Meta Launched an AI Acceleration Big Move, The AI Reasoning Chip First

HONG KONG, CHINA, May 22, 2023 /EINPresswire.com/ -- On May 18th local time, Meta announced on its website that in response to the rapid growth in demand for AI computing power in the next decade, Meta is implementing an ambitious plan to build next-generation infrastructure for AI.

Meta has announced its latest progress in building next-generation infrastructure for AI, including the first custom chip for running AI models, a new AI-optimized data center design, the first video transcoding ASIC, and an AI supercomputer RSC integrating 16,000 Gpus for accelerating AI training.

Meta sees AI as its core infrastructure. Since Meta broke ground on its first data center in 2010, AI has become the engine for the more than 3 billion people who use Meta's family of applications every day. From the Big Sur hardware in 2015 to the development of PyTorch, to the initial deployment of Meta's AI supercomputer last year, Meta is currently upgrading these infrastructures.

1. Meta's first generation AI reasoning Accelerator with 7nm process and 102.4TOPS computing power

MTIA (Meta Training and Inference Accelerator) is Meta's first in-house customized accelerator chip family for inference workloads.

AI workloads are ubiquitous in Meta's business and form the basis for a wide range of applications, including content understanding, information flow, generative AI, and AD ranking. As AI models grow in size and complexity, the underlying hardware systems need to provide exponentially increased memory and computing while maintaining efficiency. However, Meta found that the CPU could not meet the efficiency level required for its scale, so Meta designed the MTIA ASIC series of self-developed training and reasoning accelerators to address this challenge.

Since 2020, Meta has designed the first generation of MTIA Asics for its internal workloads. The accelerator uses the Taiwan Semiconductor Manufacturing 7nm process and operates at 800MHz, delivering 102.4TOPS at INT8 accuracy and 51.2TFLOPS at FP16 accuracy. It has a thermal design power (TDP) of 25W.

According to the presentation, MTIA provides more computing power and efficiency than the CPU, and by deploying the MTIA chip and GPU simultaneously, it will provide better performance, lower latency and higher efficiency for each workload.

2. Layout the next generation data center and develop the first video transcoding ASIC

Meta's next-generation data center design will support its current products while supporting training and reasoning for future generations of AI hardware. The new data center will be optimized for AI, supporting liquid-cooled AI hardware and high-performance AI networks connecting thousands of AI chips for data center scale AI training clusters.

Meta's next-generation data centers will also be faster and more cost effective to build, and will complement other new hardware, such as Meta's first internally developed ASIC solution, MSVP, designed to power Meta's growing video workloads, according to the website.

The need for video infrastructure has increased with new content technologies such as generative AI, which has prompted Meta to launch a scalable video processor called MSVP.

MSVP is the first ASIC developed internally by Meta for video transcoding. MSVPS are programmable and extensible, and can be configured to effectively support the high-quality transcoding required for on-demand, as well as the low latency and faster processing times required for live streaming. In the future, MSVP will also help bring new forms of video content -- including AI-generated content as well as VR and AR content -- to each member of the Meta app family.

3. AI supercomputer integrates 16,000 Gpus to support LLaMA large model accelerated training iteration

According to Meta's announcement, its AI supercomputer (RSC) is one of the fastest AI supercomputers in the world and aims to train the next generation of large-scale AI models to power new AR tools, content understanding systems, real-time translation technologies, and more.

The Meta RSC has 16,000 Gpus, all of which are accessible through a three-level Clos network structure, providing full bandwidth for each of the 2,000 training systems. Over the past year, the RSC has been promoting research projects like LLaMA.

LLaMA is a large language model built and open sourced by Meta earlier this year, with 65 billion parameters. Meta says its goal is to provide a smaller, higher-performance model that researchers can study and fine-tune for specific tasks without the need for significant hardware.

Meta has trained LLaMA 65B and the smaller LLaMA 33B with 1.4 trillion Tokens. Its smallest model, LLaMA 7B, has also been trained with a trillion Tokens. The ability to run at scale allows Meta to accelerate training and tuning iterations, releasing models faster than other enterprises.

4. Conclusion: The application of large model technology forces Dachang to accelerate the layout of the infrastructure
Meta customizes much of its infrastructure mainly because it enables it to optimize the end-to-end experience, from the physical layer to the software layer to the actual user experience. Because the stack is controlled from top to bottom, it can be customized to suit its specific needs. This infrastructure will enable Meta to develop and deploy larger AI models that are larger and more complex.

According to JAK Electronics, over the next few years, we will see chip design, dedicated and workload-specific AI infrastructure, specialization and increased customization of new systems and tools, and increased efficiency in product and design support. These will all provide increasingly sophisticated models and products based on the latest research, enabling people around the world to use this emerging technology.

JAK Electronics
JAK Electronics
+852 9140 9162
it@jakelectronics.com
