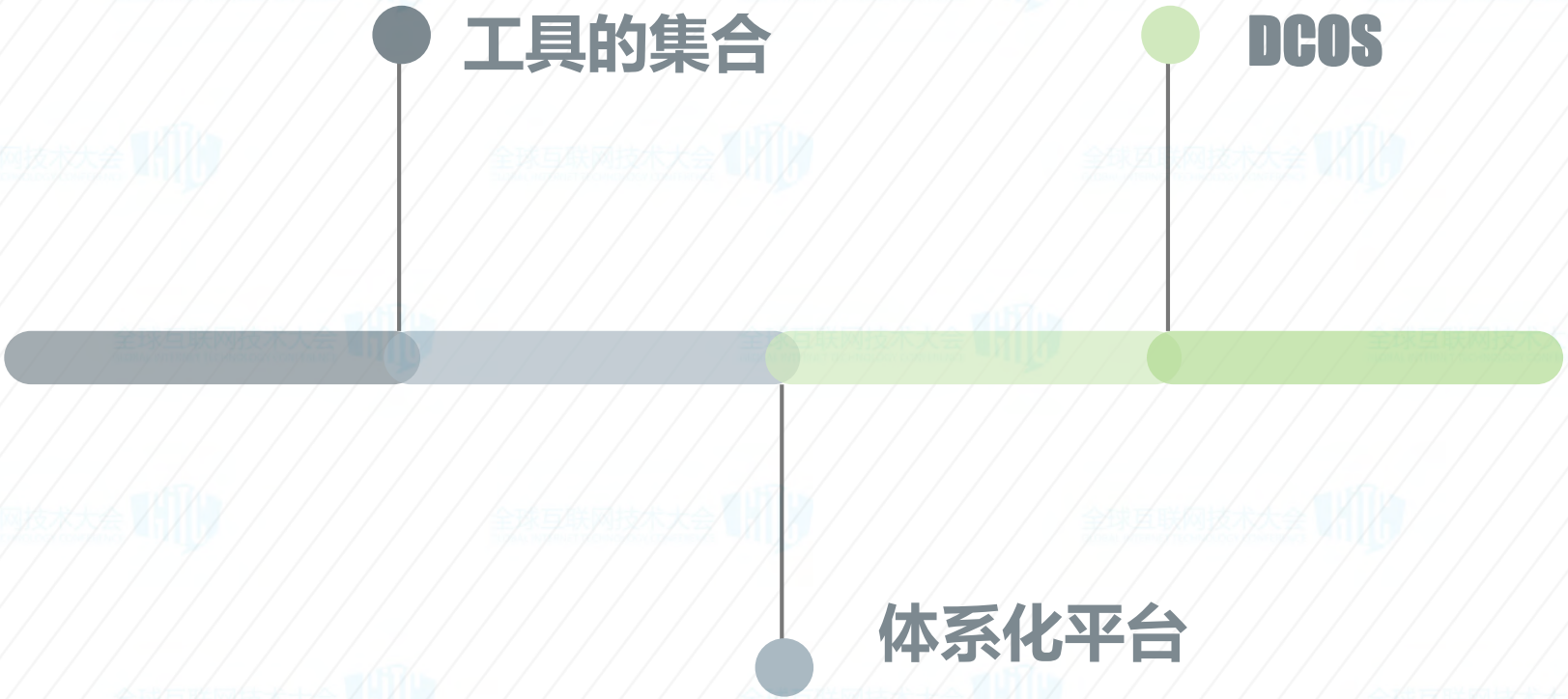
弹性伸缩的世界

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

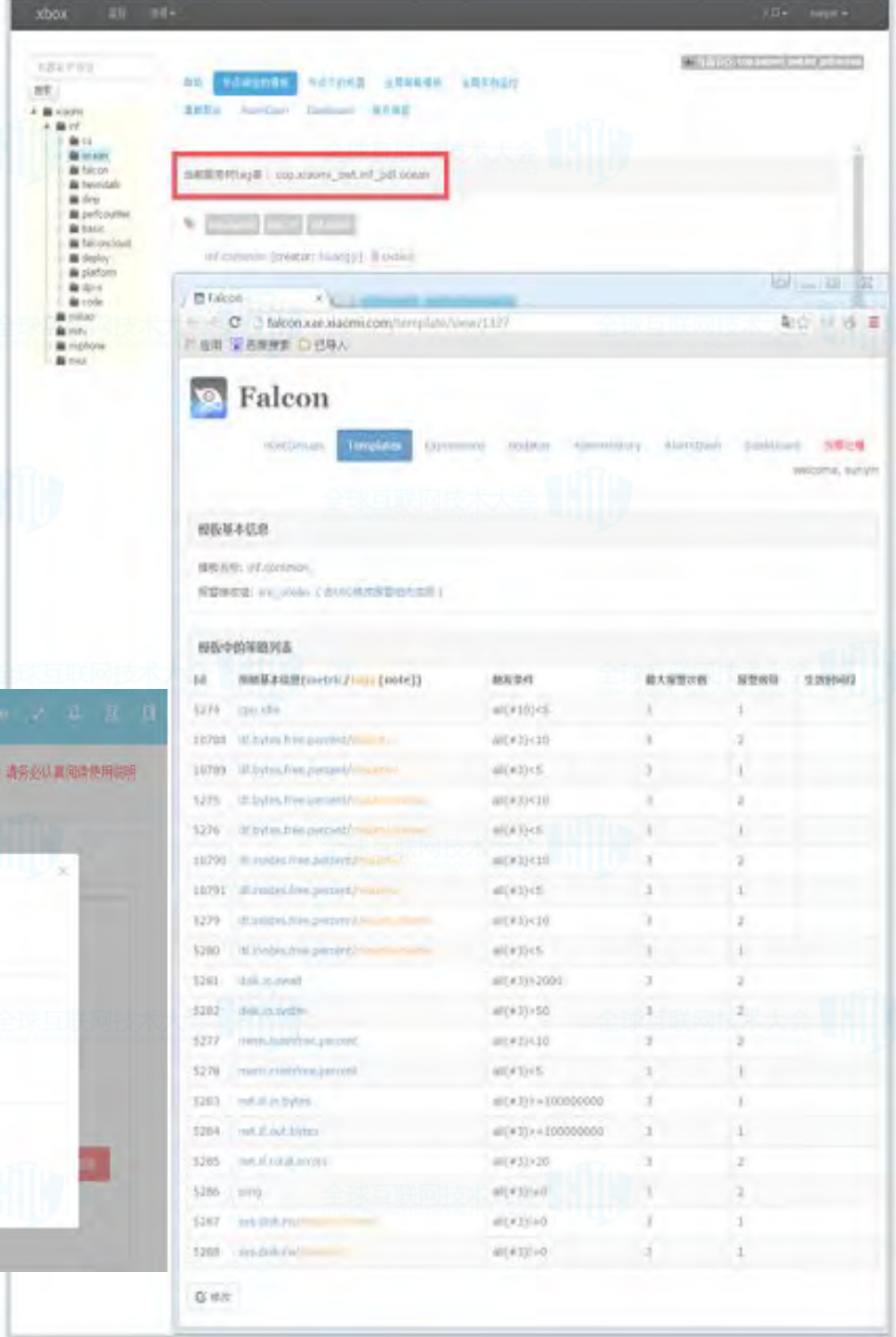


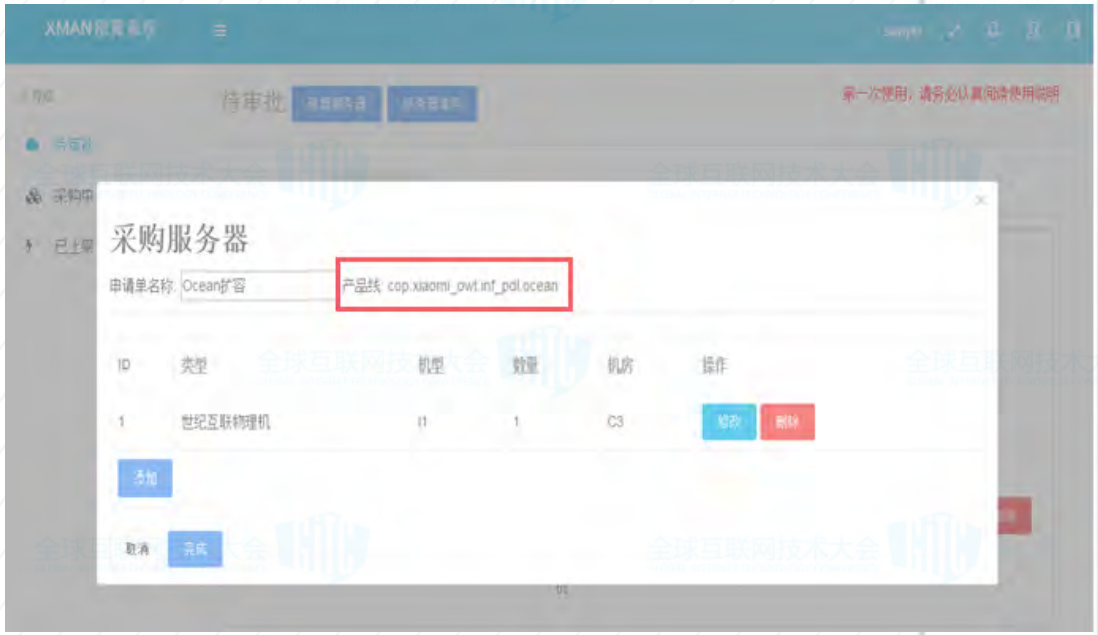
1 工具的集合

2 体系化平台

2 体系化平台

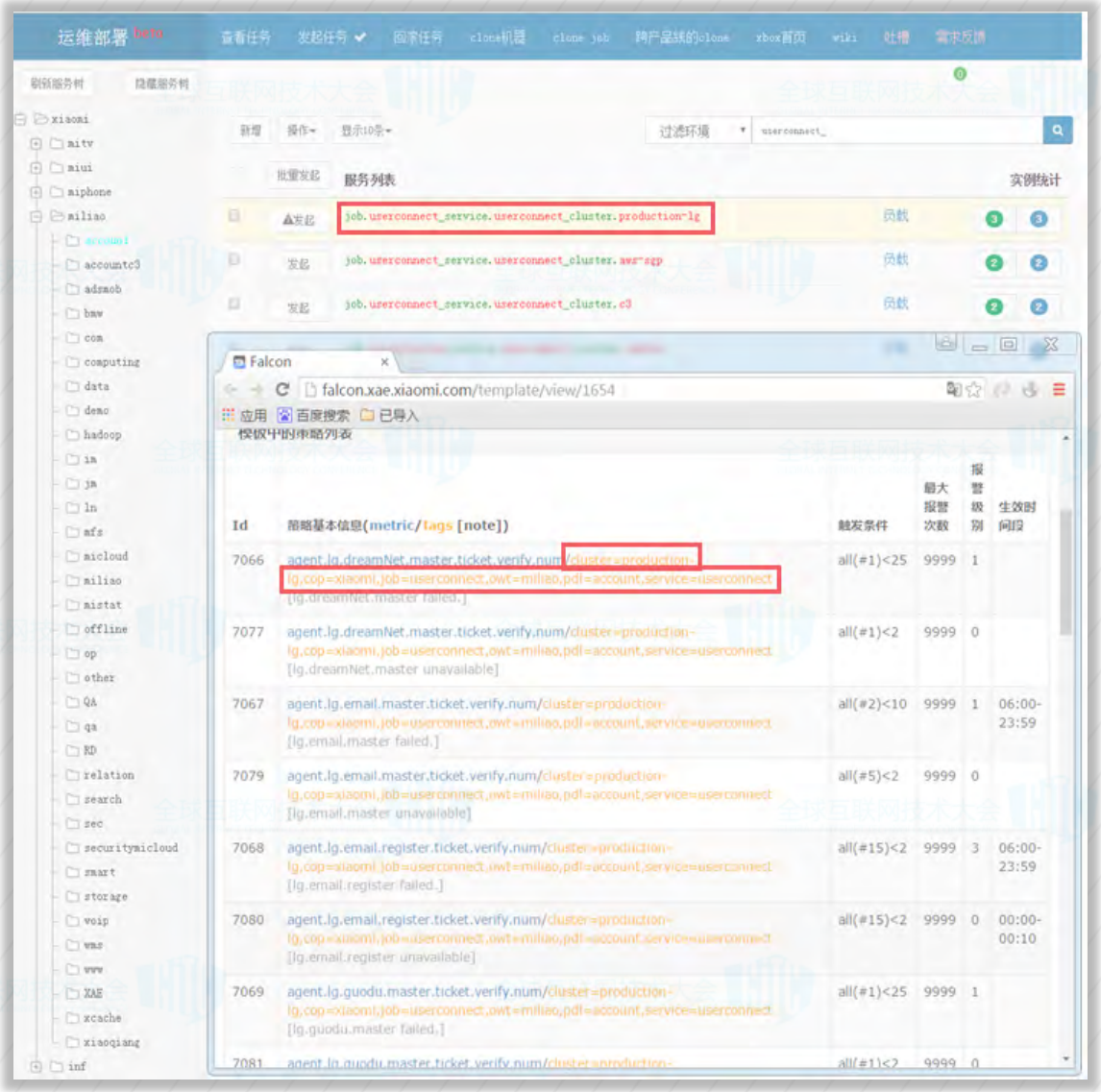
主机交付即监控



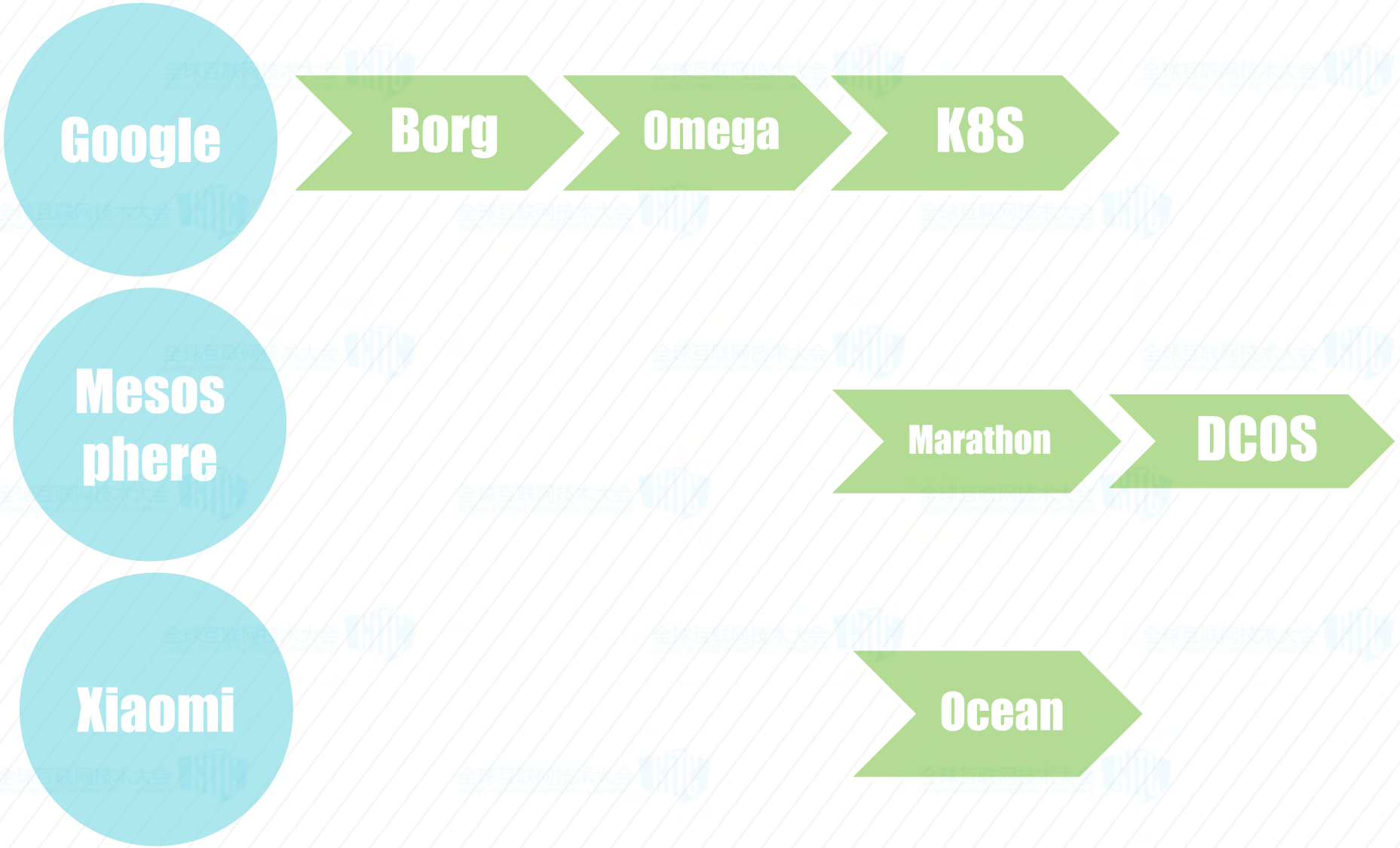


2 体系化平台

发布即监控



3 DCOS


Google
Borg
Omega
K8S
Mesos phere
Marathon
DCOS
Xiaomi
Ocean

核心技术栈

Mesos+Marathon +Docker

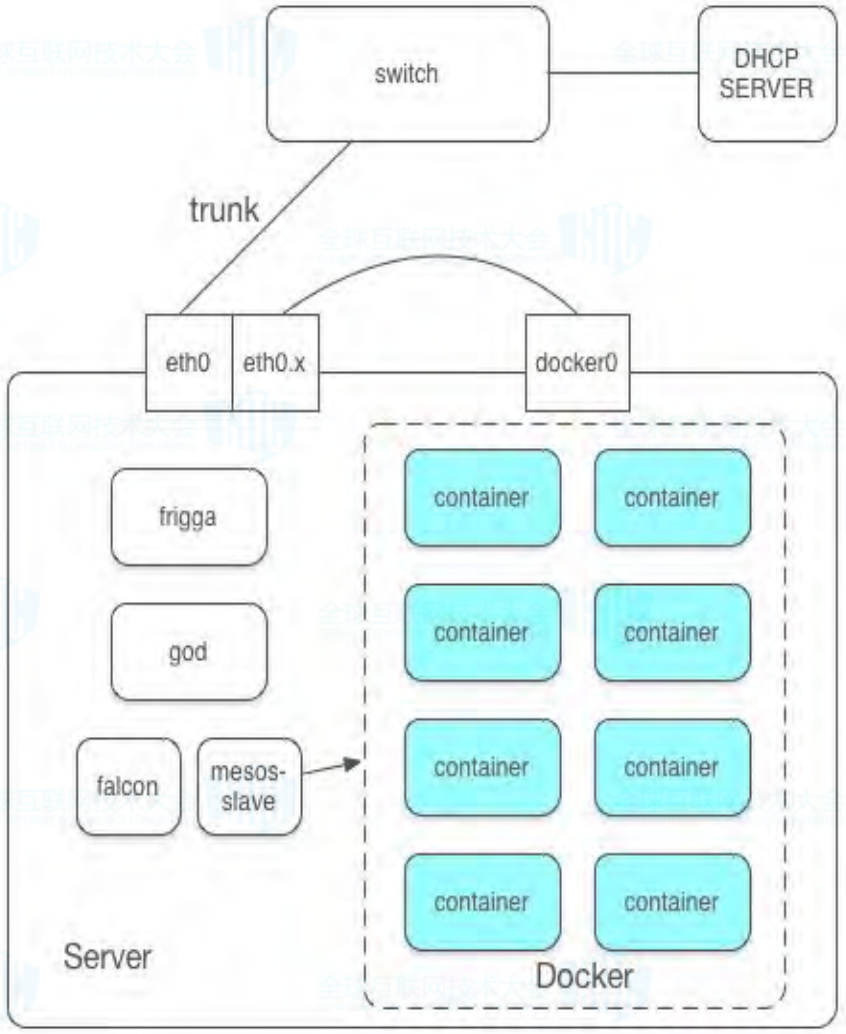
Docker

在工业环境落地

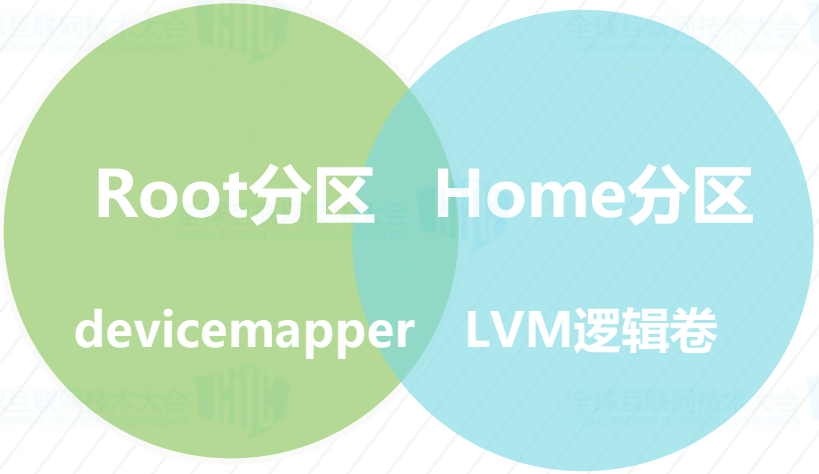
- 网络
- 文件系统
- 资源隔离
  资源汇报
- Docker-Init

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



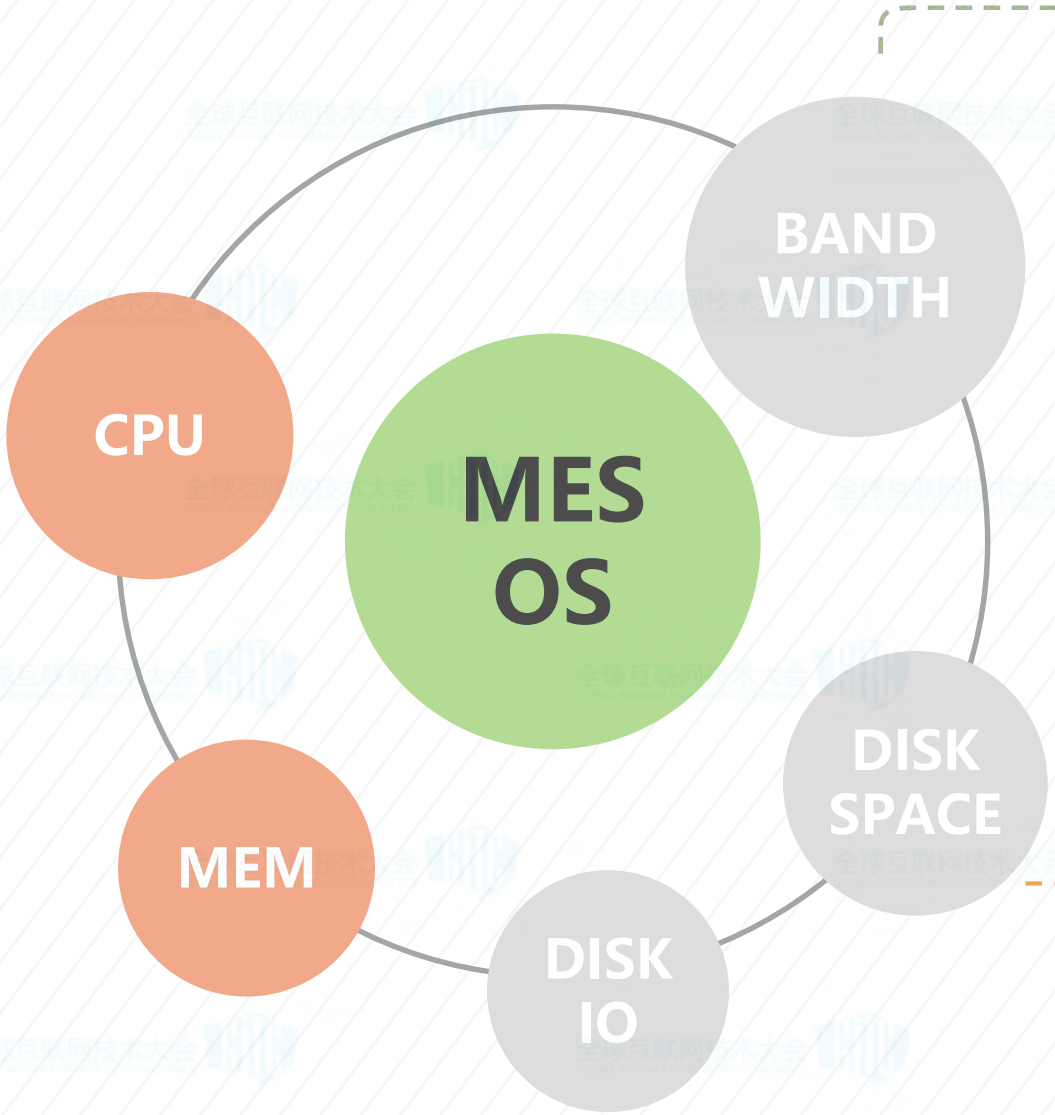
目标

- 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

4 Docker-Init

- 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

Zookeeper

Nginx Module

Mesos-DNS

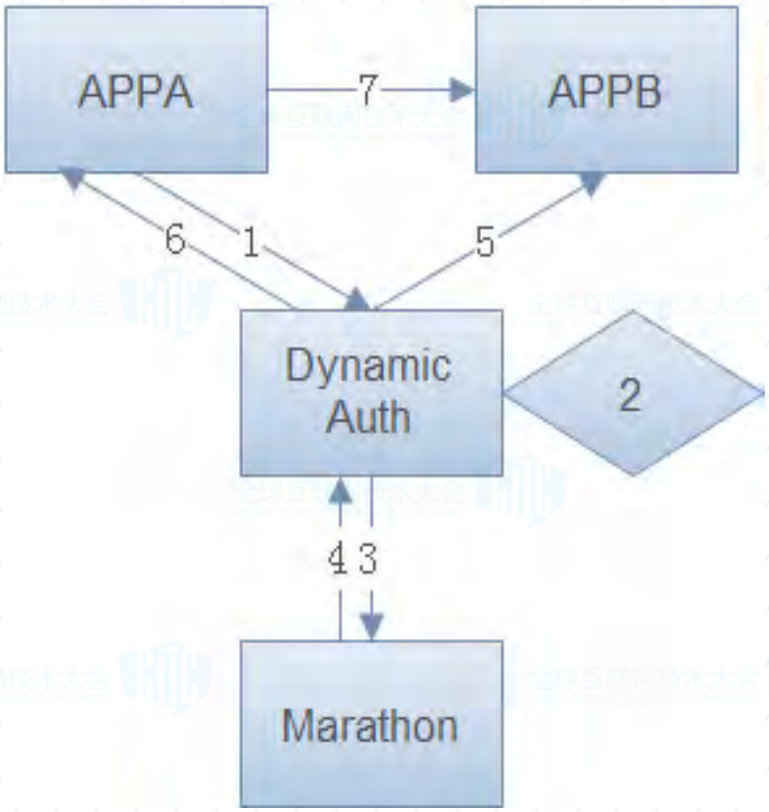
Mysql Proxy

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

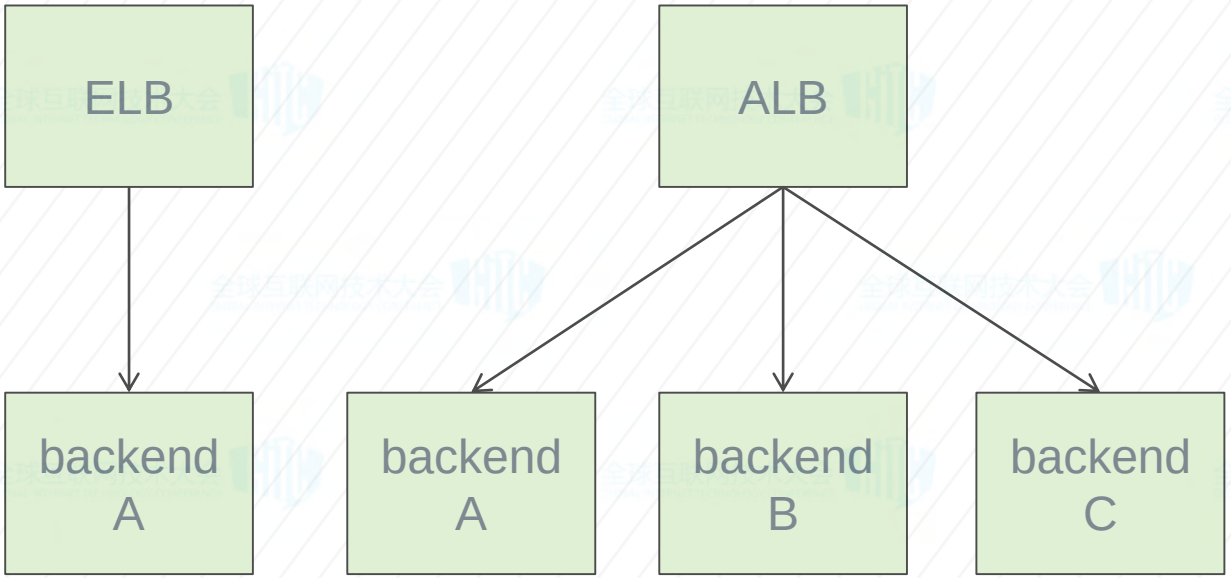
2 定时触发伸缩

- 利用Chronos调用Marathon API

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

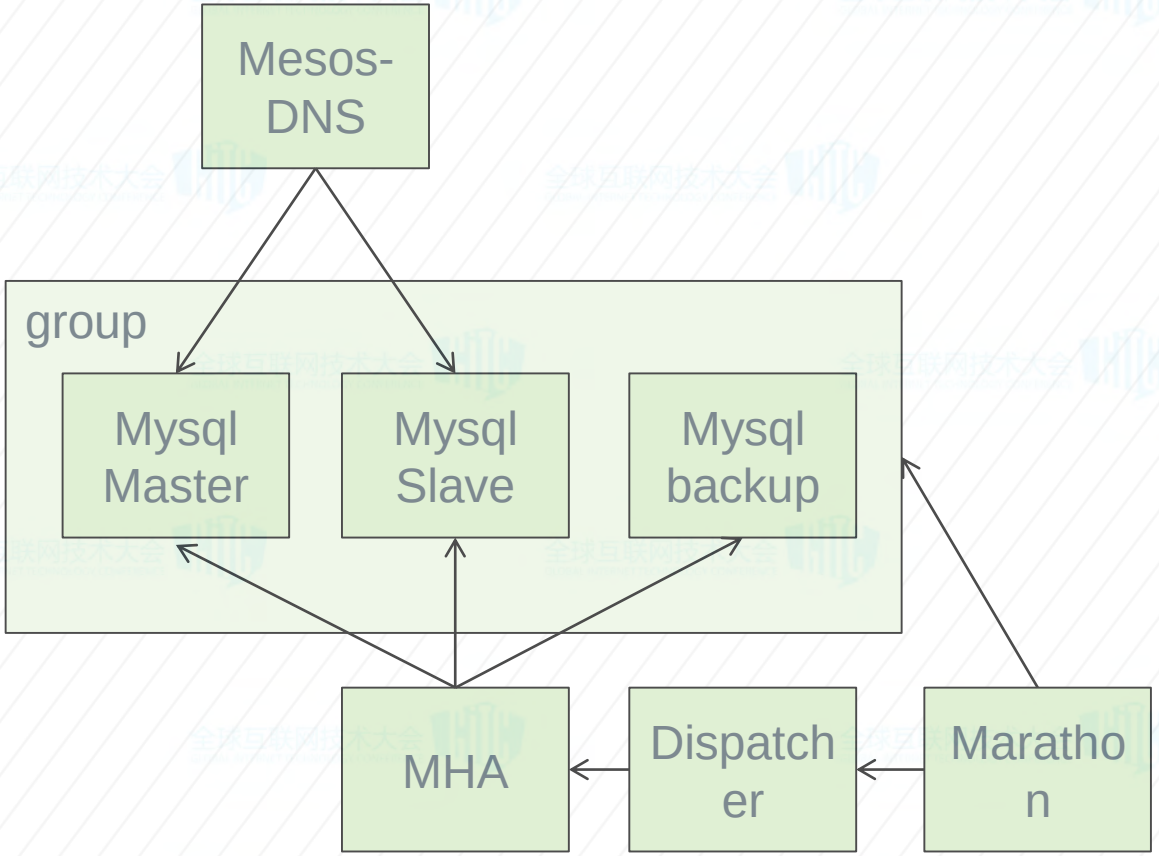
4 七层ALB



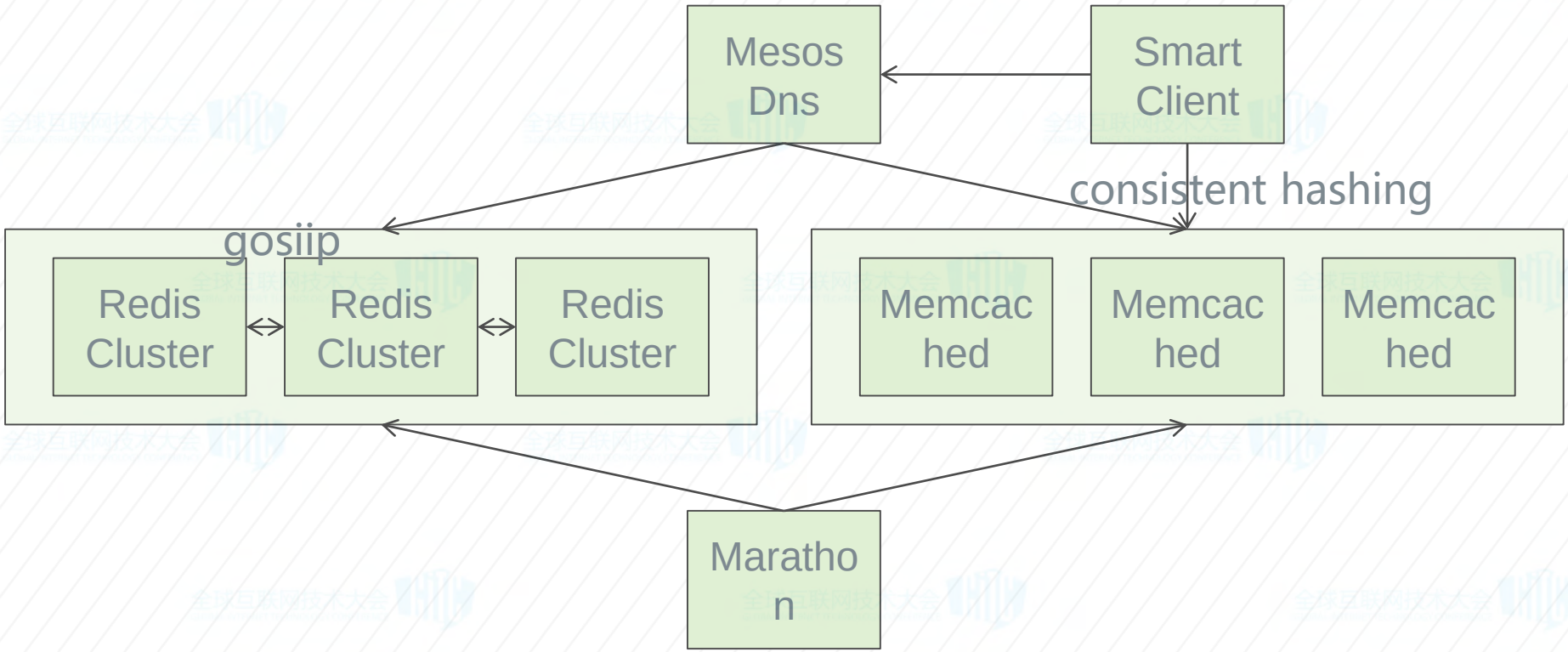
- 自动配置
- 动态更新upstream
- 根据qps自动伸缩

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

FAQ

WECHAT