5장

생성형 AI 다음 단계 성패는 연산력이 좌우

인공지능(AI) 연산의 무게중심이 학습에서 추론으로 이동하면서, AI 경쟁의 핵심은 알고리즘 효율이 아니라 대규모 추론을 감당할 수 있는 데이터센터, 고성능 칩, 전력 인프라 역량으로 재편되고 있다. 엣지 AI의 확산 가능성에도 불구하고, 단기적으로는 데이터센터 중심의 고성능 추론 구조와 고대역폭 메모리(HBM)가 결합된 그래픽처리장치(GPU) 체계가 주류를 이루며, 온프레미스·하이브리드 인프라 수요도 확대되고 있다. 로봇·드론·자율주행차 등 일부 영역에서는 엣지 AI가 필수적이지만, 2030년까지 급증하는 AI 연산 수요를 고려할 때 당분간은 인프라 투자 역량이 AI 경쟁력의 전제 조건으로 유지될 전망이다.

핵심 내용 요약 (Executive Summary)

AI 연산의 무게중심 이동: 학습에서 추론으로

- AI 컴퓨팅의 무게중심이 모델 학습(training)에서 대규모 추론(inference)으로 이동
- 2026년 전체 AI 연산의 약 3분의 2가 추론에 사용될 전망
- AI 활용이 연구·개발 단계를 넘어 일상적·상시적 서비스로 확산되고 있음을 의미

인프라의 중심은 여전히 데이터센터

- 엣지 AI 확대 기대에도 불구하고, 고성능 추론은 데이터센터·엔터프라이즈 서버가 주도
- 데이터센터 CAPEX는 2026년 4,000~4,500억 달러, 2028년 1조 달러까지 확대 전망
- 대규모·집중형 인프라가 AI 추론의 주 무대가 되는 구조가 고착

AI 칩과 연산 구조의 이중화

- 추론 최적화 칩 시장이 2026년 500억 달러 이상으로 성장하나, HBM 결합 고성능 GPU 구조는 병행 유지
- 사후 학습(post-training), 장시간 사고(long thinking) 등 고도화된 기법이 연산·전력 수요를 추가 확대
- 효율화와 고성능화가 동시에 요구되는 복합적 칩 전략이 불가피

경쟁의 본질 변화: 알고리즘에서 인프라로

- AI 경쟁의 핵심은 알고리즘 효율보다 연산 인프라, 전력 확보, 반도체 공급망을 감당할 수 있는 역량으로 이동
- 에너지·환경 부담, 보안·데이터 주권 이슈가 AI 전략의 핵심 변수로 부상
- 이에 따라 온프레미스·하이브리드 AI 인프라 도입이 확대되며, 2026년 시장 규모는 500억 달러 이상 전망

생성형 AI 컴퓨팅의 주 목적이 학습에서 추론으로 이동하고 있다. AI 연산 수요도 급증할 전망이다. 이로 인해 소비자 단말기 중심의 엣지 컴퓨팅이 확대되고 데이터센터 의존도는 감소할 것이라는 예측이 다수지만, 2026년 현재 상황은 다르게 전개되고 있다.

2026년에는 생성형 AI 컴퓨팅의 주 목적이 크게 바뀔 것으로 예상된다. 지금까지는 막대한 데이터를 학습시키는 훈련(training)이 AI 컴퓨팅의 주된 목적이었다. 하지만 앞으로는 기업과 소비자의 질문 및 프롬프트, 업무 요청에 대해 추론(inference)하고, 해석하고, 답변하는 기능이 AI 연산 수요의 핵심이 될 것이다. 상당수 전문가는 이러한 컴퓨팅 부하의 변화로 곧 추론에 최적화된 전용 칩의 필요성이 커질 것으로 본다. 이들 전문가들은 훨씬 저렴해진 칩이 엣지 디바이스에 배치되면서 결과적으로 대규모 데이터센터의 필요성이 줄거나 다른 작은 형태로 재편될 수 있으며, 비용도 감소할 수 있다고 예측한다.

그러나 딜로이트의 예측은 다소 다르다. 추론 워크로드는 2026년에 분명히 가장 빠르게 성장하는 영역으로, 전체 컴퓨팅의 약 3분의 2를 차지할 것으로 예상된다(2023년 약 3분의 1, 2025년 절반 수준에서 증가).[1] 또한 2026년 추론 최적화 반도체칩 시장 규모는 500억 달러를 넘을 전망이다. 그럼에도 불구하고, 딜로이트는 전체 연산의 대다수가 여전히 최첨단·고가·고전력 AI 칩(시장 규모 2,000억 달러 이상)에 의해 수행될 것으로 전망한다. 이러한 칩은 주로 4,000억 달러 이상 규모의 대형 데이터센터 또는 대형 데이터센터급 칩과 랙을 사용하는 500억 달러 규모의 온프레미스(on-prem) 엔터프라이즈 AI 솔루션에 설치될 것이다. 엣지 디바이스용 소형 칩이 이를 대체할 가능성은 적다. 즉, 앞으로도 계속 증가하는 데이터센터와 엔터프라이즈 온프레미스 AI 팩토리에 대한 수요는 차고 넘칠 것이고, 이들 시설이 막대한 전력을 소비하게 될 것이다.

끊임없이 증가하는 AI의 연산 수요

신규 AI 모델을 위한 훈련용 컴퓨팅 수요 증가세는 2023~2024년에 비해 둔화됐지만,[2] AI 모델은 훈련 이후에도 성능을 향상시키는 고도화 기법을 통해 지속적인 재훈련이 필요하다. 이러한 재훈련과 더불어 막대한 규모의 추론 요청량이 더해지면서, 전체적인 컴퓨팅 수요는 감소하기는커녕 계속 증가할 가능성이 크다. 달리 말하면, 무어의 법칙(Moore's Law) 덕분에 연산에 사용되는 칩의 효율이 해마다 높아지고 있음에도 불구하고, 컴퓨팅 수요는 2030년까지 매년 4~5배 증가할 것으로 예상된다.[3]

1. 초기 학습을 위한 컴퓨팅 수요 증가세는 둔화

2020년 발표된 연구에 따르면, 더 많은 데이터로 학습하고 더 발전된 AI 가속기 칩을 사용한 더 큰 모델일수록 지속적으로 더욱 뛰어난 성능을 보여주는 것으로 나타났다. 이것이 생성형 AI의 첫 번째 스케일링 법칙(scaling law)이었다.[4] 2022~2023년 사이 학습 모델은 10억 파라미터(parameter, 매개 변수)에서 1,000억을 넘어 1조 파라미터까지 성장했다.[5]

2024년에는 두 가지 문제가 부각되기 시작했다. 첫째, 세상에 존재하는 학습 데이터는 무한하지 않다는 것이다. 둘째, 모델을 계속 키울수록 성능 향상이 점점 줄어드는 한계효용 감소 현상이 나타났다는 점이다. 학습 데이터를 10배로 늘려도 최첨단 AI 모델이 기존 버전보다 약간 개선되는 데 그치거나, 경우에 따라서는 전혀 개선되지 않는 경우도 있었다.[6] 동시에, 훨씬 적은 데이터, 더 짧은 시간, 더 적은 비용과 칩으로 만든 작고 효율적인 AI 모델로도 최첨단 기능을 구현할 수 있다는 가능성이 제기됐다.[7]

AI 모델의 학습 증가세가 둔화된다면, AI 컴퓨팅의 초점은 점차 추론으로 이동하게 된다. 대규모 언어 모델(LLM)에게 문서를 요약하도록 요청하는 작업은 가장 일상적 추론의 예시이지만, 모델을 학습시키는 데 필요한 컴퓨팅의 극히 일부분만 사용한다. 그러나 수십억 명의 소비자와 근로자가 문서 요약 요청을 훨씬 더 많이 더 자주 하게 되면, 추론량이 계속 누적됨에 따라 전체 컴퓨팅 워크로드에서 추론이 차지하는 비중이 학습을 훨씬 능가하게 된다. 이러한 요청 중 일부는 소비자와 기업 사용자들이 사용하는 스마트폰이나 PC 등 단말기에서 처리될 수 있다. 실제로 딜로이트가 2024년에 정확히 예측한 대로, 2025년에는 온디바이스 AI 가속기 칩이 탑재된 PC와 스마트폰이 수억 대 판매됐다.[8] 또한 추론은 학습보다 연산 강도가 낮기 때문에, 데이터센터 내부에서도 추론 최적화 전용 칩을 사용할 수 있다. 이러한 칩은 학습을 대규모로 확장하는 데 필요한 초고성능 AI 칩보다 저렴하고, 추론당 전력 소모도 적으며, 고가의 결합형 메모리가 많이 필요하지 않을 수도 있다.[9]

2025년에 실제로 이러한 전망이 현실화됐고 2026년에도 이러한 추세가 지속될 가능성이 크다. 딜로이트가 2025년에 실시한 서베이에 따르면, 전 세계에서 생성형 AI를 사용하는 소비자가 증가하고 있으며 일일 사용률도 늘고 있다.[10] PC와 스마트폰 등 엣지 디바이스에는 점점 더 강력한 온보드 AI 가속기가 탑재되고 있다. 다수의 추론 최적화 ASIC(application-specific integrated circuit, 특정 용도 및 기능을 수행하도록 맞춤 설계된 주문형 반도체)이 설계 및 제조되어, 데이터센터와 일부 엣지 디바이스에 배치되고 있다. 메타(Meta), 구글(Google), 아마존(Amazon), 인텔(Intel), AMD, 퀄컴(Qualcomm), 그록(Groq), 삼바노바(SambaNova), 세레브라스(Cerebras), 그래프코어(Graphcore) 외에도 여러 기업이 이러한 ASIC을 출시했다. 이 중 일부는 설계사가 프로세싱 코어를 제공하는 브로드컴(Broadcom) 패키지 솔루션을 기반으로 하고 있다.[11] 이들 칩의 구체적인 판매량이 모두 공개되지는 않았지만, 딜로이트는 2025년 해당 칩 시장 규모가 200억 달러를 넘고, 2026년에는 500억 달러를 돌파할 것으로 보고 있다.[12]

그렇다면 개당 3만 달러를 넘어 2028년에는 총합 4,000억 달러 규모에 이를 것으로 추정되는 고전력 칩이 여전히 필요한 이유는 무엇인가?[13] 또한 2026년 한 해에만 4,000억 달러, 최대 1조 달러까지 소요될 수 있는 데이터센터가 여전히 필요한 이유는 무엇인가?[14]

2. 과거보다 훨씬 복잡해진 AI 모델 학습

첫 번째 스케일링 법칙의 핵심은 '더 나은' AI 모델을 만드는 것이고, 적어도 몇 년 동안은 이 법칙이 큰 효과를 가져왔다. 초기 형태의 스케일링은 과거에는 단순히 학습(training)이라고 불렸지만 이제는 '사전 학습'(pre-training)'으로 알려진 기초 모델 생성 단계를 뜻한다.

그러나 더 나은 모델을 만드는 방법은 두 가지가 더 있다. 첫째는 '사후 학습'(post-training) 스케일링으로, 파인튜닝(fine-tuning)*, 프루닝(pruning)**, 양자화(quantization)***, 지식 증류(distillation)****, 인간 피드백과 AI 피드백을 활용한 강화학습, 합성 데이터 증강 등 다양한 기법을 포함한다.[15] 두번째는 테스트 단계 스케일링 또는 장시간 사고(long thinking)로, 모델이 질문을 받은 이후 추론 과정에서 스스로 사고하도록 하는 방식이다. 여기에는 연쇄 사고(chain-of-thought) 프롬프트*****, 다수결 기반 샘플링(majority voting)******, 검색, 일부 사후 학습 기법까지 다양한 방법이 활용된다.[16] 이러한 방식은 선택지를 늘리고 정보 출처를 개선함과 동시에 환각(hallucination)을 줄임으로써 정확도를 향상시킬 수 있다는 장점이 있다.[17]

* **파인튜닝(fine-tuning)**은 사전에 학습된 모델(pretrained model)을 기반으로, 특정 목적·도메인·업무에 맞게 추가 학습시켜 성능을 정밀하게 조정하는 과정을 의미한다. 범용 모델의 추상적 지식을 실제로 쓰이는 정확한 결과로 전환하고, 특정 산업·기업 데이터를 반영하며, 톤·스타일·규칙·정책 일관성을 확보할 수 있다.

** **프루닝(pruning)**은 학습된 모델에서 중요도가 낮은 파라미터(가중치·연결·뉴런)를 제거해 모델을 경량화하고 효율을 높이는 기법을 의미한다. 연산량·메모리·전력 소모를 줄이고, 추론 속도를 개선하고(특히 엣지·모바일 환경), 과적합 완화 및 일반화 성능을 개선하고, 배포 비용을 절감할 수 있다.

*** **양자화(quantization)**는 모델의 파라미터나 연산에서 사용하는 수치 정밀도를 낮춰 경량화하는 과정이다. 모델 크기를 축소해 메모리 사용량을 줄이고, 연산을 단순화해 추론 속도를 개선하고, 엣지 및 모바일 환경에서 유리하기 때문에 전력 소모를 절감하고, 클라우드 추론 비용이 줄어 비용을 절감할 수 있다.

**** **지식 증류(distillation)**는 크고 성능이 좋은 모델이 작고 효율적인 모델에게 지식을 압축해 전달하는 경량화·고도화 기법이다. 추론 비용과 지연시간을 감소해 모델을 경량화할 수 있고, 단순 축소 대비 정확도가 우수해 성능을 유지할 수 있으며, 엣지 및 온디바이스 환경에 배포할 수 있고, 대규모 서비스 비용을 절감할 수 있다.

***** **연쇄 사고(chain-of-thought) 프롬프트**는 LLM이 문제 해결 과정에서 중간 추론 단계를 거쳐 최종 답변에 도달하도록 유도하는 프롬프트 기법을 의미한다. 단순히 정답만을 요구하는 것이 아니라, 문제 인식 → 단계별 분석 → 결론 도출의 사고 흐름을 거치도록 설계된 점이 주요 특징이다.

****** **다수결 기반 샘플링(majority voting)**은 동일한 문제에 대해 복수의 후보 답변을 생성한 뒤, 가장 빈번하게 등장하는 결과를 최종 답으로 채택하는 방법론을 의미한다. 생성형 AI 맥락에서는 하나의 질문에 대해 모델이 여러 번 독립적으로 추론·생성한 결과 중 최다 득표 답변을 선택함으로써 정확성과 안정성을 제고하는 기법으로 활용된다.

3. 새로운 AI 기술, 효율성 개선 효과보다 전력 소비 부담이 더 큰 이유

첫째, 사후 학습 스케일링과 테스트 단계 스케일링이 새로운 표준 단계로 자리 잡아 전력 소비가 커지고 있다. 대부분의 AI 기업이 이제 이러한 방법을 활용해 다양한 방식으로 AI 모델을 개선하고 있다.[18]

둘째, 두 가지 종류의 스케일링 모두 막대한 컴퓨팅 자원을 필요로 한다. 사후 학습에 투입되는 전체 연산량은 원래의 기초 모델을 학습하는 데 필요한 컴퓨팅의 30배에 달하는 것으로 추정되며, 장시간 사고는 이메일 요약과 같은 단순 추론에 비해 100배 이상의 연산을 요구한다.[19]

셋째, 이러한 스케일링 기술이 널리 사용되며 컴퓨팅 자원 요구량이 급속도로 증가함에 따라, AI 데이터센터, 데이터센터의 입지와 전력 수요, 데이터센터에 탑재되는 AI 칩 및 기타 인프라, 이전 세대 AI 칩, 엣지 디바이스 등 여타 다양한 요인들도 큰 영향을 받고 있다.

2025년 이후 AI 데이터센터 변화에 대한 딜로이트의 과거 전망 되돌아보기

데이터센터는 수십 년 전에 등장했다. 실제로 전 세계 데이터센터가 차지하는 면적은 수천만 평방피트에 달하며, 매년 수백억 달러 규모의 반도체 부품이 이러한 데이터센터를 채우기 위해 판매되고 있다.[20] 그러나 차세대 AI 데이터센터와 이를 위한 신형 반도체는 기존 데이터센터 및 반도체와는 매우 다른 경우가 많다. 말 그대로 '밤과 낮' 만큼이나 다르다.

차세대 AI 데이터센터는 매년 수천억 달러의 건설 비용이 들고, 수백 기가와트(GW)의 전력을 소비할 것으로 전망된다. 이러한 시설은 대부분 이전 세대 데이터센터와 다른 냉각 방식, 전력 공급 및 전압, 내부 통신 기술이 필요하다. 서버랙의 밀도와 무게가 훨씬 증가하기 때문에 바닥 자체도 더 두껍게 만들어야 한다. 아마도 가장 중요한 변화는 기존 데이터센터가 중앙처리장치(CPU)를 중심으로 하고 그 주변에 메모리를 배치하는 구조였던 반면, 최신 AI 서버랙은 그래픽처리장치(GPU)라 불리는 특수 칩으로 구성된다는 점이다.[21] 이 GPU에는 고대역폭 메모리(HBM)가 밀착 통합돼 있으며, 방대한 AI 연산을 조율 및 관리하기 위한 특수 설계 CPU도 함께 필요하다. 이러한 차세대 AI 데이터센터를 위해 설계된 부품은 대부분 규모와 성능의 특수성이 과거의 부품과 완전히 다르다.[22]

불과 2006년까지만 해도 고급 GPU는 게임용 컴퓨터나 콘솔에 쓰는 부품으로 여겨졌으며, 데이터센터와는 무관한 것으로 간주됐다.[23] 대부분의 데이터센터 업무는 순차적으로 작업을 처리하는 직렬 기반 CPU로도 충분히 감당 가능했다. 일부 고성능 컴퓨터, 즉 슈퍼컴퓨터에는 대량 병렬 처리기(massively parallel processor)라 불리는 특수 칩이 탑재돼 수백 개의 작업을 동시에 실행할 수 있었다. 그러나 이 칩은 게임용 GPU나 데이터센터용 CPU에 비해 수십~수백 배 이상 비쌌다.

2009년 과학자들은 게임용 고급 GPU도 병렬 프로세서라는 점에 주목하고, 동일한 고급 GPU로 머신러닝 모델을 돌려보기 시작했다.[24] 결과는 성공적이었다. 몇 년 뒤에는 게임용 GPU와는 약간 다르지만 최적화된 전용 GPU가 데이터센터와 온프레미스 장비에서 머신러닝 AI를 수행하는 데 쓰이기 시작했다.[25] 그러나 해당 시장 규모는 2018년까지만 해도 연간 수십억 달러 수준에 머물렀다.[26]

2022년에 들어 생성형 AI를 위한 LLM 개발이 활성화되면서, GPU는 더욱 고도화된 형태로 진화해야 했고, 상대적으로 새로운 형태의 특수 메모리인 HBM과 하나의 패키지로 통합되는 형태가 필요해졌다.[27] GPU와 HBM을 통합한 형태에는 방대한 데이터 흐름을 조율할 장치도 필요했다. 또한 기존 컴퓨터·스마트폰·데이터센터용 CPU와는 다르지만 핵심 아키텍처는 유사한 특수 설계 CPU가 생성형 AI 데이터센터의 주요 부품으로 추가됐다. 여기에 다른 여러 핵심 부품까지 결합됐다. 2025년 기준 전 세계 상위 500대 슈퍼컴퓨터의 거의 대부분은 GPU-특수메모리-CPU 조합이 공통적으로 탑재돼 있다.[28] 따라서 현재 구축되고 있는 메가스케일 AI 데이터센터는 특수화된 슈퍼컴퓨터의 변형 모델이라고도 볼 수 있다.

컴퓨팅 수요 증가가 AI 생태계에 가져올 변화

이제 기업은 사후 학습과 테스트 단계 스케일링이 지속되면서 대규모 데이터센터 및 엔터프라이즈 AI 팩토리의 컴퓨팅 수요가 계속 증가하는 미래에 대비해야 한다. 추론 최적화 칩과 엣지 프로세스의 성장은 계속되겠지만, 그럼에도 불구하고 하이퍼스케일 데이터센터와 엔터프라이즈 AI 장비에 대한 투자는 여전히 필요하다. '추론 최적화' 칩이라 해서 반드시 전력 소모 감소 효과가 있는 것도 아니다. 최근 출시된 한 제품은 추론 프리필(pre-fill) 연산* 최적화를 위해 HBM을 사용하지 않고 그래픽 D램 GDDR7(graphics double data rate 7)을 사용하지만, 랙 하나당 370kW의 전력 밀도가 필요하다. 이는 동일 공급업체의 훈련용 모델 대비 거의 세 배에 달하는 수준이다.[29]

* **추론 프리필(pre-fill) 연산**은 LLM이 실제 토큰 생성을 시작하기 전에 입력 프롬프트 전체를 한 번에 처리하여 내부 상태를 구축하는 초기 추론 단계를 의미한다. 겉으로 드러나지 않지만, 생성형 AI 서비스의 응답 속도, 비용 구조, 확장성을 좌우하는 핵심 기반 기술이다.

1. AI 데이터센터의 자본지출

2026년 전 세계 AI 데이터센터 자본지출(capex)은 4,000억~4,500억 달러로 예상되며,[30] 이 중 절반 이상(2,500억~3,000억 달러)은 장비 내부 칩 구매 비용이고,[31] 나머지는 토지, 건설, 전력 인프라, 인허가 등 기타 비용이 차지한다. 2028년에 이르면 AI 데이터센터 자본지출은 1조 달러까지 증가할 것으로 전망되며,[32] 이 중 AI 칩만 4,000억 달러 이상을 차지할 것으로 예측된다.[33] 사전 학습 성장세가 둔화되고 컴퓨팅의 비중이 훈련에서 추론으로 이동하고 있음에도, 사후 학습과 테스트 스케일링, 사용량 증가에 따른 컴퓨팅 수요를 감안하면, 전 세계는 여전히 방대한 데이터센터를 필요로 한다. 따라서 2025년 3,000억~4,000억 달러 규모였던 AI 데이터센터 자본지출이 2028년 1조 달러 규모로 증가한다는 전망은 실현될 가능성이 매우 크다.

2. AI 데이터센터의 입지

100조 파라미터 규모의 LLM 사전 학습은 수주에 걸친 시간 투자가 필요한 데다 잠시라도 중단되면 문제가 커진다. 핵심 부품 고장이나 프로세서 간 지연이 발생하면 모든 작업이 무효화돼 다시 시작해야 할 수도 있다. 이 때문에 대부분의 기초 모델 사전 학습은 단일 건물 또는 단일 캠퍼스 내에서 서버와 랙을 공동 배치(co-location)하는 방식으로 진행돼 왔다. 그러나 점점 더 많은 AI 연산을 미국 전역 또는 전 세계 여러 데이터센터에서 분산 처리할 수 있게 됐다.[34] 게다가 완전한 사전 학습이 아닌 최종 추론을 수행하는 소규모 추론 데이터센터는 지연을 줄이기 위해 대도시 인근에 위치하게 될 가능성이 크다. 이는 각국이 자국 내에 AI 연산 역량을 확보하려는 소버린 AI(sovereign AI)에 대한 니즈를 높이며, 엔터프라이즈 온프레미스 기반 하이브리드 클라우드 확산을 촉진하는 요인으로 작용하고 있다.[35]

3. AI 데이터센터의 전력 수요

사전 및 사후 학습과 테스트 등 세 단계 스케일링을 수행하는 AI 데이터센

터는 지속적으로 막대한 전력이 필요하다. 다만 사후 학습과 테스트 단계 스케일링은 사전 학습과 달리 비교적 중단이 용이하다는 특징이 있다. 사전 학습은 하나의 긴 작업으로 수행돼야 하지만, 사후 학습 및 테스트 단계는 부하 이전(load shifting)이 쉽기 때문에 AI 기업은 전력 수요 반응 프로그램에 참여할 수 있다. 즉, 특정 시간대에 데이터센터를 다른 지역으로 전환하거나 CPU/GPU의 클럭 속도(clock speed, 프로세서가 초당 연산 사이클을 수행하는 횟수)를 낮춰 피크 시간대 전력 수요를 줄일 수 있다.[36] 이처럼 전력 부하를 유연하게 운영하면 전력망 안정성과 비용 절감을 동시에 달성할 수 있다.[37]

또한 AI 학습 및 추론 작업을 분산할 수 있기 때문에, 데이터센터가 특정 지역에 몰리지 않고 전 세계적으로 보다 고르게 분산될 수 있다. 이는 전력 수요를 분산하는 데에도 도움이 된다.

4. AI 데이터센터의 반도체칩

일각에서는 AI 칩 시장을 '제로섬 게임'으로 보는 시각도 있다. 즉, 모델 사전 학습에는 HBM 결합 고성능 GPU에 수만 달러를 지출해야 했지만, 컴퓨팅의 비중이 추론으로 이동하면 HBM이 적게 들어가 더 저렴한 추론 최적화 칩으로 대체될 것이라는 분석이다.

그러나 실제로는 대체가 아니라 병행의 구도가 될 가능성이 크다. 추론 전용 또는 추론 최적화 칩 시장은 상당한 성장을 보이겠지만, 동시에 기초 모델 사전 학습, 사후 학습, 테스트 단계 스케일링 등 훈련과 추론이 혼합된 연산에 가장 적합한 칩은 여전히 HBM이 탑재된 대형·고성능·고전력 GPU이다. 이들 칩의 가격은 개당 수만 달러에 이른다. 게다가 첨단 공정 웨이퍼 가격이 2026년에 50% 이상 상승할 것으로 예상됨에 따라, 이러한 칩을 구매하는 기업 입장에서는 비용이 더 증가할 수 있다.[38]

5. 스마트폰·PC 등 소비자 및 기업용 디바이스의 엣지 AI

앞서 언급한 바와 같이, 신경망 처리장치(NPU)를 탑재한 수억 대의 스마트폰과 PC가 출하 및 판매되고 있다.[39] NPU는 전용 칩이나 CPU 칩의 일부로 탑재되어 AI 추론 작업을 적정 전력 소모 수준에서 처리하도록 최적화된 장치로, 가격은 몇 달러에서 수십 달러 수준이다.

그러나 NPU는 앞서 설명한 단일 추론(예: 이메일 내용을 요약해줘) 정도를 처리하는 데에는 적합하지만, 그 이상의 고도화된 작업을 수행하기는 어렵다. 따라서 딜로이트는 2026년에 수행되는 거의 모든 AI 연산이 조만간 완공될 초대형 AI 데이터센터 또는 기업이 보유한 고가의 고성능 AI 서버에서 처리될 것으로 전망한다. AI 연산이 PC나 스마트폰에서 주로 처리되지는 않을 것이라는 의미다. 적어도 선점 경쟁이 치열한 초고성장기인 지금은 하이브리드 아키텍처 기반의 비용 최적화 전략이 공급업체나 기업의 우선순위가 될 수 없을 것으로 보인다. 또한 테스트 단계 스케일링과 같은 고성능 기술은 대부분의 소비자용 사용 사례뿐 아니라 상당수의 엔터프라이즈 온디바이스 활용 사례에 적용하기에는 과도한 방식이다. 향후 어느 시점에는 컴퓨터와 스마트폰이 AI 연산에서 더 큰 비중을 차지할 수 있지만, 2026년은 아직 이른 시기라 볼 수 있다.

최근에는 한 AI 기업이 추론 능력을 갖추고 PC에서 로컬로 실행 가능한 생성형 AI 모델을 발표했다. 하지만 실제 성능이 어느 정도인지, 배터리 수명에 어떤 영향을 미치는지, 얼마나 많은 PC 사용자가 클라우드 대신 로컬 실행을 선호할지는 아직 불확실하다.[40]

6. 온프레미스 기반 엔터프라이즈 엣지 AI

전 세계 초대형 AI 데이터센터에 들어가는 고성능·고전력 GPU 및 HBM과 이를 조율하는 특수 CPU 트레이는 사후 학습을 포함한 생성형 AI 컴퓨팅을 온프레미스 기반 하이브리드 형태로 강화해 운영할 때에도 사용할 수 있다. 기업은 비용, 지식재산(IP) 보호, 주권성, 복원력, 맞춤화 요구 등을 고려해 다음과 같은 선택을 할 수 있다.

- 약 8개의 GPU(HBM·CPU 포함)가 탑재된 박스 한 대를 30만~50만 달러에 구매해 일정 수준의 AI 학습 및 추론 수행[41]
- 최대 72개의 최첨단 GPU(HBM·CPU 포함)가 장착된 랙을 300만~500만 달러에 도입해 더 높은 성능 확보[42]
- 필요할 경우 수천만 달러 규모로 여러 랙을 동시에 구축해 대규모 연산 처리[43]

딜로이트는 2026년 이러한 온프레미스 기반 하이브리드 엔터프라이즈 AI 시장 규모가 500억 달러를 넘을 것으로 전망한다.

7. 로봇·드론·자율주행차 등에 탑재되는 엣지 AI

2026년 기준 관련 시장은 아직 상대적으로 규모가 작지만, 실시간(on-device) 추론이 반드시 필요한 여러 사용 사례가 있다. 해당 사례는 드론과 로봇부터 자율주행차에 이르기까지 다양하기 때문에 현재 사용되는 칩 또한 매우 다양하다. 드론은 대부분 상대적으로 단순한 저전력 AI 추론 칩을 탑재하고 있으며,[44] 자율주행차는 대부분 데이터센터에 사용되는 GPU보다 약간 낮은 수준이지만 여전히 고성능 GPU 기반 칩 솔루션이 탑재된다.[45] 이러한 비(非)공장 기반 AI 시장 규모는 2026년에도 50억 달러를 넘지 못할 것으로 보이지만,[46] 로봇 시장이 본격적으로 성장한다면 2030년 이후 크게 확대될 가능성이 있다.[47]

인류의 AI 역사는 아직 초기 단계이다. 알고리즘 효율을 높이려는 지속적인 시도에도 불구하고, 2025년 여름 기준 AI 연산 수요와 사전 학습, 사후 학습, 테스트 단계 스케일링, 추론을 수행하기 위한 데이터센터, 엔터프라이즈 온프레미스 솔루션, 고성능 AI 칩에 대한 수요는 매우 높은 성장세를 보였다.[48] 향후 어느 시점에는 새로운 기술적 돌파구가 등장해, 개선된 AI 모델이 더 저렴한 칩에서도 원활하게 작동하고, 더 적은 데이터센터와 더 적은 전력으로도 동일한 성능을 낼 것이다. 하지만 아직 AI 초기에 해당하는 2026년에는 급증하는 AI 컴퓨팅 수요에 철저히 대비해야 한다.

Korean Perspectives

반도체와 데이터 센터가 경쟁의 본질이 된다

생성형 AI 경쟁의 핵심은 더 이상 어떤 LLM을 만들었는가가 아니다. 이제 승부를 가르는 기준은 AI 연산을 얼마나 안정적으로, 얼마나 많이 처리할 수 있느냐, 즉 컴퓨팅 인프라 역량으로 이동하고 있다. AI 연산의 중심이 모델 학습에서 실제 서비스 과정의 추론으로 옮겨가고 있지만, 사후 학습과 테스트 단계에서의 연산 증가, 그리고 이용자 수 급증으로 전체 연산 수요는 줄어들지 않고 오히려 계속 확대되고 있다. 이는 생성형 AI가 일시적 유행이 아니라, AI 반도체와 데이터센터 투자가 장기간 이어질 구조적 성장 국면에 들어섰다는 뜻이다.

추론 최적화 칩의 부상은 고성능 GPU 중심 구조를 대체하지 않는다. 오히려 사후 학습과 장시간 사고(long thinking)가 새로운 표준으로 자리 잡으면서, HBM이 결합된 고성능·고전력 GPU와 추론 전용 칩이 병행되는 이중 구조가 고착되고 있다. 이 환경에서 경쟁의 핵심은 단일 칩의 성능이 아니라, 워크로드별로 최적화된 칩 포트폴리오를 안정적으로 공급할 수 있는 역량이다.

이 지점에서 SK하이닉스의 전략적 위상은 독보적이다. 현재 글로벌 AI 연산 구조에서 가장 치명적인 병목은 연산 로직이 아니라 메모리 대역폭과 전력 효율, 즉 HBM이다. SK하이닉스는 HBM3·HBM3E 세대에서 기술, 수율, 양산 타이밍을 동시에 선점하며, 글로벌 AI 가속기 로드맵의 핵심 병목을 사실상 통제하고 있다.

오늘날 GPU 출하량과 AI 데이터센터 확장 속도는 곧 SK하이닉스의 HBM 공급 안정성에 의해 좌우되는 구조다. 이는 SK하이닉스가 단순 메모리 공급자가 아니라, AI 연산 구조와 속도를 규정하는 기준점이 되었음을 의미한다.

삼성전자 역시 이 경쟁에서 중요한 축을 형성하고 있다. 삼성전자는 HBM을 포함한 메모리 역량에 더해 파운드리와 첨단 패키징을 결합한 수직 통합 전략을 통해, AI 반도체 경쟁을 단품이 아닌 시스템 레벨 경쟁으로 확장하고 있다. 여기에 자체 GPU 개발 시도와 함께, 리벨리언스 등 AI 추론 특화 칩을 설계하는 국내 팹리스 생태계가 등장하면서 한국은 메모리·패키징·연산 보조 영역에서 대체 불가능한 국가로 자리 잡고 있다.

데이터센터 역시 비용 절감의 대상이 아니라 AI 경쟁의 인프라 전장으로 재정의되고 있다. 초대형 AI 데이터센터와 엔터프라이즈 온프레미스 AI 팩토리는 전력, 냉각, 랙 밀도, 입지 전략까지 포함한 국가 단위 산업 이슈로 확대되고 있으며, 이 과정에서 반도체 기업과 데이터센터 기업의 경계는 빠르게 흐려지고 있다. 특히 HBM을 중심으로 한 고집적 AI 서버 구조는 반도체, 서버, 데이터센터, 전력 인프라를 하나의 패키지로 묶어 경쟁하게 만든다.

결국 한국 반도체 산업은 AI 시대에 단순한 부품 공급자를 넘어섰다. SK하이닉스는 AI 연산 병목을 통제하는 필수 공급자로서, 삼성전자는 시스템 확장 플레이어로서, 한국 반도체 생태계는 이미 AI 연산 인프라의 전략적 요충지를 선점하고 있다.

앞으로의 경쟁은 이 우위를 어떻게 시스템, 데이터센터, 전력 인프라까지 확장해 나가느냐에 달려 있다. 승부는 누가 더 뛰어난 모델을 만드느냐가 아니라, 누가 끝까지 연산을 안정적으로 감당할 수 있는 구조를 구축하느냐에서 갈릴 것이다.

최호계 파트너
한국 딜로이트 그룹
TMT Industry 리더