网易云
163yun.com

IT大咖说
知识分享平台

caicloud
才云

Caicloud TaaS Introduction

TensorFlow on Kubernetes

岑鹏浩 才云科技

Agenda

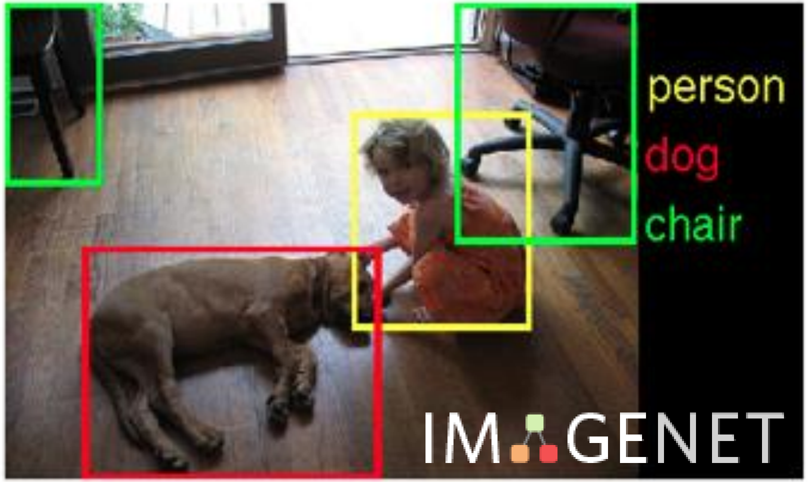
- Deep Learning Introduction
- TensorFlow Introduction
- Distributed TensorFlow on Kubernetes
- TaaS Introduction
- TaaS Demo

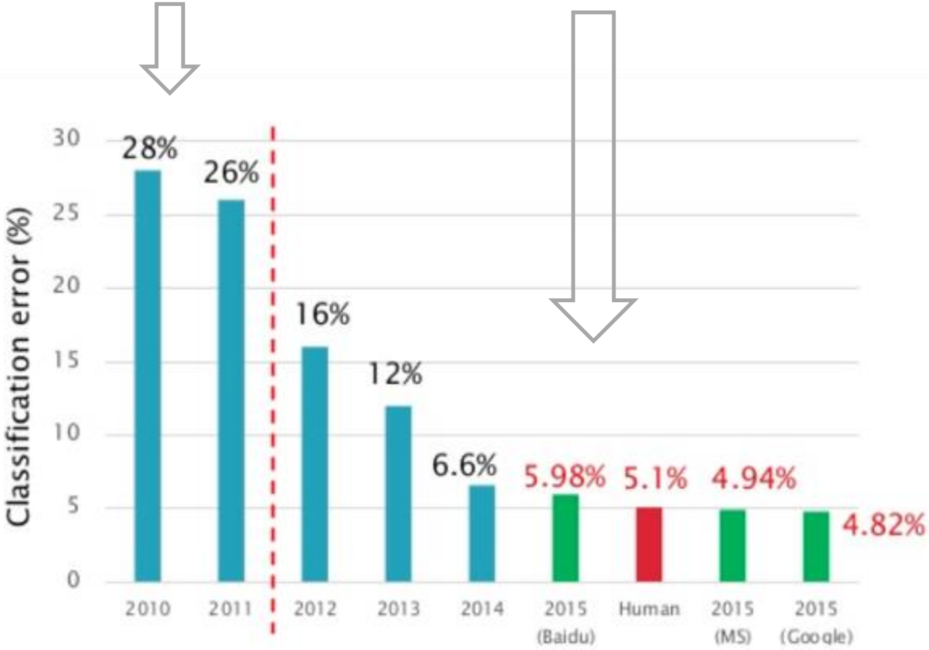
Deep Learning Introduction

Deep Learning Introduction

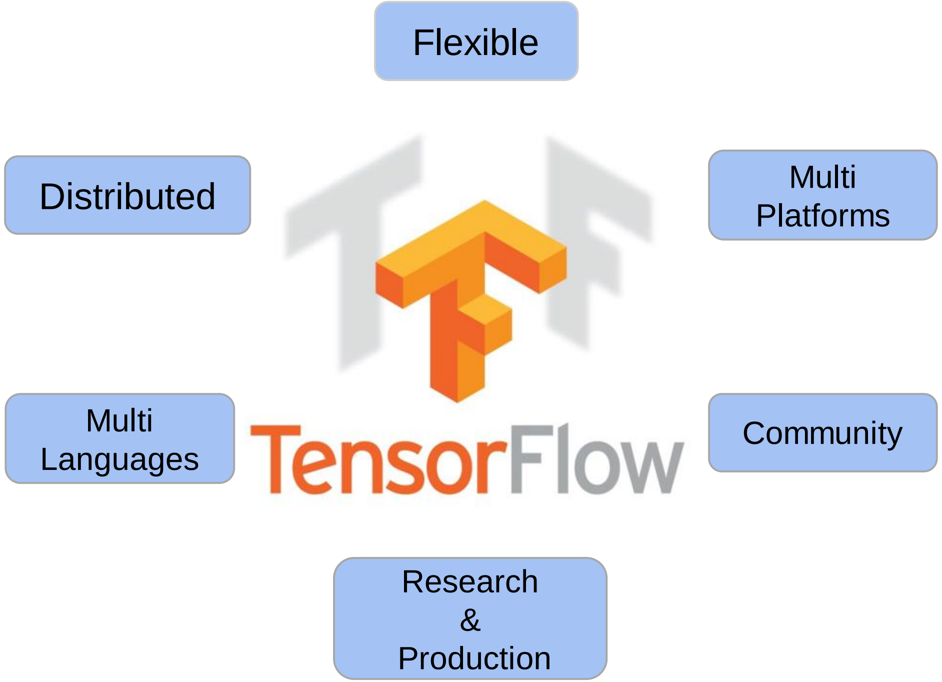
before deep learning

after deep learning





TensorFlow: Machine Learning for Everyone



Flexible

Distributed

Multi Platforms

Multi Languages

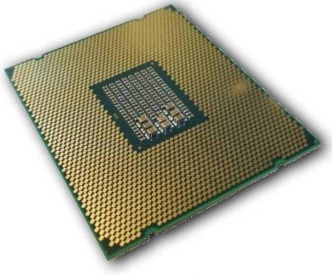
Community

Research & Production

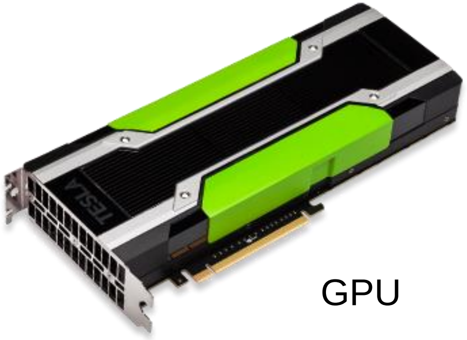
- Open Source
- Fast, Flexible, and Production-Ready
- Python and C++ API
- Distributed Processing
- Supports CPUs & GPUs
- Machine Learning & Deep Neural Network
- Based on data flow graphs

TensorFlow Introduction -- Supports Many Platforms

TensorFlow: Machine Learning for Everyone



CPU



GPU

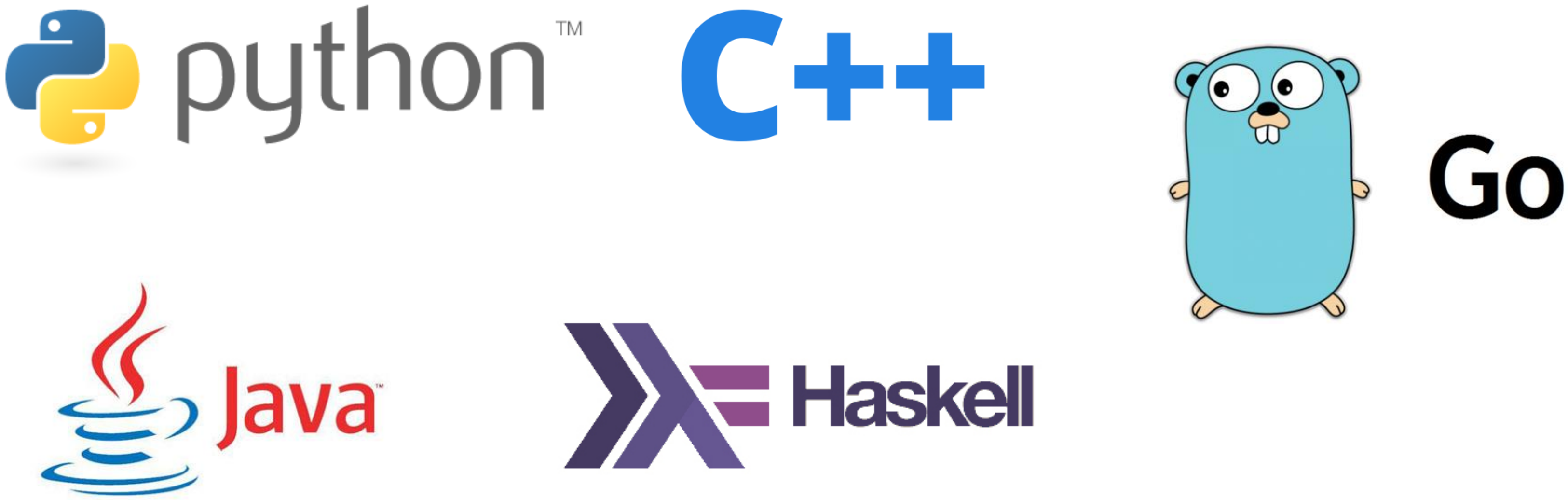
TPU

iOS

Android

TensorFlow Introduction -- Supports Many Languages

TensorFlow: Machine Learning for Everyone



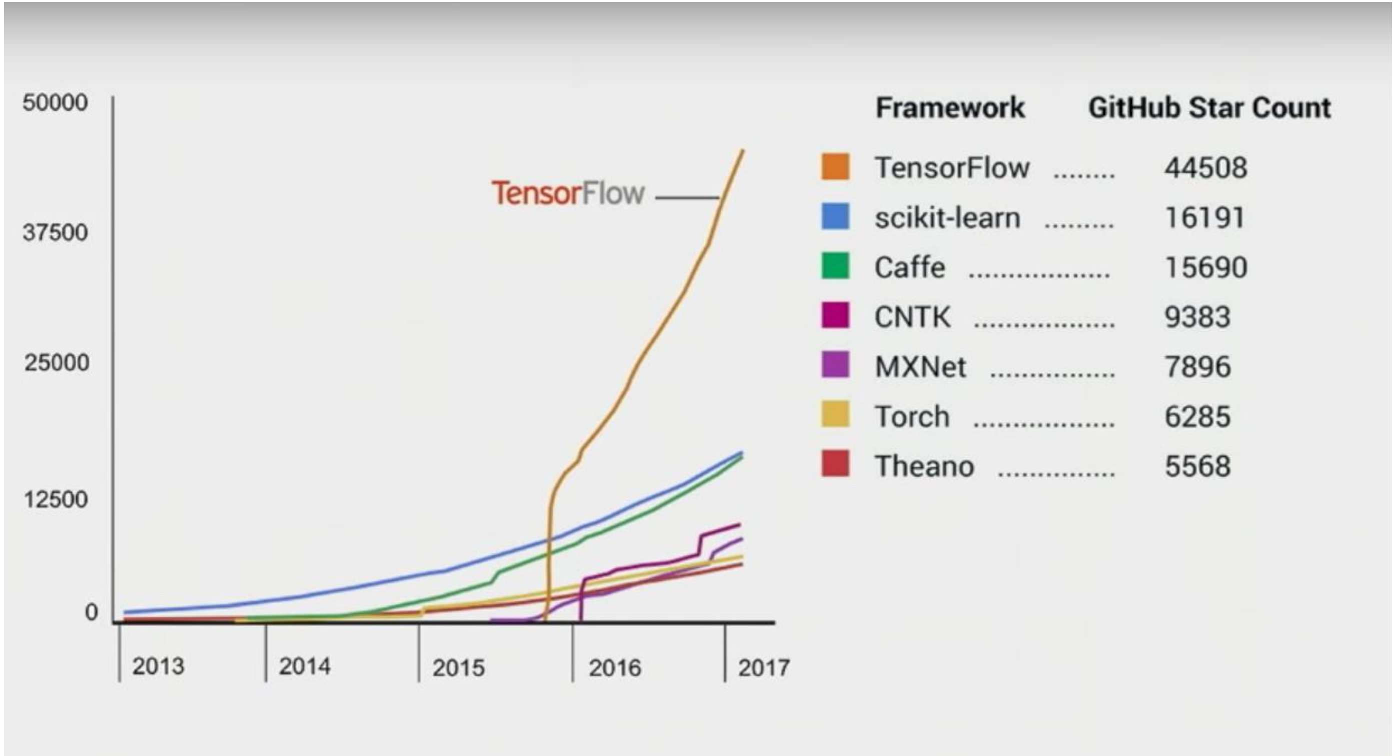
TensorFlow Introduction -- Companies using TensorFlow

TensorFlow: Machine Learning for Everyone

airbnb
AIRBUS DEFENCE & SPACE
ARM
CEVA
CI&T
DeepMind
Dropbox
ebay
Google
IBM Power Systems
intel
JD.COM 京东
mi
Movidius
QUALCOMM
quantiphi
SAP
UBER

TensorFlow Introduction -- GitHub

TensorFlow: Machine Learning for Everyone



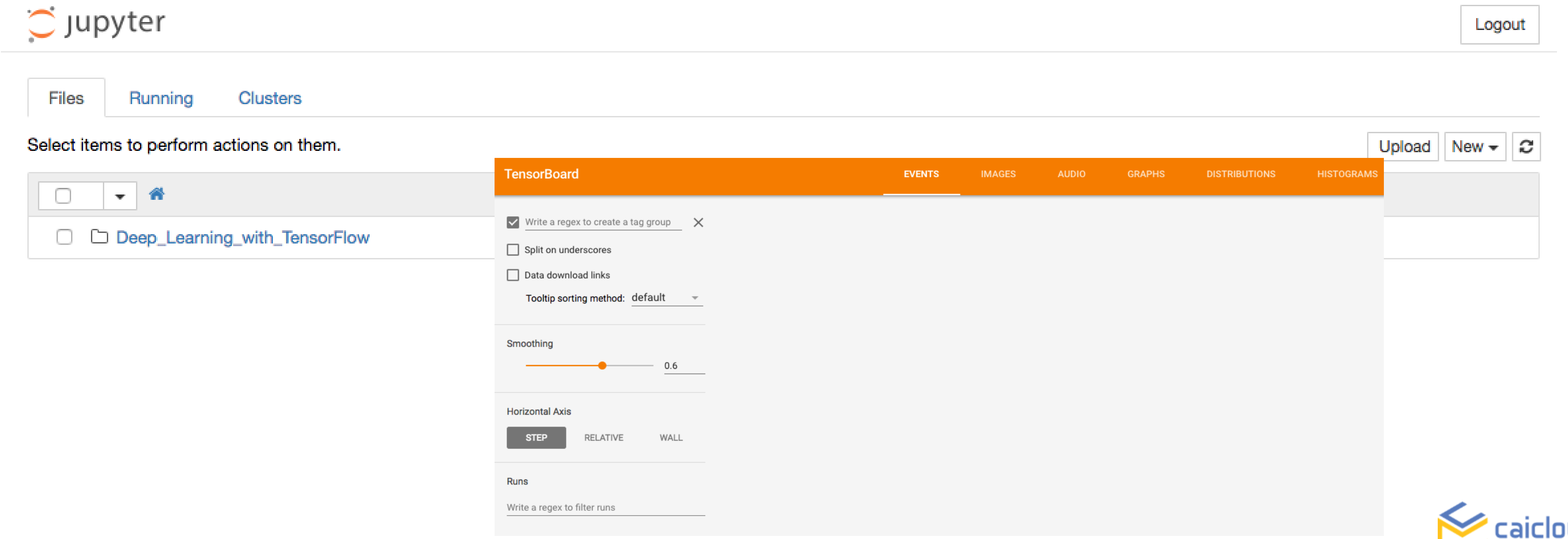
	Framework	GitHub Star Count
🟧	TensorFlow	44508
🟦	scikit-learn	16191
🟩	Caffe	15690
🟪	CNTK	9383
🟪	MXNet	7896
🟨	Torch	6285
🟥	Theano	5568

TensorFlow Introduction -- run TensorFlow on Docker

- docker run -p 8888:8888 -p 6006:6006 cargo.caicloud.io/tensorflow/tensorflow:1.0.0

Jupyter Editor Port

TensorBoard Port



Select items to perform actions on them.

□ Deep_Learning_with_TensorFlow

☑ Write a regex to create a tag group

□ Split on underscores

□ Data download links

Tooltip sorting method: default

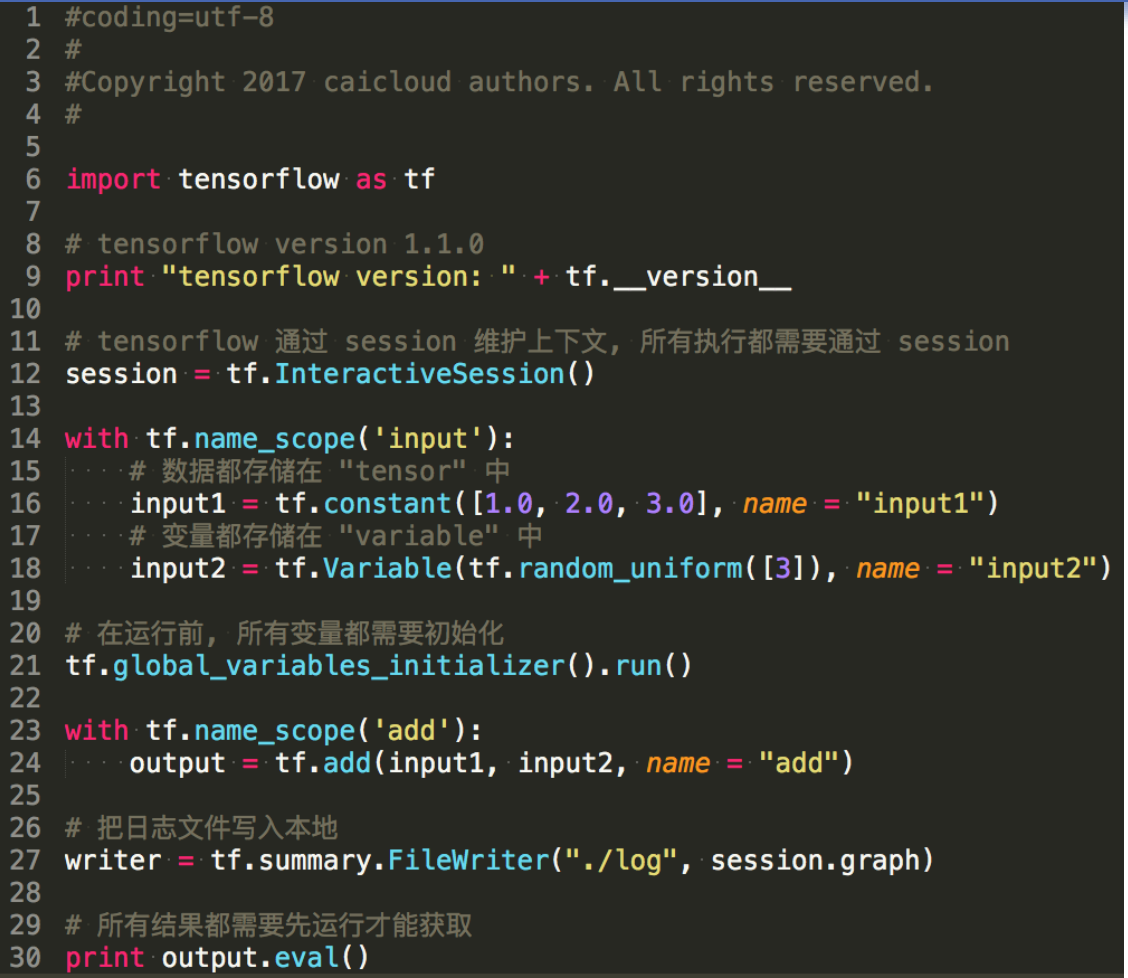
Smoothing 0.6

Horizontal Axis: STEP RELATIVE WALL

Runs

Write a regex to filter runs

TensorFlow Introduction -- Hello World



```python
#coding=utf-8
#
#Copyright 2017 caicloud authors. All rights reserved.
#

import tensorflow as tf

# tensorflow version 1.1.0
print "tensorflow version: " + tf.__version__

# tensorflow 通过 session 维护上下文，所有执行都需要通过 session
session = tf.InteractiveSession()

with tf.name_scope('input'):
    # 数据都存储在 "tensor" 中
    input1 = tf.constant([1.0, 2.0, 3.0], name = "input1")
    # 变量都存储在 "variable" 中
    input2 = tf.Variable(tf.random_uniform([3]), name = "input2")

# 在运行前，所有变量都需要初始化
tf.global_variables_initializer().run()

with tf.name_scope('add'):
    output = tf.add(input1, input2, name = "add")

# 把日志文件写入本地
writer = tf.summary.FileWriter("./log", session.graph)

# 所有结果都需要先运行才能获取
print output.eval()
```

session

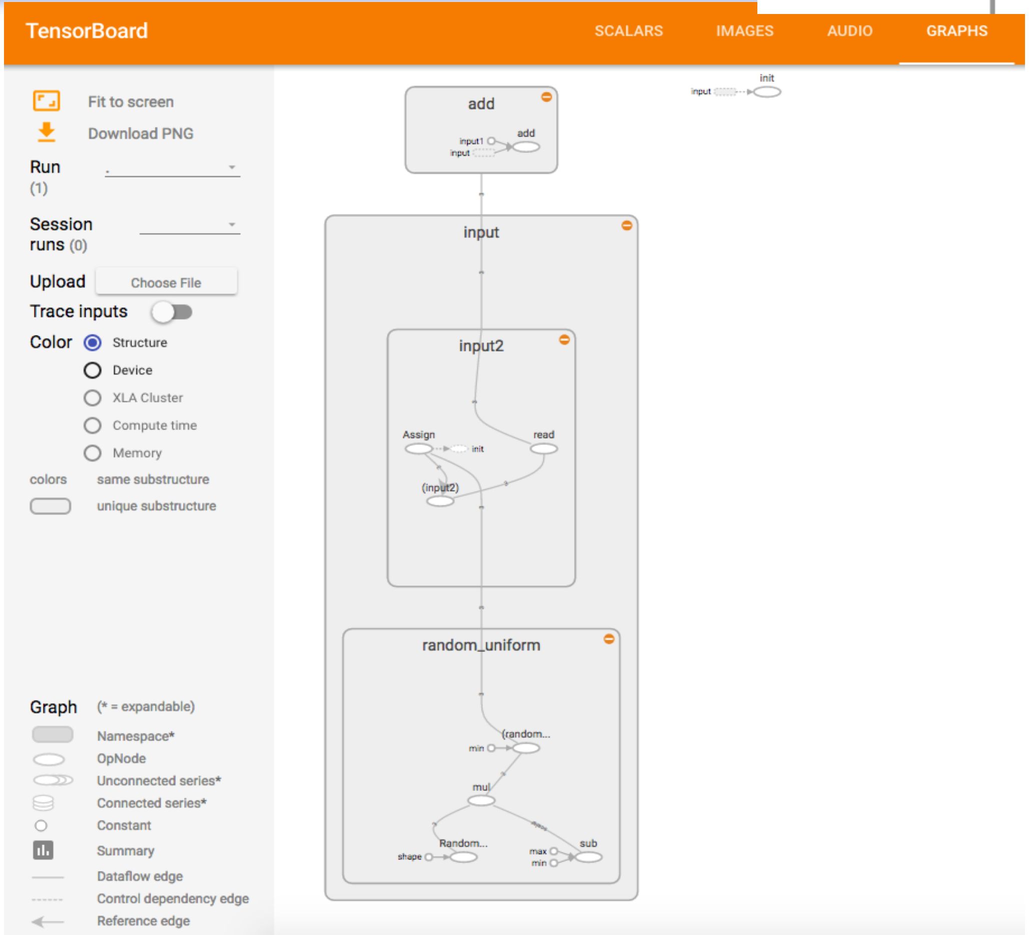
name_scope

constant

variable

TensorFlow Introduction -- TensorBoard

```
$ tensorboard --logdir ./log/
Starting TensorBoard at http://0.0.0.0:6006
```



TensorBoard   SCALARS   IMAGES   AUDIO   GRAPHS

Fit to screen

Download PNG

Run (1)

Session runs (0)

Upload   Choose File

Trace inputs

Color   ● Structure
○ Device
○ XLA Cluster
○ Compute time
○ Memory

colors   same substructure
unique substructure

Graph   (* = expandable)

Namespace*

OpNode

Unconnected series*

Connected series*

Constant

Summary

Dataflow edge

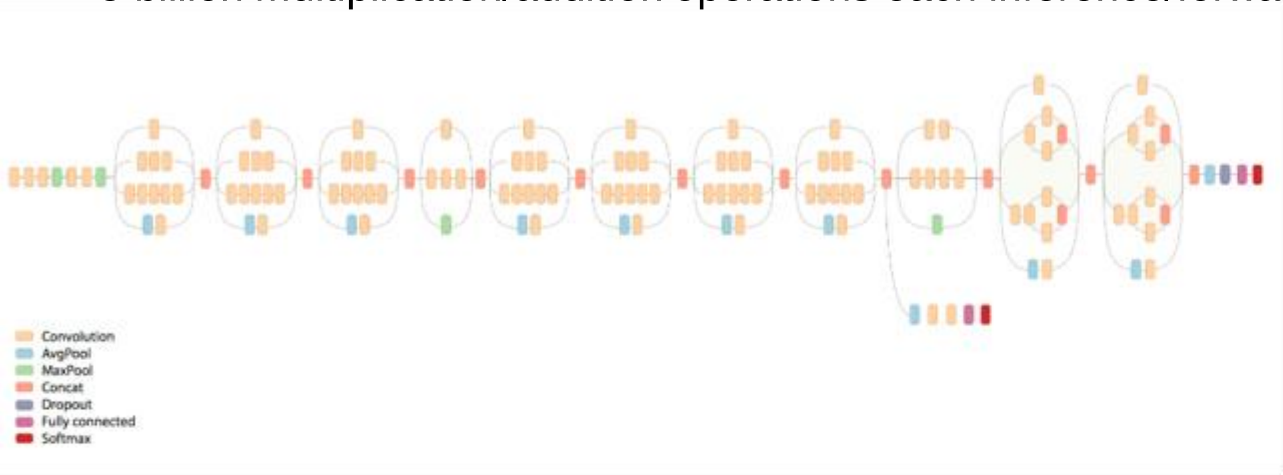
Control dependency edge

Reference edge

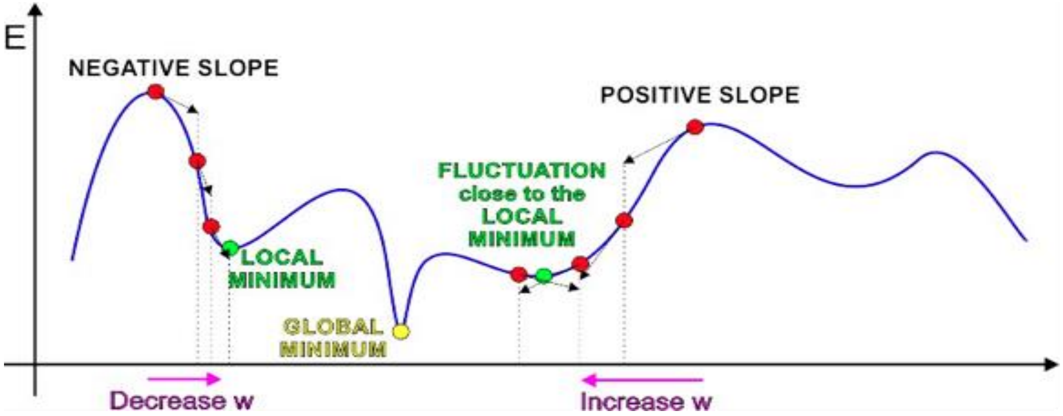
- There is no such thing as a free lunch -- Computation Problem

Inception-v3 model for ImageNet

- 25 million parameters
- 5 billion multiplication/addition operations each inference/forward-prorogate



- There is no such thing as a free lunch -- Optimizing Problem
  - Structure of neural network is complex, it is difficult to directly solve
  - Iterative optimization algorithm -- gradient descent method
  - Need massive data and massive computation
  - It takes **six months** to reach 78% accuracy on single-machine

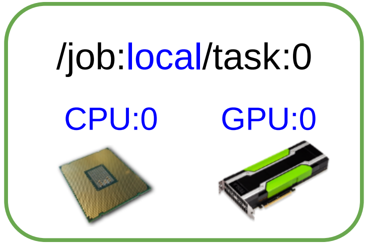


Distributed TensorFlow on Kubernetes

- on local machine

```
with tf.device("/cpu:0"):  // 参数
    var1 = tf.Variable(...)
    var2 = tf.Variable(...)
with tf.device("/gpu:0"):  // 计算
    output = tf.matmul(input + var1) + var2
    loss = loss_function(output)
```

client



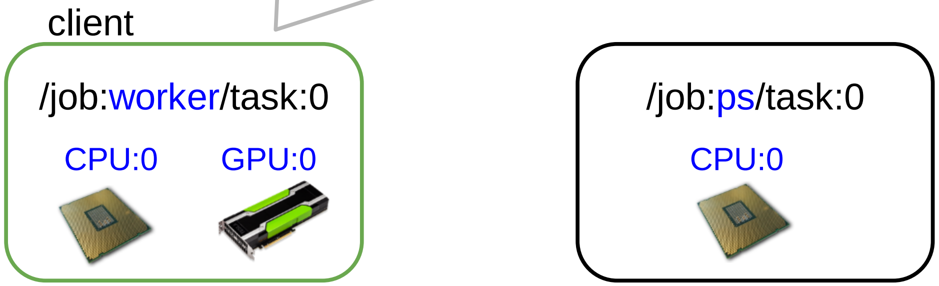
/job:local/task:0

CPU:0 GPU:0

Distributed TensorFlow on Kubernetes

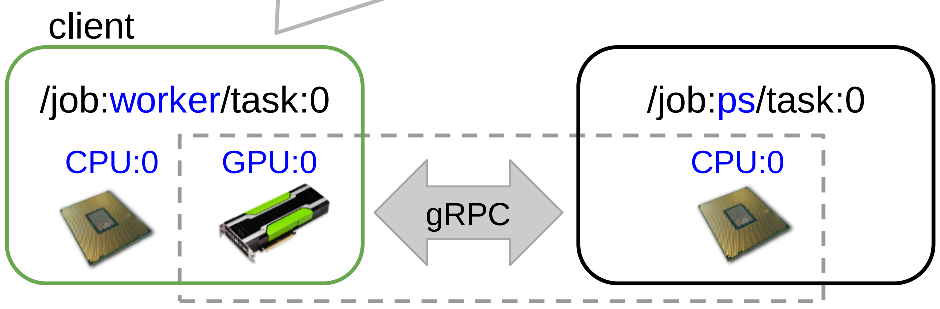
- on a cluster of servers (1 Worker + 1 PS)

```
with tf.device("/job:ps/task:0/cpu:0"):  // 参数
    var1 = tf.Variable(...)
    var2 = tf.Variable(...)
with tf.device("/job:worker/task:0/gpu:0"):  // 计算
    output = tf.matmul(input + var1) + var2
    loss = loss_function(output)
```



- on a cluster of servers (1 Worker + 1 PS)

```
with tf.device("/job:ps/task:0/cpu:0"):  // 参数
    var1 = tf.Variable(...)
    var2 = tf.Variable(...)
with tf.device("/job:worker/task:0/gpu:0"):  // 计算
    output = tf.matmul(input + var1) + var2
    loss = loss_function(output)
```



client

/job:worker/task:0

CPU:0

GPU:0

gRPC

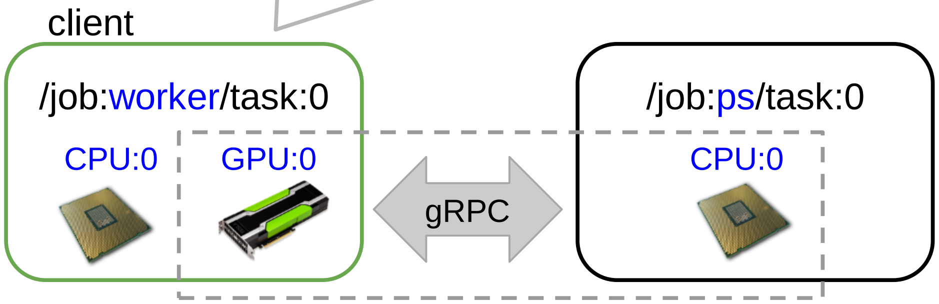
/job:ps/task:0

CPU:0

Distributed TensorFlow on Kubernetes

- on a cluster of servers (1 Worker + 1 PS)

```
with tf.device("/job:ps/task:0/cpu:0"):  // 参数
    var1 = tf.Variable(...)
    var2 = tf.Variable(...)
with tf.device("/job:worker/task:0/gpu:0"):  // 计算
    output = tf.matmul(input + var1) + var2
    loss = loss_function(output)
```



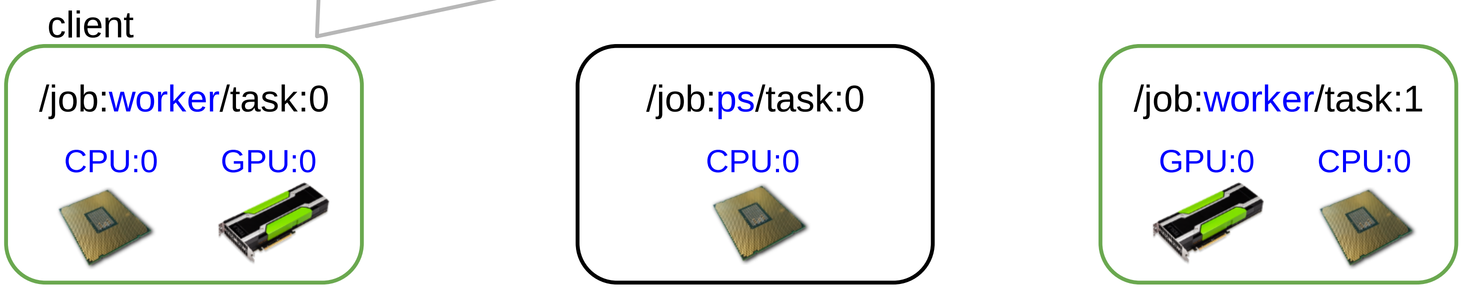
- PS task
  - Variables
  - Update parameters

- Worker task
  - Pre-processing
  - Loss calculation
  - Back-propagation

Distributed TensorFlow on Kubernetes

- on a cluster of servers (2 Workers + 1 PS)

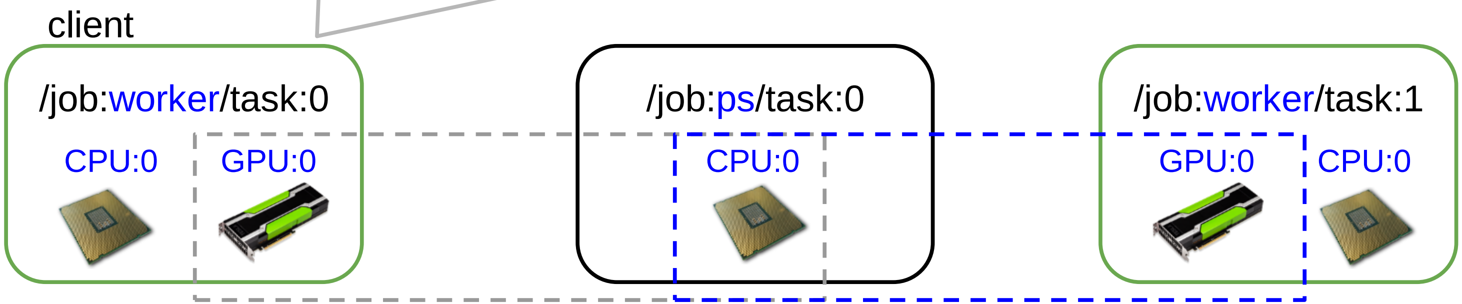
```
with tf.device("/job:ps/task:0/cpu:0"):  // 参数
    var1 = tf.Variable(...)
    var2 = tf.Variable(...)
inputs = tf.split(0, num_workers, input)
outputs = []
for i in range(num_workers)
    with tf.device("/job:worker/task:%d/gpu:0" % i):  // 计算
        outputs.append(tf.matmul(input + var1) + var2)
loss = loss_function(output)
```



Distributed TensorFlow on Kubernetes

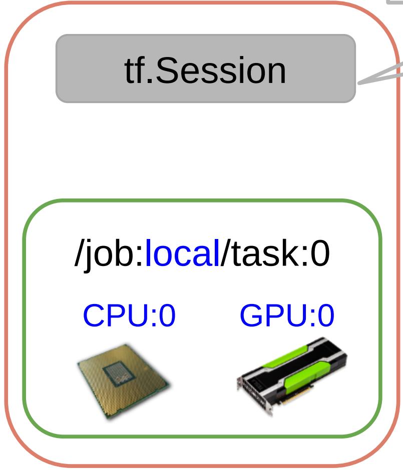
- on a cluster of servers (2 Workers + 1 PS)

```
with tf.device("/job:ps/task:0/cpu:0"):  // 参数
    var1 = tf.Variable(...)
    var2 = tf.Variable(...)
inputs = tf.split(0, num_workers, input)
outputs = []
for i in range(num_workers)
    with tf.device("/job:worker/task:%d/gpu:0" % i):  // 计算
        outputs.append(tf.matmul(input + var1) + var2)
loss = loss_function(output)
```

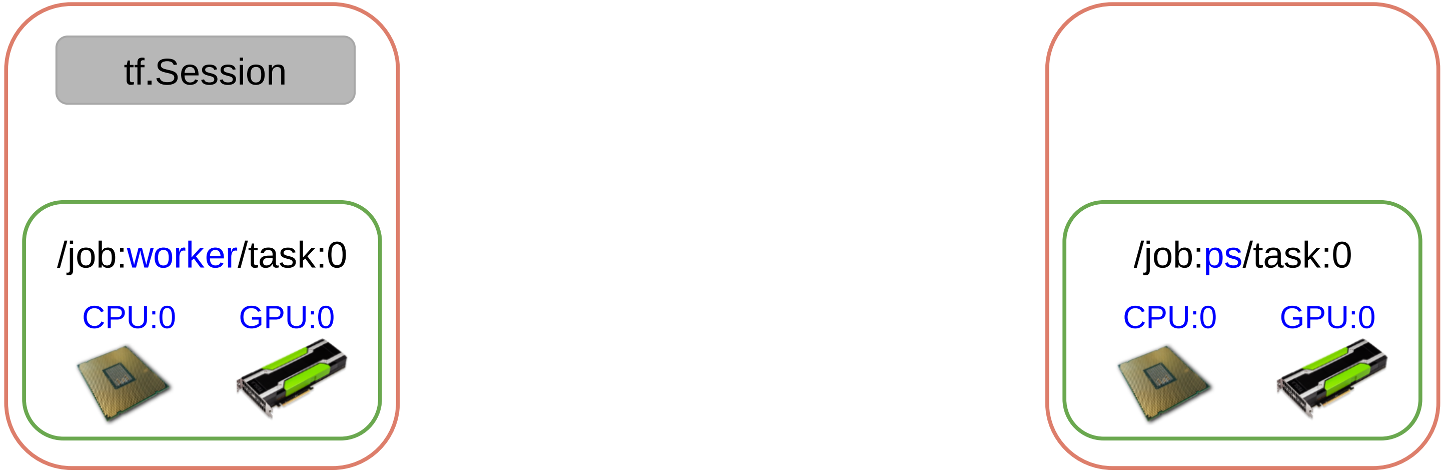


- runs on local machine

```
with tf.Session() as sess  // 声明 session
    sess.run(init_operation)  // 初始化
    for _ in range(NUM_STEPS)  // 训练轮数
        sess.run(train_operation)  // 训练
```

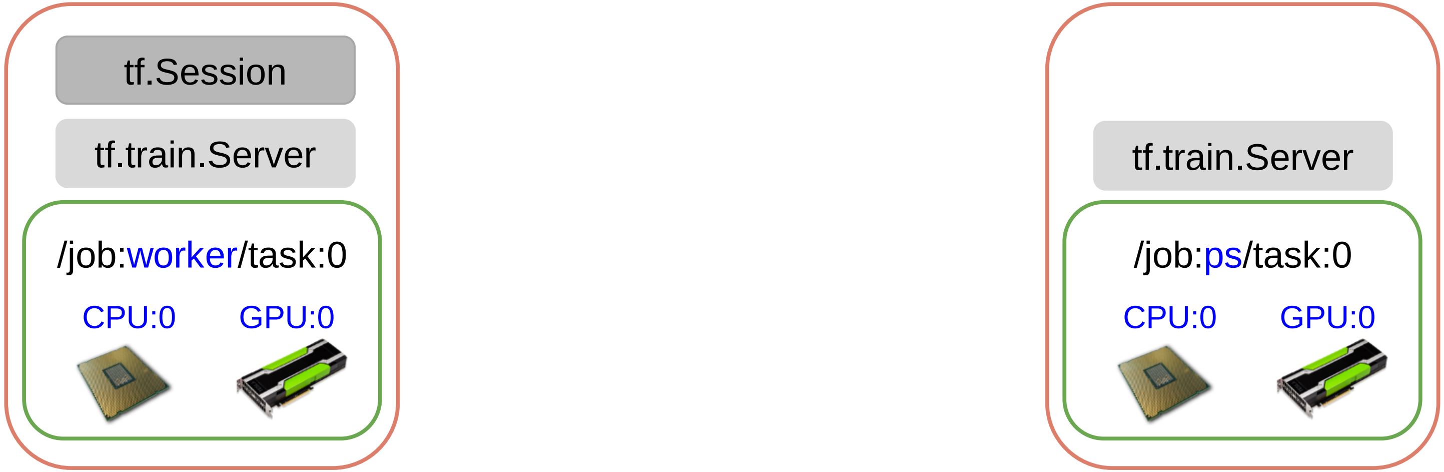


- runs on a cluster of servers



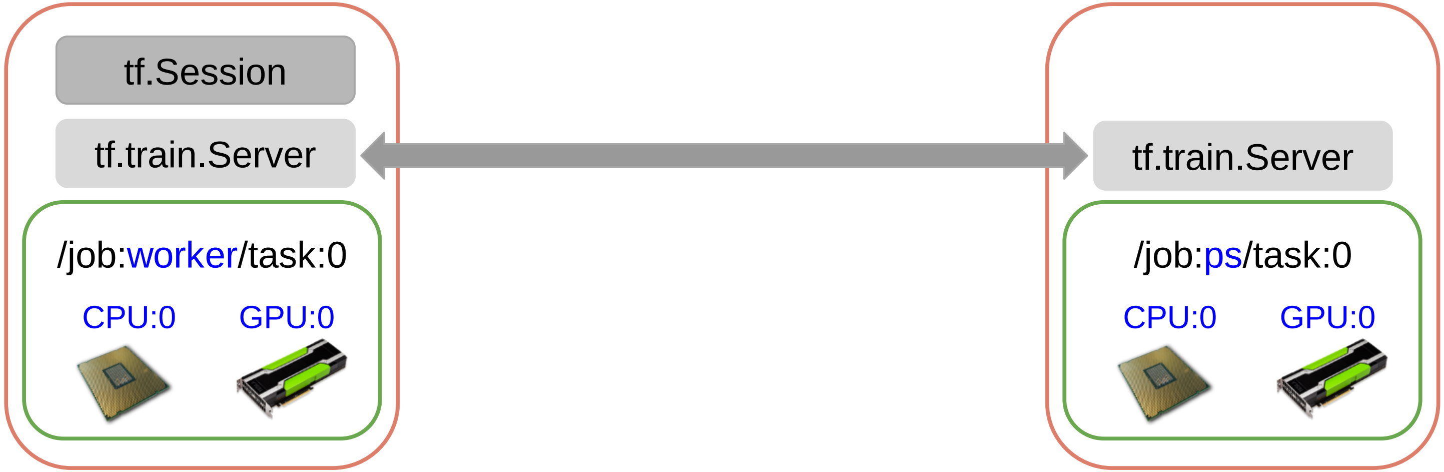
Distributed TensorFlow on Kubernetes

- runs on a cluster of servers



Distributed TensorFlow on Kubernetes

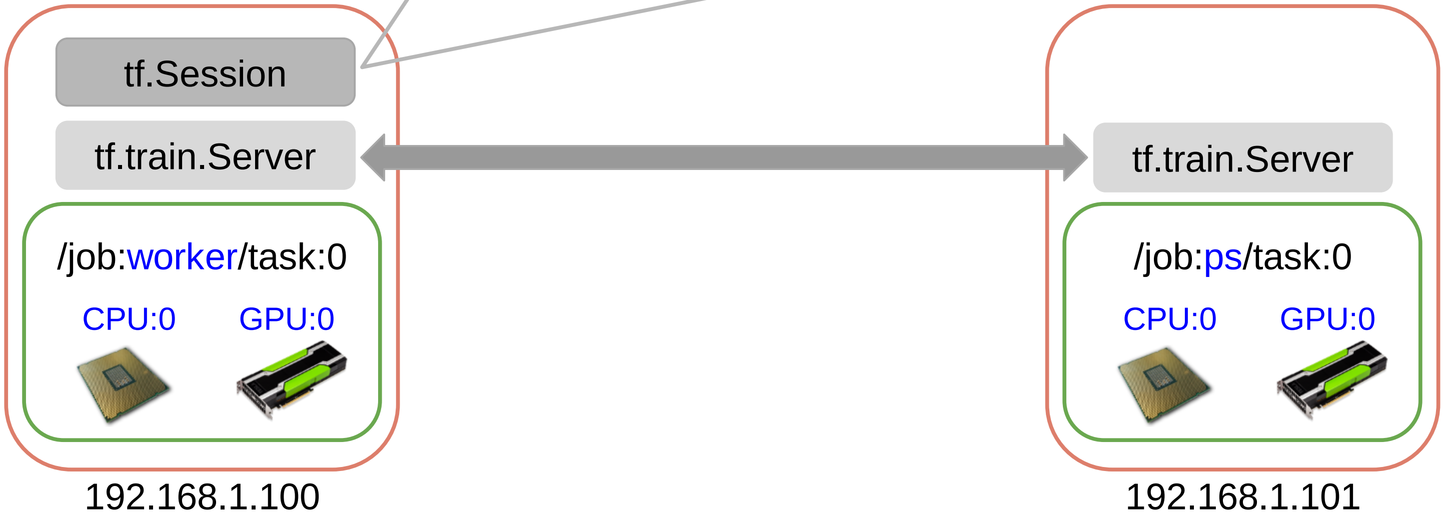
- runs on a cluster of servers



Distributed TensorFlow on Kubernetes

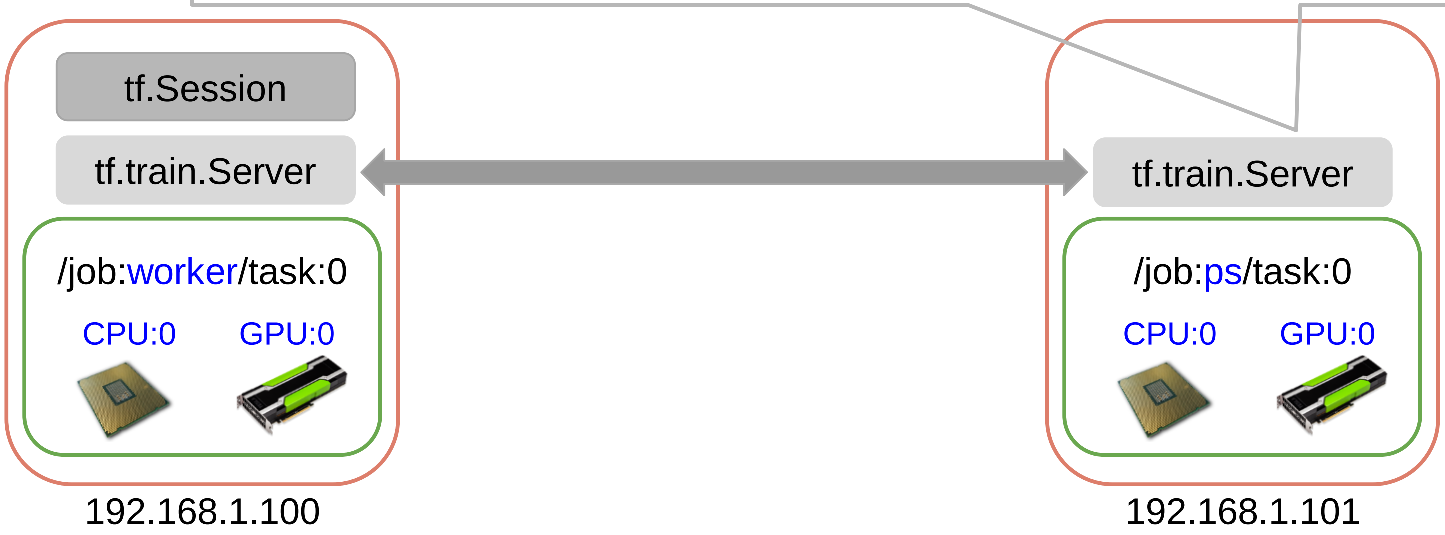
- runs on a cluster of servers

```
# in worker 0
clusterSpec = tf.train.ClusterSpec({        // 声明 cluster spec
    "worker": ["192.168.1.100:2222", ...],
    "ps": ["192.168.1.101:2222", ...]})
server = tf.train.Server(clusterSpec, job_name = "worker", task_index = 0)
with tf.Session(server.target) as sess  // 传入 server 获得 session
    # do something ...
```



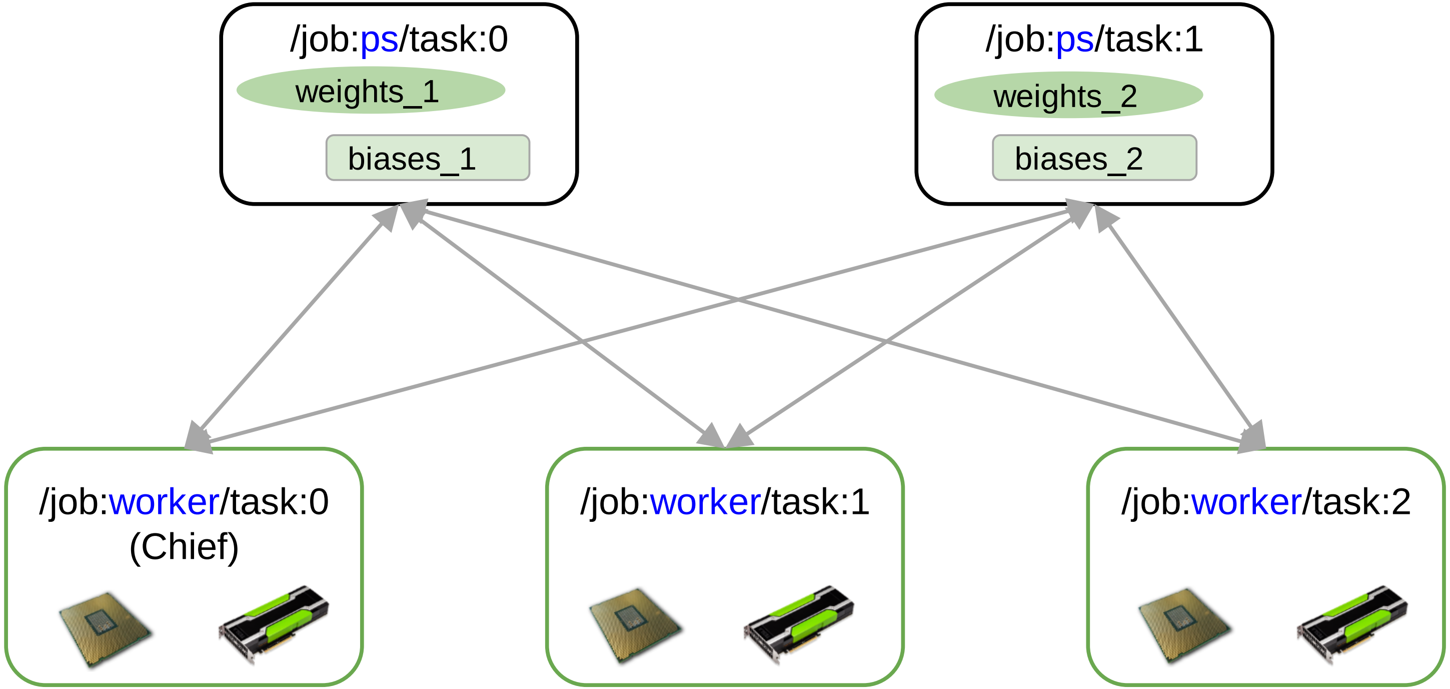
- runs on a cluster of servers

```
# in ps 0
clusterSpec = tf.train.ClusterSpec({        // 声明 cluster spec
    "worker": ["192.168.1.100:2222", ...],
    "ps": ["192.168.1.101:2222", ...]})
server = tf.train.Server(clusterSpec, job_name = "ps", task_index = 0)
 # wait for incoming connections …
server.join()     // 等待连接...
```

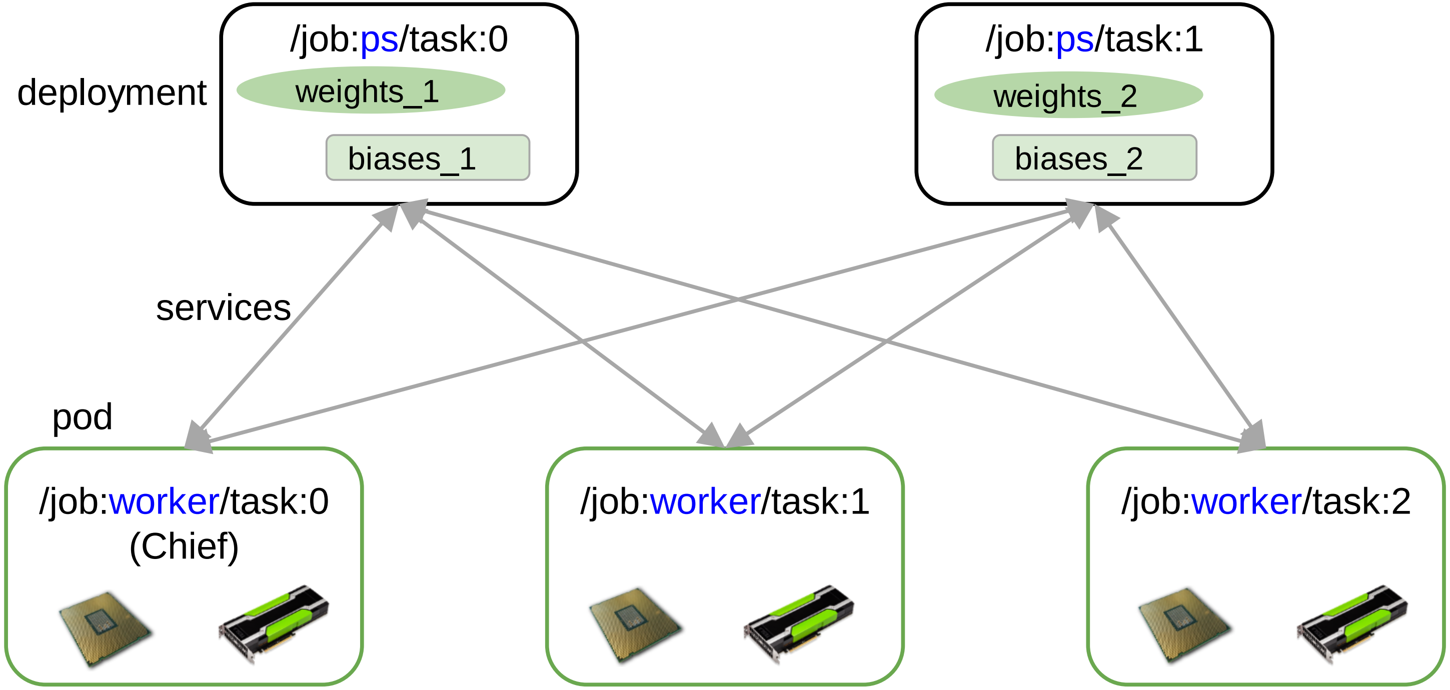


Distributed TensorFlow on Kubernetes

- on kubernetes



- on kubernetes



deployment

/job:ps/task:0

weights_1

biases_1

/job:ps/task:1

weights_2

biases_2

services

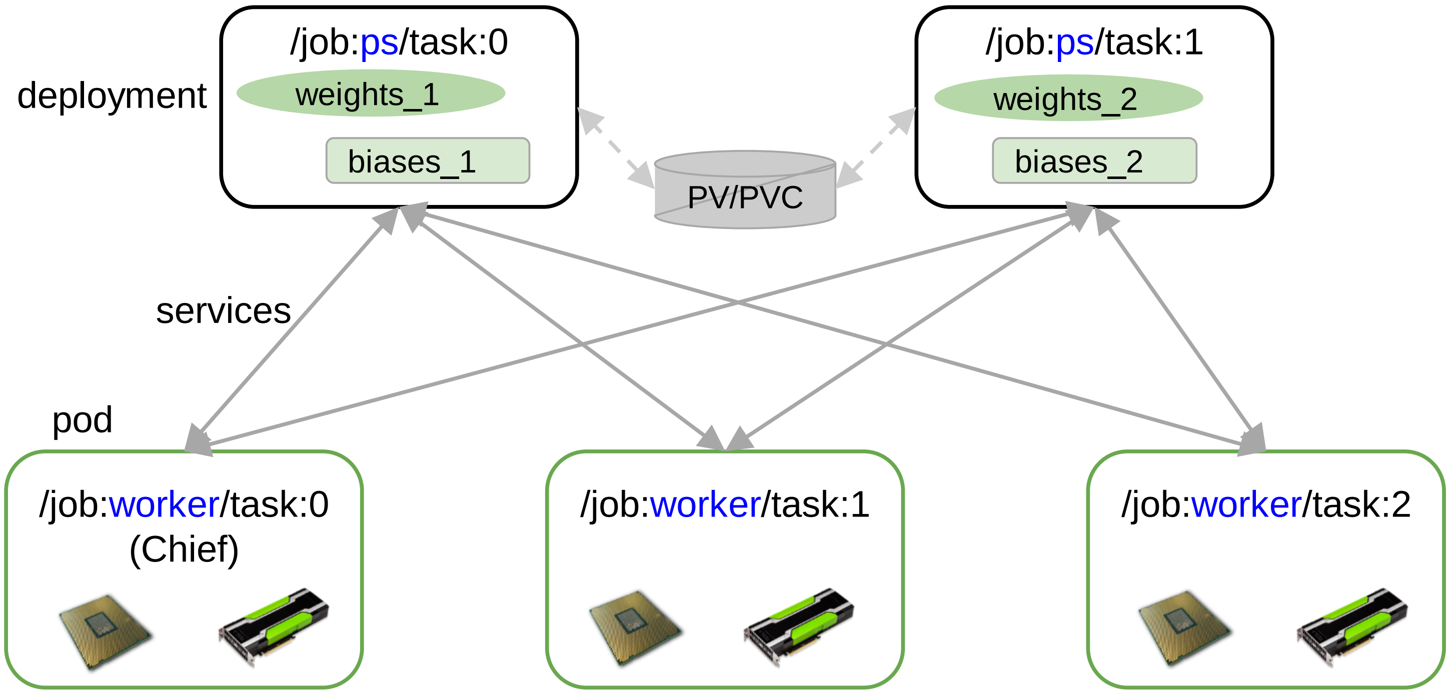
pod

/job:worker/task:0 (Chief)

/job:worker/task:1

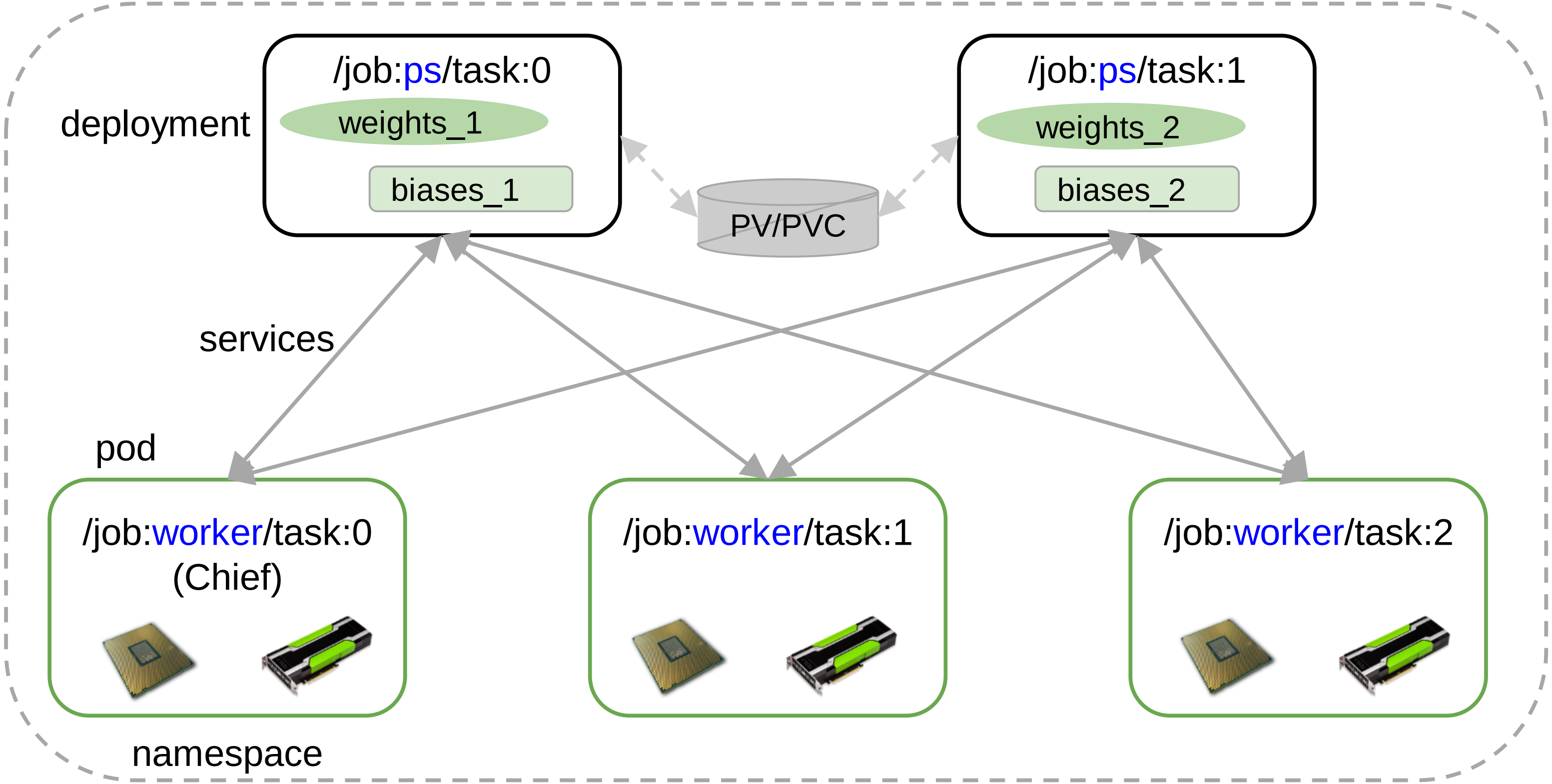
/job:worker/task:2

- on kubernetes



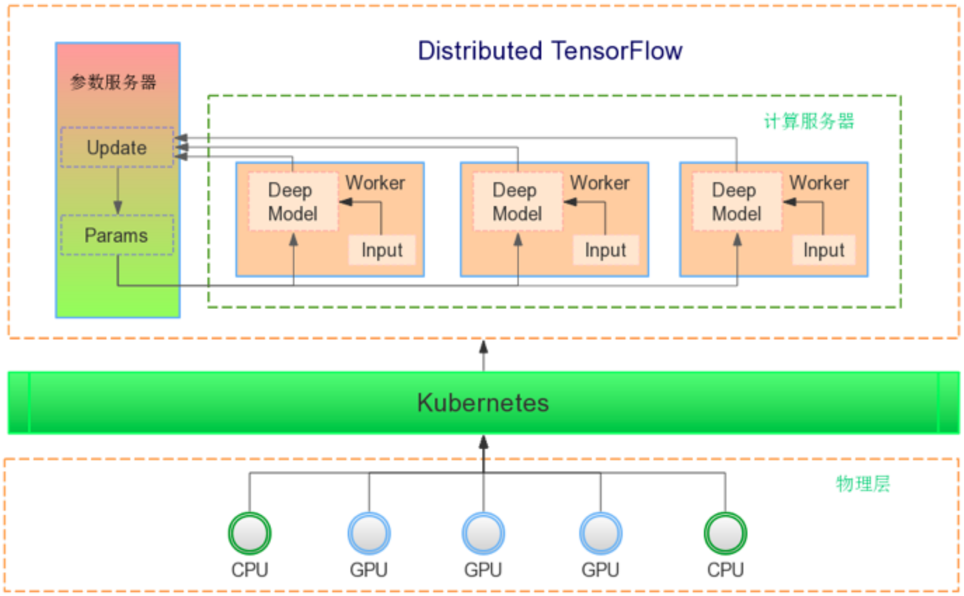
Distributed TensorFlow on Kubernetes

- on kubernetes



Distributed TensorFlow on Kubernetes

- on kubernetes



Cluster Management

- **Deployment/pod** for life cycle management
- **Namespace** for resource isolation
- Monitoring、alerting & logging

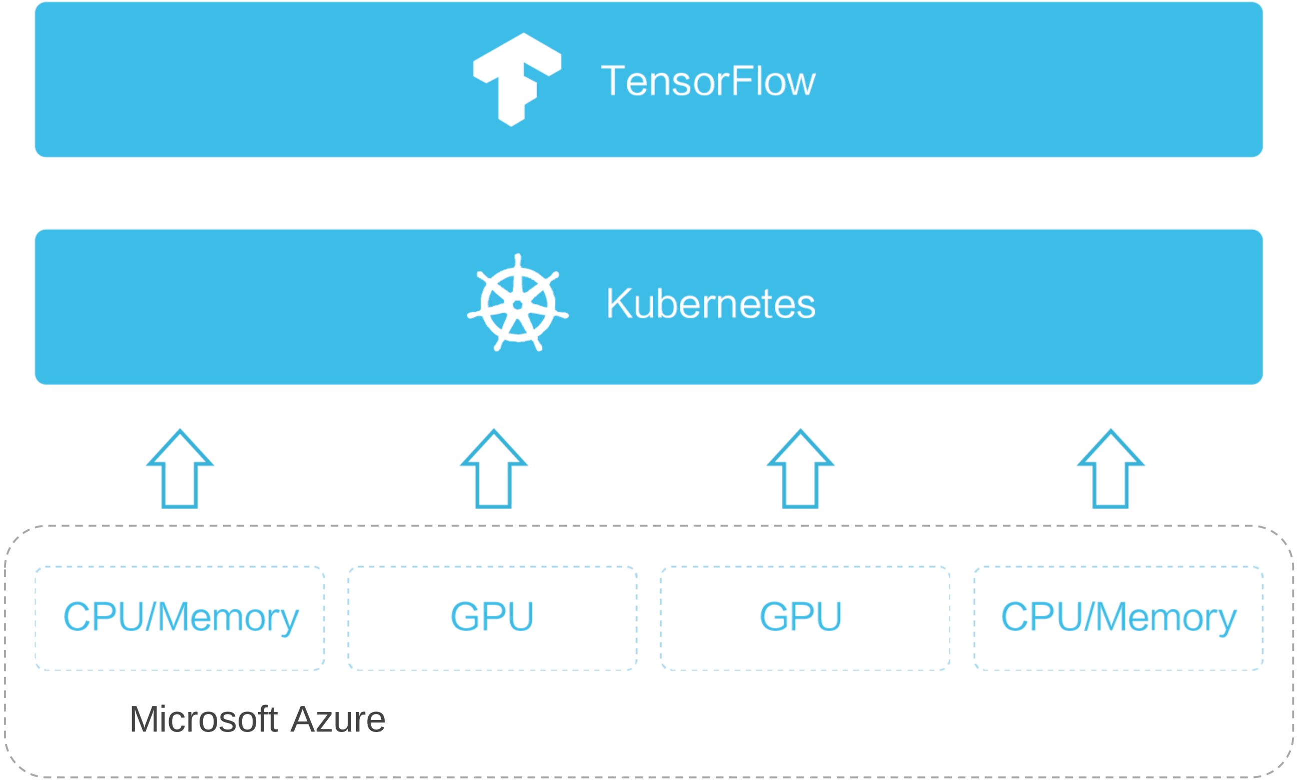
Network

- **Services** for service discovery

Storage

- **PV/PVC** for persistent storage
- GlusterFS、ceph for distributed storage

TaaS Introduction -- TensorFlow as a Service



TaaS Demo

https://taas.caicloud.io

Q&A

Q&A

网易云
163yun.com

IT大咖说
知识分享平台

caicloud
才云

Thanks！