

AI 반도체의 미래와 전환점

KISTEP 연구개발예산정책센터 정의진
기술예측센터 신동평·글로벌기술전략본부 손석호

AI 반도체의 미래와 전환점

2026.3.13. 연구개발예산정책센터 정의진 연구위원, 신동평 기술예측센터장, 손석호 글로벌기술전략본부장

요약문

- AI 확산과 함께 AI 반도체 시장도 가파른 성장이 예상되며 본 고에서는 AI 반도체 관련 현황을 살펴보고 향후 시장에 영향을 미칠 수 있는 5개의 전환점(turning point)을 제시하고자 함
- ① **(새로운 AI 모델의 등장)** 트랜스포머 기반의 범용 대형 언어모델(LLM)은 AI 시장 성장을 견인했으나 ‘포스트 트랜스포머’의 등장은 AI 반도체 시장의 판도를 근본적으로 재편할 최대 변수가 될 전망
- ② **(AI 서비스 시장의 확장)** Agentic AI 및 온디바이스 AI로 시장이 확대되면서 AI 인프라도 데이터센터, 온프레미스, 온디바이스 등으로 다변화되고 AI 인프라 시장 역시 영역별로 분화되어 성장할 것으로 전망
- ③ **(에너지 수급)** AI 서비스 확산으로 데이터센터 전력 수요가 급증하면서 향후 AI 인프라 경쟁력은 반도체 성능뿐만 아니라 안정적인 전력 공급과 저전력·고효율 기술 확보도 중요 쟁점
- ④ **(각국의 AI 풀스택 경쟁)** 국가 안보와 직결된 ‘소버린 AI’ 확보를 위해 하드웨어부터 모델까지 수직 계열화 하는 ‘AI 풀스택’ 경쟁이 가속화되고 있으며 글로벌 AI 생태계는 동맹·권역 중심으로 재편되는 양상
- ⑤ **(게임체인저 기술)** 단기적으로는 기존 하드웨어의 한계를 극복하며 성능을 극대화하는 기술이, 장기적으로는 오픈킴·뉴로모픽·양자 등 패러다임 전환 기술이나 ‘포스트 트랜스포머’ 등장에 따라 시장 판도가 근본적으로 재편 가능성
- 전 세계적 AI 인프라 구축 열풍에 따라 AI 반도체 수요는 폭발적으로 증가할 것으로 예상되며 우리나라는 메모리·제조업 강점을 기반으로 ‘AI 풀스택’ 자립 생태계를 구축하여 AI 반도체 시장 주도를 위한 전략 필요

1 작성 배경

- 챗GPT에서 시작된 AI 붐과 함께 AI 인프라 구축을 위한 데이터센터, AI 관련 기술 개발 등이 앞다투어 이루어지고 있으며 AI 반도체 시장도 가파르게 성장하고 있음
- 특히 대규모 AI 인프라 구축에 따른 반도체 수요가 폭증하면서, PC, 스마트폰 등 주요 전자기기 가격 상승 까지 이어지는, 이른바 ‘반도체 대란’ 현상이 나타나고 있음
- ※ 2026년 전 세계 반도체 매출이 전년 대비 25~30% 이상 성장하며 사상 처음 1조 달러(약 1,480조 원)를 돌파할 것으로 예상(WSTS, '25.12.)

□ 대화형 AI와 생성형 AI를 중심으로 AI 인프라는 일상과 산업 전반으로 빠르게 확산되고 있으며 관련 전문가들은 AGI(범용 인공지능) 시대도 가까운 시일 내 도래할 것으로 예상

※ 샘 올트먼(Open AI CEO) 10년 이내, 데미스 허사비스(Google DeepMind CEO) 2030년, 레이 커즈와일(미래학자) 2029년, 다리오 아모데이(Anthropic CEO) 2026~2027년 예상

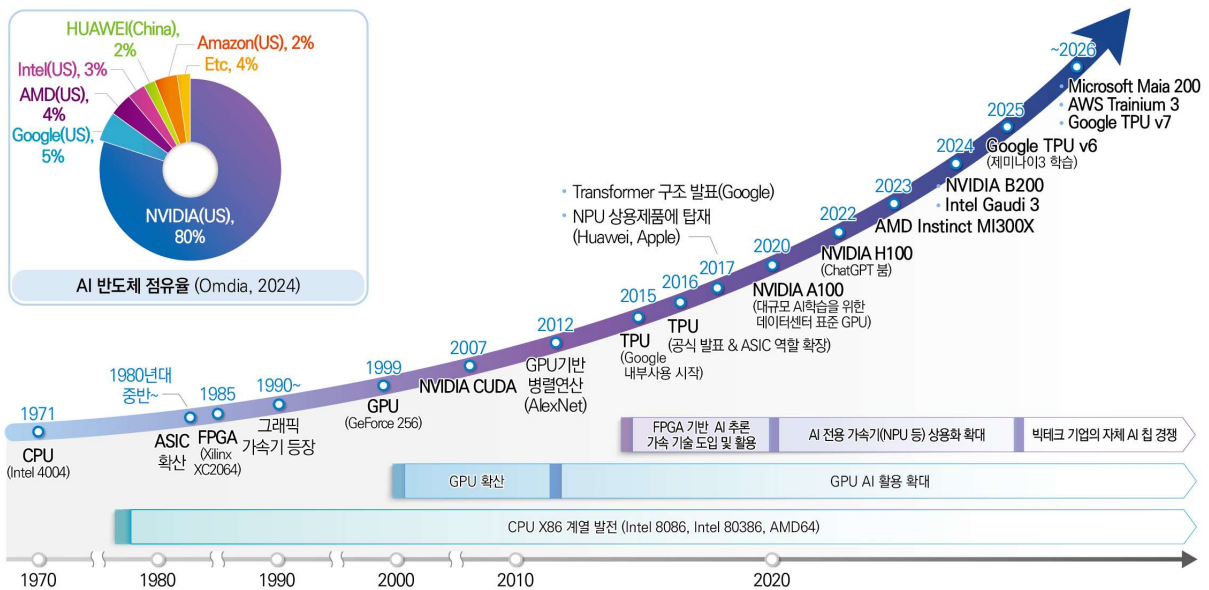
□ AI 반도체는 생성형 AI와 고성능 컴퓨팅에 대한 수요 증가로 급격한 성장을 하고 있지만 GPU의 높은 가격, 반도체 공급망 제약, 막대한 전력 소비 등의 이슈가 있으며 이 시점에서 현황과 전망을 짚어볼 필요가 있음

○ 2025년 구글의 TPU v6를 시작으로 엔비디아의 GPU 독주에서 벗어나려는 빅테크 기업의 경쟁, 자국 주도의 AI 생태계 기반을 조성하고자 하는 국가 간 AI 풀스택 경쟁도 격화되고 있음

□ 본 고에서는 ①문헌조사 및 생성형 AI를 통해 미래 전환점이 될 수 있는 후보 이슈를 발굴하고 ②전문가 FGI를 통해 향후 AI 반도체의 미래에 영향을 미칠 5개의 전환점(turning point)을 도출

2 AI 반도체의 발전 경로 및 현황

□ AI 반도체의 발전 경로



〈 AI 반도체의 발전 경로 〉

○ AI 개발 초기에는 CPU가 주요 처리장치 역할을 하며 응용 프로그램에 필요한 기본 알고리즘과 연산을 수행 했지만, 더 복잡한 작업의 처리에는 한계

※ 2012년 토론토 대학의 제프리 힌트 교수 연구팀이 '이미지넷(ImageNet)' 시물 인식 대회에서 당시 CPU만으로는 불가능한 계산량을 GPU를 사용하여 해결, 딥러닝 발전을 가속화시키는 결정적인 계기가 됨

○ 엔비디아의 GPU는 병렬컴퓨팅 작업으로 방대하고 복잡한 작업량이 처리 가능하며 GPU와 함께 사용할 수 있는 소프트웨어 생태계(CUDA)까지 구축한 덕분에 현재 AI 반도체의 80%를 점유(Omdia, '24)

※ 데이터센터 GPU 시장은 2024년 1,250억 달러 규모에 도달하였으며(loT Analytics, '25.1.), 엔비디아의 데이터센터 GPU는 4분기('25.11~'26.1.) 매출만 681억 달러에 달함('26.2.25.)

○ 지금까지는 엔비디아의 GPU가 AI 반도체 시장에서 압도적인 비중을 차지하고 있었지만, 앞으로는 AI 모델의 행방, 전력 공급, 비용 효율, 공급망 확보 등 다양한 요인이 AI 반도체 시장에 영향을 끼칠 것으로 예상

○ AI 가속기(AI accelerator)라는 개념은 2010년대 중반에 등장하였으며 AI 연산을 빠르고 효율적으로 수행하기 위해 만들어진 모든 전용 하드웨어 장치

- AI 반도체와 AI 가속기는 유사한 개념으로 쓰이고 있으며, 범용 가속기(GPU)와 목적 특화형 AI 가속기(GPU 외)로 나눌 수 있음

- GPU는 병렬 처리가 가능한 범용 가속기로 다양한 연산에 사용되는 반면, AI 가속기는 머신러닝·딥러닝 같은 AI 연산을 최적화된 특수목적 하드웨어로 특정 워크로드(workload)에서 높은 성능을 발휘

〈 AI 반도체 종류와 역할 〉

종류	역할	특징 및 한계	대표기업 예시	사용처
AI 가속기	GPU	학습 중심 + 추론 · 범용성이 높고 생태계가 강력 · 전력 소모가 크고 고비용	NVIDIA, AMD	데이터센터 AI 학습, 추론
	ASIC	특정 용도의 학습과 추론 · 워크로드 맞춤 설계, 성능/전력/비용 최적화 · 유연성 ↓, 개발비용 ↑	Amazon, Tesla, Groq	데이터센터, 자율주행
	TPU	학습 + 추론 · 구글 서비스 최적화, 대규모 처리 · 특정 플랫폼 종속	Google	Google Cloud의 AI 학습추론
	NPU	추론 중심 · 저전력·저지연, 효율성 중심 · 대규모 학습 불가	Apple, Qualcomm, Samsung	스마트폰, PC, 엣지 디바이스
	LPU	LLM 추론 · LLM의 초고속 실시간 답변 가능 · SRAM 사용으로 큰 모델에서는 고비용	Groq	데이터센터
	FPGA	추론 중심 · 극도의 저지연, 전력 효율성 · 대량 생산되는 ASIC에 비해 고비용	AMD, Altera	통신 기지국, 군사장비, 금융 등
통합 칩	AI SoC (System on chip) 추론 중심	· CPU, GPU, NPU를 하나의 칩으로 통합하여 AI 기능 내장화 · 학습 성능에 제한, 온디바이스 등에서 활용	Apple, Qualcomm, Intel	모바일, AI PC, 엣지 디바이스

※ 이 외에 AI 연산 성능 극대화를 위해 데이터 이동·보안·저장 등 흐름과 보안을 가속하는 보완적 칩인 DPU 등도 있음

〈 GPU와 기타 AI 반도체의 차이 〉

구분	GPU	기타 AI 반도체(GPU 제외)
특징	· 대규모 병렬 처리 구조를 기반으로 높은 연산 성능을 제공하며, 대규모 AI 모델 학습 작업에 강점 · 그래픽과 딥러닝·머신러닝 프로젝트를 병행하는 복잡하고 다양한 작업 환경에서는 GPU가 유연성이 높음	· 딥러닝·머신러닝을 포함한 병렬 연산을 효율적으로 수행하도록 설계된 특수 하드웨어 · 특정 AI 연산에 최적화되어 높은 전력 효율과 낮은 지연성을 바탕으로 실시간 처리에 유리

○ 2016년에 구글이 TPU를 발표하였으며 2022년 TPU를 적용한 구글 AI 모델 ‘제미나이3’가 오픈AI의 챗 GPT를 위협하는 성능을 보이자 ASIC(주문형 반도체)이 GPU의 대안으로 떠오름(’25.11.)

※ 구글 외에도 아마존, 메타 등 빅테크 기업들은 자체 서비스 맞춤형 ASIC을 제작하고 있음

□ AI 반도체 시장은 IT 및 통신, 의료, 자동차, 금융, 데이터센터 등 광범위한 산업 분야에서 가파른 성장이 예상되며 주요 기업 간의 치열한 경쟁과 혁신이 이루어질 것으로 보임

3 AI 반도체의 향후 전환점(turning point)

□ AI 반도체 시장은 2033년까지 6,000억 달러(약 800조 원) 규모로 성장할 것으로 예상되며(Bloomberg, ’26), 향후 AI 반도체 지형에 영향을 미칠 수 있는 5개의 터닝 포인트를 도출함

① **(새로운 AI 모델의 등장)** 현재 범용 대형 언어모델(LLM) 중심으로 발전하였으나 LLM의 기반이 되는 트랜스포머(Transformer) 구조를 뛰어넘는 새로운 모델의 등장 여부에 따라 AI 반도체의 판도는 완전히 바뀔 수 있음

※ 안 르쿤(전 Meta 수석과학자)은 “LLM은 막다른 길(Dead-end)”이라고 주장하며 LLM이 지능의 중심이 아닌 인터페이스나 여러 모듈 중 하나에 속하게 될 것으로 예측

○ 딥러닝은 1940년대부터 이어진 인공신경망 아이디어를 바탕으로 연구됐지만 빅데이터의 폭발적인 증가와 하드웨어 기술의 발전으로 2010년대 이후 전성기를 맞으며 초창기 알고리즘인 CNN¹⁾과 RNN²⁾이 크게 발전
- 2016년 알파고(구글 딥마인드)의 등장은 AI 기술의 발전 가능성을 전 세계에 각인시킴

○ 2017년 구글에서 트랜스포머 아키텍처를 제시하였는데 병렬화/학습 측면에서 장점을 가지고 문장의 맥락과 의미를 파악하고 새로운 문장을 생성하는데 뛰어난 능력을 보임

- 주로 자연어처리 분야에서 혁신을 일으켰지만, 컴퓨터 비전, 음성 인식 등 다양한 분야에서 활용되며 BERT, GPT와 같은 대규모 언어모델(LLM)과 생성형 AI로 더욱 발전

- 2020년 OpenAI가 발표한 GPT-3는 규모가 질적 변화를 가져올 수 있다는 것을 증명하였으며, 모델의 크기를 키우고(Scaling) 학습데이터를 늘리자, 문법 교정이나 번역과 같은 기본적인 언어 처리를 넘어서 글쓰기, 프로그래밍, 추론까지 가능해짐

○ 하지만, AI 관련 기술이 발전할수록 트랜스포머가 Agentic AI의 핵심 엔진으로는 한계가 있다는 주장이 나오고 있으며 트랜스포머 구조에서 벗어나려는 시도가 나타나고 있음

※ 비살 시카(전 SAP CTO) 연구팀은 AI의 고질적인 할루시네이션 현상이 데이터 부족 탓이 아니라 트랜스포머의 구조적 필연성에 기인한다고 지적하며 구조적 한계를 수학적으로 증명(arXiv, ’25)

※ 이스라엘 AI 스타트업 AI21 랩스의 아리 고센 CEO는 트랜스포머 구조 기반의 에이전트 시스템을 구축하기에 비용 효율성이 떨어진다고 지적(’24)

1) 합성곱 신경망(Convolutional Neural Network)

2) 순환 신경망(Recurrent Neural Network)

- 예를 들어, 지금은 학습과 추론이 구분되어 있지만 실시간 입력 데이터와 새로운 정보를 추론과 동시에 조금씩 “경험하면서 배우는” 실시간 학습형 모델(Continual Training & Inference)도 논의·연구되고 있음

○ Next 트랜스포머 시대가 언제 올지는 알 수 없으나 혁신적인 새로운 모델의 등장에 따라 AI 반도체 시장의 판도는 뒤바뀔 수 있음

2 (AI 서비스 시장의 확장) AI 기술의 발전으로 생성형 AI에서 Agentic AI, AGI(범용 인공지능) 시대가 예상보다 빠르게 도래하며 업무/생활용 온디바이스 AI까지 확산할 것으로 보임

데이터센터(성능 중심)



온디바이스(AI SoC)(효율 중심)



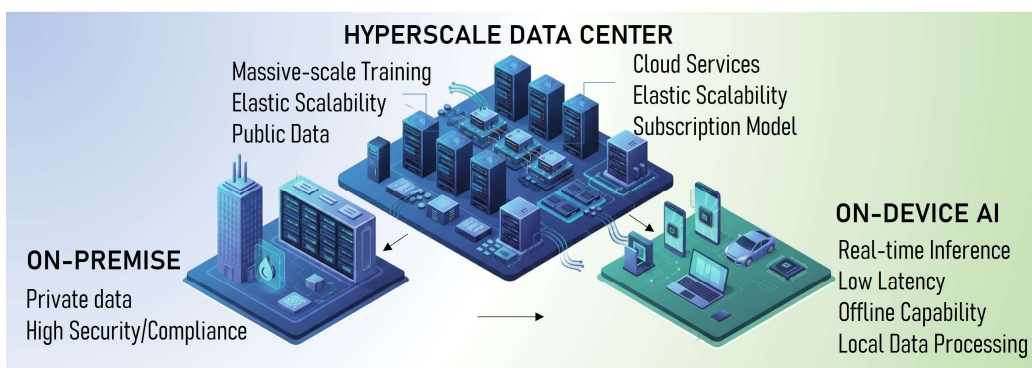
〈 AI 반도체 종류별 대표기업 예시 〉

○ AI에 대한 수익 구조가 불확실하여 ‘AI 거품론’도 한때 거론되었으나 이제는 똑똑한 ‘엔진 개발(학습)’은 충분하다는 판단과 함께 저렴하고 폭넓은 ‘서비스 운영(추론)’ 쪽으로 무게중심이 이동하고 있으며 인프라 역시 용도와 환경에 맞게 다변화되고 있음

○ AI 인프라 시장은 ①범용적인 데이터센터, ②보안·맞춤형 성능이 중요한 온프레미스(On-Premise), ③실시간성과 개인정보 보호가 강조되는 온디바이스(On-device) 영역 등 크게 3가지로 구분할 수 있음

〈 AI 반도체 시장 구분 〉

구분	하이퍼스케일 데이터센터/클라우드	온프레미스(On-Premise)	온디바이스(On-device)
용도	초거대 모델 학습 및 범용 추론	내부 데이터 기반 전문 AI	개인 서비스 및 실시간 반응
물리적 위치	외부(Public)	기업/기관 내부	스마트폰, PC, 차량, 로봇, 드론 등
응답 속도	낮음(네트워크 경유)	중간(내부망)	매우 높음(오프라인 가능)
특징	고성능 + 확장성, 유연성이 높음	운영 안정성과 데이터 보안 중심, 필요시 서버 증설	전력소모 최소화, 통합 칩에 모든 기능을 집약

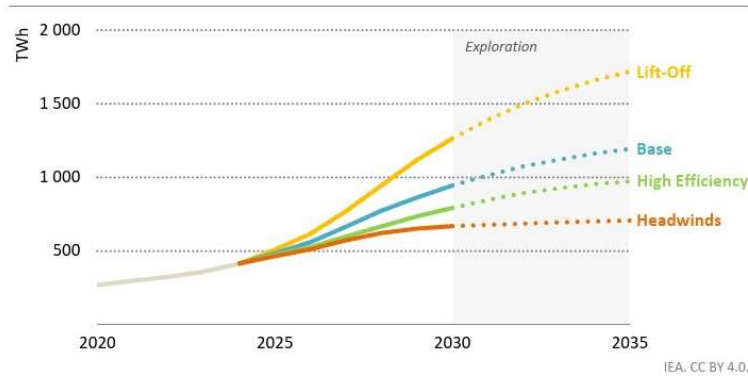


〈 AI 서비스 시장(Gemini 3.0 생성 그림) 〉

- (데이터센터) 축구장 3개 정도 되는 면적(2만 2,500m²)에 10만 대 이상의 서버를 보유한 초대형 규모로 구글, 아마존(AWS), 마이크로소프트와 같은 빅테크 기업이 초거대 모델을 학습시키고 구독형 AI 모델을 통해 전 세계 수억 명의 추론 요청을 동시에 처리
- (온프레미스) 기업/기관이 자체 시설 내에 서버, 스토리지 등 인프라를 직접 구축운영하는 방식이며 최근에는 데이터 병목 해소를 위해 현장에 분산 배치하는 엣지 서버/데이터센터 형태의 온프레미스 구축도 활발
 - 금융기관, 의료기관, 국방 분야처럼 데이터 보안이 중요한 곳이나 장기적으로 막대한 AI 연산을 24시간 수행하는 대기업의 경우, 서버를 직접 구축하는 것이 비용 측면에서 효율적
 - 과거에는 모든 데이터를 중앙 데이터센터로 보내 처리했지만, 데이터양이 폭증하면서 온프레미스는 중앙 데이터센터와 네트워크가 감당하지 못하는 병목현상을 해결하기 위한 목적도 매우 큼
- (온디바이스 AI) 클라우드를 거치지 않고 기기 자체에 탑재된 AI 반도체를 통해 실시간 정보 처리가 가능하며 데이터가 외부로 전송되지 않아 보안성이 강화됨
 - 스마트폰, 드론, 로봇, PC, 차량 등 다양한 기기에 온디바이스 AI가 적용될 수 있으며 2023년 166억 달러에서 2031년 1,181억 달러 규모로 연평균 27.9%씩 성장 예측(Deloitte Insights, '25)
- 초거대 모델의 대규모 학습, 보안, 실시간 응답성, 저전력 등 각각의 시장 수요에 따라 AI 반도체 시장이 분화되어 성장할 것으로 보임

③ (전력 수요의 급증) 서비스가 방대해질수록 AI 모델 학습 및 추론에 막대한 전력이 소비되기 때문에 AI 인프라 구축 정책이나 관련 기술 개발에 있어 원활한 전력 공급, 저전력·고효율 등도 고려해야 하는 중요한 요소

- IEA(국제에너지기구)에 따르면 전 세계 데이터센터 전력 소비량은 2024년 약 460 TWh에서 2035년 약 1,300~1,700 TWh로 증가 예정(2024년 기준, 국내 연간 전력 소비량은 549 TWh, 한국전력통계)
 - ※ 일반 구글 검색 1회(약 0.3 Wh)에 비해 챗GPT와 같은 생성형 AI의 답변 생성(약 2.9 Wh)은 약 10배의 전력 소모(A de Vries, '23)
- 가트너는 2027년까지 전 세계 AI 데이터센터의 약 40%가 전력 공급 부족으로 인해 운영에 제약 받을 것으로 예측('24)
 - 신규 데이터센터 건설보다 전력망 구축 등 인프라 확충에 시간이 소요되어 전력망에 연결하지 못해 서버가 중지되는 상황이 발생할 수도 있음



< 글로벌 데이터센터의 전력 예측치(2020~2035) (IEA, 2025) >

- 글로벌 빅테크 기업은 전력 확보를 위한 노력과 함께, 저전력·고효율 반도체 개발에 전력을 다하고 있음
 - ※ (전력 확보) (메타) 3개 원전업체와 6.6GW 전력 공급계약 체결('26.1.), (xAI) 일론 머스크의 xAI는 데이터 센터 인근에 태양광 발전소 건설 예정('25.11.), (아마존) SMR 기업 X-에너지에 5억 달러 투자('25.11.)
 - ※ (기술 개발) (엔비디아) GPU-DPU 패키지 통합을 통해 에너지 효율 최적화, (구글) TPU v7은 v6 대비 전력 효율성이 약 2배 향상, (퓨리오사) 레나게이드S의 전력 소모는 60W로 기존 대비 최대 3분의 1 수준으로 감소
- 전력 공급의 제약으로 소형 언어모델 등 AI 연산 방식 변화, 클라우드 집중형에서 개인 기기로 처리하는 온디바이스 AI로 전환, 혹은 저전력 AI 반도체 도입 등이 가속화될 수 있음

4] (AI 풀스택 확보 경쟁) AI 인프라를 구축하기 위해 하드웨어부터 소프트웨어까지 모든 단계를 수직 계열화하여 직접 갖추는 AI 풀스택 확보를 위해 치열한 경쟁 중

- 가트너는 AI 주권 확보를 위해 2027년까지 전 세계 국가 35%가 자국법과 문화, 인프라에 최적화된 '소버린 인공지능(Sovereign AI)' 체제로 전환할 것으로 전망('26.1.29.)
 - (미국) 미국의 AI가 전 세계 표준이 되는 것을 목표로 AI 기술(하드웨어, 모델, 표준 등)을 동맹국에 적극적으로 수출하고 기술 견제 대상국에는 첨단 AI 기술에 접근하지 못하도록 반도체 및 관련 기술에 대한 수출 통제를 강화(America's AI Action Plan, '25.7.)
 - ※ 트럼프 대통령은 동맹국에도 강경한 태도를 보이고 있음("엔비디아의 최첨단 반도체는 미국을 제외한 누구에게도 수출하지 않을 것('25.11.)")
 - (중국) AI 반도체 개발, 데이터센터 및 클라우드 구축 등 기술 자립과 함께 모델, 운영체제(OS) 분야를 중심으로 오픈소스 전략을 취하여 전 세계 스타트업이 중국의 오픈소스를 기반으로 기술을 개발하고 생태계를 구축하도록 하여 글로벌 AI 인프라를 주도하고자 함
 - ※ AI 오픈소스 모델 다운로드 기준, 중국산 모델이 미국산 모델을 제치고 세계 1위(MT&Hugging Face, '25.11.)
 - (한국) AI 컴퓨팅 인프라 및 데이터센터 구축, 산업별 수요 기반 AI 파운데이션 모델 확보, 피지컬 AI 전환 대응 등 글로벌 AI 강국이 되기 위한 「대한민국 인공지능 행동계획」 수립('26.2.25.)

- 빅테크 기업 또한 서비스 성능 최적화와 공급망 통제를 위해 자사 AI 모델의 니즈에 맞게 반도체 설계부터 주도권을 가지고 설계·생산·패키징 전반을 주도하는 밸류체인으로 재편 중(PWC, '25)
 - 구글(TPU), 마이크로소프트(Maia), 아마존(Trainium), 메타(MTIA)는 자사 서비스에 최적화된 맞춤형 반도체(ASIC)의 비중을 급격히 늘리고 있음

※ 자사 특정 AI 모델에 최적화·효율화, 비용 절감, 반도체 공급망 주도권 확보, 모델부터 반도체 아키텍처까지 설계하여 독보적인 서비스 제공 등을 목적으로 하고 있음

〈AI 반도체 개발에 뛰어든 전통 반도체 기업과 빅테크 기업 예시〉

반도체 기업		빅테크 기업	
기업명	특징	기업명	특징
NVIDIA (학습 ◎)	독보적 1위로 GPU뿐만 아니라 소프트웨어 플랫폼 CUDA를 통해 개발자 생태계 장악	Google (학습 ◎)	구글 클라우드에 최적화된 AI 반도체(TPU) 개발
AMD (학습 ◎)	개방형 생태계와 고대역폭 메모리(HBM)를 탑재한 고속기로 데이터센터 AI 시장 확대 시도	Amazon	클라우드 비용 최적화를 위해 학습용과 추론용 칩을 구분해 개발
Intel	파운드리 사업과 함께 Gaudi 등 자체 AI 가속기를 병행하며 제조·설계 양쪽에서 경쟁력 회복 시도	Microsoft	Open AI 모델 구동 최적화를 위한 보완적 자체 AI 반도체 개발 전략
Arm	직접 칩을 제조하지 않고 설계를 하며 컴퓨팅 서브시스템(CSS)로 빅테크 기업이 자체 AI 반도체를 더 쉽게 만들 수 있도록 돕는 파트너 전략	Meta	오픈소스 시가속화를 추진하며 라마(Llama) 모델에서 최적화된 자체 MTIA 개발 중
Broadcom	맞춤형 반도체(ASIC) 설계(구글의 TPU, 메타의 MTIA 개발 참여, Open AI와도 협력 계약), AI 데이터센터 네트워킹 분야 강자	Apple	서버보다는 아이폰, 맥북 등에서 작동하는 온디바이스 AI에 집중
Marvell	Aws의 Trainium과 Inferentia 개발에 관여, 전기 대신 빛을 이용해 데이터를 전송하는 광자 연결망 기술을 보유한 Celestial AI 인수	알리바바	피지컬 AI 전용 모델 '린브레인' 오픈소스 공개, 자회사인 핑터우거는 학습/추론이 가능한 자체 AI 칩인 '전우 810E' 공개
캠브리콘	AI 반도체 및 시스템 설계회사로 클라우드 서버, 엣지 컴퓨팅, 스마트 단말기 등 AI 반도체 설계	Huawei	Ascend 시리즈는 학습/추론용 반도체이며 단일 칩을 넘어 데이터센터 AI 아키텍처도 개발

5] (게임체인저 기술의 등장) 뉴로모픽/옵티컬/양자 컴퓨팅 등의 신기술이 개발되고 있으며 이러한 기술의 등장에 따라 AI 생태계 패러다임이 변화할 수도 있음

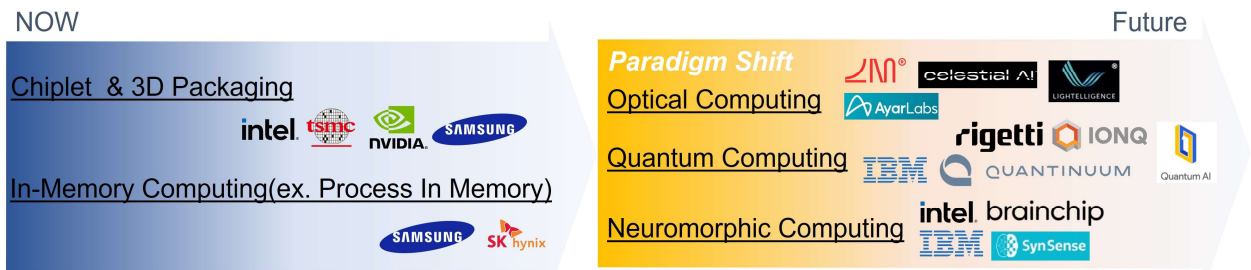
- 단기적으로는 칩렛(Chiplet) & 3D 패키징 기술이나 PIM 기술은 비용 절감, 성능 극대화, 에너지 효율성 향상 등 AI 반도체의 한계를 극복하고 성능을 좌우하는 핵심기술로 작용할 것으로 보임

* 엔비디아의 그레이스 블랙웰(GB) 칩을 잇는 과정에서 발열 관련 문제가 발생하여 AI 데이터센터 구축 업체인 오라클은 약 1억 달러의 손실이 발생('25)

- 아직은 초기 단계이지만 장기적으로는 옵티컬/뉴로모픽/양자 컴퓨팅 기술이 AI 반도체 패러다임 자체를 바꿀 수도 있음
- 하지만, AI 반도체는 AI 연산을 가속화하기 위한 목적이 크기 때문에 무엇보다도 핵심은 트랜스포머 아키텍처를 뛰어넘는 차세대 AI 모델에 따라 AI 반도체의 향방이 크게 좌우될 것으로 보임

〈 보완적 기술과 패러다임 전환 기술 예시와 설명 〉

구분	기술명	설명
①보완적 기술	칩렛 & 3D 패키징	하나의 거대한 칩을 만드는 대신, 기능별로 작은 조각(Chiplet)을 만들어 하나로 이어 붙이는 기술이며, 특히 3D 패키징은 수직으로 쌓아 올려 공간 효율성, 데이터 전송 속도와 전력 효율을 극대화하는 첨단 공법
	PIM(Process in Memory)	메모리와 프로세서를 하나로 합친 기술로 데이터를 계산하기 위해 메모리에서 CPU/GPU로 옮기는 것이 아니라, 메모리 내부에서 직접 계산을 수행 ※ 상용화 현황: 스마트폰, 노트북 등에 활용되는 저전력 메모리(LPDDR)용 PIM 개발 중
②패러다임 전환 기술	옵티컬 컴퓨팅	빛을 정보 전달의 연산과 매개체로 사용하여 빛의 속도로 대용량 데이터를 병렬 처리 하는 초고속·저전력 컴퓨터 ※ 상용화 현황: 실리콘 포토닉스 기반의 데이터 전송 및 인터커넥트 기술은 데이터 센터를 중심으로 상용화 단계에 진입하였으나 광학 연산 칩은 실험 단계
	뉴로모픽 컴퓨팅	인간의 뇌를 완전히 모방하여 뇌의 신경망처럼 특정 신호가 일정 수치를 넘을 때만 전기 스파이크를 발생시켜 정보를 전달 ※ 상용화 현황: (Intel) 차세대 뉴로모픽 연구 칩인 로이히(Loihi)3 공개('25.10.), 일반 판매 제품은 아니며, 연구 및 평가 목적의 '연구소 플랫폼'으로 제공
	양자 컴퓨팅	중첩과 얽힘이라는 양자역학적 특성을 가진 큐비트(Qbit)를 사용해, 수많은 경우의 수를 동시에 계산, 기존 연산을 완전히 대체하기 보다 일부 Task에서만 대체 가능 ※ 상용화 현황: 제약·화학 분야를 중심으로 분자 시뮬레이션 등 파일럿 수준의 활용
③와해성 기술	POST Transformer	현재 엔비디아의 GPU가 압도적인 이유는 트랜스포머의 병렬 행렬 연산을 가장 잘 수행 하기 때문에 새로운 모델에 맞춤형으로 판도가 완전히 뒤집어질 가능성



〈 보완적 기술과 패러다임 전환 기술의 대표기업(예시) 〉

- 이 외에도 Cerebras의 웨이퍼 스케일 아키텍처 등 새로운 기술이 계속해서 등장하고 있으며 에너지 효율, 데이터 전송의 병목 해소, 연산 시간 단축 등 기존 기술의 한계를 돌파한다면 언젠가 주류가 바뀔 수 있음

4 결론

□ AI 반도체 시장은 생성형 AI 확산과 함께 가파르게 성장하며 핵심 전략 산업으로 부상하고 있으며, 효율·전력·공급망 안정성 중심으로의 전환, AI 풀스택 확보를 위한 국가 간 경쟁, 차세대 기술 발전 등에 따라 시장 변화 예상

- (GPU 독주에서 'GPU + ASIC' 하이브리드 체제로의 전환) '대규모 학습(Training)'은 여전히 GPU가 주도 하되, 폭증하는 '추론(Inference)' 시장은 용도별·서비스별로 최적화된 저전력·저비용·맞춤형 ASIC이 분담 하는 하이브리드 구조가 될 것으로 예상

※ Bloomberg에 따르면, 전체 AI 반도체 시장은 연평균 16% 성장하는 반면, 맞춤형 ASIC 시장은 연평균 27%씩 성장할 것으로 예상('26)

- 다만, 향후 '포스트 트랜스포머' AI 모델의 등장은 GPU 우위의 AI 반도체 시장이 근본적으로 바뀔 수 있는 최대 변수로 작용할 수 있음

- **(국가 전략 자산으로의 'AI 풀스택'과 소버린 AI 경쟁)** AI는 단순 기술을 넘어 국가 안보와도 관련된 최첨단 전략기술로 각국은 반도체-데이터-모달-서비스 등에서 외부 의존을 최소화하고 통제하는 'AI 풀스택' 확보에 사활을 걸고 있음
 - 다만, 미국의 공급망 재편에 맞서 각국이 소버린 AI를 추진하게 되면, 글로벌 표준과 병행하여 국가·권역별로 상이한 규제 및 운영 기준이 형성되면서 부분적인 시장 분절화가 나타날 가능성도 존재
 - **(기술 제재가 촉발한 중국의 독자 생태계 강화)** 미국의 기술 제재에 중국은 오히려 반도체 자립을 앞당기고 있으며 미국의 수출 통제를 받는 나라와의 협력, 오픈소스 기반의 중국 주도 생태계를 구축하여 글로벌 AI 시장을 장악하려 하고 있음
 - ※ 중국의 딥시크(DeepSeek)는 GPU가 없어도 대규모 언어 모델이 업계 표준에 근접하는 성능을 달성('24.12.)했으며, 지푸AI는 화웨이 반도체만으로 이미지 생성 모델 학습을 완료('26.1.)
 - 중국으로 우회 수출을 막기 위한 제3국에 대한 수출 통제는 오히려 중국산 AI 반도체의 판로를 넓혀주는 부작용을 낳고 있으며, 이는 글로벌 시장에서 영향력을 확대하려는 중국에 유리한 환경을 조성
 - **(빅테크 기업의 밸류체인 개편과 '에너지' 주권 경쟁)** 빅테크 기업은 칩 구매자를 넘어 반도체 설계 단계부터 직접 참여하여 시장 변동성과 공급망 리스크에 대응하고 자사 서비스에 최적화된 고효율 인프라를 구축 중
 - 또한, 전력 확보 및 저전력화를 위한 빅테크 기업의 노력은 점점 더 심화되고 있으며 '전기'의 확보·효율을 통한 데이터센터의 원활한 운영으로 AI 서비스 시장의 주도권을 가지게 될 수도 있음
 - **(차세대 컴퓨팅 기술의 잠재력)** 뉴로모픽, 양자 컴퓨팅 등의 혁신 기술은 장기적 관점에서 AI 반도체의 지형을 바꿀 수도 있는 잠재력을 보유하고 있지만, 아직은 연구개발 및 초기 검증 단계로 기존 AI 반도체를 대체하기보다는 중·장기적으로 특정 연산에서 보완적으로 활용될 가능성이 높음
- 우리나라도 '국가대표 AI' 프로젝트, 'K-온디바이스 AI 반도체 기술개발' 프로젝트 등을 추진하고 있으며, 급증하는 AI 인프라 수요에 대응하기 위해 'SMR 특별법' 통과('26.2.) 등 에너지 기반 확보에도 노력
- 또한, AI 기본사회를 위한 생태계 조성과 산업·공공·지역 분야의 AI 전환, AI를 통한 기본사회 가치 실현 등의 전략 추진을 통해 AI 3대 강국으로 도약하고자 함(대한민국 인공지능 행동계획, '26.2.25.)
- AI 반도체 시장은 GPU+ASIC 하이브리드 구조로 전환되는 변곡점에 있으며, AI 반도체 기술 주권 확보는 산업 경쟁력을 넘어 외교·안보 자산으로 작용함에 따라 국가 차원의 AI 풀스택 전략 수립이 필요
- **(GPU+ASIC 전략적 포지션 확보)** 폭증하는 추론 시장에서는 맞춤형 ASIC과 메모리 최적화가 핵심 경쟁 요소로 한국은 메모리-제조 경쟁력을 기반으로 전략적 우위 확보 필요
 - **(초격차 메모리 주권 사수)** HBM 생산의 핵심 플레이어로 주권 사수와 동시에 메모리가 저장장치만이 아닌 추론 비용 절감과 에너지 효율 개선을 동시에 달성하는 AI 연산의 핵심 구성요소가 될 수 있도록 전략 필요
 - **(제조업 수요 연계)** 스마트폰, 자동차, 로봇 등 제조 기반 산업을 활용하여 저전력 AI 반도체, 로봇 파운데이션 모델, HW 플랫폼 등을 결합한 '피지컬 AI' 생태계 구축과 온디바이스 AI 시장 선점 필요
 - **(AI 풀스택의 운영 자립)** AI 모델, AI 반도체의 국산화뿐만 아니라 공공 데이터, 보안 환경, 서비스 운영 체계 까지 국내 기술로 안정적으로 운영할 수 있는 '소버린 AI 운영 플랫폼' 구축 필요

용어해설

- GPU(Graphic Processing Unit) 대규모 병렬 연산에 최적화된 프로세서로 본래 그래픽용이었으나, 수천 개의 코어로 데이터를 동시에 처리하는 병렬 구조가 AI 연산에 적합해 가장 대중적인 AI 가속기로 쓰임
- ASIC(Application-Specific Integrated Circuit) 특정 응용 목적이나 연산 워크로드에 최적화되어 설계된 직접 회로로 2010년대 초반 비트코인 채굴 분야에서 압도적인 효율을 증명하였으며 최근에는 대규모 AI 연산에 맞춘 특화된 전문 ASIC이 등장하기 시작
- TPU(Tensor Processing Unit) 구글이 자사 AI 워크로드 가속을 위해 설계한 AI 전용 가속기로 대규모 행렬 연산에 최적화된 맞춤형 주문형 반도체(ASIC)의 한 종류
- NPU(Neural Processing Unit) 신경망 연산에 특화된 AI 가속기로 주로 스마트폰이나 노트북 등 엣지 디바이스에서 저전력으로 AI 추론을 효율적으로 수행하기 위해 사용
- LPU(Language Processing Unit) 대규모 언어모델(LLM)의 추론 처리에 최적화된 연산 구조를 갖춘 가속기로, 메모리 접근 병목을 최소화하여 낮은 지연시간을 목표로 설계된 ASIC 계열의 칩
- FPGA(Field-Programmable Gate Array) 특정 작업에 맞게 논리 구성을 재프로그래밍할 수 있는 반도체로, 높은 유연성을 바탕으로 다양한 알고리즘이나 연산 구조 변경에 대응

※ 본문의 이해를 돕기 위한 용어 해설로 해당 용어의 공식적인 정의는 아닙니다.

참고문헌

- ArXiv, “Neuromorphic Computing-An Overview”, 2025.10.8.
- BizTech Magazine, “Intel’s New Core Ultra Signals the Next Generation of AI Computing”, 2024.4.24.
- Bloomberg, “AI Accelerator Market Looks Set to Exceed \$600 Billion by 2033, Driven by Hyperscale Spending and ASIC Adoption, According to Bloomberg Intelligence”, 2026.1.14.
- Bloomberg Intelligence, “AI Accelerator Chips 2026 Outlook”, 2026.1.12.
- BRANDSIT, “AI accelerator market in Europe: digital sovereignty vs. Nvidia’s dominance”, 2025.10.
- CLOVA, “AI 반도체”, 2026.1.30.
- Deloitte Insights, “온디바이스AI 시대: 시장 전망 및 활용 방안“, 2025.5.
- Emergen Research, “AI Accelerator Chip Market”, 2025.12.
- Fundamental Business Insights, “AI accelerator Market”, 2025.10.
- Gartner, “Gartner Predicts Power Shortages Will Restrict 40% of AI Data Centers By 2027”, 2024.11.
- IBM, “What’s the difference between AI accelerators and GPUs?”
- IBM, “AI반도체 확보에 어떻게 도전할 것인가?”, 2025.

- IEA, “Energy and AI”, 2025.10.
- IEA, “Electricity Mid-year Update 2025”, 2025.7.30.
- IEA, “World Energy Outlook 2025”, 2025.11.
- Lucintel Insights that Matter, “AI Accelerator Market Report”, 2025.12.
- McKinsey & Company, “What is quantum computing”, 2025.3.31.
- MIT Technology, “[인터뷰] 삼바노바 CEO “엔비디아는 에이전트 시대를 감당할 수 없다”, 2025.11.28.
- Polaris Market Research, “Edge AI Accelerator Market Size, Share, Trends, Industry Analysis Report”, 2025.5.
- PWC, “AI반도체 시장, 이제 진영간 전쟁이다”, 2025.7.
- PWC, “AI반도체 차세대 선두주자는 누가 될 것인가?”, 2025.7.
- Pittsburgh, “A multi-level breakthrough in optical computing”, 2024.10.23.
- Spherical Insights, “Global Edge AI Accelerator Market Size To Exceed USD 94.61 Billion by 2033: Market Study Report”, 2025.5.
- Supermicr, “What is an AI Accelerator?”
- A de Vries, “The growing energy footprint of artificial intelligence”, Joule, Volume 7, Issue 10, 18 October 2023.
- Alexandra Sasha Luccioni, et al., “Power Hungry Processing: Watts Driving the Cost of AI Deployment?”, 2023.11.
- Daniel Kokotajlo, et al., “AI 2027”, 2025.
- Dhireesha Kudithipudi et al., “Neuromorphic computing at scale”, Nature, 637, 801–812, 2025.1.
- Lohani, V, “How Transformers Changed AI Forever”, Medium, 2025.7.
- KISTEP, “인공지능 반도체” 2023.4.
- 과기정통부 블로그, “전기로 만들어지는 AI: 에너지 없는 인공지능은 없다”, 2025.6.
- 대외경제정책연구원, “세계 인공지능대회(WAIC 2025)를 통해 본 중국 AI 발전 현황 및 시사점”, 2025.9.
- 반도체공학회, “반도체 기술 로드맵 2026”, 2025.12.

저자

KISTEP 재정투자분석본부 연구개발예산정책센터 정의진 연구위원 (ejin@kistep.re.kr, 043-750-2443)

KISTEP 글로벌기술전략본부 기술예측센터 신동평 센터장/연구위원 (sheendp@kistep.re.kr, 043-750-2439)

KISTEP 글로벌기술전략본부 손석호 본부장/선임연구위원 (shson@kistep.re.kr, 043-750-2346)