zhuanlan.zhihu.com /p/29881248

数据中心网络架构浅谈（一）

不论我们在讨论SDN，NFV或者其他的虚拟网络技术，有一点需要明确，网络数据包最终都是跑在物理网络上。物理网络的特性，例如带宽，MTU，延时等，最终直接或者间接决定了虚拟虚拟网络的特性。可以说物理网络决定了虚拟网络的“天花板”。在Mirantis对OpenStack Neutron的性能测试报告中可以看出，网络设备的升级和调整，例如采用高速网卡，配置MTU9000，可以明显提高虚拟网络的传输效率。在对网络性能进行优化时，有些物理网络特性可以通过升级设备或线路来提升，但是有些与网络架构有关。升级或者改动网络架构带来的风险和成本是巨大的，因此在架设数据中心初始，网络架构的选择和设计尤其需要谨慎。另一方面，在设计虚拟网络时，不可避免的需要考虑实际的物理网络架构，理解物理网络架构对于最终理解虚拟网络是不可缺少的。

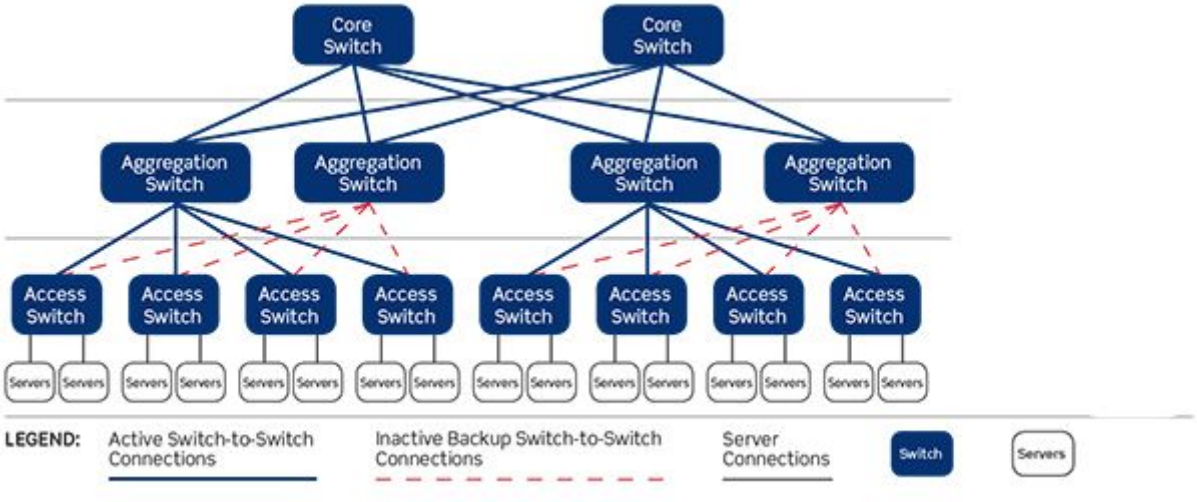
接下来我将分几次说一说自己对数据中心网络架构的认识，想到哪说到哪，不对的地方请大家指正。

传统数据中心网络架构

在传统的大型数据中心，网络通常是三层结构。Cisco称之为：分级的互连网络模型（hierarchical inter-networking model）。这个模型包含了以下三层：

- Access Layer（接入层）：有时也称为Edge Layer。接入交换机通常位于机架顶部，所以它们也被称为ToR（Top of Rack）交换机，它们物理连接服务器。
- Aggregation Layer（汇聚层）：有时候也称为Distribution Layer。汇聚交换机连接Access交换机，同时提供其他的服务，例如防火墙，SSL offload，入侵检测，网络分析等。
- Core Layer（核心层）：核心交换机为进出数据中心的包提供高速的转发，为多个汇聚层提供连接性，核心交换机为通常为整个网络提供一个弹性的L3路由网络。

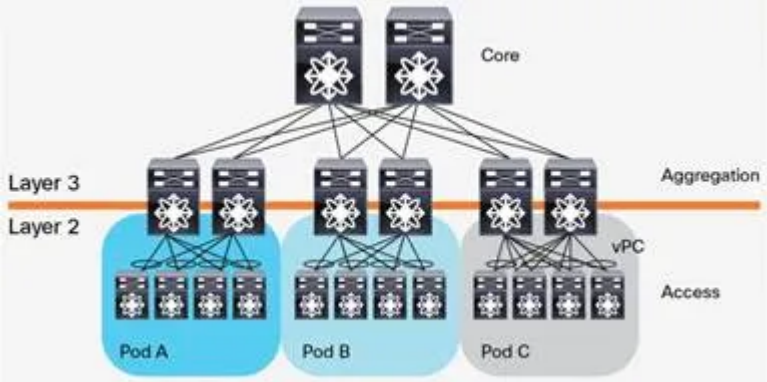
一个三层网络架构示意图如下所示：



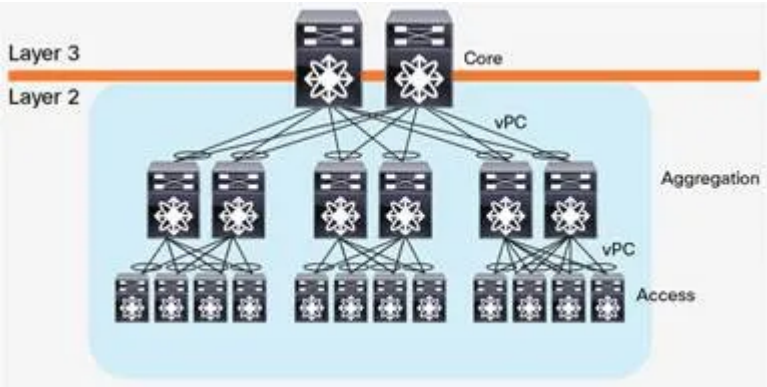
通常情况下，汇聚交换机是L2和L3网络的分界点，汇聚交换机以下的是L2网络，以上是L3网络。每组汇聚交换机管理一个POD（Point Of Delivery），每个POD内都是独立的VLAN网络。服务器在POD内迁移不必

修改IP地址和默认网关，因为一个POD对应一个L2广播域。

汇聚交换机和接入交换机之间通常使用STP（Spanning Tree Protocol）。**STP使得对于一个VLAN网络只有一个汇聚层交换机可用，**其他的汇聚层交换机在出现故障时才被使用（上图中的虚线）。也就是说汇聚层是一个active-passive的HA模式。这样在汇聚层，做不到水平扩展，因为就算加入多个汇聚层交换机，仍然只有一个在工作。一些私有的协议，例如Cisco的vPC（Virtual Port Channel）可以提升汇聚层交换机的利用率，但是一方面，这是私有协议，另一方面，vPC也不能真正做到完全的水平扩展。下图是一个汇聚层作为L2/L3分界线，且采用vPC的网络架构。



随着云计算的发展，计算资源被池化，为了使得计算资源可以任意分配，需要一个大二层的网络架构。即整个数据中心网络都是一个L2广播域，这样，服务器可以在任意地点创建，迁移，而不需要对IP地址或者默认网关做修改。大二层网络架构，L2/L3分界在核心交换机，核心交换机以下，也就是整个数据中心，是L2网络（当然，可以包含多个VLAN，VLAN之间通过核心交换机做路由进行连通）。大二层的网络架构如下图所示：



大二层网络架构虽然使得虚机网络能够灵活创建，但是带来的问题也是明显的。共享的L2广播域带来的BUM（Broadcast·，Unknown Unicast，Multicast）风暴随着网络规模的增加而明显增加，最终将影响正常的网络流量。

传统三层网络架构已经存在几十年，并且现在有些数据中心中仍然使用这种架构。这种架构提出的最初原因是什么？一方面是因为早期L3路由设备比L2桥接设备贵得多。即使是现在，核心交换机也比汇聚接入层设备贵不少。采用这种架构，使用一组核心交换机可以连接多个汇聚层POD，例如上面的图中，一对核心交换机连接了多个汇聚层POD。另一方面，**早期的数据中心，大部分流量是南北向流量**。例如，一个服务器上部署了WEB应用，供数据中心之外的客户端使用。使用这种架构可以在核心交换机统一控制数据的流入流出，添加负载均衡器，为数据流量做负载均衡等。

技术发展对网络架构的影响

数据中心是为了数据服务。随着技术的发展，数据的内容和形式也发生了变化。

- 虚拟化的流行。传统的数据中心中，服务器的利用率并不高，采用三层网络架构配合一定的超占比（oversubscription），能够有效的共享利用核心交换机和一些其他网络设备的性能。但是虚拟化的流行使得服务器的利用率变高，一个物理服务器可以虚拟出多个虚拟机，分别运行各自的任务，走自己的网络路径。因此，高的服务器利用率要求更小的超占比。Gartner的一份报告：Forecast: x86 Server Virtualization, Worldwide, 2012-2018, 2014 Update指出，在2018年，82%的服务器将是虚拟服务器。虚拟化对数据中心网络架构的影响是巨大的。
- 软件架构的解耦。传统的软件架构，采用专用模式进行部署，软件系统通常跑在一个物理服务器，与其他的系统做物理隔离。但是，模块化，分层的软件架构设计已经成为了现在的主流。一个系统的多个组件通常分布在多个虚机/容器中。最典型的就是三层WEB应用，包含了Client/Application/DB。一次请求，不再是由一个虚机/物理机完成，而是由多个服务器协同完成。这对网络的影响是，东西向流量变多了。
- 新的应用的兴起。传统数据中心是为.com应用设计的，这些流量大多是客户端和服务器之间的通信。而分布式计算，大数据渐渐兴起，这些应用会在数据中心的服务器之间产生大量的流量。例如Hadoop，将数据分布在数据中心中成百上千个服务器中，进行并行计算。据说Facebook的一个Hadoop集群有着超过100 petabytes的数据。可见对于某些应用，数据中心的东西向流量是巨大的。
- 软件定义数据中心（SDDC，Software Defined Data Center）的提出。SDDC提出软件定义的数据中心，这要求数据中心的计算存储网络都是可以软件定义的。对应于网络，就是SDN。传统的三层网络架构在设计之初并没有考虑SDN。

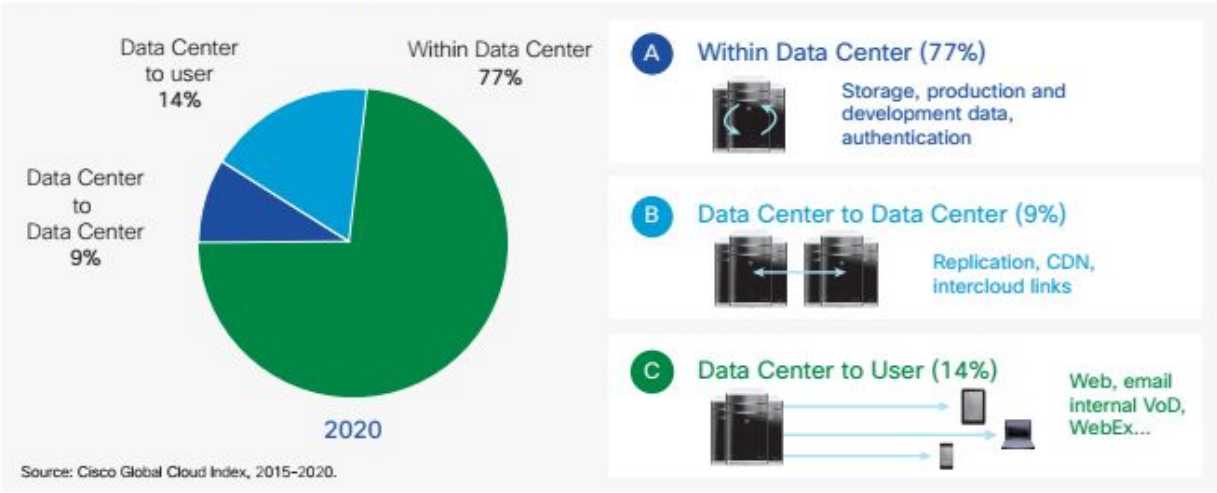
总结起来，**技术发展要求新的数据中心有更小的超占比，甚至没有超占比；更高的东西向流量带宽；支持SDN。**

在这些需求里面，更高的东西向流量支持尤为重要。前面说了南北向流量，东西向流量，这些分别是什么东东？数据中心的流量总的来说可以分为以下几种：

- 南北向流量：数据中心之外的客户端到数据中心服务器之间的流量，或者数据中心服务器访问互联网的流量。
- 东西向流量：数据中心内的服务器之间的流量。
- 跨数据中心流量：跨数据中心的流量，例如数据中心之间的灾备，私有云和公有云之间的通讯。

根据Cisco Global Cloud Index: Forecast and Methodology, 2015–2020，到2020年77%的数据中心流量将会是数据中心内部的流量，也就是东西向流量，这与上面的技术发展对网络架构的影响分析相符，这也是为什么东西向流量尤其重要。

Figure 5. Global Data Center Traffic by Destination in 2020



那传统三层网络架构下的东西向流量是怎么样的？

前面说过传统三层网络架构的诞生是在.com时代，主要是也为了南北向流量设计的。但是传统的网络架构并非不支持东西向流量，下面来分析一下传统三层网络架构中东西向流量走向。

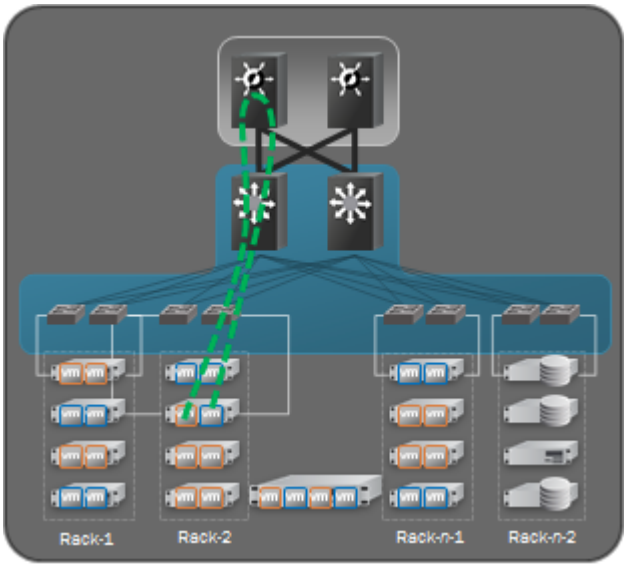
首先，东西向流量分为L2和L3流量。

东西向的L2流量，如果源和目的主机都在同一个接入层交换机下，那么可以达到全速，因为接入交换机就能完成转发。

如果需要跨机架，但仍然是在一个汇聚层POD内，则需要通过汇聚层交换机进行转发，带宽取决于汇聚层交换机的转发速率，端口带宽和同时有多少个接入层交换机共享汇聚层交换机。前面说过汇聚层和接入层之间一般使用STP，这使得一个汇聚层POD只能有一个汇聚层交换机在工作。为了满足跨机架的L2转发，汇聚层交换机的性能，例如带宽，转发速率必然要大于接入层交换机。

如果L2流量需要跨汇聚层POD（大二层架构），那必须经过核心交换机。同样的问题仍然存在，对核心交换机的要求会更高。

东西向的L3流量，不论是不是在一个接入层交换机下，都需要走到具有L3功能的核心交换机才能完成转发。如下图所示：



这是一种发卡（hair-pin）流量，它不仅浪费了宝贵的核心交换机资源，而且多层的转发增加了网络传输的延时。同样，由于超占比的存在，它也不能保证全速的L3流量。

总的来说，为了保证任意的东西向流量带宽，势必需要更高性能的汇聚层交换机和核心交换机。另一方面，也可以小心的进行设计，尽量将有东西向流量的服务器置于同一个接入交换机下。不管怎么样，这都增加了成本，降低了可用性。

市场需求变化对网络架构的影响

由于成本和运维因素，数据中心一般是大企业才有能力部署。但是随着技术的发展，一些中小型企业也需要部署数据中心。不同的是，中小型企业的需求一般是，以一个小规模启动，随着自身业务的增长再逐步的扩展数据中心。数据中心的规模很大程度上取决于网络的规模，对应网络的需求就是，以一个低成本，小规模的网络架构启动，但是要能够水平扩展到较大规模。

传统三层网络架构的规模取决于核心层设备的性能和规模，取决于交换机的端口密度。最大的数据中心对应着体积最大和性能最高的网络设备，这种规模的设备并非所有的网络设备商都能提供，并且对应的资金成本和运维成本也较高。采用传统三层网络架构，企业将面临成本和可扩展性的两难选择。

最后

传统的三层网络架构必然不会在短期内消失，但是由于技术和市场的发展，其短板也越来越明显。基于现有网络架构的改进显得非常有必要，新的网络架构最好是：由相对较小规模的交换机构成，可以方便的水平扩展，较好的支持HA（active-active模式），支持全速的东西向流量，不采购高性能的核心交换机也能去除超占比，支持SDN等等。

编辑于 2017-10-09 20:47