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·에지 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 효율 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.