(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년01월26일
(11) 등록번호 10-2207909
(24) 등록일자 2021년01월20일

(51) 국제특허분류(Int. Cl.)
G11C 7/12 (2006.01) *G06F 9/30* (2018.01)
G06N 20/00 (2019.01) *G11C 8/08* (2006.01)
(52) CPC특허분류
G11C 7/12 (2013.01)
G06F 9/3001 (2013.01)
(21) 출원번호 10-2019-0021509
(22) 출원일자 2019년02월25일
심사청구일자 2019년02월25일
(65) 공개번호 10-2020-0103262
(43) 공개일자 2020년09월02일
(56) 선행기술조사문헌
KR100814255 B1*
KR1020130058294 A*
US20160232951 A1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
정성욱
서울특별시 서대문구 연세로 50, 제3공학관 C513호(신촌동,연세대학교)
이영규
서울특별시 서대문구 연세로 50, 제3공학관 C206호(신촌동,연세대학교)
송병규
서울특별시 서대문구 연세로 50, 제3공학관 C206호(신촌동,연세대학교)
(74) 대리인
김연권

전체 청구항 수 : 총 13 항 심사관 : 박소정

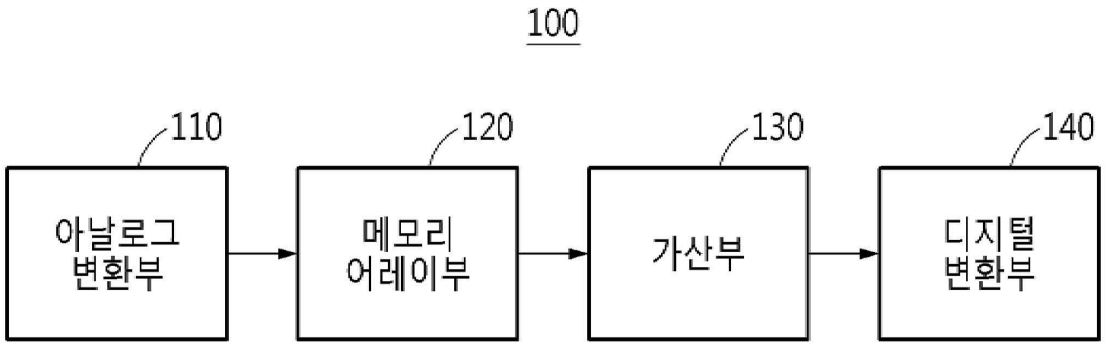
(54) 발명의 명칭 **비트라인의 전하 공유에 기반하는 CIM 장치 및 그 동작 방법**

(57) 요 약

본 발명은 비트라인에 연결된 커패시터의 비율(cap ratio)를 조절하여 멀티비트 입력(multi-bit input)과 멀티비트 웨이트(multi-bit weight)의 결합(multiply) 연산 기술에 관한 것으로서, CIM(Computation In Memory) 장치가 멀티비트의 디지털 전압을 복수의 아날로그 전압으로 변환하여 복수의 비트라인에 상기 변환된 복수의 아날로

(뒷면에 계속)

대 표 도 - 도1



그 전압을 각각 프리차지하고, 상기 복수의 비트라인에 각각 연결되고, 웨이트(weight)가 각각 저장된 복수의 메모리셀과 서로 다른 비율로 미리 결정된 복수의 커패시터를 포함하고, 상기 복수의 커패시터 각각은 상기 복수의 비트라인에 각각 연결되는 메모리 어레이부에서, 워드 라인에 입력에 따라 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하며, 상기 복수의 비트라인에 각각 연결된 스위치를 제어하여 상기 서로 다른 비율의 합이 반영된 결합 결과를 가산하는 기술에 관한 것이다.

(52) CPC특허분류
G06N 20/00 (2019.01)
G11C 8/08 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2017R1A2B2006679
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	중견연구자지원사업
연구과제명	Domain Wall Motion 시냅스 기반의 On-Chip 지도-자율 통합학습 뉴로모픽 SoC 개발(2/3)
기 여 율	1/1
과제수행기관명	연세대학교 산학협력단
연구기간	2018.03.01 ~ 2019.02.28

명 세 서

청구범위

청구항 1

멀티비트의 디지털 전압을 복수의 아날로그 전압으로 변환하여 복수의 비트라인에 상기 변환된 복수의 아날로그 전압을 각각 프리차지하는 아날로그 변환부;

상기 복수의 비트라인에 각각 연결되고, 웨이트(weight)가 각각 저장된 복수의 메모리셀과 서로 다른 비율로 정전용량(capacitance)이 미리 결정된 복수의 커패시터를 포함하고, 상기 복수의 커패시터 각각은 상기 복수의 비트라인에 각각 연결되며, 워드 라인을 통해 구동 전압이 인가될 시 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 메모리 어레이부; 및

상기 복수의 비트라인에 각각 연결된 스위치를 제어하여 서로 다른 비율의 합으로 나눈 결합 결과를 가산하는 가산부를 포함하고,

상기 가산부는 상기 복수의 비트라인에 각각 연결된 스위치를 연결하여 상기 서로 다른 비율의 합으로 상기 가산된 결합 결과를 나눈 값에 서로 다른 비율의 정전용량 합을 곱하여 상기 각각 프리차지된 아날로그 전압의 전하량을 보존하는

CIM(Computation In Memory) 장치.

청구항 2

제1항에 있어서,

상기 메모리 어레이부는 상기 복수의 커패시터의 서로 다른 비율에 기초하여 상기 각각 저장된 웨이트(weight)의 최상위 비트(most significant bit, MSB)와 최하위 비트(least significant bit, LSB)를 구분하는

CIM(Computation In Memory) 장치.

청구항 3

제1항에 있어서,

상기 아날로그 변환부는 상기 멀티비트의 디지털 전압을 상기 복수의 비트라인 개수 또는 상기 복수의 메모리셀 개수에 기초하여 동일한 간격을 갖는 복수의 아날로그 전압으로 변환하는

CIM(Computation In Memory) 장치.

청구항 4

제1항에 있어서,

상기 복수의 커패시터의 비율은 상기 복수의 비트라인과 각각 연결되는 선의 길이 차이 또는 상기 복수의 커패시터를 형성하는 금속물질 종류의 차이에 기초하여 상기 서로 다른 비율의 2의 제곱 값으로 상기 정전 용량(capacitance)이 미리 결정되는

CIM(Computation In Memory) 장치.

청구항 5

제1항에 있어서,

상기 메모리 어레이부는 상기 워드 라인을 통해 구동 전압이 인가될 시 상기 복수의 메모리셀 중 어느 하나에 저장된 웨이트(weight)가 로우 상태일 경우, 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인을 통해 상기 프리차지된 아날로그 전압을 디스차지하는

CIM(Computation In Memory) 장치.

청구항 6

제5항에 있어서,

상기 메모리 어레이부는 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인을 통해 상기 프리차지된 아날로그 전압을 디스차지하여 상기 결합 결과를 로우 상태로 출력하는

CIM(Computation In Memory) 장치.

청구항 7

제1항에 있어서,

상기 메모리 어레이부는 상기 워드 라인을 통해 구동 전압이 인가될 시 상기 복수의 메모리셀 중 어느 하나에 저장된 웨이트(weight)가 하이 상태일 경우, 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인에 프리차지된 아날로그 전압을 유지하는

CIM(Computation In Memory) 장치.

청구항 8

제7항에 있어서,

상기 메모리 어레이부는 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인에 프리차지된 아날로그 전압을 유지하여 상기 결합 결과를 상기 유지된 아날로그 전압으로 출력하는

CIM(Computation In Memory) 장치.

청구항 9

삭제

청구항 10

제1항에 있어서,

상기 가산부는 상기 복수의 비트라인에 각각 연결된 스위치를 연결하여 상기 복수의 비트라인의 전하를 공유하여 상기 각각 프리차지된 아날로그 전압의 전하량을 보존하는

CIM(Computation In Memory) 장치.

청구항 11

제1항에 있어서,

상기 가산된 결합 결과를 디지털 값으로 변환하는 디지털 변환부를 더 포함하는

CIM(Computation In Memory) 장치.

청구항 12

아날로그 변환부에서, 멀티비트의 디지털 전압을 복수의 아날로그 전압으로 변환하여 복수의 비트라인에 상기 변환된 복수의 아날로그 전압을 각각 프리차지하는 단계;

상기 복수의 비트라인에 각각 연결되고, 웨이트(weight)가 각각 저장된 복수의 메모리셀과 서로 다른 비율로 정전용량(capacitance)이 미리 결정된 복수의 커패시터를 포함하고, 상기 복수의 커패시터 각각은 상기 복수의 비트라인에 각각 연결되는 메모리 어레이부에서, 워드 라인을 통해 구동 전압이 인가될 시 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 단계; 및

가산부에서, 상기 복수의 비트라인에 각각 연결된 스위치를 제어하여 서로 다른 비율의 합으로 나눈 결합 결과

를 가산하는 단계를 포함하고,

상기 복수의 비트라인에 각각 연결된 스위치를 제어하여 서로 다른 비율의 합으로 나눈 결합 결과를 가산하는 단계는

상기 복수의 비트라인에 각각 연결된 스위치를 연결하여 상기 서로 다른 비율의 합으로 상기 가산된 결합 결과를 나눈 값에 서로 다른 비율의 정전용량 합을 곱하여 상기 각각 프리차지된 아날로그 전압의 전하량을 보존하는 단계를 포함하는

CIM(Computation In Memory) 장치의 동작 방법.

청구항 13

제12항에 있어서,

상기 워드 라인을 통해 구동 전압이 인가될 시 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 단계는,

상기 복수의 커패시터의 서로 다른 비율에 기초하여 상기 각각 저장된 웨이트(weight)의 최상위 비트(most significant bit, MSB)와 최하위 비트(least significant bit, LSB)를 구분하는

CIM(Computation In Memory) 장치의 동작 방법.

청구항 14

제12항에 있어서,

상기 워드 라인을 통해 구동 전압이 인가될 시 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 단계는,

상기 워드 라인을 통해 구동 전압이 인가될 시 상기 복수의 메모리셀 중 어느 하나에 저장된 웨이트(weight)가 로우 상태일 경우, 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인을 통해 상기 프리차지된 아날로그 전압을 디스차지하여 상기 결합 결과를 로우 상태로 출력하는 단계; 및

상기 메모리 어레이부는 상기 워드 라인을 통해 구동 전압이 인가될 시 상기 복수의 메모리셀 중 어느 하나에 저장된 웨이트(weight)가 하이 상태일 경우, 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인에 프리차지된 아날로그 전압을 유지하여 상기 결합 결과를 상기 유지된 아날로그 전압으로 출력하는 단계를 포함하는

CIM(Computation In Memory) 장치의 동작 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 머신러닝(machine learning)의 MAC(Multiply-Accumulate) 오퍼레이션(operation)에 있어, 원 비트(1-bit) 웨이트(weight) 사용에 따른 분류 정확도(classification accuracy)의 감소를 개선하는 기술에 관한 것으로, 보다 구체적으로, 비트라인에 연결된 커패시터의 비율(cap ratio)를 조절하여 멀티비트 입력(multi-bit input)과 멀티비트 웨이트(multi-bit weight)의 결합(multiply) 연산을 수행하여 분류 정확도를 향상하는 기술에 관한 것이다.

배 경 기 술

[0002] 종래 기술에 따르면, 머신러닝(machine learning)의 MAC(Multiply-Accumulate) 오퍼레이션(operation)에서 대량의 데이터 처리가 요구됨에 따라 데이터 액세스(data access) 및 전달(transfer)에 필요한 파워 소비(power consumption)가 계산력(computation power)보다 많은 양을 차지할 수 있다.

[0003] 예를 들어, 기존 폰노이만(von Neumann) 구조를 사용하는 종래 기술은 많은 양의 데이터를 처리할 경우, 계산에 사용되는 파워(power)보다 메모리 엑세스 및 데이터 이동(data movement)에 더 많은 양의 파워가 소모되어 병목(bottleneck) 현상이 발생할 수 있다.

[0004] 따라서, 데이터 엑세스 및 전달의 파워를 감소시켜 MAC 오퍼레이션의 효율성을 높이기 위해 메모리(memory) 내에서 MAC 오퍼레이션을 수행하는 방법인 CIM(Computation In Memory)이 고안되었고, CIM은 머신러닝에 필요한 MAC 오퍼레이션의 병렬(parallel)한 처리를 통한 높은 쓰루풋(throughput)이 요구된다.

[0005] 그러나, 종래의 CIM 구조(architecture)들은 원 비트 웨이트만 사용하여 분류 정확도(classification accuracy)가 소프트웨어(software)를 사용했을 때보다 감소할 수 있다.

[0006] 즉, 종래의 CIM 구조는 메모리셀에 저장된 웨이트에 따라 비트라인에 프리차지된 전압을 디스차지 또는 유지하는데, 메모리셀에 원 비트 웨이트만 저장 가능하므로 데이터의 분류 정확도가 감소할 수 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국등록특허 제10-1698632호, "전하 공유 디지털-아날로그 변환기 및 연속 근사 아날로그-디지털 변환기"

(특허문헌 0002) 미국등록특허 제8867263호, "MULTIPORT MEMORY WITH MATCHING ADDESS AND DATA LINE CONTROL"

(특허문헌 0003) 한국등록특허 제10-1907028호, "아날로그 디지털 인터페이스 SRAM 구조"

(특허문헌 0004) 한국등록특허 제10-1842322호, "공유된 비트 라인을 갖는 비휘발성 메모리에 대한 비트 라인 사전충전 스킵"

비특허문헌

[0008] (비특허문헌 0001) Avishek Biswas, Anantha P.Chandrakasan, "Conv-RAM: An Energy-Efficient SRAM with Embedded Convolution Computation for Low-Power CNN-Based Machine Learning Applications", ISSCC, pp.488-489, 2018

발명의 내용

해결하려는 과제

[0009] 본 발명은 서로 다른 비율로 미리 설정된 커패시터에 의해 서로 다른 커패시터 비율을 갖는 비트라인들의 전하 공유(charge sharing)을 통해 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행하는 것을 목적으로 한다.

[0010] 본 발명은 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행함으로써 MAC(Multiply-Accumulate) 오퍼레이션의 분류 정확도(classification accuracy)를 향상시키는 것을 목적으로 한다.

[0011] 본 발명은 서로 다른 비율로 미리 설정된 커패시터에 의해 서로 다른 커패시터 비율에 기반하여 멀티비트 웨이트의 MSB(Most Significant Bit)와 LSB(Least Significant Bit)를 구분하는 것을 목적으로 한다.

[0012] 본 발명은 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행함으로써 연산을 위한 에너지 소모를 줄이면서 머신러닝에서 요구되는 많은 양의 데이터를 처리하는 것을 목적으로 한다.

[0013] 본 발명은 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행함으로써 멀티비트 입력에 해당하는 많은 양의 데이터를 처리하면서도 병목(bottleneck)현상의 발생을 억제하는 것을 목적으로 한다.

과제의 해결 수단

[0014] 본 발명의 일실시예에 따르면 CIM(Computation In Memory) 장치는 멀티비트의 디지털 전압을 복수의 아날로그 전압으로 변환하여 복수의 비트라인에 상기 변환된 복수의 아날로그 전압을 각각 프리차지하는 아날로그 변환부, 상기 복수의 비트라인에 각각 연결되고, 웨이트(weight)가 각각 저장된 복수의 메모리셀과 서로 다른 비율로 미리 결정된 복수의 커패시터를 포함하고, 상기 복수의 커패시터 각각은 상기 복수의 비트라인에 각각 연결되며, 워드 라인에 입력에 따라 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을

결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 메모리 어레이부 및 상기 복수의 비트라인에 각각 연결된 스위치를 제어하여 상기 서로 다른 비율의 합이 반영된 결합 결과를 가산하는 가산부를 포함할 수 있다.

[0015] 상기 메모리 어레이부는 상기 복수의 커패시터의 서로 다른 비율에 기초하여 상기 각각 저장된 웨이트(weight)의 최상위 비트(most significant bit, MSB)와 최하위 비트(least significant bit, LSB)를 구분할 수 있다.

[0016] 상기 아날로그 변환부는 상기 멀티비트의 디지털 전압을 상기 복수의 비트라인 개수 또는 상기 복수의 메모리셀 개수에 기초하여 동일한 간격을 갖는 복수의 아날로그 전압으로 변환할 수 있다.

[0017] 상기 복수의 커패시터의 비율은 상기 복수의 비트라인과 각각 연결되는 선의 길이 차이 또는 상기 복수의 커패시터를 형성하는 금속물질 종류의 차이에 기초하여 상기 서로 다른 비율의 2의 제곱 값으로 미리 결정될 수 있다.

[0018] 상기 메모리 어레이부는 상기 워드 라인을 통해 구동 전압이 인가될 시 상기 복수의 메모리셀 중 어느 하나에 저장된 웨이트(weight)가 로우 상태일 경우, 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인을 통해 상기 프리차지된 아날로그 전압을 디스차지할 수 있다.

[0019] 상기 메모리 어레이부는 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인을 통해 상기 프리차지된 아날로그 전압을 디스차지하여 상기 결합 결과를 로우 상태로 출력할 수 있다.

[0020] 상기 메모리 어레이부는 상기 워드 라인을 통해 구동 전압이 인가될 시 상기 복수의 메모리셀 중 어느 하나에 저장된 웨이트(weight)가 하이 상태일 경우, 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인에 프리차지된 아날로그 전압을 유지할 수 있다.

[0021] 상기 메모리 어레이부는 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인에 프리차지된 아날로그 전압을 유지하여 상기 결합 결과를 상기 유지된 아날로그 전압으로 출력할 수 있다.

[0022] 상기 가산부는 상기 복수의 비트라인에 각각 연결된 스위치를 연결하여 상기 서로 다른 비율의 합에 상기 서로 다른 비율의 합과 상기 가산된 결합 결과의 비율을 반영하여 상기 각각 프리차지된 아날로그 전압의 전하량을 보존할 수 있다.

[0023] 상기 가산부는 상기 복수의 비트라인에 각각 연결된 스위치를 연결하여 상기 복수의 비트라인의 전하를 공유하여 상기 각각 프리차지된 아날로그 전압의 전하량을 보존할 수 있다.

[0024] 본 발명의 일실시예에 따르면 CIM(Computation In Memory) 장치는 상기 가산된 결합 결과를 디지털 값으로 변환하는 디지털 변환부를 더 포함할 수 있다.

[0025] 본 발명의 일실시예에 따르면 CIM(Computation In Memory) 장치의 동작 방법은 아날로그 변환부에서, 멀티비트의 디지털 전압을 복수의 아날로그 전압으로 변환하여 복수의 비트라인에 상기 변환된 복수의 아날로그 전압을 각각 프리차지하는 단계, 상기 복수의 비트라인에 각각 연결되고, 웨이트(weight)가 각각 저장된 복수의 메모리셀과 서로 다른 비율로 미리 결정된 복수의 커패시터를 포함하고, 상기 복수의 커패시터 각각은 상기 복수의 비트라인에 각각 연결되는 메모리 어레이부에서, 워드 라인에 입력에 따라 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 단계 및 가산부에서, 상기 복수의 비트라인에 각각 연결된 스위치를 제어하여 상기 서로 다른 비율의 합이 반영된 결합 결과를 가산하는 단계를 포함할 수 있다.

[0026] 상기 워드 라인에 입력에 따라 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 단계는, 상기 복수의 커패시터의 서로 다른 비율에 기초하여 상기 각각 저장된 웨이트(weight)의 최상위 비트(most significant bit, MSB)와 최하위 비트(least significant bit, LSB)를 구분할 수 있다.

[0027] 상기 워드 라인에 입력에 따라 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 단계는, 상기 워드 라인을 통해 구동 전압이 인가될 시 상기 복수의 메모리셀 중 어느 하나에 저장된 웨이트(weight)가 로우 상태일 경우, 상기 복수의 메모리셀 중 어느 하나에 연결된 비트라인을 통해 상기 프리차지된 아날로그 전압을 디스차지하여 상기 결합 결과를 로우 상태로 출력하는 단계 및 상기 메모리 어레이부는 상기 워드 라인을 통해 구동 전압이 인가될 시 상기 복수의 메모리셀 중 어느 하나에 저장된 웨이트(weight)가 하이 상태일 경우, 상기 복수의 메모리셀 중 어느 하나

에 연결된 비트라인에 프리차지된 아날로그 전압을 유지하여 상기 결합 결과를 상기 유지된 아날로그 전압으로 출력하는 단계를 포함할 수 있다.

발명의 효과

[0028] 본 발명은 서로 다른 비율로 미리 설정된 커패시터에 의해 서로 다른 커패시터 비율을 갖는 비트라인들의 전하 공유(charge sharing)을 통해 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행할 수 있다.

[0029] 본 발명은 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행함으로써 MAC(Multiply-Accumulate) 오퍼레이션의 분류 정확도(classification accuracy)를 향상시킬 수 있다.

[0030] 본 발명은 서로 다른 비율로 미리 설정된 커패시터에 의해 서로 다른 커패시터 비율에 기반하여 멀티비트 웨이트의 MSB(Most Significant Bit)와 LSB(Least Significant Bit)를 구분할 수 있다.

[0031] 본 발명은 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행함으로써 연산을 위한 에너지 소모를 줄이면서 머신러닝에서 요구되는 많은 양의 데이터를 처리할 수 있다.

[0032] 본 발명은 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행함으로써 멀티비트 입력에 해당하는 많은 양의 데이터를 처리하면서도 병목(bottleneck)현상의 발생을 억제할 수 있다.

도면의 간단한 설명

[0033] 도 1은 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 구성 요소를 설명하는 도면이다.

도 2는 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 회로를 설명하는 도면이다.

도 3은 본 발명의 일실시예에 따른 디지털 입력의 비트 너비를 웨이트의 비트 너비와 같은 4비트로 설정하는 동작을 설명하는 도면이다.

도 4a 내지 도 4c는 본 발명의 일실시예에 따른 메모리 어레이부의 동작을 설명하는 도면이다.

도 5는 본 발명의 일실시예에 따른 가산부의 동작을 설명하는 도면이다.

도 6은 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 동작 방법과 관련된 흐름도를 설명하는 도면이다.

도 7은 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 동작 방법과 관련된 타이밍도를 설명하는 도면이다.

도 8은 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 동작 결과를 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0034] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0035] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0036] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만, 예를 들어 본 발명의 개념에 따른 권리 범위로부터 이탈되지 않은 채, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0037] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성요소들 간의

관계를 설명하는 표현들, 예를 들어 "~사이에"와 "바로~사이에" 또는 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0038] 본 명세서에서 사용한 용어는 단지 특정한 실시예들을 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0039] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0041] 이하, 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0043] 도 1은 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 구성 요소를 설명하는 도면이다.

[0044] 도 1을 참고하면, 본 발명의 일실시예에 따른 CIM 장치(100)는 아날로그 변환부(110), 메모리 어레이부(120) 및 가산부(130)를 포함할 수 있다.

[0045] 본 발명의 일실시예에 따른 아날로그 변환부(110)는 멀티비트의 디지털 전압을 복수의 아날로그 전압으로 변환하여 복수의 비트라인에 복수의 아날로그 전압을 각각 프리차지할 수 있다.

[0046] 일례로, 아날로그 변환부(110)는 멀티비트의 디지털 전압을 복수의 비트라인 개수 또는 복수의 메모리셀 개수에 기초하여 동일한 간격을 갖는 복수의 아날로그 전압으로 변환할 수 있다.

[0047] 예를 들어, 아날로그 변환부(110)는 디지털 아날로그 변환기(Digital to Analog Converter, DAC)를 포함할 수 있다.

[0048] 본 발명의 일실시예에 따르면 아날로그 변환부(110)는 멀티비트로 구성된 디지털 전압을 전달받아 비트라인의 개수에 상응하는 아날로그 전압으로 변환하여 복수의 비트라인에 아날로그 전압을 프리차지할 수 있다.

[0049] 또한, 아날로그 변환부(110)는 멀티비트로 구성된 디지털 전압을 전달받아 메모리셀의 개수에 상응하는 아날로그 전압으로 변환하여 복수의 비트라인에 아날로그 전압을 프리차지할 수 있다.

[0050] 본 발명의 일실시예에 따르면 메모리 어레이부(120)는 복수의 비트라인에 각각 연결되고, 웨이트(weight)가 각각 저장된 복수의 메모리셀과 서로 다른 비율로 미리 결정된 복수의 커패시터를 포함할 수 있다.

[0051] 예를 들어, 복수의 커패시터는 복수의 비트라인에 각각 연결될 수 있다.

[0052] 일례로, 서로 다른 비율로 미리 결정된 복수의 커패시터는 복수의 비트라인 각각에 대한 커패시터 비율을 결정할 수 있다. 예를 들어, 커패시터 비율은 캡 비율(cap ratio)로 지칭될 수 있다.

[0053] 일례로, 메모리 어레이부(120)는 워드 라인에 입력에 따라 각각 저장된 웨이트(weight)와 각각 프리차지된 아날로그 전압을 결합하여 복수의 비트라인을 통해 결합 결과를 각각 출력할 수 있다.

[0054] 본 발명의 일실시예에 따르면 메모리 어레이부(120)는 복수의 커패시터의 서로 다른 비율에 기초하여 복수의 메모리셀에 각각 저장된 웨이트(weight)의 최상위 비트(most significant bit, MSB)와 최하위 비트(least significant bit, LSB)를 구분할 수 있다.

[0055] 예를 들어, 메모리 어레이부(120)는 복수의 커패시터의 비율 중 가장 높은 비율에 해당하는 커패시터가 연결된 비트라인과 연결된 메모리셀에 저장된 웨이트를 최상위 비트(most significant bit, MSB)로 구분할 수 있다.

[0056] 또한, 메모리 어레이부(120)는 복수의 커패시터의 비율 중 가장 낮은 비율에 해당하는 커패시터가 연결된 비트라인과 연결된 메모리셀에 저장된 웨이트를 최하위 비트(least significant bit, LSB)로 구분할 수 있다.

[0057] 따라서, 본 발명은 서로 다른 비율로 미리 설정된 커패시터에 의해 서로 다른 커패시터 비율에 기반하여 멀티비트 웨이트의 MSB와 LSB를 구분할 수 있다.

[0058] 본 발명의 일실시예에 따르면 메모리 어레이부(120)는 워드 라인을 통해 구동 전압이 인가될 시 복수의 메모리 셀 중 어느 하나에 저장된 웨이트(weight)가 로우 상태일 경우, 복수의 메모리셀 중 어느 하나에 연결된 비트라인을 통해 프리차지된 아날로그 전압을 디스차지할 수 있다.

[0059] 일례로, 메모리 어레이부(120)는 복수의 메모리셀 중 어느 하나에 연결된 비트라인을 통해 프리차지된 아날로그 전압을 디스차지하여 결합 결과를 로우 상태로 출력할 수 있다.

[0060] 본 발명의 일실시예에 따르면 메모리 어레이부(120)는 워드 라인을 통해 구동 전압이 인가될 시 복수의 메모리 셀 중 어느 하나에 저장된 웨이트(weight)가 하이 상태일 경우, 복수의 메모리셀 중 어느 하나에 연결된 비트라인에 프리차지된 아날로그 전압을 유지할 수 있다.

[0061] 일례로, 메모리 어레이부(120)는 복수의 메모리셀 중 어느 하나에 연결된 비트라인에 프리차지된 아날로그 전압을 유지하여 결합 결과를 유지된 아날로그 전압으로 출력할 수 있다.

[0062] 예를 들어, 결합 결과는 로우 상태로 출력될 경우, "0"에 상응하는 데이터를 나타낼 수 있고, 하이 상태로 출력될 경우, "1"에 상응하는 데이터를 나타낼 수 있다.

[0063] 따라서, 본 발명은 서로 다른 비율로 미리 설정된 커패시터에 의해 서로 다른 커패시터 비율을 갖는 비트라인들의 전하 공유(charge sharing)을 통해 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행할 수 있다.

[0064] 즉, 본 발명은 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행함으로써 MAC(Multiply-Accumulate) 오퍼레이션의 분류 정확도(classification accuracy)를 향상시킬 수 있다.

[0065] 일례로, 복수의 커패시터의 비율은 복수의 비트라인과 각각 연결되는 선의 길이 차이 또는 복수의 커패시터를 형성하는 금속물질 종류의 차이에 기초하여 서로 다른 비율의 2의 제곱 값으로 미리 결정될 수 있다.

[0066] 예를 들어, 복수의 커패시터의 비율은 8:4:2:1로 2의 제곱 값을 이용하여 미리 결정될 수 있다.

[0067] 본 발명의 일실시예에 따르면 가산부(130)는 복수의 비트라인에 각각 연결된 스위치를 제어하여 커패시터의 서로 다른 비율의 합이 반영된 결합 결과를 가산할 수 있다.

[0068] 일례로, 가산부(130)는 복수의 비트라인에 각각 연결된 스위치를 연결하여 서로 다른 비율의 합에 서로 다른 비율의 합과 결합 결과의 비율을 반영하여 각각 프리차지된 아날로그 전압의 전하량을 보존할 수 있다.

[0069] 또한, 가산부(130)는 복수의 비트라인에 각각 연결된 스위치를 연결하여 복수의 비트라인의 전하를 공유하여 각각 프리차지된 아날로그 전압의 전하량을 보존할 수 있다. 예를 들어, 보존되는 아날로그 전압의 전하량은 비트라인의 전하량에 상응할 수 있다.

[0070] 예를 들어, 가산부(130)는 복수의 비트라인에 각각 연결된 복수의 스위치를 포함할 수 있다.

[0071] 본 발명의 다른 실시예에 따르면 CIM 장치(100)는 디지털 변환부(140)를 더 포함할 수 있다.

[0072] 일례로, 디지털 변환부(140)는 아날로그 전압을 디지털 값(digital value)으로 변환하는 아날로그 디지털 변환기(analog to digital converter, ADC)로서, 가산부(130)에 의하여 가산된 결합 결과를 디지털 값으로 변환할 수 있다.

[0074] 도 2는 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 회로를 설명하는 도면이다.

[0075] 도 2를 참고하면, CIM 장치(200)는 아날로그 변환부(210), 메모리 어레이부(220), 가산부(230) 및 디지털 변환부(240)를 포함할 수 있다.

[0076] 본 발명의 일실시예에 따르면 아날로그 변환부(210)는 멀티비트로 구성된 디지털 전압($X_{in,1}$ 내지 $X_{in,N}$)을 전압 공급부(미도시)로부터 전달받아, 비트라인의 개수 또는 메모리셀의 개수에 상응하는 복수의 아날로그 전압으로 변환할 수 있다.

[0077] 예를 들어, 아날로그 변환부(210)는 비트라인의 개수가 4개일 경우 디지털 전압($X_{in,1}$)을 4개의 아날로그 전압($V_{in,1}$)으로 변환하여 비트라인에 프리차지할 수 있다.

[0078] 본 발명의 일실시예에 따르면 메모리 어레이부(220)는 복수의 메모리셀($W_{1,3}$ 내지 $W_{1,0}$)을 포함할 수 있다.

[0079] 일례로, 복수의 메모리셀($W_{1,3}$ 내지 $W_{1,0}$) 각각은 서로 다른 비트라인에 연결되어 서로 다른 비율의 커패시터 비율에 의해 복수의 웨이트(weight)로 구분될 수 있다. 여기서, 복수의 웨이트는 멀티 웨이트로도 지칭될 수 있으며, 복수의 웨이트는 멀티비트를 나타낼 수 있다.

[0080] 본 발명의 일실시예에 따르면 메모리 어레이부(220)는 복수의 메모리셀($W_{1,3}$ 내지 $W_{1,0}$)에 대한 웨이트를 하이 상태 또는 로우 상태로 저장할 수 있다.

[0081] 일례로, 메모리 어레이부(220)는 복수의 비트라인에 각각 연결된 복수의 커패시터(8C, 4C, 2C, C)를 포함할 수 있다.

[0082] 예를 들어, 복수의 커패시터(8C, 4C, 2C, C)는 서로 다른 비율로 설정될 수 있고, 서로 다른 비율은 8:4:2:1일 수 도 있다.

[0083] 본 발명의 일실시예에 따르면 메모리 어레이부(220)는 복수의 커패시터(8C, 4C, 2C, C)에 미리 설정된 서로 다른 비율에 기반하여 복수의 메모리셀($W_{1,3}$ 내지 $W_{1,0}$)에 저장된 웨이트를 복수의 데이터로 나타낼 수 있다.

[0084] 예를 들어, 메모리 어레이부(220)는 커패시터(8C)와 비트라인을 통해 연결된 메모리셀의 웨이트를 2^3비트로 나타내고, 커패시터(4C)와 비트라인을 통해 연결된 메모리셀의 웨이트를 2^2비트로 나타내고, 커패시터(2C)와 비트라인을 통해 연결된 메모리셀의 웨이트를 2^1비트로 나타내고, 커패시터(C)와 비트라인을 통해 연결된 메모리셀의 웨이트를 2^0비트로 나타낼 수 있다.

[0085] 또한, 메모리 어레이부(220)는 복수의 커패시터(8C, 4C, 2C, C)에 미리 설정된 서로 다른 비율에 기반하여 복수의 메모리셀($W_{1,3}$ 내지 $W_{1,0}$)에 저장된 웨이트의 최상위 비트와 최하위 비트를 구분할 수 있다.

[0086] 본 발명의 일실시예에 따르면 가산부(130)는 복수의 비트라인에 각각 연결된 복수의 스위치를 포함하고, 복수의 스위치의 연결 상태를 제어하여 비트라인의 전하를 공유하여 메모리 어레이부(220)로부터 복수의 비트라인을 통해 출력된 결합 결과를 가산할 수 있다.

[0087] 예를 들어, 가산부(230)는 가산 라인(additional line)으로 지칭될 수 도 있다.

[0088] 상술한 설명에서는 하나의 코너(single corner)만을 설명하였으나, 나머지 코너도 동일하게 동작될 수 있다.

[0089] 본 발명의 일실시예에 따르면 디지털 변환부(240)는 가산된 결합 결과를 아날로그 전압에서 디지털 값(Y_{out})으로 변환할 수 있다.

[0090] 즉, 디지털 변환부(240)는 하기 수학식1에 해당하는 디지털 값(Y_{out})을 출력할 수 있다.

[0092] [수학식1]

[0093]
$$Y_{out} = \sum_{i=1}^{N} W_i X_{in,1}$$

[0095] 수학식1에서, Y_{out}은 디지털 값을 나타낼 수 있고, W는 웨이트를 나타낼 수 있으며, X_{in}은 입력되는 디지털 전압을 나타낼 수 있다.

[0096] 즉, 디지털 변환부(240)는 1부터 N까지의 멀티비트로 구성된 W와 X_{in}의 결합 결과를 디지털 값(Y_{out})으로 출력할 수 있다.

[0098] 도 3은 본 발명의 일실시예에 따른 디지털 입력의 비트 너비를 웨이트의 비트 너비와 같은 4비트로 설정하는 동작을 설명하는 도면이다.

[0099] 도 3을 참고하면, 아날로그 변환부(300)는 디지털 입력(X_{in})이 들어오면 아날로그 전압(V_{in})으로 변환하고, 비트라인을 아날로그 전압(V_{in})으로 프리차지할 수 있다.

[0100] 즉, 아날로그 변환부(300)는 디지털 입력(X_{in})을 미리 설정한 디지털 입력(X_{in})의 비트 너비(bit width)에 비례

하도록 아날로그 전압(V_{in})으로 변환할 수 있다.

[0101] 예를 들어, 미리 설정한 디지털 입력(X_{in})의 비트 너비가 4비트라면, 디지털 입력(X_{in})은 4비트로 인가되며, 아날로그 변환부(300)는 0 내지 VDD를 2^4등분할 수 있다.

[0102] 즉, 아날로그 변환부(300)는 "1111"에 해당하는 디지털 입력(X_{in})을 입력받을 경우, "VDD"로 아날로그 전압(V_{in})을 변환하고, "1110"에 해당하는 디지털 입력(X_{in})을 입력받을 경우, "14VDD/15"로 아날로그 전압(V_{in})을 변환하며, "1101"에 해당하는 디지털 입력(X_{in})을 입력받을 경우, "13VDD/15"로 아날로그 전압(V_{in})을 변환할 수 있다.

[0103] 또한, 아날로그 변환부(300)는 "0010"에 해당하는 디지털 입력(X_{in})을 입력받을 경우, "2VDD/15"로 아날로그 전압(V_{in})을 변환하고, "0001"에 해당하는 디지털 입력(X_{in})을 입력받을 경우, "VDD/15"로 아날로그 전압(V_{in})을 변환하며, "0000"에 해당하는 디지털 입력(Xin)을 입력받을 경우, "0"로 아날로그 전압(V_{in})을 변환할 수 있다.

[0104] 예를 들어, 비트라인의 개수는 멀티비트의 웨이트와 관련되고, 멀티비트의 웨이트가 4비트일 경우, 비트라인의 개수는 4개일 수 있다.

[0105] 이 때, 웨이트의 비트 너비가 4비트라면, 커패시터 비율이 1:2:4:8로 미리 설정된 4개의 비트라인에 아날로그 변환부(300)의 출력 전압이 프리차지될 수 있다. 이럴 경우 4bit(input) x 4bit(weight)의 연산이 수행될 수 있다.

[0106] 따라서, 미리 설정된 커패시터의 비율은 웨이트의 비트 너비를 의미하며, 디지털 입력(X_{in})의 비트 너비와 웨이트의 비트 너비는 다를 수 있다. 즉, 아날로그 변환부(300)는 미리 설정된 커패시터의 비율과 무관하게 설정된 디지털 입력의 비트 너비에 따라 출력 전압을 결정 및 출력할 수 있다.

[0107] 예를 들어, 아날로그 변환부(300)는 디지털 입력이 6비트로 들어오며, 아날로그 변환부(300)는 0 내지 VDD를 2^6 등분을 할 수 있다.

[0108] 즉, 아날로그 변환부(300)는 디지털 입력이 "000000"이라면 0V를 출력하고, 디지털 입력이 "000001"라면 VDD/63V를 출력하며, 디지털 입력이 "100000"라면 VDD/63 x 32 V를 출력할 수 있다.

[0110] 도 4a 내지 도 4c는 본 발명의 일실시예에 따른 메모리 어레이부의 동작을 설명하는 도면이다.

[0111] 구체적으로, 도 4a 내지 도 4c는 메모리 어레이부가 복수의 비트라인에 프리차지된 아날로그 전압과 복수의 메모리셀에 저장된 웨이트를 결합하여 결합 결과를 출력하는 동작을 예시한다.

[0112] 도 4a를 참고하면, CIM 장치(400)는 메모리 어레이부(410)를 포함하고, 메모리 어레이부(410)는 복수의 메모리셀을 포함한다.

[0113] 복수의 메모리셀은 각각 웨이트를 저장하고 있으며, 저장된 웨이트는 하이 상태 또는 로우 상태로 구분될 수 있다.

[0114] 복수의 메모리셀은 워드 라인과 각각 연결되며, 워드 라인을 통해 구동 전압을 인가 받을 경우, 메모리셀에 저장된 웨이트와 비트라인에 프리차지된 아날로그 전압을 결합할 수 있다.

[0115] 예를 들어, 복수의 메모리셀 중 제1 메모리셀(411)은 로우 상태의 웨이트를 저장하고, 제2 메모리셀(412)은 하이 상태의 웨이트를 저장하고 있다.

[0116] 도 4b는 제1 메모리셀(411)을 나타내고, 도 4c는 제2 메모리셀(412)을 나타낼 수 있다.

[0117] 도 4b와 도 4c를 참고하면, 제1 메모리셀(411)과 제2 메모리셀(412)은 10개의 트랜지스터로 구성된 10T SRAM(Static Random Access Memory)셀에 포함될 수 있다.

[0118] 본 발명의 일실시예에 따르면 CIM 장치(400)는 워드 라인을 통해 구동 전압을 인가하고, 메모리 어레이부(410)는 제1 메모리셀(411)에 저장된 웨이트와 비트라인에 프리차지된 전압을 결합한다.

[0119] 여기서, 제1 메모리셀(411)에 저장된 웨이트는 로우 상태로, 비트라인에 프리차지된 전압은 디스차지된다.

[0120] 따라서, 메모리 어레이부(410)는 제1 메모리셀(411)에 연결된 비트라인에 프리차지된 아날로그 전압을 디스차지

하여 결합 결과를 로우 상태로 출력할 수 있다.

[0121] 한편, 메모리 어레이부(410)는 제2 메모리셀(412)에 저장된 웨이트와 비트라인에 프리차지된 전압도 결합한다.

[0122] 여기서, 제2 메모리셀(412)에 저장된 웨이트는 하이 상태로, 비트라인에 프리차지된 전압을 유지한다.

[0123] 따라서, 메모리 어레이부(410)는 제2 메모리셀(412)에 연결된 비트라인에 프리차지된 아날로그 전압을 유지하여 결합 결과를 프리차지된 전압으로 출력할 수 있다.

[0124] 본 발명의 일실시예에 따르면 메모리 어레이부(410)는 복수의 메모리셀에 저장된 웨이트에 따라 비트라인에 프리차지된 아날로그 전압을 디스차지하거나 유지하는데, 복수의 메모리셀에 각각 저장된 웨이트가 로우 상태, 하이 상태, 로우 상태 및 하이 상태일 경우, 디지털 변환 시 "0101"이 되고, "0101"은 "5"로 인식될 수 있다.

[0125] 또한, 복수의 메모리셀에 각각 저장된 웨이트가 하이 상태, 로우 상태, 로우 상태 및 하이 상태일 경우, 디지털 변환 시 "1001"이 되고, "1001"은 "9"로 인식될 수 있다.

[0127] 도 5는 본 발명의 일실시예에 따른 가산부의 동작을 설명하는 도면이다.

[0128] 도 5를 참고하면 CIM 장치 중 하나의 코너(500)는 가산부(510)를 포함하고, 가산부(510)는 복수의 스위치와 그에 연결되는 가산 라인(additional line)에 상응할 수 있다.

[0129] 본 발명의 일실시예에 따르면 가산부(510)는 복수의 스위치를 연결하여 복수의 스위치에 연결된 비트라인의 전하를 공유하여 복수의 비트라인에 프리차지된 아날로그 전압의 전하량을 보존할 수 있다.

[0130] 즉, 가산부(510)는 복수의 비트라인 각각을 통해 전달된 결합 결과($Q_{1,3}$ 내지 $Q_{1,0}$)에 커패시터 비율의 합을 반영하여 입력된 전압의 전하량을 보존할 수 있다.

[0131] 예를 들어, 가산부(510)는 하기 수학식2에 기초하여 프리차지된 아날로그 전압의 전하량을 보존할 수 있다.

[0133] [수학식2]

$$Q = (15C)\left(\frac{4+1}{15}\right)V_{in,1}$$

[0134]

[0136] 수학식2에 따르면, Q는 전하량을 나타낼 수 있고, 15C는 커패시터 비율의 합을 나타낼 수 있으며, 4와 1은 메모리셀에 하이상태로 저장된 값에 대한 결합 결과를 나타낼 수 있고, $V_{in,1}$은 프리차지된 아날로그 전압을 나타낼 수 있다.

[0137] 예를 들어, 커패시터 비율이 1:2:4:8로 설정된 비트라인과 관련된 워드 라인에 구동 전압이 인가되면, 커패시터 값이 1과 4인 비트라인에만 프리차지된 아날로그 전압이 남을 수 있다. 여기서, 비트라인을 더하기 전 총 전하량은 C(Vin + 4Vin)이 될 수 있다.

[0138] 이 상태에서 비트라인을 연결시키면 전체 커패시터 값은 15C가 되며 전하량은 보존될 수 있다. 즉, 15CVout=C(V_{in} + 4V_{in})이 되고, 비트라인을 연결한 후에 발생하는 전압 값인 Vout=(V_{in} + 4V_{in})/15가 될 수 있다.

[0139] 따라서, 분모에 있는 15는 커패시터 비율의 합에 의해 정해진다고 볼 수 있고, 전하량 보존 법칙을 만족시키기 위해 나온 값으로 볼 수 있다.

[0140] 추가적으로, 비트라인을 연결했을 때, 발생하는 전압 값인 Vout이 웨이트와 입력의 결합 결과를 나타낼 수 있다.

[0141] 예를 들어, 비트라인 연결 결과 발생하는 전압 값은 Vout=(4Vin+Vin)/15이며, 이 값은 0101에 해당하는 웨이트 값과, V_{in}에 해당하는 X_{in}값이 곱해진 결과일 수 있다.

[0142] 예를 들어, X_{in}의 비트 너비(bit width)가 4비트이고, X_{in}=1이라면 Vin=VDD/15이되고, Vout={(4+1)/15}*(VDD/15)이므로 웨이트와 입력의 결합 결과를 나타낼 수 있다.본 발명의 일실시예에 따르면 가산부(510)는 멀티비트의 웨이트와 멀티비트의 입력의 결합에 해당하는 아날로그 전압을 가산할 수 있다.

[0143] 일례로, 가산부(510)는 스위치를 연결하여 비트라인의 전하 공유에 기반하여 메모리 어레이부로부터 출력된 결

합 결과를 가산할 수 있다.

[0145] 도 6은 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 동작 방법과 관련된 흐름도를 설명하는 도면이다.

[0146] 도 6을 참고하면, 단계(601)에서 CIM 장치의 동작 방법은 복수의 비트라인에 전압을 프리차지한다.

[0147] 즉, CIM 장치의 동작 방법은 멀티비트의 디지털 전압을 복수의 아날로그 전압으로 변환하여 복수의 비트라인에 변환된 복수의 아날로그 전압을 각각 프리차지할 수 있다.

[0148] 단계(602)에서 CIM 장치의 동작 방법은 커패시터 비율에 기초하여 웨이트와 프리차지 전압을 결합하여 결합 결과를 출력한다.

[0149] 즉, CIM 장치의 동작 방법은 워드 라인에 입력에 따라 복수의 메모리셀에 각각 저장된 웨이트(weight)와 각각 프리차지된 아날로그 전압을 결합하여 복수의 비트라인을 통해 결합 결과를 각각 출력할 수 있다. 여기서, CIM 장치는 복수의 메모리셀에 각각 저장된 웨이트(weight)를 비트라인에 연결된 커패시터에 설정된 서로 다른 비율에 기반하여 최상위 비트와 최하위 비트로 구분할 수 있다.

[0150] 단계(603)에서 CIM 장치의 동작 방법은 커패시터 비율에 기초하여 결합 결과를 가산한다.

[0151] 즉, CIM 장치의 동작 방법은 복수의 비트라인에 각각 연결된 스위치를 제어하여 커패시터의 서로 다른 비율의 합이 반영된 결합 결과를 가산할 수 있다.

[0152] 여기서, CIM 장치의 동작 방법은 복수의 비트라인에 각각 연결된 스위치를 연결하여 복수의 비트라인의 전하를 공유하고, 공유된 전하에는 커패시터의 비율이 반영되어 있으며, 이에 따라 서로 다른 비율의 합이 반영된 결합 결과를 가산할 때, 입력된 전압과 관련된 전하량을 보존할 수 있다.

[0154] 도 7은 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 동작 방법과 관련된 타이밍도를 설명하는 도면이다.

[0155] 도 7을 참고하면, CIM 장치의 동작 방법은 타이밍도(700)에서 단계(700), 단계(701) 및 단계(702)로 구분될 수 있다.

[0156] 단계(700)는 디지털 전압을 아날로그 전압으로 변환하는 아날로그 변환부의 동작과 관련될 수 있다.

[0157] 예를 들어, 아날로그 변환부에 1.2V의 구동전압이 1n 내지 10n 동안 인가될 수 있다.

[0158] 단계(701)는 프리차지된 아날로그 전압과 메모리셀에 저장된 웨이트를 결합(multiply)하는 메모리 어레이부의 동작과 관련될 수 있다.

[0159] 또한, 단계(701)는 워드 라인에 구동 전압이 인가되는 시간과 관련될 수 있다.

[0160] 예를 들어, 메모리 어레이부에 워드 라인을 통해 1.2V의 구동전압이 15n 내지 60n 동안 인가될 수 있다.

[0161] 단계(702)는 결합 결과를 가산(add)하는 가산부의 동작과 관련될 수 있으며, 스위치 제어 동작과 관련될 수 있다.

[0162] 예를 들어, 스위치를 연결하는 전압으로 1.2V의 전압이 인가될 수 있다.

[0164] 도 8은 본 발명의 일실시예에 따른 CIM(Computation In Memory) 장치의 동작 결과를 설명하는 도면이다.

[0165] 도 8은 CIM 장치의 동작 결과를 예시하는 그래프(800)를 도시한다.

[0166] 도 8을 참고하면, 그래프(800)의 가로축은 멀티비트의 웨이트를 나타낼 수 있고, 세로축은 멀티비트 웨이트와 멀티비트 입력의 결합 결과를 나타낼 수 있으며, 그래프에 도신된 전압은 도 7의 타이밍도(700)에서 90n에 측정된 전압을 나타낼 수 있다.

[0167] 그래프(800)에 따르면, 본 발명은 MAC 오퍼레이션의 분류 정확도를 개선할 수 있다.

[0168] 또한, 본 발명은 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행함으로써 연산을 위한 에너지 소모를 줄이면서 머신러닝에서 요구되는 많은 양의 데이터를 처리할 수 있다.

[0169] 또한, 본 발명은 멀티비트 입력과 멀티비트 웨이트의 결합(multiply) 연산을 수행함으로써 멀티비트 입력에 해당하는 많은 양의 데이터를 처리하면서도 병목(bottleneck)현상의 발생을 억제할 수 있다.

[0171] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0172] 이상과 같이 실시예들이 비록 한정된 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.

[0173] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

부호의 설명

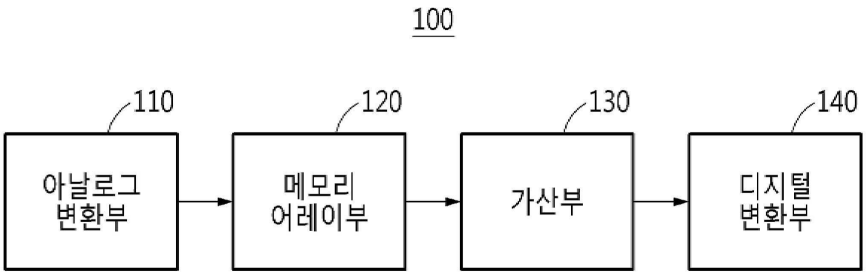
[0174] 100: CIM(Computation In Memory) 장치 110: 아날로그 변환부

120: 메모리 어레이부 130: 가산부

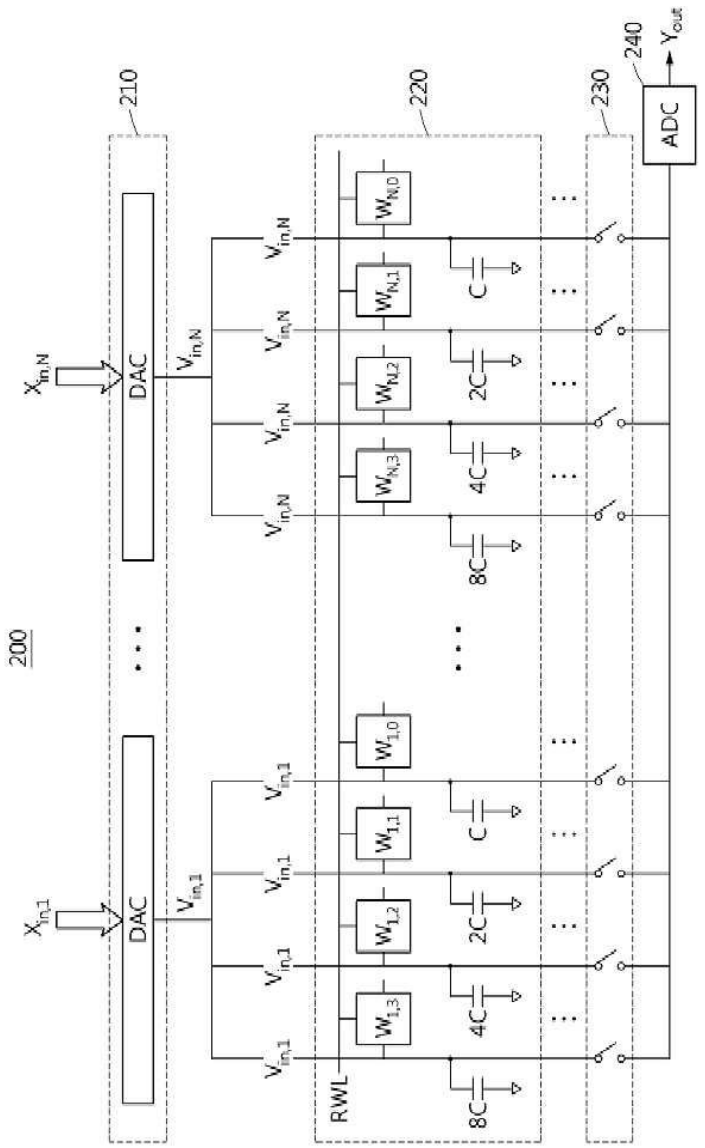
140: 디지털 변환부

도면

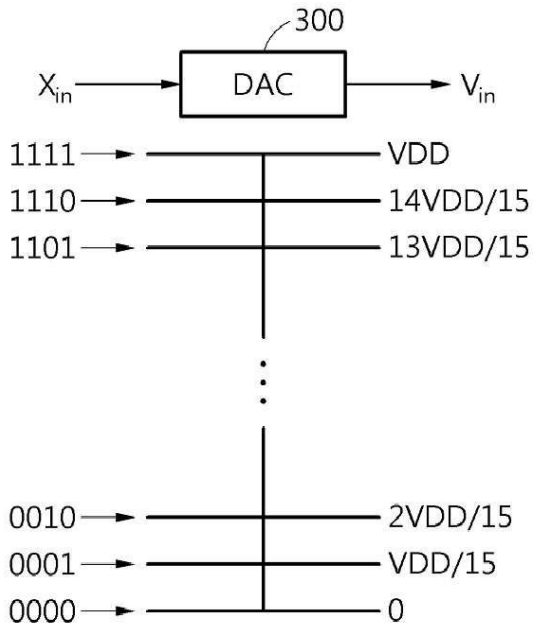
도면1



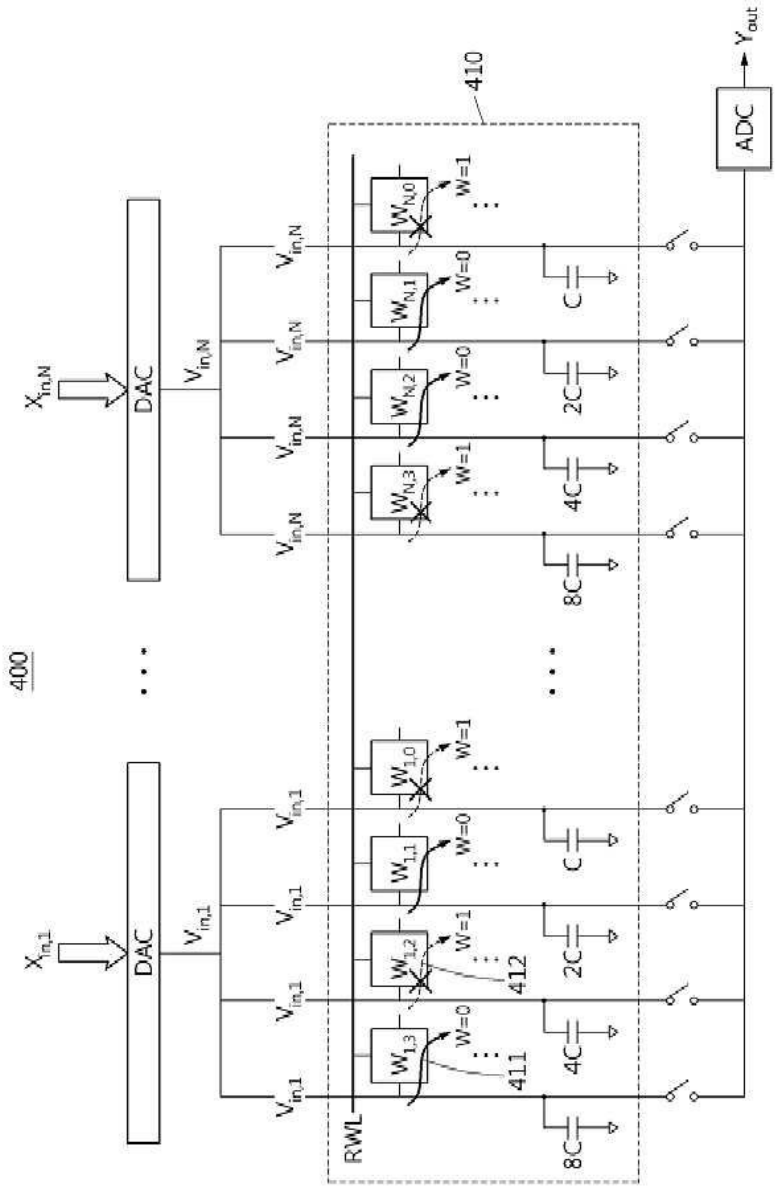
도면2



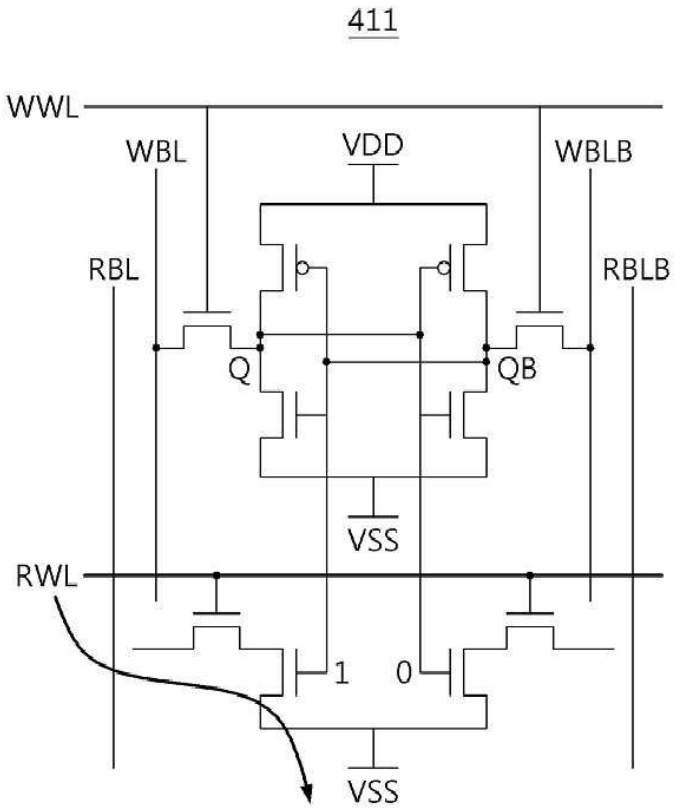
도면3


300
DAC
X_{in}
V_{in}
1111
1110
1101
VDD
14VDD/15
13VDD/15
0010
0001
0000
2VDD/15
VDD/15
0

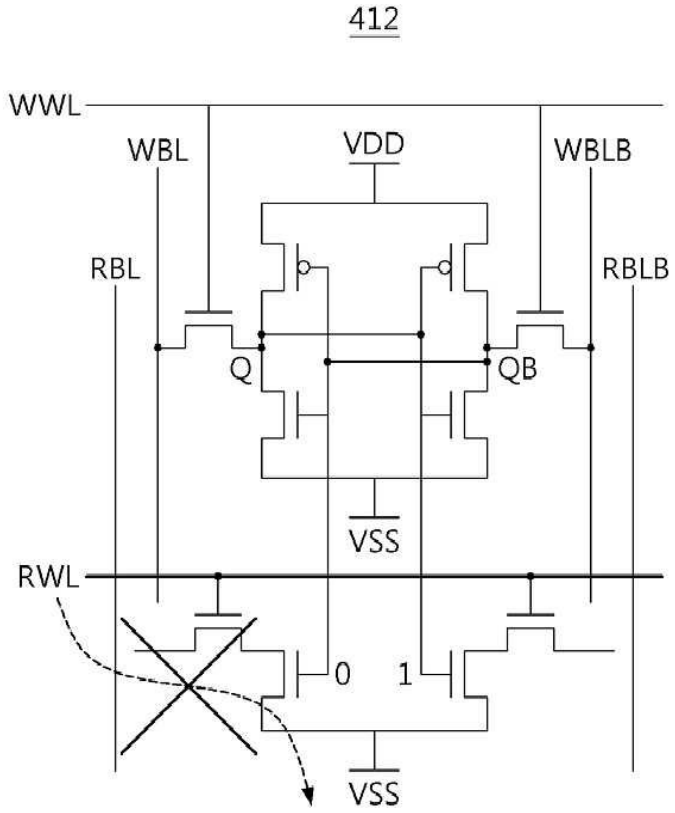
도면4a



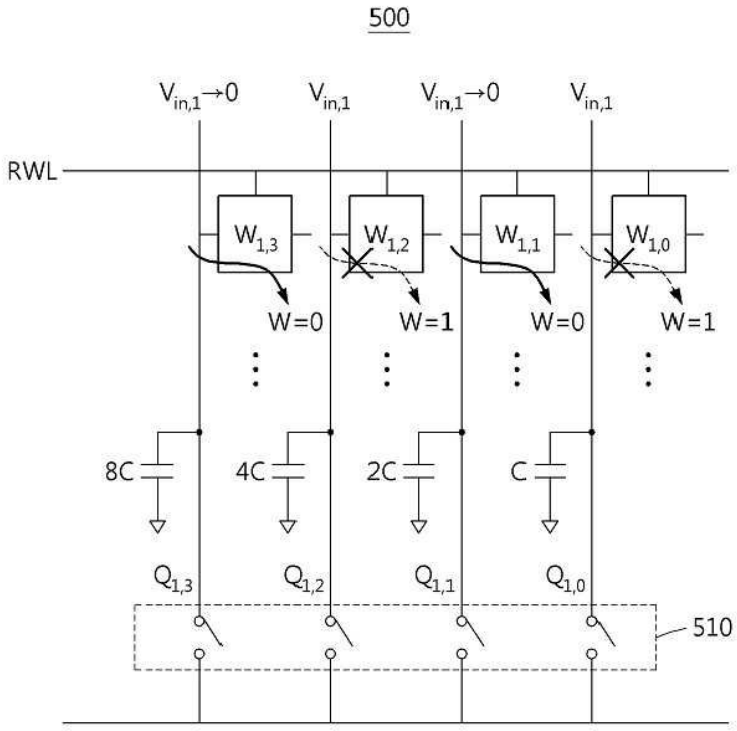
도면4b


411
WWL
WBL
VDD
WBLB
RBL
RBLB
Q
QB
VSS
RWL
1
0
VSS

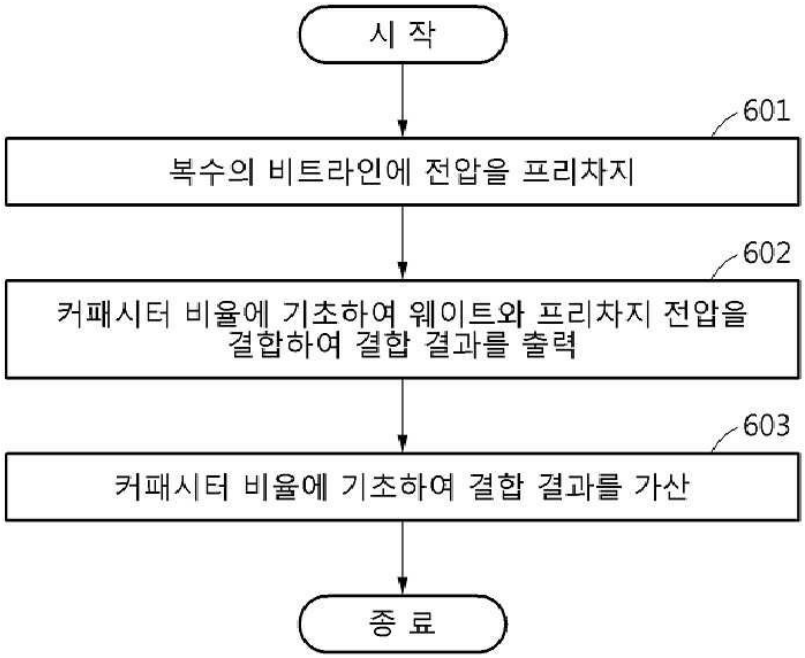
도면4c


412
WWL
WBL
VDD
WBLB
RBL
RBLB
Q
QB
VSS
RWL
0
1
VSS

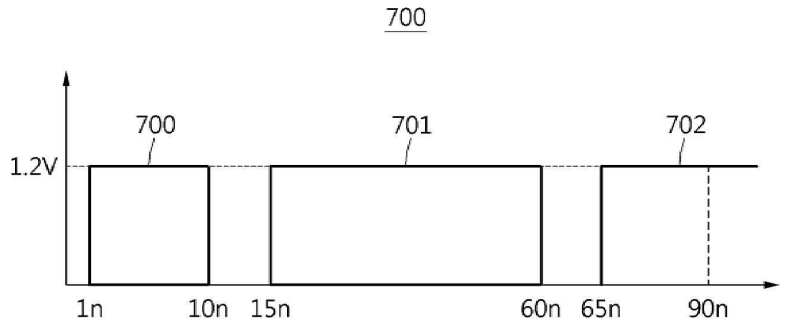
도면5


500
$V_{in,1}\rightarrow 0$
$V_{in,1}$
$V_{in,1}\rightarrow 0$
$V_{in,1}$
RWL
$W_{1,3}$
$W_{1,2}$
$W_{1,1}$
$W_{1,0}$
W=0
W=1
W=0
W=1
8C
4C
2C
C
$Q_{1,3}$
$Q_{1,2}$
$Q_{1,1}$
$Q_{1,0}$
510

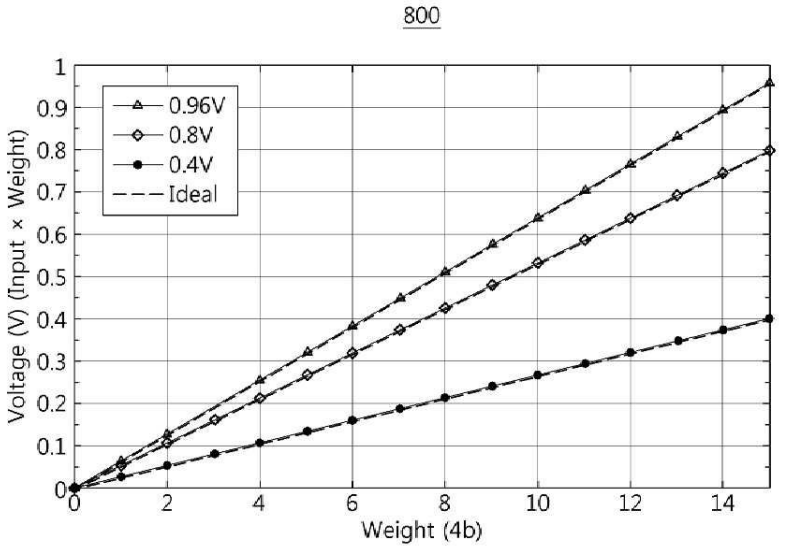
도면6


시 작
601
복수의 비트라인에 전압을 프리차지
602
커패시터 비율에 기초하여 웨이트와 프리차지 전압을 결합하여 결합 결과를 출력
603
커패시터 비율에 기초하여 결합 결과를 가산
종 료

도면7



도면8



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 1

【변경전】

멀티비트의 디지털 전압을 복수의 아날로그 전압으로 변환하여 복수의 비트라인에 상기 변환된 복수의 아날로그 전압을 각각 프리차지하는 아날로그 변환부;

상기 복수의 비트라인에 각각 연결되고, 웨이트(weight)가 각각 저장된 복수의 메모리셀과 서로 다른 비율로 정전용량(capacitance)이 미리 결정된 복수의 커패시터를 포함하고, 상기 복수의 커패시터 각각은 상기 복수의 비트라인에 각각 연결되며, 워드 라인을 통해 구동 전압이 인가될 시 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 메모리 어레이부; 및

상기 복수의 비트라인에 각각 연결된 스위치를 제어하여 서로 다른 비율의 합으로 나눈 결합 결과를 가산하는 가산부를 포함하고,

상기 가산부는 상기 복수의 비트라인에 각각 연결된 스위치를 연결하여 상기 서로 다른 비율의 합으로 상기 가산된 결합 결를 나눈 값에 서로 다른 비율의 정전용량 합을 곱하여 상기 각각 프리차지된 아날로그 전압의 전하량을 보존하는

CIM(Computation In Memory) 장치.

【변경후】

멀티비트의 디지털 전압을 복수의 아날로그 전압으로 변환하여 복수의 비트라인에 상기 변환된 복수의 아날로그 전압을 각각 프리차지하는 아날로그 변환부;

상기 복수의 비트라인에 각각 연결되고, 웨이트(weight)가 각각 저장된 복수의 메모리셀과 서로 다른 비율로 정전용량(capacitance)이 미리 결정된 복수의 커패시터를 포함하고, 상기 복수의 커패시터 각각은 상기 복수의 비트라인에 각각 연결되며, 워드 라인을 통해 구동 전압이 인가될 시 상기 각각 저장된 웨이트(weight)와 상기 각각 프리차지된 아날로그 전압을 결합하여 상기 복수의 비트라인을 통해 결합 결과를 각각 출력하는 메모리 어레이부; 및

상기 복수의 비트라인에 각각 연결된 스위치를 제어하여 서로 다른 비율의 합으로 나눈 결합 결과를 가산하는 가산부를 포함하고,

상기 가산부는 상기 복수의 비트라인에 각각 연결된 스위치를 연결하여 상기 서로 다른 비율의 합으로 상기 가산된 결합 결과를 나눈 값에 서로 다른 비율의 정전용량 합을 곱하여 상기 각각 프리차지된 아날로그 전압의 전하량을 보존하는

CIM(Computation In Memory) 장치.