White Paper

Intel® Optane™ Persistent Memory and SPDK Power Baidu's Next-generation User-mode Storage Engine

Baidu's user-mode single-node storage engine is an innovative system for an architecture that separates computing from storage, providing stable and efficient services for Baidu's online and offline products. By collaborating with Intel, Baidu greatly improves the performance of a single-node engine through hardware and software collaboration — introducing the Intel® Optane™ persistent memory and Storage Performance Development Kit (SPDK).

- Wang Yanpeng,
Chief Architect,
Infrastructure Division at Baidu

Content

Introduction

New demands for data storage are emerging with the rise of artificial intelligence (AI), cloud computing, and big data applications in information technology. Explosive growth, real-time read/write, and random access to ultra-large data sets have become commonplace challenges in the data storage arena. Developing a data storage system with high performance, stability, security, and scalability is the key factor driving business growth and innovation.

To this end, Baidu collaborated with Intel to launch the next-gen user-mode single-node storage engine built on Intel® Optane™ persistent memory featuring large capacity and low latency. This engine provides efficient, stable, low-latency, low-cost and scalable storage services to Baidu's offline and some online businesses, thereby fully tapping the value in data.

Challenges: bottlenecks in traditional data engine

As Baidu's businesses and products diversify and develop, online and offline products place ever higher demands on performance, reliability, operation and maintenance cost, as well as scalability of storage systems.

Performance improvement bottlenecks

Storage media developed rapidly in recent years, bringing higher performance capabilities to storage unit media. The read and write speed of HDD was less than 100 IOPS , but now NVMe* SSD can reach a read and write speed of 500,000 IOPS. Latency has been reduced from milliseconds to microseconds and system performance bottlenecks are no longer in the storage hardware itself but in the network and processor. Traditional file systems and schedulers do not make full use of the new storage media and therefore hinder further improvements in storage systems.

As storage media develop, the storage of data and metadata in traditional kernel-mode storage engines has created many problems. When a disk is partitioned, disk partitions must be aligned; otherwise disk performance will be significantly degraded. One sector of HDD is usually fixed at 512 bytes, while SSD aligns 4K bytes per sector. There is no guarantee however, that the user data received on file systems are aligned at a fixed size. In order to simplify upper-level software design and make full use of storage space, the storage system has to allow users to update data in arbitrary bytes, but the system reads and writes to the storage hardware with aligned IO. This type of access mode leads to excessive performance waste.

To address this issue, the unaligned data can be temporarily stored in a page cache, and then stored in the storage media once the data volume reaches the optimum alignment size. However, there is a drawback: writing data back from a page cache requires periodic scanning of the kernel; otherwise it must be done by users. Both operations take time and if a power loss happens during this period, data that haven't been written back may be lost.

Most metadata are used as indexes in storage systems. Once the data is written, it's rarely moved within a valid date, but the metadata should be frequently collated and moved, and each movement changes the index structure. Thus, an index needs frequent and timely updates to maintain the efficiency and correctness of data.

In the existing index structures, the rebuilding and reorganization of indexes may cause problems in tasks like valid index recording, invalid index marking, and index space reclamation, so rebuilding an index also involves time overhead. Such consumption of resources will be amplified in complex storage systems thus affecting overall system performance.

Difficulties in business operation and maintenance

The goal of the single-node data engine team is to develop a general-purpose storage system supporting most of Baidu's businesses and products. Baidu's businesses and products are placing increasing demands on storage system operations and maintenance as they shift from an internal distributed architecture to dynamic distributed cloud architecture, from offline storage to online real-time storage, and from a focus on system architecture to system performance. At the same time, three drawbacks in traditional kernel-mode storage systems increase the complexity and operation and maintenance costs.

First, in high business volume situations, kernel-mode storage tends to consume excessive CPU resources. As CPU resources are scheduled preemptively in the kernel, each process can be scheduled only when the CPU is idle. Therefore, as the business volume and process increase, CPU resources will be over-utilized.

Second, a storage system may shut down because of the failure on a single node. In a kernel-mode system, if a kernel error or vulnerability occurs on single-node machine, the host needs to be shut down to find the cause of failure, and fixed by patching or upgrading the kernel. A host failure will seriously affect a business running at an upper layer.

Third, in the daily maintenance when upgrading kernel components, the host should be shut down. All operations running on it are stopped to upgrade the kernel or driver, and the upgrade may introduce new systematic risks. These problems make system maintenance inconvenient, and dramatically increase operation and maintenances costs.

Solution: next-gen user-mode single-node storage engine based on Intel® Optane™ persistent memory and SPDK

Baidu's single-node engine development team has launched a user-mode storage engine based on Intel® Optane™ persistent memory and SPDK. This solution meets the data storage challenge of various business lines and achieves storage system reliability, scalability, and high performance with low operating costs.

This new single-node storage engine supports multiple APIs, including KV, file, block, etc. It is applicable in various scenarios like block, file, and object storage, to meet the storage demands of different business lines. The single-node engine also comes with an innovative structure that separates storage from computation, has an advanced distributed architecture, with flexible data layering, and a user-mode software stack with local/remote storage device compatibility.

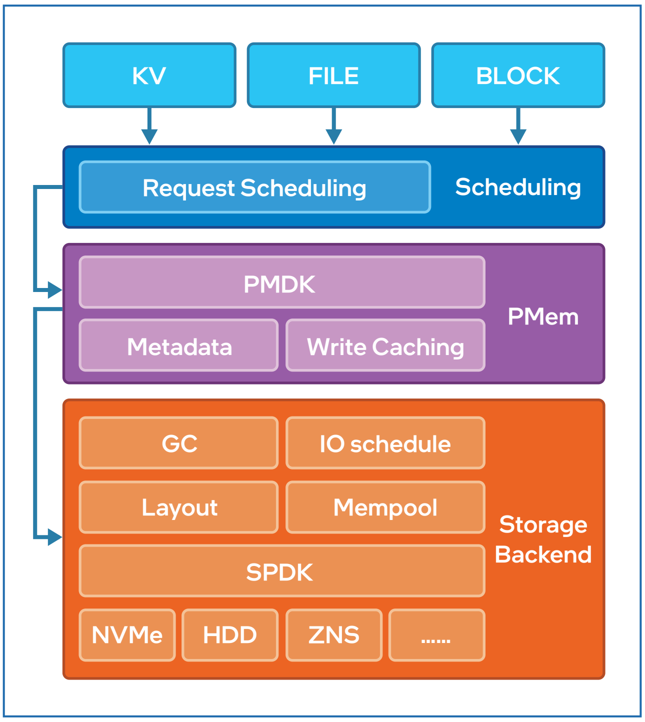


Figure 1. User-mode storage engine architecture

As shown in Fig. 1, the single-node engine consists of three layers: a scheduling layer, PMem cache layer and storage backend layer.

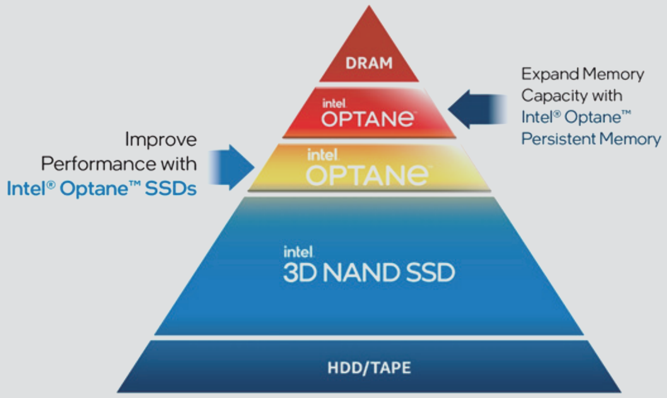

DRAM
intel OPTANE
intel OPTANE
intel 3D NAND SSD
HDD/TAPE
Improve Performance with Intel® Optane™ SSDs
Expand Memory Capacity with Intel® Optane™ Persistent Memory

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

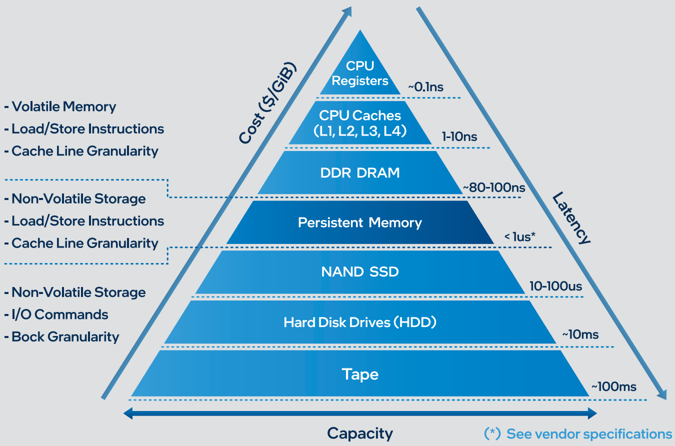


Figure 3. Capacity, price and performance of the storage hierarchy

PMem cache layer

The R&D team selected Intel® Optane™ persistent memory as the storage media for the cache layer.

Intel® Optane™ persistent memory[1] ("persistent memory" or "PMem" for short) is a revolutionary new memory product based on the 3D XPoint™ media. It offers several advantages including high speed, low latency, high cost-effectiveness, large capacity, persistent data protection, and advanced encryption.

Intel® Optane™ persistent memory changes the original storage hierarchy (see Fig. 2) and provides performance similar to DRAM and non-volatile storage like SSD. In addition, persistent memory offers a larger capacity and is cheaper than DRAM. Fig. 3 shows a comparison of capacity, price, and performance in the new storage hierarchy.

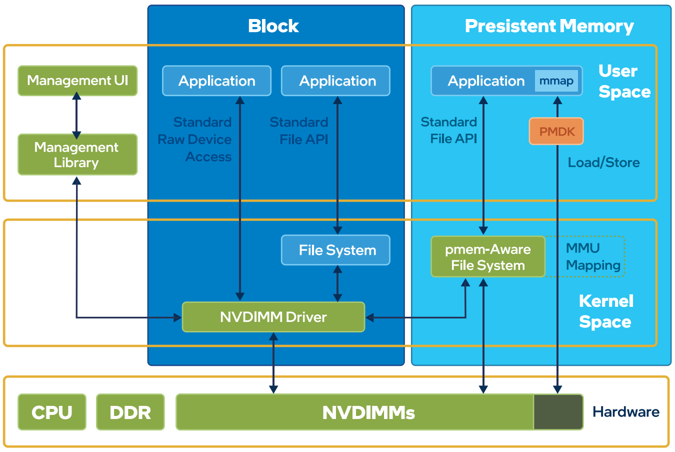


Figure 4. Persistent memory SNIA programming model and PMDK

Persistent memory follows the SNIA programming model (see Fig. 4). Intel offers a set of Persistent Memory Development Kits (PMDKs) that allow applications to directly access persistent memory devices without going through page cache systems, system calls, or drivers related to file systems. This reduces many processes, avoiding the overheads generated by data input/output (I/O), and thus significantly reducing data latency.

Storage backend layer

The new single-node engine uses persistent memory to store metadata, caches, and indexes, combined with a multiple storage SPDK backend (see Fig. 5). This approach offers users various solutions as SPDK[2] provides a set of tools, libraries and programs for writing high-performance and scalable user-mode storage applications.

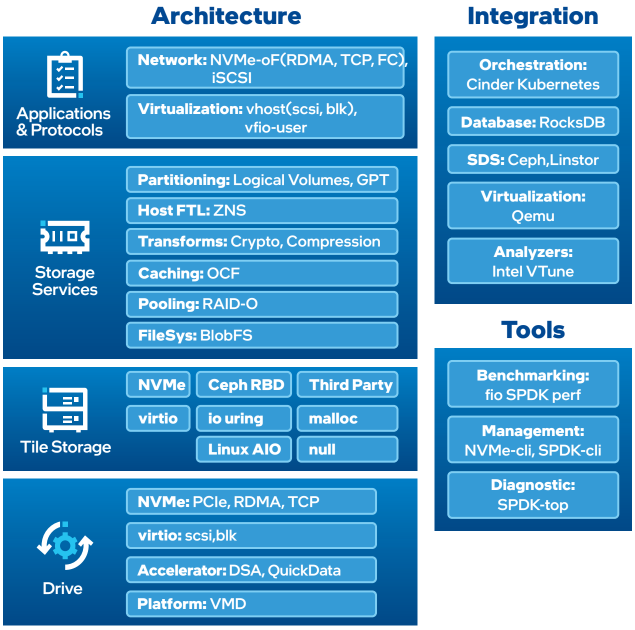


Figure 5. SPDK architecture

[*1] https://www.intel.cn/content/www/cn/zh/architecture-and-technology/optane-dc-persistent-memory.html

[*2] https://spdk.io/cn/

The engine achieves high performance and scalability through a number of key technologies. These include moving device drivers into user space; this avoids system calls and enables zero-copy access from applications. The engine implements a high-performance application framework through lockless, messaging mechanisms and asynchronous programming. This provides a unified user-mode common block device for efficient management of different storage back-end devices.

The user-mode drivers are implemented via polling hardware instead of relying on interrupts after adopting SPDK, helping reduce total latency and latency variance. Furthermore, as compared to kernel drivers, there is a significant performance advantage in terms of IOPS for each CPU core. In addition, SPDK also has a lockless high performance I/O path mode that avoids all locks in key I/O paths, and relying on message passing to share resources amongst multiple threads to achieve higher concurrency performance.

SPDK efficiently integrates Intel® CPU, storage, and networking technologies, fully unleashing the performance potential of high-performance storage media. At the same time, a high-performance framework provides unified device management to support various types of back-end storage devices.

Advantages of the new engine

The next-gen user-mode single-node storage engine based on Intel® Optane™ persistent memory and SPDK can provide solutions for various applications through a variety of configurations (see Fig. 6).

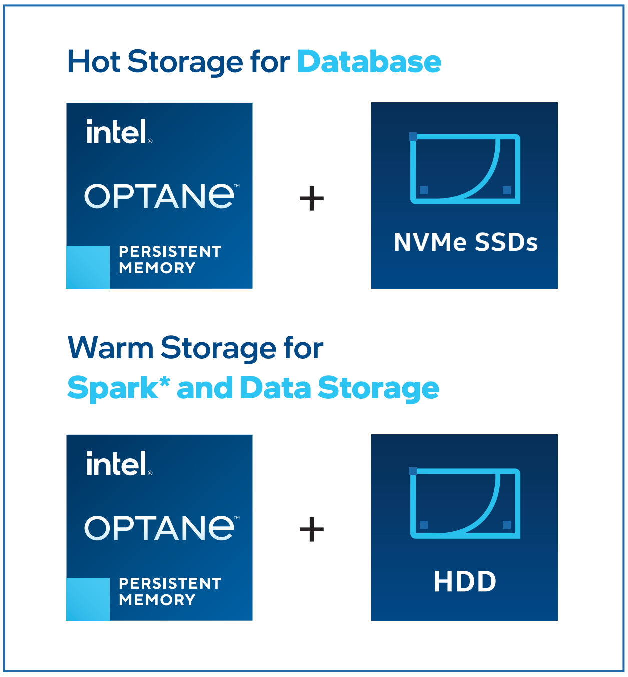

Hot Storage for Database
intel. OPTANE PERSISTENT MEMORY
+
NVMe SSDs
Warm Storage for Spark* and Data Storage
intel. OPTANE PERSISTENT MEMORY
+
HDD

Figure 6. Multiple applications of the single-node engine

Another core use for the new single-node storage engine meets the storage elasticity requirements of many applications. For local disk storage architecture, storage resources are bound to a single node. Since storage should be configured against other resources within the machine based on the application scenario, storage resources are not fully utilized in a local disk storage architecture. Furthermore, a single host node offers limited storage resources, and the size of instances deployed depends on the space constraint, so single instance cannot be expanded dynamically without limit. Essentially, the new single-node engine helps Baidu solve the problems related to the original distributed system, the lack of scalability and insufficient use of local storage resources.

Optimization: improve user-mode storage engine potential with new Intel® software and hardware technologies

Intel engineers worked closely with Baidu team in developing the new single-node engine, matching to the latest hardware products based on the performance features of the single-node engine, and tapping the full potential of hardware through iterative optimization.

As shown in the single-node engine module diagram (see Fig. 7), the Index, Log Manager (for Log Buffer management), NVDIMM Manager (for allocating space for Log Manager and Index), and IO Prepare modules are all located in the cache layer with persistent memory as the storage media.

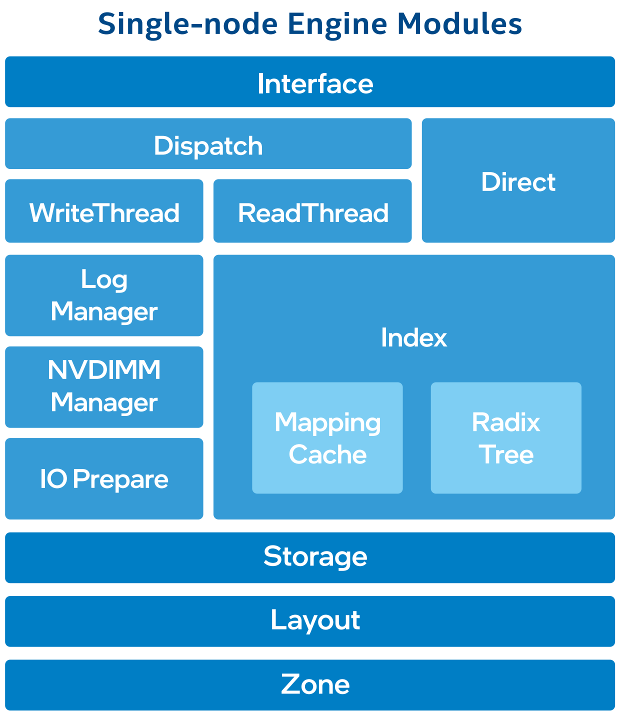


Figure 7. Single-node engine modules

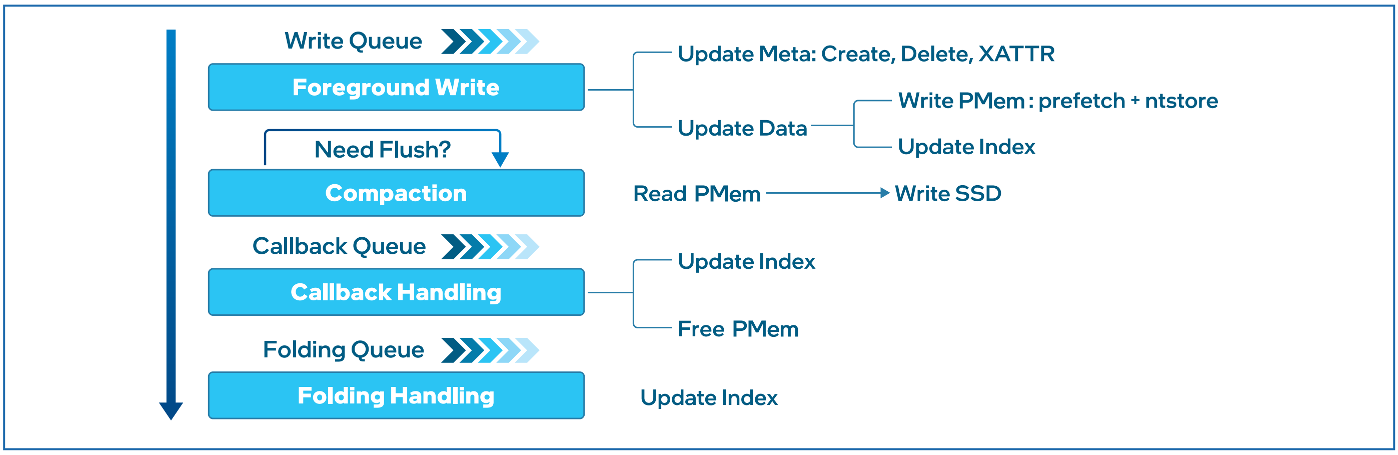


Figure 8. Working thread

As shown in the working threads of the new engine (see Fig. 8), the index is frequently modified and organized when the system reads and writes data. The time overhead resulting from the index rebuild and reorganization will consume massive system resources. With persistent memory in App Direct Mode as the cache layer however, the index data is stored on persistent memory. The memory is scheduled through PMDK and accelerates metadata read/write, to minimize resource loss.

In traditional storage systems, the buffer is written to the page cache of the file system. Hence, the size of page cache will directly affect the write performance of the system. In case of a burst of the write pressures, the data in page cache cannot be written to disk in time due to its limited space, and new data will not be written to buffer. This leads to data latency and performance degradation. However, in the new single-node engine, the buffer is written to persistent memory with large capacity, so the data read/write speed is similar to that of DRAM.

To verify the actual results, the R&D team tested the performance on data written into the buffer. 4K data were written randomly. The measured latency was 4.5us, as shown in table 1; ntstore refers to the time taken to write the data to persistent memory, which is only about 1us, while the rest of the time is consumed by other software tasks. If PMem is replaced with memory, the delay will be about 1/10 of 1us, i.e., 100ns, and the time taken to write 4K data is about 3us, close to the latency of PMem.

PMem delivers similar performance to DRAM at a much lower total cost of ownership (TCO). At the same cost, the capacity of PMem is 3X of DRAM's, so it can cache more data and improve storage system performance. What's more, as the data is cached in PMem, the data can be stored more rationally; this enhances the I/O efficiency of storage back-end devices.

In the Baidu single-node engine solution for warm storage, HDD with lower cost is used as the underlying storage media. In order to increase the engine performance, the NVMe drive is used between the cache layer and storage layer as the cache for the engine (see Fig. 9).

Table 1. Test results

Pattern	avg latency (ns)	max latency (ns)
16K	total latency	total latency
	8976	101170
	ntstore	ntstore
	3916	45808
4K	total latency	total latency
	3957	113089
	ntstore	ntstore
	1034	49686

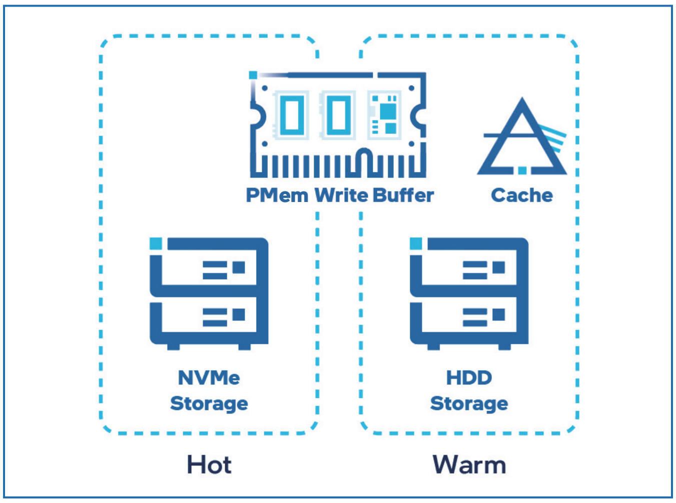


Figure 9. Warm engine accelerated by NVMe

There are drawbacks for NAND flash media in the NVMe SSD. The number of erases and writes in NAND flash is limited, so additional logic should be added to limit the frequency of cache loading during the cache design to prevent frequent cache loading, which may cause significant reduction of the service life of NVMe disks. The read latency of a NAND flash maintains at 100us level, which is not much superior to the DRAM and 3D XPoint media.

Intel® Optane™ SSDs use the same 3D XPoint media as persistent memory with no erase limit, but with the latency at 10+us, so they are good choice as the cache storage media. In addition, SPDK can leverage the high performance of Intel Optane SSDs, so it is consolidated into single-node storage engines to meet higher storage performance needs.

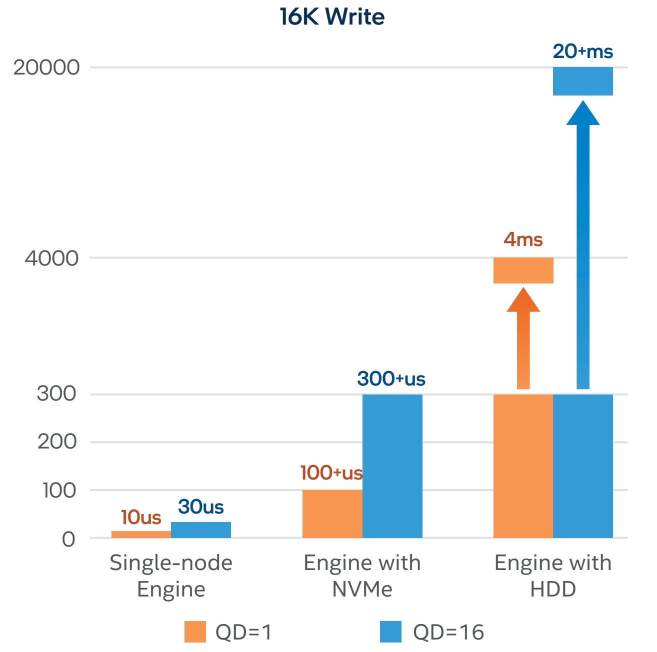
Performance advantages of the new engine

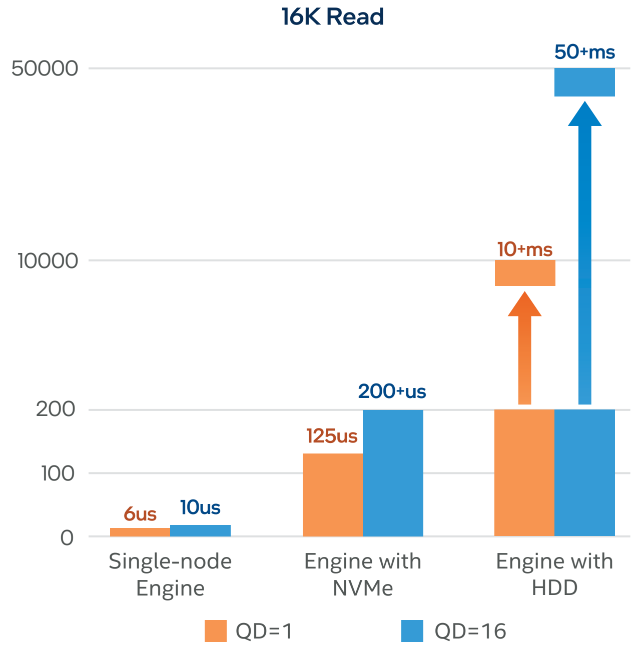
In order to verify the performance of the new single-node engine, Baidu tested two applications:

Hardware environment[3]:

	Hot Engine	Warm Engine
CPU	Intel® Xeon® Gold 6271C CPU @ 2.60GHz	Intel® Xeon® Gold 6271C CPU @ 2.60GHz
Memory	128G	128G
PMem	128G*4	128G*8
NVMe SSD	4TB P4510	4TB P4510
HDD		16T*16

Engine performance:





	QD=1	QD=16
Single-node Engine	10us	30us
Engine with NVMe	100+us	300+us
Engine with HDD	4ms	20+ms

*Lower is better

	QD=1	QD=16
Single-node Engine	6us	10us
Engine with NVMe	125us	200+us
Engine with HDD	10ms	50+ms

*Lower is better

Baidu performed random read and write tests on the new single-node engine with persistent memory, and traditional engines with NVMe and HDD with 16K data (QD is the number of threads). The test results show that the read/write performance of the new single-node engine increased by 10-20X, while fully controlling TCO.

[3] The results were from Baidu's internal test and evaluation in June 2020. For more information about these test results, please contact Baidu.
Testing platform configuration:
Hot engine: processor: Intel® Xeon® Gold 6271C CPU @ 2.60GHz; memory: 128GB DRAM (DDR4-2933, 32GB x 4), 512GB Intel® Optane™ persistent memory (128GB x 4), NVMe SSD: 4TB Intel® SSD DC P4510;
Warm engine: processor: Intel® Xeon® Gold 6271C CPU @ 2.60GHz; memory: 128GB DRAM (DDR4-2933, 32GB x 4), 1024GB Intel® Optane™ persistent memory (128GB x 8), NVMe SSD: 4TB Intel® SSD DC P4510, HDD (16TB x 16)

Envision the future

In the near future, 2nd Gen Intel® Optane™ persistent memory – the Optane PMem 200 series and 3rd Gen Intel® Xeon® Scalable processor will be equipped on Baidu's user-mode single-node storage engine, while system performance will be improved with the new CLWB (Cache Line Write Back) instruction. After the data is flushed into persistent memory, CLWB instruction is capable of keeping the data valid in CPU cashes and the data can be accessed later from caches, thus improving cache hit rate and workload performance. This new single-node engine will offer higher security, high efficiency, and lower TCO, and will provide storage services with higher scalability, reliability and added value for all Baidu products. Baidu will continue to partner with Intel to drive innovation and sustainability of storage technologies, and leading the way to the future of the data storage industry.

The results were from Baidu's internal test and evaluation in June 2020. For more information about these test results, please contact Baidu. Testing platform configuration:

Hot engine: processor: Intel® Xeon® Gold 6271C CPU @ 2.60GHz; memory: 128GB DRAM (DDR4-2933, 32GB x 4), 512GB Intel® Optane™ persistent memory (128GB x 4), NVMe SSD: 4TB Intel® SSD DC P4510;

Warm engine: processor: Intel® Xeon® Gold 6271C CPU @ 2.60GHz; memory: 128GB DRAM (DDR4-2933, 32GB x 4), 1024GB Intel® Optane™ persistent memory (128GB x 8), NVMe SSD: 4TB Intel® SSD DC P4510, HDD (16TB x 16)

Intel does not control or audit third-party data. You should review this content, consult other sources, and confirm whether referenced data are accurate.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex. See the attached sheet for workload and configuration information.

No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.