 

**Advanced Packaging for AI Hyperscale Computing:** Given the growth in large language models (LLMs) for artificial intelligence (AI) applications, a single neural network may now contain billions of parameters for computation. That number is expected to increase further at an unprecedented rate. For higher-performance, more energy-efficient neural network computing, more processor and memory dies need to be integrated in a single package. However, fabricating these integrated systems is difficult and costly. It is also challenging to increase memory bandwidth, due to the memory bottleneck arising from the relatively slow bandwidth of off-chip memory, and from the excessive energy consumption needed for frequent data access. A joint research team from KAIST-ETRI-AMKOR will describe a proposed advanced packaging solution they used to build a hyperscale processor unit (HPU) test vehicle, consisting of a pair of neural processor unit (NPU) chiplets and eight high-bandwidth memory 3 (HBM3) chiplets. The dual-NPU chiplets achieved peta-scale performance through energy-efficient tensor cores optimized for AI computation. All 10 chiplets were integrated on a large-scale RDL (redistribution layer) interposer and package substrate. High-density interconnect was used between the NPU and HBM to enable near-TB/s inter-chiplet bandwidth, allowing most data to be retrieved near the NPU for faster performance and greater energy-efficiency. The researchers will detail their overall design methodology and also will discuss thermal integrity simulations.

At left above is a conceptual view of the proposed 10-chiplet AI HPU architecture integrated above large-scale advanced package.

At right is a photo of a fabricated AI HPU daisy-chain test vehicle.

**(Interactive Presentation #38.5, “*The Energy-Efficient 10-Chiplet AI Hyperscale NPU on Large-Scale Advanced Package*,” J. Yoon et al, KAIST/ETRI/Amkor Technology)**