CASE STUDY

Intel® Optane™ Solid State Drive (SSD)
Intel® Cache Acceleration Software (CAS)
Intel® Storage Performance Development Kit (SPDK)
Enterprise Data Center

QingCloud uses Intel's futuristic storage technologies to accelerate high-performance storage systems for mission critical

QINGCLOUD

"With the approaching cloud era, how to help companies make use of their data to maintain their core competitiveness is an issue that every provider of enterprise-grade data center products and technologies is pondering. QingStor™ NeonSAN's answer is to provide solutions that boast high performance, high scalability, security and reliability as well as low TCO for storage needs under different application scenarios through different hardware combinations. Intel offers us a rich and diverse range of software and hardware products and technologies, providing powerful underlying technology support for our solutions. Intel® Optane™ SSD and Intel® CAS, in particular, are 'a match made in heaven' for their ability to help users satisfy both their needs for high performance and large capacity."

Lele Liu
Senior Technology Specialist for Storage
QingCloud

In today's world, data has become one of the most valuable core assets of corporations. The famous management guru W. Edwards Deming once said "In God we trust, all others must bring data". Modern corporate managers have fully realized that the running and development of a corporation are impossible without massive operational, product and business data, and thus they spare no efforts to build up their data centers. From centralized storage to direct attached storage (DAS), followed by storage area network (SAN), the endless stream of new storage technologies have been robustly and reliably supporting the development and innovativeness of corporations.

As technologies morph, the pace of enterprise data expansion in the Internet era has been accelerating. According to statistics, the size of Big Data industry was 470 billion RMB in 2017, a 30% year-on-year increase[1]. Data centers using SAN as their storage cornerstone are finding it hard to catch up with this trend; especially with respect to high scalability and high input/output operations per second (IOPS) which cannot be satisfied by the traditional SAN architecture which can only scale up. To tackle this problem, companies purchase in large numbers exclusive storage servers to update their data centers, but this has led to more problems: Enormous costs and highly complex operations and maintenance.

The software defined storage (SDS) that has emerged over the years is the best solution to resolve these problems. SDS abstracts the core functions of traditional storage servers and arranges them by means of software for high flexibility, security and reliability, as well as easy deployment. As a veteran in cloud computing, QingCloud* has provided all kinds of public, private and hybrid cloud services for more than 90,000 corporate customers. At the same time, QingCloud has shown accomplishments in the building of enterprise data centers. Currently it is teaming up with another cloud computing giant, Intel, to leverage Intel® Optane™ Solid State Drive (SSD), Intel® Cache Acceleration Software (CAS) and other new products and technologies to roll out the new QingStor™ NeonSAN* (hereinafter referred to as NeonSAN) distributed high capacity block storage system to create a powerful core business storage engine for enterprise data centers.

The performance of the all-new NeonSAN has been astounding, going by its laboratory assessment and user's actual deployment. Data shows that not only does its incredibly high IOPS performance and low I/O response time satisfy the stringent performance requirements of corporations for critical application loads, it also hits the mark with its ability to ensure business continuity, operational stability and reduce capacity expansion cycle. Currently, the highly acclaimed NeonSAN is deployed across numerous industries including finance, manufacturing and retail, and has successfully helped users to accelerate their business development and boost their decision-making efficiency.

What kind of database does a corporation need in the cloud era

The arrival of the cloud era has led the business world to cast a spotlight on how to effectively control and manage the massive data possessed by companies. Enterprise data may be ever-changing but to QingCloud, storage systems revolve around nothing more than these four core issues: Performance, security and reliability, scalability and costs. As one of just a handle of full stack cloud service providers, QingCloud introduced its "one-stop hybrid cloud", "hyper-converged integrated system" and other cloud service architecture products; but it did not stop there. It is now raring to go in the business of storage by partnering Intel to launch the all-new distributed block storage system "QingStor™ NeonSAN". This product leverages its enterprise-grade high performance, low latency and outstanding scale-out to help QingCloud add another milestone to its cloud services roadmap.

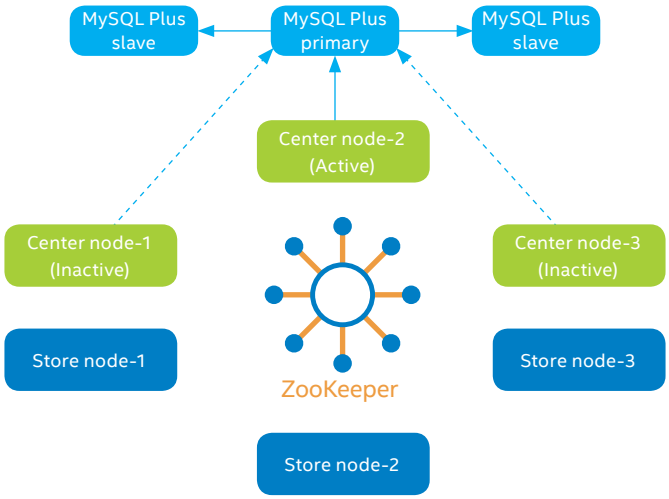

Figure 1 QingStor™ NeonSAN Cluster Architecture

As indicated in Figure 1, the NeonSAN storage cluster is composed of four types of functional components. For cluster management, QingCloud performs massive optimization on the distributed coordination services software Zookeeper* to allow it to undertake the cluster management, load balancing, active center node elections, distributed coordination and notification, and other management functions required by NeonSAN. The systems control layer in the cluster is known as the center node, which makes use of the one-master-multiple-slaves model to achieve high availability. When the active center node fails, Zookeeper will automatically generate a new active center node that takes over from it. For metadata management, NeonSAN uses distributed database cluster MySQL Plus featuring strong data consistency, second master-slave switch and other attributes, which robustly guarantees access of metadata by active center node. Lastly, NeonSAN also utilizes multiple store nodes to achieve data layer services.

To satisfy the core needs of corporate users, QingCloud provides NeonSAN with a few tricks up its sleeves. In terms of storage performance, NeonSAN is able to support remote direct memory access (RDMA) as well as a series of storage protocols and technologies, including InfiniBand*, RoCE*, NVMeoF* and iWARP*, which while enabling ultra-high performance and low latency help to reduce processor resources exhaustion. Information from QingCloud shows that when NeonSAN adopts the "all-flash memory + RDMA network" configuration, the single disk drive (volume) read/write performance of its 4K random read/write is able to achieve 100K IOPS and latency drops to 90 microseconds[2]. Such a performance can adequately satisfy the needs of financial industry users for high concurrency in mission critical, including transactions, queries and analyses.

Data security and reliability has always been a concern for corporate users. In view of this, NeonSAN sets up multiple security lines of defense to safeguard availability and security. The first line of defense is the versatile multiple replicas mechanism. Different disk drives (volumes) can assign different numbers of replicas which can be stored on different physical nodes. As seen in Figure 2, under this mechanism, after NeonSAN nodes receive a write request, besides writing into the primary database and replicas, the replicator process will also be sent from the replicas to other nodes. The nodes will respond back with a success message only after the writes to the primary database and replicas are successful. This mechanism guarantees robust consistency among the replicas.

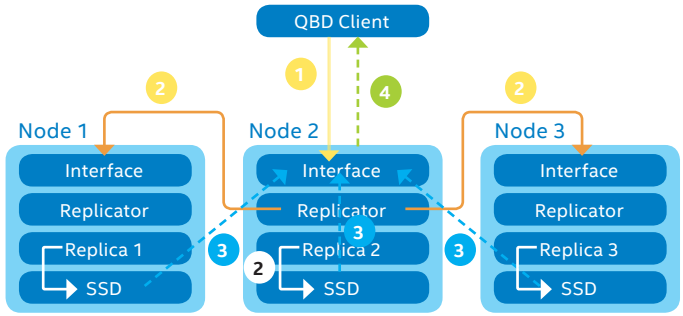

Figure 2 Data writing process under QingStor™ NeonSAN multiple replicas mechanism

The second line of defense is the automatic switch function when multiple paths and nodes fail. Every node of NeonSAN is equipped with dual network cards and each network card comes with dual ports. The four ports connect to four different switches. Among these switches, two front-end switches are redundancies for each other and constitute the front-end network that provides external data services. The other two RoCE switches are backups for each other: When a link malfunctions, the node will automatically switch to other links without disrupting the mission, ensuring high availability of services at the network layer.

The third line of defense is the "instantaneous snapshot" function and the "uninterrupted data recovery and migration" function. The former allows every single disk drive (volume) to retain 256 snapshots, such that in the event of data corruption, the snapshots can rapidly roll back to the data at a specific point in time for data recovery. The latter allows nodes to execute data recovery, data migration and capacity balancing at any time, which is not perceived by the upper-layer missions, and newly added nodes can immediately be put into work to reduce the impact of cluster capacity expansion and breakdowns on the missions.

The powerful two-directional scalability is one unique highlight of NeonSAN which places it apart from the traditional enterprise data center's SAN architecture. Built on x86 architecture-based hardware, NeonSAN makes use of the fully distributed architecture design to allow systems capacity and performance to scale horizontally. In the existing private cloud deployment plan, NeonSAN can support three to 1,024 nodes and is able to achieve smooth capacity expansion using single nodes as units without disrupting the missions. Test data from QingCloud indicates that its single disk drive (volume) can be expanded to 100TB[3], which fully satisfies the need for data storage under a Big Data context.

NeonSAN is able to satisfy another key concern of corporate users: Costs and price–performance ratio. In its private cloud deployment solutions, NeonSAN can provide thin provisioning, such as storage capacity, based on the user's actual needs. When an increase in user data leads to insufficient allocated capacities, the system will automatically make up for the shortfall from the back-end storage pool and significantly increase the storage resources utilization rate. Meanwhile, NeonSAN also supports the TCP/IP protocol, giving users the choice of the all-new RDMA protocol as well as the traditional TCP/IP protocol, which allows greater freedom and protects the users' existing investments to a large extent.

Capacity, performance balancing: Intel® Optane™ SSD + Intel® CAS

In the construction of traditional enterprise data centers, users tend to follow the simple philosophy of matching the best with the best by purchasing the most expensive computing storage and network equipment and assemble them together. However, such "ultra-luxurious combos" tend to fall short of the users' expectations, leading to a waste of resources and many compatibility problems. To QingCloud, the needs for storage systems vary across industries. Take for example the finance industry, under production contexts such as online transaction processing (OLTP), users seek high IOPS; while in strategic configuration, under the smart application contexts such as online analytical processing (OLAP), there is a higher demand for high throughput and large capacity.

To cater to the different application contexts of enterprise data center, QingCloud makes use of different combinations of technical architecture hardware to design various deployment modes for NeonSAN. First of all, for high-performance and large-capacity storage application contexts, QingCloud offers a deployment solution based on TCP/IP network and SAS/SATA HDD + SSD cache, satisfying the need for enormous data storage through tiered storage and mounting 12x SAS/SATA HDD with 4TB capacity on single nodes.

It is a well known fact that for a distributed storage system that supports tiered storage to achieve high performance, it must be able to read/write to cache with high efficiency. Factors that affect the performance of such NeonSAN deployment methods are mainly cache performance as well as systems management caching capabilities. In view of this, QingCloud introduces Intel's advanced technologies in these two areas.

The cache led by the all-new Intel® Optane™ SSD data center P4800X exhibits a mind-blowing performance on NeonSAN. This data center-grade Intel® Optane™ SSD is a unique combination of Intel® 3D XPoint™ memory media with Intel's advanced system memory controller, interface hardware, and software IP. Its low latency and stable performance far surpasses the traditional NAND media SSD, and is especially suitable for users such as e-commerce, finance and insurance companies as well as OLTP contexts with high concurrency. On top of that, the Intel® Optane™ SSD data center P4800X 375GB version currently used by NeonSAN boasts 30 drive writes per day (DWPD), which greatly guarantees the valid life cycle of users' systems[4].

To bring out greater performance from Intel® Optane™ SSD in NeonSAN, QingCloud even introduces another of Intel's exclusive technologies, Intel® CAS, designed to optimize caching performance. As shown in Figure 3, when data is first read by an application, NeonSAN will read out data from the back-end SAS/SATA storage and respond back to the application. Meanwhile, the data will also be replicated by Intel® CAS to the high-speed cache built by the data center-grade Intel® Optane™ SSD. In future reads, the application will directly read data at high-speed from the cache. And when writing data, all data will be written in sync to back-end storage and high-speed cache.

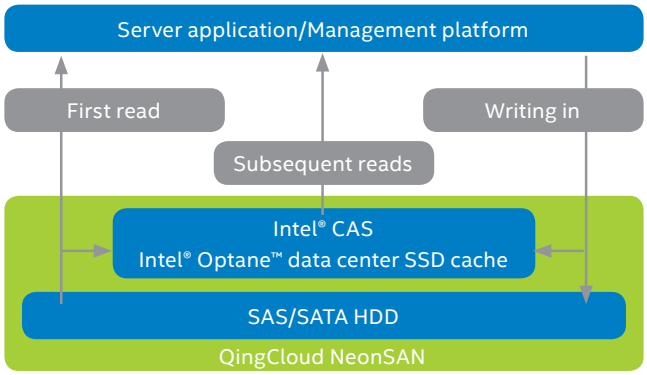

Figure 3 Diagram showing principles of Intel® CAS acceleration

When high-speed cache space is full, Intel® CAS's exclusive eviction algorithm will automatically replace the old, stale data in the high-speed cache with new active data. Evidently, with intervention by Intel® CAS, the application server will ultimately read the "hottest" data at the fastest speed. This has high practical significance for enterprise data centers which are seeing a widening gap in cold and hot data.

An Intel® CAS comparison test conducted by QingCloud also provides strong evidence to prove the above point made. In the FIO* test (a type of IO testing tool) performed on NeonSAN, a 4K random read/write test is executed and the combination of Intel® CAS + data center-grade Intel® Optane™ SSD – regardless of write back (WB) mode or write through (WT) mode – shows a performance of IOPS that far exceeds the "no-addition" combination in the comparison test group. Under WB mode, the random write performance of Intel® CAS and Intel® Optane™ SSD is even 23 times greater than the comparison test group[5].

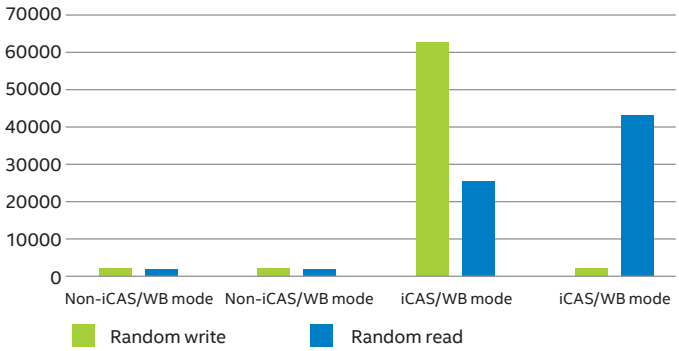

Figure 4 Comparisons of Intel® CAS FIO tests under 4K random read/write

Extreme storage performance: All flash memory + SPDK

Another deployment solution of NeonSAN is oriented towards ultra-high performance and low latency enterprise application contexts. For this, QingCloud offers an "all flash memory configuration" centering on Intel® SSD DC P4510. This SSD product complies with the standards of non-volatile memory express (NVMe) host controller interface and makes use of Intel® 3D NAND™ technology. The IOPS performance of this all flash memory configuration is a major improvement from the previous generation's SSD or traditional HDD. It also contributes to a significant decline in energy consumption and malfunction rate.

Data from a third-party evaluation organization indicates that NeonSAN configured with all flash memory exhibits 4K/8K random read/write performance of close to or more than 100K IOPS under single application stress with average response time of less than 0.8 millisecond. At the same time, as NeonSAN's volume quantity increases, its performance also exhibits linear growth. With a configuration of four NeonSAN volumes, the performances of 4K random read/write and 8K random read could hit around 300,000 IOPS, while 8K random write performance exceeds 250,000 IOPS, with average response time of less than 1 millisecond[6].

The Intel® SSD based on NVMe standards provides users with high throughput and low latency storage capabilities, and at the same time the SSD-oriented storage product – Intel® Storage Performance Development Kit (Intel® SPDK) – launched by Intel is also upgrading storage software performance using various innovative technologies.

In the HDD era, the interrupt overhead makes up only a very small proportion of the entire IO process and the impact is thus not prominent. As SSD products gain traction, especially with the arrival of NVMe SSDs, the powerful IO performance can suddenly "draw attention" to the interrupt overhead, moving the system bottleneck from storage hardware to software. In view of this, Intel® SPDK provides several innovative technologies revolving around three points: UIO/VFIO (userspace driver), asynchronous polling and lockless design, which can help upper-layer applications to make full use of the high performance brought about by the NVMe SSDs. In userspace drivers, the developed libraries made from Intel® SPDK can map the base address register (BAR) of the storage equipment in the process space of the application, thereby lowering the interrupt overhead where possible in a show of NVMe SSD's edge in performance.

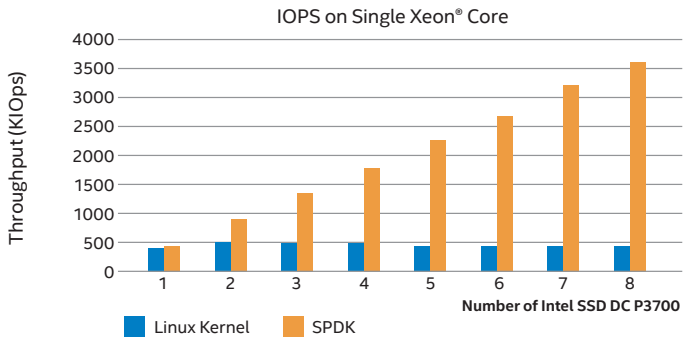

Figure 5 Comparisons between Intel® SPDK userspace driver and kernel driver[7]

On a single processor core, userspace driver can manage multiple NVMe SSD devices, thereby achieving high throughput, low latency, high-efficiency use of processor resources and other advantages. As shown in Figure 5, in a comparison test between Intel® SPDK userspace driver and kernel driver, the matching of single processor core with SPDK enables the 8 Intel® SSDs to unleash their full power; whereas in order for kernel driver to attain the same level of performance, it would require a configuration of at least eight processor cores.

The Intel® SPDK userspace driver uses asynchronous polling to test the equipment's completion state. After the application submits a read/write request, there is no need to wait for the read/write operation to complete. It can continue to send out requests based on its needs, before processing in callback completion, thereby avoiding the latency and overhead resulting from interrupt while increasing IOPS. In fact, polling for NVMe SSD equipment is highly efficient. Based on the NVMe standards, storage equipment will test whether a completed queue has new operations completed by reading the memory; and via the Intel® Data Direct I/O (DDIO) technology, the data after equipment update can be stored in the processor cache to achieve high-performance equipment access.

Intel® SPDK features a lockless design which eliminates the reliance of data channels on locks. Intel® SPDK binds one processing thread to a specific process core via the thread affinity method, while holding onto the utilization of the core via polling. Meanwhile, via the method of run to completion, the entire life cycle of the application's read/write request is bound to the specific processor core till completion. In this way, precious processor resources are not reused on data between synchronized threads and thereby IOPS performance is further enhanced.

NeonSAN's performance is significantly improved after the introduction of Intel® SPDK. QingCloud's test data shows that regardless of single-replica or multiple-replica configurations, the latency of random write drops by around 10 microseconds. And in the case of multiple replicas, the latency of random read can go down by 20 microseconds. Under a mixed read/write context (read/write ratio: 70/30), in a NeonSAN cluster with three-node configuration, the volume read/write performances of two replicas went up nearly 20%[8].

Use in practice by corporate users

Oracle* Real Application Clusters* (RAC*) is currently the most popular database integration environment in the business world. That is why QingCloud teamed up with Intel to perform massive testing and optimization on Oracle* RAC*, and proceeded with actual deployments at several users. The feedback was satisfying.

In an evaluation of a simulation of Oracle* database + NeonSAN node, it was found that the transactions per minute (TPM) of the whole storage system exceeded 1.65 million and the average transactions per second (TPS) was close to 30,000. In addition, the average latency in the completion of each transaction processing was around 15 milliseconds. Such performance can adequately support the critical application loads of a majority of enterprises[9].

During its cooperation with a major retailer, not only did QingCloud provide full stack cloud computing solutions to construct the group's cloud platform, it also helped the retailer to deploy six-node NeonSAN cluster in the production zone. As seen in Figure 6, the retailer already owned an Oracle* RAC* database in its original business environment, it only needed to add NeonSAN as a shared disk to perform industry/business data migration and the results showed good scalability.

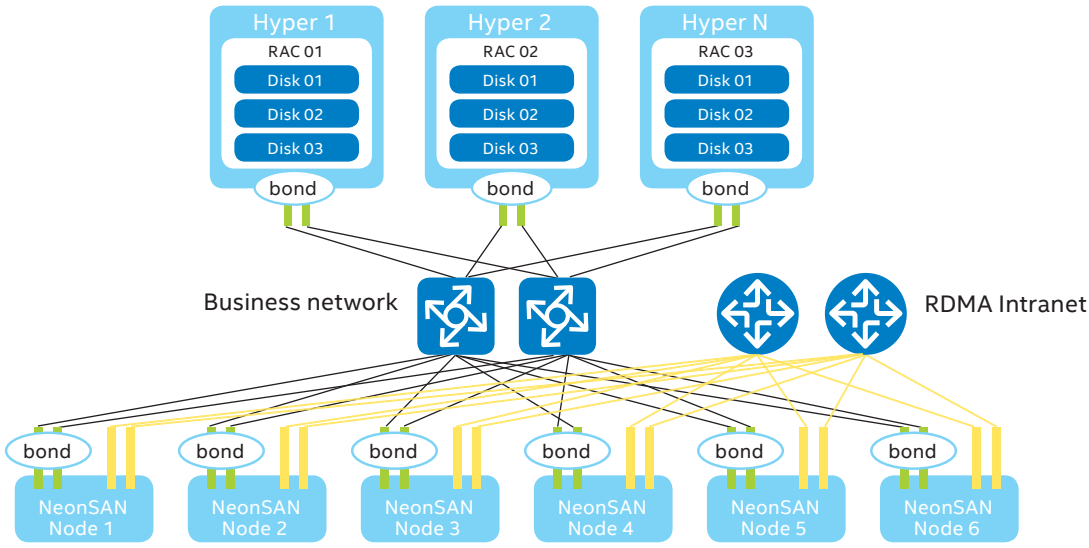

Figure 6 Storage system architecture of a major retailer: Six-node NeonSAN storage cluster deployed at back end

In its feedback, the retailer noted that the deployment of NeonSAN effectively propelled the evolution of its core ERP system towards cloud computing architecture, resulting in integrated operations and management under a private cloud environment. This helps to ensure business continuity while slashing purchasing and operating costs. At the same time, NeonSAN's scalability shortened the period for building and capacity expansion of the storage system from a few months to one week. This effectively satisfies the need for capacity expansion when there is a surge in business data volume and practically gave impetus to the smooth and rapid development of the business systems.

In the case of another corporate client in the finance sector, its various business are seamlessly connected online to QingCloud cloud platform and NeonSAN is now at the core of its businesses, especially the storage engine of OLTP business context. The finance company's actual test data shows that the NeonSAN-based complex view query time shortened by about 90%, and the execution efficiency of complex SQL words reduced from minutes to seconds[10].

Based on the positive reviews from many case studies, it is evident that the QingStor™ NeonSAN created by QingCloud and Intel has won over the market and users. The two partners will continue to deepen their cooperation in the future, using advanced products and technologies to boost the performance of enterprise data centers. Currently, QingCloud is gradually replacing the processors of NeonSAN's systems with the new-generation Intel® Xeon® Scalable Processors and has plans to further explore the underlying potential of these processors – especially the functions of Intel® Advanced Vector Extensions 512 (Intel® AVX-512), Intel® Virtual RAID on CPU (Intel® VROC), Intel® Trusted Execution Technology (Intel® TxT) – in order to cater to the need for growing computing power of enterprise storage systems and create highly efficient and differentiated cloud storage services.

[1] This data is taken from "White Paper On Big Data (2018)" published by China Academy of Information and Communications Technology

[2, 3] This figure is cited from the relevant product performance specifications on QingCloud's official website: https://www.qingcloud.com/products/qingstor-neonsan/

[4] Source of data: https://www.intel.cn/content/www/cn/zh/solid-state-drives/optane-ssd-dc-p4800x-brief.html

[5, 8] The above test data is taken from an unpublished QingCloud test report.

[6] This test data comes from the report "Facing Key Applications: QingCloud NeonSAN's performance experience"

[7] This test data is based on the following test environment: Intel® Xeon® Processor E5-2695v4 single processor core, 4K random read test, 128 queue depth, CentOS* Linux* 7.2, Linux kernel 4.10.0, with test SSD being 8x Intel® SSD P3700 NVMe (800GB).

[9] The test data is taken from the report "Key Applications: Experiencing QingCloud NeonSAN's Performance"

[10] This test data is taken from the "Performance Test Report In DT Era: What Kind Of Storage Engine Allows Oracle Database Performance to Increase By 100%?",https://www.sohu.com/a/249625689_464027