CASE STUDY

High Performance Computing (HPC)
2nd Generation Intel® Xeon® Scalable Processor
AI Cloud
Deep Learning

Seeking Smart Cores for AI

iFLYTEK optimizes their Cloud for artificial intelligence based on 2nd Generation Intel® Xeon® Scalable processors

"Intel is not only an outstanding hardware product provider, but also a leading full stack solution provider in the field of artificial intelligence. They are a trustworthy, dependable, all-round partner for our artificial intelligence strategy, helping us effectively explore paths to future innovation in artificial intelligence."

—Mr. Zhijiang Zhang, Vice President of the iFLYTEK's Cloud Computing Research Institute

Executive Summary

As a Chinese and global front-runner in the intelligent speech and artificial intelligence (AI) industry, iFLYTEK Co., Ltd. (iFLYTEK), which was founded in 1999, has established a leading international position in such technologies as speech recognition and natural language processing and has captured an over 70% share of the Chinese speech technology market.[1] Looking ahead to the future, iFLYTEK is currently planning to explore new AI applications in cognitive fields on the basis of its iFLYTEK Cloud. This platform builds upon iFLYTEK AI technology as its core engine, incorporates data and expert knowledge from industries, and will enable a more effective AI industry ecosystem.

Challenge

iFLYTEK is keenly aware of the importance of cloud computing platforms to realize these strategic goals. They also know that the development of AI technologies such as deep learning not only requires the improvement of top-layer applications, but also support in the form of computing, data processing and algorithm optimization from underlying platforms. Achieving this strategy will require iFLYTEK to build a long-term win-win relationship with a partner who has extensive integration capabilities in the aforementioned technologies. Another need is to find a more diversified and optimized hardware platform to pair with iFLYTEK's Cloud for artificial intelligence (hereafter "AI Cloud") and its applications.

Solution

Given these two considerations, iFLYTEK and Intel have entered into an in-depth collaborative relationship. Intel's AI technology portfolio includes multi-purpose, purpose-built and customizable hardware platforms, takes both hardware support and software optimization into account, and covers both end and cloud solutions. Intel has not only provided iFLYTEK with advanced hardware products like Intel® Xeon® Scalable processors as the "core" of iFLYTEK's AI Cloud, but has also been striving to accelerate the migration and optimization of relevant code and models from graphics processing units (GPGPU) to the Intel® Xeon® Scalable processor platform.

The two parties have achieved extremely positive results thus far. For one thing, iFLYTEK has already successfully run its AI Cloud on Intel® Xeon® Scalable processors. In addition, iFLYTEK has migrated large numbers of existing GPGPU-based AI applications to platforms based on Intel® Xeon® Scalable processors and has witnessed excellent performance optimization results. These underscore the first steps both parties have achieved in exploring pathways and directions for advancement in the future development of AI.

Data acquisition, development of algorithms and computing power all impact AI workload performance. Until now, iFLYTEK has met the first two requirements pretty well. However, there are still enormous computing power challenges to overcome.

In order to enhance deep learning performance, iFLYTEK's "Super Brain Project" seeks to emulate human brain neurons in order to give the company's intelligent speech devices rudimentary human thinking capabilities. To achieve this goal iFLYTEK must overcome the immense challenge of processing vast quantities of training data and model parameters. To realize this deep simulation requires an even larger-scale supercomputer cluster, better deep learning algorithms as well as customized and dedicated neural network chip systems.

iFLYTEK is therefore focusing on Intel—another AI industry leader. Mr. Zhijiang Zhang—Vice President of the iFLYTEK's Cloud Computing Research Institute—describes the cooperation between the two parties as follows: "Our partnership with Intel is for the purpose of exploring future developmental pathways and directions in the field of artificial intelligence. We are not merely looking for a hardware partner in the conventional sense."

Strategic Cooperation: Starting with "Finding an Alternative Solution"

Enabling computers to "understand" human speech is the key task of intelligent recognition. After achieving widespread acclaim for its speech recognition achievements in recent years, iFLYTEK has not only established robust speech recognition capabilities with speech recognition systems based on DNN (Deep Neural Networks), RNN (Recurrent Neural Network) and RNN-CTC (RNN with Connectionist Temporal Classification) algorithms, but has also developed several innovative intelligent speech recognition frameworks, such as FSMN (Feedforward Sequential Memory Neural Networks) and DFCNN (Deep Fully Convolutional Neural Network), which are leading the advancement of speech recognition technology.

To further refine speech recognition, massive data training must be conducted on these frameworks, which in turn will bring enormous workloads for computing. The deep learning infrastructure adopted in speech recognition is about linking computing resources to a parallel file system over high-speed networks, on which the compute engine—the core of speech recognition—is developed and used in various types of model training and computing. It's obvious that parallel computing power and high-speed network transmission capacity will directly affect the operating performance of this infrastructure.

Although iFLYTEK previously employed conventional GPGPU solutions, the impressive potential of 2nd Gen Intel Xeon Scalable processors in this type of application has made iFLYTEK's engineers eager to work with it. The engineers plan to test the performance of the processor in deep learning systems.

"The 2nd Generation Intel Xeon Scalable processor with integrated Intel Deep Learning Boost (VNNI) technology, will greatly boost our AI computing," Zhang explained. "It will also improve the TCO of our AI Cloud by allowing us to remove the previous GPU card. In real workload testing in AI cloud with 2nd Gen Intel Xeon Scalable processors, we can get similar or better performance with VNNI enabled compared to the GPU solution.

"We built a hot data cache with Intel® Optane™ SSDs in the AI Cloud to provide fast access for the models during computing, which improve the average response time of AI Cloud significantly", Zhang continued. "And we optimized the AI cloud architecture for the 2nd Gen Intel Xeon Scalable processor and Intel Optane SSD combination so that we could not only achieve the best performance for each of them, but also balance the performance from the system level.

Result

The performance of the 2nd Gen Intel Xeon Scalable processor and Intel Optane SSDs has not disappointed iFLYTEK's experts: In terms of performance, Intel Xeon Scalable processor can integrate up to 28 high-performance cores and the greater vector width obtained with Intel® Advanced Vector Extensions 512 (Intel® AVX-512) technology makes it especially well-suited to high-load parallel computing situations. At the same time, this processor also offers high scalability and reliability under high-performance workloads, making it suitable for the rapid reasoning of complex neural networks in deep learning.

> *"The 2nd Generation Intel® Xeon® Scalable processor with integrated Intel® Deep Learning Boost (VNNI) technology, will greatly boost our AI computing."*
>
> — Mr. Zhijiang Zhang

Summary

2nd Gen Intel Xeon Scalable processors and Intel Optane SSDs together can even better meet the need of the iFLYTEK AI Cloud to accelerate diverse applications. AI Cloud typically runs multiple applications programs and these programs have different hardware requirements. While a deep neural network will need high parallel floating point computing capacity, many other applications do not have these needs and may make frequent use of processing units that are suited to general purpose computing.

Accordingly, to adapt to different applications, the AI Cloud needed to be configured with different hardware. However, growing the diversity in hardware platforms would increase the complexity of purchasing, deployment, operation and maintenance, which will in turn lead to much higher costs. As a consequence, to address complex application needs, the ideal solution is to select an integrated hardware platform that is able to simultaneously perform the acceleration of general tasks and AI applications. 2nd Gen Intel Xeon Scalable processor is the choice because it can adapt to different application loads, help increase the configuration flexibility of AI Cloud, and deliver better scalability while reducing the platform's complexity and cutting costs. In addition, when applications originally running on GPGPUs are migrated to 2nd Gen Intel Xeon Scalable processor and optimized, it is also possible to reduce total cost of ownership (TCO) while ensuring high computing performance.

Migration of Applications: In-depth Cooperation between the Two Parties' Engineering Teams

In the collaboration between the two parties, Intel and iFLYTEK were aware that: The achievement of more in-depth cooperation as well as better application migration and optimization results will require extend the cooperation to the code level. Only in this way, can they attain the highest degree of optimization.

As contact and communication between the two parties increases, both parties' technical experts have been establishing highly effective communication, interchange and cooperation mechanisms in such areas as code migration and optimization as well as hardware optimization.

First, in the area of code migration and optimization, the two parties have embarked on the following tasks:

- **Clarifying roles and responsibilities.** To respect and protect iFLYTEK 's intellectual property rights, iFLYTEK abstractly summarized its deep neural network model and matrix scale etc., and then provided program code and data to Intel; Then, Intel migrated the codes and deeply optimized them based on features of Intel Xeon Scalable processor, such as high-efficiency cores, memory usage and ultra-wide vector width. This in turn ensures that the performance of applications based on Intel Xeon Scalable processor meets iFLYTEK's expectations. Afterwards, the results of optimization, including code and performance parameters, will be sent back to iFLYTEK, allowing iFLYTEK to realize the optimized solution in its actual environment, perform debugging and testing, verify the accuracy and effectiveness of optimization and ultimately help boost the performance of applications in actual production environments. Furthermore, iFLYTEK can apply its optimization methods and experience to the migration and optimization of other AI applications.
- **Exchange Ideas and Work Experiences Regularly.** Both parties' technical experts hold bi-weekly routine telephone conferences. In these conferences, the two parties report on their work progress, engage in interchange and discussion concerning technical issues, resolve any problems that have been encountered, determine optimization directions and plan next moves. In addition, the two parties have also held occasional mutual visits and meetings of high-level personnel and arranged their technical experts to have face-to-face contact to upgrade and maintain the test environment of iFLYTEK.

With regard to hardware optimization, in order to enhance the storage performance of the AI Cloud, iFLYTEK also adopted Intel Optane SSDs, which possess Intel® Optane™ technology offering a combination of high throughput, low latency, high quality of service (QoS) and high endurance. The Intel SSD technology team has helped iFLYTEK to fully optimize the performance of Intel Optane SSD in its AI Cloud and thereby transcend existing storage bottlenecks and tap the full potential of Intel Xeon Scalable processor.

Intel has also provided many reference suggestions concerning the architecture of iFLYTEK's cloud computing platform. In-depth interchanges were arranged between iFLYTEK's technical team and the Intel's cloud computing team and the big data technology team. For iFLYTEK's AI Cloud, Intel has provided all-round consulting services and gave suggestions concerning such aspects as the design of the platform's underlying architecture, optimization of Cloud software and big data software, and operational maintenance and management of the Cloud etc., helping iFLYTEK to successfully launch the AI Cloud.

Through full-scale and in-depth collaboration, Intel and iFLYTEK have achieved very satisfactory outcomes. iFLYTEK was keenly aware of Intel's full stack solution capacity in the areas of AI and cloud computing. In particular, iFLYTEK has migrated its AI applications formerly on GPGPU to Intel Xeon Scalable processor, made full use of the computing capabilities obtained from optimization and simplified the complexity of AI Cloud deployment and maintenance, all while reducing TCO. Throughout this process, Intel has had a deep understanding of the AI leader iFLYTEK's application optimization experience, views and core needs. Many AI-related software tools and libraries of Intel, such as Intel® Parallel Studio and Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN), have been significantly improved due to this collaborative effort, enabling both parties to reap the benefits of a win-win partnership.

Future: Standing Together at the Forefront of Artificial Intelligence

The foregoing series of in-depth collaboration has enabled Intel and iFLYTEK to gain a better understanding of each other's AI capabilities and visions. They were also conscious of the need to change their thinking from conventional hardware cooperation to in-depth collaboration at both commercial and strategic levels. This is an important opportunity for the two companies to embrace the future and ride the AI wave. With this background in mind, iFLYTEK has already begun testing and using more products and technologies provided by Intel. In the iFLYTEK AI Cloud provided to users and developers in various fields and industries, iFLYTEK has already begun adopting the all-new 2nd Gen Intel Xeon Scalable processor, Intel Optane SSD and Intel® FPGA which can be customized and offers flexible support for applications.

Looking ahead, Intel and iFLYTEK will continue to deepen technological collaboration on the basis of the framework specified in the two parties' strategic MOU and make joint efforts in the areas of market research and formulation of market strategies.

1 Source: http://www.mittrchina.com/news/726