

언제·어디서나 AI 구현… 韓 ‘초격차’ 메모리 중요성 커져

반도체 업계가 인공지능(AI) 시장을 둘러싼 경쟁에 돌입했다. AI 가속기 시장 절대 강자인 엔비디아가 자리를 위협받는 가운데, 국내 메모리 업계는 ‘초격차’를 바탕으로 흔들림 없는 성장을 이어갈 것으로 기대된다.

21일 업계에 따르면 인텔은 미국에서 개발자 대상 인텔 이노베이션 2023을 열고 ‘실리콘노미’ 시대를 선언하며 AI 중요성을 강조했다.

실리콘노미는 경제가 반도체와 소프트웨어로 성장한다는 개념, 인텔은 AI가 실리콘노미를 견인할 수 있다며 개발자에 언제 어디서든 AI를 구현할 수 있도록 하겠다는 비전을 밝혔다.

인텔은 이를 위해 연말 출시할 5세대 제온 프로세서(그래파이트 래피즈)를 AI에 최적화한다는 방침, 차세대 PC용 CPU인 코어 울트라 프로세서(메테오레이크)에도 처음으로 NPU를 탑재할 예정이다.

특히 인텔은 AI 반도체 업계 독보적인 점유율을 이어가는 엔비디아를 정조준했다. 가우디2를 비롯한 인텔 솔루션에서 AI 활용을 극대화할 수 있도록 개발자 클라우드도 공식 출시하고, 리눅스 재단 산하 ‘UXL’ 재단에 참여해 하드웨어간 경계를 넘을 수 있도록 통합 병렬 프로그래밍 ‘원API’ 규격을 제공키로 했다. 엔비디아 쿠다 라이브러리에 대응해 자사 AI 반도체 활용도를 늘리는 것은 물론, 엔비디아 생태계도 포용하겠다는 얘기다.

반도체업계, AI시장 본격 진출 메모리 등 고성능 장비 수요 ↑

삼성·SK, HBM 점유율 양분
GDDR·CXL 등 기술력 뽐내

앞서 AMD도 지난 6월 AI 반도체 MI300을 공개하며 시장에 도전장을 내밀었다. 엔비디아 AI 반도체와 성능 우위를 강조하며 자신감을 내비치기도 했다. 그 밖에도 구글이 TPU를 개발하는 등 엔비디아를 대체할 AI 반도체 개발 경쟁은 치열해지고 있다.

◆ AI칩 경쟁 본격화

엔비디아 H100, 인텔 가우디, AMD MI300, 구글 TPU, 테슬라 D1, 최근들어 주가를 올리는 AI 반도체는 통상 AI 가속기를 가리킨다.

AI 가속기는 병렬 연산을 하는 칩으로, 그래픽 프로세서(GPU)와 거의 같은 구조로 만들어진다. 그래픽 카드가 일찌감치 인공지능(AI)을 활용한 업스케일링 기술을 발전시켜왔던 것도 이같은 이유에서다. 최근 AI 개발자에 주목받는 엔비디아 쿠다 라이브러리가 방대한 것도 오랜 기간 활용됐기 때문이다.

다만 GPU는 그래픽 처리에 초점을



엔비디아 H100.



SK하이닉스가 소개한 GDDR6 AiM 기반 AI 가속기

맞춘 탓에 AI 가속기에 불필요한 기능을 보유하고 있어 가격이나 효율성에서 불리하다는 지적을 받아왔다.

엔비디아는 연산에 집중한 텐서 코어를 활용해 AI 가속기 효율을 높이고 있다. 그래픽카드용 칩에는 쿠다 코어를, AI 가속기에는 텐서 코어 비중을 높이는 방식으로 차별화를 둔다.

엔비디아에 이은 GPU 시장 2위인 AMD도 같은 전략이다. GPU를 여러개 합치는 방식으로 대대적인 성능 향상을 이끌어냈다는 설명이다.

인텔도 뒤늦게나마 GPU 기술을 끌어올리며 데이터센터용 GPU도 꾸준히 선보이고 있지만, AI 전용 가속기인 가우디2는 연산에 중점을 둔 텐서 프로세서를 조합해 만들었다.

그 밖에도 구글 TPU와 테슬라 D1, 국내 팹리스인 리벨리온과 퓨리오사 AI 등 역시 신경망 프로세서인 NPU로 가속기를 만들고 있다.

일단은 GPU 가속기가 높은 범용성과 방대한 라이브러리를 앞세워 시장을 대부분 장악하고 있다. 엔비디아가 전체 80% 이상 점유율을 차지하는 것으로 알려져있으며, AMD도 시장 공략에 힘을 더하고 있다.

다만 GPU 가속기가 비싸면서 물량까지 부족해지면서 AI 전용 가속기에 대한 관심이 높아지고 있다. 특히나 인텔 가우디2가 엔비디아 H100에 비견할 만한 성능을 증명하고 있고, 엔비디아 가속기를 수입하지 못하는 중국에서도 적극 도입하면서 중국에서는 인텔이 추가 주문을 넣었다는 보도가 나오기도 했다.

◆ 메모리는 삼성·SK 양분

AI 가속기 핵심은 프로세서지만, 고성능 메모리를 붙이지 않으면 제 성능을 구현할 수 없다. 병렬 처리 방식이라 처리할 수 있는 데이터도 대폭 커진 만큼, 이를 따라잡을 수 있는 용량과 속도를 내야 한다. 때문에 AI 가속기 패키지는 대부분 AI칩과 메모리를 기본으로 포함하고 있다.

고대역폭 메모리, HBM 시장이 빠르게 성장한 것도 이 때문이다. HBM은 D램에 구멍을 뚫어 직접 연결해 합친 메모리다. AI 열풍으로 갑작스럽게 고성능 메모리 수요가 늘어난 상황, 미세 공정 한계로 기술 개발이 지연되면서 대안으로 제시됐다.

HBM 시장은 삼성전자와 SK하이닉스가 거의 대부분을 점유하고 있다. 업계에서는 양사 합쳐 90% 정도로 보고 있다. SK하이닉스가 처음 개발하고 삼성전자가 뒤따르며 기술을 발전시켜왔

다. 마이크론도 HBM을 만들고는 있지만, 후공정 기술은 물론 미세공정에서도 이미 뒤쳐진 탓에 점유율 확대가 쉽지 않을 전망이다.

문제는 HBM도 가격이 너무 비싸다는 단점이 있다. 칩을 붙이는 방식이라 공정이 복잡해 부피도 크고 발열 문제도 해결하기 쉽지 않다.

그래픽카드에 사용하던 고성능 메모리 GDDR 규격이 대안으로 떠오른다. 플래그십급 AI칩보다는 보급형에

탑재가 논의 중으로 알려졌다. 삼성전자가 최근 GDDR7을 세계 최초로 공개했다.

메모리에 프로세서를 더한 PIM 기술도 주목 받고 있다. 메모리 병목현상을 최소화하는 등 효율을 대폭 끌어올리는 방식으로 AI 가속기에 사용할만한 성능을 낼 수 있다는 것. SK하이닉스는 최근 미국에서 PIM 기술을 적용한 GDDR6 AiM을 기반에 주문형 반도체(A-SIC)를 조합해 만든 AI 가속기 ‘AiM



인텔 데이터센터 GPU 프로세서를 선보이는 그렉 라벤더 인텔 CTO

X’를 소개하고 성능까지 시연했다. 추후 SK사피온과 함께 새로운 제품을 개발할 계획임도 밝혔다.

더 빠른 인터페이스, CXL(컴퓨터 익스프레스 링크) 규격 메모리도 AI 시대를 겨냥한 노력이다. 삼성전자와 SK하이닉스 모두 CXL 기반 메모리 개발을 끝마친 상태다. 국내 팹리스 기업 파네시아가 CXL 기반 AI 가속기를 처음으로 만들어 주목받기도 했다.

/김재웅 기자 juk@metroseoul.co.kr

