



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2023년06월07일

(11) 등록번호 10-2541000

(24) 등록일자 2023년06월01일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2023.01) *G06F 17/16* (2006.01)
G11C 16/04 (2006.01)

(52) CPC특허분류
G06N 3/063 (2013.01)
G06F 17/16 (2013.01)

(21) 출원번호 10-2022-0143504

(22) 출원일자 2022년11월01일

심사청구일자 2022년11월01일

(56) 선행기술조사문헌

KR1020210126335 A*

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

인하대학교 산학협력단

인천광역시 미추홀구 인하로 100(용현동, 인하대학교)

(72) 발명자

김형진

서울특별시 관악구 신림동7길 33, 902호(신림동)

이근호

인천광역시 미추홀구 인하로47번길 47, 402호(용현동)

송민석

경기도 성남시 수정구 대왕판교로 1187, 36동 305호(심곡동, 공군아파트)

(74) 대리인

양성보

전체 청구항 수 : 총 4 항

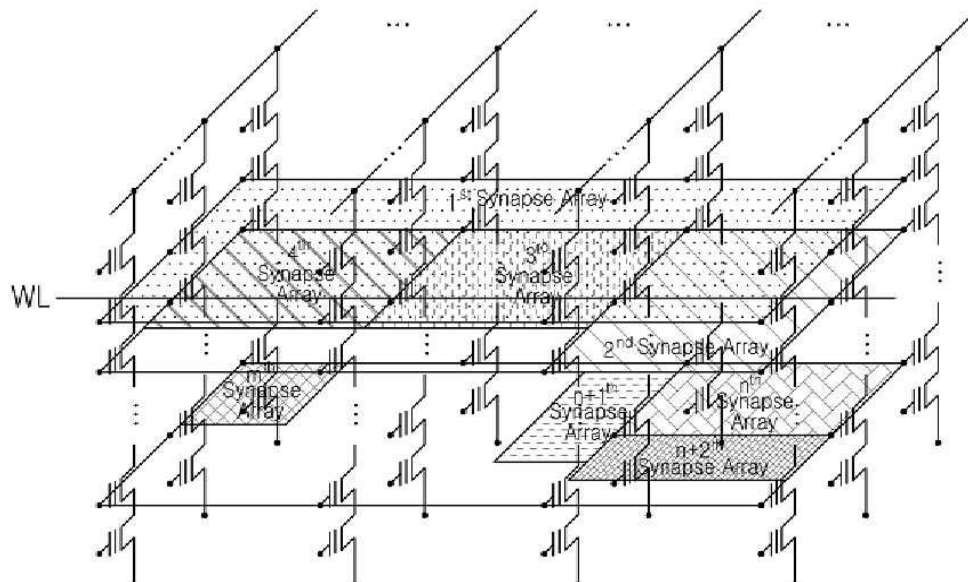
심사관 : 이준상

(54) 발명의 명칭 인공신경망 구현을 위한 3차원 적층형 시냅스 어레이 스트링

(57) 요약

인공신경망 구현을 위한 3차원 적층형 시냅스 어레이 스트링에 직렬로 연결된 3D 낸드 플래시 구조에서 시냅스 소자를 매핑 시키는 방법 및 그 구조가 제시된다. 본 발명에서 제안하는 인공신경망 구현을 위한 3차원 적층형 시냅스 어레이 스트링에 직렬로 연결된 3D 낸드 플래시 구조에서 시냅스 소자를 매핑 시키는 방법은 벡터, 행렬

(뒷면에 계속)

대표도 - 도3

급 연산을 위해 해당하는 제1 시냅스 레이어를 활성화하기 위해 리드 전압(Vread)을 상응하는 워드라인(Word Line; WL) 층에 인가하였고, 나머지 워드라인에는 패스 전압(Vpass)을 인가하는 단계, 해당하는 제1 시냅스 레이어의 크기에 맞는 제1 스트링 선택 라인(String Selection Lines; SSLs) 입력과 제1 비트 라인(Bitlines; BLs) 출력만 활성화시키는 단계 및 같은 워드라인 층을 활성화 시킨 상태에서 상기 출력된 제1 비트 라인 출력을 제2 시냅스 어레이의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 제2 시냅스 어레이에 해당하는 제2 비트 라인 출력을 활성화시키는 단계를 포함한다.

(52) CPC특허분류

G11C 11/54 (2013.01)

G11C 16/0483 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711173042

과제번호 2022M3I7A1085471

부처명 과학기술정보통신부

과제관리(전문)기관명 한국연구재단

연구사업명 원천기술개발사업

연구과제명 [Ezbaro] (2022M3I7A1078544) 전하저장형 실리콘 MOS 커패시터 기반 물리적 복제불

가 함수 개발

기 여 율 40/100

과제수행기관명 인하대학교

연구기간 2022.04.20 ~ 2022.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711154235

과제번호 2020M3F3A2A01081656

부처명 과학기술정보통신부

과제관리(전문)기관명 한국연구재단

연구사업명 원천기술개발사업

연구과제명 [Ezbaro] 고신뢰성 신경모사 시스템을 위한 실리콘 플래시 기반 전하저장형 시냅스

소자 및 어레이

기 여 율 40/100

과제수행기관명 인하대학교

연구기간 2022.01.01 ~ 2023.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711152938

과제번호 2021-0-02052-002

부처명 과학기술정보통신부

과제관리(전문)기관명 정보통신기획평가원

연구사업명 대학ITRC센터육성지원사업(교육운영수익)

연구과제명 [Ezbaro][정부] 스마트 모빌리티를 위한 인공지능 시스템반도체 핵심 기술 개발 및

인력 양성(2차년도)

기 여 율 20/100

과제수행기관명 인하대학교 산학협력단

연구기간 2022.01.01 ~ 2022.12.31

명세서

청구범위

청구항 1

제1 시냅스 레이어 및 제2 시냅스 레이어를 포함하는 시냅스 소자 매핑 방법에 있어서,

상기 제1 시냅스 레이어 및 제2 상기 시냅스 레이어 각각은 워드라인, 스트링 선택라인 및 비트라인을 각각 포함하고,

워드라인(Word Line; WL)을 통한 백터, 행렬곱 연산을 위해 해당하는 제1 시냅스 레이어를 활성화하기 위해 리드 전압(Vread)을 상응하는 워드라인(Word Line; WL) 층에 인가하고, 나머지 워드라인에는 패스 전압(Vpass)를 인가하는 단계;

제1 스트링 선택 라인(String Selection Lines; SSLs)을 통해 해당하는 제1 시냅스 레이어의 크기에 맞는 제1 스트링 선택 라인 입력과 제1 비트 라인(Bitlines; BLs) 출력만 활성화시키는 단계; 및

제1 비트 라인을 통해 같은 워드라인 층을 활성화 시킨 상태에서 상기 출력된 제1 비트 라인 출력을 제2 시냅스 레이어의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 제2 시냅스 레이어에 해당하는 제2 비트 라인 출력을 활성화시키는 단계

를 포함하고,

상기 제1 비트 라인을 통해 같은 워드라인 층을 활성화 시킨 상태에서 상기 출력된 제1 비트 라인 출력을 제2 시냅스 레이어의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 제2 시냅스 레이어에 해당하는 제2 비트 라인 출력을 활성화시키는 단계는,

상기 제2 시냅스 레이어의 행렬곱 연산을 진행하고, 복수의 시냅스 레이어 각각의 크기에 따라 유동적으로 하나의 워드라인 층에 복수의 시냅스 레이어를 매핑하여, 복수의 시냅스 레이어 각각에 해당하는 스트링 선택 라인 및 비트 라인을 순차적으로 활성화시킴으로써 인공신경망의 추론과정을 구현하여 셀 낭비를 감소시키는

시냅스 소자를 매핑 방법.

청구항 2

제1항에 있어서,

상기 제1 스트링 선택 라인(String Selection Lines; SSLs)을 통해 해당하는 제1 시냅스 레이어의 크기에 맞는 제1 스트링 선택 라인 입력과 제1 비트 라인 출력만 활성화시키는 단계는,

상기 제1 스트링 선택 라인 입력에 따른 상기 제1 비트 라인 출력의 전류를 읽음으로써 활성화된 셀만을 포함하는 상기 제1 시냅스 레이어의 행렬곱 연산을 진행하는

시냅스 소자를 매핑 방법.

청구항 3

삭제

청구항 4

제1 시냅스 레이어 및 제2 시냅스 레이어를 포함하는 시냅스 소자에 있어서,

상기 제1 시냅스 레이어 및 상기 제2 시냅스 레이어 각각은,

백터, 행렬곱 연산을 위해 해당하는 제1 시냅스 레이어를 활성화하기 위한 해당 워드라인(Word Line; WL) 층에는 리드 전압(Vread)이 인가되고, 나머지 워드라인에는 패스 전압(Vpass)이 인가되는 워드라인;

해당하는 제1 시냅스 레이어의 크기에 맞춰 활성화되는 제1 스트링 선택 라인(String Selection Lines; SSLs);

및

해당하는 제1 시냅스 레이어의 크기에 맞춰 활성화되는 제1 비트 라인(Bitlines; BLs)

을 포함하고,

같은 워드라인 층을 활성화 시킨 상태에서 상기 활성화된 제1 비트 라인의 출력을 제2 시냅스 레이어의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 제2 시냅스 레이어에 해당하는 제2 비트 라인 출력을 활성화시키고,

상기 제2 시냅스 레이어의 행렬곱 연산을 진행하고, 복수의 시냅스 레이어 각각의 크기에 따라 유동적으로 하나의 워드라인 층에 복수의 시냅스 레이어를 매핑하여, 복수의 시냅스 레이어 각각에 해당하는 스트링 선택 라인 및 비트 라인을 순차적으로 활성화시킴으로써 인공신경망의 추론과정을 구현하여 셀 낭비를 감소시키는

시냅스 소자.

청구항 5

제4항에 있어서,

상기 제1 스트링 선택 라인 입력에 따른 상기 제1 비트 라인 출력의 전류를 읽음으로써 활성화된 셀만을 포함하는 상기 제1 시냅스 레이어의 행렬곱 연산을 진행하는

시냅스 소자.

청구항 6

삭제

발명의 설명

기술 분야

[0001] 본 발명은 인공신경망 구현을 위한 3차원 적층형 시냅스 어레이 스트링에 직렬로 연결된 3D 낸드 플래시 구조에서 시냅스 소자를 매핑 시키는 방법 및 그 구조에 관한 것이다.

배경 기술

[0002] 최근, 인공 신경망은 다양한 지능형 작업에 사용되고 있으며, 심층 신경망은 높은 인지 기능 구현으로 많은 연구가 진행됐다. 하지만 기존의 소프트웨어적인 신경망은 많은 연산 양으로 메모리 오버헤드와 전력을 소모한다. 또한 폰노이만 구조는 프로세스 장치와 메모리 장치의 분리되어 많은 연산 처리에 어려움이 있다.

[0003] 따라서 이러한 폰노이만 구조의 한계를 개선하기 위해 인간의 뇌를 모방하는 뉴로모픽 시스템이 등장했다. 뉴로모픽 시스템은 저전력 시냅스 소자를 활용하여 벡터, 행렬곱 연산이 가능하다는 장점이 있다. 특히 3D 낸드 플래시 메모리는 우수한 집적도와 이미 상용화된 공정 프로세스를 바탕으로 시냅스 소자의 후보로 주목받고 있다. 하지만 스트링에 직렬로 연결된 3D 낸드 플래시 구조를 벡터, 행렬곱 연산에 활용하기 위해서는 매핑 방식이 매우 중요하다.

[0004] 3D 낸드 플래시의 스트링이라고 불리는 구조에는 낸드 플래시 셀이 직렬로 연결되어 있다. 인공신경망 연산에 필수적인 벡터, 행렬곱 연산이 병렬 연결 구조로 인해 가능한 맴리스트 기반 크로스바 어레이와는 다르게 직렬 연결 구조로 인해 제약이 존재하였다. 즉, 한 번에 하나의 셀 특성을 확인할 수 있어 순차적인 전압 펄스 인가를 통해 개별 셀의 전류를 읽은 다음 이들을 더해야 해서 많은 시간이 요구된다.

[0005] 하지만 3D NAND 플래시는 다른 시냅스 소자들 대비 우수한 집적도와 공정 성숙도를 가지고 상용화가 되어 있어 이 문제를 해결할 경우 시냅스 소자로 활용 가능성이 그 어떤 후보들보다 크다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국 등록특허 제10-2365324호(2022.02.16)

발명의 내용

해결하려는 과제

[0007] 본 발명이 이루고자 하는 기술적 과제는 스트링에 직렬로 연결된 3D 낸드 플래시 구조에서 벡터, 행렬곱 연산을 위한 시냅스 소자를 매핑 시키는 방법 및 및 그 구조를 제공하는데 있다. 제안하는 방법을 사용하여 직렬 연결의 3D 낸드 플래시 구조에서 벡터, 행렬곱 연산을 효과적으로 사용하고, 높은 집적도와 상용화된 3D 낸드 플래시를 사용하여 복잡한 구조의 심층 신경망을 효과적으로 구현하고자 한다.

과제의 해결 수단

[0008] 일 측면에 있어서, 본 발명에서 제안하는 인공신경망 구현을 위한 3차원 적층형 시냅스 어레이 스트링에 직렬로 연결된 3D 낸드 플래시 구조에서 시냅스 소자를 매핑 시키는 방법은 벡터, 행렬곱 연산을 위해 해당하는 제1 시냅스 레이어를 활성화하기 위해 리드 전압(Vread)을 상응하는 워드라인(Word Line; WL) 층에 인가하였고, 나머지 워드라인에는 패스 전압(Vpass)을 인가하는 단계, 해당하는 제1 시냅스 레이어의 크기에 맞는 제1 스트링 선택 라인(String Selection Lines; SSLs) 입력과 제1 비트 라인(Bitlines; BLs) 출력만 활성화시키는 단계 및 같은 워드라인 층을 활성화 시킨 상태에서 상기 출력된 제1 비트 라인 출력을 제2 시냅스 어레이의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 제2 시냅스 어레이에 해당하는 제2 비트 라인 출력을 활성화시키는 단계를 포함한다.

[0009] 상기 해당하는 제1 시냅스 레이어의 크기에 맞는 제1 스트링 선택 라인 입력과 제1 비트 라인 출력만 활성화시키는 단계는 상기 제1 스트링 선택 라인 입력에 따른 상기 제1 비트 라인 출력의 전류를 읽음으로써 활성화된 셀만을 포함하는 상기 제1 시냅스 레이어의 행렬곱 연산을 진행한다.

[0010] 상기 같은 워드라인 층을 활성화 시킨 상태에서 상기 출력된 제1 비트 라인 출력을 제2 시냅스 어레이의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 제2 시냅스 어레이에 해당하는 제2 비트 라인 출력을 활성화시키는 단계는 상기 제2 시냅스 레이어의 행렬곱 연산을 진행하고, 같은 워드라인 층에 복수의 시냅스 레이어가 매핑되어도 복수의 시냅스 레이어 각각에 해당하는 스트링 선택 라인 및 비트 라인을 순차적으로 활성화시킴으로써 인공신경망의 추론과정을 구현하여 셀 낭비를 감소시킬 수 있다.

[0011] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 인공신경망 구현을 위한 3차원 적층형 시냅스 어레이 스트링에 직렬로 연결된 3D 낸드 플래시 구조에서의 시냅스 소자는 벡터, 행렬곱 연산을 위해 해당하는 제1 시냅스 레이어를 활성화하기 위한 해당 워드라인(Word Line; WL) 층에는 리드 전압(Vread)이 인가되고, 나머지 워드라인에는 패스 전압(Vpass)을 인가되는 워드라인, 해당하는 제1 시냅스 레이어의 크기에 맞춰 활성화되는 제1 스트링 선택 라인(String Selection Lines; SSLs) 및 해당하는 제1 시냅스 레이어의 크기에 맞춰 활성화되는 제1 비트 라인(Bitlines; BLs)을 포함하고, 같은 워드라인 층을 활성화 시킨 상태에서 상기 출력된 제1 비트 라인의 출력을 제2 시냅스 어레이의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 제2 시냅스 어레이에 해당하는 제2 비트 라인 출력을 활성화시키다.

발명의 효과

[0012] 본 발명의 실시예들에 따른 스트링에 직렬로 연결된 3D 낸드 플래시 구조에서 벡터, 행렬곱 연산을 위한 시냅스 소자를 매핑 시키는 방법 및 및 그 구조를 통해 직렬 연결의 3D 낸드 플래시 구조에서 벡터, 행렬곱 연산을 효과적으로 사용할 수 있다. 또한, 높은 집적도와 상용화된 3D 낸드 플래시를 사용하여 복잡한 구조의 심층 신경망을 효과적으로 구현할 수 있다.

도면의 간단한 설명

[0013] 도 1은 종래기술에 따른 심층 신경망을 설명하기 위한 도면이다.

도 2는 종래기술에 따른 3D 낸드 플래시 구조에서 일반적으로 사용되는 시냅스 레이어 매핑 방식을 설명하기 위한 도면이다.

도 3은 본 발명의 일 실시예에 따른 시냅스 레이어 매핑 구조를 설명하기 위한 도면이다.

도 4는 본 발명의 일 실시예에 따른 제1 시냅스 레이어의 벡터, 행렬곱 연산을 설명하기 위한 도면이다.

도 5는 본 발명의 일 실시예에 따른 제2 시냅스 레이어의 벡터, 행렬곱 연산을 설명하기 위한 도면이다.

도 6은 본 발명의 일 실시예에 따른 시냅스 레이어 매핑 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0014] 본 발명을 이해하기 위해서는 낸드 플래시의 기본적인 동작 원리 및 뉴로모픽 시스템에서의 벡터, 행렬곱 연산이 진행되는 방식에 대해 이해할 필요가 있다. 또한, 본 발명에 사용되는 매핑 방식을 이해하기 위해 공정 프로세스에 대한 이해가 필요하다. 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.
- [0016] 도 1은 종래기술에 따른 심층 신경망을 설명하기 위한 도면이다.
- [0017] 도 1(a)는 을 참고하면, 심층 신경망의 구조를 나타내는 도면이고, 도 1(b)는 3D 낸드 플래시 구조에서 벡터, 행렬곱 연산을 설명하기 위한 도면이다.
- [0018] 3D 낸드 플래시는 스트링에 셀이 직렬로 연결되어 있지만 SSLs(String Selection Lines)에 전압을 인가하고 BLs(Bitlines) 전류를 읽어 벡터, 행렬곱 연산이 가능하다. 이 과정에서 인접한 두 BLs의 연결된 메모리 셀은 쌍을 이루게 되고 전류 차를 이용한다.
- [0019] 최종 전류는 비교회로를 통해 전류에서 전압으로 변환된다. 선택된 WL(Word Line)은 리드 전압(Vread)을 인가하고, 선택되지 않은 나머지 WLs은 패스 전압(Vpass)이 인가된다. 이 때 각각의 시냅스 셀에 흐르는 전류는 프로그램 상태에 따라 달라지므로 각각의 WL에 위치한 셀들이 심층 신경망의 가중치 맵의 역할을 하게 된다.
- [0021] 도 2는 종래기술에 따른 3D 낸드 플래시 구조에서 일반적으로 사용되는 시냅스 레이어 매핑 방식을 설명하기 위한 도면이다.
- [0022] 도 2를 참조하면, 하나의 WL에 하나의 시냅스 레이어를 매핑하는 방식이다. 3D 낸드 플래시는 홀 에칭 기술을 통해 한 번에 형성하기 때문에 직육면체의 구조를 가지게 되는데 인공신경망의 경우 층별로 가중치 맵의 크기가 다를 수 있고, 따라서 하나의 시냅스 레이어를 하나의 WL 층에 매핑하게 되면 많은 메모리 셀 낭비가 발생하게 된다.
- [0023] 대부분의 종래기술에서 가중치 맵, 즉 시냅스 어레이를 하나의 WL 층에 매핑하여 연산을 진행한다. 하지만 시냅스 레이어의 크기는 다양하게 존재하며 홀 에칭 기술로 한 번에 만들어지기 때문에 직육각형 형태를 가질 수 밖에 없는 3D 낸드 플래시에서 사용되지 않는 메모리 셀 낭비가 필연적으로 발생한다. 따라서 본 발명에서는 하나의 WL에 여러 개의 시냅스 레이어들을 배치하여 메모리 셀 효율을 높이는 방법을 제안한다.
- [0025] 도 3은 본 발명의 일 실시예에 따른 시냅스 레이어 매핑 방식을 설명하기 위한 도면이다.
- [0026] 도 3을 참조하면, 기존의 메모리 셀 낭비를 줄이기 위해 시냅스 레이어의 크기에 따라 유동적으로 한 WL에 여러 시냅스 레이어를 매핑하여 메모리 셀 낭비를 줄일 수 있다.
- [0028] 도 4는 본 발명의 일 실시예에 따른 제1 시냅스 레이어의 벡터, 행렬곱 연산을 설명하기 위한 도면이다.
- [0029] 도 4를 참조하면, 3D 낸드 플래시 구조에서 제안하는 매핑 방식을 적용하여 벡터, 행렬곱 연산을 수행하는 구체적인 실시예로 나타낸다. 도 4는 여러 개의 시냅스 레이어를 WL2에 매핑하여 메모리 셀 효율을 높이는 예시를 보여준다.
- [0030] 도 4는 제1 시냅스 레이어(도 4의 2nd Synapse Array)의 벡터, 행렬곱 연산으로 해당하는 시냅스 레이어를 활성화하기 위해 Vread를 대응되는 WL2층에 인가하였고, 나머지 WL에는 Vpass를 인가하였다. 그리고, 해당하는 제1 시냅스 레이어의 크기에 맞는 SSLs 입력(411, 412, 413)과 BLs 출력(421, 422)만 활성화시킨다. 이렇게 출력된 BLs 출력(421, 422)을 다시 같은 WL2층을 활성화 시킨 상태에서 제2 시냅스 어레이(도 4의 3rd Synapse Array)의 크기에 맞는 SSLs 입력(도 5 참조)에 인가하고 해당 BLs 출력(도 5 참조)을 활성화시킨다.
- [0032] 도 5는 본 발명의 일 실시예에 따른 제2 시냅스 레이어의 벡터, 행렬곱 연산을 설명하기 위한 도면이다.
- [0033] 도 5를 참조하면, 제2 시냅스 레이어(도 5의 3rd Synapse Array)의 해당하는 SSLs 입력(511, 512)과 BLs 출력

(521, 522)만 활성화하여 백터, 행렬곱 연산을 진행한다. 선택된 WL2에는 리드전압이 입력되고, 선택되지 않은 나머지 WLs는 패스전압이 인가된다. 이를 통해 같은 WL2 층에 복수의 시냅스 레이어가 매핑되더라도 인공신경망의 추론과정을 구현할 수 있어 기존의 3D 낸드 플래시의 구조에서 기인하는 셀 낭비를 최소화할 수 있다.

[0035] 도 6은 본 발명의 일 실시예에 따른 시냅스 레이어 매핑 방법을 설명하기 위한 흐름도이다.

[0036] 본 발명의 실시예에 따른 본 발명은 인공신경망 구현을 위한 3차원 적층형 시냅스 어레이 스트링에 직렬로 연결된 3D 낸드 플래시 구조에서 시냅스 소자를 매핑 시키는 방법은 백터, 행렬곱 연산을 위해 해당하는 제1 시냅스 레이어를 활성화하기 위해 리드 전압(Vread)을 상응하는 워드라인(Word Line; WL) 층에 인가하였고, 나머지 워드라인에는 패스 전압(Vpass)을 인가하는 단계(610), 해당하는 제1 시냅스 레이어의 크기에 맞는 제1 스트링 선택 라인(String Selection Lines; SSLs) 입력과 제1 비트 라인(Bitlines; BLs) 출력만 활성화시키는 단계(620) 및 같은 워드라인 층을 활성화 시킨 상태에서 상기 출력된 제1 비트 라인 출력을 제2 시냅스 어레이의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 제2 시냅스 어레이에 해당하는 제2 비트 라인 출력을 활성화시키는 단계(630)를 포함한다.

[0037] 단계(610)에서, 백터, 행렬곱 연산을 위해 해당하는 제1 시냅스 레이어를 활성화하기 위해 리드 전압(Vread)을 상응하는 워드라인(Word Line; WL) 층에 인가하였고, 나머지 워드라인에는 패스 전압(Vpass)을 인가한다.

[0038] 단계(620)에서, 해당하는 제1 시냅스 레이어의 크기에 맞는 제1 스트링 선택 라인(String Selection Lines; SSLs) 입력과 제1 비트 라인(Bitlines; BLs) 출력만 활성화시킨다.

[0039] 상기 제1 스트링 선택 라인 입력에 따른 상기 제1 비트 라인 출력의 전류를 읽음으로써 활성화된 셀만을 포함하는 상기 제1 시냅스 레이어의 행렬곱 연산을 진행한다.

[0040] 더욱 상세하게는, 제1 시냅스 레이어의 백터, 행렬곱 연산으로 해당하는 시냅스 레이어를 활성화하기 위해 제1 시냅스 레이어의 크기에 맞는 제1 스트링 선택 라인 입력(Vread)을 대응되는 워드라인 층에 인가하고, 나머지 워드라인에는 패스 전압(Vpass)을 인가한다. 그리고, 해당하는 제1 시냅스 레이어의 크기에 맞는 제1 스트링 선택 라인 입력과 제1 비트 라인 출력만을 활성화시킨다. 이렇게 출력된 제1 비트 라인 출력을 다시 같은 워드라인 2층을 활성화 시킨 상태에서 제2 시냅스 어레이의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 해당 제2 비트 라인 출력을 활성화시킨다.

[0041] 단계(630)에서, 상기 같은 워드라인 층을 활성화 시킨 상태에서 상기 출력된 제1 비트 라인 출력을 제2 시냅스 어레이의 크기에 맞는 제2 스트링 선택 라인 입력에 인가하고 제2 시냅스 어레이에 해당하는 제2 비트 라인 출력을 활성화시킨다.

[0042] 상기 제2 시냅스 레이어의 행렬곱 연산을 진행하고, 같은 워드라인 층에 복수의 시냅스 레이어가 매핑되어도 복수의 시냅스 레이어 각각에 해당하는 스트링 선택 라인 및 비트 라인을 순차적으로 활성화시킴으로써 인공신경망의 추론과정을 구현하여 셀 낭비를 감소시킬 수 있다.

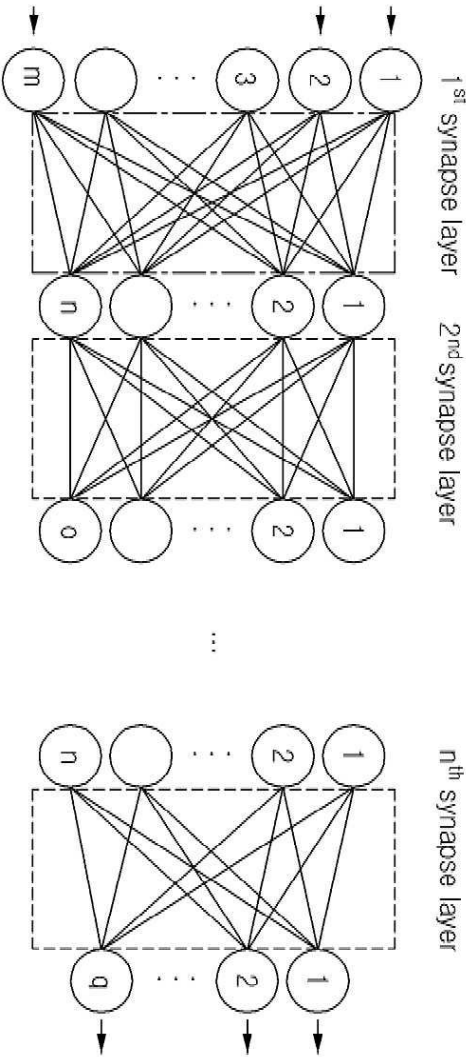
[0043] 더욱 상세하게는, 제2 시냅스 레이어의 해당하는 제2 스트링 선택 라인 입력과 제2 비트 라인만 활성화하여 백터, 행렬곱 연산을 진행한다. 선택된 워드라인에는 리드전압이 입력되고, 선택되지 않은 나머지 워드라인은 패스전압이 인가된다. 이를 통해 같은 워드라인 층에 복수의 시냅스 레이어가 매핑되더라도 인공신경망의 추론과정을 구현할 수 있어 기존의 3D 낸드 플래시의 구조에서 기인하는 셀 낭비를 최소화할 수 있다.

[0044] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

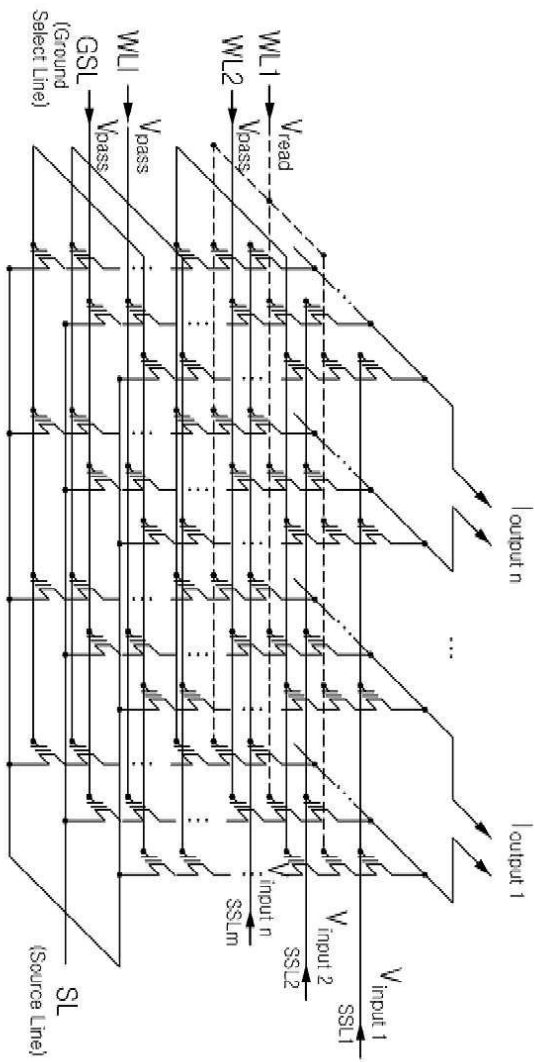
[0045] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

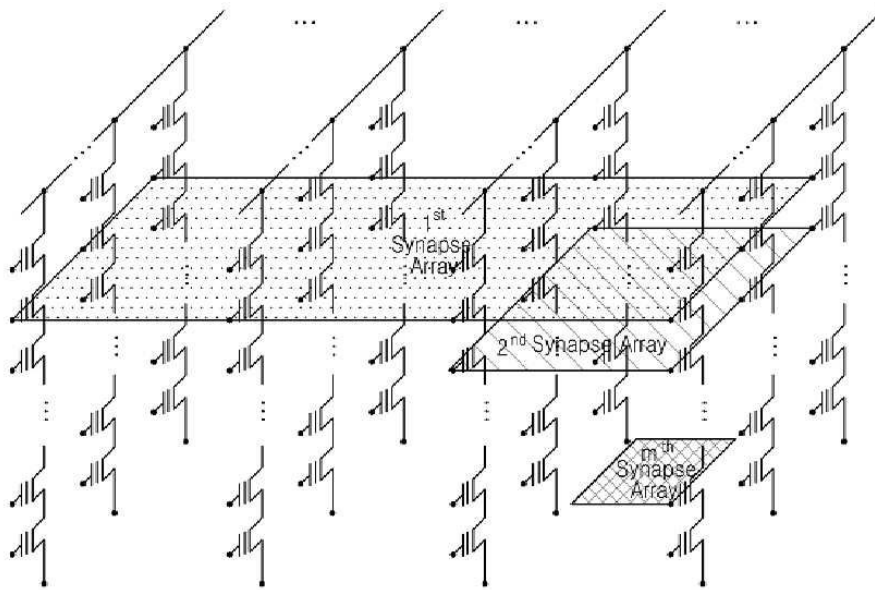
도면1a



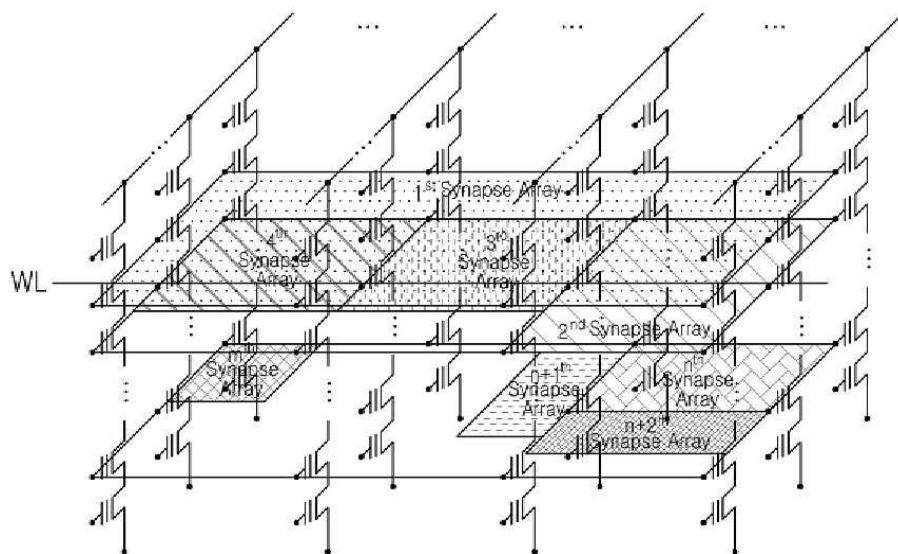
도면1b



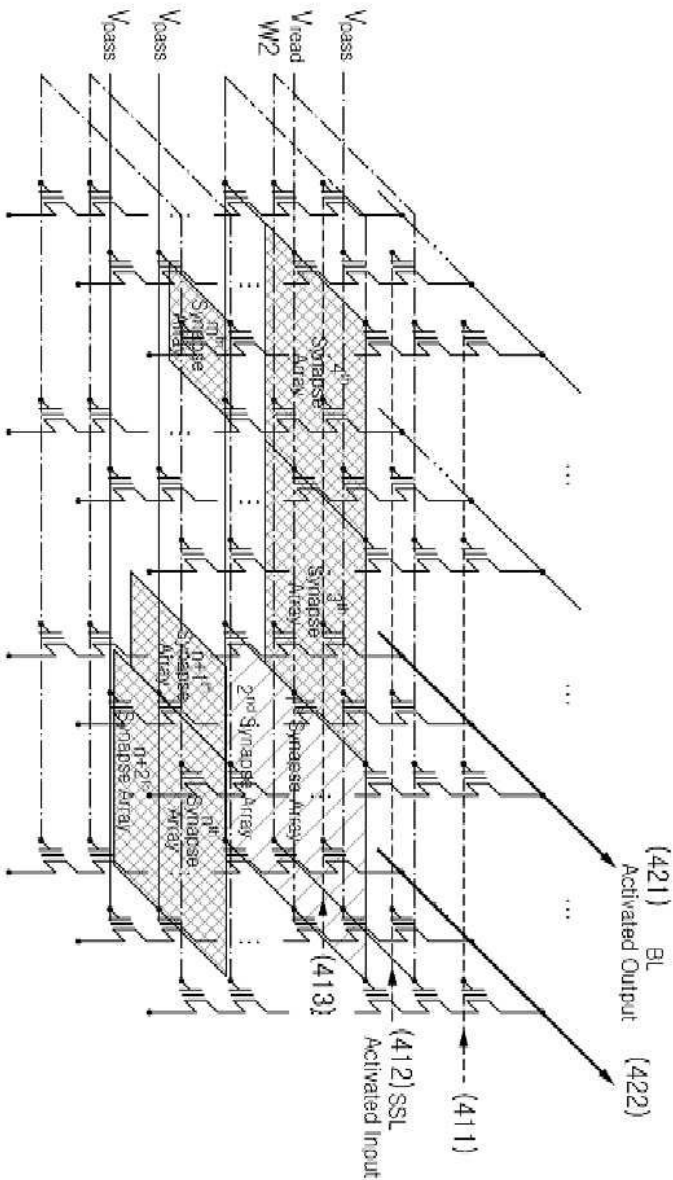
도면2



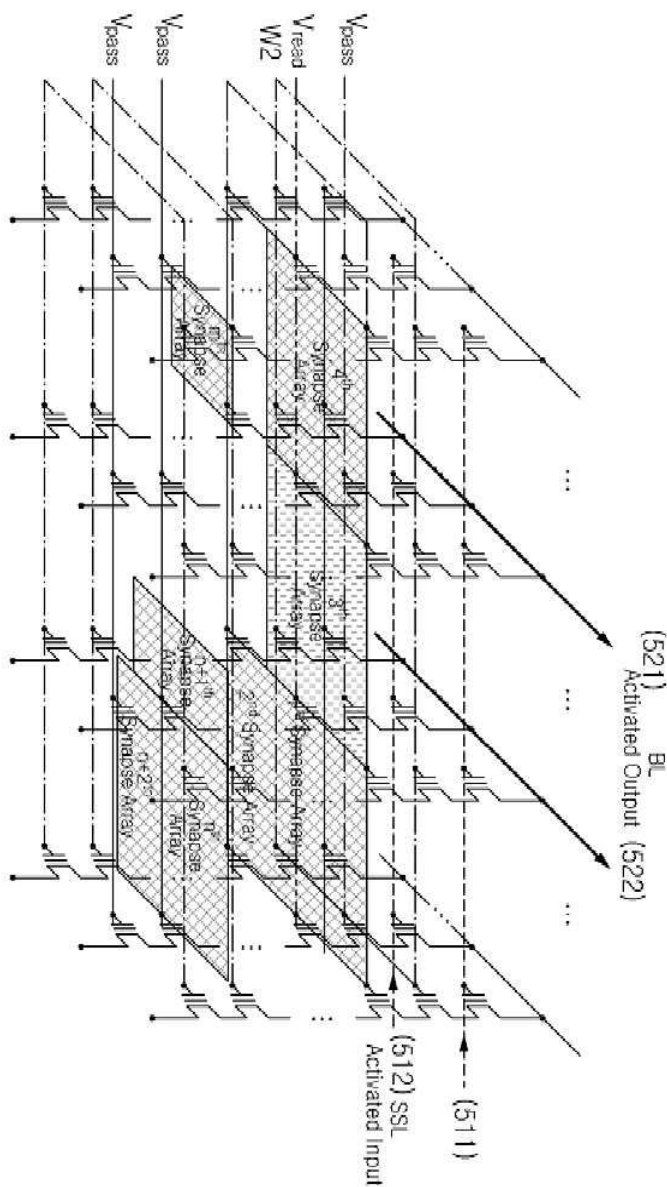
도면3



도면4



도면5



도면6

