An 89TOPS/W and 16.3TOPS/mm² All-Digital SRAM-Based Full-Precision Compute-In Memory Macro in 22nm for Machine-Learning Edge Applications

Yu-Der Chih 외 19명

목차

I. Introduction

1. 연구 배경 및 필요성

❖ 차세대 엣지 디바이스는 에너지가 제한적이므로, 메모리 접근 병목을 가진 기존 폰노이만 구조보다 CIM 구조가 더 효율적인 연산 가능

2. 기존 연구의 한계점

❖ 기존 CIM 연구는 주로 아날로그 기반으로 높은 에너지 효율을 가지지만, 낮은 SNR로 인해 정밀도가 제한되어 AI 연산 정확도에 한계가 있음

3. 본 연구 제안 구조

❖ 22nm 공정 기반, 대규모 데이터를 처리하면서도 높은 TOPS/W, TOPS/mm^2의 성능을 갖는 DCIM macro 구조 제안

❖ 기존 연구의 한계를 극복하기 위해 높은 정밀도(출력 비트폭)와 다양한 CNN 신경망 구조(topologies)을 지원하는 구조 제안

II. DCIM macro Architecture

1. 구조

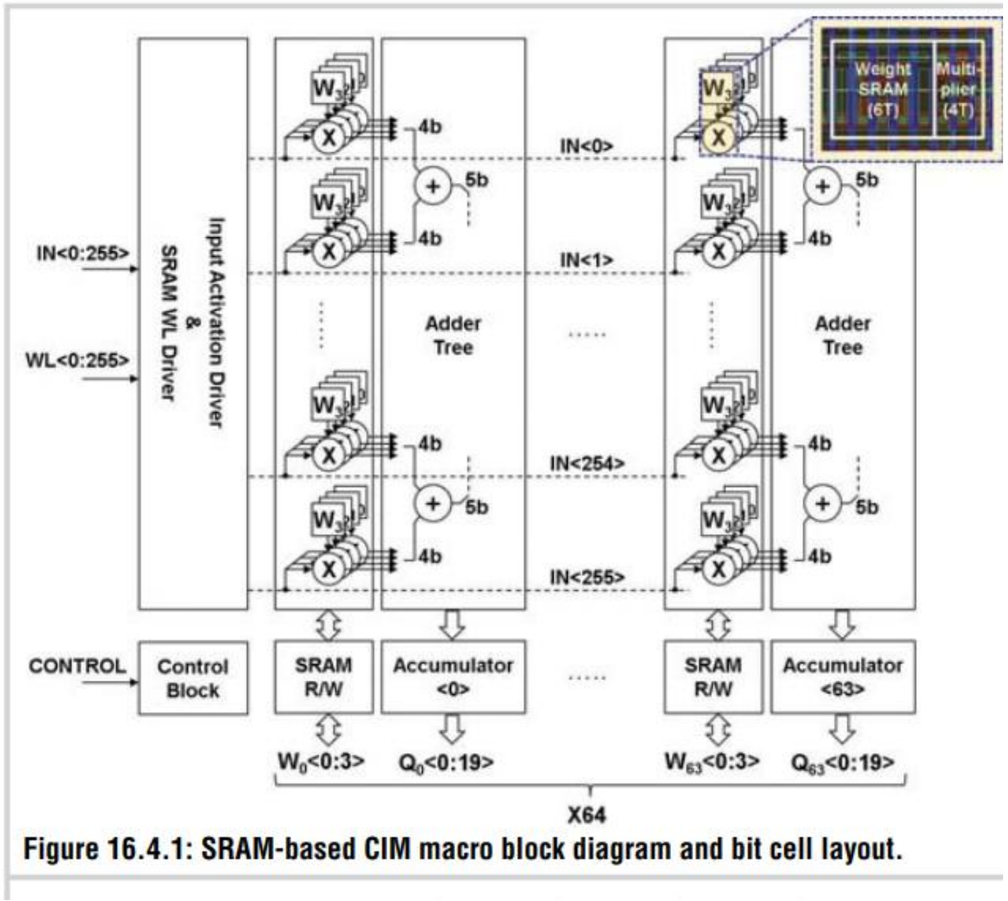


Figure 16.4.1: SRAM-based CIM macro block diagram and bit cell layout.

전체 구조

- ❖ DCIM macro는 256개 data를 제어하는 input activations, 64개의 accumulator(output), 256x64개의 weight로 구성되며, 하나의 macro는 64개의 sub-CIM unit으로 구성됨
- ❖ 해당 구조는 64x256 weight행렬과 256x1 입력 열벡터의 행렬 곱을 하드웨어로 표현
- ❖ Input activation bit-width : 1~8bit
- ❖ weight bit-width : 4/8/12/16 bit

sub-CIM unit 구조

- ❖ 256개의 4비트 가중치 저장 6T SRAM이 총 (1,024개=1kb)
- ❖ 256개의 입력과 가중치 간 bit-wise 곱셈을 위한 multiplier(nor) 1,024개
- ❖ 각 곱셈 결과를 병렬로 합산하여 partial sum 만들어내는 adder tree
- ❖ Partial sum의 MSB에 따라 비트 확장해주고 Cycle별 누적을 위한 accumulator
- ❖ SRAM을 4b weights를 각각 R/W하기위한 decoder

✓ 입력은 MSB부터 bit-serial 방식으로 순차적으로 인가되며, 이전 데이터와 변경된 비트만 switching하며 변경되지 않은 데이터는 reuse하여 에너지를 절약함

II. DCIM macro Architecture

2. 동작

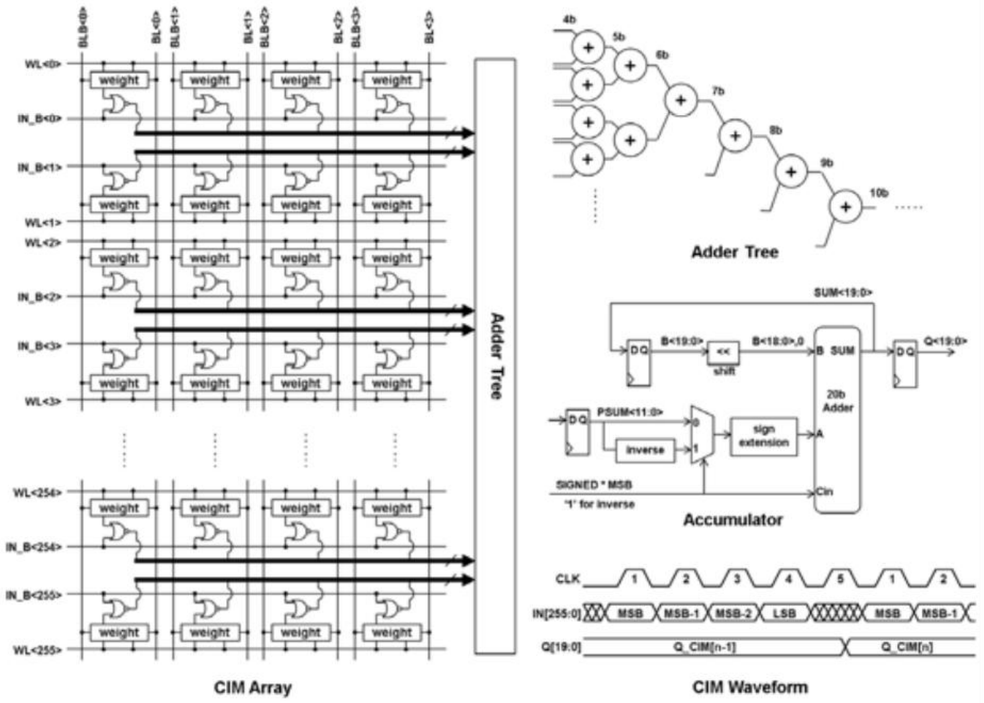


Figure 16.4.2: Schematic view of one sub-CIM unit and CIM operational timing waveforms

동작 (unsigned 4bit input, 4bit weight 기준)

1) 입력은 MSB부터 LSB까지 4개의 클럭에 걸쳐 **bit-serial 방식**으로 인가됨.
2) 첫 클럭에서 256개의 MSB가 64개의 sub-CIM unit에 동시에 입력됨.
3) 각 sub-CIM unit의 CIM Array 256개의 4bit weight와 MSB는 병렬로 곱셈을 수행함
4) 256개의 곱셈 결과는 병렬 adder tree를 통해 12비트 **partial sum**으로 합산함
5) accumulator 에서 signed partial sum일 경우엔 보수 및 증가 연산이 추가 적용됨
6) MSB sign에 따라 sign extension을 통해 상위 bit들이 0또는 1로 채워짐(16bit)
7) 이 partial sum은 4 클럭 동안 **pipeline 방식(shift and add)**으로 누산되고, 이후 추가 1 클럭에 걸쳐 최종적으로 누산됨.
8) 최종 결과값은 16비트 정밀도의 누산값으로 Q_j에 저장됨.
9) 1~8의 동작을 하며 $Q_j = 2^3 \times \Sigma(A3i \times Wji) + 2^2 \times \Sigma(A2i \times Wji) + 2^1 \times \Sigma(A1i \times Wji) + 2^0 \times \Sigma(A0i \times Wji)$을 표현함

✓ (Ani: i번째 input의 n번째 비트, Wji: j번째 출력을 위한 i번째 가중치)

III. Implementation

1. 에너지 효율성과 처리량을 향상시키기 위해 DVS 및 구성 최적화

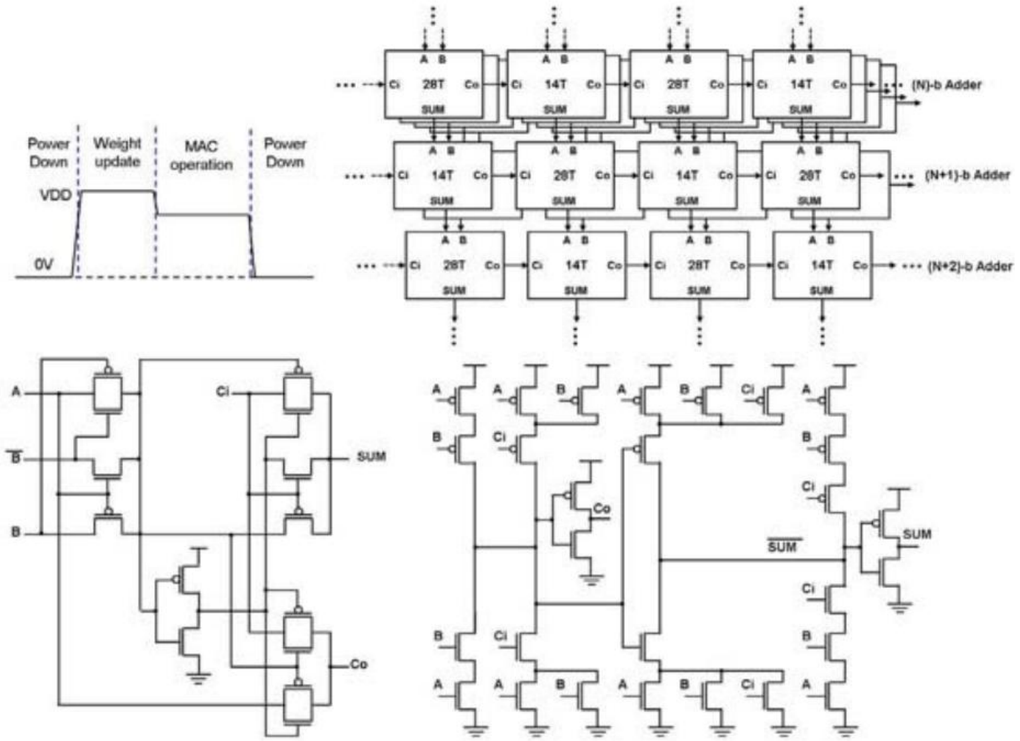


Figure 16.4.3: (Top) DVS for energy saving in MAC operation, and the proposed adder tree structure with two kinds of full adders for dynamic power saving. (Bottom) 14-T and 28-T full-adder.

Mode에 따른 DVS(Dynamic Voltage Scaling) 적용

- ❖ weight 업데이트 시에 높은 신뢰성을 위한 0.8V 사용함
  - ➢ 25℃기준, 4bit weight 256x64개를 update시 0.4μs 지연이 발생
- ❖ MAC 연산시에 이미 저장된 weight를 읽기만 하므로 0.68V에서도 동작 가능
  - ➢ 97TOPS/W(4b weight/4b input activations)성능을 가짐

덧셈 트리 전가산기 구성 최적화

- ❖ 14T FA : 빠르고, 전력소모가 적으나, 노이즈에 약함
- ❖ 28T FA : 안정적이지만 전력, 면적 소모가 큼
  - ➢ interleaving 배치하여 성능,전력,면적의 균형을 맞춰 TOPS/W 30%향상

Input activation, weights의 bits 구성

- ❖ Input activations의 toggle rate 낮을수록 전력 감소 : $P = C \times V^2 \times f \times \alpha$(toggle rate), capacitor 충/방전 횟수에 따라 에너지 소모량 증감
- ❖ weight sparsity가 높을수록 에너지 소모감소 : 곱셈결과가 0으로, 내부적으로 switching 감소

III. Implementation

2. 다양한 신경망 구조 확장성 및 Programmability

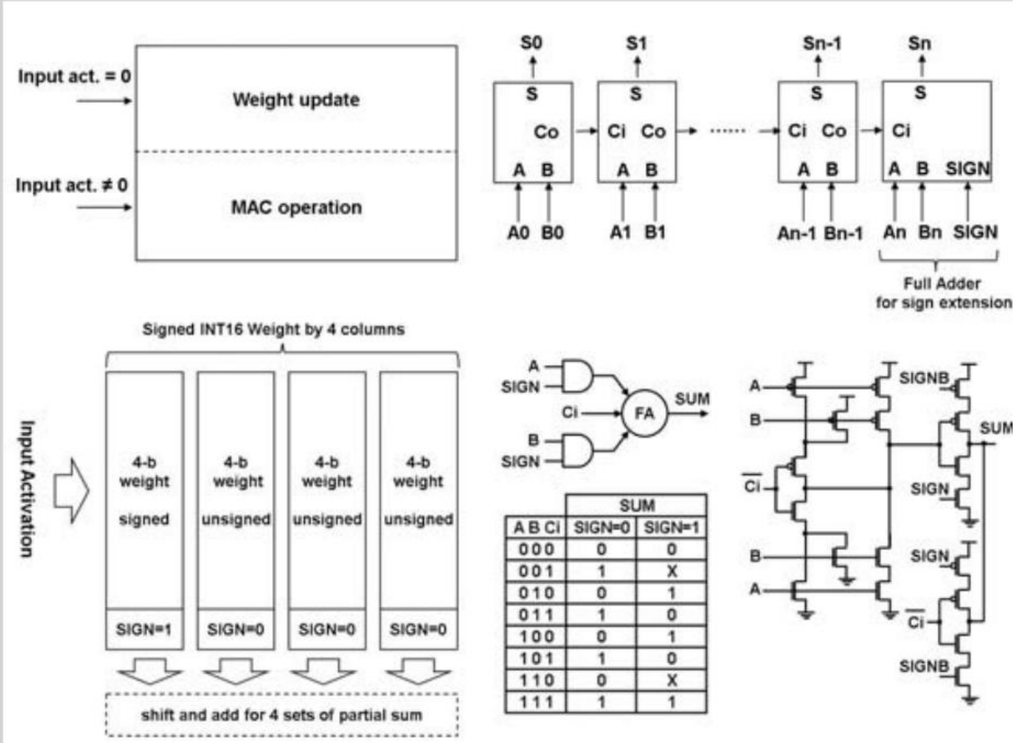


Figure 16.4.4: (Left) Concurrent weight update function and MAC operation with programmable signed/unsigned weights. (Right) Schematic and truth table for the extra full-adder used for sign extension.

구조 확장성

❖ CIM macro를 병렬, 직렬 또는 2D 배열 형태로 구성 가능

- ➢ 다양한 신경망 구조 지원 가능
- ➢ 3개의 CIM macro를 cascaded(직렬)로 연결 : 3x3 필터에 64채널 convolution 연산 가능

Programmability

❖ Input과 weight는 각각 signed/unsigned 설정 가능

❖ 4/8/12/16bit weight precision 지원

- ➢ 신경망 구조에 따라 유연한 데이터 표현 가능

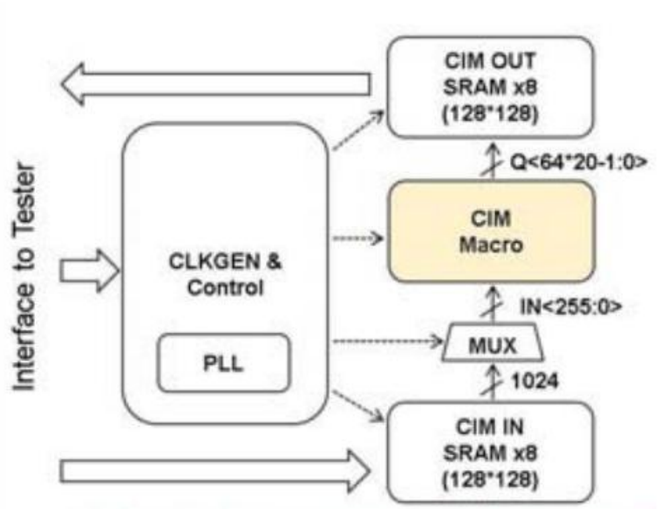
❖ MAC 연산을 하면서 input activations이 0인 weights값을 동시에 업데이트 가능

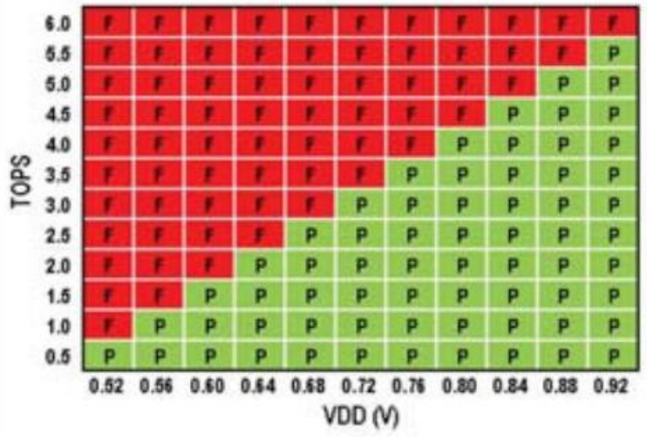
Sign extension

❖ Unsigned weight 인가 시, additional FA for sign extension은 비활성화되어 에너지 절약 가능.

IV. Results & Discussion

1. Results(1/2)





슈무(shmoo) plot 분석

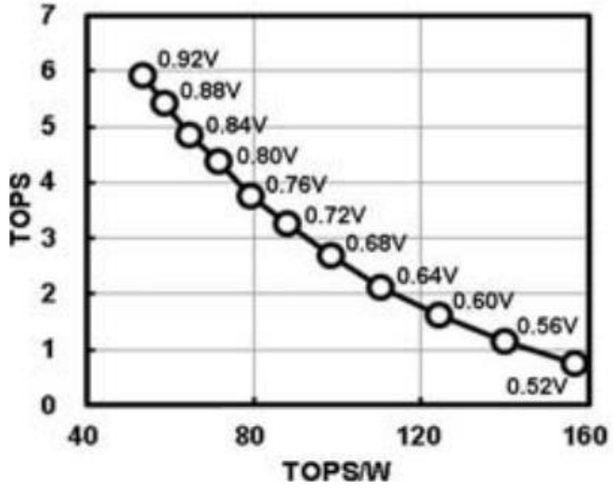
❖ DVS를 적용하여 다양한 VDD환경에서 동작 가능성을 평가

➢ AI HardWare에서 필요한 처리량에 따라 최적의 전압 (효율점)을 찾을 수 있음

DVS 분석

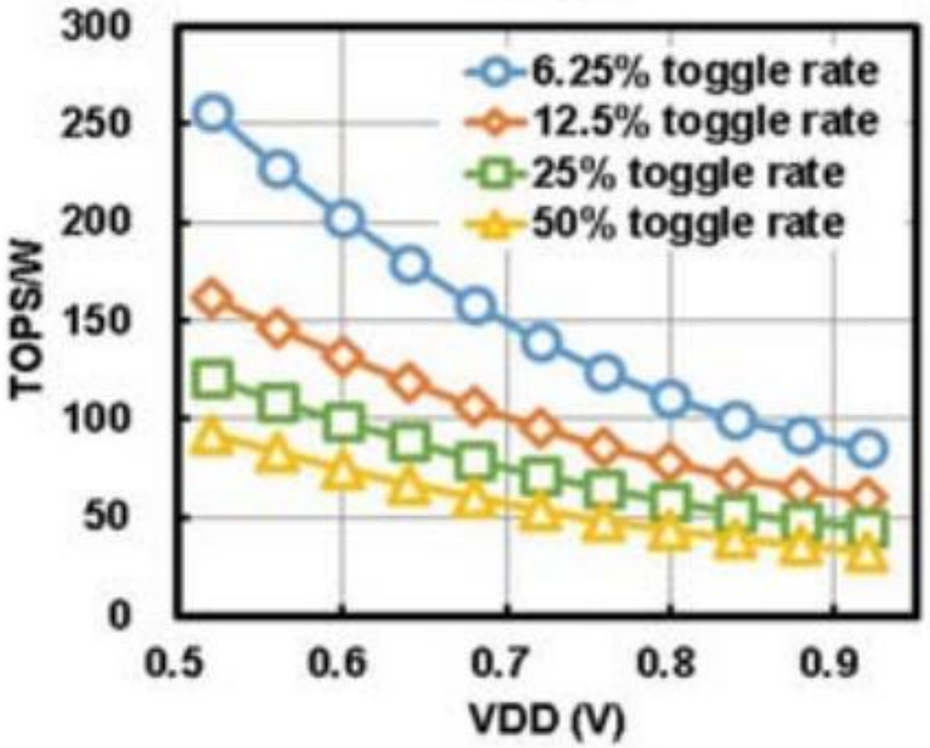
❖ 256x64 CIM array이 @25 ℃ , 0.72v 동작시 3.3TOPS 처리량 달성

➢ 전체 array가 동시에 작동했을 때 최대의 효율점을 시사



IV. Results & Discussion

1. Results (2/2)



Toggle rate measurement 분석

❖ Sparse pattern

- ➢ 입력 toggle rate 18%, weight sparsity 50% 조건
- ➢ 0.72V에서 89TOPS/W 성능 달성

❖ Dense pattern

- ➢ 입력 toggle rate 50%, weight sparsity 50% 조건
- ➢ 0.72V에서 52TOPS/W 성능 달성

✓ 입출력 타이밍(데이터 흐름 최적화) 맞추기 위해, 입력 활성화를 저장하고 출력 partial sum을 수신하기 위한 두 개의 SRAM 매크로가 추가로 구성됨

IV. Results & Discussion

2. Discussion

	ISSCC'18 [1]	ISSCC'18 [2]	ISSCC'19 [3]	ESSCIRC'19 [4]	ISSCC'20 [5]	ISSCC'20 [6]	This work
Technology	65nm	65nm	55nm	65nm	7nm	28nm	22nm
MAC operation	Analog	Analog	Analog	Digital	Analog	Analog	Digital
Array Size	4Kb	16Kb	3.8Kb	16Kb	4Kb	64Kb	64Kb
Cell Type	S6T	10T	T8T	6T	8T	6T	6T
Push rule	Yes	No	Yes	NA	Yes	NA	No
Macro size (mm^2)	NA	0.067	NA	0.2272	0.0032	NA	0.202
Bitcell Area (um^2)	0.525	NA	0.865	NA	0.053	0.25	0.379
Power Supply(V)	1&0.8	1.2&0.9	1	0.6~0.8	0.8	0.7~0.9	0.72
Inputs Bits	1	7	4	1~16	4	4~8	1~8
Weight bits	1	1	5	4/8/12/16	4	4/8	4/8/12/16
Output Bits	1	7	7	8~23	4	12 (4b/4b) 20 (8b/8b)	16 (4b/4b) 24 (8b/8b)
Cycle time (ns)	2.3	150	10.2	NA	5.5	4.1 (4b/4b) 8.4 (8b/8b)	10 (4b/4b) 18* (8b/8b)
Throughput (GOPS)	1780	10.67	17.6	567 (1b/1b)	372.4 (4b/4b)	124.88 (4b/4b) 30.48 (8b/8b)	3300 (4b/4b) 917* (8b/8b)
Energy Efficiency (TOPS/W)	55.6	28.1	18.4	117.3 (1b/1b)	262.3~610.5 (4b/4b)	68.44 (4b/4b) 16.63 (8b/8b)	89 (4b/4b) 24.7* (8b/8b)

*estimation

Figure 16.4.6: Key metric comparison table.

DCIM 구조의 장점

- ❖ Analog CIM방식에 비해 높은 bit precision을 가짐
  - ➢ DCIM은 4bit weight x 4bit or 8bit input을 사용함
  - ➢ 누산 결과를 정확도 손실 없이 표현하기 위해 출력 비트폭을 16 or 20bit으로 충분히 크게 설계함
  - ➢ 예시) 단일 곱셈 : 15x15=225 (8bit) 표현 가능
  - ➢ 예시) 누산 : 225x4=900(10bit) 이상 필요
  - ➢ 반면 ACIM의 경우 출력 비트폭이 최대 8bit로 제한되어 있음
  - ➢ 위와 같은 누산 결과를 출력 비트폭 부족으로 인해 정확도 손실 발생
- ❖ 입력과 출력의 동일한 data frame(NCHW)사용하여 데이터 형식 변환 최소화하고, input/weight를 재사용하는 구조로 에너지 및 지연시간에서 장점을 가짐
- ❖ 가변 비트폭 데이터를 해당 구조를 통해 연산을 수행할 수 있음
- ❖ 연산 효율과 면적 효율을 동시에 반영한 지표인 FOM 1450을 달성할 수 있음
  - ➢ 기존 연구의 digital 기반 CIM 중 성능이 제일 높음
  - ➢ 5nm 공정 기준으로 평가 결과, TOPS/W 2.8배, TOPS/mm^2 는 19배 향상됨
  - ➢ transistor length가 작아지면서 면적을 줄이면서 성능도 높아짐

V. Conclusion

- ❖ 22nm 공정 기반, 256 x 64 Digital CIM 매크로를 구현하여 89TOPS/W, 16.3 TOPS/mm² 성능 달성
- ❖ All Digital MAC 구조 기반으로, 입력/가중치에 대해 signed/unsigned 1~8bit bit-width 선택 가능하며 정확도 손실 없이 다양한 신경망 지원함
- ❖ DVS와 동시 weight 업데이트 기능을 갖춘 구조로 에너지 효율성과 처리량을 향상 시킴
- ❖ Input activation/weight reuse 와 input과 output의 동일한 data frame(NCHW)을 통해 메모리 접근 및 데이터 흐름에 대한 에너지와 지연을 크게 줄여줌

감사합니다!