A 5-nm 254-TOPS/W 221-TOPS/mm2 Fully-Digital Computing- in-Memory Macro Supporting Wide-Range Dynamic-Voltage-Frequency Scaling and Simultaneous MAC and Write Operations

Hidehiro Fujiwara외 17명

목차

I. Introduction

1. 연구 배경 및 필요성

❖ 엣지 AI 디바이스에서는 MAC 연산과 Data 이동이 전체 전력 소모의 주요 원인

❖ DCIM은 공정 스케일링과 정확도 손실 없이 높은 유연성으로 다양한 연산 대응 확인

2. 기존 연구의 한계점

❖ Analog CIM(ACIM)은 전력 효율이 높지만, 정확도 저하 문제 (ADC, 트랜지스터 공정 변동성 등)가 존재

3. 본 연구 제안 구조 개요

❖ 5nm 공정으로 12T 1R1W bit-cell 기반의 저전력 동작 64kb DCIM macro 제안

- ➢ MAC과 weight 쓰기 연산을 동시에 수행 가능
- ➢ DVFS(Dynamic Voltage-Frequency Scaling) 지원

II. DCIM macro Architecture

1. 전체 구조 구성 및 동작(1/2)

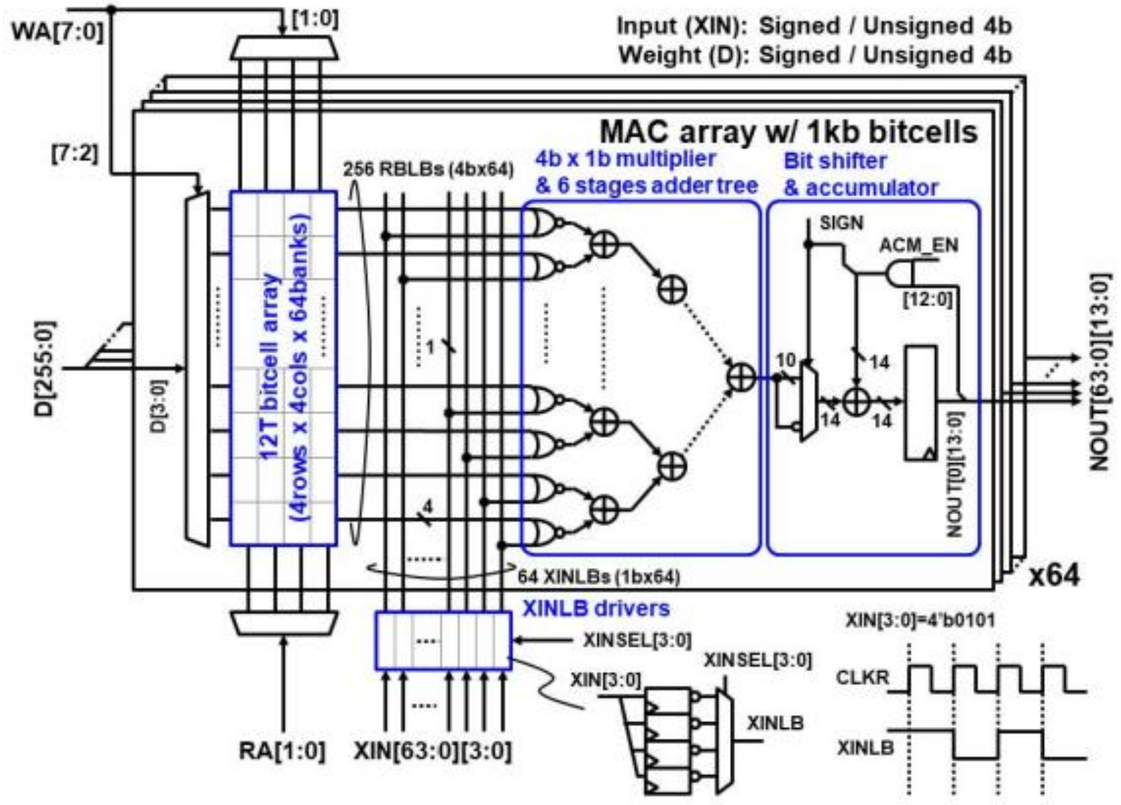


Figure 11.6.2: Overall architecture of DCIM macro.

구조

❖ 총 64개의 MAC Array, signed/unsigned weight저장을 위한 12T bit-cell, XINLB driver, Adress Decoder로 구성

❖ 각 MAC Array 구성 :

- 12T bit-cell Array
  -> 1kb = 4cols x 4rows x 64banks로 구성
  -> 1R1W 구조로 MAC연산과 weight쓰기를 동시에 지원
- XINLB driver : 64개의 4b input을 4 to 1 MUX를 통해 bit-serial 방식(MSB→LSB)으로 순차 인가
- Adress decoder : Read/Write Word line을 제어해 bit-cell에서 데이터를 읽거나 씀
- NOR multiplier : bit-cell에서 읽어온(RBL) weight와 XIN을 곱셈 연산 수행
- Adder tree : 각 banks의 곱셈(XINLB x RBLB=4bit) 결과 64개를 병렬로 합산하여 partial sum 생성
- Bit shifter : 누적 값의 bit significance를 보정하기 위해 left shift 수행
- Accumulator : shift된 값을 누산하여 최종 결과를 FF을 통해 NOUT 출력

※ 해당 DCIM 구조는 읽기 동작을 지원하며, 읽은 데이터는 바로 MAC연산을 통해 최종 출력됨

II. DCIM macro Architecture

1. 전체 구조 구성 및 동작(2/2)

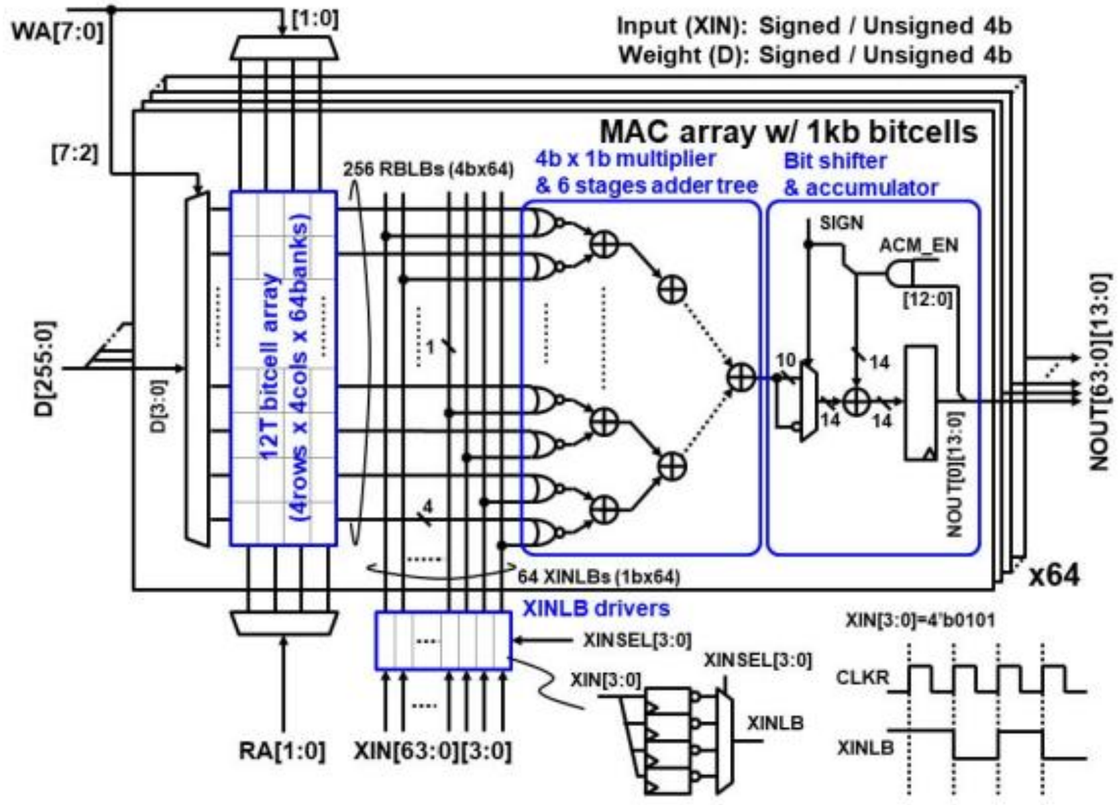


Figure 11.6.2: Overall architecture of DCIM macro.

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

1) 한 사이클마다 1bit씩 총 4사이클 동안 bit-serial방식으로 인가됨(MSB->LSB)
2) 첫 사이클에 64개의 XIN의 MSB가 각각의 64개의 MAC Array에 병렬로 인가됨
3) 각 MAC array의 bit-cell의 64개 banks에서 읽은 각각의 4bit weight와 입력된 1bit XIN을 bitwise NOR연산하여 64개의 4bit 곱셈 결과 생성
4) 4bit 곱셈 결과가 6-stg adder tree에 병렬로 합산하여 partial sum 생성
5) Partial sum의 sign에 따라 2의 보수 연산을 선택 수행함
6) 보수 연산을 통해 14bit partial sum이 누산됨
7) 다음 사이클의 14bit partial sum과 누산하기 위해, 이전 사이클의 14bit 결과는 bit shifter를 통해 left shift된 후 누산됨
8) 이 과정을 5사이클(MSB->LSB) 동안 반복
9) 연산 종료 후, 최종 결과가 FF에 저장되어 64개의 14bit NOUT 출력

II. DCIM macro Architecture

2. MAC 연산 타이밍 및 특징

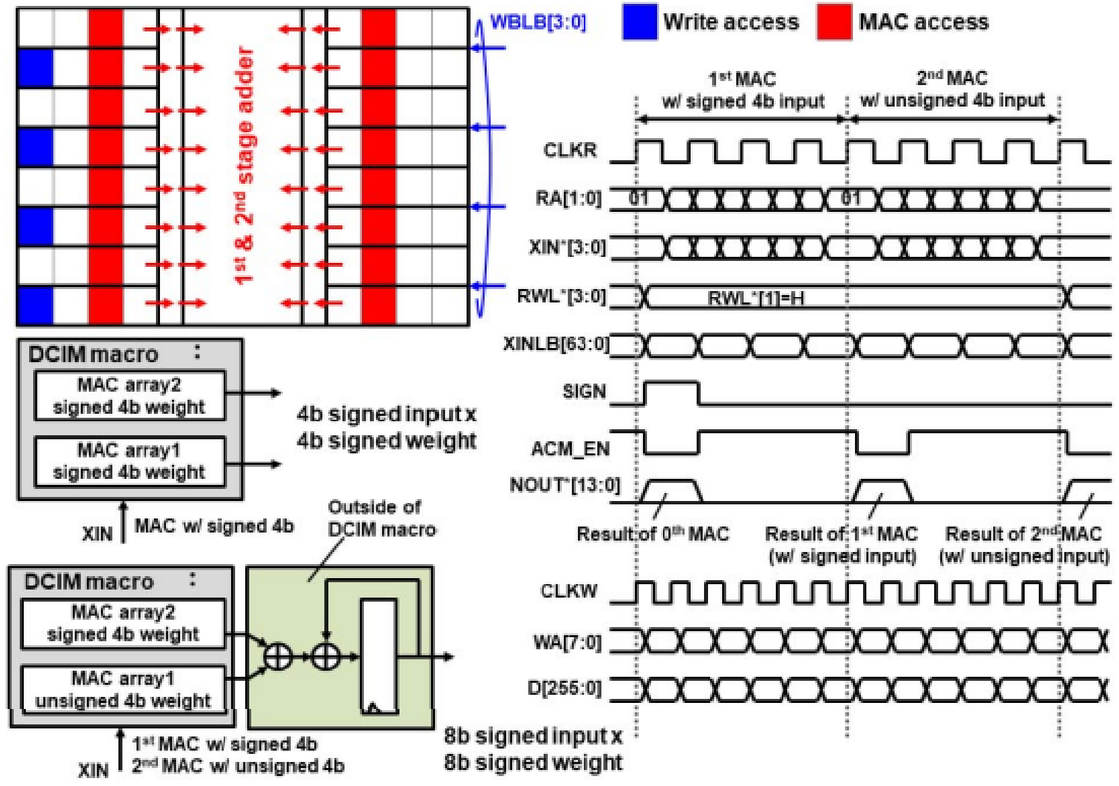


Figure 11.6.4: Timing diagram with simultaneous MAC and write accesses.

TIMING

- ❖ MAC연산과 weight 업데이트를 동시에 빠르게 수행가능
  - ➢ 읽기/쓰기 clk을 각각 사용하는 비동기 1R1W 12T bit-cell 사용
  - ➢ MAC 접근 위치와 쓰기 주소가 다를 경우에 업데이트 가능
  - ➢ 쓰기 clk을 짧게하여 빠르게 쓰기 동작 수행 가능

구조 특징

- ❖ 비트셀 배열의 각 4개 뱅크는 덧셈 트리의 1단계 및 2단계와 함께 결합되어 있음
  - ➢ Layout 설계 시, 면적을 최소화 할 수 있음
- ❖ 해당 구조 DCIM macro 외부에 추가적인 제어 로직을 추가하여 input 및 weight 비트폭의 확장 지원
  - ➢ Signed/unsigned 형식 지원
  - ➢ 8bit signed weight 저장시, array2에 signed 4bit, array1에 unsigned 4bit 저장
  - ➢ 8bit input/weight는 두 번의 MAC연산으로 처리 가능
  - ➢ 2’s complement logic을 통해 signed input을 연산할 수 있음

III. Implementation

DCIM macro 회로 및 레이아웃 최적화

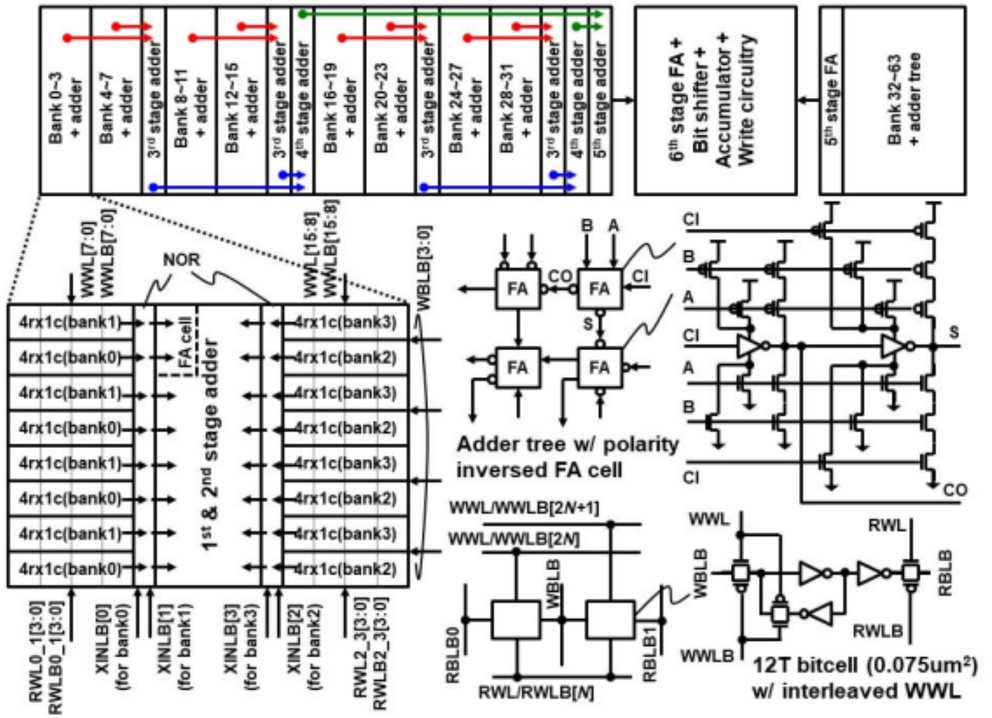


Figure 11.6.3: MAC array floor plan with polarity inversed FA cell and 12T bitcell.

회로 최적화

- CMOS형 읽기 포트(인버터와 전송 게이트)
  - RBL precharge 불필요
  - 동일 weight 재사용 시 RWL 유지 가능 → 전력 소모를 줄임
- Reversed-polarity 전가산기 도입
  - 논리 반전 구조로 트랜지스터 수 절감
  - 면적 12.5%, 전력 15% 감소

레이아웃 최적화

- 12T bit-cell 표준 셀 면적 : 0.075μm²
  - 파운드리 제공 6T/8T SRAM(dummy row 필요)보다 전체 면적 더 작을 수 있음
  - 이유 : 로직 ↔ 비트셀 전환 영역에서 dummy 영역 없이 직접 연결
- 비트셀 배열 내 셀 경계는 일반 셀 배치와 다르게 함
  - Source–Drain(S-D 공유) 기준으로 정렬하여 면적 최소화
- WWL 인터리브 + RWL 공유 구조
  - 레이아웃 라우팅 효율적으로 배치

IV. Results & Discussion

1. results

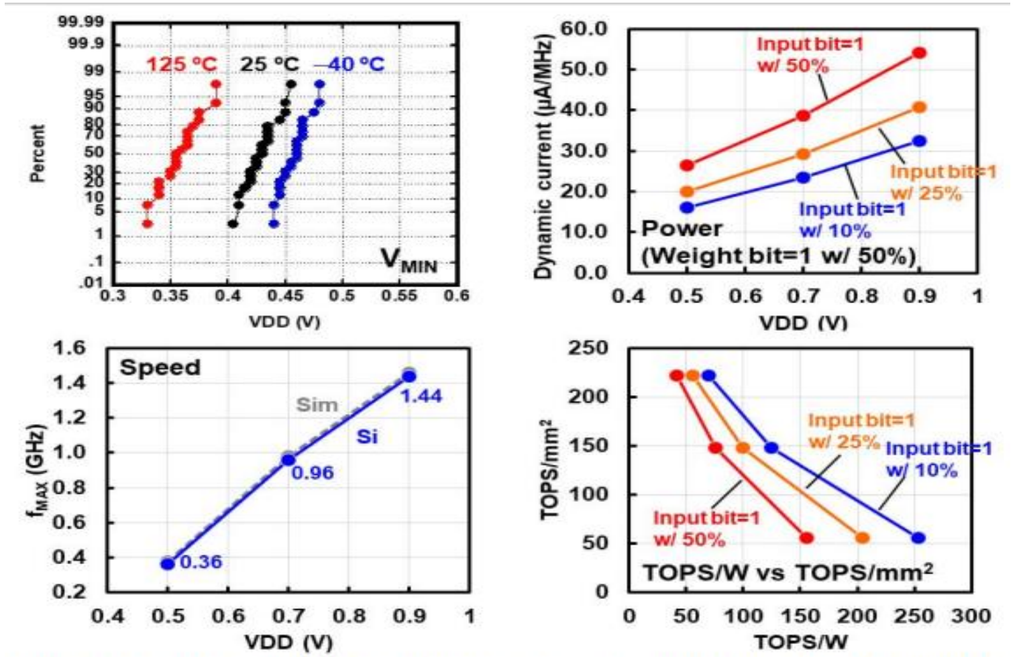
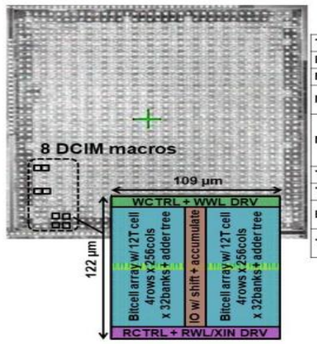


Figure 11.6.5: Measurement results: V_{MIN}, f_{MAX}, dynamic power and TOPS/mm² vs TOPS/W.



Technology	5 nm HKMG FinFET
Bitcell	Low voltage 12T (0.075μm²)
Macro size	109 μm x 122 μm
Memory config.	64 kb 256 addresses x 256 IOs w/ MUX1
MAC size	1024 MAC / CLK cycle 64 4b signed/unsigned Input 4b signed/unsigned weight w/ 4rows 64 14b singed/unsigned Output
TOPS	0.74 @0.5V ~ 2.95 @0.9V
TOPS/mm²	55.3 @0.5V ~ 221.2 @0.9V
Energy (μW/MHz)	16.2 ~ 32.5 (Input bit=1 w/ 10%) 20.0 ~ 40.7 (Input bit=1 w/ 25%)
TOPS/W	69.9 ~ 253.5 (Input bit=1 w/ 10%) 55.8 ~ 205.0 (Input bit=1 w/ 25%)

Figure 11.6.7: Test chip micrograph.

Specification

- ❖ 공정 : 5nm (HKMG) FinFET 공정 노드, DCIM macro 8개 구현
- ❖ Macro 면적 :
  - ➢ 1개당 0.0133 mm² (109μm × 122μm)
  - ➢ 유사 아키텍처의 디지털 플로우 구현 대비 3배 면적 절감된 수치
- ❖ 동작 특성 :
  - ➢ VMIN < 0.5V @ -40°C (95% yield)
  - ➢ fMAX = 0.36 / 0.96 / 1.44 GHz @ 0.5 / 0.7 / 0.9V
- ❖ 성능 :
  - ➢ 221.2 TOPS/mm² @ 0.9V, 55.3 TOPS/mm² @ 0.5V (4b/4b)
  - ➢ 메모리 용량(row)을 줄이면, TOPS/mm² 성능이 향상
  - ➢ 최대 253.5 TOPS/W @ 0.5V (전력 대비 성능이 가장 좋음)
  - ➢ 희소도 조건에 따라 전력 소모도 변화함

IV. Results & Discussion

2. 비교

	ISSCC'21 [6]	ISSCC'20 [1]	ISSCC'20 [2]	VLSIC'21 [3]	ISSCC'21 [4]	This work
Technology	5 nm	7 nm	28 nm	28 nm	22 nm	5 nm
MAC operation	MAC array in NPU core	Analog CIM (current base)	Analog CIM (charge share)	Analog CIM (charge inject)	Digital CIM	Digital CIM
Array size	2 kb	4 kb	64 kb	32 kb	64 kb	64 kb
Bitcell type	N/A	1R1W 8T	1RW 6T	10T1C	1RW 6T	1R1W 12T
Bitcell area	N/A	0.053 μm²	0.25 μm²	N/A	0.379 μm²	0.075 μm²
Macro area	0.022 mm²	0.0032 mm²	N/A	N/A	0.202 mm²	0.0133 mm²
Voltage	0.55 V ~ 0.9 V	0.8 V	0.7 V ~ 0.9 V	1 V	0.72 V	0.5 V ~ 0.9 V
# of Input Ch	16	64	16	256	256	64
# of Output Ch	16	16	16	128	64	64
Input bit	8/16	4	4~8	1/2	1~8	4
Weight bit	8/16	4	4/8	1/2	4	4
Output bit	N/A	4 (truncation)	12	4 (truncation)	16	14
Simultaneous MAC + Write	No	No	No	No	No	Yes
TOPS/mm²	28 (8b)	116	N/A	N/A	16	221 (4b) 55 (8b*)
TOPS/W	13.6 (system level)	262~610	68.4	588	89	254 (4b) 63 (8b*)

(*) Estimated value

Figure 11.6.6: PPA summary table and comparison with prior works.

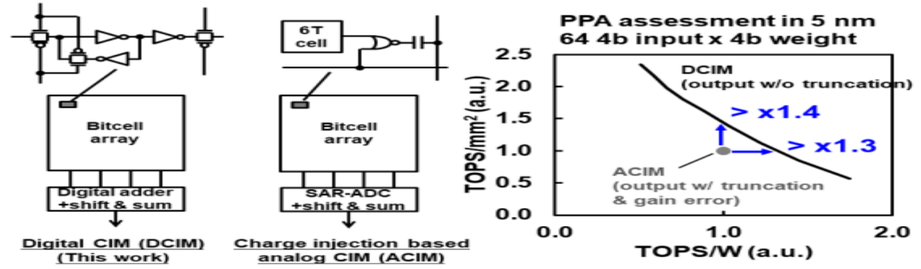


Figure 11.6.1: PPA comparison between digital CIM and analog CIM in a 5nm technology node.

기존 ACIM 대비 DCIM의 장점

- ❖ 정확도 손실 없음
  - ➢ ADC 미사용 : truncation(precision)/gain error 발생 없음
- ❖ 유연한 비트 폭 지원 (signed/unsigned)
- ❖ DVFS 가능
- ❖ 우수한 DFT(Testability) 특성 제공
- ❖ 5nm 공정 DCIM은 ACIM 대비
  - ➢ 전력 효율 1.3×, 성능/면적 효율 1.4× 향상
- ❖ 디지털 플로우 기반 구현
  - ➢ 7nm->5nm 공정 스케일링과 APR사용하여 로직 밀도 1.5~2x, 전력/성능 1.2x 개선 가능

ACIM의 한계 (고급 공정 노드에서)

- ❖ ADC 사용 : 동작 범위 제한이 되어있어 전압 스케일링 시 문제
- ❖ 전류 기반 ACIM : DIBL로 인한 드레인 전류 비선형성 증가(안정성)
- ❖ 정전용량 기반 ACIM : 기술 스케일링 시 정전용량(cap) 감소 문제

V. Conclusion

- ❖ 이전 논문의 22nm 기반 설계를 발전시켜, 5nm 공정에서 12T 1R1W bit-cell 기반 DCIM macro를 구현하여 공정 스케일링의 실효성을 입증 하였음
- ❖ DCIM이 ACIM대비 PPA, 유연성, 정확도, 확장성, 테스트 용이성 측면에서 모두 우수함
- ❖ 제안된 구조로, 다양한 비트폭의 input과 weight의 MAC연산 수행 가능
- ❖ MAC연산과 weight 업데이트 병행 수행, DVFS 지원 가능함

감사합니다!