126 부동 소수점 연산 / FLOPS

Floating point Operations Per Second

컴퓨터가 실수(Real Number) 계산을 수행하는 연산량 또는 연산 속도 지표

- AI 모델 학습·추론에 필요한 계산 능력을 나타내는 핵심 성능 지표
- GPU·AI 칩의 처리량과 시스템 효율을 비교할 때 널리 사용됨

FLOPS란?

부동 소수점 연산은 컴퓨터가 소수점이 포함된 실수를 계산하는 과정, 또는 그 연산을 수행할 수 있는 총량을 의미합니다. 정수보다 표현 범위가 넓고 다양한 크기의 수를 다룰 수 있어 과학 계산·물리 시뮬레이션·그래픽 처리처럼 복잡한 계산이 필요한 분야에서 필수적으로 사용되며, '얼마나 많은 실수 연산을 1초에 처리할 수 있는가'를 나타내는 성능 지표입니다. EU AI Act는 FLOPS를 GPAI의 법적 기준으로 사용하지는 않으나, "대규모 연산을 사용한 모델이 범용 모델일 가능성이 크다"는 취지로 참고 지표로써 언급한 바 있습니다. AI 모델은 행렬 곱셈, 벡터 연산처럼 실수 기반의 수학적 계산을 반복적으로 수행하므로, FLOPS는 AI 연산 능력을 이해하는 기본 척도가 됩니다.

AI에서 FLOPS의 활용

AI 모델 학습은 대규모 데이터와 매개변수를 기반으로 한 연속적 행렬 연산으로 구성됩니다. 이 과정에서 GPU나 AI 전용 칩은 수조 단위의 실수 계산을 처리해야 하므로, FLOPS는 AI 학습 효율과 처리 성능을 비교하는 핵심 기준으로 사용됩니다. 예를 들어 LLM은 학습 과정에서 수십억~수조 단위의 연산을 반복하기 때문에, 칩의 FLOPS 성능이 높을수록 학습 속도와 비용 효율이 크게 개선됩니다. 추론 단계에서도 FLOPS는 중요한데, 특정 모델이 사용자 요청에 얼마나 빠르게 응답할 수 있는지 판단하는 데 참고 지표가 됩니다. 다만 FLOPS만으로 실제 사용자 체감 속도를 모두 설명할 수는 없으며, HBM, 병렬 처리 구조, 최적화 알고리즘 등이 함께 작용해야 전체 성능이 확보됩니다.

FLOPS의 의의

FLOPS는 오랫동안 컴퓨터와 AI 하드웨어 성능을 비교하는 데 사용된 대표적 지표로, GPU·NPU·AI 가속기 같은 연산 장치가 얼마나 복잡한 계산을 처리할 수 있는지를 정량적으로 보여줍니다. 특히 AI 칩 경쟁에서는 TFLOPS(테라), PFLOPS(페타), EFLOPS(엑사) 같은 대규모 연산 단위가 주요 벤치마크로 활용되며, 데이터센터나 모델 개발 기업은 FLOPS 성능을 기준으로 연산 자원을 선택하는 경우가 많습니다. 그러나 FLOPS는 이론적 연산 처리량에 가깝기 때문에, 실제 성능을 과대평가할 수 있다는 한계도 있습니다. 메모리 병목, 통신 지연, 소프트웨어 최적화 부족 등은 FLOPS 수치와 별개로 성능 저하를 유발할 수 있습니다. 그럼에도 FLOPS는 AI 연구·산업 전반에서 공통적으로 사용되는 기본 성능 지표로써, 모델 규모와 연산 요구량을 이해하는 출발점으로 중요한 의미를 갖습니다.