LAMINI CHOOSES SUPERMICRO GPU SERVERS FOR LLM TUNING OFFERING

Using Supermicro GPU Servers with the AMD Instinct™ MI300X Accelerators, Lamini is Able to Offer LLM Tuning At High Speed

Supermicro GPU Server - AS -8125GS-TNMR2

INDUSTRY

AI Cloud Service Provider

CHALLENGES

- Adding Large Scale AI Tuning
- Offering AI Solutions
- Needed GPU/AI Class Servers with the latest CPUs and GPUs
- Server provider with a record of fast time-to-delivery

Introduction

Lamini is developing an infrastructure for customers to run Large Language Models (LLMs) on innovative and fast servers. End-user customers can use Lamini's LLMs or build their own using Python, an open-source programming language. Lamini has developed a software environment for customers that allows them to focus on their business needs and develop innovative AI models.

Challenges

Lamini is a growing company that is customer-driven and is dedicated to giving enterprises an easy-to-use and reliable system for developing and running generative AI models by fine-tuning their existing assets. Customers will be able to run their complex AI models at maximum speed, producing results quickly and accurately.

Lamini just raised a $25M Series A earlier this year from notable investors in AI, technology, and enterprise. Lamini has been enabling Fortune 500 enterprise companies to build and run LLMs by fine-tuning existing open models, such as Meta's Llama 3 and Mistral 2, on their proprietary data, with less development effort. Lamini is a fast-growing generative AI company that is customer-driven and is dedicated to providing enterprise companies with an easy-to-use, secure, and reliable system for developing and running generative AI models by tuning existing open source models, such as Meta Llama 3, Mistral, on their proprietary data.

More recently, Lamini has invented Lamini Memory Tuning. This research breakthrough overcomes a seeming paradox in the AI world: achieving precise factual accuracy (i.e., reduced hallucinations) while upholding the generalization capabilities that make LLMs valuable in the first place. Moreover, Lamini is the first to offer a full LLM tuning stack that can run on both AMD and NVIDIA GPUs.

Lamini needed a state-of-the-art solution built with the latest CPUs and GPUs to be able to train these models in a reasonable amount of time.

Solution

Lamini chose the new Supermicro Universal GPU Server, the AS -8125GS-TNMR2. This server contains dual AMD EPYC™ 9534 CPUs with 64 cores and 128 threads. This CPU runs at 2.45 GHz and uses a maximum of 280 watts.

A key to the AI power of this Supermicro system is 8 AMD Instinct™ MI300X GPUs. The AMD Instinct MI300X GPUs are mounted on an industry-standard universal baseboard and are connected to the CPUs via the PCIe 5.0 x16 bus. This technology gives fast access between GPUs when the CPU must issue commands or send data that resides in host memory. The AMD Instinct MI300X is a high-performance accelerator explicitly designed for artificial intelligence (AI) and high-performance computing (HPC) applications. Highlights of the AS -8125GS-TNMR2 include:

Designed for AI and HPC:

- CDNA™ 3 Architecture: Utilizes AMD's latest CDNA 3 architecture, optimized for delivering exceptional performance in AI workloads.
- 19,456 Stream Processors: Packed with a massive number of cores for parallel processing of complex AI tasks.

Memory Designed for Speed:

- 192GB HBM3 Memory: Equipped with a large pool of high-bandwidth HBM3 memory, crucial for handling demanding AI datasets.
- 256 MB Last Level Cache (LLC): Offers additional on-chip memory for frequently accessed data, improving performance.

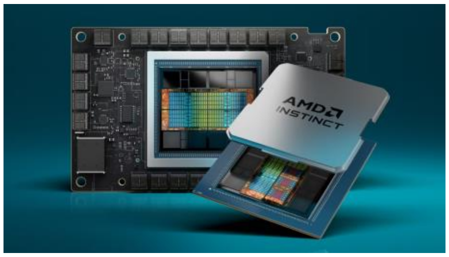
Connectivity and Power:

- PCIe 5.0 x16 Interface: Utilizes the latest PCIe 5.0 for high-speed communication between the CPUs and GPUs with the system.
- Many networking options are available, up to 400G
- Redundant power supplies are available.

Other Considerations:

- AMD ROCm™ Open Software: Offers a suite of optimizations for AI workloads and supports multiple open frameworks, models, and tools for development and deployment into AI environments.

Overall, the AMD Instinct MI300X is a powerful accelerator aimed at data centers and researchers working on intensive AI and HPC tasks. Its advanced architecture, high-bandwidth memory, and robust connectivity make it ideal for accelerating compute-heavy workloads.



SOLUTION

Supermicro Servers

AS -8125GS-TNMR2

- Dual AMD EPYC 9534 CPUs
- 1.5 TB DDR5 Memory
- 8x AMD Instinct™ MI300X GPUs

BENEFITS

- Creation of AI Offerings
- Reduced Time for Executions of AI Workloads

Benefits

Lamini quickly offered a high-end AI solution to its waiting customers. The Supermicro GPU server, the AS -8125GS-TNMR2, was the ideal initial solution for Lamini. With quick delivery, Lamini could install their customized software and set up a system where customers could easily access the system and run their AI tuning models. Supermicro was able to quickly deliver the systems due to its building block architecture approach for bringing new systems to market.

"Lamini is thrilled to be working with Supermicro on getting our AI cloud service up and running. We are ready to offer customers a leading-edge enterprise AI solution for LLM inference & tuning workloads. We look forward to an ongoing relationship with both Supermicro and AMD."

- Sharon Zhou, Co-Founder & CEO, Lamini

SUPERMICRO

Supermicro is a global leader in high performance, green computing server technology and innovation. We provide our global customers with application-optimized servers and workstations customized with blade, storage, and GPU solutions. Our products offer proven reliability, superior design, and one of the industry's broadest array of product configurations, to fit all computational need.

For more information, visit https://www.supermicro.com

LAMINI

Lamini is the LLM inference and tuning platform for the enterprise, enabling developers to build factual LLMs and deploy them anywhere quickly and securely. Lamini is used and trusted by Fortune 500 enterprise companies and leading AI startups. Our mission is to empower everyone to build their own superintelligence.

For more information, visit www.lamini.ai