Google | Privacy, Safety & Security

PyTorch in Google3

Kiuk Chung (CoreML, Google)

May 26, 2024

Agenda

1. Intro & Background
2. PyTorch in Google3
    a. Import Process
    b. Building, Packaging, Running
3. Usage
4. Work in Progress

Intro & Background

PyTorch For Alphabet

Make PyTorch a fully supported option for research at Google and targeted production use-cases

DeWitt Clinton

Kiuk Chung

Qianli (Scott) Zhu

Greg Brosman

Julia Guo

Pooja Agarwal

Michael Voznesensky

Jake Harmon

Zoe Wang

Team Part of CoreML at Google. 9 and growing!

Mission Accelerate ML innovation by supporting a community-driven path from research to production in PyTorch

Problem External AI innovation happens in PyTorch but has not been supported in google3. AI builders need it to evaluate/adopt external innovation and publish internal innovation.

PyTorch at Google

Past

Use at your own risk

Ad-hoc imported by researchers on need-to basis. Poor documentation. Flaky community-driven support.

2024

Officially supported by CoreML

Maintained and updated by the PyTorch team. Documentation revamped with gotchas, workarounds, examples. Offered at Special Availability [beta] for research and experimentation.

Q2

PyTorch as 3rd Party in Google3

Get vanilla PyTorch running smoothly at Google. Integrations to AI infrastructure (Borg, profiling, logging, data, etc).

Q3

PyTorch as 1st Party in Google3

Setup to develop and contribute to PyTorch. Easily PR [public portions] to GitHub. Similar to Meta’s setup of PyTorch in FBCode.

Q4

Contribute Features to PyTorch and Ecosystem

JAX backend for PyTorch, quantization algorithms, better support PyTorch in Colab, etc

Future

PyTorch in Google3

PyTorch in Google3

Repository (g3)

Mono-repo. Must import third_party libs as source. PyTorch imported and built from source.

google3/third_party/py/torch

Imported from PyTorch GitHub at a release tag.

torch/google/** sub-dir for g3 specific extensions.

Patches

~50 patches on upstream code. Deals with uniqueness [compared to OSS standards] of Google's internal infrastructure, filesystem, multiprocessing, logging, compiler toolchain.

Prefer Extensions over Patching

Sub-class interfaces or author custom hooks and register where PyTorch allows. No custom behaviors [yet].

google3 ~ fbcode

google3 and fbcode are similar in many ways. We run tests with PYTORCH_TEST_FBCODE=1

Exercise: search for "fbcode" on PyTorch GitHub.

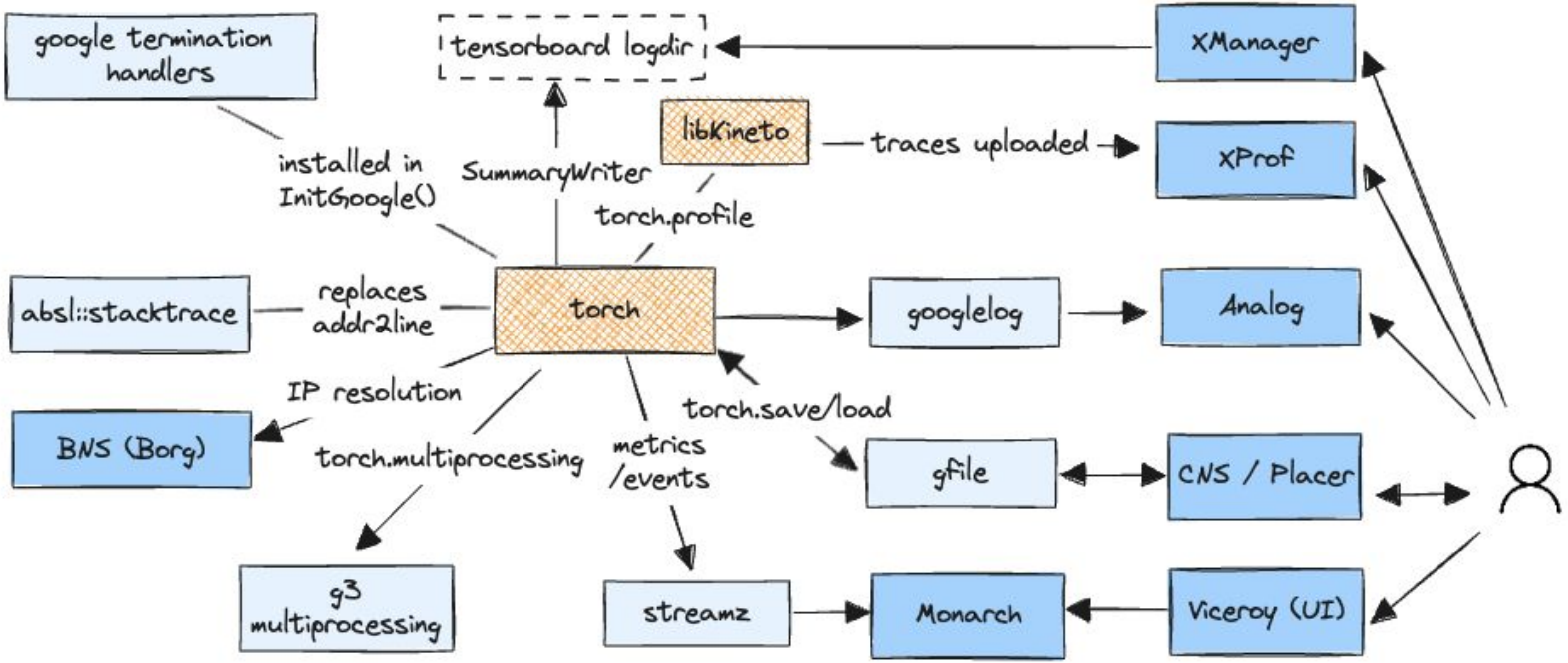
Packaging

Statically link [most] binaries. Hermetic PAR for Python binaries. All the C-extensions (e.g. libtorch) are statically linked into the interpreter itself.

PyTorch in Google3 vs OSS

	Google3	OSS
Repository	Mono-repo	Multi-repo
Build System	Blaze (Bazel)	CMake
Packaging	hermetic PAR	pip / conda
CUDA Compiler	Clang	NVCC
Logging Library	Abseil	Python built-in, glog (C++)
Flags Library	Abseil	gflags
Multiprocessing Library	g3_multiprocessing	Python built-in
Filesystem API	GFile	Python built-in

PyTorch in Google3



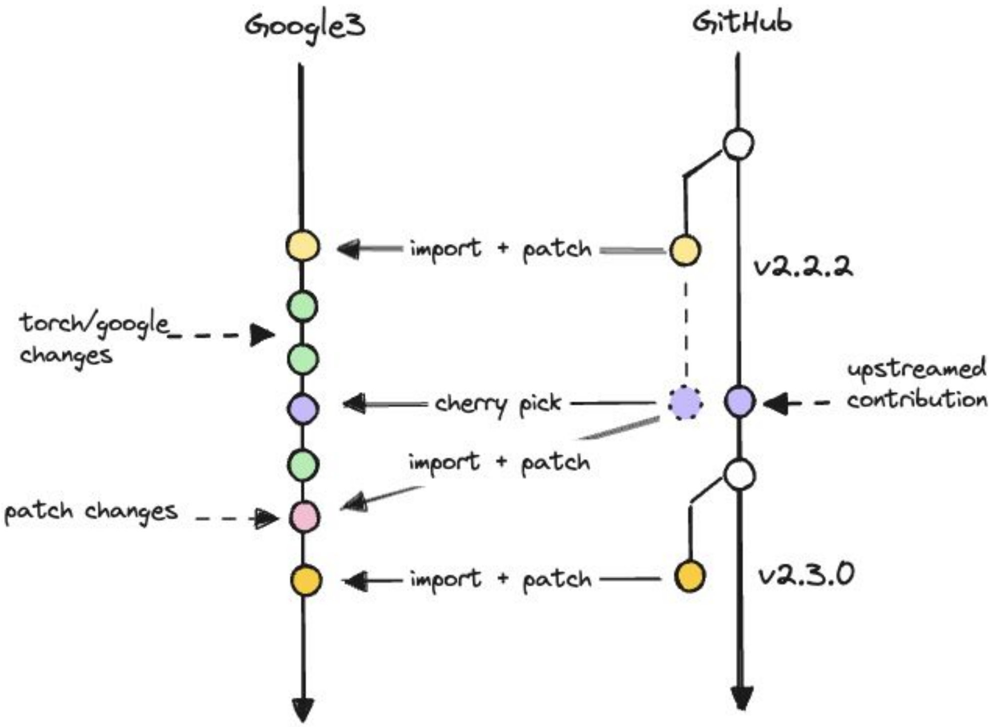
PyTorch in Google3

Import Process

Today

PyTorch as 3rd party

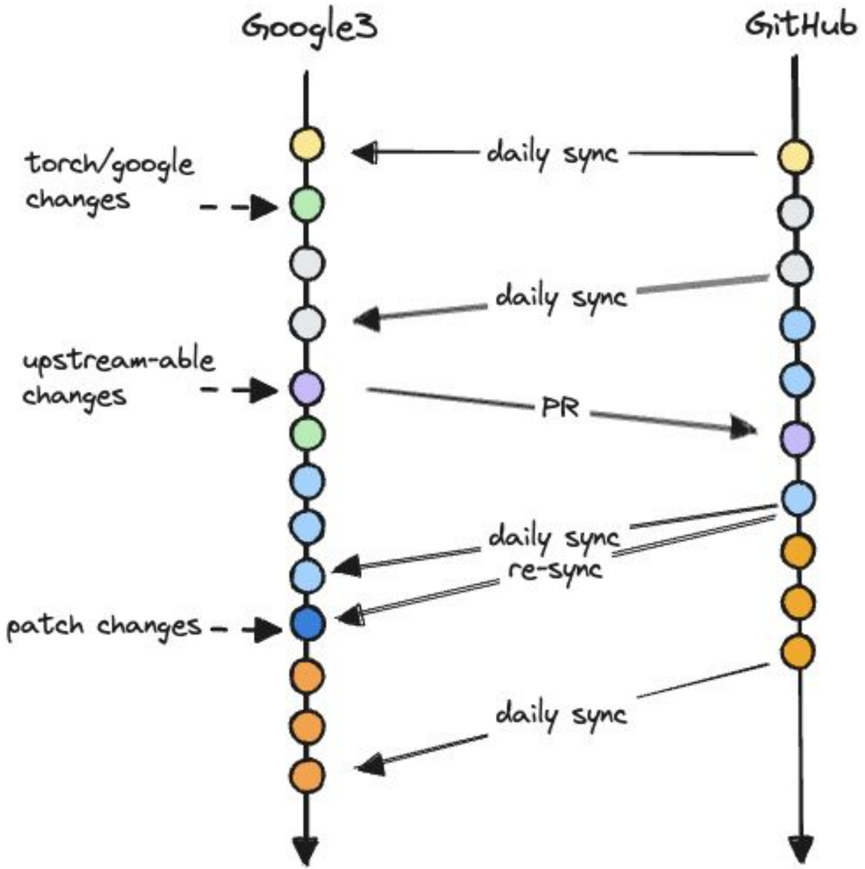
- Copybara - tool for transforming and moving code between Google3 and GitHub
- Import at release tag. Patches to upstream source code applied during import process
- torch/google [google-specific]changes not upstreamed
- Upstreamed PRs are cherry-picked
- On patch change, re-import with changed patches
- PyTorch still effectively treated as 3rd party



WIP

PyTorch as 1st party

- Still use Copybara
- Daily import from upstream main
- torch/google changes ignored by Copybara as usual
- Upstream-able changes exported as PR
- Patches still exist. Re-sync if changed)

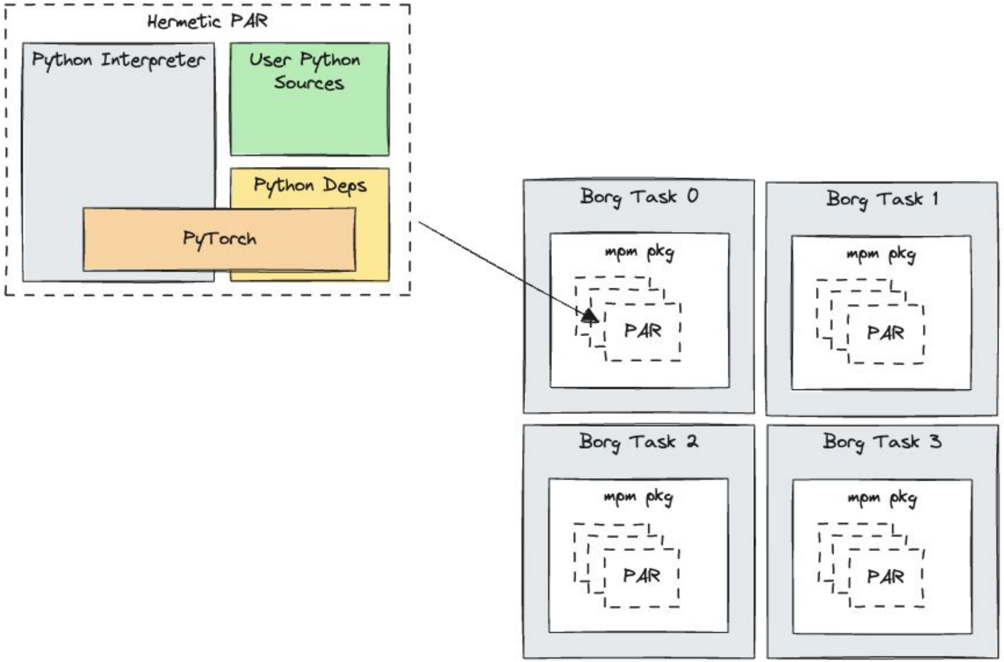


PyTorch in Google3

Building, Packaging, Running

Build, Package, Run

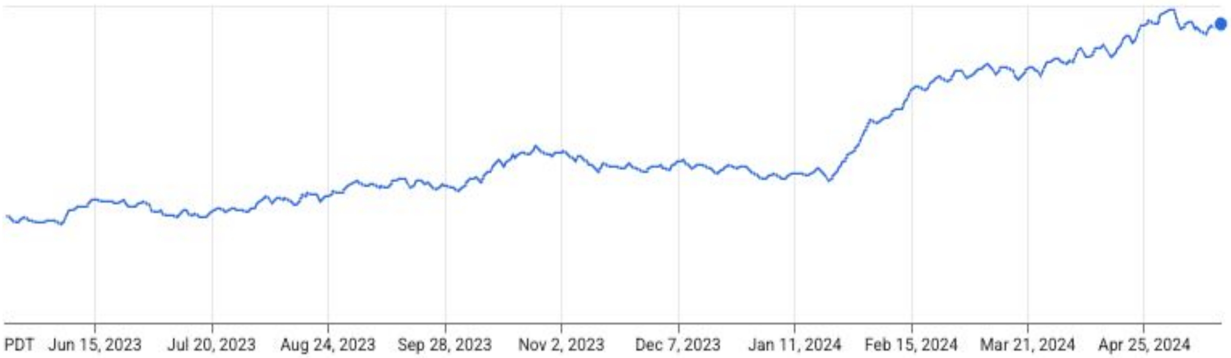
- User project packaged up as a PAR file: self-contained executable
- PAR can be run directly for local runs
- One or more PAR files packaged as MPM pkg
- Deployed to Borg as a job



Usage

Adoption Trend

Mostly research and experimentation driven. Benchmarking/evaluating papers and external models. Hints of production demand. Use JAX for LLMs at scale.



Work in Progress

Work in Progress

PyTorch with JAX as backend

Run PyTorch code using JAX under the hood. Freebie TPU, GSPMD support

Support PyTorch Ecosystem

Support PyTorch Ecosystem libs in Google3. Open source some of our own(?)

Contributions Upstream

Many Google3 patches can be upstreamed to improve PyTorch overall. Contribute to other aspects of the project: releases, CI/CD, beta-testing, etc

Thank you

Team PyTorch for Alphabet

Google CoreML

We’re Hiring!