

2026 VLSI Review

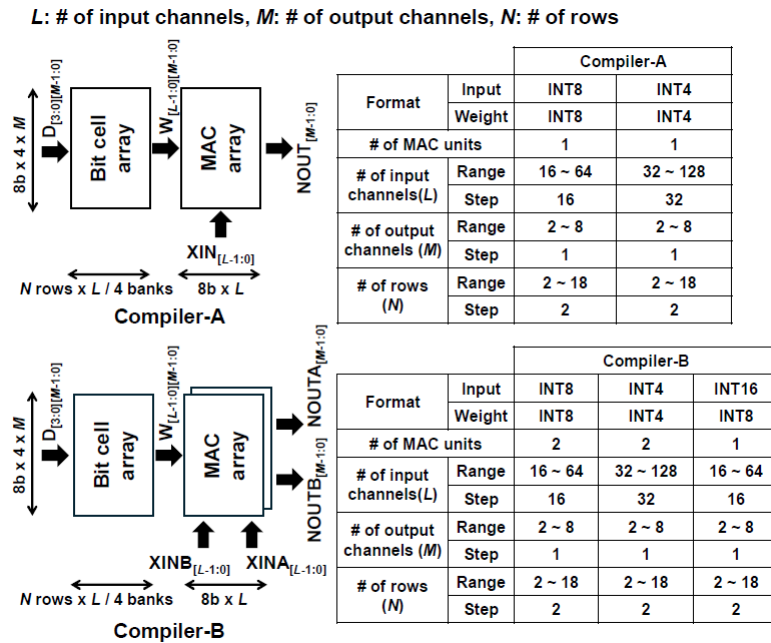
연세대학교 전기전자공학부 박사과정 여민준

Session 8 Compute-in-Memory

이번 2026 VLSI Symposium Session C8은 Compute-in-Memory(CIM)를 주제로, 2nm 첨단 공정의 디지털 CIM 컴파일러부터 부동소수점 CIM 매크로, ROM 기반 초저전력 키워드 인식 SoC, 스핀트로닉스 소자 기반 뉴로모픽 매크로에 이르기까지 메모리 내 연산의 스펙트럼 전반을 다루었다. 발표된 연구들은 다양한 데이터 포맷과 GEMM 형상에 대한 유연성, 트랜스포머 활성값의 outlier로 인한 정밀도 손실, 가산기 트리의 면적 부담, 자율주행 수준의 기능 안전 요구 등 CIM이 실제 응용에 배치될 때 마주하는 문제들을 회로-아키텍처-알고리즘 공동 설계로 해결하고자 하였다. 본 리뷰에서는 그 중 특징적인 네 편의 논문을 선정하여 핵심 아이디어와 성과를 정리한다.

#C8-1 TSMC에서 발표한 본 논문은 2nm 나노시트 공정에서 구현된 SRAM 기반 디지털 CIM(DCIM) 컴파일러로, 다양한 정수 데이터 포맷과 가중치당 복수의 MAC 유닛을 지원하여 면적 효율(TOPS/mm²)과 GEMM 활용률을 동시에 높이는 것을 목표로 한다. 기존 DCIM 매크로들이 단일 포맷과 고정된 MAC 구조에 묶여 최신 신경망의 혼합 정밀도 양자화와 다양한 GEMM 형상에 대응하기 어려웠던 한계를 극복하기 위해, 본 연구는 하나의 가중치를 두 개의 MAC 유닛이 공유하는 Compiler-B 구조를 제안하였다. 비트셀 어레이에서 읽은 가중치를 두 MAC 유닛이 공유하고 각각 독립적인 입력 활성값(XINA/XINB)을 받아 별도의 출력을 생성함으로써 가중치 재사용을 극대화하며, 행 전환(row switch) 시 로컬 SRAM 읽기 에너지가 두 MAC에 분산되어 매 사이클 행 전환 조건에서 단일 MAC 구조 대비 2% 높은 에너지 효율을 보인다. MAC 유닛은 4b×4b 곱셈기 네 개 (MUL0-MUL3)와 두 개의 가산기 트리로 구성되어 INT8/INT4의 signed-unsigned를 모두 지원하며, INT4×INT4 모드에서는 MUL1-MUL2를 게이팅하고 시프트 없이 두 트리의 부분합을 누적하여 입력 채널 수를 두 배로 확장하고, NOUTA/NOUTB를 8비트 시프트 후 결합하면 INT16 입력도 처리한다. 0.0437μm²의 래치 기반 12T 비트셀에 인터리브 WL 구조를 적용하여 4개의 가중치를 한 사이클에 갱신하는 높은 쓰기 대역폭도 확보하였다. 테스트 칩 측정 결과 -40°C에서 95% 수율로 VMIN 0.38V 이하, 1.2V에서 FMAX 2.0GHz(Compiler-A)를 달성하였으며, 0.5V에서 INT8 88.8 TOPS/W, INT4 234.4 TOPS/W, 1.2V에서 INT4 511.9 TOPS/mm²의 효율을 기록하여 3nm 선행 컴파일러 대비 면적 효율

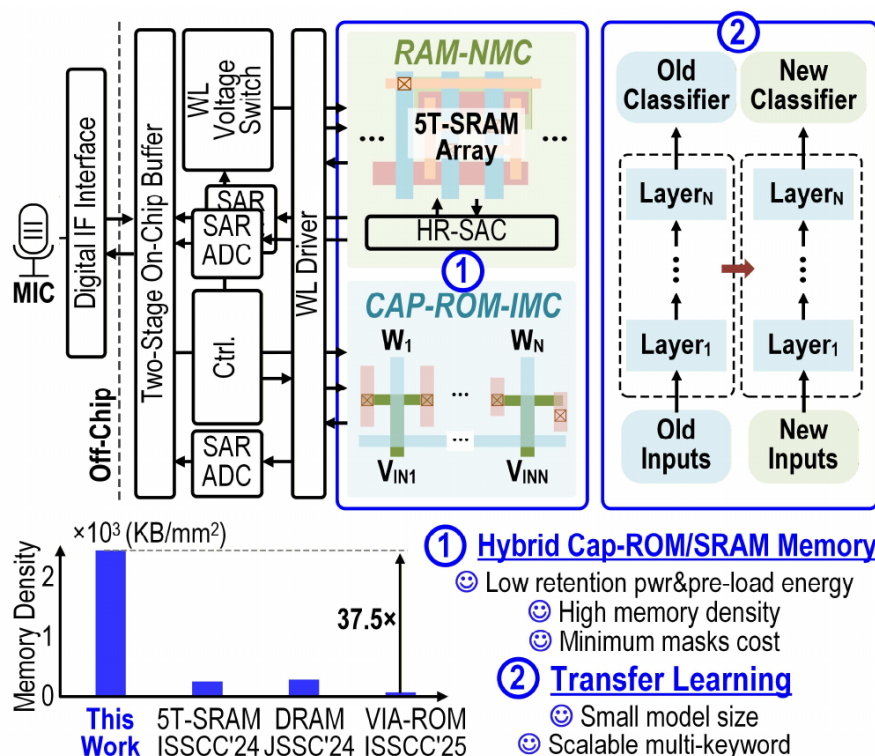
을 크게 개선하였다.



[그림 1] "A 2nm 234.4 TOPS/W and 511.9 TOPS/mm² Digital Computing-in-Memory Compiler with Multiple MAC Units per Weight and Multiple Data Format Support" 2nm CIM 컴파일러 (Compiler-A/B)의 구조 및 지원 구성 범위

#C8-2 마카오 대학교와 EPFL이 공동 발표한 본 논문은 28nm 공정에서 구현된 0.0257mm², 1.16μW의 항상 켜짐(always-on) 키워드 인식(KWS) SoC로, 어휘 수가 늘어날 수록 모델과 메모리 요구량이 급증하여 전력이 수백 μW 수준으로 치솟는 문제를 해결하는 것을 목표로 한다. 기존 10개 이상 키워드를 지원하는 SoC들이 820~949μW를 소모하고 온칩 SRAM 부족으로 오프칩 플래시 접근 에너지가 추론 에너지의 20배를 넘었던 한계에 대해, 본 연구는 Cap-ROM 기반 인메모리 연산(Cap-ROM-IMC)과 5T-SRAM 기반 근접메모리 연산(SRAM-NMC)을 결합한 하이브리드 구조를 제안하였다. 모델 가중치의 92.7%는 금속 배선의 기생 커패시턴스로 고정 저장되는 트랜지스터 없는 Cap-ROM 셀에 담겨 전하 공유 방식의 비트 직렬 MAC을 수행하며, 나머지 7.3%만 1.16KB의 상시 유지 5T-SRAM에 저장된다. Virtuoso로 커패시턴스-가중치 매핑을 추출하고 M1-M6 금속 형상을 자동 생성하는 설계 흐름을 통해 신경망에서 물리 어레이로의 오류 없는 매핑을 구현하였고, 최소 2장의 마스크만으로도 VIA-ROM 대비 5.4배, 전체로는 37.5배 높은 메모리 밀도(2372.6KB/mm²)를 달성하면서 정적 전력을 거의 0으로 유지하여 시스템 전력을 44% 절감하였다. 또한 세 개의 공유 합성곱 계층과 키워드별 분류기로 구성된 전이 학습 기반 1D-CNN을 도입하여 20개 키워드 분류기를 모두 온칩에 저장하고 재학습 없이 임의의 5개 키워드를 동시에 활성화할 수 있게 하였으며, 클래스당 파라미터를

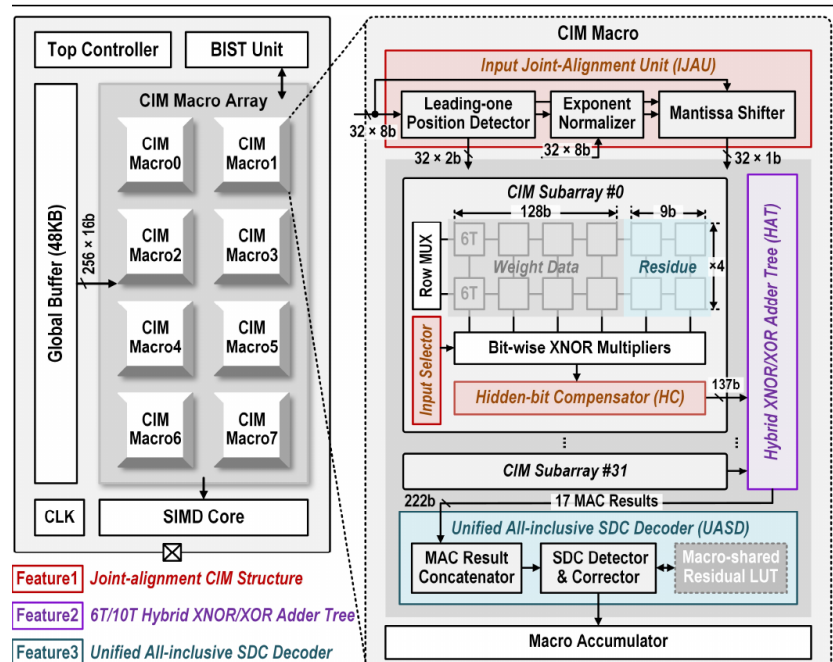
1.45KB로 줄였다. 측정 결과 GSCD-v2에서 92.1%의 정확도와 2ms 지연을 유지하면서 10개 이상 키워드 SoC 대비 700배 이상 낮은 전력을 기록하였다.



[그림 2] “A 0.0257 mm², 1.16 μ W, Mask-Efficient KWS SoC with Hybrid Cap-ROM-IMC/SRAM-NMC and Transfer Learning for Flexible Multi-Keyword Recognition” 제안 KWS SoC의 블록도와 두 가지 핵심 해법(하이브리드 Cap-ROM/SRAM 메모리, 전이학습)

#C8-3 칭화대학교와 홍콩과기대학교가 공동 발표한 본 논문은 28nm 공정에서 구현된 자율주행용 부동소수점 CIM(FP-CIM) 매크로로, 정확도·성능·신뢰성 요구가 모두 엄격한 안전 필수 응용에 FP-CIM을 적용하는 것을 목표로 한다. 기존 FP-CIM이 지수만 정렬 (exponent-only alignment)한 뒤 절단 오차를 줄이기 위해 가수 비트폭을 확장하여 비트 직렬 연산 지연이 22–44% 늘어나고, 가산기 트리가 매크로 면적의 34–71%를 차지하며, 메모리와 연산 로직을 ECC와 TMR로 따로 보호하느라 과도한 비용이 드는 세 가지 한계를 지적하였다. 이를 해결하기 위해 첫째, 히든 비트와 가수의 선행 1 사이의 본질적으로 무효인 비트 공간을 유효 비트로 재활용하는 지수-가수 joint-alignment 구조를 제안하여, 비트폭 확장 없이 1.62%의 추가 면적만으로 정확도 손실을 3.97배 줄이고 31.24%의 면적 효율 손실을 회피하였다. 둘째, Sum과 Cout에 공통된 XOR 항을 미리 계산해 재사용하는 하이브리드 XNOR/XOR 가산기 트리(HAT)를 도입하여 1단은 6T XNOR 기반, 그 외단은 10T XOR 기반 전가산기로 구성함으로써 28T 가산기 트리 대비 면적을 71.35% 절감하였다. 셋째, 가중치를 잔여 코드(residue code)로 인코딩해 저장하고 최종 MAC 결과의 나머지 검사로 오류를 검출·복원하는 통합 SDC 디코더(UASD)를 통해 8.41%의 면적

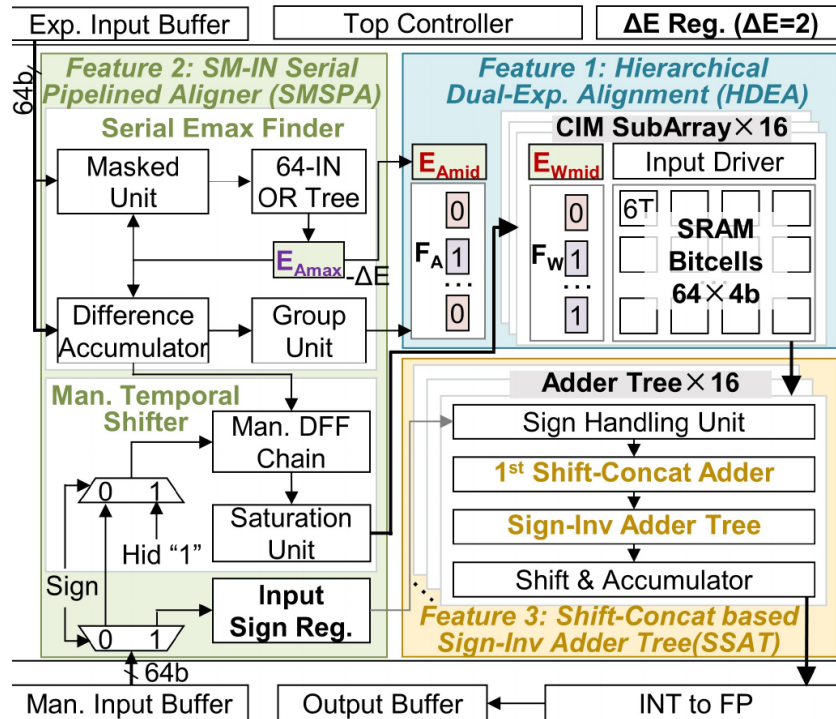
비용만으로 0.50%의 SDC율과 ISO 26262 ASIL-D 안전 등급을 달성하였다. 측정 결과 BF16 기준 44.15 TFLOPS/W, 1.57 TFLOPS/mm²로 선행 FP-CIM 대비 각각 1.51–1.62배, 1.37배의 개선을 보였다.



[그림 3] “A 28nm 44.15TFLOPS/W and 1.57TFLOPS/mm² Joint-Alignment Floating-point CIM Macro with ISO26262 ASIL-D Safety Level for Autonomous Driving” 제안 FP-CIM의 전체 아키텍처와 세 가지 핵심 기능

#C8-5 화중과기대학교와 홍콩과기대학교가 공동 발표한 본 논문은 40nm 공정에서 구현된 FP8 디지털 CIM 매크로로, 트랜스포머 활성값과 가중치의 outlier 분포로 인해 INT 양자화가 심각한 클리핑·반올림 오차를 겪는 에지 트랜스포머 가속을 목표로 한다. 기존 FP-CIM은 공유 지수로 사전 정렬(pre-alignment)하면 정확도가 떨어지고, 사후 정렬(post-alignment)은 대규모 시프터와 19비트 폭의 가산기 트리로 면적·에너지 부담이 큰 상충 관계에 갇혀 있었다. 본 연구는 이를 세 가지 기법으로 해결하였다. 첫째, 계층적 이중 지수 정렬(HDEA)은 입력을 동적으로 분류하여 outlier는 E_{max} 에, 정상값은 $E_{mid}=E_{max}-\Delta E$ 에 정렬함으로써 좌측 시프트량을 제한하고 가산기 트리 폭을 8비트로 줄이며, 설계 공간 탐색을 통해 $\Delta E=2$ 를 정확도와 면적의 최적점으로 도출하였다. 둘째, 부호-크기 입력 직렬 파이프라인 정렬기(SMSPA)는 마스크 기반 직렬 E_{max} 탐색, 포화 검사로 불필요한 시프트를 게이팅하는 차분 누산기, DFF 체인을 이용한 시간 영역 시프터로 병렬 정렬과 비트 직렬 MAC 간의 파이프라인 정체를 제거하여 지연·면적·에너지를 각각 20%, 41.1%, 34.1% 절감하였다. 셋째, Shift-Concat 기반 부호 반전 가산기 트리(SSAT)는 첫 단에서 겹치는 상위 4비트만 더하고 하위 4비트는 연결하며, 이후 단에서는 부호 비트를 반전시켜 부호 있는 덧셈을 부호 없는 덧셈으로 바꿔 불필요한 부호 확장과

스위칭을 제거하여 면적 25.9%, 에너지 18.2%를 절감하였다. 그 결과 0.8–1.1V, 100–326MHz에서 24.85–80.59 TFLOPS/W를 달성하여 최신 사후 정렬 FP-CIM 대비 2.81–6.66 배 높은 에너지 효율을 기록하였다.



[그림 4] "A 40nm 24.85-to-80.59 TFLOPS/W FP8 Computing-in-Memory Macro with Outlier-Aware Hierarchical-Alignment for Edge Transformers" 제안 CIM 매크로의 전체 아키텍처(SMSPA, HDEA, SSAT)

저자정보



여민준 박사과정 대학원생

- 소속 : 연세대학교
- 연구분야 : 저전력 고 신뢰성 SRAM 설계, Computing-in-memory 설계
- 이메일 : ymj5887@yonsei.ac.kr
- 홈페이지 : <http://vlsisys.yonsei.ac.kr/>