

案例研究

英特尔® 傲腾™ 固态硬盘
英特尔® 高速缓存加速软件
英特尔® 存储性能开发套件
企业数据中心



青云用英特尔存储“黑科技” 加速关键业务高性能存储系统



“当云时代来临，如何帮助企业利用其数据保持核心竞争力，是每个企业级数据中心产品和技术提供商都在思索的问题。QingStor™ NeonSAN 给出的答案是：通过不同的硬件组合，为不同应用场景下的存储需求提供高性能、高可扩展、安全可靠以及低 TCO 的解决方案。英特尔为我们提供了丰富多样的软、硬件产品和技术，给我们的解决方案提供了强劲的底层技术支撑。尤其是英特尔® 傲腾™ 固态硬盘与英特尔® CAS 的搭配，堪称天作之合，能帮助用户兼顾高性能和大容量两方面的需求。”

刘乐乐
存储高级技术专家
青云QingCloud

今天，数据已成为企业最具价值的核心资产之一。著名管理学大师 W. Edwards Deming (W. 爱德华·戴明) 曾言: “In God We Trust, All Others must Bring Data (除了上帝，其他皆由数据阐明)”。现代企业的经营者们已充分意识到，企业的经营与发展离不开海量的运营、产品和业务数据的支撑，因此，他们在数据中心的构建上不遗余力。从集中式存储，到直连式存储 (Direct Attached Storage, DAS)，再到存储区域网络 (Storage Area Network, SAN)，不断推陈出新的存储技术都在为企业的发展与创新提供坚实后盾。

技术在发展，互联网时代企业数据扩张的步伐也在加速。据统计，2017 年我国大数据产业规模已达 4,700 亿元人民币，同比增长 30%¹。这一趋势让以 SAN 为存储基石的数据中心逐渐力有不逮。尤其是在高可扩展性和高每秒输入输出操作 (Input/Output Operations Per Second, IOPS) 方面，传统 SAN 架构仅有的 Scale-Up (纵向扩展) 已无法满足需求。为此，企业大量采购专用存储服务器来升级数据中心，不仅带来巨大的成本开销，也让运维复杂度居高不下。

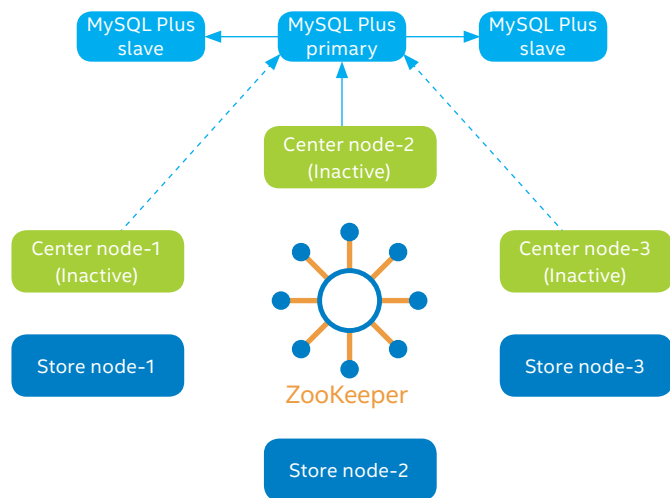
近年来逐渐兴起的软件定义存储 (Software Defined Storage, SDS) 是解决这一问题的优选。SDS 将传统存储服务器的核心功能剥离出来，并以软件的形式进行编排，从而具备高弹性、安全可靠以及易于部署等特点。作为在云计算行业耕耘多年的老牌劲旅，青云 QingCloud* 已为逾 9 万家企业客户提供了各类公有、私有和混合云服务，同时，在企业数据中心构建方面，青云 QingCloud 也颇有建树。现在，它正联手云计算领域的另一巨头英特尔，借助英特尔® 傲腾™ 固态硬盘、英特尔® 高速缓存加速软件 (Intel® Cache Acceleration Software, 英特尔® CAS) 等多项新产品与新技术，推出基于 SDS 理念的新一代分布式超大容量块存储系统 QingStor™ NeonSAN* (以下简称“NeonSAN”)，为企业数据中心打造强劲的核心业务存储引擎。

全新的 NeonSAN 产品无论是在实验室测评，还是在用户侧实际部署中，都取得了骄人的战绩。有数据表明，它不仅能以非常高的 IOPS 性能和很低的 I/O 响应时间，来满足企业关键应用负载提出的苛刻性能需求，而且在保障业务连续性、运行稳定性以及降低

扩容周期等方面也同样令人满意。现在, NeonSAN 已在金融、制造、零售等多个行业用户处进行了部署, 并成功帮助用户加速业务发展、提高决策效率, 获得了用户的一致好评。

云时代企业需要怎样的数据库

随着“云时代”的到来, 如何有效管控和利用企业拥有的海量数据已成为业界关注的焦点。在青云QingCloud 看来, 企业数据虽然千变万化, 但对于存储系统而言, 核心点无外乎四个: 性能、安全可靠、可扩展和成本。作为国内为数不多的全栈云服务提供商, 青云QingCloud 在推出“一站式混合云”、“超融合一体化设备”等云服务架构产品之余, 也在存储领域发力, 与合作伙伴英特尔公司一起, 协力推出了全新的分布式块存储系统“QingStor™ NeonSAN”。该产品凭借企业级的高性能、低延迟与出众的横向扩展能力, 帮助青云QingCloud 在云服务版图上增添了重要的板块。



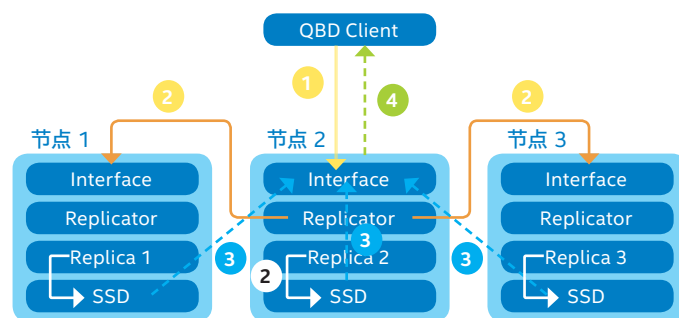
图一 QingStor™ NeonSAN 集群架构图

如图一所示, NeonSAN 存储集群由四类功能组件构成。在集群管理上, 青云QingCloud 通过对分布式协调服务软件 Zookeeper* 进行大量优化, 使之承担起 NeonSAN 所需的集群管理、负载均衡、主中心节点选举、分布式协调与通知等管理能力。集群中的系统控制层被称为中心节点 (Center Node), 其通过一主多从的模式来实现高可用性, 当主中心节点 (Active Center Node) 失效时, Zookeeper 会自动生成新的主中心节点进行接管。在元数据存管上, NeonSAN 采用了分布式数据库集群 MySQL Plus*, 其具备

的强数据一致性、秒级主从切换等特性, 有力地保障了主中心节点对元数据信息的访问; 最后, NeonSAN 还采用了多 Store 节点的方式来实现数据层服务。

面对企业用户关注的核心需求, 青云QingCloud 为 NeonSAN 提供了多项“独门秘笈”。在存储性能上, NeonSAN 可支持远程内存直连 (Remote Direct Memory Access, RDMA*), 并支持 InfiniBand*、RoCE*、NVMeoF* 及 iWARP* 等一系列存储协议和技术, 在实现超高性能、超低延时的同时, 节省对处理器资源的消耗。来自青云QingCloud 的资料显示, 当 NeonSAN 采用“全闪存 + RDMA 网络”配置时, 其 4K 随机读写的单盘 (卷) 读写性能高达 100K IOPS, 而延时低至 90 微秒², 这一表现足以满足金融行业用户在交易类、查询类、分析类等关键业务上的高并发需求。

数据的安全可靠性是企业用户心中永远紧绷的弦。为此, NeonSAN 设置了多道安全防线来保障可用性和安全性。第一道防线, 是灵活的多副本机制。不同盘 (卷) 可指定不同副本数, 并存放在不同物理节点上。如图二所示, 在此机制下, NeonSAN 节点在收到写请求后, 除了写入主副本, Replicator 进程也将会从副本发送至其他节点。只有当主、从副本都成功写入后, 节点才会返回成功消息, 这一机制保证了数据各副本之间的强一致性。



图二 QingStor™ NeonSAN 多副本机制下的数据写入流程

第二道防线, 是多路径与节点失效自动切换功能。NeonSAN 的每个节点都配备双网卡, 每个网卡都具备双端口, 四个端口分别连接到四个不同的交换机上。其中两台前端交换机互为冗余, 构成前端网络提供对外数据服务; 另两台 RoCE 交换机则互为备错, 当一条链路发生故障时, 节点能自动切换到其它链路而不中断业务, 在网络层面保障服务的高可用。

第三道防线，是“瞬时快照”功能和“无中断的数据恢复和迁移”功能。前者允许每个单盘（卷）保留 256 个快照，在数据发生损坏时，可通过快照迅速回滚到指定时间点的数据，实现数据恢复。而后者则让节点在任意时间实施数据恢复、迁移及容量均衡时，上层业务无感知，新增节点可立即投入业务，减少集群扩容和故障对业务带来的冲击。

强大的双向扩展能力是 NeonSAN 有别于传统企业数据中心 SAN 架构的主要亮点，NeonSAN 基于 x86 架构标准硬件构建，采用了全分布式的架构设计，系统容量和性能均可进行在线水平扩展。在目前的私有云部署方案中，NeonSAN 能够支持 3 至 1,024 个节点，并可在不中断业务的情况下，实现以单节点为单位的平滑扩容。来自青云QingCloud 的测试数据表明，其单盘（卷）容量可扩展到 100TB 之巨³，完全可满足大数据场景下的数据存储需求。

在企业用户关心的另一个要点——成本与性价比上，NeonSAN 也给出了令人满意的方案。在私有云部署方案中，NeonSAN 能够根据用户的实际需求来提供精简配置，例如存储容量。当用户的数据增长导致所分配的容量不够时，系统能自动从后端存储池中予以补充，大大提升了存储资源的利用率。同时，NeonSAN 还支持 TCP/IP 协议，用户在使用时，既可使用全新的 RDMA 协议，也可使用传统的 TCP/IP 协议，不必二选一，不但带来更大的自由度，也最大程度地保护了用户的既有投资。

容量与性能的均衡：英特尔® 傲腾™ 固态硬盘 + 英特尔® CAS

在传统企业数据中心的构建中，用户往往秉承“强强联手”的朴素理念，购置昂贵的计算设备、昂贵的存储设备和昂贵的网络设备，并将他们组合起来。然而，这类“超豪华”组合又往往达不到用户的心理预期，既造成资源的浪费，也带来诸多的兼容性问题。在青云QingCloud 看来，不同的行业用户对存储系统有着不同的需求。例如在财务、生产等联机事务处理（Online Transaction Processing, OLTP）场景下，用户追求的是高 IOPS，而在策略

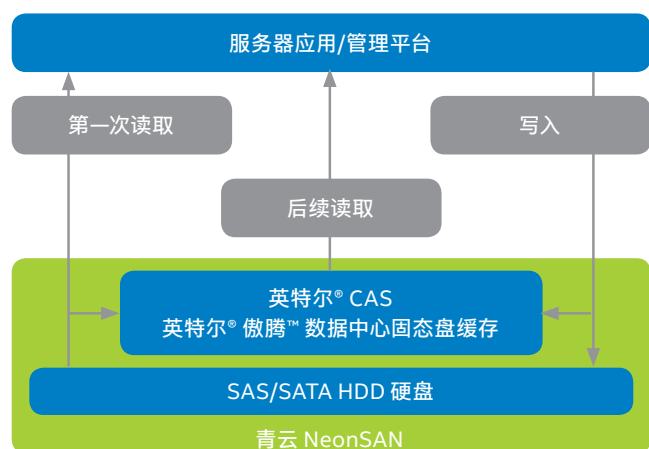
配置、智能应用等 OLAP 场景下，则对高吞吐，大容量有着较高要求。

针对企业数据中心不同的应用场景，青云QingCloud 利用不同的技术架构和硬件组合，为 NeonSAN 设计了不同的部署模式。首先，针对高性能、大容量存储的应用场景，青云QingCloud 提出了基于 TCP/IP 网络，SAS/SATA HDD 硬盘 + 固态硬盘缓存的部署方案，以分层存储（Tiered Storage）的方式，通过单节点挂载 12 块 4TB 容量的 SAS/SATA HDD 硬盘的方式，满足海量数据的存储需求。

众所周知，支持分级存储的分布式存储系统要具备高性能，就必须高效地对缓存进行读写。影响此类 NeonSAN 部署方式性能的因素，主要在于缓存的性能，以及系统管理缓存的能力。为此，青云QingCloud 引入了英特尔在这两方面的领先技术。

由全新的英特尔® 傲腾™ 固态硬盘DC P4800X 担纲的缓存（Cache）在 NeonSAN 上发挥出了惊人的性能表现。该数据中心级英特尔® 傲腾™ 固态硬盘基于创新的英特尔® 3D XPoint™ 存储介质以及英特尔先进的系统内存控制器、接口硬件和软件进行构建，其在低延迟和稳定性方面的性能表现，远远优于传统 NAND 介质固态硬盘，尤其适用于电商、金融、保险等多用户、高并发的 OLTP 场景，而且 NeonSAN 目前采用的英特尔® 傲腾™ 固态硬盘DC P4800X 375GB 版本的每天写入次数（Drive Writes Per Day, DWPD）高达 30，有力地保证了用户系统的有效生命周期⁴。

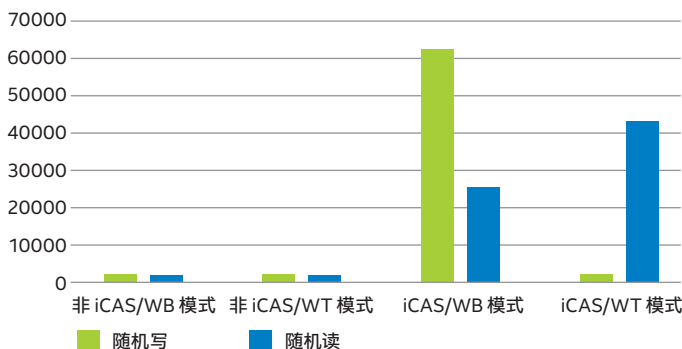
为使英特尔® 傲腾™ 固态硬盘在 NeonSAN 中发挥出更大效能，青云QingCloud 还引入了英特尔针对缓存性能优化开发的另一项专有技术：英特尔® CAS。如图三所示，当应用第一次读取数据时，NeonSAN 会从后端 SAS/SATA 存储中读出相应数据并返回给应用，同时数据也会被英特尔® CAS 复制到由数据中心级英特尔® 傲腾™ 固态硬盘构建的高速缓存中。在后续的读取中，应用直接从缓存里高速读取。而在数据写入时，所有数据都会同步写入到后端存储和高速缓存中。



图三 英特尔® CAS 技术加速原理图示

当高速缓存空间写满后，英特尔® CAS 具备的专有移出算法，会自动将新的活动数据取代高速缓存中的陈旧数据。可以看出，通过英特尔® CAS 的介入，应用服务器可始终用最快速度读到最“热”的数据，这对于冷、热数据比例越来越悬殊的企业数据中心而言，有着非常现实的意义。

青云QingCloud 进行的一项英特尔® CAS 对比测试也有力地证明了上述观点。在 NeonSAN 上进行的在 FIO*（一款 IO 测试工具）测试中，通过执行 4K 随机读写测试，英特尔® CAS + 数据中心级英特尔® 傲腾™ 固态硬盘的组合，无论是在 WB（回写）模式，还是在 WT（直写）模式，IOPS 性能都远超未加组合的对比测试组。在 WB 模式下，英特尔® CAS 与英特尔® 傲腾™ 固态硬盘的随机写性能甚至达到了对比测试组的 23 倍之多⁵。



图四 4K 随机读写下的英特尔® CAS FIO 测试对比

极致的存储性能表现：全闪存 + SPDK

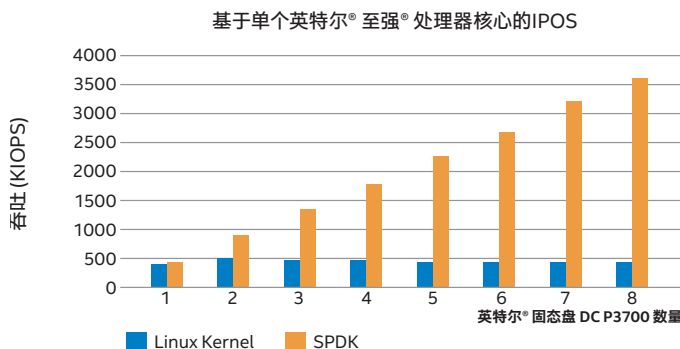
NeonSAN 的另一种部署方案是面向超高性能、低延迟的企业应用场景。为此，青云QingCloud给出了“全闪存配置”方案，而这个方案的主

角是英特尔® 固态硬盘 DC P4510。这一符合非易失性内存主机控制器接口规范 (Non-Volatile Memory express, NVMe) 的固态硬盘产品，采用了英特尔® 3D NAND™ 技术，与上一代固态硬盘或传统 HDD 硬盘相比，其 IOPS 性能有了质的飞跃，同时能耗与故障率也有显著下降。

一项来自第三方测评机构的数据显示：配置为全闪存的 NeonSAN，在单应用压力时，4K/8K 随机读写性能均接近或超过 100K IOPS，且平均响应时间低于 0.8 毫秒。同时随着 NeonSAN 卷数量的增加，其性能也会随之线性增长，在配置 4 个 NeonSAN 卷时，4K 随机读写和 8K 随机读性能可达到 30 万 IOPS 左右，8K 随机写性能则超过 25 万 IOPS，平均响应时间低于 1 毫秒⁶。

基于 NVMe 规范的英特尔® 固态硬盘在为用户提供高吞吐、低延迟存储能力的同时，英特尔推出的面向固态硬盘存储产品的英特尔® 存储性能开发套件 (Intel® Storage Performance Development Kit, 英特尔® SPDK) 也在以各种创新技术提升存储软件的性能。

在 HDD 硬盘时代，中断开销仅占整个 IO 过程的很小比例，因此影响并不突出。随着固态硬盘产品的普及，尤其是 NVMe 固态硬盘的到来，其强劲的 IO 性能让中断开销突然变得“醒目”起来，使系统瓶颈从存储硬件转移到了软件。为此，英特尔® SPDK 提供了多项创新技术，最核心的有三点：UIO/VFIO（用户态驱动），Asynchronous Polling（异步轮询机制）和无锁设计，可帮助上层应用充分利用到 NVMe 固态硬盘带来的高性能。在用户态驱动中，由英特尔® SPDK 实现的开发库能将存储设备的基地址寄存器（Base Address Register, BAR）地址映射到应用的进程空间中，从而尽可能降低中断开销，发挥出 NVMe 固态硬盘的性能优势。



图五 英特尔® SPDK 用户态驱动与内核驱动的对标测试

用户态驱动能在单处理器核心上, 管理多个 NVMe 固态硬盘设备, 实现高吞吐量、低延时以及处理器资源高效使用等优势。如图五所示, 在一项针对英特尔® SPDK 用户态驱动与内核驱动的对比测试中, 单处理器核心搭配 SPDK 后可令 8 块英特尔® 固态硬盘“火力全开”, 而内核驱动要达到类似水准, 则至少需要配置 8 个处理器核心⁷。

英特尔® SPDK 用户态驱动对设备完成状态的检测, 是采用异步轮询机制来完成, 应用在提交读写请求后, 无需等待读写操作的完成, 可以继续按需发送请求, 然后再完成回调函数中处理, 从而避免中断带来的延迟和开销, 提升 IOPS。事实上, 对 NVMe 固态硬盘设备的轮询, 是非常高效的。按照 NVMe 规范, 存储设备会通过读取内存来检测完成队列是否有新的操作完成, 通过英特尔® 数据定向 IO (Data Direct I/O, DDIO) 技术, 可实现设备更新后的数据被存放于处理器缓存中, 以此实现高性能的设备访问。

而英特尔® SPDK 的无锁设计, 是在数据通道上去掉对锁的依赖。一方面, 英特尔® SPDK 通过线程亲和性的方法, 将某个处理线程绑定到特定处理器核心上, 同时通过轮询的方式占住该核心的使用; 另一方面, 通过采用运行到完成 (Run To Completion) 的方式, 把应用读写请求的整个生命周期都绑定到特定处理器核心上完成, 从而避免再将宝贵的处理器资源用于同步线程间的数据, 以进一步提高 IOPS 性能。

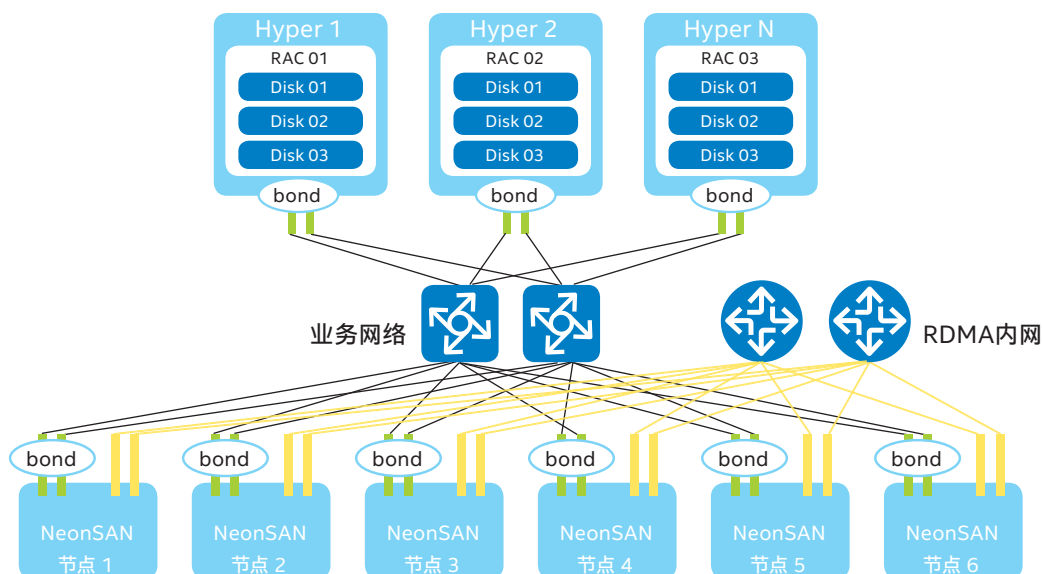
英特尔® SPDK 技术融入 NeonSAN 之后, 为其带来了显著的性能提升。来自青云 QingCloud 的测试数据表明, 无论是单副本还是多副本配置, 随机写的时延都能降低 10 微秒左右, 而多副本下随机读时延可以降低 20 微秒。混合读写场景下 (读写比: 70/30), 在 3 节点配置的 NeonSAN 集群中, 两副本卷的读写性能都提升了近 20%⁸。

企业用户的实践

Oracle® 实时应用集群* (Real Application Clusters, RAC*) 是目前企业最流行的数据库集成环境, 因此, 青云 QingCloud 也与英特尔一起, 针对 Oracle RAC 进行了大量的测试和优化, 并在多个用户处进行了实际部署, 获得了令人满意的反馈。

在一项通过模拟 Oracle 数据库 + NeonSAN 节点进行的评估中, 整体存储系统的每分钟事务处理量 (Transactions Per Minute, TPM) 性能超过 165 万, 平均每秒事务处理量 (Transactions Per Second, TPS) 性能接近 3 万, 且完成每个事务处理的平均时延为 15 毫秒左右, 这一性能能够支撑绝大多数企业关键应用负载⁹。

在与某零售巨头的合作中, 青云 QingCloud 不仅提供了全栈云计算解决方案来构建其集团云平台, 也帮助其在生产区部署了 6 节点的 NeonSAN 集群。如图六所示, 对于该企业原有业务环境中的 Oracle RAC 数据库而言, 仅需添加 NeonSAN 作为共享盘, 就可进行业务数据的迁移, 实现了良好的扩展性。



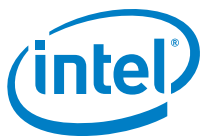
图六 某零售巨头存储系统架构, 后端部署 6 节点 NeonSAN 存储集群

该企业的反馈表明, NeonSAN 的部署有效推进了其核心 ERP 系统向云计算架构的演进, 实现了私有云环境下的一体化运营与管理, 在保障业务连续性的同时, 大幅降低了采购和运营成本。同时, NeonSAN 表现出的可扩展性, 令其存储系统的建设和扩容周期从几个月缩短至一周, 可以有效满足在业务数据量激增下的扩容需求, 切实推动了业务系统平滑、快速的发展。

而对另一大型金融企业客户而言, 通过把各项业务与青云 QingCloud 云平台实现在线无缝对接, NeonSAN 已成为其各项核心业务, 尤其是 OLTP 业务场景的存储引擎。该企业的实测数据表明, 基于NeonSAN 的复杂视图查询时间缩短了 90% 左右, 而复杂 SQL 语句的执行效率则从分钟级变成了秒级¹⁰。

来自多个实践案例的好评, 说明青云QingCloud 与英特尔携手打造的 QingStor™ NeonSAN 已充分获得市场与用户的认可。

未来, 双方还将继续深入合作, 以先进的产品与技术为企业数据中心的性能提升贡献力量。目前, 青云QingCloud 正逐步将 NeonSAN 系统的处理器更换为新一代的英特尔® 至强® 可扩展处理器, 并计划进一步挖掘该处理器所蕴含的性能潜力, 特别是利用英特尔® 高级矢量扩展 512 (英特尔® AVX-512)、英特尔® Virtual RAID on CPU (英特尔® VROC)、英特尔® 可信执行技术 (Intel® Trusted Execution Technology, 英特尔® TxT) 的功能, 来满足企业存储系统未来不断增长的算力需求, 打造高效且差异化的云存储服务。



¹ 本数据来自中国信息通信研究院发布的《大数据白皮书 (2018) 》

^{2,3} 数字引自青云 Cloud 官网的相关产品性能说明: <https://www.qingcloud.com/products/qingstor-neonsan/>

⁴ 数据来自据来自 <https://www.intel.cn/content/www/cn/zh/solid-state-drives/optane-ssd-dc-p4800x-brief.html>

^{5,6} 该测试数据援引自青云QingCloud 的内部测试报告。

^{6,9} 该测试数据来自https://www.sohu.com/a/249625689_464027

⁷ 该测试数据基于以下测试环境: 英特尔® 至强® 处理器 E5-2695 v4 单处理器核心, 4K 随机读测试, 128 的队列深度, 操作系统为 CentOS* Linux* 7.2, Linux 内核 4.10.0, 测试固态硬盘为 8 块英特尔® 固态硬盘 P3700 NVMe (800GB)。

¹⁰ 该测试数据来自《DT 时代性能测试报告: 什么样的存储引擎让 Oracle 数据库性能 100% 增长? 》, https://www.sohu.com/a/249625689_464027

英特尔技术特性和优势取决于系统配置, 并可能需要支持的硬件、软件或服务才能激活。没有计算机系统是绝对安全的。更多信息, 请见 Intel.com, 或从原始设备制造商或零售商处获得更多信息。描述的成本降低情景均旨在特定情况和配置中举例说明特定英特尔产品如何影响未来成本并提供成本节约。情况均不同。英特尔不保证任何成本或成本降低。

英特尔、Intel、至强、傲腾是英特尔公司在美国和其他国家的商标。英特尔商标或商标及品牌名称资料库的全部名单请见 intel.com 上的商标。*其他的名称和品牌可能是其他所有者的资产。