WHITE PAPER

Training 221B Parameter Korean LLM on 1,200 AMD MI250 GPU Cluster

Moreh Inc.
August 14, 2023

Moreh and MoAI Platform

With the growing interest in LLMs (large language models) and multimodal AI, many people are facing challenges in building and utilizing hyperscale computing infrastructure with thousands or more accelerators. Both excessive reliance on a specific hardware vendor and software-side difficulties in parallelization, performance scalability and portability, orchestration of heterogeneous accelerators, and failover are the major obstacles in the AI industry.

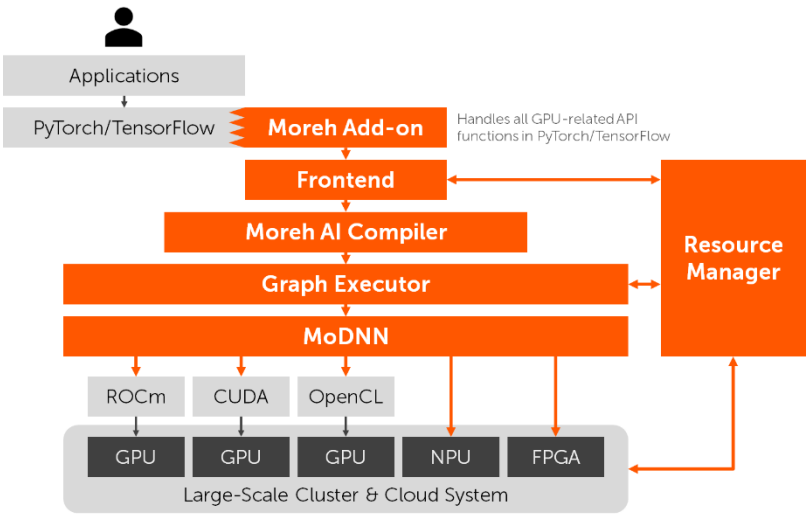

Figure 1. The software stack of the MoAI platform.

Founded in 2020, Moreh has been developing the MoAI platform, a fully integrated software stack from deep learning primitive libraries to application-level APIs. The company believes that the legacy AI software ecosystem (including CUDA and deep learning frameworks) was suitable for developing and utilizing small-scale AI models using a few GPUs but inadequate for handling modern large-scale AI models and cluster systems. Its goal is to provide a new

abstraction layer by innovative software technologies, allowing users to disregard complicated infrastructure-level problems. The key features of the MoAI platform are:

- **100% PyTorch/TensorFlow compatibility:** No code change is required. No additional preprocessing or offline compilation is required. Existing PyTorch/TensorFlow programs just run on the MoAI platform with all the remaining features.

- **AMD GPU support:** The platform is not bound to a specific hardware vendor and supports various device backends. Especially the entire software stack has been largely optimized for AMD GPUs. AI infrastructure can be built in a cost-effective way without concerning software compatibility.

- **Single device abstraction and automatic parallelization:** The platform encapsulates a large GPU cluster system as a single device. Users simply need to write programs targeting a single, very big and powerful device and do not need to care about multi-GPU and multi-node parallelization (i.e., just using a cuda:0 device in PyTorch for utilizing all the GPUs in the cluster). Our just-in-time graph compiler and runtime system do automatic parallelization.

- **Application-level GPU virtualization:** The platform does not expose physical GPUs directly to users and instead provides a virtual device that behaves in the same way as a physical GPU. The mapping between virtual and physical devices is determined by the platform, resulting in drastic improvement in the average utilization of the cloud infrastructure through efficient scheduling and placement.

- **Fault tolerance:** GPU hardware faults do not interrupt user processes, and no action is required for failover. Our runtime system automatically records checkpoints and traces and performs live migration in case of GPU failures.

AMD Instinct MI250 Accelerator

AMD Instinct™ MI250 accelerator is AMD's powerful HPC and AI accelerator for datacenters. Based on the 2nd Gen AMD CDNA™ architecture, AMD Instinct MI250 accelerator 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 accelerator 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.

Table 1 shows the hardware specifications of NVIDIA A100 and AMD Instinct MI250. Note that a single MI250 consists of two independent compute devices (named *GCD*) and all the values listed in the table are the aggregated specifications of the two devices (e.g., the memory capacity of 128 GB indicates that each device has 64 GB of memory). This shows that MI250 can be considered as a competitive alternative to market leading A100.

	NVIDIA A100	AMD MI250	MI250 over A100
Architecture	Ampere	CDNA2	
# logical devices	1	2	
# compute units	108	208	
# processing elements	6,912	13,312	
Peak performance:			
FP32 vector	19.5 TFLOPS	45.3 TFLOPS	2.32x
FP32/TF32 matrix	156.0 TFLOPS	90.5 TFLOPS	0.58x
BF16/FP16 matrix	312.0 TFLOPS	362.1 TFLOPS	1.16x
GPU memory:			
Capacity	80 GB	128 GB	1.60x
Bandwidth	Up to 2,039 GB/s	Up to 3,277 GB/s	1.61x
Cache and scratchpad:			
L1 cache + scratchpad	192 KB per SM 20.74 MB in total	16+64 KB per SM 16.64 MB in total	0.80x
L2 cache	40 MB	16 MB	0.40x
GPU interconnect:			
Technology	NVLink	Infinity Fabric	
Bandwidth	600 GB/s	800 GB/s	1.33x

Table 1. The specifications of NVIDIA A100 and AMD Instinct MI250

KT's AI Infrastructure Based on AMD GPUs

KT, formerly Korea Telecom, is one of the major telecommunications companies and the largest cloud service provider in South Korea. Since 2021, KT has been working with Moreh to design a cost-effective, scalable, and accessible AI infrastructure powered by AMD GPUs and the MoAI platform software.

As the result of the cooperation, Moreh has deployed a total of 1,600 AMD Instinct MI250 accelerators to KT for two separate cluster systems –400 MI250s for a public cloud service named *Hyperscale AI Computing*, and 1,200 MI250s for KT's internal LLM development. The former system has been operational since August 2022 and used to accommodate more than 80 end customers outside KT. The latter system was built in December 2022 and used to provide services specifically for KT's employees. It delivers the theoretical peak performance of 434.5 PFLOPS for BF16/FP16 matrix operations and 54.4 PFLOPS for FP32/FP64 vector operations. This is obviously one of the top-tier GPU cluster systems in the world, corresponding to the 15th ranking on the TOP500 list of June 2023.

LLM Training Performance

Upon the request of KT and KT Cloud's end customer, Moreh evaluated LLM training performance on AMD MI250s and the MoAI platform software.

	T5-11B	GPT-13B	T5 variant 221B Korean LLM
Customer	KT	Korean LLM startup	KT
AMD + Moreh:			
GPUs	160 MI250s	128 MI250s	1,200 MI250s
Measured performance	17.802 PFLOPS	16.668 PFLOPS	154.451 PFLOPS
Per-GPU performance	111.26 TFLOPS	130.22 TFLOPS	128.71 TFLOPS
Competitor:			
GPUs	312 A100s	256 A100s	N/A
Measured performance	29.920 PFLOPS	31.309 PFLOPS	
Per-GPU performance	95.90TFLOPS	122.30 TFLOPS	
Comparison:			
Performance	1.16x	1.06x	N/A

Table 2. A quick summary of the performance numbers below

T5-11B model training. As a preliminary work for developing Korean LLMs with hundreds of billion parameters, KT attempted to train a T5-11B model using its own Korean corpus in the last year, on two different systems – one was an NVIDIA DGX A100 cluster of 39 nodes (312 A100s in total), and another was an AMD + Moreh cluster of 40 nodes (160 MI250s in total).

Engineers from KT and NVIDIA cooperated for more than six months to port and optimize the T5-11B model implementation in NeMo Megatron for the target system (e.g., manually deciding the optimal parallelization parameters). For the AMD + Moreh system, on the other hand, we simply replaced the model class from smaller T5 to T5-11B in the existing Hugging Face Transformers-based training script targeting a single device. The MoAI platform then applied automatic parallelization and optimization (e.g., operation fusion, activation recomputation, and network topology aware placement).

Finally, KT obtained the measured performance of 29.920 PFLOPS (95.90 TFLOPS per GPU) on the NVIDIA DGX A100 cluster and 17.802 PFLOPS (117.26 TFLOPS per GPU) on the AMD + Moreh cluster. AMD MI250 powered by the MoAI platform showed 116% per-GPU throughput (trained tokens per second) compared to NVIDIA A100. The AMD + Moreh cluster turned out to be much more economical in terms of cost effectiveness (throughput per dollar).

GPT-13B model training. One of the leading LLM startups in Korea has used KT's public cloud service to meet the growing resource demands for training various language models, and especially, it attempted to train a GPT-13B model in the cloud and compare the result with the performance number previously obtained on 256 NVIDIA A100s – 31.309 PFLOPS in total (122.30 TFLOPS per GPU).

Similar to KT's T5-11B model training, we could implement a training script based on the GPT model in Hugging Face Transformers targeting a single device. As a result, we obtained the measured performance of 16.668 PFOPS on 128 AMD MI250s with the MoAI platform software. The per-GPU performance is 130.22 TFLOPS which is 1.06x higher than that of NVIDIA A100.

Scaling to 221B Parameters and 1,200 GPUs

From March to June 2023, KT and Moreh trained a largest-ever Korean LLM with 221B parameters on top of the MoAI platform and the 1,200 AMD MI250 cluster system. The cluster consists of 300 compute nodes (4 MI250s each) and 38 InfiniBand HDR 40-port switches (providing two IB HDR connections per node). Note that the scale of the InfiniBand network is much smaller than that of a typical NVIDIA DGX cluster. Communication optimization techniques in the MoAI platform allow compute nodes to be (relatively) loosely connected.

The Korean LLM is a variant of T5 adopting some novel techniques from recent research findings. Our ML engineers could easily implement such a model and a training script in two weeks (including initial testing) without considering parallelization and scalability. After that, we could train the model with a computing performance of up to 155.451 PFLOPS (128.71 TFLOPS per GPU). This is comparable to the per-GPU performance obtained on a smaller number of GPUs (160 GPUs) and for a smaller model (T5-11B), as shown in Table 2. This implies that the MoAI platform achieves a good scalability.

MOREH

The Enabler of Future AI

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