爱奇艺 Mesos 实践

杨成伟 yangchengwei@qiyi.com

@Mesos Meetup – 2015.11.29

爱奇艺

悦 享 品 质

自我介绍

- 开源软件用户、爱好者、贡献者
- Inside 英特尔
  - MeeGo, Tizen
  - Contributed to: DBus, systemd, udev …
  - DBus Committer http://people.freedesktop.org/~chengwei
- 爱奇艺
  - 负责弹性计算平台

关于爱奇艺

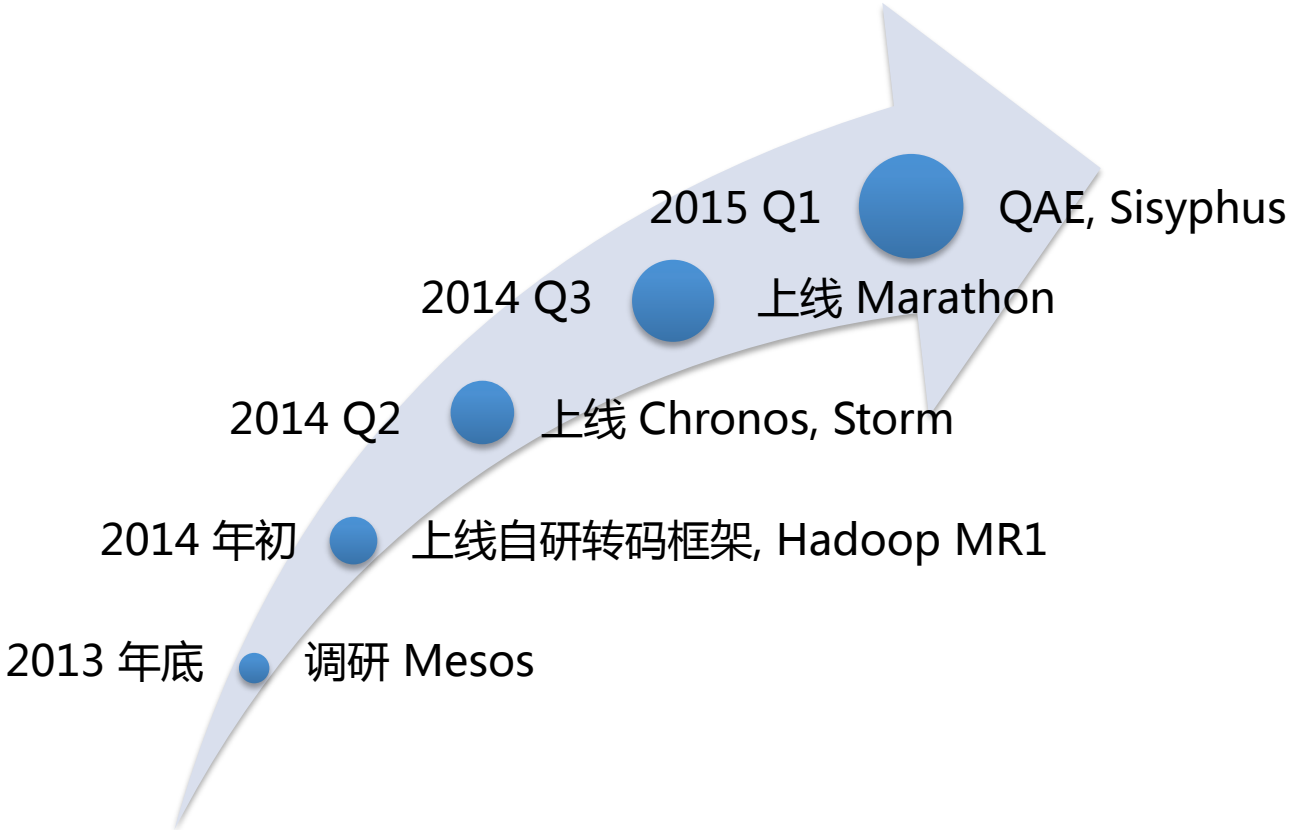
爱奇艺
悦 享 品 质

内容提要

- Mesos 在爱奇艺
- Mesos 集群建设
- 痛点

Mesos 在爱奇艺

线上 Mesos 服务



软件栈

Spark	0.8.x	1.0.1		
Hadoop	CDH 4.2.2			
Jenkins	1.600			
Storm	0.9.2	0.9.3		
Sisyphus	1.0			
Chronos	Internal release			
QAE	Internal release			
Marathon	0.5.x	0.8.0	0.9.2	
自研转码框架	Internal release			
Mesos	0.16.0	0.19.0	0.21.0	0.25.0
Docker	1.0	1.3.2	1.7.1	
CentOS	6.X	7.1		

现状

- 4 个 DC
- 超过 800 个计算结点
- 资源平均分配率 > 40%，峰值近 90%
- 框架：Storm, Marathon, Chronos, Sisyphus, QAE, Jenkins
- 周均启动 Docker 容器超过 200 万，峰值超过 350 万
- 并发峰值超过 4 千

Mesos 集群建设

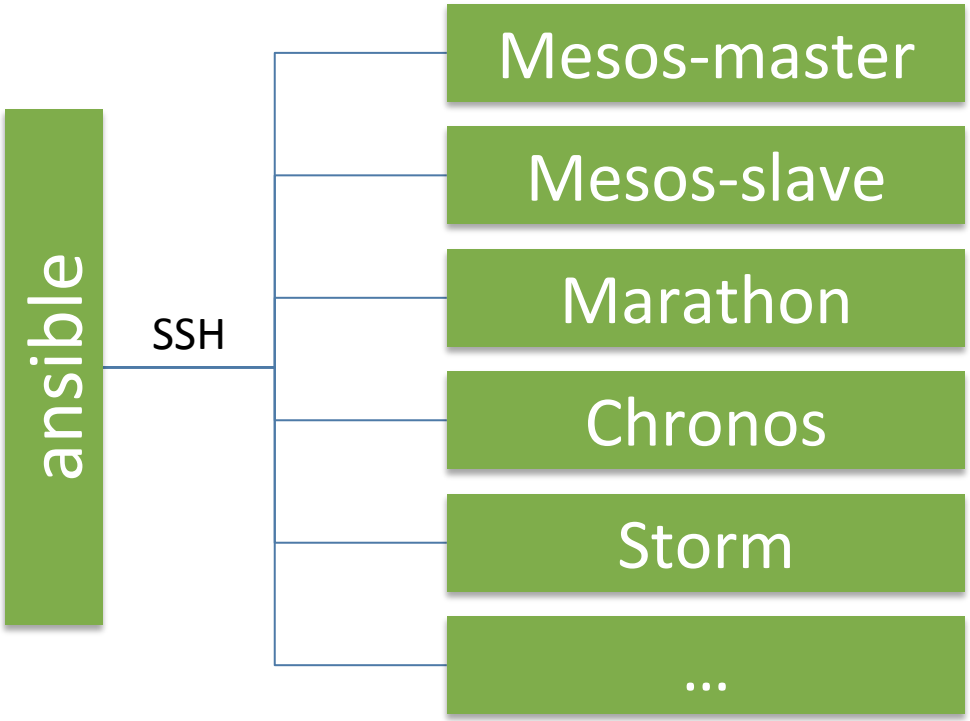
Mesos 集群建设

- 自动化部署 – Ansible
- 监控 – zabbix/cron
- 报警 – 统一报警平台
- HA(High Availability)
- 数据统计分析 – 棋布

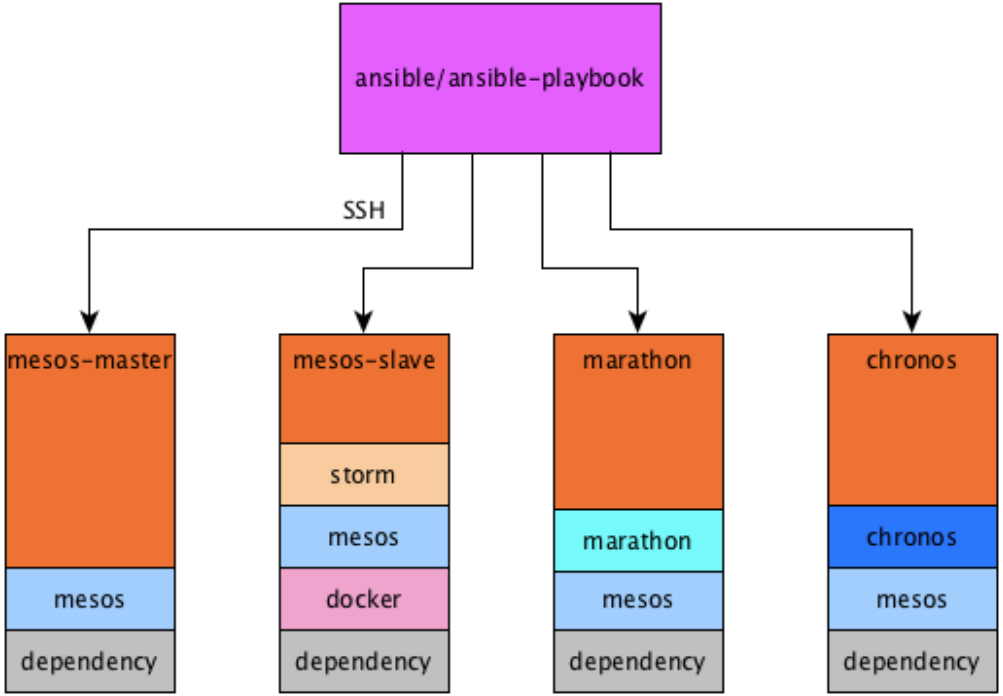
自动化部署 – Ansible

- Puppet
- 没有 agent，少一处安全风险
- 基于 SSH
- Ansible(open source) and ansible tower
- 刚被 Red Hat 收购

自动化部署 – Ansible



自动化部署 – Ansible



监控 – zabbix/cron

- 基础监控
  - 物理机、虚机
  - CPU
  - Mem
  - Disk
  - Network
  - 7x24 监控团队（mesos- 开头的主机直接重启）

监控 – zabbix/cron

- Mesos-master
  - 宕机（7x24 团队恢复）
  - Mesos 集群结点存活率
- Mesos-slave
  - 宕机（7x24 团队恢复）
  - 进程监控（和 Glusterfs 客户进程端口冲突，杀掉 Glusterfs）
  - Docker daemon 监控（daemon hang，重启）
  - 每天自动发送邮件到 7x24 团队，以免漏掉故障的机器

监控 – zabbix/cron

- Marathon
  - Marathon API 接口健康监控（重启）
- Chronos
  - Chronos 任务队列监控（在 chronos 并发上不去的时候，人工介入）
  - Chronos API 接口监控（API 可用性 99.99% 以上，每周启动超过 200 万容器，故障重启）
- Storm
  - 拓扑是否卡住，进程监控

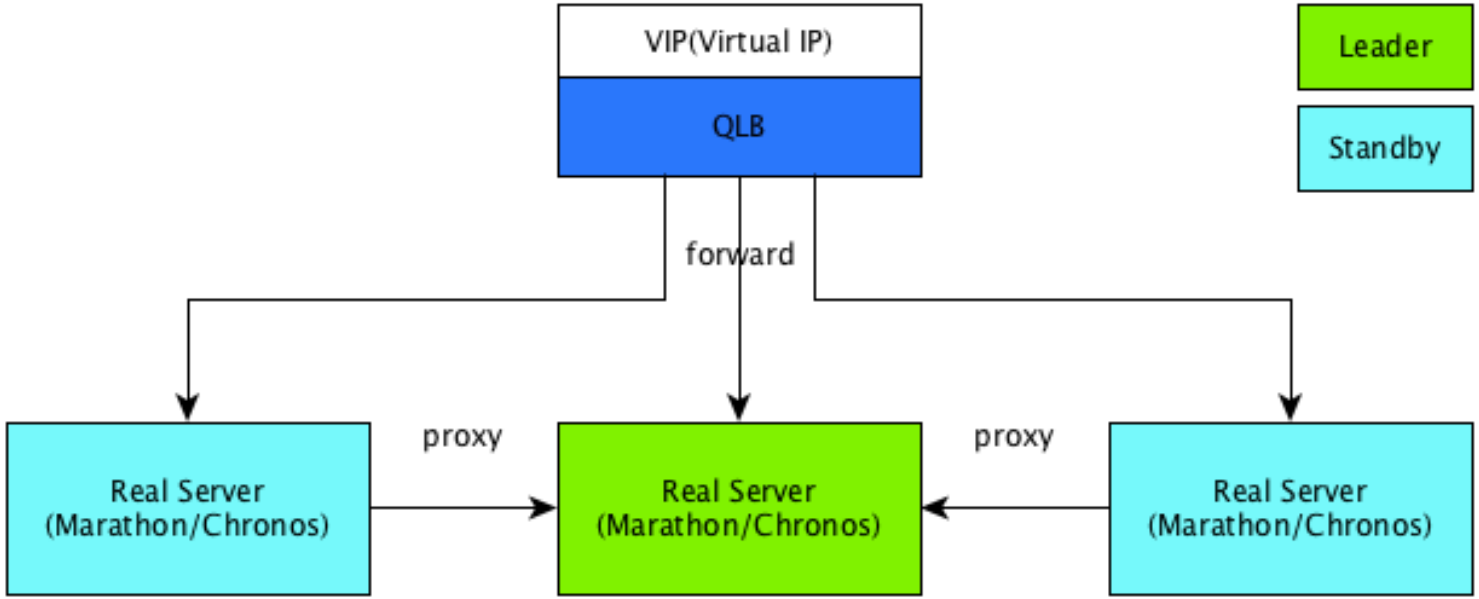
报警 – 统一报警平台

- API 接入
- 自助订阅
- 多种报警通道
  - IM
  - 短信
  - 邮件
- 可配置发送策略

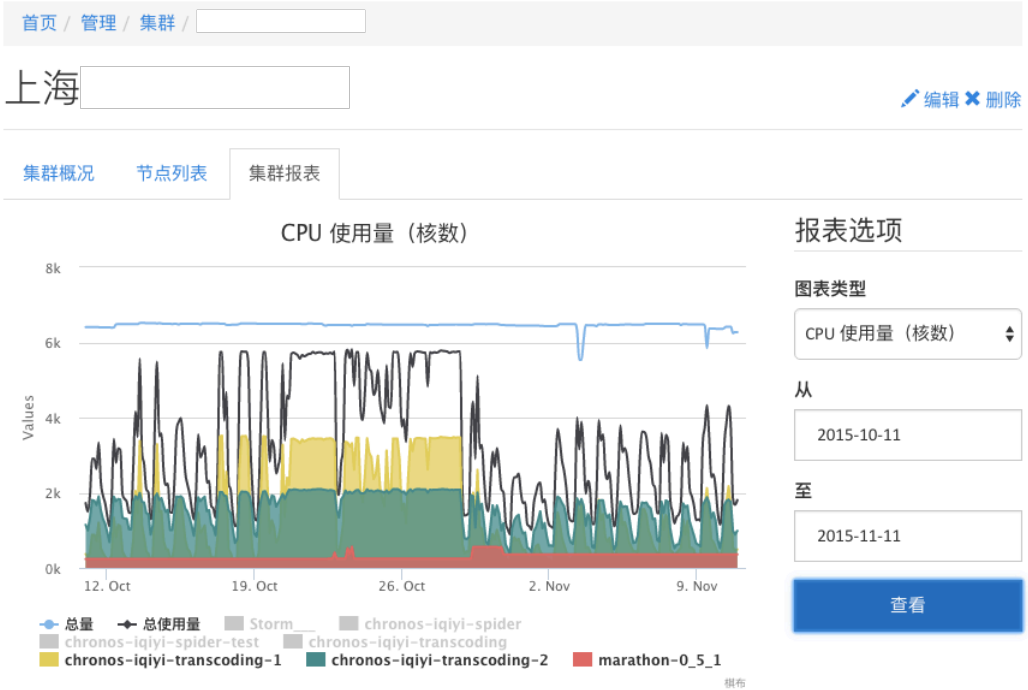
HA(High Availability)

- 3 个 Master
  - Mesos-master，容错能力：1 个 master 故障，计划增加到 5 台
  - Marathon，容错能力：2 个 marathon 故障
  - Chronos，容错能力：2 个 chronos 故障
- 服务发现
  - Mesos-master 使用 ZooKeeper
  - Marathon 及 Chronos 使用 QLB(iQIYI Load Balancer)

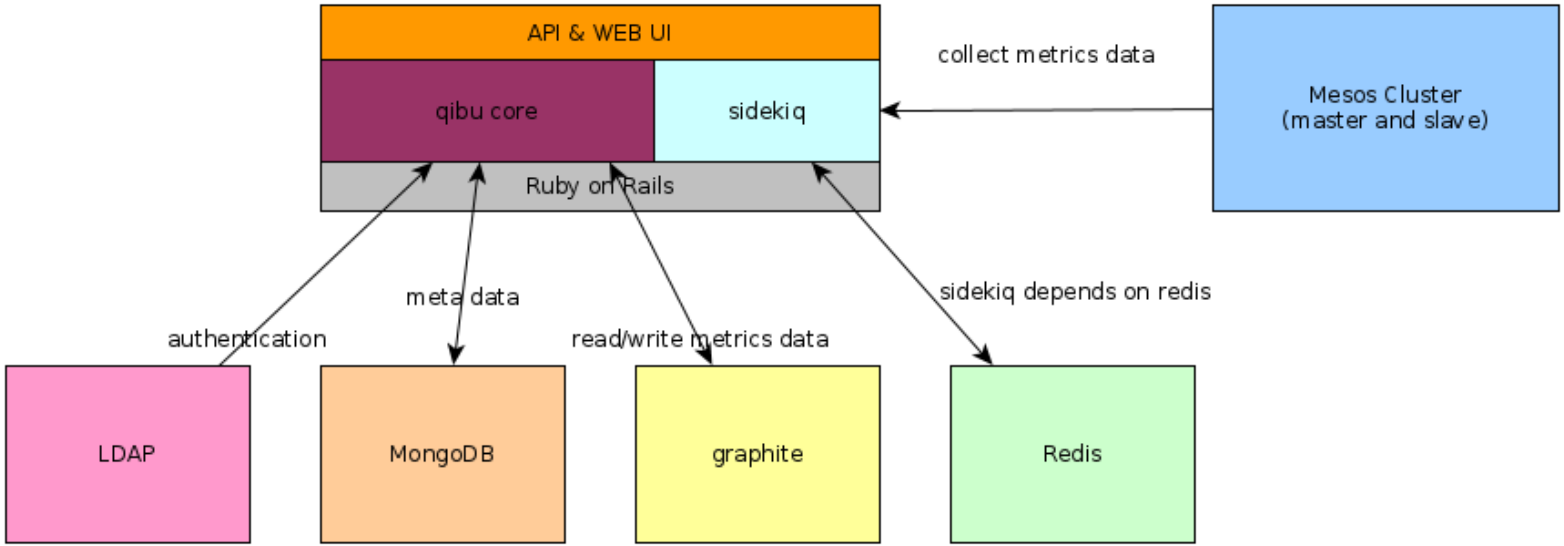
HA(High Availability)



数据统计分析 – 棋布



数据统计分析 – 棋布



痛点

静态资源预留

- 资源预留是节点的属性而非逻辑量
- 节点的资源预留不能动态更改
- 节点计算能力不同导致运维成本高

Docker 支持不好

- 各种 isolator 都只支持 mesos Containerizer
- Mesos 0.25.0 docker Containerizer 依然不支持 CFS CPU limit
- 在线离线任务几乎不可能混布
- Mesos 计划在自己的 mesos Containerizer 中兼容 Docker 镜像

Q & A

悦　　享　　品　　质