2024 IEEE VLSI Review

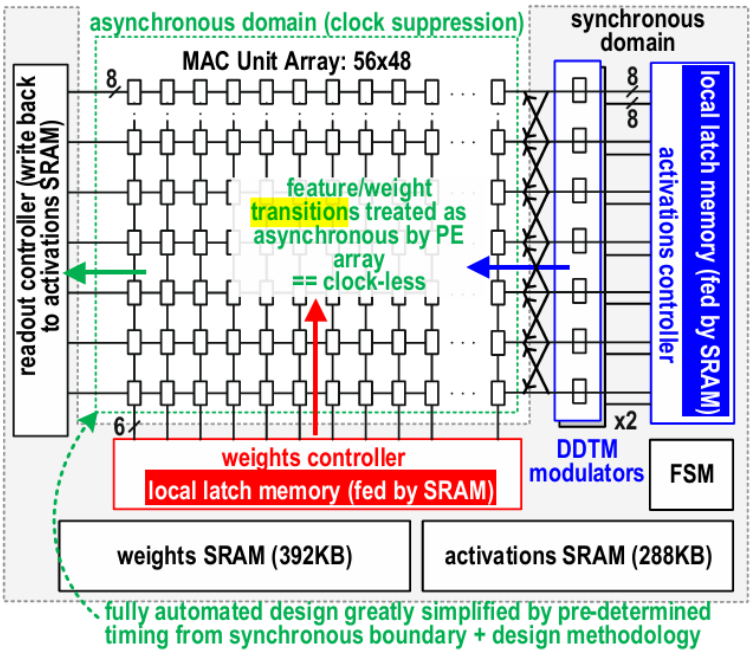
한양대학교 신소재공학과 석박통합과정 송충석

Session 3 AI/ML Accelerators and CiM

이번 2024 IEEE CICC의 Session C3은 AI/ML Accelerators and CiM라는 주제로 총 5편의 논문이 발표되었다. 본 review에서는 3-1, 3-4, 3-5를 리뷰하고자 한다.

#3-1 논문에서는 Dyadic Digital Transition Modulation (DDTM) 이라는 새로운 데이터 표현방식을 기반으로 한 DNN 가속기 구조를 발표하였다. 디지털 기반 인공신경망을 위한 transition density를 통해 저전력 아키텍쳐를 목표로 하고 있으며 비동기식 카운터를 이용하여 MAC 연산을 간단히 하고 높은 에너지 효율성을 이루었다. 비동기식 카운터를 이용하기 때문에 클럭기반으로 동작하지 않고, pseudo-sparsity를 적용하여 MAC 연산기의 activity를 줄였다. 해당 논문은 40nm 공정을 이용하여 100TOPS/W 이상의 성능을 기록하였다.

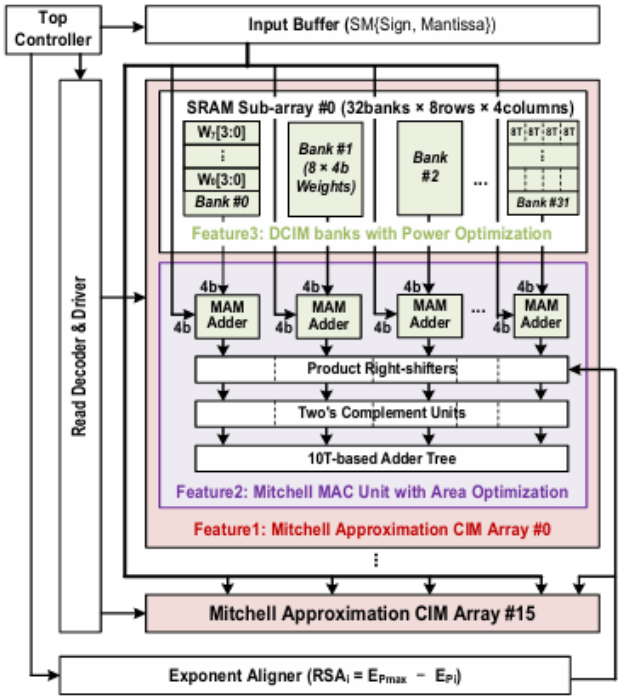
본 논문의 아키텍쳐에 이용된 비동기식 MAC array는 카운터의 동작을 98.4배 감소시켜 2bit 연산에 368.9 TOPS/W의 성능을 나타내었다. 효율을 높이기 위해 설계된 비동기식 MAC 연산기를 통해 하락된 정확도는 재학습을 통해 복구하게 된다.



[그림 3-1] 논문 3-1에서 제안한 전체 아키텍쳐

#3-4 논문에서는 Mitchell's Approximate Multiply (MAM) 기술을 적용하여 에너지와 면적 효율을 증가시킨 디지털 기반 CIM을 발표하였다. MAM은 기존에 CIM에서 사용하는 bit-serial 기반 곱셈방식으로 MAC 연산을 하지만, 본 논문은 더욱 효율적인 방법을 이용하여 부동소수점 연산을 구현하였다. MAM 방식은 복잡한 부동소수점 곱셈을 간단한 부동소수점 덧셈으로 근사화하는 방법으로 n bit 연산을 위해 n번의 cycle이 필요한 CIM 곱셈을 한번에 처리함으로써 처리량이 n 배가 증가하게 된다.

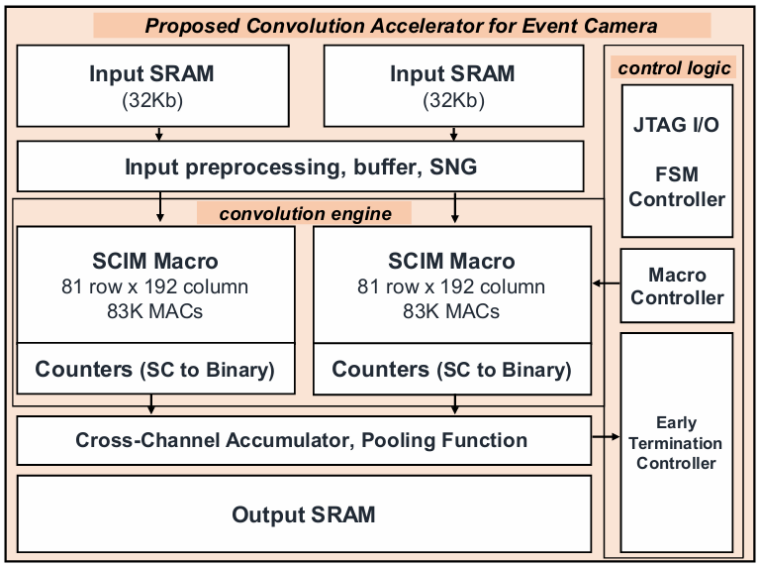
본 논문의 디지털 기반 CIM 매크로는 BF16에서 65.15TFLOPS/W의 에너지 효율과 3.04 TFLOPS/mm^2 의 면적 효율을 나타내었다. VGG16과 ResNet20 모델에서 CIFAR-10 데이터셋을 이용해 각각 90.07%, 91.71%의 정확도를 보였고, ResNet50 모델에서 ImageNet 데이터셋으로 80.52%의 정확도를 달성했다.



[그림 3-4] 논문 3-4에서 제안한 전체 아키텍쳐

#3-5 논문에서는 이벤트 카메라를 위한 ADC-less CIM 기반 컨볼루션 가속기를 발표하였다. CIM 메모리에 binary data를 저장하고 stochastic number generator (SNG)를 사용하여 실시간으로 probability 비트로 변환함으로써 기존 방식보다 10배 이상 적은 저장 공간을 사용하고 각 weight 마다 32개의 작은 MAC 연산기를 내장하고, 불필요한 0 계산을 건너뛰는 방식으로 효율성을 높였다.

본 논문에서 발표한 아키텍쳐는 485 TOPS/W의 에너지 효율과 278-514 Mevent/s의 처리량을 보여주며 기존의 이벤트 카메라를 활용한 가속기보다 60배 이상의 처리량을 기록하였다. 비 이벤트 모드에서도 다른 CIM 아키텍쳐보다 2~3배 높은 에너지 효율성을 보여주었으며 처리량은 10배 이상을 달성하였다.



[그림 3-5] 논문 3-5에서 제안한 전체 아키텍쳐

저자정보

송충석 석박통합과정 대학원생

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