Gdevops

全球敏捷运维峰会

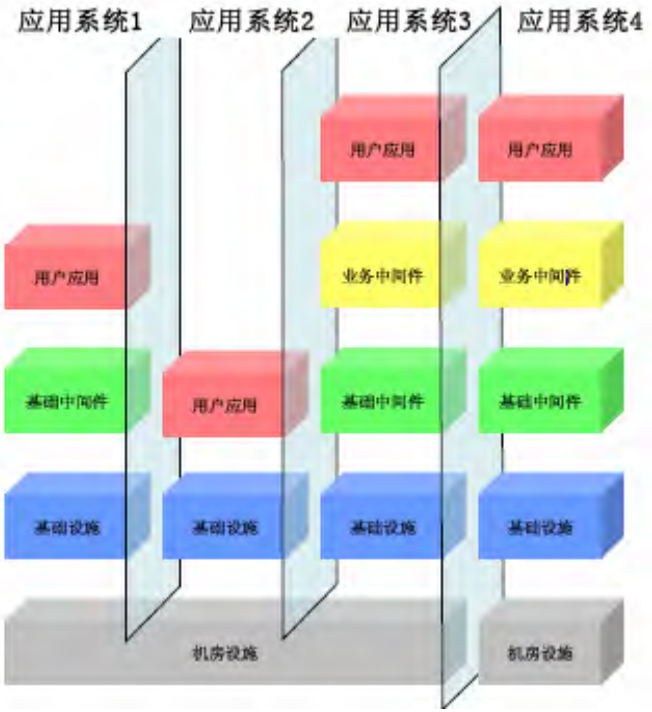
中国移动浙江公司DCOS生产实践

演讲人：朱智武

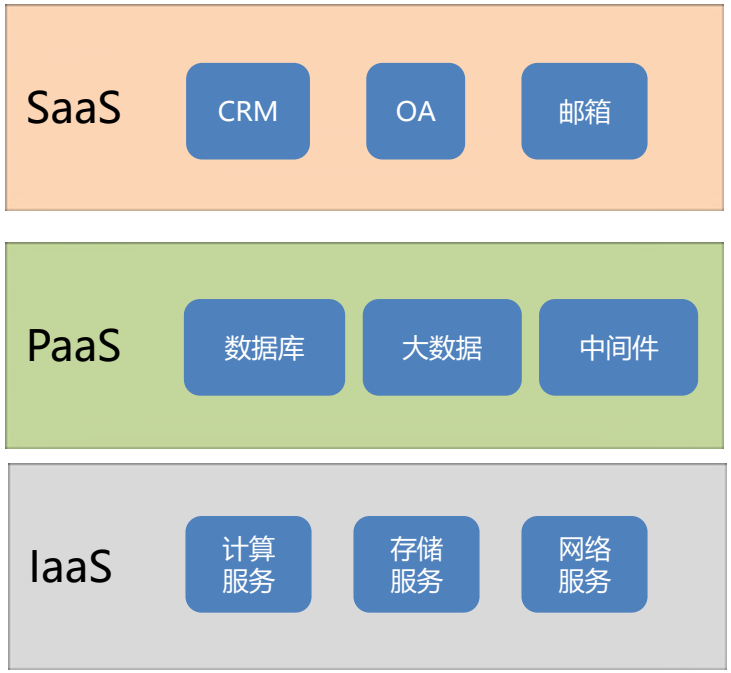
目录

- 走向DCOS
- 基于MESOS的DCOS实现
- 生产实践

云计算驱动架构演进



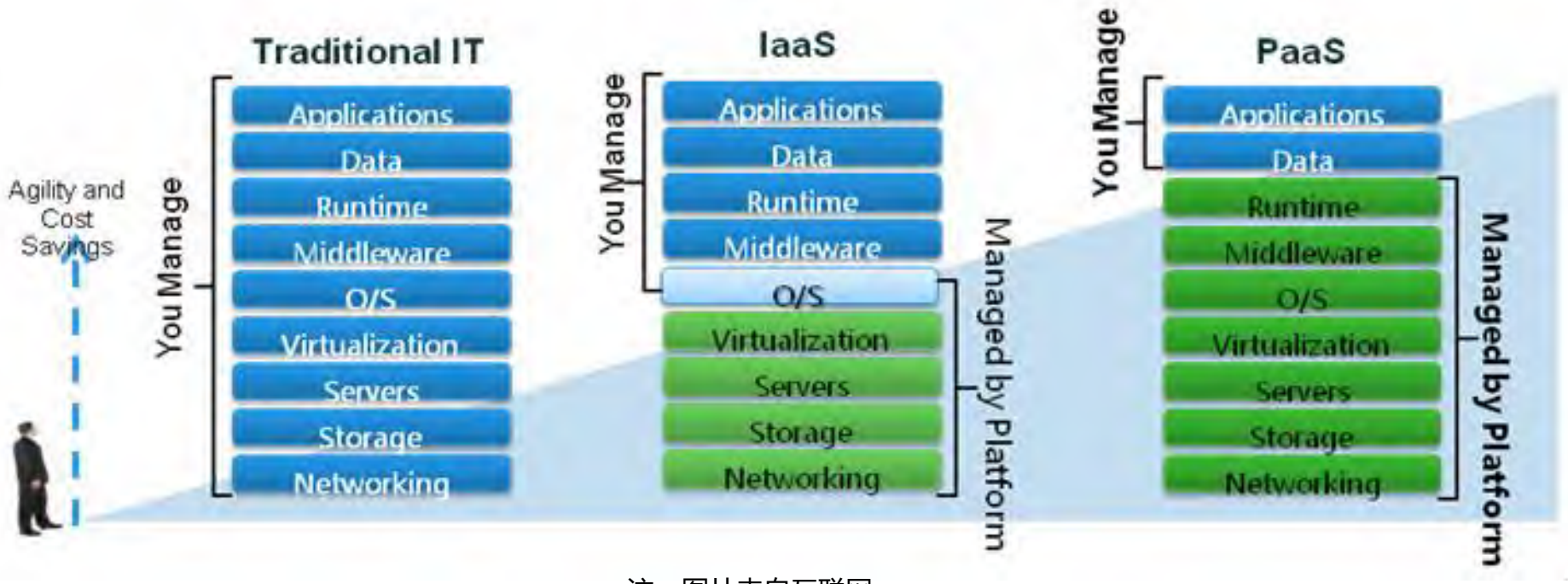
“烟囱”式IT系统架构



云化IT架构

需求

IaaS和PaaS



注：图片来自互联网

第三代PaaS

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

- 以分布式资源调度（Mesos、Yarn等）为基础，结合容器技术构建
- 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

DCOS

数据中心操作系统（DataCenter Operating System，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

注：以Mesos为例，来自互联网

DCOS解决方案

	Mesos	Yarn	Kubernetes	Docker Swarm	CloudFoundry/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）	CloudFoundry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

浙江移动云化的阶段

传统孤岛

- 对数据中心内部整体目标架构**没有统一的**规划设计

孤岛

标准化

- **标准化的**硬件和软件体系
- 业务基础架构建设以月为单位

X86化

简化

IaaS 资源池化

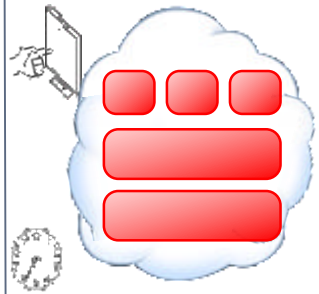
- 通过虚拟化实现共享的**基础架构**
- 业务基础架构建设以周为单位
- 实现**虚拟机级**弹性伸缩

虚拟化

高效

PaaS和应用资源池化

- 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

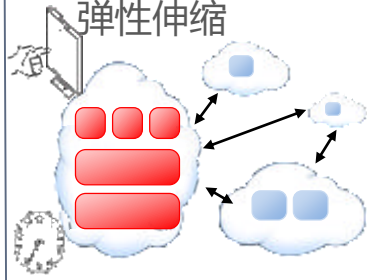


服务化

灵活

DCOS化

- 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



智能化

动态复用

浙江移动DCOS历程

2014年3-8月

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

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

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

2015年11月

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

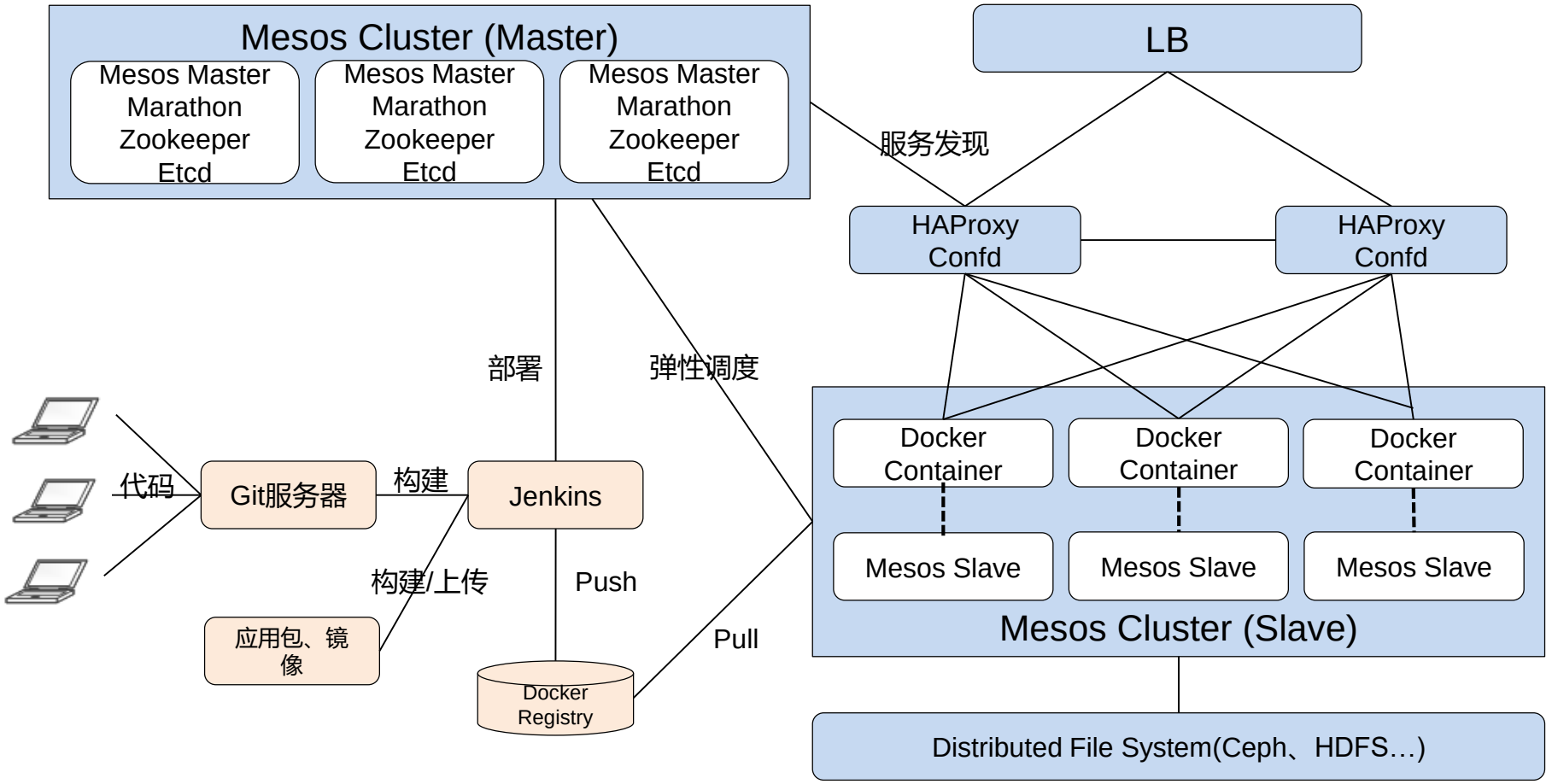
2015年12月

2015年12月10日上线CRM应用

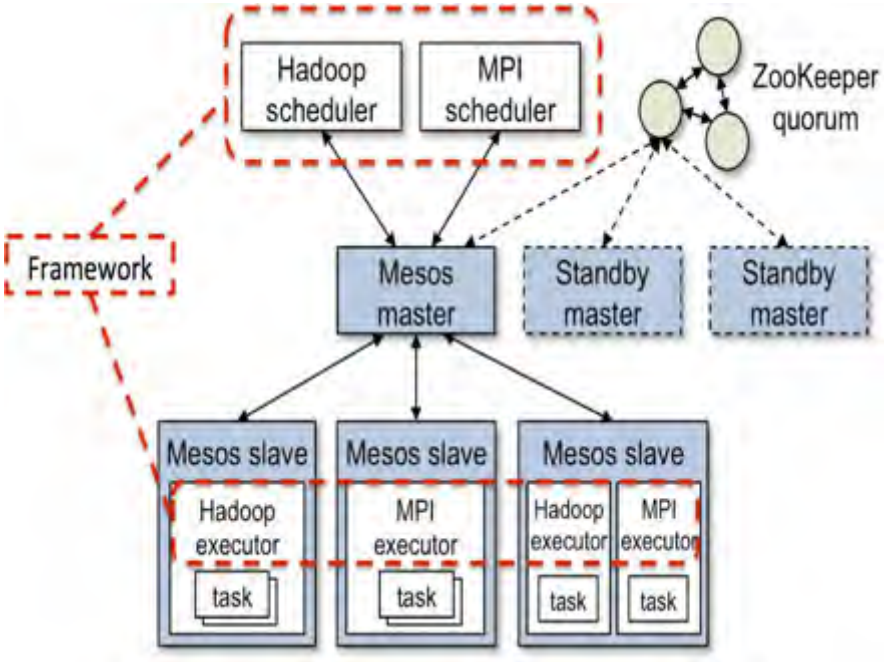
目录

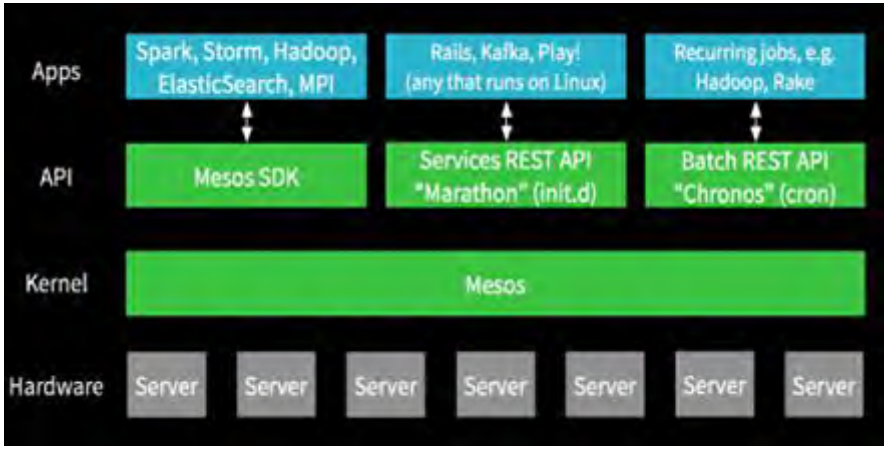
- 走向DCOS
- 基于MESOS的DCOS实现
- 生产实践

DCOS架构图

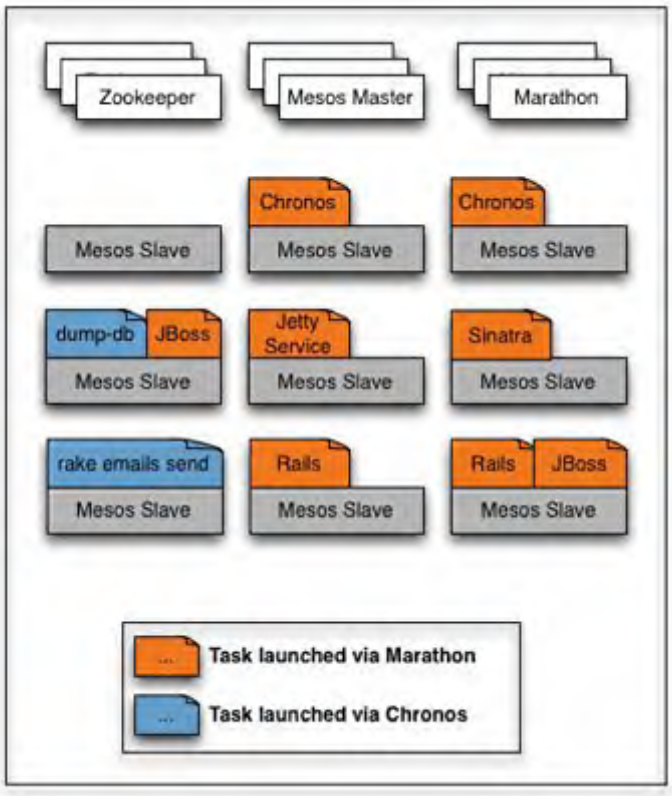

Mesos Cluster (Master)
Mesos Master
Marathon
Zookeeper
Etcd
Mesos Master
Marathon
Zookeeper
Etcd
Mesos Master
Marathon
Zookeeper
Etcd
LB
服务发现
HAProxy
Confd
HAProxy
Confd
部署
弹性调度
代码
Git服务器
构建
Jenkins
构建/上传
Push
应用包、镜像
Docker Registry
Pull
Docker Container
Docker Container
Docker Container
Mesos Slave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

资源调度 - Mesos


Hadoop scheduler
MPI scheduler
ZooKeeper quorum
Framework
Mesos master
Standby master
Standby master
Mesos slave
Mesos slave
Mesos slave
Hadoop executor
MPI executor
Hadoop executor
MPI executor
task
task
task
task


Apps
Spark, Storm, Hadoop, ElasticSearch, MPI
Rails, Kafka, Play! (any that runs on Linux)
Recurring jobs, e.g. Hadoop, Rake
API
Mesos SDK
Services REST API "Marathon" (init.d)
Batch REST API "Chronos" (cron)
Kernel
Mesos
Hardware
Server
Server
Server
Server
Server
Server
Server

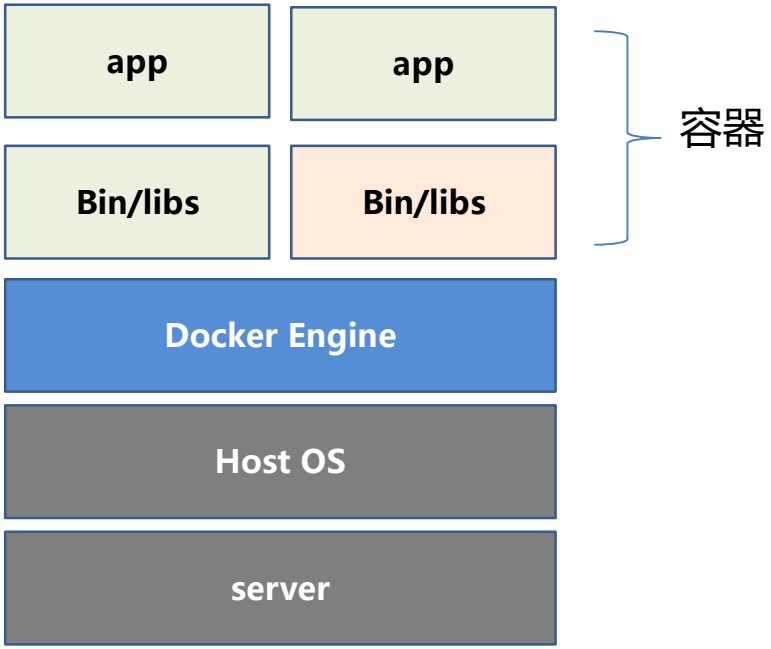
任务调度 - Marathon

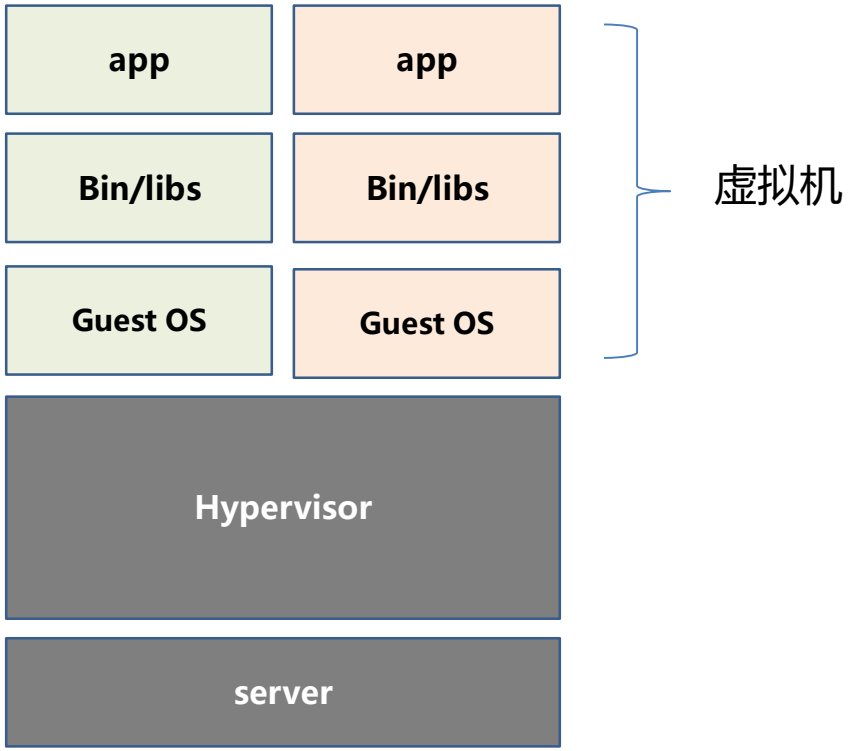


Mesos仅负责分布式集群资源分配

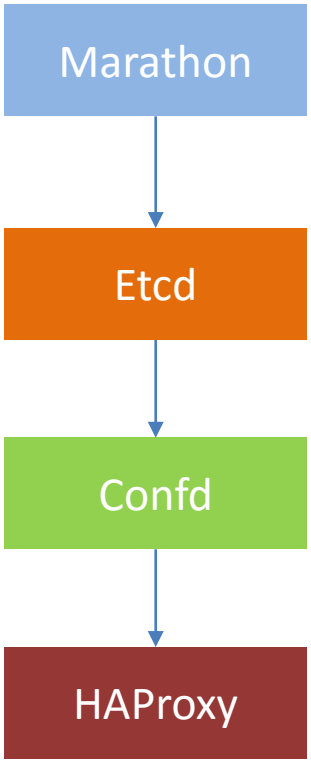
Marathon负责任务调度，故障转移

应用封装 - Docker


app
app
Bin/libs
Bin/libs
Docker Engine
Host OS
server
容器


app
app
Bin/libs
Bin/libs
Guest OS
Guest OS
Hypervisor
server
虚拟机

服务注册发现

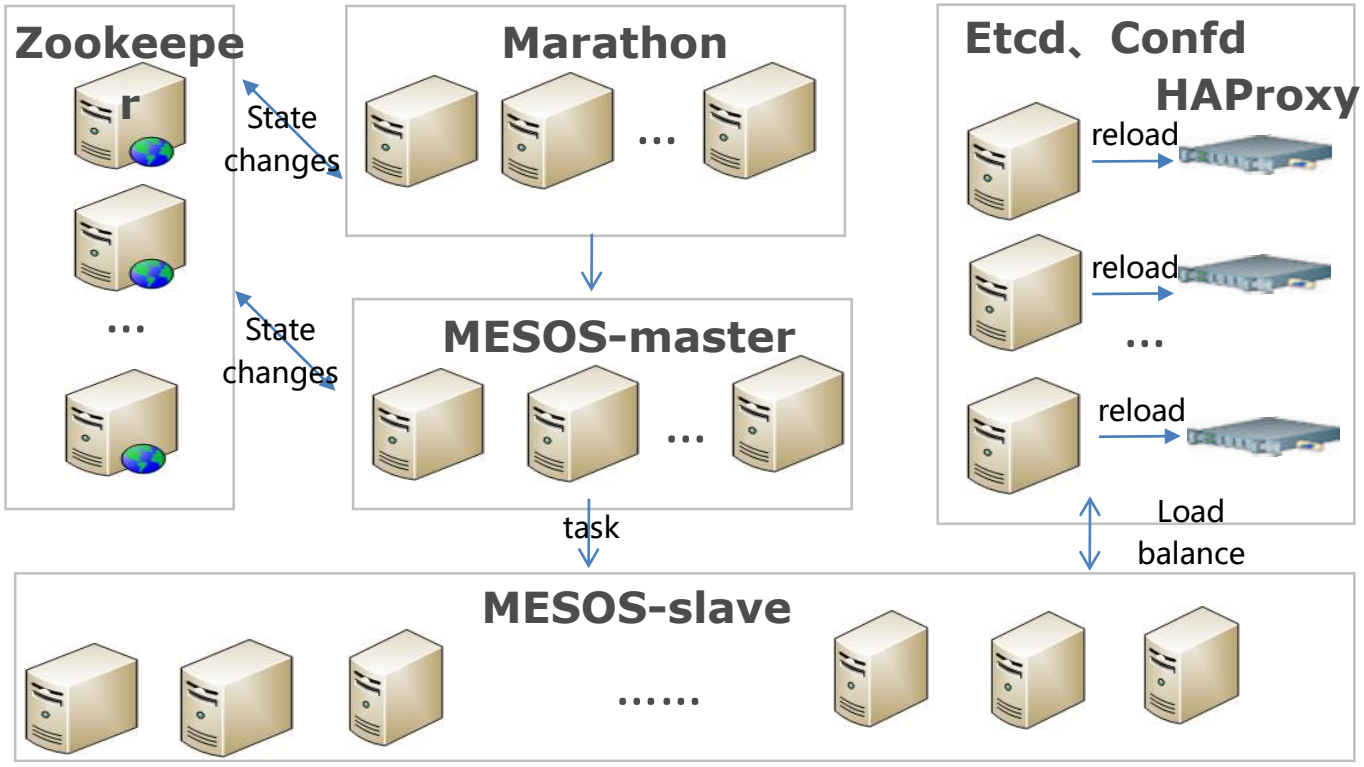


Etcd只是个独立的服务注册发现组件，只能通过在宿主机上部署Etcd发现组件，通过其发现宿主机的容器变化来发现，属于被动的发现，往往会出现发现延迟时间较长的问题，我们通过修改Etcd组件的发现接口，实现与Marathon的Event事件接口进行对接，达到Marathon的任何变动都会及时同步给Etcd组件，提高了系统的发现速度，并且避免在每个宿主机上部署Etcd发现组件。

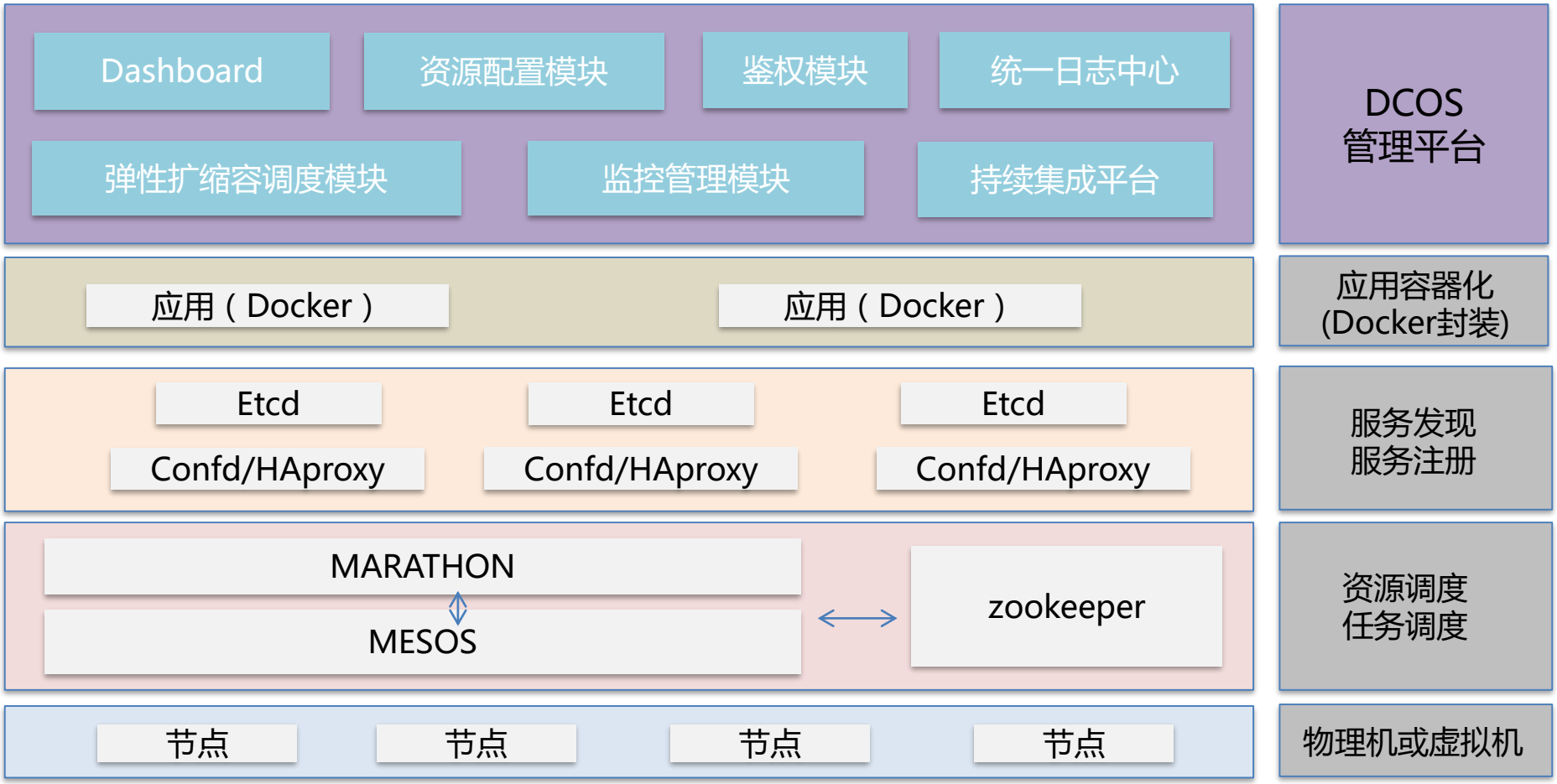
思路来自：刘天斯《构建一个高可用及自动发现的Docker基础架构-HECD》http://blog.liuts.com/post/242/

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。



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

手机营业厅试点

手机营业厅

组件版本

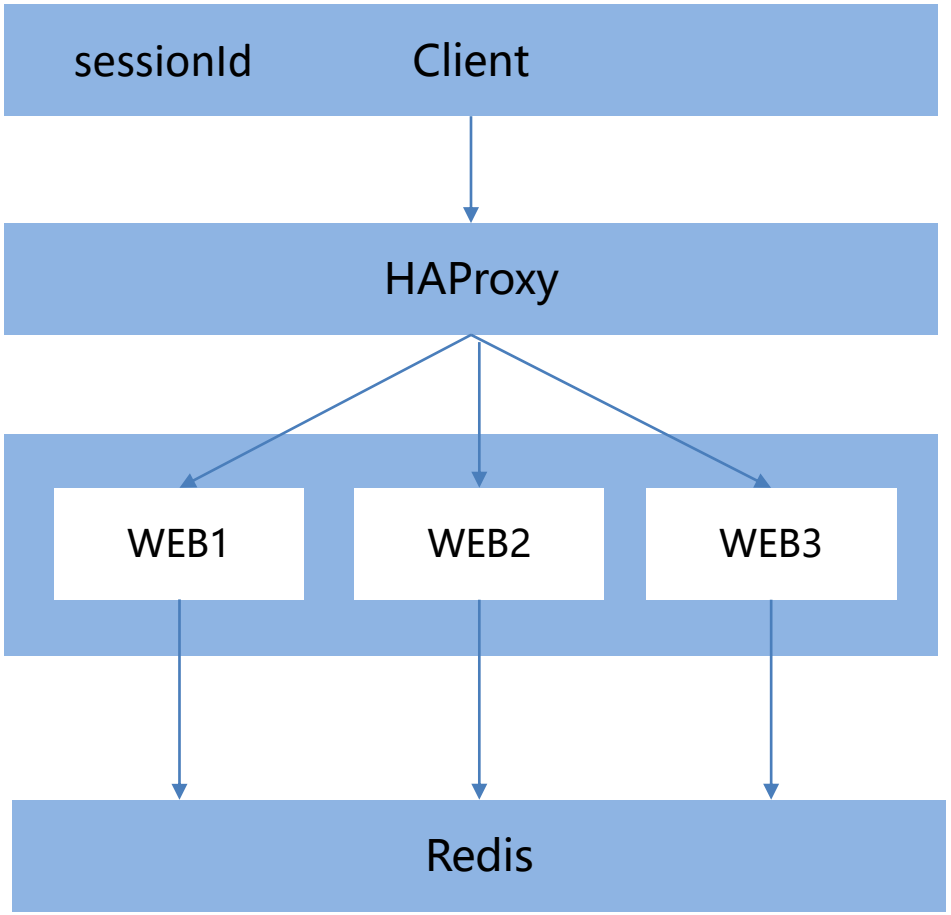
—Mesos 0.25
—Marathon 0.11
—Docker 1.8.3
—Zookeeper 3.4.6
—HAProxy 1.6.1
—Etcd 2.2.1

业务规模

—注册用户2500万
—日活跃用户数300万
— “双十一”抢购

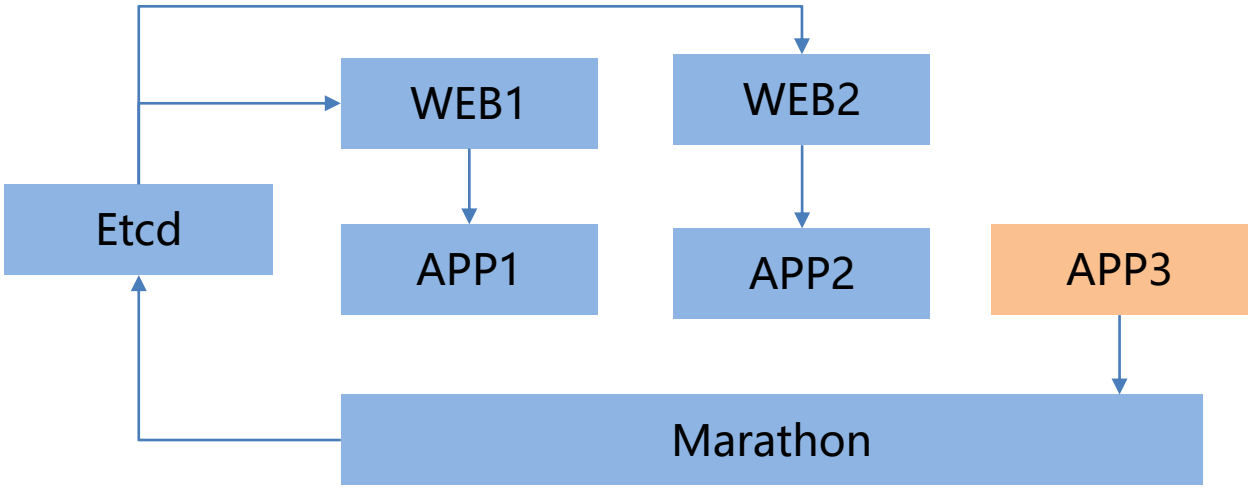
应用的改造

- 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注册发现功能，注意需要将容器外的IP和端口上报给配置中心。



目录

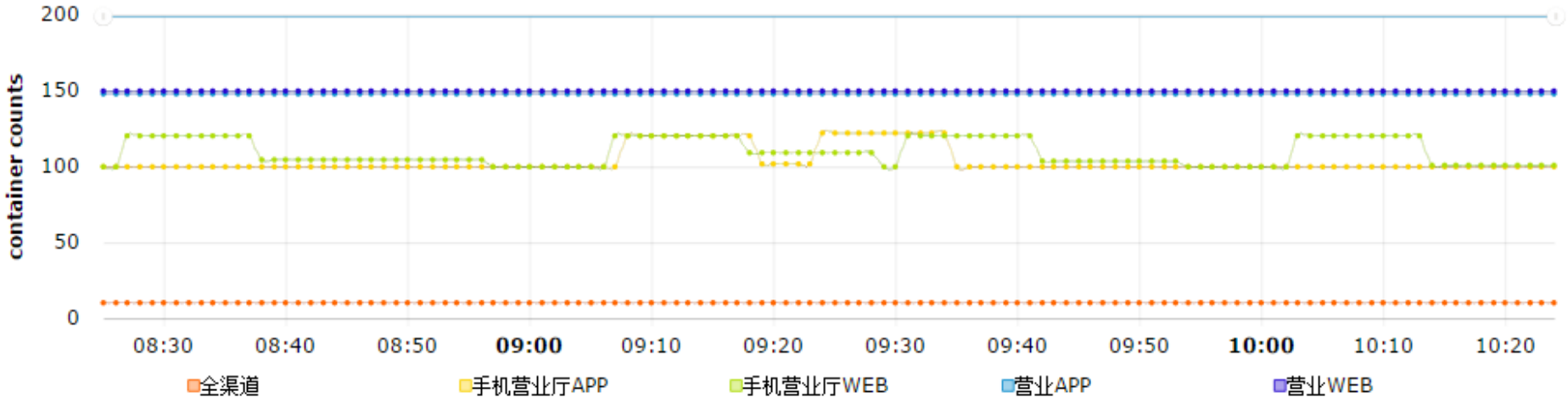
- 走向DCOS
- 基于MESOS的DCOS实现
- 生产实践

DCOS云管理平台


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
0
1434.6 / 4432.6 (GB)
存储
80
60
40
20
0
2370.1 / 7467.7 (GB)
状态
MESOS-MASTER
MARATHON
ZOOKEEPER
HAPROXY
MESOS-SLAVE
200
150
100
50
0
container counts
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

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。



跨数据中心切换


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

DCOS带来的好处

1

高资源利用率

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。

2

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。

3

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。

4

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。

三墩IT人

Gdevops

全球敏捷运维峰会

THANK YOU！