

TensorFlow 遇上 Kubernetes

中兴通讯人工智能计算平台的技术实践

韩炳涛¹ 刘光聪²

¹han.bingtao@zte.com.cn

²liu.guangcong@zte.com.cn

2017-08-05

内容

1 架构概述

2 运行模型

3 训练模型

4 中兴 AI 平台

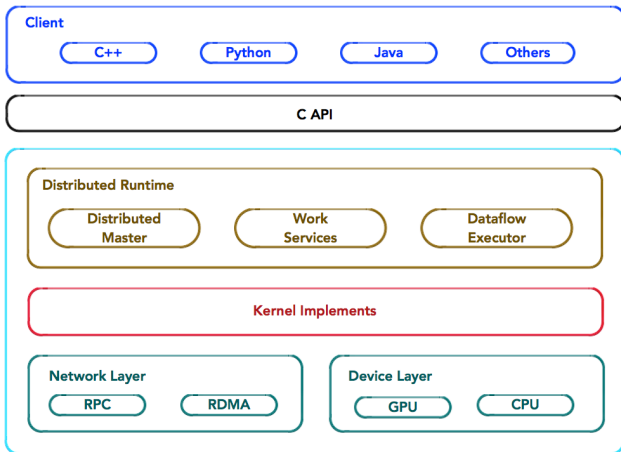
5 参考文献

架构概述

❶ 系统架构

❷ 图控制

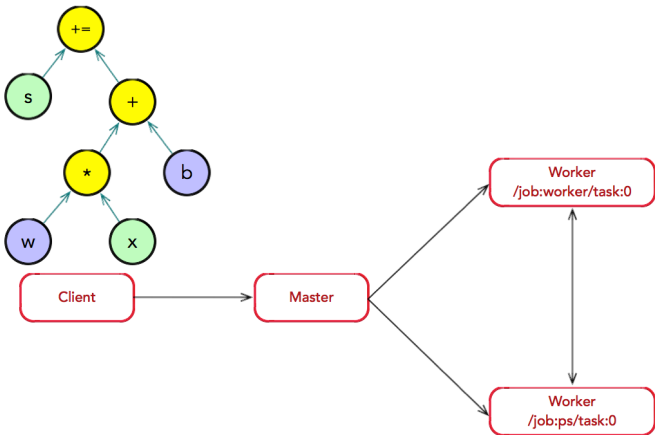
系统架构



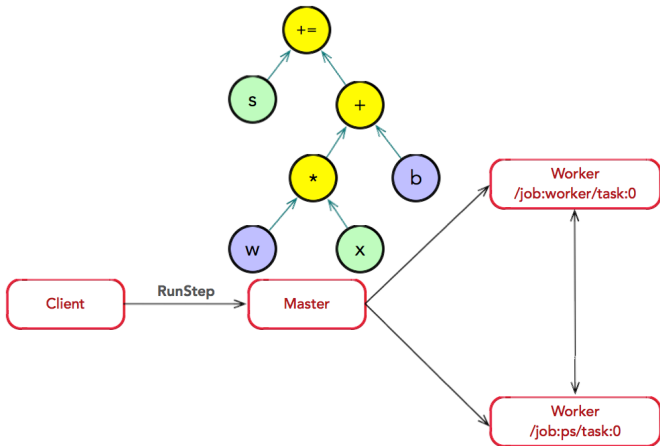
设计原则

- **延迟计算**：图的构造与执行分离，并推迟计算图的执行过程
- **原子 OP**：OP 是最小的抽象计算单元，支持构造复杂的网络模型
- **抽象设备**：支持 CPU, GPU, ASIC 多种异构设备类型
- **抽象任务**：基于 Task 的 PS，支持优化算法和网络模型的扩展

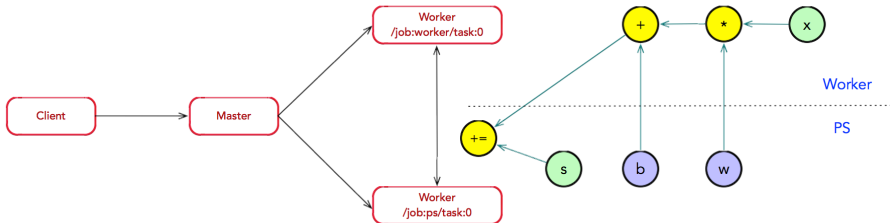
构造计算图



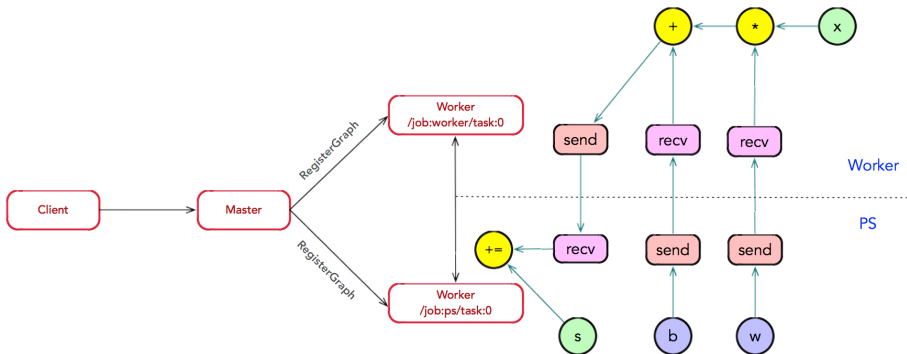
执行计算图



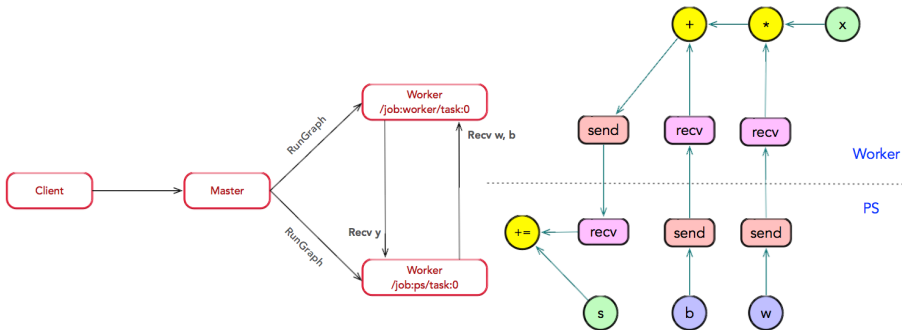
图分解：按 Task 分解



注册图



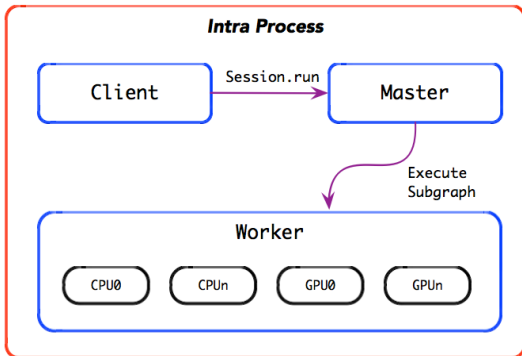
执行图



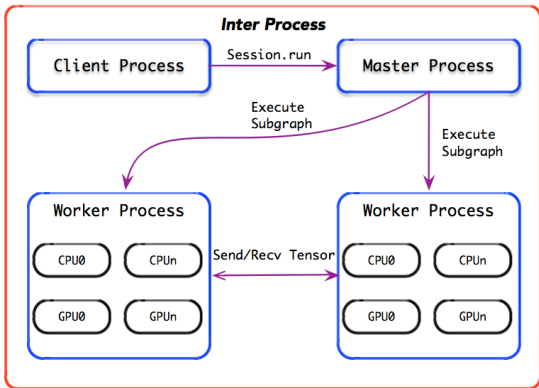
运行模型

- 1 运行模式
- 2 创建会话
- 3 RunStep 过程

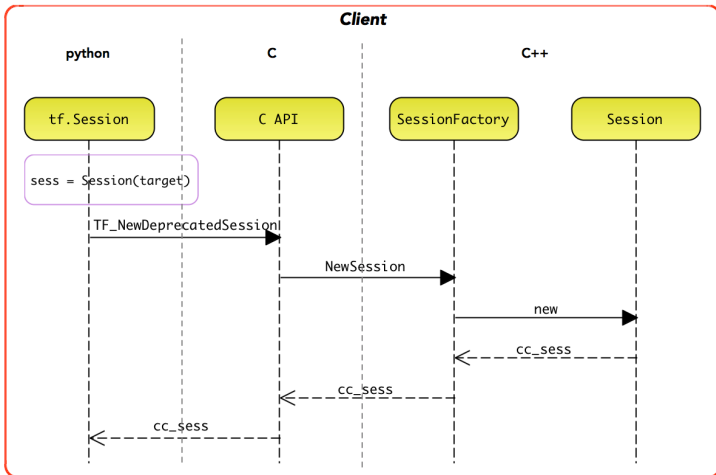
本地模式



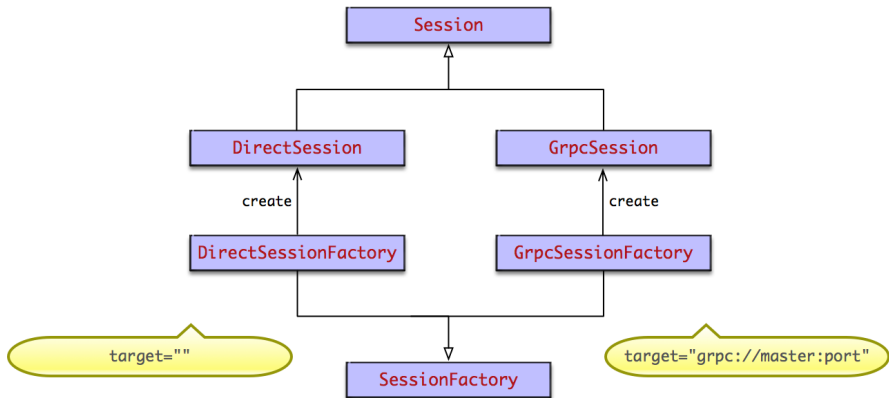
分布式模式



创建 ClientSession

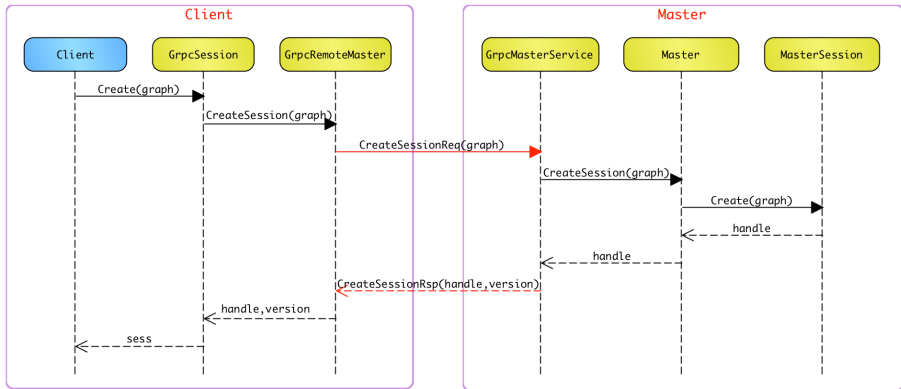


多态创建

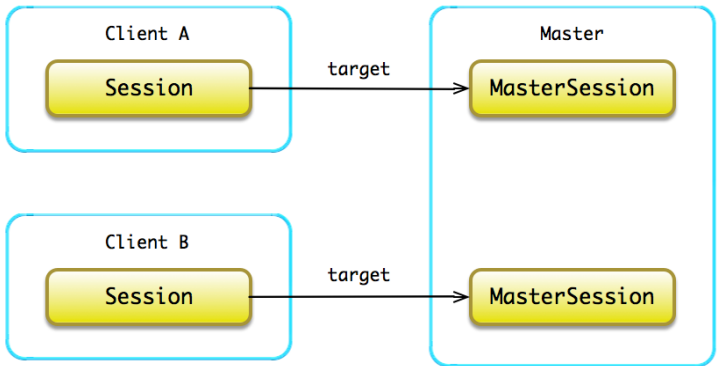


创建会话

创建 MasterSession

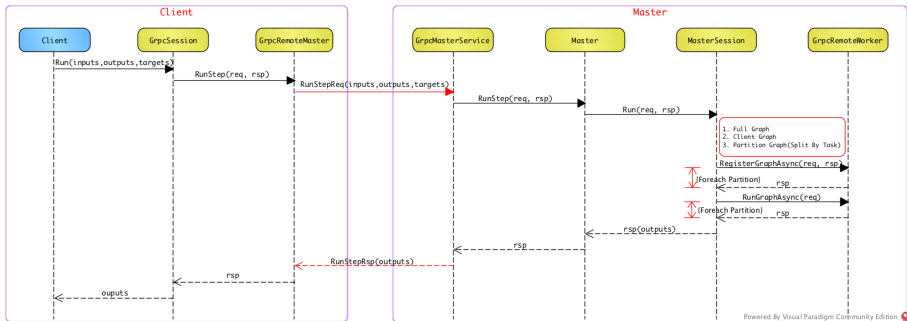


MasterSession 模型



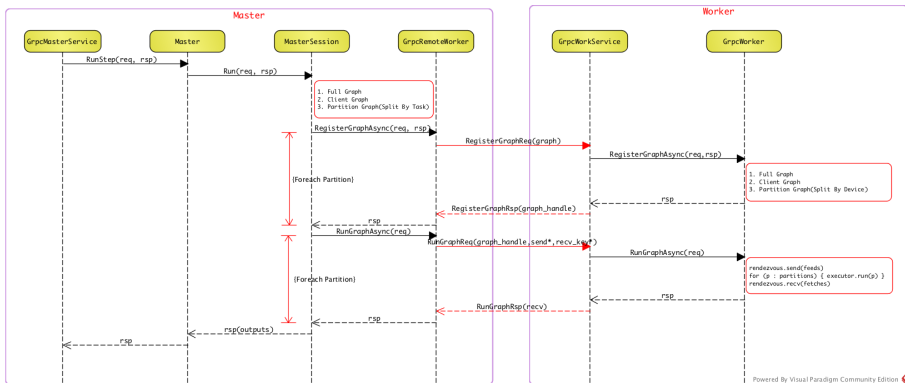
RunStep 过程

一级图分裂



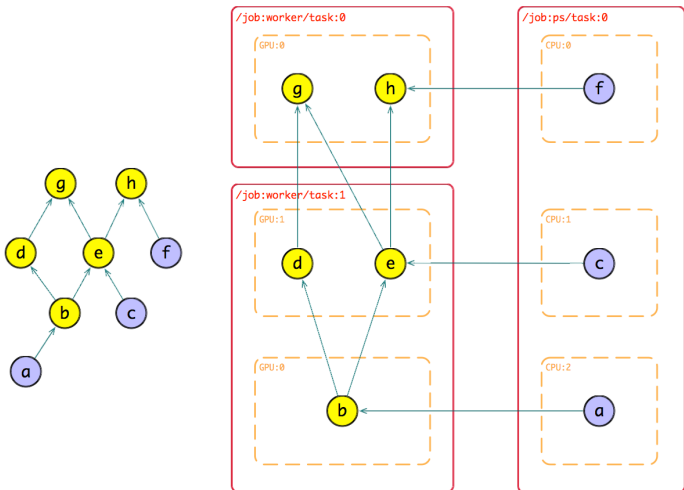
RunStep 过程

二级图分裂



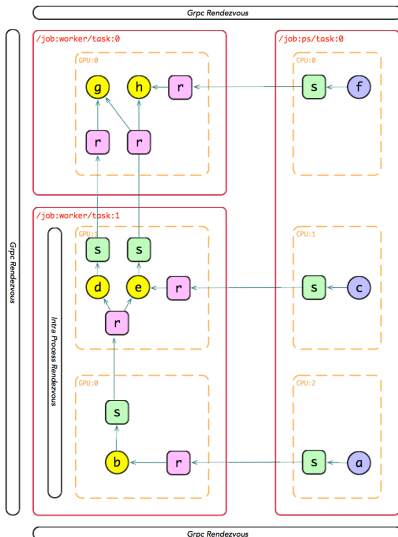
RunStep 过程

实例：图分裂



RunStep 过程

实例：交换数据



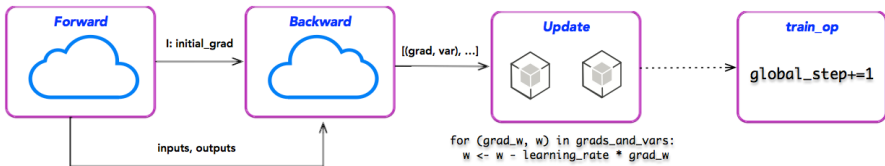
训练模型

- 1 优化算法
- 2 工作流

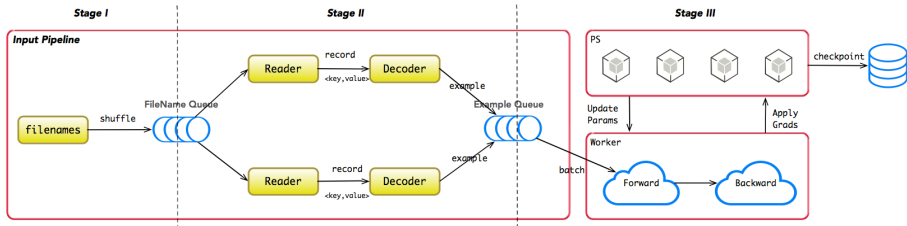
优化器

```
class Optimizer(object):  
    def minimize(self, loss, var_list=None, global_step=None):  
        grads_and_vars = self.compute_gradients(  
            loss, var_list=var_list)  
        return self.apply_gradients(  
            grads_and_vars,  
            global_step=global_step)
```

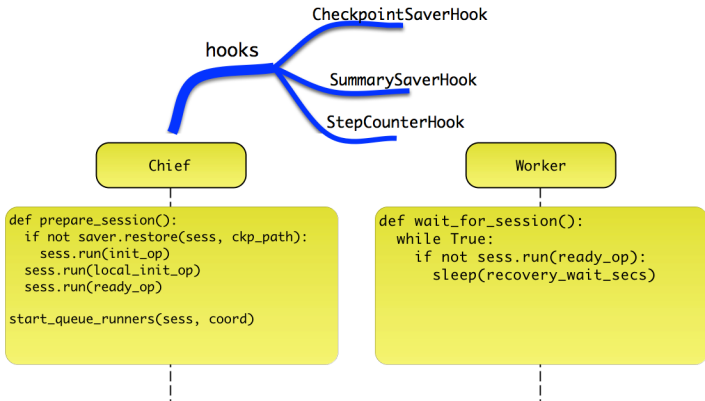
RunStep 过程



工作流



协调协议



中兴 AI 平台

❶ 平台架构

❷ 关键技术

参考文献

文献

- TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, Google Inc.
- TensorFlow: A System for Large-Scale Machine Learning, Google Inc.

致谢

Q&A



致谢

Thanks