



# 弹性伸缩的世界

孙寅，现在小米负责运维基础设施、基础平台的构建

# 平台演进

The diagram illustrates the evolution of platforms through a horizontal timeline. The timeline is represented by a bar that transitions from dark grey on the left to light green on the right. Three key stages are marked with dots and connected to labels by vertical lines. The first stage, '工具的集合' (Collection of Tools), is marked with a dark grey dot on the dark grey part of the bar. The second stage, '体系化平台' (Systematized Platform), is marked with a dark grey dot on the transition point between dark grey and light green. The third stage, 'DCOS', is marked with a light green dot on the light green part of the bar.

工具的集合

DCOS

体系化平台

# 1 工具的集合

网络  
管理

初始  
化

数据  
库

APM

CMDB

发布

监控

服务树

## 2

## 体系化平台

**JOBNAME**

公司

部门

产品

集群

服务组

服务

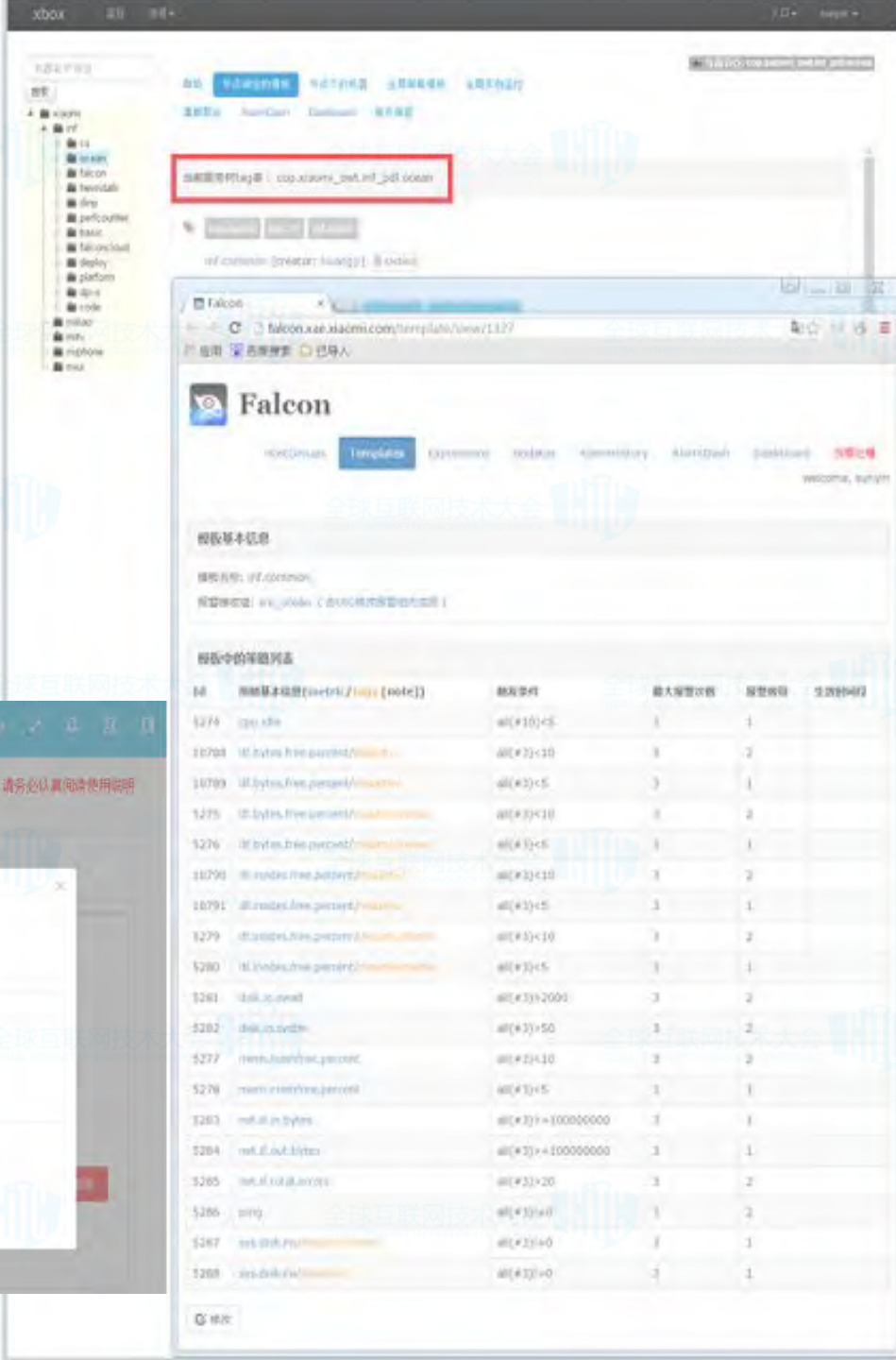
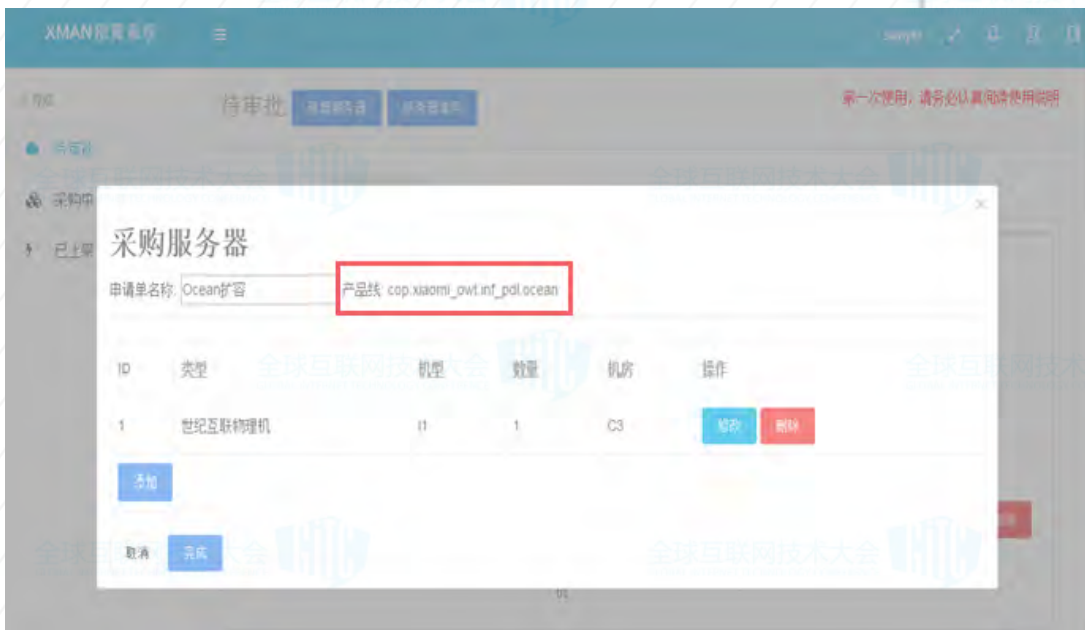
任务组

任务



## 2 体系化平台

# 主机交付即监控



## 2 体系化平台

# 发布即监控

The screenshot displays a monitoring platform interface. The top navigation bar includes links for '运维部署' (Operations & Deployment), '查看任务' (View Tasks), '发起任务' (Initiate Task), '回滚任务' (Rollback Task), 'clone机器' (Clone Machine), 'clone job', '跨产品线的clone' (Cross-product line clone), 'rbox首页' (Rbox Home), 'wiki', '吐槽' (Complaints), and '需求反馈' (Requirement Feedback).

The main content area is divided into two sections. The left section, titled '部署服务器' (Deploy Servers), shows a tree view of servers under the 'Xiaomi' category, including folders like 'mitv', 'miui', 'miipone', 'miliao', 'account1', 'account3', 'admob', 'baw', 'com', 'computing', 'data', 'demo', 'hadoop', 'ia', 'ja', 'in', 'afs', 'micloud', 'miliao', 'mistat', 'offline', 'op', 'other', 'QA', 'qa', 'RD', 'relation', 'search', 'sec', 'securitycloud', 'smart', 'storage', 'voip', 'vms', 'www', 'XAE', 'xcache', 'xiaoqiang', and 'inf'.

The right section, titled '服务列表' (Service List), shows a table of services. The first row is highlighted in yellow and contains the following information:

操作	服务名称	实例统计
发起	job.userconnect_service.userconnect_cluster.production-1g	负载 3 3
发起	job.userconnect_service.userconnect_cluster.aws-asp	负载 2 2
发起	job.userconnect_service.userconnect_cluster.c3	负载 2 2

The bottom section shows a detailed view of a template named 'Falcon'. The URL is 'falcon.xae.xiaomi.com/template/view/1654'. The table below lists the templates and their details:

Id	策略基本信息(metric/tags [note])	触发条件	最大报警次数	报警级别	生效时间
7066	agent.lg.dreamNet.master.ticket.verify.num /cluster=production-1g, cop=xiaomi, job=userconnect, owt=miliao, pdf=account, service=userconnect [lg.dreamNet.master failed.]	all(#1)<25	9999	1	
7077	agent.lg.dreamNet.master.ticket.verify.num /cluster=production-1g, cop=xiaomi, job=userconnect, owt=miliao, pdf=account, service=userconnect [lg.dreamNet.master unavailable]	all(#1)<2	9999	0	
7067	agent.lg.email.master.ticket.verify.num /cluster=production-1g, cop=xiaomi, job=userconnect, owt=miliao, pdf=account, service=userconnect [lg.email.master failed.]	all(#2)<10	9999	1	06:00-23:59
7079	agent.lg.email.master.ticket.verify.num /cluster=production-1g, cop=xiaomi, job=userconnect, owt=miliao, pdf=account, service=userconnect [lg.email.master unavailable]	all(#5)<2	9999	0	
7068	agent.lg.email.register.ticket.verify.num /cluster=production-1g, cop=xiaomi, job=userconnect, owt=miliao, pdf=account, service=userconnect [lg.email.register failed.]	all(#15)<2	9999	3	06:00-23:59
7080	agent.lg.email.register.ticket.verify.num /cluster=production-1g, cop=xiaomi, job=userconnect, owt=miliao, pdf=account, service=userconnect [lg.email.register unavailable]	all(#15)<2	9999	0	00:00-00:10
7069	agent.lg.guodu.master.ticket.verify.num /cluster=production-1g, cop=xiaomi, job=userconnect, owt=miliao, pdf=account, service=userconnect [lg.guodu.master failed.]	all(#1)<25	9999	1	
7081	agent.lg.guodu.master.ticket.verify.num /cluster=production-1g, cop=xiaomi, job=userconnect, owt=miliao, pdf=account, service=userconnect [lg.guodu.master unavailable]	all(#1)<2	9999	0	

### 3 DCOS

Google

Borg

Omega

K8S

Mesos  
phere

Marathon

DCOS

Xiaomi

Ocean





# 核心技术栈

## Mesos+Marathon +Docker





# Docker

## 在工业环境落地

● **网络**

● **文件系统**

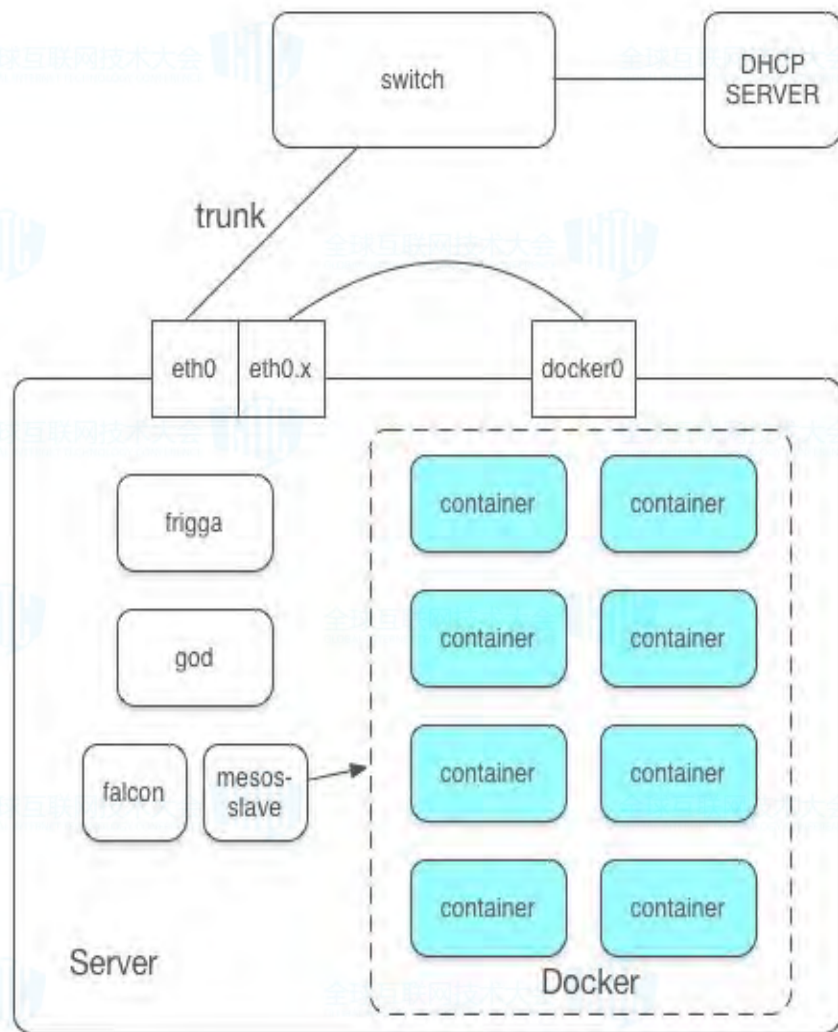
● **资源隔离  
资源汇报**

● **Docker-Init**



# 1 网络

- 真实内网IP，便于标识和定位
- 与原有物理网络天然互通
- 吞吐、延迟与纯物理网络几无差别
- 大三层网络，无广播风暴风险



## 2

# 文件系统

Root分区

Home分区

devicemapper

LVM逻辑卷

## 目标

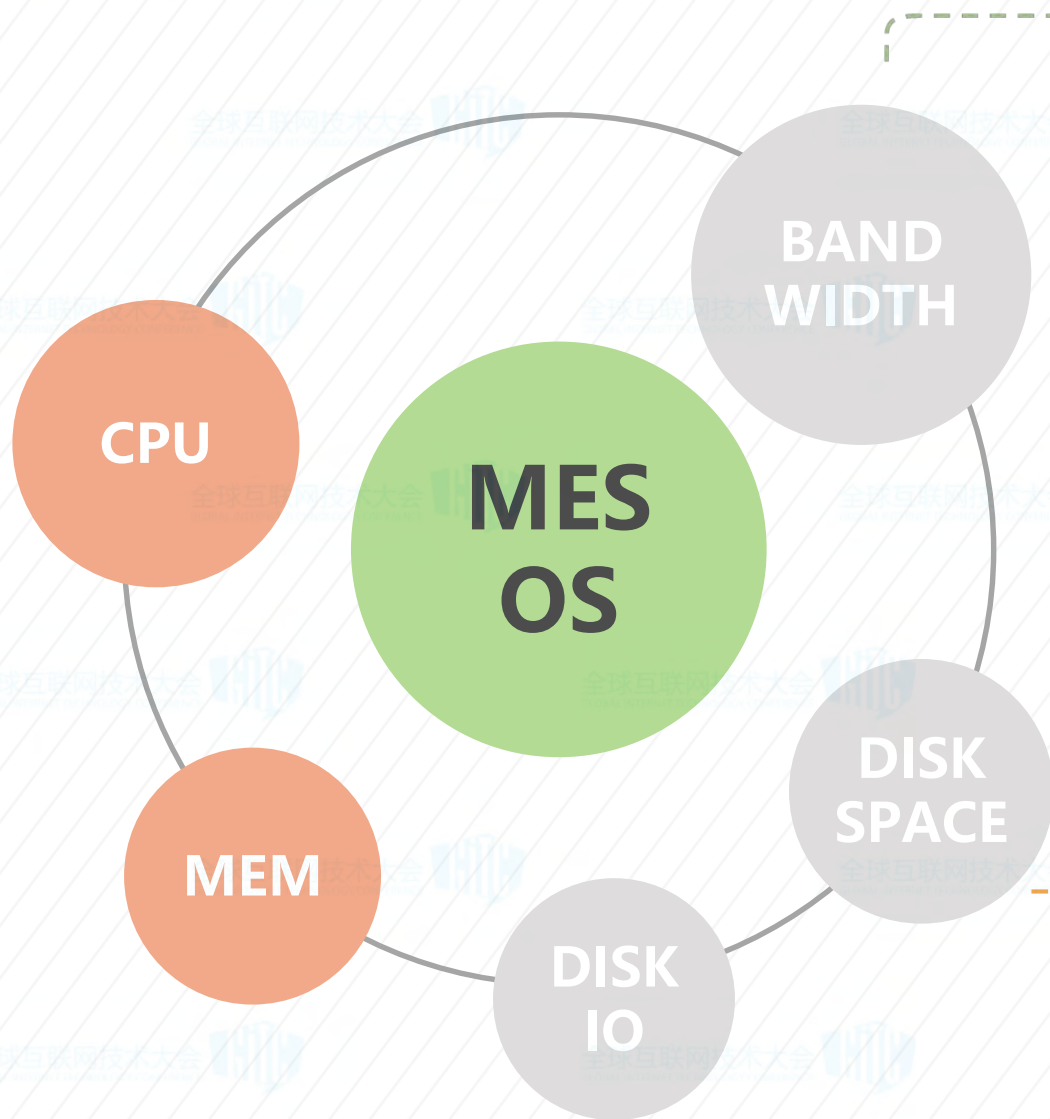
- 容器磁盘空间大小可控
- 保证磁盘IO性能不下降

## 技术考量

- aufs、btrfs等等存储方式都会损耗IO性能
- 一个宿主机DM只能设置固定大小
- Home分区引入LVM动态设置空间大小



### 3 资源隔离资源汇报



- 扩展初始Mesos Resource
- 环境变量设置资源大小
- 实时采集已使用资源
- TC限速模拟网络带宽隔离
- DM+LVM隔离磁盘空间
- 旁路清理LVM卷

➤ DHCP

➤ 内置服务管理

➤ 进程组

➤ 进程HealthCheck

➤ 进程启动退出钩子

➤ 回收僵尸进程

➤ 与内部系统对接

➤ Marathon HealthCheck

➤ Falcon

➤ DoorGod ( 安全免密登陆 )

➤ Dynamic Auth

An illustration featuring a green and white Earth globe in the center. An orange rocket with a white nose cone is shown launching from the right side of the globe, leaving a white smoke trail. Two white orbital paths are depicted around the globe. The background is a dark gray circle with several white stars. The entire image is overlaid with a repeating pattern of the text '全球互联网技术大会' and a logo.

# 工程体系的变化

## 1 状态分层

无状态

有状态

无状态服务大抵相同  
有状态服务有所不同

- 服务发现
- 数据迁移
- 分角色控制



## 2

## 服务发现

RPC框架

Mesos-DNS

Zookeeper

Mysql Proxy

Nginx Module

Redis Cluster(Gossip)

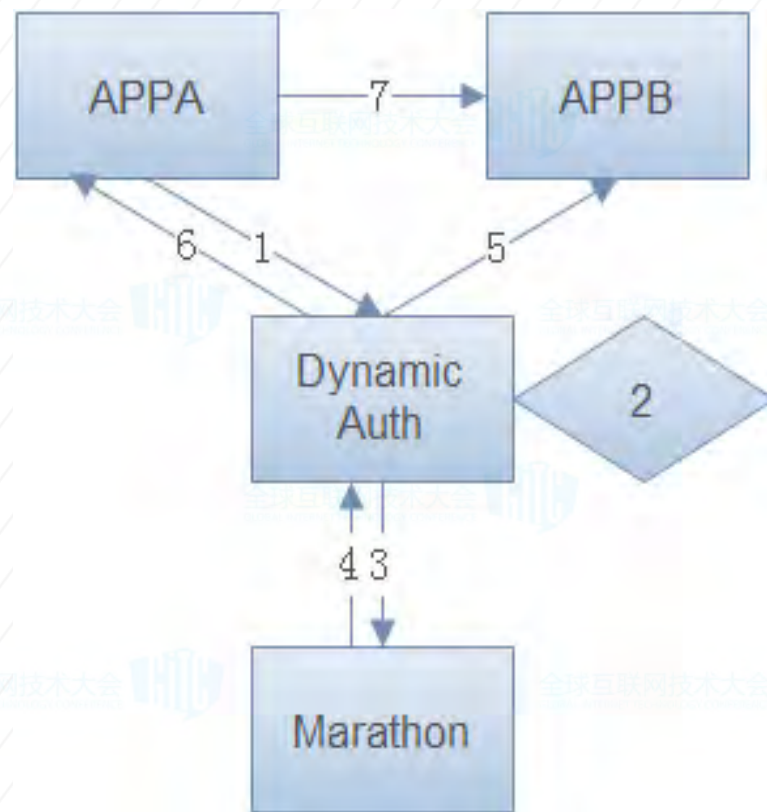
Other



## 3

## 动态安全

1. APPA表示我要访问APPB，向DynamicAuth请求私钥
2. DynamicAuth验证APPA是否对APPB有权限，无权限直接拒掉
3. DynamicAuth向Marathon验证APPA的IP是否真实属于APPA
4. 返回验证通过
5. 向APPB发放公钥
6. 向APPA发放私钥
7. APPA带着私钥访问APPB，访问通过





# 今年在做什么

## 1

## 自动伸缩

➤ Falcon自动采集并监控容器的

CPU IDLE

MEM FREE

PROC QPS

PROC DELAY

➤ Hook回调

Marathon API

➤ 最短10s触发伸缩

动态调度部署系统

job.ocean-monitor-job\_service.ocean-monitor-job\_cluster.production-ig\_pdl.ocean\_owt.inf\_cop.xiaomi

任务配置 运行时配置 自动伸缩 高级选项 取消 保存

伸缩指标 ?

CPU平均使用率

伸缩配置

实例数上限: 5 实例数下限: 1 实例增减步长: 1

是否通知: ☒ 短信通知组 (UIC): sre\_ocean

触发条件 ?

连续次数

加实例: 3 > 80 减实例: 3 < 60



## 2

## 定时触发伸缩

➤ 利用Chronos  
调用Marathon  
API

动态调度部署系统

🔍 🌐 🔔 👤 🏠

主导航

仪表盘

🌐 宿主机管理

🕒 产品线仪表盘

📋 任务管理

🗄️ 数据库管理

🗄️ XRedis

📋 模板管理

🚩 ELB

job.ocean-monitor-job\_service.ocean-monitor-job\_cluster.production-lg\_pdl.ocean\_owt.inf\_cop.xiaomi

任务配置

运行时配置

自动伸缩

高级选项

取消

保存

伸缩指标 ?

定时伸缩 ▼

定时伸缩配置 ?

+

起始时间	时间间隔	实例数	
2016-10-01 00 📅	年 月 日 02 00	2	-
2016-10-01 00 📅	年 月 日 07 00	4	-

### 3

## 四层ELB

- 自动配置内网域名，按运营商自动划分线路
- docker-init和旁路模块都会动态更新LVS配置，可重入
- 域名线路自动感知切换

动态调度部署系统

主导航

- 仪表盘
- 宿主机管理
- 产品线仪表盘
- 任务管理
- 数据库管理
- XRedis
- 模板管理
- ELB

### 新建ELB

环境

机房:

c3

内外网:

外网

类型:

4层

L4配置

JOB平台:

ocean

JOB绑定:

job.ocean-monitor-job\_

添加端口

+

VS:

80

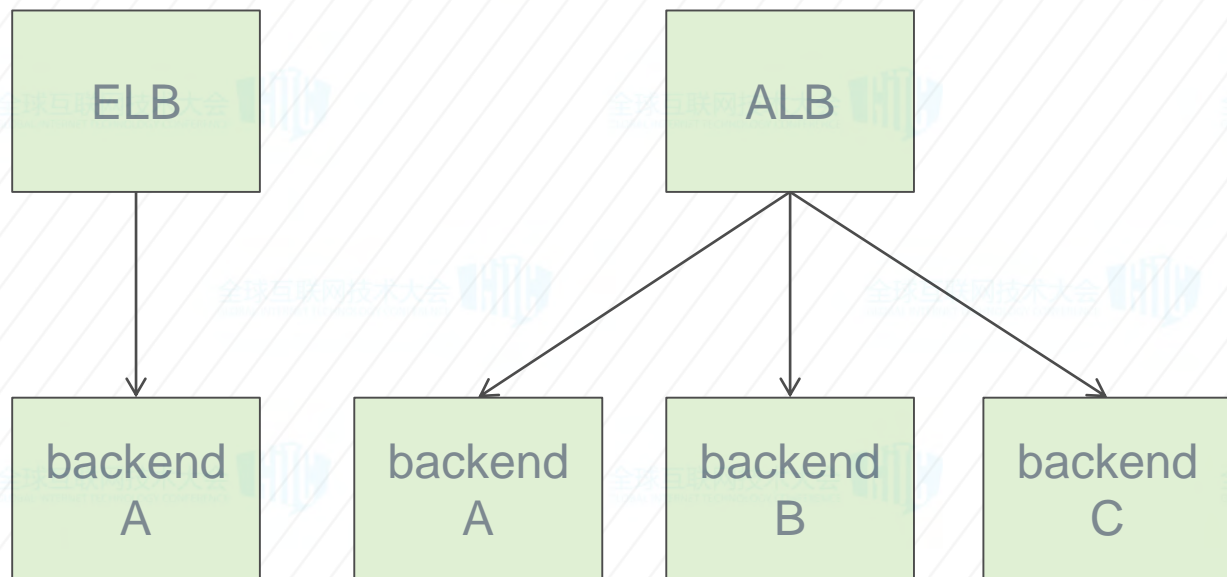
RS:

8080

提交

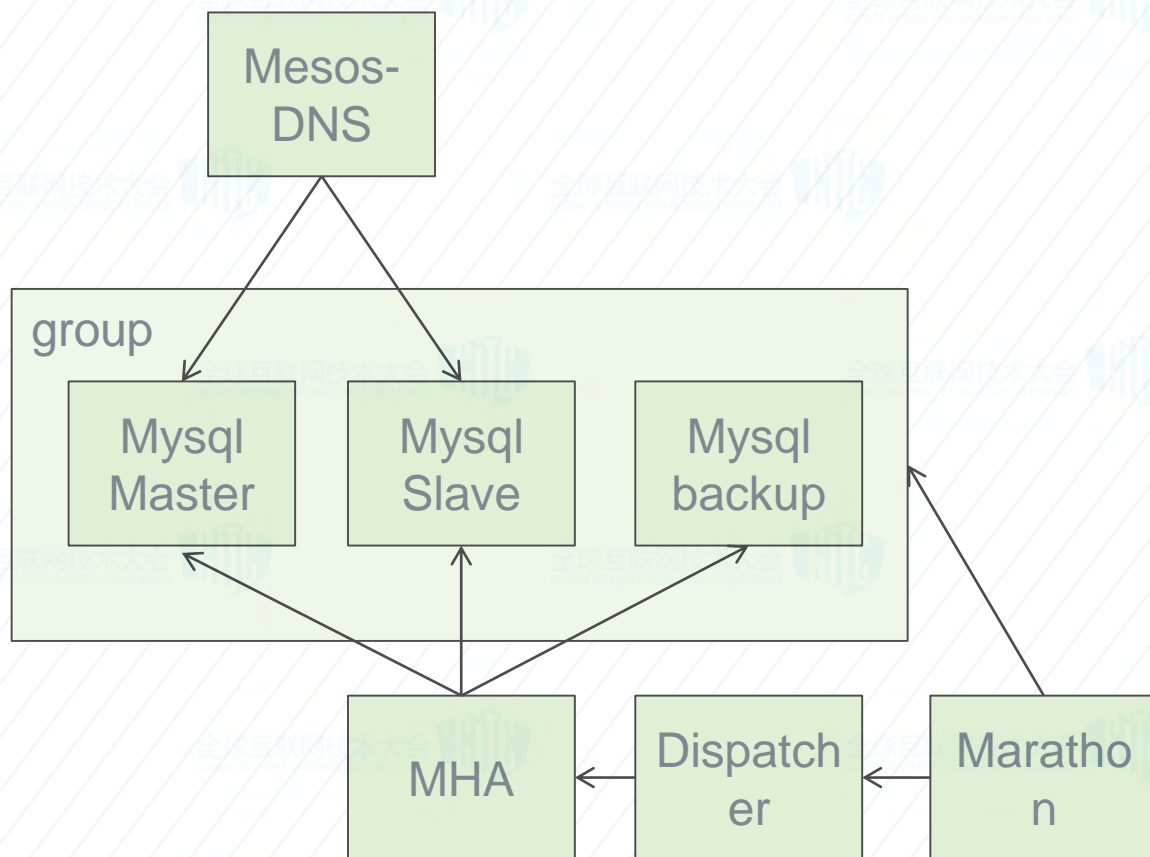
## 4

## 七层ALB



- 自动配置
- 动态更新upstream
- 根据qps自动伸缩

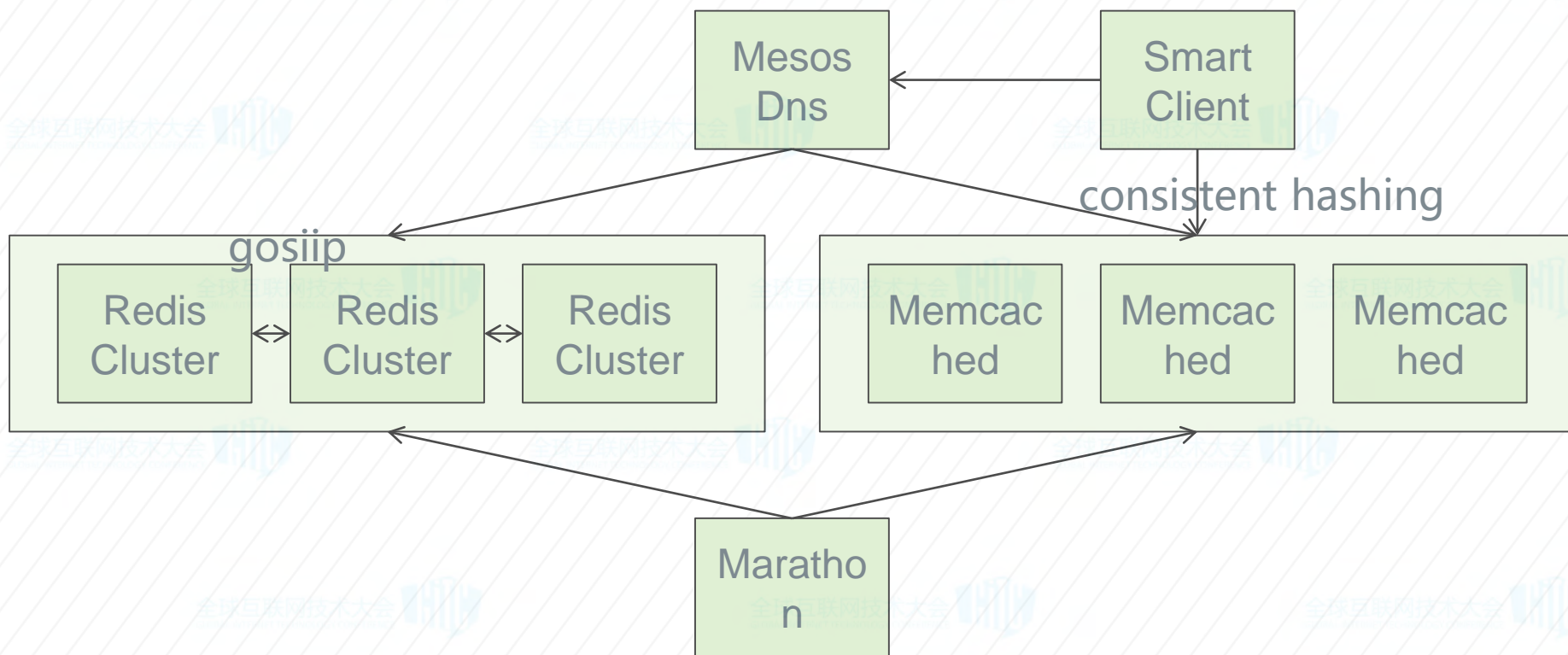
- 旁路模块接收 Marathon Events，触发自动故障处理
- 主库故障自动切换
- 备库同步复制，用于故障实例恢复数据
- Mesos-dns服务发现





## 6

## ElastiCache





# 未来规划

## 未来规划

补充完善有状态组件——Kafka、Hbase等

集成标准化CI/CD

强化故障自愈和自动容灾能力

网络设计的再优化——SDN



WECHAT

FAQ

