C3SRAM: An In-Memory-Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism

전자공학부 박종훈

목차

사전지식 개념 공부

- 인메모리 컴퓨팅(In-Memory Computing, IMC)
- IMC 설계에서 이진 가중치
- SRAM(Static Random Access Memory)
- C3SRAM
- 이진 곱셈-누산 연산(Multiply-Accumulate, MAC)
- 아날로그-혼합 신호(AMS) 용량 결합 컴퓨팅
- AMS 용량 결합 컴퓨팅의 원리와 장점
- 연산 방법 분류
  - 디지털 도메인에서의 멀티비트 활성화 연산
  - 아날로그 도메인에서의 멀티비트 활성화 연산

인메모리 컴퓨팅(In-Memory Computing, IMC)

- **전통적인 컴퓨팅의 한계**
- 폰 노이만 아키텍처에서는 데이터가 메모리와 CPU 사이를 오가며 연산 - 메모리 병목 현상
- **IMC의 기본 원리**
- IMC는 메모리 자체에서 계산을 수행함으로써 메모리와 CPU 간의 데이터 이동 ⬇
- **IMC의 장점**
- 에너지 효율성: 데이터 이동을 최소화함으로써 전력 소모⬇
- 속도: 메모리 내에서 직접 연산을 수행하기 때문에, 데이터 이동에 따른 지연 시간 ⬇
- 공간 효율성: 데이터 이동에 필요한 버스(bus)와 인터커넥트(interconnect) ⬇
  전체 시스템의 설계가 간단해지고 공간 효율성 ⬆
- **IMC 기술의 구현 방법**
- SRAM 기반 IMC: SRAM 셀을 이용해 메모리 내에서 연산을 수행하는 방법이다. 이는 주로 고속 연산이 필요한 경우에 사용된다.
- ReRAM 기반 IMC: ReRAM은 비휘발성 메모리로, 데이터 보존 능력이 뛰어나며, 전력 소모⬇
  에너지 효율성⬆, 전원을 끈 상태에서도 데이터를 보존가능
- DRAM 기반 IMC: DRAM은 대용량 데이터를 저장하는 데 유리하며, DRAM 내에서 간단한 연산을 수행함으로써 데이터 이동 ⬇

IMC 설계에서 이진 가중치

- 멀티비트 가중치 인메모리 컴퓨팅(IMC) 설계의 어려움을 해결하기 위한 주요 알고리즘적 발전 중 하나는 이진 가중치 네트워크(BWN)이다. BWN은 네트워크 가중치를 이진화하여 저장 제약을 완화하고 가중치 저장을 간단하게 만든다. BWN의 하위 집합인 이진 신경망(BNN)은 가중치와 활성화를 모두 이진화하여 연산을 단순화하고 하드웨어 구현을 용이하게 한다.
- **BWN**은 신경망의 가중치를 이진화하여 +1 또는 -1로 표현한다. 입력 및 출력 활성화는 멀티비트 값을 유지할 수 있다. 이러한 이진화된 가중치는 메모리 요구량을 줄이고, 가중치 저장을 간소화하며, 연산의 복잡성을 낮춘다.
- **BNN**은 BWN의 하위 집합으로, 가중치뿐만 아니라 활성화도 이진화한다. 이진화된 가중치와 활성화 값은 +1과 -1로 표현되며, 곱셈 연산을 단순한 XNOR 연산으로 대체할 수 있다.

$$y = XNOR(x, w)$$

- IMC(In-Memory Computing) 설계는 메모리 내에서 직접 연산을 수행함으로써 데이터 이동을 최소화하고, 전력 소모를 줄이며, 연산 속도를 향상시킨다. BWN과 BNN은 이러한 IMC 설계에서 특히 유리하다.

SRAM(Static Random Access Memory)

- SRAM의 구조와 작동 원리
  - SRAM 셀은 일반적으로 여섯 개의 트랜지스터로 구성된다. 이 트랜지스터들은 크로스 커플드 인버터(cross-coupled inverters) 형태로 연결되어 하나의 비트를 저장

- 장점
  - 빠른 읽기 및 쓰기 속도: SRAM은 데이터 접근 시간이 매우 짧아 고속 연산이 필요한 응용 분야에 적합하다.
  - 간단한 제어 회로: DRAM과 달리 리프레시 회로가 필요 없기 때문에 제어 회로가 상대적으로 간단하다.
  - 데이터의 안정성: 전원이 공급되는 동안 데이터가 안정적으로 유지된다.

C3SRAM

- **기본 개념**
- C3SRAM: 용량 결합 컴퓨팅(C3)을 기반으로 한 IMC SRAM 매크로
- C3SRAM은 인메모리 컴퓨팅 기능을 통합한 SRAM 매크로로, 신경망의 하드웨어 가속을 목표로 설계되었다. 이 매크로는 비트셀과 주변 장치에 회로가 내장된 SRAM 모듈로 구성되어 있으며, 이진화된 가중치와 활성화를 사용하는 신경망의 주요 연산을 수행한다. 특히, 아날로그-혼합 신호(AMS) 용량 결합 컴퓨팅을 활용하여 이진 신경망의 주요 연산인 이진 곱셈-누산 연산(MAC)을 평가한다.

- **메모리 셀과 행 액세스**
- 기존의 메모리 아키텍처에서는 개별 행을 선택하여 데이터를 읽고 쓰는 방식으로 작동한다. 그러나 C3SRAM은 모든 행을 동시에 어서트(activate)하여 병렬 연산을 수행한다. 이는 단일 행 접근 방식에 비해 훨씬 높은 효율성을 제공한다.

C3SRAM

- **아날로그 전압 형성**
- 저장된 가중치를 개별 행으로 액세스할 필요 없이, C3SRAM은 모든 행을 동시에 어서트하여 용량 전압 분할을 통해 읽기 비트라인 노드에서 아날로그 전압을 형성한다. 이 아날로그 전압은 이후 아날로그-디지털 변환기(ADC)를 통해 디지털 신호로 변환된다.

- **병렬 벡터-매트릭스 곱셈**
- 각 열마다 하나의 ADC를 사용하여, C3SRAM 매크로는 단일 사이클에서 완전히 병렬적인 벡터-매트릭스 곱셈을 구현한다. 이는 벡터와 매트릭스 간의 연산을 병렬로 처리함으로써 매우 높은 처리 속도를 달성할 수 있게 한다.

이진 곱셈-누산 연산(Multiply-Accumulate, MAC)

- 이 연산은 입력 데이터와 가중치를 곱한 후, 그 결과를 누적하여 최종 값을 계산하는 과정

$$MAC = \sum_{i=1}^{N} (x_i \times w_i)$$

- **연산 단계**
- 곱셈: 입력 데이터(활성화)와 가중치를 곱한다.
- 누산: 곱셈의 결과를 누적한다.
- 반복: 위 과정을 필요한 모든 입력 데이터에 대해 반복한다.

- **하드웨어 효율성**
- 이진 MAC 연산은 복잡한 곱셈 회로를 단순한 XOR 게이트와 덧셈기로 대체할 수 있다.
- 이는 하드웨어 구현 시 전력 소비를 줄이고, 연산 속도를 높이는 데 유리하다.

아날로그-혼합 신호(AMS) 용량 결합 컴퓨팅

개요

- 아날로그-혼합 신호(AMS) 용량 결합 컴퓨팅은 아날로그 신호 처리와 디지털 신호 처리를 결합하여 연산을 수행하는 기술이다. 이 접근 방식은 데이터 이동을 최소화하고 연산 효율성을 극대화할 수 있는 강력한 도구로, 특히 인메모리 컴퓨팅(In-Memory Computing, IMC)에서 큰 잠재력을 가지고 있다.

기본 구조

- AMS 시스템은 아날로그 및 디지털 신호 처리 회로를 동시에 포함한다. 이러한 시스템은 아날로그 신호의 장점(연속성 및 고해상도)을 활용하면서, 디지털 신호의 장점(정확성과 처리 용이성)을 결합하여 최적의 성능을 제공한다.
- 아날로그 신호 처리부: 신호 증폭, 필터링, 변조 등 아날로그 연산을 수행하는 회로를 포함한다. 아날로그 신호 처리부는 전압, 전류와 같은 연속적인 물리량을 직접 다루기 때문에 높은 해상도의 신호 처리가 가능하다.
- 디지털 신호 처리부: 아날로그-디지털 변환기(ADC)와 디지털 회로를 포함하여, 아날로그 신호를 디지털 신호로 변환한 후 이를 기반으로 다양한 디지털 연산을 수행한다. 디지털 신호 처리부는 데이터의 정확한 처리가 가능하며, 프로그래밍이 용이하다.

AMS 용량 결합 컴퓨팅의 원리와 장점

용량 결합 컴퓨팅의 원리

- 전압 분할: 여러 메모리 셀의 전압을 병렬로 결합하여 하나의 합산된 아날로그 전압을 생성한다. 각 셀의 전압은 해당 셀에 저장된 데이터(예: 비트 값)에 따라 다르다.
- 용량 충전/방전: 생성된 아날로그 전압을 이용해 커패시터(용량)를 충전하거나 방전한다. 이 과정에서 발생하는 전압 변화는 데이터의 가중 합에 비례한다.
- 아날로그-디지털 변환: 충전된 전압은 ADC를 통해 디지털 신호로 변환된다. 이 디지털 신호는 이후 디지털 연산에 사용된다.

AMS 용량 결합 컴퓨팅의 장점

- 에너지 효율성: 아날로그 신호 처리는 데이터 이동을 최소화하고, 연산을 메모리 내에서 직접 수행함으로써 전력 소모를 줄인다.
- 속도: 아날로그 연산은 매우 빠르게 수행되며, 디지털 변환 과정도 고속으로 이루어지기 때문에 전체 연산 속도가 크게 향상된다.
- 병렬 처리: 다수의 메모리 셀을 동시에 어서트하여 병렬 연산을 수행할 수 있어, 대규모 데이터 처리가 가능하다.
- 고밀도 집적: 아날로그 회로와 디지털 회로를 동일한 칩에 통합함으로써, 공간 효율성을 높이고 시스템 복잡성을 줄일 수 있다.

연산 방법을 디지털 도메인과 아날로그 도메인으로 분류

- 디지털 도메인:
  - Vesti 아키텍처
  - XNOR-SRAM
- 아날로그 도메인:
  - 디지털-아날로그 변환(DAC) 기술:
    - 전압 기반 DAC 회로 : Twin-8T IMC 설계, XNOR-SRAM
    - 전류 기반 DAC 회로 : PWM 기반 DAC, 전류 소스 보정
  - 전류/전하 도메인에서의 연산
    - 전류 도메인 연산: XNOR-SRAM, Zhang et al. 연구
    - 전하 도메인 연산: Valavi et al. 연구, Conv-SRAM, C3SRAM 매크로

디지털 도메인에서의 멀티비트 활성화 연산

- **디지털 도메인**: 멀티 사이클 연산과 부분 합 누적을 통해 높은 정확도를 달성할 수 있지만, 에너지 소비와 지연 시간이 증가한다.
- **Vesti 아키텍처와 XNOR-SRAM**
- Vesti 아키텍처: 멀티 사이클 연산을 통해 멀티비트 활성화 연산을 수행한다. 이 아키텍처에서는 여러 사이클에 걸쳐 부분 합을 누적하여 최종 결과를 얻는다.
- XNOR-SRAM: 이진 또는 삼진 활성화만 지원한다. 디지털 부분 합 출력을 통해 여러 사이클에 걸쳐 누적하여 멀티비트 활성화 연산 결과를 얻는다.
- **디지털 연산의 특징**
- 멀티 사이클 연산: 멀티비트 활성화 연산을 수행하기 위해 여러 사이클에 걸쳐 연산을 수행한다.
- 부분 합 누적: 각 사이클에서 부분 합을 계산하고, 이를 누적하여 최종 결과를 얻는다.
- 에너지 소모와 지연 시간: 높은 정확도를 달성하기 위해서는 더 많은 에너지와 시간이 필요하다. 이는 연산을 여러 번 반복하여 정확한 결과를 얻는 과정에서 발생한다.

아날로그 도메인에서의 멀티비트 활성화 연산

- **아날로그 도메인**: DAC 기술을 사용하여 입력 활성화를 전압 또는 전류로 변환함으로써 연산을 단순화 할 수 있다. 그러나 전압/전류 기반 DAC의 해상도 제한과 설계 과제를 극복해야 한다.
- **디지털-아날로그 변환(DAC)**
- DAC 기술: 입력 활성화를 디지털에서 아날로그로 변환하여 전압 또는 전류로 표현한다.
- **전압 기반 DAC 회로**
- Twin-8T IMC 설계: 네 가지 고유한 전압 레벨을 사용하여 입력 활성화를 나타낸다.
- XNOR-SRAM: 세 가지 전압 레벨을 지원하여 삼진 MAC 연산을 수행할 수 있다.
- **전류 기반 DAC 회로**
- PWM 기반 DAC: 입력 활성화를 펄스로 변환하고, 펄스의 창 내에서 충전 전류를 커패시티브 요소에 적용한다.
- 전류 소스 보정: 전류 소스의 비선형성을 보상하기 위해 추가 트랜지스터를 사용한다.

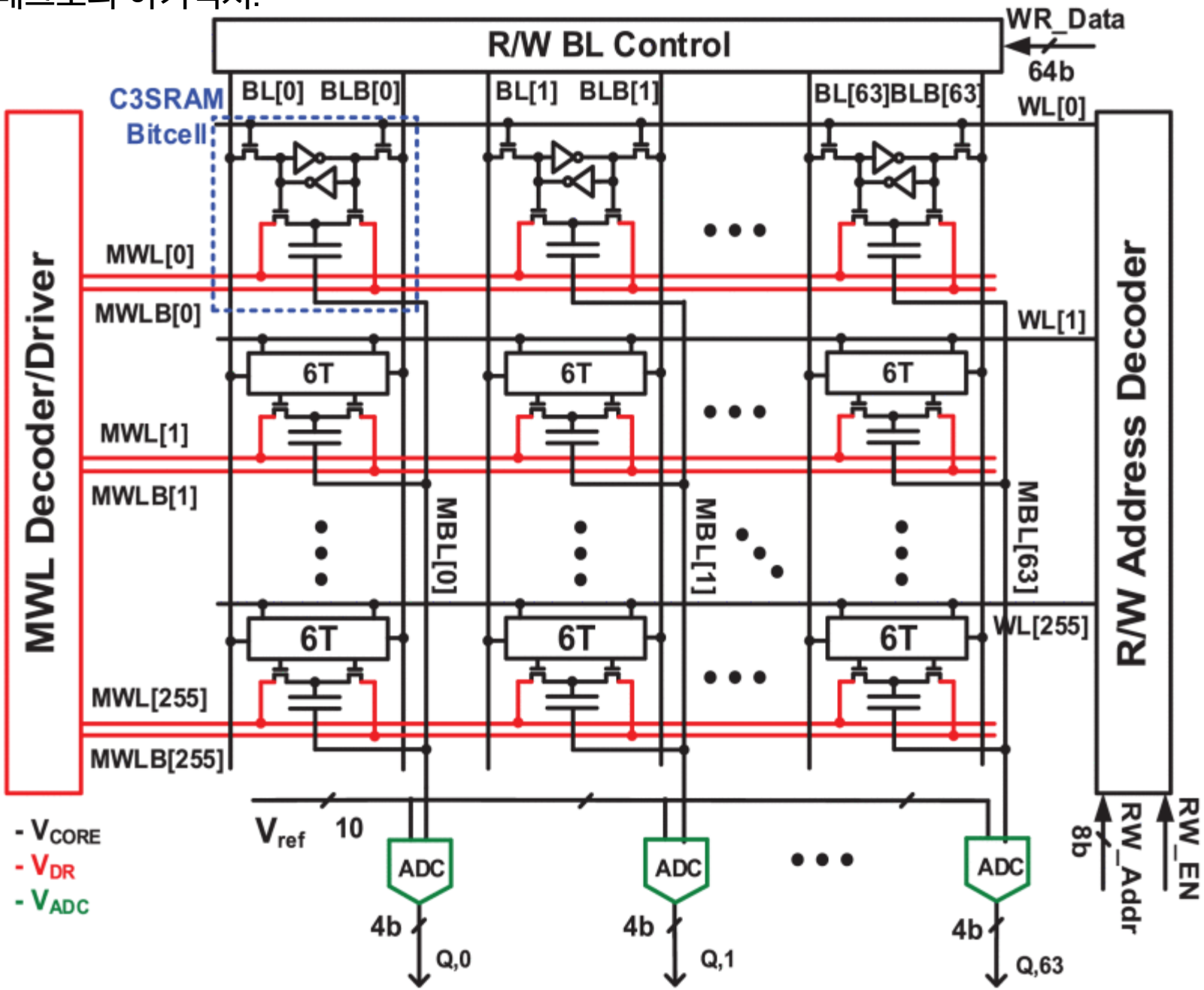
아날로그 도메인에서의 멀티비트 활성화 연산

- **전류/전하 도메인에서의 연산**
- **전류 도메인 연산:**
- XNOR-SRAM: 활성화 입력과 저장된 가중치에 따라 풀업/풀다운 트랜지스터를 켜서 저항성 전압 분배기를 생성한다.
- Zhang et al. 연구: 워드라인을 방전/충전하여 PWM 기반 DAC와 유사한 MAC 연산을 구현한다. 전류 소스의 비선형성을 보상하기 위해 추가 트랜지스터를 사용한다.
- **전하 도메인 연산:**
- Valavi et al. 연구: 전하 공유를 통해 bMAC을 수행한다. 각 비트셀의 커패시터가 충전/방전되고, 이를 통해 전하를 공유하여 누산한다.
- Conv-SRAM: 전하 공유를 통해 MAC 연산을 수행하며, 멀티비트 활성화는 PWM 기반 DAC에서 파생된다.
- C3SRAM 매크로: 전하 도메인에서 bMAC을 계산하기 위해 용량 결합을 사용한다.

논문 분석

- 그림 1. - C3SRAM IMC 매크로의 아키텍처.
- 그림 2. - C3SRAM 비트셀 설계 및 셀 내 bMAC 피연산자 테이블.
- 그림 4. - bMAC의 용량 결합 기반 인메모리 연산.
- 그림 5. - (a) TT 코너 시뮬레이션에서 온도 및 게이트 전압에 따른 MOSCAP 커패시턴스. (b) 다양한 온도에서 MOSCAP 커패시티브 전압 분배기 전송 함수.
- 그림 6. 이중 샘플링 자체 보정 단일 종단 비교기의 동작.
- 그림 7. - 아날로그 비이상성 감소를 위한 신호 전환 순서.
- 그림 17. - C3SRAM 기반 IMC에 컨볼루션 신경망을 매핑.

그림 1. - C3SRAM IMC 매크로의 아키텍처.



구성 요소 설명

- C3SRAM Bitcell:
- 각 셀은 6T (6-transistor) 구조를 가지며, 읽기 및 쓰기 연산을 지원한다.

- MWL Decoder/Driver:
- MWL (Main Word Line) 및 MWLB (Main Word Line Bar) 디코더/드라이버는 해당 워드라인을 활성화 시키는 역할을 한다.
- 각 비트셀에 접근하기 위해 주소를 디코딩하고, 적절한 워드라인을 활성화한다.

- R/W BL Control:
- 읽기/쓰기 비트라인 제어 회로이다. 이는 비트라인(BL)과 비트라인 바(BLB)를 제어하여 데이터의 읽기 및 쓰기 연산을 관리한다.
- BL과 BLB는 각각 64개의 비트라인을 가지며, 각각의 비트라인은 여러 비트셀과 연결되어 있다.

- ADC (Analog-to-Digital Converter):
- 아날로그 데이터를 디지털 데이터로 변환하는 회로이다.
- 각 ADC는 4비트 출력을 가지며, 최종적으로 디지털 데이터 Q를 생성한다.
- 두 개의 ADC는 각각 MBL (Main Bit Line)의 출력 값을 디지털 신호로 변환한다.

- Vref:
- 참조 전압을 제공하는 회로이다.
- VCORE, VDRC, VADC는 각각 코어 전압, 드레인 전압, ADC 전압을 나타내며, 회로의 동작 전압을 제어한다.

- R/W Address Decoder:
- 읽기/쓰기 주소 디코더이다.
- RW_ADDR 신호를 받아서 해당 주소에 맞는 비트셀을 선택하여 읽기 또는 쓰기 연산을 수행한다.
- RW_EN 신호는 읽기/쓰기를 활성화하는 신호이다.

동작 설명

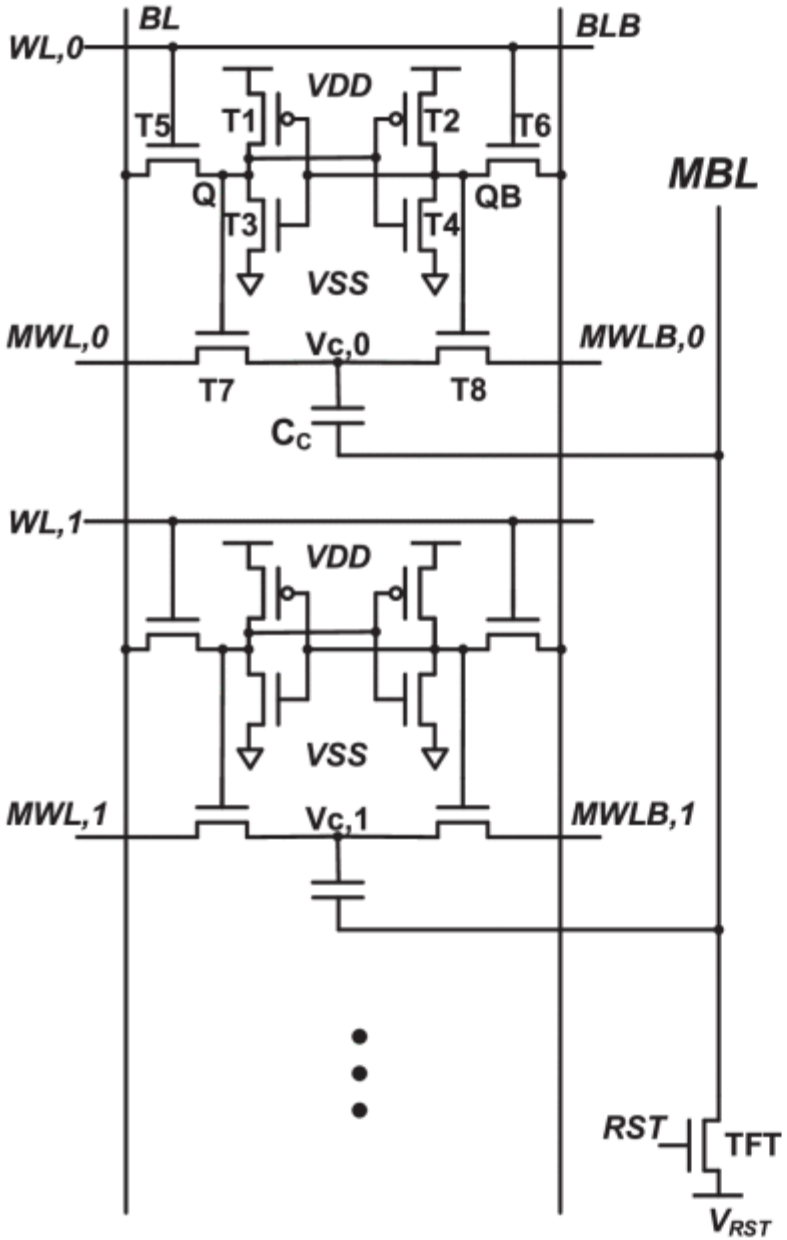
읽기 연산:

- 특정 주소에 맞는 워드라인(WL)이 활성화된다.
- 활성화된 워드라인에 의해 선택된 비트셀의 데이터가 비트라인(BL)으로 전달된다.
- 비트라인에 있는 아날로그 신호는 ADC에 의해 디지털 데이터로 변환된다.
- 변환된 데이터는 Q 출력으로 제공된다.

쓰기 연산:

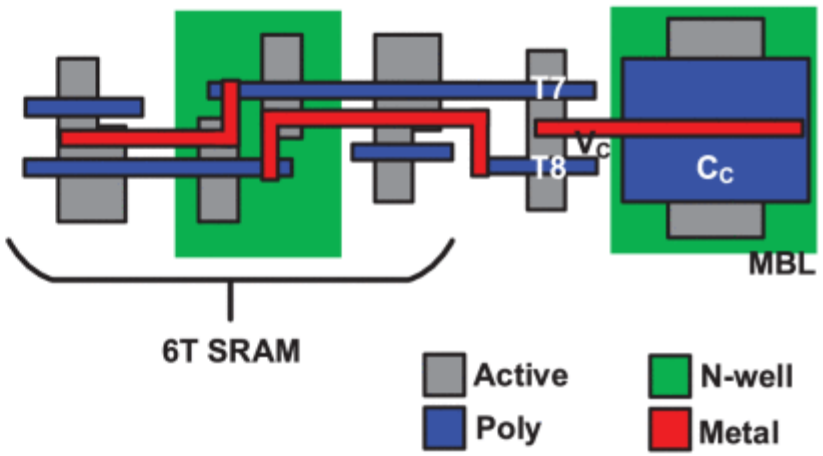
- 특정 주소에 맞는 워드라인(WL)이 활성화된다.
- R/W BL Control이 활성화되어 새로운 데이터가 비트라인(BL)을 통해 비트셀로 전달된다.
- 전달된 데이터는 비트셀에 저장된다.

그림 2. - C3SRAM 비트셀 설계 및 셀 내 bMAC 피연산자 테이블.



Input	Weight	+1	-1
	Q	VDD	0
	QB	0	VDD
MWL	MWLB		
+1	V_{DR}	0	V_{DR}	0
-1	0	V_{DR}	0	V_{DR}
0	V_{RST}	V_{RST}	V_{RST}	V_{RST}
Reset	V_{RST}	V_{RST}	V_{RST}	V_{RST}

Vc denote in red



구성 요소 설명

C3SRAM Bitcell (위쪽 회로도):

- 각 비트셀은 6T SRAM 구조를 기본으로 하고, 추가로 트랜지스터 T7, T8 및 커패시터 C_c 를 포함한다.
- 트랜지스터 (T1-T6): 6T SRAM 셀을 구성하는 트랜지스터들로, 데이터 저장 및 읽기/쓰기를 담당한다.
- 트랜지스터 (T7, T8): MWL (Main Word Line)과 MWLB (Main Word Line Bar)에 의해 제어되며, MBL (Main Bit Line)과 커패시터 C_c 를 연결한다.
- 커패시터 (C_c): 전하를 저장하고, 전하 도메인에서의 연산을 지원한다.

동작 표 (오른쪽 상단):

- 이 표는 입력과 가중치의 조합에 따른 각 노드의 전압 상태를 나타낸다.
- 입력 (+1, -1, 0)과 가중치 (+1, -1)의 조합에 따라, 각 노드 (MWL, MWLB, Q, QB)의 전압 상태가 결정된다.
- 리셋 상태: 초기화 시 V_RST가 각 노드에 인가된다.

6T SRAM 레이아웃 (오른쪽 하단):

- 6T SRAM 셀의 물리적인 레이아웃을 보여준다.
- 활성 영역, 폴리실리콘, 금속 배선 (Active, Poly, Metal): 각각의 트랜지스터와 배선이 어떻게 구성되는지를 나타낸다.
- N-well 및 P-well: 트랜지스터의 타입에 따라 N-well과 P-well이 사용된다.
- 커패시터 (C_c): 커패시터의 물리적 위치와 연결 상태를 보여준다.

이진 점 곱을 수행하기 위해, 커패시터는 저장된 가중치와 그 보완에 의해 게이트된 패스 트랜지스터를 통해 MAC 워드라인(MWL/MWLB)에 의해 충전/방전된다.

패스 트랜지스터가 NFET이기 때문에, MWL과 메모리 코어가 동일한 전압 소스를 가지면 T7/T8을 통한 매우 가변적인 문턱 전압(Vt) 드롭이 발생한다.

변동성 문제를 피하기 위해, T7/T8을 LVT 장치로 구현하고 MWL을 구동하기 위해 VCORE보다 200mV 낮은 VDR을 별도로 설정한다

(예: 1V VCORE에 대해 0.8V VDR).

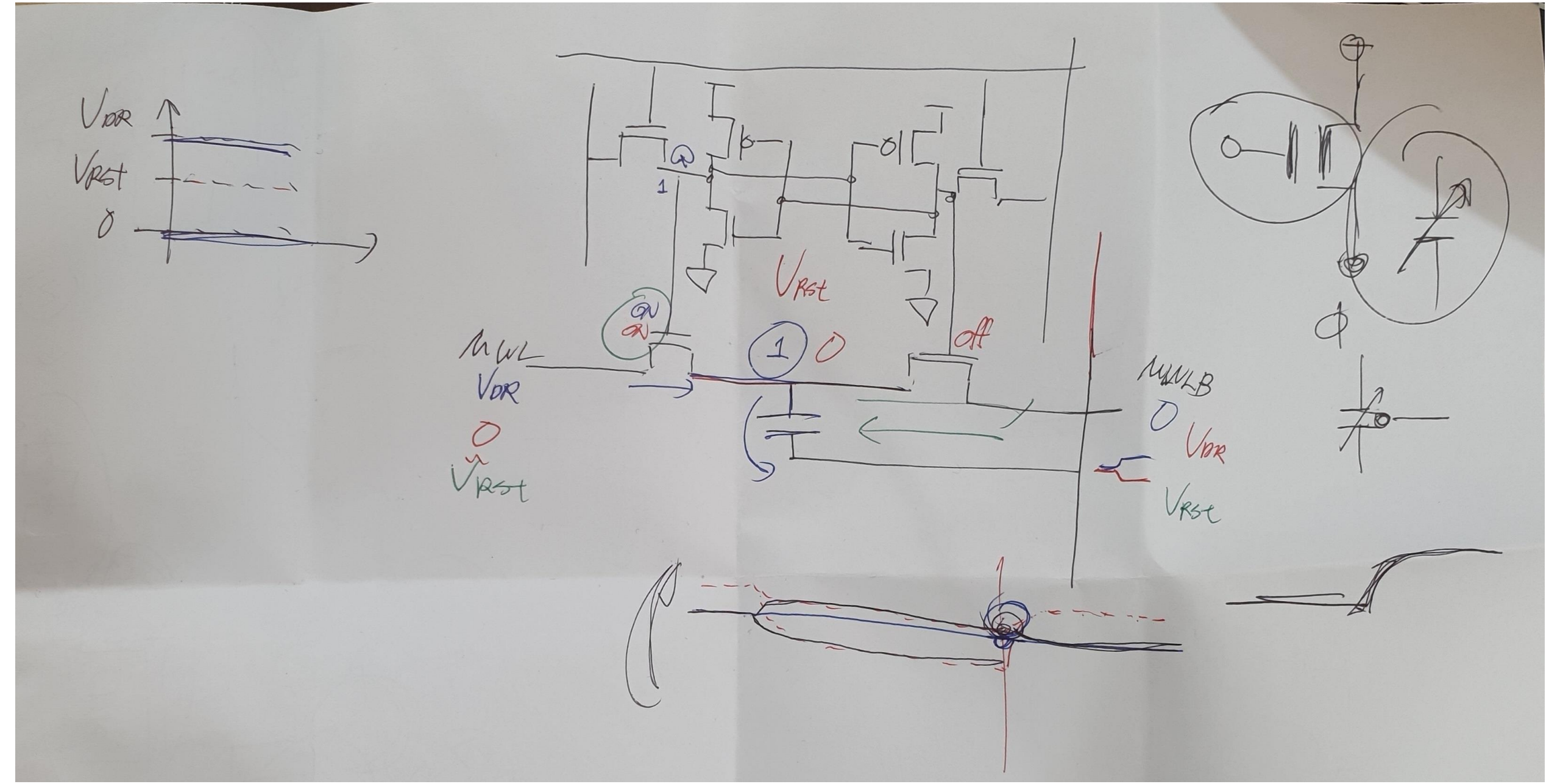
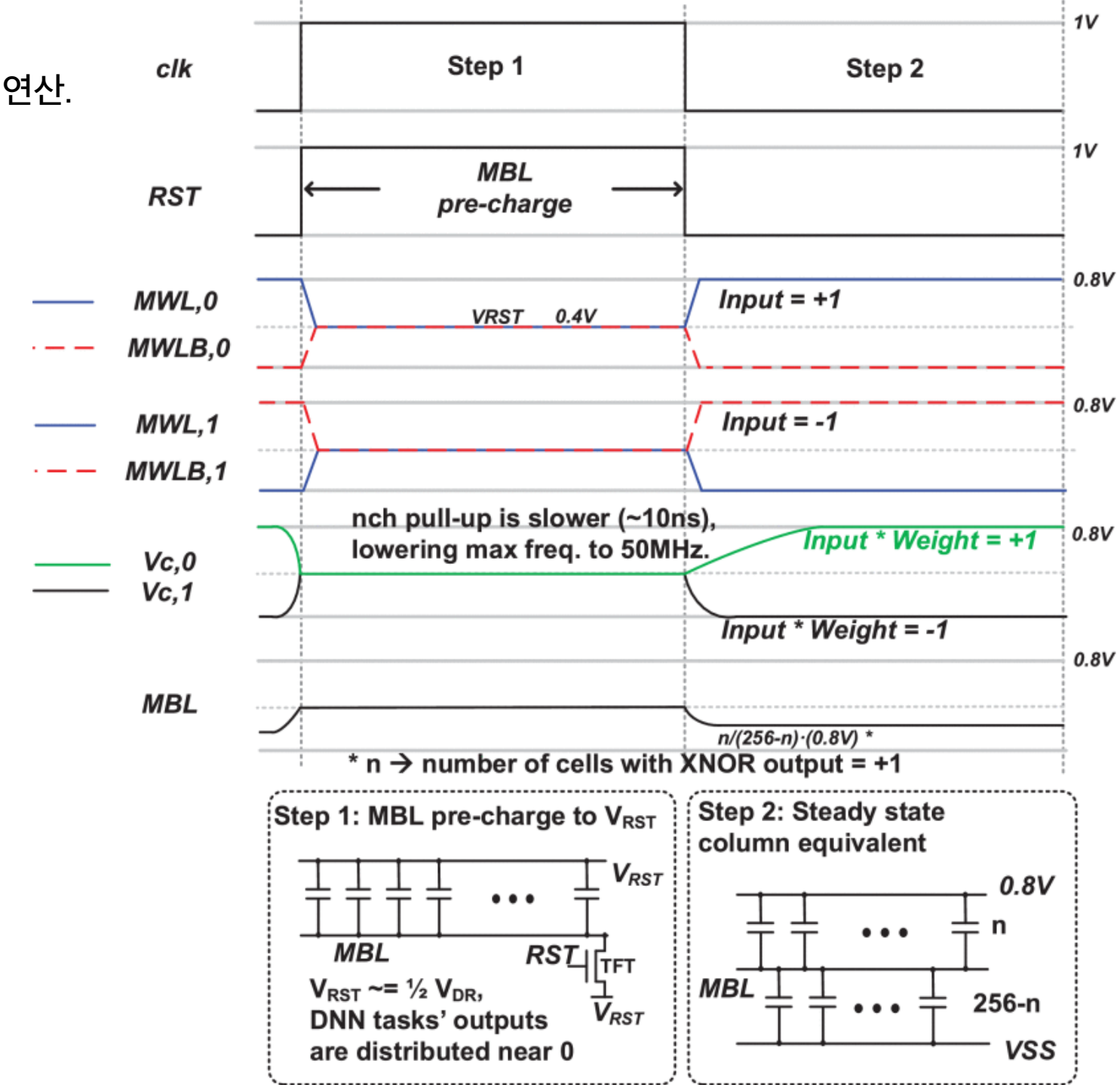

VDDR
VRST
0
1
VRST
1
0
off
MWL
VDDR
0
VRST
MWLB
0
VDDR
VRST
Φ

그림 4. - bMAC의
용량 결합 기반 인메모리 연산.



단계 1: MBL 사전 충전

MBL 사전 충전:

각 열의 메인 비트라인(MBL)은 풋터 TFT를 통해 V_RST = 0.5 * V_DR로 사전 충전된다. V_RST는 bMAC 출력이 0에 해당하는 전압(명목상 0.4V)에 가깝게 설정된다. 이는 MBL 노드의 전압 스윙을 최소화하기 위한 것이다.

MWL 및 MWLB 리셋:

각 행의 메인 워드라인(MWL) 및 메인 워드라인 바(MWLB)도 마찬가지로 V_RST로 리셋된다. 이는 비트셀 커패시터에 전압 잠재력이 없도록 하기 위함이다. 이 단계에서 커패시터는 사실상 병렬로 배열되며, 두 노드 모두 동일한 전압으로 리셋된다.

단계 2: 입력 및 가중치에 따른 연산

- **풋터가 꺼진다:**

풋터가 꺼져 MBL이 고유의 상태로 전환된다.

- **256개의 입력 활성화 적용:**

256개의 입력 활성화(In_i로 표시됨)가 256개의 MWL/MWLB에 병렬로 적용된다.

In_i가 +1인 경우, MWL은 VRST에서 VDR로, MWLB는 VSS로 구동된다.

In_i가 -1인 경우, MWL은 VRST에서 VSS로, MWLB는 VDR로 구동된다.

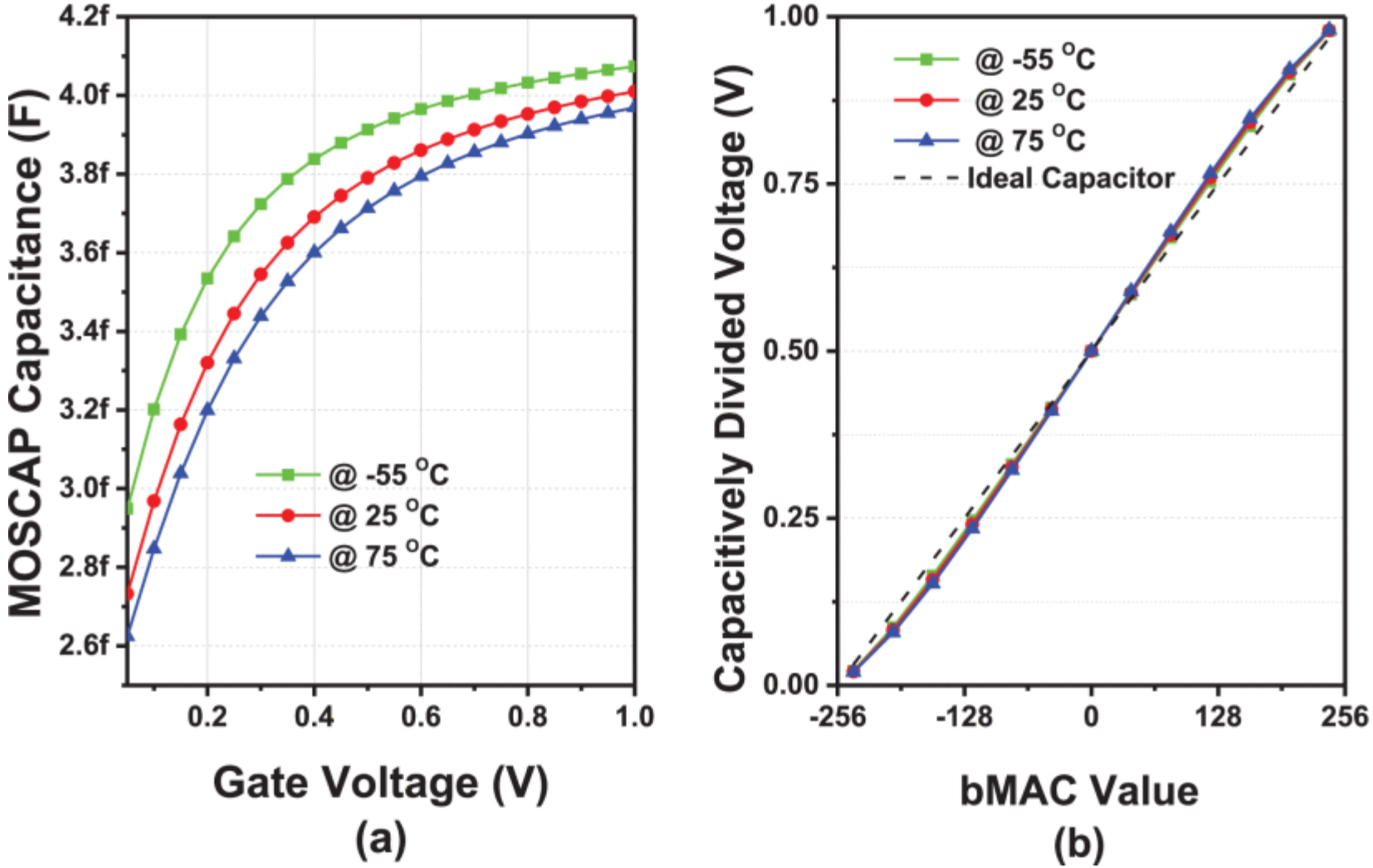
In_i가 0인 경우, MWL과 MWLB 모두 VRST에 머물러 동적 전력을 소모하지 않는다.

- **가중치 적용:**

가중치가 +1인 경우, T7을 통한 전압 램핑은 비트셀의 커패시터 CC(~4 fF)를 통해 변위 전류를 유도한다.

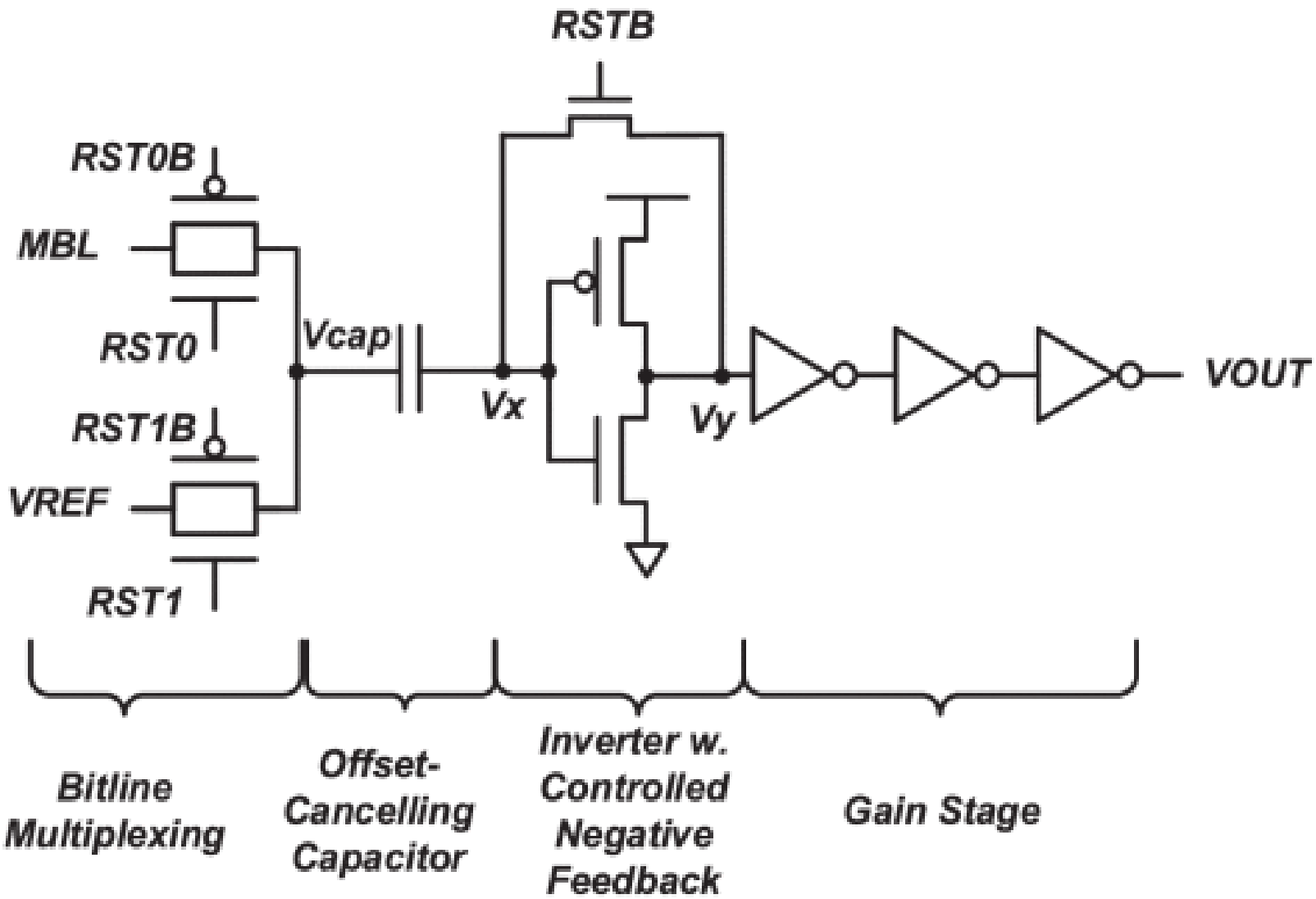
가중치가 -1인 경우, T8을 통한 전압 램핑이 발생한다.

그림 5. - (a) TT 코너 시뮬레이션에서 온도 및 게이트 전압에 따른 MOSCAP 커패시턴스.
(b) 다양한 온도에서 MOSCAP 커패시티브 전압 분배기 전송 함수.



- MOSCAP의 높은 커패시터 밀도는 bMAC 전송 곡선에 대해 MOMCAP을 사용할 때보다 더 넓은 전체 범위(FSR)를 제공한다.
- MOSCAP 커패시턴스는 온도와 게이트 전압에 따라 달라지며, 그림 5(a)는 온도 및 게이트 전압에 따른 CC 변화를 보여준다.
- 그림 5(b)는 다양한 온도에서 CC로 구성된 커패시티브 전압 분배기의 시뮬레이션 전송 함수를 보여준다.
- 온도와 관련된 비이상성은 전송 함수 안정성에 미치는 영향이 작으며, 전압과 관련된 비선형성은 전송 함수에 약간의 시그모이달 형태를 제공하여 ADC에 약간의 이점을 제공할 수 있다.

그림 6. 이중 샘플링 자체 보정
단일 종단 비교기의 동작.



구성 요소

비트라인 멀티플렉싱 (Bitline Multiplexing):

- RST0B, MBL, RST0, RST1B, VREF, RST1로 이루어진 멀티플렉싱 회로이다.
- 비교기 입력 커패시터(Vcap)에 연결된다.

오프셋 취소 커패시터 (Offset-Cancelling Capacitor):

- 오프셋 전압을 제거하여 비교기의 정확성을 높인다.

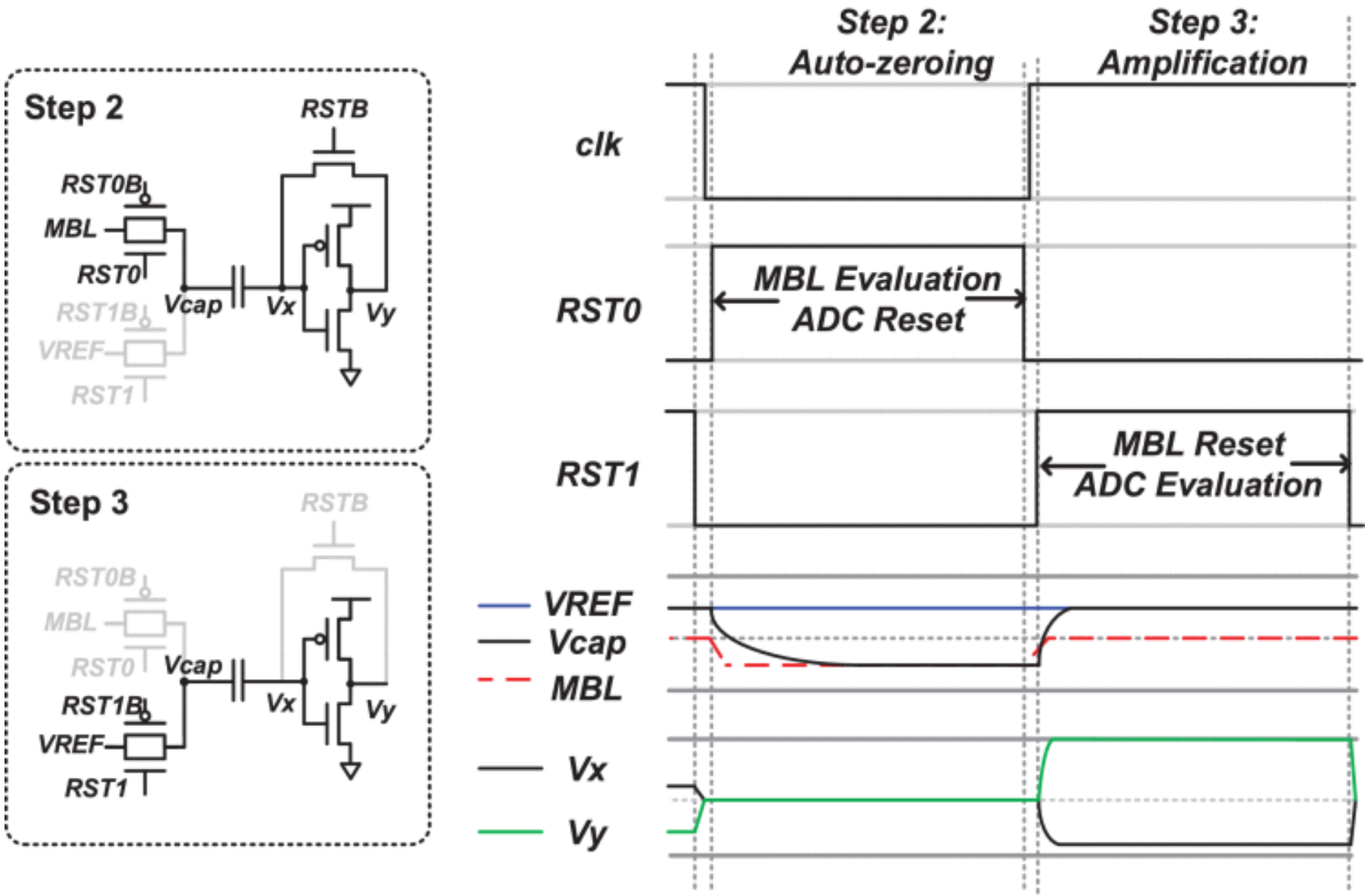
인버터와 제어된 부정적 피드백 (Inverter with Controlled Negative Feedback):

- 인버터가 증폭기로 작동하여 입력 신호를 증폭한다.
- 부정적 피드백을 통해 비교기의 성능을 안정화한다.

이득 단계 (Gain Stage):

- 인버터 체인을 통해 입력 신호를 추가로 증폭하여 디지털 도메인으로 신호를 변환한다.

그림 6. 이중 샘플링 자체 보정
단일 종단 비교기의 동작.



동작 단계

단계 2: 오토 제로잉 (Auto-zeroing):

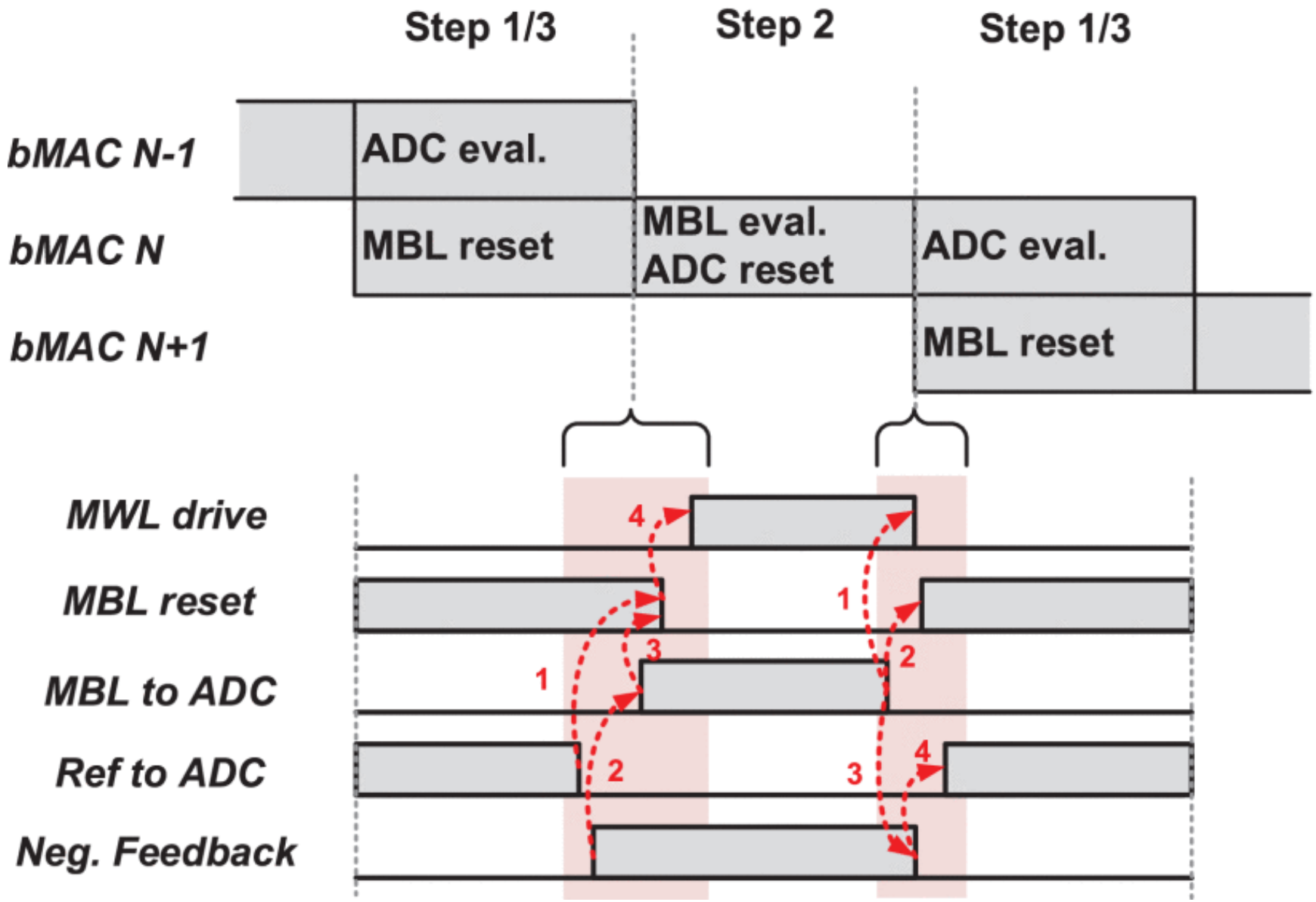
- bMAC 연산 중에 MBL이 비교기 입력 커패시터에 연결된다.
- 첫 번째 인버터의 입력과 출력이 닫혀 있어 인버터가 고전압 이득 영역에 위치한다.

단계 3: 증폭 (Amplification):

- 커패시터의 입력 노드가 기준 전압으로 전환되고, 부정적 피드백 경로가 차단된다.
- VMBL과 Vref 간의 전압 차이는 커패시터에 방전 또는 충전을 유발한다.
- 이전에 트립 포인트에서 균형을 맞춘 인버터는 유도된 전류의 방향에 따라 높거나 낮게 구동된다.
- 이득 단계 인버터 체인은 디지털 도메인으로 증폭을 완료한다.

이와 같이 C3SRAM의 ADC 동작은 두 단계로 나뉜다. bMAC 연산 중에는 MBL이 비교기 입력 커패시터에 연결되고, 오토 제로잉 과정이 진행된다. 증폭 단계에서는 기준 전압과의 차이에 따라 신호가 증폭되어 디지털 신호로 변환된다. 이를 통해 높은 정확도와 안정성을 가진 아날로그-디지털 변환이 이루어진다.

그림 7. - 아날로그 비이상성 감소를 위한 신호 전환 순서.



신호 전환 순서의 중요한 포인트

- 기준 전압의 분리:

기준 전압은 MBL이 리셋 상태를 벗어나기 전에 비교기 입력 커패시터에서 분리되어야 한다. 그렇지 않으면 기준 전압 소스가 MBL 부동 노드에 전하를 주입하게 된다.

- 부정적 피드백의 활성화:

인버터 단계의 부정적 피드백은 MBL이 입력 커패시터에 연결되기 전에 켜져야 한다. 그렇지 않으면 인버터 게이트가 트립 포인트로 구동되어 유도된 전류로 인해 VMBL이 영향을 받게 된다.

- MBL과 입력 커패시터의 연결:

MBL은 리셋 상태를 벗어나기 전에 비교기 입력 커패시터에 연결되어야 한다. 그렇지 않으면 커패시터에 저장된 기준 전압의 전하 차이가 부동 MBL에 주입될 것이다.

- MWL의 구동 시점:

MWL은 MBL이 부동 상태가 되기 전까지 구동되지 않아야 한다. 그렇지 않으면 일부 결합 전류가 방전될 수 있다.

신호 전환 순서의 예시 (그림 7 참고)

MWL 구동:

- MBL이 리셋 전압으로 전환되기 전에 MWL 구동이 이루어져야 한다.

MBL 리셋:

- MBL 리셋 풋터가 켜지기 전에 MBL이 비교기 입력에서 분리되어야 한다. 그렇지 않으면 입력 커패시터에 저장된 VMBL이 리셋되기 시작할 것이다.

부정적 피드백:

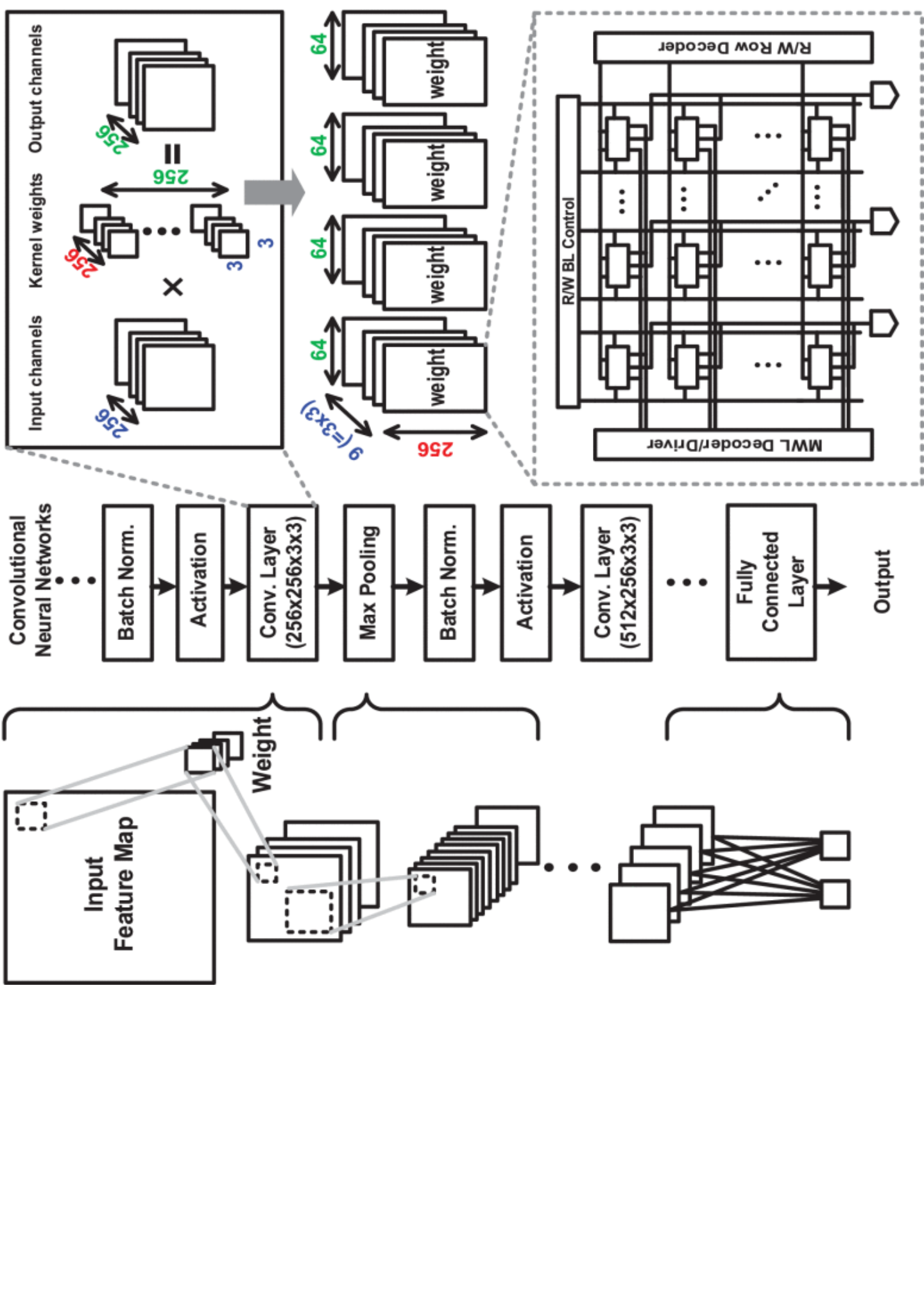
- 부정적 피드백이 꺼지기 전에 MBL이 분리되어야 한다. 그렇지 않으면 민감한 트립 포인트에 있는 인버터 입력이 교란될 수 있다.

기준 전압:

- 기준 전압이 비교기 입력에 연결되기 전에 부정적 피드백이 꺼져야 한다. 그렇지 않으면 전하 차이가 피드백 경로를 통해 방전될 것이다.

이와 같이 C3SRAM의 bMAC 연산에서 신호 전환 순서를 엄격히 준수함으로써 아날로그 비이상성을 최소화하고, 높은 정확도와 안정성을 유지할 수 있다.

그림 17. - C3SRAM 기반 IMC에 컨볼루션 신경망을 매핑.



이 그림은 C3SRAM을 이용한 신경망 작업 평가 과정을 설명하고 있다. C3SRAM은 컨볼루션 층 및 완전 연결(FC) 층의 계산을 담당하며, 기타 신경망 작업은 디지털 시뮬레이션에서 수행된다. 그림에는 입력 피처 맵, 신경망의 여러 층, 그리고 C3SRAM의 구조가 포함되어 있다.

신경망 구조와 작업

입력 피처 맵 (Input Feature Map):

- 신경망의 입력 데이터이다.

컨볼루션 신경망 (Convolutional Neural Networks):

- 배치 정규화 (Batch Norm.)
- 활성화 (Activation)
- 컨볼루션 층 (Conv. Layer)
- 예시: 256x256x3x3
- 맥스 풀링 (Max Pooling)
- 배치 정규화 (Batch Norm.)
- 활성화 (Activation)
- 컨볼루션 층 (Conv. Layer)
- 예시: 512x256x3x3x3
- 완전 연결 층 (Fully Connected Layer)
- 출력 (Output)

C3SRAM 매핑

컨볼루션 층 매핑:

- 입력 채널, 커널 가중치, 출력 채널로 구성된다.
- 예시: 256 입력 채널 x 3x3 커널 x 256 출력 채널
- 256개의 입력 채널을 3x3 커널과 곱하여 256개의 출력 채널을 생성한다.
- 3x3 커널은 9개의 가중치를 가진다.
- 각 커널 가중치는 여러 매크로에 분산된다.

완전 연결 층 매핑:

- 한 층의 가중치는 열 단위로 구성되고, 입력/활성화는 각 행에 적용된다.
- 컨볼루션 층의 매핑은 FC 층 매핑의 확장이다.
- 예: 256 뉴런 FC 층 가중치 9개를 매핑하는 것과 동일한 방식으로 컨볼루션 층의 3x3x256 필터를 매핑한다.

데이터 재사용 및 가중치 정적 매핑:

- 데이터 재사용을 최적화한 가중치 정적 매핑 스킴이 적용된다.
- ADC가 생성한 부분 합은 각 뉴런의 사전 활성화를 생성하기 위해 누적된다.

C3SRAM 구조

- MWL 디코더/드라이버 (MWL Decoder/Driver)
- 읽기/쓰기 비트라인 제어 (R/W BL Control)
- 읽기/쓰기 행 디코더 (R/W Row Decoder)

C3SRAM의 구조는 컨볼루션 신경망의 계산을 효율적으로 처리할 수 있도록 설계되었다. 대표적인 256 채널 3x3 커널 필터의 매핑은 그림에 설명되어 있으며, 이를 통해 컨볼루션 층의 계산을 효과적으로 수행할 수 있다. C3SRAM은 높은 정확도와 성능을 보장하며, 딥러닝 네트워크의 첫 번째 은닉층에서 모든 bMAC 연산을 수행한다.

감사합니다.