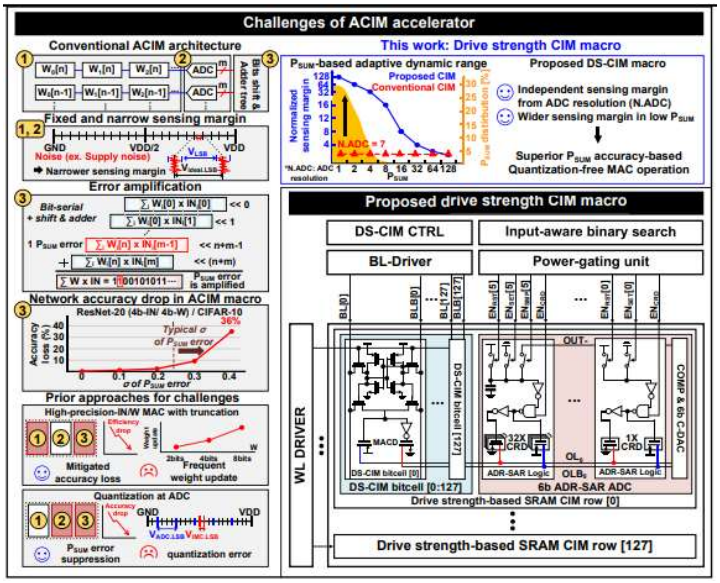
A-SSCC 2024 Review

KAIST 전기 및 전자공학부 석사과정 윤지원

Session 8 SRAM-Based Computing-in-Memory

이번 2024 IEEE ASSCC의 Session 3은 SRAM-Based Computation-in-Memory라는 주제로 총 5편의 논문이 발표되었다. 이 세션에서는 메모리 내에서 데이터를 직접 처리함으로써 데이터 이동과 연산에 소모되는 에너지와 시간을 크게 줄일 수 있도록 하는 혁신적인 아키텍처를 제안하는 것에 중점을 두었다.

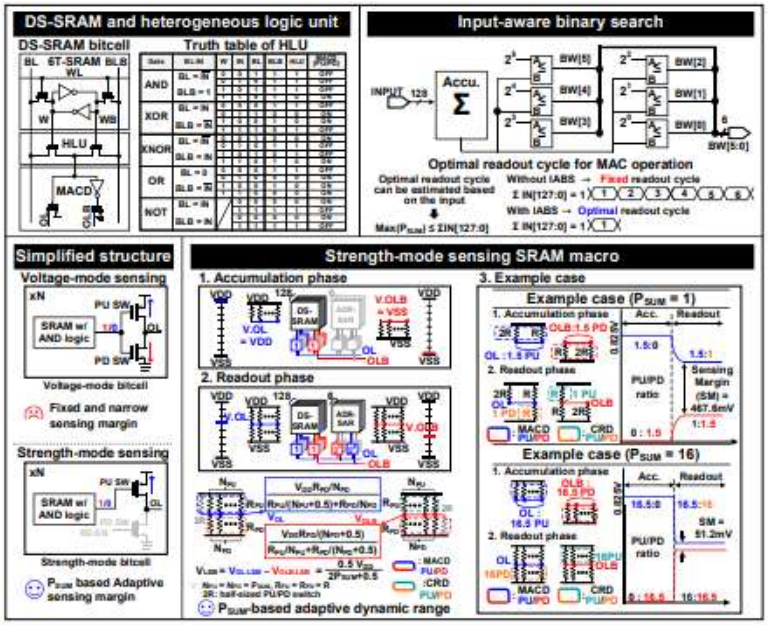
#8-3 본 논문은 DGIST, 고려대학교에서 발표한 논문으로, Computation In Memory (CIM)에서 발생하는 Partial Sum(PSUM) 오류를 해결하기 위해, 고정밀 입력/가중치 사용 및 MAC 출력을 절단하거나, 저해상도 ADC를 활용하여 Partial Sum을 양자화하는 기존의 해결 방법대신 Drive Strength 기반 SRAM CIM 접근법을 제안함으로써 정확도 손실 문제점을 해결하고자 하였다. 제안된 DS-SRAM CIM macro는, 128x128b DS-CIM bitcell array, Input-Aware Binary Search(IABS), Power Gating Unit (PGU), 128 row-wise ADR-SAR ADCs, 그리고 DS-CIM controller로 이루어져 있다.



[그림 1] 아날로그 CIM macro 설계의 난점 및 제안된 DS-SRAM CIM

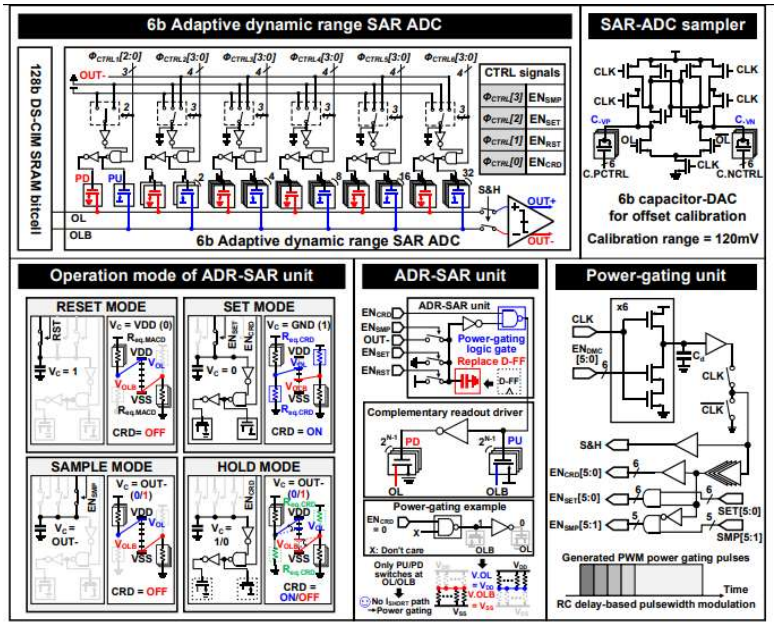
비트셀은 6T SRAM**,** Heterogeneous Logic Unit (HLU), MAC 드라이버(MACD)로 구성되며, PSUM 값은 연결된 PU 스위치 수로 표현된다. Strength 기반 MAC 연산은 저 PSUM 영

역에서 높은 감지 여유(SM)를 제공하며, 연산은 축적 단계와 읽기 단계로 수행된다. 축적 단계에서 비트별 AND 연산이 수행되고, 읽기 단계에서는 6비트 ADR-SAR ADC가 CRD와 COMP를 통해 이진 탐색으로 PSUM 출력을 읽는다. IABS는 입력 합계를 기반으로 ADC 변환 주기를 결정하고, 추가적인 스위치를 통해 충분한 SM을 확보한다. PSUM 값이 입력 합계를 초과할 경우 IABS를 비활성화해 읽기를 수행한다



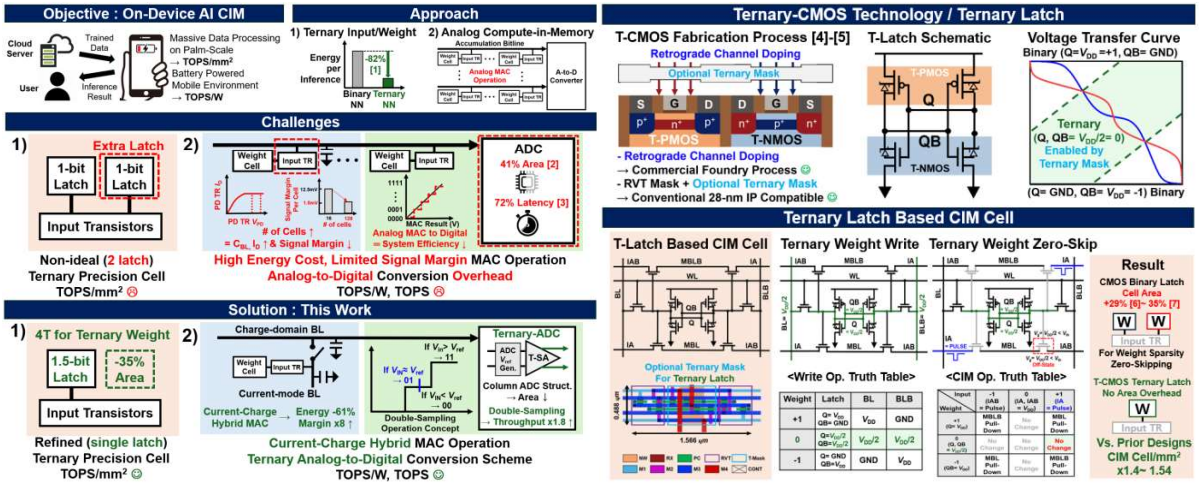
[그림 2] DS-SRAM bitcell의 Truth table과 IABS, 그리고 각 mode에서의 Sensing margin 분석

ADR-SAR ADC는 Partial Sum 출력을 읽기 위해 설계된 6비트 ADC로, 샘플링 커패시터와 NAND, 제어 스위치로 구성되며 Reset, Set, Sample, Hold 네 가지 모드로 작동한다. Reset 모드에서는 샘플링 커패시터를 초기화하고, Set 모드에서 CRD가 OL/OLB에 연결되며, Sample 모드에서는 COMP 출력을 커패시터에 저장하고 Hold 모드에서 데이터를 유지한다. IABS는 최적의 ADC 변환 주기를 예측하고 PGU는 이에 맞춰 전력 게이팅 펄스를 생성해 전력 소모를 절감한다. 이를 통해 기존 디지털 SAR ADC 대비 면적을 35% 줄이고, 전력 소모를 최대 40% 절감한다. 정밀도는 샘플링 커패시터의 불균형 보정과 하이브리드 SAR 구조로 달성되며, 변환 정확도는 98.5%에 이른다. 또한 노이즈 저감 메커니즘과 고정밀 COMP를 통해 SNR이 기존 대비 25% 이상 향상된다. 결과적으로 ADR-SAR ADC는 낮은 전력 소모와 높은 정확도를 유지해 메모리 기반 연산에서 PSUM의 오류 없는 읽기와 에너지 효율적인 성능을 제공한다.



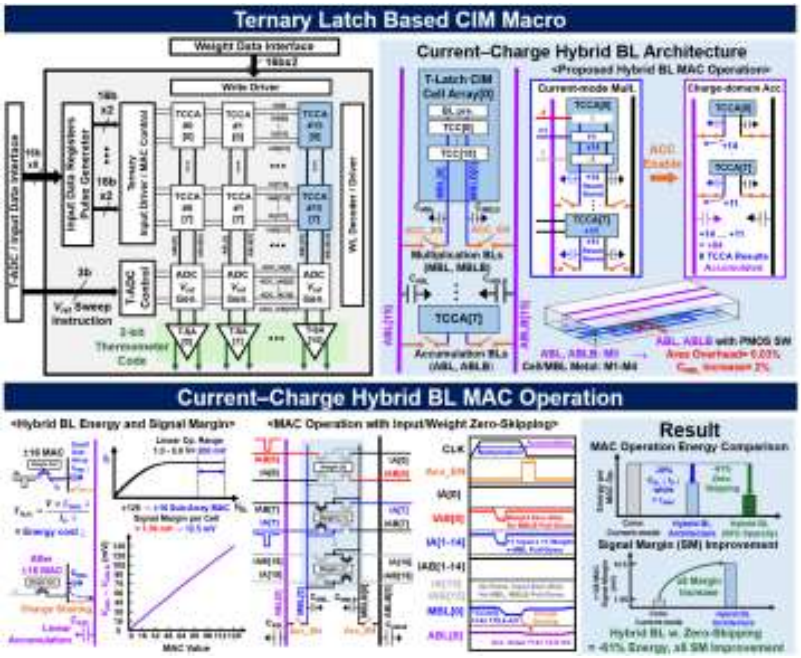
[그림 3] 제안된 ADR-SAR ADC, ADR-SAR 유닛의 동작 모드 및 Power Gating 유닛

#8-5 본 논문은 UNIST 와 삼성전자가 공동 발표한 연구로, 28nm T-CMOS 기술을 기반으로 한 단일 latch 삼진 셀과 에너지 효율 및 신호 마진을 개선한 Double-Sampling Ternary ADC 를 제안한다. 리소스가 제한된 온디바이스 AI 환경에서는 면적 효율과 전력 효율이 핵심 요소로, 네트워크 양자화와 삼진 입력/가중치 (-1, 0, +1)를 사용하면 에너지를 82% 절감할 수 있다. 그러나 삼진 가중치를 지원하기 위한 추가적인 Latch 와 아날로그 CIM 에서 MAC 연산이 증가함에 따라 효율성 저하 문제가 발생한다. 이를 해결하기 위해 T-CMOS 기술을 활용한 단일 latch 삼진 셀 (T-latch)이 제안된다. 이 설계는 29%~35%의 면적 오버헤드를 피하고, CIM 셀 밀도를 1.4~1.54 배 향상시켜 면적 효율성과 전력 효율을 크게 개선함으로써, 제한된 리소스에서 높은 성능을 유지하는 데 중요한 역할을 한다.



[그림 1,2] 본 연구의 동기 및 요약 및 Tenary-CMOS 기술과 삼진 래치 기반 CIM 셀

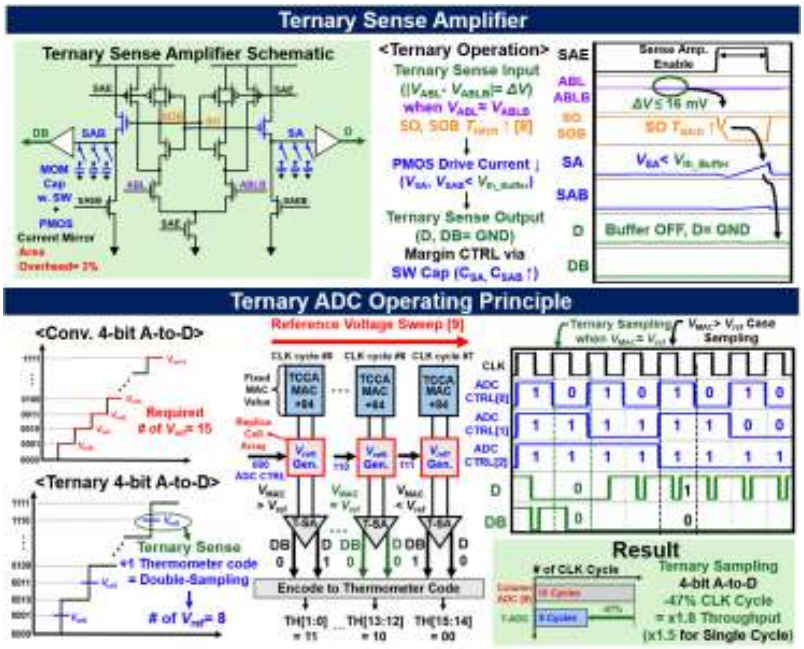
제안된 삼진 래치 기반 CIM 매크로는 128 개의 삼진 CIM 셀 배열(TCCA)로 구성되며, 각 열은 8 개의 TCCA 와 이중 샘플링 삼진 ADC 를 포함한다. 입력 데이터는 레지스터에 저장되며 값이 "+1" 또는 "-1"일 때 펄스를 생성한다. 상단과 우측에는 가중치 업데이트를 위한 쓰기 주변 회로(드라이버 및 디코더)가 위치하고, 좌측 하단에는 삼진 ADC 제어부와 3 비트 제어 코드 입력부가 배치된다. 128 개의 TCCA 는 전류-전하 하이브리드 비트라인(BL) 구조를 통해 MAC 연산을 수행한다. 기존 단일 전류 모드 비트라인 대신, 8 개의 전류 모드 곱셈 비트라인(MBL)과 1 개의 전하 도메인 누적 비트라인(ABL)으로 분리해 에너지 효율과 신호 마진 문제를 해결했다. 각 8 개의 TCCA 는 전류 모드 곱셈을 수행한 후 ±16 의 MAC 결과를 8 개의 MBL 에 저장하고, 이후 PMOS 스위치를 통해 ABL 에 ±128 의 MAC 결과를 누적한다. 이 구조는 면적 오버헤드 없이 신호 마진을 8 배 향상시키며, 전력 소모를 줄인다. 또한, 입력과 가중치가 0 일 때 동적 전력을 소모하지 않는 zero-skipping 기능을 통해 에너지 절감을 극대화한다.



[그림 3] 하이브리드 BL 구조를 갖는 삼진 래치 기반 CIM macro

삼진 센스 증폭기(T-SA)는 기존 센스 증폭기에 삼진 감지($V_{MAC} \approx V_{REF}$) 기능을 추가해 이중 샘플링을 수행한다. 이를 통해 기준 전압(V_{REF}) 단계를 기존 15 에서 8 로 줄여 처리량을 기존 방식 대비 1.5 배, 1.8 배 증가시켰다. T-SA 는 컬럼 피치에 맞게 설계되어 면적 오버헤드 없이 각 열에 통합되며, 효율적인 ADC 변환을 지원한다. 기준 전압은 TCCA 하단에 위치한 Replica T-Latch CIM 셀 배열을 통해 생성되고, 8 개의 연속적인 클럭 사이클 동안 이중 샘플링으로 2 개의 Thermometer 코드를 생성한다. 측정 결과, 제안된 CIM 매크로는 단일 클럭 사이클에서 하이브리드 MAC 연산과 삼진 A-to-D 변환을 성공적으로 수행했으며, 167 MHz 클럭 주파수에서 높은 성능을 보였다. 면적당

143 TOPS/mm²와 전력당 892 TOPS/W 를 달성했고, 비트당 정규화 시 323 TOPS/mm²/b 와 2,007 TOPS/W/b 성능을 기록했다. 이는 삼진 래치 기반 CIM 셀, 이중 샘플링 삼진 ADC, 하이브리드 비트라인 구조, 그리고 zero-skipping MAC 연산 덕분이다. 결론적으로, 제안된 CIM 매크로는 온 디바이스 AI 환경에서 높은 에너지 효율과 신호 마진을 달성하며, 면적과 전력 소모 측면에서도 뛰어난 성능을 보여준다.



[그림 4] 삼진 센스 증폭기 기반 이중 샘플링 삼진 ADC

저자정보

윤지원 석사과정 대학원생

- 소속 : 한국과학기술원 (KAIST)
- 연구분야 : 디지털 회로 설계
- 이메일 : jwyoon@kaist.ac.kr
- 홈페이지 : https://idec.or.k

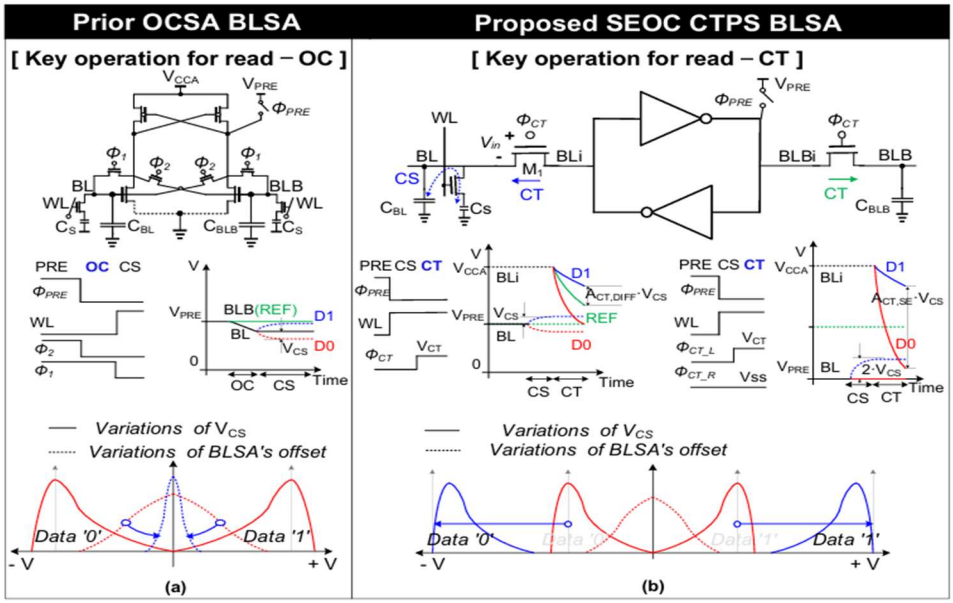
A-SSCC 2024 Review

고려대학교 전기전자공학과 박사과정 한창우

Session 28: Advanced Memory Technology

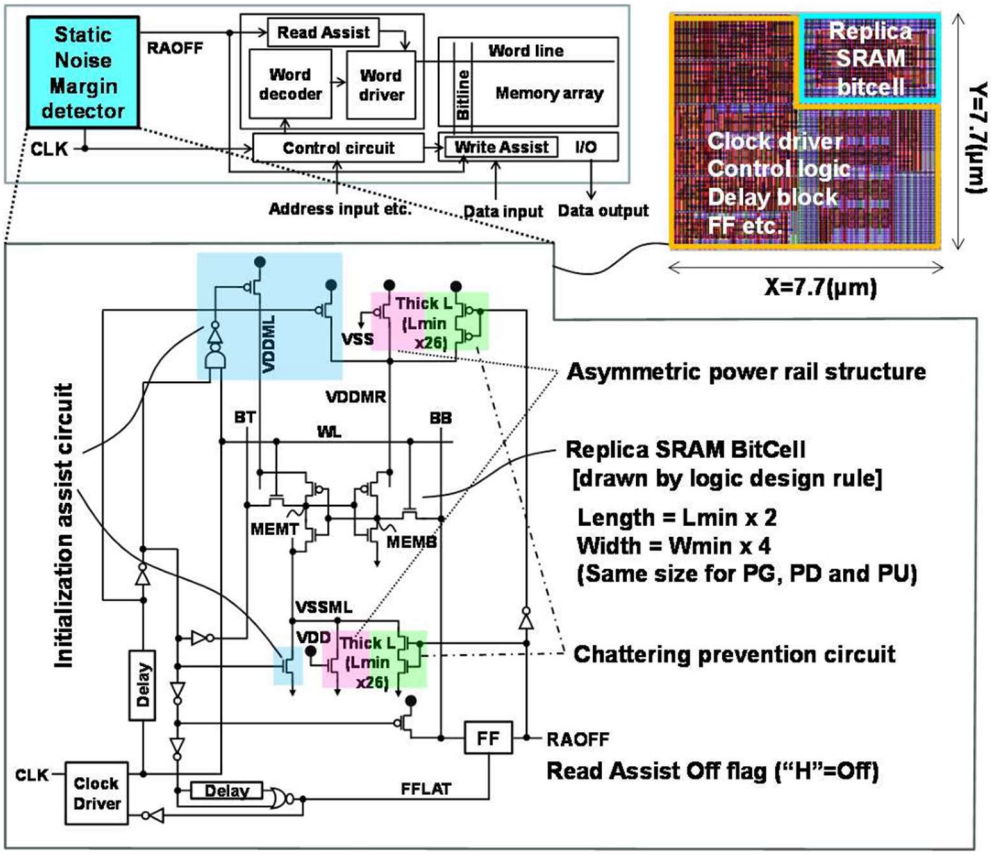
이번 A-SSCC 2024의 Session 28에서는 차세대 메모리 기술을 중심으로 총 5편의 논문이 발표되었다. 저전력, 고속 동작, 안정성을 목표로 한 다양한 메모리 기술이 소개되었으며, 특히 DRAM, SRAM과 같은 메모리 솔루션의 성능 개선과 전력 최적화 방안이 집중적으로 논의되었다. Session 28 안에서도, DRAM과 SRAM의 주요 기술을 다루는 3편의 논문을 살펴보고자 한다.

#28-1 본 논문은 삼성전자에서 발표한 논문으로, Sub-1V DRAM에서 bit-line sense amplifier (BLSA)의 성능을 개선하기 위한 single-ended offset compensation BLSA (SEOC BLSA)를 제안한다. DRAM의 에너지 효율을 위한 공급 전압이 낮아지면서 기존 cross-couple latch 기반의 BLSA는 저장 용량 감소와 트랜지스터의 임계 전압 변동성 증가로 인한 신뢰성 문제를 겪었지만, SEOC BLSA의 도입으로 앞선 문제를 해결하였다. SEOC BLSA는 ground precharge (GND PRE)와 charge-transfer (CT) topology를 도입하여 데이터 sensitivity를 대폭 향상시켰으며, PVT (프로세스, 전압, 온도) 변화에도 안정적인 동작을 보였다. 특히 데이터가 모두 0일 때 전류 소비를 최대 63% 절감했고, 기존 방식보다 최소 공급 전압을 0.85V까지 낮추어 저전력 DRAM 운영에 적합함을 입증하였다.



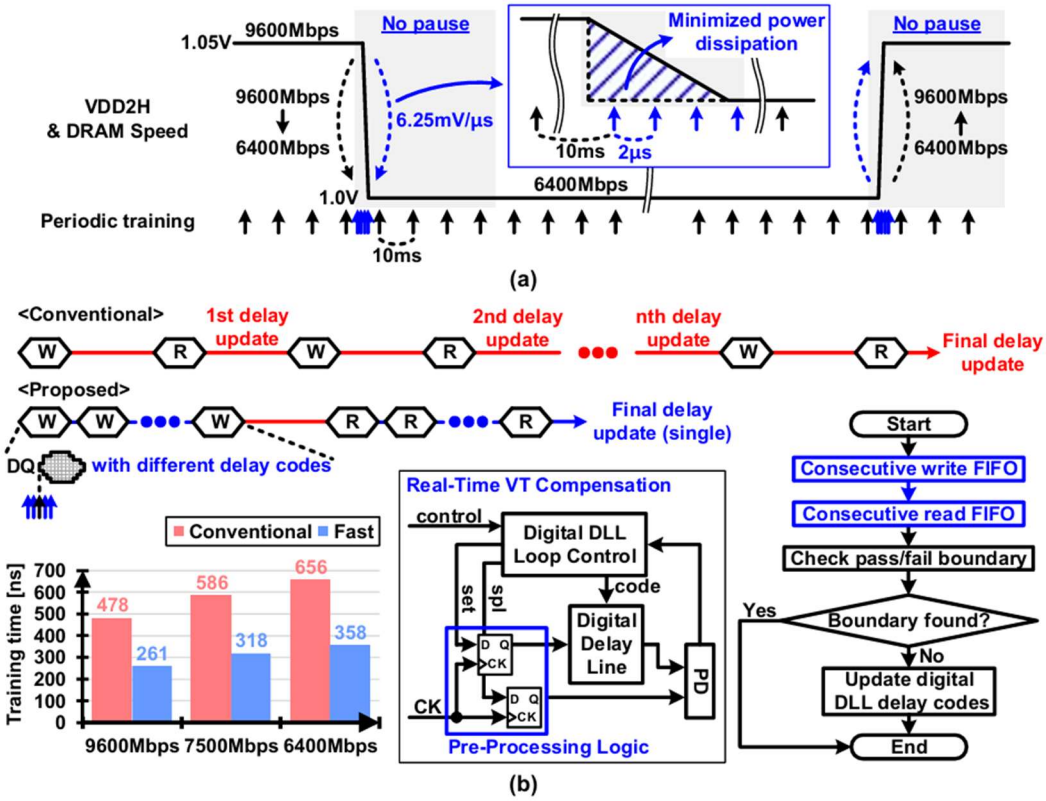
[그림 1] 기존 BLSA 구조 및 제안된 SEOC CTPS BLSA 구조 비교

#28-3 본 논문은 르네사스 일렉트로닉스에서 발표한 논문으로, High Density (HD) SRAM의 static noise margin (SNM)을 감지하는 센서를 활용한 영역 효율적인 어시스트 기술을 제안한다. SRAM은 PVT (프로세스, 전압, 온도) 조건에 아주 민감하며, 특히 HD 셀에서는 SRAM의 SNM이 감소되는 등 안정성 저하 문제가 심각한 상황에 직면해왔다. 이를 보완하기 위해 word line under drive (WLUD)와 같은 어시스트 기술을 사용했지만, 이는 전력 소비가 크고 액세스 시간을 느리게 만드는 단점이 존재했다. 또한, 기존 PVT 센서를 활용한 방식은 큰 면적을 차지하고 초기화 시간이 길어 실용성이 떨어지는 단점까지 존재했다. 본 논문에서는 앞선 문제를 모두 해결할 수 있는, 최소 안정성 셀을 복제한 구조의 SNM Detector를 제안하였고, 이를 통해 불필요한 어시스트를 제거해 SRAM의 액세스 시간을 25% 개선하고, 전력 소모를 읽기 측면에서 32.8%, 쓰기 측면에서 24.4% 감소시키는 데 성공했다. 또한, 0.015%의 무시할만한 작은 면적 오버헤드를 가졌으며, 온도 변화에 따른 어시스트 활성화/비활성화를 조절하는 효율성 또한 제공하였다. 결과적으로, 제안된 SRAM의 SNM Detector는 HD SRAM에서 공정 및 온도 변화에 따른 불안정성을 해결하고, 전력 효율과 성능을 동시에 개선할 수 있는 실용적인 솔루션을 제시할 수 있다.



[그림 2] 제안된 SNM Detector 회로도 및 레이아웃

#28-4 본 논문은 삼성전자에서 발표한 논문으로, LPDDR5X PHY 설계에서 adaptive driver strength control (ADSC)와 빠른 주기적 트레이닝(fast periodic training)을 통해 완전한 dynamic voltage and frequency scaling (DVFS)을 구현하여 메모리 성능을 유지하면서도 전력 소비를 줄이는 방법을 제안한다. 제안된 설계는 DRAM 전원 전압 (VDD2H)을 1.05V에서 1.0V로 낮추는 등 전력 절감을 달성하면서도, 전압 변동으로 인해 발생할 수 있는 쓰기 valid window margin (VWM) 저하를 극복할 수 있다. 이를 위해 빠른 주기적 트레이닝 방식을 적용하여 DVFS 전환 중에도 메모리 접근 차단(blackout) 없이 안정적인 동작을 보장한다. ADSC는 데이터 속도에 따라 드라이버 강도를 조정하여 전력 효율을 최적화하며, 고속 데이터 전송에서도 안정적인 신호 무결성을 제공한다. 실험 결과, DRAM 전력 소비는 기존 DVFSC 방식 대비 14% 이상 감소했고, 6400Mbps에서 쓰기 VWM이 26.4%, 2700Mbps에서 18.9% 개선되었다. 또한 PHY와 I/O 전력 소비가 각각 15.18%, 23.31% 감소하며, 읽기/쓰기 성능도 향상되었다. 이 기술은 고속 데이터 처리와 저전력을 모두 만족시키는 차세대 LPDDR5X DRAM 설계의 중요한 요소로 평가된다.



[그림 3] LPDDR5X PHY 설계에서 ADSC와 빠른 주기적 트레이닝을 통한 DVFS 구현

저자정보

한창우 박사과정 대학원생

- 소속 : 고려대학교 전기전자공학과
- 연구분야 : 차세대 반도체 소자 및 회로
- 이메일 : cwoo0105@naver.com
- 홈페이지 : https://sites.google.com/view/kudclab

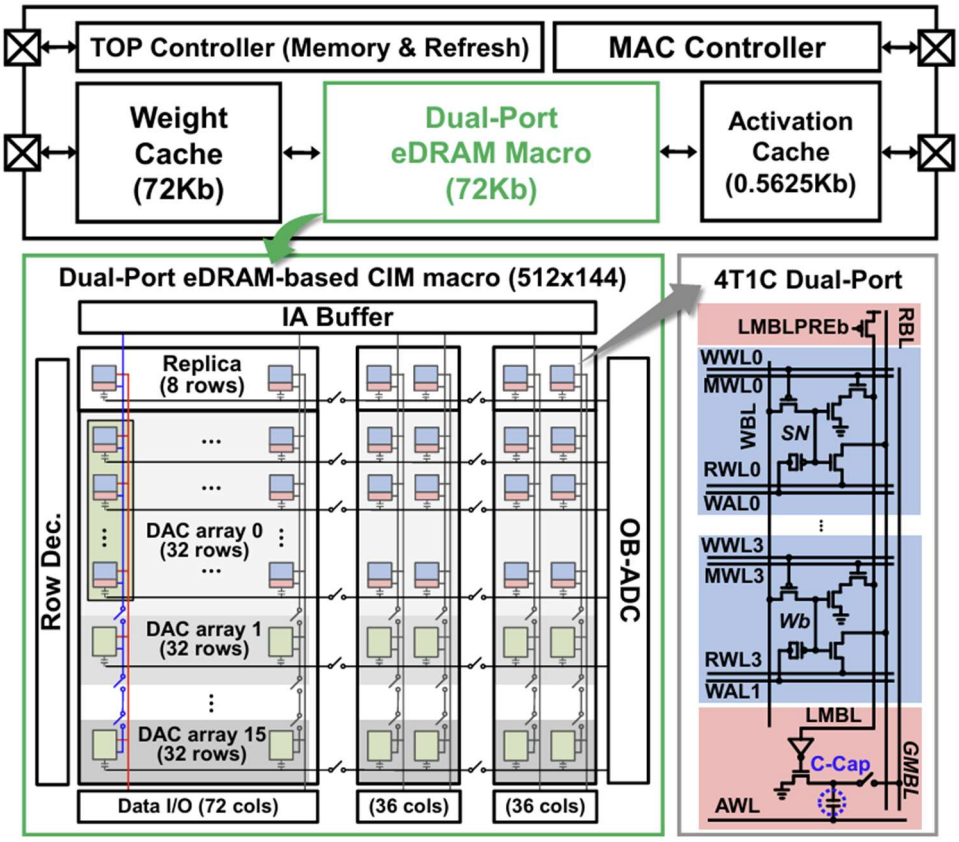
2024 A-SSCC Review

KAIST 전기및전자공학부 박사과정 엄소연

Session 25 High Density Computing-In-Memory

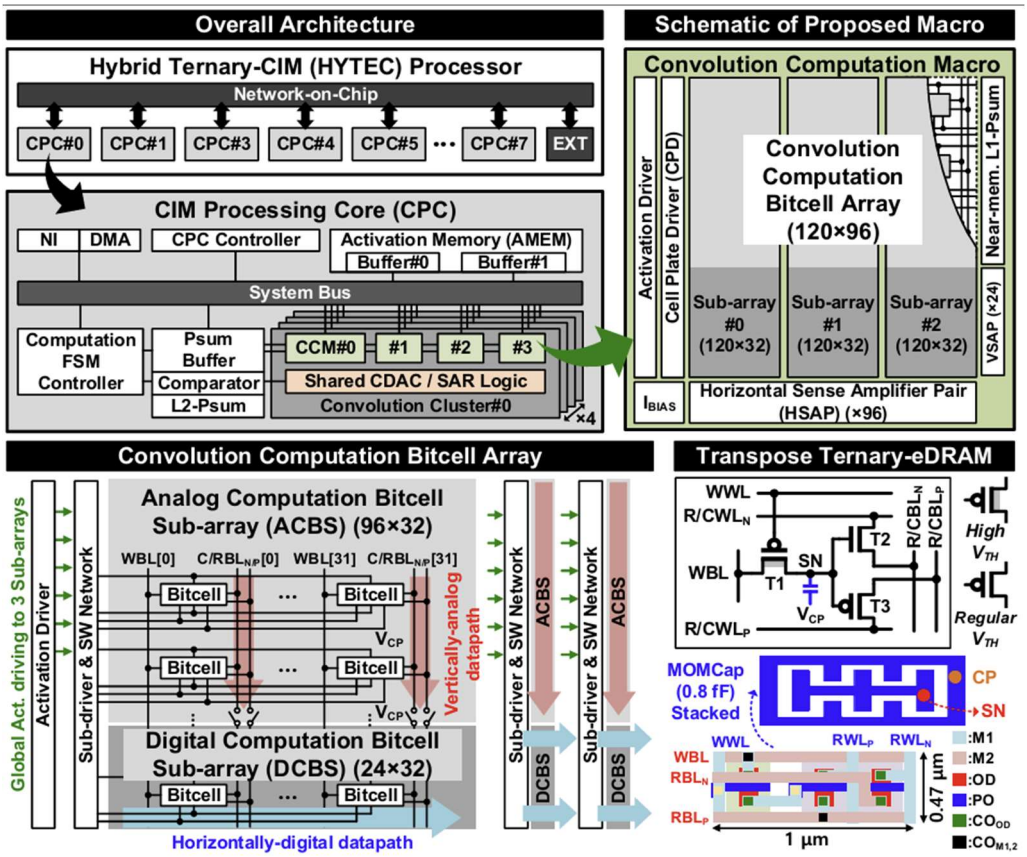
이번 2024 IEEE ASSCC의 Session 25는 High Density Computing-In-Memory 라는 주제로 총 4편의 논문이 발표되었다. 이 세션에서는 eDRAM, ROM, Flash 기반 Compute-in-Memory (CIM) 가속기와 새로운 데이터 변환 기법이 소개되었으며, FPGA를 활용한 다양한 AI 응용 사례를 중심으로 성능과 효율성을 높이기 위한 기술적 접근이 다뤄졌다. 본 후기를 통해 3개의 논문에 대해 상세히 살펴보고자 한다.

#25-1은 연세대학교에서 발표된 Dual-Port eDRAM Compute-in-Memory (CIM) 가속기로, eDRAM 기반 메모리와 연산을 통합하여 성능과 전력 효율을 극대화한 새로운 아키텍처를 제안한다. 4T1C 듀얼 포트 구조를 채택하여 데이터 리프레시와 CIM 연산을 병렬로 수행함으로써 기존의 eDRAM CIM 구조에서 발생하는 성능 병목과 리프레시 오버헤드 문제를 효과적으로 해결했다. Adaptive Refresh Tracking (ART)는 PVT(공정, 전압, 온도) 조건에 따라 최적의 리프레시 주기를 동적으로 조정하여 불필요한 리프레시를 줄이는 데 중점을 두었다. 이를 통해 리프레시 전력을 최대 60% 절감하고 시스템 전반의 에너지 효율을 높이는 데 성공했다. 또한, Data Conversion Reduction Scheme (DCRS)을 도입하여 DAC와 ADC의 면적 및 전력 소모를 최소화하였다. DAC와 ADC는 기존 CIM 시스템에서 높은 에너지 소비와 신호 마진 저하를 초래했으나, DCRS는 데이터 변환 프로세스를 간소화하고 하드웨어 자원을 줄이는 방식으로 이러한 문제를 해결하였다. 신호 마진은 8.9배 향상되었고, 메모리 효율은 이전 설계 대비 90.8% 증가했다. 이 가속기의 성능은 CIFAR-10 데이터셋과 VGG-16 모델로 평가되었다. Dual-Port eDRAM CIM은 평균적으로 기존 설계 대비 27.5% 더 높은 처리량을 제공하며, 전체 전력 오버헤드는 3.16%, 면적 오버헤드는 4.2% 증가에 불과했다.



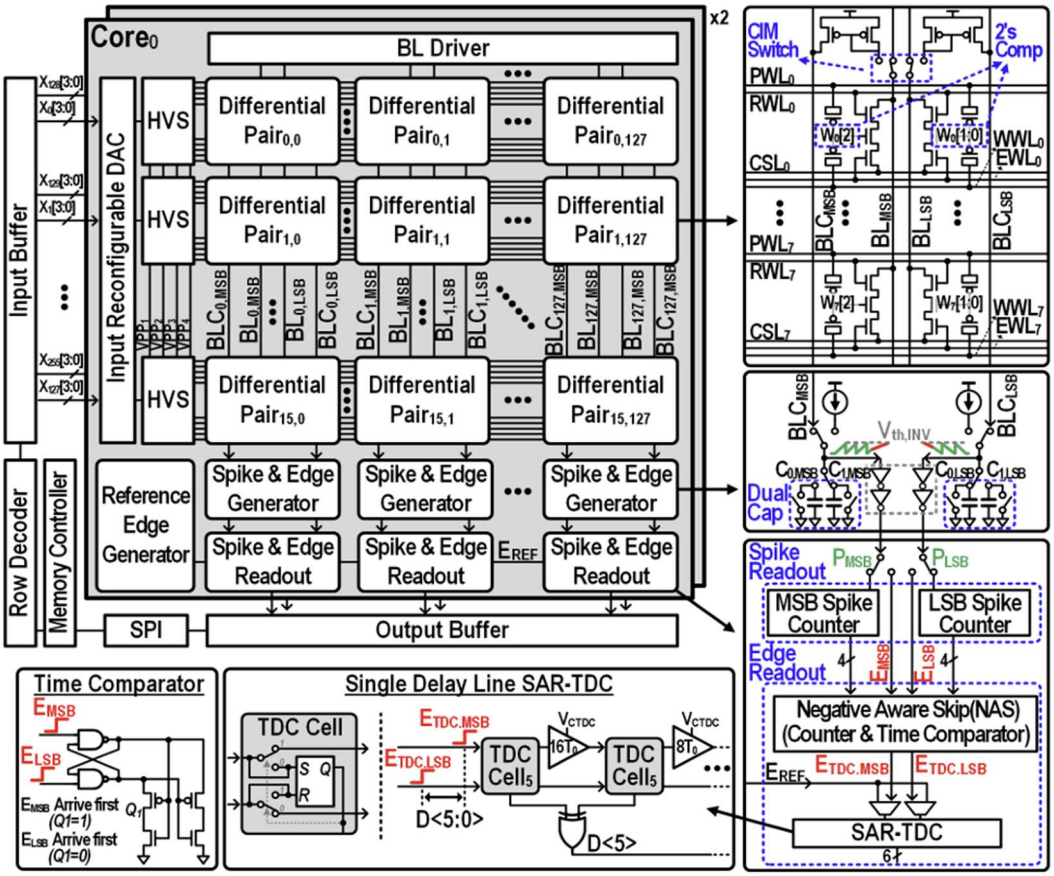
[그림 1] #25.1에서 제안한 DPe-CIM 가속기 구조

#25-2는 UNIST에서 발표한 논문으로, Hybrid-domain Ternary CIM 프로세서를 소개한다. 본 논문은 ternary neural network(TNN) 기반의 AI 연산에서 에너지 효율성을 극대화한 설계로 주목받았다. 본 연구는 eDRAM 기반 고밀도 셀과 수직 아날로그 및 수평 디지털 연산을 결합한 하이브리드 연산 구조를 통해 기존 CIM 구조의 한계를 극복하고 성능을 대폭 향상시켰다. 수직 아날로그 MAC 연산의 선형성을 높이기 위해 게이트 전압 바이어싱 기법을 적용하여 연산 정확도를 86.8% 향상시켰으며, 1.5-bit/cycle SAR ADC를 사용하여 ADC 면적을 50% 줄였다. 수평 디지털 Psum 연산은 메모리에서 데이터를 효율적으로 분배하고 병렬 연산을 수행하여 데이터 병목 현상을 방지했다. 이러한 설계는 하드웨어 자원의 활용도를 극대화하면서도 높은 처리량을 유지하도록 도왔다. TNN 모델로 CIFAR-10 데이터셋을 평가한 결과, Hybrid-domain CIM은 89.2%의 높은 정확도를 유지하면서도 기존 eDRAM CIM 대비 4.88배 높은 시스템 에너지 효율성과 1.63배 높은 매크로 셀 밀도를 달성하였다. 특히, 수직 아날로그 연산과 수평 디지털 연산을 결합한 설계는 MAC 연산의 병렬성과 속도를 극대화하여 연산 병목을 최소화하였다.



[그림 2] #25.2에서 제안한 HYTEC 구조

#25-4는 KAIST에서 발표된 Embedded Flash 기반 Compute-in-Memory (CIM) 아키텍처로, 플래시 메모리의 비휘발성 특성을 활용하여 엣지 디바이스에서 데이터 이동 지연과 전력 소비를 줄이는 동시에, 높은 에너지 효율과 처리량을 제공하도록 설계되었다. 본 연구는 Negative-Aware-Skip (NAS) 알고리즘을 통해 불필요한 연산을 줄이고, 전력 소모를 최소화하였다. NAS는 음수 출력을 제거하여 CIM 연산의 효율성을 높였으며, 이 과정에서 데이터 변환 속도를 크게 향상시켰다. SAR-TDC (Time-to-Digital Converter)는 변환 과정을 디지털화하는 데 있어 파이프라인 구조를 적용하여 변환 지연을 최소화하였다. 플래시 메모리 매크로는 65nm CMOS 공정을 기반으로 설계되었으며, 1.86mm²의 코어 면적을 차지한다. 이 구조는 플래시 메모리의 고속 데이터 읽기 및 쓰기 기능을 활용하여 연산 병목을 방지하고, 고밀도 데이터 저장과 연산을 통합하여 639.38GOPS의 처리량과 1517.16TOPS/W의 에너지 효율을 달성하였다. NAS 알고리즘은 CIM 연산에서 음수 데이터의 비효율성을 제거함으로써 전력 소비를 추가로 줄였고, MSB와 LSB 간의 차동 전류 구성은 신호의 선형성을 극대화하였다. 이 설계를 통해 기존 NVM 기반 CIM 구조 대비 전력 소비를 최대 42.5% 줄이는 동시에 높은 처리 성능을 유지하였다.



[그림 3] #25.4에서 제시한 eflash-CIM 아키텍처.

저자정보

엄소연 박사과정 대학원생

- 소속 : KAIST 전기및전자공학부
- 연구분야 : Computing-In-Memory Processor
- 이메일 : soyeon.um@kaist.ac.kr
- 홈페이지 : https://ssl.kaist.ac.kr/