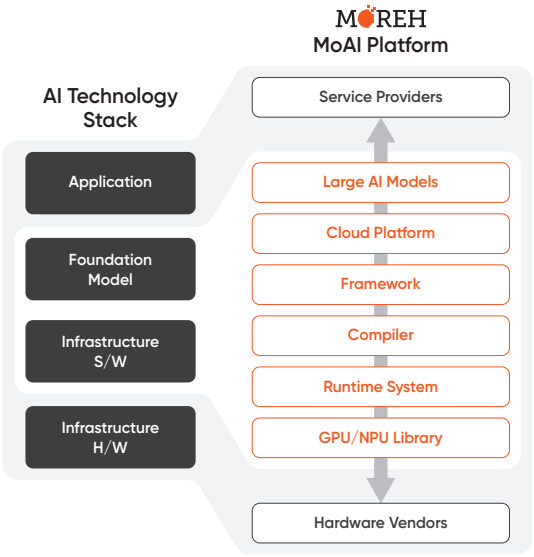
MoAI Platform

for AI Developers

MOREH

moreh.io

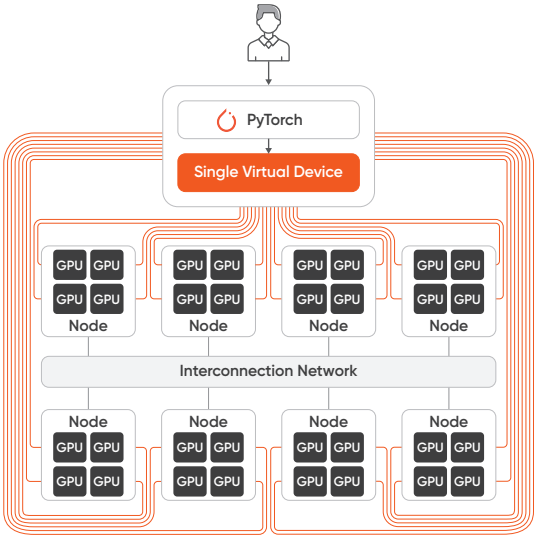
A full-stack infrastructure software from PyTorch to GPUs for the LLM era



The MoAI platform transforms the way of executing AI applications while preserving the semantics of standard deep learning frameworks including PyTorch and TensorFlow. It is powered by Moreh's proprietary software components including an on-the-fly IR constructor, a graph-level compiler, and a distributed runtime system.

- 100% PyTorch/TensorFlow compatibility:
  no code change, no additional preprocessing or offline compilation, just run existing programs on the MoAI platform
- AMD GPU support
- Triton Inference Server support with LLM optimizations
- GPU virtualization for higher utilization
- Automatic failover for GPU failures
- Tested with 130+ PyTorch models
- Served 100+ cloud customers

Programmable AI infrastructure at scale



Users can treat thousands of accelerators as a single (very large and powerful) virtual device in PyTorch and TensorFlow. Various large AI models and algorithms can be easily implemented without concerning about complex system architectures and parallelization techniques.

- Train and deploy any multi-billion or multi-trillion parameter model
- Programming anything you need for pretraining, fine tuning, compression, and serving
- Scale to thousands of GPU/NPUs by automatic parallelization and optimization

Ready for your LLM development and service

The MoAI platform has been successfully adopted for numerous LLM development projects dealing with up to 221B parameter models and 1,200 GPUs.

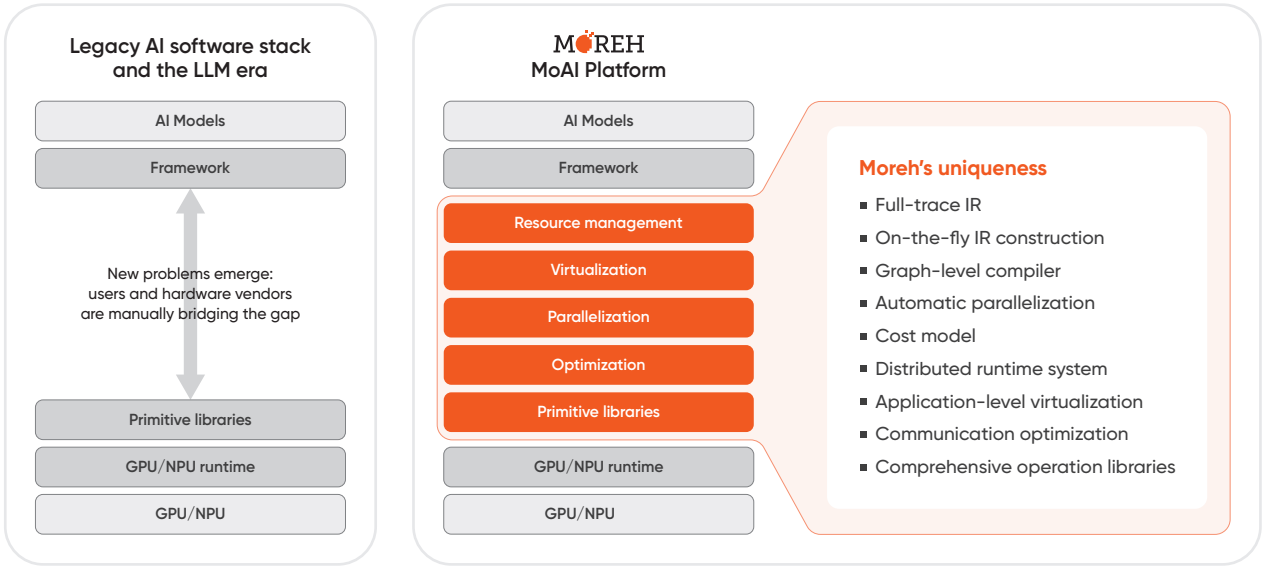
Read the success story

Moreh can deliver

- MoAI Platform, the most comprehensive infrastructure software for LLMs
- Cost-effective AMD GPU cluster
- Kubernetes-based cloud environment
- A rich set of LLM and multimodal model implementations in PyTorch running on the MoAI platform
- Technical support and co-development for LLM training

New era, new software

Handling hyperscale infrastructure is the critical enabler for developing large AI models. A very new set of software and technology is now required.

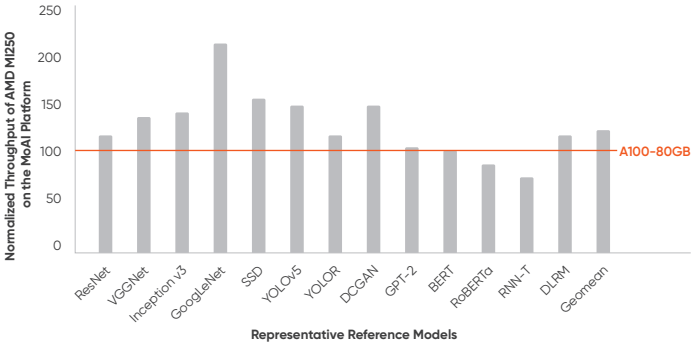


Success Stories

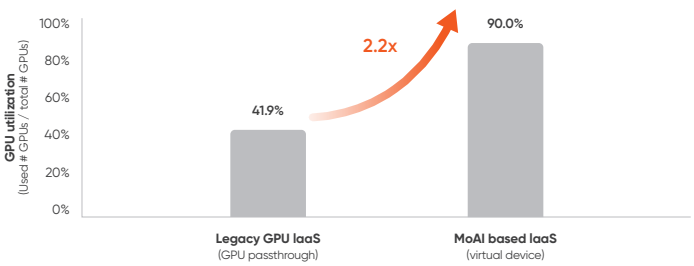
KT Cloud's AI IaaS Cloud Service

KT Cloud, the largest cloud service provider in Korea, has released an AI IaaS cloud service called HAC (Hyperscale AI Computing) powered by the MoAI platform and AMD GPUs since December 2021. HAC has accommodated more than 100 customers. It has been chosen as an official cloud service provider for the AI resource supply program of the Korean government (NIPA).

KT Cloud and Moreh reported the performance comparison between HAC and the legacy GPU system using the selective reference model implementations. The results show that the MoAI platform and AMD MI250 GPUs deliver comparable or better performance.



The GPU cluster has been consistently operating at high utilization with the help of GPU virtualization and dynamic allocation of the MoAI platform.



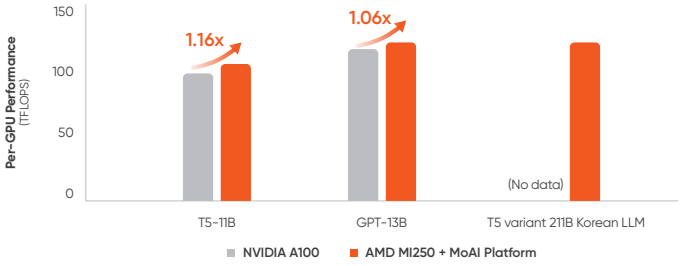
Graph 1: The throughput (trained items per second) on a HAC virtual accelerator corresponding to a single AMD MI250 GPU, normalized to that on a single NVIDIA A100-80GB GPU. The source is our whitepaper: https://moreh.io/publications/kts-success-stories-in-ai-cloud-service-and-large-ai-model-training-on-amd-instinct-mi250-and-moreh-ai-platform-221111.pdf

Graph 2: The measured average GPU utilization of the HAC service infrastructure with 400 AMD MI250 GPUs, and the simulated utilization of a legacy IaaS service using OpenStack's GPU passthrough for the same users and usage patterns. The numbers are as of December 2022.

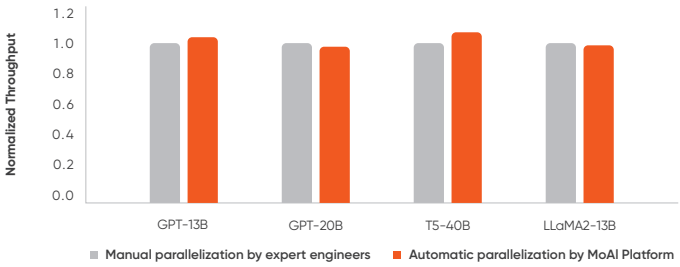
Korean LLM Development

Moreh has been working with various customers to enable and accelerate their LLM projects. The most remarkable one is the development of a largest-ever Korean language model with 221B parameters on a 1,200 AMD MI250 cluster system, done from March to June 2023. The MoAI platform enabled fast model implementation and high scalability.

The performance numbers measured from different LLM projects show that AMD + MoAI clusters are much more economical in terms of cost effectiveness (throughput per dollar).



Users do not need to care about multi-GPU and multi-node parallelization. The MoAI platform applies automatic parallelization and optimization, achieving comparable or better performance than that obtained by expert engineers. This can drastically reduce the "time-to-start-training" for large AI models, from 3-6 months to 1-2 weeks.



Graph 1: The per-GPU performance numbers of AMD MI250 + MoAI Platform measured and calculated by Moreh, and that of NVIDIA A100 provided by our customers. The source is our whitepaper: https://moreh.io/publications/training-221b-parameter-korean-llm-on-1200-amd-mi250-gpu-cluster-230814.pdf

Graph 2: The throughput (trained items per second) obtained by the MoAI platform and its automatic parallelization, normalized to that obtained also by the MoAI platform but after turning off automatic parallelization and letting our expert engineers find and apply the best parallelization schemes found by trial and error. The engineering team consists of 2 PhDs, 5 MSs, and 1 BS, having 2.3 years of experience in LLM parallelization and optimization on average. All the numbers were measured on 32 AMD MI250 GPUs.