2023 IEEE CICC Review

IEEE Custom Integrated Circuits Conference

고려대학교 반도체시스템공학과 박사과정 김현진

Session 7 Compute in Memory and Ising Machines

이번 2022 IEEE CICC의 Session 7은 Compute in Memory and Ising Machines 라는 주제로 총 8편의 논문이 발표되었다. 이 세션에서는 compute in memory 구조 및 Ising machine 를 사용하여 하드웨어적으로 알고리즘의 성능 최적화를 이끌어 내는 것에 중점을 두었고, 주로 기존 기술들이 가진 문제점을 해결하는 것을 목표로 새로운 아날로그 및 디지털 회로 기술을 사용하여 기존 기술들보다 연산 성능 혹은 하드웨어 면적 소모량을 개선시켰다. 올해는 최근 연구가 많이 진행되고 있는 eDRAM compute-in-memory 분야의 논문들 뿐만 아니라, 극저온 in-memory-computing macro, GNN 가속기, 아날로그 및 디지털 Ising machine 등 다양한 아키텍처의 논문들이 발표되었다는 것이 주목할 만하다.

#7-1 은 calibration-free 15-level/Cell eDRAM computing-in-memory (CIM) macro를 제안하는 논문이다. 이 논문에서는 현재의 multi-level-cell (MLC)-eDRAM CIM의 성능이 프로그래밍과 컴퓨팅 사이에서 발생되는 weight 값의 불일치로 심각하게 제한되고 있다는 것을 먼저 지적하고 있다. 기존의 current-based CIM의 문제점을 해결하기 위해서 본 논문의 CIM cell은 programming과 computing 모두를 current domain에서 진행하는 구조로 설계되었다. 이러한 CIM cell은 programming을 bias current branch를 사용해서 진행하므로, self-calibrated 된 게이트-소스 전압 값이 각각의 capacitor cell에 데이터로 저장된다. 따라서, self-calibrated 된 게이트-소스 전압은 트랜지스터의 I-V 비선형성 및 V_{TH} variation으로 발생하는 eDRAM cell의 nonideality를 완화해준다. 또한, 본 논문은 dynamic-cascoded read 구조를 통해서 V_{RBL} 에 대한 I_{comp}의 민감도를 4배 감소시켰으며, voltage-current two-step write driver를 통해서 정확도를 유지한 채로 프로그래밍 속도를 증가시켰다. 결과적으로 본 논문은 위와 같은 기술들을 사용하여 sub-μA current MLC 구조를 설계하였고, 233-305 TOPS/W의 높은 에너지 효율을 달성하였다.

#7-2 은 극저온 컴퓨팅을 위한 새로운 in-memory-computing (IMC) macro인 CIMC를 제안하는 논문이다. CIMC는 3가지 효과적인 기술을 사용하여 종래의 IMC macro가 극저온 환경에서 사용될 경우 발생되는 문제들을 해결한다. 먼저, 본 논문은 극저온 3T (C3T) bitcell을 통하여 full-swing 데이터 전송을 구현하였고, 이를 통해서 charge injection 등으로 생기는 기존 eDRAM의 신뢰성 문제를 해결하였다. 둘째로, adaptive reference sense amplifier (ARSA)를 사용하여 극저온 Boolean 함수들을 정확하게 구현하였다. 마지막으로, 본 논문은 최적화된 flash ADC를 사용하여 극저온 환경에서도 빠르고 저전력의 convolution 결과를 processing하였다. 극저온 환경에서의 측정 결과, CIMC는 위의 기술들을 통하여 603.1TOPS/W의 높은 평균 에너지 효율성과 284TOPS/mm^2의 높은 평균 연산 밀도를 달성하였다.

#7-3 은 많은 양의 메모리를 요구하는 인공 지능 작업의 메모리 병목 현상을 해결하기 위한 에너지-효율적인 접근 방식인 CIM(Compute-in-memory)에 관한 논문이다. 이 논문에서는 데이터 sparsity를 활용하여 높은 에너지 효율성을 달성하는 Double-Mode CIM(DM-CIM) macro를 제안한다. DM-CIM macro는 단일 계층 모드(SLM) 또는 이중 계층 모드(DLM)로 재구성될 수 있으며, SLM의 경우 charge-domain (CD) MAC array + VG ADC로 동작하고, DLM의 경우 charge-domain MAC array + VTC & TA + Time-domain MAC array로 동작하게 된다. 또한, 8T1C bitcell 배열이 data sparsity로 인한 많은 동적인 전력 손실을 감소시켜주며, ReLU-embedded VTC와 프로그래밍 가능한 선형성이 개선된 time amplifier (TA)를 통하여 CIM의 성능을 향상시켰다. DM-CIM macro는 28nm CMOS 공정에서 제작되었으며, SLM CIM macro는 CIFAR10에서 90.11% ~ 90.64%의 추론 정확도를 달성하였다.

#7-4 은 최근 소프트웨어 상으로 연구가 많이 진행되고 있는 graph neural networks (GNN)용 compute-in-memory (CIM) macro와 그를 이용한 가속기에 관한 논문이다. 논문에서는 대부분의 GNN 모델에 적용 가능한 통합된 4단계 프레임워크 및 아키텍처를 제안하고, 이에 사용되는 GNN 가속기 하드웨어를 세계 최초로 발표한다. 또한, 아날로그-디지털 하이브리드 CIM 구로를 통하여 throughput과 power 성능을 조절 가능한 matrix vector multiplication (MVM) 구조를 제안하였으며, ultra-high sparsity 그래프에서도 효율적인 연산이 가능하도록 ternary-search content addressable memory (TsCAM) macro를 제안하였다. 마지막으로, 본 논문에서는 adaptive mapping을 통하여 순차적인 메모리 접근이 가능하도록 하는 두 가지 dataflow를 발표하였다. 결과적으로, 본 논문은 기존 가속기들 대비 28.1 ~ 78.6배의 성능 향상을 보였다.

#7-5 에서는 종래의 딥러닝 어플리케이션을 위한 가속기가 겪는 문제들을 sparsity-aware compute-in-memory (CIM) 코어를 통하여 해결한 논문이다. 먼저, 본 논문은 sparsity controller와 reconfigurable shift and add를 통해서 input과 weight의 sparsity를 leverage 해주었고, 어플리케이션의 SNR 요구사항을 고려하여 WL을 adaptive하게 구성하는 방법으로 아날로그 CIM 구조의 정확도를 유지하였다. 또한, 본 논문은 sparse compute unit (SCU) 간의 load balancing을 통해서 요구되는 ADC의 resolution을 감소시켰다. 제안된 CIM 코어는 1.4 ~ 6.7 TOPS/W의 에너지 효율성을 달성하였고, baseline보다 2배의 성능 증가 및 1.78배의 에너지 효율성 증가를 달성하였다.

#7-6 은 TinyML을 사용하는 edge 딥러닝을 위한 디지털 in-memory 연산 기반의 microcontroller unit (MCU)에 관한 논문이다. 이 논문에서는 가속화 목표와 연산 flow를 고려하여 설계된 iMCU라는 면적 효율적인 MCU를 제안한다. 제안된 iMCU는 SPIKE를 통하여 모든 계층의 연산 비중을 분석하여, 가장 하드웨어적인 부담이 큰 convolution 계층과 addition 계층만을 가속화하였다. 또한, iMCU 가속기는 기존의 MCU 가속기들과는 다르게, 한번에 가속기가 연산하는 하나의 계층만을 buffering하여, 5배의 하드웨어 면적을 감소시켰다. 따라서, 본 iMCU는 기존의 최고 성능 MCU인 xG24-DK2601B보다 88배 더 나은 성능을 발휘하면서, PVT-robust한 연산을 제공하였다.

#7-7 에서는 기존의 Ising machine의 한계를 극복하기 위한 새로운 구조의 연속-시간 아날로그 Ising machine을 제안하고 있다. 이 머신은 inverter-chain 기반의 연속-시간 작동 방식을 사용하여 계산 에너지와 지연 시간을 크게 줄일 수 있으며, 외부의 random number generator (RNG) 없이도 inverter-chain 구조의 equalization 및 interaction 동작을 통해서 랜덤성 있는 연산이 가능하다. 본 Ising machine은 1920개의 spin들을 사용하여 Max-Cut 문제를 30 ns 안에 풀었으며, 본 논문에서 제시된 spin은 기존 논문들과 비교하여 224 μm^2의 가장 작은 하드웨어 면적이 요구된다. 반면에, 본 논문에서 제시하는 Ising machine은 spin 간의 coefficient이 +1, 0, -1으로 비교적 작은 범위만 가능하고, 더 복잡한 coefficient가 필요한 문제를 연산할 경우에는 process-variation으로 인한 inverter 간의 mismatch 등에 의하여 성능이 떨어질 수 있으므로, digital 기반의 Ising machine들과 직접적으로 성능 비교를 하기에는 특성이 다른 부분이 있다고 생각된다.

#7-8 은 복잡한 조합 최적화 문제를 위한 재구성 가능한 Ising machine을 제안한다. 기존의 Ising machine들은 2-body 구조로 하드웨어가 설계되었으므로, 3SAT 문제와 같은 many-body 문제를 mapping 및 연산하는데 성능 및 효율성이 떨어졌다. 따라서 본 논문은 3-body가 직접적으로 mapping 가능한 Ising machine을 소개한다. 본 논문에서 소개된 하드웨어는 King's 그래프를 기반으로 128개의 재구성 가능한 processing element (PE)

를 사용하며, 각 PE는 8개의 스핀으로 구성된다. 따라서, 각 PE는 King's 그래프를 기반으로 최대 56개의 spin interaction을 구현할 수 있다. 본 논문에서는 기존의 2-body Ising machine을 사용하여 approximate mapping을 진행한 방법보다 제안된 reconfigurable Ising machine이 11.9배의 적은 스핀 개수가 필요하고 문제 해결 성능이 83% 상승하게 되었다. 다만, area 기존의 many-body 문제를 2-body 문제로 변환시켜주는 방법이 논문에서는 자세하게 언급되지 않았는데, spin마다 3-body까지 계산을 지원하게 된다면 기존 2-body까지 연산 가능한 spin 대비 spin 자체의 하드웨어 크기가 증가하므로, spin 개수 비교만으로 machine의 정확한 성능 증가율을 규정하기는 어렵다고 생각된다.

저자정보

명예기자 김현진

- 소　　속 : 고려대학교 반도체시스템공학과 박사과정
- 연구분야 : PMIC & Ising Machines
- 이 메 일 : jamespul@korea.ac.kr
- 홈페이지 : https://kilby.korea.ac.kr

2023 IEEE CICC Review

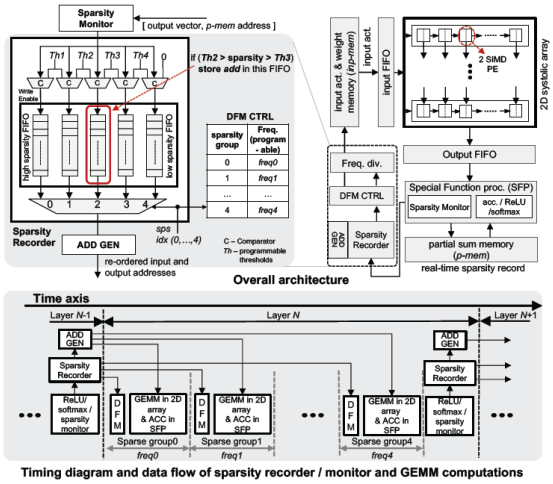
IEEE Custom Integrated Circuits Conference

포항공과대학교 전자전기공학부 박사과정 변영훈

Session 20 Machine Learning

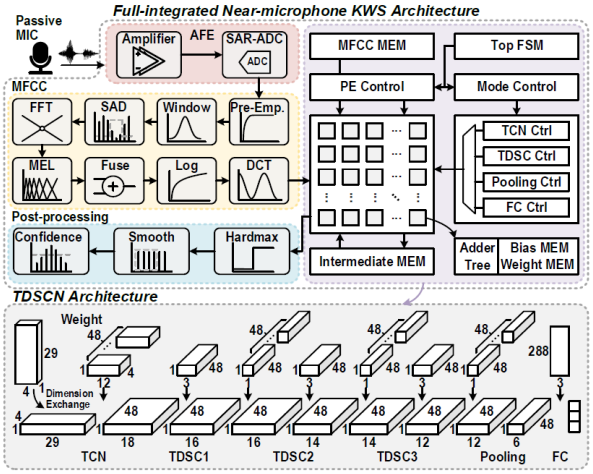
이번 2023 IEEE CICC의 Session 20은 Machine Learning이라는 주제로 총 6편의 논문이 발표되었다. 이 세션에서는 다양한 딥 러닝 모델들을 구동할 수 있는 ASIC 및 알고리즘들을 다룬다.

#20-1 본 논문에서는 CNN과 Transformer모델 동작 시 weight와 activation의 sparsity를 특정 threshold와 비교해 input sequence를 reordering하고, cluster별로 다른 clock frequency를 사용하는 기술을 제안한다. 이 때 real-time input sparsity를 파악하기 위해 이전 layer의 분포를 참고하는 방식으로 unstructured sparsity에 대응하는 방식을 활용한다. 해당 칩은 input sequence의 sparsity가 0~100%까지 범위에서 존재할 때 8-bit 기준 500~800MHz로 동작하였고, 0.6 – 1.0TOPS/W의 efficiency를 달성하였다. 이 때 sparsity recorder 모듈은 전체 프로세서에서 7% power overhead만을 차지하였다.



[그림 1] Sparsity-adaptive dynamic clock frequency modulation의 architecture와 동작 과정

#20-2 본 논문에서는 wearable이나 mobile device에서의 real-time keyword-spotting을 위한 통합 칩을 제안한다. 칩은 analog front end, MFCC feature extractor 및 temporal depthwise separable CNN (TDSCN) 모듈로 구성되며, serial FFT와 genetic algorithm (GA)을 통해 성능을 최적화하였다. 또한 TDSCN hardware를 구현할 때 parallel adder tree 기반 near-memory computing PE를 사용해 memory access 횟수를 줄였다. 해당 칩은 28nm CMOS공정으로 제작되어 0.47mm^2 면적을 차지하며 0.36-V, 8-KHz clock에서 전력 소모량 608nW를 달성하였다. 해당 칩을 사용한 결과 Google speech command dataset 기준 하나의 키워드에 대해 98.1%, 두 개의 키워드에 대해 95.8%의 정확도를 달성하였다.



[그림 2] Fully-integrated near-microphone KWS architecture (위) 와 TDSCN architecture (아래)

#20-3 본 논문에서는 최근 AI SoC design 과정에서 발생하는 다양한 문제점들 중 중요하게 여겨지는 네 가지 주제에 대해 분석하고 있다. 다루고 있는 내용은 1) Foundation model의 high-performance 학습, 2) Foundation model 학습을 위한 가속기의 dataflow 관리, 3) Tensor, model, data-level parallelism등 여러 차원을 포함하는 병렬 연산 방법, 4) Dataflow accelerator에서 data locality나 computation을 고려한 memory unit이나 energy design trade-offs인데, 각각의 분야에 대한 background부터 주요 연구 소개, 그리고 해당 연구에 대한 장단점이 서술되어 있다.

#20-4 본 논문에서는 dynamic precision을 활용해 딥 러닝 학습에서 높은 정확도와 효율성을 보여주는 training processor를 제안하고 있다. Dynamic precision은 channel이나 layer와 같이 특정 단위 별로 sensitivity를 estimation해 accuracy loss를 최소화 하면서도 연산의 효율을 올리는 방법이다. 기존 프로세서에서는 1) dynamic precision control로 인한 추가 overhead 발생, 2) low-bit FP에서 power bottleneck problem으로 인한 dynamic

precision의 scalability 한계, 3) irregular precision으로 인한 storage나 memory I/O의 낭비로 인해 dynamic precision training (DPT)로 얻을 수 있는 이득이 크지 않았다. 제안하는 DPT processor는 1) sensitivity analyzer 및 micro inst online codegen을 사용해 sensitivity search time 감소, 2) multi-level aligned BFP unit에서 INT기반 adder를 사용, 3) DPT 전용 스토리지 형식 및 I/O routing 방법을 사용함으로써 DPT를 사용해 얻을 수 있는 이득을 극대화 하였다. TSMC 28nm 공정으로 제작된 칩은 ImageNet 평균 4-bit precision으로 1%이하의 정확도 손실을 가지며 기존 4-bit 학습 대비 5.47% 더 높은 효율을 보여주었다.

#20-5 본 논문에서는 Spiking neural network (SNN)와 Artificial neural network (ANN)를 모두 연산할 수 있는 구조를 가진 in-memory neuromorphic computing chip을 제안한다. SNN의 경우 저 전력으로 동작 가능하지만 정확도가 낮고, ANN의 경우 정확도는 높지만 에너지 소모량이 크다는 점에 착안해 두 방식을 동적으로 사용해가며 높은 정확도와 낮은 에너지 소모량만으로도 기존의 ECG abnormal detection이나 voice activity detection과 같은 task에서 95%이상의 inference accuracy를 보여준다. SRAM에서 in-memory computing은 sparsity를 고려해 row 또는 column단위로 power gating을 하거나 ring-based converter를 활용하는 방식을 통해 spike의 sparsity에 power adaptive하게 동작한다. 제안하는 칩은 0.43 pJ/SOP의 energy efficiency를 달성하였다.

#20-6 본 논문에서는 dynamic workload allocation과 heat map compression/pruning 기법을 사용해 설명 가능한 Artificial intelligence (XAI)를 동작시킬 수 있는 AI processor를 제안한다. 이미지 인식과 같은 task에서 class activation mapping (CAM)과 같은 기법은 AI가 판단한 근거를 확인할 수 있는 방법 중 하나인데, 이를 위해 기존 inference에서 사용하지 않던 heat map이나 backward propagation 연산이 추가로 필요로 하게 된다. 이와 같은 연산을 최적화하기 위해 이 논문에서는 전체 gradient를 floating point로 계산하는 대신 특정 target class의 gradient만을 fixed-point로 연산하는 방식을 사용한다. 또 inference explanation scheduling이나 compressed format, 그리고 point-wise gradient pruning을 활용하였고, 최종적으로 XAI task에서 26.55TOPS/W를 달성하였다.

저자정보

명예기자 변영훈

- 소속 : 전자전기공학과
- 연구분야 : Deep learning model compression
- 이메일 : byh1321@postech.ac.kr
- 홈페이지 : sites.google.com/view/epiclab/member/yhbyun