(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0122842
(43) 공개일자 2023년08월22일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2023.01) *G06N 3/04* (2023.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/045 (2023.01)
(21) 출원번호 10-2022-0019570
(22) 출원일자 2022년02월15일
심사청구일자 2022년02월15일

(71) 출원인
서울시립대학교 산학협력단
서울특별시 동대문구 서울시립대로 163 (전농동, 서울시립대학교내)
인천대학교 산학협력단
인천광역시 연수구 아카데미로 119 (송도동)
(72) 발명자
김 윤
서울특별시 중랑구 신내로7나길 24, 211동 2002호 (상봉동, 건영2차아파트)
김정남
서울특별시 동대문구 망우로18다길 44, 202호 (휘경동)
구민석
인천광역시 연수구 먼우금로 19, 103동 1409호 (동춘동, 동남아파트)
(74) 대리인
권오준

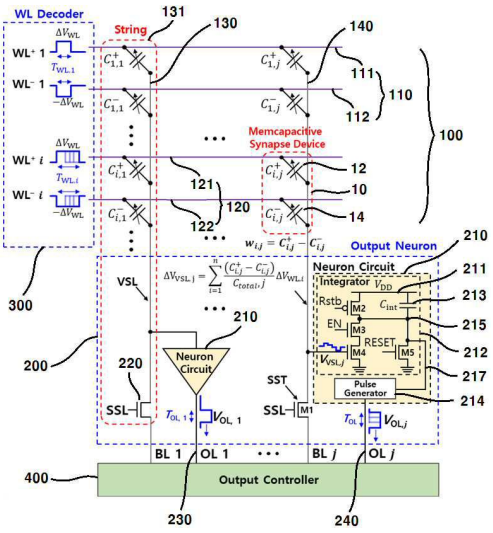
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **용량성 커플링 기반의 인공 신경망 시스템**

(57) 요 약

본 발명은 용량성 커플링 기반의 인공 신경망 시스템에 관한 것으로, 2개의 memcapacitive 소자로 하나의 시냅스 소자를 구성하고, 상기 2개의 memcapacitive 소자에 반전된 2개의 전압 변위($\pm\Delta V_{WL}$)가 각각 더해져 인가되고, 상기 2개의 memcapacitive 소자의 용량성 결합을 통하여 양과 음의 가중치로 벡터-행렬 곱 연산을 통해 VSL에 전압 합산되고, 상기 VSL의 전압으로 발화용 커패시터를 방전시켜 출력 신호가 발생되도록 함으로써, 신경망의 규모가 증가하더라도 종래 컨덕턴스 기반의 뉴로모픽 시스템보다 전력 소모를 획기적으로 줄일 수 있는 효과가 있다.

대 표 도 - 도3



(52) CPC특허분류
G06N 3/082 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711138718
과제번호 2021-0-01764-001
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 신개념 PIM반도체 선도기술개발
연구과제명 전하저장형 메모리 기반 PIM 개발
기 여 율 1/1
과제수행기관명 한국과학기술원
연구기간 2021.05.01 ~ 2024.12.31

명 세 서

청구범위

청구항 1

메모리와 캐패시터가 결합된 memcapacitive 소자를 사이에 두고 복수 개의 워드 라인과 복수 개의 비트 라인이 서로 교차하며 배치된 시냅스 어레이;

상기 시냅스 어레이의 출력측에 구비된 Output Neuron;

상기 시냅스 어레이의 입력측에 구비된 WL Decoder; 및

상기 Output Neuron의 출력측에 구비된 Output Controller를 포함하되,

상기 시냅스 어레이는 상기 복수 개의 워드 라인 중 선택된 2개의 워드 라인과 상기 복수 개의 비트 라인 중 VSL(voltage summed line)으로 선택된 1개의 비트 라인 사이에 배치된 2개의 memcapacitive 소자로 하나의 시냅스 소자를 구성하도록 형성되고,

상기 WL Decoder는 상기 선택된 2개의 워드 라인에 반전된 2개의 전압 변위($\pm \triangle V_W$)가 각각 더해져 인가되도록 구비되고,

상기 시냅스 소자는 상기 2개의 memcapacitive 소자의 용량성 결합을 통하여 양과 음의 가중치로 상기 VSL의 전압에 영향을 주도록 구비되고,

상기 Output Neuron은 발화용 커패시터(C_{int})를 포함하는 뉴런 회로를 가지고 상기 VSL의 전압으로 상기 발화용 커패시터를 방전시켜 출력 신호가 생성되도록 구비된 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템.

청구항 2

제 1 항에 있어서,

상기 WL Decoder는 외부 자극이 입력되는 동안 상기 2개의 전압 변위가 각각 더해진 상승 전압과 하강 전압을 생성하여 상기 선택된 2개의 워드 라인에 입력 신호로 동시에 인가하고, 상기 외부 자극의 강도는 상기 2개의 전압 변위의 각 펄스 유지시간(duration)을 조절하여 상기 입력 신호의 세기(intensity)로 반영한 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템.

청구항 3

제 1 항 또는 제 2 항에 있어서,

상기 Output Neuron은 상기 VSL와 상기 Output Controller 사이에 배치되어 상기 VSL에 상기 Output Controller로부터 제어 전압을 공급하거나 차단하는 SST(string select transistor)를 더 포함하고,

상기 뉴런 회로는 상기 VSL에 연결되어 상기 제어 전압에 상기 시냅스 소자의 가중치로 계산된 입력 전압이 더해져서 상기 VSL의 전압이 되도록 구비된 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템.

청구항 4

제 3 항에 있어서,

상기 뉴런 회로는 상기 VSL의 전압이 게이트로 입력되어 상기 발화용 커패시터를 방전시키는 적분기(integrator)와 상기 적분기의 출력 전압을 입력받아 상기 출력 신호를 내는 펄스 생성기(pulse generator)를

포함하여 구성된 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템.

청구항 5

제 4 항에 있어서,

상기 적분기는 동작전압 공급선과 출력단 사이에 상기 발화용 커패시터와 병렬로 연결된 충전용 트랜지스터;

상기 출력단과 접지 사이에 직렬로 연결된 동작 제어용 트랜지스터와 자극 수신용 트랜지스터; 및

상기 출력단과 접지를 연결하는 리셋용 트랜지스터를 포함하여 구성되고,

상기 자극 수신용 트랜지스터의 게이트는 상기 VSL에 연결되고,

상기 출력단은 제어선을 통해 상기 펄스 생성기에 연결된 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템.

청구항 6

제 3 항에 있어서,

상기 memcapacitive 소자는 반도체 바디 상에 제 1 유전체층, 전하 트랩층 및 제 2 유전체층으로 순차 형성되고, 상기 제 2 유전체층 상에 적층된 상기 복수 개의 워드 라인 중 하나와 체결된 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템.

청구항 7

제 6 항에 있어서,

상기 반도체 바디에는 상기 제 1 유전체층의 양 측에 소스 및 드레인이 더 형성되거나 상기 반도체 바디 밑으로 상기 반도체 바디보다 고농도로 도핑된 제2의 도핑층이 더 형성된 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템.

청구항 8

제 3 항에 있어서,

상기 Output Neuron의 출력측과 상기 Output Controller 사이에는 제2의 시냅스 어레이와 제2의 Output Neuron이 각각 하나 이상 더 연결되고,

상기 Output Neuron의 출력측 또는 상기 제2의 Output Neuron의 출력측에는 Pulse Converter가 더 구비되어, 상기 Pulse Converter를 통하여 변환된 입력 신호가 상기 제2의 시냅스 어레이로 인가되는 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템.

청구항 9

제 5 항에 의한 용량성 커플링 기반의 인공 신경망 시스템을 동작하는 방법에 있어서,

상기 인공 신경망 시스템의 추론 연산은 시간상 초기화(initialization), 적분(integration) 및 평가(evaluation)로 구분되는 3단계로 수행되는 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템의 동작 방법.

청구항 10

제 9 항에 있어서,

상기 초기화 단계는 T_{phase0} 동안 상기 SST의 SSL(string select line)에 SST 턴온(turn-on) 전압을 인가하고, 선택된 string의 비트 라인에 소정의 전압이 인가되어 상기 VSL은 프리차지되고, 상기 충전용 트랜지스터의 게이트에는 턴온 전압이 인가되어 상기 발화용 커패시터를 상기 동작전압(V_{DD})으로 충전하고,

상기 적분 단계는 상기 초기화 단계 완료 후 T_{phase1} 동안 상기 WL Decoder를 통하여 상기 시냅스 어레이의 각 선택된 2개의 워드 라인에 상승 전압(WL^{+})과 하강 전압(WL^{-})이 동시에 인가되어 상기 시냅스 소자의 가중치에 의한 벡터-행렬 곱 연산 수행 결과로 얻어진 입력 전압이 상기 프리차지에 더하여 상기 선택된 string의 VSL 전압이 결정되고, 상기 동작 제어용 트랜지스터의 게이트에 턴온 전압으로 EN 신호가 인가된 상태에서 상기 VSL 전압으로 자극 수신용 트랜지스터를 턴온시키며 상기 발화용 커패시터를 방전시키고,

상기 평가 단계는 상기 적분 단계 완료 후 T_{phase2} 동안 상기 리셋용 트랜지스터의 게이트에 턴온 전압으로 RESET 신호가 입력되어 상기 발화용 커패시터를 균일한 속도로 방전시키고,

상기 적분 단계 또는 상기 평가 단계에서 상기 출력단의 전압이 떨어지면서 상기 발화용 커패시터의 양단 전위차가 $V_{VDD} - V_{TH}$ 이상이 될 경우(V_{TH}는 펄스 생성기의 문턱 전압), 상기 펄스 생성기에서 상기 출력 신호를 내보내는 것을 특징으로 하는 용량성 커플링 기반의 인공 신경망 시스템의 동작 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공 신경망 시스템에 관한 것으로, 더욱 상세하게는 메모리와 커패시터가 결합된 memcapacitive 소자(synapse device)를 기반으로 용량성 커플링(capacitive coupling)을 이용하여 벡터-행렬 곱을 수행하는 인공 신경망 하드웨어 시스템에 관한 것이다.

배 경 기 술

[0002] 기계학습 분야가 성장함에 따라 인공 신경망(artificial neural network)을 활용하는 사례가 많아지고 있다. 특히, 이미지 인식 또는 음성 인식과 같은 복잡한 인지연산을 필요로 하는 분야에서, 도 1과 같이 하나 이상의 은닉 층(hidden layer)이 존재하는 심층 신경망(deep neural network)이 주로 사용되고 있다.

[0003] 인공 신경망을 종래의 폰 노이만(von Neumann) 아키텍처를 기반으로 설계할 경우, 폰 노이만 병목현상으로 인하여 속도 저하와 에너지 소모가 증가하는 문제가 알려져 있다.

[0004] 위와 같은 폰 노이만 병목현상으로 발생하는 문제점을 개선하기 위한 대안 중 하나로 생물학적 신경계를 하드웨어로 모방한 뉴로모픽(neuromorphic) 시스템이 활발히 연구되고 있다. 뉴로모픽 시스템은 하드웨어로 모사한 복수 개의 뉴런(neuron)이 입력 층, 은닉 층, 출력 층에 존재하고, 하드웨어 시냅스(synapse)를 통하여 각 층의 뉴런들이 연결되어 있다.

[0005] 폰 노이만 아키텍처를 기반으로 하는 디지털 컴퓨팅 시스템은 입력 신호와 시냅스 가중치(synaptic weight)값을 순차적으로 계산한다. 이에 반하여 뉴로모픽 시스템은 벡터(vector) 형태로 입력 신호가 모든 입력 층에 병렬적으로 인가되고, 시냅스 가중치 행렬(matrix)의 벡터-행렬 곱(vector-matrix multiplication) 연산이 아날로그 기반으로 동시에 수행된다. 즉, 뉴로모픽 시스템은 인공신경망에서 각층의 벡터-행렬 곱 연산을 병렬적으로 동시에 수행하기 때문에 동작 에너지 및 시간을 감소시킬 수 있다.

[0006] 통상적인 뉴로모픽 시스템에서 시냅스는 주로 메모리 소자를 통하여 구현되고, 뉴런은 CMOS 회로를 통하여 구현된다. 생물학적 시냅스는 전 뉴런의 축색돌기와 후 뉴런의 수상돌기 사이의 접합부를 의미하며, 신경전달물질의 분비 및 흡착을 통하여 전기적인 신호를 전달하는 역할을 수행한다. 이 때 시냅스가 가지는 연결강도에 따라 후 뉴런으로 전달되는 전기적 신호의 크기가 조정된다. 이러한 시냅스의 연결강도는 시냅스 가중치라 불리며, 생물학적 시냅스 가중치는 학습 과정(learning process)에 의하여 조절된다.

[0007] 메모리 소자로 시냅스를 구현할 경우, 시냅스 가중치는 메모리 소자의 컨덕턴스(conductance)로 표현된다. 생물

학적 시냅스가 갖는 음의 시냅스 가중치를 메모리 소자로 표현하기 위하여 시냅스 가중치는 두 메모리 소자의 컨덕턴스 차로 나타낸다($w = G^{+} - G^{-}$). 이 때, SRAM(static random-access memory), RRAM(resistive random-access memory), PCM(phase-change memory), STT-MRAM(spin-transfer torque magnetoresistive random-access memory), FG(floating-gate) memory, CTF(charge-trap flash) memory와 같은 메모리 소자를 이용하여 시냅스의 동작 특성을 구현 할 수 있다.

[0008] 도 2는 종래의 컨덕턴스 기반 시냅스 소자를 이용한 뉴로모픽 시스템을 나타낸 것이다. 종래의 뉴로모픽 시스템은 시냅스 소자가 각각 분리된 column line을 통하여 뉴런회로와 연결되어 있다. 컨덕턴스 기반 시냅스 어레이를 통해 추론(inference) 연산을 수행할 경우, 전압의 형태로 입력 벡터(x_1, …, x_i, …, x_n)가 시냅스 가중치 행렬로 입력된다. 가중치 행렬에 존재하는 각각의 시냅스 소자에서 입력 전압(x_i)과 컨덕턴스(G_{ij}^{+}, G_{ij}^{-})의 곱 연산이 수행되어(I_{ij}^{+}, I_{ij}^{-}), 개별 시냅스 소자와 연결된 column line을 통하여 합산된다(I_j^{+}, I_j^{-}). 각각의 column line으로 합산된 벡터-행렬 곱 연산 결과는 뉴런회로에서 차 연산이 수행되어 뉴런코어로 입력된다. 이와 같은 방식으로 구현되는 신경 연결망에 관한 기술이 한국 등록특허 제10-2126791호로 개시되어 있다.

[0009] 상기 컨덕턴스 기반 시냅스 어레이에서는 '0'의 시냅스 가중치를 갖는 경우에도 각각의 column line으로 동일한 전류가 흐르기 때문에 일정 수준의 전력 소모가 계속 발생한다는 단점이 있다. 또한 많은 수의 fan-in이 존재하는 경우 많은 전력 소모를 일으키는 원인이 된다. 인공 신경망의 은닉 층이 증가할수록 신경망에 필요한 시냅스의 개수가 증가하여 이러한 단점은 더욱 중요해진다.

발명의 내용

해결하려는 과제

[0010] 본 발명은 종래 컨덕턴스 기반 인공 신경망이 갖는 전력 소모의 문제점을 해소하기 위해, 특히 신경망의 규모에 따라 급격하게 증가하는 경우에도 에너지 소모를 획기적으로 줄일 수 있는 용량성 커플링 기반의 인공 신경망 시스템 및 그 동작 방법을 제공하고자 한다.

과제의 해결 수단

[0011] 상기 목적을 달성하기 위하여, 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템은 메모리와 캐패시터가 결합된 memcapacitive 소자를 사이에 두고 복수 개의 워드 라인과 복수 개의 비트 라인이 서로 교차하며 배치된 시냅스 어레이; 상기 시냅스 어레이의 출력측에 구비된 Output Neuron; 상기 시냅스 어레이의 입력측에 구비된 WL Decoder; 및 상기 Output Neuron의 출력측에 구비된 Output Controller를 포함하되, 상기 시냅스 어레이는 상기 복수 개의 워드 라인 중 선택된 2개의 워드 라인과 상기 복수 개의 비트 라인 중 VSL(voltage summed line)으로 선택된 1개의 비트 라인 사이에 배치된 2개의 memcapacitive 소자로 하나의 시냅스 소자를 구성하도록 형성되고, 상기 WL Decoder는 상기 선택된 2개의 워드 라인에 반전된 2개의 전압 변위($\pm \Delta V_{WL}$)가 각각 더해져 인가되도록 구비되고, 상기 시냅스 소자는 상기 2개의 memcapacitive 소자의 용량성 결합을 통하여 양과 음의 가중치로 상기 VSL의 전압에 영향을 주도록 구비되고, 상기 Output Neuron은 발화용 커패시터(C_{int})를 포함하는 뉴런 회로를 가지고 상기 VSL의 전압으로 상기 발화용 커패시터를 방전시켜 출력 신호가 생성되도록 구비된 것을 특징으로 한다.

[0012] 상기 WL Decoder는 외부 자극이 입력되는 동안 상기 2개의 전압 변위가 각각 더해진 상승 전압과 하강 전압을 생성하여 상기 선택된 2개의 워드 라인에 입력 신호로 동시에 인가하고, 상기 외부 자극의 강도는 상기 2개의 전압 변위의 각 펄스 유지시간(duration)을 조절하여 상기 입력 신호의 세기(intensity)로 반영한 것을 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템의 다른 특징으로 한다.

[0013] 상기 Output Neuron은 상기 VSL와 상기 Output Controller 사이에 배치되어 상기 VSL에 상기 Output Controller로부터 제어 전압을 공급하거나 차단하는 SST(string select transistor)를 더 포함하고, 상기 뉴런 회로는 상기 VSL에 연결되어 상기 제어 전압에 상기 시냅스 소자의 가중치로 계산된 입력 전압이 더해져서 상기 VSL의 전압이 되도록 구비된 것을 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템의 다른 특징으로 한다.

[0014] 상기 뉴런 회로는 상기 VSL의 전압이 게이트로 입력되어 상기 발화용 커패시터를 방전시키는 적분기

(integrator)와 상기 적분기의 출력 전압을 입력받아 상기 출력 신호를 내는 펄스 생성기(pulse generator)를 포함하여 구성된 것을 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템의 다른 특징으로 한다.

[0015] 상기 적분기는 동작전압 공급선과 출력단 사이에 상기 발화용 커패시터와 병렬로 연결된 충전용 트랜지스터; 상기 출력단과 접지 사이에 직렬로 연결된 동작 제어용 트랜지스터와 자극 수신용 트랜지스터; 및 상기 출력단과 접지를 연결하는 리셋용 트랜지스터를 포함하여 구성되고, 상기 자극 수신용 트랜지스터의 게이트는 상기 VSL에 연결되고, 상기 출력단은 제어선을 통해 상기 펄스 생성기에 연결된 것을 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템의 다른 특징으로 한다.

[0016] 상기 memcapacitive 소자는 반도체 바디 상에 제 1 유전체층, 전하 트랩층 및 제 2 유전체층으로 순차 형성되고, 상기 제 2 유전체층 상에 적층된 상기 복수 개의 워드 라인 중 하나와 체결된 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템의 다른 특징으로 한다.

[0017] 상기 반도체 바디에는 상기 제 1 유전체층의 양 측에 소스 및 드레인이 더 형성되거나 상기 반도체 바디 밑으로 상기 반도체 바디보다 고농도로 도핑된 제2의 도핑층이 더 형성된 것을 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템의 다른 특징으로 한다.

[0018] 상기 Output Neuron의 출력측과 상기 Output Controller 사이에는 제2의 시냅스 어레이와 제2의 Output Neuron이 각각 하나 이상 더 연결되고, 상기 Output Neuron의 출력측 또는 상기 제2의 Output Neuron의 출력측에는 Pulse Converter가 더 구비되어, 상기 Pulse Converter를 통하여 변환된 입력 신호가 상기 제2의 시냅스 어레이로 인가되는 것을 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템의 다른 특징으로 한다.

[0019] 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템의 동작 방법은 상술한 용량성 커플링 기반의 인공 신경망 시스템에서 인공 신경망 시스템의 추론 연산은 시간상 초기화(initialization), 적분(integration) 및 평가(evaluation)로 구분되는 3단계로 수행되는 것을 특징으로 한다.

[0020] 상기 초기화 단계는 T_{phase0} 동안 상기 SST의 SSL(string select line)에 SST 턴온(turn-on) 전압을 인가하고, 선택된 string의 비트 라인에 소정의 전압이 인가되어 상기 VSL은 프리차지되고, 상기 충전용 트랜지스터의 게이트에는 턴온 전압이 인가되어 상기 발화용 커패시터를 상기 동작전압(V_{DD})으로 충전하고, 상기 적분 단계는 상기 초기화 단계 완료 후 T_{phase1} 동안 상기 WL Decoder를 통하여 상기 시냅스 어레이의 각 선택된 2개의 워드 라인에 상승 전압(WL^{+})과 하강 전압(WL^{-})이 동시에 인가되어 상기 시냅스 소자의 가중치에 의한 벡터-행렬 곱 연산 수행 결과로 얻어진 입력 전압이 상기 프리차지에 더하여 상기 선택된 string의 VSL 전압이 결정되고, 상기 동작 제어용 트랜지스터의 게이트에 턴온 전압으로 EN 신호가 인가된 상태에서 상기 VSL 전압으로 자극 수신용 트랜지스터를 턴온시키며 상기 발화용 커패시터를 방전시키고, 상기 평가 단계는 상기 적분 단계 완료 후 T_{phase2} 동안 상기 리셋용 트랜지스터의 게이트에 턴온 전압으로 RESET 신호가 입력되어 상기 발화용 커패시터를 균일한 속도로 방전시키고, 상기 적분 단계 또는 상기 평가 단계에서 상기 출력단의 전압이 떨어지면서 상기 발화용 커패시터의 양단 전위차가 $V_{VDD} - V_{TH}$ 이상이 될 경우(V_{TH}는 펄스 생성기의 문턱 전압), 상기 펄스 생성기에서 상기 출력 신호를 내보내는 것을 본 발명에 의한 용량성 커플링 기반의 인공 신경망 시스템 동작 방법의 다른 특징으로 한다.

발명의 효과

[0021] 본 발명은 2개의 memcapacitive 소자로 하나의 시냅스 소자를 구성하고, 상기 2개의 memcapacitive 소자에 반전된 2개의 전압 변위($\pm \Delta V_{WL}$)가 각각 더해져 인가되고, 상기 2개의 memcapacitive 소자의 용량성 결합을 통하여 양과 음의 가중치로 벡터-행렬 곱 연산을 통해 VSL에 전압 합산되고, 상기 VSL의 전압으로 발화용 커패시터를 방전시켜 출력 신호가 발생되도록 함으로써, 신경망의 규모가 증가하더라도 종래 컨덕턴스 기반의 뉴로모픽 시스템보다 전력 소모를 획기적으로 줄일 수 있는 효과가 있다.

도면의 간단한 설명

[0022] 도 1은 심층 신경망 구조를 보인 개념도이다.

도 2는 종래의 컨덕턴스 기반 시냅스 소자를 이용한 뉴로모픽 시스템을 보여주는 회로도이다.

도 3은 본 발명의 일 실시예에 의한 용량성 커플링 기반의 인공 신경망 시스템 구조를 보여주는 회로도이다.

도 4는 도 3의 용량성 커플링 기반의 인공 신경망 시스템의 추론 시 동작 과정으로 보여주는 시간에 따른 전원 인가도이다.

도 5는 본 발명의 일 실시예에 시냅스 소자로 사용될 memcapacitive 소자들의 분류 계통도이다.

도 6 내지 도 9는 각각 본 발명의 일 실시예에 시냅스 소자로 사용되는 CTF 기반 memcapacitive 소자의 구조를 보인 단면도로, (A)는 저농도로 도핑된 반도체 바디 상에 형성된 구조를 보인 것이고, (B)는 (A)에 소스 및 드레인이 추가된 구조를 보인 것이다.

도 10은 본 발명의 일 실시예에 시냅스 소자로 사용되는 CTF 기반 memcapacitive 소자의 바디 도핑 극성에 따른 멀티레벨 전기용량 변화를 보인 전기 특성도이다.

도 11은 본 발명의 다른 실시예에 의한 용량성 커플링 기반의 인공 신경망 시스템 구조를 보여주는 회로도이다.

발명을 실시하기 위한 구체적인 내용

[0023] 이하, 첨부된 도면을 참조하며 본 발명의 바람직한 실시예에 대하여 설명한다.

[0024] 본 발명의 일 실시예에 의한 용량성 커플링 기반의 인공 신경망 시스템은, 도 3에 예시된 바와 같이, 메모리와 캐패시터가 결합된 memcapacitive 소자(12 또는 14)를 사이에 두고 복수 개의 워드 라인(111, 112, 121, 122)과 복수 개의 비트 라인(130, 140)이 서로 교차하며 배치된 시냅스 어레이(100); 상기 시냅스 어레이의 출력측에 구비된 Output Neuron(200); 상기 시냅스 어레이의 입력측에 구비된 WL(word line) Decoder(300); 및 상기 Output Neuron의 출력측에 구비된 Output Controller(400)를 포함하여 구성된다.

[0025] 상기 시냅스 어레이(100)는 상기 복수 개의 워드 라인 중 선택된 2개의 워드 라인(예를 들어, 121, 122)과 상기 복수 개의 비트 라인 중 VSL(voltage summed line)으로 선택된 1개의 비트 라인(예를 들어, 140) 사이에 배치된 2개의 memcapacitive 소자(12, 14)로 하나의 시냅스 소자(10)를 구성하도록 형성된다.

[0026] 상기 WL Decoder(300)는 상기 선택된 2개의 워드 라인(121, 122)에 반전된 2개의 전압 변위($\pm\Delta V_{WL}$)가 각각 더해져 인가되도록 구비된다. 즉, 상기 WL Decoder(300)는 하나의 워드 라인(120)이 2개의 워드 라인(121, 122)으로 나누어 각각 반전된 2개의 전압 변위($\pm\Delta V_{WL}$)가 더해진 값이 인가되도록 구비된다.

[0027] 따라서, 상기 시냅스 소자(10)는 상기 2개의 memcapacitive 소자(12, 14)의 용량성 결합을 통하여 양과 음의 가중치($C_{ij}^{\pm}$)로 상기 VSL(140)의 전압에 영향을 주도록 구비될 수 있다.

[0028] 그리고, 상기 Output Neuron(200)은 발화용 커패시터(C_{int}, 213)를 포함하는 뉴런 회로(210)를 가지고 상기 VSL(140)의 전압으로 상기 발화용 커패시터(213)를 방전시켜 출력 신호가 생성되도록 구비될 수 있다.

[0029] 구체적으로, 상기 WL Decoder(300)는, 아래 수학식 1과 같이, 외부 자극이 입력되는 동안 상기 2개의 전압 변위($\pm\Delta V_{WL}$)가 각각 더해진 상승 전압($V_{ref}+\Delta V_{WL}$)과 하강 전압($V_{ref}-\Delta V_{WL}$)을 생성하여 상기 선택된 2개의 워드 라인(121, 122)에 입력 신호로 동시에 인가하게 구비될 수 있다.

수학식 1

[0030]
$$V_{WL_i^{\pm}} = \begin{cases} V_{ref} & , T > T_{WL_i} \\ V_{ref} \pm \Delta V_{WL} & , T \leq T_{WL_i} \end{cases}$$

[0031] 여기서, V_{ref}는 0 또는 임의 값으로 자극과 무관한 기준 전압을 말한다.

[0032] 상기 외부 자극의 강도는 상기 2개의 전압 변위($\pm\Delta V_{WL}$)의 각 펄스 유지시간(duration), 즉 수학식 1에서 T_{WLi}를 조절하여 입력 신호의 세기(intensity)로 반영할 수 있다.

[0033] 상기 Output Neuron(200)은 각 비트 라인을 따라 연결된 시냅스 소자들의 String(131)을 선택하기 위한 복수 개

의 SST(string select transistor, 220)를 더 포함할 수 있다. 여기서, 각 SST(220)는 각 VSL(130)과 Output Controller(400) 사이에 배치되어 상기 VSL(130)에 상기 Output Controller(400)로부터 제어 전압(예컨대, 후술하는 프리차지 전압)을 공급하거나 차단하도록 구비될 수 있다. 이 때, 상기 VSL(130)의 전압은 상기 제어 전압에 상기 시냅스 소자의 가중치로 계산된 입력 전압이 더해지게 된다. 또한, 상기 뉴런 회로(210)는 상기 VSL(130)에 연결되어 상기 VSL(130)의 전압으로 발화용 커패시터(213)의 방전을 제어하도록 구비될 수 있다.

[0034] 상기 Output Controller(400)는 Output Neuron(200)과 연결되어, BL(bit line) 전압을 생성하여 상기 복수 개의 SST(220 등)를 통해 각 VSL(130)에 인가하거나, 도 3과 같이, 뉴런 회로(210)로부터 출력된 OL(output line; 230, 240) 신호를 입력 받아 처리하도록 구비될 수 있다.

[0035] 또한, 상기 Output Controller(400) 내에는 SSL과 BL drivers 및 program verification circuit와 같은 주변 회로들이 더 포함될 수 있고, 추론 연산 수행 또는 시냅스 가중치 학습 시 필요한 BL 전압을 생성하여 각 VSL(130)에 인가하도록 구비될 수 있다.

[0036] 상술한 실시예로 C^2-ANN(capacitive coupling artificial neural network)을 통한 전압 합산 기반의 벡터-행렬 곱 연산으로 각 VSL(130)의 전압을 구할 수 있게 된다. 이는 도 3을 참조하면 다음 과정을 통해 수행될 수 있다.

[0037] 먼저, Output Controller(400)를 통하여 SSL에 전압을 인가하여 선택된 String(131)의 SST(220)를 턴온(turn-on) 한 이후, SST의 일단과 연결된 BL에 프리차지 전압($V_{\text{precharge}}$)을 인가한다. VSL(130)과 BL 1의 전압이 동일해진 이후 SSL에 턴오프(turn-off, 예컨대 0V)를 인가하여 SST(220)를 turn-off하여 VSL(130)을 플로팅(floating) 상태로 만든다.

[0038] 이후, WL Decoder(300)를 통하여 WL^+ 및 WL^- 입력 신호를 시냅스 어레이(100)로 입력한다. 이 때, 입력 가능한 WL^+ 및 WL^- 전압은 각각 양과 음의 전압에 국한되지 않고 사용 조건에 따라 다양하게 구성할 수 있다. 일례로 0 V를 WL 기준 전압(V_{ref})으로 사용할 경우, 추론 동작 시 i번째 WL^+ 및 WL^-에 각각 $V_{\text{WL},i}$ 및 $-V_{\text{WL},i}$을 인가한다. 이 때, WL과 시냅스 소자의 용량성 결합에 의하여 j번째 string(140)에 위치한 VSL 전압은 아래 수학식 2와 같이 변하게 된다.

수학식 2

[0039]
$$V_{\text{VSL},j} = V_{\text{precharge}} + \sum_{i=1}^{N}\left\{\left(\frac{C_{i,j}^{+}}{C_{\text{total},j}}\right)V_{\text{WL},i}\right\} + \sum_{i=1}^{N}\left\{\left(\frac{C_{i,j}^{-}}{C_{\text{total},j}}\right)\left(-V_{\text{WL},i}\right)\right\}$$
$$= V_{\text{precharge}} + \sum_{i=1}^{N}\left(\frac{C_{i,j}^{+} - C_{i,j}^{-}}{C_{\text{total},j}}\right)\cdot V_{\text{WL},i}$$

[0040] 여기서, $C_{\text{total},j}$은 j번째 string(140)의 VSL의 전체 커패시턴스(전기용량)으로, 기생 전기용량 및 VSL과 연결된 모든 시냅스의 전기용량을 포함한다. $C_{i,j}^{+}$와 $C_{i,j}^{-}$는 각각 i번째 WL과 j번째 string에 위치한 양과 음의 시냅스 가중치를 전기용량으로 표현한 값이다.

[0041] 또 다른 일례로, WL 기준 전압을 임의의 전압 V_{ref}를 사용할 수도 있다. 이 경우, 벡터-행렬 곱 연산 시 i번째 WL^+ 및 WL^-에 인가되는 전압을 각각 $V_{ref}+\Delta V_{\text{WL},i}$, $V_{ref}-\Delta V_{\text{WL},i}$ 라고 할 경우, 아래 수학식 3을 통해 j번째 string(140)에 위치한 VSL의 전압을 결정할 수 있다.

수학식 3

[0042] $$V_{VSL} = V_{precharge} + \sum_{i}^{N}\left(\frac{C_{i,j}^{+} - C_{i,j}^{-}}{C_{total,j}}\right) \cdot \Delta V_{WL,i}$$

[0043] 도 3을 참조하면, 상기 뉴런 회로(210)는 각 VSL(140)의 전압이 게이트로 입력되어 발화용 커패시터(213)를 방전시키는 적분기(integrator, 212)와 상기 적분기의 출력 전압(출력단 215 전압)을 입력받아 출력 신호(OL_j 240 신호)를 내는 펄스 생성기(pulse generator, 214)를 포함하여 구성될 수 있다.

[0044] 구체적으로, 도 3과 같이, 상기 적분기(212)는 동작전압(V_{DD}) 공급선(211)과 출력단(215) 사이에 상기 발화용 커패시터(213)와 병렬로 연결된 충전용 트랜지스터(M2); 상기 출력단(215)과 접지 사이에 직렬로 연결된 동작 제어용 트랜지스터(M3)와 자극 수신용 트랜지스터(M4); 및 상기 출력단(215)과 접지를 연결하는 리셋용 트랜지스터(M5)를 포함하여 구성되고, 상기 자극 수신용 트랜지스터(M4)의 게이트는 VSL(140)에 연결되고, 상기 출력단(215)은 제어선(217)을 통해 상기 펄스 생성기(214)에 연결될 수 있다. 도 3에서 충전용 트랜지스터(M2)는 p형 모스펫(MOSFET), 나머지 SST(220, M1), 동작 제어용 트랜지스터(M3), 자극 수신용 트랜지스터(M4) 및 리셋용 트랜지스터(M5)는 n형 모스펫으로 도시되어 있으나, 모스펫 타입을 서로 바꾸거나 동일 또는 유사한 스위치 특성을 갖는 소자들로 바꾸어 실시할 수 있다. 나아가, 도 3에서 뉴런 회로(210)는 일례로 보여준 것이어서, 제시한 구조나 형태에 제한되지 않는다.

[0045] 본 발명의 일 실시예에 시냅스 소자(10)로 사용될 memcapacitive 소자들(12, 14)은 메모리와 캐패시터가 결합된 기능을 갖는 것이면 이에 사용될 수 있다. 도 5는 본 발명의 일 실시예로 사용될 memcapacitive 소자들(12, 14)의 분류 계통도를 보여준다. 이에 의하면, memcapacitive 소자들은 크게 유전 분극 기반과 전하 기반으로 구분될 수 있다.

[0046] 유전 분극 기반의 memcapacitive 소자는 유전 분극에 의해 전기용량이 변하는 소자로, 대표적인 예로 FeFET(ferroelectric field effect transistor)이 있다. 유전 분극 기반의 memcapacitive 소자는 소자 내에 존재하는 강유전체 층(ferroelectric layer)의 유전 분극 방향을 정렬하여 소자의 전기용량을 조절한다.

[0047] 전하 기반의 memcapacitive 소자는 전하 저장 노드의 성질에 따라 다시 두 가지 타입으로 구분된다. 전하 저장 노드가 유전체인 경우, 이에 해당하는 대표적인 memcapacitive 소자로 CTF(charge-trap flash) 기반의 소자가 있다. CTF 기반의 memcapacitive 소자는 소자 내에 존재하는 전하 트랩층(charge trap layer)에 전자 또는 정공을 주입하여 소자의 전기용량을 조절할 수 있다. 전하 저장 노드가 도체, 혹은 반도체인 경우, 이에 해당하는 대표적인 memcapacitive 소자로 FG(floating gate) 타입의 소자가 있다. FG 타입의 경우 공정 방식에 따라, 도 5와 같이, single poly FG 소자와 dual poly FG 소자로 나눌 수 있다. FG 타입의 memcapacitive 소자는 소자 내에 존재하는 FG에 전자 또는 정공을 주입하여 소자의 전기용량을 조절할 수 있다.

[0048] 도 5는 본 발명의 시냅스 소자(10)로 사용될 memcapacitive 소자들의 일 예시이므로, 위에 언급한 종류에 한정되는 것은 아니다.

[0049] 도 6 내지 도 9는 각각 본 발명의 일 실시예에 시냅스 소자로 사용되는 CTF 기반 memcapacitive 소자의 구조를 보인 단면도로, (A)는 저농도로 도핑된 반도체 바디 상에 형성된 구조를 보인 것이고, (B)는 (A)에 소스 및 드레인이 추가된 구조를 보인 것이다.

[0050] 도 6(A)를 참조하면, 상술한 각 실시예에서 사용될 memcapacitive 소자(12)는 저농도로 도핑된 반도체 바디를 비트 라인(VSL, 140)으로 하고, 상기 반도체 바디(140) 상에 제 1 유전체층(11), 전하 트랩층(13) 및 제 2 유전체층(15)으로 순차 형성되고, 상기 제 2 유전체층(15) 상에 적층된 상기 복수 개의 워드 라인 중 하나(121)와 체결된 구조로 실시될 수 있다. 도 6(B)는 도 6(A)의 구조에서 제 1 유전체층(11)의 양 측에 소스(142a) 및 드레인(142b)이 더 추가되어 저장 트랜지스터(charge trap transistor, CTT) 형태로 실시될 수 있음을 보여준다.

[0051] 여기서, 상기 반도체 바디(140)가 저농도로 도핑되었다고 함은 통상 모스펫의 소스/드레인보다 불순물이 낮은 농도로 도핑된 것을 말하고, 극성에 따라 P^- 또는 N^-로 표현될 수 있다. 상기 제 1 유전체층(11)은 터널링 유전체로는 SiON, SiO_2, HfO_X와 같은 물질이 사용될 수 있다. 상기 전하 트랩층(13)은 SiN_X 와 같은 물질이 사용될

수 있으며, 상기 제 2 유전체층(15)은 블록킹 유전체로는 SiO_2, Al_2O_3와 같은 물질들이 사용될 수 있다. 상기 복수 개의 워드 라인 중 하나(121)는 메모리 소자의 게이트 물질 즉, 금속 또는 고농도로 도핑된(P^+ 또는 N^+) 실리콘이 사용될 수 있다. 여기서 설명한 내용은 하나의 예시이므로, 각 구성에 사용 가능한 물질은 앞서 설명한 물질들로 한정되는 것은 아니고, 복수 개의 물질이 동시에 사용될 수도 있다.

[0052] 도 7을 참조하면, 반도체 바디(141) 밑으로 상기 반도체 바디(141)보다 고농도로 도핑된 제2의 도핑층(140)이 더 형성되어, 도 6의 각 실시예와 달리, 상기 제2의 도핑층을 비트 라인(VSL, 140)으로 할 수 있다.

[0053] 도 8과 도9를 참조하면, 각각 도 6와 도 7의 실시예에서 터널링 유전체와 블록킹 유전체를 서로 바꾸어 후자를 제 1 유전체층(15)으로, 전자를 제 2 유전체층(11)으로 각각 형성한 것에만 차이가 있다.

[0054] 도 6 내지 도 9에 도시된 CTF 기반의 memcapacitive 소자들은 전하 트랩층에 전자 혹은 정공을 주입하여 소자의 전기용량을 조절할 수 있다. 전하 트랩층으로 주입되는 전하량을 조절하여 단일 소자로 멀티레벨의 전기용량을 구현할 수 있고, 이를 기반으로 멀티레벨 시냅스 가중치를 표현할 수 있다.

[0055] 도 10은 본 발명의 일 실시예에 시냅스 소자로 사용되는 CTF 기반 memcapacitive 소자의 바디 도핑 극성에 따른 멀티레벨 전기용량 변화를 보인 전기 특성도이다. CTC의 바디의 도핑 극성에 따라 P-body와 N-body CTC으로 구분된다. 각 바디 극성에 따라 인가된 게이트 전압에 따른 전기용량 특성이 다르다. P-body CTC의 경우는 도 10(A)와 같이 게이트 전압이 커질수록 전기용량이 감소하는 특성을 갖고, N-body CTC의 경우 도 10(B)와 같이 게이트 전압이 작아질수록 전기용량이 감소하는 특성을 갖는다. 따라서 단일 소자에서 전기용량 멀티레벨을 구현할 경우, 도 10과 같이 적절한 동작 전압(V_{READ})을 설정하여야 멀티레벨 특성을 구현할 수 있다. 도 3에서, 2개의 CTF 기반 memcapacitive 소자(12, 14)로 시냅스 소자(10)를 구성할 경우 위와 같은 동작 특성을 고려하여 설계하여야 한다.

[0056] 상술한 CTF 기반 memcapacitive 소자로 시냅스 소자를 구현할 경우, 도 3에서 SSL, WL 및 BL을 선택함으로써 Fowler-Nordheim(FN) 터널링을 통하여 프로그램 및 이레이즈 동작을 수행할 수 있다. 아래 표 1은 일례로 양의 동작 전압을 기반으로 하여 FN 프로그램 및 이레이즈 방법을 나타낸 것이다.

[0057] [표 1] 전하 기반의 memcapacitive 시냅스 소자의 프로그램 및 이레이즈 방법

	FN 프로그램		FN 이레이즈	
	(A)	(B)	(C)	(D)
SSL	V_{cc}	$V_{PASS,B}$	$V_{PASS,C}$	V_{cc}
선택 BL	0 V	V_{PGM}	V_{ERS}	0 V
비선택 BL	V_{cc}	$V_{PGM,PASS}$	$V_{ERS,PASS}$	V_{cc}
선택 WL	V_{PGM}	0 V	0 V	V_{ERS}
비선택 WL	$V_{PGM,PASS}$	$V_{PGM,PASS}$	$V_{ERS,PASS}$	$V_{ERS,PASS}$

[0058]

[0059] 표 1의 (A)와 (C)는 도 6 및 도 7과 같은 게이트 유전체 구조를 갖는 전하 기반 memcapacitive 소자의 프로그램 또는 이레이즈 동작 전압이다. FN 프로그램은 종래의 NAND flash에서 사용하는 self-boosting program inhibit(SBPI) 동작과 유사한 방식으로 수행된다. 프로그램 시 표 1(A)를 참고하여 SSL에 V_{cc} 전압을 인가한다. 이후 BL 전압을 인가한다. 이 때, 선택 BL에는 0 V를 인가하고, 비선택 BL에는 V_{cc}를 인가한다. 이를 통하여 비선택 BL과 연결된 VSL들은 $V_{cc} - V_{TH,SST}$ 전압으로 프리차지 및 플로팅 된다. 이후 WL 전압을 인가하여 FN 프로그램을 수행한다. 이 때, 선택 WL에는 V_{PGM}을 인가하고, 비선택 WL에는 $V_{PGM,PAS}$를 인가한다. 이 때, 플로팅 되어 있던 VSL은 용량성 결합에 의하여 높은 전압으로 부스팅(boosting)되어 채널과 게이트 사이의 충분한 전계가 생성되지 않아 프로그램을 방지할 수 있다. 선택 WL과 선택 BL에 연결된 소자들은 채널과 게이트 사이에 충분한 전계가 생성되어 프로그램 된다.

[0060] 만일, 멀티레벨 전기용량 각각의 값을 정확히 구현하고자 할 경우 종래의 NAND flash에서 사용하는 incremental step pulse programming(ISPP) 방식을 사용할 수 있다.

[0061] 이레이즈 시 표 1(C)를 참고하여 SSL에 $V_{PASS,C}$ 전압을 인가한다. 이후 BL 전압을 인가한다. 이 때, 선택 BL에

는 V_{ERS} 전압을 인가하고, 비선택 BL에는 $V_{ERS,PASS}$를 인가한다. 이를 통하여 SSL에 인가된 $V_{PASS,C}$에 의한 SST의 게이트 유전체 breakdown을 방지한다. SSL에 인가되는 $V_{PASS,C}$ 전압은 SST의 문턱 전압($V_{TH,SST}$)을 고려하여 선택 BL에 인가되는 전압(V_{ERS}) 보다 높은 전압($V_{ERS}+V_{TH,SST}$)을 인가한다. 이후 WL 전압을 인가한다. 이 때 선택 WL에는 0 V를 인가하고, 비선택 WL에는 $V_{ERS,PASS}$ 전압을 인가한다. 비선택 BL과 비선택 WL, 혹은 선택 BL과 비선택 WL 전압이 인가되는 경우, 채널과 게이트 사이에 충분한 전계가 생성되지 않아 이레이즈를 방지할 수 있다. 반면 선택 WL과 선택 BL 전압이 인가되는 소자에서는 채널과 게이트 사이에 충분한 전계가 생성되어 이레이즈가 수행된다.

[0062] 도 8 및 도 9와 같은 게이트 유전체 구조를 갖는 memcapacitive 시냅스 소자를 사용할 경우에는 표 1의 (B)와 (D)를 참고하여 프로그램 또는 이레이즈 동작을 수행할 수 있다. 프로그램 시 표 1(B)를 참고하여 SSL에 $V_{PASS,B}$ 전압을 인가한다. 이후 BL 전압을 인가한다. 이 때, 선택 BL에는 V_{PGM} 전압을 인가하고, 비선택 BL에는 $V_{PGM,PASS}$를 인가한다. 이를 통하여 SSL에 인가된 $V_{PASS,B}$에 의한 SST의 게이트 유전체 breakdown을 방지한다. SSL에 인가되는 $V_{PASS,B}$ 전압은 SST의 문턱 전압($V_{TH,SST}$)을 고려하여 선택 BL에 인가되는 전압(V_{PGM}) 보다 높은 전압($V_{PGM}+V_{TH,SST}$)을 인가한다. 이후 WL 전압을 인가한다. 이 때 선택 WL에는 0 V를 인가하고, 비선택 WL에는 $V_{PGM,PASS}$ 전압을 인가한다. 이를 통하여 선택 WL 및 선택 BL과 연결된 소자에서만 채널과 게이트 사이에 충분한 전계가 생성되어 프로그램이 수행된다.

[0063] 이레이즈 시 표 1(D)를 참고하여 SSL에 V_{cc} 전압을 인가한다. 이후 BL 전압을 인가한다. 이 때, 선택 BL에는 0 V를 인가하고, 비선택 BL에는 V_{cc}를 인가한다. 이를 통하여 비선택 BL과 연결된 VSL들은 $V_{cc}-V_{TH,SST}$전압으로 프리차지 및 플로팅 된다. 이후 WL 전압을 인가하여 FN 프로그램을 수행한다. 이 때, 선택 WL에는 V_{ERS}을 인가하고, 비선택 WL에는 $V_{ERS,PASS}$를 인가한다. 이 때, 플로팅 되어 있던 VSL은 용량성 결합에 의하여 높은 전압으로 부스팅(boosting)되어 채널과 게이트 사이의 충분한 전계가 생성되지 않아 이레이즈을 방지할 수 있다. 이를 통하여 선택 WL 및 선택 BL과 연결된 소자에서만 채널과 게이트 사이에 충분한 전계가 생성되어 이레이즈가 수행된다.

[0064] 도 11은 본 발명의 다른 실시예에 의한 용량성 커플링 기반의 인공 신경망 시스템 구조를 보여주는 회로도이다.

[0065] 도 11을 참조하면, 상술한 도 3의 실시예에서 Output Neuron(200)의 출력측과 Output Controller(400) 사이에는 제2의 시냅스 어레이(102)와 제2의 Output Neuron(202)이 각각 하나 이상 더 연결되어 심층 신경망으로 확장한 C^2-DNN(capacitive coupling deep neural network) 구조로 실시될 수 있다.

[0066] 여기서, Output Neuron(200)은 Hidden Neuron으로 된다. 상술한 C^2-ANN 아키텍처의 Output Neuron(201)의 출력측 또는 제2의 Output Neuron(미도시)의 출력측에는 Pulse Converter(214)가 더 구비되어, 상기 Pulse Converter(214)를 통하여 변환된 입력 신호가 상기 제2의 시냅스 어레이(102)로 인가될 수 있다. 이 때, Pulse Converter(214)는 Output Neuron(201)에서 출력된 OL 신호를 다음 레이어, 즉 제2의 시냅스 어레이(102)의 레이어, 입력 신호로 변환하는 역할을 수행한다. Pulse Converter(214)로 입력된 OL 신호는 신호 변환을 거쳐 각각 WL^+ 및 WL^- 출력 신호로 변환되어 다음 층의 제2의 시냅스 어레이(102)로 인가된다.

[0067] 따라서 인지 연산 수행 시 WL Decoder를 통하여 입력 신호가 입력되면 각 층별로 순차적으로 벡터-행렬 곱 연산이 수행되고, 최종적으로 심층 신경망의 출력 뉴런에서 인지 연산의 결과가 출력된다. 이처럼 본 발명은 C^2-ANN은 간단한 구성요소 추가를 통하여 C^2-DNN을 구성할 수 있어, 심층 신경망으로의 높은 확장성을 가진다.

[0068] 다음은, 도 3 및 도 4를 참조하며, 본 발명의 다른 실시 모습인 용량성 커플링 기반의 인공 신경망 시스템의 동작 방법에 대하여 설명한다.

[0069] 도 4는 도 3의 용량성 커플링 기반의 인공 신경망 시스템의 추론 시 동작 과정으로 보여주는 시간에 따른 전원 인가도이다.

[0070] 도 4를 참조하면, 본 발명의 용량성 커플링 기반의 인공 신경망 시스템에서 상기 인공 신경망 시스템의 추론 연

산 동작은 시간상 초기화(initialization), 적분(integration) 및 평가(evaluation)로 구분되는 3단계로 나누어 수행된다.

[0071] 먼저, 상기 초기화 단계는 T_{phase0} 동안 상기 SST(220)의 SSL(string select line)에 SST 턴온(turn-on) 전압을 인가하고, 선택된 string(예컨대, 131)의 비트 라인(BL 1)에 소정의 전압이 인가되어 상기 VSL(130)은 프리차지되고, 상기 충전용 트랜지스터(M2)의 게이트에는 턴온 전압(Rstb 신호)이 인가되어 상기 발화용 커패시터(213)를 상기 동작전압(V_{DD})으로 충전한다.

[0072] 상기 적분 단계는 상기 초기화 단계 완료 후 T_{phase1} 동안 상기 WL Decoder(300)를 통하여 상기 시냅스 어레이(100)의 각 선택된 2개의 워드 라인에 상승 전압(WL^+)과 하강 전압(WL^-)이 동시에 인가되어 상기 시냅스 소자의 가중치에 의한 벡터-행렬 곱 연산 수행 결과로 얻어진 입력 전압이 상기 프리차지에 더하여 상기 선택된 string(131)의 VSL 전압이 결정되고, 상기 동작 제어용 트랜지스터(M3)의 게이트에 턴온 전압으로 EN 신호가 인가된 상태에서 상기 VSL 전압으로 자극 수신용 트랜지스터(M4)를 턴온시키며 상기 발화용 커패시터(213)를 방전시킨다.

[0073] 상기 평가 단계는 상기 적분 단계 완료 후 T_{phase2} 동안 상기 리셋용 트랜지스터(M5)의 게이트에 턴온 전압으로 RESET 신호가 입력되어 상기 발화용 커패시터(213)를 균일한 속도로 방전시킨다.

[0074] 상기 적분 단계 또는 상기 평가 단계에서 상기 출력단(215)의 전압(V_{Cint})이 떨어지면서 상기 발화용 커패시터(213)의 양단 전위차가 $V_{VDD} - V_{TH}$ 이상이 될 경우(V_{TH}는 펄스 생성기의 문턱 전압), 상기 펄스 생성기(214)에서 출력 신호(OL j)를 내보내는 것으로 동작할 수 있다. 이 때, 펄스 생성기(214)에서 출력된 출력 신호(OL j)인 펄스는 상기 적분 또는 평가 단계가 동안 계속 유지가 된다. 벡터-행렬 곱 연산 결과가 클수록 펄스 생성기(214)에서 펄스가 먼저 출력되어 출력 펄스의 유지 시간이 길어진다.

[0075] 본 발명의 C^2-ANN은 2개의 column line을 이용하는 종래의 방식과 비교하여 column line간의 cross talk 및 각 column line의 전기용량 차이에 의한 비이상적 특징에 대해 상대적으로 자유롭다. C^2-ANN 아키텍처의 시냅스 어레이는 종래의 컨덕턴스 기반의 시냅스 어레이와 다르게 시냅스 소자를 통한 DC 전류가 없다. 또한 뉴런회로는 추론 동작 시 적분 단계에서 V_{DD}로 충전된 $V_{C_{int}}$ 노드를 M4 트랜지스터를 통하여 방전하는 방식으로 시냅스 어레이에서 수행된 벡터-행렬 곱 연산 결과를 처리한다. 즉, 추론 연산 수행에 있어 뉴런 회로(210)에서 발화용 커패시터(213)의 충/방전을 제외한 다른 동작은 수행되지 않기 때문에 DC 전력 소모가 없다. 따라서 추론 연산 시 뉴런 회로(210)의 평균 전력 소모는 발화용 커패시터(213)의 전기용량(C_{int})과 V_{DD}에 의해 결정된다. 이를 통하여 종래의 전류 합산 기반의 뉴로모픽 시스템의 뉴런회로와 달리 본 발명의 C^2-ANN 아키텍처에 기반한 뉴런회로는 시냅스 어레이 크기와 무관한 전력 소모를 가진다. 즉, 본 발명의 C^2-ANN 아키텍처는 시냅스 어레이 크기가 증가하더라도, 시냅스 어레이(100), WL Decoder(300) 및 Output Neuron(200)에서 DC 전력이 증가하지 않고, 추론 동작 시 WL, VSL 및 뉴런회로의 C_{int}를 충/방전하는 과정에서만 전력소모가 발생한다. 이러한 점들을 종합하였을 때, 본 발명의 C^2-ANN은 종래 전류 합산 기반의 뉴로모픽 시스템에 비하여 전체 아키텍처의 전력 소모를 획기적으로 줄일 수 있다.

[0076] 또한, 본 발명에서 설계한 C^2-ANN 아키텍처는 추론 연산 수행 시 입력 신호 크기에 비례하는 지속 시간의 WL 펄스가 시냅스 어레이로 인가된다. 뉴런 회로를 통하여 추론 연산의 결과가 출력될 때 벡터-행렬 곱 연산 결과가 큰 string의 출력 뉴런에서 OL 펄스가 먼저 출력된다. 만일 같은 시점에 출력 펄스 전압이 0 V로 강하할 경우, 벡터-행렬 곱 연산 결과가 클수록 OL 펄스가 길게 인가된다고 볼 수 있다.

[0077] 이는 곧 C^2-ANN 아키텍처의 출력 신호는 추론 동작 시 WL Decoder를 통해 인가되는 입력 신호와 유사성을 갖는다. 이러한 유사성을 통하여, 도 11과 같이, 신경망의 규모를 증가시켜 C^2-DNN 구조로 실시할 경우, 은닉 층과 은닉 층, 혹은 은닉 층과 출력 층 사이의 신호 변환 과정을 줄일 수 있다는 장점을 갖는다. 따라서 C^2-ANN을 심층 신경망(C^2-DNN 구조)으로 확장할 경우, 뉴런 회로의 출력을 다음 층의 입력으로 변환하는 과정에서 낮은 전

력소모와 빠른 동작 속도를 기대할 수 있다.

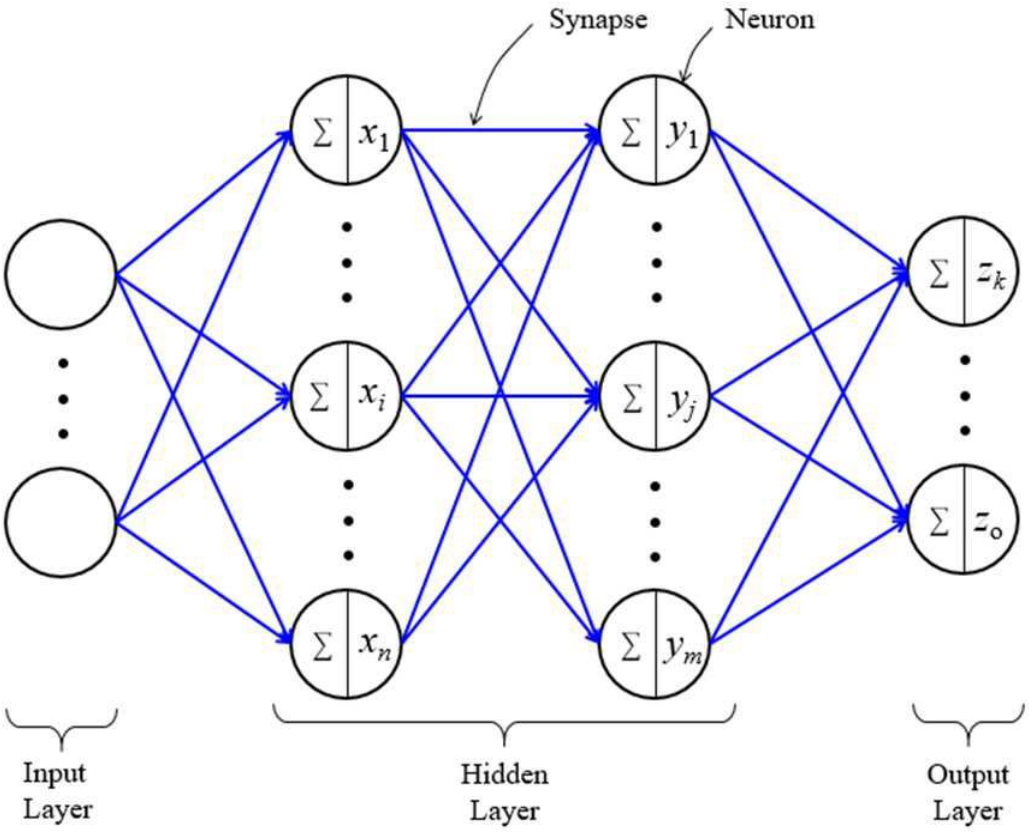
부호의 설명

[0078]

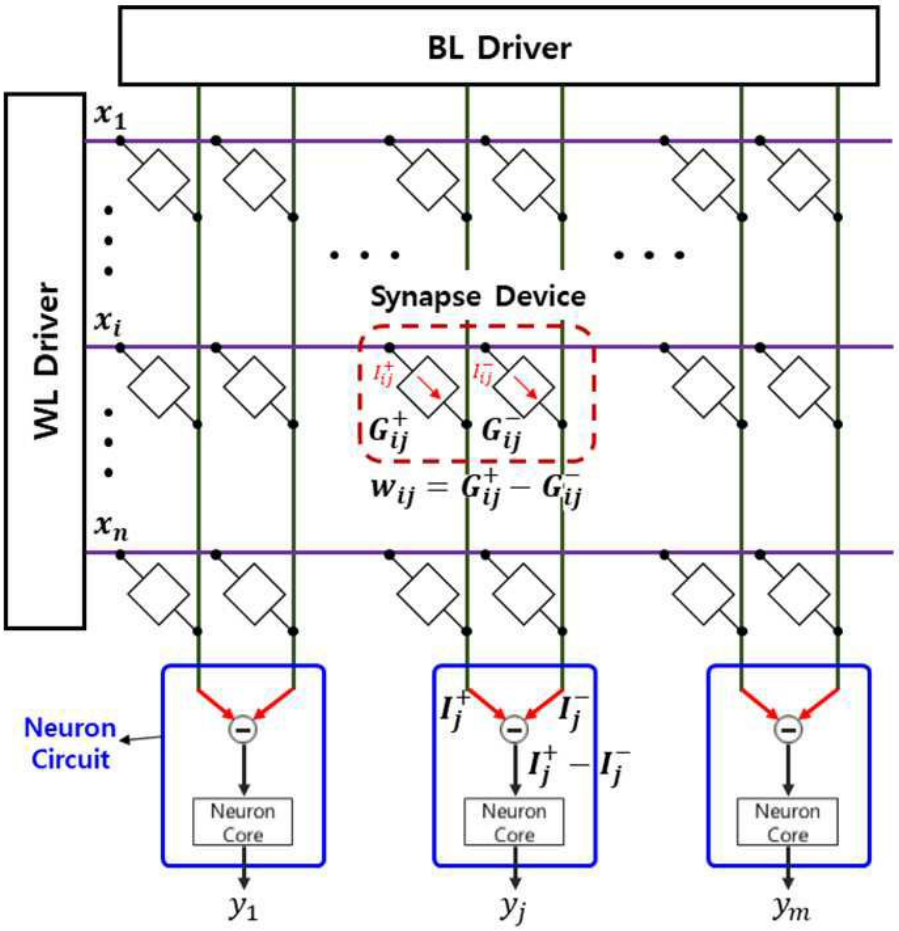
10: 시냅스 소자
12, 14: memcapacitive 소자
100: 시냅스 어레이
110, 111, 112, 120 121, 122: 워드 라인
130, 140: 비트 라인(VSL)
200, 201, 202: Output Neuron
210: 뉴런 회로
212: 적분기
213: 발화용 커패시터
214: 펄스 생성기
220: SST
230, 240: 출력 라인
300: WL Decoder
400: Output Controller

도면

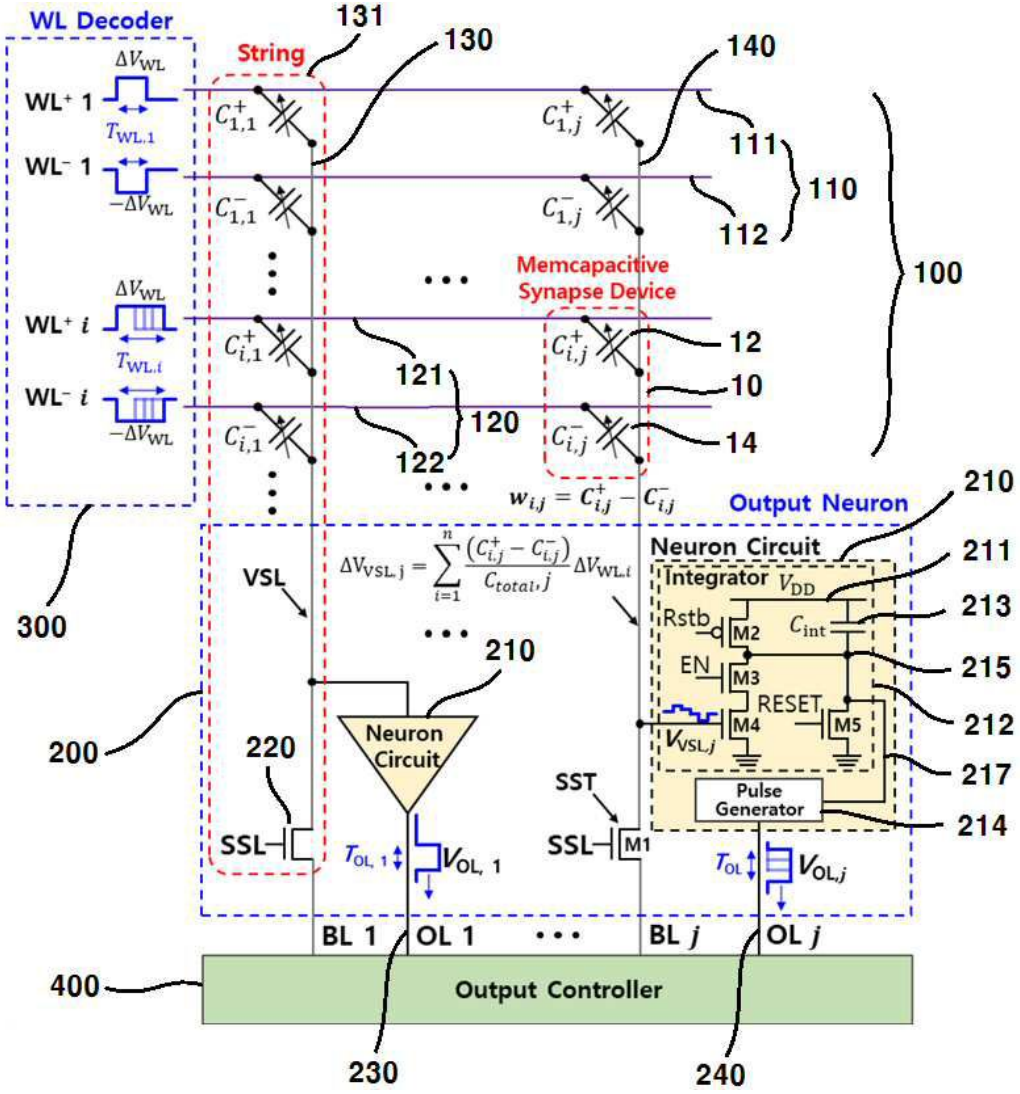
도면1



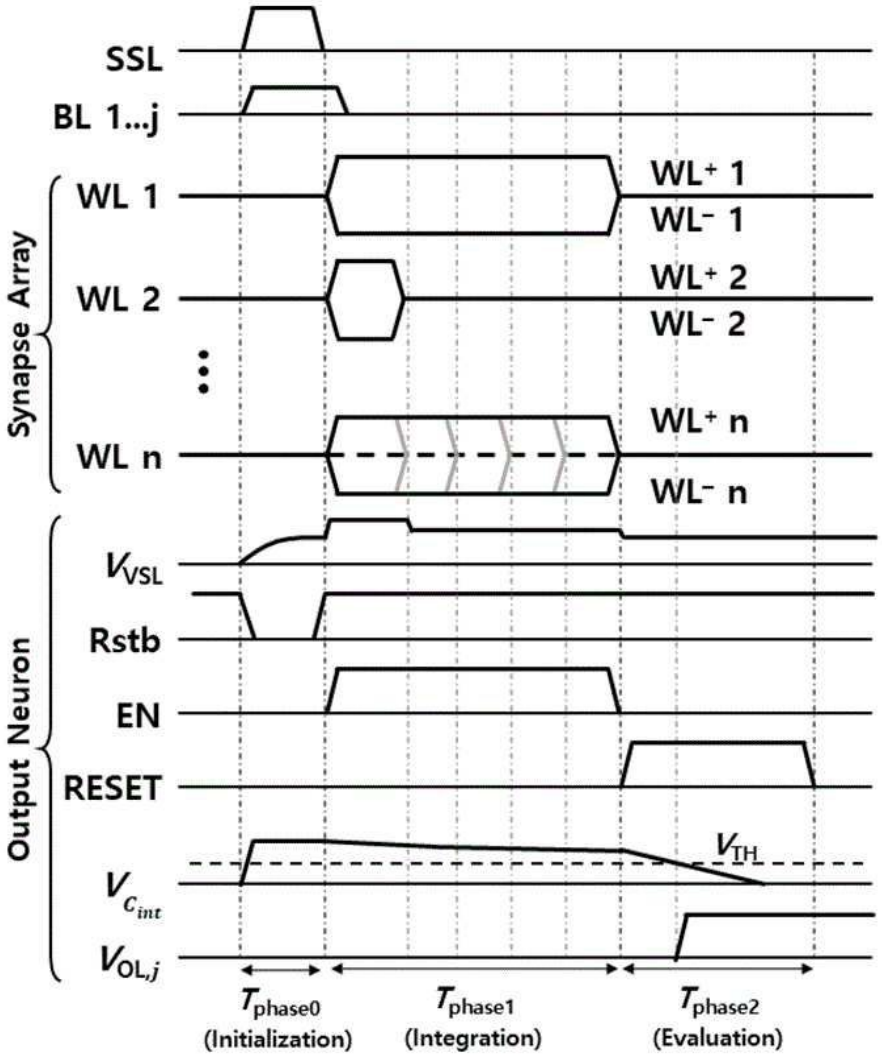
도면2


BL Driver
WL Driver
x_1
x_i
x_n
Synapse Device
I_{ij}^+
I_{ij}^-
G_{ij}^+
G_{ij}^-
$w_{ij} = G_{ij}^+ - G_{ij}^-$
Neuron Circuit
Neuron Core
I_j^+
I_j^-
$I_j^+ - I_j^-$
y_1
y_j
y_m

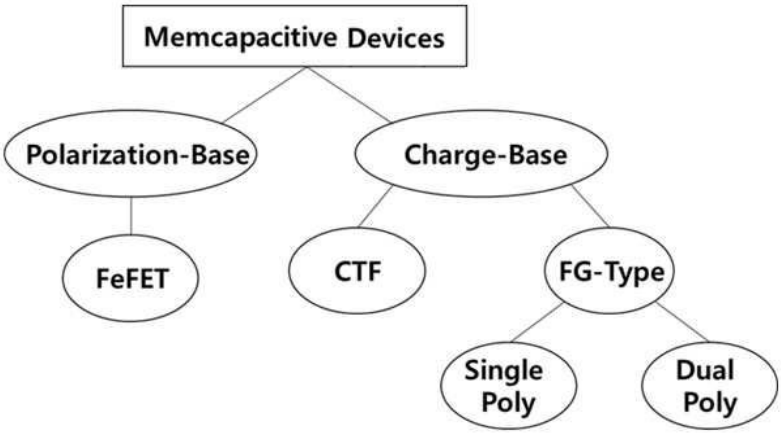
도면3


WL Decoder
String
131
130
140
111
110
112
100
Memcapacitive
Synapse Device
12
10
14
121
120
122
300
200
220
210
VSL
SSL
SST
M1
M2
M3
M4
M5
Rstb
EN
RESET
Integrator
Neuron Circuit
Pulse Generator
Output Neuron
211
213
215
212
217
214
BL 1
OL 1
BL j
OL j
400
Output Controller
230
240

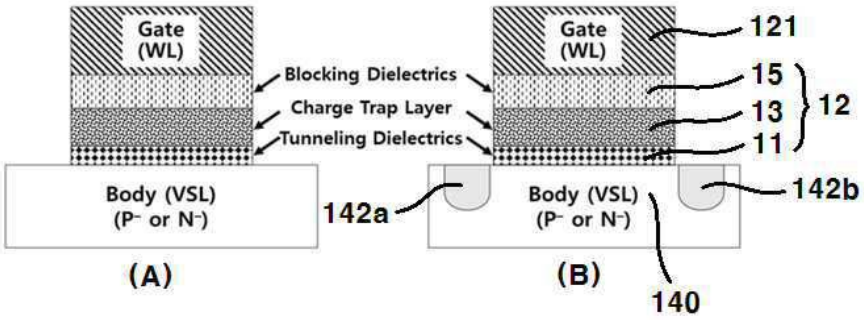
도면4


SSL
BL 1...j
Synapse Array
WL 1
WL 2
WL n
WL+ 1
WL- 1
WL+ 2
WL- 2
WL+ n
WL- n
Output Neuron
V_{VSL}
Rstb
EN
RESET
V_{TH}
$V_{C_{int}}$
$V_{OL,j}$
T_{phase0} (Initialization)
T_{phase1} (Integration)
T_{phase2} (Evaluation)

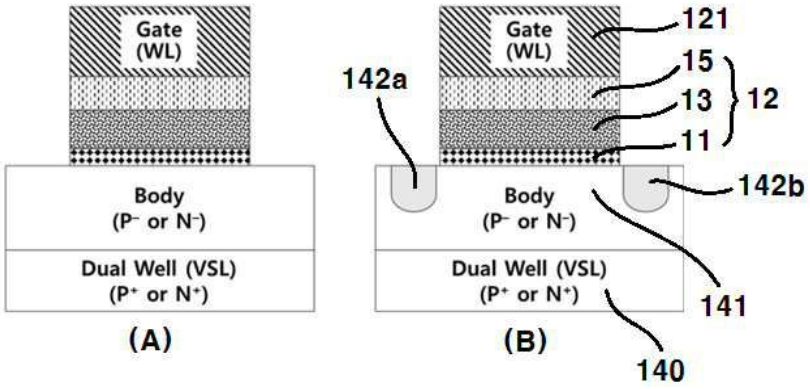
도면5


Memcapacitive Devices
Polarization-Base
Charge-Base
FeFET
CTF
FG-Type
Single Poly
Dual Poly

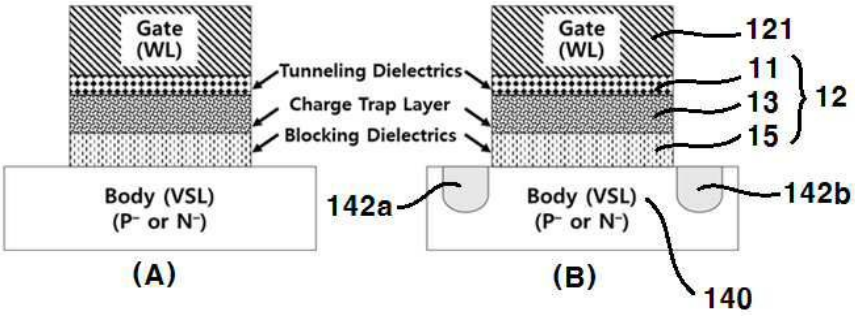
도면6



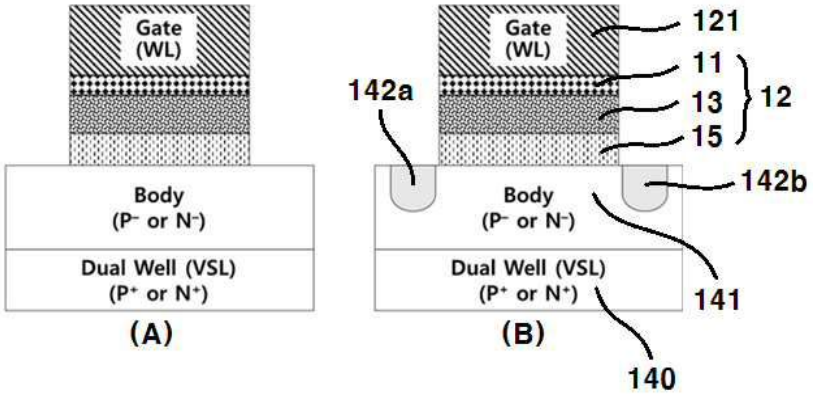
도면7



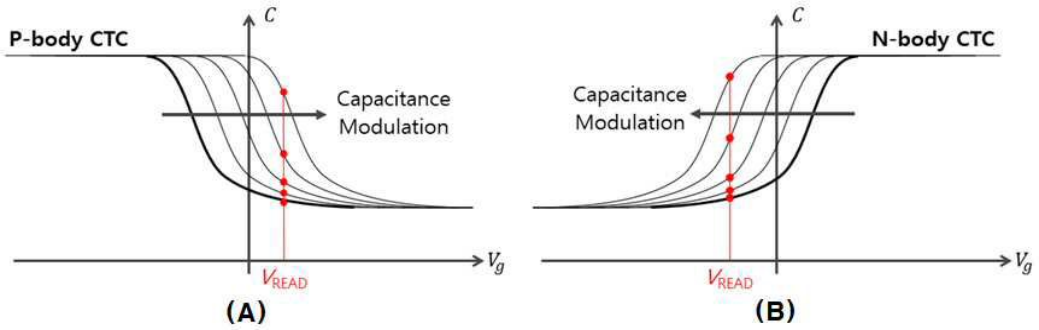
도면8



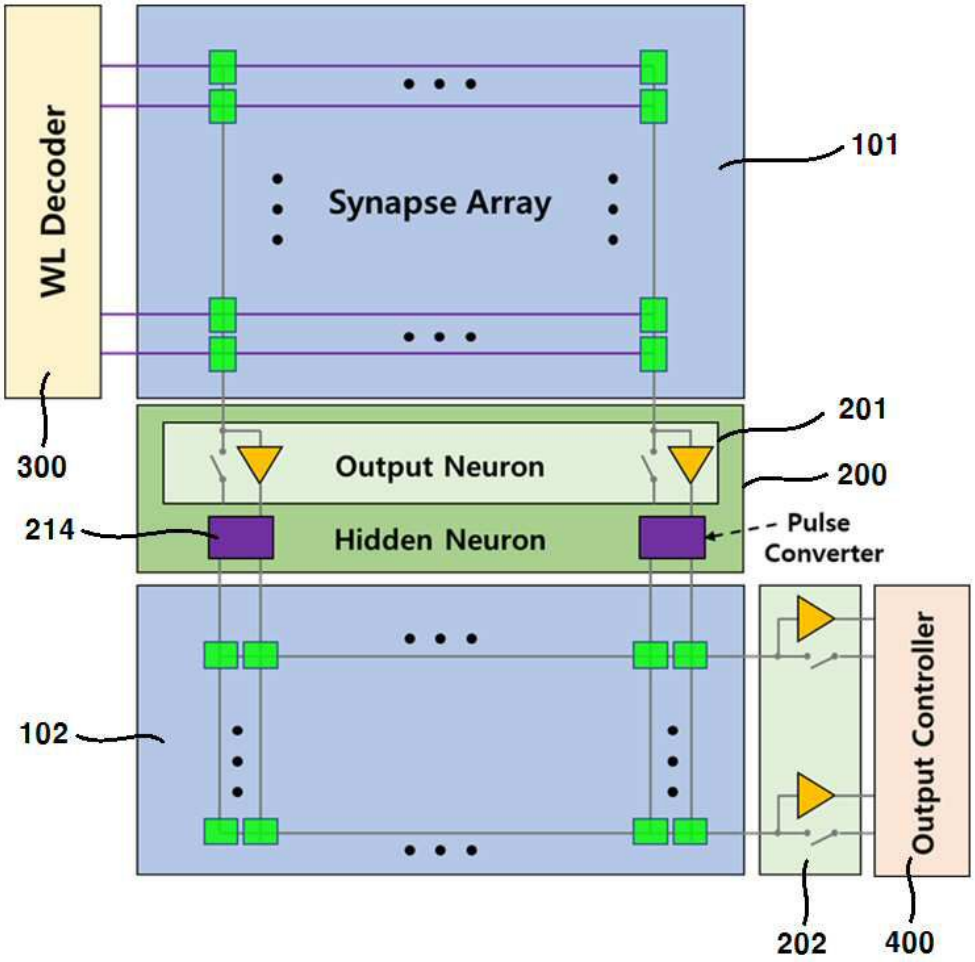
도면9



도면10


P-body CTC
C
Capacitance Modulation
V_g
V_{READ}
(A)
N-body CTC
C
Capacitance Modulation
V_g
V_{READ}
(B)

도면11


WL Decoder
Synapse Array
101
300
201
Output Neuron
200
214
Hidden Neuron
Pulse Converter
102
Output Controller
202
400