zhuanlan.zhihu.com /p/30119950

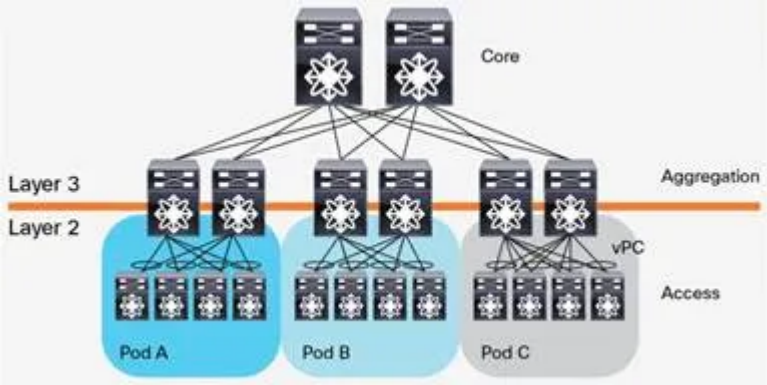
数据中心网络架构浅谈（三）

构建一个数据中心网络时，除了考虑网络硬件设备的架构，2-7层网络设计也需要考虑。这两者其实不能完全分开，硬件架构有时候决定了网络设计，网络设计有时候又限制了硬件架构。从应用场景，例如SDN/NFV来看，网络设计是最直接需要考虑的。所以这部分说说网络设计。

传统三层网络架构中的网络设计

L3架构

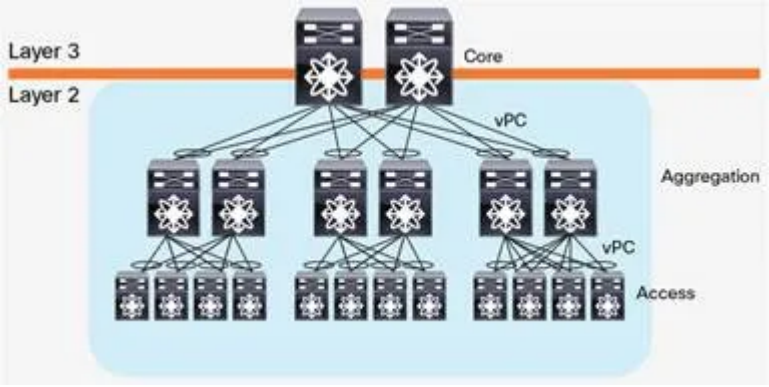
前面几部分说过，传统的三层网络架构中，通常是在汇聚层做L2/L3的分隔。这样可以在每个汇聚层POD构建一个L2广播域，跨汇聚层的通信通过核心交换机做L3路由完成。例如，在划分VLAN时，将VLAN200划分在POD A，VLAN300划分在POD B，VLAN400划分在POD C。这样设计的好处是BUM（Broadcase，Unknown Unicast，Multicast）被限制在每个POD。



由于L2广播域被限制了在汇聚层POD，所以服务器的迁移一般在POD内部完成。因为跨POD迁移，对应二层网络会变化，相应的服务器需要做一些变化，例如IP地址，默认网关。也就是说，服务器所在的网络，限制了服务器的部署范围（只能在POD内）。

大二层架构

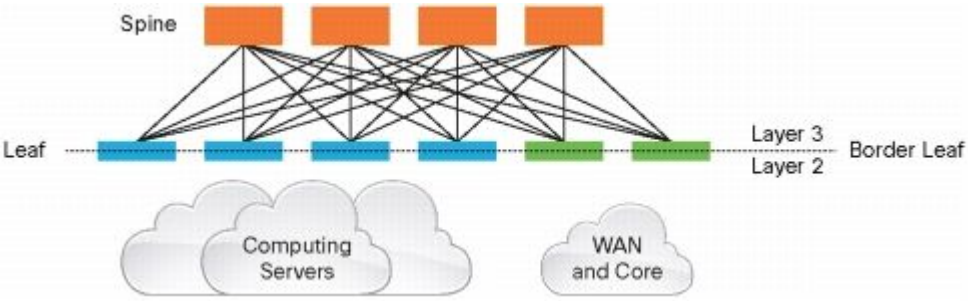
为了更灵活的管理服务器，需要服务器可以部署在数据中心任意位置，在任意位置做迁移，可以使用大二层架构。在这种架构下，整个核心交换机以下都是一个L2广播域，L2广播域中的不同L2网络，通过核心交换机的路由功能转发，同一个L2网络，服务器可以任意迁移部署。



这种架构的缺点就是，BUM会在整个数据中心传播，这最终限制了网络的规模。因为网络规模大到一定程度，BUM会严重影响正常的网络通讯。

Spine/Leaf架构中的网络设计

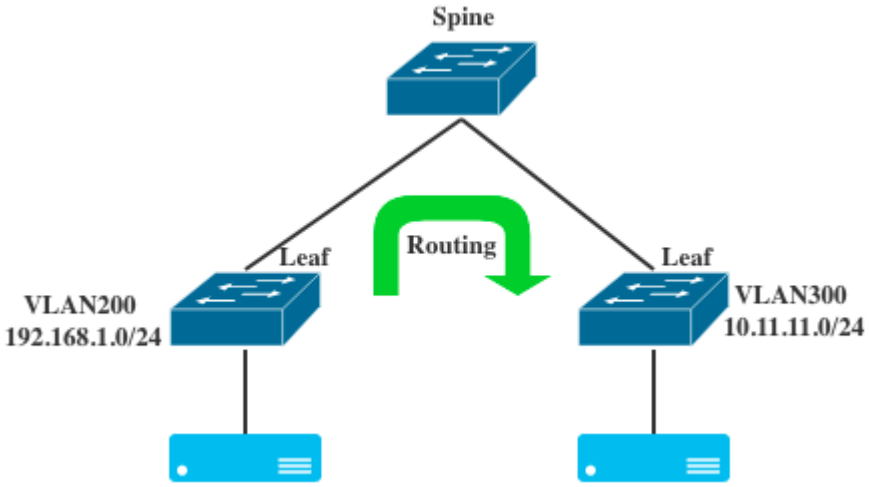
Spine/Leaf网络架构中，L2/L3的分隔通常在Leaf交换机。也就是说每个Leaf交换机下面都是个独立的L2广播域。如果说两个Leaf交换机下的服务器需要通讯，需要通过L3路由，经由Spine交换机转发。



与传统三层网络架构类似，这样的设计，能分隔L2广播域，适用的网络规模更大。但是对应的问题就是，服务器的部署不能在数据中心的任意位置。我们来进一步看这个问题。

当服务器（虚拟的或者物理的）需要被部署在数据中心时，一般需要指定特定的网络分段（Segment）中，或者说特定的L2广播域。如果Segment被局限在了某些特定的交换机下，那么服务器只能在这些交换机的管理范围内部署。也就是说，网络限制了计算资源的部署和分配。但是实际中，真正与计算资源相关的资源，例如对于物理服务器来说，机架的空间，电源，散热等，或者对于虚拟服务器来说，服务器的CPU，内存，硬盘等，这些因素才应该是决定服务器是否部署的因素。如果说对应的机架或者计算资源已经被使用了80%，而其他的机架或者计算资源还基本是空置的，但是网络只在这个高负荷的位置可用，服务器再向这个高负荷的位置进行部署明显不合适。

有什么解决办法能打破网络的限制？例如给空置的机架对应的交换机也配置上相应的网络，让新的服务器部署在它们之上，这样可行吗？举个例子，看一个最简单的spine/leaf架构：

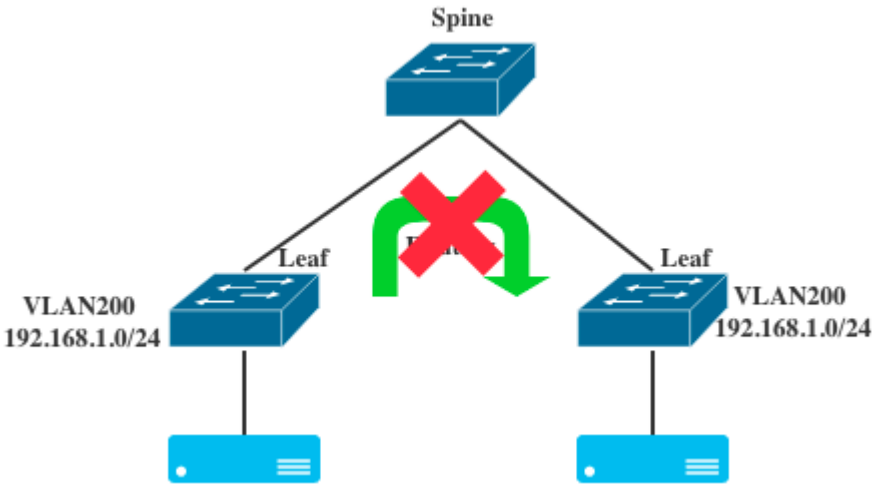


例如左边的leaf配置了VLAN200，管理的CIDR是192.168.1.0/24，右边的leaf交换机配置VLAN300，管理的CIDR是10.11.11.0/24。左右的服务器可以通过L3路由进行转发，这没问题。但是这种情况下，服务器的部署需要考虑网络的可用性。VLAN200的服务器只能在左边，VLAN300的服务器只能在右边，这部分上面说过。

那直接给右边leaf交换机也配上VLAN200，IP地址也配上192.168.1.0/24。看起来似乎可以打破网络的限制，但是实际上，这会导致：

- 两边的服务器的广播域是不通的，左边发出来的广播，Spine上的L3路由不会转发，所以右边是收不到的。
- 左边的服务器不能到达右边的服务器，因为从IP地址来看，左右服务器在一个二层网络，但是实际上两边服务器又不在一个L2广播域中，数据不会发向L3路由，本地也找不到。
- Spine交换机会感到困惑，因为当它收到目的地址是192.168.1.0/24的数据包时，它不知道该路由给左边还是右边。

实际效果如下图所示：

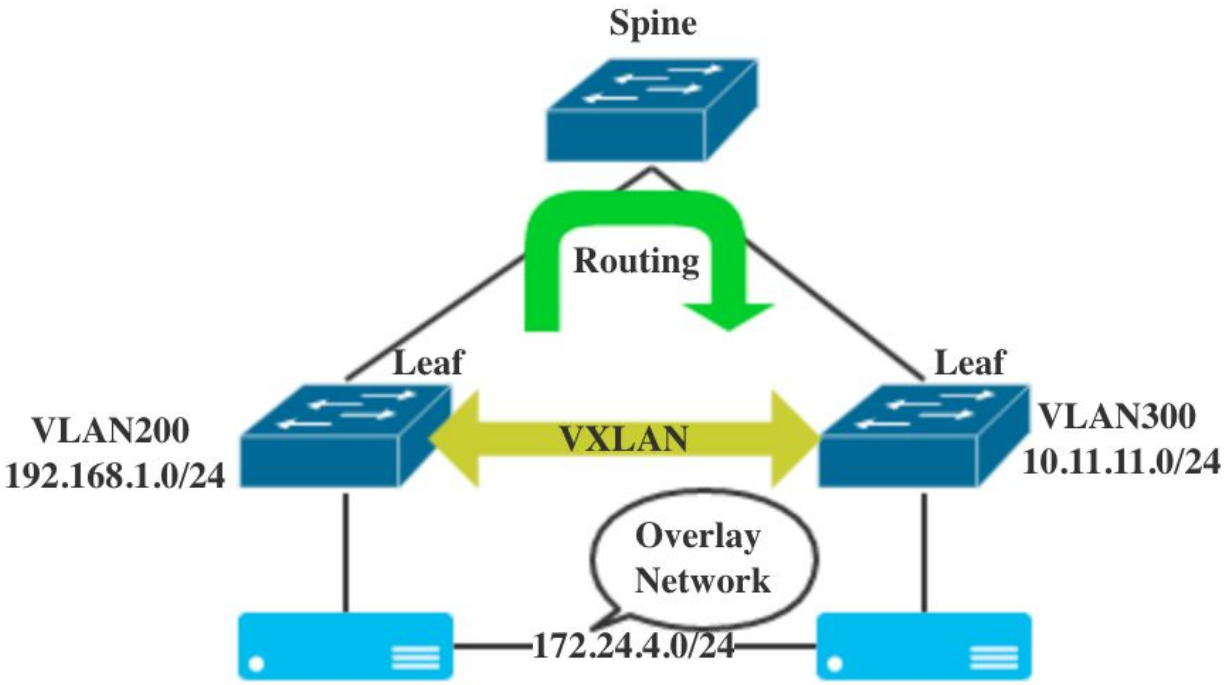


这里相当于，在左右两个Leaf交换机上，创建了两个（而不是一个）VLAN200的网络。而由于CIDR重复，这两个网络之间还不能路由。

Overlay网络

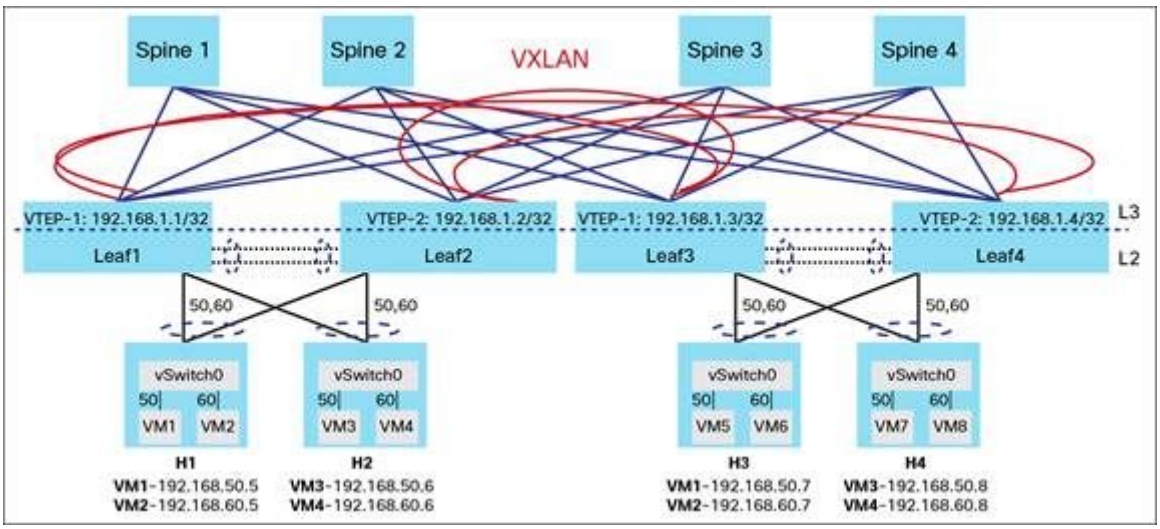
Overlay网络技术可以很好的解决上面的问题。Overlay技术有很多中，GRE，NVGRE，Geneva，VXLAN。这里就不说谁好谁不好，只以VXLAN为例说明，但是大部分内容其他的Overlay技术同样适用。

Overlay网络是在现有的网络（Underlay网络）基础上构建的一个虚拟网络。所谓的现有网络，就是之前的交换机所在的网络，只要是IP网络就行。而新构建的Overlay网络，用来作为服务器通讯的网络。Overlay网络是一个在L3之上的L2网络。也就是说，只要L3网络能覆盖的地方，那Overlay的L2网络也能覆盖。例如下图中，原有的交换机网络不变，服务器之间通过Overlay网络实现了跨Leaf交换机的L2网络。这样，在Overlay网络中，服务器可以任意部署，而不用考虑现有网络的架构。



当提起VXLAN解决了什么问题时，很多人想到的是解决了VLAN ID数量有限的问题，这的确也是VXLAN RFC7348明确说明的。但是现实中解决VLAN ID数量不够还有别的方法，例如QinQ。以VXLAN为代表的Overlay技术解决的，更多是（个人观点）提供了一个不受物理网络限制的，可软件定义的网络环境。

一个完整的Spine/Leaf网络架构配合VXLAN示意图如下所示，这个图里面以虚拟服务器（VM）为例说明，但是实际上并不局限于虚拟的服务器。对于VM来说，并不知道什么VXLAN，VM只是把Ethernet Frame发出来。Leaf交换机（或者说VTEP）将VM的Ethernet Frame封装成VXLAN（也就是一个UDP包），在原有的Spine/Leaf的Underlay网络传输。因为是一个UDP包，所以可以在原有的L3网络中任意传输。



采用Overlay，需要在Leaf交换机上集成VTEP。有时候，将这种网络架构称为VXLAN Fabric。为什么是Fabric，上一篇说过了。

VXLAN Fabric网络架构通常有两种实现，一种是基于Flood-Learn的模式，与传统的L2网络类似，另一种是基于MP-BGP EVPN作为控制层。有关数据中心内的EVPN，我在之前的多篇文章有介绍，感兴趣可以去看看，这里就不再重复了。

最后

Overlay技术并非为Spine/Leaf网络架构设计，早在传统的三层网络架构中，也有应用Overlay技术构建虚拟网络。只是说Spine/Leaf架构作为一种相对较新的网络架构，配合VXLAN或者其他Overlay技术，能够设计出更灵活的数据中心网络。在SDDC（Software Defined Data Center）架构或者SDN中，这种Overlay更是非常重要的一个部分。

最后的最后，这个系列的浅谈就先说到这里了。这个系列有一部分是受到

@阿布
同学的数据中心网络架构演进（一）和数据中心网络架构演进（二）的启发，在此感谢。之所以也写一次，是想提供一个非网工的，云计算从业人员的视角，希望对大家有帮助。