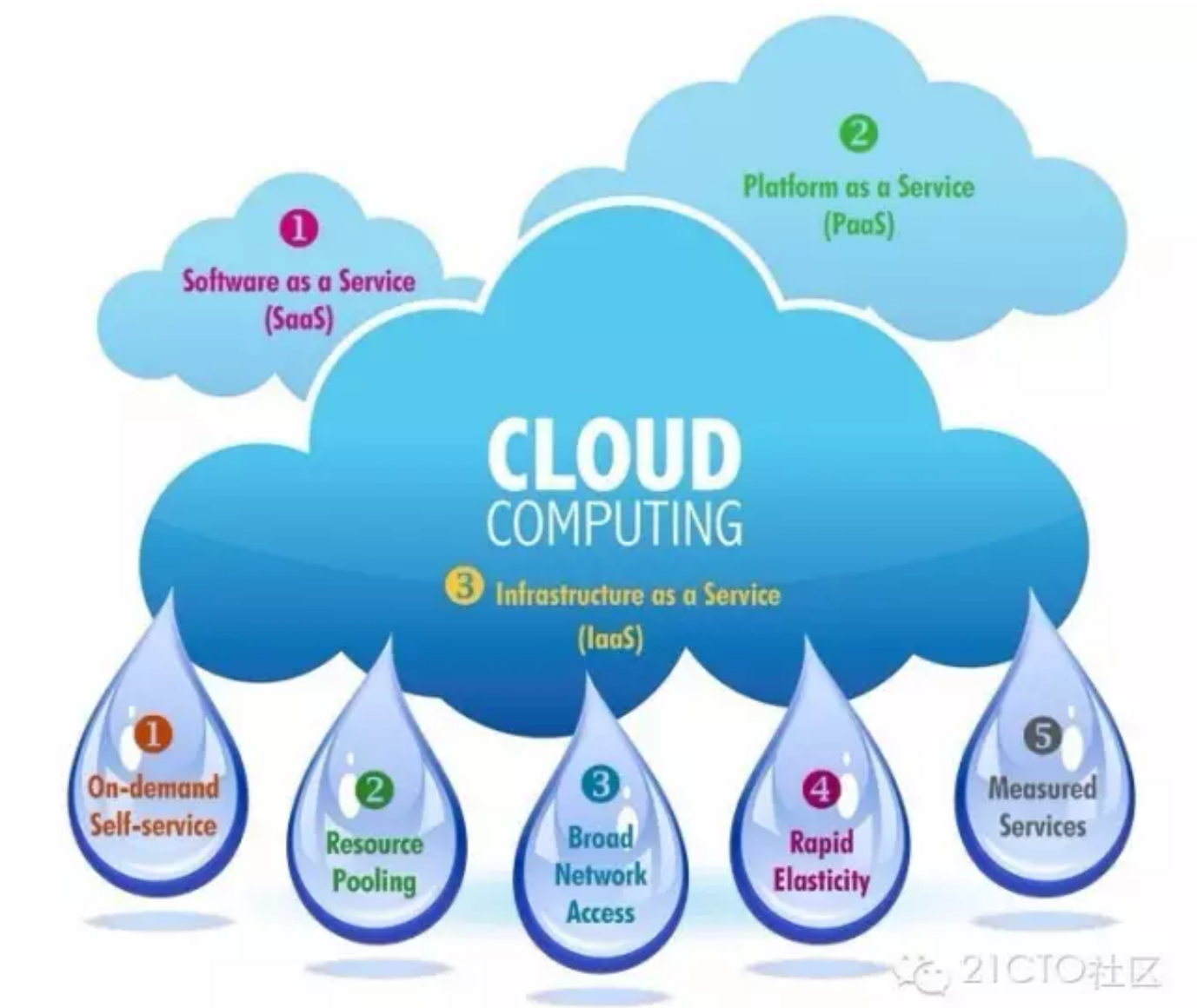
美团云的网络架构演进之路

21CTO 2016-01-19



在传统的观念里，美团似乎一直被认为是家提供吃喝玩乐的团购平台。但其实，从深入电影、外卖等领域起，美团就早已不再是一家团购公司了，打开今天的美团APP：电影、外卖、机票、酒店、上门、甚至周边游等多重垂直业务均被囊括其中。

2015年年底，美团与大众点评宣布合并，合并后新美大的年活跃用户量达到1.7亿，高速增长的业务压力和巨额交易量的背后，是美团云提供的技术支持，让其保持着平稳运营。

作为领先的O2O电商云和大数据解决方案提供商，美团云在2015年入选了“TOP100年度技术创新案例”。 而今天我们要分享的，则是**美团云的网络架构从最初到现在，是如何一路演进而来的，在这个过程中，又产生了有哪些产品和哪些思考。**

做云是水到渠成的事

总有人会问美团为什么要做云？做好电商才是美团该走的路。其实不然，亚马逊、阿里都是从电商起家的，而他们两家分别是国外、国内体量最大的，所以，大规模的电商网站具有把云做好的天然优势：

一方面，**云的核心技术一定是由规模驱动的，大规模的电商在这方面有更深的积累**。因为电商的特点除了流量大，其流量峰值波动也非常大，比如一些特殊的节日，用户访问量的峰值就会很高，所以电商在资源的弹性调度方面有更多的经验。2012年，美团始逐步创建自己的私有云平台，2013年5月正式对外推出公有云服务。2015年更是扩建了新的数据中心，并推出了更多的组件服务。

另一方面，**美团拥有的大数据相关实施经验**，使得美团云能够对外提供更有针对性的大数据融合解决方案。

从技术角度来讲，美团网是一家完全云化的电商平台，规模体量居国内最大。目前美团的交易量仅次于阿里集团，只不过，阿里巴巴的电商业务绝大部分并未完全跑在阿里云上，而美团网所有的业务和交易，从2013年开始就完全跑在美团云上。

这个过程中，美团云在虚拟化、运维等方面积累了相当多的技术经验。同时，**美团云对外输出的不仅是底层IaaS的云服务，更有大数据解决方案**。因此，美团做云是一件水到渠成的事，美团云希望成为美团网技术积累对外输出的窗口，为更多的创业者、中小企业包括正在进行”互联网＋”的传统企业提供基础设施云服务，解决大家在技术方面的后顾之忧，而能够专心业务发展。

从私有云到公有云

美团网早期架构是从私有云做起的。**目标是，资源云化和快速交付。值得一提的是，美团云从一开始就没有完全选用OpenStack，而是决定自研云平台**。原因在于当时OpenStack并不成熟，只有个别组件比如glance、keystone是合适的，所以在虚拟化、网络层，美团云进行了自主研发。

现在看来，这样做是对的。因为OpenStack偏向私有云，如果当初完全基于OpenStack，现在做公有云将比较困难。但美团云选择自研云平台，结合自身业务，所以现如今能够平稳地支撑着所有业务。

当处于私有云的阶段时，主要的事情是把资源动态管理起来，对访问控制和资源隔离没有做太多要求。最初，美团云主要通过账号登陆管理、日志进行事后审计。私有云之后，推出的是办公云。办公云主要针对研发、测试人员，进行内部的测试使用。在这个阶段，美团云已经开始为公有云做准备，建立了账号体系、计费系统等这些功能。

办公云的存在，在现在看来有一个很大的好处，就是每一个上线公有云的功能都会先在办公云上线，保证每一个功能的迭代都是稳定可靠的。也就是说，办公云实际提供了一个真实的线上测试环境。办公云之后，美团云对外推出了公有云服务。

早期的公有云和办公云的架构大体类似，拥有更用户友好、更完善的计费和消息系统、开放API等。其中，公有云最早的底层网络特点有几个，一是网络都是千兆网络，对软件性能要求不高。二是底层采用VLAN大二层，通过OVS控制器对用户进行隔离。由于早期流量不是太大，千兆的流量用OVS来控制尚可，控制器性能不够的情况尚且不多。但随着用户数量的增多，以及使用量的变大，后续开始出现问题。这也恰恰促成了美团云进入全新的网络升级时代。

从微观角度来讲，早期的公有云存在一些问题。首先，在稳定性上，内外网都是一根网线单上连一个交换机，一个地方出问题整个网络就会出问题。其次，外网、内网、管理网都是一根网线，这是在没有冗余的情况下，如果要做冗余的话，就要变成六根网线，成本太高昂。其三，千兆网络渐渐开始不能满足用户需求。还有一些隐藏问题，比如当时所有的用户都是在交换机的一个vlan网络下面。

理论上来说，这样是可行的。**但实际上，交换机对VLAN的支持能力限制了网络规模的扩展，用户数量受到限制。**再比如软件隔离占用宿主机计算资源，可能会出现响应不了或者抢占用户cpu的情况。同时，在这个网络下想实现用户自定义网络（vpc）就非常困难，灵活性低。
因此，在经过了不断地改进后，美团新的公有云网络架构在物理链路、主机网络、网关、控制器四个纬度上全面升级，大大提高了整体网络性能。

四个纬度上的性能释放

首先，从物理链路来看，性能方面，美团云实现了万兆互联；其次，在核心上实现了双机冗余，不会因为某个物理环节问题，导致网络不能启动；第三，采用了TOR交换机堆叠，双40G上联，随着日后网络流量的增加，可以再扩展。此外，**在网线的选择上，美团云还采用了10G Base-T的电口万兆网络，这个技术比较新，很多交换机厂商都还没有这样的设备。**但是它的成本较低，运维起来也会更方便。另外，在机房建设的过程中，美团云还

使用了一些目前业界领先的技术，比如核心机柜封闭冷通道、预端接，对成本的节省都是百万级的。

机房出口挖断了怎么办？同城多个互联网数据中心（IDC）之间，通过边界网关协议（BGP）来进行备份和冗余。当有一个机房的网络断掉的时候，会通过边界网关协议的流量自动转移到另一个机房。

但是底层的物理链路是万兆，不代表上层能把万兆利用起来。我们花了更多的精力，解决如何把万兆网络利用起来的问题。一部分是网关，就是整个网络出口的部分，比如DPDK技术。DPDK技术目前是被主流使用的技术方案，对释放网络性能有较大帮助。另一种，预留1-2个处理器（core）接受数据，一个处理器根据自己的逻辑负责处理控制信息，1-2个处理器负责收包，其余处理业务，自己处理数据分发。

在实际使用中，美团云根据两种模型的优势，分别都有选择。在浮动IP网关、负载均衡网关、 DDoS清洗设备三个部分，实现了全面的DPDK化，同时在四层网络上，能够并发1000w连接情况下新建连接100w / s。

> “以最小代价解决最大问题”

当网关不是瓶颈的时候，流量就能够自由通到主机上，所以接下来就是通过主机网络释放性能。美团云最早使用的OVS V1.1版本，在千兆网络下可行，但万兆网络下性能远远不够。升级到V2.3后平台后，Megaflow对高并发情况下性能有数量级的提升，创建能够满足要求。

但另一个问题出现了，在单流的情况下，对万兆网卡的利用率仅为50%。随后在升级到V2.4，支持DPDK版本后，**美团云进一步提升了单流转发性能**。在新版本的OVS下，只要10%的计算资源就可以提供万兆的网络能力，网络数据处理不影响用户计算资源。这样一来，就解决了宿主机的物理网络瓶颈。

而在控制层面，有两个选择，一个是传统工具eptables/iptables，二是OVS的方案。所谓OVS的控制方式，是配置流表，交由控制器处理。控制器决定是否放行，动态地下发对应流表，在OVS控制器对数据包进行过滤和处理过程中，美团云开发了软件层面的解决方案。针对单播，通过对SYN包检查，下发流表，并对每个不匹配的UDP包进行检查。

需要注意的是，由于发送端较难控制，而接收端对每个包处理，容易造成控制器队列积压。因此，美团云采用下发临时流表的方式解决积压问题，或者通过设置限流阈值，进行

快速恢复。

但是软件层面的解决方案无法根本解决积压的问题，**因此下一阶段的迭代就是在硬件层面进行隔离，通过VXLAN对用户进行隔离**。说到选择VXLAN，就要提到对SDN方案选用的一些思考：在底层的万兆物理链路之上，美团云选用了Overlay的网络架构。

简单来说，**Overlay的架构弹性灵活，业务与物理链接和端口分离，这就意味着网络不再受限于物理上的连接和端口数量，可以按照资源池的概念来分配网络资源**。而Underlay作为整个SDN框架的基础，充分吸取和延续了过去长期积累的物理网络优势，稳定可扩展。一方面ARP/OSFP/BGP 仍然值得信任，另一方面相关领域的运维专业人才相对储备也较多。在参考了业界最新的实践经验后，美团云选用了VXLAN的解决方案。

要做就做行业标杆

上述是在物理链路、主机网络、网关、控制器方面释放性能上，美团云所做的尝试。再上层就是让用户可以灵活地自定义自己的网络。为了应对灵活性的挑战，美团进行了相应的处理，比如分布式的 DNS。
在传统网络下，一般使用默认的DNS服务器地址，并通过源IP区分用户。但是在用户定义网络（vpc）的情况下，用户的地址是可以重复的。所以用户识别方面，**需要将VXLANID的用户信息嵌入DNS数据包**。另外在用户网络中，DNS服务器的地址也是自定义的，所以实际的DNS服务需要使用Underlay地址，这里面就需要做地址的转换和映射。

总体而言，新公有云的网络结构全面升级为万兆网络层面，管理网做Bonding，用户的内网外网overlay在管理网。VPC层面，通过VXLAN隔离用户，并实现自定义的网络。最后对外提供丰富的产品功能，比如浮动IP/负载均衡，对象存储/块存储，RDS/Redis等。

未来，运维自动化的程度会进一步提高。通过openflow或者netconf等通信手段提取到控制器上，进一步整理和分析后，能够形成可视化的网络路径图，实现更高效的网络运维管理。

这些就是**美团云网络架构一路演进的过程**，在这个过程中，美团云的团队成员始终秉承着“以最小代价解决最大问题”的思路，将软件和硬件相结合，通过开源与自研，高效地实现了网络架构的迭代，成为了行业标杆，并为千万用户提供更稳定、可靠的基础设施云服务。

> 来源：美团云

关于21CTO

21CTO是中国互联网第一技术社交与学习平台。 为CTO、技术总监，技术专家，架构师、技术经理，高级研发工程师、PM等提供学习成长，教育培训，工作机会、人脉影响力等高价值的在线教育和社交网站。

看微信文章不过瘾，请移步到网站，诚挚欢迎您加入会员，并成为21CTO学院讲师、教程作者团队。

网站：www.21cto.com

投稿：info@21cto.com

QQ群： 79309783 （欢迎扫描下列二维码关注本微信号）

阅读原文