CHINA
OpenStack Days

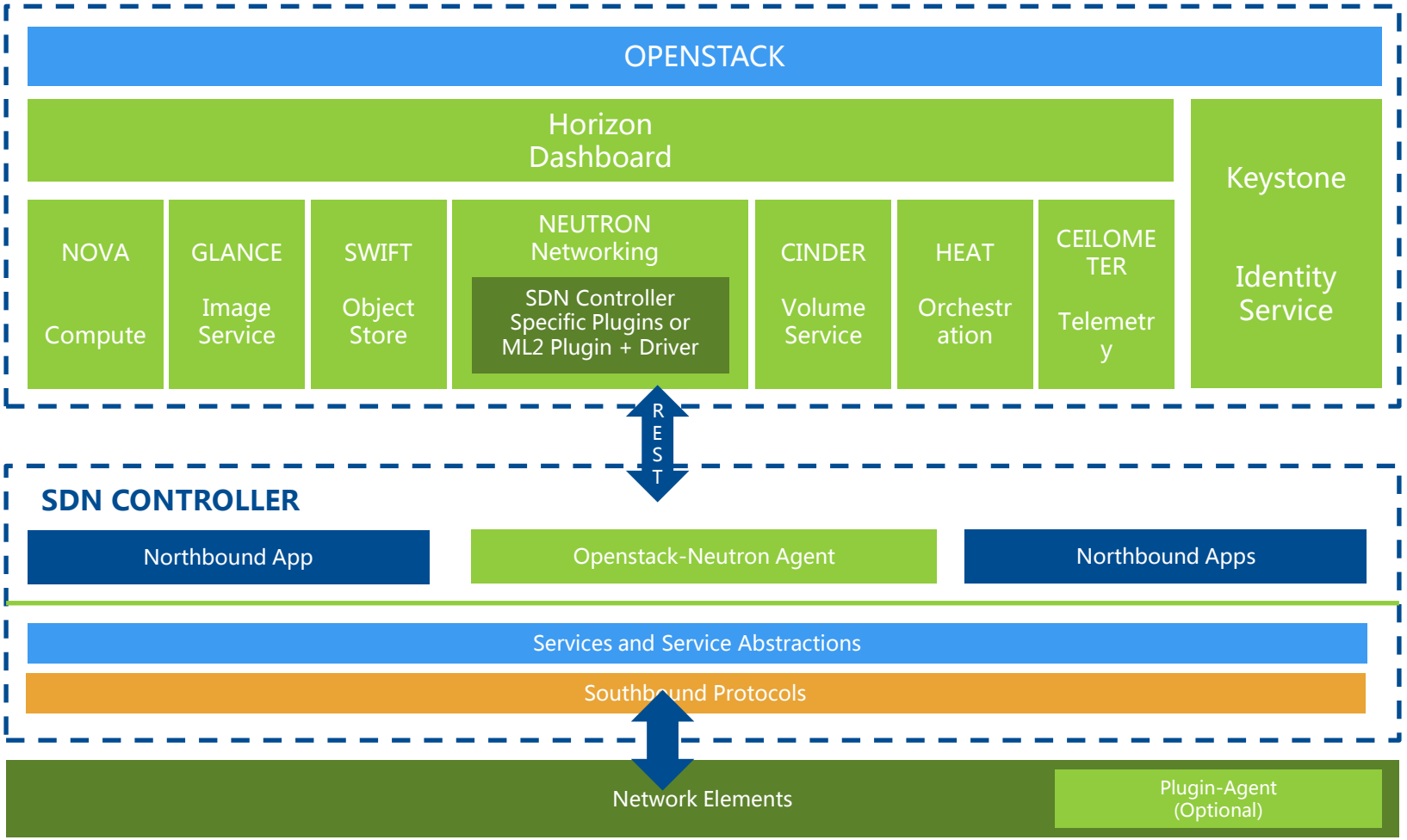
OVN支撑OpenStack全业务网络架构解析

葛建壮

数梦工场（Dt Dream）混合云产品线首席架构师

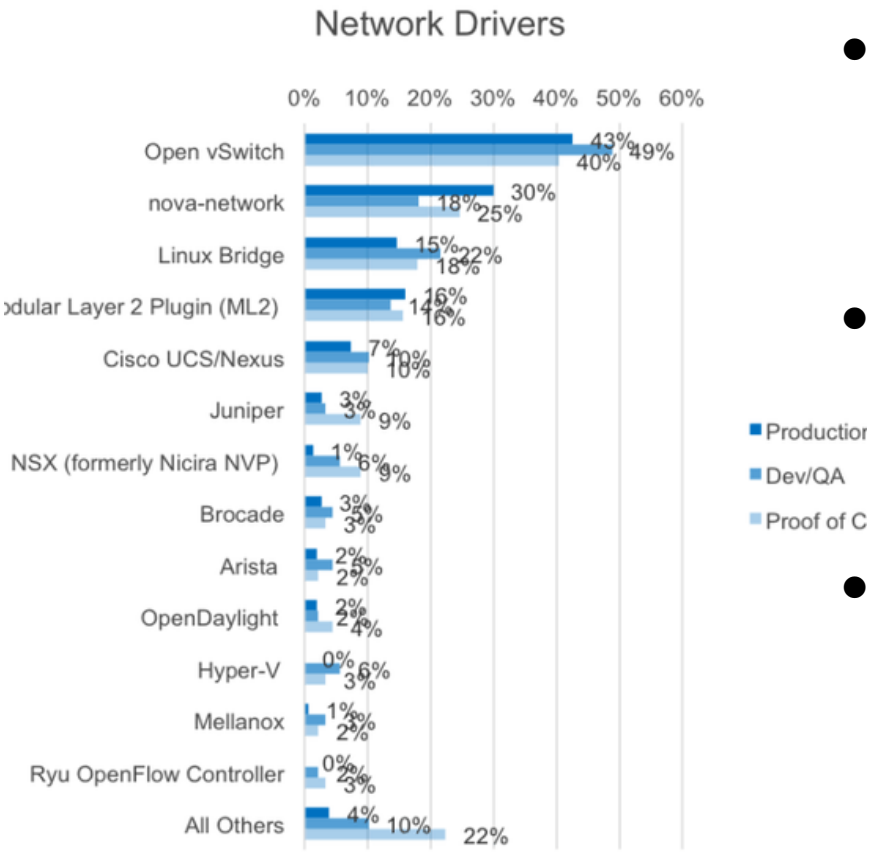
OpenStack云数据中心的SDN

云数据中心网络上最基本的技术要求就是可迁移性、隔离性和网段重叠



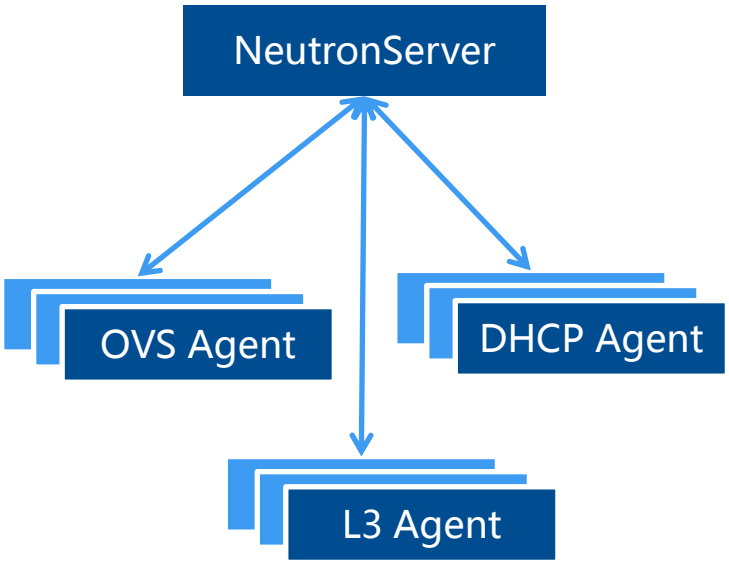
摘自：http://www.sdnlab.com/11926.html

OVN 简介和社区的发展



- OVN 是Nicria为OVS在2015早期宣布的一个子项目，实现了OVS一个简单高效的Controller，广义上OVN包括OVS、OVN南北向控制器、OpenStack Networking-OVN三部分。
- OVN是OVS社区目前以及将来的重点发展方向，在OpenStack的近期两次峰会（巴塞罗那、波士顿）的OpenvSwitch Day和Networking Day会场中，均有多个专场演讲。
- OVS的2.6是第一个支持OVN的发布版本， Newton是第一个支持 Networking-OVN的OpenStack发布版本。目前最新的OVN发布版本为OVS 2.7和OpenStack Ocata。

OpenStack中Neutron的不足



Agent过多，导致RPC成为性能瓶颈

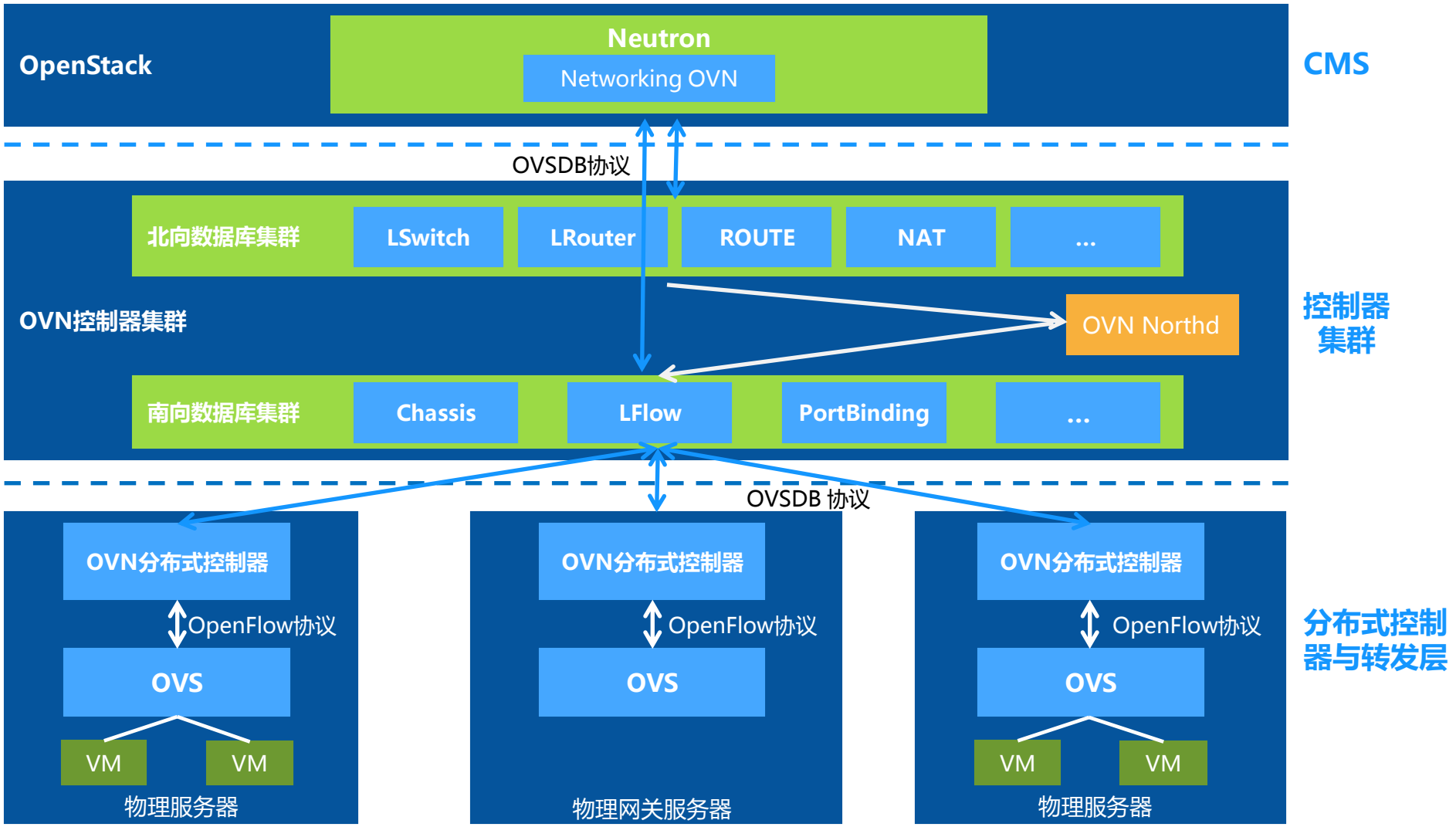
Namespace 资源有限而且系统开销比较大

OVN已经支持的功能

消除过多的Agent，不使用命名空间，全部通过流表实现

Logical_Switch	逻辑交换机，用于network和subnetwork
Logical_Router	逻辑路由器，用于router
Static Route	静态路由
DHCP	分布式DHCP
ACL	访问规则，用于Secgroup等。二到四层的 ACL，可以根据报文的 MAC 地址，IP 地址，端口号来做访问控制
Tunnel	有 Geneve，STT，VxLAN（仅Vtep支持），缺省Geneve；
NAT	集中式虚拟路由器支持SNAT和DNAT
LB	负载分担

OVN SDN基础架构


OpenStack
Neutron
Networking OVN
CMS
OVSDB协议
北向数据库集群
LSwitch
LRouter
ROUTE
NAT
...
OVN控制器集群
OVN Northd
控制器集群
南向数据库集群
Chassis
LFlow
PortBinding
...
OVSDB 协议
OVN分布式控制器
OpenFlow协议
OVS
VM
VM
物理服务器
OVN分布式控制器
OpenFlow协议
OVS
物理网关服务器
OVN分布式控制器
OpenFlow协议
OVS
VM
VM
物理服务器
分布式控制器与转发层

OVN Chassis

Hypervisors

VTEP Gateway

```
ovs-vsctl set Open_vSwitch . external-ids:ovn-remote=tcp:192.168.44.128:6642
ovs-vsctl set Open_vSwitch . external-ids:ovn-encap-type=geneve
ovs-vsctl set Open_vSwitch . external-ids:ovn-encap-ip=192.168.140.129
ovs-vsctl set Open_vSwitch . external-ids:ovn-bridge-mappings=physnet3:br-ex
```

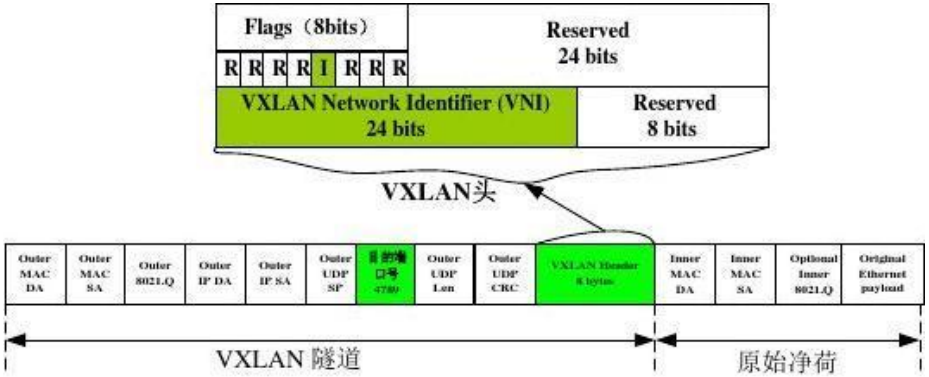
Chassis table

_uuid	encaps	external_ids	Hos tname name	vtep_logical_swithes
47eef299-406f-478b-98d4-a82c674444c8	[a9a89d8b-148b-4e55-8d67-12107ff50058]	{ovn-bridge-mappings= “physnet3:br-ex” }”	com1“	“b25bd30b-15e0-43e4-a537-3c4b53a78e6d” []
bcef2cfe-9704-473c-8fa2-998292df56bc	[d2dc1728-a81d-4d8a-b5d2-e09c96ea44c9]	{ovn-bridge-mappings= “physnet3:br-ex” }”	com2”	“ af44038a-8591-4889-b6ff-8f520b104266 ” []

OVN Tunnel

Geneve | STT | VxLAN

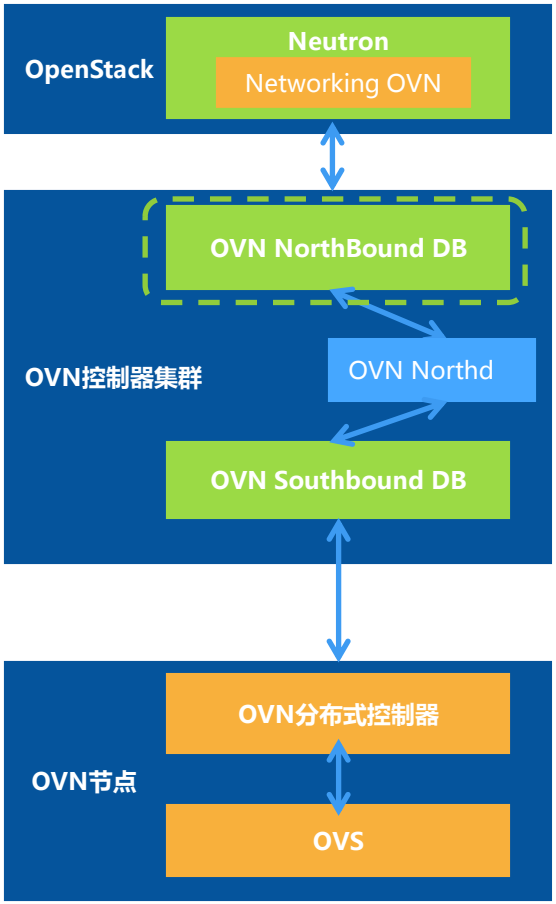
```
Outer UDP Header:
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |       Source Port = xxxx      |       Dest Port = 6081        |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |           UDP Length          |        UDP Checksum           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Geneve Header:
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |Ver|  Opt Len  |O|C|    Rsvd.  |          Protocol Type        |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |        Virtual Network Identifier (VNI)       |    Reserved   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                    Variable Length Options                    |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Inner Ethernet Header (example payload):
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                Inner Destination MAC Address                  |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | Inner Destination MAC Address |    Inner Source MAC Address   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                  Inner Source MAC Address                     |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |Optional Ethertype=C-Tag 802.1Q|  Inner VLAN Tag Information   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Payload:
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | Ethertype of Original Payload |                               |
```

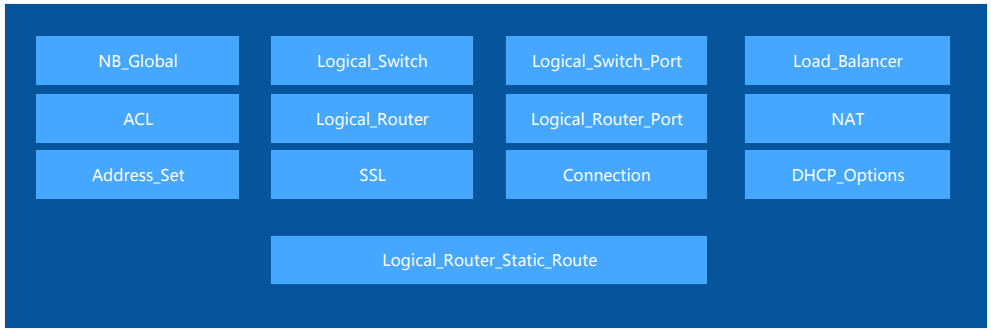


Logical input port identifier：逻辑的入端口标识符
Logical datapath identifier：逻辑的数据通道标识符
Logical output port identifier（逻辑的出端口标识符）

通过隧道传输携带的大量元数据，避免目的端再次解析报文字段，加速流表查找效率

OVN Northbound DB

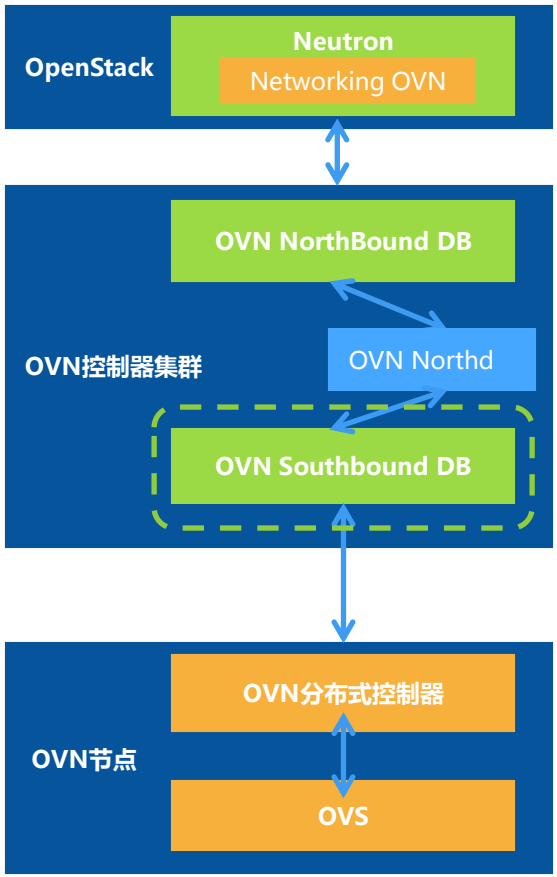


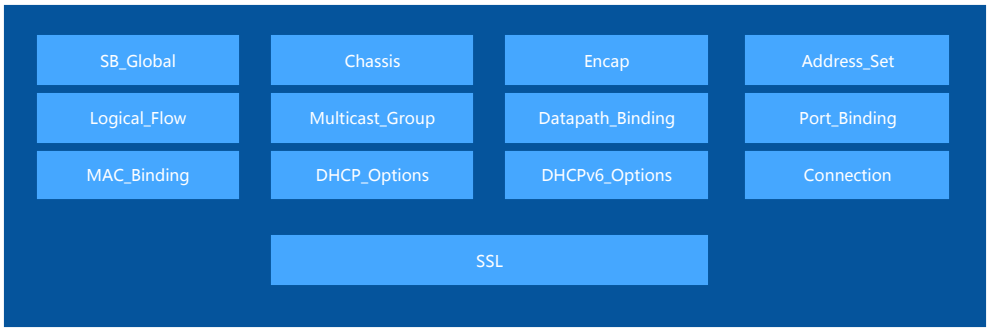


主要表说明

1. Logical_Switch：每一行代表一个逻辑交换机
2. Logical_Switch_Port：每一行代表一个逻辑端口
3. ACL：每一行代表一个应用到逻辑交换机上的 ACL 规则
4. Logical_Router：每一行代表一个逻辑路由器
5. Logical_Router_Port：每一行代表一个逻辑路由器端口
6. NAT：每行数据表示一行NAT表项，仅在集中式路由器生效

OVN Southbound DB

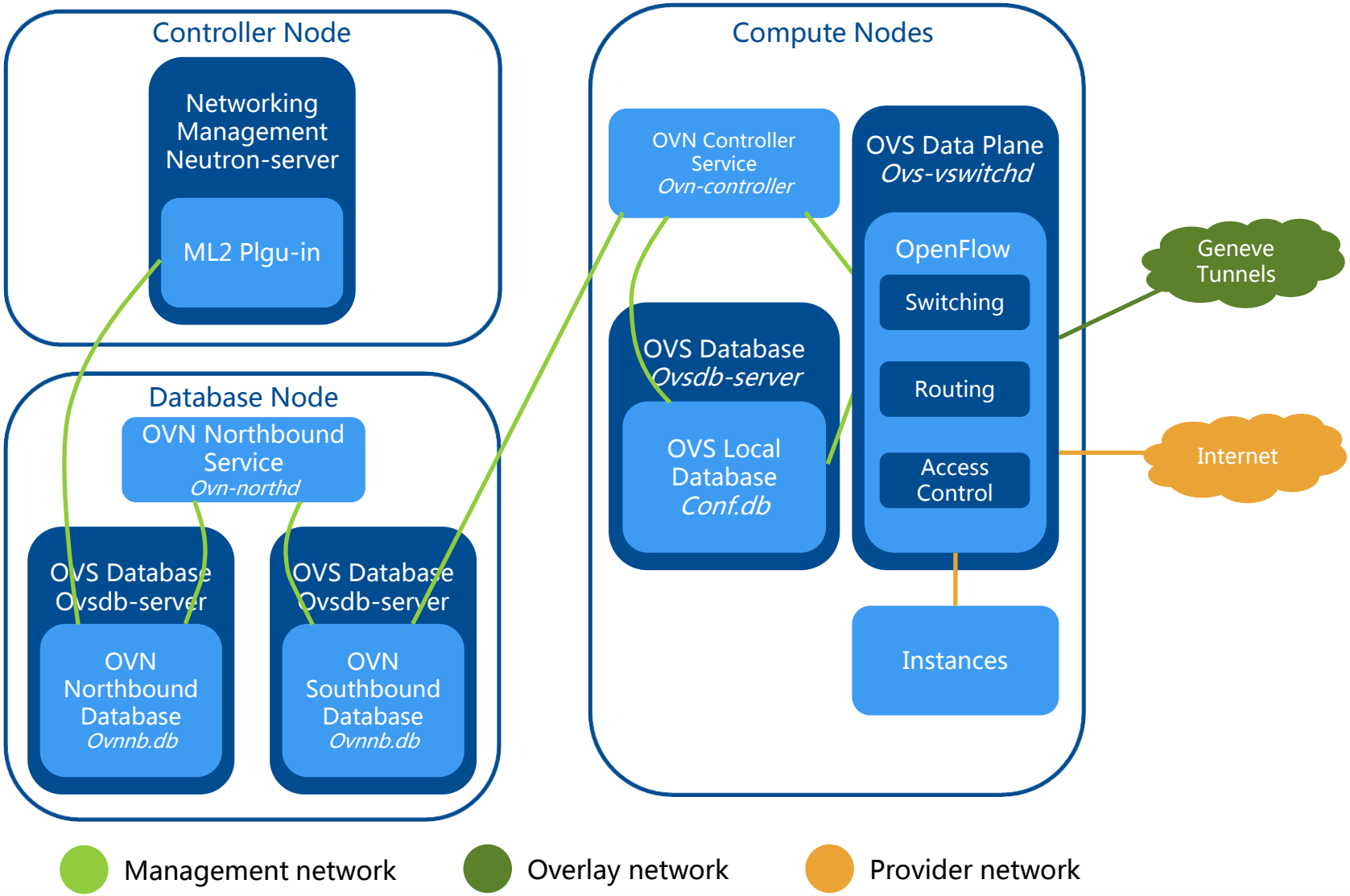




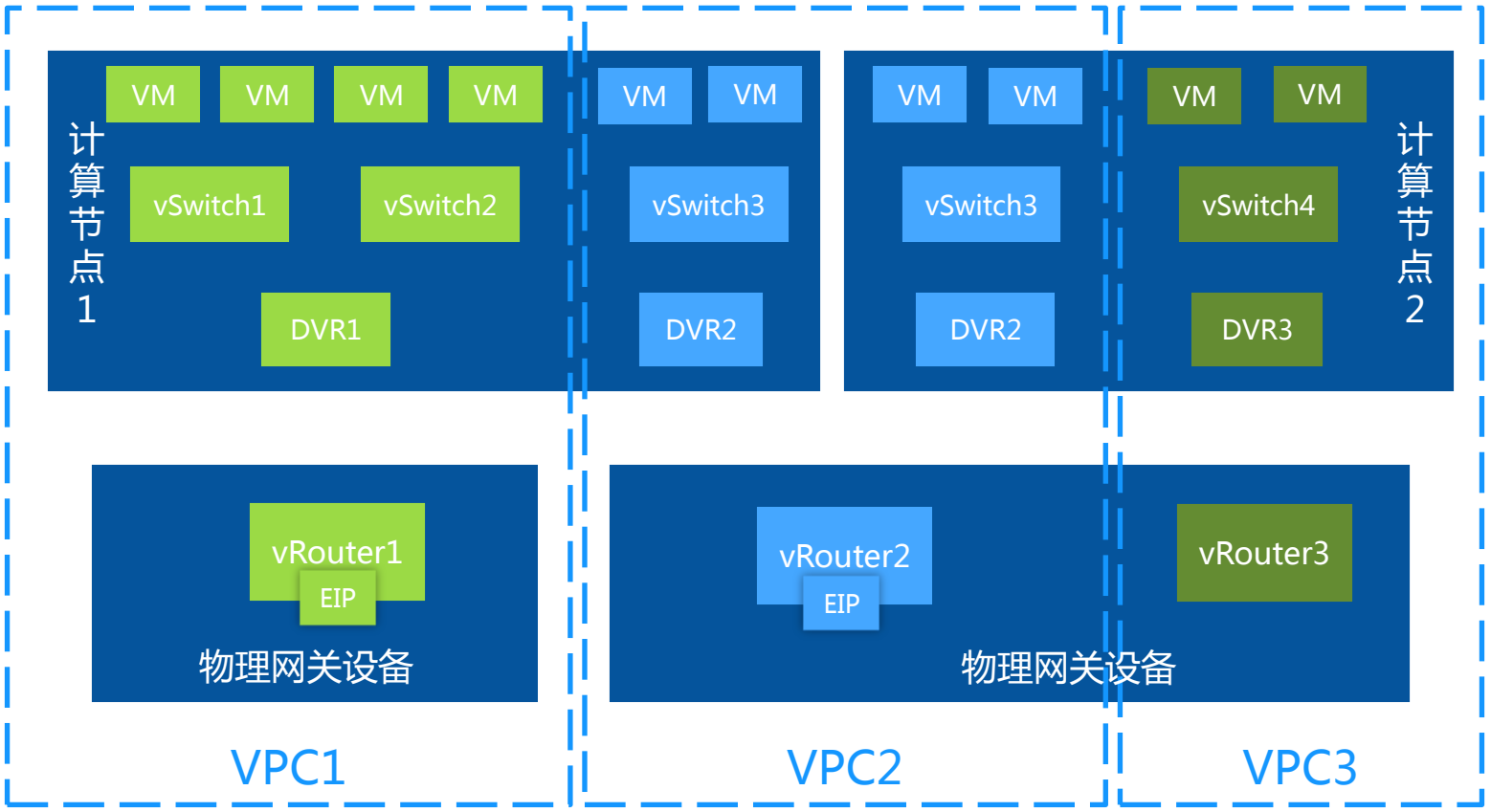
主要表说明

1. Chassis：每一行表示一个 HV 或者 VTEP 网关
2. Encap：保存着 tunnel 的类型和 tunnel endpoint IP 地址
3. Logical_Flow：每一行表示一个逻辑的流表
4. Multicast_Group：每一行代表一个组播组
5. Datapath_Binding：每一行代表一个 datapath 和物理网络的绑定关系
6. Port_Binding：这张表主要用来确定 logical port 位于哪个 chassis 上面

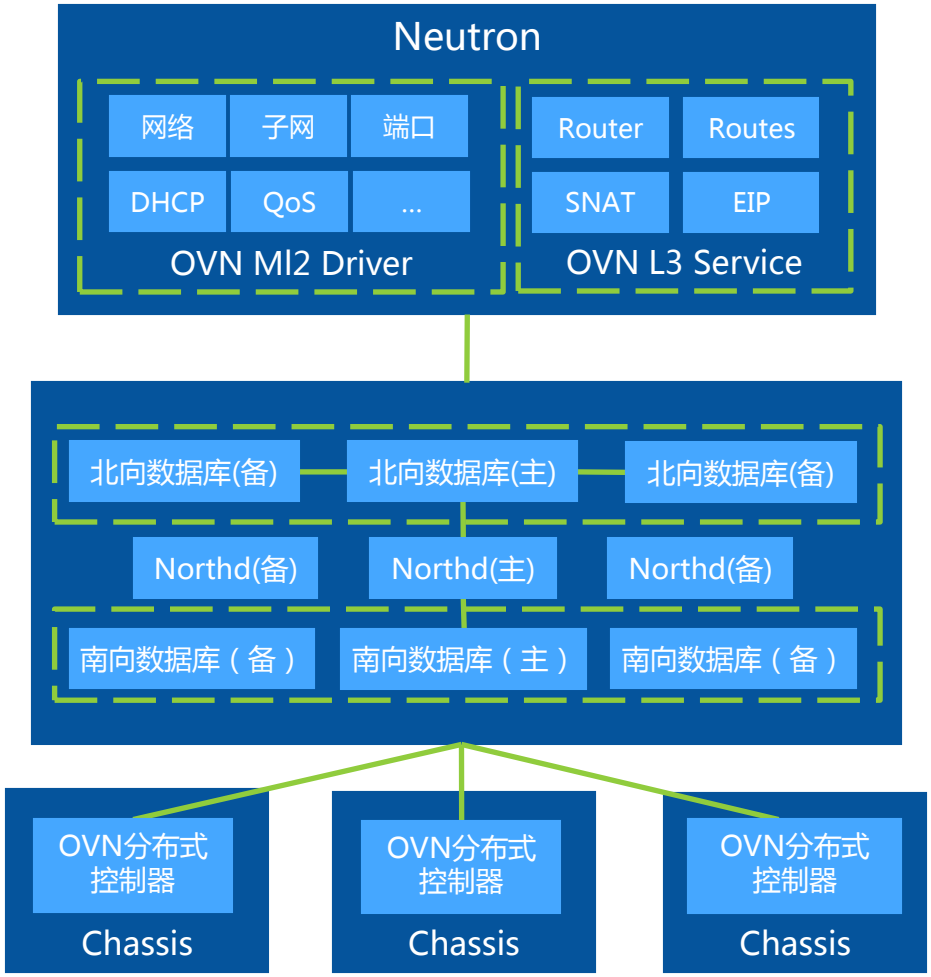
OVN与OpenStack网络集成



OVN SDN网络功能架构

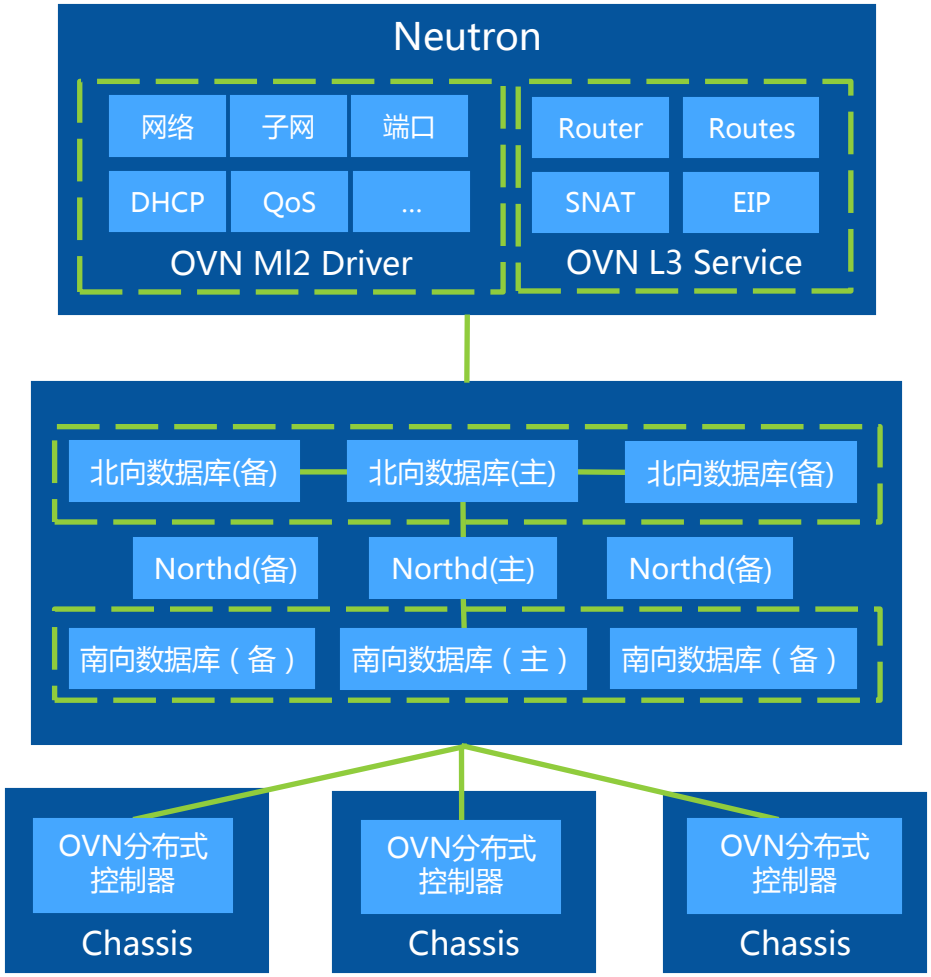


OVN 全分布式控制层



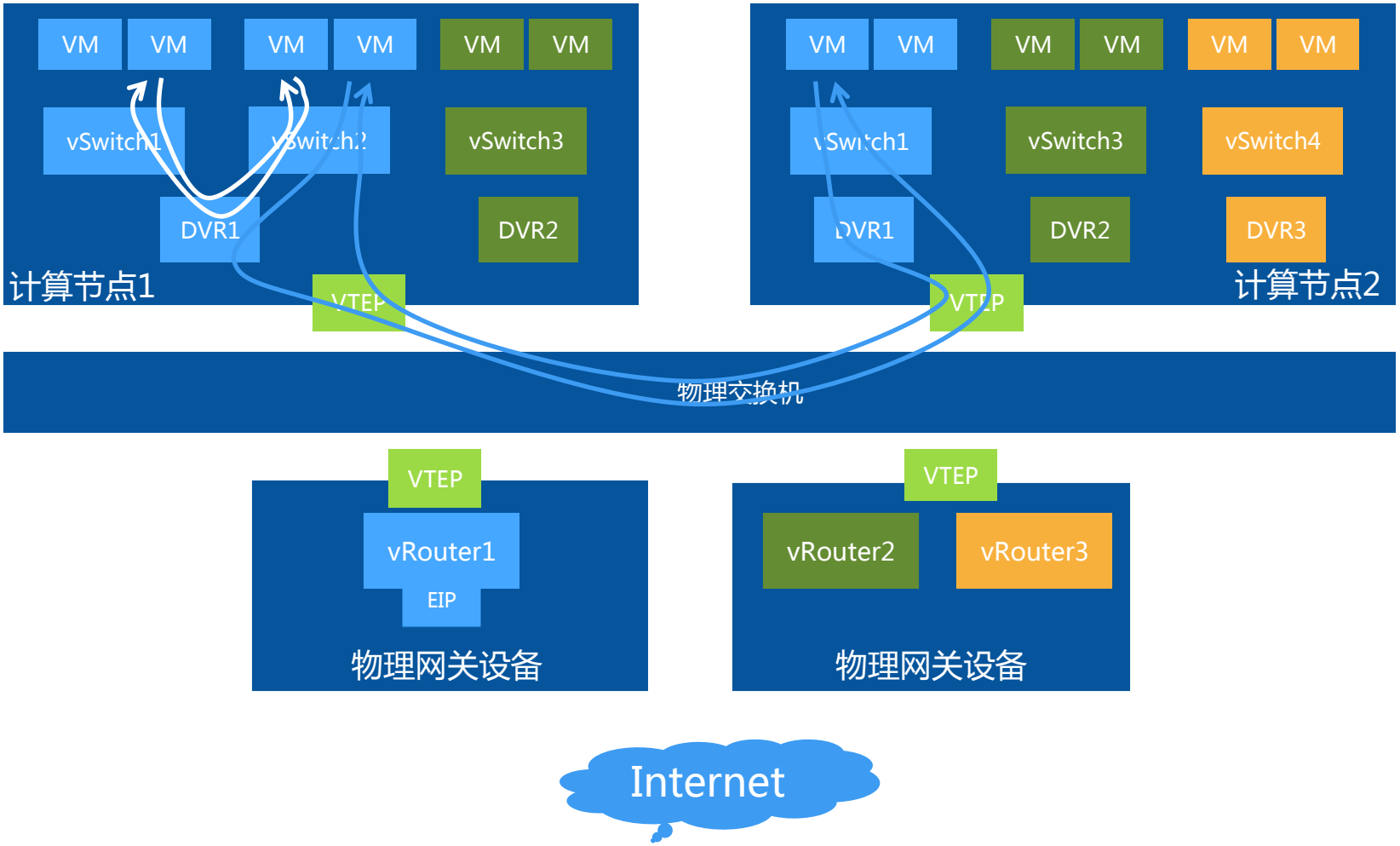
- 流表支持ARP、DHCP、Security Group等原生处理
- OVN Ml2 Driver和L3 Service向OVN北向数据库写入数据
- 北向守护进程将北向数据库翻译为南向数据库
- OVN分布式控制器将南向本Chassis表项翻译为流表下发到本机OpenFlow流表
- OVN控制连接为TCP连接，性能高于AMQP
- OVN 分布式控制器断开不影响转发功能

OVN高可用性

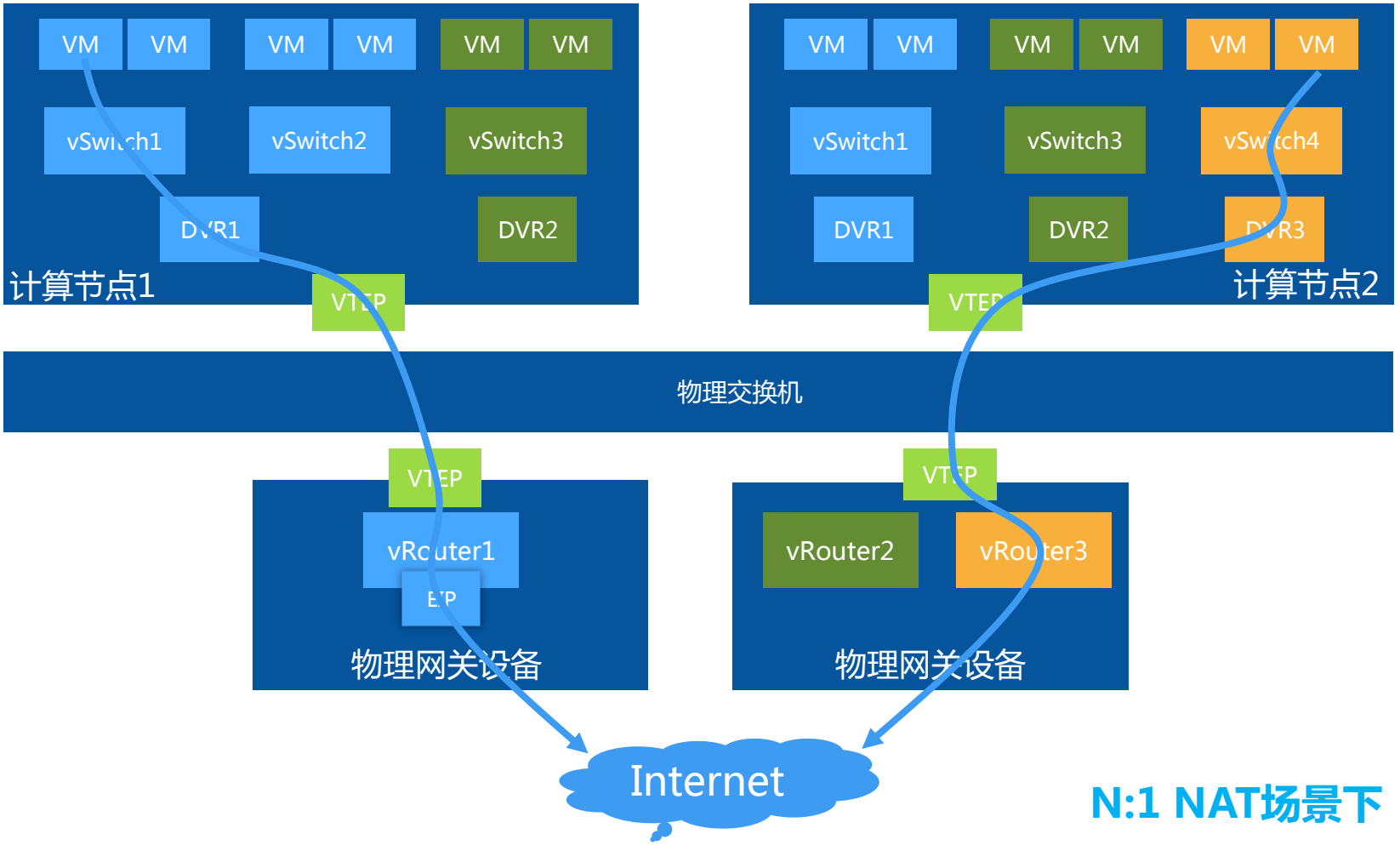


- Neutron多节点部署，北向使用SLB对外提供访问
- 北向数据库和南向数据库使用OVSDB的Active/Passive HA功能
- Northd进程在多个节点部署，使用Pacemaker进行Active/Passive HA
- OVN Controller部署于每个计算节点

OVN 支持东西向三层转发卸载

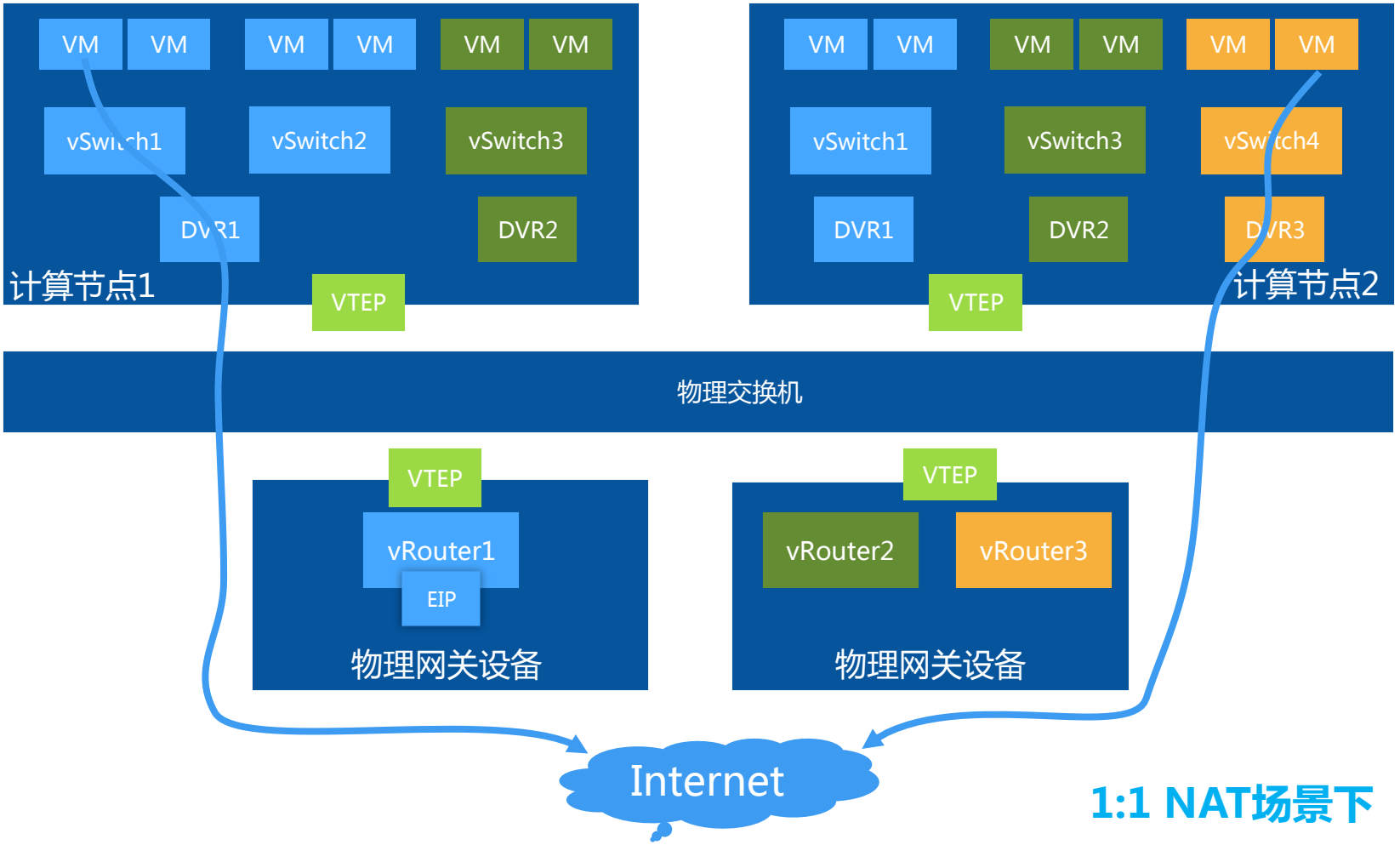


OVN南北向集中式网关



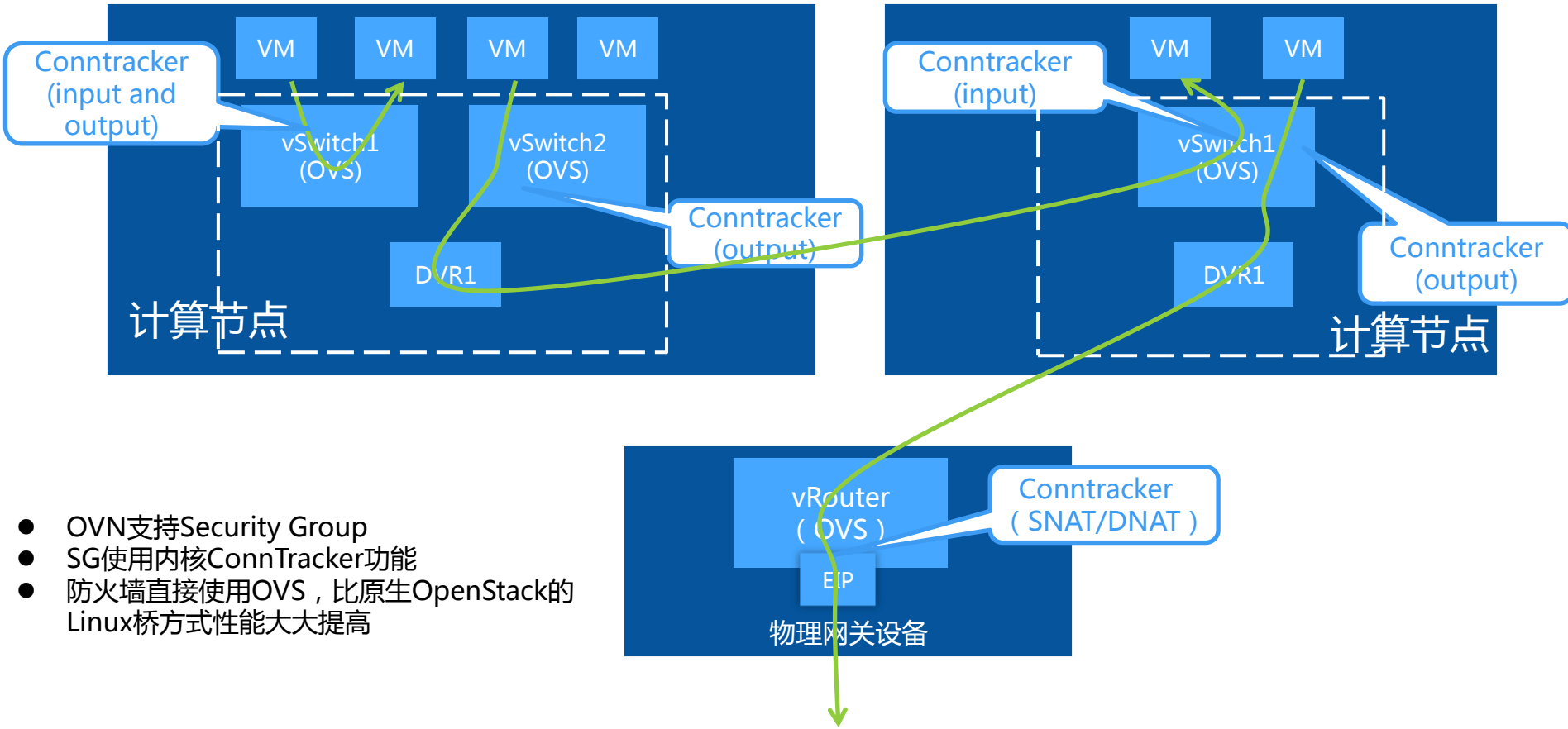
N:1 NAT场景下

OVN南北向分布式网关



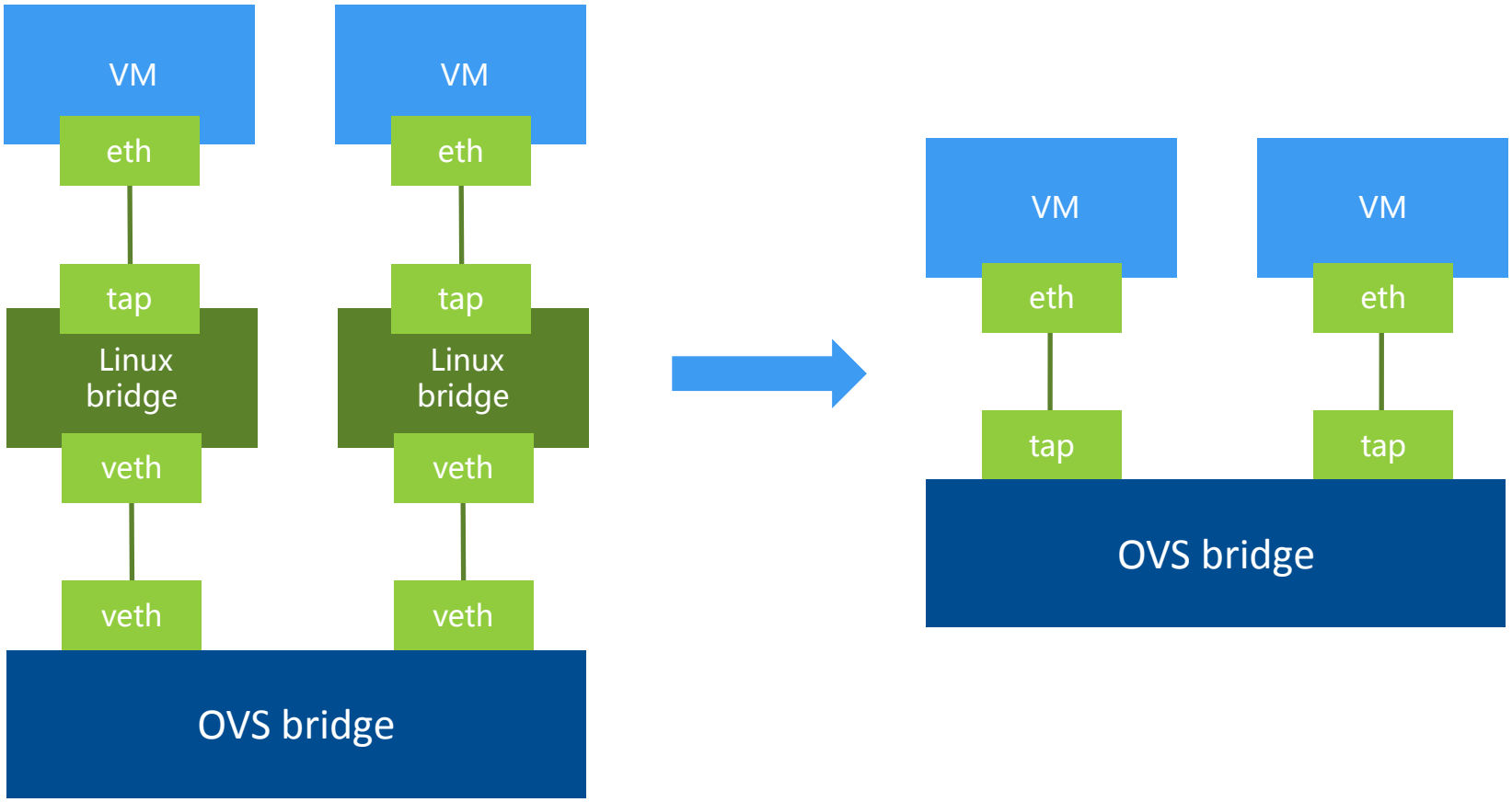
1:1 NAT场景下

OVN 基于连接状态的分布式防火墙

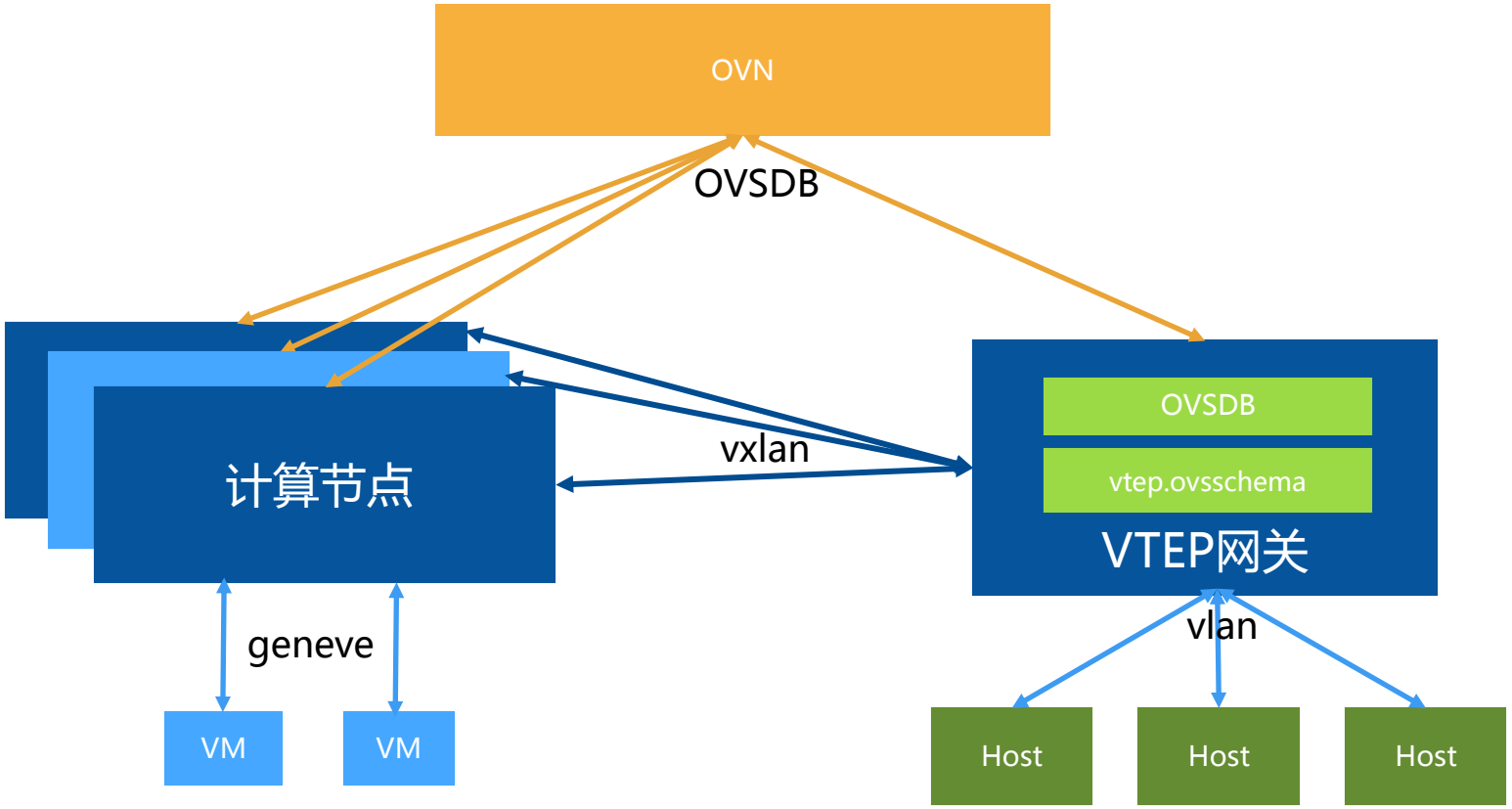


- OVN支持Security Group
- SG使用内核ConnTracker功能
- 防火墙直接使用OVS，比原生OpenStack的Linux桥方式性能大大提高

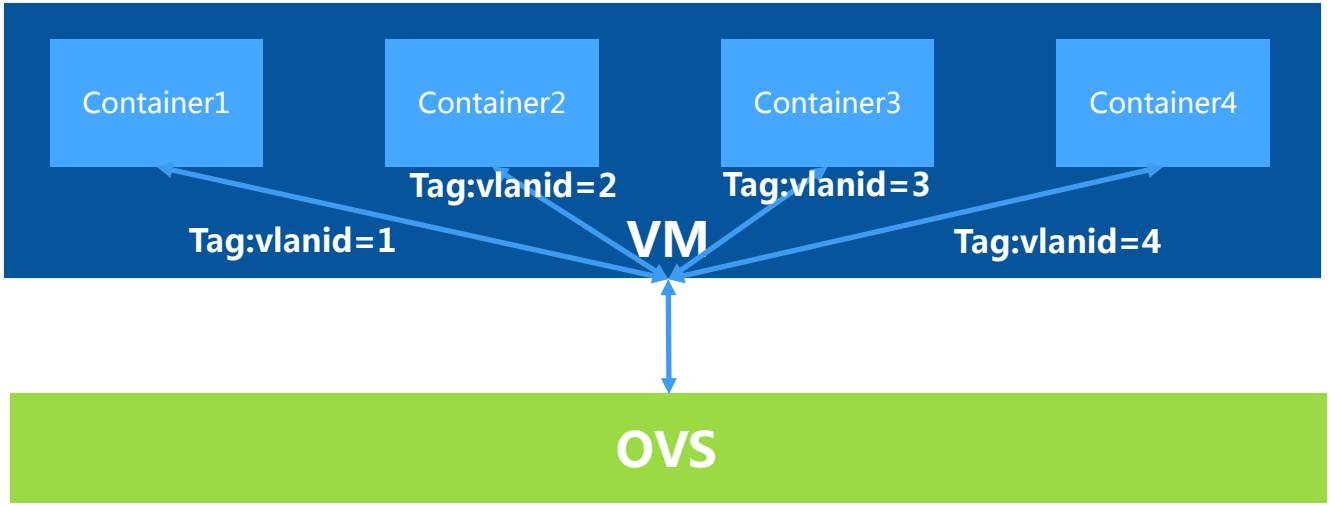
OVN对OpenStack端口安全组实现的影响



OVN对VTEP L2GW的支持



OVN对容器的支持



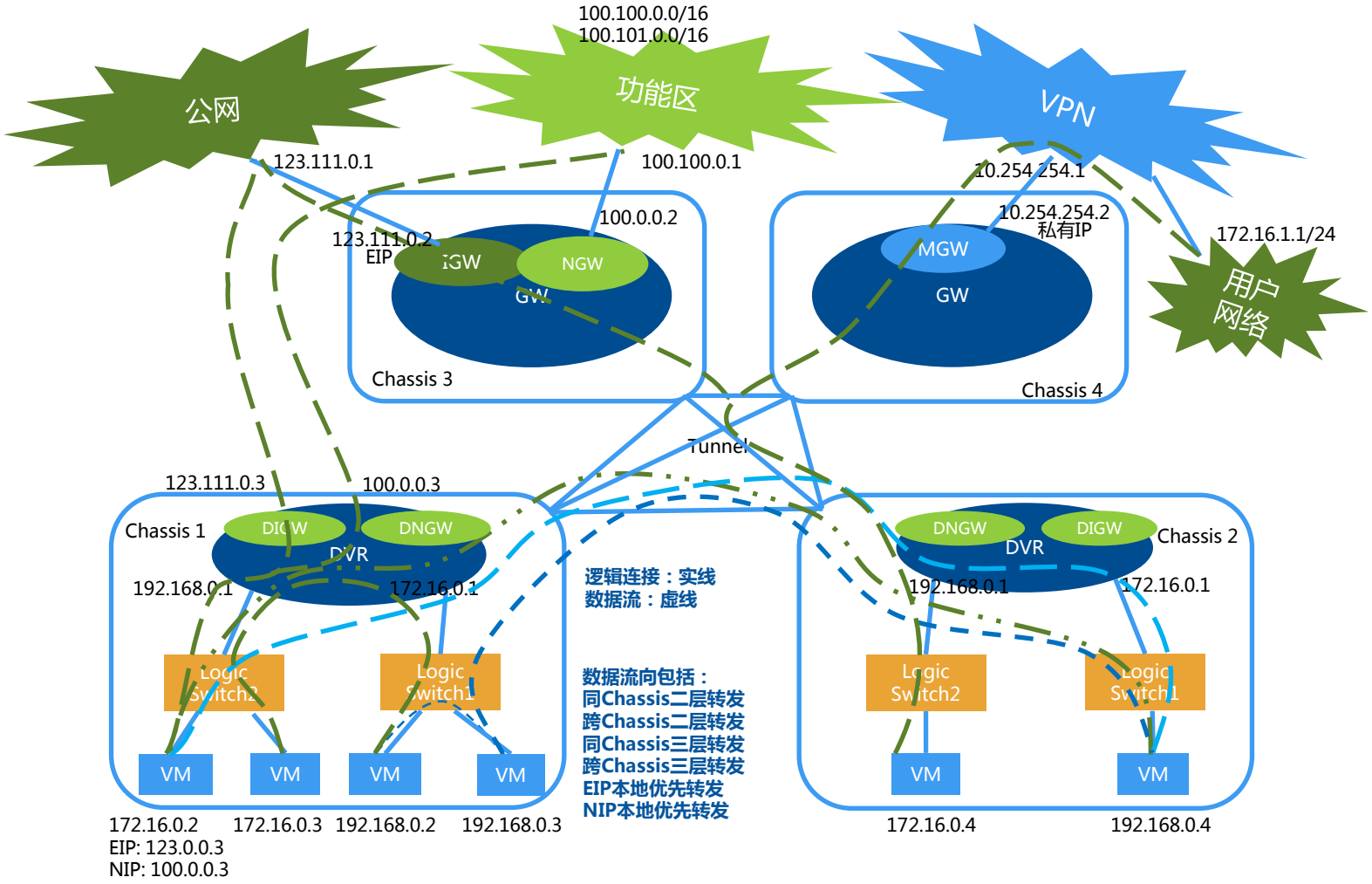

```
$ neutron port-create private
Created a new port:
+-----------------------+--------------------------------------------------------------------------------------+
| Field                 | Value                                                                                |
+-----------------------+--------------------------------------------------------------------------------------+
| admin_state_up        | True                                                                                 |
| allowed_address_pairs |                                                                                      |
| binding:vnic_type     | normal                                                                               |
| device_id             |                                                                                      |
| device_owner          |                                                                                      |
| fixed_ips             | {"subnet_id": "ce5e0d61-10a1-44be-b917-f628616d686a", "ip_address": "10.0.0.3"}      |
| id                    | 74e43404-f3c2-4f13-aeec-934db4e2de35                                                 |
| mac_address           | fa:16:3e:c5:a9:74                                                                    |
| name                  |                                                                                      |
| network_id            | f654265f-baa6-4351-9d76-b5693521c521                                                 |
| security_groups       | fe25592f-3610-48b9-a114-4ec834c52349                                                 |
| status                | DOWN                                                                                 |
| tenant_id             | db75dd6671ef4858a7fed450f1f8e995                                                     |
+-----------------------+--------------------------------------------------------------------------------------+
```

```
$ neutron port-create --binding-profile '{"parent_name":"74e43404-f3c2-4f13-aeec-934db4e2de35","tag":42}' private
Created a new port:
+-----------------------+--------------------------------------------------------------------------------------+
| Field                 | Value                                                                                |
+-----------------------+--------------------------------------------------------------------------------------+
| admin_state_up        | True                                                                                 |
| allowed_address_pairs |                                                                                      |
| binding:vnic_type     | normal                                                                               |
| device_id             |                                                                                      |
| device_owner          |                                                                                      |
| fixed_ips             | {"subnet_id": "ce5e0d61-10a1-44be-b917-f628616d686a", "ip_address": "10.0.0.4"}      |
| id                    | be155d07-ecd9-4ad7-91e5-5be60684572a                                                 |
| mac_address           | fa:16:3e:74:ef:82                                                                    |
| name                  |                                                                                      |
| network_id            | f654265f-baa6-4351-9d76-b5693521c521                                                 |
| security_groups       | fe25592f-3610-48b9-a114-4ec834c52349                                                 |
| status                | DOWN                                                                                 |
| tenant_id             | db75dd6671ef4858a7fed450f1f8e995                                                     |
+-----------------------+--------------------------------------------------------------------------------------+
```

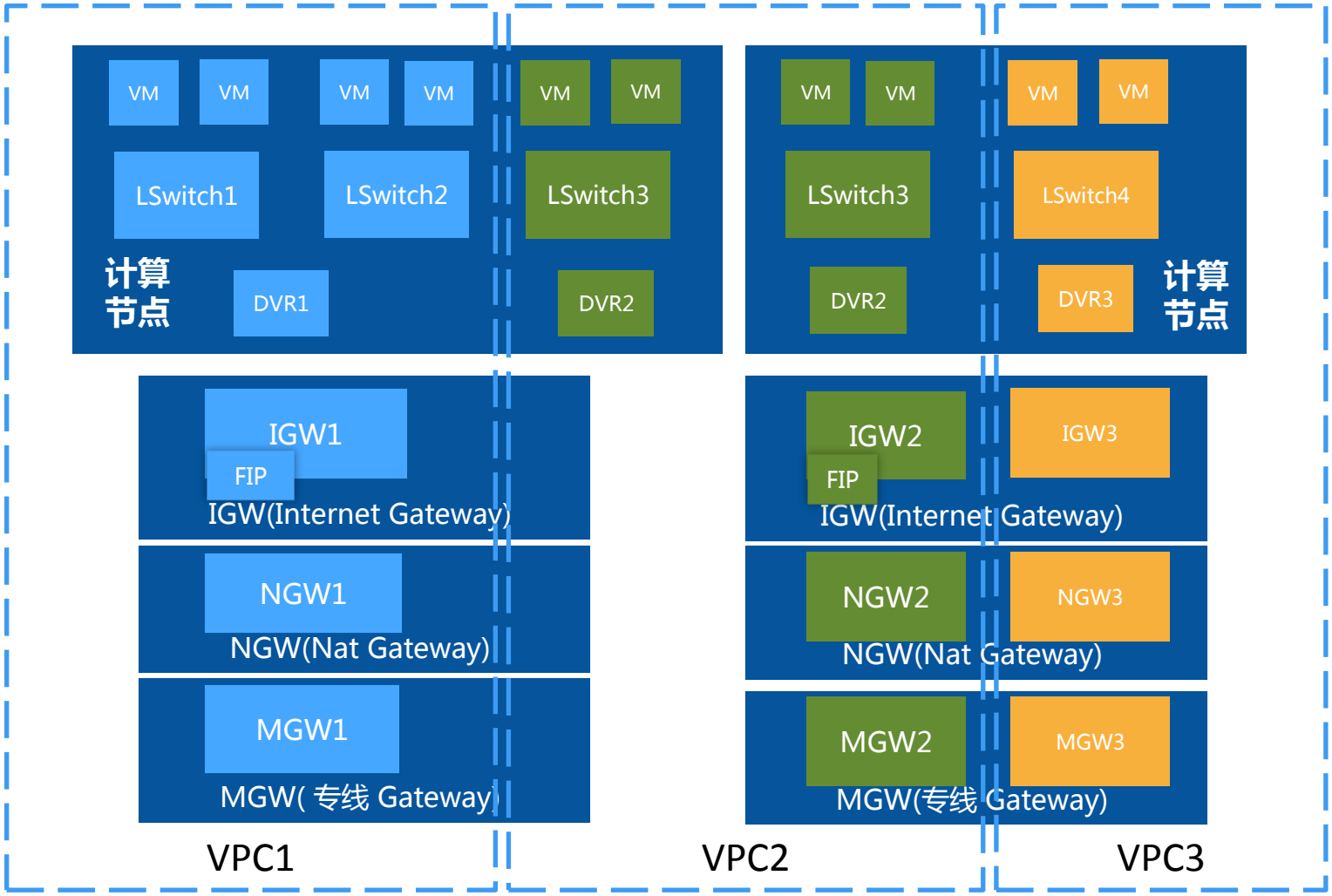
数梦工场OVN领先优势

资格优势	数梦为Openstack Networking-ovn（贡献30%）和OVN社区贡献大量功能和代码，被networking-ovn邀请为社区Core成员
社区贡献	1、分布式网关调度； 2、SSL连接； 3、DHCP功能完善； 4、70+patch； 5、CI维护
独特功能	1、支持多网关（IGW、NGW、MGW）； 2、支持Floating IP/Nat IP本地优先转发（IGW、NGW）； 3、支持隧道、VLAN的混合型组网； 4、扩展网关调度（基于Chassis Type、基于物理网络连接）； 5、支持OVN数据库多主模式；

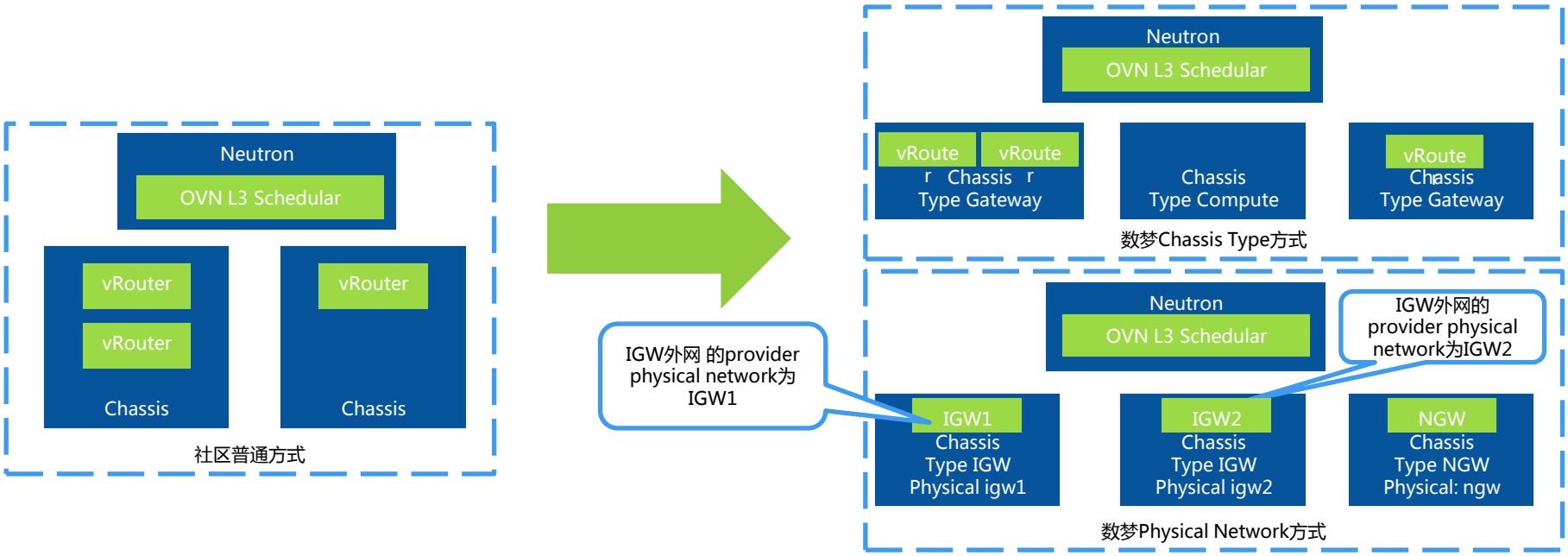
OVN SDN 数梦多网关混合网络



OVN SDN 数梦多网关混合网络架构



OVN SDN 数梦vRouter灵活调度



社区支持

1、支持所有Compute节点调度Gateway Router，此种情况所有Compute节点均存在外部连接；

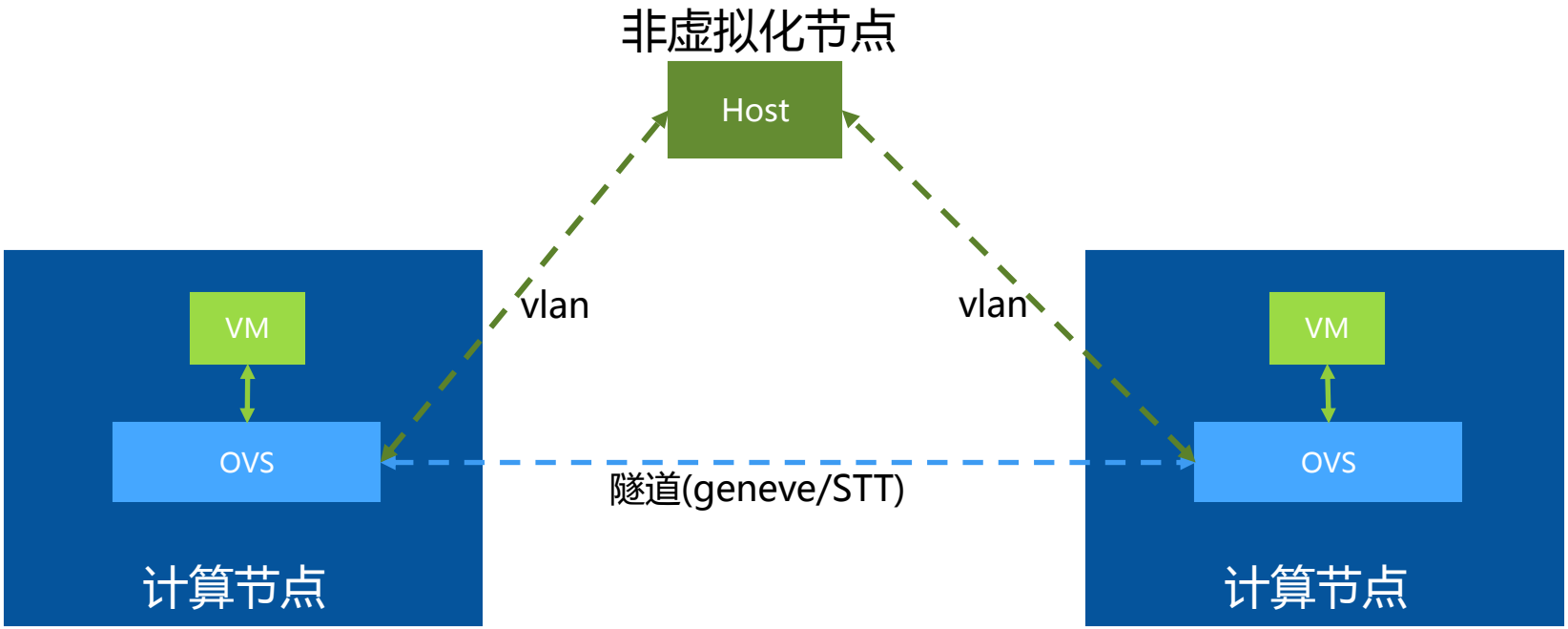
2、支持随机和最小连接方式调度Gateway Router；

数梦工场扩展支持

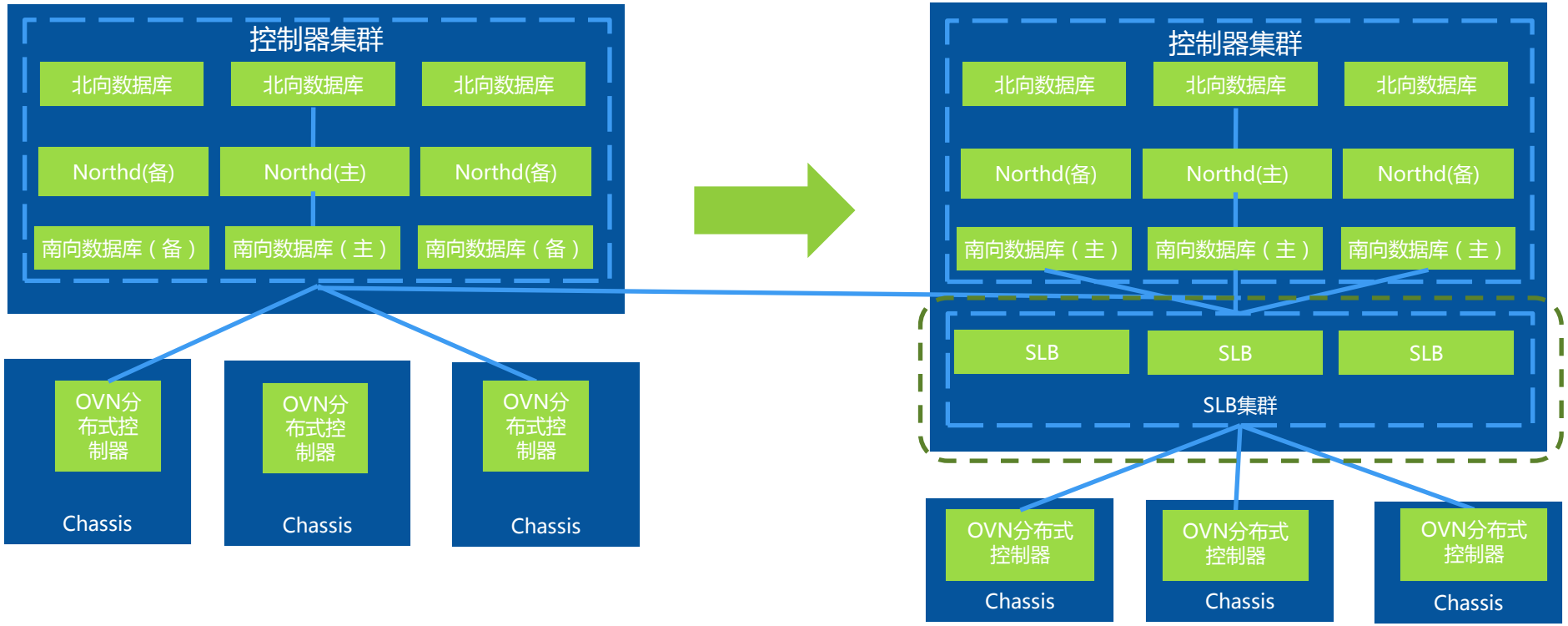
1、支持指定Chassis Type调度，满足vRouter仅在网关节点调度，提供更好的性能；

2、支持Physical Network调用；满足vRouter在多VLAN外部网络差异化调度

数梦支持隧道与VLAN的混合型组网



OVN SDN 数梦支持大规模集群部署



社区支持方式

1、北向、南向数据库单主多备；

2、OVN分布式控制器的最大压力在于连接南向数据库，获取数据库数据翻译流表。

数梦扩展支持

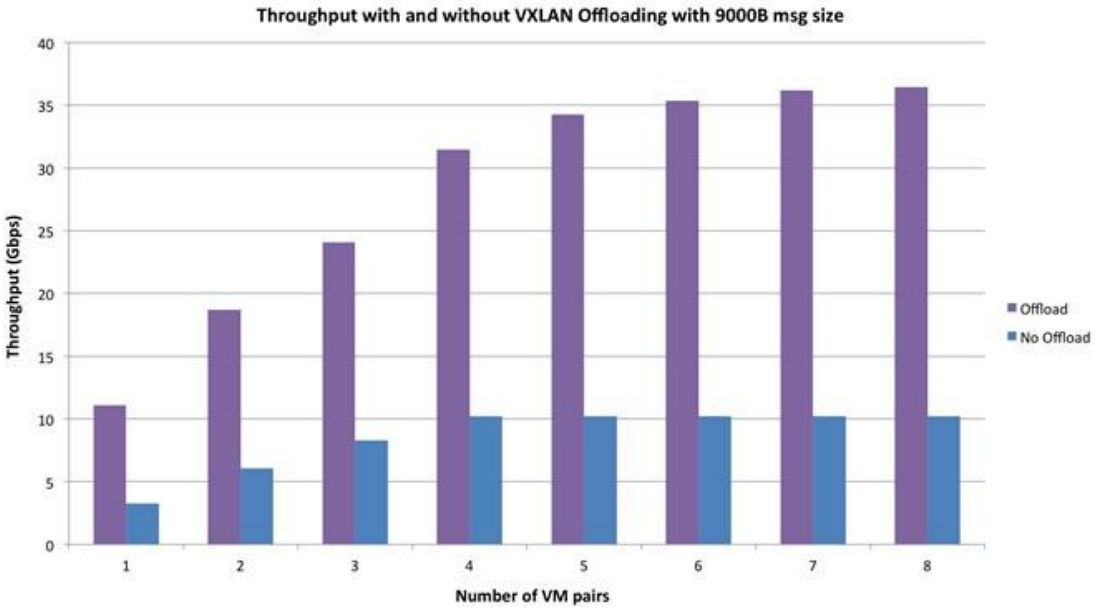
1、南向数据库读写分离（数据库只读）；

2、南向数据库SLB负载分担多主模式；

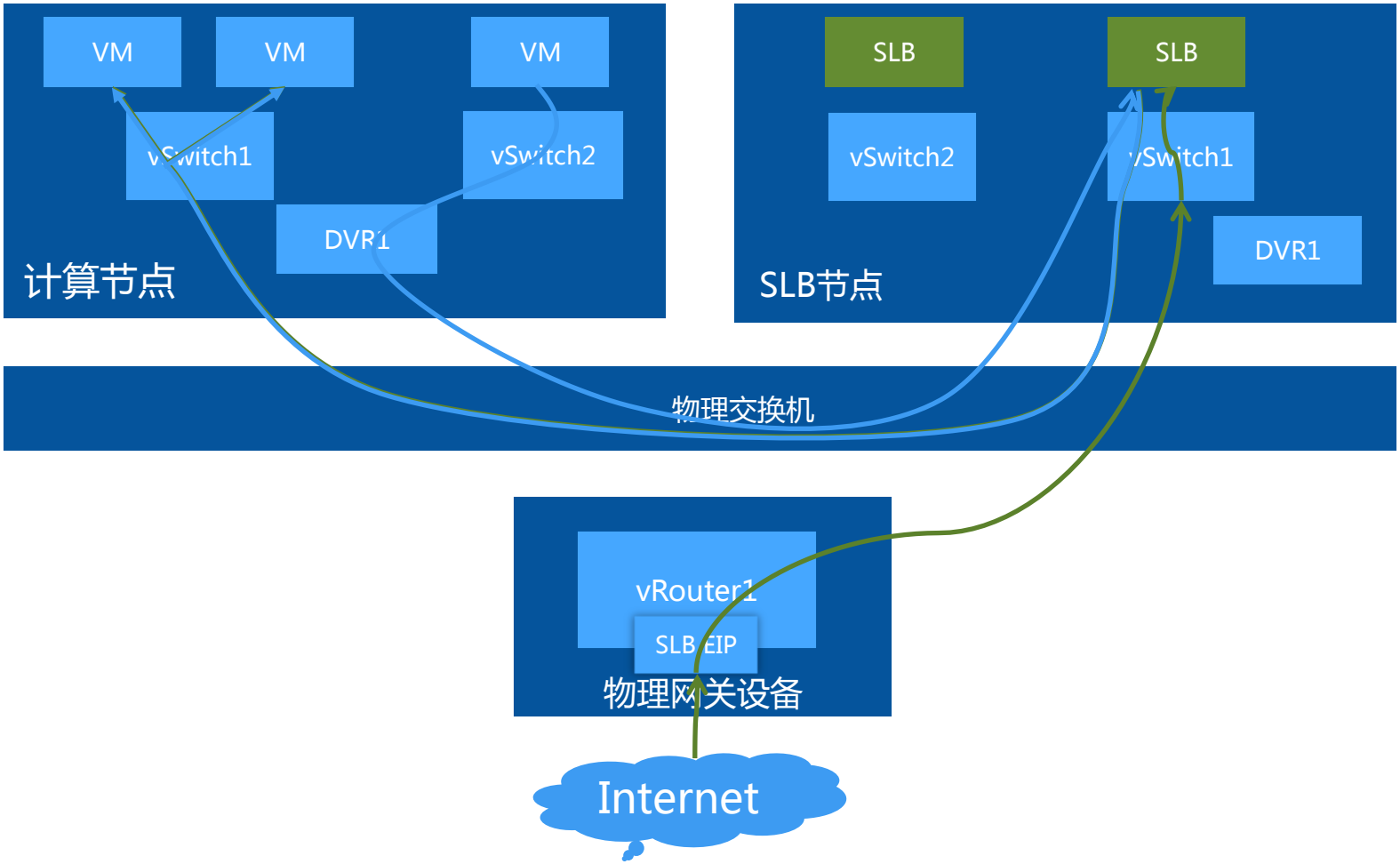
OVN SDN数梦支持网卡隧道Offload

支持Offload时，性能与TCP相近

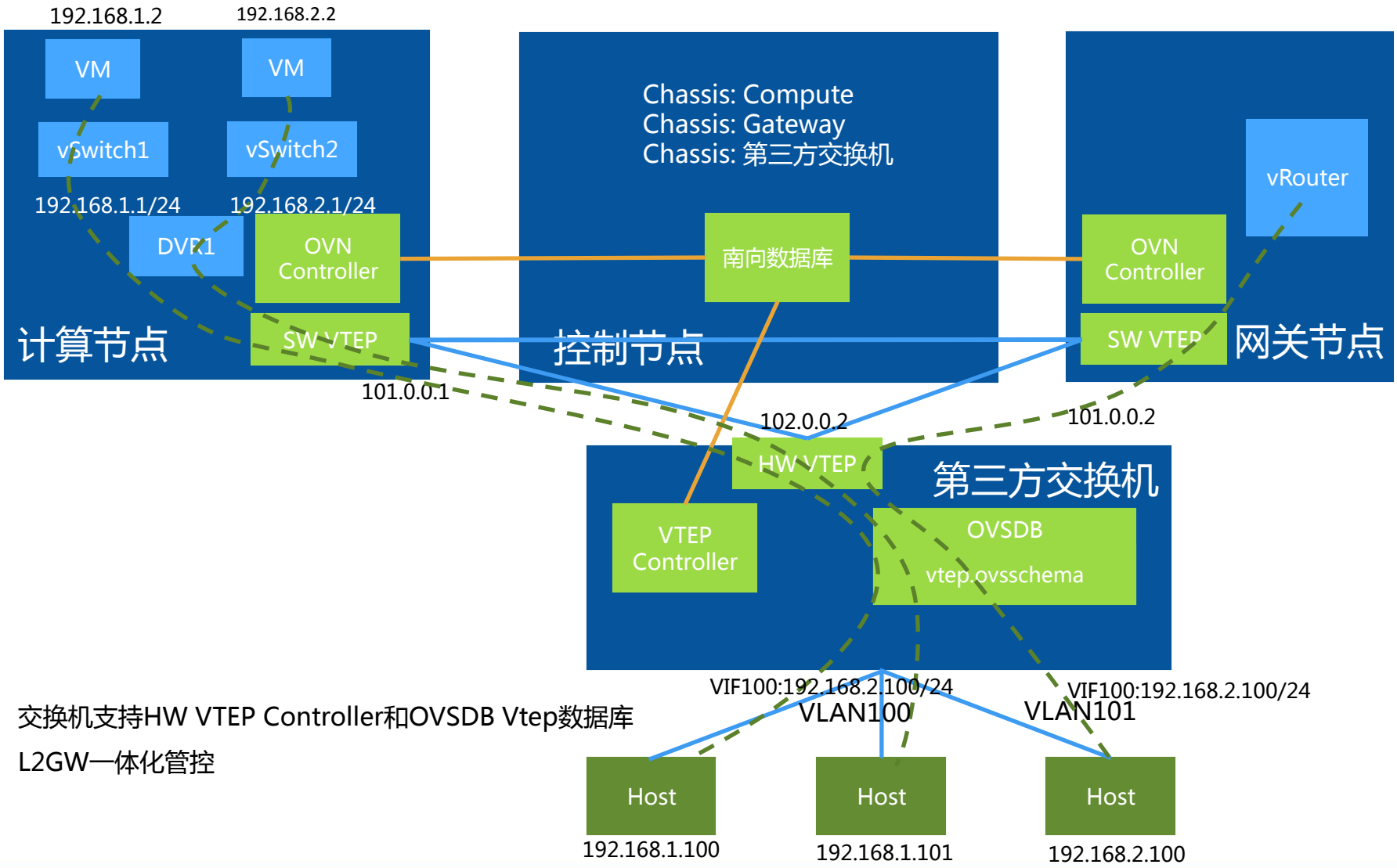
封装	offload	Gbps	Cpu(C)	Cpu(S)
tcp	on	9.41	20%	30%
vxlan	on	9.11	20%	35%
geneve	on	9.08	20%	30%



OVN SDN生态合作（第三方软件SLB）

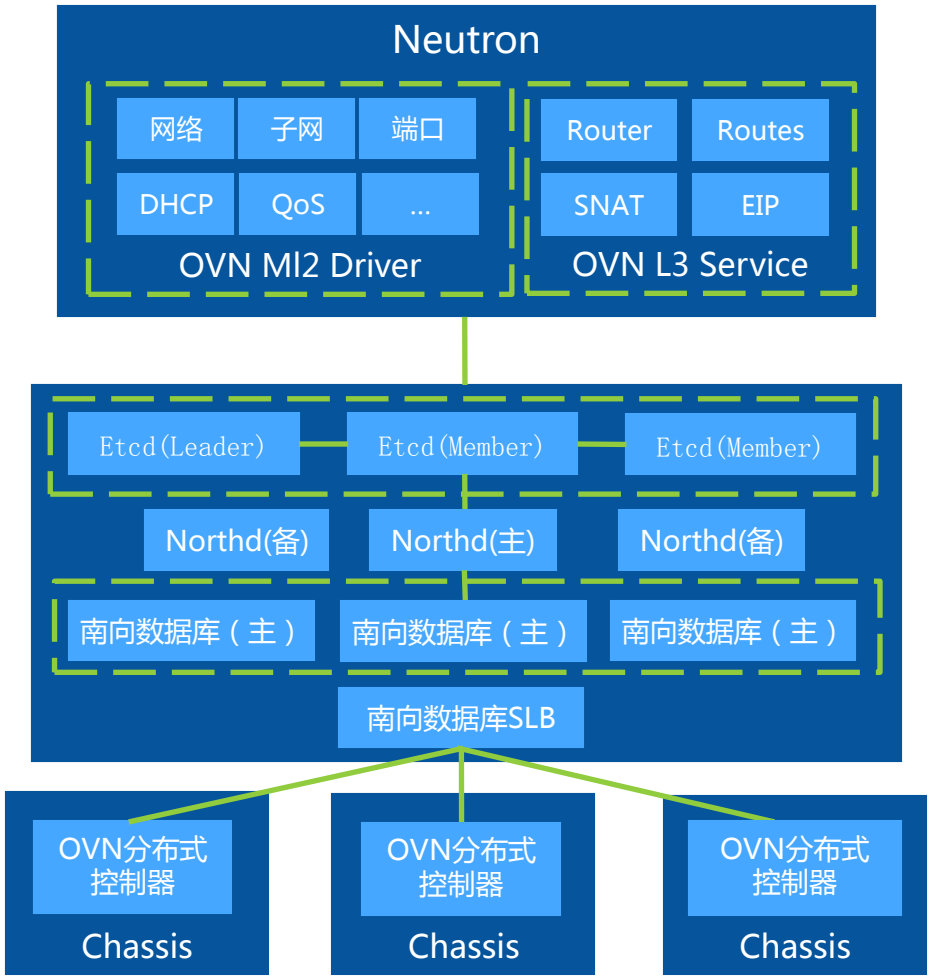


OVN SDN生态合作（第三方物理交换机）



- 交换机支持HW VTEP Controller和OVSDB Vtep数据库
- L2GW一体化管控

数梦工场正在参与的社区计划



1、提供高性能数据库，提高集群性能

备注：支持OVSDB Raft数据库（Etcd）

2、支持Metadata；

3、支持原生DNS Server；

THANK YOU