인공지능 뉴로모픽 반도체 기술 동향

Trend of AI Neuromorphic Semiconductor Technology

오광일 (K.I. Oh, kioh@etri.re.kr) 초경량지능형반도체연구실 선임연구원
김성은 (S.E. Kim, sekim@etri.re.kr) 초경량지능형반도체연구실 선임연구원
배영환 (Y.H. Bae, yhbae@etri.re.kr) 초경량지능형반도체연구실 책임연구원
박경환 (K.H. Park, khpark_2001@etri.re.kr) 초경량지능형반도체연구실 책임연구원/실장
권영수 (Y.S. Kwon, yskwon@etri.re.kr) 지능형반도체연구본부 책임연구원/본부장

ABSTRACT

Neuromorphic hardware refers to brain-inspired computers or components that model an artificial neural network comprising densely connected parallel neurons and synapses. The major element in the widespread deployment of neural networks in embedded devices are efficient architecture for neuromorphic hardware with regard to performance, power consumption, and chip area. Spiking neural networks (SiNNs) are brain-inspired in which the communication among neurons is modeled in the form of spikes. Owing to brainlike operating modes, SNNs can be power efficient. However, issues still exist with research and actual application of SNNs. In this issue, we focus on the technology development cases and market trends of two typical tracks, which are listed above, from the point of view of artificial intelligence neuromorphic circuits and subsequently describe their future development prospects.

KEYWORDS 인공지능, 뉴로모픽, 회로, 딥러닝, 반도체, SoC, DNN, SNN

Ⅰ. 서론

깊은 신경망(DNN: Deep Neural Networks)에 기반을 둔 인공지능 알고리즘은 여러 복잡한 인지 과제에서 인간과 유사하거나 인간을 뛰어넘는 성능을 입증했지만, 이러한 알고리즘을 구현하는 컴퓨팅 시스템의 에너지 효율은 인간 뇌와 비교하여 아직 상당한 차이가 존재한다. 이러한 인공지능 알고리즘 대부분은 중앙처리장치(CPU: Central Processing Unit), 그래픽처리장치(GPU: Graphics Processing Unit), Field-Programmable Gate Array(FPGA)와 같은 기존 컴퓨팅 시스템에서 실행되며, 최근에

* DOI: https://doi.org/10.22648/ETRI.2020.J.350308
* This work was supported by the ICT R&D program of MSIT/IITP[2018-0-00197, Development of ultra-low power intelligent edge SoC technology based on lightweight RISC-V processor].

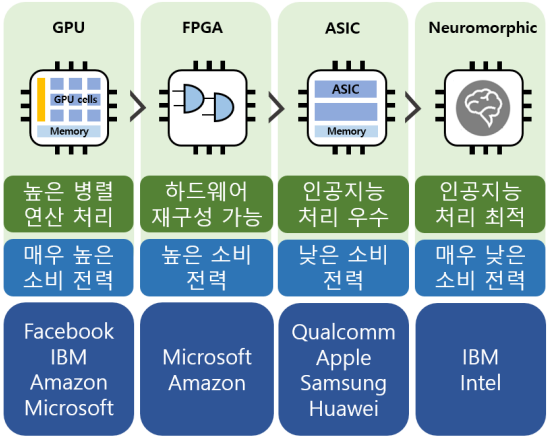


그림 1 인공지능 데이터 처리 기술별 특징

는 기계학습 가속화를 위해 digital type 또는 analog digital mixed-signal type의 Application Specific Integrated Circuit(ASIC)도 개발되고 있으나 Moore 법칙의 확장 한계가 임박함에 따라 이러한 기존 접근방식의 확장으로 얻을 수 있는 성능 및 전력 효율은 감소하고 있는 상황이다. 전통적인 폰 노이만 아키텍처에서 고성능 중앙처리장치는 메모리에서 가져온 데이터와 명령에 의하여 순차적으로 작동하며 인공지능 알고리즘을 이러한 시스템을 이용하여 수행할 때 폰 노이만 병목현상은 결국 연산유닛과 데이터 스토리지 사이에서 발생하는 성능 한계로 정의된다. 따라서 최근 인공지능 뉴로모픽 회로 기술 연구의 초점은 AI 알고리즘, 장치 기술, 통합 체계 및 아키텍처 설계뿐만 아니라 기존 컴퓨터의 연산유닛-메모리 병목현상을 극복하기 위한 노력에 집중되고 있다. 인공지능 가속기는 신경망의 효율적인 연산 최적화를 추구하는 고성능 뉴로모픽 하드웨어의 한 종류로서 그것의 아키텍처는 대규모 병렬처리를 위한 메모리와 연산유닛의 유기적인 연합을 적용하고 있다. 이러한 방법을 이용하여 복잡한 신경 네트워크에 필요한 계산은 종래 폰 노이만 아키텍처에 비해 더 빠르고 더 적은 전력으로 수행 가능하다. 신경망은 매우 규칙적인 구조를 가지므로 동일 유형의 연산유닛을 병렬로 사용하여 대규모 병렬 컴퓨팅이 가능하며 종래 병렬 컴퓨팅 아키텍처에서 알려진 아키텍처적 접근방식의 적용이 가능하다.

한편, 인공지능 컴퓨팅에 대해 기존과는 다른 접근법으로 다양하게 연구가 진행되고 있으며, 이들 중 생물학적 원리에서 영감을 받아 이를 통하여 뉴럴 네트워크를 구현하고자 하는 연구 기술로서 스파이킹 신경망(SNN: Spiking Neural Networks) 기술이 있다. 스파이킹 신경망은 뉴로사이언스 분야의 신경생물학적인 메커니즘을 인공지능 분야로 직접 접목한 기술로서, 뇌의 주요 기능 중의 하나인 시간 기반 정보 인코딩과 시간 기반 정보처리 측면을 활용하는 제3세대 인공지능 기술로 알려져 있다. 스파이킹 신경망 기반 뉴로모픽 컴퓨팅 플랫폼은 시냅스를 통한 시냅틱 연산 및 정보 저장을 이벤트 기반 비동기식 스파이크 동작 메커니즘에 의하여 정보를 전달하며, 다수의 비교적 단순한 연산 장치인 뉴런에 분산시킴으로써 생물학적인 스파이킹 신경망 메커니즘을 효율적으로 모방하는 것을 목표로 한다. 비록 현재 DNN기반의 하드웨어에 비하여 성능적인 열화 및 실제 적용 가능한 적합한 어플리케이션의 부재 등 단점들이 존재하나, 스파이킹 신경망의 이벤트 기반 비동기적 동작 특성은 메모리 및 연산유닛이 결합된 고효율 컴퓨팅 아키텍처를 도출하였으며, 병렬성을 크게 높이고 하드웨어 에너지 소모를 크게 절감할 수 있다고 알려져 있다.

본 고에서는 인공지능 뉴로모픽 회로 관점에서 기술의 동향을 상기 열거한 대표적인 두 가지 트랙의 기술 개발 사례 중심과 시장 동향 중심으로 살펴보고, 기술의 향후 전망에 대하여 논의점을 찾고자 한다.

Ⅱ. 뉴로모픽 반도체 기술 흐름

1. AI 가속기로서의 뉴로모픽 반도체

일반적으로 인공지능 가속기(Accelerator)는 인공지능 알고리즘의 빠른 처리를 위해 특별히 설계된 하드웨어 혹은 마이크로칩을 말한다. 전용의 목적을 가지는 여타 가속기와 마찬가지로 인공지능 가속기는 일반적인 x86 기반의 CPU로는 비효율적으로 동작하는 인공지능과 관련된 특정 작업을 효율적으로 수행하도록 설계되어 있다. 목적에 맞게 설계·제작된 가속기는 주어진 작업 수행 용이성을 위하여 더 높은 성능, 더 많은 기능과 더 높은 전력 효율을 제공함을 목표로 한다.

AI 알고리즘의 대부분은 대규모의 병렬 연산을 기반으로 하고 있으며, 종래 3차원 그래픽스 관련 벡터 연산에 사용되었던 GPU를 인공지능 알고리즘 처리에 활용하게 되었고, 그로 인하여 GPU에 내장된 많은 연산 코어를 인공지능 알고리즘 연산에 활용하여 인공지능과 관련된 작업의 가속이 가능하게 되었다. GPU의 구조적 특징을 활용하여 최근까지도 인공지능 알고리즘과 같은 대규모 병렬 컴퓨팅 구현에 GPU를 활용하고 있다.

한편, 인공지능 연산 가속을 위한 ASIC 칩은 TPU, NPU, VPU 등 여러 유형의 아키텍처를 가지며 각기 인공지능 응용을 위한 전용 아키텍처를 탑재하고 있다. 기본적으로 ASIC 칩은 GPU 또는 FPGA보다 더 높은 에너지 효율성을 가지고 고속으로 동작하며 die 크기가 더 작은 것이 특징이다. ASIC은 다양한 높은 연산 특성을 가지는 매우 규칙적인 데이터 처리를 에너지 효율적이고 신속하게 처리하는 것을 첫 번째 목표로 하여 궁극적으로는 CPU와 같은 범용성까지 탑재할 것을 추구하고 있으나 설계에서 제작까지 이르는 개발주기가 매우 길고, 한번 설계된 내부의 구조는 변경이 불가하여 다양한 어플리케이션에 유연하게 대응이 어려우므로 범용성 탑재는 힘들 것으로 보고 있다.

인공지능 연산 가속 측면에서의 FPGA는 고객의 요구에 따라 인공지능 연산에 필요한 칩 내부 구성의 배열을 높은 자유도로 설정할 수 있는 칩을 일컫는다. 일반적으로 FPGA는 ASIC보다 설계주기가

표 1 AI 가속기 구현방법에 따른 특징

구현 방법	장점	단점
범용 CPU	• S/W 프로그래밍 가능하여 자유도가 높음	• 낮은 에너지 효율성 • 전용 아키텍처에 비해 낮은 계산 성능
범용 CPU + 가속기	• 작업 용도에 맞게 가속기 tuning 가능	• 부착된 가속기의 개수와 성능에 따라 계산 성능이 결정됨
범용 GPU	• 서버에서 대규모 학습을 위한 사용 가능 • 병렬 연산성이 높음	• 추론에 대한 에너지 효율성이 낮음 • 소비 전력이 높음
전용 칩 (ASIC)	• 추론에 대한 저전력 동작이 가능함 • 목표 작업에 대하여 높은 성능 도출함	• 제작 이후 설계 변경이 매우 어려움 • 학습을 위한 설계가 어려움 • 높은 개발 비용과 긴 개발 기간
FPGA	• 하드웨어적 재구성 및 자유도가 높음 • 사용자가 원하는 아키텍처로 변형 가능함 • AI를 위한 전용 IP를 탑재하는 추세 • 시장에 대한 기간적 대응성이 높음	• 저전력 어플리케이션에 부적합 • 칩당 단가가 높음
DSPs	• 인공지능 연산에 가장 중요한 MAC 연산 성능이 높음	• 폰 노이만 bottleneck 단점 존재

짧으며 GPU 기반의 하드웨어보다 소비 전력이 낮으나 하드웨어적인 설계 자유도가 높으므로 내부 게이트 카운트의 규모가 커질수록 칩당 가격도 상대적으로 높다. 최근 인공지능 하드웨어 시장에서 FPGA는 에너지 효율성을 추구하는 ASIC과 구현의 유연성을 추구하는 GPU 사이에서 적절한 절충안으로 자리 잡고 있으며, 매우 급변하고 있는 AI 알고리즘에 대응할 수 있는 솔루션으로 주목받고 있다. 또한, FPGA는 ASIC 제작으로 발생할 수 있는 비용 및 기술적 제한을 피하면서 각각의 어플리케이션에 적합한 사용자에 의해 정의된 칩을 구현할 수 있는 장점이 있다. 표 1에 상기 열거한 GPU, ASIC, FPGA 외에 인공지능 가속기로서의 다양한 하드웨어 구현방법에 따른 특징과 장단점을 정리하였다.

2. 뇌를 모방한 뉴로모픽 반도체

인간 뇌 속의 '연산'이라고 함은 오늘날의 전통적인 폰 노이만 구조의 작업 처리와는 전혀 다른 패러다임을 따른다. 폰 노이만 구조를 따르고 있는 기존의 컴퓨팅 시스템은 데이터의 정확한 수치적 표현을 전송하고 수정하도록 최적화된 반면, 인간의 뇌는 행동 잠재력 또는 스파이크라고 불리는 시간적 사건에 대해 작동한다. 신경세포는 시냅스를 통해 이러한 스파이크를 받아 세포막 전위의 작은 변화로 변환하고 뉴런은 시간이 지남에 따라 이러한 잠재적 변화를 통합하며, 많은 스파이크가 짧은 시간 내에 도착하면 뉴런은 비로소 스파이크를 출력한다. 스파이크는 컴퓨팅적 의미에서 메시지로 간주될 수 있으나 해당 시점의 시간과 출처 외에 어떤 정보도 가지고 있지 않다는 점에서 종래의 메시지 개념과는 다르며 매우 단순하다. 따라서 두뇌에서의 컴퓨팅은 아주 간단한 구조의 컴퓨팅 노드(뉴런)를 기반으로, 매우 단순한 메시지인 스파이크를 통해 통신하는 이벤트 기반 동작으로 설명할 수 있다.

인간의 뇌에는 약 10^{11}개의 뉴런이 존재하는 것으로 알려져 있으며, 인간의 신피질에는 뉴런당 약 5,000~10,000개의 시냅스가 존재하는 것으로 추정된다. 하나의 뉴런이 다른 뉴런으로부터 신호를 수신할 수 있는 확률은 약 10^{-6}%로서 매우 희박한 신호 연결성을 가지며 이는 한 덴드라이트(Dendrite)가 복수의 시냅스를 형성할 수 있다는 점을 고려하면 신호 연결성은 훨씬 더 작다. 대신에 뉴런 fan-out 연결성은 매우 크며, 뇌 전체에 대한 connection은 약 10^{15}개 수준으로 추정된다. 스파이킹 신경망 구조의 인공지능 뉴로모픽 하드웨어는 대표적으로 맨체스터 대학의 SpiNaker, IBM의 TrueNorth, 인텔의 Loihi, 하이델베르크 대학의 BrainScaleS, 스탠포드 대학의 NeuroGrid, 스위스 취리히 대학의 DYNAP 등이 있으며, 모두 고에너지 효율의 SNN기반 뉴로모픽 하드웨어 구현을 추구하고 있다. 특히 digital-analog mixed type 기반의 SNN 뉴로모픽 하드웨어는 현재의 실리콘 CMOS 회로 구현 기술에서 확장하여, 나노 규모의 메모리 소자 분야의 발전과 결합되면 더욱 고효율의 에너지 소모 목표를 달성할 수 있을 것으로 전망하고 있다.

III. 뉴로모픽 반도체 개발 동향

1. 인공지능 반도체 시장 동향

각종 빅데이터 분석 및 자율주행 자동차를 포함한 제4차 산업분야에 많이 활용되고 있는 인공지능 기술은 서버 단에서 점차 벗어나 엣지 단으로 내려오고 있다. 엣지 단에서 인공지능 서비스 제공이 가능해지면 민감한 개인 데이터를 서버로 업로

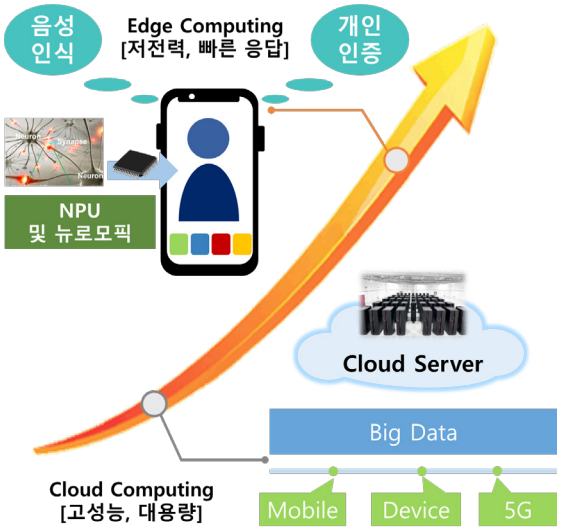


그림 2 인공지능 반도체 시장 동향

드할 필요가 없으므로 보안성이 높으며, 인공지능 서비스의 네트워크 연결에 대한 의존성이 낮아져 편리함 및 신속성을 확보할 수 있게 되므로 관련 기업들은 엣지 단에 인공지능 기능을 탑재하기 위한 인공지능 뉴로모픽 칩을 새로운 성장 동력으로 판단하고 대규모 연구개발 투자를 지속하고 있다. 그림 2는 이와 같은 인공지능 반도체 기술의 시장 동향을 나타낸다. 다만 아직 인공지능 뉴로모픽 칩에 대한 완성도는 높지 않으므로 이를 활용한 시장은 크게 열리지 않았으나, 저전력 고효율 인공지능 데이터 처리 능력에 관한 잠재력을 바탕으로 향후 사물인터넷을 포함한 다양한 엣지 컴퓨팅 시장에서 인공지능 기능을 수행할 것으로 판단된다.

대표적인 저전력 단말 시장 중 하나인 스마트폰 분야에서는 이미 인공지능 기술을 이용한 컴퓨터 비전 및 언어 처리 분야에서 활용하고 있다. 이를 위해 각 제조사들은 수년 전부터 인공지능 프로세서 유닛인 NPU를 개발하여 자사의 스마트폰에 탑재하고 있다. 스마트폰용 인공지능 프로세서 개발에는 퀄컴이 가장 앞장서고 있었으나, 최근 애플, 삼성전자, 화웨이가 자사의 인공지능 전용 NPU 개발을 발표하면서 관련 시장에서의 경쟁이 치열해지고 있다.

퀄컴은 2013년 인간의 신경망 네트워크를 모방하여 학습하는 인공지능 프로세서인 제로스(Zeroth)를 공개하였으며, 이를 기반으로 스마트폰 AP에 NPU 개념을 접목하여 스마트폰 스스로 판단하고 유용한 정보를 찾아 사용자에게 알려주도록 하였다[1]. 2016년 5월에 NPU가 탑재된 스냅드래곤(Snapdragon) 820의 발표를 시작으로 845, 855 AP를 통해 인공지능 관련 기술력을 더욱 발전시켰으며, 이를 통해 정확한 판단과 빠른 조작이 가능한 개인 비서(Personal Assistance) 서비스를 제공하고 있다. 2019년 말에는 스냅드래곤 865를 발표하면서 더욱 강력해진 인공지능 기능을 강조하였다[2].

애플은 자체 개발한 A11 바이오닉 AP를 통해 음성 인식 플랫폼 시리(Siri)와 안면인증 기술인 페이스 ID 기능을 선보였으며, 2019년 하반기에 발표한 A13 바이오닉을 통해 고성능 코어 성능 향상과 낮은 전력 사용량을 특징으로 소개하였다[3]. 더불어 미국 시애틀에 본사를 둔 저전력 엣지 기반의 인공지능 전문 업체인 엑스노.ai(Xnor.ai)를 2억 달러에 인수하면서, 애플은 사람과 동물, 물건을 감별하는 인공지능 센서 기술을 확보하며 인공지능 기술 확보를 지속적으로 추진 중이다.

삼성전자는 엑시노스(Exynos)라는 자체 AP를 보유하고 있으며, 지난해 선보인 엑시노스 9820부터 NPU 탑재를 통해 인공지능 성능을 기존 제품 대비 7배 이상 개선하면서 장기적인 뉴로모픽 반도체 개발을 목표로 향후 2030년까지 NPU 분야 전문 인력을 2,000명까지 확대하여 인공지능 프로세서 기술 기반 확보 계획을 발표했다[4]. 더불어, 인공지능 반도체 분야의 글로벌 1위 기업으로 도약하기 위하여 관련 반도체 분야의 인수합병도 적극

적으로 추진할 것을 발표하며 경쟁력을 강화하고 있다.

화웨이는 2017년 모바일 전용 인공지능 칩셋인 기린(Kirin) 970부터 NPU 탑재를 시작하였으며, 2018년 독일 IFA 박람회에서 인공지능 기능을 대폭 강화한 기린 980을 공개하였다. 2019년에는 더욱 높은 성능의 기린 990을 공개함으로써 애플과 삼성전자 경쟁 구도의 스마트폰 시장에서 화웨이가 인공지능 프로세서 분야의 주도권 획득을 전략적으로 추진하는 상황이다[5].

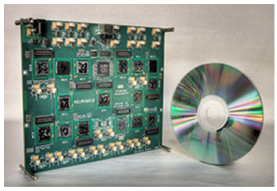
2. 뉴로모픽 반도체 연구 동향

인간의 뇌를 모방한 뉴로모픽 반도체 연구는 2000년대 중반부터 유럽과 미국 등에서 원천기술 확보를 목적으로 국가 주도 R&D 사업으로 시작되었다. 유럽연합은 2013년부터 Human Brain Project(HBP)라 불리는 인간의 뇌에 대한 대규모 원천연구를 시작하여 2023년까지 10년간 10억 유로를 투자하여 연구를 진행 중이며, 미국 역시 BRAIN Initiative 정책을 2013년부터 수립하여 인간의 뇌에 대한 심층적인 연구와 관련 원천기술 개발을 추진 중이다.

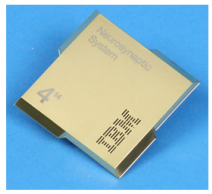
2009년에 스탠포드 대학에서 발표한 뉴로모픽 시스템인 Neurogrid에서는 기본 블록인 하나의 Neurocore 내에 256×256개의 아날로그/디지털 혼합 설계 방식의 뉴런들이 2차원 배열 형태로 구성되어 있고, 각 뉴런들은 블록 전체에 브로드캐스팅되는 버스를 통해 연결되어 있다[6]. Neurocore들은 상위 레벨에서 tree 네트워크로 연결되고 각 Neurocore 내부는 tree 네트워크를 위한 배선 스위치 및 테이블이 내장된다.

SpiNNaker는 스파이크 신경망을 실시간으로 모델링할 수 있도록 영국 맨체스터 대학을 중심으로 2012년에 개발된 대규모 병렬처리 뉴로모픽 슈퍼컴퓨터이다[7]. 130nm 공정으로 구현된 하나의 프로세서는 18개의 ARM968 코어들과 시냅스 가중치 저장을 위한 128MB SDRAM 및 주변 회로들이 NoC으로 연결된 CMP 구조를 가지는데, 각 ARM968 코어는 1,000개의 뉴런을 시뮬레이션 가능하며, 2차원 토러스(Torus) 타입 네트워크를 기반으로 최대 65,536개의 칩을 연결한 통합 시뮬레이션 수행이 가능하다. 최대 228 Dhrystone TIPS의 성능을 가지며, 518,400개의 프로세서로 확장될 수 있는 구조를 이용해 10억 개 뉴런의 시뮬레이션이 가능하다.

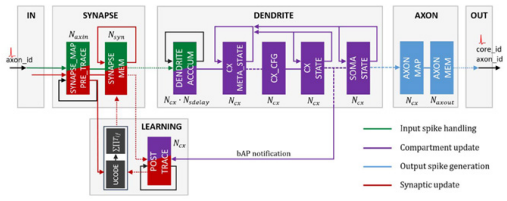
IBM은 2008년부터 미국 국방부 산하 DARPA가 주도하는 시냅스(SyNAPSE) 프로젝트에 참여하여, 2014년 TrueNorth라는 뉴로모픽 칩을 완성하였다[8]. TrueNorth 칩은 4,096개의 뉴로 시냅틱 코어로 구성되었으며, 하나의 코어에는 256개의 디지털



Stanford Univ. - Neurogrid



IBM - TrueNorth



Intel – Loihi (Architecture)

출처 https://en.wikipedia.org/wiki/File:NeuroGridBoard.jpeg, CC BY-SA 3.0

https://www.flickr.com/photos/ibm_research_zurich/26101819225 CC BY-ND 2.0

https://commons.wikimedia.org/wiki/File:Core_Top-Level_Microarchitecture.png CC BY-SA 4.0

그림 3 뉴로모픽 반도체 연구 현황[12,13]

I&F 뉴런이 256×256 크로스바(Crossbar) 네트워크를 통하여 256개의 SRAM기반의 입력 시냅스에 연결될 수 있도록 구성되었다[9]. 하나의 칩에는 64×64의 배열을 가지는 2차원 Mesh NoC로 연결된 총 4,096개의 코어가 구현되었다. PCB 보드상에서는 최대 16개의 칩이 4×4 메쉬(Mesh) 구조로 연결되도록 확장할 수 있어서, 100만 개의 뉴런과 2억 5천만 개의 시냅스가 이벤트 기반 방식으로 동작하여 초당 1,200~2,600frame의 이미지를 25mW에서 275mW 수준의 낮은 전력으로 분류할 수 있다. 이는 기존 마이크로 프로세서의 1/10,000 수준으로, 뉴로모픽 반도체의 저전력 가능성을 보여주었다.

BrainScaleS는 EU FET-Proactive FP7에서 지원하여 2011년에서 2015년까지 진행된 연구개발 프로젝트로서, 아날로그 회로로 설계된 뉴런/시냅스와 폰 노이만 구조의 petaflops급의 슈퍼컴퓨터 시스템이 혼합된 HMF 구조로 설계되었으며, 인간 뇌의 생물학적 메커니즘 연구를 목적으로 뉴런과 시냅스의 가소성 모델 에뮬레이션을 위하여 개발되었다[10]. 이후 98,304개의 시냅스와 384개의 뉴런이 내장된 analog digital mixed type 회로로 설계된 BrainScaleS-2 시스템으로 발전하여 에뮬레이션 규모를 크게 확장하였다.

DyNAPs에서는 각 코어 내의 I&F 방식의 아날로그 뉴런들과 아날로그/디지털 혼합으로 설계된 시냅스가 브로드캐스팅되는 버스를 통하여 tree 형태로 라우팅 스위치에 연결되어 있고, 각 코어들은 다음 단계의 쿼드-트리 구조와 최상위 수준에서의 2차원 메쉬 형태를 갖는 복합 계층적 구조의 NoC 구조로 연결되어 있다[11]. 특히 2단계의 태그(Tag) 필드 기반 라우팅 방식을 제안하여 배선 테이블의 크기를 획기적으로 줄이고 하나의 칩 내에 배선 테이블의 구현이 가능하게 함으로써 외부 메모리 통신으로 인한 성능 감소를 최소화하였다.

인텔은 2019년 온칩 학습이 가능한 스파이킹 뉴로모픽 칩인 로이히(Loihi)를 발표하였다[12]. 인텔 14nm CMOS 공정으로 개발된 하나의 로이히 칩에는 128개의 코어로 구성된 130만 개의 디지털 뉴런과 1억 3천만 개의 시냅스를 포함하고 있으며, 비동기 방식으로 동작한다. MNIST 기준 DNN 대비 100만 배 빠른 학습 및 실행 성능을 증명하였으며, 향후 인텔은 로이히를 활용하여 IoT용 경량 SoC뿐만 아니라 자동차, 로봇 산업 응용에 적용될 것으로 기대하고 있다.

3. 차세대 뉴로모픽 반도체

현재까지 개발된 대부분의 뉴로모픽 반도체는 기존의 실리콘 기반 CMOS 트랜지스터 기술만으로 구현되었다. 구현 소자 관점에 따라 이를 1세대 뉴로모픽 방식이라고 한다면, 차세대 뉴로모픽 소자로 구현되는 뉴로모픽 반도체는 2세대 방식이라고 할 수 있다. 1세대 뉴로모픽 반도체에서 시냅스는 기존의 CMOS 메모리 소자를 활용하여 시냅스의 가중치를 저장하였다가 읽어오는 방식으로 구현되었다. 2세대 뉴로모픽 반도체는 구현의 한 예로써, 생물학적 시냅스의 핵심적 특성을 높은 집적도를 위하여 하나의 소자로 구현할 수 있도록 메모리와 가변 레지스터의 두 가지 기능을 동시에 갖고 있는 멤리스터(Memristor) 소자를 활용하는 방식이 연구되고 있다. 멤리스터는 memory와 resistor의 합성어로 크기가 작아야 하고, 세분화된 가중치를 위하여 점진적인 스위칭 저항 특성을 갖는 것이 매우 중요하다. 2세대 뉴로모픽 반도체는 소자의 재료 및 구현 방식에 따라서 플래시 메모리(Flash Memory) 방식, RRAM(Resistive Random Access Memory) 방식, PRAM(Phase-change Random Access Memory) 방식, MRAM(Magnetic Random Access Memory) 등

이 있는데, 현재 멤리스터 방식의 연구가 가장 많이 연구되고 있으나 다양한 메모리 소자들을 대상으로도 활발한 연구가 진행되고 있다. 2세대 뉴로모픽 반도체에 대한 연구는 한국은 물론 미국, 유럽, 중국 등 전 세계적으로 활발하게 이루어지고 있으며, 현재까지 대부분 단위 기능 블록 수준에서 연구가 진행되어 왔지만 최근에는 시스템 수준에서의 구현 가능성을 테스트하는 방향으로 발전하고 있다.

IV. 결론

인공지능을 활용한 산업이 급성장하면서 인공지능 반도체 기술은 기술 도약을 위한 새로운 기회로 주목받고 있다. 인공지능 데이터를 고속 저전력으로 처리함과 동시에 시스템의 동작 효율을 높이기 위하여, GPU 대비 낮은 소비 전력으로 높은 병렬 연산 능력을 갖춘 FPGA나 낮은 소비 전력으로 인공지능 데이터를 처리할 수 있는 ASIC 기반 인공지능 가속기에 대한 관심이 높아지고 있다. 하지만 관련 기술들은 여전히 인간의 뇌 대비 높은 소비 전력 및 데이터 처리 비효율성 문제를 내포하고 있으며, 궁극적으로는 인간 뇌의 신경망 구조 및 작동 원리를 모방하여 만든 인공지능 뉴로모픽 칩 기술이 차세대 인공지능 프로세서로 각광 받을 것으로 전망되고 있다.

약어 정리

AER	Address Event Representation
AMS	Analog Mixed Signal
AP	Application Processor
ASIC	Application Specific Integrated Circuit
CMOS	Complementary Metal-Oxide Semiconductor
CMP	Chip Multi-Processor
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DNN	Deep Neural Networks
FPGA	Field-Programmable Gate Array
GPU	Graphics Processing Unit
HMF	Hybrid Multi-scale Computing Facility
I&F	Integrate and Fire
IoT	Internet of Things
MNIST	Modified National Institute of Standards and Technology
MRAM	Magnetic Random Access Memory
NoC	Network-on-Chip
NPU	Neural Processing Unit
PRAM	Phase-change Random Access Memory
RRAM	Resistive Random Access Memory
SNN	Spiking Neural Networks
SoC	System on Chip
SRAM	Static Random Access Memory
STDP	Spike Time Dependent Plasticity
TFLOPS	Terra FLoating-point Operations Per Second
TPU	Tensor Processing Unit
VPU	Vision Processing Unit

참고문헌

[1] https://www.qualcomm.com/news/onq/2013/10/10/introducing-qualcomm-zeroth-processors-brain-inspired-computing

[2] https://www.qualcomm.com/products/snapdragon-865-5g-mobile-platform

[3] https://www.samsung.com/semiconductor/minisite/exynos/products/mobileprocessor/exynos-980/

[4] https://www.macworld.com/article/3442716/inside-apples-a13-bionic-system-on-chip.html

[5] https://consumer.huawei.com/en/campaign/kirin-990-series/

[6] B. V. Benjamin et al., "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," Proc. IEEE, vol. 102, no. 5, May. 2014, pp. 699-716.

[7] E. Painkras et al., "SpiNNaker: A 1 W 18 core system-on-chip for massively-parallel neural network simulation," IEEE J. Solid-State Circuits, vol. 48, no. 8, Aug. 2013, pp. 1943-1953.

[8] F. Akopyan et al.,"TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," IEEE Trans. Comput.-Aided Desgin Integr. Circuits Syst., vol. 34, no. 10, Oct. 2015, pp. 1537-1557.

[9] L. Lapicque, "Recherches Quantitatives sur L`excitation E'lectrique des Nerfs Traite'e Comme une Polarization," J. Physiol. Pathol. Gen., vol.9, 1907, pp. 620-635.

[10] J. Schemmel et al., "Live demonstration: A scaled-down version of the BrainScaleS wafer-scale neuromorphic system," in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), Seoul, Rep. of Kroea, May 2012, doi: 10.1109/ISCAS.2012.6272131

[11] S. Moradi et al., "A Scalable Multicore Architecture with Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs)," IEEE Trans. Biomed. Circuits Syst., vol. 12, no. 1, Feb. 2018, pp. 106-122.

[12] M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," IEEE Micro, vol. 38, no. 1, Jan. 2018, pp. 82-99.

[13] https://en.wikipedia.org/wiki/File:NeuroGridBoard.jpeg

[14] https://www.flickr.com/photos/ibm_research_zurich/26101819225

[15] https://commons.wikimedia.org/wiki/File:Core_Top-Level_Microarchitecture.png