Geekbang. 极客邦科技 InfoQ

QCon 全球软件开发大会【北京站】2016

知乎 Docker 云平台架构和实践

知乎技术平台负责人 林晓峰

Intemational Software Development Conference

本“讲师”

- 前新浪高级工程师
- 高扩展网络协议栈 Fastsocket 开源作者 （Asplos 2016 Paper）
- 现知乎技术平台负责人

线索

- 初心
- 选择
- 初代
- 二代

初心

面向问题工程（Problem Oriented Engineering）

当时知乎的问题：

- 资源无法高效利用
- 虚拟机维护力不从心
- 业务扩容效率低

对策

两个重要决策：

- 确认使用容器方案
- 项目命名 Bay

现状

- 业务完全运行在 Bay 上
- 容器计算资源由平台统一管理
- 与 CI & CD 完整集成 （Build once，Run anywhere）
- 容器数量千级规模（5K）
- 秒级的业务透明伸缩（10 -> 100 35s）
- 工程师 1.5 人

选择

Candidates（2015 年 5 月）：

- Mesos
- Kubernetes
- Docker Swarm
- Self-made

选择

Factor：

- 可以快速落地（解决问题速度）
- 可持续维护和扩展（不该引入更多的问题）

Strategy：

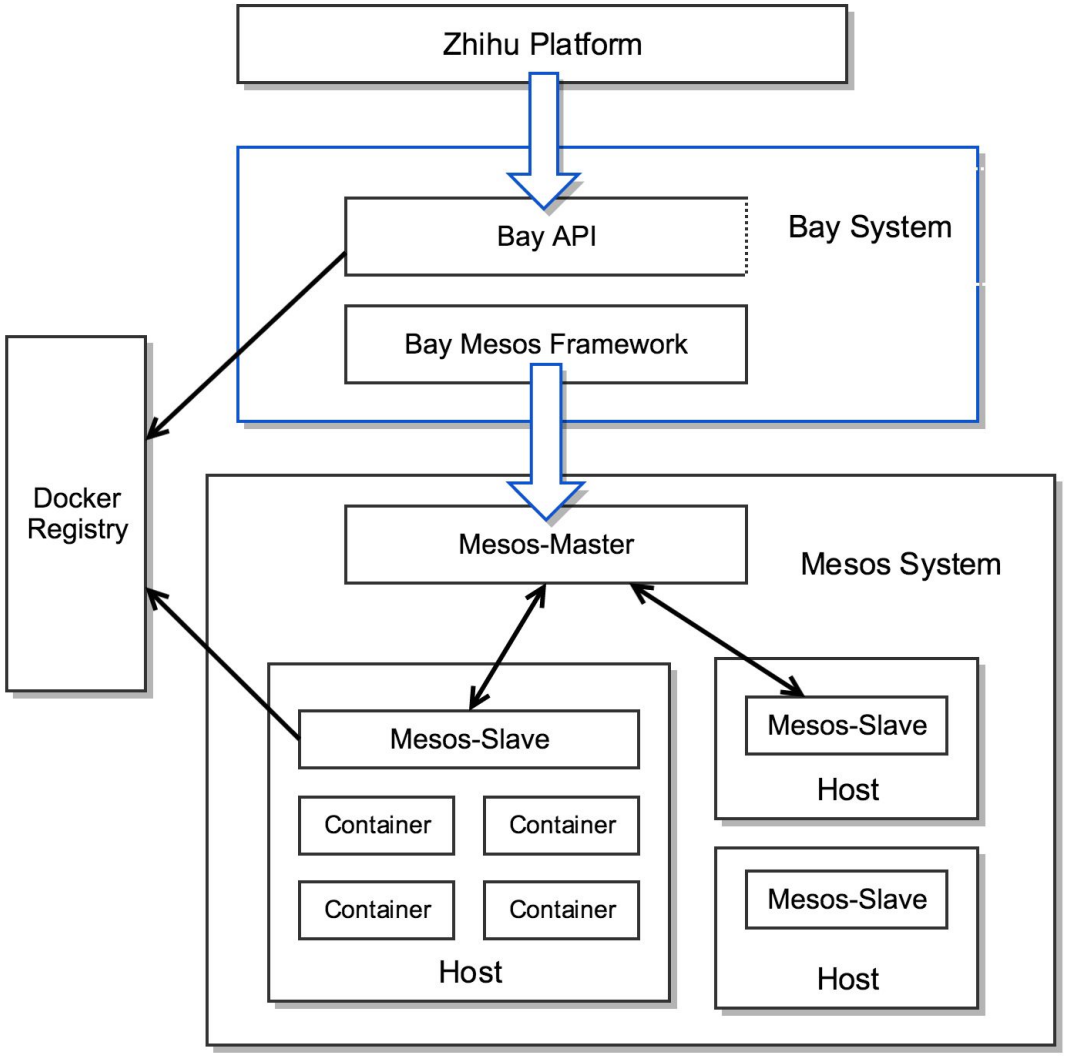
- 方案成熟情况
- 方案可控程度
- 现有设施整合难易度

选择

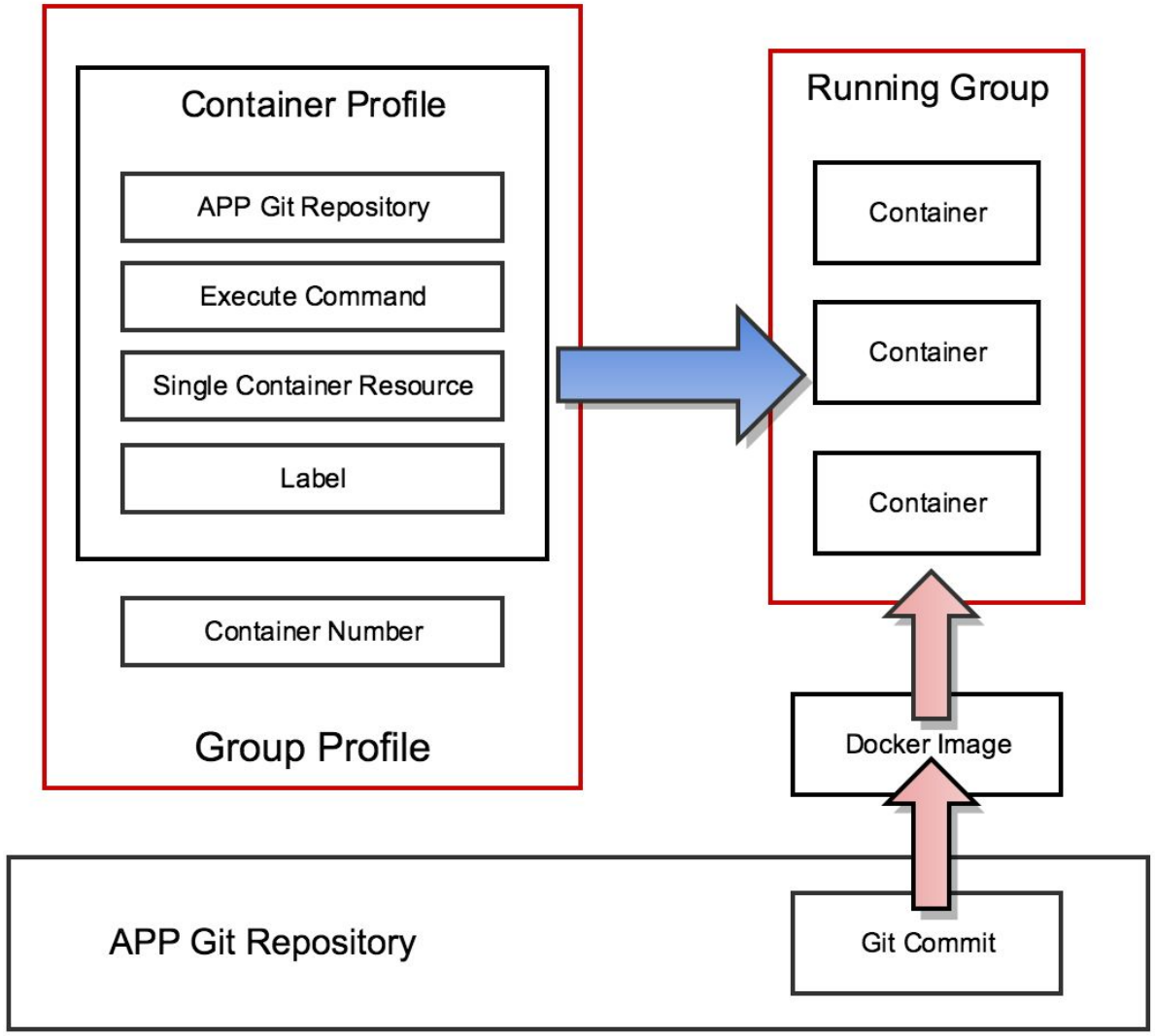
Mesos & Self-made Framework

- 更广泛的容器类型支持
- 更丰富的社区 Framework
- 数据中心资源统一管理和调度的布局

初代


Zhihu Platform
Bay API
Bay System
Bay Mesos Framework
Docker Registry
Mesos-Master
Mesos System
Mesos-Slave
Container
Container
Container
Container
Host
Mesos-Slave
Host
Mesos-Slave
Host

GROUP


Container Profile
APP Git Repository
Execute Command
Single Container Resource
Label
Container Number
Group Profile
Running Group
Container
Container
Container
Docker Image
APP Git Repository
Git Commit

API

- 描述 Group
- 制作 APP 容器镜像
- 部署镜像生成 Running Group
- 伸缩 Group 容器

Pets or Cattle

Cattle：保持健康容器的数量

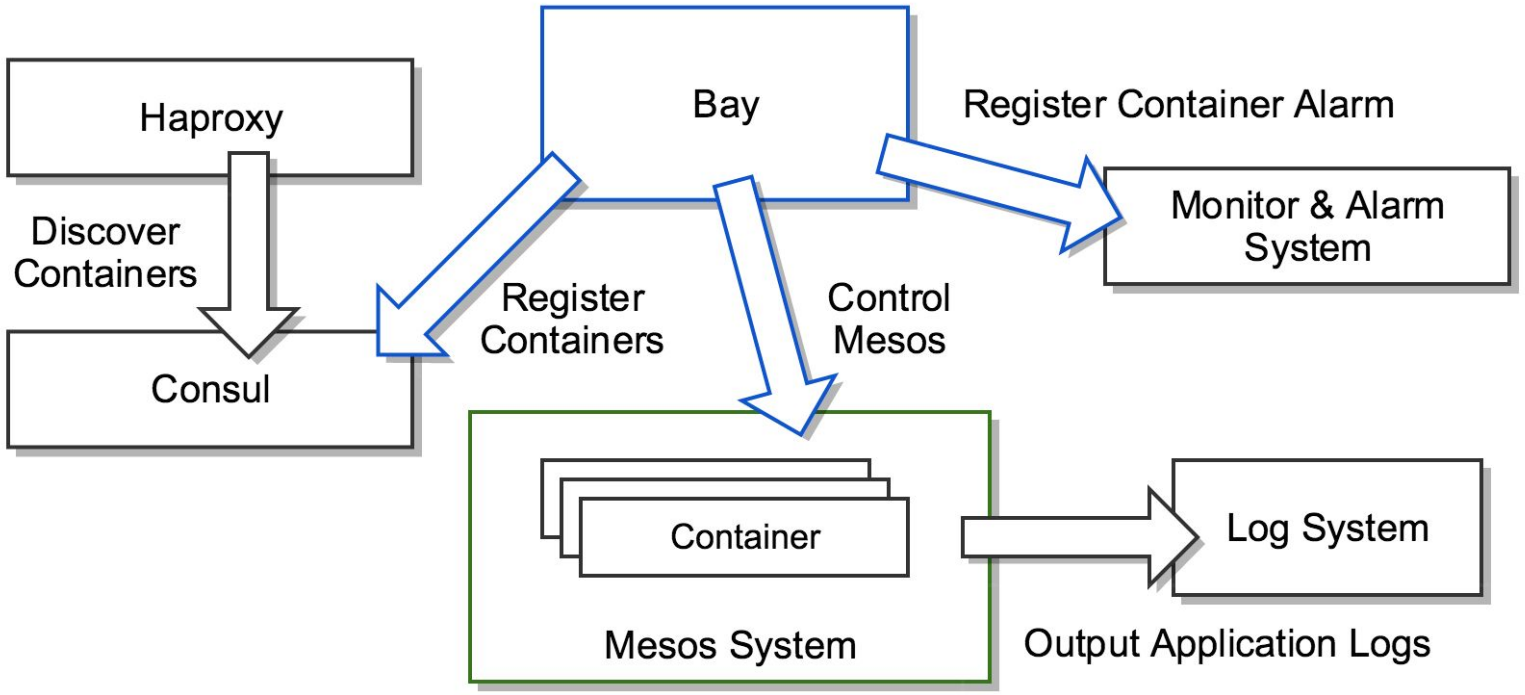
Docker Network

- Network: Bridge
- NAT is not bad
- Iptables 有些坑

系统整合

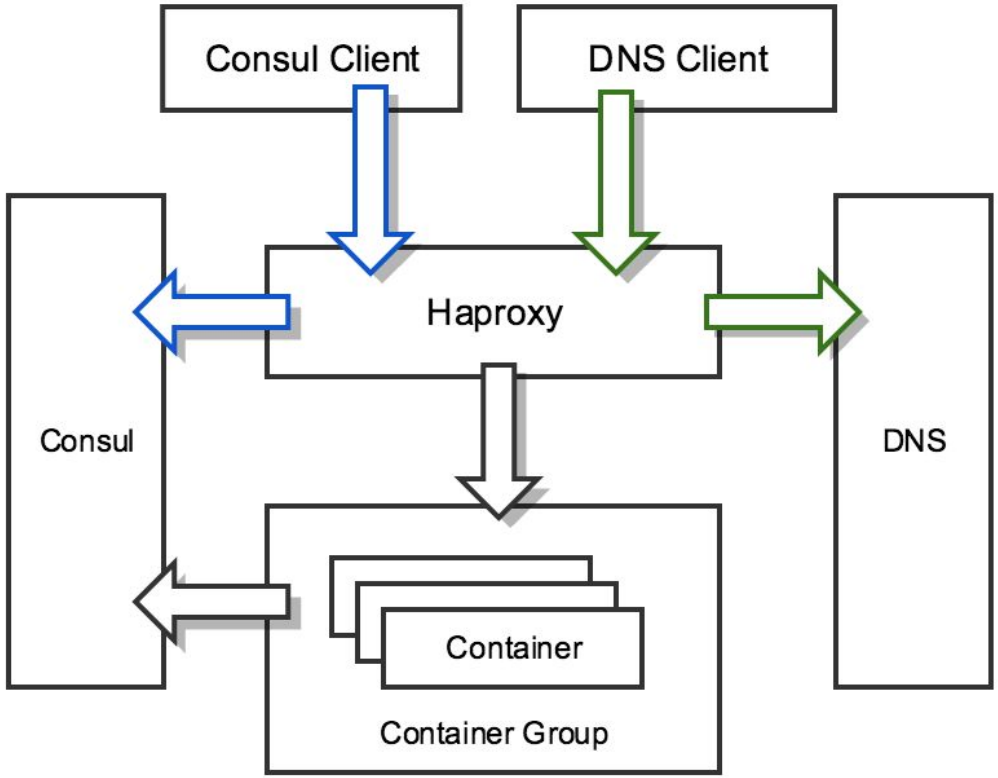
- 服务注册发现
- 负载均衡
- 日志收集
- 监控报警

系统整合



- 最大化利用现有设施
- 不断尝试简化技术方案

服务发现



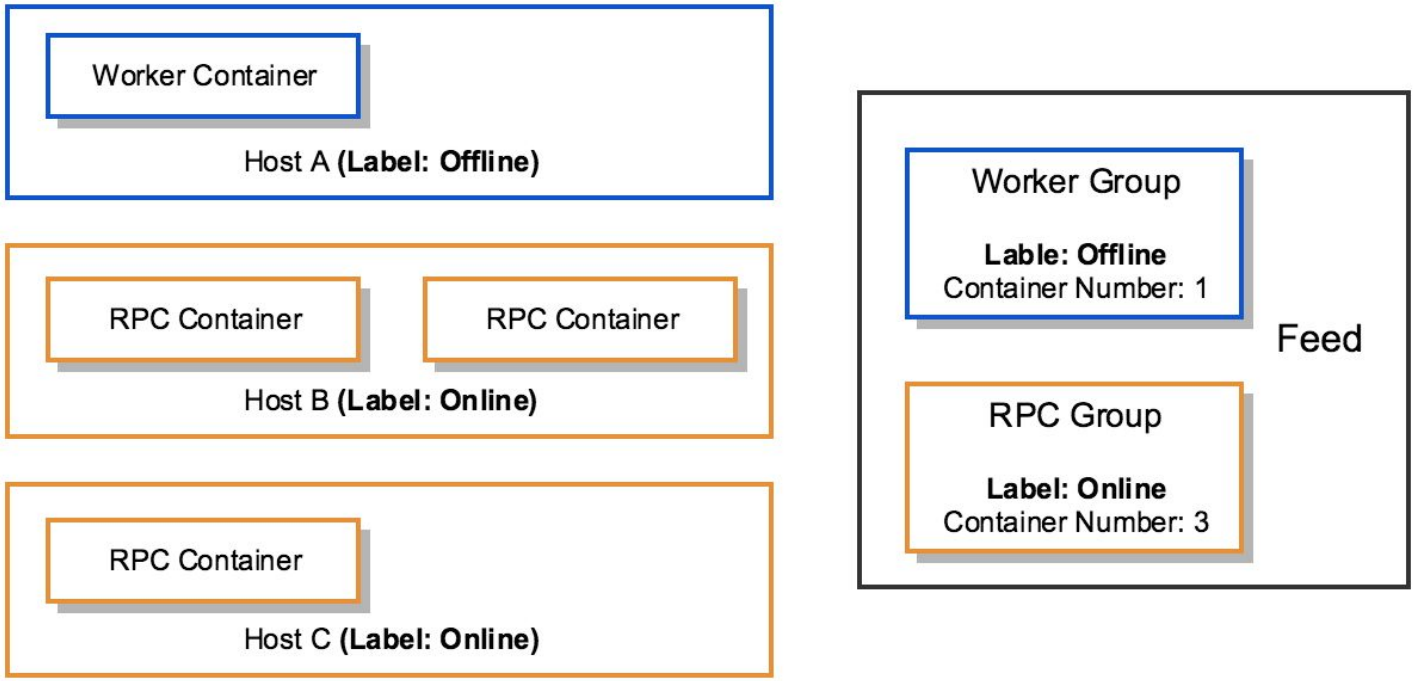
兼容传统 DNS Client

集群管理

- 集群调度细化
- 集群主机管理

Bay 通过 Label 功能来实现

Label Demo

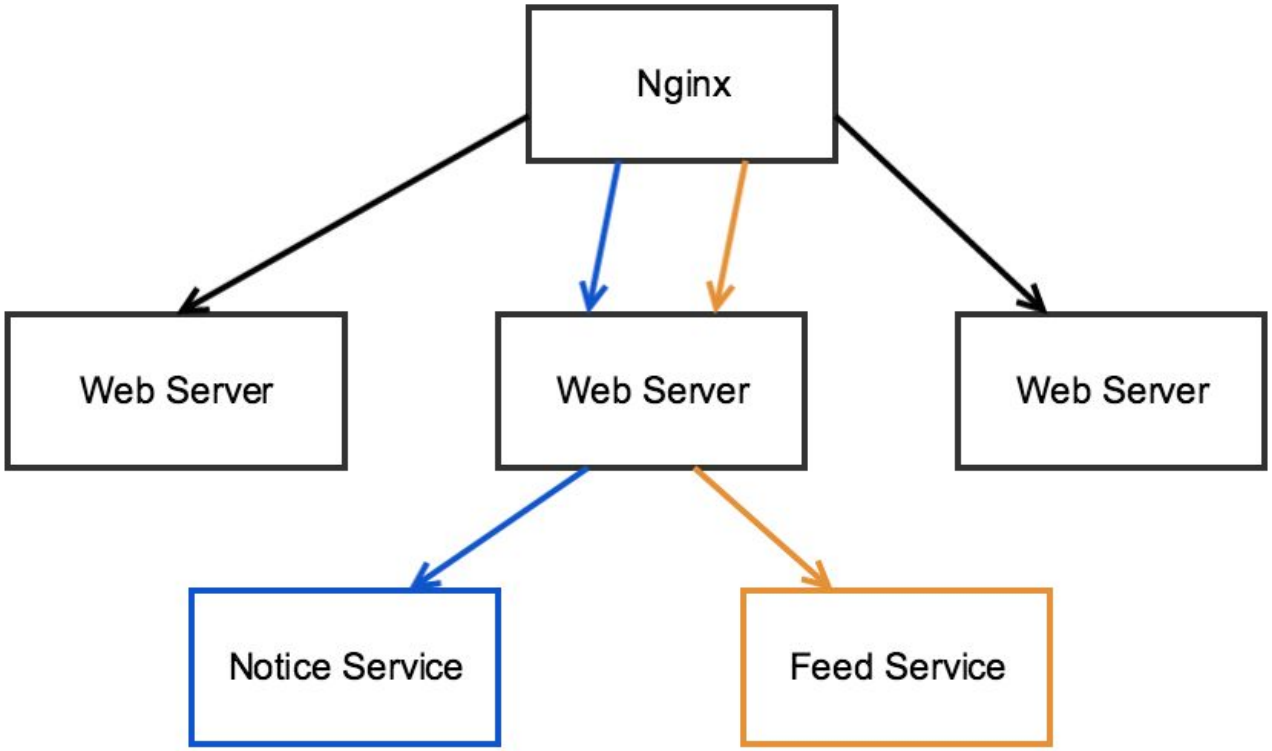


依据 Label 调度容器到相应的服务器上

First Client

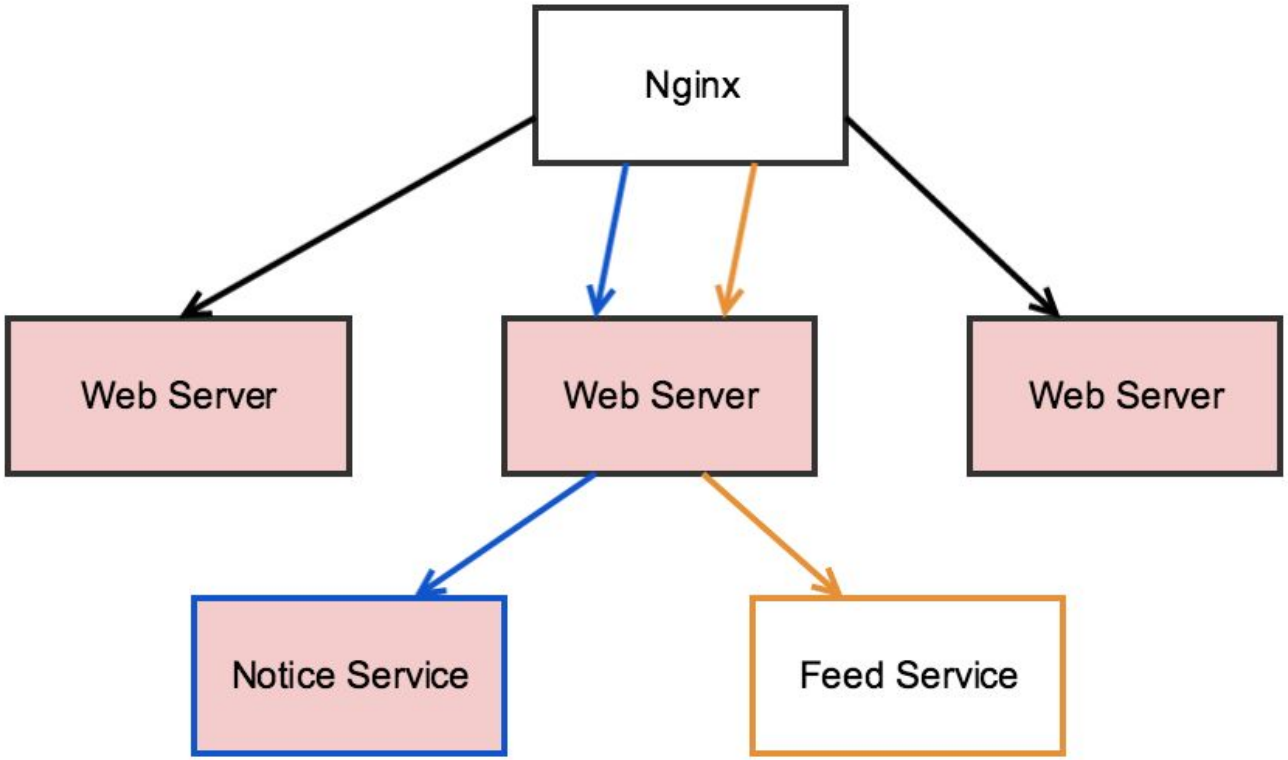
- 移动后端 Web Server
- 驱动的问题：如何做故障隔离

First Client



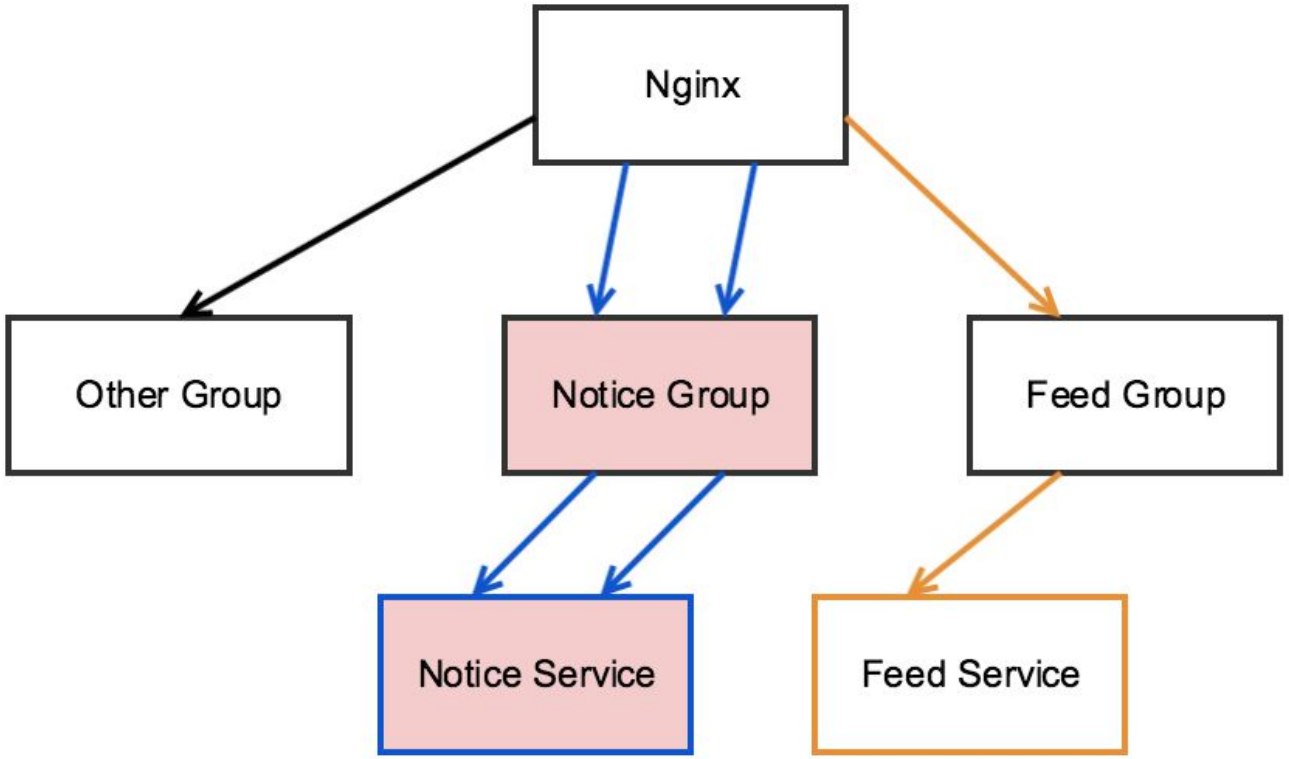
不同的请求（Notice & Feed）落在相同的 Web Server

Notice 故障



Notice 服务故障，影响所有 Web Server

Bay 方案



Notice 服务故障被隔离在 Notice Group

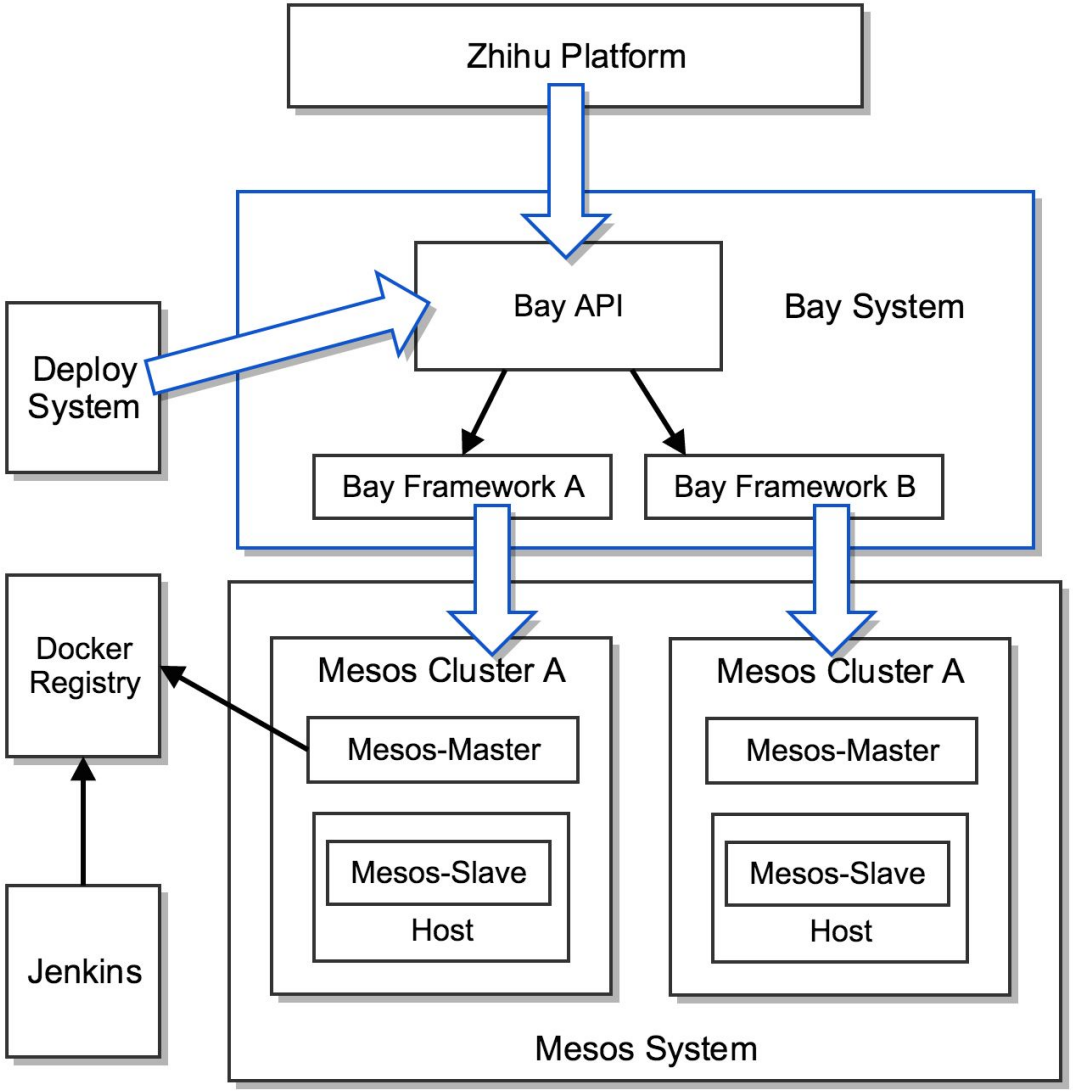
思考

- 面向问题工程
- 树立信心
- 简单 & 合适
- 持续迭代

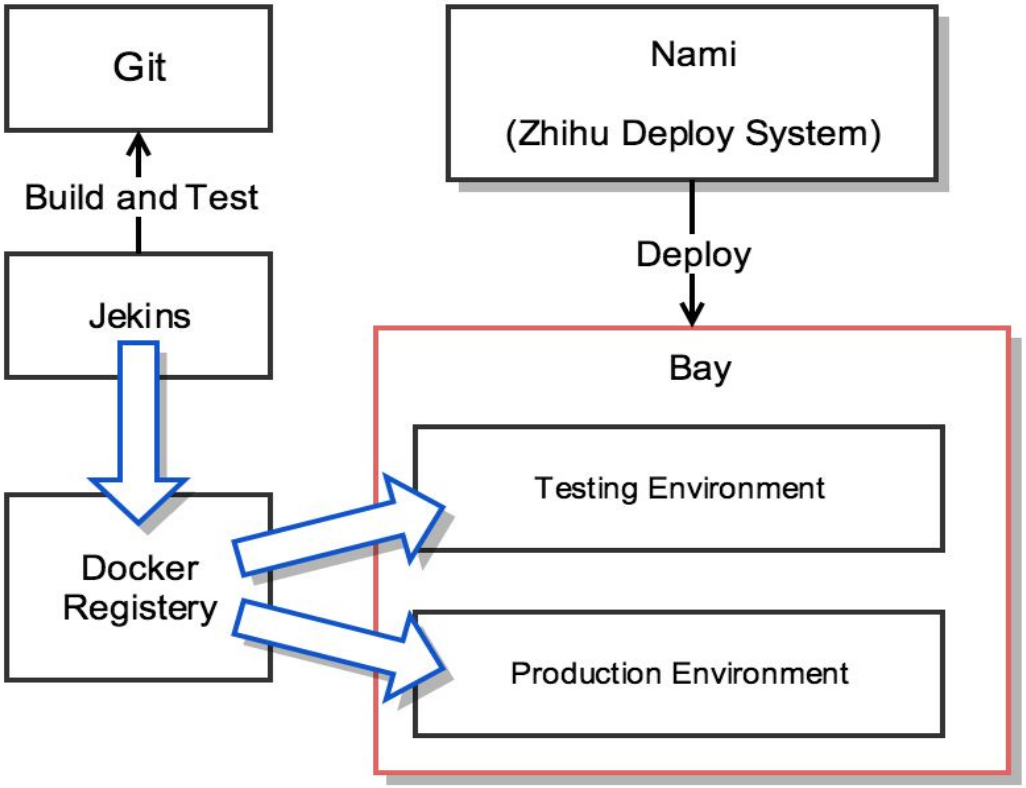
二代

- CI & CD 集成
- 支持多集群

Overview


Zhihu Platform
Bay API
Bay System
Deploy System
Bay Framework A
Bay Framework B
Docker Registry
Mesos Cluster A
Mesos-Master
Mesos-Slave
Host
Mesos Cluster A
Mesos-Master
Mesos-Slave
Host
Jenkins
Mesos System

CI & CD



Build Once， Run Anywhere

多集群

- 实现 Bay Framework 灰度发布
- 实现 Mesos 集群的横向扩展

More Feature

- 支持 Group 在集群间迁移
- 支持 Group 自动 scale
- 支持金丝雀发布
- 多种部署策略

Group 迁移

初衷：多集群的自然需求

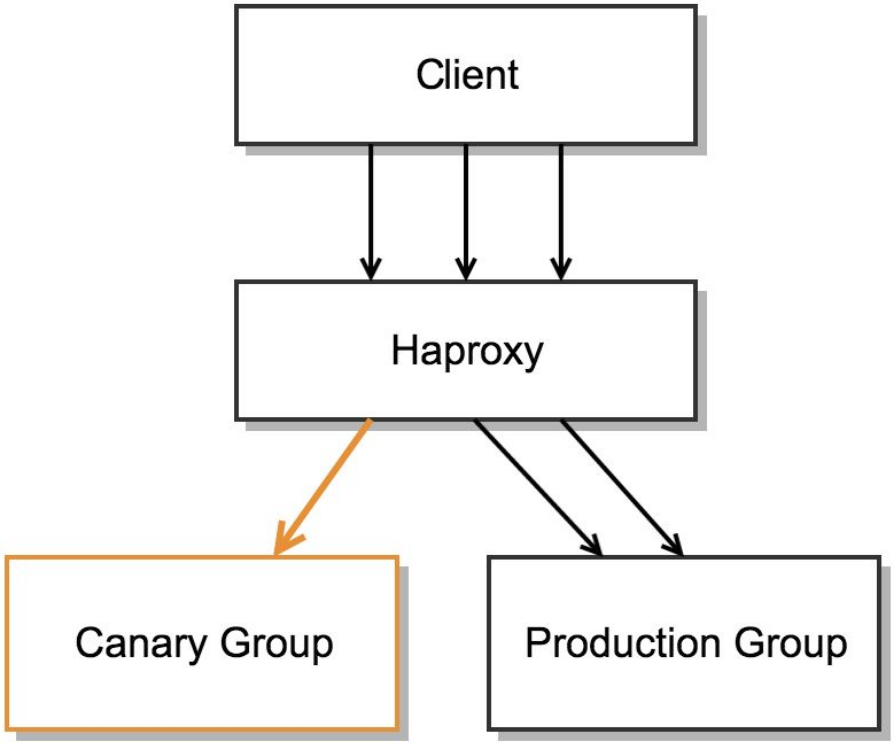
- 业务无缝迁移到 Staging 集群
- 多集群间负载调整

自动 Scale

初衷：突发响应 & 资源高效利用

- 根据 CPU 指标调整容器数量
- 快伸慢缩
- Max & Min Hard Limit
- 支持自定义指标

金丝雀发布



多 Group 注册相同服务

时间线

- 2015 年 5 月 容器弹性计算平台 Bay 立项
- 2015 年 6 月 Bay 测试上线
- 2015 年 9 月 知乎移动 Web Server 迁移 Bay
- 2016 年 2 月 服务全面迁移 Bay
- 2016 年 4 月 知乎主站 Web Server 迁移 Bay

展望

- 将基础设施纳入 Bay 进行统一管理
- 将数据业务集群纳入 Bay 进行统一管理

Geekbang 极客邦科技 InfoQ

THANKS!

小广告

jobs@zhihu.com