由器下面，有3层交换机。每台接入路由器与一台顶层交换机相连，每台顶层交换机与多台二层交换机以及一台负载均衡器相连。每台二层交换机又通过机架的TOR交换机（第三层交换机）与多个机架相连。所有链路通常使用以太网作为链路层和物理层协议，并混合使用铜缆和光缆。通过这种等级式设计，可以将数据中心扩展到几十万台主机的规模。

因为云应用提供商持续地提供高可用性的应用是至关重要的，所以数据中心在它们的设计中也包含了冗余网络设备和冗余链路（在图5-30中没有显示出来）。例如，每台TOR交换机能够与两台二层交换机相连，每台接入路由器、一层交换机和二层交换机可以冗余并集成到设计中［Cisco 2012；Greenberg 2009b］。在图5-30中的等级设计可以看到，每台接入路由器下的这些主机构成了单一子网。为了使ARP广播流量本地化，这些子网的每个都被进一步划分为更小的VLAN子网，每个由数百台主机组成［Greenberg 2009a］。

尽管刚才描述的传统的等级体系结构解决了扩展性问题，但是依然存在主机到主机容量受限的问题［Greenberg 2009b］。为了理解这种限制，重新考虑图5-30，并且假设每台主机用1Gbps链路连接到它的TOR交换机，而交换机间的链路是10Gbps的以太网链路。在相同机架中的两台主机总是能够以1Gbps全速通信，而只受限于主机网络接口卡的速率。然而，如果在数据中心网络中同时存在多条并发流，则不同机架上的两台主机间的最大速率会小得多。为了深入理解这个问题，考虑不同机架上的40对不同主机间的40条并发流的情况。具体来说，假设图5-30中机架1上10台主机都向机架5上对应的主机发送一条流。类似地，在机架2和机架6的主机对上有10条并发流，机架3和机架7间有10条并发流，机架4和机架8间也有10条并发流。如果每一条流和其他流经同一条链路的流平均地共享链路容量，则经过10Gbps的A到B链路（以及10Gbps的B到C链路）的40条流中每条流获得的速率为10Gbps/40 = 250Mbps，显著小于1Gbps的网络接口卡速率。如果主机间的流量需要穿过该等级结构的更高层，这个问题会变得更加严重。对这个限制的一种可行的解决方案是部署更高速率的交换机和路由器。但是这会大大增加数据中心的费用，因为具有高接口速率的交换机和路由器是非常昂贵的。

因为数据中心的一个关键需求是放置计算和服务的灵活性，所以支持主机到主机的高带宽通信十分重要［Greenberg 200b；Farrington 2010］。例如，一个大规模的因特网搜索引擎可能运行在跨越多个机架的上千台主机上，在所有主机对之间具有极高的带宽要求。类似地，像EC2这样的云计算服务可能希望将构成用户服务的多台虚拟机运行在具有最大容量的物理主机上，而无需考虑它们在数据中心的位置。如果这些物理主机跨越了多个机架，前面描述的网络瓶颈可能会导致性能不佳。

5.6.3 数据中心网络的发展趋势

为了降低数据中心的费用，同时提高其在时延和吞吐量上的性能，因特网云服务巨头如谷歌、脸谱、亚马逊和微软都在不断地部署新的数据中心网络设计方案。尽管这些设计方案都是专有的，但是许多重要的趋势是一样的。

其中的一个趋势是部署能够克服传统等级设计缺陷的新型互联体系结构和网络协议。一种方法是采用**全连接拓扑**（fully connected topology）来替代交换机和路由器的等级结构［Al-Fares 2008；Greenberg 2009b；Guo 2009］，图5-31中显示了这种拓扑。在这种设计中，每台第一层交换机都与所有第二层交换机相连，因此：①主机到主机的流量绝不会超过该交换机层次；②对于n台第一层交换机，在任意两台二层交换机间有n条不相交的路