2024-6

KERC
Issue Report

유럽 인공지능 반도체 연구개발 동향

ISSUE REPORT
2024-6

유럽 인공지능 반도체 연구개발 동향

[저 자] 조 권 도 책임연구원
[발행인] 조 우 현 센터장
[담당자] 송 예 일 연구원
[발행일] 2024.11.26.
[발행처] 한-EU 연구협력센터
Rue de la science 14A
1040 브뤼셀, 벨기에
http://www.k-erc.eu
+32 (0)2 880 39 05

Content

Ⅰ. 서론 ····· 4

1. 분석 배경 ····· 4
2. 인공지능의 역사 ····· 6

Ⅱ. 인공지능 반도체 개요 ····· 13

1. 인공지능 반도체의 필요성 ····· 13
2. 인공지능 반도체 분류 ····· 15
3. 인공지능 반도체 기술 동향 ····· 16

Ⅲ. EU 기관별 연구동향 ····· 35

1. 벨기에 imec ····· 36
2. 프랑스 CEA-Leti ····· 39
3. 프랑스 GML ····· 42
4. 스위스 취리히 대학 ····· 43
5. 영국 맨체스터 대학 ····· 45
6. 독일 하이델버르그 대학 ····· 46
7. 독일 프라운호퍼 IPMS ····· 48
8. 독일 TU Dresden ····· 50
9. 독일 Max Planck Institute ····· 51
10. 프랑스 UPMEM ····· 52

Ⅳ. EU 프로젝트 동향 ····· 53

1. EU human brain project 2013-2023 ····· 53
2. Horizon Europe Framework Programme ····· 56
3. Horizon 2020 Framework Programme ····· 59
4. Digital Europe Programme ····· 64

Ⅴ. 결론 ····· 66

Ⅰ. 서론

1 분석 배경

ㅇ 인공지능 기술의 발전과 확산

- 최근 ChatGPT로 대표되는 인공지능 기술의 경이로운 발전은 우리 일상의 많은 영역에서 중대한 변화를 초래

※ 언어 번역, 정보 검색, 여행계획, 헬스케어, 영상편집, 문서작성, 온라인 학습, 의료, 패션 등의 분야에서 다양한 인공지능 기반 서비스가 출시되면서 그 활용 범위가 넓어지고 사용자 수가 가파르게 증가 중

- AI 기술은 점점 더 다양한 산업에서 중요한 역할을 수행하며, 개인화된 사용자 경험 제공, 자동화, 실시간 데이터 분석과 같은 분야에서 혁신을 이루는 중

- **대화형 AI와 자연어 처리**: ChatGPT로 대표되는 대화형 인공지능 서비스는 고객 지원, 교육, 헬스케어 분야에서 고도화된 챗봇과 가상 비서로 활용되어 사용자의 요구에 실시간 응대. AI 기반 서비스가 고도화되면서 더 정교하고 자연스러운 소통이 가능해짐에 따라, 기업들은 비용 절감과 고객 만족도 제고 달성
- **생성형 AI의 부상:** AI가 단순히 데이터를 처리하는 것을 넘어, 콘텐츠를 생성(창작)하는 영역에서 활발하게 사용됨. 텍스트 생성뿐만 아니라 이미지, 비디오, 음악을 생성하는 AI는 창의적인 산업분야(광고, 미디어, 디자인 등)에 큰 변화를 일으키는 중
- **엣지용 AI:** IoT 기기와 같은 엣지 기기에 장착된 AI는 실시간 데이터 처리와 함께 저전력, 고효율 시스템에 대한 수요를 이끄는 중. 특히, 스마트폰, 드론, 자율주행차 등의 엣지 디바이스에서 AI의 실시간 추론 기능을 저전력으로 실현하는 방향으로 발전 중
- **의료용 AI:** 의료 분야에서 AI는 진단 시스템, 개인 맞춤형 치료 등의 영역에서 활용되며, 헬스케어 분야의 혁신을 견인. 질병을 예측하고 진단의 정확도를 높이고, 환자 데이터를 면밀히 분석해 개인화된 치료 계획을 제공함으로써 의료 서비스 품질 제고 역할

○ 인공지능 반도체 시장의 성장과 방향

- 최근 인공지능 서비스의 화려한 성장은 고도화된 반도체 기술이 뒷받침 되었기에 가능
- 근래의 반도체 시장 성장추세에 따르면, 기존의 반도체 시장은 성장이 지체되어 있는 반면, 인공지능과 관련된 반도체 시장(NPU, HBM 등)은 빠르게 성장 중
- 학습량이 많을수록 더 똑똑해지는 생성형 AI의 특성으로 인해 학습 데이터의 거대화, 그로 인한 연산량과 에너지 사용량의 폭증으로 인해 에너지 효율 개선이 큰 화두

• AI의 과도한 에너지 소비 문제는 중요한 환경문제로 부각되고 있으며, UN이 제정한 지속가능한 발전목표 달성을 위해 에너지 절감에 관한 관심 필요 • NPU (Neural Processing Unit)와 같은 전용 가속기 개발, 그리고 PIM (Processing In Memory)과 같은 폰 노이만 구조 개량 기술들은 에너지 효율을 개선하면서도 높은 성능을 유지할 수 있는 해결책 제공 • 다른 한 편으로는 인간의 뇌가 약 20Wh의 적은 에너지만으로 고차원적인 학습과 판단을 병렬 처리할 수 있다는 점에 착안한 뉴로모픽 컴퓨팅 연구가 활발히 진행 중. 뉴로모픽 컴퓨팅은 뇌의 동작 원리를 하드웨어적으로 모사하여 에너지 효율을 획기적으로 개선하는, AI 반도체 분야의 새로운 전환점이 될 것으로 기대

○ 유럽의 연구 동향

- 유럽은 HBP(Human Brain Project)를 통해 지난 10년간 대규모의 두뇌 연구를 지원해왔으며, 이를 통해 조성된 연구생태계를 기반으로 뉴로모픽 컴퓨팅 연구가 활발히 진행 중
- 유럽의 연구기관들은 R&I 프로그램을 통해 지속적인 연구 협력을 이어가고 있으며, 향후 인공지능 반도체 분야에서 중요한 역할을 할 것

○ 한-EU 협력의 필요성

- 내년으로 예정된 한국의 호라이즌 유럽 필라2 준회원국 승격은 한국에는 새로운 기회이자 도전. 호라이즌 유럽을 통해 더욱 활발한 한-EU 간 협력을 도모할 기회

- 유럽은 HBP를 통해 인공지능 연구생태계가 활성화되어 있으며, 한국은 메모리에 국한된 반도체 강국의 지위를 확장하는 측면에서, 아직 초기 연구 단계인 뉴로모픽 컴퓨팅을 중심으로 한-EU간 연구교류와 협력이 중요한 시점
- 본 고는 인공지능과 인공지능 반도체 기술의 발전사, 시장 동향, 주요 핵심기술을 언급하고, 인공지능 반도체가 풀어야 할 기술적 과제를 소개
- 인공지능 반도체 분야에서 유럽 연구기관들이 수행해 온 연구 내용과 프로젝트 현황을 공유하고, 한-EU 협력 주제와 파트너 발굴을 위한 기반 정보제공

2 인공지능의 역사[1]

① 초기 연구 (1940~1980년대)

1940년대	맥컬록, 피츠, 헵, 로젠블랫 같은 다양한 학자들의 노력과 아이디어로 인공신경망 연구가 진행되어 현대 인공지능 기술의 기반을 형성
1943년	워렌 맥컬록(Warren McCulloch)과 월터 피츠(Walter Pitts)는 생물학적 뉴런의 특성을 가진 최초의 인공신경망 모델인 맥컬록-피츠 모델을 발표 - 이 모델은 기계가 강력한 연산 능력을 가질 수 있는 이론적인 가능성을 제시한 것으로 인공신경망 연구의 토대를 마련
1949년	도널드 헵(Donald Hebb)은 '지속해서 활성화되는 뉴런은 연결된다'라는 헵의 가설 발표 - 이 가설은 학습 과정에서 시냅스의 연결강도가 변하며 기억이 형성된다는 사실을 설명하는 중요한 이론
1950년대 초	영국의 수학자 앨런 튜링(Alan Turing)은 기계가 생각할 수 있을지에 대한 화두 제안 - 사람이 컴퓨터와 대화를 나눈 후 컴퓨터인지 인간인지 구별하지 못하면 그 컴퓨터는 '지능'을 가졌다고 볼 수 있다는 '튜링 테스트' 개념은 인공지능 연구의 중요한 이정표
1956년	다트머스 회의(Dartmouth Conference)에서 '인공지능'이라는 용어 등장 - 본 회의에서 존 매카시(John McCarthy)를 포함한 여러 학자가 컴퓨터가 인간의 지능적 행동을 모방할 수 있다는 아이디어 논의

1) https://news.skhynix.co.kr/post/all-around-ai-1

1957년	프랑크 로젠블랏(Frank Rosenblatt)은 '퍼셉트론(Perceptron)' 모델을 통해 컴퓨터가 패턴을 인식하고 학습할 수 있다는 개념을 실증 - 프랭크 로젠블랫(Frank Rosenblatt)은 인간의 두뇌 움직임을 수학적으로 구성하여 '퍼셉트론(Perceptron)'이라는 최초의 신경망 모델을 발표. 퍼셉트론은 당시 큰 이슈를 불러일으키며 인공신경망 연구에 활력 - 이는 1943년에 신경 생리학자 워렌 맥컬록과 월터 피트가 신경세포의 상호작용을 간단한 계산 모델로 정리한 신경망 이론을 실제 테스트에 활용한 실증사례
1960년대	기계학습, 체스게임, 자연어 처리 분야에서 인공지능 연구 - 1966년, 조셉 와이젠바움(Joseph Weizenbaum)이 개발한 대화형 프로그램 '엘리자(ELIZA)'는 인간과 컴퓨터 간의 대화가 가능하다는 것을 실증
	긍정적인 연구 성과로 인해 세간의 기대는 높았으나, 컴퓨팅 성능, 논리 체계, 데이터 부족 등의 한계로 인공지능 연구는 곧 침체기에 진입 - 초기 인공지능 시스템들은 특정 분야에서 제한된 성과를 보였으나, 현실 세계에서 사용할 만한 수준에 미치지 못하는 한계 확인 - 특히, 퍼셉트론 모델이 비선형 문제를 해결할 수 없다는 한계 확인
1970년대	1970년대와 1980년대 초반은 인공지능 겨울로 불리는 인공지능 연구 침체 시기로 연구 투자와 관심 급감

② 재도약의 발판 (1980~1990년대)

- 1980년대 후반과 1990년대는 인공신경망 연구의 재도약기

1982년	존 홉필드의 홉필드 네트워크 제안. 현대 인공신경망 연구의 초석
1985년	홉필드 네트워크를 보완한 제프리 힌턴의 볼츠만 머신과 역전파 알고리즘이 발표되면서 신경망 훈련 방법에 큰 진전
1986년	제프리 힌턴은 인공신경망을 여러 겹 쌓은 다층 퍼셉트론(Multi-Layer Perceptrons) 이론에 역전파 알고리즘을 적용하여 퍼셉트론의 기존 문제를 해결할 수 있음을 증명 - 그러나 신경망의 깊이가 깊어질수록 성능이 열화되는 난제 대두

- 인공지능 연구는 **머신러닝**을 기반으로 다시 성과를 내기 시작

• 인간의 명령으로만 작동하던 인공지능은 1990년대 들어 머신러닝(기계 학습) 알고리즘 덕분에 규칙을 스스로 찾아내는 학습이 가능해짐 • 디지털과 인터넷의 등장 덕분에 웹에서 대량으로 데이터를 수집/활용할 수 있게 된 환경도 큰 도움 • AI는 스스로 규칙을 학습하고, 나아가 사람이 찾지 못하는 규칙까지 찾을 수 있게 됨

③ 딥러닝의 발전 (2000년대)

2000년대 중반	딥러닝(deep learning) 기술의 진전은 인공지능 연구에 진정한 돌파구 역할
2006년	제프리 힌턴(Geoffrey Hinton)은 심층 신경망의 효율적인 훈련 방법을 제시(비지도 학습법을 통해, 층별로 신경망을 사전 학습하는 방법) - 심층 신경망(Deep Neural Network))은 기존의 머신러닝 기법보다 훨씬 많은 계층으로 구성할 수 있게 됨으로써, 더 복잡한 학습에 사용할 수 있게 됨
2012년	2012년 개최된 '이미지 인식 경진대회 ILSVRC (ImageNet Large Scale Visual Recognition Challenge)'에서 제프리 힌턴 교수가 이끄는 알렉스넷(AlexNet) 알고리즘이 압도적 성능으로 우승하며, 딥러닝의 잠재력을 세상에 알림 - 이를 계기로 인공지능 연구는 폭발적으로 성장하며, 인공지능이 전 산업 분야에 빠르게 확산되기 시작

④ 알파고의 충격 (2010년대)

- 인공지능의 대세가 된 딥러닝은 2010년대부터 급성장

※ 급성장 배경에는 GPU(Graphics Processing Unit, 그래픽처리장치)를 비롯한 컴퓨팅 기술의 발전 그리고 인터넷, 스마트폰, IoT 기술의 확대로 인해 빅데이터 수집이 용이해진 환경 덕분

2016년	구글 딥마인드가 개발한 인공지능 알파고(AlphaGo)가 4승 1패로 바둑기사 이세돌 9단을 꺾으며 전 세계에 AI 존재감을 각인 - 알파고는 딥러닝 알고리즘과 강화학습, 몬테카를로 트리탐색 알고리즘을 결합한 형태. 수만 번의 자가 대국을 통해 스스로 학습하고, 인간의 직관을 모방하여 수를 예측하고 전략을 수립 - 인간을 꺾은 AI 바둑의 탄생은 본격적인 인공지능 시대의 시작을 알린 신호탄

⑤ AI와 함께하는 일상 (2020년대)

- 인류는 인공지능 기술로 거대한 변혁을 맞이

2022년 말	오픈 AI가 LLM(거대 언어 모델) GPT(Generative Pre-trained Transformer)를 탑재한 ChatGPT 3.5 베타를 출시하면서 생성형 AI(Generative AI) 시대 개막 - 생성형 AI는 인간의 고유 영역으로만 여겨지던 창작의 영역에 침투하여 다양한 형태의 수준 높은 콘텐츠 생성 - 데이터를 바탕으로 예측하거나 분류하는 딥러닝의 수준을 넘어 사용자의 요구에 따라 대규모 언어모델(LLM)을 활용해 스스로 결과물을 생성
2023년	새롭게 출시된 GPT-4는 GPT-3.5보다 약 500배 더 많은 데이터를 학습 - 텍스트를 넘어 이미지와 오디오, 비디오 등의 다양한 데이터를 입력받아 처리하고, 출력 데이터 역시 다양하게 생성하는 LMM(대형 멀티모달 모델)으로 진화

- ChatGPT가 촉발한 생성형 인공지능의 붐을 타고, 기업들은 너나없이 다양한 생성형 인공지능 서비스 출시 중

- 텍스트, 이미지, 오디오 등을 동시에 인식하고 이해할 수 있는 구글의 제미나이(Gemini)와 이미지 내 특정 객체를 정확하게 인식하고 분리할 수 있는 메타의 샘(SAM), 텍스트 프롬프트 기반으로 영상을 제작하는 오픈AI의 소라(Sora) 등이 대표적

- 일상 속에서 스마트폰의 음성 비서, 추천 알고리즘, 자율주행 자동차, 의료 진단 시스템 분야에서 AI 기술 활용 중 (특히, 챗봇이나 자동번역 시스템은 과거의 영화속 상상이 현실이 된 대표 사례)

- AI는 무한한 가능성을 내포하고 있으며, AI가 만드는 세상은 이미 인류의 상상을 뛰어 넘음

- AI는 무한한 가능성을 내포. 사람을 돕고, 복잡한 문제를 해결하며, 새로운 과학적 발견을 하는 데까지 활용할 수 있으며, 향후 AI는 더욱 인간과 자연스럽게 상호작용하고, 창의적인 작업에서도 더 많은 역할을 할 것으로 기대됨
- 2024년 AI의 대부 제프리 힌턴은 'AI가 산업혁명에 비견될 것'이나, '통제 불능상태가 될 수 있는 위협에 대해 우려해야 한다'며 경고
- 예상하기 어려운 미래의 변혁에 대처하기 위한 인류의 진지한 고민 필요

⑥ 2024년 노벨 화학/물리학상

- 노벨 물리학상뿐 아니라 노벨 화학상까지 인공지능과 연관된 공헌자들이 수상. 인류가 본격적인 인공지능 시대에 진입했음을 알린 사건
- 노벨 화학상은 인공지능을 이용한 단백질 구조예측과 설계에 관한 연구 성과를 인정받은 3인*에 수여

* 데이비드 베이커 미국 워싱턴대 단백질디자인연구소 교수, 구글 딥마인드의 데미스 하사비스 최고경영자 및 존 점퍼 수석연구원

2024년	1970년대부터 머신러닝의 기초를 닦은 공헌으로 존 홉필드와 제프리 힌턴이 노벨 물리학상을 수상 - 전통적인 물리학 이외에 첨단 정보기술(IT)과 관련해 노벨 물리학상을 받은 이례적인 사건

참고 ① 존 홉필드의 홉필드 네트워크

- 존 홉필드는 1933년 미국 시카고에서 태어나 1958년 미국 코넬대에서 박사학위를 취득, 현 프린스턴대 교수로 AI 학습의 기본이 되는 인공신경망 원리를 1980년대에 처음으로 제안한 공로로 2024년도 노벨물리학상을 수상

- 1982년에 제안한 '홉필드 네트워크'는 뇌 작동 방식에서 영감을 얻고, 통계물리학적인 접근을 통해 풀어낸 인공신경망으로, 현대 인공신경망 연구의 초석이 됨
- 홉필드 네트워크는 이미지와 패턴을 저장하고 재구성할 수 있는 '연관 메모리' 시스템의 토대를 마련
- 홉필드 네트워크는 물리학의 원자 스핀 개념을 활용해 패턴을 저장하고 재생성하는 원리로써, 왜곡되거나 불완전한 이미지를 입력받아 가장 유사한 이미지를 찾아낼 수 있어 패턴 인식과 기억 시스템 연구에 큰 공헌

참고 ② 딥러닝의 대부 제프리 힌턴

- 1947년 영국 런던에서 태어나 1978년 영국 에든버러대에서 박사학위 취득. 현 캐나다 토론토대 교수
- AI 4대 천황으로 꼽히는 힌턴 교수는 심층학습(딥러닝)의 개념을 처음으로 고안
- 1985년 홉필드 네트워크를 보완한 볼츠만 머신(Boltzmann machine) 고안. 데이터 속의 특징적 요소를 자동으로 인식하는 방법을 학습할 수 있어, 이미지 분류나 새로운 패턴 생성에 활용
- 이후 '역전파 알고리즘'을 개발해, 방대한 데이터 속에서 스스로 특성을 찾아내는 현대의 딥러닝 알고리즘의 기반을 다짐. 이는 출력 결과와 실제 값 간의 오차를 신경망의 각 층에 역으로 전파하여 가중치를 조정해 나가는 학습방식으로, 복잡한 학습을 가능하게 하는 기법
- 역전파 알고리즘은 현재 대부분의 딥러닝 모델에 널리 활용
- 심층 신경망에서의 효율적인 훈련 방법 제시함으로써, 신경망의 깊이가 깊어질수록 성능이 열화되는 난제를 해결한 쾌거
- 비지도 학습 기법을 통해, 층별로 신경망을 사전 학습하는 방법 제시. 이를 통해 심층 신경망(Deep Neural Network)은 기존의 머신러닝 기법보다 훨씬 많은 층으로 구성할 수 있게 되어 더 복잡한 학습이 가능해짐
- 힌턴 교수가 제시한 심층학습은 AI 기술의 토대가 되었고, 2016년 이세돌 9단을 이긴 바둑 AI '알파고'의 기반 기술이 됨

참고 ③ 단백질 구조 파악의 중요성

- 단백질의 3차원 구조는 단백질의 기능을 결정
- 단백질은 그 구조에 따라 특정한 생화학적 활동을 수행하므로, 구조를 파악하고 나면 단백질이 어떻게 작동하는지, 어떻게 특정 분자와 상호작용하는지를 이해할 수 있음

- 알츠하이머나 헌팅턴병 처럼 단백질 구조가 변형되거나 잘못 접히면서 발생하는 질병의 원인 규명과 치료제 개발에 단백질 구조 이해가 큰 도움
- 단백질 구조를 알면 특정 단백질을 표적으로 하는 신약 설계 가능
- 단백질은 20가지 아미노산의 서열로 구성되며, 서열에 따라 서로 다른 3차원 구조로 접히는 특징을 가짐
- 서열별로 단백질 구조를 알아내는 일은 랜덤한 뽑기와 같이 운에 맡겨야 해서 진보가 매우 더딘 연구분야였으나, 실험없이 계산만으로 구조를 예측하는 계산생물학에 AI가 도입되면서 혁신적인 발전

참고 ④ 알파폴드2와 로제타폴드

- 알파폴드2 등장 전에는 단백질 구조예측 연구에 이미지 프로세싱 기술과 CNN (Convolutional Neural Networks)이 활용되었음
- 알파폴드2와 로제타폴드가 기존 모델보다 뛰어난 성능을 보이는 비결은 트랜스포머(Transformer)의 일종인 어텐션(Attention)을 도입했다는 점

데미스 하사비스의 알파폴드2

- 바둑 AI 알파고로 알려진 하사비스가 이끄는 딥마인드팀이 개발한 알파폴드2는 AI를 활용한 단백질 구조예측 프로그램 (2020년 공개)
- 2020년 11월 단백질 구조예측 능력 평가 대회인 CASP에서 알파폴드2가 압도적 성능으로 우승하며 생물학계에 큰 파장
- 이미 알려진 단백질 구조와 아미노산 배열을 학습함으로써, 새로운 아미노산 배열만으로 단백질 구조를 단시간에 예측할 수 있게 됨
- 독일 막스 플랑크연구소에서 단백질 구조를 연구 중인 안드레이 루파스 박사는 '10년 동안 알아내지 못한 특정 단백질 구조를 알파폴드2는 반 시간 만에 밝혀냈다', '알파폴드2는 게임체인저이며, 앞으로 단백질 구조분석은 전적으로 컴퓨터에 의존하게 될 것'이라 강조
- 하사비스는 알파폴드를 내놓으며 '디지털 생물학'의 시대가 열렸다는 자평

데이비드 베이커 교수의 로제타폴드

- 단백질 예측 프로그램인 로제타폴드는 비밀스러운 단백질의 접힘 구조를 해독한다는 의미로 기원전 196년 고대 이집트에서 만들어진 비석 로제타스톤의 이름을 딴 것
- 알파폴드2보다 늦은 2021년 공개된 로제타폴드는 아미노산 서열을 파악하고, 아미노산들이 어떻게 연결될지 예측하고, 이를 토대로 어떤 입체 구조를 가질지 예측하는 세 과정을 반복하며 정확도를 높이는 원리

Ⅱ. 인공지능 반도체 개요

1 인공지능 반도체의 필요성

① 빅데이터 그리고 연산량의 폭발적 증가

- 빅데이터 기반으로 널리 활용되는 딥러닝 모델은 막대한 양의 데이터를 깊은 다층 신경망을 사용하여 학습하고 추론하므로 연산량이 기하급수적으로 증가
- 딥러닝 모델은 병렬처리 연산이 대부분. 인공지능 반도체는 병렬처리 기능을 대폭 강화한 구조로 설계. 순차적인 연산에 최적화된 CPU는 이러한 대규모 병렬처리 연산에 비효율적
- GPU(Graphics processing unit) 혹은 신경망 처리장치(NPU, Neural Processing Unit)는 방대한 데이터를 병렬처리를 통해 빠르게 연산이 가능
- 인공지능 신경망 모델의 복잡성 증대로 인한 연산량 폭증
- GPT-3.5 같은 대규모 자연어 처리 모델은 1,750억 개의 매개변수를 사용해 학습되며, 이는 기존의 연산장치로는 효율적인 처리가 어려운 수준
- 인공지능 반도체는 이러한 대규모 모델의 학습과 추론에 필요한 연산을 처리하기 위해 대규모 병렬연산 성능과 메모리 대역폭 개선이 필수

② 에너지 효율 측면

- 인공지능 알고리즘 실행에 연산량이 많은 만큼 에너지 소모 폭증
- 데이터센터나 클라우드 서버에서 인공지능 모델을 학습/추론할 때 대규모 전력을 소비하므로 심각한 환경비용 문제 초래
- 인공지능 반도체는 에너지 효율성을 극대화하는 구조로 진화 필요
- 적은 전력으로 더 많은 연산을 수행할 수 있어야 데이터센터의 막대한 운영 비용을 줄이고, 엣지 디바이스와 같이 전력사용에 제한이 있는 이동형 장치에서도 인공지능 활용 가능

③ 실시간 처리에 대한 요구 증가

- 자율주행차, 드론, 로봇, 스마트폰 등에서 빠른 의사결정이 필요한 응용 분야에서는 신속한 데이터 처리가 요구됨
- 예를 들어 자율주행차는 매 순간 도로 상황을 분석하고 즉각적인 반응을 해야 하므로, 판단 지연 시간 최소화 필요
- 실시간 데이터 처리 성능을 극대화를 위해 연산속도가 높은 저전력 인공지능 반도체가 필요

엣지 컴퓨팅(Edge Computing) 수요 증대

- 엣지 컴퓨팅은 데이터를 중앙 서버나 클라우드가 아닌 데이터 생성지점에서 직접 정보를 처리
  - 일반적으로 인공지능 서비스를 이용하기 위해서는 대규모 클라우드 서버와 통신하여 엣지 디바이스로 데이터를 끌어와야 하지만, 온디바이스 인공지능 반도체는 휴대폰 및 차량 같은 엣지 디바이스 자체적으로 인공지능 연산을 처리
- 엣지용 AI는 서버와의 통신에 따른 보안이나 시간 지연 문제가 개선되어 개인화된 인공지능 서비스가 가능한 장점으로 인해 큰 성장 가능성
  - 음성인식, 이미지 처리, 증강현실 등과 같은 엣지 디바이스에서의 실시간 인공지능 서비스에 대한 수요 증가
- 이동성이 필요한 엣지 디바이스의 특성상 적은 배터리 소모로 인공지능 연산을 실행하기 위해서는 저전력, 고성능의 인공지능 반도체가 필수

④ 특정 작업에 최적화된 성능 제공 필요

- 특정 인공지능 연산(딥러닝, 머신러닝, 이미지 및 음성 인식 등)에 국한된 최적화 설계 시 개량된 성능 달성 가능
- CPU와 일반 GPU와 같은 장치는 범용으로 설계되었기에 다양한 영역에 활용가능하나, 특정 인공지능 연산에 최적화되어 있지 않아 개선의 여지 있음
- 특정 인공지능 전용 하드웨어는 해당 인공지능 서비스에 특화된 모델을 더 빠르게 처리할 수 있음

2 인공지능 반도체 분류[2]

※ 인공지능 반도체는 ①용도, ②서비스 플랫폼, ③최적화 여부로 분류

① 용도에 따른 분류

학습용(Learning)	추론용(Inference)
방대한 데이터를 대상으로 러닝 알고리즘을 통해 지식을 습득하는 학습 단계에 적합하게 설계된 인공지능 반도체	학습한 내용을 토대로 외부 명령이나 상황에 따라 적절한 해답을 신속히 내놓는 추론용 인공지능 반도체

② 서비스 플랫폼에 따른 분류

데이터센터 서버용	엣지 디바이스용
데이터센터용 인공지능 반도체는 빠른 병렬연산 능력과 에너지 효율이 중요하며, 서버 운영 측면에서는 확장성과 유연성도 중요 고려 요소	자율주행차, 드론, IoT 등 개별 인공지능 서비스에 특화된 엣지 디바이스용 인공지능 반도체는 저전력, 경량화, 소형화, 연산속도, 제조원가 등이 주요 경쟁력 요소

② 최적화 여부에 따른 분류

범용 구조	특정 용도
CPU, GPU와 같이 일반적인 컴퓨팅 용도로 널리 활용 가능한 프로세서로써, 대규모 인공지능 연산 시 성능과 에너지 효율이 낮음	NPU*, 뉴로모픽 반도체, AI용 GPU와 같이 특정 인공지능 연산에 필요한 기능 위주로 특수 제작된 반도체로써, 소모전력의 고효율화와 개선된 연산속도 제공

* NPU: 구글이 자사의 데이터센터에 이용하기 위해 개발한 TPU(Tensor Processing Unit) 프로세서, 자율주행용으로 개발된 NVIDIA의 Xavier Soc 와 Tesla의 차량용 프로세서 FSD (Full Self Driving), 그리고 퀄컴의 스마트폰용 AP인 스냅드래곤에 포함된 엣지 디바이스용 NPU인 헥사곤 등이 시장에 발표된 NPU의 대표적 예

2) AI 반도체 시장 동향 및 우리나라 경쟁력 분석, ETRI Insight, 기술정책 트렌드 2020-12

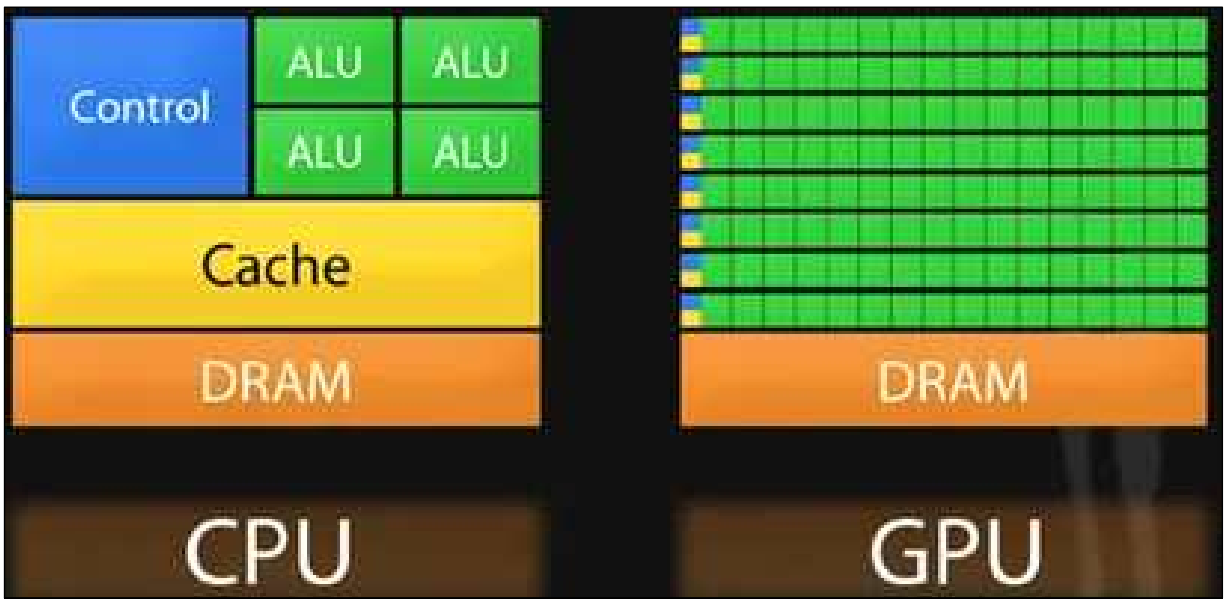
3 인공지능 반도체 기술 동향

① GPU의 활용

ㅇ **GPU는 대규모 데이터를 병렬처리하고 다양한 구조의 신경망을 학습하는 데에 CPU보다 성능 탁월**

- GPU는 당초 고화질 그래픽 연산을 효율적으로 처리하기 위한 장치였으나, 인공지능의 반복적인 연산 대부분을 GPU의 병렬 연산유닛에서 처리할 수 있는 특징 덕분에 인공지능 핵심 반도체 역할 수행
- GPU는 이미지와 동영상 같은 데이터를 처리할 때 많은 수의 픽셀을 동시에 연산해야 하는 그래픽 작업에 유리하도록 설계
- 이러한 병렬처리 능력은 딥러닝 모델의 행렬 연산에 효과적이어서, AI 연산용으로 널리 활용
- GPU는 많은 수의 범용 연산 코어를 통해 계산을 수행하므로, 유연성이 높고 다양한 종류의 AI 모델과 데이터에 적응할 수 있는 장점이 있지만, 특정 AI 연산에 맞춘 최적화된 구조는 아님

<그림: GPU와 CPU의 차이 (출처:doi:10.3389/fgene.2013.00266)>



ㅇ **GPU는 학습과 추론 두 영역에서 모두 높은 성능을 발휘하는 범용성이 특징**

- 에너지 효율보다 성능이 우선시 되는 대규모 데이터센터나 클라우드 인프라에 널리 활용되는 경향

ㅇ **NVIDIA와 AMD는 인공지능 학습과 추론 용도에 맞는 GPU를 지속적으로 출시해 왔음**

- NVIDIA V100, A100와 같은 데이터센터용 고성능 GPU부터 Jetson 시리즈와 같은 엣지 AI용 소형 모듈에 이르기까지 다양한 제품 출시
- AMD는 MI100, MI200, MI300과 같은 인공지능용 GPU 출시
- AI 연산을 효율적으로 처리하기 위한 하드웨어를 포함하여 제작된 GPU는 NPU로도 분류 가능

② NPU 시대 (Neural-network Processing Unit, 신경망 처리장치)

ㅇ **GPU와 같은 병렬처리 기능 외에 AI 연산*에 사용되는 특정 수학적 연산에 맞춤형으로 최적화된 프로세서**

* 예: 합성곱 연산, 활성화 함수 계산 등

- NPU는 인공지능 플랫폼의 용도에 따라 요구되는 인공지능 연산에 특화된 특정용도 전용 반도체
- 특정용도로 제작되는 만큼 필요한 인공지능 프로세서 외에 각종 센서, DSP, 통신모듈 등이 추가로 포함된 시스템반도체(SoC) 형태로 출시되는 경향

ㅇ **신경망 처리장치(NPU)는 머신러닝 전용으로 설계된 칩이므로, 범용으로 설계된 GPU보다 에너지 효율과 성능 측면에서 우월**

- NPU는 다양한 회사에서 각자의 용도에 맞게 제작하므로, 성능은 우수하나 범용성은 떨어짐
- 인공지능 가속기는 NPU의 또 다른 이름

ㅇ **대용량 데이터를 병렬처리하고, 적은 전력으로 복잡한 연산, 추론, 학습이 가능한 NPU는 온-디바이스* 인공지능 구현에 필수 장치**

* 스마트폰, 노트북, 가전 기기 등

- 스마트폰, 자율주행차, IoT 기기와 같이 에너지 효율이 중요한 요소인 엣지 디바이스에서는 일반 GPU보다 인공지능 추론연산에 최적화된 NPU가 널리 사용되는 경향

- 스마트폰에서 인공지능 기반의 사진 보정, 음성 인식, 얼굴 인식을 빠르게 처리하거나, 자율주행차에서 실시간으로 데이터를 분석하고 의사 결정을 내리는 작업에 활용. 또한, IoT 장치, 웨어러블 기기, 드론, 자율 로봇 등과 같이 전력사용이 제한된 이동 환경에서 실시간 추론에 널리 활용

<회사별 NPU 개발 현황>[3)]

회사	NPU 개발 현황
Meta	- 자체 개발한 AI반도체 MTIA(Meta Training & Inference Accelerator)를 자사 데이터센터에 탑재 계획 ▪ 2023년 5월, 1세대 MTIA를 공개. 2024년 4월 2세대 칩 사양 공개 - 1세대 제품은 페이스북, 인스타그램 등의 콘텐츠 추천 서비스 등에 사용되나 Meta는 자체 개발한 인공지능 반도체를 궁극적으로 생성형 인공지능 학습용으로 활용하려는 목표
Google	- 세계 3위의 클라우드 업체로써, 2016년 클라우드용 인공지능 반도체 TPU(Tensor Processing Unit)를 발표 후 6세대 Trillium까지 진화 중 - TPU는 구글의 인공지능 엔진 TensorFlow에 최적화 - 2024년 하반기에 자체 개발한 ARM 아키텍처 기반의 서버용 CPU Axion 출시계획
Micorosoft	- 세계 2위의 클라우드업체로써, 2019년부터 '아테나' 프로젝트를 통해 자체 AI칩 개발을 추진 - 2023년 11월에 자체 개발한 AI반도체 첫 공개 ▪ GPU를 대체할 인공지능 가속기 Maia 100, 클라우드 Azure용 CPU인 Cobalt 100 발표 ▪ 설계단계부터 제품 테스트까지 OpenAI와 협력 ▪ Maia는 외부 판매 계획은 없으나 Cobalt는 향후 타사에도 판매할 수 있다는 입장
AWS (Amazon Web Service)	- AWS는 세계 1위 클라우드 업체로써, 2018년부터 인공지능 반도체 개발. 타 클라우드사 대비 자사 칩 개발/활용에 적극적 ▪ AWS와 AMD는 서버용 프로세서 개발에 협력한 바 있으나 성능 문제로 결별 ▪ 이스라엘 팹리스 Annapurna Labs 인수('15) - 2018년 출시된 서버용 CPU Graviton은 ARM 아키텍처를 채용하여 인텔, AMD가 사용하는 기존 X86 아키텍처 기반 CPU 대비 높은 에너지 효율 달성. 2023년 11월에 4세대 제품 공개 - 2018년에 추론용 칩 Inferentia 개발 - 2020년에 머신러닝에 특화된 Trainium 개발 - 2023년 초 Inferentia2, 11월에 Trainium2 공개

3) 이슈보고서 2024 AI반도체 시장 현황 및 전망, 한국수출입은행 해외경제연구소

Intel	- 인텔은 CPU에서 GPU, FPGA로 사업을 확대하며 데이터센터 사업 강화 추진 - FPGA 기업 Altera, AI반도체 스타트업 하바나랩스 인수 등을 통해 데이터센터 공략 ▪ 2016년 인공지능 가속기 스타트업 Nervana를 인수했으나 Nervana 기반 제품은 실패 - 2019년 하바나랩스, 차량용 반도체기업 모빌아이 등 인수 - 하바나랩스는 인공지능 학습용 반도체 Gaudi를 3세대까지 개발 ▪ Gaudi3는 2024년 3분기에 출시 예정으로 델, HP, 레노버, 슈퍼마이크로 등이 채택
Graphcore	- Graphcore는 영국을 대표하는 AI반도체 스타트업. 한때 엔비디아의 대항마로 부상했으나 대형 고객사 이탈과 미국의 중국 규제 강화 등의 악재 - 2016년 엔비디아 출신 개발자가 설립한 기업으로 IPU(Intelligent Processing Unit, 지능형 처리장치)를 설계하며 마이크로소프트, 삼성전자 등으로부터 투자 받음 ▪ IPU는 벡터 연산을 지원하는 GPU와 달리 그래프 연산을 지원하며 칩 안에 데이터를 저장할 수 있는 메모리를 넣어 연산 지연을 해소함으로써 유니콘 기업으로 부상 ▪ 주요 구매자는 마이크로소프트, Cirrascale, NHN 클라우드 등 - 마이크로소프트의 구매 중단, 미국의 대중국 반도체 수출 규제('23), AI반도체 기술 경쟁 심화 등으로 매출이 급감하면서 매각 추진 ▪ 마이크로소프트는 자체 AI반도체 개발을 추진하며 Graphcore와 거래 중단. Graphcore의 경쟁력은 다수 AI반도체 스타트업의 등장, 제품 출시 지연 등으로 약화 - 2022년 매출은 전년 대비 46% 감소한 2.7백만달러. 중국이 주력시장(매출비중 20~25%)이었으나 2023년에는 미국의 규제로 중국에서 철수 ▪ OpenAI, ARM, 소프트뱅크 등이 인수를 검토하였으며, 2024년 소프트뱅크가 인수 발표
NVIDIA	- 2018년 Xavier Soc 공개 ▪ 자율주행 차량과 엣지 컴퓨팅을 위한 고성능 AI 프로세서 ▪ ARM CPU, NVIDIA의 GPU, NVIDIA DLA(Deep Learning Accelerators), PVA(Programmable Vision Accelerators), ISP(Image Signal Processor)를 융합 ▪ 드론, 로봇 및 자율주행에 활용성이 높도록 다양하고 복잡한 AI 연산은 실시간으로 처리하도록 설계 - 2020년 A100 출시 ▪ 3세대 Tensor 코어를 탑재한 Ampere 아키텍처 기반의 고성능 GPU 또는 인공지능 가속기 - 2022년 H100 출시 ▪ A100의 후속 모델로 4세대 Tensor 코어 탑재 ▪ 트랜스포머 엔진 탑재로 AI 모델 학습 및 추론 성능 개선 ▪ HBM3 사용으로 A100 대비 2배 가까운 메모리 대역 지원

AMD	- Instinct MI300 시리즈는 AI 및 고성능 컴퓨팅(HPC)을 위해 설계 ▪ Instinct MI300X : 대규모 AI 추론 및 훈련, 특히 대형 언어 모델(LLM)과 생성형 AI 용도 ▪ Instinct MI300A : AI와 HPC의 융합을 위한 가속 프로세싱 유닛(APU)
Tesla	- 차량용 프로세서 FSD(Full Self-Driving)은 CPU, GPU, Dual core NPU를 융합한 자율주행용 NPU ▪ 50 TOPS(Tera Operations/Sec)의 계산 속도에서 40W/Chip의 저전력 소모
삼성전자	- 스마트폰용 AP인 엑시노스(Exynos)에 NPU 기능을 탑재한 9820 모델을 2018년 공개 - 갤럭시 S24에 탑재된 엑시노스 2400의 후속 모델 엑시노스 2500 준비 중
Qualcomm	- 2013년 인간의 신경망을 모방하여 학습하는 인공지능 프로세서 제로스(Zeroth) 공개 - 퀄컴의 헥사곤은 DSP와 인공지능 가속기가 결합된 NPU. 스마트폰용 AP인 스냅드래곤에 포함된 엣지 디바이스용 NPU - 강력해진 인공지능 기능이 탑재된 스냅드래곤8 4세대는 2024년 가을 공개예정 ▪ 스냅드래곤8 4세대 칩은 이전 모델과 달리 ARM CPU가 아닌 퀄컴이 자체 개발한 맞춤형 오라이온(Oryon) CPU 코어를 채택
Apple	- 2017년 NPU가 탑재된 애플의 첫 번째 스마트폰용 바이오닉 AP인 A11 공개 - 미국 시애틀에 본사를 둔 저전력 엣지 기반의 인공지능 전문업체 Xnor.ai 인수를 통해 인공지능 기술확보에 노력 (2020년) - 2024년 후속작 M4와 A18 출시를 통해 코어의 성능향상과 낮은 소모전력의 특징을 소개
대기업 외에도 호라이즌로보틱스(Horizon Robotics, 중국), 헤일로(Hailo Technologies, 이스라엘), 크네론(Kneron, 미국), 그린웨이브(GreenWaves Technologies, 프랑스), 에타 컴퓨트(Eta Compute, 미국), Cerebras systems(미국), Sambanova systems(미국), 퓨리오사AI(한국), 리벨리온(한국), 모빌린트(한국), 딥덱스(한국), 사피온(한국), 뉴블라(한국), 비전넥스트(한국), 에임퓨처(한국), 보스반도체(한국) 등 다양한 스타트업들이 AI 반도체 개발 경쟁 중	

③ HBM 메모리의 진화 (High Bandwidth Memory, 고대역폭 메모리)[4]

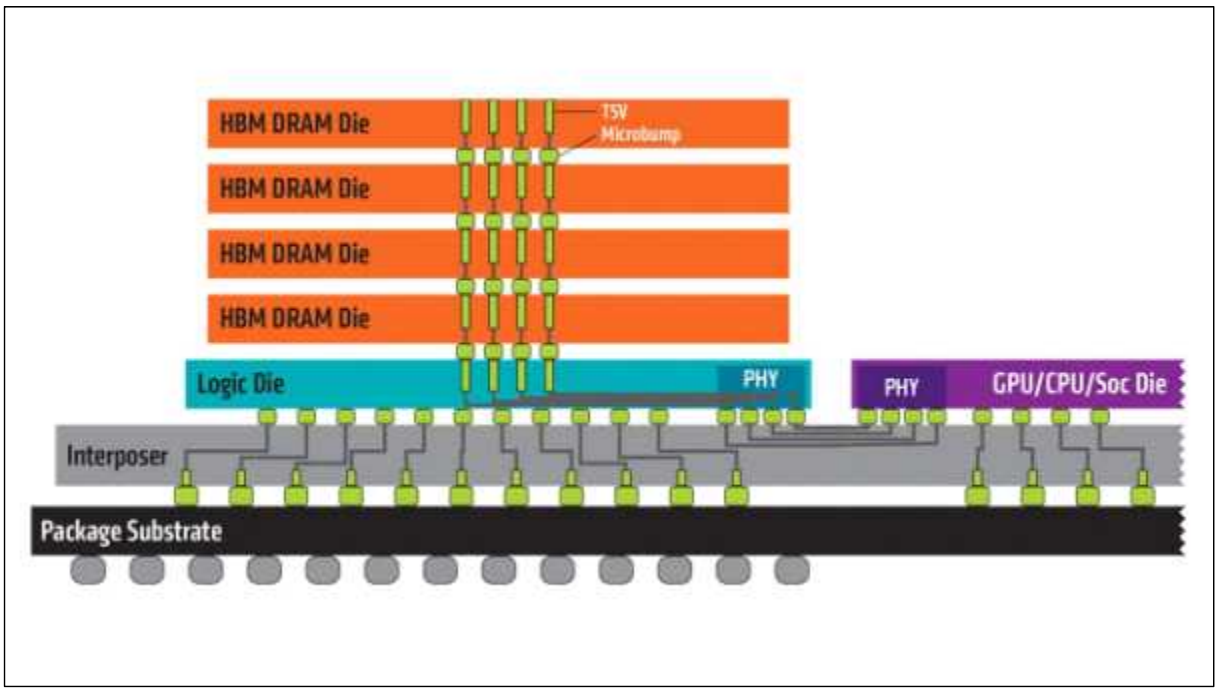
ㅇ 메모리 회사들은 읽기/쓰기 속도는 높이면서 전력 소모를 줄일 수 있는 개량된 구조의 메모리 개발에 역량 집중

- 메모리 속도 향상은 AI 빅데이터 처리속도를 개선하는 효과 제공

ㅇ HBM은 여러 개의 D램을 수직으로 연결해 접근속도를 높인 고성능 D램

- 3차원 적층 구조로 작은 공간에 많은 메모리를 배치할 수 있어 공간 효율 향상
- 메모리 대역폭[5]이 극대화됨에 따라 인공지능에서 요구되는 병렬연산의 속도를 개선하는 효과. 그래픽 처리 장치(GPU), 인공지능 가속기, 고성능 컴퓨팅(HPC) 등에서 고성능 연산에 널리 사용
- HBM은 AI의 추론과 학습 분야 모두에 활용될 수 있으나, 인공지능 학습 과정에서 GPU로 전달하는 데이터가 가장 크고 많은 특징으로 인해, 추론보다는 학습용 연산에 더 효과적

<HBM 개요(출처:AMD)>



4) https://news.skhynix.co.kr/
5) 메모리 대역폭(Bandwidth) : 메모리의 단위 시간당 입출력 데이터 교환량

ㅇ **HDM은 메모리 다이를 적층하여 실리콘을 관통하는 통로(TSV)를 통해 주 프로세서와 통신하는 방식**[6]

- 메모리를 적층하는 만큼 배선수가 증가하는 문제를 인터포저를 활용하는 패키징 기술로 해결
- HDM은 기존의 GDDR 계열 SGRAM을 대체하고, 더 넓은 대역폭의 메모리 성능을 달성하기 위해 제안되어, 2013년 반도체 표준협회인 JEDEC에서 채택
- 한국 SK하이닉스, 삼성전자, 미국의 마이크론이 치열한 글로벌 경쟁 중(2024년 현재 삼성전자, SK하이닉스, 마이크론 모두 12단 5세대 HBM 양산 또는 샘플 공급 단계)

④ PIM: 메모리와 컴퓨팅의 융합 (Processing-In-Memory)

ㅇ **프로세서를 메모리 내부에 분산 배치함으로써 데이터 이동 시 발생하는 정체 문제를 개선한 인공지능용 메모리**

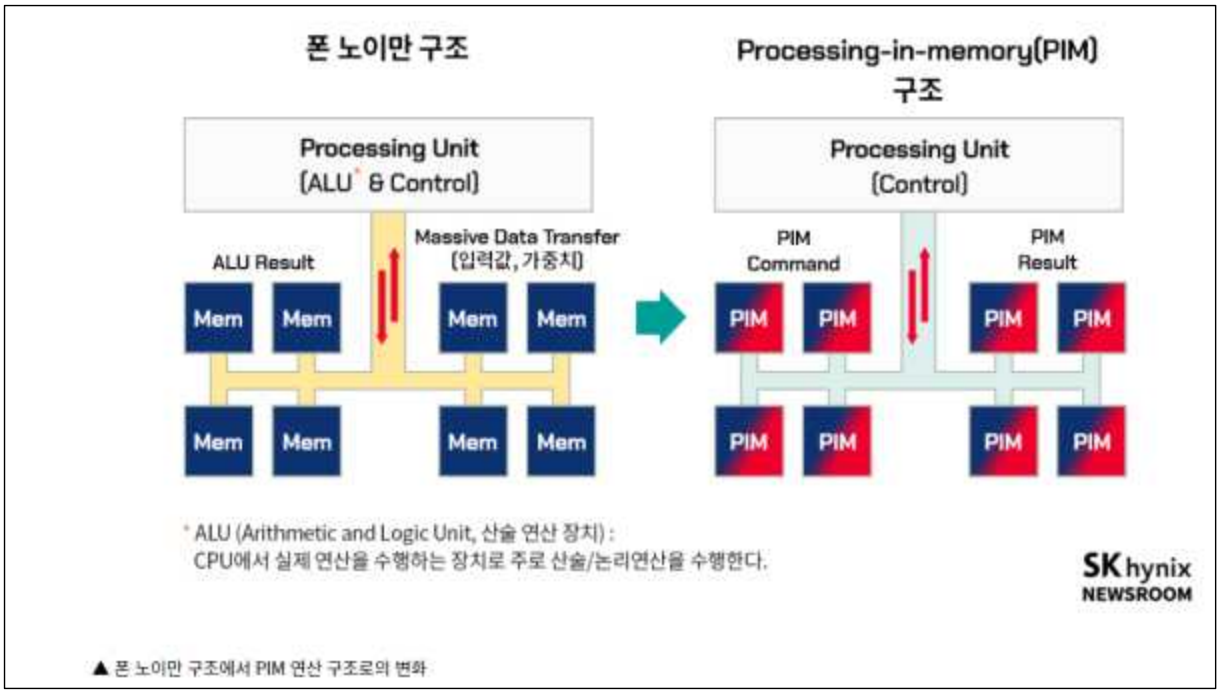
- PIM은 메모리 내부에서 연산하도록 작은 프로세서들을 메모리 내부에 분산 배치한 구조. 반면 GPU와 CPU는 메모리에서 데이터를 불러와 연산하는 폰노이만 구조 기반
- 메모리 영역에 연산장치가 분산 배치되고 연산장치와 메모리 간 물리적 거리가 가까워짐에 따라 병목현상 완화 효과
- PIM 내부에서 데이터를 빠르게 교환하며 연산을 수행하고, 연산 결과 정보만 외부 프로세서로 전달함으로써, 폰노이만 병목을 완화, 연산속도 향상, 줄어든 데이터의 이동거리 만큼 소모전력 절감 효과 제공

ㅇ **PIM은 인공지능 연산 중 추론 영역에 적합**

- 학습은 시간당 얼마나 많은 데이터를 처리하는지가 중요하고, 추론은 데이터를 적은 시간지연으로 빠르게 처리하는 것이 중요
- 메모리가 연산한 값을 신속히 처리할 수 있어 PIM은 AI 추론에 효과적
- 데이터 이동에 소비되는 에너지를 크게 절감하는 장점으로 인해 엣지 디바이스에 효과적

6) https://namu.wiki/w/HBM

<폰노이만 구조와 PIM 연산 구조(출처:SK하이닉스, https://news.skhynix.co.kr/post/dgist-series-1)>



▲ 폰 노이만 구조에서 PIM 연산 구조로의 변화

용어해설

- **폰노이만 아키텍처**: 1940년대 후반에 수학자 폰노이만이 제안하여 현재까지 대부분의 컴퓨터에서 사용되고 있는 컴퓨팅 방식. CPU가 공통 메모리에 있는 프로그램 명령어와 데이터를 버스를 통해 접근하는 구조로써, CPU는 프로그램의 명령어를 순차적으로 실행하고, 필요한 데이터를 메모리에서 읽고 처리하는 방식
- **폰노이만 병목 현상**: 메모리와 CPU가 동일한 데이터 버스를 공유하는 구조로 인해, CPU가 명령어와 데이터를 메모리에서 읽고 쓸 때 발생하는 병목현상. 이로 인해 실질적인 메모리 접근속도가 제한되는 결과 초래
- **뉴로모픽 컴퓨팅:** 인간 뇌의 구조와 기능을 모방하여 설계된 컴퓨팅 아키텍처의 하나로, 고도의 에너지 효율성과 자율성을 통해 기존 컴퓨팅 시스템을 대체할 것으로 기대되는 기술. 향후 자율주행차, 로봇, 의료 등 다양한 산업 분야에서 중요한 역할을 할 것으로 기대됨

⑤ 뉴로모픽 컴퓨팅 - 3세대 인공지능 기술

ㅇ 뉴로모픽 컴퓨팅의 원리는 인간 뇌의 신경세포(뉴런)와 시냅스 간의 통신 방식을 하드웨어적으로 모방하는 것

- 메모리 회사들은 메모리의 구조 개량을 통해 인공지능 연산속도는 높이면서 전력 소모를 줄이려고 노력하는 반면, 뉴로모픽 연구는 인공지능 연산을 위한 컴퓨팅 아키텍처를 인간의 신경망에 더 가깝도록 근본적인 구조 변경을 통해 극단적으로 에너지를 절감하고자 함
- 3세대 인공지능 기술로 간주되는 뉴로모픽 컴퓨팅은 현재 인간 신경망에 가장 가까운 것으로 알려진 스파이킹 신경망(SNN, Spiking Neural Network)을 활용하여 뉴런과 시냅스 기능의 시간적 요소를 물리적으로 추가로 복제하여 에너지 효율을 더욱 높이려는 시도
- 현재 널리 활용되는 인공지능 반도체는 심층 신경망(DNN)을 SRAM이나 플래시 메모리 같은 종래의 반도체 기술로 구현하여, 두뇌의 병렬처리와 계산 방식을 모방하는 개념
- 궁극적으로 두뇌와 유사한 레벨의 높은 에너지 효율을 갖는 컴퓨팅을 실현할 수 있다는 측면에서, 뉴로모픽 컴퓨팅은 향후 폭발적인 시장 성장을 이끌 것으로 예상됨

<뉴로모픽 컴퓨팅 주요 특징>

스파이크 신경망 적용	인간 뇌 속의 뉴런은 외부 자극을 받을 때만 신호(스파이크)를 출력. 이를 모방한 스파이킹 신경망(Spiking Neural Network, SSN)에서는 특정 조건이 만족 되었을 때만(예를 들어 전압 임계값을 넘었을 때만) 인공뉴런이 신호를 출력
비동기 신호 처리	전통적인 디지털 컴퓨팅에서 전자회로는 매 클럭 신호에 맞추어 지속적으로 동작하지만, 뉴로모픽 시스템은 비동기적으로 작동. 즉, 클럭에 의존하지 않고 이벤트가 발생할 때만 회로가 동작하므로, 극단적인 에너지 절감이 가능
메모리와 연산의 융합	전통적인 컴퓨팅 아키텍처에서는 데이터가 메모리에서 연산기로 이동하여 처리되지만, 뉴로모픽 시스템에서는 신경망 내에서 데이터를 저장하고 처리하는 기능이 결합되어 있어 데이터 이동에 소요되는 에너지를 크게 절감할 수 있음. 따라서 기존 아키텍처에서 빅데이터 처리 시 심화될 수 밖에 없는 메모리 병목현상에서 근본적으로 해방

○ 스파이킹 신경망 동작 원리[7][8]

- 현재의 뉴로모픽 시스템 또는 반도체는 스파이킹 신경망(SNN)으로 뇌를 모사

※ 스파이크란 인간 뇌 속에서 뉴런과 뉴런이 신호를 주고 받을 때 사용하는 전기 신호로 뉴로모픽 칩의 인공 뉴런들도 스파이크 신호로 동작

- 스파이크의 강도는 신호의 빈도와 타이밍으로 표현되며, 뇌와 유사하게 빈도와 타이밍에 따라 시냅스간 연결을 강화하거나 약화시킴

※ 스파이크들의 간격이 좁을수록, 즉, 잦은 입력 스파이크로 인해 결과 스파이크가 빠르게 나타날수록 연관성이 크다고 인식되어 뉴런간 연결이 강화됨

- 스파이크 타이밍 의존 가소성을 이용한 학습은 스파이크 인공신경망에서 널리 사용되는 학습 알고리즘 중 하나

스파이크 타이밍 의존 가소성(STDP, Spike Timing Dependent Plasticity)

- 학습을 통해 시냅스 연결이 강화되거나 약화되는 특성을 시냅스 가소성이라 함
- STDP는 뉴런들이 전달하는 신호의 변화에 따라 연결이 재구성되는 원리로써, 학습과 기억을 가능하게 하는 SNN의 중요 메커니즘
- 스파이크의 시간 간격이 짧으면 시냅스 결합이 강해지고, 길면 약해짐
- STDP는 뉴런들의 연결을 강화하거나 약화시켜 정보를 장기기억으로 전환하거나 빨리 잊도록 만들 수 있음

- 뉴런에서 스파이크가 발생하는 상대적인 시간 간격이 연결의 강화와 억제에 영향

짧은 시간 간격 -> 연결 강화	긴 시간 간격 -> 연결 약화
시냅스 전 스파이크가 시냅스 후 스파이크보다 먼저 발생하면, 시냅스 전 뉴런의 신호가 시냅스 후 뉴런의 신호에 영향을 준 것으로 간주하여 시냅스 연결 강화	시냅스 전 스파이크가 시냅스 후 스파이크 이후에 발생하면, 연관이 적다고 인식하여 시냅스 연결 억제

7) https://post.naver.com/viewer/postView.nhn?volumeNo=30387060&memberNo=10728965&vType=VERTICAL
8) TTA 저널 204호, 2022, 11/12월호 권오현 뉴로모픽

- 전통적인 신경망과 달리 SNN은 빈도와 타이밍으로 동작하므로, 인공지능 학습과 추론 과정 또한 기존 인공지능 시스템과 차이가 있음

차이점	SNN	DNN
신호 전달 방식	뉴런은 입력값이 특정 임계값을 넘을 때만 '스파이크' 출력 신호 발생	인공신경망 내의 모든 뉴런이 활성화되어 연속적으로 신호를 전달
에너지 효율	입력값이 임계값을 초과하는 뉴런만 활성화되므로 에너지 절감 효과 큼. 이러한 뉴런 특성은 하드웨어로 구현 가능 (e.g. 멤리스터 소자)	모든 뉴런이 매 순간의 계산에 참여하므로 에너지 소비가 큼

ㅇ SSN 기반 뉴로모픽 컴퓨팅의 기술적 난제[9)]

- SSN 기반 뉴로포픽 컴퓨팅을 효율적으로 구현하는 데에 적합한 최적의 하드웨어 플랫폼을 만들기 위한 수십 년간의 연구에도 불구하고, 아직 해결할 기술적 난제 존재
- 하나의 디바이스로 뉴런과 시냅스 동작을 모사할 수 있는 뉴모로픽 반도체를 구현하는 것이 이상적이나, 추가적인 인터페이스 회로를 필요로 하지 않는 실용적인 구현 방법이 없어서(면적 및 에너지 효율 저하 측면) 아직은 디지털 회로 기반으로 구현하는 방법이 대세
- 현재까지 상업적으로 가장 성공적인 뉴로모픽 칩은 인텔의 Loihi와 IBM의 TrueNorth로 기존의 디지털 회로 기반 방식으로 일반적인 CMOS(상보성 금속 산화물 반도체) 기술로 구현

※ 디지털 회로 방식은 디지털 회로가 제공하는 노이즈 내성 및 손쉬운 프로그래밍이 가능하다는 장점 덕분에 대량 생산이 상대적으로 용이하고, 소형 폼팩터로 수십억 개의 신경 시냅스 소자를 단일 칩으로 집적 가능

- 디지털 기술을 활용한 뉴로모픽 컴퓨팅의 구현은 고밀도, 초전력효율, 고성능 인공지능 시스템 구현에 적합하지만, 높은 오프상태 누설전류와 저하된 부 임계값 스윙 특성은 에너지 효율 추가 개선의 걸림돌

9) https://www.nature.com/articles/s41467-024-46397-3

- 멤리스터를 활용한 아날로그 구현 방식이 이를 개선할 것으로 기대되지만 아직은 초기 연구 단계

뇌의 작동 원리[10]

- 인간의 뇌 신경계에서는 외부 자극이 뉴런의 수상돌기에 입력되면, 뉴런은 이를 전기 신호로 변환하여 축색돌기 말단으로 전달
- 축색돌기 말단에서는 자극의 세기에 비례하여 신경전달물질 분비
- 이 신경전달물질은 시냅스를 통해 다른 뉴런의 수상돌기에 있는 수용기(인체가 외부 자극을 받아들이는 기관과 세포를 통칭)로 전달
- 이 과정에서 전기 신호는 신경전달물질에 의해 화학적 신호로 변환되어 전달되며, 다른 뉴런의 수용기에서 다시 전기 신호로 환원

☞ ***뉴런**: 전기를 발생시켜 다른 세포에 정보를 전달하는 신경계의 단위. 핵이 있는 신경세포체, 다른 뉴런으로부터 신호를 받는 수상돌기, 그리고 다른 뉴런에 신호를 주는 축색돌기로 구성*

☞ ***신경전달물질**: 뉴런에서 다른 뉴런으로 신호를 전달하기 위해 분비되는 화학 물질*

☞ ***시냅스**: 한 뉴런의 축색돌기 말단과 다음 뉴런의 수상돌기가 만나는 부분으로, 뉴런과 뉴런 사이에서 신호를 전달하는 역할 수행*

⑥ 대표적인 뉴로모픽 컴퓨팅 시스템 및 반도체

ㅇ 인텔 로이히(Loihi)

- 인텔은 2017년 테스트용 뉴로모픽칩 '로이히(Loihi)' 공개

- 128개의 컴퓨팅 코어로 구성
- 각 코어에는 1024개의 인공 뉴런이 있어 13만 개 이상의 뉴런과 1억 3,000만 개의 시냅스 연결 가능
- 바닷가재의 뇌보다 조금 더 복잡한 수준으로 알려짐

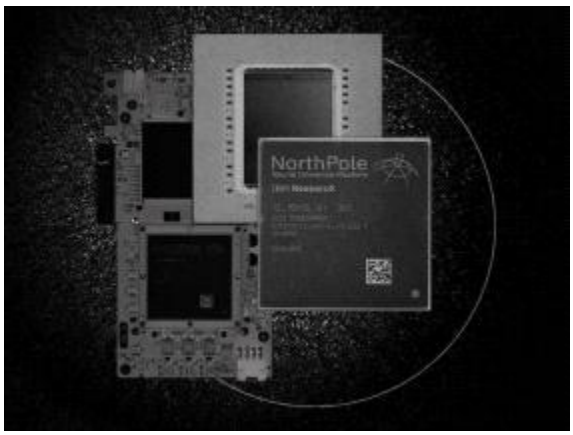
10) https://news.skhynix.co.kr/post/skirmion-based

- 2021년에 발표된 로이히2 (일명 Pohoiki Beach)

- 전 모델보다 약 8배 높은 칩당 최대 100만개의 뉴런을 장착하여 처리 속도를 최대 10배 개선
- 64개의 코어를 장착하고, 더 빠른 속도, 확장성 향상을 위한 고대역폭 칩 간 통신, 칩당 용량 증가, 더 작은 크기, 향상된 프로그래밍 기능 등이 특징
- 데이터 저장과 연산에 들어가는 전력 소모도 줄여 높은 에너지 효율성 달성 (4나노 공정)
- 뉴로모픽 연구를 협업할 수 있는 공통 소프트웨어 프레임워크인 '라바 소프트웨어 프레임워크'와 함께 조기 상용화 시도

<인텔 로이히2
(출처: 인텔, http://intel.com)>

<IBM NorthPole (출처:IBM,
https://research.ibm.com/blog/northpole-ibm-ai-chip)>



ㅇ IBM 뉴로모픽칩 Truenorth

- IBM은 2014년 S램 기술을 활용해 인간의 뇌를 모방 한 트루노스칩 발표

- 미국 방위고등연구계획국(DARPA)의 시냅스(SyNAPSE) 프로젝트의 일환
- 4,096개의 뉴로코어(Neurocore)는 독립적으로 작동하면서 다른 뉴로코어와 연결되어 복잡한 병렬연산 가능
- 100만 개의 뉴런과 2.56억 개의 시냅스로 구성. 소모전력 70mW
- 신경 시냅스 코어는 완전히 비동기 방식으로 상호 연결. 본 신경망은 전송할 스파이크(이벤트)가 있을 때만 활성화되는 이벤트 기반 동작 방식. 따라서 SpiNNaker와 유사하게 글로벌 클럭 사용 불필요
- 대부분의 뉴로모픽 칩과 유사하게, 본 칩은 인메모리 컴퓨팅 아키텍처로 구성되어, 중앙 메모리와 중앙 처리 장치가 없는 대신, 저장 및 연산 회로가 골고루 분산된 탈 폰노이만 구조

- 2023년 공개된 IBM NorthPole 칩은 IBM의 2014 트루노스 시스템의 접근 방식과 최신 하드웨어 설계를 결합하여 트루노스보다 약 4,000배 빠른 속도를 달성

• 224MB의 RAM과 256개의 프로세서 코어 포함 • 8비트 정밀도에서 코어당 2,048회, 2비트 정밀도에서 8,192회의 연산 수행 가능 • 이 칩은 추론용으로 활용 가능하고, GPT-4는 처리 불가한 수준으로 알려짐

ㅇ 맨체스터 대학 SpiNNaker

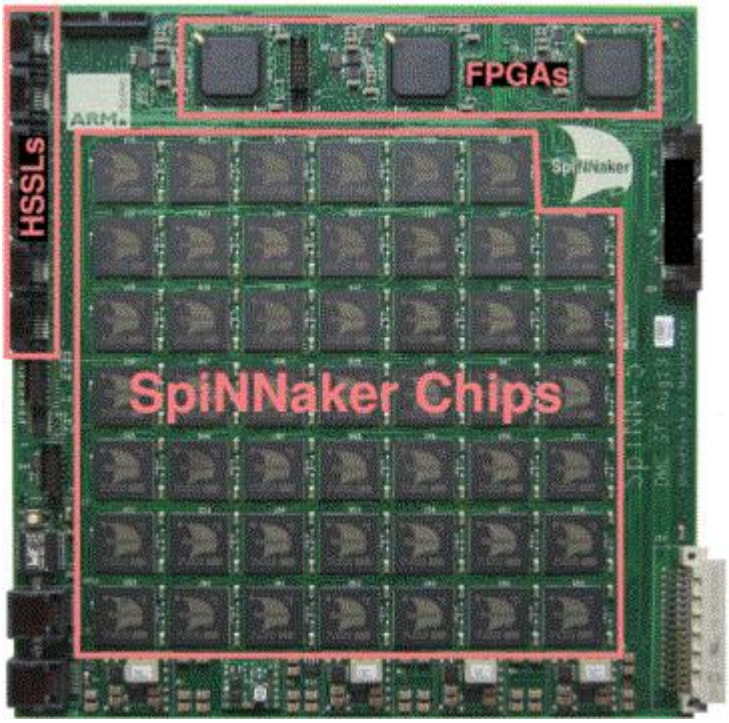
- SpiNNaker (Spiking Neural Network Architecture)는 생물학적 뇌의 신경망을 대규모로 시뮬레이션할 수 있는 뉴로모픽 컴퓨팅 플랫폼

• 영국 맨체스터 대학교 Steve Furber 교수 주도로 차세대 프로세서 기술 연구팀이 설계한 스파이크 신경망 기반 슈퍼컴퓨팅 아키텍처 • 10개의 19인치 랙으로 구성되며, 각 랙에는 100,000개 이상의 코어 장착 • 각 코어는 1,000개의 뉴런을 에뮬레이션. 최종 목표는 최대 10억 개의 뉴런 집합체의 동작을 실시간으로 시뮬레이션하는 것 • 이 시스템은 에어컨 환경에서 240V 100kW 전력 소모

- SpiNNaker는 휴먼 브레인 프로젝트의 뉴로모픽 컴퓨팅 플랫폼으로 활용 중

• 2011년 첫 번째 시스템 완성 이후 여러 단계의 개선을 거쳐 2018년에는 100만 개의 코어를 포함하는 대규모 시스템으로 업그레이드 • 하나의 칩으로 800만 개의 가소성 시냅스가 있는 16,000개의 뉴런을 1W의 전력으로 실시간 시뮬레이션 가능 • 2019년 휴먼 브레인 프로젝트는 2세대 머신(SpiNNcloud) 건설을 위한 800만 유로의 보조금을 드레스덴 공과대학교에 지급 • SpiNNaker의 후속 시스템인 SpiNNaker2 칩은 드레스덴 공과대학교의 스핀오프 회사인 SpiNNcloud Systems GmbH를 통해 2021년에 공개 • 독일 연방 혁신 기관인 SPRIN-D와 유럽 혁신 위원회(EIC)로부터 자금 지원과 함께 BMW, 인피니언 등 글로벌 기업의 지원을 확보하여 뉴로모픽 컴퓨팅기법을 산업에 적용하는 시범 프로젝트 진행 중

<SpiNNaker Chips>



(출처: https://ieeexplore.ieee.org/document/9079810/figures#figures)

ㅇ 하이델베르그 대학 BrainScales

- BrainScales는 하이델베르크 대학교에서 Karlheinz Meier 교수 주도로 개발된 뉴로모픽 컴퓨팅 플랫폼

• 2016년 발표된 BrainScales 시스템(NM-PM-1)은 20개의 실리콘 웨이퍼로 구성되며, 각 웨이퍼에는 4천만 개의 플라스틱 시냅스와 20만 개의 뉴런이 통합

- BrainScaleS-2는 2022년에 출시된 차기 뉴로모픽 컴퓨팅 시스템으로, 아날로그 및 디지털 신호를 결합한 혼합 신호 아키텍처를 기반

• 약 512개의 뉴런과 130,000개의 시냅스를 칩에 집적 • 온칩 학습 기능과 저전력 소비가 특징

ㅇ 취리히 대학 Dynap

- DYNAP는 ETH 취리히와 취리히 대학교의 뉴로모픽 연구 성과를 활용하여 SynSense가 2017년부터 개발해온 뉴로모픽 컴퓨팅 시스템

※ DYNAP: Dynamic Neuromorphic Asynchronous Processor

- 아날로그와 디지털 방식을 혼합한 구조로 개발된 SNN 기반 뉴로모픽 칩
- 2019년 첫 상용품인 DYNAP-CNN 출시
- 1백만 개의 스파이킹 뉴런 지원
- 저전력 및 저지연 신경망 처리를 목표로 설계
- 자율주행, 드론, 스마트 보안 등 실시간 비전 처리 분야에 응용 가능
- 기존 방식보다 100~1,000배 더 낮은 전력으로 작동.
- 이벤트 기반으로 연산을 처리하고, 5ms 이하의 짧은 지연 시간이 특징
- DYNAP-CNN은 비전 센서와 통합되어 실시간 영상 데이터 처리 지원

- 2023년 DYNAP-CNN2 칩 출시

- 저전력 소모와 저지연 신경망 처리 측면에서 개량된 성능 제시

⑦ 맴리스터 - 뉴로모픽 반도체 차세대 기술[11)12)]

ㅇ 뇌의 학습방식과 멤리스터의 관계

- 뇌의 각 뉴런에는 들어오는 연결이 다수 있고 나가는 연결이 하나 존재
- 들어오는 시냅스는 뉴런에서 합산된 신호가 수신되며, 이 값이 임계값에 도달하면 출력 시냅스가 발화
- 다만, 모든 시냅스가 똑같이 기여하는 것은 아니고, 뉴런의 발화에 대한 기여도를 결정하는 가중치가 있으며, 이러한 가중치는 시냅스를 통과하는 앞선 신호의 영향을 받아 시간에 따라 변하는 것이 뇌가 학습하는 방식

11) https://www.imec-int.com/en/articles/what-if-we-could-design-chips-that-behave-like-brains
12) 인공지능 뉴로모픽 반도체 기술 동향, 전자통신동향분석 35권 제3호, 한국전자통신연구원, 2020

- 따라서 시냅스 동작을 구현해줄 새로운 전자소자가 필요
- 멤리스터는 이전 경험에 따라 저항을 아날로그 방식으로 변경하여 전류를 전달하는 역할을 수행하는 새로운 전자소자

ㅇ 뉴로모픽 반도체 2세대 기술

- 뉴로모픽 반도체는 실리콘 기반 CMOS 트랜지스터 기술만을 이용하는 1세대를 지나 멤리스터 소자를 활용하는 2세대로 진화 중
- 1세대에서는 기존의 CMOS 메모리 소자를 활용하여 시냅스의 가중치를 저장하고 읽어오는 방식으로 구현

※ IBM의 TrueNorth 칩의 경우 CMOS의 SRAM을 시냅스로 사용하며, 54억 개의 트랜지스터를 사용하여 100만 개의 뉴런과 2억 5,600만 개의 시냅스 내장

- 뇌에서 일어나는 신호전달 과정을 기존의 반도체 소자 CMOS로 구현하는 것은 소모전력과 집적도 측면에서 한계에 직면

※ CMOS 메모리 소자 대부분은 휘발성 메모리인 SRAM을 사용하나, 비휘발성 메모리 사용으로 에너지 절감 가능

- 시냅스 역할을 충실히 수행할 뉴로모픽 소자, 멤리스터

• 전원 없이도 정보가 소멸하지 않는 비휘발성 특성을 가지면서, 여러 단계의 시냅스 신호 강도를 표현할 수 있어야 하며, 시냅스가 실제 수행하는 학습방식을 쉽게 구현할 수 있어야 함 • 이 소자는 인가되는 전압에 따라 저항값이 변화하는 트랜지스터의 특성과 함께 저항값을 유지하는 메모리 특성도 지녀야 함 • 이런 특성을 갖는 소자가 멤리스터(Memristor)이며, 이는 메모리(Memory)와 저항(Resistor)의 합성어

ㅇ 멤리스터 연구 동향[13]

- 시냅스의 핵심 특성을 높은 집적도로 구현할 수 있으면서, 메모리와 가변 레지스터 두 가지 특성을 동시에 지니는 멤리스터(Memristor) 소자를 활용하는, 인공지능 컴퓨팅 방식에 대한 연구가 활발
- 맴리스터는 기존 SRAM과 달리 간단한 소자(금속/산화물)로 구성되는 $10nm^2$ 이하의 매우 작은 소자로, 변화된 저항 특성을 유지(저장)할 수

13) https://news.skhynix.co.kr/post/skirmion-based

있어, 시스템 레벨에서 더 높은 집적도로 시냅스 구현이 가능

- 멤리스터라 불리는 시냅스 모방 소자로 FRAM(Ferroelectric RAM), MRAM(Magnetic RAM), PRAM(Phase-change RAM), RRAM(Resistive RAM) 등이 구조가 간단하면서 저전력 및 고집적 인공 시냅스로 구현이 가능하다는 사실이 알려지면서, 2세대 뉴로모픽 반도체용 소자 연구가 활발

시냅스 모방 소자기술[14]

(1) PRAM(Phase-change memory, 상변화 메모리)

- DRAM과 NAND의 장점을 모두 가진 차세대 메모리로 속도 빠르고 비휘발성
- 기록층 물질의 결정상(Crystalline state)과 비정질상(Amorphous state)의 결합 상태 변화를 통해 0과 1을 저장하는 소자
- 기록 물질로는 Ge-Sb-Te 등의 칼코겐 화합물을 주로 사용
- 결정상과 비정질상이 광학적 반사도 및 전기 저항에서 차이가 나는 성질을 이용해 정보 구분

(2) RRAM (Resistive Random Access memory, 저항변화 메모리)

- 재료에 전압을 인가하여 변경되는 저항을 저장하는 비휘발성 메모리
- 저항변화 메모리는 산화물 재료 내 산소 이온이나, 이온이 빠진 빈 공간을 의미하는 산소간극(Oxygen vacancy)과 같은 결함들의 움직임에 의한 저항 변화 특성을 이용하는 소자
- 인가전압에 의해 재료는 고저항 또는 저저항 상태로 천이. 이를 저장 데이터 0과 1로 간주
- 인가전압에 따라 국부적으로 만들어지는 물질의 산소 간극 필라멘트는 절연체의 산화물 내 전자가 흐를 수 있는 길을 만들어 저항의 변화가 나타나는 원리

(3) FRAM (Ferroelectric random access memory, 강유전체 메모리)

- 강유전체는 강유전성(Ferroelectric)을 가진 재료를 의미
- 강유전체 물질은 전기장을 가하면 그 안의 전기 쌍극자가 정렬되는데, 쌍극자 정렬은 전기장이 제거된 후에도 일정 시간 동안 유지될 수 있는 성질 보유
- 전기적 분극을 유지한다는 것은 극성을 바꿔 데이터를 저장할 수 있다는 의미로써, 극성에 따라 0과 1로 정보 매핑

(4) MRAM (Magnetic memory, 자성 메모리)

- 자성 메모리는 N극과 S극으로 구분된 자석의 특성을 이용한 메모리로, 자기 터널 접합(MTJ, Magnetic Tunnel Junction) 원리 이용

- 절연체로 구분된 2개의 자석이 서로 평행한 방향으로 정렬되면 저항값이 작고, 반대 방향으로 정렬되면 저항이 높아지는 성질을 이용해 0과 1 구분
- 자성 메모리는 원리상 저항값이 두 가지로만 구분되고 저항값 차이가 미미하여, 높은 전류를 이용해야 하는 특징으로 인해 저전력 설계에 적합하지 않은 한계. 최근 전력소모를 기존의 방식보다 1/10 수준으로 감소시킨 소자가 개발되어 연구 활발

- 한 예로, 전자의 회전(Spin)성질을 이용한 자기 메모리(Magnetic Memory) 소자는 비휘발성이면서 고속 정보 입출력이 가능. 또한, 높은 수준의 집적이 가능하며, 뉴로모픽 소자에 요구되는 가소성, 선형성, 대칭성 측면에서 유리

※ 스커미온을 이용한 시냅스 소자 연구가 MRAM 연구의 한 사례

- 멤리스터 연구는 PIM 분야의 차세대 메모리 연구분야와 유사하여 동반 진화 중
- 2세대 뉴로모픽 반도체 연구는 한국, 미국, 유럽, 중국을 포함하여 전 세계적으로 활발히 전개 중

※ 현재까지 대부분은 단위 기능 블록 수준의 연구로 진행되어 왔으나, 최근에는 시스템 수준에서 구현 가능성을 테스트하는 방향으로 발전 중

용어해설

- **가소성**: 고체가 외부에서 탄성 한계 이상의 힘을 받아 형태가 바뀐 뒤 그 힘이 없어져도 본래의 모양으로 돌아가지 않는 성질. 인공지능 분야에서는 학습을 통해 시냅스 연결이 강화되거나 약화되는 특성을 시냅스 가소성 이라 함
- **선형성**: 어떤 양의 변화가 다른 양의 변화에 비례적인 변화를 가져오는 성질
- **대칭성**: 시냅스 가중치의 증가량과 감소량이 대칭적으로 같아야 하는 성질. 대칭적이지 않을 경우, 특정 학습 방향에서 치우친 결과가 발생
- **스커미온(Skyrmion)**: 소용돌이 모양으로 스핀들이 배열되어 형성되는 스핀 구조체

14) https://renewableenergyfollowers.org/4485

Ⅲ. 유럽 기관별 연구동향

□ 개요

ㅇ **차세대 인공지능 반도체인 뉴로모픽 반도체 연구는 2000년대 중반부터 유럽과 미국 등에서 원천기술 확보 목적으로 국가 주도 R&D 사업으로 전개**[15)]

- 특히, EU는 인간 뇌에 관한 대규모 원천연구(Future and Emerging Technologies (FET) 플래그쉽 프로젝트)인 Human Brain Project(HBP)를 2013년부터 2023년까지 10년간 10억 유로를 투자하여 진행
- HBP 투자 덕분에 유럽에 조성된 관련 연구 생태계로 인해 차세대 인공지능 반도체 연구가 활발히 진행 중

ㅇ **본 고에서는 유럽 주요국의 대학과 연구소, 그리고 EU의 H2020 및 Horizon Europe 등 유럽 내 연구 프로젝트에서 진행된 인공지능 반도체 연구 현황을 소개**

<목차>

1. 벨기에 IMEC
2. 프랑스 CEA-Leti
3. 프랑스 GML
4. 스위스 취리히 대학
5. 영국 맨체스터 대학
6. 독일 하이델버르그 대학
7. 독일 프라운호퍼 IPMS
8. 독일 TU Dresden
9. 독일 막스플랑크 연구소
10. 프랑스 UPMEM

15) 인공지능 뉴로모픽 반도체 기술 동향, 전자통신동향분석 35권 제3호, 한국전자통신연구원, 2020

1 벨기에 IMEC

imec

- imec(Interuniversity Microelectronics Centre)은 벨기에 루벤에 본사를 두고 나노전자공학 및 디지털 기술 분야에서 활동하는 국제적 연구혁신 기관
- 1984년 설립 이후, 글로벌 협력 생태계 에코시스템 활용한 최첨단 R&D 인프라와 5,500여 명의 직원으로 5G 통신 및 센싱 기술을 넘어 시스템 스케일링, 실리콘 포토닉스, 인공지능 등 첨단반도체 연구개발
- 인텔, 삼성, SK하이닉스, 퀄컴, TSMC 등 세계 유수의 기업들과 프로토타입 개발 및 차세대 기술 연구 협력 중이며, 최첨단 반도체 제조사들이 사용하는 세계 최고 수준의 EUV(극자외선) 노광장비를 제조하는 ASML과 긴밀한 협력관계

ㅇ 뉴로모픽 컴퓨팅 아키텍처 SENeCA 발표[16)17)]

- imec은 2022년 인공지능 회로 및 시스템(AICAS) 학회에서 뉴로모픽 컴퓨팅 아키텍처 SENeCA 소개

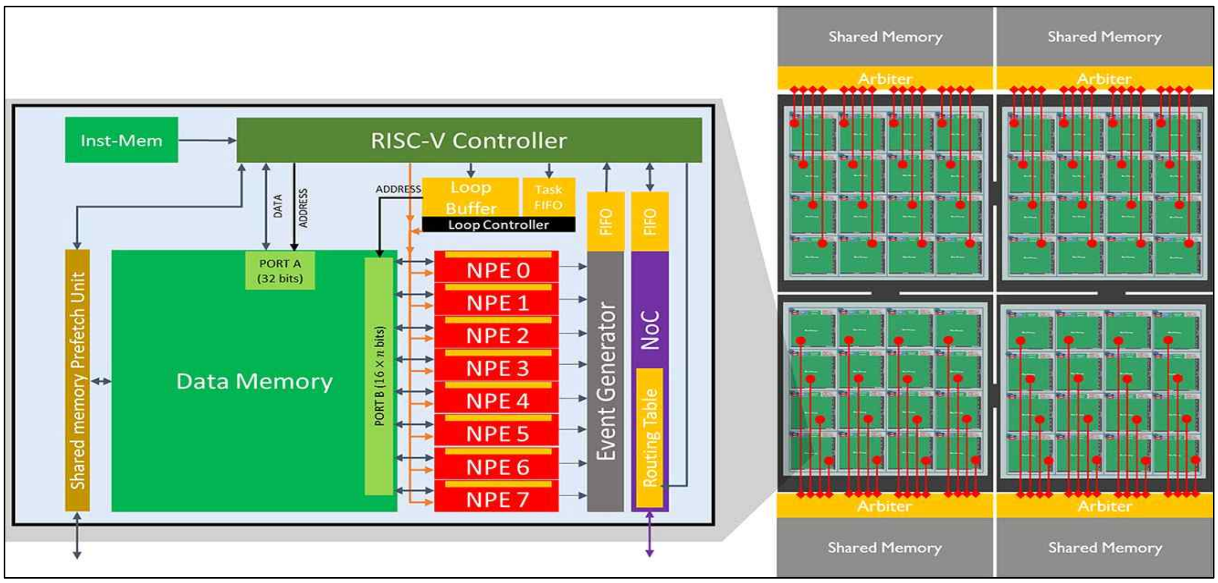
- SENeCA는 최초로 RISC-V를 기반으로 하는 스파이킹 신경망 구조의 디지털 뉴로모픽 프로세서
- 오픈소스 RISC-V 소형 프로세서를 컨트롤러 용도로 사용(이벤트 처리 용도 아님)하고, 최적화된 가속기와 낮은 오버헤드 메시형 멀티캐스팅 NoC (Network-on-Chip)를 사용하여 이벤트 기반 연산을 병렬로 수행
- SENeCA에는 상호 연결된 뉴런 클러스터 코어와 RISC-V 기반 명령어 세트, 최적화된 뉴로모픽 코프로세서, 이벤트 기반 통신 인프라가 포함됨
- 하나의 SENECA 코어는 22nm 공정에서 0.47mm^2 크기이며, 시냅스 동작당 2.8 pJ의 에너지 소모

- 두뇌 시뮬레이션을 위해 설계된 SpiNNaker[18)]와 달리, SENeCA의 개발 목적은 하드웨어와 소프트웨어를 모두 개방하여 엣지용 인공지능 연산의 최적화와 혁신을 도모하는 것

※ 현재 SENeCA 플랫폼은 교육 목적으로 개방

16) DOI: 10.1109/AICAS54282.2022.9870025
17) https://doi.org/10.3389/fnins.2023.1187252
18) 영국의 맨체스터 대학교 주도로 개발된 뉴로모픽 컴퓨팅 연구용 플랫폼의 하나

<SENeCA 구조 (출처:doi 10.3389/fnins.2023.1187252)>



○ 아날로그 방식의 추론 가속기 개발[19)]

- imec은 전문 파운드리 업체 글로벌파운드리(GF)와 함께 인공지능 칩 발표 (2020년 7월)

• 신경망의 가중치를 아날로그로 처리하기 위해 멤리스터 소자 활용 • 스파이크 신경망 모델에는 수천만 또는 수억 개의 가중치가 있을 수 있으므로 메모리와 프로세서 간에 데이터를 주고받는 것은 비효율적 (폰노이만 병목현상 발생) • 아날로그 컴퓨팅은 메모리 어레이를 사용하여 가중치를 저장하고 MAC(곱셈-누적) 연산도 수행하므로, 메모리와 프로세서간 정보교환 불필요 • 각 멤리스터 소자(예: ReRAM 셀)는 필요한 가중치에 비례하는 아날로그 레벨로 저항값 표현

- 메모리 셀에서 아날로그 연산을 수행함으로써 폰 노이만 병목 현상을 제거한 AiMC(Analog in Memory Computing) 아키텍처를 기반으로 하는 22FDX 칩은 아날로그 추론 가속기(AnIA)를 내장

• 아날로그 방식의 인메모리 컴퓨팅 구조에서 심층 신경망 연산수행에 적합한 구조 • ReRAM 계열의 멤리스터 소자 적용 • 벡터 행렬 곱셈을 아날로그 기술로 다소 낮은 정밀도로 수행해도 신경망에서는 비교적 정확한 결과를 얻을 수 있음 • 대규모 벡터 행렬 연산에 의존하는 신경망에서 디지털 컴퓨팅은 정확도는 높지만 폰 노이만 병목으로 인해 연산속도에 제한이 있고 에너지 소모가 큰 문제

19) https://gf.com/gf-press-release/imec-and-globalfoundries-announce-breakthrough-ai-chip-bringing-deep-neural-network/

- 1초 1와트당 2,900 테라연산(TOPS/W)의 기록적인 에너지 효율. 저전력 엣지 디바이스에서 추론기능 수행 용도로 적극 활용 가능

※ 일반적으로 데이터센터에서 머신러닝으로 구동되던 소형 센서 및 저전력 엣지 장치의 패턴 인식 기능을, 로컬 엣지에서 수행 가능

- imec의 머신러닝 프로그램 디렉터인 Diederik Verkest는 AnIA의 성공적인 테이프 아웃은 메모리 내 아날로그 컴퓨팅(AiMC) 기술 검증을 향한 중요한 진전이며, 아날로그 인메모리 계산이 실현 가능함을 보여줄 뿐만 아니라, 디지털 가속기보다 10~100배 더 나은 에너지 효율 달성 가능함을 강조

ㅇ 스파이크 순환 신경망 칩 개발[20)]

- 2020년 4월, imec은 스파이크 순환 신경망을 사용하여 레이더 신호를 처리하는 칩 발표

• 저전력의 멀티 센서 인식 시스템으로 드론의 장애물 식별 속도를 높이는 솔루션 • 기존 구현 방식보다 최대 100배 적은 전력을 소모하는 반면, 대기 시간은 10배 감소(드론의 의사 결정 속도 개선 효과) • 이 칩은 온전한 '이벤트 기반' 디지털 아키텍처 기반이며, 저비용 40nm CMOS 기술로 구현 • 이벤트 기반 처리를 위해 로컬 온디맨드 오실레이터를 사용하여 글로벌 클록 사용 배제 • 별도의 메모리 블록을 활용하지 않는 대신 메모리와 연산기가 같은 영역에 배치되어, 데이터 액세스 시간 지체와 에너지 오버헤드 문제 경감

- 네덜란드 수학 및 컴퓨터 과학 국립 연구소(CWI)와의 연구에 따르면, 적응형 임계값을 가지는 스파이크 뉴런을 훈련하여 최고 수준의 추론 정확도를 달성할 수 있음

- imec과 CWI가 수행한 연구는 적응형 임계값을 가지는 뉴런을 사용하는 SNN을 다른 여섯 개의 신경망과 비교한 결과, 적응형 임계값이 있는 뉴런을 사용하는 SNN이 에너지를 적게 소비하면서 추론 정확도가 높다는 의견

20) https://www.imec-int.com/en/articles/imecs-snn-chip-combines-low-latency-energy-consumption-high-inference-accuracy

2 프랑스 CEA-Leti

ㅇ 맴리스터를 활용한 뉴로모픽 칩 설계[21)22)]

- 프랑스 CEA-LETI는 유명 학술지 Nature electronics에 2022년 CEA-List와 프랑스 국립과학연구센터(CNRS)와 공동 발표한 'A memristor-based Bayesian machine' 논문을 통해, 실제로 활용이 가능한 멤리스터 기반 베이지안 신경망을 구현했다는 발표로 주목받음

• 제한적인 데이터만으로 효과적인 추론을 가능하게 하는 기존 Baysian machine 설계는 많은 컴퓨팅 자원을 필요로 하여, 저전력 로컬 디바이스에서는 구동이 어렵다는 문제 • 베이지안 추론기법은 일반 신경망과 달리 멤리스터 기반 아키텍처로 변환이 어렵다는 문제 • 이러한 문제를 해결하기 위해 저항기반 메모리(멤리스터)를 활용한 새로운 Baysian machine 설계를 도입하여, 적은 에너지로 연산을 수행하는 엣지 AI 기능을 성공적으로 실현

- 제안된 아키텍처는 분산된 메모리와 베이즈 법칙을 활용한 확률적 컴퓨팅의 원리를 이용

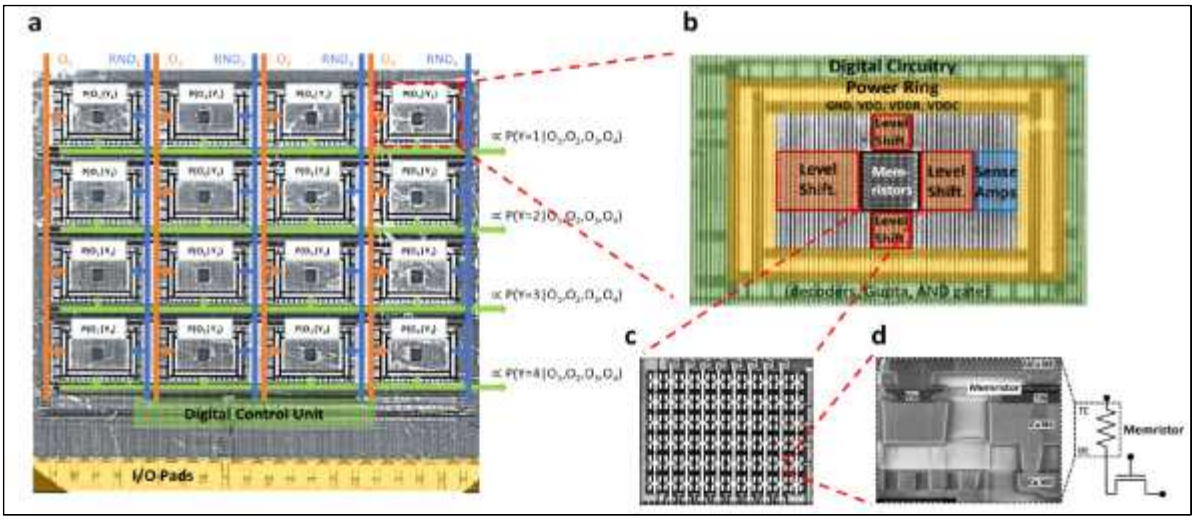
※ 제안된 베이지안 머신은 분산 배치된 멤리스터에서 로컬 계산을 수행함으로써 에너지 이동을 최소화하여, 일반 마이크로 컨트롤러 장치보다 세 배 이상 높은 에너지 효율로 베이지안 추론 가능

- 제작된 칩은 하이브리드 CMOS/멤리스터 공정을 사용하여 2,048개의 멤리스터와 30,080개의 트랜지스터를 통합한 프로토타입
- 기술성숙도가 높은 CMOS 기술과 저항기반 메모리 혹은 2D 물질의 장점을 동시에 취할 수 있는 이기종 통합 기술을 활용한 뉴로모픽 시스템 연구 또한 많은 주목을 받음

베이지안 추론
▪ 베이지안 추론은 베이즈 정리를 바탕으로 데이터나 관찰된 증거를 통해 확률을 갱신하는 방법론 ▪ 인공지능 분야에서 베이지안 추론은 불확실성을 감수하고 데이터에 기반한 의사결정을 내리는 데 중요한 역할

21) https://k-erc.eu/2024/04/publication/18602/
22) https://www.nature.com/articles/s41928-022-00886-9

<Memristor-based Bayesian machine (출처: CEA-LETI, DOI:10.48550/arXiv.2112.10547)>



○ 비휘발성 메모리 연구

- 비휘발성 메모리 연구의 필요성

- 데이터 생산 속도가 기하급수적으로 증가 추세. IoT 서비스와 데이터센터의 폭증으로 2025년에는 2015년의 10배에 달하는 1억 7,500만 테라바이트 생성 예상. 따라서 메모리의 속도, 크기, 에너지 효율이 중요
- 데이터 저장과 전송은 컴퓨팅 시스템의 에너지 소비의 최대 90% 점유
- 에너지 소비를 줄이기 위해서는 전원이 공급되지 않을 때도 정보가 유지되는 비휘발성 메모리가 유리
- 폰 노이만 아키텍처를 벗어나기 위해서는 뉴로모픽 컴퓨팅 시스템을 메모리를 중심으로 완전한 재설계 필요
- CEA-Leti는 비휘발성 메모리로 유망한 기술인 상변화 메모리(PCRAM, PCM), 강유전체 메모리(FeRAM, FRAM), 저항성 메모리(ReRAM, RRAM), 메모리(SOT-MRAM)를 연구개발

<뉴로모픽용 메모리 기술 비교 (출처: CEA-Leti)>

Promising memory technologies

	NOR FLASH	MRAM	PCRAM	RRAM	FeRAM (HfO_2)
Storing capacity	~200 pJ / bit	~20 pJ / bit	~300 pJ / bit	~100 pJ / bit	~10 fJ / bit
Write speed	20 µs	20 ns	10-100 ns	10-100 ns	4 ns @ 4.8 V
Endurance	10^5-10^6	10^6-10^{15}	10^9	10^5 - 10^6 on 16 Mbit	10^8-10^7 on 16 kbit
Data retention	> 125° C	85° C - 165° C	165° C	> 150° C	> 125° C
Manufacturing complexity	Complex	Medium	Medium	Simple	Simple
Multi-level cell	Yes	No	Yes	Yes	No
Industrial scale-up	Bad	Medium	High	High	High
		++ Reliability	++ Endurance	++ Low cost	++ Consumption

- 위상변화 메모리 (PCM) 연구

- 칼코게나이드 소재의 상변화 성질을 이용한 메모리 연구
- 소재 개발 : 칼코게나이드를 포함한 새로운 화합물 탐색
- 설계 및 통합 : 소형화, 에너지 효율 증대, 스토리지 밀도 및 성능 최적화를 위한 PCM 디바이스 설계
- 프로그래밍 및 테스트 : 개발되는 각 메모리에 적합한 동작 프로토콜 연구

칼코게나이드 소재

- 칼코게나이드 소재는 주기율표의 16족 원소인 칼코겐 원소들, 즉 황(S), 셀레늄(Se), 텔루륨(Te)을 포함하는 화합물
- 칼코게나이드 소재의 가장 중요한 특징 중 하나는 상변화 능력
- 열을 가했을 때 비정질(무질서) 상태와 결정질(질서) 상태간 천이가 가능하며, 이 과정에서 전기적 및 광학적 특성이 크게 변함. 이러한 특성은 상변화 메모리(PCM)와 같은 데이터 저장 기술, 재기록 가능한 광디스크(CD, DVD, 블루레이 등)에 널리 사용
- 칼코게나이드 장점
  - ☞ 소형화 : 나노 스케일에서 상전이 메커니즘이 동작
  - ☞ 수명 : 플래시 메모리보다 100~1,000배 높음
  - ☞ 동작 속도: 수십 나노초 수준
  - ☞ 아날로그 정보 저장 가능 : 0과 1 사이의 중간 상태를 사용하면 저장 밀도를 3~4배까지 높일 수 있음

- 강유전체 메모리 (FeRAM, FRAM) 연구

- 메모리는 두 개의 금속 전극을 분리하는 강유전체 절연 재료로 구성
- 전기 편광 벡터의 방향에 따라 정보(0 또는 1)가 물리적으로 변환되어 저장되는 원리
- 소재 개발 : PZT 기반 활성 소재를 무연 소재인 하프늄 산화물(HfO_2)로 대체하여 CMOS와 호환되고, 매우 얇은 층으로 쉽게 쌓을 수 있어 소형화 가능성 큼

강유전체 메모리 장점

- 매우 낮은 소비전력 : FeRAM은 다른 메모리보다 훨씬 적은 에너지 소모
- 수명 : 최대 10^{15} 사이클의 수명으로 다른 메모리보다 월등
- 매우 빠른 쓰기 및 읽기 속도 : 4.8V에서 4ns에 불과함
- 낮은 제조 비용

- 저항성 메모리 (ReRAM, RRAM) 연구

• 전기장의 영향으로 두 개의 금속 전극 사이에 끼워진 절연 재료 내부에 전도 채널이 형성 또는 폐쇄되는 성질을 이용한 메모리 • CEA-Leti는 여러 종류의 RRAM을 연구 : 2010년대에는 CBRAM에 관한 연구 수행 경험. 현재는 속도와 통합 측면에서 더 나은 HfO2 소재 기반의 OxRAM 연구 중

저항성 메모리의 장점

- 속도 : 100ns 이하의 스위칭 시간
- 저렴한 비용
- 낮은 소비전력 : 약 100pJ/비트

3 프랑스 스타트업 GML[23][24][25]

※ GML(GrAI Matter Labs)은 2016년 프랑스 파리에서 설립된 스타트업

ㅇ 2019년 초저지연, 저전력 엣지 처리를 위한 인공지능(AI) 프로세서 GrAI One 발표

- NeuronFlow 기술을 기반으로 한 GrAI One 칩은 딥러닝 네트워크의 엔드투엔드 지연시간을 획기적으로 단축 (자율주행에서 20μs이며, 키워드 인식에서 10μs, 손 제스처 인식은 약 1μs 수준)
- GML의 NeuronFlow 기술은 코어와 로컬 뉴런/시냅스 메모리를 메시 형태로 배치하여 인메모리 컴퓨팅을 실현
- 자율주행, 인간과 기계의 상호작용, 스마트 헬스케어 분야와 같은 빠른 응답이 중요한 엣지 애플리케이션에 활용기대
- 이 디지털 칩은 TSMC 28nm 공정으로 20mm^2 면적에 196개의 뉴런 코어와 로컬 뉴런/시냅스 메모리로 구성된 메쉬 형태로 구현되어 총 20만 개의 뉴런을 처리 (최대 전력 소비량은 35mW)

23) https://www.eetimes.eu/startup-launches-its-first-low-latency-edge-ai-chip/
24) https://www.eetimes.com/neuromorphic-chip-gets-1-million-in-pre-orders/
25) https://bits-chips.nl/article/grai-matter-labs-quietly-snapped-up-by-snap/

ㅇ 2022년 발표된 3세대 GrAI VIP는 풀스택 인공지능 SoC 플랫폼으로 이벤트 기반 컴퓨팅 기법

- 16비트 부동 소수점 기능을 탑재한 업계 최초의 근접 센서 인공지능 솔루션
- 증강현실 영상과 오디오 처리를 위한 엣지 디바이스에 활용
- 20만 개의 뉴런 코어에서 약 1,800만 개의 뉴런으로 확장되어 총 4,800만 개의 신경망 파라미터 처리 가능. 온칩 메모리는 4MB에서 36MB로 확장
- 칩 면적 7.6 x 7.6mm
- 파리에 본사를 두고 아인트호벤과 실리콘밸리(산호세)에 지사를 두고 있는 GML(GrAI Matter Labs)은 2023년 10월 미국의 인스턴트 메시징 앱 Snapchat 개발사인 Snap에 인수

4 스위스 취리히 대학[26]

ㅇ 2019년 출시한 첫 상용품 DYNAP-CNN

※ DYNAP(Dynamic Neuromorphic Asynchronous Processor)은 취리히 연방 공과대학교와 취리히 대학교의 뉴로모픽 연구 성과를 활용하여 SynSense가 2017년부터 개발해 온 뉴로모픽 컴퓨팅 시스템

- DYNAP-CNN은 22nm 기술로 제작된 12mm^2 칩으로, 100만 개 이상의 스파이크 뉴런과 400만 개의 프로그래밍 가능한 파라미터를 수용하며, 컨볼루션 신경망 구현에 적합한 아키텍처로 구성
- SNN으로 변환된 CNN[27]을 효율적으로 실행하기 위해 출시된 디지털 컴퓨팅 방식의 상용칩
- 전통적인 CNN과 달리 이벤트 기반으로 동작하는 스파이킹 뉴런을 이용하여 동작하므로, 더 적은 전력으로 동일한 성능을 얻을 수 있어, 실시간 비전 처리나 입력된 센싱 정보의 신속처리에 적합
- 일반 컴퓨팅 방식보다 100~1,000배 더 낮은 전력으로 동작
- 고속 클럭을 사용하지 않고, 영상이 변할 때만 트리거되는 이벤트 기반 방식으로 처리하므로, 프레임 개념 없이 5ms 이하의 짧은 지연

26) https://doi.org/10.48550/arXiv.2205.13037
27) CNN - Convolutional Neural Networks

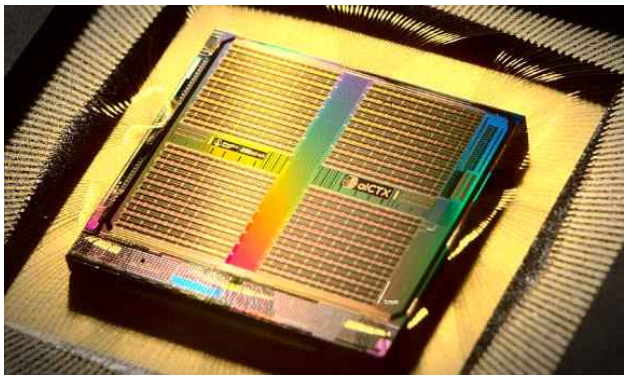
시간으로 처리 가능

- 비전 센서와 통합되어 실시간으로 영상 데이터 처리 가능. 자율주행, 드론, 스마트 보안 등 실시간 비전 처리와 같은 응용분야에 적합

ㅇ 2023년 DYNAP-CNN2 칩 출시

- 기본 아키텍처는 유지하면서 처리속도 및 에너지 효율성을 높이는 최적화를 통해, 저전력 소모와 저지연 신경망 처리 측면에서 개량된 성능 제시

<DYNAP-CNN 칩(좌), Dynap-SE2(우)>



※ 출처: https://www.synsense.ai/ 및 https://doi.org/10.1088/2634-4386/ad1cd7

ㅇ 2023년 Dynap-SE2 소개

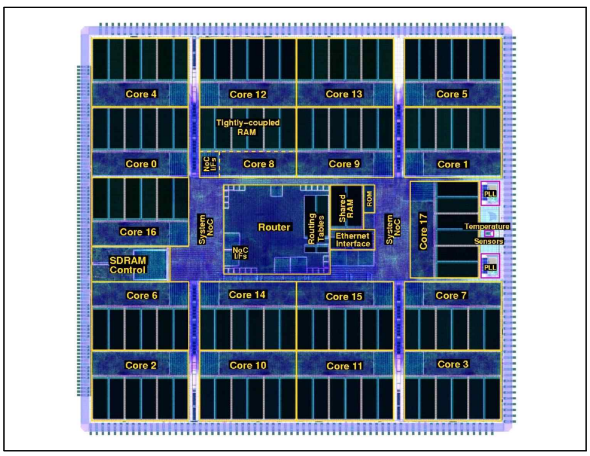
- 2023년 소개된 DYNAP-SE2는 생체 신호 증폭기를 통합하여 모바일 의료 및 로봇 애플리케이션 용도로 활용 가능하도록 제작
- 칩에는 1024개의 혼합(아날로그와 디지털) 신호 스파이크 뉴런과 64k 개의 혼합 신호 시냅스 포함
- 각 칩은 재설계된 1k 개의 아날로그 저전력 스파이크 뉴런과 지연, 무게 및 단기 가소성을 재구성할 수 있는 65k 개의 개량된 시냅스로 구성
- 비동기식 저지연 통신 인프라를 통해 각 뉴런은 최대 230만 개의 주변 뉴런과 통신할 수 있으며, 릴레이 뉴런을 통해 무한한 확장성을 제공하므로, 대규모 네트워크 구축에 활용 가능

5 영국 맨체스터 대학

ㅇ 2019년 SpiNNaker1 발표

- 맨체스터 대학은 2006년 스파이크 신경망을 실시간으로 시뮬레이션하는 데에 최적화된 대규모 병렬 시스템을 설계하는 것을 목표로 SpiNNaker 프로젝트 개시
- 첫 버전인 SpiNNaker1은 2019년에 최종 완성되어 1,200개의 48노드 보드에 배치된 1백만 개의 ARM9 프로세서 코어 장착 목표를 달성
- 본 시스템은 외부 공개되어 뇌 모델링과 뉴로모픽 연구에 널리 활용
- 스파이크 신경망을 실시간으로 시뮬레이션할 수 있도록 개발된 SNN 연구 전용 플랫폼
- 대규모 병렬처리 뉴로모픽 슈퍼컴퓨팅이 가능한 최초의 하드웨어 플랫폼 (2012년 첫 공개)
- 130nm 공정으로 구현된 하나의 프로세서는 18개의 ARM968 코어들과 시냅스 가중치 저장을 위한 128MB SDRAM 및 주변 회로들이 NoC으로 연결된 멀티코어 구조
- 각 ARM968 코어는 1,000개의 뉴런을 시뮬레이션 가능
- 2차원 토러스(Torus) 타입 네트워크를 기반으로 최대 65,536개의 칩을 연결한 통합 시뮬레이션 수행 가능
- 518,400개의 프로세서로 확장할 수 있는 유연한 구조를 이용해 최대 10억 개의 단순 뉴런 혹은 복잡한 구조의 수백만 개의 뉴런 시뮬레이션 가능

<SpiNNaker Chip(좌), SpiNNaker 1 million core machine(우)>



※ 출처(좌): https://apt.cs.manchester.ac.uk/projects/SpiNNaker/SpiNNchip/

※ 출처(우): https://www.humanbrainproject.eu/en/science-development/focus-areas/neuromorphic-computing/

ㅇ SpiNNaker의 후속작 SpiNNaker2

- 성능과 에너지 효율을 개선하는 목표로 HBP의 지원을 받아 2013년에 맨체스터 대학교와 드레스덴 공과대학교와의 협력으로 시작되어 여러 실리콘 프로토타입을 출시
- 스파이킹 신경망을 시뮬레이션하는 순수한 뉴로모픽 컴퓨팅 플랫폼인 SpiNNaker1과 달리 이벤트 기반 심층 신경망(DNN)을 추가로 지원[28]

- SpiNNaker2 칩에는 153개의 ARM 코어와 19MB 온칩 SRAM, 2GB DRAM, 머신러닝 및 뉴로모픽 가속기 탑재
- SpiNNaker2는 단일 머신에서 1,000만 개의 ARM 코어 장착을 목표
- 하나의 SpiNNaker2 칩에는 152개의 코어에 152,000개의 뉴런과 152백만 개의 시냅스 포함
- 22nm 제조 공정의 칩은 아키텍처 개선을 통해, 비슷한 에너지로 SpiNNaker1보다 10배 이상의 신경망 시뮬레이션 가능
- SpiNNaker2 칩은 드레스덴 공대에서 분사한 SpiNNcloud Systems GmbH를 통해 2021년 공개

6 독일 하이델베르그 대학[29][30][31][32][33][34]

ㅇ BrainScaleS 뉴로모픽 컴퓨팅 아키텍처

- BrainScaleS는 EU FET-Proactive FP7의 지원을 받아 2011년에서 2015년까지 독일 하이델베르크 대학교(Heidelberg University)의 연구그룹 주도로 진행된 프로젝트의 결과로 탄생한 뉴로모픽 컴퓨팅 아키텍처
- 아날로그 회로로 설계된 뉴런/시냅스와 폰 노이만 구조의 petaflops급의 슈퍼컴퓨터 시스템이 혼합된 구조로 설계되어, 아날로그 연산을 수행하는 스파이크 뉴런의 기능을 모사
- 인간 뇌의 생물학적 메커니즘 연구를 목적으로 뉴런과 시냅스의 가소성 모델 에뮬레이션에 활용

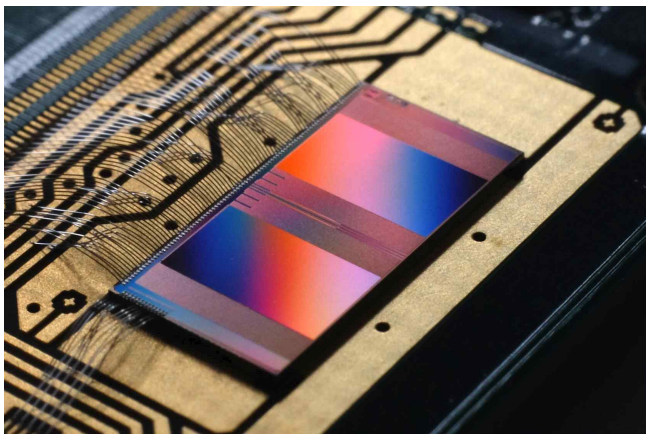
28) https://open-neuromorphic.org/neuromorphic-computing/hardware/spinnaker-2-university-of-dresden/
29) https://doi.org/10.48550/arXiv.2205.13037
30) https://www.kip.uni-heidelberg.de/vision/previous-projects/facets/neuromorphic-hardware/
31) https://www.humanbrainproject.eu/en/science-development/focus-areas/neuromorphic-computing/
32) https://brainscales.kip.uni-heidelberg.de/
33) https://wiki.ebrains.eu/bin/view/Collabs/neuromorphic/BrainScaleS/
34) 인공지능 뉴로모픽 반도체 기술 동향, 전자통신동향분석 35권 제3호, 한국전자통신연구원, 2020

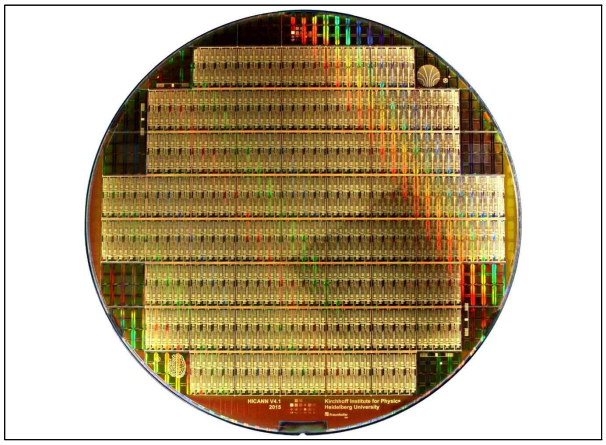
<BrainScaleS 연구 참여 기관>

독일	Uniklinik RWTHAachen, Debreceni Egyetem, Technische Universität Dresden, Ruprecht-Karls-Universität Heidelberg, Forschungszentrum Jülich GmbH
네덜란드	Nederlandse Akademie van Wetenschappen
노르웨이	NMBU
영국	University of Cambridge, The University Of Manchester
스페인	Universitat Pompeu Fabra
프랑스	CNRS-UNIC, CNRS-INT, AMU, INRIA
오스트리아	TUG (오스트리아)
스위스	EPFL LCN, EPFL-BBP, Universität Zürich

- 2016년 발표된 BrainScales 시스템(NM-PM-1)은 20개의 실리콘 웨이퍼로 구성되어, 각 웨이퍼에는 4천만 개의 플라스틱 시냅스와 20만 개의 뉴런이 통합되어 있음

<BrainScaleS-2 single chip(좌), 1세대 웨이퍼 스케일 시스템(우)>





※ 출처: Human Brain Project[35] 및 https://wiki.ebrains.eu/bin/view/Collabs/neuromorphic/BrainScaleS/

- 이후 BrainScales는 2013년에 승인받은 HBP의 지원을 통해 하이델베르그 대학 연구소 'Kirchhoff-Institute for Physics'에서 2023년까지 후속 연구를 진행

ㅇ 2022년 차기 뉴로모픽 컴퓨팅 시스템 BrainScaleS-2 발표

- 아날로그 및 디지털 신호가 결합된 하이브리드 신호 아키텍처 기반
- 단일 칩으로 동작 가능. 약 512개의 뉴런과 130,000개의 시냅스를 칩에 집적
- 온칩 학습 기능과 저전력 소비가 특징. 생물학적 뉴런 속도의 1,000배로 실행

※ BrainScaleS과 SpiNNaker 계열의 시스템은 현재 서비스되고 있는 EBRAINS 플랫폼을 통해 연구목적으로 원격접근 가능

35) https://www.humanbrainproject.eu/en/science-development/focus-areas/neuromorphic-computing/

7 독일 프라운호퍼 IPMS[36]

ㅇ 프라운호퍼 IPMS는 뉴로모픽 컴퓨팅 분야에서 다양한 해법 연구 진행

- 빅데이터와 인공지능(AI) 애플리케이션을 구현하는 데 있어 처리속도와 에너지 효율을 중요한 요소로 설정
- 특히 엣지 애플리케이션을 위해 에너지 효율이 높은 소재, 기술 및 하드웨어 솔루션을 개발하기 위해 다양한 프로젝트 참여 중
- 뉴로모픽 하드웨어를 위해 비휘발성 메모리인 강유전체 전계 효과 트랜지스터를 기반으로 하는 크로스바 아키텍처 연구
- 다수의 유럽 프로젝트 (TEMPO, ANDANTE, STORAIGE) 참여 및 프라운호퍼 내부 자금으로 연구 진행
- MEMION (작센주 후원) 프로젝트에서는 미래형 SNN을 위해 리튬물질을 활용한 혁신적인 소재 연구

① TEMPO 프로젝트

※ Technology and Hardware for Neuromorphic Computing

- 뉴로모픽 하드웨어의 에너지 효율을 크게 개선하여 산업, 의료, 지원 시스템 등 다양한 분야에서 새로운 애플리케이션을 구현하려는 목표
- 멤리스터 연구의 일환으로 강유전체 커패시터를 이용한 트랜지스터(FeFET)로 만든 비휘발성 메모리 연구
- 전력 소비가 적은 마이크로전자 부품 구현 연구. 인공지능, IoT 및 엣지 컴퓨팅 분야의 애플리케이션에 활용기대

② ANDANTE 프로젝트

- 강유전체 트랜지스터(FeFET)를 사용하여 시냅스를 모사 (TEMPO 프로젝트와 유사)
- 프라운호퍼 IIS 및 EMFT와의 협력을 통해 글로벌파운드리의 22FDX® 기술로 FeFET 칩을 개발 중
- 에너지 효율 높은 인공지능 칩 출시가 목표

36) https://www.ipms.fraunhofer.de/en/Strategic-Research-Areas/Neuromorphic-Computing.html

③ StorAIge 프로젝트

※ New storage technology for edge AI applications

- 에너지 효율적인 메모리와 높은 컴퓨팅 성능을 갖춘 인공지능 칩 개발
- 와트당 10TOPS 목표
- 유전체 트랜지스터(FeFET)를 일반 CMOS 반도체에 통합 구현

④ MEMION 프로젝트

※ Memristive Redox Transistors for Neuromorphic Computer Architectures

- 프로젝트 기간 : 2020 - 2023
- 목표는 뇌의 기능을 최대한 충실하게 재현하고 높은 수준의 가소성을 구현하는 아키텍처 설계. 특히, 뉴로모픽 컴퓨팅 네트워크에 사용할 수 있는, 다단계 스위칭 동작이 가능한 에너지 효율적인 트랜지스터 개발을 목표
- 특히, 멤리스티브 레독스(Redox) 트랜지스터 연구에 집중

• 레독스 트랜지스터는 산화환원반응을 이용하여 작동하는 트랜지스터 • Redox는 산화 환원(oxidation-reduction)의 줄임말. 전자가 한 물질에서 다른 물질로 이동하는 현상을 이용하는 이 트랜지스터는 산화환원반응을 통해 전도도를 제어하거나 전기 신호를 증폭하는 역할 수행 • 레독스 트랜지스터는 산화 상태의 변화에 따라 전류의 흐름을 제어하는 원리를 통해 신호 증폭과 스위칭이 가능한 성질 보유. 사용되는 전기화학적 반응이 낮은 전압에서도 매우 효과적으로 작동하므로 저전력 소자로 활용 가능

- 필라멘트의 형성 때문에 채널의 전도도가 변하는 일반적인 RRAM 소자보다 레독스 트랜지스터는 훨씬 더 정밀하고 정확하게 동작 가능
- 본 과제는 레독스 트랜지스터를 리튬 물질을 이용하여 제작 예정

8 독일 드레스덴공대(TU Dresden)[37]

○ BrainScaleS와 SpiNNaker 프로젝트 참여

- SpiNNaker 연구 참여

• HBP의 지원으로 2013년부터 맨체스터 대학교와 협력관계 • 특히, SpiNNaker의 후속 버전인 SpiNNaker2 칩 개발은 드레스덴 공과대학의 세바스찬 회프너(Sebastian Höppner) 박사가 깊게 관여 • SpiNNaker2칩은 드레스덴 공대에서 분사한 SpiNNcloud Systems GmbH를 통해 2021년 공개 • SpiNNaker2 칩은 맨체스터 대학교의 스티브 퍼버 교수가 개발한 다중 코어 아키텍처로 구성. 퍼버 교수는 오늘날 모바일 기기에서 널리 사용되는 ARM 아키텍처 선구자 • 2019년 휴먼 브레인 프로젝트는 2세대 머신(SpiNNcloud)의 건설을 위한 800만 유로의 보조금을 드레스덴 공과대학교에 지급 • 현재 드레스덴 공대에 설치 중인 SpiNNaker2 머신은 720개의 48 노드 회로 기판에 520만 개의 코어를 탑재하고, 최대 8개의 서버 랙 크기의 슈퍼컴퓨터와 같은 시스템으로 구성될 예정. 2024년 완성 예정

- BrainScalesS 연구 참여

• EU FET-Proactive FP7의 지원을 받아 2011년에서 2015년까지 독일 하이델베르크 대학교(Heidelberg University) 연구그룹 주도로 진행된 프로젝트의 결과로 탄생한 뉴로모픽 컴퓨팅 아키텍처 • 드레드덴 공대는 취리히 대학, 맨체스터 대학 등과 함께 EU FET-Proactive FP7의 지원을 받아 2011년에서 2015년까지 독일 하이델베르크 대학교(Heidelberg University)의 연구그룹 주도로 진행된 BrainScaleS 프로젝트 참여

○ 맴리스터 연구[38]

- 뉴로모픽 연산에 광범위하게 활용될 수 있는 멤리스터 소자 연구
- ReRAM의 설계 방안과 신뢰성 및 내결함성 개선을 위한 회로의 다각도 연구
- 단일 디바이스로 메모리와 컴퓨팅에 모두 활용할 수 있는, 또 다른 잠재력이 높은 멤리스터 소자의 후보인 강유전체 트랜지스터(FeFET) 연구

37) https://tu-dresden.de/tu-dresden/newsportal/news/computer-lernen-das-lernen?set_language=en
38) https://ieeexplore.ieee.org/document/9473976

- 새로이 떠오르는 나노기술 개념인 재구성 가능한 트랜지스터(RFET)와 강유전체 트랜지스터(FeFET)를 하나로 통합하는 연구 수행

9 독일 막스플랑크 연구소(Max Planck Institute) 39)40)

ㅇ 2021년 유명학술지 Nature Electronics를 통해 높은 에너지 효율을 제공하는 커패시터형 메모리 기술 시연 내용 발표

※ 드레스덴의 할레(Saale)에 위치한 막스플랑크 미세구조 물리연구소는 SEMRON GmBH와 공동 발표한 해당 논문을 통해 주목을 받음 (제목: Energy-efficient memcapacitor devices for neuromorphic computing)

- 멤리스티브 소자를 사용하는 대신, 전하 차폐 원리를 활용하는 멤커패시티브 소자를 활용하여 병렬 곱셈-누적 연산을 높은 에너지 효율로 실현할 수 있음을 소개
- 156개의 마이크로스케일 멤캐패시터 소자로 구성된 크로스바 어레이를 만들어 신경망을 훈련
- 와트당 1초에 29,600테라 연산이라는 높은 에너지 효율을 제공하면서 높은 정밀도(6~8비트)를 보장함을 강조
- 이 기술은 기존의 실리콘 제조 공정을 사용하므로 구현이 용이하여 즉시 상용화가 가능한 장점

멤리스티브와 멤커패시티브

- **멤리스티브(Memristive) 소자**는 저항이 전류 히스토리에 따라 변하고 과거의 상태를 기억하는 소자. 멤리스티브 소자는 저항뿐만 아니라 시스템의 다른 특성(예: 전압, 전류, 시간)의 변화에 대한 기억 능력으로, 뉴로모픽 컴퓨팅과 비휘발성 메모리 기술 분야에서 주목받는 소자
- **멤커패시티브(Memcapacitive) 소자**는 메모리 기능을 가진 커패시터. 메모리(memory)와 커패시터(capacitor)의 성질을 결합한 소자. 기억된 과거 전압 히스토리를 기반으로 커패시턴스가 변함. 이 소자는 전류 히스토리를 기반으로 저항이 변화하는 멤리스터(Memristor)와 유사하지만, 전압 히스토리에 따라 저항이 아닌 커패시턴스가 변하는 특징이 활용됨

39) https://www.mpi-halle.mpg.de/ultra-low-power-memcapacitor-device-for-neuromorphic-computing
40) https://www.nature.com/articles/s41928-021-00649-y

0 프랑스 스타트업 UPMEM[41)42)43)]

ㅇ 2019년 DDR4 메모리 기반 인공지능 가속기용 PIM 반도체 출시

- PIM 개념을 상용으로 널리 쓰이는 DRAM에 도입한 상용칩(PIM-DRAM)

※ PIM은 CPU와 메모리 간의 폰노이만 병목현상을 줄이는 효과

- 2019년 발표된 Hot Chips 31 4GB DRAM 칩은 인공지능 연산을 기준으로 기존 DRAM보다 20배의 성능 제고, 10배의 에너지 절감 효과
- 본 칩은 메모리 모듈 내부에 인공신경망에 필요한 MAC 연산기능을 구현하여, PCIe와 같이 시스템 버스에 연결되는 GPU 및 NPU와는 달리, 상대적으로 속도 빠른 메모리 버스에 연결 가능한 장점
- UPMEM는 서버의 DRAM 모듈을 PIM-DRAM 모듈로 단순 교체할 수 있게 하는 것을 목표로 함

ㅇ ODYSSAI 프로젝트 참여

- 프랑스 2030 투자 플랜이 지원하는 ODYSSAI 프로젝트 참여 (2023년 12월 프로젝트 개시)
- 탈레스가 주관하는 본 프로젝트는 고성능 인공지능 엣지 컴퓨팅 시스템 개발에 중점
- 기존 DDR4 기반의 PIM을 DDR5 기반으로 업그레이드하고, 더 다양한 컴퓨팅 플랫폼을 지원하도록 호환성을 확장하고, 오픈 소스 기술을 기반으로 하여 개방적이고 안전한 플랫폼 확보하는 목표

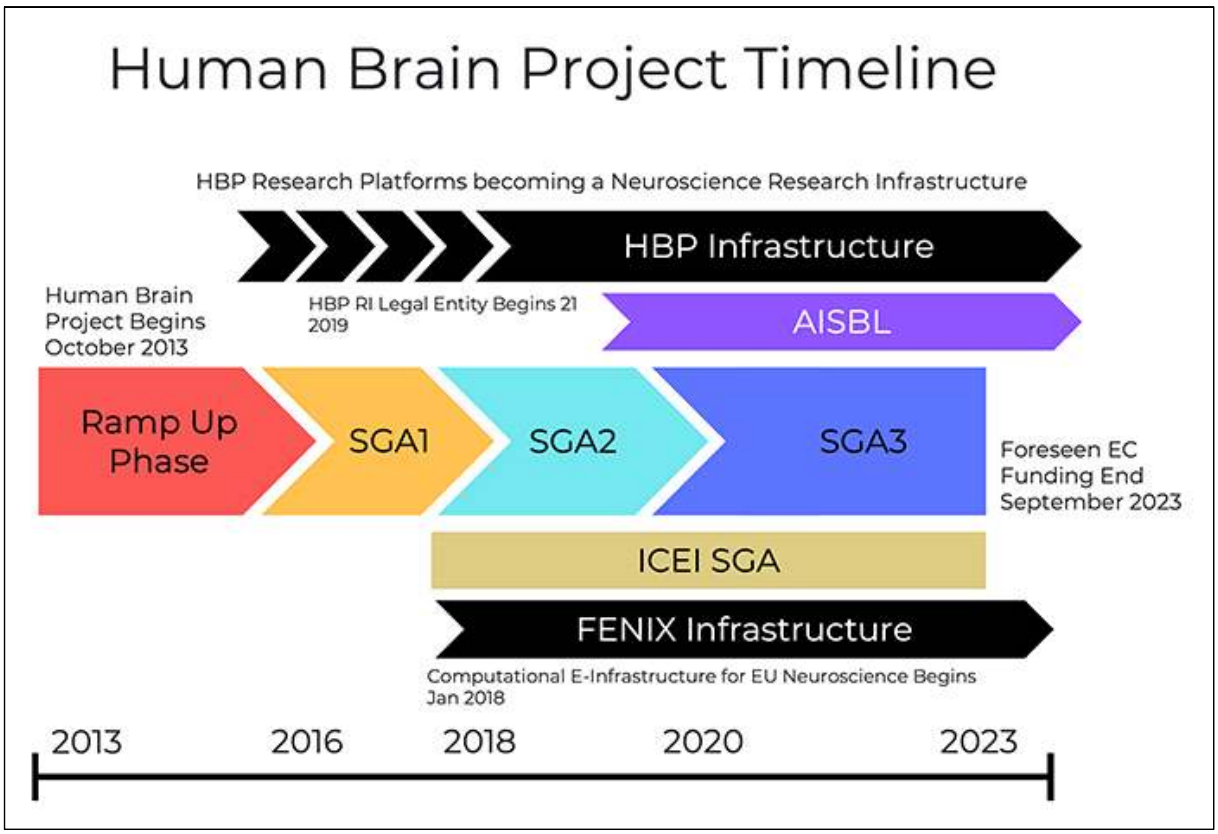
41) https://www.upmem.com/
42) https://www.anandtech.com/show/14750/hot-chips-31-analysis-inmemory-processing-by-upmem
43) https://ethz.ch/en/industry/industry/news/data/2022/03/mehr-daten-schneller-und-energiesparender-verarbeiten.html

Ⅳ. EU 프로젝트 동향

1 EU human brain project 2013-2023[44)45)46)]

ㅇ 인간의 뇌에 대한 대규모 미래원천기술 연구를 위한 EU 프로젝트

- Human Brain Project(HBP)는 2013년 7차 EU R&D 프레임워크 프로그램(FP7)에서 시작되어 2023년까지 10년간 10억 유로가 투자된 대표적 연구
- 2000년대 중반부터 유럽과 미국 등에서 원천기술 확보를 목적으로 국가 주도로 진행된 인공지능 반도체 연구 사업의 일환



(출처: https://www.humanbrainproject.eu/en/about-hbp/project-structure/human-brain-project-ec-grants/)

- 대규모 HBP 연구투자 덕분에 조성된 연구생태계로 인해 인공지능 반도체 연구가 유럽에서 활발히 지속 중

※ 앞서 소개된 BrainScales와 SpiNNaker는 HBP를 통해 개발되어 뇌와 인공지능 연구용으로 널리 활용된 대표적인 EU의 뉴로모픽 컴퓨팅 시스템. 이들은 차기 시스템으로 업그레이드되어 현재까지 연구개발용으로 널리 활용 중

44) https://www.humanbrainproject.eu/en/follow-hbp/news/2023/09/12/human-brain-project-celebrates-successful-conclusion/
45) https://www.humanbrainproject.eu/en/
46) https://www.humanbrainproject.eu/en/science-development/scientific-achievements/brochures/

○ HBP 개요

사업 목표	• ICT 기술을 활용하여 뇌 연구의 새로운 기술적 토대를 마련. 다양한 분야의 데이터와 지식의 통합을 촉진하고, 뇌에 대한 새로운 이해, 뇌 질환에 대한 새로운 치료법 및 새로운 뉴로모픽 컴퓨팅 기술개발을 위한 연구 생태계 조성
사업 규모	• 9개국 155개 협력 기관과 총 6억 7천만 유로의 예산이 투입된 유럽 최대 규모의 연구 프로젝트이자, 최초의 플래그십 프로젝트의 하나
사업 성과	• 상세한 3차원 디지털 인간 뇌 지도 • 간질 및 파킨슨병과 같은 질환자들의 뇌에 대한 개인화된 가상 모델링 • 인공지능 분야의 획기적인 기술 성과 (BrainScales, SpiNNaker 등) • 사업종료 후에도 전체 신경과학 커뮤니티에 귀중한 자원으로 남게 될 개방형 디지털 연구 인프라, EBRAINS • 사업 기간 동안 발전된 컴퓨팅기술과 디지털화된 데이터로 인해 뇌 연구 수행 방식에 근본적인 변화 견인 • 뇌에 대한 이해가 높아지면서, 뇌에서 영감을 얻은 AI와 뉴로모픽 컴퓨팅, 인지 로봇 공학 분야에까지 큰 발전 견인

○ HBP 주요연구 - 스파이크 신경망 연구

- 딥러닝으로 대표되는 인공지능 시스템은 연산량이 방대하므로, 에너지 효율이 중요한 문제
- 본 사업에서 네덜란드, 오스트리아, 독일, 스위스의 HBP 과학자들은 인간의 뇌를 모사하여 인공지능의 에너지 효율을 개선하는 방안 연구

※ 이 연구팀은 GPU를 활용한 기존의 딥러닝 구현방식과 다르게, 뇌를 더 정밀하게 모사하는 알고리즘을 개발하고 최적화 수행. 스파이크 신경망 도출

- 연구결과로 도출된 스파이크 신경망 연구를 위해서는 대규모 뉴로모픽 컴퓨팅 시스템이 필요하게 되어, 대표적인 두 개의 시스템(SpiNNaker와 BrainScaleS)이 사업의 일환으로 개발되었음
- EBRAINS를 통해 관련 연구자들이 액세스할 수 있도록 공개된 상태

○ HBP 주요 참여 기관

- 총 100개 이상의 연구그룹 중에서 뉴로모픽 분야의 연구에 기여했던 기관들은 다음과 같음

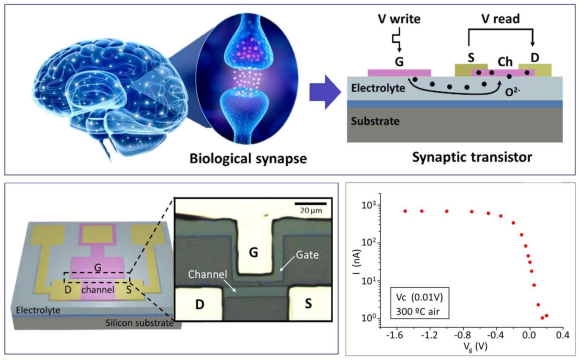
스위스 연방 공과대학(ETH Zürich)	• DYNAP과 같은 뉴로모픽 칩 개발을 주도한 바 있는 본 대학은, 스파이킹 신경망(SNN)를 기반으로 저전력, 고성능 신경망 연산을 가능하게 하는 컴퓨팅 아키텍처 연구 수행 • 특히 BrainScaleS 및 SpiNNaker 세부 프로젝트와 협력하여 뉴로모픽 칩 설계 및 하드웨어 개발 지원
독일 하이델베르크 대학교(University of Heidelberg)	• BrainScaleS 뉴로모픽 플랫폼 개발 주도 • BrainScaleS는 아날로그 회로를 사용하여 뇌의 신경망을 물리적으로 구현하고 시뮬레이션할 수 있도록 설계되어, 신경과학적 데이터를 기반으로 한 뇌 시뮬레이션과 뉴로모픽 칩 연구에 널리 활용 가능한 공개된 플랫폼
영국 맨체스터 대학교 (University of Manchester)	• 병렬처리를 통해 대규모 스파이킹 신경망을 시뮬레이션할 수 있는 시스템인 SpiNNaker (Spiking Neural Network Architecture) 연구개발 주도 • 뉴로모픽 컴퓨팅 분야에서 큰 기여
독일 율리히 연구소 (Forschungszentrum Jülich)	• 세계적인 뇌 연구기관이자 율리히 슈퍼컴퓨팅 센터를 보유한 율리히 연방 뇌연구소가 10년간의 프로젝트에서 중요 역할 수행 • 보유한 슈퍼컴퓨터(JUQUEEN 및 JURECA) 인프라를 활용하여 뉴로모픽 컴퓨팅 및 스파이킹 신경망(SNN)의 시뮬레이션 및 테스트 지원 • 신경과학 데이터 관리를 위한 인프라 개발에 기여 • 연구자들이 대규모 신경과학 데이터를 관리하고 분석할 수 있도록 지원하는 EBRAINS 플랫폼 개발에 기여 (EBRAINS는 HBP의 주요 성과 중 하나로 뇌 연구에 필요한 데이터를 저장하고 공유할 수 있는 연구 인프라)
스위스 로잔공대 (EPFL)	• 휴먼 브레인 프로젝트에서 신경정보 이론 연구, EBRAINS 및 뉴로보틱스 플랫폼 및 고성능 컴퓨팅 플랫폼 개발에 기여 • 한편 EPFL은 스위스에서 자금을 지원하는 대규모 과학 이니셔티브, 블루 브레인 프로젝트(BBP) 수행
독일 드레스덴 공대 (TUD)	• 병렬처리 VLSI 시스템과 뉴로모픽 회로 연구
영국 University of Sussex	• 뇌 기능을 분석을 통해, 인지적인 뇌 모델을 적용한 로봇 시스템 개발에 기여 • 뉴로로봇공학(Neurorobotics)과 인지시스템 분야 연구

2 Horizon Europe Framework Programme

① TRANSIONICS 프로젝트[47]

※ TRANSIONICS: Solid-State Ionics Synaptic Transistors for Neuromorphic Computing

주관기관	스페인 FUNDACIO INSTITUT DE RECERCA DE L'ENERGIA DE CATALUNYA
연구기간	2022.06 - 2023.11
연구내용	멤리스터 소자 연구



<이미지: Transionics Project Results(출처: https://cordis.europa.eu/project/id/101066321/reporting)>

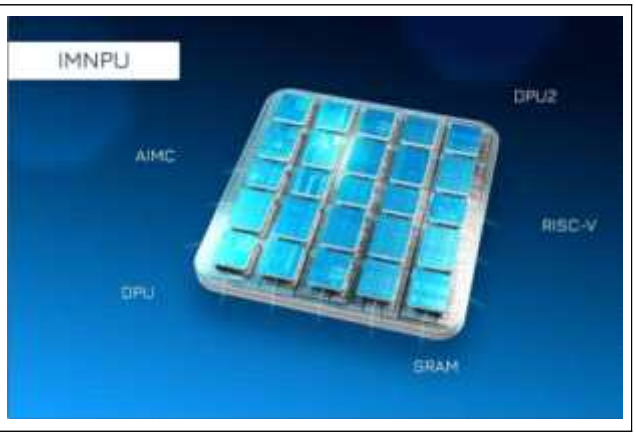
- SNN 기반의 뉴로모픽 컴퓨팅 분야의 주요 목표는 기존 트랜지스터를 신경 시냅스와 유사한 방식으로 학습할 수 있는 시냅스 트랜지스터로 대체하는 것. 즉 다중 상태 비휘발성 트랜지스터를 개발하는 것
- 그러나 기존의 시냅스 트랜지스터는 이온성 액체나 양성자 전도성 고분자와 같이 본질적으로 불안정하고 집적하기 어려운 전해질이라는 문제
- 본 과제는 ERC CoG 지원금 (ULTRASOFC 과제)을 통해 개발된 TRANSIONICS 트랜지스터를 대상으로 연구
- TRANSIONICS 트랜지스터는 산화물 이온 전해질을 이용하여, 상온에서도 매우 안정적으로 비휘발성 특성을 보이며 실리콘 공정을 사용하여 확장이 용이한 장점
- TRANSIONICS 트랜지스터는 혼합된 이온-전자 도체를 환원/산화하여 실제 뉴런처럼 외부 자극으로 채널 특성 변조 가능
- 또한 TRANSIONICS는 널리 사용되는 칩 제조 기술과 호환되므로 시장 진입에 용이하므로, 본 과제는 상용화를 염두에 두고 해당 기술의 타당성 검토를 목표로 함

47) https://codasip.com/2023/04/25/customized-risc-v-core-for-neuromorphic-computing/

② NeuroSoC 프로젝트[48]

※ A multiprocessor system on chip with in-memory neural processing unit

주관기관	이탈리아 STMICROELECTRONICS SRL
연구기간	2022.09 - 2026.02
연구내용	상용화 가능한 인공지능 반도체 제작 (위상변화방식의 메모리와 복수의 RISC-V 프로세서를 집적한 PIM 구조의 SoC)



<이미지: IMNPU (출처: https://cordis.europa.eu/project/id/101070634)>

- 뉴로모픽 및 인메모리 컴퓨팅을 연구하는 여러 프로젝트와 제품이 등장하였으나, 대량 생산이 가능한 수준에 미치지 못함
- 또한 스파이킹 신경망과 같이 아직 완전히 입증되지 인공지능 컴퓨팅 알고리즘을 대상으로 하는 프로토타입이 대부분
- 본 사업은 산업적으로 입증된 공정 기술을 활용하여 상용화 가능성이 큰 SoC (28nm CMOS 공정의 멀티프로세서 시스템 온 칩) 프로토타입 칩을 개발하는 것이 목표
- 시장에서 검증된 ST마이크로일렉트로닉스의 고밀도 임베디드 PCM 셀 공정 기술 활용하여 상용화 시도

③ HYBRAIN 프로젝트[49]

※ Hybrid electronic-photonic architectures for brain-inspired computing

주관기관	Netherlands, UNIVERSITEIT TWENTE
연구기간	2022.05 - 2026.04
연구내용	컨볼루션 신경망(CNN)의 연산 가속을 위한 하이브리드 아키텍처 연구

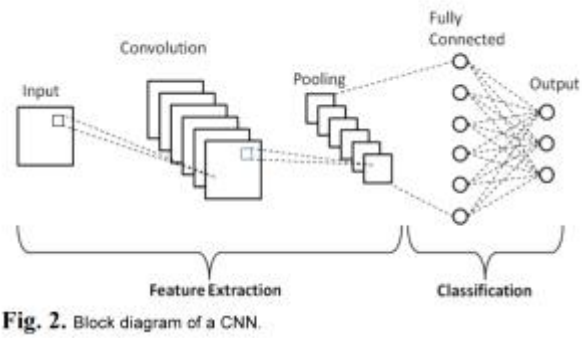


Fig. 2. Block diagram of a CNN.

<이미지: CNN 구성도 (출처: https://doi.org/10.6109/jicce.2018.16.3.173)>

48) https://cordis.europa.eu/project/id/101070634
49) https://cordis.europa.eu/project/id/101046878

- 기존의 컴퓨터 아키텍처는 컨볼루션 신경망(CNN) 연산에 최적화 되어 있지 않음

※ 따라서 지연시간과 소모전력 측면에서 높은 요구사항을 충족하기 어려우며 특히, CNN에서 가장 심각한 병목현상은 초기 컨볼루션 레이어에서 발생

- 본 연구는 초기 컨볼루션 레이어에 새로운 위상변화 물질을 사용하는 광자 컨볼루션 프로세서(PCP)를 활용하여 처리속도 개선

※ HYBRAIN 프로젝트에서 개발한 광자 텐서코어(포토닉 크로스바 어레이 형태의 포토닉 매트릭스-벡터 곱셈이 가능)는 데이터 처리 능력이 탁월

- PIM 구조 : 포유류의 뇌처럼 가까이 얽혀있는 메모리와 프로세싱 유닛은 시냅스 가중치 이동에 소요되는 시간을 줄이는 효과

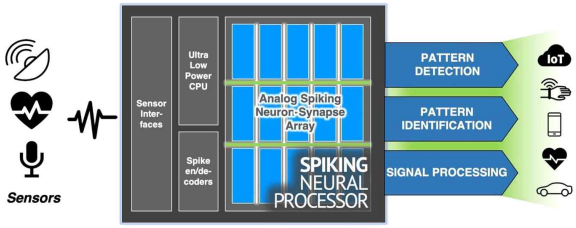
- 출력부는 각각 멤리스티브(상변화 메모리) 크로스바 어레이와 도펀트(dopant) 네트워크 처리장치를 기반으로 하는 캐스케이드 전자 분류기 레이어를 적용

- 검토된 여러 기술을 복합적으로 활용하는 하이브리드 아키텍처 구조

※ 초고속(1마이크로초 미만) 및 에너지 효율적인(1와트 미만) 엣지 인공지능 추론 성능개선 연구

④ Spiking Neural Processor 프로젝트[50]

주관기관	Netherlands, INNATERA NANOSYSTEMS BV
연구기간	2023.05 - 2025.04
연구내용	스파이킹 신경망 프로세서 개발



<이미지: Spiking Neural Processor (출처: https://cordis.europa.eu/project/id/190123060/reporting>

- Innatera는 델프트 공과대학교에서 분사한 신경망 프로세서 반도체 회사

- Innatera의 스파이킹 뉴럴 프로세서(SNP)는 시간 개념이 내재된 SNN을 구현한 프로세서

- SNP는 전력 및 지연 시간에 민감한 엣지 센서 애플리케이션에서 상시 패턴 인식 기능 지원 가능

50) https://cordis.europa.eu/project/id/190123060

⑤ spiNets 프로젝트[51]

※ Functionalised dense spintronics oscillator networks for neuromorphic computing

주관기관	포르투갈, INESC MN
연구기간	2024.09 - 2026.08
연구내용	뉴로모픽 컴퓨팅용 메모리 구현에 자기 터널 접합 스핀트로닉스 나노소자 활용 연구

- 뇌의 핵심 구성 요소 중 하나인 시냅스는 뉴런끼리 통신할 수 있게 하고 뉴런 간의 연결 강도가 저장되는 역동적이고 가소성 있는 메모리 역할을 하는 중요한 요소

※ 이러한 시냅스의 특성은 뉴런간 연결을 비휘발적이고 가역적인 방식으로 제어하면서 복잡한 학습 및 기억을 가능하게 함

- 자기 터널 접합(MTJ)과 같은 스핀트로닉스 나노소자는 견고성, 다기능성 및 금속 산화물 반도체(CMOS) 기술과의 호환성으로 인해 최근 뉴로모픽 컴퓨팅의 유망한 소자로 부각

- 전통적인 트랜지스터에 비해 전력 소모가 낮고 스위칭 속도가 빨라 인공 시냅스와 뉴런 구현에 적합한 후보 소자
- 스핀트로닉스 소자는 나노스케일로 크기를 줄일 수 있어 대규모 신경망 구성에 유리

3 Horizon 2020 Framework Programme

① MENESIS 프로젝트[52]

※ Memristor-Enabled NEuromorphic System for Intelligence in Space

주관기관	영국, UNIVERSITY OF SOUTHAMPTON
연구기간	2021.09 - 2023.08
연구내용	멤리스터 소자를 활용한 인공위성 탑재용 뉴로모픽 시스템

- 저궤도 위성 시장의 확대로 인공위성의 수가 폭증하면서 위성통신 트래픽을 효율적으로 운영할 필요성 증대

51) https://cordis.europa.eu/project/id/101180621
52) https://cordis.europa.eu/project/id/101029535

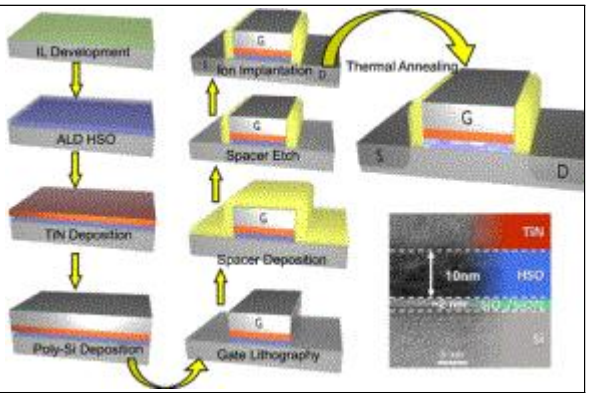
- 위성통신용 지구국에는 인공지능을 도입하여 트래픽 처리 역량이 강화되었으나, 위성 탑재체에도 인공지능 도입이 필요함
- 위성에서 인공지능을 활용하기 위해서는 전력소모가 적으면서 처리속도가 높은 인공지능 칩 필요
- 이 프로젝트는 인공지능을 위성에 탑재하여, 위성이 지구국과 통신하기 전에 신호와 잡음을 구별할 수 있는 자율적인 계산 능력을 갖추게 하는 혁신 개념 제안

• 사용할 인공지능 가속기에는 새로운 멤리스터 기술을 활용 • 멤리스터 아키텍처는 가볍고 전력 소모가 적으며 빠르고 집적도 높은 칩을 구현할 수 있을 뿐 아니라 온도변화에도 강인 • 우주에서 활용할 장비의 단가와 운용 비용 절감 기대

② TEMPO 프로젝트[53)54)]

※ Technology and Hardware for Neuromorphic Computing

주관기관	imec (벨기에)
연구기간	2019.05 - 2023.01
연구내용	SNN을 지원하면서 새로운 멤리스터 기술을 포괄하는 고도화된 심층 신경망(DNN) 엔진 개발



<이미지: FeFET 소자 제작 예 (출처: DOI 10.1021/acsaelm.2c00771>

- 뉴로모픽 하드웨어의 에너지 효율을 크게 개선하여 산업, 의료, 지원 시스템 등 다양한 분야에서 새로운 애플리케이션 구현을 목표
- 기존 메모리를 대체할 수 있는, 강유전체 커패시터를 이용한 트랜지스터(FeFET)로 만든 비휘발성 메모리 연구

• 멤리스터 연구의 또 다른 변형 • FeFET는 일반적인 트랜지스터 게이트에 강유전체 층을 추가하여 전하 저장 기능이 부가됨. 전원이 꺼져도 정보가 유지되는 비휘발성 메모리로 활용 가능

53) https://k-erc.eu/2023/06/europe-trends/15610/
54) https://cordis.europa.eu/project/id/826655

- 전력 소비가 적은 소형 전자 부품으로 구현이 가능하므로, 인공지능, IoT 및 엣지 컴퓨팅 분야의 애플리케이션에 활용 기대

③ ANDANTE 프로젝트[55]

※ AI for New Devices And Technologies at the Edge

주관기관	프랑스 STMICROELECTRONICS GRENOBLE 2 SAS
연구기간	2020.06 - 2024.01
연구내용	인공 및 스파이크 신경망 연산을 돕는 가속기를 활용하여 저전력 고성능 인공지능 하드웨어와 소프트웨어 플랫폼을 구축

<이미지: ANDANTE Consortium Members (출처: https://cordis.europa.eu/project/id/876925/reporting>

- MRAM, PCM, RRAM, OXRAM, FeFET과 같은 미래 메모리 기술을 기반으로 혁신적인 하드웨어/소프트웨어 딥러닝 솔루션을 개발하여 극한의 에너지 효율을 가지는 강력한 인공지능 컴퓨팅 제품 실현 목표

※ 유럽의 주요 파운드리, 칩 설계, 시스템 하우스, 애플리케이션 회사 및 연구 파트너간 협력 생태계를 구축

- 강유전체 트랜지스터(FeFET)를 사용하여 시냅스를 모사 (TEMPO 프로젝트와 유사)
- 강유전체 트랜지스터는 매우 높은 동적 범위와 매우 낮은 전파 지연이 특징

> • 뉴런의 신호를 매우 낮은 손실로 빠르게 전송하는 장점
> • 본 트랜지스터는 기존 트랜지스터와 함께 칩에 통합할 수 있어 확장 가능한 엣지 인공지능 가속기로 구현 가능

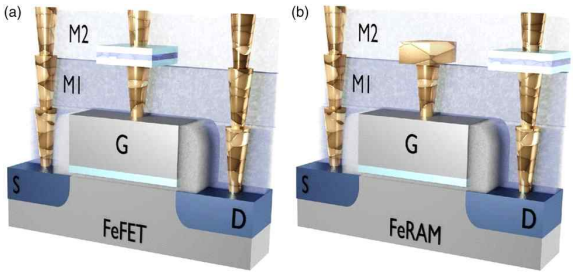
- 프라운호퍼 IIS 및 EMFT와의 협력을 통해 글로벌파운드리의 22FDX® 기술로 FeFET 칩 개발 중. 에너지 효율 높은 인공지능 칩 출시 목표

55) https://cordis.europa.eu/project/id/876925

④ StorAIge 프로젝트[56]

※ Embedded storage elements on next MCU generation ready for AI on the edge

주관기관	프랑스 STMICROELECTRONICS GRENOBLE 2 SAS
연구기간	2021.07 - 2024.06
연구내용	다양한 차세대 비휘발성 메모리 기술(PCM, OxRAM, FeRAM)과 가속기를 활용한 인공지능 반도체 제작



<이미지: FeFET와 FeRAM 구조 연구 (출처: https://doi.org/10.1002/pssr.202200168>

- 28nm FD-SOI 공정으로 차세대 위상변화 메모리(PCM)를 상용화하여 성능, 에너지 효율, 보안 측면에서 경쟁력 있는 시스템온칩(SoC) 프로토타입을 제작하여, 와트당 10 TOPS의 컴퓨팅 파워를 갖춘 칩셋과 솔루션 확보 목표

> • Fully Depleted Silicon on Insulator (FD-SOI)는 반도체 제조 기술의하나로, 고성능 저전력 소자 제작에 유리. 기존의 CMOS 공정과 비교하여 소비전력, 스위칭 속도, 온도 민감성, 공정의 단순성 측면에서 장점
> • FinFET 같은 최신 3D 트랜지스터의 구조에 비해 공정이 단순하여 제조 비용이 낮고 높은 수율 제공

- 유전체 트랜지스터(FeFET)를 일반 CMOS 반도체에 통합하여 구현
- 인메모리 컴퓨팅(IMC)기법 적용 검토

※ 고성능/고에너지효율의 실리콘 기반 AI칩 플랫폼을 개발하여 엣지장치에 활용

⑤ MELON 프로젝트[57]

※ Memristive and multiferroic materials for emergent logic units in nanoelectronics

주관기관	프랑스, UNIVERSITE DE PICARDIE JULES VERNE
연구기간	2020.04 - 2025.03
연구내용	산화물질을 이용하는 멤리스터 소재 연구

<이미지: 맴리스터 연구 사례 (출처: https://www.melon.ferroix.net/mission>

56) https://cordis.europa.eu/project/id/101007321
57) https://cordis.europa.eu/project/id/872631

- 뇌 신경간 연결을 모방하기 위해 신호의 이력에 따라 전도도가 달라지는 실리콘 기반의 메모리스티브 산화물 연구

• 나노 스케일의 다중강성(multiferroic) 재료를 사용하여 다중 논리값을 지니는 소재 구현 • 멤리스티브 시스템과의 연결을 위한 2차원 산화물 인터페이스 연구

- 기초과학, 응용 화학 및 물리학, 재료 과학, 나노 기술에 이르기까지 견고한 학제 간 교류가 가능한 컨소시엄 구성

※ EU 회원국, 프랑스, 네덜란드, 스페인, 파트너 국가인 아르헨티나, 준회원국인 우크라이나의 중소기업 포함

⑥ PHOTON-NeuroCom 프로젝트[58]

※ Memristor-Enabled NEuromorphic System for Intelligence in Space

주관기관	덴마크, AARHUS UNIVERSITET
연구기간	2017.08 - 2019.07
연구내용	스핀트로닉스 소자를 활용하여 뉴로모픽 컴퓨팅의 성능을 개선하기 위해, 시스템 내 큰 바이어스 전류를 짧은 편광 레이저 펄스로 대체하는 새로운 접근 방식 연구

- 자기-광자 상호작용을 모델링하고, 추출된 모델을 적용한 NCS 설계 및 시뮬레이션
- 현재 뉴로모픽 컴퓨팅 시스템(NCS)은 회로 면적과 소모전력 측면에서 효율성이 다소 떨어지는 CMOS 기술로 구현
- NCS의 면적과 전력효율 개선을 위한 방법의 하나로 스핀트로닉스 기반의 소자가 유망

• 스핀트로닉스 소자 기반 NCS는 에너지 효율 개선에 유리하지만, 바이어스 전류를 사용하여 자기 모멘트 상태를 변경하는 일반적인 방법은 NCS 전력소모의 90%를 차지하므로 개선의 여지가 있음 • 바이어스 전류를 제거하거나 줄이는 기술이 개발되면 NCS의 에너지 효율을 크게 개선 가능

- 본 과제는 NCS의 큰 바이어스 전류를 짧은 편광 레이저 펄스로 대체하는 새로운 접근 방식 제안

※ 최신 스핀트로닉스 기반 NCS보다 최소 2~3배 더 낮은 에너지 소비와 더 빠른 동작 속도 기대

58) https://cordis.europa.eu/project/id/751089

4 Digital Europe Programme

○ PREVAIL 프로젝트[59][60]

※ A multi-hub Test and Experimentation Facility for edge AI hardware

주관기관	프랑스, CEA
연구기간	2022년 말부터 42개월간
연구내용	• 엣지 AI 하드웨어 연구를 위한 멀티 허브 실험 시설을 구축 • 유럽의 디지털 혁신을 지원하기 위해 고성능, 저전력 엣지 구성 요소와 기술을 검증할 수 있는, 신뢰성 높고 차별 없는, 테스트 시설을 유럽에 구축하려는 목표 • 4개의 주요 유럽 연구 및 기술 기관이 컨소시엄을 구성하여 첨단 300mm 제조, 설계 및 테스트 시설을 상호 협력을 통해 구축할 예정 • 각 기관의 기술을 더욱 발전시켜, 유럽이 혁신 기술을 기반으로 AI 프로토타입을 쉽게 제작하고 테스트할 수 있도록 지원
주요 파트너기관	CEA-Leti, Fraunhofer-Gesellschaft, imec, VTT Technical Research Centre of Finland Ltd

- 세부 연구 내용

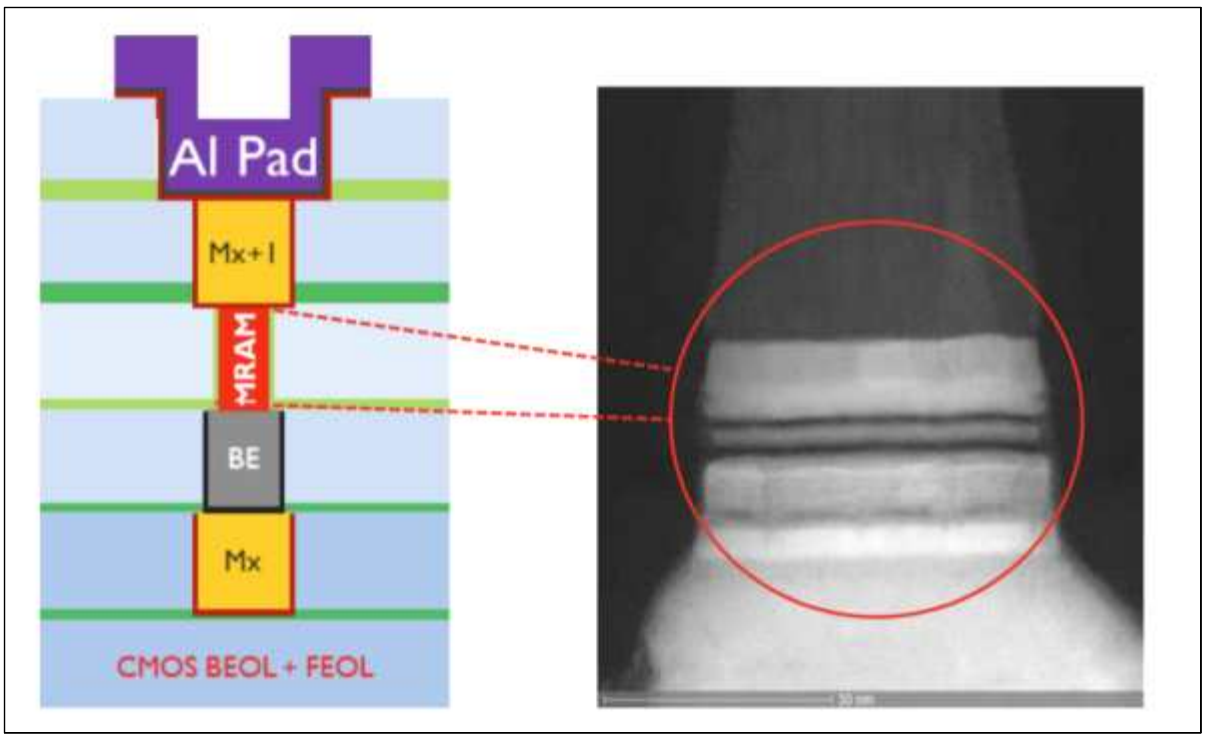
- **비휘발성 임베디드 메모리(MRAM, OxRAM, FeRAM)** : 초기에는 CMOS 기본 웨이퍼에 저항성 임베디드 비휘발성 메모리를 구현을 목표. 추후 MRAM(자기저항 랜덤 액세스 메모리), OxRAM(산화물 기반 RAM) 및 FeRAM(강유전성 RAM)과 같은 진보된 기술 연구
- **3D 이기종 통합 및 조립(다이 대 다이, 다이 대 웨이퍼, 다이 대 기판, 웨이퍼 대 웨이퍼)** : 3D 및 고급 패키징 기술을 사용한 이종 부품의 통합은 미래 시스템 확장의 핵심. 고성능, 저전력 시스템에 다양한 기능(컴퓨팅, 메모리, 캐시, IO)을 각각 최적화된 기술로 구현하고 3D 첨단 패키징으로 통합. 본 프로젝트에서는 인터포저, 다이-웨이퍼, 웨이퍼-웨이퍼, TSV(Through-Silicon Via)와 같은 다양한 상호연결 기술을 연구
- **통합 포토닉스 및 RF 부품** : 최근 몇 년 동안 최첨단 포토닉스 집적 회로(PIC)는 초고속 인공신경망 시스템의 계산 속도와 전력 효율성을 높였음. 보유 중인 세계 최고 수준의 첨단 PIC 프로세스를 활용하고, 최신 기술이 적용된 엣지 AI칩 기반 솔루션을 검증할 수 있는 진보된 테스트베드를 제공

59) https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/digital-2021-cloud-ai-01-tef-edge
60) https://www.imec-int.com/en/press/eu-consortium-developing-next-gen-edge-ai-technologies-accepting-design-proposals

- 컨소시엄 기관별 역할

CEA	에너지, 국방 및 보안, 정보 기술 및 보건 기술 분야를 연구하는 프랑스 기관
CEA-Leti	컨소시엄 관리 역할을 담당하는 본 기관은 약 11,000m^{2}의 클린룸 공간과 300mm 제조 능력 보유
Fraunhofer-Gesellschaft	프라운호퍼의 다수의 연구소가 연합하여 프로젝트에 참여 중. 프라운호퍼 IPMS의 나노전자공학 기술 센터(CNT)와 프라운호퍼 IZM-ASSID (Center All Silicon System Integration Dresden)은 CMOS 나노전자 기술과 진보된 3D 웨이퍼-레벨 시스템 통합을 위한 300mm 제조라인을 지원. Fraunhofer EMFT는 신뢰성 있는 물리적 해석과 칩-투-포일 통합 기여. Fraunhofer IIS는 22nm 수준의 칩 설계 및 테스트 경험을 바탕으로 CMOS IC를 설계, 측정 및 테스트에 기여
imec	300mm 제조 인프라를 포함하여, 약 12,000m^{2}의 클린룸 공간을 갖춘 imec은 메모리, 3D 및 포토닉스 연구팀과 단위 공정 모듈 연구팀을 통해 PREVAIL 과제에서 제공 가능한 기술을 정의하는 역할 수행

<Imec의 STT-MRAM 통합기술 (출처: https://prevail-project.eu/offer/non-volatile-memories/)>



Ⅴ. 결론

1 요약

- **100여년 전부터 한 걸음씩 진보를 거듭해온 인공지능 기술은 인터넷 활용의 확대와 첨단반도체 기술의 발달 덕분에 급성장 단계에 진입**
- **딥러닝으로 대표되는 인공지능 기술혁신과 더불어 인공지능 반도체 시장은 1세대(CPU+GPU)를 지나 2세대 솔루션인 NPU, HBM, PIM을 통해 폭발적인 성장**
  - 메모리 반도체 강국인 우리나라는 인공지능 반도체 2세대 시장에서도 HBM 기술 고도화를 통해 시장에서 중요한 역할 담당
  - 삼성전자는 2018년 출시한 엑시노스9 AP부터 자체 개발한 NPU를 탑재해왔으며, 최근 갤럭시 S24에는 개량된 NPU를 포함한 엑시노스 2400 AP를 장착하여 실시간 통역기능을 제공
- **최근 급성장세를 보이는 생성형 AI의 경우처럼 인공지능은 학습량이 많을수록 더 똑똑해지는 특성으로 인해 더 높은 지능을 얻기 위해 업체간 무한 경쟁 상태**
  - 학습량을 경쟁적으로 늘리면서 에너지 소모와 시스템 구성 비용이 천문학적으로 증가
  - 차세대 인공지능 반도체의 화두는 에너지 절감과 소형화
- **3세대 인공지능 반도체 기술(뉴로모픽 컴퓨팅, 멤리스터)은 여전히 연구단계**
  - 차세대 인공지능 반도체로 주목받는 뉴로모픽 컴퓨팅 기술에 관한 연구는 전 세계적으로 활발
  - IBM은 2014년 26억 5600만 개의 시냅스를 가진 뉴로모픽 칩인 '트루노스(TrueNorth)'를, 퀄컴은 2013년 뇌와 같이 학습하는 연산처리장치 '제로스(Zeroth)'를, 인텔은 2017년, 2021년 뉴로모픽 칩 '로이히(Loihi)'와 '로이히2'를 발표

ㅇ **3세대 뉴로모픽 인공지능 반도체 기술은 스파이킹 신경망과 멤리스터 소자 분야에서 연구 활발**

- EU의 프로젝트 사례로 살펴본 3세대 연구는 크게 SNN에 기반한 학습과 추론 알고리즘을 개발하는 분야와 멤리스터로 대표되는 PRAM, MRAM, RRAM, FRAM 등과 같은 뉴런의 신경전달 체계를 모사하기 위한 비휘발성 메모리 소자를 개발하는 분야로 나뉘어 진행 중

2 시사점

ㅇ **스파이킹 신경망 기반 뉴로모픽 반도체는 높은 에너지 효율과 소형화가 가능해 향후 인공지능 반도체 시장에 큰 파급효과**

- PIM, NPU, HBM과 같은 2세대 기술의 성능개량은 시장을 주도 중인 세계적 기업들이 심혈을 기울이는 형국
- 현재 유럽의 인공지능 관련 연구 프로젝트는 3세대 기술에 집중되어 있음
- 뇌를 모델로 한 소자와 시스템 기술은 에너지 효율, 신뢰성, 소형화 측면에서 근본적인 한계를 극복하고, 인간과 같은 지능과 에너지 효율을 갖춘 시스템으로 진화할 것임
- 뇌의 생물학적인 동작을 최대한 유사하게 모사하려는 뉴로모픽 연구는 알고리즘, 재료와 소자기술, 집적기술, 플랫폼 기술의 통합을 통해 기술 진보로 이어질 것

ㅇ **인공지능 반도체로 인한 국제 질서 변화에 대응**

- 산업 전 분야에 인공지능의 활용이 급속히 확산되면서 인공지능 반도체 기술이 전략자산으로 인식되어 통상과 정치 및 외교 분야에서 경쟁과 견제 심화
- 인공지능 반도체 기술을 선도하는 주요국의 하나인 대만과 미국은 자국 산업생태계 육성을 중점으로 R&D 협력을 추진하는, 반면 EU는 국가 간 협력을 근간으로 기술 역량을 확장하는 구조이므로 비유럽 국가와의 협력에도 개방적인 성격이 있음

- 미국을 필두로 대만 및 EU와 같은 주요국은 이른바 반도체 법을 제정하여, 기술개발과 자국의 생산역량 확대 및 인력양성에 지원을 확대
- 대내외 환경 변화에 유연하게 대응하기 위해서 우리나라는 글로벌 협력과 국내 산업육성, 미국의 기술 블록화 대응 측면을 면밀히 고려해야 함

ㅇ 국제협력 생태계 참여 활성화

- 인공지능 반도체 시장은 하나의 국가에서 기술과 시장을 독점하는 것이 아닌 국가별 협력과 연대를 통해 발전
- 인공지능 반도체 기술 난이도 상승에 따라 국가 간 분업이 심화되는 특성을 보이는 첨단반도체 산업에서는 상호 협력의 중요성이 더욱 강조됨. 유럽, 미국, 일본, 대만 등 반도체 선진국들과의 협력 확대 필요
- 한국은 국제협력 생태계에 참여 또는 조성에 노력 필요. 이를 위해 정치·외교적인 측면에서 정부의 지원과 함께 산학연의 관심 중요
- 국제 협력과 상생을 위해서는 우리나라에 높은 기술력과 잘 갖추어진 연구 인프라 보유가 전제되어야 함
- 인공지능 반도체 분야에서 유럽은 인간 뇌 프로젝트(HBP)를 통해 기술 경쟁력을 키워왔고, 호라이즌 프레임워크 프로그램과 EU 국가별 지원 아래 연구개발 진행 중. 우리나라는 호라이즌 유럽 필라2의 준회원국 가입을 유럽과의 협력 확대를 위한 발판으로 활용
- 국제 공동연구와 협력을 통해 뉴로모픽 컴퓨팅 플랫폼과 지식을 공유
- 인공지능 반도체 3세대 시장은 발전 가능성 크지만, 아직 미성숙 단계에 있어 미래 경쟁력 강화를 위한 적극 투자 필요

ㅇ 인공지능 반도체는 우리의 국가안보와 경쟁력 제고를 위한 중요기술

- 연구기관과 국내 파운드리, 대학 간 밀접한 협력관계는 국내 반도체 산업의 발전적인 장기 비전을 위해 필요
- SNN과 같은 뉴로모픽 신경망 알고리즘 연구의 활성화를 위해서는 유럽의 SpiNNaker와 BrainScales와 같이, 관련 연구원이 쉽게 접근할 수 있는 개방된 뉴로모픽 컴퓨팅 플랫폼이 필요하며, 국가차원의 지원이 효과적

- 인공지능 반도체 분야는 고도의 기술력과 창의적인 아이디어를 보유한 핵심인력이 경쟁력을 좌우하는 기술집약적 산업인 만큼, 대학 교육 강화와 학·연·산 협력을 통한 실무역량 제고가 필요하며, 글로벌 석학 유치 노력과 함께 국내 전문가의 국외 유출 우려가 없는 건강한 생태계 조성을 위한 노력 필요
- 인공지능 반도체 기술은 체계적이고 전략적인 산업육성을 통해서 우리나라의 미래 핵심 성장동력이 될 것으로 기대