CASE STUDY

2nd Generation Intel® Xeon® Scalable Processor
Intel® Optane™ DC persistent memory
Search Engine for In-Memory Databases

Baidu Feed Stream Services Restructures Its In-Memory Database with Intel® Optane™ Technology

In today's comprehensive coverage of mobile Internet, everyone hopes to access their favorite news, short videos and promotional information just by clicking or tapping on their mobile phones. Therefore, Feed Stream services which aggregate information and deliver personalized content to users are becoming an increasingly important tool for internet companies to win over customers. As a leading enterprise in global IT and the Internet, Baidu* is committed to becoming a hi-tech company which knows its users best. Its Feed Stream services, which were deployed and developed several years ago, are becoming one of its key business growth engines.

To provide users with more efficient and smoother Feed Stream services, Baidu leveraged its technical advantage in search engines and artificial intelligence to build a highly efficient core in-memory database - Feed-Cube*. Feed-Cube provides data storage and access services with high concurrency, large capacity and high performance. As Baidu's businesses expand further, Feed-Cube also needs to deploy larger memory to support the explosive growth of data. However, the high cost of Dynamic Random-Access Memory (DRAM) brings increasing pressure on the Total Cost of Ownership (TCO) for memory scaling.

In order to lower the TCO while ensuring outstanding performance, Baidu and Intel have started an in-depth collaboration that introduces Intel Optane DC persistent memory and migrates the core working scenario of Feed-Cube to a new memory platform built using it. Baidu's internal test data shows that the Feed-Cube based on Intel Optane DC persistent memory can maintain its performance advantage in a highly concurrent Feed Stream business scenario and significantly reduce costs. This also prompted Baidu to conduct more verification and testing on the feasibility and practical application value of Intel Optane DC persistent memory in more critical application scenarios apart from Feed Stream.

"The Feed Stream services create user profiles and provide personalized content for users as per their preferences. It requires high-performance online storage support. Intel® Optane™ DC persistent memory helps Feed-Cube, the core module of the Feed Stream services, to greatly reduce TCO while ensuring high concurrency, large capacity and high performance."

Tao Wang
Chief Architect, Recommendation Technology Architecture
Baidu

Benefits of the solution realized by Baidu:

- The configuration of Feed-Cube, the core module of the Feed Stream services, changes from DRAM only to a hybrid mode using both DRAM and Intel Optane DC persistent memory, and finally to a configuration using Intel Optane DC persistent memory only. The performance and resource consumption of these changes under the pressure of large concurrent access are in line with Baidu's expectations, and tests show that it can fully support the Feed Stream services with high-performance data access;
- As Feed-Cube gradually migrates from DRAM only to a configuration using Intel Optane DC persistent memory only, its system construction cost also decreases, helping Baidu to reduce TCO.

The Future of Baidu Feed Stream Services

Mobile Internet has become one of the most important approaches to network connection. According to statistics from CNNIC* (China Internet Network Information Center*), the proportion of Internet users, who access the Internet on their mobile phones, had reached 98.6% in China by the end of 2018[1]. With a smart phone, users are more inclined to obtain information through simply "swiping the screen rather than the traditional text method", resulting in an increasing need for Feed Stream services for automatic aggregation and accurate information feed.

Feed Stream is an internet service that aggregates content and continuously presents it to users. It can be implemented by Timeline, PageRank or specific artificial intelligence algorithms. Feed stream services can provide users with more personalized information and help avoid providing irrelevant data. At the same time, advertisers on the platform can achieve better marketing results.

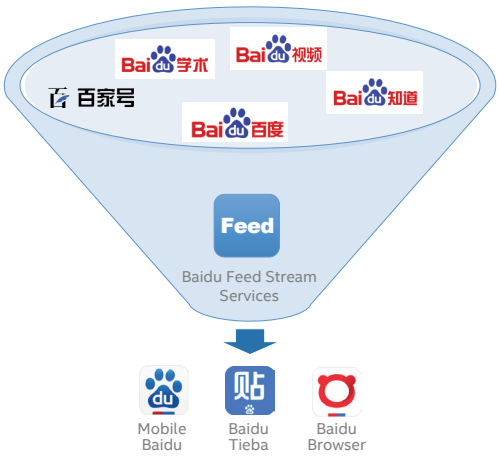


Figure 1. Feed Stream Services in Baidu Core Apps

With hundreds of millions of users, Baidu must consider millions of concurrent services and lower latency for data processing when building its Feed Stream service system. And the key to this is in the building of its data storage and information retrieval capabilities. To optimize this, Baidu uses the advanced core in-memory database Feed-Cube to provide key support for data storage and information retrieval for the Feed Stream services.

Feed-Cube is built based on memory and uses the storage structure of "Key-Value Pair". As shown in Figure 2, keys and the storage offset values of the files where keys and values are located, are stored in hash tables, while the values themselves are stored separately in another data file. Both hash tables and data files are stored in memory. With the high-speed I/O capability of the memory, Feed-Cube provides excellent read-and-write performance and ultra-low latency.

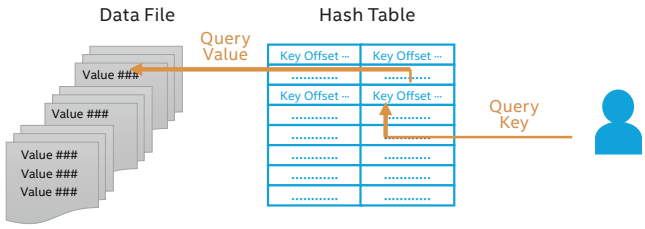


Figure 2 Baidu Feed-Cube Working Diagram

When the front-end application needs to query certain data, it accesses the hash table one or more times by querying the Key value, obtains the offset of the data file storage location where the Value is located and finally accesses the data file to get the desired Value.

Although the Baidu Feed-Cube that is built on DRAM has always had excellent performance in the environment of high/large concurrency (millions of queries per second) and massive data storage (petabyte level), it has to face-up to new and emerging challenges along with the continuous expansion of the Baidu Feed Stream services. The use of expensive DRAM to build a large memory pool results in Baidu's TCO soaring while the limited capacity of DRAM also restricts further enhancement of the processing capability of Feed-Cube streaming.

Enhancement with Intel® Optane™ DC persistent memory

In response to these challenges, Baidu tried higher-performance non-volatile memory (NVM)-based storage devices, such as NVMe SSDs, to store data files and hash tables in Feed-Cube. To verify the system performance with NVMe SSDs, Baidu conducted comparative testing with two Feed-Cube clusters based on DRAM and NVMe SSD respectively.

The test results show that there are three key issues with Feed-Cube using NVMe SSD compared to Feed-Cube using DRAM:

- In a scenario where a large concurrent application was used, the NVMe SSD experienced serious queuing delay and 100% QoS guarantee could not be achieved in a high queue depth (for example, greater than 1,024);
- In a scenario where large-capacity data storage was tested, the marginal effect of the NVMe SSD was poor. The more data that was deployed, the longer the query execution time was, and the disk space utilization rate was lower as well;
- There is still a big gap between the I/O speed of the NVMe SSD and that of DRAM. Therefore, it is still necessary to deploy a large amount of DRAM as a cache in the system to ensure performance.

Intel Optane DC persistent memory provides a new way to resolve these issues. Compared to SSDs, this new product, which revolutionizes memory and storage architecture, has higher read-and-write performance, lower latency and higher endurance and has comprehensive application advantages in a multi-user, high-concurrency and high-capacity environment.

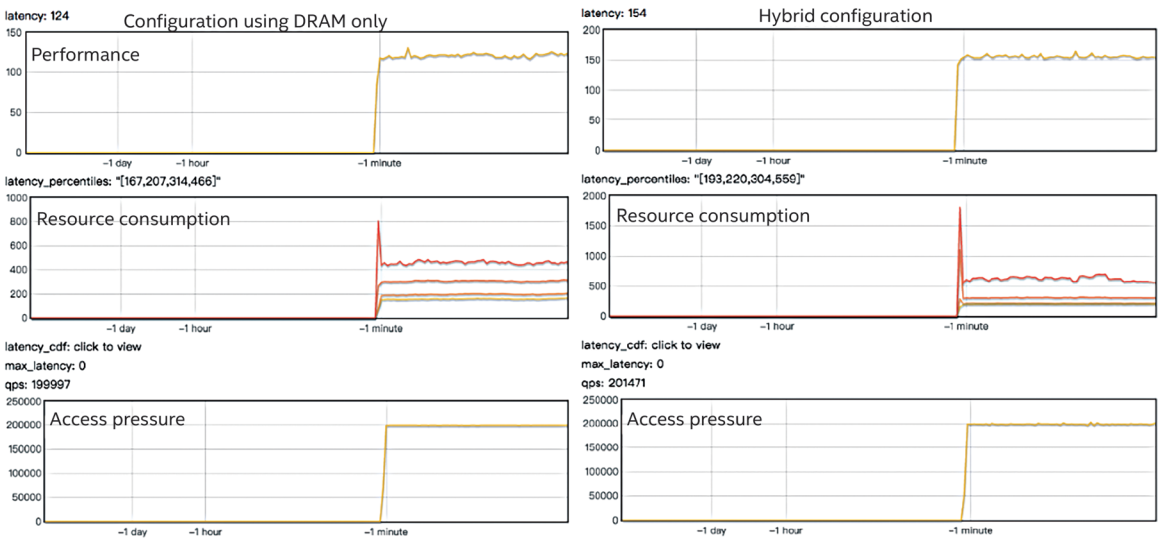
In view of this, Baidu first introduced Intel Optane DC persistent memory to store data files in Feed-Cube, while still using DRAM to store the hash tables. The purpose of this hybrid configuration was to verify the performance of Intel Optane DC persistent memory in Feed-Cube while at the same time minimizing impact on the performance of Feed-Cube. This was achieved by replacing the memory that stores data files first since the number of times that Feed-Cube reads hash tables is much higher than its reading of data files when querying values.

To enable Intel Optane DC persistent memory to be successfully applied to Feed-Cube, Baidu and Intel carried out all-round optimization of the system hardware, the operating system, cores and other components. Both parties first deployed Feed-Cube on a platform built with 2nd Generation Intel® Xeon® Scalable processors which not only offer strong computing power, but also are a 'good match' for Intel Optane DC persistent memory. Secondly, Intel introduced a driver to support Intel Optane DC persistent memory to the BIOS of the server according to the Feed-Cube application requirements and added related patches to the foundation of Baidu's self-developed Linux* kernel 4x to be able to fully unleash the performance potential of the new hardware.

After completing this series of optimizations, Baidu carried out a comparison test between the configuration using DRAM only and the hybrid configuration, simulating the large-scale concurrent access that can happen in a real-life scenario. In the test a setting of 200,000 QPS (Queries Per Second) was used with 100 sets of Key-Value pairs being retrieved per access and therefore, the total access pressure on the system was 20 million. The test results are shown in Figure 3 and Table 1.

Using the hybrid configuration, Feed-Cube has an average access time increase of only about 24% (30 microseconds) under the pressure of 20 million concurrent accesses[3], and the CPU utilization rate increases by 7%[4], which means performance fluctuations are within acceptable limits for Baidu. At the same time, single-server DRAM usage drops by more than half, which will undoubtedly reduce costs in terms of the petabyte-level storage capacity of Feed-Cube.

As shown in Figure 4, the success of the above hybrid configuration prompted Baidu to further try the configuration using Intel Optane DC persistent memory only. There was, however, a particular problem to overcome - the hash tables in DRAM usually use memory allocation commands such as malloc/free, so new commands are required to replace them after introducing Intel Optane DC persistent memory. To address this, Baidu used a self-developed space allocation library based on the libmemkind library* to improve the space utilization while providing space allocation capability.



	Configuration using DRAM only	Hybrid configuration using both DRAM and Intel Optane DC persistent memory
Performance	Average time spent - 124 microseconds, 99th percentile/314 microseconds	Average time spent - 154 microseconds, 99th percentile/304 microseconds
Resource consumption	CPU utilization rate accounts for 40.2% 13GB DRAM-only memory usage	CPU utilization rate accounts for 47.2% 6.3GB DRAM-only memory usage

Figure 3 & Table 1. Test result comparison between Feed-Cube running with DRAM only and running in the hybrid configuration with DRAM + Intel Optane DC persistent memory[2]

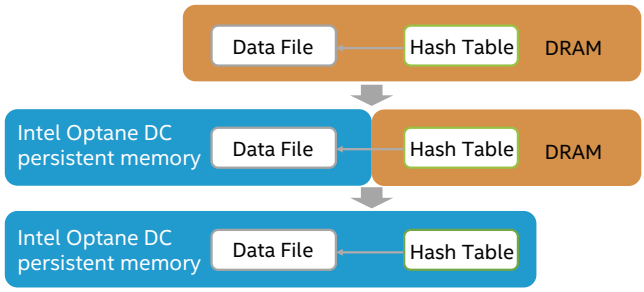


Figure 4 Baidu Feed-Cube Memory Hardware Change Path

After building Feed-Cube using only Intel Optane DC persistent memory, Baidu also tested its performance and resource consumption. As shown in Figure 5, the access pressure of 500,000 queries per second (QPS) was used as an example. The test result shows that the average latency for the configuration with only Intel Optane DC persistent memory is about 9.66% higher than that for the configuration with only DRAM[5]. Performance fluctuations are also within the acceptable range for Baidu.

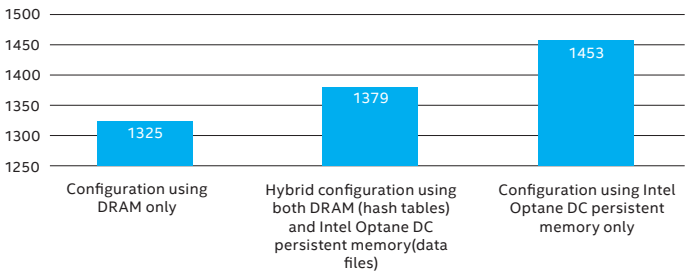


Figure 5. Comparison of processing latency in different configurations

Verifying the Value of Optane™ in more Application Scenarios

While conducting the verification and testing related to the Feed Stream services, Baidu also explored the feasibility and options for using Intel Optane DC persistent memory in many other business scenarios.

For instance, in the case of the fault recovery of a core business module system, when the business module was originally configured with DRAM, it was necessary to reload data from the backend SSD/HDDs for service recovery in case of a power failure and downtime etc. This process could take as long as 10 minutes and severely affect the launch and release of the services. However, with the high-speed read-and-write performance and non-volatility of Intel Optane DC persistent memory, the loading time is now reduced to just seconds[6].

The combination of Intel Optane DC persistent memory featuring high capacity and the 2nd Generation Intel Xeon Scalable processor supporting large memory helps Baidu improve memory density per device and computing efficiency while greatly reducing TCO in Redis* in-memory database and the Spark* distributed in-memory computing engine. Spark deployed in certain key business areas uses the system storage built on DRAM and local/cloud HDDs. When the system is processing data, the data needs to be read from the local/cloud HDDs into the DRAM. However, with the expansion of computing demand, the movement of data between local/cloud HDDs and DRAM has become a bottleneck. To resolve this, Baidu is planning to add Intel Optane DC persistent memory to the Spark system to significantly increasing the system's memory density per device.

In these scenarios, it can clearly be seen that the high-density memory and warm boot features of Intel Optane DC persistent memory are contributing to the continuous innovation and development of "Function as a Service (FaaS)" at Baidu. As an important part of future cloud services, FaaS is based on the "serverless" function framework. Its short lifecycle (even just a few seconds) requires higher booting speed and more memory. Baidu's exploration of FaaS has already yielded benefits in terms of boot speed and TCO with the help of Intel Optane DC persistent memory.

Outlook

In future, Baidu and Intel will conduct more in-depth technical exchange and collaboration around a series of advanced products and technologies including Intel® Optane™ DC persistent memory and 2nd Generation Intel Xeon®Scalable processors. Baidu and Intel will work together to enable these products and technologies to play an increasingly important role in growing core Internet business scenarios as well as critical applications and services. The deeper collaboration between Baidu and Intel will help Baidu provide more diverse and engaging user experience.

[1] Data cited from the 43rd "Statistical Report on Internet Development in China" issued by CNNIC (China Internet Network Information Center).

[2,3,4,5,6] Data cited from Baidu's internal verification and testing based on 2nd Generation Intel® Xeon® Scalable Processor and Intel® Optane™ DC persistent memory. For more details on these tests, please contact Baidu.