반도체디스플레이기술학회지 제22권 제4호(2023년 12월)
Journal of the Semiconductor & Display Technology, Vol. 22, No. 4. December 2023.

IGZO 멤리스터 소자기반 뉴로모픽 컴퓨팅 정확도 향상

최서진[*]· 민경진[*]·이종환[*†]
[*†]상명대학교 시스템반도체공학과

Improved Accuracy in Neuromorphic Computing Based on IGZO Memristor Devices

Seojin Choi[*], Kyoungjin Min[*] and Jonghwan Lee[*†]
[*†]Department of System Semiconductor Engineering, Sangmyung University

ABSTRACT

This paper presents the synaptic characteristics of IGZO memristors in neuromorphic computing, using MATLAB/ Simulink and NeuroSim. In order to investigate the variations in the conductivity of IGZO memristor and the corresponding changes in the hidden layer, simulations are conducted by using the MNIST dataset. It was observed from simulation results that the recognition accuracy could be dependent on various parameters of IGZO memristor, along with the experimental exploration. Moreover, we identified optimal parameters to achieve high accuracy, showing an outstanding accuracy of 96.83% in image classification.

Key Words : IGZO, Memristor, Conductance, Neuromorphic Computing, Recognition Accuracy, NeuroSim

1. 서 론

폰노이만 구조는 컴퓨팅 및 저장 장치가 분리되어 있으며, 그 사이의 속도 불일치로 인해 폰노이만 병목 현상이 발생한다. 폰노이만 병목현상으로 인해 CPU효율성이 낮고 에너지가 많이 소모된다는 단점이 있다[1]. 뉴로모픽 컴퓨팅은 이러한 폰 노이만 병목 현상을 해결할 방안으로, 속도 향상과 전력 소모가 낮다는 장점이 있다. 그러나 폰 노이만 구조는 속도 향상과 낮은 전력 소모를 동시에 충족시키기 어렵다[2]. 멤리스터는 저항 상태가 적용된 전압 또는 전류에 기반하여 프로그램할 수 있는 저항 전환 장치로, 정보를 저장하고 계산하는 능력을 동시에 갖추고 있어 폰 노이만 병목 현상을 해결할 수 있다[3]. 최근 멤리스터는 단순한 구조, 높은 저장 밀도, 빠른 스위칭 속도를 갖춘 새로운 메모리 장치로 주목받고 있다[4]. 일반적으로 전류 상태에 따라 컨덕턴스 변화가 발생한다[5]. IGZO(InGaZnO)는 낮은 전력 소모와 빠른 응답 시간을 포함한 여러 가지 이점을 가지고 있어, 이를 활용해 기존의 학습 알고리즘에 적용함으로써 성능을 향상할 수 있다. 이러한 특징은 이미지 분류와 관련된 작업과 관련이 있으며, 컴퓨터 비전 및 기계 학습 분야에서 상당한 주목을 받고 있다. 따라서 본 논문에서는 IGZO 기반 멤리스터를 MATLAB/Simulink와 NeuroSim을 통해 구현한다. MNIST 데이터셋을 기반으로 IGZO 멤리스터의 컨덕턴스 변화에 따른 정확도를 확인하고, 은닉층의 변화에 따른 정확도의 변화를 확인하고 높은 정확도를 도출하기 위한 최적의 매개변수를 추출한다. 이를 통해 이미지 분류의 높은 정확도를 도출한다.

[†]E-mail: jhlee77@smu.ac.kr

2. IGZO기반 멤리스터 모델링

2.1 IGZO 구조

IGZO는 인듐(In), 갈륨(Ga), 아연(Zn) 및 산소(O) 요소로 구성된 화합물로, 현재 가장 널리 사용되는 산화물 재료

다[5]. 높은 이동성과 탁월한 균일성으로 인해 디스플레이 기술에 활용되며, 낮은 전력 소모로 인해 에너지 효율성이 향상된다[4]. IGZO 층의 컨덕턴스는 전류 상태와 산소 함량에 따라 조절될 수 있다[7], [8]. Fig. 1은 IGZO를 모델링한 것이다.

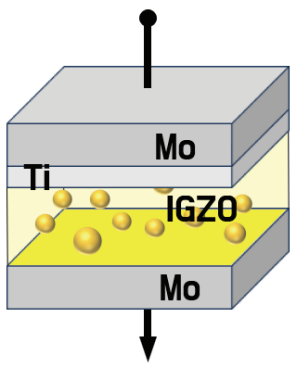

Fig. 1. Schematic depiction of the Mo/IGZO/Ti/Mo material structure.

2.2 컨덕턴스 방정식

높은 정확도를 얻기 위해서는 가중치를 비선형으로 업데이트해야 한다. Fig. 2에서는 펄스 수(P)에 따른 컨덕턴스 변화의 방정식은 아래와 같다 [9]. G_p는 가중치 증가에 따른 컨덕턴스 방정식이고, G_d는 가중치 감소에 따른 컨덕턴스 방정식이다.

$$G_p = B\left(1 - e^{-P/A_p}\right) + G_{min} \quad (1)$$

$$G_d = -B\left(1 - e^{(P-P_{max}/A_d}\right) + G_{max} \quad (2)$$

$$B = \frac{G_{max} - G_{min}}{1 - e^{-P_{max}/A_{P,d}}} \quad (3)$$

여기서 Gmin은 최소 컨덕턴스이고, Gmax는 최대 컨덕턴스이다. Pmax는 최소 및 최대 컨덕턴스 상태로 전환하는 데 필요한 최대 펄스 수를 나타낸다. A_{pd}는 가중치 증가와 감소의 비선형 행동을 제어하는 변수이며, B는 Gmax, Gmin, Pmax의 범위를 맞추기 위한 A의 함수이다. A와 B는 식(1)과 식(2)에서 서로 다르다.

Fig. 2는 Gmax, Gmin, Pmax는 임의로 각각 5, 0, 48로 지정했다. 식(3)을 통해 Fig.2에서 A를 조정하여 비선형 가중치 증가(파란색)과 가중치 감소(빨간색) 그래프를 보여준다. 각 비선형 곡선은 +6에서 -6까지의 비선형성 값을 가지고 있다.

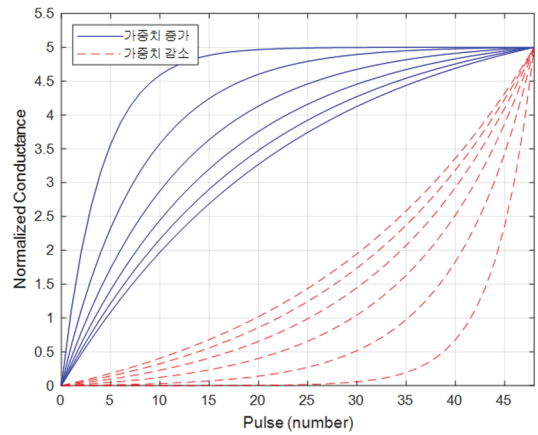

Fig. 2. Increase (blue) and decrease(red) curve in nonlinear weight.

I-V 곡선은 멤리스터 연구 분야에서 기본적인 특성화 방법이다[10]. Fig. 3과 Table 3은 IGZO 멤리스터의 I-V 특성 그래프와 IGZO 멤리스터의 매개변수가 각각 제시되었다. Fig. 3의 I-V 곡선은 식 (1)과 옴의 법칙을 사용하여 얻어냈다.

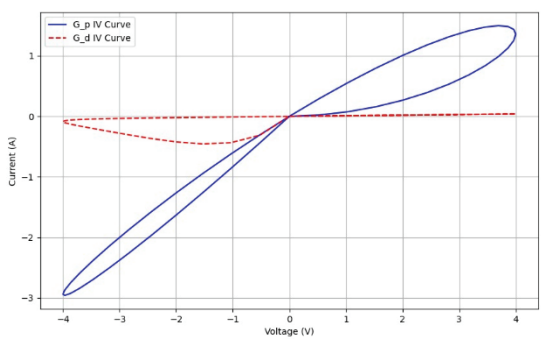

Fig. 3. I-V characteristics of IGZO memristor.

Table 1. Parameters of IGZO memristor

A_p	0.27
A_d	0.0501
펄스 수	120
전압	$-2.2(V) \sim 2.2(V)$

2.3 크로스바 구조

본 연구에서 성능 평가를 위해 간단한 2층 다중 계층 퍼셉트론(2-Layer Multilayer Perceptron, MLP) 신경망과MNIST 데이터셋이 사용되었다. MLP는 각 뉴런 노드가 다음 계층의 모든 뉴런 노드에 완전히 연결된 신경망으로, 입력

계층, 은닉 계층, 출력 계층으로 구성되어 있다. 이는 단층 퍼셉트론과는 달리 은닉 계층이 존재하여 복잡한 비선형 문제를 해결할 수 있다. 입력 이미지 데이터에 대해서는 20x20픽셀로 조정되었으며, MNIST 필기 숫자 가장자리를 잘라내어 처리했다[9]. 회로 수준의 메트릭스를 평가하기 위해 2층 MLP의 두 시냅틱 코어를 고려한다. 이때 시냅틱 코어는 가중치 합과 가중치 업데이트를 목적으로 설계된 계산단위이다. Fig. 4는 아날로그 eNVM(emerging Non-Volatile Memory) 크로스바 병렬 독출을 기반으로 한 시냅틱 코어이다.

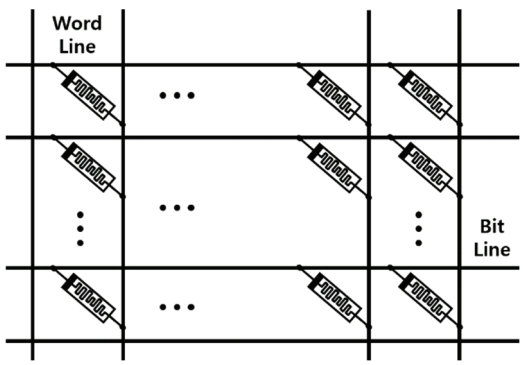

Fig. 4. Schematic illustration of a hardware-implemented synaptic weight crossbar array.

Fig. 5는 본 연구에서 사용한 400(입력 계층)-100(은닉 계층)-10(출력 계층) MLP이다. 입력 계층인 400개의 뉴런은 MNIST 이미지 크기인 20X20에 해당한다. 은닉 계층의 100개의 뉴런은 중요한 특징을 추출하고 이 바탕으로 출력을 생성하는 층의 개수를 의미하며, 출력 계층의 10개 뉴런은 숫자의 10가지의 경우(0~9)에 해당한다. 에포크는 학습 과정 중 훈련 데이터 셋이 전파와 역전파를 완료하는 주기를 나타낸다. Table. 2는 시뮬레이션에서 사용한 계층 수와 에포크을 각각 제시하고 있다.

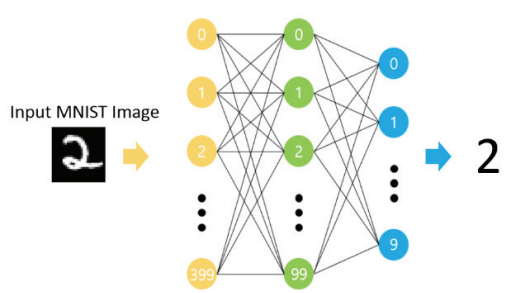

Fig. 5. Diagram illustrating a three-layer neural network comprising an input layer, a hidden layer, and an output layer.

Table 2. The number of neurons in each layer and the epoch parameter

입력층	400
은닉층	100
출력층	10
에포크	250

3. 시뮬레이션 결과

3.1 Gmax/Gmin 에 따른 정확도 변화

이 연구에서 사용한 주된 파라미터 중 하나는 최대 컨덕턴스(Gmax)와 최소 컨덕턴스(Gmin)의 비율에 따른 조건이다. 이때 사용한 Gmax와 Gmin의 비율은 Gmax/Gmin이며, 0부터 100까지 5씩 증가시키며, 총 20개의 컨덕턴스 비를 사용했다. Fig. 6과 Table. 3은 각각 NeuroSim을 사용한 에포크에 대한 정확도 변화 그래프와 각 Gmax/Gmin의 값에 따라 학습 도중 나온 최대 정확도가 제시되었다. 이때 학습을 반복할수록 정확도가 증가한다는 것을 Fig. 6을 통해 확인할 수 있다. 따라서 MNIST 데이터셋을 사용하여, Gmax/Gmin이 5일 때, 정확도는 10.11%, 15일 때, 정확도는 49.10%이다. Gmax/Gmin이 20일 때, 정확도는 79.74%, Gmax/Gmin이 25일 때, 정확도는 95.88%라는 결과를 도출하였다. 결과적으로 Gmax/Gmin가 80일 때, 최대 분류 정확도를 달성하였다. 이러한 Gmax/Gmin 값의 최적화는 소비 전력 감소가 기여하게 되며, 이는 궁극적으로 정확도를 향상시킬 수 있다.

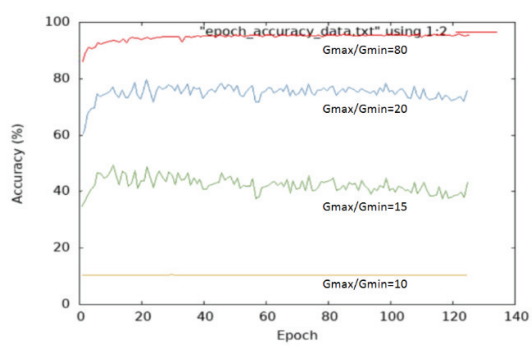

Fig. 6. Impact of Gmax/Gmin on the training accuracy of memristors.

컨덕턴스는 전기를 전달하는 물질의 물리적 속성이며, 환경 요인에 따라 변한다는 특징이 있다. 환경 요인으로 산소 분압과 온도가 있으며, 두 요인이 주로 컨덕턴스에 영향을 미친다. 일반적으로 온도가 증가할수록 전도성이 증가한다. 이러한 다양한 요인에서 Gmax/Gmin 값의 최적화는 소비 전력이 감소시켜 궁극적으로 정확도를 향상시

켜준다. Fig. 7은 Gmax/Gmin 값에 따라 학습 중 나타난 최대 정확도를 나타낸 그래프이다. 이 그래프는Gmax/Gmin 값이 증가할수록 정확도가 증가한다는 것을 나타내고 있으며, 이를 통해 두 요소 간 강한 연관성을 알 수 있다. 그러나 Gmax/Gmin의 비율이 높은 부분에서 주의할 필요가 있다. Gmax/Gmin가 높아질 경우에 과적합 현상이 발생할 수 있다[11]. 과적합은 기계 학습 모델이 학습 데이터에 지나치게 적합 되어 학습 데이터에 대해서 높은 성능을 보이지만 새로운 데이터에 대해 일반화하지 못하는 현상이다[12]. 이 현상은 학습데이터에 지나치게 의존하여 다른 데이터에 대한 적응력을 떨어뜨리기 때문에 정확성의 감소를 야기할 수 있다.

Table 3. Results of learning accuracy about conductance ratio

G_{max}/G_{mix}	정확도
10	10.11%
15	49.10%
20	79.74%
80	95.88%

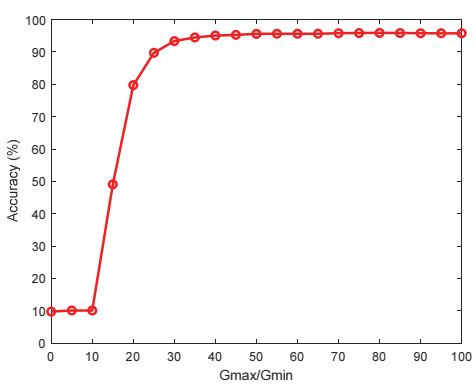

Fig. 7. Accuracy variation plot based on the Gmax/Gmin.

3.2 은닉층에 따른 정확도 변화

본 연구에서 사용된 또 다른 파라미터는 은닉층의 뉴런 수에 대한 조건이다. 많은 연구에서 정확도를 향상시키기 위해 적절한 은닉층의 뉴런 수를 찾는데 노력을 기울이고 있다. 하지만 아직 최적의 뉴런 수 공식을 발견한 연구는 없다. 비교를 위해, 앞서 시뮬레이션을 진행한 Gmax/Gmin에 대한 20가지의 값과 10부터 325까지 25씩 증가시킨 15가지의 은닉층의 뉴런 수에 대한 정확도를 추출하여 약 300개의 정확도를 얻었다. Fig. 8과 Table. 4는 NeuroSim을 사용하여 Gmax/Gmin가 80일 때, 은닉층의 뉴런 수에 대한 정확도 변화이다. MNIST 데이터셋을 사용한 결과, 은닉층의 뉴런 수가 10개일 때 정확도는 82.43%이다. 은닉층의 뉴런 수가 250개일 때, 정확도는 96.76%, 은닉층의 뉴런 수가 325개일 때 정확도는 57.70%이며, 이로써 최대 분류 은닉층의 뉴런 수가 250일 때 정확도가 가장 높게 나오는 것을 확인할 수 있다.

은닉층의 뉴런 수가 문제 데이터의 뉴런 수가 데이터의 복잡성에 비해 적으면 과소적합이 발생할 수 있다. 과소적합은 복잡한 데이터 세트에서 신호를 충분히 감지할 수 있는 숨겨진 계층의 뉴런이 너무 적을 때 발생한다. 불필요하게 많은 뉴런이 네트워크에 존재하면 과적합이 발생할 수 있다. Fig. 9는 Gmax/Gmin가 80일 때, 은닉층의 뉴런 수에 따라 학습 중 나타난 최대 정확도를 나타낸 그래프이다. Fig. 9를 통해 뉴런 수가 적을 때는 과소적합, 뉴런 수가 많을 때는 과적합이 발생하여 정확도가 급격히 낮아지는 것을 확인할 수 있다 [13].

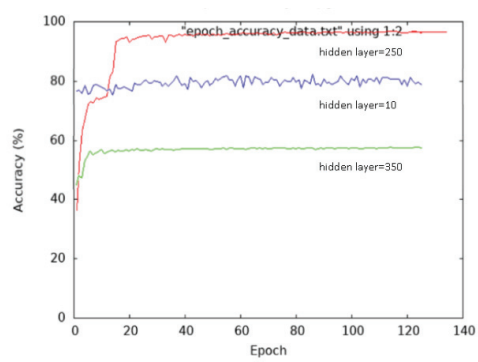

Fig. 8. Impact of the number of neurons in the hidden layer on the training accuracy of memristors.

Table 4. Results of learning accuracy about the number of neurons in the hidden

은닉층	정확도
10	82.43%
250	96.76%
325	57.70%

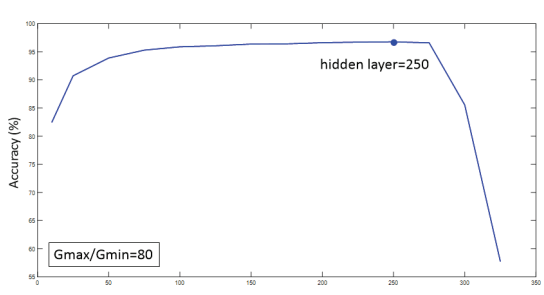

Fig. 9. Accuracy variation plot based on the number of neurons in the hidden layer.

3.3 최적의 정확도

Gmax/Gmin과 은닉층의 뉴런 수는 멤리스터의 정확도를 향상시키는 데 중요한 매개 변수이다. Fig. 10은 에포크가 125일 때, Gmax/Gmin와 은닉층의 값에 따른 정확도 그래프이다. 먼저, Gmax/Gmin의 값이 증가할수록 정확도는 증가하지만, 일정 비율 이상부터는 과적합이 발생하여 정확도가 조금씩 감소한다. 또한, 은닉층의 뉴런 수는 10부터 325까지 훈련해 본 결과, 250일 때, 정확도가 가장 높다. 앞의 두 매개 변수의 특성을 합쳐 Gmax/Gmin이 70이고, 은닉층의 뉴런 수가 250일 때가 96.83%로 가장 높은 정확도가 나오는 것을 확인했다. 그에 따른 정확도 추이는 Fig.11과 같다.

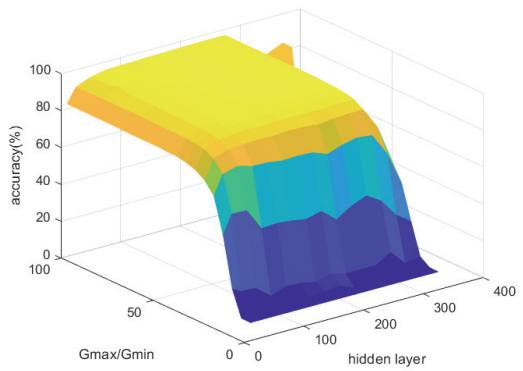

Fig. 10. Accuracy variation plot based on the Gmax/Gmin and the number of neurons in the hidden layer.

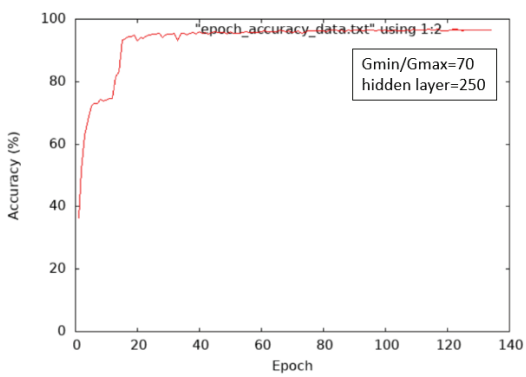

Fig. 11. Impact of the number of the Gmax/Gmin and neurons in the hidden layer on the training accuracy of memristors.

4. 결 론

본 논문에서는 IGZO 기반 멤리스터를 펄스 수에 따른 컨덕턴스 수식을 통해 MATLAB/Simulink와 NeuroSim으로 모델링하였다. 모델링한 IGZO 기반 멤리스터의 Gmax/Gmin과 은닉층을 변화시키고 시뮬레이션을 진행하였다. 시뮬레이션을 통해 위에서 언급한 2개의 매개변수에 따른 정확도 곡선을 얻어내고 이를 통해 높은 정확도를 도출하기 위한 최적의 매개변수를 추출하였다. 그 결과 Gmax/Gmin가 70이고 은닉층이250일 때, 정확도가 96.83%로 높은 정확도를 도출하였다.

감사의 글

This work is funded by a 2023 research Grant from Sangmyung University.

참고문헌

1. Park, Geon-Woo, et al. "A Review of RRAM-based Synaptic Device to Improve Neuromorphic Systems." Journal of The Korean Society of Semiconductor & Display Technology, Vol. 21, pp. 50-56, 2022.
2. Lu, Tian, et al. "Optimal Weight Models for Ferroelectric Synapses Toward Neuromorphic Computing." IEEE Transactions on Electron Devices (2023).
3. Chen, Wenbin, et al. "Essential Characteristics of Memristors for Neuromorphic Computing." Advanced Electronic Materials 9.2 (2023): 2200833.
4. Agarwal, Sapan, et al. "Resistive memory device requirements for a neural algorithm accelerator." 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016.
5. Zhang, Li, et al. "Resistive switching performance improvement of InGaZnO-based memory device by nitrogen plasma treatment." Journal of Materials Science & Technology 49 (2020): 1-6.
6. Pereira, Maria Elias, et al. "Tailoring the synaptic properties of a-IGZO memristors for artificial deep neural networks." APL Materials 10.1 (2022).
7. Choi, Hyun-Woong, et al. "Zinc oxide and indium-gallium-zinc-oxide bi-layer synaptic device with highly linear long-term potentiation and depression characteristics." Scientific reports 12.1 (2022): 1259.
8. Agarwal, Sapan, et al. "Resistive memory device requirements for a neural algorithm accelerator." 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016.
9. Chen, Pai-Yu, Xiaochen Peng, and Shimeng Yu. "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 37.12 (2018): 3067-3080.
10. Liu, Huan, Wei, Min and Chen, Yuzhong. "Optimization

of non-linear conductance modulation based on metal oxide memristors" Nanotechnology Reviews, vol. 7, no. 5, 2018, pp. 443-468.
11. Karsoliya, Saurabh. "Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture." International Journal of Engineering Trends and Technology 3.6 (2012): 714-717.
12. Lee, Yong-Hwan and Kim, Heung-Jun. "Implementation of Fish Detection Based on Convolutional Neural Networks." Journal of The Korean Society of Semiconductor & Display Technology, Vol. 20, pp. 124-129, 2020.
13. Ying, Xue. "An overview of overfitting and its solutions." Journal of physics: Conference series. Vol. 1168. IOP Publishing, 2019.

접수일: 2023년 12월 5일, 심사일: 2023년 12월 18일, 게재확정일: 2023년 12월 19일