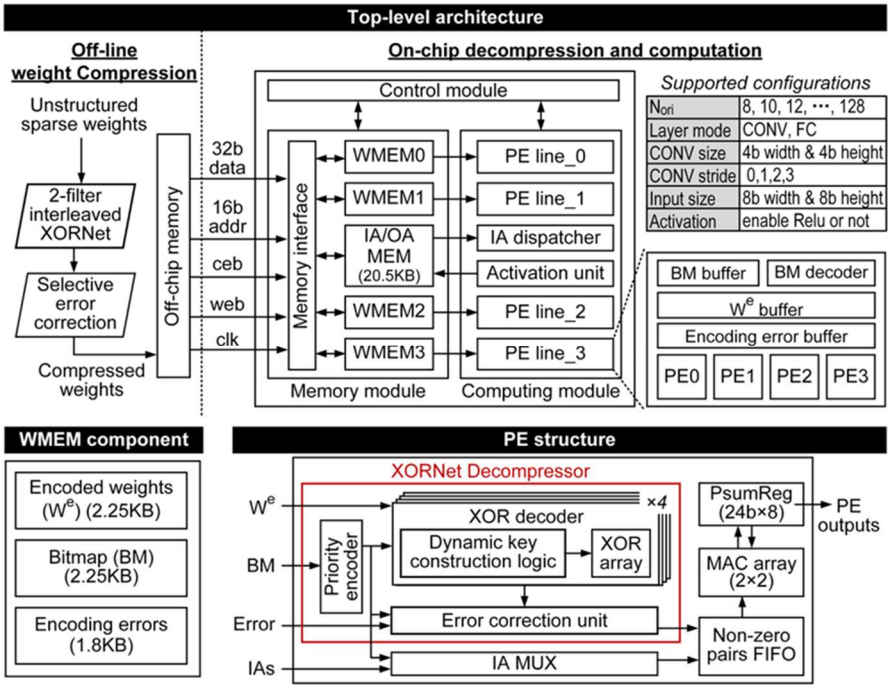
2025 IEEE CICC Review

포항공과대학교 반도체대학원 박사과정 박은빈

Session 11 ASIC and Accelerator

이번 CICC 2025의 Session 11에서는 AI 가속기 및 특화된 ASIC 설계를 주제로 총 8편의 논문이 발표되었다. 다양한 응용 분야에서 요구되는 고효율, 저전력 연산을 실현하기 위해, 각 논문은 특화된 데이터 표현 방식, 연산 아키텍처, 그리고 하드웨어-소프트웨어 협력 최적화 전략을 제안하였다. 특히 신경망 압축, reinforcement learning, spiking neural networks, LLM fine-tuning, 그리고 mmWave 기반의 OTA 레이더 처리까지 폭넓은 기술 스펙트럼이 다루어졌다.

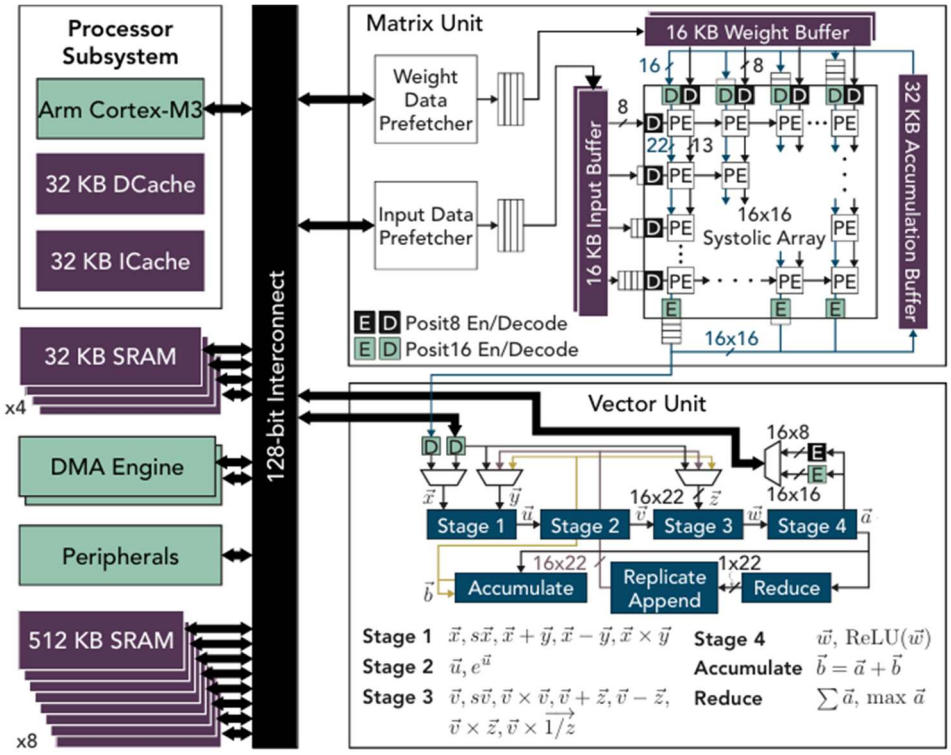
#11-3 본 논문은 상하이교통대와 콜롬비아대 공동 연구팀이 발표한 것으로, 고압축 희소 신경망 모델의 온칩 비복원 압축 해제 구조를 갖춘 NN 가속기 SparseTrim을 제안한다. 기존 희소 가중치 압축 방식은 압축률이 낮거나 직렬 해제가 병렬 연산 구조에 비효율적인 문제가 있었으나, SparseTrim은 XORNet 기반 압축 포맷을 활용해 이러한 문제를 해결하였다. XORNet은 fine-grained sparsity를 활용하여 압축 중 일부 오류를 허용한 뒤, 중요 오류만 선택적으로 복구하는 방식으로 고압축을 실현하며, 최대 5.3배 (INT8 기준)의 압축률을 제공한다.



[그림 1] SparseTrim 과 PE의 전반적인 구조

이 논문에서는 해당 포맷에 맞춘 경량 동적 해제 키 생성 하드웨어를 설계하여 기존 LUT 기반 해제 방식 대비 면적을 52배 감소시켰고, PE 간 연산 부하 불균형 문제를 완화하기 위해 filter-pair 기반 부하 균형 기법도 함께 제안하였다. SparseTrim은 28nm 공정으로 제작되었으며, INT8 기준 최대 10.1 TOPS/W의 시스템 에너지 효율을 달성하였다. 이는 기존 COO, RLC, CFO 기반 방식 대비 평균 23~24% 높은 효율을 보였고, ResNet50 기준 throughput이 22% 개선되는 효과도 함께 입증되었다. 본 연구는 압축된 희소 신경망의 실시간 처리에 적합한 가속기 구조를 제시함으로써, 메모리 대역폭 제약과 에너지 병목을 동시에 해결한다는 점에서 의의가 크다.

#11-4 본 논문은 스탠퍼드대학교에서 발표한 것으로, 확장현실(XR) 환경에서 실시간 인식 작업을 위한 통합형 가속기 SoC인 Aspen을 제안한다. 기존 XR 기기에서는 시각관성 측위(VIO), 시선 추적, 객체 인식 등 다양한 인식 파이프라인을 별도의 하드웨어로 처리했으나, 이로 인한 전력 소모 증가가 문제였다. 이를 해결하기 위해 본 논문은 모든 인식 작업을 DNN 기반으로 통합 처리할 수 있는 통합형 DNN 가속기 구조를 설계하였다.

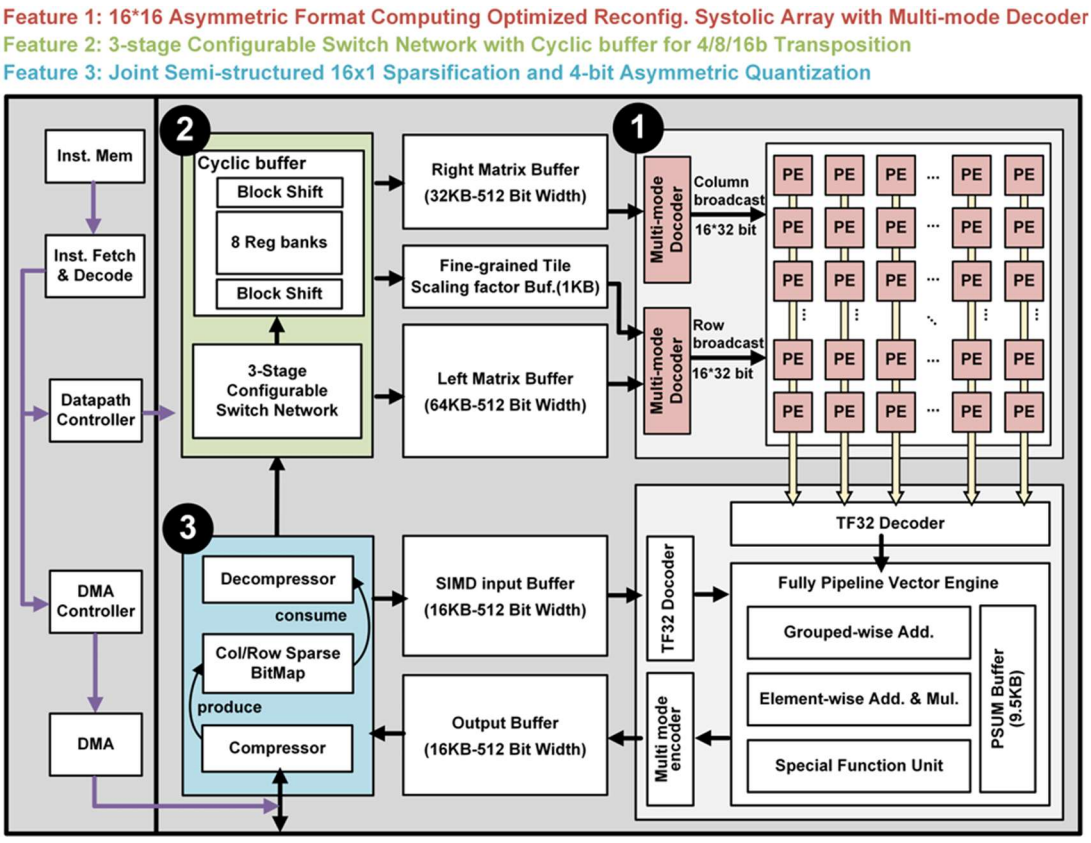


[그림 2] Aspen의 전반적인 구조

Aspen의 주요 기술로는, 정밀도 손실 없이 메모리 사용량을 줄이기 위한 Posit 기반의 mixed-precision quantization 전략이 있다. 특히 정밀도가 중요한 VIO의 경우, 입력 특성에 따라 Posit8과 Posit16을 계층적으로 혼합하여 활용하며, 전용 vector unit과 systolic array 기반의 matrix unit을 통해 고성능 연산을 영역별로 최적 수행한다. 또한, 데이터

layout 일관성 유지와 효율적인 prefetching을 통해 메모리 접근 병목을 해소하였다. 그 결과, Aspen은 VIO 98.9 FPS, 시선 추적 630 FPS, 객체 인식 31 FPS의 실시간 성능을 달성하였고, 모델 전체를 온 칩 4MB SRAM에 수용함으로써 외부 메모리 접근을 제거하였다. 특히 Posit quantization을 통해 FP32 대비 74% 메모리 절감, 오차는 0.0002m/0.014° 수준으로 억제하여 정확도도 유지하였다. 본 논문은 다양한 실시간 XR 인식 작업을 하나의 하드웨어 구조로 통합 처리할 수 있는 실용적 DNN SoC를 구현하였다는 점에서 높은 실용성과 확장 가능성을 지닌다.

#11-6 본 논문은 칭화대학교에서 발표한 것으로, 대규모 언어 모델(LLM)의 로컬 파인튜닝을 위한 고효율 전용 프로세서 구조를 제안한다. 최근 QLoRA, QA-LoRA와 같은 Parameter-Efficient Fine-Tuning(PEFT) 기법이 부상하면서, 메모리 제약이 있는 AI PC 환경에서도 LLM을 미세 조정하려는 수요가 증가하고 있다. 그러나 기존의 신경망 가속기(NPU)는 비대칭 양자화 연산에 대한 지원이 미비하고, 낮은 정밀도의 데이터 전치(transposition) 및 외부 메모리 접근 비용이 크다는 문제를 안고 있다.



[그림 3] 본 논문에서 제안하는 fine-tuning 과정

이를 해결하기 위해 본 논문은 QLoRA 연산 흐름을 분석한 뒤, 세 가지 핵심 기술을 기반으로 한 전용 파인튜닝 프로세서를 설계하였다. 첫째, 제안된 텐서 어레이는 비대칭 양자화 연산에 최적화된 구조로, 4비트 곱셈기와 5비트 누산기를 기반으로 다양한 연산 모

드(BF16×INT4, FP9×FP16 등)를 지원하는 재구성 가능한 연산 소자(PE)를 포함한다. 이를 통해 높은 연산 효율과 유연성을 동시에 확보하였다. 둘째, 다중 형식 지원 전치 엔진은 3단계 스위칭 네트워크와 순환 버퍼(cyclic buffer)를 활용하여 4/8/16비트 데이터에 대한 전치 연산을 효율적으로 수행하며, 기존 GPU 대비 8배 높은 어레이 활용도를 달성한다. 셋째, 반구조화 sparsity 기반의 압축 기법을 적용하여, QLoRA의 INT4 가중치 및 BF16 어댑터 연산에서 외부 메모리 접근을 최소화하면서도 정확도를 유지할 수 있도록 하였다. 해당 프로세서는 28nm 공정으로 구현되었으며, 3.14 TFLOP/W의 에너지 효율을 기록하였다. LLaMA2-7B 모델에 대해 GPTQ 기반의 QLoRA 파인튜닝을 수행한 결과, 기존 GPU 대비 1.475배 높은 에너지 효율과 8배 향상된 데이터 전치 성능을 보였으며, 누적 오차는 -95dB 수준으로 억제되었다. 본 연구는 LLM의 로컬 파인튜닝이라는 최신 응용에 대응하기 위해 비대칭 양자화, 형식 전치, 압축을 통합 설계한 점에서 큰 의의가 있으며, 향후 AI PC 및 엣지 디바이스에서의 실용적 LLM 적용에 중요한 기여를 할 수 있다.

저자정보

박은빈 박사과정 대학원생

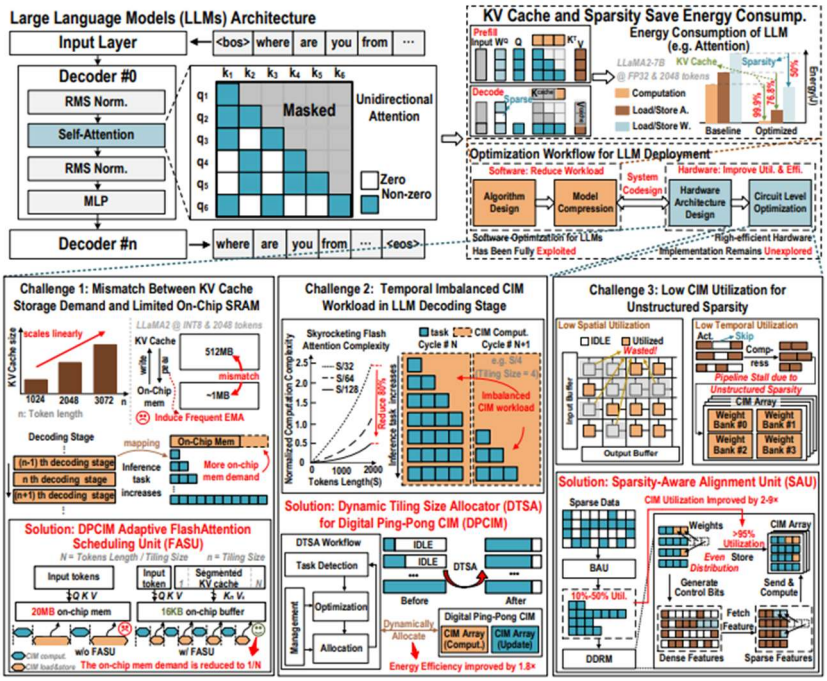
- 소속 : 포항공과대학교
- 연구분야 : 딥러닝 가속기 설계
- 이메일 : eunbin@postech.ac.kr
- 홈페이지 : https://sites.google.com/view/epiclab

2025 IEEE CICC Review

KAIST 전기및전자공학부 석사과정 박민하

Session 18 Digital Compute-in-Memory

#18-4 본 논문은 중국 Southeast University와 EDA 국가혁신센터가 공동으로 발표한 연구로, FlashAttention 기반 DCIM(Digital Compute-in-Memory) 가속기를 제안한다. 수십억 개의 파라미터와 복잡한 연산을 요구하는 LLM의 추론을 edge 환경에서도 효율적으로 수행하기 위해, 본 구조는 FlashAttention, KV 캐시 최적화, 희소성 정렬(sparsity-aware alignment) 기법을 통합한 DCIM 아키텍처로 설계되었다.



[그림 1] LLM 가속기에서의 주요 과제와 제안된 해결 방안

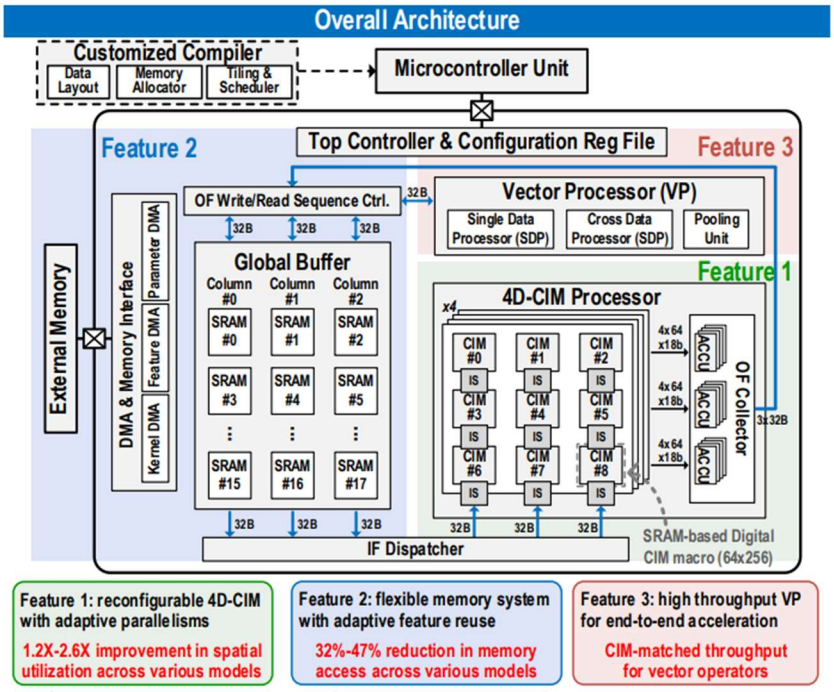
Fig.1은 제안된 구조가 decoder-only 기반의 LLM 아키텍처임을 보여주며, unidirectional attention 방식이 적용된 다수의 디코더 계층으로 구성되어 있음을 나타낸다. 핵심 연산은 Query/Key/Value 생성을 위한 WEngine, Attention 연산을 위한 AEngine, SAU, DTSA, Post Processing Unit(PPU) 등으로 구성된다.

WEngine은 16개의 1Kb 디지털 Ping-Pong CIM(DPCIM) 매크로로 구성되어 있으며, 8-bit 가중치와 2-bit 활성화 간 병렬 곱셈 연산을 수행한다. DPCIM은 연산과 가중치 업데이트를 병행하며, 근사화와 오차 보상 기법을 통해 에너지 및 면적 효율을 높인다. AEngine은 FlashAttention 방식에 따라 QKT, PV 연산을 처리하며, KV 캐시는 on-chip에 저장된다.

또한 FASU는 이전 스테이지의 K/V를 재사용하고, Q만 계산에 포함함으로써 연산량을 줄이며, 타일링 및 세그먼트화된 KV 캐시 구조는 DRAM 접근을 93.8%까지 줄이고 on-chip 메모리 사용량도 크게 감소시킨다. 전체적으로 기존 FlashAttention 대비 최대 3.1× 속도 향상을 달성하였다.

비정형 희소성에 대응하기 위한 SAU 구조는 최대 87.5% sparsity를 지원하며, prefix popcount 기반의 정렬 방식과 weight-stationary 구조를 통해 3.2× 처리량 개선을 이끌었다. 28nm 칩 측정 결과, 최대 52.03 TOPS/W의 연산 효율과 기존 SOTA 대비 최대 2.54× 에너지 효율을 달성했다.

#18-5 본 논문은 CNN과 Transformer 모델을 모두 고효율로 가속할 수 있는 22nm 공정 기반 CIM-utilization-aware 가속기를 제안한다. 다양한 연산 차원과 접근 패턴을 갖는 최신 네트워크 모델을 위해, 재구성 가능한 4D-CIM 구조, 유연한 feature reuse 메모리 시스템, 고속 벡터 프로세서 등을 결합하였다.



[그림 1] 제안된 재구성 가능한 가속기의 전체 아키텍처

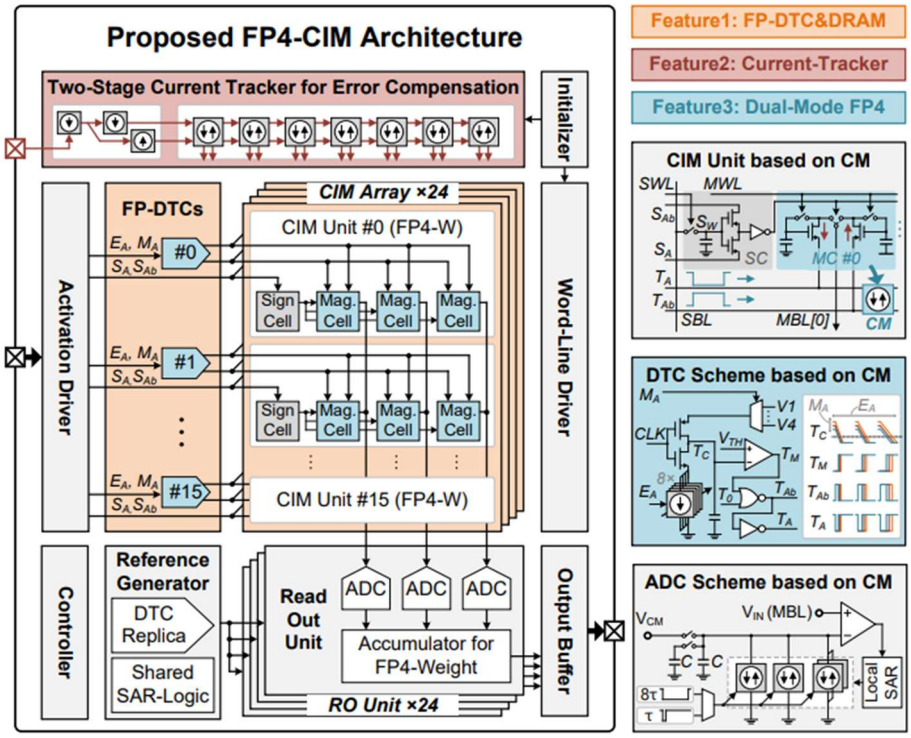
Fig. 1는 전체 아키텍처를 보여주며, 본 구조가 4D-CIM 프로세서, 글로벌 버퍼, SIMD 벡터 프로세서, DMA 모듈, 설정 레지스터 등으로 구성되어 있음을 나타낸다. 4개의 CIM 코어는 각각 9개의 64×256 SRAM 기반 CIM 매크로를 포함하고 있으며, H-tree 기반 네트워크를 통해 동적으로 병렬도를 조절할 수 있다. 입력과 출력 feature는 dispatcher와 collector를 통해 관리되며, SIMD 벡터 프로세서는 Softmax, LayerNorm, Pooling 등 다양한 연산을 수행한다.

본 구조는 다양한 연산자에 맞춰 병렬도를 적응적으로 조절할 수 있어, ResNet-50과 BERT-base에서 기존 고정형 구조 대비 각각 2.6배, 1.5배의 CIM 활용률 향상을 달성하였다. 또한 feature reuse를 위한 유연한 메모리 계층 설계로 VGG-16과 BERT-base에서 최대 47%의 메모리 접근 감소 효과를 보였다. Softmax 및 LayerNorm 같은 연산에서는 online 알고리즘 기반의 벡터 프로세서를 적용하여 지연 시간은 59%, 메모리 접근은 33% 줄였다.

최종적으로 22nm에서 구현된 본 칩은 VGG-16 기준 최대 29.3TOPS/W, BERT-base 기준 25.4TOPS/W의 에너지 효율을 달성하였으며, CNN과 Transformer 모두를 고효율로 가속할 수 있는 최초의 범용 CIM 기반 구조로 평가된다.

#18-6 본 논문은 Edge 환경에서 대규모 언어 모델(LLM)의 효율적인 추론을 가능하게 하기 위해, FP4 기반의 One-Shot Compute-in-Memory(CIM) 매크로를 제안한다. 최근 LLM의 복잡성과 매개변수 수가 급격히 증가함에 따라, 클라우드에서 에지로의 실행 전환이 요구되고 있으나, 에지 디바이스는 제한된 전력과 저장 용량으로 인해 고정밀 연산을 감당하기 어렵다. FP4는 INT8 대비 더 나은 정확도를 제공하면서도 절반의 저장 공간만 요구하여, LLM 추론을 위한 유망한 저정밀 포맷으로 주목받고 있다.

이 논문에서 제안하는 아키텍처는 입력값을 시간 신호로 변환하는 FP-DTC, 곱셈 누산을 수행하는 CIM array, 아날로그 출력을 디지털로 변환하는 Read-out 유닛, PVT 및 mismatch 보정을 위한 2단계 Current Tracker(CT), 그리고 입출력 스케줄링을 담당하는 제어 로직 블록으로 구성된다. 특히 CIM array는 부호 셀과 크기 셀로 구성된 전류 기반 아날로그 구조로, one-shot 방식의 FP MAC 연산을 수행하며, 이 구조는 이후 소개될 세 가지 핵심 기술의 기반이 된다. 전체 구성은 Fig. 2에 나타나 있다.



[그림 1] 제안된 FP4-CIM 매크로의 전체 아키텍처

기존 FP-CIM은 주로 FP16/FP32 포맷에 최적화되어 있어 FP4 연산에 비효율적이며, pre-alignment에 의한 전력 소모, 아날로그 회로의 PVT 및 mismatch 민감성, LLM 계층별 hybrid-FP4 처리가 어려운 문제가 있다.

이를 해결하기 위해 본 논문은 세 가지 핵심 기술을 통합하여 설계된 아키텍처를 제안한다. 먼저, 입력값과 가중치를 아날로그로 변환해 곧바로 곱셈-누산(MAC) 연산을 수행하는 One-Shot FP-MAC 구조를 도입하였다. 이 구조에서는 FP-DTC가 디지털 FP4 입력을 시간 신호로 변환하고, 변환된 신호는 24개의 CIM array로 전달되어 eDRAM 기반 전류 연산을 거친다. 연산 결과는 Read-out 유닛을 통해 디지털로 변환된 후 병합되며, 전체 연산 경로의 전류 일관성은 CT 회로를 통해 보정된다.

또한, **전류 기반 연산 경로 전반의 PVT 변화 및 디바이스 mismatch에 대응하기 위한 2단계 전류 보정 회로(CT)**를 설계하였다. 이 회로는 Current Memory 기반으로 모든 연산 유닛의 전류를 정규화하며, 공급 전압의 ±10% 변동과 –20℃~85℃의 온도 변화에도 MAC 출력의 안정성을 유지한다. 이를 통해 칩 간 편차는 1LSB 이하로 억제된다.

마지막으로, LLM 계층별 정밀도 요구에 대응할 수 있도록 Dual-mode Hybrid-FP4 MAC 구조를 채택하였다. 제안된 구조는 E2M1과 E1M2 포맷을 모두 지원하며, FP-DTC와 FP-DRAM의 간단한 설정 변경만으로 포맷 전환이 가능하다. 이를 통해 다양한 정밀도 요구

에 유연하게 대응하면서도 양자화 오차를 최소화할 수 있다.

28nm CMOS 공정에서 구현된 본 매크로는 0.012mm²의 면적, 0.8~1.0V 구동 전압에서 동작하며, 최대 581.8 TFLOPS/W 및 9.1 TFLOPS/mm²의 에너지·면적 효율을 기록하였다. 기존 INT8-CIM 대비 최대 5.2배 더 높은 에너지 효율과 2.2배의 면적 효율, 그리고 더 우수한 LLM 추론 정확도를 달성하였다.

따라서 본 논문은 one-shot 연산, 전류 기반 보정, hybrid-FP4 지원이라는 세 가지 기술을 통합함으로써, 에지 LLM 추론을 위한 고효율·고정밀·고신뢰 하드웨어 솔루션을 제시한다.

저자정보

박민하 석사과정 대학원생

- 소속 : KAIST
- 연구분야 : 디지털 회로 설계
- 이메일 : mhpark@ics.kaist.ac.kr
- 홈페이지 : https://idec.or.k

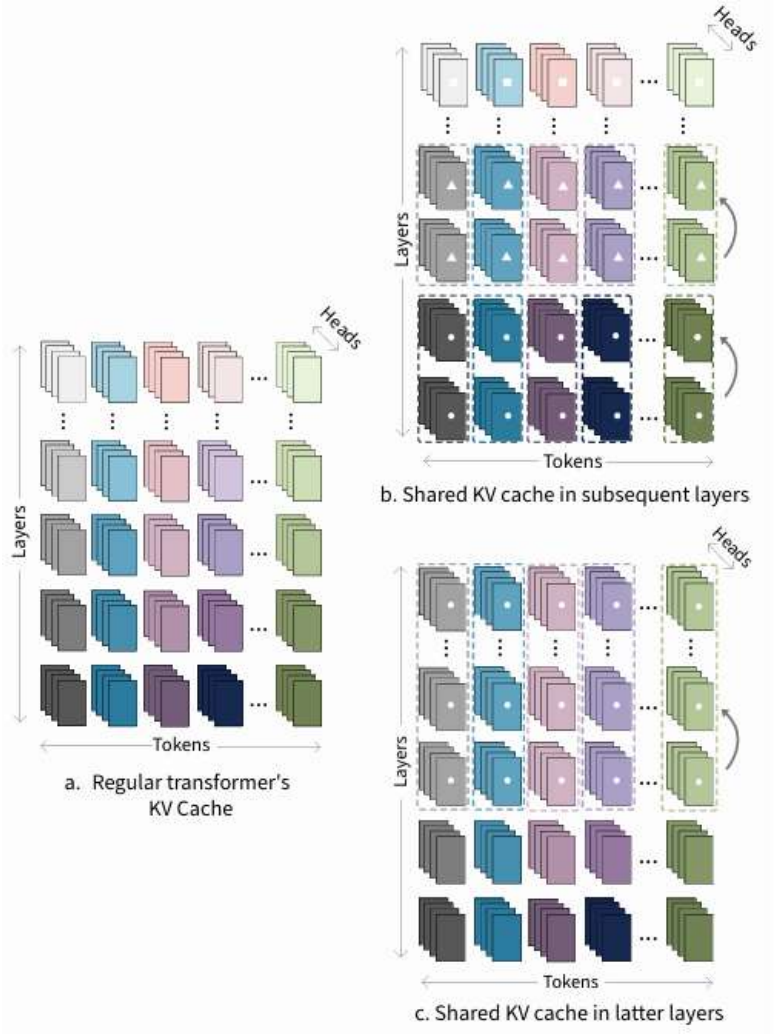
2025 IEEE CICC Review

KAIST 전기및전자공학부 석사과정 권재훈

Session 37 Machine Learning and Energy Efficient SoCs

이번 2025 IEEE CICC의 Session 37은 Machine Learning and Energy Efficient SoCs라는 주제로 총 8편의 논문이 발표되었다. 이 세션에서는 고성능 신호 처리 및 효율적인 메모리 활용을 통해 inference throughput과 energy efficiency를 극대화하는 데 중점을 두었다.

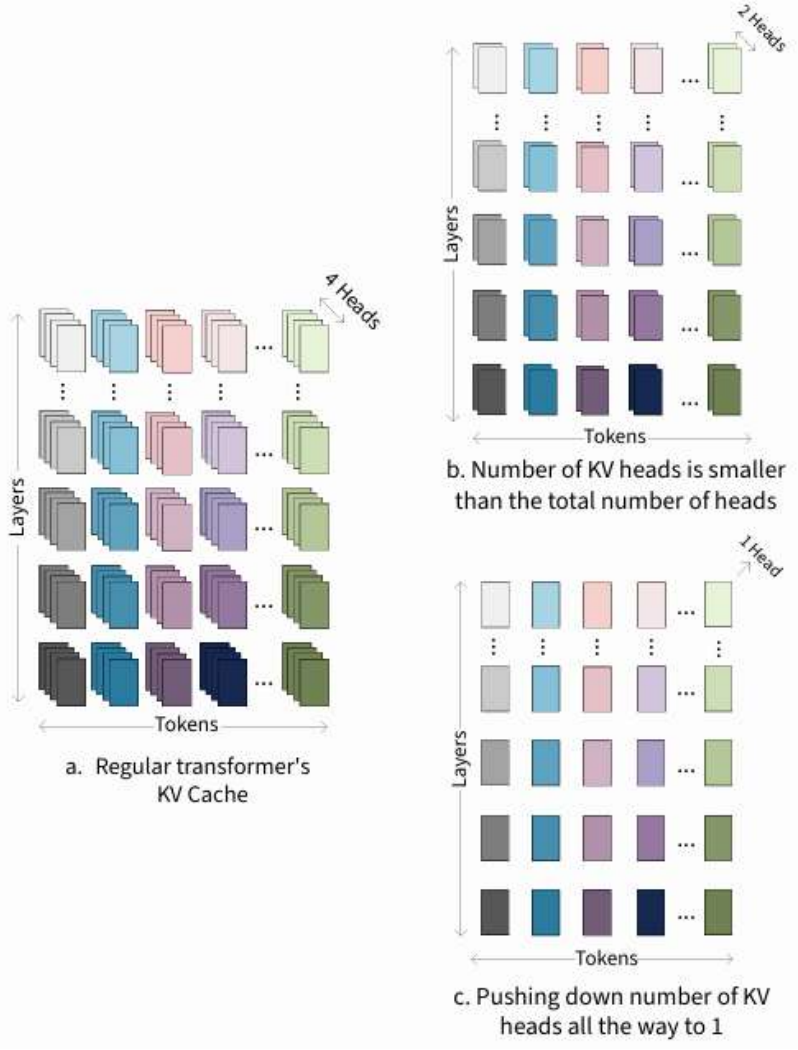
#37-1 본 논문은 UC San Diego와 NVIDIA의 공동 연구로, Key-Value (KV) cache compression에 대한 분석과 implementation strategy를 다룬 논문이다. 최근 Large language models (LLMs)에서 긴 context를 효율적으로 처리하는 것이 점점 더 중요해짐에 따라, KV cache의 저장 및 연산 효율을 개선하기 위한 다양한 방법들이 제안되어 왔다. 본 연구에서는 KV cache compression method를 layer, head, token, hidden dimension이라는 4개의 category로 분류하고, 각각의 method들이 inference latency와 model 정확도에 미치는 영향을 평가하였다. 특히 본 논문에서는 post-training과 training-free 방식 모두를 포함하며, 실제 LLM benchmark를 통해 compression 기법 간 trade-off를 정량적으로 비교하고자 하였다. 실험 결과, YOCO 기반 cross-layer compression 기법으로 latency를 최대 2.87배 가속할 수 있었으며, CPC 및 LongLLMLingua 같은 training-free token compression 기법은 end-to-end inference를 최대 10.93배 까지 단축함으로써, long-context scenarios에서 GPU memory footprint를 획기적으로 절감하고 inference throughput을 개선할 수 있음을 확인했다.



[그림 1] Cross-Layer Attention이 적용된 KV Cache

[그림 1]은 Cross-Layer Attention을 통해 KV cache를 layer 간에 공유하는 세 가지 방식을 보여준다. regular transformer는 각 layer가 독립적으로 KV cache를 저장하고, subsequent layers 방식은 첫 layer의 key-value를 이후 모든 layer에서 재사용하며, latter layers 방식은 decoder 후반 layer에만 공유 범위를 제한해 메모리와 연산을 줄이면서 성능 저하를 최소화한다. 본 논문의 contribution을 구체적으로 정리하면 다음과 같다. 일단 Layer dimension에서는 Cross-Layer Attention을 통해 KV cache를 공유하는데, subsequent layers sharing 방식과 latter layers sharing 방식을 도입하여 redundant storage와 compute overhead를 줄였다. 다음으로 dimension에서는 Multi-Query Attention을 통해 모든 head에 하나의 key/value를 공유하거나, Grouped-Query Attention을 통해 head를 그룹별로 묶어서 key/value 수를 줄임으로써 메모리 사용량을 감소시켰다. 그리고 Token dimension에서는 Compressed Past Context (CPC)와 LongLLMLingua를 활용하여 유사도가 낮은 과거 Token 의 KV entry를 pruning하거나 summarization하여 Token 수를 크게 축소함으로써 end-to-end inference latency를 단축하였다. 마지막으로 Hidden Dimension에

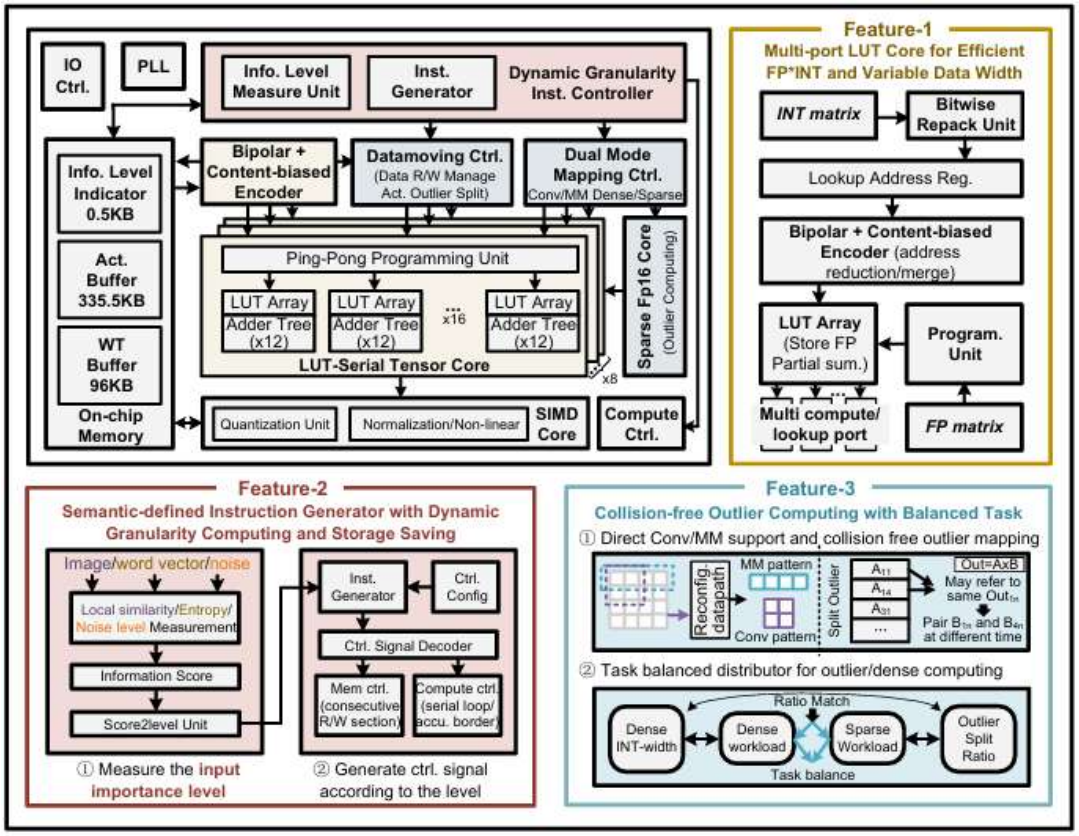
서는 low-rank factorization과 8-bit quantization 기반의 압축 기법을 적용해 KV representation의 차원을 축소하고 GPU memory footprint를 효과적으로 낮추었다.



[그림 2] Head Dimension에서의 KV Cache Compression (Multi-Query & Grouped-Query Attention)

#37-4 본 논문은 Tsinghua University에서 발표한 연구로, 비디오 생성을 위한 Content Creation Engine (CCE)의 회로 구현 및 최적화를 다룬다. 최근 diffusion model, transformer, super-resolution (SR) model과 같은 대규모 생성 모델들이 고화질 동영상 생성에 활용됨에 따라, 이를 효율적으로 가속할 수 있는 저전력, 고성능 하드웨어 설계가 요구되고 있다. 본 연구에서는 asymmetric INT*FP 연산을 지원하는 LUT-based tensor core와 semantic-driven instruction generator, collision-free outlier mapper를 통합한 구조를 제안하여, heterogeneous 연산 특성을 통합적으로 처리할 수 있는 아키텍처를 구현하였다. 특히 LUT-serial core는 bipolar-content encoding 및 multi-port 구조를 통해 power와 area overhead를 최소화하였으며, input redundancy 제거를 위한 information-level

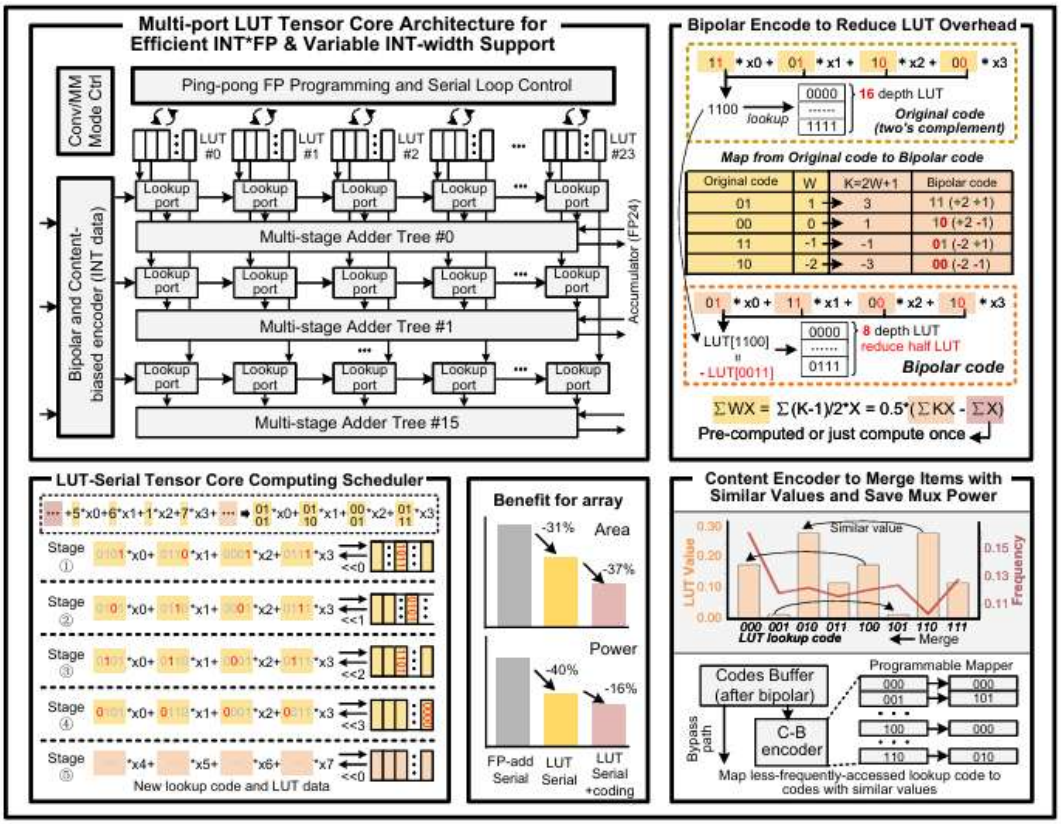
scoring과 granularity control을 통해 연산 효율을 극대화하였다. 실험 결과, diffusion task에서 3.58배, SR에서 2배, transformer에서는 12%의 평균 energy efficiency 향상시켰고, inference power는 최대 87.6%까지 감소되었다.



[그림 3] Head Dimension에서의 KV Cache Compression (Multi-Query & Grouped-Query Attention)

본 논문의 contribution을 구체적으로 정리하면 다음과 같다. 먼저 bipolar-content encoding 및 multi-port LUT-serial tensor core 기반의 FP*INT computing cell을 도입하여, 입력 feature를 ±1 bit로 표현함으로써 accumulation 연산을 XNOR 및 popcount 연산으로 대체하고, multi-port LUT-serial tensor core 구조를 통해 하나의 LUT에서 여러 input stream을 serial processing하도록 설계하여, 기존 FP MAC와 비교했을 때 area 및 power 절감을 동시에 달성하였다. 또한 semantic-defined instruction generator를 도입하여, 모델 레벨에서 feature map의 channel 중요도를 semantic score로 계산한 뒤 낮은 점수의 채널 블록을 pruning하고, 남은 블록에 대해 dynamic granularity control을 적용하여 연산 분해 수준을 조정함으로써 instruction issue 시점에 필요한 compute와 on-chip storage를 최소화하며, 16-entry deep FIFO와 4-stage pipelined control path로 매 사이클 최적화된 명령어 스트림을 연속 발행할 수 있도록 설계하였다. 그리고 Collision-free Sparse Mapper 기반의 unified dense/outlier architecture를 제안하여, input feature의 non-zero sparsity 정보를 사전 분석해 8-bank interleaved on-chip SRAM의 write 주소를 remapping함으로써 dense core와 outlier core 간 memory bank write 충돌을 방지했다.

마지막으로 Diffusion, Super-Resolution, Transformer 워크로드별로 compute와 memory 요구 특성을 분석하여 adaptive resource allocation 기반의 balanced scheduling 알고리즘을 제안하고, 여기에 dynamic voltage and frequency scaling (DVFS)를 결합하여 각 task의 critical path에 맞춰 전압, 주파수를 실시간 조정함으로써 energy efficiency 개선과 높은 throughput을 구현하였다.



[그림 4] bipolar-content encoder를 활용한 Multi-port LUT-serial core

저자정보

권재훈 석사과정 대학원생

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

2025 IEEE CICC Review

KAIST 인공지능반도체대학원 석사과정 윤지원

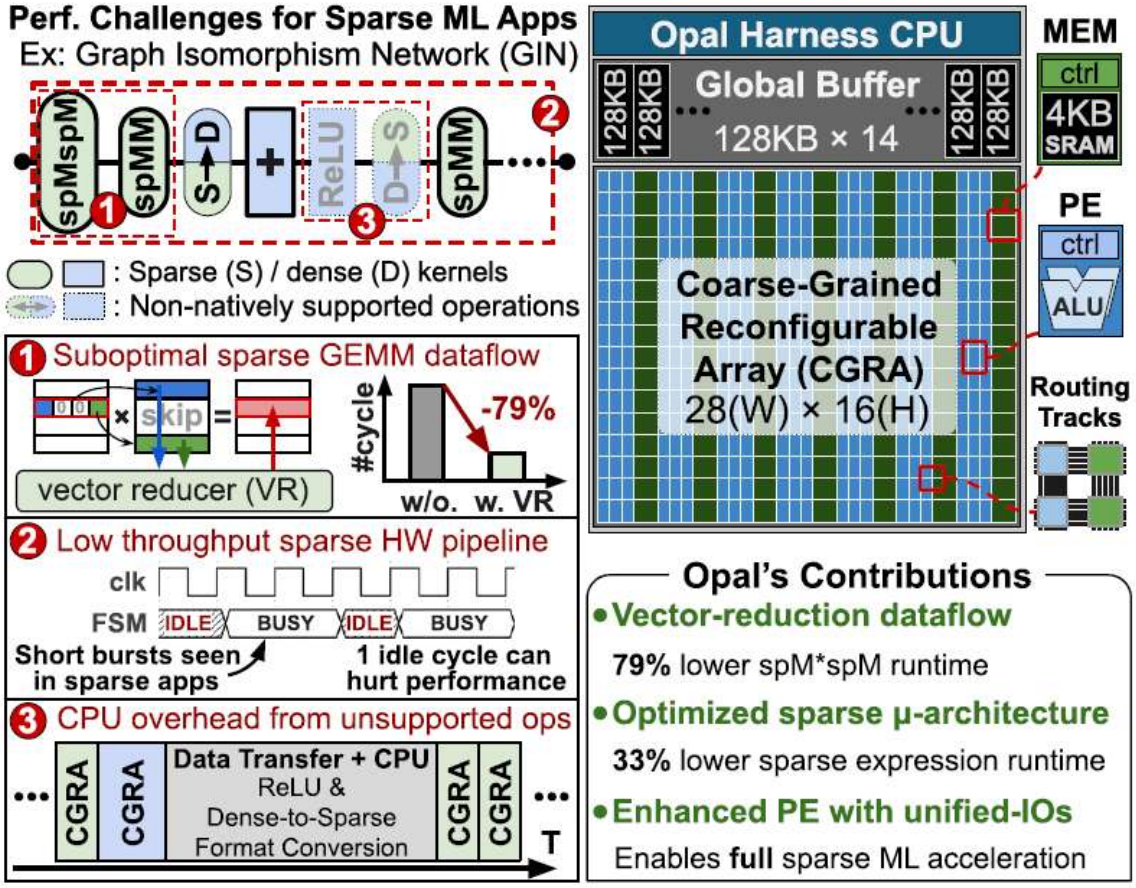
Session 37 Machine Learning and Energy Efficient SoCs

이번 2025 IEEE CICC Session 37은 Machine Learning and Energy Efficient SoC라는 주제로 총 8개의 논문이 발표되었다. 이 Section은 인공지능 연산의 고도화와 함께, 엣지 디바이스 및 임베디드 시스템에서의 전력 효율성과 연산 성능 간의 균형을 달성하기 위한 다양한 하드웨어 최적화 연구를 중심으로 소개되었다.

#37-8 본 논문은 스탠포드 대학교에서 발표한 내용으로, Gustavson dataflow와 통합 PE(Processing Element)를 이용하여 sparse와 dense 연산 모두를 효율적으로 처리할 수 있는 Opal을 제안한다. 최근 ML 모델은 정확도를 높이기 위하여 연산량과 자원 소모가 크게 증가하고 있다. 이에 입력과 가중치에 Sparse 연산을 도입함으로써 이러한 비용을 줄이고자 하였으나 기존 Sparse Accelerator는 특정 모델에 종속되며, CGRA (Coarse-Grained Reconfigurable Array)는 dense 연산에 최적화되어 있어 sparse 연산에서 성능 저하가 발생하는 문제가 있었다.

본 논문에서 제안하는 Opal은 sparse ML 연산을 위해 특화된 CGRA SoC로 sparse, dense 연산을 함께 가속할 수 있는 유연성을 가지며 그림 2에서 제시된 바와 같이 다음의 세 가지를 주요 기여로 제안한다. 첫 번째, vector reducer primitive를 도입해 Gustavson dataflow를 지원하여, sparse matrix multiplication에서도 최대 79% runtime감소를 보였다. 두 번째, Coordinate Dropper, Repeater, Level Scanner, Level Writer 등에서 발생하는 주요 Pipeline Bubble을 제거해 최대 33%의 성능 향상을 달성했다. 셋째, 통합형 PE구조를 통해 RELU, SoftMax와 같은 비선형 연산 및 dense와 sparse 사이 데이터 변환을 CPU 호출 없이 직접 수행 가능하게 하여 최대 89%의 실행 시간 단축 효과를 얻었다.

이러한 구조는 실제 sparse ML 모델인 GCN 과 GIN에도 적용되어 각각 68%, 79%의 runtime과 에너지 소비 감소를 이끌었으며, 주요 8개 sparse 연산 kernel에서도 평균 35%의 실행 시간 단축을 보였다. 결론적으로 Opal은 높은 동작 주파수 (720MHz)와 유연한 연산 지원을 바탕으로, 차세대 sparse ML 가속을 위한 유망한 CGRA 기반 아키텍처로 자리매김한다.



[그림 1] Opal의 SoC아키텍처와 sparse ML 애플리케이션의 성능 문제를 해결하기 위한 본 논문의 기여 내용

#37-6 본 논문은 카네기 멜론 대학교에서 발표한 내용으로, 다양한 연산 집약적 애플리케이션을 효율적으로 처리할 수 있도록 RISC-V CPU와 고도로 최적화된 eFPGA를 밀접하게 통합한 SoC 아키텍처를 제안한다. 기존의 fixed-function accelerator는 유연성이 부족하고, 범용 eFPGA는 성능과 에너지 효율이 낮은 한계가 있었던 반면, 본 아키텍처는 두 방식의 장점을 통합함으로써 높은 성능과 유연성을 동시에 확보할 수 있도록 하였다.

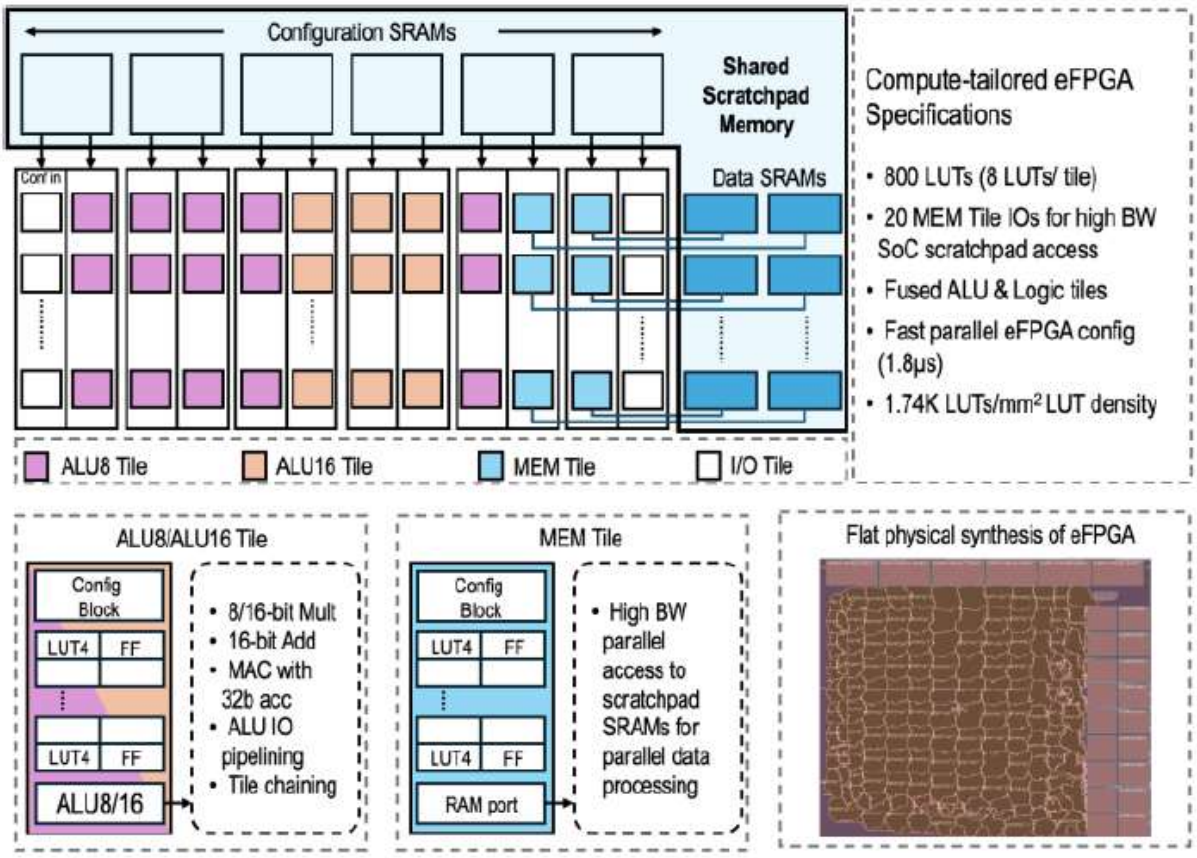
제안된 SoC는 VexRISC-V CPU와 eFPGA를 CFU (Custom Function Unit) 인터페이스로 연결하고, 비동기 FIFO를 통해 서로 다른 클럭 도메인에서도 안정적인 통신이 가능하도록 설계되었다. 또한, eFPGA 내부에 ALU8/ALU16 유닛 (MAC + Adder + Accumulator)을 logic tile에 직접 융합하여, 기존의 LUT구조보다 훨씬 높은 계산 성능과 밀도를 확보하였다. 이러한 구조는 22.3GOPS/mm^2의 연산 밀도를 달성하며, ALU chaining과 Pipelining을 통해 260MHz의 고속 동작이 가능하다.

Scratchpad Memory (96KB)는 CPU와 eFPGA 사이에서 고속 병렬 데이터 공유를 가능하게 하며, DMA (Direct Memory Access) 와 CFU Manager를 통해 bitstream 전송 및 제어

가 자동화되도록 설계되었다 특히 scratchpad를 통해 eFPGA 전체를 병렬로 구성함으로써 1.8μs 이내의 초고속 재구성이 가능하며, 이는 문헌상 최단 시간으로 초당 1000회까지 애플리케이션 전환이 가능하다.

실제 구현된 애플리케이션 성능 측정 결과, NTT(Number Theoretic Transform) – 256은 CPU 대비 29배의 처리량, 66배의 에너지 효율, FFT(Fast Fourier Transform) – 64는 각각 108배, 228배 향상을 기록하였다. FIR 필터와 CRC32 또한 RISC-V기반 소프트웨어 구현 대비 높은 성능을 보였다.

마지막으로 본 연구는 기존 상용 및 학계의 eFPGA SoC들과 비교해, 에너지 효율(747.83 GOPS/W for INT8), 연산 밀도(22.3 GOPS/mm^2), 재구성 속도 (1.8μs) 측면에서 모두 최고 수준의 결과를 달성하였으며, 이는 다양한 엣지 연산 환경에서 실시간 가속기 전환 및 높은 처리 성능이 요구되는 응용에 매우 적합함을 입증하였다.



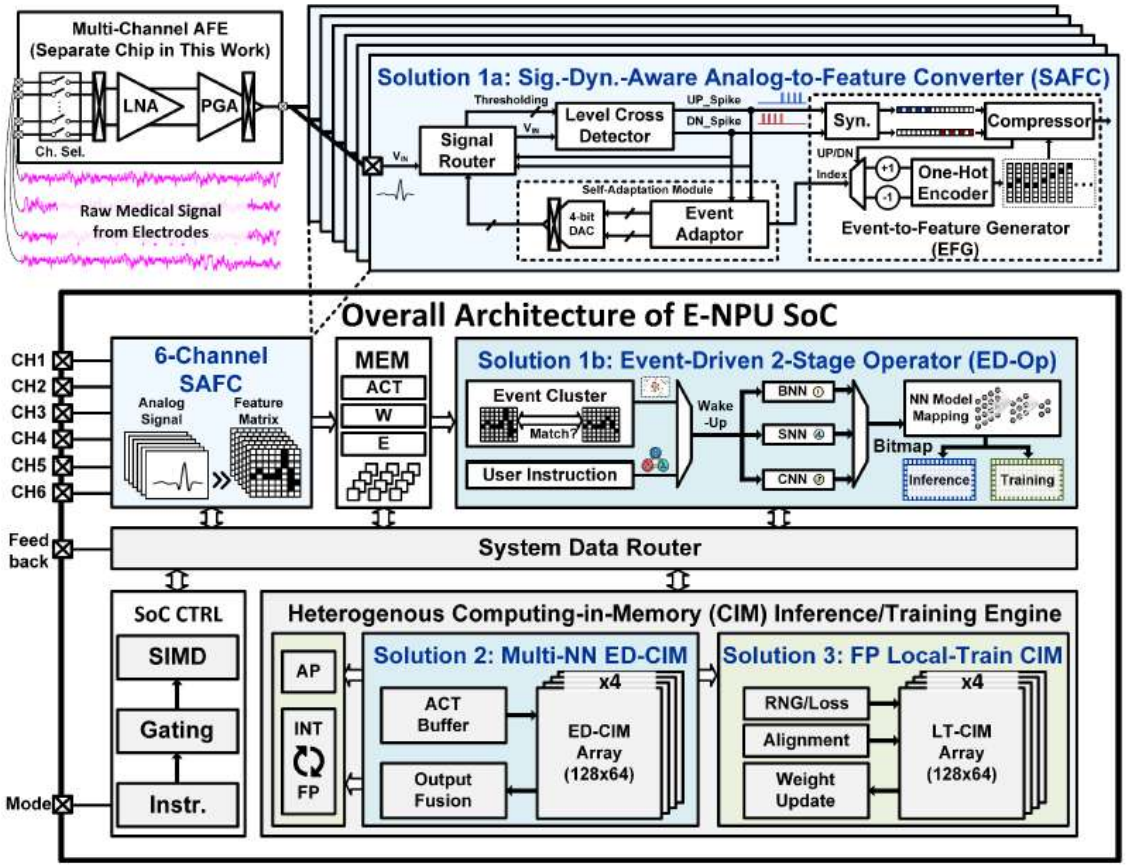
[그림1] 통합 ALU 및 scratchpad 접근 기능을 갖춘 연산 중심 eFPGA 블록 다이어그램

#37-7 본 논문은 칭화대학교에서 발표한 내용으로, 신호 변화에 따라 연산을 수행하고, In-Memory 추론 및 환자 맞춤형 On-Chip 학습을 지원하는 초 저전력 의료 웨어러블용 neural SoC(E-NPU)를 제안한다. 의료 웨어러블 기기는 생체 신호를 장시간 모니터링하고 정확하게 해석해야 하므로 낮은 에너지 소비, 높은 정확도, 사용자 맞춤형 학습이 가능한

연산 구조를 요구한다. 최근에는 BNN, SNN, CNN 등의 경량 신경망과 NAS(Neural Architecture Search) 기반 설계 기법이 발전하고 있지만, 신호 특성에 따라 유연하게 작동하고, 다양한 신경망 모델을 동시에 지원하며, 환자별 특성에 적응할 수 있는 SoC 설계는 여전히 도전 과제로 남아있다.

이를 해결하기 위해 E-NPU는 그림 1과 같이 설계되었으며, 다음 세가지 핵심 기술을 통합하였다. 첫째, SAFC (Signal-Dynamics-Aware Analog-to-Feature Converter) 기반의 동적 신호 처리 아키텍처를 통해 신호 변화가 있을 때만 연산을 수행함으로써, 기존의 항상-활성화 방식 대비 평균 1.7배 이상의 에너지 절감을 달성한다. 둘째, 다양한 정밀도 설정이 가능한 다중 모델 In-Memory 연산 (ED-CIM) 구조를 통해 BNN, SNN, CNN을 모두 처리하며, 병렬 Column 구조와 ping-pong SRAM 기반의 파이프라인으로 고속 및 고효율 연산을 지원한다. 셋째, Direct Feedback 기반의 On-chip 학습 엔진 (LT-CIM)을 통해 cross-layer dependency 없이 layer 단위로 병렬 학습을 수행하며 정밀도 변환, 에러 예측, 가중치 업데이트까지 전체 학습 흐름을 칩 내부에서 완결할 수 있다.

E-NPU는 40nm CMOS 공정으로 제작되었으며, 수정된 LeNet 모델 기반 실험에서 추론 지연 3.03배 감소, 학습 지연 2.49배 감소, 외부 메모리 접근 3.41배 감소 등의 성능 향상을 보였고, EEG, EMG, ECG 관련 4개 공개 데이터셋에서 ED-Op 기반 처리로 평균 1.7배 이상의 에너지 효율을 보였다. 결론적으로, E-NPU는 신호 변화 기반 연산, 다중 정밀도의 In-Memory 추론, 환자 맞춤형 On-Chip 학습을 하나의 SoC에 통합하여, 개인 의료 웨어러블 기기에 최적화된 고성능 및 초저전력 뉴럴 하드웨어 플랫폼을 구현하였다.



[그림 1] 제안된 event-driven neural SoC (e-NPU)의 전체 아키텍처 및 세 가지 기여점

저자정보

윤지원 석사과정 대학원생

- 소속 : 한국과학기술원 (KAIST)
- 연구분야 : 디지털 회로 설계
- 이메일 : jwyoon@kaist.ac.kr
- 홈페이지 : https://ics.kaist.ac.kr