中国移动DC/OS实践

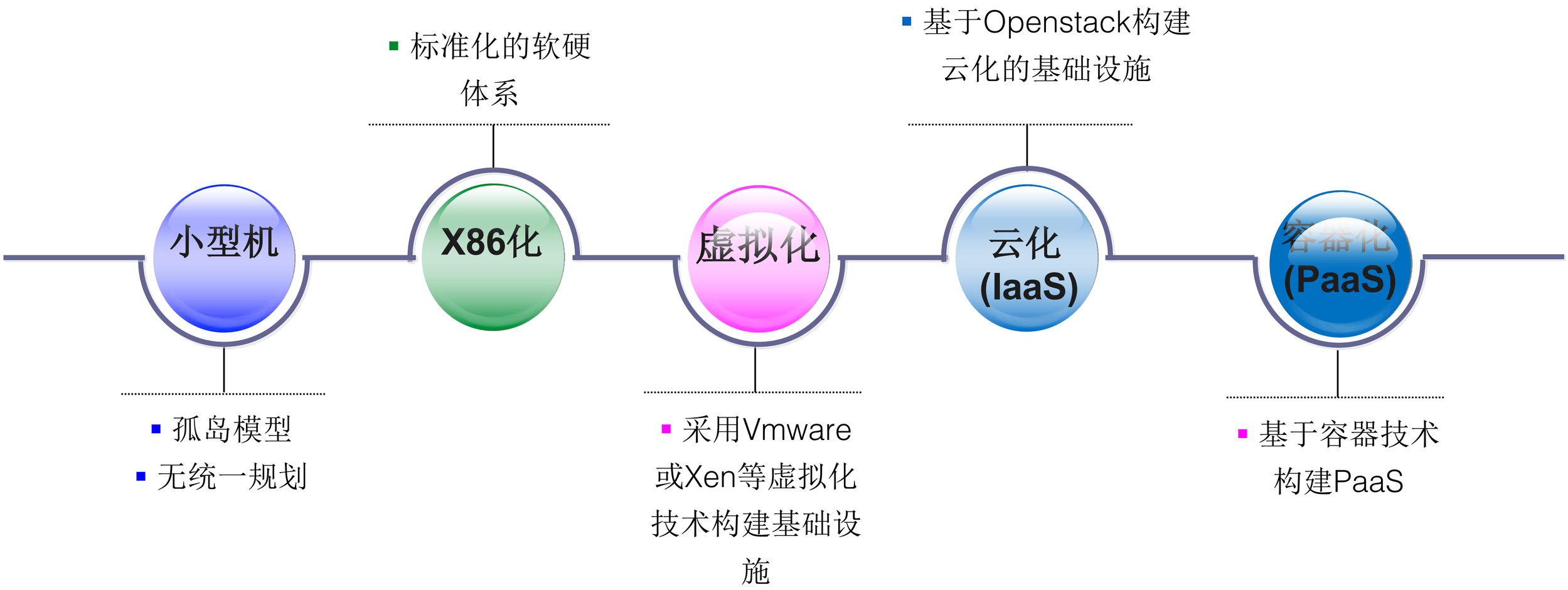
苏州研发中心

2016年9月

主要内容

- 中移苏研DC/OS研发背景
- 中移苏研DC/OS介绍
- 中移苏研DC/OS实践

数据中心操作系统（DC/OS）— 背景



中国移动 China Mobile

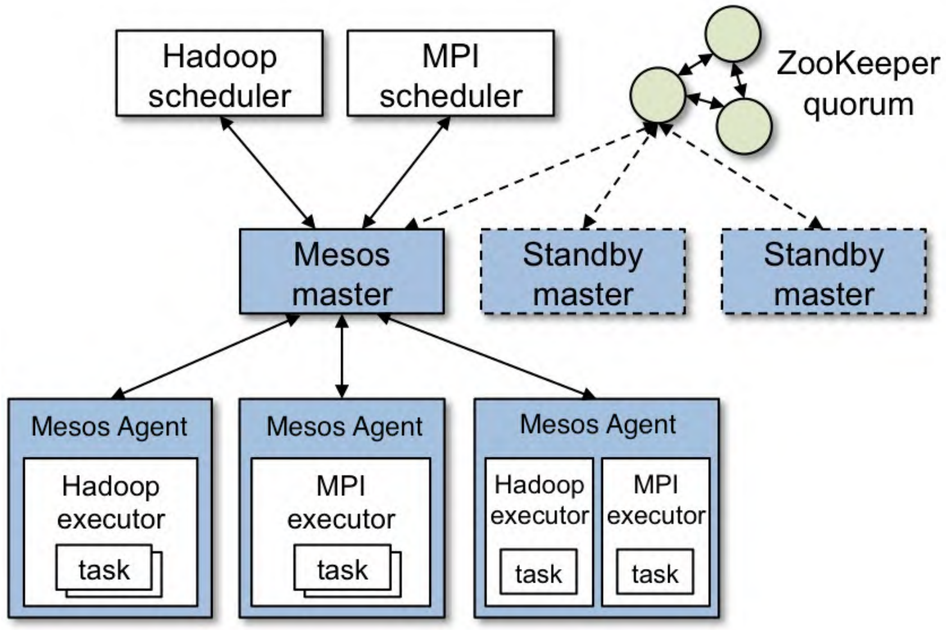
面临的问题

- 资源静态划分，整个数据中心资源利用率低
- 部署效率低下，无法满足业务的快速上线
- 应用弹性扩缩能力不足，应对互联网模式的业务显得能力不足
- 缺少业务生命周期统一管理的模式，运维复杂度高

数据中心操作系统（DC/OS）— Mesos+Marathon

- Mesos线性可扩展，可支持**10,000节点**
- Kubernetes/Swarm大规模生产案例较少
- 支持多种容器Docker、Appc等；可插拔的isolator：能够支持CPU、内存、磁盘、Port、GPU等隔离，可自定义isolator
- **两层调度**：Mesos负责资源管理与分配；上层framework负责在分配的资源上调度任务，因此framework也叫作scheduler

Power By Mesos

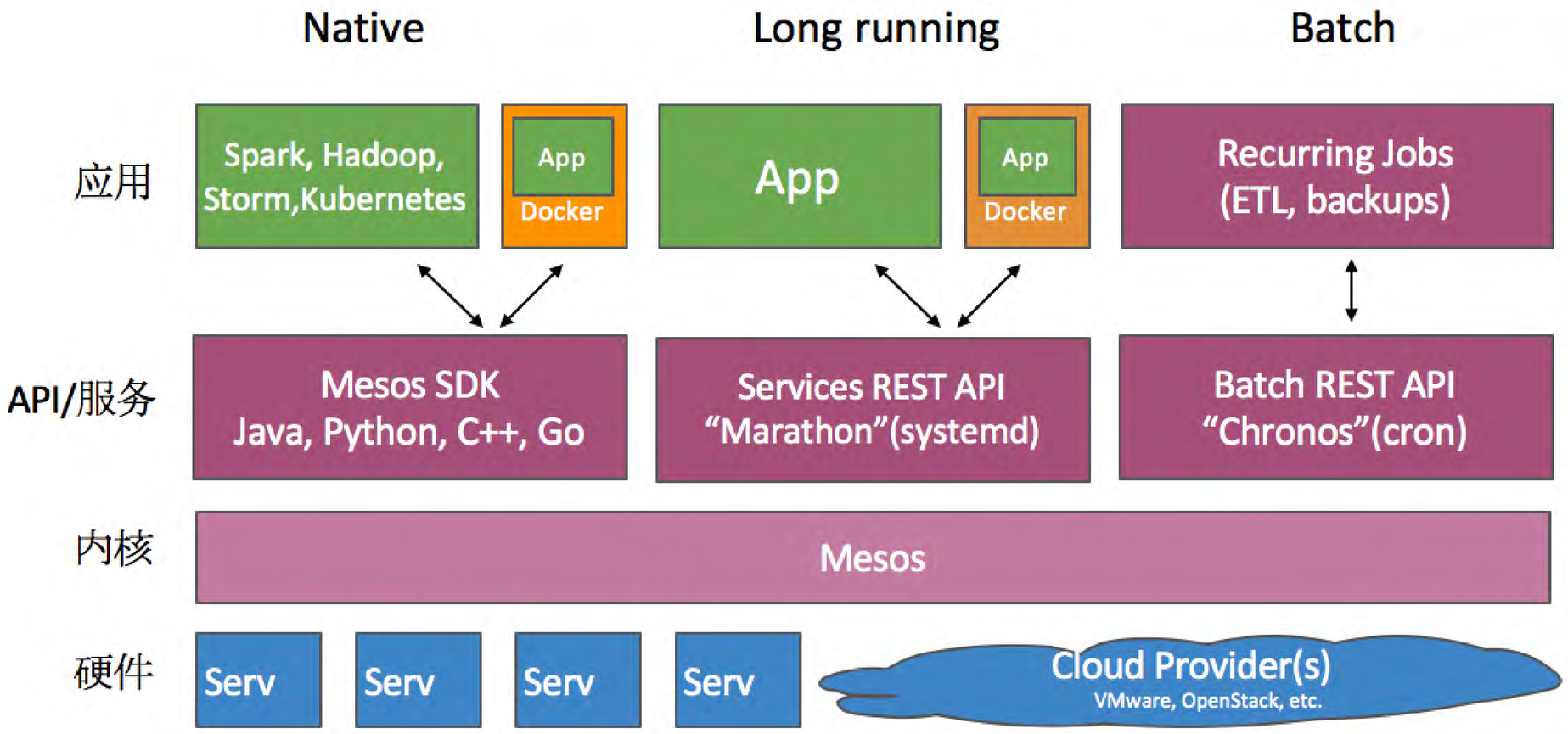


DevOps tooling	vamp
Long Running Services	**Aurora**、**Marathon**、**Swarm**、**Kubernetes**、Sigularity、SSP
Big Data Processing	Cray Chapel 、Dpark、Exelixi、**Hadoop**、Hama、MPI、**Spark**、**Storm**
Batch Scheduling	**Chronos**、**Jenkins**、JobServer、GoDocker、Cook
Data Storage	Alluxio、**Cassandra**、**Elasticsearch**、Hypertable、MrRedis

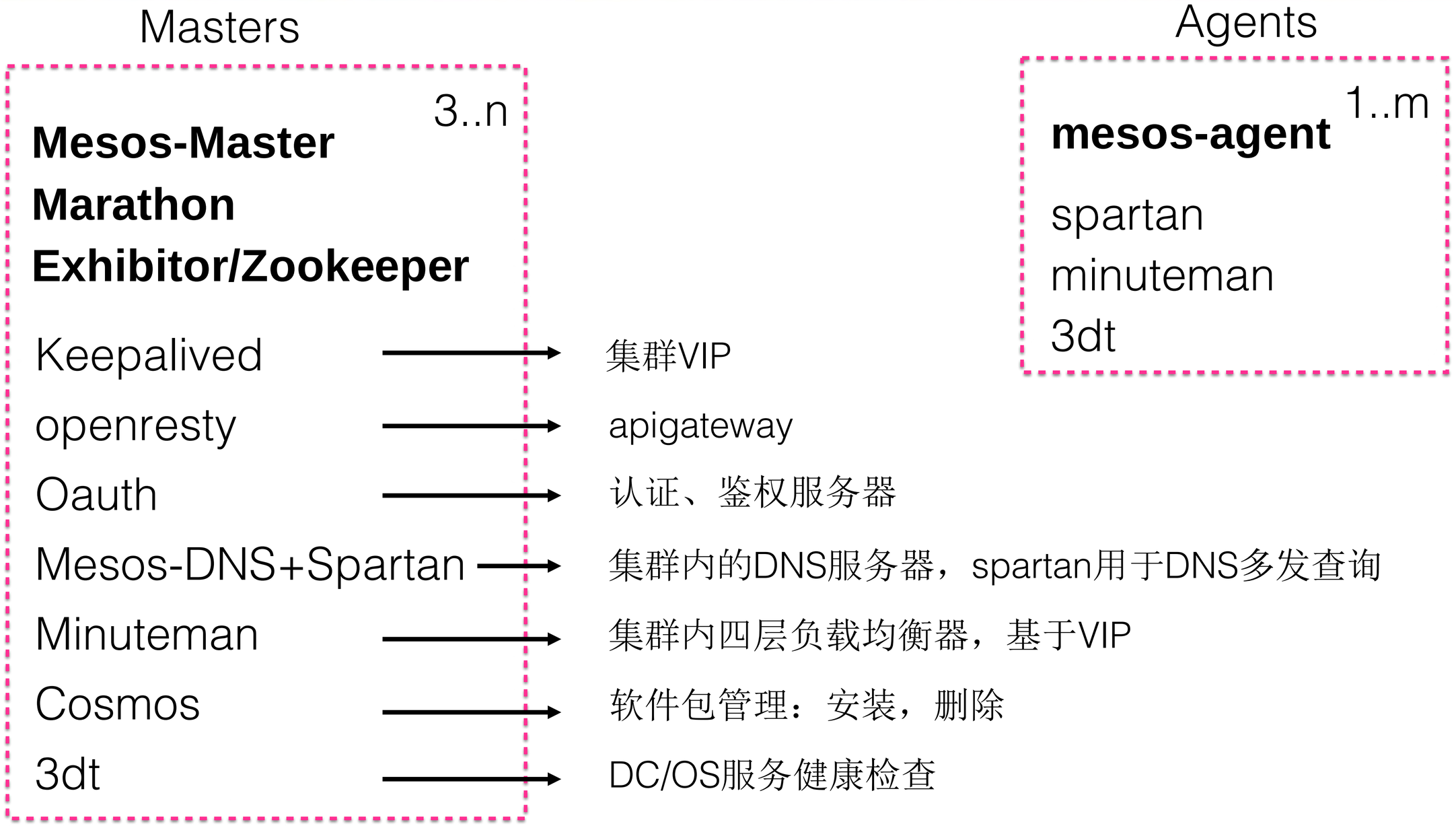
frameworks

数据中心操作系统（DC/OS）— Open DC/OS

Open DC/OS是Mesosphere DCOS的开源版本，是围绕着Mesos + Marathon的软件栈（Bundle），提供开箱即用的DC/OS。



数据中心操作系统（DC/OS）— Open DC/OS



数据中心操作系统（DC/OS）—SY-DC/OS

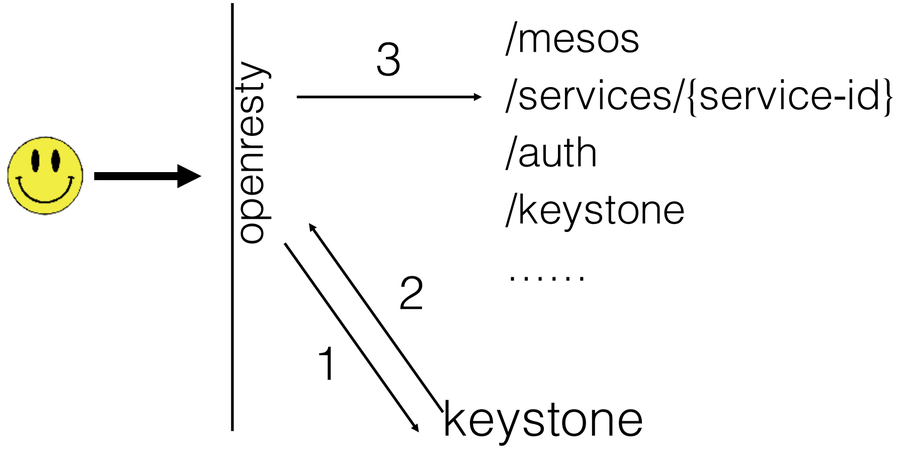
Open DC/OS不满足我们的需求：

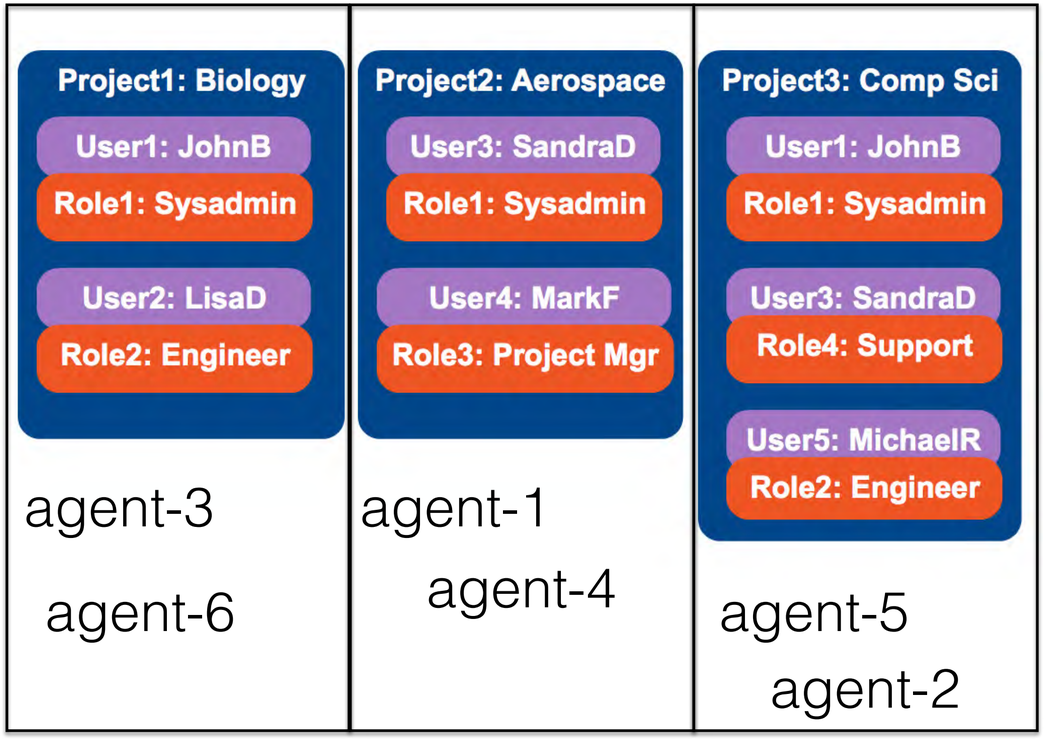
- 缺少多租户支持；
- 没有提供监控、告警、和日志的解决方案
- 不支持资源的管理，如分配主机；
- 没有镜像仓库解决方案；
- 没有离线的用户管理机制，dcos-oauth对接的是auth0的API
- 缺少LB的集成化展示
- 软件仓库不支持跨marathon部署
- 缺少k8s的支持
- GUI业务流程的定制化

数据中心操作系统（DC/OS）— 多租户

用户模型(Openstack Keystone)

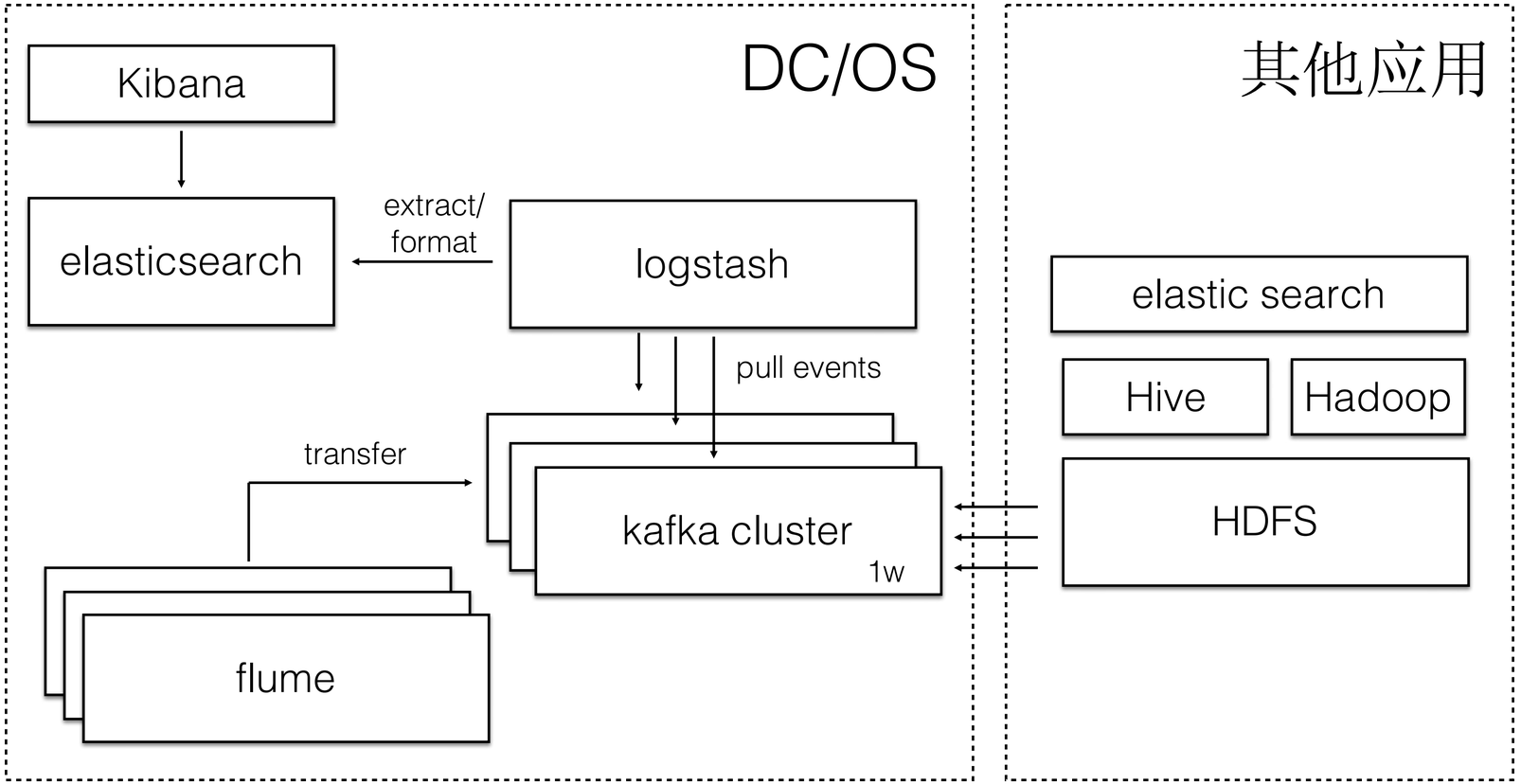
- 一个mesos role对应keystone的project
- 用户模型：dc-admin, project-admin, member
- dc-admin是超级管理员，拥有最大的权限，可以分配资源、CURD project等
- 默认配置下有dc-admin-role、sys role以及*role，dc-admin-role的资源只能dc-admin使用，*的资源可以公用；各个project都有对应project name的mesos-role。
- dc-admin以物理节点为单位为租户分配资源
- project-admin可以单独通过软件仓库部署服务
- project-admin可以安装服务，如marathon、k8s、swarm等；在DC/OS中，service指的就是framework。frameworks只能使用本project内的资源





数据中心操作系统（DC/OS）— 日志

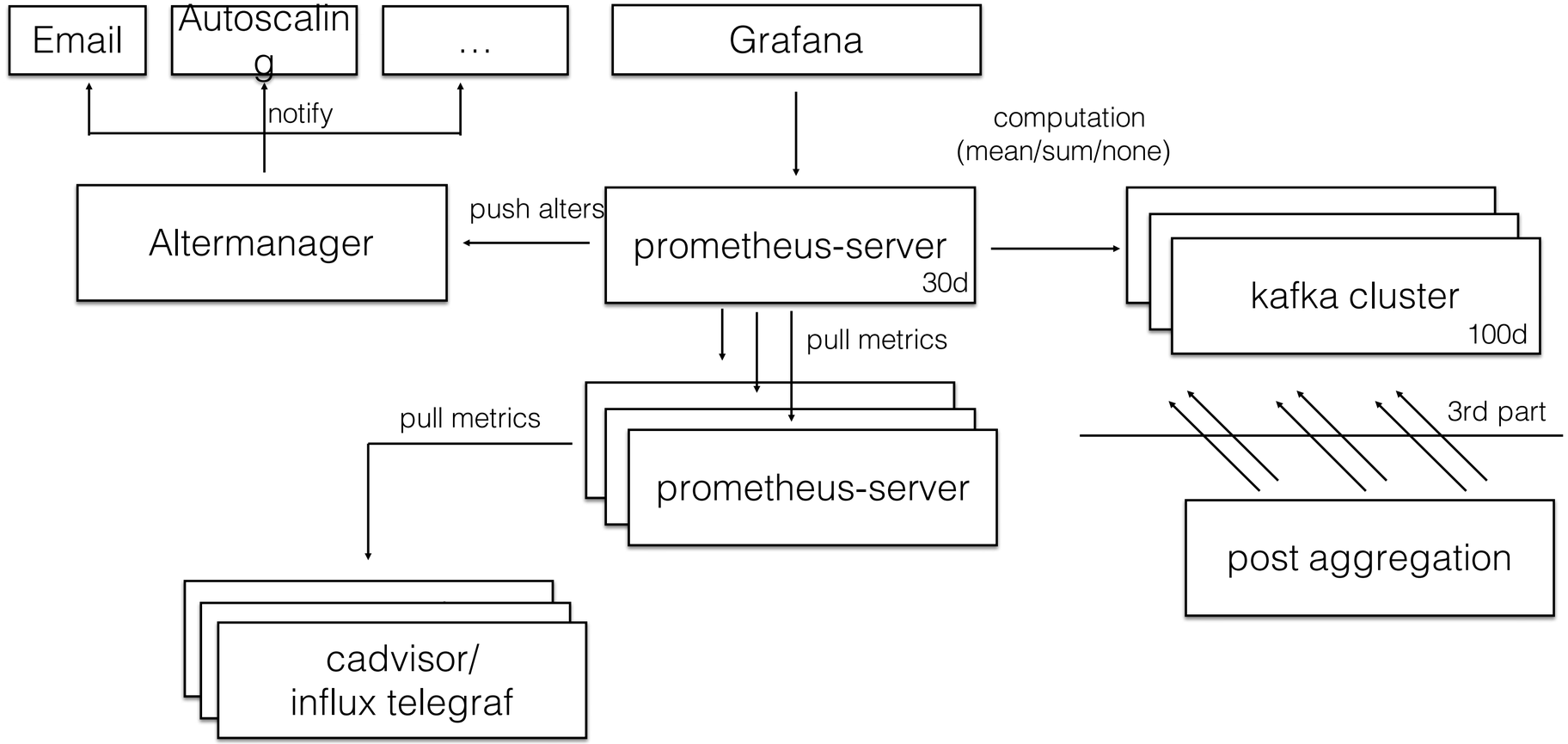
日志



- 每台主机上安装flume采集日志；
- 每个project对应一个topic：dc-admin-role对应dc_admin_topic
- 每条日志都是一个kafka event，header标识为：hostname+path等
- 应用日志必须写到sandbox中；
- 租户的日志自己解析（elasticsearch的日志是同一存放）

数据中心操作系统（DC/OS）— 监控

监控

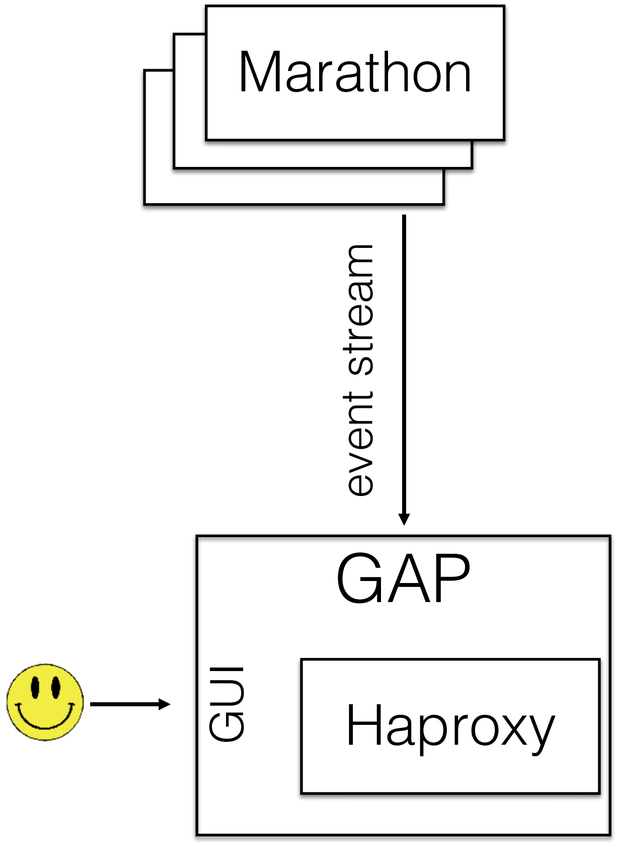


- 主机/应用metrics通过cadvisor/influx telegraf采集
- cadvisor采集通用指标，如CPU、Ram、Network指标
- influx telegraf采集自定义指标，如haproxy的session数、线程数
- prometheus组成级联的集群，定时pull metrics存储本地；
- prometheus不断evaluate 各项指标，并通过altermanager发布告警
- prometheus 本地保存30天的记录，本把历史记录通过计算，或直接传送都kafka持久化保存100天

数据中心操作系统（DC/OS）— 服务发现

服务发现GAP

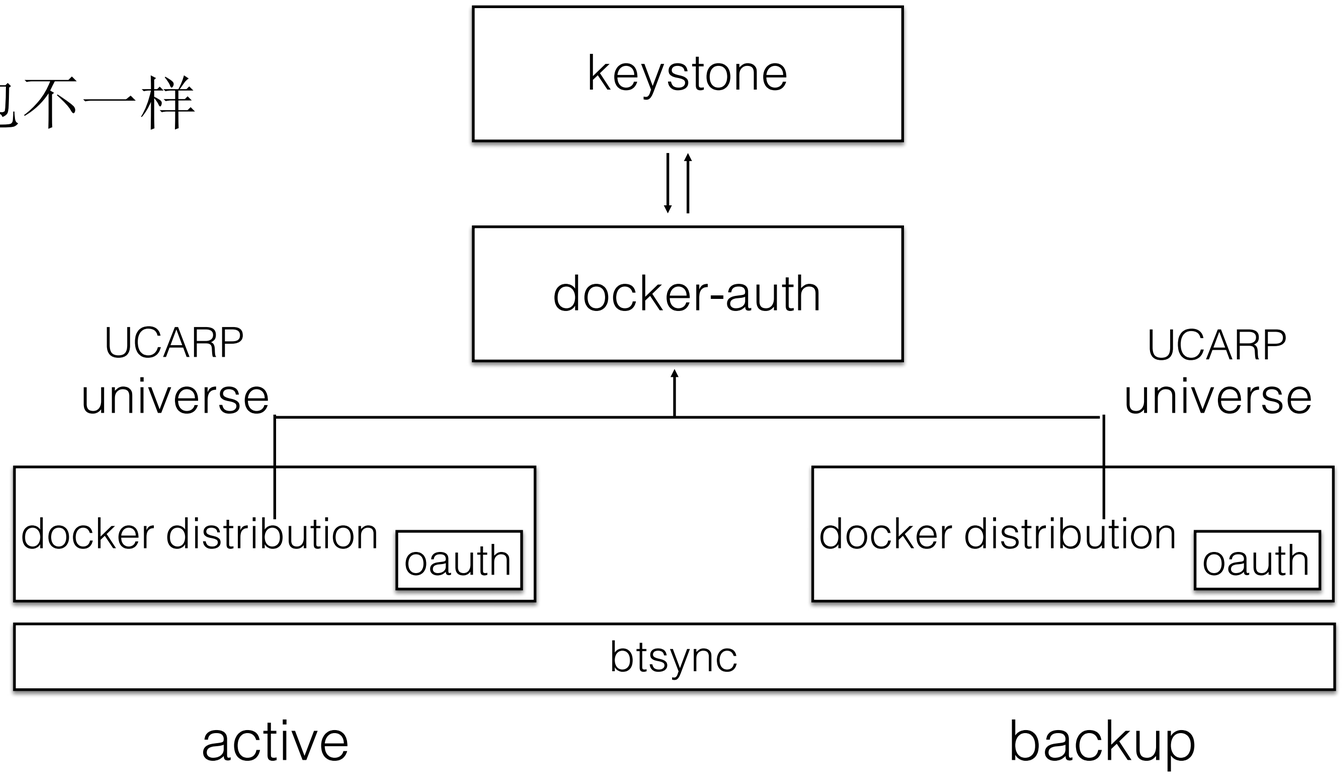
- 集群内部使用基于DNS和VIP的服务发现机制
- 外部访问集群内应用使用GAP(定制化得haproxy)
- GAP是软件仓库中的一个软件，用户可以直接使用界面安装到指定的某台agent节点；
- 通过给GAP增加label，使得TASC能够在界面上的haproxy上展示；
- haproxy的性能数据能够被收集，并运用到autoscaling、灰度发布等；
- 通过GAP页面能够隔离特定的容器（故障隔离、维护等）；
- 可通过marathon的label，或GAP页面Haproxy的参数，如ACL规则、负载均衡策略



数据中心操作系统（DC/OS）— 镜像仓库

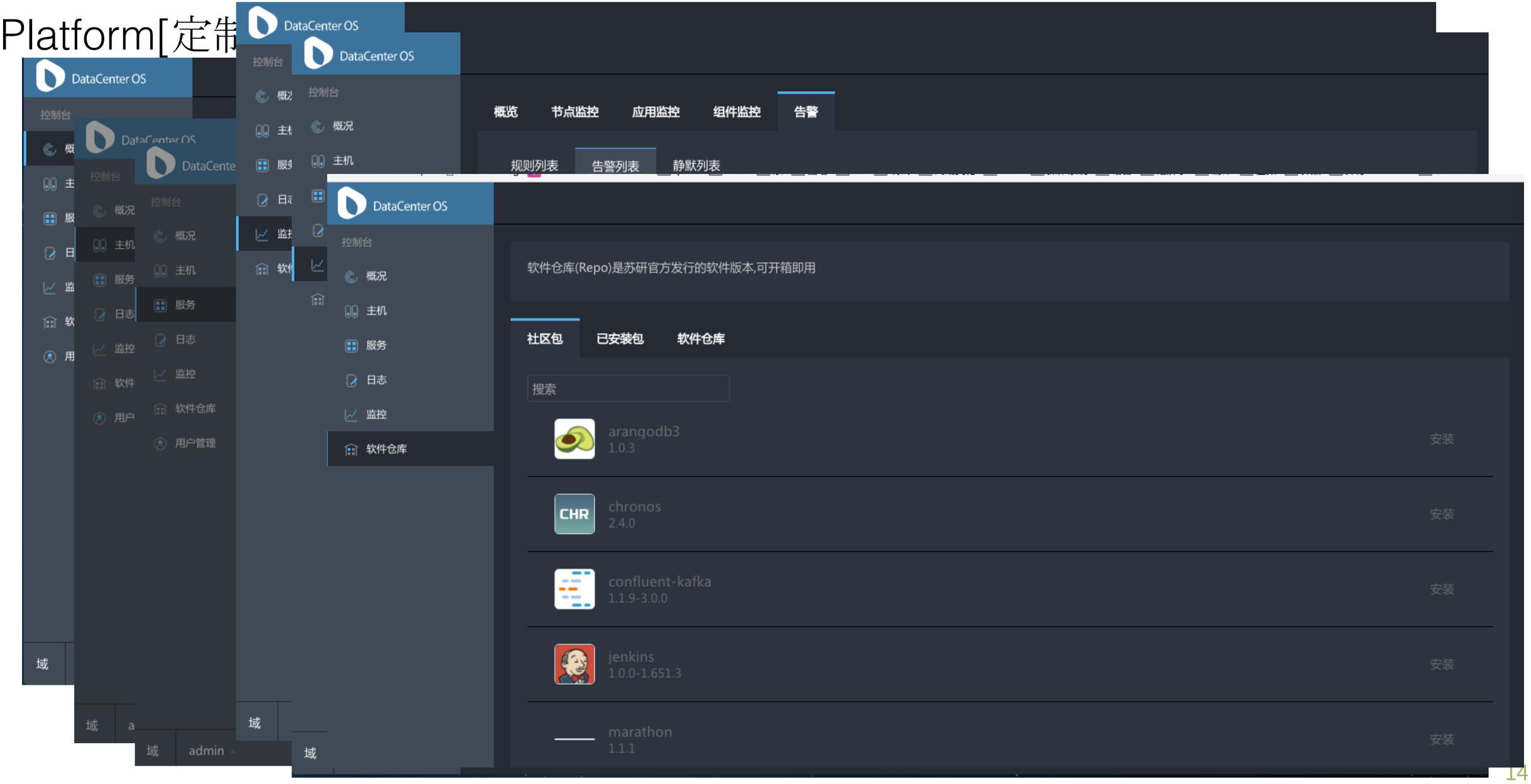
镜像仓库&&软件仓库

- 主备模式；软件仓库基于mesosphere的universe
- UCARP提供VIP，btsync提供增量备份
- 增强版的cosmos[Open DC/OS软件包管理器]
  - 可在不同的marathon上部署软件
  - 根据角色的不同，能够看到的软件包不一样



数据中心操作系统（DC/OS）— Platform

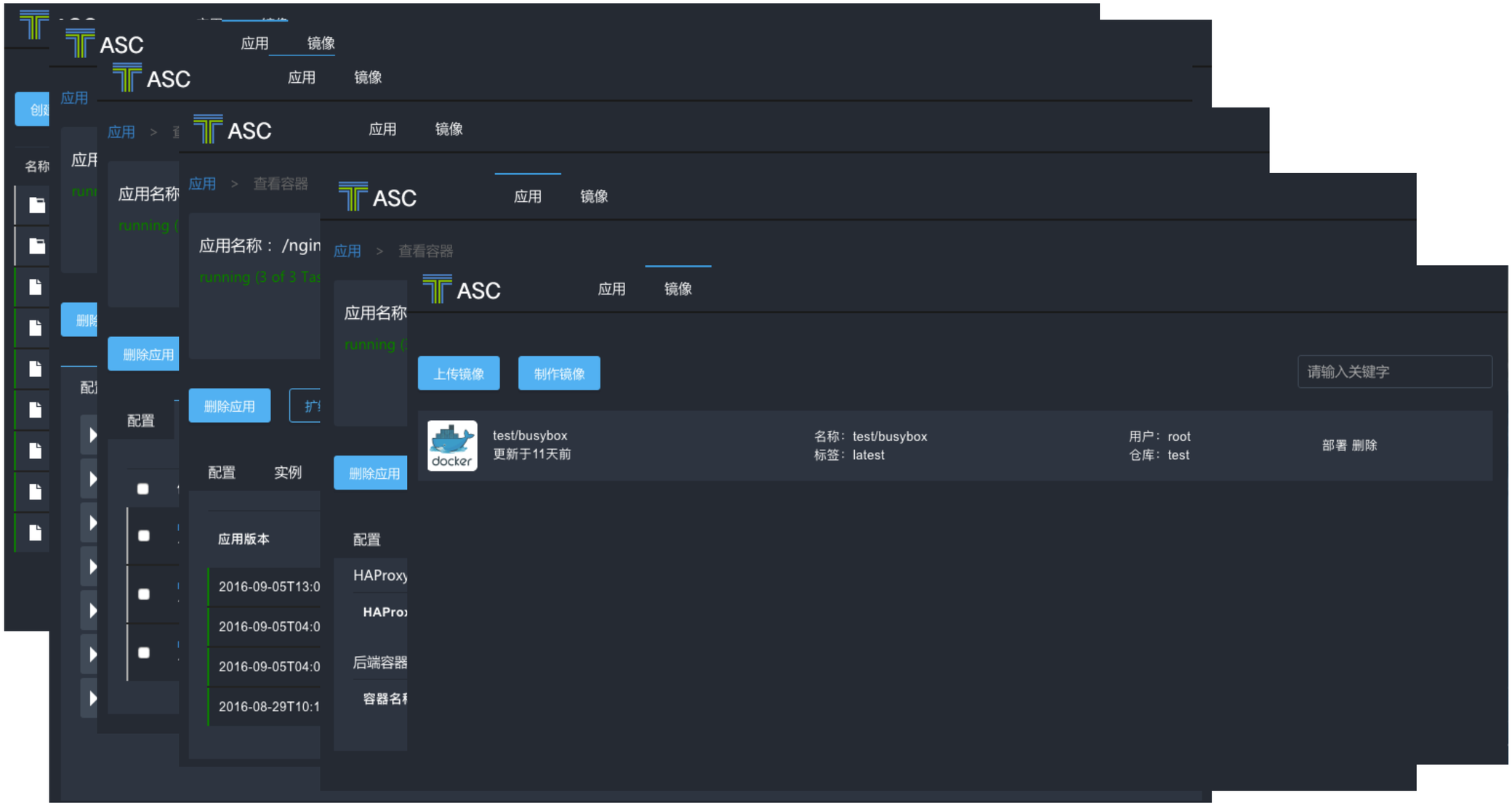
Platform[定制



数据中心操作系统（DC/OS）—TASC

中国移动 China Mobile

TASC[定制化的marathon]



数据中心操作系统（DC/OS）— 实践之应用场景

中国移动 China Mobile

访问量集中，突发流量大

- 符合分布式无状态应用系统特征的应用
- 能够自动化配置资源到最有效被利用的地方，实现资源弹性伸缩
- 优化开发、调测、部署操作，实现应用程序敏捷开发，快速部署上线

数据中心操作系统（DC/OS）— 实践之经验

中国移动 China Mobile

- Mesos的API存在性能问题，并发性能很低，需要做缓存
- 通过压力测试，推算应用性能拐点，合适设置容器的资源配置
- 给应用打上合理的标签，以便监控、日志系统能够区分
- CentOS系统上推荐使用XFS文件系统，使用overlayfs driver和ext4文件系统时，ubuntu的镜像可能会有问题
- Host机器尽量使用X86架构

谢谢