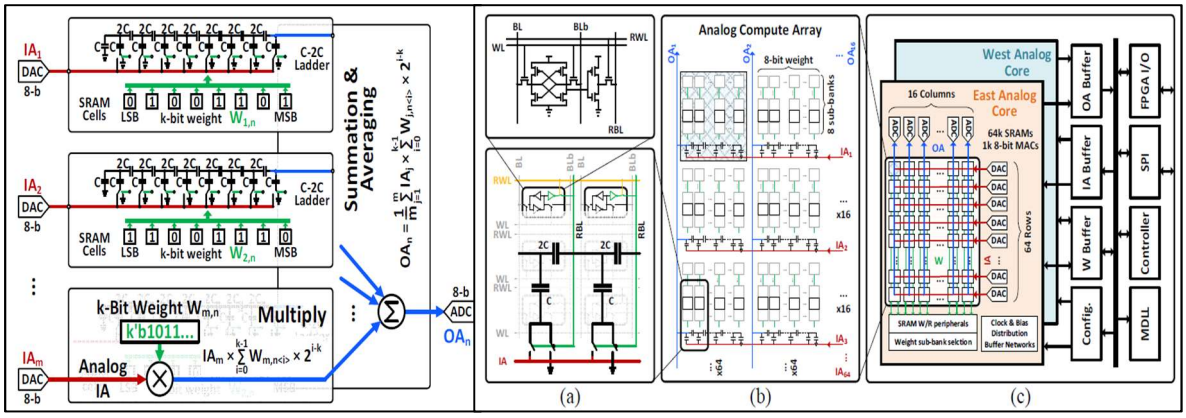
# 2022 Symposia on
# VLSI Technology and Circuits Review

서울대학교 전기정보공학부 석박통합과정 민정우

**Topic : Devices and accelerators for machine learning**
**Session 4 : ML(machine learning) Processors**

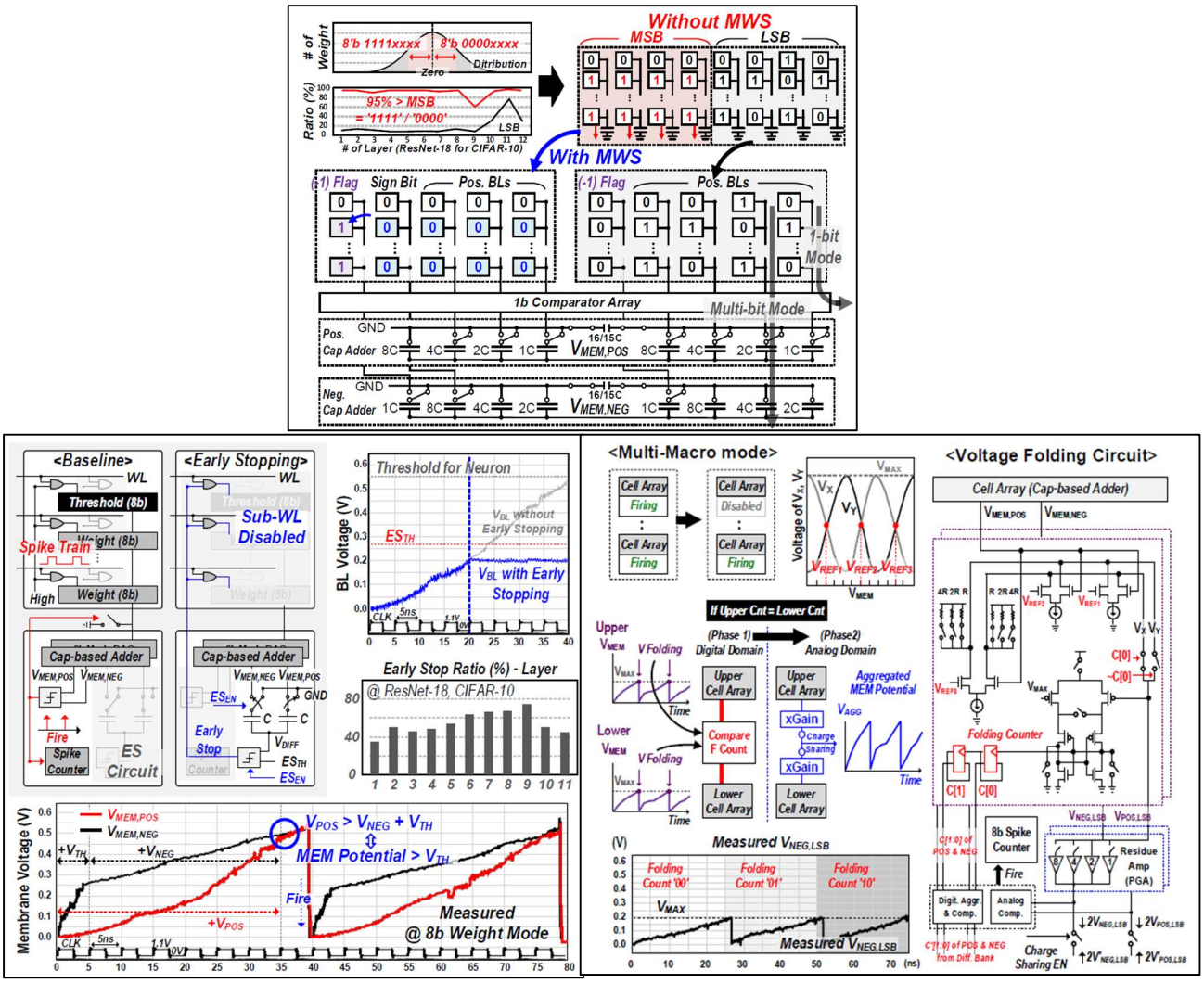
이번 VLSI 2022에서 Devices and accelerators for machine learning 분야와 관련해서 Session 2, Session 4, Session 12에서 소개됐다. 그 중에 특히 Session 4는 ML processors 라는 주제로 총 4개의 논문으로 소개됐다. 본 세션에서는 여러 ML application에 맞게 설계된 ML processor에서 높은 에너지 효율을 확보하기 위해 적용한 테크닉들에 대해 소개한다.

**#4-1**은 Intel의 Intel Labs에서 발표한 논문으로 compute-in-memory (CiM) 구조의 메모리 어레이 안에서의 arithmetic unit (AU)의 제한되는 면적을 해결하기 위해 새로운 scheme을 소개한다. 이전 연구는 XNOR을 이용해서 연산을 진행하지만 정확도 저하 문제가 있고 multibit을 지원하기 위해 binary-weighted capacitor ladder를 제안했지만 스케일링에 문제가 있다. 그래서 해당 논문에서는 C-2C ladder를 사용해서 이를 해결하려고 한다. C-2C ladder 구조는 그림1의 왼쪽과 같고 SRAM셀의 weight값에 따라 스위치를 조절해 capacitive charge sharing을 통해 덧셈과 평균이 자동으로 진행되게 된다. 아날로그 연산에 영향을 끼치는 capacitive parasitic은 unit capacitor 크기를 키워 effective capacitor비를 조절했고, $\mathrm{k_B}T/C$ 노이즈는 quantization 노이즈보다 적게 설정하고, mismatch에 대한 효과를 최소화하기 위해 deviation을 LSB/2보다 작게 구성하였다. 또한 weight loading에 발생할 수 있는 dead time을 줄이기 위해 그림1의 오른쪽과 같이 sub-bank multiplexing 구조를 도입했다. 3개의 추가 트래지스터를 SRAM에 부착해 읽기와 쓰기 동작을 분리해 입력 재사용을 가능하게 했고 하나의 C-2C ladder를 공유 가능하게 구성하여 면적 효율을 높였다. 22nm CMOS 공정으로 구현하여 10k-point 임의 matrix-vector-multiplication (MVM) 테스트를 진행했을 때, $1\sigma$ 에러로 0.65%의 좋은 정확도를 보여준다. I/O TF, INL, DNL 커브를 통해 선형성을 확인했고 32.2TOPS/W, 4.0TOPS/mm$^2$의 높은 에너지와 면적 효율을 달성하였다.




[그림 1] 제안하는 C-2C CiM MAC유닛 (좌), 9T SRAM셀과 weight sub-bank multiplexing 구조 (우) (a), 아날로그 코어 안의 CiM어레이 (우) (b), 칩레벨 CiM아키텍처 (우) (c)

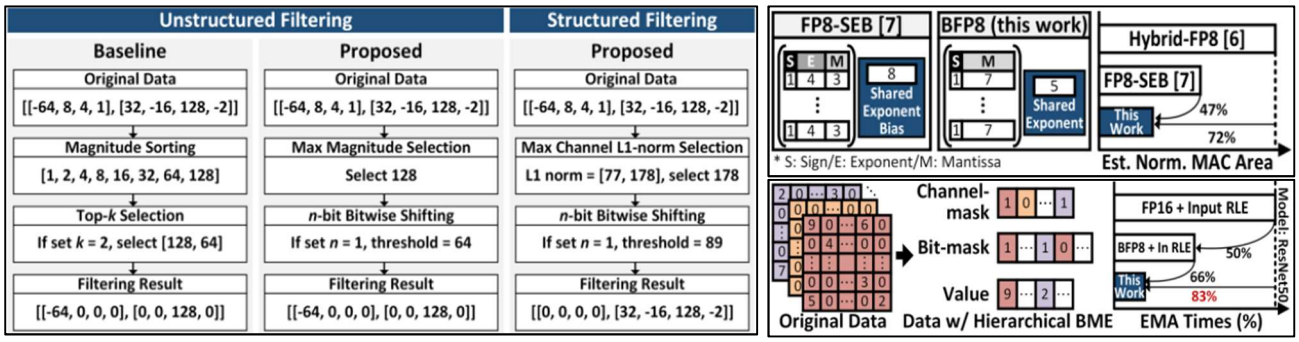
**#4-2**는 KAIST와 취리히연방공과대학교에서 발표한 논문으로 CiM에서 여러 고정확 ADC로 인한 제한된 에너지 효율과 실용적인 예제에서 높지 않은 sparsity 문제를 해결하고자 한다. 해당 논문에서는 실제 뉴로모픽 동작 방식을 활용한 Neuro-CIM을 제안해서 높은 에너지 효율을 달성하였다. 입력 데이터를 action potential (AP) 스파이크 train으로 변환해 membrane potential ($V_{mem}$)에 integrate-and-fire (I&F)하는 SNN 동작방식으로 동작하게 된다. 여기서 에너지 효율을 높이기 위해 3가지 scheme을 제시한다. 그림2의 윗부분과 같이 weight의 가우시안분포 때문에 4b MSB는 대부분 '0000'이거나 '1111'이어서 각 bank의 (-1) flag 셀을 이용해 bitline (BL) 활동을 줄여주는 MSB word skipping (MWS)를 제안했다. 전압 integration 후에 bridge capacitor가 있는 음의 reconfigurable capacitor-DAC adder로 넘어가 $V_{mem,NEG}$를 만들고 나머지는 양의 capacitor-DAC adder로 넘어가 $V_{mem,POS}$를 만들어 연산이 진행된다. 또한 그림2의 왼쪽과 같이 early stopping (ES) scheme을 통해 뉴런이 출력을 하지 않을 것을 예상해 $V_{mem}$ integration을 일찍 멈춰줘 파워 소모를 줄인다. $ES_{en}$ 스위치를 통해 $V_{diff}$를 구해 $ES_{th}$보다 작으면 해당 word line (WL) 드라이버를 꺼서 실제 입력 sparsity보다 더 높은 sparsity를 만들어 적은 정확도 저하로 파워를 절약시킨다. 마지막으로 그림 2의 오른쪽과 같이 folding 카운트와 folded 전압을 이용해 1b dynamic 비교기만으로 dynamic range를 늘릴 수 있다. Programmable gain amplifier (PGA)를 통해 folded 전압을 증폭시켜 charge sharing이 일어나고 난 다음에 resolution을 유지시켜 고정확 ADC를 필요로 하지 않게 된다. 해당 아키텍처를 28nm CMIS로 구현해 CIFAR-10, ImageNet을 ResNet-18, ResNet-50 네트워크에서 테스트해 90.7%-92.1%, 72.5% 정확도를 달성하고 310.4TOPS/W, 546.1TOPS/mm$^2$의 에너지 효율과 면적 효율을 확인하였다.



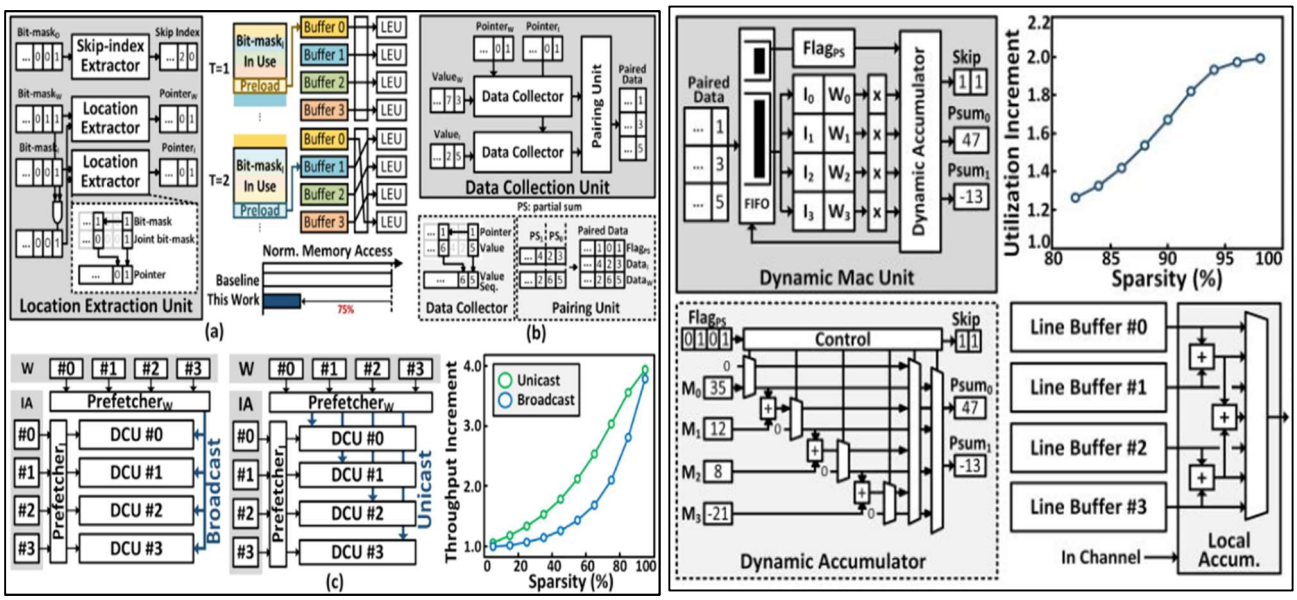

[그림 2] 제안하는 MSB word skipping (상), early stopping 동작 (좌), mixed-mode 뉴런 firing과 folding 회로 (우)

**#4-3**은 대만국립대학교와 Qualcomm에서 발표한 논문으로 edge AI application의 보안문제가 중요해져서 on-device 학습이 중요해지고, 이에 따른 높은 학습 복잡도를 효율적으로 해결하고자 한다. 특히 external memory access (EMA)를 줄이고 더 높은 sparsity를 활용하기 위해 sparsity-scaling training (SST) scheme을 제시한다. 그림3의 왼쪽과 같이 unstructured와 structured sparsity를 위한 filtering scheme을 통해 비교 동작을 줄였다. 또한 SST를 통해 데이터 range가 줄어 그림3의 오른쪽과 같이 BFP8을 사용해 회로면적을 줄였다. 그리고 sparsity를 활용하기 위해 그림3의 오른쪽과 같이 bit-mask encoding (BME)을 통해 hierarchical format으로 데이터를 압축해 EMA를 줄였다. HW를 살펴보면, 그림4의 왼쪽과 같이 match checker의 location extraction unit (LEU)를 통해 non-zero 데이터 포인터를 생성해 인코딩이 진행되고 pair collector의 data collection unit (DCU)를 통해 데이터를 가져온다. Sparsity에 따른 work-load를 지원하기 위해 broadcast와 unicast 모드의 dataflow 아키텍처로 구성됐다. 연산은 그림4의 오른쪽과 같이 sparse

computation engine의 dynamic MAC unit (DMU)의 dynamic accumulator를 통해 sparsity에 알맞게 이뤄지게 되고 마지막으로 post-processing engine을 거쳐서 후처리 돼서 external memory로 보내지게 된다. 해당 아키텍처를 40nm CMOS로 구현해 ImageNet에 대해 ResNet과 VGG 네트워크에서 테스트해 205.2-646.6TOPS/W 에너지 효율을 달성해, 이전 연구에 비해 에너지 효율과 면적 효율 측면에서 3.7배, 4.9배 개선됐다.



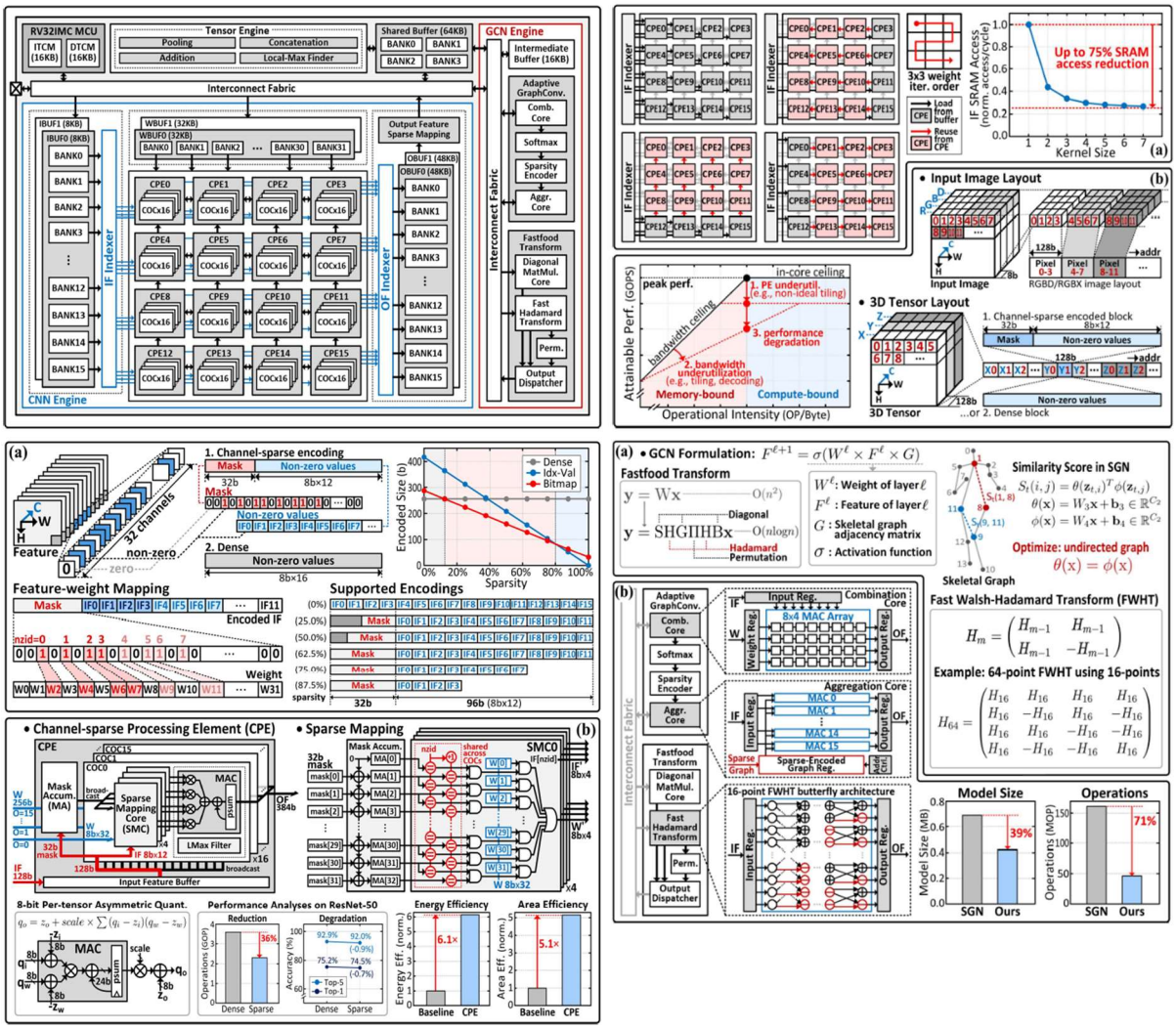

[그림 3] 제안하는 학습 프로세서의 SST scheme (좌) BFP8 data와 hierarchical BME format (우)




[그림 4] 제안하는 학습 프로세서의 LEU, DCU, data collection 어레이 구성 (좌), DMU와 local accumulator 디자인 (우)

**#4-4**는 대만국립대학교와 TSRI에서 발표한 논문으로 제한된 베터리 환경에서 실시간 mobile AR관련 DL 알고리즘이 가능하게 하기 위한 에너지 효율적인 프로세서를 구현하고자 한다. 제안하는 SoC는 그림5의 상단 좌측과 같이 CNN, GCN, tensor 엔진과 RISC-V MCU로 구성된다. CNN 엔진은 channel-sparse processing element (CPE)로 구성되어 그림5의 하단 좌측과 같은 channel-sparse 인코딩을 통해 여러 sparsity 상황을 지원하고 sparse mapping core (SMC)를 통해 입력 feature에 맞는 weight를 선택하면서 효율적으로 모든 CPE가 사용될 수 있다. Sparsity 분포에 따라 인코딩 모드를 선택하여 에너지 효

율을 적은 정확도 저하로 확보할 수 있다. 또한 burst-mode transaction을 위해 제안하는 인코딩 scheme은 3D tensor layout와 같이 설계되어, 그림 5의 상단 우측과 같이 H-W plane의 feature partitioning을 가능하게 하여 성능을 향상시켰다. GCN 연산을 최적화시키기 위해 그림5의 하단 우측과 같이 Fastfood 변환을 통해 연산 복잡도를 낮추고 GCN 엔진을 adaptive graph convolution core (AGC)와 Fastfood transform core (FTC)로 구성해서 해당 연산을 효율적으로 가속시켰다. 해당 아키텍처를 28nm CMOS로 구성해 CNN 엔진은 3.3TOPS, GCN은 72 action/s 성능을 달성했다. 이전 연구에 비해 6% 더 높은 cross-subject와 cross-view 정확도를 달성했고 0.89 action/mJ 에너지 효율을 보이며, 4.5배 개선됐다.




[그림 5] 제안하는 AR SoC 아키텍처 (상단 좌측), CPE 어레이 dataflow, 3D tensor layout에 의한 성능 향상 (상단 우측), channel-sparse feature-weight 맵핑과 인코딩 scheme (하단 좌측), GCN 최적화와 GCN 엔진 아키텍처 (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