
KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2019

基于Kubernetes进行深度学习训练推理的成本优化实践

演讲人：韩沛、王磊

AI on Kubernetes

现状

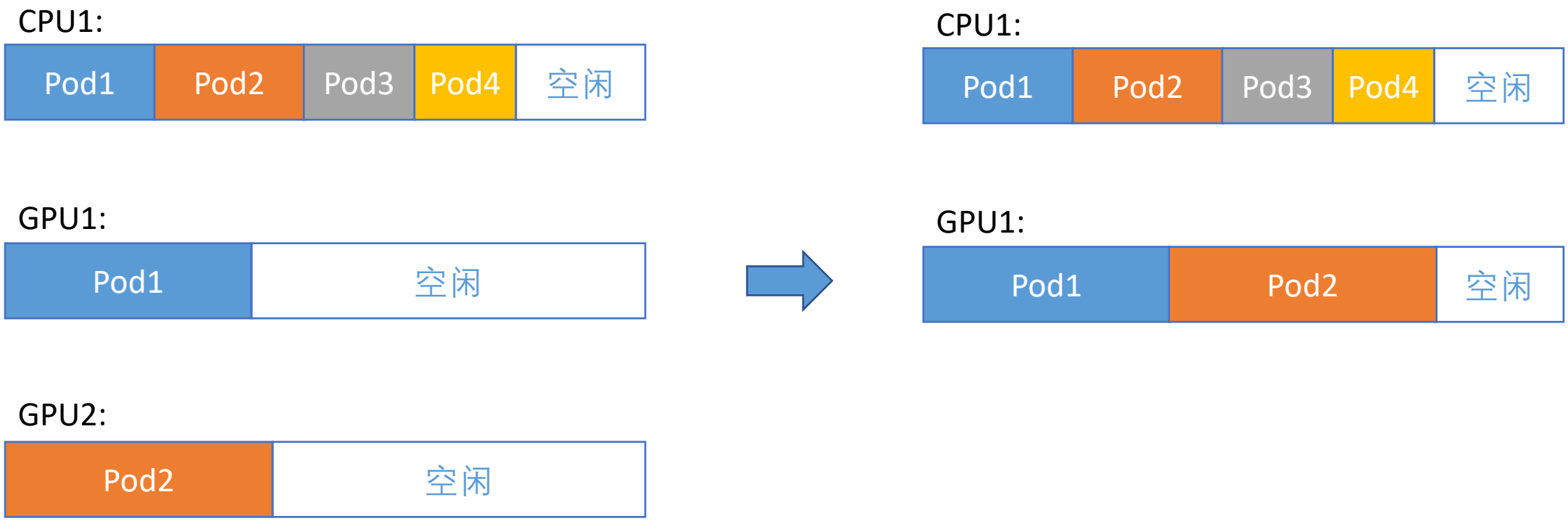
- Kubernetes 在国内已经获得认可
- TensorFlow、PyTorch、Caffe等 DL framework 均已提供 Kubernetes Operator
- 针对AI场景的 PAAS 产品在云上日渐丰富
- DL framework + Kubernetes 已经成为构建这些平台型产品的最佳选项之一

趋势预测

- DL framework 在 Kubernetes 上的配套功能会越来越强大和丰富
- AI PAAS 平台会逐渐转向或兼容 Kubernetes 生态
- 会有更成熟的针对多机多卡场景的容器网络方案
- GPU虚拟化技术会带动整个行业进行产品升级
- Serverless 形态的产品由于运维成本和自动伸缩优势会展露头脚

GPU 虚拟化的目标

像 CPU 一样规划 GPU 资源分配，提高昂贵的 GPU 资源利用率



GPU 虚拟化的标准

性能

执行操作的速度

资源复用

能够使多容器共享同一个物理GPU的能力以及隔离性

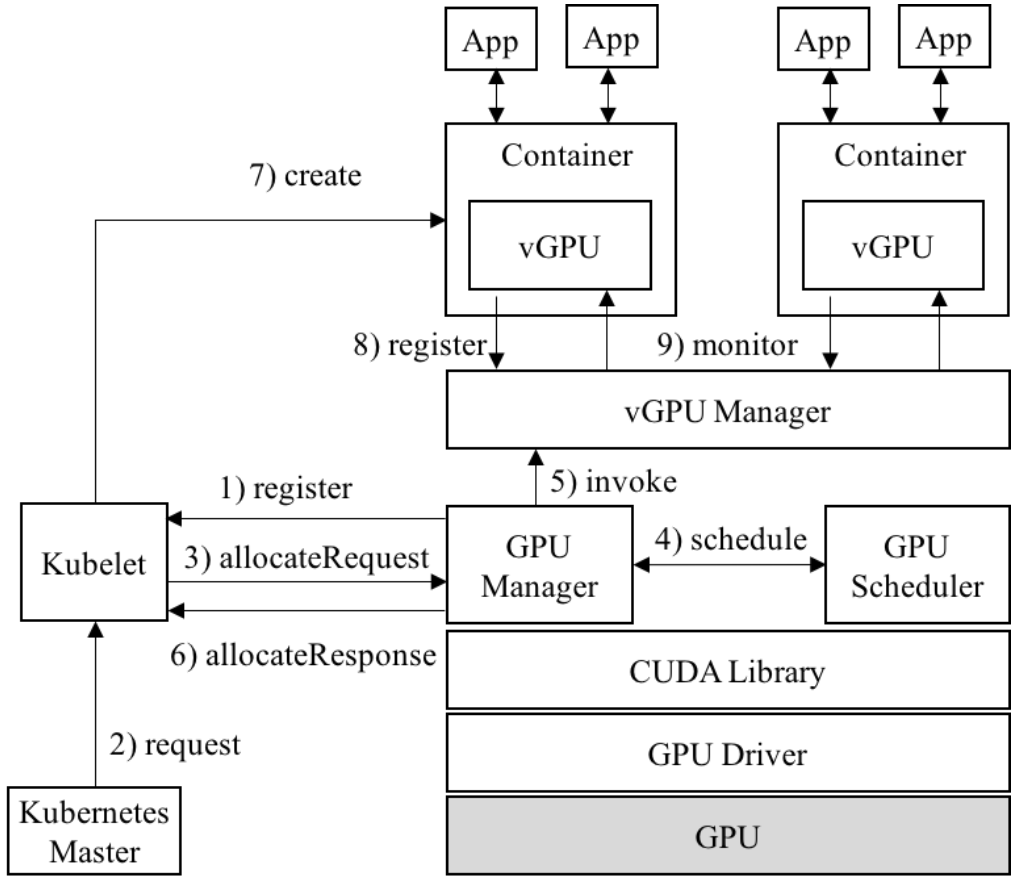
保真度

支持多少GPU提供的特性，以及对这些特性的支持的质量

是否支持虚拟化的特性

是否能支持虚拟化技术提供的容器与物理机之间的中介过程

一种适配K8S生态的GPU虚拟化技术



主要考虑：

性能 —— 保证vGPU的性能与原生GPU性能相近。

资源复用 —— 可以有效的分配和回收每个容器使用的GPU资源并实现不同容器间的资源隔离。

保真度 —— 支持物理GPU的大部分主要功能。

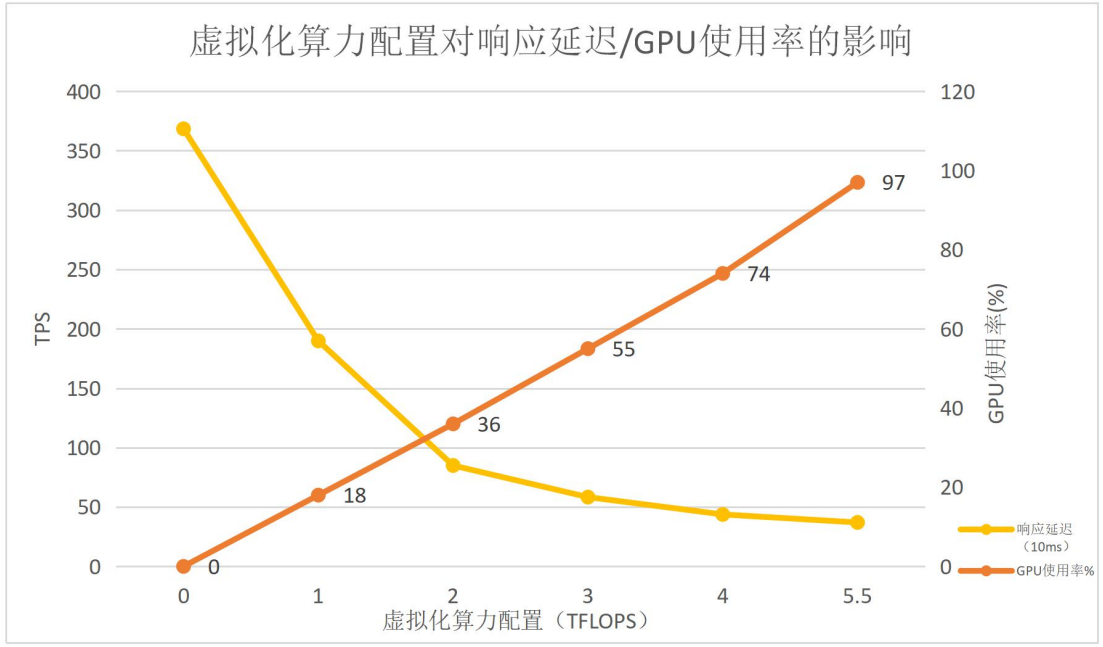
	Native_time (seconds)	vGPU_time (seconds)	Difference (%)
Tensorflow	47.82	47.88	0.13
Caffe	22.47	22.50	0.15
PyTorch	69.33	69.64	0.44
CNTK	7.39	7.41	0.27

原生兼容

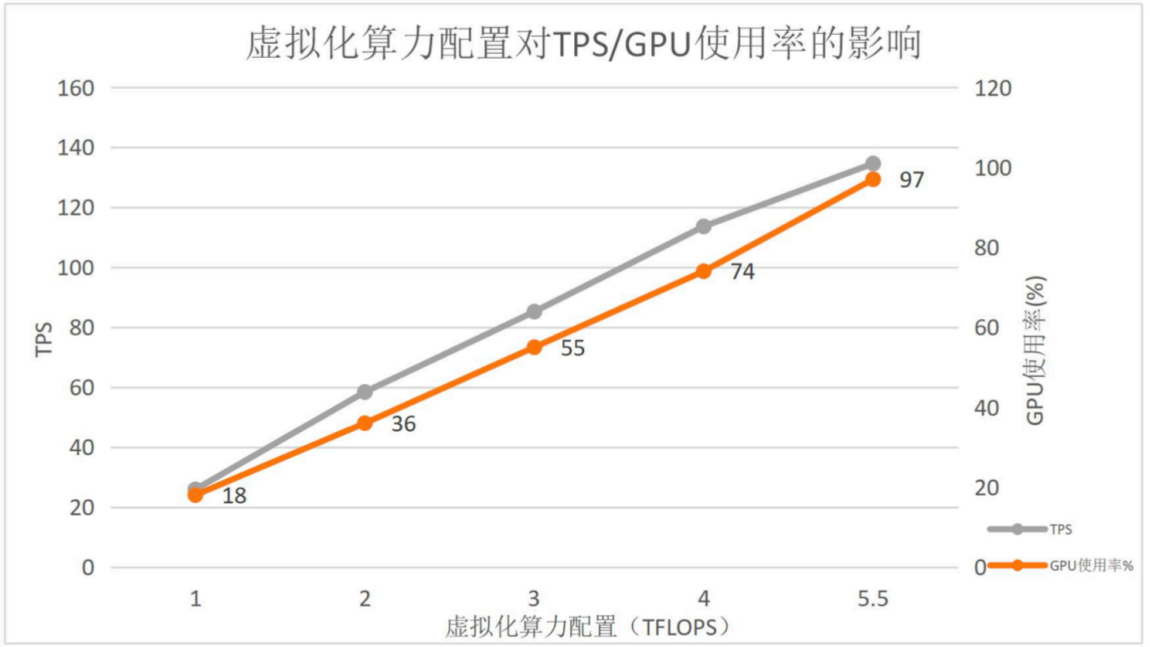

```
    imagePullPolicy: Always
    name: main
    resources:
      limits:
        cpu: "1"
        memory: 2Gi
        tencent.com/vcuda-core: 10
        tencent.com/vcuda-memory: 8
      requests:
        cpu: "1"
        memory: 2Gi
        tencent.com/vcuda-core: 10
        tencent.com/vcuda-memory: 8
    volumeMounts:
    - mountPath: /data/model/m
      name: model
```

- 原生兼容，不需要修改Kubernetes代码或容器镜像。
- 使用共享GPU执行应用程序应该就像在物理GPU上执行一样。
- GPU Manager 支持以可插拔的扩展组件形式安装、管理和删除。

算力分配效果



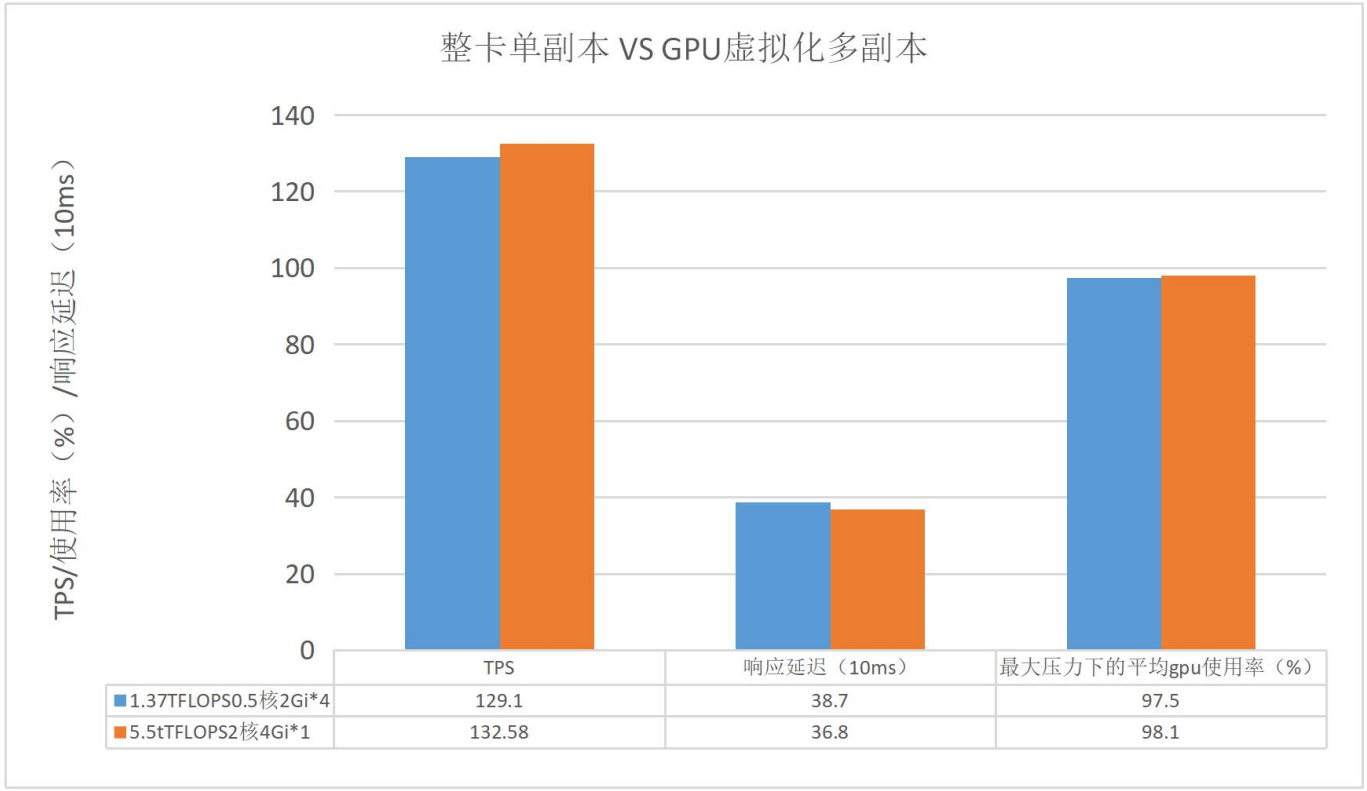
高QPS场景，GPU利用率和运算速度和分配的GPU资源息息相关



高QPS场景，服务吞吐量和分配的GPU量呈正相关关系

*数据来自生产环境实测

多副本并行使用效果



- TFLOPS总量相同，低算力多副本的情形和高算力单副本的响应延迟，TPS 接近相同。
- 多副本情形提供了更小的资源粒度，可以根据业务需求动态扩缩容，从而提高在云环境下的资源利用效率，节省成本。

*数据来自生产环境实测

商用状况

Tencent Cloud

TI·EMS

Tencent Intelligence Elastic Model Service

Tencent Kubernetes Engine

Tencent 腾讯

腾讯游戏 Tencent Games

腾讯优图

腾讯自动驾驶 Tencent Autonomous Driving

QQ空间

腾讯课堂

TEG Technology and Engineering Group 专业 服务 伙伴


KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2019
Thank
You