errors

Grammar correction @ 200K QPS

https://www.databricks.com/blog/how-superhuman-and-databricks-built-200k-qps-inference-platform-together



- *Routing:* Power of two choices
- Scaling: Faster container image loads
- Steady state: Quantization, CPU-side overhead reduction, async scheduling

databricks

If you want to geek out on:

- Fast inference at scale
- RL infra and post-training
- CUDA kernels and authoring frameworks

RSVP here

Wednesday, May 20
6:30 PM - 9:00 PM

We are hiring in:

- ML systems
- Agent systems
- Scaling & efficiency
- Retrieval

Apply here