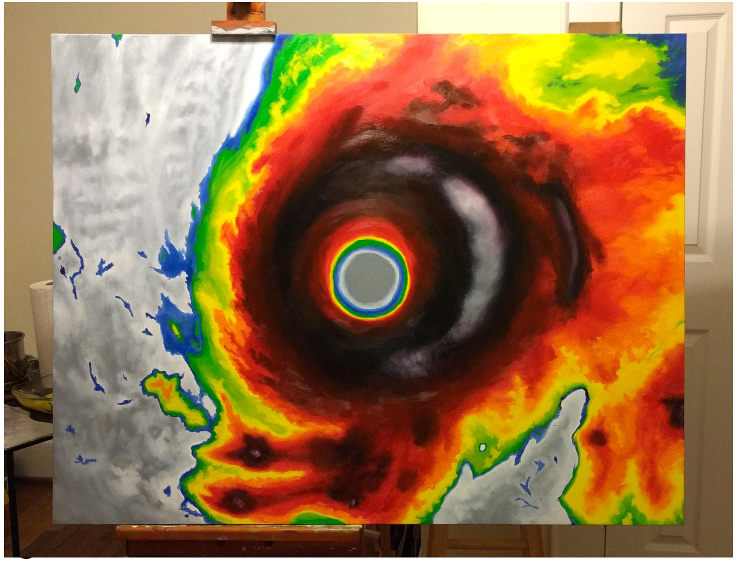
Making Humans & Code GPU-Capable

Data Council Austin 2022

Emily May Curtin
Senior ML Ops Engineer, Mailchimp/Intuit
@emilymaycurtin

Howdy, I'm Emily

- 👽 ATLien (don't call it Hotlanta)
- ❌ #NotADataScientist
- 🎨 Oil painter by passion
- 💾 MLOps by day job (btw we're hiring!)
- ❤️ Big fan of Ryan Curtin



Our Goal:

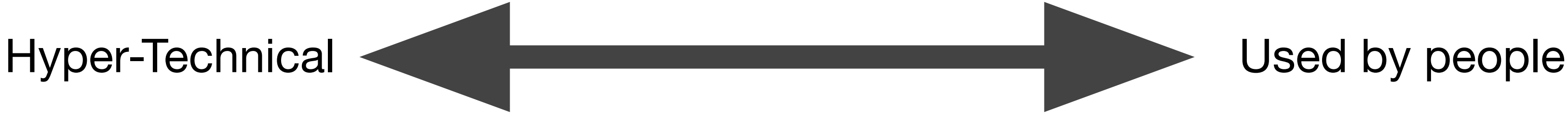
Help Data Scientists produce higher quality work faster

MLOps

is a hyper-technical field that is

all about *people*

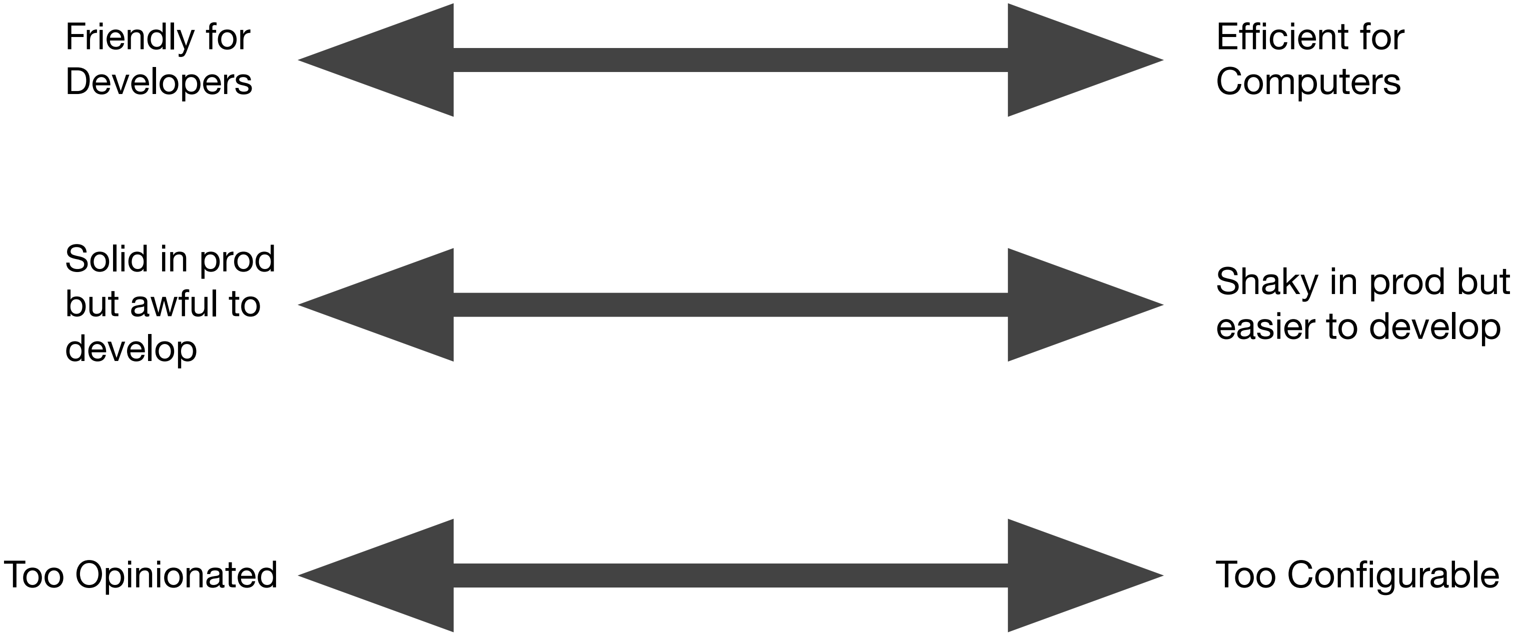
Inherent Design Tradeoff



Hyper-Technical

Used by people

Other Design Tradeoffs



Friendly for Developers ↔ Efficient for Computers

Solid in prod but awful to develop ↔ Shaky in prod but easier to develop

Too Opinionated ↔ Too Configurable

Let’s talk about ML stacks

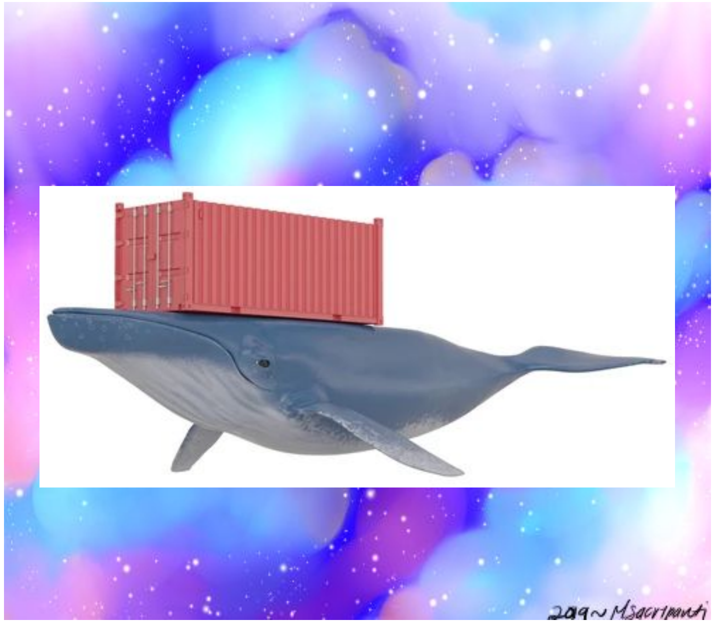
Typical ML Tech Stack

- Python
- Pytorch, HuggingFace, Tensorflow
- Docker
- Cloud infrastructure (we happen to use GCP)
- Kubernetes either directly or indirectly

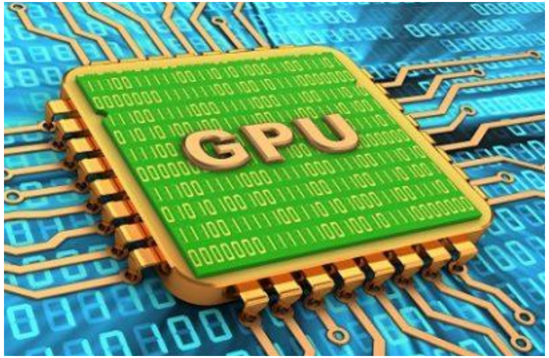
Benefits

- Good scalability, reproducibility
- Cloud infra good for spiky ML workloads (vs. more consistent, predictable web service)

But...



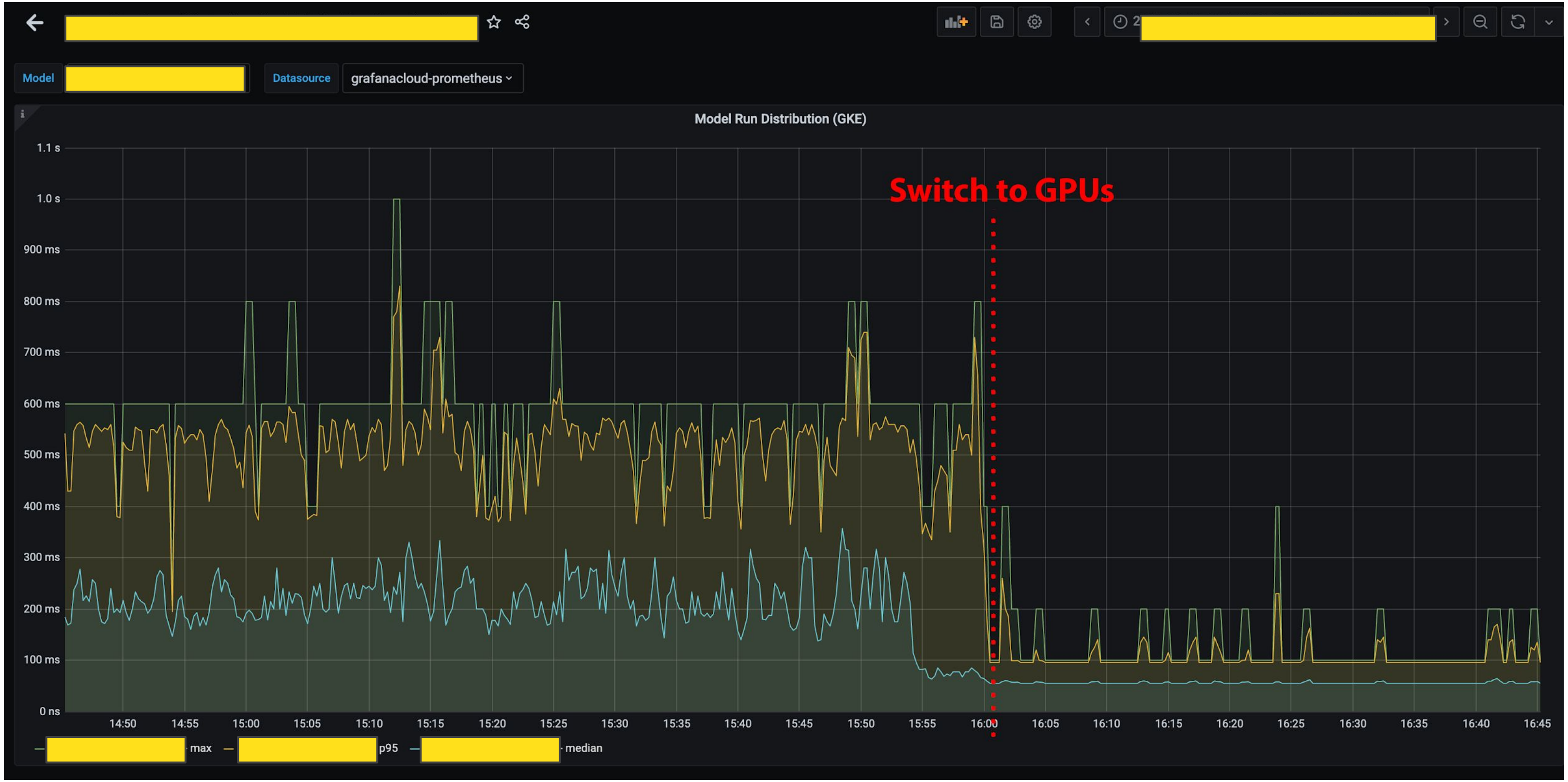
?????????



Let’s talk about GPUs

GPU
GPU
GPU
ANIMATED LASERS
GEFORCE
HDMI
VGA
DVI

GPUs Can Be *Really* Awesome



[redacted]

Model [redacted]

Datasource grafanacloud-prometheus

Switch to GPUs

GPUs . . .

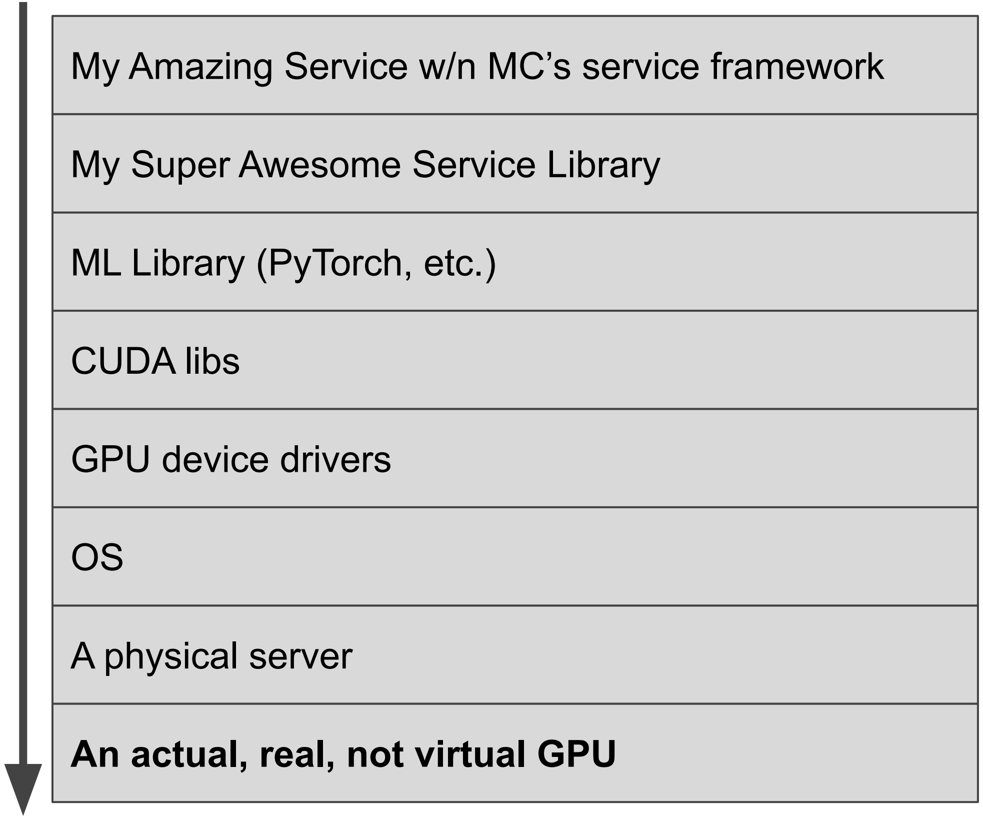
- Are optional hardware peripherals
- Require special drivers
- Rely on system buses for I/O

GPUs . . . Are Printers

GPUs . . . Are Printers

That are very good at linear algebra

Call Stack on a plain server



My Amazing Service w/n MC’s service framework

My Super Awesome Service Library

ML Library (PyTorch, etc.)

CUDA libs

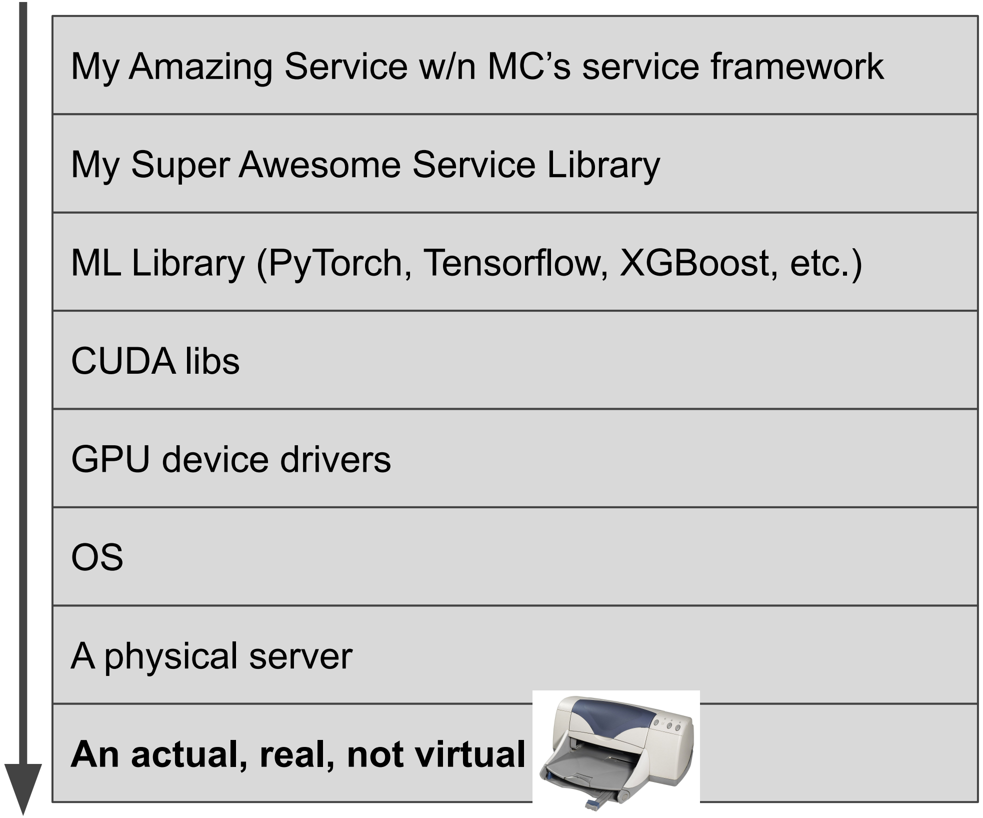
GPU device drivers

OS

A physical server

An actual, real, not virtual GPU

Call Stack on a plain server



My Amazing Service w/n MC’s service framework

My Super Awesome Service Library

ML Library (PyTorch, Tensorflow, XGBoost, etc.)

CUDA libs

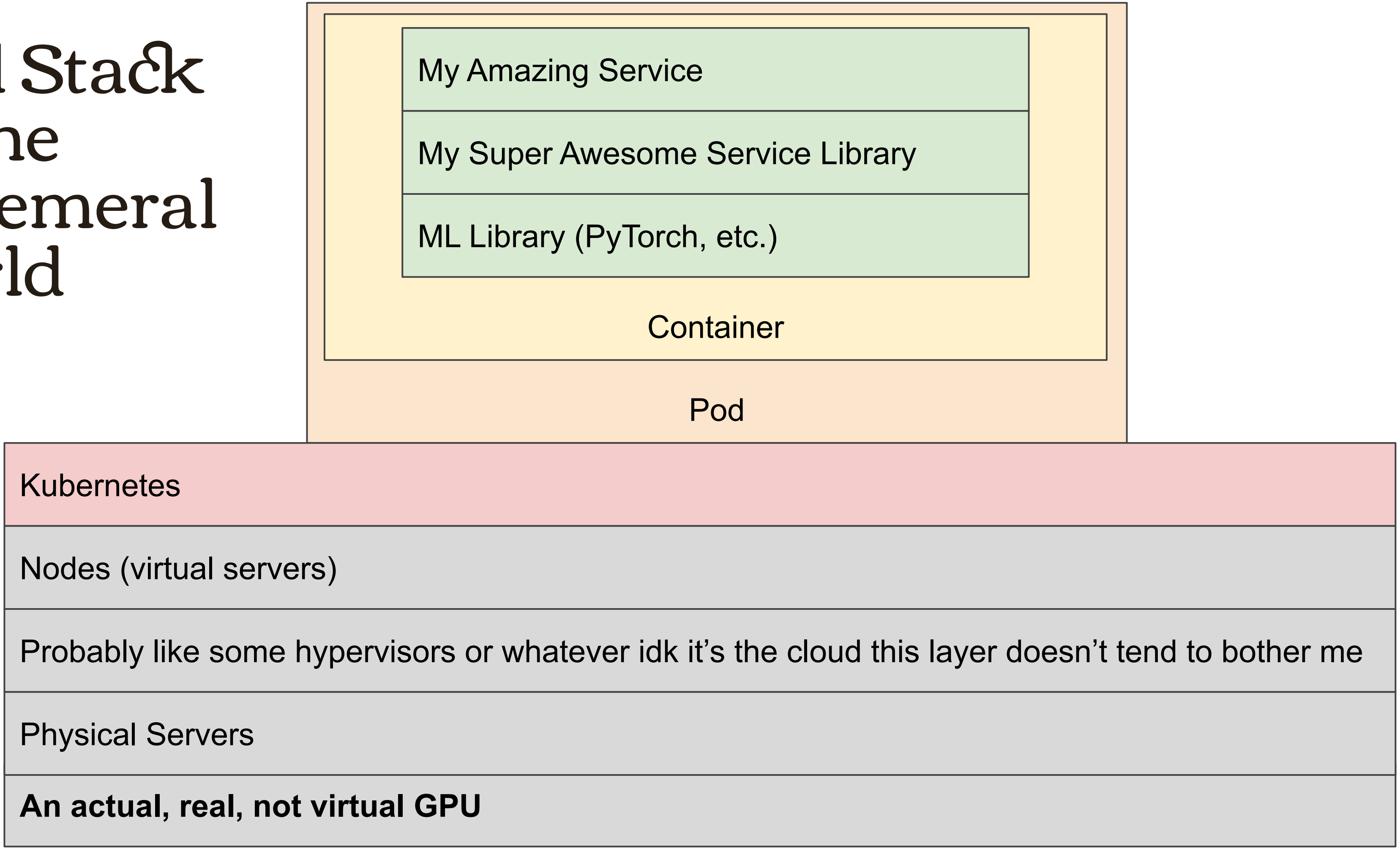
GPU device drivers

OS

A physical server

An actual, real, not virtual

Call Stack in the ephemeral world



My Amazing Service

My Super Awesome Service Library

ML Library (PyTorch, etc.)

Container

Pod

Kubernetes

Nodes (virtual servers)

Probably like some hypervisors or whatever idk it’s the cloud this layer doesn’t tend to bother me

Physical Servers

An actual, real, not virtual GPU

What you need to talk to a GPU

- GPU
- Drivers
  - `nvidia.ko` - Kernel mode GPU driver
  - `libcuda.so` - User mode GPU driver (aka low-level API)
- CUDA Toolkit
  - `libcudart.so` - Runtime API (aka high-level API)
  - `cuBLAS`, `cuRAND`, `cuSOLVER`, and other toolkit libs

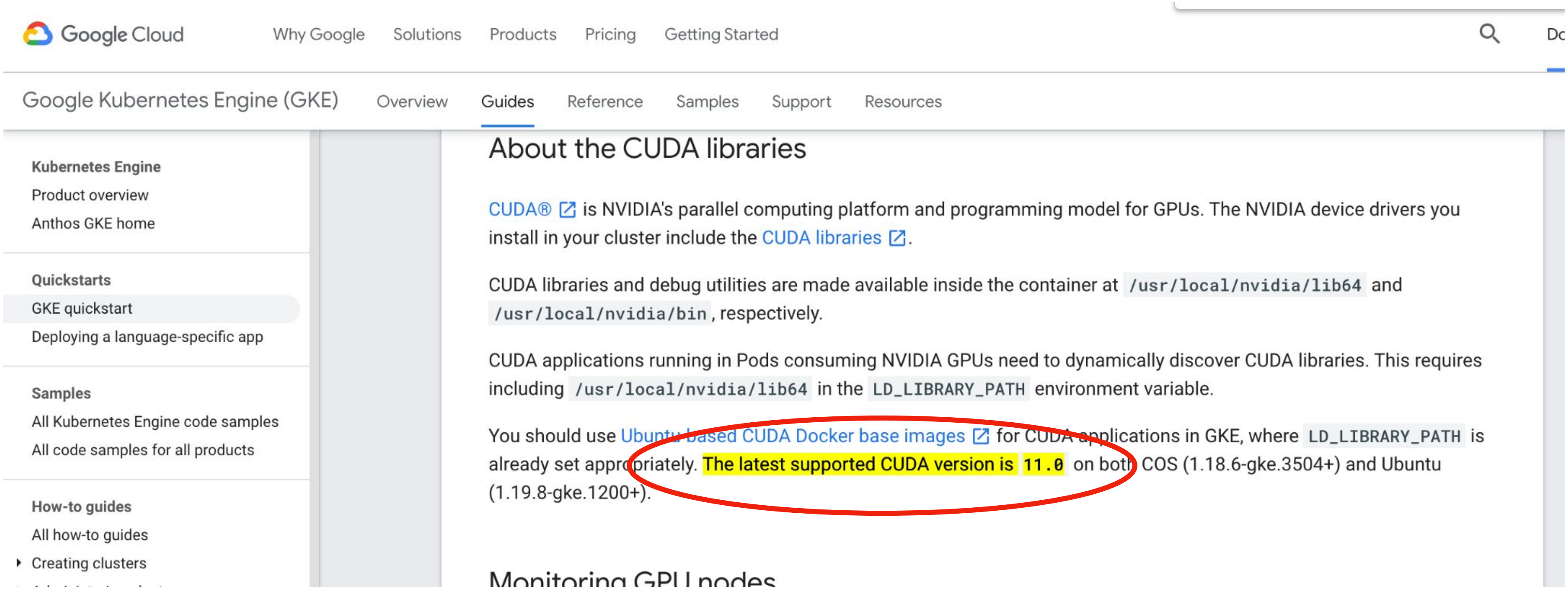
GPUs and Device Drivers

2022

These come from your k8s service provider, GKE in my case

GKE Provides

- Configurable GPUs and GPU pools
- DaemonSet for device drivers



About the CUDA libraries

CUDA® is NVIDIA's parallel computing platform and programming model for GPUs. The NVIDIA device drivers you install in your cluster include the CUDA libraries.

CUDA libraries and debug utilities are made available inside the container at `/usr/local/nvidia/lib64` and `/usr/local/nvidia/bin`, respectively.

CUDA applications running in Pods consuming NVIDIA GPUs need to dynamically discover CUDA libraries. This requires including `/usr/local/nvidia/lib64` in the `LD_LIBRARY_PATH` environment variable.

You should use Ubuntu-based CUDA Docker base images for CUDA applications in GKE, where `LD_LIBRARY_PATH` is already set appropriately. The latest supported CUDA version is `11.0` on both COS (1.18.6-gke.3504+) and Ubuntu (1.19.8-gke.1200+).

Monitoring GPU nodes

Various CUDA APIs and other libs

- Some Python ML Libs ship with binaries in the wheels
  - Dependent on Python package manager (pip, anaconda, etc)
  - Usually does not include `libcuda.so`
- Might be made available via your device driver Daemonset
  - Set `LD_LIBRARY_PATH` to access
  - Usually only API binaries, not other toolkit libs
- Might have to DIY via base container or custom install step
- Might have to combine all of the above

Matching CUDA Versions Matters

- CUDA version supported by your ML library of choice
- CUDA version in your base docker image
- CUDA version available on your k8s nodes, exposed through Daemonset

Matching CUDA Versions Matters*

Matching CUDA Versions Matters*

*Sometimes. Depending. Maybe not.

Matching CUDA Versions Matters*

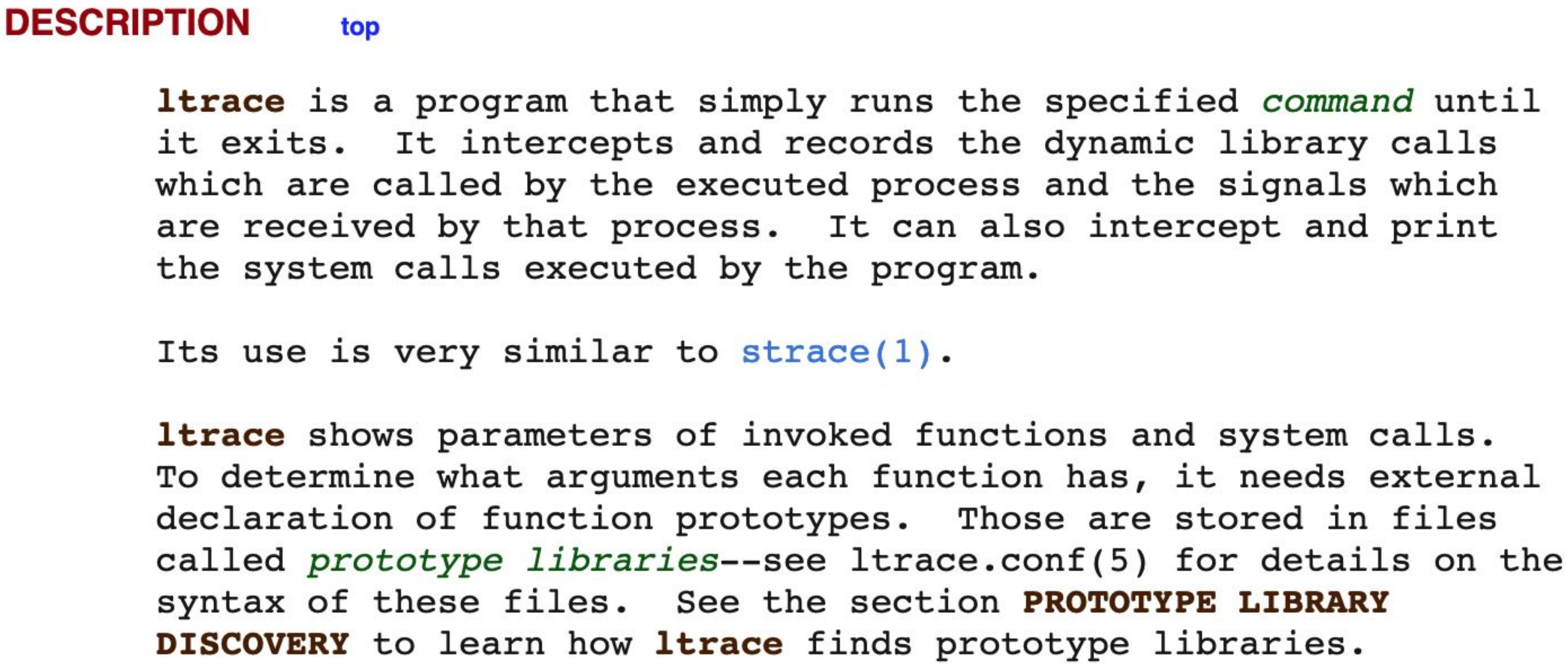
*Sometimes. Depending. Maybe not.

YMMV depending on your library

- PyTorch does a lot of stuff to support 10.x and 11.x
- Tensorflow is very picky about everything

CUDA has complex forward and backward compatibility scenarios

ltrace and strace rock



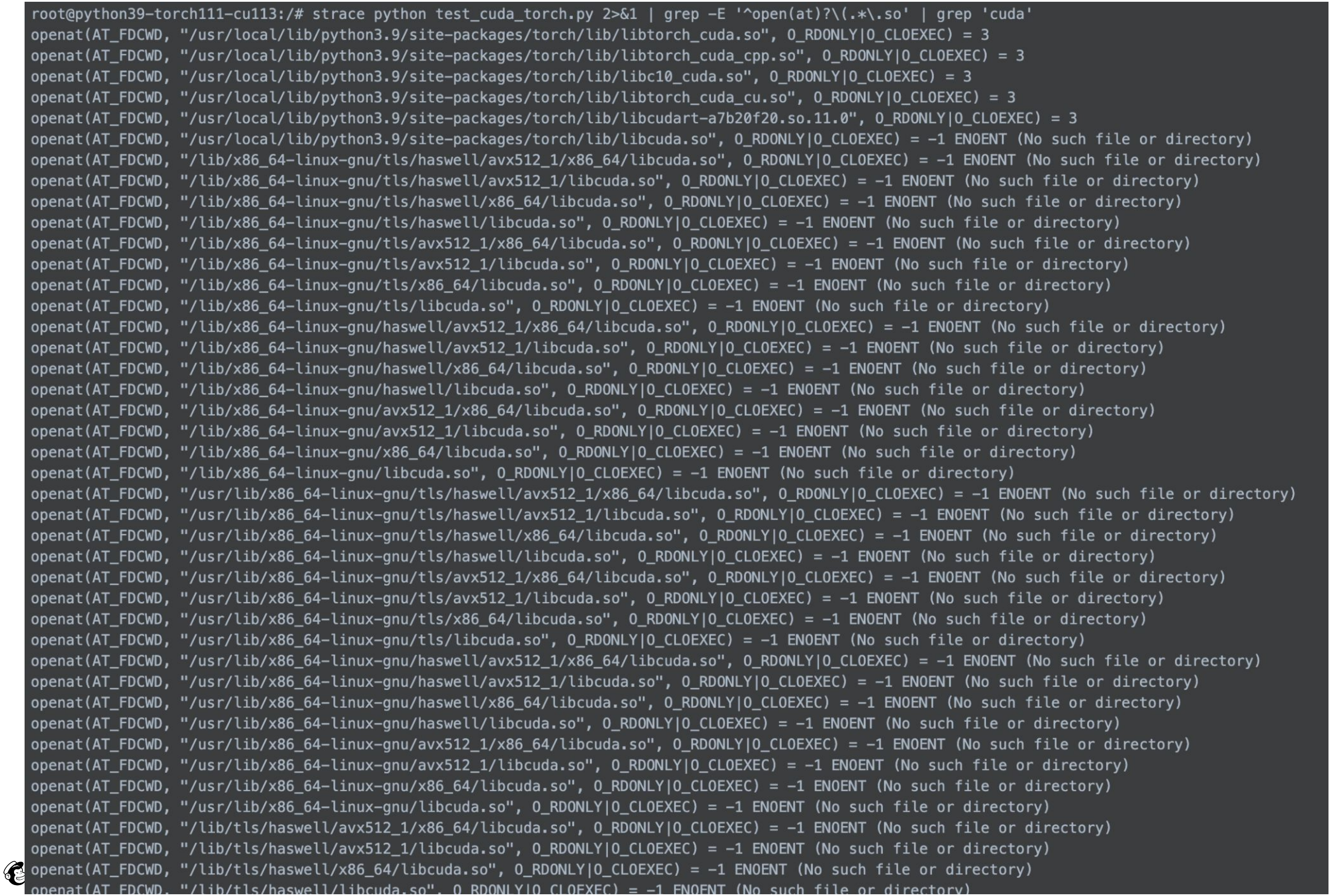
DESCRIPTION

top

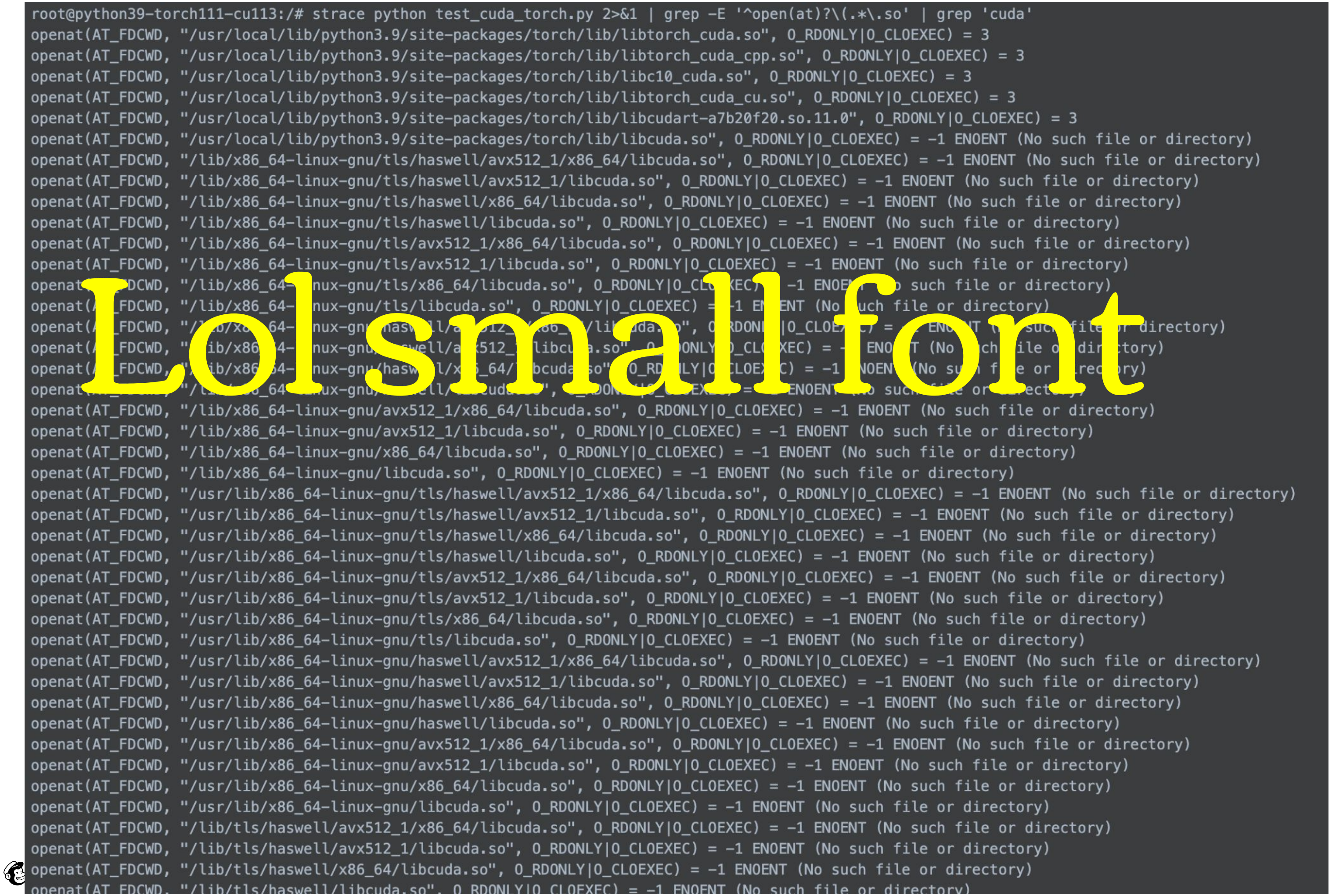
ltrace is a program that simply runs the specified *command* until it exits. It intercepts and records the dynamic library calls which are called by the executed process and the signals which are received by that process. It can also intercept and print the system calls executed by the program.

Its use is very similar to strace(1).

ltrace shows parameters of invoked functions and system calls. To determine what arguments each function has, it needs external declaration of function prototypes. Those are stored in files called *prototype libraries*--see ltrace.conf(5) for details on the syntax of these files. See the section **PROTOTYPE LIBRARY DISCOVERY** to learn how **ltrace** finds prototype libraries.

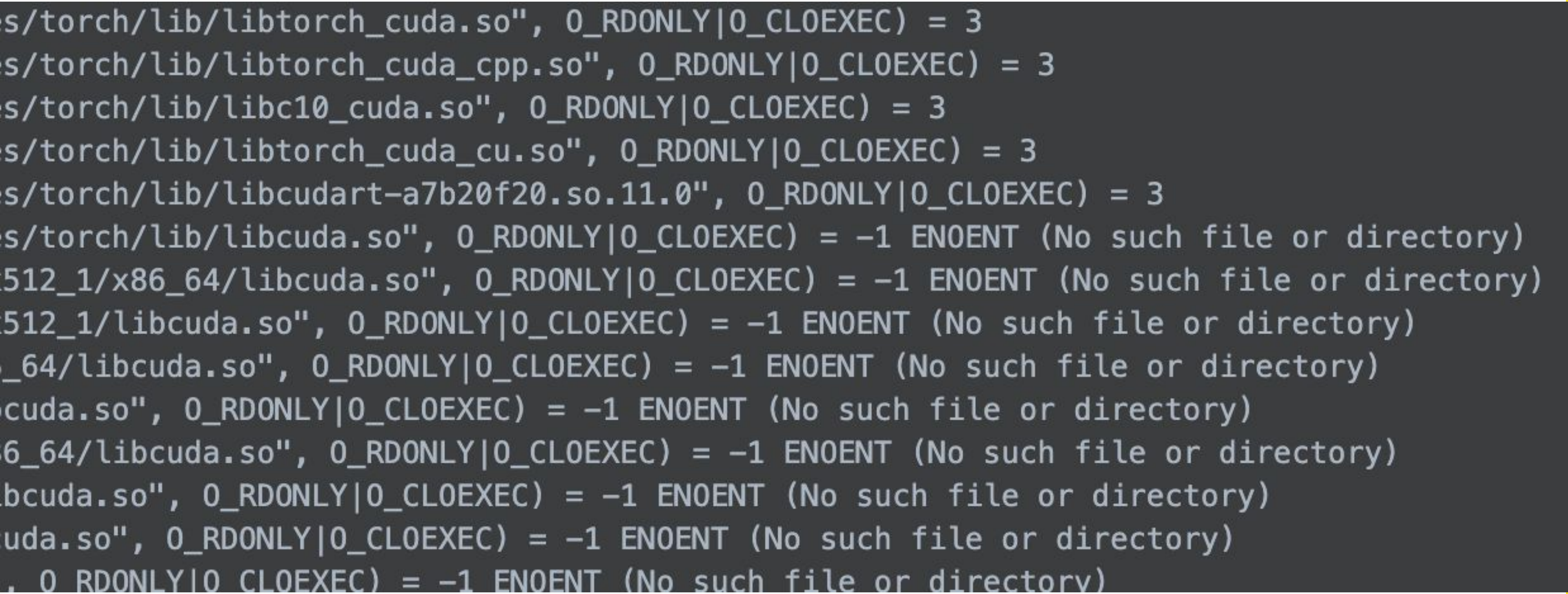


```
root@python39-torch111-cu113:/# strace python test_cuda_torch.py 2>&1 | grep -E '^open(at)?\(.*\.so' | grep 'cuda'
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libtorch_cuda.so", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libtorch_cuda_cpp.so", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libc10_cuda.so", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libtorch_cuda_cu.so", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libcudart-a7b20f20.so.11.0", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/tls/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/tls/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/tls/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/tls/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
```



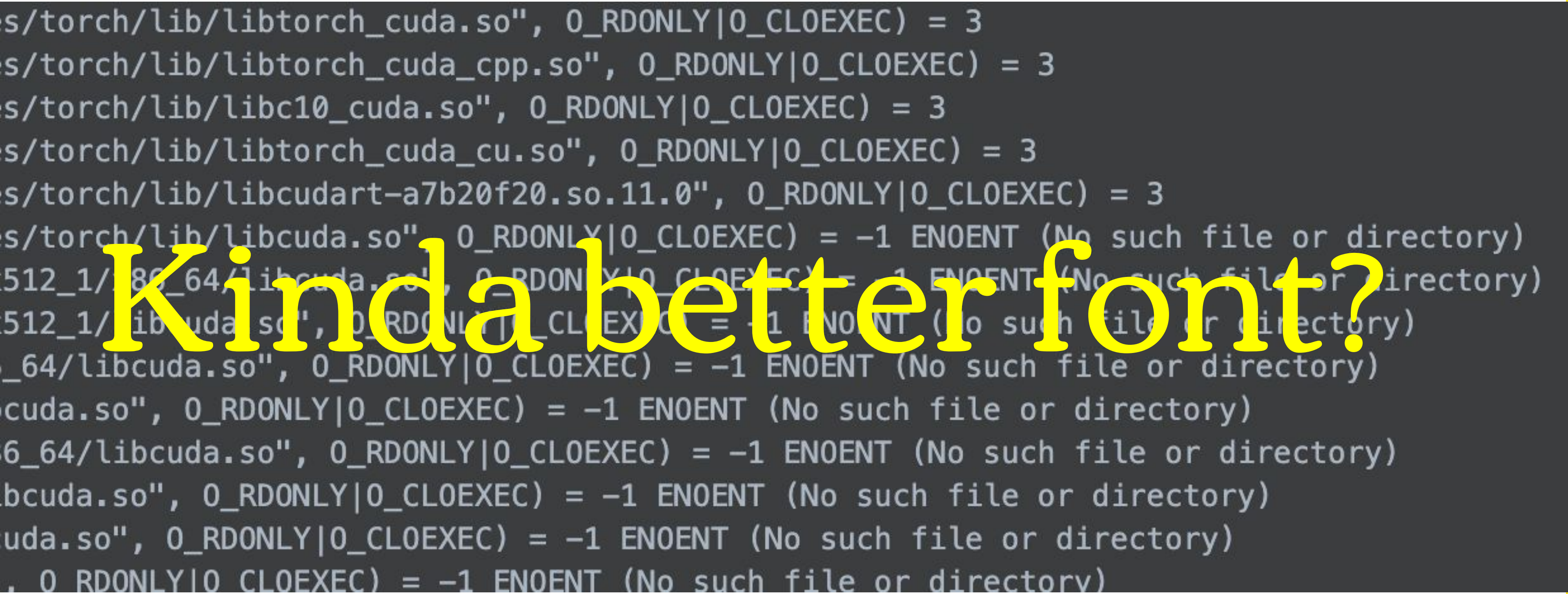
```
root@python39-torch111-cu113:/# strace python test_cuda_torch.py 2>&1 | grep -E '^open(at)?\(.*\.so' | grep 'cuda'
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libtorch_cuda.so", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libtorch_cuda_cpp.so", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libc10_cuda.so", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libtorch_cuda_cu.so", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libcudart-a7b20f20.so.11.0", O_RDONLY|O_CLOEXEC) = 3
openat(AT_FDCWD, "/usr/local/lib/python3.9/site-packages/torch/lib/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/tls/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/tls/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/usr/lib/x86_64-linux-gnu/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/tls/haswell/avx512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/tls/haswell/avx512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/tls/haswell/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/tls/haswell/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
```

Lol small font

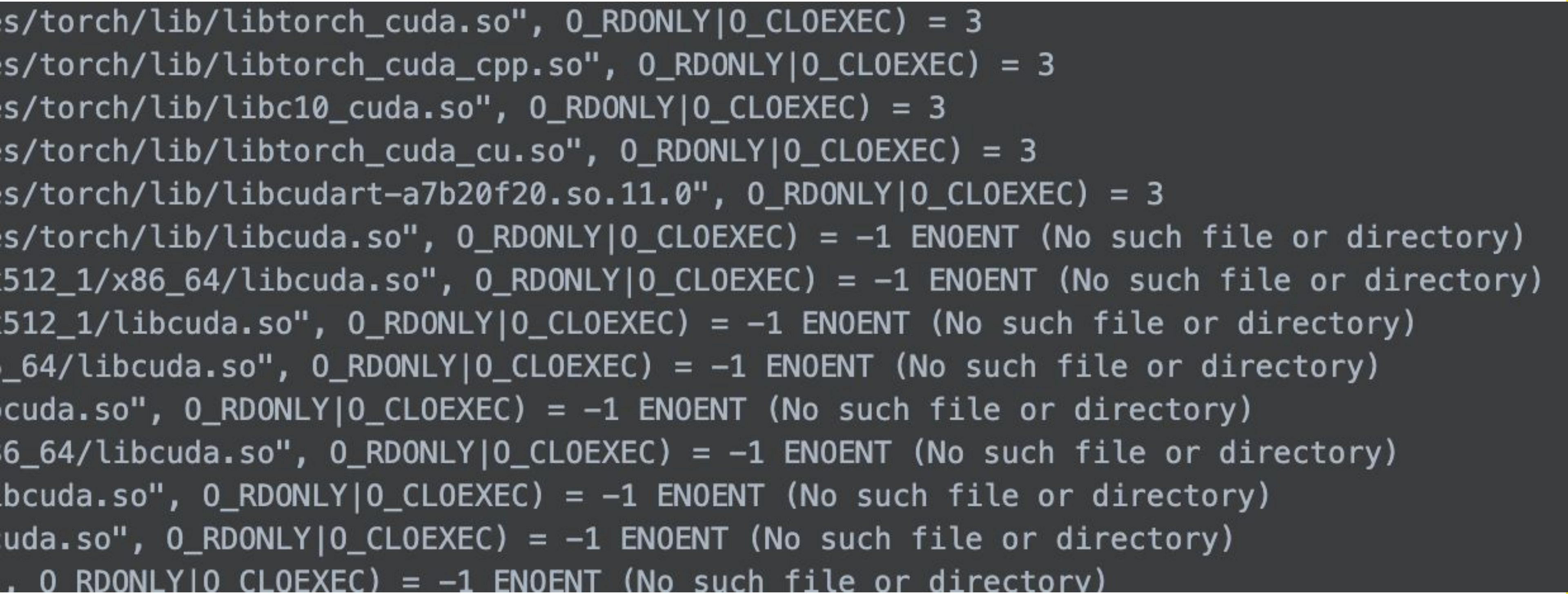


```
s/torch/lib/libtorch_cuda.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libtorch_cuda_cpp.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libc10_cuda.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libtorch_cuda_cu.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libcudart-a7b20f20.so.11.0", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
cuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
6_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
bcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
uda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
, O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
```

Kinda better font?



```
s/torch/lib/libtorch_cuda.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libtorch_cuda_cpp.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libc10_cuda.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libtorch_cuda_cu.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libcudart-a7b20f20.so.11.0", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
cuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
6_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
bcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
uda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
, O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
```

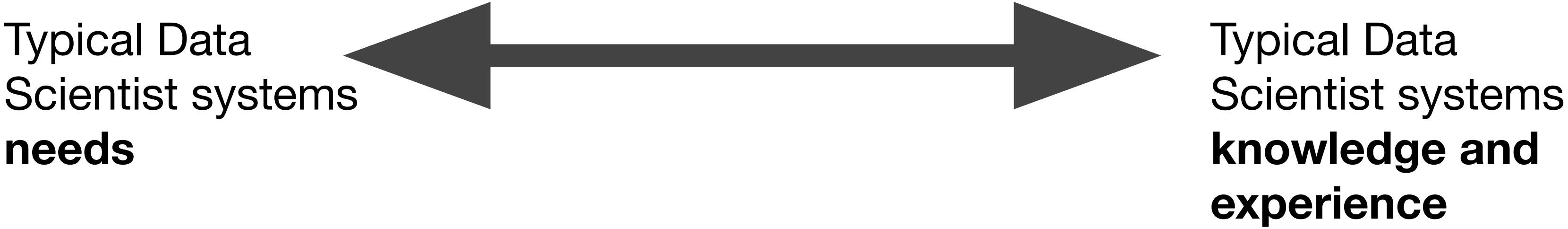


```
s/torch/lib/libtorch_cuda.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libtorch_cuda_cpp.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libc10_cuda.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libtorch_cuda_cu.so", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libcudart-a7b20f20.so.11.0", O_RDONLY|O_CLOEXEC) = 3
s/torch/lib/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
512_1/x86_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
512_1/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
cuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
6_64/libcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
bcuda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
uda.so", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
, O_RDONLY|O_CLOEXEC) = -1 ENOENT (No such file or directory)
```

MLOps

is a hyper-technical field that is

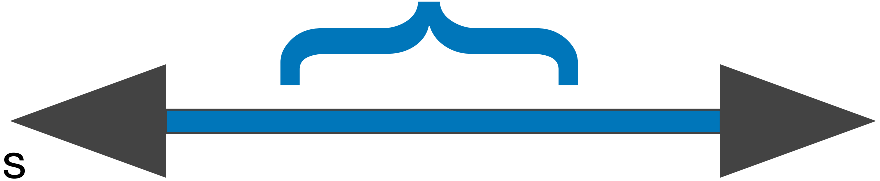
all about *people*



Typical Data Scientist systems **needs**

Typical Data Scientist systems **knowledge and experience**

Where we (MLOps) come in



Typical Data Scientist systems **needs**

Typical Data Scientist systems **knowledge and experience**

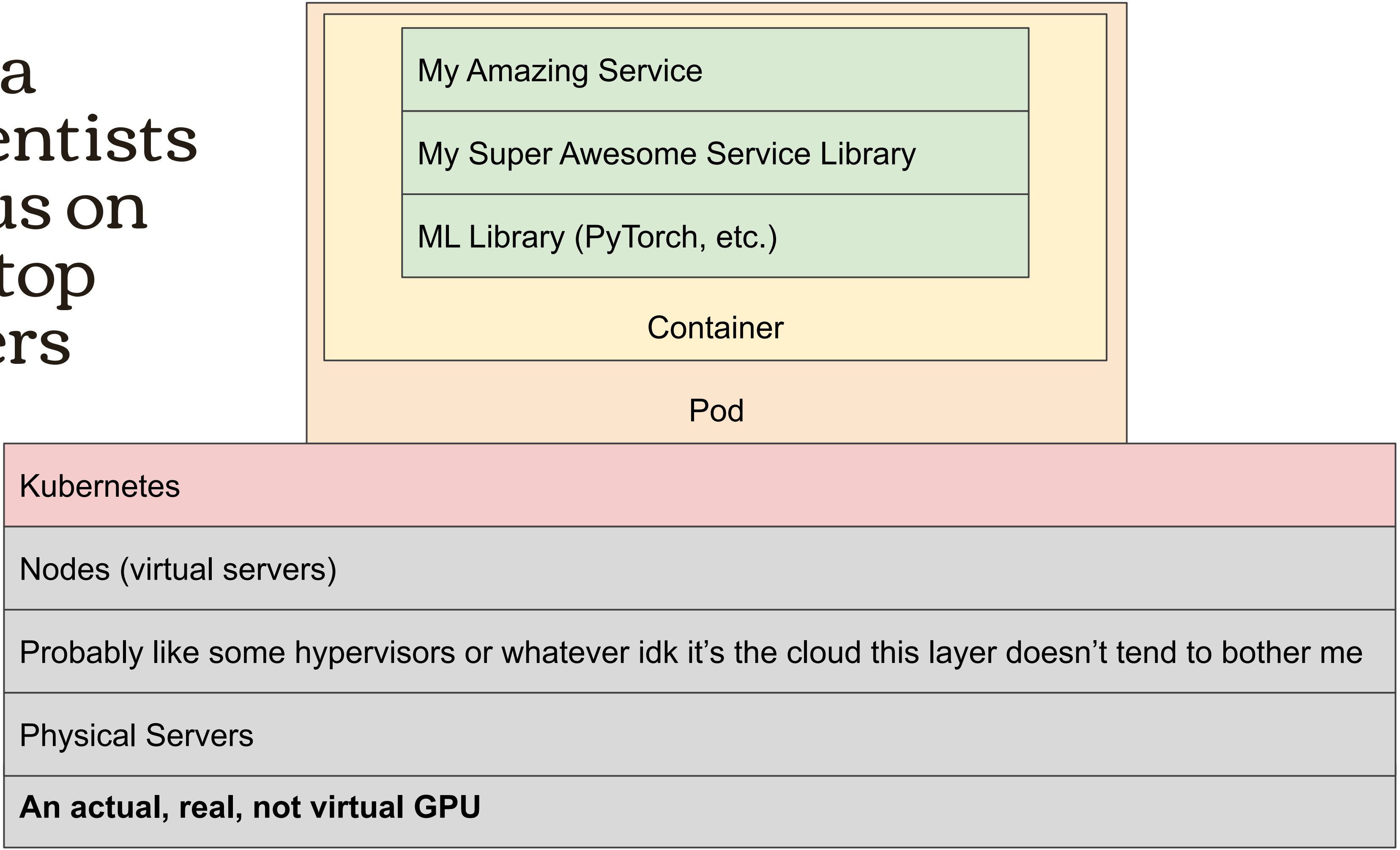
Systems Abstraction

Providing a good enough encapsulation of the system so Data Scientists can focus on the application layers.

It’s really hard.

Most MLOps systems are *full* of leaky abstractions.

Data Scientists focus on the top layers



My Amazing Service

My Super Awesome Service Library

ML Library (PyTorch, etc.)

Container

Pod

Kubernetes

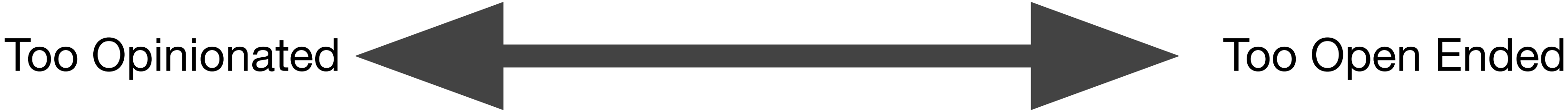
Nodes (virtual servers)

Probably like some hypervisors or whatever idk it’s the cloud this layer doesn’t tend to bother me

Physical Servers

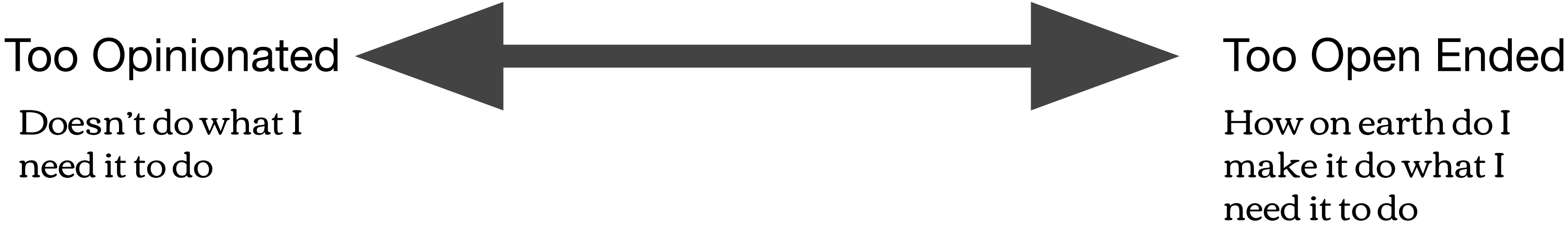
An actual, real, not virtual GPU

Design Tradeoffs


Too Opinionated

Too Open Ended

Design Tradeoffs



Too Opinionated

Doesn't do what I
need it to do

Too Open Ended

How on earth do I
make it do what I
need it to do

To enable high tech, go low tech

GPUs for ML

… via repo templating

@lowcost_cosplay

Repo templating is not cool. And it works.

Repo Templating

- Provide a good enough, general enough base for the majority
- Includes
  - Base container to encapsulate the runtime environment
  - Places to integrate custom Python code
  - Basic run scripts for applications
  - Basic CI/CD stuff (ex: Jenkinsfile)
- GPU capability built in via base container(s)

Challenges

- Is your base container general enough? Will it match prod?
- Differences between libraries, batch jobs, live services, etc.
- How do children of a template get updates from the parent?
- How do we provide general GPU capability to everything using the template(s)?

Some Hard-Won Wisdom

- One template per project type (library, batch job, etc.) with shared base containers.

- Allow massive flexibility in ML lib choice within your language

- One base container is probably not good enough. Have curated options. (ex: tensorflow breaks everything)

- Design for the 90% cases, don’t generalize the other 10%

In Conclusion

- MLOps is a super technical role that’s **all about people**
- `strace` is your friend
- Repo templating is your friend
- Be uncool to do cool stuff

@emilymaycurtin

Thank you.