

Arista构建低延迟云网络

前言：在云网络环境中，应用的响应时间至关重要。

随着数据中心迁移到下一代虚拟化弹性云架构，为确保应用永远在线、始终可用，高性能、灵活的云云网络成为必须。在过去十年，对应用响应时间的要求越来越严苛（见图1），在一些情况下甚至提高了二到三个数量级。除金融、教育/科研机构 and Web 2.0 公司之外，越来越多的垂直行业客户开始将快速的响应时间视为首要的业务要求之一，这其中包括多媒体公司（如从事视频后期制作、数字动画和广播业务的客户）、石油天然气生产商和医疗机构。

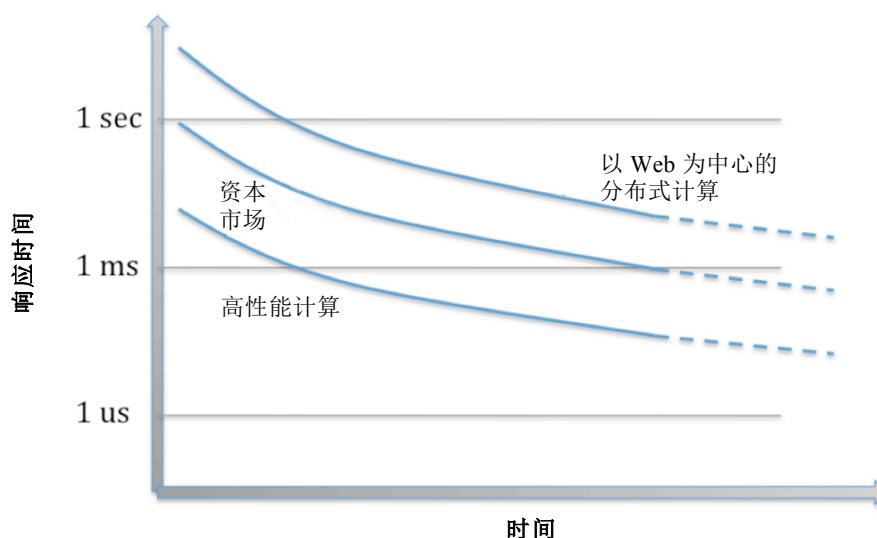


图1：应用响应时间发展趋势

随着响应时间逐渐缩至几分之一秒，网络和计算延迟也必须同步降低。因此，我们需要采取一种有效的架构方法，以便大幅度改进各种云应用的响应时间。与以前部署针对性技术、甚至采用专有技术的做法相比，目前的关键要求是采用基于标准和广泛普及的基础设施架构来部署低延迟的云网络。同时，由于延迟优化是在基础设施级别实施，这种优化必须适用于未来发展，以满足3到5年之内的延迟要求。

构建低延迟云网络

延迟敏感型工作负载的一个主要特点是它们采用分布式计算架构。每个工作负载交易都将在计算节点之间产生大量交互，在某些情况下甚至涉及几千台机器和数据存储库，如图2所示。举例来说：

- 在算法交易中，机器在收到来自其他机器的市场数据后，将计算交易决策，并在另一组机器上开展交易。
- 在HPC中，几千个商用服务器汇集成一个集群，作为一个“超级计算机”执行操作，它与一个集群文件系统（如Lustre）相连，以处理复杂的数学公式和模拟运算。
- 在基于Web的云应用中，工作负载在由几千个服务器组成的动态集群上运行，并使用多级处理机制，以筛选整理数拍字节（petabytes）的数据。相关例子包括Web 2.0和互联网搜索公司使用的的应用，以及Hadoop等新兴的开源计划。Hadoop为分布式计算(MapReduce)、分布式文件系统和分布式数据库提供了可扩展的可靠框架。

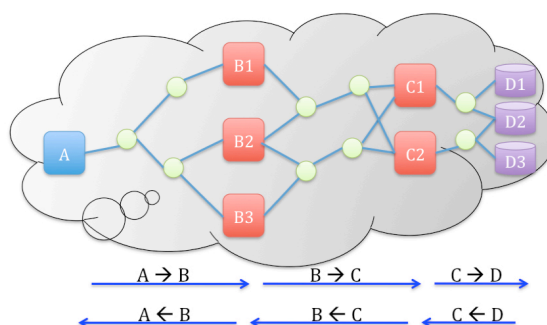


图2：基于网络的单个工作负载的机器间交互：一个交易

为显著改进系统级响应时间，必须最大限度缩小计算和网络延迟。通常会选择最好的商用服务器，使其处理延时小到极限。用户通常会选择配备密度更高、速度更快的多核心CPU；更大、更快的内存/高速缓存和速度更快的外设连接服务器。另一方面，在云网络中，必须通过平台级(fabric-wide)优化来实现传输延迟最小化。具体来说，应在四个重要领域优化网络延迟：

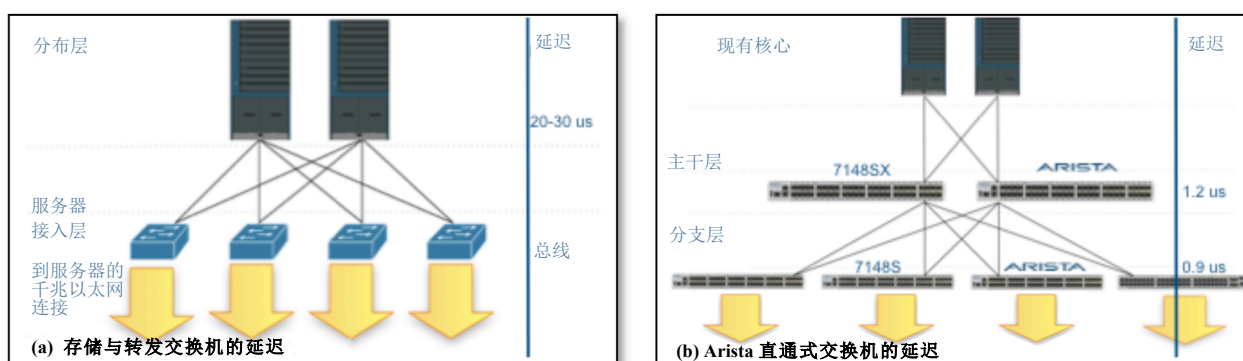
- 降低每个网络节点的延迟
- 减少从一级转至另一级所需穿越的网络节点的数量
- 消除网络拥塞
- 降低传输协议延迟

Arista的优势：

Arista Networks采取了一种体系架构方法来构建低延迟云网络。利用Arista的线速、无阻塞、超低延迟千兆/万兆以太网交换机以及可扩展的操作系统(EOS)，构建灵活的云网络，用于传输数据、多媒体、存储和计算流量。采用Arista的7xxx系列交换机构建云网络具有以下优势：

- 从四个重要方面降低网络延迟（详见下文）；
- 适应未来发展的网络基础设施，可满足当前和未来的延迟要求；
- 部署应用广泛且易于操作的标准以太网。

网络节点延迟：对于金融应用和HPC等领域的超低延迟工作负载而言，缩短每一微秒的延迟都至关重要。通过部署最佳的千兆/万兆以太网交换机，可将1500B以太网帧的传输延迟缩短10微秒以上（如果使用9KB巨型帧，延迟还可降低更多）。此外，与传统的存储与转发设计相比，直通式交换架构还可将延迟缩短数十微秒。Arista推出的直通式架构高密度万兆交换机，可将交换延迟保持在0.6 - 1.2微秒之间（见图3）。



(a) 传统的存储与转发设计	(b) Arista直通式设计
机架内部延迟: 8us	机架内部延迟: 0.6us
机架间延迟: 36us	机架间延迟: 2.4us

图3: Arista直通式交换机相对于传统存储与转发交换机的延迟优势

网络跳数：典型的数据中心架构采用三层架构：即服务器接入层、分布层和核心交换层。通常数据包在服务器机架之间传输时最终都要到达核心层。在Arista采用分支和主干交换机的双层拓扑中（见图4），由于数据包需要穿越的跳数减少，因此延迟也大为降低。

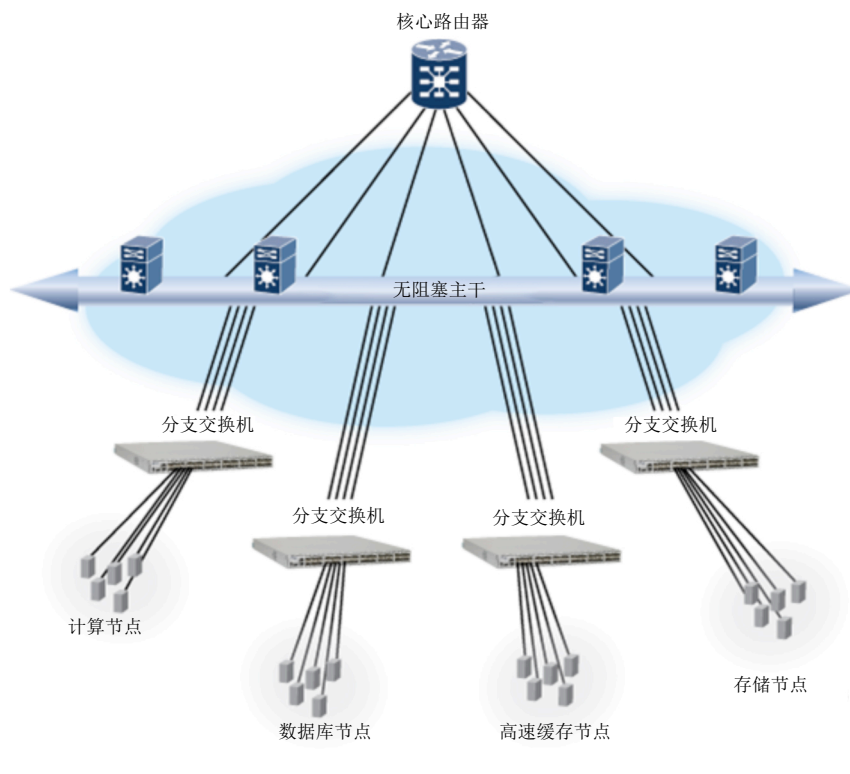


图4: Arista双层(分支与主干)拓扑的延迟低于传统的3层（接入层、分布层和核心层）拓扑

网络拥塞：在多数应用中，IP/以太网数据包通过TCP协议来传输。为保证传输可靠性和吞吐率，TCP状态机利用窗口功能来应对动态的网络拥塞。在拥塞期间，TCP采用较小的窗口，从而降低吞吐率并使总体传输延迟加大。传统的数据中心架构通常都超额配置，因此容易导致拥塞和丢包现象，从而产生较长的TCP延迟。Arista的7100交换机是线速交换机（面向每个万兆以太网端口的线速性能），因此可支持非超额配置的云网络架构，提供1:1的对分带宽（如图4所示）。构建大型的无阻塞架构(fabric)，允许TCP能够使用大型窗口尺寸，从而提高吞吐率，降低延迟。成百或上千个应用工作负载（不管是Web、文件/存储、多媒体还是数据库）机器间TCP交互都能显著受益于TCP吞吐率和延迟的改进。

传输协议延迟：在某些低延迟环境中（如HPC，采用Lustre文件系统算法交易的应用），最好的解决方案是放弃TCP，转而采用基于消息传递（MPI）或远程直接内存存取(RDMA)的直接机器间通信。使用Infiniband和其他专有互连的原因就是它们具有超低延迟、高吞吐率和无损的特点。客户渴望利用基于标准的以太网来实现此类功能和性价比优势，从而能够跨公共云或私有云经济高效地管理单一的以太网架构。Arista的7100系列万兆以太网交换机可提供线速和直通式低延迟性能，能够帮助建立一个统一的低延迟云网络。Arista的架构还支持IEEE和IETF关于第2/3层多路径与拥塞控制的当前及新兴的以太网标准，能够覆盖大型的无阻塞域。大型第2层域和Arista的低延迟线速无阻塞云网络相结合，允许用户快速迁移应用工作负载到云网络中任意地方。

低延迟应用举例：资本市场

算法交易和市场数据馈送是使用超低延迟设计的最典型应用，据说在这些环境中，“对于一家大型经纪公司，交易应用的延迟缩短1毫秒将意味着每年节省1亿美元”¹。数毫秒的延迟缩短对于经纪公司至关重要，为此他们甚至将交易系统放置在交易所内，以便能够直接获取市场数据馈送，从而消除距离造成的延迟。当前的设计力争将交易的响应时间降至1毫秒，并计划在未来几年实现低于1毫秒的延迟设计。出于对速度的追求，用户经常部署非主流的互连技术（比如Infiniband），而在云架构中，这些互连的部署和管理都非常困难。

Arista的线速以太网交换架构和Solace Systems等生态系统合作伙伴共同推出了一个延迟为31微秒的算法交易解决方案，如图5(b)所示。

¹ 《Business at Light Speed》，Richard Martin，《信息周刊》，2007年4月
(<http://www.informationweek.com/news/infrastructure/showArticle.jhtml?articleID=199200297>)

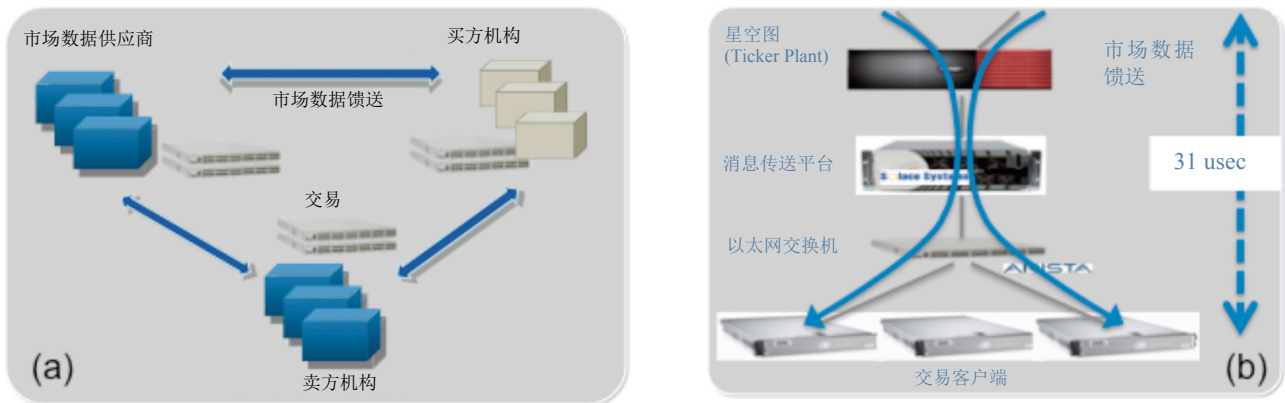


图5(a): 市场数据馈送和算法交易 5(b): Arista + Solace Systems提供超低延迟算法交易解决方案

总结

可扩展的高效云网络的兴起使万兆以太网成为首选的传输接口，用于在一个公共以太网云上同时传输数据、存储和计算信息。许多种应用，包括资本市场的应用、高性能计算和以Web为中心的计算等，都依赖于延迟优化型架构来实现超快速的响应。响应时间的改进能够加速价格浮动期间算法交易的执行速度，从而确保用户的服务粘性，增强HPC集群的可扩展性和提高交易利润。Arista Networks 7XXX系列交换机的设计宗旨就是应对现实世界中的云计算延迟和规模扩展问题。

总部

5470 Great America
Parkway
Santa Clara, California

支持

support@aristanetworks.com
408-547-5502
866-476-0000

销售

sales@aristanetworks.com
408-547-5501
866-497-0000

www.aristanetworks.com