GOPS2016 Shenzhen

GOPS

2016 全球运维大会 深圳站

会议时间：3月25日-3月26日

会议地点：深圳 • 南山区 圣淘沙酒店(翡翠店)

主办单位：开放运维联盟 OOPSA Open OPS Alliance　高效运维社区 GreatOPS Community

指导单位：数据中心联盟 Data Center Alliance

协办单位：中国新一代IT产业推进联盟

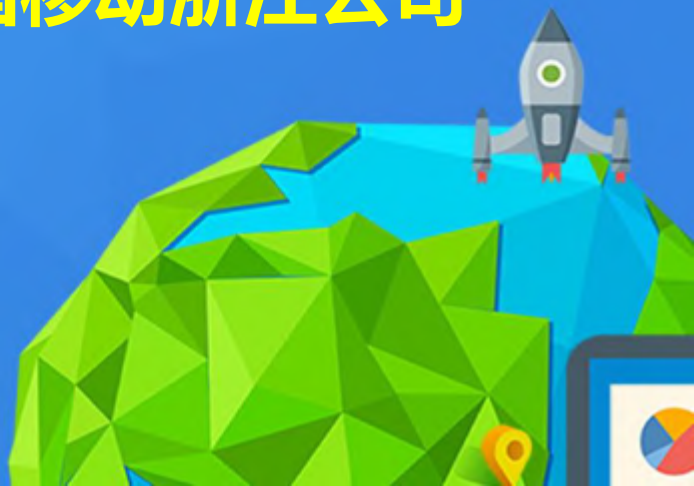
GOPS 2016 Shenzhen

GOPS

2016 全球运维大会 深圳站

中国移动浙江公司Mesos生产实践

钟储建，中国移动浙江公司

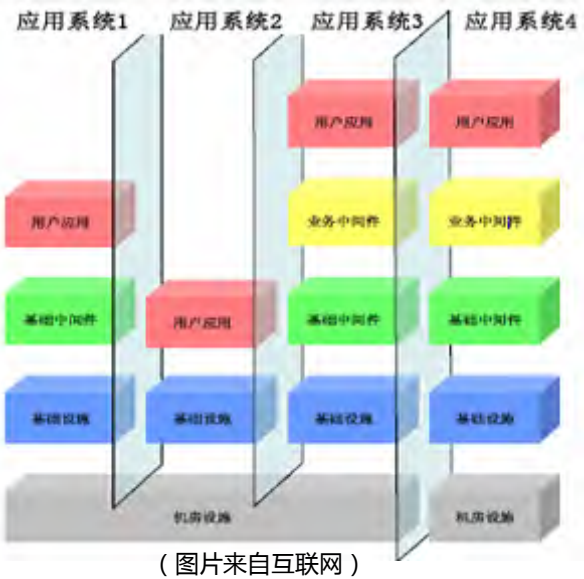


内容

- **为什么使用MESOS**
- 基于MESOS的DCOS实现
- 实践经验

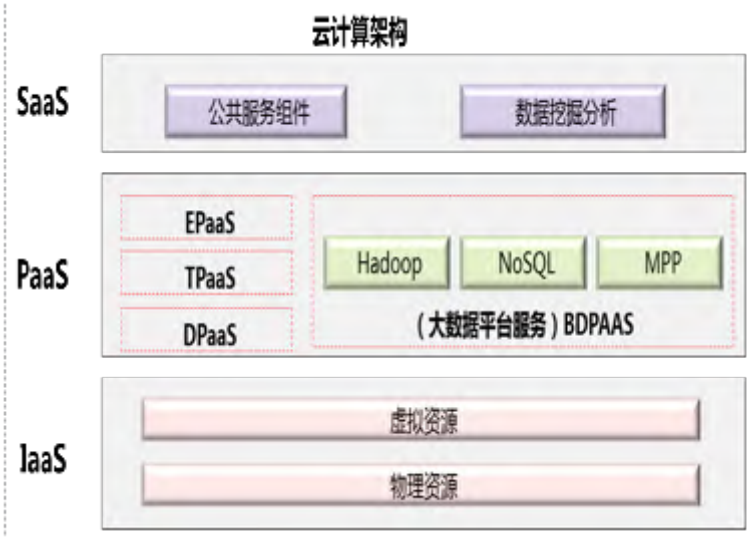
云计算驱动企业IT架构演进

企业IT架构演进



（图片来自互联网）

“烟囱”式IT系统架构

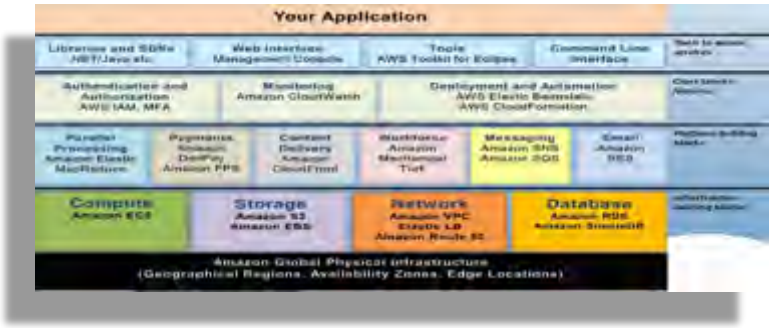


云化IT架构

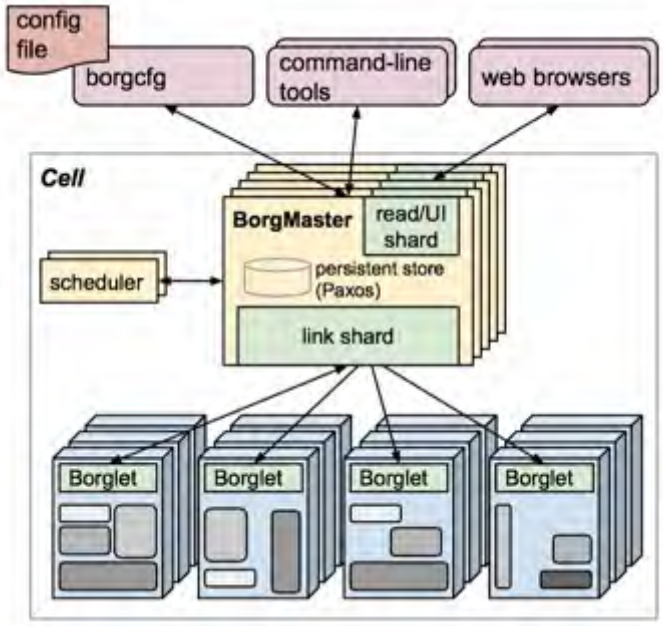
打破竖井、应用和平台解耦
打破供应商绑定
加强企业自身核心能力掌控
敏捷建设、聚焦支撑业务
统一管理建设运营，提升运维效率、提升资源利用率，降低TCO

典型的云计算平台

亚马逊的AWS服务：以虚拟化为基础，提供IaaS、PaaS及跨服务功能层和服务访问工具。

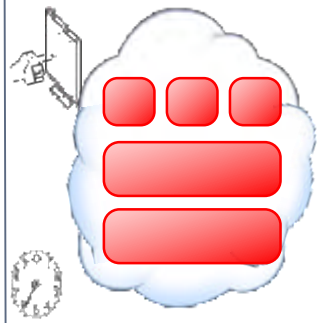
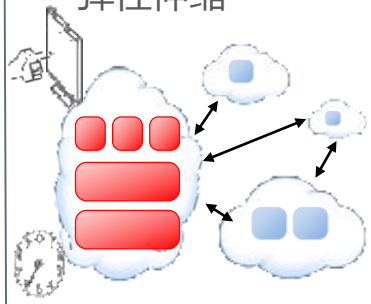


Google云计算平台：基于操作系统层面轻量级隔离技术的数据中心操作系统（Borg/Omega），在数以万记的PC服务器上进行集中的资源分配和调度。



（图片来自互联网）

浙江移动云化的阶段

传统孤岛	标准化	IaaS 资源池化	PaaS和应用 资源池化	DCOS化
▪ 对数据中心内部整体目标架构**没有统一的**规划设计	▪ **标准化的**硬件和软件体系 ▪ 业务基础架构建设以月为单位	▪ 通过虚拟化实现共享的**基础架构** ▪ 业务基础架构建设以周为单位 ▪ 实现**虚拟机级**弹性伸缩	▪ 通过服务化实现共享的**平台架构** ▪ 业务基础架构建设以日为单位 ▪ 实现**集群级**弹性伸缩	▪ 通过核心云构件实现**进程级资源共享** ▪ 业务基础架构建设以分钟为单位 ▪ 实现**数据中心级**弹性伸缩
				
孤岛	X86化	虚拟化	服务化	智能化
	简化	高效	灵活	动态复用

存在的问题

基于IaaS层的云化先天不足：

- **静态部署**
- **只能大切小不能小聚大**
- **不能维持应用环境的自动化封装**
- **应用的快速部署开通受到极大制约**
- **传统虚拟化只能实现虚机级弹性伸缩，效果极其有限**
- **资源利用率低**

DCOS – 第三代PaaS的核心

PaaS 1.0

- Google App Engine、SAE等
- 早期的PaaS概念，提供软件开发平台和运行环境

PaaS 2.0

- Cloud Foundry、OpenShift等
- 允许用户运行自己的PaaS，将平台进行标准化、服务化。

PaaS 3.0

- 以分布式集群资源调度（Mesos、Yarn）为基础，结合容器（Docker）技术构建
- 支持多种计算框架，具备敏捷开发、快速部署和弹性伸缩特性

- 第三代PaaS的核心是DCOS（DataCenter Operating System）

数据中心操作系统DCOS

数据中心操作系统（DataCenter Operating System，简称**DCOS**）是为整个数据中心提供分布式调度与协调功能，实现数据中心级弹性伸缩能力的软件堆栈，它将所有数据中心的资源当做一台大型计算机来调度，可以视作这个大型主机的操作系统。

	Linux OS	DCOS
Resource Management	Linux Kernel	Mesos
Process Management	Linux Kernel	Docker
Job Scheduling	init.d, cron	Marathon, Chronos
Inter-Process Communication	Pipe, Socket	RabbitMQ
File System	ext4	HDFS, Ceph

（以Mesos为例）

（图片来自互联网）

DCOS的特征

数据中心操作系统终极目标是提供一个通用的标准化运维系统高效率可靠安全地管理数据中心，同时简化应用程序的开发、部署难度，协调各类资源，确保各类资源随着应用的需求动态调度

- **数据中心级的弹性伸缩**
- **自动化调度、故障自愈**
- **细粒度的资源分配**
- **高资源利用率**
- **敏捷开发、快速部署**

DCOS解决方案

- **典型案例**

  Google：**Borg/Omega**

  Twitter、Apple、Netflix：**Mesos**

- **解决方案**

Mesos ： Mesos由加州大学伯克利分校AMPLab开发，后在Twitter广泛使用，成熟度高。Mesosphere公司DCOS，以Mesos为核心，支持多领域的分布式集群调度框架：Marathon、Chronos和Hadoop、Spark等的集群调度框架，实现系统的资源弹性调度。

Apache Hadoop YARN：一种新的 Hadoop 资源管理器，它是一个通用资源管理系统，可为上层应用提供统一的资源管理和调度 。

Kubernetes：是Google多年大规模容器管理技术的开源版本，面世以来就受到各大巨头及初创公司的青睐，社区活跃。

Docker Machine + Compose + Swarm：Docker公司的容器编排工具。

传统PaaS产品：CloudFoundry/OpenShift等传统PaaS解决方案。

Why Mesos

根据对适合构建DCOS的各种技术架构的评估，选择以Mesos为基础的方案。优点是成熟度高、两级调度框架、适合多种应用场景、混合部署、应用与平台耦合度低

	Mesos	Yarn	Kubernetes	Docker Swarm	CF/OpenShift
调度级别	二级调度（Dominant Resource Fairness）	二级调度（FIFO，Capacity Scheduler，Fair Scheduler）	二级调度（基于Predicates和Priorities两阶段算法）	一级调度（提供Strategy 和Filter两种调度策略）	CF一级调度（基于Highest-scoring调度策略）/OpenShift使用Kubernetes
生态活跃	活跃	活跃	非常活跃	活跃	一般
适用场景	通用性高，混合场景	大数据生态场景	目前较单一	较单一	较单一
成熟度	高	高	中	低	中
应用与平台耦合度	低	中	中	低	高
应用案例分析	Twitter、Apple、Airbnb、Yelp、Netflix、ebay、Verizon	Hadoop生态圈应用	目前快速发展中，生产环境应用较少	很少	较少，PaaS整体解决方案，应用与平台的耦合度较高

中国移动浙江公司DCOS建设历程

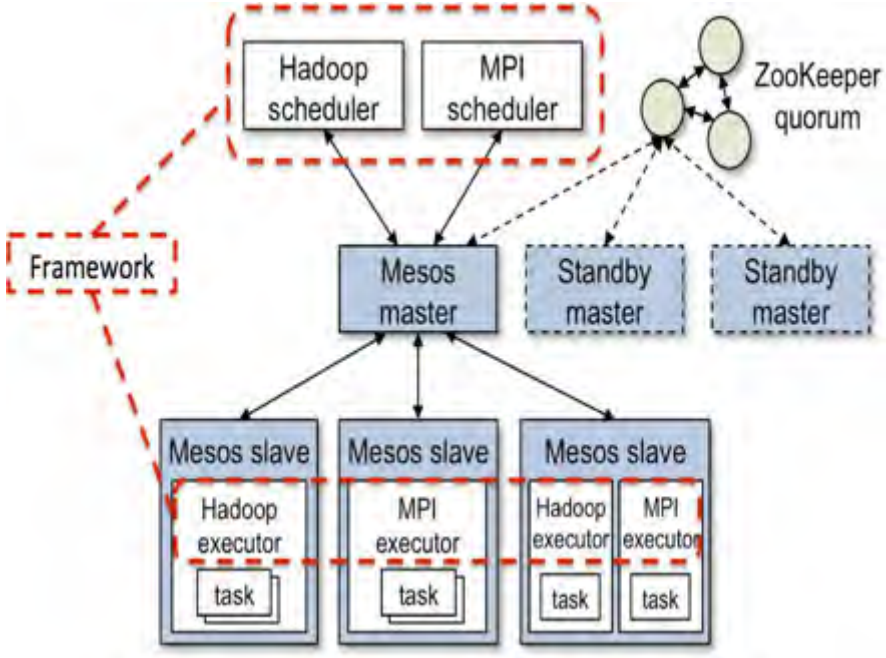
1. 2014年3月开始关注Docker容器化技术，2014年8月启动Docker应用的技术验证
2. 2014年11月将核心系统CRM的一个完整集群迁移到容器运行，Docker正式投入生产
3. 2015年8月，提出数据中心操作系统的设想，建设DCOS验证网，使用Mesos+Marathon+Docker方案
4. 2015年11月4日中国移动浙江公司DCOS验证网上线，11月11日支撑手机营业厅“双11”活动
5. 2015年12月10上线CRM应用

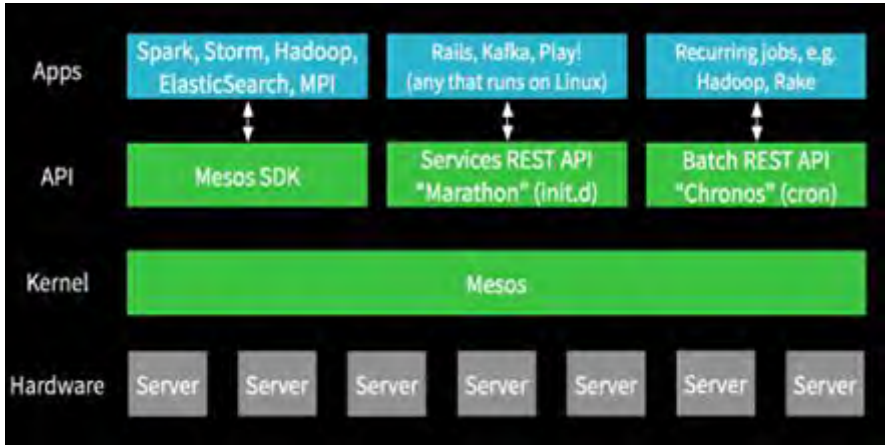
内容

- 为什么使用MESOS
- **基于MESOS的DCOS实现**
- 实践经验

关键技术选型 - 资源调度

Mesos

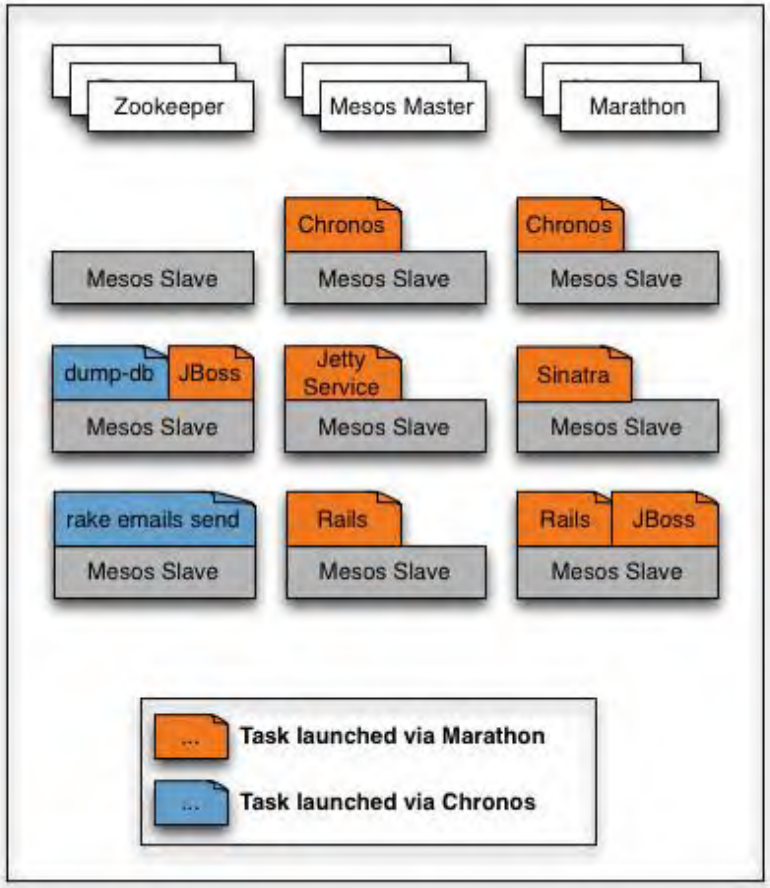




（图片来自互联网）

关键技术选型 - 任务调度

Marathon



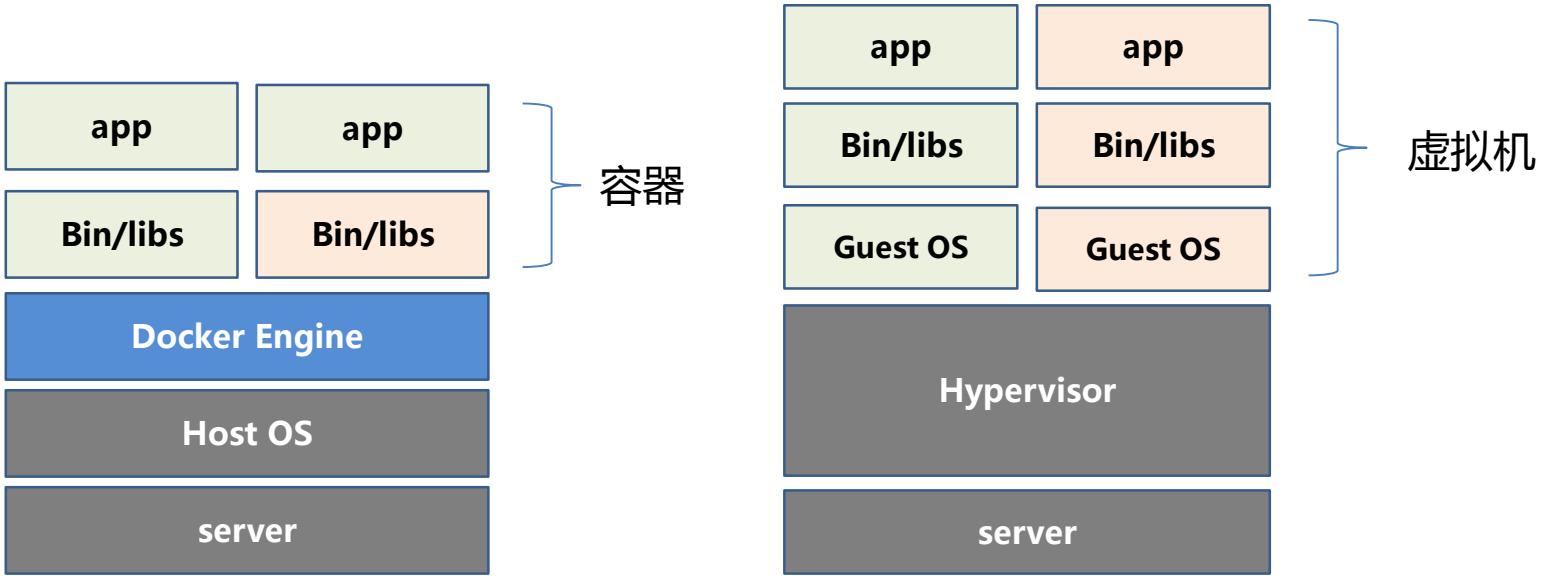
Mesos仅负责分布式集群资源分配

Marathon做任务调度，故障转移

（图片来自互联网）

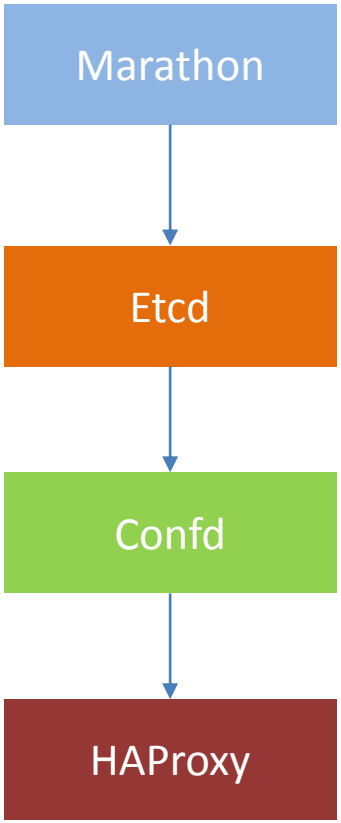
关键技术选型 – 应用封装

Docker



（图片来自互联网）

关键技术选型 - 服务发现与注册

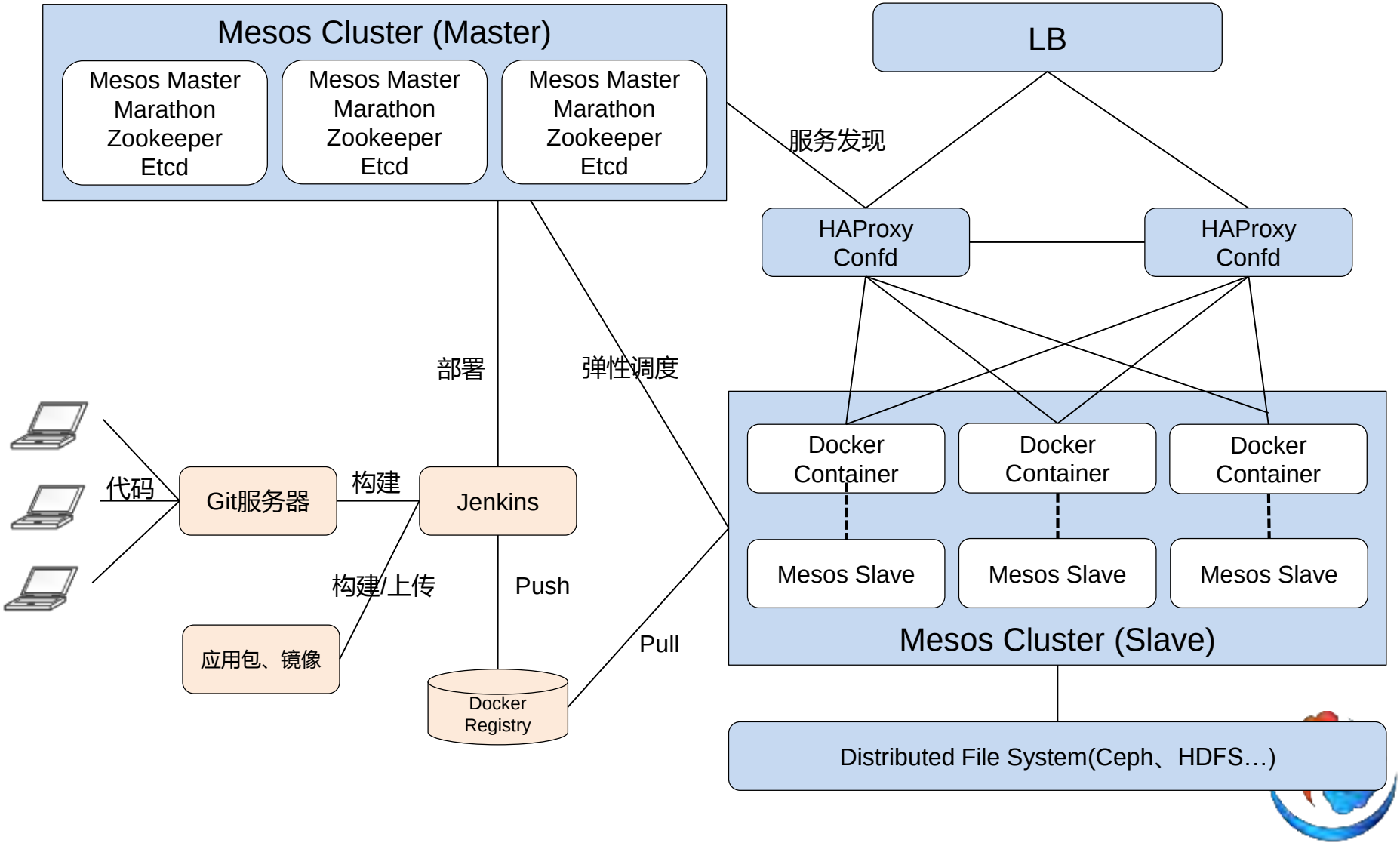


HAProxy业务负载的分发

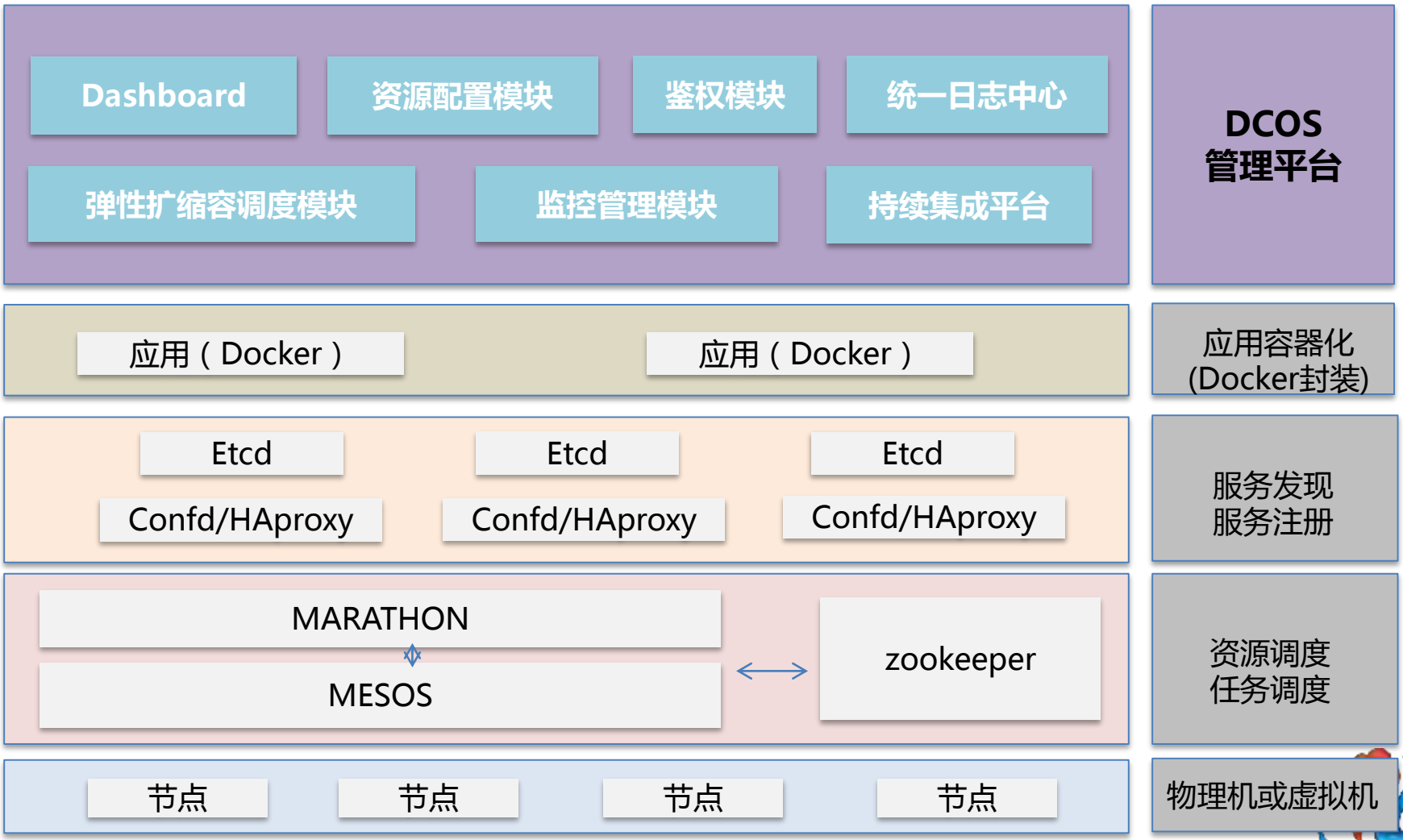
Marathon将服务通过Confd注册到HAProxy

思路来自：刘天斯《构建一个高可用及自动发现的Docker基础架构-HECD》
http://blog.liuts.com/post/242/

DCOS架构图


Mesos Cluster (Master)
Mesos Master
Marathon
Zookeeper
Etcd
Mesos Master
Marathon
Zookeeper
Etcd
Mesos Master
Marathon
Zookeeper
Etcd
LB
服务发现
HAProxy
Confd
HAProxy
Confd
部署
弹性调度
代码
Git服务器
构建
Jenkins
构建/上传
Push
应用包、镜像
Docker
Registry
Pull
Docker
Container
Docker
Container
Docker
Container
Mesos Slave
Mesos Slave
Mesos Slave
Mesos Cluster (Slave)
Distributed File System(Ceph、HDFS…)

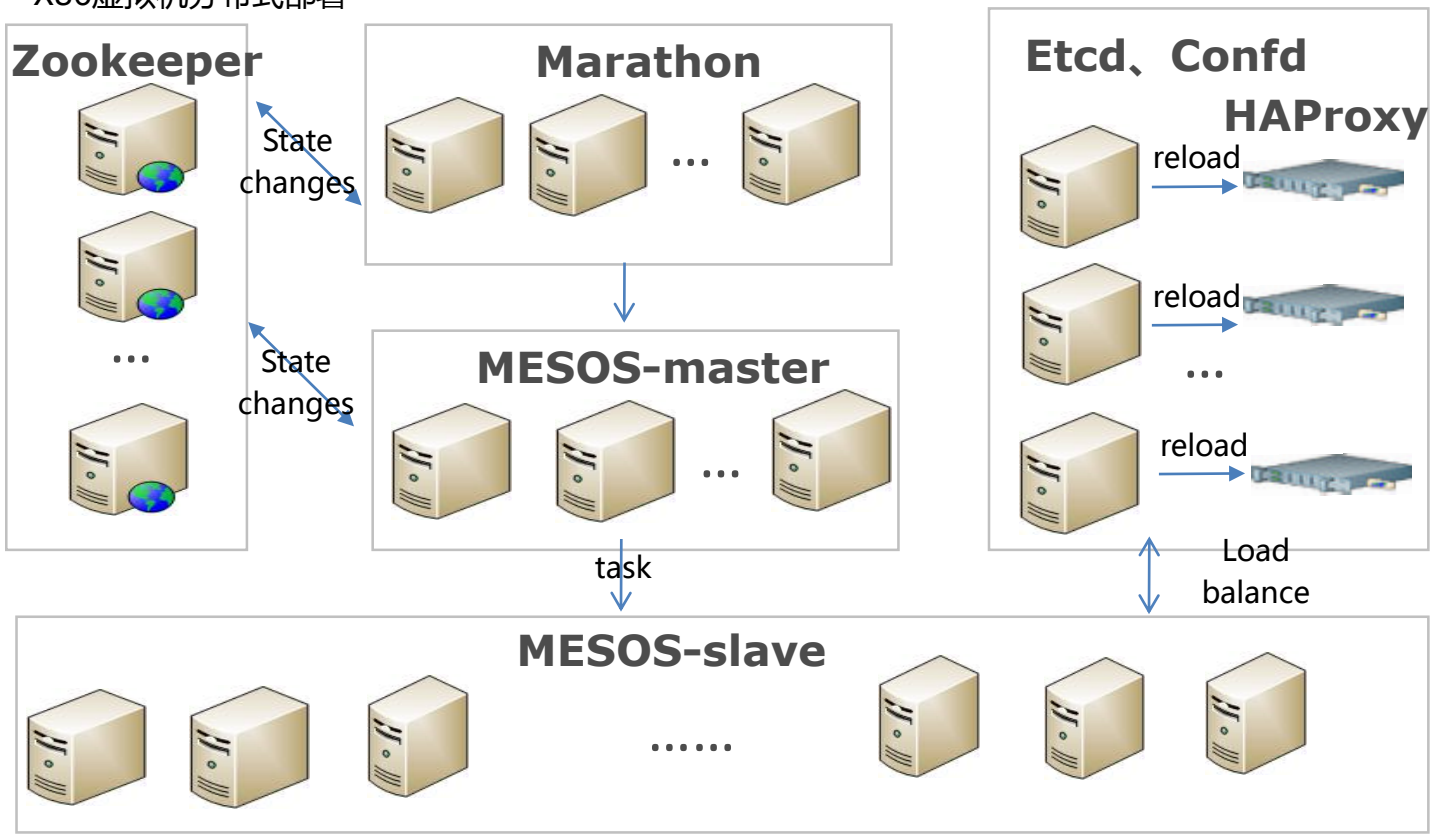
DCOS功能架构图


Dashboard
资源配置模块
鉴权模块
统一日志中心
弹性扩缩容调度模块
监控管理模块
持续集成平台
DCOS
管理平台
应用（Docker）
应用（Docker）
应用容器化
(Docker封装)
Etcd
Etcd
Etcd
Confd/HAproxy
Confd/HAproxy
Confd/HAproxy
服务发现
服务注册
MARATHON
MESOS
zookeeper
资源调度
任务调度
节点
节点
节点
节点
物理机或虚拟机

DCOS物理部署

浙江移动DCOS平台采用93个主机节点，其中平台部分由5个节点构成Mesos Master Cluster，8个节点构成Haproxy Cluster，80个计算节点，平台和计算节点均跨机房部署

Mesos master、Marathon 、Zookpeer、Etcd 共采用5台X86虚拟机分布式部署



HAProxy 和Confd共采用8台X86虚拟机

Mesos slave节点采用80台X86虚拟机

所有组件容器化部署

试点

- 组件版本
  - Mesos 0.25
  - Marathon 0.11
  - Docker 1.8.3
  - Zookeeper 3.4.6
  - HAProxy 1.61
  - Etcd 2.2.1

- 业务规模（手机营业厅）
  - 注册用户2500万
  - 日活跃用户数300万
  - “双十一”抢购

Dashboard


DCOS云管理平台
数据中心视图
Dashboard
Help
管理员
服务管理
服务管理
服务模板
容器模板
服务扩缩
服务重启
持续集成
集成
测试
性能管理
容器性能
路由性能
主机性能
统一告警
统一日志
中心总数: 2
系统总数: 5
计算节点: 80
容器总数: 528
CPU
100
50
0
空闲率
分配率
使用率
269 / 1040 (Core)
内存
80
60
40
20
1434.6 / 4432.6 (GB)
存储
2370.1 / 7467.7 (GB)
状态
MESOS-MASTER
MARATHON
ZOOKEEPER
HAPROXY
MESOS-SLAVE
container counts
200
150
100
50
0
08:20
08:30
08:40
08:50
09:00
09:10
09:20
09:30
09:40
09:50
10:00
10:10
全渠道
手机营业厅APP
手机营业厅WEB
营业APP
营业WEB

数据中心容器视图

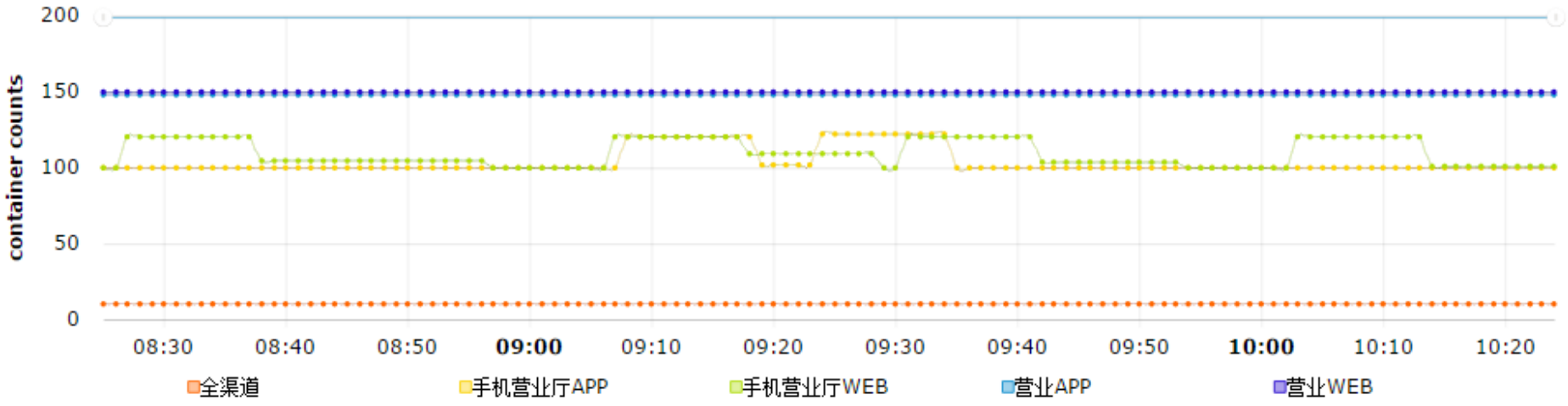

数据中心视图
全渠道
营业APP
营业WEB
手机营业厅APP
手机营业厅WEB
三墩 节点数40 容器数224
石桥 节点数40 容器数285

内容

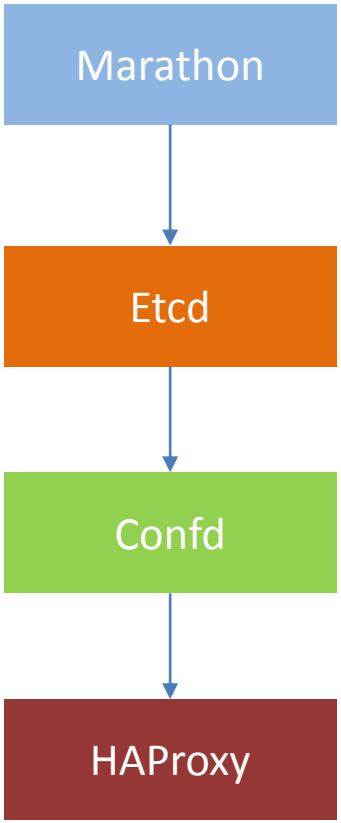
- 为什么使用MESOS
- 基于MESOS的DCOS实现
- **实践经验**

自动弹性扩缩容

- Marathon的扩缩容默认只能根据用户需要进行手动调整，我们结合多年的系统运维经验，实现基于并发数、响应时间、CPU和内存使用率等容量指标进行自动弹性扩缩容调度的算法。



Marathon Etcd联动实现服务发现注册



- Etcd只是个独立的服务注册发现组件，只能通过在宿主机上部署Etcd发现组件，通过其发现宿主机的容器变化来发现，属于被动的发现，往往会出现发现延迟时间较长的问题，我们通过修改Etcd组件的发现接口，实现与Marathon的Event事件接口进行对接，达到Marathon的任何变动都会及时同步给Etcd组件，提高了系统的发现速度，并且避免在每个宿主机上部署Etcd 发现组件。

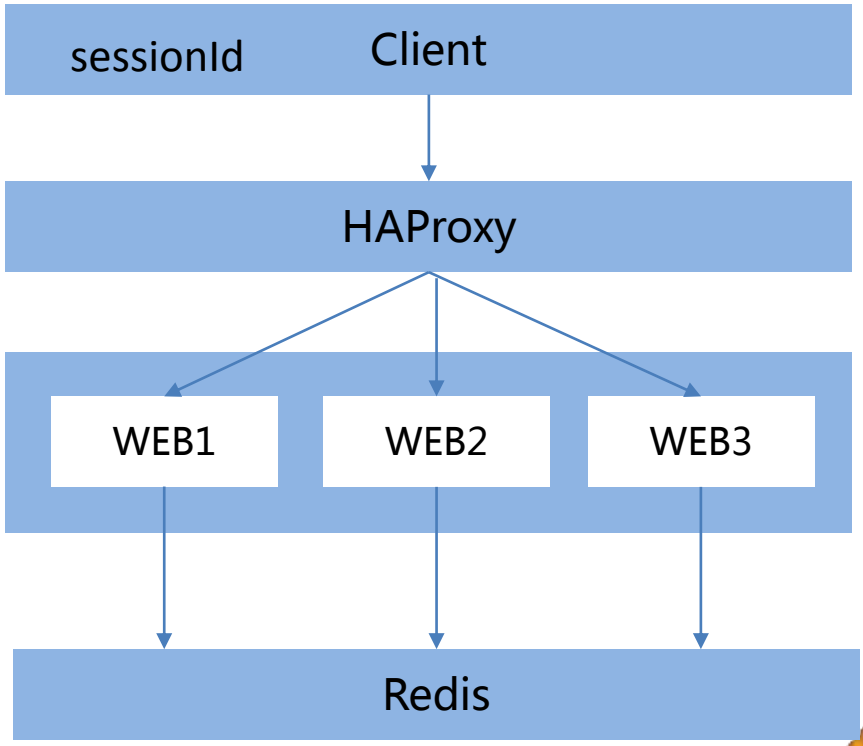
数据中心切换


数据中心视图|跨数据中心切换
数据中心名称
三墩
容器总数
224
应用名称
营业WEB 72
手机营业厅APP 43
全渠道 5
手机营业厅WEB 34
营业APP 70
关闭
跨数据中心切换

应用的改造

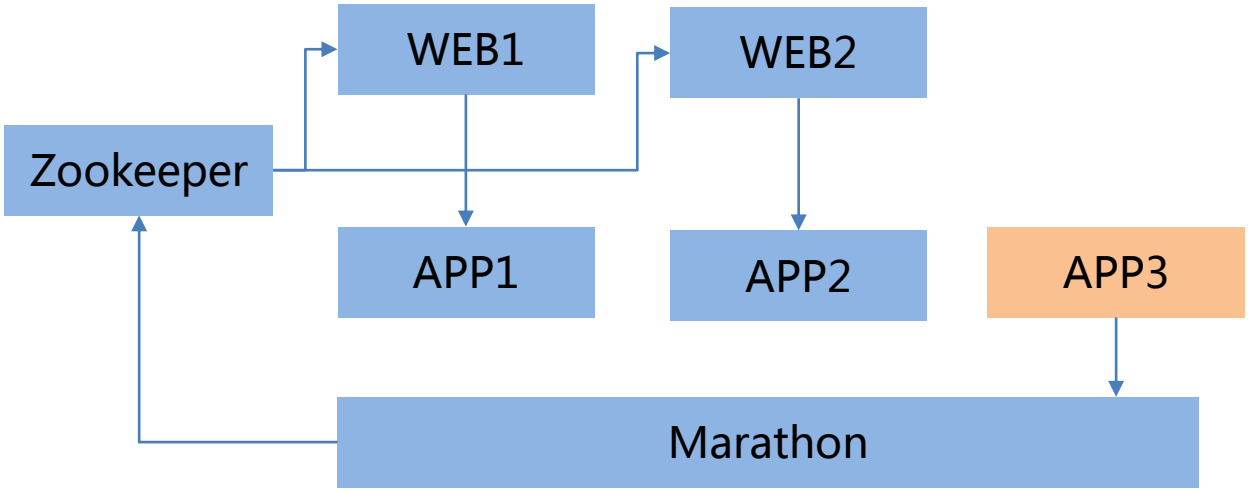
自动弹性扩缩容对应用的要求：**无状态化**

- 接入层的无状态化改造
  - 去http session
  - 交互用http+json短连接
  - Session信息放缓存



应用的改造

- 内部服务调用的改造
  - **HTTP接口**：同接入层一样使用负载均衡方案HAProxy+Confd+Etcd；
  - **服务化框架**：使用服务化框架服务的发现和注册功能，注意需要将容器外的IP和端口上报给配置中心；



DCOS带来的好处

■ 高资源利用率

DCOS相较于虚拟机有着基于CPU、内存、IO的更细粒度的资源调度，多个计算框架或应用程序可共享资源和数据，提高了资源利用率。

■ 高效的跨数据中心的资源调度

DCOS平台展现了其在线性扩展、异地资源调度等方面的优异性能，无需大二层网络实现跨机房的资源调度。

■ 弹性扩缩容

彻底解决应用的扩缩容问题，容量管理从“给多少用多少”向“用多少给多少”转变，被动变主动。应用的扩缩容时间从传统集成方式的2-3天缩短到秒级，可以根据业务负载自动弹性扩缩容。

■ 高可用性、容灾

DCOS平台所有组件采用分布式架构，应用跨机房分布式调度。自动为宕机服务器上运行的节点重新分配资源并调度，保障业务不掉线，做到故障自愈。

谢谢

三墩IT人