지능형 반도체 신소자 기술 동향

김판길 가천대학교 나노물리학과 석사
배준호 가천대학교 나노물리학과 부교수

1. 머리말

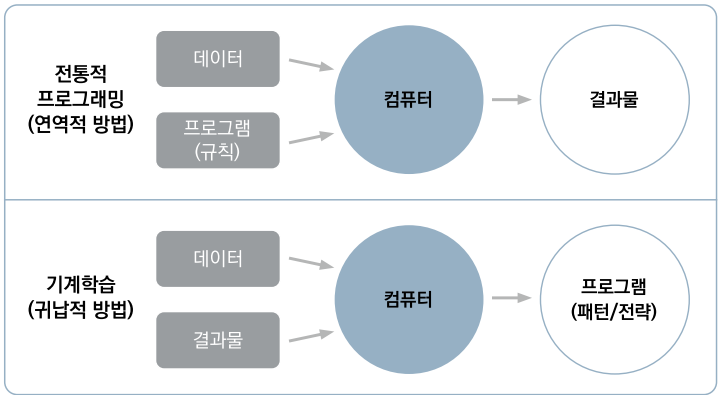
지능형 반도체는 통상적으로 인공지능 기능에 최적화된 소프트웨어와 시스템 반도체가 융합된 새로운 패러다임의 반도체를 말한다. 주지하다시피 인공지능의 발전으로 기존 컴퓨팅 패러다임이 사람이 짜놓은 규칙에 따라 데이터를 처리하는 방식에서 컴퓨터 스스로 학습하여 규칙을 습득하는 방식으로 변하게 되었다. 인공지능에 의해 컴퓨터가 스스로 학습하면 사람이 프로그램을 만든 것보다 더 좋은 결과를 얻게 되었다. 기존의 반도체 패러다임이 메모리 반도체와 비메모리 (시스템)반도체였다면, 지능형 반도체는 이를 뛰어넘어 단순한 반도체소자의 고성능 외에 초전력, 저가, 초연결성, 소자에 특화된 차별화 기능이 가능하게 되었다. 또한, 지능형 반도체에 융합된 인공지능에 의해 인간의 뇌처럼 대량의 기억과 연산을 동시에 처리하고 데이터를 학습하며, 판단(추론)이 이뤄지는 것이다.

※ 출처: 한국반도체 산업협회(2018b)

[그림 1] 지능형 반도체 구현을 위한 핵심 기술

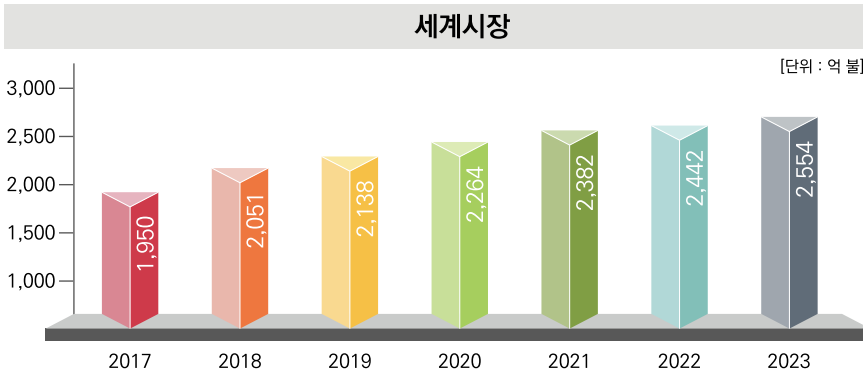
인공지능의 도래는 기존 반도체의 근본적인 한계를 드러냈다. 인공지능 연산은 학습과 추론의 2단계 과정으로 이뤄진다. 신경망의 최적 가중치를 찾는 학습(Training)은 많은 양의 데이터를 이용하여 신경망의 최적 가중치를 찾아가는 과정으로 신경망의 복잡도와 학습되는 데이터가 많을수록 분석 정확도가 높아지지만 요구되는 계산능력이 급속하게 높아진다(예: 2012년 이미지넷에서 우승한 AlexNet은 신경망이 8개, 2015년 우승한 ResNet은 신경망이 152개 층의 심층망을 가졌음). 신규 데이터를 적용하는 추론(Inference) 단계는 학습(Training)이 끝난 신경망에 신규 데이터를 적용하는 과정으로 실행(Run, execution)이라고도 표현한다. 이 단계는 신경망 최적화 수준에 따라 품질에 큰 차이가 발생한다. 학습은 데이터 센터에서 진행된 후 사용자 단말로 전달되었으나, 머지않아 사용자 단말에서 추론처리가 일어날 것으로 전망된다. 데이터를 입력 순서에 따라 순차적으로 처리하는 기존 반도체는 많은 데이터를 처리할 경우 중앙처리장치(CPU)와 메모리 사이에서 생기는 병목현상으로 인해 학습과 추론 수행 시 속도 저하와 전력의 한계가 존재한다.

관련 시장에서 지능형 반도체는 이미 급속한 성장세를 보이고 있다. 2018년 시장조사기관인 Marketsandmarkets은 지능형 반도체 세계 시장이 2018년 70.6억 달러에서 2025년 592.6억 달러 규모로 연평균 35.5% 성장할 것으로 전망했다. 같은 해



※ 출처: STEPI(2016)

[그림 2] 전통적 프로그래밍과 기계학습



※ 출처 : Gartner 2018, 2023년은 성장률을 고려한 추정치

[그림 3] 지능형 반도체의 시장규모 및 전망

Gartner는 세계 지능형 반도체 시장은 2017년 1,950억 달러에서 2023년 2,554억 달러 규모로 성장하여 전체 반도체 시장 중 70% 이상을 차지할 것으로 예측했다. 2015년 Marketsandmarkets은 한국의 지능형 반도체 시장의 세계 시장 비중이 2016년 15.8%에서 2022년에는 19.5%로 성장할 것으로 전망했다.

본고에서는 이른바 4차 산업혁명의 키워드인 빅데이터, 가상현실, 자율주행, 그리고 인공지능에 의해 부상한 지능형 반도체의 기술적 동향을, 전 세계 기업들의 개발동향과 제품을 중심으로 소개하고 향후 발전 방향을 전망해본다.

2. 지능형 반도체 기술동향

2.1 인공지능으로 인한 기존 반도체 패러다임의 한계와 해결

인공지능은 컴퓨팅 기술의 새로운 가능성을 제시하면서 기존 반도체 및 컴퓨팅의 한계를 야기하였는데 이는 크게 2가지 영역에서 고찰된다. 첫 번째는 현재 컴퓨터 구조, 즉 폰 노이만(von-Neumann) 구조의 데이터 병목 문제이다. 주기억 장치, 중앙 처리 장치, 입·출력 장치로 이어지는 직렬처리(Serial Processing)에서 최근 요구되는 고속 병렬 연산 수행 시 심각한 데이터 병목 현상(특정 부분에 사용량이 많아 일부분의 성능 저하로 이어져 전체적인 시스템이 마비되는 현상)이 발생한다. 하나의 중앙연산장치(CPU)가 중앙에서 모든 데이터를 처리 및 제어하여 연산량이 많을수록 메모리와 CPU 사이의 병목 현상이 심각해진다. 두 번째는 무어의 법칙(Moore's Law)으로 대변되는 반도체 집적도의 물리적 한계를 들 수 있다. 이전까지는 무어의 법칙이 얘기하는 것처럼 반도체 집적도를 높여 데이터 처리 속도를 증가시킬 수는 있으나, 고도화된 미세공정으로 인해 발열 및 간섭 등 물리적 한계에 봉착하게 되었다.

이러한 문제점들을 해결하기 위해 대규모 데이터의 고속처리가 필요하고, 기존 CPU에 특정 데이터 연산에 특화된 가속프로세서를 추가하는 이(異) 기종 시스템 구조가 대규모 데이터 고속처리의 해결책으로 제시된다. CPU는 기억, 해석, 연산, 제어를 하는 메인 역할을 담당한다. 이런 CPU가 특정 연산에 특화된 GPU(그래픽 처리장치) 및 다른 장치들과 함께 메모리를 공유하여 하나의 연산 장치처럼 작동한다. 다른 기종 시스템 구조는 메모리와 프로세서의 병렬 연결 확대를 위해서 3D HBM(High Bandwidth Memory)의 구조로 진화할 전망이다. 3D HBM 구조는 메모리와 GPU 사이의 수많은 병렬 연결선을 확보하기 위해 2차원 구조에서 3차원 구조로 진화하고, 전자파 도파관 구조로 된 연결선에 의해 1개 연결선당 100Gbps 이상을 확보하게 된다.

2.2 인공신경망 전용 반도체

기존 반도체의 한계를 극복하고 인공지능의 핵심인 인공신경망을 이식해 학습과 추론을 수행하는 인공신경망 반도체가 개발되고 있다. 인공신경망 반도체는 크게 소프트웨어 기반과 하드웨어 기반으로 나뉘지며 각각 다음과 같은 특징이 있다.

2.2.1 소프트웨어 기반 인공신경망 전용 반도체

인공신경망 모사 소프트웨어를 더 효율적으로 처리할 수 있도록 대규모 병렬 컴퓨팅과 연산이 가능한 인공지능 전용 반도체이다. 현재까지 소프트웨어를 통해서 인공신경망을 구축하고 이를 기존 CPU와 GPU와 같은 일반 반도체를 사용한 컴퓨터를 이용해 연산하는 것이 대세였다. 그러나, 독립적인 기억장치 인터페이스로 인해 발생하는 병목현상으로 인해 많은 전력소모, 저속 동작이라는 문제점이 발생해 학습을 보다 효율적으로 진행하기 위해 특수한 목적을 가진 인공지능 전용 반도체인 Neural Processing

Unit(NPU)과 Tensor Processing Unit(TPU)이 2015년부터 등장하였다.

인공지능 전용 반도체는 일반적인 메모리나 폭넓은 데이터 전송 대역폭을 가진 메모리를 사용하며 메모리에 연산기를 집적하는 구조로 이용될 것으로 예상된다. 프로그램 기능을 최대화하기 위해 인공지능 가속 반도체와 초병렬 프로세서를 이용하여 각양각색인 인공지능 알고리즘의 병렬 구현이 가능한 반도체 기술이다.

초고성능의 병렬 프로세서 시장 안정화 및 차세대 메모리 기술의 출현에 따른 프로세서와 메모리 통합 거대 병렬 컴퓨팅 시장이 성장하고 있다. 또한 데이터 처리 대역폭과 에너지효율을 증가시킬 기술로 메모리 내부와 주변에 연산을 포함하는 로직을 추가하는 PIM에 대한 연구가 진행되고 있다. 인공지능 알고리즘 중 하나로서 딥러닝 알고리즘을 구현하는 심층 연결망(DNN, Deep Neural Network)은 외부 메모리에서 많은 양의 신경망 데이터를 읽어 특징을 추출하여 처리하는 데이터 집약적인 특성을 지녀 메모리, 컴퓨팅 자원, 에너지를 절약하는 기술이 매우 중요하다.

2.2.2 하드웨어 기반 인공신경망 전용 반도체

인공신경망을 하드웨어에 직접 구현하는 새로운 개념의 반도체이다. 소프트웨어 기반의 인공지능 시스템은 뇌의 시냅스와 뉴런의 기능을 수식적으로 정의하고 코딩하여 기존 폰노이만 구조의 컴퓨터를 통해 연산한다. 그렇기 때문에 전용 인공지능 반도체를 채용한다고 하더라도 궁극적으로는 컴퓨터 가격, 학습시간, 소비전력 면에서 한계를 보일 것이다. 이러한 한계를 해결하기 위해, 뇌의 시냅스와 뉴런의 기능을 모방한 하드웨어 기반 인공지능 시스템 기술이 부각된다.

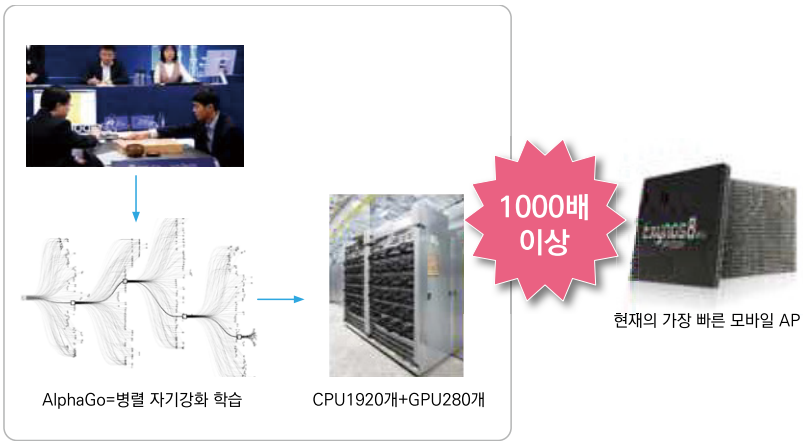
인공지능 하드웨어의 기본 구성요소(building block)는 뉴런과 시냅스를 모방한 수학적 연산을 하는 소자와 회로이다. 이 구성요소에 뉴런의 막 전위가 문턱 전압보다 높을 때만 시냅스 간의 정보를 전달하는 방식을 추가하여 저전력 동작이 가능한 스파이킹 인공 신경망(SNN, Spiking Neural Network)을 구현하였다. 이를 통한 새로운 컴퓨터 아키텍처와 구현 방식으로 컴퓨팅 기술의 패러다임이 변할 것으로 예상된다.

하드웨어 기반 인공신경망 전용 반도체는 새로운 형태의 메모리 소자나 생물학적 시냅스를 모방하는 소자에서 창의적인 아이디어를 추가하여 사용하는 소자가 필요하며 각종 형태의 소자, 회로, 아키텍처 및 알고리즘이 하나의 시스템으로 합쳐지는 형태로 발전된다. 현재 기술은 CMOS 기반의 전원을 공급하는 한 저장된 데이터가 보존되는 램(SRAM, Static Random access memory), 부동 게이트(floating gate)를 시냅스로 활용하는 신경 세포 모방 회로기술과 이전의 상태를 모두 기억하는 메모리 소자인 멤리스터, 전자의 회전(spin) 방향을 이용하는 스핀트로닉스 등 최근에 만들어진 소자와 이를 뉴런이나 시냅스로 사용하는 기술(신경 세포 모방 소자 기술)로 구분 할 수 있다.

2.3 지능형 반도체의 목표와 전개방향

지능형 반도체 기술이 최종적으로 달성할 목표는 가격, 성능과 전력문제를 해결하기 위한 기술 개발을 통해 '모든 ICT 기기에 인공지능을 부여하는' 사회로의 발전을 위한 기반 기술을 실현하는 것이라고 할 수 있다. 구체적으로 살펴보면 다음과 같다.

- 4차 산업혁명의 핵심기술인 인공지능 반도체와 인공지능 소프트웨어로 정보기기화 혁명(IDX, Intelligent Digital Transformation)이 나타나는 것을 의미한다.
- 현재 구현된 인공지능 서비스는 빅데이터 서버와 서비스 클라이언트가 상호 연결된 상태로 간단한 서비스를 이용할 때에도 많은 에너지가 소모된다.
- 작은 반도체 내부에서 완전한 인공지능 구현 및 모바일과 내장형 환경에도 적용할 수 있는 지능형 반도체의 기술 개발로 4차 산업혁명에 의한 사회 구현이 가능해진다.



※ 출처: 중소기업벤처기업부 2019-2021 중소기업 전략기술로드맵

[그림 4] 인공지능 서비스가 요구하는 성능 및 전력 소모량

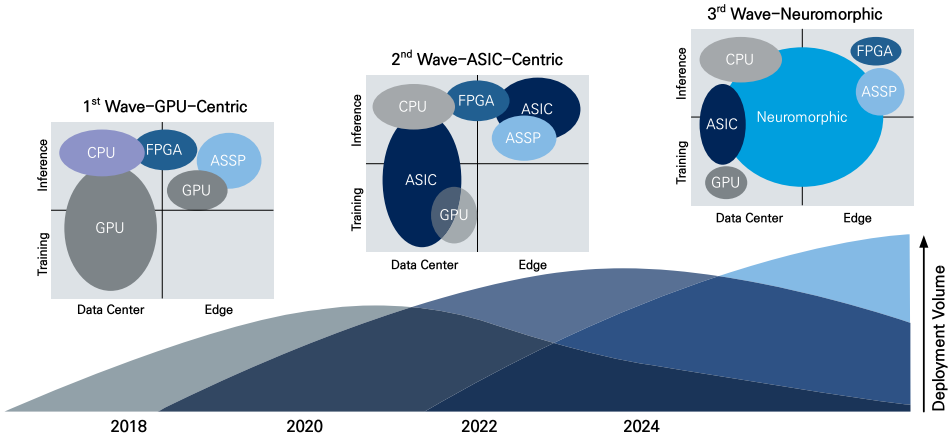
지능형 반도체의 발전은 크게 3단계로 전개된다.

- **1단계**: 현재의 인공지능 반도체는 GPU 중심으로 데이터 센터 및 에지 디바이스[1] 에 탑재되어 주로 활용.
- **2단계**: 주문형 반도체(ASIC, Application Specific Integrated Circuit) 방식으로 저전력 및 고성능인 반도체가 인공지능을 지원하는 형태로 변화.
- **3단계**: 초저전력·초고성능의 뉴모로픽 반도체로 추론과 학습, 데이터 센터와 장치 등 인공지능 시스템의 많은 기능이 지원 가능한 방향으로 발전.

2.4 세계 지능형 반도체 제품군

2.4.1 엔비디아(NVIDIA)

인공지능의 발전으로 GPU(Graphics Processing Unit)가 각광을 받고 있다. 직렬(순차) 처리 방식인 CPU는 한 개의 일을 수행하기 위해서 여러 단계를 거쳐야 한다. 그러나 여러 일을 동시에 처리 가능한



※ 출처: Gartner(2018), 정보통신정책연구원(2018) 재인용

[그림 5] 반도체 종류에 따른 사용 분야 전망

1) 에지(edge) 디바이스: 클라우드 컴퓨팅과 같이 데이터센터에서 집중하여 데이터를 처리하는 방식이 아닌, 단말기에서 개방형 아키텍처를 사용하여 데이터를 처리하는 방식의 디바이스

〈표 1〉 지능형 반도체 종류에 따른 장점과 단점

유형	GPU	FPGA	ASIC
장점	• 병렬처리에 최적화된 프로세서로, CPU에 비해 빠른 가속 성능 • NVIDIA社의 CUDA등 개발자 환경이 잘 갖춰져 있으며, 적용 사례가 많아 지원받기 용이	• ASIC보다 초기 개발 비용이 저렴 • CPU와 병렬 작동이 용이하여 전체 시스템 병목현상 발생 없음 • 회로 재구성이 가능. 개발 중인 AI 알고리즘을 유연하게 적용 가능 *(예) A라는 업무에 최적화하여 사용하다 반도체 회로 구성을 다시 설정(재프로그래밍)하여 B라는 업무에 맞춰 사용 가능	• GPU, FPGA 대비 매우 빠른 속도와 우수한 전력효율
단점	• FPGA, ASIC 대비 낮은 전력 효율 • 기존 x86 시스템(CPU)에 추가 구축시, 확장성과 호환성에 한계 *(예) 데이터 전송 병목문제, 시스템 호환 문제 등	• ASI보다 연산속도가 느리고 CPU나 GPU 같은 범용 프로세서 대비 프로그래밍 전문성을 요함	• 매우 비싼 초기 제작비용, 장시간의 개발소용 시간 • 특정 연산에 최적화 되었기 때문에 응용분야가 한정

※ 출처 : 한국전자통신연구원(2017)

병렬 처리 구조(Parallel Processing)를 가진 GPU는 방대한 데이터를 연산하기 위한 반도체로 주목받고 있다. GPU는 연산을 하는 ALU(Arithmetic Logic Unit)의 코어가 수천 개로 구성되어 있지만 CPU는 소규모로만 구성되어 있어 일부 GPU의 성능을 낼 수 있지만 연산 능력차이가 매우 크게 발생한다.

그래픽카드 지포스로 익숙한 기업인 엔비디아는 PC용 그래픽 반도체인 GPU(Graphic Processing Unit) 분야에서 독보적인 세계 1위 기업이다. DNN 기술을 도입한 지 3년 만에 그래픽스 처리 프로세서(GPU)의 장점을 이용하여 병렬 작업 기술을 최적화시켜 훈련 속도를 50배 증가시켰다. 또한, 범용성 컴퓨팅 폼팩터를 위한 GPU를 토대로 첨단 운전자 보조 시스템(ADAS, Advanced Driver Assistance System) 프레임워크(framework)를 개발하고 있으며, 쉐이더(shader) 프로그램이 도입된 이후의 가장 큰 기술 발전으로 꼽히는 RTX 기술은 GPU 아키텍처, 알고리즘, 딥 러닝을 자사만의 방법으로 구현하였다.

2016년 GTC(GPU Technology Conference)에서 데이터센터에 최적화된 테슬라(Tesla) P100 GPU를 발표, 1년 뒤인 2017년에는 볼타 기반의 최초의 프로세서인 엔비디아 테슬라(NVIDIA Tesla) V100 데이터센터 GPU가 공개되었다. 볼타의 성능은 2014년에 발표된 GPU 아키텍처인 파스칼(Pascal) 대비 5배, 2년 전 출시된 맥스웰 대비 15배 향상되었고 이는 무어의 법칙으로 예측된 수준을 4배 가량 넘어선 성능 개선으로 일반 CPU 100대 정도의 성능으로 딥러닝을 구현할 수 있다. 테슬라 V100은 클라우드 서버(데이터센터)용으로 Microsoft, 바이두, IBM 등의 회사에 사용되었다.

2.4.2 구글(Google)

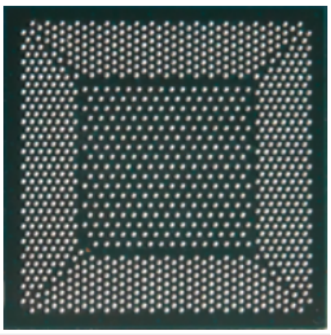
소프트웨어 및 빅데이터 센터를 토대로 인터넷 서비스를 메인으로 하는 기업이었지만 빅데이터 센터에 사용되는 서버에 인공지능 등 새로운 서비스를 추가할 필요성이 증가되어 병렬 프로세서 반도체를 개발하기 시작하였다. 2016년 5월 GPGPU에 비해서 벡터와 행렬연산의 병렬처리에 특화된 TPU(Tensor Processing Unit)를 공개하였다. 2017년 학습에도 사용 가능한 TPU를 개선하여 TPU2를 개발, 45테라플롭스(TFLOPS)짜리 성능을 내는 TPU칩 4개를 탑재한 Cloud TPU 개발과 이를 최대 64개 연결하여 제작하는 TPU Pod를 발표하였다.

2.4.3 인텔(Intel)

대형, 고속의 데이터센터 서버용 프로세서 개발을

중심으로 다양한 프로세서 제품들을 생산하면서 소비자 기기용 AI 반도체 영역에도 상당한 투자를 하고 있다. 또한 DNN 가속 기술을 보유한 업체들과 협력 및 인수 합병을 진행하고 있다. 2015년 FPGA 선두 제조업체인 알테라를 인수한 후 시스템 설계자가 원하는 대로 칩 구성이 가능한 FPGA를 이용한 자동차, IoT, 인공지능과 같이 용도에 따른 프로세서를 제공하고 있다. 2016년에는 인수한 인공지능 반도체 너바나 시스템즈(Nervana Systems)의 기술을 응용하여 2017년에 AI 프로젝트와 관련된 딥러닝 전용 NNNP(Nervana Neural Network Processor) 칩 개발을 발표했다.

인텔은 지능형 반도체인 뉴로모픽 칩을 개발하고 있다. 2017년에는 인간의 뇌신경세포를 실리콘에 구현하여 스스로 학습하는 뉴로모픽 칩인 로이히(Loihi)를 개발하였다. 총 13만 개의 뉴런과 1억3000만 개 시냅스로 구성된 0.47mm^2 크기의 코어 128개로 로이히를 제작하였고, 반도체 집적도는 14nm(나노미터·1nm는 10억분의 1m) 수준으로 최고 수준의 기술력을 요구하는 뉴로모픽 반도체이다.



※ 출처: https://en.wikichip.org/wiki/intel/loihi

[그림 6] 패키징된 인텔의 뉴로모픽 칩 로이히(Loihi)

2.4.4 퀄컴

2013년에 세계 최초로 뇌의 신경세포처럼 스파이크 형태의 신호를 주고받으며 시냅스 연결 강도를 조절하여 정보를 처리하는 프로세서인 제로스(Zeroth)를 개발했다. 그러나 DSP(Digital Signal Processor) 코어 중심의 뉴럴넷 컴퓨팅 환경을 제공하는 형태로 개발 방향을 전환하였다. 스냅드래곤 820에 내장된 제로스(Zeroth)는 기계 학습에 사용되었고, 2016년에는 제로스 플랫폼을 지원하는 스냅드래곤 뉴럴 프로세싱 엔진용 디자인 플랫폼을 개발하여 배포하였다. 2017년 12월 스냅드래곤 845를 발표하였고 기계 학습과 인공 지능을 위해 만들어진 코프로세서(coprocessor) 기관이 진행하는 '인공두뇌 만들기 프로젝트'에 참가하여 트루노스(TrueNorth)라는 뉴로모픽칩 개발에 성공하였다. 트루노스 칩은 약 54억 개의 트랜지스터를 내장한 4,096개의 프로세서로 구성되어 있고 전자회로 소자들이 인간 두뇌의 신경망처럼 연결되어 있어서 인간 두뇌 활동을 모방한다.

2.4.5 IBM

2008년에 미국 국방부 산하 방위고등연구계획국(DARPA)이 주도하는 '인공두뇌 만들기 프로젝트'에 참여하여 트루노스라는 뉴로모픽 반도체를 2014년 8월에 만드는 데 성공하였다. 트루노스는 54억 개 트랜지스터를 내장한 4,096개의 프로세서로 이루어져 전자 회로 소자들을 인간의 신경망처럼 연결해 인간 두뇌 활동을 모방한다.

100만 개의 뉴런과 2억 5000만 개의 시냅스로 얽혀 초당 1,200프레임에서 2,600프레임으로 이미지를 분류 소비하는 전력은 25mW에서 275mW 수준으로 매우 적은 전력을 소비하며 이는 기존 마이크로프로세서의 1만분의 1의 전력을 소요한다.

2.4.6 SK 하이닉스, 네패스(NEPES), 바이두 및 삼성전자

SK 하이닉스는 2016년 10월 미국 스탠퍼드 대학과 강유전체(Ferroelectrics) 물질을 활용한 '인공신경망 반도체 소자 공동 연구개발' 협약을 체결하고, 전하 유입 여부를 통해 0과 1을 구분하는 기존의 방식 대

신 전압 크기로 다양한 신호를 저장할 수 있는 유기 물질인 강유전체를 사용하여 뉴로모픽 반도체 개발을 추진 중이다.

1990년 설립된 네패스는 에지(edge) 디바이스용 인공지능 반도체인 NM500을 상용화를 위해 미국의 반도체 설계 업체인 제너럴비전사와 협업을 하였다. NM500은 0.4mm 반도체에 576개의 인공뉴런을 집적하여 고속·병렬연산 처리를 수행한다.

중국의 바이두는 2018년 AI 개발자 컨퍼런스(BaiduCreate 2018) 행사에서 AI 연산용 ASIC '쿤룬(Kunlun)'을 공개하였으며, 쿤룬은 초당 512GB 데이터를 주고받아 무인자동차부터 데이터센터까지 모든 곳에 사용 가능할 정도의 수준으로 260 TFLOPS 급의 성능이다.

삼성전자는 초고속 모뎀을 탑재하면서도 인공지능(AI) 연산 기능을 증가시켜 고성능 모바일 애플리케이션 프로세서(AP) '엑시노스9(9810)'를 개발하였다. 삼성전자와 바이두의 협력으로 2020년 초에 14나노 AI칩 쿤룬(KUNRUN)이 생산 예정이라고 한다.

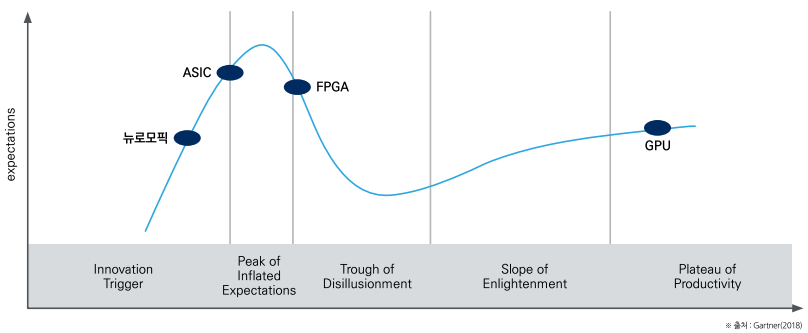
2.4.7 Neo^{2}C

2021년까지 자가 학습이 가능한 뉴로모픽 반도체 'Neo^{2}C' 개발을 목표로 KIST, KAIST, 서울대, 포스텍, 울산과학기술원(UNIST), 국민대, 어바인 캘리포니아대의 7개 기관 외의 연구단을 구성하여 연구 추진 중이다. Neo^{2}C의 반도체 집적도는 55nm, 소모전력은 56mW 수준으로 코어당 뉴런 수는 1,024개로 '로이히'와 유사하지만 로이히는 128개 코어이며 Neo^{2}C는 단일코어라는 차이점이 있다.

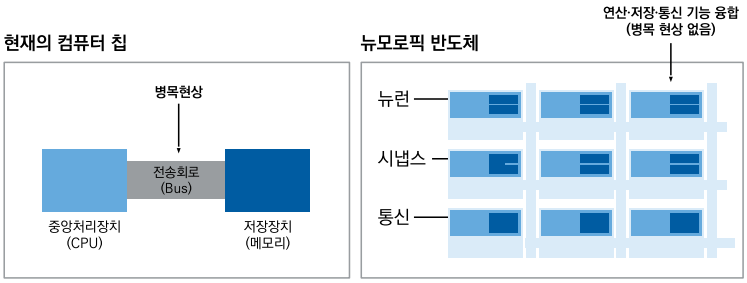
3. 맺음말

지금까지 반도체 소자의 새로운 패러다임인 지능형 반도체 소자의 개념과 기술 동향을 세계 기업들의 제품을 중심으로 살펴보았다. 지능형 반도체 소자는 초고성능, 초저전력을 중심으로 빠르게 진화하고 있다. 전 세계 산학연 기관들이 수많은 기술을 선보이고 있으며, Gartner社(2018)의 'Technology Hype Cycle'에 따르면 지능형 반도체 기술은 성숙도가 GPU 〉 FPGA 〉 ASIC 〉 뉴로모픽 반도체 순으로 높게 나타난다([그림 7]).

이러한 관측에서 현재 기술적 성숙도와 상용화 정도에서 GPU 기술이 시장을 선도하고 있으나, 2장에서 살펴보았듯이 현재는 개발 초기단계에 있는 뉴로모픽이 지능형 반도체 소자를 주도해 나갈 것으로 전망된다. 기존 컴퓨터인 폰 노이만 구조는 순차적 직렬적인 컴퓨팅으로서 데이터가 많을수록 처리시간과 전력소모가 급속히 증가한다. 반면, 뉴로모픽 소자에 있는 여러 개의 코어는 뇌의 신경세포인 뉴런 역할을 하며, 인간 뇌의 특징인 저전력 소모가 가능하고, 학



[그림 7] 인공지능 반도체 Technology Hype Cycle



	기존 반도체	뉴모로픽 반도체
구조	• 셀(저장 · 연산), 밴드위스(연결)	• 뉴런(신경기능), 시냅스(신호전달)
강점	• 저장과 연산	• 이미지와 소리 느끼고 패턴 인식
기능	• 각각의 반도체가 정해진 기능만 수행	• 저장과 연산 등을 함께 처리
데이터 처리 방식	• 직렬(입출력을 한번에 하나씩)	• 병렬(다양한 데이터 입출력을 동시에)

[그림 8] 기존 반도체와 뉴모로픽 반도체의 비교

〈표 2〉 국내외 주요 뉴로모픽 반도체 성능비교

	Neurogrid (2009) Stanford Univ.	SpiNNaker (2012) Manchester Univ.	SyNAPSE TrueNorth (2014) IBM, HRL	Zeroth (2014) Qualcomm	Loihi (2017) Intel	Neo^{2}C (2016-2021) KIST & etc.	Brain Human (Biology)
Neurons	10^6	2×10^7	10^6	–	13×10^4	10^2(2018) 10^6(2021)	10^{10}-10^{12}
Synapses	8×10^9	2×10^{10}	256×10^6	–	13×10^7	2×10^8(2021)	2×10^{14}
Energy Consumption (mW/cm^2)	50	1,000	20	–	–	56	10
Manufacturing (nm)	180	130	28	–	14	55	–

※ 출처: 융합연구정책센터(2017) 재구성

습기능에 의해 많은 데이터에서 연산성능이 더 높아진다([그림 8], 〈표 2〉 참조).

뉴로모픽 소자는 현재 인텔, 퀄컴, IBM에서 상용화가 가능한 시제품을 내놓고 있으며 향후 10여년 내에 시장에 본격 진입하여 자율주행자동차, 지능형 로봇 및 각종 인공지능 기기에 탑재될 것으로 기대되고 있다(Gartner, Technology Hype Cycle).

참고문헌

[1] ETRI Insight, '지능형 반도체의 주요 응용분야 시나리오와 핵심가치', 2019. 2.

[2] 정보통신정책연구원, '지능형 반도체 기술개발을 위한 기획 연구', 2018. 2.

[3] 한국과학기술기획평가원, 2019-01호, '인공지능(반도체)'

[4] 정보통신기술진흥센터(주간기술동향), '반도체 시장의 새로운 바람, 지능형 반도체', 2017. 5.

[5] 중소기업벤처기업부, '2019-2021 중소기업 전략기술로드맵'

[6] 뉴스핌, '인공지능 반도체의 미래', 2019. 2.

[7] https://brownbears.tistory.com/431

[8] 테크월드, '새 시대의 GPU, 그리고 인공지능', 2019. 8.

[9] https://en.wikichip.org/wiki/intel/loihi

[10] KIISE Transactions on Computing Practices 'STDP 알고리즘과 스파이크 간의 시간적 상호작용에 따른 SNN의 학습 성능 및 시간 분석' 2018. 9.