

Publisher
토스증권 리서치센터

Analyst
이영곤

Date
2026.04.28



GTC DIVE

Deep

01 Overview

우리가 GTC 현장에 다녀온 이유

지난 3월, 저희는 GTC 현장에서 보고 듣고 느낀 것들을 최대한 빠르게 전해드리기 위해 미국 실리콘밸리 현지에서 원격으로 <GTC LIVE> 자료를 발행했습니다. 실제로 가보지 않으면 느끼기 어려운 생생한 현장감을 전달해 드리고 싶었기 때문입니다.

현장에서 받은 인상과 감각은 분명 중요하지만, 투자 관점에서는 한 단계 더 나아가야 합니다. 발표된 기술이 어떤 의미를 가지는지, 산업 구조를 어떻게 바꾸는지, 기업들의 경쟁 구도를 어떻게 재편하는지, 여기까지 연결해야 비로소 '투자 인사이트'가 됩니다.

그래서 이번에는 'GTC Deep Dive' 자료를 준비했습니다.

'LIVE'가 현장에서 본 것이라면, 'Deep Dive'는 그걸 어떻게 이해할 것인가에 대한 이야기입니다. 조금 더 차분하게, 그리고 조금 더 깊이 있게 GTC에서 나온 기술과 메시지를 하나씩 뜯어보며, 그 안에 담긴 의미를 정리해보려 합니다.

이 자료가 엔비디아를 중심으로 AI라는 거대한 흐름을 이해하는 데, 그리고 그 안에서 자신만의 투자 판단 기준을 세우는 데 작은 도움이 되기를 바랍니다.

투자가 필수인 시대, 여러분이 조금 더 편안하게 오래도록 투자하는 데 도움이 되면 좋겠습니다.

낮선 땅에서 담아온 순간들이 투자 여정에 따뜻한 숨결 하나 보태기를 바라며,

2026. 04. 28

토스증권 리서치센터

Intro. GTC 에서 발견한 AI 트렌드 3 가지

GTC가 다가오면 샌프란시스코행 항공권을 구하기가 힘들어집니다. 최소 한 달 전에 예매를 하고 숙소도 잡아야 하죠. 저희도 2월 초에 항공권을 예매했는데요. 행사가 한 달 넘게 남았음에도 남아 있는 비행편이 거의 없어서 남은 좌석을 확보하느라 고생했던 기억이 납니다. 행사 개막 3일 전 샌프란시스코 산호세 인근에 도착한 저희는, 간단히 짐을 풀고 저녁을 먹으러 나갔습니다. GTC 프로그램과 AI 산업의 미래에 대해 열띤 토론을 벌이는 식당 옆자리 사람들에게서 흥분과 기대감이 느껴졌습니다.

실리콘밸리 아침은 늘 분주하지만, GTC가 열리는 날 공기는 조금 더 달랐습니다. 행사장에는 연구원, 투자자, 기자 등 다양한 사람들이 모여 있었는데요. 인도 포함 아시아에서 방문한 이들도 눈에 띄게 많았습니다. 입장을 기다릴 땐 일본에서 온 반도체 연구원과 GTC에 대해, 그리고 AI 반도체 산업에 대해 짧게 이야기를 나누기도 했습니다. 한때 엔비디아 GPU 홍보 행사였던 GTC는 이제 AI 반도체, 고성능 컴퓨터, 데이터센터, 로봇틱스 등 산업 전반의 방향성을 가늠하는 무대가 되고 있습니다.

GTC의 하이라이트는 단연 **젠슨 황의 기조연설(Keynote)**이었습니다. 발표 중 청중들의 가장 큰 호응을 끌어냈던 건 2027년 매출 전망치로 '1 trillion dollar(1조 달러)'를 제시한 순간입니다. 사실 행사 전까지만 해도 AI 수요가 둔화되는 것 아니냐는 우려가 나오고 있었는데요. **젠슨 황은 발표 내내 'AI 산업은 이제 시작 단계에 불과하다'**라는 자신감을 내비치며 확신에 찬 모습을 보였습니다.

GTC는 이제 엔비디아만의 행사가 아닙니다. AI 산업 생태계가 어떻게 구축되고 있으며, 이 산업의 핵심 플레이어들이 어디에 주목하고 있는지 알 수 있는 기회입니다. 이는 투자와도 직접적인 관련이 있을 텐데요.

저희가 GTC 2026에서 확인한 핵심 포인트 3가지는 다음과 같습니다.

- 첫째, AI 산업은 모델 경쟁을 넘어, 생태계와 플랫폼 중심으로 확장될 것입니다.
- 둘째, AI 수요 확대와 효율 개선은 사용량 증가로 이어져, 메모리 수요도 늘어날 것입니다.
- 셋째, AI 팩토리 시대를 맞아 인프라 병목 해소가 중요해질 것입니다.

[Data-1] GTC 2026에 참석한 토스증권 애널리스트



출처: 토스증권

[Data-2] GTC 2026에 참석한 토스증권 애널리스트



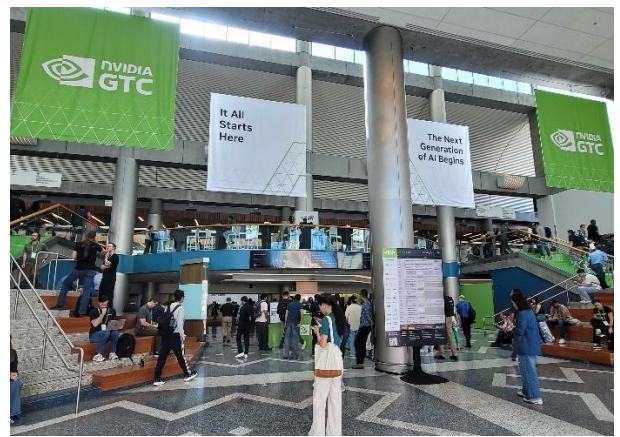
출처: 토스증권

[Data-3] GTC 2026에 참석한 토스증권 애널리스트



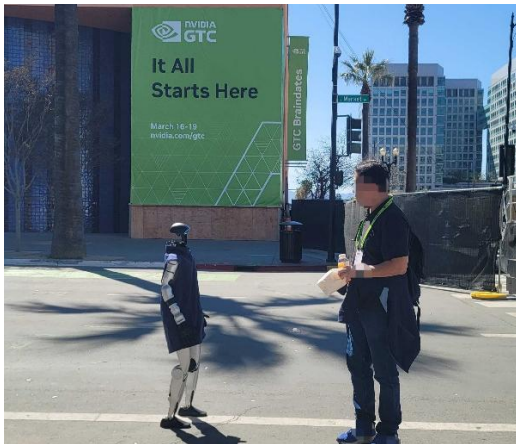
출처: 토스증권

[Data-4] GTC 행사장에는 다양한 세션과 첨단 기업들의 부스가 있다



출처: 토스증권

[Data-5] GTC 행사장 주변에서 거리를 돌아다니는 로봇을 볼 수 있다



출처: 토스증권

[Data-6] GTC 안내데스크에 있는 인공지능 로봇과 대화하는 토스증권 애널리스트



출처: 토스증권

1. '좋은 모델'에서 '좋은 생태계'로

젠슨 황 CEO는 매년 GTC 첫날, 키노트 연설을 통해 엔비디아의 기술 로드맵과 AI 산업 전반의 방향성을 직접 제시합니다. 실제로 투자자와 업계 관계자들은 이 키노트를 통해 향후 AI 인프라 수요, 데이터센터 투자, 기술 경쟁 구도를 가늠합니다.

이번 GTC 2026에서는 **젠슨 황의 키노트 못지않게 중요한 세션이 하나 더 있었습니다.** 바로 3일차에 열린 스페셜 세션입니다. 엔비디아를 비롯해 퍼플렉시티, 커서, 미스트랄 등 AI 산업을 선도 중인 기업의 CEO 11명이 한 자리에 모여 토론을 벌였는데요. 현재 AI 산업의 핵심 플레이어들이 모였다는 점에서 의미가 컸습니다. **젠슨 황은 이례적으로 사회자로서 세션을 이끌기도 했죠.**

[Data-7] GTC 2026 공식 홈페이지에서 메인 이벤트로 소개된 스페셜 세션, '오픈 프런티어 모델(Open Frontier Models)'



젠슨 황이 진행하는 스페셜 패널 세션

3월 18일 (수) 12:30-2 p.m. PDT, Civic Center

지난 1년 사이 AI 분야에서 일어난 가장 중대한 변화를 꼽으라면 단연 오픈 프런티어 모델의 비약적인 발전일 것입니다. 이러한 개방형 혁신은 산업 전반의 발전 속도를 비약적으로 끌어올리고 있으며, 이제 AI가 우리 일상의 모든 영역으로 스며드는 것은 거스를 수 없는 흐름이 되었습니다.

NVIDIA의 CEO 젠슨 황이 진행하는 이번 스페셜 세션에서는 Ai2, Cursor, LangChain, Mistral 등 업계 최고의 리더들이 한자리에 모입니다. 오픈 프런티어 모델이 도달한 현재의 기술적 정점은 어디인지, 그리고 우리가 마주할 다음 단계는 무엇인지에 대한 이들의 솔직하고 심도 있는 인사이트를 직접 확인해 보세요.

출처: 엔비디아 GTC 2026 공식 홈페이지

이 스페셜 세션의 주제가 바로 '오픈 프런티어 모델'(Open Frontier Models)이었습니다.

- 여기서 프런티어 모델이란 GPT, 제미니(Gemini)와 같은 초대형 AI 모델을 의미합니다. 모델 개발에 막대한 자본을 쏟아부은 만큼 독점적으로 운영되었고, 모델 성능 자체가 곧 경쟁력이었죠.
- 이제 그 앞에 '오픈'이 붙은 것입니다. 누구나 다운받아 수정하고 활용할 수 있게끔 모델이 공개되고 있다는 거죠. 실제로 최근 메타의 라마, 미스트랄, 딥시크 등 보다 개방적인 형태의 AI 모델들이 빠르게 확산되고 있습니다.

즉, 공개된 AI 모델을 기반으로 더 많은 기업이 AI를 독자적으로 직접 활용하고 서비스에 적용할 수 있는 환경이 빠르게 만들어지고 있다는 것입니다.

세션 참석자들 역시, 과거엔 소수 기업만 최고 성능의 모델을 갖췄지만 지금은 후발주자들의 모델 성능이 개선되면서 진입장벽이 낮아졌다고 말했습니다.

오픈 프런티어 모델의 등장은 AI 산업 경쟁 구도를 바꾸고 있습니다. 모델 성능이 비슷해질수록 '누가 더 좋은 모델을 만들었는가'보다 '그 모델을 어떻게 서비스와 업무에 연결하는가'가 더 중요해지기 때문입니다.

이러한 구조에서 핵심은 인프라와 플랫폼입니다. 모델이 범용화될수록 이를 빠르고 효율적으로 실행할 수 있는 GPU, 데이터센터, 네트워크가 곧 경쟁력이 됩니다. 엔비디아가 GPU, CPU, 네트워크, 소프트웨어를 묶은 풀스택 AI 인프라를 강조하는 것도 그래서입니다.

[Data-8] AI, 폐쇄형 모델과 개방형(Open) 모델의 개념 및 장단점 비교

	폐쇄형 모델	개방형 모델
개념	모델을 공개하지 않고 API 형태로만 제공	모델을 공개하거나 직접 실행 가능한 형태로 제공
사용 방식	API로 요청해서 사용	다운로드 후 직접 실행 가능
주요 기업	오픈AI, 구글, 엔트로픽	메타(라마), 미스트랄, 딥시크
장점	성능 높고 안정적, 사용이 쉬움	자유롭게 수정 및 활용 가능, 비용 절감 가능
단점	비용 발생, 해당 기업에 의존하게 됨	직접 운영 필요, 관리 부담
핵심 방향	높은 성능과 독점적 모델 바탕으로 수익 창출	모델 확산을 통해 생태계 중심이 되려는 전략

출처: 토스증권

1 최고 수준의 성능에 가까운 모델이면서도, 외부에서 활용, 수정, 배포가 가능한 형태로 개방된 모델.

[Data-9] 젠슨 황과 스페셜 세션 패널로 참석한 글로벌 AI 리더



출처: 엔비디아

[Data-10] 글로벌 AI 모델 주요 특성과 강점, 역할 비교

	오픈AI	구글	앤트로픽	메타	미스트랄	딥시크
모델 전략	폐쇄형	폐쇄형	폐쇄형	개방형	개방형	개방형
대표 모델	GPT 시리즈	제미나이(Gemini)	클로드(Claude)	라마(Llama)	미스트랄(Mistral), 믹스트랄(Mixtral)	딥시크 R1(DeepSeek R1)
접근 방식	API 기반 제공	API + 제품 통합	API 기반	모델 공개	모델 공개	모델/코드 공개
활용 방식	제한적 접근 (API)	서비스 통합 중심	기업용 중심	자유로운 수정/배포	상업적 활용 가능	높은 개방성
주요 강점	성능, 생태계	검색/클라우드 연계	안정성, 기업 특화	생태계 확장	효율성, 경량화	고성능 + 개방성
주요 역할	프런티어 모델	플랫폼 + AI 통합	기업용 AI	오픈 생태계 확산	오픈 모델 경쟁	비용/성능 혁신

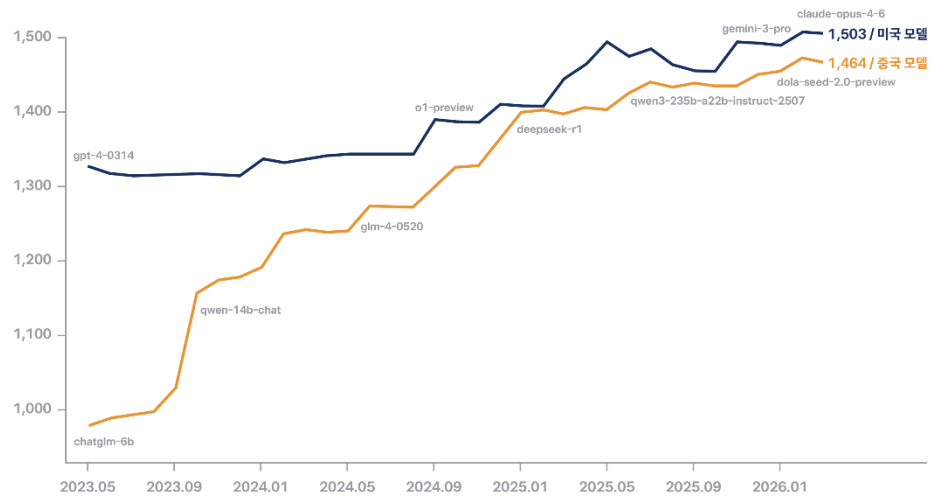
출처: 토스증권

투자 관점에서도 마찬가지입니다.

오픈 모델이든 독점 모델이든 AI 사용량이 늘어날수록 연산량은 기하급수적으로 증가하고, 그에 따라 인프라 수요는 더욱 높아질 것입니다. 엔비디아 같은 인프라 구축 기업을 주목해야 하는 이유입니다.

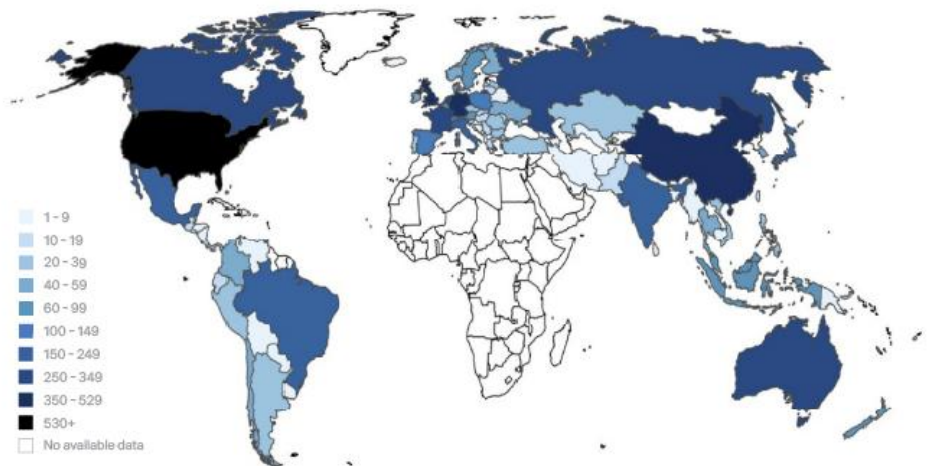
AI 산업은 한 기업이 독식하기보다는 인프라, 플랫폼, 모델, 서비스 등 각 부문에서 핵심 플레이어가 따로 생겨나는 방향으로 갈 가능성이 높습니다. 인프라 부문에서 압도적인 강점을 보유한 엔비디아는 특정 모델의 흥망과 관계없이 산업 성장의 직접적인 수혜를 받을 것으로 보입니다.

[Data-11] AI 모델 경쟁 구도가 초기 단계를 지나 성숙 단계에 진입하면서, 최고 수준 AI 모델 간 성능 격차는 줄어들고 있다



출처: Arena, 2025

[Data-12] AI 수요 확산과 함께 가속화되고 있는 데이터센터 구축 경쟁



출처: Cloudscene, 2025

2. 효율 개선 통해 사용량 더욱 증가할 것

GTC 키노트가 진행된 SAP 센터에서 약 2km 정도 떨어진 곳에 400여 개 기업들의 부스가 열렸습니다. 마이크로소프트, 구글, 아마존 같은 빅테크뿐 아니라 델, HP 같은 하드웨어 기업, 데이터브릭스, 스노우플레이크, 서비스나우 같은 소프트웨어 기업까지 AI 생태계 속에 있는 다양한 기업들을 만날 수 있었습니다.

가장 눈에 띈 곳은 삼성전자와 SK하이닉스 부스였습니다. 두 곳 모두 부스 규모가 꽤 컸는데요. 넓은 공간을 꽉 채울 정도로 많은 사람들이 모여들었습니다. 부스에 배치된 연구원과 이야기를 나누려면 줄을 서서 기다려야 할 정도였습니다. 두 기업의 부스는 다른 부스들과 성격이 조금 달랐는데요. 상당수가 AI 기술과 활용에 초점을 맞춘 반면, 삼성전자와 SK하이닉스는 AI 기술에 필요한 '인프라', 즉 반도체를 만드는 기업이라는 점이었습니다.

AI 연산이 늘어날수록 데이터를 빠르게 처리할 수 있는 고성능 메모리(HBM 등)의 중요성도 함께 커지고 있습니다. 특히 최근 AI 산업이 '학습(Training)' 중심에서 '추론(Inference)' 중심으로 바뀌면서 고성능 메모리는 더욱 주목받는 상황입니다. 젠슨 황도 이번 GTC 2026에서 '추론 중심'으로의 변화를 "inference inflection"이라 표현하며 '추론'의 중요성을 여러 번 강조했는데요.

그렇다면 '추론'은 무엇이고, 왜 이토록 중요해졌을까요?

[Data-13] 학습(Training)과 추론(Inference) 비교

	학습(Training)	추론(Inference)
개념	AI가 데이터를 통해 스스로 규칙을 배우는 과정	학습한 내용을 바탕으로 실제 결과를 만드는 과정
적용 시점	AI를 만드는 과정인 초기 단계 (개발 단계)	AI 학습 이후 실제로 사용하는 단계 (서비스 단계)
수행 횟수	제한적	매우 반복적
비용 구조	한 번에 큰 비용 발생	사용량에 따라 지속적으로 비용 발생
연산 특징	대규모 데이터 처리, 장시간 연산	빠른 응답, 반복 연산
중요 요소	GPU 연산 성능	속도, 지연(latency), 효율
인프라 영향	초기 투자 중심	지속적인 인프라 수요 증가

출처: 토스증권

[Data-14] 기업 전시장에서는 각 기업의 핵심 제품과 전략을 한눈에 확인할 수 있다



출처: 토스증권

[Data-15] 주목도 높은 기업들엔 발 디딜 틈조차 없을 만큼 많은 인파가 몰렸다



출처: 토스증권

[Data-16] 삼성전자 전시장에 들어가고 있는 토스증권 애널리스트



출처: 토스증권

[Data-17] 삼성전자 HBM4를 승인했다고 자필 서명한 젠슨 황 CEO



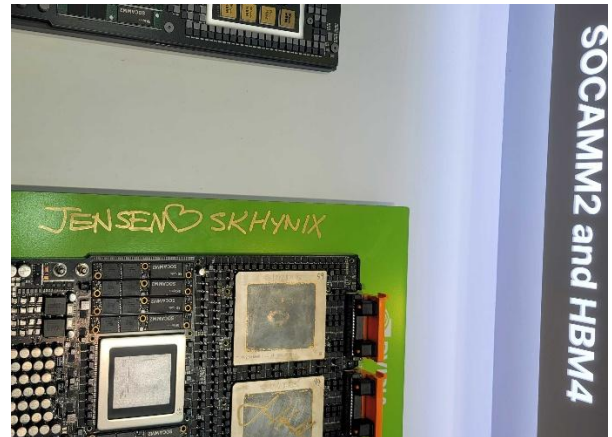
출처: 토스증권

[Data-18] 큰 규모로 자리하고 있는 SK하이닉스 전시장



출처: 토스증권

[Data-19] SK하이닉스 부스의 HBM에는 'JENSEN SKHYNIX'라고 적혀 있다



출처: 토스증권

AI를 작동시키는 건 학습과 추론, 2가지입니다. 학습이 AI 모델을 '잘 만드는 과정'이라면, 추론은 학습된 모델이 요청에 '잘 반응하도록 사용하는' 과정이죠. 쉽게 비유하면 학습은 AI가 공부하는 단계, 추론은 AI가 시험을 보는 단계와 비슷합니다.

추론은 모델 크기가 커질수록 필요한 데이터 양도 늘고, 요청 한 건을 처리할 때마다 상당한 양의 데이터에 접근해야 합니다. 학습과 다른 점이죠. 이때 중요한 것은 필요한 데이터를 얼마나 빠르게 읽어와 GPU에 공급할 수 있느냐입니다. GPU의 계산 속도가 아무리 빠르더라도, 데이터를 제때 전달받지 못하면 전체 처리 속도는 메모리 성능에 맞춰 제한됩니다.

특히 추론 환경에서는 다수의 사용자 요청을 짧은 시간 내에 반복적으로 처리해야 하므로, 단위 시간당 데이터를 얼마나 많이 전달할 수 있는지(메모리 대역폭)가 핵심 요소로 작용합니다. HBM 같은 고대역폭 메모리가 더욱 중요해질 수밖에 없겠죠. 메모리 반도체가 AI 인프라의 필수 요소로 자리잡게 되는 것입니다.

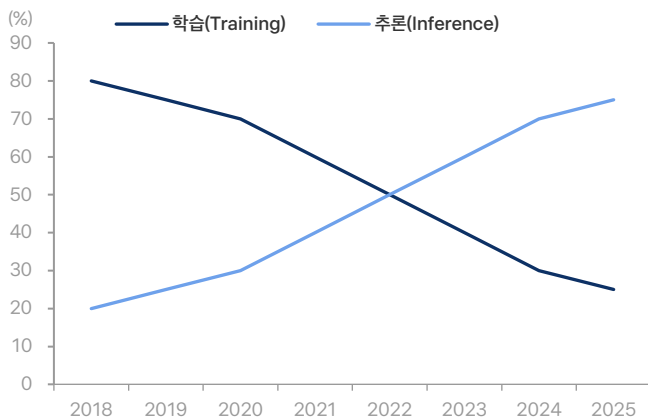
GTC 2026에서 공개된 엔비디아의 루빈(Rubin) 아키텍처는 이러한 흐름을 잘 보여줍니다. 추론 단계에서는 대규모 모델을 한 번 돌리는 것이 아니라 수많은 사용자 요청을 계속해서 반복적으로 처리해야 합니다. 모델의 절대적인 성능보다는 같은 모델로 얼마나 많은 요청을 처리할 수 있느냐, 즉 '효율'이 핵심 경쟁력이 됩니다. 루빈(Rubin)은 기존 대비 최대 10배, 특정 조합에서는 35배 수준의 효율 개선을 제시했습니다. 같은 비용으로 10~35배 많은 요청을 처리할 수 있다는 의미입니다.

최근 엔트로픽, 오픈 AI 등 주요 AI 기업들은 서비스 사용에 대한 과금 체계를 강화하는 움직임을 보이고 있습니다. 이는 모델 운영 비용 증가에 따른 결과로, 헤비유저 포함 프리미엄 AI 서비스 가격은 일부 상승할 수도 있습니다. 그러나 저가 범용 모델은 AI 연산 효율 개선과 단위 비용 하락에 힘입어 오히려 사용량이 증가할 가능성이 높습니다.

AI 사용량이 증가하면 활용 범위가 넓어지고, 그럼 전체 데이터 처리량과 연산 수요가 늘고, 이는 다시 인프라 수요를 자극하는 선순환 구조가 만들어질 것으로 보입니다.

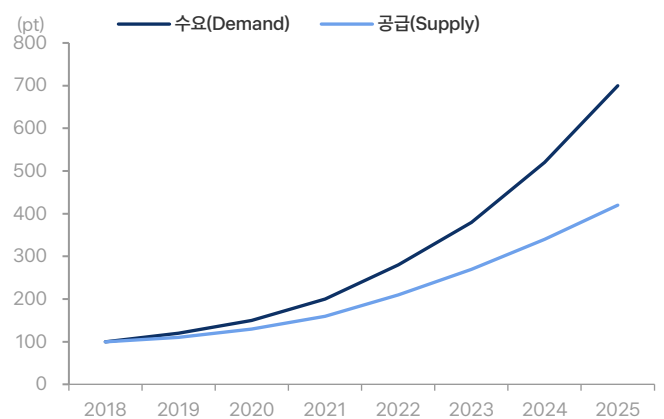
이러한 흐름의 수혜는 AI 인프라 밸류체인 전반으로 확산될 것으로 예상됩니다. 특히 SK 하이닉스, 삼성전자의 HBM은 대규모 데이터 처리 환경에서 필수적인 요소로 자리잡고 있기 때문에, 추론 수요 확대와 AI 사용량 증가의 직접적인 수혜를 받을 가능성이 높아 보입니다.

[Data-20] AI 컴퓨팅 자원의 비중은 학습에서 추론으로 전환되고 있다



출처: 스탠포드 AI Index, 엔비디아, 오픈AI, 토스증권

[Data-21] AI 데이터센터 인프라 수요가 공급을 앞질러, 더욱 격차를 벌리고 있다



출처: IEA, 블룸버그, 맥킨지, 토스증권 추정 (2018년=100으로 환산한 지수)

3. AI 팩토리


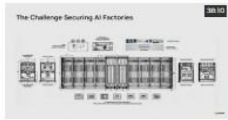


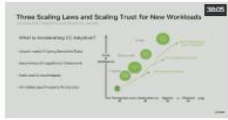

이번 GTC 에서 가장 인상적인 단어는 'AI 팩토리(Factory)'였습니다. AI 팩토리를 주제로 한 세션만 수십 개에 이를 정도로, 이번 행사의 핵심 화두 중 하나였습니다.

젠슨 황은 이전까지 데이터센터가 '창고' 또는 '계산기' 역할을 했다면, 이제는 AI 소프트웨어 및 인프라가 더해져 '공장'으로 진화하고 있음을 강조했습니다. 데이터를 입력하면 텍스트, 이미지, 코드와 같은 결과물을 만들어내는 디지털 생산 시설이 되었다는 의미입니다. 이 공장에서 데이터와 전력은 원재료 역할을 하고 GPU 는 기계, AI 모델은 생산 라인이라 할 수 있습니다.

[Data-22] GTC 2026에서 열린 주요 AI 팩토리 세션

Enterprise AI Factory Conference Sessions <

17 sessions

	<p>Operationalizing AI at Scale: NVIDIA's End-to-End Journey to an Enterprise AI Factory</p> <p>March 2026</p> <p>Ashwin Jha, Senior Director, Enterprise Productivity Engineering, NVIDIA Nic Borenstein, Distinguished Solution Architect, NVIDIA Rama Akkiraju, VP, AI for IT, NVIDIA</p> <p>Learn how NVIDIA's IT organization engineered an on-premises AI factory to deliver agentic AI with enterprise-grade security, reliability, and governance. We'll walk through the end-to-end technical stack behind NVIDIA's internal AI platform, from our on-prem enterprise AI factory infrastructure to containerized agentic ...</p>	<p><</p> <p>☆</p> <p>🔍</p>
	<p>Reinventing Security for AI at Scale</p> <p>March 2026</p> <p>Ofir Arkin, Sr. Distinguished Engineer, NVIDIA Rich Campagna, SVP for Network Security, Palo Alto Networks</p> <p>As enterprises build AI factories, massive data flows, distributed compute, and real-time inferencing push infrastructure to new limits while introducing new security risks. An AI factory is a specialized data center built for intelligence at scale, yet traditional defenses struggle to secure dynamic AI pipelines and emergent model ...</p>	<p><</p> <p>☆</p> <p>🔍</p>
	<p>The Builder's Toolkit: Scaling Enterprise AI Factories</p> <p>March 2026</p> <p>Bert Condensa, Vice President, Enterprise AI Factory Segment Sales, NVIDIA Peter Lillian, Sr. Director of Product Management, NVIDIA</p> <p>This session explores the architectural foundations required to build and scale enterprise AI factories for LLMs, agentic AI, physical AI, and HPC workloads. We'll outline the full-stack infrastructure and software requirements of an AI factory and demonstrate how NVIDIA's NV-Certified systems, enterprise reference architectures, validated ...</p>	<p><</p> <p>☆</p> <p>🔍</p>
	<p>How to Build Planetary-Scale AI Infrastructure</p> <p>March 2026</p> <p>Catherine Kniker, Chief Marketing and Sustainability Officer, PTC Chris Dolan, Chief Data Center Officer, Crusoe Energy Systems, Inc. Natasha Nelson, CTO of Services and VP of EcoStructure Power, Schneider Electric USA Scott Wallace, Director Data Center Engineering, NVIDIA Vivik Mishra, Corporate VP, Cadence Design Systems, Inc.</p> <p>Gigawatt-scale AI facilities are pushing data center infrastructure beyond the limits of traditional design and operations workflows. Sited approaches to buildings, power, cooling, and compute make it difficult to scale efficiently while meeting energy and sustainability requirements. This panel explores how simulation-based co...</p>	<p><</p> <p>☆</p> <p>🔍</p>
	<p>From Isolation to Integration: Evolving Confidential Computing for a Scalable, Secure Future</p> <p>March 2026</p> <p>Emily Sakata, Product Manager, NVIDIA Nelly Porter, Director of Product Management, Trusted Cloud, Google</p> <p>Generative AI is rapidly becoming the defining workload of modern computing, but securing these powerful systems without compromising performance is still a challenge for most enterprises. This session dives into how you can lock down your most valuable AI assets—models, data, and prompts—while continuing to push the limits ...</p>	<p><</p> <p>☆</p> <p>🔍</p>
	<p>Building and Scaling AI Factories With Digital Twins and Robotics</p> <p>March 2026</p> <p>Leo Guo, GM, Hon Hai Technology Group (Foxconn)</p> <p>Foxconn is collaborating with NVIDIA to build an AI factory in Texas for the production of AI servers. Foxconn will leverage NVIDIA Omniverse libraries to build a digital twin environment for simulation and real-time monitoring to rapidly scale on multiple lines, and will introduce humanoid robots and a high degree of automation for the first</p>	<p><</p> <p>☆</p> <p>🔍</p>

출처: 엔비디아

이러한 관점에서 AI 인프라의 핵심 경쟁력은 안정성과 효율입니다. 얼마나 많은 전력을 안정적으로 확보했는지, 얼마나 많은 결과를 만들어낼 수 있는지가 중요한 거죠.

AI가 확장될수록 GPU 외에도 이를 수용할 데이터센터, 안정적인 전력, 냉각, 네트워크 등 물리적 인프라 전반이 함께 필요합니다. 특히 AI 연산이 증가하면 GPU 간 데이터 이동이 급증해 전력 소모 및 지연(latency) 문제가 발생할 확률이 높아지는데요. 그래서 데이터 전송 효율을 높이기 위한 새로운 기술이 필요해지고 있습니다. 기존의 전기 기반 연결은 거리와 속도가 증가할수록 전력 손실 및 발열 문제가 생긴다는 한계가 있기 때문입니다.

이 과정에서 하나의 대안으로 주목받는 것이 바로 광통신입니다. 광 기반 연결은 데이터를 더 빠르고 멀리 전달하는 데 유리하며, 전력 효율 측면에서도 개선 효과를 기대할 수 있습니다. 특히 대규모 AI 클러스터 환경에서는 데이터 이동이 빈번하게 발생하기 때문에, 이러한 효율 개선의 중요성이 더욱 커집니다.

최근 엔비디아가 데이터센터 네트워킹에서 광 기반 연결을 강조한 것도 이러한 흐름과 맞닿아 있습니다. AI 데이터센터가 점점 대규모 클러스터 구조로 확장되면서, 연산 성능뿐 아니라 GPU 간 연결과 데이터 이동 효율이 전체 성능과 비용에 미치는 영향이 점차 커지고 있기 때문입니다.

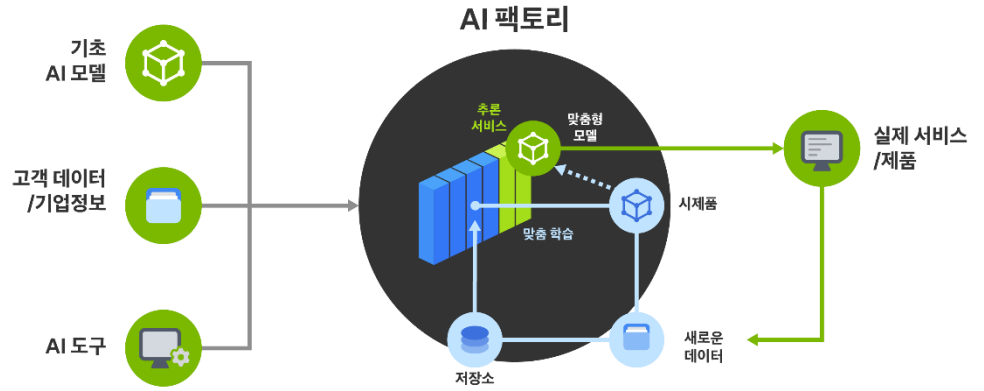
이를 전력 밸류체인과 연결해 보면, 초기 수혜는 송배전과 데이터센터 인프라에서 먼저 나타날 가능성이 높습니다. 초기 단계가 지나면, 같은 전력으로 더 많은 연산을 처리하기 위해 데이터 이동 효율이 중요해지면서 광트랜시버, 실리콘 포토닉스, 광케이블, 스위치 등 광통신 장비의 역할이 커지는 흐름으로 이어질 것으로 보입니다.

[Data-23] 광통신 인프라의 주요 장비와 역할

장비	주요 내용	역할	주요 포인트
광트랜시버 (Optical Transceiver)	전기 신호와 빛 신호를 맞바꾸는 장치	서버·스위치 간 데이터 송수신	광통신의 핵심 인터페이스
실리콘 포토닉스 (Silicon Photonics)	빛으로 데이터를 처리하는 반도체 기술	고속·저전력 데이터 전송 구현하는 반도체	차세대 광통신 핵심 기술
광케이블 (Optical Fiber)	빛이 지나가는 통로	장거리·고속 데이터 전달하는 매체	전력 손실 적고 속도 빠름
스위치 (Switch)	데이터 흐름을 연결, 분배하는 장비	서버·GPU 간 데이터 경로 제어 네트워크	AI 클러스터 연결 핵심
광인터커넥트 (Optical Interconnect)	장비끼리 광케이블로 직접 연결	데이터 이동 효율 개선 장치	대규모 AI 클러스터에 중요
광모듈 (Optical Module)	트랜시버 포함 통합 장치	장비 간 연결 단위	데이터센터 필수 구성 요소

출처: 토스증권

[Data-24] 엔비디아의 AI 팩토리 개념도



출처: 엔비디아, 토스증권

마치며

AI 산업의 핵심은 모델 성능 경쟁을 넘어, 병목을 얼마나 효과적으로 해소하느냐로 이동하고 있습니다. 전력, 데이터 이동, 냉각 등의 인프라가 AI 확산의 기반이 되고, 비용 하락과 추론 확대는 이러한 인프라 수요를 구조적으로 높여지고 있습니다.

이 과정에서 경쟁력은 자연스럽게 서비스와 활용으로 이동합니다. 같은 모델이라도 어떤 문제를 해결하고, 이를 어떻게 사용자 경험에 연결하느냐에 따라 만들어내는 가치가 달라지기 때문입니다.

이제 AI는 '잘 만든 기술'에서 '많이 쓰이는 산업'으로 전환되고 있습니다. 이 흐름 속에서 인프라와 고성능 메모리, 그리고 병목을 해소하는 기술들이 구조적인 수혜를 받을 것으로 판단됩니다.

다음 리포트 예고 – 엔비디아와 차세대 컴퓨팅

다음 리포트에서는 <GTC 2026> 현장에서 보고 들은 이야기를 토대로 엔비디아가 차세대 컴퓨팅을 어떻게 바라보고 있는지, 미래의 컴퓨팅에 본인들의 기술력을 어떻게 적용하고 있는지 전해 드리겠습니다.

Compliance Note

- 당사는 발간일 기준 지난 1년간 위 조사분석자료에 언급된 종목의 지분증권 발행에 참여한 적이 없습니다.
- 당사는 발간일 기준 위 조사분석자료에 언급된 종목의 지분을 1% 이상 보유하고 있지 않습니다.
- 본 자료의 애널리스트는 발간일 기준 위 조사분석자료에 언급된 종목에 재산적 이해관계가 없습니다.
- 본 자료는 기관투자자 등 제 3자에게 사전 제공된 사실이 없습니다.
- 본 자료에는 외부의 부당한 압력이나 간섭 없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.
- 본 자료는 당사의 저작물로서 모든 저작권은 당사에게 있으며, 당사의 동의 없이 어떠한 경우에도 복제, 배포, 전송, 변형, 대여할 수 없습니다.
- 본 자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다.