



弹性伸缩的世界

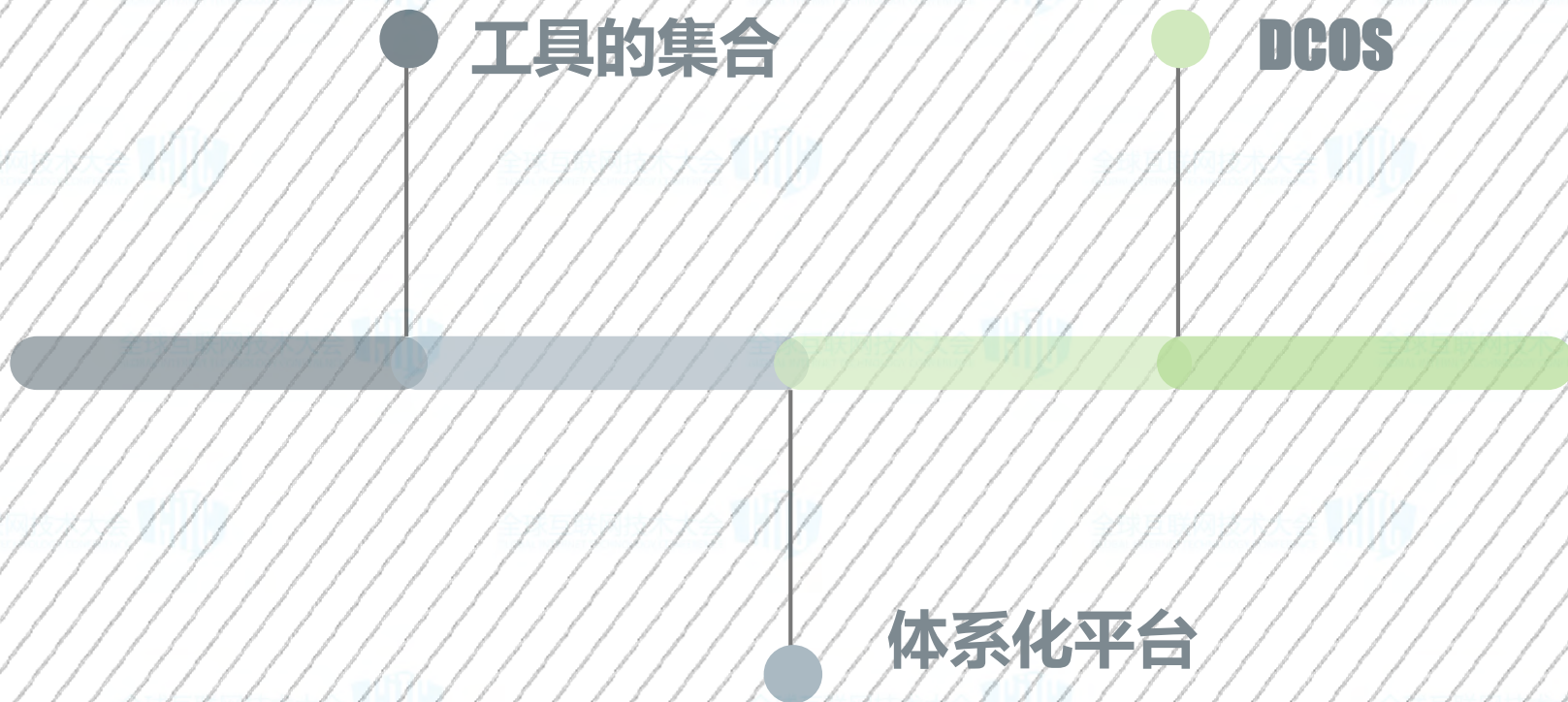
孙寅，现在小米负责运维基础设施、基础平台的构建

平台演进

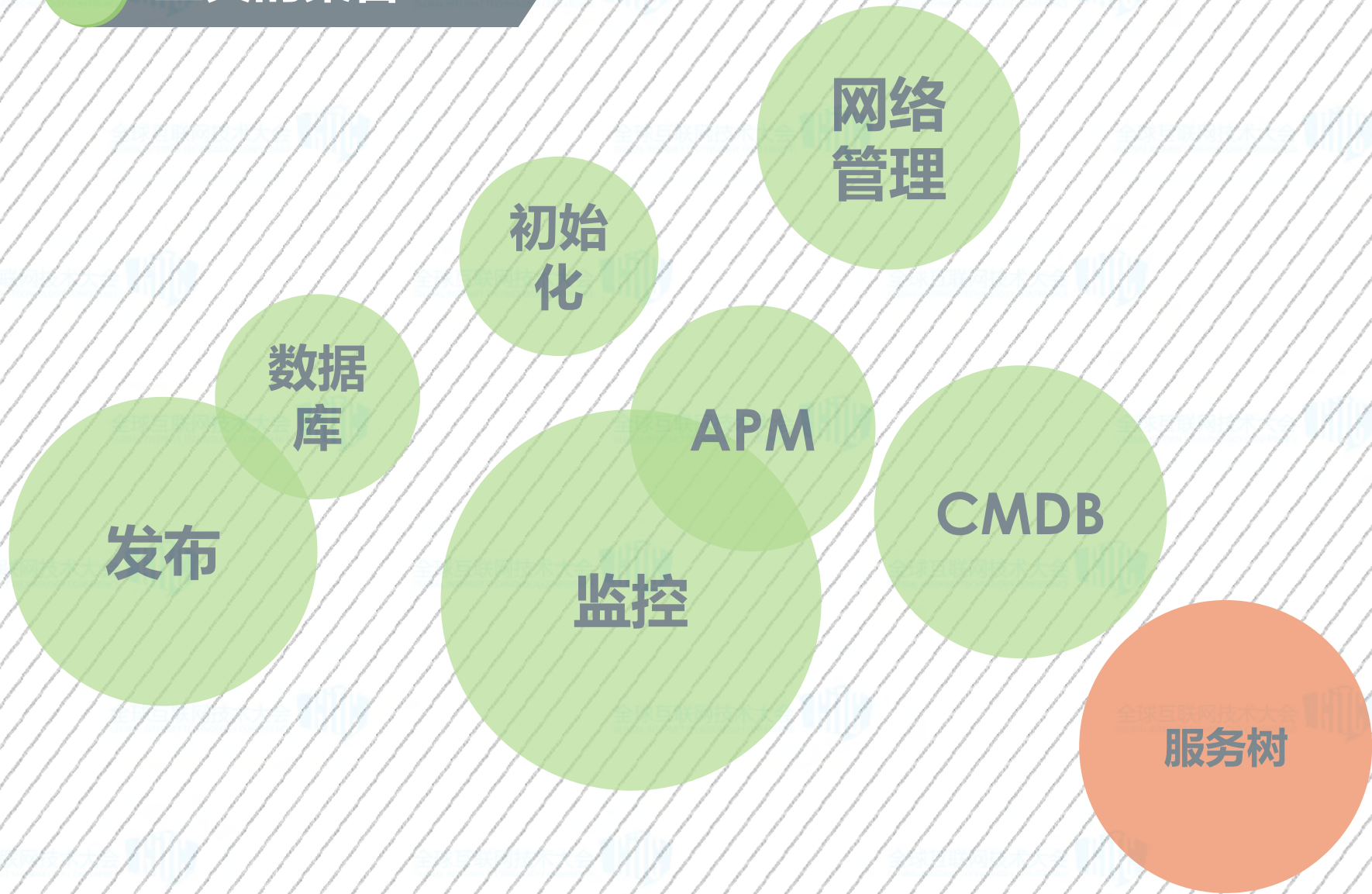
工具的集合

DCOS

体系化平台



1 工具的集合



2 体系化平台

JOBNAME

公司

部门

产品

集群

服务组

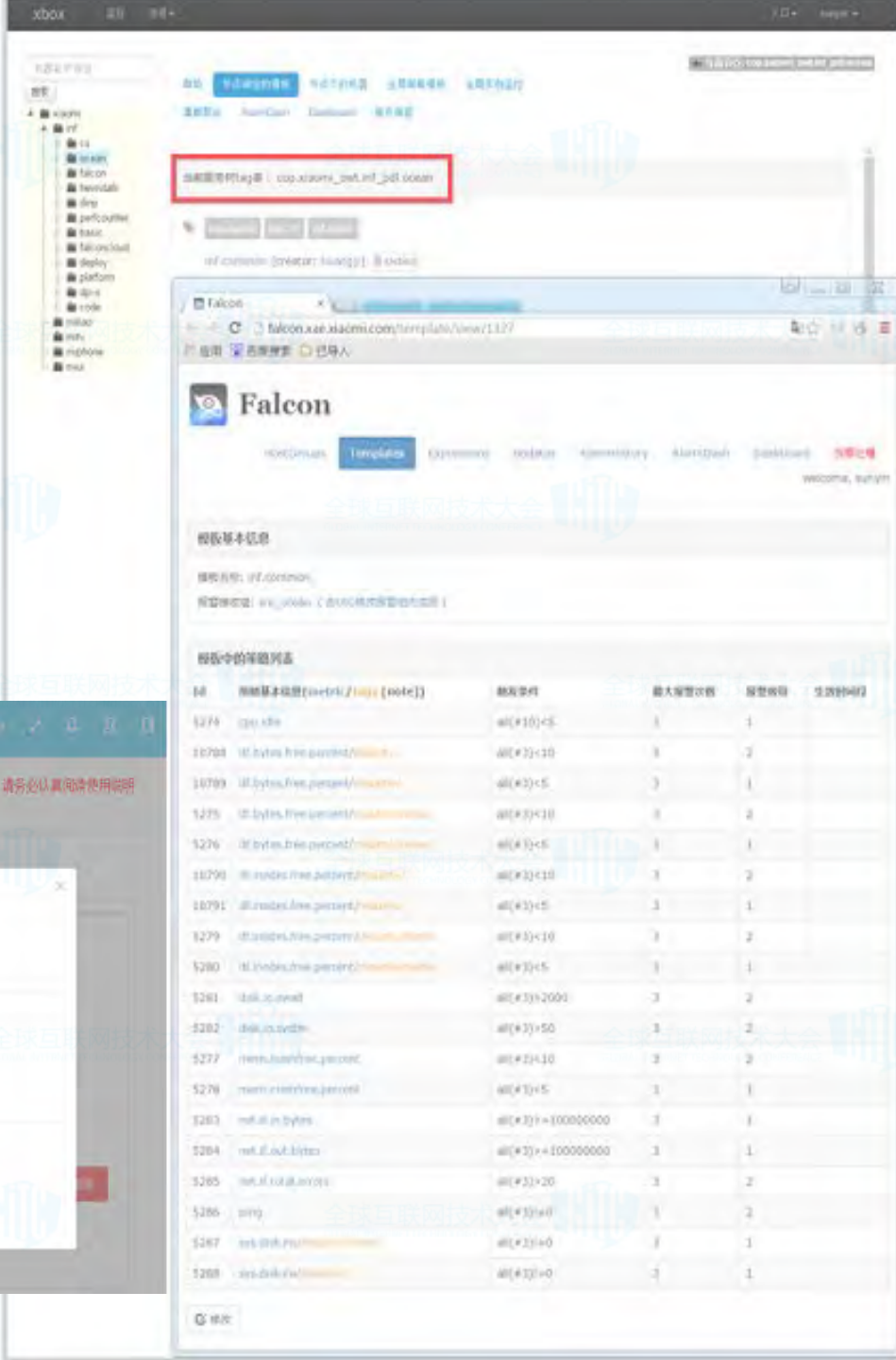
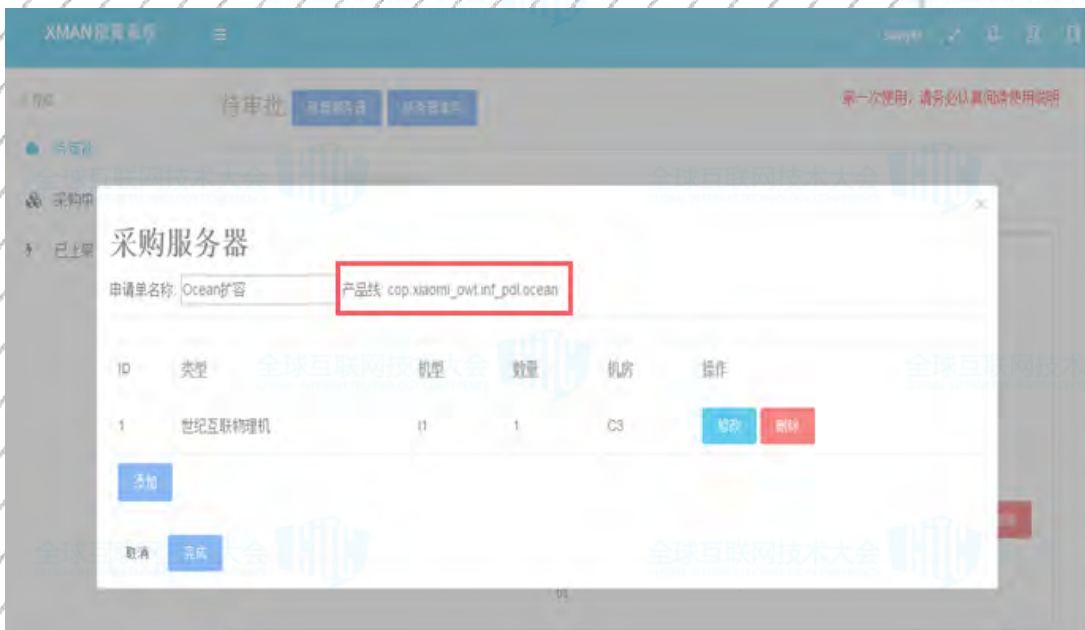
服务

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 the Falcon 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), '共享任务' (Share Task), '知识库' (Knowledge Base), '创建 job' (Create Job), '知识库的 job' (Jobs in Knowledge Base), 'ybox 群聊' (ybox Chat), 'wiki' (Wiki), '社群' (Community), and '需求反馈' (Feedback).

The main content area is divided into two sections. The left section, titled '数据源分类' (Data Source Classification), shows a tree view of data sources including 'Xiaomi', 'Mi TV', 'Mi UI', 'Mi Phone', 'Mi Mix', 'account3', 'admob', 'baw', 'com', 'computing', 'data', 'demo', 'hadoop', 'ia', 'ja', 'in', 'sfs', 'micloud', 'miliao', 'miostat', 'offline', 'op', 'other', 'qa', 'sa', 'relation', 'search', 'sec', 'securitycloud', 'smart', 'storage', 'vosp', 'vuo', 'vuv', 'x2d', 'xcache', 'xiaohuang', and 'xud'.

The right section, titled '服务列表' (Service List), shows a table of services. The first service, 'job-userconnect_service, userconnect_cluster-production-ig', is highlighted with a red box. The table columns include '服务名称' (Service Name), '状态' (Status), '实例数' (Instance Count), and '实例统计' (Instance Statistics).

Below the service list, a detailed view of a specific alert is shown. The alert is titled 'Falcon' and has a URL of 'falcon.xae.xiaomi.com/template/view/1654'. The alert details include:

- Id:** 7066
- Alert Name:** agent.lg.dreamhet.master.ticket.verify.num/cluster=production-ig,cpu=xiaomi,job=userconnect_prod+release,pb=account_service+userconnect
- Alert Description:** [lg.dreamhet.master.failed.]
- Alert Type:** alert
- Alert Value:** all(#1)<25
- Alert Count:** 9999
- Alert Status:** 1
- Alert Time:** 06:00-23:59

The alert details are displayed in a table format with columns for 'Id', 'Alert Name', 'Alert Description', 'Alert Type', 'Alert Value', 'Alert Count', 'Alert Status', and 'Alert Time'.

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

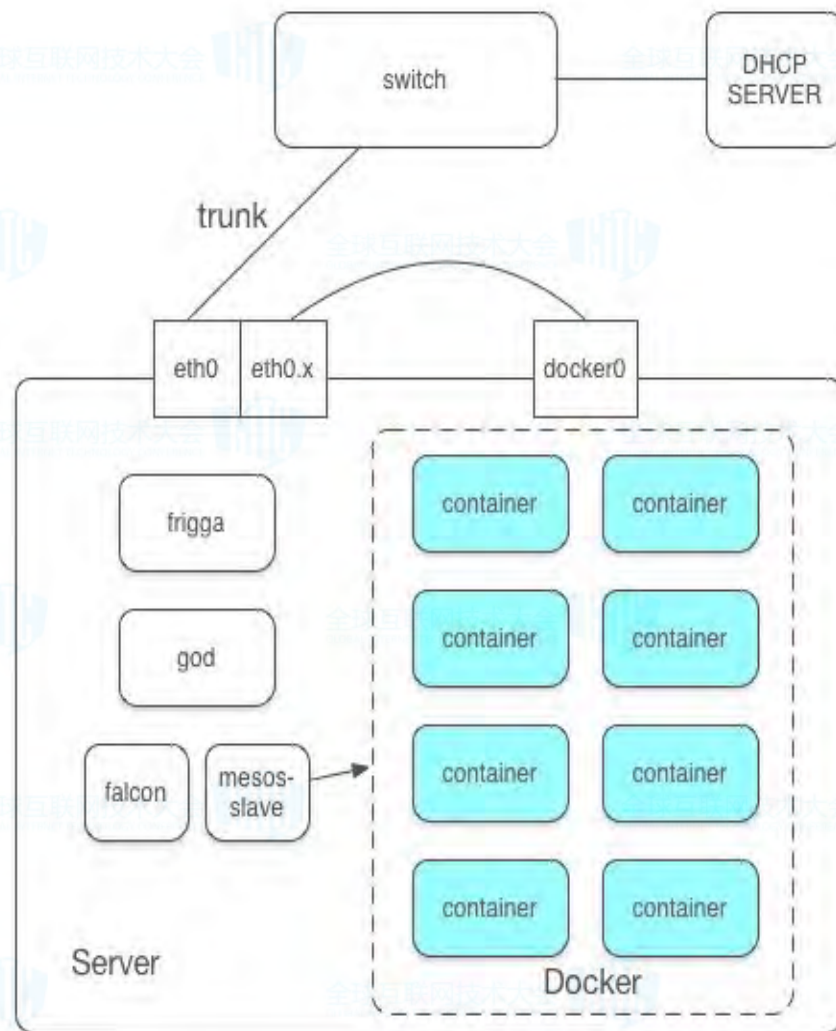
● 文件系统

● 资源隔离
资源汇报

● Docker-Init

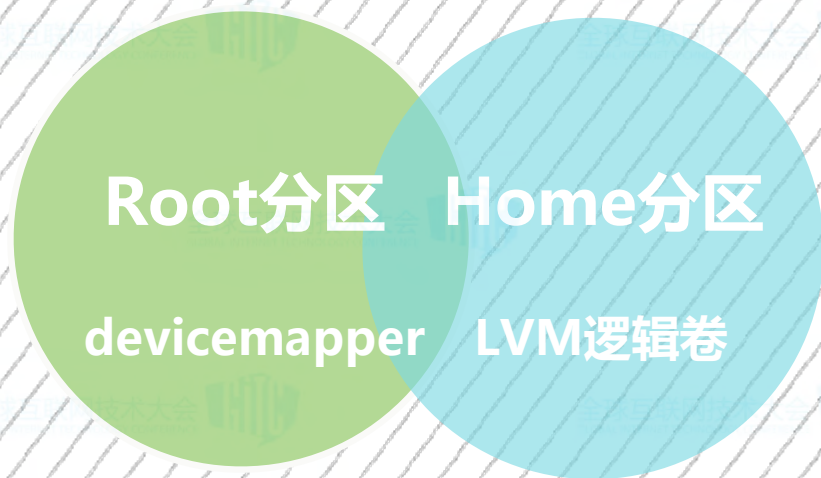
1 网络

- 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



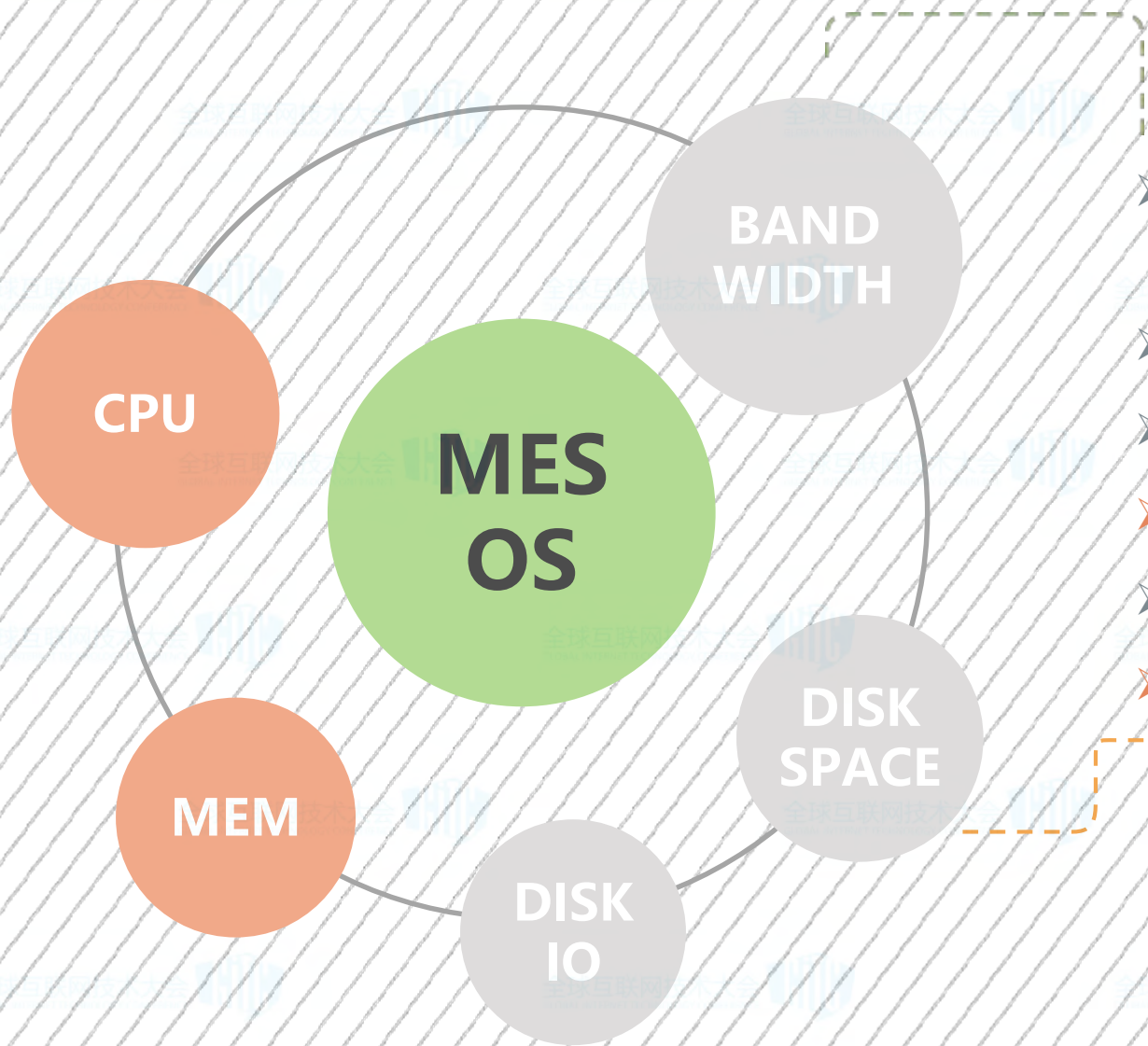
目标

- 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



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

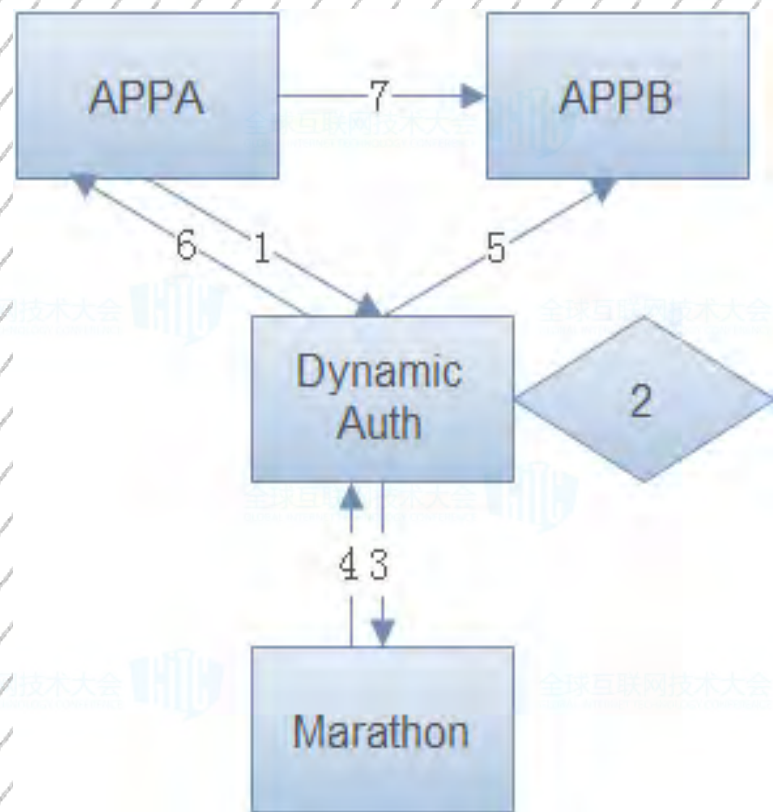
Redis Cluster(Gossip)

Other

服务发现

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_pdi.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平均值

伸缩配置

实例数 5 实例数 1 实例增 1

上限: 下限: 减步 长:

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

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 ?

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

新建ELB

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

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

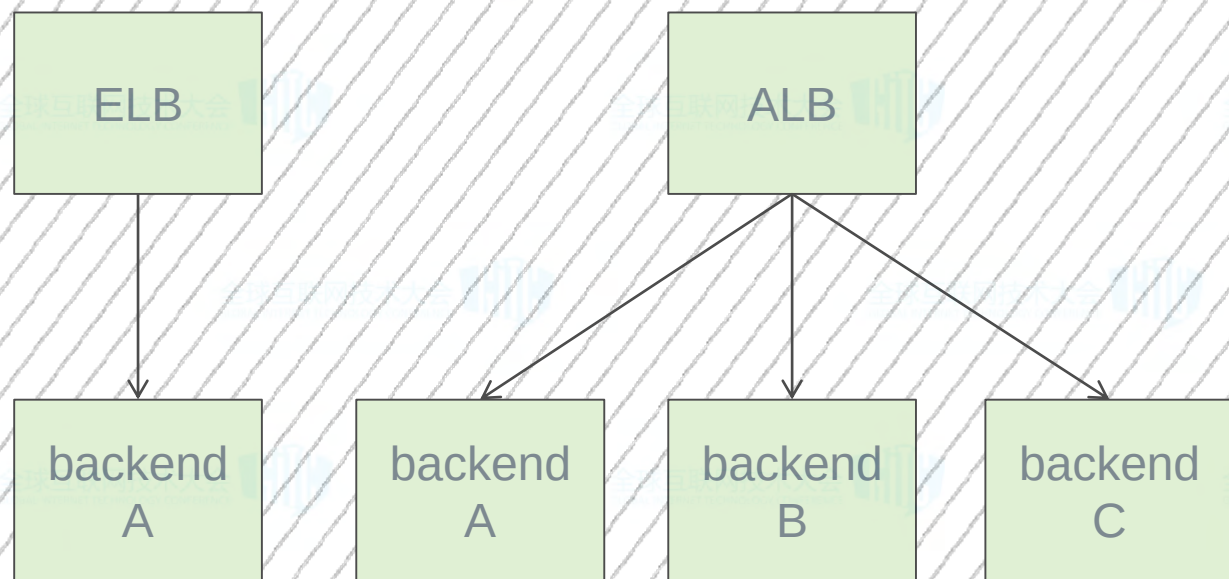
RS:

8080

提交

4

七层ALB

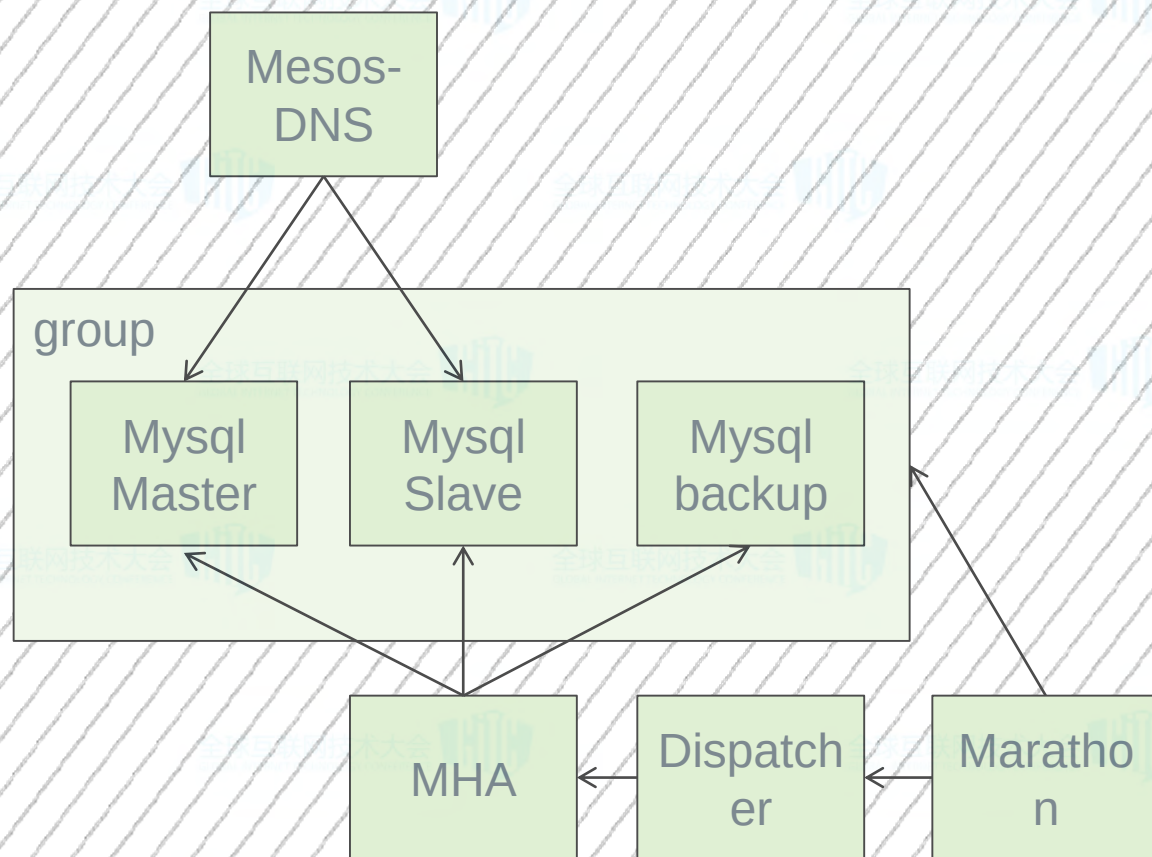


- 自动配置
- 动态更新upstream
- 根据qps自动伸缩

5

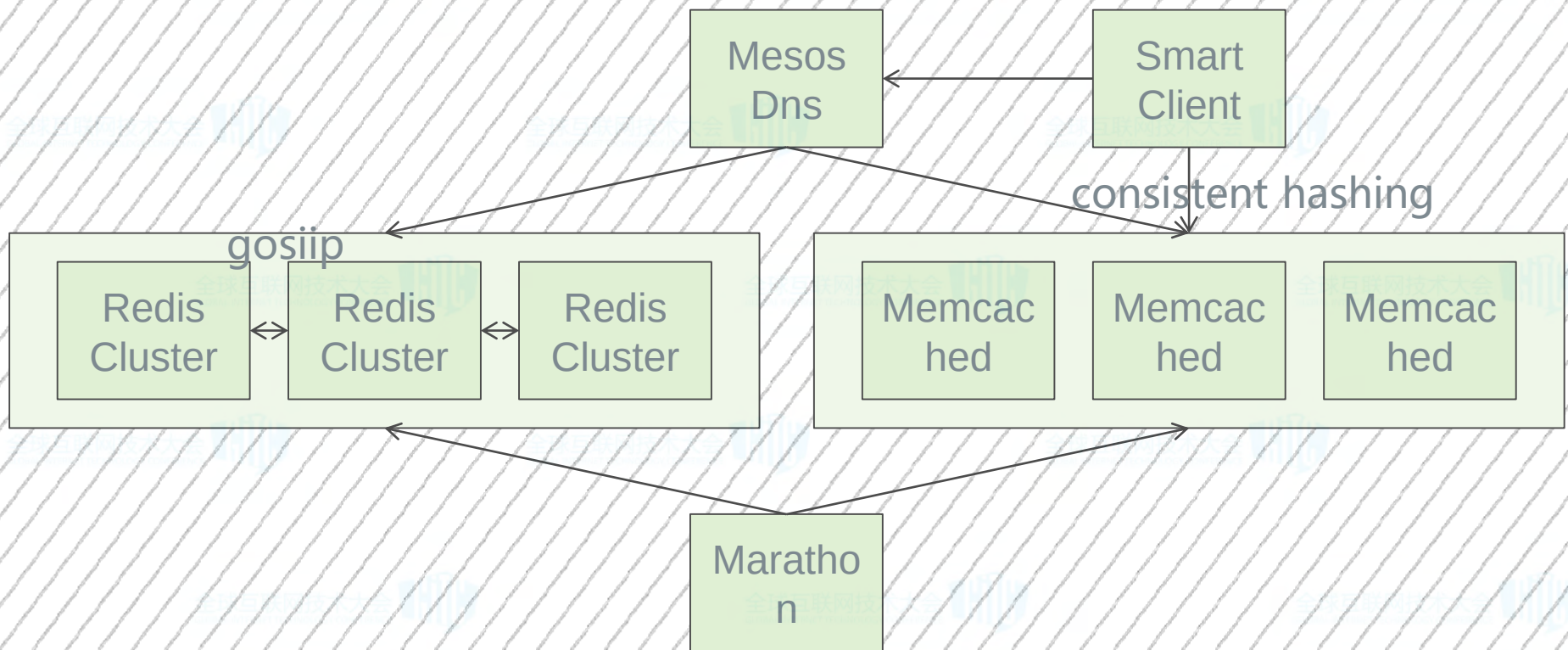
RDS

- 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

