2024 IEEE CICC Review

한양대학교 신소재공학과 박사과정 송충석

Topic : Digital circuits, SoCs, and Systems

Session 7. Mixed-Signal Compute in Memory

이번 2024 IEEE CICC의 Session 7은 Mixed-Signal Compute in Memory라는 주제로 총 7편의 논문이 발표되었다. Compute in memory (CIM) macro는 딥러닝을 가속하기 위해 아날로그 도메인과 디지털 도메인을 나누어서 연산을 하게 된다. 아날로그 도메인을 사용할 경우 연산효율성은 증가하지만 아날로그 회로의 특성상 노이즈에 약하여 정확도가 떨어지는 문제가 있으며 디지털 도메인을 사용할 경우 한 번에 많은 연산을 처리하기 위해 많은 면적을 요구하므로 연산효율성은 상대적으로 낮지만 디지털 회로의 특성상 노이즈에 강한 특징이 있다. 따라서 두 도메인을 적절하게 사용한 혼성신호 기반 CIM macro가 많이 연구가 되고 있다. 본 세션에서는 이러한 혼성신호 기반 CIM macro 를 중국에서 5편, 한국에서 2편 발표하였고, 본 리뷰에서는 7-1, 7-2, 7-3, 7-5, 7-7 을 리뷰하고자 한다.

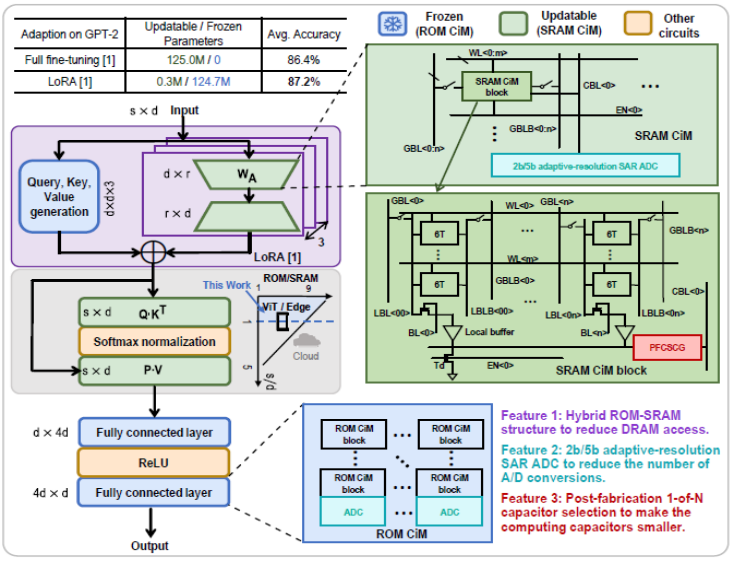
#7-1 논문은 depthwise separable neural network (DSNN)를 CIM에서 가속하기 위한 논문으로 기존 SRAM 기반 CIM에서 DSNN을 구현할 시 3가지 문제점을 지적하였다. (I) 외부 cache에서 activation을 불러오기 위해 반복되는 접근으로 인해 memory energy utilization이 낮고, (II) 짧은 채널 길이로 인해 array temporal utilization이 낮으며, (III) input과 output activation의 분리된 cache로 인해 memory가 두배로 사용되어 memory spatial utilization이 낮다. 이러한 문제점을 해결하기 위해 본 CIM macro인 MixCIM은 computing near memory (CNM)과 CIM을 조합하였다. Depthwise convolution 시 eDRAM array에서 sliding을 가능하게 설계하여 inference시 요구되는 cycle 회수를 감소시켰고, mix memory computing으로 인해 데이터 이동 및 접근 회수를 낮추어 에너지 효율을 증가시켰다. 그리고 input과 output activation을 공유하는 memory sharing scheme을 개발하였다.

본 macro는 28nm CMOS 공정을 이용하여 8bit 데이터를 이용하여 pointwise 연산 시 40.99 TOPS/W, depthwise 연산 시 15.31 TOPS/W 의 에너지 효율을 나타냈고, MobileNet v2 모델을 사용하여 CIFAR-10 데이터 셋을 92.47%의 정확도로 추론해냈다. 더불어

baseline과 비교하여 연산속도를 2.19배 증가시켰고, 유효한 memory utilization을 32.4% 상승시켰다.

#7-2 논문은 transformer 모델을 CIM에 적용하기 위한 논문으로 transformer model의 경우 엄청난 수의 파라미터를 사용하여 연산을 하기 때문에 데이터 이동을 줄이는 것이 필수적이며 CIM은 이를 해결하기 위한 좋은 대안으로 각광받고 있다. 그러나 transformer model을 CIM에 접목시키기에 3가지 문제점이 있고 이를 해결하기 위한 방안을 제안했다: (I) macro 상의 on chip memory의 한계로 인해 DRAM 접근 횟수가 증가하는데 사전에 학습된 weight는 변경되지 않는다는 점을 이용하여 고정된 weight를 위한 ROM과 data를 유연하게 사용할 수 있는 SRAM을 결합하는 구조를 고안해 내었다. ROM은 multi-level을 표현할 수 있는 1T 구조로 밀집도를 증가시켰다. (II) macro의 높은 비율의 전력소모와 면적을 사용하는 ADC의 해상도를 낮추기 위한 adaptive resolution ADC를 고안하였다. SAR-ADC 기반으로 선택적으로 2bit 혹은 5bit A/D 변환을 하게 된다. (III) 마지막으로 아날로그 기반 연산 시 요구되는 면적이 큰 capacitor에서의 에너지 소모를 낮추기 위해 post-fabrication 1-of-N capacitor selection (PFCS)를 적용하여 capacitor의 전체 면적을 감소시켰다.

본 macro는 28nm CMOS 공정으로 1.1V 동작전압에서 220MHz, 0.7V 동작전압에서 120MHz로 동작하였다. Macro의 처리량(throughput)은 0.22 ~ 0.40 TOPS (tera operations per second) 달성하였고, 연산효율은 42.0 TOPS/W를 달성하였다. 더불어 ROM의 사용으로 인해 weight density 또한 높일 수 있었는데, CMOS SRAM 기반 CIM 설계에 비해 17.3 배 이상 증가한 8928Kb/mm^2의 면적효율을 달성하였다.

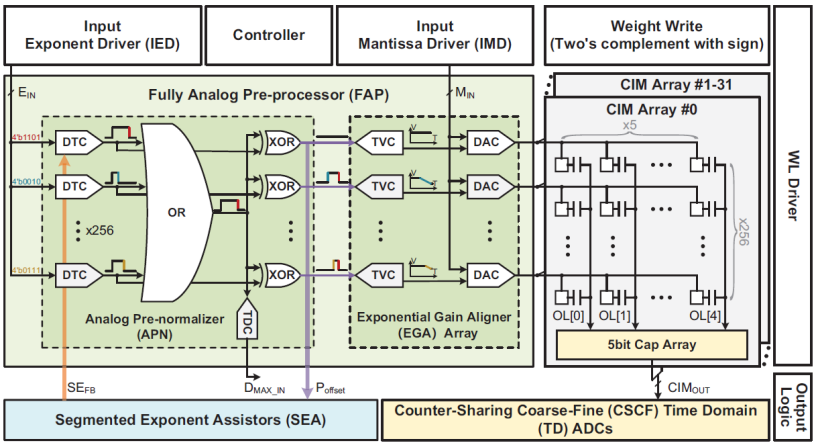


[그림 1] #7-2에서 제안한 hybrid CiM의 전체 구조

#7-3 논문은 연산효율을 증가시키기 위해 아날로그 기반으로 multiply-accumulate (MAC) 연산을 수행하는 CIM macro에서 DAC와 ADC로 인해 생기는 파워 및 면적 overhead를 줄이는 방안을 고안한 macro를 개발하였다. MAC 연산을 위해 DAC과 shift and adder unit (SnA)을 추가로 설계하는 것이 아닌 SRAM bit-cell 내부에서 DAC 과 SnA 동작을 수행할 수 있는 10T2C bit-cell SRAM을 고안하였다. 추가로 여러 bit의 MAC 연산을 수행하기 위해 ADC 동작을 반복적으로 수행하는 기본적인 SRAM 기반 CIM 대비 본 논문에서 발표한 macro는 추가적인 capacitor 없이 하나의 스위치만을 이용하여 SnA 동작을 bit-cell에서 수행하게 하였다. 따라서 DAC과 SnA 동작을 기존 CIM과 거의 동일한 면적을 사용하여 구현하여 아날로그 MAC 연산 시 요구되는 에너지 소모를 감소시켰다.

본 논문의 macro는 28nm CMOS 공정으로 제작되었고, 128x128 array size를 이용하여 0.7V의 동작전압에서 130.0TOPS/W, 1.1V의 동작전압에서 108.1GOPS를 달성하였다.

#7-5 논문은 정수형 데이터를 연산하는 것이 아닌 부동소수점형(floating point) 데이터를 연산하기 위한 아날로그 기반 CIM macro를 발표하였다. 부동소수점형 데이터를 이용하여 모델을 학습시킬 경우 모델의 정확도를 향상시킬 수 있다는 장점이 있는 것에 비해 정수형 데이터를 연산하는 것보다 연산 효율성이 낮아지는 문제가 있다. 특히, 부동소수점형 계산 시 지수부분을 정렬하는 alignment 동작에 성능 병목현상이 나타나는데, 이는 부동소수점형 계산을 아날로그 기반 CIM으로 구현하는데 큰 문제점으로 작용하고 있다. 이러한 문제점을 해결하기 위해 본 논문에서는 아날로그 도메인에서 지수부분을 gain 값을 통해 예측하여 alignment를 수행케 하였다. 더불어 time domain ADC를 이용하여 지수부분의 alignment 과정에서 생기는 시간차이를 이용하여 기존 ADC 대신 적용함으로써 전력효율성을 증가시켰다.



[그림 2] #7-5에서 제안한 부동소수점형 데이터를 연산할 수 있는 fully analog CIM

본 논문의 macro는 28nm 기반으로 제작되었으며 FP8과 BF16을 지원한다. 또한 0.75V의 동작전압에서 83MHz, 0.9V 동작전압에서 125MHz로 동작한다. 본 논문의 macro는 부동소수점형 연산을 완전히 아날로그 도메인에서 수행함으로써 데이터 타입의 변환에 필요한 latency와 area 과부하를 줄였다는 점에서 의의가 있다. FP8 데이터 연산 시 전력효율은 1.78배 증가시켰고, 면적효율은 1.94배 증가시켜, 최대 전력효율은 314.6 TFLOPS/W, 최대 면적효율은 12.13 TFLOPS/mm^2를 달성하였다.

Session 14. Domain-Specific Accelerators

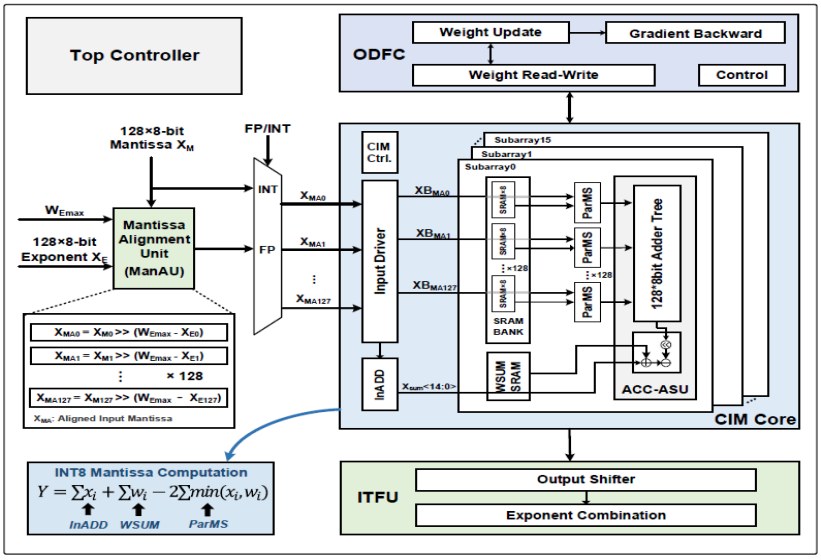
이번 2024 IEEE CICC의 Session 14는 Domain-Specific Accelerators라는 주제로 총 6편의 논문이 발표되었다. CPU 및 GPU와 같은 general purpose의 하드웨어가 아닌 설계의 자유도가 상대적으로 높은 ASIC 기반 하드웨어의 경우에는 특정 application에 최적화하여 설계할 수 있어 더 높은 연산효율성을 기대할 수 있다. 본 세션에서도 특정 domain에 최적화된 하드웨어 설계를 다루고 있다. #14-1 에서는 Edge device를 위한 on-device 튜닝이 적용된 CIM에 관한 macro를 발표하였고, #14-2 에서는 특징 추출과 관심영역 감지(region of interest detection)를 위한 near-sensor convolutional imager SoC를, #14-3 에서는 다양한 형태로 사용되는 행렬곱을 지원하기 위해 설계된 flexible한 프로세서를, #14-4 에서는 multi agent 뉴로모픽 가속기를, #14-5 에서는 의료분야에서 사용할 수 있는 프로세서를, 마지막으로 #14-6 에서는 unstructured sparsity를 구현할 수 있는 인공신경망 가속 프로세서를 발표하였다. 본 리뷰에서는 14-1, 14-3, 14-6을 리뷰하고자 한다.

#14-1 논문은 부동소수점형 데이터를 CIM에서 연산을 가속하기 위한 macro를 발표하였다. 부동소수점형 연산은 복잡한 task를 수행하기 위해 필요하지만 기존 CIM에서 부동소수점형 연산을 하기 위해서는 다음과 같은 문제점이 있다. (I) mantissa 부분의 multiply-accumulate (MAC) 연산 시 비트별로 수행하는 병렬연산은 많은 cycle을 소모하기 때문에 throughput이 낮아지고, (II) mantissa normalization 과정이 많은 비교기를 통해 이루어지기 때문에 면적과 에너지의 과부하가 발생하며, (III) 이전의 부동소수점형 연산을 지원하는 CIM의 경우 디바이스 내에서 fine-tuning을 지원하지 못하여 실제 어플리케이션에 적용했을 때 정확도가 감소하는 문제가 있었다.

따라서 이를 해결하기 위하여 본 논문의 macro는 one-shot compute scheme을 적용하여 mantissa MAC 연산의 throughput을 8배 상승시켰으며, 비교기와 select unit을 이용한 parallel minimal selector를 도입하여 adder tree에서의 잦은 데이터 flip 현상을 줄여 일반

적인 8bit 곱셈기에 비해 면적과 에너지소모를 11.5배, 8.2배 각각 감소시켰다. 또한 디바이스 내에서 fine-tuning을 지원하는 core를 설계하여 어플리케이션이 변경될 때마다 감소하는 정확도를 개선시켰다.

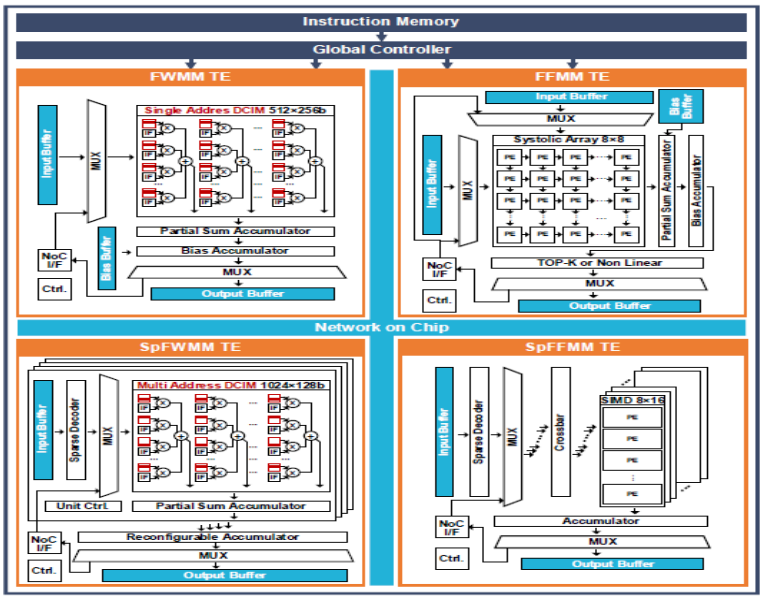
본 논문의 macro는 28nm CMOS 공정으로 제작되었으며 0.55V의 동작전압에서 20MHz 구동을, 0.9V의 동작전압에서 180Mhz까지 구동되었다. Unsigned BF16으로 연산 시 최고 성능은 128TFLOPS/W, 7.02TFLOPS/mm^2를 달성하였다. 이는 기존 연구에 비해 에너지/면적 효율을 4.1배, 3.4배 증가시킨 결과이다.



[그림 3] #14-1에서 제안한 부동소수점형 연산 가속과 on chip tuning이 가능한 macro

#14-3 논문은 딥러닝에서의 핵심연산인 행렬곱을 효율적으로 연산하는 macro를 발표하였다. 행렬곱 연산은 transformer, GCN, 컨볼루션 등 다양한 모델에서 사용되지만 CNN의 경우 CIM에서 고정된 weight를 반복 사용하여 데이터 이동을 줄일 수 있는 반면, transformer의 attention layer, GCN, GNN 등에서의 행렬곱 연산은 인풋이 지속적으로 변하기 때문에 CIM에서 구현하기에 비효율적이다. 이렇듯 다양한 인공신경망 모델에서, 그리고 다양한 인공지능 연산 등을 모두 효율적으로 다룰 수 있는 가속기가 필요하며 본 논문에서 발표한 가속기는 다음과 같은 특징을 가진다: (I) 4개의 서로 다른 tensor engine을 사용하여 병렬 연산 혹은 파이프라인이 가능하게 하였으며, (II) 각각의 tensor engine은 multi-address digital CIM, single-address digital CIM, systolic array, SIMD 로 구성되어 있어 다양한 데이터를 유연하게 처리할 수 있도록 구성되었다. (III) 그리고 각각의 tensor engine에는 로컬 버퍼를 구성하여 NoC를 통해 데이터를 이동시켜 off chip 메모리와의 접근을 최소화했다.

본 논문에서 제안한 가속기는 22nm CMOS 공정에서 제작되었으며 0.76V 동작전압에서 75MHz, 0.92V 동작전압에서 152MHz로 동작하였다. Dense한 INT8 데이터에 대하여 0.8V, 121MHz 환경에서 4.52TOPS/W의 성능을 달성하였고, GCN의 성능은 비교군보다 3.47배 이상의 성능향상을 이루었다. 또한 sparse한 데이터에 대해서도 zero skipping을 통해 높은 성능을 달성하였고 서로 다른 4개의 tensor unit을 적극 활용하여 다양한 데이터셋에 대하여 높은 성능을 이루었다.



[그림 4] #14-3에서 제안한 4개의 tensor engine

#14-6 논문은 Spiking Neural Network (SNN) 기반 모델을 가속시키기 위한 프로세서를 발표했다. SNN은 sparsity가 높기 때문에 효율적으로 accumulation 연산이 가능하지만 반대로 sparsity가 조금이라도 낮은 모델의 경우 효율성이 급감한다는 단점이 있다. 더불어 서로 다른 time step간의 인풋과 가중치를 불러오기 위해 반복적으로 메모리에 접근해야 하고, unstructured spike의 경우 throughput 개선효과가 없으며, 모델 내 서로 다른 연산에 대해 latency의 불균형으로 인한 최적화된 스케쥴링을 하기 어렵다는 문제점이 있다. 이러한 문제점을 해결하기 위해 본 논문에서는 3D computation array를 이용하여 timestep별로 연산을 동시에 하도록 했고, unstructured non-zero 데이터를 다루기 위해 non-zero 주소 생성을 위한 fetcher와 비동기식으로 동작하는 sparse skipper를 위한 스케쥴러를 도입하였다.

본 논문은 40nm CMOS 공정으로 제작되었으며 0.56V의 동작전압에서 50MHz, 1.1V의 동

작전압에서 200MHz로 구동된다. 0%에서 97%까지의 sparsity에 대하여 20.3 ~ 132.4 TOPS/W의 성능을 달성하였으며 다른 최신 SNN 가속기에 비해 약 3.57배의 연산성능을 보였다.

저자정보

송충석 박사과정 대학원생

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

2024 IEEE CICC Review

KAIST 전기및전자공학부 박사과정 엄소연

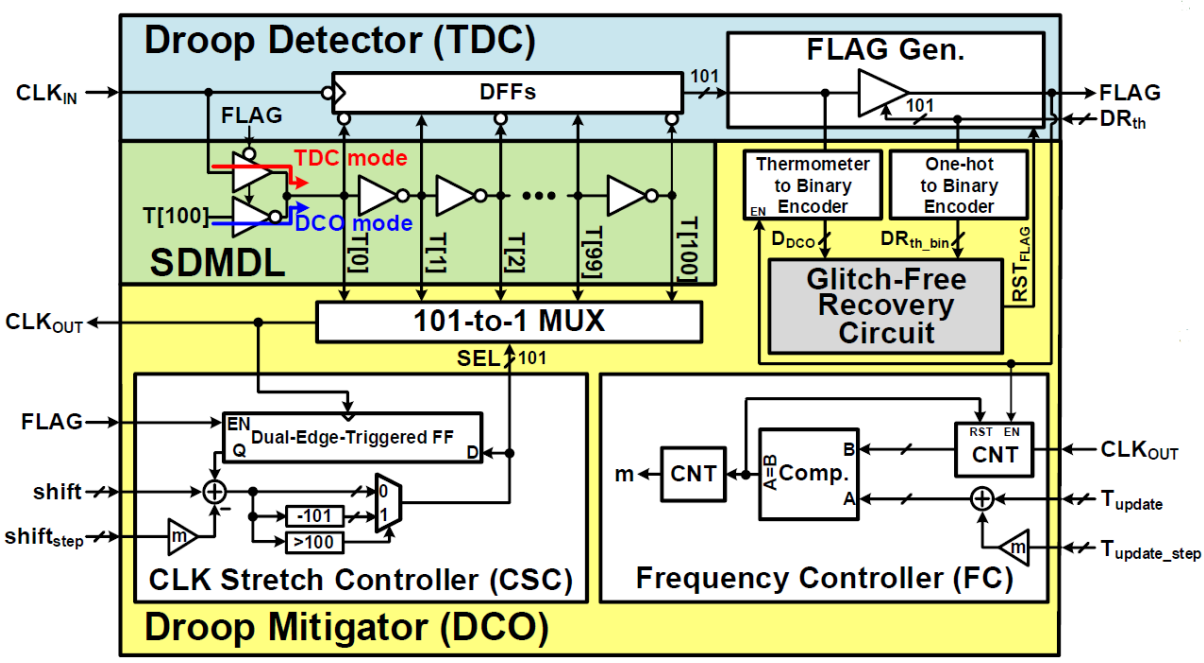
Session 20 Digital Circuit Techniques

이번 2024 CICC의 Session 20은 Digital Circuit Techniques라는 주제로 총 6편의 논문이 발표되었다. 이 세션에서는 Droop Detection 및 Protection, Hardware Security, 그리고 Post-Quantum Cryptography 주제로 발표되었으며, 해당 논문에 대해 간략하게 살펴보고자 한다.

#20.2는 Southeast University와 Shandong Yunhai Guochuang Cloud Computing Equipment Industry Innovation Co., Ltd.에서 발표한 논문으로, 28nm CMOS 공정에서 제작된 적응형 넓은 전압 범위 droop 감지 및 보호 시스템을 소개한다. 기존 고해상도 droop 감지기는 낮은 전압에서 낮은 견고성과 코드 오류를 겪는다. 이 논문은 EDAC가 지원하는 넓은 전압 범위, 고해상도, 고속 droop 감지기, 타이밍 정보를 포함하는 전압 코드 값, 신뢰할 수 있는 임계 값 자기 보정 메커니즘을 통해 성능 손실을 줄이는 방법을 제안한다. 제안된 시스템은 RISC-V 마이크로프로세서에 적용되어 0.48 ~ 1.1V의 감지 범위를 달성하며, E203 CPU의 LITTLE 코어가 0.65 ~ 0.9V, 56 ~ 400MHz에서 동작하게 한다. 링 오실레이터 기반 droop 감지기는 데이터 처리 회로와 EDAC를 통합하여 타이밍 오류를 감지하고 수정한다. 샘플링 모드 제어 모듈은 타이밍 경고에 따라 샘플링 주파수나 ECFF 모드를 조정한다. droop 보호 메커니즘은 전압 코드와 임계 값을 비교하여 타이밍 오류를 방지하며, 2단계 임계 값 자기 보정 방법을 통해 성능 손실을 줄인다. 측정 결과, 제안된 DD는 샘플링 모드 전환을 통해 2GHz에서 0.68V, 1GHz에서 0.48V까지 전압을 낮추며, 고속, 고해상도 및 견고한 전압 코드 값을 유지한다. 자기 보정 접근 방식은 0.9V, 400MHz에서 125mV 전압 증가와 0.55%의 성능 손실을 달성했다.

#20.3은 서울대학교에서 발표한 논문으로, 28nm 공정에서 제작된 고속 디지털 droop 감지 및 완화 회로를 소개한다. 이 회로는 고성능 마이크로프로세서에서 발생하는 급격한 전류 변동으로 인한 전압 droop을 감지하고 완화하여 타이밍 오류를 방지한다. 기존 아날로그 방식은 큰 면적과 추가 전원 공급이 필요하지만, 이 논문은 작은 면적과 동일한 전원을 공유하는 디지털 방식을 제안한다. 주요 기여는 공유 듀얼 모드 지연 라인을 사용한 droop 감지 및 완화, 재구성 가능한 다중 위상 디지털 제어 오실레이터로 짧은 지연 시간으로 droop을 완화하고 높은 해상도로 주파수를 최적화하며, glitch 없는 복구 회로 (GFRC)로 glitch와 전원 전압 overshoot를 방지한다. 실험 결과, 전압 droop이 발생하

면 설계된 회로가 이를 즉시 감지하고 clock 주기를 늘려 타이밍 오류를 방지하며, 복구 과정에서 주파수를 점진적으로 증가시켜 정상 상태로 복귀한다. GFRC는 두 clock 간의 위상 차이를 감시하여 안전하게 clock 소스를 전환한다. 본 논문은 droop 감지 및 완화 기술 중 가장 낮은 반 주기 지연과 60%의 가장 큰 주파수 조정 범위를 제공하며, 14.8%의 V_{min} 감소와 42.9%의 처리량 증가를 달성했다.

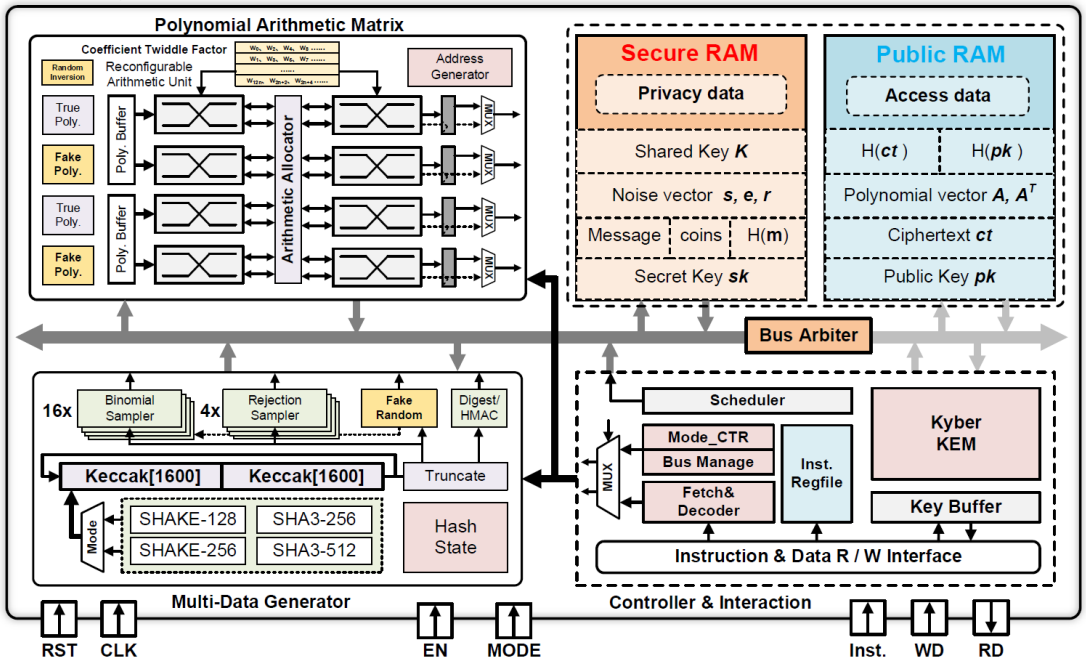


[그림 1] #20.3에서 제안한 Droop Detector와 Droop Mitigator 구조

#20.4는 Shenzhen University에서 발표한 논문으로, 높은 PVT(공정, 전압, 온도) 내성과 주파수 주입 공격에 대한 내성을 가진 저전력 current-starved ring oscillators (CSRO) 기반 난수 생성기 TRNG를 소개한다. 이 TRNG는 데이터 암호화, 키 생성 등 하드웨어 보안 관련 응용에 필수적이다. 기존의 TRNG 디자인은 높은 실리콘 면적과 낮은 에너지 효율성을 가지며, 복잡한 후처리 단계가 필요했다. 이 논문에서는 40nm CMOS 공정을 사용한 3단계 CSRO 기반의 TRNG를 제안한다. 이 논문은 전류를 제한하고 진동 소음을 증폭시켜 98fJ/bit의 초고 에너지 효율을 달성하며, 0.6V~1.3V의 넓은 전압 범위와 –65°C~140°C의 온도 범위에서 높은 Shannon 엔트로피를 유지한다. 또한, 1V의 노이즈 주입 공격에 대한 내성을 제공한다. 제안된 TRNG는 CSRO의 비표준 출력을 CMOS 인버터 기반 버퍼를 통해 표준 VDD/0으로 변환하고, D Flip-Flop을 사용해 순수 비트를 생성한다. 여러 CSRO 기반 엔트로피 소스를 결합하여 주기 전체를 활용하고, N-to-1 XOR 트리를 사용해 랜덤 비트를 연속적으로 생성한다. 이 논문은 복잡한 후처리 단계를 제거하고, FI 공격에 대한 높은 내성을 가지며, 0.6V에서 40Mb/s의 최대 처리량을 제공한다. NIST SP

800-22 및 800-90B 무작위성 테스트를 모두 통과했으며, 에너지 효율은 기존 설계보다 1.2~278배 향상되었다.

#20.5는 Huazhong University of Science and Technology에서 발표한 논문으로, 에너지 효율이 높은 CRYSTALS-KYBER 포스트 양자 암호 프로세서를 소개한다. 이 논문은 포스트 양자 암호화로의 전환에서 키 크기 증가, 복잡한 계산과 스케줄링, 사이드 채널 공격 위험성 문제를 해결하기 위해 설계되었다. 이 프로세서는 네 가지 주요 기능을 포함한다. 첫째, 재구성 가능한 산술 유닛을 갖춘 효율적인 다항식 행렬 연산. 둘째, 다양한 데이터 생성을 지원하는 콤팩트한 해시 계수 생성기. 셋째, 개인 정보 보호와 키 도난 방지를 위한 물리적으로 분리된 버스와 저장 영역. 넷째, 자가 생성 랜덤 Keccak 코어 기반의 에너지 절약 보호 방법을 채택하여 SCA를 방지한다. 40nm 기술로 제작된 이 프로세서는 0.43mm²의 면적을 차지하며, 0.65V에서 10MHz, 1.2V에서 180MHz로 작동한다. 총 302k 개의 등가 게이트와 9KB의 메모리를 사용하며, 단일 KYBER 작업당 1.26μJ/Op의 에너지를 소비한다. 이는 포스트 양자 암호화 칩 중 가장 높은 효율을 달성한다. SCA 보안은 약 1000배 향상되었다. 시스템 아키텍처는 컨트롤러, 다중 데이터 생성기, 다항식 산술 행렬, 두 개의 물리적으로 분리된 메모리로 구성된다. 컨트롤러는 칩 내부 및 외부의 데이터와 명령을 관리하며, 다중 데이터 생성기는 1600비트 상태 레지스터를 사용하여 면적과 전력 소비를 줄인다.



[그림 2] #20.5에서 제안한 양자 암호와 프로세서의 전체 구조

#20.6는 MIT와 IBM T.J. Watson Research Center에서 발표한 논문으로, 사이드 채널 공격

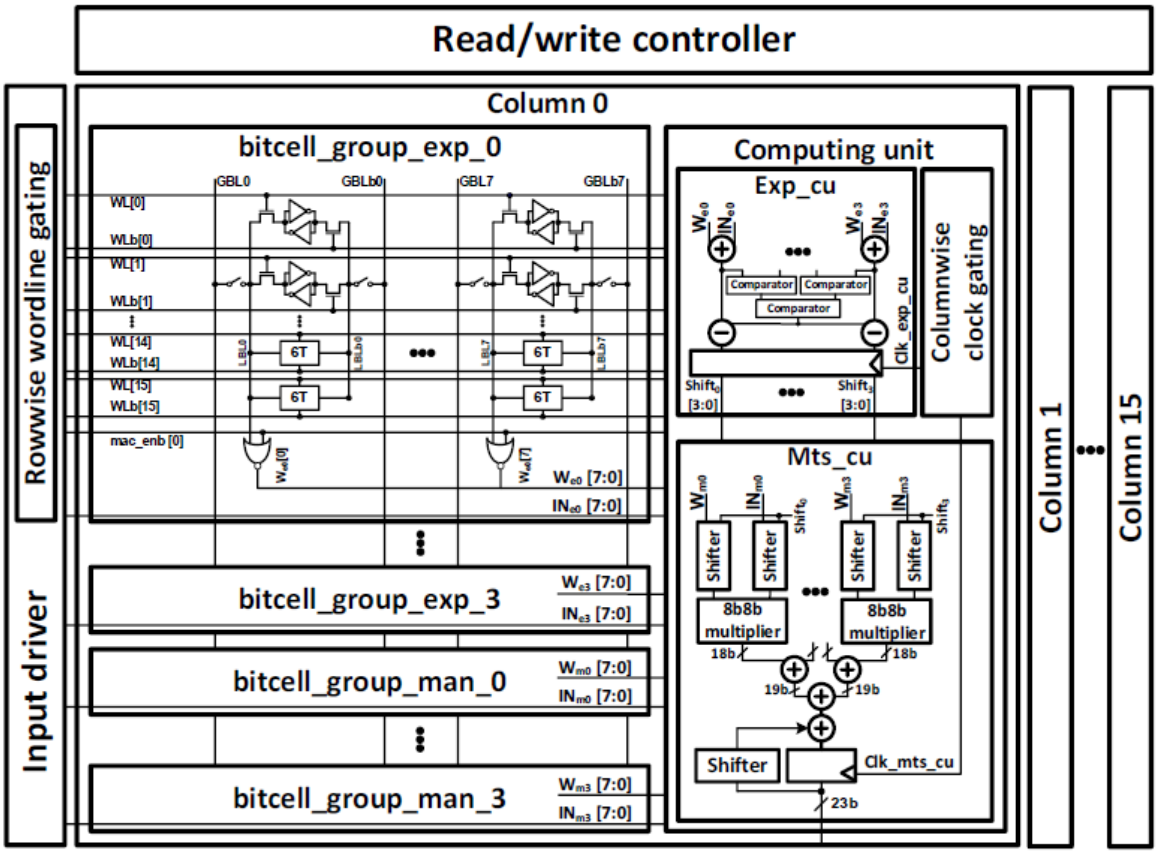
(SCA)과 버스 프로빙 공격(BPA) 보호 기능을 갖춘 안전한 디지털 인-메모리 컴퓨팅(IMC) 매크로를 소개한다. 디지털 IMC는 데이터 전송 에너지를 줄이면서도 높은 비트폭과 정확성을 유지하기 위해 제안되었다. ML 워크로드의 프라이버시는 SCA와 BPA에 의해 악용될 수 있으며, 이 논문은 이러한 위험을 완화하는 IMC 매크로를 제안한다. 논문은 세 가지 주요 과제를 해결한다. 첫째, 기존 기술의 오버헤드를 피하면서도 높은 병렬성을 유지하고 보안성을 확보하는 SCA-secure MAC 연산 방식을 제안하였다. 둘째, NIST 표준 경량 ASCON 암호를 사용하여 오프-칩 저장 및 전송 중 가중치가 평문으로 존재하지 않도록 보장하여 BPA 보안을 강화하였다. 셋째, IMC 메모리를 재사용하여 키 생성에 피드백 컷 Physical Unclonable Function를 도입하였다. 논문에서는 난수 생성기가 필요 없는 비트 직렬 곱셈을 수행하는 공유 계산 기술을 제안하며, XNOR을 사용한 곱셈으로 보안성을 유지하면서 계산 지연을 최소화하였다. CSA 트리를 사용해 부분 곱을 집계하고, 마지막 몇 단계의 덧셈을 비트 직렬 축적과 동시에 수행하여 지연을 줄였다. 이 논문은 14nm CMOS 기술로 구현되었으며, 0.50V에서 8.1 TOPS/W의 성능을 달성하였다. 이 논문은 운영 중 난수 생성기가 필요 없고 정확도 제한이 없는 최초의 IMC 매크로로, 다양한 ML 애플리케이션에서 프라이버시를 보장하는 안전한 IMC 솔루션을 제공한다.

Session 26 Digital Compute in Memory

이번 2024 CICC의 Session 26은 Digital Compute in Memory라는 주제로 총 5편의 논문이 발표되었다. 이 세션에서는 deep neural network acceleration, integer linear programming solving, 그리고 edge AI applications를 위한 가속기가 발표되었다. 이번 후기를 통해 5개의 논문에 대해 간략하게 살펴보고자 한다.

#26.1은 Columbia University에서 발표한 논문으로, 에너지 효율, 연산 밀도, 가중치 밀도를 모두 개선시킨 SRAM 기반의 디지털 CIM이다. 본 논문은 BF16을 지원하고 있으며, 앞서 말한 3가지에 대해 state-of-the-art를 달성하기 위하여 다음과 같은 아이디어를 제시했다. 첫번째로, 에너지 효율, 연산 밀도, 가중치 밀도 이 세가지의 관계를 정리하여 16개의 가중치가 하나의 곱셈기를 공유하는 것이 최적의 디자인 포인트임을 밝혀냈다. 두번째로, 연산 시 누적전에 자릿수를 맞추는 일반적인 방식과 달리 곱셈점에 자릿수를 맞추는 방법을 채택하여 최소한의 오차로 연산 밀도와 가중치 밀도를 향상시켰다. 마지막으로 0이나 무시할 수 있을 정도로 작은 입력 및 가중치에 대해 열 방향으로는 클락 게이팅, 행 방향으로는 워드라인 게이팅 기술을 제안하여 에너지 효율을 약 2배 가량 향상시켰다. 이를 통해 에너지 효율, 연산 밀도, 가중치 밀도를 곱했을 때 이전 논문 대비 17.7배 높은 Figure-of-Merit (FoM)을 달성했으며, CIFAR100용 ResNet18을 매핑하는 동안

77.36%의 추론 정확도를 달성했다.

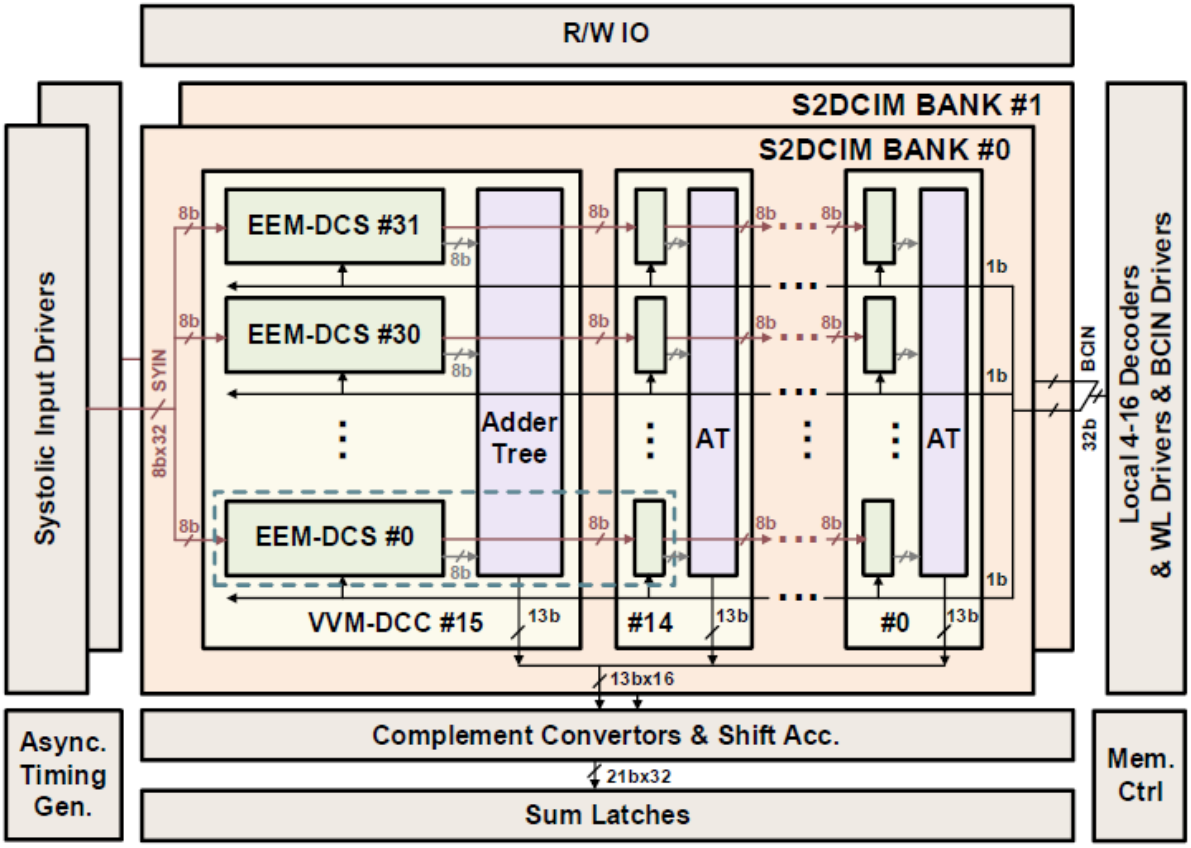


[그림 3] #26.1에서 제안한 CIM의 전체 구조

#26.2은 Johns Hopkins University와 Cornell Tech에서 발표한 논문으로, 매우 희소한 DNN 워크로드를 위한 가변 희소성 표현을 갖춘 CIM을 소개한다. 본 논문은 CIM 설계에서 희소성 압축 기법을 도입하여 저장 용량과 연산 효율을 개선했다. 특히 세가지 인기 있는 희소성 압축 형식인 COO(Coordinate Representation), RL(Run Length Encoding), N:M 희소성을 통합한 CIM를 설계하여, 저장 밀도를 높이고 에너지 효율을 극대화하였다. CIM 매크로는 64x128 비트 셀로 구성되며, 16개의 컬럼 그룹(CG)과 32개의 로우 그룹(RG)으로 나뉜다. 각각의 RG는 16개의 비트 셀을 가지고 있으며, 곱셈 디코드 및 비교 블록에 의해 두 개의 8비트 행으로 나뉜다. 10T 비트 셀은 입력 활성화와 저장된 가중치의 AND 연산을 병렬로 수행하여 입력 활성화 병목 현상을 줄였다. 또한, 이 논문은 COO, RL, N:M 희소성 압축 형식에서 직접 가중치를 처리하여 에너지 효율성을 크게 향상시킨다. 이를 통해 기존 비희소 설계에 비해 3-40배의 면적 감소를 가져오며, 최초로 COO, RL, N:M 압축 형식을 모두 지원하고 다양한 DNN 워크로드에 대해 높은 성능을 제공한다.

#26.3은 Columbia University에서 발표한 논문으로, 온라인 압축 및 압축해제를 사용하여 BF16 지원이 가능한 디지털 기반의 CIM이다. 본 논문은 DNN 가속기의 오프-칩 메모리 접근 에너지 소비 문제를 해결하기 위해 세 가지 새로운 기술을 제안하였다. 첫번째로, 오프-칩 데이터의 이동을 줄이기 위해 CPR4라는 오프라인 압축 알고리즘을 개발하였다. 이는 BF16 가중치를 16개의 그룹으로 묶어 각각 4b 인덱스로 인코딩한다. 압축된 가중치 데이터를 오프칩 메모리에서 가져와 실시간으로 BF16으로 압축해제해 오프-칩 데이터 접근 에너지를 4배 줄인다. 둘째, CPR4 기반의 온라인 압축 하드웨어를 개발하여 각 층의 FP32 활성화 값을 압축한 후 메모리에 저장한다. 이를 통해 온-칩 활성화 메모리 요구량을 4배 줄인다. 셋째, 정확한BF16 MAC 연산을 지원하는 CIM을 개발하였다. 제안된 가속기는 오프-칩 데이터 접근 에너지를 포함하여 1 TFLOPS/W의 에너지 효율을 달성하였다. VGG16 기반 추론 작업에서 127MHz의 동작 클럭 주파수를 유지하며, 기존의 DNN 가속기와 비교하여 1.96배 이상의 에너지 효율을 보여주었다.

#26.4는 Peking University에서 발표한 논문으로, 엣지 AI 애플리케이션을 위한 22nm 128Kb Systolic Digital Compute-in-Memory (S2D-CIM)를 소개한다. 본 논문은 유연한 벡터 연산과 2D 가중치 업데이트를 통해 에너지 효율과 면적 효율을 개선하는 설계를 제안하였다. 이 논문은 여러 단계 도미노 데이터 경로를 설계하여 각 열에 대해 개별 입력 벡터를 지원함으로써 벡터-벡터-곱셈 연산의 유연성을 높였다. 또한, 비동기 타이밍 기법을 통해 작업량에 따라 입력 데이터 흐름과 활성화된 열을 적응적으로 조정하여 에너지 효율성을 향상시켰다. 비동기 타이밍 제어 논리와 도미노 데이터 경로 셀은 다양한 벡터 연산을 지원한다. Systolic 모드에서는 최대 16개의 8b 입력이 16개의 연산기로 전송되며, Broadcast 모드에서는 단일 벡터가 비트 직렬로 전송되어 연산을 수행한다. S2D-CIM의 유연한 데이터 흐름 조합은 효율성을 크게 향상시켰다. 이 논문은 다른 최첨단 SRAM 기반 디지털 CIM과 비교하여 1.67배 향상된 효과적인 에너지 효율성을 보여주었다. Broadcast 모드와 Systolic 모드의 결합 아키텍처는 다양한 뉴럴 네트워크 모델에 대해 단일 데이터 흐름 대비 1.25-2.84배 향상된 에너지 효율성을 제공하여 엣지 AI 애플리케이션의 다양한 연산 요구를 충족했다.



[그림 4] #26.4에서 제안한 S2D-CIM의 전체 구조

#26.5는 The University of Texas at Austin에서 발표한 논문으로, 임의 비트 정밀도 디지털 인-메모리 컴퓨팅 기반 Integer Linear Programming (ILP) solver를 소개한다. 본 논문은 ILP 문제를 효율적으로 해결하기 위해 8T SRAM 셀을 활용한 CIM 기반 solver를 제안하였다. 8T SRAM 셀을 최소한의 주변 장치 변경으로 ILP 계산에 재활용하여 재구성 가능성을 유지하고, 메모리 내 제약 조건 검사, 목적 함수 계산, 변수 업데이트를 통해 데이터 이동을 최소화했다. 또한, 임의 비트 정밀도 지원을 통해 다양한 규모와 정밀도에 적응할 수 있는 재구성 가능한 계산 유닛을 설계하고, 사용자 정의 알고리즘을 지원하는 재구성 가능한 제어 및 업데이트 구성 요소를 제공하여 ILP 해결의 타당성과 최적화 단계를 모두 지원했다. 대규모 ILP 문제를 지원하는 메가비트 수준의 캐시를 활용한 확장 가능한 설계를 했다. 본 논문은 기존 소프트웨어 solver에 비해 최대 471배의 클럭 사이클 절감을 달성하였으며, FPGA 기반 가속기에 비해 최대 7.3배의 클럭 사이클 절감과 106배의 에너지 절감을 이뤘다.

저자정보

엄소연 박사과정 대학원생

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