



OSC 源创会第 90 期【在线直播】

# AI 专场

</>

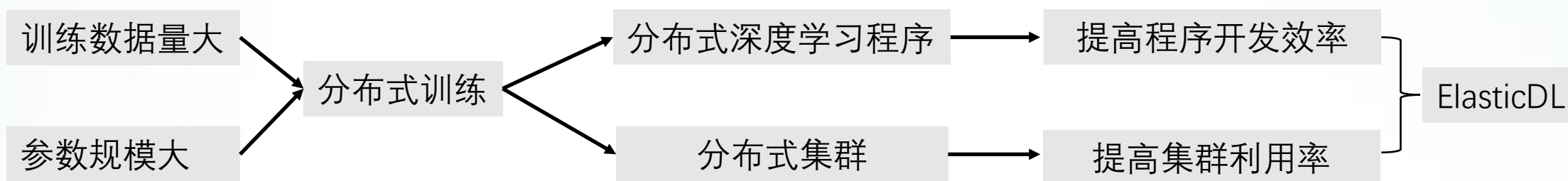
  直播时间：本周一至周四，每晚 20:00 - 21:00

# ElasticDL : Kubernetes-native 的弹性分布式深度学习框架

项目地址 <https://github.com/sql-machine-learning/elasticdl>

## 背景

在工业界生产场景中，深度学习应用具有训练数据量大和参数规模大的特点。



为了提供分布式程序开发效率和集群利用率，ElasticDL 提供了如下解决方案：

- 提供简单的分布式深度学习编程框架，让用户像写单机程序一样写分布式深度学习程序
- 弹性分布式深度学习系统，提高分布式作业的执行效率和集群资源利用率

ElasticDL 是基于 Kubernetes 和 TensorFlow 2.x 实现的分布式深度学习训练系统。

使用 TensorFlow/Keras API 开发深度学习程序



- TensorFlow 是业界应用最广泛的深度学习框架
- end-to-end 的深度学习平台
- eager execution 大幅提升了TensorFlow 的易用性

Kubernetes 集群上运行分布式程序



- 目前最先进的分布式操作系统
- 公有云和私有云的事实工业标准
- Docker 容器隔离

ElasticDL : 像写单机程序一样写分布式深度学习程序

ElasticDL : Kubernetes-native 弹性分布式训练系统

ElasticDL 在蚂蚁集团 CTR 预估场景的实践

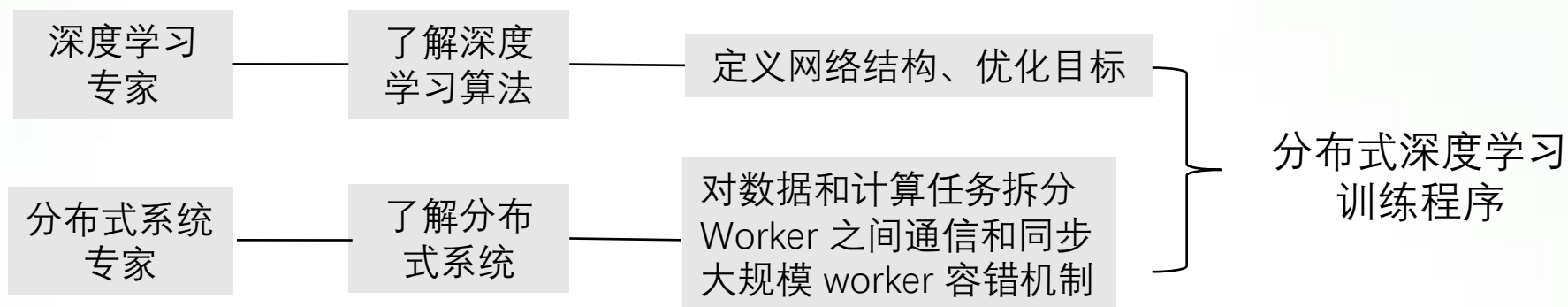
➤ ElasticDL : 像写单机程序一样写分布式深度学习程序

ElasticDL : Kubernetes-native 弹性分布式训练系统

ElasticDL 在蚂蚁集团 CTR 预估场景的实践

## ElasticDL：像写单机程序一样写分布式深度学习程序

分布式深度学习程序难写的原因：



Kubernetes 上运行 TensorFlow 分布式训练程序的一些开源解决方案：

	模型定义	Kubernetes 调度工具	缺点
方案一	TensorFlow Estimator API	Kubeflow TF-operator	仅支持 graph execution, 不支持 eager execution
方案二	TensorFlow Keras API	Kubeflow TF-operator	不支持 Parameter Server 分布式
方案三	Horovod with TensorFlow	Kubeflow MPI-operator	了解 TensorFlow 和 Horovod API

使用 Kubeflow 提供的 Kubernetes Operator 在集群上运行分布式训练作业，还需要用户对 Kubernetes 有所了解。

## ElasticDL：像写单机程序一样写分布式深度学习程序

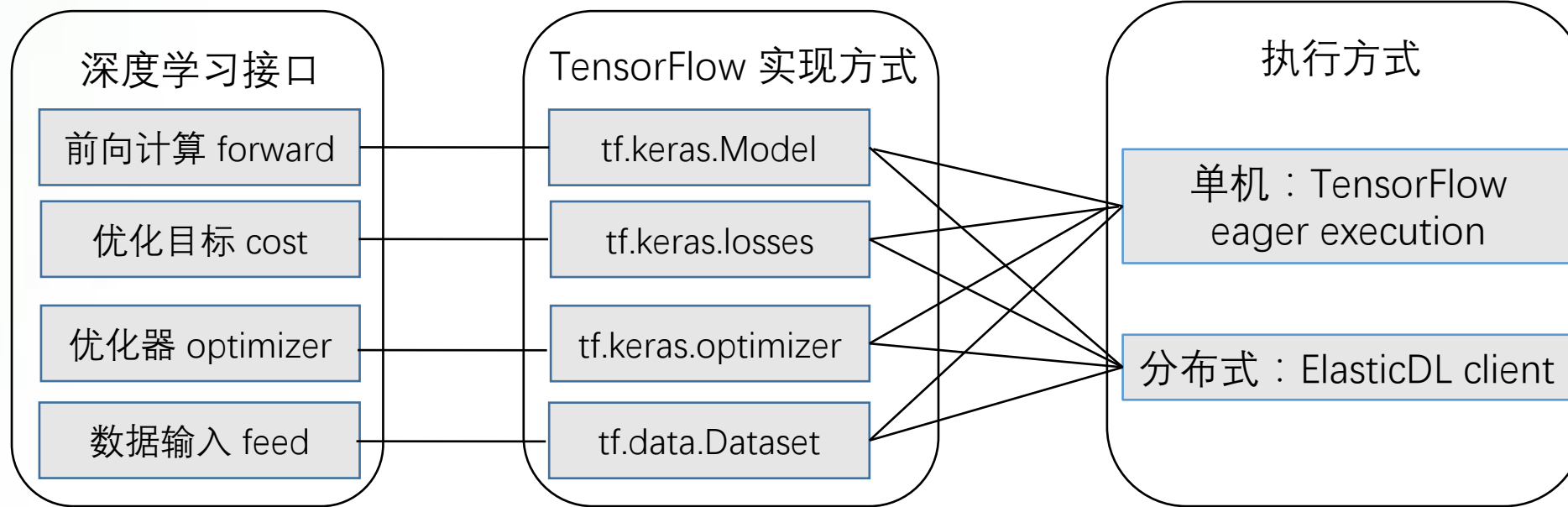
- ElasticDL 使用 TensorFlow eager execution 开发了分布式的 training loop，用户无需关心分布式系统。

```
w = tf.Variable([[1.0]])
with tf.GradientTape() as tape:
    loss = w * w

grad = tape.gradient(loss, w)
print(grad) # => tf.Tensor([[ 2.]], shape=(1, 1), dtype=float32)

tf.Tensor([[2.]], shape=(1, 1), dtype=float32)
```

- 针对深度学习抽象了4种接口，用户只需使用 TensorFlow/Keras API 定义这些接口即可。





# ElasticDL：像写单机程序一样写分布式深度学习程序

## mnist 数据集手写数字识别模型

```
def forward():
    inputs = tf.keras.Input(shape=(28, 28), name="image")
    x = tf.keras.layers.Reshape((28, 28, 1))(inputs)
    x = tf.keras.layers.Conv2D(32, kernel_size=(3, 3), activation="relu")(x)
    x = tf.keras.layers.Conv2D(64, kernel_size=(3, 3), activation="relu")(x)
    x = tf.keras.layers.BatchNormalization()(x)
    x = tf.keras.layers.MaxPooling2D(pool_size=(2, 2))(x)
    x = tf.keras.layers.Dropout(0.25)(x)
    x = tf.keras.layers.Flatten()(x)
    outputs = tf.keras.layers.Dense(10)(x)
    return tf.keras.Model(inputs=inputs, outputs=outputs, name="mnist_model")

def loss(labels, predictions):
    labels = tf.reshape(labels, [-1])
    return tf.reduce_mean(
        input_tensor=tf.nn.sparse_softmax_cross_entropy_with_logits(
            logits=predictions, labels=labels
        )
    )

def optimizer(lr=0.1):
    return tf.optimizers.SGD(lr)

def dataset_fn(dataset, mode, _):
    dataset = dataset.map(_parse_data)

    if mode == Mode.TRAINING:
        dataset = dataset.shuffle(buffer_size=1024)
    return dataset
```

## 生成模型镜像

```
FROM tensorflow
RUN pip install elasticdl
COPY model_zoo /model_zoo
```

## ElasticDL client 提交分布式训练

```
elasticdl train \
  --image_name=elasticdl:ci \
  --model_zoo=model_zoo \
  --model_def=mnist_functional_api.mnist_functional_api.custom_model \
  --training_data=/data/mnist/train \
  --num_epochs=1 \
  --master_resource_request="cpu=1,memory=4096Mi,ephemeral-storage=1024Mi" \
  --worker_resource_request="cpu=1,memory=4096Mi,ephemeral-storage=1024Mi" \
  --ps_resource_request="cpu=1,memory=4096Mi,ephemeral-storage=1024Mi" \
  --minibatch_size=64 \
  --num_ps_pods=1 \
  --num_workers=2 \
  --job_name=test-train \
  --distribution_strategy=ParameterServerStrategy \
  --output=model_output
```

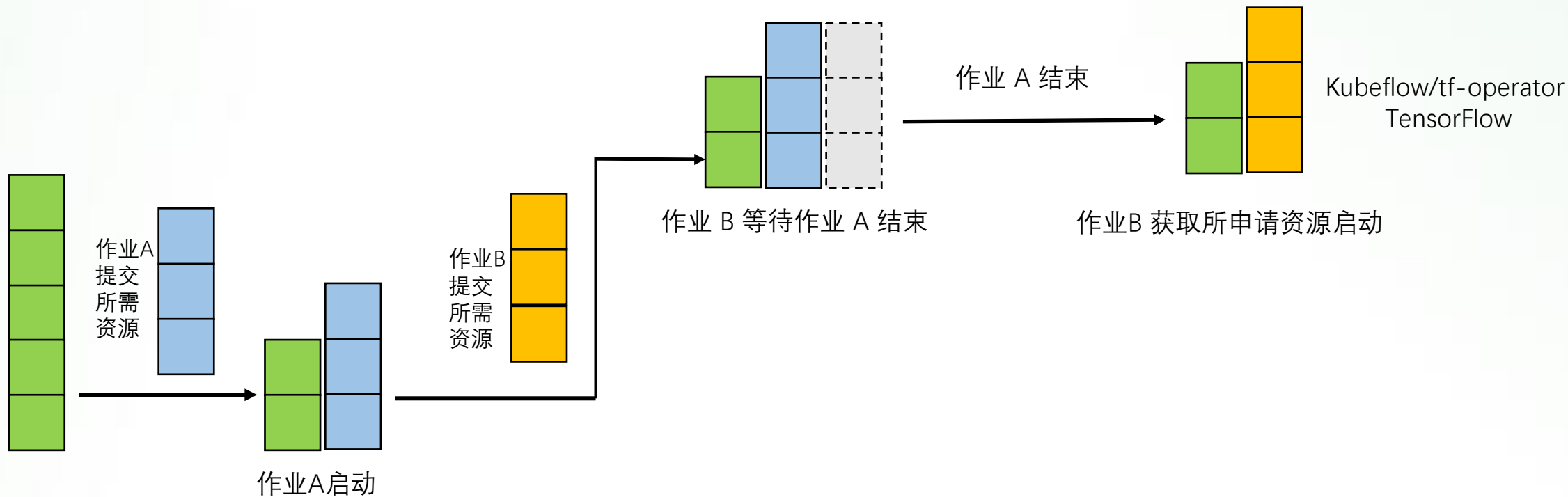
ElasticDL : 像写单机程序一样写分布式深度学习程序

➤ ElasticDL : Kubernetes-native 弹性分布式训练系统

ElasticDL 在蚂蚁金服 CTR 预估场景的实践

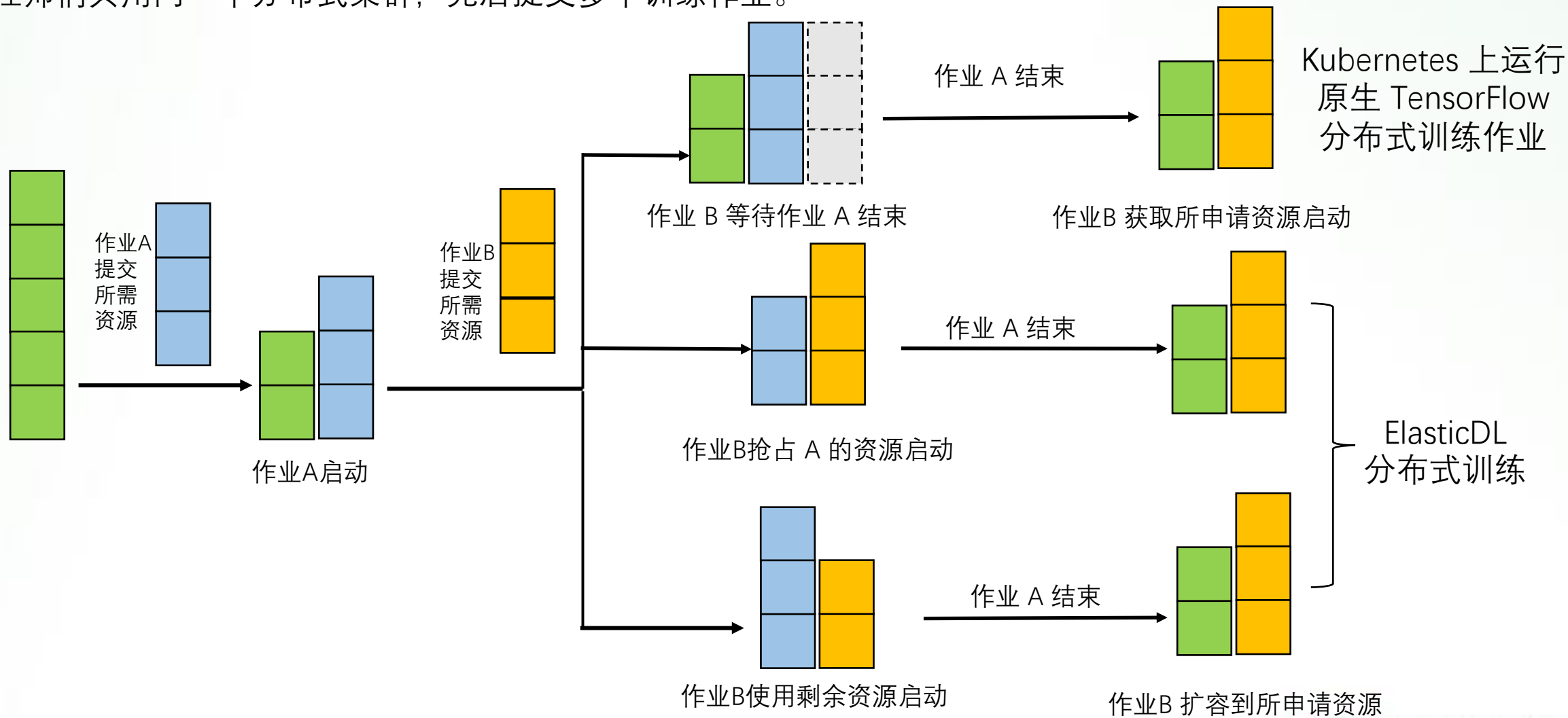
## 共用集群资源，资源等待时间长，资源利用率低

工程师们共用同一个分布式集群，先后提交多个训练作业。



## 共用集群资源，资源等待时间长，资源利用率低

工程师们共用同一个分布式集群，先后提交多个训练作业。



- 调用 TensorFlow eager execution 构建分布式训练系统
- 调用 Kubernetes API 来调整训练过程中的进程数量
- 数据分片的动态分发机制

## ElasticDL: 调用 TensorFlow eager execution 构建分布式训练系统

基于 TensorFlow 的分布式训练系统大致分为以下四类，ElasticDL 位于田字格的右下角是为了容错和弹性。

	TensorFlow 1.x graph mode	TensorFlow 2.x eager execution
in TensorFlow runtime	TensorFlow's parameter server	TensorFlow distributed strategy
above TensorFlow API	Horovod	ElasticDL, Horovod

TensorFlow runtime: 与平台无关，不会调用集群管理系统的来调整训练资源，无法实现主动的弹性调度。

TensorFlow API：ElasticDL 和 Horovod 都是基于 TensorFlow API 来实现分布式训练。

相同点：

- 每个进程有完成的计算逻辑
- 通过 TensorFlow API 获取梯度

不同点：

- Horovod 与平台无关，可以运行在多种分布式集群上，但是不能利用集群管理系统来调整进程数量进行弹性调度
- ElasticDL 是 Kubernetes-native，只能运行在 Kubernetes 集群上，能利用 Kubernetes API 来调整作业进程数量

## ElasticDL：调用 Kubernetes API 来调整训练过程中的进程数量

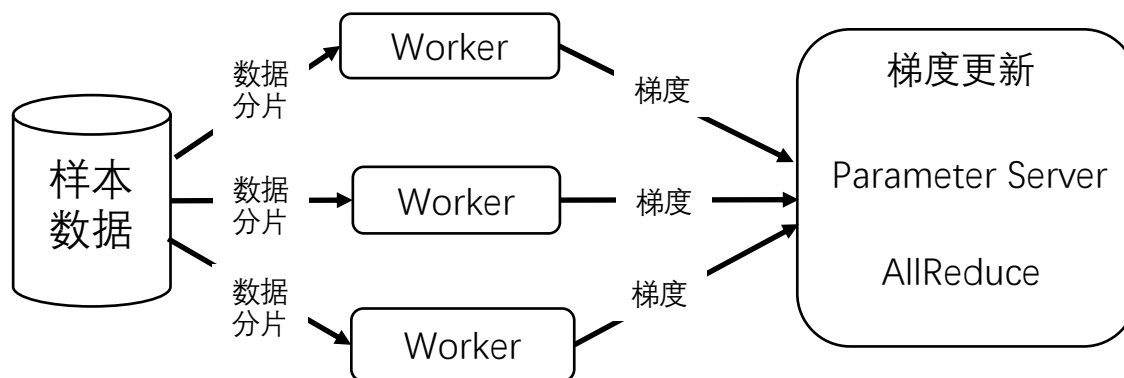
ElasticDL 通过在 Kubernetes 上创建 master 进程来协调训练数据分配、通信、梯度同步和容错，从而支持弹性调度。



- ElasticDL Client 通过 Kubernetes API 启动 Master 进程
- Master 通过 Kubernetes API 启动 PS 和 Worker 进程
- Master 给 worker 分发数据分片
- Master 通过 Kubernetes API 来监听 Worker 状态
- Master 通过 Kubernetes API 重启被抢占的 Worker

## ElasticDL：数据分片的动态分发机制

分布式训练作业启动多个 worker 进程后，需要给 worker 进程分配数据分片。



为了提高分布式训练的效率，数据分片要尽可能均匀，防止 worker 间的负载不均衡。为了达到弹性调度，数据分片还需满足如下条件：

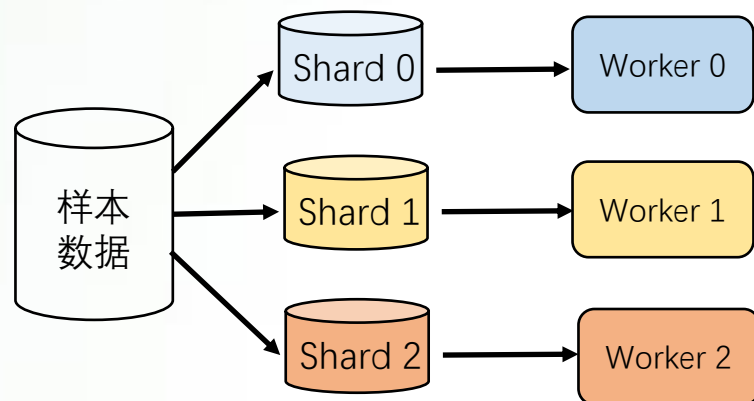
- Worker 失败后，其计算的数据分片能分配给其他 Worker 重新计算
- 有新的 Worker 加入后，能获取为计算的数据分片进行计算



## ElasticDL：数据分片的动态分发机制

### 静态数据分发

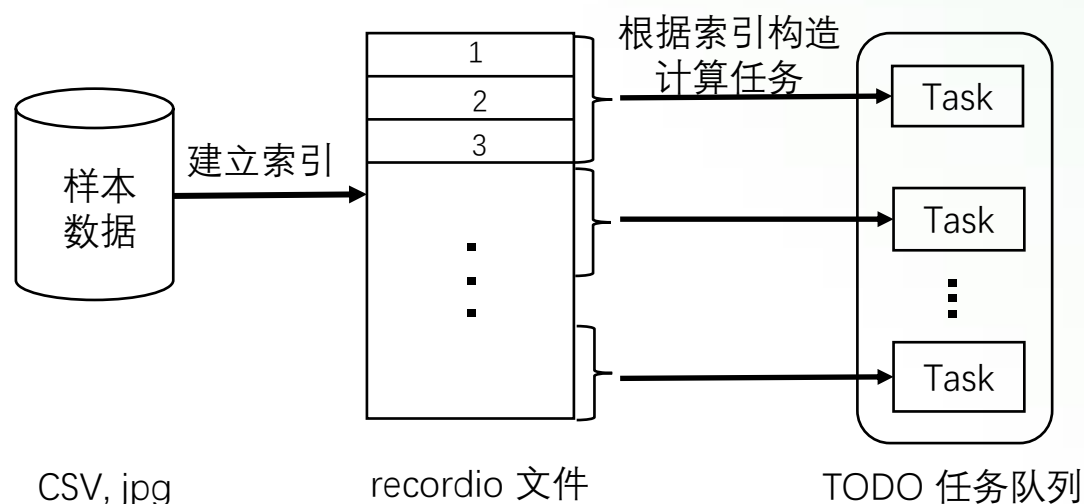
在作业开始前，将数据进行分片，然后指定给某个 Worker



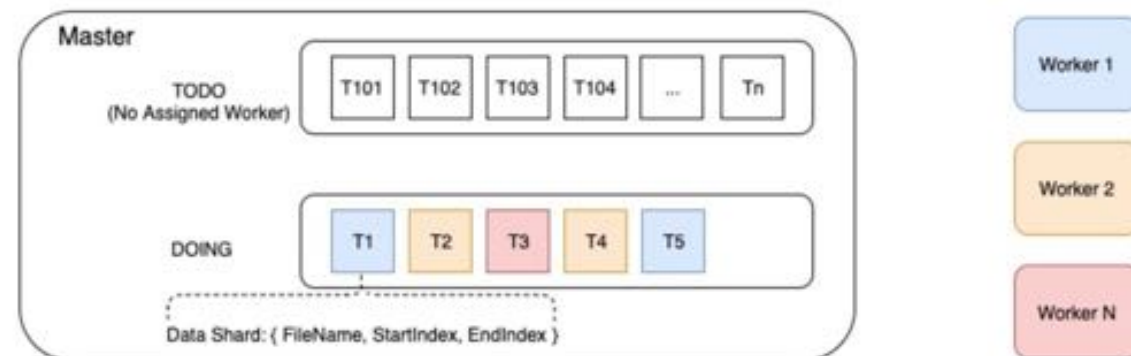
缺点：

- Worker 失败后，其分配的数据没法给其他 Worker 训练
- 慢 Worker 会拖慢整个训练作业
- 数据分片遍历次数不均衡，影响收敛性

### ElasticDL 动态数据分发



Dispatch the task(a shard of data) to each worker dynamically at runtime



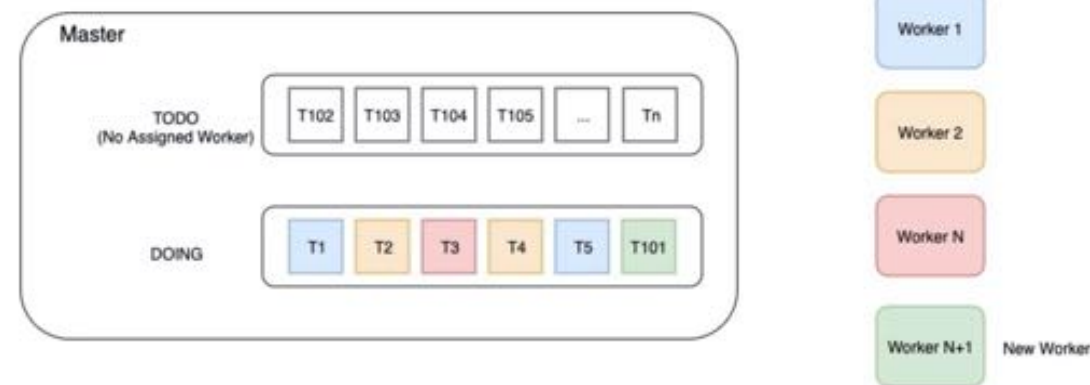
## ElasticDL：数据分片的动态分发机制

### ElasticDL：调整进程数量时的数据分发

新的 Worker 加入作业（扩容）

Master 将新的 Worker 加入作业时，会从 TODO 队列里给新 Worker 分配 Task，Worker 根据 Task 拿到数据分片进行计算

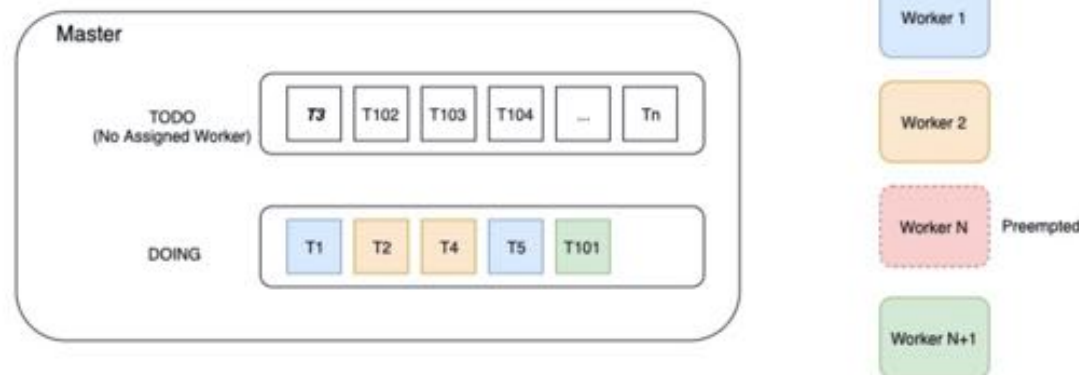
While a new worker joins, dispatch a todo task to this worker



Worker 被其他作业抢占（缩容）

- Master 回收被抢占 Worker 的 Task 到 TODO 队列
- Master 将 Task 分配给其他正常 Worker 计算

While a worker is preempted, recover the task of this worker to the todo list



为了验证 ElasticDL 弹性调度的能够提高深度学习作业的研发效率和集群资源利用率，分别做了三组实验：

- 多个深度学习训练作业同时在集群上启动，验证弹性调度能缩短作业等待时间
- 深度学习作业与高优先级的在线服务混布，验证弹性调度可以提升集群资源利用率
- 训练时调整 worker 数量观察模型收敛性，验证 ElasticDL 弹性调度不影响模型收敛

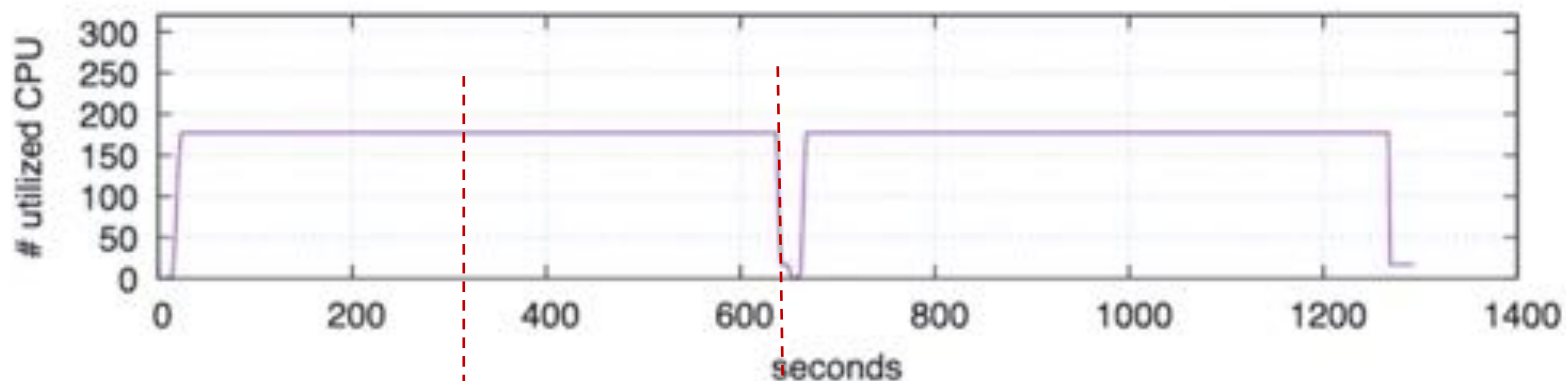
### ElasticDL benchmark 实验——多个深度学习训练作业同时在集群上启动

两个深度学习训练作业需要的资源总和略超过集群的情况

实验设置：

- 集群总 CPU 数 320 个
- 先后提交两个训练作业
- 每个作业需要 175 个CPU

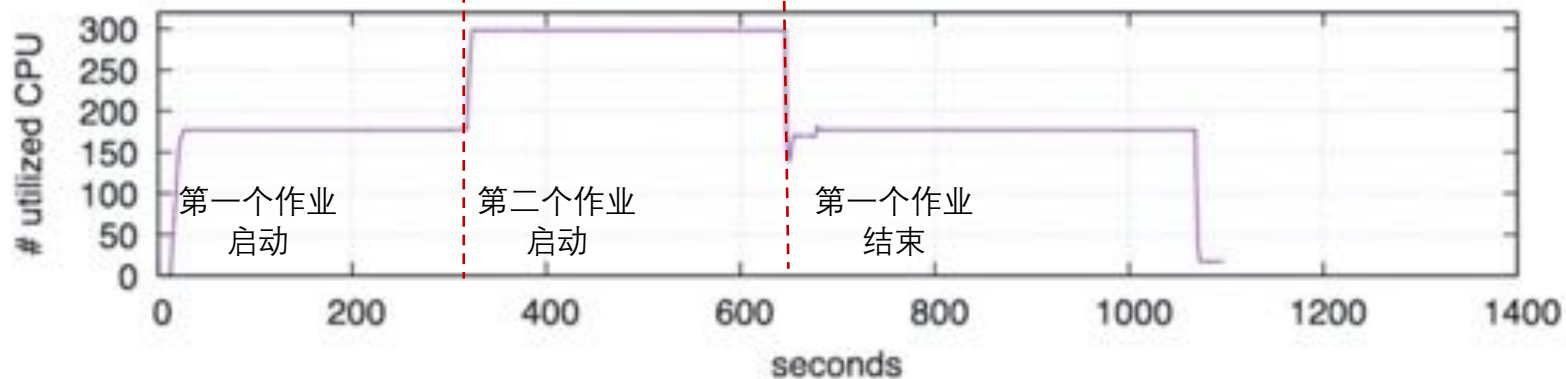
Kubeflow/tf-operator先后运行两个作业



实验结论：

- 用户作业等待时间几乎为 0
- 集群利用率高
- 作业完成时间更快

ElasticDL 先后运行两个作业



### ElasticDL benchmark 实验——深度学习作业与高优先级在线服务混布

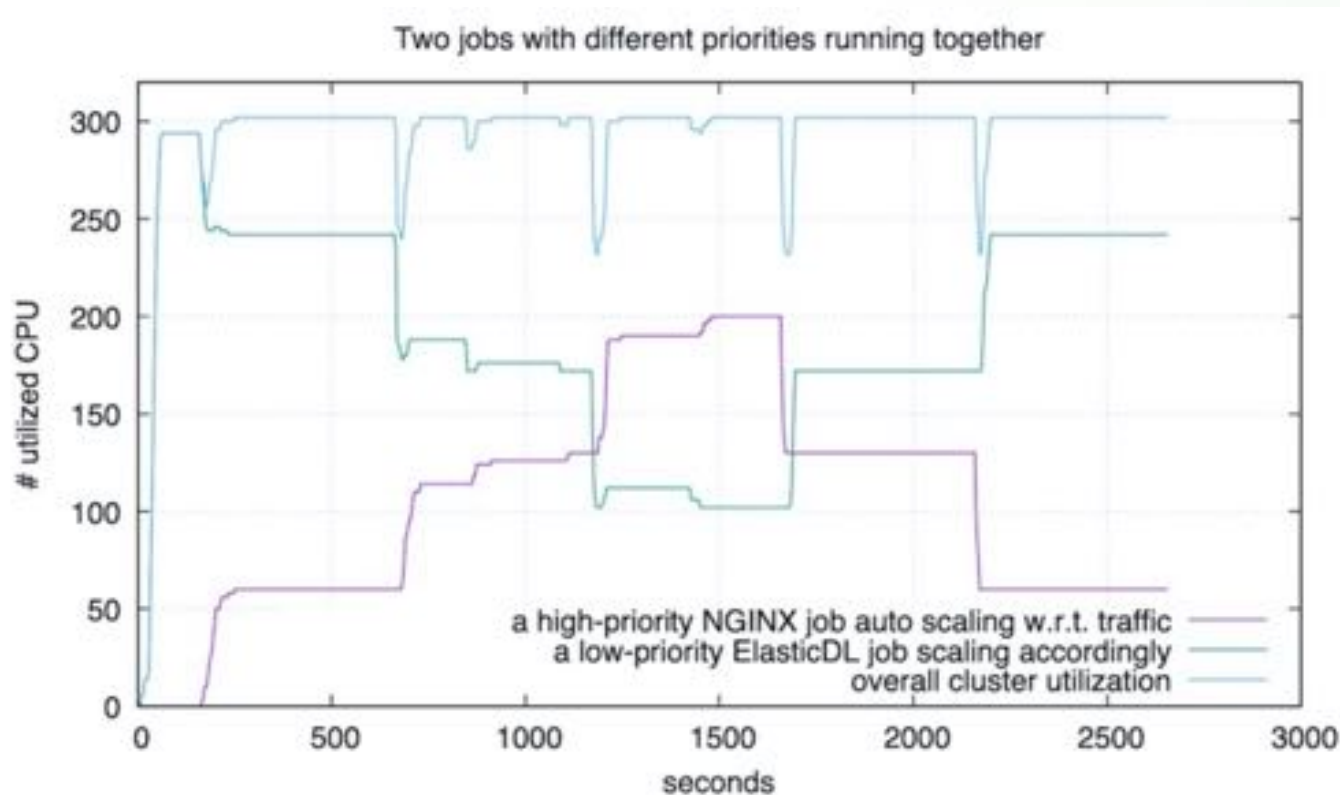
ElasticDL 作业与在线 nginx 服务运行在同一个 Kubernetes 集群上，nginx 作业的优先级高于ElasticDL作业。

实验设置：

- 集群总 CPU 数 320 个
- 训练任务 DeepFM 二分类模型
- 训练时长约 40 min

实验结论：

- 当用户请求增加时，Kubernetes 自动扩容 nginx 服务，ElasticDL 作业释放部分资源
- 流量高峰过去后，nginx 服务释放资源，ElasticDL 扩充资源



## ElasticDL: 弹性调度性能验证

### ElasticDL benchmark 实验——训练时调整 worker 数量不影响模型收敛性

训练过程中，ElasticDL 调整的 worker 数量，不会超过用户配置的最大 worker 数。防止 worker 数量过多，梯度更新的参数滞后严重，影响模型收敛性。

实验数据集：

Kaggle Display Advertising Challenge

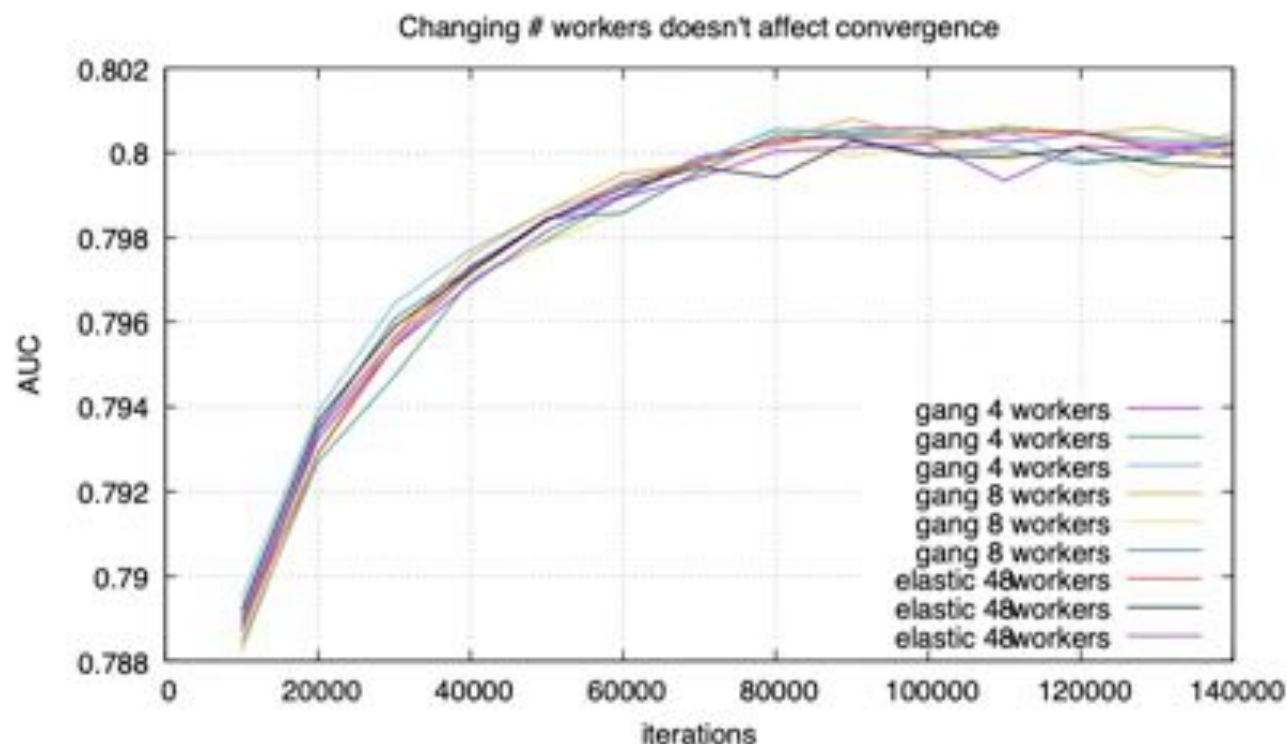
数据列	描述
Label 列	1 或 0, 1表示点击, 0 表示未点击
I1-I13	整数型特征
C1-C26	字符串型类别特征

样本数量：约4千万

模型：Wide & Deep Learning

参数更新策略：Parameter Server 下的异步 SGD

实验结论：弹性调度的模型 AUC 和kubeflow/tf-operator 提交的原生 TensorFlow 分布式训练作业持平。



ElasticDL : 像写单机程序一样写分布式深度学习程序

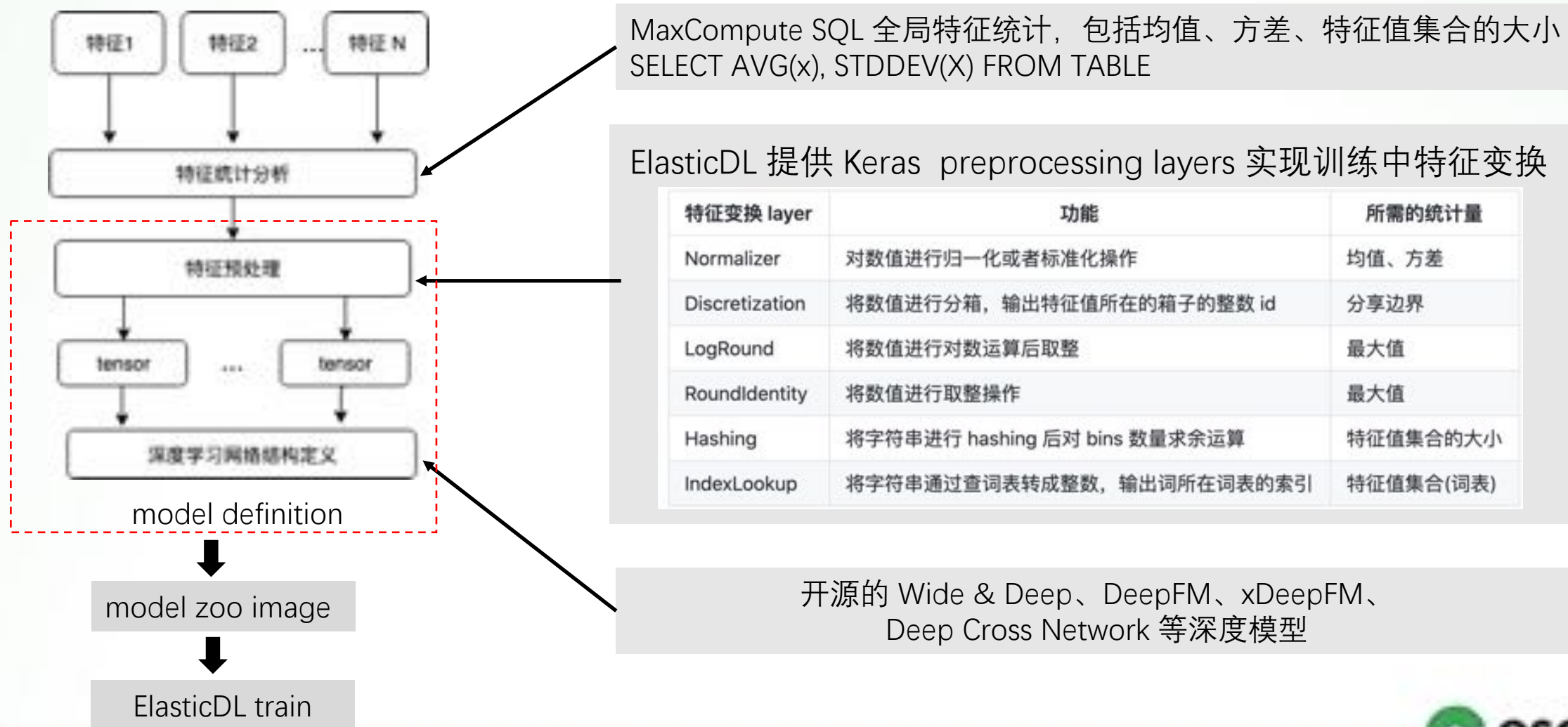
ElasticDL : Kubernetes-native 弹性分布式训练系统

➤ ElasticDL 在蚂蚁集团 CTR 预估场景的实践



## ElasticDL 在蚂蚁集团 CTR 预估场景的实践

蚂蚁集团的搜索推荐广告场景应用到了大量的 CTR 预估模型。其数据以结构化的形式存储在阿里云的 MaxCompute 上。CTR 预估模型建模流程如下：

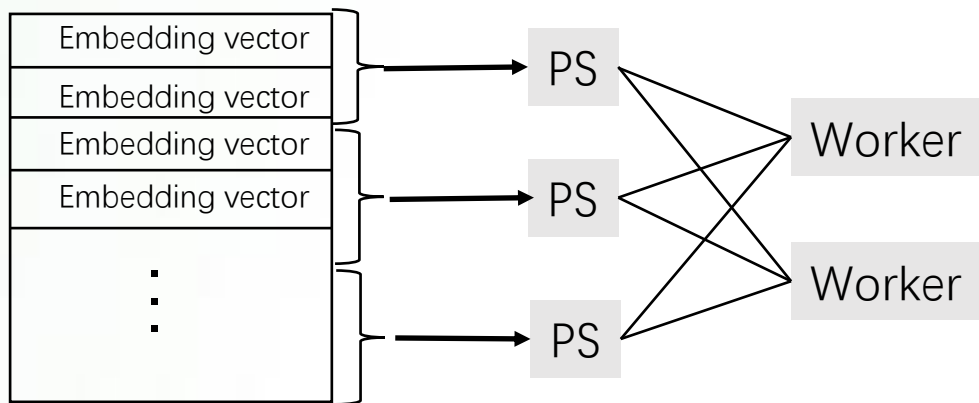




## ElasticDL 在蚂蚁集团 CTR 预估场景的实践

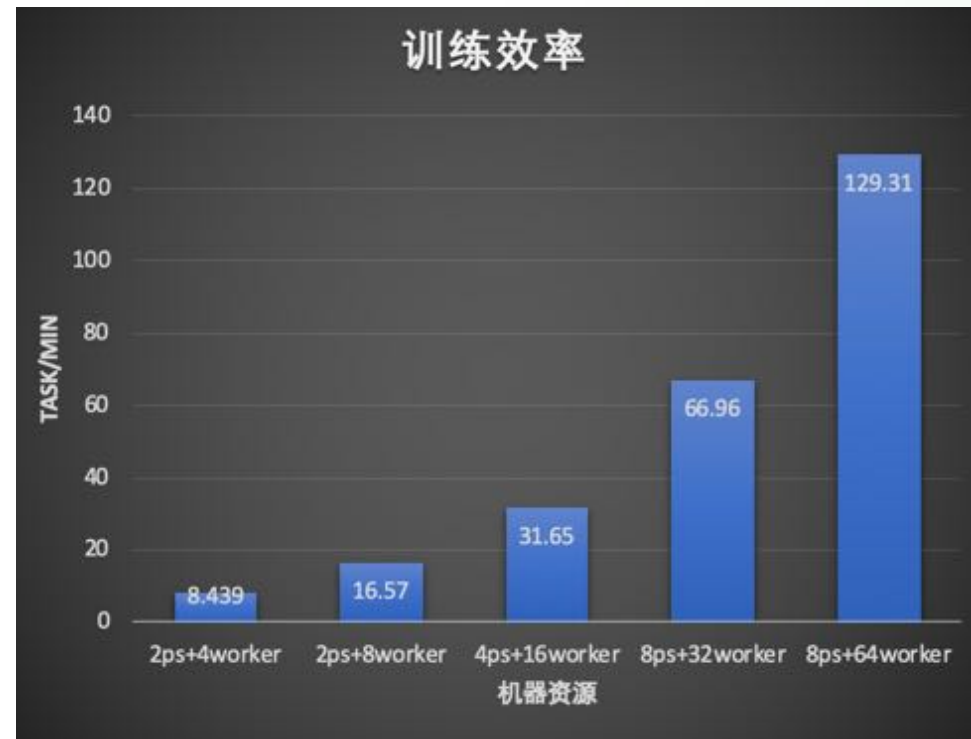
CTR 预估模型的输入一般含有用户和商品属性的高维稀疏特征，需要使用大规模的 embedding，通常采用 Parameter Server (PS) 来训练这种模型。ElasticDL 针对 PS 的优化：

### ElasticDL Embedding 的优化



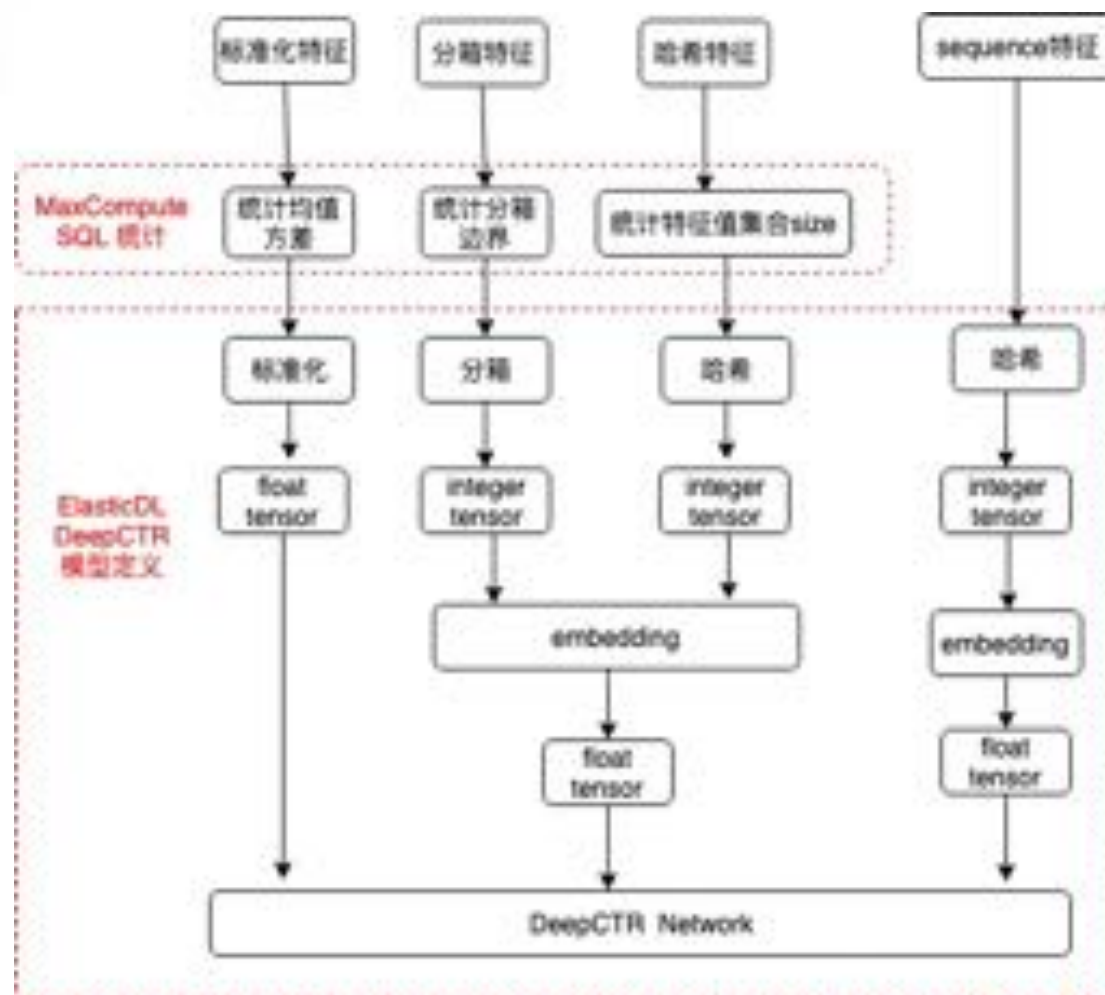
- Embedding vector 惰性初始化，无需提前指定 embedding 的大小；
- Embedding 表分拆到多个 PS 上存储和更新，均衡存储和通信负载；
- Worker 从 PS 请求参数和发送梯度时，针对重复 ID 做合并，减少通信量。

PS 并发度优化：训练速率与worker数量线性相关



## ElasticDL 在蚂蚁集团 CTR 预估场景的实践

为了进一步提升易用性，ElasticDL 结合 DeepCTR 预估模型，开发了自动特征预处理的建模方案，用户只需配置特征处理方式即可完成 CTR 模型的分布式训练：



# ElasticDL 在蚂蚁集团 CTR 预估场景的实践

为了验证方案的性能，选用 Kaggle Display Advertising Challenge 数据集进行测试。

样本数量：训练集约 4000 万条，测试集约 600 万条



## 模型性能

CTR 算法	测试集 logloss
xDeepFM	0.45634
Wide & Deep	0.45998
Deep Cross Network	0.45988
Kaggle Best	0.44463

模型代码地址：[https://github.com/sgl-machine-learning/elasticdl/tree/develop/model\\_zoo/dac\\_ctr](https://github.com/sgl-machine-learning/elasticdl/tree/develop/model_zoo/dac_ctr)

## 总结：

---

- ElasticDL 基于 TensorFlow API 提供了分布式深度学习编程框架，降低分布式程序编程难度。
- ElasticDL 基于 Kubernetes 和 TensorFlow 实现了弹性分布式训练，来提高作业研发效率和集群资源利用率。
- ElasticDL 针对蚂蚁集团 CTR 预估场景，提供了包括特征处理和模型训练的开箱即用方案，已在蚂蚁集团多个搜索推荐场景落地。

ElasticDL 项目 github 地址

<https://github.com/sql-machine-learning/elasticdl>



# Q & A