

011 그래픽 처리 장치 / GPU

Graphics Processing Unit

대규모 병렬 연산을 수행해 AI 연산 속도를 높이는 고성능 연산 장치

- 수천 개의 코어를 통해 동시에 연산을 수행하며,
딥러닝·AI 학습·과학 계산 등 대규모 데이터 처리에 필수적인 하드웨어
- 고속 연산 능력과 메모리 대역폭으로 AI 모델 훈련·추론의 효율성을 크게 향상시키는 핵심 인프라

GPU 개요

GPU는 원래 그래픽 렌더링을 위해 개발된 장치였으나, 현재는 대규모 병렬 연산에 최적화된 범용 연산 플랫폼으로 발전했습니다. 수천 개의 코어가 동시에 작동해 대량의 데이터를 병렬 처리할 수 있어, CPU보다 훨씬 빠르게 연산을 수행합니다. 이러한 구조 덕분에 GPU는 이미지 처리뿐 아니라 AI 학습·자연어 처리·과학 계산 등 고속 연산이 필요한 모든 영역에서 핵심 장치로 사용되고 있습니다. 특히 AI 모델 학습에 필요한 매개변수 억 단위 이상의 연산을 효율적으로 수행하여 AI 발전의 연산 기반을 제공합니다.

GPU의 활용

GPU는 현재 AI 반도체 산업과 클라우드 인프라의 핵심 축으로 자리 잡고 있습니다. 데이터센터와 클라우드 기업은 GPU 서버를 통해 거대 AI 모델을 학습·운영하고 있으며, 주요 반도체 기업이 GPU 시장을 주도하고 있습니다. 또한 GPU는 자율주행, 로보틱스, 의료 영상 분석, 기후 시뮬레이션 등에서 활용되며, AI 서비스의 품질과 효율을 좌우하는 핵심 요소가 되고 있습니다. 앞으로 GPU는 단순한 가속기를 넘어, AI 연산의 표준 플랫폼이자 산업 경쟁력의 핵심 인프라로 기능할 것으로 전망됩니다.

관련 용어

중앙처리장치 (Central Processing Unit, CPU)

CPU는 컴퓨터 시스템의 중앙 연산 장치로, 명령어를 해석하고 연산·제어·입출력 관리 등 전반적인 처리를 담당합니다. 복잡한 연산을 순차적으로 수행하도록 설계되어 범용성과 안정성이 높지만, 대량의 데이터를 동시에 처리하는 병렬 연산에는 한계가 있습니다. AI나 그래픽 연산처럼 연산량이 많은 작업에서는 GPU나 NPU에 비해 속도가 느리지만, 여전히 운영체제 제어·프로그램 실행·논리 연산 등 시스템의 핵심 역할을 수행합니다. 최근에는 CPU와 GPU, NPU를 결합한 이기종 연산 구조(Heterogeneous Computing)도 많이 활용되고 있습니다.

관련 용어

신경망처리장치 (Neural Processing Unit, NPU)

NPU는 AI 연산, 특히 인공신경망 구조의 학습과 추론에 특화된 연산 장치입니다. GPU보다 연산 구조가 단순하지만, 행렬 연산과 벡터 계산을 병렬로 처리하는 데 최적화되어 있습니다. 이러한 구조 덕분에 전력 소모가 적고, 모바일·애지 기기에서도 AI 연산을 실시간으로 수행할 수 있습니다. 스마트폰의 이미지 인식, 음성 비서, 자율주행 센서 제어 등 저전력 환경에서 고속 연산이 필요한 분야에서 주로 활용됩니다. 최근에는 클라우드 서버용 NPU도 등장해 AI 가속기 생태계의 핵심 구성 요소로 성장하고 있으며, GPU 대비 에너지 효율을 중심의 차세대 AI 연산 아키텍처로 주목받고 있습니다.

관련 용어

텐서처리장치 (Tensor Processing Unit, TPU)

TPU는 구글이 개발한 AI 전용 연산 프로세서로, 딥러닝 모델 학습과 추론에서 많이 사용되는 행렬·텐서 연산을 빠르게 처리하도록 설계된 장치입니다. GPU가 다양한 병렬 작업을 수행하는 범용 가속기라면, TPU는 신경망 연산 흐름을 효율적으로 처리하는 데 초점을 둔 전용 구조를 갖고 있습니다. 구글 발표에 따르면 TPU는 대규모 모델 학습이나 특정 워크로드에서 높은 처리 속도와 에너지 효율을 보였다고 하지만, 이러한 차이는 모델 종류나 환경에 따라 달라질 수 있습니다. TPU는 주로 텐서플로우(TensorFlow) 기반의 대규모 학습 환경에서 활용되며, 클라우드 데이터센터에서 AI 성능을 높이기 위해 활용되고 있습니다.

관련 용어

데이터 처리장치 (Data Processing Unit, DPU)

DPU는 대규모 데이터 이동·저장·네트워크 처리와 같은 데이터 관리 업무를 전담하도록 설계된 프로세서입니다. AI 연산을 담당하는 GPU나 NPU와 달리, DPU는 데이터 패킷 처리, 암호화·압축, 스토리지 관리, 네트워크 가상화 등 시스템 운영에 필요한 주변 작업을 하드웨어 수준에서 가속합니다. 특히 AI 모델을 대규모로 운영하는 데이터센터에서는 연산 처리보다 데이터 이동과 I/O 병목이 성능을 좌우하는 경우가 많아, DPU가 이를 분리해 처리함으로써 전체 시스템 효율을 크게 높일 수 있습니다. 최근에는 CPU·GPU와 함께 데이터센터 3대 가속기로 불리며, 클라우드 인프라와 AI 서비스의 확장성을 뒷받침하는 핵심 장치로 주목받고 있습니다.



CPU

- 소형 모델 및 데이터셋
- 설계공간탐색에 유용



NPU

- 뇌구조 모방
- 저전력, 저지연
- 실시간 처리
- 병렬 처리 특화



DPU

- 범용 병렬 처리
- 네트워킹, 스토리지,
- 데이터 이동



GPU

- 중대형 모델 및 데이터셋
- 이미지, 비디오 처리



TPU

- 행렬 연산
- 밀집 벡터 처리
- 사용자 정의 연산
- 사용 불가