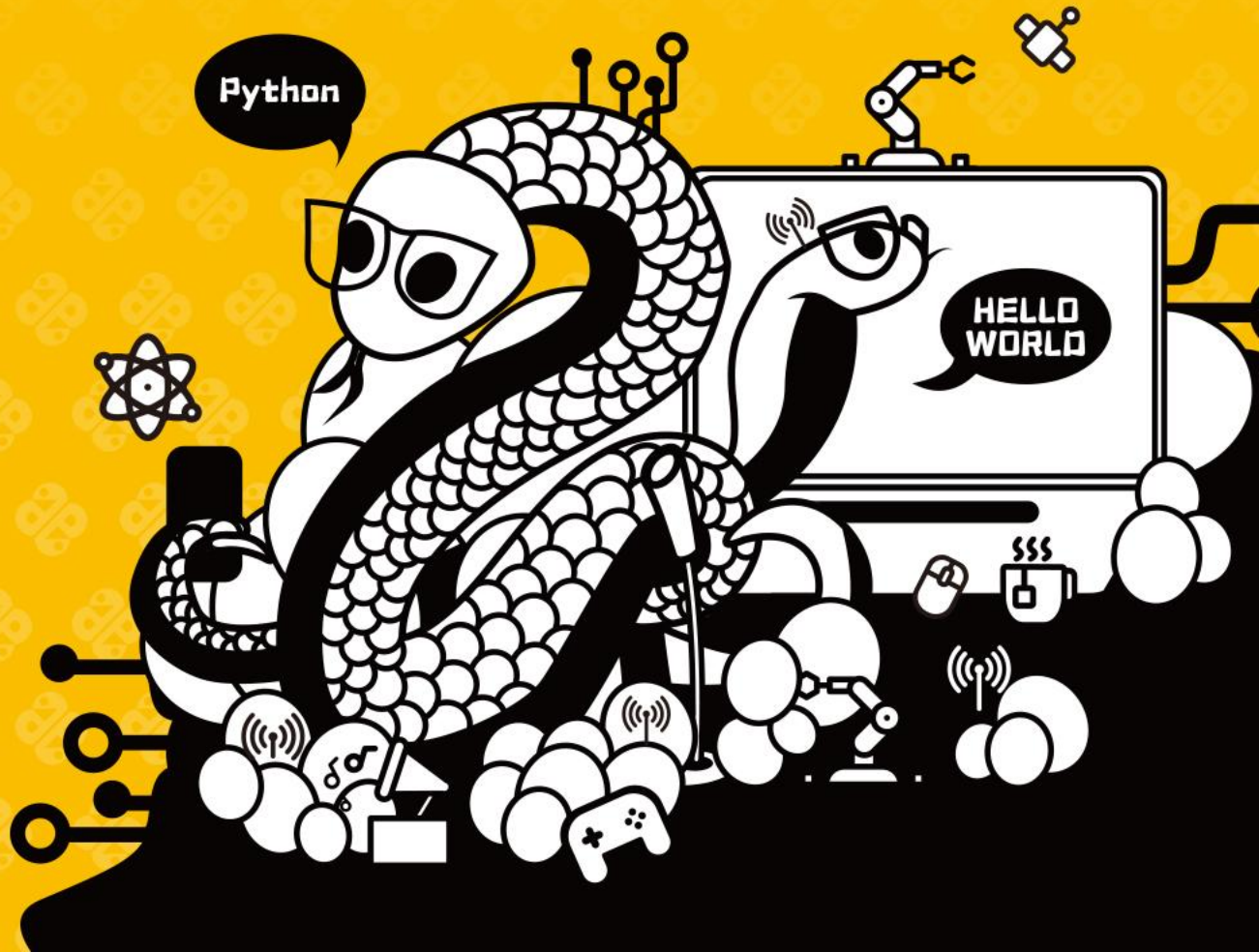
python × PYCHINA Python 中国社区

Python for Good

PyCon China 2022

云原生场景下的AI落地

主讲人： 刘宇宙 – 工程师



个人简介

刘宇宙，资深开发工程师，目前就职于某硬件智能，从事硬件智能相关研发。先后供职于上海海鼎、广州棒谷等科技公司。

先后从事过云计算IaaS、大数据、物联网和人工智能等的研发，对Python有深入研究。

目前已经出版《Python3.5 从零开始学》、《Python3.7 从零开始学》、《Python3.8 从零开始学》、《Python 实用教程》、《Python 实战之数据库应用和数据获取》、《Python 实战之数据分析与处理》、《好好学Python 从零基础到项目实战》、《Python进阶编程：编写更高效、优雅的Python代码》、《左手Python，右手Excel，带飞Excel的Python绝技》等Python书籍。

大纲

一、云原生概述

二、AI之痛

三、云原生下的AI

一、云原生概述

1、云原生的关键技术包括

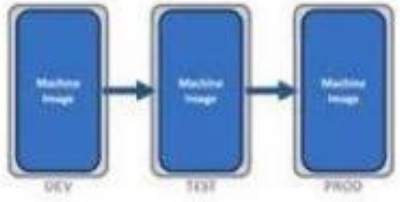
微服务架构

容器

容器编排

服务网格

申明式API



不可变基础设施

DevOps

云原生-关键技术

微服务架构：服务与服务之间通过高内聚低耦合的方式交互；

容器：作为微服务的最佳载体，提供了一个自包含的打包方式；

容器编排：解决了微服务在生产环境的部署问题（k8s）；

服务网络：作为基础设施，解决了服务之间的通信（istio）；

不可变基础：设施提升发布效率，方便快速扩展；

声明式 API：让系统更加健壮；

命令式 API：可以直接发出让服务器执行的命令，例如：“运行容器”、”停止容器”等；

声明式 API：可以声明期望的状态，系统将不断地调整实际状态，直到与期望状态保持一致。

DevOps：缩短研发周期，增加部署频率，更安全地方便：

Culture ：达成共识

Automation：基础设施自动化

Measurement：可度量

Sharing：你中有我，我中有你

2、云原生主流组件


Prometheus
Grafana
EFK
(Elasticsearch+Fluentd+Kibana)
Jaeger
Chaos Engineering

Prometheus

- 开源监控解决方案
- CNCF第02个毕业项目

Grafana

Grafana

开源的度量分析与可视化套件

elasticsearch fluentd kibana

EFK
(Elasticsearch + Fluentd + Kibana)

- 日志的查询、采集和展示

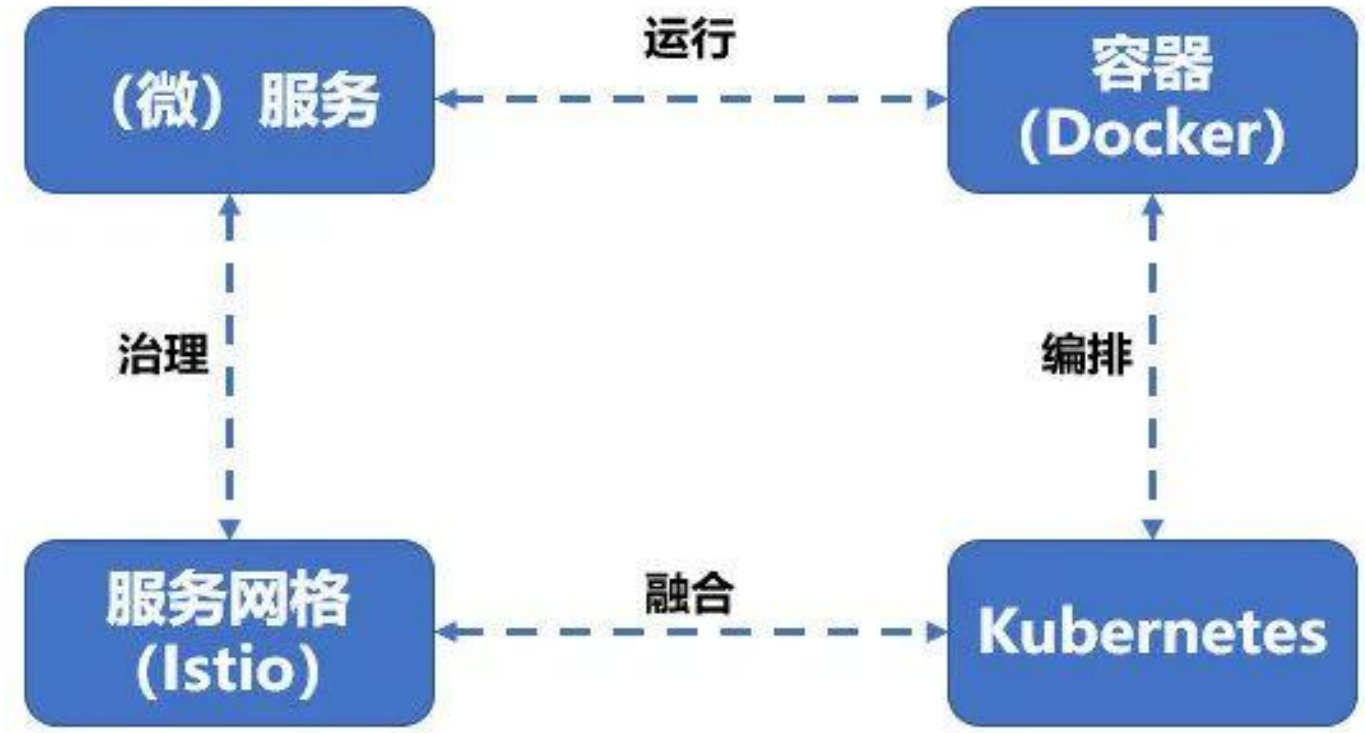
Jaeger

- 分布式调用链跟踪平台

Chaos Engineering

- 主动注入异常，提前找到弱点，提升系统自愈和鲁棒性。

3、云原生——总结

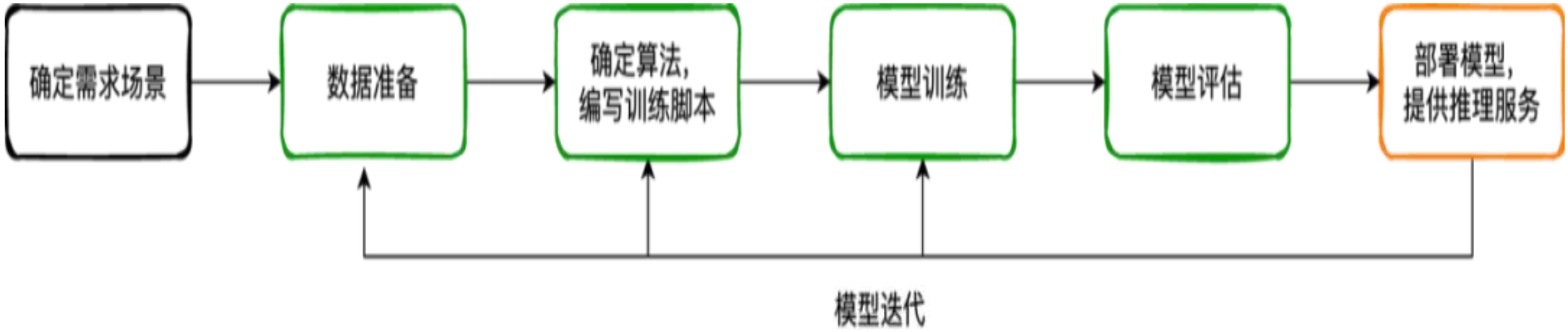


微服务 & Docker & Kubernetes & Istio 的关系

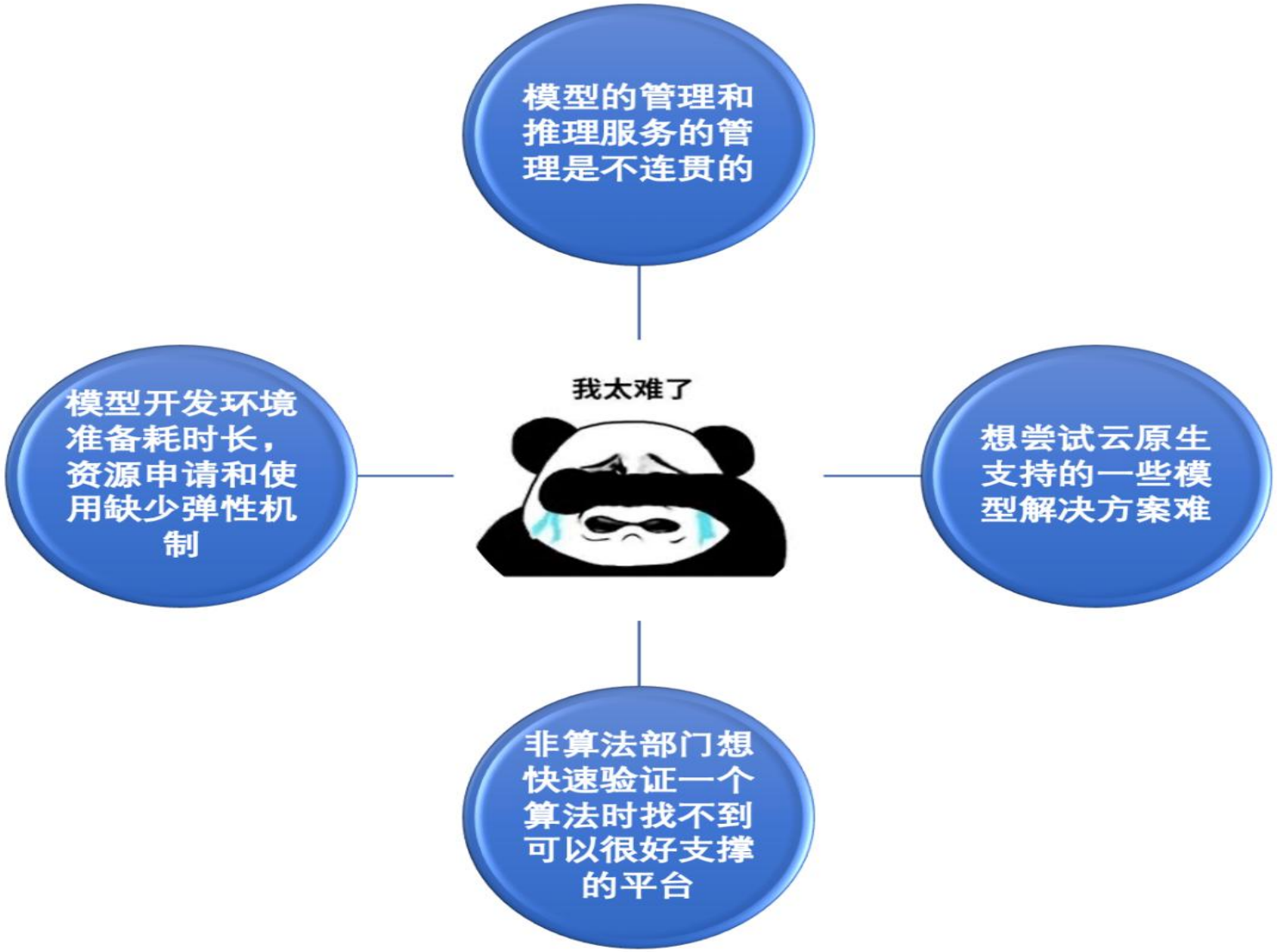
云原生应用：docker 应用打包、发布、运行，Kubernetes 服务部署和集群管理，Istio 构建服务治理能力。

二、AI之痛

AI算法模型开发流程繁杂



算法开发与部署的痛点

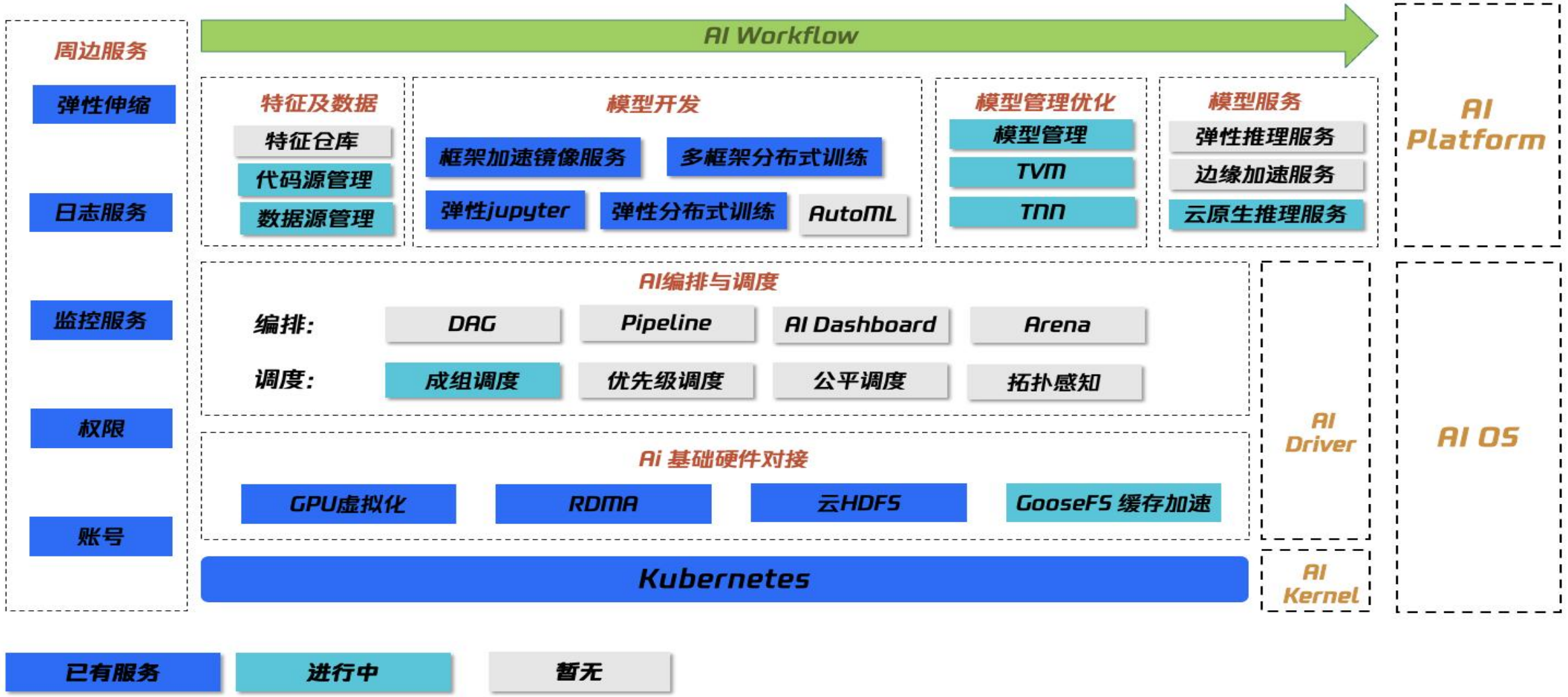

模型的管理和推理服务的管理是不连贯的
模型开发环境准备耗时长，资源申请和使用缺少弹性机制
我太难了
想尝试云原生支持的一些模型解决方案难
非算法部门想快速验证一个算法时找不到可以很好支撑的平台

三、云原生下的AI

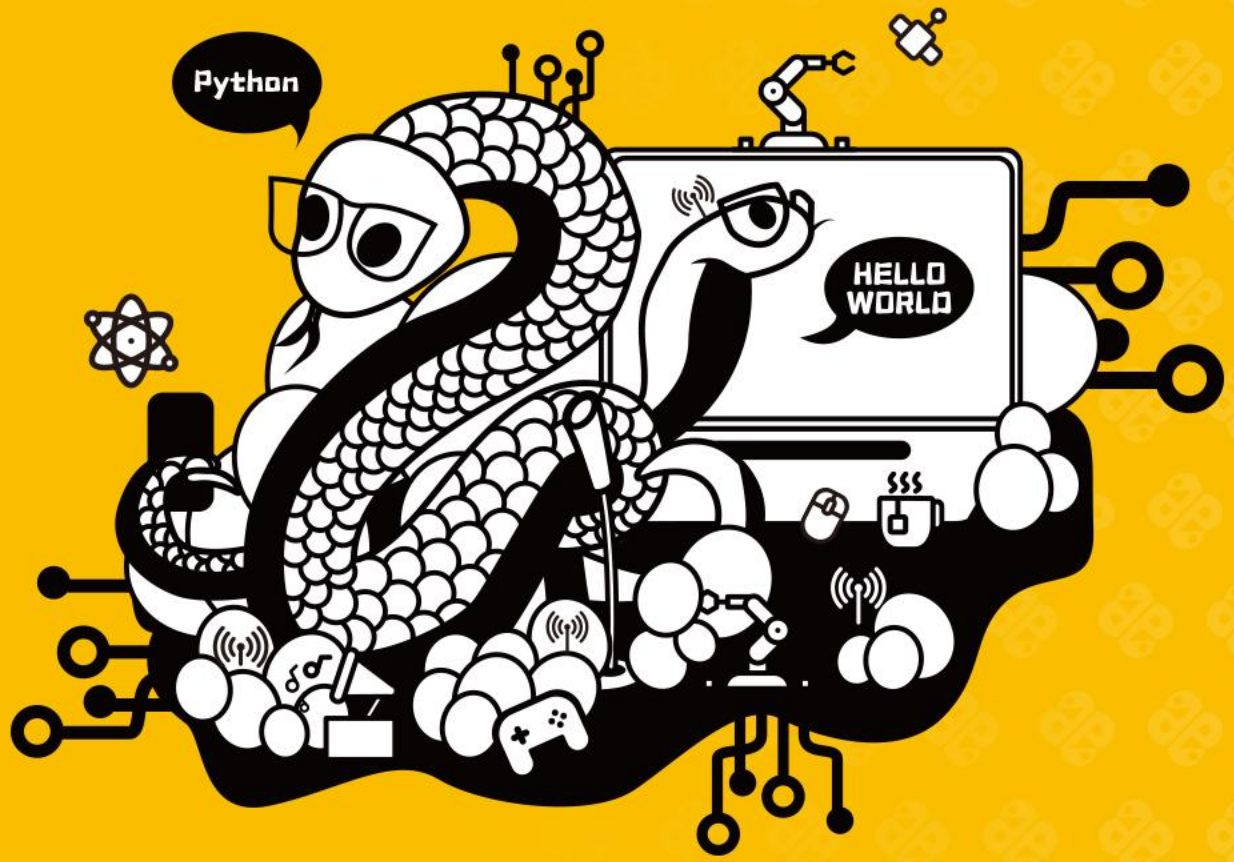
云原生AI在资源弹性、跨节点架构感知，训练推理效率等多方面的能力显著提升，可最大化地帮助企业实现AI应用的快速交付与落地。

AI 加速库	分布式训练加速	自研通信库ECCL	推理加速					
AI 任务管理	AI 任务管理	工作流任务	AI 任务细粒度监控	AI 资源画像				
AI 资源调度	精细化调度	共享调度	优先级调度	资源超发	亲和调度	弹性训练	节点内架构感知	节点间架构感知
AI 资源管理	GPU虚拟化用户态(算力和显存的隔离)	GPU虚拟化用户态(显存超发，编解码)	GPU虚拟化内核态(算力、显存隔离、混布)	NPU(算力、显存的共享)	PFS CSI	RapidFS CSI	RDMA网络	GPU细粒度监控

云原生 AI 旨在利用云原生的思想和技术，为 AI 场景的数据处理、模型训练、模型上线推理等需求构建弹性可扩展的系统架构的技术，在支持更广泛、多样的用户需求的同时，提高开发、运维和设备的效率。


周边服务
弹性伸缩
日志服务
监控服务
权限
账号
AI Workflow
特征及数据
特征仓库
代码源管理
数据源管理
模型开发
框架加速镜像服务
多框架分布式训练
弹性jupyter
弹性分布式训练
AutoML
模型管理优化
模型管理
TVM
TNN
模型服务
弹性推理服务
边缘加速服务
云原生推理服务
AI Platform
AI编排与调度
编排:
DAG
Pipeline
AI Dashboard
Arena
调度:
成组调度
优先级调度
公平调度
拓扑感知
Ai 基础硬件对接
GPU虚拟化
RDMA
云HDFS
GooseFS 缓存加速
AI Driver
Kubernetes
AI Kernel
AI OS
已有服务
进行中
暂无

云原生下的AI落地优势



Thanks!

感谢观看