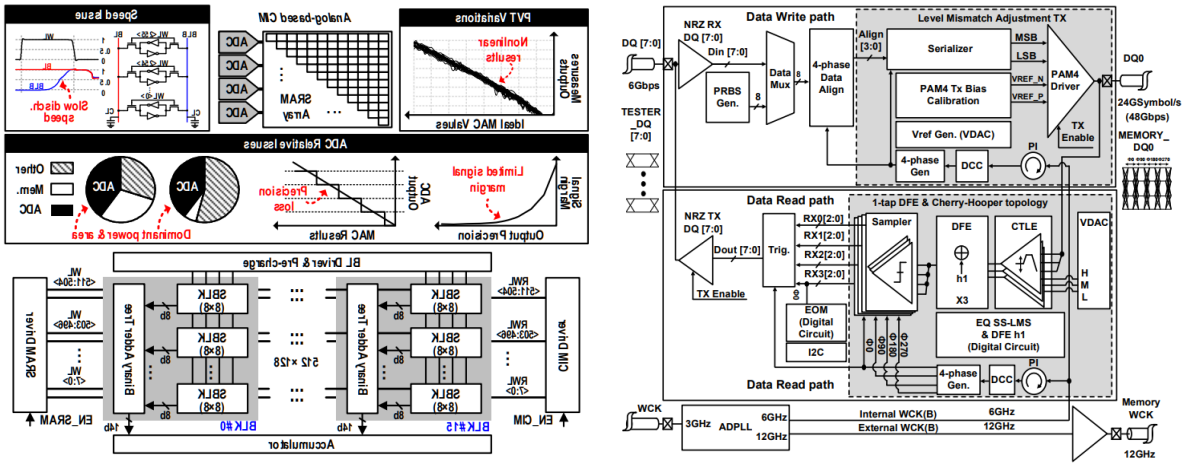
2023 IEEE ASSCC Review

서울대학교 전기정보공학부 박사과정 박현준

Session 5 Advanced Wireline Transceiver Techniques

Session 5는 고성능 메모리 인터페이스 및 메모리 테스팅 기술에 초점을 맞춘 5개의 혁신적인 논문이 발표되었으며, 각 논문은 메모리 기술의 다양한 측면을 탐구했다. 이 세션을 통해 메모리 기술의 최신 동향과 발전된 설계 및 테스팅 방법에 대한 통찰력을 얻을 수 있다.



[그림 1] (좌) 5-3 (우) 5-5

#5-1 이 논문은 고대역폭 메모리(HBM)를 위한 고속 저잡음 LDO를 제안한다. HBM3 표준을 지원하는 이 LDO는 빠른 시간(4ns)내에 정착되어, IO 타이밍 마진 감소로 인한 전력 공급 유발 지터(PSIJ)를 줄이는 데 효과적이다. 또한, 반전된 전압 추종자(FVF) 구조는 추가 전류 없이 출력 임피던스를 낮추고 대역폭을 늘려 IO 회로에서 발생하는 전력 소비와 내부 전력 공급 잡음을 줄일 수 있다. 성능 면에서, 제안된 LDO는 HBM3의 높은 데이터 전송률과 많은 IO 수를 지원하면서, 전력 소비를 줄이고 전력 공급 잡음을 효과적으로 관리할 수 있기 때문에 고성능 컴퓨팅 및 시스템에서의 전력 소비 및 전력 공급 잡음 관리에 큰 기여를 하였다.

#5-2 이 논문은 HBM을 위한 효율적인 아날로그 CIM 칩을 제안한다. 이 칩은 디지털 빔

포밍(DBF) 및 펄스 압축과 같은 디지털 신호 처리(DSP)에 사용되며, 8T-SRAM 비트 셀을 기반으로 HDR 데이터에 대한 양자화 손실 없이 병렬 입력 부동 소수점 곱셈을 지원하여 고해상도(HDR) 신호 처리 응용 프로그램에 적합하다. 또한, 입력 피크 감지 및 고차 가수 블록 인덱싱을 통해 적응형 전력 절약 기능을 구현하여 아날로그 연산에서의 전력 소비를 줄인다. 또한, 각 처리 요소에서 입력 데이터와 FMIM 가중치를 재사용하여 전송 에너지를 크게 줄일 수 있다. 제안된 FMIM은 메모리 내 곱셈 및 누적 후 5.8배의 에너지 효율과 95%의 지연 시간 감소를 달성하고, ARA는 FP32 펄스 압축 테스트에서 55%의 전력 절감을 제공한다. 이 연구는 이러한 특성들 덕분에 높은 에너지 효율성과 스루풋을 제공하여 복잡한 DSP 알고리즘을 처리하는데 적합하다.

#5-3 이 논문은 컨볼루션 신경망(CNN)의 가속화를 위한 SRAM 기반의 디지털 컴퓨트-인-메모리(CIM) 매크로를 제안한다. 이 매크로는 벡터-행렬 곱셈을 효율적으로 수행하며, 메모리와 처리 요소 사이의 데이터 전송을 최소화하여 전력 소비를 줄입니다. 또한, 1비트에서 8비트까지 완전히 재구성 가능한 가중치 정밀도를 지원하며, 감지 증폭기(SA) 없는 SRAM 읽기 작동으로 실리콘 면적을 절약한다. 이 매크로는 8비트 가중치와 1비트 입력에서 819.2 GOPS의 처리량을 달성하며, 249.1TOPS/W의 에너지 효율을 보여주는데, 현재 보고된 디지털 CIM 중에서 가장 높은 에너지 효율을 나타낸다. 이러한 특징들은 다양한 신경망 크기와 다중 계층을 필요로 하는 신경망 가속기에 적합하다.

#5-4 이 논문은 NAND 플래시 메모리의 워드 라인(WL) 누설 전류 문제에 대한 해결책을 제시한다. 고성능 컴퓨터와 인공 지능(AI)과 같은 고급 산업의 발전으로 대용량 데이터 처리의 중요성이 증가함에 따라, VNAND 플래시 메모리의 성능 및 용량 요구사항이 높아지고 있고 이 요구사항을 충족하기 위해 WL의 층 수가 증가하고 있지만, 인접한 WL 사이의 공간 감소와 채널 홀 식각 공정의 한계로 층수를 늘리기 어렵다. 이 연구는 주변 회로의 기생 누설 전류 성분을 제거하여 메인 셀 영역에서 WL 누설 전류를 독립적으로 탐지하는 새로운 방법을 제안한다. 제안된 회로는 WL 결함을 정확하게 탐지하고, 기존 방법에 비해 더 효율적으로 메모리 제품의 품질을 향상시키는데 이는 VNAND 플래시 메모리의 제조 과정에서 중요한 발전을 나타내며, 대용량 데이터 처리 및 관리에 필요한 높은 성능의 플래시 메모리 제품을 제공한다.

#5-5 이 논문은 GDDR6X와 같은 PAM4(Pulse Amplitude Modulation-4) 신호를 사용하는 메모리의 테스트를 위한 새로운 브리지 아키텍처를 제안합니다. 이 아키텍처는 기존 테스트 장비와 결합하여 확장된 기능과 높은 대역폭을 제공합니다. 특히, 메모리의 operation 초기 training을 통해 드라이버의 PMOS 및 NMOS 너비와 게이트 입력 전압을 조정하여 replica 드라이버가 메인 드라이버의 목표 입력 레벨에 도달하도록 PAM4

드라이버의 출력 레벨을 조정하여 레벨 불일치(Ratio of Level Mismatch, RLM)를 개선한다. 48 Gbps 속도로 작동하며, 읽기 및 쓰기 모드에서 각각 1.85 pJ/bit와 2.97 pJ/bit의 전력을 소비하며, 16 Gbps에서 0.73에서 0.98로 개선된 RLM을 보여준다. 또한, 내부 EOM 회로에서 얻은 데이터 아이 다이어그램과 48 Gbps에서 개방된 아이를 통해 PAM4 RX의 성능을 입증한다.

저자정보

박현준 박사과정 대학원생

- 소속 : 서울대학교
- 연구분야 : HBM, Chord Signaling, Information Theory
- 이메일 : spp098@snu.ac.kr
- 홈페이지 : https://sites.google.com/view/wschoi?pli=1

2023 IEEE ASSCC Review

한양대학교 신소재공학과 박사과정 송충석

Topic : Memory

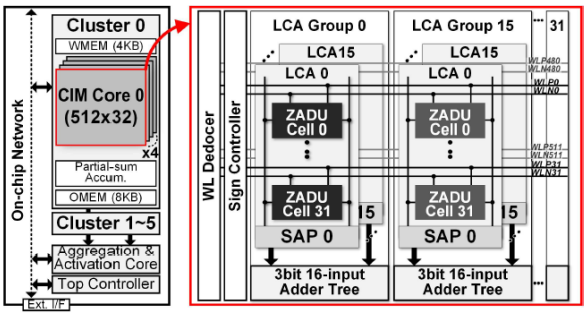
Session 16. Intelligent Memory and Logic Circuit Techniques

이번 2023 IEEE ASSCC의 Session 16은 Intelligent Memory and Logic Circuit Techniques라는 주제로 총 4편의 논문이 발표되었다. 머신러닝의 여파로 연산량이 기하급수적으로 늘어남에 따라 메모리에 요청하는 데이터의 양도 늘어나고 있다. 데이터의 메모리-프로세서간 물리적 이동에서 생기는 병목현상과 에너지소모는 현재 무시할 수 없는 수준으로 그에 따라 메모리에서의 추가적인 동작들이 요구되는 상황이다. 본 세션에서는 메모리 내에서 연산을 수행할 수 있는 computing in memory(CIM) 기술을 총 3편(16-1, 16-2, 16-3) 발표했고, 전압/주파수 스케일링을 통한 효율적인 메모리를 구현하는 기술을 1편(16-4) 발표했다. 본 리뷰에서는 16-1, 16-2, 16-4를 다뤄보고자 한다.

#16-1 논문은 메모리 내에서 로그 스케일 양자화(logarithmic quantization, LOGQ)를 연산하는 CIM 프로세서이다. 같은 크기로 양자화를 진행하는 INT 양자화에 비해 값에 따라 양자화 간격이 다른 로그 양자화는 bit serial이 단 1개의 '1'값만 가진다는 것과 곱셈연산을 bit shift 연산으로 대체할 수 있어 cost를 줄일 수 있다는 장점이 있다. 본 논문에서는 3가지 특징이 있는데: (i) MAC연산시 여러 값 들을 align후에 accumulation을 해야 하는데, 이 작업을 최소한의 연산 cycle로 구현함으로써 효율적인 동작을 할 수 있는 유닛을 만들었고, (ii) WL이 분리된 6T-SRAM을 사용하여 동시에 두개의 cell을 활성화시켜 BL과 BLB를 모두 읽을 수 있도록 하여 전력소모를 감소시켰고(BL/BLB의 pre charging 재사용), (iii) 로그 양자화 값을 연산하기 위해 bit shift 후 덧셈을 하는 주변부 회로 로직을 만들었다.

첫 번째 특징으로 연산 cycle을 37퍼센트 감소시켰고, 에너지 효율은 51.5% 증가시켰다. 두 번째 특징 덕분에 90% sparsity에서 60.7% 전력을 감소시켰다. 세 번째 곱셈 유닛 대신 bit shift 유닛의 사용으로 에너지효율성을 최대 3.17배 증가시켰다. 본 CIM 프로세서는 28nm CMOS 공정으로 48KB의 온 칩 CIM을 탑재하였다. 로그양자화를 적용시켜 116.4 TOPS/W(Trillion Operation Per Second Per Watt)의 성능을 이루었다. 로직을 많이 둘

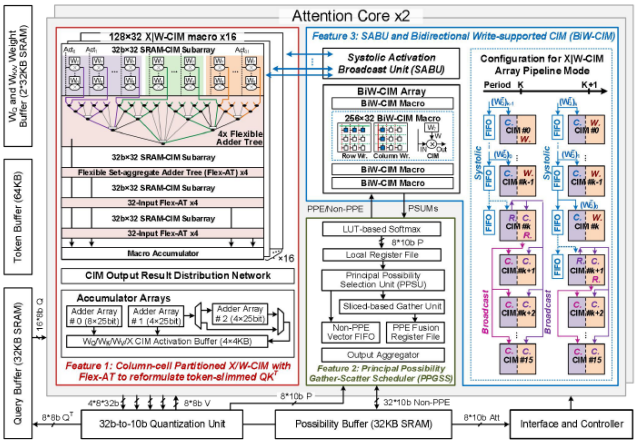
수 없는 CIM의 특성 상 곱셈기는 큰 오버헤드가 될 수밖에 없는데 이를 로그 양자화를 적용하여 곱셈기를 없앴다는 것이 눈에 띄는 점이다.



[그림 4] #16.1에서 제안한 LOG-CIM 전체 구조

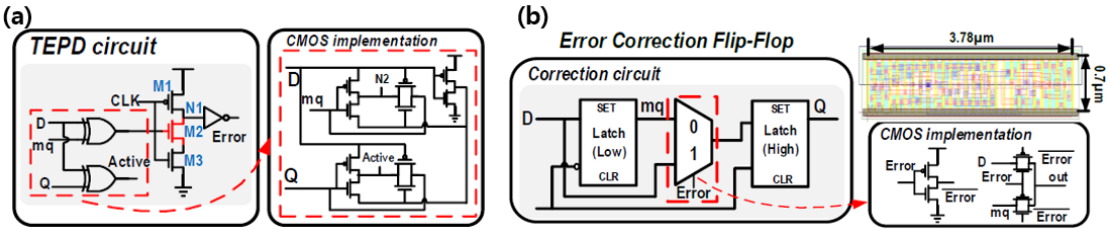
#16-2 논문은 CIM에 트랜스포머 알고리즘을 시스톨릭을 이용하여 구현한 CIM 프로세서이다. 인풋 토큰으로부터 Q(query), K(key), V(value) 행렬을 생성하고 이를 이용해 attention score를 구하는 트랜스포머 구조는 CIM에서 구현하기 어려운 부분이 있다. K 행렬을 이용해 P(=$Q \cdot K^T$)를 구할 때, CIM macro에 access를 자주 해야 한다는 점과, 인풋을 bit-serial로 넣어주는 CIM 매크로의 특성상 attention value의 값은 너무 많은 '0'을 포함해 '0'을 연산하는 의미 없는 연산 cycle이 생기게 되고, 마지막으로 CIM 매크로의 SRAM 메모리에 모든 가중치를 저장할 수 없다는 것이 문제다. 본 논문은 위의 문제점을 언급하면서 다음 세가지 특징을 가지는 CIM 프로세서를 만들었는데 : (i) $Q \cdot K^T$의 연산을 재구성하여 과도하게 K 행렬의 값을 반복적으로 불러오는 것을 줄였고, (ii) 데이터의 Bit 시퀀스가 계속 '0'이 나오지 않는 길이인 effective bit-width가 작은 데이터에 대해서 연산 latency를 줄였고, (iii) 마지막으로 시스톨릭 CIM을 지원하여 양방향 행렬 곱셈이 가능케 하였다.

본 CIM 프로세서는 28nm CMOS 공정으로 6.98mm^2의 면적으로 제작되었다. 동작 전압 0.6V ~ 1.0V에서 80MHz~240MHz 까지 동작하고, INT8연산으로 최고 2663.4GOPS, 38.9 TOPS/W의 성능을 이루었다. 트랜스포머 알고리즘은 최종 attention score까지 얻기 위해 많고 다양한 데이터를 불러오고, 연산을 해야 하는데 그 데이터의 흐름을 효율적으로 잘 처리한 것으로 보여 눈에 띄는 논문이다.



[그림 5] #16.2에서 제안한 CIMFormer 전체 구조

#16-4 논문은 메모리 사용량에 따라 전압과 주파수를 조절해서(AVFS) 전력효율성을 증가시키는 방법을 발표했다. 메모리가 활성화(activate) 되어있는지를 판단함과 동시에 timing 에러를 탐지하는 TEPD와(즉, 활성화가 되어있음과 동시에 에러가 없어야 정상적으로 동작한다고 판단) 그 오차를 수정하는 ECFF, 그리고 오차발생에 따라 전압 및 주파수를 조절하는 과정을 통해 메모리 전력효율을 높였다. TEPD는 15개의 트랜지스터를 이용해 구성했고 다른 SOTA 논문과 비교해서 상대적으로 낮은 전력소모를 달성했다(0.185nW). ECFF는 DFF(D-flip flop)와 multiplexer를 이용해 구성하였다. 본 논문에서 발표한 회로는 0.54~0.9V 범위에서 10~525MHz까지 구현되었으며 일반적인 timing margin을 둔 baseline 회로보다 37~52퍼센트의 전력소모 감소를 달성했다. 메모리 내에서 오차를 감지하고 수정하는 것과 더불어 그 오차에 따라 동작전압과 주파수를 스케일링 하려는 의도는 앞으로도 다방면으로 활용될 가능성이 높아 보인다.



[그림 6] #16.4에서 제안한 TEPD와 ECFF 구조

저자정보

송충석 박사과정 대학원생

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