北京郵電大學

报告：CLOS & Fat-Tree & Spine-Leaf

	分工贡献
张晨阳	完成研讨汇报 PPT 的制作、研讨报告的格式调整
金建名	完成研讨报告的所有内容
共同完成	所有参考资料的寻找和收集

2025 年 3 月 31 号

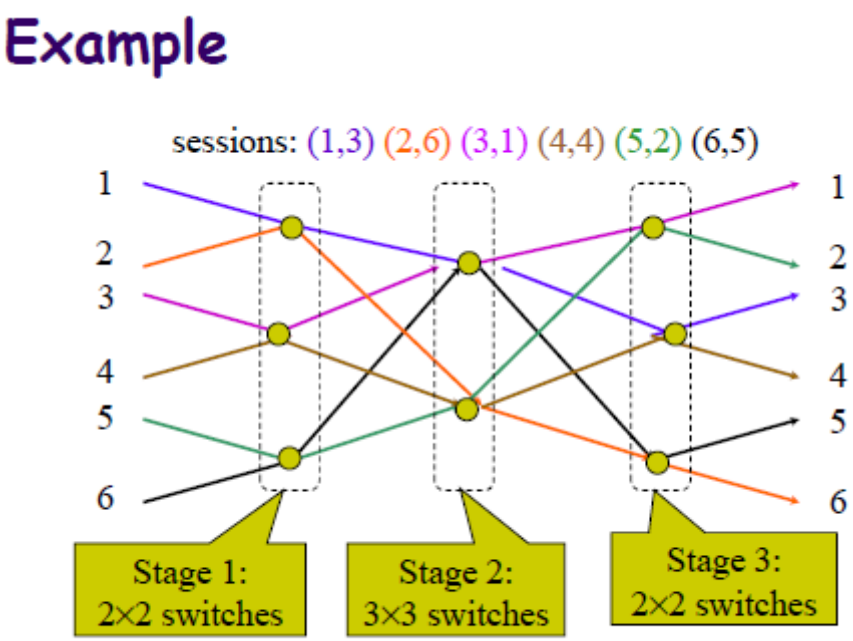
目录

1.什么是 CLOS

随着全球数据量的不断扩大，更大的带宽、更高的速度、更低的延迟逐渐成为用户刚需，这就对底层的数据交换网络抛出了巨大的挑战，我们迫切的需要更灵活、更高性能的交换网络架构。而 CLOS 架构正是一个很好的解决方案。

CLOS 架构是一种多级交换网络拓扑结构，由贝尔实验室的工程师 Charles Clos 在 1950 年代提出。这个架构主要描述了一种多级电路交换网络的结构，它对传统的 Crossbar 结构进行了改进，这使得 CLOS 架构的交换网络可以提供无阻塞的网络。

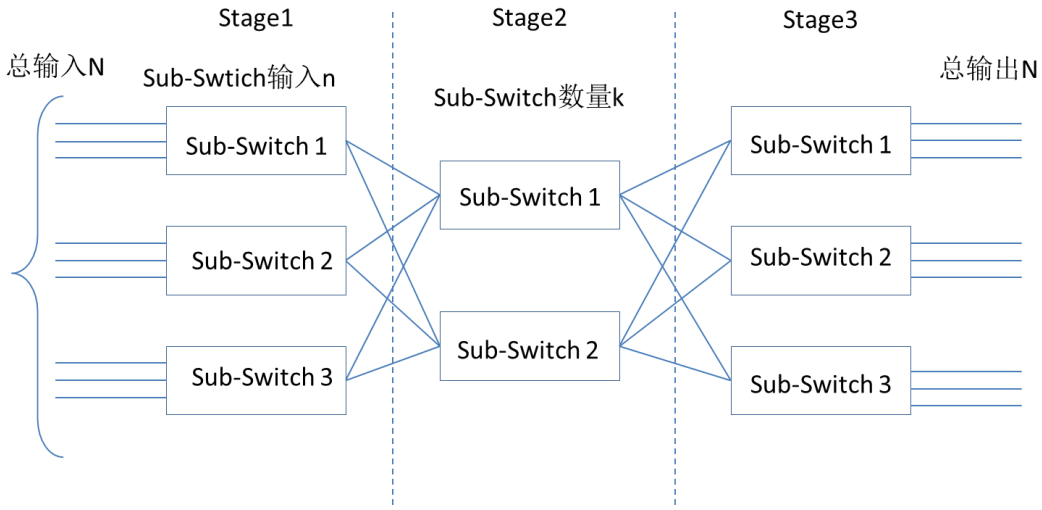
我们来看一个 CLOS 架构的例子：



在两个 stage 之间，复数的连接将两个 stage 的节点全连接，确保每一个输入口的信号都可以通过每一个输出口。中间的节点则是传统的 n*n 的交换单元。

让我们来分析这种结构的优势：

先来计算其所有的节点个数：



公式 Crosspoint = $2\frac{N}{n} * nk + k(\frac{N}{n})^2 = 2Nk+k\frac{N^2}{n^2}$

结果如图所示，我们计算总共 N=20 个输入，stage1 上 sub-switch 上承担 n=10 个输入，stage2 采用 k=3 个 sub-switch 的情况，我们共需要 crosspoint 为 136 个，而采用 Clossbar 需要 N*N，即 400 个 crosspoint。取一个 N 相当大的情况，经过 stage1 承担的输出，到达是 stage2 的线数仍然很多，对于这个 subswitch，我们仍旧可以采用 CLOS 的架构继续替换，这就可以使交换单元的个数大大降低。

1. 每一个 Session 都有冗余的链路。

2. 任何输入都可以找到没有在同时使用的线路，故也叫做无阻塞架构。

总结：

CLOS 的特点有：1.多级交换，最常见的时三级交换架构；2.到指定目的地，在第 1 级交换单元存在多条路由，而后续交换单元都只存在唯一的一条路由；3.严格意义上的无阻塞；4.支持递归，可以无限拓展。

2.胖树和叶脊介绍

2.1. 胖树架构介绍

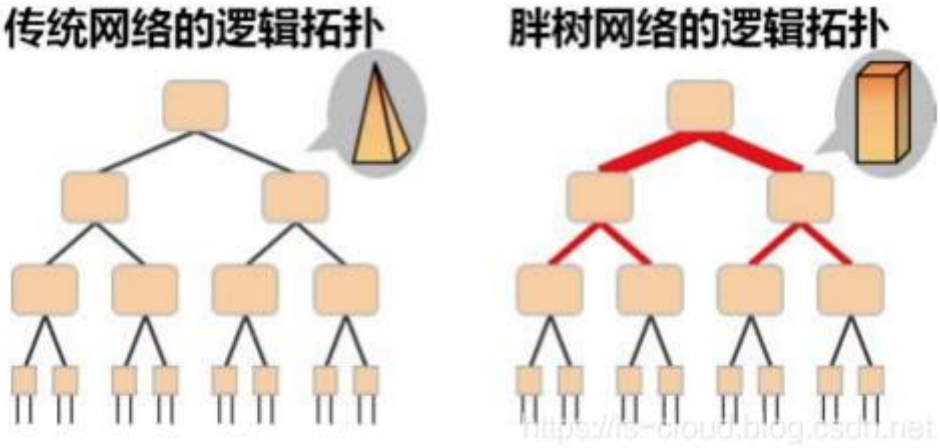
2000 年之后，互联网从经济危机中复苏，以谷歌和亚马逊为代表的互联网巨头开始崛起。他们开始推行云计算技术，建设大量的数据中心（IDC），甚至超级数据中心。

面对日益庞大的计算规模，传统树型网络肯定是不行的了。于是，一种改进型树型网络开始出现，它就是**胖树（Fat-Tree）架构**。

胖树（Fat-Tree）就是一种 CLOS 网络架构。

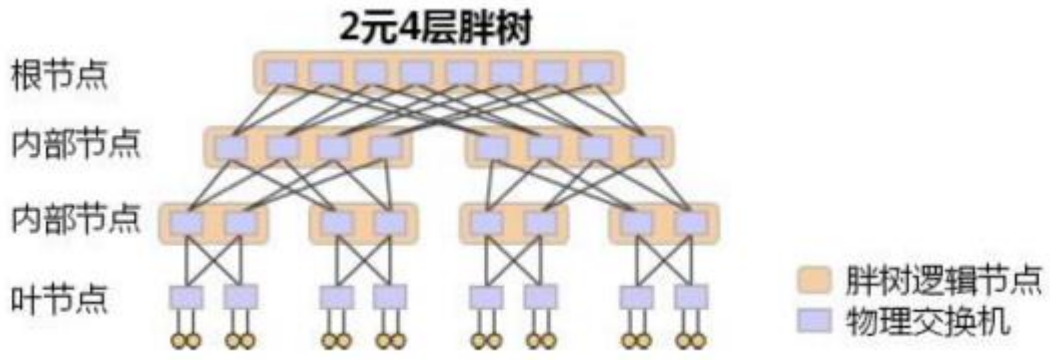
胖树架构的基本理念是：使用大量的低性能交换机，构建出大规模的无阻塞网络。对于任意的通信模式，总有路径让他们的通信带宽达到网卡带宽。

相比于传统树型，胖树（Fat-Tree）更像是真实的树，越到树根，枝干越粗。从叶子到树根，网络带宽不收敛。这是 Fat-Tree 能够支撑无阻塞网络的基础。



如上图所示，为了实现网络带宽的无收敛，Fat-Tree 中的每个节点（根节点除外）都需要保证上行带宽和下行带宽相等，并且每个节点都要提供对接入带宽的线速转发的能力。

下图是一个 2 元 4 层 Fat-Tree 的物理结构示例（2 元：每个叶子交换机接入 2 台终端；4 层：网络中的交换机分为 4 层），其使用的所有物理交换机都是完全相同的。



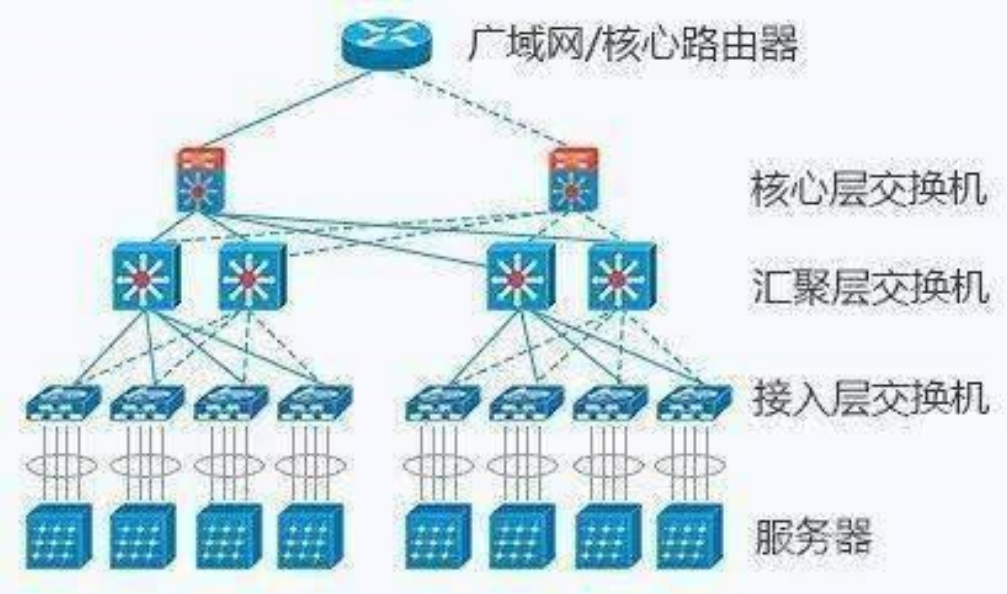
假设一个 k-ary（每个节点有不超过 k 个子节点）的三层 Fat-Tree 拓扑：

- 核心交换机个数 (k/2)^2
- POD 个数 k
- 每个 POD 汇聚交换机 k/2
- 每个 POD 接口交换机 k/2
- 每个接入交换机连接的终端服务器 k/2
- 每个接入交换机剩余 k/2 个口连接 POD 内 k/2 个汇聚交换机，每台核心交换机的第 i 个端口连接到第 i 个 POD，所有交换机均采用 k-port switch。

可以计算出，支持的服务器个数为 $k*(k/2)*(k/2)=(k^3)/4$，不同 POD 下服务器间等价路径数 $(k/2)*(k/2)=(k^2)/4$。

上图为最简单的 k=4 时的 Fat-Tree 拓扑，**连在同一个接入交换机下的服务器处于同一个子网，他们之间的通信走二层报文交换。不同接入交换机下的服务器通信，需要走路由**。

胖树架构被引入到数据中心之后，数据中心变成了传统的三层结构：



接入层：用于连接所有的计算节点。通常以机柜交换机（TOR，Top of Rack，柜顶交换机）的形式存在。

汇聚层：用于接入层的互联，并作为该汇聚区域二三层的边界。各种防火墙、负载均衡等业务也部署于此。

核心层：用于汇聚层的的互联，并实现整个数据中心与外部网络的三层通信。

项目	网络设计分层	连接对象	性能要求	目的	路由功能
核心层交换机	三级	三层以上交换设备	最高	路由和高速转发	支持
汇聚层交换机	二层/三层	交换机和路由器	中等	提供策略连接	支持
接入层交换机	二层	电脑	最低	终端接入	支持

2.2. Fat-Tree 的缺陷

Fat-Tree 的扩展规模在理论上受限于核心层交换机的端口数目，不利于数据中心的长期发展要求；

对于 POD 内部，Fat-Tree 容错性能差，对底层交换设备故障非常敏感，当底层交换设备故障时，难以保证服务质量；

Fat-Tree 拓扑结构的特点决定了网络不能很好的支持 One-to-All 及 All-to-All 网络通信模式，不利于部署 MapReduce、Dryad 等高性能分布式应用；

Fat-Tree 网络中交换机与服务器的比值较大，在一定程度上使得网络设备成本依然很高，不利于企业的经济发展。

因为要防止出现 TCP 报文乱序的问题，难以达到 1:1 的超分比。

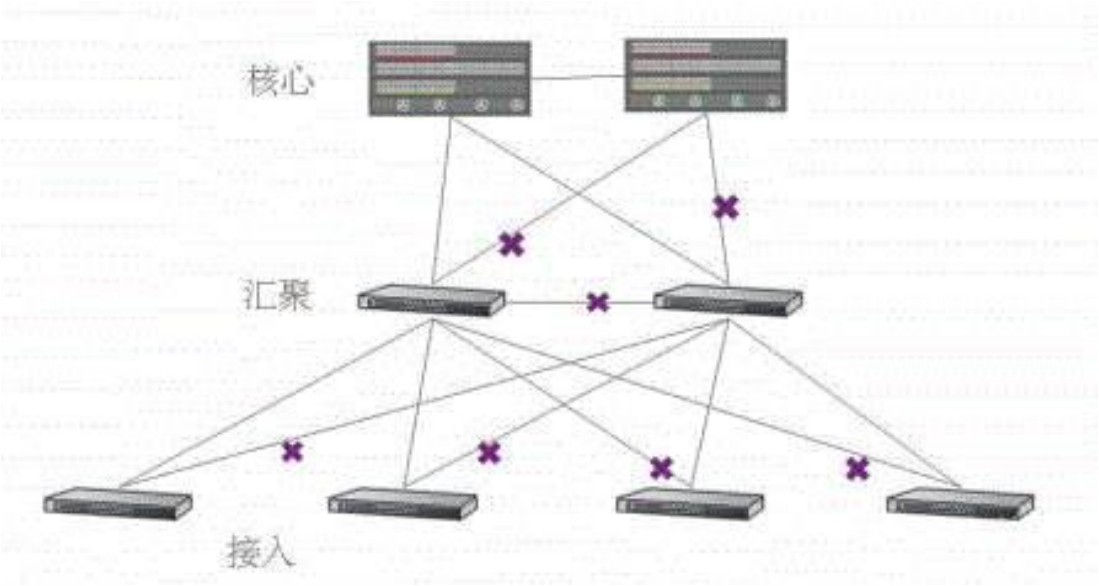
2.3. 叶脊架构介绍

在很长的一段时间里，三层网络结构在数据中心十分盛行。在这种架构中，铜缆布线是主要的布线方式，使用率达到了 80%。而光缆，只占了 20%。

然而，在三层网络结构的应用中逐渐出现了一些问题：

1. 资源浪费

传统三层结构中，一台下层交换机会通过两条链路与两台上层交换机互连。



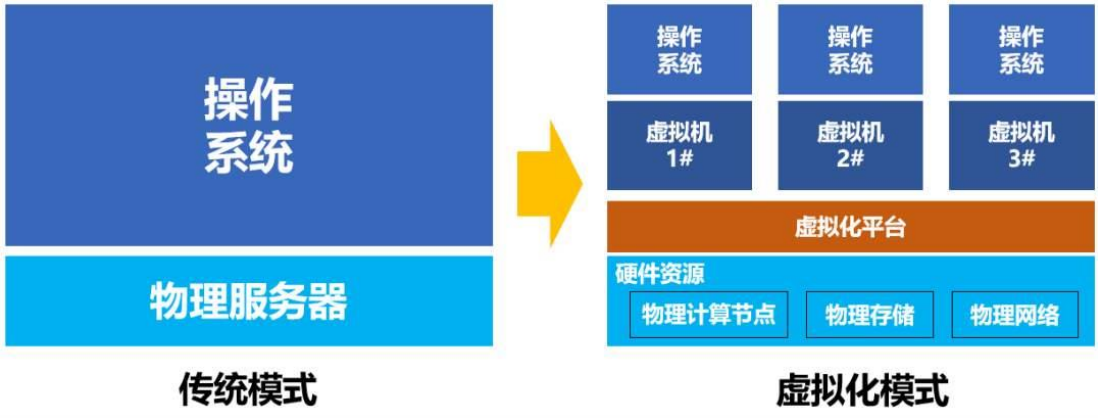
由于采用的是 STP 协议（ Spanning Tree Protocol，生成树协议），实际承载流量的只有一条。其它上行链路，是被阻塞的（只用于备份）。这就造成了带宽的浪费。

2. 故障域比较大

STP 协议由于其本身的算法，在网络拓扑发生变更时需要重新收敛，容易发生故障，从而影响整个 VLAN 的网络。

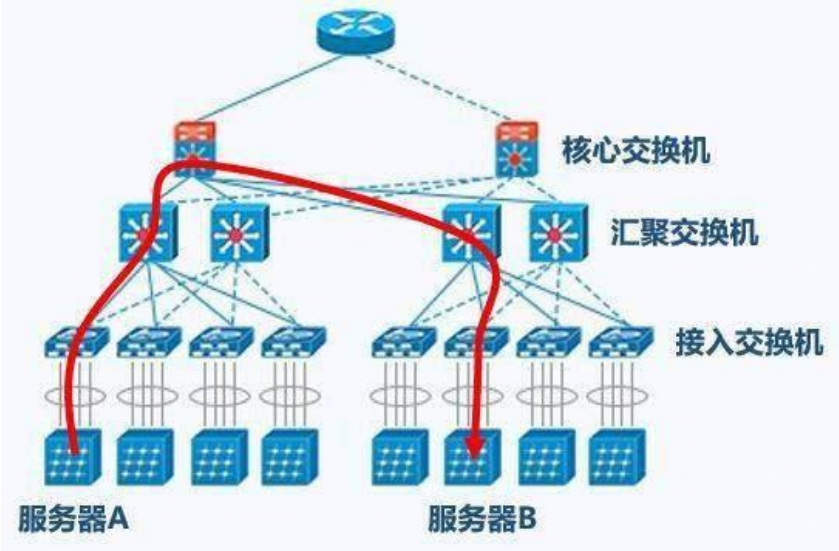
3. 数据中心流量走向发生了巨大的变化

2010 年之后，为了提高计算和存储资源的利用率，所有的数据中心都开始采用**虚拟化技术**。网络中开始出现了大量的虚拟机（VM，Virtual Machine）。



与此同时，微服务架构开始流行，很多软件开始推行功能解耦，单个服务变成了多个服务，部署在不同的虚拟机上。虚拟机之间的流量，大幅增加。

这种数据流量的大幅增加，给传统三层架构带来了很大的麻烦——因为服务器和服务器之间的通信，需要经过接入交换机、汇聚交换机和核心交换机。

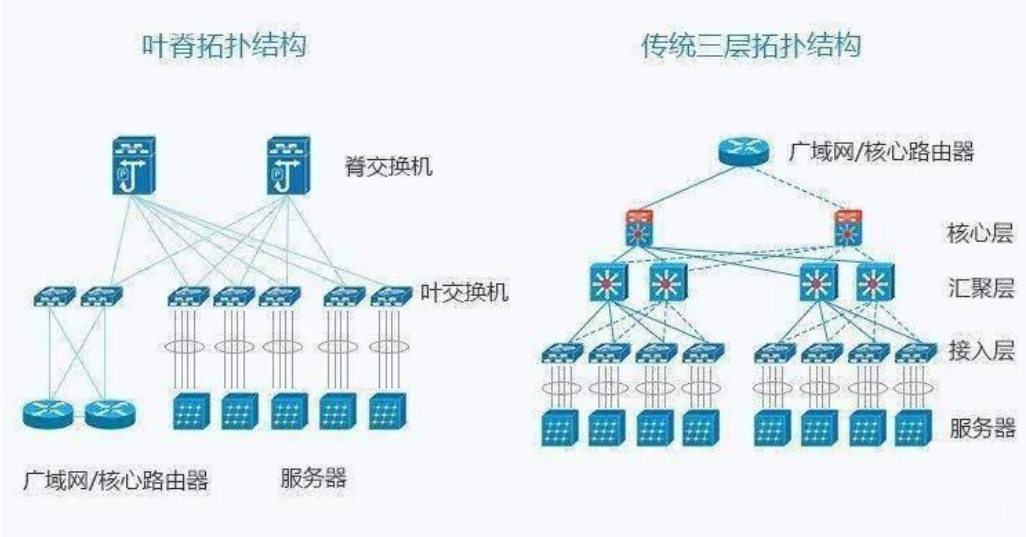


这意味着，核心交换机和汇聚交换机的工作压力不断增加。要支持大规模的网络，就必须有性能最好、端口密度最大的汇聚层核心层设备。这样的设备成本高，价格非常昂贵。并且难以应对实时膨胀的数据量。

因此，网络工程师们提出了**“Spine-Leaf 网络架构”，也就是叶脊网络**（有时候也被称为脊叶网络）

Spine-Leaf 网络架构，也称为分布式核心网络，由于这种网络架构来源于交换机内部的 Switch Fabric，因此也被称为 Fabric 网络架构，同属于 CLOS 网络模型。事实已经证明，Spine-Leaf 网络架构可以提供高带宽、低延迟、非阻塞的服务器到服务器连接。

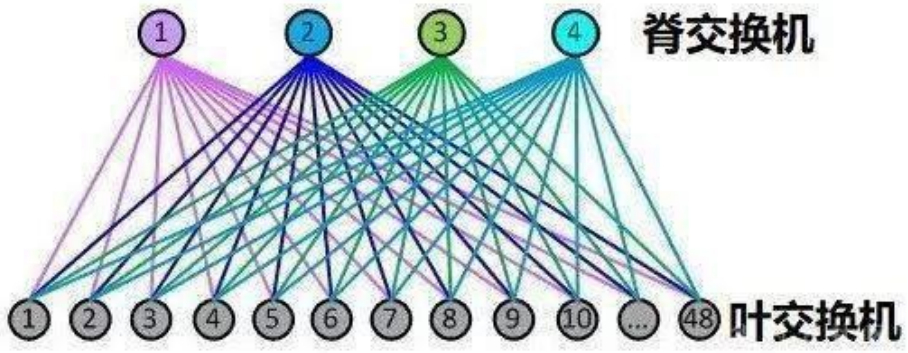
相比于传统网络的三层架构，叶脊网络进行了扁平化，变成了两层架构。如下图所示：



叶交换机，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。叶交换机之上是三层网络，之下都是个独立的 L2 广播域。如果说两个叶交换机下的服务器需要通信，需要经由脊交换机进行转发。

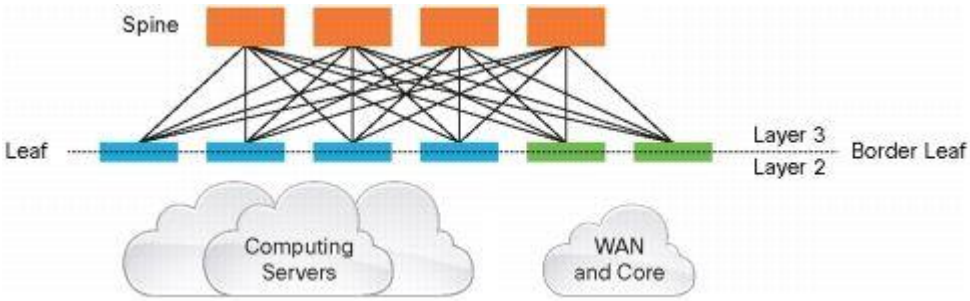
脊交换机，相当于核心交换机。叶和脊交换机之间通过 ECMP（Equal Cost Multi Path）动态选择多条路径。

脊交换机下行端口数量，决定了叶交换机的数量。而叶交换机上行端口数量，决定了脊交换机的数量。它们共同决定了叶脊网络的规模。

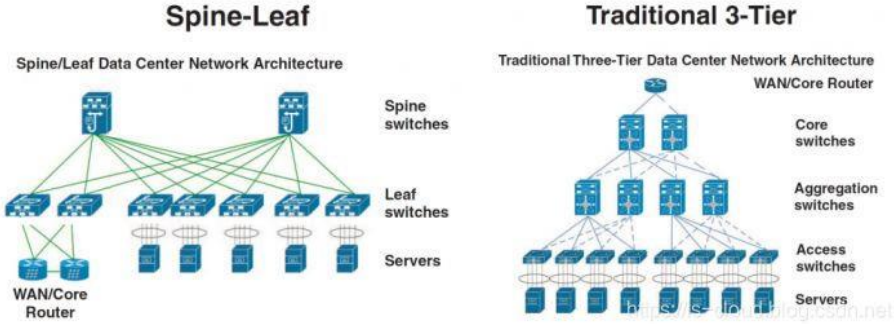


2.4. Spine-Leaf 的工作原理

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，Leaf 交换机之下都是个独立的 L2 广播域，这就解决了大二层网络的 BUM 问题。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 交换机进行转发。



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 交换机并行的交机（edge switch）再接到 WAN router 出去。



Fabric 中的 Leaf 层由接入交换机组成，用于接入服务器，Spine 层是网络的骨干（Backbone），负责将所有的 Leaf 连接起来。每个低层级的 Leaf 交换机都会连接到每个高层级的 Spine 交换机上，即每个 Leaf 交换机的上行链路数等于 Spine 交换机数量，同样，每个 Spine 交换机的下行链路数等于 Leaf 交换机的数量，形成一个 Full-Mesh 拓扑。当 Leaf 层的接入端口和上行链路都没有瓶颈时，这个架构就实现了**无阻塞（Nonblocking）**。并且，因为任意跨 Leaf 的两台服务器的连接，都会经过相同数量的设备，所以保证了**延迟是可预测的**，因为一个包只需要经过一个 Spine 和另一个 Leaf 就可以到达目的端。

因为 Fabric 中的每个 Leaf 都会连接到每个 Spine，所以，如果一个 Spine 挂了，数据中心的吞吐性能只会有轻微的下降（Slightly Degrade）。如果某个链路的流量被打满了，Spline-Leaf 的扩容过程也很简单：添加一个 Spine 交换机就可以扩展每个 Leaf 的上行链路，增大了 Leaf 和 Spine 之间的带宽，缓解了链路被打爆的问题。如果接入层的端口数量成为了瓶颈，那就直接添加一个新的 Leaf，然后将其连接到每个 Spine 并做相应的配置即可。这种易于扩展（Ease of Expansion）的特性优化了 IT 部门扩展网络的过程。

由此可见，叶脊网络的优势非常明显：

1、带宽利用率高

每个叶交换机的上行链路，以负载均衡方式工作，充分的利用了带宽。

2、网络延迟可预测

在以上模型中，叶交换机之间的连通路径的条数可确定，均只需经过一个脊交换机，东西向网络延时可预测。

3、扩展性好

当带宽不足时，增加脊交换机数量，可水平扩展带宽。当服务器数量增加时，增加脊交换机数量，也可以扩大数据中心规模。总之，规划和扩容非常方便。

4、降低对交换机的要求

南北向流量，可以从叶节点出去，也可从脊节点出去。东西向流量，分布在多条路径上。这样一来，不需要昂贵的高性能高带宽交换机。

5、安全性和可用性高

传统网络采用 STP 协议，当一台设备故障时就会重新收敛，影响网络性能甚至发生故障。叶脊架构中，一台设备故障时，不需重新收敛，流量继续在其他正常路径上通过，网络连通性不受影响，带宽也只减少一条路径的带宽，性能影响微乎其微。

但是，Fabric 架构并非完美。叶子节点网络设备无论是性能要求还是功能要求，均高于传统架构下的接入设备，其作为各种类型

的网关（二三层间、VLAN/VxLAN 间、VxLAN/NVGRE 间、FC/IP 间等等），芯片处理能力要求较高，目前尚无满足所有协议间互通的商用芯片；由于不存在相关的标准，为了实现各种类型网络的接入，其骨干节点与叶子节点间的转发各个厂商均采用了私有封装，这也为将来的互通设置了难题。除此之外，还有：

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。

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。

3.CLOS、胖树以及叶脊架构的对比

3.1. 架构设计

架构类型	设计特点
CLOS 架构	三级或多级交换架构，包含输入级、中间级和输出级，每级之间通过全连接实现无阻塞。
胖树架构	三级 CLOS 架构的变种，通常包含接入层、汇聚层和核心层，带宽不收敛，但通过增加交换机数量实现无阻塞。
叶脊架构	两层架构，由叶交换机（Leaf）和脊交换机（Spine）组成，叶交换机连接服务器，脊交换机作为网络核心。

3.2. 优势

架构类型	优势
CLOS 架构	严格无阻塞，支持递归扩展，适用于大规模网络。
胖树架构	使用标准化交换机，成本较低，能够实现无阻塞网络。
叶脊架构	高带宽、低延迟、易于扩展，适合现代数据中心的高吞吐量需求。

3.3. 缺点

架构类型	缺点
CLOS 架构	设计复杂，成本较高，需要更多的交换机和连接。
胖树架构	扩展性受限于核心交换机端口数量，容错性较差，对分布式应用支持不足。
叶脊架构	叶交换机需要支持 L2/L3 功能，对设备性能要求较高。

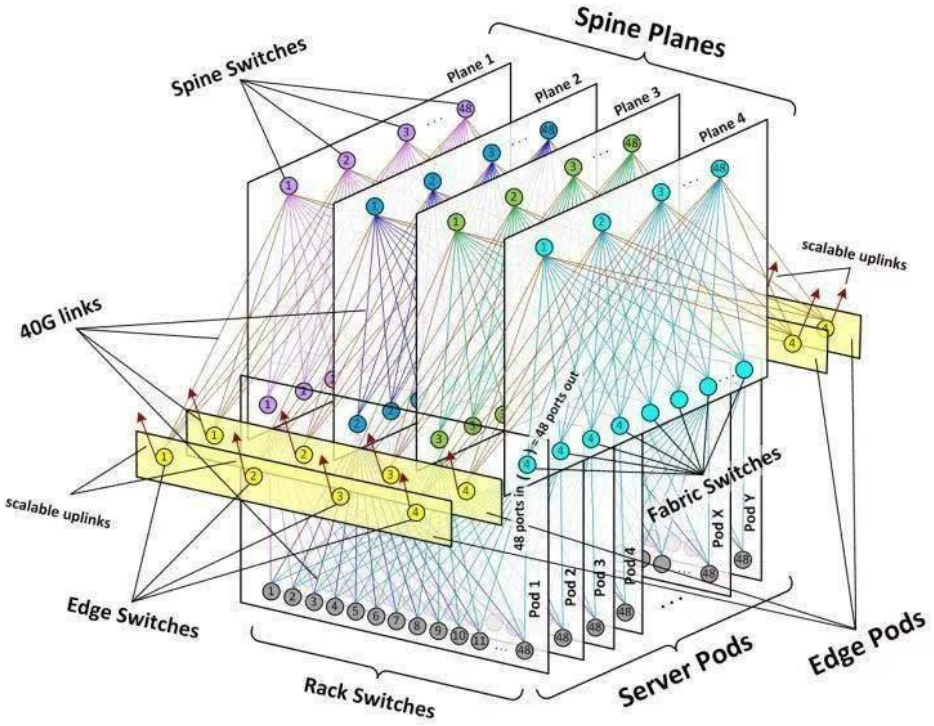
3.4. 适用场景

架构类型	使用场景
CLOS 架构	大规模数据中心、高性能计算集群。
胖树架构	早期大规模数据中心、对成本敏感的场景。
叶脊架构	现代数据中心、云计算环境。

4.架构的具体应用

最具有代表性的，是 Facebook 在 2014 年公开的数据中心架构。

Facebook 使用了一个五级 CLOS 架构：



除了 Facebook 之外，谷歌公司的第五代数据中心架构 Jupiter 也大规模采用了叶脊网络，其可以支持的网络带宽已经达到 Pbps 级。谷歌数据中心中 10 万台服务器的每一个，都可以用任意模式以每秒 10 千兆比特的速度互相通信。

除此之外，中国电信提出的实现固移融合的新型城域网借鉴了数据中心采用的叶脊网络架构。