(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0090078
(43) 공개일자 2022년06월29일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01) *G11C 11/412* (2006.01)
G11C 11/413 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G11C 11/412 (2013.01)
(21) 출원번호 10-2020-0180915
(22) 출원일자 2020년12월22일
심사청구일자 2020년12월22일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
정성우
서울특별시 강남구 개포로 264, 124동 2505호(개포동, 개포래미안포레스트)
최정환
서울특별시 동대문구 무학로49길 35, 203호(용두동)
(74) 대리인
송인호, 윤형근, 최관락

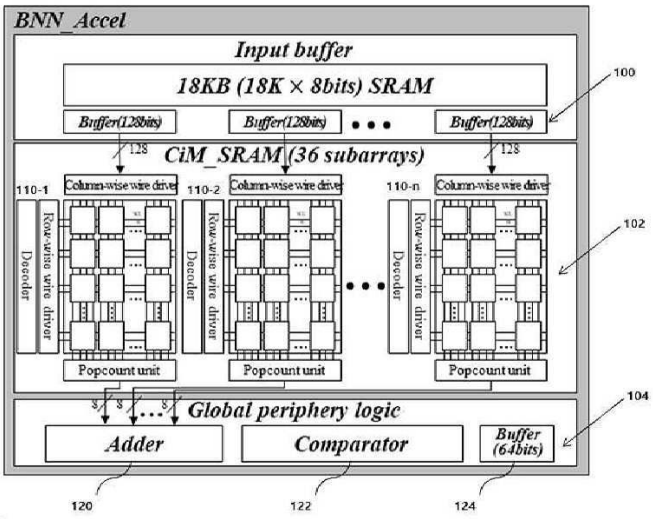
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 **모놀리식 3D 집적 구조를 이용한 인-메모리 SRAM 기반 이진 신경망 가속기**

(57) 요 약

본 발명은 모놀리식 3D 집적 구조를 이용한 인-메모리 SRAM 기반 이진 신경망 가속기를 개시한다. 본 발명에 따르면, 이진 신경망의 합성곱 층 연산 수행을 위한 입력 데이터를 인-메모리 SRAM에 공급하는 입력 버퍼; 상기 합성곱 층 연산 수행을 위한 커널 데이터를 저장하며, 복수의 서브어레이를 통해 상기 커널 데이터와 상기 입력 데이터를 이용하여 XNOR 연산과 팝카운트(Popcount) 연산을 포함하는 XAC 연산을 병렬적으로 수행하는 인-메모리 SRAM; 및 상기 병렬적으로 수행된 XAC 연산 결과를 가산 및 비교하여 상기 합성곱 층의 최종 연산 결과를 출력하는 글로벌 주변 로직을 포함하는 이진 신경망 가속기가 제공된다.

대 표 도 - 도1



(52) CPC특허분류
G11C 11/413 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711110769
과제번호 2020R1A2C2003500
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 온도를 고려한 3D 적층 메모리 기반 인메모리 가속기
기 여 율 1/2
과제수행기관명 고려대학교 산학협력단
연구기간 2020.03.01 ~ 2021.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711121315
과제번호 2020M3F3A2A01082329
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 차세대지능형반도체기술개발(R&D)
연구과제명 저온 공정 기반 Si/SiGe M3D 집적 소자 및 회로 플랫폼 기술개발
기 여 율 1/2
과제수행기관명 고려대학교 산학협력단
연구기간 2020.07.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

이진 신경망의 합성곱 층 연산 수행을 위한 입력 데이터를 인-메모리 SRAM에 공급하는 입력 버퍼;

상기 합성곱 층 연산 수행을 위한 커널 데이터를 저장하며, 복수의 서브어레이를 통해 상기 커널 데이터와 상기 입력 데이터를 이용하여 XNOR 연산과 팝카운트(Popcount) 연산을 포함하는 XAC 연산을 병렬적으로 수행하는 인-메모리 SRAM; 및

상기 병렬적으로 수행된 XAC 연산 결과를 가산 및 비교하여 상기 합성곱 층의 최종 연산 결과를 출력하는 글로벌 주변 로직을 포함하는 이진 신경망 가속기.

청구항 2

제1항에 있어서,

상기 인-메모리 SRAM의 개별 서브어레이는 N개의 셀을 포함하고,

상기 N개의 셀 각각은 XNOR 연산을 위한 3개의 트랜지스터 및 상기 커널 데이터 저장을 위한 6개의 트랜지스터를 포함하고,

상기 3개의 트랜지스터는 제1 상위 레이어에 배치되고, 상기 6개의 트랜지스터는 제1 하위 레이어에 배치되며,

상기 제1 상위 레이어와 상기 제1 하위 레이어는 모놀리식 3D 비아를 통해 연결되며,

상기 N개의 셀 중 N/2개의 셀은 상위 레이어에 배치되고, 나머지 N/2개의 셀은 하위 레이어에 배치되는 이진 신경망 가속기.

청구항 3

제2항에 있어서,

상기 3개의 트랜지스터는 트랜지스터 NL, NR 및 NB를 포함하고, 상기 제1 상위 레이어는 행 방향으로 배치되는 제1 배선 및 열 방향으로 배치되는 제2 및 제3 배선을 포함하는 이진 신경망 가속기.

청구항 4

제3항에 있어서,

상기 개별 서브어레이는,

입력 주소를 디코딩하는 디코더;

상기 입력 주소를 디코딩하여 선택된 상기 제1 배선이 상기 트랜지스터 NB를 활성화하도록 하는 행 단위 배선 드라이버;

상기 입력 버퍼를 통해 공급된 입력 데이터를 기반으로 열 방향으로 배치되는 제2 및 제3 배션이 상기 트랜지스터 NL 및 NB를 활성화하도록 하는 열 단위 배선 드라이버; 및

상기 N개의 셀 각각에 저장된 커널 데이터를 통해 수행된 XNOR 연산 결과를 기반으로 팝카운트 연산을 수행하는 팝카운트 유닛을 포함하는 이진 신경망 가속기.

청구항 5

제4항에 있어서,

상기 글로벌 주변 로직은,

상기 복수의 서브어레이 각각의 팝카운트 유닛에서 출력하는 연산 결과를 가산하는 가산기; 및

상기 가산기에서 출력하는 가산 결과를 미리 설정된 값과 비교하여 상기 합성곱 층의 최종 연산 결과를 출력하는 비교기를 포함하는 이진 신경망 가속기.

발명의 설명

기 술 분 야

[0001] 본 발명은 모놀리식 3D 집적 기술을 활용한 인-메모리 SRAM 기반 이진 신경망 가속기에 관한 것이다.

배 경 기 술

[0002] 최근 이미지 인식을 수행하기 위한 신경망(Neural Network) 알고리즘으로 단일 비트(Single-bit) 매개변수를 활용하는 이진 신경망(Binary Neural Network, BNN)이 등장하였다.

[0003] 이진 신경망은 단일 비트 매개변수를 기반으로 하므로, 기존 합성곱 신경망(Convolution Neural Network, CNN)의 덧셈 곱셈 연산(Multiply-and-Accumulation, MAC)을 단순한 XNOR 연산과 Population Count(Popcount) 연산으로 대체할 수 있었다.

[0004] 상기한 XNOR 연산과 Popcount 연산을 XAC 연산이라 한다.

[0005] XAC 연산은 연산 복잡도가 단순하기 때문에, 데이터 이동이 지연시간과 에너지 소모의 대부분을 차지하게 된다.

[0006] 데이터 이동에 소요되는 지연시간을 줄이고 에너지를 절감하기 위하여, 최근 기존 6T SRAM에 추가적인 트랜지스터(Transistor)를 활용한 XAC 연산을 수행하는 인-메모리 (Compute-in-Memory) SRAM이 등장하였다.

[0007] 하지만, 기존의 2D 인-메모리 SRAM는 이차원 평면에 추가적인 트랜지스터를 활용한 구조로 인하여, 기존 6T SRAM과 비교하여 SRAM 셀(Cell)의 면적이 커져 배선이 길어진다. 또한, 인-메모리 SRAM은 병렬 연산 수행을 위해 다수의 서브어레이(Subarray)를 활용하기 때문에 서브어레이 간 배선이 길어진다. 이러한 배선 길이 증가는 배선에서 소모하는 지연시간과 에너지 소모가 커지게 된다는 문제점을 야기한다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 한국등록특허 제10-1913930호

발명의 내용

해결하려는 과제

[0009] 상기한 종래기술의 문제점을 해결하기 위해, 본 발명은 지연시간 및 에너지 소모를 줄일 수 있는 모놀리식 3D 집적 구조를 이용한 인-메모리 SRAM 기반 이진 신경망 가속기를 제안하고자 한다.

과제의 해결 수단

[0010] 상기한 바와 같은 목적을 달성하기 위하여, 본 발명의 일 실시예에 따르면, 이진 신경망의 합성곱 층 연산 수행을 위한 입력 데이터를 인-메모리 SRAM에 공급하는 입력 버퍼, 상기 합성곱 층 연산 수행을 위한 커널 데이터를 저장하며, 복수의 서브어레이를 통해 상기 커널 데이터와 상기 입력 데이터를 이용하여 XNOR 연산과 팝카운트(Popcount) 연산을 포함하는 XAC 연산을 병렬적으로 수행하는 인-메모리 SRAM 및 상기 병렬적으로 수행된 XAC 연산 결과를 가산 및 비교하여 상기 합성곱 층의 최종 연산 결과를 출력하는 글로벌 주변 로직을 포함하는 이진 신경망 가속기가 제공된다.

[0011] 상기 인-메모리 SRAM의 개별 서브어레이는 N개의 셀을 포함하고, 상기 N개의 셀 각각은 XNOR 연산을 위한 3개의 트랜지스터 및 상기 커널 데이터 저장을 위한 6개의 트랜지스터를 포함하고, 상기 3개의 트랜지스터는 제1 상위 레이어에 배치되고, 상기 6개의 트랜지스터는 제1 하위 레이어에 배치되며, 상기 제1 상위 레이어와 상기 제1 하위 레이어는 모놀리식 3D 비아를 통해 연결되며, 상기 N개의 셀 중 N/2개의 셀은 상위 레이어에 배치되고, 나

머지 N/2개의 셀은 하위 레이어에 배치될 수 있다.

[0012] 상기 3개의 트랜지스터는 트랜지스터 NL, NR 및 NB를 포함하고, 상기 제1 상위 레이어는 행 방향으로 배치되는 제1 배선 및 열 방향으로 배치되는 제2 및 제3 배선을 포함할 수 있다.

[0013] 상기 개별 서브어레이는, 입력 주소를 디코딩하는 디코더, 상기 입력 주소를 디코딩하여 선택된 상기 제1 배선이 상기 트랜지스터 NB를 활성화하도록 하는 행 단위 배선 드라이버, 상기 입력 버퍼를 통해 공급된 입력 데이터를 기반으로 열 방향으로 배치되는 제2 및 제3 배션이 상기 트랜지스터 NL 및 NB를 활성화하도록 하는 열 단위 배선 드라이버 및 상기 N개의 셀 각각에 저장된 커널 데이터를 통해 수행된 XNOR 연산 결과를 기반으로 팝카운트 연산을 수행하는 팝카운트 유닛을 포함할 수 있다.

[0014] 상기 글로벌 주변 로직은, 상기 복수의 서브어레이 각각의 팝카운트 유닛에서 출력하는 연산 결과를 가산하는 가산기 및 상기 가산기에서 출력하는 가산 결과를 미리 설정된 값과 비교하여 상기 합성곱 층의 최종 연산 결과를 출력할 수 있다.

발명의 효과

[0015] 본 발명에 따르면, 모놀리식 3D 집적 구조를 이용한 인-메모리 SRAM을 기반으로 하는 이진 신경망 가속기를 제공하여 이진 신경망의 연산의 대부분을 차지하는 이진 합성곱 층 연산에 소모되는 지연시간과 에너지 소모를 줄일 수 있는 장점이 있다.

도면의 간단한 설명

[0016] 도 1은 본 발명의 일 실시예에 따른 이진 신경망 가속기의 구성을 도시한 도면이다.

도 2는 본 발명의 일 실시예에 따른 인-메모리 SRAM을 도시한 도면이다.

도 3은 본 실시예에 따른 인-메모리 SRAM의 개별 서브어레이 구조를 도시한 도면이다.

도 4는 본 실시예에 따른 인-메모리 SRAM이 XAC 연산을 수행하는 과정을 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0017] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다.

[0018] 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0020] 도 1은 본 발명의 일 실시예에 따른 이진 신경망 가속기의 구성을 도시한 도면이다.

[0021] 도 1에 도시된 바와 같이, 본 실시예에 따른 이진 신경망 가속기는 입력 버퍼(Input Buffer, 100), 모놀리식 3D 집적 구조를 이용한 XAC 연산을 수행하는 인-메모리 SRAM(CiM_SRAM, 102) 및 글로벌 주변 로직(Global Periphery Logic, 104)을 포함할 수 있다.

[0022] 여기서, XAC 연산은 XNOR 연산과 Population Count(Popcount) 연산을 의미한다.

[0023] 입력 버퍼(100)는 이진 신경망의 합성곱 층 연산 수행을 위한 입력 데이터 (Input Feature Map)를 저장하고 인-메모리 SRAM(102)에 공급한다.

[0024] 인-메모리 SRAM(102)은 합성곱 층 연산 수행을 위한 커널 데이터(Kernel Weights)를 저장한다.

[0025] 인-메모리 SRAM(102)은 복수의 서브어레이(110-1 내지 110-n)으로 구성될 수 있고, 복수의 서브어레이(110)에서 미리 저장된 커널 데이터와 입력 버퍼(100)로부터 공급된 입력 데이터를 이용하여 XAC 연산을 병렬적으로 수행한다.

[0026] 글로벌 주변 로직(104)은 이진 합성곱 층의 최종 결과를 추출한다.

[0027] 본 실시예에 따른 글로벌 주변 로직(104)은 가산기(Adder, 120), 비교기(Comparator, 122) 및 버퍼(124)를 포함할 수 있다.

[0028] 인-메모리 SRAM(102)의 각각의 서브어레이(110)의 XAC 연산 결과는 부분적인 XAC 연산 결과이므로, 병렬적으로 수행된 XAC 연산 결과를 가산기(120)를 통해 가산하여 최종 XAC 연산 결과를 출력한다.

[0029] 그 후, 비교기(122)를 통해서 미리 설정된 값(Threshold)과 비교하여 이진 합성곱 층의 최종 연산 결과를 출력한다.

[0030] 도 2는 본 발명의 일 실시예에 따른 인-메모리 SRAM을 도시한 도면이다.

[0031] 도 2에 도시된 바와 같이, 인-메모리 SRAM의 서브어레이는 디코더(Decoder, 200), 행 단위 배선 드라이버(Row-wise Wire Driver, 202), 열 단위 배선 드라이버(Column-wise Wire Driver, 204), 복수의 셀 어레이(Cell Array, 206) 및 팝카운트 유닛(208)을 포함할 수 있다.

[0032] 도 2를 참조하면, 본 실시예에 따른 인-메모리 SRAM(102)은 트랜지스터 수준과 서브어레이 수준에서 모놀리식 3D 집적 구조가 적용된다.

[0033] 도 2는 4층 구조의 인-메모리 SRAM을 도시한 도면이다.

[0034] 먼저, 트랜지스터 수준의 모놀리식 3D 집적 구조를 이용하여, 연산에 필요한 3개의 트랜지스터(3T)와 데이터를 저장하는 6개의 트랜지스터(6T)를 3D 구조로 적층한다.

[0035] 모놀리식 3D 집적 소자 형성 기술은 하나의 웨이퍼에 소자 형성 후 epi 성장, recrystallization 또는 절연체 층을 증착한 후 얇은 웨이퍼 위에 bonding을 통해 새로운 substrate layer를 형성 후 그 위에 소자를 시퀀셜하게 형성하는 공정 기술이다.

[0036] 모놀리식 3D 집적 소자 형성 기술은 각 레이어의 소자를 매우 얇은 채널 (50~500nm) 형성으로 Via specification의 변경이 가능 [(Via diameter (10~50nm), Via pitch (100~300nm)] 하여 TSV(Through Si Via)의 한계점으로 지적되는 웨이퍼 의 두께에 따른 문제점들을 해결 가능한 새로운 대안 제조 기술이다.

[0037] 이와 같은 모놀리식 3D 집적 구조에서는 각 레이어의 소자를 연결하는 비아를 인터-레이어 비아(inter-layer via) 또는 인터-타이어 비아(inter-tier via)로 정의한다. 본 실시예에서는 소자를 연결하는 비아를 모놀리식 3D 비아(MIV)로 정의한다.

[0038] 본 실시예에 따른 모놀리식 3D 비아를 통해 연산에 필요한 3개의 트랜지스터(3T)와 데이터를 저장하는 6개의 트랜지스터(6T)를 연결하기 때문에 인-메모리 SRAM(102)은 기존의 6T SRAM과 동일한 셀 면적을 가지게 된다. 또한, 서브어레이 수준에서도 모놀리식 3D 비아를 통해 서브어레이를 반으로 분할하여 적층하기 때문에 면적이 커지지 않는다.

[0039] 본 발명에서는 인-메모리 SRAM(102)의 구조가 열(column) 단위 배선의 길이가 길고 많음을 고려하여 비트라인 분할 방식의 서브어레이 수준 모놀리식 3D 집적 공정을 활용한다.

[0040] 이를 통해 본 실시예에 따른 인-메모리 SRAM(102)은 열 단위 배선의 길이 뿐 아니라, 서브어레이의 면적이 줄어들어 서브어레이 간 배선의 길이가 짧아진다.

[0041] 결과적으로, 본 실시예에 따른 인-메모리 SRAM(102)은 모놀리식 3D 집적 공정을 활용하여 2D 기반 인-메모리 SRAM 대비 배선 길이 감소로 인해 지연시간 및 에너지 소모를 절감할 수 있다.

[0042] 즉, 3개의 트랜지스터와 6개의 트랜지스터를 2층이 아닌 4층 구조로 형성하여 서브어레이 간 배선의 길이를 짧게 할 수 있는 것이다.

[0043] 보다 상세하게, 인-메모리 SRAM(102)의 개별 서브어레이(110)는 N개의 셀을 포함하고, N개의 셀 각각은 XNOR 연산을 위한 3개의 트랜지스터 및 커널 데이터 저장을 위한 6개의 트랜지스터를 포함한다.

[0044] 여기서, 트랜지스터 수준에서, 3개의 트랜지스터는 상위 레이어에 배치되고, 상기 6개의 트랜지스터는 하위 레이어에 배치되며, 상기 상위 레이어와 상기 하위 레이어는 모놀리식 3D 비아를 통해 연결된다.

[0045] 또한, 도 3에 도시된 바와 같이, 서브어레이 수준에서 N개의 셀 중 N/2개의 셀은 상위 레이어에 배치되고, 나머지 N/2는 하위 레이어에 배치되며, 이들 상하위 레이어는 모놀리식 3D 비아를 통해 연결된다.

[0046] 본 실시예에 따른 인-메모리 SRAM의 동작은 도 3을 참조하여 상세하게 설명한다.

[0047] 도 3은 본 실시예에 따른 인-메모리 SRAM의 개별 서브어레이 구조를 도시한 도면이다.

[0048] 도 3에서, 셀 어레이는 512-by-128 (512개의 행과 128개의 열)로 구성된 것을 예시적으로 도시한다.

[0049] 본 실시예에 따른 인-메모리 SRAM(102)은 기존 6T SRAM과 비교하여 3개의 추가적인 트랜지스터(NL, NR, NB)와 3개의 배선(Signal L, R, B)이 추가된 9T SRAM의 구조를 가진다.

[0050] 여기서, 3개의 배선 중, Signal B(제1 배선)은 행 방향으로 배치되는 배선이고, Signal L/R(제2 및 제3 배선)은 열 방향으로 배치되는 배선이다.

[0051] 도 4는 본 실시예에 따른 인-메모리 SRAM이 XAC 연산을 수행하는 과정을 도시한 도면이다.

[0052] 입력 주소에 따라 디코더(200)를 통해 선택된 Signal B(제1 배선)는 해당 행 (Row)의 모든 트랜지스터 NB를 활성화한다(단계 400).

[0053] 단계 400에서, 트랜지스터 NB의 활성화는 행 단위 배선 드라이버(202)에 의해 수행될 수 있다.

[0054] 서브어레이 외부에서 들어온 입력 데이터를 기반으로, Signal L/R(제2 및 제3 배선)은 트랜지스터 NL/NR를 활성화/비활성화한다(단계 402).

[0055] 단계 402에는 열 단위 배선 드라이버(204)에 의해 수행될 수 있다.

[0056] 각각의 셀에 저장되어 있는 값(Q/QB)은 Cout으로 이동하며 XNOR 연산이 수행된다(단계 404).

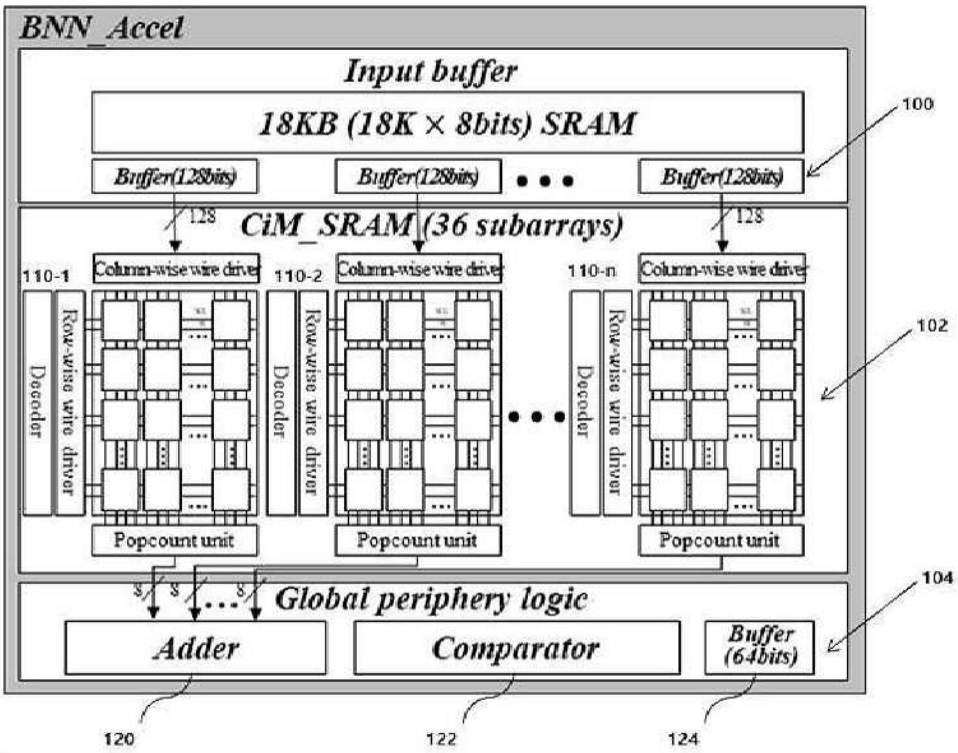
[0057] 선택된 행의 모든 셀에서의 XNOR 값은 비트라인(Bitline)을 통해 서브어레이 하단부의 팝카운트 유닛(208)으로 전달된다(단계 406).

[0058] 팝카운트 유닛(208)은 XNOR 값을 기반으로 팝카운트 연산을 수행한다(단계 408).

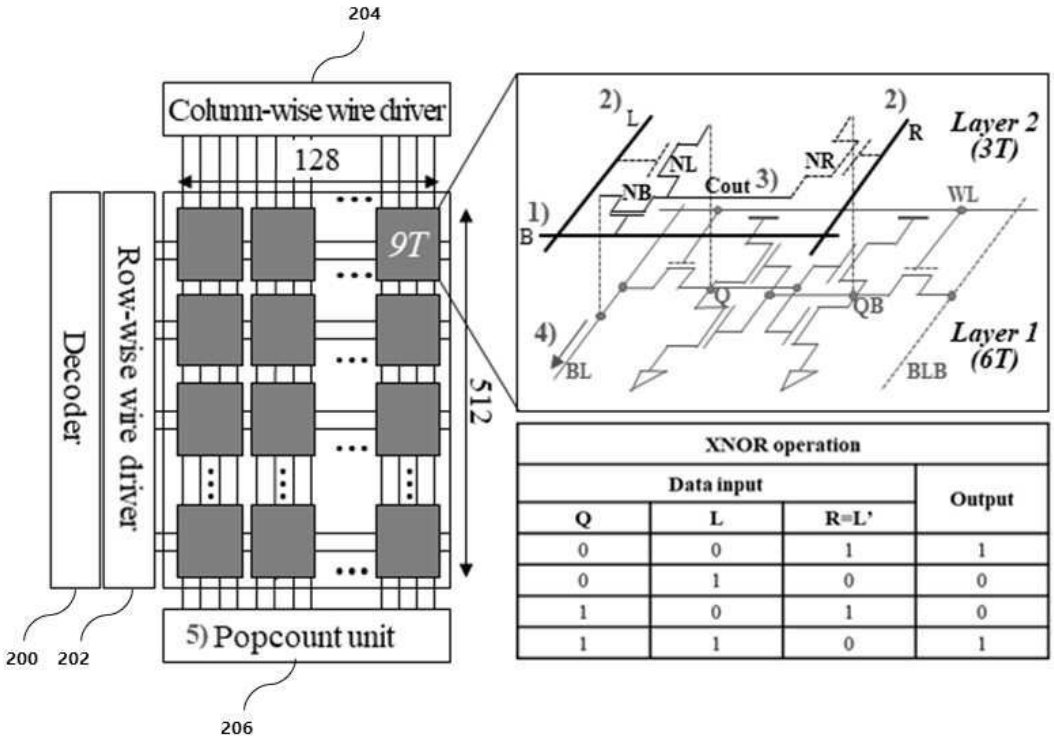
[0059] 상기한 본 발명의 실시예는 예시의 목적을 위해 개시된 것이고, 본 발명에 대한 통상의 지식을 가지는 당업자라면 본 발명의 사상과 범위 안에서 다양한 수정, 변경, 부가가 가능할 것이며, 이러한 수정, 변경 및 부가는 하기의 특허청구범위에 속하는 것으로 보아야 할 것이다.

도면

도면1

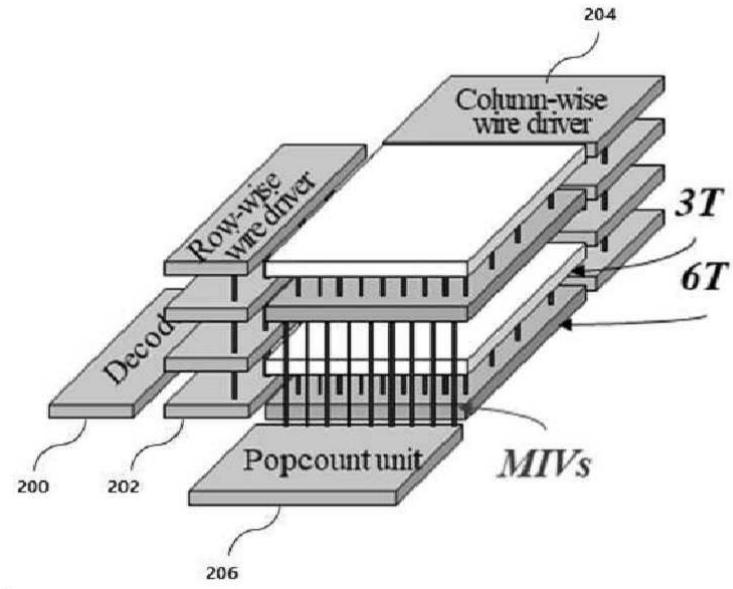


도면2

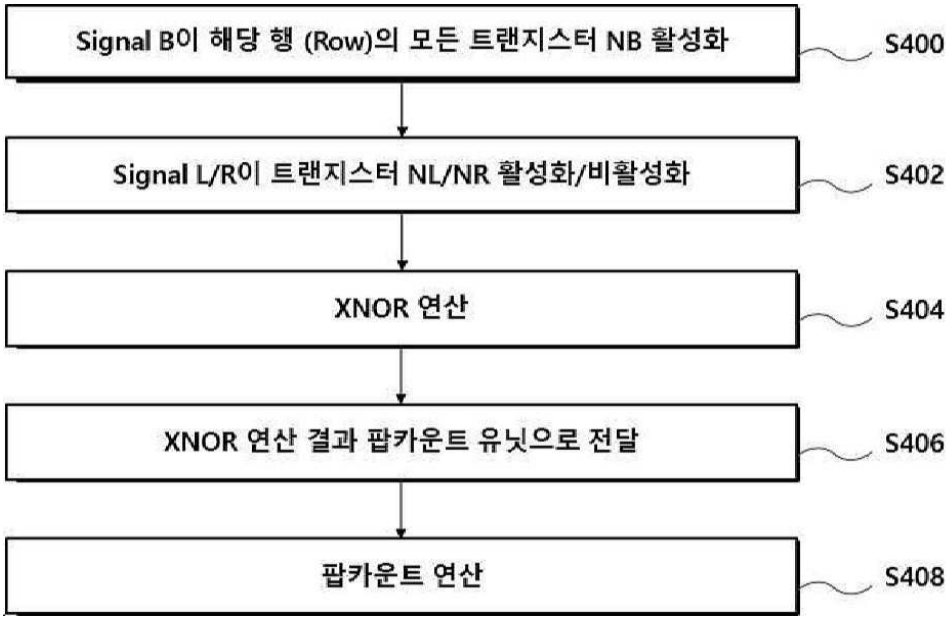


XNOR operation			
Data input			Output
Q	L	R=L'	
0	0	1	1
0	1	0	0
1	0	1	0
1	1	0	1

도면3


204
Column-wise wire driver
Row-wise wire driver
3T
6T
Decod
Popcount unit
MIVs
200
202
206

도면4


Signal B이 해당 행 (Row)의 모든 트랜지스터 NB 활성화
S400
Signal L/R이 트랜지스터 NL/NR 활성화/비활성화
S402
XNOR 연산
S404
XNOR 연산 결과 팝카운트 유닛으로 전달
S406
팝카운트 연산
S408