China · Shenzhen

微博机器学习平台混合云应用实践

韩冰

新浪微博 资深产品运维工程师

China · Shenzhen

SPEAKER
INTRODUCE

韩冰　资深产品运维工程师

- 微梦创科网络科技有限公司平台保障服务部担任资深产品运维工程师，10年的系统运维服务经验，长期关注 OpenStack, Kubernetes 等技术的服务应用。

- 目前在微博从事混合云和算法平台的技术保障工作。

TABLE OF
CONTENTS 大纲

微博平台的特点

- 春晚，红包飞等运营活动
- 频发的热点事件带来三到四倍乃至更高的访问量
- 个性化、兴趣化的信息流
- 短视频、直播等多媒体内容

微博 + 机器学习

业界成果

采用智能算法的信息流逐步取代人工规则信息流

视觉听觉语义算法部分能力超过人类

人工智能激活机器人产业，全面改造社会

微博

社交媒体以人为本，有效连接人-内容-人

- 内容-人：feed流、热门流、push等
- 人-内容：反黄、垃圾过滤等
- 人-人：用户推荐、亲密度等

微博+机器学习

机器学习可以对微博的主流业务带来巨大价值，同时在智能潮流大时代逐渐成为决定社交媒体业务的生死因素之一

微博机器学习平台的系统体量

- 每日上亿用户，百亿分钟阅读时间
- 万亿级样本、近百亿级特征的超大规模机器学习平台

机器学习平台的效果衡量

- 用户点击量
- 用户互动行为（转发、评论、赞）
- 用户使用时长

TABLE OF
CONTENTS 大纲

机器学习平台的挑战

- 机器学习平台要成为支撑万亿级样本、近百亿级特征的超大规模平台，平台对存储、计算有非常大的需求和挑战
- 对于成熟的互联网企业，机房机架位资源更加趋于紧张，机器学习平台需要超大规模服务器资源，有较大的资源压力和成本压力

TABLE OF
CONTENTS 大纲

- 微博机器学习平台的业务特点、规模
- 微博机器学习平台建设的挑战
- DCP 混合云调度平台在机器学习中的应用实践

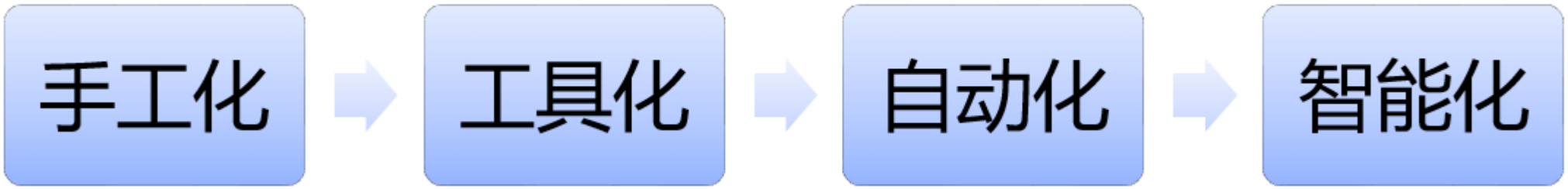
混合云DCP

- 基于Docker的云资源管理与调度平台
- 支撑镜像仓库、多云支持、服务编排、服务发现
- 支持服务池的快速扩缩容，有长期的热点峰值流量的弹性调度实践经验

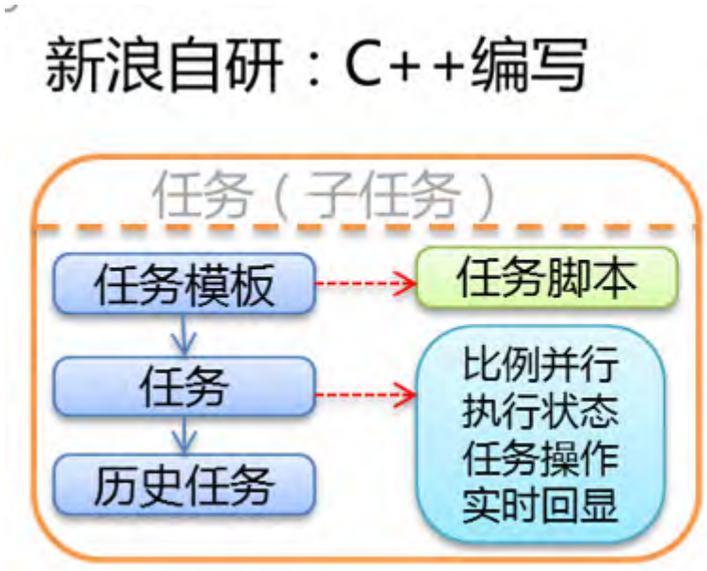
混合云DCP实践

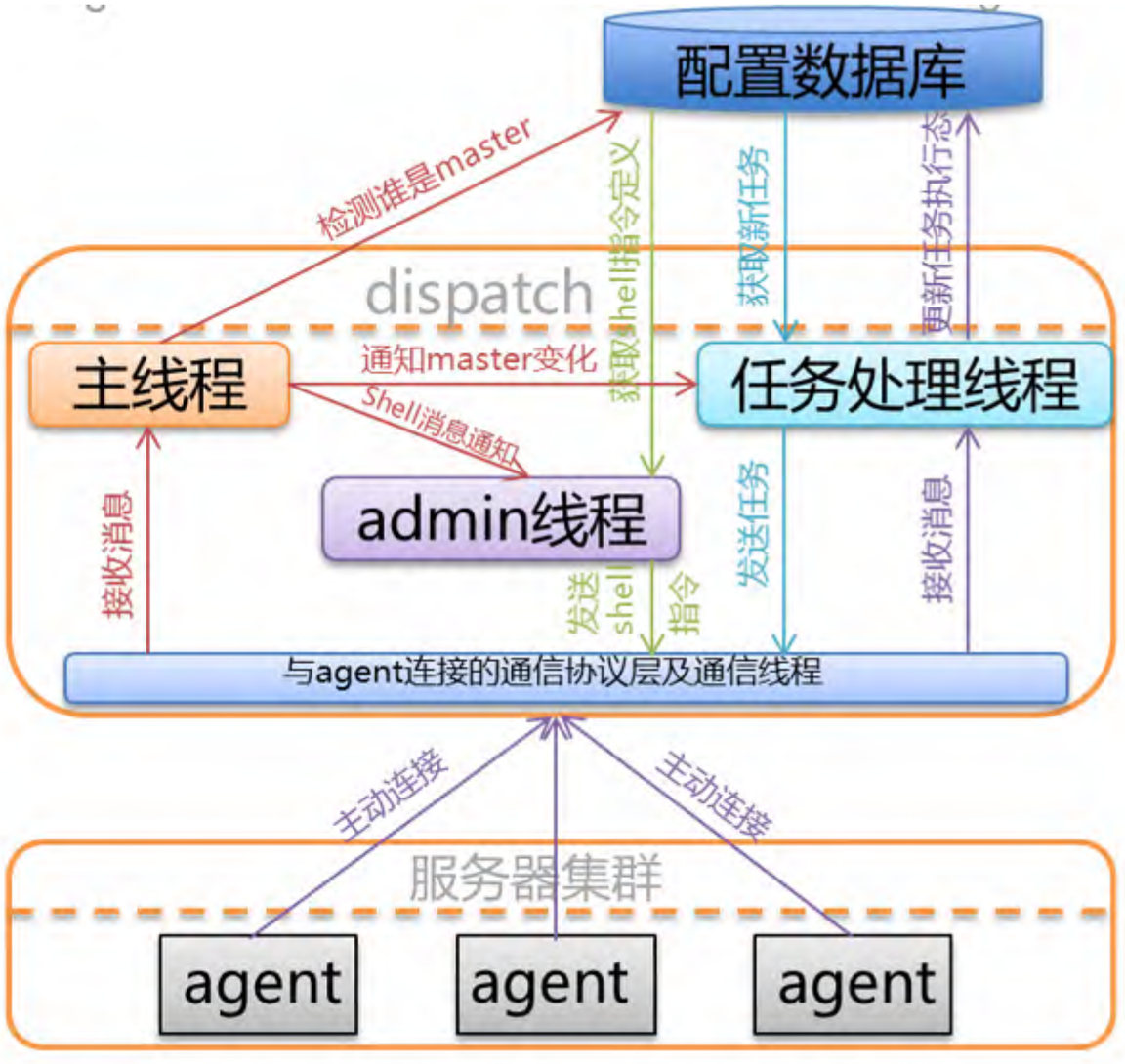
- 已完成基于Docker的全自动化运维平台的建设，并完成了微博主要业务的改造覆盖
- 2017年春晚保障中，在不到一天的时间内完成了近5000台服务器的创建和部署上线，帮助微博实现了在新的流量峰值的情况下整体服务无降级的成绩

混合云DCP平台的演进



Dispatch任务执行引擎

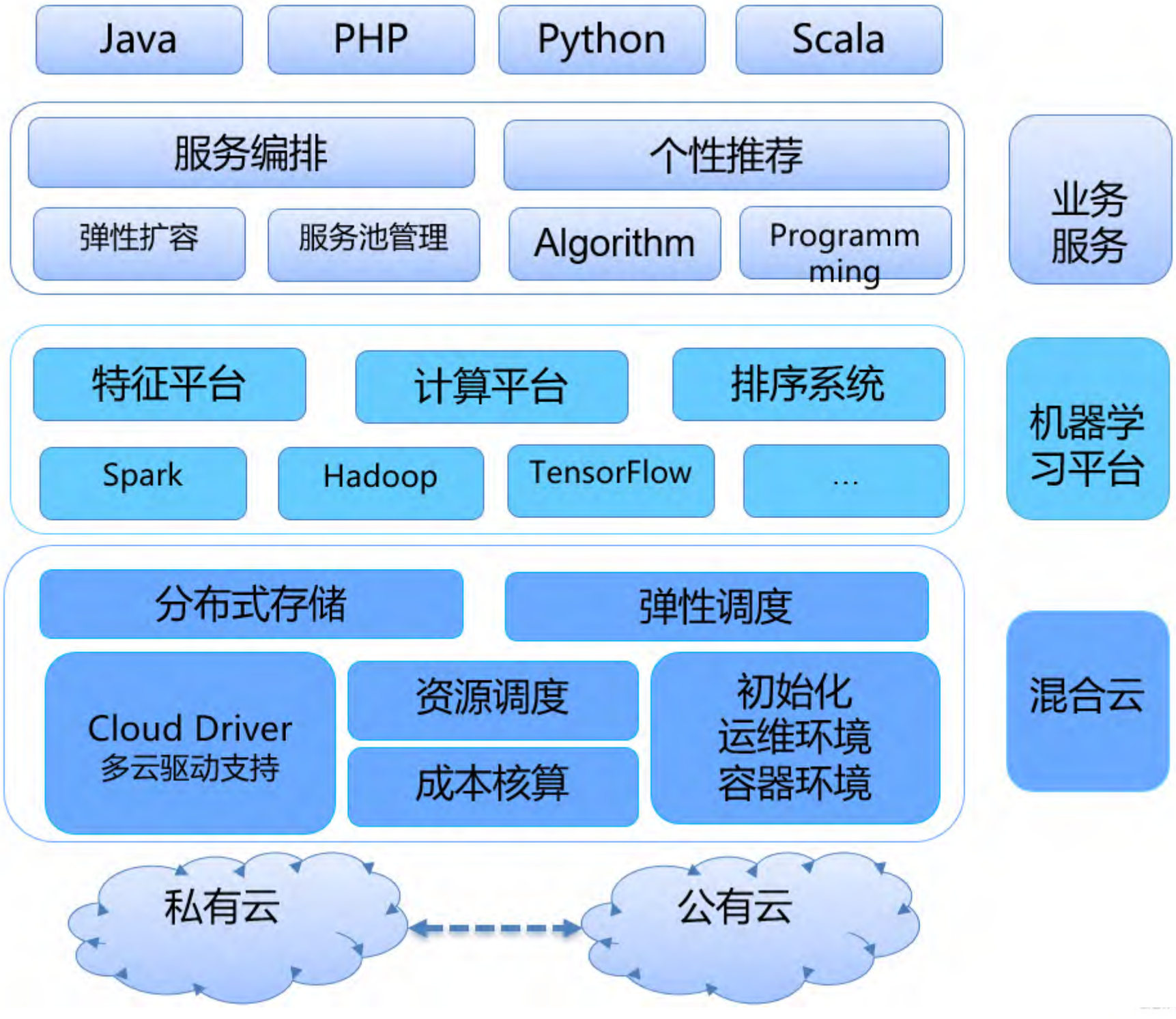




混合云DCP的本轮演进

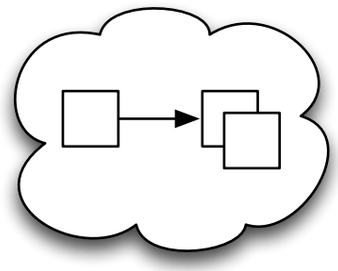
- 支持内网虚拟化引擎
- 支持大规模学习集群

整体架构

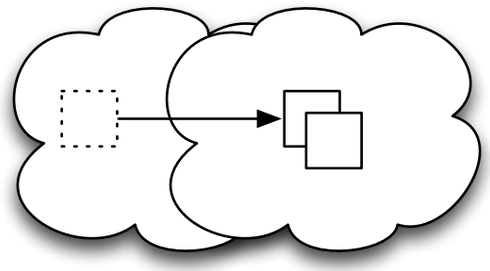

Java
PHP
Python
Scala
服务编排
个性推荐
弹性扩容
服务池管理
Algorithm
Programm
ming
业务
服务
特征平台
计算平台
排序系统
Spark
Hadoop
TensorFlow
...
机器学
习平台
分布式存储
弹性调度
Cloud Driver
多云驱动支持
资源调度
成本核算
初始化
运维环境
容器环境
混合云
私有云
公有云

混合云DCP公有云调度

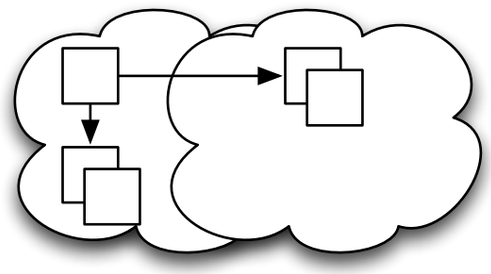
私有内弹性扩容



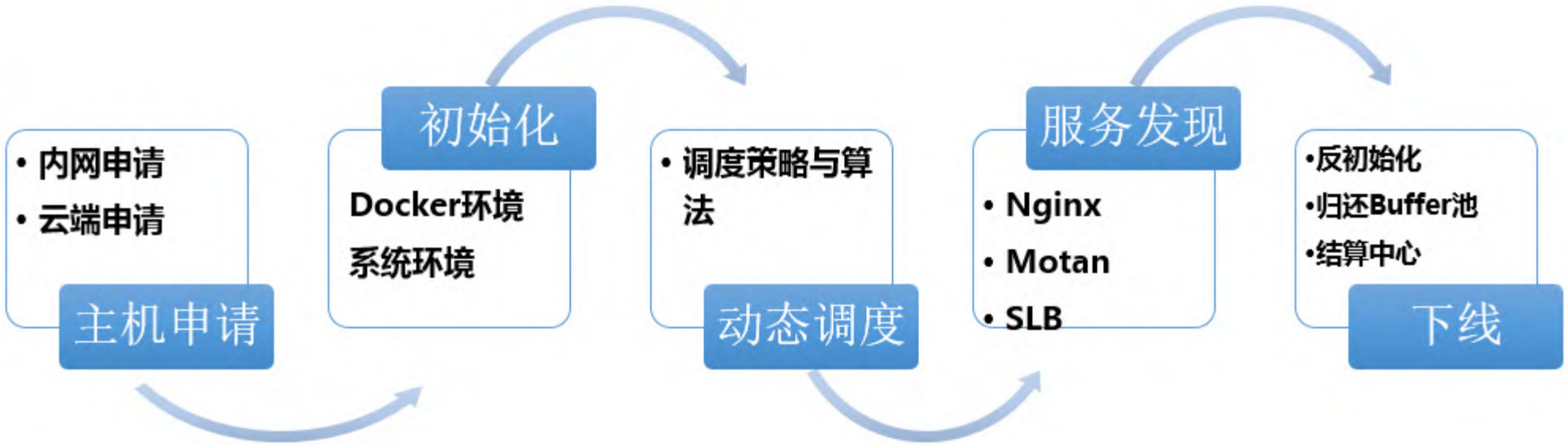
扩容到公有云弹性



私有云、公有云同时弹性扩容



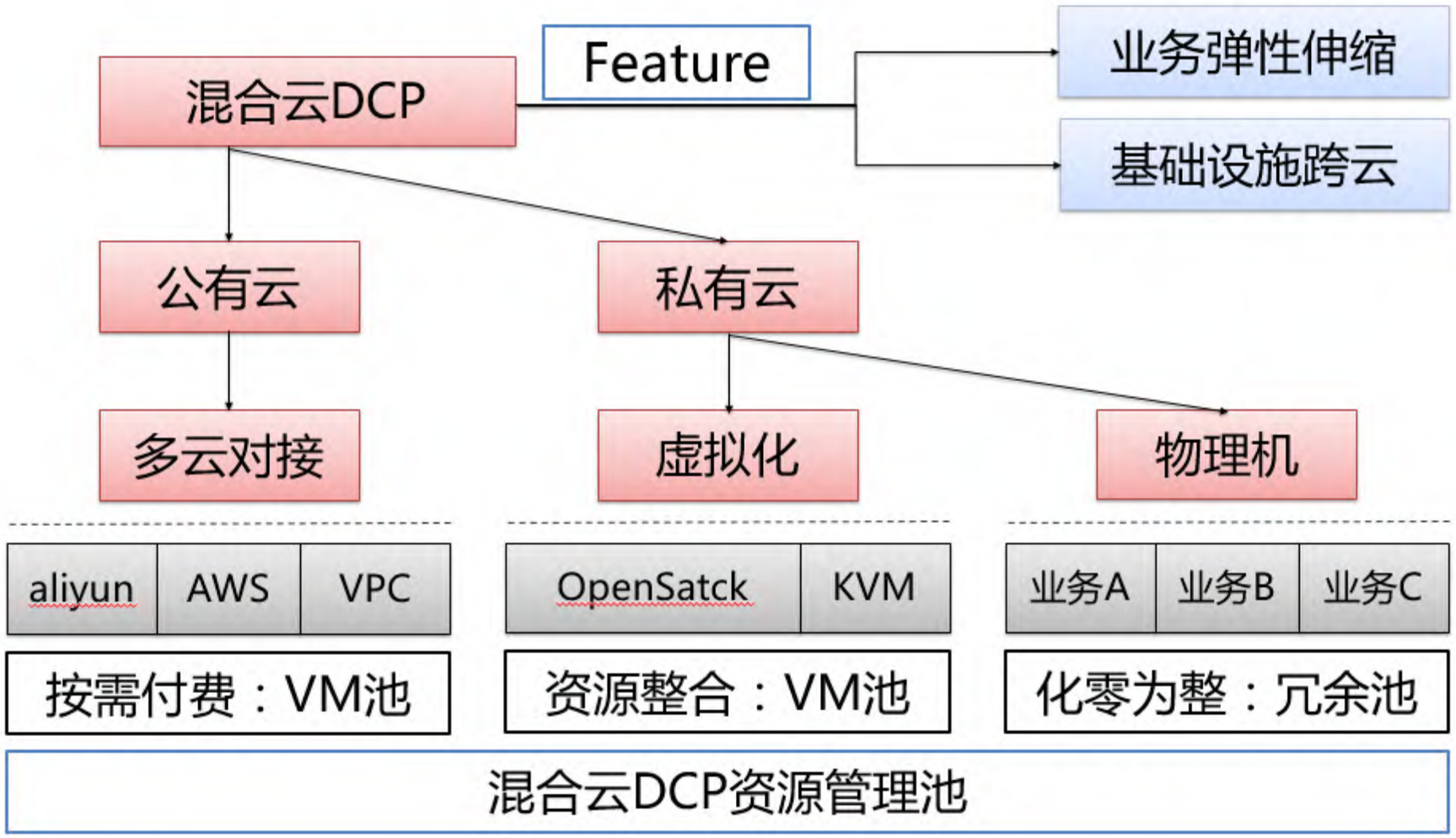
混合云DCP流程


主机申请
• 内网申请
• 云端申请
初始化
Docker环境
系统环境
动态调度
• 调度策略与算法
服务发现
• Nginx
• Motan
• SLB
下线
•反初始化
•归还Buffer池
•结算中心

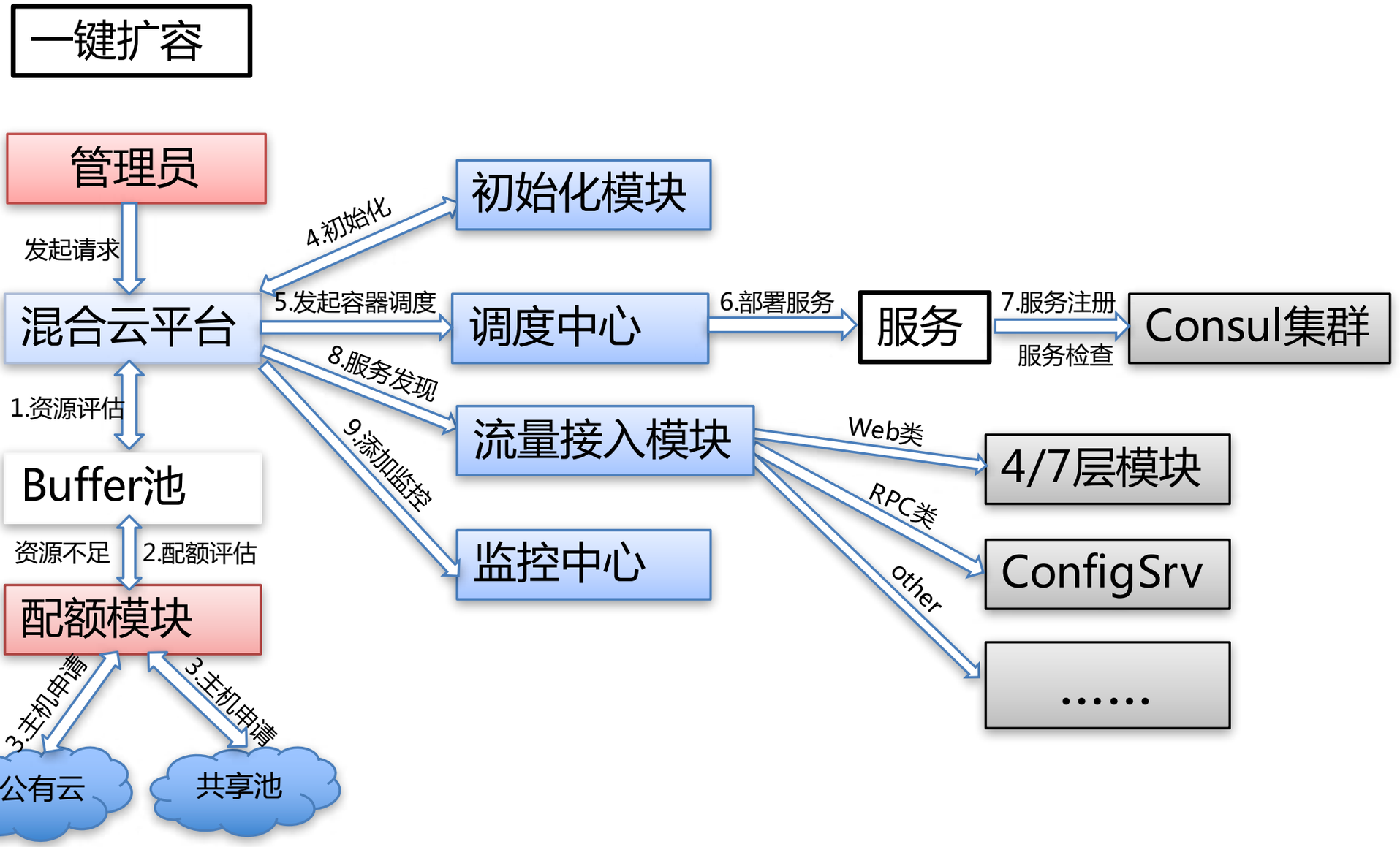
机器学习平台组件的Docker化

- 特征工程平台的各组件
- 流式计算Storm，Spark，Spark Streaming
- 热门微博
- Tensorflow Cluster

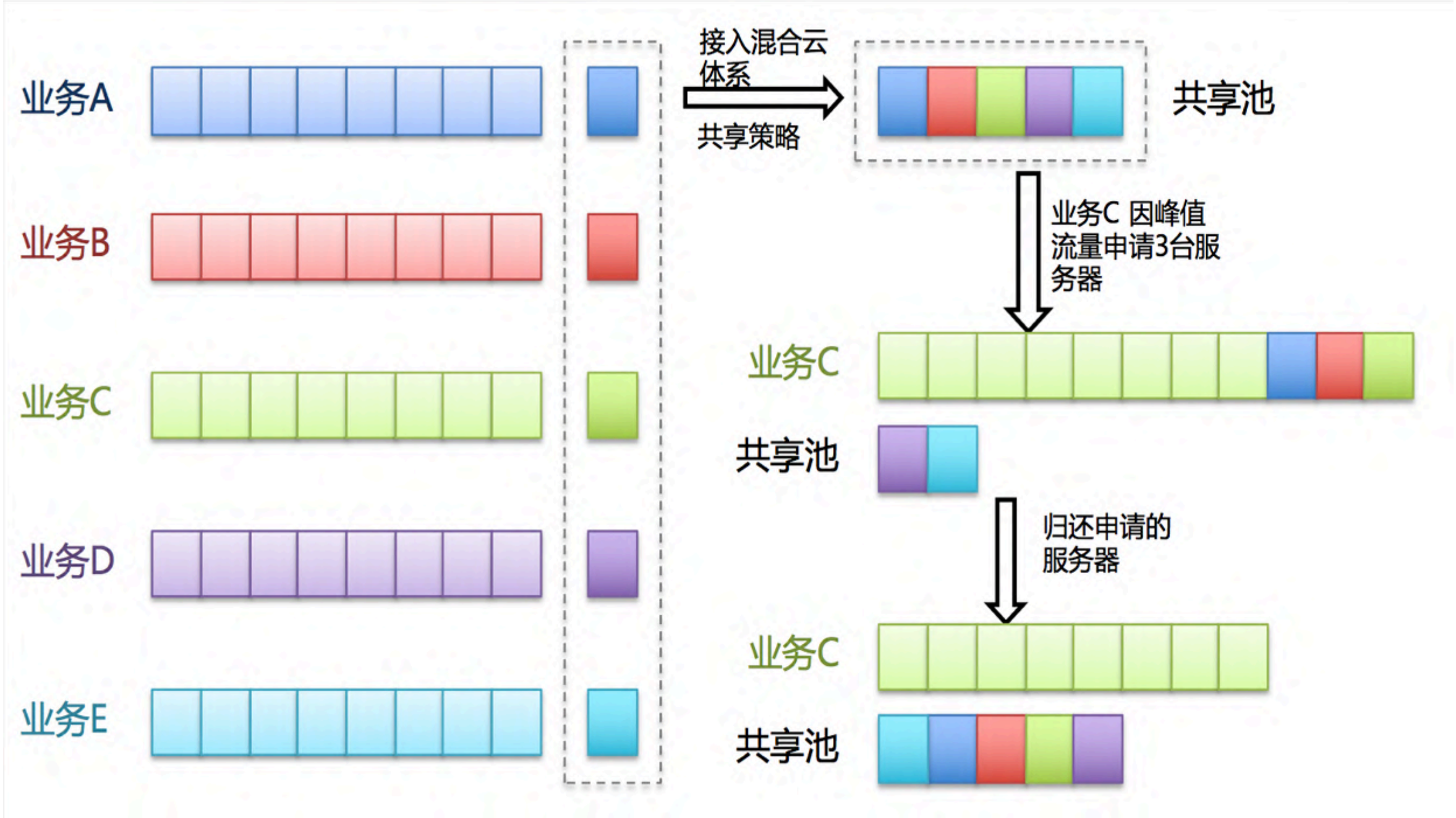
混合云DCP平台


混合云DCP
Feature
业务弹性伸缩
基础设施跨云
公有云
私有云
多云对接
虚拟化
物理机
aliyun
AWS
VPC
OpenSatck
KVM
业务A
业务B
业务C
按需付费：VM池
资源整合：VM池
化零为整：冗余池
混合云DCP资源管理池

混合云DCP平台弹性扩容



弹性调度下的资源共享


业务A
业务B
业务C
业务D
业务E
接入混合云体系
共享策略
共享池
业务C 因峰值流量申请3台服务器
业务C
共享池
归还申请的服务器
业务C
共享池

机器学习平台-公有云调度

- 使用阿里云OSS扩展大规模存储
- 使用阿里云ECS支撑如热门微博弹性扩容的计算资源需求
- 在阿里云IaaS上使用预制镜像，快速扩充Spark/Storm类流式分析服务

机器学习平台 - 内网虚拟化建设

- 置换陈旧设备，采用提高计算密度的方式，用于机架资源的高效利用
- 计算资源单元化，柔性支持业务扩展和缩容

OpenStack选型

- OpenStack Mitaka
- 可管理性
  - Keystone更细粒度的控制（Domains, Projects, Users, Groups, and Roles）
  - 单一的命令行客户端
  - SDK进展
  - Neutron API易用性

OpenStack选型

- 可扩展性
  - Heat的可扩展性
  - 调度器
  - LBaaS v2
- Nova
  - 热迁移

OpenStack选型

- KVM
- Ceph分布式存储

OpenStack部分调优

- 调整QEMU版本
- 系统参数优化
- 使用Rally项目进行扩容瓶颈压测

混合云DCP的内网虚拟化实践

- 生产引流压测
- 虚机类型配比模板
- 在同一宿主机上混合不同业务类型的虚机，尽可能利用宿主机的物理资源
- 目前规模300台物理宿主

OpenDCP

- 开源地址：https://github.com/weibocom/opendcp
- 综合性的运维管理平台。涵盖运维配置、发布、上线变更等运维管理主要功能，而不局限于容器集群管理，可适配Kubernetes、Mesos、Swarm等
- 功能覆盖镜像市场、多云对接、服务编排、服务发现等云资源管理主要环节
- 支持阿里云、AWS、私有云等主流云厂商
- 支持Nginx、SLB等服务发现方式
- 支持Java、PHP、C/C++、Go等主流语言

小结

- 解决了目前阶段的机架资源紧张
- 利用不同业务时段特征，实现资源的潮汐调度
- 快速部署机器学习的功能集群

近期的方向

- 基于Ocata版本的集群
  - Nova: Cell v2
  - Cinder: Active/Active HA
  - Neutron: Resource Tag Mechanism
- OVS-DPDK优化网络性能
- Kubernetes GPU集群

THANKS!