

Building AI Platform Based on Kubernetes and TensorFlow

阿里云容器服务 王涛

主办方：



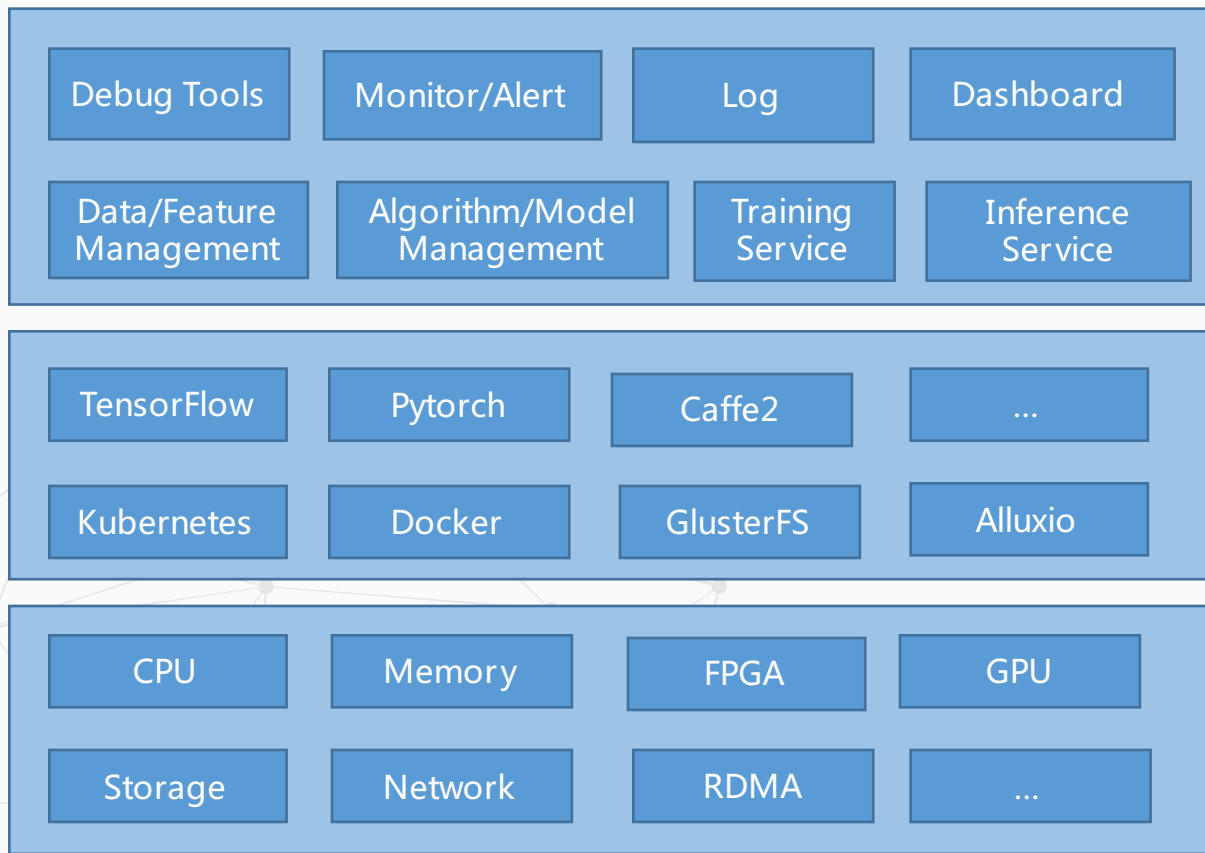
协办方：





TensorFlow on Kubernetes

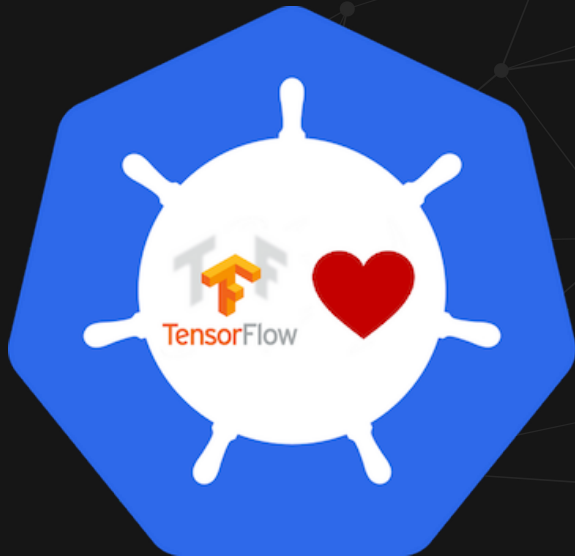
AI Platform Based on Kubernetes



Why TensorFlow on Kubernetes ?



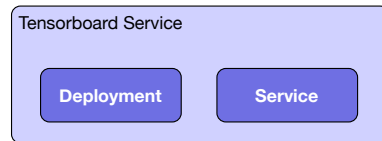
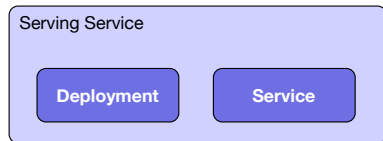
- Task-Level 资源隔离 ;
- 资源调度(GPU) ;
- 大规模任务管理 ;
- 训练日志 ;
- 大规模集群的资源管理 ;
- Serving的持续部署和弹性伸缩 ;



TensorFlow on Kubernetes



Kubernetes Cluster



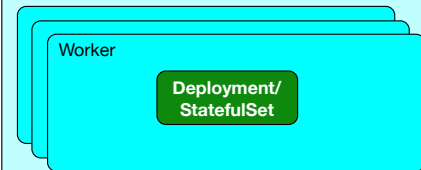
TensorFlow Distributed Training Cluster (PS Mode)

namespace



TensorFlow Distributed Training Cluster (MPI Mode)

namespace



需要注意的问题



- 问题worker，无限重启，Kubernetes 1.7；
- 大规模训练集群，KubeDNS性能引发的TensorFlow Session问题；
- 做好集群防雪崩保护措施；
- 清理历史训练任务（namespace→deployment/job→pod→container）；
- 创建训练集群前，检查集群剩余可用资源（Mutex）；
- 不可忽视的grpc分布式通信性能损耗，合理规划task number；
- ...

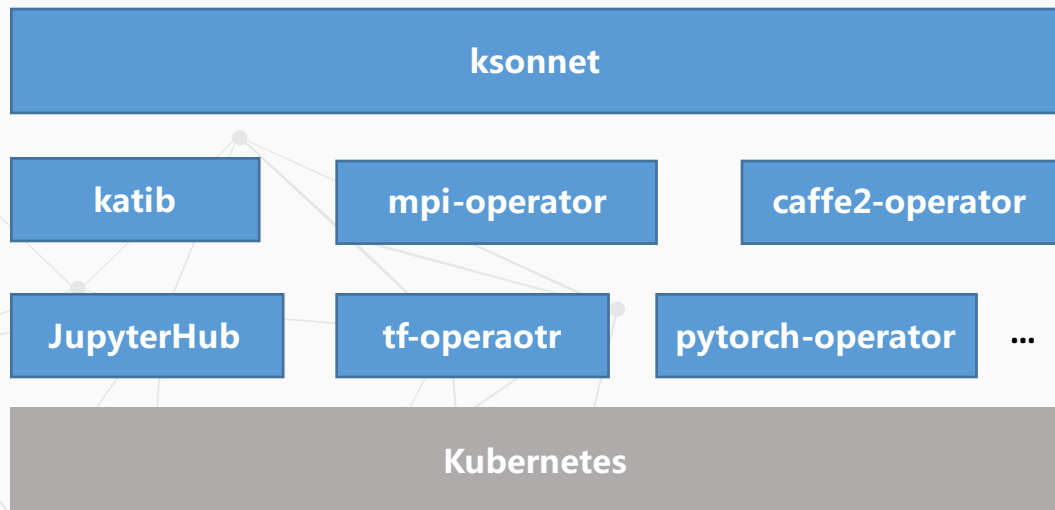


Introduction to Kubeflow

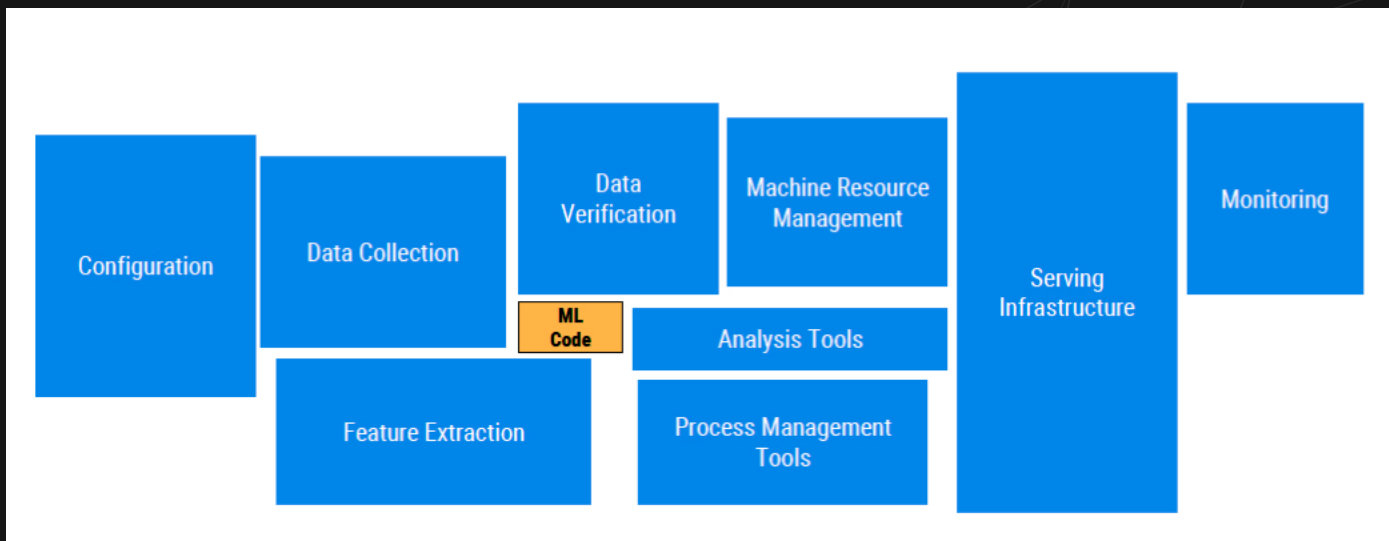
What is Kubeflow



- The Kubeflow project is dedicated to making deployments of machine learning (ML) workflows on Kubernetes simple, portable and scalable.



Kubeflow – ML Toolkit for K8S

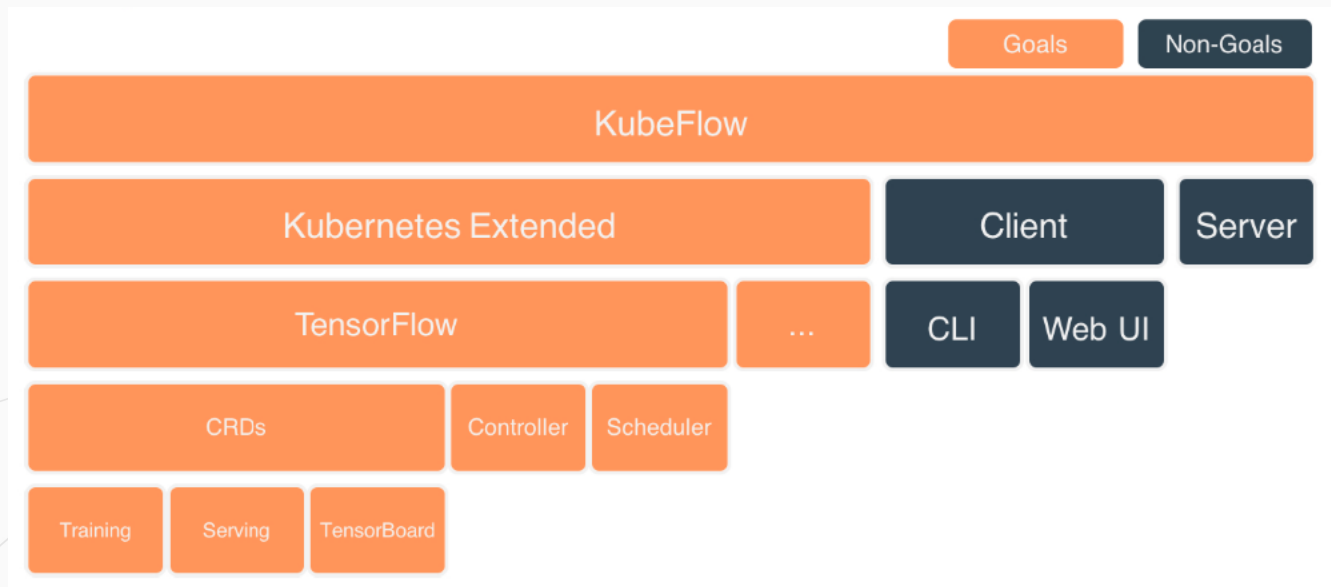


[PR: define cleanup policy](<https://github.com/kubeflow/tf-operator/pull/685>)



Arena-Alibaba Cloud DL Tools for Kubernetes

为什么需要Arena



What is Arena



➤ 用一个工具屏蔽所有底层资源、环境管理和任务调度的复杂性

- 深提供数据科学家最熟悉的使用体验
- 兼容多种深度学习框架 – Tensorflow, Caffe, MPI, Hovorod, Pytorch

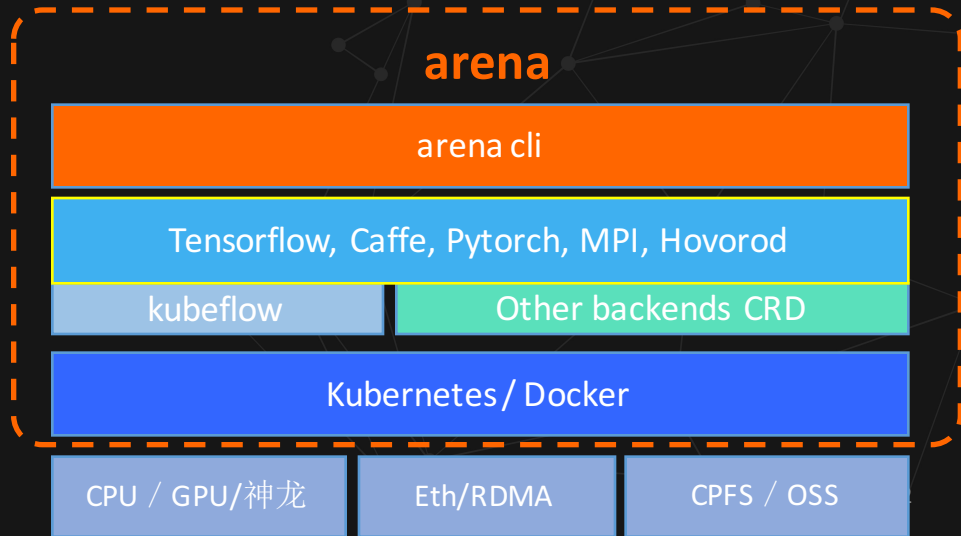
➤ 端到端工作流

- 深度学习生产流水线 – 集成训练数据管理，实验管理，模型开发，持续训练、评估，上线预测等全流程

➤ 轻量级、弹性、高效

➤ 支持实时训练运维

- 支持CPU，GPU等资源监控
- 可查看GPU资源、任务分配情况
- 可查看实时训练日志



What Can Arena Do



提交训练任务

arena submit

--name=myhvd \

--mode=horovod \

--workers=3 \

--gpus=2 \

--sshPort=33 \

--env=num_batch=100 \

--env=batch_size=80 \

--image=registry.cn-hangzhou.aliyuncs.com/tensorflow-samples/ali-horovod:gpu-tf-1.6.0 \

"/root/hvd-distribute.sh 6 2"

查看任务列表

arena list

查看任务详情

arena get jobname

查看任务日志

arena logs jobname

arena logviewer jobname

查看集群资源分配情况

arena top node -d

Arena Show



```
arena submit --name=tf-dist \
  --gpu=1 \
  --workerCpu=2 \
  --workerMemory=4Gi \
  --workerPort=22222 \
  --mode=tensorflow \
  --workers=2 \
  --workerImage=registry-vpc.cn-zhangjiakou.aliyuncs.com/tensorflow-samples/tf-worker:1.0.2-gpu \
  --psPort=22223 \
  --ps=2 \
  --workingDir=/root \
  --psImage=registry-vpc.cn-zhangjiakou.aliyuncs.com/tensorflow-samples/tf-ps:1.0.2 \
  --psCpu=1 \
  --psMemory=2Gi \
  "python /var/tf_dist_mnist/dist_mnist.py --data_dir=/mnist_data"
```

Arena Show



```
arena list
```

NAME	STATUS	TRAINER	AGE	NODE
[REDACTED]	FAILED	STANDALONE	4h	N/A
[REDACTED] inception-73-1	FAILED	STANDALONE	5h	N/A
tf-dist	SUCCEEDED	TensorFlow	5h	N/A
k8s-logotest-60	SUCCEEDED	STANDALONE	5d	N/A
[REDACTED] spam-07112	FAILED	STANDALONE	5d	N/A
[REDACTED]	SUCCEEDED	MPI	5d	N/A
mytest1001	SUCCEEDED	STANDALONE	5d	N/A
[REDACTED]	RUNNING	MPI	5d	[REDACTED]
[REDACTED]	SUCCEEDED	MPI	5d	N/A
[REDACTED]	SUCCEEDED	MPI	5d	N/A

```
# arena get tf-tensorboard
```

NAME	STATUS	TRAINER	AGE	INSTANCE	NODE
tf-tensorboard	RUNNING	TensorFlow	3h	tf-tensorboard-tfjob-5646449556-kf27w	192.168.1.119
tf-tensorboard	SUCCEEDED	TensorFlow	3h	tf-tensorboard-tfjob-worker-0	N/A

```
TENSORBOARD URL:  
192.168.1.117:30212
```

Arena Show



arena top node

NAME	IPADDRESS	ROLE	GPU(Total)	GPU(Allocated)
cn-shanghai.i- 	192.168.168.120	master	0	0
cn-shanghai.i- 	192.168.168.124	worker	2	2
cn-shanghai.i- 	192.168.168.123	worker	2	0
cn-shanghai.i- 	192.168.168.125	worker	2	2
cn-shanghai.i- 	192.168.168.133	worker	2	2
cn-shangha 	192.168.168.122	master	0	0
cn-shanghai.i- 	192.168.168.121	master	0	0

Allocated/Total GPUs In Cluster:

6/8 (75%)

Arena Show



```
arena top node -d
```

```
NAME:      cn-shanghai.i-uf61psgsvkxjp9o64gua
IPADDRESS: 192.168.168.120
ROLE:      master
```

```
Total GPUs In Node cn-shanghai.i-uf61psgsvkxjp9o64gua:    0
Allocated GPUs In Node cn-shanghai.i-uf61psgsvkxjp9o64gua: 0 (0%)
```

```
NAME:      cn-shanghai.i-uf62q95pbpe10t76c4lm
IPADDRESS: 192.168.168.124
ROLE:      worker
```

NAMESPACE	NAME	GPU REQUESTS	GPU LIMITS
default	myhvd123-tf-horovod-1	2	2

```
Total GPUs In Node cn-shanghai.i-uf62q95pbpe10t76c4lm:    2
Allocated GPUs In Node cn-shanghai.i-uf62q95pbpe10t76c4lm: 2 (100%)
```



Particularity of The Scheduler in AI Platform

AI场景下的调度有何不同？



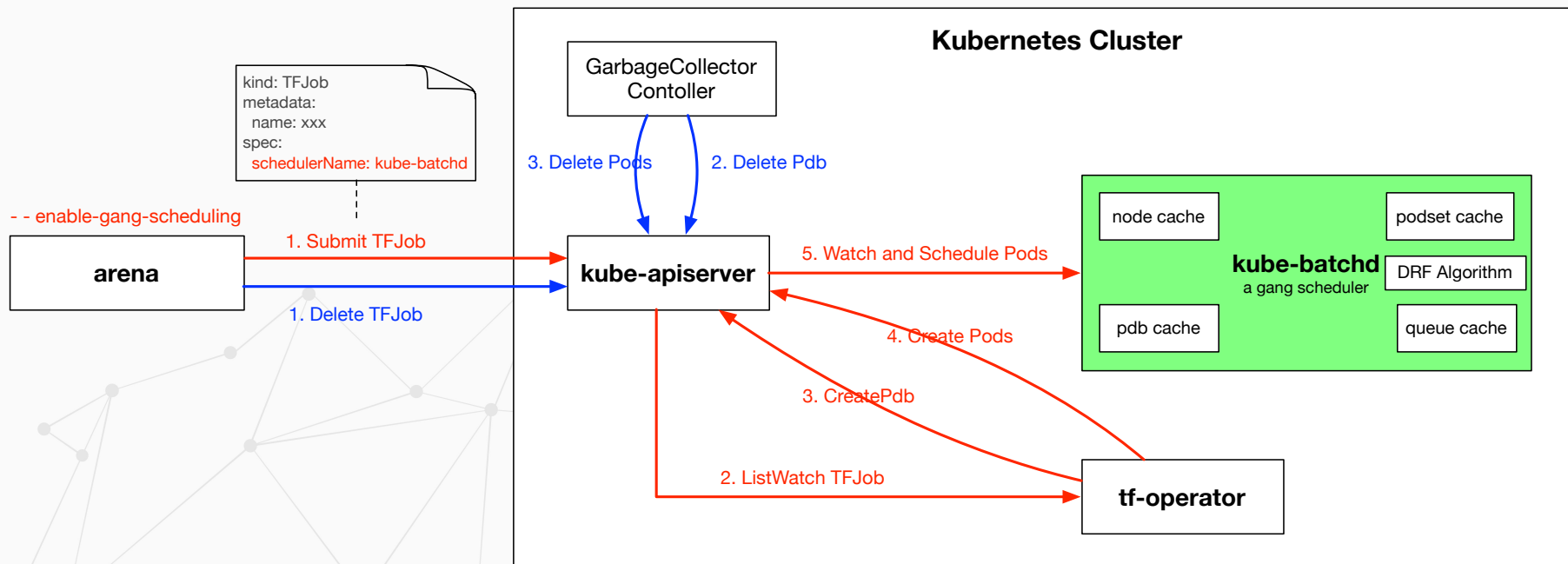
- **gang schedule (all or nothing) ;**
- multi-tenant task queue ;
- task priority ;
- preemption ;
- gpu affinity ;

kube-arbitrator



- kube-arbitrator is a policy based resource sharing system for Kubernetes cluster, it includes two sub projects: kube-batchd and kube-quotalloc.
- **kube-batchd**
 - gang scheduler
- kube-quotalloc
 - dynamic quota allocator
- <https://github.com/kubernetes-incubator/kube-arbitrator>

Create TFJob Workflow



Schedule TFJob with Kube-batchd



```
apiVersion: "kubeflow.org/v1alpha2"
kind: "TFJob"
metadata:
  name: "dist-mnist-for-e2e-test"
spec:
  tfReplicaSpecs:
    PS:
      replicas: 2
      restartPolicy: Never
      template:
        spec:
          schedulerName: kube-batchd
          containers:
            - name: tensorflow
              image: registry.cn-shenzhen.aliyuncs.com/walton/tf-dist-mnist-test:1.0
              resources:
                limits:
                  nvidia.com/gpu: "1"
    Worker:
      replicas: 8
      restartPolicy: Never
      template:
        spec:
          schedulerName: kube-batchd
          containers:
            - name: tensorflow
              image: registry.cn-shenzhen.aliyuncs.com/walton/tf-dist-mnist-test:1.0
              resources:
                limits:
                  nvidia.com/gpu: "1"
```

Schedule TFJob with Kube-batchd



- 感知gpu资源的伸缩；
- arena支持用户提交TFJob, MPIJob作业时使用kube-batchd进行gang schedule；

```
[root@iZwz91h5stiwetzn1plphZ dist-mnist]# kubectl get po
```

NAME	READY	STATUS	RESTARTS	AGE
dist-mnist-for-e2e-test-ps-0	0/1	Pending	0	6s
dist-mnist-for-e2e-test-ps-1	0/1	Pending	0	6s
dist-mnist-for-e2e-test-worker-0	0/1	Pending	0	6s
dist-mnist-for-e2e-test-worker-1	0/1	Pending	0	6s
dist-mnist-for-e2e-test-worker-2	0/1	Pending	0	6s
dist-mnist-for-e2e-test-worker-3	0/1	Pending	0	5s
dist-mnist-for-e2e-test-worker-4	0/1	Pending	0	5s
dist-mnist-for-e2e-test-worker-5	0/1	Pending	0	5s
dist-mnist-for-e2e-test-worker-6	0/1	Pending	0	4s
dist-mnist-for-e2e-test-worker-7	0/1	Pending	0	4s



```
[root@iZwz91h5stiwetzn1plphZ dist-mnist]# kubectl get po
```

NAME	READY	STATUS	RESTARTS	AGE
busybox3-c757dc546-p6thm	1/1	Running	0	9d
dist-mnist-for-e2e-test-ps-0	1/1	Running	0	8m
dist-mnist-for-e2e-test-ps-1	0/1	Running	0	8m
dist-mnist-for-e2e-test-worker-0	0/1	Running	0	8m
dist-mnist-for-e2e-test-worker-1	0/1	Running	0	8m
dist-mnist-for-e2e-test-worker-2	0/1	Running	0	8m
dist-mnist-for-e2e-test-worker-3	1/1	Running	0	8m
dist-mnist-for-e2e-test-worker-4	1/1	Running	0	8m
dist-mnist-for-e2e-test-worker-5	1/1	Running	0	8m
dist-mnist-for-e2e-test-worker-6	1/1	Running	0	8m
dist-mnist-for-e2e-test-worker-7	1/1	Running	0	8m



Build Your AI Platform on ACK

阿里云容器服务提供一系列最佳实践



- [在阿里云Kubernetes容器服务上打造TensorFlow实验室](#)
- [阿里云Kubernetes 1.9上利用Helm玩转TensorFlow模型预测](#)
- [阿里云Kubernetes 1.9上利用Helm运行TensorFlow 分布式模型训练](#)
- **Kubeflow实战系列：**
 - [阿里云上使用JupyterHub](#)
 - [阿里云上小试TFJob](#)
 - [利用TFJob运行分布式TensorFlow](#)
 - [利用TFJob导出分布式TensorFlow模型](#)
 - [利用TensorFlow Serving进行模型预测](#)



谢 谢



主办方：



caicloud 才云



 TensorFlow

协办方：



网易云