Intel Helps Tencent Cloud Deeply Optimize its Cloud Block Storage (CBS) to Create Ultra-fast Cloud Storage Experience

Preface

As many of today's enterprises are adopting cloud technologies in their core systems, cloud storage has become an important carrier of business data, and its performance has been attracting more and more attention. Tencent Cloud is one of the world's leading cloud service providers. Its cutting-edge Cloud Block Storage (CBS) provides highly efficient and reliable persistent block storage services for many industrial users and is widely deployed and used in various scenarios such as core database, Content Distribution Network (CDN), and e-commerce systems.

To provide users with high-performance enterprise-level cloud storage services, Tencent Cloud collaborated deeply with Intel to reconstruct and optimize its ultra-fast solid-state drive CBS with a brand-new storage engine design and Intel® Optane™ Persistent Memory. It has been verified that with better bandwidth, lower latency, and higher Input/Output Per Second (IOPS), the new solution can create an extremely fast cloud storage experience for performance-intensive business scenarios.

Challenges: Fast-evolving Cloud Services Place Greater Demands on Cloud Storage Performance

Whether for emerging industries such as Internet, big data, and artificial intelligence, or traditional sectors such as finance, medicine and manufacturing, cloud services have gradually become one of the standards for the next-generation IT infrastructure. Cloud storage products and solutions such as cloud disks will be an important carrier of future business data, so their performance is a key factor to consider for enterprises when choosing cloud services.

As one of the world's leading cloud service providers, Tencent Cloud has been providing persistent block storage services for users with its cutting-edge CBS cloud disk. Figure 1 shows a typical Tencent Cloud CBS storage system

architecture. The system consists of the CBS access, MDS control cluster, and CBS storage cluster. When receiving a data read-write request from the CVM cloud host cluster, the CBS access forwards the request to the corresponding CBS storage node according to the cluster routing information provided by the MDS.

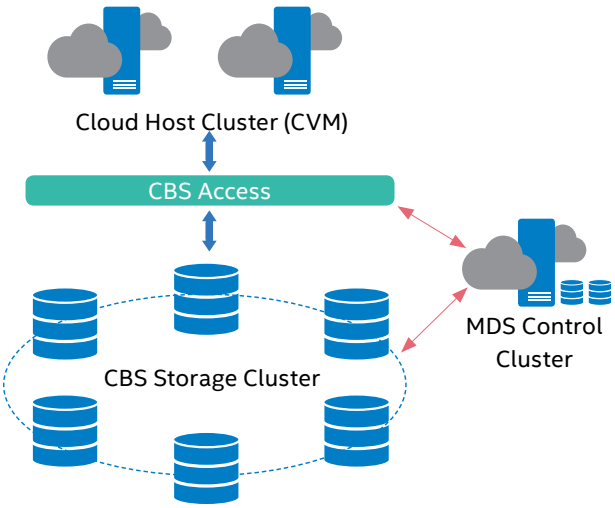

Figure 1. Storage System Architecture of Tencent Cloud's CBS

Built upon its long-standing technology accumulation and continuous technical optimization and evolution, Tencent Cloud's CBS offers excellent performance, availability, reliability, and scalability:

- **High Performance:** With the effective combination of Intel's high-performance NVMe SSDs and Tencent Cloud's innovative proprietary storage engine, CBS provides up to 1.1 million random IOPS on a single disk and a bandwidth capability of up to 4Gbps per second for users' business scenarios.
- **High Availability:** With its high availability and disaster recovery design, CBS can effectively minimize the probability of the system being unavailable, and back up user data through snapshots, which prevents data loss due to tampering or accidental deletion and ensures quick rollbacks in the event of business failure.
- **High Reliability:** Using the 3-copy distributed mechanism, CBS provides users with up to 99.9999999% data reliability.
- **High Scalability:** CBS allows users to configure storage capacity based on their business needs and scale on demand. Currently, a single disk supports up to 32 TB capacity, and a single cloud host can mount a total of 640 TB, enabling users to easily deal with the TB/PB-level big data processing.

With these advantages, Tencent Cloud's CBS performs well across various business scenarios, including the high-load On-line Transaction Processing (OLTP) financial transaction systems, high-throughput e-commerce systems, data analytics systems for artificial intelligence, and high-concurrency CDN, etc. Its performance has received positive feedback from customers.

However, it can be seen from the CBS architecture that the distributed storage cluster could cause latencies in network access and transmission, which will reduce its overall performance, leading to a performance gap with that of the local storage. This is also one of the reasons why users sometimes hesitate to choose CBS in performance-sensitive scenarios such as core databases and CDN. As cloud services are gradually rising to become a core carrier of business systems, the demands for data read-writes in increasing and more complex core businesses are driving Tencent Cloud to further optimize its ultra-fast CBS to boost performance and eliminate such concerns.

Based on the CBS architecture, storage engine, and hardware infrastructure, Tencent Cloud introduced the Remote Direct Memory Access (RDMA) protocol and partnered with Intel to carry out comprehensive optimizations, including:

- Adding mechanisms such as Round Robin, algorithm optimization, lock and contention elimination, to optimize the CBS storage engine.
- Introducing the Storage Performance Development Kit (SPDK) provided by Intel, to optimize the IOPS and latency of NVMe SSDs.

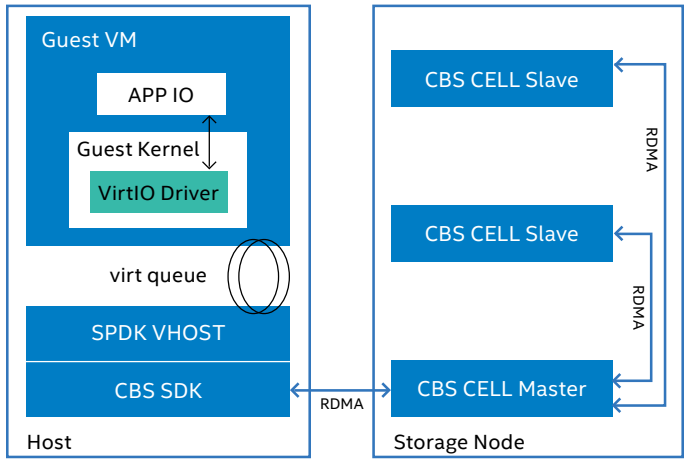

Figure 2. Ultra-fast CBS Architecture

After optimizations on the architecture, engine, and software, Tencent and Intel found the latency of the SSD itself can also be an obstacle for further performance enhancement. The most effective solution to address the issue is to find a storage medium with better performance.

Therefore, Tencent Cloud and Intel leveraged the Intel® Optane™ Technology, a cutting-edge memory and storage technology, and used Intel® Optane™ Persistent Memory as the storage core of the next-generation ultra-fast CBS. They also reconstructed how data is written to the storage medium to meet the latency requirement in performance-intensive scenarios.

Solution: Intel's cutting-edge Storage Technology Empowers Ultra-fast CBS for Better Performance

Figure 3 shows how data is written to the storage in the existing CBS design. The cloud host data from the computing cluster is hashed or allocated to the corresponding block node, and then cached to different Pages. Next, the system needs to perform two write operations, one for the business data to be written to the corresponding data area in SSD, and the other for the metadata to be written (wAppend) to the SSD in the form of LOG.

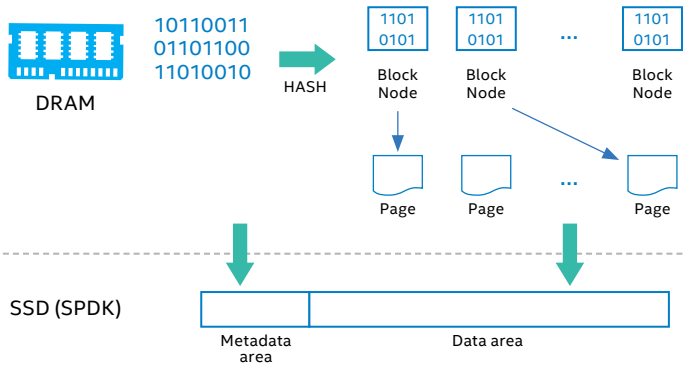

Figure 3. Data Writing in the Existing Ultra-fast CBS Design

As shown in the figure, two writes are needed for this process. The write latency of SSD based on NAND flash memory is usually tens of microseconds, so two writes add up to a latency of tens or even hundreds of microseconds. It seems minor, but in the 5G era where end-to-end network latency can be only 1 millisecond (1,000 microseconds), this latency level will obviously constrain the overall performance of CBS.

Meanwhile, data is written to the NAND SSD in blocks, and erase operation is needed before writing. This will drag down writing efficiency, and greatly reduce the service life of SSD (namely the "write amplification" issue). Moreover, LOG recycle will cause glitches.

Intel® Optane™ Persistent Memory, developed based on Intel® Optane™ Technology, can help CBS to effectively handle those issues. Intel® Optane™ Technology uses a whole new transistor-less storage architecture that stacks storage grids in a three-dimensional matrix to improve density, increase read-write performance and provide persistence. Intel® Optane™ Persistent Memory is also byte-addressable, so it can control the location and size of data read and write just like memory does.

Figure 4. Intel® Optane™ Persistent Memory 200 Series

Compared with the traditional DRAM, the Intel® Optane™ Persistent Memory built on Intel® Optane™ Technology, together with other advanced storage control technologies, hardware interface, and software enhancements by Intel, has two significant advantages: first, it has higher storage density and lower unit storage cost, allowing users to expand their cloud storage capabilities more economically; secondly, with the persistence feature, the Intel® Optane™ Persistent Memory configured in App Direct mode can serve as an effective persistent storage carrier for CBS.

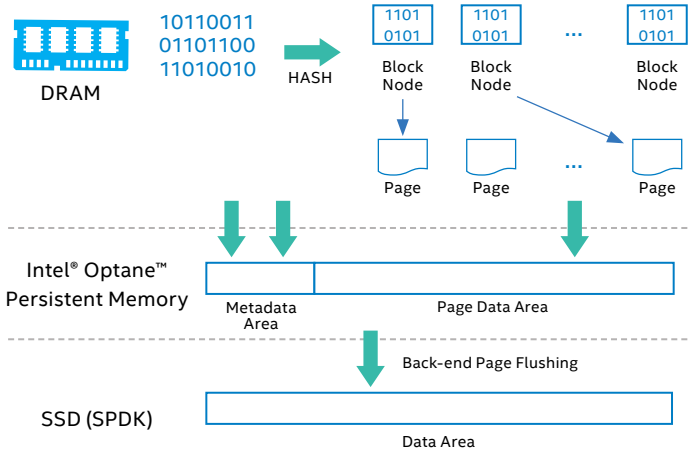

Figure 5. Data Writing in the Optimized Ultra-fast CBS Design

Thanks to the innovative features of Intel® Optane™ Persistent Memory, data writing in the ultra-fast CBS has been optimized,

as shown in Figure 5. The data from the computing cluster is first hashed to the corresponding block node and cached in Page, and then instantly stored persistently into Intel® Optane™ Persistent Memory. Meanwhile the Page/Block metadata is updated locally to the corresponding data area.

In addition to data writing optimization, users can customize their strategies and algorithms, which will allow them to decide whether or not to further flush the data from Intel® Optane™ Persistent Memory to the SSD. For example, "hot data" that needs to be read and written frequently can be stored in the persistent memory, while "cold data" that is not accessed frequently can be transferred to the SSD by the back-end to effectively reduce the Total Cost of Ownership (TCO) of CBS.

In addition to the advanced storage hardware, the Persistent Memory Development Kit (PMDK) offers a programming model and environment for Intel® Optane™ Persistent Memory.

For example, libpmem is a low-level library in PMDK that supports accessing persistent memory through memory mapping, so that files in persistent memory can be mapped to the virtual memory space of applications for further operation. This saves the overhead inflicted by kernel participation and context switching, and applications can directly benefit from the high performance of persistent memory.

Moreover, the libpmem library can detect the features of the processor and use the most efficient instructions (e.g., CLWB, CLFHASHOPT) to write data into persistent memory. The CLWB instruction allows for multiple cache-line write-backs to proceed in parallel while also maintains the validity of processor cache after refreshing the data. The libpmem is also packaged with the Non-Temporal Write (NTW) instruction, which can bypass the processor cache using write combining, and write data from the Store Buffer directly to the WPQ of the memory controller to improve performance.

These features not only allow the libpmem library to help users achieve more refined and accurate control of the entire write process with rich interfaces, but also improve the overall write performance of the entire system using NTW write instructions and memory mapping to access the persistent memory, maximizing the benefits of Intel® Optane™ Persistent Memory features in the new CBS design.

Results: Innovative Hardware and Optimized Design Bring Comprehensive Benefits to CBS

Compared to the existing solutions, the implementation of the optimized CBS based on Intel® Optane™ Persistent Memory has delivered tremendous improvements and benefits, including:

Significant reduction in read-write latency:

- Compared with the read-write latency of NAND SSD, which is measured in tens of microseconds, Intel® Optane™ Persistent Memory latency can be controlled within 1 microsecond.
- With the libraries and tools from PMDK, Intel® Optane™ Persistent Memory can achieve more refined and accurate control of the entire write process, and effectively improve the write performance of the system.

Increased system service life:

- The byte addressability of Intel® Optane™ Persistent Memory resolves the "write amplification" problem observed with NAND SSDs. Recurring writes and erases are no longer needed, contributing to longer service life of devices.
- The unique storage structure enabled by Intel® Optane™ Technology also helps increase the service life expectancy of Intel® Optane™ Persistent Memory.

Enhanced efficiency of storage space usage:

- Intel® Optane™ Technology allows memory cells to be individually accessed and updated, so Intel® Optane™ Persistent Memory does not need to perform garbage collection, avoiding the glitch problems with NAND SSD, hence improving the efficiency of storage space usage.

To verify the benefits of the new hardware and optimized design to CBS, Tencent Cloud and Intel carried out multi-faceted tests. As shown in Figure 6, the optimized CBS built with Intel® Optane™ Persistent Memory reduces the overall write latency from 120 microseconds to 60 microseconds and the overall read latency from 130 microseconds to 40 microseconds. Moreover, IOPS reaches over 2 million. The performance has been effectively improved[1].

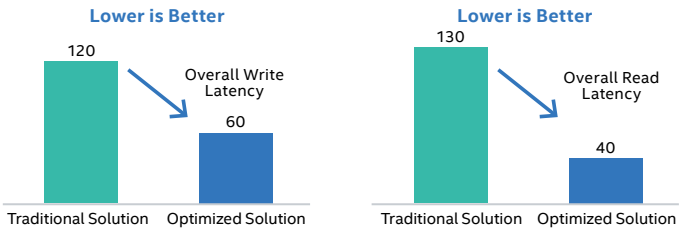

Figure 6. Read-write Latency Significantly Reduced with Optimized CBS

Looking Ahead: Creating Better Cloud Storage Experience for Users with Advanced Products and Technologies

With the continuous improvement of cloud computing and cloud storage technologies, cloud services are more and more important in enterprise-level business systems. It is believed that users will have more and higher requirements on cloud service performance. These technologies and their applications in various scenarios are driving the constant evolution and optimization of related products and technologies. As critical participants and leaders in the cloud service industry, Tencent Cloud and Intel collaborated on the optimization of CBS based on Intel® Optane™ Persistent Memory, which has been a great success. This is just a typical example of the trend.

In the future, building on this success, Tencent Cloud and Intel will conduct wider cooperation in cloud computing, cloud storage, and other fields, and use more advanced products and technologies to continuously optimize cloud service products such as CBS. For example, both sides plan to add the RDMA protocol to solutions based on Intel® Optane™ Persistent Memory to bring down processor and memory overheads. In addition, the new 3rd Gen Intel® Xeon® Scalable Processors, with more cores, optimized architecture and larger memory capacity, will enable stronger performance for cloud service products. It will work better with the next-generation Intel® Optane™ Persistent Memory to enhance cloud storage experience, making cloud storage products such as CBS to be a reliable pillar for future enterprise-level business data storage.