
DLRover
DLRover
Open Day
大模型时代的 AI 基建
2023 年 7 月 8 日
北京市朝阳区东三环环球金融中心（WFC）东塔 9F 大厅
OSCHINA
segmentfault

阿里云

ACK：云上面向AI的Infrastructure

阿里云云原生容器服务 云原生AI DLRover Open Day分享

霍智鑫

阿里云云原生容器服务

2023/07/08

云原生AI的能力提升

	工作项	原有方式： 从底层资源到上层框架，全手动	云原生AI解决方案： 自动、集成、端到端
深度学习环境搭建	安装配置	Make，Bazel或者pip安装，或者通过容器镜像	无需安装，支持cuda，TensorFlow，Pytorch，Caffe等环境
	分布式环境	通过SSH登录到每台机器上手工部署	无需挨个部署，一键完成整个分布式训练集群的构建
	GPU资源调度	手动记录管理，整机分配，使用效率低	自动化统一管理和调度
数据准备	数据集存储、共享	手动拷贝数据到每台机器上	利用OSS、NAS、CPFS等分布式存储一键挂载为本地数据卷； 并提供通用的数据缓存加速
模型开发	开发	手动安装Jupyter(或其他WebIDE)+Tensorboard	自动集成Jupyter(或其他WebIDE)+Tensorboard
模型训练	训练	通过SSH登录到每台机器上手工／脚本启动	命令行、图形界面设置训练参数，一键启动训练
	监控	GPU资源监控,需要登录执行nvidia-smi； 训练效果监控, 手动启动TensorBoard	GPU资源全方位监控；自动启动Tensorboard，提供访问链接
	checkpoint和模型导出	手动保存checkpoint和导出模型	自动将模型导出到分布式存储，支持checkpoint自动恢复
模型推理	模型发布上线	用户需自定义发布流程和系统	容器化CI/CD流程和工具链，支持各种发布策略（灰度、金丝雀、分批）
	线上运维	用户自建运维系统	内置微服务架构支持，完善的容器化运维体系集成云端监控、日志、弹性伸缩、负载均衡服务

01 ACK云原生AI套件能力

- AI套件产品能力全景
- GPU能力
- 调度能力
- 弹性数据集Fluid
- AI工具Arena

02 AI套件中的弹性训练

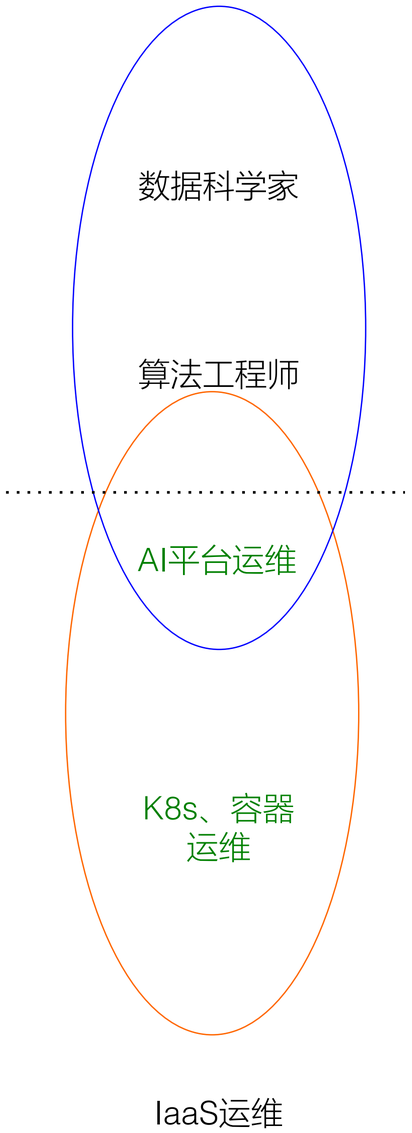
- Elastic Horovod
- Elastic Training Operator
- ACK上的弹性训练

03 技术展望

- 规划中的能力支持

ACK云原生AI套件能力

云原生AI套件产品能力全景图



AI领域模型

Transformer, ResNet, InsightFace, DeepStream, CTR, …

AI计算框架

Tensorflow, Pytorch, Horovod, TensorRT, TF-Serving, Triton, Seldon, PAI, …

AI容器仓库

弹性分布式训练

弹性推理服务

数据集管理

分布式缓存加速

大数据服务集成

任务生命周期管理

工作流

批作业调度

拓扑感知调度

数据感知调度

GPU / vGPU管理

GPU共享

NPU调度

RDMA调度

AI运维体系

监控
自愈
日志
容错
可视化
弹性伸缩

场景一：异构计算

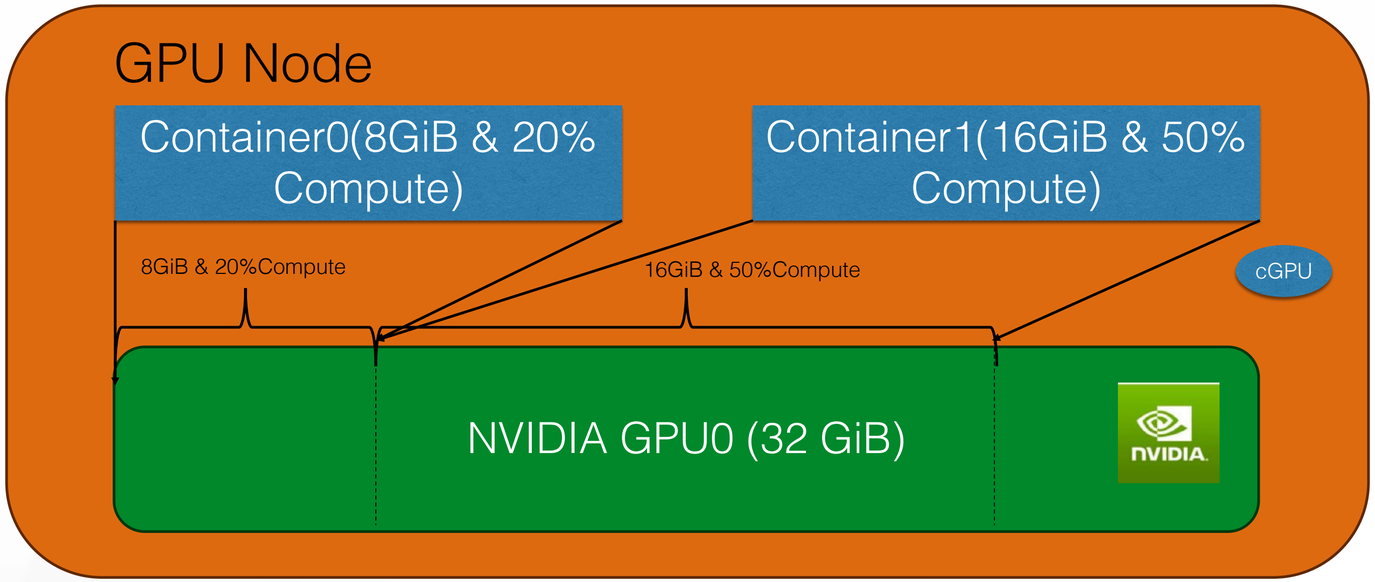
- 一键部署CPU/GPU/vGPU/NPU/RDMA/FPGA异构计算集群，统一管理运维
- 多维度GPU监控、健康检查和告警
- 多种GPU调度策略（共享+隔离、优先级、Topology aware）
- 自动挂载共享存储（NAS, CPFS, OSS）
- 自动弹性伸缩灵活配置

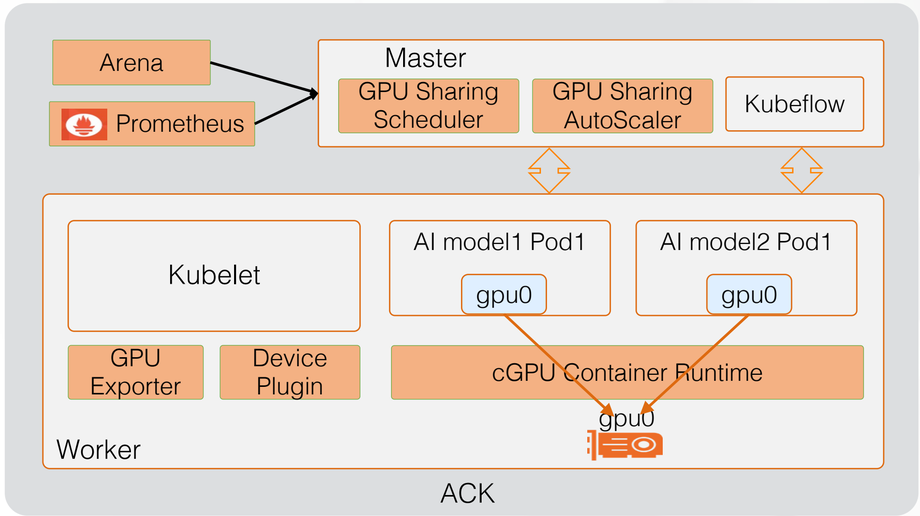
场景二：深度学习任务

- 屏蔽底层结构快速开启深度学习任务
- 端到端的深度学习任务生命周期（模型开发-训练-推理）
- 支持TensorFlow, Pytorch, MXNet, Horovod等和PAI，AIACC阿里自研优化框架
- 支持Spark, Flink, Presto等大数据服务
- 任务级调度策略（Gang, Binpack, Capacity, 优先级队列）
- 弹性数据集管理和分布式数据缓存加速
- 集成阿里云基础服务，综合性能优化

共享GPU调度

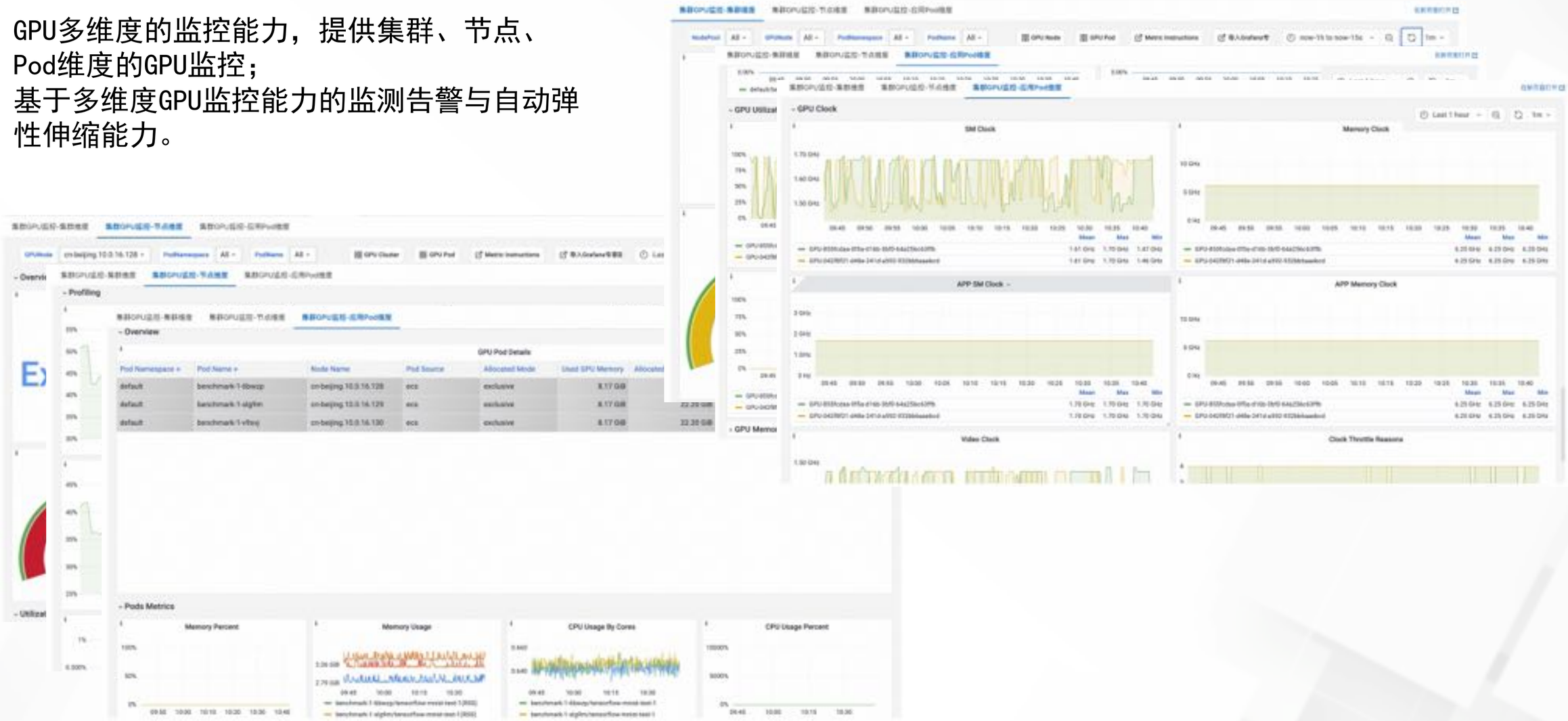
- 业界首款Kubernetes GPU独占和共享调度方案，应用代码0侵入
- 结合cGPU技术支持多模型共享显存时严格隔离，以及GPU算力分时调度，同时避免虚拟化开销
- 为AI推理服务提供安全地GPU利用率提升与优化



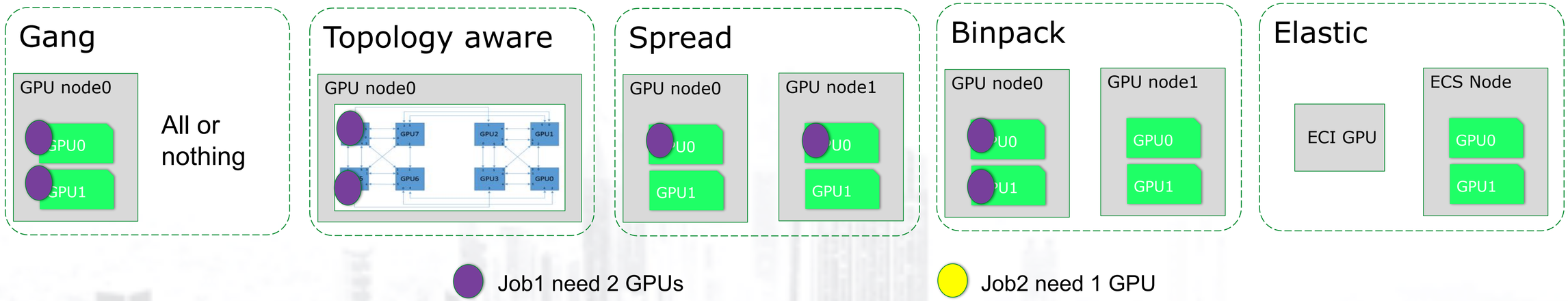


GPU监控能力

- GPU多维度的监控能力，提供集群、节点、Pod维度的GPU监控；
- 基于多维度GPU监控能力的监测告警与自动弹性伸缩能力。

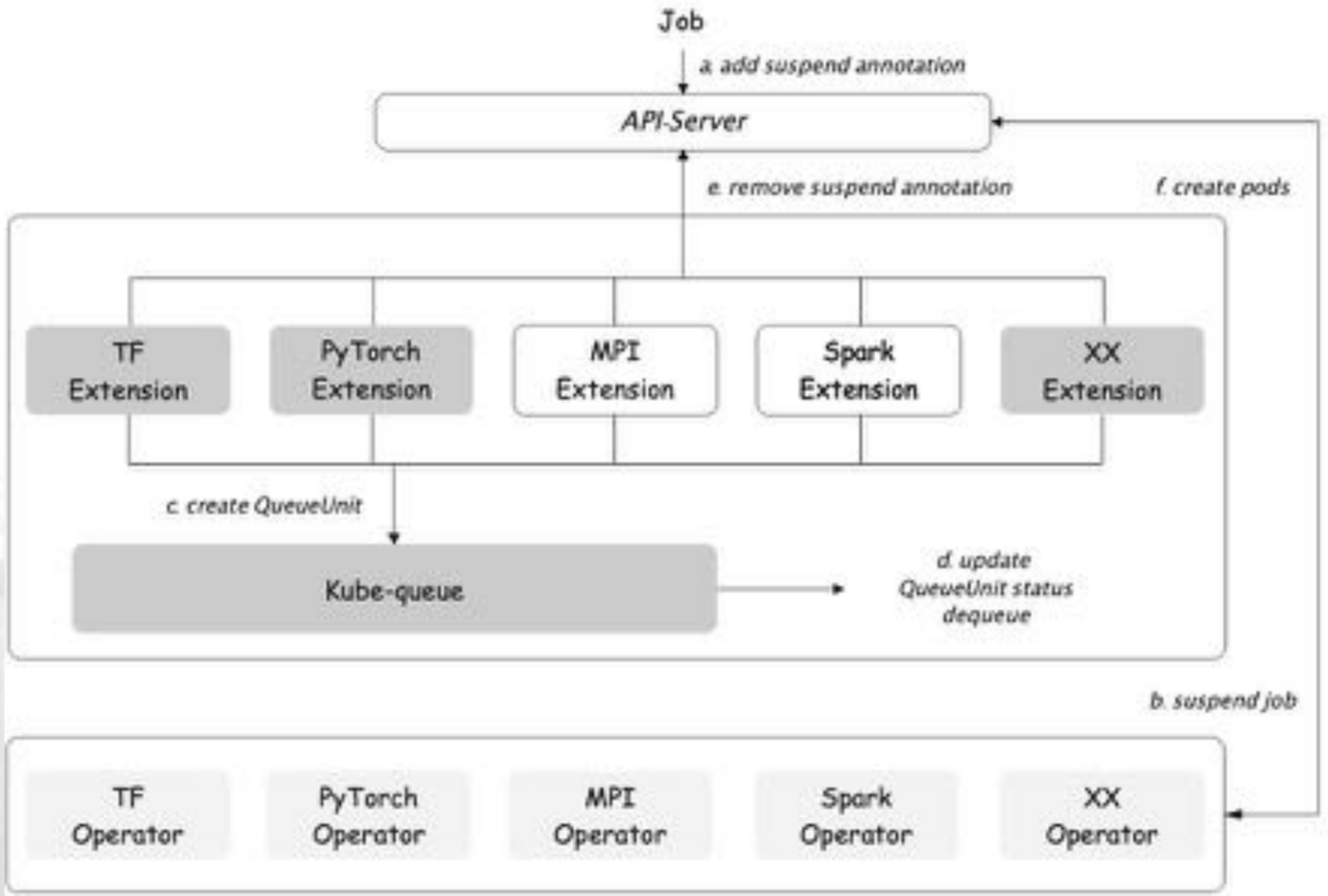


多种资源调度策略能力



- **Gang**: 作业的所有子任务都能满足资源需求才整体分配，否则不分配任务资源。防止大作业挤占小作业。
- **Topology Aware**: 多卡需求自动选择最佳P2P连接（NVLink, NUMA, RDMA, PCIE）分配
- **Spread**: 作业均匀分配在各个节点，集群利用率较平均
- **Binpack**：作业优先集中分配在某个节点，当节点资源不足时，依次在下一节点集中分配，适合单机多卡训练任务，避免跨机数据传输。防止资源碎片。
- **Elastic**：在应用发布或扩容过程中，自定义资源策略（ResourcePolicy），设置应用实例Pod被调度到不同类型节点资源的顺序（ECI、ECS）。同时在缩容过程中按照原调度顺序逆序缩容。

Kube-Queue任务调度队列

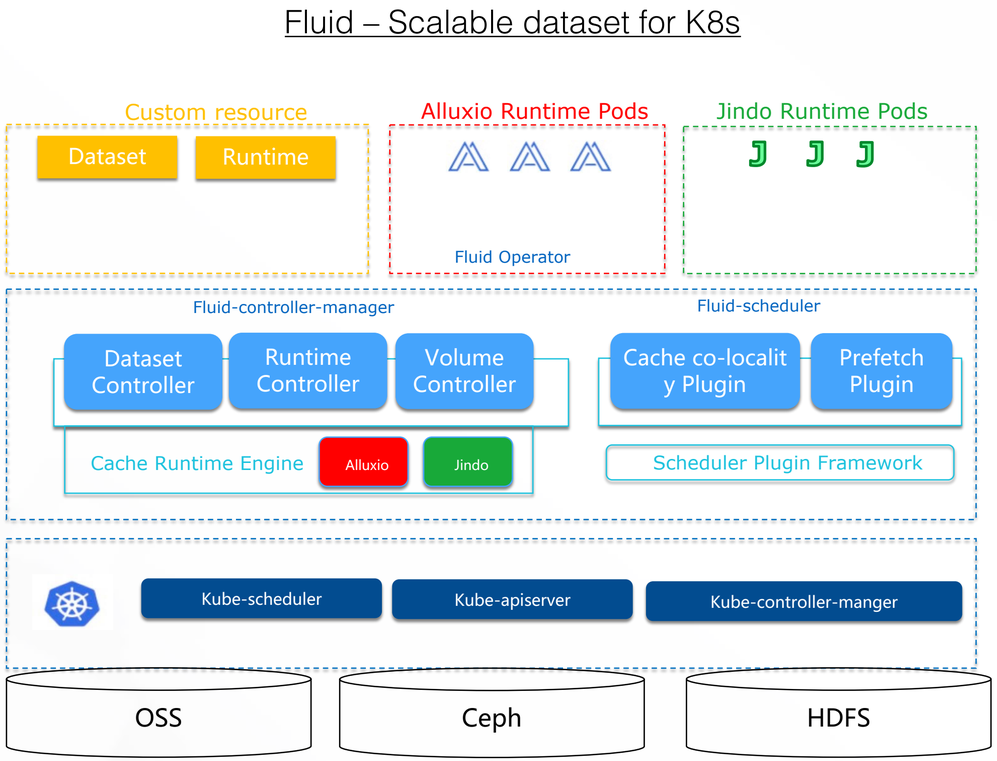


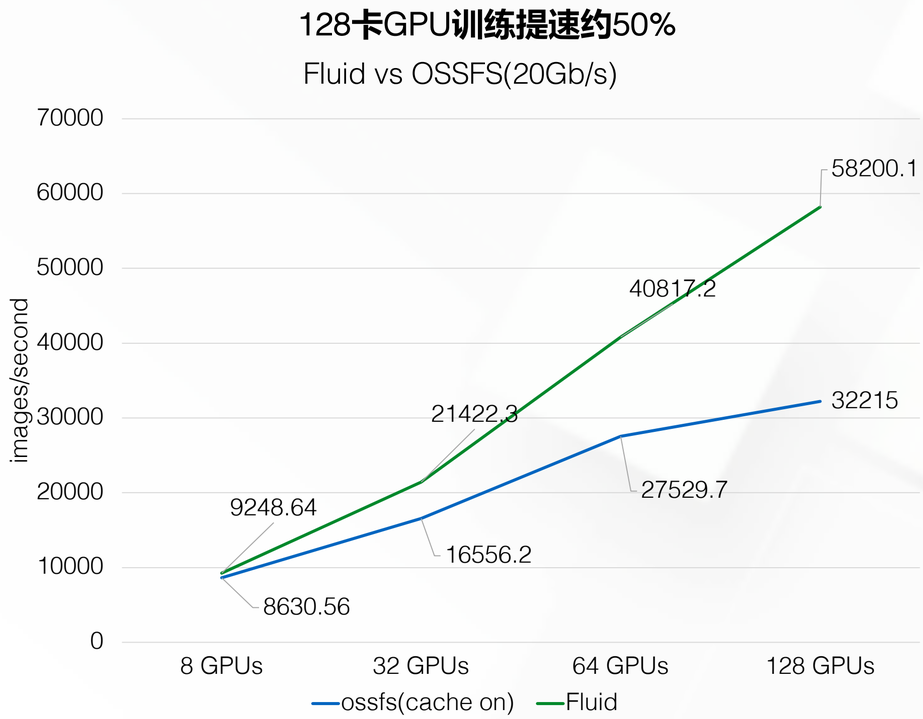
Kube-Queue旨在管理Kubernetes中的AI/ML工作负载和批处理工作负载。允许系统管理员使用自定义队列的作业队列管理，以提高队列的灵活性。结合Quota系统，Kube-Queue自动优化了工作负载和资源配额管理，以便最大化利用集群资源。

https://github.com/kube-queue/kube-queue

Kubernetes弹性数据集-Fluid

- 通用的深度学习训练数据加速方案，提升数据读取速度，改善GPU计算效率
- 统一提供FUSE接口，支持OSS / HDFS / NAS / PVC等多种后端存储
- 分布式缓存, 横向扩展，本地支持RAM / SSD / HDD多级cache
- 支持冷加载、预热，缓存配比等丰富控制策略，适配不同场景
- 插件扩展机制，支持各种分布式缓存引擎，如Jindofs，Alluxio



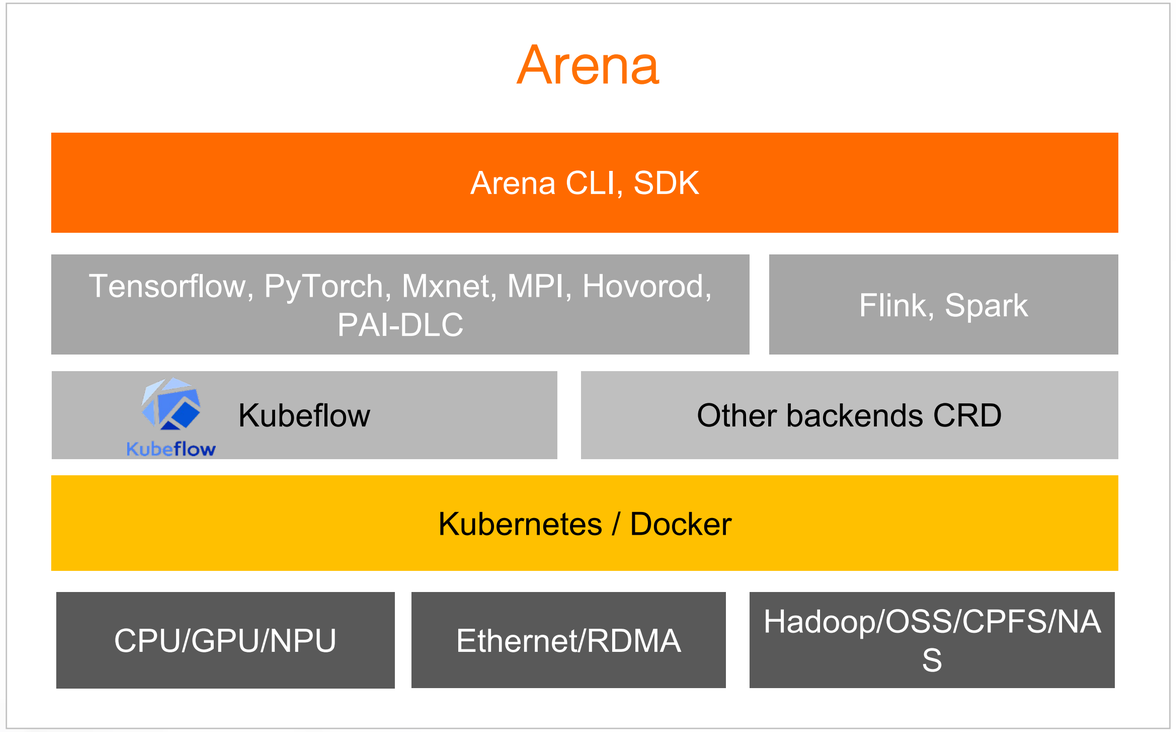


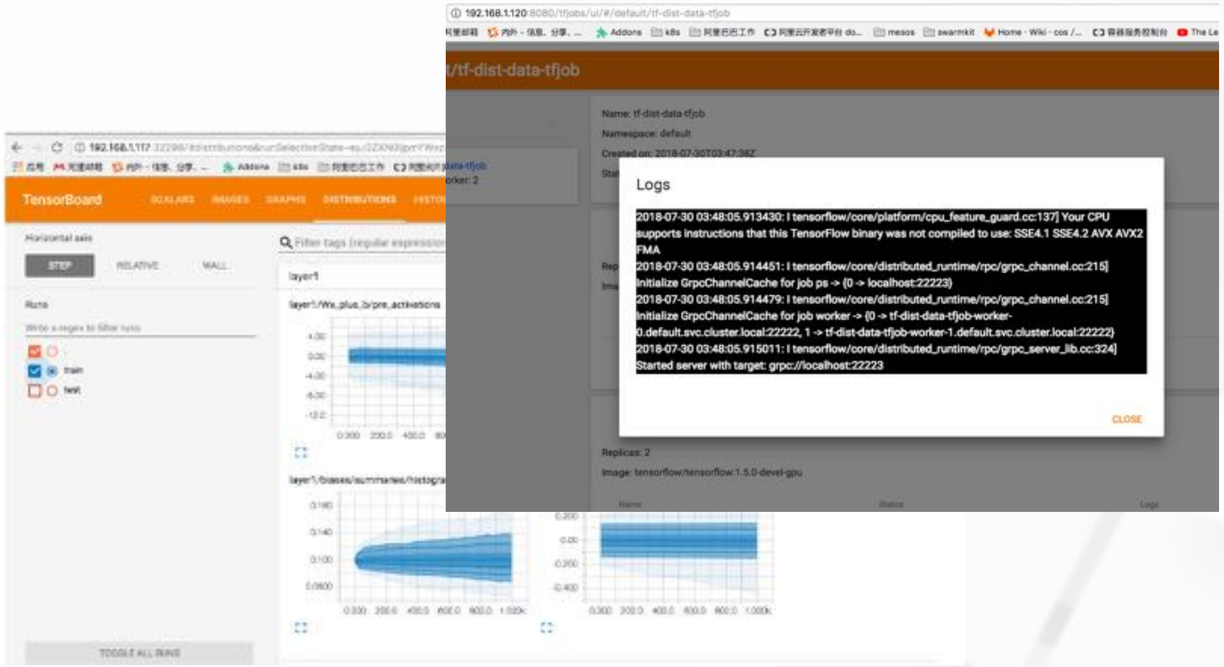
https://github.com/fluid-cloudnative/fluid

Arena基本能力

Kubeflow/arena

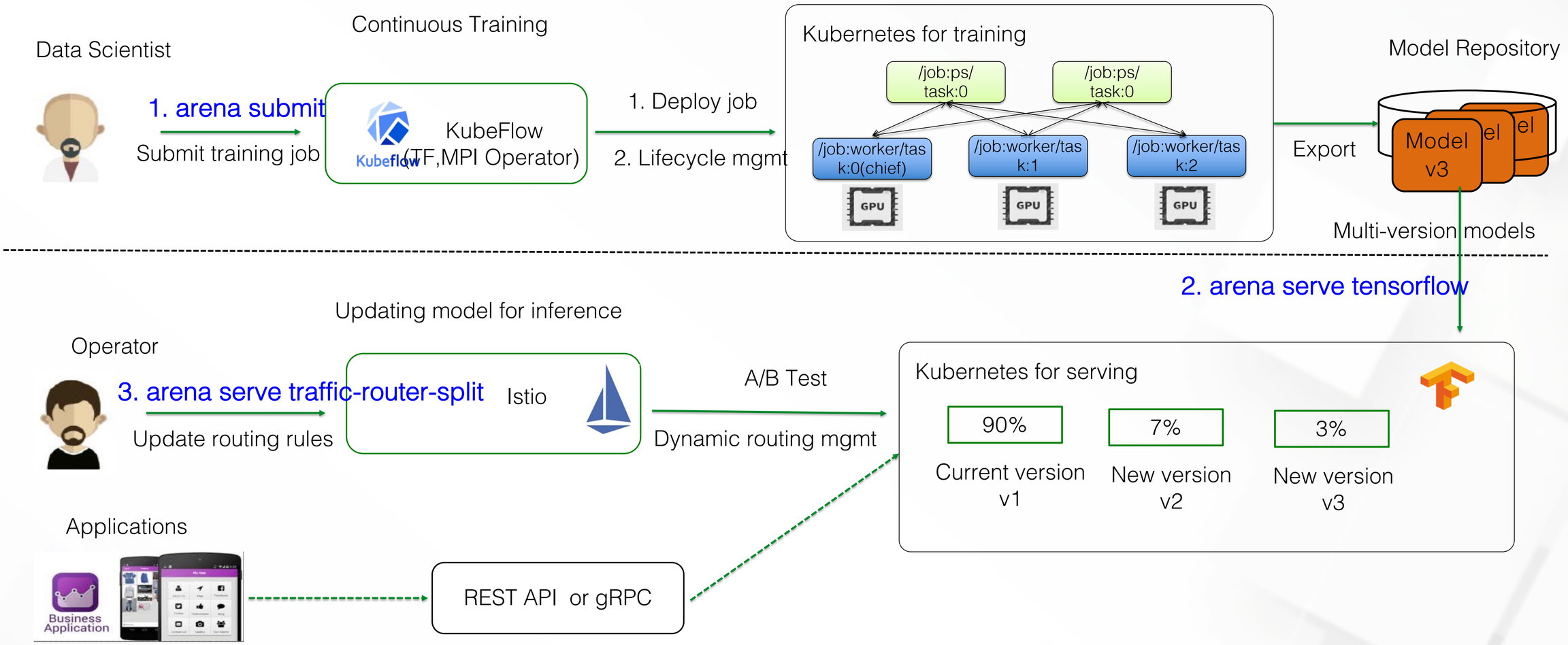
- 用一个工具屏蔽所有底层资源、环境管理、任务调度、GPU分配和监控的复杂性
- 兼容多种深度学习框架 – Tensorflow, Caffe, MPI, Hovorod, Pytorch, DeepSpeed等
- 提供命令行，Golang/Java/Python SDK，
- 深度学习生产流水线 – 训练数据管理，任务管理，模型开发，分布式训练、评估，推理上线等全流程
- 已开源贡献到Kubeflow社区





阿里云开源贡献给kubeflow社区 https://github.com/kubeflow/arena

Arena支持从训练到推理



不同视角下的AI系统工程

数据科学家 / 算法工程师

Configuration

Data Collection

Data Verification

ML Code

Feature Extraction

Machine Resource Management

Analysis Tools

Process Management Tools

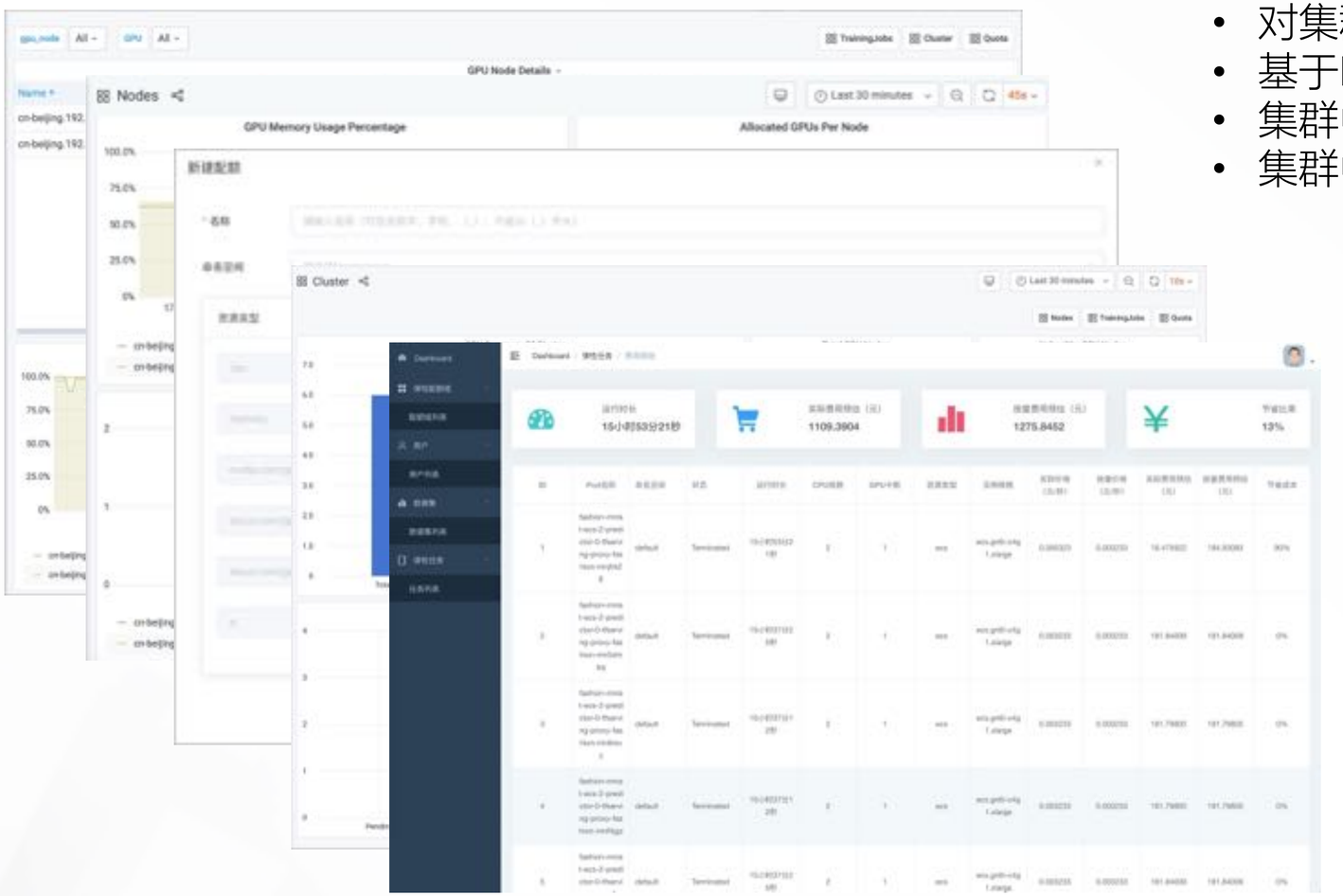
Serving Infrastructure

Monitoring

AI平台运维

控制台能力

ACK集群管理员（运维控制界面）

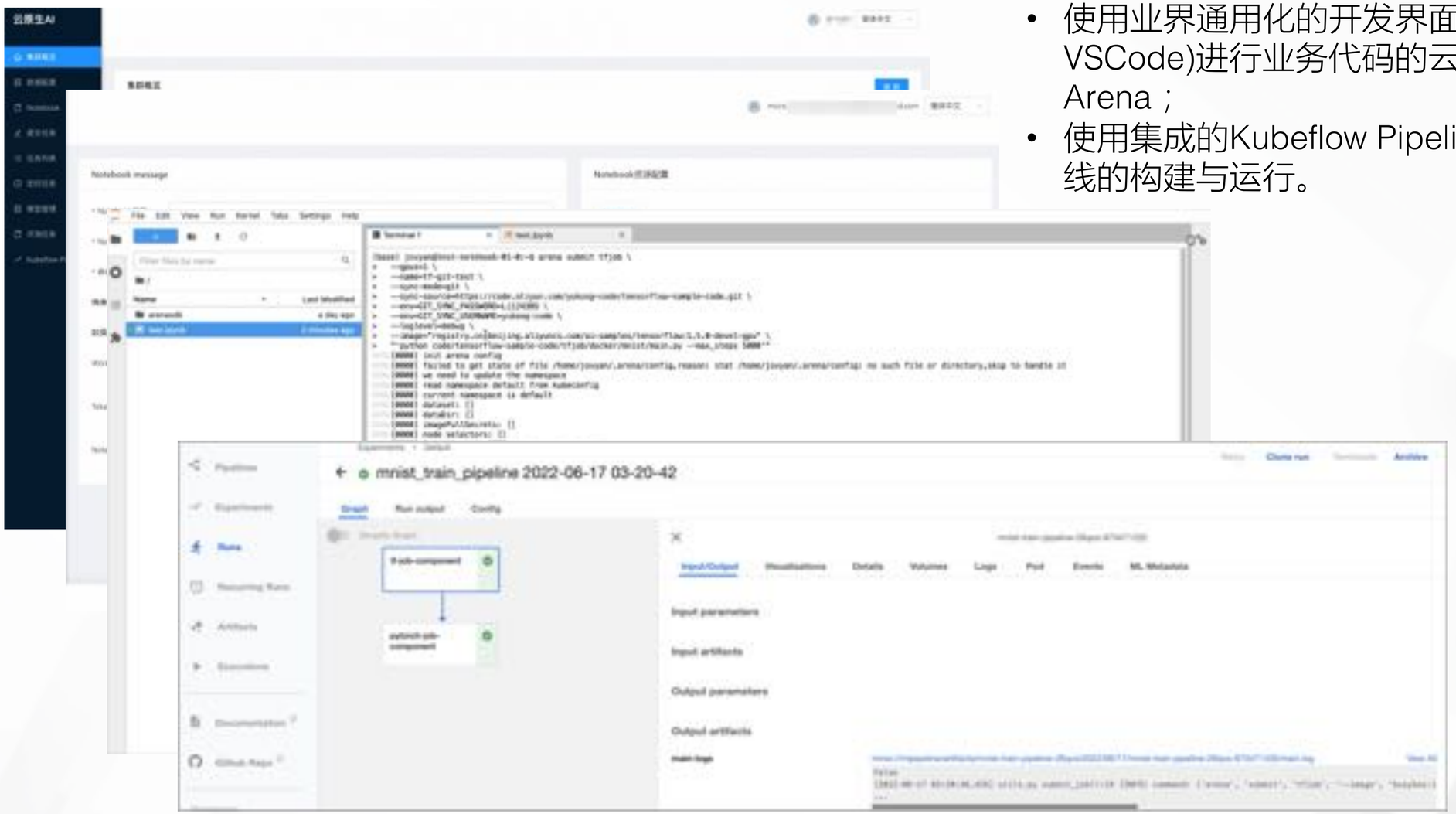


面向ACK运维管理员的控制台使用界面

- 对集群中训练任务和资源的精细化监控大盘；
- 基于Elastic Quota的多租户隔离用户组的配额管理；
- 集群中弹性数据集的统一创建与管理；
- 集群中训练任务的成本管控与资源管理。

控制台能力

算法开发工程师（算法开发与提交使用界面）



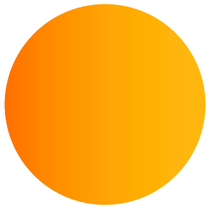
面向算法开发工程师的控制台使用界面

- 使用便捷化的UI界面进行训练任务的提交与状态查看；
- 使用业界通用化的开发界面Notebook(支持Jupyter和VSCode)进行业务代码的云上开发，并在其中操作Arena；
- 使用集成的Kubeflow Pipeline进行机器学习流程化流水线的构建与运行。

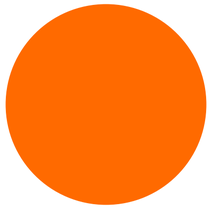
AI套件中的弹性训练

弹性训练的意义

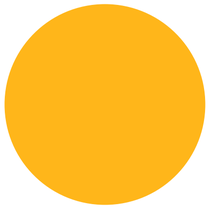
阿里云



大规模分布式训练的容错



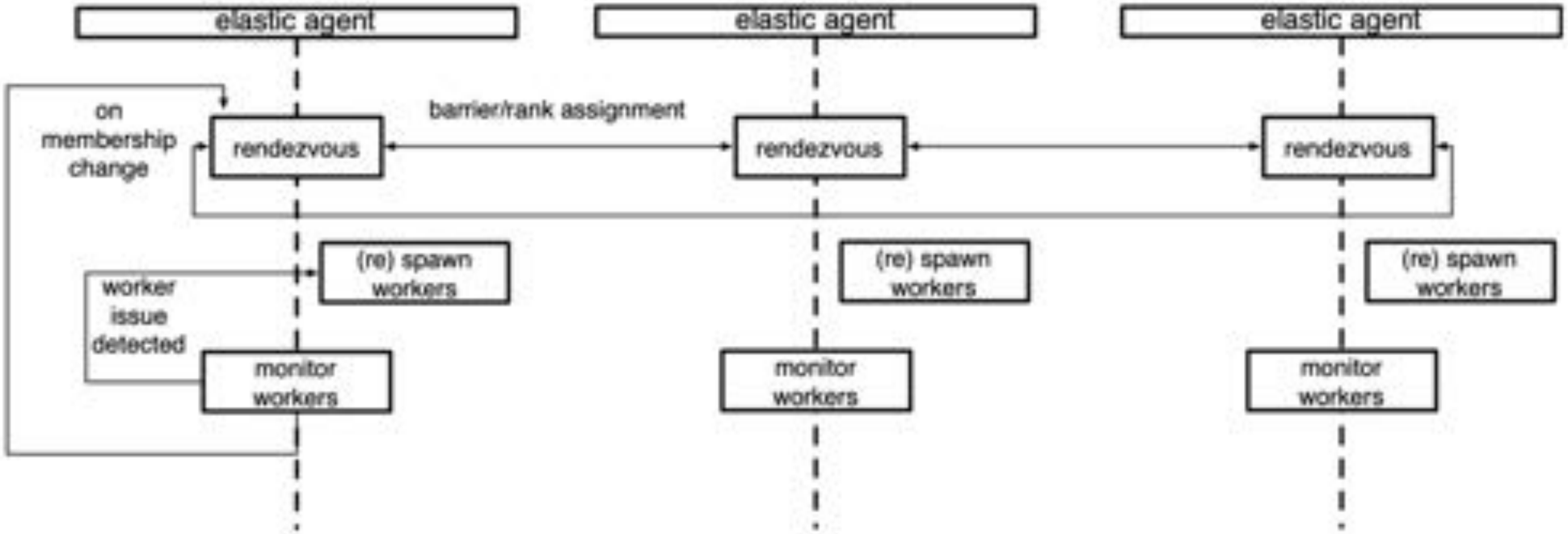
集群算力资源利用率的提升



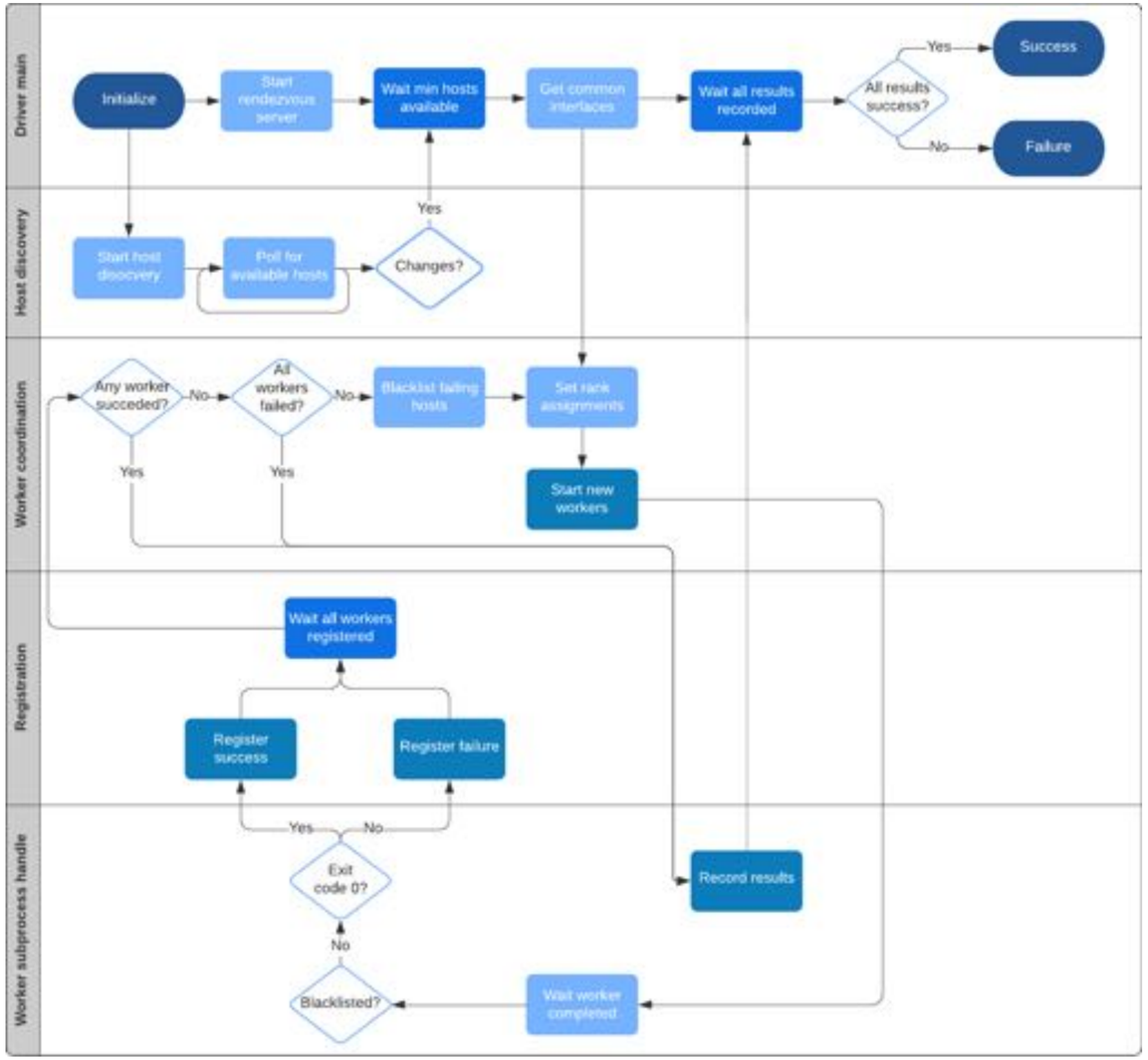
训练任务成本的降低

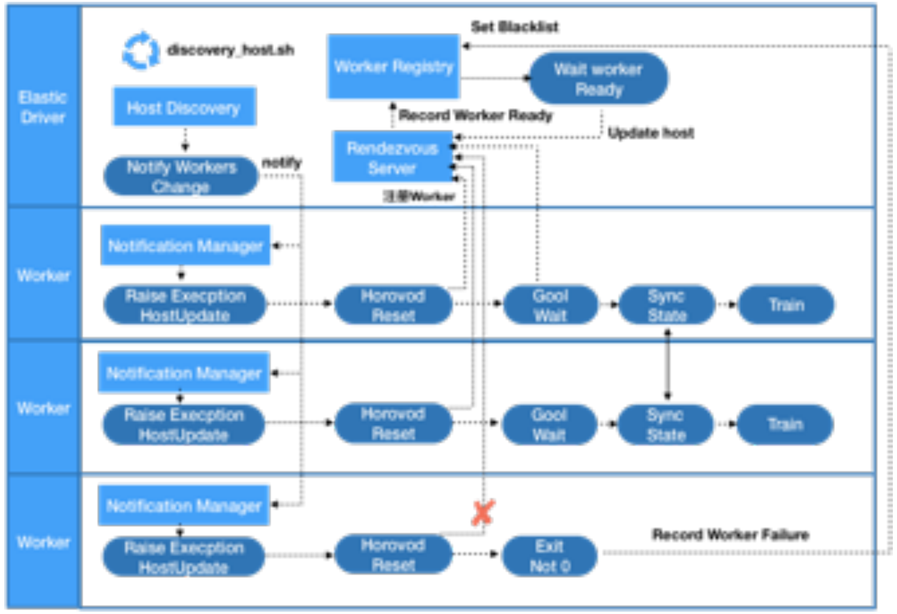
Elastic Pytorch

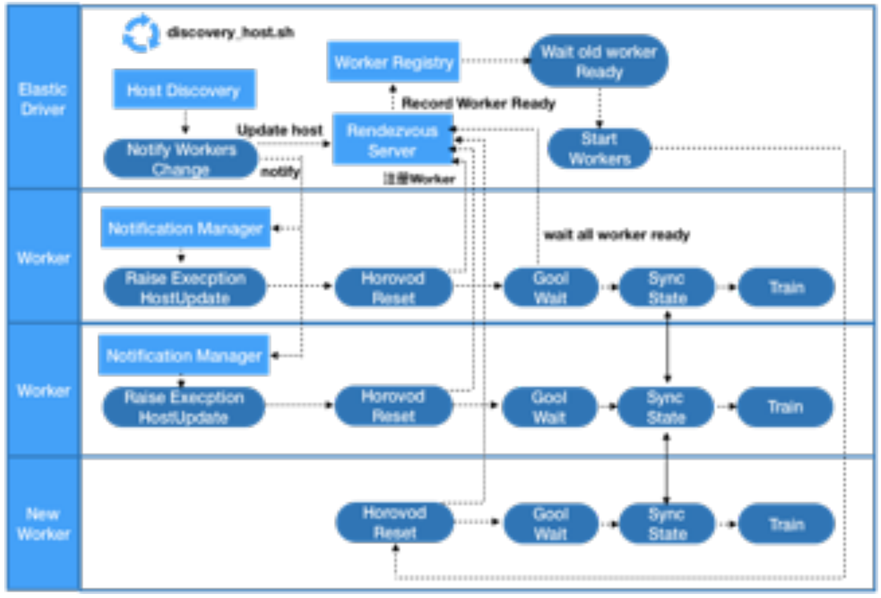
PyTorch



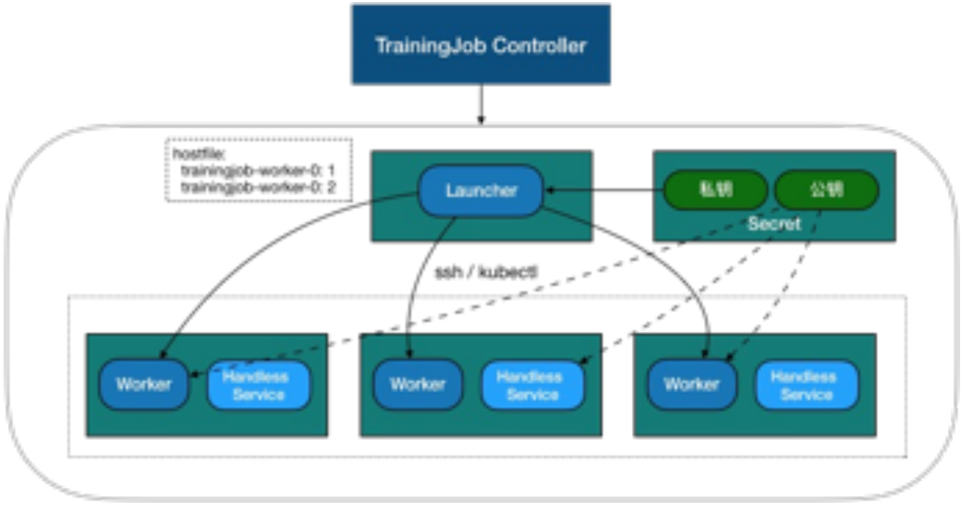
Elastic Hovorod

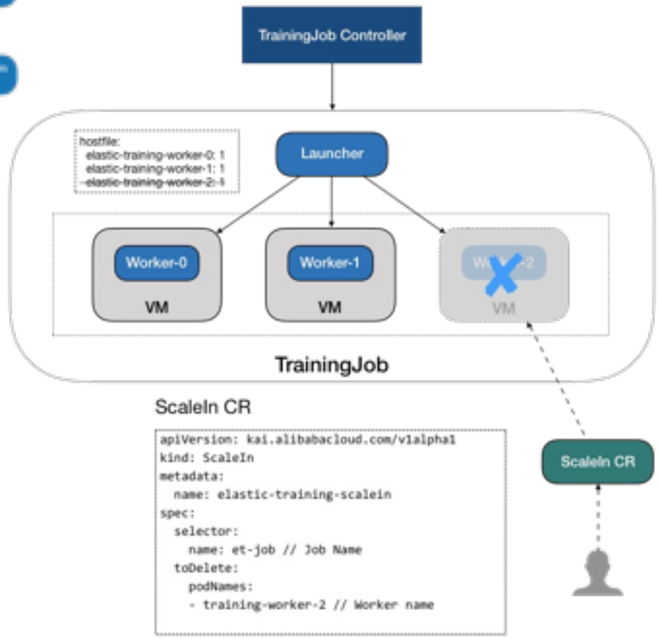


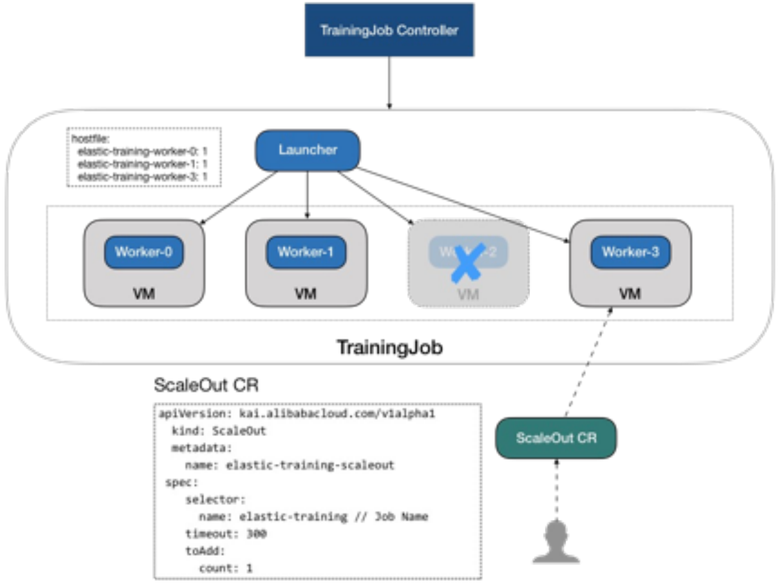




ACK上基于Elastic Hovorod的弹性训练







开源地址： https://github.com/AliyunContainerService/et-operator

弹性训练场景

随着模型不断增大，AI作业训练成本不断判断攀升，节省成本逐渐称为各行各业的关键命题。

面向在ACK上做AI模型训练且成本敏感的客户，我们在ACK上期望推广的弹性训练场景为基于抢占式实例Spot的弹性节点池作为底层训练资源的云原生AI弹性训练解决方案。

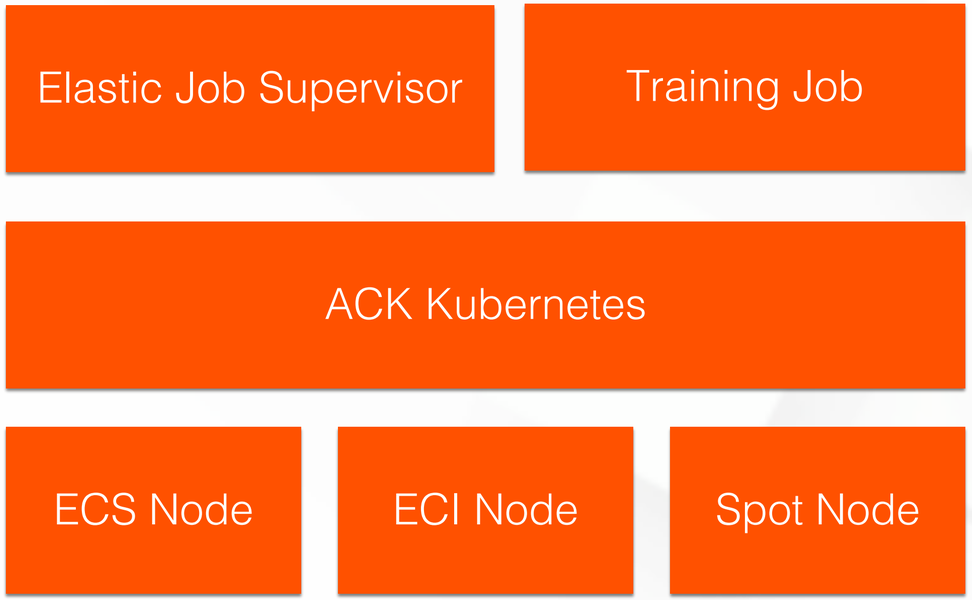
整体方案的目标在于以下几点：

1. 期望将更多类型更多训练场景的AI训练任务在集群中以弹性的方式尽可能多的运行在成本更低的抢占式实例上；

2. 这些训练任务可以根据客户需求动态的占用集群中空闲的资源，以达到资源利用率提升的目的；

3. 使用该种弹性训练方式对客户AI训练任务的精度影响处于一个可以接受的范围内，不影响其最终的效果表现；

4.使用该种弹性训练方式可以使得客户的训练任务不会因为资源回收或者其他原因而导致整个任务进程的中断，进而丢失训练结果。



目前在ACK上，云原生AI套件提供了对Elastic Horovod、DLRover (Tensorflow PS)、Elastic Pytorch的支持，可以覆盖对NLP、CV、搜推广场景的AI训练任务的支持，基本上涵盖了目前市面上的绝大多数的AI任务训练场景，具有很好的可推广性。

现有弹性训练能力

ACK上基于抢占式实例Spot的弹性训练能力

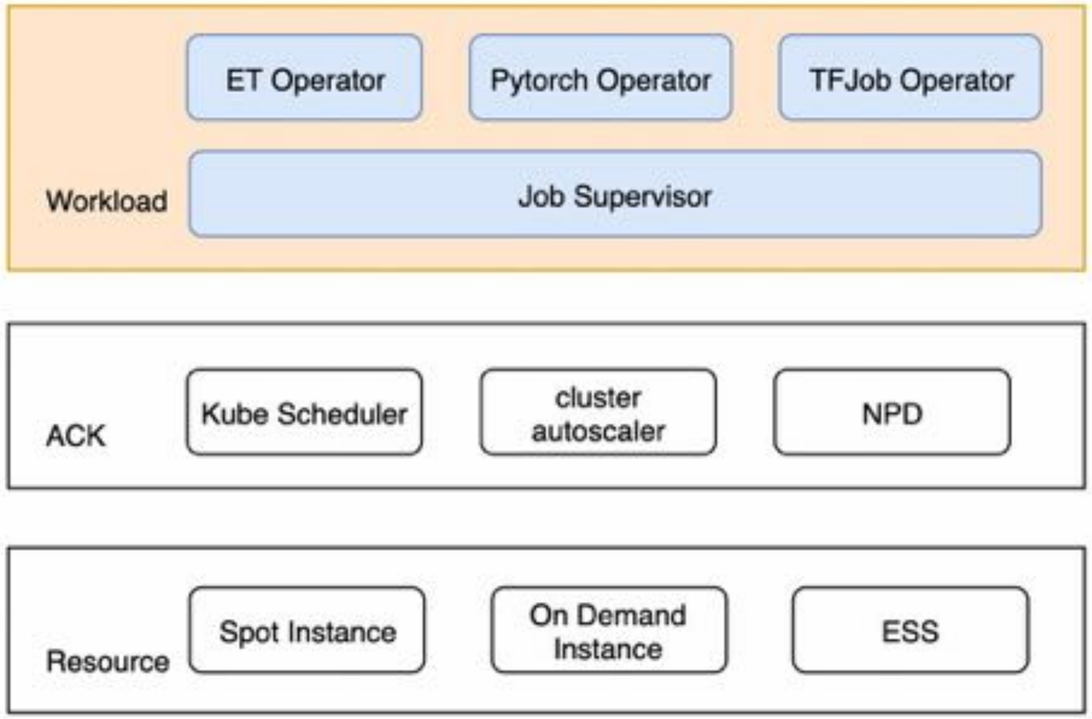
- Max Wait Time：若最大等待时长之前无法满足训练任务的资源请求时，则任务终止资源的等待，避免部分worker申请资源后造成的浪费；
- Checkpoint Saving：拥有实例回收通知机制，使得训练任务在接收到抢占式实例回收的通知时进行自动的Checkpoint Save操作，以避免训练结果的丢失；
- Fail Tolerance：提交了一个分布式弹性训练任务，当部分实例被回收时，该分布式训练任务可以做到仍继续运行，不会因为部分Worker的回收而导致中断；
- Job Recovery：当集群中重新加入训练可用资源时，之前由于资源不足而Suspended的任务可以重新拉起继续进行训练，或者之前被缩容的分布式训练任务可以自动扩容到预设的Replica进行训练；
- Cost Observability：在使用抢占式实例进行训练时，可以通过对整体的训练成本的监控计算，展示基于抢占式实例Spot的弹性训练带来的成本介绍。

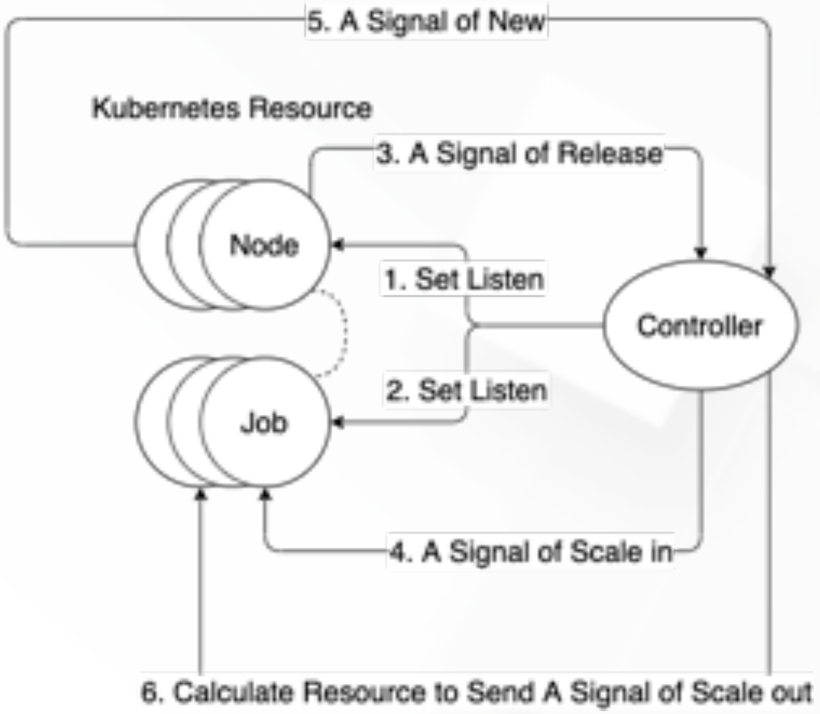
DLRover
Deep Learning Rover

PyTorch

弹性训练能力架构原理

Elastic Job Supervisor的弹性训练实现架构

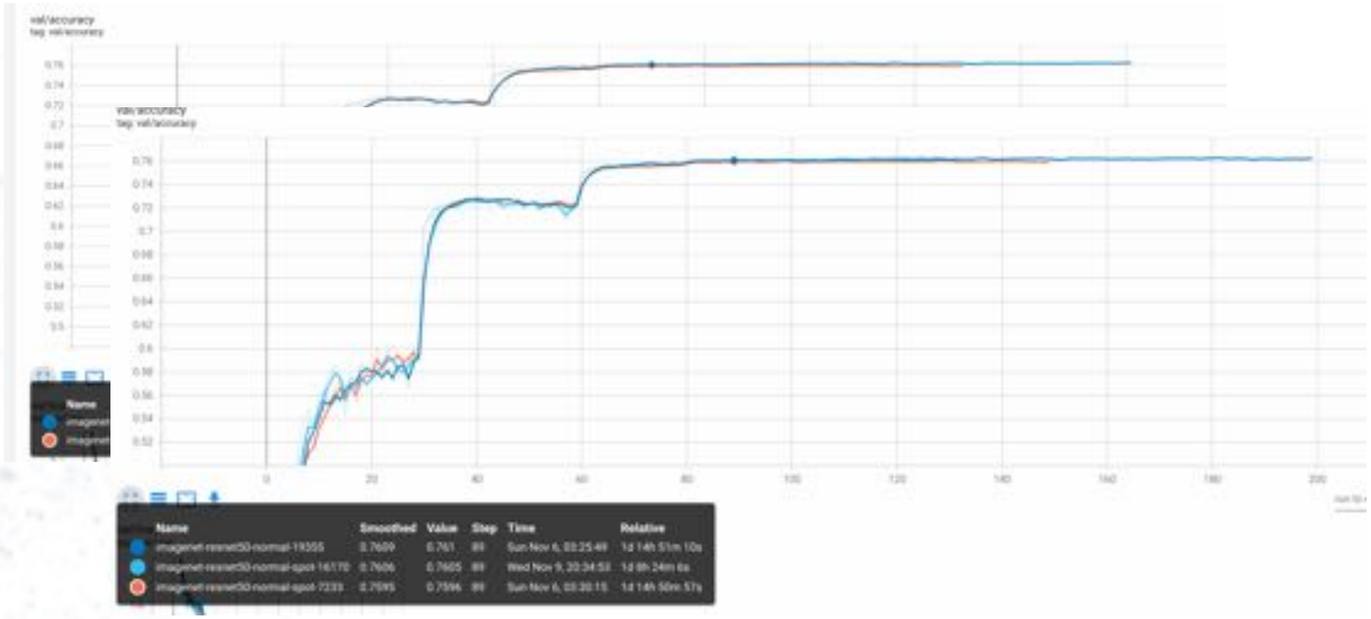


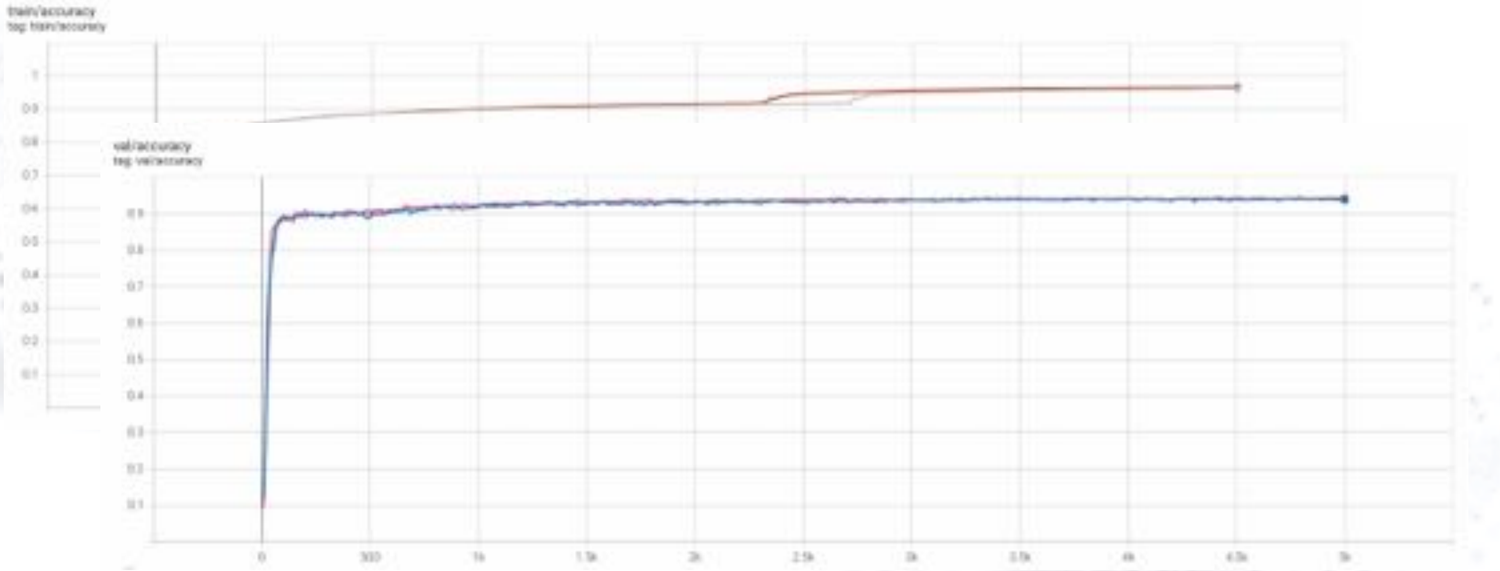


弹性训练实验效果

分别在三个场景下进行弹性训练的实验：

1. ResNet-50（Computer Vision）
2. BERT（NLP）
3. Deep & Wide（Recommend）

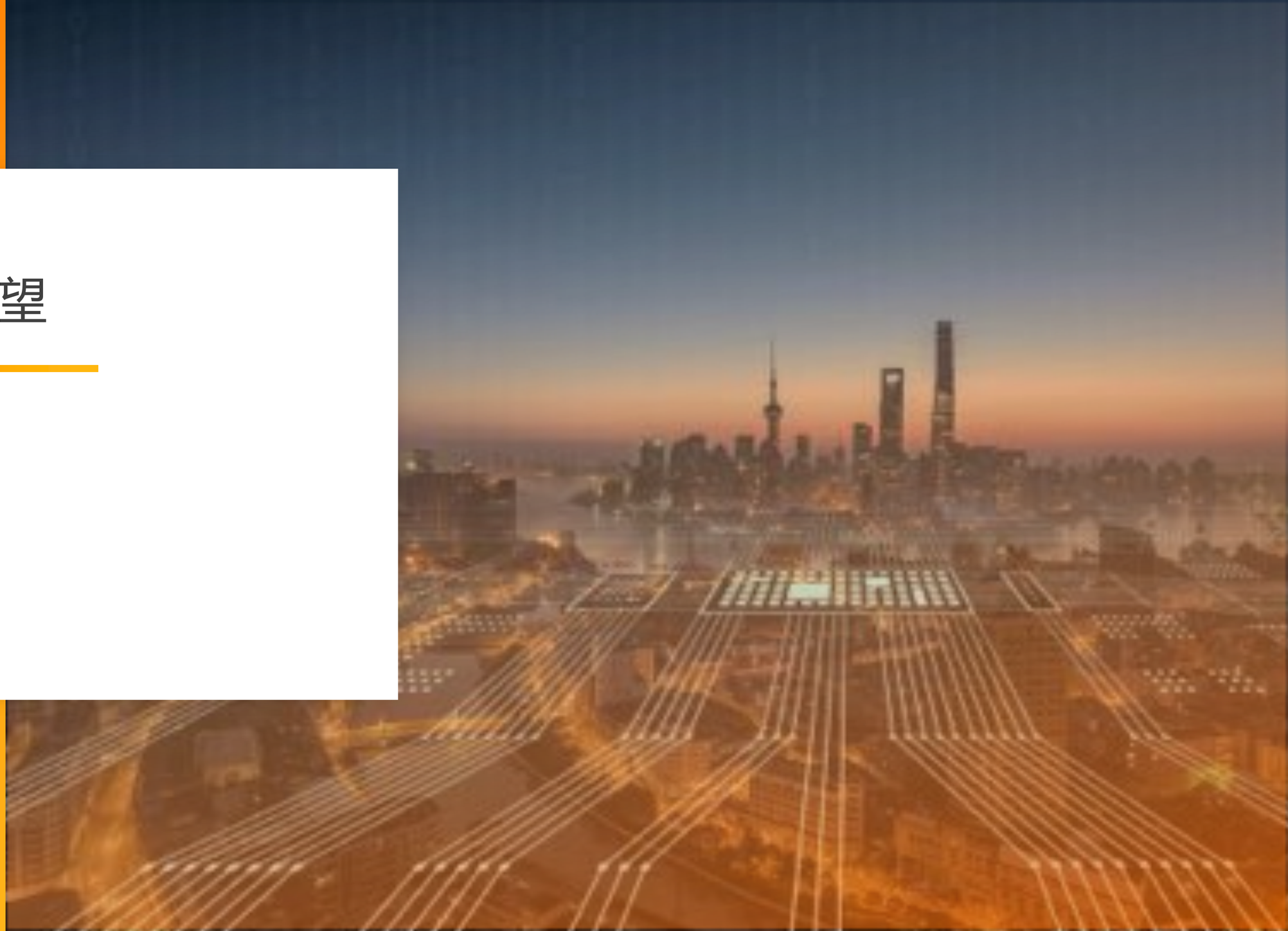




（精度）通过实验验证，不同类型的分布式训练的副本数在一定的范围进行弹性的改变，其对精度的影响都处于可以接受的范围之内。

（成本）通过在抢占式实例上进行弹性变化可以在整体上将整个训练任务的花费成本降低到一个比较可观的值。与正常的按量付费云资源比起来，在ResNet上的测试可以达到92%的成本节省，在BERT上的测试可以达到81%的成本节省。

技术展望



挑战:

- 对大模型的能力支持（Arena、弹性训练）；
- 多种资源模式的扩展与支持；
- 异构设备利用率的进一步提升与Profiling。

应对:


阿里云