I'm a SWE and eng leader specializing in making GPUs run fast. I've led and helped write most of Google's ML compilers stack, in clang, LLVM, and XLA. I'm looking for the equivalent of a Google L7 role.

Justin Lebar
justin.lebar@gmail.com

Updated 2023 Aug

Timeline

2020–present Waymo tech lead and manager focused on GPU infra and performance. I have impact across Waymo, ranging from strategic leadership to individual technical contributions.

- I set Waymo's GPU software strategy, e.g. deciding which compilers we use and which optimization opportunities (quantization, distillation, Triton, etc.) we focus on.
- I've built a team of 5 formal and about 5 informal members. We:
  - own Waymo's LiDAR processing pipeline (written in CUDA),
  - build infra for programming in CUDA (has cool portability and safety features), and
  - build ML optimization infra (e.g. XLA and quantized-aware training).
- I'm involved in most ML model launches across the company, debugging latency issues and often fixing issues in XLA.
- I'm the customer for the GPU part of Waymo's planned custom hardware. For example I defined the GPU's programming model.
- I act as a tech lead and senior SWE for the Google XLA project. I write and review code for XLA, clang, and LLVM and help grow the SWEs and TLs on the team.
- I'm a C++ expert; I teach a class at Waymo, and I'm one of the top 10 C++ readability reviewers by volume across all of Alphabet.

Fall 2019 Taught a semester of intro CS at Hampton University.

2016–2019 Google tech lead and manager for GPU compilers and XLA:CPU/GPU.

- I was part of the team of two that added support for CUDA in clang and made LLVM's PTX support usable.
- Later I moved to working on XLA, where I led the XLA:CPU and XLA:GPU projects as the tech lead and one of the managers. During this time we transformed XLA from a toy into a product with real users.

2013–2016 Google SWE on Census, a performance-monitoring and debugging tool.

2011–2013 Gecko hacker at Mozilla. I worked on many aspects of the platform — I was an official reviewer of seven modules. I specialized in optimizing the browser's memory usage.

2006–2011 Stanford University, BS/MS in Computer Science.

Etc.

I'm good at C++, CUDA, LLVM, Python, bash, gdb, git, hg, and vim.

I love to swing dance and teach.

I care deeply about inclusivity. I was involved with the *Yes, at Google* project and work on inclusivity in the swing dance scene.

Work Samples

A large fraction of my work these days is helping others (writing and reviewing designs, reviewing code, etc.), most of which isn't public. Here's what I can show:

- My CppCon 2016 talk about GPUs and CUDA support in clang
- I have some old stuff on my blog. The post on rvalue references is a good example of my approach to teaching C++.
- Recent patch to rewrite an LLVM pass. This helps with int8 kernels generated by XLA.
- Patch adding support for ahead-of-time cudnn/cublas autotuning. This is a big startup-time win for Waymo, plus it makes us more deterministic. There's a whole infrastructure at Waymo to generate these AoT autotuning choices, which I designed but didn't implement myself.
- "Fun" optimization to XLA:GPU compile time. This was part of a long chain of compile-time optimizations.

All my commits to XLA, clang, and LLVM are open-source; Google doesn't maintain an internal branch. My github activity tracker isn't showing most of my contributions for some reason, so the best I have to offer is running `git log` over the repos:

```
$ git clone https://github.com/tensorflow/tensorflow.git
$ git -C tensorflow --author lebar

$ git clone https://github.com/llvm/llvm-project.git
$ git -C llvm-project --author lebar
```