WHITE PAPER

KT's Success Stories in AI Cloud Service and Large AI Model Training on AMD Instinct MI250 and Moreh AI Platform

Moreh Inc.
November 11, 2022

Challenges of AI Service Providers

KT, formerly Korea Telecom, is one of the major telecommunications companies and the largest cloud service provider in South Korea. As virtually all industries have been infused with Artificial Intelligence (AI) technology, KT intends to provide various levels of services from IaaS (infrastructure-as-a-service) to SaaS (software-as-a-service) for customers who want to adopt AI. At the same time, KT has been working on building and deploying in-house AI models to increase the competitiveness of its existing products and services, including smart speakers, call center services, and so on. In achieving such goals, KT faces several challenges including:

Traditionally high cost of accelerators in the AI GPU market. Many AI training and inference tasks require large amounts of computation, and GPU clusters are the key infrastructure to process them. To meet the rapidly growing internal and external demands for GPU resources, KT expects hundreds to thousands of new GPU servers will be required each year. However, the price and the supply of AI GPUs have traditionally been very expensive and without room for negotiation. This causes excessive investment costs for computing infrastructure.

Inefficient use of GPU resources in the cloud. The usual GPU cloud services of KT and other providers assign one or more physical GPUs exclusively to each user instance (i.e., virtual machine). In other words, each GPU is visible only to a single virtual machine by PCI passthrough and cannot be shared by multiple users. Hardware-assisted GPU virtualization solutions such as NVIDIA GRID are not widely adopted in commercial cloud services because their overhead outweighs the advantages. This is obviously very different from how the cloud provides other kinds of resources including CPU cores, main memory, network, and storage.

Users must pay for the GPU resources according to the whole lifetime of instances, regardless of their actual utilization. This hurts the economics of GPU cloud services compared to on-premises servers. From the cloud service providers' point of view, too many GPU resources are needed to accommodate a growing number of customers. This further aggravates the aforementioned investment cost problem.

Difficulties in using large GPU clusters efficiently. Among various in-house AI models, KT particularly focuses on developing Korean language models. It plans to train Transformer-based encoder-decoder models with tens to hundreds of billions of parameters using Korean corpora, and the final goal is to build a model larger than GPT-3. Training will take place using more than 1,000 GPUs.

As is well known, parallelizing and training such a large-scale model on a GPU cluster is not trivial. Users need to manually apply various parallelization strategies and optimization techniques to reduce the GPU memory footprint. Whenever the scale or structure of the target model changes, the optimal "mix" of the parallelization and optimization techniques also changes. KT absolutely needed a smarter way for utilizing a large GPU cluster to deliver numerous AI-based products and services in a timely manner.

KT's Cooperation with AMD and Moreh

KT has collaborated with AMD and Moreh to overcome the challenges in public cloud services and in-house AI model development. The three companies have designed a novel AI platform architecture powered by powerful AMD Instinct™ MI250 accelerators and Moreh's software technology. Their key ideas are:

- Adopting AMD Instinct MI250 accelerators to deliver leading-edge performance, cost-effectiveness, and versatility for a variety of AI applications. AMD Instinct accelerators are a compelling option to accommodating various AI workloads in a cloud environment.

- Developing the entire AI software stack from PyTorch and TensorFlow APIs to GPU-accelerated primitive operations to overcome the current limitations of AI cloud services and large AI model training.

AMD Instinct MI250 Accelerator

AMD Instinct MI250 is AMD's powerful HPC and AI accelerator for datacenters. Based on the 2nd Gen AMD CDNA™ architecture, AMD Instinct MI250 delivers leading-edge performance, memory capacity, and cost effectiveness. Matrix cores in an AMD Instinct MI250 accelerator support a full range of precisions including int8, fp16, bf16, and fp32 for accelerating various AI training and deployment tasks. AMD Instinct MI250 provides 128 GB of high bandwidth HBM2e memory with ECC support to help support large AI modes

and datasets. AMD Infinity Architecture enables advanced peer-to-peer connectivity of up to 800 GB/s bandwidth between Instinct MI250 accelerators and even AMD EPYC™ processors with up to eight 3rd Gen AMD Infinity Fabric™ links.

Moreh and MoAI Platform

Founded in 2020, Moreh is a startup aiming to make it easier to build and utilize AI infrastructure at scale. The company believes that many infrastructure-level challenges that KT and other customers face are mainly due to the limitations of the legacy AI software stacks, specifically deep learning frameworks and parallel computing platforms. From this insight, Moreh has been developing the MoAI platform, a set of fully integrated software components from deep learning primitive libraries to application-level APIs. The platform bridges AI applications and underlying accelerators in a more efficient, scalable, and flexible way.

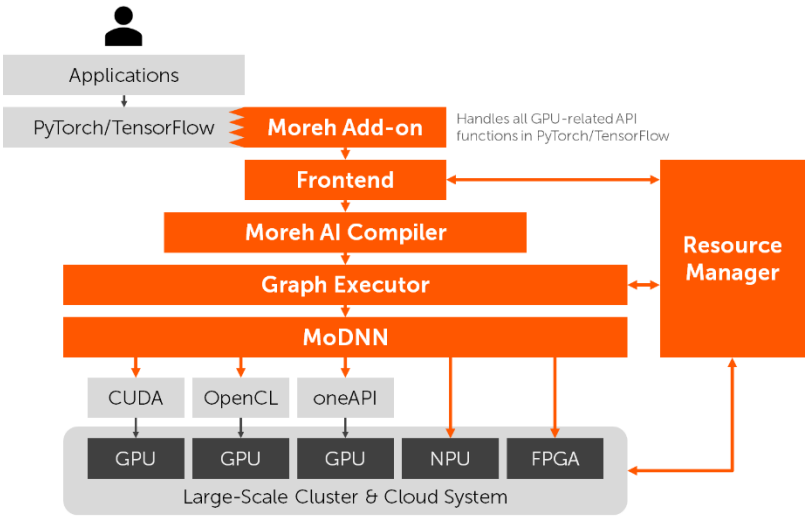

Figure 1. The software stack of the MoAI platform.

Framework-level compatibility. The key design goal of the MoAI platform is to provide 100% PyTorch/TensorFlow compatibility and to ensure the same user experience as traditional AI software stacks. Users do not need to insert or modify even a single line of existing source code for the MoAI platform. They also do not need to change the method of running a PyTorch/TensorFlow program. For example, no additional preprocessing or offline compilation is required. The architectural changes in the MoAI platform are completely hidden from the users.

Accelerator portability. The MoAI platform supports various device backends including the industry standard OpenCL, and its structure is not bound to a specific hardware vendor. Thus, it can run existing AI applications on various GPUs and other types of accelerators (e.g., NPU). AI infrastructure can be built using the most cost-effective hardware without concerning software compatibility.

Runtime IR construction and compilation. Modern deep learning frameworks including PyTorch and TensorFlow 2.0 support imperative APIs and run applications using eager execution by default, i.e., they execute individual operations immediately instead of constructing a static graph and executing it later. Eager execution is much more intuitive and flexible than (static) graph execution, which is especially helpful for researchers. However, it greatly reduces optimization opportunities because the underlying software stack cannot obtain information about the semantics of an application.

To take advantage of both, the MoAI platform gives the illusion of eager execution to users and internally records all tensor operations in a computational graph. After enough operations are collected, the graph is parallelized and optimized by the just-in-time graph compiler and executed at once. This process is not exposed to users, and applications can use the existing imperative-style programming APIs as they are. Moreh IR, a novel graph-level intermediate representation designed for the MoAI platform, is used to represent all stateless and even stateful tensor operations in a directed acyclic graph. It contains the contextual information of tensors and operations for the interaction between a running program and the IR constructor.

Application-level virtualization. The MoAI platform does not expose physical accelerators directly to users by decoupling application processes from the accelerators and their low-level software such as device drivers. Instead, it provides the users with a virtual device that behaves in the same way as a physical GPU. The mapping between virtual and physical devices is solely managed by the MoAI platform. For example, GPU resources are allocated to a user only while the program is doing something on the virtual device. After the program finishes, they are freed and can be used for serving other users. This enables more flexible AI cloud services and drastically improves the average utilization of accelerators.

This technology is called *application-level virtualization* because it is implemented in the platform software without hardware or driver support. It is intended specifically for PyTorch and TensorFlow programs and is not as universal as hardware-level virtualization. Instead, it works with much less overhead and is acceptable for AI purpose.

Single device abstraction. The most important feature of the MoAI platform is to encapsulate a large cluster system as a single device. Users can implement AI applications (e.g., a training code for a large-scale model) on a single device without considering parallelization across multiple accelerators and nodes. Then the just-in-time graph compiler automatically parallelizes the applications, and the runtime system distributes data and computation to the accelerators.

Various parallelization schemes (e.g., data parallelism, inter-layer parallelism, and intra-layer parallelism) and optimization techniques (e.g., operation fusion, activation recomputation, and gradient partitioning) are implemented as the graph-to-graph translation passes of the compiler. The compiler finds the optimal combination of the different techniques based on its own cost model. The compilation process is not specific to a particular AI model and can be applied to a variety of models and domains.

KT Cloud's New AI Cloud Service

KT Cloud has released a new infrastructure-as-a-service (IaaS) level AI cloud service named *Hyperscale AI Computing* based on the MoAI platform. This service was first launched in December 2021 and has become paid for since August 2022. It has been chosen as an official cloud service for the high-performance computing resource supply program of the Korean government (National IT Industry Promotion Agency).

The service provides users with a virtual machine containing a single virtual accelerator named *KT AI Accelerator*. This virtual accelerator is recognized as a **cuda:0** device in PyTorch and a **GPU:0** device in TensorFlow. The underlying infrastructure of the service is mainly equipped with hundreds of AMD Instinct MI250 accelerators. KT Cloud expected that the use of cost-effective accelerators and the pay-as-you-go pricing model can reduce the effective price of its GPU cloud service by 70%.

The users can easily scale up the number of GPUs that each virtual machine can use by selecting one of the available flavors of the virtual accelerator from *small.64gb* to *48xlarge.24576gb* using a simple command line interface. The different flavors only affect the GPU scheduling policy of the MoAI platform and do not affect virtual machine configuration. Applications also do not need to be changed depending on the flavor because parallelization is totally done by the MoAI platform.

The Hyperscale AI Computing service provides a set of reference model implementations covering the representative vision and NLP models and the corresponding sample training datasets. This enables users to immediately begin AI model training and deployment without the burden of software configuration. The reference model implementations are just normal PyTorch programs based on various open-source programs and libraries (e.g., Torchvision and Hugging Face Transformers) and can even be run on other platforms and accelerators.

KT Cloud and Moreh reported the performance comparison between the new Hyperscale AI Computing service and the legacy GPU cloud service using the selective reference model implementations. The results show that the MoAI platform and AMD Instinct MI250 accelerators generally deliver comparable or better performance than NVIDIA A100 GPUs that powered the legacy service.

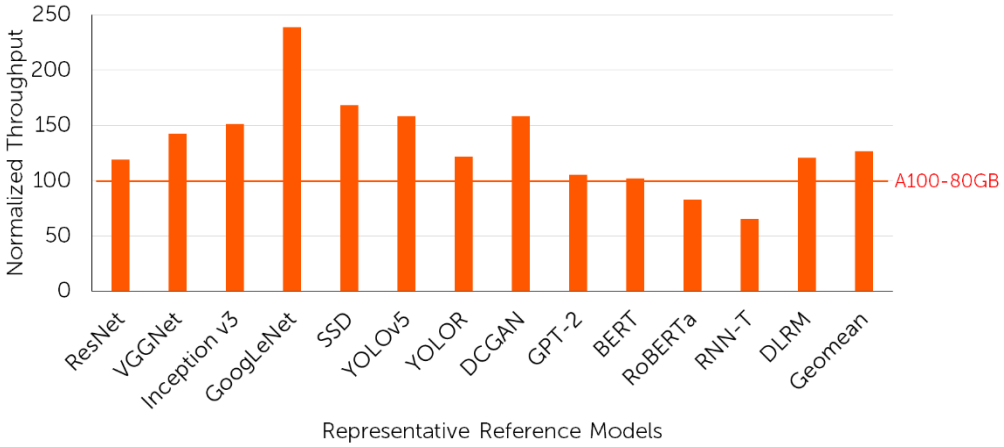

Figure 2. The throughput (trained items per second) of the representative reference models on a virtual accelerator corresponding to a single AMD Instinct MI250 accelerator, normalized to that on a single NVIDIA A100 GPU provided in the legacy GPU cloud service.[1]

11B Language Model Pretraining

KT recently developed its first Korean language model with 11 billion parameters. The training took place on two machines – one was an NVIDIA DGX A100 cluster, and another was an AMD + Moreh cluster equipped with AMD Instinct MI250 accelerators, Supermicro's universal GPU systems (SYS-4124GQ-TNMI), and the MoAI platform software. In addition to building the language model itself, KT has also evaluated two different systems in terms of training methods, results (i.e., perplexity and downstream task scores) and performance (i.e., training time).

For the NVIDIA DGX A100 system, KT used the T5-11B model implementation in the NeMo Megatron library. NeMo Megatron provides manually parallelized implementations for few well-known Transformer models, and the T5-11B model is the only available option for the encoder-decoder architecture. Engineers from KT and NVIDIA cooperated for more than six months to resolve many training issues (some due to the differences of English and Korean) and to manually decide the optimal parallelization strategy for the target system.

For the AMD + Moreh system, on the other hand, the T5 model implementation in the Hugging Face Transformers library was used. KT and partner research groups have a lot of experiences in training small-sized Transformer models using Korean corpora as a preliminary study for large language models, and they used the Hugging Face Transformers for development. What the Moreh engineering team does is replace the model class from smaller T5 to T5-11B in the existing Hugging Face-based training code. Both the original code and the modified one target a single device. Then, the MoAI platform does:

[1] The performance of an NVIDIA A100 GPU was measured on a virtual machine with 8 vCores, 90 GB of main memory, and 4 NVIDIA A100-80GB GPUs, provided by KT Cloud. The performance of a KT AI Accelerator was measured on a HAC virtual machine with 8 vCores and 128 GB of main memory using the *medium.128gb* flavor. Ubuntu 18.04, CUDA 11.2 (for the A100 GPU), PyTorch 1.7.1, Torchvision 0.8.0, and Hugging Face Transformer 4.11.3 were used.

- Applying inter-layer parallelism and intra-layer parallelism, by splitting a single iteration into multiple micro-batches and multiple pipeline stages, and by further splitting some of the operations into smaller ones.

- Distributing input data and computational tasks to different nodes and GPUs.

- Inserting communication operations (e.g., all-reduce, all-gather, send, and recv) between computational operations to guarantee the correct synchronization, and deciding their optimal timing to improve the overlapping of computation and communication.

- Applying operation fusion and activation recomputation techniques.

- Assigning workloads to appropriate nodes and GPUs by considering the interconnection network topology to minimize communication via L2 switches (i.e., to maximize communication within individual L1 clusters).

In the NVIDIA DGX A100 cluster, nodes are tightly coupled with 8-port InfiniBand connections per node (1.6 Tb/s per node). This minimizes the bandwidth gap between intra-node GPU-to-GPU communication and inter-node communication. However, it is not a scalable architecture because it needs impractically many InfiniBand switches as the number of nodes increases to hundreds or thousands. In most cases, moreover, such a tightly coupled interconnection network is overkill. The AMD + Moreh cluster has an interconnection network with adequate bandwidth – two InfiniBand connections per node (400 Gb/s per node). Instead, it applies many software techniques to user applications for minimizing communication overhead.

Finally, KT obtained the equivalent training results for the 11B language model in both an NVIDIA DGX A100 cluster of 40 nodes (320 A100 GPUs in total) and an AMD + Moreh cluster of 40 nodes (160 MI250 GPUs in total). The AMD + Moreh system running T5-11B in Hugging Face Transformers showed 116% throughput (trained tokens per second) compared to NVIDIA DGX A100 running T5-11B in NeMo Megatron, according to the number of GPUs. Considering the construction cost of both systems, the cost effectiveness (throughput per dollar) is 2.05x higher for the AMD + Moreh cluster.

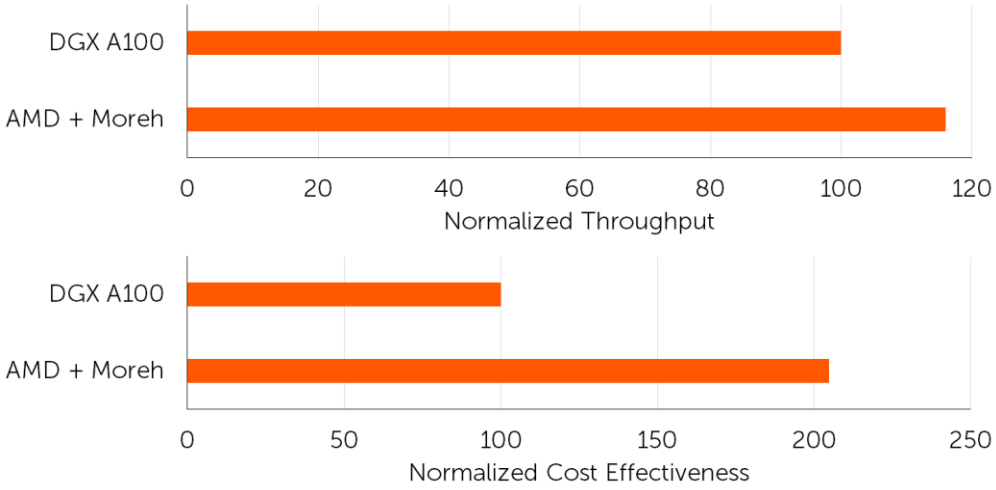

Figure 3. The training throughput normalized according to the number of GPUs and the cost effectiveness of the AMD + Moreh cluster running T5-11B in Hugging Face Transformers and the NVIDIA DGX A100 cluster running T5-11B in NeMo Megatron.[2]

Based on the evaluation results, KT decided to build a larger AMD + Moreh cluster of 300 nodes (1,200 MI250 GPUs in total) for training the next version of the Korean language model with 200 billion parameters. This cluster will be operational in December 2022. It delivers the theoretical peak performance 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. It is one of the top-tier GPU supercomputers in the world, as it is expected to be about 30-40th place on the TOP500 list of June 2023. KT, AMD, and Moreh believe that the new system will contribute to the successful development of the largest-ever Korean language model.

[2] The NVIDIA DGX A100 cluster is comprised of 40 compute nodes. Each node of the NVIDIA DGX A100 cluster contains 2 AMD EPYC 7742 CPUs, 1 TB of main memory, 8 NVIDIA A100 40 GB GPUs, and 8 single-port Mellanox ConnectX-6 VPI. The AMD + Moreh cluster is comprised of 40 compute nodes and a front node. The MoAI platform runs a PyTorch process on the front node while the actual GPU computations are performed on the computing nodes. Each computing node contains 2 AMD EPYC 7413 CPUs, 512 GB of main memory, 4 AMD Instinct MI250 GPUs, and 2 single-port Mellanox ConnectX-6 VPI. The front node is a virtual machine with 32 vCores, 512 GB of main memory, and dual InfiniBand HDR virtual interfaces provided through SR-IOV. Note that the MoAI platform does not enforce the front node to be a virtual machine. Since the NVIDIA DGX A100 cluster has twice as many GPUs as the AMD + Moreh cluster, we adjusted the measured throughput according to the number of GPUs for the first graph.

MOREH

The Enabler of Future AI

To learn more, please visit our website (https://moreh.io) or contact us (contact@moreh.io).

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions, and typographical errors, and Moreh Inc. is under no obligation to update or otherwise correct this information. Moreh Inc. makes no representations or warranties with respect to the accurcy or completeness of the contents of this document and assumes no liability of any kind for the consequences or use of such information or for any infringement of patents. Moreh Inc. reserves the right to make corrections modifications, enhancements, improvements, and other changes to this information, at any time and/or to discountinue any service without notice.

Copyright ©2022 Moreh Inc. All rights reserved.