2024 ISSUE REPORT

이슈보고서

혁신성장금융1부

VOL.2024-이슈 (2024.5)

AI반도체 시장 현황 및 전망

CONTENTS

작성
선임연구원 이미혜 (6255-5404)

한국수출입은행 | 해외경제연구소

< 요 약 >

Ⅰ. AI 및 AI반도체 개요

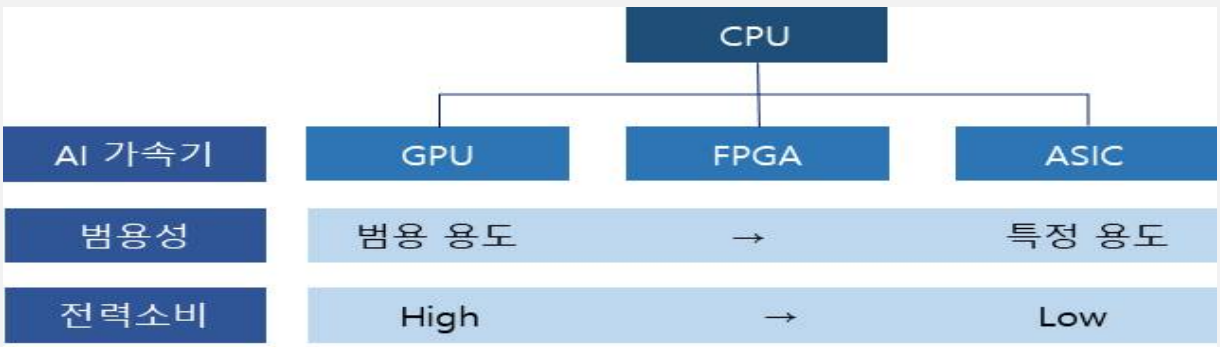
AI반도체는 AI 알고리즘을 실행할 수 있는 반도체로 광의로는 CPU를 포함하나 협의로는 AI 연산 가속(AI Accelerator)이 목적인 GPU, FPGA, ASIC을 포함

- AI 연구 초기에는 CPU를 사용했으나 딥러닝 등장 이후 데이터의 양과 매개변수의 수가 가파르게 증가하면서 GPU(Graphic Processing Unit)가 사용됨

- GPU는 단순 계산을 병렬(parallel)로 처리하여 대규모 데이터 처리에 효율적이나 AI 연산에 최적화된 구조가 아니어서 전력소모가 큰 단점이 있음

- GPU를 보완하기 위해 AI의 특정 연산에 특화된 FPGA(Field-Programmable Gate Array), ASIC(Application Specific Integrated Circuit) 사용 증가

AI 구현에 사용되는 반도체



Ⅱ. AI반도체 시장 현황 및 전망

시스템반도체의 차세대 성장동력인 AI반도체 시장규모는 2022년 411억 달러에서 2028년 1,330억 달러로 연평균 21.6% 성장할 전망

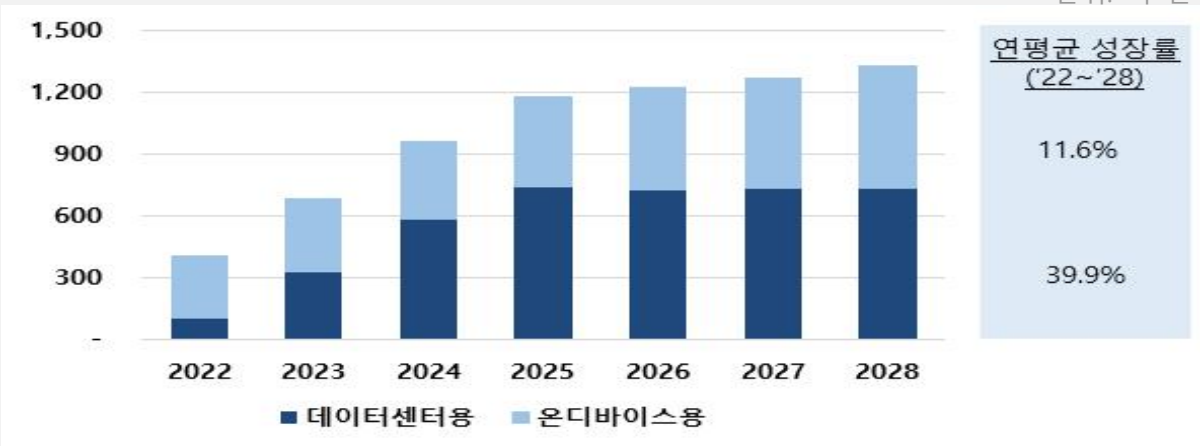
- 데이터센터용 AI반도체 시장은 2022년 97억 달러에서 2028년 730억 달러로 연평균 39.9% 성장할 전망

- 온디바이스용 AI반도체 시장은 2022년 314억 달러에서 2028년 602억 달러로 연평균 11.6% 성장할 전망

- 스마트폰, PC, 자동차, 사물인터넷(IoT) 등이 수요를 견인할 전망

AI반도체 시장 전망

단위: 억 달러



자료: 옴디아.

데이터센터용 반도체 시장은 엔비디아가 독점, 온디바이스용 AI반도체 시장은 반도체 기업과 수요기업들이 참여하며 스마트폰, 자동차 등 수요처별로 상이한 경쟁구도를 형성

- (데이터센터 GPU) 시장점유율은 2022년 엔비디아 97.3%, AMD 1.2%, 인텔 0.8% 순이나 2027년에는 엔비디아 87.0%, AMD 7.7%, 인텔 5.3% 순으로 전망
- (스마트폰) 반도체기업 퀄컴·미디어텍과 스마트폰 기업 애플, 삼성전자 등이 참여
- (PC) 인텔, AMD, 퀄컴 등이 PC용 고성능 AI반도체를 출시
- (자동차) 인텔, 엔비디아, 퀄컴, 테슬라, 차량용 반도체기업 NXP 등이 참여

엔비디아는 후발주자들의 거센 추격에도 불구하고 기술, 생태계, 재무적 우위 등으로 상당 기간 AI반도체 시장을 주도할 전망

- 엔비디아는 AI반도체 시장 지배력을 유지하기 위해 제품 출시 주기 단축, AI 관련 종합 솔루션 제공, 가격인하, 고객 맞춤형 AI반도체를 설계 지원 등을 추진

- 엔비디아 경쟁사의 제품 출시 주기는 2년내외, 엔비디아의 제품 출시 주기는 1년으로 경쟁사가 고성능 칩을 출시해도 엔비디아의 신제품에 역전당할 가능성이 큼
- 엔비디아는 AI 관련 종합 솔루션을 제공하고 있으나 AI반도체 스타트업은 AI 종합 솔루션을 제공하기에는 자원 또는 역량이 부족

Ⅲ. 주요국 육성정책

(미국) 민간기업이 AI반도체 개발을 주도하며 국방부는 차세대 반도체 리더십 확보를 위한 장기적인 기술과제 해결, 상무부는 국내 반도체 제조시설 구축 등을 지원

- 국방부 산하 방위고등연구계획국(DARPA)은 산·학·연 중심의 중장기 프로젝트를 통해 AI반도체 기초·원천기술 개발에 주력
- 2022년 8월, 미국은 반도체법(Chips & Science Act)를 제정하고 미국내 첨단 반도체 제조기반 구축, 산업생태계 조성 등을 지원

(중국) 2030년 세계 1위 AI 국가로 도약을 추진했으나 미국의 제재 등으로 AI 연산을 위한 컴퓨팅 자원 확보가 어려워져 자국 AI반도체 육성 지원을 강화

- '14차 5개년 계획('21~'25)'에서는 인공지능, 반도체 등의 핵심기술 개발을 강조하며, 자국 AI반도체 산업 육성을 위해 제조 지원, 데이터센터 프로젝트 등을 추진

- 중국은 미국의 제재로 AI반도체 구매가 제약을 받자 중국산 AI반도체 사용시 보조금 지급, AI반도체 수요 창출을 위한 데이터센터 프로젝트 등을 추진

Ⅳ. 한국의 AI반도체 산업 현황

국내 AI반도체기업은 10여개로 모바일, 가전 등 온디바이스 부문에서 일부 제품을 상용화 했으며 데이터센터 부문은 Reference를 구축하고 사업 본격화 단계

- 데이터센터용 AI반도체는 퓨리오사AI, 리벨리온, 사피온 등이 NPU를 개발해 클라우드에 시범 적용했으며 매출은 2024~2025년부터 본격화될 전망
- 온디바이스용 AI반도체는 DeepX, 텔레칩스, 삼성전자, LG전자 등이 참여

한국의 AI반도체 기술수준은 선도국인 미국 대비 80.0%, 기술격차는 2.5년

- 한국의 AI반도체 기술수준은 미국의 기술수준을 100로 볼 때 80.0로 중국(90.0), 유럽(85.0)보다 낮으나 일본(70.0)보다 높은 수준

- 한국의 AI반도체 특허 출원 건수는 세계 3위이나 질적 수준은 미흡

정부는 '인공지능 반도체 산업 발전전략('20)'을 수립하고 2030년 글로벌 시장점유율 20% 달성을 추진

- 인공지능 반도체 산업 발전전략은 First Mover형 혁신 기술·인재 확보와 혁신성장형 산업 생태계 조성을 지원
- 2024년, 정부는 AI G3 도약을 위해 'AI-반도체 이니셔티브' 추진 계획을 발표하고 AI모델, AI반도체*, AI서비스 관련 기술혁신을 추진

  * 차세대 AI반도체 Processing in Memory(PIM), 저전력 K-AP, 신소자 및 첨단 패키징

Ⅴ. 결론 및 시사점

AI반도체는 국가안보와 경쟁력 제고 등을 위한 핵심기술로 한국의 주력산업 경쟁력 제고에 기여할 수 있어 정책적 지원이 필요

- 한국은 메모리반도체 강국이나 시스템반도체 경쟁력은 상대적으로 취약한 상황이며 AI반도체는 성장 초기 단계로 한국이 시스템반도체 경쟁력을 제고할 기회
- AI반도체 경쟁력 제고는 한국의 주력 수출산업인 휴대폰, 자동차, 조선, 가전 등을 Smart하게 만들어주어 산업 경쟁력 제고에 기여할 수 있음

AI반도체 기업을 육성하기 위해 개발자금 지원, Reference 구축, 수요산업과 협력 강화, '팹리스-파운드리'의 유기적 협력관계 구축 등이 요구됨

- 국내 AI반도체 기업은 대부분 중소기업으로 개발비 부담 등을 경감시켜주기 위해 첨단 파운드리 공정 이용 등에 대한 정책적 지원이 필요
- '팹리스-파운드리'간 유기적 협력관계 구축을 통해 AI반도체 기업의 성장이 국내 파운드의 성장을 견인하는 선순환 구조 구축 가능

한국은 AI반도체, 초거대 AI 플랫폼 등을 기반으로 소버린 AI를 추진하나 국가안보, 지정학적 이슈 등에 민감하거나 기술력이 부족한 국가와 협력할 기회가 있을 전망

- 중동·유럽 등이 국가안보, 지정학적 이슈 등으로 AI 반도체 공급처 다변화를 추진할 것으로 예상되어 한국기업의 해외진출에 대한 정부의 외교적 지원 등이 필요

Ⅰ. AI 및 AI반도체 개요

1. 인공지능의 부상

인공지능은 1950년대에 처음 등장한 이후 머신러닝(Machine Learning), 딥러닝(Deep Learning)으로 발전

- 인공지능은 기계가 독립적으로 문제 해결, 추론, 학습, 지식표현 등을 내릴 수 있게 하는 컴퓨터 과학의 한 분야로 1950년대에 처음으로 등장
  - 초기 인공지능 연구는 규칙과 논리를 사용하여 지식을 표현하고 추론했으며 제한적인 컴퓨팅 파워 등으로 인해 단순한 알고리즘과 작은 데이터 세트에 국한되었음
- 1980년대에 머신러닝이 부상했으며 2010년대에는 머신러닝의 한 방법인 딥러닝이 부상
  - 머신러닝은 인터넷의 등장과 데이터 양의 급증 등으로 주목받기 시작했으며 정형 데이터(데이터베이스 등) 분석을 통해 패턴(특징)을 찾고 패턴을 기반으로 추론하며 사람이 소프트웨어가 분석해야 하는 특징 집합을 수동으로 결정
  - 딥러닝은 비정형 데이터 등도 인간의 뇌 신경망(Neural network)을 모방한 인공신경망을 이용하여 스스로 학습
    - 인간의 두뇌와 비슷한 계층 구조로 상호 연결된 노드 또는 뉴런을 사용하며 기본 신경망은 입력 계층, 은닉 계층, 출력 계층(레이어)으로 인공 뉴런을 상호 연결

인공지능의 역사

자료: 삼성SDS.

딥러닝의 주류 인공신경망 알고리즘은 CNN(Convolutional Neural Network), RNN (Recurrent Neural Network)에서 Transformer로 진화

- CNN(Convolutional Neural Network, 합성곱[1] 신경망)은 이미지 인식 분야에서 뛰어난 성능을 발휘하는 알고리즘으로 객체인식, 분류 등의 작업에 활용

- 합성곱 신경망의 은닉 계층은 합성곱을 수행하여 이미지에서 관련 특징을 추출하며 타 알고리즘 대비 빠른 학습 속도 등이 장점

- RNN(Recurrent Neural Network, 순환신경망)은 자연어 처리 등에 활용되며 입력 노드에서 출력 노드까지 한 방향으로 시간 순서에 따라 데이터를 처리

- 순차적 데이터 처리로 연산속도가 느리고, 긴 문장 번역시 오역이 증가하는 단점 발생

- Transfomer는 2017년 구글이 발표한 모델로 Attention 매커니즘을 사용하여 문장 내 단어의 중요성을 가중치를 고려해 상관관계를 효과적으로 표현하고 학습

- RNN(순환 신경망)의 한계를 극복하기 위해 문장 전체를 병렬구조로 처리하여 텍스트와 음성을 거의 실시간으로 옮김

  * Chat-GPT(Generative Pre-trained Transfomer) 등 생성형 AI는 Transformer 알고리즘을 기반으로 개발됨

- Transformer는 대규모 데이터 학습에 적합한 방식이나 데이터의 양이 작으면 성능 한계가 있어 아직 CNN(합성곱 신경망), RNN(순환 신경망)도 다수 사용

딥러닝 신경망 비교

종류	내용
합성곱 신경망 (CNN)	·합성곱 계층을 통해 데이터의 특징을 추출하고 그 특징들을 기반으로 분류 ·이미지, 영상 데이터 처리에 주로 사용
순환 신경망 (RNN)	·순차적 데이터(문장, 시계열 데이터 등) 처리에 특화된 모델 ·자연어 처리에 사용되며 연산속도가 느리고 긴 문장 번역시 오역 증가
Transformer	·Attention 매커니즘을 사용해 멀리 있는 단어까지도 연관성을 만들어 단어간 상관관계를 효과적으로 표현하고 학습하며 처리 속도가 빠름 ·현재 이미지나 언어 번역 기능으로 폭넓게 사용됨

자료: AI타임즈.

1) 하나의 함수를 반전 이동한 것과 또 다른 함수를 곱한 후 구간에 대해 적분하여 새로운 함수를 구하는 수학 연산

딥러닝의 성능은 AI 연산력의 기준이 되는 매개변수(Parameter)가 증가할수록 향상되며 매개변수의 수가 가파르게 증가하면서 AI반도체의 중요성이 커짐

- 매개변수는 AI 모델 내부에서 결정되는 변수로 AI 훈련 과정에서 Input이 원하는 Output으로 변환될 수 있도록 조정하는 가중치

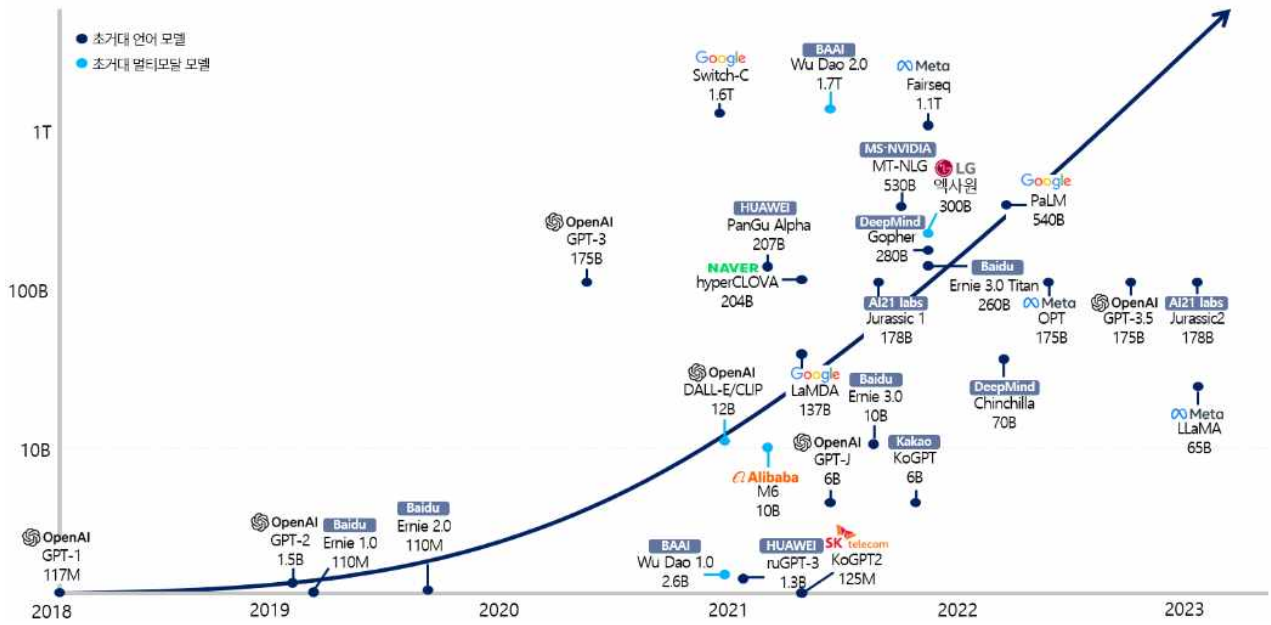
- AI 모델은 Input이 다수의 딥러닝 레이어를 통과하여 Output을 도출하며 AI 모델 학습은 각층에 있는 가중치를 찾는 것으로 가중치에 따라 모델의 정확성이 결정됨

- 1950년대부터 2018년까지 개발된 AI 모델들의 매개변수가 두 배 증가하는데 걸린 시간은 39.7개월이었으나 2018년부터 2022년까지는 불과 4개월이 소요됨[2)]

- GPT-3의 매개변수는 1,750개, GPT-4의 매개변수는 1조개 이상, 개발중인 GPT-5의 매개변수는 125조개

  · 주요 AI기업은 텍스트 중심의 대규모언어모델(LLM), 이미지·음성·영상을 생성하는 멀티모달(Multi Modal) 기반의 AI를 넘어 인간과 모든 분야에서 비슷하거나 똑똑한 AI(Artificial General Intelligence, AGI) 개발을 추진

초거대 AI 개발 동향(파라미터 수)



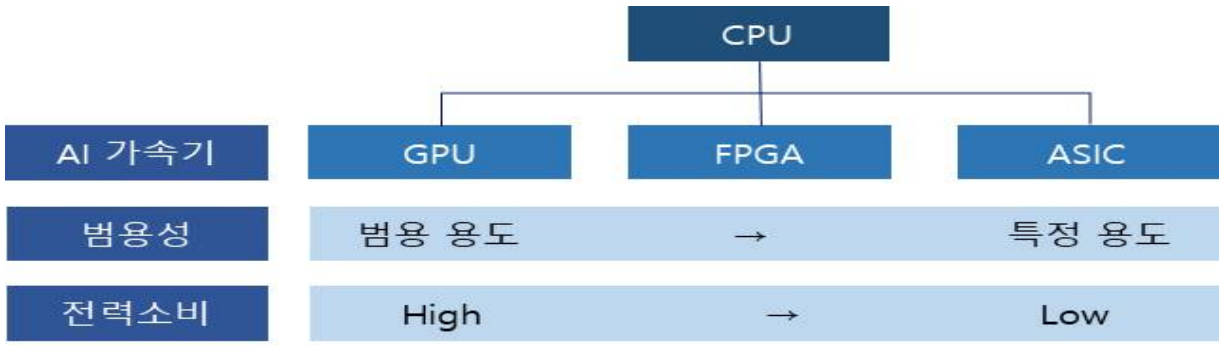
자료: KDI.

2) 한화투자증권

AI반도체는 AI 알고리즘을 실행할 수 있는 반도체로 광의로는 CPU를 포함하나 협의로는 AI 연산 가속(AI Accelerator)이 목적인 GPU, FPGA, ASIC을 포함

- CPU(Central Processing Unit)는 컴퓨터의 주 연산장치로 복잡한 연산을 순차적(sequential)으로 처리하여 대량의 데이터 연산에 긴 시간이 소요됨
  - CPU는 컴퓨터의 두뇌역할을 담당하며 운영체제(OS) 등을 원활하게 작동하도록 컨트롤
- AI 연구 초기에는 CPU를 사용했으나 딥러닝 등장 이후 데이터의 양과 매개변수의 수가 증가하면서 GPU(Graphic Processing Unit)가 사용됨
  - GPU는 단순 계산을 병렬(parallel)로 처리하여 대규모 데이터 처리에 효율적이며, 범용성(Flexibility)이 높아 다양한 AI 알고리즘을 사용할 수 있음
  - 그러나 GPU는 그래픽 처리를 위해 개발된 프로세서로 AI 연산에 최적화된 구조가 아니어서 전력소모가 큰 단점이 있음
- GPU를 보완하기 위해 AI의 특정 연산에 특화된 FPGA(Field-Programmable Gate Array), ASIC(Application Specific Integrated Circuit)[3] 사용 증가
  - FPGA는 소프트웨어를 업데이트해 용도에 맞게 내부 회로를 바꿀 수 있는 반맞춤형 반도체로 AI 알고리즘이 확정되지 않은 상황에 적합
    - ASIC 대비 칩 설계 시간 단축, 초기 비용 절감으로 소량 생산, 맞춤형 반도체 설계를 테스트하는데 적합
  - ASIC은 특정 AI 알고리즘에 특화된 반도체로 범용성이 낮으나 저전력 등이 장점이며 대표적인 프로세서는 NPU(Neural Processing Unit, 신경망처리장치)
    - NPU는 인공신경망 연산에 특화된 반도체로 저정밀 연산으로 고속 연산, 저전력 구현이 가능하나 구조상 다른 AI 알고리즘 사용시 연산 자체가 불가능하거나 속도가 크게 떨어짐
    - 기업들은 FPGA를 이용해 를 개발하고 성과가 좋으면 ASIC으로 대량 생산

AI 구현에 사용되는 반도체



3) NPU(Neural Processing Unit), TPU(Tensor Processing Unit), VPU(Vision Processing Unit), IPU (Intelligence Processing Unit) 등을 포함

AI반도체는 서비스 플랫폼에 따라 데이터센터용과 온디바이스용(엣지디바이스용), 사용 목적에 따라 학습용(Training)과 추론용(Inference)으로 분류

- AI반도체는 서비스 플랫폼에 따라 데이터센터용(서버용)과 온디바이스용(엣지디바이스용)으로 분류

- 데이터센터용 AI반도체는 병렬연산 처리능력과 저전력*이 중요하며, 운영 측면에서 확장성과 유연성도 고려 대상

  * OpenAI가 Chat-GPT를 운영하는데 사용되는 전기료는 연간 6,000억원 수준으로 추정

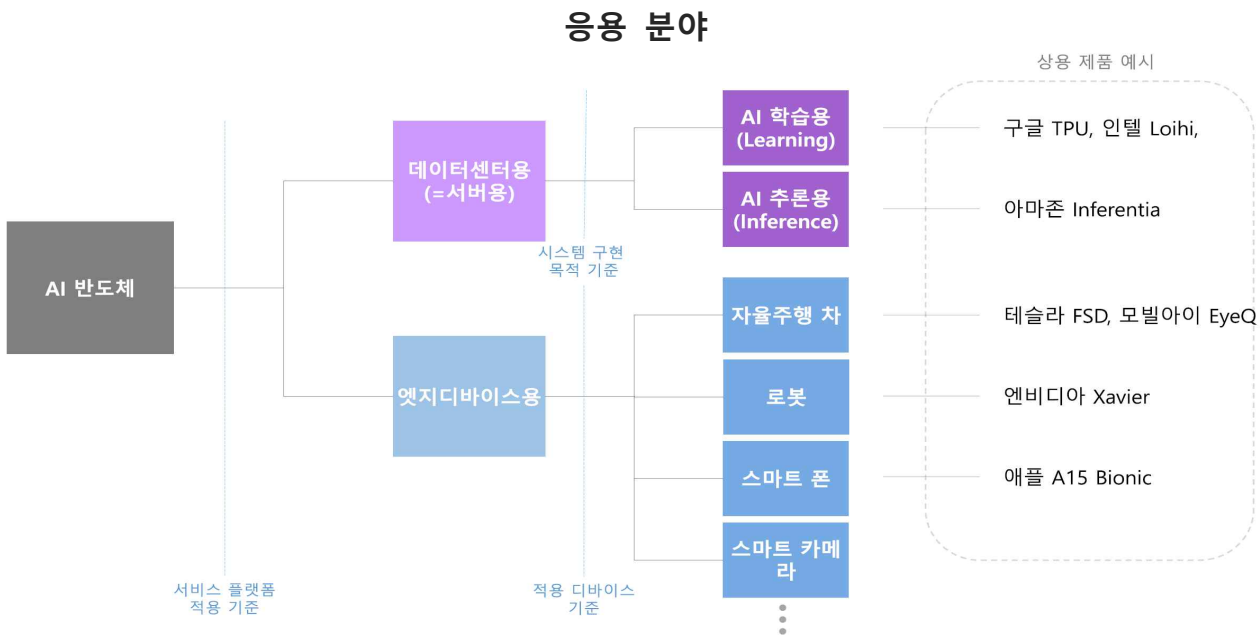
- 스마트폰, 자동차, 드론, IoT 등 개별 서비스에 특화된 온디바이스용 AI반도체는 저전력, 가격, 크기 등이 핵심 경쟁요소

- 학습은 대량의 데이터를 기반으로 AI 모델을 개발, 추론은 AI 모델을 기반으로 데이터를 분석

- 학습용 AI반도체는 주로 클라우드[4]에서 사용되며 연산속도 등이 중요, GPU 중심

- 추론용 반도체는 클라우드와 엣지에서 모두 사용되며 정확도, 저지연성 등이 중요하며 CPU, GPU, NPU[5]가 사용됨

  · 추론은 사용되는 디바이스별로 필요한 연산 능력이 다양함



자료: 네이버.

4) 클라우드 컴퓨팅은 하드웨어, 소프트웨어 등의 정보자원을 직접 구축·운영하지 않고 네트워크에 접속해 이용하는 기술로 이용자가 사용한 만큼 과금되는 구조
5) NPU는 독립적인 반도체 칩으로 설계되기도 하고 혹은 반도체 칩 내부의 일부분으로 설계되기도 함

Ⅱ. AI반도체 시장 현황 및 전망

1. 시장규모

시스템반도체의 차세대 성장동력인 AI반도체 시장규모는 2022년 411억 달러에서 2028년 1,330억 달러로 연평균 21.6% 성장할 전망[6)]

- 데이터센터용 AI반도체 시장은 2022년 97억 달러에서 2028년 730억 달러로 연평균 39.9% 성장할 전망

- AI서버 출하량은 2023년 1.2백만대에서 2026년 2.4백만대로 확대되면서 서버 출하량중 AI서버 비중은 '23년 약 9%에서 2026년 15%로 증가할 전망

- 온디바이스용 AI반도체 시장은 2022년 314억 달러에서 2028년 602억 달러로 연평균 11.6% 성장할 전망

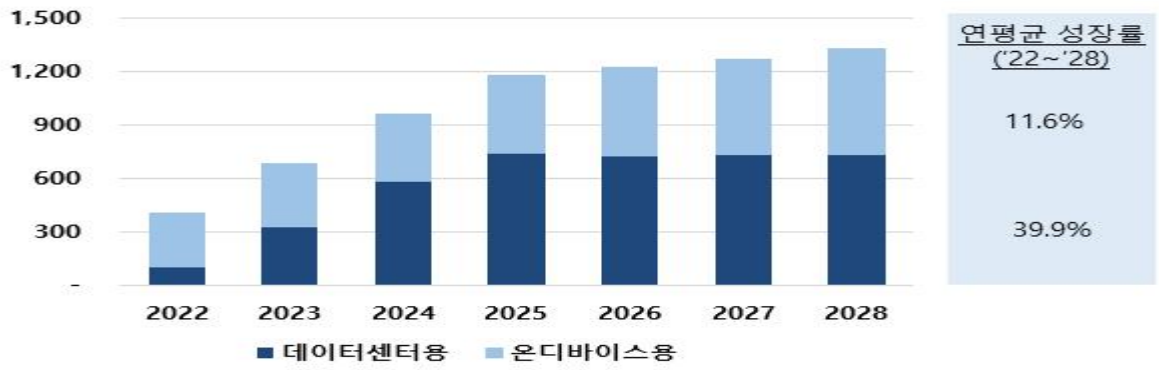
- (스마트폰) 생성형 AI 스마트폰 출하량은 2023년 47백만대에서 2027년 5.2억대로 연평균 83% 성장하여 2027년 스마트폰 출하량의 40%를 차지할 전망
  · 2017년 아이폰8의 AP(Application Processor)가 NPU를 처음 탑재한 이후 스마트폰 AP의 NPU 탑재가 증가했으며 생성형 AI를 지원하는 고성능 AP는 2023년 하반기부터 출시됨

- (PC) 온디바이스 AI PC 출하량은 2024년 약 5천만대에서 2027년 1.7억 대로 성장하여 AI PC 비중은 2024년 19%에서 2027년 약 60%로 확대될 전망

- (자동차) ADAS(Advanced Driver Assistance Systems, 운전자 지원 시스템), 인포테인먼트 등이 성장을 견인할 전망

시장 전망

단위: 억 달러



자료: 옴디아.

6) 기관별로 AI반도체의 범주, 시장규모 전망 등이 상이하나 모두 고성장을 전망(AMD는 AI반도체시장이 2023년 450억 달러에서 2027년 약 4,000억 달러로 성장 전망)

2. 경쟁구도

데이터센터용 반도체 시장은 엔비디아가 독점하는 가운데 AMD, 인텔 등이 엔비디아를 추격

- 엔비디아의 분기별 데이터센터 부문 매출은 2022년 1분기에 20억 달러에서 2023년 4분기 184억 달러로 9배 증가, 동기간 AMD의 매출은 2배 증가, 인텔의 매출은 감소

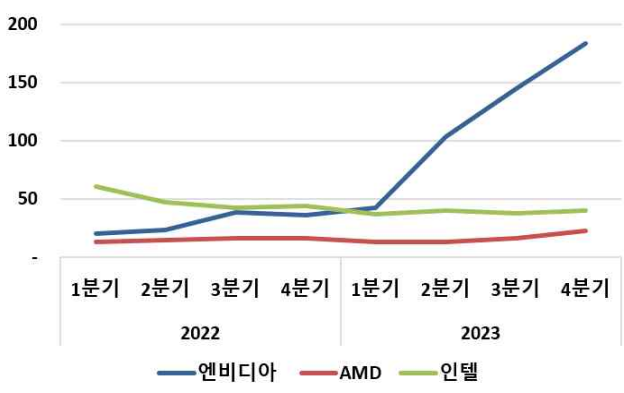
- 데이터센터 CPU 시장점유율은 2021년 인텔 80.7%, AMD 11.7% 순이였으나 2022년에는 인텔 70.8%, AMD 19.8%, 아마존웹서비스(AWS) 3.2% 순
- 데이터센터 GPU 시장점유율은 2022년 엔비디아 97.3%, AMD 1.2%, 인텔 0.8% 순이나 2027년에는 엔비디아 87.0%, AMD 7.7%, 인텔 5.3% 순으로 전망

- 엔비디아의 핵심 고객인 대형 클라우드는 2023년 생성형 AI 붐으로 GPU 공급부족 등을 경험하면서 2024년부터 엔비디아 의존도를 낮추기 위한 노력을 강화할 전망

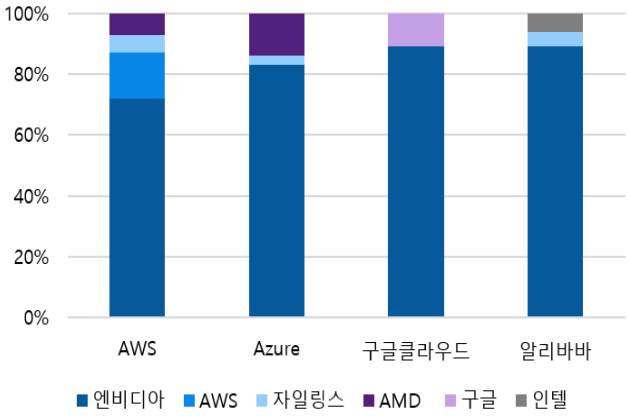
- 2023년, 엔비디아 H100의 수요가 급증하자 H100의 납품주기는 8~11개월, 가격도 3~4만 달러 수준을 형성하여 칩 구매도 어렵고 가격 부담도 증가
  · AI서버 비용의 70%는 GPU가 차지하며 일반 서버 대비 약 10배 높은 투자 비용이 소요됨
- 클라우드는 AMD 비중 확대, 자체 개발한 AI반도체 사용 노력 등을 강화
  · 주요 클라우드는 자사 workload에 최적화된 반도체를 개발했으나 엔비디아 생태계 탈피의 어려움 등으로 일부만 사용되었음. 향후 자체 개발한 반도체 사용 확대 추진
  · 메타는 2024년에 총 60만개의 H100급 GPU를 추가할 계획이며 이중 35만개(약 60%)는 엔비디아 H100, 25만개는 AMD 또는 자체 개발 칩 탑재를 추진

주요 기업의 데이터센터 부문 매출

단위: 억 달러



클라우드의 탑재 비중 ('22.6)



주: Azure는 마이크로소프트의 클라우드사업이며, 자일링스는 AMD에 인수됨('22)
자료: 블룸버그, Liftr insights.

온디바이스용 AI반도체는 반도체기업과 수요기업들이 참여하며 스마트폰, 자동차 등 수요처별로 상이한 경쟁구도를 형성

- (스마트폰) 반도체기업 퀄컴, 미디어텍과 스마트폰기업 애플, 삼성전자 등이 참여

- 2023년 하반기부터 생성형AI를 지원하는 스마트폰용 AP인 퀄컴의 스냅드래곤8 3세대[7], 미디어텍의 디멘시티 9300, 삼성전자의 엑시노스 2400 등이 출시됨

  · 생성형AI를 지원하는 스마트폰은 30TOPS[8] 이상의 NPU를 활용하며, 퀄컴의 2024~2025년 시장점유율은 80% 이상이 될 전망

- (PC) 인텔, AMD, 퀄컴 등이 PC용 고성능 AI반도체를 출시

- 인텔의 Meteor Lake 칩은 CPU, GPU, NPU를 결합해 최대 34TOPS를 제공하며 2024년말까지 출시 예정인 차세대 Lunar Lake는 40TOPS 이상의 계산 능력 보유할 전망

  · 2023년말 인텔은 노트북용 CPU 메테오레이크에 처음으로 NPU를 탑재

- AMD의 Ryzen 8000은 45TOPS를, 2024년에 출시될 퀄컴의 Snapdragon X Elite는 약 45 TOPS를 제공할 전망

- (자동차) 인텔, 엔비디아, 퀄컴, 테슬라, 차량용 반도체기업 NXP[9] 등이 사업을 영위

- 인텔은 이스라엘 팹리스 모빌아이 인수를 통해 ADAS 시장지배력을 강화했으며, 퀄컴, 미디어텍 등 무선통신 등에 집중한 기업들이 차량용 칩 사업을 강화하는 추세

- (TV) AI반도체가 TV의 화질·음질 향상 등에 사용되며 미디어텍, 삼성전자 등이 사업을 영위

엣지용 기업 현황

수요처	주요 기업
스마트폰	퀄컴, 미디어텍, 삼성전자, 애플
PC	인텔, AMD, 퀄컴,
자동차	인텔(모빌아이), 엔비디아, 퀄컴, 테슬라, Horizon Robotics(중)
스마트 TV	미디어텍, 삼성전자, LG전자
가정용 보안카메라	Ambarella(미), 퀄컴

자료: Yole.

7) 퀄컴 최초로 생성형AI에 최적화된 AP로 100억개 매개변수의 생성형AI 모델을 지원하며 퀄컴이 제공하는 Meta의 LLama2를 사용하거나 타 AI 모델을 넣을 수 있음
8) Tera Operations Per Second. 1초에 1조번의 연산을 처리
9) 2023년 하반기에 차량용 반도체기업 NXP와 ST마이크로는 AI MCU(Micro Controller Unit)를 출시

3. 주요 사업자

1) 반도체기업

엔비디아는 반도체기업에서 AI 서비스 개발에 필요한 하드웨어, 소프트웨어, 플랫폼을 제공하는 종합 AI 기업으로 도약

- 엔비디아는 1993년에 설립된 이후 20년 이상 GPU를 개발했으며 AI 프로그램 개발을 지원하는 소프트웨어 CUDA(Compute Unified Device Architecture)를 통해 강력한 진입장벽을 형성

- 엔비디아의 대표 상품은 소비자용 GPU '지포스(GeForce)', 컴퓨터 그래픽 전문가용 GPU '쿼드로(Quadro)', 데이터센터에서 사용되는 GPU '테슬라'(Tesla)' 시리즈 등
- CUDA는 AI 개발자들이 프로그래밍을 위해 사용하는 소프트웨어로 엔비디아의 GPU에서만 구동되며 CUDA가 업계 표준이 되면서 개발자들은 엔비디아 생태계에 종속

- 엔비디아는 AI 발전 가속화, 경쟁사의 등장 등에 대응하기 위해 GPU 출시 주기 단축, 경쟁력 있는 가격책정, 제품 포트폴리오 확대, 맞춤형 반도체사업 등을 추진

- 2023년 생성형 AI 붐으로 인해 요구되는 반도체의 성능이 높아지자 서버용 GPU 출시 주기를 2년에서 1년으로 단축
- 2024년말에 출시 예정인 Blackwell GPU의 가격은 3~4만 달러로 전작 대비 추론 성능은 30배 향상되나 H100의 가격(2.5~4만 달러) 수준으로 책정[10)]
- GPU에서 CPU, DPU(Data Processing Unit, 데이터처리장치)로 제품 포트폴리오 확대
  · Arm 기반의 데이터센터 CPU인 Grace를 개발하고 GPU와 하나로 결합한 Superchip을 통해 정보를 전송하는 과정을 단축해 저전력을 구현
  · DPU는 데이터 처리를 가속화하여 데이터센터 운영 효율성을 높이고 비용 절감을 지원
- Big Tech는 직접 AI반도체를 개발하고 있으나 성능, 대량 생산 등에 어려움을 겪고 있어 엔비디아는 고객 맞춤형 AI반도체를 설계하는 사업부 신설을 추진

엔비디아의 고성능 GPU 개발 계획

	2021	2022	2023	2024		2025
GPU	A100	-	H100	H200	B100	X100
CPU+GPU(Superchip)	-	-	-	GH200	GB200	GX200

주: 1) A는 Ampere, H는 Hopper, B는 Blackwell, X는 Xavier, G는 Grace 아키텍처
2) GB200은 Blackwell GPU 2개와 CPU를 연결
자료: 트렌드포스.

10) CNBC, Nvidia's latest AI chip will cost more than $30,000, CEO says, 2024.3.19

- 엔비디아는 AI 개발 및 활용에 필요한 하드웨어(칩), 소프트웨어(CUDA), 플랫폼을 제공하는 종합 AI 기업으로 사업모델 확대

- AI를 처음 적용하는 기업들이 엔비디아의 하드웨어, 소프트웨어, 플랫폼을 이용해 쉽게 AI 서비스를 구성하도록 지원

  · NeMo(Nvidia Enterprise Modular AI)는 특정 분야 AI 모델 구축을 지원하는 소프트웨어 플랫폼으로 산업별·기업별 AI 모델 개발을 지원

  · 신약 개발을 위한 AI 플랫폼 BioNeMo 등 특정 분야의 AI 모델 개발을 지원

엔비디아의 사업구조

자료: 한국경제.

AMD는 CPU에서 GPU, FPGA 등으로 사업을 확대했으며 Pervasive AI[11] 전략하에 향후 5년내 모든 제품에 AI엔진 탑재 추진

- AMD는 CPU와 GPU 시장에서 각각 인텔과 엔비디아에 이어 2위이며 GPU는 고성능 제품 개발과 엔비디아 대비 경쟁력 있는 가격 등을 통해 시장점유율 확대 추진

- AMD는 엔비디아의 동급 모델 대비 60~70% 수준의 가격을 제시하며, 마이크로소프트와 메타가 2023년말에 출시된 GPU MI300 채용을 추진

  · AMD는 2024년 AI반도체 매출을 40억 달러로 전망

- AMD는 FPGA 기업 Xilinx, 네트워킹 칩 및 관련 소프트웨어 스타트업 Pensando Systems 인수 등을 통해 제품 포트폴리오를 강화

- FPGA 시장은 인텔과 양강 구도를 형성하고 있으며 시장점유율 기준으로 1위를 기록

- Pensando Systems 인수('22)를 통해 DPU(Data Processing Unit) 사업을 강화

- Pervasive AI 전략하에 향후 5년내에 개인용 PC부터 HPC, 클라우드에 이르기까지 모든 제품에 AI엔진 탑재 추진

11) 모든 곳에 존재하는 AI

인텔은 CPU에서 GPU, FPGA로 사업을 확대하며 데이터센터 사업 강화 추진

- FPGA 기업 Altera, AI반도체 스타트업 하바나랩스 인수 등을 통해 데이터센터 공략

- 2016년 AI 가속기 스타트업 Nervana를 인수했으나 Nervana 기반 제품은 실패, 이후 2019년 하바나랩스, 차량용 반도체기업 모빌아이 등을 인수
- 하바나랩스는 AI 학습용 반도체 Gaudi를 3세대까지 개발했으며 Gaudi3는 2024년 3분기에 출시 예정으로 델, HP, 레노버, 슈퍼마이크로 등이 Gaudi3을 채택

2) Big Tech

AWS(Amazon Web Service)는 세계 1위 클라우드로 2018년부터 AI반도체를 개발했으며 타 클라우드 대비 자사 칩 개발 및 사용에 적극적

- 아마존과 AMD는 서버용 프로세서 개발을 위해 협력했으나 낮은 성능 등으로 양사는 결별하고 아마존은 이스라엘 팹리스 Annapurna Labs를 인수('15)
- 아마존은 자체 칩 개발을 추진하면서 ARM 아키텍처 기반의 CPU Gravition, 머신러닝에 특화된 Trainium, 추론용 칩 Inferentia 등을 발표

- 서버용 CPU Graviton은 ARM 아키텍처를 채용하여 인텔, AMD가 사용하는 기존 X86 아키텍처 기반 CPU 대비 저전력을 구현
- Graviton은 2018년 1세대 제품이 출시되었으며 2023년 11월에 4세대 제품을 공개, 5년 동안 4개 제품을 개발
  · AWS의 서버 CPU중 Gravition 비중은 2022년 3%이나 2024년에는 40%로 확대될 전망
- 2018년에 추론용 칩 Inferentia, 2020년에 머신러닝에 특화된 Trainium을 개발했으며 2023년초 Inferentia2, 11월에 Trainium2 공개

마이크로소프트는 세계 2위의 클라우드로 2019년부터 AI반도체를 개발했으며 2023년 11월에 처음으로 자체 개발한 AI반도체를 공개

- 2019년부터 '아테나' 프로젝트를 통해 자체 AI칩 개발을 추진했으며 2023년 11월에 GPU를 대체할 가속기 Maia 100, 클라우드 Azure용 CPU인 Cobalt 100를 발표

- 자체 반도체 개발을 위해 AI기업 OpenAI와 설계단계부터 제품 테스트까지 협력했으며 Maia는 외부 판매 계획이 없으나 Cobalt는 향후 타사에도 판매할 수 있다고 밝힘
- 현재 2세대 Maia, Cobalt 칩 설계에 착수

구글은 세계 3위의 클라우드로 2016년부터 AI반도체를 발표, 5세대 제품까지 개발

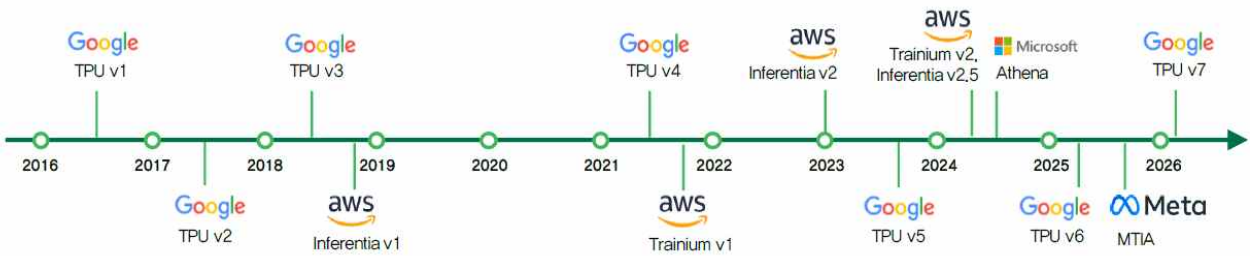
- 2016년 발표한 구글의 자체 AI반도체 TPU(Tensor Processing Unit)[12]는 구글의 AI 엔진 TensorFlow에 최적화되어 있으며 5세대 제품 출시
- 2024년 하반기에 자체 개발한 ARM 아키텍처 기반의 서버용 CPU Axion을 출시할 계획
- 구글은 자사 스마트폰 픽셀에 2021년부터 자체 설계한 AP 텐서를 탑재하고 있으며, 현재 3세대 제품을 사용중
- 구글의 AI 연구기업 DeepMind는 AI를 이용해 반도체 설계를 지원할 계획

Meta는 자체 개발한 AI반도체 MTIA(Meta Training & Inference Accelerator)를 자사 데이터센터에 탑재할 계획

- 2023년 5월, 1세대 MTIA를 공개했으며 2024년 4월 2세대 칩 사양을 공개

- 1세대 제품은 페이스북, 인스타그램 등의 콘텐츠 추천 서비스 등에 사용되나 Meta는 자체 개발 AI반도체를 궁극적으로 생성형 AI 학습으로 확대하는 것이 목표

주요 클라우드의 AI반도체 개발 계획



자료: 트렌드포스, DB금융투자.

12) 구글에서 설계한 NPU

3) 스타트업

Cerebras Systems는 세계 최대 크기의 AI반도체[13]를 통해 거대 AI 모델 실행을 지원하며 반도체를 별도로 판매하지 않고 시스템으로 판매

- 2015년에 설립된 미국 AI반도체 스타트업으로 세계 최초로 모놀리식(Monolithic) Wafer-Scale Engine(WSE)를 개발
  - 모놀리식 설계는 웨이퍼를 잘라 칩을 만들지 않고 웨이퍼 전체를 단일 칩으로 제작
  - 하나의 칩 안에 모든 구성 요소가 통합되어 칩 내부의 통신거리가 짧아 신호 전달 속도가 빠르고 저전력이 장점이나 높은 제조 난이도, 낮은 수율 등은 단점
  - 2021년 2세대 WSE-2, 2024년 3월 3세대 WSE-3(5나노) 공개했으며 WSE-3은 이론상 엔비디아 H100 62개에 해당하는 성능을 제공
  - WSE는 CS(Cluster Scale)라는 데이터센터 기기의 일부로 판매되며, CS 시스템에는 냉각 장비, 네트워킹 하드웨어 등이 탑재됨
- Cerebras는 UAE의 AI·클라우드 기업 G42, 미국 아르곤국립연구소, 제약회사 GSK 등을 고객으로 확보, 동사 매출은 매년 2배 성장중이며 손익분기점에 도달
  - 2023년 7월, Cerebras는 G42에 64개의 CS-2를 판매, AI 슈퍼컴퓨터 3대중 첫 번째를 공급하는 1억 달러 규모의 계약을 체결

삼바노바(Sambanova Systems)는 반도체를 외부에 판매하지 않고 AI반도체 Farm을 구축하고 이를 클라우드 형태로 기업에 임대하는 사업모델을 보유

- 2017년에 스탠퍼드대 교수 2명이 창업한 미국 AI반도체 스타트업으로 FPGA처럼 재구성이 가능한 NPU를 개발
  - 소프트웨어 개발로 사업을 시작했으나 회사가 구상한 소프트웨어를 운용할 수 있는 반도체가 없어 직접 AI반도체를 개발
  - AI 학습에서 CPU, GPU가 나누거나 순차적으로 작업하는 것을 단일 시스템으로 처리하여 효율성을 제고
  - 2023년에는 기업용 AI 모델 구축 및 서비스 플랫폼 'SambaNova Suite', 2024년에는 1조개 이상의 매개변수를 갖춘 대형언어모델(LLM) Samba-1*을 출시

    * 단일 모델이 아니라 전문가 구성(Composition of Experts) 방식으로 56개의 AI 모델을 조합

13) A4 크기

Groq는 AI 추론용 반도체 LPU(Language Processing Unit)을 출시

- 2016년 구글 출신 엔지니어들이 설립한 미국 AI반도체 스타트업으로 대형언어모델(LLM)의 실행속도를 높힌 LPU(Language Processing Unit)을 설계

- GPU 대비 빠르고 저렴하게 답변을 생성하면서 주목받고 있는 회사로 Chat-GPT가 Groq 칩 사용시 13배 이상 빠르게 실행될 수 있다는 벤치마크 테스트 결과를 발표

Graphcore는 영국을 대표하는 AI반도체 스타트업으로 한때 엔비디아의 대항마로 부상했으나 대형 고객사 이탈, 미국의 중국 규제 강화 등으로 생존 경쟁에 직면

- 2016년에 엔비디아 출신 개발자가 설립한 기업으로 IPU(Intelligent Processing Unit, 지능형처리장치)를 설계하며 마이크로소프트, 삼성전자 등의 투자를 받음

- IPU는 벡터 연산을 지원하는 GPU와 달리 그래프 연산[14]을 지원하며 칩 안에 데이터를 저장할 수 있는 메모리를 넣어 연산 지연을 해소하면서 유니콘 기업으로 부상

- 주요 구매자는 마이크로소프트, Cirrascale, NHN클라우드 등

- 동사는 마이크로소프트의 구매 중단, 미국의 대중국 반도체 수출 규제('23), AI반도체 기술 경쟁 심화 등으로 매출이 급감하면서 매각 추진

- 마이크로소프트는 자체 AI반도체 개발을 추진하며 Graphcore와 거래를 중단했으며 Graphcore의 경쟁력은 다수 AI반도체 스타트업의 등장, 제품 출시 지연 등으로 약화됨

- 2022년 매출은 전년 대비 46% 감소한 2.7백만달러를 기록했으며 중국이 주력시장(매출비중 20~25%)이었으나 2023년에는 미국의 규제로 중국에서 철수

- OpenAI, ARM, 소프트뱅크 등이 동사 인수를 검토중

캠브리콘은 중국을 대표하는 AI반도체 스타트업으로 주목받았으나 지속적인 R&D 투자에도 불구하고 고객사 확보의 어려움 등을 겪으면서 투자자 이탈 발생

- 2016년에 설립된 기업으로 화웨이를 기반으로 급성장했으나 화웨이의 이탈, 지속적인 R&D 투자 등에도 불구하고 고객 확보의 어려움 등으로 영업손실 지속

- 2020년 상해 증시에 상장하면서 '중국의 엔비디아'로 불렸으며, 상장 이후에도 공격적인 R&D 투자를 지속했으나 2017년부터 2023년까지 영업손실 지속

14) 벡터 연산은 3D 그래픽스 등을 지원, 그래프 연산은 소셜네트워크에서 인물과의 관계망을 만들거나 화합물, 단백질 구조 등 다양한 종류의 데이터 처리를 지원(KIPOST)

4. 향후 전망

엔비디아는 후발주자들의 거센 추격에도 불구하고 기술, 생태계, 재무적 우위 등으로 상당 기간 AI반도체 시장을 주도할 전망

- 엔비디아는 AI반도체 시장 지배력을 유지하기 위해 제품 출시 주기 단축, AI 관련 종합 솔루션 제공, 가격인하, 고객 맞춤형 AI반도체를 설계 지원 등을 추진

- 엔비디아 경쟁사의 제품 출시 주기는 2년내외, 엔비디아의 제품 출시 주기는 1년으로 경쟁사가 고성능 칩을 출시해도 엔비디아의 신제품에 성능을 역전당할 가능성이 큼
  - 엔비디아는 다수 AI 스타트업, 클라우드와 협력해 AI모델의 발전방향을 예상하고 반도체를 설계
  - 엔비디아는 연간 80억 달러를 R&D에 투자, Blackwell GPU 개발에 약 100억 달러를 투자했으나 다수의 AI반도체 스타트업 등은 자금 제약 등이 R&D 투자를 제약할 수 있음
- 엔비디아는 AI 관련 종합 솔루션을 제공하고 있으나 AI반도체 스타트업은 AI 종합 솔루션을 제공하기에는 자원 또는 역량이 부족
  - 엔비디아의 CUDA에 대응하기 위해 구글, 인텔, 퀄컴 등은 기술 컨소시엄 UXL(Unified Acceleration Foundation)을 구성('23)했으며 현재 개발 단계
  - 다수의 AI반도체 스타트업은 상대적으로 진입장벽이 낮은 NPU 사업을 약 10년간 영위해 왔으나 소프트웨어 개발 역량 등에서 엔비디아 대비 열위
- 엔비디아는 AI반도체 시장지배력을 유지하기 위해 신제품 가격을 전작 수준으로 책정
  - 엔비디아의 신모델 Blackwell은 성능 향상에도 불구하고 전작 수준의 가격을 책정하여 AMD 등은 가격전략 재검토 필요

- NPU는 GPU 대비 개발 역사가 짧아 GPU 수준의 완성도를 갖추지는 못했지만 추론 시장의 성장, GPU 대비 낮은 가격, 저전력 등의 장점으로 시장이 확대될 전망

- 2023년 엔비디아의 데이터센터 사업 매출의 40% 이상이 추론에서 발생했으며 생성형 AI 서비스 확산으로 AI반도체 성장의 축이 학습용 칩에서 추론용 칩으로 이동할 전망
- NPU의 가격은 엔비디아 GPU 가격 대비 ~1/10 수준으로 가격경쟁력을 보유

Ⅲ. 주요국의 육성정책

(미국) 민간기업이 AI반도체 개발을 주도하며 국방부는 차세대 반도체 리더십 확보를 위한 장기적인 기술과제 해결, 상무부는 국내 반도체 제조시설 구축 등을 지원

- 국방부 산하 방위고등연구계획국(DARPA)은 산·학·연 중심의 중장기 프로젝트를 통해 기초·원천기술 개발에 주력

- 2008년부터 차세대 AI반도체인 뉴로모픽 반도체[15] 개발을 위해 Systems of Neuromorphic Artificial Intelligence Components(SyNAPSE) 프로젝트를 추진, 2014년 TrueNorth를 개발

  * IBM은 2017년부터 TrueNorth를 상용화했으나 아직 초기 단계 기술

- 인공지능 NEXT 캠페인('19)을 통해 AI와 이종 칩의 적층·통합, 뉴로모픽 칩 등 정부 주도의 차세대 R&D 및 기업의 장기 투자를 지원

- 2018년 산학연 차세대 반도체 R&D 프로그램인 Electronics Resurgence Initiative (ERI)를 발표하고 AI 하드웨어 등 총 6개 분야를 지원

  · 인텔, 퀄컴, IBM, 반도체장비기업 Applied Materials, EDA(Electronic Design Automation) 기업 시놉시스, 대학 등이 참여하며 2022년까지 약 15억 달러를 지원

  · ERI 2.0은 2022년 2월 출범한 ERI의 후속 프로그램으로 5년동안 30억 달러 지원 계획

ERI의 6대 중점 지원분야

자료: DARPA.

15) 인간의 뉴런구조를 하드웨어 신경세포로 모사한 반도체로 뉴로는 신경, 모픽은 형상을 의미

- 2022년 8월, 미국은 반도체법(Chips & Science Act)를 제정하고 미국내 첨단 반도체 제조기반 구축, 산업생태계 조성 등을 지원

- 반도체법은 반도체산업에 대한 재정지원 527억 달러, 투자세액공제 25% 등을 규정하며 재정지원은 미국내 제조시설 구축 390억 달러, R&D 110억 달러 등을 포함
- AI반도체 등에 사용되는 첨단 공정 반도체 팹은 아시아에 집중되어 있어 미국 정부는 반도체 공급망의 지정학적 리스크 경감 등을 위해 2030년까지 미국 내에 최소 2개의 대형 첨단공정 파운드리 클러스터 보유 등을 추진
  · TSMC, 삼성전자, 인텔이 미국내 첨단공정 파운드리 팹을 건설중

(중국) 2030년 세계 1위 AI 국가로 도약을 추진했으나 미국의 제재 등으로 AI 연산을 위한 컴퓨팅 자원 확보가 어려워져 자국 AI반도체 육성 지원을 강화

- 2017년 '차세대 인공지능 발전계획'을 통해 2030년 세계 1위 AI 국가로 도약을 추진

- 2020년까지는 AI 기술·응용 수준을 선진국 수준으로 높이고, 2025년까지 일부 AI 기술·응용 분야에서 세계 최고 수준으로 도약, 2030년까지 AI 중심 국가로 도약 추진
  · '차세대 AI사업 발전 3개년 행동계획('18~'20)'을 통해 AI 소프트웨어·하드웨어(파운드리, NPU 등) 발전을 위한 기반 구축 등을 추진
- 2018년, 중국과학원 산하 자동화연구소는 인공지능혁신연구원을 설립하고 중국 반도체 기업 Spreadtrum(현 UniSOC) 등과 협력해 AI반도체 개발 등을 추진
  · 중국의 AI반도체 생태계는 정부 지원을 바탕으로 산학연관이 상호 협력하는 구조로 정부와 주요 인터넷기업(알리바바 등)이 AI반도체 스타트업에 적극 투자

- '14차 5개년 계획('21~'25)'에서는 인공지능, 반도체 등의 핵심기술 개발을 강조하며, 자국 AI반도체 산업 육성을 위해 제조 지원, 데이터센터 프로젝트 등을 추진

- 차세대 인공지능은 전용 칩 개발, AI 알고리즘 플랫폼 구축 등, 반도체는 설계도구(EDA), 소재, 화합물 반도체 등의 기술 개발을 추진
- 중국 정부는 AI반도체 설계와 제조 역량 확보를 위해 화웨이와 파운드리기업 SMIC에 보조금을 지급하는 것으로 추정[16]
  · 화웨이는 TSMC에서 반도체를 위탁생산했으나 TSMC는 미국의 제재가 강화되자[17] 2020년 5월 15일부터 화웨이의 신규 반도체 수주를 중단

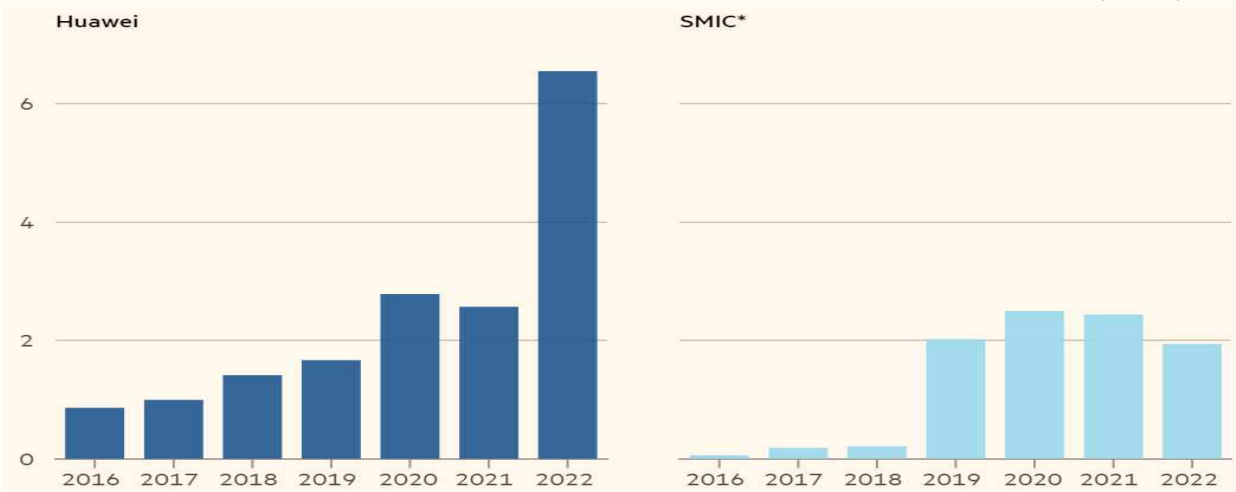
16) 미국 상무부는 우려거래자 명단(Entity List)에 화웨이('19)와 SMIC('20)를 등재
17) 2020년 5월, 미국 상무부는 자국 기술을 이용하는 해외 반도체기업이 화웨이에 제품 공급시 사전 승인을 요구

· 화웨이의 AI반도체 Ascend 910B는 SMIC의 N+2(7나노)에서 위탁생산하나 낮은 수율로 경제성 확보가 어려워 중국 정부가 보조금을 지급하는 것으로 추정

· 2022년 기준 화웨이는 65.5억 RMB(9.5억 달러), SMIC는 2020~2022년에 68.8억 RMB의 보조금을 수령했으며 SMIC는 중국 정부의 지원하에 5나노 칩 제조 라인 구축을 추진

중국 정부의 화웨이와 SMIC 보조금 지원 규모

단위: 10억 RMB



자료: Financial Times.

- 중국은 미국의 제재로 고성능 AI반도체 구매가 제약을 받자 중국산 AI반도체 사용시 보조금 지급, AI반도체 수요 창출을 위한 데이터센터 프로젝트 등을 추진

  · 미국은 2022년 10월 고성능 AI반도체의 대중국 수출을 규제(엔비디아의 A100·H100), 2023년 10월에는 중국 판매용 AI반도체(엔비디아의 A800·H800 등)의 수출 금지

  · 중앙정부는 '컴퓨팅 인프라 고품질 발전을 위한 행동계획('23)'을 통해 2025년까지 컴퓨팅 인프라 고도화 추진, 지방정부는 이에 발맞추어 데이터센터 건설 등을 추진

  · 중앙정부의 정책에 발맞추어 북경시 등 일부 지방정부는 AI반도체 개발 지원정책을 발표, 중국산 AI반도체를 사용하는 기업에 보조금 지급 등 추가 대책을 모색중

- 중국은 보안 강화 등을 이유로 정부기관이나 공기업에서 미국기업의 반도체가 탑재된 컴퓨터와 서버 사용 금지, 사용 기기는 2027년까지 중국 제품으로 교체 지시

  · 2023년 12월, 중국 공업정보화부는 안전하고 신뢰할 수 있는 프로세서를 갖춘 제품만을 구매하도록 지침을 수립했으며 해당 제품은 화웨이 등 중국기업 제품

  · 2024년 3월 양회에서 기술대표단은 소버린 AI(Sovereign AI)[18]를 확보하기 위해 중국은 하드웨어, 데이터 인프라, 인재개발의 3가지 병목 현상 해소가 필요하다고 권고

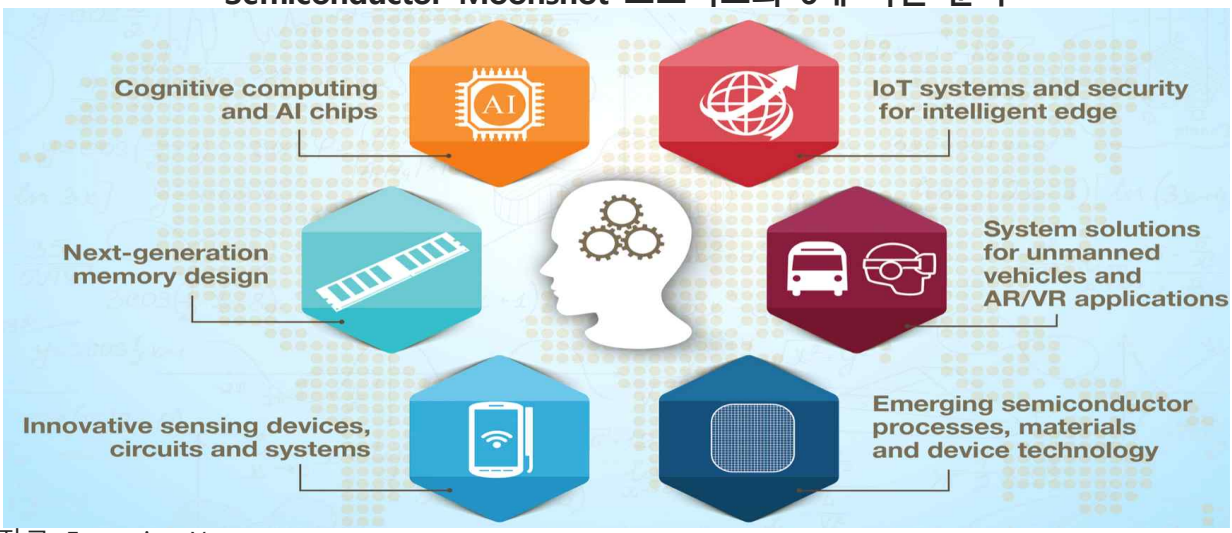
18) 국가의 디지털 주권을 보장하기 위해 자국 인프라, 데이터 등을 사용해 AI를 구축하는 국가의 역량

(대만) AI반도체 기술력 제고를 위해 핵심기술 개발 지원, 산학연 협력 플랫폼 조성, 글로벌 선도기업과 협력 강화 등을 지원

- 2018년, 대만 과학기술부는 'Semiconductor Moonshot Project'를 통해 엣지 컴퓨팅용 AI반도체 핵심기술 개발을 지원

- 4개년('18~'21) 동안 총 1.3억 달러를 AI반도체, 차세대 메모리반도체, IoT 시스템 및 보안, 자율주행차·AR/VR 적용 소자, 차세대 반도체 공정·소재 개발 등에 지원

Semiconductor Moonshot 프로젝트의 6대 지원 분야



자료: Executive Yuan.

- 2019년, 산학연 협력 플랫폼 'AI반도체 연맹(AI on Chip Taiwan Alliance, AITA)'을 설립

- AI반도체 연맹의 설립목표는 AI반도체 R&D 비용 절감, 개발시간 단축, 성능 향상, 핵심 IP(Intellectual property) 개발, AI 생태계 구축

- 회원사는 파운드리 기업 TSMC와 UMC, 팹리스 미디어텍 등 100여개 기업 및 단체

- 대만은 글로벌기업의 R&D 센터 유치를 통해 국내기업과 협력을 강화하고 대만을 첨단기술 R&D 센터로 조성 추진

- 대만은 글로벌 R&D 혁신 파트너 프로그램 등을 통해 미국 EDA 기업 Synopsys, 엔비디아, 마이크로소프트의 R&D 센터 등을 유치

  · 대만 공업기술연구원(ITRI)은 Synopsys와 대만에 AI반도체 설계 연구소를 설립('20)

  · 마이크로소프트는 아시아 최초의 AI R&D센터를 대만에 설립

  · 엔비디아는 2023년 12월에 대만에 AI R&D센터를 설립했으며 68억 대만 달러(약 3천억원)의 보조금을 지원받음

(일본) '반도체·디지털 산업전략 개정안('23)'[19]을 통해 글로벌 협력을 통한 첨단 반도체 제조기반 구축, 차세대 반도체 기술기반 확립, 미래 기술기반 확보를 추진

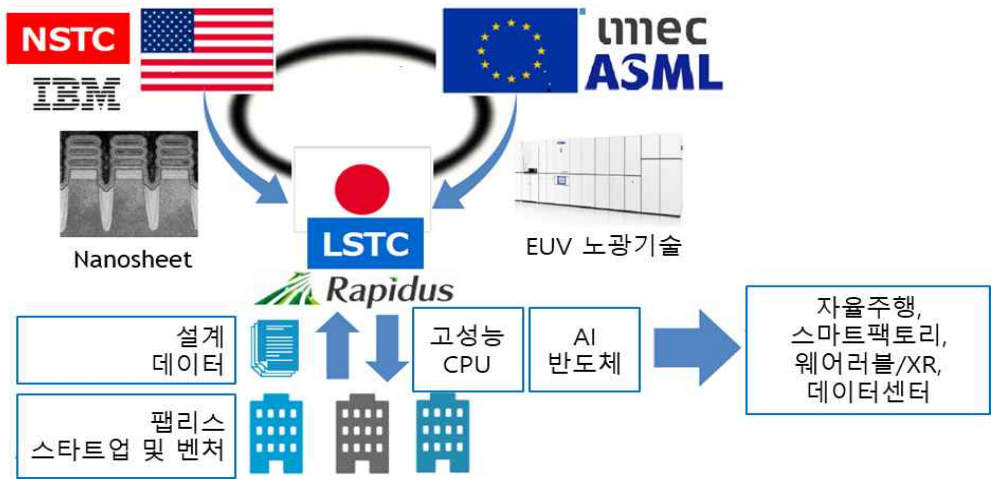
- (1단계) 첨단반도체 제조기반 구축 등을 위해 글로벌 반도체기업의 일본 투자를 지원

- 경제산업성 산하 신에너지·산업기술개발기구(New Energy and Industrial Technology Development Organization, NEDO)에 반도체기금('21)을 조성하고 보조급을 지급
  · 보조금은 일본에서 10년 이상 반도체를 생산하는 조건으로 지급되며, 보조금 규모는 설비 투자의 최대 1/3
- 일본의 반도체 제조기술은 40나노 수준에 머물러 일본의 역량만으로는 선도기업 추격이 어려워 TSMC의 일본 투자를 유치(1단계 28나노, 2단계 12나노)

- (2단계) 2나노 이하 차세대 반도체 기술을 확보하기 위해 미국 등과 협력

- 2022년 12월, 첨단반도체기술센터(Leading-edge Semiconductor Technology Center, LSTC)를 설립하고 첨단반도체 설계, 제조장비, 소재 등의 연구개발을 추진
  · 첨단반도체기술센터는 엣지(로봇, 자동차 등) 추론용 AI반도체 기술개발 등을 추진하며, 캐나다 AI반도체 스타트업 Tenstorrent와 2나노 AI반도체 설계를 공동 연구
- 민관합동기업 라피더스*를 설립하고 첨단반도체기술센터(LSTC), 미국 국립반도체기술센터(NSTC), IBM, 벨기에 반도체연구소 IMEC 등과 협력해 2027년 2나노 양산 추진
  * 키옥시아, 도요타, NTT, 소니, NEC, 소프트뱅크, 덴소, 미쓰비시UFJ은행의 출자금과 정부 지원금으로 설립되어 IBM이 2나노 공정 기술 제공

일본의 차세대 반도체 기술 확보 전략



자료: 경제산업성.

- (3단계) 글로벌 협력을 통한 미래기술개발, 차세대 소재 개발 등을 추진

19) 경제산업성은 2021년에 '반도체·디지털 산업전략', 2023년에는 동 전략의 개정안을 발표

Ⅳ. 한국의 AI반도체 산업 현황

1. AI반도체 기업 현황 및 경쟁력

국내 AI반도체 기업은 10여개로 모바일, 가전 등 온디바이스 부문에서 일부 제품을 상용화했으며 데이터센터 부문은 Reference를 구축하고 사업 본격화 단계

데이터센터용 AI반도체는 퓨리오사AI, 리벨리온, 사피온 등이 NPU를 개발해 클라우드에 시범 적용했으며 매출은 2024~2025년부터 본격화될 전망

- (퓨리오사AI) 2017년에 설립되었으며 2023년에 1세대 칩 워보이 양산, 2024년 9월에 2세대 칩 레니게이드 정식 출시 추진

- 워보이는 Vision 연산에 특화된 NPU로 타겟 시장은 엔트리급 데이터센터와 엔터프라이즈 서버이며 2024년부터 대만 PC·서버 제조사 Asus와 협력
  · 워보이는 14나노에서 생산되며 데이터 처리 속도는 64TOPS, 가격은 2백만원대
  · Asus는 서버의 세부 옵션으로 엔비디아 등과 더불어 퓨리오사 칩을 포함
- 레니게이드는 하이퍼스케일 데이터센터가 타겟이며 국내 AI반도체(NPU) 최초로 HBM3를 탑재

- (사피온) 2021년 SK텔레콤에서 분사했으며[20] 2020년 1세대 칩 X220, 2023년말 2세대 칩 X330을 출시했으며 2026년에 차세대 칩 X430을 출시할 계획

- X220는 추론 전용 모델로 SK텔레콤, NHN클라우드 등에 공급되었으며 매출의 50%는 계열사에서 발생
- X330은 데이터센터용 반도체이며 주요 고객사의 시제품 테스트와 신뢰성 검증 작업이 종료되면 2024년 상반기부터 양산을 시작할 계획
  · X330은 7나노에서 생산되며, 가격은 엔비디아의 가성비 GPU L40('22년 출시) 가격의 70~80% 수준으로 책정할 예정
  · 일본 통신사 NTT도코모의 자회사 도코모이노베이션과 X330 기술검증 수행 계약 체결
  · X330 매출은 Reference를 확보한 2025년부터 발생할 전망
- 차세대 칩 X430은 사피온 시리즈중 처음으로 HBM 탑재 예정

20) SK텔레콤, SK하이닉스, SK스퀘어가 투자

- (리벨리온) 2020년에 설립되었으며 2021년 금융 트레이딩에 특화된 칩 Ion, 2023년 데이터센터를 타겟으로 아톰 출시, 차세대 칩은 2025년 상반기에 출시 예정

- 아톰은 리벨리온의 투자사 KT클라우드에 일부 적용되었으며 IBM은 성능 평가 진행 중으로 2024년 상반기부터 본격 양산 추진
- 차세대 칩인 Rebel은 1000억개 이상의 매개변수를 가진 생성형 AI 모델까지 지원하며 삼성전자와 공동 개발중으로 HBM3E 탑재 예정
  · 삼성전자가 설계, 생산, 검증까지 모든 개발 과정에 참여하며 4나노 공정 이용

엔비디아와 국산 비교

회사		엔비디아	사피온	퓨리오사AI		리벨리온
제품명		A100(암페어)	X220-엔터프라이즈	레니게이드(추정치)	워보이	아톰
생산공정		TSMC 7nm	TSMC 28nm	TSMC 5nm	삼성전자 14nm	삼성전자 5nm
연산	정수	전방위 지원 (FP32~INT4)	INT4~16	INT4~8	INT8	INT2~8
	부동소수점		미지원	FP8~16	미지원	FP8~16
트랜지스터 규모	정수	624 TOPS	174 TOPS	512 TOPS	64 TOPS	128 TOPS
	부동소수점	312 TFLOPS	미지원	256 TFLOPS	미지원	32 TFLOPS
전력소모(TDP)		300W	135W	데이터 없음	40~60W	30~75W
인터페이스	메모리	HBM2e	LPDDR4	HBM3	LPDDR4	GDDR6
	CPU 연결	PCIe 4세대(16배속)	PCIe 3세대(16배속)	PCIe 5세대(16배속)	PCIe 4세대(8배속)	PCIe 5세대(16배속)
MLPerf (낮을수록 우수)	영상모델(ResNet50)	0.48ms	미제출	데이터 없음	0.71ms	0.24ms
	언어모델(Bert-Large)	1.57ms	미지원	데이터 없음	미지원	4.30ms

주: 1) TOPS는 Tera Operations Per Second로 1초에 1조번 정수 연산, TFOPS는 Tera Floating-Point Operations Per Second는 1초에 1조번 실수 연산
2) 부동소수점(Floating Point, FP)은 컴퓨터에서 실수를 표시하는 방법으로 FP32, FP16 등이 사용되며 FP 다음의 숫자가 클수록 정밀도는 높지만 속도는 느림
3) MLPerf는 MLCommons에서 측정하는 벤치마크 테스트
자료: 아주경제.

- (삼성전자) 대형언어모델(LLM)을 지원하는 추론용 반도체 마하1(MACH-1)을 2024년말~2025년초에 양산 예정

- 2022년 12월, 삼성전자와 네이버는 AI반도체 솔루션 개발 협력을 시작하고 대형언어모델(LLM)용 AI반도체를 공동 개발
- 마하1은 기술 검증을 완료했으며 네이버는 2024년에 마하1의 안정성 테스트를 진행할 계획
  · 물량은 15~20만개, 가격은 엔비디아 H100대비 약 1/10 수준인 500만원대로 예상
- 삼성전자는 범용인공지능(AGI) 반도체를 개발하기 위해 미국과 한국에 AGI 컴퓨팅랩을 설립하고 추론과 서비스 애플리케이션에 초점을 두고 LLM용 칩을 개발을 추진

- (한화뉴블라) 한화 계열사로 저전력 서버용 NPU 개발

엣지용 AI반도체는 DeepX, 텔레칩스, 삼성전자, LG전자 등이 참여

- (DeepX) 2018년 설립된 NPU 설계 기업으로 지능형 영상 분석 및 보안 시스템 시장을 타겟으로 2024년 하반기에 1세대 제춤 양산 추진
  - 2023년에는 현대자동차·기아 로보틱스랩과 로봇 플랫폼용 AI반도체 탑재를 위한 업무협약(MOU)을 체결
- (텔레칩스) 사피온의 IP를 기반으로 ADAS(첨단운전자보조시스템)용 칩을 만들어 완성차 업체에 공급 추진
  - ADAS용 칩은 상용 샘플 출시, 자율주행칩은 2024년 7월경에 샘플 출시 예정
- (LG전자) TV, 가전 등에 탑재되는 AI반도체를 개발하며 캐나다 AI반도체 스타트업 Tenstorrent와 TV, 자동차, 데이터센터용 반도체 개발 협력 계획
  - 스마트TV용 AI칩 알파11은 색감 보정, 보이스 ID를 통한 가족 구성원의 목소리를 구별해 개인별 취향을 반영한 개인 맞춤형 화질 등을 제공
  - 가전 전용 온디바이스 AI칩 DQ-C은 세탁기, 건조기, 에어컨 등에 탑재됨
- (삼성전자) 모바일, 자동차 등 다양한 엣지 분야에 적용되는 AI반도체를 개발
  - 스마트폰용 AP에 NPU 탑재, 자동차용 AP 엑시노스오토(ADAS 지원 등) 등
- (비전넥스트) 한화 계열 팹리스로 비전 CCTV 등에서 수집한 영상과 이미지를 분석하기 위해 Vision AI에 특화된 반도체를 2024년에 양산 예정

자료: 정보통신정책연구원.

한국의 AI반도체 기술수준은 선도국인 미국 대비 80.0%, 기술격차는 2.5년

- 한국의 AI반도체 기술수준은 미국의 기술수준을 100로 볼 때 80.0로 중국 90.0, 유럽 85.0보다 낮으나 일본 70.0보다 높은 수준

- 선도국인 미국 대비 기술 격차는 중국이 1.0년, 유럽 2.0년이나 한국은 2.5년, 일본은 3.5년으로 추정
- AI반도체 특허 출원 건수를 살펴볼 때 한국의 특허 출원 건수는 세계 3위이나 질적 수준은 미흡

  * AI반도체 특허 건수('11~'20): 중국 6,067건, 미국 2,625건, 한국 534건, 일본 308건 (KAIST)

  * AI반도체 핵심기술[21] 보유 비중('11~'20): 싱가포르 16.7%, 스위스 12.5%, 미국 12.0%, 한국 3.7%, 중국 1.1%
- AI반도체가 발전하기 위해서는 상생 관계에 있는 인공지능 기술이 경쟁력을 갖추어야 하나 우리나라의 인공지능 기술 수준은 주요국 대비 뒤쳐진 상황

  · 인공지능 기술수준은 선도국인 미국을 100으로 볼 때 중국은 90.0, 유럽 87.5, 한국 78.8

주요국 AI반도체 기술수준

	한국	미국	일본	중국	유럽
기술수준	80.0	100	70.0	90.0	85.0
기술격차(년)	2.5	0.0	3.5	1.0	2.0

자료: 과학기술정보통신부, 2022년도 기술수준평가 결과, 2024.2

2. 정부의 육성정책

정부는 '인공지능 반도체 산업 발전전략('20)'을 수립하고 2030년 글로벌 시장점유율 20% 달성을 추진

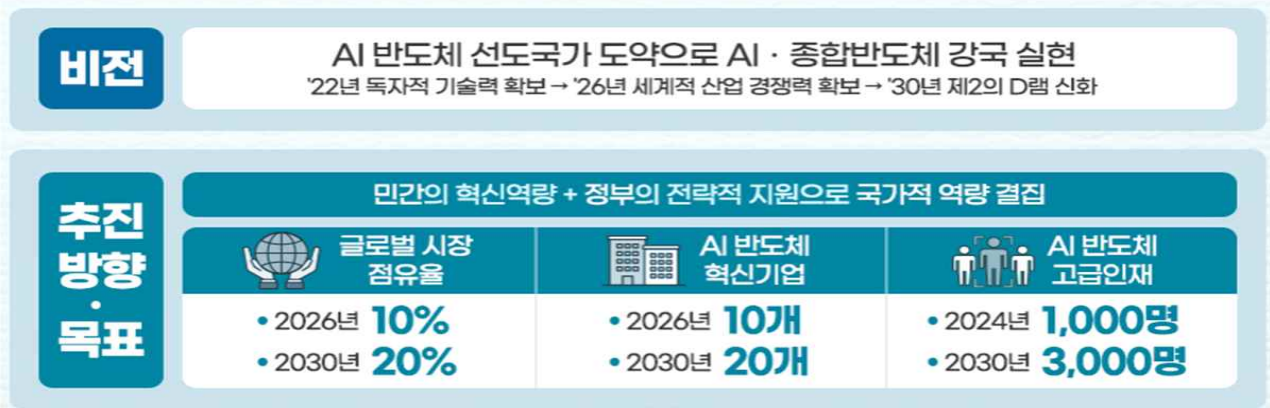
- 인공지능 반도체 산업 발전전략은 First Mover형 혁신 기술·인재 확보와 혁신성장형 산업 생태계 조성을 지원

- '차세대 지능형 반도체 기술개발 사업('20~'29, 약 1조원)'을 활용해 NPU 등의 개발을 지원, PIM(Process in Memory, 메모리+프로세서 통합) 핵심기술개발사업('22~'28, 총 4,000억원) 등을 통해 PIM 반도체 개발 착수

21) 특허기술의 인용수가 전세계 상위 10%에 해당되는 특허 보유 비중

- 대규모 공공 컴퓨팅 인프라에 AI반도체 도입 및 민관 협력을 통해 국산 AI반도체의 민간 AI 데이터센터 도입·검증 지원
- 영상·음성·언어 등 다양한 AI 응용분야 지원을 위한 공통 라이브러리, 컴파일러 등 시스템 소프트웨어 개발 지원

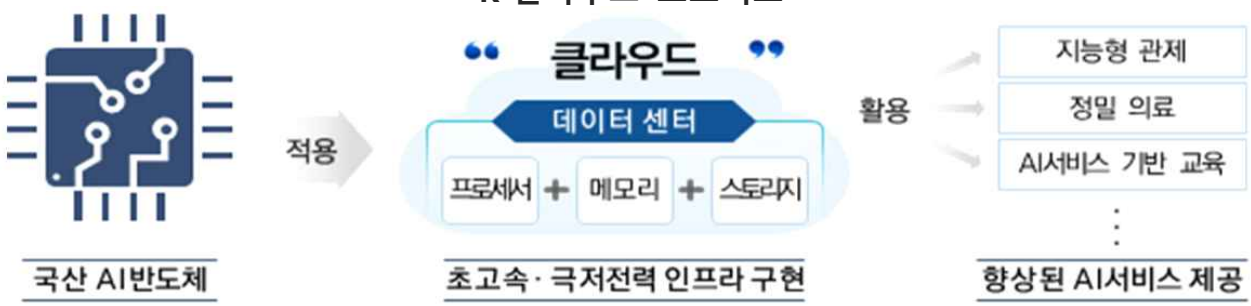
인공지능 반도체 산업 발전전략



자료: 과학기술정보통신부, 산업통상자원부.

- 국산 AI반도체를 활용한 K-클라우드 추진방안('22)을 통해 2030년까지 국내 데이터센터 시장의 국산 AI반도체 점유율 목표를 80%로 수립
- 2023년부터 국산 AI반도체를 단계별로 데이터센터에 적용해 국산 AI반도체의 국내 시장 창출, Reference 확보를 통한 글로벌 진출을 지원
  · 국산 AI반도체는 상용화 초기 단계로 시장 진입을 위해 종합성능뿐 아니라 기존 시스템과의 호환성, 안정성 등에 대한 검증이 필요

K-클라우드 프로젝트



자료: 과학기술정보통신부.

- 2024년 4월, 정부는 AI G3 도약을 위해 'AI-반도체 이니셔티브'를 추진 계획을 발표

- (AI 모델) 범용인공지능(AGI), 소형거대언어모델(sLLM)*의 원천기술 확보 추진

  * 초거대 AI 모델의 크기를 10% 수준으로 축소해도 기존 성능을 유지

- (AI반도체) 차세대 AI반도체 Processing in Memory(PIM), 저전력 K-AP, 신소자 및 첨단 패키징 기술혁신을 추진

- (HW+SW→AI서비스) AI슈퍼컴퓨팅을 지향하는 K-클라우드2.0을 추진해 범부처 AI 서비스 확산, 온디바이스 AI를 위한 핵심기술 확보 및 유망시장 선점을 위한 플래그십 프로젝트를 추진

AI-반도체 이니셔티브

자료: 산업통상자원부, 과학기술정보통신부.

Ⅴ. 결론 및 시사점

AI반도체는 국가안보와 경쟁력 제고 등을 위한 핵심기술로 한국의 주력산업 경쟁력 제고에 기여할 수 있어 정책적 지원이 필요

- 한국은 메모리반도체 강국이나 시스템반도체 경쟁력은 상대적으로 취약한 상황이며 AI반도체는 성장 초기 단계로 한국이 시스템반도체 경쟁력을 제고할 기회
  - 우리나라의 시스템반도체 세계시장점유율('22)은 3.3%으로 지난 10여년간 3% 수준에서 정체
  - 시스템반도체 시장규모는 메모리반도체 시장 대비 2배 크고 변동성이 낮아 시스템반도체 경쟁력 제고시 한국 반도체산업의 변동성이 완화될 전망
- AI반도체 경쟁력 제고는 한국의 주력 수출산업인 휴대폰, 자동차, 조선, 가전 등을 Smart하게 만들어주어 경쟁력 제고에 기여할 수 있음
  - 한국의 주력 수출산업 대부분은 중국의 추격 등으로 입지가 좁아지고 있어 AI반도체 등을 통한 제품 차별화 필요
    - 2023년에는 애플이 스마트폰 출하량 1위로 도약했으나 2024년에는 삼성전자가 온디바이스 AI폰을 출시하면서 Game의 Rule을 바꾸어 양사간 치열한 경쟁이 예상됨
  - AI반도체는 스마트폰, 가전 등을 중심으로 탑재되고 있으나 자동차, 조선, 로봇 등 다양한 산업이 AI반도체 탑재를 확대할 전망
    - ADAS(첨단운전 보조시스템) 시장은 2020년 130억 달러에서 2030년 430억 달러[22], 자율주행 선박시장은 상용화 예정인 2025년에 180조원, 2030년 330조원으로 성장 전망

AI반도체 기업을 육성하기 위해 개발자금 지원, Reference 구축, 수요산업과 협력 강화, '팹리스-파운드리'의 유기적 협력관계 구축 등이 요구됨

- 국내 AI반도체 기업은 대부분 중소기업으로 개발비 부담 등을 경감시켜주기 위해 첨단 파운드리 공정 이용 등에 대한 정책적 지원이 필요
  - 반도체 설계에서 양산까지 약 2년이 소요되며 설계에서 양산까지 최소 수천억원의 비용 소요
    - 최신 공정 이용시 비용 부담이 커 우리기업이 엔비디아보다 앞선 공정 사용이 쉽지 않음

22) 맥킨지

- 주요 AI반도체 스타트업은 2024~2025년부터 매출이 본격화될 것으로 예상되지만 AI 모델의 발전 속도 등에 발맞추기 위해 차세대 제품 개발에 지속적인 투자 필요
  · 2023년 퓨리오사와 리벨리온[23]의 매출은 각각 36억원, 27억원, 영업손실은 각각 600억원, 159억원을 기록

• 주력 수출산업인 자동차, 조선 등과 연계 강화를 통해 수요기업이 요구하는 반도체 적기 개발 필요

- AI반도체를 개발해도 실제 서비스에 활용하지 못한다면 향후 칩 개발 방향을 설정하기 어려움
- 주력 수출산업이 변화하는 환경에 신속히 대응하지 못할 경우 경쟁력이 빠르게 하락할 가능성이 있음

• '팹리스-파운드리'간 유기적 협력관계 구축을 통해 AI반도체 기업의 성장이 국내 파운드의 성장을 견인하는 선순환 구조 구축 가능

한국은 AI반도체, 초거대 AI 플랫폼* 등을 기반으로 소버린 AI를 추진하나 국가안보, 지정학적 이슈 등에 민감하거나 기술력이 부족한 국가와 협력할 기회가 있을 전망

* 세계 3번째 초거대 AI 개발: 美 GPT3('20)→中 판구('21)→韓 하이퍼클로바('22)[24]

• 세계 주요국은 자국의 언어, 문화 등을 학습시켜 자국의 환경에 맞는 AI 모델을 자체 구축하는 '소버린 AI' 전략을 추진

- 소수의 Big Tech 기업이 AI 기술을 독점하고 AI 모델이 대부분 영어 데이터를 기반으로 학습하면서 AI모델에 서구의 가치관이 내재화되는 것 등에 대한 우려 증가
- 소버린 AI는 각국이 자체 인프라, 데이터 등을 활용해 독자적인 AI 역량을 구축하는 것으로, 특정 국가·기업의 영향력에서 벗어나 AI 주권을 지킬 수 있음

• 중동·유럽 등이 국가안보, 지정학적 이슈 등으로 AI반도체 공급처 다변화를 추진할 것으로 예상되어 한국기업의 해외진출에 대한 정부의 외교적 지원 등이 필요

- 사우디아라비아는 AI분야에서 중국과 협력해왔으나 미국이 중국과의 협력을 경계하여 중국외 국가와 협력을 모색할 전망
  · 킹압둘라과학기술대학은 중국 인력과 공동으로 AI모델을 개발했으며 엔비디아 H100을 3천개 이상 구매하여 미국은 중국이 사우디를 우회하여 GPU를 확보하는지 의구심을 갖음
  · 이에, 미국은 고성능 반도체 수출 통제 국가에 중동 일부 국가(사우디, UAE)를 포함

23) 2023년 5월부터 KT에 아톰 초도 물량(K-클라우드 프로젝트 등 정부 플젝)을 납품하면서 매출 발생 시작
24) 하이퍼클로버X는 한국판 AI 성능 평가 체계 'KMMLU(Measuring Massive Multitask Language Understanding in Korean)'에서 오픈AI의 GPT-3.5-Turbo, 구글의 제미나이 프로보다 높은 점수를 기록

- UAE는 '국가 인공지능 전략('17)'에 따라 2031년까지 AI 선도국으로 도약할 계획이며 아랍어 대형언어모델(LLM) '자이스' 출시, 기술투자회사 MGX 설립 등을 추진
- 이탈리아는 자국어 대형언어모델(LLM) 구축을 추진중이며, 인공지능 프로젝트를 촉진하기 위해 10억 유로 규모의 투자 펀드 조성 계획

주요국의 데이터센터용 AI반도체 기업

자료: Yole.

참고문헌

DB금융투자, 'CES2024: The Age of AI', 2024.1.16.

KOTRA, '일본의 新반도체 산업 전략과 글로벌 공급망 구축', 2023.5.31.

대외경제정책연구원, '한·일 반도체 전략 및 협력방안', 2023.9

테크월드, '日 라피더스 2나노 반도체 양산 실현되나...LSTC 본격 시동', 2024.2.13

테크월드, '대만 '칩스법' 2월 발효...TSMC 등 일부 기업만 혜택 지적', 2024.1.30.

Center for Security and Emerging Technology, 'AI Chips: What They Are and Why They Matter', 2020.4

Deloitte, '2024 global semiconductor industry outlook', 2024

Forbes, 'New MLPerf Benchmarks Show Why NVIDIA Reworked Its Product Roadmap', 2023.11.8.

SiliconANGLE, 'Ampere and other chipmakers form AI Platform Alliance', 2023.10.17.

The Wall Street Journal, 'How a Shifting AI Chip Market Will Shape Nvidia's Future', 2024.2.25.

Industrial Technology Research Institute, 'ITRI and Synopsys' AI Chip Design Lab Launches Soft Opening', 2020.10.21.

Financial Times, 'China offers AI computing 'vouchers' to its underpowered start-ups', 2024.3.4.

China Daily, 'China to step up efforts for AI chips', 2023.10.21.