

(19) 대한민국특허청(KR)  
(12) 공개특허공보(A)(11) 공개번호 10-2023-0133051  
(43) 공개일자 2023년09월19일

(51) 국제특허분류(Int. Cl.)  
G06N 3/063 (2023.01) G11C 16/04 (2006.01)  
(52) CPC특허분류  
G06N 3/063 (2013.01)  
G11C 16/0441 (2013.01)  
(21) 출원번호 10-2022-0030083  
(22) 출원일자 2022년03월10일  
심사청구일자 2022년03월10일

(71) 출원인  
포항공과대학교 산학협력단  
경상북도 포항시 남구 청암로 77 (지곡동)  
(72) 발명자  
지원재  
경상북도 포항시 남구 청암로 77 (포항공과대학교) 기숙사 11동 315호  
김세영  
경상북도 포항시 남구 효자동 효자로 62 302동 203호 (효자동, 테라비아타)  
정윤영  
경상북도 포항시 남구 청암로 77 LG연구동 203호  
(74) 대리인  
정부연

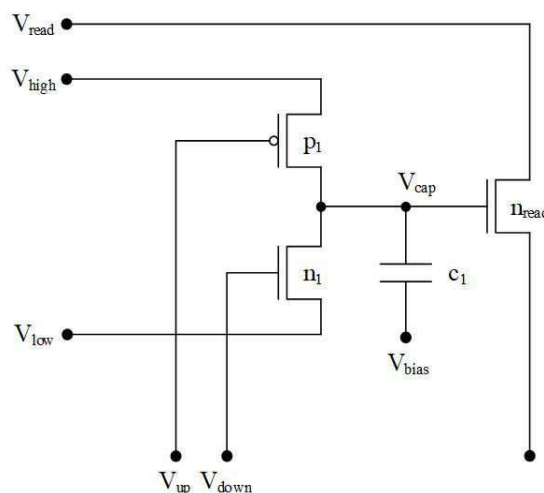
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 뉴로모픽 반도체 소자 및 동작 방법

## (57) 요약

본 발명은 뉴로모픽 반도체 소자의 동작 방법에 관한 것으로, 직렬로 연결되어 시냅스 소자의 가중치 값의 업데이트를 진행하는 제1 스위치 트랜지스터와 제2 스위치 트랜지스터와, 상기 제1 스위치 트랜지스터와 상기 제2 스위치 트랜지스터 사이에 연결되어 상기 시냅스 소자에 가중치 값을 저장하고, 저장된 가중치 값을 리드(read)하는 읽기용 3단자 메모리 소자를 포함하는 시냅스 소자에 있어서, 상기 제1 스위치 트랜지스터 및 상기 제2 스위치 트랜지스터 각각의 소스 단자와 게이트 단자에 전압을 인가하는 단계와, 상기 소스 단자 및 게이트 단자에 특정 전압이 인가되는 경우, 상기 시냅스 소자에 저장된 가중치 값이 증가 또는 감소되는 단계와, 상기 저장된 가중치 값의 증가 또는 감소에 따라 전압 펄스를 이용하여 가중치 값 업데이트를 진행하는 단계를 포함하는 것을 특징으로 한다.

대표도 - 도6



이 발명을 지원한 국가연구개발사업

|             |                                                             |
|-------------|-------------------------------------------------------------|
| 과제고유번호      | 1711130173                                                  |
| 과제번호        | 2020M3F3A2A01081240                                         |
| 부처명         | 과학기술정보통신부                                                   |
| 과제관리(전문)기관명 | 한국연구재단                                                      |
| 연구사업명       | 차세대지능형반도체기술개발(R&D)                                          |
| 연구과제명       | 고성능 온칩 훈련용 초저누설전류 산화물 반도체와 커패시터 기반 전하저장형 시냅스 소자 어레이 및 뉴런 회로 |
| 기 여 율       | 1/1                                                         |
| 과제수행기관명     | 서울대학교                                                       |
| 연구기간        | 2021.01.01 ~ 2021.12.31                                     |

---

## 명세서

### 청구범위

#### 청구항 1

직렬로 연결되어 시냅스 소자의 가중치 값의 업데이트를 진행하는 제1 스위치 트랜지스터와 제2 스위치 트랜지스터; 및

상기 제1 스위치 트랜지스터와 상기 제2 스위치 트랜지스터 사이에 연결되어 상기 시냅스 소자에 가중치 값을 저장하고, 저장된 가중치 값을 리드(read)하는 읽기용 3단자 메모리 소자를 포함하는 시냅스 소자에 있어서,

상기 제1 스위치 트랜지스터 및 상기 제2 스위치 트랜지스터 각각의 소스 단자와 게이트 단자에 전압을 인가하는 단계; 및

상기 소스 단자 및 게이트 단자에 특정 전압이 인가되는 경우, 상기 시냅스 소자에 저장된 가중치 값이 증가 또는 감소되는 단계; 및

상기 저장된 가중치 값의 증가 또는 감소에 따라 전압 펄스를 이용하여 가중치 값 업데이트를 진행하는 단계를 포함하는 것을 특징으로 하는 뉴로모픽 반도체 소자의 동작 방법.

#### 청구항 2

제1 항에 있어서, 상기 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터는

각각 산화물 반도체 기반의 PMOS 소자 및 NMOS 소자로 구성되는 것을 특징으로 하는 뉴로모픽 반도체 소자의 동작 방법.

#### 청구항 3

제1 항에 있어서, 상기 읽기용 3단자 메모리 소자는

상기 시냅스 소자의 가중치 값을 저장하는 캐패시터 및 상기 캐패시터에 저장된 가중치 값을 리드(read)하는 트랜지스터로 구성되는 것을 특징으로 하는 뉴로모픽 반도체 소자의 동작 방법.

#### 청구항 4

제1 항에 있어서, 상기 읽기용 3단자 메모리 소자는

전류 또는 전압에 의해 메모리 상태가 변하는 ECRAM(electrochemical random access memory), PRAM(Phase-change random access memory) 및 FeRAM(ferroelectric random access memory)로 구성되는 것을 특징으로 하는 뉴로모픽 반도체 소자의 동작 방법.

#### 청구항 5

제1 항에 있어서,

상기 제1 스위치 트랜지스터의 소스 단자로 인가된 제1 전압과, 상기 제1 스위치 트랜지스터의 게이트 단자로 인가된 제2 전압을 비교하여 제1 전압 값이  $V_{dd}$ , 제2 전압 값이  $V_{dd} - V_{th,p}$ 인 경우, 캐패시터가 충전되어 시냅스 가중치 값이 증가되는 것을 특징으로 하는 뉴로모픽 반도체 소자의 동작 방법.

## 청구항 6

제1 항에 있어서,

상기 제2 스위치 트랜지스터의 소스 단자로 인가된 제1 전압과, 상기 제1 스위치 트랜지스터의 게이트 단자로 인가된 제2 전압을 비교하여 제1 전압 값이  $V_{ss}$ , 제2 전압 값이  $V_{ss} + V_{th,n}$ 인 경우, 캐패시터가 방전되어 시냅스 가중치 값이 감소되는 것을 특징으로 하는 뉴로모픽 반도체 소자의 동작 방법.

## 청구항 7

제1 항에 있어서, 상기 가중치 값 업데이트를 진행하는 단계에서

상기 제1 스위치 트랜지스터 소스 단자 및 게이트 단자에 인가되는 증가 전압 펄스가 겹치는 경우 가중치 값이 증가되는 방향으로 업데이트가 발생하고, 제2 스위치 트랜지스터의 소스 단자 및 게이트 단자에 인가되는 감소 전압 펄스가 겹치는 경우 가중치 값이 감소되는 방향으로 업데이트가 발생하는 것을 특징으로 하는 뉴로모픽 반도체 소자의 동작 방법.

## 청구항 8

가중치 값의 증가 업데이트를 진행하는 제1 스위치 트랜지스터; 및

상기 제1 스위치 트랜지스터와 직렬로 연결되며, 가중치 값의 감소 업데이트를 진행하는 제2 스위치 트랜지스터; 및

상기 제1 스위치 트랜지스터와 상기 제2 스위치 트랜지스터 사이에 연결되어 상기 시냅스 소자에 가중치 값을 저장하고, 저장된 가중치 값을 리드(read)하는 읽기용 3단자 메모리 소자

를 포함하는 것을 특징으로 하는 뉴로모픽 반도체 소자.

## 청구항 9

제8 항에 있어서, 상기 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터는

각각 산화물 반도체 기반의 PMOS 소자 및 NMOS 소자로 구성되는 것을 특징으로 하는 뉴로모픽 반도체 소자의 동작 방법.

## 청구항 10

제9 항에 있어서,

상기 산화물 반도체 기반의 PMOS 소자는 구리 산화물( $Cu_2O$ ), 주석 산화물( $SnO$ ) 및 이들의 조합 중 선택된 어느 하나를 사용하며, 상기 산화물 반도체 기반의 NMOS 소자는 IGZO(In-Ga-Zn Oxide), GZO(Ga-Zn Oxide), IZTO(Indium-Zn-Tin Oxide), ZnO 및 이들의 조합으로 구성된 아연 산화물 기반의 산화물 반도체 소재를 사용하는 것을 특징으로 하는 뉴로모픽 반도체 소자.

## 청구항 11

제9 항에 있어서, 상기 읽기용 3단자 메모리 소자는

상기 시냅스 소자의 가중치 값을 저장하는 캐패시터 및 상기 캐패시터에 저장된 가중치 값을 리드(read)하는 트랜지스터로 구성되는 것을 특징으로 하는 뉴로모픽 반도체 소자.

## 청구항 12

제9 항에 있어서, 상기 읽기용 3단자 메모리 소자는

전류 또는 전압에 의해 메모리 상태가 변하는 ECRAM(electrochemical random access memory), PRAM(Phase-change random access memory) 및 FeRAM(ferroelectric random access memory)로 구성되는 것을 특징으로 하는 뉴로모픽 반도체 소자.

## 청구항 13

제12 항에 있어서,

상기 ECRAM(electrochemical random access memory), PRAM(Phase-change random access memory) 및 FeRAM(ferroelectric random access memory)의 게이트 전극에 전압을 인가하여 시냅스 소자의 가중치 값을 증가 또는 감소시키는 것을 특징으로 하는 뉴로모픽 반도체 소자.

## 발명의 설명

### 기술 분야

[0001] 본 발명은 뉴로모픽 반도체 소자 및 동작 방법에 관한 것으로, 보다 상세하게는 산화물 반도체 기반의 아날로그 시냅스 회로를 구성하여 병렬적인 가중치 업데이트 특성을 갖는 뉴로모픽 반도체 소자 및 그 동작 방법에 관한 것이다.

### 배경 기술

[0003] 최근 수 년간 급격히 발달한 심층 신경망(Deep Neural Network, DNN) 기술은 이전까지 컴퓨터로 구현이 어려웠던 영상 및 음성 분석, 자연어 처리, 자율 시스템 등 인간의 인지능력 구현을 가능케하며 인공 지능의 획기적 성능 향상을 이끌어 왔다.

[0004] 심층 신경망은 모델의 규모가 클수록 계층의 수가 많을수록 성능이 향상되는 경향을 보이므로 인공지능의 성능 향상을 위해서 더욱 크고 깊은 신경망이 개발되고 있다. 현재, 심층 신경망의 학습에 요구되는 막대한 계산량은 GPU(Graphics Processing Unit)를 이용한 병렬 컴퓨팅에 의존하고 있다. 그러나, GPU를 이용한 연산 가속은 메모리와 연산 장치가 분리된 폰 노이만 구조에서 발생하는 메모리 병목 현상이라는 근본적인 한계가 있다. 따라서 전통적인 반도체 집적도 향상을 통한 심층 신경망 학습의 속도나 에너지 효율의 획기적인 향상을 기대하기 어려우며, 심층 신경망 연산에 특화된 새로운 패러다임의 컴퓨팅 방식이 절실히 요구되고 있다. 이에 심층 신경망 연산의 가장 큰 비중을 차지하는 행렬곱 연산을 아날로그 연산을 통해 효율적으로 수행할 수 있는 저항형 연산 장치(Resistive Processing Unit, RPU)의 개념이 제안되었다. 이러한 저항형 연산 장치를 크로스 포인트 어레이 형태로 집적하게 되면, 기존의 디지털 시스템 대비 획기적으로 높은 에너지 효율을 얻을 수 있다. 이렇게 높은 신경망 학습 성능을 달성하면서도 학습 소요 시간을 획기적으로 가속할 수 있는 이상적인 저항형 연산 장치 소자는 1000개 이상의 메모리 단계를 가지며 저항 상태의 업데이트 방향에 대한 비대칭성이 5% 이내의 조건을 만족해야 하므로, 이러한 요구 조건을 완벽히 만족하는 이상적인 소자의 개발이 요구되고 있는 실정이다.

[0005] 공개특허 10-2020-0100286호는 효율적인 음수 가중치 구현이 가능한 뉴로모픽 회로 시스템에 있어서, 다수의 프리 뉴런, 다수의 포스트 뉴런, 상기 프리 뉴런의 각각에서 로우 방향으로 연장되는 다수의 로우 라인과, 상기 포스트 뉴런의 각각에 대응되는 다수의 컬럼 라인의 교차점들 상에서 배치되어, 시냅스 어레이를 형성하는 다수의 시냅스, 상기 다수의 프리 뉴런의 입력에 시프트 가중치를 가중하여 합산하고 합산된 결과를 출력하는 시프트 회로, 및 상기 다수의 컬럼 라인 각각의 출력에서 상기 시프트 회로의 출력을 차감하여, 차감된 출력을 상기 포스트 뉴런 각각에 출력시키는 차감 회로를 포함하고, 상기 다수의 시냅스 각각은 원래 가중치에서 상기 시프트 가중치로 시프트한 가중치를 가지는 것을 특징으로 한다.

## 선행기술문헌

### 특허문헌

[0007] (특허문헌 0001) 공개특허 제10-2020-0100286호(2020.08.26)

## 발명의 내용

### 해결하려는 과제

[0008] 본 발명의 일 실시예는 산화물 반도체 트랜지스터와 캐패시터 기반으로 전하 저장형 아날로그 시냅스 회로를 구성할 수 있으며, 산화물 반도체 트랜지스터를 스위치 소자로 사용하여 캐패시터의 전하 저장 특성을 향상시킬 수 있는 뉴로모픽 반도체 소자 및 동작 방법을 제공한다.

[0009] 또한, 읽기용 3단자 소자를 ECRAM, PRAM 또는 FeRAM과 같은 비휘발성 소자로 사용함에 따라 가중치 값 저장 시 누설 효과가 없는 시냅스 회로를 구성하여 회로 특성을 향상시킬 수 있는 뉴로모픽 반도체 소자 및 동작 방법을 제공한다.

### 과제의 해결 수단

[0011] 본 발명의 일 실시예에 따른 뉴로모픽 반도체 소자 및 동작 방법은 직렬로 연결되어 시냅스 소자의 가중치 값의 업데이트를 진행하는 제1 스위치 트랜지스터와 제2 스위치 트랜지스터와,

[0012] 상기 제1 스위치 트랜지스터와 상기 제2 스위치 트랜지스터 사이에 연결되어 상기 시냅스 소자에 가중치 값을 저장하고, 저장된 가중치 값을 리드(read)하는 읽기용 3단자 메모리 소자를 포함하는 시냅스 소자에 있어서,

[0013] 상기 제1 스위치 트랜지스터 및 상기 제2 스위치 트랜지스터 각각의 소스 단자와 게이트 단자에 전압을 인가하는 단계와,

[0014] 상기 소스 단자 및 게이트 단자에 특정 전압이 인가되는 경우, 상기 시냅스 소자에 저장된 가중치 값이 증가 또는 감소되는 단계와,

[0015] 상기 저장된 가중치 값의 증가 또는 감소에 따라 전압 펄스를 이용하여 가중치 값 업데이트를 진행하는 단계를 포함하는 것을 특징으로 한다.

[0016] 상기 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터는 각각 산화물 반도체 기반의 PMOS 소자 및 NMOS 소자로 구성된다.

[0017] 상기 읽기용 3단자 메모리 소자는 상기 시냅스 소자의 가중치 값을 저장하는 캐패시터 및 상기 캐패시터에 저장된 가중치 값을 리드(read)하는 트랜지스터로 구성된다.

[0018] 상기 읽기용 3단자 메모리 소자는 전류 또는 전압에 의해 메모리 상태가 변하는 ECRAM(electrochemical random access memory), PRAM(Phase-change random access memory) 및 FeRAM(ferroelectric random access memory)로 구성된다.

[0019] 상기 제1 스위치 트랜지스터의 소스 단자로 인가된 제1 전압과, 상기 제1 스위치 트랜지스터의 게이트 단자로 인가된 제2 전압을 비교하여 제1 전압 값이  $V_{dd}$ , 제2 전압 값이  $V_{dd} - V_{th,p}$ 인 경우, 캐패시터가 충전되어 시냅스 가중치 값이 증가된다.

[0020] 상기 제2 스위치 트랜지스터의 소스 단자로 인가된 제1 전압과, 상기 제1 스위치 트랜지스터의 게이트 단자로 인가된 제2 전압을 비교하여 제1 전압 값이  $V_{ss}$ , 제2 전압 값이  $V_{ss} + V_{th,n}$ 인 경우, 캐패시터가 방전되어 시냅스 가중치 값이 감소된다.

[0021] 상기 가중치 값 업데이트를 진행하는 단계에서

[0022] 상기 제1 스위치 트랜지스터 소스 단자 및 게이트 단자에 인가되는 증가 전압 펄스가 겹치는 경우 가중치 값이

증가되는 방향으로 업데이트가 발생하고, 제2 스위치 트랜지스터의 소스 단자 및 게이트 단자에 인가되는 감소 전압 펄스가 겹치는 경우 가중치 값이 감소되는 방향으로 업데이트가 발생하는 것을 특징으로 한다.

- [0023] 본 발명의 일 실시예에 따른 뉴로모픽 반도체 소자는 가중치 값의 증가 업데이트를 진행하는 제1 스위치 트랜지스터와,
- [0024] 상기 제1 스위치 트랜지스터와 직렬로 연결되며, 가중치 값의 감소 업데이트를 진행하는 제2 스위치 트랜지스터와,
- [0025] 상기 제1 스위치 트랜지스터와 상기 제2 스위치 트랜지스터 사이에 연결되어 상기 시냅스 소자에 가중치 값을 저장하고, 저장된 가중치 값을 리드(read)하는 읽기용 3단자 메모리 소자를 포함하는 것을 특징으로 한다.
- [0026] 상기 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터는 각각 산화물 반도체 기반의 PMOS 소자 및 NMOS 소자로 구성되며, 상기 산화물 반도체 기반의 PMOS 소자는 구리 산화물( $\text{Cu}_2\text{O}$ ), 주석 산화물( $\text{SnO}$ ) 및 이들의 조합 중 선택된 어느 하나를 사용하며, 상기 산화물 반도체 기반의 NMOS 소자는 인듐-갈륨-아연 산화물(In-Ga-Zn Oxide, IGZO)을 사용할 수 있다.
- [0027] 상기 읽기용 3단자 메모리 소자는 상기 시냅스 소자의 가중치 값을 저장하는 캐패시터 및 상기 캐패시터에 저장된 가중치 값을 리드(read)하는 트랜지스터로 구성된다. 상기 읽기용 3단자 메모리 소자는 전류 또는 전압에 의해 메모리 상태가 변하는 ECRAM(electrochemical random access memory), PRAM(Phase-change random access memory) 및 FeRAM(ferroelectric random access memory)로 구성될 수 있다.
- [0028] 상기 ECRAM(electrochemical random access memory), PRAM(Phase-change random access memory) 및 FeRAM(ferroelectric random access memory)의 게이트 전극에 전압을 인가하여 시냅스 소자의 가중치 값을 증가 또는 감소시키는 것을 특징으로 한다.

### 발명의 효과

- [0030] 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.
- [0031] 본 발명의 일 실시예에 따른 뉴로모픽 반도체 소자 및 그 동작 방법은 산화물 반도체 트랜지스터와 캐패시터 기반으로 전하 저장형 아날로그 시냅스 회로를 구성할 수 있으며, 산화물 반도체 트랜지스터를 스위치 소자로 사용하여 캐패시터의 전하 저장 특성을 향상시킬 수 있는 효과를 얻을 수 있다.
- [0033] 또한, 본 발명의 다른 실시예에 따른 뉴로모픽 반도체 소자 및 그 동작 방법은 읽기용 3단자 소자를 ECRAM, PRAM 또는 FeRAM과 같은 비휘발성 소자로 사용함에 따라 가중치 값 저장 시 누설 효과가 없는 시냅스 회로를 구성하여 회로 특성을 향상시킬 수 있는 효과를 얻을 수 있다.

### 도면의 간단한 설명

- [0035] 도 1은 전 결합 계층 뉴럴 네트워크(Fully Conncted Layer neural network)를 도시한 것이다.
- 도 2는 도 1과 같은 전 결합 계층 인공 신경망 중 2층 구조의 전 결합 계층 인공 신경망을 도시한 것이다.
- 도 3은 도 2와 같은 인공 신경망 구조를 적용할 수 있는 시냅스 어레이를 나타내는 도면이다.
- 도 4 및 도 5는 인공 신경망의 시냅스 누설 특성에 따른 인공 신경망의 성능을 설명하기 위한 그래프이다.
- 도 6은 본 발명의 일 실시예에 따른 뉴로모픽 반도체 소자의 시냅스 회로를 도시한 것이다.
- 도 7 및 도 8은 본 발명의 일 실시예에 따른 뉴로모픽 반도체 소자의 동작 원리를 설명하기 위한 도면이다.
- 도 9 및 도 10은 본 발명의 일 실시예에 따른 뉴로모픽 반도체 소자의 시냅스 회로의 업데이트 동작 원리를 설명하는 도면이다.

도 11 내지 도 13은 본 발명의 다른 실시예에 따른 뉴로모픽 반도체 소자의 시냅스 회로를 도시한 것이다.

### 발명을 실시하기 위한 구체적인 내용

- [0036] 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.
- [0037] 한편, 본 출원에서 서술되는 용어의 의미는 다음과 같이 이해되어야 할 것이다.
- [0038] "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.
- [0039] 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.
- [0040] 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.
- [0041] 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.
- [0042] 본 발명은 컴퓨터가 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현될 수 있고, 컴퓨터가 읽을 수 있는 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.
- [0043] 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.
- [0044] 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.
- [0046] 인공 신경망은 생물학적인 신경망을 수학적 모델로 모방한 것으로, 하나의 뉴런이 여러 개의 뉴런으로부터 입력을 받아들여 출력을 내보낸다. 이 때, 특정 뉴런 간의 연결, 혹은 시냅스는 변동 가능한 특정 가중치 값을 갖는다. 즉, 하나의 뉴런은 여러 개의 뉴런으로부터 가중치가 반영된 입력 값을 받아들인다. 뉴런에 입력된 값에 가중치를 반영하여 다음 뉴런으로 출력하는 간단한 형태의 인공 신경망은 선형 분류기로써 동작할 수 있다. 인공 신경망은 음성 인식, 이미지 분류 등을 수행하는 인공지능을 구현하기 위해 사용되는 대표적인 모델이며, 특정 문제를 해결하기 위해서는 특정 데이터로 신경망을 학습시켜 가중치를 최적화한다. 일반적으로, 인공 신경망은 특정한 데이터로부터 특정한 시냅스 값을 최적화하기 위해 오차 역전파 알고리즘을 사용한다. 오차 역전파 알고



리즘은 오차의 계산과 이의 역전파를 통한 가중치 갱신을 반복함으로써 신경망의 가중치 값을 최적화하는 알고리즘이다. 오차 역전파 알고리즘을 사용하는 경우, 데이터를 인공 신경망에 입력하여 인공 신경망의 출력을 정답 데이터와 비교함으로써 오차를 구한 후, 오차를 인공 신경망의 마지막 층으로부터 첫 번째 층의 순서로 가중치를 갱신하는 데에 사용한다. 따라서, 학습된 인공 신경망은 특정 데이터셋으로부터 최적화된 가중치 값을 저장한다. 학습된 인공 신경망은 데이터에 대한 정답을 추론할 수 있다. 학습된 인공 신경망은 학습 데이터셋으로부터 정답을 출력하는 것에 최적화된 가중치를 갖고 있기 때문에, 학습된 인공 신경망에 입력된 데이터는 가중치 값을 거쳐 정답으로 출력된다.

[0048] 도 1은 전 결합 계층 뉴럴 네트워크(Fully Connected Layer neural network)를 도시한 것이다.

[0049] 도 1을 참조하면, 여러 뉴런이 모인 단위를 계층(layer)라고 하며, 전결합 계층 구조는 각 계층의 모든 경우들이 연결되어 있는 구조이다. 입력 계층의 뉴런들과 출력 계층의 뉴런들이 연결될 수 있는 모든 경우의 수와 동일하게 연결이 되어 있으면 이를 전결합 계층(Fully Connected Layer)라고 한다. 이러한 뉴럴 네트워크는 입력 계층(Input layer)(100), 은닉 계층(hidden layer)(110) 및 출력 계층(output layer)(120)로 구성된다.

[0050] 입력 계층은 입력을 받아서 다음 계층인 은닉 계층으로 넘기는 역할을 하고, 은닉 계층은 입력 계층과 연결된 전결합 계층이며, 복잡한 문제를 해결할 수 있게 하는 핵심적인 계층이라고 할 수 있다. 마지막으로 출력계층은 은닉 계층 다음에 오는 전결합 계층으로, 신경망의 외부로 출력 신호를 전달하는 데에 사용하며, 신경망의 기능은 출력 계층의 활성 함수에 의해 결정된다.

[0051] 여기서는, MNIST 데이터베이스(Modified National Institute Standards and Technology)를 구분하는 작업을 수행하는 2단 전 결합 계층 뉴럴 네트워크를 도시한 것으로, 예를 들어, 사용되는 데이터가 총 784개의 픽셀로 구성되어 있다면, 입력 노드 역시 784개로 구성된다.

[0052] 또한, 출력으로는 0 내지 9 사이의 숫자를 구분해야 하기 때문에 총 10개의 출력 노드로 구성된다. 여기서는 두 개의 은닉층을 도시하고 있으며, 첫번째 은닉층에 256개의 노드, 두번째 은닉층에 128개의 노드로 설정하였으나, 은닉층을 구성하는 노드의 개수는 이에 한정하지는 않는다. 트레이닝 과정은 forward pass 및 backward pass로 구성되어 있으며, 이후 가중치 업데이트가 이루어진다. 이때 행렬 연산이 가장 많은 비중을 차지한다.

[0054] 도 2는 도 1과 같은 전 결합 계층 인공 신경망 중 2층 구조의 전 결합 계층 인공 신경망을 도시한 것이고, 도 3은 도 2와 같은 인공 신경망 구조를 적용할 수 있는 시냅스 어레이를 나타내는 도면이다.

[0055] 도 2 및 도 3의 시냅스 어레이는 아날로그 하드웨어 가속기에서 시냅스 소자를 사용한  $n \times m$  크기의 시냅스 어레이를 나타내며, 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 소자가 위치한다.  $n \times m$  크기의 시냅스 어레이는  $i$ 번째 층의  $n$ 개의 뉴런과  $i+1$ 번째 층의  $m$ 개의 뉴런을 연결하는  $nm$ 개의 시냅스를 요소로 갖는다. 시냅스 어레이는 인공 신경망의 가중치 값을 각 시냅스에 저장하며, 가중치 값을 병렬적으로 갱신할 수 있다.

[0056] 이러한 어레이로 구성된 뉴럴 네트워크 장치는 행렬 연산의 비중이 높으며, 해당 연산의 행렬의 각 요소들의 값을 각각의 메모리 소자의 전도도로 치환할 수 있으며, 전압 펄스(Voltage pulse)를 주어 흘러나오는 커런트를 통합하여 행렬 곱을 연산할 수 있다. 이때, 각 시냅스 소자에 저장된 컨덕턴스의 값은 뉴럴 네트워크에서의 가중치로 사용되며, 가중치가 정확하게 기록될수록 학습 성능 특성이 우수한 뉴럴 네트워크를 제공할 수 있다.

[0058] 도 4 및 도 5는 인공 신경망의 시냅스 누설 특성에 따른 인공 신경망의 성능을 설명하기 위한 그래프이다.

[0059] 도 4 및 도 5를 참조하면, 시냅스 회로가 가중치를 보유하는 특성인 라이프 타임이 감소할수록, 인공 신경망의 성능이 감소하는 것을 알 수 있다. 또한, 시냅스 회로의 라이프 타임이 감소할수록 인공 신경망의 학습 성능이 이상적인 성능에 비해 저하되는 것을 알 수 있다. 이와 같이 뉴로모픽 시스템을 사용한 인공 신경망 학습 성능은 시냅스 소자의 가중치 보유 특성에 따라 좌우됨을 알 수 있다. 이에 따라 본 발명에서는 가중치 보유 특성이 향상된 시냅스 회로로 구성된 뉴로모픽 반도체 소자를 제공함으로써 인공 신경망의 학습 성능을 향상시키고자

한다.

- [0061] 도 6은 본 발명의 일 실시예에 따른 뉴로모픽 반도체 소자의 시냅스 회로를 도시한 것이다.
- [0062] 도 6을 참조하면, 본 발명의 일 실시예에 따른 시냅스 소자는 제1 스위치 트랜지스터(p1), 제2 스위치 트랜지스터(n1) 및 읽기용 3단자 소자를 포함한다. 여기서, 읽기용 3단자 소자는 캐패시터(c1) 및 읽기용 트랜지스터( $n_{\text{read}}$ )로 구성될 수 있다.
- [0063] 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터는 전류 소스로 사용되며, 게이트 전극과 소스 전극의 전도도 증가 또는 감소 상황에 따라 가중치 값을 증가 또는 감소하는 방향으로 업데이트 동작을 구현할 수 있다. 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터의 게이트 전극은 입력 단자( $V_{\text{up}}$ ,  $V_{\text{down}}$ )와 전기적으로 연결되며, 제1 스위치 트랜지스터의 소스 전극은 전원 전압( $V_{\text{high}}$ )과 전기적으로 연결될 수 있고, 제2 스위치 트랜지스터의 소스 전극은 그라운드 전압( $V_{\text{low}}$ )과 전기적으로 연결될 수 있다.
- [0064] 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터는 직렬로 연결되며, 각각 산화물 반도체 기반의 PMOS 소자 및 NMOS 소자로 구성될 수 있다. 예를 들어, P형 산화물 반도체 산화물로는 구리 산화물( $\text{Cu}_2\text{O}$ ), 주석 산화물( $\text{SnO}$ ) 및 이들의 조합 중 선택된 어느 하나를 사용하여 PMOS 소자를 구성할 수 있다. 또한, N형 반도체 산화물로는 IGZO(In-Ga-Zn Oxide), GZO(Ga-Zn Oxide), IZTO(Indium-Zn-Tin Oxide), ZnO 및 이들의 조합으로 구성되는 것이 바람직하다. 그러나, N형 반도체 산화물의 구성을 명시된 물질로 한정하지 않으며 아연 산화물 기반의 반도체 산화물이라면 어떤 형태라도 N형 반도체 산화물로 구성할 수 있다.
- [0065] 이러한 PMOS 소자 및 NMOS 소자는 실리콘 기반의 MOSFET 소자에 비해 낮은 누설 전류 특성을 가지고 있으므로, 메모리 상태의 변화를 막고, 긴 리텐션(retention) 시간을 확보할 수 있으며, 전류에 따른 캐패시터의 방전 속도를 감소시킬 수 있다.
- [0066] 캐패시터(c1)는 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터와 전기적으로 연결되어 구성된다. 캐패시터는 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터의 동작에 따라 충전 또는 방전이 진행되며, 시냅스의 가중치 값을 저장할 수 있다. 캐패시터에 저장된 전하는 캐패시터 양단의 전위차( $V_{\text{cap}}$ )을 발생시킨다.
- [0067] 읽기용 트랜지스터( $n_{\text{read}}$ )는 캐패시터와 전기적으로 연결되며, 시냅스에 저장된 가중치 값을 읽어낼 수 있다. 캐패시터 양단에 발생된 전위차( $V_{\text{cap}}$ )에 의해 읽기용 트랜지스터의 게이트 전극에 전압이 인가된다. 이에 따라, 읽기용 트랜지스터의 소스 단자 및 드레인 단자를 통해 읽기용 트랜지스터에 흐르는 전류가 바뀌는 현상을 확인함으로써 캐패시터에 저장된 전하, 시냅스 가중치 값을 읽을 수 있다.
- [0069] 도 7 및 도 8은 본 발명의 일 실시예에 따른 뉴로모픽 반도체 소자의 동작 원리를 설명하기 위한 도면으로, 시냅스에 저장된 가중치 값이 증가 또는 감소하는 경우의 동작을 나타낸 것이다.
- [0070] 도 7은 시냅스에 저장된 가중치 값이 증가하는 경우의 동작을 나타내는 것이다. 도 7과 같이 제1 스위치 트랜지스터(p1)의 소스 전극( $V_{\text{high}}$ )에  $V_{\text{dd}}$ 가 인가되고, 제1 스위치 트랜지스터(p1)게이트 전극( $V_{\text{up}}$ )에  $V_{\text{dd}}-V_{\text{th,p}}$ 가 인가되는 경우, 제1 스위치 트랜지스터는 온(on) 상태가 된다. 한편, 제2 스위치 트랜지스터(n1)의 소스 전극( $V_{\text{low}}$ )에  $V_{\text{ss}}$ , 제2 스위치 트랜지스터(n1)의 게이트 전극( $V_{\text{down}}$ )에  $V_{\text{ss}}$ 가 인가되면 제2 스위치 트랜지스터는 오프(off) 상태가 된다. 따라서, 캐패시터는 제1 스위치 트랜지스터의 소스 전극( $V_{\text{high}}$ )으로부터 전하를 충전하며 캐패시터의 전압( $V_{\text{cap}}$ )이 증가된다. 이때, 캐패시터의 단자에  $V_{\text{bias}}$ 를 인가함으로써 캐패시터 양단에 발생하는 전압의 범위를  $V_{\text{bias}}$ 와  $V_{\text{dd}}$  사이로 정할 수 있다.
- [0071] 도 8은 시냅스에 저장된 가중치 값이 감소하는 경우의 동작을 나타내는 것이다. 도 8과 같이 제1 스위치 트랜지스터(p1)의 소스 전극( $V_{\text{high}}$ )에  $V_{\text{dd}}$ 가 인가되고, 제1 스위치 트랜지스터(p1)게이트 전극( $V_{\text{up}}$ )에  $V_{\text{dd}}$ 가 인가되는 경우, 제1 스위치 트랜지스터는 오프(off) 상태가 된다. 한편, 제2 스위치 트랜지스터(n1)의 소스 전극( $V_{\text{low}}$ )에  $V_{\text{ss}}$ , 제2 스위치 트랜지스터(n1)의 게이트 전극( $V_{\text{down}}$ )에  $V_{\text{ss}}+V_{\text{th,n}}$ 이 인가되면 제2 스위치 트랜지스터는 온(on) 상

태가 된다. 따라서, 캐패시터는 제2 스위치 트랜지스터의 소스 전극( $V_{high}$ )으로부터 전하가 방전되며 캐패시터의 전압( $V_{cap}$ )이 감소된다. 이때, 캐패시터의 단자에  $V_{bias}$ 를 인가함으로써 캐패시터 양단에 발생하는 전압의 범위를  $V_{bias}$ 와  $V_{dd}$  사이로 정할 수 있다.

[0073] 도 9 및 도 10은 본 발명의 일 실시예에 따른 뉴로모픽 반도체 소자의 시냅스 회로의 업데이트 동작 원리를 설명하는 도면으로, 전압 펄스를 이용한 어레이 병렬 업데이트 동작을 나타낸다.

[0074] 도 9를 참조하면, 시냅스의 가중치 값을 증가시키는 동작을 나타내는 것으로, 제1 스위치 트랜지스터의 소스 전극( $V_{high}$ ) 및 게이트 전극( $V_{up}$ )에 인가된 전압에 따라 캐패시터의 충전 여부가 결정된다. 이때, 인가되는 전압은  $V_{dd}$ 와  $V_{dd}-V_{th,p}$  전압을 갖는 전압 펄스를 사용할 수 있으며, 전압 펄스는 확률적으로 계산된 무작위적인 전압 펄스를 이용할 수 있다.

[0075] 아래의 <표 1>와 같이 제1 스위치 트랜지스터의 소스 전극( $V_{high}$ , Col<sub>up</sub>)에  $V_{dd}$ 가 인가되고, 제1 스위치 트랜지스터의 게이트 전극( $V_{up}$ , Row<sub>up</sub>)에  $V_{dd}-V_{th,p}$ 가 인가되는 경우에만 캐패시터가 충전된다. 즉, 소스 전극 및 게이트 전극의 전압 펄스가 겹칠 때 가중치 값의 증가 업데이트가 발생한다.

[0076] < 표 1 >

| Col <sub>up</sub> \ Row <sub>up</sub> | $V_{DD}$ | $V_{DD}-V_{th,p}$ |
|---------------------------------------|----------|-------------------|
| $V_{DD}$                              | 꺼짐       | 충전                |
| $V_{DD}-V_{th,p}$                     | 꺼짐       | 꺼짐                |

[0077]

[0078] 또한, 도 10을 참조하면, 시냅스의 가중치 값을 감소시키는 동작을 나타내는 것으로, 제2 스위치 트랜지스터의 소스 전극( $V_{low}$ )에 인가되는 전압과 게이트 전극( $V_{down}$ )에 인가되는 전압에 따라 캐패시터의 방전 여부가 결정된다. 이때, 인가되는 전압은  $V_{ss}$ 와  $V_{ss}+V_{th,n}$  전압을 갖는 전압 펄스를 사용할 수 있으며, 전압 펄스는 확률적으로 계산된 무작위적인 전압 펄스를 이용할 수 있다.

[0079] 아래의 <표 2>와 같이 제2 스위치 트랜지스터의 소스 전극( $V_{low}$ , Col<sub>dn</sub>)에  $V_{ss}$ 가 인가되고, 제2 스위치 트랜지스터의 게이트 전극( $V_{down}$ , Row<sub>dn</sub>)에  $V_{ss}+V_{th,n}$ 이 인가되는 경우에만 캐패시터가 방전된다. 즉, 소스 전극 및 게이트 전극의 전압 펄스가 겹칠 때 가중치 값의 감소 업데이트가 발생한다.

[0080] < 표 2 >

| Col <sub>dn</sub> \ Row <sub>dn</sub> | $V_{SS}$ | $V_{SS}+V_{th,n}$ |
|---------------------------------------|----------|-------------------|
| $V_{SS}$                              | 꺼짐       | 방전                |
| $V_{SS}+V_{th,n}$                     | 꺼짐       | 꺼짐                |

[0081]

[0083] 도 11 내지 도 13은 본 발명의 다른 실시예에 따른 뉴로모픽 반도체 소자의 시냅스 회로를 도시한 것이다.

[0084] 도 11, 도 12 및 도 13을 참조하면, 시냅스 소자는 제1 스위치 트랜지스터(p1), 제2 스위치 트랜지스터(p2) 및 읽기용 3단자 소자로 구성될 수 있다.

[0085] 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터는 직렬로 연결되며, 각각 산화물 반도체 기반의 PMOS 소자 및 NMOS 소자로 구성될 수 있다. 예를 들어, P형 산화물 반도체 산화물로는 구리 산화물( $Cu_2O$ ), 주석 산화물( $SnO$ )

및 이들의 조합 중 선택된 어느 하나를 사용하여 PMOS 소자를 구성할 수 있다. 또한, N형 반도체 산화물로는 IGZO(In-Ga-Zn Oxide), GZO(Ga-Zn Oxide), IZTO(Indium-Zn-Tin Oxide), ZnO 및 이들의 조합으로 구성된 아연 산화물 기반의 산화물 반도체 소재를 사용할 수 있다. 그러나, N형 반도체 산화물의 구성을 명시된 물질로 한정하지 않으며 아연 산화물 기반의 반도체 산화물이라면 어떤 형태라도 N형 반도체 산화물로 구성할 수 있다.

[0086] 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터는 전류 소스로 사용되며, 게이트 전극과 소스 전극에 전도도 증가 또는 감소 상황에 따라 가중치 값을 증가 또는 감소하는 방향으로 업데이트 동작을 구현할 수 있다. 제1 스위치 트랜지스터 및 제2 스위치 트랜지스터의 게이트 전극은 입력 단자와 전기적으로 연결되며, 제1 스위치 트랜지스터의 소스 전극은 전원 전압 노드와 전기적으로 연결될 수 있고, 제2 스위치 트랜지스터의 소스 전극은 그라운드 전압 노드와 전기적으로 연결될 수 있다.

[0087] 읽기용 3단자 소자는 전류나 전압에 의해서 메모리 상태가 변하는 비휘발성 메모리 소자로 사용할 수 있으며 예를 들어, ECRAM(electrochemical random access memory), PRAM(Phase-change random access memory) 및 FeRAM(ferroelectric random access memory)를 사용할 수 있다.

[0088] 도 11을 참조하면, 읽기용 3단자 소자로 ECRAM 소자를 사용한 것으로, ECRAM 소자의 게이트 전극에 전압을 인가하여 ECRAM 소자에 저장된 시냅스 가중치를 증가 또는 감소시킬 수 있다.

[0089] 또한, 도 12를 참조하면, 읽기용 3단자 소자로 PRAM을 사용한 것으로, PRAM 소자 일측의 게이트 전극에 전압을 인가함으로써 PRAM 소자에 저장된 시냅스 가중치를 증가 또는 감소시킬 수 있다.

[0090] 도 13을 참조하면, 읽기용 3단자 소자로 FeRAM 소자를 사용한 것으로, FeRAM 소자의 게이트 전극에 전압을 인가함으로써, FeRAM 소자에 저장된 시냅스 가중치를 증가 또는 감소시킬 수 있다.

[0091] 도 11 내지 도 13과 같은 ECRAM, PRAM 및 FeRAM 소자를 적용하는 경우에는 트랜지스터의 게이트 전극을 메모리 저장용으로 활용할 수 있으며, 읽기용 소자를 하나의 트랜지스터로 줄일 수 있어 추가적인 셀 면적을 감소시킬 수 있다. 이러한 시냅스 소자를 어레이 형태로 구성하여 전체 시냅스 어레이 값의 병렬적인 업데이트와 읽기가 가능하다. 이때,  $V_{high}$  또는  $V_{low}$ ,  $V_{up}$  또는  $V_{down}$  단자에 전압 펄스를 인가함으로써 전체 소자가 병렬적으로 동작하도록 할 수 있다.

[0093] 상술한 바와 같이 본 발명의 실시예에 따른 뉴로모픽 반도체 소자는 산화물 반도체 트랜지스터와 캐패시터 기반으로 전하 저장형 아날로그 시냅스 회로를 구성할 수 있으며, 산화물 반도체 트랜지스터를 스위치 소자로 사용하여 캐패시터의 전하 저장 특성을 향상시킬 수 있다.

[0094] 또한, 읽기용 3단자 소자를 ECRAM, PRAM 또는 FeRAM과 같은 비휘발성 소자로 사용함에 따라 가중치 값 저장 시 누설 효과가 없는 시냅스 회로를 구성하여 회로 특성을 향상시킬 수 있다.

[0096] 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

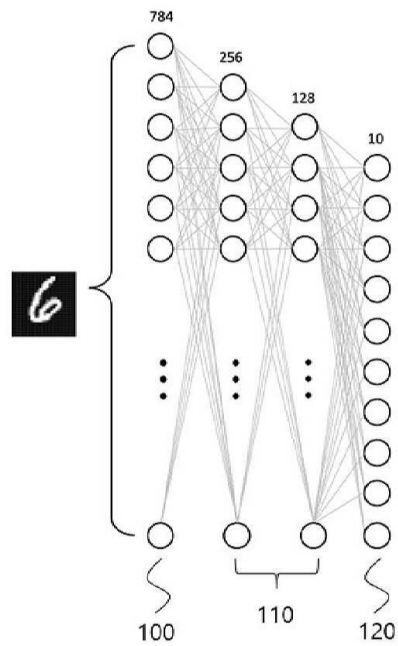
[0098] 100 : 입력 계층    110 : 은닉 계층

120 : 출력 계층

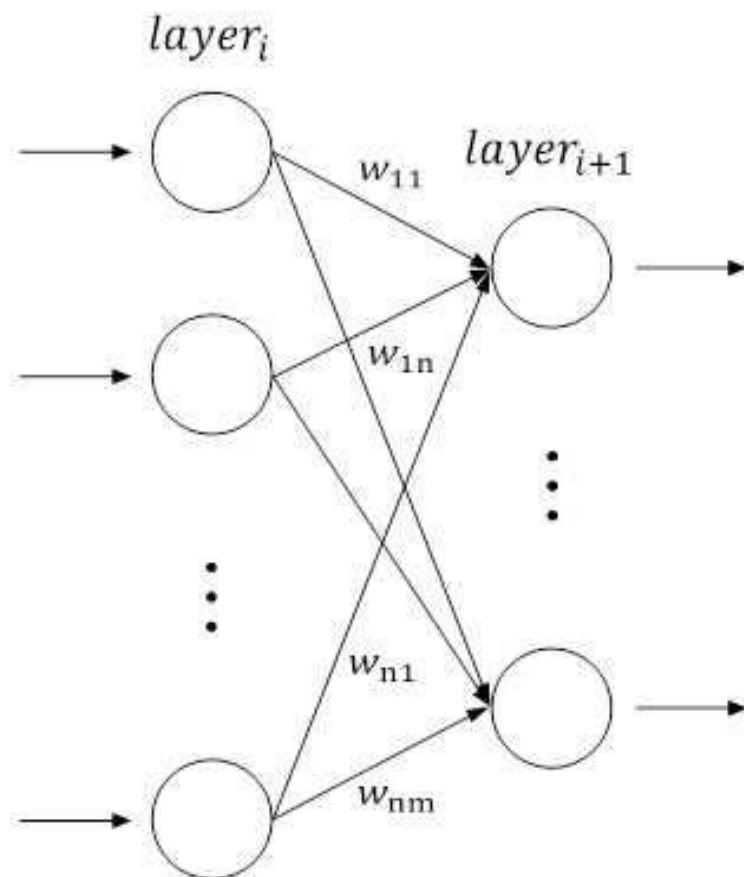
p1 : 제1 스위치 트랜지스터    n1 : 제2 스위치 트랜지스터

도면

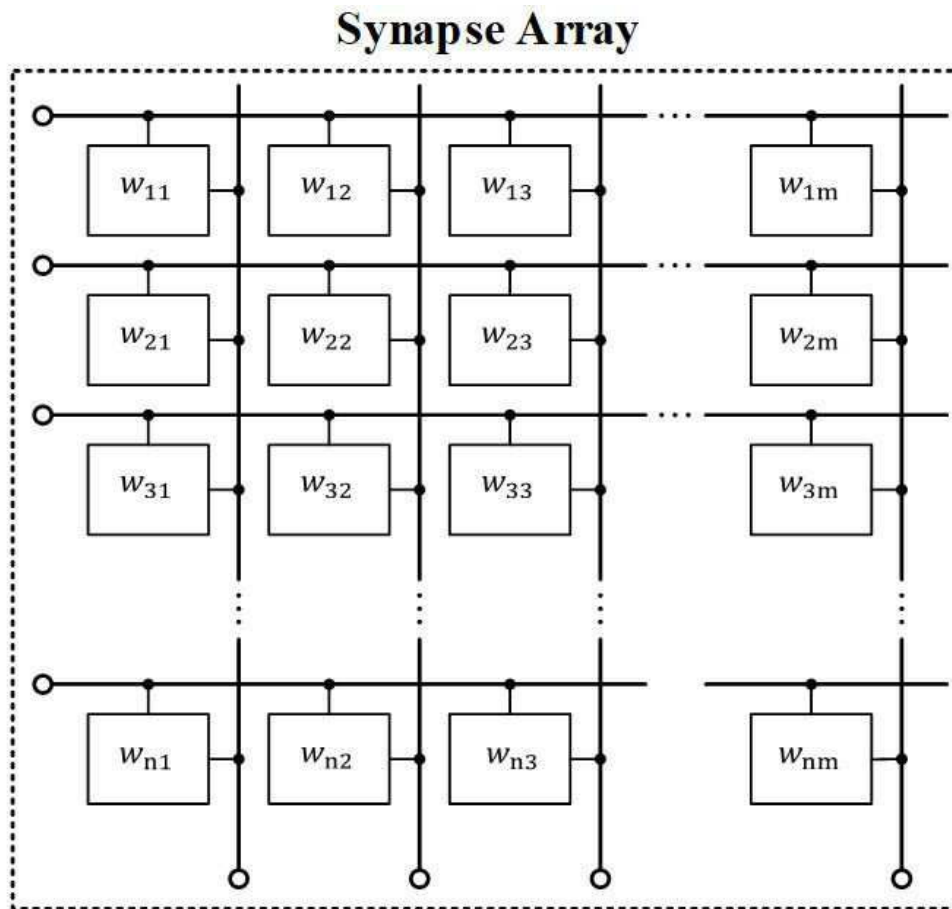
도면1



도면2

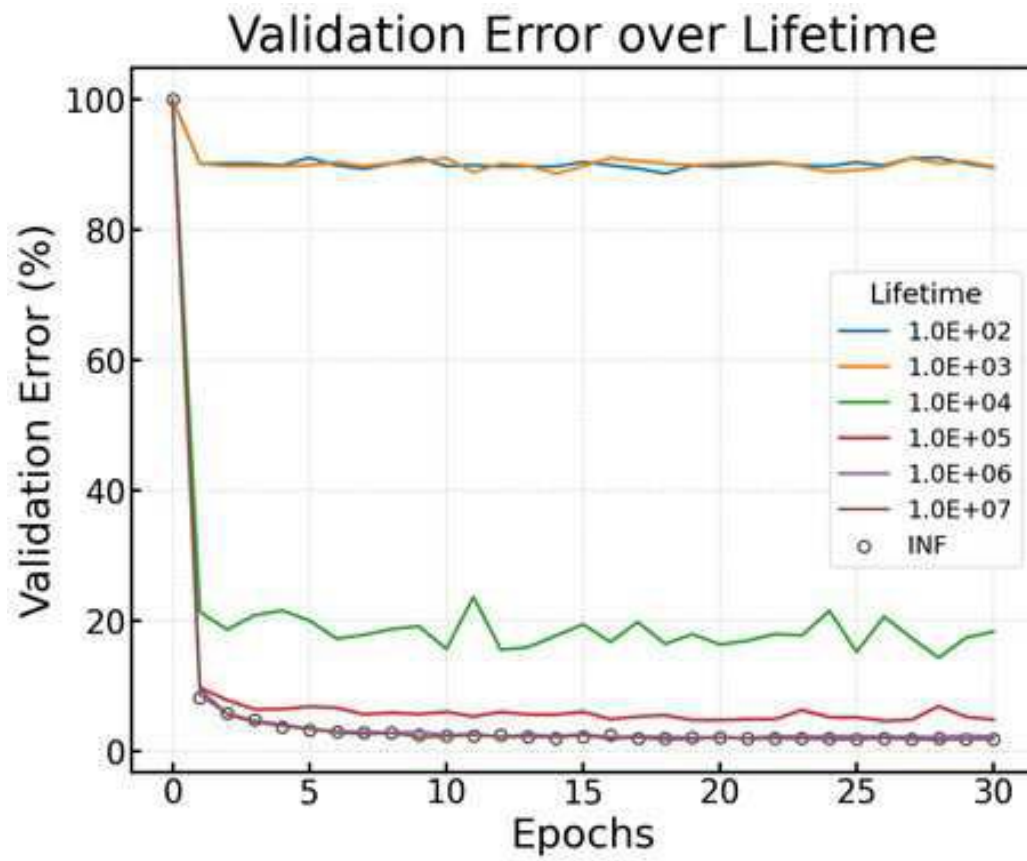


도면3

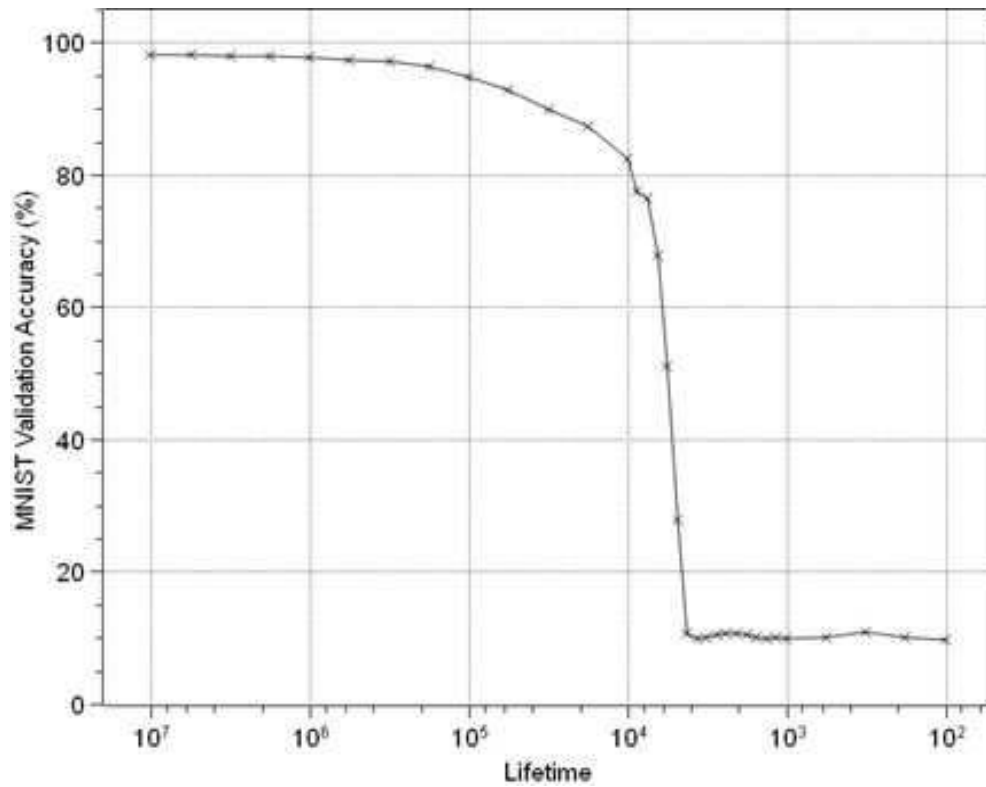




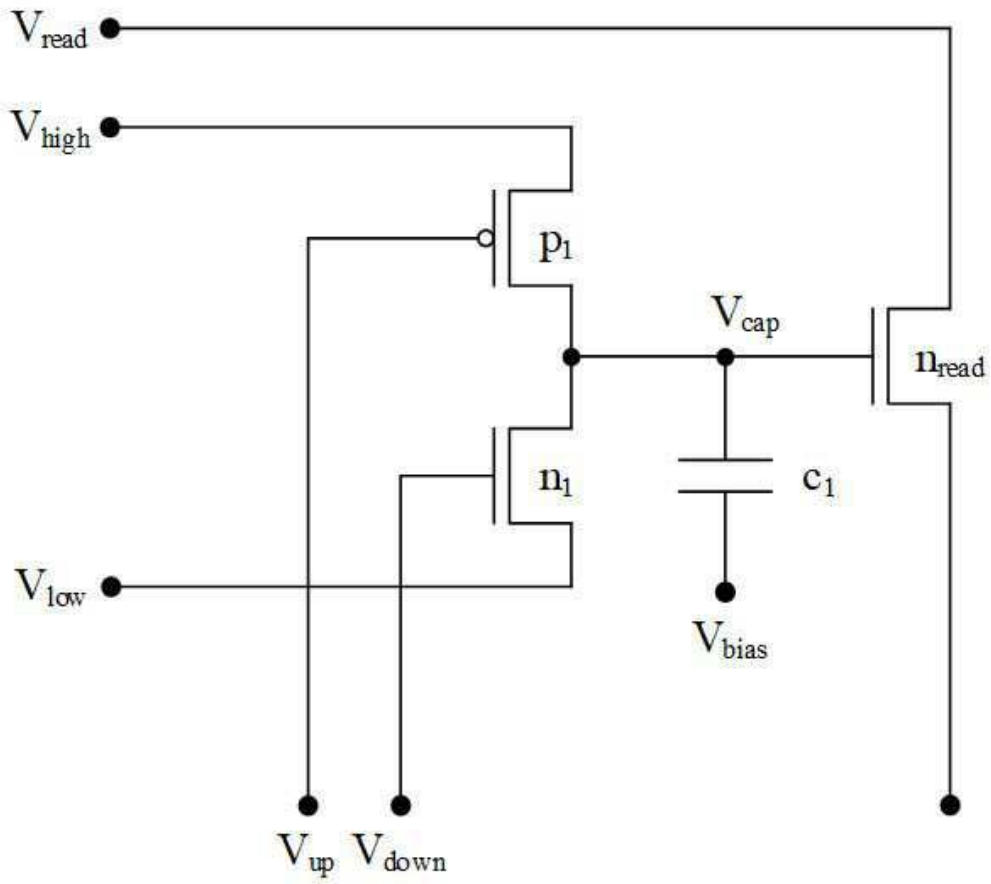
도면4



도면5

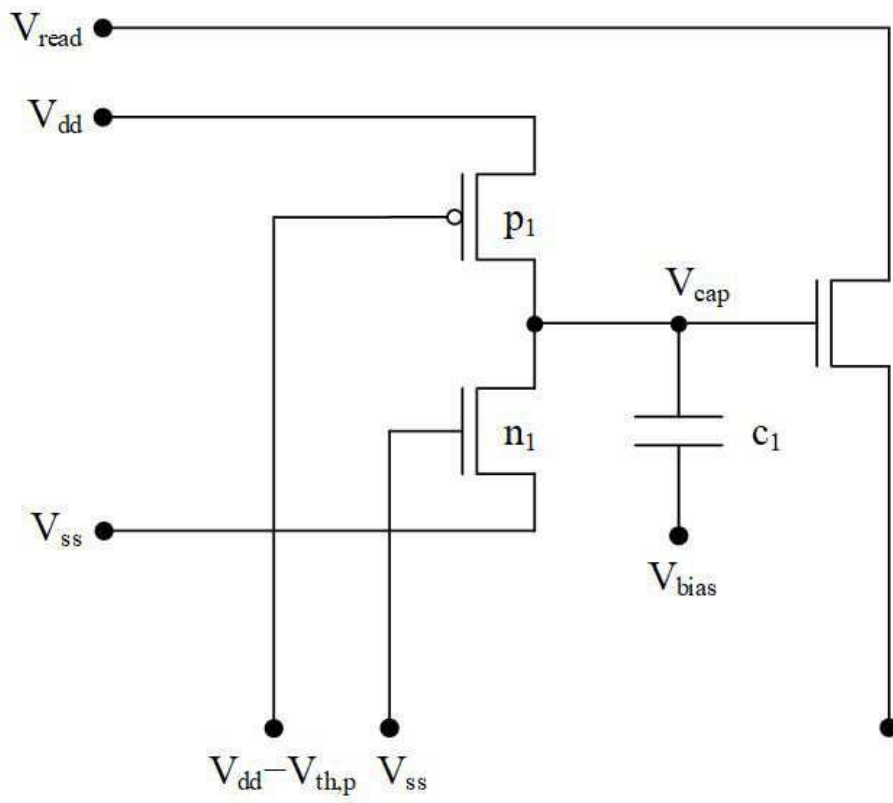


도면6

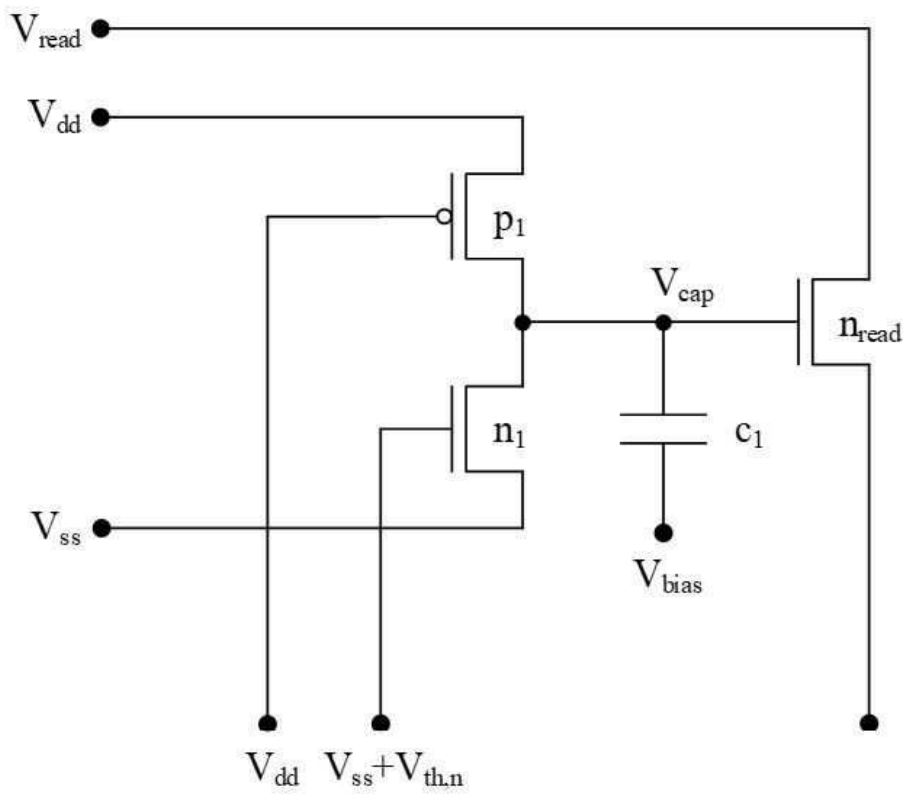




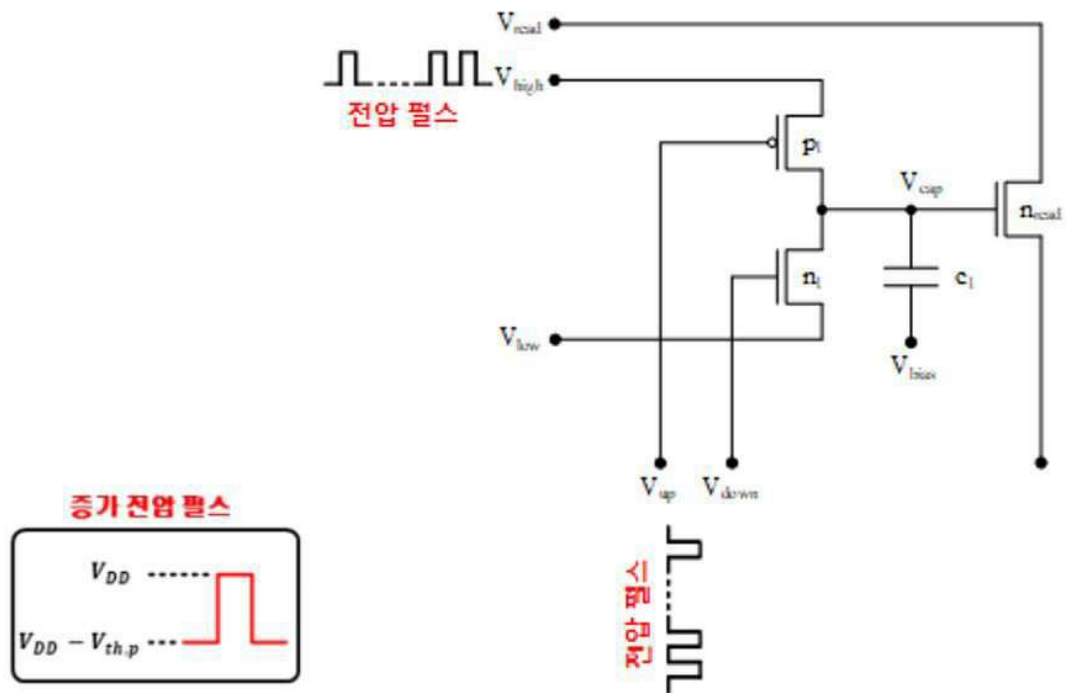
도면7



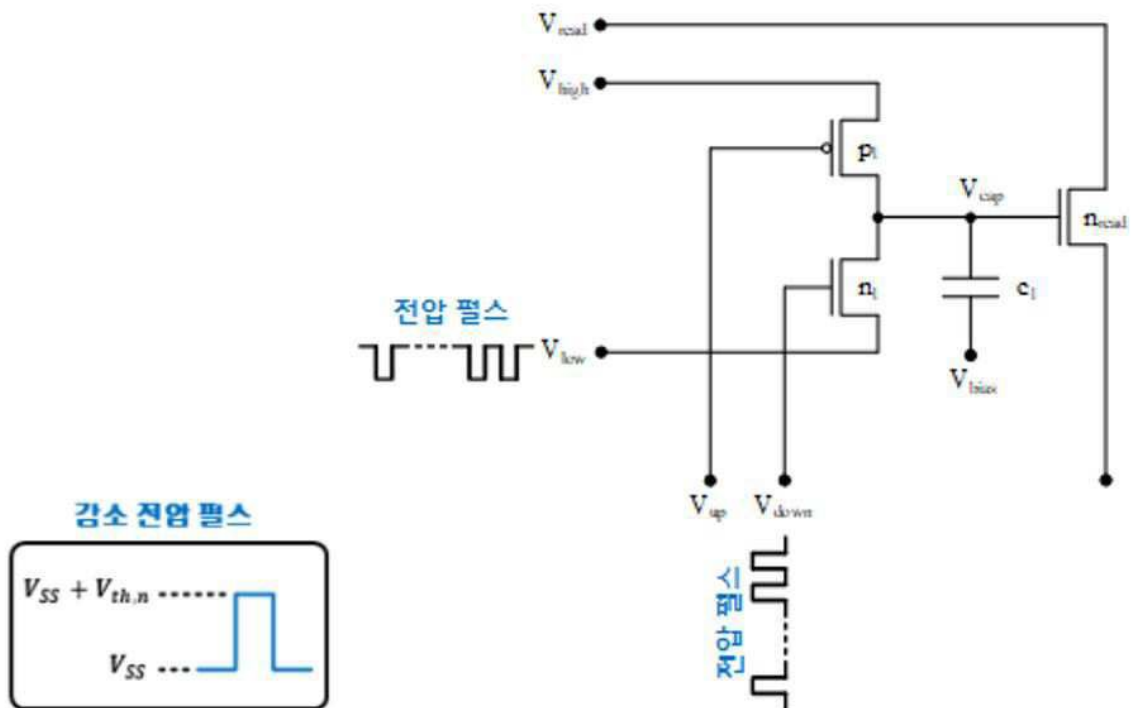
도면8



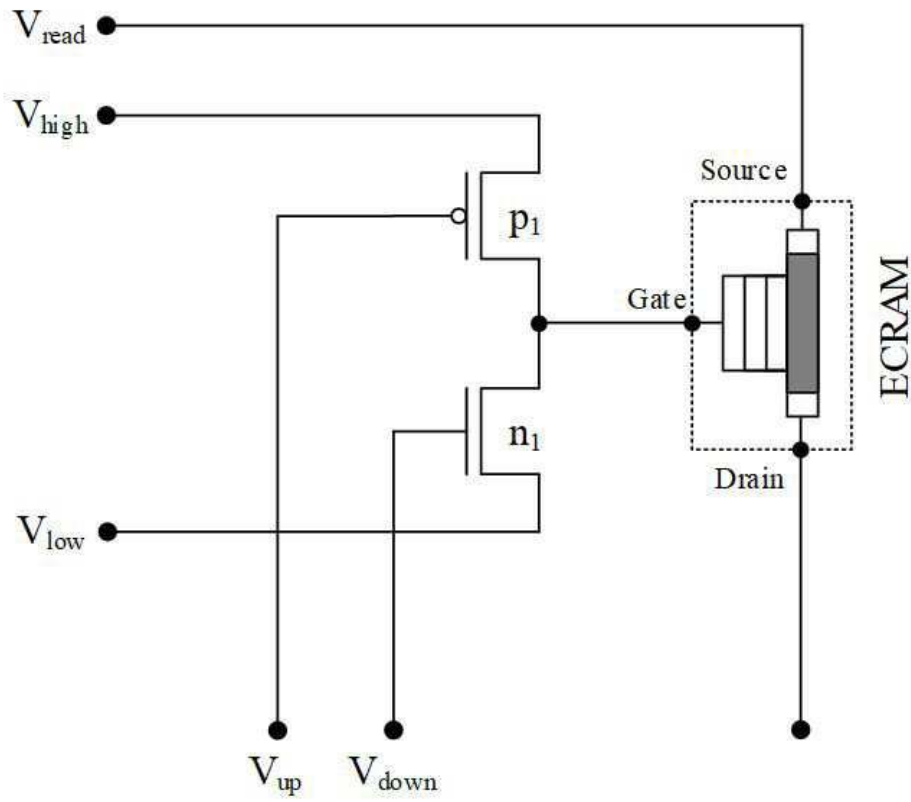
도면9



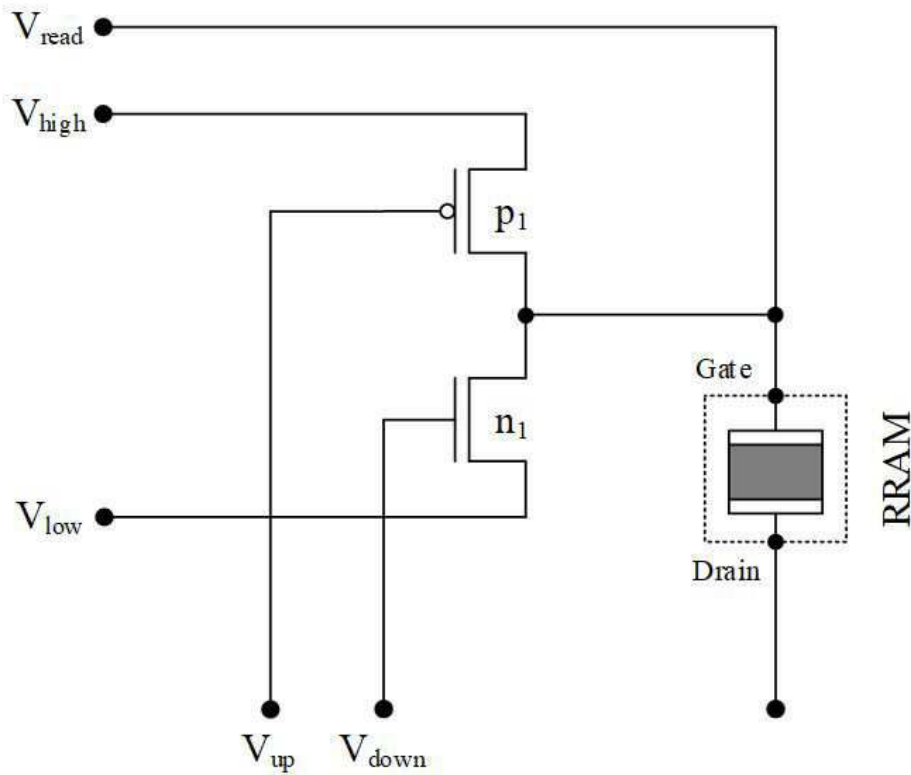
도면10



도면11



도면12



도면13

