참조 용어 | 그래픽 처리 장치(GPU) p.19

087 AI 반도체

AI Semiconductor

AI 학습·추론을 위한 대규모 병렬 연산용 특화 반도체

- 대량의 데이터를 동시에 계산해 AI 학습·추론 속도를 높이고 전력 소모를 줄이는 핵심 반도체 기술
- 클라우드와 데이터 센터, 에지 기기 등에서 AI 서비스 성능과 비용을 좌우하는 핵심 부품

AI 반도체 개요

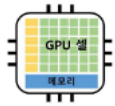
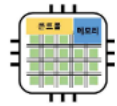
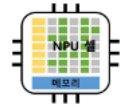
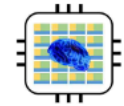
AI 반도체는 인공지능이 방대한 데이터를 학습하고 추론할 때 필요한 연산을 효율적으로 처리하도록 설계된 칩입니다. 크게 범용 반도체(CPU·GPU)와 특수 목적 반도체(ASIC)로 구분되며, ASIC에는 TPU·NPU 등이 포함됩니다. 기존 중앙처리장치(CPU)가 순차적으로 작업을 수행한다면 GPU, TPU, NPU 등은 대규모 행렬·벡터 연산을 병렬로 처리하도록 설계돼 속도와 효율이 더욱 높습니다.

AI 반도체의 중요성

AI 반도체는 기술 발전 속도와 품질을 좌우하는 핵심 하드웨어입니다. 대규모 모델 학습과 추론은 막대한 연산량과 데이터 이동을 요구하므로, AI 반도체의 성능은 학습 시간·비용·정확도에 직접적인 영향을 줍니다. 초거대 모델과 생성형 AI의 확산으로 연산 수요가 폭발적으로 증가함에 따라, 범용 칩만으로는 대응이 어렵습니다. 이에 따라 고성능·저전력·고밀도 연산이 가능한 AI 전용 칩 수요가 급증하고 있습니다.

AI 반도체의 전망

AI 반도체 시장은 기존 GPU 중심 구조에서 TPU·NPU 등 특정 AI 작업에 최적화된 ASIC 기반 구조로 다양화되고 있습니다. 미국·유럽·중국·한국 등 주요국은 반도체를 전략 산업으로 지정하며 연구개발과 공급망 투자를 강화하고 있고, 기업들 역시 AI 모델 특성에 맞춘 맞춤형 ASIC 개발을 통해 성능 우위를 확보하고 있습니다. 여기에 메모리 내 연산으로 데이터 이동을 줄이고 속도·효율을 높이는 인메모리 컴퓨팅, 여러 개의 소형 칩을 조립해 확장성과 생산 효율을 높이는 칩렛 아키텍처와 같은 차세대 기술이 확산되면서 성능과 효율 측면의 경쟁이 더욱 치열해지고 있습니다. 향후 에너지 효율, 탄소 저감, 보안성이 핵심 과제로 부상할 것으로 보이며, 특히 독자적 칩 설계 능력을 갖춘 국가와 기업이 AI 산업의 주도권을 확보할 것이라 전망됩니다.

종류	CPU(1세대)	GPU(1세대)	FPGA(2세대)	ASIC(2세대)	뉴로모픽(3세대)
특징					
	복잡 계산 순차처리	단순 계산 병렬처리	목적별 하드웨어 재구성	용도 맞춤형 고효율 전용칩	뉴런과 시냅스를 모방한 新구조

세대별 반도체

출처 : 특허청