2015 Container
技术峰会

容器技术在SNG中间层的应用

funnychen（陈芳录）

2015@OpenCloud

关于我

- funnychen（陈芳录）
- @腾讯社交网络运营部平台技术运营中心
- 擅长大规模组件运维、运维自动化建设

提纲

- 中间层介绍及其运维体系
- LXC虚拟化实践
- 蜂巢集群和Docker尝试

中间层在社交架构中的位置

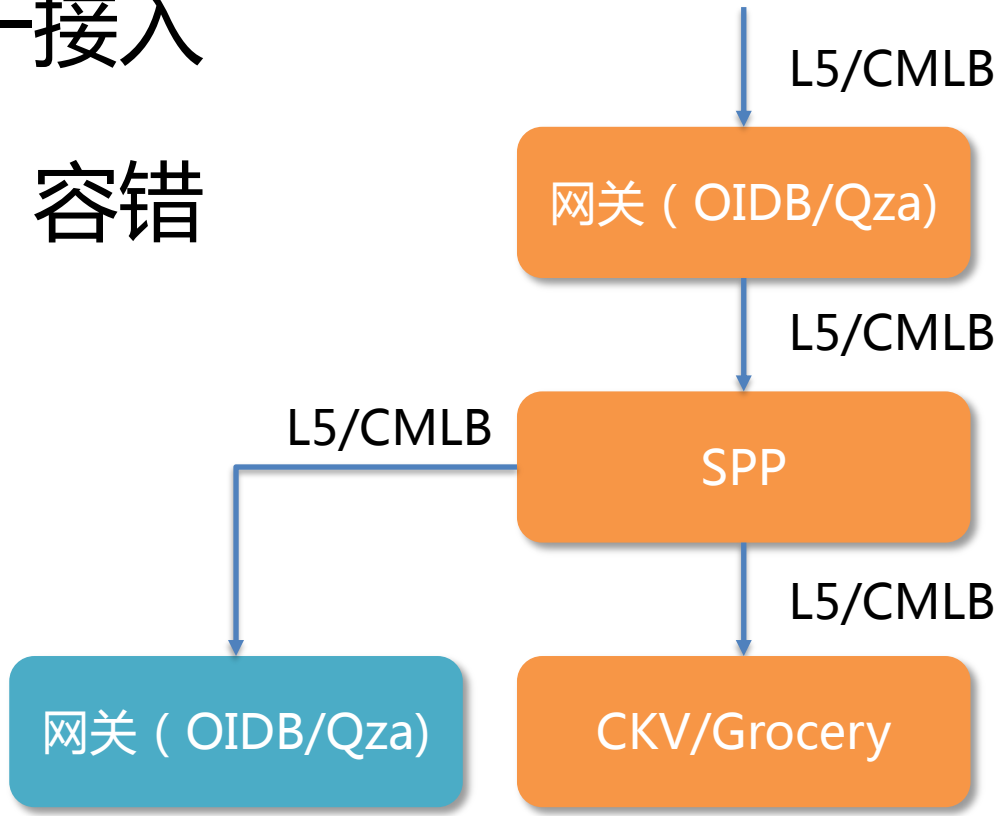
Qzhttp Nginx

SPP

CKV Grocery CDB

典型业务架构

- 网关：鉴权、统一接入
- 路由组件：寻址、容错



现状

20000+服务器

30000+服务实例

1000+业务

三地部署

QQ群 QUN.QQ.COM

QQ空间

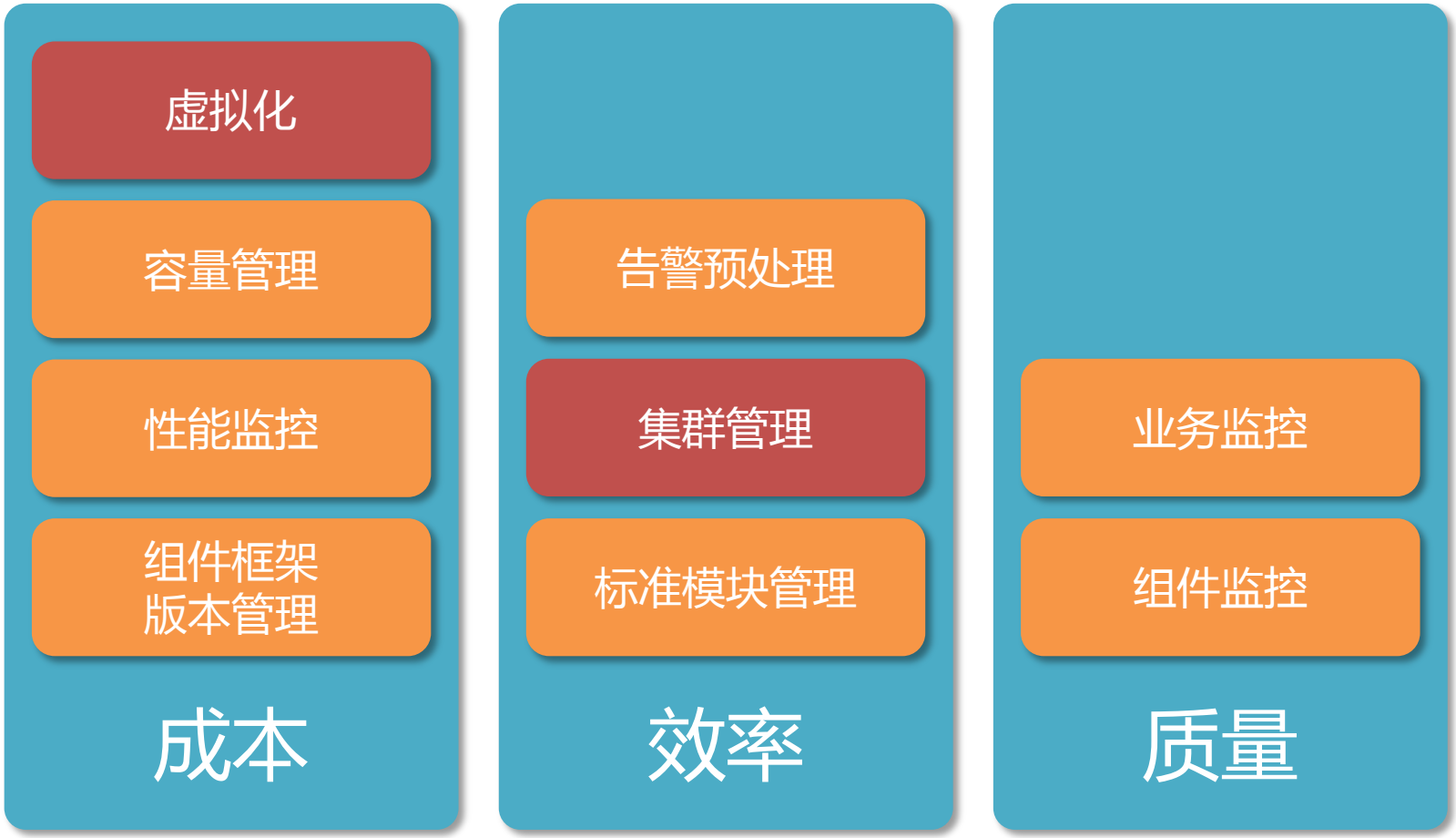
广点通 社交效果营销平台

QQ音乐

QQ Show SHOW.QQ.COM

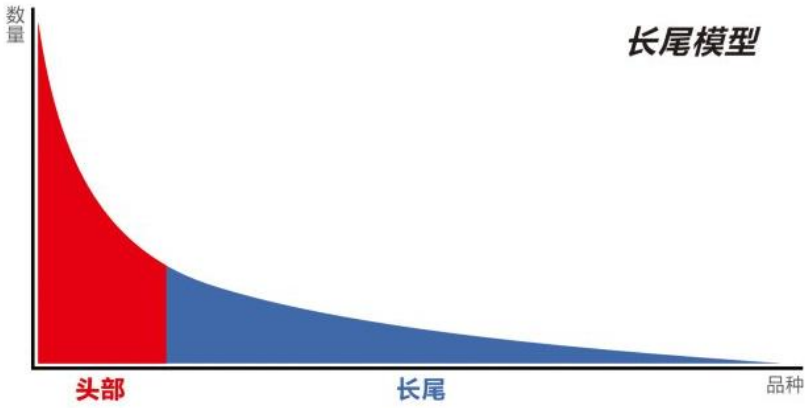
QQ相册 PHOTO.QQ.COM

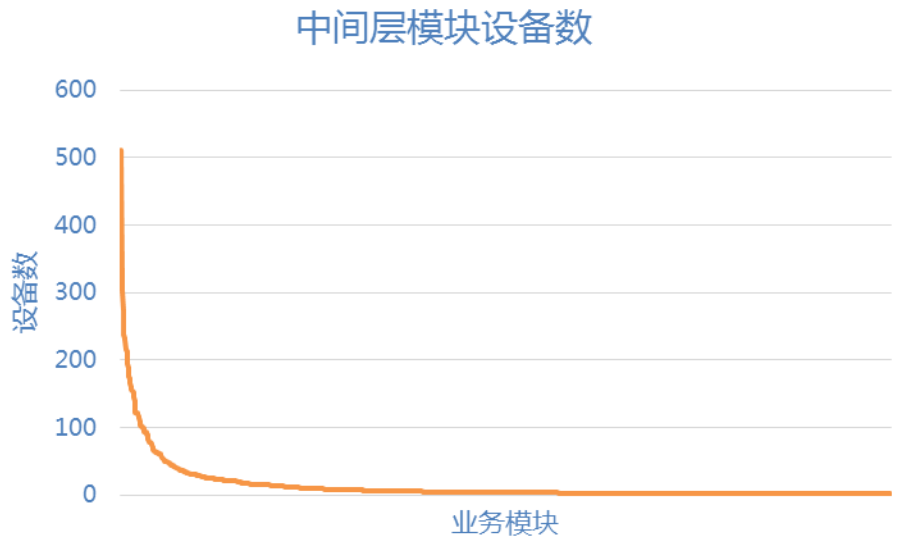
中间层运维体系



引入虚拟化技术

- 提供细粒度资源单位，解决长尾问题
- 统一机型，提升资源流转效率





虚拟化选型

- 成本最低：海量服务对性能的要求
- 尽量不增加监控、运维成本
- 业务运行环境可控：自研业务

	实体机	XEN	LXC
网络	-	veth + bridge	veth + bridge
资源隔离	完全隔离	强	弱
性能损耗	无	有	极小
镜像与迁移	无	有	无

隔离方案

- 资源隔离
  - namespace（隔离uts,ipc,mnt,pid,user,net）
  - cgroup（限制进程组资源）
    - 子系统：

      cpuset,cpu,cpuacct,memory,devices,freezer,net_cls,blkio,tstat

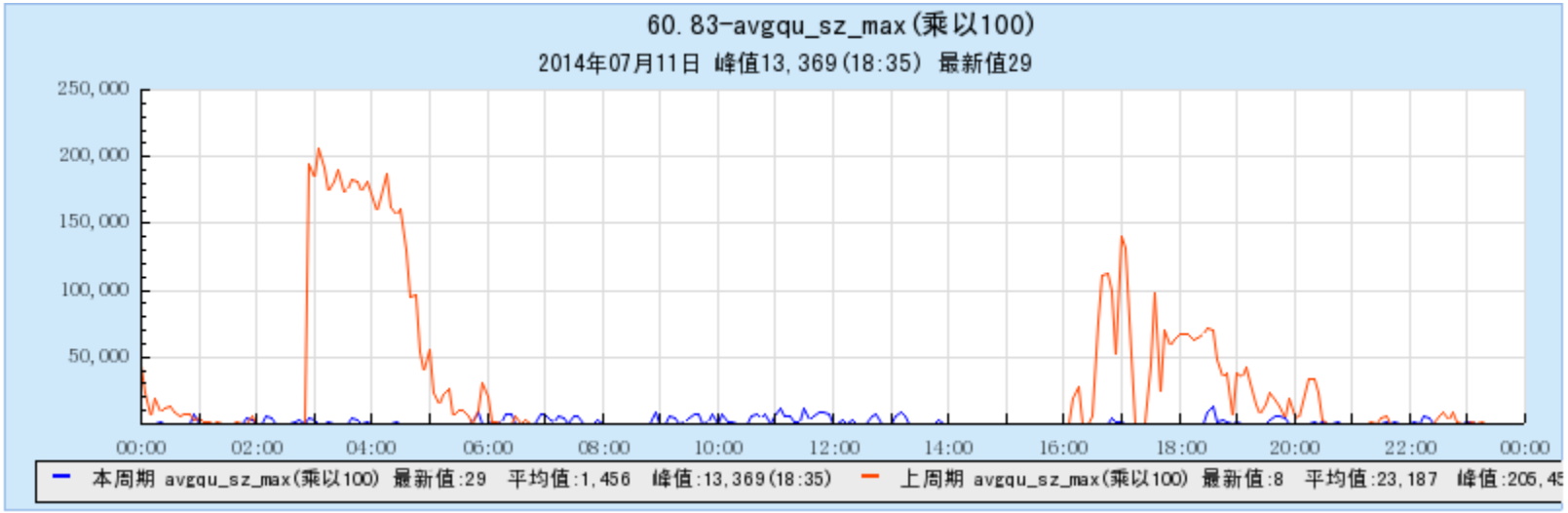
CPU和内存的隔离细节

- CPU：通过cpuset给子机指定CPU
- 内存：
  - 根据子机CPU核心数比例等额分配
  - 禁用swap：memory.limit_in_bytes = memory.memsw.limit_in_bytes

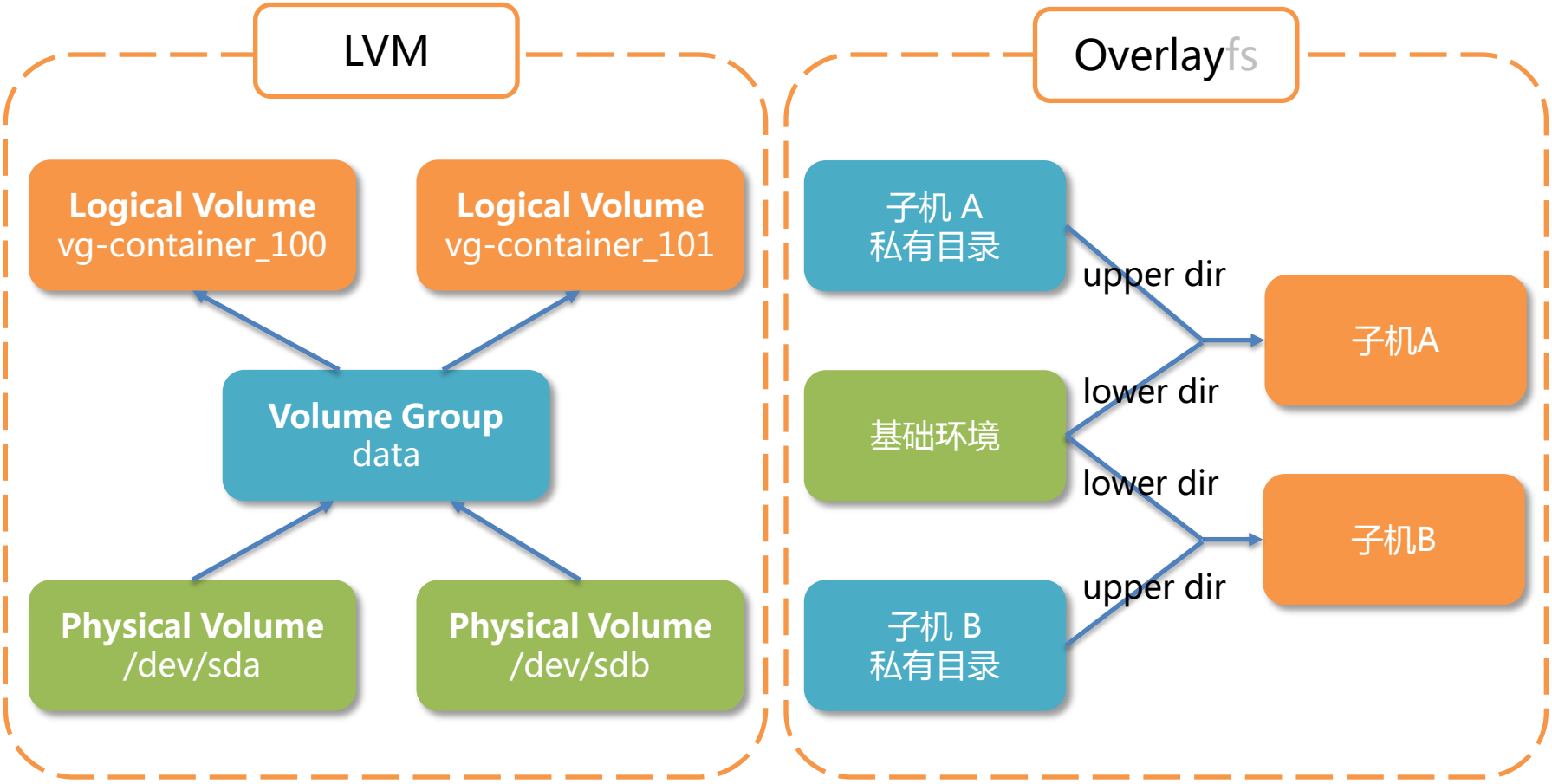
IO限速

检测avgqu-sz → 检测子机 iops、r/w bps → 动态调整 blkio.throttle.*

检测avgqu-sz → 连续N次空闲 → 恢复

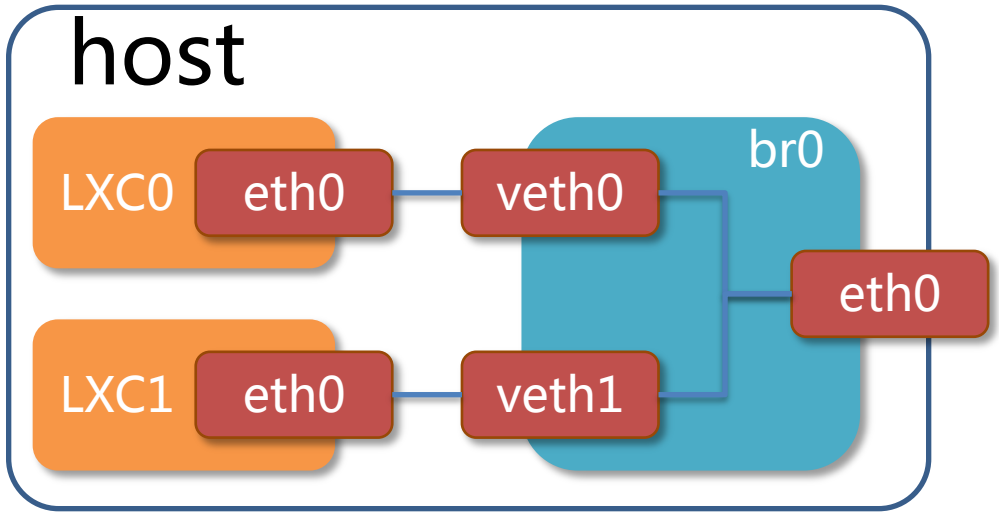


磁盘方案


LVM
Logical Volume
vg-container_100
Logical Volume
vg-container_101
Volume Group
data
Physical Volume
/dev/sda
Physical Volume
/dev/sdb
Overlayfs
子机 A
私有目录
基础环境
子机 B
私有目录
upper dir
lower dir
lower dir
upper dir
子机A
子机B

网络方案

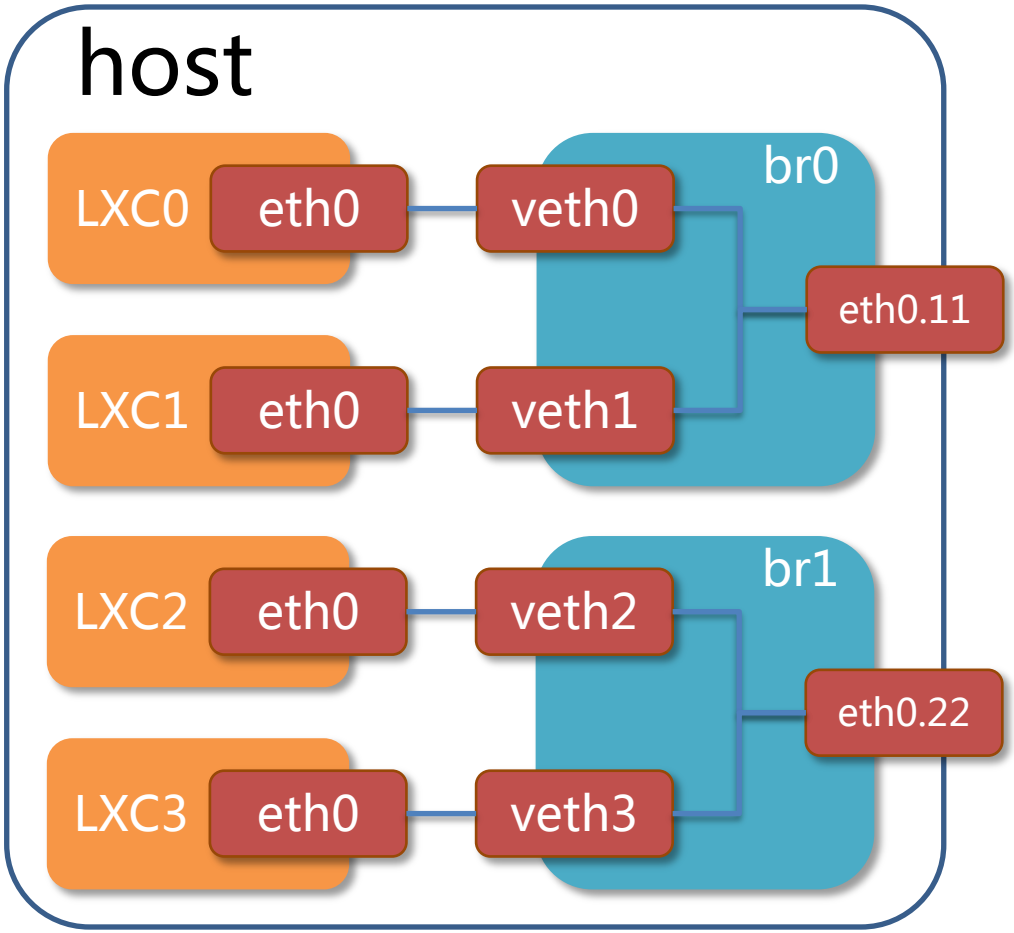
- veth+bridge
  - 可直接ssh登录
  - 对业务程序透明
  - 兼容网管监控



- net_cls + tc（htb）
  - 保证子机最小带宽

网络方案升级

- VLAN太小！



内核增强

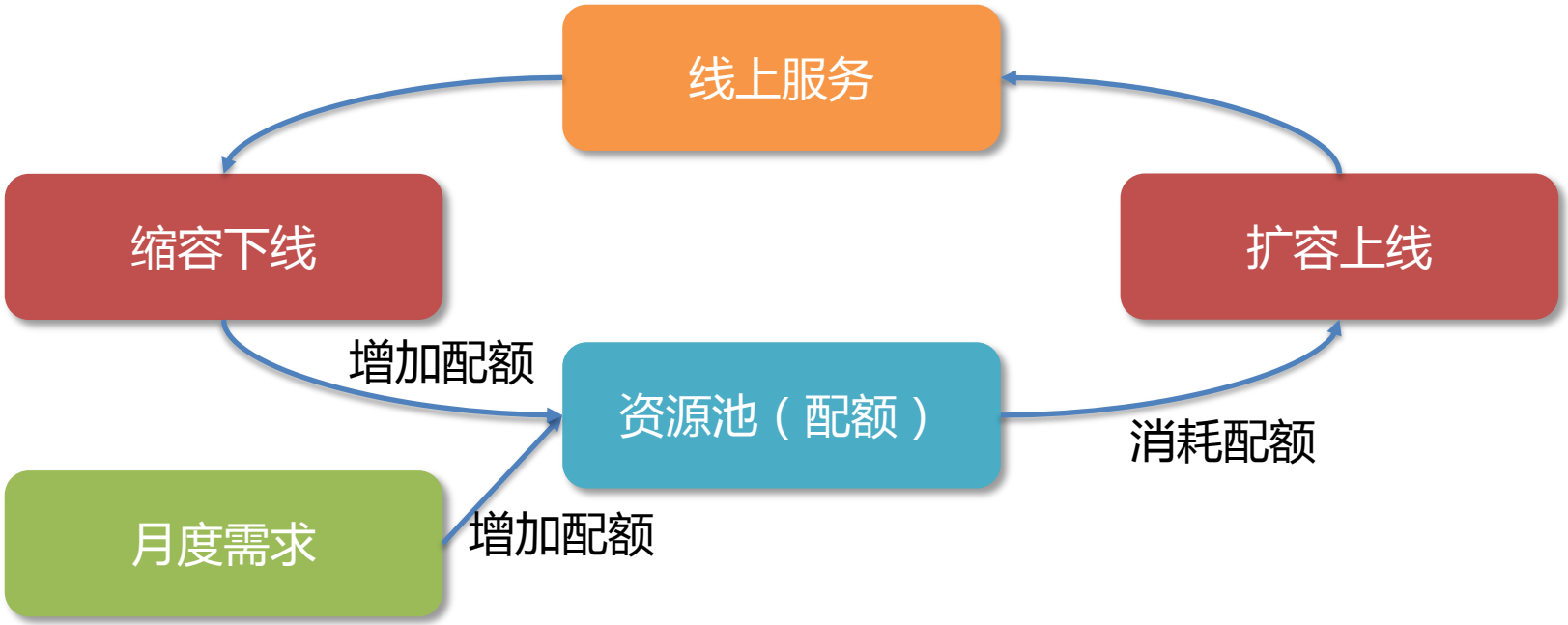
- /proc 虚拟化
  - 使子机只看到自己的CPU和内存情况，兼容网管监控
- 移植Task counter
  - 限制子机进程数量

生产管理

- 固定机型：2核/4核/8核
- 独占CPU：子机间无CPU竞争
- 打散：同个业务的子机分散到不同母机
- CPU动态调整：母机内CPU使用率均衡
- veth / VG 命名规范化，易于辨认缩绑定子机

配额制度

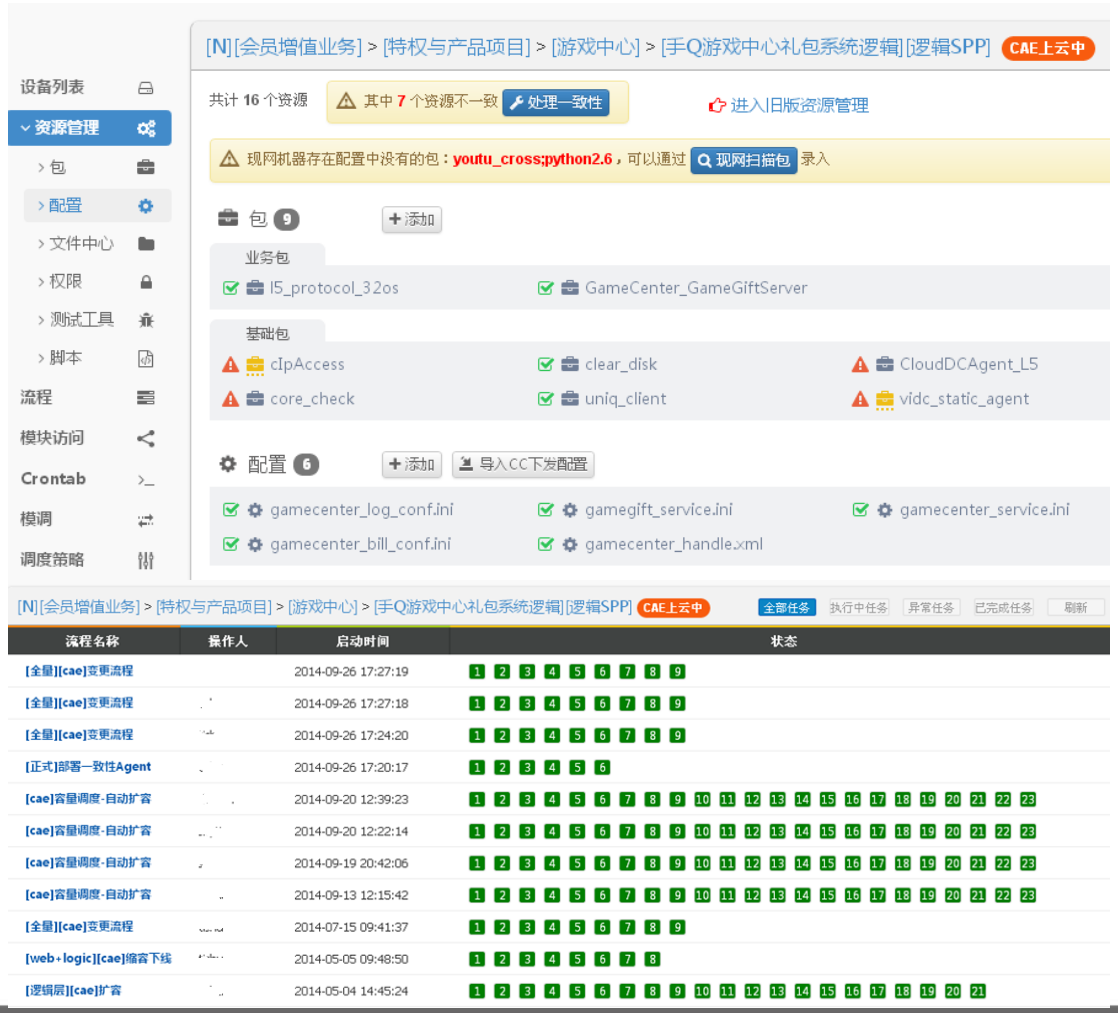
- 配额关联业务低负载，推动持续成本优化

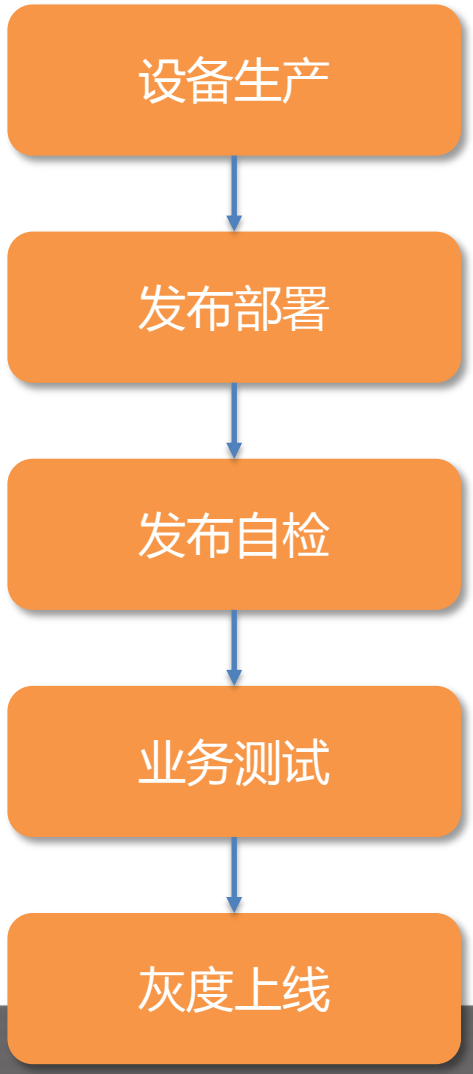


织云

- 内部云管理平台
- 分享：
  - 《梁定安：解密腾讯SNG云运维平台“织云”》
  - 《腾讯SNG织云自动化运维体系》

织云虚拟机自动扩容流程

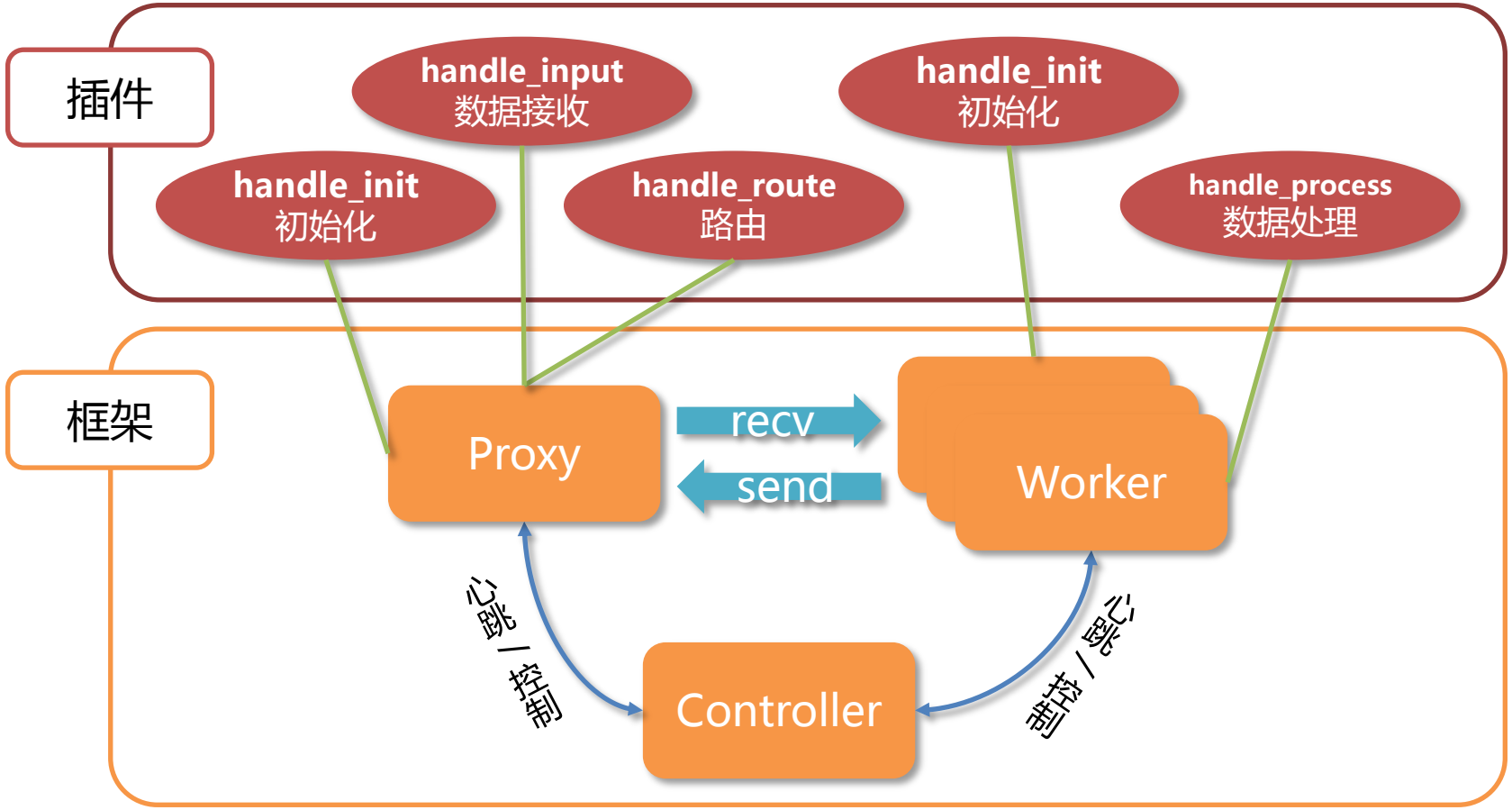




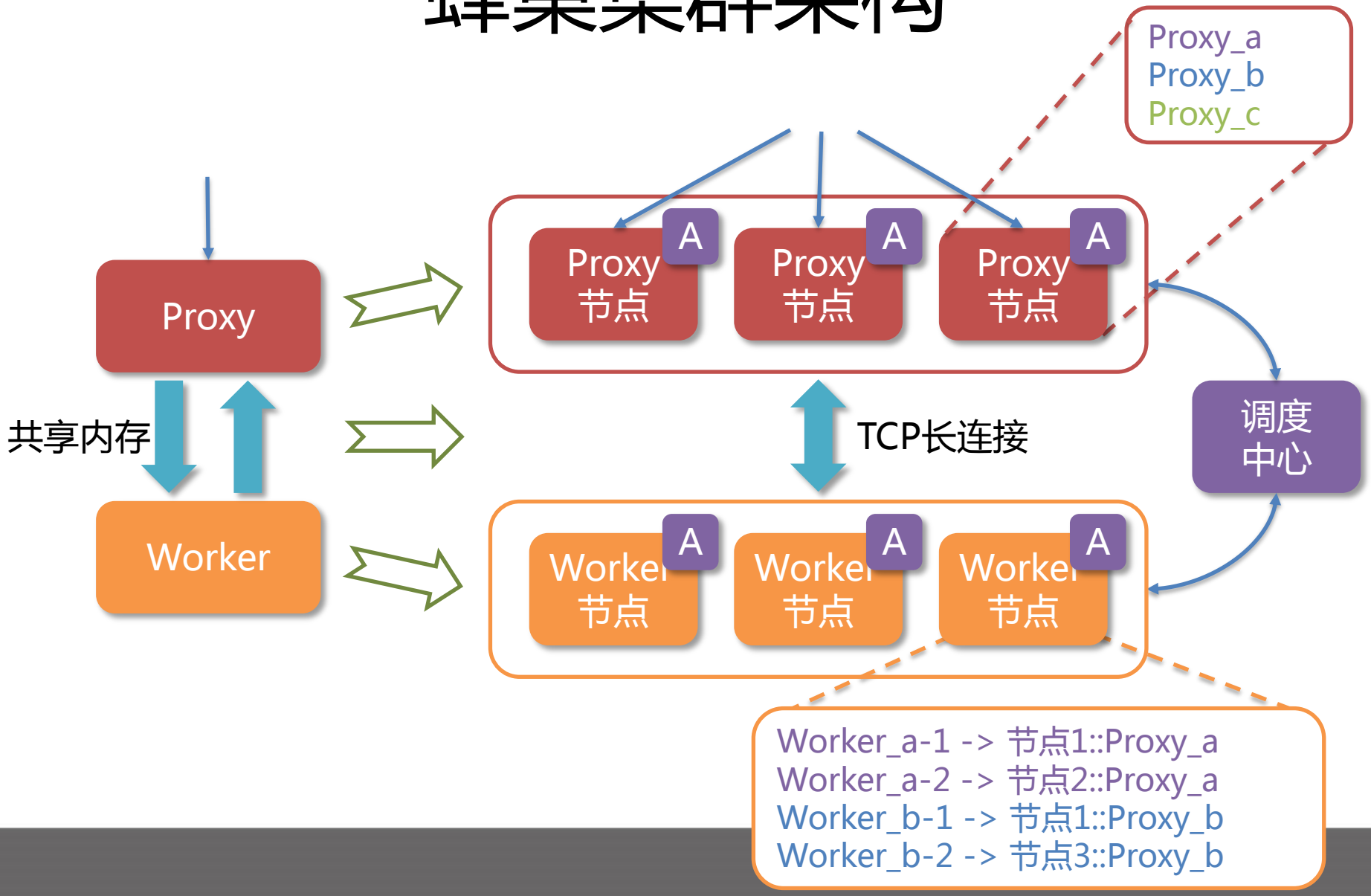
效率的痛

- 人均近万台设备、数百个模块
- 高强度扩容：春节、元宵、清明、五一……

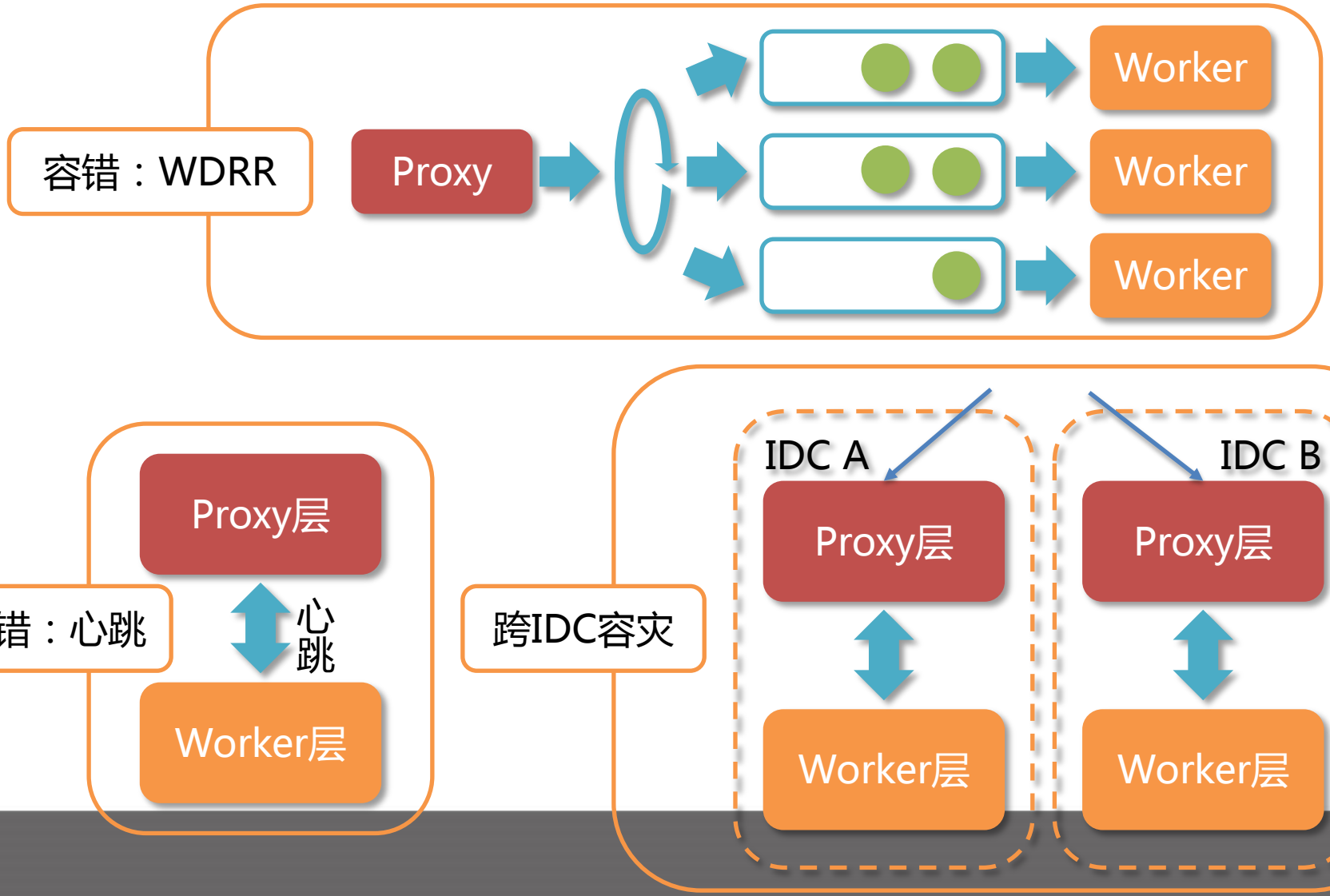
SPP组件架构


插件
handle_input
数据接收
handle_init
初始化
handle_init
初始化
handle_route
路由
handle_process
数据处理
框架
Proxy
recv
send
Worker
心跳 / 控制
Controller
心跳 / 控制

蜂巢集群架构

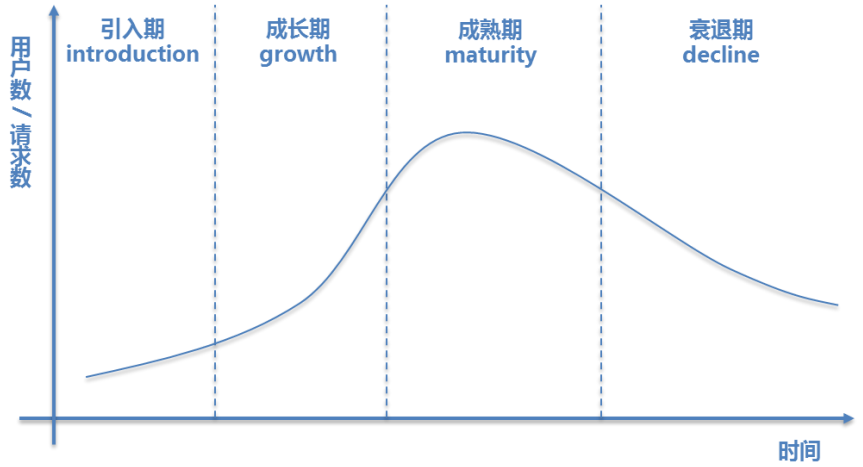

Proxy_a
Proxy_b
Proxy_c
Proxy
共享内存
Worker
Proxy 节点
A
Proxy 节点
A
Proxy 节点
A
TCP长连接
调度中心
Worker 节点
A
Worker 节点
A
Worker 节点
A
Worker_a-1 -> 节点1::Proxy_a
Worker_a-2 -> 节点2::Proxy_a
Worker_b-1 -> 节点1::Proxy_b
Worker_b-2 -> 节点3::Proxy_b

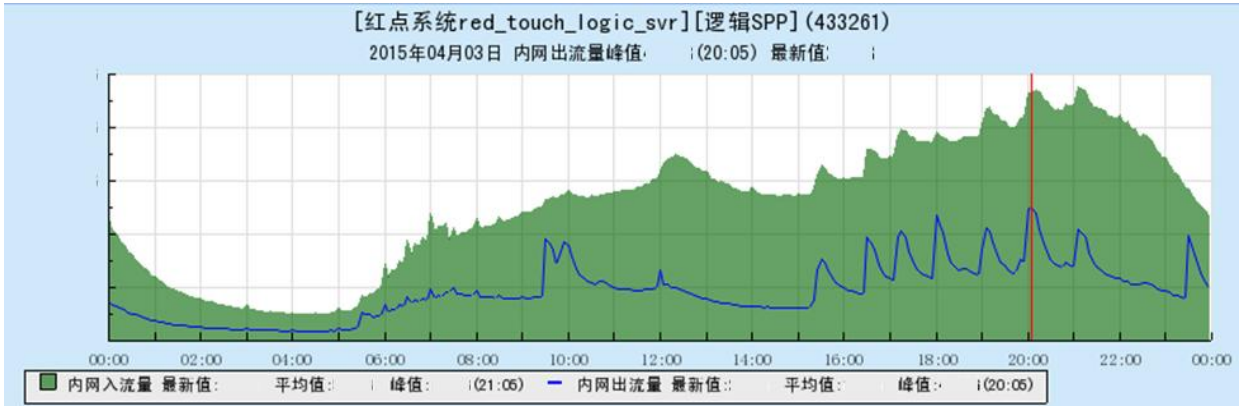
蜂巢高可用


容错：WDRR
Proxy
Worker
Worker
Worker
容错：心跳
Proxy层
心跳
Worker层
跨IDC容灾
IDC A
Proxy层
Worker层
IDC B
Proxy层
Worker层

典型场景

- 衰退期业务
- 业务活动

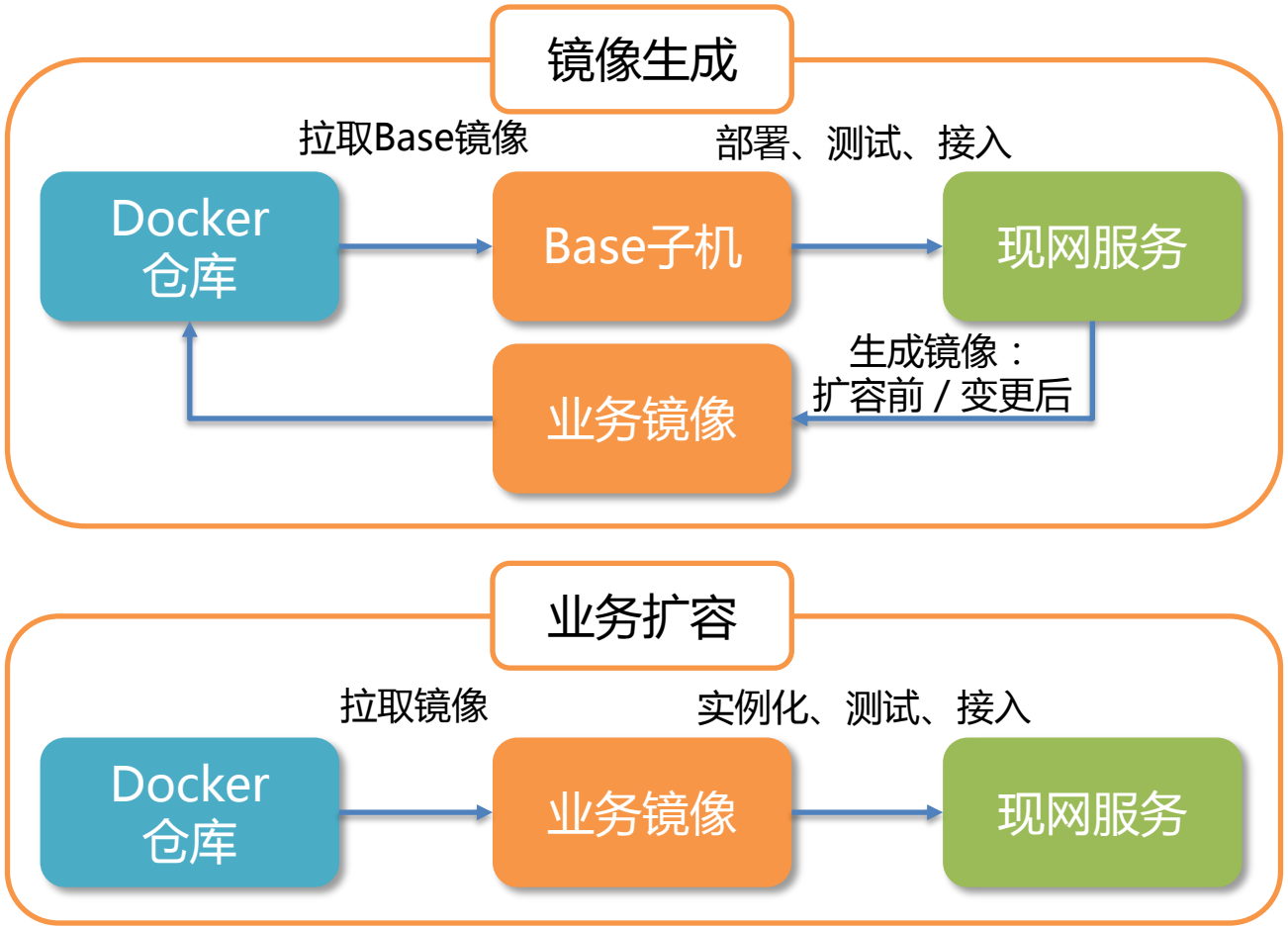




业务迁移绊脚石

- 包、配置、文件，如何关联管理？

引入Docker


镜像生成
拉取Base镜像
部署、测试、接入
Docker 仓库
Base子机
现网服务
生成镜像：
扩容前 / 变更后
业务镜像
业务扩容
拉取镜像
实例化、测试、接入
Docker 仓库
业务镜像
现网服务

未来挑战：中间层遇见Docker

- Carrier：Docker平台
- 全路径Docker化
  - 开发 -> 测试 -> 预发布 -> 发布
- 蜂巢调度Docker实例
- 蜂巢Docker定制化
  - 不隔离pid、mnt、uts、user

回顾

- 中间层介绍及其运维体系
  - 万级设备和服务、千级业务
  - 成本、效率、质量
- LXC虚拟化实践
  - 隔离方案、磁盘方案、网络方案、内核增强、生产管理
- 蜂巢集群和Docker尝试
  - 集群架构、服务打包、Docker应用方案

OpenCloud | 2015 Container 技术峰会

THANKS

加入我们：

做一流的互联网运维团队，

为用户创造优质的服务体验。

funnychen@tencent.com