대규모 뉴로모픽 칩 및 시스템 소개

DOI: 10.3938/PhiT.28.012

박 종 길

Introduction to Large-Scale Neuromorphic Systems

Jongkil PARK

Neuromorphic engineering has been studied for the last few decades as a way to mimic biological systems by using electronics. Many researchers have used electronics to demonstrate various systems mimicking neural and synaptic dynamics. In this article, we introduce the idea of neuromorphic engineering from the viewpoint of circuits and systems design. We explain spiking neural networks, which are run on neuromorphic systems. We introduce the principle components of neuromorphic system design, such as neuron and synapse circuits. In addition, we introduce various large-scale neuromorphic systems that have been developed by various worldwide neuromorphic research groups.

뉴로모픽 공학의 시작, 말하고자 하는 바

1. 뉴로모픽 공학의 시작

뉴로모픽 공학은 생물학적 시스템을 전기적인 시스템으로 모사하고자 하는 연구로 1980년대 후반 캘리포니아 공과대학의 Carver Mead 교수에 의해 처음 제안되었다.^[1,2] 생물학적 시스템을 모사하고자 하는 것의 근본적인 동기는 전력 효율이 뛰어난 연산 기계를 만들고자 함에 있다. 일례로 인간의 두뇌는

저자약력

박종길 박사는 2014년 미국 University of California, San Diego에서 전자공학 박사학위를 받았으며, 한국전자통신연구원(ETRI)에 2017년까지 재직 후, 2018년부터 현재까지 한국과학기술연구원(KIST)에 재직 중이다. 뉴로모픽 시스템 설계 및 스파이크 기반 온라인 학습법을 이용한 스파이킹 신경망 알고리즘 등을 활발히 연구 중이다.(jongkil@kist.re.kr)

Fig. 1. Hierarchy of neuromorphic engineering research domain.

인지 및 판단 등을 수행하는데 뛰어난데 불구하고 20 W의 전력을 소모한다고 알려져 있지만, 기존 폰노이만 방식의 컴퓨터는 단순한 패턴 인식을 위해 필요한 전력의 양도 상당하다는 점을 근거로 들 수 있다. Carver Mead 교수는 첫 논문에서 트랜지스터의 특정 동작 영역(sub-threshold region)에서 발현되는 물리적 현상이 생물학적 뉴런의 막에서 발현하는 것과 유사하다는 것을 밝혔다. 이를 근거로 트랜지스터를 디지털 스위치 소자가 아닌 시냅스 소자로 사용할 수 있을 것이라고 제안한다. 이를 잘 활용하여 구조적으로 배열하여 연결하면 생물학적 시스템의 뉴런의 구조적 특징 및 기능을 모사할 수 있을 것이라고 생각하였다.

2. 뉴로모픽 공학에서 다루고 있는 분야

초기의 뉴로모픽 공학은 기존 Complementary metal-oxide-semiconductor(CMOS) 공정을 이용한 아날로그 회로 설계 및 시스템 관점에서 연구가 많이 진행되었다. CMOS 트랜지스터

REFERENCES

- [1] C. Mead, "Neuromorphic electronic systems," Proceedings of the IEEE **78**, 1629 (1990).
- [2] G. Indiveri *et al.*, "Neuromorphic silicon neuron circuits," Frontiers in Neuroscience **5**, Article 73 (2011).

를 이용하여 단일 뉴런 및 시냅스의 동작 특성을 전기적으로 모사할 수 있음을 보여주는 연구로 시작하였다. 이를 확장하여 뉴런을 작은 규모(128~1,024개) 집적하여 간단한 생물학적 네트워크의 동작을 모사할 수 있음을 보여주기도 하였다. 하지만 뇌를 모사하는 연구 주제의 복잡성 등으로 인해 현재는 다양한 관점에서의 연구가 진행되면서 뉴로모픽 공학의 연구 범위가 넓어지고 있다. 먼저, 시냅스 및 뉴런의 동작 특성을 기존의 실리콘 기반 CMOS 공정이 아닌 새로운 소자로 구현하고자 하는 연구가 진행 중이다.^[3] 또한, 다양한 형태로 구현되는 뉴런 및 시냅스를 이용하여 제작되는 뉴런과 시냅스를 연결하여 대규모로 집적화된 시스템으로 구성하기 위한 회로 설계 및 시스템, 아키텍처 관점에서의 연구가 진행 중이며,^[4-9] 이러한 설계된 시스템을 이용하여 인공지능을 구현하기 위한 알고리즘이 동시에 연구가 되고 있다.(그림 1 참조)

이와 같이 뉴로모픽 공학은 다양한 관점에서의 접근이 필요한 분야지만 본 원고에서는 CMOS 공정을 이용한 뉴로모픽 회로 및 시스템 설계와 관련된 연구 주제들과 최근의 연구 동향에 대하여 간략히 소개한다. 아래 글 내용은 다음과 같이 구성하였다. 처음, 뉴로모픽 시스템을 이용하여 구동하고자 하는 스파이킹 신경망이 무엇인지 알아본다. 뉴로모픽 시스템 설계에 필요한 구성요소들인 뉴런과 시냅스를 설계하는 방법에 대하여 알아본다. 또한, 대규모 뉴로모픽 시스템 구성을 위해 필요한 시스템 아키텍처 설계 관점에 대하여 논한 뒤, 최근에 발표된 다양한 시스템에 대한 간략한 소개를 다룬다. 마지막으로 뉴로모픽 시스템의 미래에 대하여 논하는 것으로 글을 마친다.

뉴로모픽 시스템의 구현

1. 스파이킹 인공 신경망(Spiking neural network, SNN)

뉴로모픽 시스템을 소개하기 앞서 먼저 뉴로모픽 시스템을 이용하여 실행하고자 하는 알고리즘이 무엇인지 이해할 필요가 있다(그림 2). 뇌의 동작을 모사한 인공지능 구현을 위해 시냅스를 통해 연결된 뉴런들의 연결망으로 구성된 인공 신경망(Artificial neural network, ANN) 구조가 제안되었다. 하나의 뉴런은 수많은 다른 뉴런들과 시냅스를 통해 연결이 되어있고 시냅스를 통해 전달되는 입력에 따른 다중합의 결과로 다음 뉴런으로의 신호 전달 여부를 결정한다. 현재 대중적으로 알려져 있는 다양한 기계학습 어플리케이션은 대부분 이러한 인공 신경망 구조에서 기인한 심층 신경망(Deep neural network, DNN)으로 구현되어 있다. 이와 달리 뉴로모픽 시스템에서 구현하고자 하는 하드웨어는 스파이킹 신경망(Spiking neural network, SNN)을 구현하는 것을 목표로 한다.(그림 2 참조) 스파이킹 신경망은 차세대 인공 신경망이라고도 일컬어지며 신

Fig. 2. Comparison of deep neural network (DNN) and spiking neural network (SNN).

경과학 관점에서 생물학적 타당성을 더욱 근접하게 모사하여 인공지능 구현을 하고자 한다. 스파이킹 신경망은 실제 두뇌에서 뉴런 간의 상호 스파이크 발화 신호를 주고받듯이 실시간으로 스파이크 입력을 받는다. 스파이크 입력에 따른 뉴런의 상태 변수가 실시간으로 변화하고 연속적으로 입력되는 스파이크 입력들에 따라 출력이 결정된다. 생물학적 뉴런의 실제 동작 특성을 모사하였기 때문에 인공지능 구현에 더 적합한 모델로 여겨지지만 입력 정보에 시간 정보가 추가되어서 해석에 어려움이 존재한다.

스파이킹 신경망은 뉴런의 상태 변수가 실시간으로 변화하고 스파이크 입력이 가지는 시간 정보가 중요하기 때문에 스파이크 입력을 실시간으로 처리해 주어야 하는 점이 중요하다. 뉴런의 상태 변수를 매 시간정확도마다 계산해 주어야 하기 때문에 뉴런 상태 변수를 항상 저장할 수 있는 메모리 공간도 필요하다. 이러한 점 때문에 기존에 심층 신경망 구조처리에 도움이 되는 그래픽카드(GPU) 등을 이용한 연산 방식으로는 스파이킹 신경망 구현이 어렵다. 그래서 스파이킹 신경망 구현

REFERENCES

- [3] Jongkil Park, Vacuum magazine 4(3), 21-24 (2017).
- [4] P. A. Merolla *et al.*, Science 345, 668 (2014).
- [5] M. Davies *et al.*, IEEE Micro 38, 82 (2018).
- [6] A. Necker *et al.*, "Braindrop: A mixed-signal neuromorphic architecture with a dynamical systems-based programming model," Proceedings of the IEEE 107, 144 (2019).
- [7] J. Park *et al.*, IEEE Transactions on Neural Networks and Learning Systems 28, 2408 (2017).
- [8] S. Furber *et al.*, Proceedings of the IEEE 102, 652 (2014).
- [9] S. Schmitt *et al.*, "Neuromorphic hardware in the loop: training a deep spiking network on the BrainScaleS wafer-scale system," 2017 International Joint Conference on Neural Networks (IJCNN) (2017).

Fig. 3. Diagram of address event representation protocol.

에 적합한 구조의 연산 장치를 개발하여 스파이킹 신경망 연 구를 가속시키고자 하는 것이 뉴로모픽 시스템 설계가 가지는 하나의 중요한 의미라고 할 수 있다.

2. 뉴런의 설계 방법

뉴로모픽 시스템을 구성하기 위한 기본 연산 단위는 뉴런과 시냅스이다. 뉴런은 시냅스를 통해 전달된 이온들로 인해 상승 된 막전위가 특정 문턱전압을 넘어서면 스파이크를 발현하게 된다. 이러한 뉴런의 동작 특성을 물리적으로 분석하여 수학적 으로 모델링하는 방법은 다양하게 존재한다. 뉴런의 막에서 벌 어지는 이온채널의 현상까지 모델링한 복잡도가 높은 모델도 존재하며 뉴런의 함축적인 동작 특성인 시냅스의 정보 집합에 따른 뉴런 스파이크 발화만을 모사하여 계산 복잡도를 단순화 한 모델이 있다. Leaky-integrate-and fire(LIF) 뉴런 모델이 대표적으로 많이 사용되는 모델이다. LIF 뉴런을 CMOS 공정 을 이용하여 구현하기 위한 설계 방법은 크게 아날로그로 설 계하는 방법과 디지털로 설계하는 방법으로 나뉜다.

아날로그로 설계할 때는 캐패시터를 이용하여 뉴런의 막전위 를 저장한다. 트랜지스터로 구현된 시냅스를 통해 캐패시터에 전하가 주입되고 캐패시터의 전위가 높아지게 된다. 아날로그 비교기가 캐패시터의 전위 값을 측정하여 특정 전위 값을 넘 어가면 스파이크를 발현한다. 아날로그로 뉴런을 설계하면 캐 패시터의 자연적인 전하 방출로 막전위의 시상수 구현이 용이 하며 실제 뉴런의 동역학 구조를 잘 설명할 수 있다는 장점이 있다. 하지만 아날로그 소자의 산포에 따른 동작 특성이 매우 다르기 때문에 뉴런을 집적화하였을 때 신경망의 결과를 분석 하기 매우 어렵다는 단점이 있다.

뉴런을 디지털로 설계할 때는 메모리 소자에 디지털 값으로 뉴런의 막전위 값을 저장한다. 디지털 값으로 입력되는 시냅스 값을 이용하여 메모리에 저장되어 있는 뉴런의 막전위 값을 조정한다. 막전위 값이 특정 값을 넘어서면 뉴런이 스파이크를 발현한다. 가산기, 감산기 등의 디지털 연산기를 이용하여 뉴 런의 동역학을 설명하는 수식을 계산하는 방식으로 디지털 뉴 런은 동작한다. 디지털 방식으로 설계하는 뉴런은 설계 및 구 현이 용이하고 신경망 결과 분석이 용이한 장점이 있지만 시

Fig. 4. Various structures of multichip neuromorphic systems.

상수 구현을 위해서는 매 시간 메모리 값을 읽어 들여 계산해 주어야 하는 단점이 존재하고 메모리가 칩 구현 시 차지하는 면적이 넓어서 뉴런을 집적할 때 단점이 있다.

3. 가상 시냅스 연결을 위한 주소기반 표현 방식

뉴로모픽 시스템에서는 시냅스를 통해 뉴런과 뉴런 사이를 연결해 주어야 한다. 실제 두뇌의 뉴런 한 개는 약 1,000여 개 의 다른 뉴런과 연결이 되어 있으며 이러한 연결성이 규칙적 이지 않다. 또한 입력에 따른 학습의 결과물로 연결성이 달라 질 수도 있다. 이러한 이유로 인해 뉴런들 간의 연결을 물리적 인 연결로 구현하기 매우 어려운 점이 존재한다. 이를 극복하 기 위해 뉴로모픽 시스템에서는 주소기반 표현 방식(Address event representation)을 뉴런 간의 통신 프로토콜로 제안하여 문제를 해결하였다(그림 3). 주소기반 표현 방식은 디지털 패킷 을 이용하여 가상의 시냅틱 연결을 구성하는 방법이다. 어레이 형태로 집적되어 있는 뉴런이 스파이크를 발화하면 뉴런이 어 레이 상에서 가지고 있는 주소가 이벤트로 표현되어 전달된다. 예를 들어, 뉴런 어레이 상의 '10'번 뉴런이 발화를 하면 디지 털 버스를 통해 '10'이라고 부호화된 이벤트를 전송한다. 이 이 벤트는 뉴런 어레이 외부에 있는 메모리 테이블(Lookup table)로 전달되어 '10'번 뉴런과 연결되어 있는 시냅스 정보를 읽어 오게 된다. 저장되어 있는 시냅스 정보는 연결되어 있는 뉴런 의 주소 및 시냅스의 강도 등이 될 수 있다. 읽어 들인 시냅스 이벤트들을 다시 뉴런 어레이로 전달하여 가상의 시냅스 연결 을 구성한다.

4. 시스템 아키텍처 설계

뉴로모픽 시스템 설계의 목표는 대규모로 뉴런을 집적하는 것에 있다. 예를 들어 성인의 뇌에는 평균적으로 10^{11} 개의 뉴런이 존재하고 하나의 뉴런은 약 $10^3 \sim 10^4$ 개의 다른 뉴런과 연결되어 있다고 알려져 있다. 이처럼 뇌를 모사하고자 하는 목표치에 근사하기 위해서도 대규모로 뉴런을 집적하고자 하는 이유가 있고 실제로 연구가 진행되면서 구현하고자 하는 스파 이킹 신경망의 크기 또한 커지고 있기 때문에 이에 대응하기 위해서는 대규모로 뉴런을 집적한 시스템 설계가 필요하다. 주

소기반 표현 방식을 통해 작은 규모로 집적된 뉴런들의 시냅스 연결성을 쉽게 확보할 수 있다. 하지만 뉴로모픽 시스템을 더욱 확장한 시스템(1,000,000개 이상 뉴런)을 만들기 위해서는 아키텍처 레벨에서 해결책이 필요하다.

하나의 칩에 집적할 수 있는 뉴런의 개수는 물리적으로 집적할 수 있는 공간에 제약을 받는다. 따라서 여러 개의 칩을 연결하여 뉴런의 집적도를 시스템 차원에서 높이고자 한다. 앞서 언급한 바와 같이 스파이킹 신경망에서는 스파이크 신호가 가지는 발화 타이밍이 하나의 중요한 정보를 담고 있다. 따라서 단순히 여러 칩을 물리적으로 붙이는 것에 그치지 않고 시스템 전체 내에서 발화하는 스파이크 신호가 시스템 내에서 시간적으로 동기화될 필요가 있다. 이를 위해서는 스파이크 이벤트들이 효율적으로 전달할 수 있는 시스템 아키텍처가 제안되어야 한다. 먼저 가장 단순하게는 뉴런 칩 여러 개를 직렬로 연결하는 방법이 있을 수 있다. 이러한 직렬구조를 2차원적으로 표현한 메쉬(Mesh) 형태의 아키텍처도 있으며 3차원 형태로 연결한 원환면 구조로 아키텍처를 설계하여 칩을 배열하기도 한다. 또한 트리 형태의 계층 구조로 아키텍처를 설계하여 전체 시스템의 확장성을 확보하는 구조도 널리 제안되고 있다. (그림 4)

5. 설계해 놓은 시스템들을 소개

앞서 언급한 뉴런과 시냅스, 그리고 이를 시스템 상에서 연동하는 아키텍처를 반영하여 다양한 시스템들이 제작되었다. 미국에서는 정부 주도 과제로 시작된 뉴로모픽 과제들이 성공적으로 안착하여 최근 들어 IBM이나 Intel과 같은 반도체 대기업들이 시작품을 출시하기 시작하였다. 또한 Stanford University, UCSD 등 학계에서도 다양한 형태의 시스템을 발표하였다. 유럽에서는 EU 주도의 Human Brain Project (HBP)의 결과물로 SpiNNaker와 BrainScale 같은 시스템을 발표하였다.

IBM에서 발표한 TrueNorth는 디지털 방식의 뉴로모픽 칩으로 뉴런 코어를 메쉬 형태의 아키텍처로 연결하였다.^[4] 집적도를 높이기 위해 시냅스의 정확도를 1 bit로 낮추었고 학습기능은 구현되지 않았다. Intel에서도 비슷한 디지털 방식의 뉴로모

픽 칩인 Loihi를 2018년 발표하였다.^[5] TrueNorth와는 달리 신경망 구조가 재구성 가능하며 제한적인 학습 기능이 구현되었다. 디지털 방식의 뉴로모픽 시스템과는 달리 학계에서는 아날로그 방식의 뉴로모픽 시스템도 많이 연구되었다. Stanford에서는 BrainDrop으로 명명한 시스템을 선보였다.^[6] 생물학적 뉴런과 신경망의 구조를 아날로그 형태로 모사하였다. 트리 형태의 계층 구조로 아키텍처를 구성하였다. UCSD에서는 HiAER-IFAT이라는 계층적 시스템을 개발하였다.^[7]

유럽의 HBP에서 수행하는 뉴로모픽 시스템 설계는 크게 두 가지가 있다. 영국의 맨체스터 대학에서 개발하는 SpiNNaker라는 시스템이 있다. 이 시스템은 기존 CPU에서 사용되는 ARM 코어를 기반으로 설계되었으며 ARM 코어들 사이에서 스파이크 이벤트 교환을 최적화한 라우터 설계에 초점을 맞추어 설계되었다.^[8] 기존 ARM 코어 기반으로 설계되었기 때문에 소프트웨어 인터페이스 구성이 용이하다는 장점이 있지만 뉴로모픽 공학에서 추구하는 저전력 시스템 구성과는 멀어지는 경향이 존재한다. 또 하나의 시스템으로는 BrainScaleS이라는 뉴로모픽 시스템이 있다.^[9] 아날로그 뉴런을 기반으로 설계되었으며 멀티 칩 라우팅 문제를 해결하기 위해 웨이퍼 스케일에서 라우팅 문제를 해결하고자 시도하였다.

뉴로모픽 공학 시스템의 미래

본 원고에서는 뉴로모픽 공학의 한 분야인 회로 설계 및 시스템 관점에서 뉴런이 대규모로 집적된 뉴로모픽 시스템에 대하여 논하였다. 현재 개발된 다양한 대규모 뉴로모픽 시스템들은 뉴로모픽 공학 연구를 통해 스파이킹 인공 신경망 알고리즘을 검증할 수 있는 플랫폼으로 의미가 크다. 이러한 시스템들은 기존 컴퓨터를 이용하여 연구하기 힘들었던 스파이킹 인공 신경망 알고리즘 등을 검증하기에 매우 용이한 시스템이 될 것이다. 하지만 뇌를 모사한 인공지능을 구현하기 위해서는 하드웨어가 진정한 비-폰노이만 방식으로 구현되어야 한다. 이를 위해 다양한 분야의 학문적 발전이 동반되어야 한다. 앞서 언급한 소재, 소자 분야에서의 발전과 함께 계산 신경학 분야에서의 이해도 함께 발전하면 진정한 뉴로모픽 시스템이 등장할 것으로 기대된다.