2024 IEEE VLSI Review

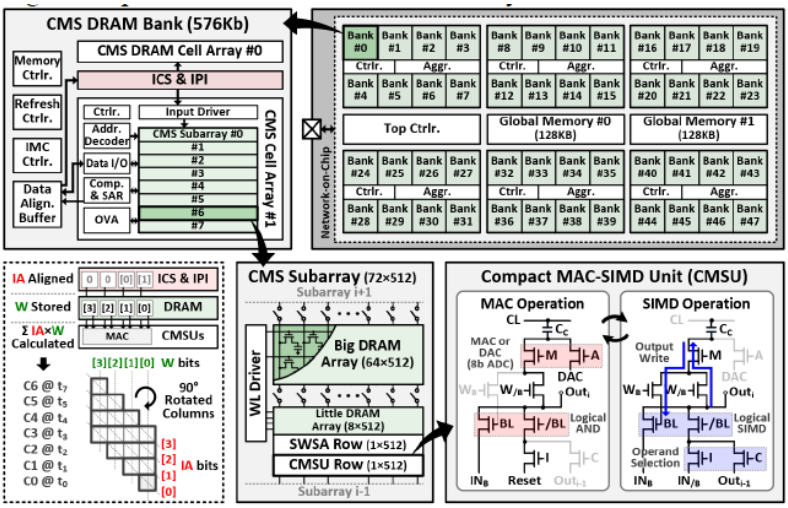
한양대학교 신소재공학과 석박통합과정 송충석

Session 20 Processing for AI

이번 2024 IEEE CICC의 Session C20은 Processing for AI 라는 주제로 총 4편의 논문이 발표되었다. 본 review에서는 20-1, 20-2, 20-3을 리뷰하고자 한다.

#20-1 논문에서는 Dyamond라는 새로운 1T1C DRAM 기반 CIM을 발표했다. Dyamond는 column방향으로 accumulation을 하는 column addition (CA) 데이터플로우를 사용하고 높은 메모리 사용 효율성과 에너지 효율성을 기록하였다. LSB쪽 연산에 사용되는 LSB-CA는 ADC 사용을 최소화 하여 에너지 효율을 높이고, MSB쪽 연산에 사용되는 MSB-CA는 signal enhanced MAC과 signal shifted ADC를 통해 SQNR을 향상시켰다. 또한 switchable sense amplifier를 사용하여 저전력 기반 CIM 연산을 구현하였다.

28nm CMOS 기반으로 제작된 본 논문의 Dyamond는 6.48mm^2 면적에 27Mb DRAM 메모리를 implementation하여 최대 27.2 TOPS/W의 에너지 효율성을 달성하였으며 ResNet, BERT, GPT-2 와 같은 최신모델에서도 성능을 입증하였다. 기존의 eDRAM 기반 CIM과 비교하여 Dyamond는 1.8배 높은 효율성을 보였고, figure of merit은 7.9배 높였다.

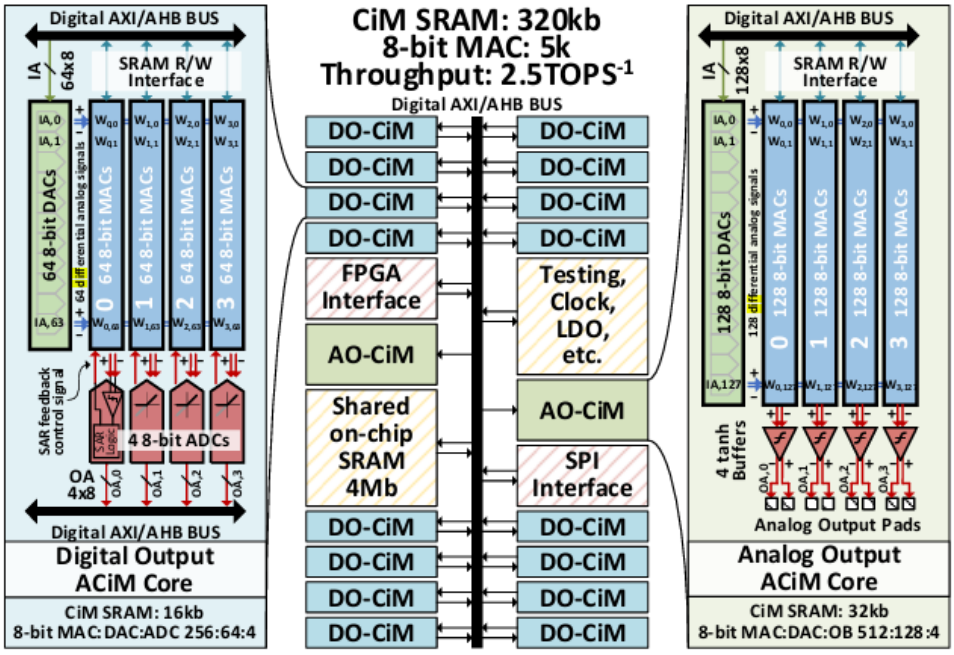


[그림 20-1] 논문 20-1에서 제안한 전체 아키텍쳐

#20-2 논문에서는 아날로그 컴퓨팅 방식 CIM에서 발생하는 빈번한 데이터 변환 문제를 해결하기 위해 signed 8bit MAC 연산기를 새롭게 제안하였다. 더불어 ReLU는 SAR-ADC에서 LSB 스킵 방식으로, tanh은 아날로그 버퍼를 통해 구현하여 이를 MAC 연산기에 통

합시켜 효율성을 극대화시켰다.

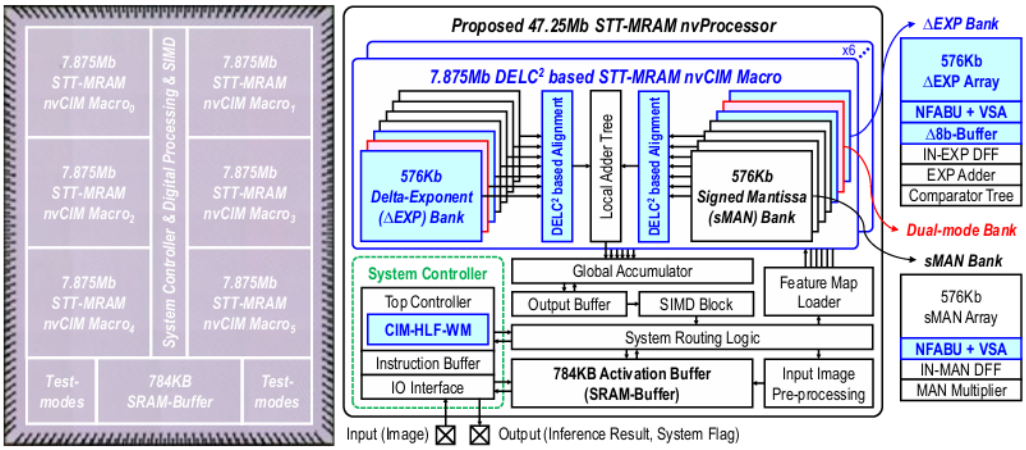
본 논문에서 제안한 프로세서는 디지털 및 아날로그 연산코어를 포함한 하이브리드 형태를 띄고 있으며 2.5TOPS의 throughput을 달성하였다. 64개의 8bit DAC, 16Kb SRAM 셀, 256개의 8bit MAC 유닛 및 4개의 8bit ADC가 통합된 DO-CIM 코어와, 데이터 변환기를 사용하지 않고 tanh 활성화 함수를 사용할 수 있는 AO-CIM 코어를 통해 에너지 효율과 면적 효율을 상승시켰다. 실험 결과, 55TOPS/W의 성능을 보이며, AO-CiM 코어에서는 최대 104.5 TOPS/W 를 기록하였다.



[그림 20-2] 논문 20-2 에서 제안한 전체 아키텍쳐

#20-3 논문은 22nm CMOS 공정 기반 AI 엣지 프로세서를 제안하였다. 비휘발성 메모리인 STT-MRAM을 사용하여 near memory computing을 구현하여 47.25Mb 메모리의 21.4TFLOPS/W 의 에너지 효율성을 기록하였다. 연산 효율성을 위해 compression을 데이터 손실 없이 구현하였다.

본 논문에서 개발한 프로세서는 428.58us의 짧은 응답속도를 가지며 압축된 데이터를 만들기 위한 하드웨어 오버헤드를 줄이고 SRAM 버퍼 접근을 감소시켜 전체적인 시스템 성능을 향상시켰다. 이러한 방법으로 인해 기존 비휘발성 메모리 프로세서보다 2.55배 이상 높은 시스템 수준 에너지 효율성을 달성했다.



[그림 20-3] 논문 20-3 에서 제안한 전체 아키텍쳐

저자정보

송충석 석박통합과정 대학원생

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

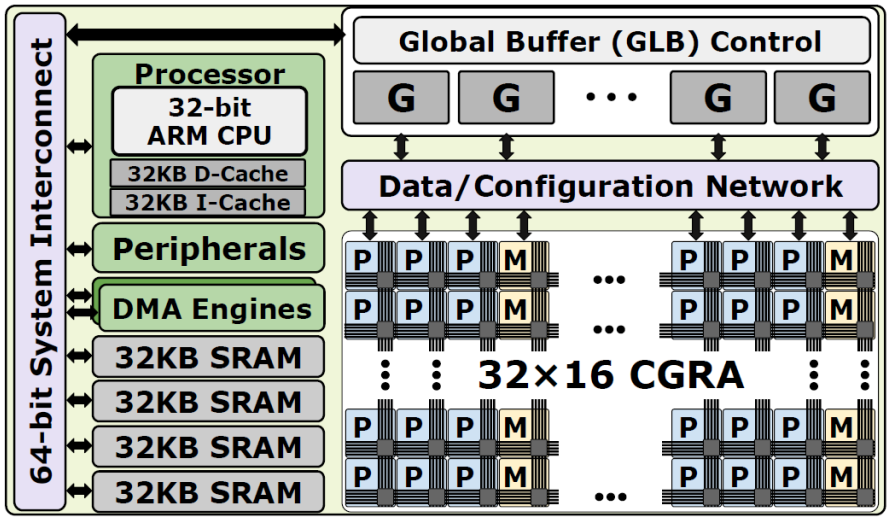
2024 IEEE VLSI Review

KAIST 전기및전자공학부 박사과정 엄소연

Session 7 Processors I

이번 2024 VLSI의 Session C7은 Processor라는 주제로 총 4편의 논문이 발표되었다. 이 세션에서는 dense/sparse application, diffusion model, stochastic analog SAT solver, 그리고 stencil/sparse linear algebra를 위한 하드웨어가 발표되었다. 이번 후기를 통해 4개의 논문에 대해 간략하게 살펴보고자 한다.

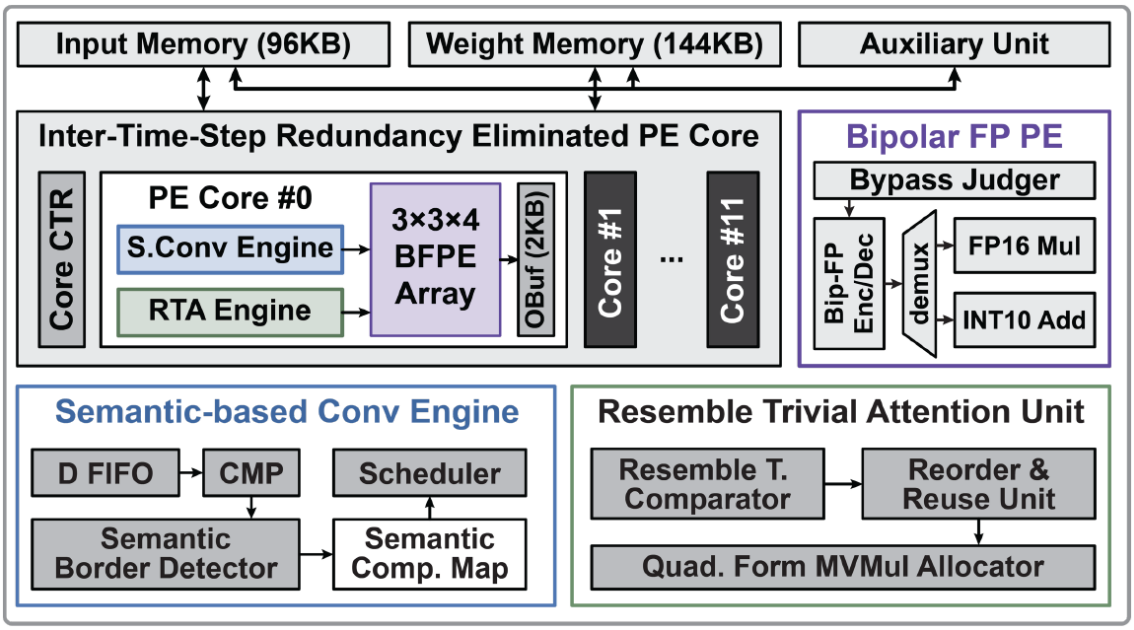
#7-1 이 논문은 Stanford University에서 발표한 연구로, Onyx라는 최초의 완전 프로그래머블 Coarse-Grained Reconfigurable Array (CGRA)를 소개한다. Onyx는 밀집(dense) 및 희소(sparse) 응용 프로그램 모두를 가속할 수 있으며, 임의의 고차원 텐서 연산을 지원한다. Onyx는 압축된 텐서를 처리할 수 있는 메모리 및 계산 프리미티브를 사용하여 불필요한 연산을 제거하고, 이미지 처리 및 머신러닝에서 효율성을 높인다. 28nm 공정에서 제작된 Onyx SoC는 384개의 프로세싱 타일과 128개의 메모리 타일을 사용하며, 메모리 타일은 피버트리(fibertree) 기반 스트리밍 구조로 희소 데이터를 효율적으로 처리한다. 제안된 시스템은 희소 행렬 곱셈에서 CPU 대비 최대 565배, 이미지 처리와 머신러닝에서 각각 76%, 85%의 에너지-지연 제품(EDP) 개선을 달성한다.



[그림 1] Onyx 시스템-온-칩 아키텍처

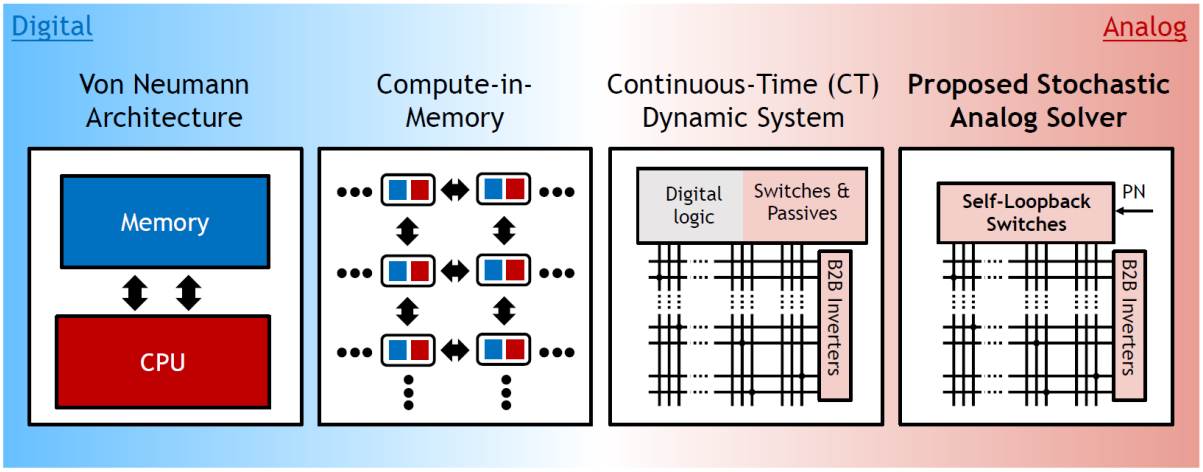
#7-2 이 논문은 Tsinghua University에서 발표한 에너지 효율적인 디퓨전 모델 프로세서를 소개하며, 22nm CMOS 공정에서 제작되었습니다. 이 프로세서는 시간 단계 간의 중복

을 제거하는 방식을 활용하여 성능과 에너지 효율을 크게 향상시킵니다. 주요 특징으로는 세 가지가 있습니다. 첫째, semantic-segment sparse convolution 엔진을 통해 중요하지 않은 이미지 부분에서 발생하는 중복된 convolution 연산을 88.5%까지 제거하여, 중요한 부분만 계산함으로써 연산 효율을 극대화합니다. 둘째, resemble trivial attention exponent inheritance 기법을 통해 어텐션 레이어의 계산 중복을 줄여 16.7배의 연산 효율을 달성합니다. 셋째, bipolar floating-point multiplier는 mantissa 곱셈의 비효율성을 줄여 25.4%의 연산량을 절감합니다. 결과적으로 이 프로세서는 CIFAR-10 및 ImageNet 데이터셋을 대상으로 실험되었으며, 실험 결과 평균 52.01 TFLOPS/W의 에너지 효율을 달성했습니다. 더불어 이미지 품질 손실은 1% 미만으로 유지되었습니다. 이 diffusion model 프로세서는 convolution layer와 attention layer에서 각각 11.76배, 22.51배의 에너지 효율 향상을 기록했으며, 기존 가속기 대비 최대 23.14배의 성능 향상을 보여줍니다.



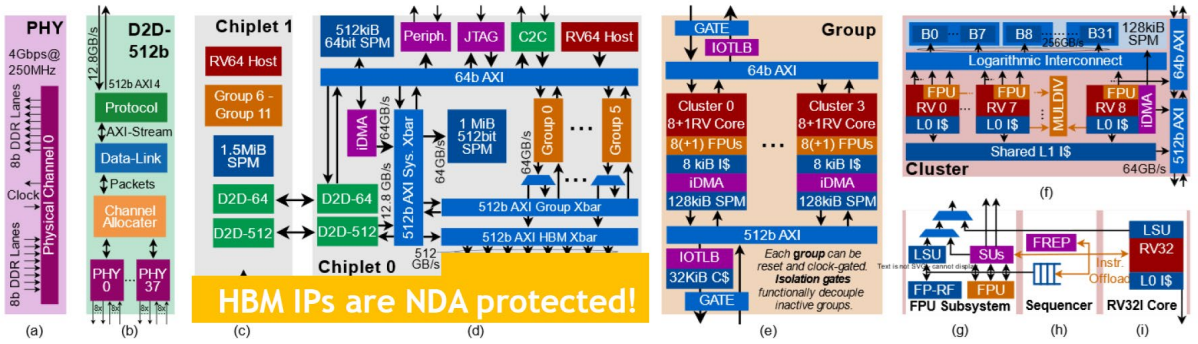
[그림 2] 디퓨전 모델 아키텍처

#7-3 이 논문은 65nm CMOS 공정에서 구현된 확률적 아날로그 SAT 솔버를 제시하며, 기존 디지털 및 아날로그 SAT 솔버에 비해 속도와 에너지 효율성을 개선했다. 제안된 솔버는 병렬 연산을 가능하게 하는 아날로그 오픈 루프 구조를 사용하며, 연속 시간 셀프 루프백 풀업 스위치와 Pseudo Noise(PN) 기반 스크램블링 기법을 도입해 해결 가능성을 높인다. 하이브리드 PN 생성기를 통해 비용을 절감하면서 성능 최적화를 달성했다. 제안된 솔버는 3-SAT 문제 해결을 위한 크로스바 구조를 기반으로 하며, 각 열은 변수, 각 행은 절을 나타낸다. 3비트 PN 제어 코드를 사용하여 절을 난수화된 상태로 제어하며, 이를 통해 문제 해결 가능성을 높인다. 그 결과, 기존 아날로그 SAT 솔버보다 1000배, 디지털 SAT 솔버보다 10배 빠른 속도를 기록했으며, 평균 해결 시간은 6.6μs였다.



[그림 3] SAT Solver 카테고리

#7-4 이 논문은 희소 선형대수와 스텐실 연산에 최적화된 432코어 RISC-V 기반 듀얼 칩렛 시스템 Occamy를 소개한다. Occamy는 FP64, FP32, FP16, FP8 연산을 지원하며, 희소하고 불규칙한 메모리 접근을 효율적으로 처리하기 위해 설계되었다. 각 칩렛은 16GiB HBM2E 메모리 스택과 216개의 RISC-V 코어로 구성되어 있으며, 클러스터당 64 GiB/s의 높은 메모리 대역폭을 제공한다. 이 시스템의 주요 특징은 다양한 precision 연산 코어와 희소 스트리밍 유닛을 사용하여 간접 주소 지정과 희소 데이터 병합을 가속화하는 것이다. 또한, 확장 가능한 지연 허용 아키텍처와 분산 DMA 유닛을 통해 데이터와 제어 트래픽을 효율적으로 처리할 수 있다. 실리콘 테스트 결과, Occam는 스텐실 연산에서 최대 571 GFLOP/s, 28.1 GFLOP/s/W의 성능을 기록했으며, 희소-밀집 행렬 연산에서 307 GFLOP/s, 희소-희소 연산에서는 187GCOMP/s의 성능을 보여주었다. 이 시스템은 CPU와 GPU 대비 최대 11배 높은 연산 밀도와 에너지 효율을 제공하며, 스텐실 및 희소 연산에서 뛰어난 성능을 보였다.

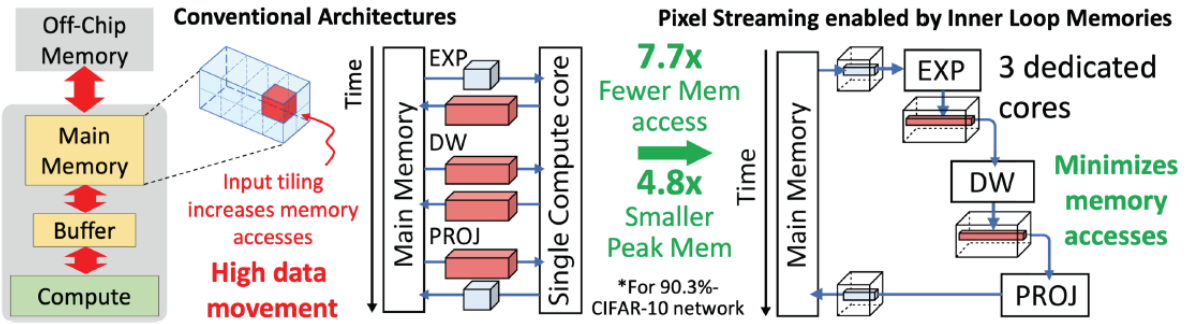


[그림 4] 듀얼 칩렛 아키텍처

Session 28 Processors II

2024 VLSI의 Session C28 또한 Processor라는 주제로 총 4편의 논문이 발표되었다. 이 세션에서는 tinyML 작업을 위한 프로세서, 에너지 효율적인 이종 SoC, 시각적 문맥 이해를 위한 Scene Graph Generation 프로세서, 그리고 고속 광통신을 위한 소프트 디시전 오류 수정 디코더가 발표되었다. 이번 후기를 통해 4개의 논문에 대해 간략하게 살펴보고자 한다.

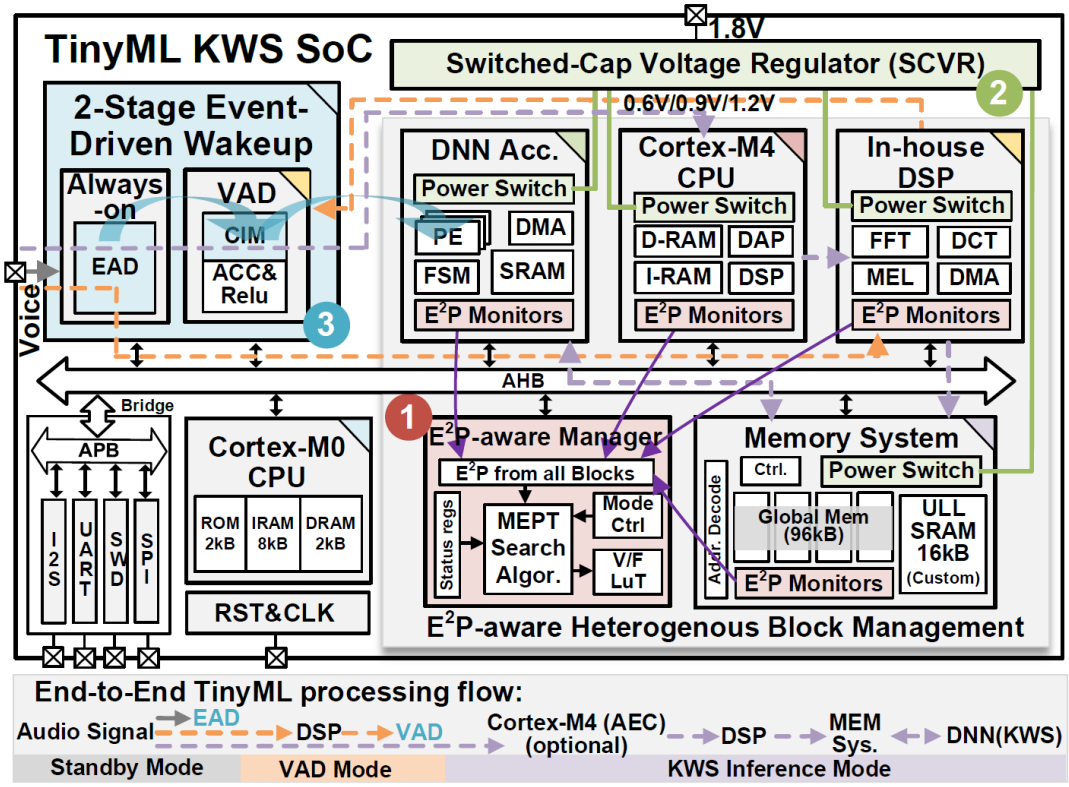
#28-1 Medusa는 28nm CMOS 공정에서 제작된 프로그래머블 8비트 프로세서로, 항상 켜져 있는 tinyML 작업에서 업계 최고의 추론 에너지를 달성합니다. Medusa는 6T 래치 기반의 Inner Loop Memories (ILMs)를 사용하여 15 fJ/Byte의 낮은 읽기 에너지를 제공하며, Pipelined Pixel Streaming (PPS) 아키텍처를 통해 시스템 레벨의 메모리 접근 에너지를 최대 9.5배 절감합니다. 주요 특징으로는 메모리 집약적인 병목층에서 에너지 소비를 줄이고, ILMs와 통합된 비트 직렬 곱셈기 및 덧셈 트리를 사용하여 연산 영역과 에너지를 최소화하며, 열 게이팅 및 파이프라인 우회 체계를 통해 불필요한 스위칭 에너지를 줄입니다. Medusa는 CIFAR-10에서 0.83/4.6 μJ/프레임과 86.2/91.6%의 정확도를 기록하며, 추론 에너지에서 3.4배/4.9배의 향상을 보여줍니다. Google Speech Commands에서는 0.23 μJ/프레임, Visual Wake Words에서는 5.0 μJ/프레임을 달성합니다. Medusa는 전압-주파수 스케일링을 지원하여 에너지와 지연 시간을 최적화할 수 있으며, 전체 메모리 접근 에너지는 시스템 총 에너지의 20%에 불과합니다.



[그림 5] Pipelined Pixel Streaming 아키텍처

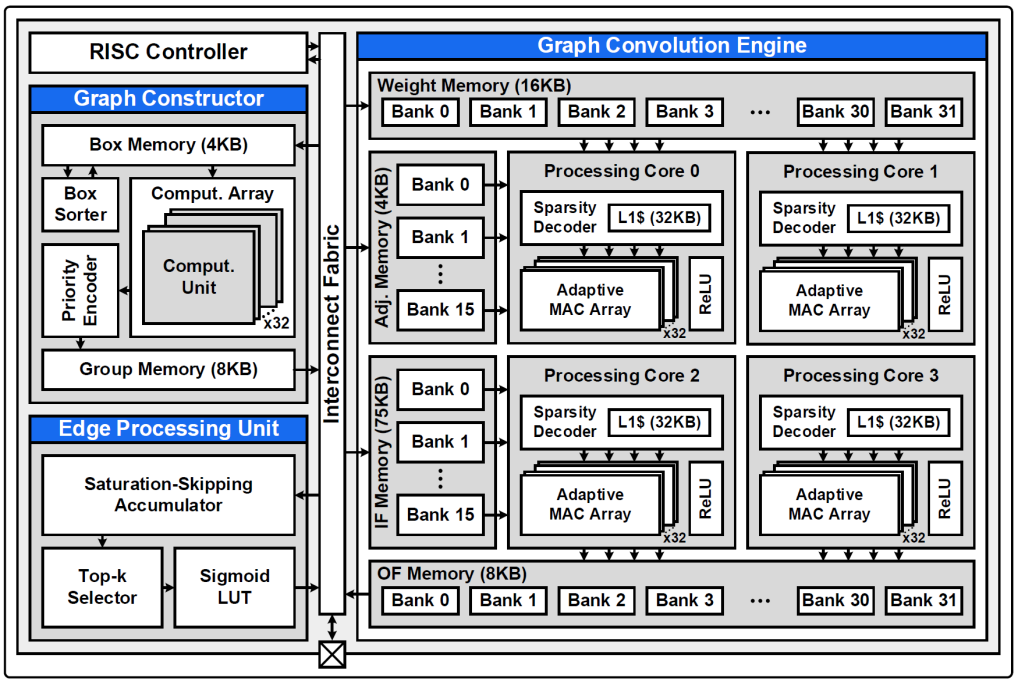
#28-2 이 논문은 Peking University에서 발표한 연구로, 에너지 이벤트 성능(E2P)-인식 관리 기능과 CIM 기반 2단계 이벤트 구동 웨이크업 방식을 갖춘 이종 TinyML SoC를 소개한다. 이 시스템은 리소스가 제한된 엣지 AI 기기에서 효율적으로 동작하도록 설계되었으며, 최소 전력 소비는 3.5μW로, 피크 대 대기 전력 비율은 최대 30,000배에 이른다. 특히, E2P 인식 시스템은 다양한 이종 블록의 런타임 상태를 실시간으로 모니터링하고, 계층적 전압 조정 메커니즘을 통해 시스템 수준에서 최소 에너지 지점(MEPsys)을 찾아냄으

로써, 에너지 효율을 28% 이상 향상시킨다. 또한, 이 연구에서는 CIM 기반 2단계 이벤트 구동 웨이크업 방식을 제안하여, 항상 켜져 있어야 하는 모드의 에너지를 87.3% 절감하는 데 성공했다. 이 SoC는 딥러닝 가속기와 같은 다양한 도메인 특화 가속기를 포함한 이종 SoC 구조를 채택하여, 엣지 AI 애플리케이션에서 초저전력 성능을 보여준다.



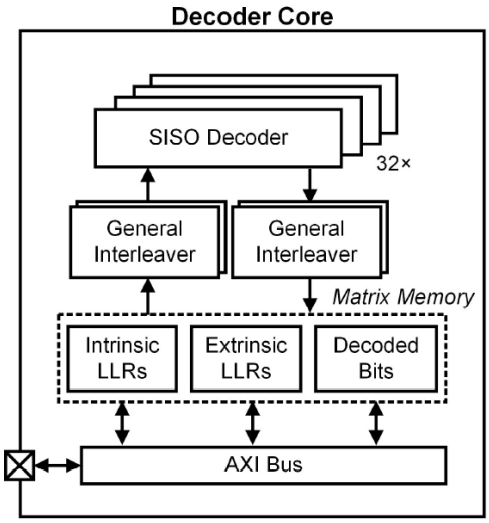
[그림 6] TinyML SoC 아키텍처

#28-3 이 논문은 National Taiwan University에서 발표된 연구로, 시각적 문맥 이해를 위한 최초의 전용 Scene Graph Generation (SGG) 프로세서를 소개한다. SGG는 객체 간 관계를 파악하여 고수준의 시각적 정보를 추출하는 작업으로, 기존 객체 감지보다 훨씬 복잡하다. 본 연구는 알고리즘-아키텍처 공동 최적화를 통해 계산 복잡성을 줄이고, 하이브리드 희소성 인코딩과 재구성 가능한 MAC 배열 구조를 활용해 메모리 사용량을 최소화하고 에너지 효율성을 높였다. 제안된 프로세서는 40nm CMOS 공정에서 제작되었으며, 280fps의 처리 속도와 0.36mJ/frame의 에너지 소비를 달성했다. 이는 GPU 대비 154배 높은 처리 속도와 1,800배 적은 전력 소모를 기록하며, 총 2.7×10^5배 더 높은 에너지 효율을 보여준다.



[그림 7] Scene Graph Generation 시스템 아키텍처

#28-4 이 논문은 University of Michigan에서 발표한 연구로, 고속 광통신을 위한 소프트 디시전 오픈 전방 오류 수정(oFEC) 디코더를 제안한다. oFEC 디코더는 타일 기반 아키텍처와 차이 기반 소프트 디시전 디코딩을 사용해 성능을 극대화하며, 32개의 파이프라인 SISO BCH 디코더를 통해 256비트 코드워드를 병렬로 처리한다. 또한, 매트릭스 메모리와 일반 인터리버(GIL)를 최적화해 메모리 영역을 83.6% 줄였고, 지연 시간을 33% 줄였다. 이 디코더는 인텔 16 공정에서 제작되었으며, 11.4mm²의 면적을 차지하고 40.2Gbps의 처리량과 17.43pJ/b/iteration의 에너지 효율을 보인다. 기존 FPGA 기반 oFEC 디코더와 비교해 10배 높은 처리량과 낮은 전력 소모를 달성해 차세대 광통신을 위한 솔루션으로 주목받고 있다.



[그림 8] Decoder Core아키텍처

저자정보

엄소연 박사과정 대학원생

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