(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년06월07일
(11) 등록번호 10-2540226
(24) 등록일자 2023년05월31일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2023.01) *G11C 16/30* (2006.01)
H10B 51/30 (2023.01) *H10B 51/50* (2023.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G11C 16/30 (2013.01)
(21) 출원번호 10-2022-0094203
(22) 출원일자 2022년07월28일
심사청구일자 2022년07월28일
(56) 선행기술조사문헌
KR1020220044642 A
(뒷면에 계속)
전체 청구항 수 : 총 9 항

(73) 특허권자
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
서영교
경기도 부천시 길주로77번길 19-45, 920호
권대웅
서울특별시 서초구 주흥13길 38, 102동 103호
(74) 대리인
김남식, 이인행

심사관 : 이준상

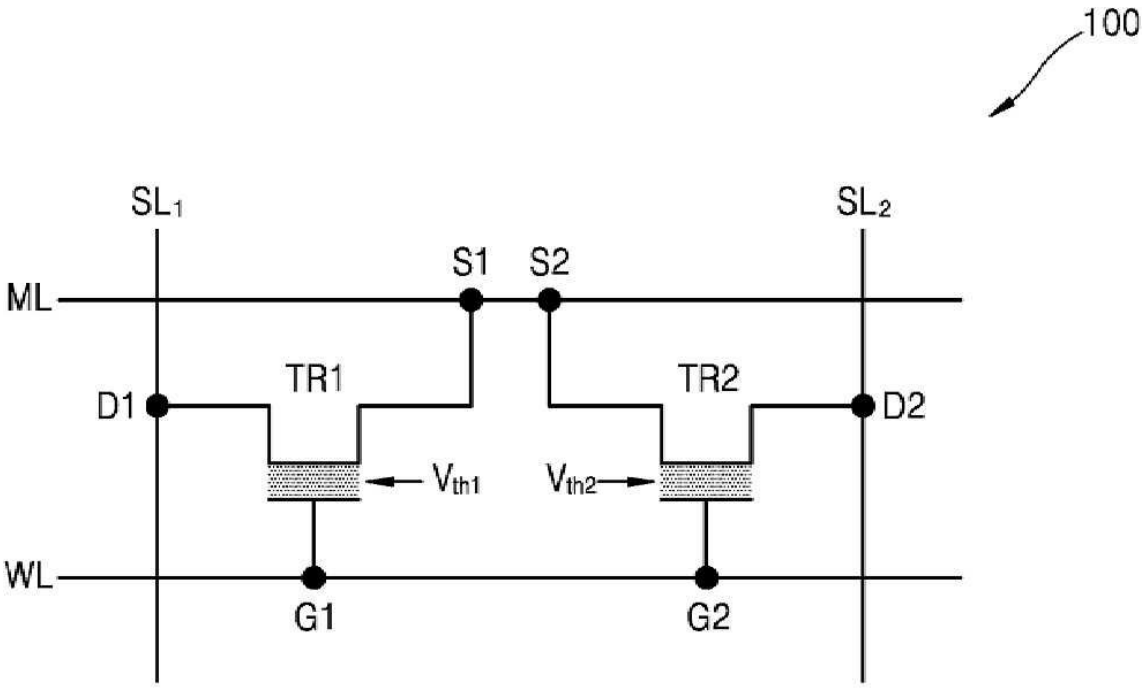
(54) 발명의 명칭 **3진 신경망 가속기 소자 및 그 동작 방법**

(57) 요 약

본 발명의 일 관점에 의한 3진 신경망 가속기 소자는, 제 1 소오스 단자, 제 1 드레인 단자 및 제 1 게이트 단자를 포함하는 제 1 반도체 소자와, 제 2 소오스 단자, 제 2 드레인 단자 및 제 2 게이트 단자를 포함하는 제 2 반도체 소자와, 상기 제 1 드레인 단자에 연결된 제 1 서칭 라인과, 상기 제 2 드레인 단자에 연결된 제 2 서칭 라

(뒷면에 계속)

대 표 도 - 도1



인과, 상기 제 1 소오스 단자 및 상기 제 2 소오스 단자에 공통으로 연결된 매칭 라인을 포함하고, 3진 처리를 위한 3진 웨이트 및 3진 입력이, 상기 제 1 반도체 소자의 제 1 문턱전압과 상기 제 2 반도체 소자의 제 2 문턱전압을 상대적으로 낮은 로우 문턱전압과 상대적으로 높은 하이 문턱전압으로 조합된 세가지 스테이트 중의 하나로 변경하는 제 1 동작과, 상기 제 1 서칭 라인 및 상기 제 2 서칭 라인에 인가되는 전압을 접지전압, 동작전압 및 중간 동작전압으로 조합된 세가지 조합 중의 하나로 설정하는 제 2 동작 중 서로 다른 어느 하나로 수행되고, 상기 3진 웨이트 및 상기 3진 입력의 조건에 따라서 9개의 연산 결과가 상기 매칭 라인을 통해서 출력된다.

(52) CPC특허분류
H10B 51/30 (2023.02)
H10B 51/50 (2023.02)

(56) 선행기술조사문헌
KR102348169 B1
KR101928223 B1
Ashok Srivastava, “Ternary input signal to binary bit output conversion CMOS integrated circuit design using neuron MOSFETs“, Microsystem Technologies (2022) 28:101-108 (2019.04.26.)*
S. Heo et al., "High-Speed Ternary CMOS Inverter by Monolithic Integration of NbO2 Threshold Switch with MOSFET," 2021 IEEE IEDM, 2021, pp.32.2.1-32.2.4,(2021.12.11.)*
*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업
과제고유번호 1711139247
과제번호 2021-0-02052
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 대학ICT연구센터육성지원사업
연구과제명 스마트 모빌리티를 위한 인공지능 시스템반도체 핵심 기술 개발 및 인력 양성
기 여 율 40/100
과제수행기관명 인하대학교 산학협력단
연구기간 2021.07.01 ~ 2028.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711169337
과제번호 2020R1F1A1051529
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 기본연구
연구과제명 CMOS 메모리와 차세대 메모리를 이용한 저전력 저면적 PIM (Processing In Memory) 설계
기 여 율 15/100
과제수행기관명 인하대학교
연구기간 2022.03.01 ~ 2023.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1415179211
과제번호 20009972
부처명 산업통상자원부
과제관리(전문)기관명 한국산업기술평가관리원
연구사업명 차세대지능형반도체기술개발
연구과제명 고효율 초저전력 경량 엣지 디바이스용 소자회로 및 SoC 개발(3차년도)
기 여 율 20/100
과제수행기관명 인하대학교
연구기간 2022.01.01 ~ 2022.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711173059
과제번호 2022M3I7A2085479
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 신재료PIM소자
연구과제명 [Ezbaro] (2022M3I7A2079155) 강유전체 메모리 기반 PIM 소자, 어레이, 구동회로 개발
기 여 율 20/100
과제수행기관명 인하대학교
연구기간 2022.04.20 ~ 2022.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345353993
과제번호 2022R1A6A1A03051705
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축
연구과제명 3D 나노융합소자 연구 센터
기 여 율 5/100
과제수행기관명 인하대학교
연구기간 2022.06.01 ~ 2023.02.28

명 세 서

청구범위

청구항 1

제 1 소오스 단자, 제 1 드레인 단자 및 제 1 게이트 단자를 포함하는 제 1 반도체 소자;

제 2 소오스 단자, 제 2 드레인 단자 및 제 2 게이트 단자를 포함하는 제 2 반도체 소자;

상기 제 1 드레인 단자에 연결된 제 1 서칭 라인;

상기 제 2 드레인 단자에 연결된 제 2 서칭 라인;

상기 제 1 게이트 단자 및 제 2 게이트 단자에 공통으로 연결된 워드라인; 및

상기 제 1 소오스 단자 및 상기 제 2 소오스 단자에 공통으로 연결된 매칭 라인을 포함하고,

3진 처리를 위한 3진 웨이트 및 3진 입력이,

상기 제 1 반도체 소자의 제 1 문턱전압과 상기 제 2 반도체 소자의 제 2 문턱전압을 상대적으로 낮은 로우 문턱전압과 상대적으로 높은 하이 문턱전압으로 조합된 세가지 스테이트 중의 하나로 변경하는 제 1 동작과, 상기 제 1 서칭 라인 및 상기 제 2 서칭 라인에 인가되는 전압을 접지전압, 동작전압 및 중간 동작전압으로 조합된 세가지 조합 중의 하나로 설정하는 제 2 동작 중 서로 다른 어느 하나로 수행되고,

상기 3진 웨이트 및 상기 3진 입력의 조건에 따라서 9개의 연산 결과가 상기 매칭 라인을 통해서 출력되고,

상기 제 1 동작에서 상기 세가지 스테이트는,

상기 제 1 문턱전압과 상기 제 2 문턱전압이 모두 하이 문턱전압인 제 1 스테이트;

상기 제 1 문턱전압이 하이 문턱전압이고 상기 제 2 문턱전압이 로우 문턱전압인 제 2 스테이트; 및

상기 제 1 문턱전압이 로우 문턱전압이고 상기 제 2 문턱전압이 하이 문턱전압인 제 3 스테이트를 포함하는,

3진 신경망 가속기 소자.

청구항 2

삭제

청구항 3

제 1 항에 있어서,

상기 제 2 동작에서 상기 세가지 조합은,

상기 제 1 서칭 라인과 상기 제 2 서칭 라인에 모두 중간 동작전압이 인가되는 제 1 조합;

상기 제 1 서칭 라인에 접지 전압이 인가되고, 상기 제 2 서칭 라인에 동작 전압이 인가되는 제 2 조합; 및

상기 제 1 서칭 라인에 동작 전압이 인가되고, 상기 제 2 서칭 라인에 접지 전압이 인가되는 제 3 조합을 포함하는,

3진 신경망 가속기 소자.

청구항 4

제 1 항에 있어서,

상기 3진 웨이트는 상기 제 1 동작으로 수행되고,

상기 3진 입력은 상기 제 2 동작으로 수행되는,

3진 신경망 가속기 소자.

청구항 5

제 1 항에 있어서,

상기 3진 웨이트는 상기 제 2 동작으로 수행되고,

상기 3진 입력은 상기 제 1 동작으로 수행되는,

3진 신경망 가속기 소자.

청구항 6

제 1 소오스 단자, 제 1 드레인 단자 및 제 1 게이트 단자를 포함하는 제 1 반도체 소자;

제 2 소오스 단자, 제 2 드레인 단자 및 제 2 게이트 단자를 포함하는 제 2 반도체 소자;

상기 제 1 드레인 단자에 연결된 제 1 서칭 라인;

상기 제 2 드레인 단자에 연결된 제 2 서칭 라인;

상기 제 1 게이트 단자 및 제 2 게이트 단자에 공통으로 연결된 워드라인; 및

상기 제 1 소오스 단자 및 상기 제 2 소오스 단자에 공통으로 연결된 매칭 라인을 포함하고,

3진 처리를 위한 3진 웨이트 및 3진 입력이,

상기 제 1 반도체 소자의 제 1 문턱전압과 상기 제 2 반도체 소자의 제 2 문턱전압을 상대적으로 낮은 로우 문턱전압과 상대적으로 높은 하이 문턱전압으로 조합된 세가지 스테이트 중의 하나로 변경하는 제 1 동작과, 상기 제 1 서칭 라인 및 상기 제 2 서칭 라인에 인가되는 전압을 접지전압, 동작전압 및 중간 동작전압으로 조합된 세가지 조합 중의 하나로 설정하는 제 2 동작 중 서로 다른 어느 하나로 수행되고,

상기 3진 웨이트 및 상기 3진 입력의 조건에 따라서 9개의 연산 결과가 상기 매칭 라인을 통해서 출력되고,

상기 매칭 라인은 상기 중간 동작 전압으로 미리 충전된, 3진 신경망 가속기 소자.

청구항 7

제 1 항에 있어서,

상기 제 1 반도체 소자 및 상기 제 2 반도체 소자는 강유전체 전계효과 트랜지스터(FeFET)인, 3진 신경망 가속기 소자.

청구항 8

제 7 항에 있어서,

상기 제 1 동작에서, 상기 하이 문턱전압은 상기 강유전체 전계효과 트랜지스터가 프로그램된 상태에서 문턱전압이고, 상기 로우 문턱전압은 상기 강유전체 전계효과 트랜지스터가 소거 상태에서 문턱전압인, 3진 신경망 가속기 소자.

청구항 9

제 1 항에 있어서,

상기 제 1 반도체 소자 및 상기 제 2 반도체 소자는 플래시(flash) 메모리 소자인, 3진 신경망 가속기 소자.

청구항 10

제 1 소오스 단자, 제 1 드레인 단자 및 제 1 게이트 단자를 포함하는 제 1 반도체 소자;

제 2 소오스 단자, 제 2 드레인 단자 및 제 2 게이트 단자를 포함하는 제 2 반도체 소자;

상기 제 1 드레인 단자에 연결된 제 1 서칭 라인;

상기 제 2 드레인 단자에 연결된 제 2 서칭 라인;

상기 제 1 게이트 단자 및 제 2 게이트 단자에 공통으로 연결된 워드라인; 및

상기 제 1 소오스 단자 및 상기 제 2 소오스 단자에 공통으로 연결된 매칭 라인을 포함하는 3진 신경망 가속기 소자의 동작 방법으로서,

상기 3진 신경망 가속기 소자에 3진 웨이트 및 3진 입력의 조건을 설정하고, 상기 3진 웨이트 및 상기 3진 입력의 조건에 따라서 9개의 연산 결과를 상기 매칭 라인을 통해서 출력하는 3진 동작을 수행하되,

상기 3진 웨이트 및 상기 3진 입력은,

상기 제 1 반도체 소자의 제 1 문턱전압과 상기 제 2 반도체 소자의 제 2 문턱전압을 상대적으로 낮은 로우 문턱전압과 상대적으로 높은 하이 문턱전압으로 조합된 세가지 스테이트 중의 하나로 변경하는 제 1 동작과, 상기 제 1 서칭 라인 및 상기 제 2 서칭 라인에 인가되는 전압을 접지전압, 동작전압 및 중간 동작전압으로 조합된 세가지 조합 중의 하나로 설정하는 제 2 동작 중 서로 다른 어느 하나로 수행되고,

상기 제 1 동작에서 상기 세가지 스테이트는,

상기 제 1 문턱전압과 상기 제 2 문턱전압이 모두 하이 문턱전압인 제 1 스테이트;

상기 제 1 문턱전압이 하이 문턱전압이고 상기 제 2 문턱전압이 로우 문턱전압인 제 2 스테이트; 및

상기 제 1 문턱전압이 로우 문턱전압이고 상기 제 2 문턱전압이 하이 문턱전압인 제 3 스테이트를 포함하는,

3진 신경망 가속기 소자의 동작 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 반도체 소자에 관한 것으로서, 더 상세하게는 신경망 가속기 소자 및 그 동작 방법에 관한 것이다.

배 경 기 술

[0002] 컴퓨터의 중앙처리장치(CPU)는 일반적으로 속도가 느리고 전력이 부족한 상태에 있다. 이에 반해서, 특정 작업 집합만 수행하도록 최적화된 프로세서가 저전력에서 빠르게 실행될 수 있으며 특정 알고리즘을 더 빠르게 실행하는 데 도움이 된다는 것이 알려져 있다. 이에 따라, 많은 기기들이 특정 카테고리 및 애플리케이션을 위해서 맞춤 제작되었으며, 디지털 신호 처리기(DSP), 그래픽 처리 장치(GPU) 및 신경망 가속기(neural network accelerator) 등이 개발되고 있다.

[0003] 특히, 4차 산업의 핵심기술로 꼽히는 인공지능(AI) 시대에서는 신경망 가속기와 같이 딥러닝과 추론 등의 AI 처리에 특화된 반도체 소자가 필수적이다. 신경망은 머신 러닝 응용 프로그램에서 데이터의 클러스터링 및 분류를 수행하는 알고리즘 세트의 이름이다. 인간의 뇌의 뉴런이 어떻게 작동하는 지와 비교하여 신경망이라고 칭하게 되었다. 신경망 가속기는 신경망 워크로드를 처리하도록 특별히 최적화된 프로세서로서, 매우 빠른 속도로 데이터를 수집하고 클러스터링하고 분류하는 효율적인 프로세서이다.

[0004] 기존의 신경망 가속기는 대부분 경량 네트워크를 통해서 2진 처리에 최적화되어 있다. 하지만, 이러한 2진 신경망 가속기는 2진 데이터만 처리하므로 연산 정확성이 낮다는 문제가 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 전술한 문제점을 해결하기 위한 것으로서, 기존 2진 신경망 가속기와 같은 간단한 구조로 경량 네트워크의 장점을 가지면서 3진 데이터 처리를 통해서 연상 정확성을 높일 수 있는 3진 신경망 가속기 소자 및 그 동작 방법을 제공하는 것을 목적으로 한다. 그러나 이러한 과제는 예시적인 것으로, 이에 의해 본 발명의 범위가 한정되는 것은 아니다.

과제의 해결 수단

[0006] 상기 과제를 해결하기 위한 본 발명의 일 관점에 의한 3진 신경망 가속기 소자는, 제 1 소오스 단자, 제 1 드레인 단자 및 제 1 게이트 단자를 포함하는 제 1 반도체 소자와, 제 2 소오스 단자, 제 2 드레인 단자 및 제 2 게이트 단자를 포함하는 제 2 반도체 소자와, 상기 제 1 드레인 단자에 연결된 제 1 서칭 라인과, 상기 제 2 드레인 단자에 연결된 제 2 서칭 라인과, 상기 제 1 소오스 단자 및 상기 제 2 소오스 단자에 공통으로 연결된 매칭 라인을 포함하고, 3진 처리를 위한 3진 웨이트 및 3진 입력이, 상기 제 1 반도체 소자의 제 1 문턱전압과 상기 제 2 반도체 소자의 제 2 문턱전압을 상대적으로 낮은 로우 문턱전압과 상대적으로 높은 하이 문턱전압으로 조합된 세가지 스테이트 중의 하나로 변경하는 제 1 동작과, 상기 제 1 서칭 라인 및 상기 제 2 서칭 라인에 인가되는 전압을 접지전압, 동작전압 및 중간 동작전압으로 조합된 세가지 조합 중의 하나로 설정하는 제 2 동작 중 서로 다른 어느 하나로 수행되고, 상기 3진 웨이트 및 상기 3진 입력의 조건에 따라서 9개의 연산 결과가 상기 매칭 라인을 통해서 출력된다.

[0007] 상기 3진 신경망 가속기 소자에 따르면, 상기 제 1 동작에서 상기 세가지 스테이트는, 상기 제 1 문턱전압과 상기 제 2 문턱전압이 모두 하이 문턱전압인 제 1 스테이트와, 상기 제 1 문턱전압이 하이 문턱전압이고 상기 제 2 문턱전압이 로우 문턱전압인 제 2 스테이트와, 상기 제 1 문턱전압이 로우 문턱전압이고 상기 제 2 문턱전압이 하이 문턱전압인 제 3 스테이트를 포함할 수 있다.

[0008] 상기 3진 신경망 가속기 소자에 따르면, 상기 제 2 동작에서 상기 세가지 조합은, 상기 제 1 서칭 라인과 상기 제 2 서칭 라인에 모두 중간 동작전압이 인가되는 제 1 조합과, 상기 제 1 서칭 라인에 접지 전압이 인가되고, 상기 제 2 서칭 라인에 동작 전압이 인가되는 제 2 조합과, 상기 제 1 서칭 라인에 동작 전압이 인가되고, 상기 제 2 서칭 라인에 접지 전압이 인가되는 제 3 조합을 포함할 수 있다.

[0009] 상기 3진 신경망 가속기 소자에 따르면, 상기 3진 웨이트는 상기 제 1 동작으로 수행되고, 상기 3진 입력은 상기 제 2 동작으로 수행될 수 있다.

[0010] 상기 3진 신경망 가속기 소자에 따르면, 상기 3진 웨이트는 상기 제 2 동작으로 수행되고, 상기 3진 입력은 상기 제 1 동작으로 수행될 수 있다.

[0011] 상기 3진 신경망 가속기 소자에 따르면, 상기 매칭 라인은 상기 중간 동작 전압으로 미리 충전될 수 있다.

[0012] 상기 3진 신경망 가속기 소자에 따르면, 상기 제 1 반도체 소자 및 상기 제 2 반도체 소자는 강유전체 전계효과 트랜지스터(FeFET)일 수 있다.

[0013] 상기 3진 신경망 가속기 소자에 따르면, 상기 제 1 동작에서, 상기 하이 문턱전압은 상기 강유전체 전계효과 트랜지스터가 프로그램된 상태이고, 상기 로우 문턱전압은 상기 강유전체 전계효과 트랜지스터가 소거 상태일 수 있다.

[0014] 상기 3진 신경망 가속기 소자에 따르면, 상기 제 1 반도체 소자 및 상기 제 2 반도체 소자는 플래시(flash) 메모리 소자일 수 있다.

[0015] 본 발명의 다른 관점에 따른 3진 신경망 가속기 소자의 동작 방법은, 전술한 3진 신경망 가속기 소자를 이용한 것으로서, 상기 3진 신경망 가속기 소자에 3진 웨이트 및 3진 입력의 조건을 설정하고, 상기 3진 웨이트 및 상기 3진 입력의 조건에 따라서 9개의 연산 결과를 상기 매칭 라인을 통해서 출력하는 3진 동작을 수행하되, 상기 3진 웨이트 및 상기 3진 입력은, 상기 제 1 반도체 소자의 제 1 문턱전압과 상기 제 2 반도체 소자의 제 2 문턱전압을 상대적으로 낮은 로우 문턱전압과 상대적으로 높은 하이 문턱전압으로 조합된 세가지 스테이트 중의 하나로 변경하는 제 1 동작과, 상기 제 1 서칭 라인 및 상기 제 2 서칭 라인에 인가되는 전압을 접지전압, 동작전압 및 중간 동작전압으로 조합된 세가지 조합 중의 하나로 설정하는 제 2 동작 중 서로 다른 어느 하나로 수행한다.

발명의 효과

[0016] 상기한 바와 같이 이루어진 본 발명의 일 실시예에 따른 3진 신경망 가속기 소자 및 그 동작 방법에 의하면, 기존 2진 신경망 가속기와 같은 간단한 구조로 경량 네트워크의 장점을 가지면서 3진 데이터 처리를 통해서 연상 정확성을 높일 수 있다. 물론 이러한 효과는 예시적인 것이고, 이러한 효과에 의해서 본 발명의 범위가 한정되는 것은 아니다.

도면의 간단한 설명

[0017] 도 1은 본 발명의 일 실시예에 따른 3진 신경망 가속기 소자를 보여주는 개략적인 회로도이다.

도 2는 도 1의 3진 신경망 가속기 소자 문턱전압 특성을 보여주는 그래프이다.

도 3은 도 1의 3진 신경망 가속기 소자 내 반도체 소자의 소거 및 프로그램 동작 조건을 보여주는 데이터이다.

도 4는 도 1의 3진 신경망 가속기 소자 내 반도체 소자의 스테이트에 따른 문턱전압의 변화를 보여주는 V-I 그래프이다.

도 5a 내지 도 5c 및 도 6은 도 1의 3진 신경망 가속기 소자의 일 실시예에 따른 3진 동작 특성을 보여주는 데이터들이다.

도 7a 내지 도 7c는 도 1의 3진 신경망 가속기 소자의 3진 동작을 보여주는 회로도들이다.

도 8a 내지 도 8c는 도 1의 3진 신경망 가속기 소자의 다른 실시예에 따른 3진 동작 특성을 보여주는 데이터들이다.

도 9는 도 1의 3진 신경망 가속기 소자의 3진 동작 특성을 보여주는 그래프이다.

도 10은 본 발명의 다른 실시예에 따른 3진 가속기 소자를 보여주는 개략적인 회로도이다.

발명을 실시하기 위한 구체적인 내용

[0018] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세히 설명하면 다음과 같다. 그러나 본 발명은 이하에서 개시되는 실시예에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있는 것으로, 이하의 실시예는 본 발명의 개시가 완전하도록 하며, 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이다. 또한 설명의 편의를 위하여 도면에서는 적어도 일부의 구성 요소들이 그 크기가 과장 또는 축소될 수 있다. 도면에서 동일한 부호는 동일한 요소를 지칭한다.

[0019] 다르게 정의되지 않는 한, 여기에 사용된 모든 용어들은 해당기술 분야에서 통상의 지식을 가진 자에 의해서 통상적으로 이해되는 것과 같은 의미로 사용된다. 도면에서, 층 및 영역의 크기는 설명을 위해 과장되었고, 따라서 본 발명의 일반적인 구조들을 설명하기 위해 제공된다.

[0020] 동일한 참조 부호들은 동일한 구성 요소를 나타낸다. 층, 영역, 또는 기판과 같은 한 구성이 다른 구성 상(on)에 있다고 지칭할 때, 그것은 다른 구성의 바로 상부에 있거나 또는 그 사이에 다른 개재된 구성이 또한 존재할 수 있는 것으로 이해될 것이다. 반면에, 한 구성이 다른 구성의 “바로 위에(directly on)” 있다라고 지칭할 때는 중간 개재 구성들이 존재하지 않는다고 이해된다.

[0021] 도 1은 본 발명의 일 실시예에 따른 3진 신경망 가속기 소자(100)를 보여주는 개략적인 회로도이다.

[0022] 도 1을 참조하면, 3진 신경망 가속기 소자(100)는 제 1 반도체 소자(TR1) 및 제 2 반도체 소자(TR2)를 포함할 수 있다. 예를 들어, 제 1 반도체 소자(TR1)는 제 1 소오스 단자(S1), 제 1 드레인 단자(D1) 및 제 1 게이트 단자(G1)를 포함하고, 제 2 반도체 소자(TR2)는 제 2 소오스 단자(S2), 제 2 드레인 단자(D2) 및 제 2 게이트 단자(G2)를 포함할 수 있다.

[0023] 나아가, 3진 신경망 가속기 소자(100)는 제 1 서칭 라인(SL1), 제 2 서칭 라인(SL2), 워드 라인(WL) 및/또는 매칭 라인(ML)을 더 포함할 수 있다. 예를 들어, 제 1 서칭 라인(SL1)은 제 1 드레인 단자(D1)에 연결되고, 제 2 서칭 라인(SL2)은 제 2 드레인 단자(D2)에 연결되고, 매칭 라인(ML)은 제 1 소오스 단자(S1) 및 제 2 소오스 단자(S2)에 공통으로 연결될 수 있다.

[0024] 제 1 반도체 소자(TR1)는 제 1 문턱전압(V_{th1})을 갖고, 제 2 반도체 소자(TR2)는 제 2 문턱전압(V_{th2})을 가질 수 있다. 도 2에 도시된 바와 같이, 제 1 반도체 소자(TR1)의 제 1 문턱전압(V_{th1}) 및 제 2 반도체 소자(TR2)의 제 2 문턱전압(V_{th2})은 듀얼 스테이트, 예컨대 상대적으로 낮은 로우 문턱전압(L-V_{th}) 상대적으로 높은 하이 문턱 전압(H-V_{th})을 가질 수 있다.

[0025] 여기에서, 로우 문턱전압(L-V_{th})과 하이 문턱 전압(H-V_{th})은 둘 사이에서 상대적으로 그 크기를 비교하여 그 명칭이 부여될 수 있다. 예를 들어, 제 1 반도체 소자(TR1) 및 제 2 반도체 소자(TR2) 각각은 서로 다른 문턱전압을 갖는 적어도 두 스테이트를 포함할 수 있고, 로우 문턱전압(L-V_{th})은 그 중에서 낮은 문턱전압을 지칭하고 하

이 문턱 전압(H-V_{th})은 그 중에서 상대적으로 높은 문턱전압을 지칭할 수 있다.

[0026] 제 1 반도체 소자(TR1) 및 제 2 반도체 소자(TR2)는 데이터 저장이 가능한 메모리 소자일 수 있다. 예를 들어, 제 1 반도체 소자(TR1) 및 제 2 반도체 소자(TR2)는 강유전체 전계효과 트랜지스터(ferroelectric field effect transistor, FeFET) 또는 플래시(flash) 메모리 소자를 포함할 수 있다.

[0027] 일부 실시예에서, 하이 문턱전압(H-V_{th})은 강유전체 전계효과 트랜지스터(FeFET) 또는 플래시 메모리 소자가 프로그램된 상태에서 문턱전압이고, 로우 문턱전압(L-V_{th})은 강유전체 전계효과 트랜지스터(FeFET) 또는 플래시 메모리 소자가 소거 상태에서 문턱전압일 수 있다.

[0028] 이하에서는 3진 신경망 가속기 소자(100)에서 프로그램 상태 및 소거 상태의 조건을 예시적으로 설명한다.

[0029] 도 3은 도 1의 3진 신경망 가속기 소자(100) 내 반도체 소자의 소거 및 프로그램 동작 조건을 보여주는 데이터이고, 도 4는 도 1의 3진 신경망 가속기 소자(100) 내 반도체 소자의 스테이트에 따른 문턱전압의 변화를 보여주는 V-I 그래프이다.

[0030] 도 3을 참조하면, 제 1 반도체 소자(TR1)의 프로그램(program) 동작을 위해서, 제 1 서칭 라인(SL1)은 접지(GND)에 연결되고, 제 2 서칭 라인(SL2)에는 프로그램 방지 전압(V_{INH})이 인가되고, 워드 라인(WL)에는 프로그램 전압(V_{PRG})이 인가되고, 매칭 라인은 접지(GND)에 연결될 수 있다. 예를 들어, 프로그램 방지 전압(V_{INH}) 및 프로그램 전압(V_{PRG})은 하나의 동작 전압(V_{DD})으로 동일할 수 있다. 이에 따르면, 제 1 반도체 소자(TR1)의 프로그램(program) 동작 시, 제 2 반도체 소자(TR2)의 프로그램 동작은 방지될 수 있다.

[0031] 한편, 제 1 반도체 소자(TR1)의 소거(erase) 동작을 위해서, 제 1 서칭 라인(SL1)에는 소거 전압(V_{ER})이 인가되고, 제 2 서칭 라인(SL2)에는 소거 방지 전압(V_{ER})이 인가되고, 워드 라인(WL) 및 매칭 라인은 접지(GND)에 연결될 수 있다. 예를 들어, 소거 방지 전압(V_{ER})은 접지 전압일 수 있고, 소거 전압(V_{ER})은 동작 전압(V_{DD})일 수 있다. 이에 따르면, 제 2 반도체 소자(TR2)의 소거 동작 시, 제 1 반도체 소자(TR1)의 소거 동작은 방지될 수 있다.

[0032] 도 4를 참조하면, 제 1 반도체 소자(TR1) 또는 제 2 반도체 소자(TR2)의 데이터 스테이트에 따라서 문턱전압이 달라지는 것을 알 수 있다.

[0033] 이하에서는 3진 신경망 가속기 소자(100)를 이용한 3진 동작에 대해서 설명한다.

[0034] 3진 신경망 가속기 소자(100)의 3진 동작에 따르면, 3진 웨이트 및 3진 입력의 조건을 설정하고, 3진 웨이트 및 3진 입력의 조건에 따라서 9개의 연산 결과가 매칭 라인(ML)을 통해서 출력될 수 있다.

[0035] 예를 들어, 3진 처리를 위한 3진 웨이트 및 3진 입력은, 제 1 반도체 소자(TR1)의 제 1 문턱전압(V_{th1})과 제 2 반도체 소자(TR2)의 제 2 문턱전압(V_{th2})을 상대적으로 낮은 로우 문턱전압(L-V_{th})과 상대적으로 높은 하이 문턱전압(H-V_{th})으로 조합된 세가지 스테이트 중의 하나로 변경하는 제 1 동작과, 제 1 서칭 라인(SL1) 및 제 2 서칭 라인(SL2)에 인가되는 전압을 접지전압(GND), 동작전압(V_{DD}) 및 중간 동작전압(V_{DD}/2)으로 조합된 세가지 조합 중의 하나로 설정하는 제 2 동작 중 서로 다른 어느 하나로 수행될 수 있다.

[0036] 일부 실시예들에서, 제 1 동작에서 세가지 스테이트는, 제 1 문턱전압(V_{th1})과 제 2 문턱전압(V_{th2})이 모두 하이 문턱전압(H-V_{th})인 제 1 스테이트와, 제 1 문턱전압(V_{th1})이 하이 문턱전압(H-V_{th})이고 제 2 문턱전압(V_{th2})이 로우 문턱전압(L-V_{th})인 제 2 스테이트와, 제 1 문턱전압(V_{th1})이 로우 문턱전압(L-V_{th})이고 제 2 문턱전압(V_{th2})이 하이 문턱전압(H-V_{th})인 제 3 스테이트를 포함할 수 있다.

[0037] 일부 실시예들에서, 제 2 동작에서 세가지 조합은, 제 1 서칭 라인(SL1)과 제 2 서칭 라인(SL2)에 모두 중간 동작전압(V_{DD}/2)이 인가되는 제 1 조합과, 제 1 서칭 라인(SL1)에 접지 전압 GND)이 인가되고, 제 2 서칭 라인(SL2)에 동작 전압(V_{DD})이 인가되는 제 2 조합과, 제 1 서칭 라인(SL1)에 동작 전압(V_{DD})이 인가되고, 제 2 서칭 라인(SL2)에 접지 전압 GND)이 인가되는 제 3 조합을 포함할 수 있다.

[0038] 일부 실시예들에서, 매칭 라인(ML)은 중간 동작전압(V_{DD}/2)으로 미리 충전될 수 있다. 따라서, 제 1 반도체 소자

(TR1) 및 제 2 반도체 소자(TR2)가 모두 턴-오프되는(turn-off) 경우, 매칭 라인(ML)은 미리 충전된 중간 동작 전압($V_{DD}/2$)을 그대로 출력할 수 있다. 반면, 제 1 반도체 소자(TR1) 또는 제 2 반도체 소자(TR2)가 모두 턴-온되는(turn-on) 경우 충전 또는 방전 전류가 흐르면서 제 1 서칭 라인(SL1) 또는 제 2 서칭 라인(SL2)의 전압이 출력될 수 있다.

[0039] 먼저, 3진 웨이트가 제 1 동작으로 수행되고, 3진 입력이 제 2 동작으로 수행되는 실시예에 대해서 설명한다.

[0040] 도 5a 내지 도 5c 및 도 6은 도 1의 3진 신경망 가속기 소자(100)의 일 실시예에 따른 3진 동작 특성을 보여주는 데이터들이다.

[0041] 도 5a를 참조하면, 3진 웨이트는 제 1 반도체 소자(TR1) 및 제 2 반도체 소자(TR2)의 문턱전압을 조절하여 설정될 수 있다. 예를 들어, 3진 웨이트 설정을 위한 제 1 동작 시 세가지 스테이트는 제 1 스테이트(Weight=0), 제 2 스테이트(Weight=1) 및 제 3 스테이트(Weight=-1)를 포함할 수 있다. 제 1 반도체 소자(TR1) 및 제 2 반도체 소자(TR2)의 문턱전압의 조절은 사전 설정이 필요하지만, 3진 웨이트는 자주 변경되지 않는다는 점에서 한번 설정으로 오래 유지될 수 있다.

[0042] 도 5b를 참조하면 3진 입력은 제 1 서칭 라인(SL1) 및 제 2 서칭 라인(SL2)에 인가되는 전압을 조합하여 설정될 수 있다. 예를 들어, 세가지 조합은 제 1 조합(iFMAP=0), 제 2 조합(iFMAP=1) 및 제 3 조합(iFMAP=-1)을 포함할 수 있다.

[0043] 도 5c를 참조하면, 매칭 라인(ML)을 통한 출력은 세가지 상태를 나타낼 수 있다. 예를 들어, 제 1 상태(Act=0)는 중간 동작전압($V_{DD}/2$)을 나타내고, 제 2 값(Act=1)은 충전(charging) 상태를 나타내고, 제 3 값(Act=-1)은 방전(discharging) 상태를 나타낼 수 있다.

[0044] 도 6을 참조하면, 3진 웨이트(Weight) 및 3진 입력(iFMAP)의 조건에 따라서 9개의 연산 결과가 출력될 수 있다. 출력은 크게 매칭 라인(ML)이 충전 상태를 나태는 매치(match) 상태, 매칭 라인(ML)이 방전 상태를 나태는 미스매치(mismatch) 상태 및 매칭 라인(ML)이 중간 동작전압($V_{DD}/2$)을 갖는 제로(0) 상태를 가질 수 있다.

[0045] 도 7a 내지 도 7c는 도 1의 3진 신경망 가속기 소자의 3진 동작을 보여주는 회로도들이다.

[0046] 도 7a를 참조하면, 3진 웨이트(Weight)가 제 1 스테이트(Weight=0)이거나(case a-1) 또는 3진 입력(iFMAP)이 제 1 조합(iFMAP=0)인 경우(case a-2), 매칭 라인(ML)이 중간 동작전압($V_{DD}/2$)을 갖는 제로(0) 상태가 출력될 수 있다.

[0047] 도 7b를 참조하면, 3진 웨이트(Weight)가 제 2 스테이트(Weight=1)이고 3진 입력(iFMAP)이 제 2 조합(iFMAP=1)인 경우(case b-1) 또는 3진 웨이트(Weight)가 제 3 스테이트(Weight=-1)이고 3진 입력(iFMAP)이 제 3 조합(iFMAP=-1)인 경우(case b-2), 매칭 라인(ML)이 충전(charging) 상태를 나타내는 매치(match) 상태가 출력될 수 있다.

[0048] 도 7c를 참조하면, 3진 웨이트(Weight)가 제 3 스테이트(Weight=-1)이고 3진 입력(iFMAP)이 제 2 조합(iFMAP=1)인 경우(case c-1) 또는 3진 웨이트(Weight)가 제 2 스테이트(Weight=1)이고 3진 입력(iFMAP)이 제 3 조합(iFMAP=-1)인 경우(case c-2), 매칭 라인(ML)이 방전(discharging) 상태를 나타내는 미스매치(mismatch) 상태가 출력될 수 있다.

[0049] 한편, 일부 실시예들에서, 3진 웨이트가 제 2 동작으로 수행되고, 3진 입력이 제 1 동작으로 수행될 수도 있다.

[0050] 도 8a 내지 도 8c는 도 1의 3진 신경망 가속기 소자(100)의 다른 실시예에 따른 3진 동작 특성을 보여주는 데이터들이다.

[0051] 도 8a를 참조하면, 3진 입력은 제 1 반도체 소자(TR1) 및 제 2 반도체 소자(TR2)의 문턱전압을 조절하여 설정될 수 있다. 예를 들어, 3진 입력 설정을 위한 제 1 동작 시 세가지 스테이트는 제 1 스테이트(iFMAP=0), 제 2 스테이트(iFMAP=1) 및 제 3 스테이트(iFMAP=-1)를 포함할 수 있다.

[0052] 도 8b를 참조하면 3진 웨이트는 제 1 서칭 라인(SL1) 및 제 2 서칭 라인(SL2)에 인가되는 전압을 조합하여 설정될 수 있다. 예를 들어, 세가지 조합은 제 1 조합(Weight=0), 제 2 조합(Weight=1) 및 제 3 조합(Weight=-1)을 포함할 수 있다.

[0053] 도 8c를 참조하면, 매칭 라인(ML)을 통한 출력은 세가지 상태를 나타낼 수 있다. 예를 들어, 제 1 상태(Act=0)

는 중간 동작전압($V_{DD}/2$)을 나타내고, 제 2 값(Act=1)은 충전(charging) 상태를 나타내고, 제 3 값(Act=-1)은 방전(discharging) 상태를 나타낼 수 있다.

[0054] 이 실시예에서도, 도 6에서 설명한 바와 마찬가지로, 3진 입력과 3진 웨이트를 조합하여 9개의 연산 결과가 출력될 수 있다.

[0055] 도 9는 도 1의 3진 신경망 가속기 소자(100)의 3진 동작 특성을 보여주는 그래프이다.

[0056] 도 9를 참조하면, 매칭 라인(ML)을 미리 중간 동작전압($V_{DD}/2$)으로 충전한(ML precharge) 후, 제 1 및 제 2 서칭 라인들을 활성화하여, 매칭 라인(ML)이 중간 동작전압($V_{DD}/2$)을 갖는 제로(0) 상태, 매칭 라인(ML)이 충전(charging) 상태를 나타내는 매치(match) 상태 및 매칭 라인(ML)이 방전(discharging) 상태를 나타내는 미스매치(mismatch) 상태가 감지될 수 있다.

[0057] 도 10은 본 발명의 다른 실시예에 따른 3진 가속기 소자(200)를 보여주는 개략적인 회로도이다.

[0058] 도 10을 참조하면, 3진 가속기 소자(200)는 도 1의 3진 가속기 소자(100)를 단위셀로 하여, 이러한 단위셀들을 4X4 어레이 형태로 배열된 2차원 구조를 포함할 수 있다. 예를 들어, 3진 가속기 소자(200) 내에서, 복수의 서칭 라인들(SL1 ~ SL8)이 복수의 열로 배치되고, 복수의 매칭 라인들(ML1 ~ ML4)와 복수의 워드 라인들(WL1 ~ WL4)이 복수의 행으로 배치되고, 각 단위셀은 복수의 서칭 라인들(SL1 ~ SL8)의 인접한 둘 사이, 복수의 매칭 라인들(ML1 ~ ML4) 중 하나 및 복수의 워드 라인들(WL1 ~ WL4) 중 하나와 각각 연결될 수 있다. 3진 가속기 소자(200)의 동작은 전술한 3진 가속기 소자(100)의 동작을 참조하여 이해될 수 있다.

[0059] 3진 가속기 소자(200) 내 단위셀들의 수는 적절하게 선택될 수 있고 이 실시예의 범위를 제한하지 않는다.

[0060] 전술한 3진 가속기 소자들(100, 200)에 따르면, 기존 2진 신경망 가속기와 같은 간단한 구조로 경량 네트워크의 장점을 가지면서 3진 데이터 처리를 통해서 연상 정확성을 높일 수 있다.

[0061] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 당해 기술분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 다른 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 진정한 기술적 보호 범위는 첨부된 특허청구범위의 기술적 사상에 의하여 정해져야 할 것이다.

부호의 설명

[0062] 100, 200: 3진 신경망 가속기 소자

TR1, TR2: 반도체 소자

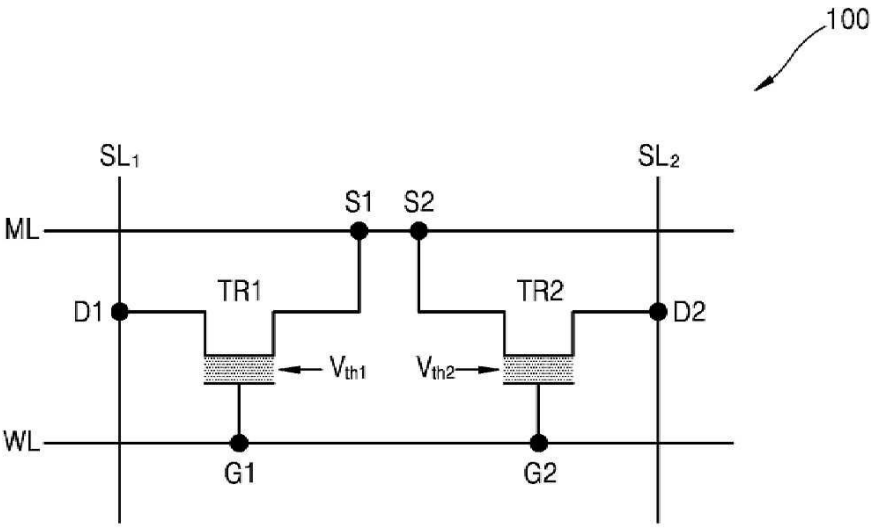
SL1, SL2: 서칭 라인

ML: 매칭 라인

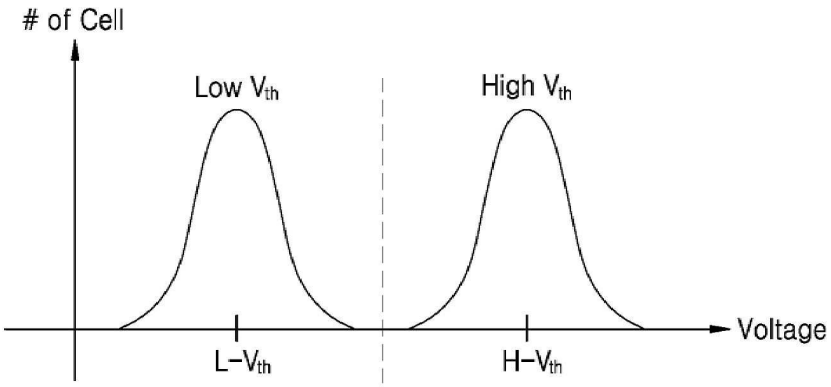
WL: 워드 라인

도면

도면1



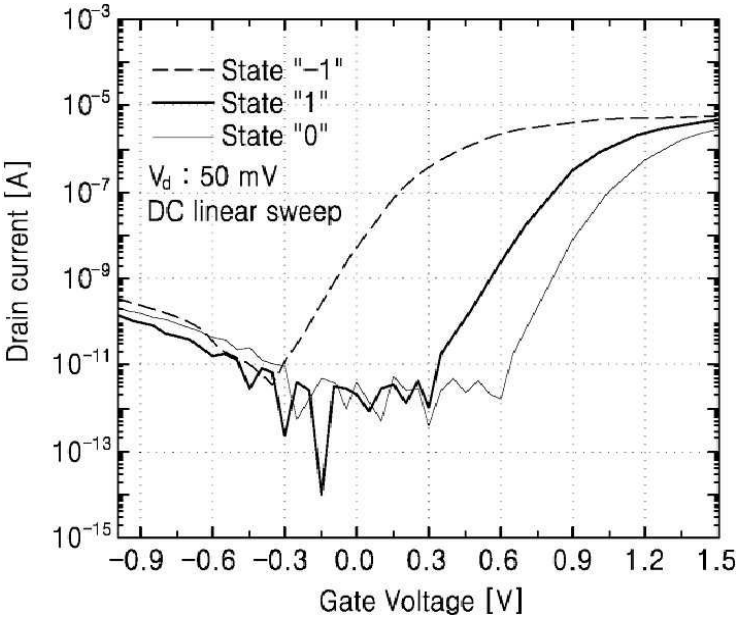
도면2



도면3

	Erase	Program
SL_1	V_{ER} (erase)	GND (program)
SL_2	GND (inhibit)	V_{INH} (inhibit)
WL	GND	V_{PRG}
ML	GND	GND

도면4



도면5a

	Weight = 0	Weight = 1	Weight = −1
V_{th1}	H−V_{th}	H−V_{th}	L−V_{th}
V_{th2}	H−V_{th}	L−V_{th}	H−V_{th}

도면5b

	iFMAP=0	iFMAP=1	iFMAP=−1
SL_1	V_{DD}/2	GND	V_{DD}
SL_2	V_{DD}/2	V_{DD}	GND

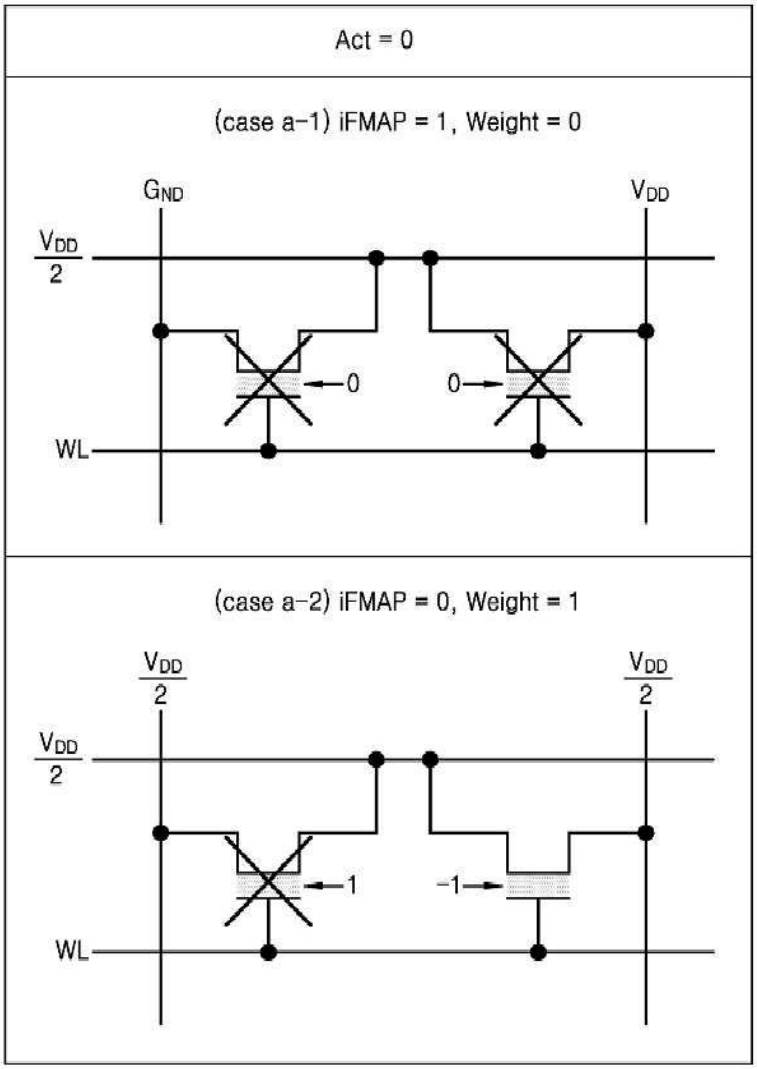
도면5c

	Act=0	Act=1	Act=−1
ML	V_{DD}/2	Charging	Discharging

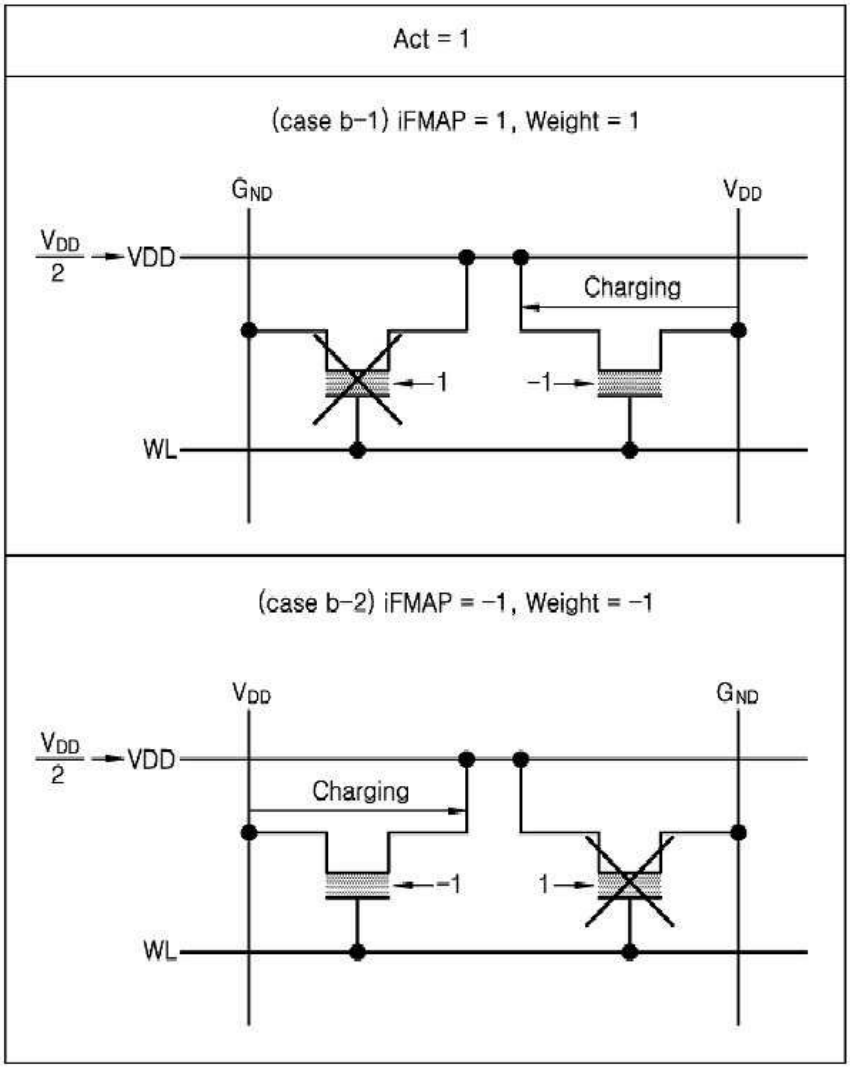
도면6

iFMAP	Weight	Calculation	Status	ML
−1	−1	1	Match	Charging
−1	0	0	0	$V_{DD}/2$
−1	1	−1	Mismatch	Discharging
0	−1	0	0	$V_{DD}/2$
0	0	0	0	$V_{DD}/2$
0	1	0	0	$V_{DD}/2$
1	−1	−1	Mismatch	Discharging
1	0	0	0	$V_{DD}/2$
1	1	1	Match	Charging

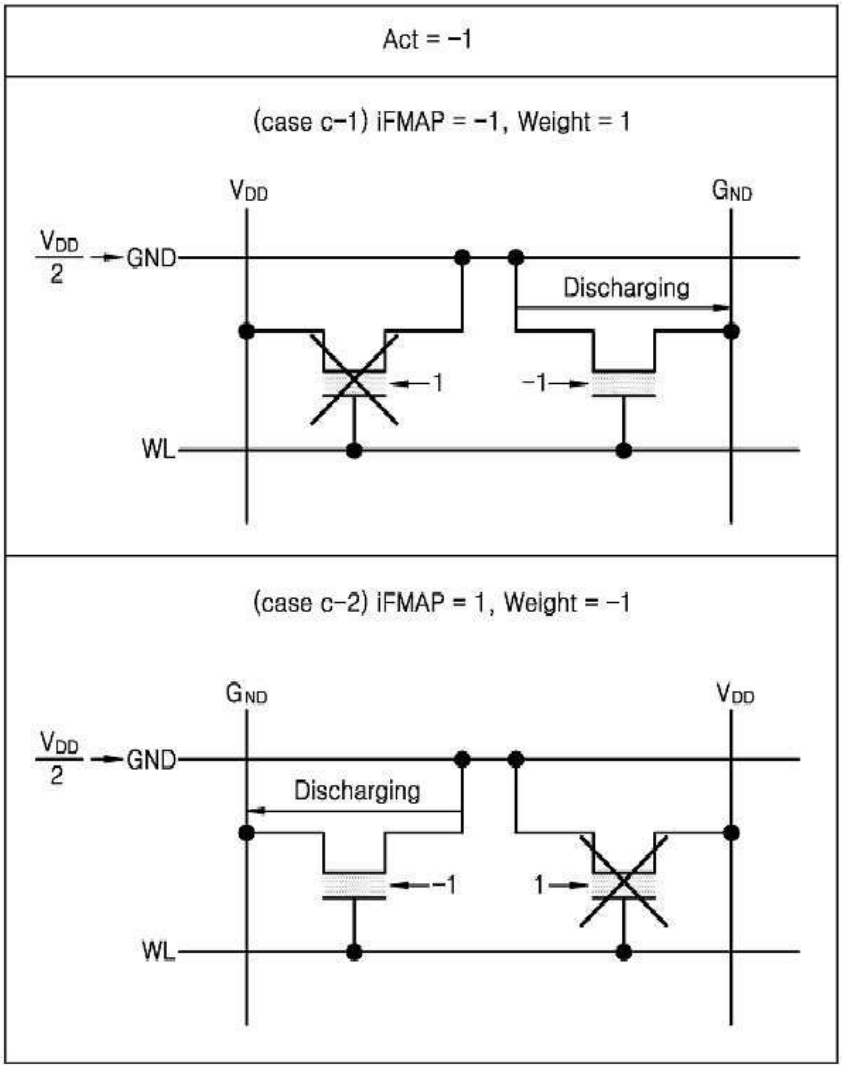
도면7a



도면7b


Act = 1
(case b-1) iFMAP = 1, Weight = 1
GND
VDD
VDD/2
VDD
Charging
1
-1
WL
(case b-2) iFMAP = -1, Weight = -1
VDD
GND
VDD/2
VDD
Charging
-1
1
WL

도면7c



도면8a

	iFMAP=0	iFMAP=1	iFMAP=−1
V_{th1}	H−V_{th}	H−V_{th}	L−V_{th}
V_{th2}	H−V_{th}	L−V_{th}	H−V_{th}

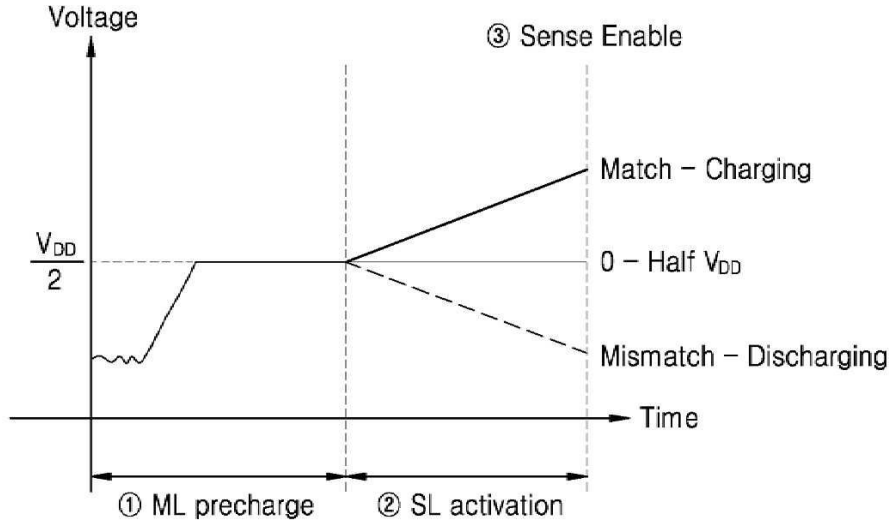
도면8b

	Weight = 0	Weight = 1	Weight = −1
SL_1	$V_{DD}/2$	GND	V_{DD}
SL_2	$V_{DD}/2$	V_{DD}	GND

도면8c

	Act=0	Act=1	Act=−1
ML	$V_{DD}/2$	Charging	Discharging

도면9



도면10

