



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0168681
(43) 공개일자 2022년12월26일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01) G06N 3/04 (2006.01)
H01L 27/11563 (2017.01) H01L 29/792 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2021-0078468
(22) 출원일자 2021년06월17일
심사청구일자 2021년06월17일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
최우영
서울특별시 용산구 원효로 216, 109동 1303호(신계동, 용산e편한세상)
이장우
경기도 안양시 동안구 부림로 55 초원대원아파트 304동 403호
우재승
서울특별시 서대문구 연희로41길 114, 3동 102호(홍은동, 미성아파트)
(74) 대리인
정부연

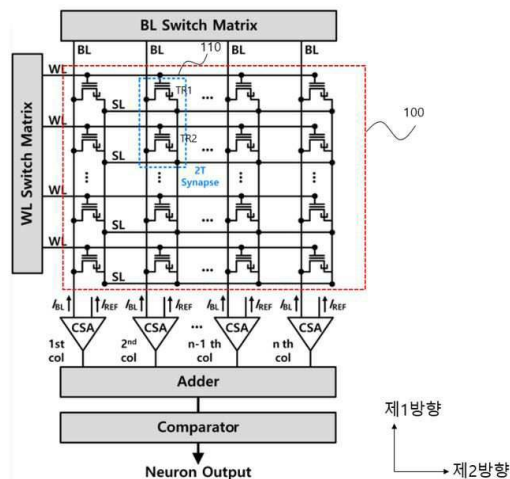
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 인공신경망 연산 장치

(57) 요약

본 발명은 인공연산망 장치에 관한 것으로, 제1 방향 및 상기 제1 방향과 수직한 제2 방향을 따라 일정간격 이격되어 배치된 시냅스 어레이를 포함하는 이진 신경망 아키텍처에 있어서, 상기 시냅스 어레이는 다수의 시냅스 소자를 포함하며, 각각의 시냅스 소자는 상기 제1 방향을 따라 인접하여 배치된 제1터널링 전계 효과 트랜지스터(TFET) 및 제2 터널링 전계 효과 트랜지스터(TFET)들을 포함하고, 상기 제1 터널링 전계 효과 트랜지스터 및 상기 제2 터널링 전계 효과 트랜지스터는 각각 게이트 전극, 드레인 단자 및 소스 단자를 포함하는 것을 특징으로 한다.

대표도 - 도1



(52) CPC특허분류

H01L 27/11563 (2013.01)

H01L 29/792 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711133834
과제번호	10080575
부처명	과학기술정보통신부
과제관리(전문)기관명	한국산업기술평가관리원
연구사업명	전자정보디바이스산업원천기술개발(R&D, 정보화)
연구과제명	초저전력 로직응용을 위한 SiGe 다중 적층 채널 GAA TFET 개발
기 여 율	1/4
과제수행기관명	아주대학교산학협력단
연구기간	2021.01.01 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711116161
과제번호	2018-0-01421-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	인공지능 서비스 실현을 위한 지능형 반도체 설계 핵심기술 개발
기 여 율	1/4
과제수행기관명	서강대학교산학협력단
연구기간	2021.01.01 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711134833
과제번호	2021M3F3A2A01037927
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	차세대지능형반도체기술개발(R&D)
연구과제명	SNN 성능 향상을 위한 시냅스 소자와 신소자 기반 뉴런 회로
기 여 율	1/4
과제수행기관명	서울대학교
연구기간	2021.04.09 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711127062
과제번호	2021R1A2C1007931
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	초저전력 비휘발성 메모리 소자의 모노리식 삼차원 집적을 이용한 두뇌모방 연산 기술 개발
기 여 율	1/4
과제수행기관명	서강대학교
연구기간	2021.03.01 ~ 2022.02.28

명세서

청구범위

청구항 1

제1 방향 및 상기 제1 방향과 수직한 제2 방향을 따라 일정간격 이격되어 배치된 시냅스 어레이를 포함하는 이진 신경망 아키텍처에 있어서,

상기 시냅스 어레이는 다수의 시냅스 소자를 포함하며,

각각의 시냅스 소자는 상기 제1 방향을 따라 인접하여 배치된 제1터널링 전계 효과 트랜지스터(TFET) 및 제2 터널링 전계 효과 트랜지스터(TFET)들을 포함하고, 상기 제1 터널링 전계 효과 트랜지스터 및 상기 제2 터널링 전계 효과 트랜지스터는 각각 게이트 전극, 드레인 단자 및 소스 단자를 포함하는 것을 특징으로 하는 인공연산망 장치.

청구항 2

제1항에 있어서, 상기 이진 신경망 아키텍처는

상기 제1 방향을 따라 연장되어 배치된 복수의 비트라인; 및

상기 제2 방향을 따라 연장되어 배치된 복수의 워드라인을 포함하며, 상기 시냅스 소자는 상기 비트라인과 워드라인이 교차하는 영역에 위치된 것을 특징으로 하는 인공연산망 장치.

청구항 3

제2항에 있어서,

상기 제1 터널링 전계 효과 트랜지스터 및 상기 제2 터널링 전계 효과 트랜지스터의 상기 게이트는 상기 워드라인과 접속하는 것을 특징으로 하는 인공신경망 연산 장치.

청구항 4

제2항에 있어서,

상기 제1 터널링 전계 효과 트랜지스터 및 상기 제2 터널링 전계 효과 트랜지스터의 상기 드레인 단자는 상기 비트라인과 연결된 것을 특징으로 하는 인공신경망 연산 장치.

청구항 5

제2항에 있어서,

상기 제1 터널링 전계 효과 트랜지스터 및 상기 제2 터널링 전계 효과 트랜지스터의 상기 소스 단자는 메탈 라인과 접속되어 그라운드(GND)에 연결된 것을 특징으로 하는 인공신경망 연산 장치.

청구항 6

제1항에 있어서, 상기 제1 터널링 전계 효과 트랜지스터 및 제2 터널링 전계 효과 트랜지스터는

반도체 기판 상부에 형성된 게이트 패턴;

상기 게이트 패턴 하부에 위치한 채널 영역; 및

상기 게이트 패턴 양측의 상기 반도체 기판 내에 형성된 소스 영역 및 드레인 영역을 포함하며, 상기 게이트 패턴은 하부 산화막, 질화막, 상부 산화막 및 게이트 전극층의 적층 구조로 형성된 것을 특징으로 하는 인공신경망 연산 장치.

청구항 7

제6항에 있어서,

상기 제1 터널링 전계 효과 트랜지스터 및 제2 터널링 전계 효과 트랜지스터의 상기 드레인 영역은 콘택 플러그를 통해 비트라인과 연결되고, 상기 소스 영역은 콘택 플러그를 통해 메탈 라인과 연결되며, 상기 게이트 전극층은 콘택 플러그를 통해 워드라인과 연결되어 시냅스 소자를 구성하는 것을 특징으로 하는 인공신경망 연산 장치.

청구항 8

제1항에 있어서,

상기 시냅스 소자로 활용되는 터널링 전계 효과 트랜지스터는 프로그램(program) 및 이레이즈(erase) 동작을 수행하며, XNOR 동작 구현을 통해 이진 신경망 연산을 수행하는 것을 특징으로 하는 인공신경망 연산 장치.

청구항 9

제1항에 있어서,

상기 각각의 시냅스 소자들의 가중치(Weight) 값을 개별적으로 제어하여 입력(Input) 값에 따른 출력(Output) 값을 나타내는 것을 특징으로 하는 인공신경망 연산 장치.

청구항 10

제1항에 있어서, 상기 시냅스 어레이는

비트라인에 흐르는 전류 값(I_{BL})과 기준 전류 값(I_{REF})에 따라 출력 값이 결정하는 CSA 블록(Conditional sum adder); 및

각 시냅스 소자에 대한 비교 결과값들을 순차적으로 입력 받아 처리하고, 최종 결과값을 출력하는 가산기(Adder)와 비교회로(Comparator)를 더 포함하는 것을 특징으로 하는 인공신경망 연산 장치.

발명의 설명

기술 분야

[0001] 본 발명은 인공신경망 연산 장치에 관한 것으로, 보다 상세하게는 터널링 전계 효과 트랜지스터(TFET)를 이용한 초저전력 시냅스 소자를 기반으로 하는 인공신경망 연산 장치에 관한 것이다.

배경 기술

[0003] 무어의 법칙이 종말을 맞이하고 기존의 von Neumann 구조의 한계가 명확해짐에 따라 이를 극복하고자 인공 신경망 (인공지능) 및 이를 구현하기 위한 시냅스 소자에 대한 연구가 활발하게 이루어지고 있다. 기존 von Neumann 구조의 경우 비록 CPU를 통한 고속 연산이 가능하지만 근본적으로 메모리와 CPU가 분리되어 있기에 병렬 연산이

어려운 단점이 존재하며, 이는 향후 big data를 처리하는데 있어서 장애물이 될 것으로 예상된다. 그러나 신경망 구조의 경우에는 개별 시냅스 소자의 속도는 기존 CPU의 연산 처리 속도에 훨씬 미치지 못하지만 병렬 연산이 가능하다는 장점이 있으며, 이에 하드웨어 기반 인공 신경망 연산 체계는 neuromorphic computing system이라 불리며, 기존 반도체 소자 기술을 이용하여 하드웨어가 인간의 신경계인 synapse, neuron, axon 등을 모방하고자 하는 새로운 개념으로 제안되고 있다.

[0004] 하드웨어 기반의 시냅스 소자를 통해 학습을 진행하는 방식에는 크게 on-chip learning과 off-chip learning의 두 가지 방식이 존재하며, 전자는 생물학적 학습 특성을 그대로 모방하는 방식으로 학습의 전반적인 과정이 각각의 소자 내부에서 동시 다발적으로 일어나며 주로 STDP 방식을 차용한다. 그러나 그 학습 방법이 완벽하다 할지라도 학습 횟수가 이를 구성하는 시냅스 소자의 endurance 특성에 의해 크게 제한된다는 명확한 단점이 존재한다. 반면, 후자는 기존 von Neumann 기반의 소프트웨어 인공지능 알고리즘을 통해 미리 학습을 진행하고 그 결과 (synaptic weight)를 하드웨어에 이식하는 형태로, 이때 하드웨어 상에 이식된 결과에 대해 주로 추론만을 진행하므로 전자의 방식과는 달리 소자의 신뢰성에 의한 영향을 크게 줄일 수 있다는 장점이 있다. 이는 기존 소프트웨어와 하드웨어 기반의 인공 신경망 형성 기술의 장점을 결합한 형태로 주로 edge computing 구현을 목적으로 하여 활발하게 연구되고 있다.

[0005] 기존 하드웨어 기반 인공지능 연구에서는 생물학적 시냅스 특성을 구현하기 위해 저항성 메모리 (RRAM)을 비롯한 MRAM, PCRAM 등의 다양한 메모리 소자들에 대한 연구가 활발히 진행 중이나, 이와 같은 memristor라 불리는 메모리 소자들은 모두 2단자 형태로 구성되어 있으므로 이에 대한 제어(read/write)를 위한 회로들이 반드시 필요하며, 또한 소자의 신뢰성이 지극히 낮다는 단점이 있다. 그리고 memristor를 형성하기 위해 사용되는 대부분의 물질들은 기존 반도체 공정과 호환되지 않는다는 큰 단점이 존재하여 이에 기존 MOSFET 기반의 flash memory를 이용한 시냅스 소자 연구 또한 진행되고 있는 상태이다.

[0006] 한국등록특허 제10-1950786호는 분산처리용 인공신경망 연산 가속화 방법에 관한 것으로, 단일의 호스트 가속화 장치와 복수개의 슬레이브 가속화 장치를 포함하여 입력 뉴런들에 대한 입력데이터가 M개의 깊이와 N개의 계층으로 구성되는 인공신경망 처리를 가속화하는 방법에 있어서, 인공신경망을 구성하는 뉴런들에 대한 입력데이터와 시냅스 가중치를 입력데이터를 구성하는 깊이, 계층 구조, 신경망 네트워크 또는 이들의 조합 단위로 나누어 각각의 가속화장치에서 연산이 가능하게 하는 것을 특징으로 한다.

[0007] 한국 공개특허 제10-2019-0007143호는 뉴로모픽 연산 장치에 관한 것으로, 펄스폭 변조기, 시냅스 어레이, 발진기, 카운터, 및 보상기를 포함한다. 펄스폭 변조기는 입력 데이터 값에 근거하여 펄스폭 변조 신호를 생성한다. 시냅스 어레이는 펄스폭 변조 신호 및 가중치의 곱셈 연산에 근거하여 시냅스 연산 신호를 생성한다. 발진기는 시냅스 연산 신호의 크기에 의존하는 출력 주파수를 갖는 발진 신호를 생성한다. 카운터는 발진 신호에 포함된 펄스의 개수를 카운팅하여 카운팅 데이터를 생성한다. 보상기는 선형적으로 증가하는 값을 갖는 기준 데이터가 저장된 메모리를 포함한다. 보상기는 기준 데이터와 카운팅 데이터를 비교하여 카운팅 데이터 값을 보상하는 것을 특징으로 한다.

선행기술문헌

특허문헌

- [0009] (특허문헌 0001) 한국등록특허 제10-1950786호(2019.02.15)
(특허문헌 0002) 한국 공개특허 제10-2019-0007143호(2019.01.22)
(특허문헌 0003)

발명의 내용

해결하려는 과제

[0010] 본 발명의 일 실시예는 CMOS와 공정 호환성이 매우 높은 TFET를 이용한 초저전력 시냅스 소자를 제안함으로써, 하드웨어 기반의 인공 신경망 형성 시 전력 소모를 절감하고 신뢰성을 향상시키는 인공신경망 연산 장치를 제공

하고자 한다.

- [0011] 본 발명의 일 실시예는 TFET를 이용한 초저전력 시냅스 소자를 제안함으로써, 휴대용 전자기기, 사물 인터넷 등과 함께 항공, 우주, 의료 분야 등 높은 에너지 효율이 필요하면서 실시간으로 학습이 필요한 응용분야까지 기술 확장이 가능한 인공신경망 연산 장치를 제공하고자 한다.

과제의 해결 수단

- [0013] 실시예들 중에서, 본 발명은 제1 방향 및 상기 제1 방향과 수직한 제2 방향을 따라 일정간격 이격되어 배치된 시냅스 어레이를 포함하는 이진 신경망 아키텍처에 있어서, 상기 시냅스 어레이는 다수의 시냅스 소자를 포함하며, 각각의 시냅스 소자는 상기 제1 방향을 따라 인접하여 배치된 제1터널링 전계 효과 트랜지스터(TFET) 및 제2 터널링 전계 효과 트랜지스터(TFET)들을 포함하고, 상기 제1 터널링 전계 효과 트랜지스터 및 상기 제2 터널링 전계 효과 트랜지스터는 각각 게이트 전극, 드레인 단자 및 소스 단자를 포함하는 것을 특징으로 한다.
- [0014] 상기 이진 신경망 아키텍처는 상기 제1 방향을 따라 연장되어 배치된 복수의 비트라인과, 상기 제2 방향을 따라 연장되어 배치된 복수의 워드라인을 포함하며, 상기 시냅스 소자는 상기 비트라인과 워드라인이 교차하는 영역에 위치된 것을 특징으로 한다.
- [0015] 상기 제1 터널링 전계 효과 트랜지스터 및 상기 제2 터널링 전계 효과 트랜지스터의 상기 게이트는 상기 워드라인과 접속하는 것을 특징으로 한다.
- [0016] 상기 제1 터널링 전계 효과 트랜지스터 및 상기 제2 터널링 전계 효과 트랜지스터의 상기 드레인 단자는 상기 비트라인과 연결된 것을 특징으로 한다.
- [0017] 상기 제1 터널링 전계 효과 트랜지스터 및 상기 제2 터널링 전계 효과 트랜지스터의 상기 소스 단자는 메탈 라인과 접속되어 그라운드(GND)에 연결된 것을 특징으로 한다.
- [0018] 상기 제1 터널링 전계 효과 트랜지스터 및 제2 터널링 전계 효과 트랜지스터는 반도체 기판 상부에 형성된 게이트 패턴, 상기 게이트 패턴 하부에 위치한 채널 영역, 상기 게이트 패턴 양측의 상기 반도체 기판 내에 형성된 소스 영역 및 드레인 영역을 포함하며, 상기 게이트 패턴은 하부 산화막, 질화막, 상부 산화막 및 게이트 전극층의 적층 구조로 형성된 것을 특징으로 한다.
- [0019] 상기 제1 터널링 전계 효과 트랜지스터 및 제2 터널링 전계 효과 트랜지스터의 상기 드레인 영역은 콘택 플러그를 통해 비트라인과 연결되고, 상기 소스 영역은 콘택 플러그를 통해 메탈 라인과 연결되며, 상기 게이트 전극층은 콘택 플러그를 통해 워드라인과 연결되어 시냅스 소자를 구성하는 것을 특징으로 한다.
- [0020] 상기 시냅스 소자로 활용되는 터널링 전계 효과 트랜지스터는 프로그램(program) 및 이레이즈(erase) 동작을 수행하며, XNOR 동작 구현을 통해 이진 신경망 연산을 수행하는 것을 특징으로 한다.
- [0021] 상기 각각의 시냅스 소자들의 가중치(Weight) 값을 개별적으로 제어하여 입력(Input) 값에 따른 출력(Output) 값을 나타내는 것을 특징으로 한다.
- [0022] 상기 시냅스 어레이는 비트라인에 흐르는 전류 값(I_{BL})과 기준 전류 값(I_{REF})에 따라 출력 값이 결정하는 CSA 블록(Conditional sum adder)과, 각 시냅스 소자에 대한 비교 결과값들을 순차적으로 입력 받아 처리하고, 최종 결과값을 출력하는 가산기(Adder)와 비교회로(Comparator)를 더 포함하는 것을 특징으로 한다.

발명의 효과

- [0024] 개시된 기술은 다음의 효과를 가질 수 있다. 다만, 특정 실시예가 다음의 효과를 전부 포함하여야 한다거나 다음의 효과만을 포함하여야 한다는 의미는 아니므로, 개시된 기술의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.
- [0025] 본 발명의 일 실시예에 따른 인공신경망 연산 장치는 CMOS와 공정 호환성이 매우 높은 TFET를 이용한 초저전력 시냅스 소자를 제안함으로써, 하드웨어 기반의 인공 신경망 형성 시 전력 소모를 절감하고 신뢰성을 향상시키는 효과를 얻을 수 있다.

[0026] 본 발명의 일 실시예에 따른 인공신경망 연산 장치는 TFET를 이용한 초저전력 시냅스 소자를 제안함으로써, 휴대용 전자기기, 사물 인터넷 등과 함께 항공, 우주, 의료 분야 등 높은 에너지 효율이 필요하면서 실시간으로 학습이 필요한 응용분야까지 기술 확장이 가능한 효과를 얻을 수 있다.

도면의 간단한 설명

[0028] 도 1은 본 발명의 일 실시예에 따른 인공신경망 연산 장치의 이진 신경망 아키텍처를 도시한 개념도이다.

도 2는 본 발명의 일 실시예에 따른 인공 신경망 연산 장치의 시냅스 소자를 구성하는 터널링 전계 효과 트랜지스터를 도시한 단면도이다.

도 3은 도 2의 터널링 전계 효과 트랜지스터를 이용한 시냅스 소자를 포함하는 시냅스 어레이를 도시한 사시도이다.

도 4는 본 발명의 일 실시예에 따른 인공 신경망 장치의 시냅스 소자로 활용되는 SONOS TFET에 대한 프로그램, 이레이즈 매커니즘을 설명하기 위한 도면이다.

도 5는 본 발명의 이진 신경망을 이용한 XNOR 동작 구현 방식을 설명하기 위한 개념도이다.

도 6은 본 발명의 일 실시예에 따른 인공 신경망 장치를 구성하는 SONOS TFET의 특성을 나타내는 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0029] 본 발명에 관한 설명은 구조적 내지 기능적 설명을 위한 실시예에 불과하므로, 본 발명의 권리범위는 본문에 설명된 실시예에 의하여 제한되는 것으로 해석되어서는 아니 된다. 즉, 실시예는 다양한 변경이 가능하고 여러 가지 형태를 가질 수 있으므로 본 발명의 권리범위는 기술적 사상을 실현할 수 있는 균등물들을 포함하는 것으로 이해되어야 한다. 또한, 본 발명에서 제시된 목적 또는 효과는 특정 실시예가 이를 전부 포함하여야 한다거나 그러한 효과만을 포함하여야 한다는 의미는 아니므로, 본 발명의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0030] 한편, 본 출원에서 서술되는 용어의 의미는 다음과 같이 이해되어야 할 것이다.

[0031] "제1", "제2" 등의 용어는 하나의 구성요소를 다른 구성요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다.

[0032] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결될 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다고 언급된 때에는 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 한편, 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0033] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함하다" 또는 "가지다" 등의 용어는 실시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0034] 각 단계들에 있어 식별부호(예를 들어, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0035] 여기서 사용되는 모든 용어들은 다르게 정의되지 않는 한, 본 발명이 속하는 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한 이상적이거나 과도하게 형식적인 의미를 지니는 것으로 해석될 수 없다.

[0036] 이하 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 이하 도면상의

동일한 구성 요소에 대하여는 동일한 참조 부호를 사용하고, 동일한 구성 요소에 대해서 중복된 설명은 생략한다.

- [0038] 도 1은 본 발명의 일 실시예에 따른 인공신경망 연산 장치의 이진 신경망 구조를 도시한 개념도이다.
- [0039] 도 1을 참조하면, 인공신경망 연산 장치는 터널링 전계 효과 트랜지스터(TFET)를 이용한 시냅스 소자를 기반으로 하는 이진 신경망(binary neural network) 구조를 포함한다.
- [0040] 이진 신경망 구조는 비트라인 스위치 매트릭스(BL switch matrix), 워드라인 스위치 매트릭스(WL switch matrix), 시냅스 어레이, CSA 블록(Conditional sum adder), 합산기(Adder) 및 비교 회로(Comparator)로 구성된다.
- [0041] 먼저, 비트라인 스위치 매트릭스는 다수의 비트라인(BL)과 연결되며, 워드라인 스위치 매트릭스는 비트라인 스위치 매트릭스와 수직한 방향에 위치하며, 워드라인 스위치 매트릭스에 다수의 워드라인(WL)이 연결된다.
- [0042] 시냅스 어레이(100)는 제1 방향으로 연장된 다수의 비트라인(BL)과 상기 제1 방향과 수직한 제2 방향으로 연장되어 배치된 다수의 워드라인(WL)으로 구성되고, 비트라인과 워드라인이 교차하는 영역마다 시냅스 소자(110)가 위치한다.
- [0043] 시냅스 소자(110)는 적어도 두 개 이상의 트랜지스터를 포함한다. 여기서, 트랜지스터는 터널링 전계 효과 트랜지스터(TFET)로 구성되는 것이 바람직하며, TFET는 NOR 타입으로 설계할 수 있다. 시냅스 소자(110)를 구성하는 두개 이상의 트랜지스터는 제1 방향을 따라 인접하여 배치된다. 여기서는 제1 트랜지스터(TR1) 및 제2 트랜지스터(TR2)로 정의하도록 한다. 제1 트랜지스터(TR1) 및 제2 트랜지스터(TR2)의 게이트는 각각 워드라인과 연결되고, 제1 트랜지스터(TR1) 및 제2 트랜지스터(TR2)의 드레인 단자는 비트라인과 연결된다. 또한, 제1 트랜지스터(TR1) 및 제2 트랜지스터(TR2)의 소스 단자는 메탈라인과 접속되어 그라운드(GND)에 연결된다.
- [0044] CSA 블록(Conditional sum adder)은 비트라인에 흐르는 전류 값(I_{BL})과 기준 전류 값(I_{REF})에 따라 출력 값이 결정된다. 가산기(Adder) 및 비교회로(Comparator)는 각 시냅스 소자(110)에 대한 비교 결과값들을 순차적으로 입력 받아 처리하고, 최종 결과값을 출력한다. 이때 한 column 당 하나의 출력 값을 내는 과정에서 시냅스 소자들에 각기 저장된 가중치(weight)값에 따라 해당 column에 흐르는 전류량이 결정되며 (summation of synaptic current), 일반적으로 가장 많은 전류가 흐르는 column의 결과가 출력 값(neuron output)으로 결정된다. 따라서 다양한 입력 및 출력 값(neuron input/output)을 가지는 정교한 인공 신경망 연산 체계를 구축하기 위해서는 매트릭스의 크기가 가능한 한 커질 수밖에 없으며, 한 시냅스 소자에 흐르는 전류가 지극히 작아야 유리하다. 한편, I_{off} 또한 마찬가지로 인공 신경망 전체 전력 소모량에 영향을 미치므로 가능한 한 아주 작은 값을 가져야 한다. 본 발명에서 사용된 TFET은 I_{off} 가 $\sim 10^{-14}$ $\mu A/\mu m$ 수준의 매우 낮은 값을 가지며 dynamic range는 2,700 정도로 상당히 우수한 특성을 가지므로, TFET 기반의 초저전력 시냅스 소자를 하드웨어 기반의 인공 신경망 연산 장치에 활용함으로써, 전력 소모를 절감하며 동시에 신뢰성이 향상되는 효과를 얻을 수 있다.
- [0046] 도 2는 본 발명의 일 실시예에 따른 인공신경망 연산 장치의 시냅스 소자를 구성하는 터널링 전계 효과 트랜지스터의 구조를 도시한 단면도이다.
- [0047] 도 2를 참조하면, 터널링 전계 효과 트랜지스터는 매몰 산화막(200)을 포함하는 반도체 기판(210) 상부에 게이트 패턴(220)이 형성된다. 반도체 기판(210)은 Si, SiGe, Ge, CNT, graphene, MoS₂ 등을 포함하는 물질로 형성될 수 있다. 게이트 패턴(220)은 하부 산화막(220a), 절화막(220b), 상부 산화막(220c) 및 게이트 전극층(220d)의 적층 구조를 포함한다. 하부 산화막(220a) 및 상부 산화막(220c)은 실리콘 산화막(SiO₂), 스트론튬 산화막(SrO), 알루미늄 산화막(Al₂O₃), 마그네슘 산화막(MgO), 스칸듐 산화막(Sc₂O₃), 가돌리늄 산화막(Gd₂O₃), 이트륨 산화막(Y₂O₃), 사마륨 산화막(Sm₂O₃), 하프늄 산화막(HfO₂), 지르코늄 산화막(ZrO₂), 탄탈 산화막(Ta₂O₅), 바륨 산화막(BaO) 및 비스무스 산화막(Bi₂O₃) 중에서 선택하여 형성할 수 있다. 게이트 패턴(220)은 산화막, 절화막 및 산화막이 적층된 ONO 구조를 적용하며, 게이트 패턴(220) 하부의 반도체 기판(200)은 실리콘 기판(Silicon) 및 매몰 산화막(Buried OXide)을 포함하는 구조이므로, 본 발명의 일 실시예에 따른 TFET는 SONOS 구조인 것이 바람직하나 반드시 이에 한정하지는 않는다.

- [0048] 또한, 터널링 전계 효과 트랜지스터는 게이트 패턴(220) 양측의 반도체 기판(210) 상에 소스 영역(230) 및 드레인 영역(240)이 형성된다. 소스 영역(230)은 P+ 영역이고, 드레인 영역 (240)은 N+ 영역으로 형성할 수 있다. 소스 영역(230) 및 드레인 영역 (240)은 확산(Diffusion) 공정 또는 이온 주입(Ion Implantation) 공정을 통해 불순물을 도핑하여 형성할 수 있다.
- [0049] 소스 영역(230) 및 드레인 영역 (240) 사이의 반도체 기판(210) 상에는 채널 영역(250)이 형성된다. 채널 영역 (250)은 소스 영역(230)보다 P형 불순물이 약하게 도핑(P-)되거나, 도핑되지 않은 진성 영역(intrinsic region)으로 형성될 수 있고, 드레인 영역(240)보다 N형 불순물이 약하게 도핑(N- 영역)되거나, 도핑되지 않은 진성 영역으로 형성될 수도 있다.
- [0050] 기타 공정이나 각 단계의 미설명된 부분은 알려진 일반적인 CMOS 공정이나 TFET 공정에 따르면 되므로, 자세한 설명은 생략한다. 상기와 같은 제조 공정은 III-V족 반도체 물질, CNT, 그래핀(graphene), 이황화몰리브덴(MoS2) 등을 기반으로 하는 TFET 소자에도 적용될 수 있다.
- [0052] 도 3은 도 2의 터널링 전계 효과 트랜지스터를 이용한 시냅스 소자를 포함하는 시냅스 어레이를 도시한 사시도이다.
- [0053] 도 3을 참조하면, 시냅스 어레이는 매물 산화막(300)을 포함하는 반도체 기판(310) 상부에 다수의 트랜지스터가 배치되는 구조로 형성된다. 시냅스 어레이를 구성하는 트랜지스터(TR1, TR2)는 도 2에 도시된 바와 같은 터널링 전계 효과 트랜지스터(TFET)로 형성할 수 있으며, 다수의 트랜지스터는 제1 방향 및 제1 방향과 수직한 제2 방향을 따라 일정 간격 이격되어 배치된다. 이때, 제1 방향을 따라 인접한 두 개의 트랜지스터(TR1, TR2)가 하나의 시냅스 소자(320)로 사용된다. 도시된 바와 같이, 하나의 시냅스 소자(320)는 두 개의 트랜지스터로 구성되며, 여기서는 프로그램 동작을 진행하는 제1 트랜지스터(TR1) 및 이레이즈 동작을 진행하는 제2 트랜지스터(TR2)로 정의하도록 한다. 시냅스 소자(320)는 적어도 두 개의 트랜지스터를 포함하는 것이 바람직하나 이진신경망(binary neural networks)이 구현 가능한 범위 내에서 트랜지스터의 개수를 조절할 수 있다.
- [0054] 제1 트랜지스터(TR1) 및 제2 트랜지스터(TR2)의 드레인 영역(Drain)은 콘택 플러그를 통해 제1 방향으로 연장된 비트라인(330)과 연결되며, 소스 영역(Source)은 콘택 플러그를 통해 비트라인(330)과 평행하게 위치하는 메탈라인(340)과 접속되며, 메탈라인(340)은 그라운드에 연결된다. 그리고, 제1 트랜지스터(TR1) 및 이레이즈 동작을 진행하는 제2 트랜지스터(TR2)의 게이트 전극은 메탈 콘택(350)을 통해 비트라인(330)과 수직한 제2 방향으로 연장된 워드라인(360)과 연결된다.
- [0055] 도 2 및 도 3에 도시된 TFET는 기존 실리콘 기반의 MOSFET를 형성하기 위한 과정을 그대로 이용 가능하며, 동시 집적도 가능하기 때문에 시냅스 소자로의 활용에 용이한 장점이 있다. 도 2 및 도 3에 도시된 TFET를 이용한 시냅스 소자 기반의 인공신경망 장치들을 통해 이진 신경망(binary neural network)을 구축할 수 있다.
- [0057] 도 4는 본 발명의 일 실시예에 따른 인공 신경망 장치의 시냅스 소자로 활용되는 TFET에 대한 프로그램(program) 및 이레이즈(erase) 매커니즘을 설명하기 위한 도면이다.
- [0058] 먼저, 도 4a는 TFET 기반 시냅스 소자의 프로그램 동작 특성에 대한 설명을 나타낸 것으로, 도 4a (i)는 A - A'에 따른 측면 에너지 밴드 다이어그램을 나타내며, 도 4a (ii)는 B - B'에 따른 수직 에너지 밴드 다이어그램을 나타낸다. TFET는 도 4a에 도시된 같이 밴드간 터널링(BTBT; Band-to-band tunneling)에 의해 캐리어 이동이 일어난다. TFET는 MOSFET과 달리 소스, 드레인이 비대칭적인 구조로 형성되어 있으므로 핫 캐리어(hot carrier) 주입 효율이 우수하다. 따라서, 프로그램 동작 시 소수의 핫 캐리어 만으로도 메모리 특성을 얻을 수 있으며, 이에 따라 프로그램, 이레이즈 동작 과정에서 소모되는 에너지 또한 MOSFET에 비해 작게 나타난다.
- [0059] 또한, 도 4b는 TFET 기반 시냅스 소자의 이레이즈 동작 특성에 대한 설명을 나타낸 것으로, B - B'에 따른 수직 에너지 밴드 다이어그램을 도시한 것이다. TFET는 MOSFET과 마찬가지로 어레이 형태로 구성되어 있으므로 페이지 이레이즈 방식을 사용하며, 도 4b에 도시된 바와 같이 FN-터널링(tunneling) 방식으로 이레이즈가 진행된다.
- [0061] 도 5는 본 발명의 이진 신경망을 이용한 XNOR 동작 구현 방식을 설명하기 위한 개념도이다.
- [0062] 도 5를 참조하면, TFET를 이용한 초저전력 시냅스 소자를 사용하여 이진 신경망의 XNOR 동작을 구현하기 위한

개념도이다. 구체적으로 설명하면, 시냅스 소자를 구성하는 2개의 트랜지스터와 전압이 인가되는 비트라인(BL)과 반대의 전압이 인가되는 2개의 워드라인(WL)이 배치된다. 입력(Input)과 가중치(Weight)는 +1 또는 -1의 두 가지 값을 가질 수 있다. 도 5a 내지 도 5d를 참조하여 입력 값과 가중치 값에 따른 출력 값을 설명하면 다음과 같다.

[0063] 도 5a 및 도 5b는 가중치(Weight)값이 +1로 프로그램된 상태를 나타낸 것이다. 먼저, 도 5a와 같이 이레이즈 동작을 수행하는 제1 트랜지스터(TR1)에 저전압(Low V_t)이 인가되고, 프로그램 동작을 수행하는 제2 트랜지스터(TR2)에 고전압(High V_t)이 인가된다. 그리고 제1 트랜지스터와 연결된 제1 워드라인(WL1)의 전압이 온(On) 상태이고, 제2 트랜지스터와 연결된 제2 워드라인(WL2)의 전압이 오프(Off) 상태인 경우, 입력(Input) 값이 +1이 된다. 이 경우, 입력 값과 가중치 값이 일치하므로, 출력(Output)값은 +1로 표현된다.

[0064] 한편, 도 5b와 같이 제1 트랜지스터에 저전압, 제2 트랜지스터에 고전압 형태의 가중치가 인가되고, 제1 트랜지스터에 연결된 제1 워드라인의 전압이 오프 상태이고, 제2 트랜지스터와 연결된 제2 워드라인의 전압이 온 상태인 경우, 입력 값이 -1된다. 이 경우, 입력 값과 가중치 값이 일치하지 않으므로, 출력값은 -1로 표현될 수 있다.

[0065] 도 5c 및 도 5d는 가중치(Weight)값이 -1로 프로그램된 상태를 나타낸 것이다. 먼저, 도 5c와 같이 프로그램 동작을 수행하는 제1 트랜지스터에 고전압이 인가되고, 이레이즈 동작을 수행하는 제2 트랜지스터에 저전압이 인가된다. 그리고 제1 트랜지스터와 연결된 제1 워드라인의 전압이 온 상태이고, 제2 트랜지스터와 연결된 제2 워드라인의 전압이 오프 상태인 경우, 입력 값이 +1이 된다. 이 경우, 입력 값과 가중치 값이 일치하지 않으므로, 출력(Output)값은 -1로 표현된다.

[0066] 한편, 도 5d와 같이 제1 트랜지스터에 고전압, 제2 트랜지스터에 저전압이 인가되고, 제1 트랜지스터에 연결된 제1 워드라인의 전압이 오프 상태이고, 제2 트랜지스터와 연결된 제2 워드라인의 전압이 온 상태인 경우, 인풋 값이 -1이 되며, 입력 값과 가중치 값이 일치하므로 출력값은 +1로 표현될 수 있다.

[0067] 이와 같이, 각각의 시냅스 소자들의 가중치 값을 개별적으로 제어함과 동시에 해당 소자들에 대해서만 리드(read)가 가능한 것을 알 수 있다.

[0069] 도 6은 본 발명의 일 실시예에 따른 인공 신경망 장치를 구성하는 SONOS TFET의 특성을 나타내는 그래프로, SONOS MOSFET 소자와 SONOS TFET 소자의 전류-전압 전달 특성 곡선(transfer characteristics)을 도시한 것이다.

[0070] 도 6을 참조하면, SONOS TFET는 기존에 사용되던 MOSFET 대비 매우 낮은 구동 전류(I_{on})와 오프 상태 전류(I_{off})를 나타내는 것을 알 수 있다. TFET는 I_{on} 이 MOSFET 대비 약 10^5 가량 낮으며, 이는 TFET이 시냅스 소자로서 활용 가능성이 매우 높다는 점을 의미한다.

[0071] 또한 TFET의 경우 전류량이 매우 낮은 대신 MOSFET과 달리 소스, 드레인이 비대칭적인 구조로 형성되어 있으므로 TFET은 소스-채널 접합에서의 전기장이 MOSFET 대비 더 크기 때문에 핫 캐리어(hot carrier) 주입 효율 또한 더 크게 나타난다. 따라서 프로그램 시 소수의 핫 캐리어 만으로도 메모리 특성을 얻을 수 있는 장점이 있다.

[0073] 상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

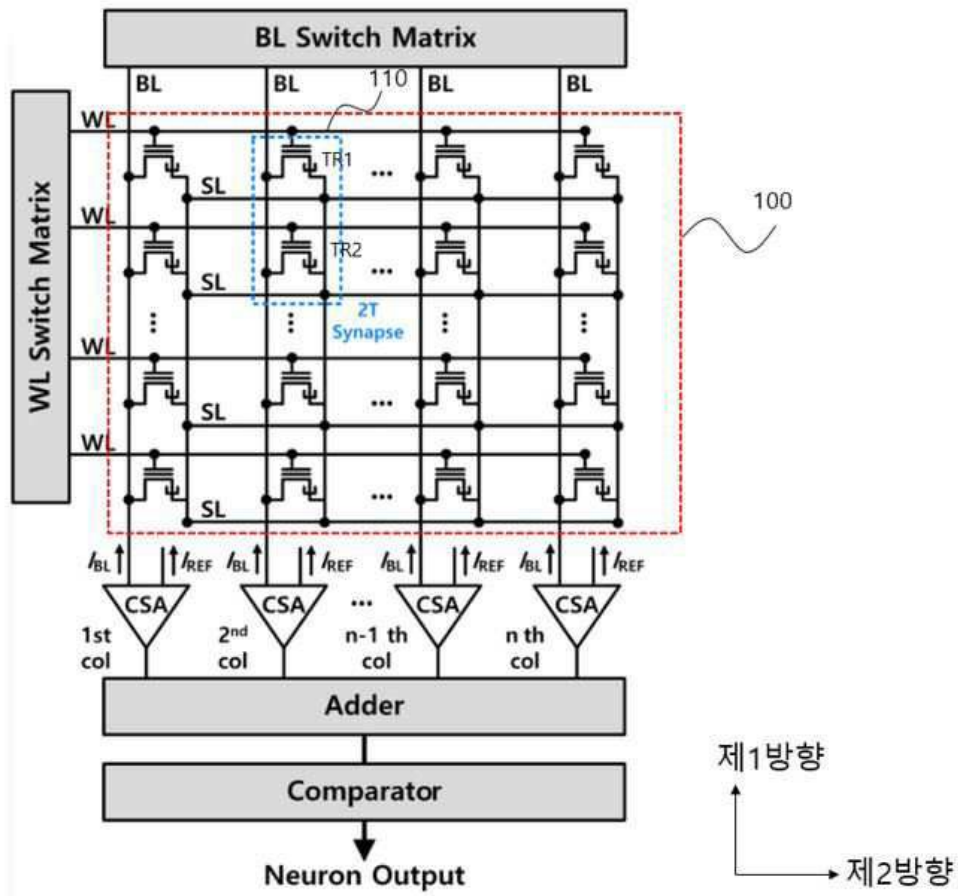
부호의 설명

[0075] 100 : 시냅스 어레이 110, 320 : 시냅스 소자
200, 300 : 매몰 산화막 210, 310 : 반도체 기판
220 : 게이트 패턴 220a : 하부 산화막

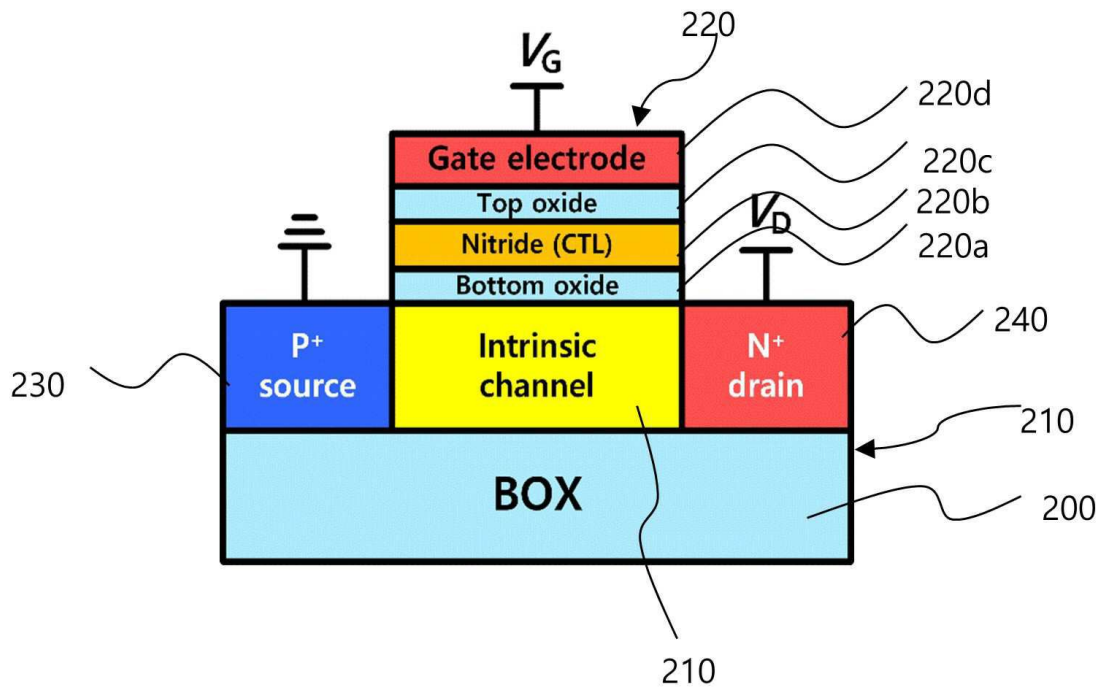
220b : 질화막 220c : 상부 산화막
 220d : 게이트 전극층 230 : 소스 영역
 240 : 드레인 영역 250 : 채널 영역
 330 : 비트라인 340 : 메탈 라인
 350 : 메탈 콘택 360 : 워드라인

도면

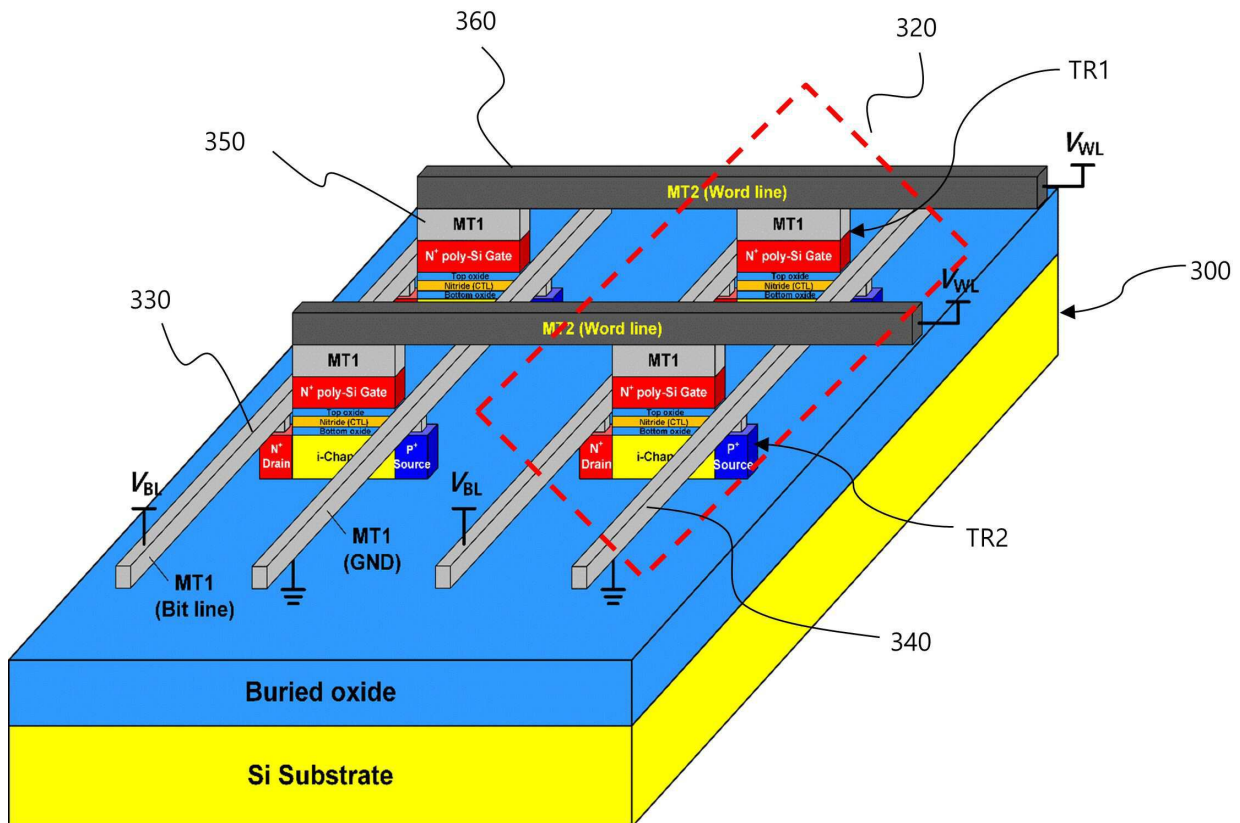
도면1



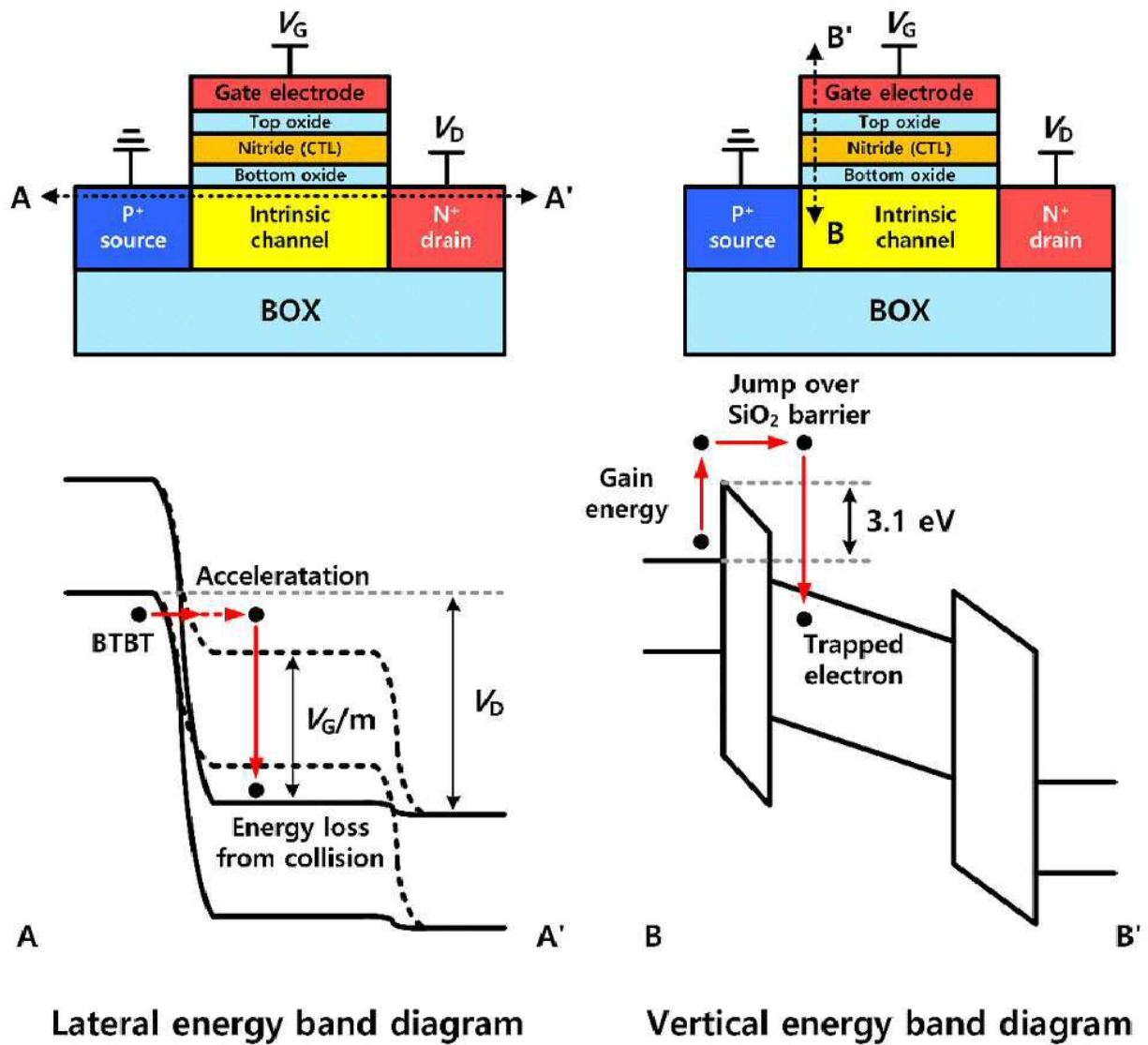
도면2



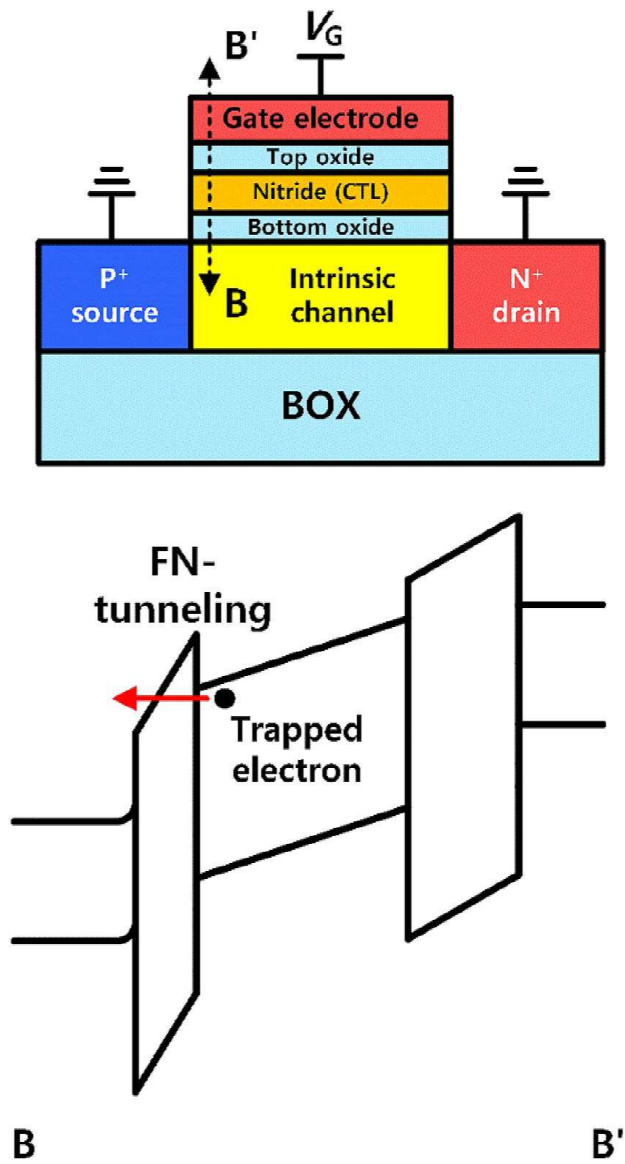
도면3



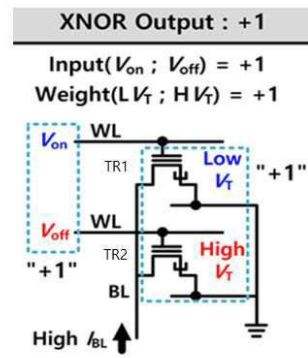
도면4a



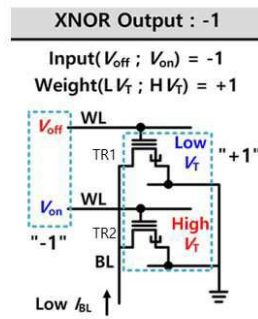
도면4b



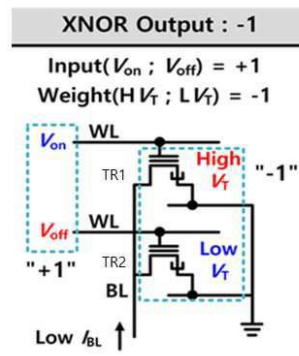
도면5a



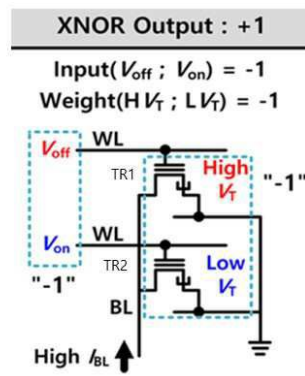
도면5b



도면5c



도면5d



도면6

