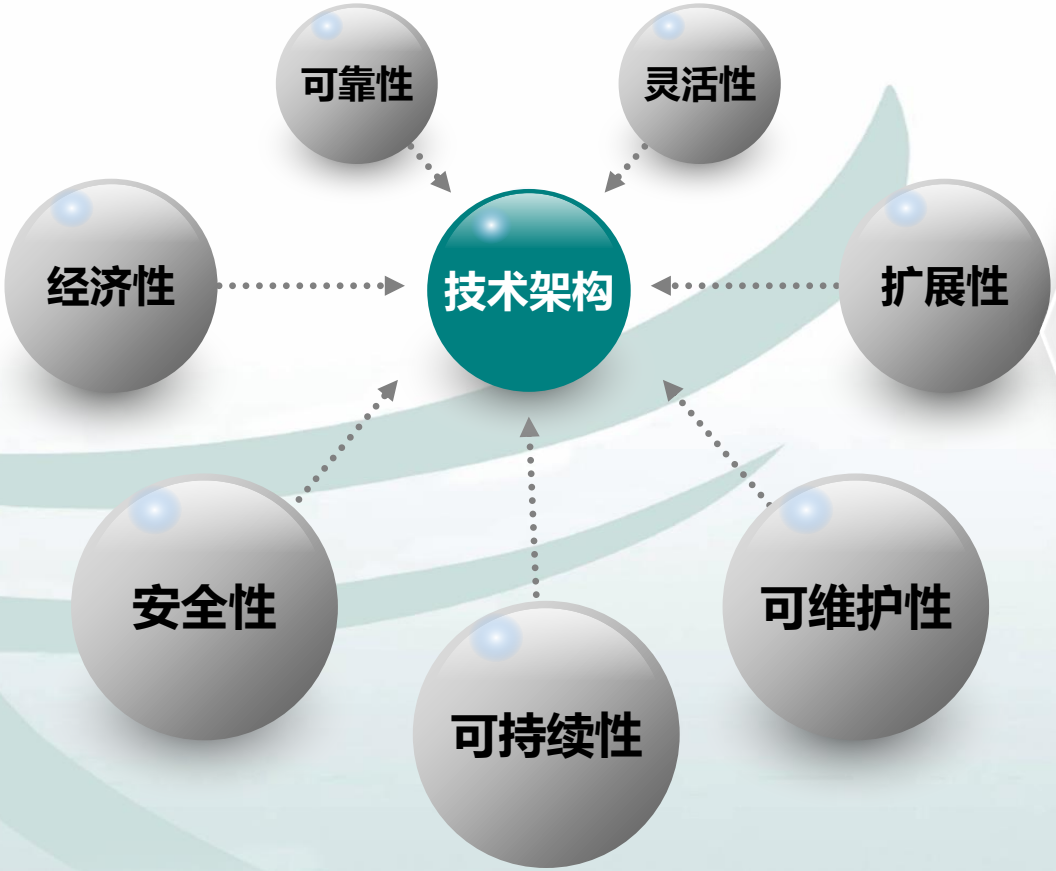
国家电网
STATE GRID

龙江电力云计算实践之路

2016年1月

目录

1、技术架构



基于国网SG186典型架构，龙江电力共运行2套云平台，一套以**中间件**为主，一套以**数据库**为主。其中中间件型云平台主要以高CPU计算，大内存为标志，数据库型云平台以高存储容量、高I/O处理为主。 中间件云平台采用Openstack框架进行建设。

什么是OpenStack？

OpenStack是一个开源的云计算管理平台项目，旨在为公共及私有云的建设与管理提供软件的开源项目。

OpenStack的首要任务是简化云的部署过程并为其带来良好的可扩展性，为企业业务系统快速部署、简化运维复杂度、降低运维成本、提高业务系统安全性、方便业务系统扩展提供了良好的基础平台。

- 国家电网SG186典型架构 Middleware+Database
- 互联网公司的架构解救不了传统企业的信息系统
- 接受短期问题，更看重远期路径可达性
- 不要用旧观念指导新世界

传统架构主导

基于现有的架构进行重构和改造，羁绊太多，进展慢

互联网架构主导

不能完全把**互联网**的东西搬过来

渐进式架构

新应用，新架构，渐进融合老架构

建设思路

为保护龙江公司原有的IT投资，在本项目实施过程中将大量采用利旧的方式来进行建设。过程如下：

- 项目前期只需要采购20台高性能X86架构PC服务器做为基础云平台环境；
- 迁移部分非核心业务系统至新的云平台环境上。将迁移后的X86服务腾退；
- 改造腾退后的X86服务器，使其内存、硬盘、网卡满足云化要求，将改造后服务器加入云平台中；
- 再次迁移部分业务系统至新的云平台环境上。将迁移后的X86服务腾退；
- 改造腾退后的X86服务器，使其内存、硬盘、网卡满足云化要求，将改造后服务器加入云平台中；
- 通过迁移、腾退、改造、加入云平台、再迁移的跌代过程，将龙江公司的所有业务系统全部迁移至云平台中。最终实现700个节点的私有云。

2、主要技术组件

KVM的开源性及兼容性，KVM是基于硬件的完全虚拟化而进行虚拟环境的构建，是当前主流的主流VMM；	CEPH可以提供企业级的对象存储，存储系统可轻松扩展到数 PB 容量，支持多种工作高负载的高性能，并提供高可靠性；	Neutron可提供云计算环境下的虚拟网络功能，同时支持多种物理网络类型，更多的网络设备，支持防火墙服务，节点间 VPN 服务支持和开源 SDN 网络的实现。

对于OpenStack的云平台建设，龙江公司在调研的基础上，采用如下建设方案，在计算节点方面采用KVM，分布式存储方面采用CEPH，在网络方面采用Neutron。

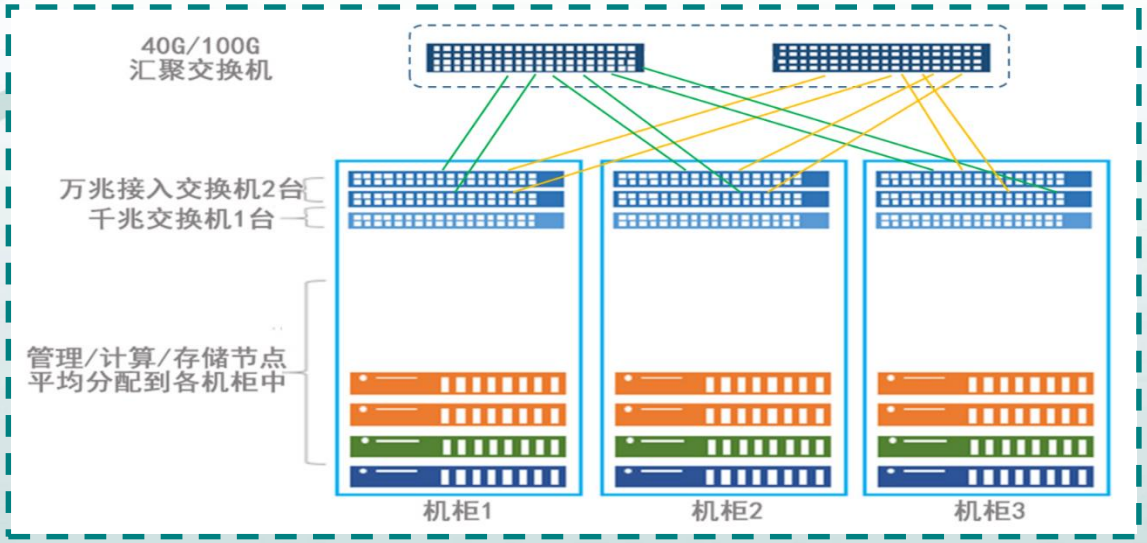
目录

1、物理设备情况

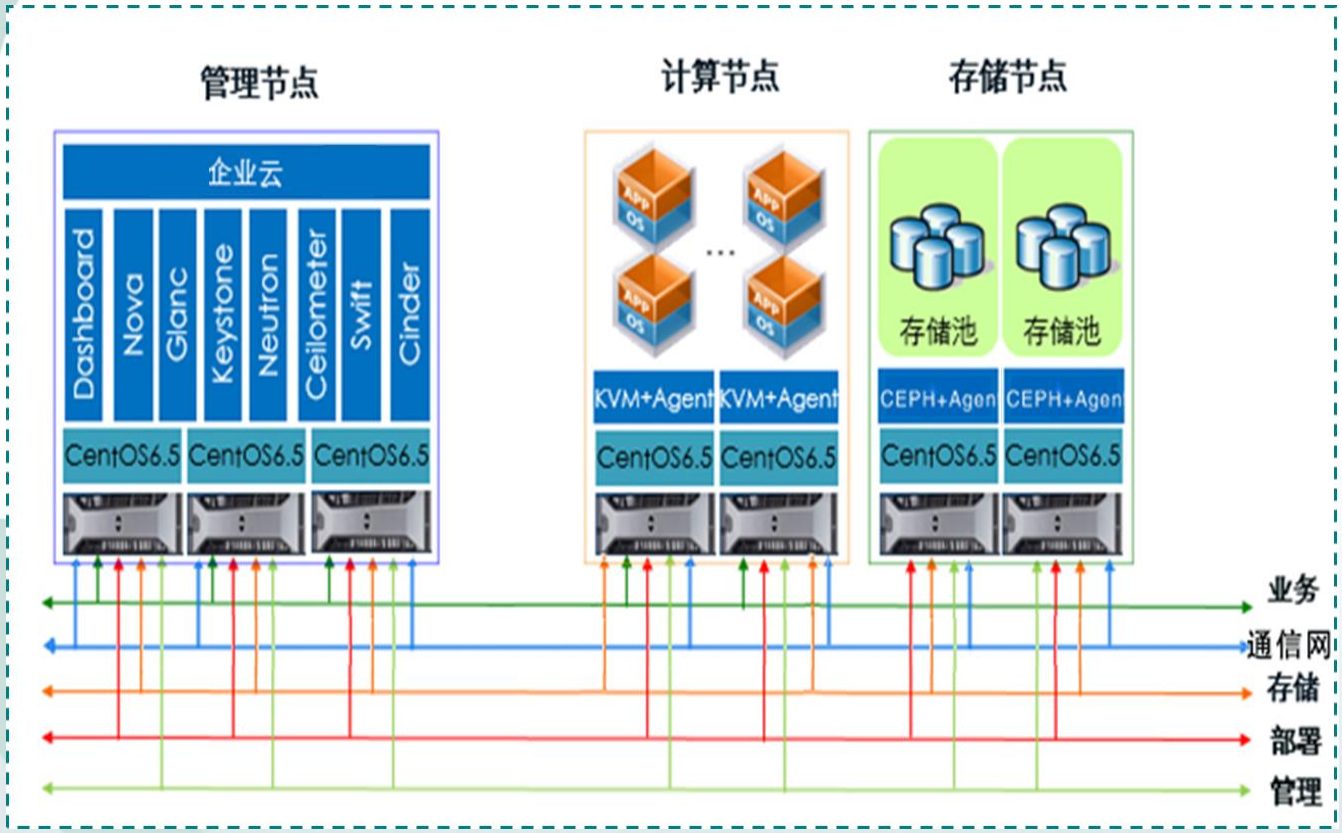
本期规划700台物理服务器，本期资源清单如下：

设备型号	用途	数量	CPU（Core）	内存	SATA盘	SSD盘	网卡
x86	控制节点	21	2*8Core E5 2.3GHz	128G	1000G	4*128GB	2*10Gbps 2*1Gbps
x86	计算节点	468	2*10Core E5 2.3GHz	256G	1000G	2*128GB	2*10 Gbps 2*1 Gbps
x86	存储节点	210	2*8Core E5 2.4GHz	128G	1600G	2*128GB	2*10Gbps 2*1Gbps
4万兆交换机	核心交换机	3					
万兆交换机	数据交换机	60					

- 中间件云平台控制节点5台、计算节点237台、存储节点117台
- 数据库云平台采用混合模式即结合oracle rac 和adg技术加上基于scn号级的数据保护机制。数据复写8/16 份



2、架构设计

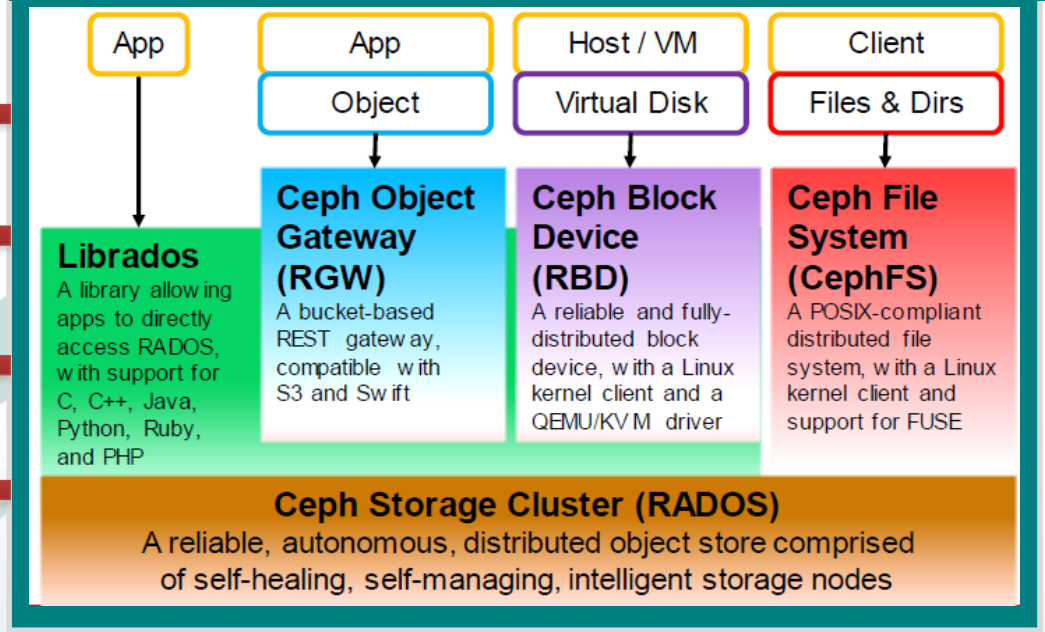


计算资源池设计架构图

- 通信网，用于云内部虚拟机间通信；
- 业务网，用于虚拟机提供外部业务访问；
- 存储网，用于Ceph集群间数据拷贝；
- 部署网，用于物理主机云环境部署；
- 管理网，用于OpenStack与被管理主机间通信。

3、存储设计

本期云平台规划采用Ceph分布式存储架构，为了提高虚拟机可靠性，采用分布式块存储为虚拟机的磁盘，分布式块存储通过在提供存储资源的节点上部署Ceph集群组成。

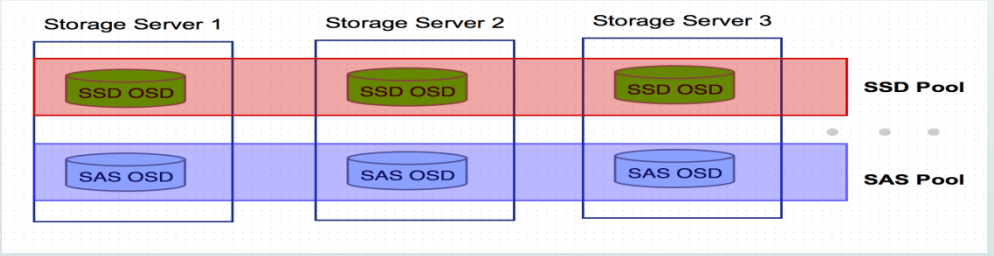


每台机器有个数不同的SAS盘和SSD盘或连接闪存柜，其中SSD盘做为系统盘。其他每个做为一个OSD（Object Storage Device）。

根据节点数量和网络带宽情况，设置Ceph集群中每份数据有三份，一份主本，两个副本。

根据以上资源特点，设计两个不同的资源池来区别不同性能的资源，并为不同的应用和数据提供服务。

- ●SSD极速资源池
- ●HDD高容量资源池



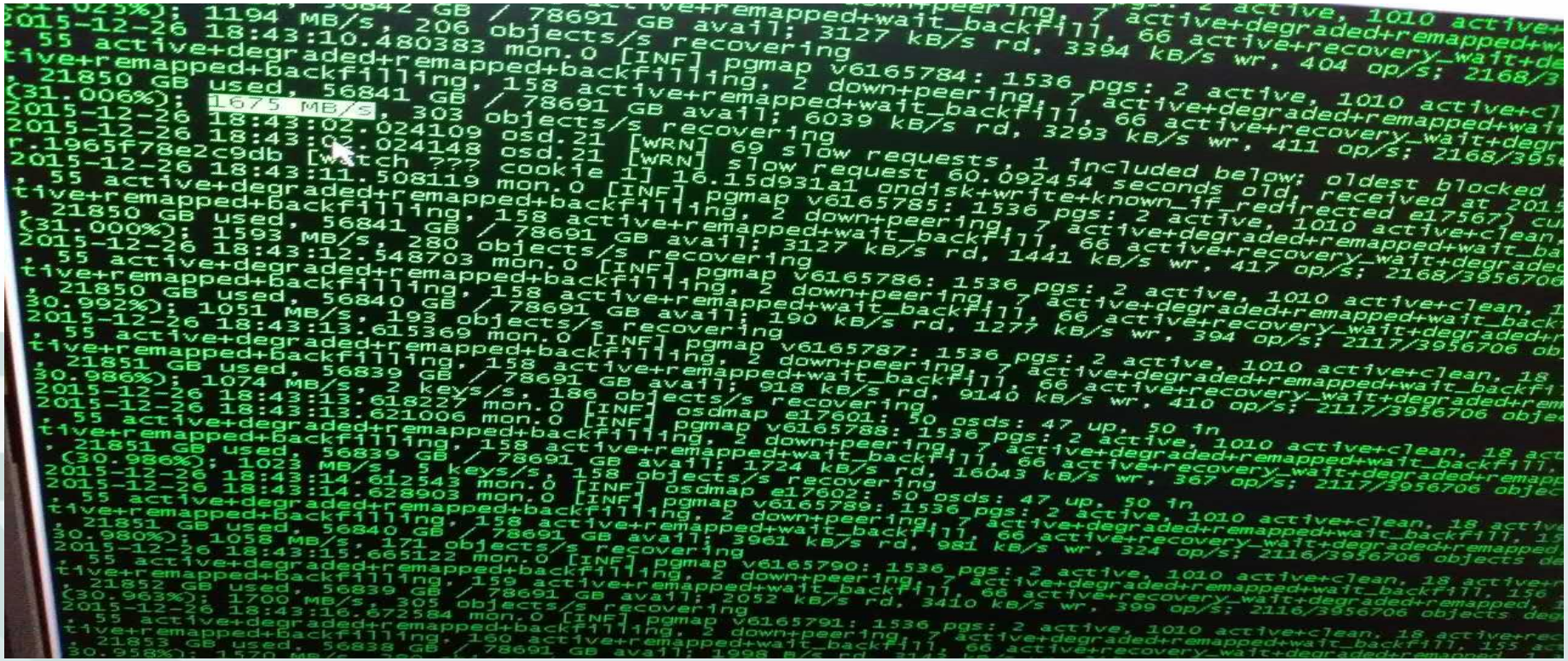
4、存储设计

基于Ceph的分布式高性能存储方案，极大的提升云主机的IO性能，足以应对各种苛刻的企业应用需求。其云平台可以在10秒内完成一台云主机的创建；为云主机提供最高提供6000IOPS及170MB/s的高性能磁盘卷（一般相同配置的单机可提供1608IOPS及78MB/s写操作）；支持实时快照，对1T硬盘的快照的操作耗时不超过2秒。为了实现这一特性，龙江公司在Ceph系统中采取了：

1. 增大文件标识符数量，将文件名描述符全部缓存在内存里；
2. 启用Ceph的Sparse Write（减少写过程），并对XFS（高性能的日志文件系统）进行patch解决filemap的BUG；
3. Ceph的默认参数并不适合SSD，按照SSD的要求重新进行参数调优；
4. 打开RBD Cache（客户端的缓存），可以获得明显的性能提升；
5. 通过选择性的延长特性线程的活跃时间，大大减少Context Switch（上下文切换）的次数，降低I/O延迟；
6. 针对Simple Messenger（单消息）线程数过多、延迟较高的问题，开发了Async Messenger（异步消息）组件，实现更低的延迟，并大大减少了线程数；
7. Ceph默认的Cache机制并龙江OpenStack架构的缓存特点，引入RandomCache（随机缓存）解决这一问题，Cache访问速度比以往提高一个量级。

4、存储设计

2015年12月26日龙江对云平台模拟掉电试验，下午14时，人为操作10台机器掉电，1小时后，恢复。恢复时大约产生51Tb数据数据量，造成网络拥堵。峰值数据达1675M/s。



5、安全设计

为保证云平台的安全，避免虚拟化后的信息安全，龙江公司在云平台建设中采用了自主可控安全管理系统对云平台进行防护：

目录

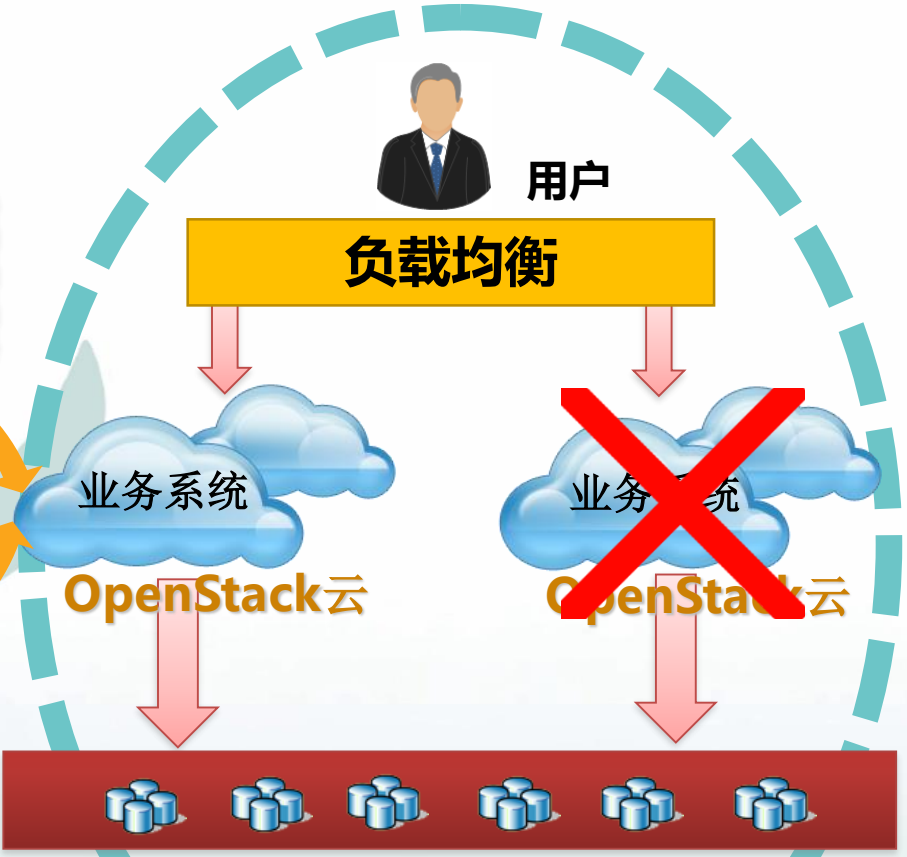
1 V To V

虚拟机对虚拟机：完成原部署在其他不同版本下VMWare中业务系统至新的云平台环境的迁移。

业务系统迁移方式

2 P To V

物理机对虚拟机：完成原部署在不同物理主机环境中（包括小型机、PC服务器）业务应用至新云平台环境的迁移。



当其中一片云因故障整体宕机的情况下，也可以保证服务的连续性。对于特殊的业务系统，同时保留实体机，将云主机与实体机通过负载均衡设备共同做成集群进行协同工作。

三、云平台实施

- **第一阶段：**建立起云平台环境。
- **第二阶段：**跌代过程，迁移其他业务系统至新的云平台环境中，并将把腾退的物理机加入到云平台中。
- **第三阶段：**跌代过程，迁移其他业务系统至新的云平台环境中，并将把腾退的物理机加入到云平台中。
- **第四阶段：**全面建成，在本阶段国网黑龙江电力计划将全部业务系统（注：除特殊环境要求的业务系统外）迁移至云平台中。

2014年7月 — 2015年3月 — 2015年7月31日 — 2016年1月30日

在本项目中，实施过程分为四个阶段

在第四阶段中

- 验证云平台的技术架构、业务系统及接口的迁移方案、计划检修方案及操作方案。
- 按照时间表的要求，分批分步实施业务系统迁移。
- 测试迁移后业务系统及接口稳定性，持续优化云平台。

网黑龙江电力云计算平台
共计部署物理主机272台：

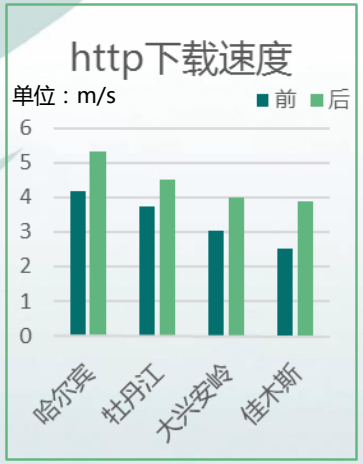
1. 内网事务云200台
2. 内网计算云60台
3. 外网事务云12台

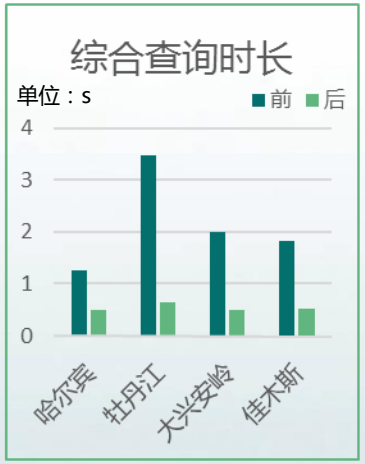
创建385台虚拟机，完成营销系统、财务管控、PMS2.0、电力交易、GIS平台等80%系统迁移。营销系统已连续运行11个月，运行效率提升近30%。预计在2016年1月底完成100%业务系统入云。

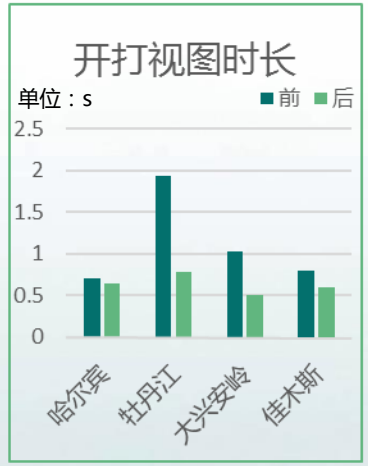
目前，共部署物理主机272台，创建385台虚拟机，完成营销系统、财务管控、PMS2.0、电力交易、GIS平台等80%系统迁移。营销系统已连续运行11个月，运行效率提升近30%。预计在2016年1月底完成100%业务系统入云。

单位	http下载速度		综合查询		打开视图	
	前	后	前	后	前	后
哈尔滨供电公司	4.17m/s	5.33 m/s	1.26s	0.50s	0.7s	0.64s
牡丹江供电公司	3.74m/s	4.52 m/s	3.47s	0.65s	1.93s	0.78s
大兴安岭供电公司	3.02m/s	4.01 m/s	1.99s	0.49s	1.03s	0.51s
佳木斯供电公司	2.50m/s	3.89 m/s	1.82s	0.52s	0.8s	0.6s

系统迁移前后测试对比表







系统迁移前后测试对比图

四、工作成果

自2014年7月，国网黑龙江电力已营销生产数据库、营销历史数据库、营销费控库、营销ODS数据中心、财务管控、GIS系统、PMS系统、IMS数据库等入云，已稳定运行15个月。效率提升明显：营销月结时计算时长从原12小时缩短到72分钟，效率提升900%； 营销系统合帐报表耗时从原1小时47分钟缩短到4分钟效率提升2575%；高峰页面访问响应时长由6-8秒缩短至1-2秒，效率提升300%。

四、工作成果

完成Hadoop平台搭建，实现离线数据分析；在线统一日志分析平台（基于SPARK）实现对服务器、网络设备、安全设备、数据库、系统中间件、权限管理系统、端设备的日志收集，对即时通讯系统信息分析。

1. • 通过分析网络“五元组”数据，实现网络行为建模（白环境）；
2. • 更好为用户提供直观的数据关联，实现数据可视化分析；
3. • 更好的分析员工的需求和想法，使公司产品能够更加适合生 产需要；
4. • 更好的定制化产品或为用户提供服务；
5. • 在实时数据的趋势和预测上更加主动，为状态检修提供数据支持。

目录

五、经验与建议

1. 建议充分保证足够冗余，包括交换机、网卡，并采用高速网络以保证足够带宽及I/O；
2. 尽量避免服务器品牌过杂、型号过多；
3. 建议计算节点采用4路PC服务器，大容量内存;
4. 建议存储节点采用2路PC服务器，128G内存，多磁盘盘位机型，同时RAID带有直通模式或是采用直通卡；
5. 云平台单节点建议采用2块双光口网块用于云平台数据交换及对外提供服务，同时保持单节点网络冗余；一个千兆电口用于管理网络；
5. 单个千兆PXE接口，用于部署；

五、经验与建议

7 • 云平台机柜内接入交换机建议采用两台万兆光交换机，单台千兆电口交换机，可与信息网通用

8 • 汇聚交换机建设采用4万兆或是10万兆交换机；

9 • 业务系统在迁移时，可通过虚拟机与实机相结合的方式，即在云下部署虚拟机，并加入到负载集群中，通过负载集群转发业务，并验证新环境是否有问题；

10 • 对于VMware虚拟机迁移，建议不采用VToV的方式迁移，直接采用重新部署模式；

11 • 为保证业务系统顺利迁移，建议在云环境建设时，充分考虑IP段的使用情况，保证有冗余的IP以方便部署新环境；

12 • 为保证两片云间可以业务互漂，需要考虑好云的性能冗余，以免由于性能不足无法全部接管另一片云业务。

谢谢！