A-SSCC 2024 Review

한국과학기술원 바이오및뇌공학과 박사과정 석동열

Session 21 Application-Specific Processors

올해 2024 아시아고체회로학회(ASSCC)의 특수목적 프로세서(Application-Specific Processors) 부문(Session 21)에 선정된 논문은 5개로 주로 일반적인 연산 프로세서를 활용할 때 시간, 메모리, 전력 등의 측면에서 비효율적으로 계산이 이루어 질 수밖에 없는 다양한 응용 분야에서 특정 형태의 연산을 효과적으로 효율적으로 수행하기 위한 하드웨어 연구를 다루고 있다. 본 부문에 게재된 논문의 제목과 연구 목적과 방법론을 요약해 보면(표2), 생체신호나 시각적 정보처리를 웨어러블 장치 또는 인공지능 로봇 등에서 빠르고 전력 효율적으로 수행하기 위한 특수목적 프로세서, 데이터 압축해제, 조합론적-최적화문제(COP) 등 기존의 프로세서가 수행 할 경우 시간이 오래 걸리는 연산을 가속하는 하드웨어 등이 소개되어 있다. 이 리뷰에서는 게재된 다섯 건의 연구 중 #21-3 한 건을 살펴보고자 한다.

[표 2] 특수목적 프로세서 부문 (Session 21) 선정 논문 5개

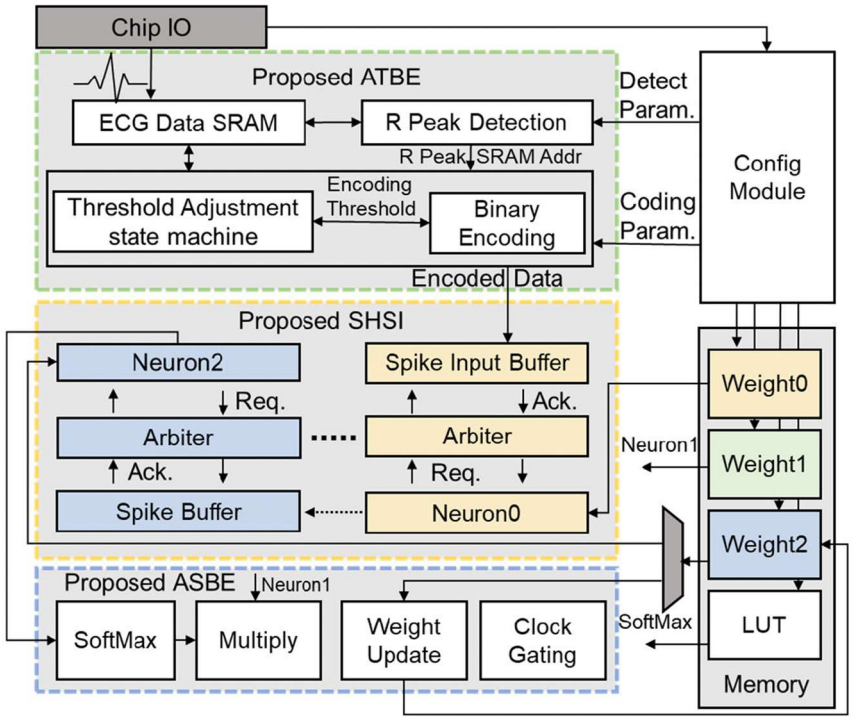
번호	제목	요약
#21-1	ROC-Spin: A 28nm 2,000 Ring-Oscillator-Collapse Spins for Solving Combinatorial Optimization Problems	조합론적-최적화문제(COP) 해결을 위한 이징 모델의 스핀을 링-오실레이터의 듀티 사이클로 구현, 복잡도 높은 COP 문제를 빠르게 해결할 수 있음을 보임
#21-2	A Unified Microrobotic Visual-Perception Processor with 62.2-FPS/mm2 and 103-uJ/frame Navigation in 28nm	마이크로로봇의 시각인지프로세서에서 이루어지는 Cholesky 분해를 가속화 하는 연산모듈 ReMMDA를 고안하여 시간적, 하드웨어 측면에서 연산성능 향상
#21-3	A 0.04 μJ/Classification High-Accuracy Energy-Efficient ECG Processor with SNN On-Chip Backpropagation and Adaptive Threshold Encoding	SNN 층 간 불필요한 연산을 줄이고, 심전도파형 인코딩 방식 제안, 마지막 층 가중치를 역전파보정 하는 방법으로 지연을 줄이고 정확도를 높인 연구
#21-4	A 1.12nJ/Pixel High-Accuracy and Memory-Efficient Real-Time Object Detection Processor for Neuromorphic Vision Sensors	하이브리드 다운 샘플링, 배경제거, 관심영역 설정 등을 하드웨어로 구현하여 1.12nJ/pixel 수준의 고메모리 효율의 실시간 사물 감지 프로세서 개발
#21-5	A 43.3 bit/cycle Inflate Accelerator Featuring Static-Dynamic Huffman Decoder with Multiple Checkpoints and Optimized End-Of-Block Control for Hyperscale data	데이터 압축 해제를 위한 inflate 알고리즘을 하드웨어 가속기 형태로 개발한 연구로 해당 과정에 사용되는 Huffman Decoder, LZ77 Decoder가 압축 데이터를 병렬로 처리하도록 하여 처리 속도를 향상함

#21-3 SNN(spike neural network)은 신경세포의 작동방식을 모방한 인공 신경망 네트워크로 이벤트 발생시 반응하는 특성을 가지고 있어서 저전력 방식의 생체신호 분류 목적의 프로세서에 자주 사용된다. 본 연구는 중국 전자과기대학, Hangzhou Shiguang Xight Inc.

Nanjing Houmo AI Inc.의 공동연구로 진행되었으며, SNN 방식의 심전도(ECG) 분류 하드웨어에서 적응형 임계 값을 활용한 심전도 파형 디코딩 방식(ATBE, Adaptive Threshold Binary Encoding) 과 시간 효율적인 각 층(layer)간 Handshaking 방식의 연산 모듈(SHSI, SNN Hand Shaking Inference), 역전파를 통한 마지막 층 가중치 보정(ASBE, Adaptive SNN Backpropagation Engine)을 도입하여 지연시간을 줄이고, 정확도를 높일 수 있음을 보여주었다.

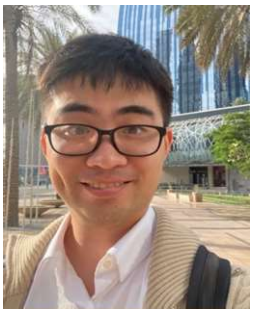
ATBE는 심전도 전위 값의 변화를 추적하여 정상적인 심전도의 파형을 구성하는 P, Q, R, S, T파의 위치 정보를 96 비트의 벡터에 담아내는 알고리즘으로 전위 값의 변화를 검출하는 한계를 적응조절(adaptive control) 하여, 인코딩 된 데이터의 유효성을 높이는 모듈이며, SHSI는 앞 층에서 전달되는 스파이크 정보를 입력 버퍼(spike input buffer)에 저장한 후 전달 된 경우에 대해서만 연산하고, 연산 중에 앞뒤 레이어에서의 스파이크를 방지하는 방식으로 SNN 연산을 단순화하여 지연 시간을 줄였다. 또한, 소프트맥스(softmax) 모듈을 이용하여, 입출력 값 비교를 통한 가중치 변화를 계산하여 마지막 층 가중치를 보정하는 ASBE를 통해 전체 시스템의 정확도를 향상시켰다.

55nm CMOS 공정을 활용하여 0.31mm^2 의 칩을 제작하였으며, 분류 건 당 0.04μJ 수준의 에너지 소모와 98.6%의 분류 정확도 0.3ms 수준의 지연을 성능지표로 보고하였다.



[그림 3] #21-3 ECG 분류 목적의 SNN에서 지연시간 감소, 저전력화 및 정확도 향상 연구 개요도

저자정보



석동열 박사과정 대학원생

- 소속 : 한국과학기술원
- 연구분야 : 바이오메디컬 응용회로 설계(센서 및 신호처리)
- 이메일 : sukd10@kaist.ac.kr

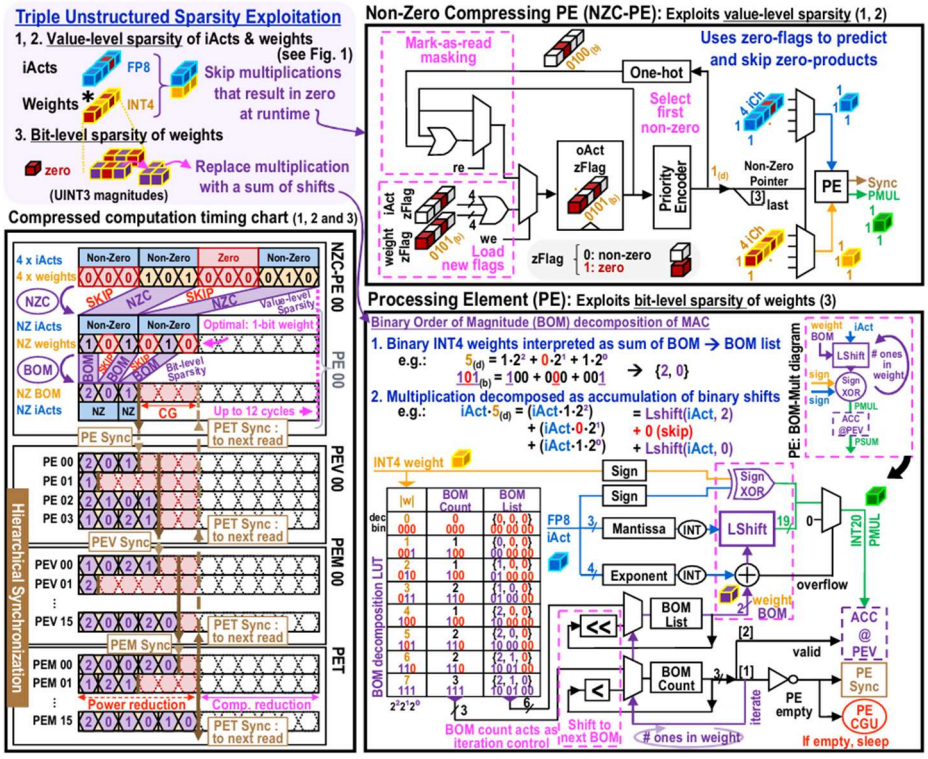
A-SSCC 2024 Review

포항공과대학교 반도체대학원 박사과정 박은빈

Session 3 Efficient AI and DSP Processors

이번 2024 IEEE A-SSCC 학회의 Session 3에서는 Efficient AI and DSP Processor라는 주제로 총 5편의 논문이 발표되었다. 이 세션은 주로 고효율 AI 및 DSP 프로세서 설계를 통해 다양한 응용 시나리오에서 전력 소비를 줄이고 성능을 극대화하는 데 중점을 두었다.

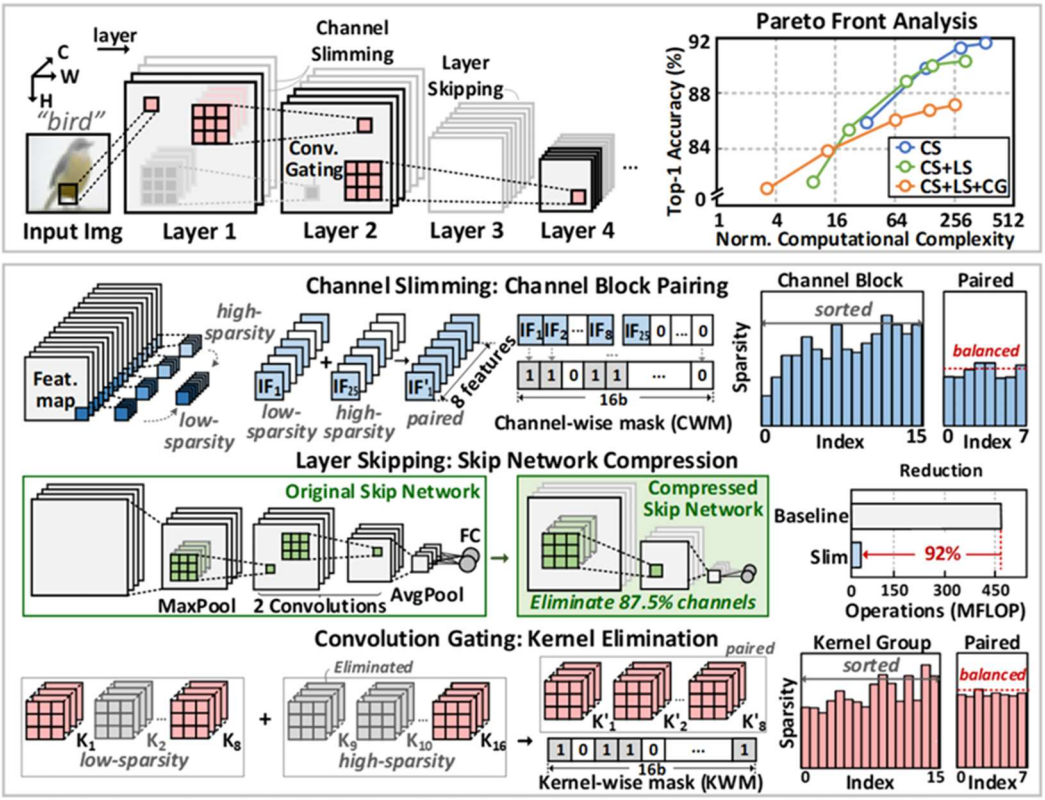
#3-2 논문에서는 WhiteDwarf라는 고효율 신경망 추론 엔진을 발표하였다. 이 엔진은 **트리플 비정형 희소성(triple unstructured sparsity)**과 모델 압축 기술을 결합하여 CNN 및 MLP 모델의 연산 효율성을 극대화한다. 특히, 활성화, 가중치, 비트 수준의 희소성을 활용해 데이터 처리 효율성을 높이고, INT2-4와 FP8 같은 적응형 정밀도를 통해 메모리 접근과 전력 소비를 줄였다. 이를 통해 ResNet-50 모델은 원래 크기의 5.7%로 압축되었고, 정확도 손실 없이 ImageNet 데이터셋에서 74.7%의 성능을 유지하며, 12.24 TFLOPS/W의 에너지 효율성을 달성하였다.



[그림 1] triple unstructured sparsity에 관한 설명

WhiteDwarf는 RISC-V 기반의 커스텀 컨트롤러와 Huffman 코딩 기술을 사용하여 다양한 신경망 모델을 지원하고 메모리 접근을 최소화한다. 또한, 클럭 게이팅(clock gating)과 같은 전력 최적화 기술을 통해 유휴 상태의 전력 소비를 줄였다. 본 논문은 트리플 희소성 활용과 적응형 정밀도를 통해 엣지 컴퓨팅과 같은 자원 제약 환경에서의 딥러닝 적용 가능성을 크게 확대하며, 차세대 신경망 가속기의 설계 방향성을 제시한다는 점에서 큰 의의를 가진다.

#3-5 논문에서는 동적 신경망(dynamic neural network)의 하드웨어 지원을 통해 에너지 효율을 극대화하는 딥러닝 가속기를 제안하였다. 이 가속기는 채널 슬리밍(channel slimming), 레이어 스키핑(layer skipping), 컨볼루션 게이팅(convolution gating) 같은 최신 최적화 기법을 결합하여 다양한 운영 시나리오에 적응할 수 있는 설계를 구현하였다. 특히, 채널 및 커널 데이터를 페어링하여 연산 효율을 극대화하고, 불필요한 연산을 줄여 메모리 접근과 전력 소비를 효과적으로 감소시켰다. 이러한 기술을 통해 8.1-to-353 TOPS/W의 에너지 효율성과 높은 처리량을 실현하였다.



[그림 2] 채널 슬리밍, 레이어 스키핑, 컨볼루션 게이팅에 관한 설명

이 가속기는 CIFAR-10 데이터셋을 기반으로 실험한 결과, 83~93%의 정확도를 유지하면서도 최대 3.94배 높은 에너지 효율과 55.8배 높은 면적 효율을 기록하였다. 또한, 동적

데이터 흐름(dynamic dataflow)을 활용해 프로세싱 유닛(PE)의 활용도를 극대화하고, 레이어 스키핑을 통해 비핵심 연산을 효율적으로 제거하였다. 이러한 설계는 엣지 AI와 같은 자원 제약 환경에서 딥러닝의 효율적 운영을 가능하게 하며, 하드웨어 차원에서 동적 신경망의 잠재력을 효과적으로 실현한 점에서 큰 의의를 가진다. 본 논문은 다양한 에너지-성능 요구를 충족시키는 차세대 딥러닝 가속기의 설계 방향성을 제시하며, AI 응용 분야에서의 실용적 기여 가능성을 입증하였다.

저자정보

박은빈 박사과정 대학원생

- 소속 : 포항공과대학교
- 연구분야 : HW설계 및 딥러닝 최적화
- 이메일 : eunbin@postech.ac.kr
- 홈페이지 : https://sites.google.com/view/epiclab/member/ebpark

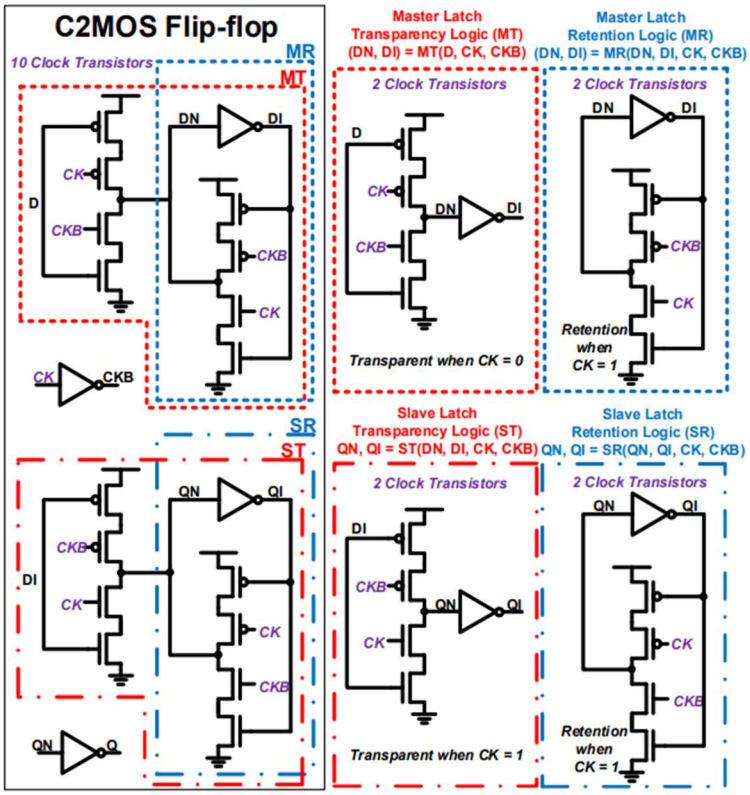
A-SSCC 2024 Review

KAIST 전기및전자공학부 석사과정 박민하

Session 15 Energy-Efficient Circuit Level Techniques

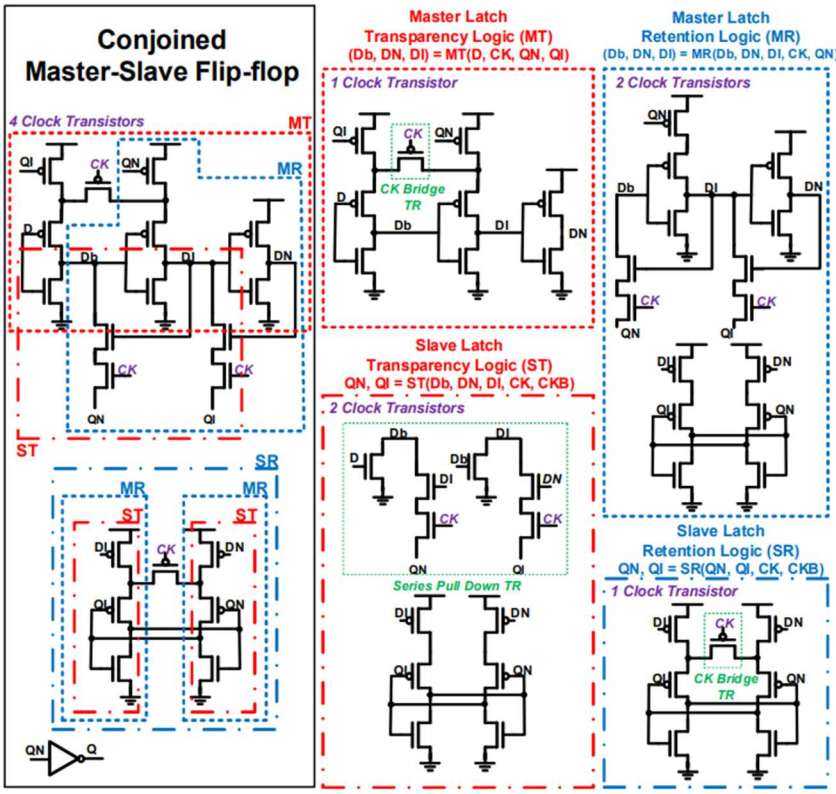
Session 15 Energy-Efficient Circuit Level Techniques에서는 Asynchronous Spiking Neural Network, Time-multiplexed Random-Access Processing-in-Memory, Static Contention-free Conjoined Master-Slave Flip-flop, Asynchronous Non-Volatile-Memory-based Computing-In-Memory Neuromorphic Processor 등 총 5편의 논문이 발표되었다. 이 중 Static Contention-free Conjoined Master-Slave Flip-flop, Asynchronous Non-Volatile-Memory-based Computing-In-Memory Neuromorphic Processor에 대해 살펴보고자 한다.

#15-1 본 논문은 Low-Voltage, Low-Power, 그리고 Low-Area를 갖는 Contention-free Master-Slave Flip-Flop을 제안한다. 기존 C2MOS 또는 flip-flop의 경우 static contention-free와 낮은 전력 소비를 동시에 달성하지 못하는 문제점을 갖고 있거나, 이를 달성하더라도 큰 면적을 차지하는 문제점이 있다.



[그림 1] 독립적인 두개의 래치(Master/Slave)로 구성된 기존 C2MOS 구조

위와 같이 기존의 C2MOS 구조는 총 24개의 트렌지스터와 10개의 클록 트렌지스터를 요구한다. 그러나 제안된 CMSFF의 구조는 다음과 같다.



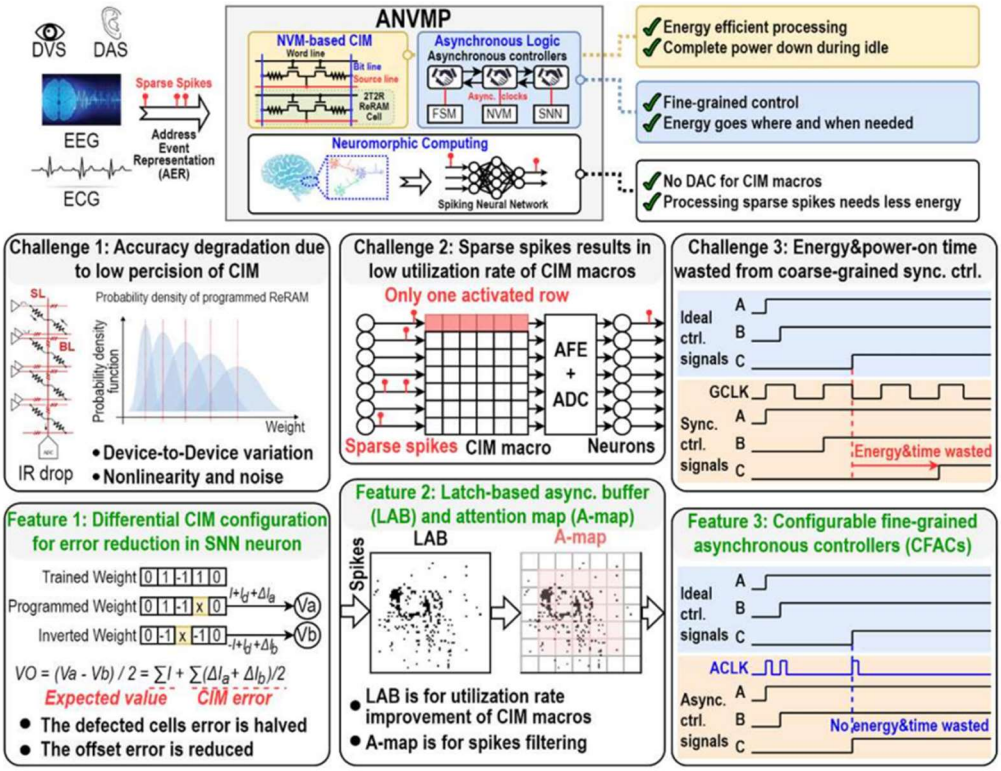
[그림 2] 제안된 CMSFF 구조

CMSFF는 Master와 Slave 래치를 유기적으로 결합하여 트렌지스터와 클록 제어 회로를 줄였다. Master latch의 transparency와 데이터 유지 기능을 slave latch와 공유하며 클록 트렌지스터를 기준으로 2개에서 1개로 줄여, Transparent 상태와 retention 상태 간의 전환을 효율적으로 수행하도록 하였다. 결과적으로, 트렌지스터 수를 22개로 줄였으며, 면적 및 전력을 절감할 수 있었다.

28nm 공정을 통해 test chip을 제작하였고, CMPSFF를 포함한 7개의 플립플롭과 비교 실험을 수행한 결과, CMSFF는 클록 전력을 1.0V에서 60%, 0.4V에서 63%까지 절감하였다. 면적 측면에서는 REFF 대비 20%, TGFF 대비 6%의 면적 감소를 달성하였다. 또한 몬테카를로 시뮬레이션에서 CMSFF는 static contention-free를 입증하였고, 저전압 환경에서도 안정적으로 동작함을 보여주었다.

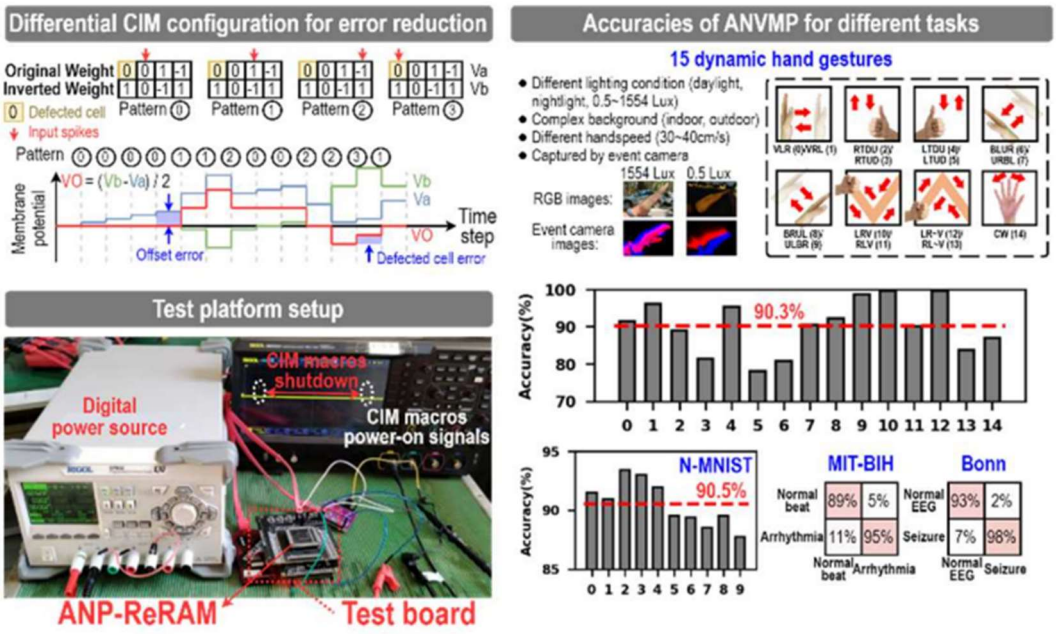
#15-2 본 논문은 저전력 엣지 AI 애플리케이션을 위해 설계된 비휘발성 메모리 기반 컴퓨팅-인-메모리(CIM) 뉴로모픽 프로세서(ANVMP)에 관한 연구이다. ANVMP는 28nm 공정으로 제작되었으며, 다양한 작업에서 높은 에너지 효율성과 낮은 전력 소비를 달성합니

다. 엣지 AI 장치에서 저전력과 다기능성에 대한 요구가 증가하고 있고, 기존 연구는 mW급 전력 소모 또는 단일 작업에 국한된 문제를 가진다. 다음 그림 3은 ANVMP가 해결하고자 하는 주요 과제와 이를 해결하기 위한 핵심 기술을 시각적으로 나타내고 있다. 제안된 ANVMP가 기존 뉴로모픽 설계의 한계를 어떻게 극복하였는지를 한 눈에 보여주는 중요한 도식으로, 본 연구의 주요 기요를 나타낸다.



[그림 3] 제안된 asynchoronous NVM-based CIM neuromorphic

ANVMP 설계의 도전 과제는 NVM 기반 CIM의 오차와 SNN 뉴런의 누적 오차, 희소 스파이크로 인한 CIM 매크로 활용률 저하, 그리고 글로벌 클럭 사용으로 인한 에너지 낭비였다.



[그림 4] ANVMP의 오차 보정 능력과 성능 평가

따라서, 제안된 설계에서는, Differential CIM configuration으로 CIM 오차를 줄이고 평균 정확도를 2% 향상시켰다. 그리고 비동기 버퍼(LAB)와 A-map으로 불필요한 전력 소비를 줄였다. 이를 통해 power-on 시간을 최대 45.1%까지 절약할 수 있었다. 또한 세분화된 비동기 컨트롤러CFAC)를 통해 동적 주파수 제어로 CIM 매크로의 power-on 시간을 41% 단축하였다.

제안된 ANVMP는 비휘발성 메모리(NVM) 기반 CIM, 스파이킹 신경망(SNN), 비동기 logic를 통합하였고, 결과적으로 보면, 이미지 분류에서 90.5%, 손 제스처 인식에서 90.3%, 심전도(ECG) 부정맥 감지에서 94%, 뇌파(EEG) 발작 감지에서 95.6%의 정확도를 보였다. 또한 전력 소비는 평균적으로 52.6μW였으며, 디지털 코어가 전체 전력의 63.5% 차지하였다.

저자정보

박민하 석사과정 대학원생

- 소속 : KAIST
- 연구분야 : 디지털 회로 설계
- 이메일 : mhpark@ics.kaist.ac.kr
- 홈페이지 : https://idec.or.k