LG경영연구원

2024.12.

AI 혁신의 게임 체인저, AI 반도체

김은수 · 김종대
euncarp2@lgbr.co.kr · jdkim@lgbr.co.kr

인공지능(AI)은 인간 수준의 능력을 향한 준비 단계에 들어섰다. Open AI의 GPT-o1은 추론 시간을 늘려 인간의 사고를 모방하는 새로운 패러다임을 제시했다. AI의 발전은 이제 고성능 학습과 고효율 추론이라는 두 축을 중심으로 재편되고 있으며, 이 과정에서 AI 반도체의 역할은 그 어느 때보다 중요해졌다.

학습용 반도체는 칩 결합과 고속 통신 기술로, 추론용 반도체는 맞춤 설계와 오픈소스 기술로 고성능화와 비용 절감을 동시에 실현하고 있다. 이러한 혁신은 금융, 의료, 보험 등 보수적인 산업의 AI 도입을 가속할 뿐만 아니라 휴머노이드의 실용화 가능성까지 제시한다. AI 추론 비용 구조의 혁신은 시장 접근성을 넓히는 동시에 무료 서비스 기반 성능 개선, 다단계 심층 추론 통한 신뢰성 확보라는 새로운 방향성을 열어가고 있다.

AI 시대가 본격화하는 지금, AI 반도체 기술 발전 및 투자, 시장 변화에 주목할 필요가 있다. 이를 통해 새롭게 떠오르는 전후방 산업에서 기회를 선점할 수 있을 것이다.

"기술이 우리를 석기 시대에서 산업 시대로 이끌었듯이, 이제는 컴퓨팅, 에너지, 인간의 의지가 지능 시대를 향한 길을 열고 있다."

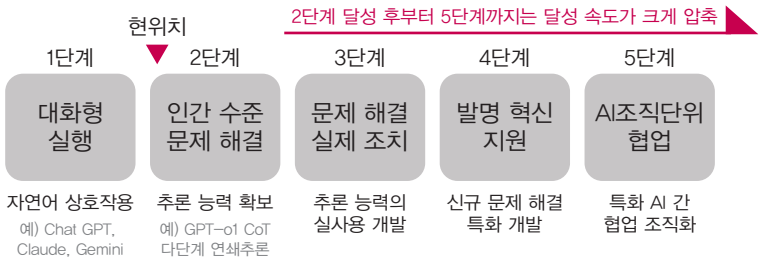
AI 혁명의 최전선에 있다는 평가를 받는 Open AI의 샘 알트만 CEO는 지난 9월 GPT-o1 공개 기자회견에서 현 시점을 '지능 시대(The Intelligence Age)'[1]라 명명했다. 그는 수천 일 내 인간 수준의 인공지능(AGI)이 일상화될 것이며, 더 많은 AI 반도체[2]가 필요할 것이라 예견했다.

AI 혁신, AGI 준비 단계 진입

Chat GPT로 시작된 AI 혁신은 새로운 국면을 맞았다. AI가 단순 자연어 소통을 넘어 인간 수준의 추론 능력 확보에 도전하게 된 것이다. 9월 공개된 GPT-o1은 다단계 연쇄 추론(CoT, Chain-of-Thought) 방식을 통해 '이 문제를 풀려면 어떤 과정을 거쳐야 할까, 적용할 만한 방법이 있을까'와 같은 논리적 사고 과정을 모방한다. 복잡한 질문 하나를 여러 하위 문제로 나눠 순차적으로 추론하고, 각 단계의 결과를 종합해 정확도를 높인다. 이는 추론 시간을 늘려 AI 모델의 성능을 획기적으로 향상시키는 새로운 패러다임이다.

엔비디아는 이러한 변화를 포착해 한 달 뒤인 10월 투자자 발표에서 AI 발전이 두 가지 핵심 축으로 재편될 것이라고 내다봤다. 한 축은 대형 모델, 멀티 모달리티(Modality), 고품질 합성 데이터를 활용한 고성능 학습이다. 다른 한 축은 다양한 답변 후보 생성과 검증을 위한 고효율 추론이다[3].

AGI로의 다섯 단계 중 가장 소모적인 도약 구간에 진입



현재 AI는 추론 능력 확보 단계(2단계) 진입. AGI 구현 위한 다섯 단계 중 가장 오래 걸리고 소모적이나, 2단계 달성 후부터는 각 단계 달성 속도가 크게 압축될 것이라고 OpenAI는 전망

AI는 초거대 모델의 매개변수 증가와 더불어, 추론 과정에도 추가적 계산과 정보 검색을 수행하는 새로운 단계로 진화하고 있다. 학습에 이어 추론에서도 컴퓨팅의 중요성이

1 AI가 인류의 문제해결 능력과 창조적 잠재력을 획기적으로 증폭시키는 시대. 인간 수준의 지능과 자율학습 능력을 갖춘 AI가 개인 비서로서 복잡한 과제를 해결하고, 전문 AI 팀이 과학 · 의료 등 전 분야 혁신을 주도하는 시기

2 AI 모델 학습과 추론에 특화된 칩으로서, AI의 인공신경망 연산을 가속할 수 있도록 설계된 반도체를 지칭

3 앤트로픽의 Andy L. Jones는 작은 모델의 15회 추론이 10배 큰 모델의 1회 추론과 동등한 성능을 보임을 입증하며, AI 성능 향상을 위한 학습과 추론 간의 효율적 자원 배분 가능성을 제시

커지면서 AI 반도체가 결정적인 역할을 할 것으로 예상된다. 인간 수준 인공지능 시대로의 진입을 이끄는 AI 반도체의 발전 방향과 이에 따른 사용자 경험의 혁신, 그리고 향후 전망을 살펴보자.

학습용 반도체, 칩 결합이나 고속통신으로 고성능화

AI 모델의 고도화로 파라미터 수는 1조 개를 돌파했고, 학습 데이터도 텍스트에서 음성, 이미지, 동영상으로 다변화하며 연산량이 급증하고 있다. OpenAI, Google DeepMind, Meta 등 AI 기업의 컴퓨팅 사용량은 2010년 이후 연간 4~5배씩 증가했으며, 모델의 기본 학습 기간[4]도 매년 1.2배씩 늘어나는 추세다. 더욱이 인간 수준 AI를 개발하기 위해 동시에 여러 개의 모델을 학습시켜 융합하거나, 대규모의 강화 학습을 도입하는 등 학습 방법론이 고도화됨에 따라 그 기간이 더욱 늘어날 전망이다. 이에 AI 반도체 업계는 고성능화를 통한 학습 기간 단축을 핵심 과제로 삼았는데, 크게 두 가지 방향에서 접근하고 있다.

첫째는 여러 개의 작은 칩을 수평 또는 수직으로 결합해 단일 칩의 한계를 뛰어넘는 칩렛(Chiplet) 방식이다. 주로 엔비디아, 인텔, AMD 등 선도 기업에서 고성능 칩 생산의 집적도 한계를 극복하려 적극 도입하고 있다. 다만, 고난도의 패키징 기술이 요구되며, 칩 간 데이터 통신이 성능의 병목이 될 수 있다는 과제도 있다. 엔비디아의 최근 행보는 칩렛 기술의 잠재력을 잘 보여준다. 지난 3월 공개된 블랙웰 아키텍처의 B200은 엔비디아 최초로 멀티 GPU를 칩렛 방식으로 구현했다. 2개의 GPU와 8개의 HBM3E[5] 고대역폭 메모리를 초당 10TB의 데이터 전송이 가능한 인터페이스로 결합해 단일 칩처럼 작동하게 만들었다. 트랜지스터 수는 2,080억 개로 H100 대비 2.6배 증가했고, 학습 속도는 4배 향상됐다.

다른 접근방식으로는, 반도체 간 통신 병목을 없애기 위해 웨이퍼 레벨 칩(Wafer-Level Chip)을 만드는 것이다. 웨이퍼를 잘라 개별 칩으로 만들고 결합하는 것이 아닌 웨이퍼 전체를 하나의 거대한 칩으로 제작한다. 칩 간 연결이 불필요해 패키징이 용이하고 데이터 병목 현상도 쉽게 해결된다. 다만, 웨이퍼의 어느 한 부분에서라도 불량이 발생하면 전체를 사용할 수 없어 수율 향상과 생산 비용 절감이 과제다. 미국 스타트업인 세레브라스(Cerebras)는 Wafer Scale Engine(WSE)이라는 웨이퍼 한 장 크기의 세계 최대 단일 칩을 생산한다. 지난 3월 공개된 WSE-3는 4조 개의 트랜지스터를 탑재해 엔비디

4 기초 모델(foundation model)의 사전학습 기간. 이후 미세조정(fine-tuning)에 소요되는 시간은 제외
5 최신 고대역폭 메모리(High Bandwidth Memory) 제품으로 초당 최대 1,280GB의 대역폭을 제공

아 B200 28개에 맞먹는 학습 성능을 보였다. WSE-3의 핵심 경쟁력은 별도 HBM 없이 칩 내에 메모리를 통합해 데이터 처리 효율성을 극대화한 데에 있다. WSE-3의 메모리는 B200 대비 2,000배의 대역폭을 제공하며, 웨이퍼상 배선을 통해 데이터 통신 지연을 근본적으로 감소시켰다. 세레브라스는 메타의 Llama 2(70B) 규모 모델의 학습 기간을 평균 한 달에서 하루로 단축할 수 있다고 발표했다.[6]

추론용 반도체, 맞춤화나 오픈소스 설계로 저원가 구현

학습용 AI 반도체는 학습해야 하는 데이터가 다양하고 모델도 빨리 변해 AI 반도체 역시 범용성이 중시되지만, 추론은 이미 학습된 모델이 있어 최적화한 반도체를 활용할 경우 성능과 비용 개선을 기대할 수 있다. 추론용 AI 반도체는 효율화를 위해 범용적인 GPU보다 AI에 특화된 NPU(Neural Processing Unit)가 주로 쓰이고 있다. 추론에 필수적인 행렬 연산에 최적인 하드웨어 구조로 돼 있어서, 추론용 AI 연산 시 GPU보다 훨씬 빠르고 소모하는 전력도 매우 적다.

NPU 개발에 가장 적극적인 기업은 AI 서비스나 서비스 사업자를 위한 클라우드 사업을 하는 빅테크들이다. 구글은 2015년 자체 NPU 칩인 TPU(Tensor Processing Unit)를 통해 실시간 음성 검색, 사진 객체 인식, 번역 서비스에 적용하는 등 NPU 개발을 선도해 왔다. 2020년 이후부터는 아마존도 자체적으로 추론용 NPU 칩을 개발하고 있다. 최신 칩 인퍼런시아2의 경우 엔비디아 GPU로 구축한 서버[7] 대비 AI 연산 처리량을 2.6배, 응답 속도는 8.1배, 와트당 성능은 1.5배만큼 높였다. 그 결과, 아마존 클라우드 서비스(AWS) 고객사들은 동일 성능 기준, 추론 비용을 50~80% 가량 절감하고, 응답 대기시간을 30%~90% 단

6 Llama 2 발표 당시 메타가 보유한 GPU 클러스터(엔비디아 GPU A100 16,000개로 구성)와 세레브라스의 CS-3 클러스터(WSE-3 2048개로 구성)를 비교

7 AWS의 EC2 G5 서버로, 엔비디아 GPU 기반 크기가 가장 큰 서버 (클라우드 전용칩 A10G GPU 8로 구축)

축할 수 있었다. 최근에는 마이크로소프트의 마이아100, 메타의 MTIA 등 그동안 NPU 개발에 소극적이었던 주요 빅테크들도 대부분 자체 개발에 뛰어들고 있다.

한편, 스타트업들도 추론용 AI 반도체 개발에 뛰어들고 있는데 에치드(Etched)는 NPU에서 한 단계 더 들어가, 트랜스포머[8]라는 한 종류의 AI 모델에만 특화한 NPU인 Sohu칩을 개발했다. 하나에 집중한 만큼 연산 효율을 극단적으로 높일 수 있어서, 엔비디아 GPU H100이 수행하는 연산작업을 약 20분의 1 가격으로 저렴하게 구현했다. 트랜스포머 계열이 아닌 AI 모델에는 활용하기 어려운 극단적인 효율화 방식이지만, 현존하는 생성형 AI 모델이 대부분 트랜스포머 계열이라는 점을 고려할 경우, 유의미한 접근이라고 할 수 있다.

반도체 산업의 전설로 알려진 짐 켈러가 주도하는 스타트업인 텐스토렌트(Tenstorrent)도 주목할 필요가 있다. 짐 켈러는 차별적 설계 기술을 통해 고대역 메모리(HBM)를 삭제하고 오픈 소스(RISC-V[9])를 활용해 가성비 높은 AI 반도체 개발을 추진하고 있다. 고대역 메모리 대신 범용 메모리를 쓸 경우 30% 수준의 원가절감이 가능하며, 2010년 UC버클리 연구진에 의해 개발된 RISC-V는 라이선스 비용이 무료라는 점을 고려할 때 비용을 크게 절감할 수 있는 것으로 보인다.

NPU 및 특화 NPU와 같은 하드웨어 맞춤화, 오픈 소스(RISC-V), 차별적 반도체 설계 시도 등 다양한 AI 반도체 혁신이 병렬적으로 진행되면서 향후 추론 비용은 획기적으로 낮아질 수 있다. 자산운용사 ARK Invest에 따르면, 2030년 추론에 필요한 연산비용은 2022년 수준의 1만 5천분의 1 수준으로 떨어질 것으로 예상된다.

8 자연어 처리 분야 혁신의 근간이 되는 모델. 문장의 이해와 생성에 효과적이며, 대규모 비정형 데이터의 병렬처리가 가능해 자연어처리, 번역, 이미지 인식 등 다양한 분야에 적용 중

9 RISC-V는 자주 사용되는 명령어만 탑재하고, 명령어 조합 설계를 최적화해 메모리 요구도를 낮추고 연산을 효율적으로 수행

AI, 인간과 닮아갈수록 커지는 시장

구글 딥마인드 'AlphaProteo'는 노벨 화학상 수상작 'AlphaFold'의 단백질 구조 예측을 넘어 구조 설계 영역까지 진화했다. 특히 최적 물질 후보를 찾는 정확도가 연구원의 수작업 대비 3배에서 최대 300배에 달한다고 한다. AI 반도체 혁신이 진행되면 AI 기반 고객 경험 혁신은 더욱 깊어지고 다양한 산업으로 퍼질 수 있어 금융, 의료, 보험 등 AI 도입에 신중했던 산업에서도 AI 적용이 확대될 것이다. 더불어 인식과 행동 제어 기술 부족으로 실험실에만 머물던 디바이스들은 AI 반도체 혁신으로 상용화로 이어질 수 있는 중대한 전환점을 맞이할 수 있다.

휴머노이드를 예로 들어보자. 과거의 휴머노이드는 카메라와 비전 AI를 통해 주변 물체의 형상은 확인할 수 있었지만, 각 물체의 특성과 쓰임새 등을 이해하기는 어려웠다. 공장 바닥에 떨어져 있는 물체를 밟아도 되는지, 피해 가야 하는지, 집어서 사용해야 하는 부품인지, 버려야 하는 쓰레기인지 등을 판단할 수 없었다. 그러나 최근의 휴머노이드는 LLM(Large Language Model)과 비전 AI가 결합된 형태의 멀티모달 AI인 VLM(Visual Language Model)을 적용하여, 마치 ChatGPT에 물어보고 답변을 받듯이 수많은 물체의 특성과 쓰임새 등을 쉽게 이해할 수 있게 됐다. 제어 역시 많은 변화가 진행 중이다. 지금까지는 대부분 사람이 수작업으로 작성한 프로그래밍 코드에 의해 이루어졌다. AI를 활용할 경우, 대량 연산에 따른 지연 시간 발생으로 제어가 불가능했기 때문이다. 그러나 엔비디아의 '젯슨 토르' 등 휴머노이드용 고성능 AI 칩셋이 등장하면서, 제어에도 AI를 활용하려는 시도가 늘어나고 있다. 주변 환경에 대한 센싱 데이터와 관절의 액츄에이터 제어 신호 데이터까지 함께 학습해서 사람과 같이 유연한 동작이 가능해지면, 휴머노이드는 조만간 어떤 돌발 상황에도 안정적으로 대처할 수 있게 될 것이다.

추론 비용 절감은 적용 시장을 넓힘과 동시에 새로운 혁신으로 이어질 것이다. 올해 AI 애플리케이션 및 솔루션 제공 기업이 전년 대비 42% 급증했는데, 업무용 문서 요약부터

지난 10월 열린 테슬라의 'We Robot' 행사, 복잡한 인파 속에서도 유연하게 활동하는 3세대 옵티머스
출처: 테슬라 공식 유튜브

게임, 엔터테인먼트, 영상 및 음성 콘텐츠 편집까지 AI 적용 영역이 확대된 결과다. 이는 지난해 말부터 시작된 AI 추론 비용의 대폭 하락(약 1년간 최대 100분의 1 수준)에 기인한다. 추론 비용 절감으로 고객에게 제공할 수 있는 무료 이용 범위가 넓어지면서 성능 개선의 새로운 가능성이 열리고 있다는 점도 주목할 필요가 있다. '데이터 플라이휠[10]' 방식으로 사용자 활용 행태를 수집해 추론 성능을 개선할 수 있게 됐으며, 이는 늘어난 사용자들의 프롬프트를 기반으로 정확한 답변을 찾는 추론 과정을 지속적으로 고도화할 수 있음을 뜻한다. 더불어 'GPT-o1', 'AlphaProteo'에서 단일 작업을 여러 단계로 분리해 추론하고 검증하는 방식이 등장해 AI 신뢰도가 향상되면서, 의료계 AI 활용 가능성도 주목받고 있다.

AI 반도체 혁신이 가져올 사업기회들

AI 발전은 AI 반도체 기술의 진보와 맥을 같이해 왔다. 가속 컴퓨팅이라는 새로운 아키텍처를 시작으로, 엔비디아의 A100, H100, B200 등 AI 반도체의 연이은 출시는 지난 10년간 AI 연산 성능을 100만배 향상시키며 현재의 AI 붐을 이끌었다. 하지만 GPU 수급난 등 반도체 산업의 병목현상이 AI 산업 전반의 발전을 지연시키고 있었다는 점을 유념할 필요가 있다. AI를 활용한 고객경험 및 생산성 개선은 AI 반도체 혁신과 같은 속도로 진행될 수 있으므로 반도체 혁신과 연계해 그 미래 잠재력을 살펴봐야 할 것이다.

AI 반도체의 미래 변화를 읽으려면 관련 스타트업들의 움직임을 민감하게 살필 필요가 있다. 이들은 대규모 자본을 투입할 수 없기 때문에 GPU 중심의 접근을 지양하고, 오히려 그들의 약점인 저전력 고성능, 온디바이스 AI 등 신시장을 개척하고 있다. 향후 시장 경쟁력은 AI 반도체의 발전 궤적을 정확히 예측하고, 이에 기반한 혁신적 서비스를 선제적으로 개발하는 기업이 주도할 것이다. 주요 기업과 스타트업의 AI 반도체 개발 동향을 면밀하게 모니터링하면서 협력을 통해 역량을 확보하는 방안을 모색할 필요가 있다.

또한 AI 산업 가치사슬에서 파생되는 사업 기회도 무시할 수 없다. AI 반도체를 많이 소비하는 대형 클라우드 기업에게 대규모 데이터센터와 서버를 운영하는 데 에너지 및 열관리가 핵심 이슈다. 유리 기판 및 GAN 전력소자 등 고내열 · 고효율의 소재, 액침 냉각 기술, 데이터센터 특화 HVAC 시스템 등 과거 반도체 가치사슬에서 존재하던 솔루션들이 AI 시대를 맞아 새롭게 조명받고 있다. 이런 기회에도 주목하면서 사업 연계 가능성을 저울질해봐야 한다. LG경영연구원

10 사용자들이 입력한 프롬프트와 AI 모델의 출력 결과 등 이용 데이터를 확보해 더 나은 알고리즘을 설계하고, AI 성능 개선을 가속하는 방식