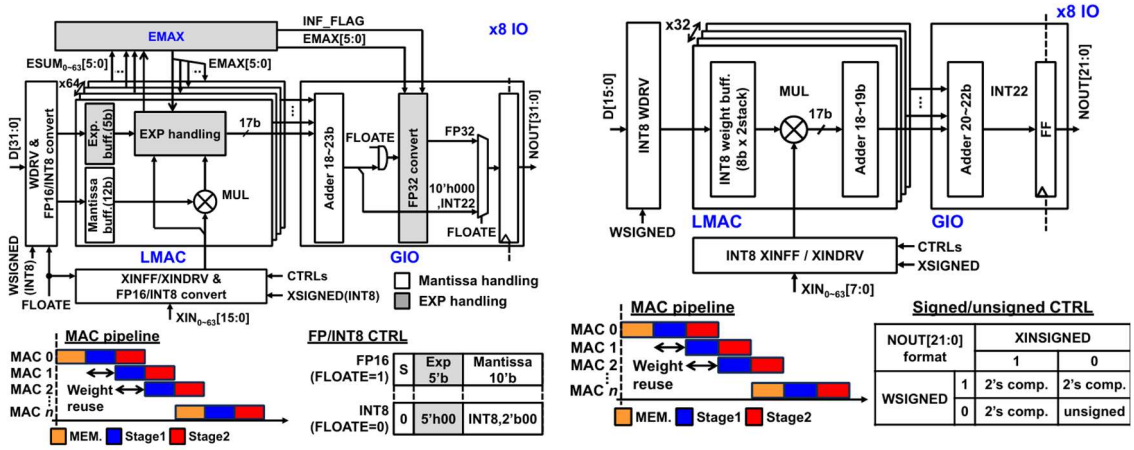
2025 IEEE VLSI Review

포항공과대학교 반도체대학원 박사과정 박은빈

Session 1 CIM and Quantum-inspired Computing

이번 VLSI 2025의 Session C1 CIM and Quantum-inspired Computing은 인메모리 컴퓨팅의 확장성과 양자 영감형 아키텍처의 진화를 동시에 보여준 세션이었다. 발표된 연구들은 초미세공정 기반의 디지털 CIM부터 뉴로모픽 SNN 전용 구조, 그리고 대규모 이징 머신까지 폭넓게 다루며, 공통적으로 데이터 이동 최소화와 계산 효율 극대화를 목표로 하고 있었다. 특히, 최근의 요구인 LLM·SNN과 같은 응용 특화 과제와 NP-hard 최적화 문제를 겨냥해, CIM은 더 높은 정밀도와 재구성 가능성을 확보하고, 이징 머신은 대규모 스핀 연결성과 멀티칩 확장성을 통해 기존 한계를 뛰어넘고자 했다. 종합적으로 이번 세션은 CIM이 범용 가속기에서 응용 맞춤형 설계로 확장되고, 양자 영감형 접근이 실용적인 최적화 솔루션으로 자리잡아가는 흐름을 잘 보여주었다.

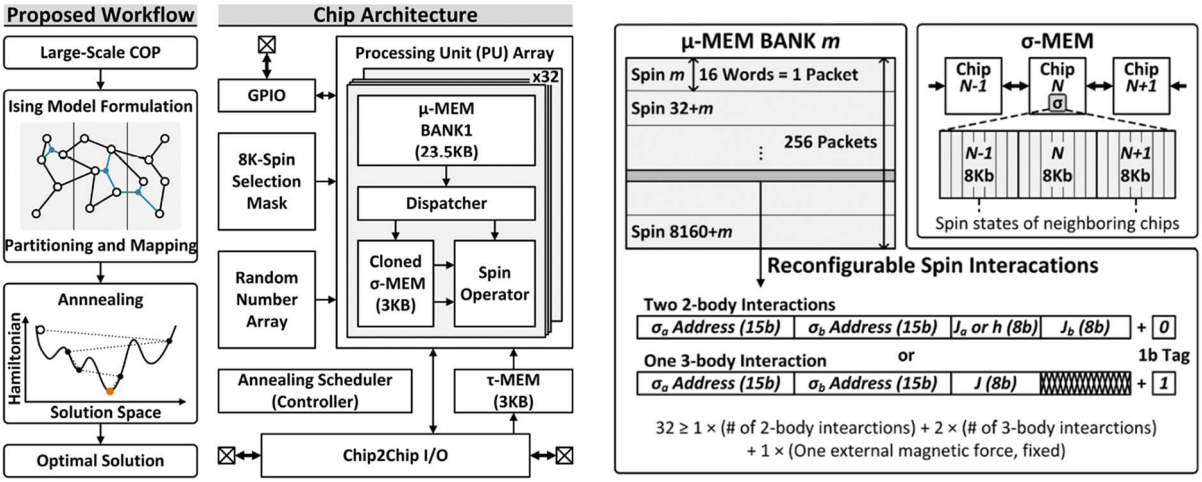
#C1-1은 TSMC 연구팀에서 발표한 3nm 공정 기반 SRAM Digital-CIM Compiler로, 기존 CIM과 달리 INT8과 FP16을 모두 지원하여 범용성과 유연성을 크게 확장한 것이 특징이다. 특히 FP16 연산에서 기존의 Alignment-First (AF) 방식 대신 Multiply-First (MF) 방식을 채택해 연산 정확도를 크게 개선하였으며, weight update를 주기당 여러 번 수행하는 multi-weight update per cycle 기법을 통해 메모리 대역폭 병목을 완화하고 활용도를 극대화하였다.



[그림 1] DCIM complier design의 전반적인 구조 및 동작원리

이러한 구조적 최적화 덕분에 INT8 모드에서 124.6 TOPS/W, FP16 모드에서 28.6 TFLOPS/W라는 세계 최고 수준의 에너지 효율을 달성하였다. 다만 상대적으로 간단한 벤치마크 환경에서의 검증에 머물렀다는 점은 아쉬움으로 남지만, 초미세공정 기반 범용 CIM이 실용화 단계에 진입했음을 보여주는 대표적 성과라 할 수 있다.

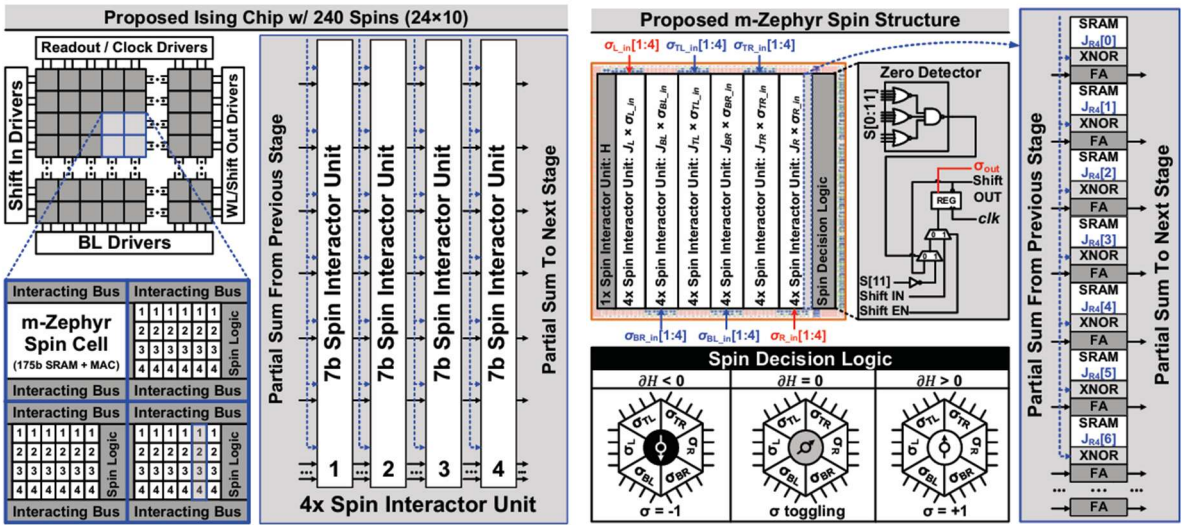
#C1-4는 POSTECH에서 발표한 28nm 공정 기반 8K-Spin Ising Machine IC로, 기존 이징 머신들이 주로 2체 상호작용(two-body)에 한정된 것과 달리 최대 31개의 2·3체 상호작용(many-body interactions)을 지원하는 것이 큰 특징이다. 또한 다수 칩을 직렬로 연결해 사실상 무한 확장이 가능한 limitless 1D multichip extension 구조를 도입해, 단일 칩의 한계를 넘어 대규모 최적화 문제를 직접 풀 수 있도록 설계되었다. 칩 내부에서는 σ-MEM과 μ-MEM을 기반으로 스핀 상태와 상호작용을 병렬적으로 처리하며, 스토캐스틱 셀룰러 오토마타(SCA)와 비율 제어 병렬 어닐링(RPA) 알고리즘을 지원해 빠른 수렴을 가능하게 했다. 실제 측정에서는 4개의 칩을 연결해 총 32,760 변수를 포함하는 대규모 3SAT 문제를 풀어내며, 기존 이징 머신 대비 문제 규모와 확장성에서 가장 앞선 성능을 입증하였다. 다만 복잡한 상호작용 지원을 위해 비교적 큰 하드웨어 자원이 요구된다는 점은 한계로 남지만, NP-hard 문제 해결을 위한 양자 영감형 컴퓨팅의 실질적 확장 가능성을 보여준 점에서 중요한 성과라 할 수 있다.



[그림 1] 본 논문에서 제안한 대규모 최적화문제를 풀기위한 workflow 및 Ising model 3차 하드웨어 구현

#C1-5는 UCSB와 KAIST가 공동으로 발표한 m-Zephyr Digital In-Memory Ising Chip으로, 기존 이징 머신의 제한된 연결성을 극복하기 위해 수정된 3D Zephyr 토폴로지를 도입하여 스핀 당 24개의 상호작용을 지원하는 것이 특징이다. 이는 lattice나 King's graph 기반 구조보다 훨씬 복잡한 문제를 직접적으로 매핑할 수 있게 하여, 재구성 가능성과 표

현력을 크게 높였다. 또한 7비트 정밀도의 완전 디지털 인메모리 연산을 통해 정확도를 유지하면서도 <180ns의 빠른 수렴 시간과 <350nJ의 초저전력 동작을 달성하였다. 실제 측정 결과, 기존 Metropolis 알고리즘 대비 106배 빠른 문제 해결 속도와 더 낮은 평균 Hamiltonian 값을 보여주며, NP-hard 최적화 문제에서의 실효성을 입증하였다. 다만 65nm 공정 기반이라 최신 초미세공정보다 집적도는 낮지만, 연결성·정밀도·에너지 효율의 균형을 모두 잡은 디지털 이징 머신의 진화형 아키텍처라는 점에서 의의가 크다.



[그림 1] 제안한 m-Zephyr Ising chip의 전반적인 구조 및 schematic diagram

저자정보

박은빈 박사과정 대학원생

- 소속 : 포항공과대학교
- 연구분야 : 임베디드 시스템 및 지능형 반도체
- 이메일 : eunbin@postech.ac.kr
- 홈페이지 : https://sites.google.com/view/epiclab/member/ebpark

2025 IEEE VLSI Review

한국과학기술원 전기및전자공학부 석사과정 권재훈

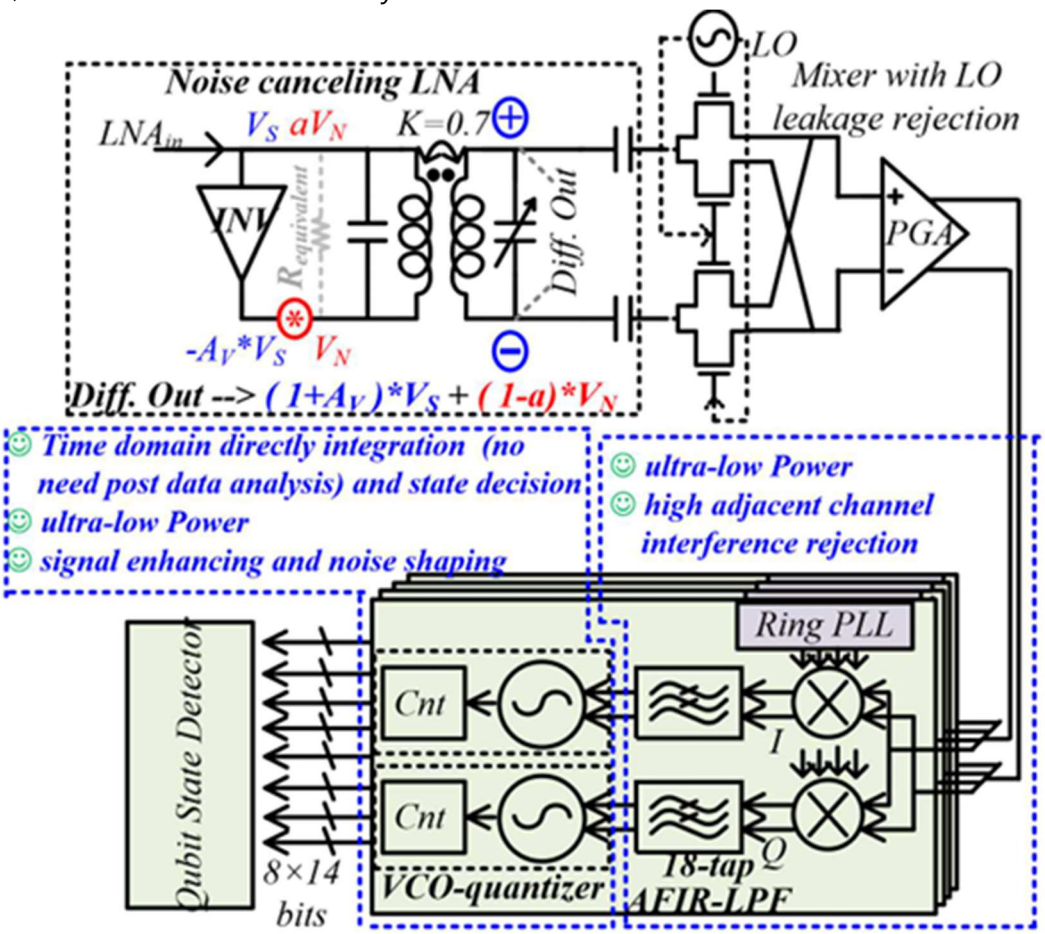
Session C30 Cryo-CMOS Circuit

이번 2025 IEEE VLSI의 Session C30은 Cryo-CMOS Circuit이라는 주제로 총 4편의 논문이 발표되었다. 이 세션에서는 cryogenic 환경에서 동작하는 fully-integrated 4-channel Frequency-Division-Multiplexing (FDM) transmon qubit state readout ASIC과 dual-stage injection-locked oscillator(IL-DCO)로 구성된 cryo-CMOS signal selector 등을 제안하였는데, 두 circuit 모두 cyro-CMOS의 특징을 이용한 power comsuption을 줄이는 것에 중점을 두었다.

#C30-3 본 논문에서는 cryogenic 환경에서 동작하는 fully-integrated 4-channel Frequency-Division-Multiplexing (FDM) transmon qubit state readout ASIC을 제안하였다. Cryogenic CMOS ASIC은 표준 CMOS 공정으로 만든 ASIC을 극저온인 cryogenic 온도에서도 안정적으로 동작하도록 설계한 chip이다. 본 논문에서는 이러한 cryogenic CMOS ASIC의 특성을, 양자컴퓨팅의 대규모 확장에서 일어나는 문제점인 배선, I/O 병목의 해소에 이용하였다. 이전 연구는 Cryogenic CMOS ASIC을 일부에만 적용하여 여전히 전력 소모가 높으므로, 본 논문은 더 낮은 readout 전력을 달성하기 위한 연구를 한 것이다.

Cryo-CMOS Fully-integrated 4 Channel FDM qubit state readout architecture는 크게 3가지 특징을 가지고 있는데, 1번째는 Two-step down-conversion sliding-IF를 통해 high-frequency quadrature down-mixer가 요구하는 전력소모를 절감한 것이다. 구체적으로는 기존 수신기는 quadrature down-mixer를 사용하여 한 번에 frequency를 down-conversion하여 전력 소모가 컸는데, sliding-IF 구조로 바꿔서 frequency를 한 번에 down-conversion하지 않고 Intermediate frequency를 중간에 거쳐서 two-step으로 down-conversion하여 전력 소모를 줄인것이다. 2번째는 Narrow-band zero-IF topology에 18-tap AFIR을 결합하여 FDM의 adjacent channel interference를 줄인것이다. 기존 FDM readout에서는 adjacent channel 누설로 인해 SNR이 저하되는 문제가 있었는데, AFIR(Analog FIR) LPF(low pass filter)가 narrow한 band만 통과시켜서 channel selectivity를 높인 것이다. 이때 zero-IF는 direct-conversion 또는 homodyne이라고도 불리며 한 번의 mixing으로 baseband로 down-conversion하는 구조이며, 1번째에서 얘기한 IF에서 basedband로 frequency를 내릴 때 zero-IF 구조를 사용한 것이다. 3번째는 VCO-based analog signal conversion으로 time-domain state decision을 수행해 noise shaping 장점을

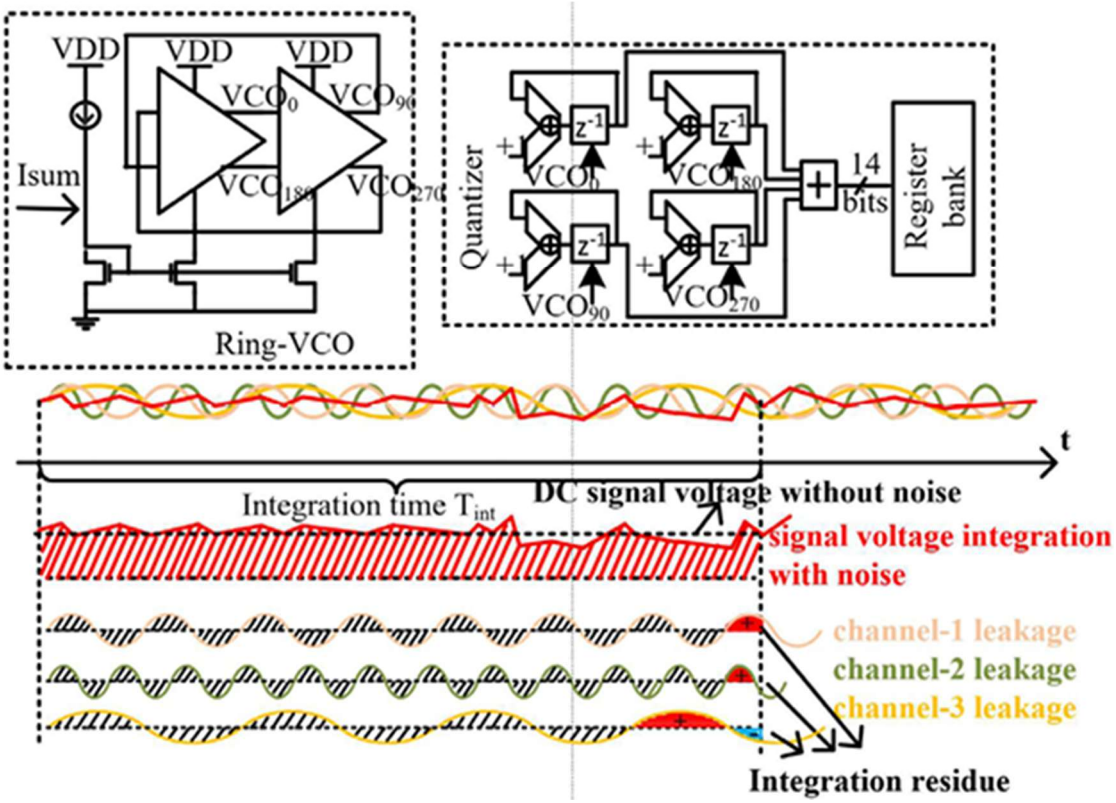
얻고 직접 qubit 상태를 판독한 것이다. 기존에는 ADC입장에서 FDM의 대역폭을 한 번에 봐야하므로 고속, wideband를 지원해야 하는 부담이 있었는데, 본 논문에서는 VCO를 통해 입력 전압을 frequency 변화로 mapping하여 counter로 period를 측정하는 것으로 바꿔서, ADC를 사실상 대체하여 system의 부담을 낮춘 것이다.



[그림 1] fully-integrated 4-channel FDM homodyne demodulation readout receiver의 architecture

그림 1에 fully-integrated 4-channel FDM transmon qubit state readout ASIC의 전체 architecture가 나와있고, 크게 RF-(Mixer1)-IF-(IF Mixer2)-zero-IF의 구조를 갖는다는 것을 알 수 있다. 구조에 대해 살펴보면, 먼저 좌측 상단의 Noise canceling LNA(low noise amplifier)는 입력으로 $V_S + V_N$이 들어오고, transformer를 통해 differential 신호로 변환하며 신호(V_S)는 증폭하고 noise(V_N)는 낮추는 역할을 한다. 이때의 식은 $(1 + A_v)V_S + (1 - a)V_N$ 이다. 우측 상단에 나와있는 Mixer with LO(Local Osillator) leakage rejection과 PGA(Programmable Gain Amplifier)는 LO leakage를 줄이기 위한 architecture이며 self-mixing과 DC-offset 등을 줄이는 역할을 한다. 이때 LO leakage는 mixer의 LO signal이 RF-IF path로 leakage되는 현상이다. 또한 여기 있는 mixer는 1번째 down-conversion으로 수백 MHz 대역의 IF로 내리고, PGA가 필요한 gain만큼 gain을 보정하는 역할도 겸한다. 우측 하단에 나와있는 Ring PLL은 4개의 on-chip PLL로 구성되고, 1번째 mixing 후에 4개의 channel에서 channel마다 IF frequency가 조금씩 다른데, 이것을 딱 맞춰주는 역할

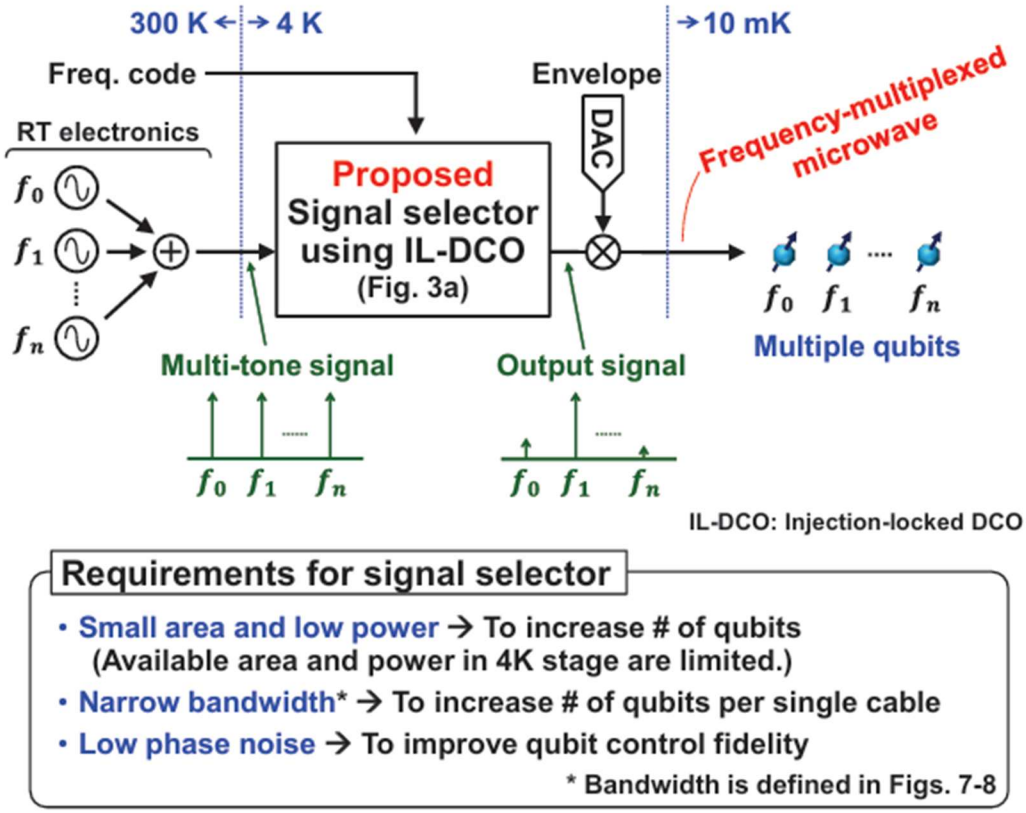
을 한다. 이 과정을 거쳐야 AFIR-LPF가 narrow bandwidth로 adjacent channel interference를 효과적으로 rejection할 수 있다. Ring PLL바로 아래의 IF mixer는 2번째 down-conversion으로 각 channel을 baseband 근처인 zero-IF로 정밀하게 낮춰주는 역할을 한다. IF mixer의 바로 왼쪽에 있는 18-tap AFIR-LPF는 FDM adjacent channel interference를 효과적으로 제거하여, 뒷 단의 dynamic range와 power등의 constraint를 완화시킨다. 이때 area와 power를 줄이기 위해, time-interleaved 방식으로 하나의 AFIR-LPR를 4-channel이 공유한다.



[그림 2] VCO-quantizer와 time-domain integration

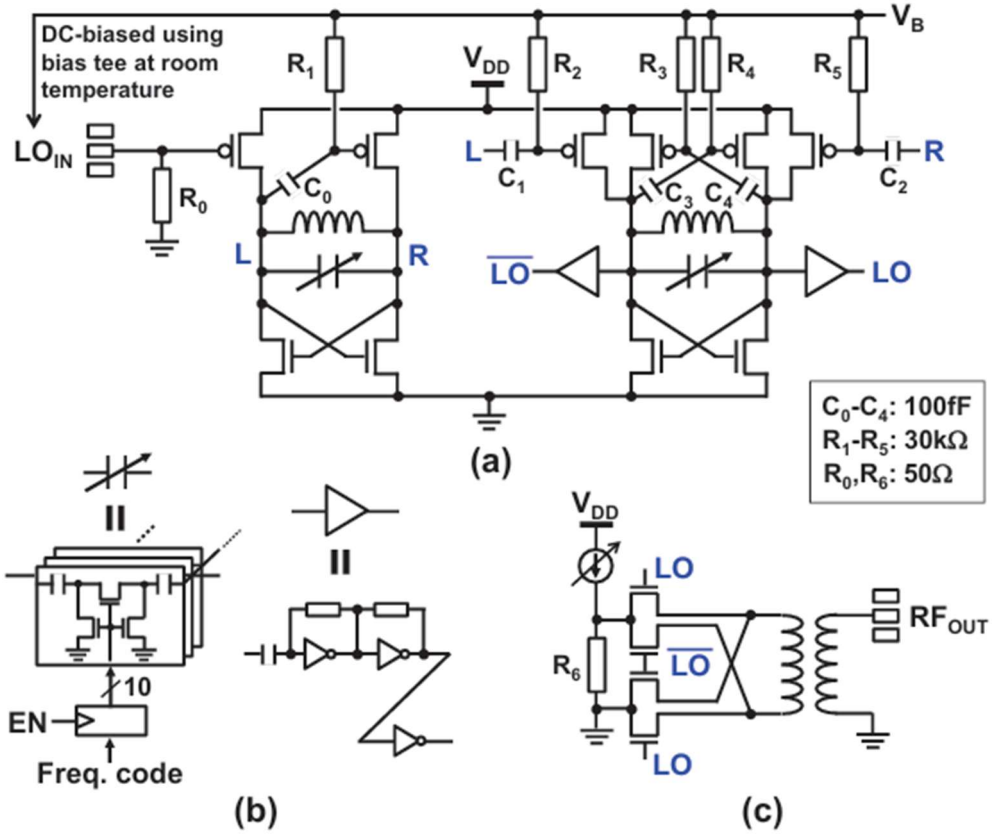
다음으로 그림 하단의 VCO-quantizer는 baseband 전압을 Gm(transconductance) cell을 이용하여 전류로 바꿔서 current-biased ring-VCO에 누적하고, 그 결과 생기는 frequency/phase 변화를 T_{int}동안 counter로 세어 14-bit으로 encoding하여 이를 직접 판정하는데 사용한다. 이 과정은 그림 2에 상세히 나와있고, 우측 상단의 Quantizer와 Register bank를 통해 각 phase signal의 edge를 감지하여, counter가 T_{int}동안 세서 14-bit code로 encoding되는 것을 확인할 수 있다. VCO-quantizer의 output은 8*14-bit인데, 이는 channel마다 I,Q 2가지 path가 있으므로 8개의 14-bit값이 출력되기 때문이다. 마지막으로 그림 1 좌측 하단의 Qubit State Detector는 counter 출력을 직접 비교하여 qubit 상태를 분류하는 역할을 한다. 그림에 표시된 것처럼 복잡한 digital channelizer나 high resolution ADC등의 post data anlysis가 별도로 필요 없다는 것을 알 수 있다.

#C30-4 본 논문에서는 dual-stage injection-locked oscillator(IL-DCO)로 구성된 cryo-CMOS signal selector를 제안하였다. 현재 양자컴퓨터에서는 qubit마다 unique한 control frequency를 가지고 있어서, refrigerator의 cryogenic stage에 oscillator를 여러 개 둬야 하고, 이로인해 power 소모가 큰 문제가 있다. 따라서 논문에서는 selector를 통해 multi-tone microwave 신호에서 원하는 단일 tone만 추출해서 qubit 제어에 쓰고 power 소모를 줄인다는 idea이다.



[그림 3] 제안된 cyro-CMOS signal selector architecture

그림 3에는 논문에서 제안한 cyro-CMOS signal selector architecture가 나와있고, 그림의 왼쪽에서 보이는 것처럼 RT(300K)에서 여러 oscillator가 만든 multi-tone microwave를 single 케이블로 CT(4K)까지 내린다. 그 다음 IL-DCO based signal selector가 지정된 single-tone만 선택하여, envelope DAC와 mixer를 통해 pulse shaping한 후 10mK의 multiple qubits에 공급하는 frequency-multiplexed qubit control을 하는 idea가 설명된 것이다. 논문에서 제안한 signal selector는 5.7~7.5 GHz 대역을 목표로 한 dual-stage IL-DCO(Injection-Locked Digitally Controlled LC Oscillator) 구조이며 그림 4에 selector의 circuit implementation이 나와있다. (a)는 Dual IL-DCO이며 핵심 circuit이다. 먼저 1번째 IL-LDO에서 cross-coupled LC tank가 injection-locking으로 resonance f_0' 에 가장 가까운 입력 tone에 lock시킨다. 이때 L, R node는 1번째 DCO의 differntial output이고, C1~C4를 통해 2번째 DCO에 전달된다.



[그림 4] 제안된 selector의 circuit implementation

2번째 IL-DCO에서는 1번째와 같은 f_0'에 맞춰져서, 선택되지 않은 tone을 weak하게 만들어서, select된 tone을 더 refine한다. signal flow는 다음과 같다. LOIN(multi-tone) - 1번째 IL-DCO가 f_0' 근처 tone에 lock - L/R – 2번째 IL-DCO injection&refinement - mixer에서 LO로 gating – balun - RFOUT(single-tone). 실험은 single-tone 입력의 frequency relationship evaluation, 4-tone 입력의 tone selection, 2-tone에서의 SFDR 측정으로 크게 3가지 경우에 대해 진행되었으며, 각각의 실험으로 frequency를 정밀하게 선택할 수 있는지. multi-tone을 실제로 support하는지, SFDR requirement를 만족하는지를 증명한 것이다.

저자정보

권재훈 석사과정 대학원생

- 소속 : 한국과학기술원 전기및전자공학부
- 연구분야 : Digital Circuit Design, ECC Hardware Design
- 이메일 : jhkwon@ics.kaist.ac.kr
- 홈페이지 : https://ics.kaist.ac.kr/

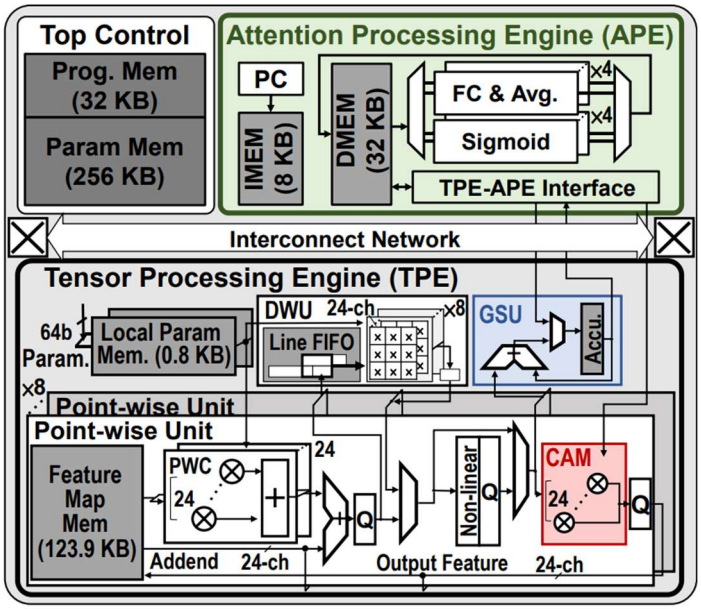
2025 IEEE VLSI Review

경북대학교 전자전기공학부 박사과정 박승현

Session 10 AI Accelerators 1

VLSI 2025 10 세션의 네 편 논문은 영상 처리와 엣지 컴퓨팅, 온디바이스 학습, 로보틱스 시스템 등 차세대 응용을 위한 새로운 아키텍처적 접근을 제시한다. 10-2는 코덱과 후처리를 통합한 비디오 프로세서를 통해 영상 스트리밍 전반의 효율을 근본적으로 개선할 수 있음을 보여주었다. 10-3은 하이퍼디멘셔널 컴퓨팅을 활용해 지속 학습이 가능한 온디바이스 가속기를 구현함으로써, 에너지 효율적인 적응형 학습의 가능성을 제시했다. 10-4는 타일 기반의 유연한 가속기 구조를 통해 워크로드 변화에 실시간으로 대응할 수 있는 엣지 컴퓨팅 플랫폼의 새로운 방향을 제안하였다. 마지막으로 10-5는 로보틱스 응용에 특화된 이종 SoC를 설계해, 복잡한 인지 및 제어 파이프라인을 하나의 칩에서 통합적으로 처리할 수 있는 가능성을 열었다.

#10-2 NuVPU: A 4.8~9.6 mJ/frame Progressive NTT-based Unified Video Processor for Stable Video Streaming and Processing with Neural Video Codec



[그림 1] Overall System Architecture

본 논문은 Neural Video Codec (NVC)을 기반으로, 영상 스트리밍과 후처리를 동시에 지원하는 통합형 비디오 프로세서 NuVPU를 제안한다. 기존 연구들은 주로 후처리에 초점

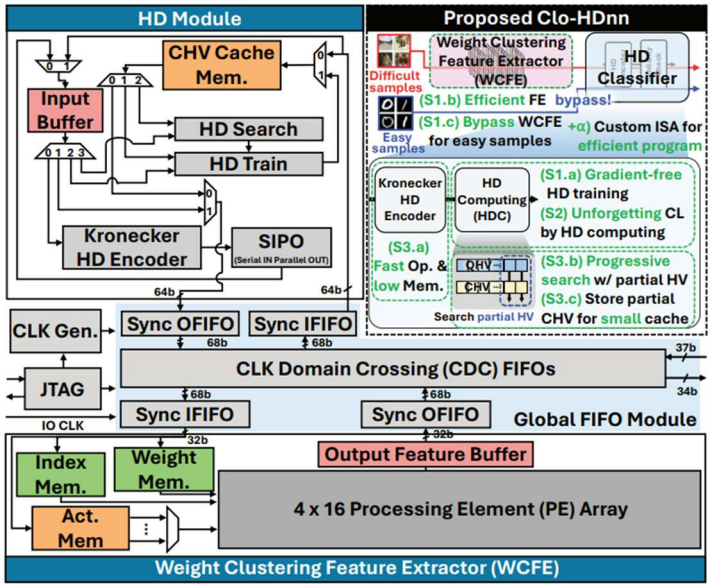
을 맞춰 코덱 정보 활용이 미흡했는데, 저자는 코덱 내부의 모션 벡터 등 중간 정보를 적극적으로 활용하여 전체 파이프라인 효율을 높이고자 하였다.

핵심 기여는 세 가지이다. 첫째, Selective Convolution-mode Neural Engine(SCNE)은 공간 도메인과 NTT 도메인 연산을 상황에 따라 전환하며, 이를 통해 평균 1.69~3.35배의 처리량 향상을 달성하였다. 둘째, Progressive NTT Unit (PNTU)는 불필요한 연산과 메모리 접근을 줄여 도메인 변환 시 연산량을 44.8%, 메모리 접근을 80% 절감하였다. 셋째, Frequency-aware Compressor (FAC)와 Adaptive Tile Scheduler (ATS)는 warping 기반 프레임 재활용 과정에서 발생하는 EMA를 81.3%까지 줄였다.

28nm 공정으로 제작된 칩은 최대 250MHz에서 동작하며, 4.89.6 mJ/frame의 에너지 소모로 36.9 TOPS/W의 탁월한 에너지 효율을 달성하였다. 이는 기존 비디오 프로세서 대비 9.22.3배 향상된 수치다. 또한 4K 영상 스트리밍에서 0.37 SSIM, 0.7 PSNR의 품질 향상을 입증하며, 실제 응용에서의 우수한 성능을 보여주었다.

본 연구는 코덱과 후처리를 통합적으로 고려하여, NVC 기반 차세대 영상 스트리밍에 적합한 새로운 프로세서 아키텍처를 제시한다는 점에서 의의가 크다. 특히 도메인 변환 최적화와 메모리 효율화 기법을 결합해, 실시간 고해상도 영상 서비스의 연산·전력 병목을 근본적으로 해결할 수 있음을 입증했다는 점이 인상적이다.

#10-3 Clo-HDnn: A 4.66 TLOPS/W and 3.78 TOPS/W Continual On-Device Learning Accelerator with Energy-efficient Hyperdimensional Computing via Progressive Search



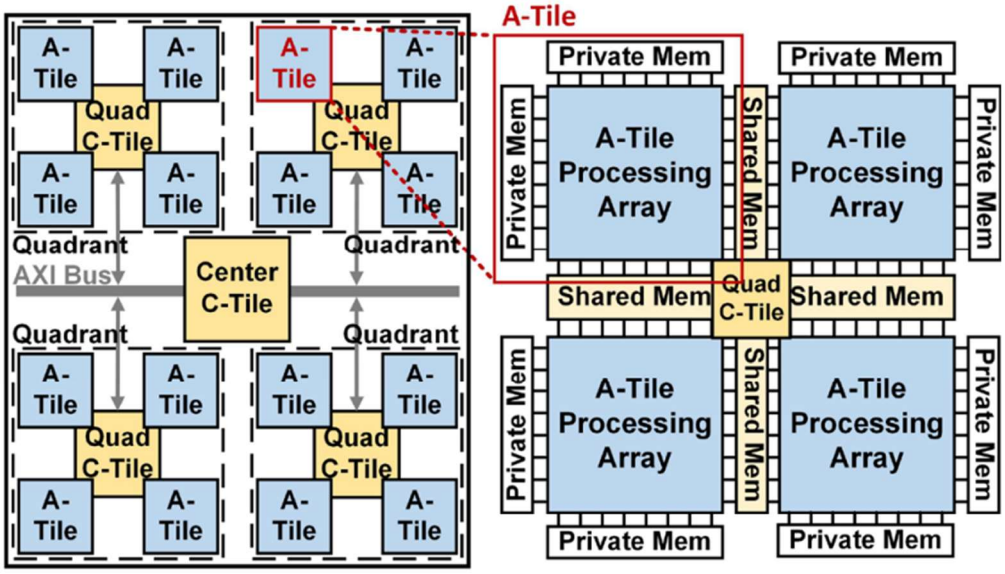
[그림 2] Proposed Clo-HDnn Architecture and Key Features

본 논문은 지속 학습 (Continual Learning, CL)을 지원하는 온디바이스 학습 가속기 Clo-HDnn을 제안한다. 기존 온디바이스 학습은 그래디언트 기반 학습으로 인한 높은 연산량과 메모리 소모, 그리고 새로운 데이터 학습 시 기존 지식을 잃는 문제 (catastrophic forgetting)를 겪어왔다. 저자는 이러한 한계를 극복하기 위해 뇌 영감을 받은 하이퍼디멘셔널 컴퓨팅 (HDC)을 도입하였다.

핵심 아이디어는 세 가지이다. 첫째, Gradient-free HDC 기반 학습을 적용해 연산 복잡도를 크게 줄이고, 필요 시 저비용의 Weight Clustering Feature Extractor (WCFE)를 통해 효율적인 특징 추출을 수행한다. 둘째, Kronecker HD 인코더와 프로그레시브 서치 (progressive search) 기법을 통해 입력 데이터를 부분적으로만 인코딩·탐색해 불필요한 연산을 줄이면서도 정확도를 유지한다. 셋째, 듀얼 모드 동작을 지원하여 간단한 데이터셋에서는 특징 추출 과정을 생략하고, 복잡한 데이터셋에서는 정교한 추출을 수행함으로써 상황에 맞는 유연한 학습이 가능하다.

이를 통해 Clo-HDnn은 기존 CL 가속기의 한계를 넘어서, 지속적 적응 학습과 지식 보존을 동시에 달성할 수 있는 새로운 아키텍처적 패러다임을 제시한다. 본 연구는 온디바이스 환경에서 에너지 효율적이고 경량화된 CL 구현 가능성을 입증하며, 향후 IoT, 웨어러블, 로보틱스 등 동적 환경에서의 지능형 시스템 설계에 중요한 기여를 한다.

#10-4 EVA: A 16mm2 1.54TFLOPS Tiled-Based Accelerator for Evolvable Edge Computing



[그림 3] Hierarchically Connected EVA Architecture

본 논문은 엣지 컴퓨팅의 빠르게 변화하는 워크로드에 대응하기 위해 설계된 타일 기반 가속기 EVA를 제안한다. 기존의 도메인 특화형 가속기는 특정 커널에는 최적화되어 있으나, 워크로드가 변화할 때 자원 활용도가 떨어지는 한계가 있었고, 반대로 CGRA 기반 구조는 유연성을 제공하지만 재구성이 느려 실시간 적응성에 제약이 있었다.

EVA는 이를 해결하기 위해 프로그래머블 PE 어레이 (A-Tile)와 RISC-V 기반 제어 타일 (C-Tile)을 결합한 이종 타일 아키텍처를 도입하였다. A-Tile은 다양한 연산 커널을 높은 활용도로 수행할 수 있고, C-Tile은 수십~수백 ns 수준에서 빠른 런타임 적응 및 재구성을 담당한다. 또한 분산 메모리 구조와 다층 인터커넥트 패브릭을 통해 데이터 이동의 효율성을 높였으며, 워크로드를 공간적/시간적 매핑으로 최적화할 수 있는 유연한 실행 환경을 제공한다.

이러한 설계를 통해 EVA는 고성능·고밀도 연산과 동시에 실시간 적응성을 달성하였으며, 엣지 환경에서의 진화하는 다양한 작업을 단일 플랫폼에서 효율적으로 처리할 수 있는 가능성을 보여주었다. 본 연구는 엣지 컴퓨팅 아키텍처가 앞으로 성능과 유연성의 균형을 어떻게 맞춰야 하는지를 제시하는 중요한 방향성을 제공한다.

저자정보

박승현 박사과정 대학원생

- 소속 : 경북대학교
- 연구분야 : 딥러닝 가속기 설계
- 이메일 : ijjh0435@gmail.com
- 홈페이지 : https://ai-soc.github.io/

2025 IEEE VLSI Review

한양대학교 신소재공학과 석박통학과정 송충석

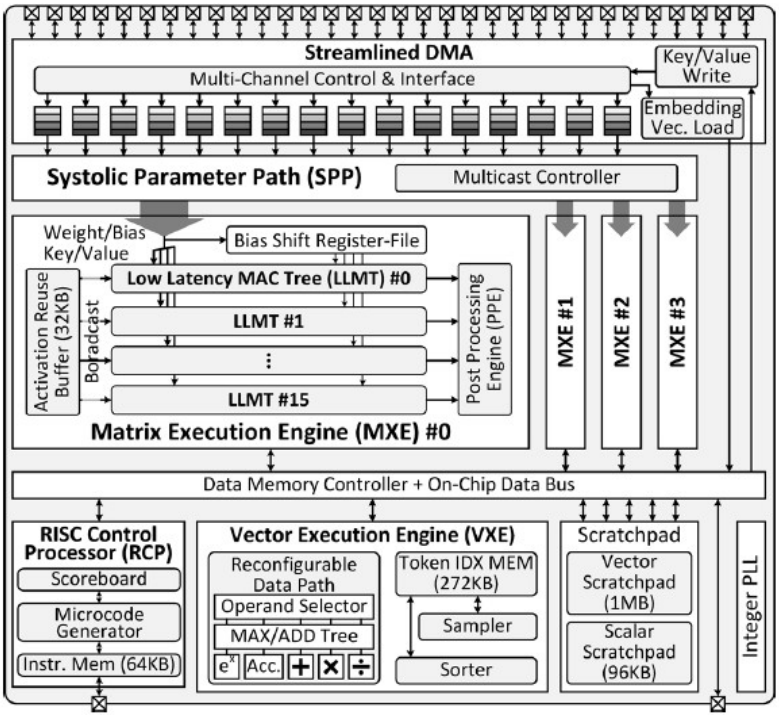
Session 13 AI Accelerators 2

이번 2025 IEEE VLSI의 Session 13은 AI Accelerators 2라는 주제로 총 4편의 논문이 발표되었다. 현재 가장 활발히 연구되고 있는 transformer 구조를 가속하기 위한 하드웨어가 발표되었으며, #13-1은 디코더 기반 transformer, #13-2는 multi-task transformer, #13-3은 transformer 학습, #13-4는 diffusion 기반 transformer를 타겟으로 정하였다. 본 review에서는 #13-1, #13-2, #13-3을 review한다.

#13-1 논문에서는 Adelia라는 대규모 언어 모델을 가속하는 칩을 발표하였다. 디코더 기반 대규모 언어 모델의 추론은 prefill 단계와 decode 단계로 구분되며 prefill 단계에서는 높은 연산 집약도를, decode 단계에서는 메모리 접근 병목을 보이는 상반된 특성을 가진다. 이러한 특성으로 인해 기존 GPU에서는 메모리 대역폭 활용률과 연산 자원 활용률이 낮으며, 다른 전용 가속기들 또한 decode 단계에 최적화되었으나 prefill 처리에서 성능이 떨어지는 문제를 가지고 있다. 이러한 문제를 해결하기 위해 본 논문에서는 streamlined dataflow와 dual-mode parallelization을 제안하고 이를 Adelia에 탑재하였다.

Streamlined dataflow는 외부 메모리와 연산 엔진 사이의 대역폭을 정확히 매칭시켜 주는 방법으로 데이터 이동 경로를 불필요한 버퍼링 없이 단순화함으로써 연산 엔진이 항상 메모리 대역폭에 맞춰 최대로 동작하도록 하는 구조다. 더불어 긴 context를 분산 처리하는 context mode와 여러 요청을 동시에 처리하는 batch mode를 적용한 Dual-mode parallelization을 Adelia에 구현하였다.

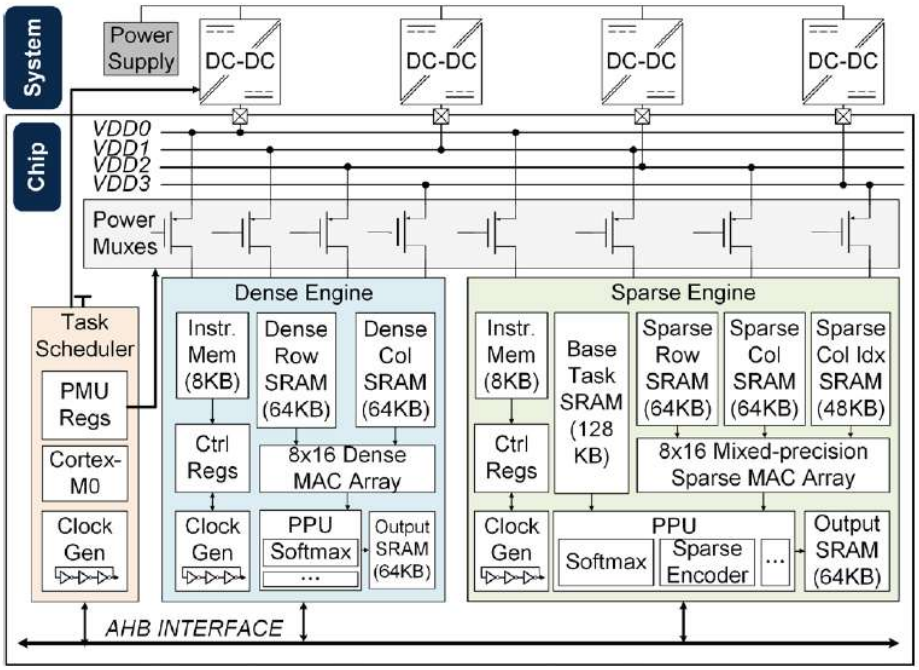
Adelia는 4nm CMOS 공정으로 제작되었으며, 면적은 5.28mm^2, 동작 전압은 0.65 – 0.9V, 최대 주파수는 1GHz이다. 측정 결과 GPU 대비 메모리 대역폭 효율은 1.59배, 처리량 효율은 2.51배 향상되었다. 그 결과 Llama-7B 추론에서 최대 latency를 35.1퍼센트 감소를 보여줌으로 인해 단일 최적화(decode 단계)에 집중했던 기존 연구 대비 더 균형 잡힌 성능을 보여주었다.



[그림 1] Adelia의 전체 하드웨어 구조

#13-2 논문에서는 최초의 multi-task transformer 가속기를 칩으로 제작한 결과를 발표하였다. Multi-task transformer는 "base task" 와 그에 대한 차이(delta)를 계산하는 "sub task" 구조를 활용하여 "sub task" 연산 시 base task를 기준으로 차이만을 활용하여 재연산하기 때문에 효율적인 네트워크로 주목받고 있다. 그러나 delta 행렬의 불규칙한 희소성, 빠른 전력/성능 변화, base task와 delta의 서로 다른 정밀도 요구로 인해 하드웨어 구현이 까다로운 문제가 있었다. 본 논문은 세가지 방식을 하드웨어에 적용해 multi-task transformer를 가속하였다: 첫번째는 block-wise structure sparsity를 도입하여 delta 행렬의 희소성에 제약을 가해 재 학습시켜 불규칙성을 완화시켰고, 두번째는 DVFS (dynamic voltage frequency scaling) 기법을 접목시켜 workload에 따른 빠른 전압 변환을 통해 효율적인 전력공급을 가능케 하였고, 마지막으로 mixed precision 연산을 적용해 base task를 INT8, sub task와 delta는 INT4/INT8, softmax 연산은 INT32로 수행하여 전력 소모를 줄이면서 정확도를 유지시켰다.

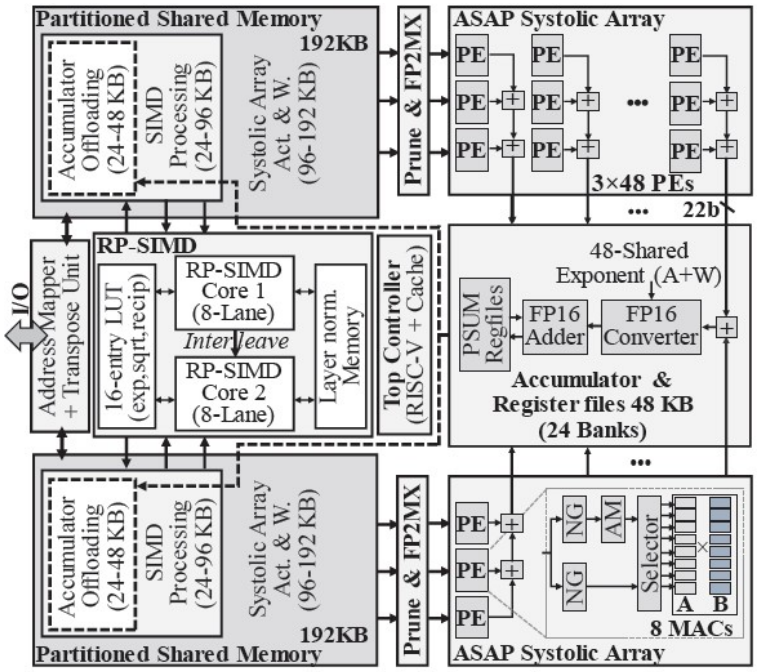
제안된 칩은 22nm CMOS 공정으로 제작되었으며, 면적은 5.8mm^2를 차지하였다. 실제 multi-task 시나리오에서 평균적으로 sub-task는 41%에너지 감소와 68% 딜레이 감소를 달성했으며, DVFS 기법을 적용시켜 추가로 21.4%의 에너지 절감을 이루었다. 특히 sub task의 수가 증가할수록 base task 계산 비용이 분산되어 효율이 더 높아졌으며, 이는 기존 single-task transformer 가속기와 차별화되는 지점이라고 볼 수 있다.



[그림 2] 제안한 칩의 하드웨어 전체 구조.

#13-3 논문에서는 ASAP이라는 transformer 학습을 가속화할 수 있는 칩을 발표하였다. 기존 GPU에서 희소성을 포함한 네트워크를 학습시킬 때, 고정된 N:M 희소성 제약으로 인해 중간 단계의 희소성 수준을 적용하기 어려운 문제가 있다. 예를 들어, GPU에서 2:4 희소성만 (4개의 데이터 중 2개는 0으로 설정) 지원한다면, 중간 수준의 3:8, 5:8 희소성 등은 적용할 수 없다. 이를 해결하기 위해 ASAP은 다양한 N:M 희소성 연산을 지원함과 동시에 비대칭 정밀도를 동시에 지원하는 접근법을 제안하였다.

ASAP은 동적 정밀도 할당 알고리즘을 사용해 가중치의 중요도에 따라 여러 정밀도로 표현할 수 있는 알고리즘을 (E2M2, E2M5. Exponent n-bit, Mantissa n-bit) 적용하였다. 특히 이 알고리즘을 활용하여 높은 희소성에 대해서는 높은 정밀도, 낮은 희소성에 대해서는 낮은 정밀도를 적용해 각 연산기마다 탑재된 8개의 3b*3b 곱셈기가 최대로 활용될 수 있도록 하였다. 28nm CMOS 공정으로 제작된 ASAP 칩은 1.64mm^2의 면적에서 26.6TOPS/W의 최대 에너지 효율과 2.9TOPS 성능을 이루었으며 이는 기존 연구 대비 훈련 에너지 효율에서 2.02-5.07배 향상을 보여주었다. GPU에서 구현하지 못한 다양한 희소성 수준을 모두 효율적으로 지원가능한 하드웨어로 효율적인 학습에 활용가능성을 보여주어 추후 연구가 더욱 기대되는 연구이다.



[그림 3] 제안한 ASAP의 하드웨어 전체 구조

저자정보

한양대 석박통합과정 대학원생

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