## A Scalable, Commodity Data Center Network Architecture

传统的数据中心网络架构采用树形拓扑，即使是采用高端的网络设备，顶层的设备都难以支撑网络边缘的数据总流量。面临着流量的非均匀化，软件协议设计的复杂性，使得成本大大提高。在这篇论文中，作者展示了利用传统的网络交换设备并结合胖树网络架构可以换来高性能的传输效率，并且完全兼容现有的网络协议和硬件接口。

在以往的数据中心里，通常具备如下几个特点：

1. 数据中心有成千上万的计算机构成，且带宽需求量巨大。
2. 网络架构由路由和交换机组成树形结构，越来越多的高端设备的加入使得网络层次结构越发的明晰。
3. 即使顶层采用最高端的网络设备也只能支撑起整个网络50%的带宽。
4. 流量大小的不一致性导致传统的算法很难适配数据中心，或者说需要设计很复杂的算法。

而在新型的数据中心架构中，应当具备的特点是：

1. 通过大规模使用传统的网络设备就支撑起大量的网络流量。
2. 采取新型架构可以以较低的成本换取高性能。
3. 完全向后兼容，不必更改现有网络设备和协议。

**问题分析：**

商用集群的出现使得任何一个组织都可以轻松实现大量的计算和PB级的存储，但该网络的关键瓶颈在于节点间的通讯，譬如在web请求、云计算、并行计算等诸多应用中都存在这样的问题。

作者认为解决这样问题的思路是，一是通过利用新型互联硬件和软件，如infiniBand，Myrinet，实现成本高，且无法兼容现有网络协议，因此在现有网络中进行改进的可扩展性不强。二是采用传统的网络设备和协议，但随着网络节点数量的增加，管理成本和利用率、可扩展性效果都变得极差，这在现行的ip网络下也是如此。本文采用的是第二种方案，此时，网络整体的性能受限于核心层节点的处理能力，因此需要从网络架构和设备、软件协议方面入手进行适当的改进。

**设计目标：**

1. 具有良好的可扩展性，且任意加入主机都能够以本地连接处的全部带宽与其它主机进行通信。
2. 廉价性，能够利用传统设备集群实现大规模部署。
3. 向后兼容，新的架构能够在不修改网络接口和软件协议的基础上与现有网络兼容。

**背景知识介绍：**

拓扑结构，本文数据中心采用三层网络架构，包括核心层、汇聚层和边缘层，也有的数据中心采用二层网络架构，相比二层架构，三层能够支持更多的接入主机，但线路较为复杂，且管理成本也更加高昂。图1所示是数据中心网络的三层架构。

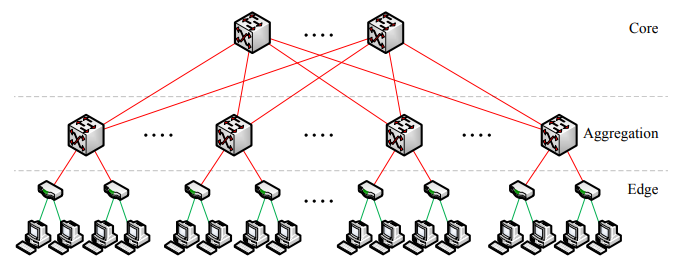


图1 数据中心网络的三层架构

超额订阅，通常理想的情况下，超额订阅应当为1：1，这说明全部的主机都能以本地接口处的全部带宽接入到网络，这是一个衡量网络传输性能的一个很重要的指标，如果超额订阅为5：1，则说明数据中心里只有20%的接入主机可以达到全速率传输。

等价路由，数据中心具备多条冗余链路，使得其容错及负载性都有了较大提升，通常实现等价路由的关键技术为ECMP，可以使得任意主机间采取不同的传输路径，即使是同一个流。然而ECMP基于静态划分，没有动态运行时监测做调控，使得多条链路负载很大，难以实现较高的超额订阅比，此外，ECMP提供的等价路由数较少，无法满足数据中心网络的需要，随着路由数的增多，路由表项的个数也将成倍的增加。

花销，不光需要考虑网络架构部署时的开销，还应当考虑为实现不同超额订阅比所需要增加的成本，其中包括冗余链路的成本，等价路由计算的成本，图2所示是不同超额订阅比所造成的数据中心网络整体的开销。可以看到，最理想的超额订阅比1：1，是所有方案中耗费成本最高的一个。

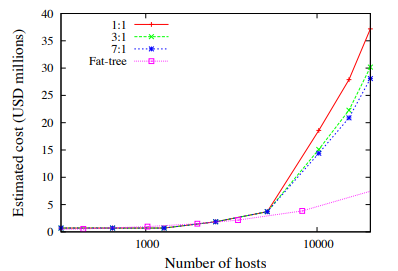


图2 不同超额订阅比所带来的网络整体开销

**网络设计架构：**

网络的设计架构，包含网络本身的拓扑结构设计，ip地址划分，路由表结构设计，路由算法的设计。这些都是为了追求数据中心网络链路的充分利用率和超额订阅所需要详细考虑的设计实现因素，除此之外，作者也提到为了实现某一目标，还综合考虑了其他诸多方案，譬如为实现等价路由的选取，提出了流分类和流调度技术。

首先来看数据中心网络所采用的网络拓扑结构。本文提出的数据中心网络大致采用的是三层结构，包括顶部的核心层，中间的汇聚层，以及最底下的边缘层，详情如图3所示。图中含有4个pod，称之为4-ary的胖树结构，每个pod所管辖的区域属于同一个网段，这样，由于有中间汇聚层的多个交换机做冗余，因此可以实现多条等价路由的选择，如果数据中心网络包含k个pods，则两台主机间就有k/2条路径。

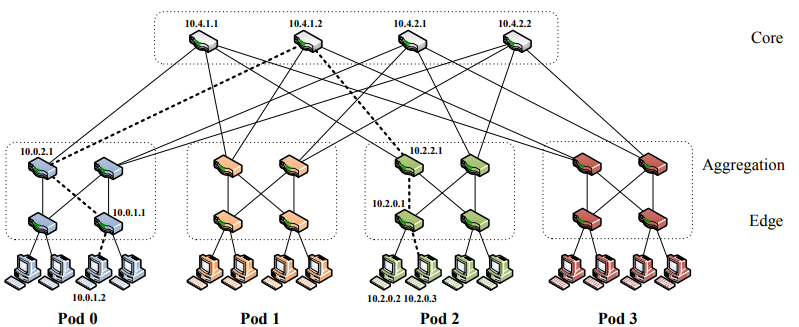


图3 k-ary胖树架构

说完网络的拓扑结构，紧接着来分析一下地址是如何分配的。汇聚层中第二个byte代表pod数，第三个byte代表哪一个交换机，边缘层，根据所连接的交换机，其第四个byte依次+1。这种ip地址的划分虽然造成了部分浪费，但也足以够数据中网络节点主机的分配了，另外，这种分配方式有利于后面谈到的两级路由表的实现，方便了路由的计算。

通常数据中心设计的难点基本都是围绕如何划分多条等价路由，以实现多条链路上的流量均匀分配。这里作者首先提到了传统的OSPF算法，这是一种基于跳数选择的“最短路径”，然而基于K-ary的胖树结构本来拥有k/2个最短路径，但却只选择了一个，网络流被集中到一个端口上，无法利用等价路由和改善超额订阅。而OSPF-ECMP改进版则会导致路由表项数量的增加。最终提出了基于两级路由查找的算法来充分利用网络带宽。

两级路由表包含两层机构，一层是前缀表项，决定路由的南向出口，一层是后缀表项，决定路由的北向出口，后缀表项通常包含一个指向前缀表项的指针，以指向更小层次的路由，当路由没有在本pod下查找到，则要通过两级路由表决定下一个转发路径。

然后就是路由算法的实现，胖树结构中，边缘层和汇聚层都含有指向该机架中的子网前缀列表，所以一个子网中的数据包如果被发送到另一个机架中但位于不同子网的主机上，则该机架中所有上层的交换机都将有一个指向目标子网的前缀表项。对于所有其他往上传输的pod的间通信，pod交换机有一个默认的/0前缀，带有一个匹配主机id的辅助表(目标IP地址的最低有效字节)。这将导致流量均匀地向上分布到核心交换机。

**提升传输效率的其他措施**

基于两级路由转发的方案有一个问题，即如果两个流的源ip和目的ip相同，则会将其分配到通过一端口进行输出，这会导致端口竞争，从而无法发挥冗余链路的优势。但如果完全的动态路由，将数据流引导至不同交换机进行转发，则又会出现数据包重排序的问题，所以流分配，就是将不同流拆分到不同交换机上进行传输，而同一个流则始终在相同路径上传输，这样，即实现了多条等价路由的分配，与此同时，也保证了同一个流的数据总是经过相同的交换机进行转发，从而避免后续的包进行重排序。

一些研究表明，互联网流量的分布具有少数大的长生命流和许多小的短生命流的特征。可以认为，大流量在确定网络里实现流量的均分可以显著提升网络整体的性能，因此值得特殊处理。在流调度中，针对大型流进行调度，以尽量减少彼此之间的重叠。中心调度器利用对网络中交换机上传的元信息，判断是否有新增的大流量，平判断是否分配一条独占的路径。

具体实现是，边缘开关在本地分配一个新的占用最少的端口给新增的流。同时，边缘开关还会检测大小超过预定义阈值的任何传出的流，并向中心调度器发送通知，以请求将该流放置在一条非竞争的路径中。注意，边缘交换机只能请求分配路径，但无法私自决断使用哪条路径。中心调度器会跟踪所有的请求，并尽可能为其分配非冲突的路径，同时维护一个布尔变量，以指示该路径是否已被一条大流量所占用。

**网络测验：**

测量目标包括交换机、流分类、流调度器，同时基于3.6：1的超额订阅与现有的数据中心进行比较。利用的是3层胖树结构，有四台机器，每台机器下连四个主机，四台机器运行四个pod交换机，每个交换机有一个上行链路。四个pod开关连接到一个在专用机器上运行的4端口核心开关。为了在从pod交换机到核心交换机的上行链路上实现3.6:1的超额订阅，这些链路的带宽限制为106.67Mbit/s，而所有其他链路限制为96Mbit/s。

测试了五种通信样例，并记录了不同样例下的拓扑结构、二级路由表、流分类和流调度所贡献的效率比，这五种测试样例如下：

1. 随机通信对，pod内任意主机发送数据到其他pod内的任意主机。
2. 固定通信对，pod内x编号的主机发送数据到其他pod内编号为(x+i)%16的主机。
3. 概率通信 (SubnetP, PodP)，即以podP的概率向SubnetP内的主机发送数据，以1-podP的概率向其他子网中发送数据。
4. 多个pod发送数据到同一个子网中的不同主机，该情况下，最坏的超额订阅比为(k-1)：1。
5. 同一子网不同主机发送数据到其他pod的主机，该情况下静态路由使得网络流竞争到同一个端口上，这是，流调度可以最大限度的改善着这种情况。

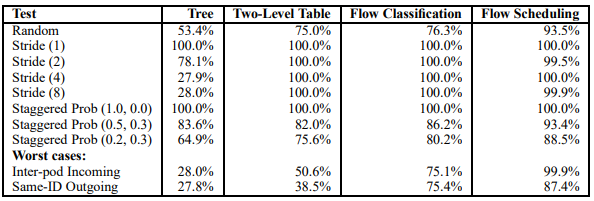


图4 不同测试样例下的网络测试结果

图4显示了实验测试结果，可见当同一子网主机向外发送数据产生链路竞争时，流调度总能很好的完成工作，而二级路由表提供的是多条等价路由，因此在相同源ip和目的ip的情况下，起到了链路负载的作用，流分类更多提供的是一种避免重排序的作用，所以其提供的收益是始终存在的。在多个pod向同一子网发送数据的情况下，最终达到了28%的链路利用率，这也验证了3.6：1的超额订阅比情况。