Case Study

OpenVINO™ Model Server
2nd Gen Intel® Xeon® Scalable Processor
AIaaS (AI as a Service, AI cloud service)

CDS Global Cloud Uses OpenVINO™ Model Server to Accelerate AI Cloud Service Deployment

As public clouds gradually become major carriers for the delivery of artificial intelligence(AI) services in various industries, the potential of AIaaS (AI as a Service) solutions have gained increasing attention. CDS Global Cloud,a cloud service brand under the well-known services provider Capital Online Data Service Co., Ltd. (Capital Online), is leveraging its own advantages to develop an AI cloud service solution with better performance, scalability, easier deployment and lower Total Cost of Ownership (TCO).

Model deployment and inference is an important part of AI, and its efficiency directly affects the overall performance of solutions. However, the deployment model tools that come with single AI frameworks usually can hardly support different deep learning frameworks for different application scenarios. Furthermore, it is difficult to further optimize inference for different infrastructures. Based on this observation, CDS Global Cloud and Intel have worked together to introduce OpenVINO™ Model Server to its K8S high performance container platform to enable quick deployment of AI models and to improve inference efficiency.

This all-new solution has already been preliminarily deployed and verified in CDS Global Cloud's internal AI applications such as in inappropriate content detection. Results have shown that the OpenVINO™ Model Server-based all-new solution is not only far better than traditional AI model deployment tools in user concurrent access capacity, it also has outstanding performance in key performance metrics such as detection delay.

> "Providing high-performance, scalable, easy-to-deploy, and more cost-effective AI deployment capabilities to users is an important strategy for increasing the competitiveness of AI services on public clouds, and for improving AI application inference performance. Thanks to OpenVINO™ Model Server, we are able to further simplify the deployment process and improve user-friendliness for end users while also significantly improving the production performance of our AI cloud service solution".
>
> – Zhao Ercheng
> **Software Architect**
> **CDS Global Cloud**

Application benefits of using CDS Global Cloud's all-new cloud service solution based on OpenVINO™ Model Server:

- OpenVINO™ Model Server can be optimized for Intel-based infrastructure, which allows for access to even more AI services. Compared to the TensorFlow Serving service framework, the concurrent access capability is improved by 2.4 times[1] when using OpenVINO™ Model Server.
- When integrated with Kubernetes, OpenVINO™ Model Server can provide even better performance support for the AI cloud service solution on CDS Global Cloud's high-performance container platform. Detection delay for all concurrent tasks is less than 30 milliseconds[2] for inappropriate video content detection applications which satisfies real-time detection requirements.
- OpenVINO™ Model Server provides good support for multiple deep learning frameworks, which allows CDS Global Cloud users to overcome the limitations of using one framework. It supports gRPC, REST, and other standard APIs (Application Programming Interface) and further increases the availability of the AI cloud service solution.

Allowing for more advanced AI technologies and models to be introduced into core services is the secret of many companies to implement intelligent transformation, promote innovation in business models, and maintain core competitiveness. Compared to creating the AI application infrastructure and deploying and optimizing the AI models by enterprises themselves, providing direct and fast AI technology and modeling capability support for the service system via public cloud-based AI services not only addresses the problem of different enterprises having different AI technologies and talents with different technical levels, but also allows the enterprises to enjoy higher deployment efficiency, easier-to-use applications, and more flexible expansion, along with other advantages that are inherent to public cloud services. In view of this, an increasing number of public cloud service providers have begun to actively invest into AI cloud services and increased their efforts in R&D into "diversified", "differentiated", and "segmented" services.

Using the real-time detection of inappropriate images and video content in internet and mobile internet applications as an example, the timely identification and/or deletion of inappropriate content is very important. For applications that involve a large amount of video and image content, if inappropriate content is not handled in a timely manner, there may be user complaints and other serious consequences. To prevent these situations, the providers of the corresponding applications mainly relied on manual review to eliminate these hidden dangers. However, since the amount of content has grown exponentially, this method can no longer meet the requirements of high-speed detection. Therefore, companies engaged in online storage, e-commerce, online education, and gaming are all increasingly reliant on AI for the timely, efficient, and accurate detection of inappropriate content in their own products. Furthermore, when these companies select a third-party public cloud service, they increasingly more inclined to choose cloud service providers that have powerful AI cloud service capabilities.

As a well-known cloud service provider in the industry, CDS Global Cloud has built a solid user base in the areas of gaming, video, e-commerce, education, and others over many years, and provided high standard and flexible integrated cloud and internet products and services including compute, internet, and others. At the same time, it has also seized the urgent requirements for AI cloud services from users, and has worked with Intel and other partners to introduce the OpenVINO™ Model Server, which has better overall performance in deployment efficiency, compatibility, and performance optimization, to its high performance Kubernetes container platform. This has resulted in higher efficiency, more convenient, and more complete AI cloud service capabilities available to users.

Creating High-efficiency AI Cloud Services: Completing the Masterpiece

Like other cloud services which provide platform or application-level capabilities, AI cloud services (including all AI application-oriented optimizations and enhancement cloud services) cannot be limited only to providing infrastructure hardware oriented for AI application acceleration. It must also include the deployment of a cloud platform capable of allocating and scheduling infrastructure resources efficiently and be able to provide the middleware or frameworks in a complete operating environment for the AI applications and models. The better matched the cloud platform with the AI operating environment, the more efficient the collaboration, and the more outstanding the performance of the overall cloud service.

CDS Global Cloud adhered to the above principles when creating the AI cloud service solution and, as we can see from the figure, a series of advanced compute, storage and network products were provided by Intel in the infrastructure layer to provide powerful data processing, storage, and transmission capabilities to the solution. On top of this infrastructure layer is the Kubernetes virtualization layer which is responsible for providing node management and scaling services. While making the best use of the capabilities in the infrastructure layer, this solution also has great scalability. Finally, the AI service provided by the solution utilizes containers to provide AI capabilities to the upper layer which can be used in a variety of application scenarios.

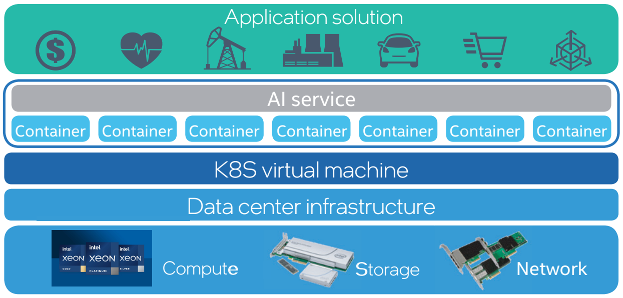

Figure 1. Framework of the CDS Global Cloud AI cloud service solution

Based on this framework, CDS Global Cloud's AI cloud service is able to provide users (who have either chosen to use a cloud platform or a "bare-metal" platform) with additional local services as well as remote AI capabilities via APIs. Using inappropriate content detection as an example for illustrative purposes, the user can choose to build online education, online gaming, or other such application systems, using either CDS Global Cloud's cloud platform or a "bare-metal" platform while also including this AI solution. They can also connect their own system with CDS Global Cloud using the provided APIs to acquire real-time online content detection capabilities.

If the process of building this AI cloud service solution is a work of art, then we can say that the introduction of OpenVINO™ Model Server completes the masterpiece. This software tool is an important component of the OpenVINO™ toolkit. If the purpose of the OpenVINO™ toolkit is to help users implement model optimization and for acceleration in order to achieve the goal of reducing costs and increasing efficiency, then the goal of OpenVINO™ Model Server is to provide computing power by utilizing Intel-based infrastructures to help CDS Global Cloud's new solution to greatly simplify the model deployment process in order to achieve higher efficiency, and to make it easier to deploy AI models to production environments, and to effectively improve inference performance.

The reason why OpenVINO™ Model Server can play this important role is due to its unique framework design and work modes. As shown in Figure 2, it provides gRPC and REST standard network APIs externally so that users can call functions in different scenarios no matter if they are running local or remote AI tasks. These tasks will be connected to the OpenVINO™ Model Server services in the deployed container. After that, the service system scheduler will allocate the task to a certain OpenVINO™ inference engine. On the one hand, the inference engine will use the corresponding plug-in according to the Intel® -based infrastructure component, such as the 2nd Gen Intel® Xeon® scalable processor, Intel® Server GPU, Intel® FPGA (Field Programmable Gate Array), etc.; on the other hand, it will utilize the Model Optimizer (provided by the OpenVINO™ toolkit) to optimize the converted Intermediate Representation (IR) and combine it with the device optimization plug-in to provide high performance inference services within the container and enable user models to run even more efficiently on different Intel® -based hardware infrastructures.

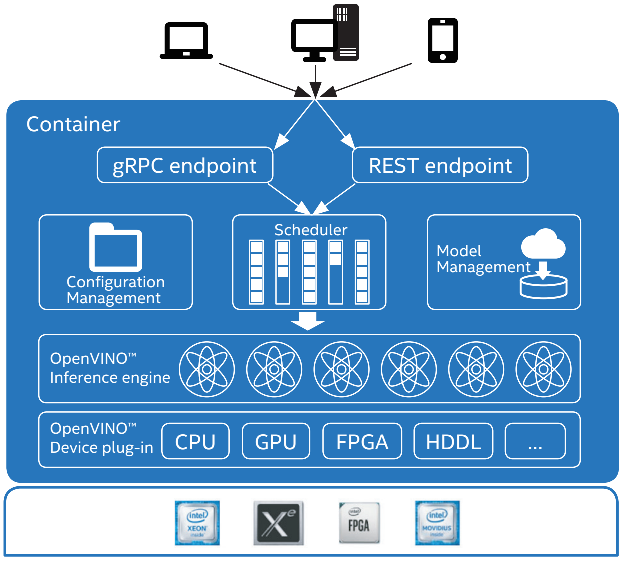

Figure 2. Framework of OpenVINO™ Model Server

The framework design and process shown above enables OpenVINO™ Model Server to have the four following major benefits in AI cloud service deployment and application:

- Can be optimized for different types of hardware infrastructure (Intel® architecture) for performance. As shown in Figure 3, this benefit allows users to deploy models in different scenarios for quick inference after model training has been completed, and effectively reduces deployment and maintenance costs.
- By integrating with Kubernetes, OpenVINO™ Model Server can be used to quickly deploy, easily maintain and expand models using mirroring, while also enabling better horizontal scalability. This may make it easier to provide intensive computing hosted services to users.
- It provides good support for common mainstream deep learning frameworks and can help users overcome framework constraints during the design and deployment of the AI solutions. The best frameworks are introduced based on demands and the AI service capabilities are enhanced.
- The universal network API access allows end users deploy localized AI capabilities on CDS Global Cloud's cloud platform or on "bare-metal" platforms, and users can call these functions remotely via the internet. Support for gRPC, REST, and other APIs also greatly enhances the solution's availability.

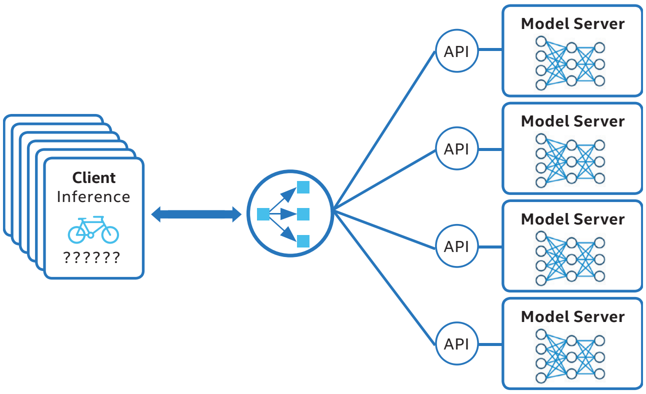

Figure 3. Implementing quick deployment using OpenVINO™ Model Server mirrors

Practical Test: High-Performance and Low-Cost Inappropriate Content Detection

After creating the all-new AI cloud service solution, CDS Global Cloud immediately started testing and verifying the solution in a targeted manner, and used inappropriate content detection, which urgently needed by the company, for testing purposes with its own internal applications. As mentioned previously, as AI technology develops rapidly, deep learning-based image segmentation and image recognition has been widely adopted in various application scenarios, and has proven to perform no worse than manual methods in terms of recognition precision, accuracy, etc. They also have advantages that cannot be matched by manual methods in terms of work efficiency, work persistence, etc. Inappropriate content detection is one of the major applications for this type of technology, and since CDS Global Cloud provides cloud storage services, performing inappropriate content inspection on its stored data is an essential operation.

To meet this internal requirement, CDS Global Cloud has created a deep learning, high-efficiency, complete, and scalable detection service for inappropriate image and video content a long time ago by using its deep technical knowledge along with a variety of technologies. The process for this solution is shown in Figure 4. During the model training stage, the training server will take the uploaded or collected images (that require detection) or the segmented video frames from the user and perform sample annotation, model training, optimization verification and other processes in iterations, to finally obtain a usable AI detection model. During the model inference stage, the inference server will select a suitable detection model for the inference process according to the detection application, and the required results are finally obtained.

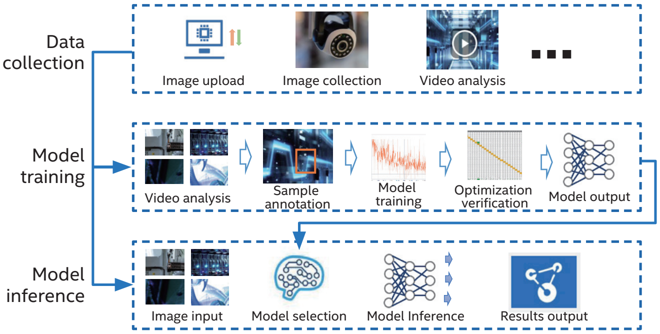

Figure 4. CDS Global Cloud's inappropriate content detection process

The process above shows that the AI model obtained via training can be viewed as the "brain" of the entire detection service which determines whether the service can successfully detect inappropriate content. The inference process, which implements detection using a huge amount of image and video content according to the model, is a key part for the "productivity" of this service. In the past, CDS Global Cloud mainly used the TensorFlow Serving service framework and other such open source tools to deploy deep learning models to production environments to perform inference. However, users usually face the following challenges: the first is the difference in performance and accuracy of models during the training and inference stages no matter which tools are used. If tuning is not done appropriately, there will be no way of maximizing the tool's productivity; the second challenge has to do with changes in the application scenario where users may need to choose different deep learning frameworks such as TensorFlow, PyTorch, Kaldi, etc. TensorFlow Serving and other such model deployment tools are usually only used with only one framework, which restricts the availability and scalability of the solution.

Of particular importance is the fact that there will be significant differences in the inference performance of AI models built in production environments running on different architecture-based infrastructures. And if the user needs to perform repetitive and cumbersome configuration and tuning for each different type of hardware environment, it will inevitably lead to a waste of time and lower work efficiency.

All of these challenges are addressed appropriately on the all-new AI cloud service solution built by CDS Global Cloud. Thanks to the benefits from the deployment and application of OpenVINO™ Model Server, the new solution not only allows users to enjoy highly available, easy-to-maintain, and one-button-deployment AI application capabilities, it also increases inference performance significantly while also helping users to reduce TCO.

To verify the actual performance of the AI cloud service after the introduction of OpenVINO™ Model Server, CDS Global Cloud conducted a real verification test using the inappropriate content detection application on its Kubernetes high performance platform. The design of the solution verification is shown in Table 1:

Test scenario	Real-time inappropriate video content detection (30fps)
Deep learning model	MobileNetV2 (BS = 1)
Test group	Using OpenVINO™ Model Server 21.1
Comparison group	Using TensorFlow Serving 2.3.0
Testing standards	Number of concurrent users

Table 1. Design of OpenVINO™ Model Server verification

The test is based on common real-time inappropriate video content detection and utilizes the MobileNetV2 model (BS=1). The TensorFlow Serving service framework is used as the comparison group for this verification test. As shown in Figure 5, the test results show that using OpenVINO™ Model Server in the solution resulted in a much higher number of supported concurrent users compared to the comparison group and was 2.4 times that of the solution using TensorFlow Serving. Furthermore, the delay time for each concurrent client was controlled to within 30 milliseconds. This means that the detection speed can match the video playback speed for 30 fps videos, and the goal of real-time detection is achieved[3].

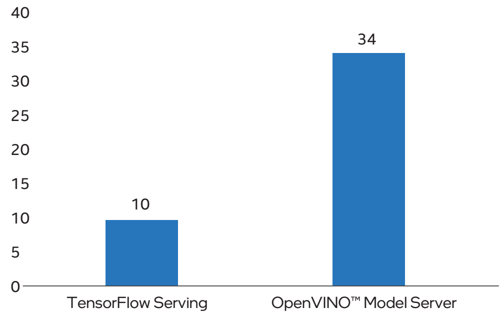

Figure 5. Verification test results for inappropriate video content detection on CDS Global Cloud

In the Future

CDS Global Cloud's all-new AI cloud service solution has been preliminarily tested in internal inappropriate content detection applications and is only at the beginning of its development. The goal is to create a cornerstone for CDS Global Cloud to provide customized and personalized AI cloud services to users. The solution benefits from using OpenVINO™ Model Server will also become a competitive advantage for the company's pivot to expanding into the public cloud market. With the help of this software tool, AI cloud services are easier to deploy, more flexible and scalable, better performing, and the OpenVINO™ Model Server is expected to become the standard amongst users.

In the future, besides further performance optimization of the new solution for even more application scenarios, CDS Global Cloud also plans to work with Intel to integrate the OpenVINO™ Model Server-based AI cloud service with edge computing. This type of solution can further expand AI cloud service applications to the areas of security, automated detection, access control systems, and other applications to provide more diversified and differentiated cloud services to users.

[1, 2, 3] Data comes from internal tests from CDS Global Cloud's own data and models. If you want to learn more, please consult CDS Global Cloud. Testing configuration: test group: processor: Intel® Xeon® Gold 6240Y processor, 18 core/36 threads, Hyperthreading enabled, Turbo Boost disabled; RAM: 6 * 16GB 2666MHz DDR4; Storage: 2 * Intel® S4510 480GB: BIOS: 2.48: Operating system: CentOS Linux Release 7.8.2003 (Core); Kernel version: 3.10.0-1127.19.1.el7.x86_64; OpenVINO™ Model Server version: 21.1; Comparison group: Processor: Intel® Xeon® Gold 6240Y processor, 18 core/ 36 threads, Hyperthreading enabled, Turbo Boost disabled; RAM: 6 * 16GB 2666MHz DDR4; Storage: 2* Intel® S4510 480GB; BIOS: 2.48: Operating system: CentOS Linux Release 7.8.2003 (Core); Kernel version: 3.10.0-1127.19.1.el7.x86_64; TensorFlow Serving version: 2.3.0.

Intel does not control or audit third party data. You should review this content, consult other sources, and confirm whether referenced data are accurate.
Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.
Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.