



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0096296
(43) 공개일자 2019년08월19일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01) G06N 3/04 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2019-0015142
(22) 출원일자 2019년02월08일
심사청구일자 2019년02월08일
(30) 우선권주장
1020180015670 2018년02월08일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
정성우
서울특별시 강남구 도곡로28길 8, 104동 1802호 (도곡동, 도곡1차아이파크아파트)
남상준
서울특별시 성북구 개운사길 42-4, 201호 (안암동5가)
공영호
서울특별시 동대문구 무학로49길 25, 206호 (용두동, 대명빌딩)
(74) 대리인
송인호, 윤형근, 최영중, 최관락

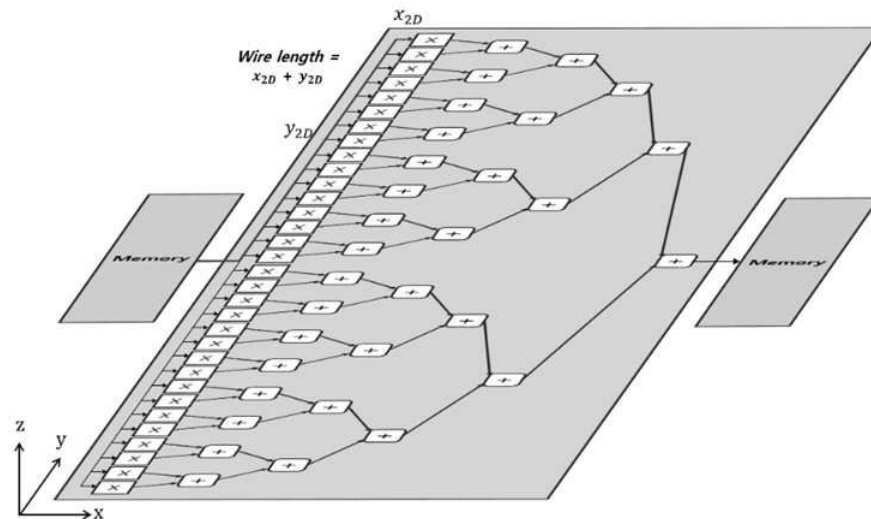
전체 청구항 수 : 총 4 항

(54) 발명의 명칭 모놀리식 3D 기반 뉴로모픽 칩

(57) 요약

본 발명은 모놀리식 3D 기반 뉴로모픽 칩을 개시한다. 본 발명에 따르면, 메모리 및 수직으로 적층된 복수의 레이어로 구성되고, 각 레이어는 상기 메모리에서 출력된 입력 뉴런 및 그에 상응하는 가중치를 입력 받아 곱셈 및 덧셈 연산을 수행하는 곱하기 연산자 및 더하기 연산자를 포함하는 로직을 포함하되, 상기 복수의 레이어는 모놀리식 3D 비아로 연결되는 모놀리식 3D 기반 뉴로모픽 칩이 제공된다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 1711057659
 부처명 과학기술정보통신부
 연구관리전문기관 한국연구재단
 연구사업명 나노·소재기술개발
 연구과제명 초절전 IoT 소자 결합 3차원 플랫폼 집적 공정과 최적 아키텍처 개발, 검증
 기 여 율 1/2
 주관기관 인하대학교 산학협력단
 연구기간 2017.07.01 ~ 2018.04.30

이 발명을 지원한 국가연구개발사업

과제고유번호 1711048678
 부처명 과학기술정보통신부
 연구관리전문기관 한국연구재단
 연구사업명 나노·소재기술개발
 연구과제명 모노리틱 3D 기반 3차원 고집적 뉴로모픽 인터커넥트 설계/분석 플랫폼
 기 여 율 1/2
 주관기관 동국대학교 산학협력단
 연구기간 2017.03.01 ~ 2018.02.28

명세서

청구범위

청구항 1

모놀리식 3D 기반 뉴로모픽 칩으로서,

메모리; 및

수직으로 적층된 복수의 레이어로 구성되고, 각 레이어는 상기 메모리에서 출력된 입력 뉴런 및 그에 상응하는 가중치를 입력 받아 곱셈 및 덧셈 연산을 수행하는 곱하기 연산자 및 더하기 연산자를 포함하는 로직을 포함하되,

상기 복수의 레이어는 모놀리식 3D 비아로 연결되는 모놀리식 3D 기반 뉴로모픽 칩.

청구항 2

제1항에 있어서,

상기 복수의 레이어 중 최하위에 배치된 제1 레이어는 k개의 곱하기 연산자 및 k개의 더하기 연산자를 포함하고,

나머지 레이어는 k개의 곱하기 연산자 및 k-1개의 더하기 연산자를 포함하는 모놀리식 3D 기반 뉴로모픽 칩.

청구항 3

제2항에 있어서,

상기 k는 2의 배수로 설정되는 모놀리식 3D 기반 뉴로모픽 칩.

청구항 4

제2항에 있어서,

상기 k는 8 이상이고, 상기 복수의 레이어는 2 내지 4의 레이어를 포함하는 모놀리식 3D 기반 뉴로모픽 칩.

발명의 설명

기술 분야

[0001] 본 발명은 모놀리식 3D 기반 뉴로모픽 칩에 관한 것이다.

배경 기술

[0002] 각종 사물에 센서와 통신 기능을 내장하는 사물인터넷(Internet of Things) 시장이 빠르게 성장하고 있다.

[0003] 최근에는 각각 사물이 인공지능처럼 생각을 할 수 있는 기능까지 고려되어지고 있고, 이에 따라, 뉴럴 네트워크(Neural Network) 연산만을 효율적으로 처리하기 위한 뉴로모픽 칩(Neuromorphic Chip)이 제안되고 있다.

[0004] 뉴로모픽 칩을 사용함으로써, CPU와 GPU 같은 기존 프로세서(Processor) 보다 작은 면적으로 전력 소모는 줄이면서 성능은 더욱 향상시킬 수 있다. 뉴로모픽 칩은 크게 입력 뉴런(Input Neuron), 가중치(Weight), 출력 뉴런(Output Neuron)을 저장하는 메모리(Memory)와, 입력 뉴런과 가중치를 특정 알고리즘에 따라 더하고 곱하여, 출력 뉴런을 계산하는 로직(Logic)으로 구성되어 있다.

[0005] 뉴로모픽 칩의 성능은 처리량(Throughput)이기 때문에, 최근 뉴로모픽 칩의 로직은 곱하기 연산자(Multiplier)와 더하기 연산자(Adder)의 개수를 크게 늘리고 있다.

[0006] 하지만 기존의 2D구조 기반의 뉴로모픽 칩에서 무작정 곱하기 연산자 및 더하기 연산자들의 개수를 늘리는 것은

한계가 있다.

[0007] 공급하기 연산자 및 더하기 연산자의 개수를 늘리는 경우 물리적인 크기가 증가하기 때문에 와이어의 길이가 길어지게 되고, 이는 데이터(Data)가 이동하는 latency도 길어질 뿐만 아니라, 와이어에서 소모하는 전력도 크게 증가하게 된다.

[0008] 즉, 뉴로모픽 칩을 사물인터넷 기기에 내장하기 위해서는, 크기가 작아야할 뿐만 아니라, 전력 소모도 작아야한다. 또한, 작은 면적에서도 충분한 계산능력을 보유해야 하지만, 2D구조 기반의 뉴로모픽 칩에서는 공정 미세화의 한계와 와이어 오버헤드 증가로 인해서 작은 면적에서 성능을 높이는 것은 한계가 많다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) KR 공개특허 10-2018-0086152

발명의 내용

해결하려는 과제

[0010] 상기한 종래기술의 문제점을 해결하기 위해, 본 발명은 전력 소모는 줄이면서도 성능은 향상시킬 수 있는 모놀리식 3D 기반 뉴로모픽 칩을 제안하고자 한다.

과제의 해결 수단

[0011] 상기한 바와 같은 목적을 달성하기 위하여, 본 발명의 일 실시예에 따르면, 모놀리식 3D 기반 뉴로모픽 칩으로서, 메모리; 및 수직으로 적층된 복수의 레이어로 구성되고, 각 레이어는 상기 메모리에서 출력된 입력 뉴런 및 그에 상응하는 가중치를 입력 받아 곱셈 및 덧셈 연산을 수행하는 공급하기 연산자 및 더하기 연산자를 포함하는 로직을 포함하되, 상기 복수의 레이어는 모놀리식 3D 비아로 연결되는 모놀리식 3D 기반 뉴로모픽 칩이 제공된다.

[0012] 상기 복수의 레이어 중 최하위에 배치된 제1 레이어는 k개의 공급하기 연산자 및 k개의 더하기 연산자를 포함하고, 나머지 레이어는 k개의 공급하기 연산자 및 k-1개의 더하기 연산자를 포함할 수 있다.

[0013] 상기 k는 2의 배수로 설정될 수 있다.

[0014] 상기 k는 8 이상이고, 상기 복수의 레이어는 2 내지 4의 레이어를 포함할 수 있다.

발명의 효과

[0015] 본 발명에 따르면, 뉴로모픽 칩을 수직으로 적층된 복수의 레이어로 구성하고, 이들을 모놀리식 3D 비아로 연결하여 전력 소모를 한층 줄일 수 있다.

도면의 간단한 설명

[0016] 도 1은 기존 뉴로모픽 칩의 구조를 도시한 도면이다.

도 2는 본 발명의 바람직한 일 실시예에 따른 모놀리식 3D 기반 뉴로모픽 칩 구조를 도시한 도면이다.

도 3은 FPGA에서 PE-8 내지 PE-32를 2D 구조(1-layer)로 구성한 경우와 3D 구조(2-layer, 4-layer)로 구성한 경우의 처리 성능을 비교한 것이다.

도 4는 ASIC에서 PE-8 및 PE-16을 2D 구조(1-layer)로 구성한 경우와 3D 구조(2-layer)로 구성한 경우의 처리 성능을 비교한 것이다.

발명을 실시하기 위한 구체적인 내용

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

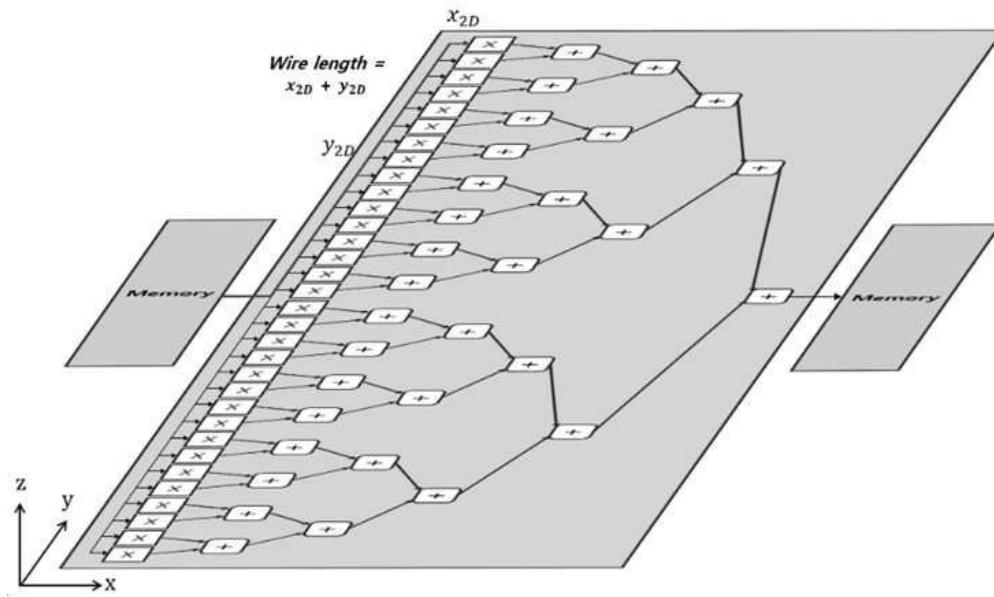
- [0018] 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.
- [0020] 도 1은 기존 뉴로모픽 칩의 구조를 도시한 도면이다.
- [0021] 도 1은 32개의 곱하기 연산자와 31개의 더하기 연산자를 포함하는 뉴로모픽 칩의 구조를 도시한 것이다.
- [0022] 이는 PE(Processing Element)-32 뉴로모픽 칩으로 명명된다.
- [0023] 도 1의 뉴로모픽 칩은 2D 구조(1-layer)를 가지며, 메모리에서 하나의 입력 뉴런과 그에 상응하는 하나의 가중치가 출력되고, 각각 곱하기 연산자에 입력되어 곱셈 연산이 이루어진다.
- [0024] 메모리의 출력 포트와 거리가 가까운 곱하기 연산자에 데이터가 입력되는 경우, 와이어(wire)의 길이가 짧기 때문에 지연(latency)도 작을 뿐만 아니라, 전력 소모도 적다.
- [0025] 그러나, 메모리 출력 포트와 거리가 먼 곱하기 연산자에 데이터가 입력되는 경우, 와이어의 길이(데이터 전송 길이)가 길어진다.
- [0026] 이에, 와이어를 따라 전송되는 데이터의 지연이 길어질 뿐만 아니라, 와이어에서 발생하는 전력 소모도 커진다.
- [0027] 도 1에서 와이어의 길이(wire length)는 $x_{2D} + y_{2D}$ 이다.
- [0028] 뉴로모픽 칩의 특성상 데이터의 이동이 많기 때문에 긴 와이어에서 발생하는 전력 소모와 짧은 와이어에서 발생하는 전력 소모의 차이가 크다. 또한, 2D 구조 상 뉴로모픽 칩의 연산자 수가 증가하는 만큼 면적이 비례하여 증가하기 때문에 IoT 소자로 활용하기에는 큰 단점이 있다.
- [0029] 본 발명의 바람직한 일 실시예에 따르면, 2D 구조의 뉴로모픽 칩의 크기가 증가함에 따라 발생하는 문제점을 해결하기 위해 모놀리식 3D 비아를 활용하여 3차원 구조를 가진 뉴로모픽 칩을 제안한다.
- [0030] 도 2는 본 발명의 바람직한 일 실시예에 따른 모놀리식 3D 기반 뉴로모픽 칩 구조를 도시한 도면이다.
- [0031] 도 2에 도시된 바와 같이, 본 실시예에 따른 모놀리식 3D 기반 뉴로모픽 칩은 메모리(200) 및 메모리(200)의 출력 포트에 연결되는 복수의 레이어(202-k)로 구성된 로직을 포함할 수 있다.
- [0032] 도 2는 4개의 레이어(202-1 내지 202-4)를 포함하는 뉴로모픽 칩 구조를 도시한 것으로서, 도 2에서 와이어의 길이는 $x_{2D} + \frac{1}{4} y_{2D} + z_{2D}$ 이다. 여기서, 높이방향으로의 와이어 길이 z_{2D} 는 $\frac{1}{2} y_{2D}$ 보다 작다.
- [0033] 본 실시예에 따르면, 최하위에 배치된 레이어(202-1)를 제외한 나머지 레이어(202-2 내지 202-4)에는 k개의 곱하기 연산자와 k-1개의 더하기 연산자가 포함된다.
- [0034] 그리고, 최하위에 배치된 제1 레이어(202-1)에는 k개의 곱하기 연산자와 k개의 더하기 연산자가 포함된다.
- [0035] 바람직하게, k는 2의 배수로 설정될 수 있고, 보다 바람직하게, k는 8이상일 수 있다.
- [0036] 예를 들어, PE-32 뉴로모픽 칩에서 k는 8일 수 있고, 이에 따라 4개의 레이어가 포함될 수 있다.
- [0037] 본 발명의 일 실시예에 따르면, 수직으로 나란하게 배치된 복수의 레이어(202-n)는 모놀리식 3D 비아(Monolithic 3D Via, 204)를 통해 연결된다.
- [0038] 여기서 모놀리식 3D 집적 구조는 3차원 시퀀셜 집적구조로 정의된다.
- [0039] 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.
- [0040] 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 웨이퍼의 두께에 따른 문제점들을 해결 가능한 새로운 대안 제조 기술이다.
- [0041] 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.

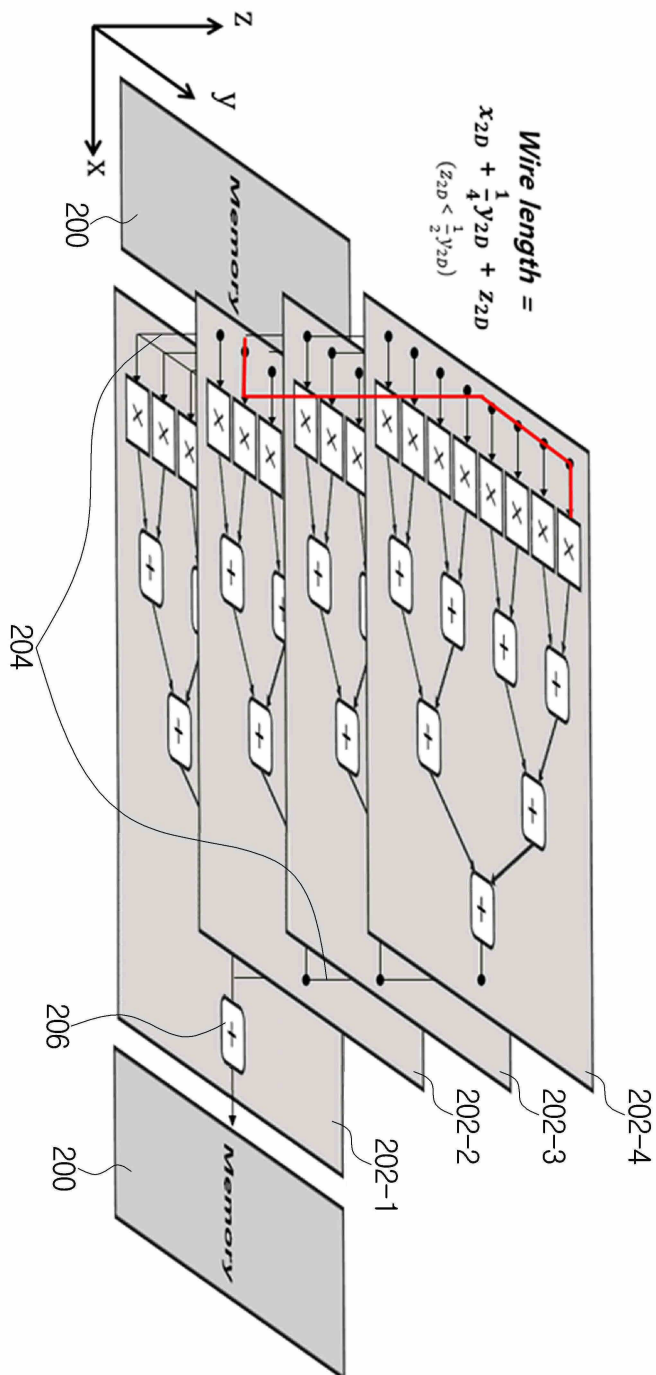
- [0042] 메모리(200)에서 하나의 입력 뉴런과 그에 상응하는 하나의 가중치가 출력되는 경우, 이값은 제1 레이어(202-1) 및 이와 모놀리식 3D 비아(204)로 연결된 상위 레이어들(202-2 내지 202-4)의 곱하기 연산자들로 입력된다.
- [0043] 모놀리식 3D 비아를 통해 수직으로 적층된 구조를 이용하는 경우, 메모리(200)에서 출력된 데이터가 곱하기 연산자로 입력되는 거리가 짧아지기 때문에 전력 소모가 줄어들게 된다.
- [0044] 또한, 각 레이어(202-n)의 마지막 덧하기 연산이 이루어진 후, 해당 값은 모놀리식 3D 비아를 통해 최하위에 배치된 제1 레이어(202-1)의 최종 덧하기 연산자(206)에 입력된다.
- [0045] 도 3은 FPGA에서 PE-8 내지 PE-32를 2D 구조(1-layer)로 구성한 경우와 3D 구조(2-layer, 4-layer)로 구성한 경우의 처리 성능을 비교한 것이다.
- [0046] 도 3을 참조하면, 1-layer와 비교하여 2-layer의 경우 PE-8, PE-16, 그리고 PE-32에서 각각 6.2%, 7.1%, 11.0% 씩 향상된다.
- [0047] 1-layer와 비교하여 4-layer의 경우 PE-8, PE-16, 그리고 PE-32에서 각각 11.5%, 16.0%, 16.7% 씩 향상된다.
- [0048] 상기한 결과를 볼 때, 하나의 레이어에 PE를 8 이상으로 구성하는 경우 성능이 대폭 향상되는 것을 알 수 있다.
- [0049] 또한, 4개의 레이어 이상으로 적층하는 경우에는 과도한 발열 현상이 발생할 수 있기 때문에 레이어의 적층 수는 2 내지 4가 바람직하다.
- [0050] 도 4는 ASIC에서 PE-8 및 PE-16을 2D 구조(1-layer)로 구성한 경우와 3D 구조(2-layer)로 구성한 경우의 처리 성능을 비교한 것이다.
- [0051] 도 4를 참조하면, 1-layer와 비교하여 2-layer의 경우 PE-8과 PE-16에서 각각 14%, 17% 씩 향상된다.
- [0052] 여기서 2-layer일 때, Cell delay가 1-layer일 때와 다른 값인 것을 확인할 수 있는데, 이는 비교한 두 Critical path delay가 같은 Path가 아니기 때문이다.
- [0053] 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

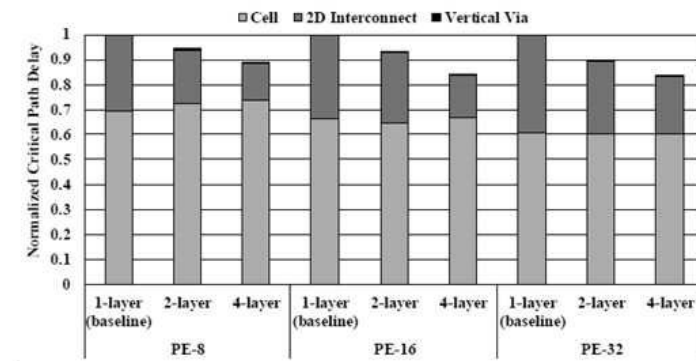
도면1



도면2



도면3



도면4

