面向大规模网络应用的

青云SDN 2.0

技术分析报告

摘 要

云计算的兴起正使得数据中心规模越来越大，网络性能已经成为制约数据中心规模扩展的瓶颈之一。传统物理网络解决方案越来越难满足用户提出的快速部署、灵活多变等新需求，尤其是公有云服务市场所面对的多租户混合负载环境；公有云服务供应商转而使用部署管理更加灵活的虚拟网络（如SDN、NFV），但在大规模环境下，常见虚拟网络解决方案的性能难以提升，同样制约数据中心的规模。如何在保持虚拟网络灵活性的基础之上，尽可能提升网络性能水平，已经成为诸多用户面临的一大难题。

青云（QingCloud）率先提出了 SDN 2.0 的概念，采用集中管理、分布式控制的新技术方案，利用网状网络替代传统树状网络，并引入 VXLAN Offload、LSO/LRO 等技术，有效提升青云网络子系统性能。SDN 2.0 已经在青云北京 3 区云数据中心正式商用，在这之前，企事录与青云对北京 3 区中的 SDN 2.0 进行了针对性测试。测试结果表明：

1、青云 SDN 2.0 采用分布式控制器构建的网状网络，可有效避免传统树状网络“多跳”产生的延迟，平均响应延迟在 1 毫秒以下；

2、物理服务器内部带宽最高可达 25Gbps，相当于物理服务器背板所能提供的实际带宽极限。不同物理机间传输带宽最高可达 6.31Gbps，跨 VPC 网络传输带宽亦可达到 3.85Gbps；

3、通过 VXLAN Offload 技术有效降低青云网络子系统的性能开销。在极限测试环境下，服务器网卡端口数据转发流量占满时，青云 SDN 2.0 系统服务器总 CPU 占用率为 23.4%，相比上一代技术，CPU 占用率降低了 50% 以上；

4、全局 SDN 控制器迁移测试中，同主机内部网络连接无中断，主机间网络连接 42 秒后恢复可用；

5、在邮件服务器集群测试中，随着邮件服务器数量的增加，其性能线性增长。在测试中，7 个 Exchange Server 可支持 2.1 万邮箱用户，相比于传统的物理服务器解决方案，其成本更低，扩展更加方便。

6、青云 SDN 2.0 所采用的分布式控制器架构在大规模部署环境下，可消减非必要广播包出现，有效避免广播风暴发生。

云计算以低成本和高扩展性，解决无限增长的海量信息存储和计算问题，使得 IT 基础设施能够实现资源化和服务化，用户可以按需定制，从而改变了传统 IT 基础设施的交付方式。

据 IHS2015 年调查报告显示，云服务提供商和通信服务提供商在数据中心部署 SDN（Software Define Network，软件定义网络）将从 2015 年的 20% 提高到 60%，同时，企业采用率预计将从 6% 提升至 23%。2015 年应用于数据中心和企业局域网 SDN 领域的以太网交换机和控制器收入为 14 亿美元，预计 2019 年，该数字将攀升至 122 亿美元。2016 年服务提供商对开源技术可编程软件、SDN 应用和控制器的需求将提高到 45%，企业转向虚拟厂商或第三方 SDN 相关软件的数量将达到 39%。

阻碍云计算发展的两大问题

但目前企业及用户向云计算转型时，还存在着两方面问题亟待解决。

传统网络架构无法满足云计算网络应用需求，设备无法延用

传统数据中心三层网络架构，受到 STP 协议限制，只能通过“树”状节构进行网络数据传输。树状网络结构的规模越大，节点之间转发路径越长，相互间数据传输延迟随之增加。并且，在三层网络核心节点上，也存在引发全网数据传输中断的风险。同时在传统网络中还存在 VLAN 划分数量限制、同网段主机数量不能过大等问题。导致传统网络企业用户难以向云计算转型。

云计算大二层网络的管理控制能力存在瓶颈

有别于传统网络架构，SDN 希望通过将 Control Plane（控制平面）与 Data Plane（数据平面）分离来实现数据高效转发和灵活管控的目的。这种对传统网络的颠覆式创新技术，在实际环境中还有诸多问题亟需解决。例如在大规模虚拟网络部署环境下，

虚拟网络数据传输与管理控制难以兼顾：

此前在国内评测实验室进行的 SDN 白盒交换机测试中，发现测试交换机可以满足 200G 的数据转发，但第三方控制器却无法处理 1000 条数据流的同时请求。一旦发生故障，全网数据包转发均无法进行。

SDN 解决方案的新趋势

作为网络领域面向未来的必然技术趋势，越来越多的新兴 SDN 解决方案供应商开始得到市场认可并逐渐壮大，进而驱动传统网络厂商亦加入其中，推动 SDN 技术加速向前发展。但各自市场地位的不同，使得新兴网络厂商和传统网络厂商在 SDN 产品化方式上大相径庭，大致来看，其可分为基于硬件向上发展和立足软件向下兼容两大阵营。

这两者都有共同的目的，但实现方式的不同也就意味着其具体的 SDN 产品在性能和管控方面各有千秋。SDN 创新性地将 Control Plane 与 Data Plane 分离以实现灵活管控和按需分配的目的，但在虚拟网络的大规模部署情况下，数据传输与管理控制通常会有难以协调的问题。

基于硬件向上发展的 SDN

这种解决方式以传统的路由器交换机厂商为主，利用服务于云计算的专用网络设备，将 Data Plane 与 Control Plane 交由专用网络设备实现。这也是几年前软件定义网络——SDN 技术一经提出，各大网络厂商积极参与的主要原因之一。但是目前通用型 SDN 管理控制器为了兼容更多的底层网络设备，过于追求通用性的开源管理控制器设计。使得管理控制器对网络连接请求处理性能十分低下，管理控制规模和稳定性上的问题始终无法解决，技术发展出现瓶颈，为了寻求解决方案，有很多网络厂商已经开始放弃通用型协议，另行开发适用自身产品的 SDN 管理控制设备。

这方面以思科的 ACI 技术为代表，现在还有很多厂商在借鉴 OpenFlow 与 SDN 控

制器技术，向这个目标进行研发。但到目前为止，基本上都处于实验室阶段，距离成熟商用还有一段距离。

专用网络硬件的优点是 Data Plane 转发性能好，但 Control Plane 不够灵活，很难满足云计算环境下多租户实时多变的应用需求。集中式管理模式也避免不了性能及稳定性方面的瓶颈，网络厂商所提供的解决方案中，管理控制设备多采用自身私有协议设计，很难与其他硬件厂商产品兼容。

立足软件向下兼容的 SDN

第二种解决方式就是利用软件对云计算的虚拟网络进行管理控制。这也是 SDN——软件定义网络的本意所在：网络设备仅需提供最基础的二层数据转发功能，其它一切由软件进行定义。

软件管理控制的优点是灵活性好，研发能力较强的云计算企业可以直接开发出满足其网络应用需求的软件对网络进行控制管理。然而当云计算网络规模扩大后，容易发生以下几个问题：

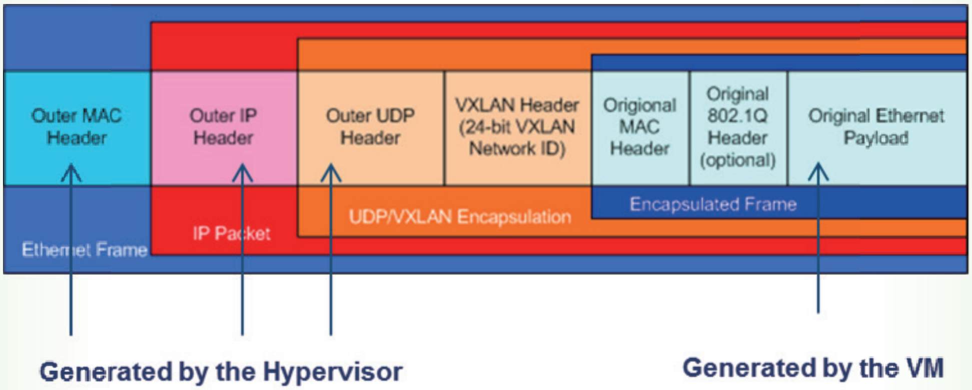
网络连接管理控制问题：

受到 TCP/IP 传输控制协议影响，网络连接建立需要向网络发送广播确定端口位置，从而建立连接。在传统三层网络架构中，可以通过划分子网的方式，将 IP 进行隔离，跨网段时通过三层路由方式进行连接。这样，一个 C 类子网中，可以将 IP 地址数量限制在 254 个 IP 以内。从而有效限制了网络中广播包发送数量，确保网络传输的稳定进行。

而在云计算的虚拟网络中，为了解决虚机在云中迁移的“东西流量”问题，通常做法是构建一个“大二层”网络环境。但是在大二层网络中，无法通过划分子网的形式对广播包传播范围进行规范，而且受到 4096 个 VLAN 划分数量的限制，无法满足多租户 VLAN 划分的应用需求。多租户问题和网络广播风暴问题难以得到解决。

虽然在虚拟化大二层的网络中，可以采用 Overlay 技术来解决 VLAN 划分数量不足

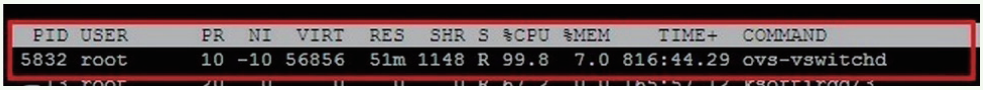
的问题，（目前以 L2 over UDP 模式实现的 VXLAN 技术具备较大优势，现已经成为主流的 Overlay 技术选择）。但是由于它需要对每个数据包都进行封装、解封装等操作，导致基于软件的解决方案效率不高。



VXLAN 数据包封装结构

网络传输性能问题：

在网络数据转发层面，传统网络中，数据转发是由网络设备（交换机、路由器等）承担，芯片和操作系统都经过专门优化，在虚拟网络中通常使用 x86 服务器来实现，但标准通用的 x86 处理器很难针对数据包转发这种逻辑简单但负载繁重的应用场景进行优化，在规模较大环境下必然导致计算资源争用等问题。例如在较常见的“OVS(Open VSwitch)”所构建的虚拟网络环境中，很容易发生 CPU 计算资源占用过高的问题，从而引发抖动异常、延迟过高、丢包乃至于全网瘫痪的恶劣后果。



```
  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
 5832 root      10 -10 56856  51m 1148 R 99.8  7.0 816:44.29 ovs-vswitchd
```

OVS(Open vSwitch)CPU 占用率

在经历诸多探索之后，越来越多人意识到，在保持 Control Plane 灵活性的基础上，同时提高 Data Plane 的高效处理，就必须在“软硬”两方面同时入手，软硬协作以发挥各自的优势。例如，在青云提出的 SDN 2.0 中，其通过分布式、去中心化的软件来提高管理控制的灵活性，同时又通过将数据转发等 Data Plane 的操作从

x86 处理器卸载到网络硬件之上来提高性能，这为 SDN 产品化提供了一种新的解决思路。

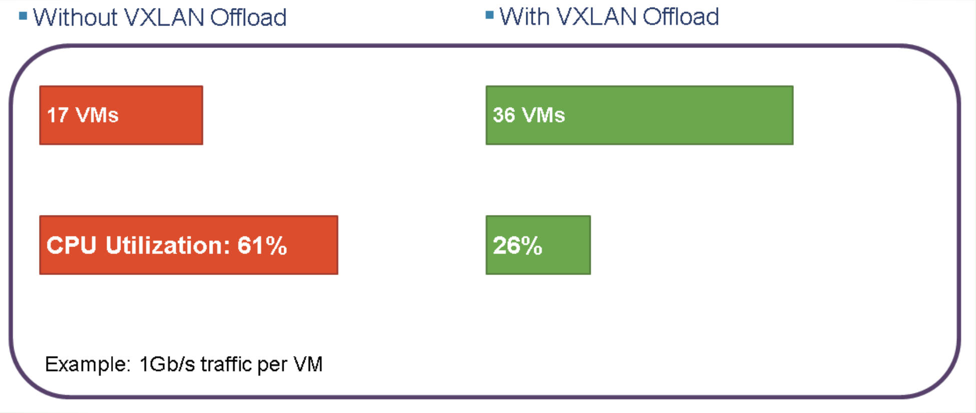
青云 SDN 2.0 技术解决方案

青云在 2015 年末提出了 SDN 2.0 的技术理念，并在其最新发布的北京 3 区云数据中心内规模部署。

本次青云全新采用的青云 SDN 2.0 网络架构，与其上一代产品相比，有以下几方面的重大变化：

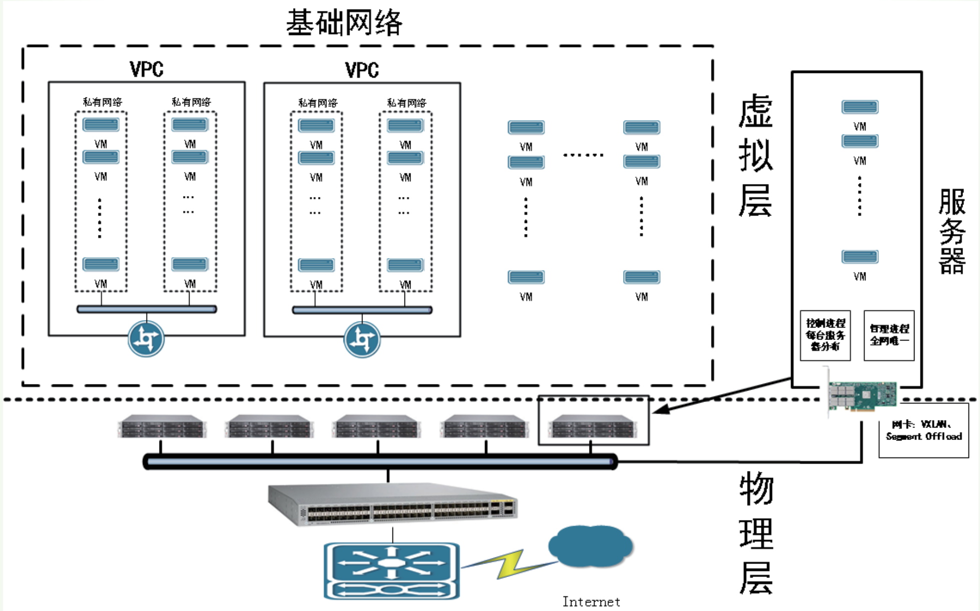
管理控制器方面，将 SDN 控制器升级为扩展能力更强、可靠性更高的分布式管理控制器；网络传输方式上，采用有利于大二层网络管理控制的 VXLAN 取代 OVS，进而在大二层上将虚拟网络与物理设备进行打通，形成一个点对点任意互联的全网状基础网络架构；网络数据转发方面，采用博通（Broadcom）、英特尔（Intel）、迈络思 Mellanox 等主流网卡厂商已经普遍提供支持的 VXLAN 卸载（offload）功能，对 VXLAN 进行一些辅助操作，包括封装、解封装，以及校验和（checksum）计算等工作，有效的减少了虚拟网络数据传输时的 CPU 负载。

（下图为 Mellanox 提供的 VXLAN Offload 前后对比数据。通过对比可以了解，启用网卡 VXLAN Offload 功能后，网络带宽明显增加，CPU 占用有效降低。）



青云 SDN 2.0 网络架构

青云整体网络架构分成三大部分：物理层、虚拟层以及系统层。三个层面逻辑独立但有相互关联，共同构建成了青云 SDN 2.0 网络架构。参见青云 SDN 2.0 网络架构拓扑。



青云 SDN 2.0 网络架构拓扑

系统层——分布式管理控制

青云 SDN 2.0 网络架构中，有一个隐藏在服务器中的层面：系统层——Linux Kernel。它的重要性不仅在于，将物理层和网络层整合为了一个整体。更重要的是，依靠它实现了青云 SDN 2.0 分布控制和集中管理。

在常规 SDN 网络架构设计中，Control Plane 由独立的管理控制设备进行全网统一的管理和控制。而青云 SDN 2.0 架构中，创新性的将 SDN 控制器转变为 Linux 内核中的管理进程和控制进程。控制进程分布在在青云 SDN 2.0 架构中的每台服务器上，各自处理服务器虚机的网络连接。管理进程全网唯一，负责每个服务器内控制

进程的协调管理。

统一管理与分布控制的再次细分，解决了分布式控制与统一管理无法协调并存的现实问题，SDN 网络管理控制能力可以随网络规模拓展而同步提升。使得 SDN 网络控制器控制能力不再成为 SDN 网络规模扩大的发展瓶颈。

正是由于有这一套独特的系统层面存在，才使得青云 SDN 2.0 不再需要一个独立的、性能受限的 SDN 管理控制器存在。安装青云 SDN 2.0 系统的每一台服务器都是一个轻量级、分布式可无限扩展的 SDN 控制器。

物理层——简化网络功能

在物理层中，青云 SDN 2.0 没有采用专用私有的大二层网络产品，也没有使用高度依靠控制器的白盒 SDN 设备。而是采用了最普通的二层网络交换产品让数据中心服务器相互互连，交换设备只需提供两个最基本的功能，即转发和互通。

虚拟层——全网状互通

物理层上承载的是青云 SDN 2.0 的虚拟层。虚拟层由三部份组成：最底层是基于 VXLAN 技术构建的大二层全网状点对点互通基础网络，替代原来系统资源占用高，传输性能低下的 OVS 技术。在基础网络之上，用户可以采用 VPC 网络的方式，预配置出一个采用逻辑隔离的部分，使用户在自己定义的虚拟网络中启动云资源，在 VPC 网络之上，用户还可以根据自身需求，创建自管或拖管的私有网络，满足用户不同形式的网络应用需求。

例如，用户可以为可访问 Internet 的 Web 服务器创建私有网络，而将数据库或应用程序服务器等后端系统放在不能访问 Internet 的私有网络中。可以利用安全组和网络访问控制列表等多种安全层，对各个私有网络中的主机设备进行不同安全策略的访问控制。

企事录验证及评估

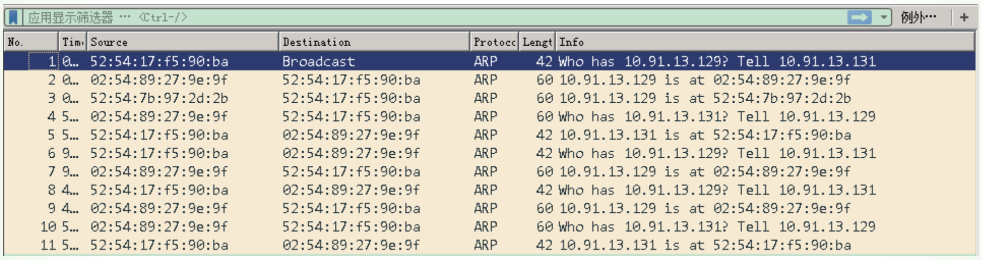
控制器处理能力及可靠性

作为智能的核心，控制器在整个网络中的地位不言而喻，其不仅直接决定网络子系统的规模，还对整体网络的综合性能发挥息息相关。在这一测试方案中，企事录从广播包消减、响应延时以及全网控制器迁移等方面进行多个测试项目，从多个角度考虑青云 SDN 2.0 网络子系统在控制平面的综合表现。

广播包消减

广播包是网络通信的必备组成要素，在网络连接建立的时候，必然会从源端口向网络发出广播包，询问目的端口所在的位置，广播包被正确响应后，数据连接才可以建立，并进行数据传输。但网络中如果有过多网络端口出现，势必会发出过多广播。过多的广播包不仅占用网络带宽资源，还有可能引发网络故障，如引发 ARP 风暴的原因之一就是网络内过多广播包存在，影响了网络中的数据通讯的传输。

下面是企事录对青云 SDN 2.0 基础网络与私有网络抓包后信息对比：



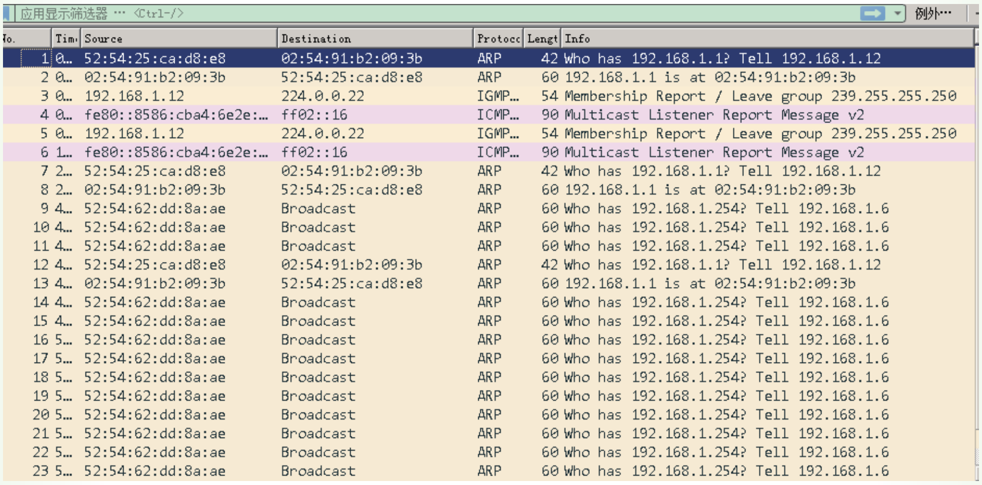
```
No.  Tim Source             Destination        Protocc Lengt Info
   1 0… 52:54:17:f5:90:ba  Broadcast          ARP        42 Who has 10.91.13.129? Tell 10.91.13.131
   2 0… 02:54:89:27:9e:9f  52:54:17:f5:90:ba  ARP        60 10.91.13.129 is at 02:54:89:27:9e:9f
   3 0… 52:54:7b:97:2d:2b  52:54:17:f5:90:ba  ARP        60 10.91.13.129 is at 52:54:7b:97:2d:2b
   4 5… 02:54:89:27:9e:9f  52:54:17:f5:90:ba  ARP        60 Who has 10.91.13.131? Tell 10.91.13.129
   5 5… 52:54:17:f5:90:ba  02:54:89:27:9e:9f  ARP        42 10.91.13.131 is at 52:54:17:f5:90:ba
   6 9… 52:54:17:f5:90:ba  02:54:89:27:9e:9f  ARP        42 Who has 10.91.13.129? Tell 10.91.13.131
   7 9… 02:54:89:27:9e:9f  52:54:17:f5:90:ba  ARP        60 10.91.13.129 is at 02:54:89:27:9e:9f
   8 4… 52:54:17:f5:90:ba  02:54:89:27:9e:9f  ARP        42 Who has 10.91.13.129? Tell 10.91.13.131
   9 4… 02:54:89:27:9e:9f  52:54:17:f5:90:ba  ARP        60 10.91.13.129 is at 02:54:89:27:9e:9f
  10 5… 02:54:89:27:9e:9f  52:54:17:f5:90:ba  ARP        60 Who has 10.91.13.131? Tell 10.91.13.129
  11 5… 52:54:17:f5:90:ba  02:54:89:27:9e:9f  ARP        42 10.91.13.131 is at 52:54:17:f5:90:ba
```

青云基础网络抓包截图

上图是在青云 SDN 2.0 的基础网络内部虚机进行的一次抓包。结果发现除虚机与控制器网关的 ARP 广播外，没有其它 IP 地址的广播信息出现，其基础网络内部广播包异常“干净”。即便有意向一个错误 IP 地址发送 Ping 包，也没有 ARP 广播出现。在青云基础网络内，有效地对广播包进行了消减。

在青云 SDN 2.0 网络架构中曾经介绍过，青云基础网络是采用 VXLAN 技术所搭建。VXLAN 所采用的是一种隧道封装技术，连接建立时，形成的是点对点连接的隧道，

自然不会有杂乱的广播包接收进来或传播出去。而当主机端口向错误地址发出连接请求时，广播包首先会向网关（分布式 SDN 控制器）询问，当请求地址不在控制器管理列表内时，广播包就会被丢弃。不会在基础网络内进行广播，自然起到了广播消减的做用。



o.	Tim	Source	Destination	Protocc	Lengt	Info
1	0…	52:54:25:ca:d8:e8	02:54:91:b2:09:3b	ARP	42	Who has 192.168.1.1? Tell 192.168.1.12
2	0…	02:54:91:b2:09:3b	52:54:25:ca:d8:e8	ARP	60	192.168.1.1 is at 02:54:91:b2:09:3b
3	0…	192.168.1.12	224.0.0.22	IGMP…	54	Membership Report / Leave group 239.255.255.250
4	0…	fe80::8586:cba4:6e2e:…	ff02::16	ICMP…	90	Multicast Listener Report Message v2
5	0…	192.168.1.12	224.0.0.22	IGMP…	54	Membership Report / Leave group 239.255.255.250
6	1…	fe80::8586:cba4:6e2e:…	ff02::16	ICMP…	90	Multicast Listener Report Message v2
7	2…	52:54:25:ca:d8:e8	02:54:91:b2:09:3b	ARP	42	Who has 192.168.1.1? Tell 192.168.1.12
8	2…	02:54:91:b2:09:3b	52:54:25:ca:d8:e8	ARP	60	192.168.1.1 is at 02:54:91:b2:09:3b
9	4…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
10	4…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
11	4…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
12	4…	52:54:25:ca:d8:e8	02:54:91:b2:09:3b	ARP	42	Who has 192.168.1.1? Tell 192.168.1.12
13	4…	02:54:91:b2:09:3b	52:54:25:ca:d8:e8	ARP	60	192.168.1.1 is at 02:54:91:b2:09:3b
14	4…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
15	4…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
16	5…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
17	5…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
18	5…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
19	5…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
20	5…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
21	5…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
22	5…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6
23	5…	52:54:62:dd:8a:ae	Broadcast	ARP	60	Who has 192.168.1.254? Tell 192.168.1.6

青云私有网络抓包截图

点对点连接，有效消减了网络中广播包的产生。但还有一些企业级应用，需要在传统三层网络中运行。为此青云为其提供具备传统三层网络功能的私有网络以满足传统网络业务应用的需求。在青云提供的私有网络中，重复上述测试时，向错误 IP 地址发送 Ping 包时，会有正常的 ARP 请求报文出现。但广播包会控制在青云私有网络划分的网段之内，不会对整体网络构成威胁。

网络中过多广播包的存在，会极大增加控制器对网络管理的负载，青云虚拟网络规模庞大，如果不能对广播包进行有效抑制，就会存在引发网络广播风暴的风险。上面的测试结果表明，青云 SDN 2.0 在保障网络传输正常进行的同时，有效地对不必要广播进行了消减，极大减小了广播风暴发生的可能性。

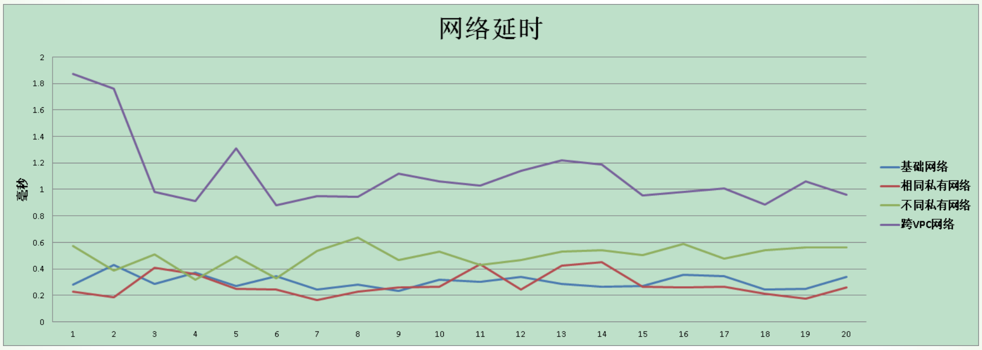
网络延时

网络延时，是网络中源端口向目的端口发送数据包之间的响应时间。响应时间过长

会造成传输性能下降、网络连接中断等一系列不良影响。

为此，企事录对青云 SDN 2.0 基础网络内部虚机间点对点网络延时、同一私有网络内虚机间点对点延时、不同私有网络跨路由器虚机点对点延时、两个 VPC 网络内私有网络经过基础网络跳转跨两个路由器后的网络延时分别进行了测试。

下图是企事录针对青云北京 3 区进行延迟测试后获得的结果：



从网络延时的响应时间上也可以看出，青云 SDN 2.0 网络的延时响应基本平稳，除了在私有网络跨 VPC 连接时，由于网络连接需要经过两个虚拟路由进行转发响应延时略高以外，在跨一个路由的不同私有网络时延时基本就已保持在 0.5 毫秒左右，而在无需经过路由的基础网络和私有网络内部，响应延时更是全部保持在 0.5 毫秒以下。不同位置虚拟机通信延迟都在一个数量级上，低延迟有利于提高网络性能，充分显示出青云 SDN 2.0 点对点传输时出色的连接请求响应能力。

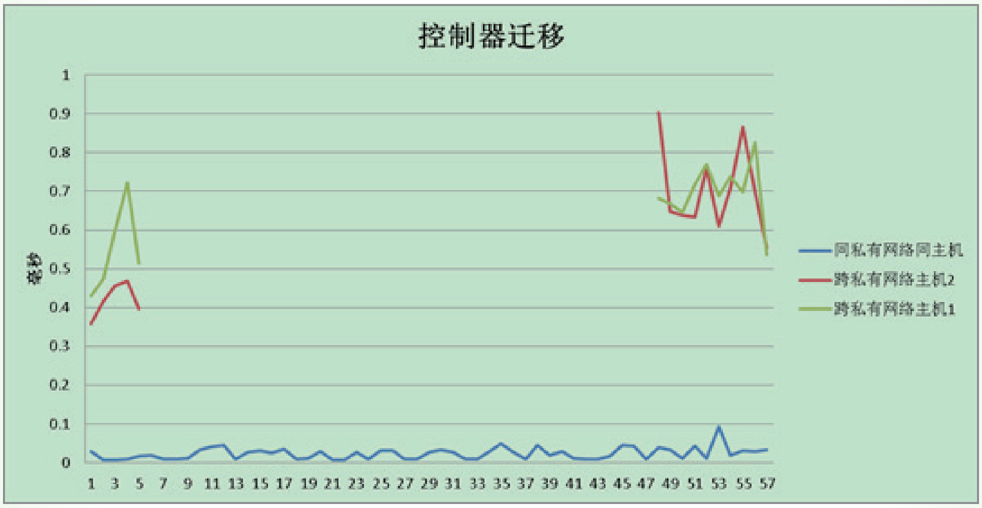
测试结果表明：青云 SDN 2.0 采用分布式控制器构建的网状网络，任意两台虚拟机可以自由进行点对点通信，可以有效避免传统树状网络“多跳”产生的延迟。

管理进程迁移

集中式 SDN 管理控制器，网络管理控制能力无法随网络规模扩大而进一步拓展，在大规模网络应用中会出现网络控制能力拓展性不足的问题。青云 SDN 2.0 通过集中式管理分布式控制的方式，解决了控制器处理能力拓展的问题。但是为了统一协

调管理分布式控制器，必然会有一个全网整体协调的集中式管理进程存在。当这个管理进程出现故障而发生迁移时时，青云 SDN 2.0 的网络会受到什么影响？

为此，企事录对青云 SDN 2.0 在管理进程迁移的时候，网络延时响应时间进行了测试。在测试过程中，同时对同一物理服务器内两个同一私有网络虚机间响应延时和一台虚机在另一私有网络内跨私有网络数据连接的双向响应延时进行了测试。测试结果如下：



测试结果显示：同主机同私有网络内两台虚机在管理进程迁移时，主机上分布式控制进程正常工作，主机内虚机间网络传输未受到任何影响。由于是在同一主机内部，响应延时始终保持在 0.1 毫秒以下。

跨私有网络间主机网络连接需要由管理进程统一调度，因此当管理进程发生迁移后，两个虚机间双向发出的连接请求在出现 42 秒的中断，当管理进程迁移成功能，两虚机间连接请求重新恢复。

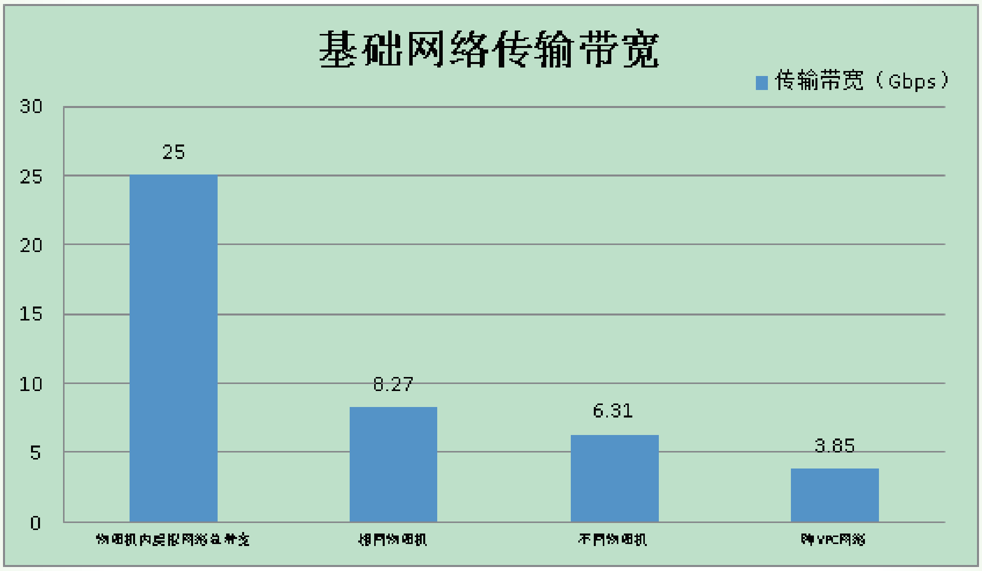
通过以上测试可以了解，当青云 SDN 2.0 管理进程迁移时，分布式控制进程所管主机内部网络传输不受影响，服务器之间网络传输中断时间较短，可在很短时间内重新恢复连接。青云 SDN 2.0 管理控制的稳定性和可靠性得到有效提升。

网络传输带宽及系统资源占用

网络的进化史，就是一部网络传输带宽的进化史，更高的网络带宽为我们带来更多种类的网络应用。大规模的虚拟化网络应用背后必然要有更高的网络带宽进行支持。然而虚拟网络传输对 CPU 计算资源的占用，阻碍了虚拟化网络规模的进一步扩展。同时 VXLAN 技术的采用在为虚拟网络带来更大扩展规模的同时，也会对 CPU 资源带来更高的负载。青云 SDN 2.0 所利用网卡硬件加速提升虚拟网络带宽的效果如何，网卡进行 VXLAN offload 卸载后 CPU 占有用情况怎样，企事录同样也进行了测试。

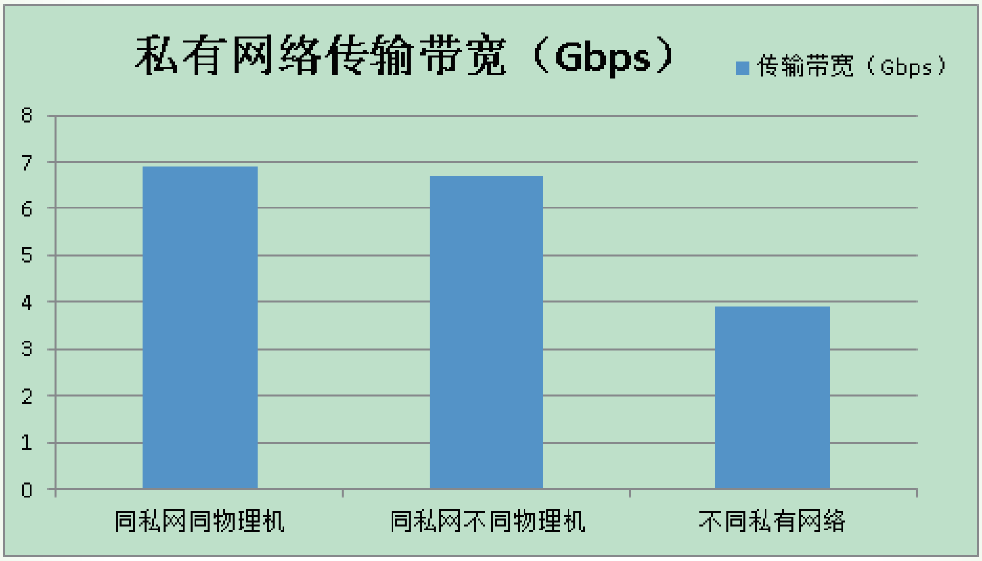
网络传输带宽

在网络传输带宽测试中，企事录对青云 SDN 2.0 系统基础网络内同一物理机内部最大传输带宽、同一物理机内虚机对虚机传输带宽、不同物理机间虚机对虚机传输带宽以及两个虚机在不同物理机上通过两个私有网络路由后基于基础网络数据传输带宽情况进行了测试。



在基础网络传输测试中，采用发送缓冲区长度为 8Kbyte 的流量，进行传输性能测

试。测试结果显示：在相同物理机内部，虚拟网络总带宽可以达到 25Gbps，点对点传输时，传输带宽可以达到 8.27Gbps。在不同物理机之间传输带宽可以达到 6.31Gbps，跨 VPC 网络通过基础网络传输时由于受到多路由器夸 VPC 网络影响，传输带宽降为 3.85Gbps。



在私有网络传输带宽测试时，对同一 VPC 中，两个虚机在同一物理机上采用相同私有网络时传输带宽、两个机在不同物理机上采用同一私有网络传输带宽、两个虚机在不同私有网络传输带宽进行了测试。

在私有网络传输测试中，同样发送缓冲区长度为 8Kbyte 的流量，进行传输性能测试。测试结果显示：在相同私有网络相同物理机内点对点传输时，青云 SDN 2.0 可以达到 6.90Gbps 的传输带宽。在相同私有网络不同物理机之间传输可以达到 6.71Gbps 的平均传输带宽，不同私有网络数据传输由于要通过路由器进行转发，传输带宽为 3.90Gbps 的平均传输带宽。

单个物理机内部 25Gbps 的传输带宽为多虚机间网络数据传输打下了很好的网络基

础，在同一物理机内部，两个虚机不限传输流量下，可以达到8.27Gbps与6.71Gbps的出色传输能力。在不同物理机之间，两虚机不限传输流量下，均可以达到 6Gpbs 以上的网络传输性能。在跨路由器传输时，由于受到路由器转发影响，传输性能有所下降，可最低也可以保持在 3.85Gbps。

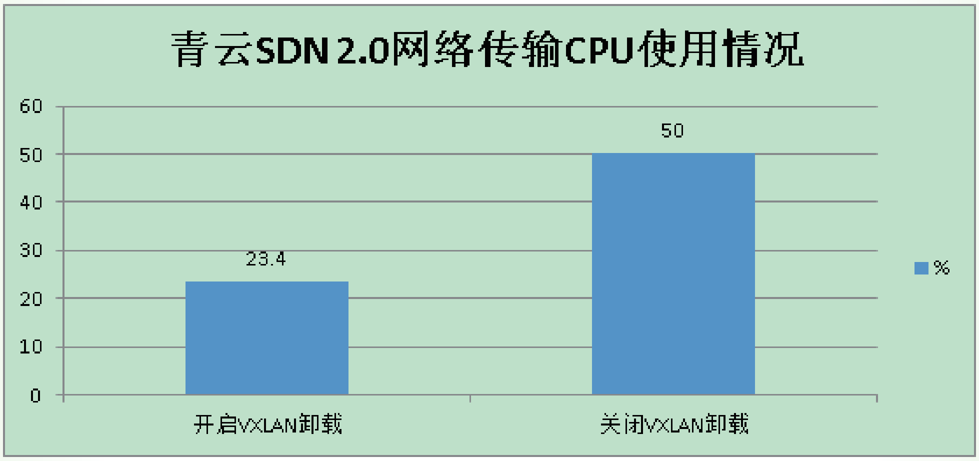
测试结果表明，青云 SDN 2.0 网络系统可以为网络传输提供不弱于传统物理网络的传输性能保障。充裕的网络传输带宽，为青云云计算系统的稳健运营提供了坚实的网络带宽保障。

网络传输 CPU 资源占用

在青云 SDN 2.0 网络架构中曾经介绍，青云采用主流网卡厂商已经普遍支持的 VXLAN Offload 网卡硬件加速功能对 CPU 处理资源进行卸载。以减轻青云 SDN 2.0 网络传输时的 CPU 占用率。CPU 卸载后起到的效果如何，企事录同样进行了测试。

在本次测试中，企事录对青云 SDN 2.0 开启网卡硬件 VXLAN Offload 功能和关闭网卡硬件 VXLAN Offload 时的 CPU 占用率进行了统计。

测试结果表明，青云 SDN 2.0 开启网卡硬件 VXLAN Offload 功能后，在服务器网卡端口数据转发流量占满的情况下，青云 SDN 2.0 系统服务器总的 CPU 占用率为 23.4%，而关闭网卡硬件 VXLAN Offload 功能时，CPU 占用率达到 50%。网卡硬件 VXLAN Offload 效果明显，功能开启后，服务器可以为虚拟系统提供 70% 以上的 CPU 资源处理能力，可以充分满足多种不同的虚拟化网络应用需求。不必再受网卡端口数据转发流量占满后，CPU 资源使用率下降一半的不良影响。

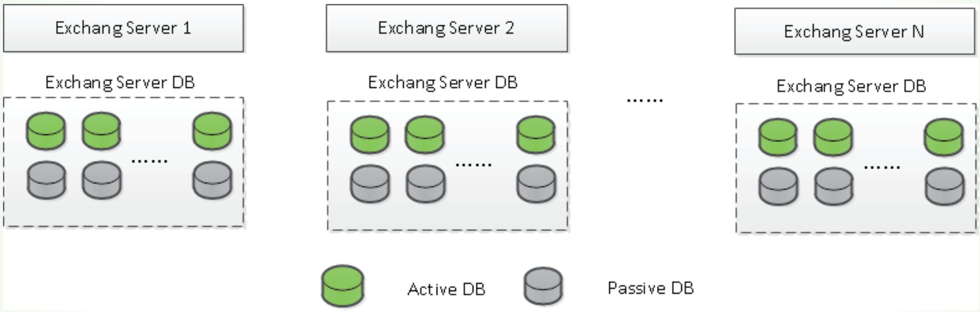


青云 SDN 2.0 系统之所以能提供如此高的 CPU 可用性能，一方面是采用 VXLAN Offload 技术，将繁杂的 VXLAN 打包解包工作卸载到网卡上来进行入理之外，还通过 LSO/LRO（ Large Segment Offload / Large Receive Offload ）技术，尽量加大转发数据容量，从而有效减少数据转发处理速率。因此，起到了降低 CPU 占用，提升网络传输带宽的作用。

青云（QingCloud）邮件服务器集群性能验证与评估

云服务，尤其是公有云服务之所以受到越来越多企业用户的青睐，除了云计算本身具有的池化间隔的物理资源、统一按需分配，以及理论上无限可扩展的特点之外，其很大部分原因还在于，云计算给用户提供了一套整体解决方案，用户可以聚焦在与企业业务息息相关的应用层的研发之上，而无需额外关注底层基础设施的建设与运维。

企事录实验室通过构建一个尽可能真实的企业关键应用环境来验证青云的综合性能表现，并评估随着需求的增长，青云的水平扩展能力与性能表现。下图企事录构建的邮件服务器集群架构示意图：



企事录实验室基于青云构建的邮件服务器集群示意图。安装多个同等配置的 Exchange Server（均为 4vCPU，8GB 内存）以构建邮件服务器集群。正常情况下，每个 Exchange Server 上均设置多个 Exchange 数据库，并均分为活跃数据库（Active DB）和备用数据库（Passive DB）以满足高可用需求。但在本次测试中，Exchange 数据库均为活跃数据库

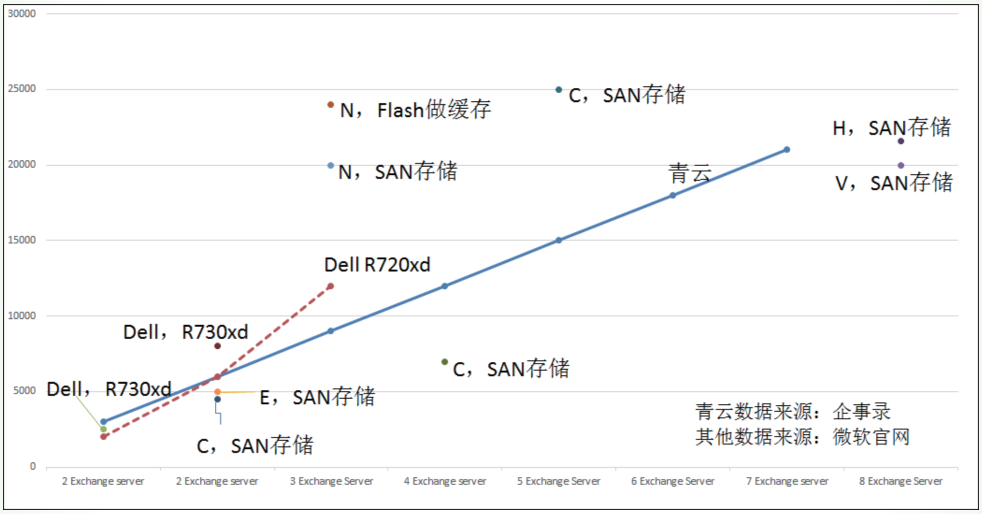
尽管对于中小企业而言，采用公有云提供的电子邮件应用（如软件即服务）更加

简单便捷，但对于中大型规模企业用户而言，自建邮件系统仍具有很大意义，也是企业用户不可或缺的关键应用之一。微软公司推出的电子邮件（Exchange Server）解决方案以其部署简单、功能多样、管理便捷以及扩展方便等特点而受到企业市场广泛接受，使用基于 Exchange Server 的 ESRP（Exchange Solution Reviewed Program，微软公司推出)性能评估方案来测试和评估 IT 系统(包括计算、存储和网络子系统）的综合性能水平，也广受解决方案供应商和企业用户的认可。

基于 Exchange Server 所构建的邮件服务器集群测试环境，虽然对计算子系统有一定的要求，但更多的是评估存储子系统的综合表现（包括性能和容量），同时，在集群环境下，其对网络子系统的性能也有很高的需求（包括带宽和延时）。并且，因为企事录所构建的 Exchange Server 集群是一个贴近真实的测试环境，在测试过程中，缓存（Cache，这里主要是内存和闪存）作为整体解决方案的一部分，也将对测试结果带来较大的影响。

在测试环境构建方面，企业录实验室采用微软公司的 Exchange Server 2013 产品来构建高可用的邮件服务器集群，每个 Exchange Server 上均挂载 5 个 300 GB 大小的卷（Volume）用以 Exchange Server 数据库，共 1.5TB（可以更多，主要受限青云存储策略，单个 VM 的性能卷容量最大为 1.5TB）。

值得注意的是，在真实应用环境中，Exchange Server 通常会有活跃数据库（Active DB）和备用数据库（Passive DB）两种设置，以满足某一个或多个 Exchange Server 故障停机后的业务不中断需求。但在本次测试中，Exchange Server 上的所有数据库均为活跃数据库（Active DB），即假定某一台或几台 Exchange Server 故障后，在满负载环境下，邮件服务器集群是否能够按照既定设计目的正常运行，从而实现业务连续性。企事录实验室以最小 2 个 Exchange Server 为一个集群（一个 DAG）起步进行测试，随着测试的深入，不断增加集群中 Exchange Server 的数量，来验证青云的水平扩展能力和综合性能表现。在 2-7 个 Exchange Server 测试环境下，青云的综合性能表现（如下图）：



随着 Exchange Server 数量的增加，邮件服务器集群所能支持邮箱总数随之增长（蓝线），在 7 个 Exchange Server 环境下，青云可支持 2.1 万个邮箱，每个邮箱大小为 500MB。图中橙色虚线为参考线，数据源自微软官网公布的物理服务器所支持的最大邮箱数量，企事录整理，并不代表实际的极限性能水平；散列点则表示使用 SAN 或者融合设施（闪存用作缓存）实现的最大邮箱数量，数据同样来自微软官网

测试结果表明，随着标准 Exchange Server 的增加，邮件服务器集群规模增加，其所能支持的邮箱用户总量亦线性增长。需要注意的是，ESRP 方案重点考虑存储子系统的综合表现，包括性能和容量。在上图中横坐标中的两个“2 Exchange Server”表示了两种容量配置下的最大邮箱测试数量，前者在测试设置中为每 Mailbox 的容量为 1GB，由于青云的策略设置，每 VM（具体到本次测试则为 Exchange Server）的性能盘总量不能超过 1.5TB，所以单个 Exchange Server 的最大邮箱数量为 15000 个。

当把 Mailbox 的容量设置为 500MB 时，这意味着对性能的要求翻倍，每 Exchange Server 的最大邮箱数量可达到 3000 个。但其存储性能仍有富余，瓶颈限制主要在于存储容量。之所以没有再继续调小 Mailbox 的容量，其原因在于，更小的容量对于实际的用户业务环境没有参考意义；二者，考虑到在公有云多租户环境下，某个物理主机上的某个 VM 不可能也不必要独占物理资源的极限性能。

在本次测试中，当 Exchange Server 数量从 2 个逐步增加到 7 个时，其存储子系

统性能（包括容量）随之线性增长。这主要得益于青云存储子系统中的两大技术：分布式存储和数据本地化。来自应用的数据除了存储在本地以外，还将在其他节点上保存完整镜像，通过类似多副本机制来实现冗余，提高数据的可用性；同时应用所在的物理主机上也将存储一份完整的数据，通过应用与数据位置尽量贴近来避免跨节点读请求导致的延迟增大问题。数据本地化的另一个特点还在于可以进行存储性能隔离，以避免多租户环境下某个租户 / 应用突发性能争抢导致其他租户 / 应用性能受到影响。除此之外，利用缓存(内存)来优化 I/O 以提高整体系统的综合性能水平，也是青云的一大重要手段。

由 7 台 Exchange Server 构成的单一应用服务器集群已属于中大规模集群，为了实现业务连续性，通常也会从应用层面保证数据的高可用，比如本次测试中所使用的 DAG，以及 2 副本设置，会对网络子系统提出一定要求（如带宽和延时）。青云 SDN 2.0 中的分布式网络设计也为此测试项目提供了必要保障，同时也验证了青云 SDN 2.0 在实际应用中的可用性。