

2014  
年度

# PPTV系统平台建设实践

演讲人：陈文春

2014-11-10

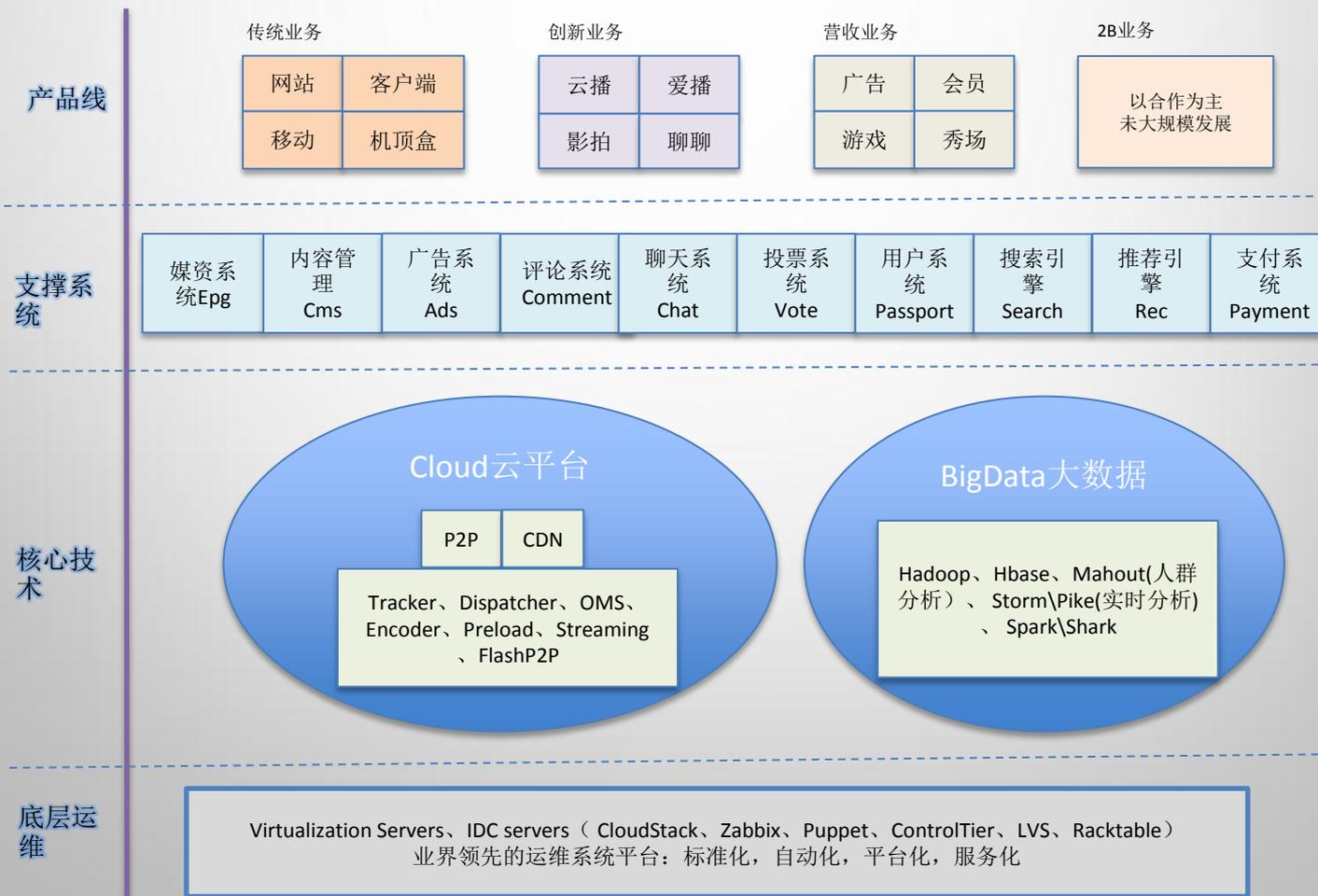


# 自我介绍—陈文春

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	



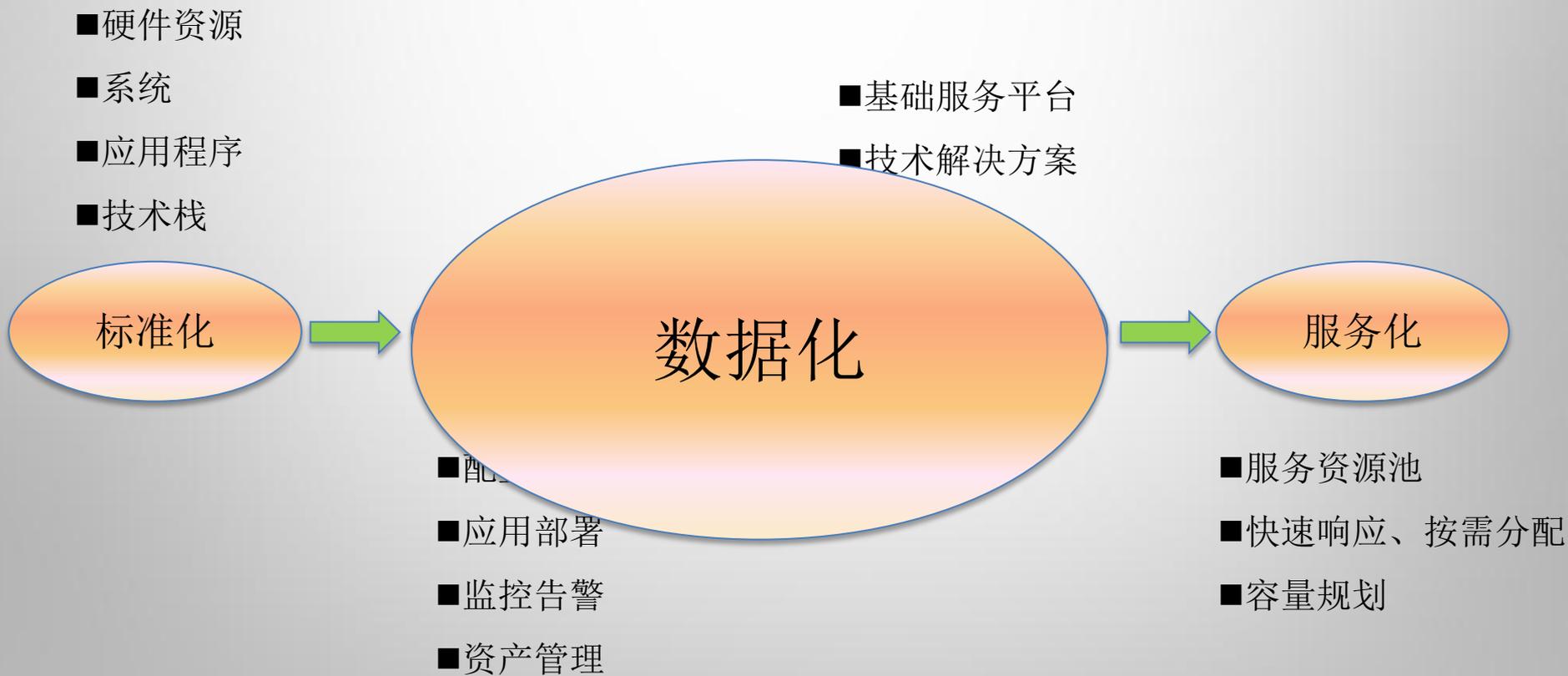
?TB 视频CDN带宽	>100GB 网页CDN带宽	~ 10k 服务器	> 60次/周 自动化上线
> 200个 CDN节点	~10个 核心IDC节点	> 400台 网络设备	> 10K 计算云CPU core
>10GB 核心机房带宽	150万个 监控点	40万个 告警点	> 4k 计算云节点
>10PB 分布式存储服务	> 10TB 分布式缓存服务	421个 核心IDC服务池	> 20亿次/天 单个服务池访问量

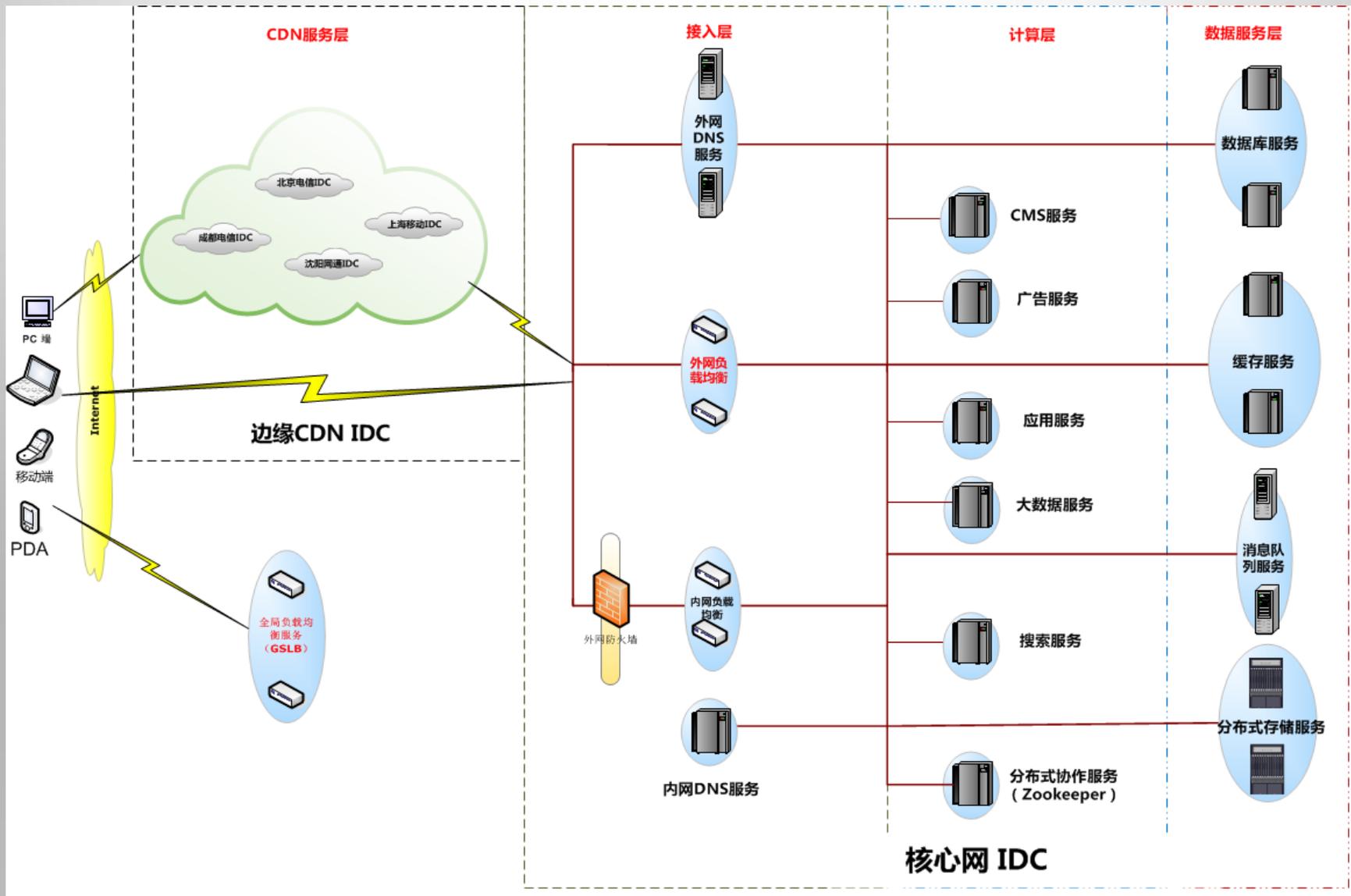


大中型互联网系统的特点：

- 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

- 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

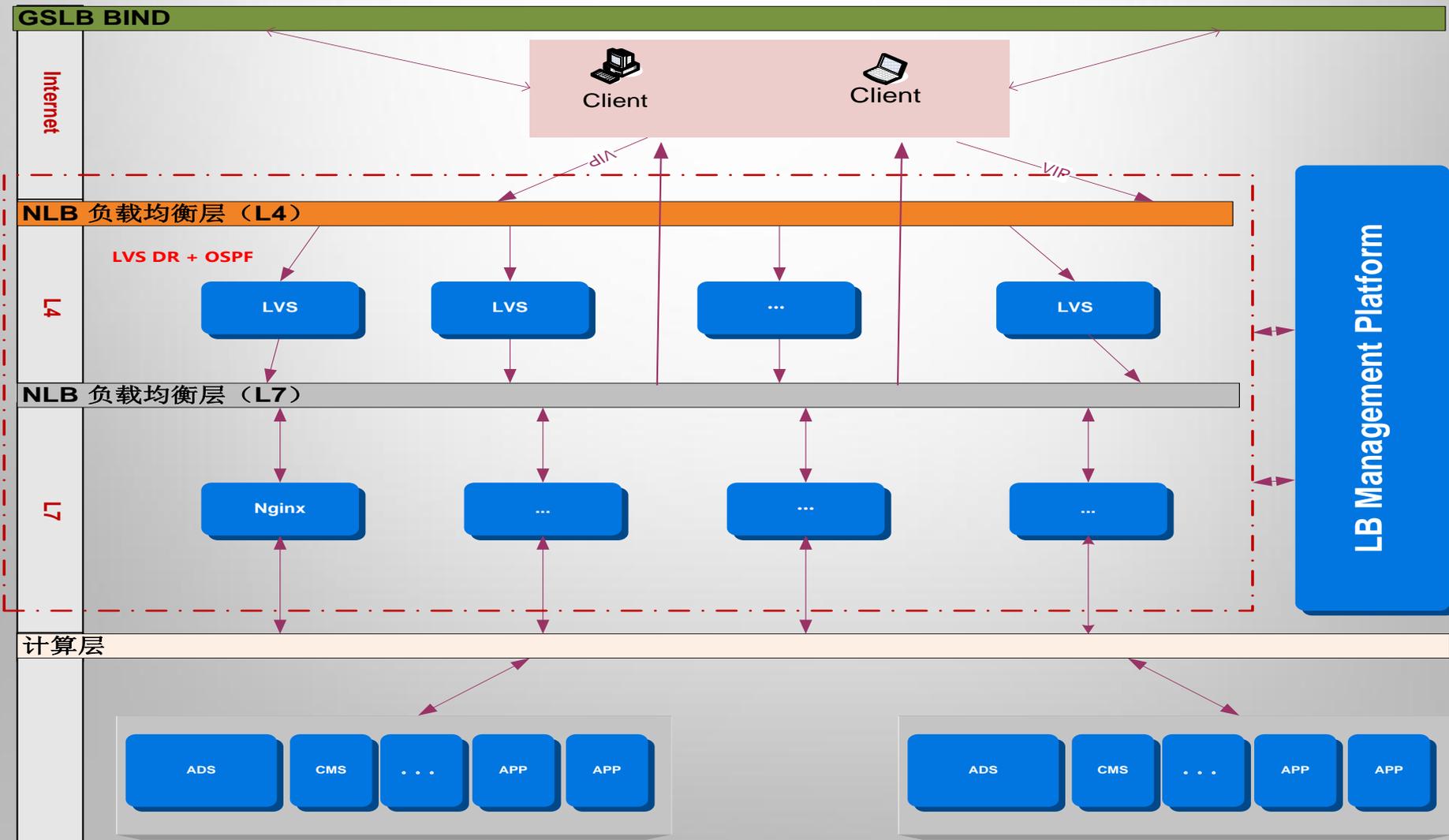


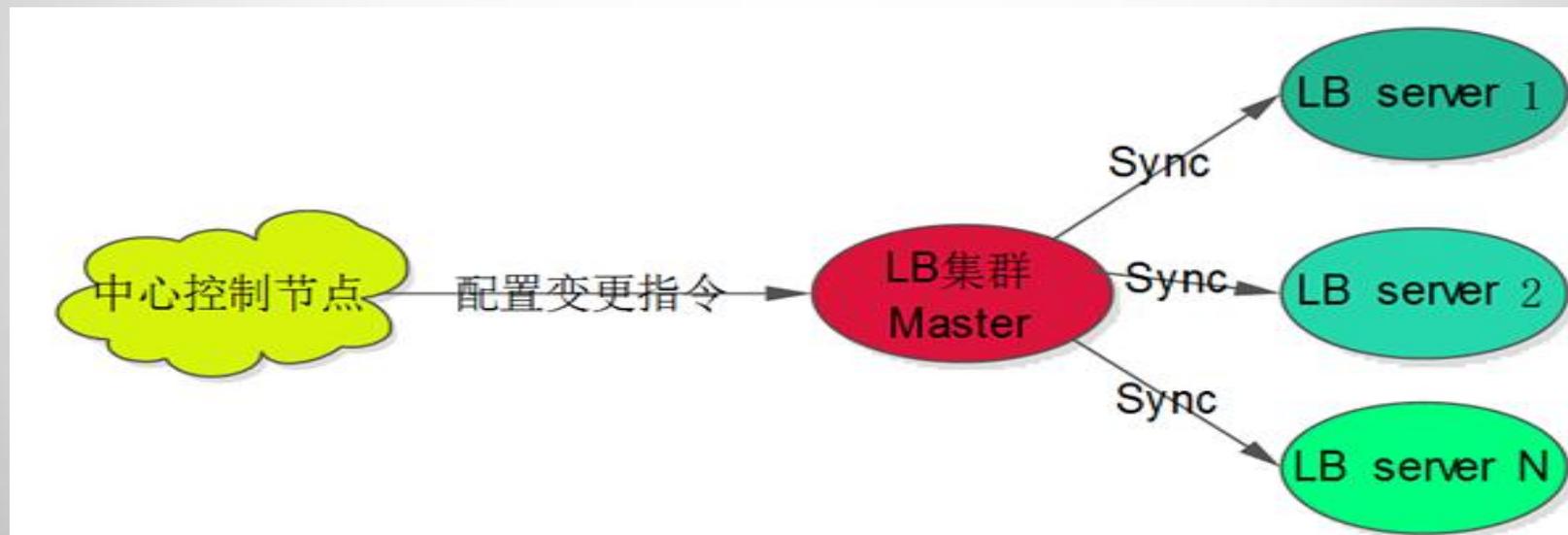


## ➤ IaaS&PaaS

- 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 )







- 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



- 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

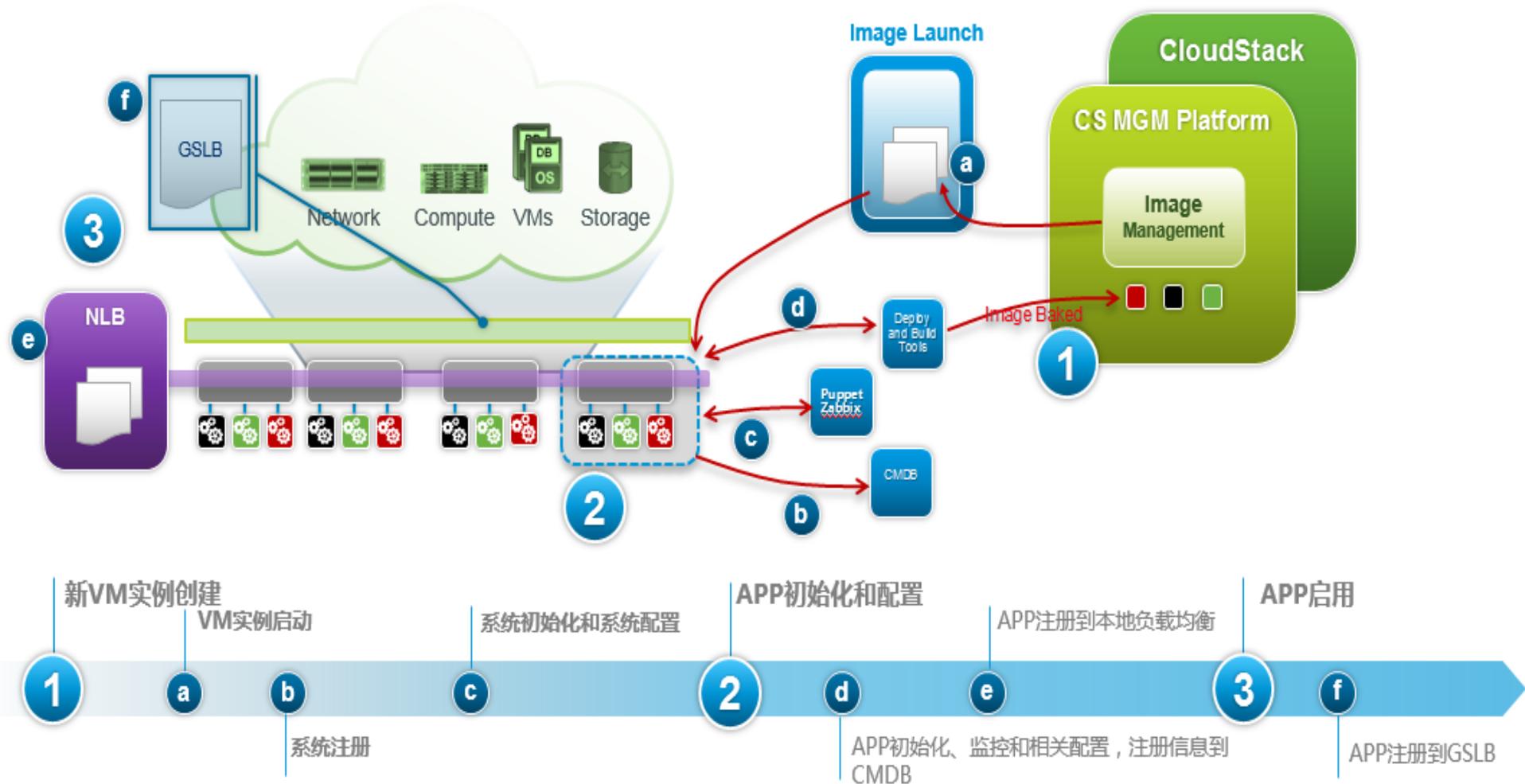
## 基于CloudStack的部署管道演进

Provisioned when needed

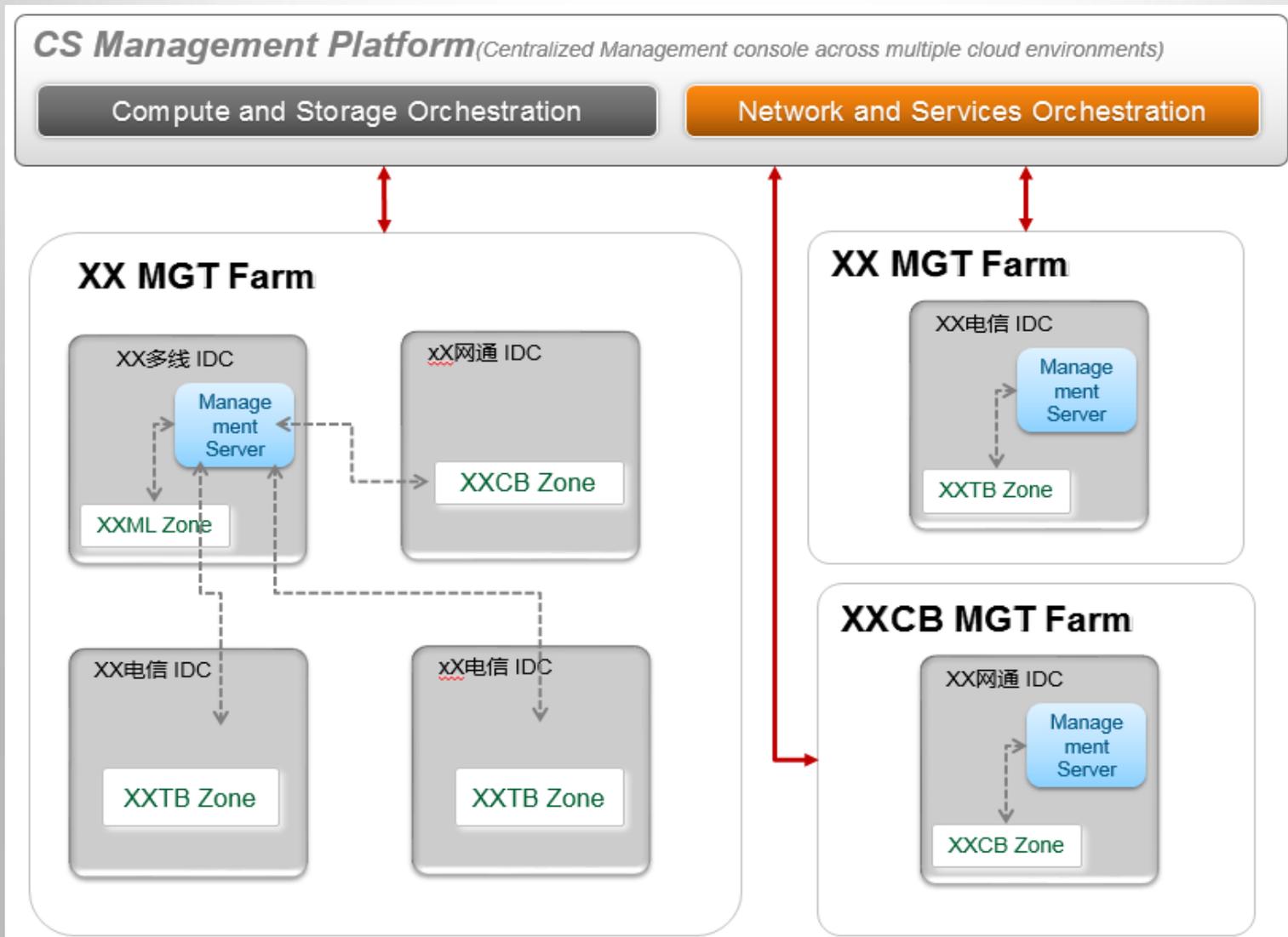


- Out of the box
- Parallel building
- Self Service
- One-button for All
- Elastic

## 基于CloudStack的部署管道演进



## 基于CloudStack的部署架构在PPTV



## 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



- 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

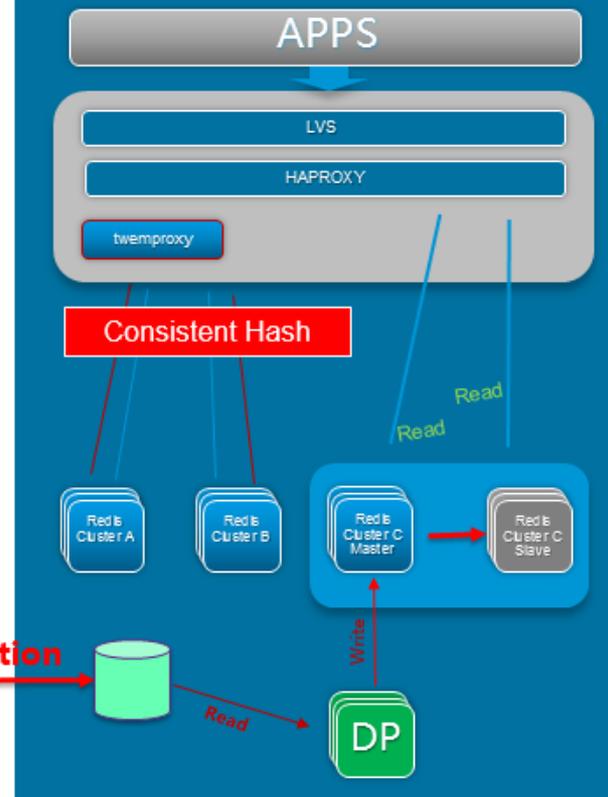
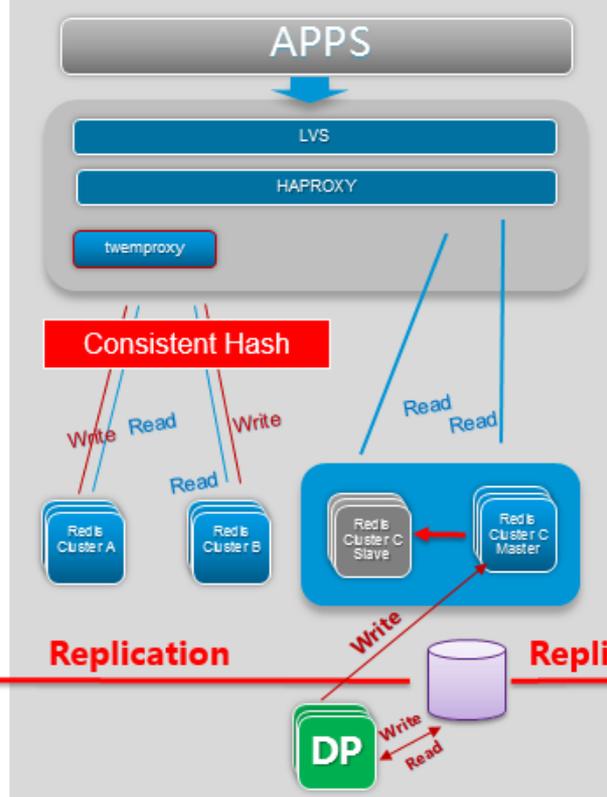
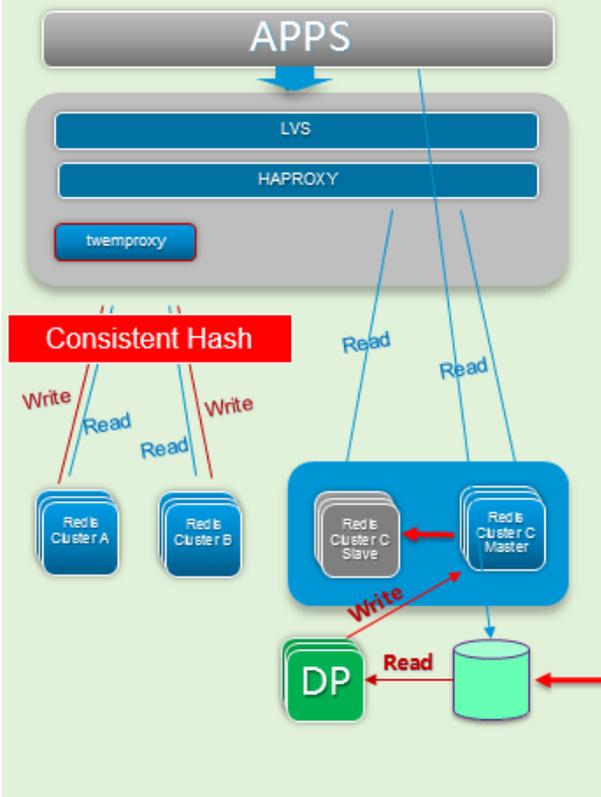
GSLB/NLB



电信 IDC

BGP/多线 IDC

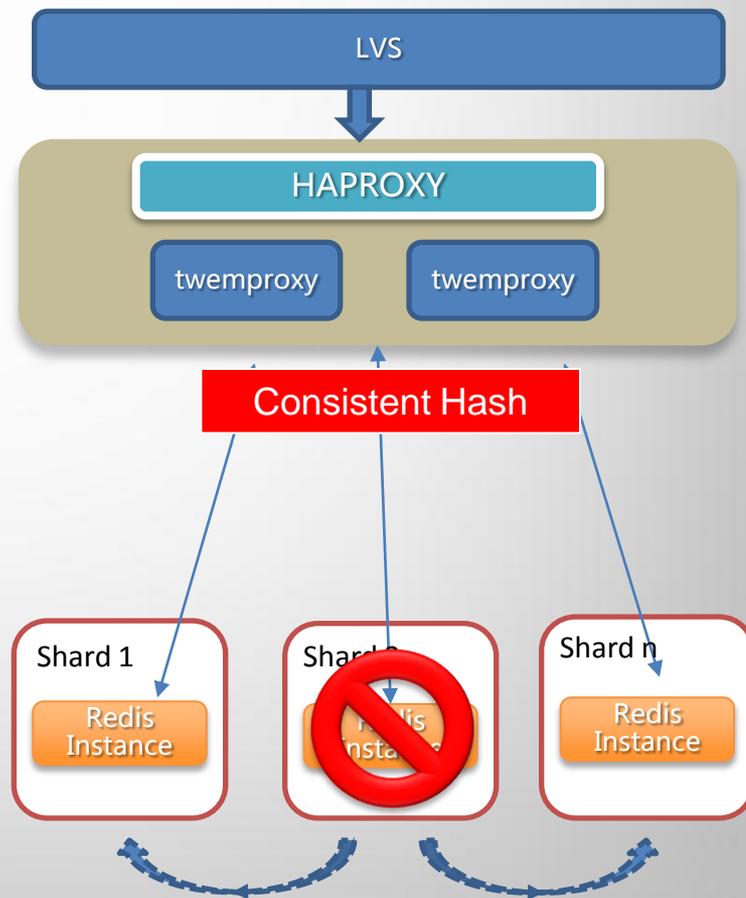
联通 IDC



## 缓存服务-HASH模式

### ➤ HASH模式

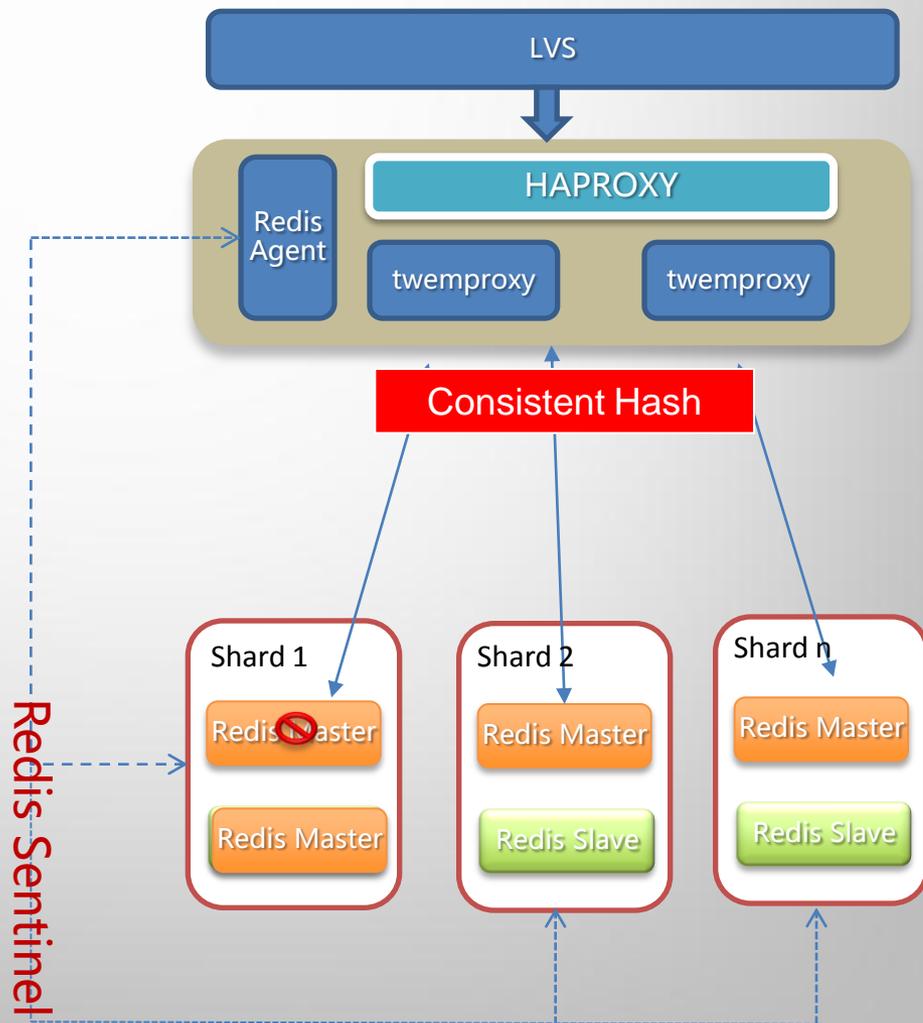
- 提供缓存服务，非持久化存储
- 集群提供自动数据分片
- 单个Redis实例宕机只影响部分数据
- 动态扩容



## 缓存服务—HASH+HA模式

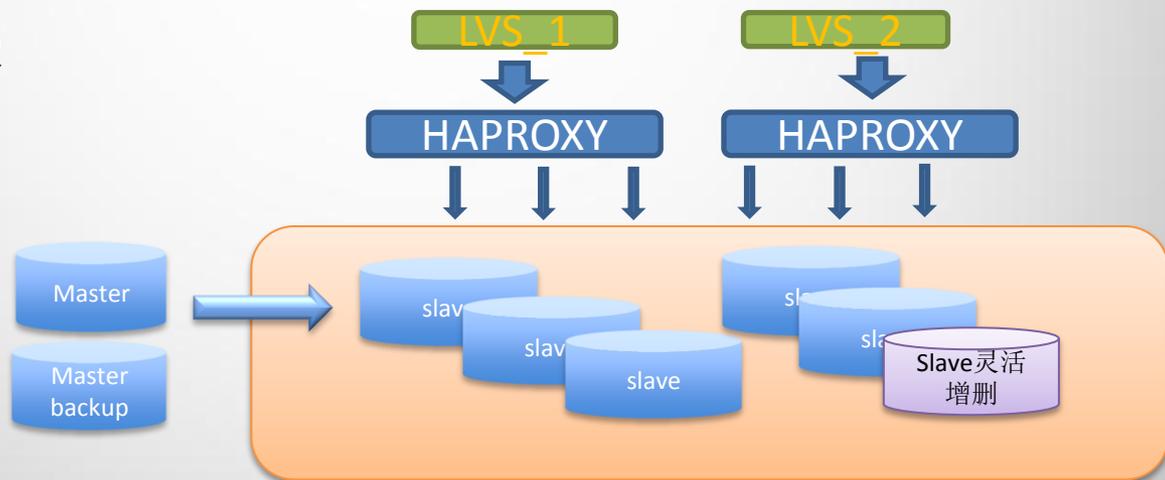
### ➤ HASH+HA

- 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

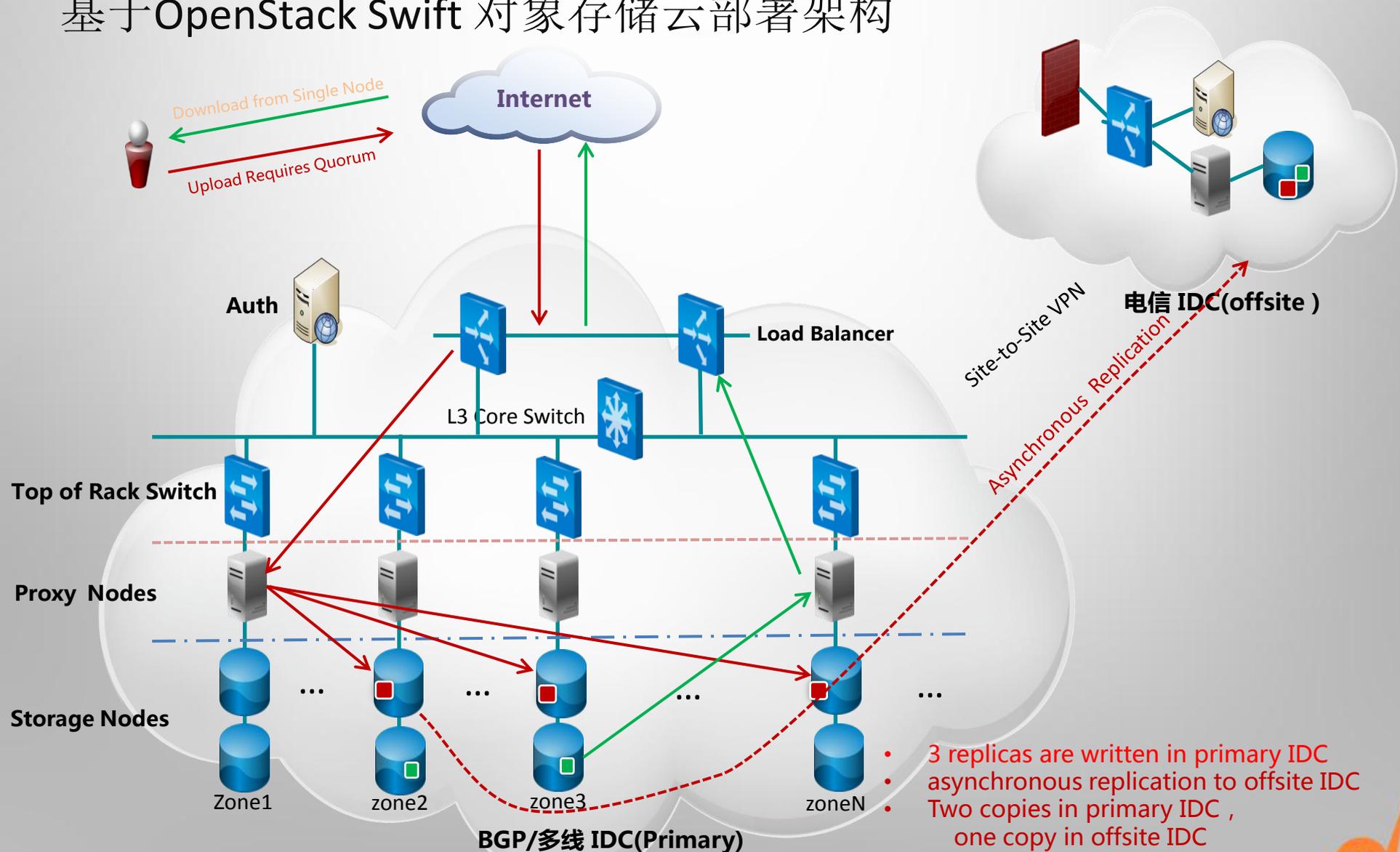


## 读写分离的缓存服务

- 多机高可用
- 故障节点自动下线
- 权重控制
- 在线水平扩容



## 基于OpenStack Swift 对象存储云部署架构

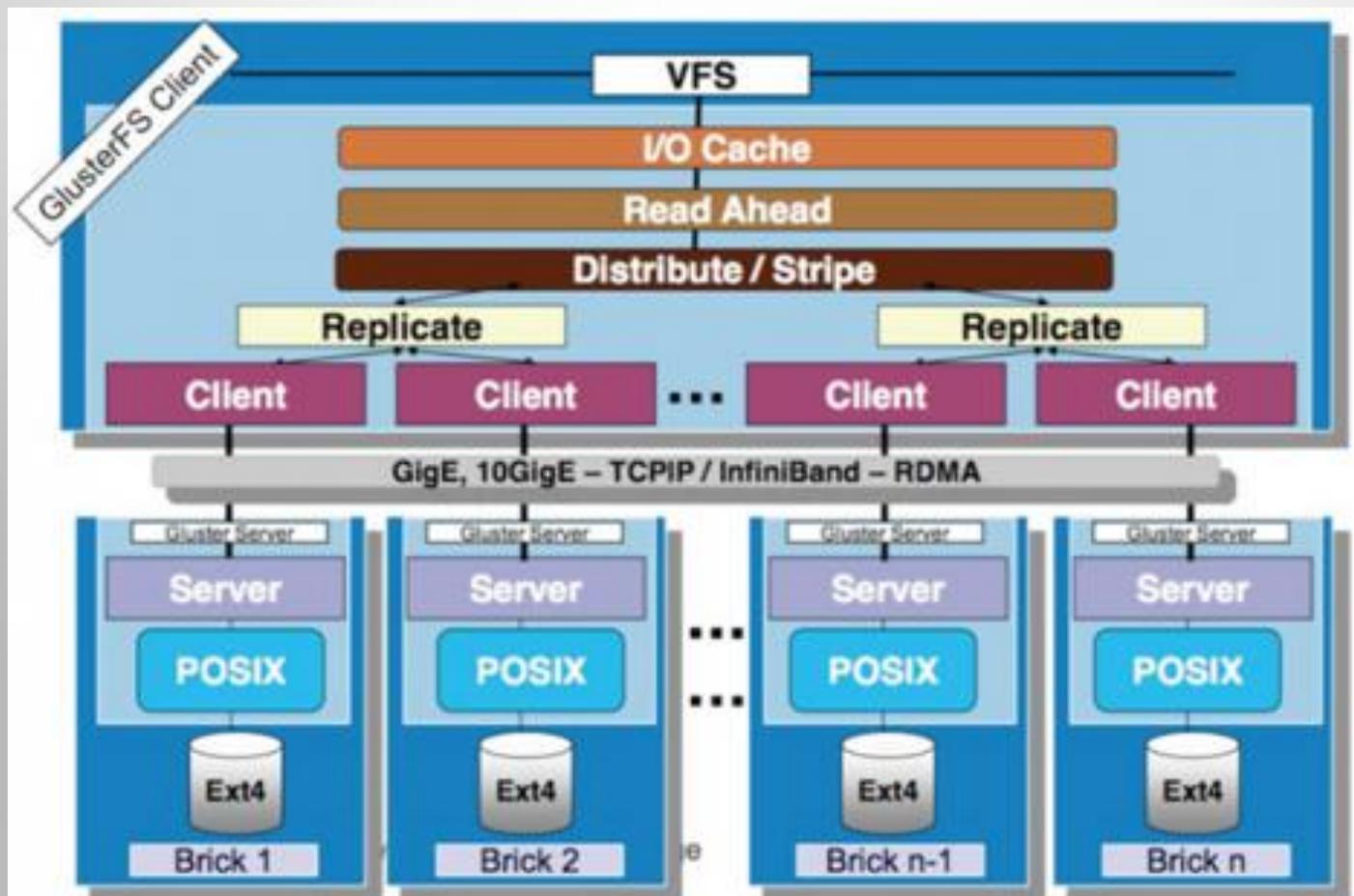


## OpenStack Swift 部署最佳实践

- 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 )

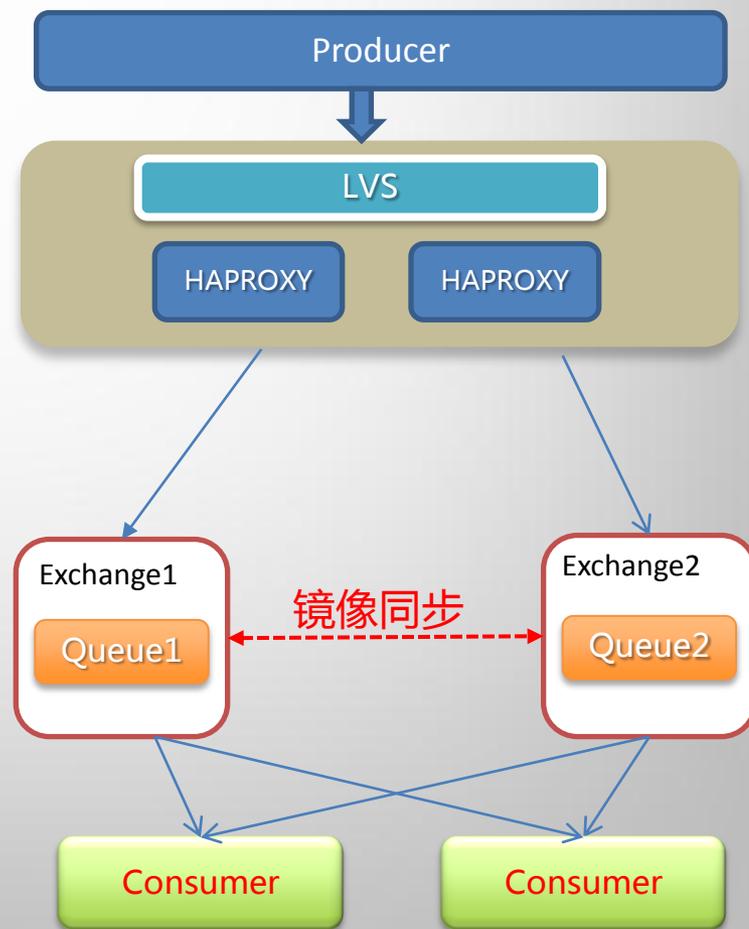
## GlusterFS分布式文件系统

- 多点挂载写入
- 图片等静态资源的CDN源站



## 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



## PPKeeper(Zookeeper As A Service)

### ➤ PPKeeper

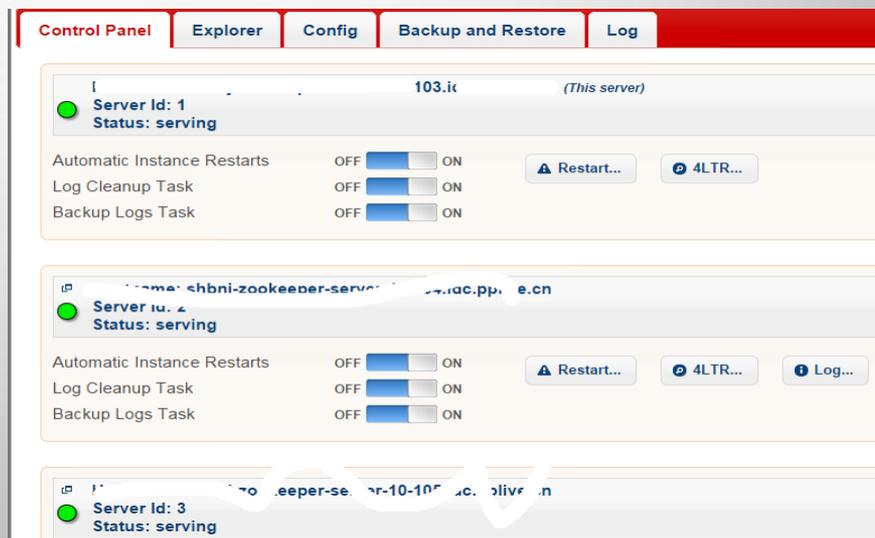
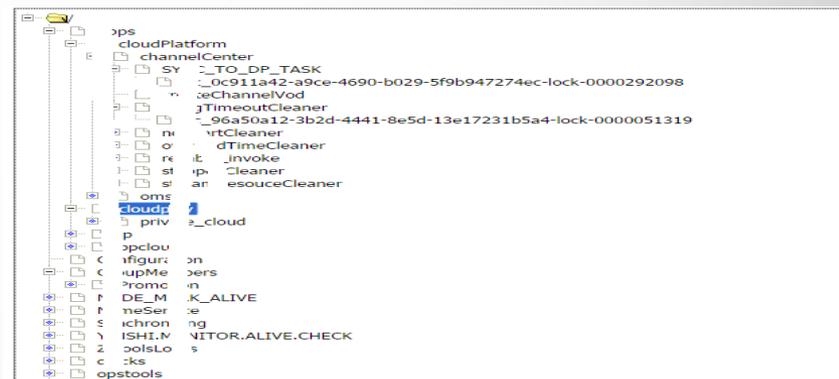
- Zookeeper3.4.5
- Java Client: Curator
- 服务端管理: Exhibitor
- Znode可视化管理
- 在线扩容、备份和恢复

### ➤ 分布式锁

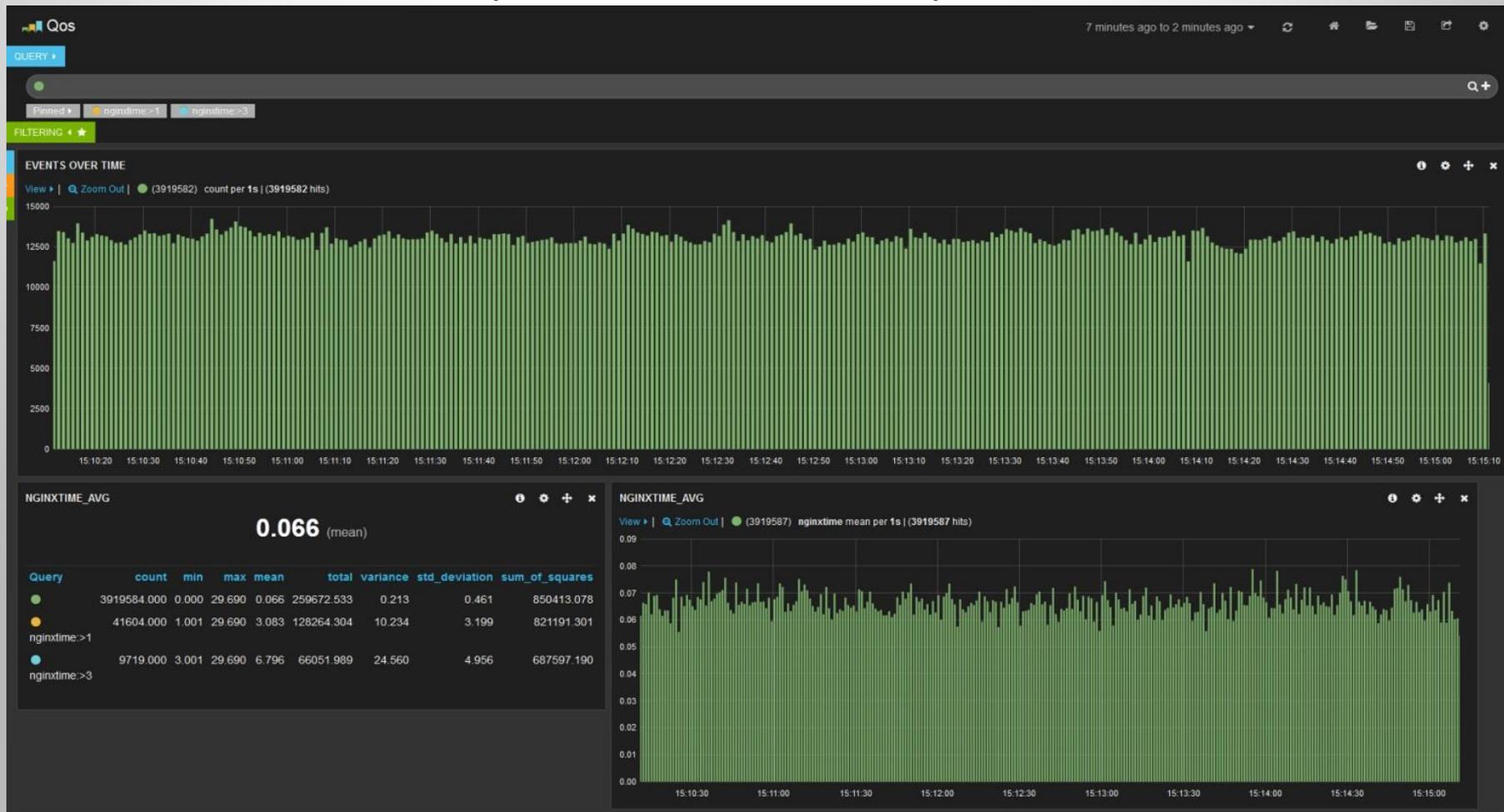
### ➤ 命名服务

### ➤ 集群管理

### ➤ 数据发布与订阅



## 分布式日志收集管道(Events As A Service )



## 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

- 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

## IDC机房管理

- 机柜位置
- 服务器信息
- 网络连接拓扑



Main page : Rackspace    Object

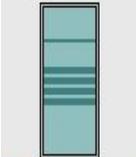
[Browse](#)   [Manage locations](#)   [Manage rows](#)

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

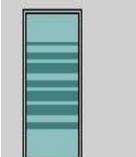
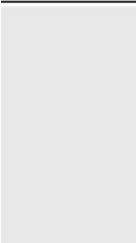
Local name	Visible label	Interface	L2 address	Remote object and port	Cable ID
kvm		KVM (host)			
eth0		1000Base-T	00:BD:81:D9:45:54		
eth1		1000Base-T	00:BD:81:D9:45:55	WA00638H1205 GigabitEthernet1/0/21	

## summary

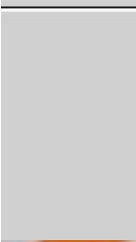
**名称:** JA03125L1109  
**类型:** Server  
**资产编号:** JA03125L1109  
**CPU品牌:** Intel  
**CPU信息:** 2.40G\*2  
**远程管理卡IP:** 10.208.254.138  
**购买日期:** 08/31/2011  
**内存大小:** 32G(8G\*4)  
**品牌及型号:** LenovoR520 G7  
**设备尺寸:** 2U  
**设备归属:** 有产权有使用权  
**设备序列号:** empty  
**设备种类:** 机架服  
**所属机房:** 上海多线  
**使用期限:** 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



11号楼3F-A08



shbnj\_3K09



**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

The screenshot shows the PPTV MaaS platform interface. The top navigation bar includes 'PPTV-自动化装机平台', user 'hunterhe', '平台管理', and 'Log Out'. The main content area is titled '服务器装机 上海多线'. Below the title, there are navigation tabs: '服务器-重装' (selected), '服务器-新装', '服务器-重启', '系统版本管理', and '历史记录'. The '服务器重装' section has buttons for '开启自动刷新' and '关闭自动刷新'. A table displays server installation records with columns: '资产编号', 'MAC地址', '远程管理卡IP', 'Cobbler IP', and 'Status'. The table shows 3 entries, all with '待装机' status. A '下一步' button is visible at the bottom.

资产编号	MAC地址	远程管理卡IP	Cobbler IP	Status
JA03667D1106	eth0@d4:ae:52:63:95:cb	10.208.253.4	10.208.101.42	待装机
JA03213L1109	eth0@00:E0:81:D9:4B:A2	10.208.255.178	10.208.101.42	待装机
JA03079H1109	eth0@54:89:98:02:5f:4f	10.208.254.102	10.208.101.42	待装机

This screenshot shows the same '服务器重装' page, but the 'Status' column now displays progress bars for each server, indicating they are 6% installed. A red box highlights the 'Status' column. At the bottom, a log message is visible: 'username:hunterhe==>zcbh:JA03079H1109==>eth:eth1==>IP:10.208.10.102==>time:2014-02-27 11:52:38==>conn:生成重装网络配置文件成功,现在开始在cobbler上添加system。'.

资产编号	MAC地址	远程管理卡IP	Cobbler IP	Status
JA03667D1106	eth0@d4:ae:52:63:95:cb	10.208.253.4	10.208.101.42	6%
JA03213L1109	eth0@00:E0:81:D9:4B:A2	10.208.255.178	10.208.101.42	6%
JA03079H1109	eth0@54:89:98:02:5f:4f	10.208.254.102	10.208.101.42	6%

```
username:hunterhe==>zcbh:JA03079H1109==>eth:eth1==>IP:10.208.10.102==>time:2014-02-27 11:52:38==>conn:生成重装网络配置文件成功,现在开始在cobbler上添加system。。
```



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



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-sqlpg	dba	在保

磁盘信息

[返回](#)

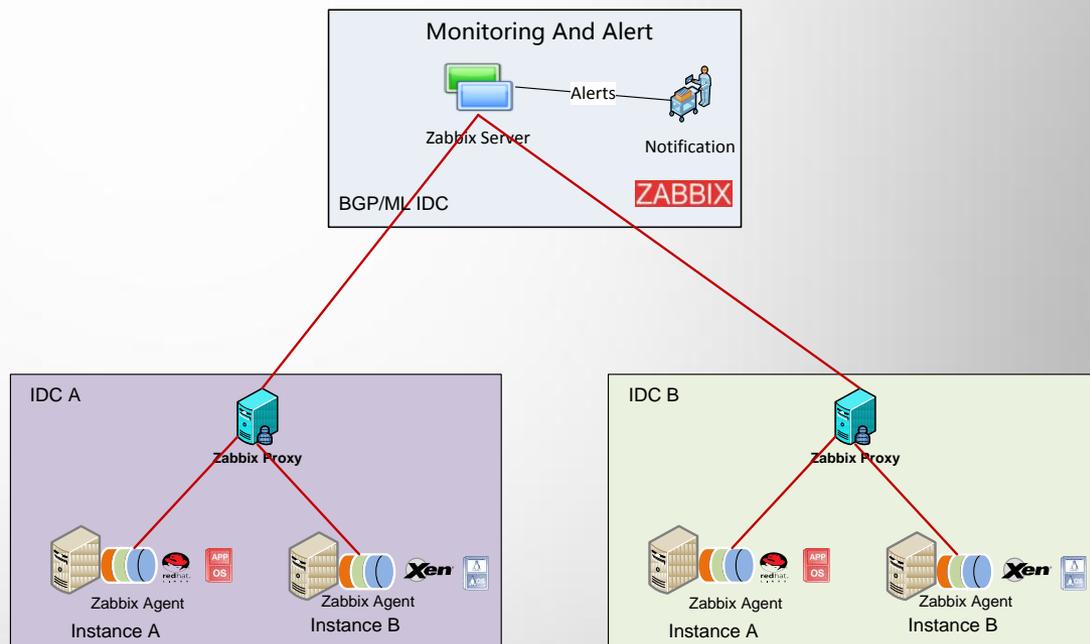
磁盘											
每页显示 20 条记录											快速搜索
Raid Id	Raid Level	盘符	槽位号	SN	接口类型	容量(G)	健康状态	使用状态	磁盘灯	记录	
0	10	sda	0	D004PB1077YP	SAS	299	ok	在线	不闪	记录	
0	10	sda	1	6XP1FRDQ	SAS	299	ok	在线	不闪	记录	
0	10	sda	2	WX61EC1WRD82	SAS	299	ok	在线	不闪	记录	
0	10	sda	3	D004PB107698	SAS	299	ok	在线	不闪	记录	
0	10	sda	4	6XP1G37Y	SAS	299	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



- 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

PPTV 聚力

Event Console

查询:

警报级别过滤选择: P4 P3 P2 P1 P0

警报状态过滤选择: New Acked Snoozec fixing fix\_succ fix\_failed Closed 按过滤条件获取数据

批量处理操作: 请选择批量动作 执行批量操作

第 1 页 共 1 页 显示 1 - 16条, 共 16条

服务器	警报名	级别	次数	状态	报警创建时间	最后一次报警时间	负责	功能按钮
<input type="checkbox"/> SYT-WEB-CDN-192-167.idc.ppliv...	syslog error on SYT-WEB-CDN-192-167.idc.pplive.cn	P3	1	new	2014-04-22 18:38:42	2014-04-22 18:38:42		
<input type="checkbox"/> JNT-WEB-CDN-50-148.idc.pplive...	syslog error on JNT-WEB-CDN-50-148.idc.pplive.cn	P3	1	new	2014-04-22 18:38:38	2014-04-22 18:38:38		
<input type="checkbox"/> SYT-WEB-CDN-192-166.idc.ppliv...	syslog error on SYT-WEB-CDN-192-166.idc.pplive.cn	P3	1	new	2014-04-22 18:37:42	2014-04-22 18:37:42		
<input type="checkbox"/> HZC-WEB-CDN-184-169.idc.ppliv...	syslog error on HZC-WEB-CDN-184-169.idc.pplive.cn	P3	1	ack	2014-04-22 18:29:59	2014-04-22 18:31:45	iyuanhe@pptv.c...	
<input type="checkbox"/> HZC-WEB-CDN-184-168.idc.ppliv...	syslog error on HZC-WEB-CDN-184-168.idc.pplive.cn	P3	1	ack	2014-04-22 18:29:15	2014-04-22 18:31:45	iyuanhe@pptv.c...	
<input type="checkbox"/> SGT-WEB-CDN-44-62.idc.pplive.cn	syslog error on SGT-WEB-CDN-44-62.idc.pplive.cn	P3	1	ack	2014-04-22 18:27:11	2014-04-22 18:31:45	iyuanhe@pptv.c...	
<input type="checkbox"/> SGT-WEB-CDN-44-63.idc.pplive.cn	syslog error on SGT-WEB-CDN-44-63.idc.pplive.cn	P3	1	ack	2014-04-22 18:24:52	2014-04-22 18:31:45	iyuanhe@pptv.c...	
<input type="checkbox"/> CST-media-155-228.windows	[UNREACHABLE_hezuo]CST-media-155-228.windows ...	P1	1	ack	2014-04-22 18:24:31	2014-04-22 18:31:45	iyuanhe@pptv.c...	
<input type="checkbox"/> SGT-WEB-CDN-44-64.idc.pplive.cn	syslog error on SGT-WEB-CDN-44-64.idc.pplive.cn	P3	1	ack	2014-04-22 18:23:30	2014-04-22 18:31:45	iyuanhe@pptv.c...	
<input type="checkbox"/> SHBNJ-WEB_QIPAO-PHP-169-1...	CPU load is too high on SHBNJ-WEB_QIPAO-PHP-169-...	P2	1	ack	2014-04-22 18:17:04	2014-04-22 18:31:45	iyuanhe@pptv.c...	
<input type="checkbox"/> SGT-WEB-CDN-44-61.idc.pplive.cn	syslog error on SGT-WEB-CDN-44-61.idc.pplive.cn	P3	1	ack	2014-04-22 18:14:25	2014-04-22 18:17:50	iyuanhe@pptv.c...	
<input type="checkbox"/> GZC-switch-37.65-10G	[XGigabitEthernet0/1/1] outgoing traffic dropped on ...	P2	2	ack	2014-04-22 18:07:03	2014-04-22 18:27:23	iyuanhe@pptv.c...	
<input type="checkbox"/> SHBNJ-WEB_NewView-python-...	[UNREACHABLE_feihezuo]SHBNJ-WEB_NewView-pyth...	P1	1	ack	2014-04-22 18:00:00	2014-04-22 18:00:42	iyuanhe@pptv.c...	
<input type="checkbox"/> htiyo-h-tomcat-169-Q3 nsrm-filt-	[UNREACHABLE_feihezuo]htiyo-h-tomcat-169-Q3 nsr...	P1	1	ack	2014-04-22 17:59:31	2014-04-22 18:00:30	iyuanhe@pptv.c...	

autofix\_failed邮件

IDC网络质量 = 【外部质量+内部质量】 = 互联网业务的**基石**

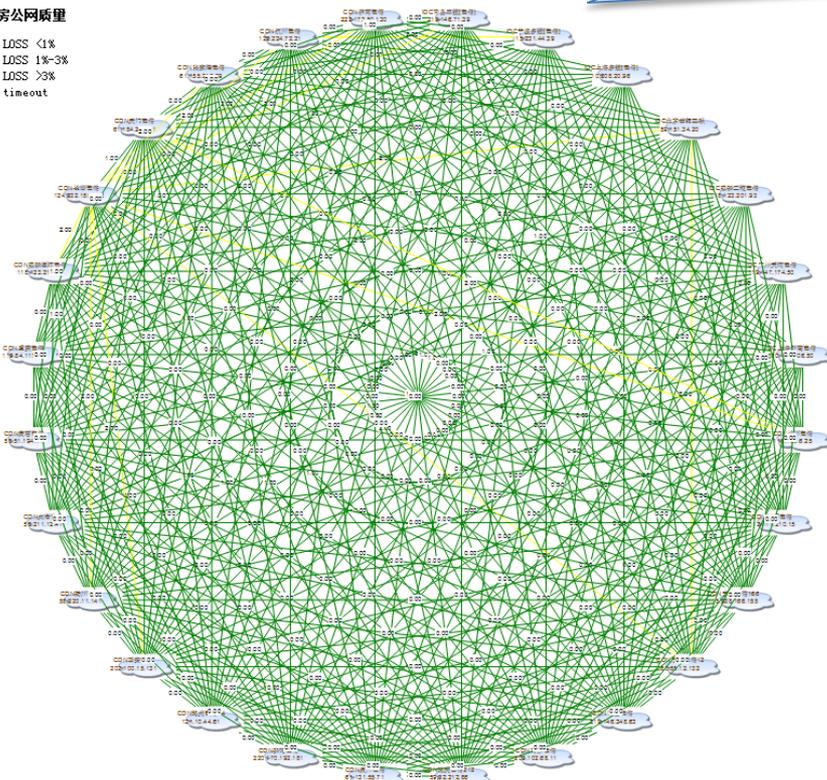
**直观**: 颜色反映质量, 一目了然  
**全覆盖**: 核心网+CDN+内网VPN  
**自动报警**: 邮件+短信  
**历史查询**: 任意时间追溯

Nightwatch-外部网络质量监控系统

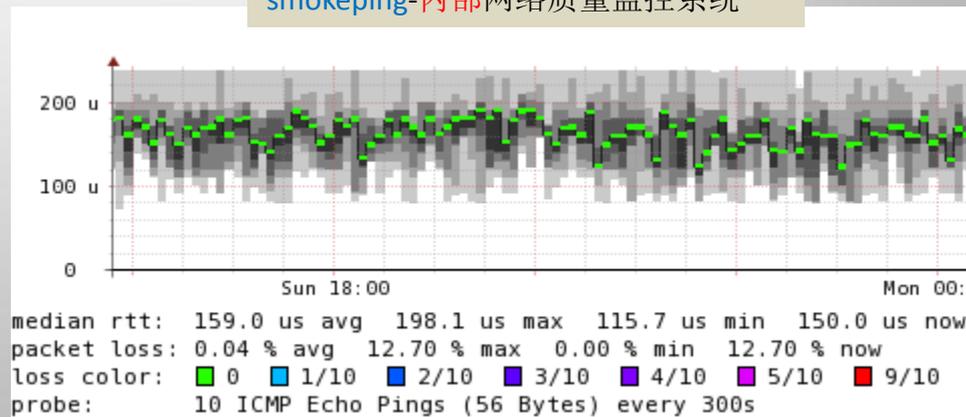
Last update: 2014-11-18 1

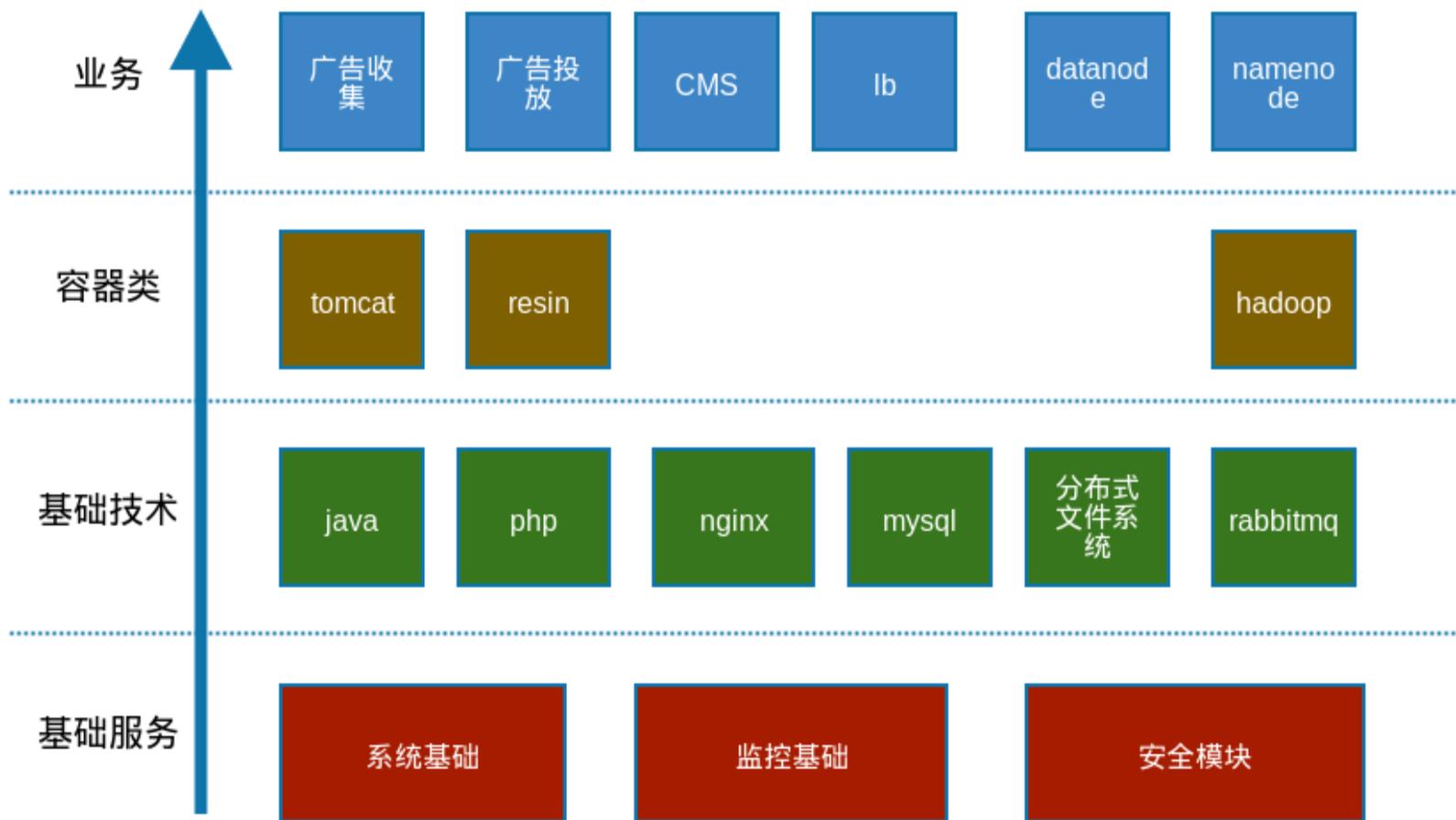
电信机房公网质量

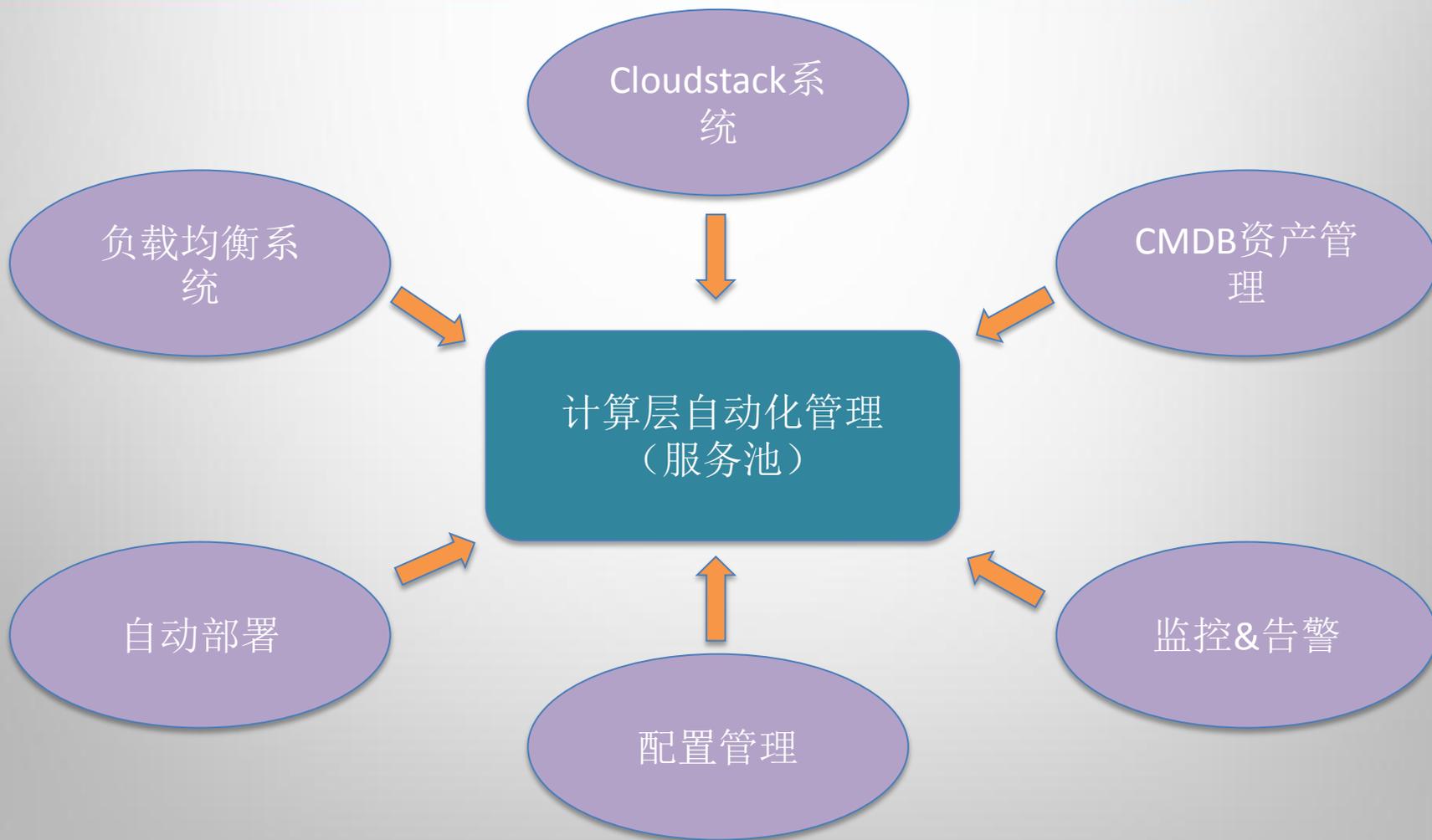
Green: LOSS <1%  
Yellow: LOSS 1%-3%  
Red: LOSS >3%  
Grey: timeout



smokeping-内部网络质量监控系统







# MaaS—计算层自动化管理



The screenshot displays the PPTV Ops Platform V3.0 interface. The main content area shows a '服务器列表' (Server List) table with columns for host name, IP, monitoring, maintenance, deployment, online status, traffic, version, link, and progress. A dropdown menu is open over the table, listing various actions like '添加监控' (Add Monitoring) and '回收服务器' (Reclaim Server). The interface includes a sidebar with navigation items and a top navigation bar with user information and platform management options.

主机名	主机IP	监控	维护	发布	在线	流量	版本	链接	进程
shnj-b-tomcat-189-56.ad-delv-tf.lin.idc.pplive.cn	10.208.189.56	●	●	●	●	10/5	20141125.1	zabbix / 负载 / 信息页	开发中
shnj-b-tomcat-189-54.ad-delv-tf.lin.idc.pplive.cn	10.208.189.54	●	●	●	●	10/5	20141125.1	zabbix / 负载 / 信息页	开发中
shnj-b-tomcat-189-51.ad-delv-tf.lin.idc.pplive.cn	10.208.189.51	●	●	●	●	10/5	20141125.1	zabbix / 负载 / 信息页	开发中
shnj-b-tomcat-189-50.ad-delv-tf.lin.idc.pplive.cn	10.208.189.50	●	●	●	●	10/5	20141125.1	zabbix / 负载 / 信息页	开发中
shnj-b-tomcat-189-49.ad-delv-tf.lin.idc.pplive.cn	10.208.189.49	●	●	●	●	10/5	20141125.1	zabbix / 负载 / 信息页	开发中
shnj-b-tomcat-189-37.ad-delv-tf.lin.idc.pplive.cn	10.208.189.37	●	●	●	●	20/10	20141125.1	zabbix / 负载 / 信息页	开发中
shnj-b-tomcat-189-250.ad-delv-tf.lin.idc.pplive.cn	10.208.189.250	●	●	●	●	10/5	20141125.1	zabbix / 负载 / 信息页	开发中
shnj-b-tomcat-189-249.ad-delv-tf.lin.idc.pplive.cn	10.208.189.249	●	●	●	●	10/5	20141125.1	zabbix / 负载 / 信息页	开发中
shnj-b-tomcat-189-242.ad-delv-tf.lin.idc.pplive.cn	10.208.189.242	●	●	●	●	10/5	20141125.1	zabbix / 负载 / 信息页	开发中
shnj-b-tomcat-189-239.ad-delv-tf.lin.idc.pplive.cn	10.208.189.239	●	●	●	●	10/5	20141125.1	zabbix / 负载 / 信息页	开发中

操作菜单:

- 请选择操作
- 添加监控
- 开启告警
- 加入发布
- 删除发布
- 添加LB流量
- 移除LB流量
- 上线服务器
- 下线服务器
- 回收服务器
- 添加canary
- 取消canary
- 进程重启
- 请选择操作

## 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

PPTV-自动发布管理平台

控制台中台 > platform.services.passport > platform.services.passport.service > platform.services.passport.service.api

任务列表

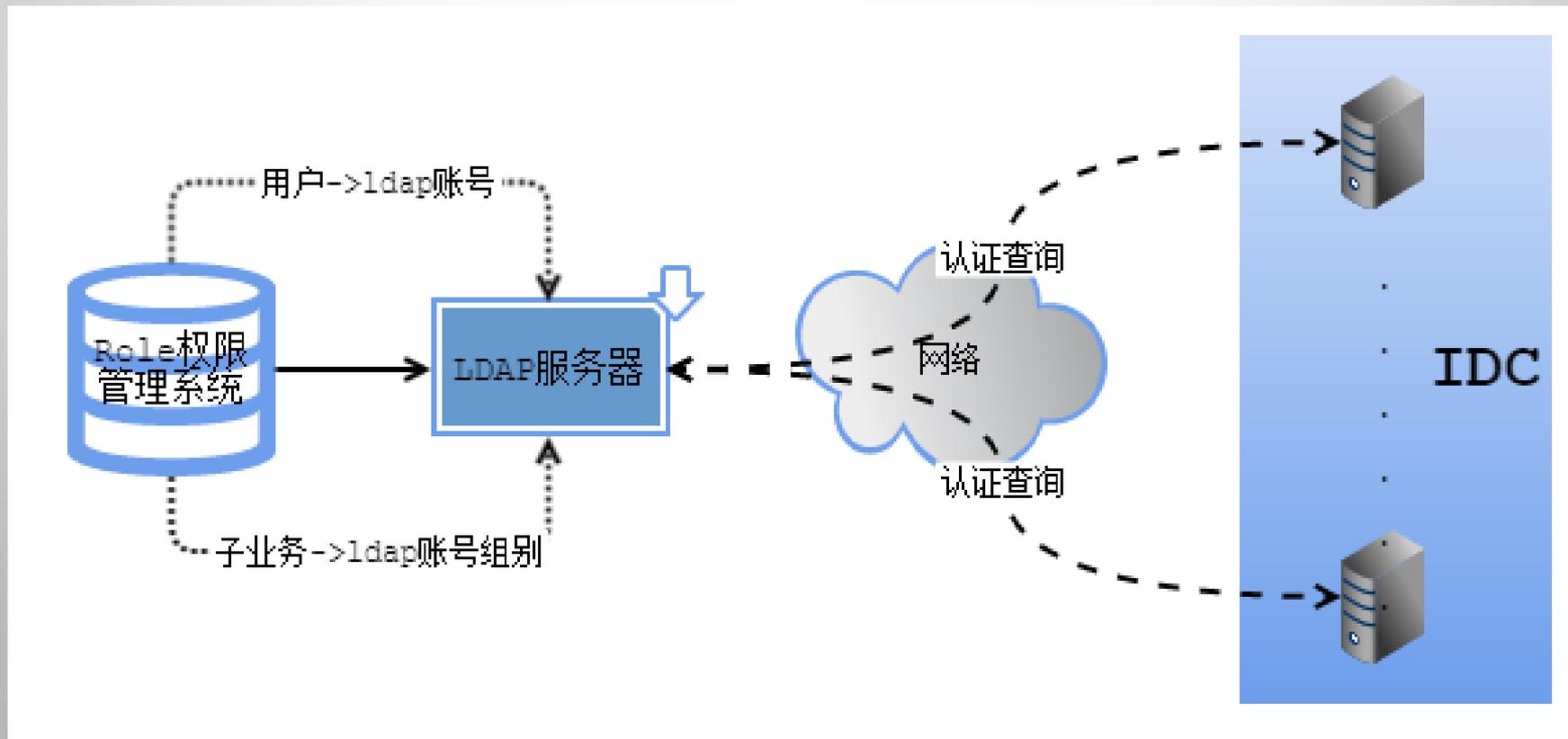
选择时间: 2014-11-03 至 2014-11-17 查询 当前版本号: 20141113.1 可回滚版本号: 20141112.1 目录

任务列表

每页显示 10 条记录 快速搜索

ID	业务	类型	创建人	创建时间	版本	审批状态	部署人	最后操作时间	部署成功率	部署状态	操作
1280	platform.services.passport.service.api	resin	shuaiju	2014年11月13日 14:46:30 星期四	20141113.1	审批通过	shuaiju	2014年11月13日 15:26:16 星期四	40%灰度-100% 70%灰度-0% 10%灰度-100% 100%灰度-0%	部署成功	🔍 ✎
1277	platform.services.passport.service.api	resin	joanliu	2014年11月13日 13:49:24 星期四	20141113.1	审批通过	joanliu	2014年11月13日 14:04:47 星期四	40%灰度-0% 70%灰度-0% 10%灰度-0% 100%灰度-0%	已瘫痪	🔍 ✎
1275	platform.services.passport.service.api	resin	qinggao	2014年11月13日 12:55:49 星期四	20141113.1	审批通过	qinggao	2014年11月13日 12:59:33 星期四	40%灰度-0% 70%灰度-0% 10%灰度-28.57% 100%灰度-0%	已瘫痪	🔍 ✎
1256	platform.services.passport.service.api	resin	shuaiju	2014年11月12日 15:30:11 星期三	20141112.1	审批通过	shuaiju	2014年11月12日 15:41:31 星期三	40%灰度-100% 70%灰度-100% 10%灰度-100% 100%灰度-100%	部署成功	🔍 ✎
1255	platform.services.passport.service.api	resin	shuaiju	2014年11月12日 15:25:54 星期三	20141112.1	审批通过	shuaiju	2014年11月12日 15:28:08 星期三	40%灰度-0% 70%灰度-0% 10%灰度-37.5% 100%灰度-0%	已瘫痪	🔍 ✎
1254	platform.services.passport.service.api	resin	joanliu	2014年11月12日 14:50:10 星期三	20141112.1	审批通过	joanliu	2014年11月12日 14:56:44 星期三	40%灰度-0% 70%灰度-0% 10%灰度-0% 100%灰度-0%	部署失败	🔍 🗑️ ✎





- 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