新浪微博 weibo.com

微博DCP系统基于Docker容器混合云架构应用实践

@it_fuwen 微博平台研发中心

主要分享内容

新浪微博 weibo.com

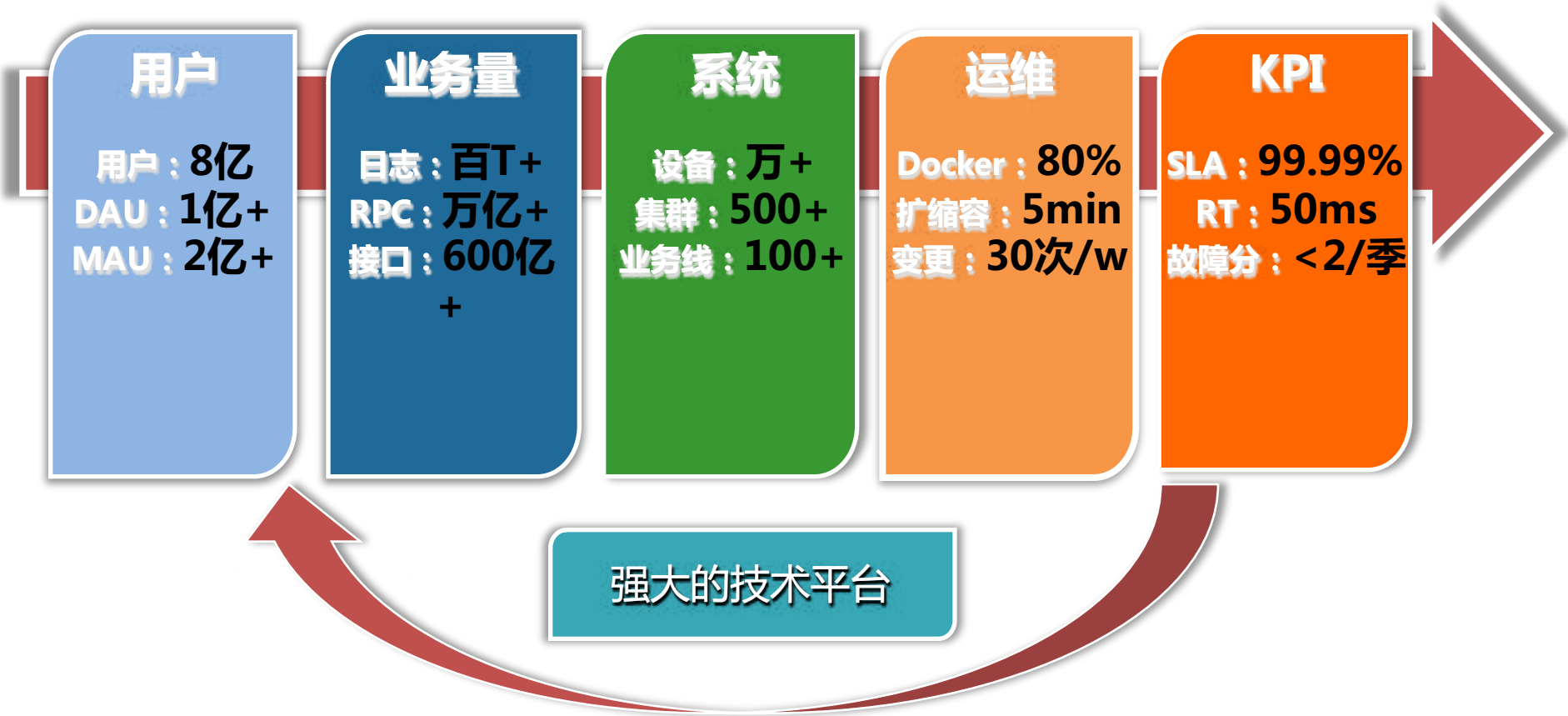
一、微博业务与混合云背景

二、DCP技术架构

三、弹性调度

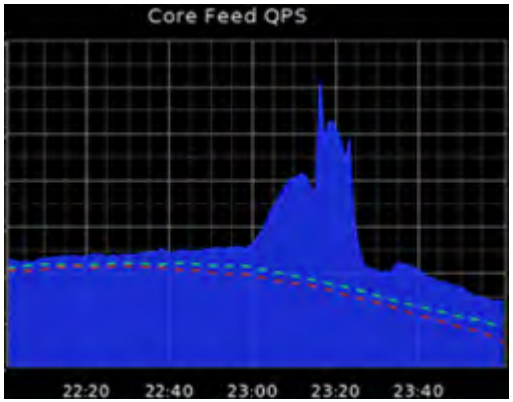
四、混合云三节实战

微博业务规模


用户
用户：8亿
DAU：1亿+
MAU：2亿+
业务量
日志：百T+
RPC：万亿+
接口：600亿+
系统
设备：万+
集群：500+
业务线：100+
运维
Docker：80%
扩缩容：5min
变更：30次/w
KPI
SLA：99.99%
RT：50ms
故障分：<2/季
强大的技术平台

微博业务现状

- 春晚峰值流量应对
  - 机架位不足，上千台服务器库存不足
  - 千万级采购成本巨大
  - 采购周期长，运行三个月只为一晚
- 李晨娱乐事件等热点突发峰值应对
  - 突发性强无预期、无准备
  - PUSH常规化，短时间大量设备扩容需求



如何10分钟内完成1000节点扩容能力？

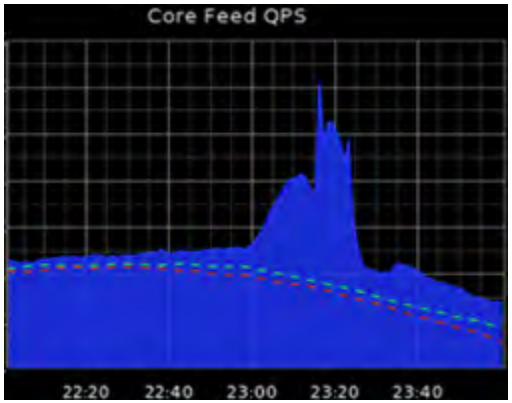
服务扩缩容流程繁琐

项目评审 设备申请 → 入CMDB → 装机 上架 → 初始化 → 服务部署 → 报修 下架

微博业务现状与解决方案

新浪微博 weibo.com

弹性快速扩缩容



混合云弹性调度可伸缩业务成本节省数倍

场景	私有云方案	公有云方案	预估总成本对比（私/公）
短期峰值应对	部分调度，部分采购	按小时付费	数十倍
日常常规部署	按月摊销	包月付费	1.0–1.2
离线计算	按月摊销	按小时付费	数十倍

配置	按量（元/小时）	包月（元/月）
4核4G，20G磁盘	1.13	302
4核8G，20G磁盘	1.77	402
4核16G，20G磁盘	3.05	602
8核8G，20G磁盘	2.25	598
8核32G，20G磁盘	6.09	1198
16核16G，磁盘20G	4.52	1207
16核64G，磁盘20G	12.17	2390
备注	小时计费按月为包月3～4倍（64G必须选择16核）	

扩无可扩？技术体系升级

产品更新迭代快，系统变更代码指数增长

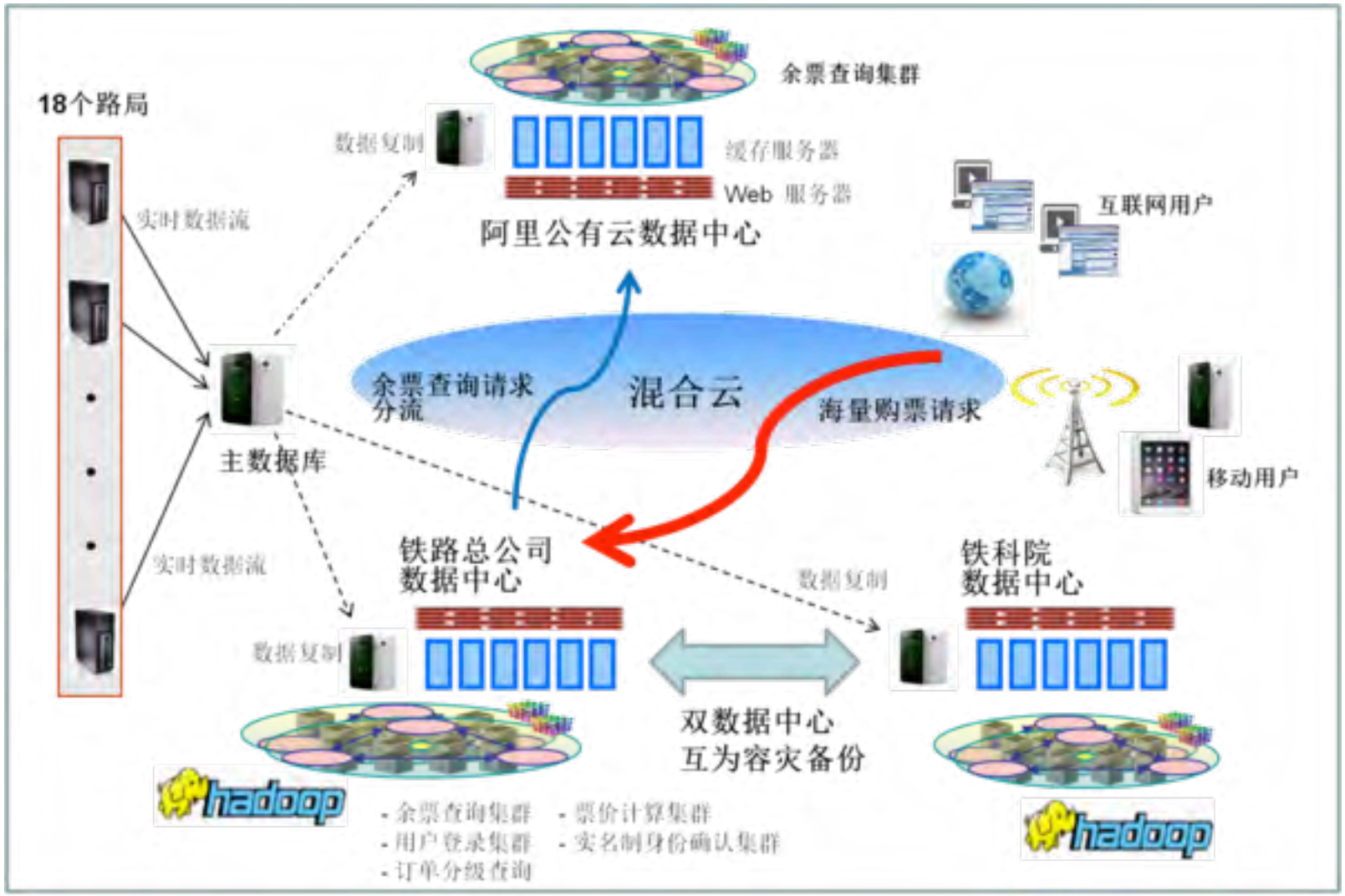
扩！扩！扩！

业界趋势

混合云趋势：安全、可扩展性、成本...

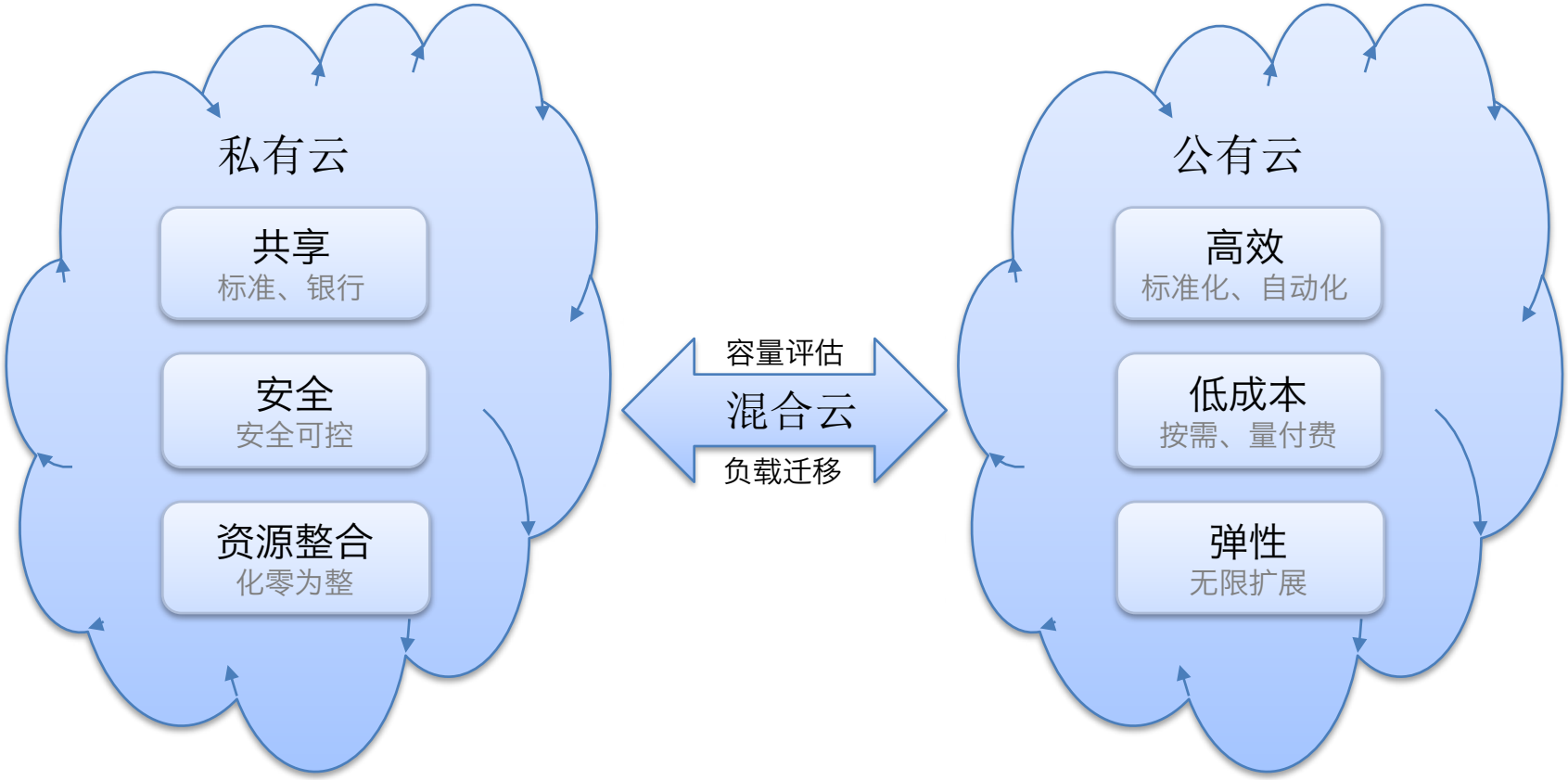
- AWS、阿里云等公有云平台趋于成熟
  - 国外Zynga、Airbnb、Yelp等使用AWS进行部署
  - 国内阿里云12306、高德、快的已部署，陌陌等部署中
  - 12306借助阿里云解决饱受诟病的春节余票查询峰值问题
- Docker、Mesos等容器新技术使大规模动态调度成为可能
  - 京东618大促借助Docker为基础的弹性云解决峰值流量问题

12306混合云案例



12306 两地三中心 混合云架构

混合云趋势



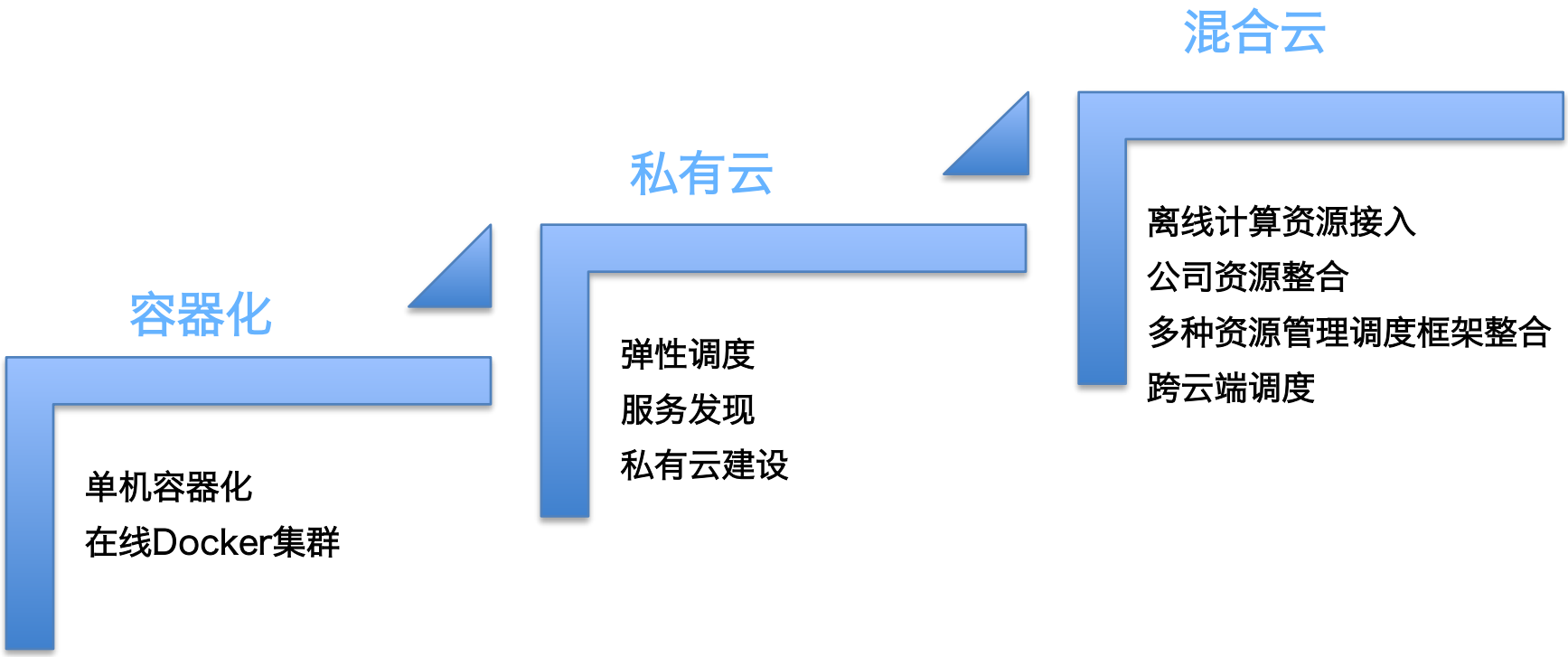
私有网络

VPC网络

弹性调度

基础设施跨云

DCP系统技术架构演进



混合云DCP技术架构

基础设施

服务发现
Nginx、motan
SLB、DNS

镜像中心
Docker
Registry

监控中心

容量评估

……

Java | PHP | … | Hadoop

Jpool | 扩缩容 | 上线发布

Web | RPC | 服务池管理 | 回滚降级

编排

Roam

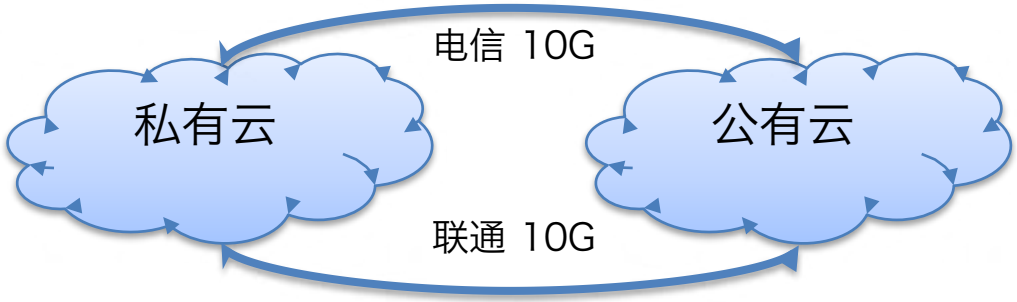
Mesos | Swarm | Docker Deamon

调度

Pluto

创建主机 | 成本中心 | 初始化

主机



混合云DCP技术栈

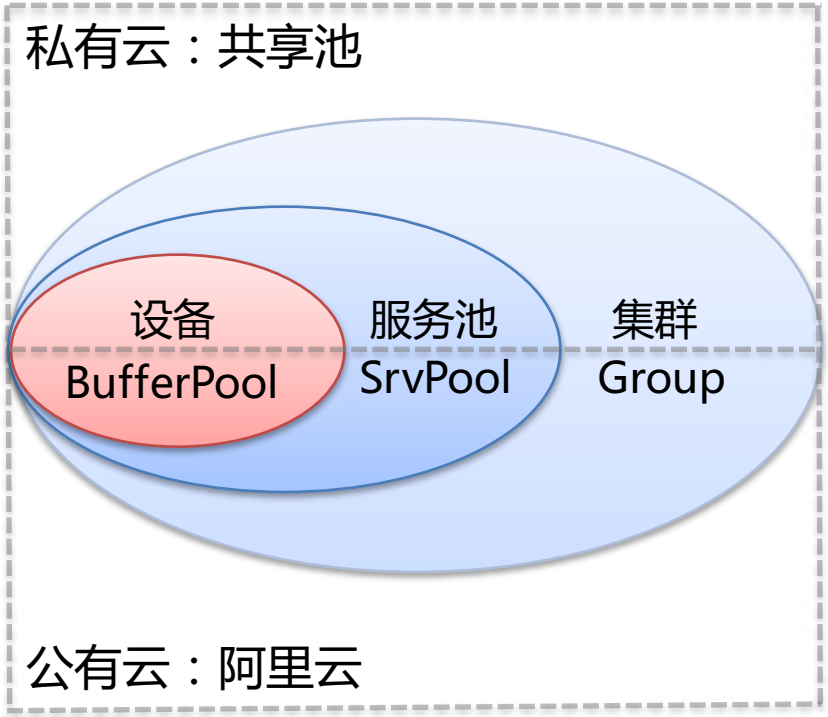
主机 or VM	私有云裸主机	公有云VM(ECS)		
OS	CentOS 6.5	CentOS 7.0		
Docker	1.6.2	Host	iptables=false	DeviceMapper
Docker Registry	V1	V2		
Swarm	1.0.0			
Mesos	0.25			
Marathon	0.13.0			
Consul	0.6.0			

混合云DCP功能模块


业务方
红包飞
MAPI
广告
有信
Feed
用户
通讯
平台架构
PAAS
Swarm
Docker 调度策略
Mesos 调度管理
容量评估
容器监控
调度监控
IAAS
共享池管理
成本核算
ECS管理
SLB管理
四七层解决方案
专线保障
公有云流量管理
审批流程
基础框架
软件安装
工程框架
安全保障
账户体系
监控体系
DNS管理
配置管理
阿里云Yum/日志
Docker Register
Docker 镜像市场
Consul 工具管理
OS升级自动化
Docker 工具体系

混合云DCP共享机制

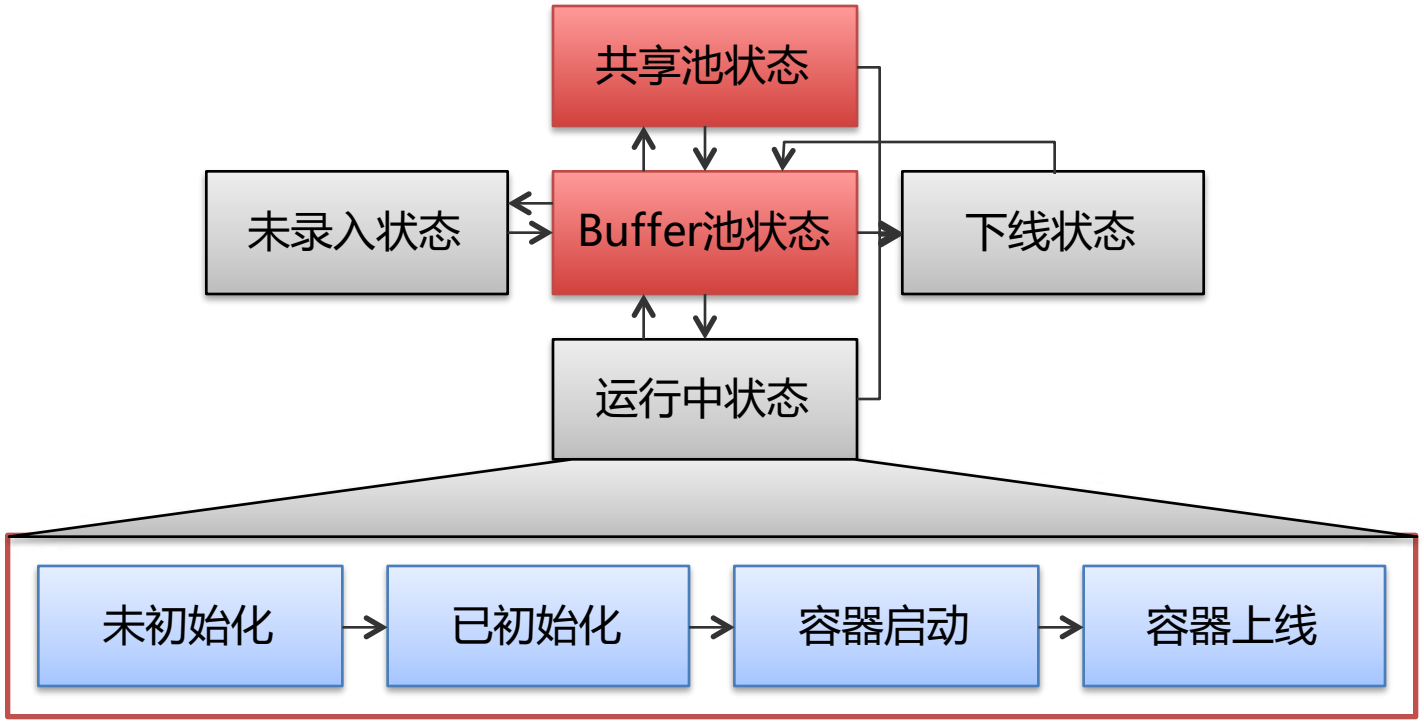
- 核心思想：借鉴于银行的运作机制
- 弹性方案：内网共享池 + 公有云
- 服务：IP + Port



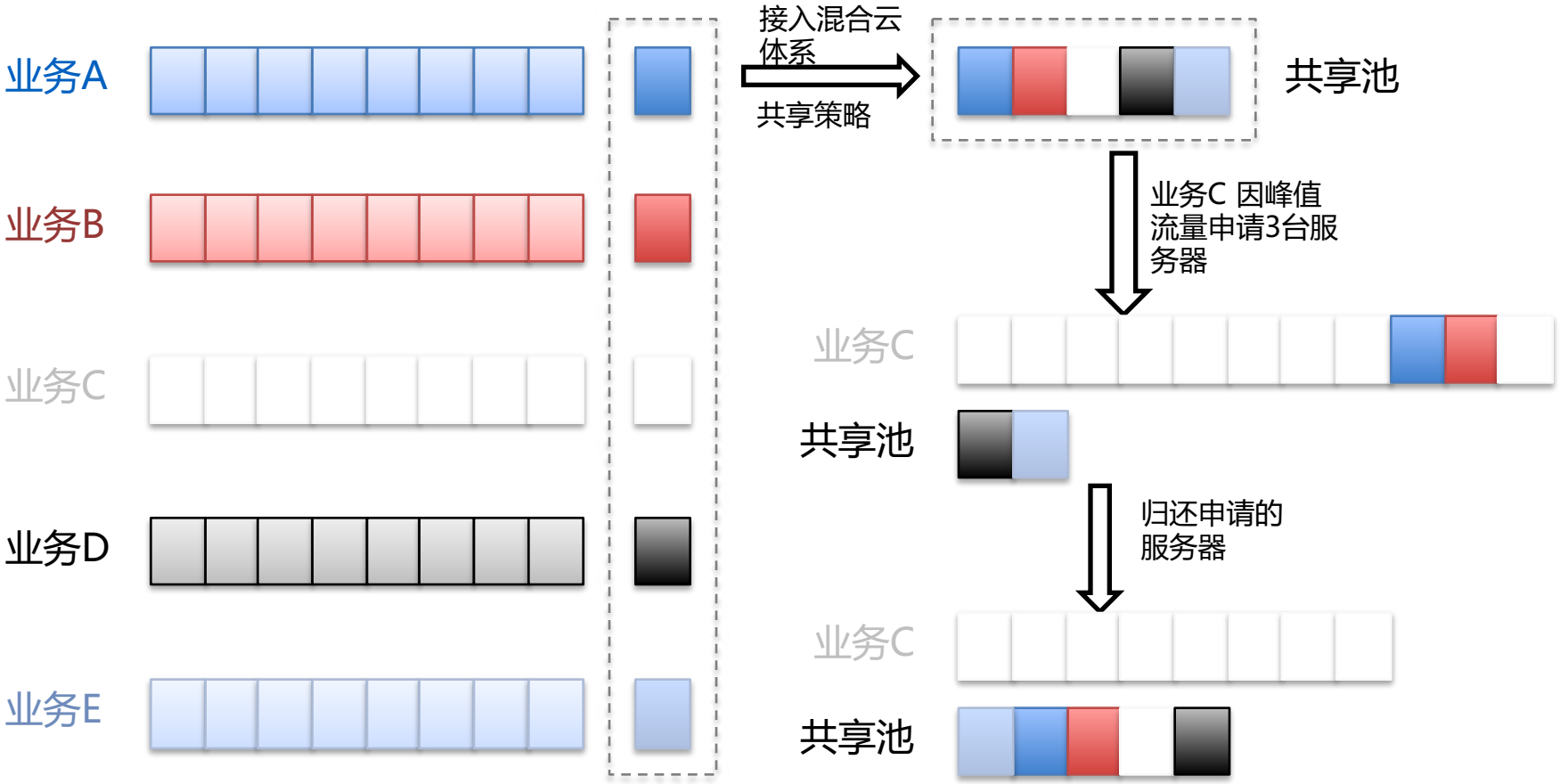
层级关系

- DCP：分为多个集群
- 集群：为独立平台，对应业务线
  - 集群内：自由调度(跨Pool)
  - 集群外：配额调度
- 服务池：同一业务线的同构服务
- 设备：共享池 + buffer池 + ECS

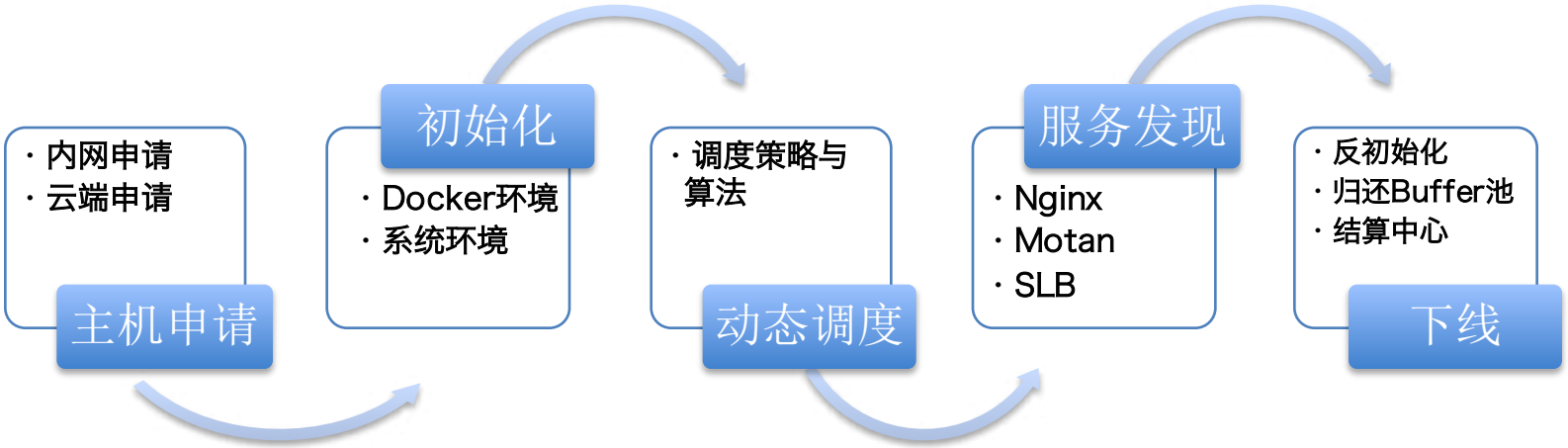
混合云DCP容器生命周期管理

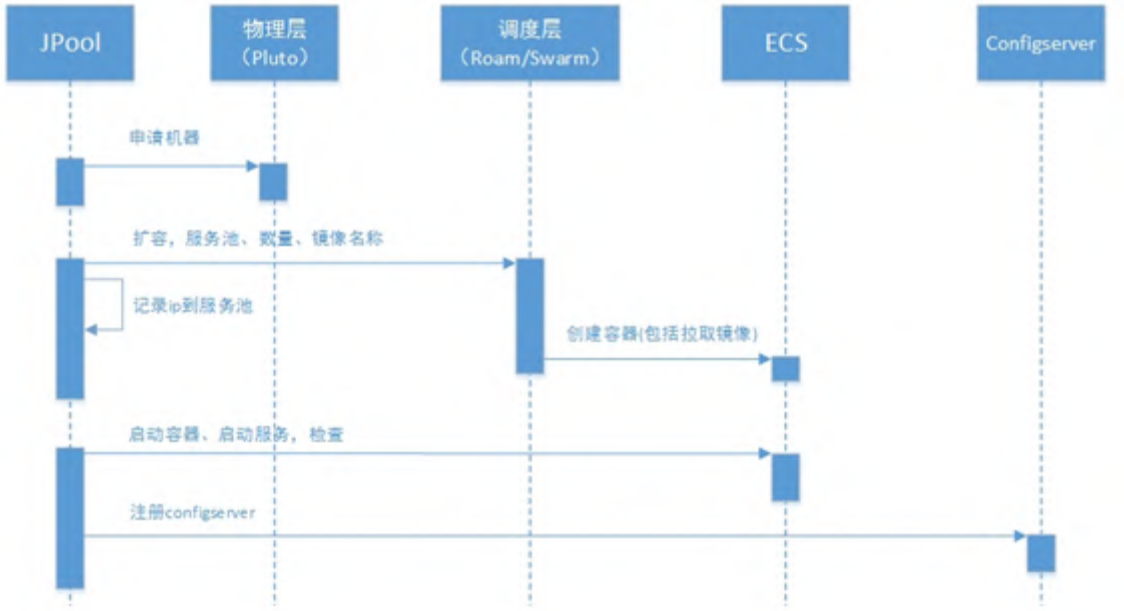


混合云DCP容器资源共享池



混合云DCP流程


· 内网申请
· 云端申请
主机申请
初始化
· Docker环境
· 系统环境
· 调度策略与算法
动态调度
服务发现
· Nginx
· Motan
· SLB
· 反初始化
· 归还Buffer池
· 结算中心
下线


JPool
物理层（Pluto）
调度层（Roam/Swarm）
ECS
Configserver
申请机器
扩容，服务池、数量、镜像名称
记录ip到服务池
创建容器(包括拉取镜像)
启动容器、启动服务，检查
注册configserver

混合云DCP - 主机申请

新浪微博 weibo.com

交易单 | 请输入交易单 | Go! | 资源申请 + | 阿里云ECS创建 +

#	交易单	发起方	接收方	原因	交易数量	机型	CPU	MEM	开始时间	结束时间	状态	#

发起申请

发起方	WeiboPatform_Platform	可用地域	北京地区
可用区域	北京可用区C	VPC	weibo_vpc
交换机	Weibo_Switch_C_1	安全组	beijing-a-default
机器规格	16Cores-16GB	CPU	16
内存	16	cmdb服务类型	JAVA应用
申请机器数量	100	镜像	weibo_hongbao_v5_4
结束时间	2016-01-21T01:29:00Z		

混合云DCP - 初始化

系统环境：dnsmasq、ntp、cron…

软件环境：Docker、Swarm、Consul
Mesos、SinaWatch…

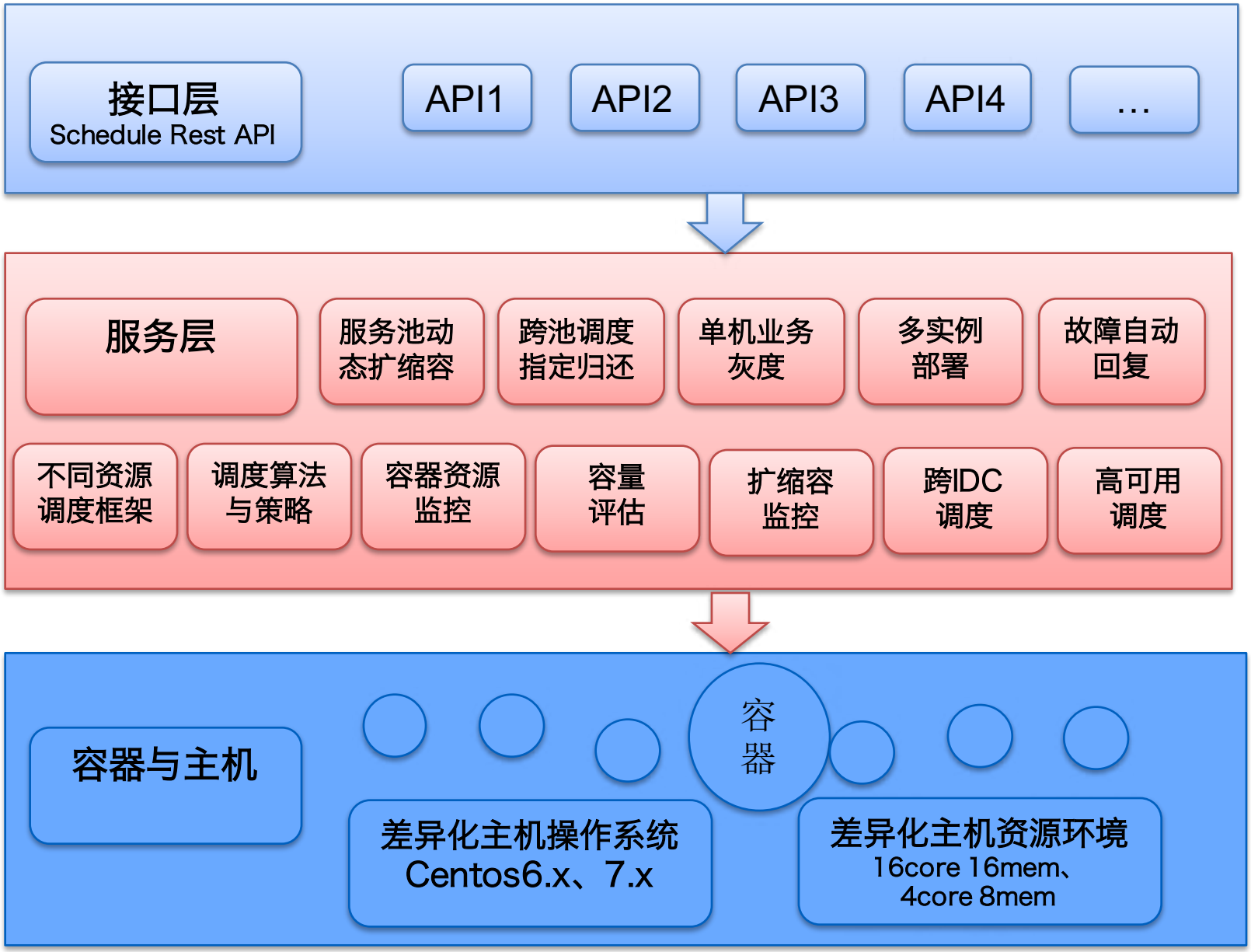
20	new : install docker	详细	false	Ok	详细	0	2016-01-21 02:24:01	2016-01-21 02:24:01
21	new : remove conflict package	详细	false	Ok	详细	1	2016-01-21 02:24:00	2016-01-21 02:24:01
22	new : yum state=present name=nc	详细	false	Ok	详细	1	2016-01-21 02:23:59	2016-01-21 02:24:00
23	new : command	详细	true	Ok	详细	1	2016-01-21	2016-01-21

弹性调度 - 选型

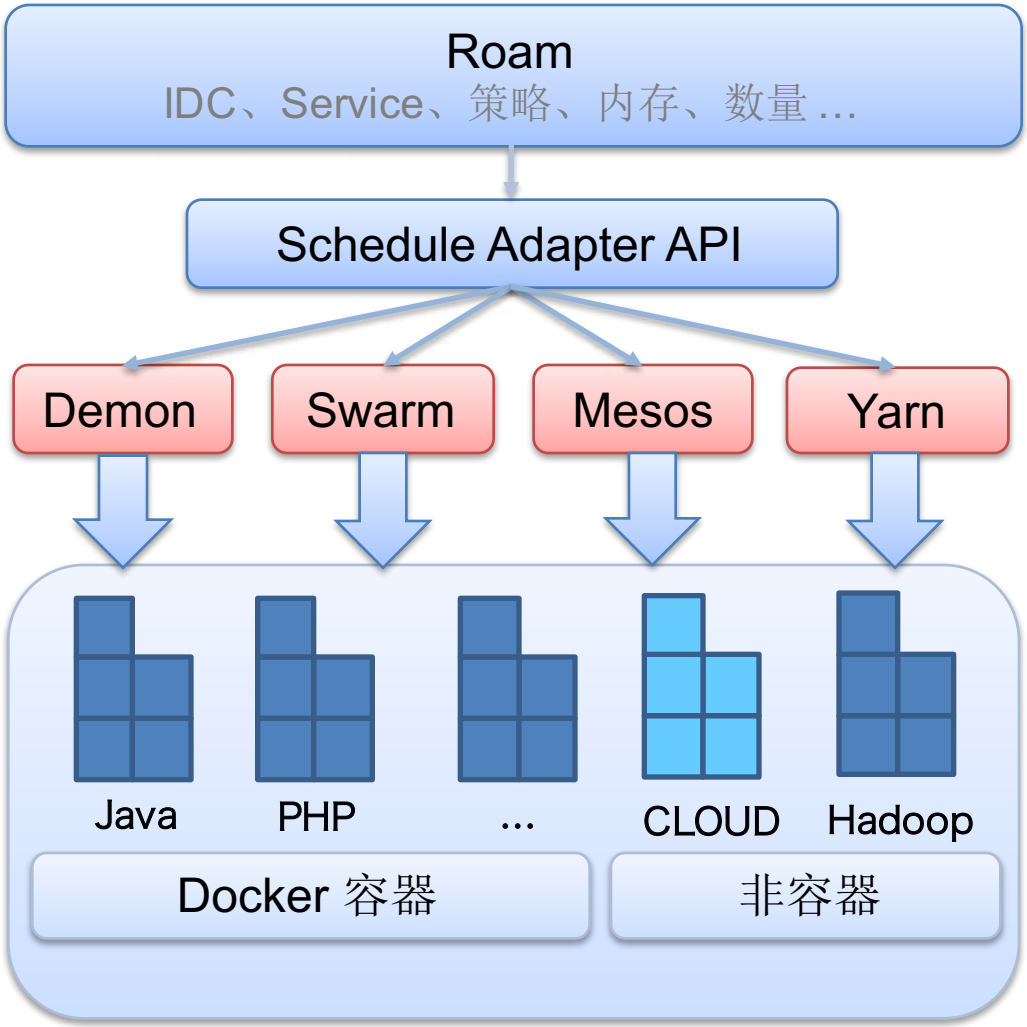
-	Swarm	Mesos	k8s
架构特点	针对Docker体系	借鉴Borg理念	源自Borg，原生Docker
与Mesos结合	社区完善中	成熟	社区积极推荐
隔离机制	Docker	Mesos/Docker/其它	Docker
资源类型	内存 CPU 端口	内存 CPU 端口 Ulimit	内存 CPU 端口 Ulimit
调度非Docker	No	Yes	No
主机分组	Docker进程标签	Slave配置	Slave配置
打分策略	资源使用情况	资源使用情况	资源使用情况+应用节点均衡
端口编排	No	Yes	Yes
网络模式	Docker原生	支持自建	支持自建
主高可用	双主切换	zk	etcd
扩缩容	Roam二次开发	API修改实例数	API修改实例数
应用节点健康检测	No	Yes	Yes
应用节点自动故障转移	No	Yes	Yes
服务发现	Consul	zk	etcd
负载均衡	No	Haproxy	Kube-proxy
DNS	No	No	skydns
使用复杂度	低	中	中
集群规模	小	中	中，在提升
生产使用	尚无大规模使用	业界大规模使用	业界大规模使用

弹性调度 - 选型

需求：
快速迭代
实现内网
计算资源
统一管理
调配，公
有云上获
得计算资
源，快速
自动化资
源调度与
应用部署

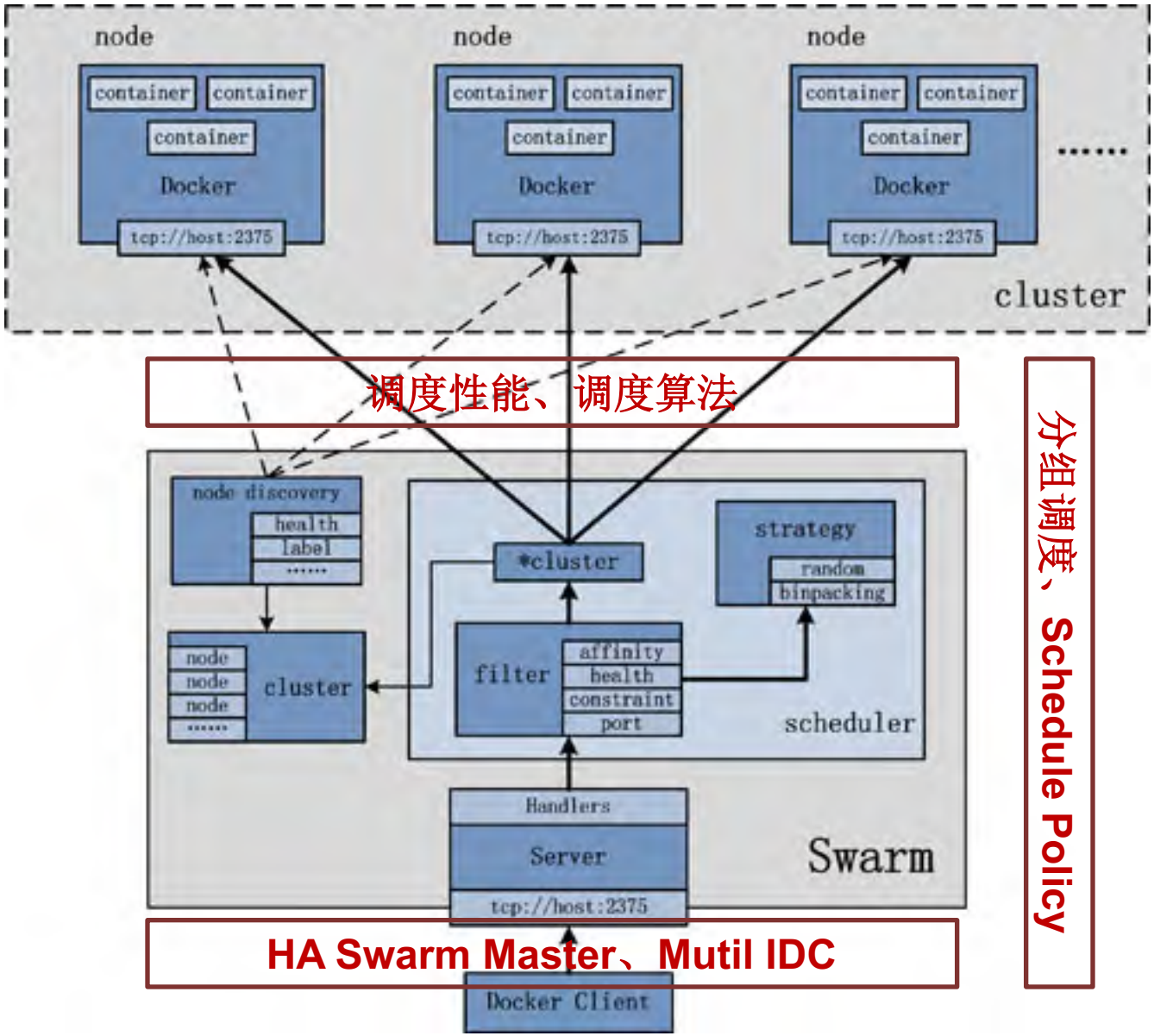


弹性调度系统

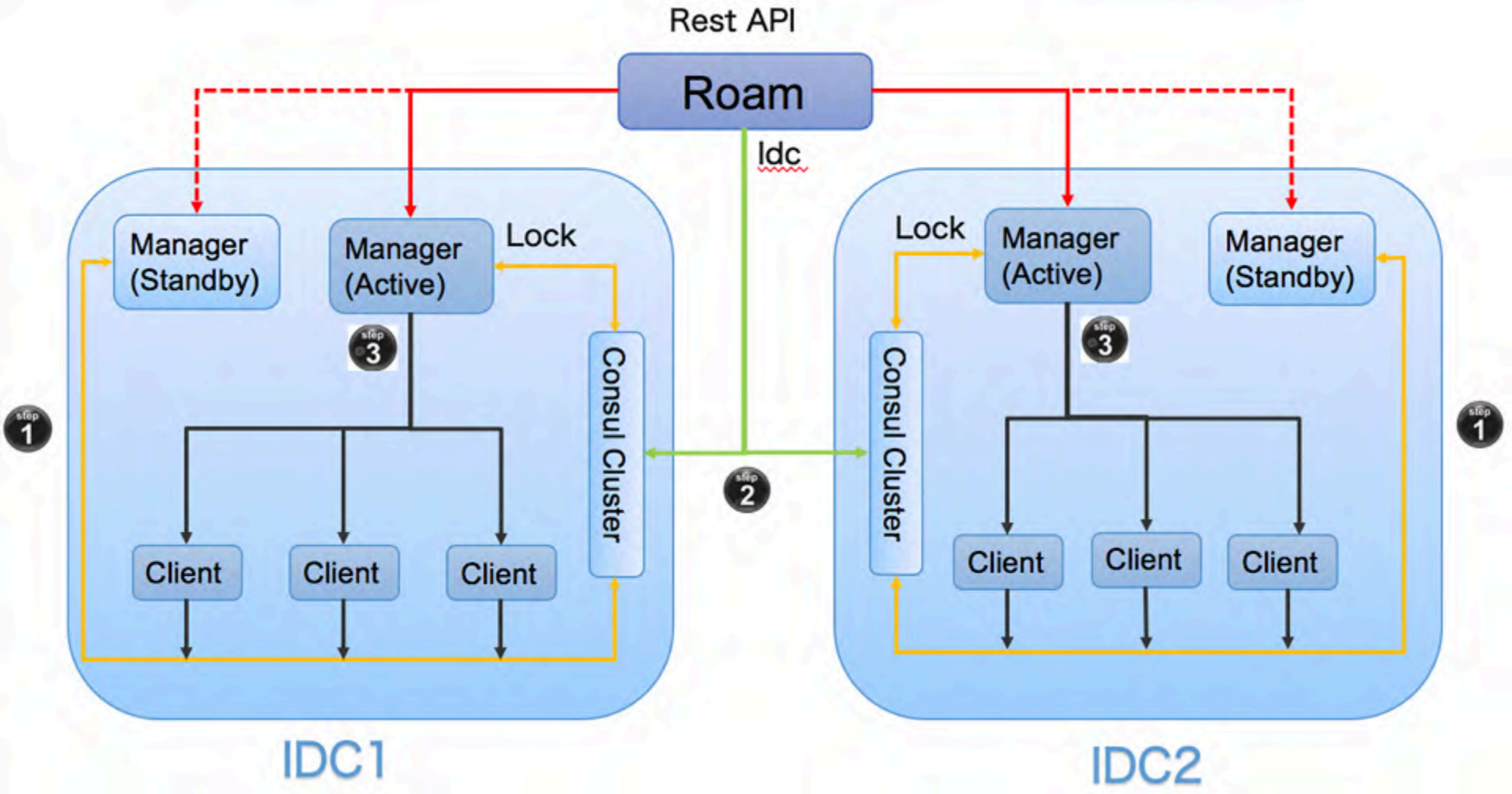


- 服务发现
- 调度框架
- 自定义调度
- 资源管理
- 容量评估
- 监控报警

动态调度 - Swarm架构



动态调度 - 多IDC、高可用、可扩展



Swarm – 调度策略

- 调度=主机 or 容器过滤 +策略选择
- 过滤器filter
  - Node Filters：health（会根据节点状态进行过滤，会去除故障节点）、constraint（约束过滤器、Label分组调度）
  - Container Configuration Filters：affinity（亲和性过滤器）、denpendency（依赖过滤器）、Port（会根据端口的使用情况过滤）
- 调度策略
  - 根据各个节点的可用的CPU, Mem及正在运行的容器的数量来计算应该运行容器的节点进行打分，剔除掉资源不足的主机，然后策略选择：spread、binpack、random
  - Binpack：在同等条件下，选择资源使用最多的节点
  - Spread：在同等条件下，选择资源使用最少的节点
  - Random：随机选择一个
- 调度颗粒度
  - Memory：docker run -m 1g …
  - CPU：docker run -c 1 …

Swarm - 调度算法

新浪微博 weibo.com

```
if config.CpuShares  > 0 {
        cpuScore  = (node.UsedCpus  + config.CpuShares)  * 100 / nodeCpus
        //cpuScore=（物理机已用CPU+本次需用CPU）*100/物理机CPU
}
if config.Memory  > 0 {
        memoryScore  = (node.UsedMemory  + config.Memory)  * 100 / nodeMemory
        //memScore=（物理机已用内存+本次需用内存）*100/物理机内存
}
```

```
if cpuScore <= 100 && memoryScore <= 100 {同时满足可用内存、CPU
   weightedNodes = append(weightedNodes, &weightedNode{Node: node, Weight: cpuScore + memoryScore})
}
```

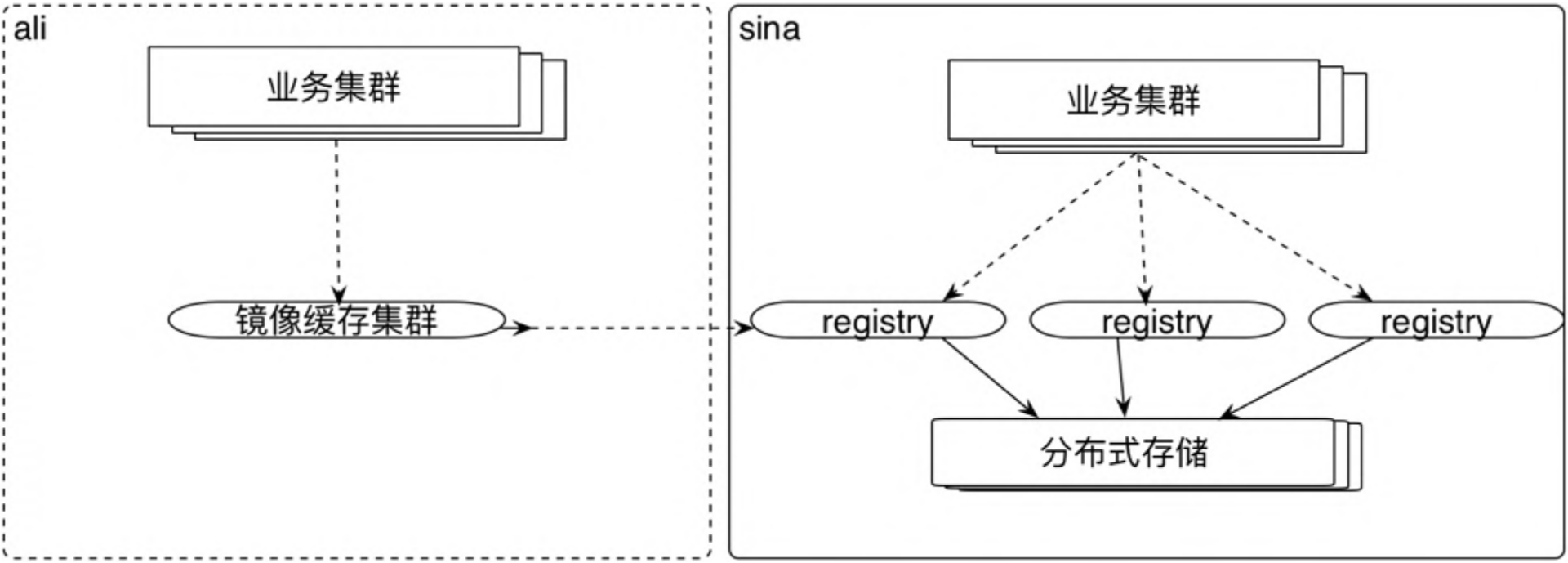
资源只与容器Create时配置有关，与运行时实际使用资源情况无关。无论容器是否由Swarm创建，无论容器处在何种状态，只要配置了资源限额，调度时均会计算在内！

动态调度 - Roam

- 提供通用HTTP API，适配不同分布式资源调度框架
- 不同调度策略与算法
- HA Swarm Master高可用
- 多机房自动适配
- 单IP Docker Deamon下发执行机制
- 容器资源监控
- 扩缩容监控
- 容器资源评估
- …

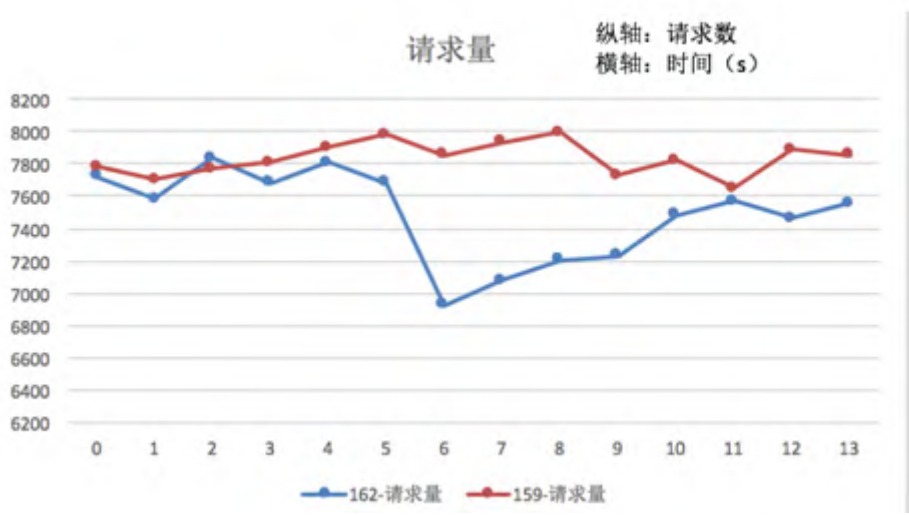
跨云动态调度挑战 - Docker Registry

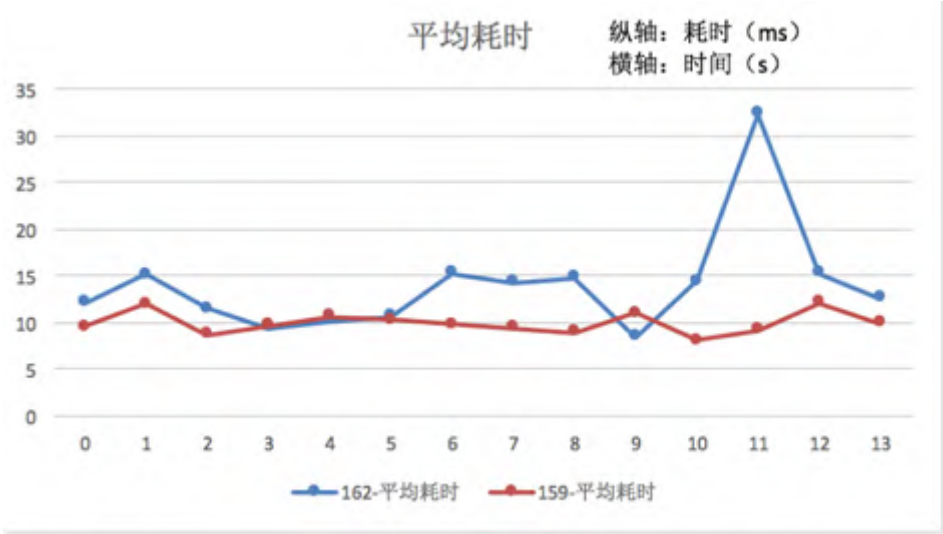
- Docker Image
  - Docker Image分层设计
  - 自动打包机制
- 构建私有Registry Hub
  - Docker Registry：V1 → V2
  - Storage Driver：Ceph



问题：Reload损耗

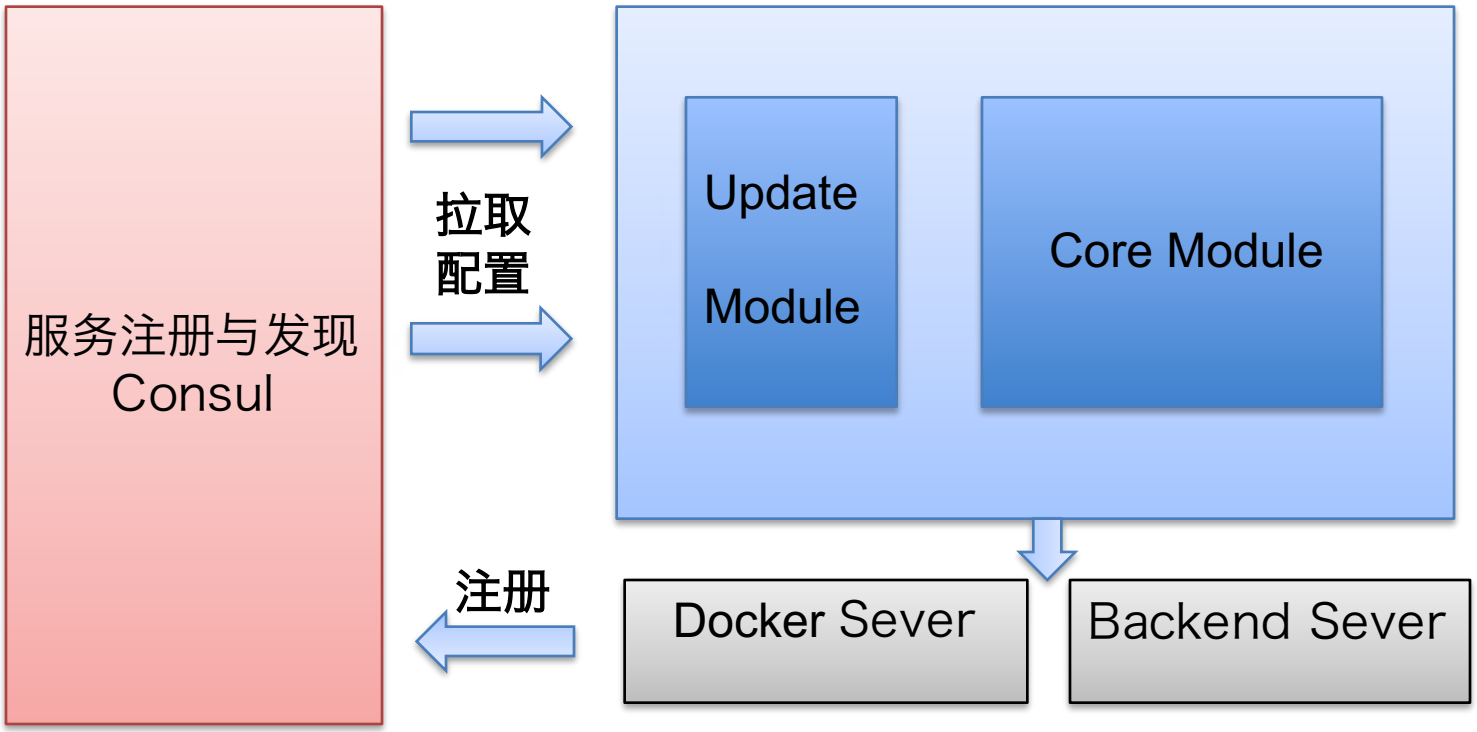
- 开源解决方案大多利用Nginx的Reload机制
- 性能损耗情况:
- 请求量：普通reload会导致吞吐量下降10%
- 平均耗时： 差异不大，新模块稍有优势





微博方案 - nginx-upsync-module

- Nginx Plus的开源版
- 支持基于Consul自动服务发现
- 开源：https://github.com/weibocom/nginx-upsync-module



服务发现

Nginx Upsync - 自动适配后端处理能力

- 弹性节点的处理能力不对等
  - server 10.xx.xx.xx:xxxx max_fails=0 fail_timeout=30s weight=20; #同样的权重导致单点性能恶化
- 节点注册计算能力
  - 所有节点默认权重是20;
  - 公有云有20%性能损耗,权重=16;

Motan RPC服务发现

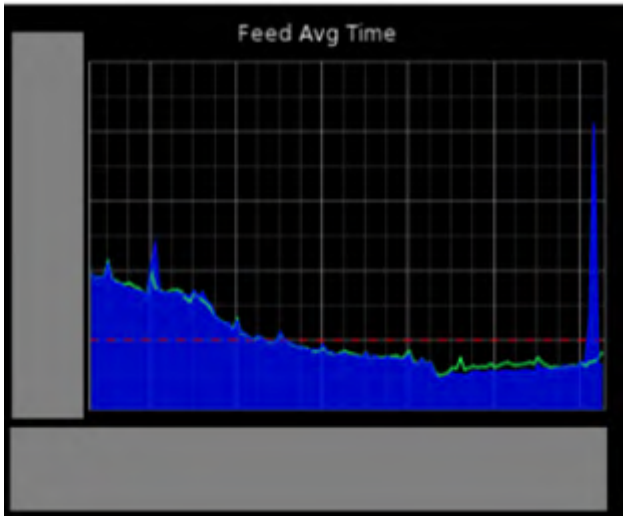
- 支持跨IDC流量切换
- 支持按流量权重配置定向路由

DNS 服务发现

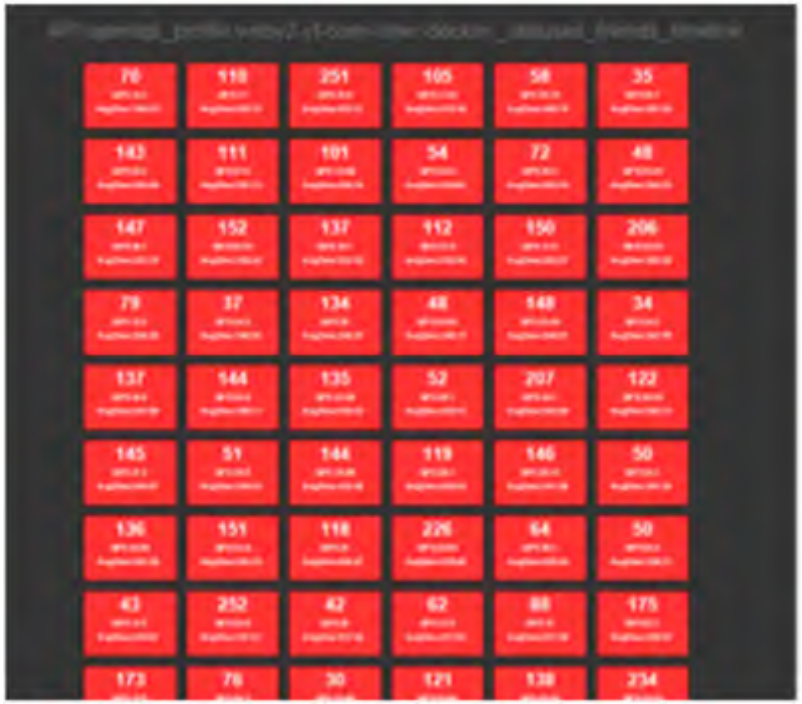
监控体系

针对于混合云体系，我们提供了一套完整的监控告警解决方案，实现对于云上IT基础架构的整体监控与预警机制，最大程度地保证了微博体系能够稳定地运行在混合云体系上，不间断地为用户提供优质的服务。监控告警解决方案实现了四个级别上的监控与预警：

- 系统级监控
- 业务级监控（全网监控到单机容器监控）
- 资源级监控
- 专线网络监控



单机性能恶化

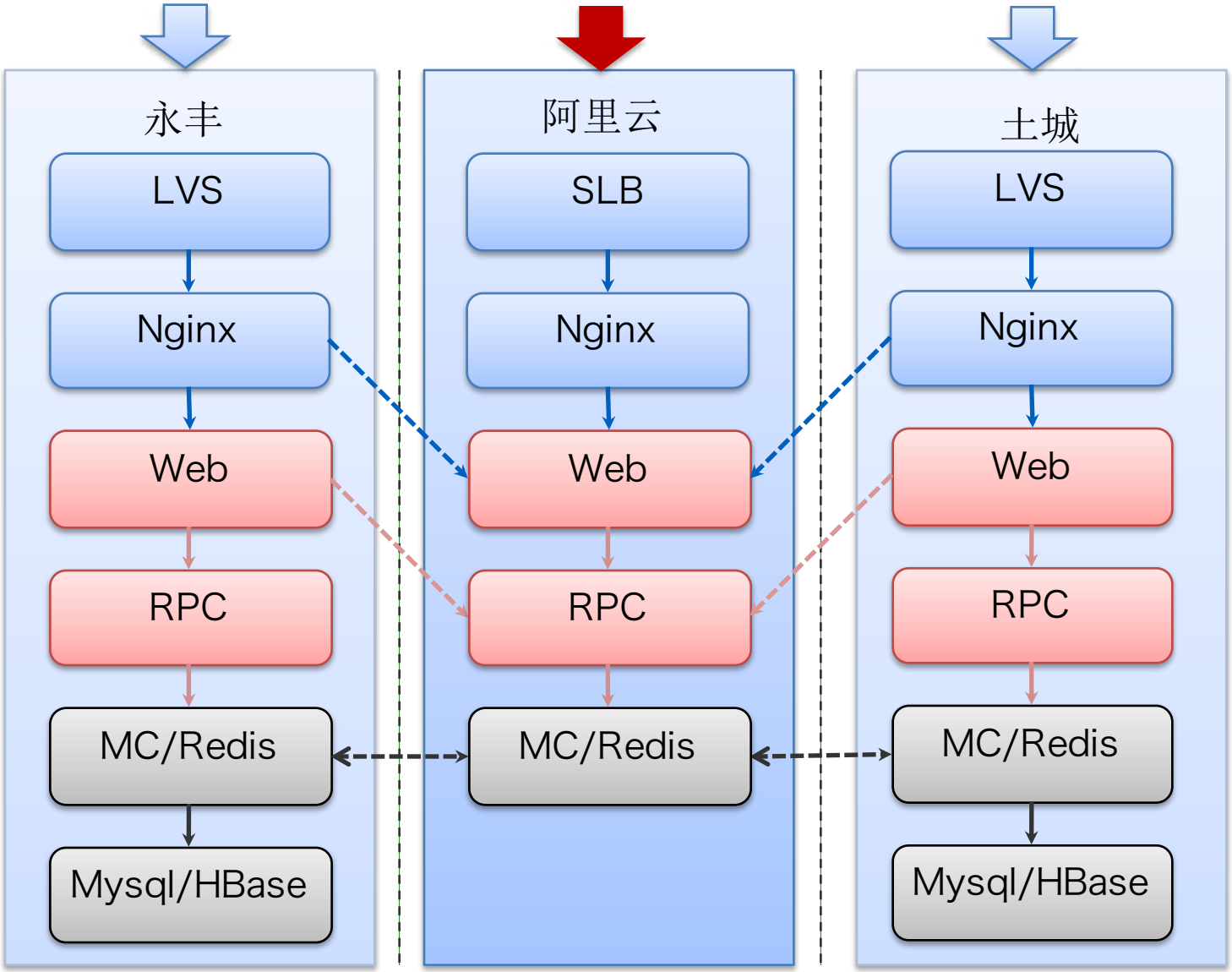


容量评估

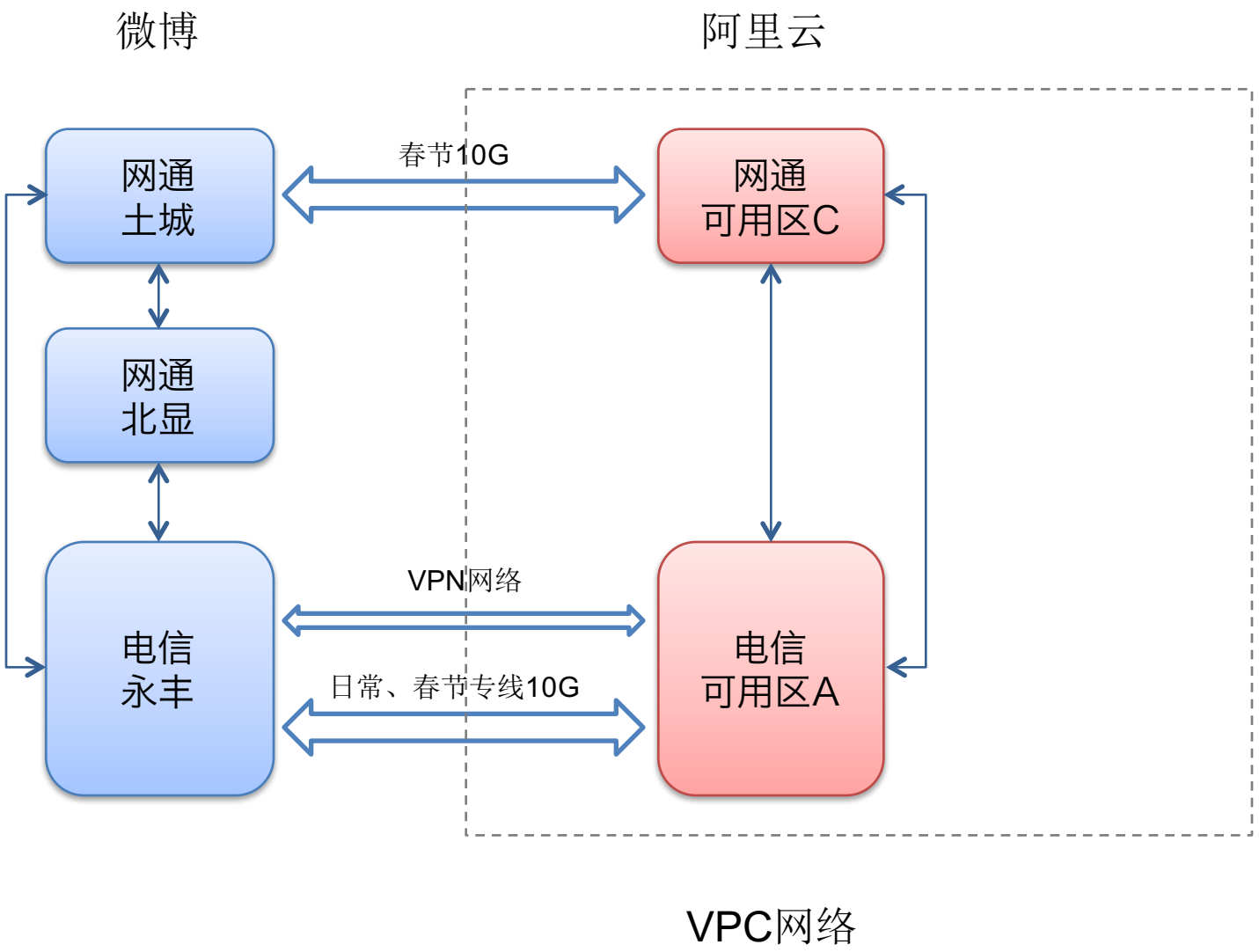
- 根据服务池单机平均系统指标（CPU idle、mem load）、QPS、带宽、业务SLA综合指标容量评估

出发

三节保障与阿里云部署


永丰
LVS
Nginx
Web
RPC
MC/Redis
Mysql/HBase
阿里云
SLB
Nginx
Web
RPC
MC/Redis
土城
LVS
Nginx
Web
RPC
MC/Redis
Mysql/HBase

微博混合云专线网络架构



总结 - 微博混合云DCP进展

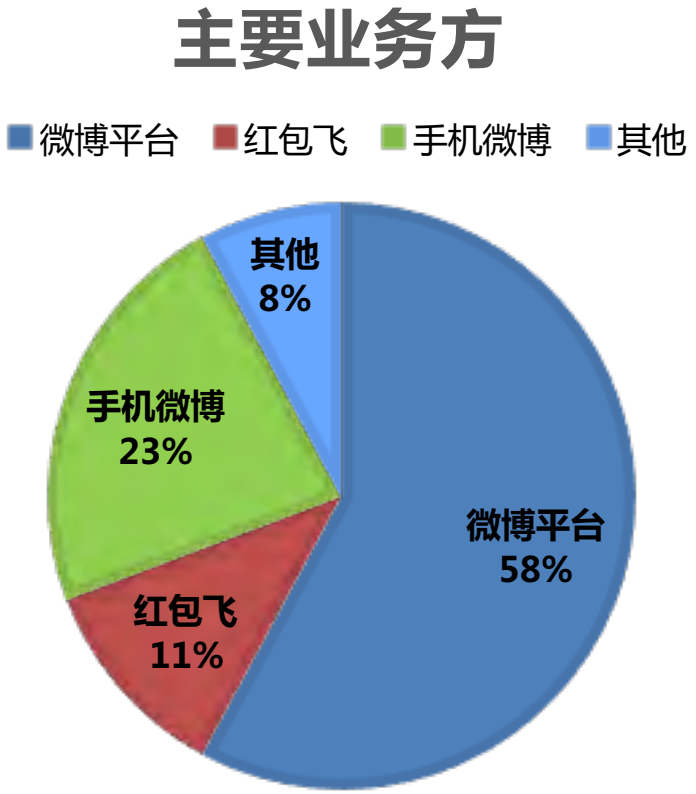
新浪微博 weibo.com

- 混合云进展：
  - 上线：2015.10
  - 容器数：3000+
  - Swarm集群：三IDC 2500+ Containers
  - Mesos集群：100+

- 双十一考验：
  - 单日十次扩缩容
  - 单次扩缩容时间：< 5分钟

- 元旦考验：
  - 当天数十次扩缩容
  - 跨云端动态扩缩容

- 春晚备战：
  - 10分钟混合云扩容1000节点技术能力
  - 春晚提供3000节点扩容能力，Feed、红包飞、手机微博均可支持



总结 - 遇到的坑

问题	解决
Ulimit：Docker容器启动ulimit1024未获取到主机设置的ulimit值导致业务容器启动不成功	Docker Demon启动时设置ulimit或重启操作系统预先加载Ulimit
Swarm全局锁：Swarm全局锁，并发变为串行	Roam二次开发预先拉取镜像，提高并发度。Swarm1.0.0改为分布锁已修复
僵尸容器：Docker Create设置-m，导致资源无法使用并调度	Roam探测处理僵尸容器
安全性问题：Docker Demon开启2376端口引起容器安全问题	开启iptables，同时关掉nf_conntrack连接跟踪，添加ip和端口过滤规则
Consul网络波动：Consul通过UDP协议跨数据中心广播，设置-advertise IP不在可达网段，导致整个Consul集群Down掉	设置Consul -advertiseIP为实际通信网卡IP

总结

- 从业务需求出发设计系统架构
- 整体架构通用性设计
- 实践去做，技术架构迭代升级
- 周边技术体系建设重要性

Thanks

以微博之力 让世界更美！

weibo.com