창작의 신기원인인가, 열린 사회의 적인가?

인공지능 반도체

권오현 계간 스켑틱 편집자

우리가 직장과 학교, 일상생활에서 손쉽게 이용하는 챗-GPT(Chat-GPT)는 초거대 AI로서 대용량 데이터를 학습해 작동한다. 최신 버전인 GPT-4o의 파라미터 수는 공개되지 않았으나 과거 모델인 GPT-3의 1,750억 개를 월등히 뛰어넘으며, 오디오와 비디오까지 처리하는 멀티모달 능력을 갖추고 있다. 이런 초거대 AI의 고도화는 AI의 데이터 학습과 연산을 수행하는 하드웨어인 반도체의 혁신 없이는 불가능하다. 최근 챗-GPT를 만든 오픈AI(OpenAI)의 CEO 샘 올트먼(Sam Altman)은 AI 반도체 자체 생산을 위해 9,000조 원 투자 유치에 나섰다. 마이크로소프트(Microsoft)와 메타(META) 등 글로벌 빅테크 기업도 AI 반도체를 자체 생산하겠다고 발표했다. 반도체 선도 국가로서 우리나라도 예외는 아니다. 미래 AI 산업 전쟁은 곧 반도체 전쟁이다. 그렇다면 AI 반도체란 도대체 무엇인가?

AI의 분석과 추론에 특화된 전용 반도체의 출현

우선 반도체는 일종의 두뇌로서 그 기능에 따라 메모리 반도체와 비메모리, 즉 시스템 반도체로 나뉜다. 시스템 반도체는 연산을, 메모리 반도체는 이름처럼 정보 저장을 담당한다. 시스템 반도체에는 우리가 익히 아는 중앙처리장치(CPU), 한때 비트코인 채굴로 널리 쓰여 품귀 현상이 일었던 그래픽처리장치(GPU), 스마트폰에 들어가는 애플리케이션 프로세서(AP)가 있다.

AI 특화 반도체가 개발되기 전에는 AI의 두뇌 역할을 CPU와 GPU가 맡아 왔다. 이를 1세대 AI 반도체라고 할 수 있다. CPU는 범용 프로세서이기에 다양한 계산 작업을 수행할 수 있고 정확도가 높다. 그러나 주로 대규모 병렬 연산을 필요로 하는 초거대 AI 작업에는 속도가 느리다. 반면 GPU는 병렬식 구조 반도체로서 계산 속도가 매우 빠르다. 다만 정확도가 CPU만큼 높지 않다.

AI 반도체는 이 1세대의 단점을 극복하고 장점은 통합하고자 한다. AI 서비스를 구동하는 데 필요한 대량 연산을 초고속으로 그리고 초전력으로 실행하는 특화된 비메모리 반도체를 만드는 것이다. 이를 2세대 AI 반도체라 할 수 있다. 2세대 AI 반도체는 AI 딥러닝에 최적화됐다는 의미에서 NPU(신경망처리장치, Neural Processing Unit)라 부르기도 한다.

대표적인 NPU에는 프로그래밍 가능한 집적 회로(FPGA, Field-Programmable Gate Array)와 특정 용도용 집적 회로(ASIC, Application Specific Integrated Circuit)가 있다. 그 이름에서 알 수 있듯

이 특정 영역에서 AI 작업을 하기 위한 반도체다.

FPGA는 목적에 따라 반도체 칩 하드웨어를 재프로그래밍할 수 있는 반도체로 개발 시간이 짧고 유연성이 높아 AI를 학습시키는 알고리즘 변화에 효과적으로 대응 가능하다. 수요 기업이 AI 기능에 따라 맞춤 설계를 할 수 있고 고성능·저전력 구현이 가능하다. ASIC 역시 명확한 애플리케이션 기능과 목적을 가진 시스템을 저전력으로 구동하기 위해 활용하는 주문형 반도체다. CPU보다 범용성이 낮음에도 주요 빅테크나 스타트업 기업이 자신들의 제품 및 서비스에 특화된 AI 반도체를 개발하기에 유리하도록 돼 있다.

아직 개발 중인 3세대 AI 반도체는 기존 컴퓨팅 구조를 완전히 탈피해 인간 신경망 구조를 모사한 뉴로모픽 반도체다. 인간의 뇌는 1,000억 개가 넘는 신경세포가 시냅스를 통해 다른 뉴런과 서로 신호를 주고받으며 순식간에 정보를 처리한다. 또한 이런 시냅스는 병렬적으로 연결돼 있어 약 20W 수준 저전력으로도 복잡한 연산을 해낸다. 따라서 뉴로모픽 반도체는 컴퓨터가 특히 이해하기 어려운 비정형 문자, 이미지 같은 데이터 처리에 강점이 있으며 딥러닝 알고리즘의 효율성을 크게 향상한다. 뉴로모픽 반도체는 인간처럼 학습과 추론을 동시에 수행해 동시다발적인 연산과 정보 처리를 수행할 수 있는 것이다.

HBM과 AI 가속기

그런데 AI 반도체는 메모리 반도체와는 상관없는 걸까? 그렇지 않다. 기억을 저장하는 능력이 있어야 비로소 두뇌라고 할 수 있듯이 말이다. 2013년 우리나라 SK하이닉스가 최초로 개발한 HBM(고대역폭 메모리, High Bandwidth Memory)이 AI 시대 패권을 잡는 게임 체인저가 될 수 있다. HBM은 이름 그대로 넓은 대역폭을 지닌 메모리를 뜻한다. 여기서 '대역폭'이란 주어진 시간 내에 데이터를 전송하는 속도나 처리량으로서, 데이터 운반 능력을 의미한다. HBM은 현재 메모리 시장에서 가장 넓은 대역폭을 지닌 메모리 반도체인데, 단순하게 이야기하면 메모리 중 데이터를 가장 빠르게 처리하고 전송할 수 있다는 것이다.

HBM은 메모리 반도체인 D램을 여러 개 쌓아 만든 고성능 메모리다. 단순히 쌓아올린다고 해서 고성능인 것은 아니다. HBM의 핵심은 메모리 반도체(D램) 중간에 데이터가 오가는 일종의 도로라 할 수 있는 관통 전극을 만들어, 데이터 전송 속도를 대폭 늘린 것이다. 비유하자면, 일반적인 D램이 32차선에서 64차선 도로라면 HBM은 1,024차선 이상이다. 또한 HBM은 이런 D램을 여러 층으로 쌓아서 AI가 학습하는 데 필요한 방대한 데이터를 한 번에 저장하고 전달할 수 있다.

이것이 왜 AI 향상에 중요할까? 연산 속도가 아무리 빨라봤자 수많은 데이터를 저장하고 이를 빠르게 갖다 쓸 수 있는 메모리가 없다면 그 결과물은 보잘것없다. HBM은 데이터 병목 현상이 발생하지 않도록 연산 속도를 좇아 민첩하게 데이터를 전달해 주는 역할을 한다. 그래서 HBM은 AI 반도체를 포함해 AI 특화 컴퓨터를 구성하는 'AI 가속기'의 주요 부품이다.

현재 AI 반도체 시장을 선도하기 위해 전통적인 반도체 기업인 퀄컴(Qualcomm), 인텔(Intel), 엔비디아(NVDIA)는 물론 우리나라 삼성전자와 SKT 그리고 구글(Google), 아마존(Amazon) 등 글로벌 빅테크 기업들도 AI 반도체 개발에 뛰어들었다. 이미 AI에 사용하는 반도체 중 범용성이 높은 CPU, GPU 시장은 기술성숙 단계에 접어들었다. 때문에, 향후 제2, 제3세대 AI 반도체와 HBM 메모리 혁신을 일궈낸 기업이 패권을 잡을지도 모른다.