

2026 VLSI Review

포항공과대학교 반도체대학원 박사과정 박은빈

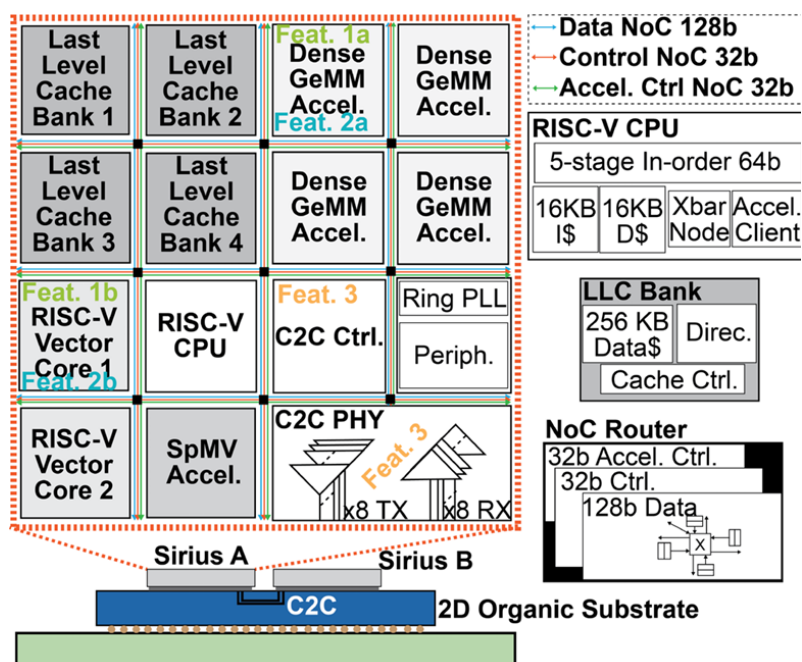
Session 21 SoCs for Reasoning and Robotics

2026 IEEE Symposium on VLSI Technology and Circuits의 Session 21에서는 초저전력 on-device reasoning, edge 기반 agentic AI, multimodal embodied AI, 그리고 Vision-Language-Action(VLA) 기반 실시간 로봇 제어를 중심으로 총 네 편의 논문이 발표되었다. 최근 AI 시스템은 단순한 이미지 인식이나 언어 모델 추론을 넘어, 주변 환경을 지속적으로 인식하고(reasoning), 판단하며, 실제 물리적 행동까지 수행하는 Physical AI 및 embodied AI 방향으로 빠르게 확장되고 있다. 이러한 변화에 따라 AI 프로세서 또한 높은 TOPS를 제공하는 범용 행렬 연산 가속기에서 벗어나, Transformer/vision/robotics workload의 서로 다른 연산 특성을 동시에 지원하고 제한된 메모리 및 전력 환경에서 end-to-end inference를 수행할 수 있는 시스템 수준의 설계가 요구되고 있다. Session 21의 논문들은 이러한 흐름을 반영하여, wearable의 always-on reasoning부터 agentic AI 및 실제 robot perception-to-action까지 다양한 형태의 차세대 edge AI SoC 구현 사례를 제시하고 있다.

#21-3 논문에서는 Transformer 기반 language reasoning, vision backbone, robotics perception과 같이 연산 특성이 서로 다른 multimodal workload를 효율적으로 처리하기 위한 dual-chiplet heterogeneous AI SoC인 Sirius를 제안하였다. 각 chiplet은 4개의 Dense GeMM accelerator, 1개의 SpMV accelerator, 2개의 RISC-V Vector(RVV) core를 통합하며, programmable LUT 기반 non-uniform quantization과 Dense+Sparse decomposition을 통해 dense 연산과 sparse outlier 연산을 각각 적합한 accelerator에서 수행한다. 또한 RVV와 GeMM accelerator의 local memory를 logically shared, physically distributed SPAD로 연결하여 intermediate data가 LLC나 DRAM을 거치지 않고 직접 전달되도록 함으로써 heterogeneous kernel fusion에서 최대 약 3×의 성능 향상을 달성하였다.

Sirius는 두 개의 chiplet을 custom chip-to-chip interconnect로 연결하여 multimodal pipeline을 병렬 실행할 수 있도록 확장성을 확보하였다. Intel 16 nm FinFET 공정으로 제작된 시스템은 Dense GeMM에서 최대 2.24 TFLOPS/W, SpMV에서 7.11 TFLOPS/W를 달성하였으며, 실제 robotics workload인 ViPlanner에서 0.97 TFLOPS/W, Qwen2.5-0.5B와 같은 SLM에서는 최대 16 tokens/s, 23 mJ/token의 성능을 보였다. 본 연구는 단일 AI accelerator의 peak throughput을 높이는 것보다, dense/sparse/irregular workload를

heterogeneous compute engine에 분배하고 chiplet 단위로 확장하는 embodied-AI SoC 구조를 제시했다는 점에서 의미가 있다.

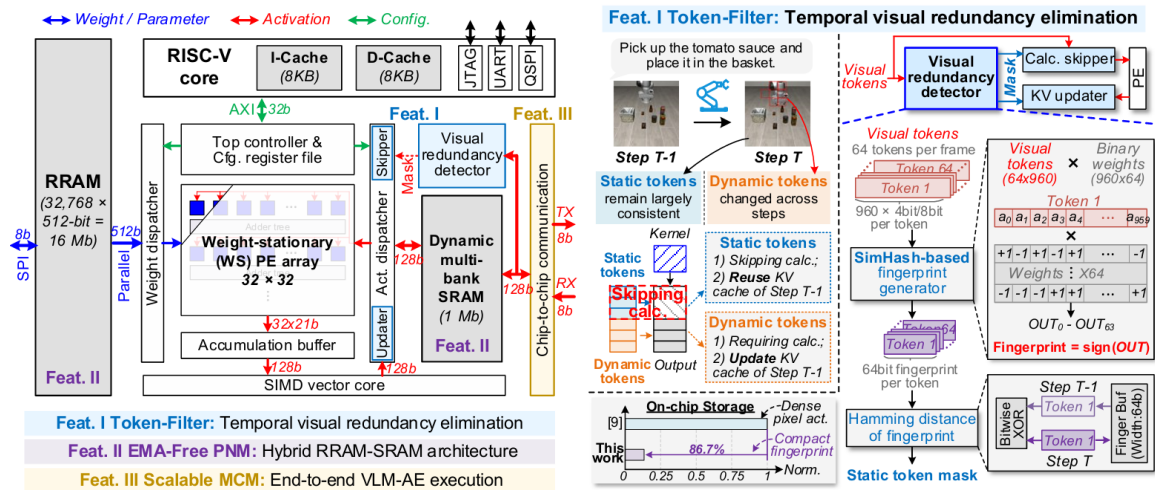


[그림 1] Sirius dual-chiplet SoC 구조 및 heterogeneous accelerator 구성

#21-4 논문에서는 Vision-Language Model(VLM)의 perception과 Diffusion Transformer 기반 Action Expert(AE)를 edge에서 end-to-end로 실행하기 위한 실시간 Vision-Language-Action(VLA) processor를 제안하였다. VLM에서는 연속된 robot action step 사이에 반복되는 static visual token이 많은 점에 착안하여, 960-dimensional token을 64-bit fingerprint로 변환하는 SimHash 기반 temporal token filter를 적용하였다. 이를 통해 static token의 연산과 KV update를 생략하여 평균 59.6%의 연산량을 줄였으며, success rate는 90%에서 87%로 소폭 감소하였다. 반면 memory-bound 특성을 갖는 AE에는 16 Mb RRAM과 1 Mb SRAM을 결합한 hybrid Processing-Near-Memory(PNM) 구조를 적용하여 weight와 activation의 external memory access를 제거하였다.

또한 16-chip MCM 구조를 적용하고 chip 간 activation transfer와 computation을 overlap하여 VLM과 AE의 전체 pipeline latency를 크게 감소시켰다. 55 nm CMOS로 제작된 prototype은 Libero benchmark에서 5.31–6.6ms의 end-to-end perception-to-action latency, 10.46–12.74 mJ/inference, 그리고 최대 150 Hz의 robot control rate를 달성하였다. 특히 기존 accelerator가 VLM 또는 Action Expert 중 일부만 가속하던 것과 달리, 본 연구는 perception부터 action generation까지 전체 VLA pipeline을 실제 silicon에서 구현했다는 점이 특징적이다. 따라서 algorithm-level token filtering, memory-centric PNM, multi-

chip scheduling을 함께 최적화하여 실시간 Physical AI 및 robotics를 위한 end-to-end VLA hardware의 가능성을 보여준 연구로 평가할 수 있다.



[그림 1] SimHash 기반 temporal token filtering 및 VLA processor 전체 구조

저자정보



박은빈 박사과정 대학원생

- 소속 : 포항공과대학교
- 연구분야 : HW-SW co-optimization 및 양자오류정정부호
- 이메일 : eunbin@postech.ac.kr, eunbin.epiclab@gmail.com
- 홈페이지 : <https://sites.google.com/view/epiclab/member/ebpark>