zhuanlan.zhihu.com /p/64445886

数据中心网络架构浅谈（四）

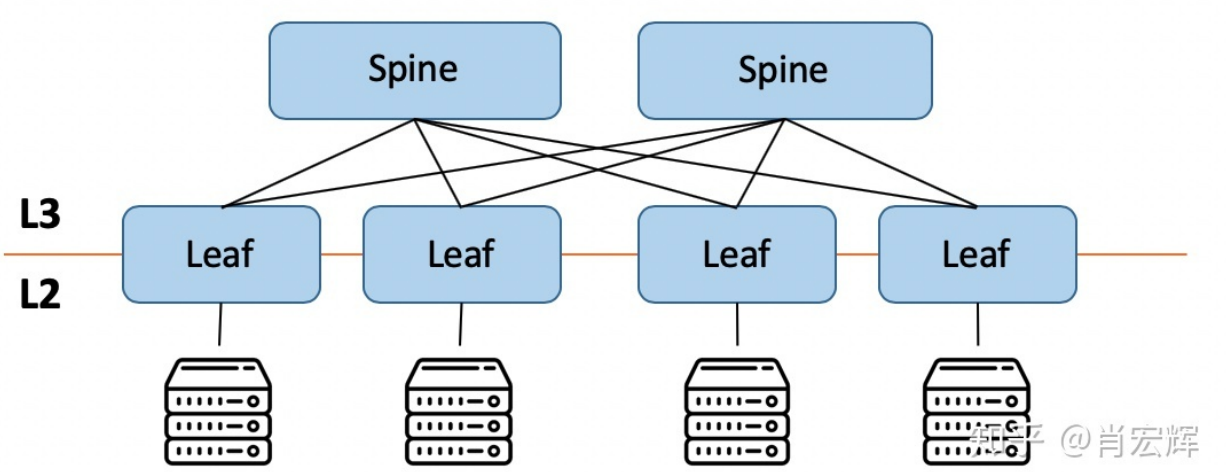
通常来说，如果一个数据中心服务器规模超过10万台，就可以称large-scale datacenter，也就是常说的大规模数据中心。大规模数据中心对于网络的要求有很多，但是最突出的就在于稳定和简单。

这两点要求本身也有一定的关联性。比如，大规模数据中心因为网络设备数量多，所以从统计学的角度来说，出故障的频率也更高。这里说的故障，不仅包括设备本身出现的硬件软件问题，还包括因为运维过程中对设备误操作引起的故障。因此，一个简单的网络设计，例如采用统一的硬件连接方式，使用有限的软件功能，能减少故障概率，从而一定程度提升整个网络架构的稳定性。

但是，或许不只对于IT行业，对于任何领域，用简单的方法去解决一个复杂的问题，本身就不简单。因此，这一次分析一下如何用CLOS架构，来“简单的”管理大规模数据中心的网络。

CLOS架构

CLOS架构被广泛应用在现代的数据中心，因为它提供了数据中心的水平扩展能力和大规模数据中心所需要的稳定和简单。下图就是一个最基本的CLOS单元，Spine和Leaf交换机共同组成数据中心网络，其中Leaf交换机作为TOR交换机，连接服务器；Spine交换机，为Leaf交换机提供网络连接。

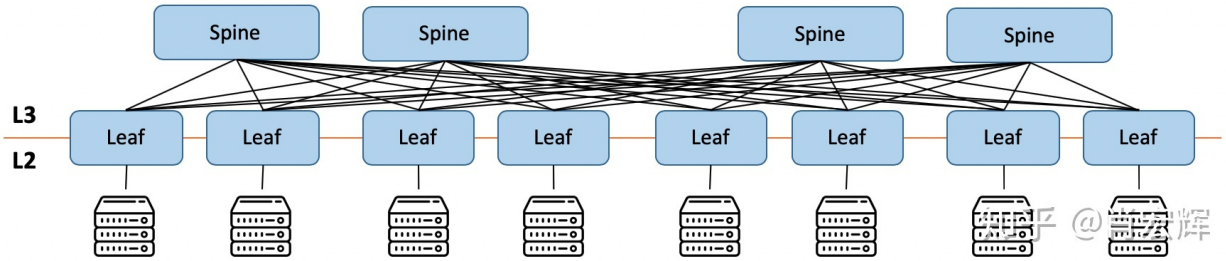


水平扩展能力

--

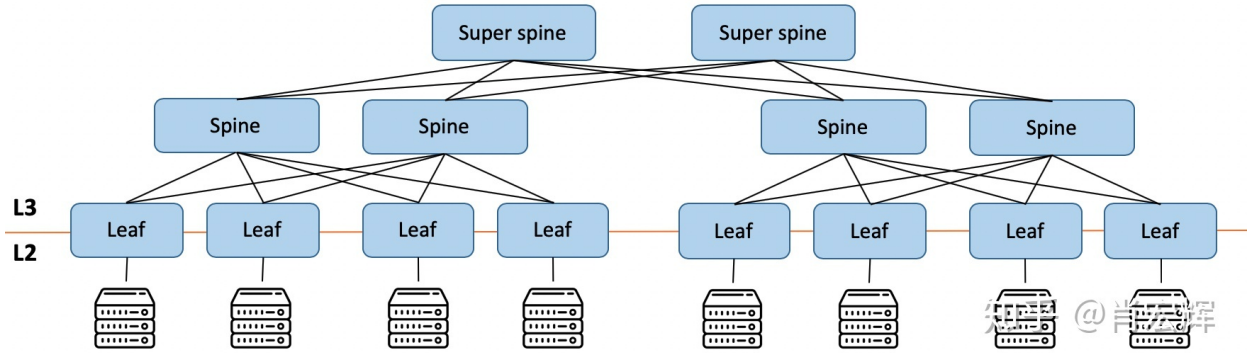
想要扩展一个CLOS网络架构，通常有两种方法，第一就是增加设备的端口数；第二就是增加更多的层级。

通过增加交换机端口数量，可以连接更多的服务器，如下图所示，端口数量扩大一倍，数据中心规模也扩大了一倍。



也可以增加CLOS架构的层级数。上面图中都是3-stages CLOS架构，虽然只有两层交换机，但是因为对应CLOS的理论，是一个对折了的架构，所以被称为3-stages。

在现有的spine-leaf基础上，再增加一层super-spine交换机，就可以构成一个5-stages CLOS架构。如下图所示，增加了一层super-spine交换机，数据中心规模也水平扩大了一倍。



稳定简单

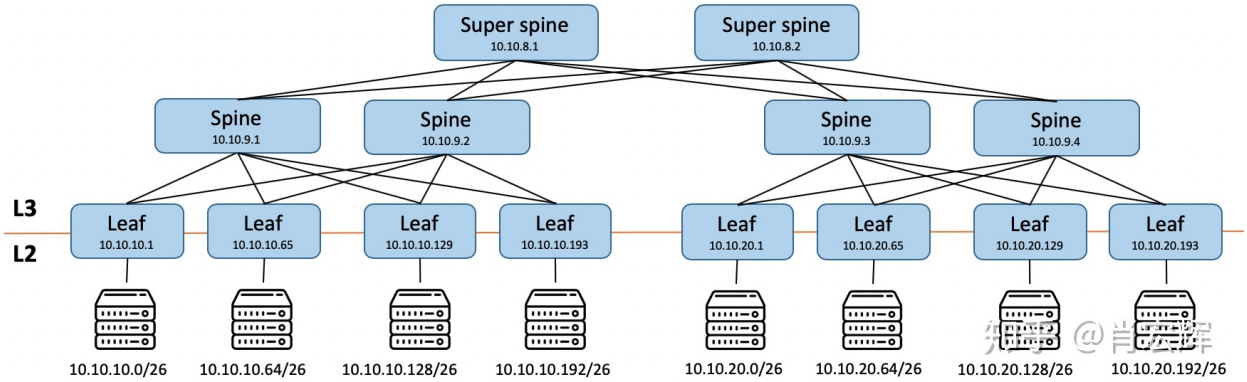
--

从第一眼来看，CLOS架构是简单的。根据CLOS理论，所有的交换机，不论是Super Spine，Spine和是Leaf，都应该采用同质的交换机。虽然实际应用和文中图里面，都不一定严格按照这个要求来，但是至少是照着这个样式去实现。所以从硬件构成来看，较为简单。

其次，CLOS架构采用的是一个纯L3网络的架构，也就是说所有的交换机都是三层交换机，交换机之间都通过IP网络连接的。所以，从网络连接来看，也较为简单。毕竟，传统三层网络架构，要涉及到L2连接，L3连接，VLAN配置等等。

稳定包含很多，除了相对简单的设计，还有就是减少故障范围。0故障是不可能的，这辈子都不可能0故障。我们能做的是限制故障的范围，而CLOS架构下，每个Leaf交换机下都是一个独立的L2 Domain，这样可以将所有二层网络的问题，例如BUM风暴，限制在一个Leaf交换机范围内。

所以最终网络架构的IP地址分布如下：



看起来似乎很美好，但是相比较传统的三层网络，CLOS架构也有自己的问题，其中包括但不限于以下几点：

- 独立的L2 Domain限制了依赖L2 Domain应用程序的部署。要求部署在一个二层网络的应用程序，现在只能部署下一个机架下了。
- 独立的L2 Domain限制了服务器的迁移。迁移到不同机架之后，网关和IP地址都要变，这谁也受不了。
- 子网数量大大增加了。每个子网对应数据中心一条路由，现在相当于每个机架都有一个子网，对应于整个数据中心的路由条数大大增加，并且这些路由信息要怎么传递到每个Leaf上，也是一个复杂的问题。

前两点都可以通过VxLAN来解决，VxLAN将应用网络和基础网络解耦开，这是其最大价值和流行的最主要原因（个人认为没有之一）。至于解决VLAN ID不够的问题，只能说是一个肤浅的标签，毕竟不是谁都做公有云，企业数据中心内部署，大部分连4000个TAG都用不完。

路由协议选择

配置各种协议最终在物理链路连通的基础上将整个数据中心网络打通，是数据中心网络运维最复杂的工作，也是各位网工朋友的主要负担。传统的三层网络架构，因为是一个L2和L3混合的网络，因此STP（以及STP的改良版RSTP，MST），VLAN，OSPF，EIGRP，IBGP都需要做相应的配置。

CCIE就是在一个约50台网络设备的环境里，在6个小时内，将这些配置配通。在一个大规模的数据中心，对应的网络设备是数千台，如果还按照这个配置跑的话，运维成本太高，这也是传统三层网络架构被取代的原因之一。

那如何为CLOS架构选取一个合适的路由协议？常规的选项是使用且仅使用EBGP。

BGP一直以IBGP的形式来构建数据中心内部网络，而且是构建在IGP，例如OSPF之上的。而EBGP一般用来连接不同的数据中心。但是在CLOS架构中，EBGP，却是最合适的一个协议，因为它能极大简化实现。

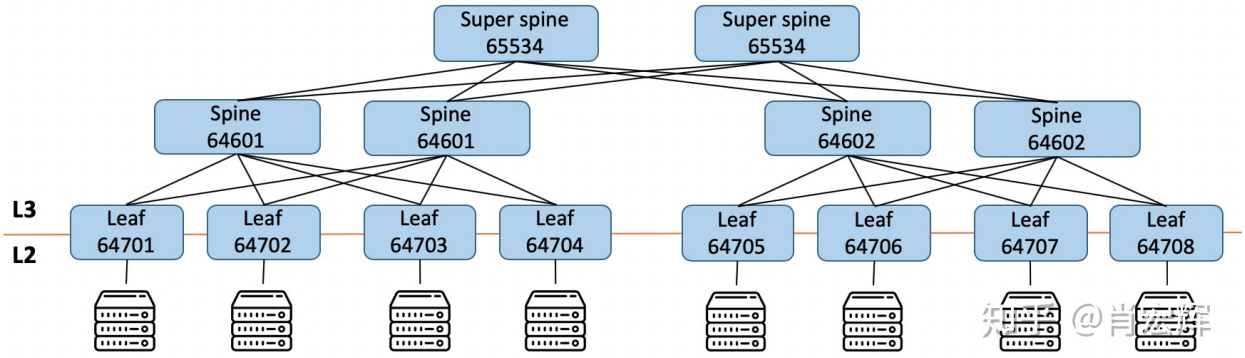
首先因为现在路由条目数增多，这样的量级只有BGP能稳定维护。

其次，因为现在每个Leaf Switch，都管理一个独立的子网。而数据中心内网络连通的前提是，每一个Leaf Switch的子网，都需要传给其他所有的Leaf Switch。这样，相当于每个Leaf Switch都是一个自治域（AS），现在要实现的就是实现所有的自治域的连通。这个问题，就是EBGP在互联网上正在解决的问题。

因此在CLOS架构下，采用了EBGP作为路由协议，具体细节有以下几点：

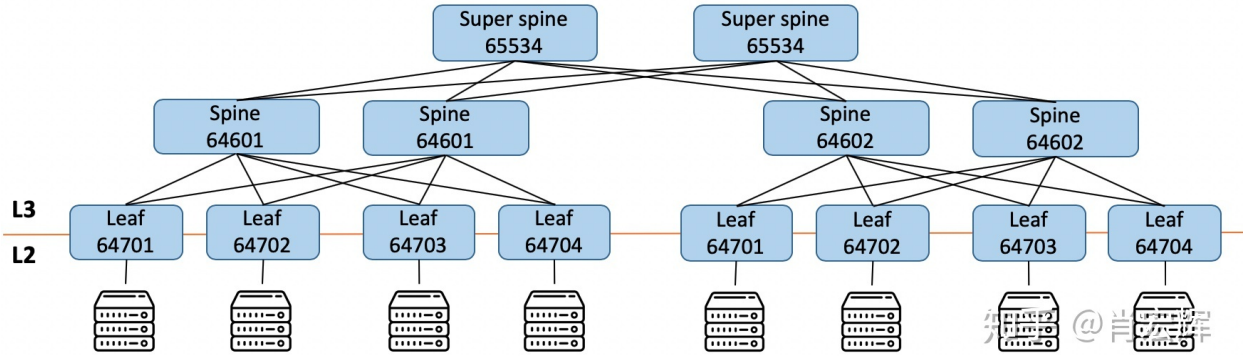
- EBGP连接都是单跳。这样就不用依赖IGP构建nexthop网络，EBGP的nexthop都在链路的另一端。
- 采用ASN中保留给数据中心内部的ASN 64512到65534，共1023个ASN。
- 所有Super Spine共用一个唯一的ASN。
- 每组Spine共用一个唯一的ASN。
- 每个Leaf有一个唯一的ASN。

所以最后整个CLOS网络架构中EBGP连接示意图如下。



EBGP与IBGP的一个最大不同在于，EBGP会转发路由，因此借助Spine和Super Spine上面EBGP程序的转发，一个Leaf的子网信息，可以发布给所有其他Leaf交换机，从而实现全数据中心内网络联通。

但是这里有一个问题，在大规模数据中心里面，按照10万条服务器，一个机架40台服务器算的话，总共会有2500个Leaf交换机，这样，光是Leaf就把 ASN消耗完了。为了解决这个问题，可以使用4字节的ASN（RFC6793）；也可以在一组Spine下面，复用ASN，如下图所示：



复用ASN的话，为了避免EBGP的 AS loop detection，需要把EBGP的allowas-in给打开。这样，左边64701对应的Leaf发出的路由，就可以被右边64701对应的Leaf接收。

通过这样的EBGP连接，很自然就实现了ECMP，而ECMP又是CLOS架构的核心所在。

所以，仅通过EBGP，就是实现了CLOS架构中网络连接需要的全部内容。相比较传统三层网络架构，CLOS架构这里又以简单胜出。

发布于 2019-05-02 05:55