Chapter 01

저전력 인공지능 반도체 기술 동향

박성모_한국전자통신연구원 책임연구원
최병건_한국전자통신연구원 책임연구원
구본태_한국전자통신연구원 책임연구원
박경환_한국전자통신연구원 책임연구원
권영수_한국전자통신연구원 책임연구원

AI 산업이 확대됨에 따라 AI 반도체를 효과적으로 구현하기 위한 AI 반도체 기술 개발이 진행 중이며 AI 반도체가 산업의 핵심으로 떠오르고 있다. 특히, 뉴로모픽(Neuromorphic) 기술은 드론, 인텔리전스 엣지, IoT 디바이스, 웨어러블 디바이스, 자율주행차, 로봇, 모바일 단말을 중심으로 한 인공지능 반도체 시장의 급성장으로 성능과 저전력 기술의 중심으로 괄목하게 성장을 하고 있으나 아직은 초기 단계이다. 본 고에서는 저전력 인공지능 반도체 기술을 중심으로, 저전력 기술인 뉴로모픽 기술, 스파이킹 뉴럴 네트워크(Spiking Neural Networks: SNN), 프로세싱-인-메모리(Processing-In-Memory: PIM) 기술, 뉴로모픽을 위한 차세대 디바이스인 멤리스터(Memristor) 및 스커미온(Skyrmion) 기술에 대해 언급하고자 한다. 인간의 뇌처럼 저전력으로 동작하는 뉴로모픽 SoC는 다양한 기술의 융합을 통해 개발될 것으로 예상된다.

I. 서론

전세계 기업, 연구기관 및 스타트업에서는 인간과 유사한 서비스, 또는 이상의 욕구에 의해서 이를 가능하게 하는 인공지능 칩인 뉴로모픽(Neuromorphic) 개발을 진행하고 있다. 많은 과학자와 연구자는 인간의 뇌와 유사한 기능을 하는 칩을 만드는 것을 생각하고 구현을 하고 있다. 뉴로모픽 칩은 시냅스와 뉴런으로 구성되어 있으며 이를 스파이크의 자극에 의해서 스스로 상황에 따라 유기적으로 조절하는 것을 목표로 하고 있다. 현재는 특정 응용 분야에 우수한 특성을 나타내고 있으나 인간의 뇌와 같은 저전력으로 효율적인 학습을 하는 단계

* 본 내용은 박성모 책임연구원(☎ 042-860-5203, smpark@etri.re.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

OPEN 공공누리 출처표시 상업용금지 공공저작물 자유이용허락

에 아직 많은 연구가 필요하다[1],[2]. 인간의 뇌는 뉴런과 시냅스의 조합으로 연결되어 있으며 약 100조 개의 연결을 만들 수 있는 복잡한 구조이다. 최근 뇌를 모방하는 인공지능 반도체인 뉴로모픽 하드웨어 개발이 가속화되고 있다. 기존의 폰노이만 구조의 한계를 극복하고 인간의 뇌의 저전력 구조를 모방하기 위해 뉴로모픽은 스파이킹 뉴럴 네트워크(Spiking Neural Networks: SNN)와 기존 메모리를 대체할 새로운 소자를 사용하여 인간의 뇌를 모방한다. 실시간 저전력 연산을 제공하므로 사용자가 자기 학습, 에너지 효율적인 장치를 구현할 수 있다. 인공지능, 인공지능 내장 사물인터넷(AIoT), 자율 주행 및 이를 적용하는 엣지 장치와 같은 저전력 소비로 인간과 유사한 추론을 확장하고 다양한 서비스가 가능하다. 지능형 반도체는 크게 지능화, 저전력, 고신뢰 기술로 구분을 한다[2]. 저전력 기술은 인공지능 연산을 가진 마이크로 컨트롤러 등에서의 핵심 IP로서 전력소모량을 최소화하기 위한 회로 설계 기술인 저전력 로직 아키텍처 및 IP, 온칩 오프칩에서 메모리 아키텍처 변경에 의한 저전력 차세대 메모리 아키텍처, 칩 간의 데이터 전송시 전력소모량을 최소로 하는 저전력 무선 통신 설계 기술, 및 공정 기술로 분류할 수 있다.

뇌는 정보를 처리하기 위해 신호를 전송 및 수신을 한다. 이때 각 시냅스의 가중치 값에 의해 뉴런에 전달되며, 정보의 종류에 따라 웨이트가 달라진다. 새로운 정보에 따른 시냅스 가소성의 특성을 이용하여 학습이 이루어지며, 정보에 따라 가중치의 강도가 적절하게 조정된다. 두뇌는 학습, 판단, 기억을 동시에 처리하며 이때 약 20W의 에너지를 소비하고 시냅스 작동을 위해 한 번에 소비되는 에너지는 10fJ의 에너지 효율을 갖는 것으로 알려져 있다[2].

본 고에서는 먼저 II장에서 스파이킹 뉴럴 네트워크 기반의 뉴로모픽 칩 및 메모리 중심의

[표 1] 지능형 반도체 기술 분류

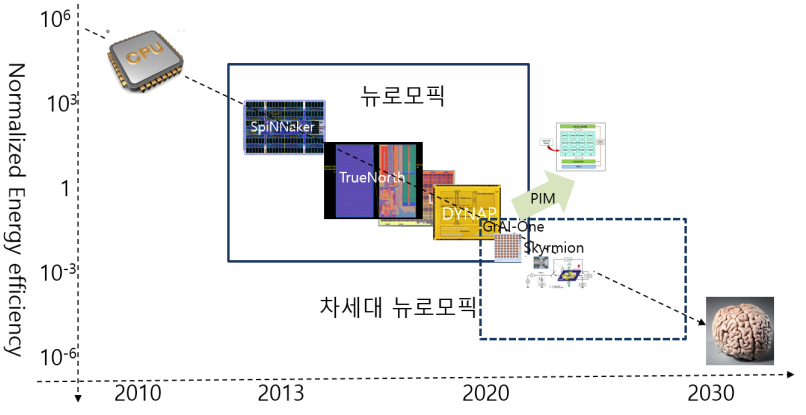
중분류	소분류	개념
저전력기술	저전력 로직 아키텍처 및 IP	인공지능 등에 널리 사용되는 핵심 IP로서 전력 소모량과 면적 최소화 회로 설계 기술
	저전력 차세대 메모리 아키텍처	메모리의 전력소모량을 줄이기 위한 대용량 메모리 집적, 동적전압 조절, 메모리 구조의 최적화를 통해 데이터 송수신을 최소화하는 저전력 메모리 기술
	저전력 무선 통신	칩 간의 데이터 전송 시 소모되는 전력소모량을 최소로 하는 저전력 통신 반도체
	저전력 소자 및 공정 기술	반도체의 전력소모량을 최소로 하기 위한 반도체 소자(Device)로서 트랜지스터 레벨의 저전력 기술

〈자료〉 하영욱, 김태완, "지능형반도체의 주요응용분야 시나리오와 핵심가치", ETRI, Insight Report, Feb., 2019. p.8.

뉴로모픽 칩인 SRAM 기반 프로세싱 인 메모리(Processing in Memory: PIM)에 대한 기술 동향을 살펴보고, III장에서는 맴리스트 반도체 소자기술 및 저전력 스커미온 소자기술을 살펴보고자 한다. 끝으로 IV장에서는 본 고의 결론을 제시한다.

II. 뉴로모픽 칩

현재 뉴로모픽 에너지 효율 측면에서 살펴보면 IBM에서 개발한 트루노스 칩은 CPU 대비 에너지 효율이 1/84000 정도로 나타나며, 이는 인간의 뇌와 비교하면 약 30만 배의 에너지 효율을 가진다[3]. [그림 1]은 에너지 효율 측면에서 CPU, 뉴로모픽, 차세대 뉴로모픽 및 인간 두뇌의 관계를 정규화하여 재구성한 것이다. 현재의 에너지 효율 측면에서 상당한 발전을 이루었지만 아직도 인간의 뇌와 비교를 하면 많은 연구가 필요하다. 앞으로 차세대 새로운 소자 기반의 뉴로모픽 칩의 개발을 통해 인간의 뇌와 비슷한 에너지 효율을 가질 것으로 예상이 되며 이는 2030년 이후가 될 것으로 기대된다.



〈자료〉 F. Akopyan et al., "TrueNorth: Design and tool flow of a 65mW 1 million neuron programmable neurosynaptic chip," IEEE Trans. Comput.-Aided Design Integr, Circuits Syst., Vol.34, No.10, 2015. pp.1537-1557.

[그림 1] 에너지 효율 측면에서 CPU, 뉴로모픽, 인간 두뇌의 관계

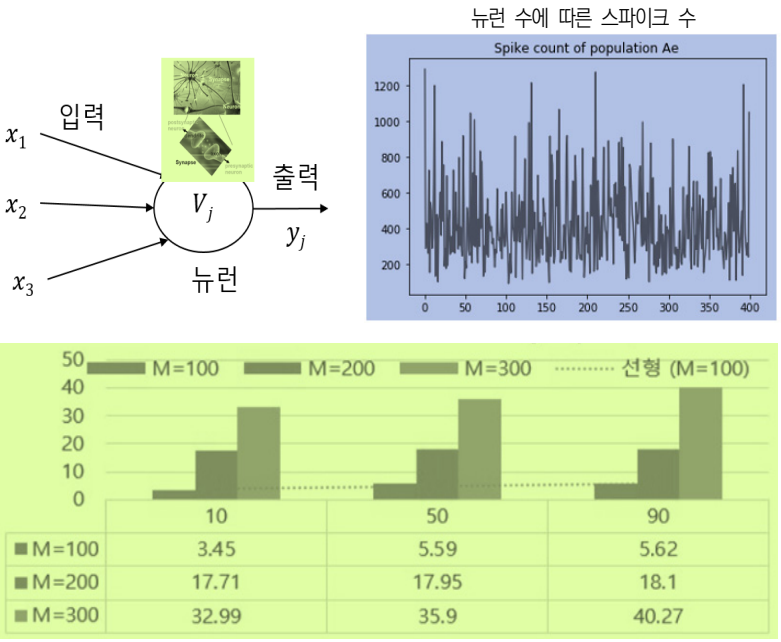
1. 저전력 스파이킹 뉴럴 네트워크

스파이킹 신경망(SNN)은 기존 컴퓨터 구조의 단점을 해결하기 위해 생물학적 신경 네트

워크 구조를 유사하게 모방하였다. 1세대에서는 다층 퍼셉트론 신경망을 위한 이진 입출력을 가지며 디지털 계산을 위한 매우 직관적인 해석과 최적화를 하였으나 아날로그 특성을 가지는 문제점이 있었다. 이런 1세대의 단점을 개선하기 위해 2세대에서는 디지털 계산이 가능하고 좀 더 생물학적인 접근 및 최근화 기법을 향상시켰으나 파라미터 값에 민감하게 동작하는 오류가 있었다. 이의 예제로는 합성 곱 신경망, 순환신경망 등이 있다. 제 3세대에서는 2세대의 문제점인 생물학적인 접근이 가능하고 집적회로 구현이 가능함으로써 연산의 대규모 병렬화가 가능하였으나 아직 확실한 이론이 없으며 다양한 모델과 코딩 체계에 대한 일관된 프레임워크가 부족한 단점을 가지고 있다[4].

2. 스파이킹 네트워크 시뮬레이터

뇌신경 모사를 위해서 스파이킹 네트워크를 모델링하고 시뮬레이션할 수 있는 대표적인 시뮬레이터가 브라이언[4] 시뮬레이터이다. 특히, 브라이언 시뮬레이터에서는 뉴런모델, 시



M=시냅스/뉴런 Ratio

〈자료〉 박성모 외, “스파이킹 신경망 기반 뉴로모픽 기술 동향”, TTA, 188호, 3월 2020년, pp.28-33.

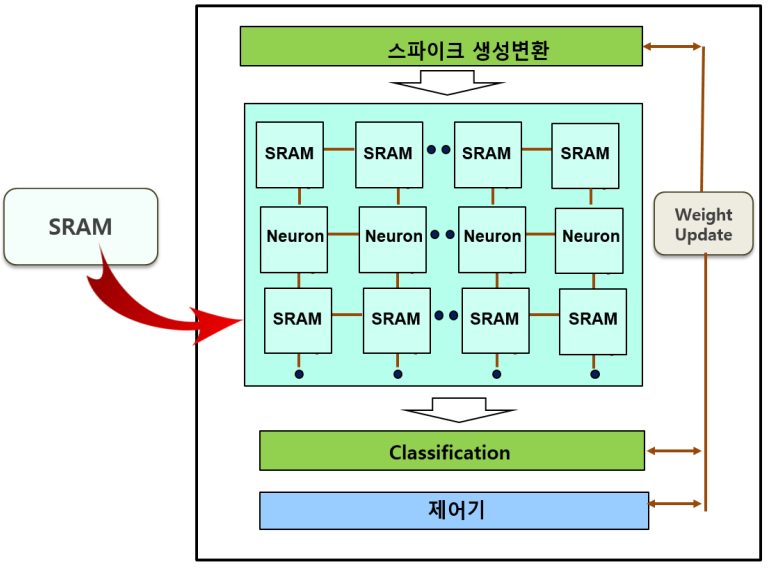
[그림 2] 스파이킹 뉴럴 네트워크 동작 및 시냅스/뉴런 Ratio 관계

냅스모델, 가소성, 통합모델 및 종합적인 플랫폼[5]을 제공한다. 이를 이용한 뉴로모픽 개발에서 상위수준 모델링 및 칩의 검증을 위한 개발 툴로 활용하고 있다[7].

3. SRAM 기반 프로세싱 인 메모리

프로세서와 메모리가 분리된 폰노이만 구조에서는 데이터의 양이 증가하면 데이터 병목현상이 필연적으로 발생하는데, 이러한 이유로 프로세서와 메모리를 통합하자는 PIM 구조는 이미 1970년대부터 알려져 있다[6]. 프로세서와 메모리가 분리된 구조의 폰노이만 구조는 사용자의 성능 요구량이 커지면서 로직과 메모리 사이의 데이터량에 의해 한계에 직면했다. 특히, 다양한 구조의 연구가 진행되고 있으나 그 중에 SRAM 기반의 PIM 구조에 대해서 언급을 하고자 한다. SRAM 메모리 밀도를 기존의 향상 한 트랜지스터 SRAM 셀 기술이 연구되고 있다.

[그림 3]은 SRAM 기반의 PIM 구조이다. 뉴런, SRAM, 스파이크 생성변환, Classification, 제어기 등의 모듈로 구성이 이루어져 있다. 스파이크 입력을 변환하는 생성변환 블록과 뉴런의 어레이에서는 뉴런의 가소성과 스파이크의 입력 및 입출력 뉴런의 시간에 따라서 맴브레



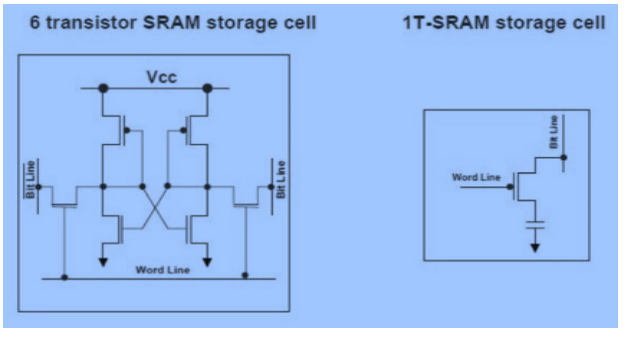
〈자료〉 한국전자통신연구원 자체 작성

[그림 3] SRAM 기반의 PIM 구조

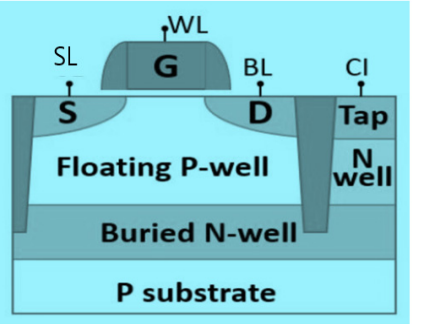
인 전압이 올라가거나 또는 감소되는 동작을 수행하고, 일정 전압 이상이 되면 스파이크가 발화를 하는 동작으로 학습을 하며, 학습이 완료된 후에 이러한 특성을 입력을 분류하는 동작을 수행한다. 웨이트 값은 정기적으로 업데이트하며 웨이트의 값은 메모리(SRAM)에 저장한다.

MoSys사(미국)는 1990년대 후반 고밀도 임베디드 메모리에 대한 수요 증가에 대응하기 위해서 고밀도 응용을 위한 1T-SRAM 기반 1세대 메모리 기술을 1998년에 발표하였다. 1T-SRAM 기술은 표준 CMOS 로직 프로세스를 기반으로 구축된 고밀도 임베디드 메모리 솔루션을 제공하며 기존의 6T 메모리보다 2배 더 작고 비용이 50~70% 낮은 1 트랜지스터, 1 커패시터 비트 셀로 전력 소비를 최대 75%까지 줄이는 비트 라인과 워드 라인의 특징을 가진다.

2001년에 발표된 1T-SRAM-M 기술은 극적으로 낮은 대기 전력 특징을 가진다. 1T-SRAM 아키텍처를 기반으로 하는 1T-SRAM-M은 완전한 데이터 보존과 함께 10μA/Mbit의 낮은 대기 전류를 달성하는 고유한 누설 억제 회로를 갖추고 있으며, 모바일 애플리케이션에 필요한 감소된 대기 전력 요구사항과 높은 수율, 밀도 및 성능을 결합하는 표준 로직 프로세스를 사용하였다. 모바일 애플리케이션에 대한 수요가 급격히 증가함에 따라 저전력 임베디드 메모리 솔루션에 대한 요구도 증가하였다. SoC 설계자는 메모리 크기가 증가하더라도 대기 모드에서 전력을 줄여 휴대폰, 핸드헬드 컴퓨터 및 기타 메모리 집약적인 모바일 기기에서 배터리 수명을 최대화할 수 있는 기능을 내장하였으며, 6T SRAM의 4배 밀도를



[50~70% less silicon area]



[1T SRAM 단면]

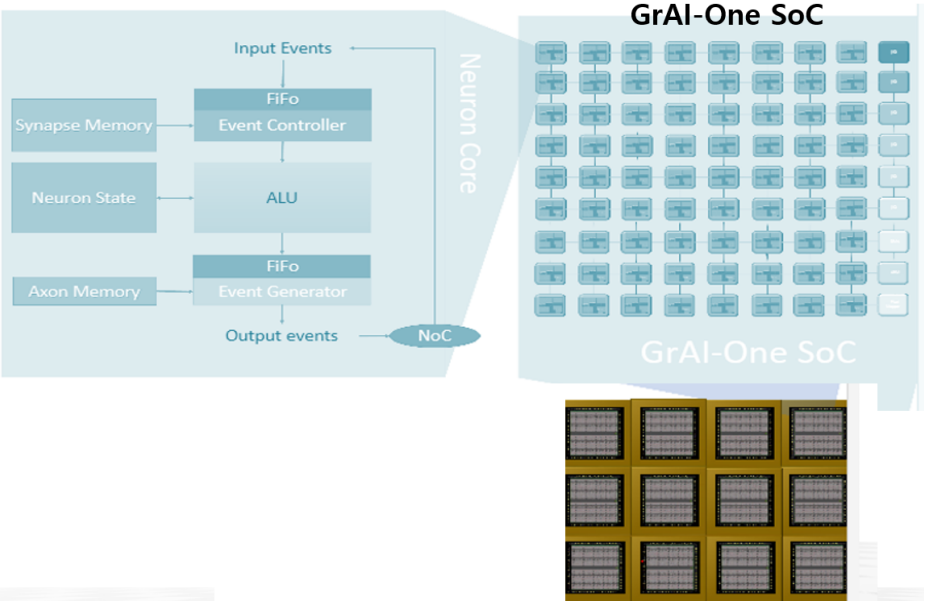
〈자료〉 Jin-Woo. Han, et. al., "A Novel Bi-stable 1-Transistor SRAM for High Density. Embedded Applications," IEDM Conference Procedings, 2015, pp.684-687.

[그림 4] 1T SRAM 회로도 및 단면도

달성하는 동시에 미래 나노미터 노드에 뛰어난 확장성을 제공하는 새로운 메모리 기술인 1T-SRAM-Q 메모리를 개발하였다. Folded Area CapacitorTM(FACTM) 기술이라고 하는 혁신적인 커패시터 기술을 사용하는 1T-SRAM-Q 메모리는 매우 높은 집적도의 임베디드 메모리 솔루션을 제공하며, 0.13μm에서 128Mb 이상, 90nm에서 256Mb 이상의 어레이를 통합하여 시스템 레벨 메모리 솔루션을 구현하였다([그림 4] 참조). 1T- SRAM-Q 메모리는 1T-SRAM 셀에서 수평이 아닌 수직으로 확장되는 새로운 고효율 커패시터 기술로 Shallow Trench Isolation(STI) 층에서 Well을 에칭하고 이를 폴리실리콘으로 채우기 위해 두 개의 추가 단계를 사용하여 1T-SRAM-Q의 접힌 영역 커패시터를 생성하며 면적을 최소화하는 셀을 구현하였다[11].

4. 저전력 GrAI 원 인공지능 반도체

GrAI 원칩은 TSMC 28nm 공정의 20mm^2 크기이며, 총 20만 개의 뉴런에 대해 로컬 뉴런·시냅스 메모리가 장착된 196개의 뉴런 코어로 되어있다[8]. Bio-inspired Edge 추론

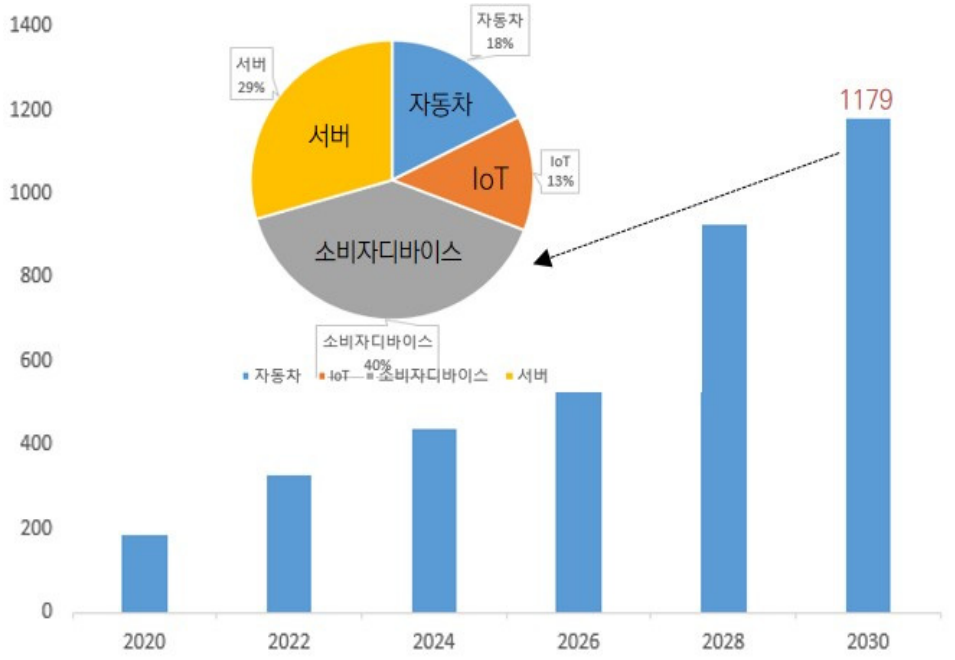


〈자료〉 Amirreza Y. et al., "The GrAI One Neuromorphic Many-core Processor," Event-based Asynchronuous Neuro-Congnitive Control Workshop, August 2019,

[그림 5] GrAI 인공지능 반도체

이 가능한 칩으로 이벤트 처리 당 10pJ 미만이어서 범용 및 유연한 프로그래밍을 위한 Dataflow 작업 지원, 인기있는 스파이킹 뉴런 모델 및 딥 뉴럴 네트워크 지원, 다양한 뉴런으로 확장 등이 가능하며, 모든 신경망 구성에 적합하다. 에지 추론, 시간적 희소성, 비동기 처리, 이벤트 중심 처리가 가능하며, 응용 프로그램은 비디오 처리, 자율 스티어링, 3D 심도 맵을 위한 스테레오 비전, 항상 사용하는 손 제스처 인식, 자율 드론, 오디오 처리, 항상 키워드 검색, 및 소음 제거 등의 다양한 분야에서의 사용이 가능하다.

인공지능 시장은 급격히 성장하고 있다. 서버, 자동차, AIoT 및 소비자 가전(CE) 등에서 인공지능 컴퓨팅을 실현하는 주체는 인공지능 반도체로서 인공지능 반도체 시장은 향후 10년 간 6배 이상 성장할 것으로 예상된다. 컴퓨팅 반도체 시장의 대부분이 인공지능 반도체로 전환될 것으로 전망되어 미래의 컴퓨팅 반도체는 인공지능 반도체가 대부분을 차지할 것이다. 인공지능 반도체 시장규모는 2030년에 1,179억 달러로 예상되며, 시장에서 소비자 디바이스가 40%, 서버용이 29%를 차지할 것으로 예측되고 있다([그림 6] 참조).



〈자료〉 MarketsandMarkets, 2018 재구성

[그림 6] 인공지능 반도체 시장의 성장(단위: 억 달러)

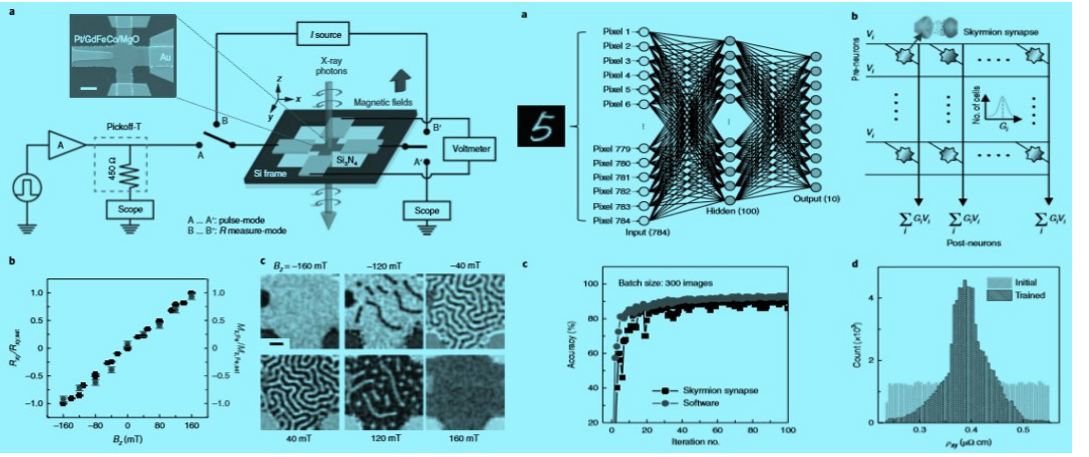
III. 저전력 차세대 메모리 소자

1. 뉴로모픽 소자

멤리스터는 메모리와 저항의 합성어로 저항 특성을 가진 소자의 저항 값이 일정하지 않고 양단에 인가되는 특정 전압 펄스에 따라 저항 값이 변하는 특징을 이용하여 인공지능 반도체의 뉴런 시냅스의 핵심 소자로 개발되고 있다[9]. 멤리스터 특성을 보여주는 대표적인 장치로는 ReRAM(Resistive RAM), PCRAM(Phase Change RAM) 및 FeRAM(Ferroelectric RAM)이 있다[9]. 현재 다양한 멤리스터가 개발되고 있으나 소자의 신뢰성 및 특성의 재현성이 아직 문제로 남아 있다.

2. 스커미온 기반 저전력 인공지능 반도체

스커미온은 각각 고유한 전기 저항을 가져 스커미온 개수에 따른 저항 변화를 아날로그적으로 조절하고 측정이 가능한 장점으로 인해 스커미온 기반의 인공 시냅스(Synaptic) 소자의 개발에 대한 관심이 높았으나 스커미온을 전기적으로 제어하는 기술적 어려움으로 인해 현재까지 이론적으로만 예측되었다. 페리 자성 다층에서 자기 스카이 러 미온의 축적 및



〈자료〉 Kyung Mee Song et al., "Skyrmion-based artificial synapses for neuromorphic computing," KIST, Nature Electronics, Vol.3, Mar., 2020, pp.148-155.

[그림 7] 스커미온 기반 인공지능 반도체

소산은 전기 펄스로 제어되어 시냅스 가중치의 변화를 나타낼 수 있다. 칩 수준 시뮬레이션을 사용하여 자기 스카이 러 미온을 기반으로 한 인공 시냅스가 패턴 인식과 같은 신경 형태 컴퓨팅 작업에 사용될 수 있음을 보여주었다. 손으로 쓴 패턴 데이터 세트의 경우 시스템은 89%까지의 인식 정확도를 달성하는데 이는 소프트웨어 기반의 이상적인 교육으로 달성한 정확도(~93%)와 비슷한 결과이며, 이 인공 시냅스 소자를 이용하여 손글씨 숫자 패턴(MNIST) 인식 학습을 진행한 결과, 90%의 높은 인식률을 증명했다[10].

IV. 결론

지금까지 저전력 기술을 중심으로 뉴로모픽 기술, 스파이킹 뉴럴 네트워크(SNN), 프로세싱-인-메모리(PIM) 기술, 뉴로모픽을 위한 차세대 디바이스인 멤리스트 및 스커미온에 기술에 대해 기술하였다.

인공지능 하드웨어에서 생물학적 행동을 모방하는 두뇌를 구현하려면 알고리즘, 구조, 회로 설계, 장치 및 플랫폼에 대한 종합적인 기술 개발이 필요하다. 스파이킹 신경망 기반 뉴로모픽 칩은 매우 낮은 전기적 신호와 높은 에너지 효율로 인해 향후 인공지능 반도체 시장의 급속한 성장에 기여하는 등 업계 전반에 큰 파급 효과를 미칠 것으로 예상된다. 인간의 뇌처럼 저전력으로 동작하는 뉴로모픽 SoC는 뉴로모픽 디바이스를 기반으로 SNN, PIM 등의 기술과 융합을 통해 개발될 것으로 예상된다.

참고문헌

[1] The Science Times, “사람 뇌 닮은 반도체칩, 뉴로모픽”, 10. June, 2019.
[2] 하영욱, 김태완 “지능형반도체의 주요응용분야 시나리오와 핵심가치”, ETRI, Insight Report, Feb. 2019. p.8.
[3] F. Akopyan et al., “TrueNorth: Design and tool flow of a 65mW 1 million neuron programmable neurosynaptic chip,” IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., Vol.34, No.10, 2015, pp.1537-1557.

* Acknowledgment
This research was supported by Nuclear Technology Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(No. 2018M2A8A1083094)

[4] 박성모 외, “스파이킹 신경망 기반 뉴로모픽 기술 동향”, TTA, 188호, 2020. 3, pp.28-33.
[5] CARLsim, http://www.socsci.uci.edu/~jkrichma/CARLsim/
[6] Harold S. Stone, “A Logic-in-Memory Computer,” IEEE Transactions on Computer, Vol.19, Jan. 1970, pp.73-78.
[7] Paul A. Merolla at al., “A million spiking-neuron integrated circuit with a scalable communication network and interface,” Science, Vol.345, 2014, pp.668-673.
[8] Amirreza Y. et al., “The GrAI One Neuromorphic Many-core Processor,” Event-based Asynchronuous Neuro-Congnitive Control Workshop, August 2019.
[9] S. G. Hu et al., “Review of nanostructured resistive switching memristor and its applications,” Nanoscience and Nanotechnology Letters, Vol.6, 2014, pp.729-757.
[10] Kyung Mee Song et al., “Skyrmion-based artificial synapses for neuromorphic computing,” KIST, Nature Electronics, Vol.3, Mar., 2020, pp.148-155.
[11] Jin-Woo. Han, et. al., “A Novel Bi-stable 1-Transistor SRAM for High Density. Embedded Applications,” IEDM Conference Procedings, 2015, pp.684-687.