2022 IEEE CICC Review

아주대학교 전자공학과 석사과정 이동근

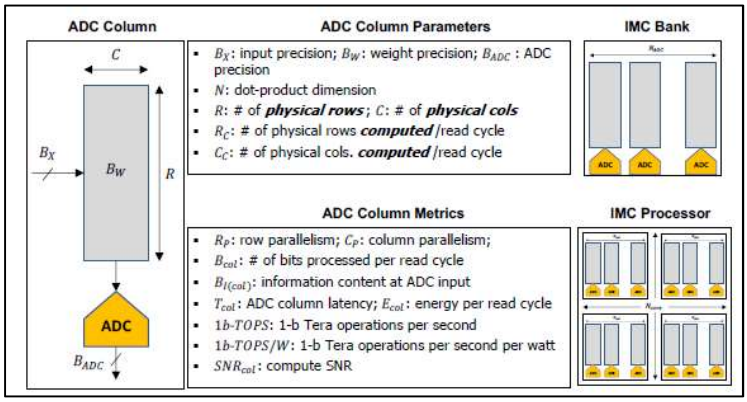
Topic : Digital Circuits, SoCs, and Systems

Session 7 : Digital Circuit

2022 IEEE CICC에서 발표된 논문에 의하면, session 7인 Digital circuit 분야에서는 in-memory computing (IMC)의 에너지 효율과 성능 향상을 목표로 연구가 진행되고 있음을 짐작할 수 있다. 현재 Artificial intelligence (AI) 연산에서는 수많은 메모리 입출력이 발생하기 때문에 기존의 폰 노이만 (Von Neumann) 구조에서는 입출력 속도와 병목현상 문제로 인한 AI 성능 개발에 한계가 있음이 드러났다. 하여, 메모리 내에서 연산을 처리하는 IMC에 대한 개발이 매우 중요하게 되었다. IMC는 메모리 내에서 연산을 진행하기 때문에 메모리 입출력이 최소화되면서 입출력으로 인해 소모되는 에너지도 감소함으로써 높은 에너지 효율과 전체적으로 향상된 성능을 보여준다. 이러한 이유로 인하여, 현재 IMC 관련 많은 연구가 진행되고 있다.

#7-1 Comprehending In-memory Computing Trends via Proper Benchmarking

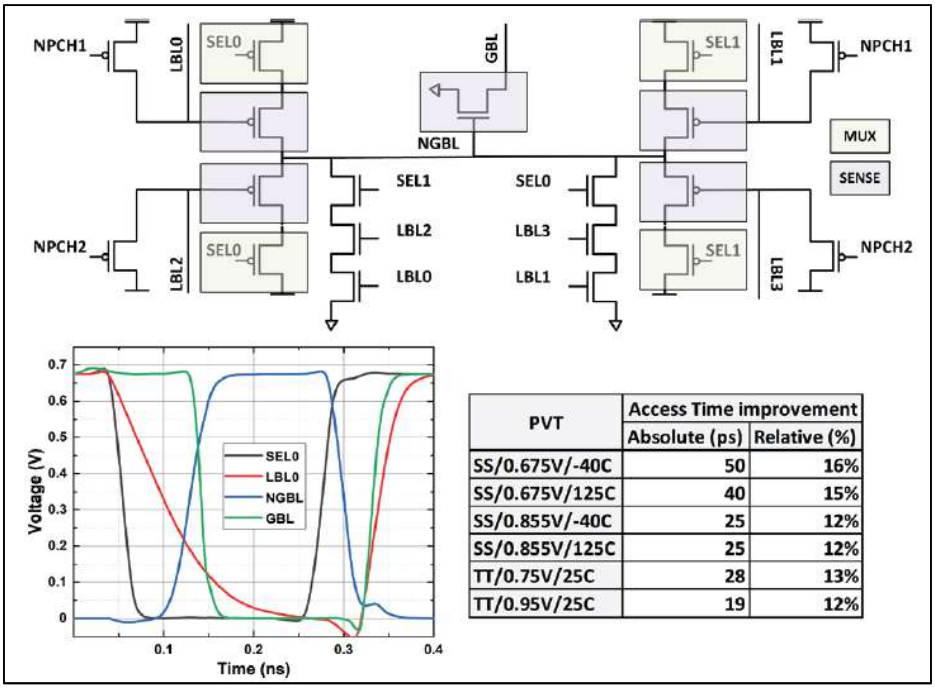
AI 분야에 있어 IMC 개발이 매우 중요해진 만큼 IMC 관련 circuit design에 대한 많은 연구가 진행되고 있다. IMC의 에너지 효율 및 성능을 향상시키 위해 circuit design에 대한 다양한 아이디어가 발표되지만 현재 IMC design과 관련된 benchmarking methodology가 없어 IMC design 관련 연구들을 평가할 수가 없다. 하여 해당 논문에서는 IMC design 관련 연구들을 조사 및 분석하여 IMC benchmarking methodology를 발표하였다.



[그림 1] Proposed hierarchical view of IMCs

#7-2 5GHz SRAM for High-Performance Compute Platform in 5nm CMOS

IMC 분야만 아니라 high performance compute (HPC) 분야에 대한 연구도 진행되었다. 해당 논문에서는 기존 SRAM의 cycle time과 access time 문제와 dynamic voltage frequency scaling (DVFS) 문제로 인해 HPC platform 개발에 한계를 직면하게 되었다. HPC platform의 전체적인 성능을 향상시키기 위해서는 이러한 SRAM의 한계를 극복해야 했고, 그리인해 6T 기반 SRAM에 비해 최소동작전압 (V_{min})이 더 낮고 write time과 access time이 더 빠른 8T 기반 SRAM에 대한 연구를 진행하였고 그 결과 최근에 연구된 SRAM에 비해 더 높은 max frequency를 가지는 SRAM 개발에 성공하게 되었다.

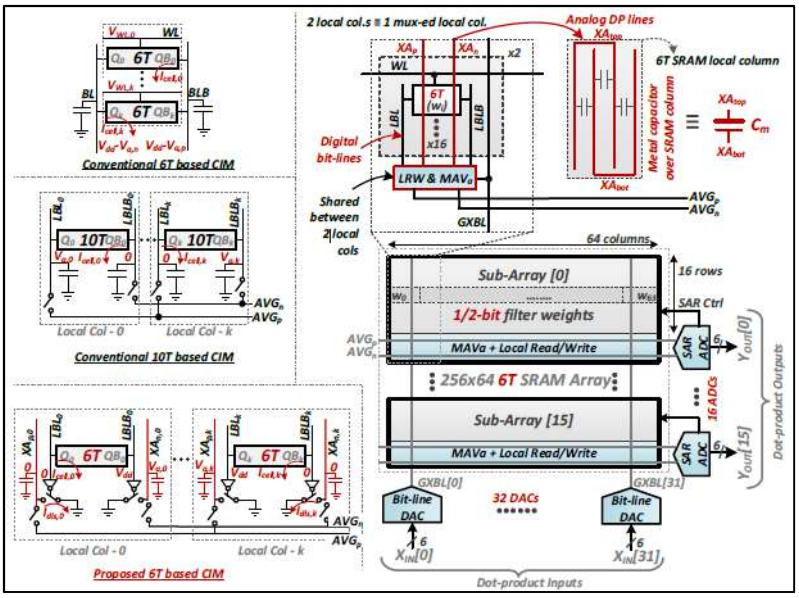


PVT	Access Time improvement	
	Absolute (ps)	Relative (%)
SS/0.675V/-40C	50	16%
SS/0.675V/125C	40	15%
SS/0.855V/-40C	25	12%
SS/0.855V/125C	25	12%
TT/0.75V/25C	28	13%
TT/0.95V/25C	19	12%

[그림 2] Diagram and signal waveform of the proposed compact BL evaluation circuitry

#7-3 An area-efficient 6T-SRAM based Compute-In-Memory architecture with reconfigurable SAR ADCs for energy efficient deep neural networks in edge ML applications

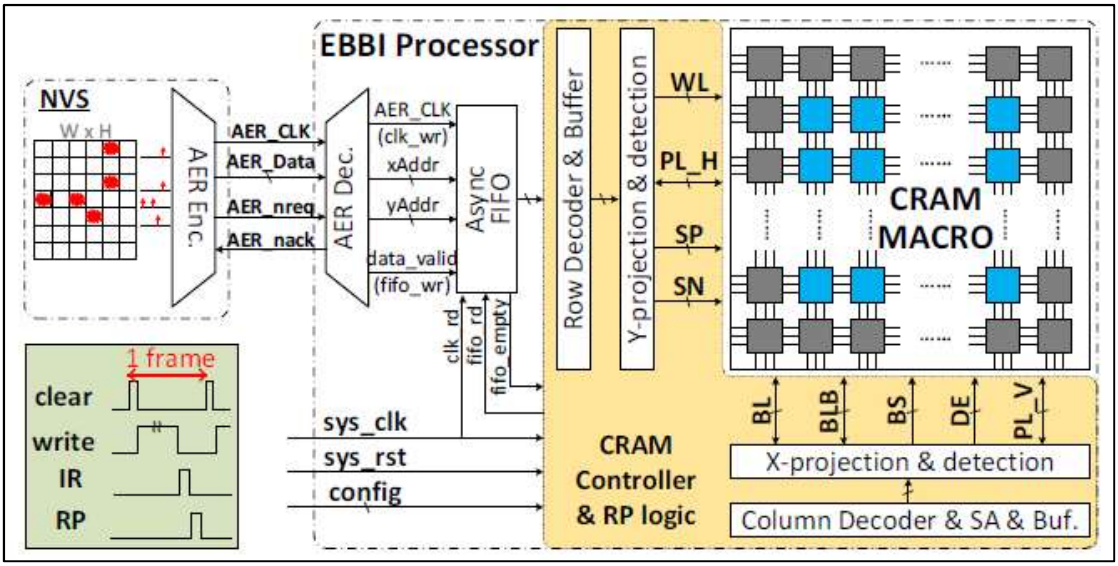
기존 6T SRAM 기반의 Compute-in-memory (CIM) 구조의 경우 SRAM의 bit-cell disturb 문제와 제한적인 dynamic voltage 범위 문제를 가지고 있다. 이러한 6T SRAM의 문제점을 해결하기 위해 10T SRAM 기반 CIM이 제안되었지만 cell 하나가 차지하는 area가 너무 커진다는 문제가 있다. 하여 해당 논문에서는 metal capacitor를 사용하여 아날로그 dot-product computation과 6T cell read를 분리시킨 새로운 6T SRAM 기반 CIM 구조를 제안하였다. 그 결과, 최신 CIM에 비해 집적도와 에너지 효율 면에서 가장 우수하다는 점을 확인할 수 있었다.



[그림 3] 기존 6T SRAM과 10T SRAM 기반의 CIM과 제안된 6T SRAM 기반 CIM 구조

#7-4 An area-efficient 6T-SRAM based Compute-In-Memory architecture with reconfigurable SAR ADCs for energy efficient deep neural networks in edge ML applications

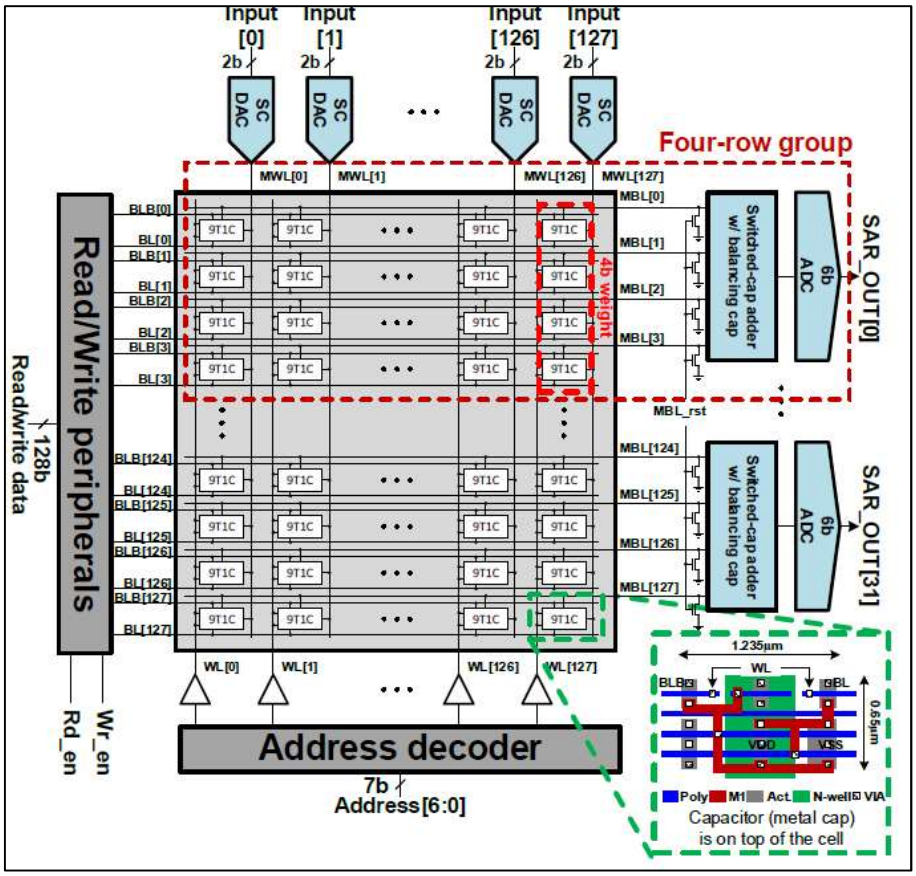
이러한 IMC는 neuromorphic vison sensor (NVS) 연구에도 적용되고 있다. IMC 기반의 NVS는 에너지 효율면에서 매우 우수하다는 장점이 있다. 하지만 IMC의 간단한 알고리즘으로 인해 정확성이 현저히 떨어진다는 문제가 있다. 그러므로 해당 논문에서는 11개의 트랜지스터를 이용하여 SRAM과 DRAM으로 구성된 CRAM 구조를 제안한다. 이러한 CRAM 기반의 IMC로 NVS를 동작시킬 경우, 기존 IMC 기반 동작에 비해 에너지 효율과 성능 면에서 매우 우수하다는 결과를 확인할 수 있게 된다.



[그림 4] The top-level architectural diagram of the proposed CRAM

#7-5 A 177 TOPS/W, Capacitor-based In-Memory Computing SRAM Macro with Stepwise-Charging/Discharging DACs and Sparsity-Optimized Bitcells for 4-Bit Deep Convolutional Neural Networks

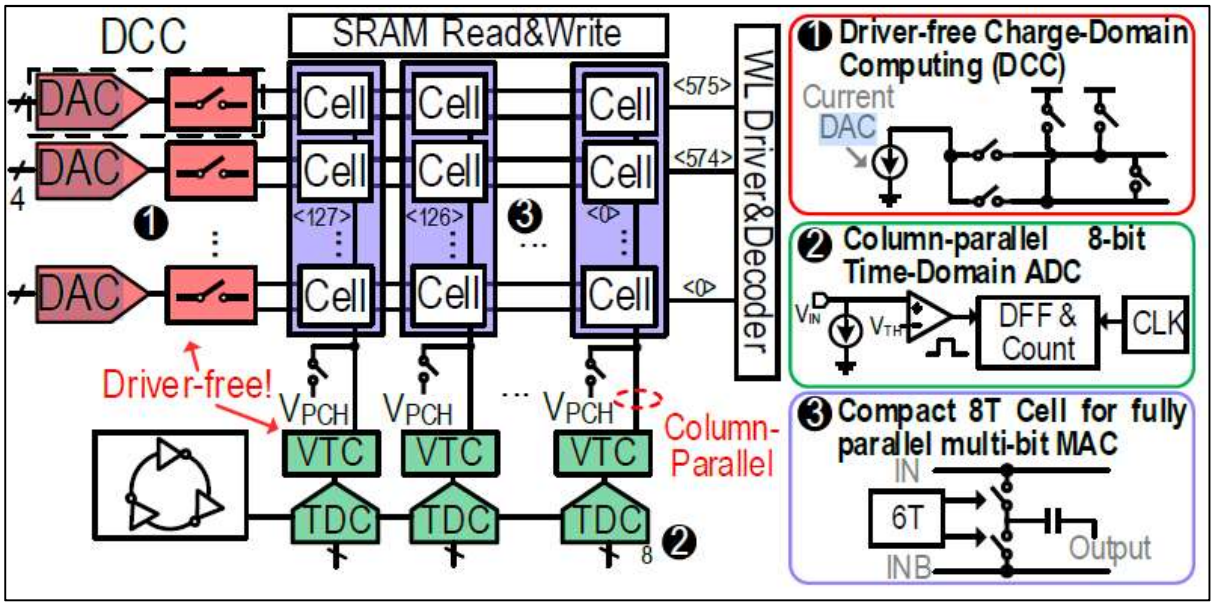
해당 논문에서는 IMC의 에너지 효율성을 더욱 향상시키기 위해 stepwise charging digital-to-analog converter (SCDAC)와 analog adder-first 구조 등 여러 방법들을 도입한 IMC 구조를 제안하였다. 해당 논문에서 제안한 IMC는 다른 기존 IMC에 비해 에너지 효율과 성능 면에서 매우 우수한 것을 확인할 수 있었다.



[그림 5] The proposed IMC SRAM macro

#7-6 DCT-RAM: A Driver-Free Process-in-Memory 8T SRAM Macro with Multi-Bit Charge-Domain Computation and Time-Domain Quantization

SRAM 기반 process-in-memory (PIM)은 전하 영역으로 연산할 때 우수한 아날로그 multiply-and-add computations (MAC)와 quantization을 나타낸다. 하지만 DAC driver PIM의 고밀도 집적에 대한 한계가 존재한다. 그러므로 해당 논문에서는 driver가 없는 driver-free PIM SRAM (DCT-RAM)를 제안하였다. DCT-RAM은 driver가 없을 뿐만 아니라 charge 영역 MAC와 시간 영역 quantization을 수행하기 때문에 높은 에너지 효율과 정확한 MAC가 가능하다.



[그림 6] Proposed driver-free PIM macro with charge computing and time quantization

명예기자 이동근

- 소 속 : 아주대학교 전자공학과 석사과정
- 연구분야 : Low power device
- 이 메 일 : roog258@ajou.ac.kr
- 홈페이지 : https://sites.google.com/a/ajou.ac.kr/edl/home

2022 IEEE CICC Review

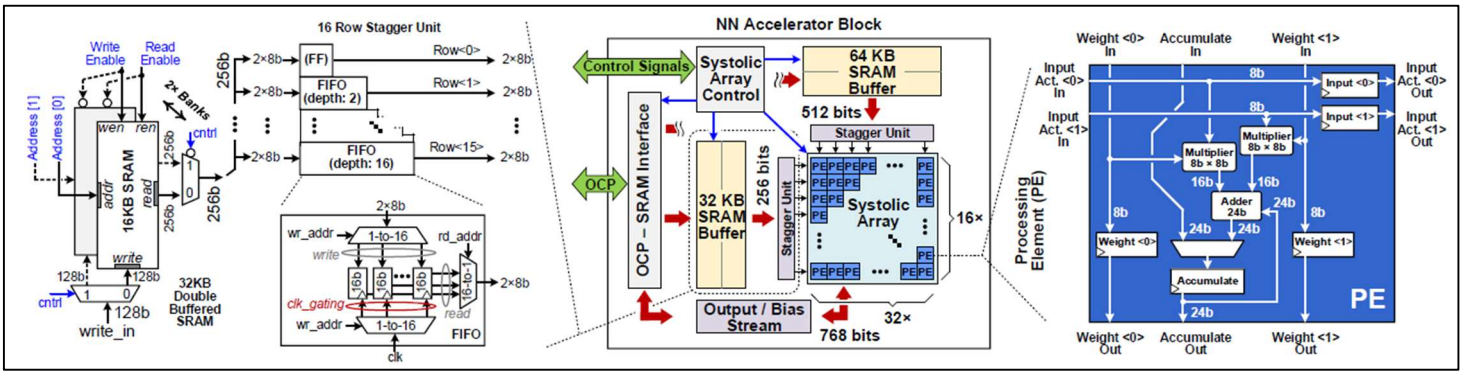
서울대학교 전기정보공학부 석박통합과정 민정우

Topic : Digital Circuits, SoCs, and Systems

Session 19 : High Performance Digital

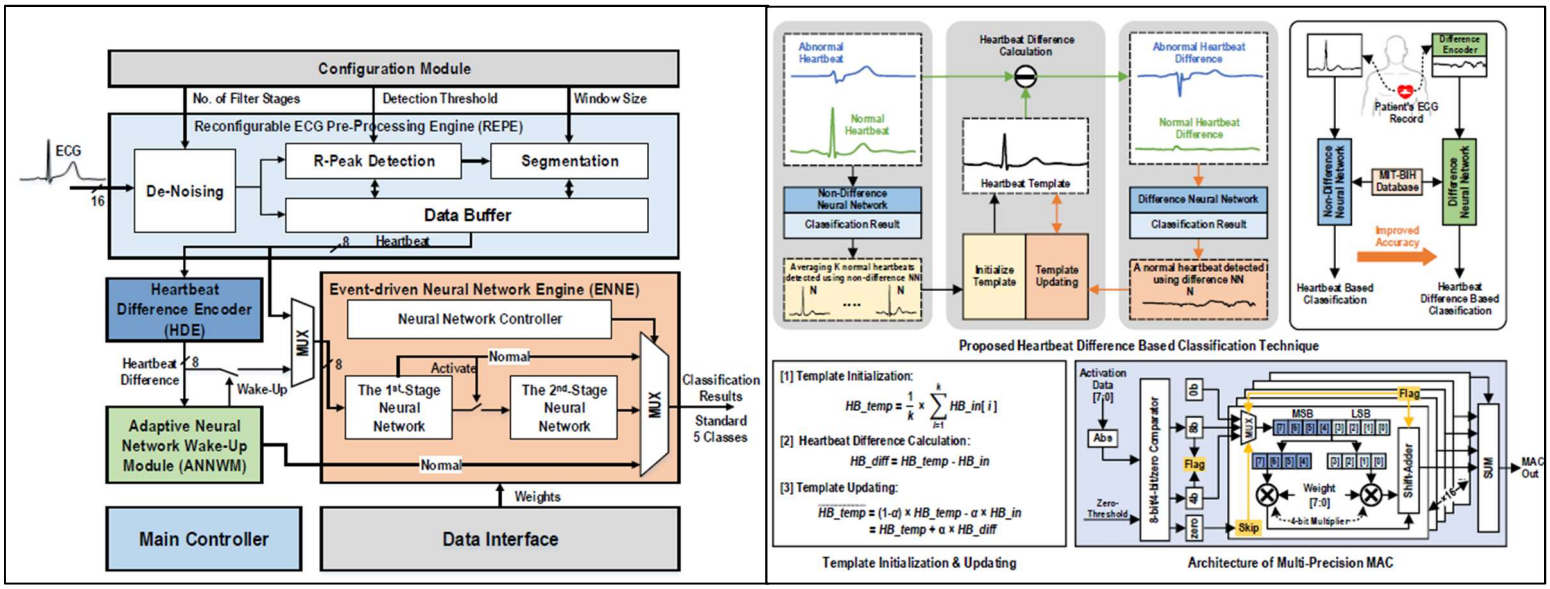
이번 CICC 2022에서 Digital Circuits, SoC, and Systems 분야와 관련해서 Session 7, Session 13, Session 19에서 소개됐다. 그 중에 특히 Session 19는 High Performance Digital이라는 주제로 총 7개의 논문으로 소개됐다. 본 세션에서는 여러 분야에 대한 효율적인 디지털 아키텍처에 대해서 소개하며, 주로 저전력 동작을 위한 여러 방법들을 소개한다. 본 리뷰에서는 7개의 논문중 4개의 논문에 대해 리뷰해보겠다.

#19-1은 Meta의 Reality Labs에서 발표한 논문으로 Codec Avatar을 위한 VR 시스템의 커스텀 SoC 구현으로, 커스텀 NN 아키텍처, 칩에서의 배치, 전체 시스템 레벨에서의 PCB를 소개한다. 프로토타입 PCB는 크게 DNN 가속기, FPGA, DRAM, Flash 메모리로 구성되고 특히 DNN 가속기에 대해 소개한다. CNN 아키텍처에서 offline batch-normalization folding, integer기반 quantization, shortcut 연결을 뺀 ResNet 아키텍처로 효율적인 inference를 가능케했다. 그리고 관련 커스텀 컴파일러 및 C 언어 함수를 구현해 HW-aware 커스텀화를 진행했다. NN 가속기 구조를 살펴보면 그림 1과 같고 systolic-array 기반의 1024 MAC으로 구성됐다. MAC 아키텍처는 output-stationary 방식이고 각 PE는 reduction-of-2 방식으로 2개의 곱셈기와 공유된 adder 구조이다. 입력 matrices는 double buffering 구조로 banking되어 throughput을 높이고 FIFO – staggering unit을 통해 에너지 소모가 큰 re-organization 작업을 피한다. 7nm 공정으로 이러한 아키텍처를 구성한 결과, 30fps 타겟 스펙에 맞춰 입력당 22.7mW 파워와 16.5ms 소요하고 375uJ/frame/이미지 에너지 효율을 달성하였다.



[그림 1] NN 가속기 block 구성: Double-buffering와 staggering unit 회로, systolic-array 기반 MAC block, PE 회로

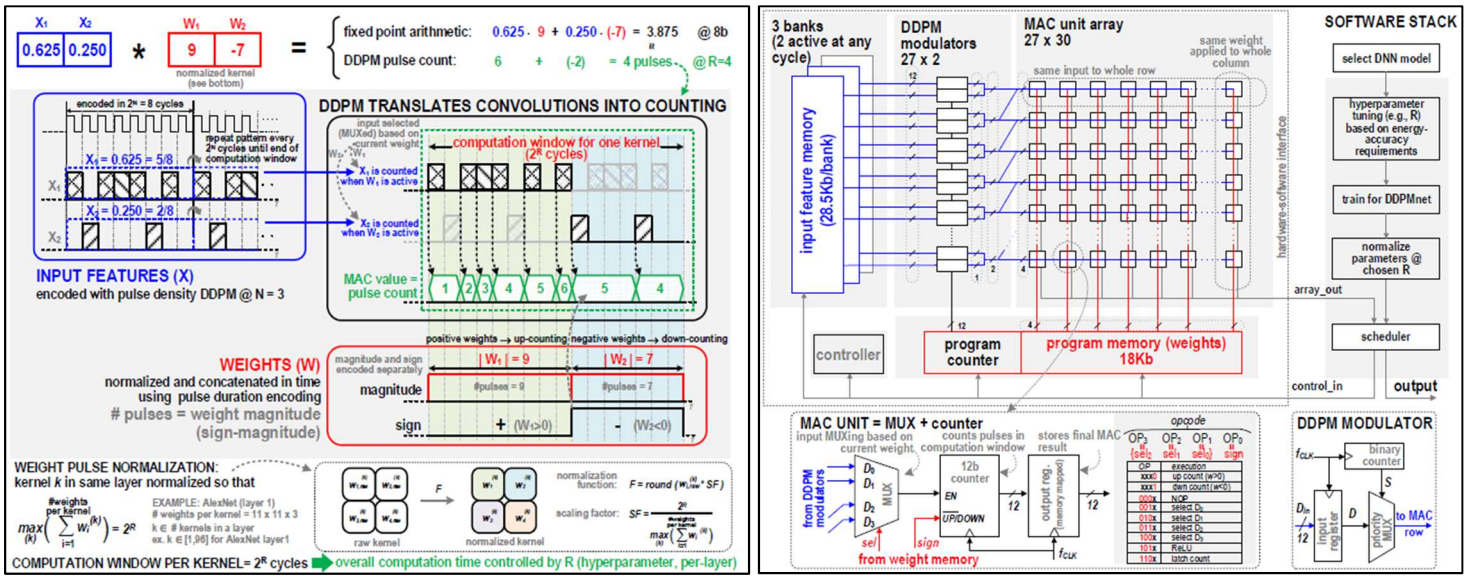
#19-4는 전자과기대학과 둥난대학교에서 발표한 논문으로 arrhythmia 분류 문제를 NN으로 해결할 때, 높은 에너지 소모와 patient-to-patient 변동성에 의한 정확도 저하 문제를 풀고자 한다. 전체 프로세서의 구조는, 그림2의 좌측에서 볼 수 있듯이 REPE, HDE, ENNE, ANNWM, 데이터 인터페이스로 구성된다. REPE를 통해 분할된 심장박동수 신호가 HDE로 전송되어 차이 인코딩이 진행되고, ENNE로 가서 분류작업이 이뤄지고, 이 때 ANNWM 통해 조건적으로 wake-up하게 된다. 그림2의 우측에서 볼 수 있듯이 유저 등록 단계에서 non-difference 네트워크 모델을 통해 초기 정상 심장박동수 템플릿을 추출해 이에 대한 차이를 계산해 학습하여 유저간의 변동성을 반영해 정확도를 향상시킨다. 또한 심장박동수 차이는 절댓값보다 값이 적어서 MAC 연산을 4b multiplier나 zero-data skip을 통해 에너지를 아낄 수 있다. 그리고 2stage event-driven NN 연산 테크닉을 통해 첫 단계에서는 작은 네트워크로 정상여부만 판단하고 두번째 단계에서는 전체 5-class 분류 작업을 진행해 concatenation을 통해 두번째 단계 연산량을 줄여줬다. 마지막으로 ANNWM에서 approximate variance를 계산해 쉽게 분류할 수 있는 정상 심장박동수인지 판단하여 에너지 소모량을 줄여줬다. 해당 프로세서는 MIT-BIH 데이터셋에 대해 평균적으로 8.1%의 정확도 향상, 0.17uJ/분류 에너지 소모, NN-wakeup 테크닉까지 도입하면 0.09uJ/분류 에너지 소모량을 0.2% 정확도 저하로 달성하였다.



[그림 2] Cardiac arrhythmia 분류 프로세서 아키텍처 (좌), 심장박동수 차이 기반 분류 테크닉 (우)

#19-5는 싱가폴국립대학교와 Politechnico di Torino 대학교에서 발표한 논문으로 저비용 edge 디바이스에서의 DNN 가속기가 flexibility가 필요한 상황에서 발생하는 면적 및 에너지 문제에 대해 다룬다. 해당 문제를 DDPM을 통해 펄스 밀도 도메인에서의 MAC 연산으로 해결하려 한다. 그림3의 좌측에서 볼 수 있듯이 DDPM을 통해 convolution 연산을 펄스 카운팅으로, 2^R사이클-윈도우 안에 끝날 수 있도록 weight를 정규화해서 진행된

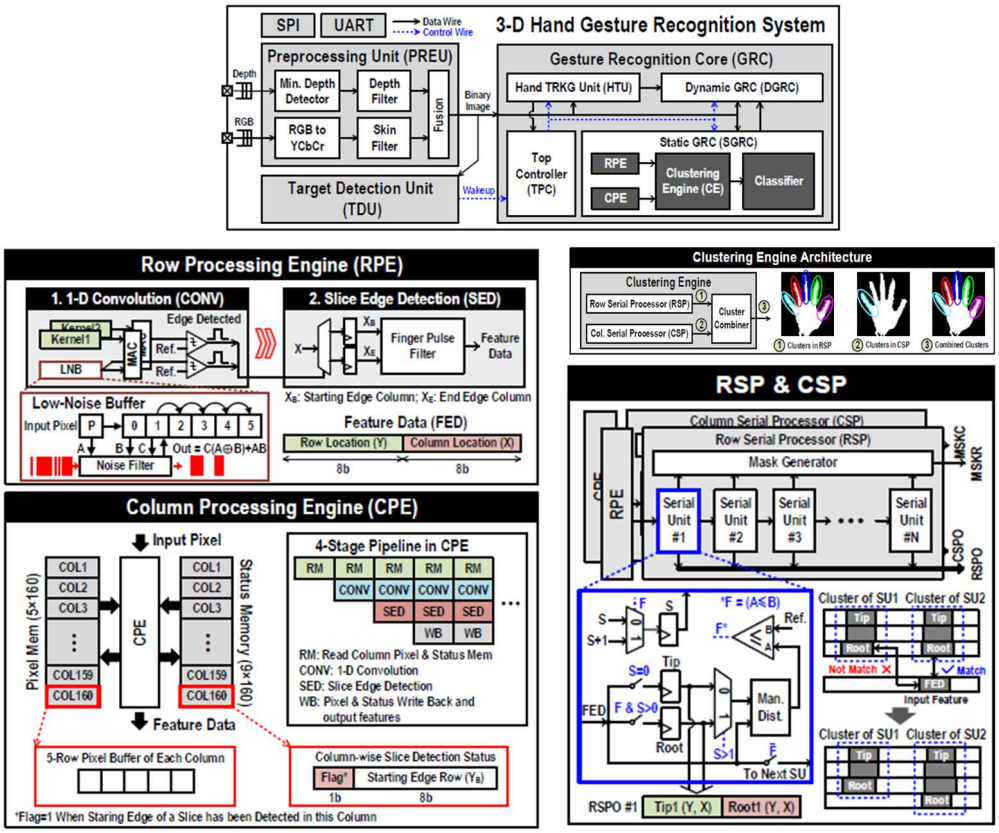
다. 그림3의 우측에서 볼 수 있듯이 27*30 MAC, DDPM modulator, 3 bank 입력 메모리를 통해 three-way time interleaving 방식으로 동작하고 SW stack을 통해 R 최적화, 재학습, weight 펄스 정규화 작업을 진행해준다. 특히 weight 펄스 정규화 작업에서 quantization 에러를 줄이기 위해 DDPM 연산 루프에서 재학습 방법과 평균 residual 에러의 bias 뺄셈 테크닉을 제안했다. MNIST, CIFAR-10, ImageNet 데이터셋에 대해 각각 LeNet-5, AlexNet, AlexNet 구조를 이용했을 때, 각각 22.39/1.22, 12.48/0.84, 9.77TOPS/W/0.52TOPS/mm2 평균 에너지/면적 효율을 달성하였다. 이전 연구에 비해 평균적으로 1.9-27x/152-6285x 에너지/면적 효율 개선을 확인하였다.



[그림 3] Dyadic Digital Pulse Modulation (DDPM) MAC 연산 (좌), DDPMnet 아키텍처 및 SW stack (우)

#19-6은 난양공과대학교와 싱가폴국립대학교에서 발표한 논문으로 NN기반 HGR 시스템이 파워 소모와 시스템의 안정성 사이의 균형을 잡기 어려운 문제를 해결하고자 한다. 제안하는 전체 HGR 구조는 그림4의 상단 부분과 같이 PREU, TDU, GRC와 주변 회로로 구성된다. IDLE 단계에서 PREU를 통해 hand region (HR)의 깊이와 색깔 정보를 fusion해 GRC에 넘겨준다. 이 때, HR이 어느 정도 커야지 TDU를 통해 wake-up해 GRC가 동작하기 시작해, 불필요한 신호 토글링을 없애 에너지 소모를 줄여준다. Static gesture를 인식할 수 있는 SGRC의 내부에는 그림 4의 하단 좌측 부분과 같이 RPE와 CPE는 bi-directional convolution 방식을 통해 손가락 slice를 pipelined 방식으로, 이미지에 회전이 있어도 정확히 인식할 수 있다. 그림 4의 하단 우측 부분에는 공간적인 상관관계에 따라 FED를 grouping 해주는 CE가 있고 mask 생성기와 serial unit으로 구성되어 있다. 각 FED가 어떠한 클러스터에 소속되어 있는지 판단하기 위해 입력 FED와 가장 최근 FED간의 manhattan 거리 계산을 통해 판단해 iterative cluster centroid 업데이트 방식보다 간

편하게 연산을 진행시켰다. TSMC 65nm 공정으로 제작해 타겟 거리 범위인 20~60cm에서 177~183uW 최소 파워를 달성하고 30~35cm에서 96.6% 평균 정확도와 60cm에서 92.8% 최악 정확도를 확인하였다.



[그림 4] 전체 HGR 시스템 아키텍처 (상) Row와 Column Processing Engine(RPE와 CPE) 코어 구조 (하단 좌측), Clustering Engine 아키텍처와 Row와 Column Serial Processor(RSP와 CSP) (하단 우측)

명예기자 민정우

- 소　　속 : 서울대학교 전기정보공학부 석박통합과정
- 연구분야 : 딥러닝 가속기 설계
- 이 메 일 : jwmin@mics.snu.ac.kr
- 홈페이지 : https://mics.snu.ac.kr