RayServe + Tensorflow Demo

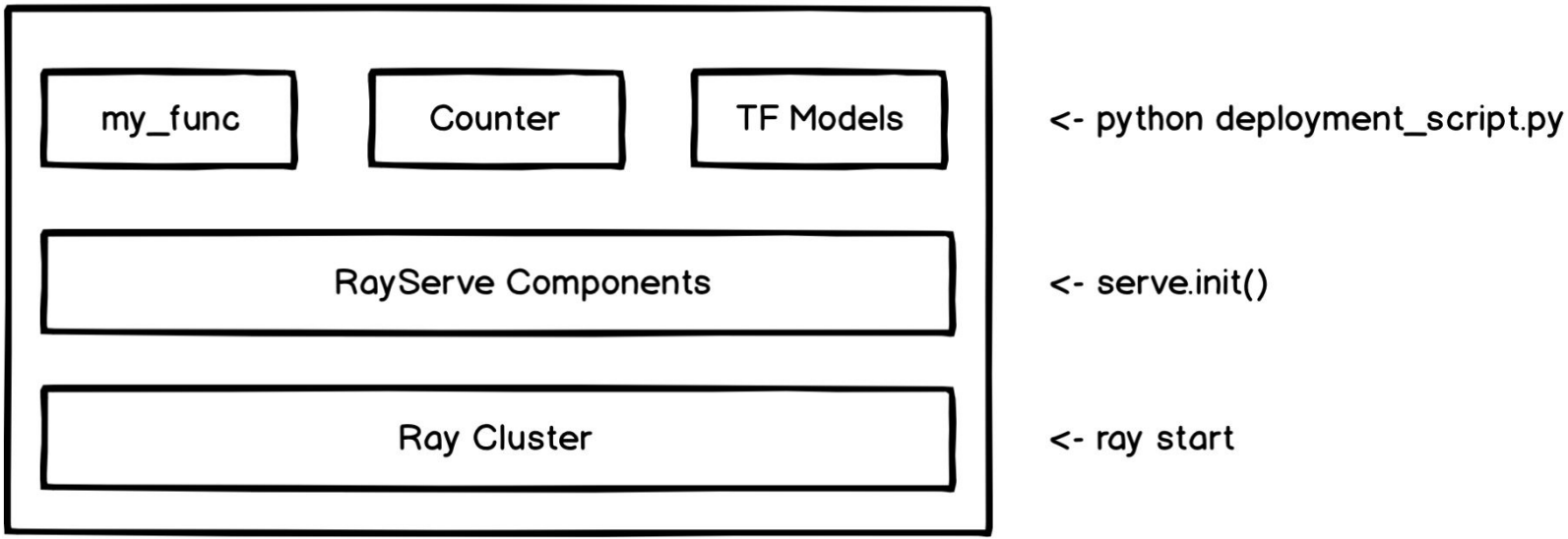
Simon Mo

anyscale

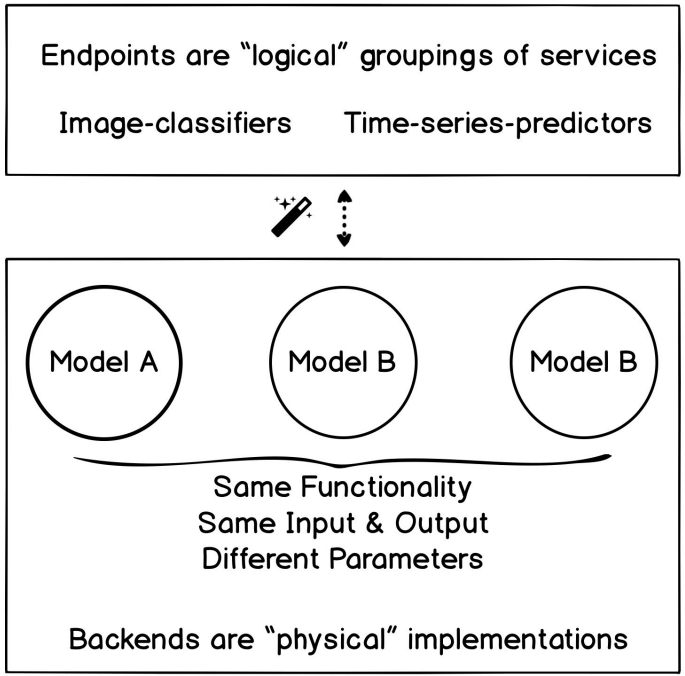
Introduction

- This will be a short Demo (5min)
- RayServe is designed to be a *simple* tool
- We cover how real deployment will look like
- Please interrupt me anytime!

Setup


my_func
Counter
TF Models
<- python deployment_script.py
RayServe Components
<- serve.init()
Ray Cluster
<- ray start

Concepts: Endpoints and Backends



Demo

In Summary

- Deploy *stateless* functions and *stateful* workload
- Concept of *endpoint* and *backend*
- Deploy multiple TensorFlow models under the same *endpoint* and how this enables rolling update and A/B testing
- **We did all these without shutting down the services or the cluster.**

- We didn't show you:
    - Create more replicas and scale to multiple machines
    - Monitoring, batching, etc...