<Professor Joo-Young Kim's Lab.>

CastLab

■ **Contact information**

Professor : E3-2 #4202 TEL : 042-350-7461
Lab. : E3-2 #4209 TEL : N/A
Website : https://castlab.kaist.ac.kr

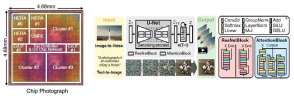
■ **Current state of the Lab. (in 2025 Spring Semester)**

Postdoctoral Fellows : 0 PhD Students: 22 Master's Student: 21

■ **Research Areas**

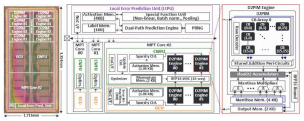
1. Neural Processing Unit

Neural Processing Unit (NPU) is AI-specialized hardware vital for edge and cloud computing. As AI usage grows, dedicated hardware becomes crucial for faster computations. In the edge scenarios like robotics, reinforcement learning, AR/VR demands real-time, energy-efficient processing, highlighting the need for dedicated hardware solutions.



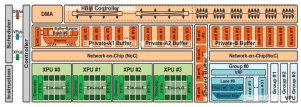
2. Processing-in-Memory

Traditionally, CPUs performed arithmetic and logic calculations, while memory stored data. However, technology scaling now results in compute units outpacing memory in speed, making data movement the bottleneck. The memory-centric approach, such as processing-in-memory (PIM), integrates computation into memory to avoid data movement.



3. Encryption

Privacy-preserving technology has become an increasingly crucial aspect in current information technology, in which private data are constantly being shared, processed, and stored online. Fully Homomorphic Encryption (FHE) enhances data privacy through encrypted computation. However, current hardware acceleration is insufficient for FHE due to complexity, necessitating specialized architecture.



■ **Recommended courses & Career after graduation**

- **Recommended Courses:** Digital System Design (EE303), Computer Architecture (EE312), Digital Electronic Circuits (EE372), Courses related to deep learning algorithms.
- **Career:** Silicon companies (Samsung, Apple, IBM) and IT companies (Microsoft, Google, Meta).

■ **Introduction to other activities besides research**

Beyond research , we enjoy a lot of activities including gatherings like strawberry parties, and lunch buddies; celebratory events for graduations and birthdays; sports like football and basketball.

■ **Introduction to the Lab.**

We aim to innovate modern computing systems through hardware specialization. To this end, we are focusing on co-design of multiple layers of computing system such as application, architecture, circuit, and technology.

■ **Recent research achievements ('24~'25)**

"Adelia: A 4nm LLM Accelerator with Steamlined Dataflow and Dual-Mode Parallelization for Efficient Generative AI Inference", IEEE Symposium on VLSI Techonology Circuits (VLSI), 2025.

"Hybe: GPU-NPU Hybrid System for Efficient LLM Inference with Million-Token Context Window", ACM/IEEE International Symposium on Computer Architecture (ISCA), 2025.

"Oaken: Fast and Efficient LLM Serving Online-Offline Hybrid KV Cache Quantization", ACM/IEEE International Symposium on Computer Architecture (ISCA), 2025.

"LightNobel: Improving Sequence Length Limitation in Protein Structure Prediction Model via Adaptive Activation Quantization", ACM/IEEE International Symposium on Computer Architecture (ISCA), 2025.

"ABC-FHE: A Resource-Efficient Accelerator Enabling Bootstrappable Parameters for Client-Side Fully Homomorphic Encryption", IACM/IEEE Design Automation Conference (DAC), 2025.

"EXION: Exploiting Inter- and Intra-Iteration Output Sparsity for Diffusion Models", ACM/IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2025.

"HuMoniX: A 57.3 fps 12.8 TFLOPS/W Text-to-Motion Processor with Inter-Iteration Output Sparsity and Inter-Frame Joint Similarity", IEEE International Solid-State Circuits Conference (ISSCC), 2025.