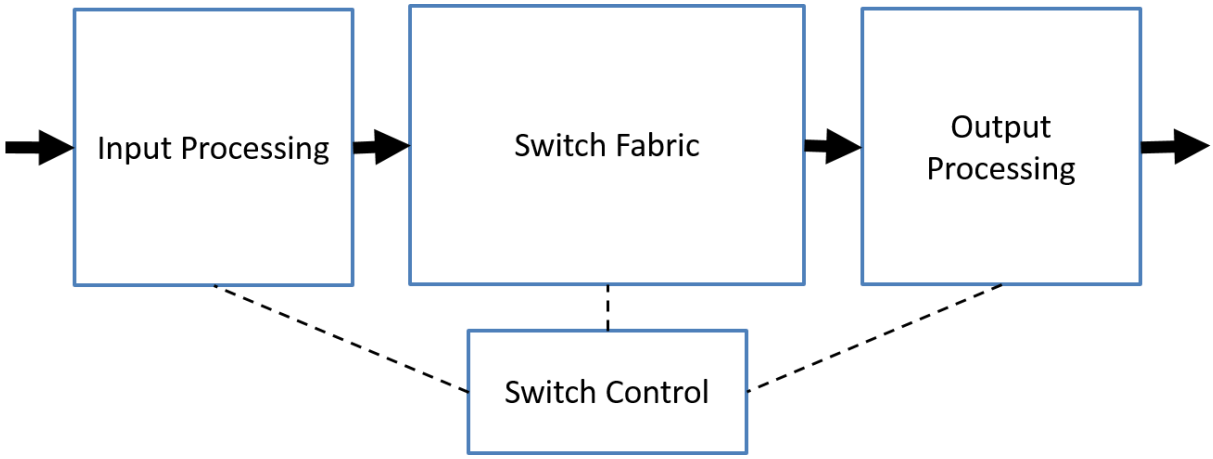
zhuanlan.zhihu.com /p/29975418

数据中心网络架构浅谈（二）

上一篇说了传统三层网络架构，这一次来看看近些年开始流行的Fabric网络架构。

Fabric

Fabric一词来源于网络交换机。网络交换机就是将输入端口的数据，经过判断，转发到输出端口。其架构大体如下图所示：



交换机内部连接输入输出端口的是Switch Fabric。最简单的Switch Fabric架构是crossbar模型，这是一个开关矩阵，每一个crosspoint（交点）都是一个开关，交换机通过控制开关来完成输入到特定输出的转发。一个crossbar模型如下所示：

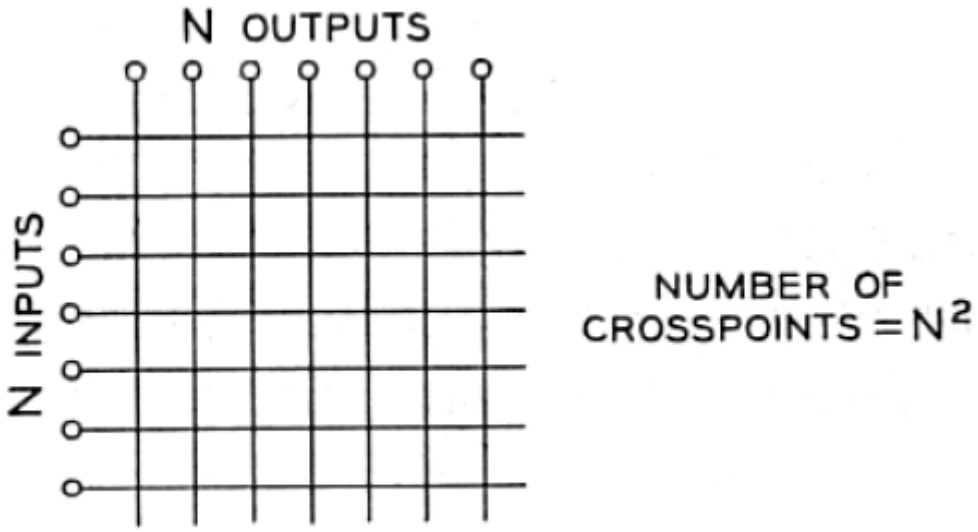


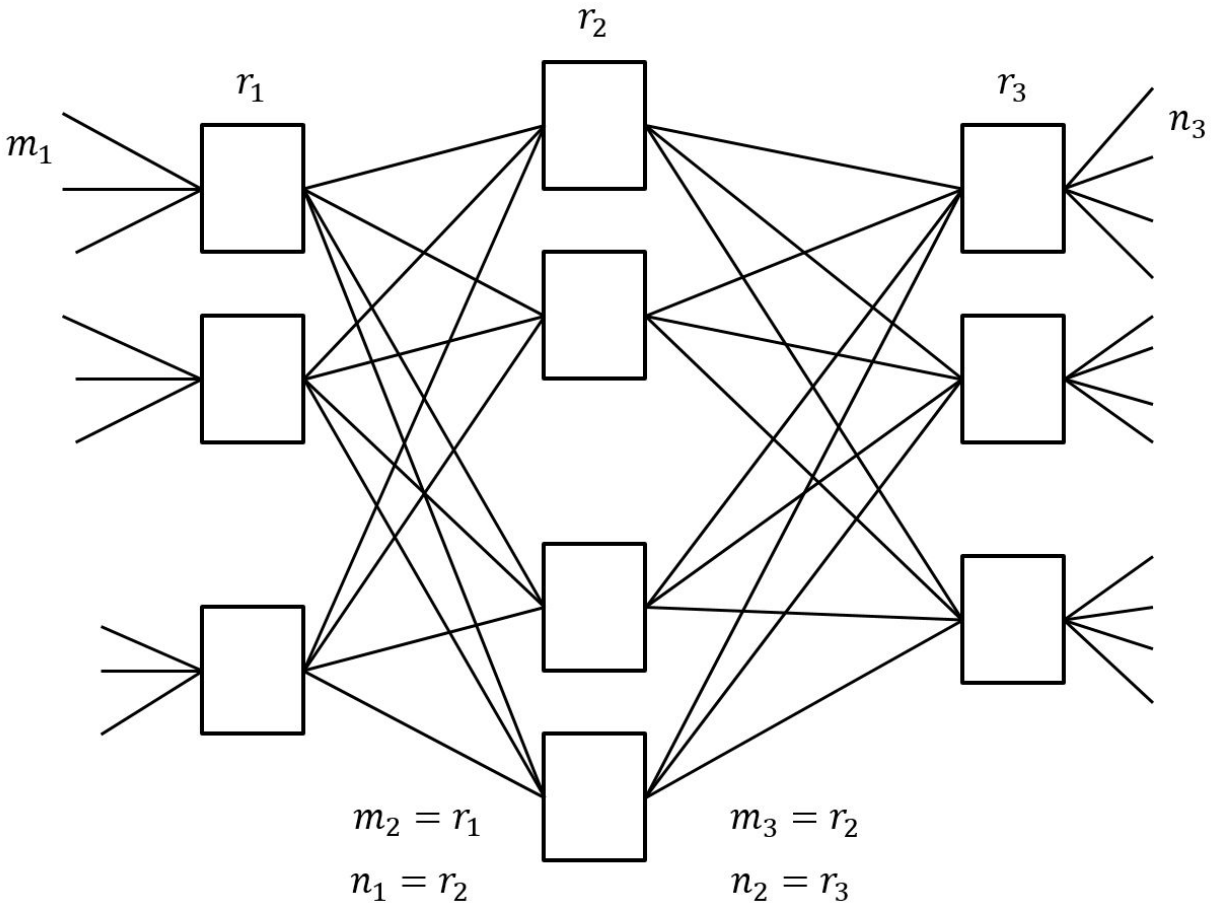
Fig. 1 — Square Array.

可以看出，这里的开关矩阵类似于一块布的纤维，所以交换机内的架构被称为Switch Fabric（纤维）。这是Fabric一词在网络中的起源。

Clos架构

Charles Clos曾经是贝尔实验室的研究员。他在1953年发表了一篇名为“A Study of Non-blocking Switching Networks”的文章。文章里介绍了一种**用多级设备来实现无阻塞电话交换的方法**，这是Clos架构的起源。

简单的Clos架构是一个三级互连架构，包含了输入级，中间级，输出级，如下图所示：



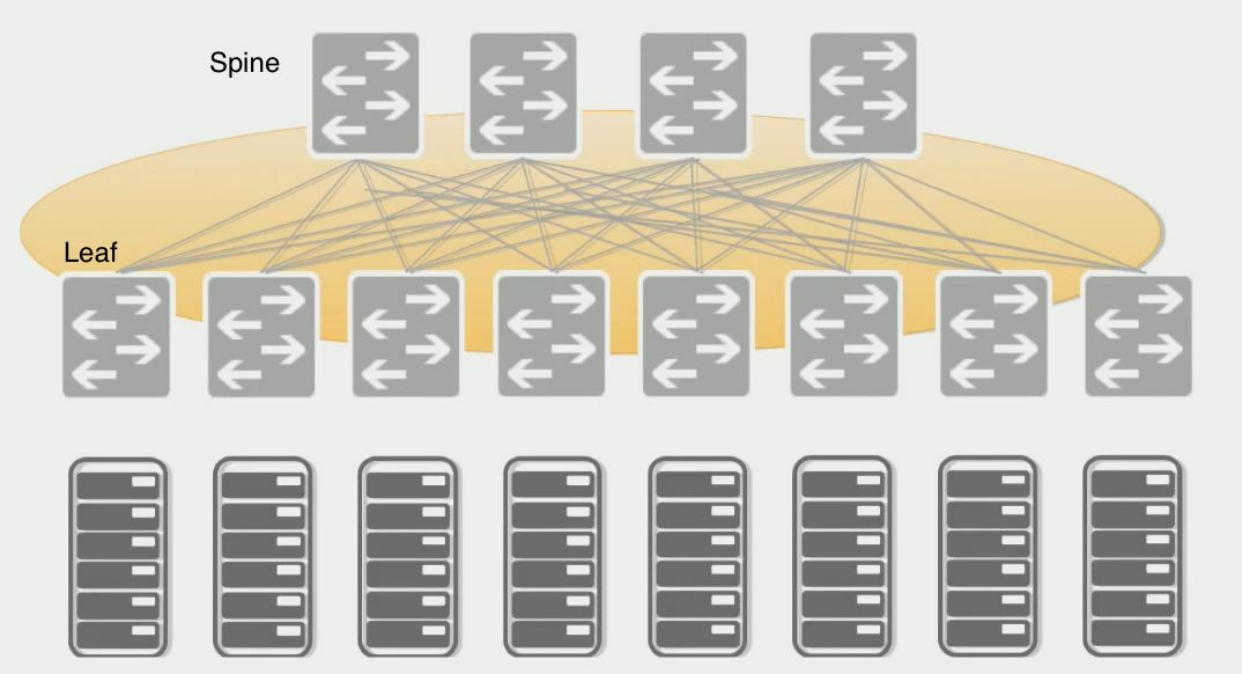
图中的矩形是规模小得多的转发单元，相应的成本小得多。**Clos架构的核心思想是：用多个小规模、低成本的单元构建复杂，大规模的架构。**上图中，m是每个子模块的输入端口数，n是每个子模块的输出端口数，r是每一级的子模块数，经过合理的重排，只要满足r2≥max(m1,n3)，那么，对于任意的输入到输出，总是能找到一条无阻塞的通路。

回到交换机架构，随着网络规模的发展，交换机的端口数量逐渐增多。crossbar模型的交换机的开关密度，随着交换机端口数量N呈 O(N2)O(N^{2}) 增长。相应的功耗，尺寸，成本也急剧增长。在高密度端口的交换机上，继续采用crossbar模型性价比越来越低。大约在1990年代，Clos架构被应用到Switch Fabric。应用Clos架构的交换机的开关密度，与交换机端口数量N的关系是 O(N3/2)O(N^{3/2})，所以在N较大时，Clos模型能降低交换机内部的开关密度。这是Clos架构的第二次应用。

Clos网络架构

由于传统三层网络架构存在的问题，在2008年一篇文章A scalable, commodity data center network architecture，提出将Clos架构应用在网络架构中。2014年，在Juniper的白皮书中，也提到了Clos架构。这一回，Clos架构应用到了数据中心网络架构中来。这是Clos架构的第三次应用。

现在流行的Clos网络架构是一个二层的spine/leaf架构，如下图所示。spine交换机之间或者leaf交换机之间不需要链接同步数据（不像三层网络架构中的汇聚层交换机之间需要同步数据）。每个leaf交换机的上行链路数等于spine交换机数量，同样的每个spine交换机的下行链路数等于leaf交换机的数量。可以这样说，spine交换机和leaf交换机之间是以full-mesh方式连接。

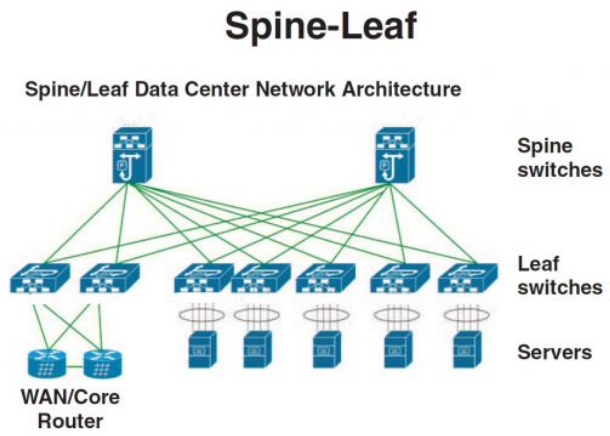


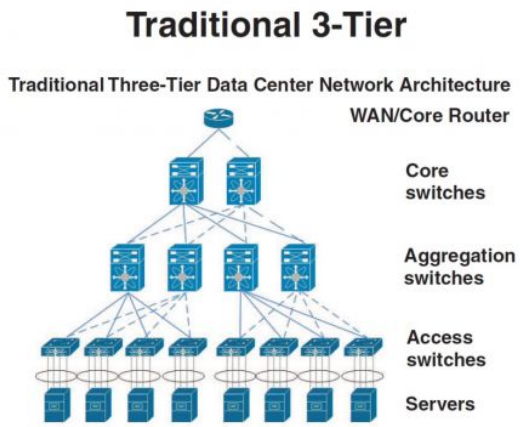
可前面不是说Clos架构是三级设备架构吗？为什么这里只有两层网络设备？这是因为前面讨论Clos架构的时候，都是讨论输入到输出的单向流量。网络架构中的设备基本都是双向流量，输入设备同时也是输出设备。因此三级Clos架构沿着中间层对折，就得到了二层spine/leaf网络架构。由于这种网络架构来源于交换机内部的Switch Fabric，因此这种网络架构也被称为Fabric网络架构。

在spine/leaf架构中，每一层的作用分别是：

- leaf switch：相当于传统三层架构中的接入交换机，作为TOR（Top Of Rack）直接连接物理服务器。与接入交换机的区别在于，L2/L3网络的分界点现在在leaf交换机上了。leaf交换机之上是三层网络。
- spine switch：相当于核心交换机。spine和leaf交换机之间通过ECMP（Equal Cost Multi Path）动态选择多条路径。区别在于，spine交换机现在只是为leaf交换机提供一个弹性的L3路由网络，数据中心的南北流量可以不用直接从spine交换机发出，一般来说，南北流量可以从与leaf交换机并行的交换机（edge switch）再接到WAN router出去。

对比spine/leaf网络架构和传统三层网络架构

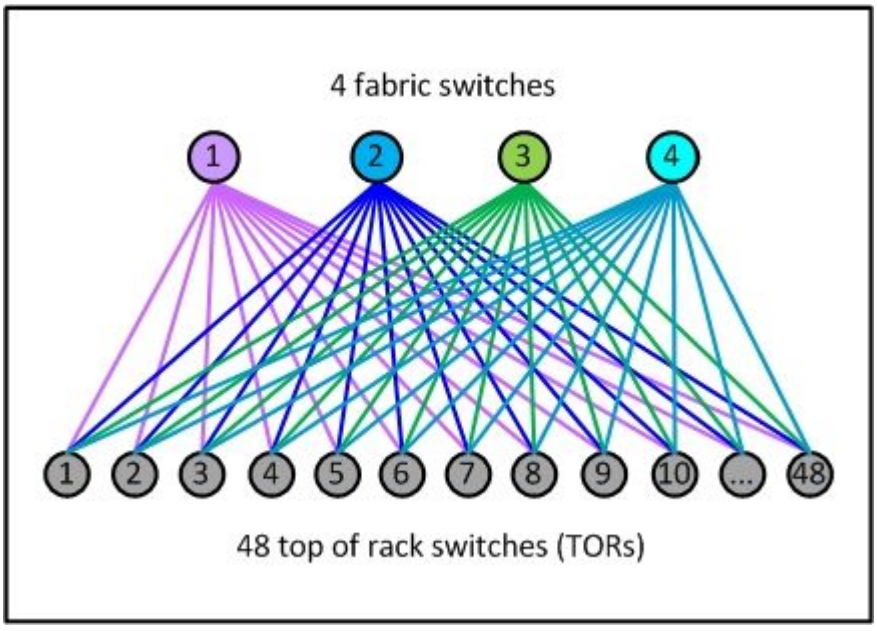




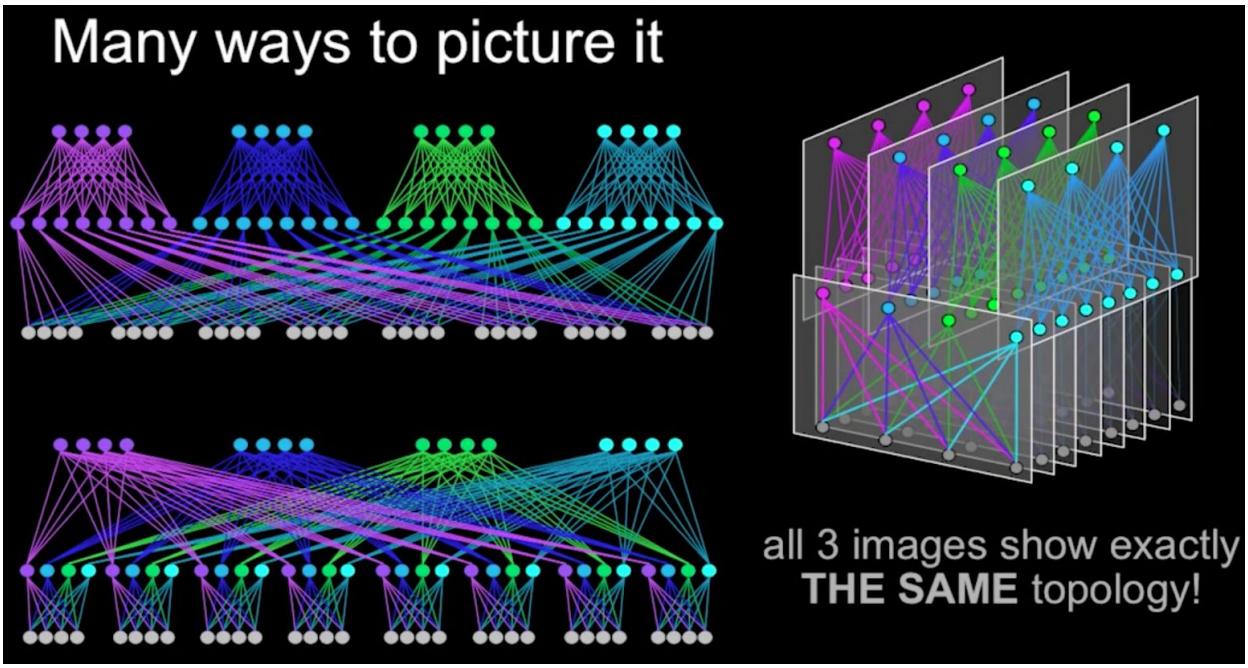
可以看出传统的三层网络架构是垂直的结构，而spine/leaf网络架构是扁平的结构，从结构上看，spine/leaf架构更易于水平扩展。

Facebook Fabric Datacenter

Fabric网络架构最具有代表性的就是Facebook在2014年公开的其数据中心架构：Introducing data center fabric, the next-generation Facebook data center network。Facebook使用了一个五级Clos架构，前面说过实际的网络设备即是输入又是输出，因此五级Clos架构对折之后是一个三层网络架构，虽然这里也是三层，但是跟传统的三层网络架构完全是两回事。对应于上面介绍的架构，Facebook将leaf交换机叫做TOR，间添加了一层交换机称为fabric交换机。fabric交换机和TOR构成了一个三级Clos结构，如下图所示，这与前面介绍的spine/leaf架构是一样的。Facebook将一组fabric交换机，TOR和对应的服务器组成的集群称为一个POD（Point Of Delivery）。POD是Facebook数据中心的最小组成单位，每个POD由48个TOR和4个fabric交换机组成，下图就是一个POD的示意图。

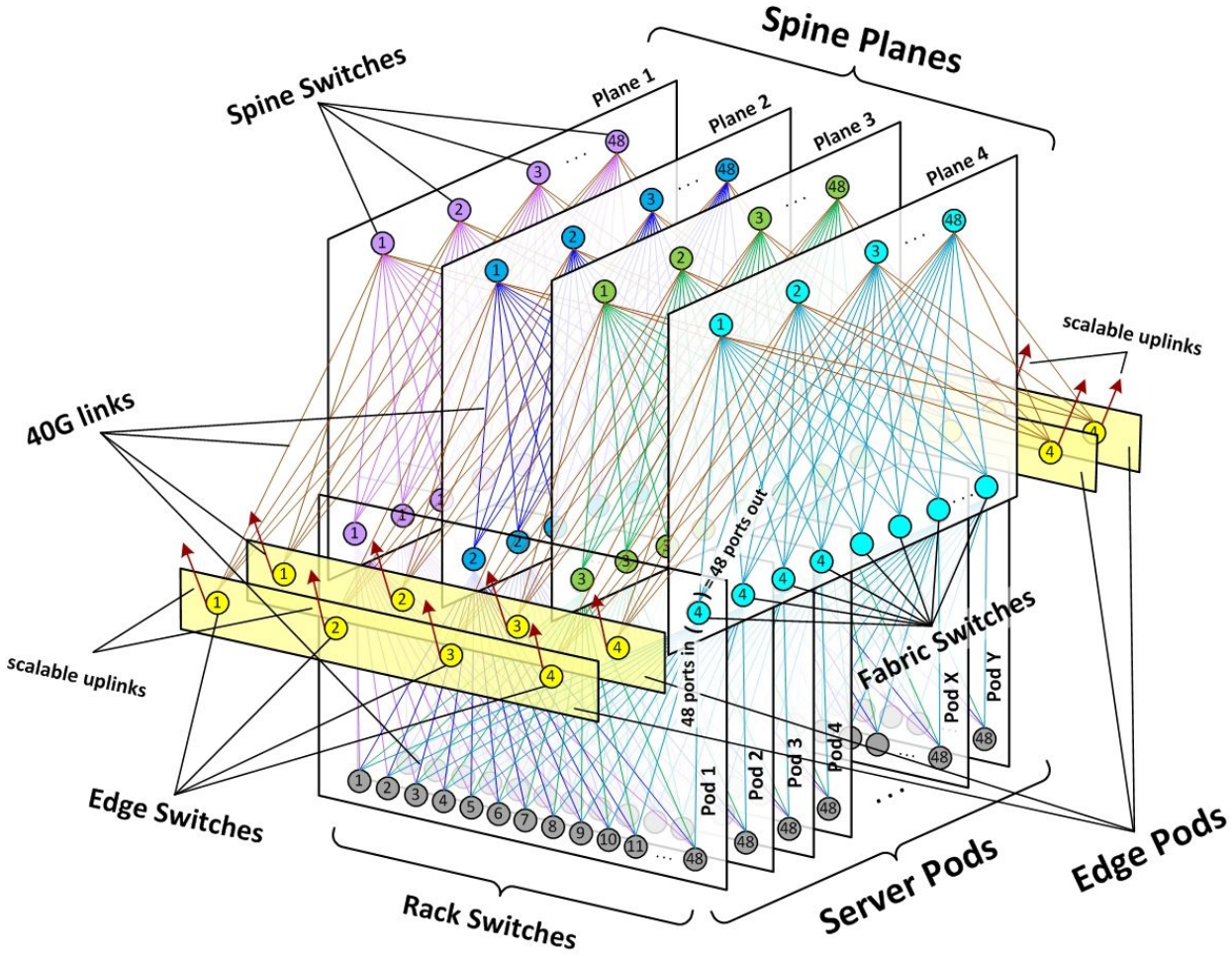


在Facebook的Fabric架构中，spine交换机与多个fabric交换机连接，为多个POD提供连通性。其整体网络架构如下图所示。下图中用三种方式表示了同一种网络架构。最上层是spine交换机，中间是fabric交换机，最下面是TOR。



在Fabric架构中，存在着两个切面，左右切面是一个个POD，前后切面被称为Spine Plane。总共有4个Spine Plane，每个Spine Plane也是一个三级Clos架构。在Spine Plane中，leaf是Fabric交换机，Spine就是Spine交换机。每个Spine Plane中，由48个spine交换机和N个fabric交换机相连组成，N等于当前数据中心接入的POD数。Spine Plane的三级Clos架构和POD的三级Clos架构，共同构成了数据中心的五级Clos架构。因为在POD内，fabric交换机通过48个口与TOR连接，所以在Spine Plane的Clos架构中，fabric交换机的输入输出端口数都是48，对应上面的公式，m1=n3=48。根据Clos架构的特性，在Spine plane中，Spine交换机只要大于等于48个，不论N（POD数）等于多少，都可以保证网络架构无阻塞。当然实际中N还受限于spine交换机的端口密度。

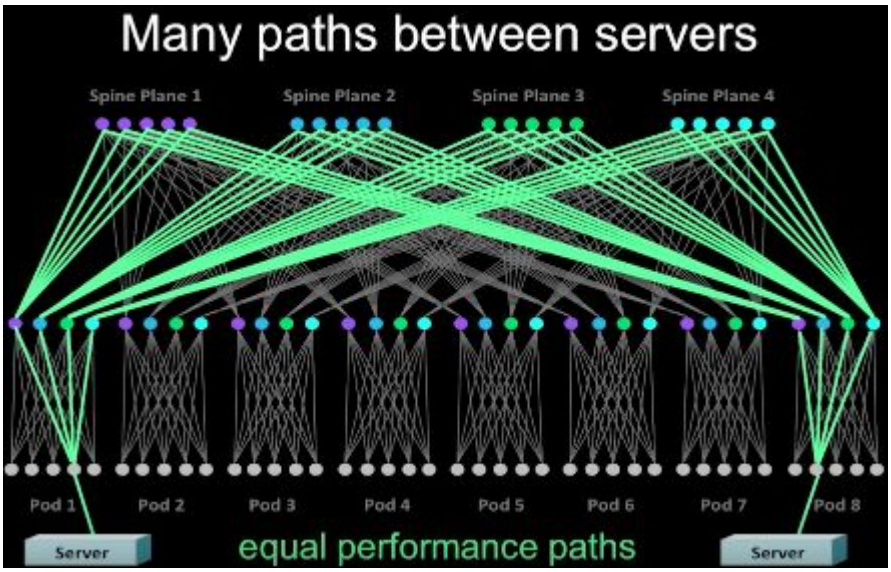
由于每个POD有4个fabric交换机，所以总共有4个Spine Plane。完整的架构如下图所示：



除了前面描述的POD和Spine Plane，上图中还有黄色的Edge Plane，这是为数据中心提供南北向流量的模块。它们与Spine交换机的连接方式，与二层的spine/leaf架构一样。并且它们也是可以水平扩展的。

采用Clos架构的数据中心网络架构的优势：

- 弹性可扩展。数据中心可以以POD为单位构建，随着规模的增加，增加相应的POD即可。在Spine交换机端口数可承受的范围内，增删POD并不需要修改网络架构。
- 模块化设计。不论是POD，Spine Plane还是Edge Plane，都是一个个相同的模块，在水平扩展的时候，不需要新的设计，只是将原有的结构复制一份即可。
- 灵活。当对网络带宽要求不高的时候，Spine交换机和Edge交换机可以适当减少。例如Facebook表示，在数据中心的初期，只提供4：1的东西向流量超占比，这样每个Spine Plane只需要12个Spine交换机。当需要更多带宽时，再增加相应的Spine交换机即可。同样的模式也适用于Edge交换机。这符合“小规模启动，最终适用大规模”的思想。
- 硬件依赖性小。传统三层网络架构中，大的网络规模意味着高端的核心汇聚交换机。但是在Fabric架构中，交换机都是中等交换机，例如所有的fabric交换机只需要96个端口，中等规模的交换机简单，稳定，成本低，并且大多数网络厂商都能制造。
- 高度高可用。传统三层网络架构中，尽管汇聚层和核心层都做了高可用，但是汇聚层的高可用由于是基于STP（Spanning Tree Protocol），并不能充分利用多个交换机的性能，并且，如果所有的汇聚层交换机（一般是两个）都出现故障，那么整个汇聚层POD网络就瘫痪。但是在Fabric架构中，跨POD的两个服务器之间有多条通道（4*48=192），除非192条通道都出现故障，否则网络能一直保持连通，下图是一个跨POD服务器之间多通道示意图。



需要指出的是，这种网络架构并非Facebook独有（是不是原创无从考证），例如Cisco的Massively Scalable Data Center (MSDC)，Brocade的Optimized 5-Stage L3 Clos Topology都是类似的5级Clos架构。其中一些组成元素，各家叫法不一样，不过原理都是类似的。

最后

技术发展的过程中，有一些技术提出，应用，流行，消逝，过了一段时间，在新的领域，被人又重新提出，应用，流行，这本身就是一种非常有意思的现象。Clos架构就是这么一种技术，从最开始的电话交换架构，到交换机内部模型，到现在的网络架构，它在不同的领域解决着同样的问题。

发布于 2017-10-10 08:01