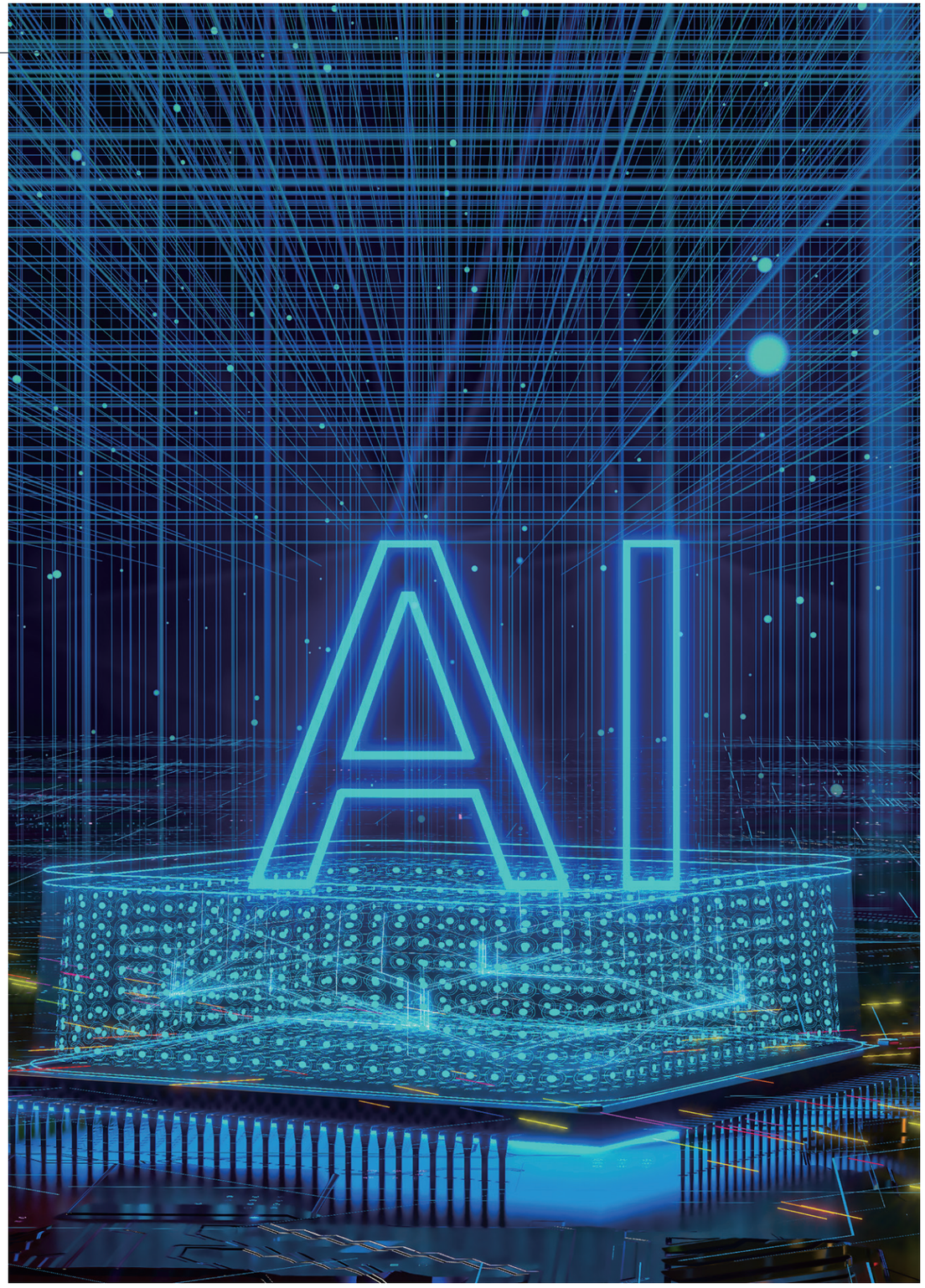
글로벌 빅 테크의 AI 칩 전쟁

엔비디아의 독주를 막아라

AI 태풍이 휘몰아치는 현재, 최고의 승자는 엔비디아다.
그러나 최후의 승자를 가리는 게임은 이제부터 시작이다.
글로벌 빅 테크 중 어느 회사도 AI를 외면할 수 없기 때문이다.

글 **김인순** 사진 Getty Images



2022년 11월 30일 공개된 챗GPT는 1년여 만에 1억 명의 주간 활성 사용자 수를 기록했다. 챗GPT 애플리케이션 프로그래밍 인터페이스(API)를 이용하는 개발자는 200만 명에 달한다.

챗GPT가 촉발한 생성 AI(Generative AI) 혁명이 1년간 휘몰아쳤다. 이러한 생성 AI 서비스의 기반인 대형언어모델(LLM)을 개발하고 실행하는 데 필요한 핵심 인프라는 두 가지가 있다. 바로 AI 전용 반도체 칩과 클라우드다.

AI 반도체란 무엇인가

AI 반도체란 AI 서비스 구현에 필요한 대규모 연산을 효율적으로 처리하는 데 특화한 비메모리 반도체를 말한다. AI 반도체는 기존의 중앙처리장치(CPU)와 그래픽처리장치(GPU)가 하는 일을 담당하는 등 'AI의 두뇌' 역할을 한다고 해도 과언이 아니다. AI 반도체는 수많은 데이터 연산과 추론을 초고속으로 처리한다.

지난해부터 올해까지 생성 AI 관련 애플리케이션이 속속 개발되고 있지만 관련 인프라는 그만큼 빠르게 성장하지 못했다. 본래 대부분의 비즈니스는 인프라가 만들어지고 콘텐츠와 생태계가 형성되는데, 생성 AI 시장은 콘텐츠가 먼저 나오고 인프라가 따라가는 추세다. 그리고 관련 시장을 엔비디아(NVIDIA)가 80% 이상 점유하는 등 인프라 시장 우위를 점령하며 성장 중이다.

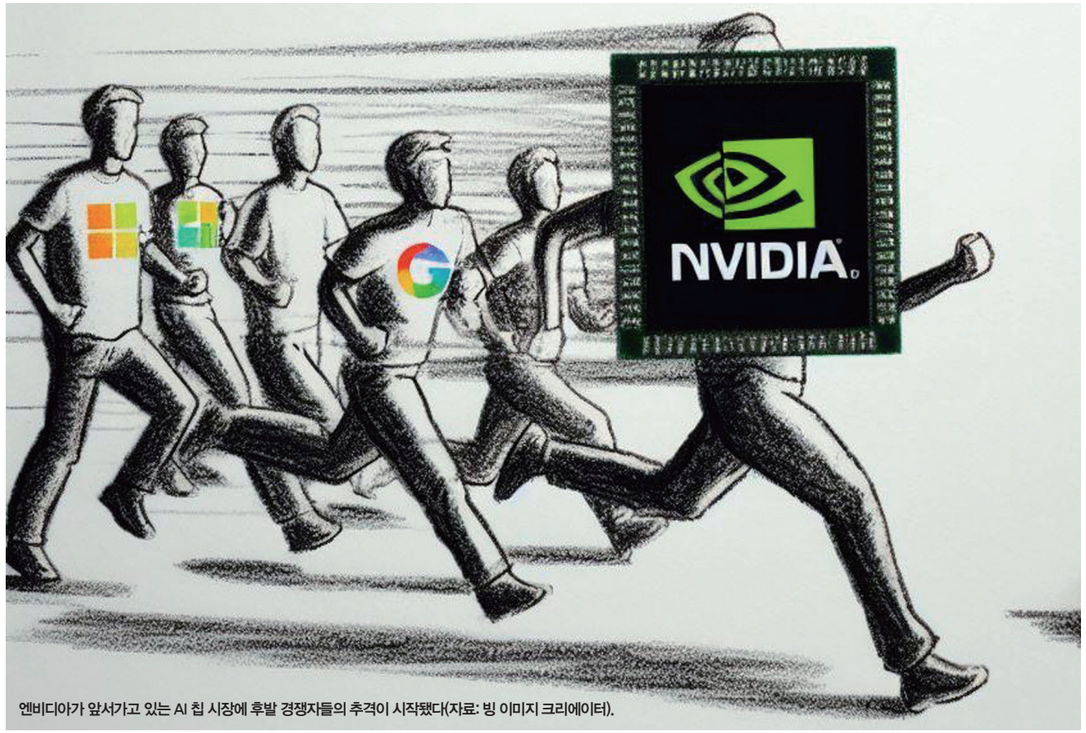
엔비디아는 생성 AI 붐 시장의 초기 승자라고 해도 과언이 아니다. 엔비디아와 대표 반도체 기업인 AMD와 인텔의 2023년 2분기 데이터센터용 칩 판매량을 비교해 보자. 데이터센터용 칩은 CPU와 GPU, 데이터처리장치(DPU) 등으로 구성되는데 3개 기업의 데이터센터 매출을 비교한 결과, 엔비디아 데이터센터 매출은 2년 동안 4배 증가하며 나머지 두 기업을 앞질렀다. 엔비디아의 독주 속에 AMD가 그나마 힘겨운 싸움을 하고 있다. 엔비디아 칩의 공급이 부족해지자, AMD는 이를 기회로 삼아 AI 전용칩 MI300X를 내놓고 반전을 노리고 있다. 반면, 인텔은 매년 데이터센터 매출이 감소하는 등 관련 시장에서 전혀 힘을 쓰지 못하고 있다.

한편 반도체 기업보다 적극적으로 엔비디아를 견제하는 곳은 따로 있다. 바로 빅테크 기업이다. 생성 AI 서비스와 클라우드 서비스를 가진 마이크로소프트, 구글, 아마존 등이 AI 칩 자체 제작에 뛰어들었다. 엔비디아의 고객들이 스스로 칩을 만들겠다고 나선 것이다.

빅테크 기업 입장에서는 엔비디아 독점 상황을 이대로 두고 볼 수 없는 것이 당연하다. 이 상황이 계속되면 생성 AI 서비스의 로드맵을 짜는 데도 엔비디아의 눈치를 보아야 하기 때문이다. 더군다나 실제로 엔비디아 칩 공급이 원활하지 않아 생성 AI 서비스 확장에 어려움도 겪고 있다. 과거 모바일 시장에서 퀄컴이 스마트폰 제조사의 라인업을 좌우했던 경험이 있는 만큼, 비슷한 상황이 펼쳐지는 것을 미연에 방지하겠다는 계산이다.

이처럼 빅테크는 미래 비즈니스를 위해 AI 칩 개발을 꼭 해야만 하는 상황에 놓인 한편, 스타트업 입장에서는 제2의 엔비디아가 될 수 있는 새로운 기회가 열린 셈이다. AI 칩 시장은 빅테크와 스타트업까지 뛰어들어 엔비디아에 도전장을 던지는 구도다.

김인순
인싸이트아웃 대표. 김인순 대표는 전자신문 ICT융합부 데스크 출신으로 20년간 보안 소프트웨어 분야를 전문적으로 취재했다. 기자협회 '이달의 기자상'을 두 차례 수상했다. 실리콘 밸리의 혁신기업을 취재한 "파괴자들 ANTI의 역습"을 집필했다. 더밀크코리아 대표를 역임하고 테크 커뮤니케이션 기업 인싸이트아웃을 운영하고 있다.



엔비디아가 앞서가고 있는 AI 칩 시장에 후발 경쟁자들의 추격이 시작됐다(자료: 빙 이미지 크리에이터).

GPU 시장에서의 엔비디아 독주

AI 칩 시장을 독점하고 있는 엔비디아는 11월 14일(현지 시각) 뉴욕증권거래소(NYSE)에서 전 거래일 대비 2.13% 오른 496.56달러에 거래를 마치며 사상 최고치를 경신했다. 이전 최고가는 지난 8월 31일 기록한 493달러다. 시가총액은 1조 2,270억 달러(약 1,603조원)로 집계됐다.

2023년 초부터 11월 14일까지 엔비디아의 주가는 247% 폭등했다. 같은 기간 S&P 500지수와 나스닥지수가 각각 18%, 36% 상승한 것과 비교하면 엄청난 상승이다.

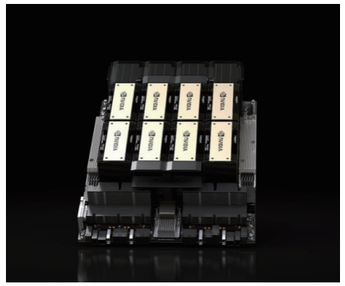
엔비디아는 본래 컴퓨터 게임용 그래픽을 처리하는 칩을 개발했다. 그런데 이 GPU가 AI 시대에 꼭 필요한 제품이 됐다. PC 시대에 인텔 칩이 필수였다면 AI 시대에는 엔비디아 GPU가 필수라는 말이 나온다. 글로벌 리서치업체 CB인사이트의 최근 보고서에 따르면 엔비디아는 기계 학습(Machine Learning)용 GPU 시장의 약 95%를 점유하고 있다.

엔비디아는 어떻게 AI 혁명의 핵심 플레이어가 됐을까? 전문가들은 엔비디아의 기술력과 대담한 베팅을 핵심적 요인으로 꼽는다. 엔비디아 CEO이자 창립자 중 한 명인 젠슨 황은 게임과 애플리케이션을 위한 그래픽 개선에 집중했고, 1999년 컴퓨터 이미지 디스플레이를 향상시키는 GPU를 개발했다. 이때 개발한 GPU는 많은 작업을 동시에 처리하는 데에도 탁월한 성능을 보였다. 이후 2006년, 스탠퍼드 대학 연구원들이 GPU가 일반 칩이 할 수 없는 방식으로 수학 연산을 가속화하는 것을 발견했다. 이때 젠슨 황은 그래픽 이외 용도로 병렬 처리하는 기능 개발에 투자했다. 젠슨 황의 이 같은 결정이 AI 혁신을 촉발하는 데 기여했다.

2012년 이미지를 분류할 수 있는 AI 시스템(아키텍처)인 '알렉스넷'(Alexnet)이 나왔다. 알렉스넷은 엔비디아의 GPU 두 개로 훈련됐는데, 일반 프로세서를 사용하면 몇 달이 걸리는 과정을 단 며칠 만에 처리해 냈다. 이후 컴퓨터 과학자 사이에 GPU가 신경망 처리를 대폭 가속화할 수 있다는 소문이 퍼졌고 GPU에 대한 수요가 증가했다. 이후 엔비디아는 AI에 더 적합한 새로운 GPU를 개발하고 기술을 쉽게 사용할 수 있는 소프트웨어를 개발했다. 생성 AI 시대에 미리 대비한 것이다. 엔비디아는 11월 13일 생성 AI 모델의 기반이 되는 LLM을 훈련하도록 설계한 GPU 'H200'을 공개했다. H200은 오픈AI의 GPT-4 훈련에 적용되기도 하였으며, 세계 기업들이 확보하기 위해 경쟁을 벌이는 H100의 업그레이드 버전이다. 현재 H100 칩 1개당 가격은 2만 5천~4만 달러로 추정된다. LLM을 구동하는 데에는 수천 개의 칩이 필요하다. H200의 가격은 아직 알려지지 않았다.

H200에는 141기가바이트(GB)의 차세대 메모리 반도체 'HBM3'가 들어갔다. 고대역폭 메모리를 뜻하는 HBM(High Bandwidth Memory)은 여러 개의 D램을 수직으로 연결해 데이터 처리 속도를 혁신적으로 끌어올린 고성능 제품이다. 엔비디아는 신제품 H200이 H100보다 2배 빠른 출력을 낸다고 설명했다.

H200은 H100과 호환된다. 때문에 기존에 H100을 확보한 AI 기업이 이후 H200을 구매하여 사용하더라도 이를 위해 서버 시스템이나 소프트웨어를 바꿀 필요가 없다.



> 이미지를 분류할 수 있는 AI 시스템인 알렉스넷은 엔비디아 GPU 두 개로 훈련됐는데, 일반 프로세서를 사용하면 몇 달이 걸리는 과정을 단 며칠 만에 처리해 냈다.

엔비디아의 독주를 막기 위한 AI 칩 개발 레이스

이처럼 엔비디아는 GPU시장을 독점하면서도 끊임없이 혁신을 이어가고 있다. 그리고 이에 대항해 빅테크가 AI 칩 시장에 도전장을 내밀었다. 엔비디아의 가격 정책에 휘둘리는 GPU 시장과 공급 부족이라는 문제가 수면 위로 떠오르면서, GPU를 대체할 대항마를 찾고자 하는 것이다. 물론 AI 칩 시장 자체가 도전을 불러일으키는 비교적 새롭고 치열한 장(場)이기도 하다.

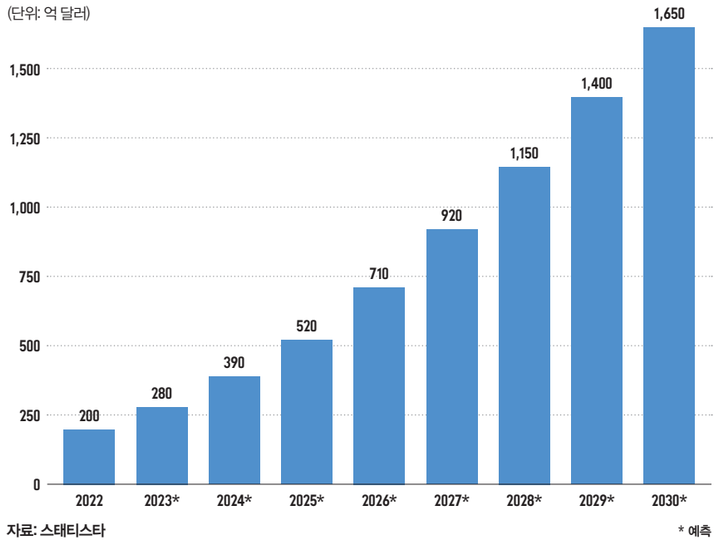
AI 반도체는 다양한 일을 수행하지 않는다. 하나의 하드웨어를 사서 다양한 일을 하고 싶으면 GPU가 더 효율적이다. 하지만 지금과 같은 추세로 AI 트래픽이 폭발하면 AI 반도체만 서비스하는 인프라가 규모의 경제를 형성할 것이다. 이 때문에 AI 반도체 스타트업들이 생기고 있다.

글로벌 시장조사 기업 스태티스타(Statista)에 따르면 2022년 AI 칩 시장 규모는 200억 달러였다. AI 칩 시장은 연평균성장률(CAGR)이 약 30.3%에 달해 2030년까지 1,650억 달러로 증가할 전망이다.

엔비디아의 GPU는 챗GPT나 메타(Meta)의 '라마-2'(Llama-2) 와 같은 대규모 AI 모델을 훈련하는 데 필수적이지만 추론하는 기능을 위해 설계된 반도체는 아니다. 이제 훈련된 AI 모델이 최적화될 수 있도록 칩의 효율성이 더욱 중요해지는 시점이다.

구글과 아마존 등이 대규모 언어 모델 교육을 완료하고 이를 음성 도우미와 같은 제품에 통합할 준비가 되면, 추론 기능 중심의 칩에 대한 수요가 급증할 수 있다. 그때는 AI 칩이 인공지능/기계학습 분야에서 GPU를 넘어 가장 중요한 반도체가 될 수 있다. 과연 AI 칩이 대세로 떠올라 엔비디아의 GPU 독주를 막을 수 있을지 귀추가 주목된다.

글로벌 AI 칩 시장 규모 예측



자료: 스태티스타 * 예측