Whitepaper

Low Costs and High Performance: You Can Have Both — Intel® Optane™ Persistent Memory Empowers Alibaba Cloud Services

The persistent memory-optimized instance re7p is built on Alibaba Cloud's 3rd-generation X-Dragon architecture and Intel® Optane™ persistent memory 200 series. Compared with the previous generation instance re6p, re7p has doubled the network bandwidth and storage bandwidth and increased the computing power by over 40%. In addition, the performance-enhanced local disk instance i4p, powered by Intel® Optane™ persistent memory, has a read/write latency of as low as 170 nanoseconds, delivering a performance 2 orders of magnitude higher than that of conventional NVMe SSDs. Furthermore, i4p supports over 1,400,000 IOPS per disk and a throughput of up to 9 GB/s per disk, greatly improving the performance of I/O-intensive applications. For example, with i4p, the performance of the RocksDB database improves by 2.5 times, the performance of the ClickHouse database improves by 2 times, and the recovery time of NSQ message Queue service by 2-3 times.

— Tang Xianghua,
expert of Alibaba Cloud elastic computing products

Contents

Introduction

According to the 2021 China Cloud Computing Market Report released by Canalys, a global analyst firm, [1] China's cloud infrastructure market registered an annual growth of over 30% in 2021 to a total of US$27.4 billion, making it one of the fastest-growing markets in the world.

As China ushers in the "Smart+" era, cloud computing has become the foundation of the digital economy and the driver of the digital development of small and medium-sized enterprises. Especially since the outbreak of COVID-19, the demand for telecommuting, online education, and online conferences has exploded, which has further advanced the rapid growth of the cloud computing market.

Alibaba Cloud is a global leader in cloud computing and artificial intelligence. Relying on Alibaba's robust technical strengths and business scenarios, Alibaba Cloud has gathered top experts in the field of cloud computing at home and abroad and is committed to building a public, open cloud computing platform. Driven by technological innovation, Alibaba Cloud will further improve its computing power and economies of scale to turn cloud computing into a real public service.

Elastic Compute Service (ECS) is a high-performance, stable, reliable, and scalable IaaS (infrastructure-as-a-service) service provided by Alibaba Cloud. ECS eliminates the need for upfront investments in IT hardware and allows you to scale computing resources on demand. This makes ECS instances more convenient and efficient than physical servers. ECS provides a variety of instance types that suit various business needs and help boost business growth.

Challenges

With the advancement of digital transformation in China, most Chinese enterprises have chosen to migrate to the cloud. Digital transformation enables continuous business growth and helps enterprises maintain sustainable development. However, the rapid development of digital technology has also increased the IT budget of enterprises, putting them under pressure to cut costs while improving efficiency.

For example, in the enterprise cloud transformation journey, business lines are getting increasingly complex and data volume is growing rapidly. This puts higher requirements for data storage capacity, real-time response to data, and system stability. In previous solutions, the hot or active data are stored in the memory for guaranteed performance, but the price per GB of DRAM remains high, resulting in high DRAM costs, and existing DRAM DIMM densities also restrict the physical capacity of system memory. As a result, it is difficult for companies to store all hot data to the memory.

[1] Quoted from the "2021 China Cloud Computing Market Report" released by Canalys

For warm data stored in the storage, in some nearline storage scenarios, the growing demand for low latency exceeds the limit of NAND SSDs because the performance and durability of NAND media are inadequate.

In this case, enterprises have to either expand capacity or reduce costs, and a new demand gap is formed between memory and storage, as shown in Figure 1.

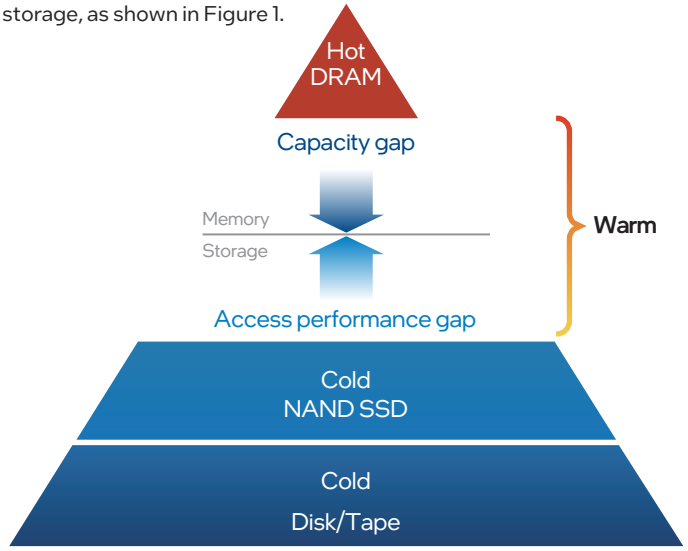

Figure 1

Therefore, as enterprises migrate to the cloud, the balance between costs and security and the full integration of various elements in the cloud computing ecosystem for rapid business and technological innovation have become the major problems. As the provider of core products at Alibaba Cloud's IaaS layer, ECS needs to provide a suite of stable, reliable, and cost-effective solutions with excellent performance.

To cater to the demand for cost-effective, high-performance solutions, Alibaba Cloud chooses to bring the persistent memory technology to the cloud and integrate Alibaba Cloud's X-Dragon architecture and Intel Optane persistent memory to build persistent memory products (ecs.i4p and ecs.re7p) in the cloud.

Solutions

Alibaba Cloud is the first vendor in the world to launch cloud servers powered by Intel Optane persistent memory. The latest persistent memory-optimized instances — re7p, r7p, and i4p — launched in 2021 are built on Alibaba Cloud's 3rd-generation X-Dragon architecture and the Intel Optane PMem 200 Series. Compared with previous-generation products, they have doubled the network bandwidth and storage bandwidth and increased the computing power by over 40%.

In addition, Alibaba Cloud also launched the persistent memory-optimized and performance-enhanced local disk instance i4p. Combined with Alibaba Cloud's X-Dragon architecture, i4p has achieved significant performance improvement compared with conventional NVMe local disk instances. i4p supports a read/write latency of as low as 170 nanoseconds, over 1,400,000 IOPS per disk, and a throughput of up to 9 GB/s per disk, greatly improving the performance of I/O-intensive applications.

Intel Optane persistent memory

PMem is a revolutionary memory product built on 3D XPoint media. It features high speed, low latency, large capacity, cost-effectiveness, persistent data storage, and advanced encryption.

Intel Optane persistent memory helps bridge the gap between performance and capacity that DRAM and NAND storage just cannot deal with. The breakthrough Intel® Optane™ technology allows users to choose configuration modes as needed for better total memory capacity and higher virtual machine density to increase the number of virtual machines a server can maintain and improves the server consolidation rate. Intel Optane persistent memory offers a reliable, cost-effective solution that provides data centers and cloud applications with sufficient capacity and performance to cater to the new wave of data demand.

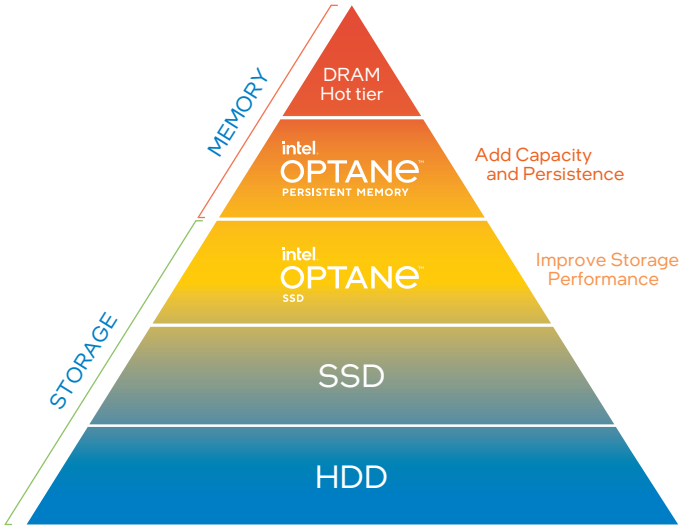

Figure 2: In the storage hierarchy,
Intel® Optane™ persistent memory is just below DRAM.

Intel Optane persistent memory also excels in its byte addressability. Conventional databases mainly rely on disks to store data, but disks are not byte-addressable, so databases are generally read and written in blocks (such as 4 KiB). By comparison, PMem supports the direct access (DAX) capability, which allows applications to directly access persistent memory media without kernel participation, interruption, or context switch. DAX fully exposes the performance of persistent memory to applications, thereby greatly simplifying underlying I/O and speeding up query speed.

Intel Optane persistent memory supports two modes: memory mode and App Direct (AD) mode.

In the memory mode, the CPU memory controller treats persistent memory as volatile memory and DRAM memory as the cache for persistent memory. Given the greater memory capacity, when data is requested in this mode, the memory controller will first check DRAM memory. If the data is present, it will be retrieved directly from DRAM memory; if not, it will access persistent memory.

In the AD mode, software and applications that support the standard SNIA persistent memory programming model can communicate directly with Intel Optane persistent memory and utilize its byte addressability to access smaller block files (such as 64-byte, 128-byte, and 256-byte files). This reduces latency significantly and prepare enterprises for high-speed business scenarios. Most data is stored in persistent memory with large capacity, while indexes are stored in DRAM, so that all keyword queries are performed in DRAM memory with higher speed. In this way, DRAM memory and persistent memory can work together to deliver better performance.

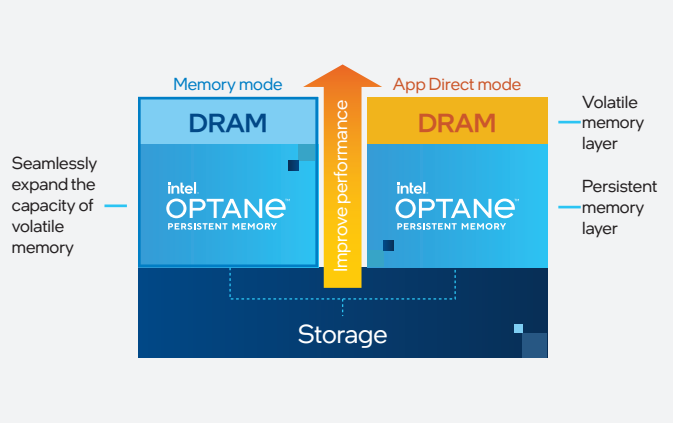

Figure 3: Two modes of PMem

Memory-optimized instance re7p

The memory-optimized instances r7p and re7p of ECS are built on the 3rd-generation X-Dragon architecture developed by Alibaba Cloud and Intel Optane PMem 200 series. The instances offer a vCPU_number-to-memory_capacity ratio of up to 1:20, that is, 20 GB of memory per vCPU, including 4 GB of common memory and 16 GB of persistent memory. For memory-intensive applications, greater memory ratio means higher cost performance.

The instances are suitable for: in-memory database Redis; applications that require a large page cache, such as RocketMQ; Hadoop clusters, Spark clusters, and other memory-intensive enterprise applications. Such instances can help greatly reduce the total cost of ownership (TCO) per GiB.

To test the performance of re7p powered by Intel Optane PMem 200 series, Alibaba Cloud's test team deployed an in-memory database Redis application. Redis is an in-memory but persistent on-disk database that supports not only simple key-value pairs, but also more complex data structures such as list, set, zset, and hash. By relying on the main memory of computer for data storage, Redis can provide a higher data throughput bandwidth and a lower latency in data processing than SSDs do, significantly improving the speed of data processing. The 1 ms response latency is unmatchable in the database world.

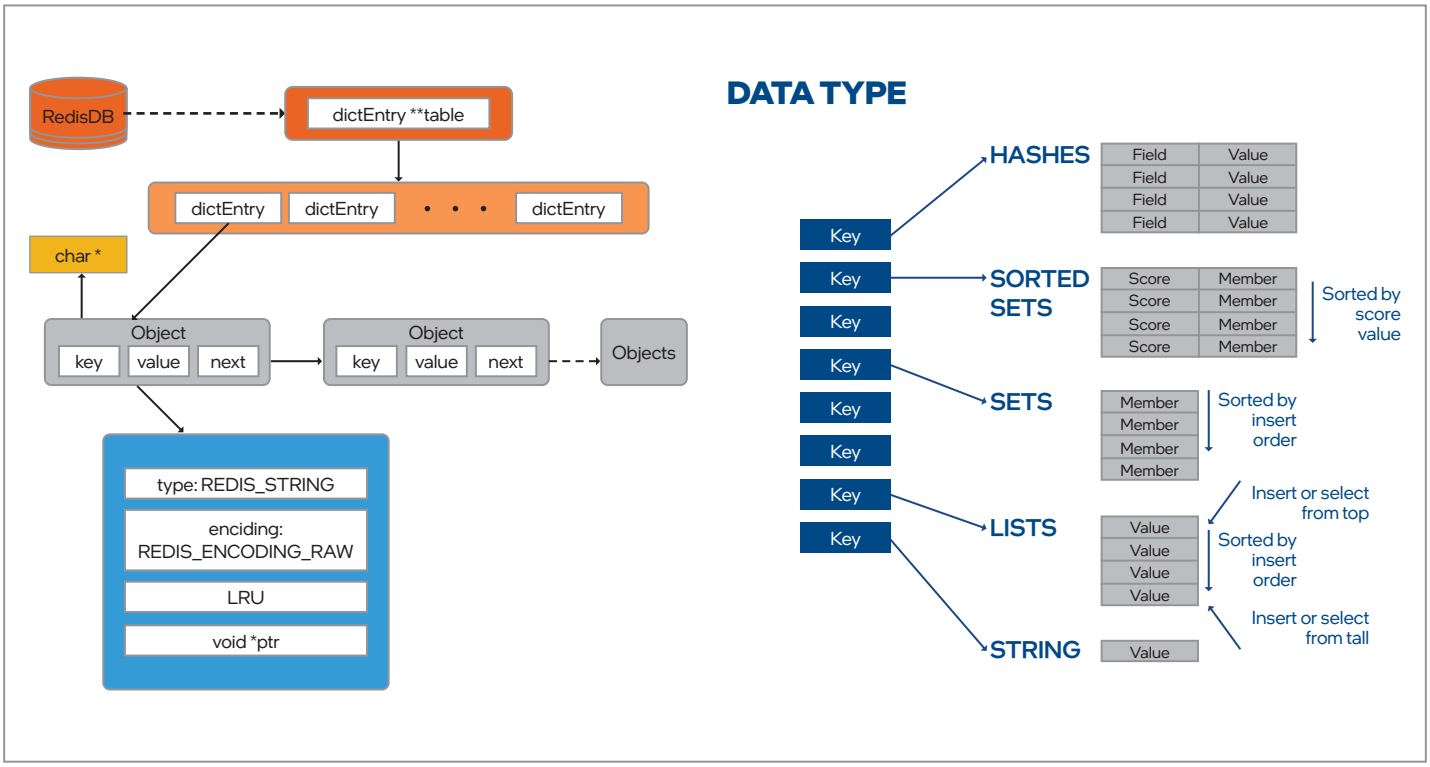

Figure 4: Redis architecture

However, it was found that during the deployment, the Redis community of Antirez was unable to meet the rapid iteration requirements of Chinese internet customers in a timely manner. Therefore, Alibaba Cloud's technical team used TieredMemDB [2] instead. TieredMemDB is an open-source Redis fork that takes full advantage of DRAM and Intel Optane technology. It is fully compatible with Redis and supports all the structures and features of Redis.

With the KMEM DAX feature provided by Linux kernel 5.1 or higher, TieredMemDB uses a dynamic threshold algorithm to manage data distribution. Data larger than or equal to the threshold is stored in persistent memory, while data smaller than the threshold is stored in DRAM. TieredMemDB regularly monitors the allocator statistics related to DRAM and persistent memory to adjust the dynamic threshold accordingly, so that the distribution of data in DRAM and persistent memory always conforms to the preset ratio for optimal performance. In addition, different ratios of DRAM and persistent memory are set for different customer instances to manage customers and QoS by hierarchy.

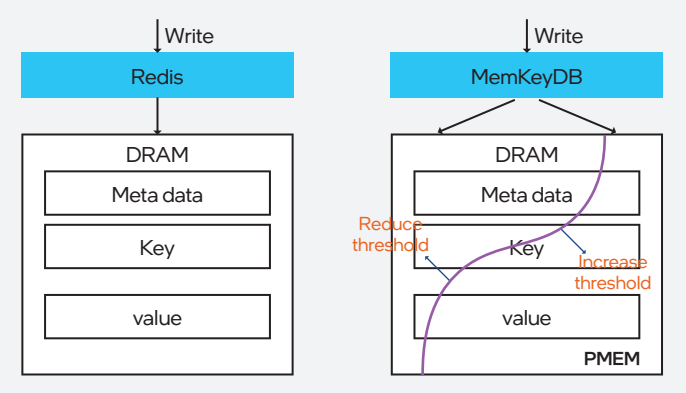

Figure 5: Controlling the ratio of data in DRAM and persistent memory via a dynamic threshold

The test compared a re7p instance with a r7 instance that does not use persistent memory. The standard OpenSource Redis6.0.5 ran on r7.2xlarge, which was the test baseline, and the TieredMemDB of the same version of Redis ran on re7p.2xlarge. The test result is as follows:

Instance Type	re7p.2xlarge	r7.2xlarge
vCPU	8	8
Memory (GiB)	160	64
Read QPS	848445.52	767780.45
Write QPS	804766.93	780125.3
Read Latency (ms) (p999)	1	1
Write Latency (ms) (p999)	1	1
Memory Price per GB (CNY/month)	11.87	22.46

According to the above table, the memory-optimized instance re7p can achieve similar or even higher read QPS and write QPS than the OpenSource Redis-based instance r7 at a lower price per GB of memory. The re7p also meets the service level agreement (SLA) (read/write latency < 1 ms) and provides more memory capacity.

Performance-enhanced local disk instance i4p

The performance-enhanced local disk instance i4p of ECS, based on Alibaba Cloud's 3rd-generation X-Dragon architecture and Intel Optane PMem 200 series, offers high-performance local disk. It has a maximum of 3 million IOPS, and a single-access latency as low as 30 microseconds.

This product is suitable for on-disk KV databases including RocksDB and ClickHouse; OLTP and high-performance relational databases (for WAL optimization); NoSQL databases such as Cassandra, MongoDB, and HBase; Elasticsearch and other search scenarios; and other I/O-intensive applications such as message-oriented middleware and containers.

Compared with the traditional way of directly storing data into local disk, the read/write performance of running RocksDB on i4p is improved by 2-2.8 times, the latency falls by 23%, and the overall data reliability is higher:

a. SST data is stored in ESSDs with a 3-replica protection mechanism, which frees you from the worry of data loss;

b. WAL logs are stored in BPS to ensure data reliability since local disk based on the BPS simulation have a failure rate much lower than conventional SSDs and HDDs.

The RocksDB project started at Facebook as an experiment to enable the full potential of data storage for fast storage. RocksDB borrowed core code from the open-source LevelDB project and important ideas from Apache HBase. The initial code is forked from open-source leveldb 1.5.

Based on the LSM-Tree data structure, RocksDB can be tuned to run on a variety of production environments (pure memory, flash, hard disks, or HDFS) and supports different compression algorithms. The main design point of RocksDB is its superior performance under fast storage and high service pressure, so this DB needs to fully exploit the read and write rates of flash and RAM.

RocksDB supports efficient point lookup and range scan operations and needs to support the configuration of various parameters to optimize performance when high-pressure random reads, random writes, or both cause heavy traffic.

[2] https://github.com/TieredMemDB/TieredMemDB

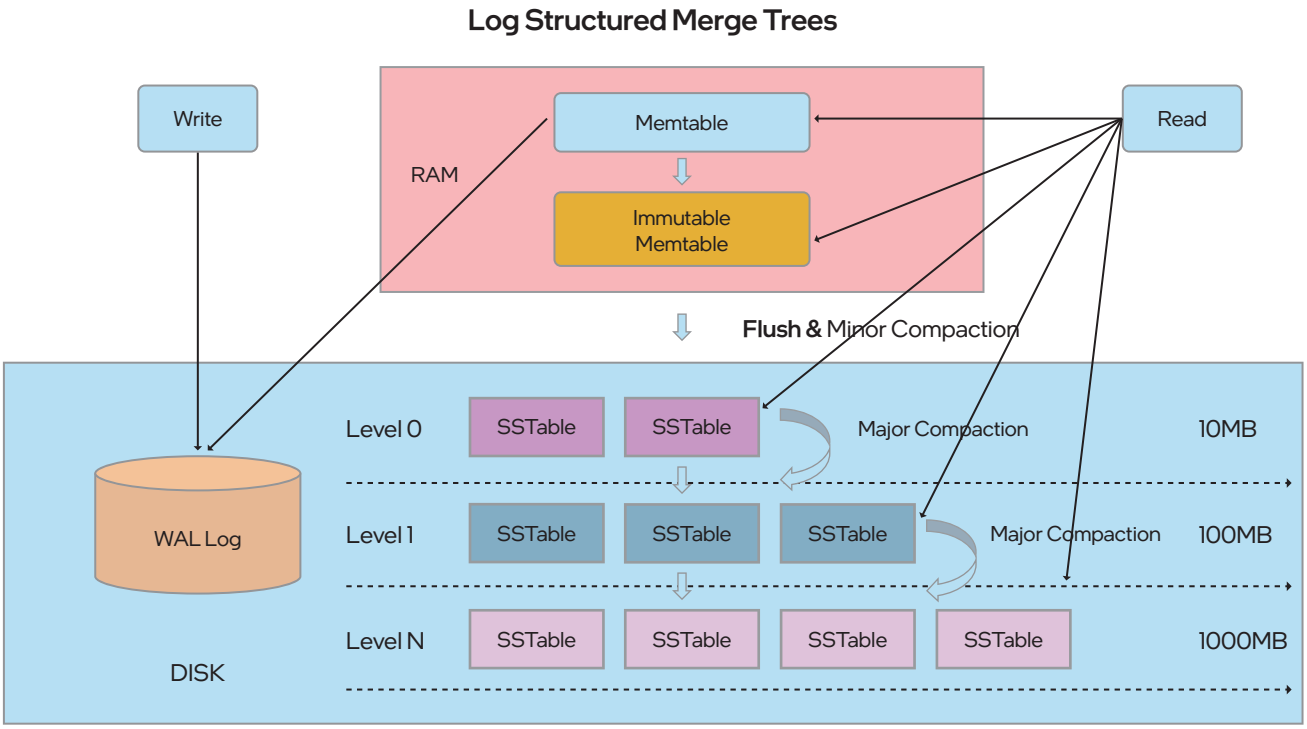

Figure 6: LSM-Tree persistent memory optimization solution

Optimization based on Intel Optane persistent memory

▪ Write optimization

In RocksDB, a write request first writes data to the WAL and then to the Memtable. In case of a system crash, the WAL log can be used to recover the data in the Memtable to ensure data integrity. Under the default configuration, RocksDB guarantees consistency by calling flush on the WAL after each write operation. In a scenario using persistent memory, the mmap function can be used to map the WAL file to the virtual memory space of the application for reading and writing, and the NT-Store instruction can be used to increase the write efficiency, as shown in Figure 7:

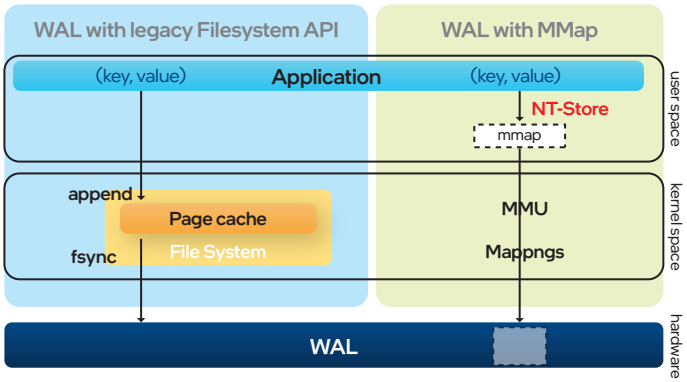

Figure 7: WAL write optimization

▪ Read optimization

RocksDB's persistent cache is designed to improve the read performance on low-latency devices. PMem, as a low-latency device, comes in significantly greater capacities than DRAMs and has compressed data on it. Therefore, storing persistent cache in persistent memory can reduce the DRAM space consumption by block cache, thereby improving the cache hit rate in the whole read process of RocksDB.

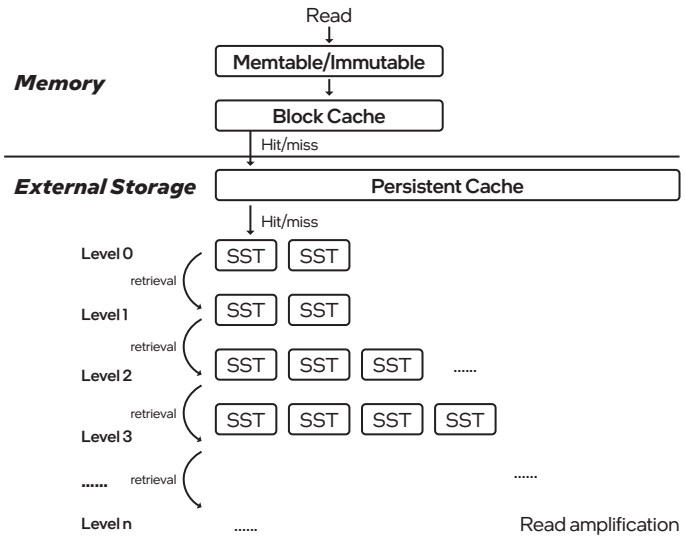

Figure 8: Read optimization

Test

To verify the performance of the high-end local disk i4p, Alibaba Cloud's test team conducted read and write performance tests on instance i3 (without persistent memory) and instance i4p that both run RocksDB.

▪ Test environment

Instance Type	i3	i4p
Physical Cores	26	32
Memory Capacity	384	256
BPS Capacity	0	1T
ESSD/SSD Capacity	1.8T	1.8T

▪ Test scheme

- i3 is the baseline. Both SST and WAL files point to NVMe SSD. DRAM is used as block cache for reading.
- In i4p, the WAL/Persistent cache path points to persistent memory, and the SST data file points to NVMe SSD. WAL utilizes libpmem to optimize writes, and the persistent memory stores persistent cache which comes in a greater capacity than block cache to optimize reads.
- Persistent cache is set to 128 GB and block cache is set to 10 GB.
- For a fair comparison between persistent memory and NVMe SSD, the io-direct option of RocksDB must be enabled to avoid read and write page cache.

▪ Version compilation

Source code: https://github.com/pmem/pmem-rocksdb
make db_bench ROCKSDB_ON_DCPMM=1 DEBUG_LEVEL=0 -j

Test steps: (1) Generate a 100 GB dataset.

(2) Perform write operations using db_bench.

Main parameters:

Dataset:	100GB
Key size:	16B
Value size:	128B
Threads:	96
Writes/Reads:	500000

▪ Write performance test results

On i4p, the write performance achieved by storing WAL in persistent memory and optimizing WAL with mmap was higher than that obtained by storing WAL in the NVMe SSD on i3. When sync was disabled (data might be lost in case of a power outage), the write rate to the NVMe SSD on i3 doubled, but it was still much lower than that on i4p.

Figure 9: i3 vs. i4p in write performance

▪ Read performance test results

As shown in the figure, the PMem of i4p took a relatively long time in loading data into persistent cache. After the cache is filled, the QPS of i4p mainly stayed at about 1,100,000. The loading process of the 10 GB block cache on the NVMe SSD of i3 was very short. After loading, the QPS of i3 stayed at about 300,000.

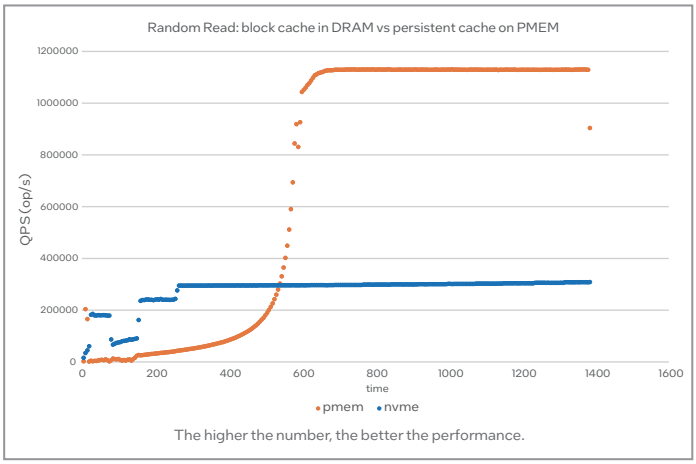

Figure 10: i3 vs. i4p in read performance

To sum up, Intel Optane persistent memory can improve the read and write performance of RocksDB. WAL-based optimization results in a 3.4X higher write performance, and persistent cache-based optimization results in a 3.7X higher read performance.

Results

Alibaba Cloud ECS re7p and i4p instances built on Intel Optane persistent memory have brought huge changes and benefits:

- Improved performance: Compared with similar products, i4p keeps the read/write latency within 1 microsecond, which effectively improves the system's read/write performance and reduces latency.
- Reduced total cost of ownership: for enterprises with insufficient cash flow and a limited budget for IT departments, Alibaba Cloud ECS re7p and i4p instances provide bigger memory capacity and cheaper memory allowing them to save costs while improving efficiency.

Looking Ahead

As cloud computing becomes an important infrastructure that supports the growth of the digital economy and the digital transformation of various industries, user demands for cloud service performance are even higher. This will drive the continuous innovation and development of products.

As a leading cloud service provider in China, Alibaba Cloud will work closely with Intel to update the re7p and i4p instances built on Intel® Optane™ persistent memory. Relying on its technical strengths and huge presence, Alibaba Cloud will keep its persistent memory-optimized instances stay competitive in the market, providing stable, efficient, and cost-effective cloud services for more enterprise users.