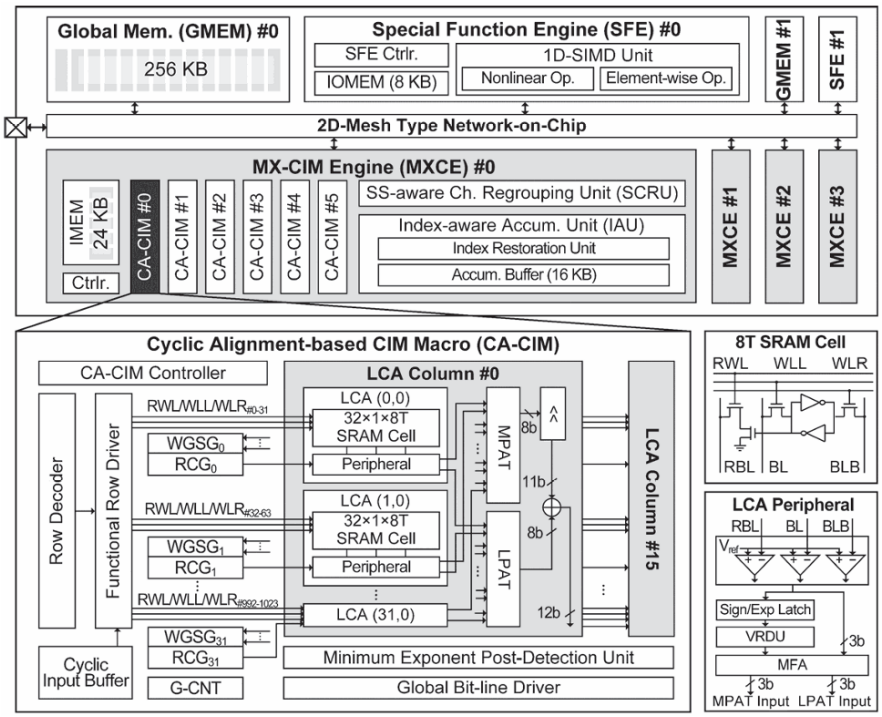
A-SSCC 2025 Review

연세대학교 전기전자공학부 박사과정 여민준

Session 11 Advanced Digital Compute-In-Memory For Edge AI

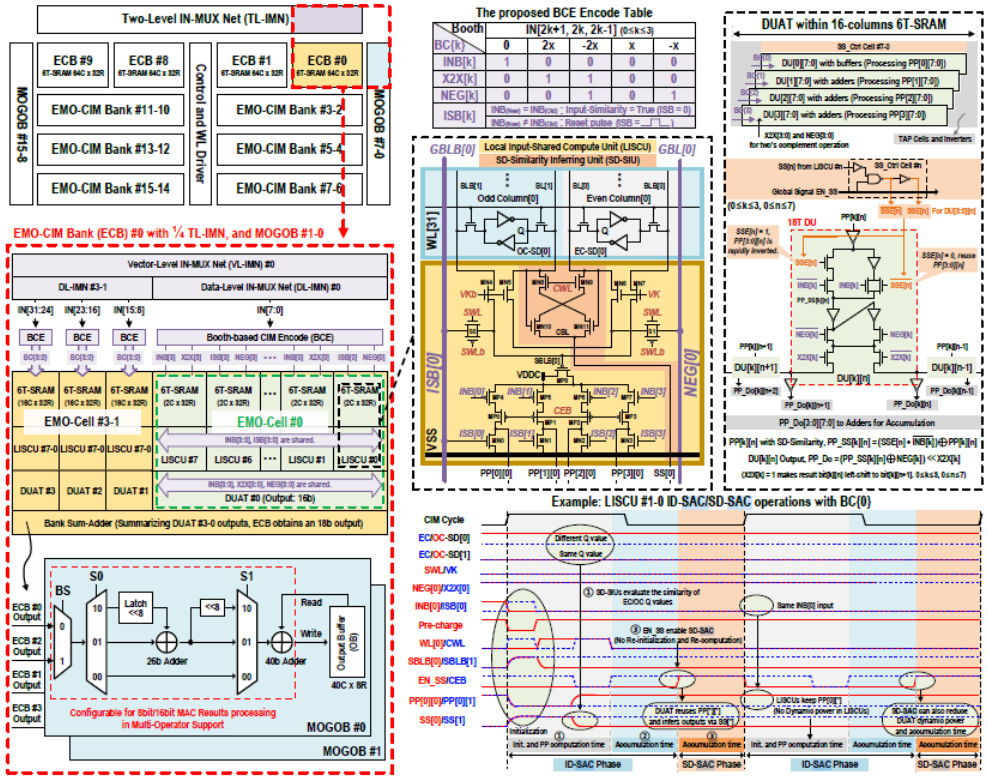
이번 A-SSCC 2025의 Session 11은 Edge AI를 위한 디지털 인메모리 컴퓨팅(Compute-In-Memory, CIM)을 주제로 총 4편의 논문이 발표되었다. 하나의 세션 안에서 생성형 AI용 sub-8bit Microscaling data format 기반 디지털 CIM 가속기(MIDAS), CNN/Attention/DW를 단일 매크로에서 지원하는 유사도 기반 EMO-CIM, RRAM–eDRAM 융합 메모리를 사용하는 부동소수점(FP) CIM 엔진(CENTAUR), 그리고 BF16 포맷을 사용하는 초고효율 디지털 CIM 매크로까지, 정수–BF16–부동소수점 전 영역을 포괄하고 있다는 점이 특징이다. 공통적으로는 (1) CNN뿐 아니라 Attention, Depthwise Conv, Diffusion 등 최신 네트워크의 다양한 연산자를 단일 하드웨어에서 지원하려는 "multi-operator" 지향, (2) 데이터 포맷(Microscaling, BF16, FP)과 Dataflow(3D-MAC, SAC, RCBS 등)을 활용해 에너지·면적 효율(EF/AF)을 극대화하려는 시도가 돋보인다.

#11-1 KAIST에서 발표한 28nm 디지털 CIM 기반 생성형 AI 가속기 MIDAS로, Microscaling data format과 "Cyclic Alignment" data flow를 활용해 가중치 저장과 CIM Array 효율을 동시에 개선한 점이 핵심이다. Microscaling 포맷에 맞춘 Cyclic alignment 기반 CIM은 동일한 메모리 용량에서 더 많은 가중치를 수용하면서 에너지 효율 1.72배, 면적 효율 2.47배 향상을 달성하였다. 또한 최소 지수(min-exponent)를 검출하는 relative counter 기반 포스트 프로세싱을 통해 디지털 구현 대비 전력 68.8%, 면적 63.2%를 절감하였고, 센스앰프 기반 dynamic bit-significance 처리 및 shared scale-aware 채널 리그룹핑으로 추가 27.3%의 에너지 절감을 얻었다. 28nm 실리콘 측정 결과, 기존 최첨단 디지털 CIM 가속기 대비 시스템 레벨 에너지 효율 1.58배, 매크로 레벨 에너지 효율 3.32배를 보고하였다. 기존 디지털 CIM 연구들이 주로 이미지 분류와 같은 classification 중심 워크로드를 대상으로 했던 것에 비해, MIDAS는 Microscaling 포맷과 Cyclic alignment를 활용해 생성형 AI까지 적용 범위를 확장했다는 점에서 의미가 크다



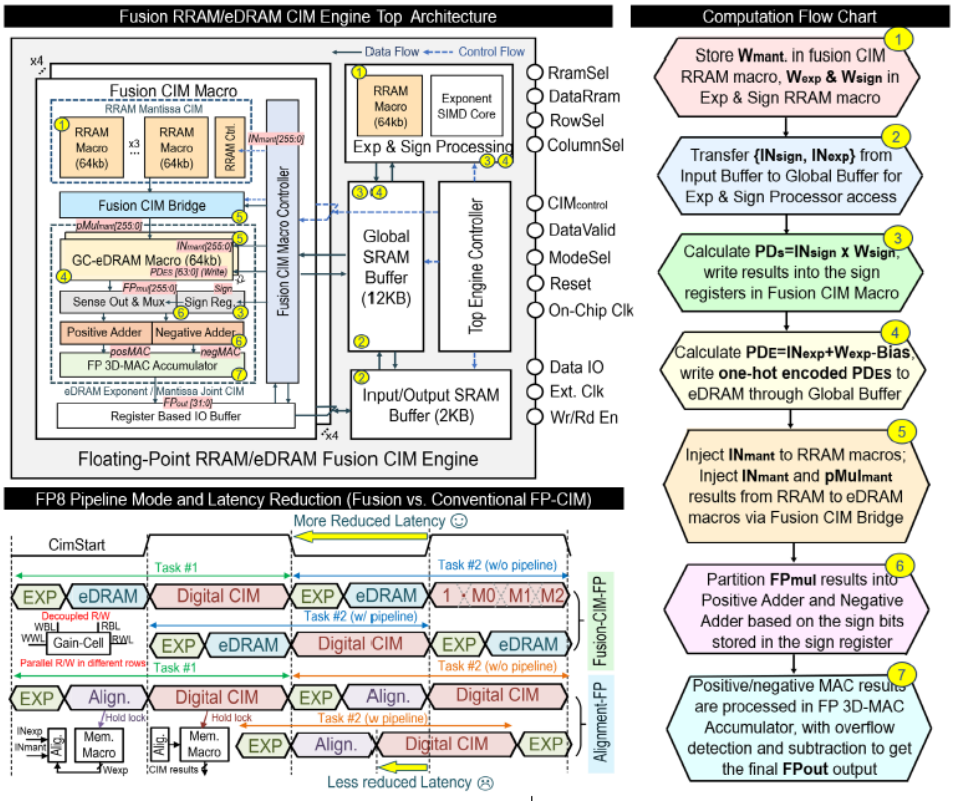
[그림 1] "MIDAS: An Energy-Efficient Microscaling Digital Compute- In-Memory-Based Accelerator with Spatio-Temporal Cyclic Alignment for Generative AI Inference" 전체 아키텍처 구조

#11-2 Southeast University의 28nm SRAM 기반 EMO-CIM 매크로로, 하나의 CIM 매크로에서 CNN·Attention·Depthwise Conv를 모두 지원하는 "multi-operator" 구조를 제안한다. 입력과 정지 데이터의 비트 유사도를 추출하는 LISCU와 Booth 기반 인코더를 결합한 SAC 연산을 통해 불필요한 MAC을 줄이고, RVSA 기반 data flow로 연산 길이가 달라져도 에너지·면적 효율을 유지하도록 설계한 것이 특징이다. 또한 16개의 EMO-CIM bank와 TL-IMN, MOGOB로 구성된 37-kb 유닛 매크로 구조를 통해 CNN/Attention/DW 모드 간 전환을 하드웨어 차원에서 유연하게 지원한다. 입력/정지 데이터 유사도에 기반한 SAC 기법 덕분에 동일 공정의 기존 디지털/하이브리드 CIM 대비 에너지·면적 동시 지표(FoM)를 향상시킨 것으로 보고되며, 0.7 V에서 CNN/Attention 93.5 TOPS/W, DW 39.3 TOPS/W 및 최대 4.97 TOPS/mm² 수준의 효율을 달성한다. Mobile-ViT, YOLO-v10 기반 Edge-AI 벤치마크에서도 실용적인 정확도를 달성하였다.



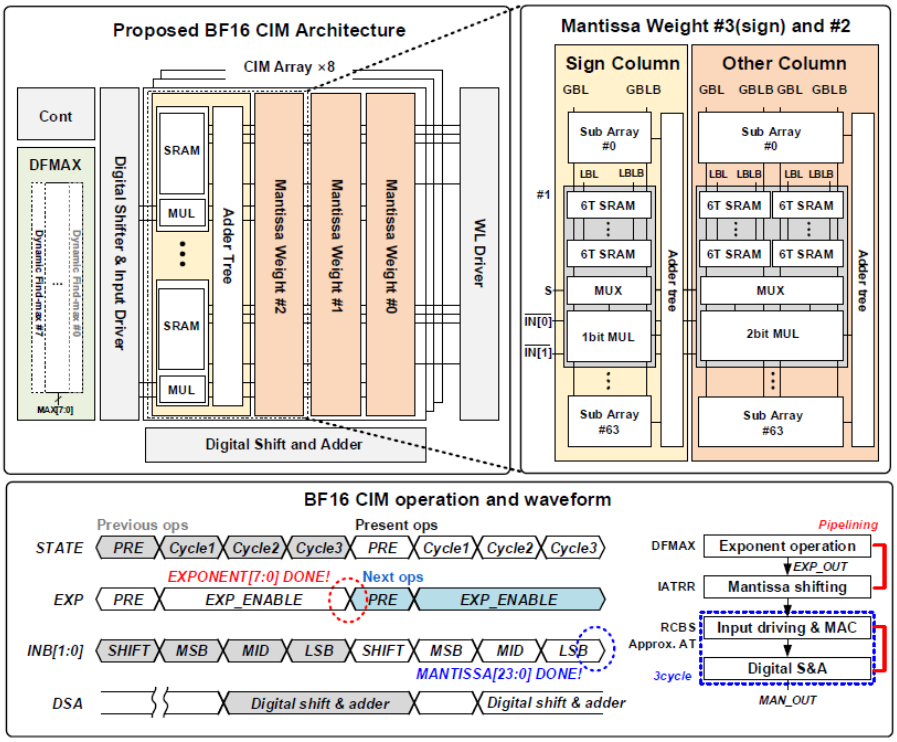
[그림 2] "EMO-CIM: An Input/Stationary-Data Similarity-Aware Computing-In-Memory Macro for Variable Vector-Wise Computation in Edge Multi-Operator AI Acceleration" 전체 아키텍처 구조

#11-3 Purdue University 주도의 40nm RRAM–eDRAM 융합 부동소수점 CIM 엔진(CENTAUR)으로, Mantissa 연산은 RRAM, Exponent 처리는 eDRAM에서 담당하는 "FP 3D-MAC" dataflow를 통해 기존 NVM 기반 FP-CIM의 병목인 지수 정렬(pre-alignment) 오버헤드와 낮은 동작 주파수 문제를 완화한 점이 특징이다. 3T1C gain-cell 기반 shift-vector 3-operand MAC 구조를 도입해 메모리 내에서 곱셈·가산·시프트를 결합적으로 수행함으로써, 외부 디지털 로직으로의 데이터 이동을 줄이고 MAC 파이프라인을 깊게 가져갈 수 있도록 했다. 그 결과, 40nm 공정임에도 불구하고 600 MHz에서 38.5 TFLOPS/W의 에너지 효율을 달성하고, 동일 세대의 기존 SRAM-/RRAM-CIM 대비 종합 FoM이 각각 약 4.8배, 8.5배 향상시켰다. Tiny-ViT 기반 비전 워크로드를 대상으로 baseline 대비 약 1.7% 수준의 정확도 손실만을 보고하여 실용적인 정확도를 확인하였지만, 더 복잡한 dataset으로 검증 범위를 확장한다면, 제안한 엔진의 효용성이 더욱 뚜렷해 질 것으로 보인다.



[그림 3] "CENTAUR: A 38.5-TFLOPS/W 600MHz Floating-Point Digital Compute-In-Memory Engine with 40nm Fusion RRAM-eDRAM Macros Featuring 3D-MAC Operation" 전체 아키텍처 구조

#11-4 성균관대학교에서 발표한 28nm BF16 디지털 CIM 매크로로, BF16 포맷의 넓은 지수 범위와 상대 오차 특성을 활용해 지수·가수 연산을 근사화하면서 에너지와 레이턴시를 동시에 줄인 설계이다. Dynamic Find-Max(DFMAX) 유닛을 통해 64개 입력 쌍에서 최대 지수를 동적 로직과 마스킹으로 효율적으로 탐색하고, Reduced-Cycle Bit-Serial(RCBS) 구조를 도입해 비트-직렬 연산을 기존 4사이클에서 3사이클로 단축함으로써 연산 지연과 에너지 소모를 크게 줄인다. 여기에 Vth drop 문제를 완화한 16T/18T full-adder 기반 approximate adder tree와 input-aware toggling rate reduction(IATRR) 기법을 적용하여, 동일 면적 대비 스위칭 활동을 낮추면서 연산량을 극대화한 것이 특징이다. 그 결과, 0.0576 mm² 매크로에서 0.4 V 기준 77.2 TFLOPS/W, 0.324 TFLOPS/mm²의 높은 에너지·면적 효율을 달성하고, CIFAR-100/ResNet-18 벤치마크에서도 약 0.21% 이내의 정확도 손실만을 보여 BF16 기반 근사 CIM의 실용성을 입증하였다.



[그림 4] "A 28nm 77.2 TFLOPS/W Digital Floating-Point Compute-In-Memory Macro Employing Dynamic Find-Max and Reduced-Cycle Bit-Serial Architecture with Approximation" 전체 아키텍처 구조

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/

A-SSCC 2025 Review

한양대학교 신소재공학과 석박통합과정 송충석

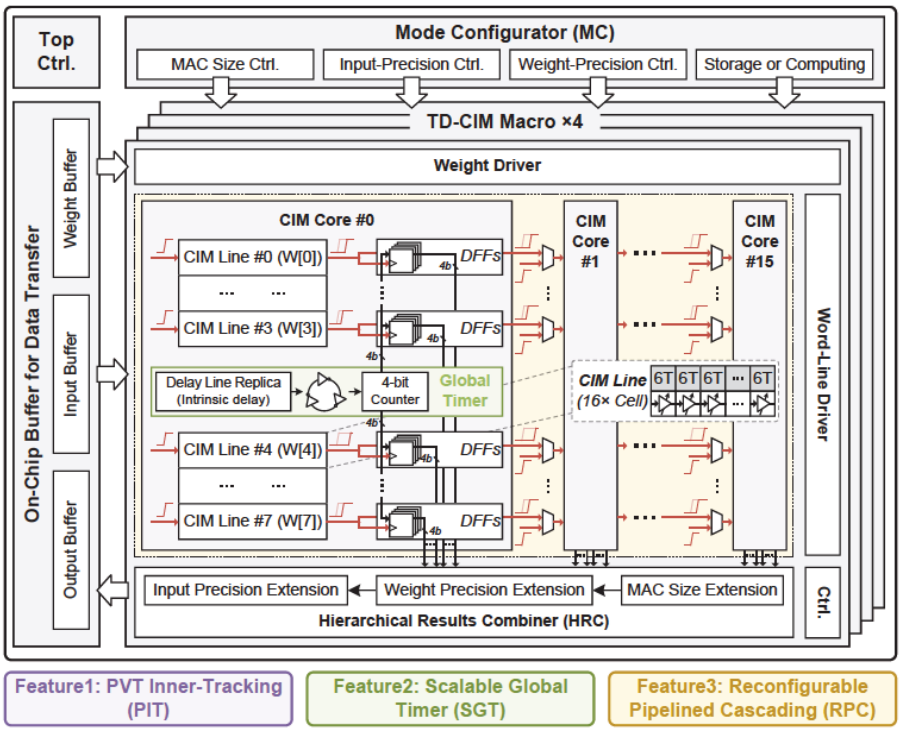
Session 25 Energy Efficient Mixed Signal CIM Circuits

이번 2025 IEEE A-SSCC의 Session 25는 Energy Efficient Mixed Signal CIM Circuits 라는 주제로 총 5편의 논문이 발표되었다. 높은 효율성을 달성하기 위해 아날로그 기반 CIM이 주로 채택되었으며 아날로그 연산의 고질적인 문제인 낮은 정확도, PVT 변동, IR drop, Sensing margin 부족 등을 해결하기 위한 아이디어가 제시되었다. 본 Review에서는 25-1, 25-2, 25-4, 25-5 총 4편을 리뷰하고자 한다.

#25-1 논문은 time-domain (TD)에서 multiply-accumulate (MAC) 연산을 하는 CIM 매크로로 낮은 전압에서도 높은 에너지 효율과 확장성을 제공하는 장점이 있지만, PVT (공정, 전압, 온도) 변화에 매우 민감하고, TD 양자화로 인한 추가회로가 에너지 및 면적에서 오버헤드를 유발하며. 다양한 CNN 모델에서 CIM의 utilization이 떨어져 에너지효율성 측면에서 최대 성능과 평균 성능의 차이가 크다는 문제점이 있었다. 이를 해결하기 위해 본 논문에서는 PVT Inner-tracking 방법을 도입해 CIM cell과 TD Quantizer의 PVT 반응을 동일하게 맞춰 PVT 강건성을 확보하고, 기존 TD Quantizer들의 높은 에너지소모와 큰 면적을 감소시킬 수 있는 SGT회로를 적용해 유연한 누적계산을 가능케 하였다. 마지막으로 RPC 기법을 통해 여러 CIM 라인을 유연하게 병렬연산시켜 속도 저하 없이 큰 누적계산을 처리하였다.

28nm 공정으로 제작된 본 논문의 칩은 4개의 매크로와 매크로당 16개의 코어로 구성되었다. -10% 에서 10% 까지의 전압변화와 -50℃ 에서 100℃ 까지의 온도변화에서도 MAC 에러가 거의 없는 PVT 안정성을 보여주었으며 4b 모드에서 82.82 – 236.5 TOPS/W, 8b 모드에서 20.4 – 58.7 TOPS/W의 매우 높은 에너지 효율을 달성하였다.

CIM 매크로에서 아날로그 연산의 정확도를 높이기 위한 회로를 PVT 강건설계를 중심으로 제안하였으며 정확도 측면에서 유의미한 결과를 남겼다. 그러나 edge device에 적용할 목적으로 depthwise CNN으로 테스트하였지만 이는 다소 제한적인 환경에서만 결과를 도출했으며, 좀 더 다양한 네트워크 환경에서의 정확도 입증이 필요해 보인다.



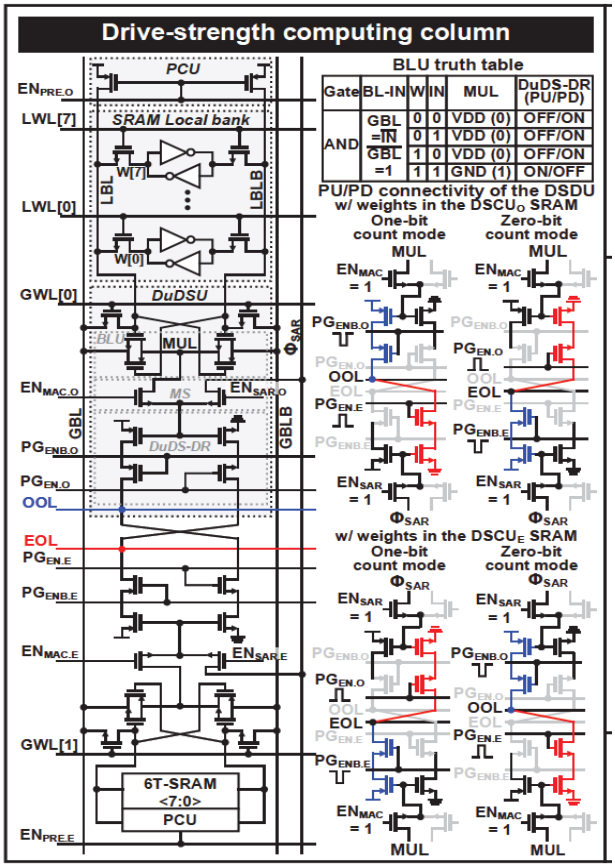
[그림 1] 논문 25.1의 제안한 TD-CIM 의 전체 구조

#25-2 논문은 SRAM 기반 아날로그 방식으로 MAC 연산을 하는 CIM macro이다. 기존 SRAM 기반 CIM 매크로는 vision transformer와 같이 전체 attention 연산을 수행하여 partial sum (PSUM) 오차가 쉽게 증폭되어 정확도 저하를 야기시키며, 특히 비트라인의 제한된 sensing margin으로 인해 PSUM 영역에서 오차가 계속 증가하고, 이는 shift-and-add 단계에서 이러한 에러가 누적 및 증폭되는 문제가 있었다. 이러한 문제를 해결하기 위해 본 논문에서는 Dual Drive Strength 기반 구조를 제안한다. PSUM 값이 커질수록 비트라인에 연결되는 스위치의 수가 많아져 sensing margin이 감소하게 되는데 one-bit/zero-bit count mode를 사용해 '1'의 개수에 따라 PSUM이 작을 때는 '1의 개수', PSUM이 클 때는 '0'의 개수를 읽는 방식으로 전환하여 sensing margin을 유지하게 한다.

제안하는 DuDS-CIM 매크로는 65nm LP CMOS에서 제작되었으며, 측정결과 높은 PSUM 영역에서 RMSE를 2.46에서 0.51로 낮추었으며 ResNet-20 (CIFAR-10) 에서 91.26%, ViT-S/16 (CIFAR-10) 95.57%의 높은 정확도를 달성하여 dense 및 sparse 모델 모두에서 안정적인 PSUM 연산이 가능함을 입증하였다. 에너지 효율은 942 TOPS/W, 면적 효율은 2.13 TOPS/mm^2, 동작 전압은 1.1V, 주파수 50MHz에서 동작한다.

SRAM 기반 CIM의 근본적인 취약점인 sensing margin 부족 문제를 PSUM 분포 기반의 대칭형 구조로 해결했다는 점에서 기술적 기여도가 높다. 특히, ViT 네트워크 환경에서 안정적인 동작을 실측한 것에 의의가 있다고 본다. 그러나 50MHz라는 상당적으로 느린

동작 주파수에서의 결과만 보여주고 있으며 shmoo plot을 게재하지 않았다는 점이 아쉬운 부분으로 남는다.

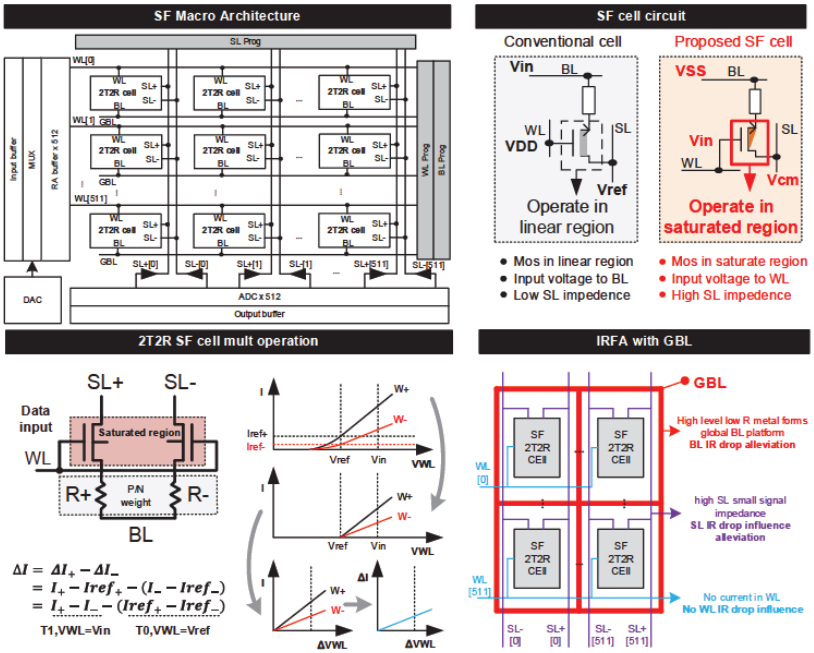


[그림 2] 논문 25.2의 제안한 Drive-strength computing column 회로

#25-4 논문은 RRAM 기반 CIM 매크로로, RRAM은 비휘발성 저장과 아날로그 MAC 연산을 수행할 수 있어 Edge 기반 장치에 차세대 메모리로써 각광받고 있지만 실제 하드웨어에서는 IR drop, weight cell의 선형영역 동작, 아날로그 MAC 연산의 결과전압 오차, 높은 주변부회로의 전력 소모 등의 구조적 문제가 있다. 이러한 문제를 해결하기 위해 본 논문에서는 Source-Follower (SF) 기반 weight cell과 IR-Drop-Freed Array (IRFA) 구조를 새롭게 제안한다. SF cell은 2T2R 구조의 트랜지스터를 포화영역에서 동작시켜 연산 전류를 MAC 연산의 결과전압과 분리시키고, IRFA는 1T1R의 각 라인(BL, SL, WL)에서의 전압 강하를 구조적으로 제거하여 대규모 병렬 MAC 연산에서도 안정적인 연산과 높은 처리량을 확보할 수 있도록 설계한 것이 특징이다.

제안하는 매크로는 28nm 공정으로 1Mb RRAM 어레이로 구성하여 17.0 TOPS/mm^2의 높은 연산 밀도를 달성했으며 RRAM 특유의 아날로그 변동성을 극복하여 실사용 모델에서 의미 있는 정확도를 유지하였다. 다만 실사용 모델에 대해서의 결과는 시뮬레이

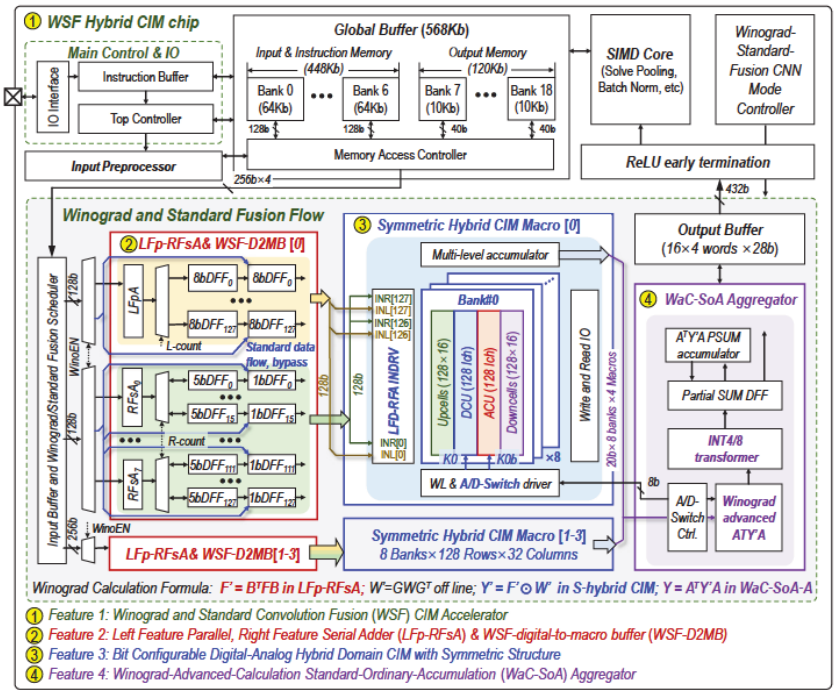
션을 기반으로 진행된 점이 아쉬운 점이다. 그럼에도 불구하고 RRAM 기반 CIM의 가장 근본적인 문제인 IR drop 과 전압 의존성 문제를 해결한 점에서 기술적 완성도가 있었으며, 추가적인 회로에 대한 면적과 파워 오버헤드에 대한 데이터, 병렬성(Parallelism)에 따른 데이터 또한 제시하고 있어 실제 칩 기반 end-to-end 연산 결과가 제시된다면 설득력이 더 강해질 것으로 판단된다.



[그림 3] 논문 25.4에서 제안한 2T2R 구조의 RRAM기반 CIM 매크로와 IRFA.

#25-5 논문은 CNN 연산에서 널리 사용되는 Winograd Convolution 연산 시 발생하는 하드웨어 오버헤드를 줄이는 회로를 제안한다. Winograd Convolution은 연산량을 최대 2.25배 절감할 수 있지만, 실제 하드웨어 적용에서는 추가적인 변환 연산과 메모리 요구량 증가, 정밀도 손실 등 해결해야 과제가 남아있다. 이를 해결하기 위해 본 논문은 Winograd Domain (WD) 와 Standard Domain (SD) 연산을 하나의 하드웨어 안에서 효율적으로 결합시키기 위한 Winograd-Standard Fusion 아키텍처를 제안한다. 이 아키텍처는 CIM 내에서 WD/SD 양쪽 계산을 모두 처리할 수 있도록 구성하였다. 특히 WD transform에 유리한 부분과 SD에서 정밀도를 유지해야 하는 부분을 대칭적으로 그룹화된 CIM 도메인에서 분산 수행하여 정확도 손실을 최소화하면서도 전체 CNN 연산 효율을 극대화하도록 설계하였다.

제안하는 CIM 매크로는 28nm CMOS 공정에서 제작되었으며, 기존 Winograd 전용 가속기 대비 변환 오버헤드를 줄이고, SD 연산보다 높은 연산 밀도를 확보하여 244.45 TOPS/W의 매우 높은 에너지 효율을 달성하였다. 그러나 WD와 SD를 하나의 도메인에서 처리함으로써 발생하는 추가적인 회로의 오버헤드 및 차지하는 면적 비율 등의 추가적인 데이터가 필요해 보인다.



[그림 4] 논문 25.5의 제안한 매크로의 전체 구조

저자정보

송충석 석박통합과정 대학원생

- 소속 : 한양대학교
- 연구분야 : 딥러닝 가속기 설계
- 이메일 : scs940430@hanyang.ac.kr
- 홈페이지 : https://sites.google.com/site/dsjeonglab1/home

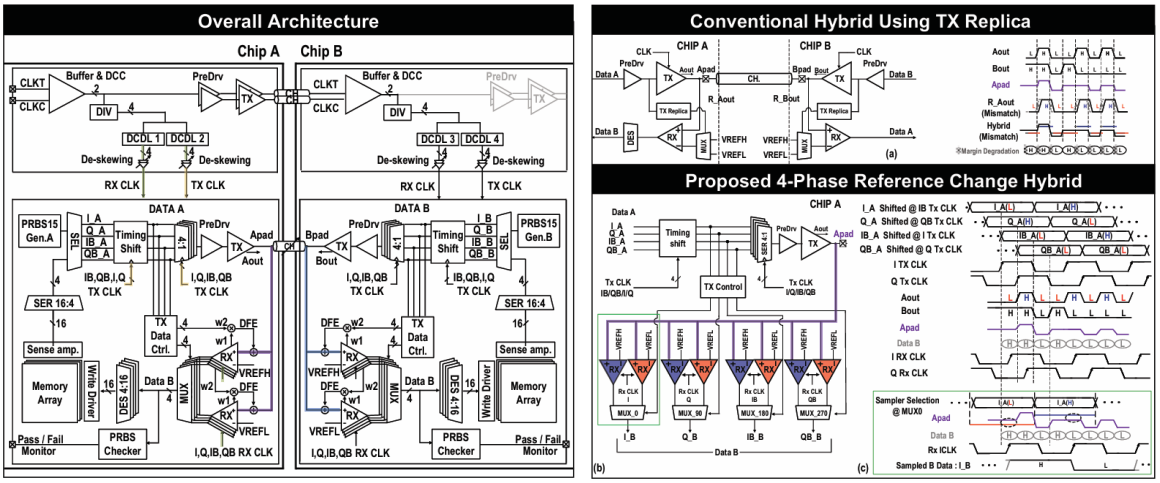
A-SSCC 2025 Review

한국과학기술원 전기및전자공학부 박사과정 윤웅노

Session 29 High Speed Circuit and Interface for Memory

이번 2025 IEEE A-SSCC의 Session 29에서는 High Speed Circuit and Interface for Memory 라는 주제로 총 4편의 논문이 발표되었다. 발표된 논문 중 3편(29.1, 29.2, 29.4)는 기존 채널들의 물리적인 한계를 극복하기 위하여 NRZ 방식을 탈피하거나 개선하는 연구이다. 나머지 한 편(29.3)은 high speed SRAM 동작에 필수적인 ECC 회로에 대한 논문으로 고속 동작으로 생기는 에러에도 데이터의 신뢰성을 보장하기 위한 연구를 보여주었다.

#29-1 해당 논문은 좁은 면적에서 고속 통신을 요구하는 차세대 HBM4 인터페이스를 위한 기술로 주목받는 SBD (Simultaneous Bidirectional) signaling에서 생기는 문제들을 해결하고자 하였다. SBD signaling으로 핀 당 통신 속도를 2배로 높였으나 channel delay로 인하여 신호 파형이 복잡하여 signal eye가 좁아지는 문제가 존재하였다. 이를 해결하기 위해 첫째, timing alignment 기법을 적용하였다. DCDL을 이용하여 channel delay를 타겟 주파수의 1/2 UI 배수가 되도록 조절함으로써, superimposed signal의 파형을 단순화시켜 신호 복원 난이도를 낮췄다.



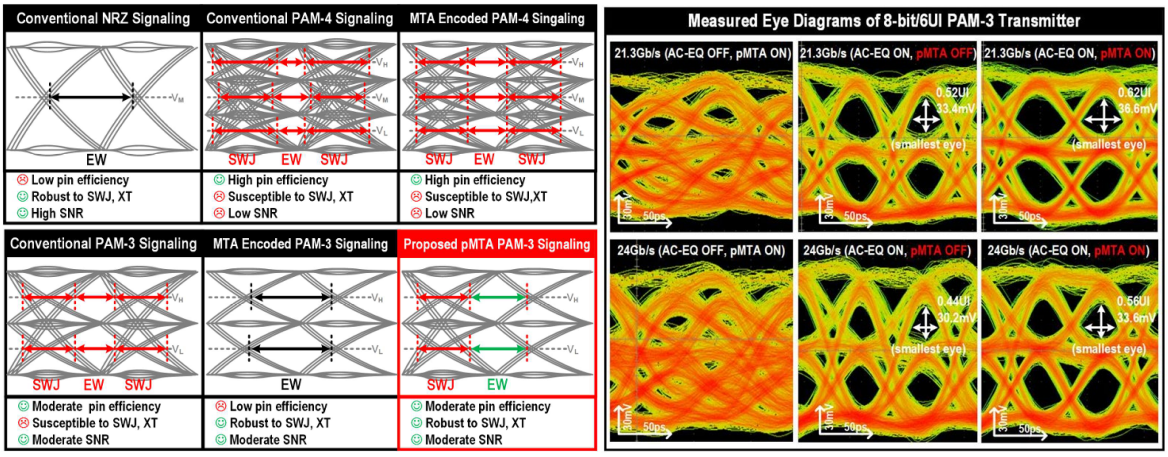
[그림 1] 제안된 SBD signaling 방식의 전체 아키텍처 (좌), 기존 TX replica 방식과 제안된 4-Phase 기준 방식의 비교 (우)

둘째, 기존 TX Replica 방식의 mismatch 한계를 극복하기 위하여 4-phase hybrid scheme을 제안하였다. 이는 replica 회로 없이, 이미 알고 있는 4-UI 단위의 송신 데이터에 따라

서로 다른 기준 전압을 가진 수신기를 선택하는 방식이다. 이를 통하여 타이밍 마진을 대폭 개선하였다. 마지막으로, dual equalization과 Ron 최적화를 적용하였다. IR drop의 영향을 줄이기 위해 TX driver 저항을 45 ohm으로 최적화하고, 이로 인해 약해진 신호 구동력과 ISI 이슈는 수신 신호용 DFE와 송신 신호용 XOR 기반 equalizer를 결합하여 보상하였다. 이러한 방식으로 18.4Gb/s/pin의 속도를 달성할 수 있었다.

#29-4 차세대 메모리 인터페이스는 높은 데이터 전송율에 더불어 에너지 효율성을 극대화하는 것이 핵심 과제이다. 이를 위해서 NRZ 대비 높은 대역폭 효율을 지니는 다중 레벨 signaling을 위한 PAM-4, PAM-3 구조가 도입되었으나 그 한계가 존재하였다. PAM-4는 eye height가 NRZ의 1/3로 줄어들어 SNR이 낮고, switching jitter (SWJ)와 crosstalk(XT)에 취약해 signal integrity (SI) 성능 하락 이슈가 있다. PAM-3는 PAM-4에 비해서는 좋은 SNR을 지녔으나 여전히 SI 하락 문제가 남아있었다. 이에 더불어 기존 PAM-3 방식은 이론적으로 NRZ 대비 150%의 속도 향상이 가능하지만, DDR5나 HBM3와 같은 실제 메모리 시스템의 burst length가 2의 지수 단위로 동작하기 때문에 데이터 매핑 시 실질적인 효율은 133%로 제한되는 구조적 비효율성이 존재하였다.

본 논문은 메모리 인터페이스에서 다중 레벨 송신기를 구현함에 있어서 생기는 이슈들을 저전력으로 해결하고자 partial Maximum transition avoidance (pMTA) 방식을 제안하였다. 기존의 MTA 방식과는 다르게, 신호가 2-UI 내에서의 최대로 변화하는 경우를 제거하도록 LUT를 통해 인코딩한다. 잘 최적화된 encoding mapping을 통하여 전반적인 전류 소모를 낮추면서도, 8-bits/6-UI가 가능하도록 구현하였다. 이는 곧 data rate per pin은 유지하며 switching jitter가 발생하더라도 eye width가 기존에 비하여 13.8%P 향상되는 결과로 이어졌다.



[그림 2] NRZ, PAM-4, PAM-3 방식과 TMA, pMTA에 따른 eye diagram의 비교 (좌), Signaling 방식에 따라 측정된 eye diagram (우)

또한, 해당 논문은 0.54pJ/bit의 최고 수준의 에너지 효율과 0.051pJ/bit/dB의 FoM을 달성하였는데 이는 0.3V의 무척 낮은 VDDQ를 이용하였기 때문으로 사료된다. 특히, 낮은 VDDQ에서 드라이빙을 위하여 N-over-N voltage-mode driver를 사용하여 low voltage swing terminated logic (LVSTL)을 구현한 것도 주목할만한 부분이다.

저자정보

윤응노 박사과정 대학원생

- 소속 : 한국과학기술원 전기및전자공학부
- 연구분야 : Sensor Interface ICs, Frequency Generation ICs
- 이메일 : voogi3925@kaist.ac.kr
- 홈페이지 : https://impact.kaist.ac.kr/