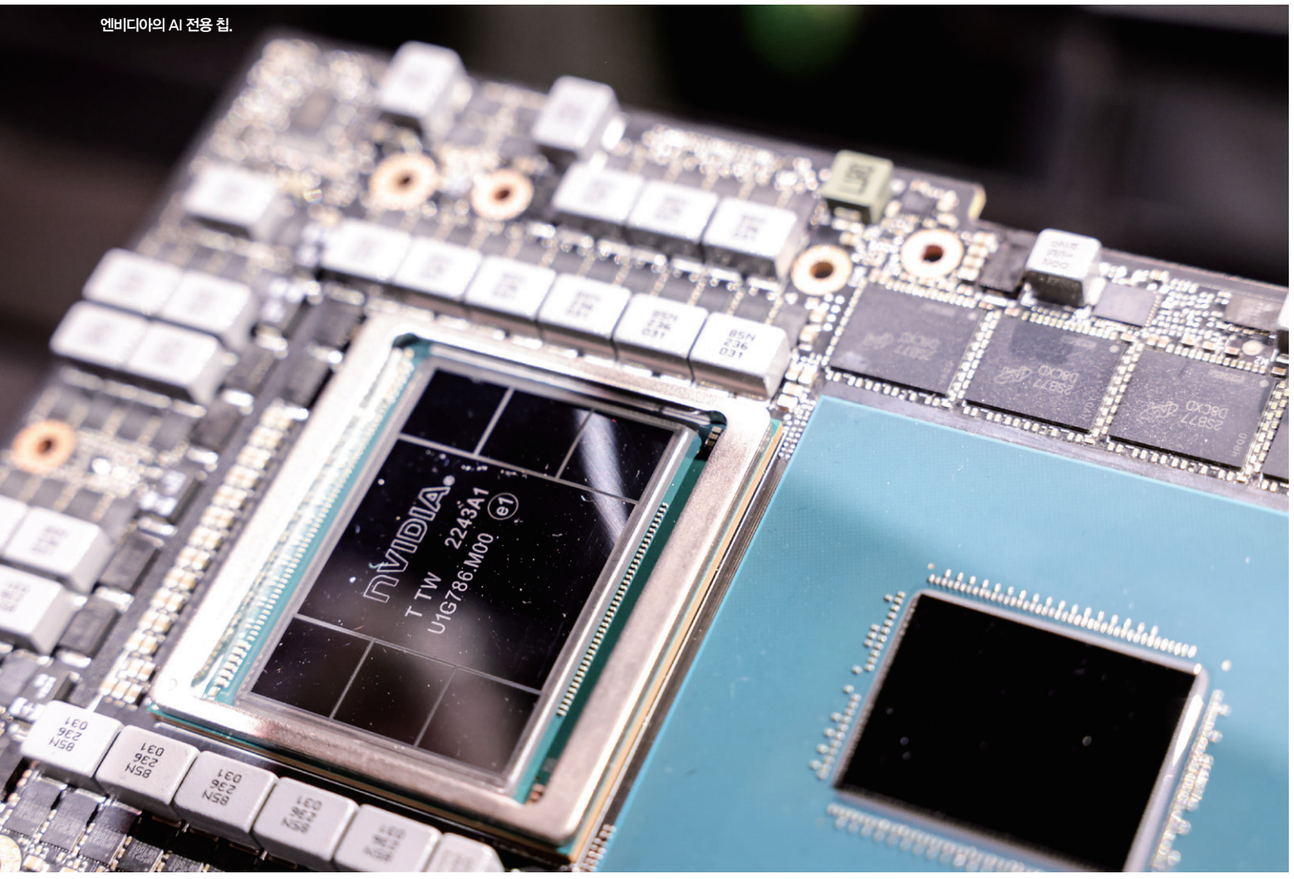
AI 반도체 시장에 뛰어든 빅테크와 스타트업

어떤 산업도 AI와 무관할 수 없다.
어떤 테크 기업도 AI 반도체를 외면할 수 없다.

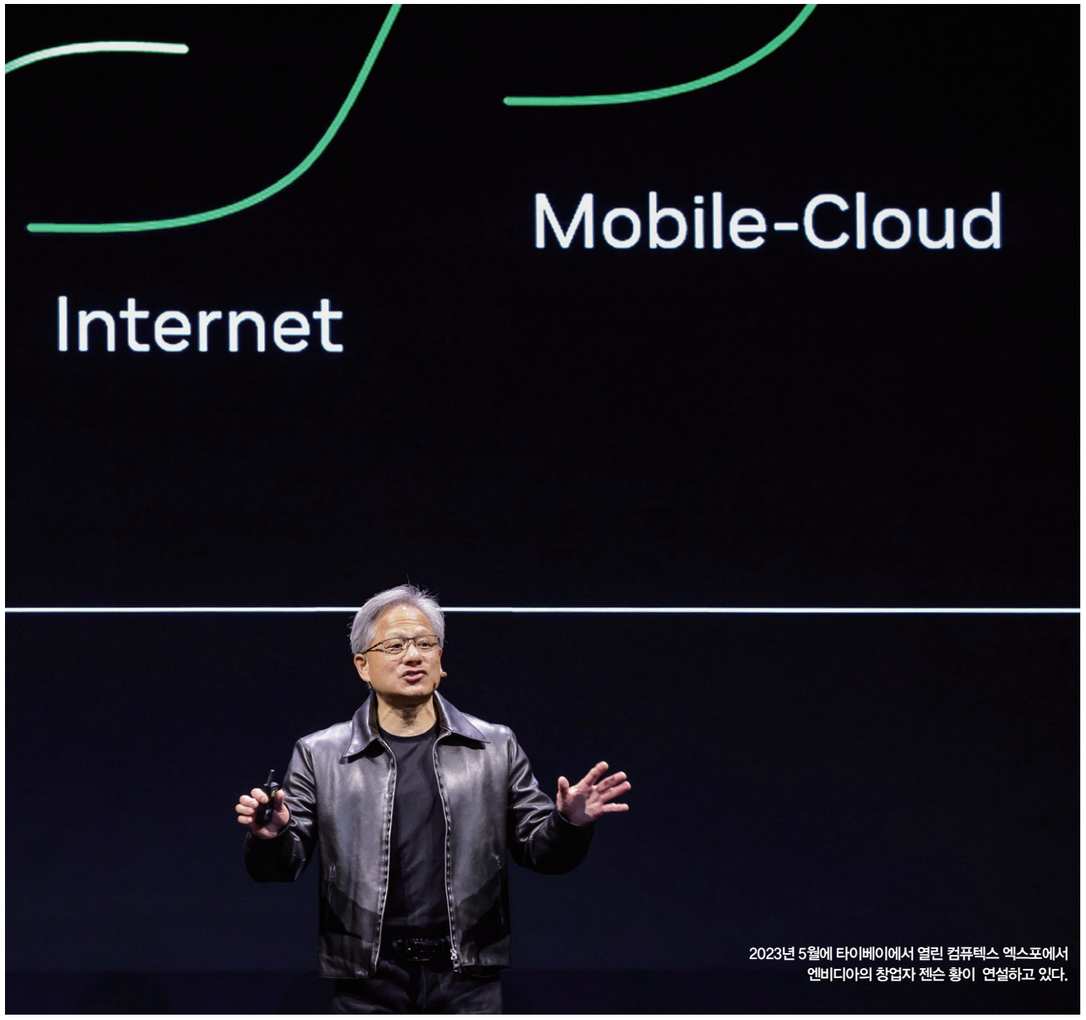
글 **김인순** 사진 Getty Images

마이크로소프트는 소프트웨어 기업으로 알려졌지만 실제로는 시스템 회사다. 그리고 시스템 회사의 마지막 퍼즐을 완성하는 것은 반도체 설계다. 마이크로소프트는 11월 15일 연례 개발자 회의 '이그나이트'(Ignite)에서 맞춤 설계한 두 개의 칩을 공개했다. 소프트웨어부터 하드웨어까지 최적의 생성 AI 인프라 구축을 위해 탄생한 작품들이었다. 그리고 이로써 마이크로소프트는 생성 AI 인프라 최강자가 되는 마지막 단추를 끼웠다.

첫 번째 칩은 생성 AI에 최적화된 '마이크로소프트 애저 마이아(Azure Maia) AI 엑셀러레이터'다. 엔비디아와 직접적으로 경쟁하는 AI GPU이다. 두 번째 제품은 마이크로소프트 애저 코발트(Cobalt) CPU다. 이 칩은 클라우드에서 범용 컴퓨팅 워크로드를 실행하도록 맞춤화된 기반 프로세서다.



엔비디아의 AI 전용 칩.



2023년 5월에 타이베이에서 열린 컴퓨텍스 엑스포에서 엔비디아의 창업자 젠슨 황이 연설하고 있다.

> 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.

마이크로소프트, 생성 AI 인프라 구축의 마지막 단추를 끼우다

마이크로소프트는 2016년 이전에 클라우드 센터를 기성품 서버 등으로 구성했다. 이후 클라우드 서비스 수요가 증가하면서 자체 서버와 랙을 맞춤 제작해 비용을 절감하고 고객에게 일관된 경험을 제공하기 시작했다. 마이크로소프트가 자체 데이터센터용 서버와 랙을 만들면서 가장 아쉬웠던 부분이 반도체였다. 반도체 공급 부족 문제가 발생하면 데이터센터까지 영향을 받았다. 마이크로소프트는 이때부터 자체 칩 개발에 집중했다.

마이크로소프트는 반도체 개발도 섭렵하며 소프트웨어, 서버, 냉각시스템에 이르기까지 모든 인프라를 수직 계열화했다. 마이아 100 GPU는 2024년 초 마이크로소프트 데이터센터에 들어간다. 초기에는 마이크로소프트 코파일럿(Microsoft Copilot)이나 애저 오픈AI 서비스(Azure OpenAI Service)와 같은 서비스를 지원할 계획이다.

마이크로소프트는 전용칩 개발을 완성하면서 AI 혁신을 지원하는 인프라를 재구성할 계획이다. 반도체는 클라우드 서비스와 밀접하게 관련된다. 데이터센터는 사실상 1과 0의 디지털 흐름을 처리하는 수십억 개의 트랜지스터를 제어하는 곳이다. 마이크로소프트가 개발한 칩은 데이터 센터 내부에 쉽게 들어갈 수 있는 맞춤형 랙(Rack) 안 서버 보드에 부착된다. 마이크로소프트는 자체 데이터센터에 최적화된 칩으로 전력은 낮추고 성능과 지속가능성, 비용을 최적화하는 데 집중할 예정이다.

두 번째 칩인 코발트 100 CPU는 ARM 아키텍처를 기반으로 개발됐다. 클라우드 네이티브 제품에서 더 높은 효율성과 성능을 제공하도록 최적화했다.

마이크로소프트는 데이터센터 전체에서 '와트당 성능'을 최적화하려 한다. 이는 본질적으로 소비되는 각 에너지 단위에 대해 더 많은 컴퓨팅 성능을 얻는 것을 의미한다.

마이크로소프트는 오픈AI와 협력해 생성 AI를 다양한 제품에 내재화한 데 이어 반도체칩까지 개발하며 소프트웨어부터 하드웨어까지 서비스 최적화를 시도했다. 이로써 현재 데이터센터 자산 사용을 최적화하고 기존 설치 공간 내에서 서버 용량을 최대화하는 길을 열었다고 볼 수 있다.

> 마이크로소프트는 오픈AI와 협력해 생성 AI를 다양한 제품에 내재화한 데 이어 반도체칩까지 개발하며 소프트웨어부터 하드웨어까지 서비스 최적화를 시도했다.

구글의 TPU에 이어 아마존, 메타, 화웨이도 자체 칩 개발 나서

구글은 2013년부터 AI 칩 '텐서 프로세서 유닛'(TPU: Tensor Processing Unit) 성능 향상에 집중 중이다. 구글의 TPU는 기계학습을 위해 특별히 설계된 칩이다. TPU는 초당 수조 개의 작업을 처리한다. 기존 CPU나 GPU 보다 빠르다. 여기에 에너지 효율이 높고 전력 소모가 적다.

구글은 4월 자체 개발한 TPU v4를 약 4천여 개 집어넣은 AI 개발용 슈퍼컴퓨터 '팜'(PaLM)을 선보였다. 9월에도 기존 제품보다 AI 훈련·추론 성능이 좋아진 5세대 'TPU v5e'를 공개했다. 나아가, 구글은 마벨(Marvell Technology)과 함께 네트워크 인터페이스 칩 '그래닛 리덕스'(Granite Redux)를 출시할 계획이다.

구글 TPU는 텐서플로우(TensorFlow) 소프트웨어와 함께 사용되도록 설계됐다. 텐서플로우는 머신러닝을 위한 오픈소스 소프트웨어 라이브러리다. 구글은 이를 인공지능 모델을 훈련하는 데 사용한다. TPU를 사용하면 인공지능 모델을 빠르고 효율적으로 학습시킬 수 있다. 이는 의료연구나 자율주행차, 일기예보 등 다양한 분야에 적용할 수 있다.

아마존은 2015년 이스라엘 반도체 기업 안나푸르나 랩스(Annapurna Labs)를 인수하고 자체 AI 반도체인 '인퍼런시아'(Inferentia)와 '트레이니엄'(Trainium)을 내놓았다. 아마존의 데이터센터와 AI 음성인식 서비스에 적용 중이다. 엔비디아 칩 공급이 부족한 틈을 타 자사 클라우드 고객을 공략하기 위해서다. AWS(Amazon Web Services)는 엔비디아 칩의 큰손이었지만 엔비디아가 자체 클라우드 서비스를 개발하면서 회사 간 갈등도 생겼다.

메타도 지난 5월 'MTIA'라는 자체 개발 AI 반도체를 공개했다. 중국의 화웨이도 엔비디아 A100 수준의 AI 반도체를 개발한 것으로 알려졌다.

AI 칩 스타트업, 삼바노바부터 리벨리온까지

소프트뱅크와 인텔 등은 2017년 설립된 AI 칩 스타트업 삼바노바(Sambanova)에 10억 달러를 투자했다. 삼바노바는 9월 4세대 SN40L프로세서를 출시했다. 이 칩에는 1,020억 개 이상의 트랜지스터가 들어있다. 최대 638테라플롭스(TFlops) 컴퓨팅 속도를 제공한다. AI와 관련된 데이터 흐름을 효율적으로 처리할 수 있도록 새로운 3계층 메모리 시스템도 갖췄다. 삼바노바는 AI 모델 크기가 커지면서 데이터 이동이 성능에 영향을 미친다고 설명했다. 메모리가 AI 칩의 주요 차별요소가 된다는 뜻이다.

또 다른 스타트업 디-매트릭스(d-Matrix)는 GPT(Generative Pre-Trained Transformer) 배포를 지원하는 칩을 개발하고 있다. 디-매트릭스는 9월 시리즈 B 라운드에서 삼성, 미래에셋, 마이크로소프트와 미국의 스타트업 자문기업 플레이그라운드 글로벌(Playground Global) 등에서 1억 1천만 달러를 투자받았다. 앞서 SK하이닉스가 투자를 진행한 바도 있다.

2019년 설립된 디-매트릭스는 챗GPT와 같은 생성 AI 애플리케이션을 지원하는 최적화 칩을 설계한다. 특히 AI 컴퓨터 코드가 더 효율적으로 실행될 수 있도록 하드디스크가 아닌 메인 메모리를 활용하는 '디지털 인메모리 컴퓨팅'을 도입해 활용 중이다. 이 기술은 필요한 데이터를 처리할 때 에너지를 더 적게 사용토록 한다.

디-매트릭스는 올해 주로 평가용 칩을 구매하는 고객으로부터 1천만 달러의 매출을 올릴 것으로 예상한다. 2년 후에는 연간 7천만~7,500만 달러의 매출을 올릴 전망이다.

텐스토렌트(Tenstorrent)는 2016년 설립된 AI 칩 스타트업이다. 삼성과 현대 등이 3억 3,500만 달러를 투자했다. 기업가치는 10억 달러로 추산된다. 이 회사는 삼성전자와 생산 협력을 체결했다.

한국 스타트업 리벨리온(Rebellions)도 경쟁에 참여했다. 박성현 칩 설계 엔지니어가 지난 2020년 설립한 리벨리온은 포브스 선정 가장 빠르게 성장하고 있는 스타트업 중 하나다. 인공지능 모델 학습을 기반으로 모델을 통한 추론에 최적화된 시스템 반도체를 설계한다.

리벨리온은 데이터센터용 AI 반도체 '아톰'(Atom)을 개발했다. 아톰은 현재 KT의 AI 전략 핵심인 신경망 처리장치(NPU) 인프라를 담당하고 있다. 리벨리온은 창업 2년 만에 글로벌 수준의 AI 반도체를 출시했고, 지난해 620억원의 시리즈A 투자를 유치하며 누적 투자금 1,120억원을 달성했다.

챗GPT를 개발한 오픈AI도 자체 GPU 제작을 검토하고 있는 것으로 알려졌다. 오픈AI는 관련 제조사 인수를 위해 기업 실사 단계까지 진행한 것으로 전해진다. 물론 오픈AI가 인수를 포함, 맞춤형 칩에 대한 계획을 실제 진행하더라도 개발되기까지는 수년이 걸릴 가능성이 높다. 직접 제작을 추진하는지도 아직 결정되지 않았다. AI 칩은 높은 기술력이 필요해 연간 비용이 수억 달러에 이를 수 있는 막대한 투자가 필요하지만, 성공이 보장되지는 않는다. 마이크로소프트가 11월 15일 이그나이트에서 자체 칩을 발표한 상황에서 오픈AI까지 칩을 설계할지는 미지수다.