ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 24-02-16
112
j.inst.Korean.electr.electron.eng.Vol.28,No.2,234~238,June 2024
https://doi.org/10.7471/ikeee.2024.28.2.234

3차원 적층 구조 저항변화 메모리 어레이를 활용한 CNN 가속기 아키텍처

CNN Accelerator Architecture using 3D-stacked RRAM Array

이 원 주*, 김 윤*, 구 민 석**★

Won Joo Lee*, Yoon Kim*, Minsuk Koo**★

Abstract

This paper presents a study on the integration of 3D-stacked dual-tip RRAM with a CNN accelerator architecture, leveraging its low drive current characteristics and scalability in a 3D stacked configuration. The dual-tip structure is utilized in a parallel connection format in a synaptic array to implement multi-level capabilities. It is configured within a Network-on-chip style accelerator along with various hardware blocks such as DAC, ADC, buffers, registers, and shift & add circuits, and simulations were performed for the CNN accelerator. The quantization of synaptic weights and activation functions was assumed to be 16-bit. Simulation results of CNN operations through a parallel pipeline for this accelerator architecture achieved an operational efficiency of approximately 370 GOPs/W, with accuracy degradation due to quantization kept within 3%.

요 약

본 논문은 낮은 구동 전류 특성과 3차원 적층 구조로 확장시킬 수 있는 장점을 가진 3차원 적층형 이중 팁 RRAM을 CNN 가속기 아키텍처에 접목하는 연구를 수행한 논문이다. 3차원 적층형 이중 팁을 적층 형태의 병렬연결로 시냅스 어레이에 사용하여 멀티-레벨을 구현하였다. 이를 Network-on-chip 형태의 가속기 내에 DAC, ADC, 버퍼 및 레지스터, shift & add 회로 등 다양한 하드웨어 블록들과 함께 구성하여 CNN 가속기에 대한 시뮬레이션을 수행하였다. 시냅스 가중치와 활성화 함수의 양자화는 16-bit으로 가정하였다. 해당 가속기 아키텍처를 위한 병렬 파이프라인을 통해 CNN 연산을 시뮬레이션한 결과, 연산효율은 약 370 GOPs/W를 달성하였으며, 양자화에 의한 정확도 열화는 3 % 이내가 되는 결과를 나타냈다.

Key words : RRAM, neuromorphic computing, 3D-stacked, artificial neural network, artificial intelligence

Ⅰ. 서론

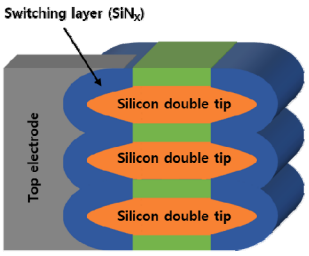
저항성 랜덤 액세스 메모리(RRAM)는 신경 모방 시스템을 위한 시냅스 장치로 주목받고 있다. 그러나 RRAM은 일반적으로 높은 형성 전압과 스위칭 동작의 큰 변동성과 같은 여러 문제를 겪는다. 이를 해결하기 위해 다양한 연구가 진행되고 있으나 소자의 단편적 특성의 향상을 보고하는데 그치고 있다[1]. 제안한 소자를 활용한 아

* (Graduate student, Professor) Dept. of Electronical and Computer Engineering, University of Seoul
** (Professor) Dept. Computer Science and Engineering, Incheon National University
★ Corresponding author
Email : koo@inu.ac.kr, Tel : +82-32-835-8499
※ Acknowledgment
This work was supported by Incheon National University Research Grant in 2024.
Manuscript received Jun. 25, 2024; revised Jun. 27, 2024; accepted Jun. 28, 2024.

키텍처 및 시스템 시뮬레이션을 통해서만 대형 인공 신경망을 구현하였을 때의 효율성 및 실효성을 검증할 수 있으며 여기에는 실제 제작된 소자의 특성을 반영되어야 한다. 본 연구에서는 본 연구팀이 제안한 저전압으로 동작하며 기존의 3차원 어레이에 비해 높은 집적도를 달성할 수 있는 적층 이중 팁 실리콘 나노와이어를 기반으로 한 3차원 적층 RRAM을 활용하는 합성인공신경망(CNN) 가속기 아키텍처를 제안한다. 실제 제작된 소자의 성능을 반영한 시스템 시뮬레이션을 통해 제안한 가속기 구조의 ImageNet 분류 정확도 및 전력효율을 도출해봄으로써 해당 소자의 효율성 및 실효성을 검증하였다.

II. 본론

1. CNN 가속기 아키텍처



State	Conductance	
	Average	Standard deviation
HRS	0.45 nS	51.6 %
LRS	7.02 nS	128 %

Fig. 1. 3D-stacked double-tip Si nanowire RRAM & its READ characteristics.
그림 1. 3차원 더블팁 RRAM 구조와 읽기동작 특성

가. 3차원 적층형 RRAM과 크로스바 어레이 구조

본 연구에 사용된 시냅스 소자는 기존의 기둥형 수직 RRAM의 일관성 없는 전도성 필라멘트의 형성의 단점을 극복하면서 2배 이상의 집적도를 달성할 수 있는 본 연구팀에서 개발한 더블팁 구조의 RRAM을 사용하였다[2]. 이러한 더블팁 구조의 RRAM의 장점은 동작 전류의 편차를 적게 만들어 크로스바 배열을 통해 Vector - Matrix - Multiplication(VMM) 연산을 수행하여 인공지능 분류 작업을 수행할 시 더블 팁 없는 RRAM에 비해 우수한 성능을 보인다고 보고되었다. 또한 기존 RRAM의 MIM(Metal - Insulator - Metal) 구조가 아닌 MIS (Metal - Insulator - Silicon) 구조로 되어 있기에 기존 RRAM에 비해 낮은 동작 전류를 가진다. 전극의 넓이가 적층된 실리콘 두께로 결정되기 때문에, 리소그래피를 통해 패터닝된 전극에 비해 전극 넓이를 많이 줄일 수 있다. 이러한 구조적 특징들로 인해 구동 전류가 작아져 전력 효율을 증가시킬 수 있다[2].

그림 1. 해당 더블팁 RRAM의 구조와 제작된 RRAM의 상태별 읽기 특성을 나타낸다. 각 층의 실리콘 더블팁 전극은 동일한 상부 전극 한 개와 연결되어 있으므로, 3개의 적층 소자들은 병렬로 연결되어 있음을 알 수 있다. 이후 CNN 가속기 아키텍처에서 시냅스 소자의 특성은 이 값들이 사용되었다.

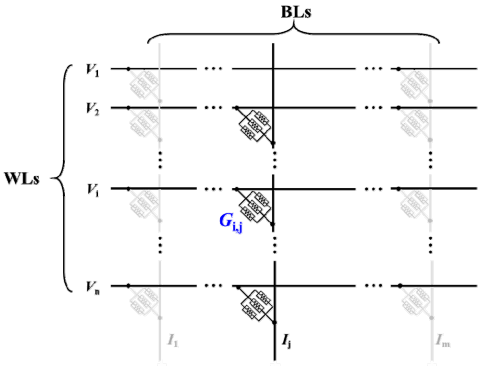


Fig. 2. 3D-stacked RRAM crossbar array.
그림 2. 3차원 적층형 RRAM 크로스바 어레이

그림 2는 3차원 적층형 더블 팁 저항성 변화 메모리를 크로스바 배열 형태로 구성한 Multiply-and-Accumulate (MAC) 연산 구조를 나타낸다. 각 BL(Bit Line)은 저항성 변화 메모리를 통해 모든 WL(Word Line)과 연결된 구조를 갖추고 있고 i와 j는 crossbar 배열 내의 인덱스를 나타낸다. 시냅스 배열 내 시냅스 소자의 전도도는 시냅스 가중치를 나타내며, 이는 Ohm의 법칙과 Kirchhoff의 전류 법칙을 기반으로 VMM 연산을 수행한다. WL의 활성화 전압(V1~Vn)과 시냅스 소자의 전도도의 곱은 Ohm의 법칙에 의해 BL로 들어가는 전류로 표현된다. 모든 WL에 전압이 동시에 인가되어 BL에 연결된 모든 시냅스 소자에 전류가 흐르게 되면, Kirchhoff의 전류 법칙에 의해 전류들이 합쳐져 VMM 연산이 수행되며 이는 다음과 같은 식으로 표현할 수 있다.

$$I_j = \sum_{i=1}^{n} V_i \times G_{i,j} \quad (1)$$

크로스바 배열은 3층으로 적층된 더블 팁 저항성 변화 메모리를 하나의 시냅스 소자로 사용하는데, 이는 2-bit 동작을 위해서이다. 적층된 소자들은 서로 병렬로 묶여 있으며, 개별 소자는 HRS/LRS의 두 상태만 활용하여 4개의 전도도 레벨을 표현할 수 있다. 예를 들어, 2-bit로 00은 3개의 병렬 소자가 모두 HRS인 경우로 나타낼 수 있고, 01은 1개의 소자만 LRS, 10은 2개의 소자가 LRS, 마지막으로 11은 3개의 소자가 모두 LRS인 경우로 볼 수 있다.

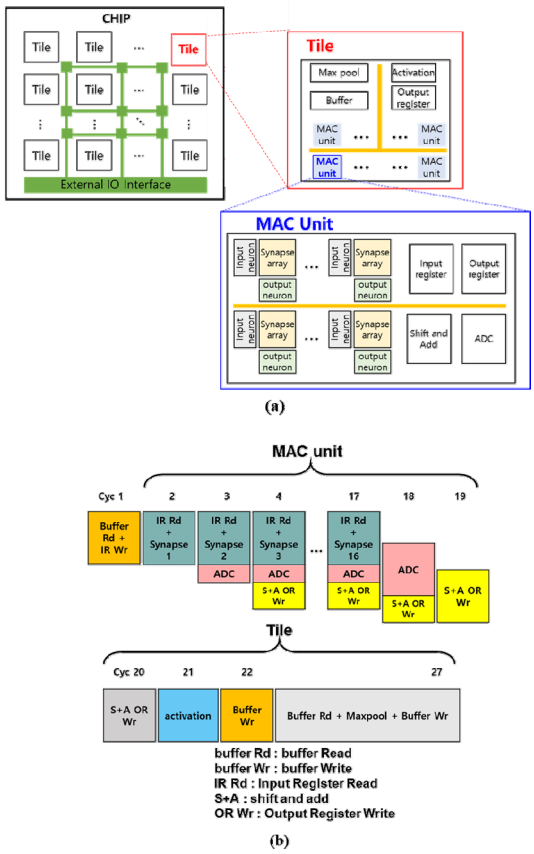


Fig. 3. (a) The proposed NoC architecture using 3D-stacked double-tip RRAM and (b) pipeline for parallel computation.

그림 3. (a) 제안된 3차원 더블팁 RRAM을 활용한 NoC기반 아키텍처와 (b) 병렬연산을 위한 파이프라인

나. 더블팁 RRAM을 활용하는 NoC 아키텍처

그림 2. (a)는 제안된 3차원 더블 팁 RRAM을 활용한 Network-on-Chip(NoC) 구조의 아키텍처를 나타낸다. 하나의 칩 안에 168개의 타일이 엔진 역할을 하며 각 타일은 12개의 Multiply-and-Accumulate(MAC) 유닛과 풀링 및 활성함수 기능을 수행하는 유닛과 버퍼로 구성되어 있다. 이러한 구성은 제한된 버퍼와 시냅스 배열 크기에 따라 최적화된 ISSAC[3] 논문을 참조하였다. MAC 유닛 내부에는 128×128 더블 팁 RRAM 배열 8개, Shift-and-Add 블록, Analog-to-Digital Converter (ADC), 그리고 버퍼가 포함되어있다.

그림 2. (b)는 그림 2. (a) 아키텍처의 효율적인 병렬연산을 위해 파이프라인이다. 각 MAC 유닛 및 타일에서 파이프라인을 통해 각 사이클별로 가속기 내부 하드웨어 블록들이 병렬적으로 동작한다. 각 사이클의 시간은 아키텍처 내부의 모든 하드웨어 블록 중 가장 시간이 오래 걸리는 ADC를 기준으로 정해지는데, 이 시간은 1.2 GHz 주파수 클럭을 사용할 시에 1.28 Gsps(Giga sample per second)로 128 개 BL의 결과를 입력으로 받아 디지털 신호로 변환하는데 100 ns가 소모된다[4].

파이프라인 동작에 대한 자세한 설명은 다음과 같다. 첫 번째 사이클에서는 타일 내부의 버퍼로부터 입력데이터를 읽어와 MAC 유닛의 입력 레지스터에 쓰는 동작을 수행한다. 사이클 2~17은 입력 레지스터에 쓰여진 입력 데이터를 읽어와 1-bit DAC를 통해 시냅스 배열을 활용하여 MAC 연산을 수행하는 과정이다. 16 사이클을 소모하는 이유는 1-bit DAC를 활용하여 16-bit를 표현하기 위해 MAC 연산을 수행하고 shift and add 과정을 통해 자릿수를 맞추기 때문이다. 사이클 2에서는 첫 번째 자릿수의 MAC 연산을 수행하고 그 결과는 사이클 3에서 ADC로 들어간다. 이때 새로운 자릿수의 입력이 시냅스 배열로 입력되어 MAC 연산이 수행되며 이 결과는 첫 번째 자릿수의 ADC 연산이 완료되기 전까지 출력 뉴런의 sample and hold 블록에 임시적으로 저장된다. 사이클 4에서는 첫 번째 자릿수 연산 결과가 ADC를 거쳐 shift and add 회로를 통해 자릿수를 맞추고 MAC 유닛 내부의 출력 레지스터에 쓰여진다. 이때 앞선 과정과 동일하게 두 번째 자릿수 연산 결과는 ADC를 거치고 세 번째 자릿수 입력데이터가 시냅스 배열에 입력되어 MAC 유닛 및 타일들이 병렬적으로 연산을 수행한다. 이러한 과정이 사이클 19까지 이어지며 시냅스 배열을 활용한 모든 자릿수의 16-bit MAC 연산이 끝나면 타일 내의 출력 레지스터로 새로운 데이터가 쓰여지고, 다른 MAC 유닛의 결과와 섞이지 않게 타일 내의 shift and add 회로를 통해 타일 내 출력 레지스터에 쓰여진다. 사이클 21에서 타일 내 모든 MAC 유닛 결과들은 한 번에 활성화 함수를 적용하는 회로를 거치며, 사이클 22에서는 이 결과들을 타일 내 버퍼에 쓰는 동작을 수행한다. 이후 사이클 23 ~ 27은 버퍼에 저장된 결과들을 읽어와 maxpool 연산을 수행한다.

2. CNN 가속기 시스템 시뮬레이션 결과

3차원 적층형 더블팁 RRAM을 활용한 제안된 CNN 가속기 시스템의 성능을 평가하기 위해 RRAM 소자의 특성과 32nm CMOS 공정으로 제작된 주변회로들의 동작속도와 전력소모를 고려하여 시스템 시뮬레이션을 수행하였으며, 사용된 회로들의 Spec은 표 1에 나타나 있다.

그 결과, 그림 4에서 확인 할 수 있듯이 전체 66.23 W의 전력을 소모하면서 370 GOPs/W(Giga operation per second per Watt)의 연산효율을 보였다. 이는 기

Table 1. Parameters of CMOS circuits for system simulation.

표 1. 시스템 시뮬레이션을 위한 CMOS 회로 파라미터

Off-chip links			
HyperTransport	links/freq link bw	4/1.6 GHz 6.4 GB/s	10.4 W
Tile at 1.2 GHz			
Component	**Params**	**Spec**	**Power**
Buffer	Size Num_banks Bus_width	64KB 4 256b	20.7 mW
Router	Flit size Num_port	32 8	42mW
Activation S+A Maxpool Output Register	Number Number Number Size	2 1 1 3KB	0.52 mW 0.05 mW 0.4 mW 1.68 mW
MAC unit property (12 MAC unit per tile)			
ADC	Resolution Frequency Number	8 bits 1.2 GSps 8	16 mW
DAC	Resolution Number	1 bit 8 X 128	4 mW
S+H	Number	8 X 128	10 uW
synapse array	Number Size	8 128 X 128	0.6 mW
S+A Input Register Output Register	Number Size Size	4 2 KB 256 B	0.2 mW 1.24 mW 0.23 mW

존 DaDianNao 아키텍처의 연산효율(286.4 GOPs/W)보다 우수한 연산효율이며. DaDianNao 아키텍처는 NVIDIA K20M GPU에 비해 450배 빠른 스피드 및 150배 낮은 에너지 소모를 달성했다고 보고된 바 있다[5].

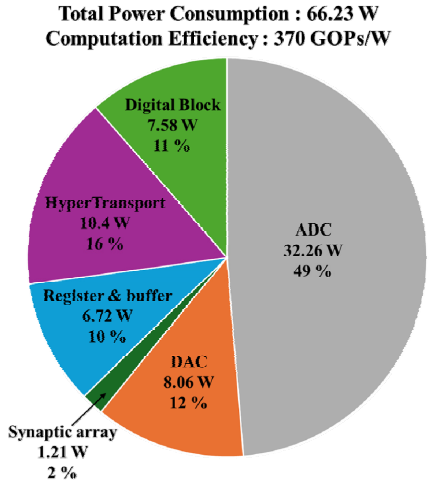


Fig. 4. System simulation Results with power breakdown.

그림 4. 시스템 시뮬레이션 결과와 블록별 전력소모

그 결과, 그림 4에서 확인 할 수 있듯이 전체 66.23 W의 전력을 소모하면서 370 GOPs/W(Giga operation per second per Watt)의 연산효율을 보였다. 이는 기존 DaDianNao 아키텍처의 연산효율(286.4 GOPs/W)보다 우수한 연산효율이며. DaDianNao 아키텍처는 NVIDIA K20M GPU에 비해 450배 빠른 스피드 및 150배 낮은 에너지 소모를 달성했다고 보고된 바 있다[5].

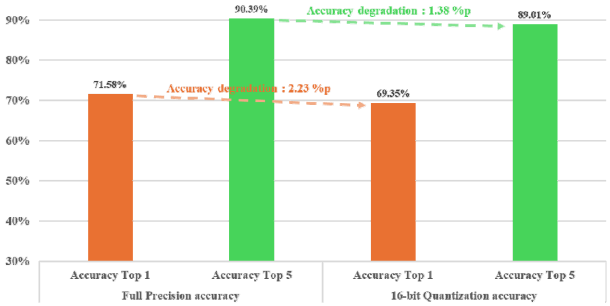


Fig. 5. Classification accuracy for ImageNet dataset using VGG-16 with and without 16-bit quantization.

그림 5. 16-bit 양자화를 적용하였을 때와 하지 않았을 때의 VGG-16에서의 ImageNet 분류 정확도

그림 5는 VGG-16을 사용하여 ImageNet dataset에 대한 정확도를 살펴보았을 때 full-precision에 비해서 16-bit 양자화를 적용하였을 때 정확도의 열화에 대한 시뮬레이션 결과로 본 아키텍처에서 적용한 16-bit 양자화가 3% 이내의 성능 열화만을 유발한다는 것을 확인할 수 있었다.

Ⅲ. 결론

본 논문에서는 3차원 적층형 저항변화 메모리를 활용하여 CNN 가속기 아키텍처를 시뮬레이션하여 그 연산효율과 정확도에 대해 분석하였다. 시뮬레이션 결과 연산효율은 370 GOPs/W로 비슷한 면적의 DaDianNao 아키텍처에 비해 30% 우수한 연산효율을 낼 수 있었다. 또한 하드웨어 특성을 반영해 16-bit으로 가중치 및 활성화 함수 결과를 양자화 하였음에도 정확도 측면에서 3% 이내의 낮은 정확도 감소만 나타남을 보였다. 해당 연구를 통해 새로운 메모리 소자의 개발이 단순히 소자 특성을 증명하는 데에서 그치지 않고 인공지능 연산을 위한 컴퓨팅 시스템에 어떠한 영향을 미칠 수 있는지를 증명할 수 있었다.

References

[1] Rehman, Muhammad Muqeet, et al. “Decade of 2D-materials-based RRAM devices: a review,” *Science and technology of advanced materials,*

vol.21, no.1, pp.147-186, 2020.
DOI: 10.1080/14686996.2020.1730236
[2] Lee, Won Joo, et al. "Three-Dimensional Resistive Random-Access Memory Based on Stacked Double-Tip Silicon Nanowires for Neuromorphic Systems," *ACS Applied Electronic Materials,* vol.6, no.4, pp. 2232-2241, 2024. DOI:10.1021/acsaelm.3c01680
[3] Shafiee, Ali, et al. "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *ACM SIGARCH Computer Architecture News,* vol.44, no.3, pp.14-26, 2016. DOI: 10.1145/3007787.3001139
[4] L. Kull, et al. "A 3.1 mW 8b 1.2 GS/s Single-Channel Asynchronous SAR ADC with Alternate Comparators for Enhanced Speed in 32 nm Digital SOI CMOS," *Journal of Solid-State Circuits,* vol. 44, no.12, pp.3049-3058, 2013.
DOI: 10.1109/JSSC.2013.2279571
[5] Y. Chen, et al., "DaDianNao: A Machine-Learning Supercomputer," *in Proceedings of 47th Annual IEEE/ACM International Symposium on Microarchitecture,* 2014, pp.609-622.
DOI: 10.1109/MICRO.2014.58

BIOGRAPHY

Won Joo Lee (Member)

2016 : BS degree in Nanomaterial Engineering, Pusan National University.
2018 : MS degree in Nanomaterial Engineering, Pusan National University.
2022~ : PhD degree in Electrical and Computer Engineering, University of Seoul.

Yoon Kim (Member)

2006 : BS degree in Electrical Engineering, Seoul National University.
2012 : PhD degree in Electrical Engineering, Seoul National University.
2012~2015 : Senior Engineer, Samsung Electronics.
2015~2018 : Professor, Pusan National University.
2018~ : Professor, University of Seoul

Minsuk Koo (Member)

2007 : BS degree in Electrical Engineering, KAIST.
2009 : MS degree in Electrical Engineering, Seoul National University.
2009~2012 : Senior Engineer, Radiopulse Inc.
2020 : PhD degree in Electrical and Computer Engineering, Purdue University.
2020~ : Professor, Incheon National University.