Case Study

2nd Gen Intel® Xeon® Scalable Processors
Intel® Optane™ Persistent Memory
Intel® Deep Learning Boost
Database Cloud Service
AI as a Service (AIaaS)

Tencent Cloud: Creating Diverse and Efficient Cloud Services on the Intel® Xeon® Scalable Platform

Tencent Cloud

With more than 10 years of rapid development, cloud services have become the cornerstone for many organizations to work with data and support business development. At the same time, the types of services they cover and the improvement of their service capabilities have also gone broader and deeper. Especially when innovative applications like cloud-based artificial intelligence (AI) and big data analysis are getting mature, a new generation of diverse cloud services, represented by Database as a Service (DBaaS) or Data Analysis as a Service (DaaS) and AI as a Service (AIaaS), has become an enabler for enterprises to implement digital transformation. As a leader in China's cloud service industry, Tencent Cloud takes innovation as its mission and is committed to providing users with these cloud services with higher agility, efficiency, reliability and diversity.

Such highly effective cloud services are inseparable from a strong IT infrastructure. As a long-term innovation partner of Tencent Cloud, the Intel has provided cutting-edge hardware products and technologies for compute, storage and network in the hope that its delivery of more comprehensive data computation, stronger data storage and faster data transmission can help Tencent Cloud further optimize its application and technological frameworks, upgrading, expanding and optimizing its cloud services.

With this partnership, Tencent Cloud for the first time introduced the core products and technologies on Intel's next generation Xeon® Scalable platform, including the Tencent-customized 2nd Gen Intel® Xeon® Scalable processors with Intel® Deep Learning Boost (Intel® DL Boost) and Intel® Optane™ persistent memory. The combination of these technologies not only enables Tencent Cloud's innovative cloud services, such as AI-based intelligent video analysis, to achieve significant

"It has been Tencent Cloud's vision and mission to provide users with a better service experience on its industry-leading hardware architecture. Through in-depth collaboration with Intel, especially with the introduction of the 2nd Gen Intel® Xeon® Scalable processors and Intel® Optane™ persistent memory, Tencent Cloud is able to take the lead continuously in innovations such as AI-based cloud services and database cloud services, providing users with more diversified, differentiated and efficient cloud services."

Ying Liu
Vice President
Tencent Cloud

The benefits from Tencent Cloud's new cloud services:

- The 2nd Gen Intel Xeon Scalable processor supports Intel Optane persistent memory, enabling Tencent Cloud to increase the single-instance memory capacity of cloud database services for Redis by up to 1.34 times under the same SLA and corresponding cost[1].
- Intel® DL Boost built in the Tencent-customized 2nd Gen Intel Xeon Scalable processor brings excellent acceleration capabilities to the deep learning model for Tencent Cloud's intelligent video analysis. With the same hardware and algorithms as well as similar accuracy, it helps improve the efficiency of video analysis by 3.26 times[2].

efficiency improvements, but also allows it to provide users with more affordable cloud database services for Redis.

As one of the major cloud service providers in China, Tencent Cloud is continuously leading innovation with its in-depth expertise, rich industry experience and great infrastructure capabilities, providing agile, efficient, reliable and diverse cloud services for users.

As the saying goes, "you cannot make bricks without straw". Efficient and reliable cloud services are inseparable from strong IT infrastructure capabilities. To help Tencent Cloud further upgrade, improve or revolutionize the performance and experience of its cloud services, Intel has provided it with a new Intel Xeon Scalable platform. The platform is based on the 2nd Gen Intel Xeon Scalable processor. With the Intel DL Boost built in the processor and the processor's strong support for Intel Optane persistent memory, Tencent Cloud's various innovative cloud services, such as TencentDB for Redis based on in-memory database technology and AI-based intelligent video analysis, have achieved better performance and availability.

More Affordable Cloud Database Services for Redis

The cloud database for Redis based on in-memory database technology has been popular with users for its high performance, high flexibility, low latency and rich data structure types. It plays a key role in application scenarios such as transactional cache, session storage, message queues, information publishing, etc. As one of Tencent Cloud's core businesses, the TencentDB for Redis can effectively help users optimize their business processes and improve operational efficiency when it is used with other cloud service capabilities.

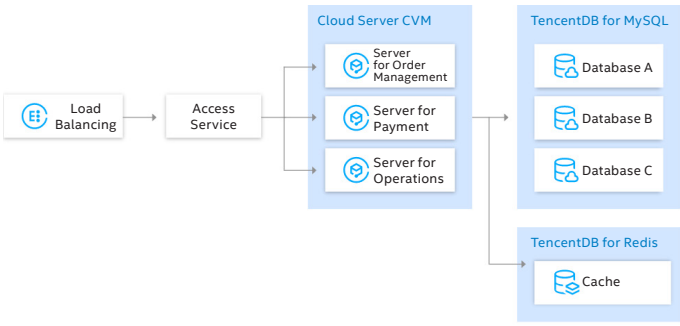

Figure 1. Application of TencentDB for Redis in e-commerce scenarios

Taking the e-commerce industry as an example, the latency-sensitive key data, such as online sellers' product displays and shopping recommendations, can be stored in the cache built by the TencentDB for Redis to achieve faster access, as shown in Figure 1. At present, TencentDB for Redis is already able to deliver 100,000 queries per second (QPS), which is sufficient for high concurrent access in big sales and flash sales[3].

However, the expensive Dynamic Random Access Memory (DRAM) greatly increases the cost of TencentDB for Redis, which in turn limits its application. To help users access more high-availability cloud databases for Redis, Tencent Cloud worked with Intel to introduce the 2nd Gen Intel Xeon Scalable processor with Intel Optane persistent memory, providing TencentDB for Redis with a more affordable memory extension solution.

Intel Optane persistent memory based on Intel® 3D Xpoint™ technology brings users a new memory application experience with larger capacity and data persistence. On one hand, it provides up to 512 GB of memory capacity; on the other hand, data persistence ensures that data remains intact in the event of a power outage, improving data security.

In addition to large capacity and persistent storage, Intel Optane persistent memory provides TencentDB for Redis with higher affordability. The comparative test data from Tencent Cloud shows that under the same service level agreement (SLA), the single-instance memory capacity on the platform with Intel Optane persistent memory can be increased to 1.34 times compared to that on the platform with DRAM-only memory[1]. This allows end users to have better cloud databases for Redis at the same cost.

In this comparative test, Tencent used two platforms, both of which were based on the 2-socket Intel® Xeon® Platinum 8260 processor. This processor has 24 cores/48 threads and is clocked at 2.4 GHz. With similar total cost of ownership (TCO), each CPU socket was only configured with 384 GB DRAM (32 GB*12) for the control group but 96 GB DRAM + 512 GB Intel Optane persistent memory for the experimental group which used a mixed configuration of DRAM and Intel Optane persistent memory at the ratio of 1:5.3. In the test, 88 instances were launched on both platforms.

Model	Experimental Group Mixed configuration of DRAM physical memory and Intel® Optane™ persistent memory	Control Group DRAM-only physical memory
Single-instance memory capacity (GB)	11.27	8.36
Maximum write throughput (TPS/instance)/P99 latency at 1K (ms)	51k/1.86	58k/1.25
Maximum read throughput (TPS/instance)/P99 latency at 1K (ms)	63k/0.82	61.5k/0.71
TCO	0.986	1
Single-instance memory capacity/Total cost of ownership	1.36	1

Table 1. Tencent Cloud's comparative test results on DRAM-only physical memory configuration vs. the configuration of DRAM memory + Intel® Optane™ persistent memory

The test results are shown in Table 1. The two platforms configured with DRAM and Intel Optane persistent memory have similar performance in terms of maximum read and write throughput and P99 latency at 1K. In the case of similar TCO, the platform with a mixed configuration of DRAM and Intel Optane persistent memory is significantly better than the one with only DRAM in terms of single-instance memory capacity. The former's single-instance memory capacity is 1.34 times the latter; and for single-instance memory capacity/TCO, the former is 1.36 times the latter[1].

Higher-performance Intelligent Video Analysis

Traditionally, AI inference is mainly performed based on 32-bit floating point computation, which ensures the accuracy of inference, but brings a huge amount of calculation and deployment complexity. And in application scenarios such as image recognition and image categorization, INT8 and other low-precision fixed-point computations are totally comparable to 32-bit floating-point computation in terms inference accuracy. At the same time, it can also significantly accelerate inference. The Tencent-customized 2nd Gen Intel Xeon Scalable processor not only provides more powerful compute for Tencent Cloud with a better micro-architecture, more cores, and faster and larger memory support, its new integrated Intel DL Boost technology also significantly accelerates inference of INT8-based deep learning models. When used with Intel® MKL-DNN, it greatly improves the inference speed of deep learning models without affecting the accuracy of inference.

The effectiveness of Intel DL Boost derives from its VNNI instruction sets which provide several new wide fused multiply-add (FMA) core instructions for deep learning models and can be used to multiply 8-bit or 16-bit low-precision values. This is particularly important for inference processes that require a large number of matrix multiplications. The introduction of this technology enables the deep learning system to reduce its memory requirements by up to 75% when performing INT8 inference[4]. The reduction in memory and bandwidth requirements can greatly speed up the lower numerical precision operations.

In the AI-based cloud services provided by Tencent Cloud, Intel DL Boost quickly finds its place. For example, Tencent Cloud's newly added intelligent video analysis in its audio and video solutions allow users to categorize, tag and extract highlight images in 25 scenarios including live streaming of online games, beauty & makeup and football matches. This helps users develop a variety of applications and services based on the video content more easily.

Tencent Cloud's intelligent video analysis capabilities are built on the deep learning method (Inception v3 model, partially modified with the Intel® Optimization for Caffe). Therefore, with the introduction of the 2nd Gen Intel Xeon Scalable processor and the use of its built-in Intel DL Boost, the overall efficiency of intelligent video analysis has been improved greatly as Intel DL Boost accelerates inference efficiency based on the INT8 fixed-point computation.

To verify the effectiveness of Intel's technology for intelligent video analysis, Tencent Cloud tested on the 2nd Gen Intel Xeon Scalable processor platform customized for Tencent. The processor has 24 cores/48 threads and is clocked at 2.5GHz. The platform is equipped with 192 GB of memory and uses Intel MKL-DNN with Intel Optimization for Caffe v1.1.3.

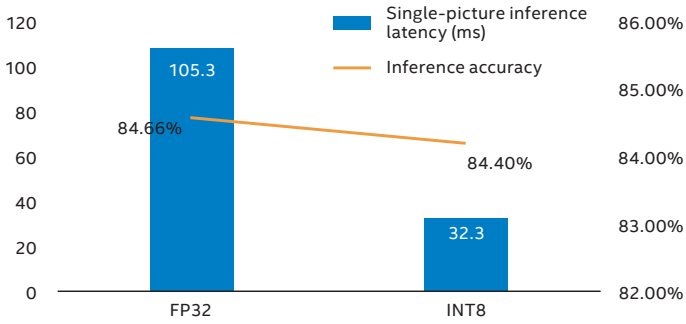

Figure 2. Comparison of inference accuracy and speed of the Deep Learning Model for Tencent Cloud's intelligent video analysis at different computational precision

This test compares inference accuracy and speed of the deep learning model for intelligent video analysis of Tencent's private tag sample data sets (about 40,000 pictures), using 32-bit floating-point computation and INT8 fixed-point computation based on Intel® Deep Learning Boost respectively with the configuration of a single-core processor and Minibatch=1. The test results are shown in Figure 2. The inference accuracy while using INT8 based on Intel DL Boost is only 0.26% lower than that while using 32-bit floating point computation, which means the accuracy of the two is basically the same. At the same time, when INT8 is used for inference, the single-picture inference latency is 32.3ms, which is only 30.7% of the inference latency when 32-bit floating point computation is used – 105.3ms. It can be seen that with the same hardware and algorithm, the introduction of the new solution of Intel DL Boost has improved the efficiency of Tencent Cloud's intelligent video analysis by 3.26 times[2].

Look Ahead

The above tests based on the actual cloud service environment show that the introduction of several core products and/or technologies on the next-generation Intel Xeon Scalable platform, including 2nd Gen Intel Xeon Scalable processors, Intel Deep Learning Boost and Intel Optane persistent memory, has brought Tencent Cloud a more comprehensive, better and more balanced IT infrastructure capability. Based on this, Tencent Cloud will continue to work with Intel to optimize the capabilities of its existing platforms and develop more extensive and in-depth cooperation around high-quality cloud services on high-performance infrastructure in the future.

[1] The data is quoted from Tencent Cloud's Redis cloud service test based on 2nd Gen Intel® Xeon® Scalable processors. Test configuration: 2-socket Intel® Xeon® Platinum 8260 processor, 24 cores/48 threads, HT/Turbo on, BIOS 1.018, 96 GB DRAM and 512 GB Intel® Optane™ persistent memory per socket, single 25 GbE network adapter, Linux kernel 4.14.68-1-tlinux3-nvdimm-0005, Redis 4.10, raw data volume 11.27 GB, totally 88 VM instances vs. 2-socket Intel® Xeon® Platinum 8260 processor, 24 cores/48 threads, HT/Turbo on, BIOS 1.018, 384 GB DRAM physical memory per socket, single 25 GbE network adapter, Linux kernel 4.14.68-1-tlinux3-nvdimm-0005, Redis 4.10, raw data volume 8.36 GB, totally 88 VM instances launched.

[2] The data is quoted from Tencent Cloud's cloud service test for intelligent video analysis based on 2nd Gen Intel® Xeon® Scalable processors. Test configuration: Tencent-customized 2nd Gen Intel® Xeon® Scalable processor @ 2.5GHz, 24 cores/48 threads, HT/Turbo on, 192 GB memory, OS: CentOS 7.6 with kernel 3.10.0-957.el7.x86_64, compiler: GCC4.8.5. The workload in the test was run based on Tencent Cloud's video analysis, using the Intel® MKL-DNN with Intel® Optimization for Caffe V1.1.3. The comparison tests of FP32 and INT8 data were conducted with the configuration of a single-core processor and Minibatch=1.

[3] Data quoted from Tencent Cloud official website: https://cloud.tencent.com/act/pro/redis?fromSource=gwzcw.1345398.1345398.1345398

[4] Data quoted from https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training