数据中心流量调优项目

2016-8-10

作者：朱林峰

目录

数据中心流量调优

1 vxlan 介绍

现在数据中心都会通过虚拟机的方式为租户提供云计算服务，虚拟机能方便的对计算资源和存储资源按序分配，从而极大地发挥数据中心的效率，同时虚拟机本身的隔离特性，也为租户提供足够的安全性。但基于虚拟机的云计算，也对数据中心的网络造成不少挑战。

1.1 云计算数据中心对网络的挑战

1. 虚拟机迁移范围受到网网络架构限制

由于虚拟机迁移的网络属性要求，其从一个物理机上迁移到另一个物理机上，要求虚拟机不间断业务，则需要其IP地址、MAC地址等参数维保持不变，如此则要求业务网网络是一个二层网络(局域网)，且要求网络本身具备多路径链路的冗余和可靠性。传统的网络生成树 (STPSpaning Tree Protocol)技术不仅部署繁琐荣，且协议复杂，网络规模不宜过大，限制了虚拟化的网络扩展性。

2. 虚拟机规模受网络规格限制

在大二层网络环境下，数据流均需要通过明确的网络寻址以保证准确到达目的地，因此网络设备的二层地址表项大小(即MAC地址表)，成为决定了云计算环境下虚拟机的规模的上限，并且因为表项并非百分之百的有效性，使得可用的虚机数量进一步降低，特别是对于低成本的接入设备而言，因其表项一般规格较小,限制了整个云计算数据中心的虚拟机数量，但如果其地址表项设计为与核心或网关设备在同一档次，则会提升网络建设成本。虽然核心或网关设备MAC与ARP规格会随着虚拟机增长也面临挑战，但对于此层次设备能力而言，大规格是不可避免的业务支撑要求。减小接入设备规格压力的做法可以是分离网关能力，如采用多个网关来分担虚机的终结和承载，但如此也会带来成本的上升。

3. 网络隔离能力限制

当前的主流网络隔离技术为VLAN(或VPN)，在大规模虚拟化环境部署会有所限制：VLAN数量在标准定义中只有12个比特单位，即可用的数量为4000个左右，这样的数量级对于公有云或大型虚拟化云计算应用而言微不足道，其网络隔离与分离要求轻而易举会突破4000。

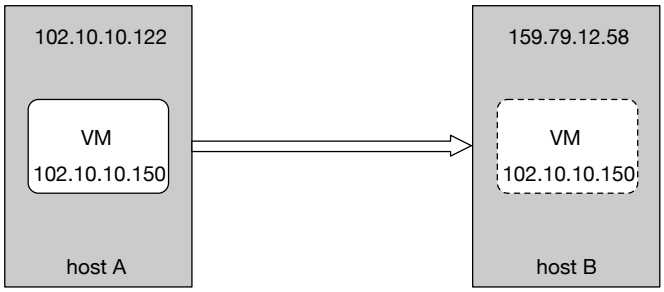
数据中心一般要支持多租户，每个租户都有自己独立的地址空间，如 MAC 地址，可能两个租户的虚拟机会使用同样的 MAC 地址，由此会导致物理网络上的地址冲突。

由此引发对 Overlay 网络的需求，在实际二层或三层网络里面携带虚拟网络的二层包。有多种方式实现 Overlay，而 VXLAN 无疑是最成熟的。

4. 总结

云计算数据中心面对的挑战总结如下：

1. 因为要求虚拟机的迁移保证不中断虚拟机的业务，因此虚拟机的 ip 地址和 MAC 地址必须不变，如果虚拟机 V 是从主机 A 迁移到主机 B 上，则A 、B 和 V 都必须是在一个局域网内，这是很难满足的。因为数据中心往往很庞大，甚至不在同一个地域，运行虚拟机的主机 A 和 B 往往不能保证是在同一个局域网内。

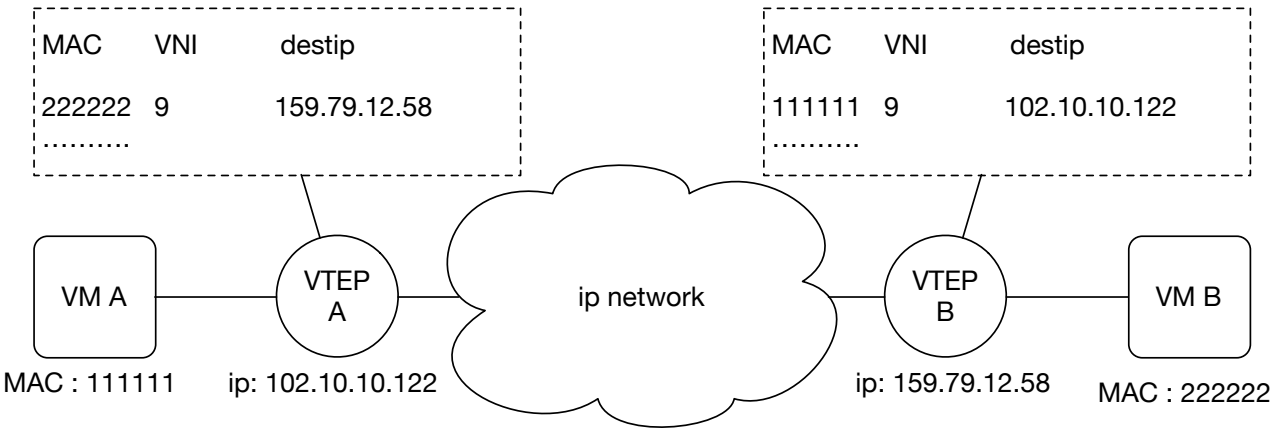


图中，虚拟机 VM 运行在主机 host A 上，并且其 ip 和主机是一个网段的，
这样虚拟机能被外界访问到，就像 host A 被外界访问一样。
现在 VM 需要迁移到主机 host B 上，而 host B 的 ip 和 vm 的 ip 却不是一个网段的，
外界将无法访问到 VM。
因此要求 host A 和 host B 的 ip 是同一个网段的。

2. 虚拟机和其宿主主机在同一个局域网内，而我们知道云计算中虚拟机的数量是相当庞大的，这就会造成一个相当庞大的局域网，这样庞大的局域网会对交换机造成巨大的压力。
3. 一个租户可能会租用一批虚拟机 M，另一个租户也会租用 一批虚拟机 N，那就要求 M 和 N 的虚拟机之间不能直接通信，因此需要进行局域网的隔离。传统的 VLAN 因为隔离能力有限而无法满足。

1.2 VXLAN

VXLAN 的思想是用三层物理网络的 IP 包承载虚拟机之间的链路包，将虚拟机之间通信的数据由三层物理网络传递，而虚拟机不感知。具体流程如下：



1. 虚拟机 A 发送数据包

   VM A 给 VM B 发送数据包，数据包格式为 M = [链路头][源 MAC 111111][目的 MAC 222222][数据][链路尾]

2. 数据包进入三层网络前被封装

   数据包进入三层网路前经由 VTEP A 封装。VETP A 收到数据包 M，根据数据包 M 的目的 MAC 查表，发现目的 MAC 222222 属于 VNI=9 的虚拟局域网，且对端 VETP 的 IP 地址为 159.79.12.58。于是 VTEP A 将数据包 M 封装称为数据包 X = [源 IP: 102.10.10.122][目的 IP: 159.79.12.58][VNI: 9][数据包 M]

3. 数据包走三层网络转发

   数据包 X 就是一个普通的三层网络的数据包了，有源 IP 地址，也有目的 IP 地址，因此按照普通的路由转发方式发送给 VTEP B。

4. 数据包在离开三层网络前被解封装

   数据包 X 到达 VTEP B 后，VTEP B 对数据包解封装，将数据包 M 提取出来。

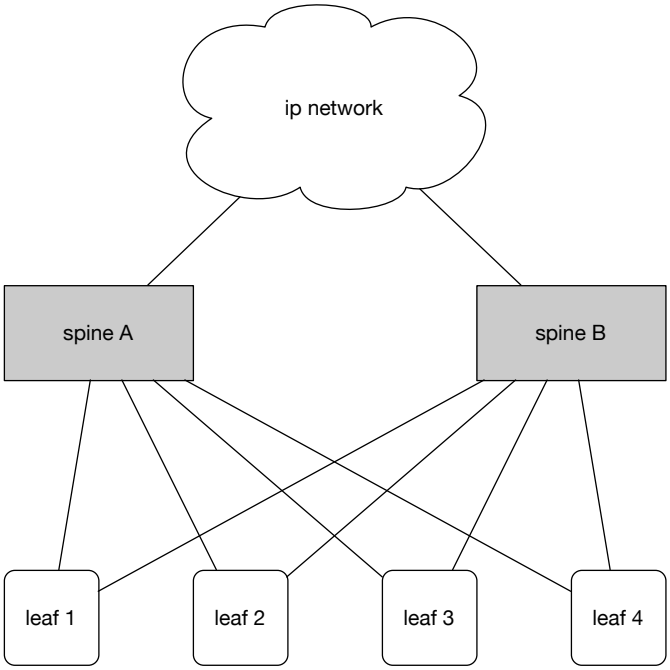
5. 数据包传达目的虚拟机

   VTEP B 检查 M 的目的 MAC 地址为 222222，于是直接将数据发送给了 VM B。

1.3 三层物理网络的负载问题

数据中心虚拟机之间的数据全部由三层物理网络承载，很容易造成三层物理网络的负载不均衡。比如某个 VXLAN 内的虚拟机之间进行大量数据传输，会造成某些物理链路流量十分拥塞，甚至影响其他业务，而其他某些链路则可能十分空闲。因此需要对基于 VXLAN 的数据中心实施流量调优。

2 基于 opendaylight 的流量调优项目



组网介绍:

数据中心的三层物理网络是一种层次架构，底层 leaf 是三层交换机，既同时具有交换能力和路由能力的网络设备，leaf 下面连接大量的虚拟机。spine 是核心路由器，也是整个数据中心的网关，可以连接到外网。

2.1 拓扑发现

我们必须知道整个三层网络的拓扑结构才能实施流量调优。三层网络的拓扑结构是 opendaylight 的 BGPLS 模块自动处理的，而且还能动态响应网络拓扑的变化。BGPLS 将收集到的拓扑信息存储于 opendaylight 的数据库，拓扑信息以点集和链路集的方式组织（nodes and links）。

我们需要从数据库读取拓扑信息，并传递给客户端以直观的图形化方式进行展示。

2.2 流量监控

为了及时发现网络中存在的拥塞并及时进行流量调优，我们需要监控网络的链路状态，方法是通过 netconf 协议连接到网络设备，读取网络设备每个接口的带宽和当前速率，每秒钟读取一次，并保存结果。

客户端可以通过我们提供的 rest 接口请求链路状态，并结合拓扑信息进行网络拥塞的分析。

2.3 路径计算

当发现网络中存在拥塞，如果把某两个 VTEP（leaf 都是 VTEP）之间通信的数据转移到其他负载很轻的路径上会减缓网络拥塞。这就需要计算这两个 VTEP 之间所有可能存在的路径，并选择其中负载较轻的路径进行流量转移。我们的路径计算算法是采用广度优先算法。

2.4 策略下发

当计算出 VTEP 间负载较轻的链路后，就需要针对这条链路进行策略路由的下发，从而将此 VTEP 间的流量引导到这条链路上来。客户端下发一条流量调优的策略格式为:

```
{
    “scheduleName” : “leaf1-leaf2”,
    “scheduleId” : 1,
    “sourceIP” : “12.10.10.112”,
    “destIp” : “150.79.12.58”,
    “vni” : 9,
    “hops” :
    [“x.x.x.x”, “x.x.x.x”, “x.x.x.x”]
}
```

一条规则描述了一条完整的路径，控制器通过 netconf 协议为路径上的每个网络节点进行策略路由的下发，每个策略路由的信息如下:

```
{
    “scheduleName” : “leaf1-leaf2”,
    “scheduleId” : 1,
    “destIp” : “150.79.12.58”,
    “vni” : 9,
    “nexthop” : “x.x.x.x”
}
```

这样当数据包通过网络设备转发时，会对数据包进行 ACL 匹配，如果发现数据包的 destIp 和 VNI 与此网络设备的某条策略路由匹配，则按照策略路由指定的 nexthop 进行转发；如果不能匹配，则按照普通的路由表进行转发。如此，某个 VXLAN 两个 VTEP 间的所有数据流量都会通过特定的路径转发，从而实现网络的负载均衡。

2.5 其他特性

1. 动态响应网络拓扑结构的变化
2. 客户端下发的流量调优策略持久化，即便控制器重启也不影响，仍然会恢复至以前的状态
3. 对客户端下发的策略进行校验，如果与当前的拓扑不符合则不下发
4. 调优策略动态响应拓扑变化，例如某条规则 A 下发时与拓扑不匹配，而后网络拓扑变化了，发现 A 能够与之匹配，则自动下发此策略；反之，也会自动撤销此策略
5. 事务性：一条策略的下发涉及到对若干网络设备的配置，如果对某个设备的配置失败，则不再继续配置，而且之前配置成功的设备也撤销配置。也就是说，一条策略，要么完全下发成功，要么完全失败。
6. 幂等性：同一条策略不会下发多次