ASP-DAC 2024 - Designers' Forum

Session (1) Next-Generation AI Semiconductor Design

Time	23 January, 2023 (Tuesday)
Location	Room 110/111
Organizer	Kyumyung Choi (Seoul National University)
Chair	Changho Han (Kumoh National Institute of Technology)

1. Building the programmable, high performance and energy-efficient AI chip for ChatGPT

Speaker: Joon Ho Baek (CEO, FuriosaAI, Korea)

Abstract:

With the advent of ChatGPT and generative AI models, the demand for deep learning inference in data centers is exploding. While energy efficiency is important to reduce TCO (total cost of ownership), high performance is also essential to serve large models in production. Hyperscalers, on the other hand, emphasized the importance of programmability and flexibility for inference accelerators to track DNN progress. This talk will introduce high performance AI chips developed by FuriosaAI, designed to tackle all these challenge

2. Enabling AI Innovation through Zero-touch SAPEON AI Inference System

Speaker: Soojung Ryu (CEO, SAPEON, Korea)

Abstract:

SAPEON, a leading player in the AI semiconductor industry, has achieved remarkable success in bringing server-grade semiconductors to market through its X220 platform. These semiconductors have gained wide recognition for their exceptional performance in MLPerf benchmarks. SAPEON's Zero-Touch AI Inference System, powered by state-of-the-art semiconductor technology, provides an AI model inference SDK and a cloud-based Inference Serving Platform. This comprehensive solution enables Customer Engineers to perform AI model inference on NPUs with minimal involvement. SAPEON is actively engaged in collaborations with key stakeholders in the AI industry, playing a pivotal role in driving AI innovation forward as we prepare for the widespread adoption of AI inference using our cutting-edge X330 platform.

ASP-DAC 2024 - Designers' Forum

3. Processing-in-Memory in Generative AI Era

Speaker: Kyomin Sohn (Master, Samsung Electronics, Korea)

Abstract:

With the advancement of neural networks, particularly the emergence of LLM(large language models), a solution to address memory bottlenecks and improve system energy efficiency is required strongly. Currently, HBM DRAM is the only memory solution to meet high bandwidth requirements. In this talk, we will look at HBM DRAM that are currently actively used and discuss the next generation of HBM DRAM and what technologies are needed. However, the memory bottleneck caused by the Von Neumann architecture is no exception in the case of HBM, so we look at the PIM technology that is actively discussed to overcome this limitation. The concept and implementation cases of the recently developed HBM-PIM will be examined, and the next generation of DRAM-PIM will be discussed.

4. AiMX: Cost-effective LLM accelerator using AiM (SK hynixʹs PIM)

Speaker: Euicheol Lim (Fellow, SK hynix, Korea)

Abstract:

AI chatbot service has been opening up the mainstream market for AI services. But problems seem to exist with considerably higher operating costs and substantially longer service latency. As the LLM size continued to increase, memory intensive function takes up most of the service operation. Thatʹs why even latest GPU system does not provide sufficient performance and energy efficiency. To resolve it, we are introducing shorter latency and operating cost effective LLM accelerator using AiM (SK hynixʹs PIM) Weʹd like to introduce how to reduce service latency and decrease energy consumption through AiM, as well as explain the architecture of AiMX, an accelerator using AiM. Please come and see for yourself that AiM is no longer a future technology, but can be deployed to the existing system right now.