2014
年度

PPTV系统平台建设实践

演讲人：陈文春

2014-11-10

www.pptv.com

2014
年度

自我介绍—陈文春

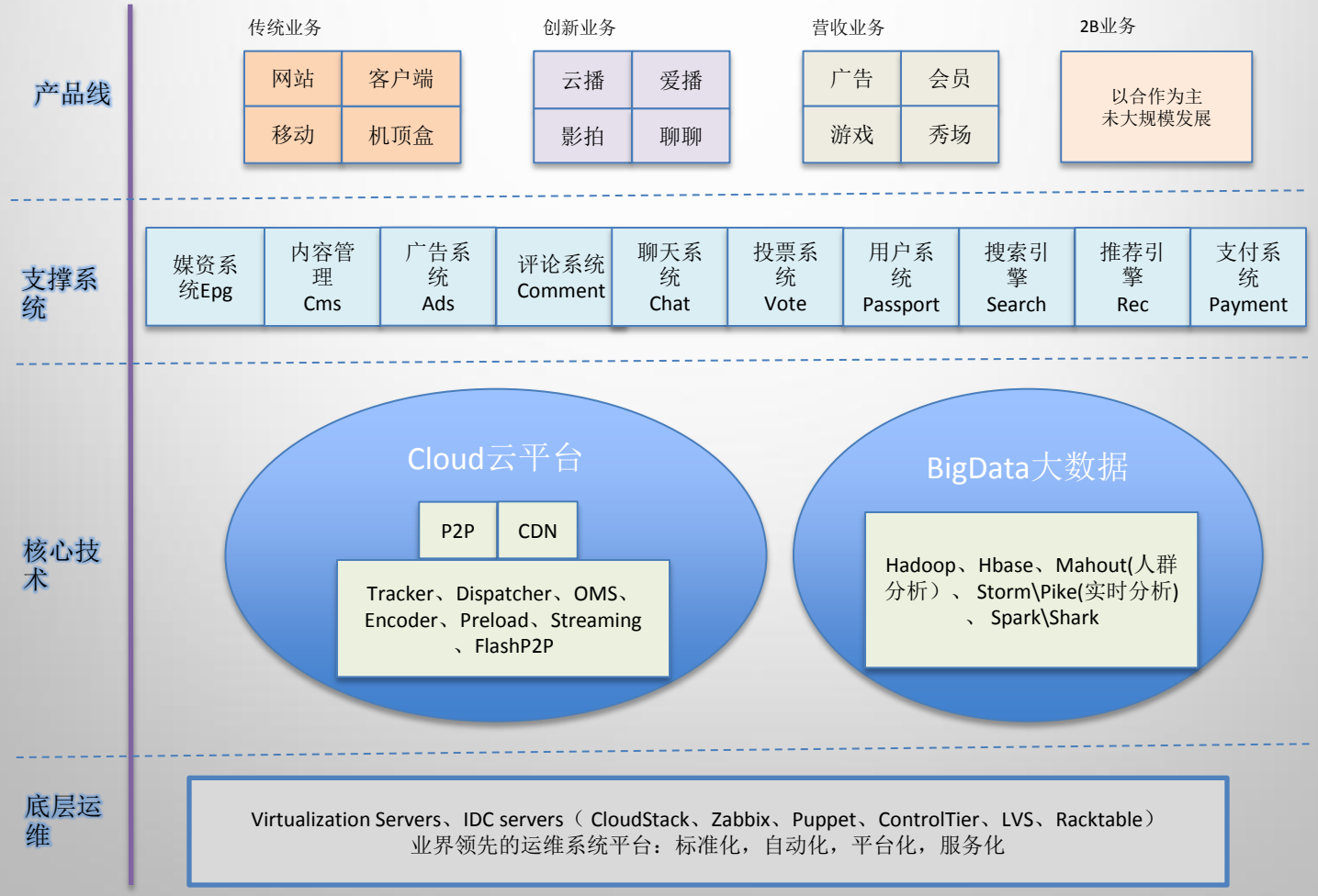
1998-2002	中国	复旦大学（FuDan Univ） 理学学士(B.S)

2011-now	PPTV	技术产品部-系统平台部
2007-2011	eBay	中国运营中心（COC）
2004-2007	EbaoTech	全球第一大保险IT解决方案提供商
2002-2004	金仕达卫宁	国内最大医疗IT系统提供商

PPTV的生产环境规模

?TB 视频CDN带宽	>100GB 网页CDN带宽	~ 10k 服务器	> 60次/周 自动化上线
> 200个 CDN节点	~10 个 核心IDC节点	> 400台 网络设备	> 10K 计算云CPU core
>10GB 核心机房带宽	150万个 监控点	40万个 告警点	> 4k 计算云节点
>10PB 分布式存储服务	> 10TB 分布式缓存服务	421个 核心IDC服务池	> 20亿次/天 单个服务池访问量

产品技术体系


产品线
传统业务
网站
客户端
移动
机顶盒
创新业务
云播
爱播
影拍
聊聊
营收业务
广告
会员
游戏
秀场
2B业务
以合作为主
未大规模发展
支撑系统
媒资系统Epg
内容管理Cms
广告系统Ads
评论系统Comment
聊天系统Chat
投票系统Vote
用户系统Passport
搜索引擎Search
推荐引擎Rec
支付系统Payment
核心技术
Cloud云平台
P2P
CDN
Tracker、Dispatcher、OMS、Encoder、Preload、Streaming、FlashP2P
BigData大数据
Hadoop、Hbase、Mahout(人群分析）、Storm\Pike(实时分析)、Spark\Shark
底层运维
Virtualization Servers、IDC servers（CloudStack、Zabbix、Puppet、ControlTier、LVS、Racktable）
业界领先的运维系统平台：标准化，自动化，平台化，服务化

挑战与特点

大中型互联网系统的特点：

- 高并发，高流量
- 高可用
- 海量数据
- 用户分布广泛，网络情况复杂
- 需求快速变更、发布频繁
- 快速扩张
- 业务调用复杂，容易产生级联风暴

目标

- 可伸缩性(Scalability)
  - 弹性伸缩
- 可用性(Availability)
  - 服务拆分
  - 多数据中心、灾备机房
  - Canary预警
- 敏捷性(Agility)
  - 持续部署
  - 自助服务和管理
  - 基础组件服务化(应用开发关注应用逻辑实现，底层组件服务化、组件化)
- 效率(Efficiency)
  - 资源池
  - 容量规划替代容量预测、
- 可监控性
  - 自动监控注册
  - 健康告警

路线

■硬件资源
■系统
■应用程序
■技术栈

标准化

数据化

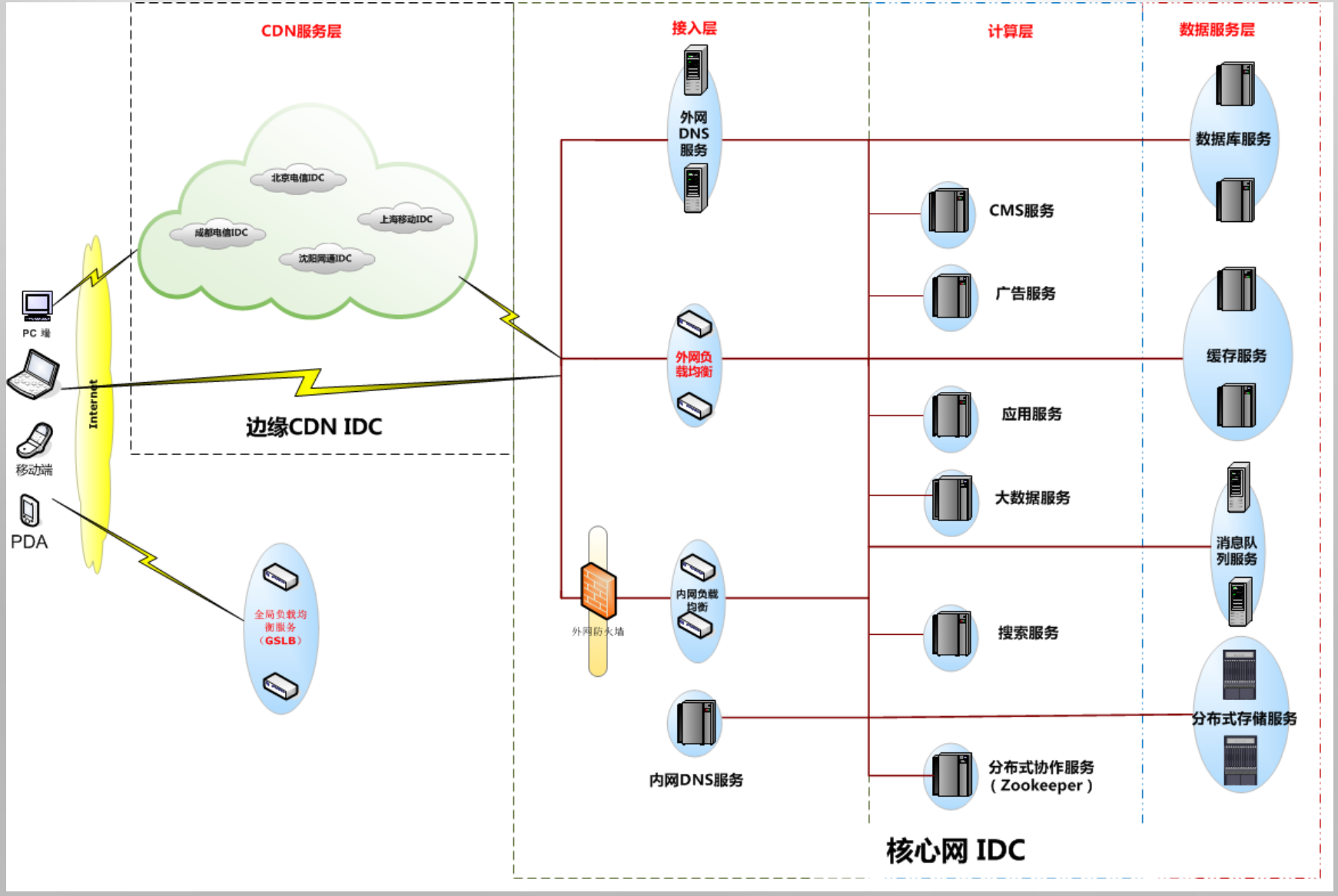
■基础服务平台
■技术解决方案

■配
■应用部署
■监控告警
■资产管理

服务化

■服务资源池
■快速响应、按需分配
■容量规划

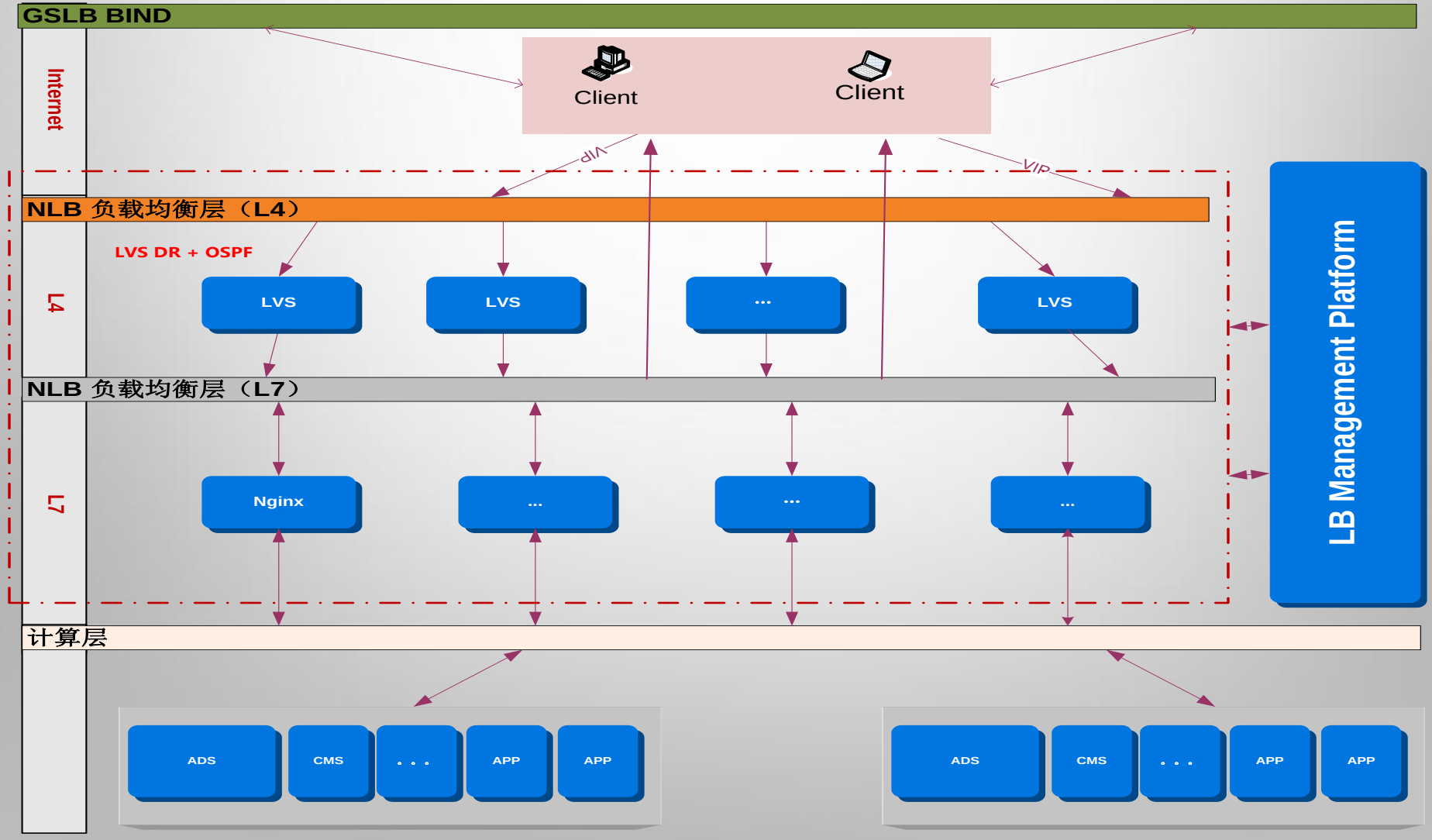
web服务技术架构抽象



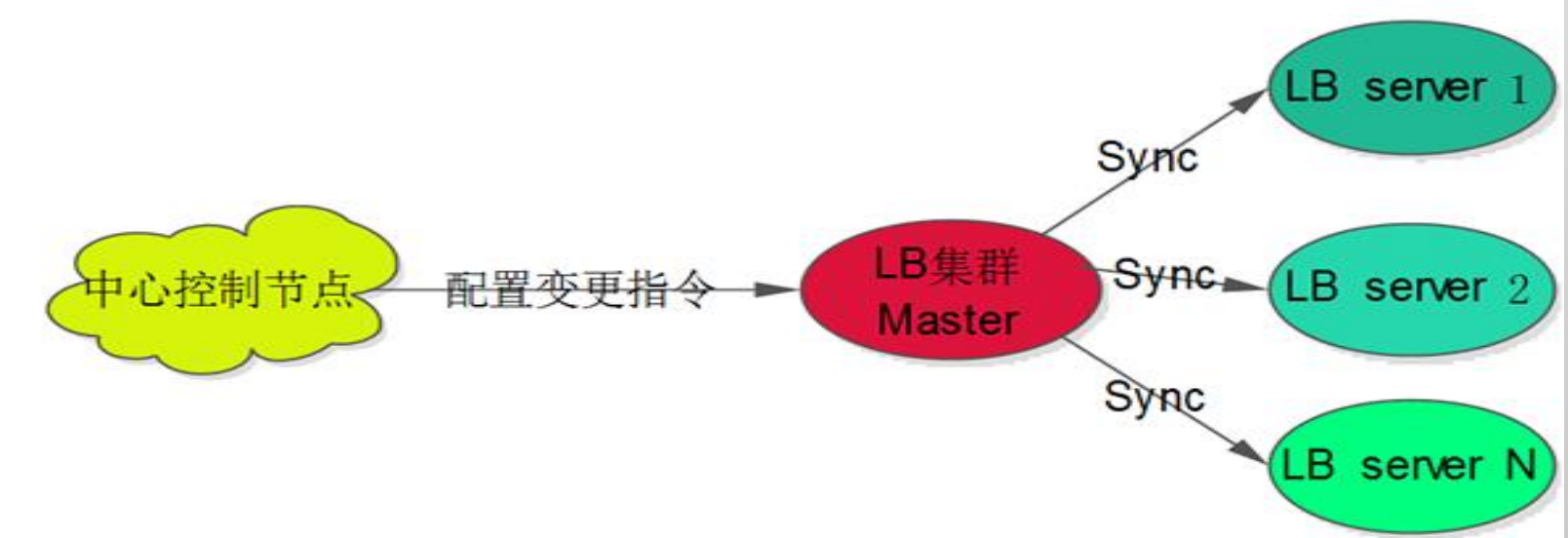
总览

- IaaS&PaaS
  - 计算云(Cloudstack)
  - 分布式缓存服务(Cache As A Service)
  - 存储云(Openstack Swift,GlusterFS,HDFS)
  - 分布式消息队列服务(MQ As A Service)
  - PPKeeper(Zookeeper As A Service)
  - 分布式日志收集服务(Events As A Service)

Iaas&PaaS—负载均衡服务



Iaas&PaaS—负载均衡服务



- 标准化、批量配置
- 水平扩容
- API调用接口
  - 发布系统
  - 计算节点online\offline
  - 权重

Iaas&PaaS—Cloudstack计算云

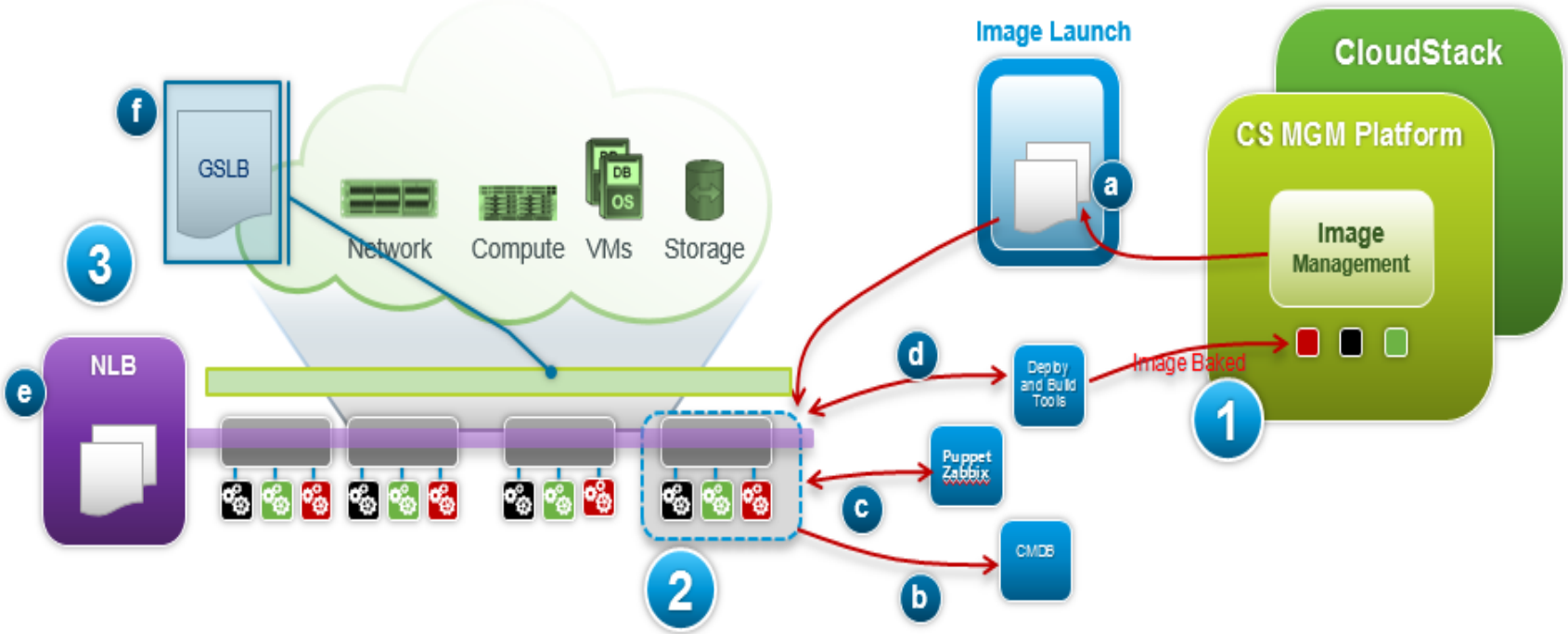
- 非标准化，质量不可控
- 流程长，路径依赖多，人力成本高
- 缺乏弹性

Iaas&PaaS—Cloudstack计算云

基于CloudStack的部署管道演进

Iaas&PaaS—Cloudstack计算云

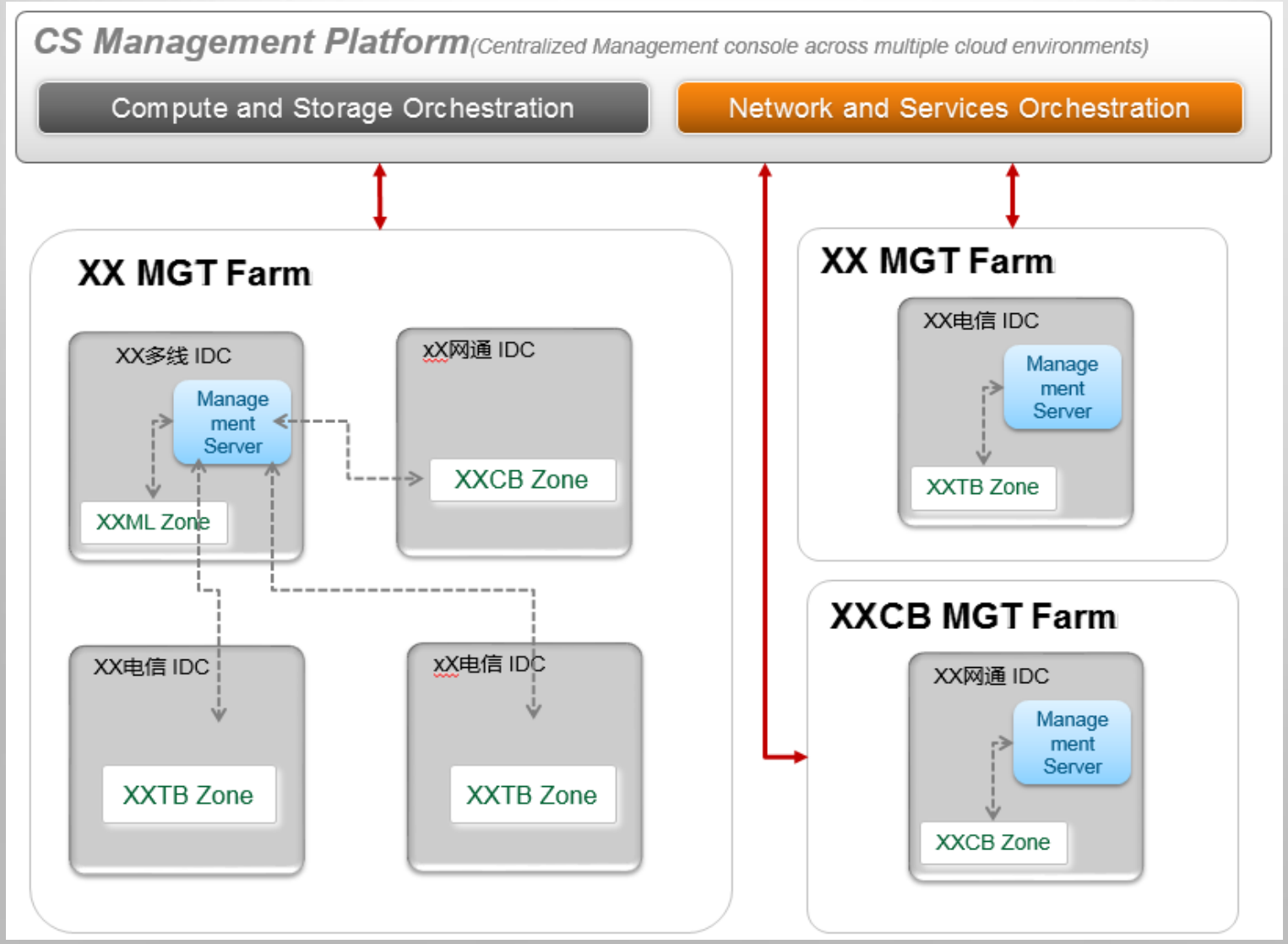
基于CloudStack的部署管道演进



1 新VM实例创建
a VM实例启动
b 系统注册
c 系统初始化和系统配置
2 APP初始化和配置
d APP初始化、监控和相关配置，注册信息到CMDB
e APP注册到本地负载均衡
3 APP启用
f APP注册到GSLB

Iaas&PaaS—Cloudstack计算云

基于CloudStack的部署架构在PPTV



Iaas&PaaS—Cloudstack计算云

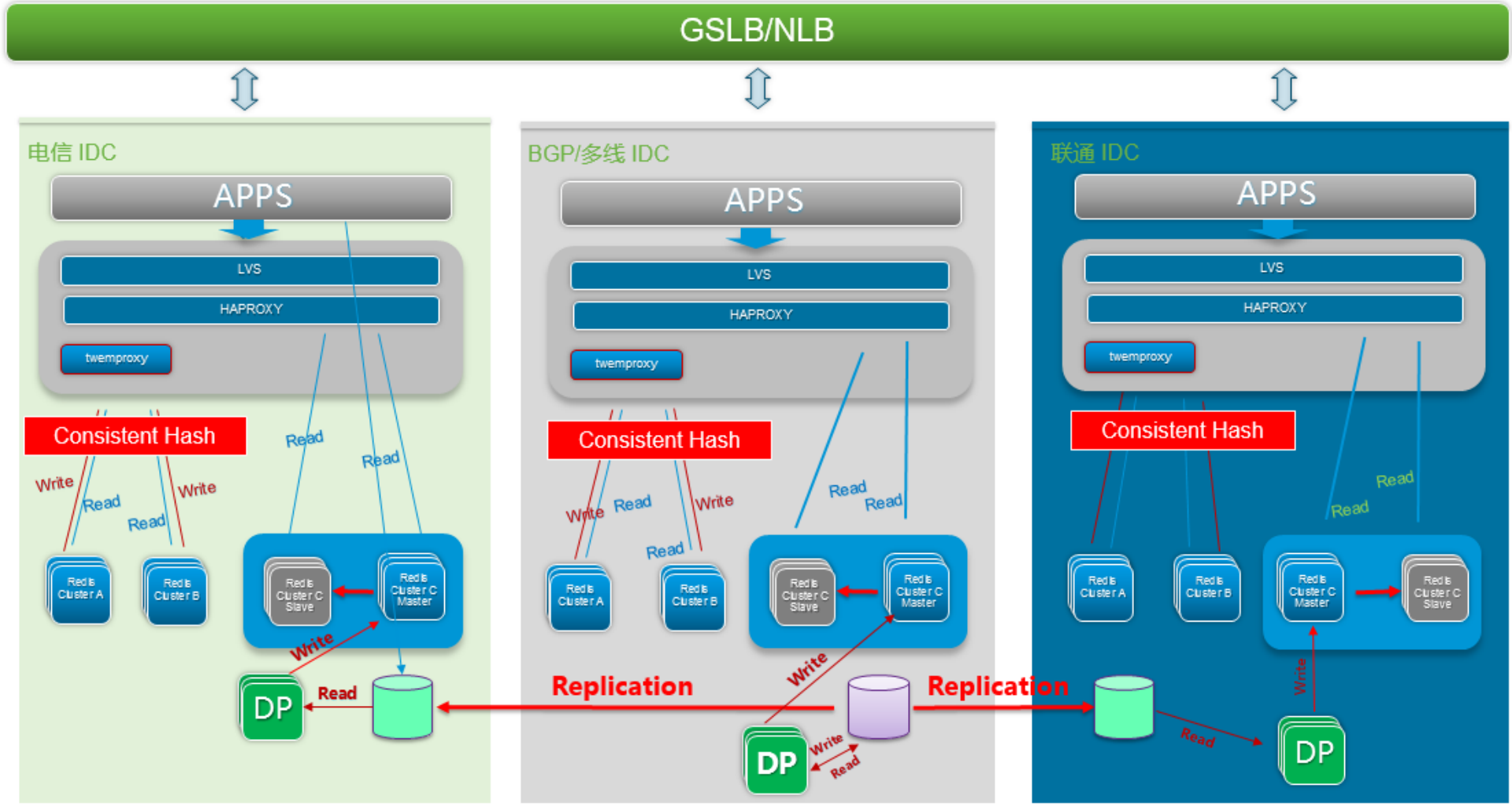
Iaas带来益处

- 自助服务
  - 知识和技能可以通过工具和系统转移替代
  - 解决路径依赖问题、缩短环节
- 管理自动化
  - 降低手动操作和犯错的机会
  - 减少40%基础服务管理流程
- 标准化和效率
  - 10倍应用部署效率提升
  - 确保应用部署一致性及质量
- 敏捷和弹性
  - 弹性化更好地为计划外停机时间做好准备
  - 减少故障发现和恢复时间
  - 基于资源池的容量规划

Iaas&PaaS—缓存服务

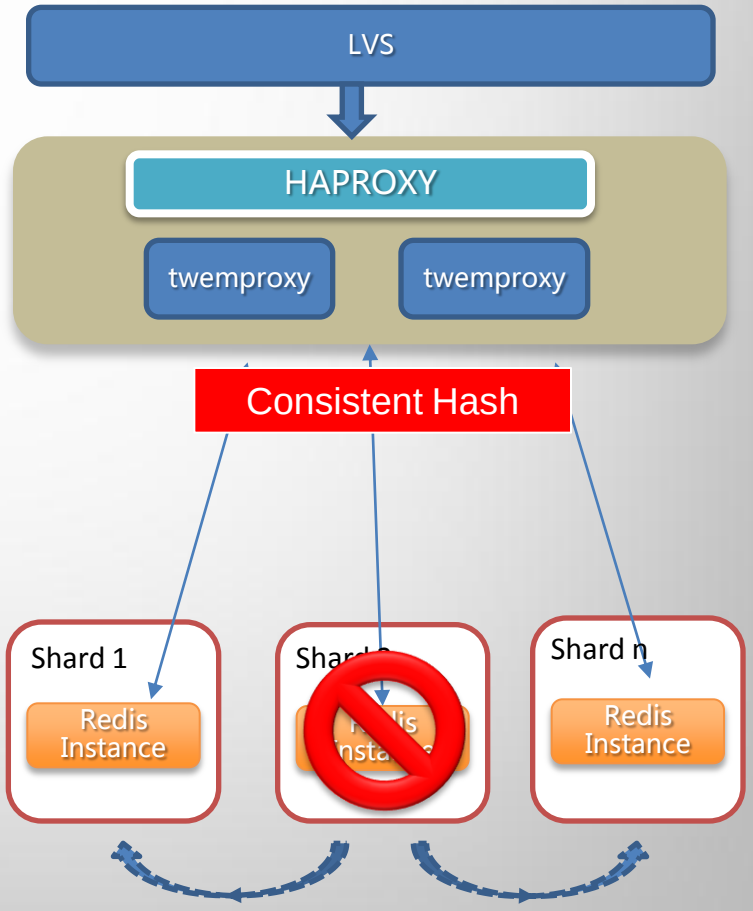
- 跨IDC分布式全局缓存服务
- 数据分区(Consistent Hash)
- 集群成员维护和失败检测
- 高可用
- 弹性容量管理
- 简化的应用调用和配置

Iaas&PaaS—缓存服务



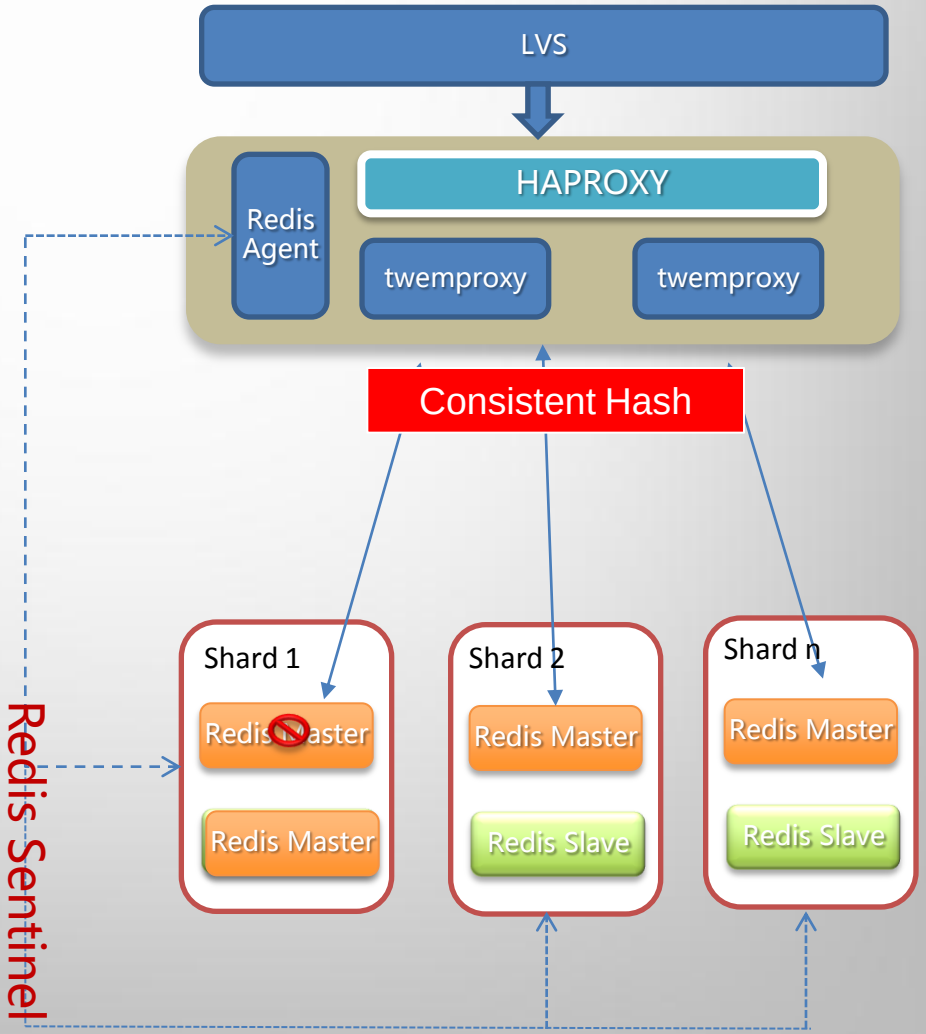
缓存服务-HASH模式

- HASH模式
  - 提供缓存服务，非持久化存储
  - 集群提供自动数据分片
  - 单个Redis实例宕机只影响部分数据
  - 动态扩容



缓存服务—HASH+HA模式

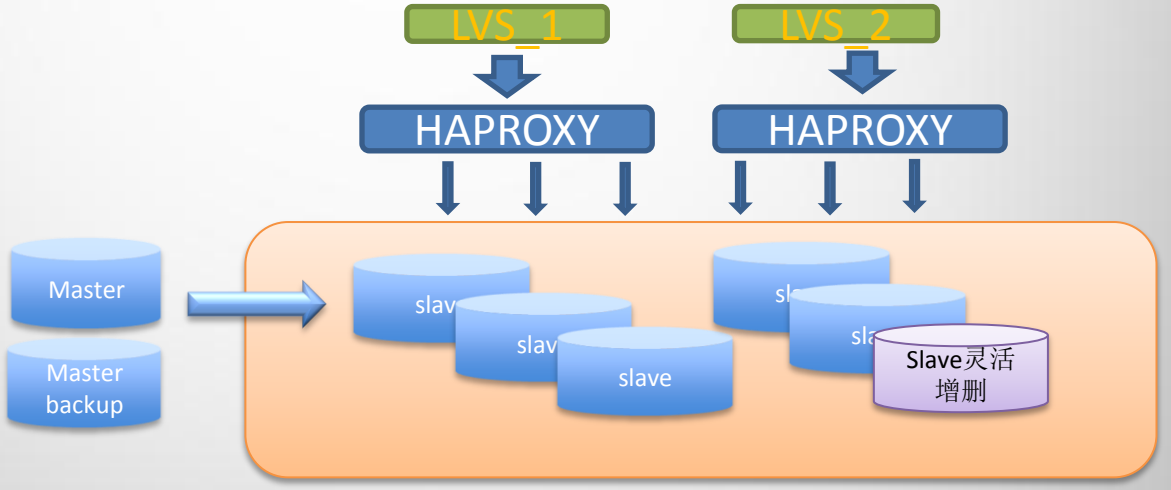
- HASH+HA
  - 适用于有存储需求的业务
  - 在Hash模式基础上增加单实例镜像
  - 基于Redis Sentinel实现单实例故障自动主从切换
  - 支持动态扩容



Iaas&PaaS—缓存服务

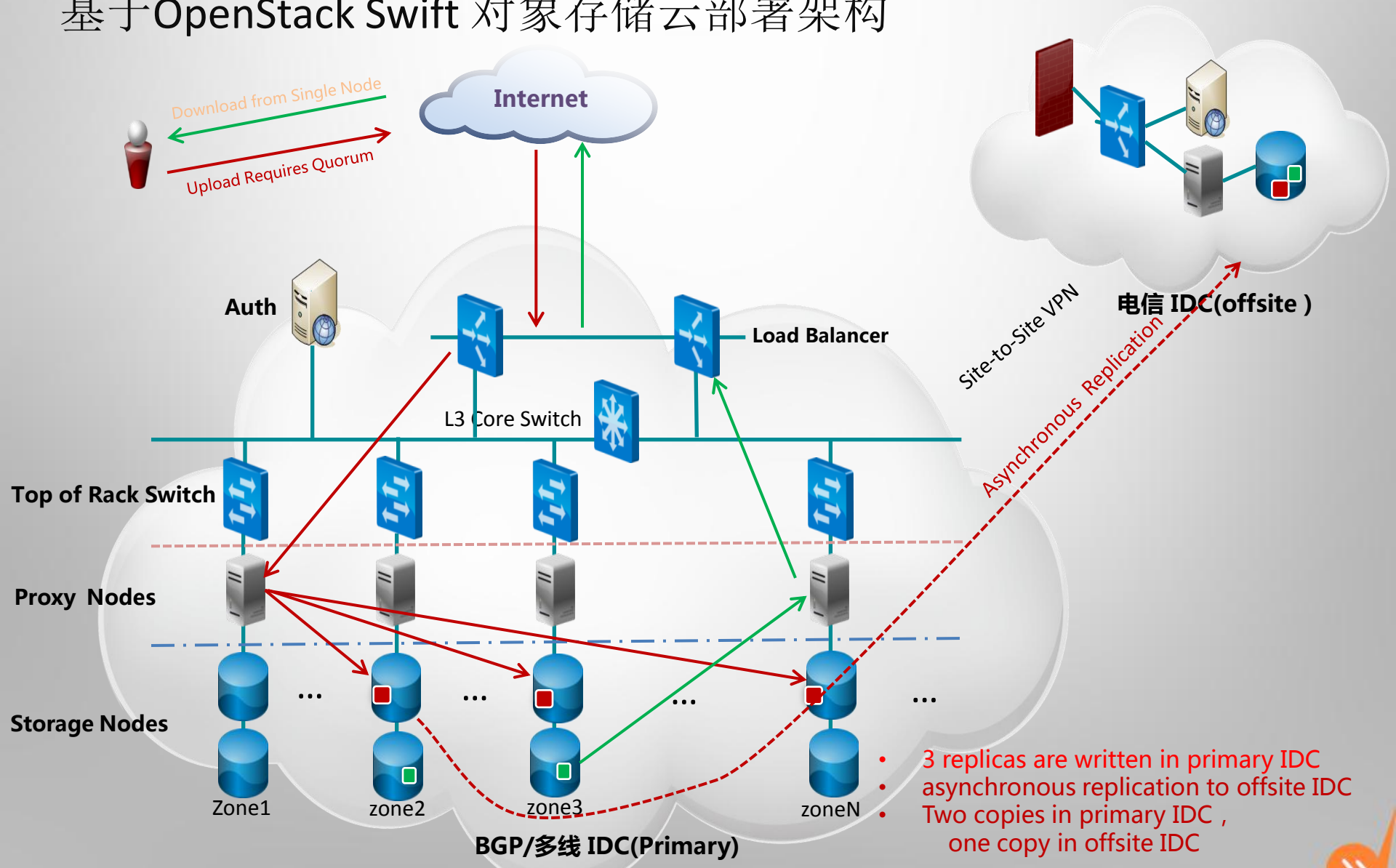
读写分离的缓存服务

- 多机高可用
- 故障节点自动下线
- 权重控制
- 在线水平扩容



Iaas&PaaS—分布式对象存储

基于OpenStack Swift 对象存储云部署架构



- 3 replicas are written in primary IDC
- asynchronous replication to offsite IDC
- Two copies in primary IDC ,
  one copy in offsite IDC

Iaas&PaaS—分布式对象存储

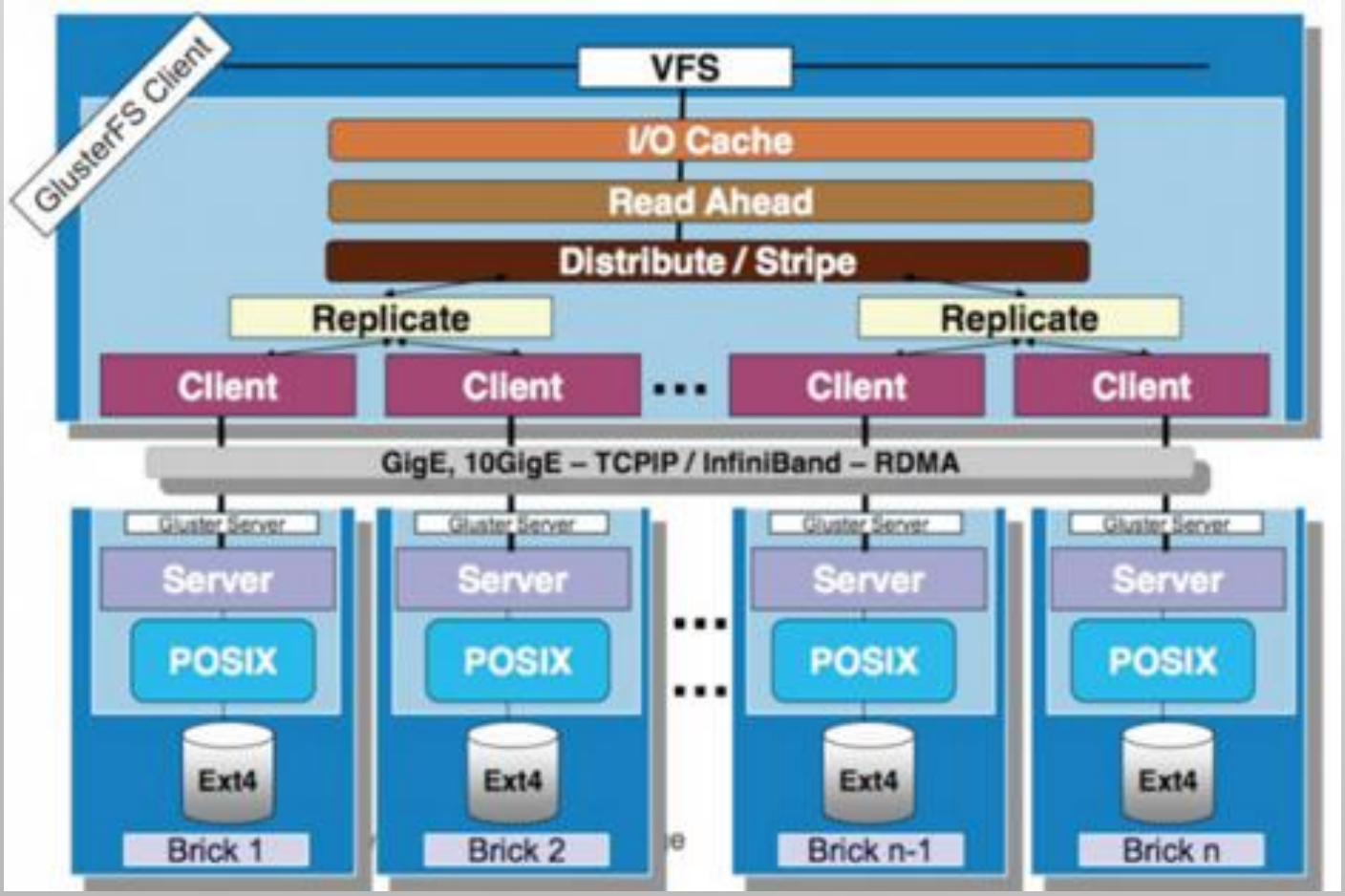
OpenStack Swift 部署最佳实践

- 主节点部署在多线或者BGP机房
- Swift 使用DR集群模式构建
  - 2份数据存储在主节点，1份数据存储在DR节点
  - 主节点IDC故障，服务降级切换到DR站点只读模式
- 负载均衡使用LVS DR模式
  - 使用OSPF
- Linux 内核调优
  - Network buffer
  - 中断均衡
  - 借鉴和引入TaoBao Kernel调优经验
- Swift Proxy Node使用10G网络
- 单个Container存储数据量小于100w
- 使用SSD在Account和Container server
- 所有日志发送到远程存储，避免使用本地syslog方式收集日志（Python Bug）

Iaas&PaaS—分布式文件存储

GlusterFS分布式文件系统

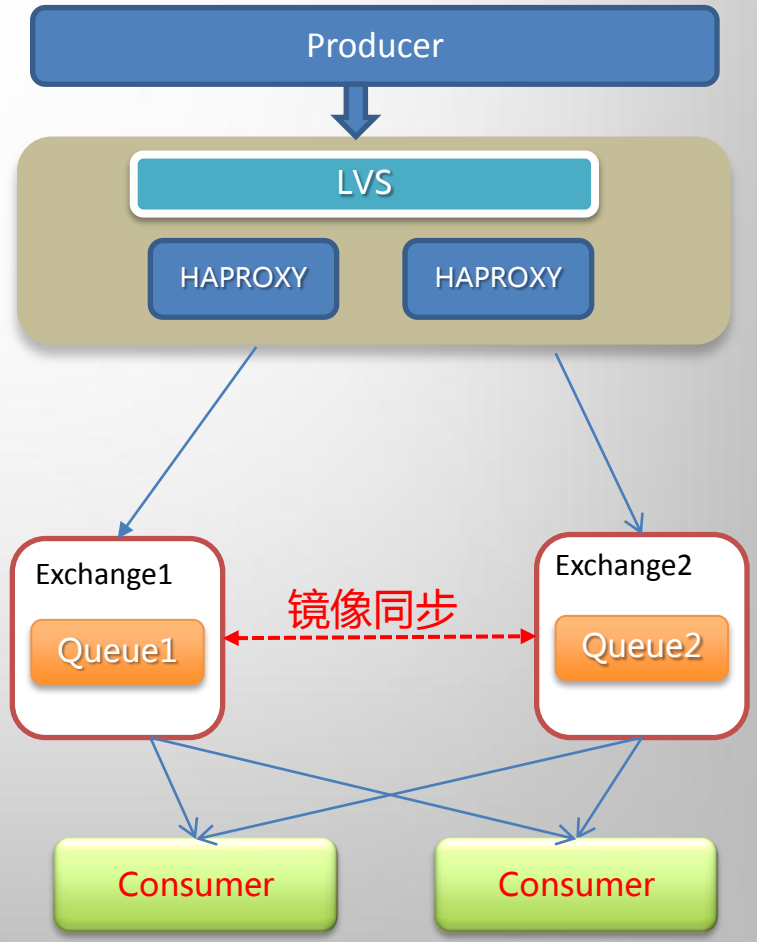
- 多点挂载写入
- 图片等静态资源的CDN源站



Iaas&PaaS—消息队列服务

分布式消息队列服务(MQ As A Service)

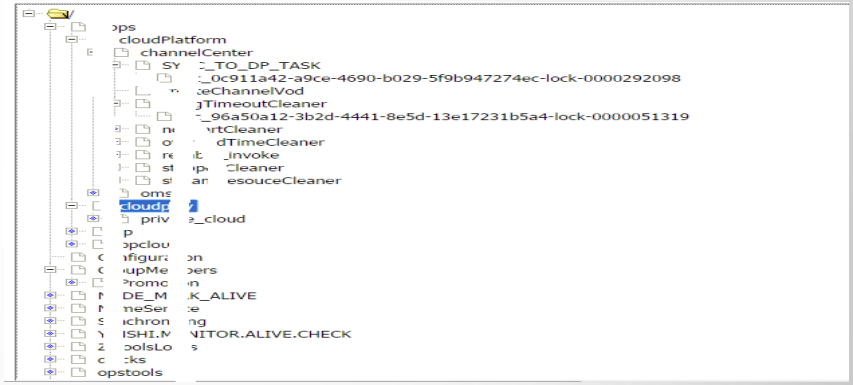
- RabbitMQ
- 部署模式
  - 单机房部署，集中式管理
  - 镜像模式: 提供高可用
  - 直接路由模式：提供高性能
- 统一MQ管理平台
  - MQ实例快速部署
  - 监控
  - 容量管理

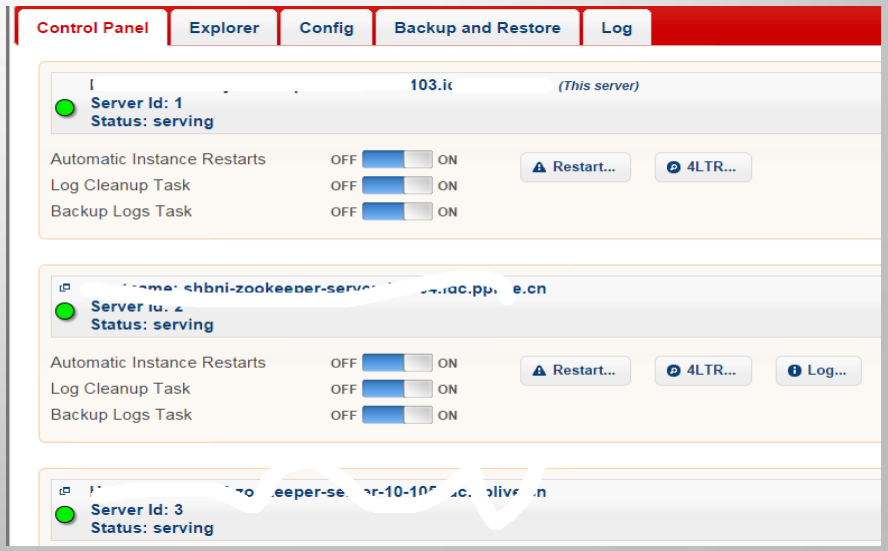


Iaas&PaaS—PPKeeper服务

PPKeeper(Zookeeper As A Service)

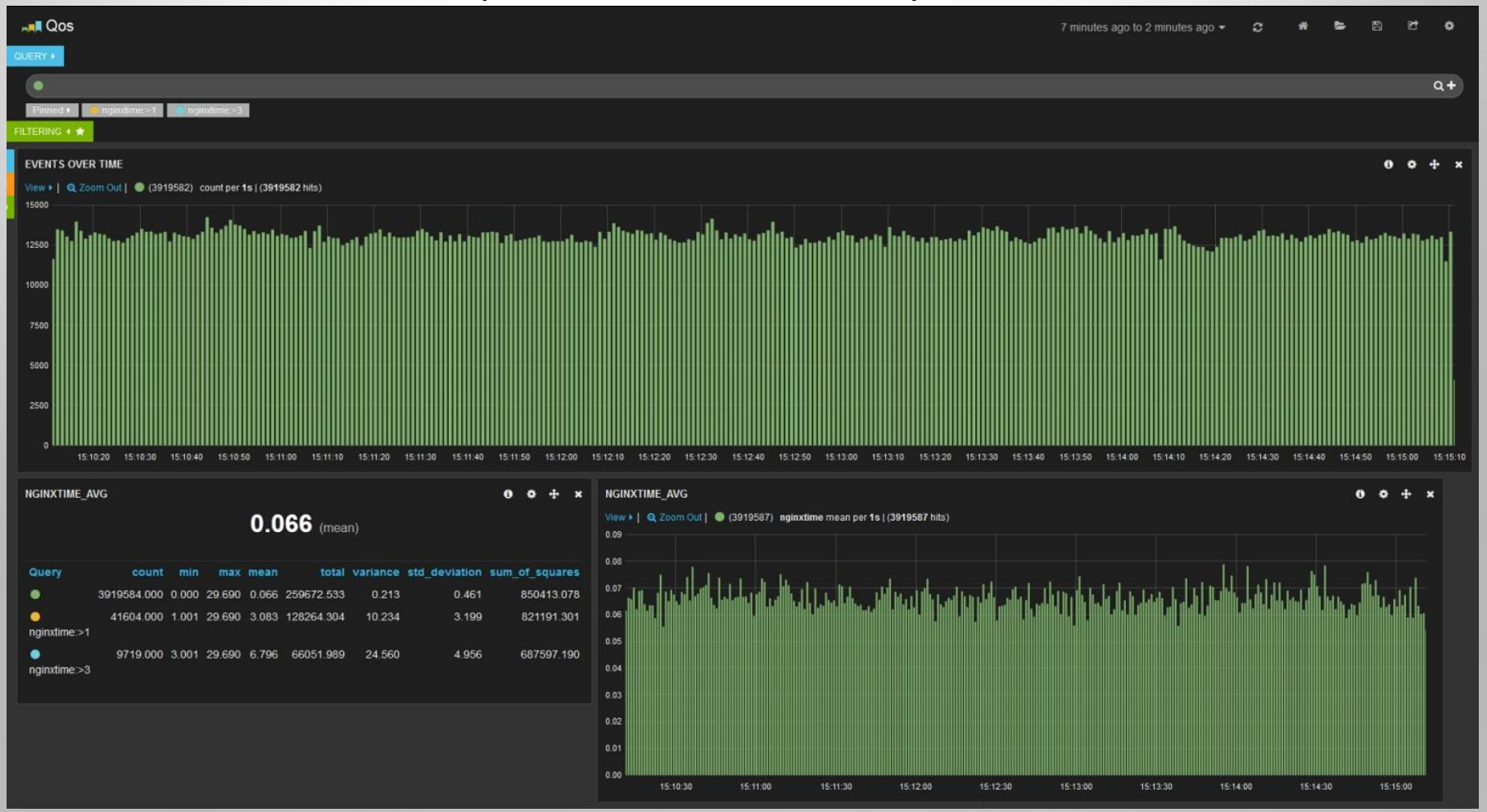
- PPKeeper
  - Zookeeper3.4.5
  - Java Client: Curator
  - 服务端管理: Exhibitor
  - Znode可视化管理
  - 在线扩容、备份和恢复
- 分布式锁
- 命名服务
- 集群管理
- 数据发布与订阅





Iaas&PaaS—日志收集服务

分布式日志收集管道(Events As A Service)



Iaas&PaaS—日志收集服务

分布式日志收集服务(Events As A Service)

- 事件传输和捕获：Flume-NG
  - 扩展SDK添加业务相关属性，简化app端配置和部署
  - 每个App Server启用两个Flume agent，分别提供APP层和系统安全日志传输
  - 基于avro 压缩传输到Flume集群会聚和路由，部分敏感日志使用加密传输
- Sinks: Kafka,HDFS,File System
- Data Bus：Apache Kafka
- 流处理(Storm)：实时访问统计、Security和系统告警、DNS、防火墙日志
- 实时索引和搜索(ElasticSearch)：APP日志分析查询、QOS统计分析
- 可视化展现(Kibana)
- HDFS：M/R 离线计算；长期数据存储

总览

- MaaS(Management As A Service)
  - Racktable数据中心管理
  - 自动化服务器安装平台
  - 配置管理平台
  - 监控和告警管理平台
  - 计算层自动化管理平台
  - 自动化发布系统
  - 权限及消息订阅系统

MaaS— RackTable数据中心管理

IDC机房管理
- 机柜位置
- 服务器信息
- 网络连接拓扑

数据中心资产管

Main page : Rackspace Objec

Browse | Manage locations | Manage rows

Location	Row	Rac
上海多线	11号楼3F-A	11
	11号楼3F-A (continued)	11
上海多线	shbnj_3K	sh
		sh

11号楼3F-A08

shbnj_3K09

Local name	Visible label	Interface	L2 address	Remote object and port	Cable ID
kvm		KVM (host)			
eth0		1000Base-T	00:E0:81:D9:45:54		
eth1		1000Base-T	00:E0:81:D9:45:55	WA00638H1205 GigabitEthernet1/0/21	

summary

名称:	JA03125L1109
类型:	Server
资产编号:	JA03125L1109
CPU品牌:	Intel
CPU信息:	2.40G*2
远程管理卡IP:	10.208.254.138
购买日期:	08/31/2011
内存大小:	32G(8G*4)
品牌及型号:	LenovoR520 G7
设备尺寸:	2U
设备归属:	有产权有使用权
设备序列号:	empty
设备种类:	机架服
所属机房:	上海多线
使用期限:	01/01/1970
网卡1速度:	千兆
网卡1详细:	Intel Corporation 82574L Gigabit Network Connection
网卡2详细:	Intel Corporation 82574L Gigabit Network Connection
硬盘类型:	SAS/SATA
硬盘容量:	22T(300G/SAS+2T/SATA*11)
已贴资产标签:	Yes
资产状态:	在线中

MaaS—自动化服务器安装平台

高效率： 安装100台服务器，只需1个人花10分钟

标准化： 无论谁安装，系统都统一线上标准

MaaS—磁盘管理平台

磁盘信息状态直观展现
故障磁盘**自动屏蔽/上线**
通知/记录/报表

监控告警 → 触发事件 → 自动下线 → 故障修复 → 自动上线

机器

资产编号	品牌及型号	ip	位置	业务code	业务负责人	维保结束日期
JA01393D1103	DELLR710	10.203.4.3	shtb_G10	osd-sqlepg	dba	在保

磁盘信息

返回

磁盘

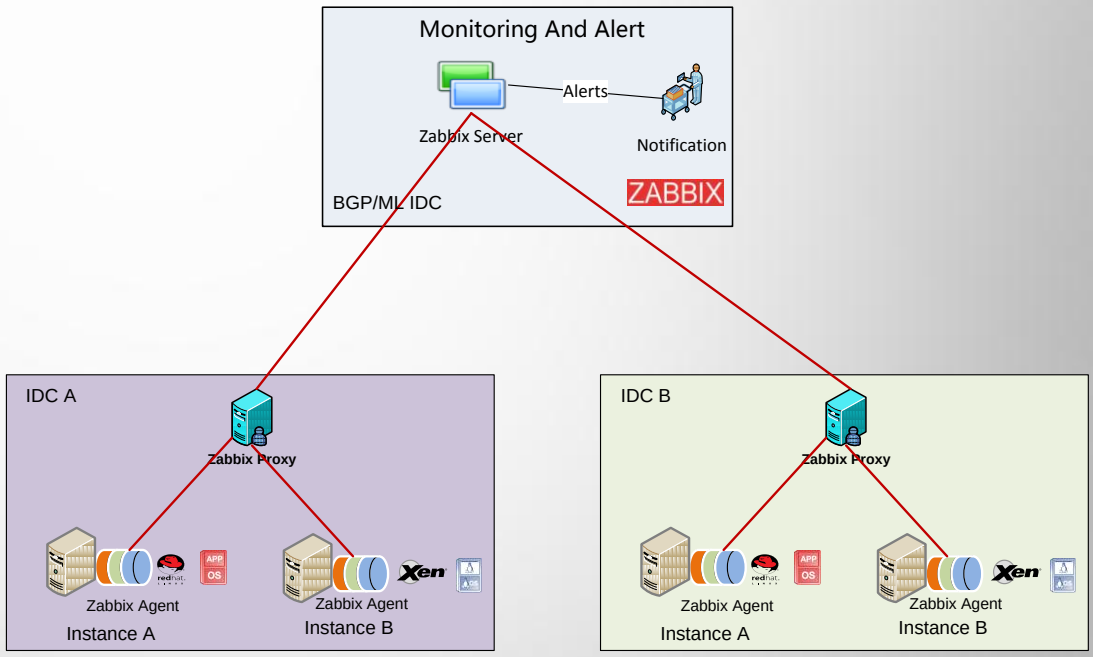
每页显示 20 条记录　　快速搜索

Raid Id	Raid Level	盘符	槽位号	SN	接口类型	容量(G)	健康状态	使用状态	磁盘灯	记录
0	10	sda	0	D004PB1077YP	SAS	299	ok	在线	不闪	记录
0	10	sda	1	6XP1FRDQ	SAS	299	ok	在线	不闪	记录
0	10	sda	2	WX61EC1WRD82	SAS	299	ok	在线	不闪	记录
0	10	sda	3	D004PB107698	SAS	299	ok	在线	不闪	记录
0	10	sda	4	6XP1G37Y	SAS	299	ok	在线	不闪	记录
0	10	sda	5	6XP1FSCZ	SAS	299	ok	在线	不闪	记录
			6	D004PB10780Y	SAS	299	ok	在线	不闪	记录
			7	D004PB107807	SAS	299	ok	在线	不闪	记录

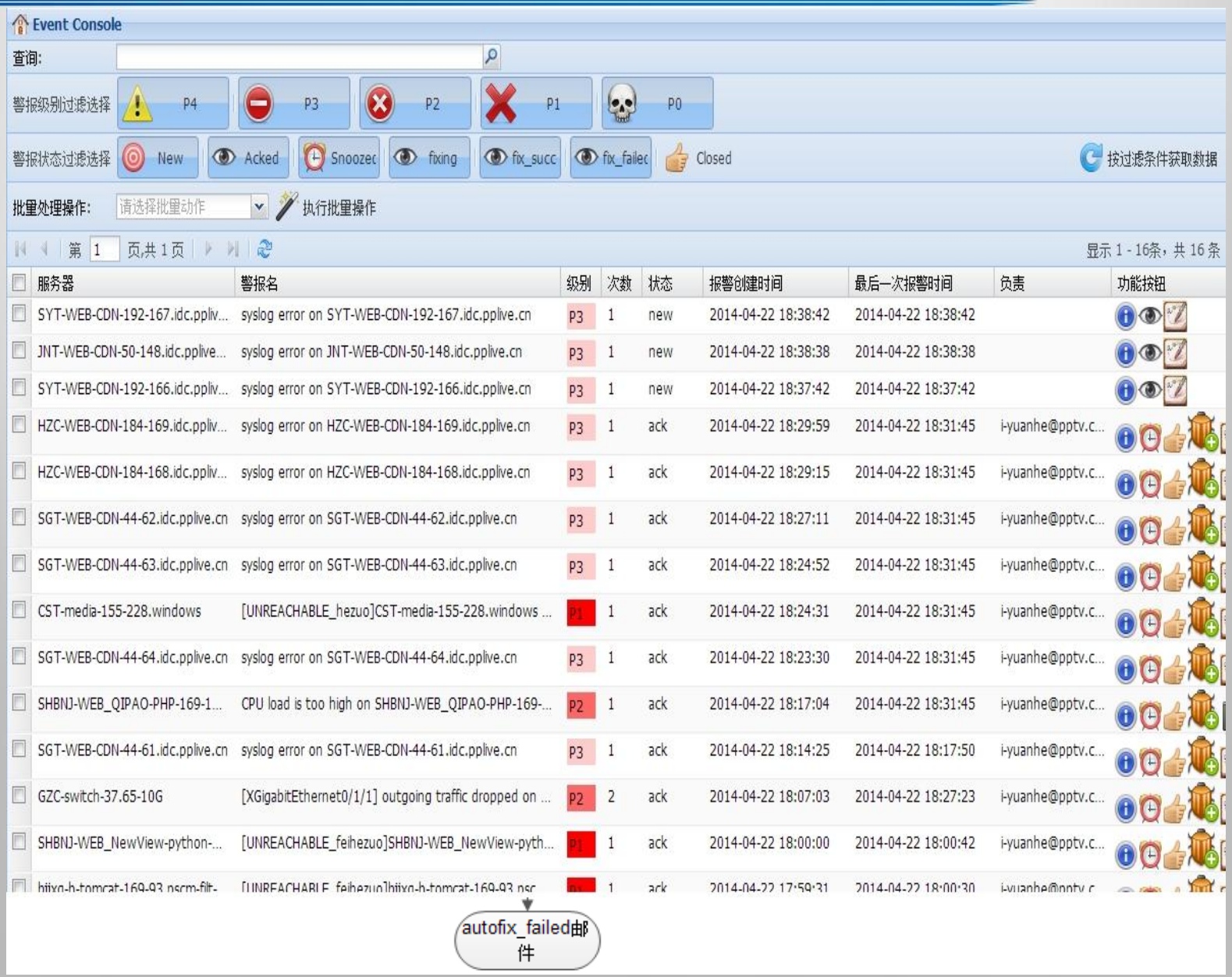
当前显示 1 到 8 条，共 8 条记录　　上一页 1 下一页

MaaS—监控和告警服务

- 硬件层监控
  - 物理硬件健康状况
- 系统层监控
  - OS级别相关监控
  - 安全
- 应用层监控
  - 应用相关健康状况监控
- 容量监控和规划
- 告警服务
- 监控信息订阅



MaaS—监控和告警服务

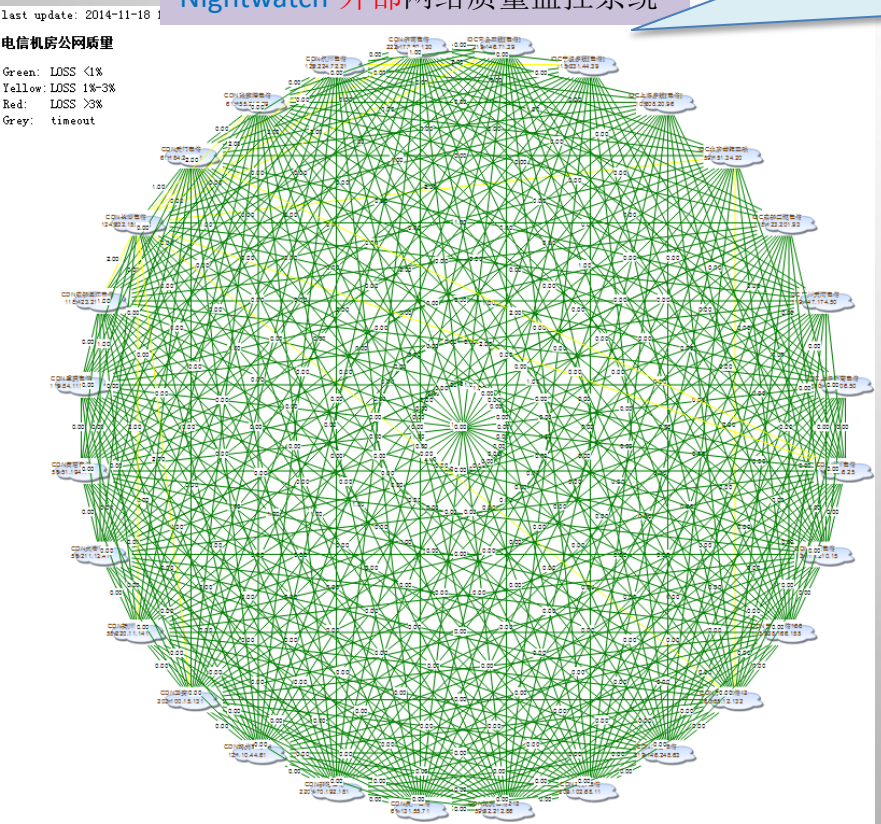


网络质量监控系统

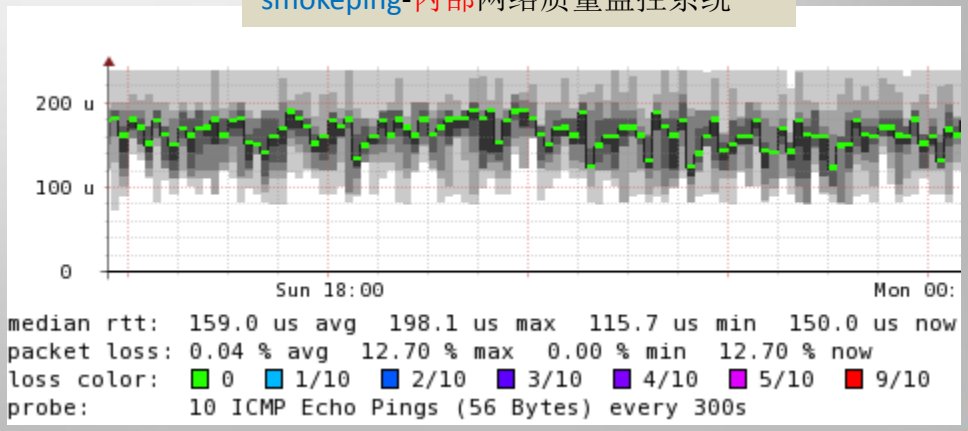
IDC网络质量 = 【外部质量+内部质量】 = 互联网业务的**基石**

直观：颜色反映质量，一目了然
全覆盖：核心网+CDN+内网VPN
自动报警：邮件+短信
历史查询：任意时间追溯

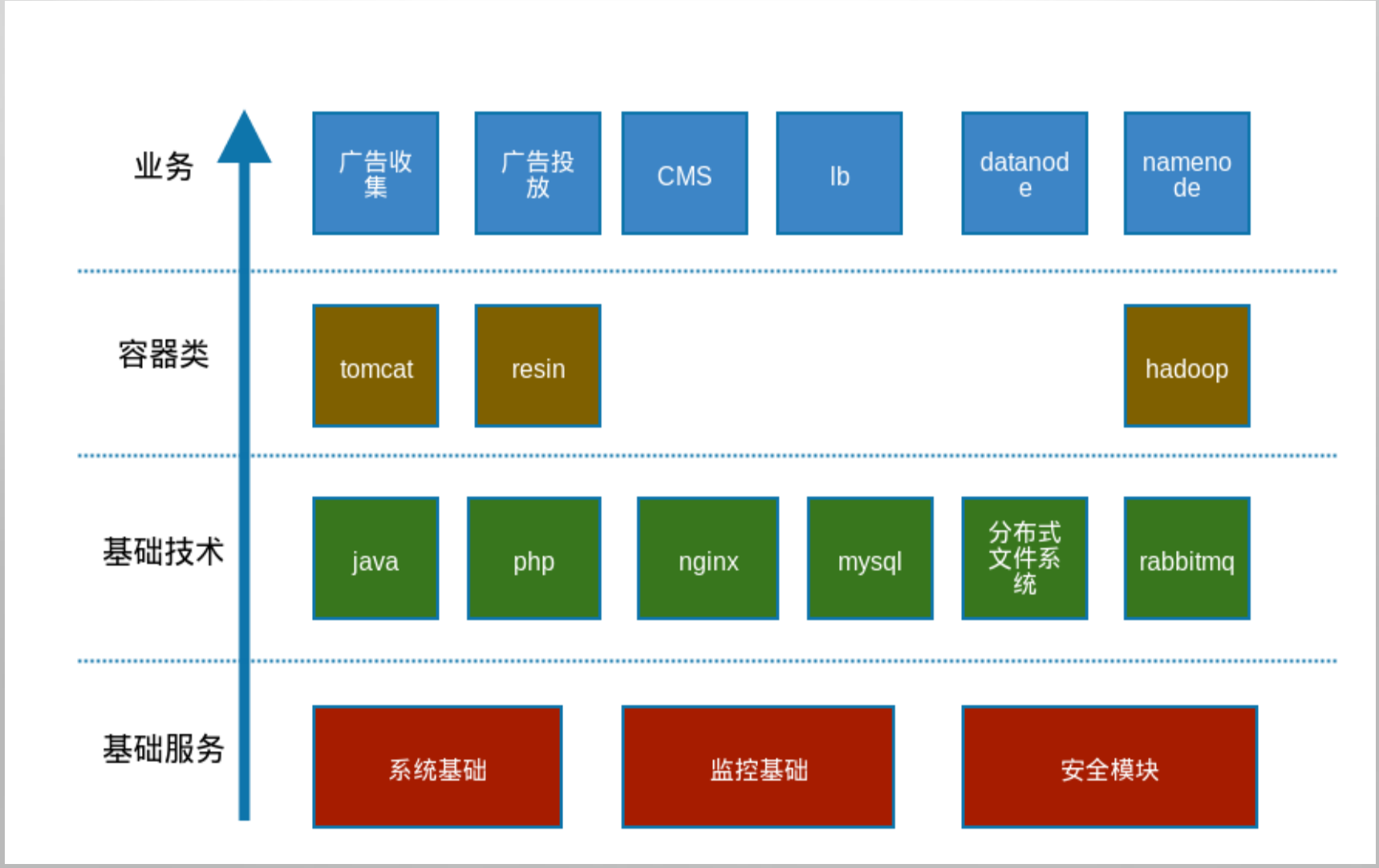
Nightwatch-外部网络质量监控系统



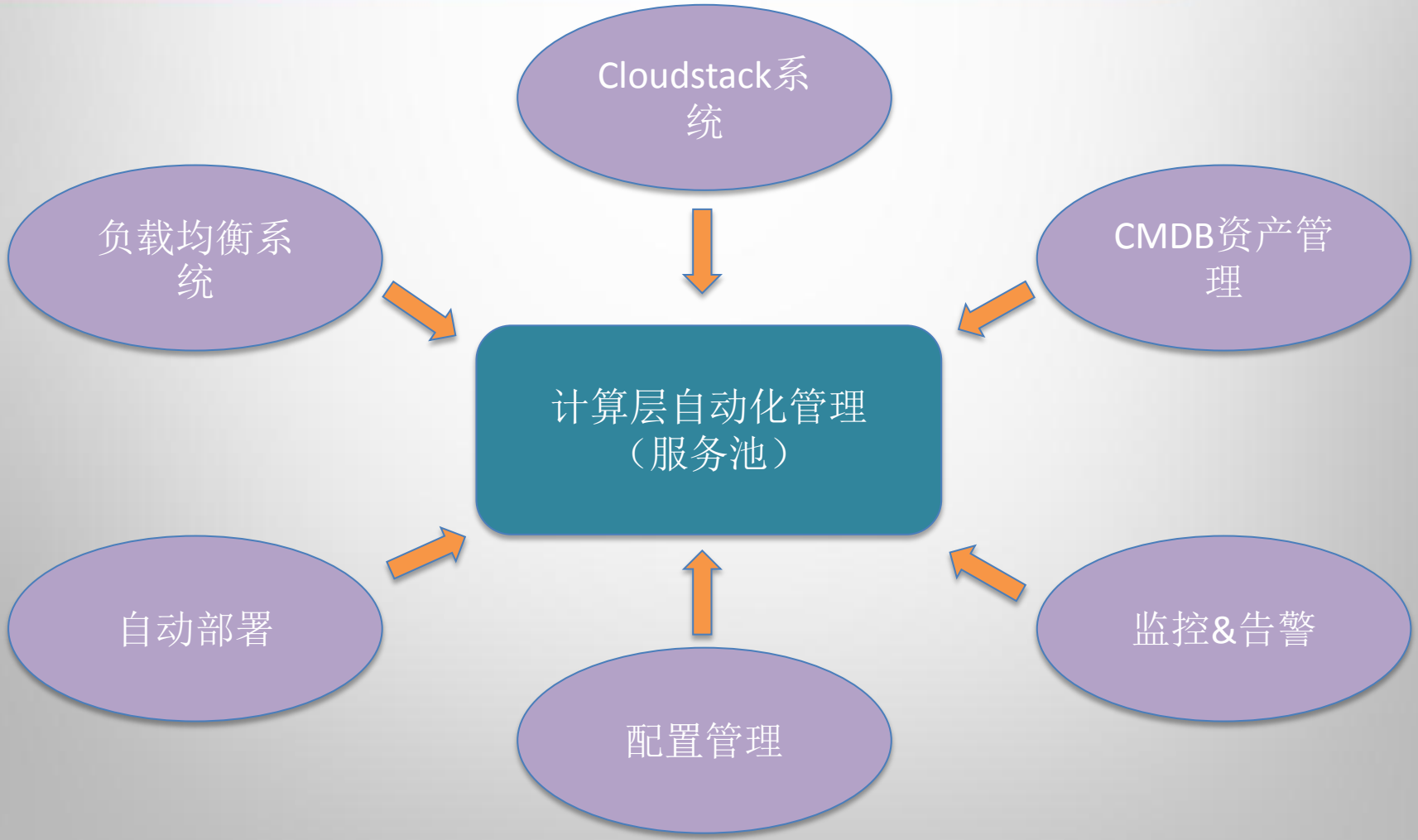
smokeping-内部网络质量监控系统



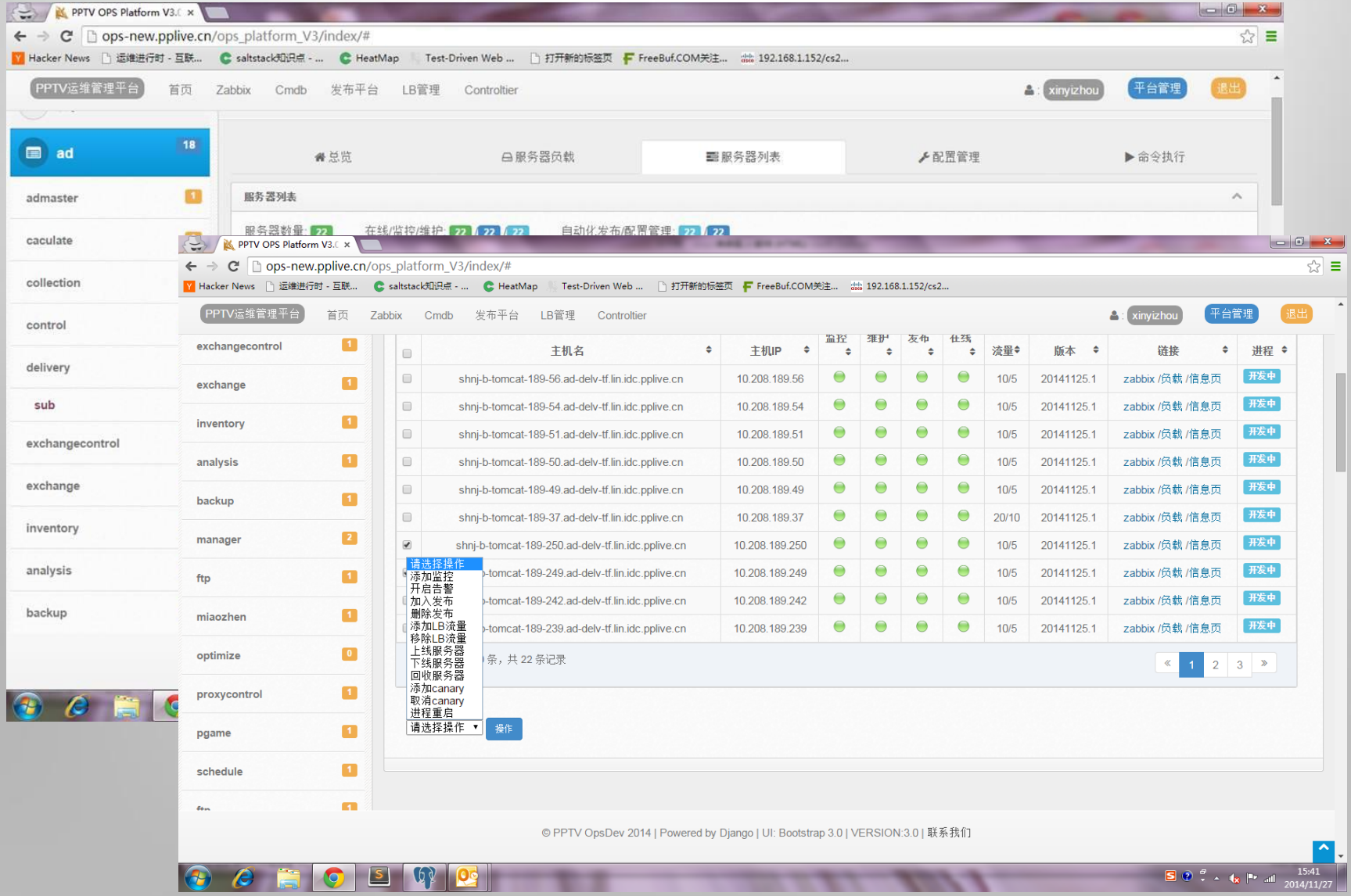
MaaS—配置管理平台


业务
广告收集
广告投放
CMS
lb
datanode
namenode
容器类
tomcat
resin
hadoop
基础技术
java
php
nginx
mysql
分布式文件系统
rabbitmq
基础服务
系统基础
监控基础
安全模块

MaaS—计算层自动化管理



MaaS—计算层自动化管理



MaaS—自动化部署系统

多服务池类型支持
- JAVA/PHP/NodeJS/软件包

机房串并行控制

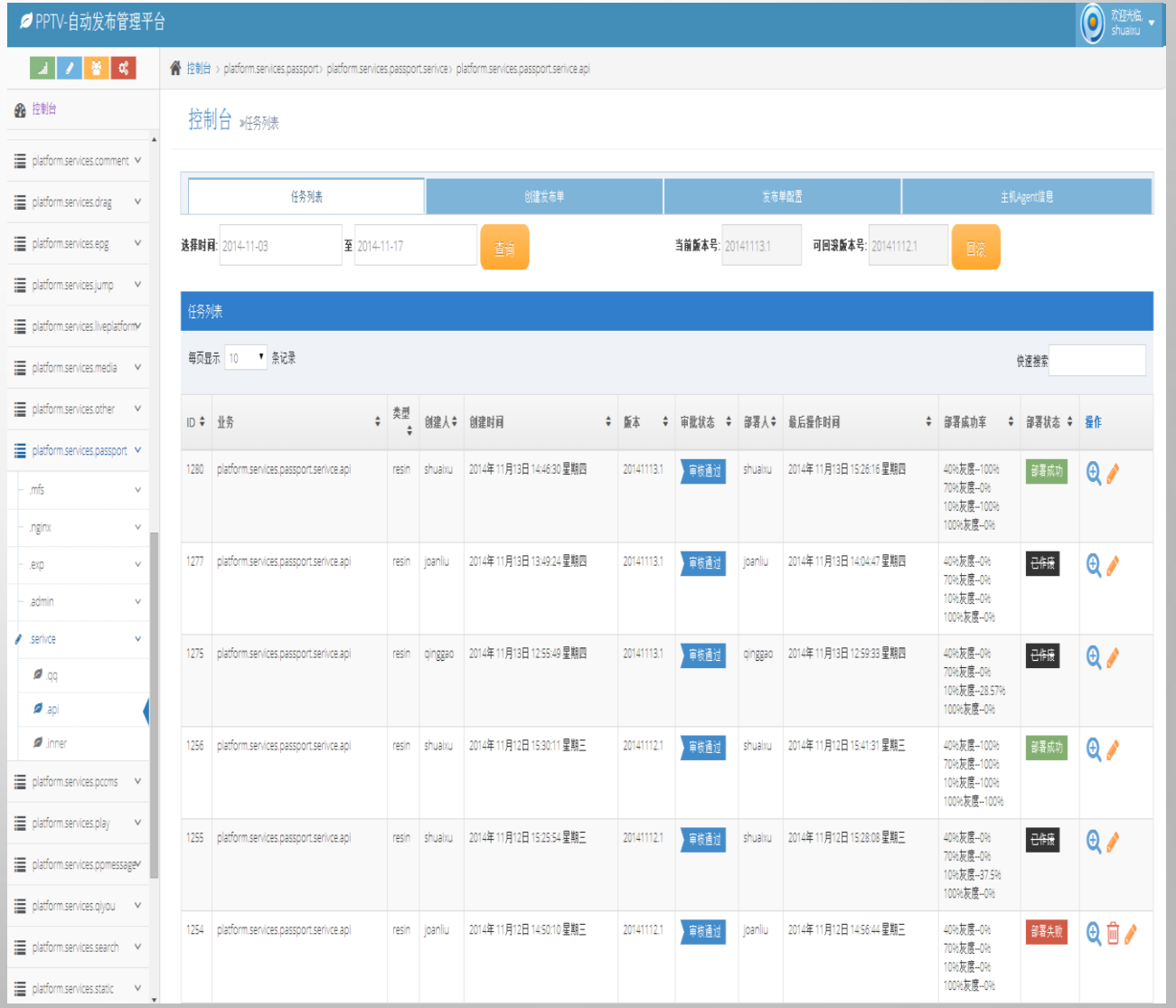
灰度发布

集成负载均衡自动化管理
- 离线，发布，验证，上线一体化

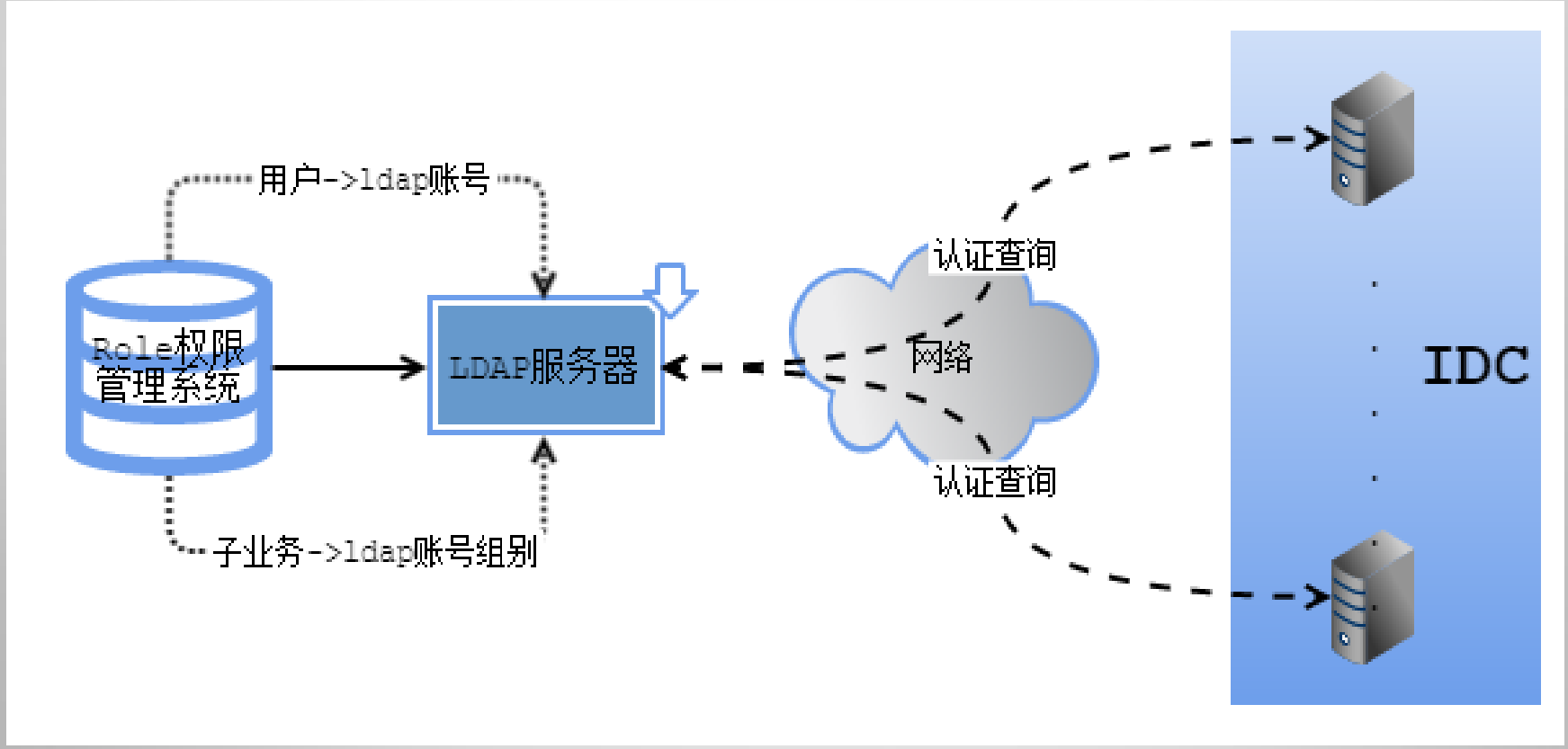
集成权限及消息订阅系统
- 通知到服务池相关测试、产品、研发

集成监控告警系统
- 自动屏蔽发布过程中产生的误告警

集成计算层管理系统
- 自动加入、删除计算节点



MaaS—权限及消息订阅系统



- 分业务及服务池进行权限管理
- 自动发布系统权限
- 服务器登录LDAP账号管理
- 生产环境消息订阅（告警、发布、变更）

➢ 体会和经验

- ◆ 高内聚，松耦合的多系统松散结构
- ◆ 小步快跑，快速迭代
- ◆ 80:20原则
- ◆ 紧跟时代潮流，新技术提前预研

➢ Next Step

- ◆ 平台化
- ◆ 服务化
- ◆ 数据化

Q&A

谢谢！

joechen@pptv.com