START

2018 THINK IN CLOUD BEIJING

如何利用公有云快速落地AI应用

宋 翔　LabU

人脸识别

图片分类

语音识别

语音合成

物体追踪

OCR

物体检测

语义分析

特征分析

AI行业赋能

零售

教育

安防

游戏

医疗

金融

什么场景需要AI赋能?

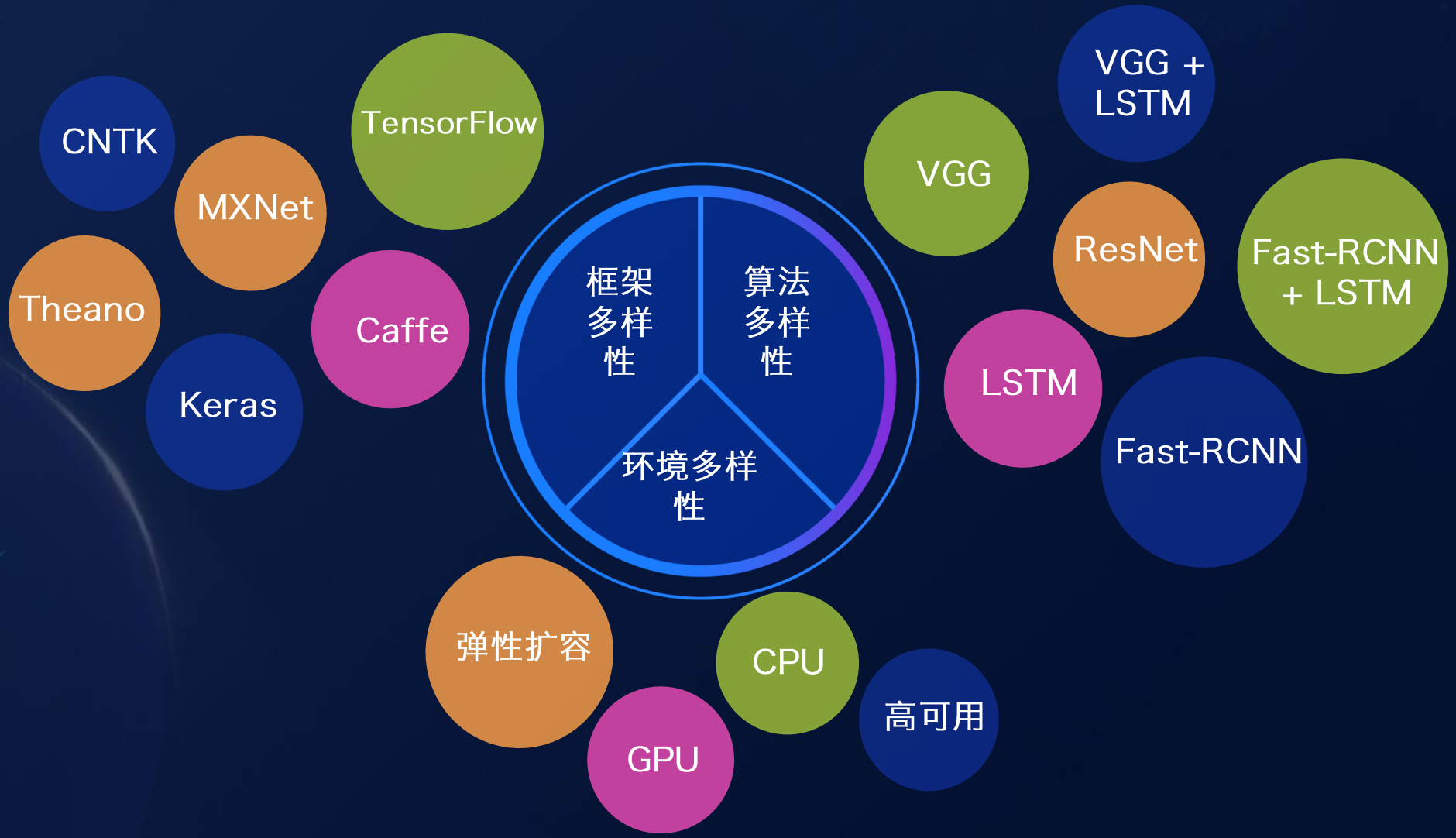
如何快速、低投入地验证AI技术?

如何快速展开AI应用业务?

如何高效实现AI应用迭代?

目 录

AI 的选择

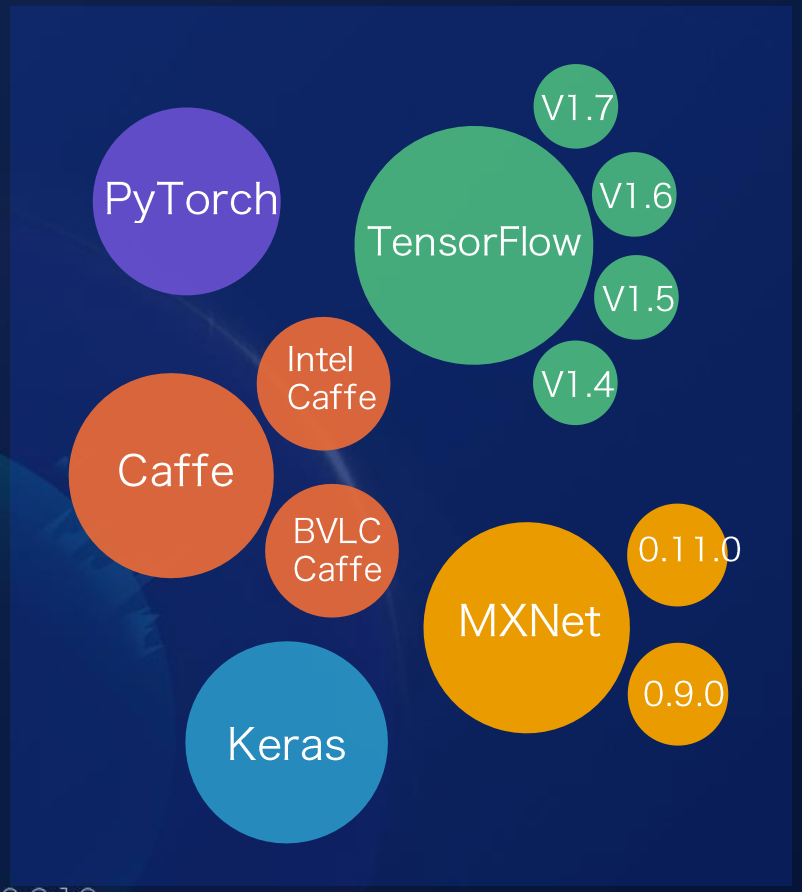


AI落地的挑战

- AI技术验证
- 展开AI应用业务
- AI应用迭代

- AI算法的选择
- AI框架的选择
- 硬件设备的选用
- 软件环境配置
- 在线服务环境的部署
- AI集群的成本
- 在线服务系统的可扩展性
- 在线服务系统的可靠性
- AI算法模型的优化
- Ai应用更新的发布

AI的挑战I: 基础环境

AI 框架	依赖环境	硬件设备	存储
	Python 2.7	K80	对象存储
	Python 3.6	P40	HDFS
	numpy	P40 X 4	NFS
	cv2	V100	本地存储
	pillow	Xeon	
	cuda9	KNL	
	cudnn 6		
	cudnn 7		

AI的挑战II: AI系统建设

算法兼容性	平台扩展性	分布式化	纵向扩展
更好地兼容各类AI框架和算法	平台具备横向扩展能力，支持业务规模的不断扩大	具备弹性伸缩的能力以及容灾能力	支持CPU、GPU 支持S3、NFS、HDFS等多种存储

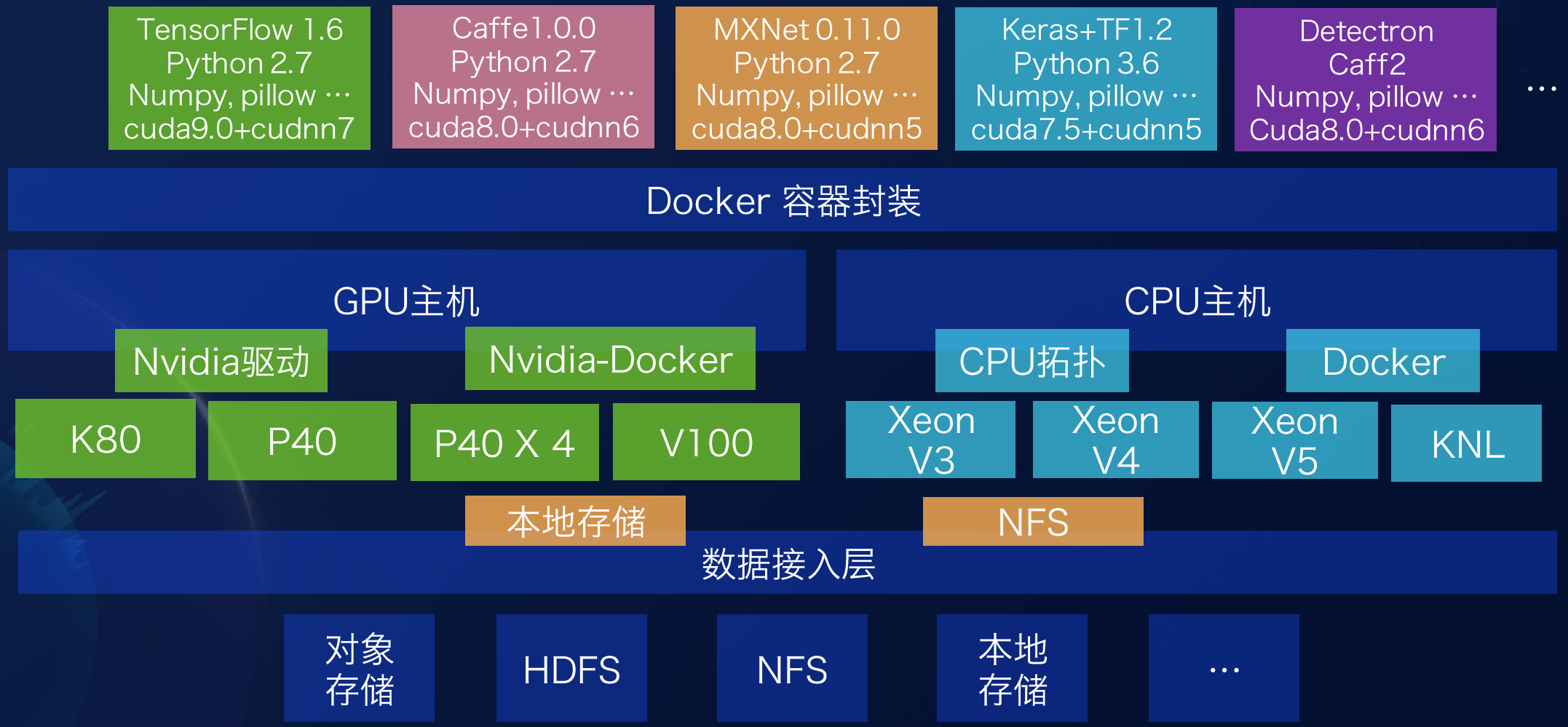
AI的挑战III: 投入产出

调研投入	研发成本	资源成本	运营成本
高效、低投入 快速调研、验证	专注AI应用研发	降低训练资源成本 降低在线服务资源成本	降低资源运营管理成本

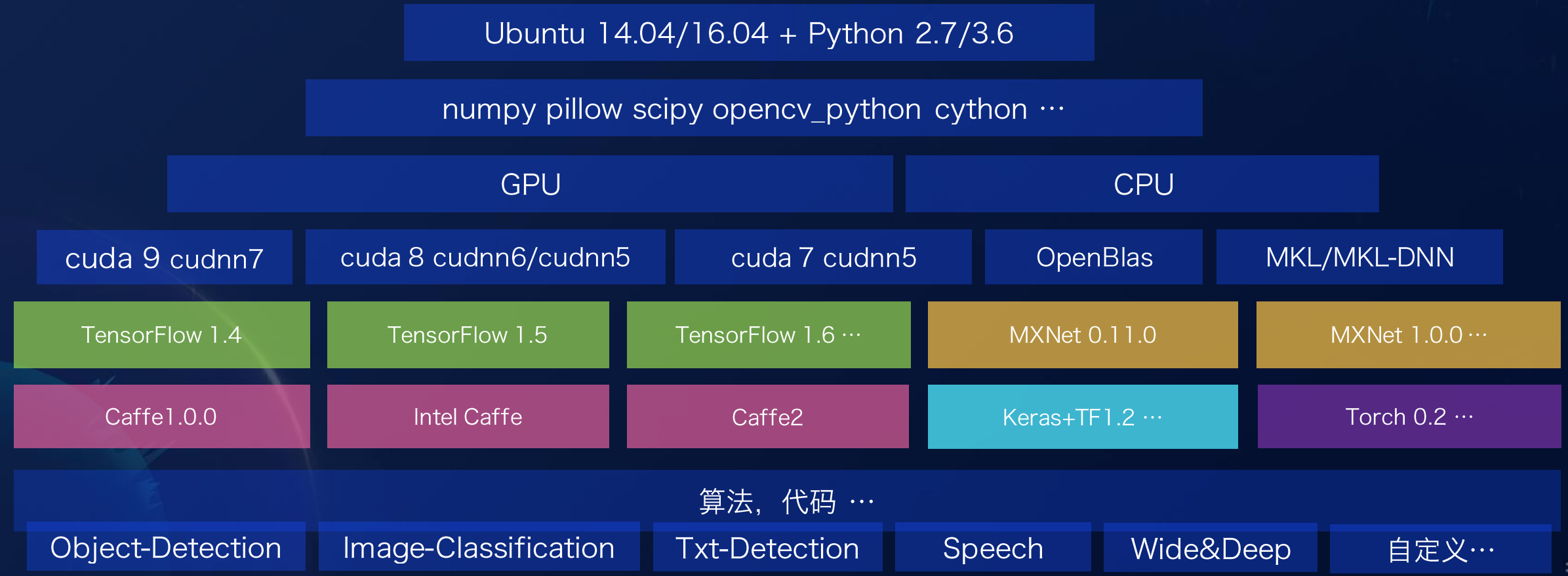
目 录

AI落地技术挑战的解决思路

环境分离



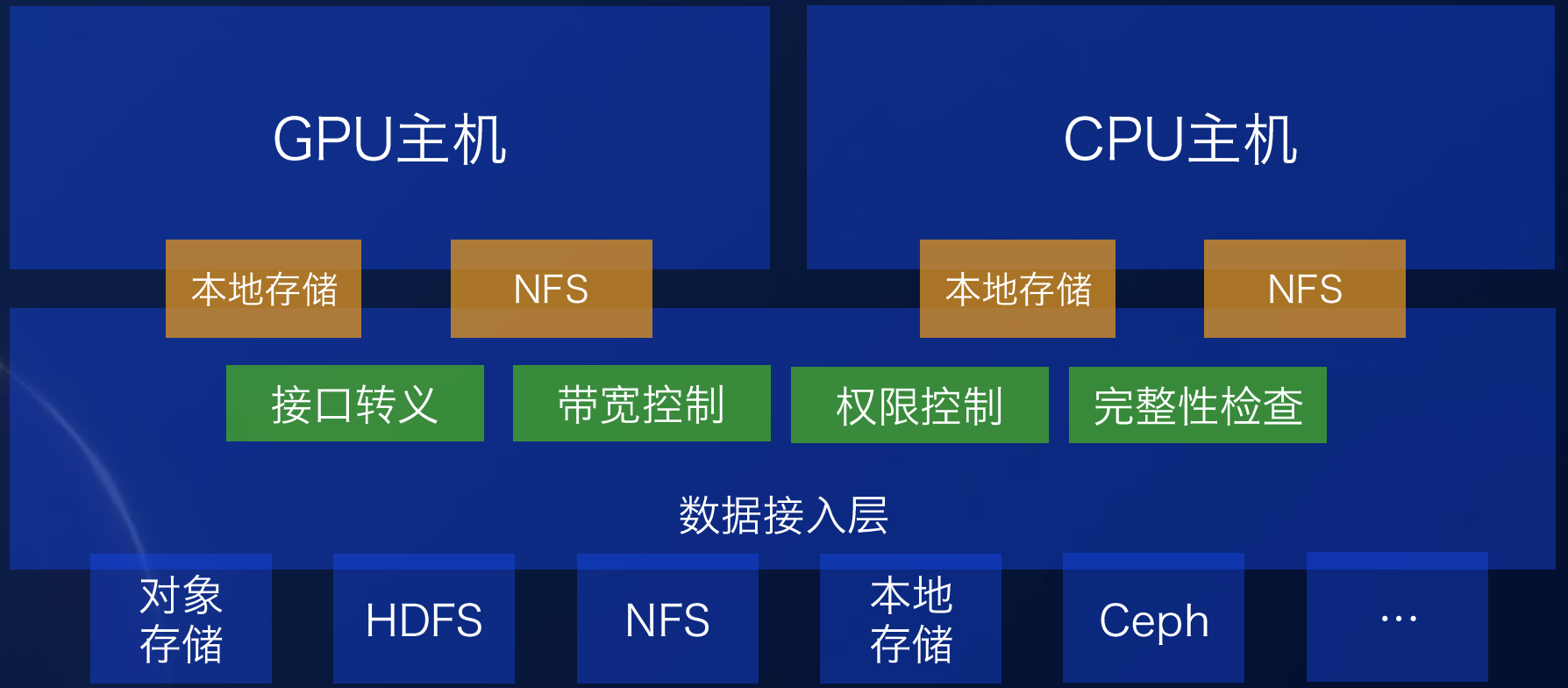
环境分离I: 容器



环境分离I: 容器

封装	运行环境完全隔离，不同任务之间不会产生软件冲突
预装	基础镜像内置各类基础软件环境，减少使用者环境准备开销
自由	可以自由安装各类软件包，封装各类算法
可重用	算法的容器镜像可以重复使用
兼容性	GPU容器镜像可以在任意类型GPU节点运行 CPU容器镜像可以在任意类型CPU节点运行

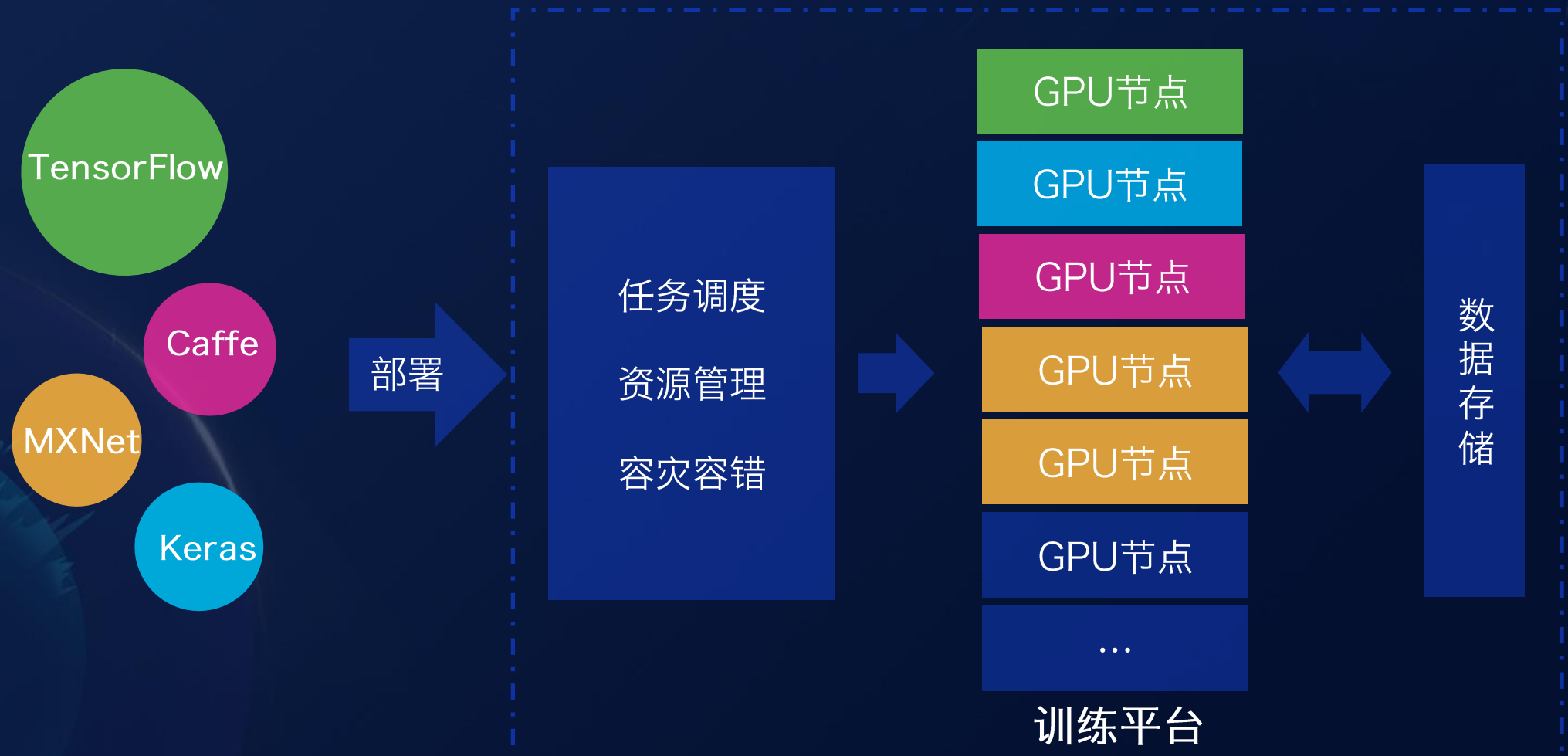
环境分离II: 数据接入


GPU主机
CPU主机
本地存储
NFS
本地存储
NFS
接口转义
带宽控制
权限控制
完整性检查
数据接入层
对象存储
HDFS
NFS
本地存储
Ceph
…

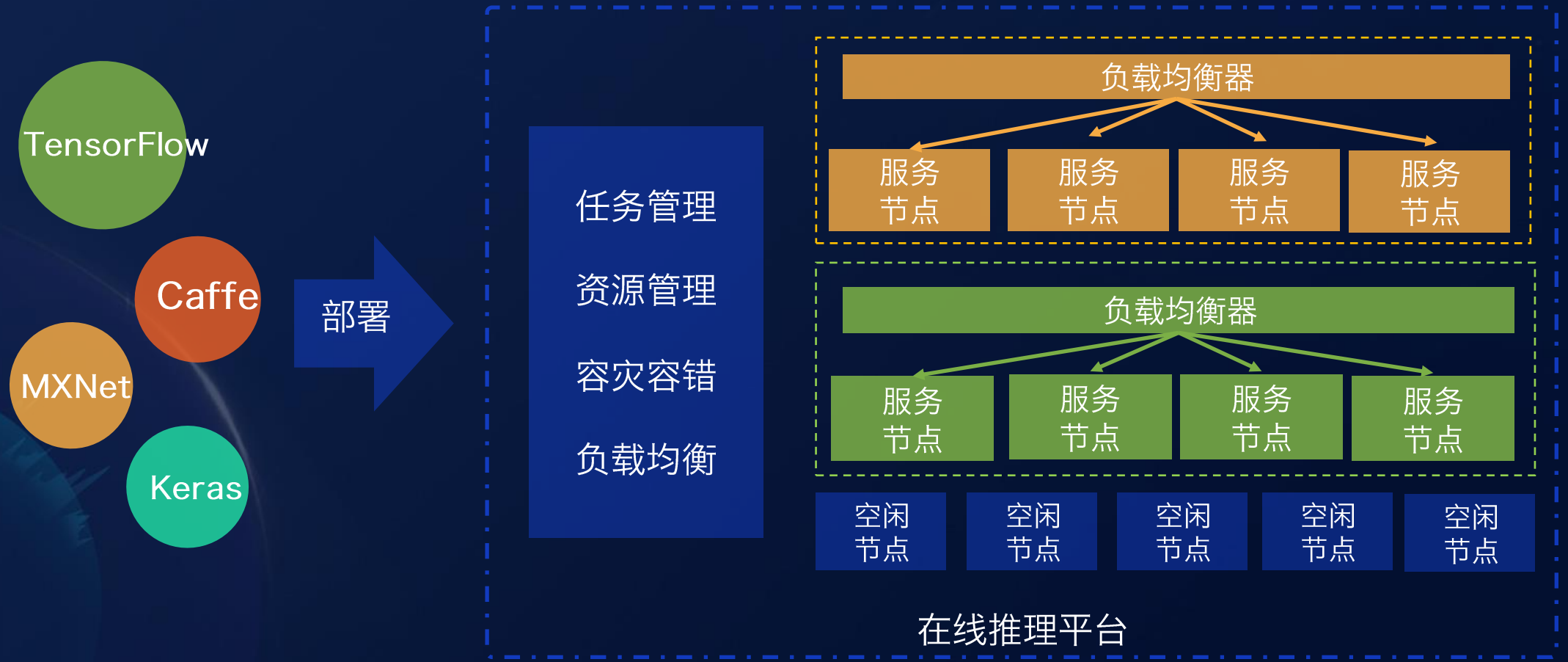
环境分离II: 数据接入

封装	计算节点逻辑不需要支持各种存储接口，仅需要通过2-3种（例如本地存储、NFS）接口就可以对接各类存储类型
灵活	通过扩展数据接入层可接入的存储类型，也就可以扩展AI平台的数据接入类型
稳定	数据接入层可以做数据流量控制，确保各个任务的SLA，同时对后端的数据存储系统进行带宽、流量保护
安全	数据访问权限控制，确保数据安全性

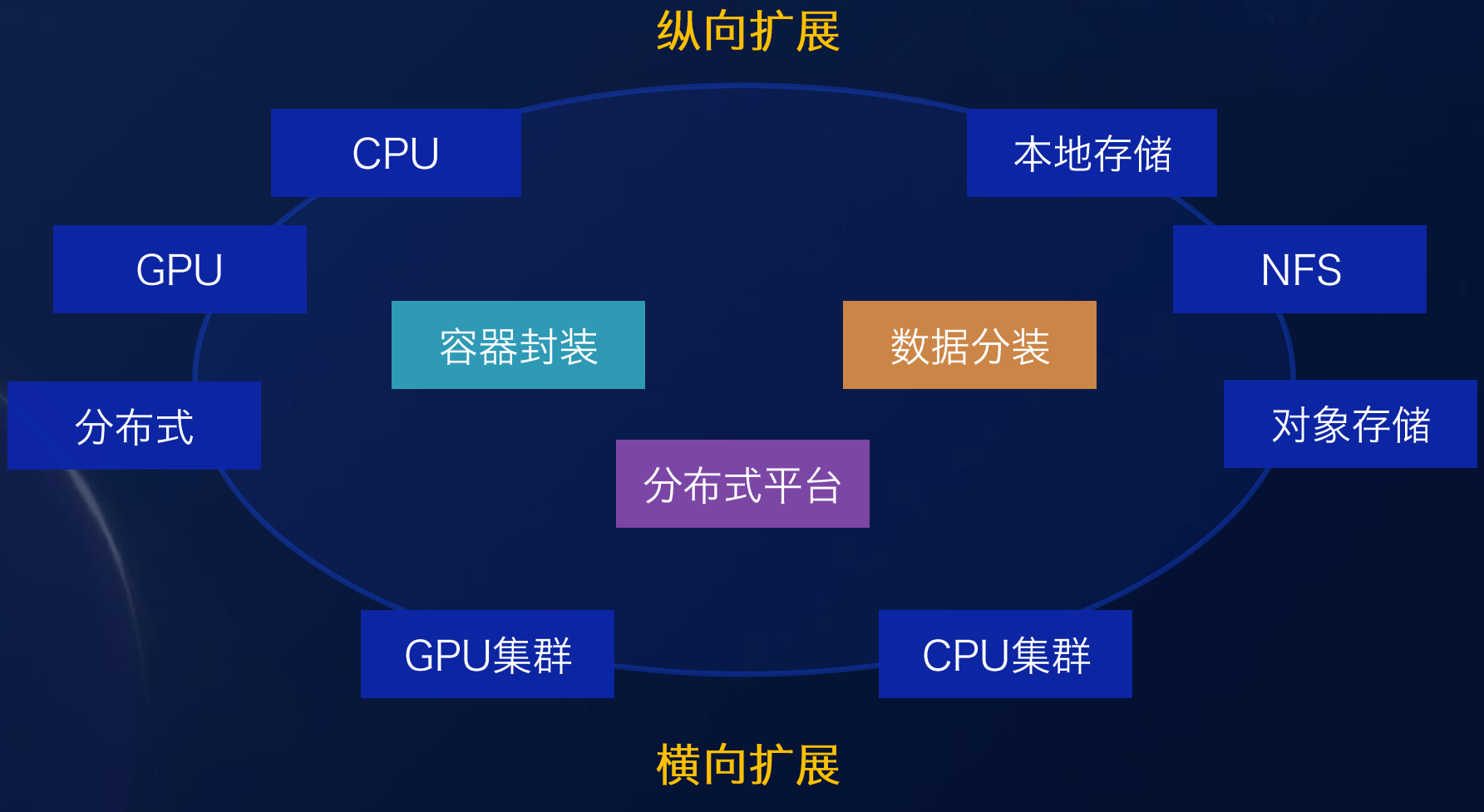
分布式化—训练平台


TensorFlow
Caffe
MXNet
Keras
部署
任务调度
资源管理
容灾容错
GPU节点
GPU节点
GPU节点
GPU节点
GPU节点
GPU节点
…
训练平台
数据存储

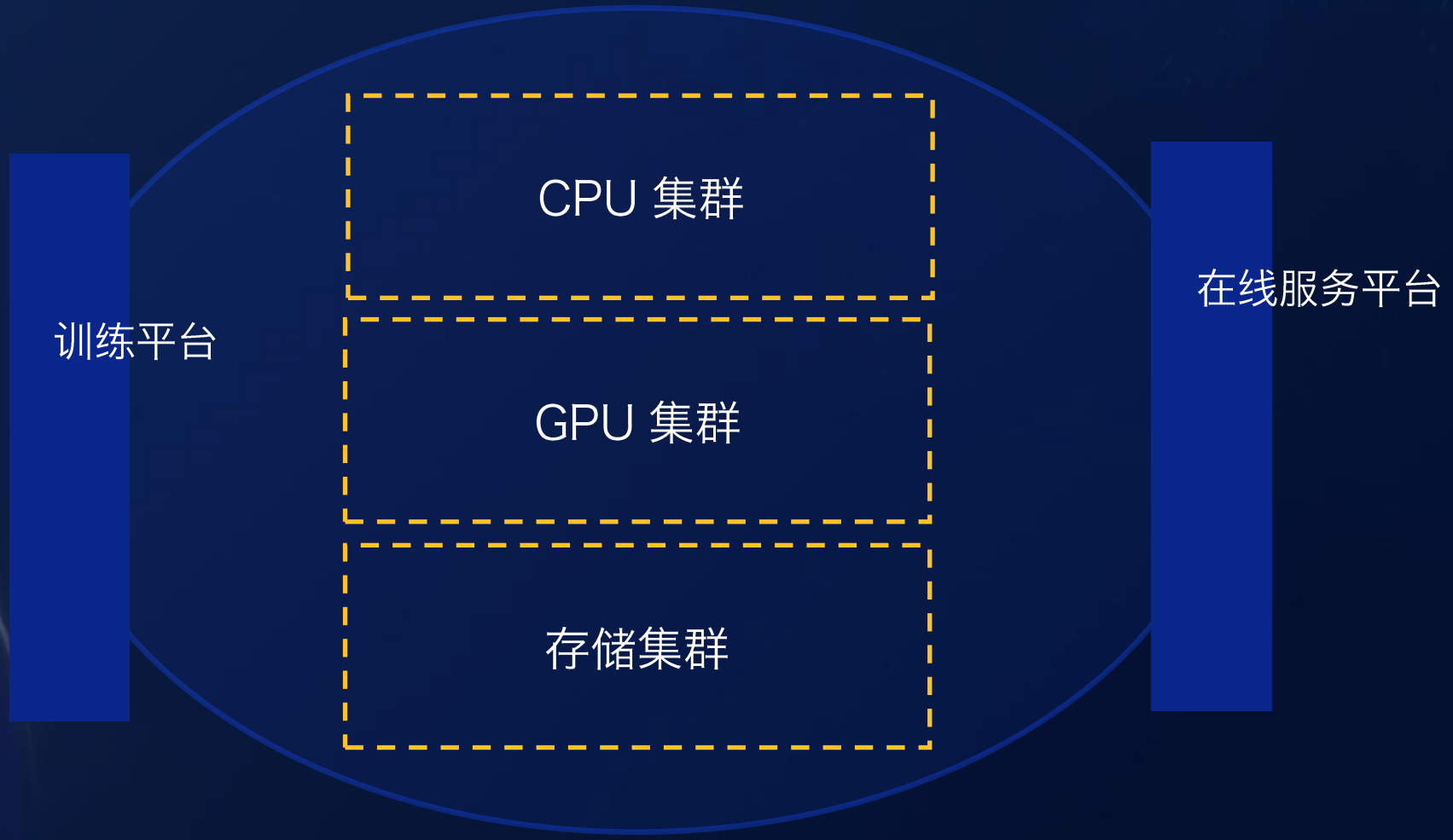
分布式化—在线推理平台



可扩展性


纵向扩展
CPU
本地存储
GPU
NFS
容器封装
数据分装
分布式
对象存储
分布式平台
GPU集群
CPU集群
横向扩展

资源共享

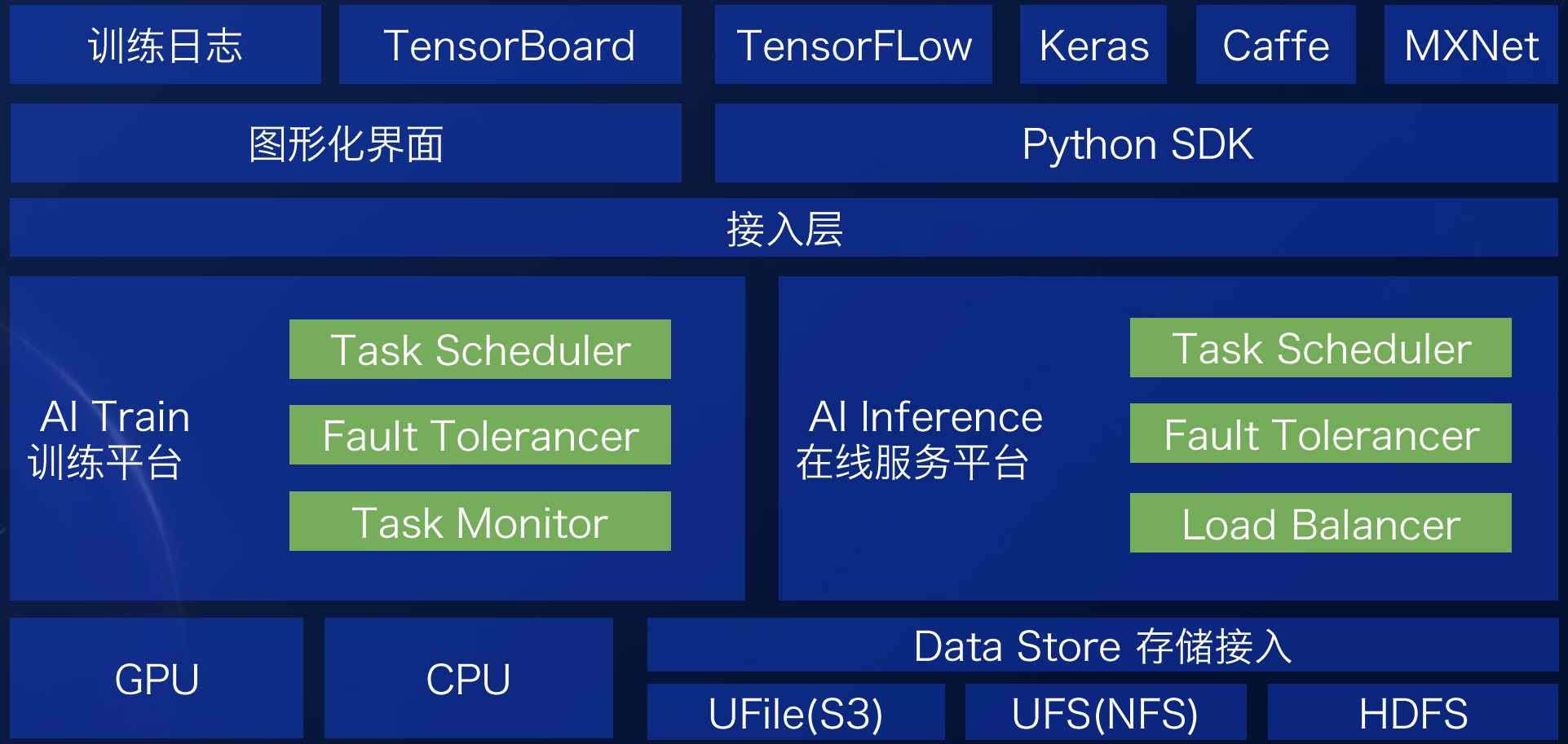


目 录

公有云支持AI落地

资源	基础架构	PaaS服务
计算、存储、网络 多机房、跨地域	云主机、docker容器 负载均衡、共享存储	训练平台 在线服务平台

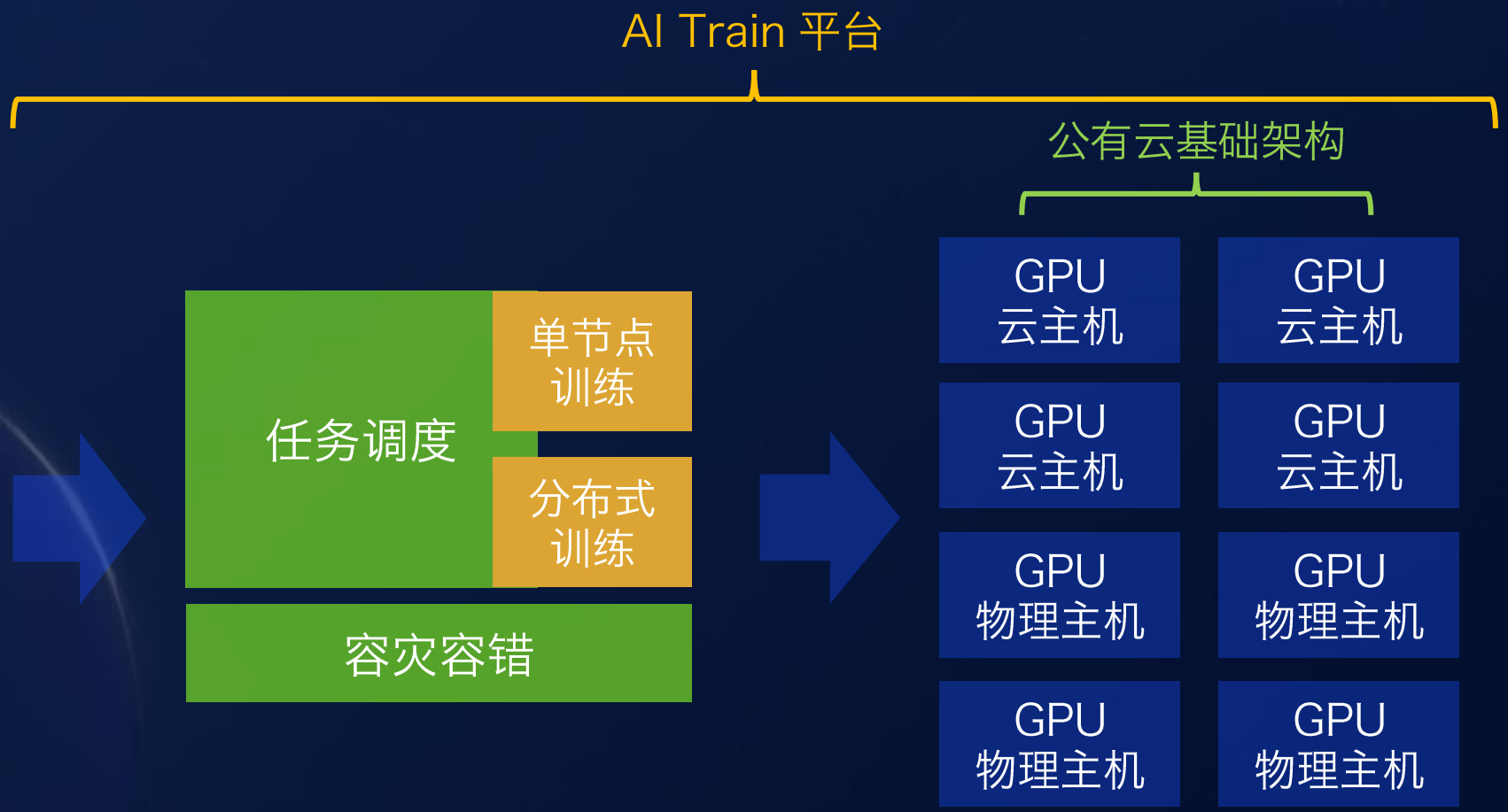
公有云支持AI落地: UAI PaaS平台



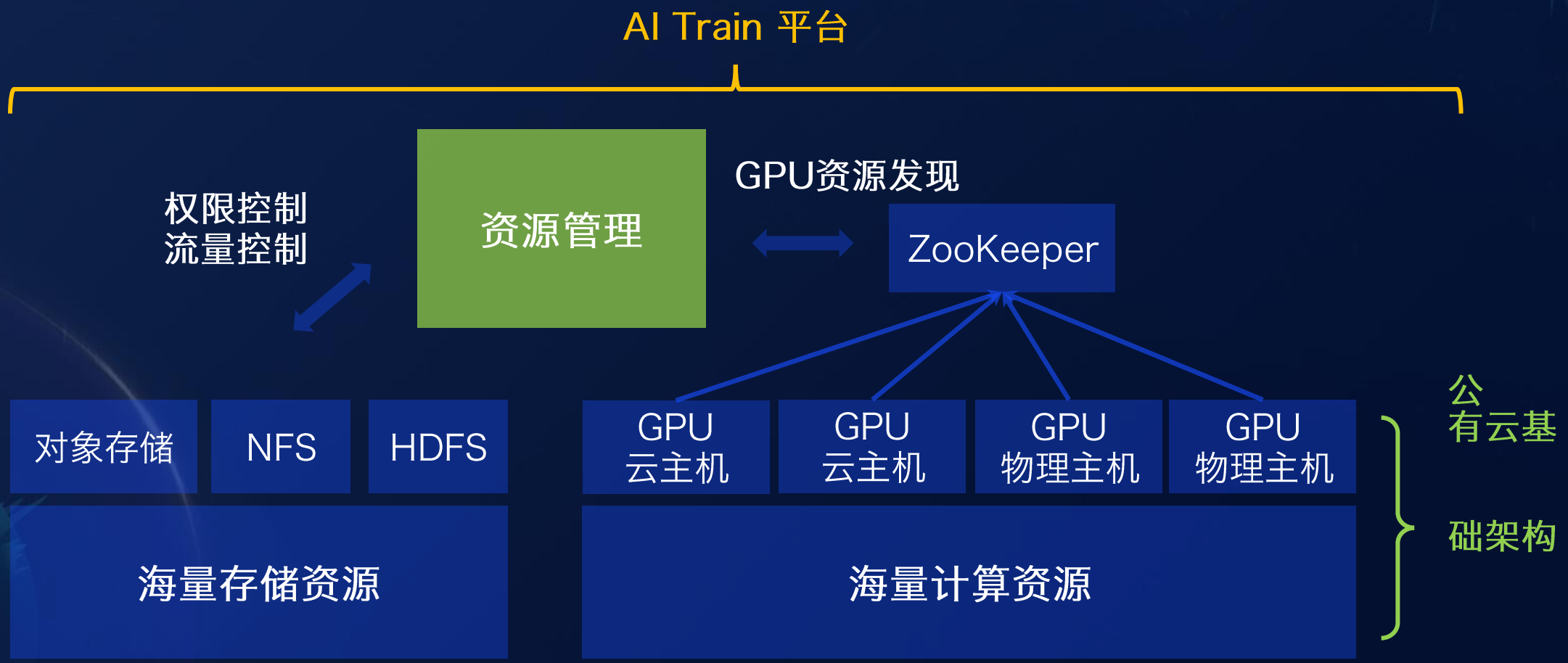
基础环境封装

算法容器	开源算法训练镜像：TF-Slim, East, Detectron, Wide&Deep … 开源算法服务镜像：CTPN, Wide&Deep, Inception …
基础容器	基础AI框架镜像：TensorFlow, MXNet, Caffe, Torch … 基础GPU容器镜像：Cuda9+Cudnn7，Cuda8+Cudnn6 …
系统镜像	基础GPU镜像：NV Driver + NV Docker 基础CPU镜像:：Docker

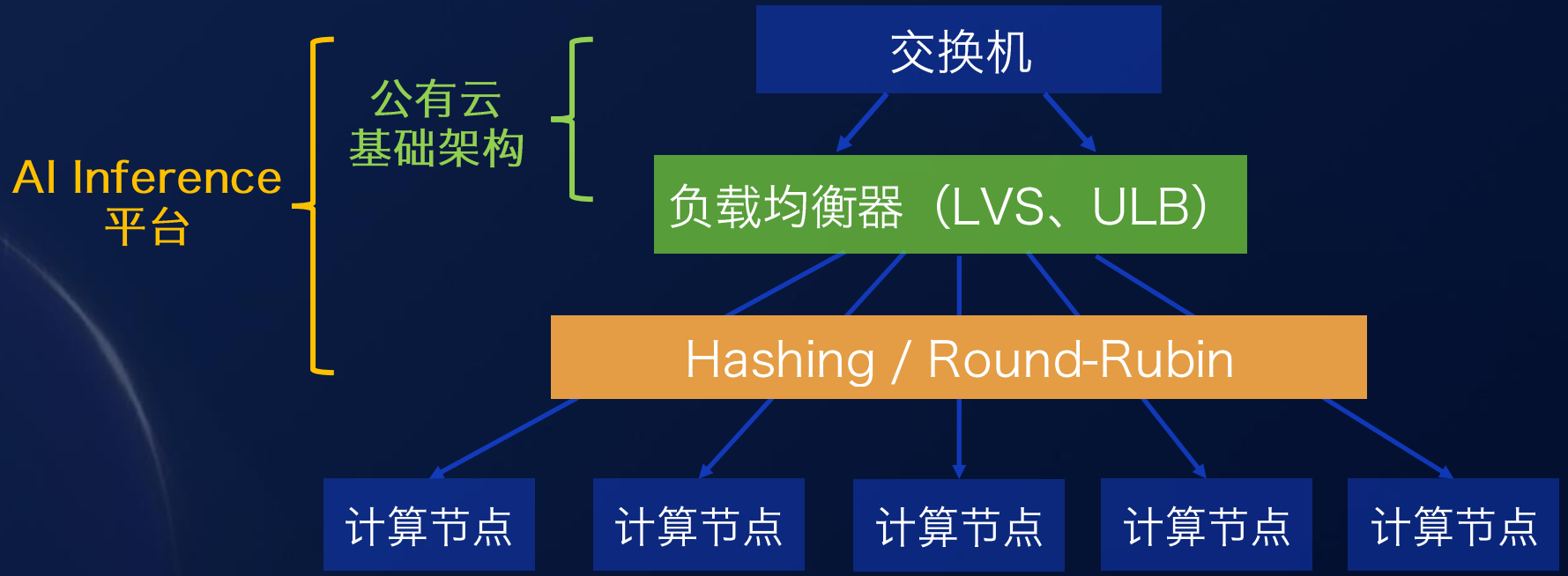
分布式AI训练平台——自动调度



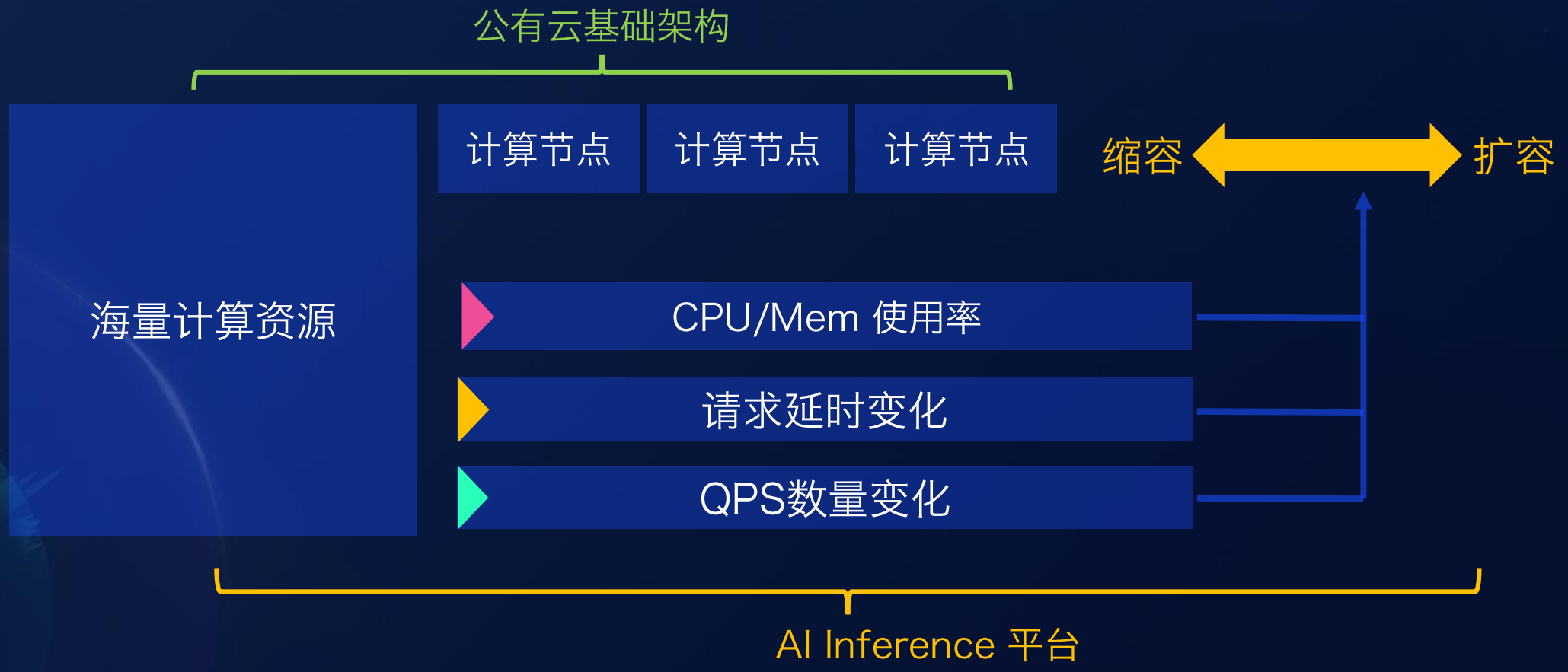
分布式AI训练平台—— 可扩展



分布式AI在线服务平台——负载均衡

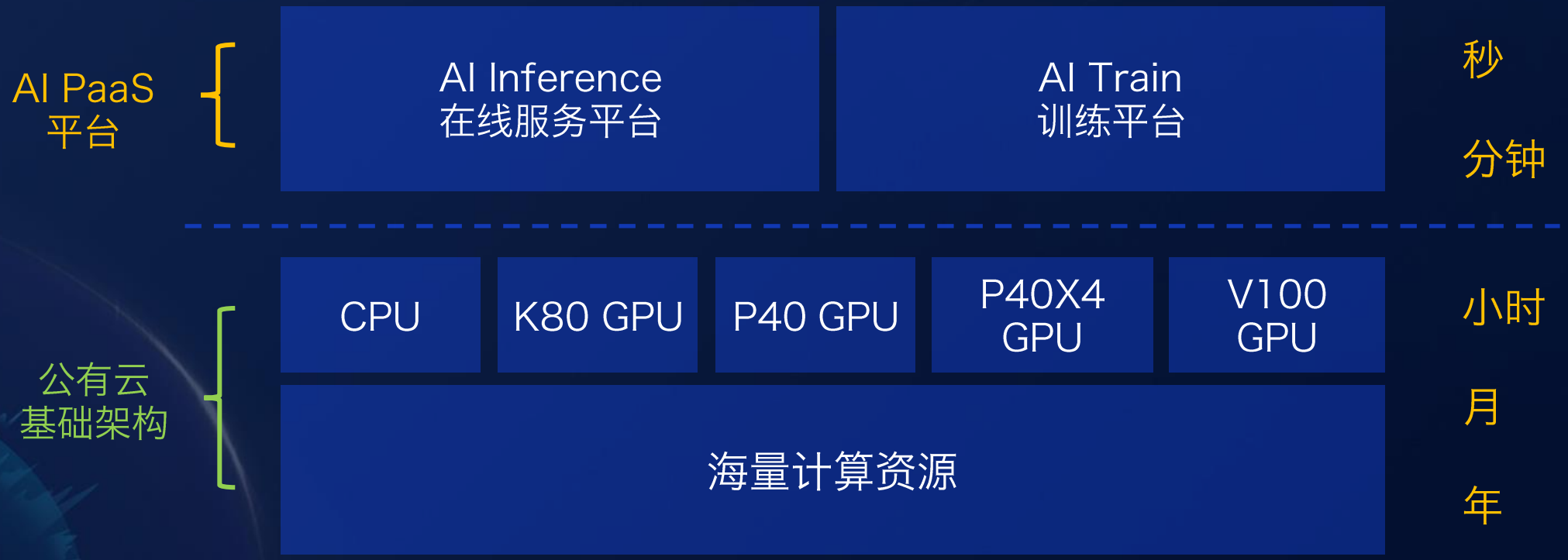


分布式AI在线服务平台——弹性可扩展



分布式AI在线服务平台——高可用

海量资源共享


AI PaaS
平台
AI Inference
在线服务平台
AI Train
训练平台
秒
分钟
公有云
基础架构
CPU
K80 GPU
P40 GPU
P40X4
GPU
V100
GPU
海量计算资源
小时
月
年

公有云支持AI落地之IaaS服务

资源	充足的计算资源、存储资源、网络资源 降低AI研发过程资源采购、维护的成本
基础环境	提供虚拟机镜像、容器镜像等服务 降低AI研发、应用过程中AI环境部署的难度
基础服务	提供诸如负载均衡(ULB) 、分布式存储等基础服务 降低AI应用产品化过程的研发成本

公有云支持AI落地之PaaS服务

环境封装	提供预置AI基础环境，包括NV GPU驱动、Cuda、TensorFlow/MXNet等框架 用户无须进行复杂的环境安装、配置工作
分布式	提供AI训练平台和AI在线服务平台，提供一站式AI服务 用户无须自行搭建复杂的AI平台
横向扩展	提供充足CPU/GPU资源，可自由横向扩展 用户无需担心资源问题
纵向扩展	支持多种计算、存储、网络资源类型 用户可自由选择合适组合
计费灵活	基于秒级/分钟级的计费规则，按需收费 用户无需担心资源浪费

目 录

案例分享I: 图片特征标签

海量用户
图片信息

用户标签信息

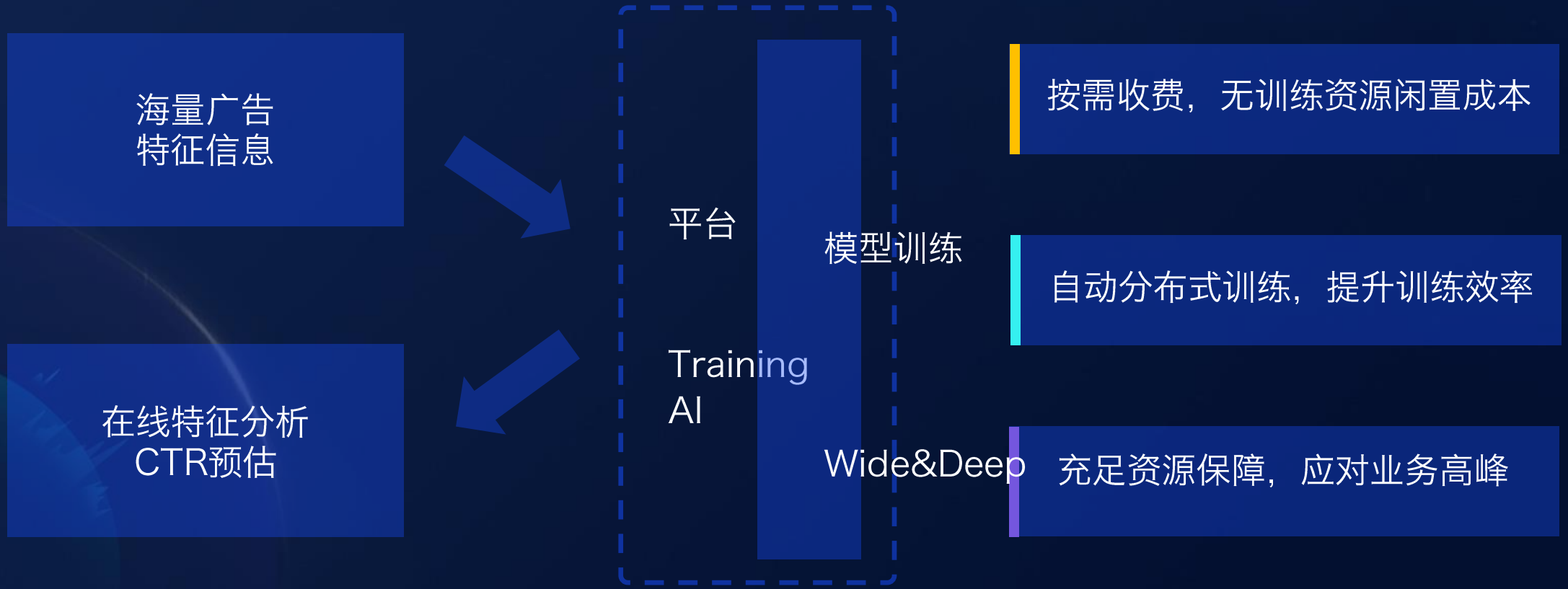
平台

Inference AI

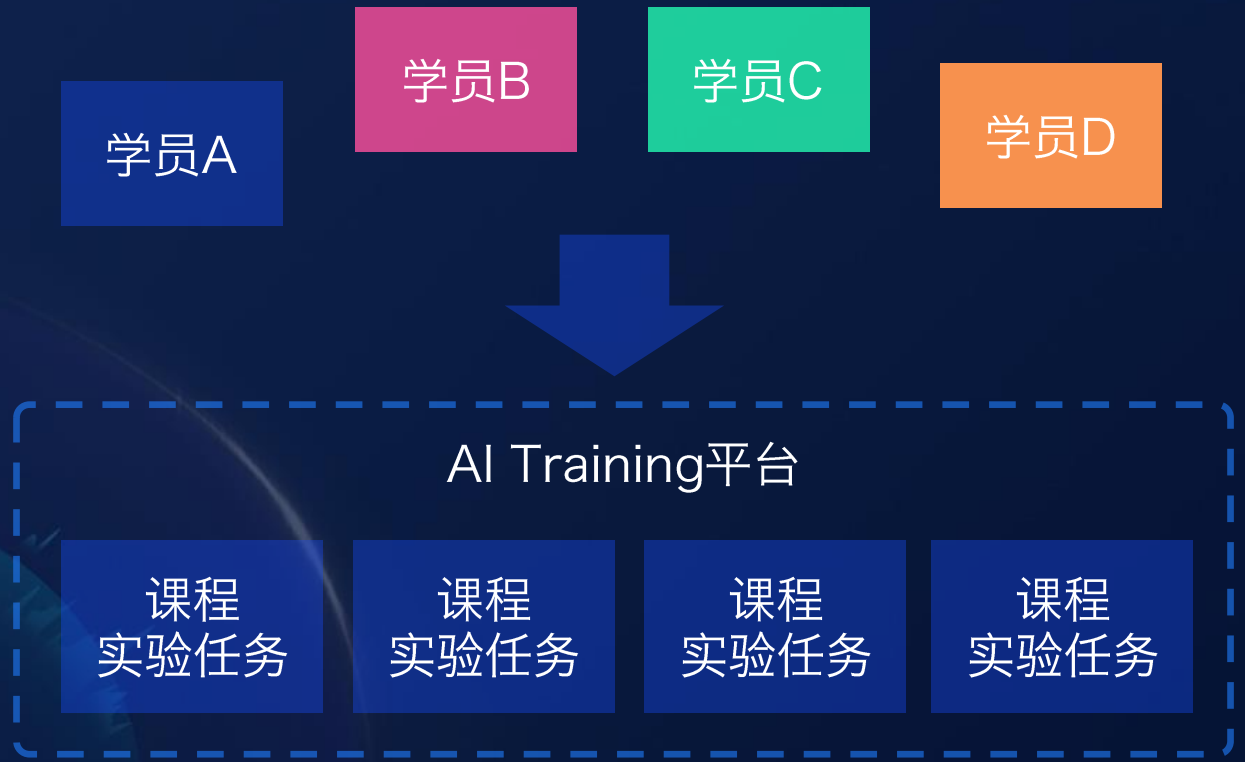
图片识别在线服务

- 快速部署图片识别服务
- 分布式AI在线服务平台研发0投入
- 充足资源保障，应对业务高峰
- 按需收费

案例分享II: 客户特征分析



案例分享III: AI 培训



- 训练资源学员间共享，按需收费
- 基础AI镜像封装，降低实验任务准备难度
- 充足资源保障，应对高峰

联系我们

UAI Inference
UAI Train

Github:

https://github.com/ucloud/uai-sdk/

UAI Train:

https://www.ucloud.cn/site/product/uaitrain.html

UAI Service:

https://www.ucloud.cn/site/product/uaiservice.html

Contact: 4000188113

加入我们团队

Contact:

charlie.song@ucloud.cn

john.hu@ucloud.cn

Think in Cloud UCLOUD

THANKS