



中国移动
China Mobile



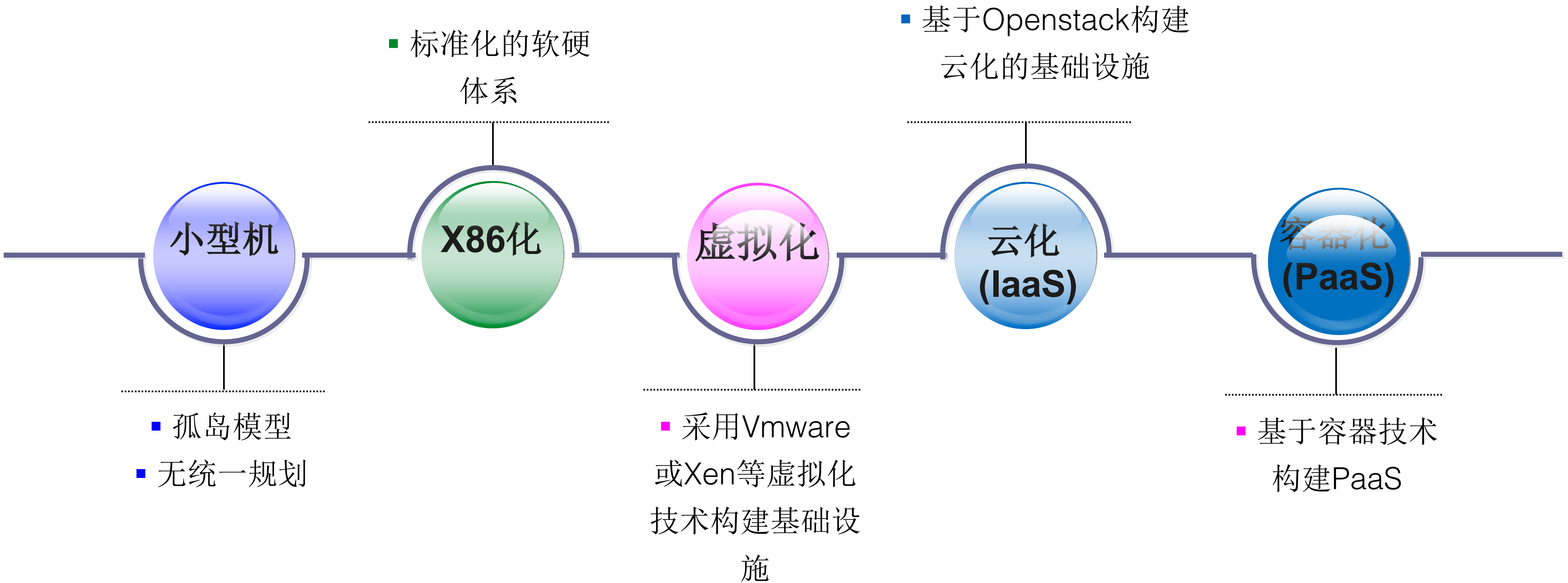
中国移动DC/OS实践

苏州研发中心

2016年9月

中国移动内部资料，
未经允许不得复制、转发、传播。

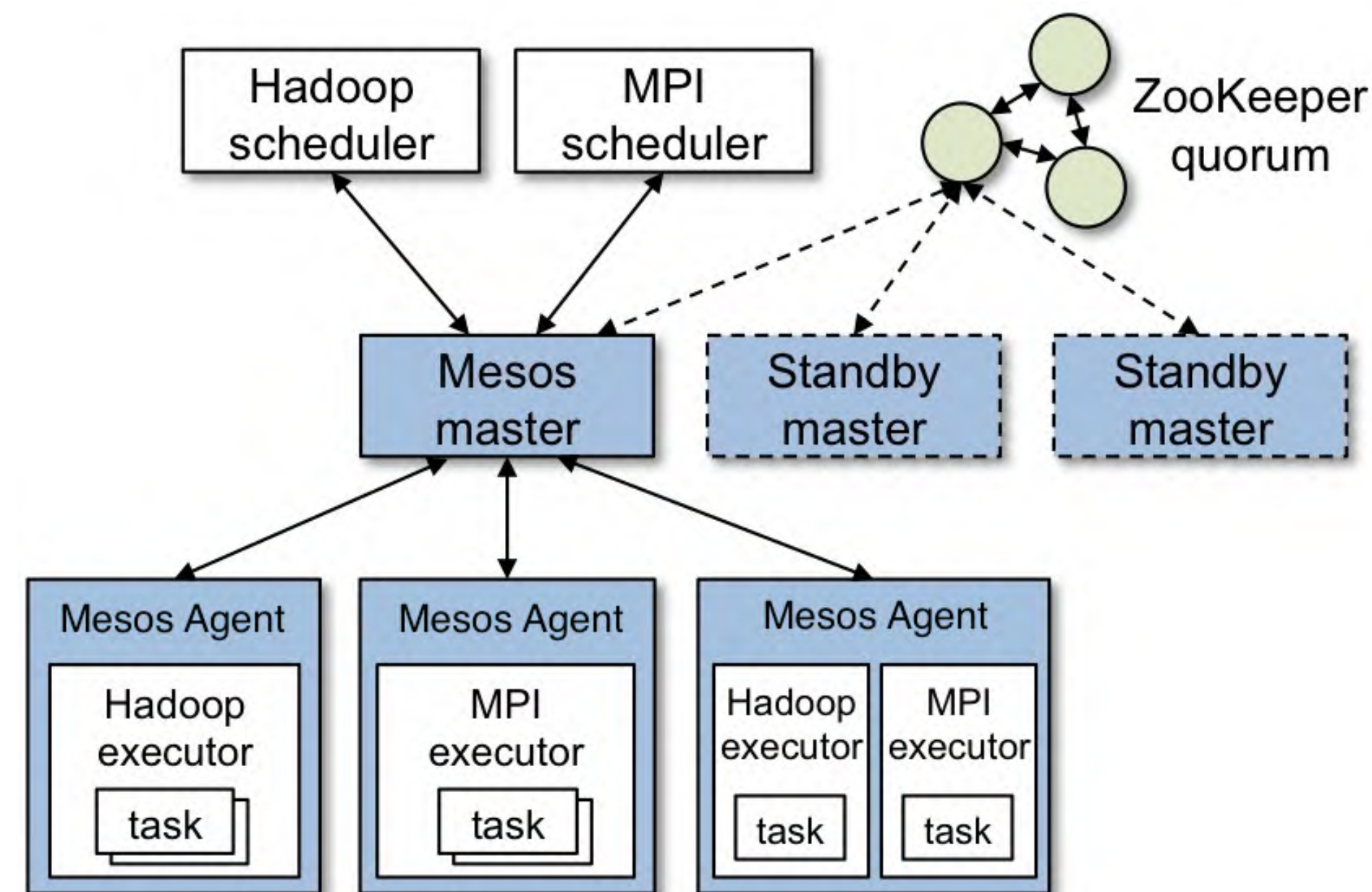
- 中移苏研DC/OS研发背景
- 中移苏研DC/OS介绍
- 中移苏研DC/OS实践



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

- ▶ 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



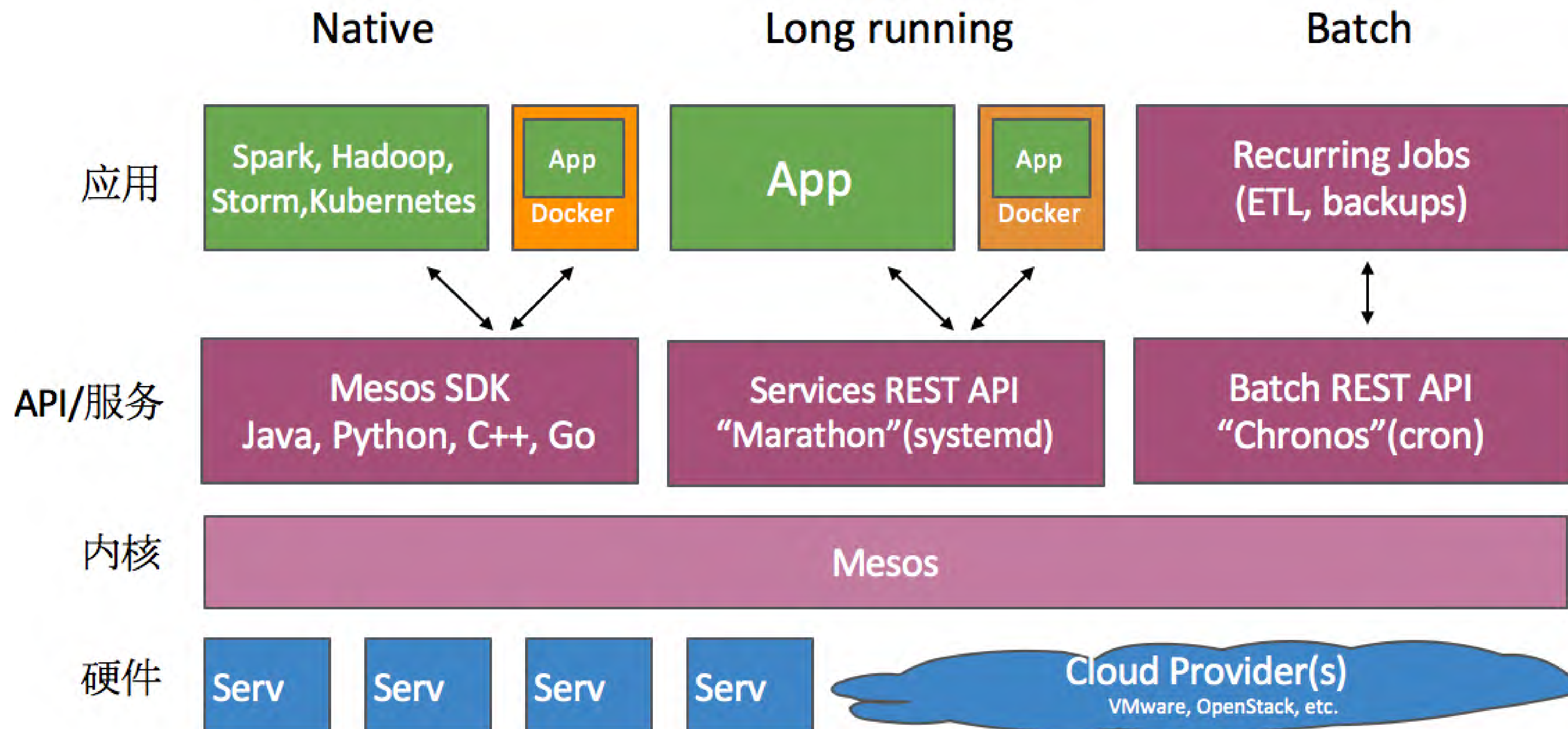
| | |
|------------------------------|--------------------------------------------------------------------------------------|
| DevOps tooling | vamp |
| Long Running Services | Aurora , Marathon , Swarm , Kubernetes , Sigularity, SSP |
| Big Data Processing | Cray Chapel, Dpark, Exelixi, Hadoop , Hama,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



Power By Mesos

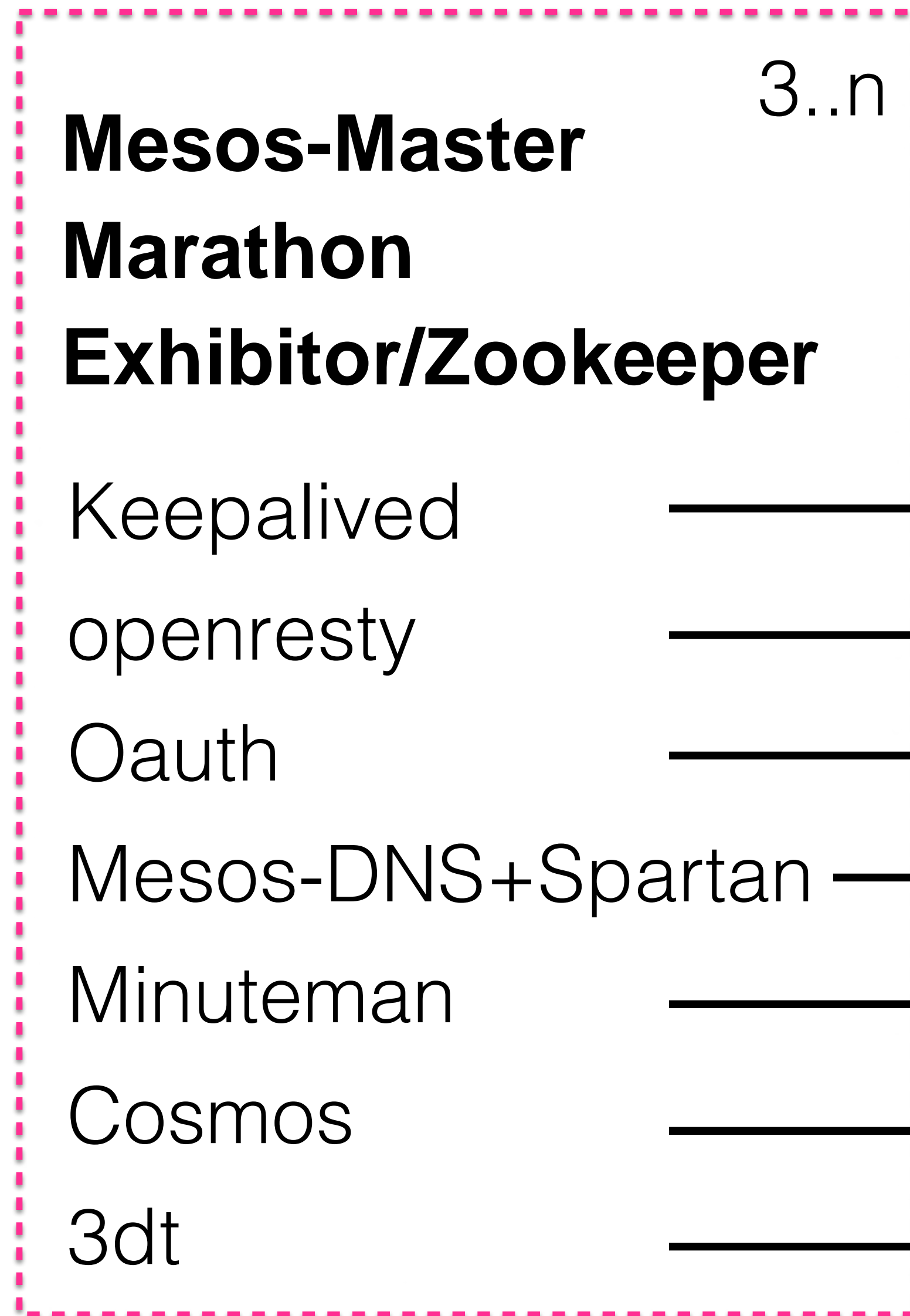
frameworks

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。



Masters

Agents



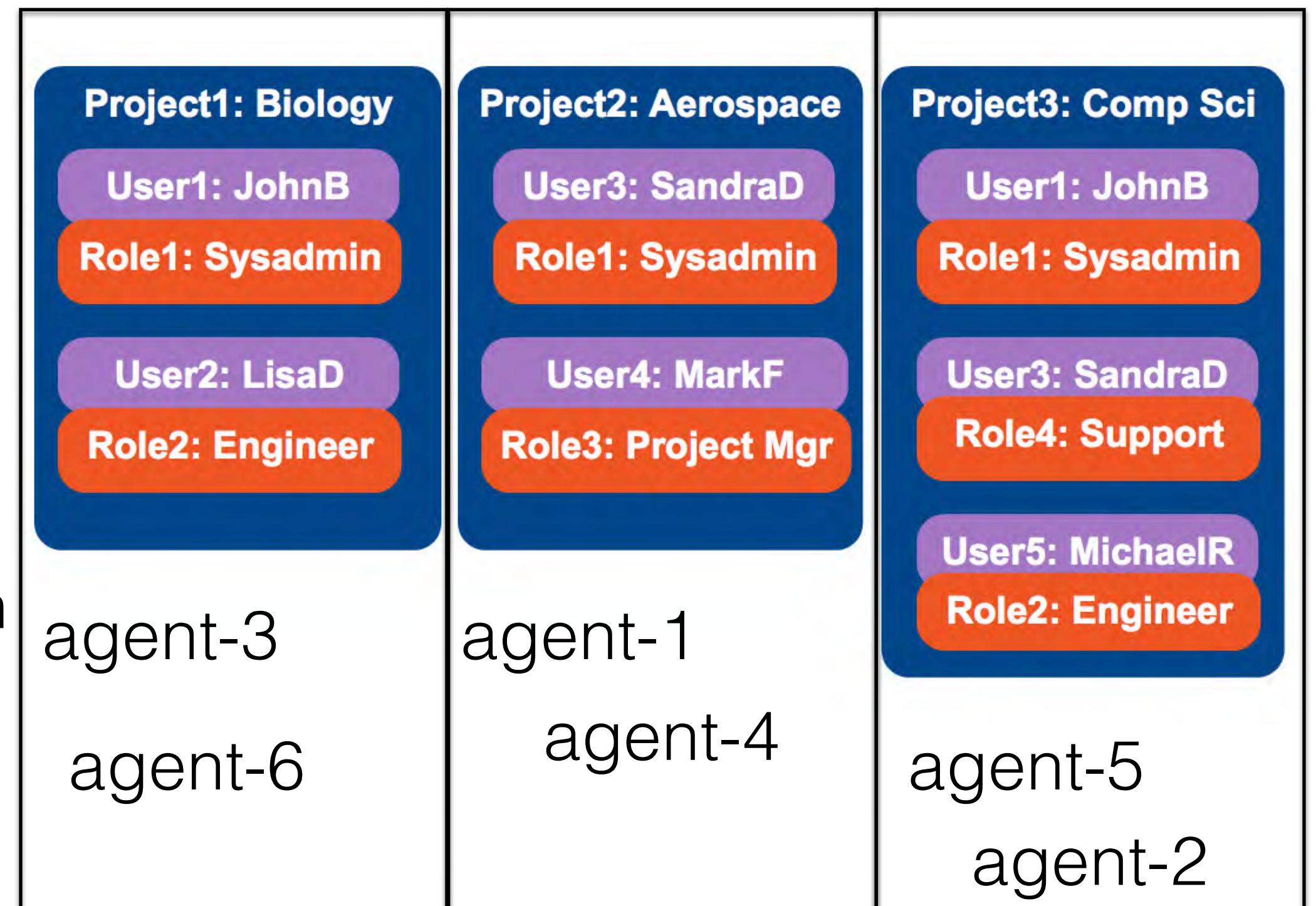
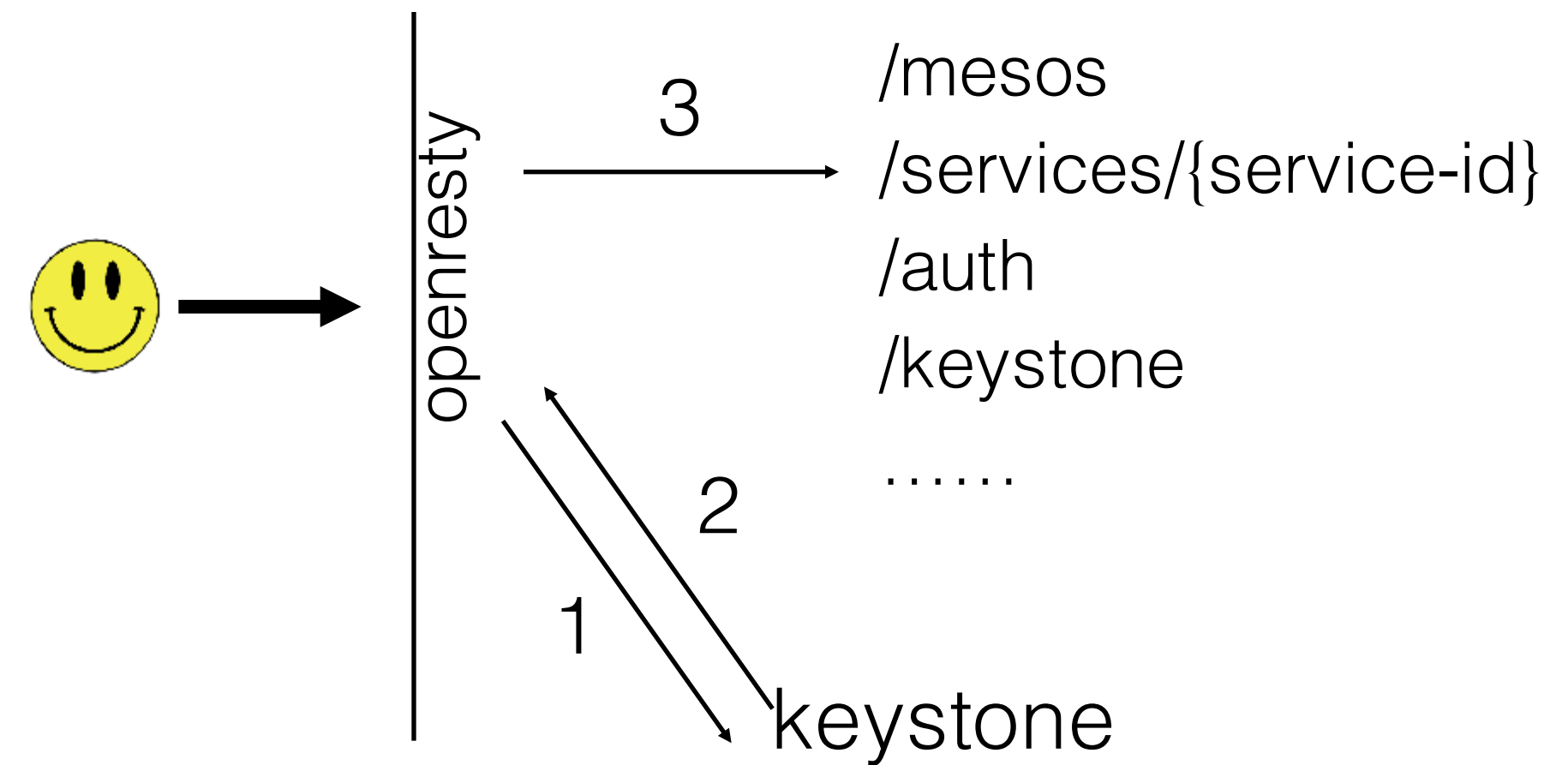
- 集群VIP
- apigateway
- 认证、鉴权服务器
- 集群内的DNS服务器, spartan用于DNS多发查询
- 集群内四层负载均衡器, 基于VIP
- 软件包管理: 安装, 删除
- DC/OS服务健康检查

Open DC/OS不满足我们的需求：

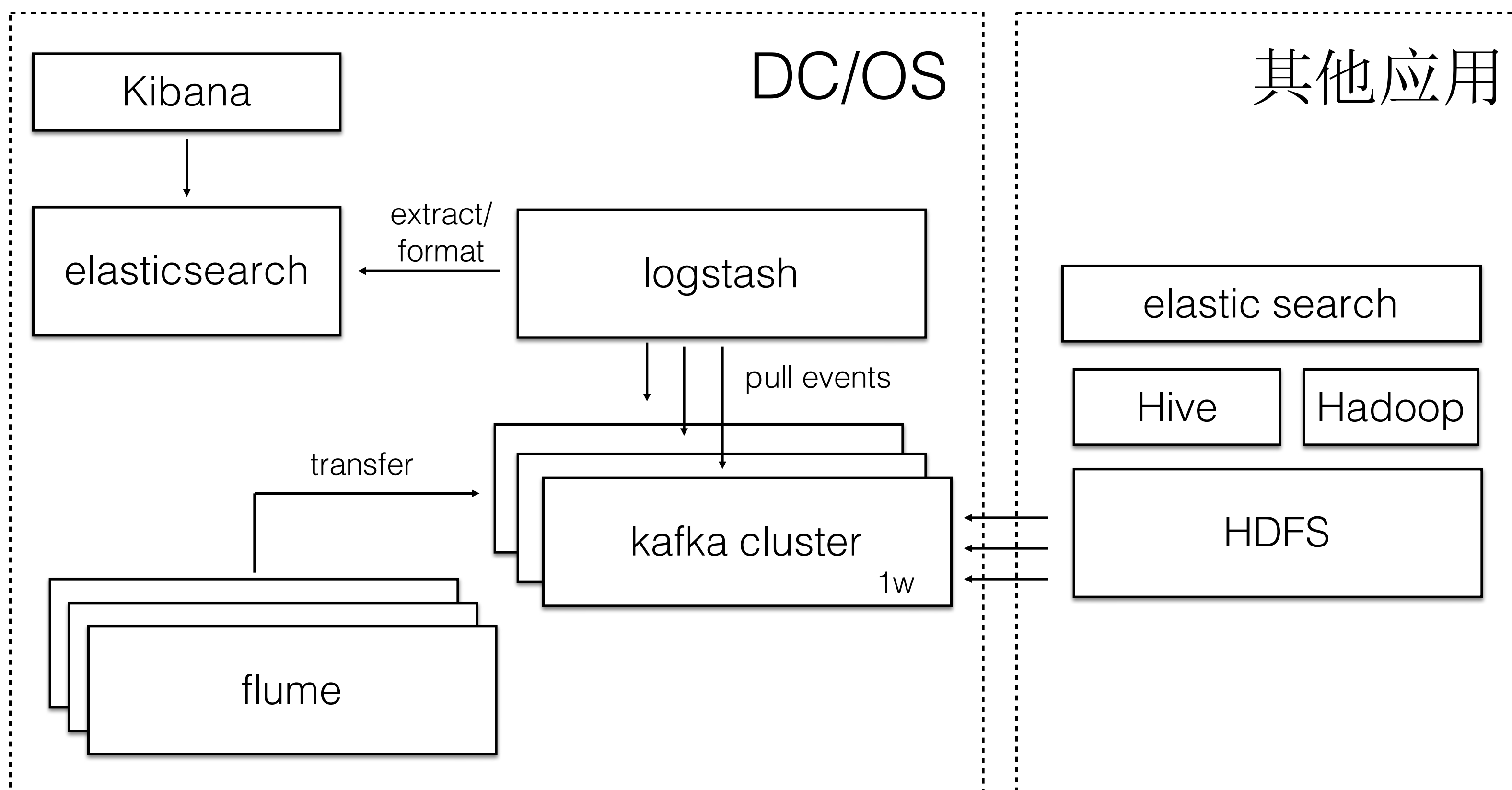
- 缺少多租户支持；
- 没有提供监控、告警、和日志的解决方案
- 不支持资源的管理，如分配主机；
- 没有镜像仓库解决方案；
- 没有离线的用户管理机制，dcos-oauth对接的是auth0的API
- 缺少LB的集成化展示
- 软件仓库不支持跨marathon部署
- 缺少k8s的支持
- GUI业务流程的定制化

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

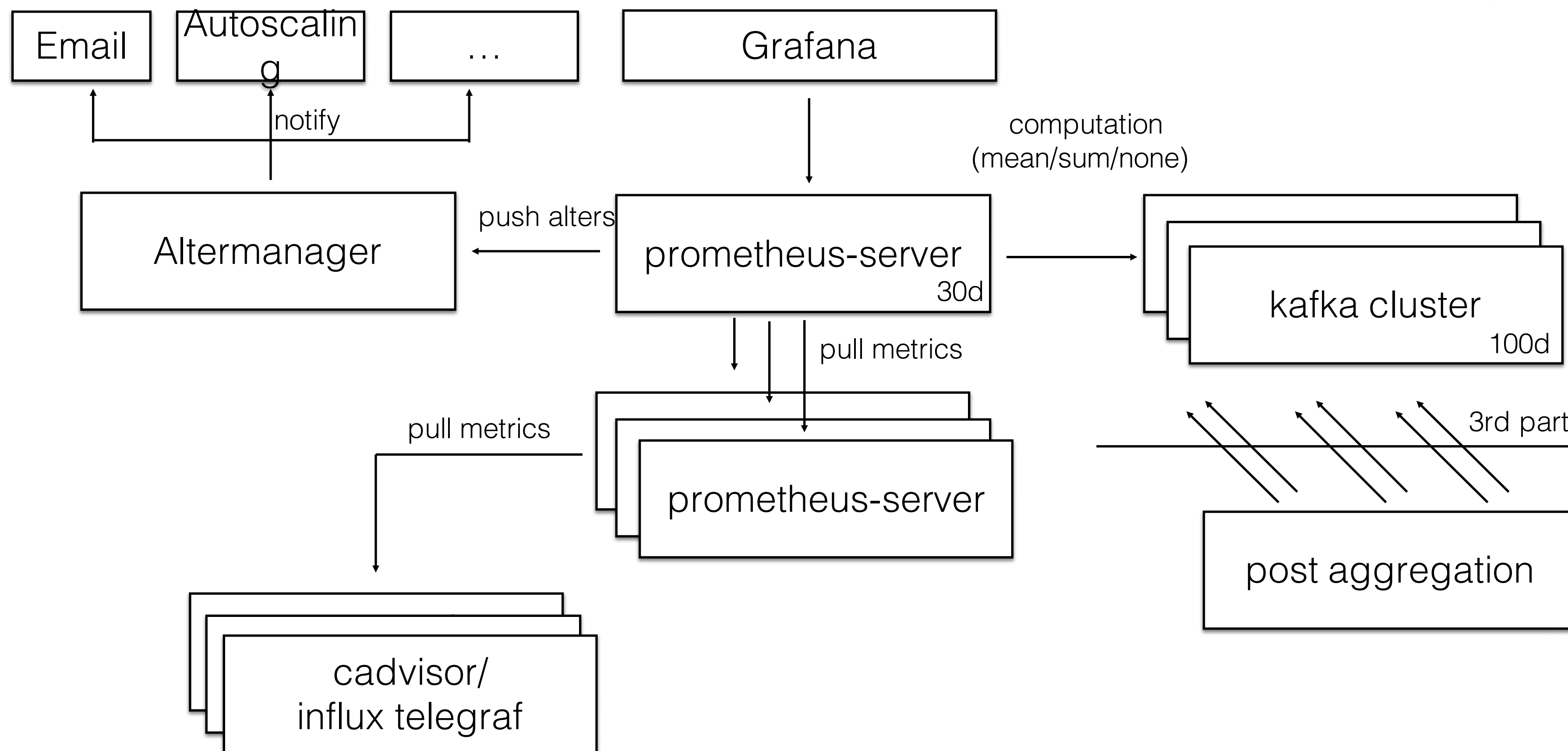


日志



- 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）

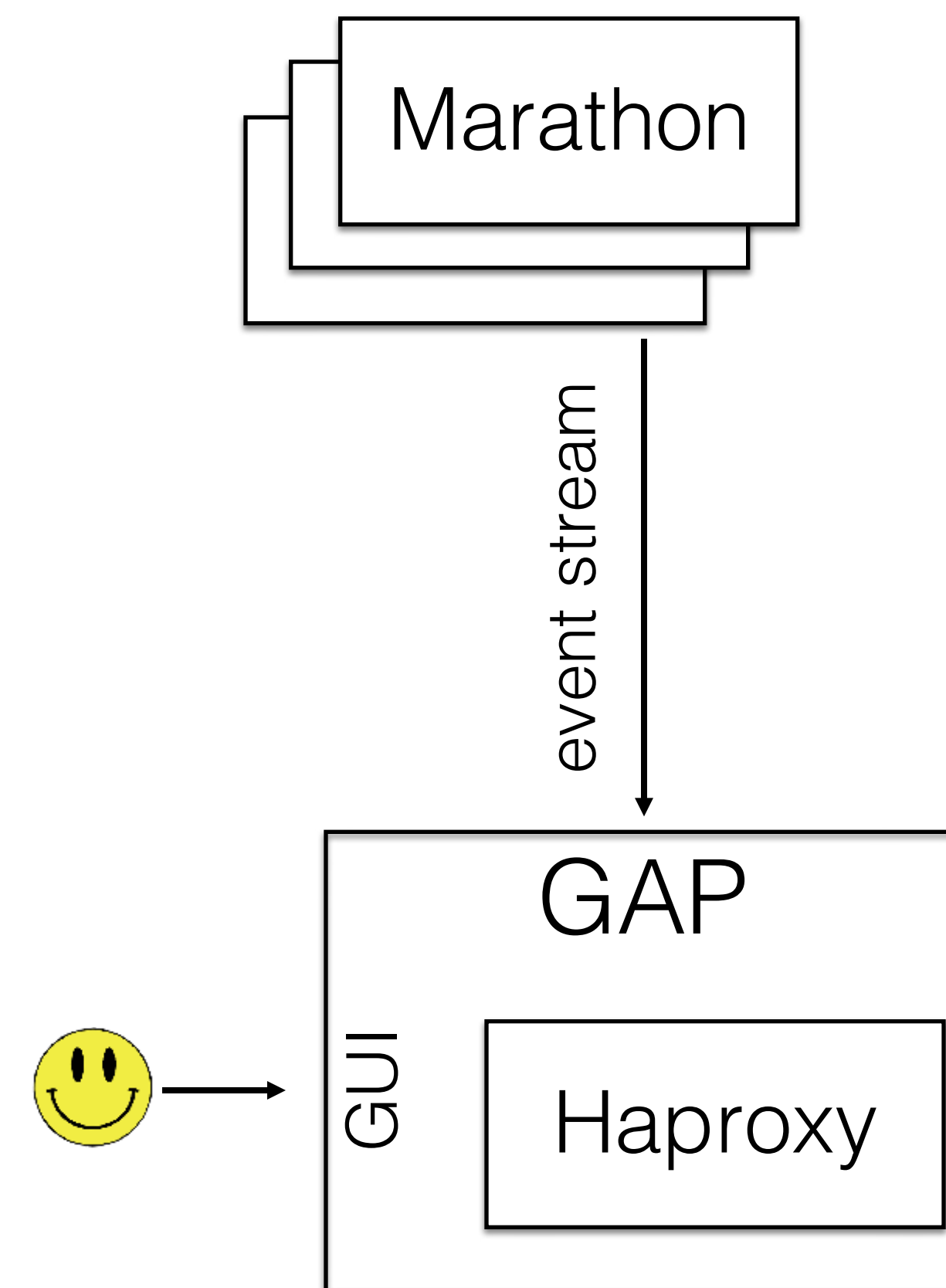
监控



- 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

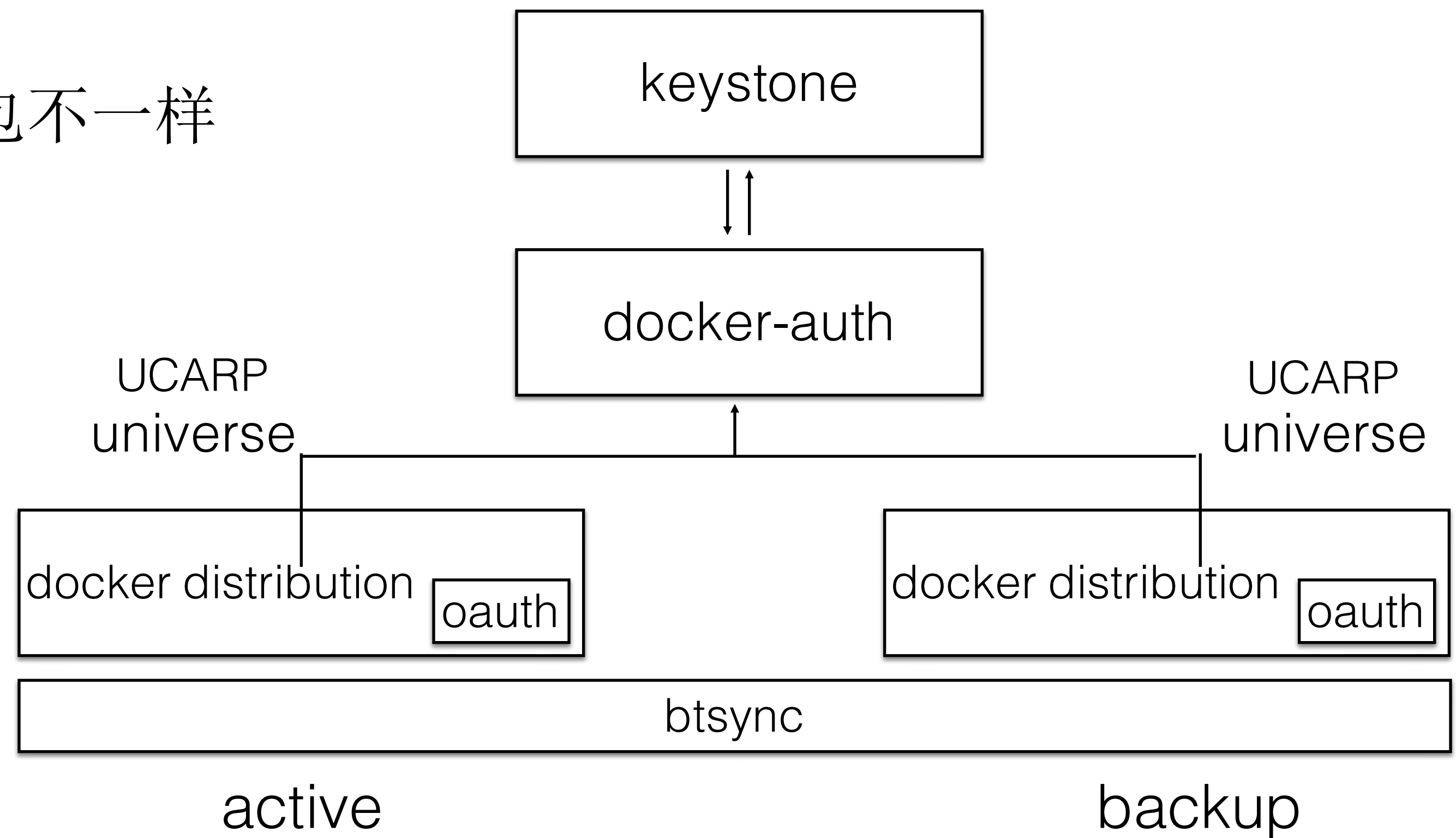
服务发现GAP

- 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

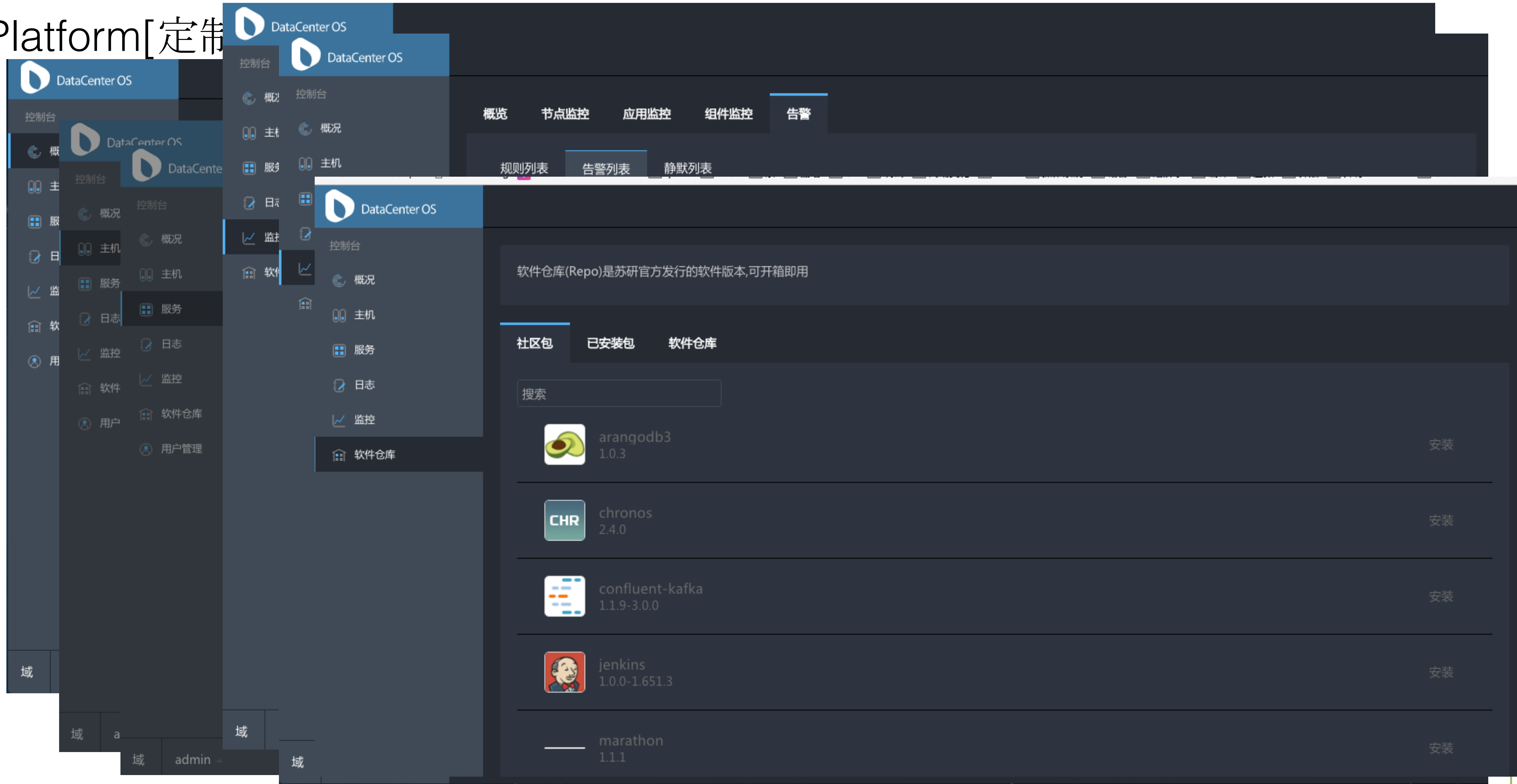


镜像仓库&&软件仓库





- 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



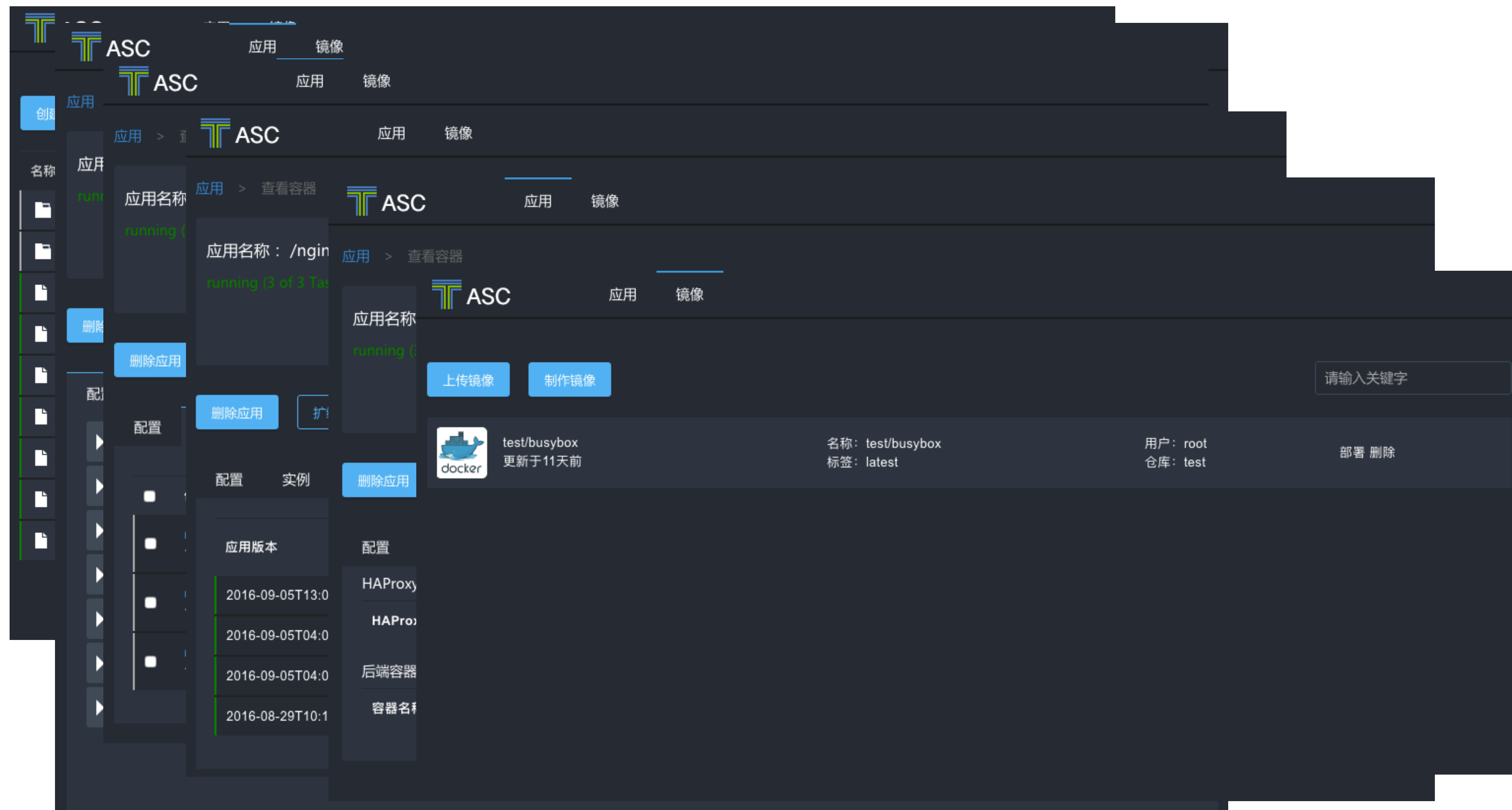
Platform[定制]



The screenshot displays the DataCenter OS Platform interface. The top navigation bar includes tabs for 概览 (Overview), 节点监控 (Node Monitoring), 应用监控 (Application Monitoring), 组件监控 (Component Monitoring), and 告警 (Alerts). The 告警 (Alerts) tab is currently active, showing sub-tabs for 规则列表 (Rule List), 告警列表 (Alert List), and 静默列表 (Silence List). The main content area is titled "软件仓库(Repo)是苏研官方发行的软件版本,可开箱即用" (Software Repository (Repo) is the software version issued by the official research institute, ready to use out of the box). Below this, there are tabs for 社区包 (Community Packages), 已安装包 (Installed Packages), and 软件仓库 (Software Repository). A search bar is present, and a list of software packages is shown, each with an icon, name, version, and an 安装 (Install) button.

| 社区包 | 已安装包 | 软件仓库 |
|---------------------------------------------------------------------------------------|------|--------------------------------------|
| 搜索 | | |
|  | | arangodb3 1.0.3 安装 |
|  | | chronos 2.4.0 安装 |
|  | | confluent-kafka 1.1.9-3.0.0 安装 |
|  | | jenkins 1.0.0-1.651.3 安装 |
| | | marathon 1.1.1 安装 |

TASC[定制化的marathon]



The screenshot displays the TASC web interface for managing applications. The interface is dark-themed and includes a sidebar on the left with navigation options like '应用' and '配置'. The main content area shows a list of applications with columns for '应用名称', '配置', and '实例'. One application is highlighted, showing details like '应用名称: /nginx', '名称: test/busybox', '用户: root', and '仓库: test'. There are buttons for '上传镜像', '制作镜像', and '删除应用'.

| 应用名称 | 配置 | 实例 |
|---------------------|----|----|
| running 1 | | |
| running 2 of 3 Test | | |
| running 3 | | |
| running 4 | | |
| running 5 | | |
| running 6 | | |
| running 7 | | |
| running 8 | | |
| running 9 | | |
| running 10 | | |
| running 11 | | |
| running 12 | | |
| running 13 | | |
| running 14 | | |
| running 15 | | |
| running 16 | | |
| running 17 | | |
| running 18 | | |
| running 19 | | |
| running 20 | | |
| running 21 | | |
| running 22 | | |
| running 23 | | |
| running 24 | | |
| running 25 | | |
| running 26 | | |
| running 27 | | |
| running 28 | | |
| running 29 | | |
| running 30 | | |
| running 31 | | |
| running 32 | | |
| running 33 | | |
| running 34 | | |
| running 35 | | |
| running 36 | | |
| running 37 | | |
| running 38 | | |
| running 39 | | |
| running 40 | | |
| running 41 | | |
| running 42 | | |
| running 43 | | |
| running 44 | | |
| running 45 | | |
| running 46 | | |
| running 47 | | |
| running 48 | | |
| running 49 | | |
| running 50 | | |

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

- 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