

2026 VLSI Review

연세대학교 전기전자공학과 석박통합과정 김동욱

Session 23 ASICs for Emerging Workloads

2026 VLSI의 Session 23은 ASICs for Emerging Workloads 라는 주제로 총 4편의 연구가 발표되었다. 연구들은 소형(compact) video encoder 시스템, spatiotemporal reconstruction, 웨어러블 AI 및 dynamic RF 신호 분석을 대상으로, 각 응용의 연산과 데이터 흐름에 최적화된 하드웨어 솔루션을 제시했다. 본 리뷰에서는 새로운 워크로드의 병목을 해결하는 아키텍처적 기여에 주목하여, spatiotemporal reconstruction을 위한 4DGS 프로세서, 다양한 AI 모델을 지원하는 Hybrid CIM-NPU 프로세서, 그리고 dynamic RF 환경에 대응하는 하드웨어 스케줄링 기반 스트리밍 프로세서를 다룬 3편의 연구를 중심으로 살펴본다.

#23-2 KAIST 연구진이 발표한 본 논문 [Specter: 376 fps 2.67 nJ/pixel 4DGS Processor for Spatiotemporal Reconstruction with Contribution-Ordered Rendering and Deformation-Aware Modeling]은 시간에 따라 변화하는 scene의 재구성을 위한 4D Gaussian Splatting (4DGS) 프로세서인 Specter를 제안했다. 4DGS는 3차원 Gaussian으로 표현한 공간 정보에 시간에 따른 변형을 반영하는 기술이지만, 매 시점 수십만 개의 Gaussian을 처리해야 하므로 외부 메모리 접근과 연산 비용이 크다는 한계가 있었다. 특히, 실제 픽셀 색상에 기여하지 않는 Gaussian까지 타일에 할당하고 처리하는 과정에서 상당한 낭비가 발생했다. 본 연구에서는 이를 해결하기 위해, Gaussian의 불투명도(opacity)와 형상(shape)에 따라 타일 할당 범위를 조정하는 Adaptive Tile Allocation Core (ATAC)를 설계하여 baseline 대비 외부 메모리 접근을 67.3% 줄였다. 또한, Gaussian의 기여도가 높은 픽셀부터 처리하고 이후의 불필요한 연산을 타일 및 서브타일 단위로 생략하는 Contribution-ordered Hierarchical Skipping Core (CHSC)를 도입했다. Modeling 단계에서는 타일별 Gaussian ID 목록의 해시(hash)값을 비교하여 변화가 없는 것으로 판단한 타일의 재계산을 생략함으로써 추가적인 연산 절감을 달성했다. 결과적으로 28nm CMOS 공정으로 제작된 12.96mm² 칩은 1352×1014 해상도에서 최대 376fps의 rendering 성능과 최대 112.3 iterations/s의 modeling 성능을 기록했다. 이는 dynamic scene rendering뿐 아니라 modeling까지 지원하면서, scene의 공간적·시간적 중복성을 하드웨어 수준의 연산 생략으로 연결했다는 점에서 의미가 있다.

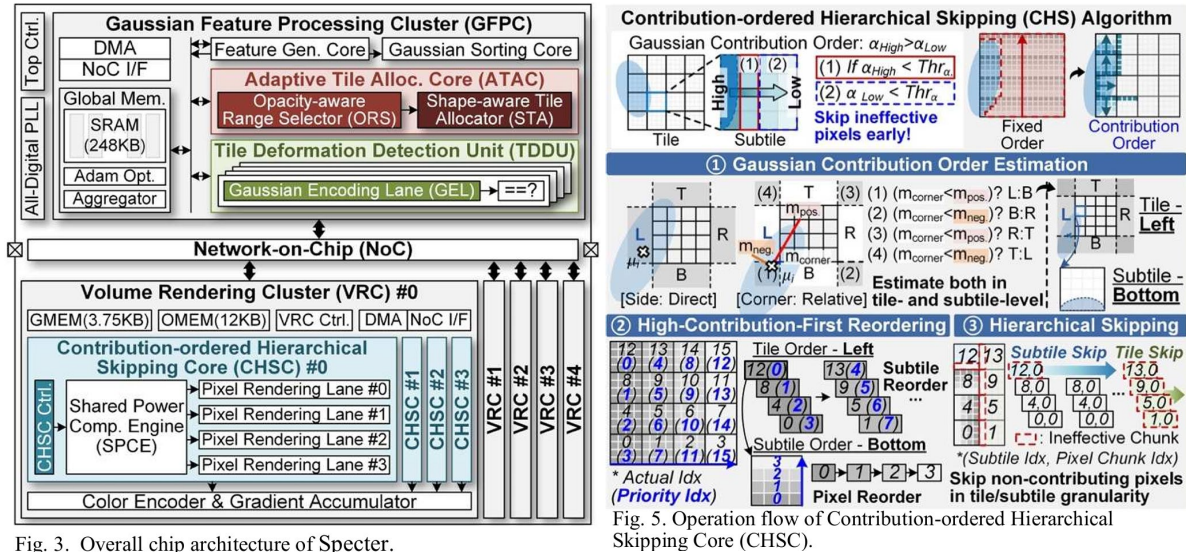


Fig. 3. Overall chip architecture of Specter.

Fig. 5. Operation flow of Contribution-ordered Hierarchical Skipping Core (CHSC).

[그림 1] Specter 전체 구조와 기여도 기반 계층적 렌더링 생략 기법 (원문 Fig. 3, Fig. 5)

#23-3 UNIST와 연세대학교 연구진이 발표한 본 논문 [HCNP: A 70.2 TOPS/W Hybrid CIM-NPU Processor with In-Streaming Processing for Energy-Efficient CNN/Transformer Acceleration in HMD]는 전력 및 발열 제약이 큰 Head-Mounted Display (HMD)를 위한 Hybrid CIM-NPU 프로세서를 제안했다. Compute-in-Memory (CIM)는 고정된 가중치를 사용하는 대규모 행렬 곱셈에 높은 에너지 효율을 제공하지만, 다양한 보조 연산과 데이터 흐름을 단독으로 처리하기에는 제약이 있다. 반면 NPU는 연산의 유연성이 높지만, 두 구조를 별도 칩으로 구성하면 중간 데이터의 외부 메모리 접근이 새로운 병목이 된다. 본 연구에서는 CIM과 NPU를 단일 칩에 통합하고, 데이터가 연산 블록 사이를 이동하는 동안 필요한 처리를 수행하는 In-Streaming Processing Core (ISPC)를 도입했다. ISPC는 nonlinear activation function과 depth-wise convolution 등의 연산을 스트리밍 과정에서 수행하여 layer fusion을 지원하며, 모바일 CNN에서 외부 메모리 접근을 최대 79.7% 줄였다. 또한, attention 기반 중요도에 따라 불필요한 KV cache 항목을 제거하는 Runtime KV Cache Eviction (RKVe)을 NPU와 co-design하여 KV cache 메모리 요구량과 attention MAC 연산량을 각각 최대 79.8%, 32.1% 절감했다. 여기에 flexible-route network-on-chip (FR-NoC)과 single-column-multiple-output (SCMO) CIM mapping 구조를 결합해, 평가한 모바일 CNN 및 LLM에서 평균 CIM 활용도를 각각 77.2%, 81.8%까지 높였다. 결과적으로 28nm CMOS 공정으로 제작된 칩은 8-bit 연산에서 최대 70.2 TOPS/W를 달성했으며, 16-bit MobileNet-v2와 LLaMA2-7B 워크로드에서는 각각 15.0, 17.9 TOPS/W의 효율을 보였다. 이는 CIM 연산기 자체의 효율을 넘어, 이종 연산 블록 사이의 데이터 이동과 활용도까지 함께 최적화하는 co-design을 제시했다는 점에서 주목할 만하다.

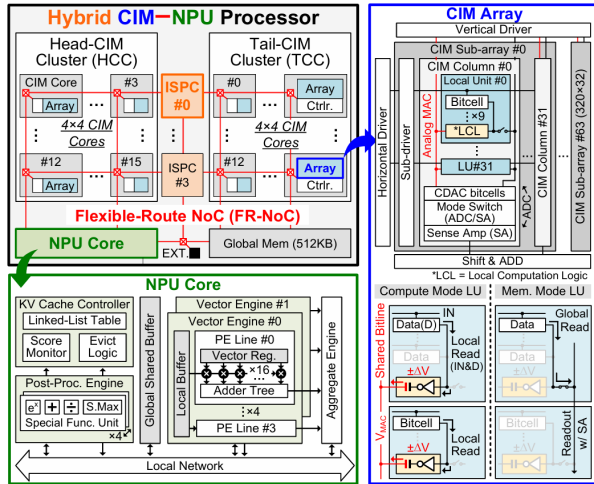


Fig. 3. Overall architecture of the proposed highly energy-efficient Hybrid CIM-NPU Processor for HMD platform.

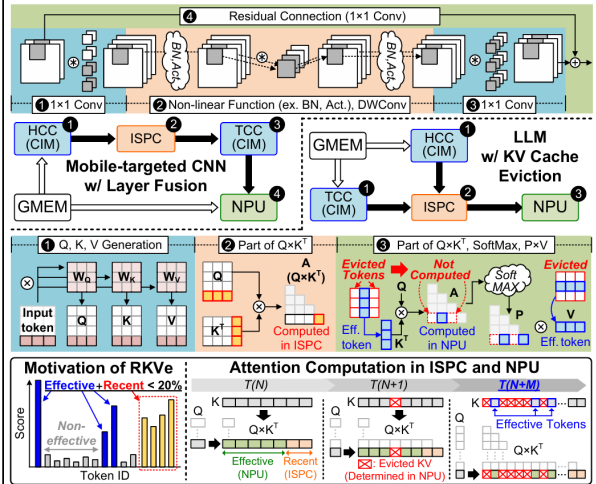


Fig. 4. Layer fusion dataflow of the proposed HCNP for MCNN and LLM, with the motivation and operation of the proposed RKVe.

[그림 2] HCNP의 Hybrid CIM-NPU 구조와 layer fusion을 위한 데이터 흐름 (원문 Fig. 3, Fig. 4)

#23-4 University of Michigan 연구진이 발표한 본 논문 [A 248.5 GFLOPS/mm², 1.67 TFLOPS/W Streaming Processor with Hardware-Level Scheduling for Advanced Spectrum Sensing]은 무선 환경에서 빠르게 변하는 RF 채널에 대해, 다수 채널을 분석하기 위한 스트리밍 프로세서를 제안했다. Spectrum sensing에서는 현재 신호의 분석 결과에 따라 다음에 실행할 커널이 결정되므로, 실행 순서와 자원 배치를 사전에 스케줄링하기 어렵다. 또한, 개별 커널의 실행 시간이 짧아 host CPU가 매번 커널을 배치하고 loading하는 방식은 큰 오버헤드를 유발한다. 본 연구에서는 이를 해결하기 위해, 77개의 Processing element (PE)와 96개의 Switch로 구성된 2차원 배열에 하드웨어 Kernel Launcher (KL)를 통합했다. KL은 커널 종료 시 실행 그래프의 분기를 판단하고, PE-Switch-DMA 자원의 가용성을 확인하여 다음 커널을 배치한다. 특히, 실행 중인 커널의 후속 분기 후보를 미리 살펴보고 프로그램을 preload 함으로써, host CPU의 개입 없이 최단 10ns의 reconfiguration 시간을 달성했다. PE 내부에는 복소수 MAC 등을 포함한 VLIW 구조를 적용하여 FFT butterfly를 단일 PE에 mapping할 수 있도록 했으며, 사용하지 않는 연산기와 메모리에 대한 세밀한 clock/power gating으로 전력 소모를 줄였다. 결과적으로 28nm 공정으로 제작된 4.78mm² 칩은 FP16 기준 최대 248.5 GFLOPS/mm²의 연산 밀도와 1.67 TFLOPS/W의 에너지 효율을 기록했다. 다중 채널의 동시 실행으로 PE 활용도를 약 5배 높였고, 제안한 스케줄링을 적용하지 않은 baseline 대비 최대 8.3배의 end-to-end 성능 향상을 보였다. 이는 데이터에 따라 실행 경로가 달라지는 워크로드에서, 연산기 성능뿐 아니라 작업 전환과 자원 관리의 지연까지 하드웨어 설계의 핵심 대상으로 삼았다는 점에서 의미가 있다.

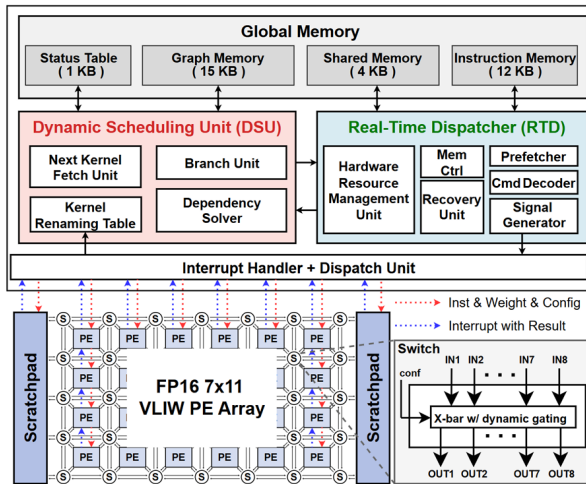


Fig. 3. Overview of proposed architecture. Kernel launcher (KL) is shown in upper part including Global Memory, DSU, RTD, interrupt handler and dispatch unit.

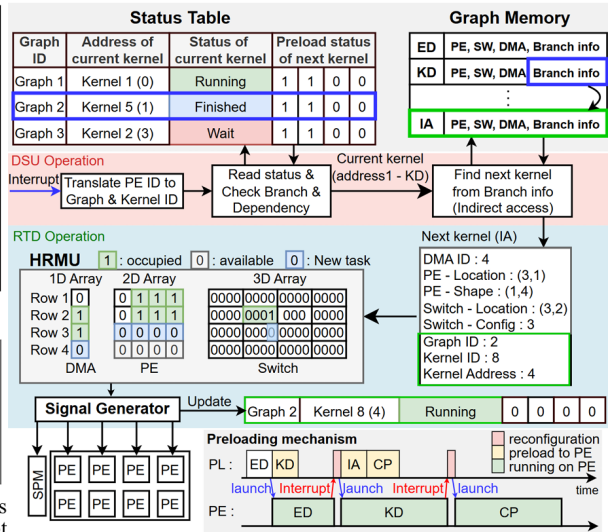


Fig. 5. Detailed explanation of kernel launching protocol

[그림 3] 스트리밍 프로세서 전체 구조와 하드웨어 기반 커널 실행 프로토콜 (원문 Fig. 3, Fig. 5)

저자정보



김동욱 석박통합과정 대학원생

- 소속 : 연세대학교
- 연구분야 : 메모리 시스템, PIM
- 이메일 : dwkim3852@yonsei.ac.kr
- 홈페이지 : <https://dtl.yonsei.ac.kr>