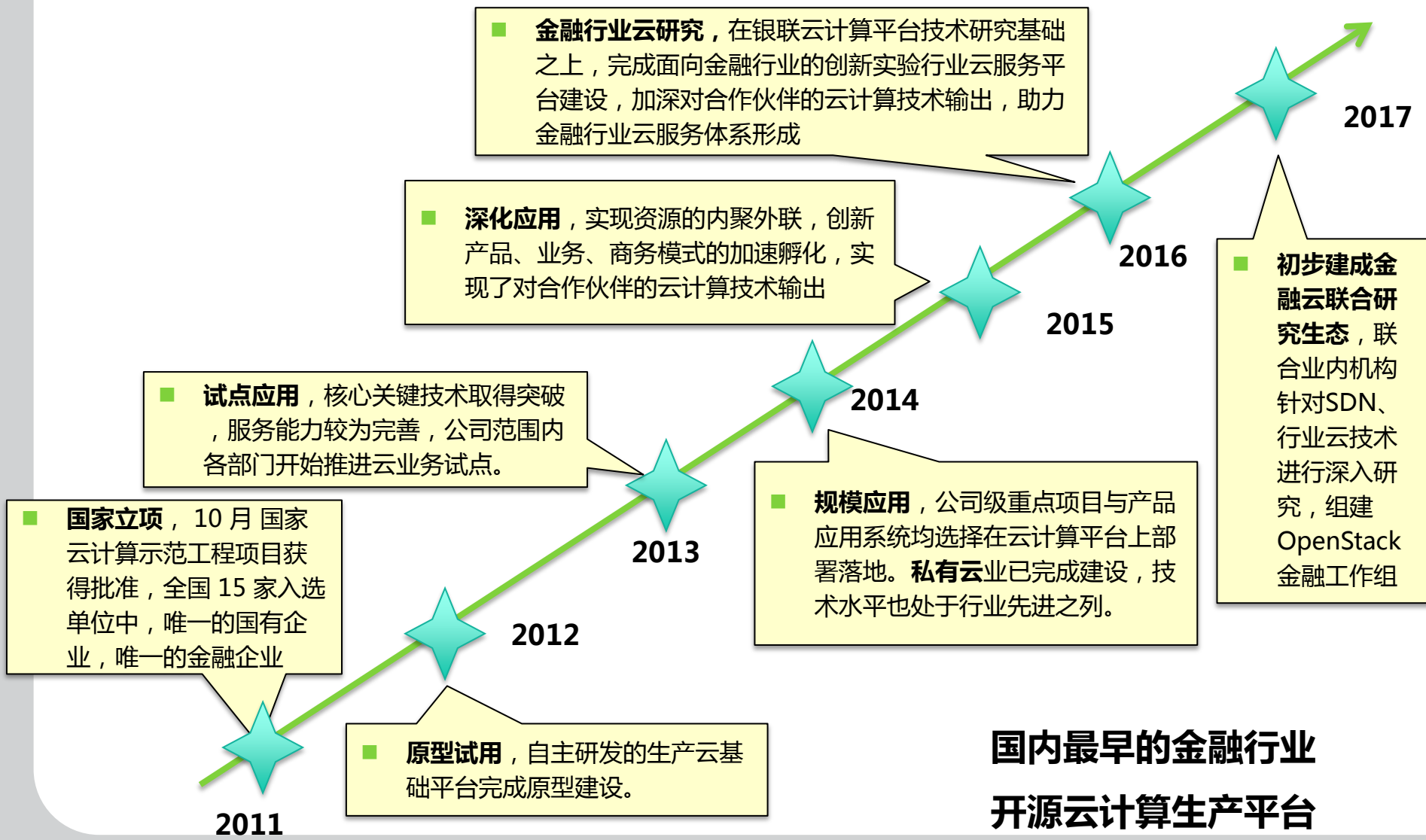
基于开源SDN控制器的金融云网络方案研究

2018 6

中国银联 China UnionPay

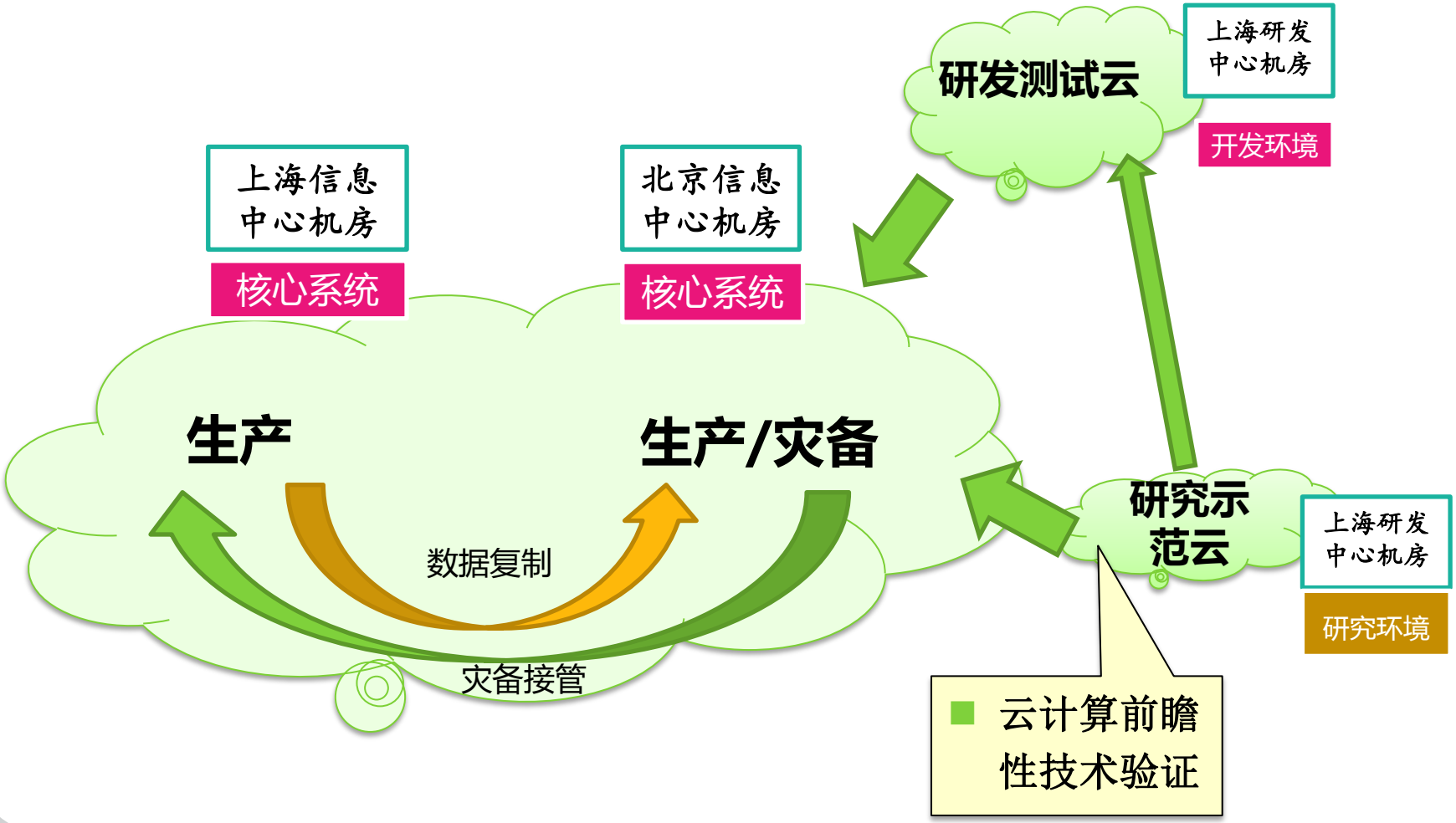
银联云计算实践历程



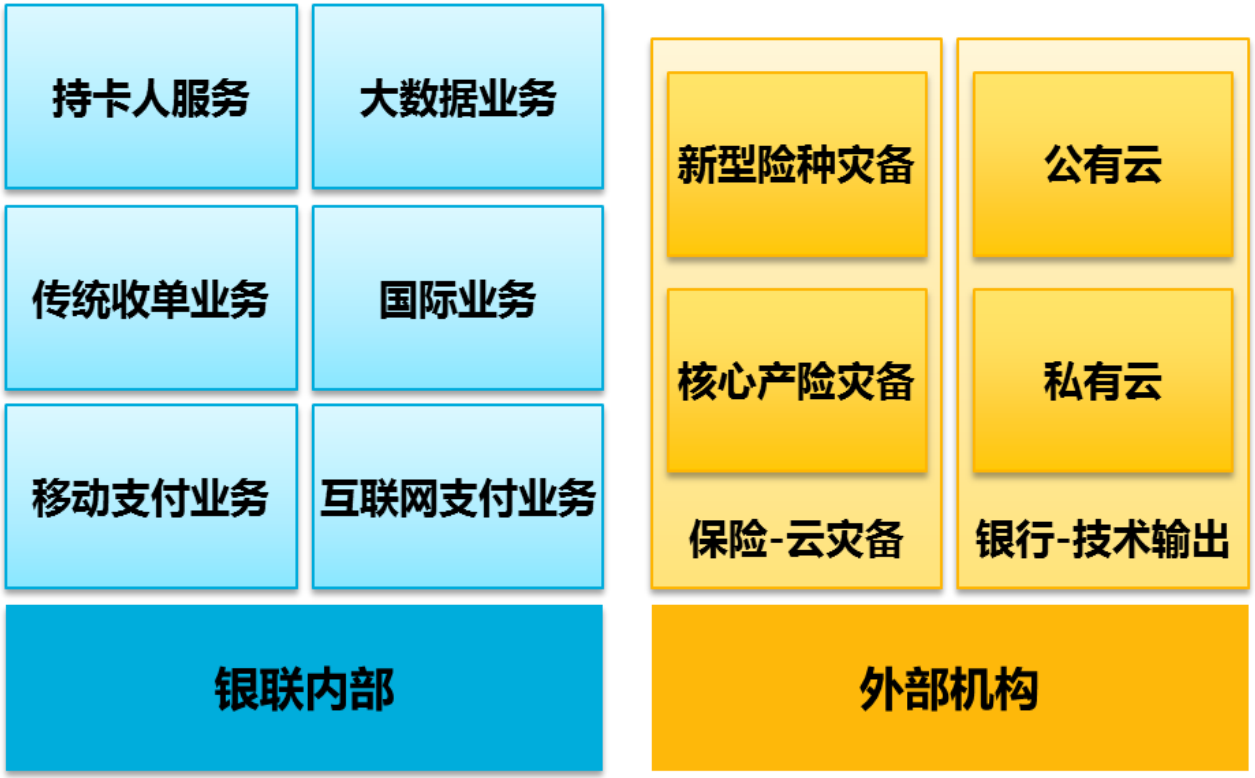
国内最早的金融行业开源云计算生产平台

云平台部署情况

生产灾备云、研发测试云、研究示范云三朵子云互为联动



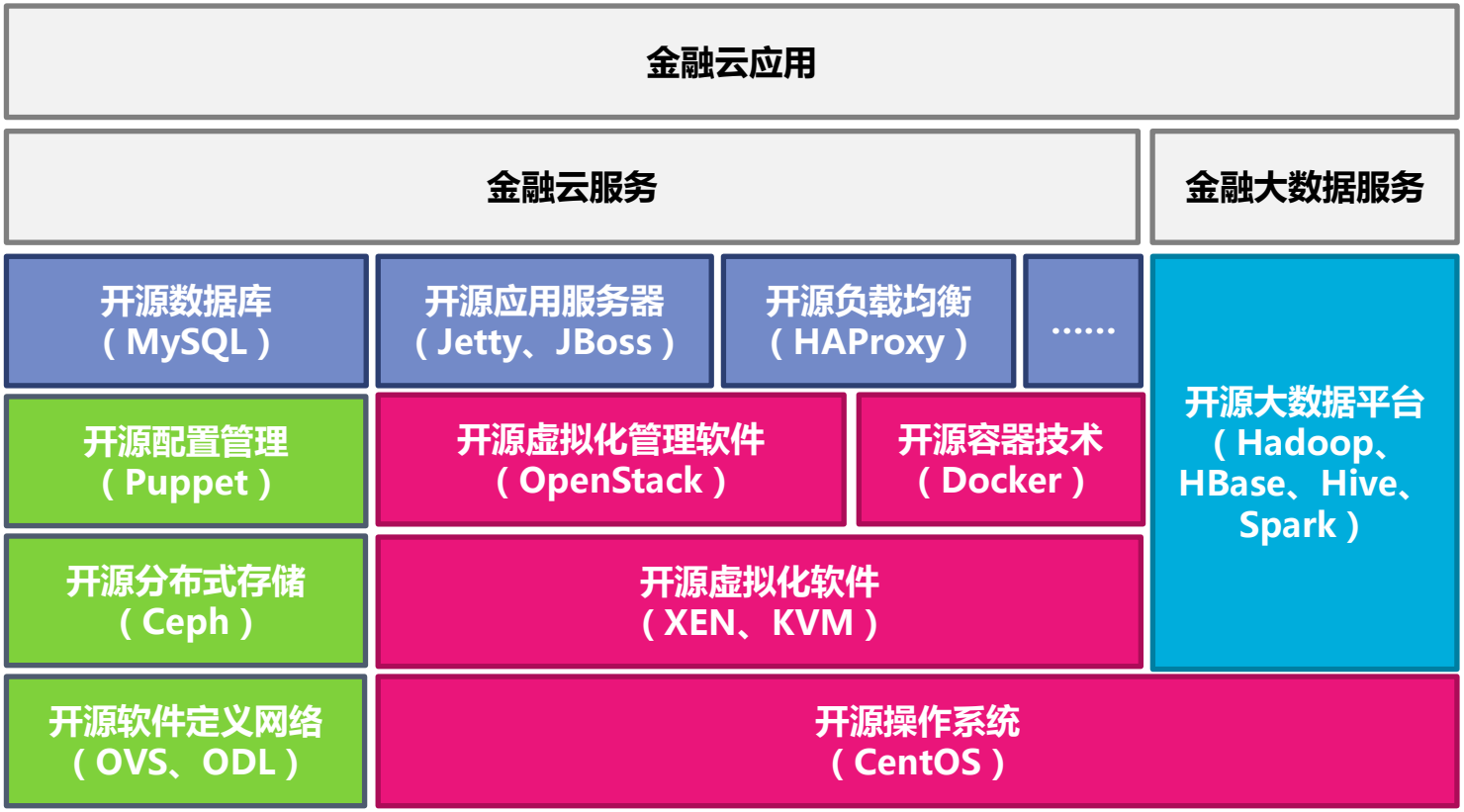
业务应用情况



- ✓ 银联内部：总/分/子公司业务支撑，云闪付统一APP、移动支付(ApplePay)、银联国际卡权益等关键应用均基于云平台
- ✓ 外部机构：在银行与保险业的合作伙伴中提供云计算各类服务，形成技术输出初步辐射

银联云技术路线

- 基于业界领先的开源技术和开源软件，自主研发基础平台
- 站在开源技术最前沿，即有继承又有创新



✓ *功能强大*

✓ *成本节约*

✓ *响应及时*

17年关键技术突破

2017年在生产环境实现新一代基于SDN技术的金融云平台应用

- **商业硬件方案：华为**
- **开源软件方案：Neutron**

OpenStack版本 E版 -> L版

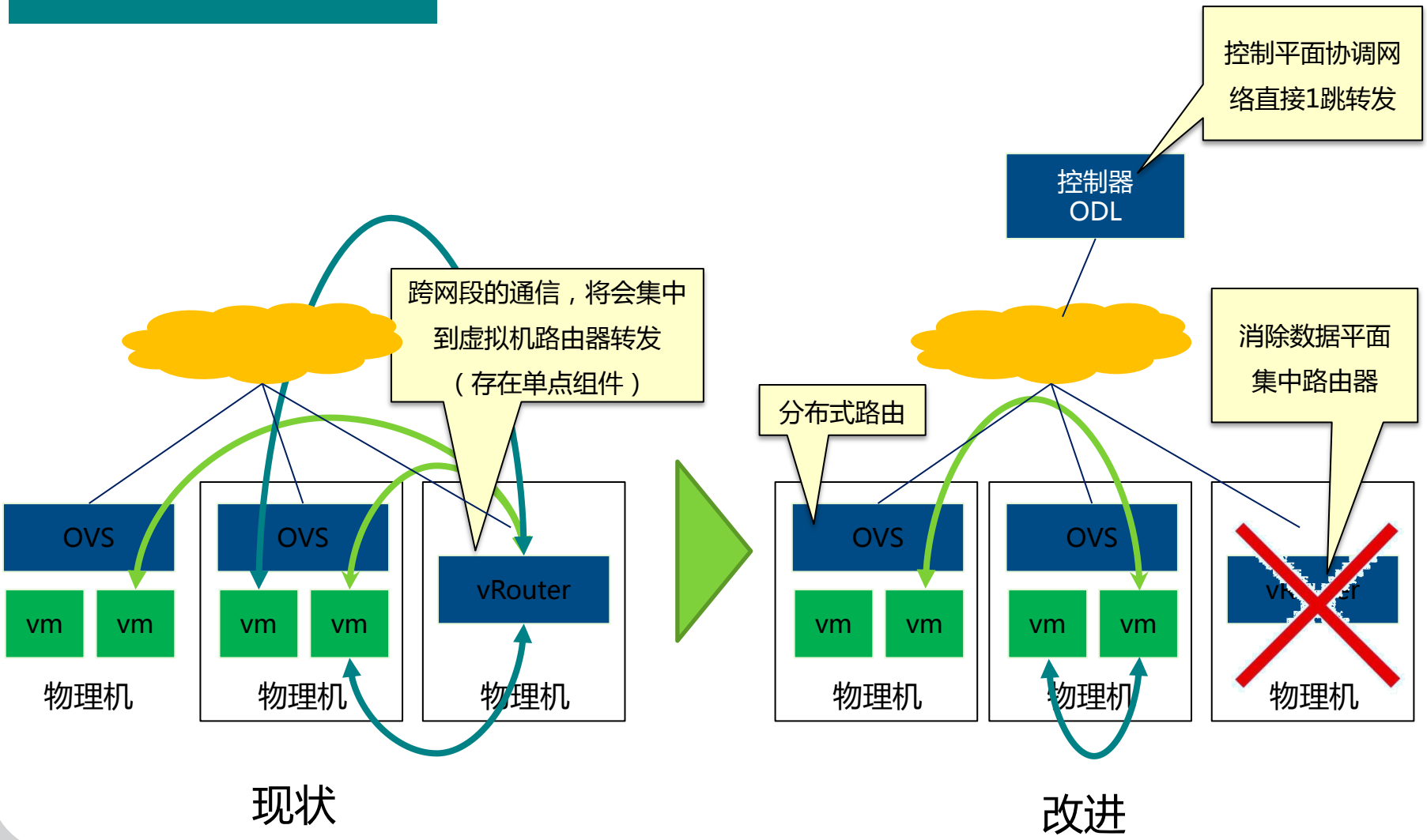
银行业云闪付统一APP承载验证

当前基础设施即服务的技术攻关聚焦网络

异构SDN区域互联（RI）	开源SDN控制器(ODL)	云网监控（Hadoop）
数据中心内各云区域网络一致性架构	软件SDN技术的自主优化方案	SDN技术应用后网络运维强化

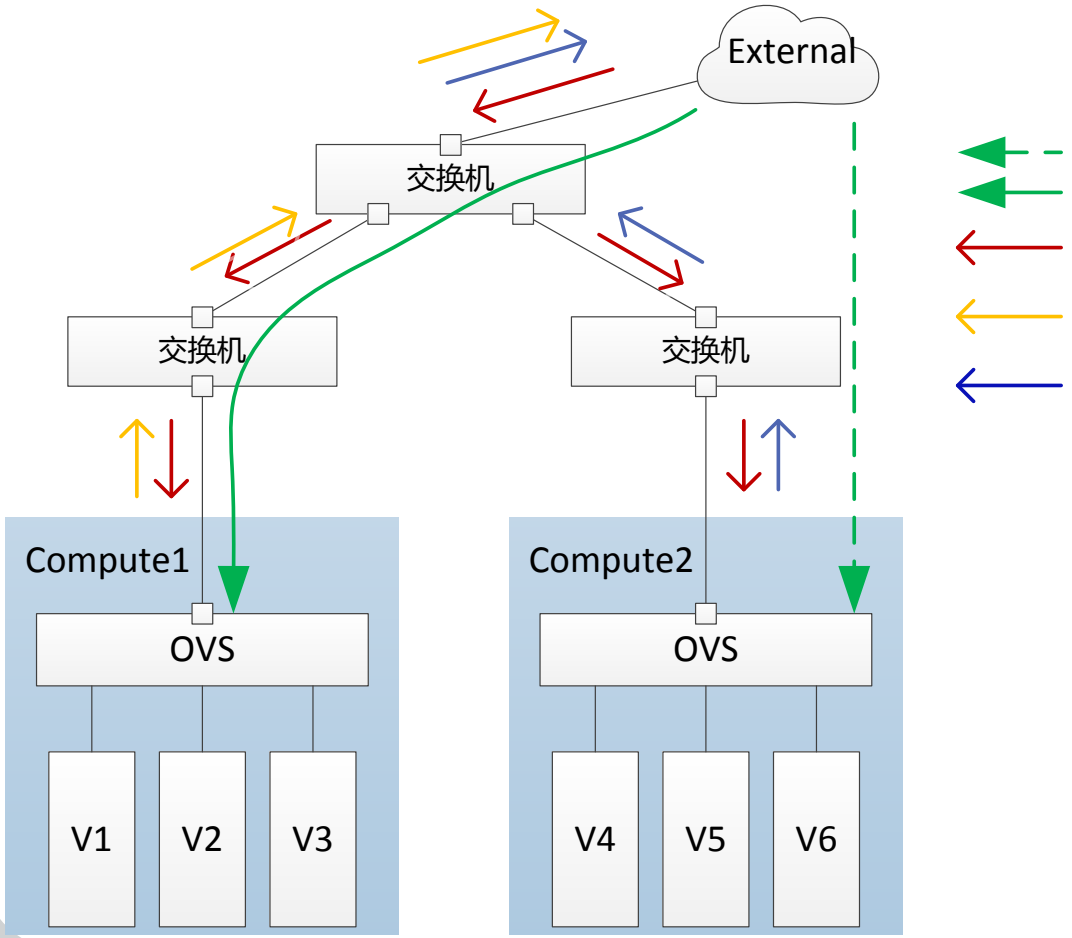
优化点：可靠性

分布式路由



优化点：可靠性

ARP代答



ARP代答实现前

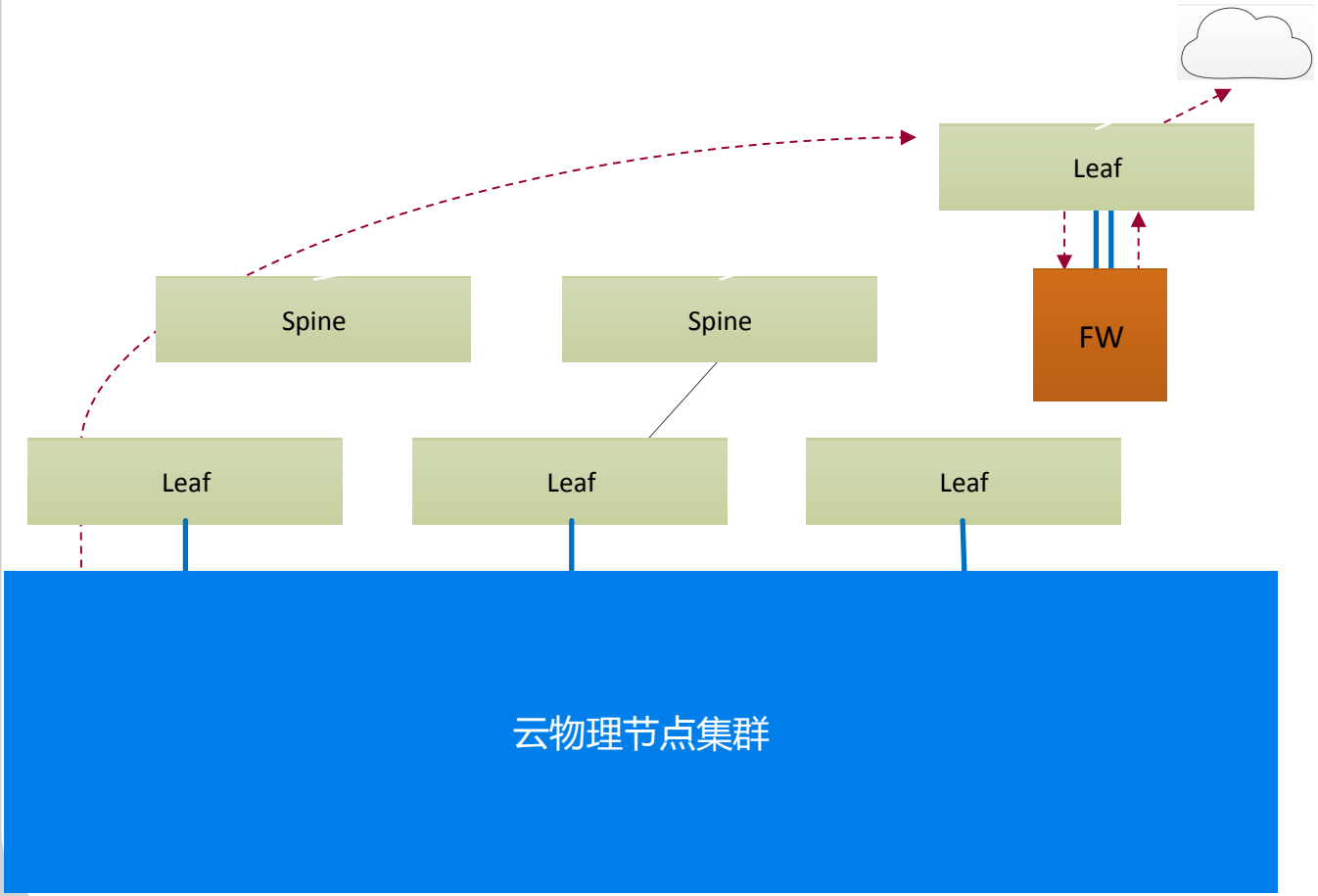
由于软件SDN方案中没有使用硬件路由器设备，所以整个二层域都会成为ARP包的广播域，影响面较广。

ARP代答实现后

由ARP请求发起虚拟机所连接的OVS中的流表直接进行ARP响应，该过程中ARP请求不会进行广播。

优化点：可靠性

防火墙并联接入

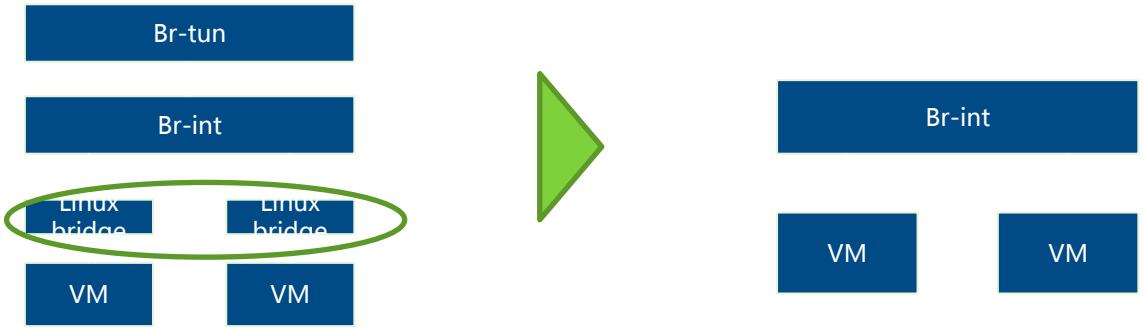


- ✓ 提升防火墙稳定性，外部网关不能设置在防火墙上；
- ✓ 通过引流做到防火墙物理并联，逻辑串联的效果；
- ✓ 防火墙故障流量可不通过防火墙直接出区域；

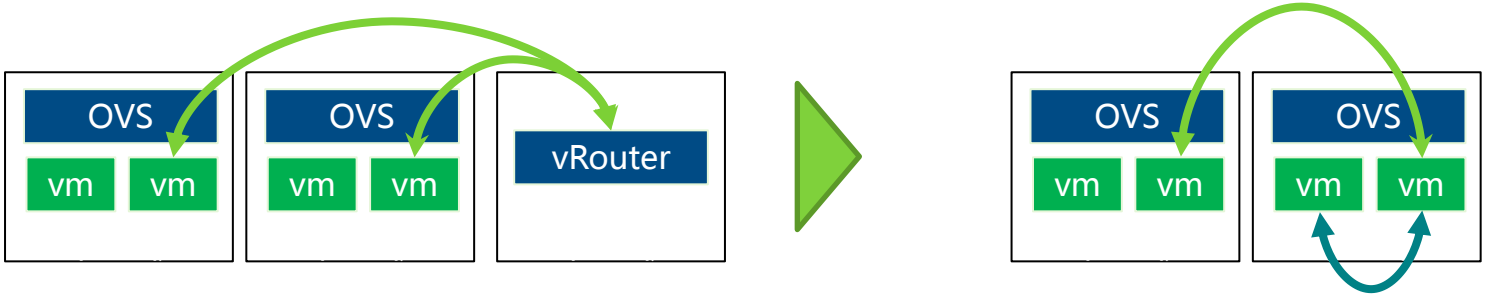
优化点：性能

精简数据传输路径

- 服务器系统网络软件精简

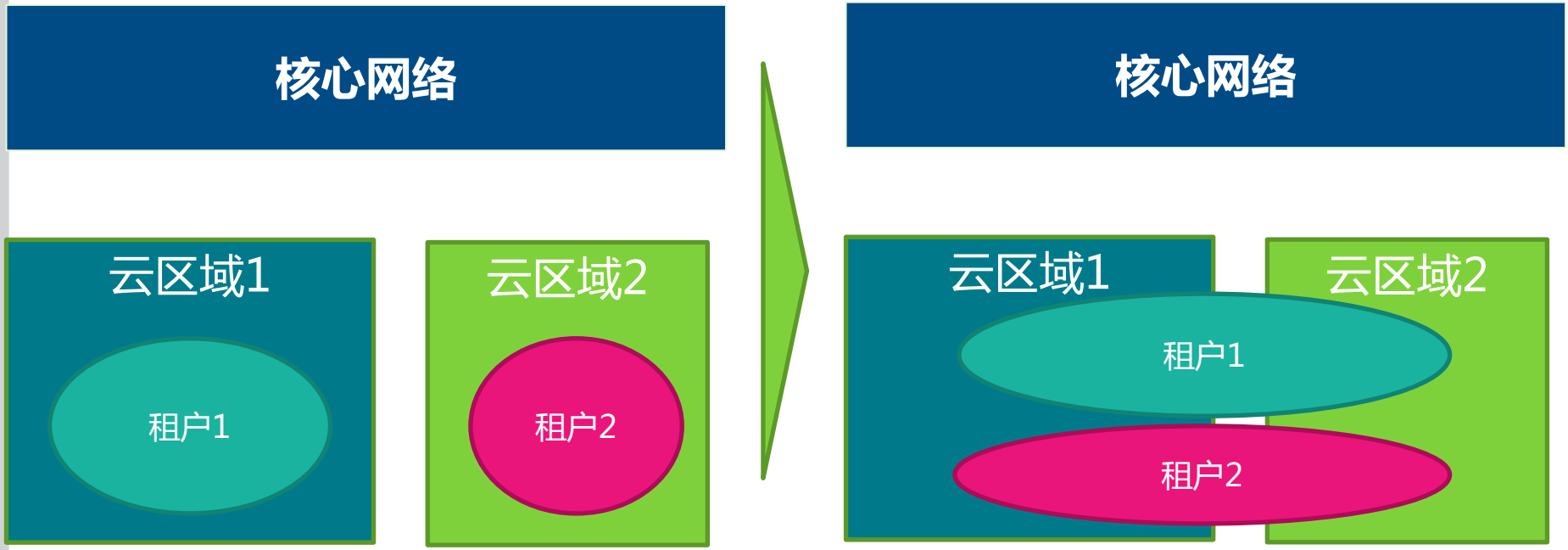


- 网络传输跳数减少

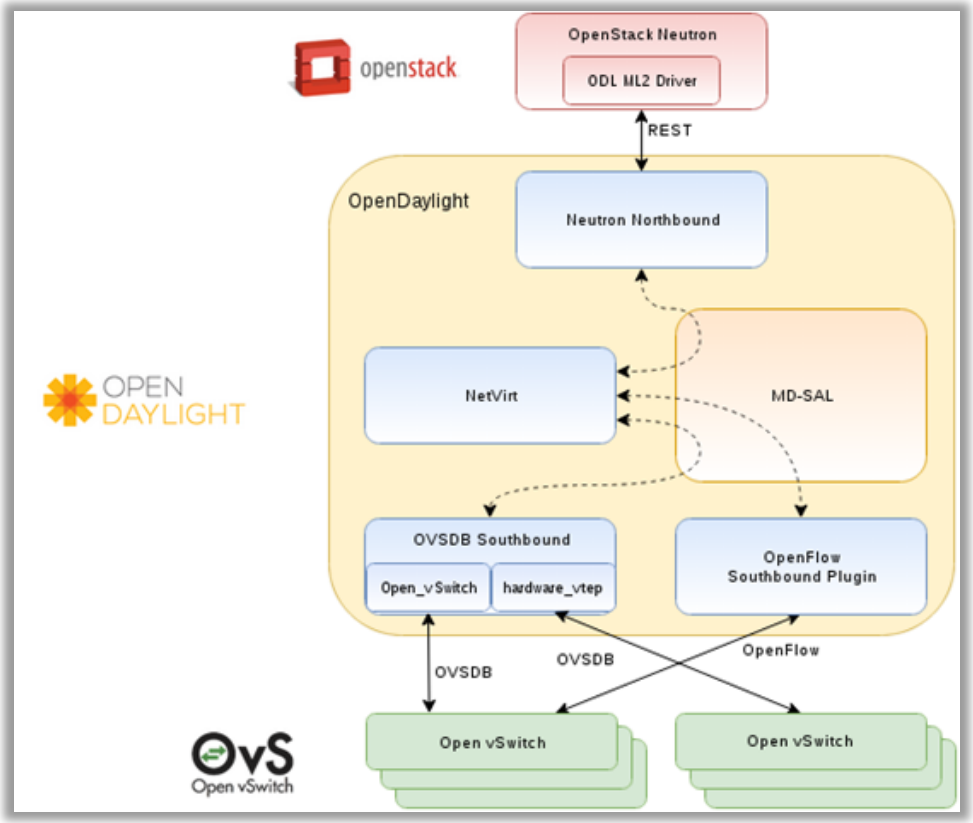


优化点：扩展性

租户资源跨区域互联



基于ODL的开源网络架构



- 管理平面：Openstack

  ML2对接

- 控制平面：ODL

  OVSDB & Openflow

- 数据平面：OVS

能力规划

ODL原生能力

分布式路由

分布式ARP代答

物理机内部链路精简

附加能力

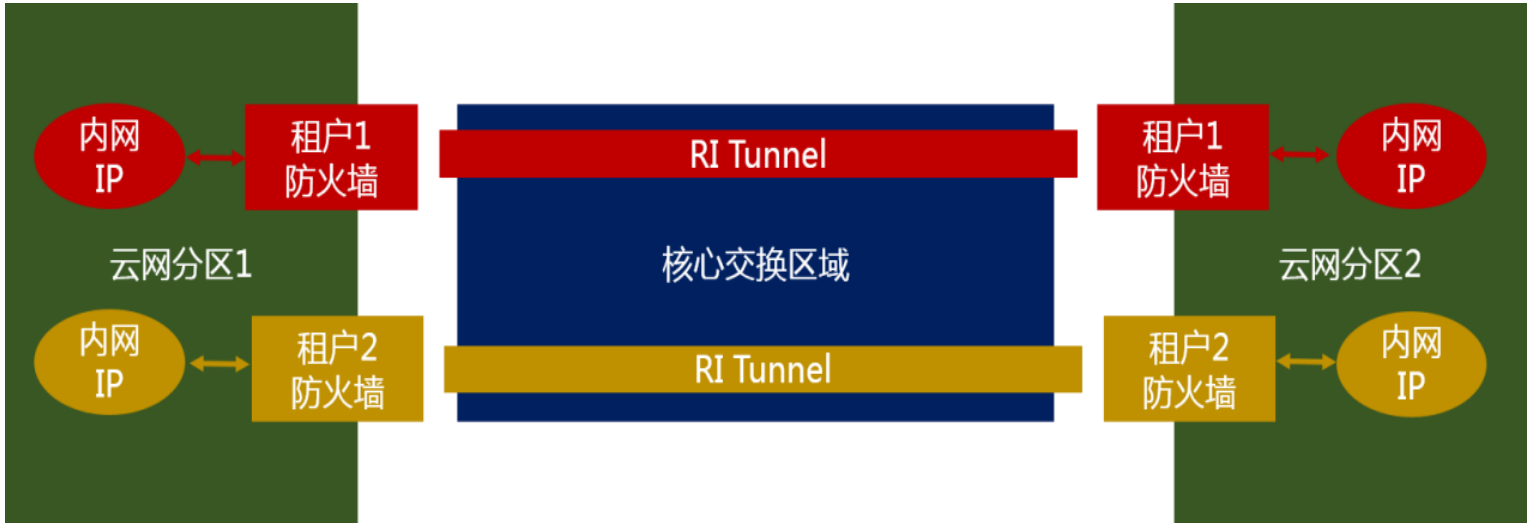
跨区域互联

防火墙引流

为实现附加能力，需要对ODL原生能力进行增强

能力实现一：跨区域互联

区域外部对接RI，实现在租户数据在核心交换区域传输



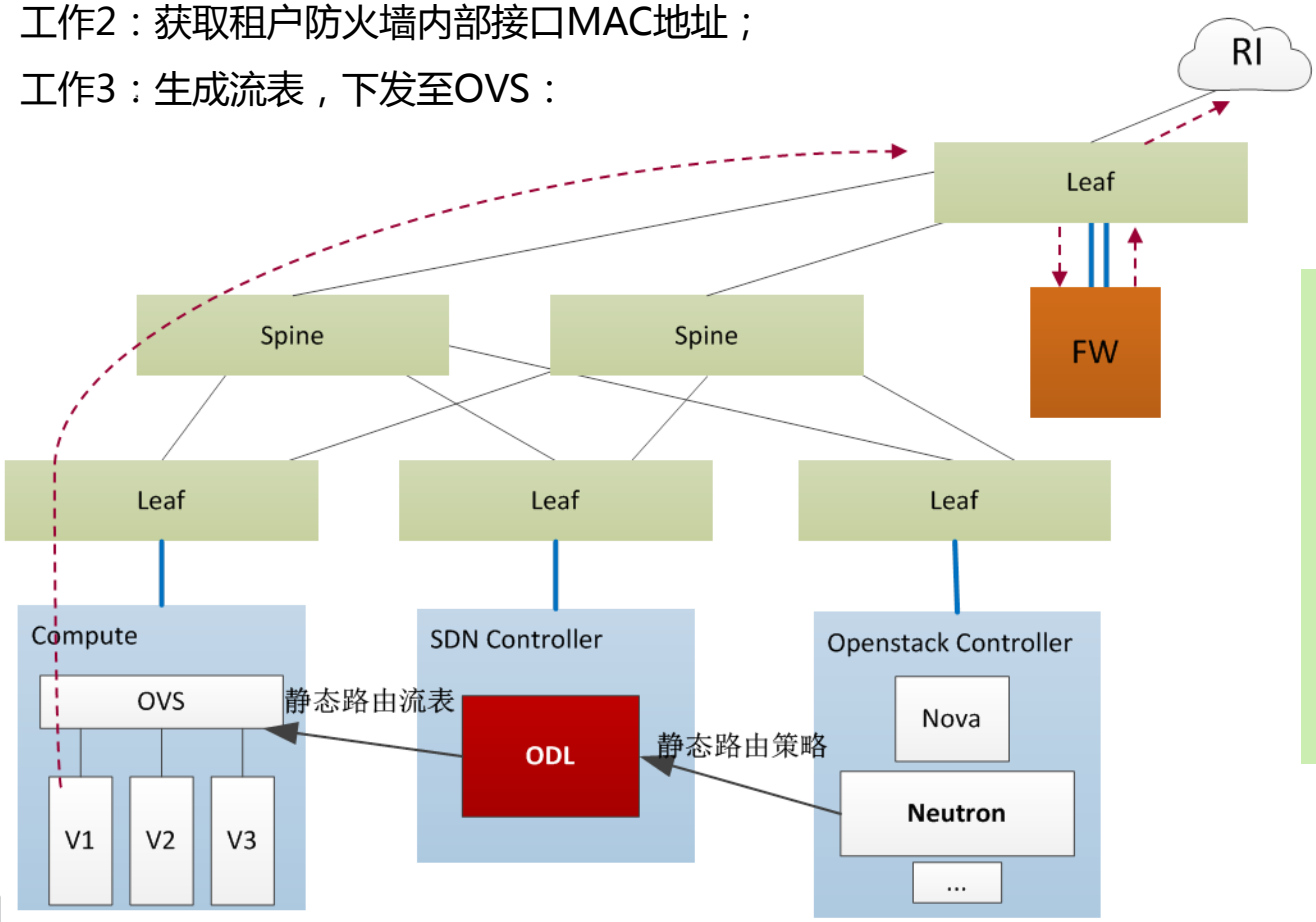
能力实现二：防火墙引流

集成Openstack静态路由功能，通过配置相关静态路由，生成相应的OpenFlow流表，下发至OVS中进行数据传输。

工作1：对接静态路由功能，监控并获取静态路由数据；

工作2：获取租户防火墙内部接口MAC地址；

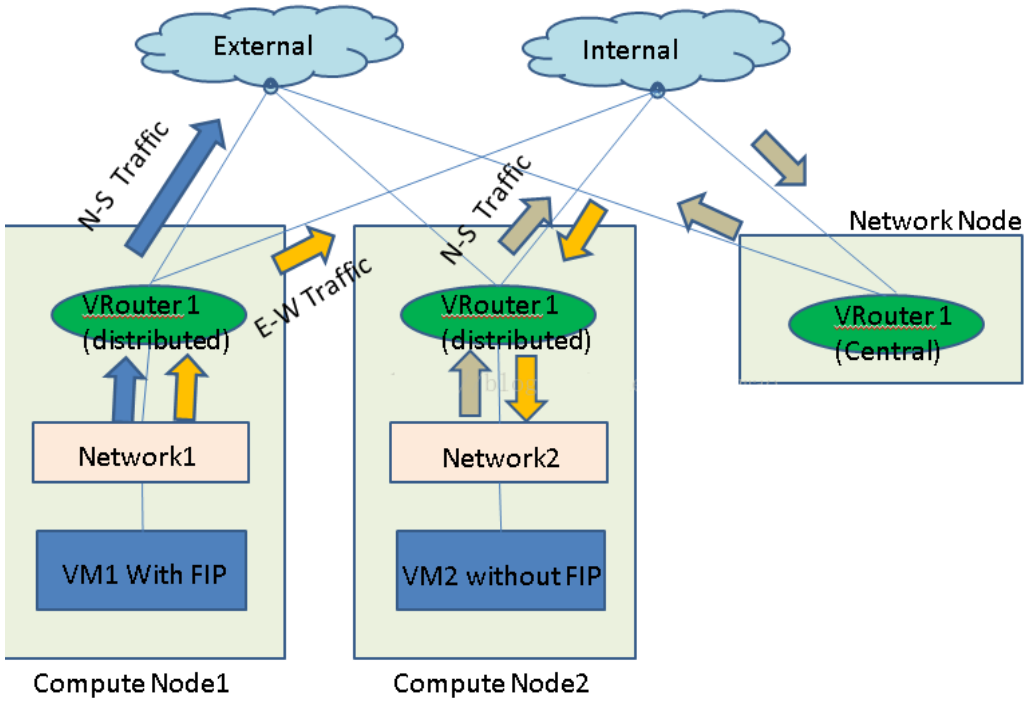
工作3：生成流表，下发至OVS：




```
table=60,priority=4096,ip
,tun_id=0x1e,nw_src=10.
1.1.0/24,nw_dst=10.2.1.0/
24,actions=set_field:f8:4a:
bf:5a:2b:ea -
>eth_dst,dec_ttl,mod_vla
n_vid:211,output:3
```

能力实现三：支持去floating IP的分布式路由

ODL原生分布式路由



区域内部跨Network的东西向流量，直接发送；

有floating ip的南北向流量，直接发送；

不支持无floating ip的南北向流量直接发出。

原因分析

1.分布式Vrouter外部接口不具备接收外部流量的能力；

2.分布式状态下无法对虚拟机位置进行定位；

能力实现三：支持去floating IP的分布式路由

1.实现路由器外部接口对外来流量的数据接收能力

之所以路由器外部接口不能接收外部数据，主要是软件SDN分布式路由在设计的时候，没有赋予该接口的ARP相应的能力，所以第一步就是在ovs中加入外部接口相应ARP请求的流表。流表如下：

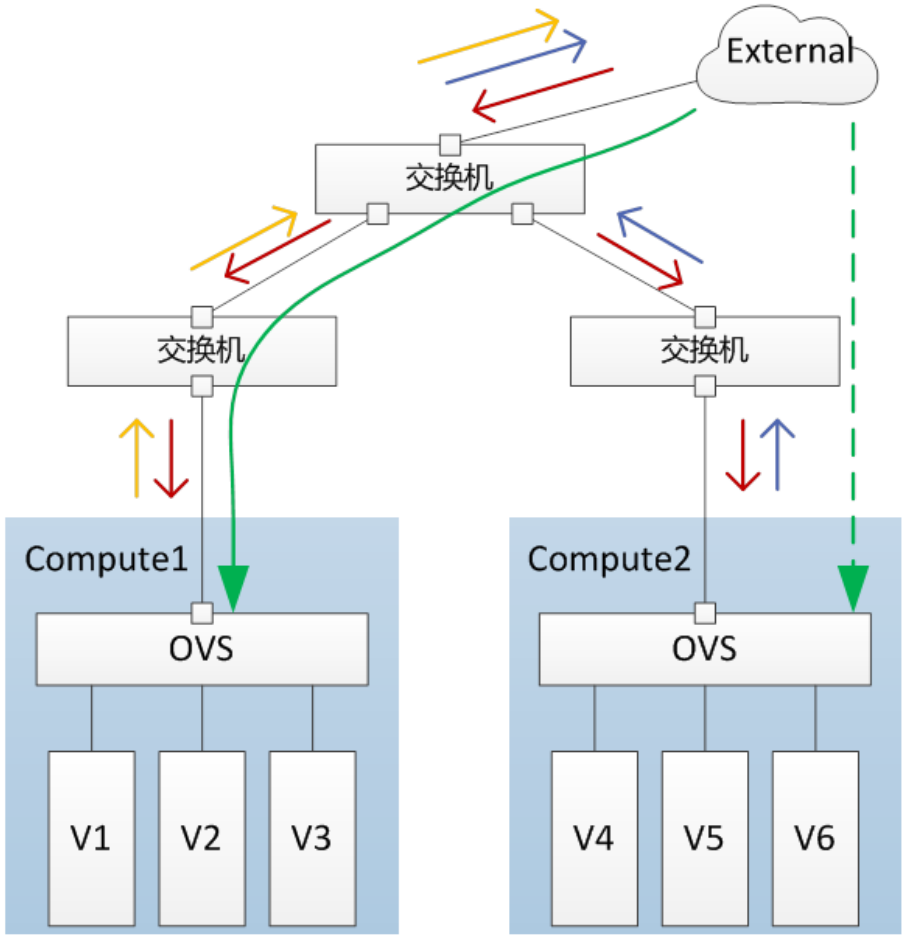
```
table=20,priority=1024,arp,arp_tpa=172.16.1.3,arp_op=1 actions=move:NXM_OF_ETH_SRC[]->NXM_OF_ETH_DST[],set_field:f8:4a:bf:5a:2b:ea->eth_src,load:0x2->NXM_OF_ARP_OP[],move:NXM_NX_ARP_SHA[]->NXM_NX_ARP_THA[],move:NXM_OF_ARP_SPA[]->NXM_OF_ARP_TPA[],load:0xf84abf5a2bea->NXM_NX_ARP_SHA[],load:0xac100164->NXM_OF_ARP_SPA[],IN_PORT
```

上面的流表的主要作用就是为外部接口构造了一个ARP的相应包，在接收到ARP请求的时候，OVS会根据该流表生成一个ARP相应包，发回给请求方。当请求方接收到该ARP回包后，就会将数据包发送到该接口。

下发方式：全量OVS下发

能力实现三：支持去floating IP的分布式路由

2.虚拟机的定位能力



数据包

ARP请求

ARP响应

ARP响应

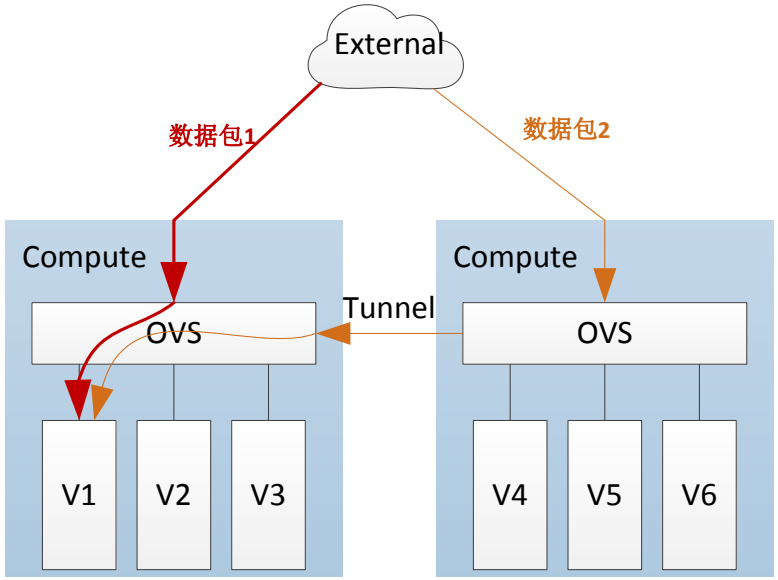
因为路由器的外部接口ARP响应流表是通过全量下发的方式下发到区域内所有的OVS中，所以所有的OVS都会对外部数据的ARP请求进行响应。

收到ARP响应后，外部网络就会将数据包发出，发出后数据包就会按照物理交换机上的mac表进行转发，最终发送到平台中的某一个物理节点的OVS上。

1.目标虚拟机恰好在该物理节点中

2.目标虚拟机不在该物理节点中

能力实现三：支持去floating IP的分布式路由



目标虚拟机恰好在该物理节点中

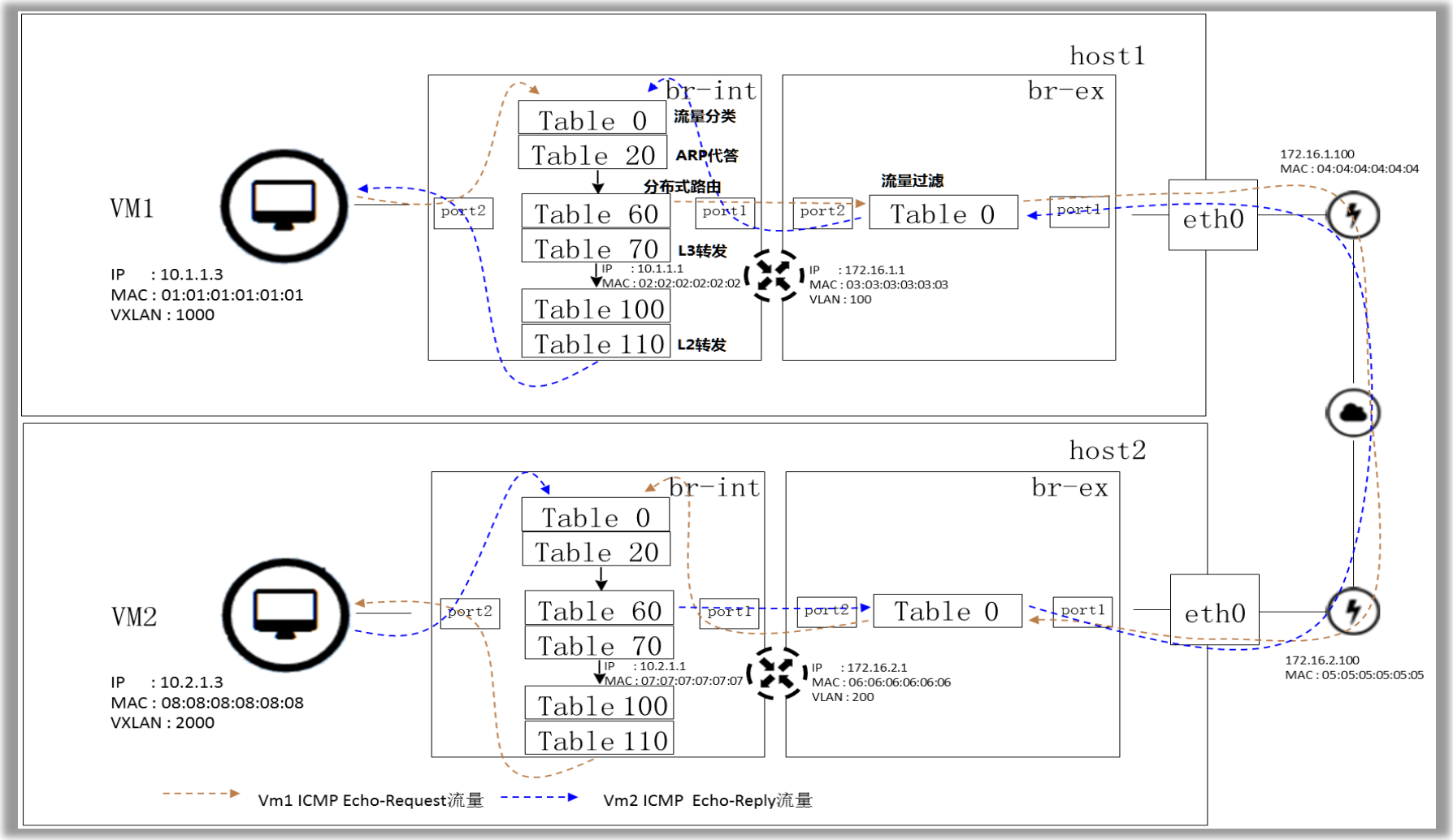
table=70,priority=1024,ip,tun_id=0x5a,nw_dst=10.0.0.3 actions=set_field:fa:16:3e:99:df:47->eth_dst,goto_table:80（三层转发）
table=110, tun_id=0x5a,dl_dst=fa:16:3e:99:df:47 actions=output:23（二层转发到虚拟机，23口与是虚拟机连接的ovs的端口）

目标虚拟机不在该物理节点中

本地物理机流表
table=70,priority=1024,ip,tun_id=0x5a,nw_dst=10.0.0.3 actions=set_field:fa:16:3e:99:df:47->eth_dst,goto_table:80（三层转发）
table=110, tun_id=0x5a,dl_dst=fa:16:3e:99:df:47 actions=output:3（通过Tunnel转发到对应物理机，后面的output：3代表从3口发出，3口即为隧道的端口）

对端物理机流表
table=110, tun_id=0x5a,dl_dst=fa:16:3e:99:df:47 actions=output:23（二层转发到虚拟机）

整体流表架构


host1
br-int
br-ex
Table 0
流量分类
Table 20
ARP代答
分布式路由
Table 60
Table 70
L3转发
Table 100
Table 110
L2转发
流量过滤
Table 0
port1
port2
VM1
IP : 10.1.1.3
MAC : 01:01:01:01:01:01
VXLAN : 1000
IP : 10.1.1.1
MAC : 02:02:02:02:02:02
IP : 172.16.1.1
MAC : 03:03:03:03:03:03
VLAN : 100
eth0
172.16.1.100
MAC : 04:04:04:04:04:04
host2
VM2
IP : 10.2.1.3
MAC : 08:08:08:08:08:08
VXLAN : 2000
IP : 10.2.1.1
MAC : 07:07:07:07:07:07
IP : 172.16.2.1
MAC : 06:06:06:06:06:06
VLAN : 200
172.16.2.100
MAC : 05:05:05:05:05:05
Vm1 ICMP Echo-Request流量
Vm2 ICMP Echo-Reply流量

性能测试对比：单对虚拟机性能

ODL较Neutron：时延平均降低68.8%；带宽平均提升39.3%

数据包大小		256		512		1024		1456	
场景/测试项目		延时（ms）	带宽（G）	延时（ms）	带宽（G）	延时（ms）	带宽（G）	延时（ms）	带宽（G）
同网段同主机	ODL	0.463	28.9	0.449	28.7	0.599	28.8	0.451	29
	Neutron	1.298	18.5	1.267	19.1	1.016	23.4	1.031	22.9
同网段不同主机	ODL	0.92	0.329	1.376	0.778	2.078	1.75	1.861	2.41
	Neutron	3.765	0.197	4.78	0.501	4.491	1.2	3.613	1.71
跨网段同主机	ODL	0.437	28.3	0.42	29.2	0.677	27.9	0.356	30.1
	Neutron	2.989	0.141	3.962	0.321	4.506	0.689	4.793	1.02
跨网段不同主机	ODL	1.161	0.34	1.725	0.801	1.869	1.7	1.665	2.32
	Neutron	2.898	0.181	2.929	0.422	3.273	0.907	3.353	1.28
跨区域通信	ODL	0.696	0.276	0.851	0.706	1.168	1.57	1.235	2.25
	Neutron	4.246	0.206	4.577	0.551	4.561	1	3.096	1.29

工作总结与后续工作计划

方案尚需完善：

1.支持去floating ip方案仍需优化

2.控制器高可用不成熟

3.L4-L7层方案不完善

后续工作：

已发起跨数据中心多云协同资源管理技术

联合研究课题，希望更多合作伙伴参与

谢谢大家！

中国银联
China UnionPay