KT CLOUD SET TO EXPAND AI POTENTIAL WITH AMD INSTINCT™ ACCELERATORS

AMD Instinct MI250 accelerators helping KT Cloud optimize its infrastructure to build a wide range of new AI capabilities

AMD EPYC | INSTINCT

CUSTOMER

kt cloud

INDUSTRY

Cloud Services

CHALLENGES

Deliver cost-effective and optimized cloud-based GPU resources for AI

SOLUTION

Create a new AI platform powered by AMD Instinct™ MI250 accelerators

RESULTS

Achieved a 1.4x performance increase for AI tasks and 70% reduction in GPU cloud service price

AMD TECHNOLOGY AT A GLANCE

AMD Instinct MI250 accelerators
2nd Generation AMD CDNA architecture
AMD Infinity Architecture

TECHNOLOGY PARTNER

MOREH

For over a decade, KT Cloud (formerly Korea Telecom) has delivered secure and reliable cloud-based solutions to businesses. The company's technology enables organizations to leverage powerful cloud computing capabilities without sacrificing control or flexibility. Their commitment to providing top-notch services and solutions has made them one of the most respected names in the cloud services industry.

KT Cloud has ambitious plans to introduce several new offerings including AI Cloud service for public cloud users in the form of Infrastructure-as-a-Service (IaaS). KT also plans to develop Software-as-a-Service (SaaS) capabilities to support its own requirements, such as its automated call center. In addition, KT Cloud intends to provide Application Programming Interfaces (APIs) to AI application companies to support commercial applications such as Chatbots.

"With cost-effective AMD Instinct accelerators and a pay-as-you-go pricing model, KT Cloud expects to be able to reduce the effective price of its GPU cloud service by 70%."

JooSung Kim, VP of KT Cloud

GPUs (Graphics Processing Units) excel in parallel processing, making them highly suitable for AI and deep learning tasks. They can also provide the computational power required to accelerate complex calculations and algorithms involved in AI model training, language processing, and other computationally intensive applications. With this in mind, KT Cloud partnered with AMD and Moreh to create a new AI platform powered by AMD Instinct MI250 accelerators.

Based on the 2nd Generation AMD CDNA architecture, AMD MI250 accelerators are designed for deep learning and machine learning applications, offering exceptional performance, support for a range of precisions for AI tasks, and significant cost-effectiveness. With 128 GB of high bandwidth HBM2e memory with ECC support, AMD Instinct MI250 accelerators offer leading-edge performance. They can be connected to other accelerators and EPYC™ processors through AMD Infinity Architecture, offering up to 800 GB/s of bandwidth.

KT Cloud set two immediate objectives for its new AI platform. One requirement is to optimize the use of its GPU cluster for its AI-based products and services. The second is to create a large language model to serve commercial application requirements in the Korean market.

Transforming cloud computing with KT Cloud's hyperscale AI computing and AMD Instinct MI250 accelerators

Cloud service providers face several challenges when offering GPU resources as a service to customers. They need to charge for GPU resources regardless of whether they are in use, making the service less cost-effective for some customers than on-premises servers. Providers also must invest heavily in GPU resources to meet customer demand. The lack of widespread adoption of hardware-assisted GPU virtualization solutions exacerbates the challenges.

Intent on solving those issues, KT Cloud launched Hyperscale AI Computing, an IaaS-level AI cloud service based on Moreh's MoAI software platform and hundreds of AMD Instinct MI250 accelerators. "With cost-effective AMD Instinct accelerators and a pay-as-you-go pricing model, KT Cloud expects to be able to reduce the effective price of its GPU cloud service by 70%," says JooSung Kim, VP of KT Cloud.

Optimizing AI applications with Moreh's MoAI platform

KT Cloud recognized the importance of seamless migration and accelerated development times for AI developers.

To meet those requirements, KT Cloud embraced AMD Instinct GPUs as a vendor lock-in-free alternative to proprietary NVIDIA CUDA-based options. AMD Instinct GPUs are compatible with industry-standard programming frameworks and libraries, enabling developers to write hardware-agnostic code. Leveraging existing codebase and expertise, AI developers can smoothly transition and deploy on the AMD platform, ensuring a seamless experience.

The Moreh MoAI platform provides a layer of abstraction that eases access to accelerators, increasing the speed and accuracy of training models while reducing manual effort. Single-device abstraction enables users to access AI applications from a single device easily. A just-in-time graph compiler parallelizes and optimizes performance across multiple AI models and domains. The compiler also utilizes a novel intermediate representation called Moreh IR to optimize and execute tensor operations recorded in a computational graph, providing users with a more efficient way to build models compared to conventional uses of PyTorch/TensorFlow. MoAI enables users to use imperative APIs in deep learning frameworks like PyTorch and TensorFlow 2.0 for more intuitive and flexible execution.

KT Cloud Hyperscale AI Computing offers customers a range of virtual accelerator options—from one GPU and 64 GB to 48 GPUs and 24,576 GB—to easily scale the number of GPUs in their virtual machines. The option selected does not affect the virtual machine configuration, so applications do not need to be altered. Instead, the service presents a virtual machine containing a single virtual accelerator, allowing users to build models as if they were using one GPU while KT Cloud handles parallelization and cluster environment setup.

"The results showed that the new service based on the MoAI software and AMD Instinct MI250 accelerators was on average 1.4x faster than A100-based servers."

JooSung Kim, VP of KT Cloud

Putting KT Cloud hyperscale AI computing to the test

KT Cloud and Moreh compared the performance of their new Hyperscale AI Computing service, which uses MoAI and AMD Instinct MI250 accelerators, to KT Cloud's legacy GPU cloud service with NVIDIA A100 GPUs. "We tested each platform using an identical set of approximately 40 open source models," says Mr. Kim. "The results showed that the new service based on the MoAI software and AMD Instinct MI250 accelerators was on average 1.4x faster than A100-based servers."

Tapping machine learning's potential with KT Cloud's 11 billion parameter Korean language model

KT Cloud's project to develop its own large language model for Korean requires substantial computational resources, a significant CAPEX investment that KT Cloud is uniquely positioned to accomplish. KT Cloud is leveraging the power of over 1,000 AMD Instinct GPUs, which offer the immense compute "horsepower" needed to efficiently power KT Cloud's massive Transformer-based encoder-decoder models and train them using billions of parameters. The goal is to offer an API-based service with widespread potential commercial application. For example, KT plans to support an AI chatbot that offers psychological counseling based on a famous Korean counselor.

KT Cloud's first Korean language model used 11 billion training parameters and was evaluated by two different training methods. The first method used the legacy NVIDIA DGX A100 cluster, which has 40 nodes (using 320 GPUs) closely connected for high 1.6 Tb/s bandwidth per node. The second method used the AMD cluster featuring 160 AMD Instinct MI250 accelerators and MoAI platform software, with a more balanced interconnection network featuring two InfiniBand connections per node at 400 Gb/s, along with software that efficiently handles communication overhead in user applications. KT successfully achieved identical training results on both clusters.

"The AMD Instinct accelerator-based system requires only 25% of its counterpart's network switches and cables," explains Mr. Kim. "Each switch installation costs around $20K, so as KT's clusters grow, this benefit will become increasingly noticeable."

"In terms of cost-effectiveness, the AMD Instinct-based cluster using Moreh software exhibited 1.9 times higher throughput per dollar compared to the NVIDIA cluster while improving results by up to 117%," says Mr. Kim.

AI performance leaps forward with KT Cloud's 1200 GPU supercomputer cluster

Encouraged by its results, KT Cloud announced the construction of a new supercomputer cluster—featuring 1200 AMD Instinct MI250 GPUs—for training the next version of its Korean language model. With 200 billion parameters, this cluster will have a theoretical peak of 434.5 PFLOPS for fp16/bf16 matrix operations, 108.6 PFLOPS for fp32/fp64 matrix operations, and 54.4 PFLOPS for fp32/fp64 vector operations, potentially making it one of the top GPU supercomputers in the world.

About KT Cloud

KT Cloud provides cloud computing services to businesses, including infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), and software-as-a-service (SaaS) solutions. Their services include cloud storage, cloud computing, big data analysis, artificial intelligence, and more. KT Cloud is one of the leading cloud service providers in South Korea and has been recognized for its reliability and security. For more information visit cloud.kt.com.

About AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics, and visualization technologies. Billions of people, leading Fortune 500 businesses, and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible. For more information about how AMD is enabling today and inspiring tomorrow, visit the AMD (NASDAQ: AMD) website, blog, LinkedIn, and Twitter pages.

All performance and cost savings claims are provided by KT Cloud and have not been independently verified by AMD. Performance and cost benefits are impacted by a variety of variables. Results herein are specific to KT Cloud and may not be typical. GD-181