Mark’s wishlist

What I’m trying to build

discord.gg/gpumode

Community of ~15K developers that deeply cares about ML Systems

Weekly lectures

Working groups

[NEW]: Kernel competition platform

Our goal is to make ML cheaper by dramatically increasing the quality of open source ML systems software

What’s slowing me down?

Interoping PyTorch and custom CPP code is not great

1. AOT binary is huge and complicated to ship (in torchao spin up a machine per wheel)
2. JIT is unnecessarily slow
3. Custom kernel code in PyTorch often just using raw pointers to avoid overhead (see examples from DeepGemm and Bits&Bytes)

Dream: PyTorch is a frontend language for DL **but** make it trivial to plug in and ship more performant backends

3 Positive things

Code that’s easy for humans to understand, machines to generate and easy to ship

1. Bindings we codegen into: see Triton, CUTLASS backend
2. Triton code that’s easy to mix and match with PyTorch code
3. TORCH_LOGS=”output_code” easy to understand backend

3 things not working well

1. Shipping Python code that's multiplatform is suffering (ABI, different CUDA version, different OS, vendor specific languages)
2. AOT shipping a python binary is too hard unless team is strong at Dev infra
3. We need more JIT but PyTorch load_inline() has a crazy startup time

Bonus

1. PyTorch binary size is too damn high
2. No real "conda" replacement

How I’d like to contribute

Kickstart community adoption

1. Give a talk on the server
2. Get early adopters
3. Get early developers
4. Share updates in public