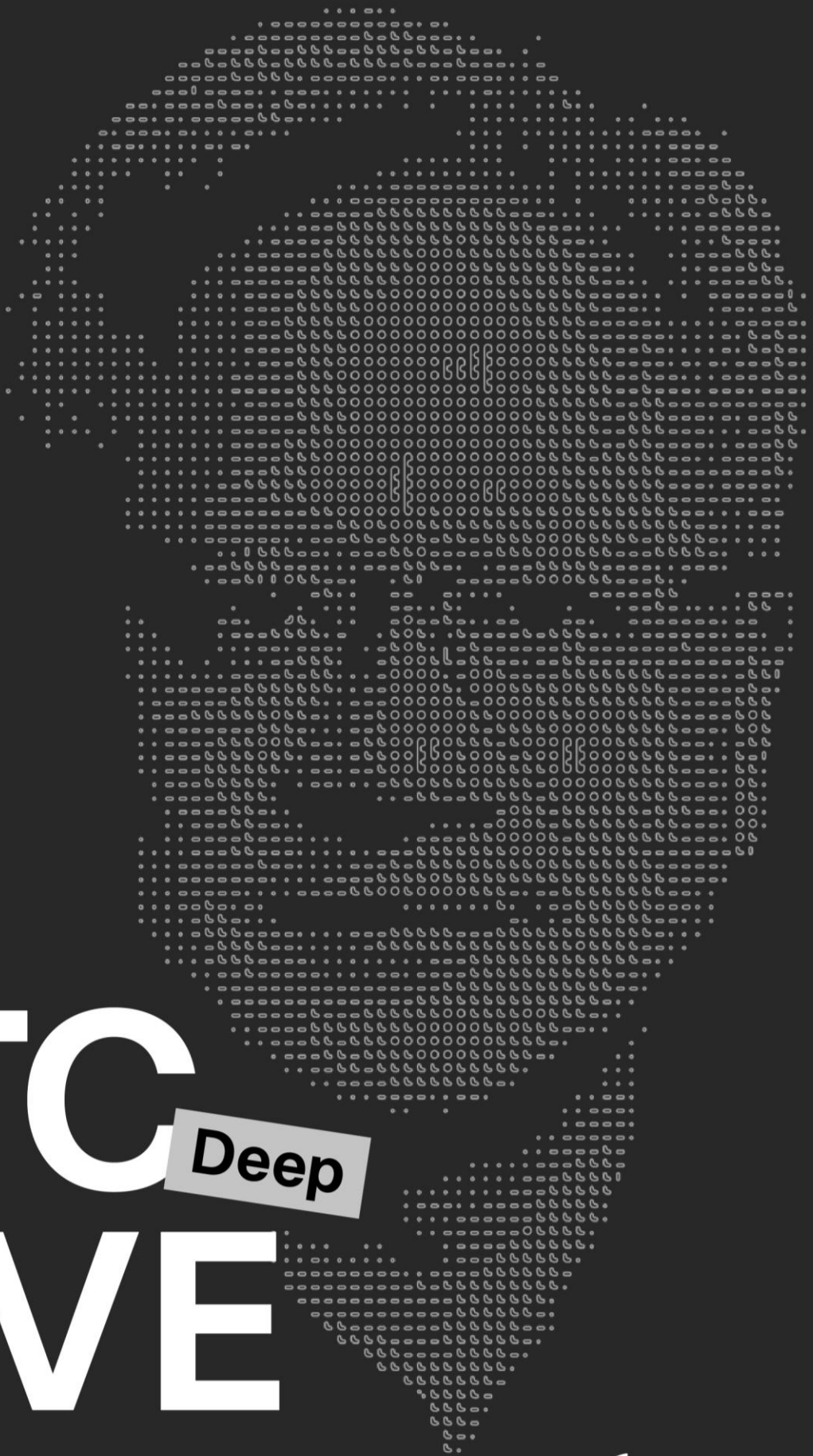
Publisher
토스증권 리서치센터

Analyst
이영곤 이지선 한상원

Date
2026.05.12



GTC Deep DIVE

차세대 컴퓨팅, 추론, 그리고 AI 에이전트

토스증권

GTC Deep Dive

- 차세대 컴퓨팅, 추론, 그리고 AI 에이전트

Intro

3 우리가 GTC 현장에 다녀온 이유

4 요약

Overview

8 '좋은 모델'에서 '좋은 생태계'로

12 효율 개선 통해 사용량 더욱 증가할 것

15 AI 팩토리

차세대 컴퓨팅

22 하이브리드 컴퓨팅, 벌써 시작됐다

26 양자컴퓨팅과 엔비디아 GPU 협업의 의미

27 왜 지금 양자컴퓨팅일까?

추론, 그리고 AI 에이전트

33 오케스트레이션(Orchestration)

36 베라 루빈(Vera Rubin)

38 네모클로(NemoClaw)

우리가 GTC 현장에 다녀온 이유

지난 3 월, 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 GTC 에서 나온 기술과 메시지를 하나씩 뜯어보며, 그 안에 담긴 의미를 정리해보려 합니다.

이번 출장에서 주목한 키워드는 차세대 컴퓨팅, 추론, 그리고 AI 에이전트입니다.

엔비디아는 차세대 컴퓨팅에 꼭 필요한 인프라를 제공함으로써 양자 기술 생태계의 핵심이 되려 합니다. 그리고 추론의 중요성 확대, 에이전트로의 전환 등 AI 산업 내에서 나타나고 있는 다양한 변화에 적극적으로 대응하고 있습니다. GTC 현장에서 보고 들은 내용 중 세 명의 애널리스트가 특히 주목한 내용을 추려 이번 자료에 담았습니다.

이 자료가 엔비디아를 중심으로 AI 라는 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.

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

2026. 05. 12

토스증권 리서치센터

요약

01. AI 는 '잘 만든 기술'에서 '많이 쓰이는 산업'으로 전환되고 있습니다.

GTC 는 한때 엔비디아 GPU 홍보 행사에 불과했지만, 이제 AI 반도체, 고성능 컴퓨터, 데이터센터, 로보틱스 등 산업 전반의 방향성을 가늠하는 무대가 되고 있습니다. 이는 투자와도 직접적인 관련이 있습니다. 저희가 이번 GTC 2026 현장에서 확인한 핵심 포인트 3 가지는 다음과 같습니다. 첫째, AI 산업은 모델 경쟁을 넘어, 생태계와 플랫폼 중심으로 확장될 것입니다. 누구나 다운받아 활용할 수 있게끔 공개되는 '오픈 프런티어 모델'이 확산되면서, 모델을 빠르고 효율적으로 실행할 수 있는 GPU, 데이터센터, 네트워크의 중요성이 커지고 있습니다. 둘째, AI 수요 확대와 효율 개선은 사용량 증가로 이어져, 메모리 수요도 늘어날 것입니다. 특히 SK 하이닉스, 삼성전자의 HBM 은 대규모 데이터 처리 환경에서 필수적인 요소로 자리 잡고 있기 때문에 직접적인 수혜를 받을 가능성이 높아 보입니다. 셋째, AI 팩토리 시대를 맞아 인프라 병목 해소가 중요해질 것입니다. 데이터를 더 빠르고 멀리 전달할 수 있고 전력 효율 측면에서도 유리한 광 기반 연결이 대안으로 주목받고 있습니다.

02. 양자컴퓨팅은 더 이상 먼 미래의 이야기가 아닙니다.

현장에서 특히 인상적이었던 부분은 양자컴퓨팅이 먼 미래의 이야기가 아니라, 현재 AI 인프라 논의의 연장선 위에 올라와 있다는 점입니다. 엔비디아 내부 디렉터가 직접 양자컴퓨팅 방향성을 설명했고, 관련 내용이 컨퍼런스의 상당한 비중을 차지했습니다. 투자 관점에서 양자컴퓨팅이 주목받는 이유는 지금까지 계산 자체가 어려웠던 문제를 풀 수 있는 가능성을 열어주기 때문입니다. 그럼에도 양자컴퓨팅은 '오류가 자주 발생하고 비효율적'이라는 이유로 지금껏 상용화되지 못했는데요. 엔비디아는 그동안 쌓아온 노하우로 이러한 단점을 보완하려 합니다. QPU 와 GPU 를 연결하는 NVQLink, 개발 환경을 제공하는 CUDA-Q, 핵심 병목인 오류 정정과 보정을 자동화하는 Ising 등 양자 하드웨어가 작동하기 위해 반드시 필요한 인프라를 제공하겠다는 것입니다.

03. '학습'에서 '추론'으로, '챗봇'에서 '에이전트'로 흐름이 바뀌고 있습니다.

AI 산업의 무게중심은 이제 학습에서 추론으로 넘어왔습니다. 이젠 개별 칩의 성능이 아니라 여러 칩을 시스템 내에서 어떻게 최적화하느냐, 그 시스템을 누가 더 효율적으로 관리하느냐가 경쟁력이 될 것입니다. 이를 위한 핵심 키워드로 오케스트레이션(Orchestration), 베라 루빈(Vera Rubin), 네모클로(NemoClaw)를 꼽을 수 있습니다. 오케스트레이션은 비효율 문제를 해결하기 위해 등장한 개념으로, AI 에이전트가 효율적으로 돌아가도록 전체 흐름을 설계하고 역할을 나누고 실행까지 관리하는 것을 의미합니다. 베라 루빈은 루빈 GPU, 베라 CPU, 그록 LPU 등 7 개의 칩으로 구성된 엔비디아의 통합 GPU 플랫폼입니다. 처음부터 7 개의 칩이 최적화될 수 있도록 설계해 효율을 높였습니다. 네모클로는 AI 에이전트를 구축하고 운영하는 과정에서 보안 문제가 생기지 않도록 관리하는 운영 플랫폼입니다. 엔비디아는 네모, 네모트론, 네모클로 등을 묶음으로 제공해 락인(lock-in) 효과를 거두고자 합니다.

Publisher
토스증권 리서치센터

Analyst
이영곤

Date
2026.05.12

GTC Deep DIVE

01 Overview

토스증권

Intro. GTC 에서 발견한 AI 트렌드 3 가지

GTC 가 다가오면 샌프란시스코행 항공권을 구하기가 힘들어집니다. 최소 한 달 전에 예매를 하고 숙소도 잡아야 하죠. 저희도 2 월 초에 항공권을 예매했는데요. 행사가 한 달 넘게 남았음에도 남아 있는 비행편이 거의 없어서 남은 좌석을 확보하느라 고생했던 기억이 납니다. 행사 개막 3 일 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, GTC 가 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 GTC 에 대해, 그리고 AI 반도체 산업에 대해 짧게 이야기를 나누기도 했습니다. 한때 엔비디아 GPU 홍보 행사였던 GTC 는 이제 AI 반도체, 고성능 컴퓨터, 데이터센터, 로보틱스 등 산업 전반의 방향성을 가늠하는 무대가 되고 있습니다.

GTC 의 하이라이트는 단연 젠슨 황의 기조연설(Keynote)이었습니다. 발표 중 청중들의 가장 큰 호응을 끌어냈던 건 2027 년 매출 전망치로 '1 trillion dollar(1 조 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.

GTC 는 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 2026 에서 확인한 핵심 포인트 3 가지는 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

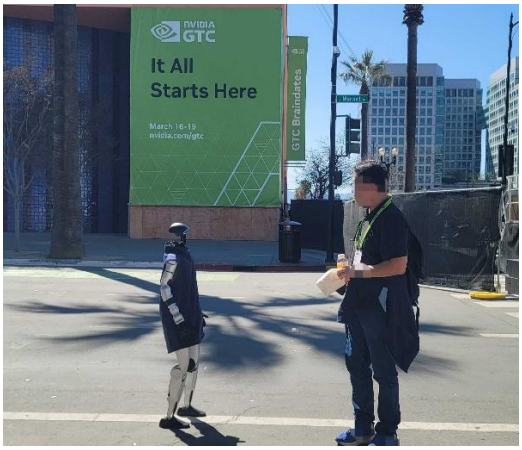
[Data-3] GTC 2026에 참석한 토스증권 애널리스트

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 2026 에서는 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 프론티어 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 프론티어 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[1](Open Frontier Models)'이었습니다.

- **여기서 프런티어 모델이란** GPT, 제미나이(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, 네트워킹, 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, 앤트로픽	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

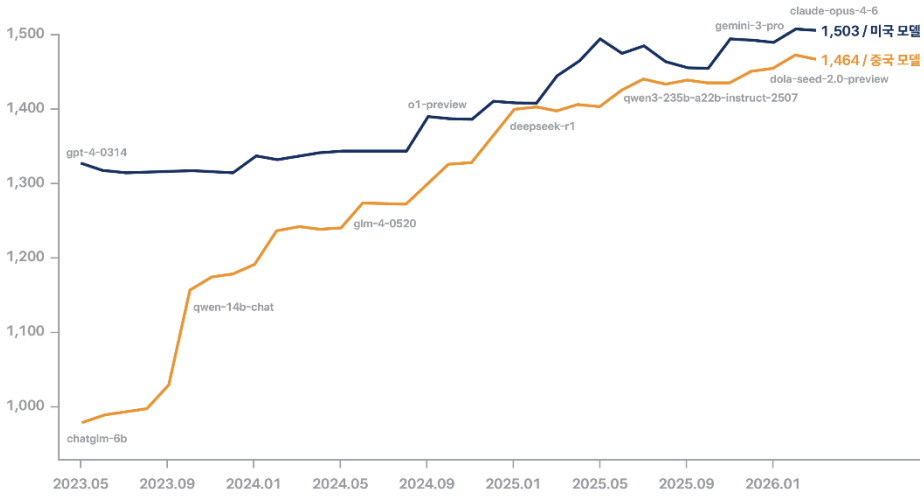
출처: 토스증권

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.

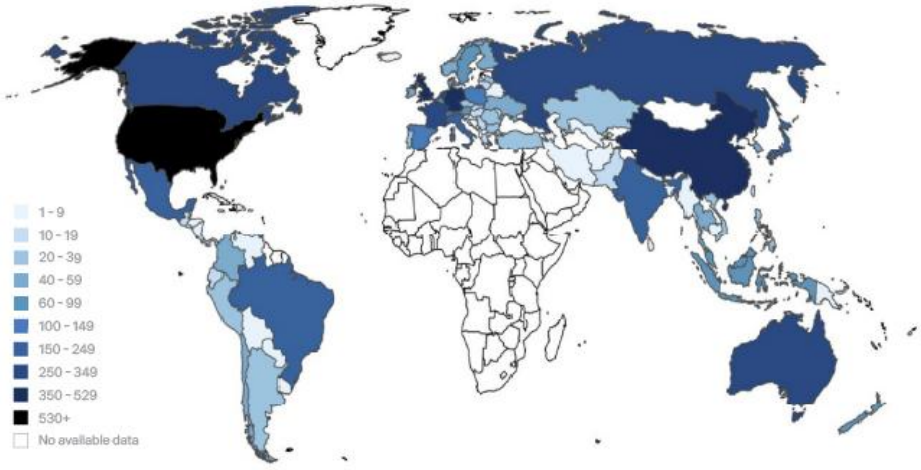
AI 산업은 한 기업이 독식하기보다는 인프라, 플랫폼, 모델, 서비스 등 각 부문에서 핵심 플레이어가 따로 생겨나는 방향으로 갈 가능성이 높습니다. 인프라 부문에서 압도적인 강점을 보유한 엔비디아는 특정 모델의 흥망과 관계없이 산업 성장의 직접적인 수혜를 받을 것으로 보입니다.

[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 SK 하이닉스 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 SK 하이닉스는 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

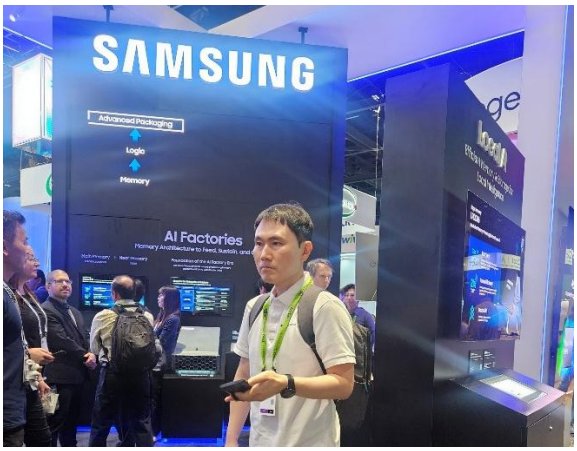
[Data-14] 기업 전시장에서는 각 기업의 핵심 제품과 전략을 한눈에 확인할 수 있다

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

AI 를 작동시키는 건 학습과 추론, 2가지입니다. 학습이 AI 모델을 '잘 만드는 과정'이라면, 추론은 학습된 모델이 요청에 '잘 반응하도록 사용하는' 과정이죠. 쉽게 비유하면 학습은 AI 가 공부하는 단계, 추론은 AI 가 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 GPU 에 공급할 수 있느냐입니다. GPU 의 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.

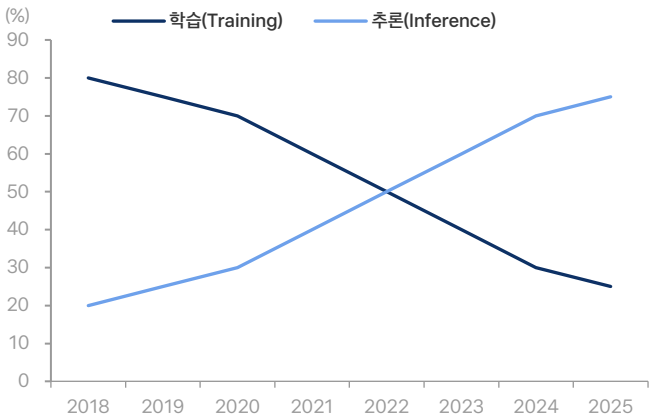
GTC 2026에서 공개된 엔비디아의 루빈(Rubin) 아키텍처는 이러한 흐름을 잘 보여줍니다. 추론 단계에서는 대규모 모델을 한 번 돌리는 것이 아니라 수많은 사용자 요청을 계속해서 반복적으로 처리해야 합니다. 모델의 절대적인 성능보다는 같은 모델로 얼마나 많은 요청을 처리할 수 있느냐, 즉 '효율'이 핵심 경쟁력이 됩니다. 루빈(Rubin)은 기존 대비 최대 10배, 특정 조합에서는 35배 수준의 효율 개선을 제시했습니다. 같은 비용으로 10~35배 많은 요청을 처리할 수 있다는 의미입니다.

최근 앤트로픽, 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.

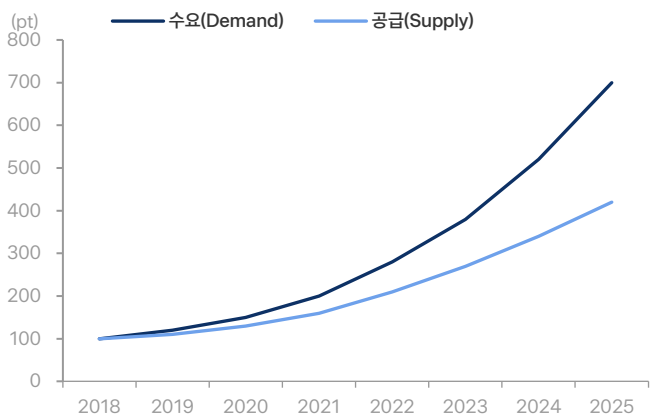
이러한 흐름의 수혜는 AI 인프라 밸류체인 전반으로 확산될 것으로 예상됩니다. 특히 SK 하이닉스, 삼성전자의 HBM 은 대규모 데이터 처리 환경에서 필수적인 요소로 자리잡고 있기 때문에, 추론 수요 확대와 AI 사용량 증가의 직접적인 수혜를 받을 가능성이 높아 보입니다.

[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



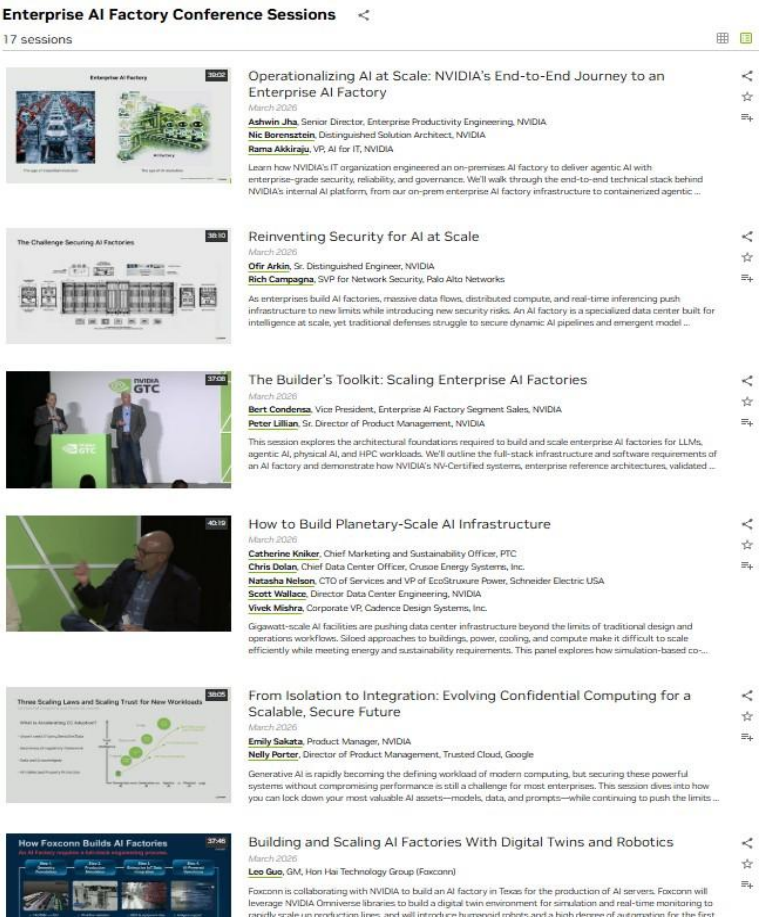
출처: IEA, 블룸버그, 맥킨지, 토스증권 추정 (2018년=100으로 환산한 지수)

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



출처: 엔비디아

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

AI 가 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.

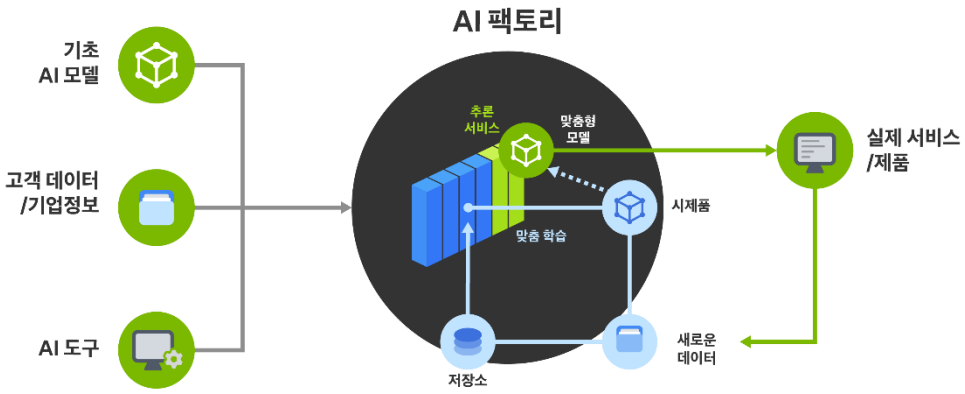
이를 전력 밸류체인과 연결해 보면, 초기 수혜는 송배전과 데이터센터 인프라에서 먼저 나타날 가능성이 높습니다. 초기 단계가 지나면, 같은 전력으로 더 많은 연산을 처리하기 위해 데이터 이동 효율이 중요해지면서 광트랜시버, 실리콘 포토닉스, 광케이블, 스위치 등 광통신 장비의 역할이 커지는 흐름으로 이어질 것으로 보입니다.

[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 높아지고 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.

Publisher
토스증권 리서치센터

Analyst
이지선

Date
2026.05.12

GTC Deep DIVE

02 차세대 컴퓨팅

토스증권

Intro. GTC 에서 만난 양자컴퓨팅

AI 산업은 지금, 모델 경쟁에서 인프라 경쟁으로 빠르게 이동하고 있습니다. 과거엔 더 정교한 모델을 만드는 것이 핵심이었다면, 이제는 그 모델을 얼마나 빠르고 효율적으로 운영할 수 있는지가 경쟁력을 결정합니다. 이 흐름 속에서 대규모 연산을 처리할 수 있는 HPC(High Performance Computing)가 AI 산업의 핵심 인프라로 자리잡았고, 엔비디아는 GPU 기반 HPC 를 통해 이 변화를 주도해왔습니다.

이번 GTC 에서는 그 다음 단계에 대한 논의가 본격적으로 시작됐습니다. GPU 기반 HPC 만으로는 해결하기 어려운 영역이 존재하고, GTC 에서도 그 한계를 보완할 여러 기술이 소개됐는데요. 그중 특히 눈에 띈 것이 바로 양자컴퓨팅이었습니다.

현장에서 가장 인상적이었던 부분은 양자컴퓨팅이 먼 미래의 이야기가 아니라, 현재 AI 인프라 논의의 연장선 위에 올라와 있다는 점입니다. 엔비디아 내부 디렉터가 직접 양자컴퓨팅 방향성을 설명했고, 관련 기업들의 부스 및 발표가 컨퍼런스의 상당한 비중을 차지했습니다. 기술을 넘어, 이미 하나의 산업으로 자리잡기 시작한 느낌이었습니다.

현재의 양자컴퓨팅은 기존 컴퓨팅의 대체재가 아니라 보완재에 가깝습니다. 문서 처리, 데이터 분석, AI 학습과 같은 일반적인 연산에서는 여전히 GPU 기반 HPC 가 훨씬 효율적입니다. 양자컴퓨팅이 투자 관점에서 주목받는 이유는 지금까지 계산 자체가 어려웠던 문제를 풀 수 있는 가능성을 열어주기 때문입니다.

이 리포트는 GTC 에서 보고 듣고 경험한 내용을 바탕으로 양자컴퓨팅의 현주소를 정리한 글인데요. GTC 는 엔비디아의 행사인 만큼 엔비디아의 관점이 상당 부분 반영될 수밖에 없다는 점을 미리 밝힙니다. 다만, 새로운 기술에서 투자 기회를 찾으려 할 때는 엔비디아처럼 시장을 주도하는 플레이어가 가리키는 방향을 먼저 파악하는 것이 중요하다는 생각입니다.

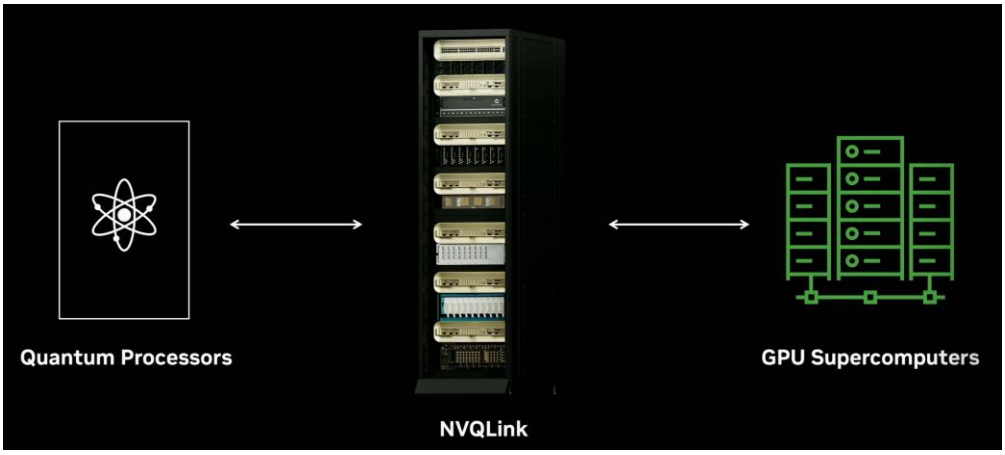
비전공자도 이해할 수 있도록 쉽게 설명하되, 투자 판단에 필요한 최소한의 구조와 맥락을 담으려 했습니다. 빠르게 현실이 되고 있는 양자컴퓨팅 기술을 이해하는 출발점으로 삼으셨으면 합니다.

[Data-1] GTC에 참여한 주요 양자컴퓨팅 기업의 핵심 기술 및 사업 방향 비교

기업명	상장 여부	핵심 기술	설명
리게티 컴퓨팅 (Rigetti Computing)	상장 (RGTI)	풀스택, 하드웨어(초전도)	양자 컴퓨터 칩을 직접 만들고, 클라우드 기반 양자 컴퓨팅 제공. 여러 칩을 연결해 성능을 높이는 기술을 보유.
인플렉션 (Infleqtion Inc.)	상장 (INFQ)	하드웨어(중성원자)	중성원자 기반 양자 기술 리더. 컴퓨팅 외에도 RF 시스템, 양자 시계, 항법 등 상업 포트폴리오 보유.
아이큐엠 퀀텀 컴퓨터스 (IQM Quantum Computers)	RAAQ와 스팩 합병 진행중	풀스택, 하드웨어(초전도)	초전도 양자 컴퓨터를 개발하며 온프레미스 및 클라우드 접근 제공 지향. HPC 및 연구기관 중심 고객 기반.
퀀티넘 (Quantinuum)	비상장 (상장 추진 중)	풀스택, 하드웨어(이온트랩)	세계 최대 규모의 통합 양자컴퓨팅 기업. 고성능 양자컴퓨터, 첨단 소프트웨어를 동시에 개발하며 풀스택 기술 기반 확장 주도.
파스칼 (Pasqal)	비상장 (상장 추진 중)	하드웨어(중성원자)	중성원자 양자컴퓨팅 산업화를 주도. 최적화/시뮬레이션/AI 문제 해결용 시스템 제공.
클래시큐 (Classiq)	비상장	소프트웨어 및 인프라	양자컴퓨터 전문 지식이 없어도 알고리즘을 짤 수 있게 해주는 도구를 만드는 회사. 양자컴퓨팅의 개발 편의성을 높이는 역할.
사이퀀텀 (PsiQuantum)	비상장	하드웨어(포토닉스)	빛(광자)을 이용해 오류 없이 작동하는 대규모 양자컴퓨터 구현이 목표인 회사. 반도체 공정 기반 확장 전략 추진.
큐블록스 (Qblox)	비상장	소프트웨어 및 인프라	양자컴퓨터가 제대로 작동하려면 필요한 정밀 제어 장비를 만드는 회사.
퀀텀머신 (Quantum Machines)	비상장	소프트웨어 및 인프라	양자와 고전 연산을 통합한 제어 플랫폼을 통해 실시간 제어 플랫폼을 만드는 회사. 연구 속도를 높여주는 인프라 역할.

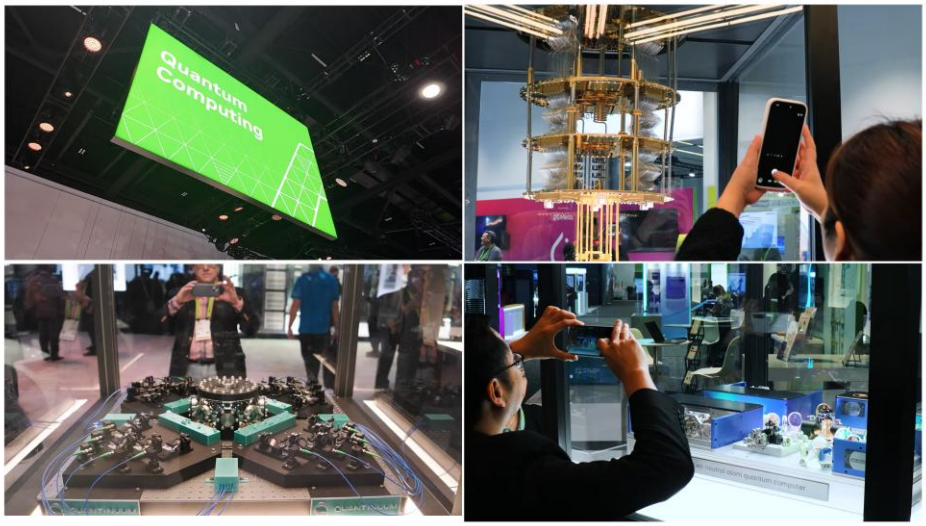
출처: 토스증권

[Data-2] 엔비디아의 NVQLink: QPU (양자 기술 기반 프로세서)와 GPU를 연결하는 기술이다



출처: 엔비디아, 토스증권

[Data-3] GTC 내 양자 컴퓨팅 전시관 모습



출처: 엔비디아, 토스증권

1. 하이브리드 컴퓨팅, 벌써 시작됐다

젠슨 황은 "미래의 슈퍼컴퓨터는 양자와 GPU 가 결합된 형태일 것"이라고 선언했습니다. 이 선언은 GTC 2026 보다 몇 달 앞선 2025 년 말에 이미 나왔는데요. 당시 엔비디아는 "신기술 NVQLink 를 전 세계 10 여 개 슈퍼컴퓨팅 센터 및 연구소에 도입한다"는 내용의 보도자료를 발표하기도 했습니다.

반년도 채 지나지 않아 GTC 2026 에서 이 비전이 다시 한번 강조되었습니다. 엔비디아 중역 세션, 협력사 발표, 부스 투어 등 컨퍼런스 곳곳에서 양자와 GPU 의 연결고리를 확인할 수 있었습니다.

NVQLink: QPU 를 현존하는 컴퓨팅 인프라에 연결한다

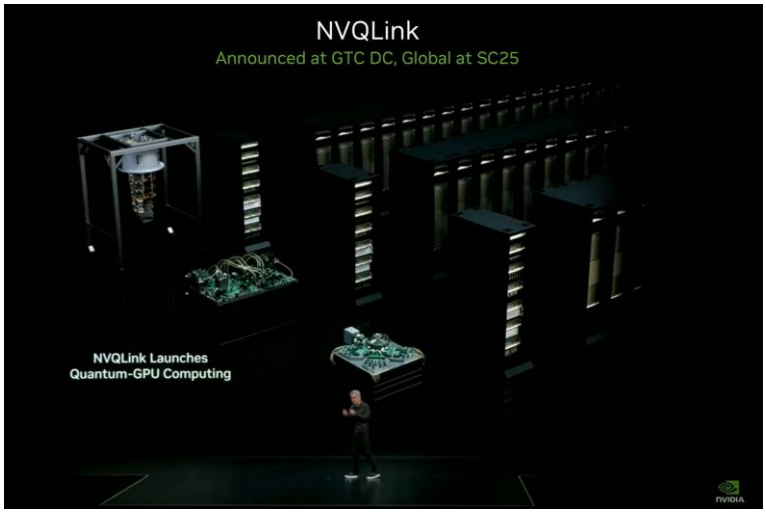
NVQLink 는 양자컴퓨터(QPU)와 GPU 기반 슈퍼컴퓨팅을 연결하는 엔비디아의 인터커넥트 기술입니다. 하이브리드 컴퓨팅을 데이터센터 환경에서 구현하기 위한 핵심 요소로, QPU 와 GPU 를 초저지연으로 연결해 GPU 가 QPU 의 오류를 실시간으로 처리할 수 있도록 설계됐습니다. 2025 년 10 월 첫 발표 후, 이번 GTC 2026 에서는 NVQLink 가 실제 연구 및 프로젝트에 적용된 사례들이 공개됐는데요. 아직 상용 서비스나 대규모 실용 계산 단계는 아니지만, 연구소와 기업 환경에서의 실증 단계로 진입하고 있습니다.

현장에서 공개된 주요 사례는 다음과 같습니다.

- **퀀티넘(Quantinuum):** 하드웨어부터 소프트웨어까지 아우르는 풀스택 양자컴퓨팅 기업으로, 기업 가치가 약 100 억 달러(2025 년 기준) 수준입니다. NVQLink 를 통해 이온 트랩 방식의 QPU 를 엔비디아 GPU 와 통합하고, 양자 오류 정정을 실제로 구현했습니다.
- **국립에너지연구과학컴퓨팅센터(NERSC):** 미국 에너지부 산하 기관으로, 기존 HPC 워크로드에 양자컴퓨팅을 통합하는 방향을 추진하고 있습니다. 이는 엔비디아의 전략 방향과 일치합니다.
- **폭스콘(Foxconn):** 애플 아이폰의 협력사로 잘 알려진 대만 기업입니다. 이온 트랩 기반 양자컴퓨팅 연구를 산업화 방향으로 확장하면서, 데이터센터 인프라와의 연결을 위해 엔비디아의 CUDA-Q 플랫폼을 활용하고 있습니다.
- **디랙(Diraq):** 호주의 양자컴퓨팅 스타트업으로, 기존 실리콘 반도체 공정을 활용해 대규모 양자컴퓨터를 개발하는 것이 핵심 강점입니다. 엔비디아와의 협업을 통해 NVQLink 기반의 GPU-QPU 결합 구조를 구체화하고 있습니다.

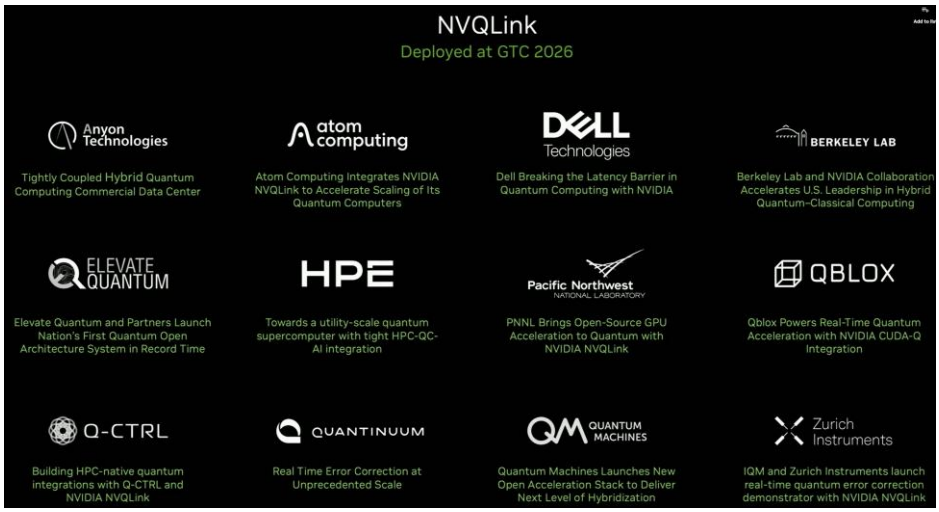
이들 사례에서 중요한 것은 개별 기업의 기술 선택보다 공통적으로 나타나는 방향입니다. 현재 **양자컴퓨팅은 독립적인 시스템으로 발전하기보다, GPU 기반 HPC 와 결합된 형태로 구현되고 있는데요.** 이 흐름 속에서 엔비디아는 NVQLink 와 CUDA-Q 를 통해, 연구소에서나 작동하던 양자컴퓨팅이 실제 데이터센터 인프라 위에서 구현될 수 있도록 환경을 만들어가고 있습니다.

[Data-4] NVQLink: 2025년 말에 처음 소개되었다



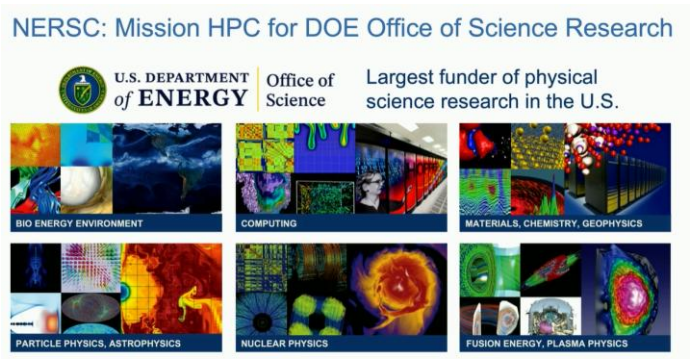
출처: 엔비디아, 토스증권

[Data-5] NVQLink: 2026년 3월 GTC에서 구체적인 적용 사례 공개, 다양한 기업들에 빠르게 퍼지고 있음을 알 수 있었다



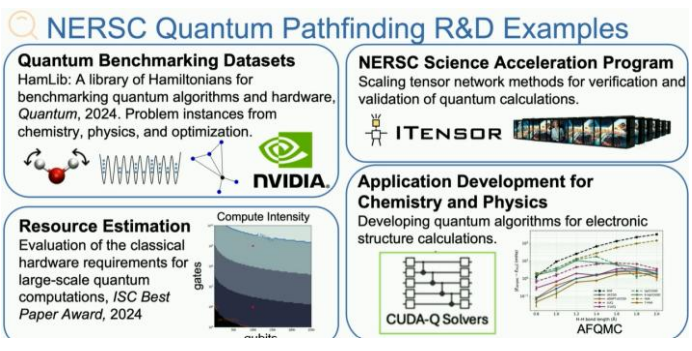
출처: 엔비디아, 토스증권

[Data-6] NERSC: 미국 에너지부 산하 컴퓨팅 연구 기관



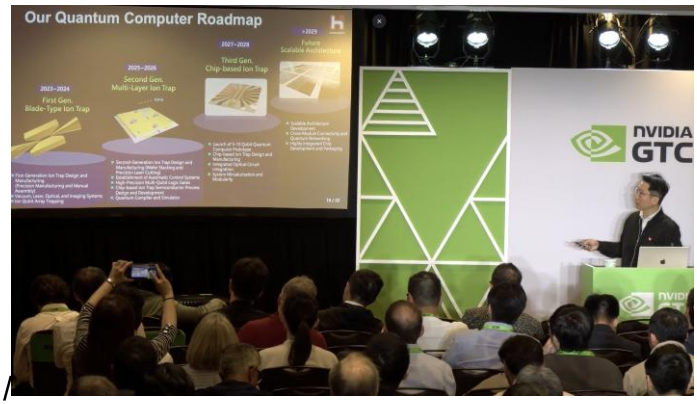
출처: 엔비디아 GTC, 토스증권

[Data-7] NERSC: 엔비디아의 하드웨어 및 소프트웨어가 이미 차세대 양자 컴퓨팅 연구를 가능하게 하고 있다고 발표했다



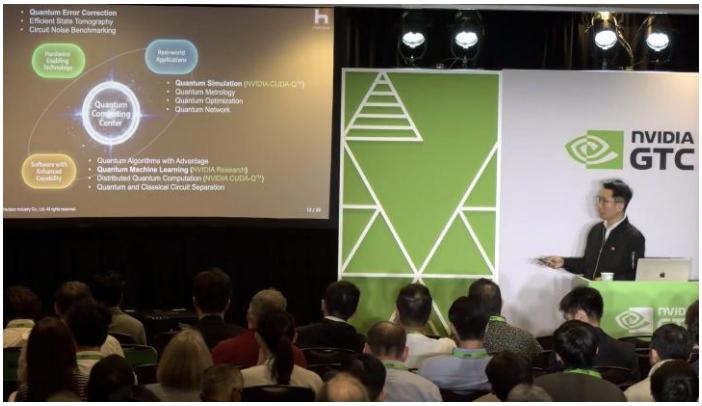
출처: 엔비디아 GTC, 토스증권

[Data-8] 폭스콘: 칩 형식의 3세대 이온트랩 퀀텀 컴퓨터 프로토타입 소개를 앞두고 있다 (2027~28년 목표)



출처: 엔비디아 GTC, 토스증권

[Data-9] 폭스콘: CUDA-Q로 하이브리드 컴퓨팅 연구 가속화 중



출처: 엔비디아 GTC, 토스증권

[Data-10] 퀀티넘: 붐비는 부스의 모습에서 높은 관심을 짐작할 수 있었다

출처: 토스증권

[Data-11] 리게티 컴퓨팅: 부스에서 관계자로부터 리게티 컴퓨팅의 QPU인 Cepheus에 대한 설명을 들었다
(해당 사진에서 보이는 Cepheus-1-108Q는 현재 아마존의 퀀텀 컴퓨팅 서비스인 Amazon Braket에서 이용 가능하다)

출처: 엔비디아, 토스증권

양자 생태계 접근 위해, CUDA 플랫폼 전략을 한 번 더

GTC 에 다녀와 한국에서 리포트를 준비하던 중, 4 월 14 일 세계 양자의 날에 엔비디아가 발표한 아이징(Ising) 소식이 눈에 들어왔습니다.

아이징은 양자컴퓨팅의 핵심 과제인 프로세서 보정(Calibration)과 오류 정정(Error Correction)을 자동화하기 위해 개발된 AI 모델입니다.

지금까지 양자컴퓨터는 온도나 진동 같은 미세한 환경 변화에도 오류가 발생하는, 매우 민감한 시스템이었습니다. 그래서 전문가가 수시로 보정 작업을 수행해야 했는데요. 아이징은 AI 로 이 작업을 대체하려는 시도입니다. 발표 이후 주요 연구기관과 기업을 중심으로 초기 테스트가 빠르게 진행됐고, 일부 환경에서는 실제로 개선되는 모습을 보였습니다. 양자컴퓨팅 상용화의 핵심 병목이 해결될 수 있겠다는 기대감은 주식시장에도 반영되어, 양자컴퓨팅 기업들의 주가가 급등하기도 했습니다.

엔비디아는 CUDA 로 AI 생태계를 장악했던 것처럼, 양자컴퓨터에서도 비슷한 전략을 펼치고 있습니다. QPU 와 GPU 를 연결하는 NVQLink, 개발 환경을 제공하는 CUDA-Q, 핵심 병목인 오류 정정과 보정을 자동화하는 Ising 등 양자 하드웨어가 작동하기 위해 반드시 필요한 인프라를 제공하겠다는 것입니다.

[Data-12] 엔비디아의 양자컴퓨팅 생태계

출처: 토스증권

2. 양자컴퓨팅과 엔비디아 GPU 협업의 의미

GTC 에서 엔비디아가 강조한 차세대 컴퓨팅 구조에서, GPU 와 양자컴퓨터는 협력 관계입니다. GPU 로 세계 최정상의 입지를 굳힌 엔비디아가 자신의 컨퍼런스에서 대대적으로 양자컴퓨팅을 선보인 이유도 여기에 있습니다. 차세대 컴퓨팅 시장에서도 칩 메이커들의 핵심 파트너 자리를 선점하고, 락인 효과까지 기대하는 것입니다.

물론 엔비디아가 제시하는 이 구조는 아직 초기 단계입니다. GPU 같은 고전 컴퓨팅[2]이 양자컴퓨팅 인프라에서 어떤 비중을 차지하게 될지는 향후 기술 발전 방향과 경쟁 환경에 따라 달라질 수 있습니다.

엔비디아 GPU, QPU 의 고질적인 문제를 해결한다

엔비디아가 제시하는 차세대 컴퓨팅 구조는 이렇습니다.

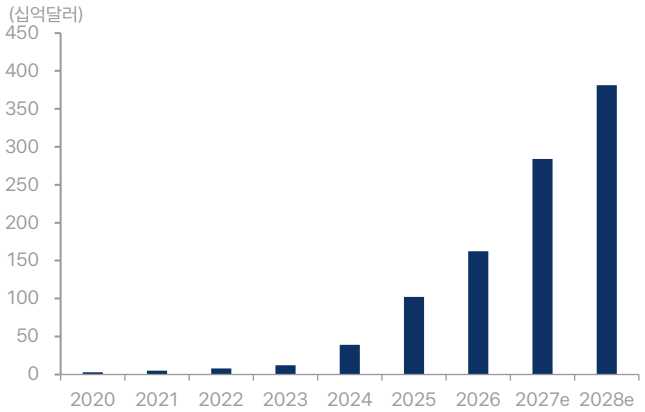
- **GPU** 가 데이터 전처리/후처리/제어를 담당하고,
- **양자컴퓨터**는 특정 고난도 연산을 수행하며,
- **NVQLink** 는 이 둘을 하나의 시스템으로 연결한다.

이 구조에서는 GPU 가 양자컴퓨터의 고질적인 결함을 메웁니다. 양자컴퓨팅의 오류 정정은 고전 연산으로 처리되어야 하는데, 이 과정에서 GPU 가 중요한 역할을 하는 것입니다. GPU 가 양자컴퓨팅 인프라 역할을 하는 셈입니다. 양자 알고리즘 개발, 시스템 보정, 시뮬레이션 과정에서도 마찬가지입니다.

장기적으로 양자컴퓨팅이 확장되면 GPU 의 역할이 더욱 커질 수 있습니다. 큐비트[3] 수가 증가할수록 오류 정정 및 제어에 필요한 고전 연산량도 함께 증가하기 때문입니다.

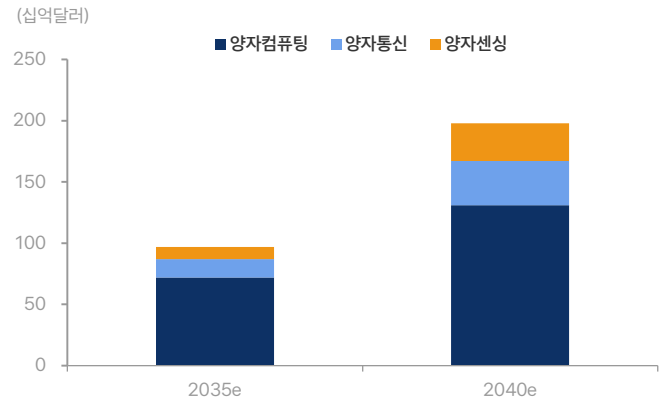
즉, 엔비디아는 AI 에 이어 양자컴퓨팅에서도 인프라 공급자로서의 역할이 유지 또는 강화될 가능성이 높습니다.

[Data-13] 엔비디아 데이터센터 매출 추이: 엔비디아의 성장은 AI 기술 수요 덕분



출처: 블룸버그, 토스증권

[Data-14] 양자 기술 경제적 효과 10년 내 2조달러 전망[4]



출처: 맥킨지, 토스증권

2 고전 컴퓨팅(Classical Computing)은 양자 역학 원리를 이용하지 않는 모든 컴퓨팅으로 CPU, GPU 모두 해당된다.
3 양자컴퓨터의 연산 단위.
4 2035년에 최대 1,000억달러 매출, 타 산업 추가 매출과 비용 감소 효과 고려한 경제적 효과는 최대 2조달러 규모 전망.

3. 왜 지금 양자컴퓨팅일까?

AI 기술의 발전과 상용화 속도가 너무나도 빠르기 때문입니다.

AI 의 쓰임이 텍스트와 이미지 처리를 넘어 물리적 세계로 확장되면서, 기존 컴퓨팅으로는 해결하기 어려운 문제들이 실제로 드러나기 시작했습니다. 신약 개발, 소재 설계, 에너지 문제 같은 영역에서는 분자와 전자의 상호작용을 계산해야 하는데, 이때 경우의 수가 기하급수적으로 증가합니다. 더 빠른 컴퓨터를 만드는 것만으로는 해결되지 않는, 계산 방식 자체의 한계에 가깝습니다.

동시에 GPU 기반 HPC 와 데이터센터 인프라가 충분히 발전하면서, 양자컴퓨팅을 보조하고 제어할 수 있는 고전 컴퓨팅 기반이 갖춰지기 시작했습니다. 과거에는 이론으로만 가능했던 구조가 실제 시스템으로 구현될 수 있는 조건이 만들어진 것입니다.

3-a 양자컴퓨팅을 쉽게 설명한다면

고전 컴퓨터의 연산 단위가 비트라면, 양자컴퓨터의 연산 단위는 큐비트입니다. 비트는 반드시 0 과 1 중 하나의 값을 지니지만, 큐비트는 0 과 1 의 가능성을 동시에 품고 있습니다. 이러한 큐비트의 특성을 '중첩'이라 부릅니다.

큐비트의 중첩 특성을 활용한 양자컴퓨팅의 계산 과정은 고전 컴퓨터의 그것과는 완전히 다릅니다. 고전 컴퓨터는 문제 하나를 풀고 나서 그 다음 문제를 풀고, 이런 식으로 계산합니다. 하지만 양자컴퓨터는 여러 가능성을 동시에 가진 채로 정답에 가까운 상태가 어디일지 확률을 높여갑니다.

GPU 에 익숙한 투자자 입장에서는 '병렬 처리'와 비슷하게 느껴질 수 있는데요. 분명한 차이가 있습니다.

GPU 가 '여러 사람이 각각 하나의 문제를 푸는 방식'이라면, 양자컴퓨터는 '한 사람이 여러 가능성을 동시에 고려하는 방식'에 가깝습니다. GPU 는 연산 속도를 높이는 기술이고, 양자컴퓨팅은 문제 푸는 방식 자체를 달리하는 기술인 것입니다.

3-b. 양자컴퓨팅, 어떤 계산을 할 수 있나?

양자컴퓨팅이 주로 언급되는 영역은 계산 복잡도가 매우 높은 분야입니다.

- **신약 개발이 대표적입니다.** 특정 분자가 인체 내에서 어떻게 반응하는지 전자 수준까지 계산해야 하는데, 현재 슈퍼컴퓨터로도 수십 년이 걸리거나 아예 불가능합니다. 양자컴퓨팅은 중첩을 통해 여러 분자를 동시에 평가하는 방식으로 이 문제에 접근할 수 있습니다. 알츠하이머나 특정 암 치료제 개발, 신약 개발 시간 단축, 부작용 예측 등에서 활용이 기대됩니다.
- **경우의 수가 너무 많은 물류, 항공, 우주의 난제를 풀어줄 것도 기대됩니다.** 내비게이션 앱이 가끔 이상한 경로를 안내하는 이유는, '근삿값을 빠르게 찾는 방식'으로 처리되기 때문입니다. 계산이 너무 오래 걸려 모든 경우의 수를 다 따질 수 없는 거죠. 대형 물류, 위성 관측, 우주선 경로 설계도 구조적으로 비슷한 문제입니다. 다양한 제약 조건을 동시에 만족하는 최적의 조합을 찾아야 한다는 점에서, 양자컴퓨팅의 계산 방식이 유용하게 쓰일 수 있습니다.

3-c. 핵심은 '성능 향상'이 아니라 '잠재력'

2026년 현재 기준, 양자컴퓨팅이 모든 것을 대체하는 범용 기술이 되긴 어렵습니다. 문서 처리, 데이터 분석, AI 학습과 추론 같은 일반적인 연산에서는 여전히 GPU 기반 HPC가 훨씬 효율적입니다.

미래의 차세대 컴퓨팅이 양자컴퓨팅에만 국한되는 것도 아닙니다. 뉴로모픽 컴퓨팅, 광자 컴퓨팅, 메모리 중심 컴퓨팅 등 다양한 접근 방식이 동시에 발전하고 있습니다.

그럼에도 양자컴퓨팅이 주목받는 이유는 새로운 계산 패러다임을 제시하기 때문입니다. 다른 차세대 접근 방식들이 기존 컴퓨팅의 효율을 개선하는 방향으로 발전하는 것과 달리, 양자컴퓨팅은 기존 방식으로는 해결이 어려웠던 문제에 접근합니다. 양자컴퓨팅의 핵심은 '성능 향상'이 아니라, 풀 수 있는 문제의 종류를 늘리는 '잠재력'에 있습니다.

마치며

양자컴퓨팅은 여전히 쉽지 않은 기술이고, 증명된 상업적 성과도 많지 않습니다.

다만 지금까지의 흐름을 보면, 이 기술이 최근 빠르게 산업 구조 안으로 들어오고 있다는 점은 분명해 보입니다.

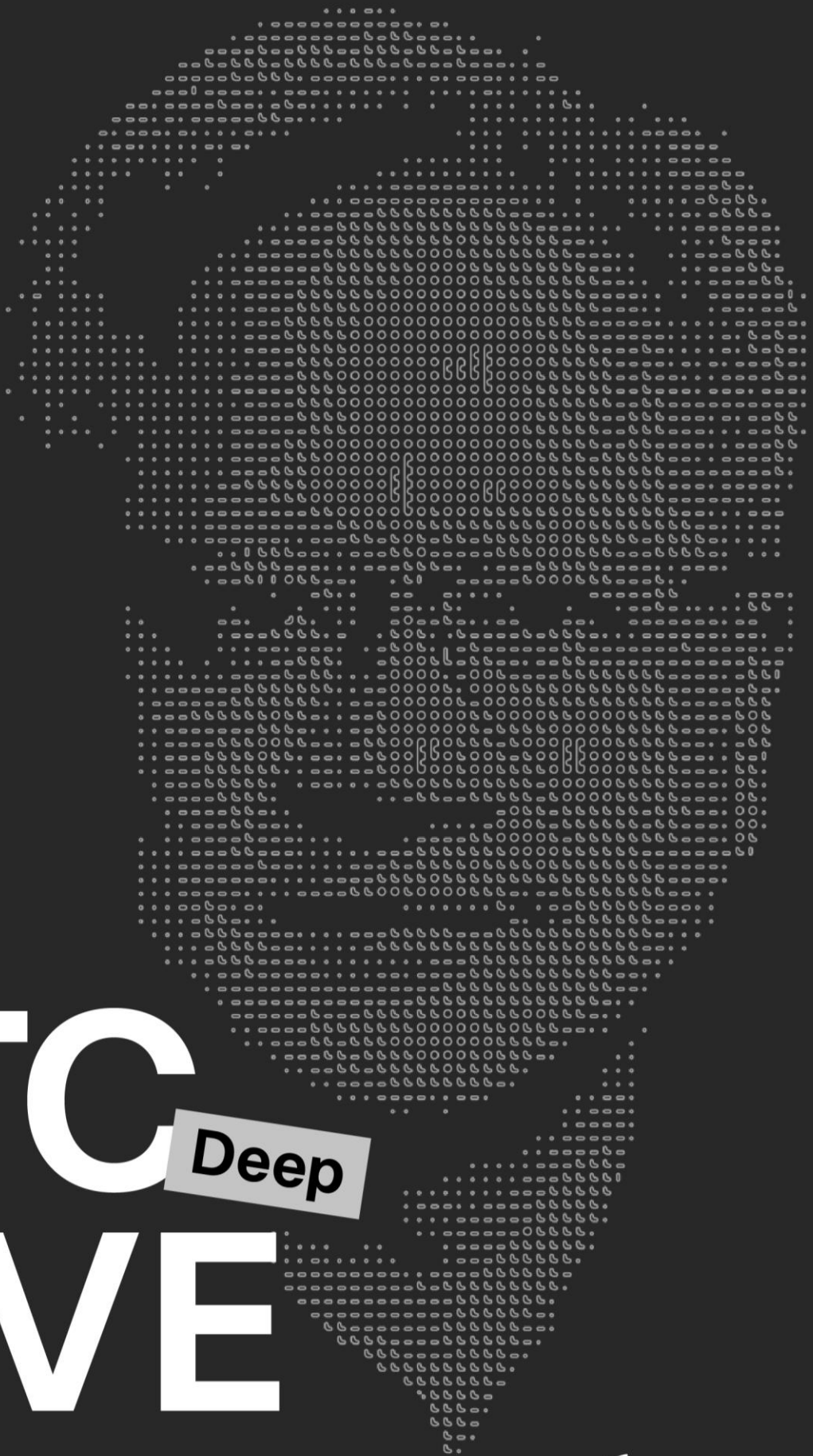
AI 기술의 빠른 발전은 기존 컴퓨팅의 한계를 드러냈고, 새로운 문제 해결 방식이 필요해졌습니다. 실험실에서나 가치가 있는 것으로 여겨지던 양자컴퓨팅이 다시 한번 주목받는 이유입니다. 양자컴퓨팅의 단점은 오류가 쉽게 발생한다는 것인데요. 엔비디아는 그동안 쌓아온 노하우로 그 단점을 보완하려 합니다. GPU, NVQLink, CUDA-Q, Ising 까지 엔비디아의 인프라/플랫폼이 양자컴퓨팅의 상용화를 앞당길 수 있을지 시장 또한 주목하고 있습니다.

투자에서 중요한 것은 기술 그 자체보다, 언제 '연구'에서 '수익'으로 넘어가는지 판단하는 기준입니다. 이 기준에 대해서는 별도의 산업 리포트에서 보다 구체적으로 다뤄볼 예정이니 많은 관심 부탁드립니다.

Publisher
토스증권 리서치센터

Analyst
한상원

Date
2026.05.12



GTC Deep DIVE

03 추론, 그리고 AI 에이전트

토스증권

Intro. 우리가 추론의 왕이다

GTC 에서 가장 주목받는 이벤트는 언제나 젠슨 황 엔비디아 CEO 의 기조연설(Keynote)입니다. 이번 GTC 2026 에서 그가 AI 와 엔비디아 다음으로 가장 많이 꺼낸 단어는 추론(Inference)과 에이전트(Agent)였습니다.

"Now it's in the field of inference(지금부터는 추론의 시대입니다)."

그는 이 한 문장으로 학습(Training)의 시대가 지나고 추론의 시대가 시작됐음을 선언했습니다. AI 가 읽고, 생각하고, 실행하는 모든 과정에서 추론이 훨씬 더 많이 필요해진다는 것입니다. 발표 도중 "Inference King!"을 외치며 챔피언 벨트 이미지 아래서 양손을 번쩍 들어올린 장면은, 지금 엔비디아가 어디에 집중하고 있는지를 분명하게 보여줬습니다.

학습의 시대에 AI 성능을 높이는 방법은 비교적 단순했습니다. 더 좋은 GPU, 더 많은 데이터, 더 긴 학습 시간. 특히 성능 좋은 GPU 를 만들고 확보하는 것이 AI 성능을 높이는 가장 결정적인 요인이었습니다.

추론의 시대에는 성능 좋은 GPU 를 갖추는 것만으로는 부족합니다. GPU 를 효율적으로 잘 쓰는 방법을 찾아야 합니다.

추론에서는 왜 효율이 중요할까요? 효율을 위해 GPU 성능 외에 뭐가 더 필요한 걸까요? 엔비디아는 어떻게 스스로 추론왕이라고 자신할 수 있었을까요? 이번 리포트에는 이러한 질문에 대한 답이 담겨 있습니다.

GTC 2026 현장에서 보고 듣고 느낀 점을 토대로 추론의 시대를 이해하는 3 가지 키워드를 추렸습니다. 1)오케스트레이션(Orchestration), 2)베라 루빈(Vera Rubin), 3)네모클로(NemoClaw)입니다. 이 3 가지가 지금 왜 중요한지, 지금부터 함께 살펴보겠습니다.

[Data-1] 추론의 왕(Inference King) 엔비디아, 그리고 젠슨 황 CEO

출처: 엔비디아, 토스증권

1. 오케스트레이션(Orchestration)

AI가 에이전트가 된다는 건 사용자가 AI에게 기대하는 것 자체가 달라진다는 의미입니다. 엔비디아는 챗봇과 에이전트의 차이를 아래와 같이 설명합니다.

- **챗봇(Chatbot):** 프롬프트가 주어지면 → 응답을 생성
- **에이전트(Agent):** 작업 목표가 주어지면 → 맥락을 파악하여 → 무엇을 할지 결정하고 → 계획을 세워서 → 결과를 확인하고 → 이를 반복하며 만족스러운 결과가 나왔을 때 종료

에이전트는 작업 목표 달성을 위해 스스로 움직이는 존재입니다. AI 에이전트는 요청 하나에 대해서도 여러 차례 판단과 실행을 반복해야 하고, 따라서 추론 능력의 중요성이 더욱 높아집니다.
학습(Training)이 AI 모델을 만드는 과정이라면, 추론은 학습된 모델을 잘 사용하는 과정인데요. 쉽게 말해, 학습은 공부하는 단계, 추론은 시험을 보는 단계와 비슷합니다.

AI 산업의 흐름이 학습에서 추론으로 확장되면서 '효율성'이 새로운 평가 지표로 자리 잡고 있습니다. 시험을 볼 때 정해진 시간 안에 답을 찾아내는 것이 중요하듯, AI 역시 같은 시간이나 같은 전력을 썼을 때 더 많은 일을 해낼 수 있느냐가 중요해지고 있다는 뜻입니다.

바꿔 말해, 지금까지 AI가 추론하는 과정은 다소 비효율적이었다는 건데요. 엔비디아의 엔지니어들은 그 이유에 대해 다음과 같이 설명합니다.

- AI가 질문에 답하는 과정은 2단계로 나눌 수 있다.
- 1단계 프리필(Prefill): 답변을 준비하는 단계로, 입력된 질문과 앞선 맥락을 한꺼번에 처리해야 하기 때문에 대규모 연산 능력이 필요하다.
- 2단계 디코드(Decode): 답변을 생성하는 단계로, 과거에 저장해둔 정보를 빠르게 불러와야 하기 때문에 메모리 능력을 필요로 한다.

문제는 이 두 단계에서 필요로 하는 능력이 완전히 다르다는 것입니다. 연산 능력을 필요로 하는 프리필과 메모리 능력을 필요로 하는 디코드가 같은 GPU에서 처리됨으로써 비효율이 발생하거든요.

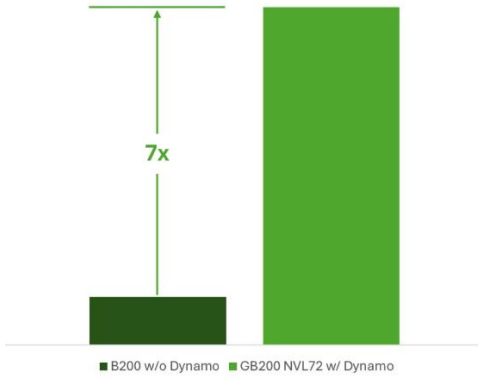
이러한 비효율 문제를 해결하기 위해 등장한 개념이 바로 오케스트레이션(Orchestration)입니다. AI 에이전트가 효율적으로 돌아가도록 전체 흐름을 설계하고 역할을 나누며, 실행까지 조율하고 관리하는 것을 의미합니다. 엔비디아는 이러한 역할을 하는 소프트웨어 다이나모(Dynamo) 1.0을 내놓기도 했습니다.

다이나모는 프리필과 디코드를 분리했습니다. 프리필은 연산 성능이 강한 GPU 클러스터에, 디코드는 메모리 대역폭이 넓은 하드웨어에 맡긴 것입니다. 젠슨 황 CEO는 이를 통해 추론 성능이 최대 7배 향상되었다고 밝혔습니다.

[Data-2] AI가 효율적으로 돌아가도록 전체 흐름을 설계하는 오케스트레이션

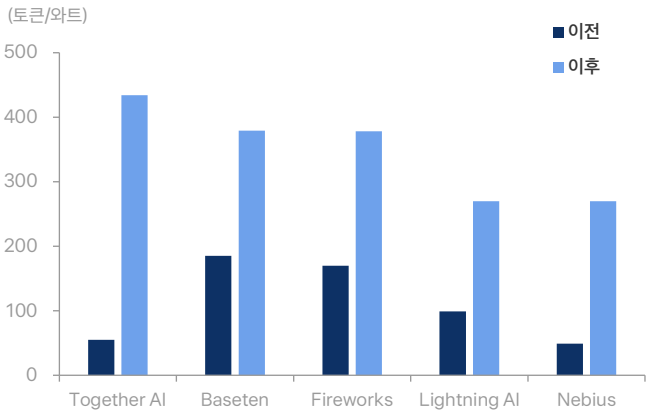
출처: AI ESPRESSO

[Data-3] 다이나모(Dynamo)를 활용해 토큰 처리량 크게 증가



출처: 엔비디아

[Data-4] 주요 추론 서비스의 성능 개선 효과

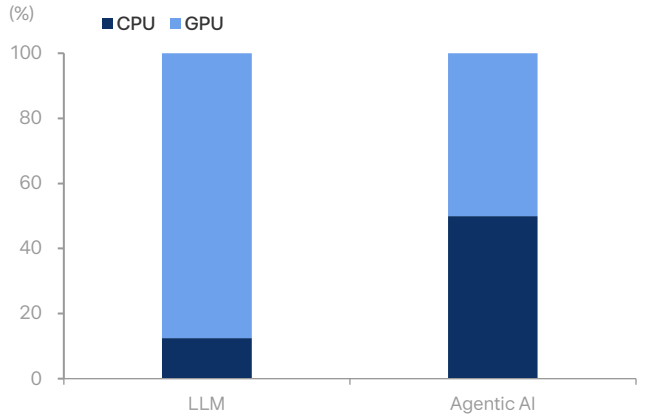


출처: 엔비디아, 토스증권

오케스트레이션은 CPU(Central Processing Unit)에서 주로 실행됩니다. 단순한 작업을 동시에 여러 개 처리하는 데 특화된 GPU 와 달리, CPU 는 복잡한 작업 하나를 빨리 끝내는 데 특화되어 있는데요. 이러한 강점을 살려, CPU 는 전체 흐름을 조율하며 어떤 GPU 에게 언제 무슨 일을 맡길지 결정합니다. 레스토랑에 비유하면 GPU 는 요리사 1 천명, CPU 는 1 명의 뛰어난 셰프인 셈입니다. 스스로 판단해야 하는 AI 에이전트가 더 많은 CPU 를 필요로 하는 이유입니다. 이런 점이 주목받으며 올해 들어 CPU 기업들의 주가가 크게 오르기도 했습니다.

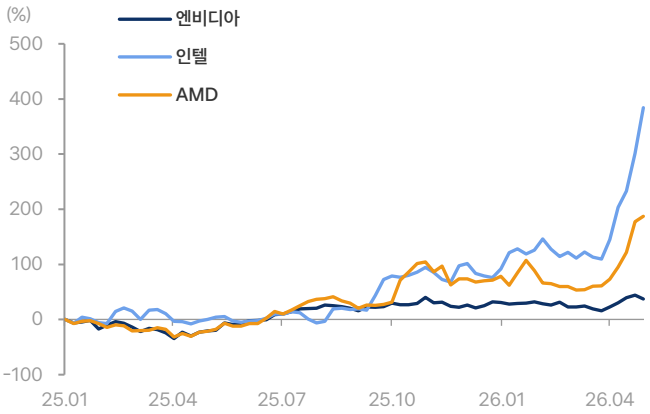
추론의 시대에, 높은 효율성은 수익성과도 직결되어 있습니다. 기업 입장에서 하드웨어를 더 효율적으로 운영해 처리 비용을 낮추면 그만큼 수익의 기회도 커지기 때문입니다. 이 과정에서 오케스트레이션의 중요성은 더욱 높아지고 있습니다.

[Data-5] AI 에이전트에서는 더 많은 CPU가 필요하다



출처: TrendForce, 토스증권

[Data-6] 2026년 들어 CPU 기업들 주가가 더 크게 상승했다



출처: Bloomberg, 토스증권

2. 베라 루빈(Vera Rubin)

젠슨 황 CEO 의 기조연설이 진행되는 중, 잠시 장내가 조용해졌습니다. 엔비디아의 GPU 아키텍처가 하나씩 순서대로 등장하며 다음 세대 제품 베라 루빈(Vera Rubin)에 대한 기대가 고조되던 순간이었습니다. 젠슨 황 CEO 는 전작 블랙웰(Blackwell)에 비해 베라 루빈의 성능이 35 배 좋아졌다고 밝혔습니다.

여기서 '35 배'라는 수치가 의미하는 바는 '효율 개선'입니다. 같은 비용(전력)으로 35 배나 더 많은 일(토큰 생성)을 해낼 수 있다는 거죠. 그는 베라 루빈을 "AI 에이전트를 위한 플랫폼[5]"이라 표현하기도 했는데요. 앞서 살펴봤듯이 AI 에이전트는 하나의 요청에 대해서도 더 많은 추론을 반복적으로 수행해야 하고, 그만큼 효율이 중요하기 때문입니다.

베라 루빈은 루빈 GPU, 베라 CPU, 그록(Groq) LPU 등 7 개의 칩으로 구성된 통합 플랫폼입니다.

이전에는 각 분야에서 가장 좋은 칩과 부품을 사서 조합하는 방식이었는데요. 하지만 최고의 선수들이 모였다고 해서 최고의 팀이 되는 것은 아니듯, 성능이 뛰어난 칩을 쓰더라도 칩끼리 데이터를 주고받는 과정에서 병목이 발생할 수 있습니다.

반면, 베라 루빈은 처음부터 7 개의 칩이 함께 최적화될 수 있도록 설계됐습니다. 젠슨 황 CEO 가 밝힌 '35 배 향상'도 단일 GPU 의 성능 개선에 의한 것이 아니라, 플랫폼 관점에서 여러 칩의 조화가 이뤄졌기에 가능한 것입니다. 실제로 베라 루빈 내의 역할 분담은 명확합니다. 예를 들어 루빈 GPU 가 연산 집약적인 프리필 단계를 처리하면, 그록 LPU 가 메모리 집약적인 디코드 단계를 맡는 식입니다. 앞서 설명한 오케스트레이션이 소프트웨어로 효율을 높이는 거라면, 베라 루빈은 하드웨어 설계 단계부터 최적화했습니다.

여기에는 엔비디아의 전략적 판단도 녹아 있습니다. 7 개의 칩이 맞물려 최적화된 상황에서는 개별 칩을 바꾸는 순간 최적화 설계 역시 깨집니다. 가령, 다른 GPU 로 바꿀 경우 NVLink 스위치와 연동이 끊기고, NVLink 를 교체하면 전체 대역폭 설계가 흔들리는 거죠. 부품 하나를 바꾸려면 사실상 전체를 바꿔야 하는 구조입니다.

즉, 엔비디아는 베라 루빈 같은 하드웨어 플랫폼을 통해 락인(Lock-in) 효과를 노리고 있는 것입니다.

[Data-7] 베라 루빈을 구성하는 7개의 칩

이름	특징
베라 CPU	시스템 전체를 조율하고 데이터 이동을 담당하는 '두뇌' 역할
루빈 GPU	AI 학습과 추론 같은 무거운 계산을 처리하는 초고성능 연산 엔진
NVLink 6 스위치	여러 GPU를 하나처럼 묶어 동시에 작동하게 만드는 초고속 연결 장치
CX9 Super NIC	GPU끼리 데이터를 빠르고 지연 없이 주고받게 해주는 네트워크 인터페이스
BP4 DPU	네트워크, 보안, 스토리지 같은 부가 작업을 대신 처리해 GPU 부담을 줄여주는 칩
스펙트럼-X 이더넷 스위치	데이터센터 전체를 효율적으로 연결해 전력과 안정성을 크게 개선하는 네트워크 스위치
그록 3 LPU	메모리 속도를 극대화해 AI 언어 처리에 특화된, GPU와 다른 방식의 연산 칩

출처: 토스증권

5 "Compute demand continues to grow exponentially. And now Vera Rubin, architected for every phase of Agentic AI(컴퓨팅 수요는 계속해서 기하급수적으로 증가하고 있습니다. 그리고 베라 루빈은 AI 에이전트의 모든 단계를 위해 설계된 아키텍처입니다)."

베라 루빈의 사례는 소프트웨어뿐 아니라 하드웨어 영역에서도 AI 에이전트를 위한 효율화가 이뤄지고 있다는 것을 보여줍니다.

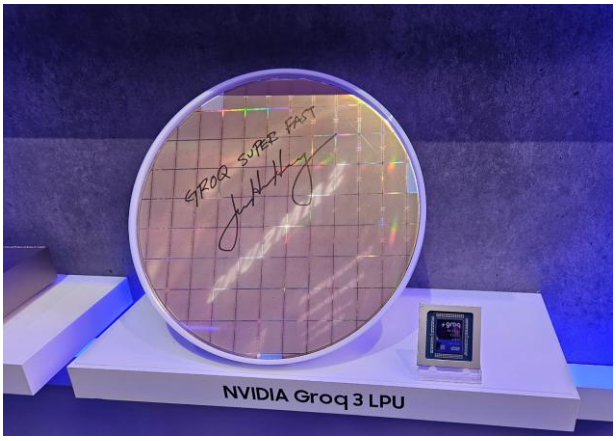
사람들의 관심이나 AI 에 대한 투자도 이미 AI 밸류체인 전반으로 넓어지고 있는데요. 이러한 사실은 GTC 2026 기업 부스에서도 확인할 수 있었습니다.

- **삼성전자의 부스**는 GTC 2026 에서 가장 많은 인파가 몰린 곳 중 하나였습니다. 젠슨 황 CEO 는 기조연설에서 삼성전자가 그록 LPU 를 생산할 거라며 "I want to thank Samsung(삼성에 감사한다)."이라 말하기도 했는데요. 삼성전자 부스에는 젠슨 황 CEO 의 사인이 담긴 그록 LPU 와 함께, 약 50% 성능 개선이 기대되는 차세대 HBM4E(7 세대)도 전시되어 있었습니다.
- **SK 하이닉스의 부스** 역시 많은 인파가 몰렸습니다. SK 하이닉스의 부스는 젠슨 황 CEO 와 최태원 SK 그룹 회장이 함께 방문한 것으로 알려졌는데요. 둘은 HBM 을 포함해 앞으로의 협력 로드맵에 대해서도 논의한 것으로 전해졌습니다. 국내 반도체 기업들이 핵심 부품을 공급하는 것을 넘어 엔비디아 생태계의 전략적 파트너로 주목받고 있다는 사실을 다시 한번 확인할 수 있었습니다.

[Data-8] 삼성전자 부스를 방문한 토스증권 애널리스트

출처: 토스증권

[Data-9] 삼성전자는 베라 루빈 플랫폼의 그록 LPU를 생산

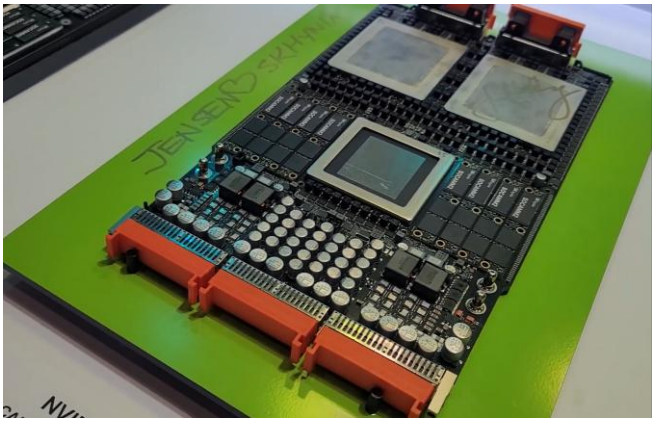


출처: 토스증권

[Data-10] SK하이닉스 부스를 방문한 토스증권 애널리스트

출처: 토스증권

[Data-11] SK하이닉스의 SOCAMM2와 HBM4가 사용된 베라 루빈 Superchip



출처: 토스증권

3. 네모클로(NemoClaw)

젠슨 황 CEO는 오픈클로(OpenClaw)의 성과에 대해 여러 차례 언급했습니다. "인류 역사상 가장 인기 있는 오픈소스 프로젝트"라 소개하며, 리눅스(Linux)가 30년 걸린 일을 단 몇 주 만에 달성했다고 강조했습니다. 이야기는 자연스레 엔비디아의 네모클로(NemoClaw)로 이어졌습니다. 그의 말에 따르면, 네모클로를 통해 누구나 안전하게 AI 에이전트를 만들 수 있게 됩니다.

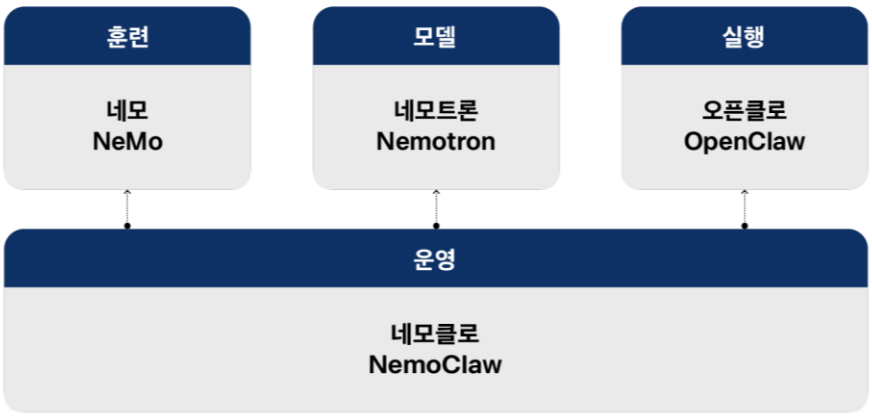
AI 에이전트를 만들고 운영하는 과정은, 네모(NeMo)부터 시작되어 네모트론(Nemotron), 오픈클로(OpenClaw), 네모클로까지 이어지는데요. 엔비디아는 이러한 과정 전체를 묶음으로 제공하려 합니다.

AI 에이전트를 구축하고 운영하는 일은, 회사에서 유능한 인재를 육성해 효과적으로 일하도록 돕는 과정과 비슷한데요. 지금부터 각 단계를 하나씩 자세히 설명드리겠습니다.

- **네모(NeMo, 직원 교육 프로그램):** AI 모델을 개발하고 학습시키기 위한 엔비디아의 프레임워크입니다. 'NeMo'라는 이름은 신경망 모듈을 의미하는 'Neural Modules'에서 유래했는데요. AI의 여러 기능을 모듈화해 필요에 따라 조합할 수 있습니다. 회사에서 유능한 직원을 키우기 위해 필요한 여러 교육을 수행하는 시스템과 비슷합니다.
- **네모트론(Nemotron, 네모로 훈련된 전문가 팀):** '네모'를 활용해 개발된 대형 언어모델(LLM)입니다. 네모트론은 다양한 목적에 최적화되어 있어, '한 명의 유능한 직원'이라기보다는 '각자 강점을 지닌 전문가 팀'에 가깝습니다. 이 '전문가 팀'은 AI 에이전트의 임무 수행에서 두뇌 역할을 합니다.
- **오픈클로(OpenClaw, 업무 도구 및 인프라):** AI 에이전트가 실제로 행동할 수 있게 해주는 실행 엔진입니다. 웹 검색, 파일 읽기 및 쓰기, 외부 시스템 연동 등 실질적인 작업을 수행할 수 있도록 돕습니다. 아무리 유능한 직원이라도 업무할 때 컴퓨터가 꼭 필요하듯, 오픈클로 또한 AI 에이전트 작동을 위한 필수 도구입니다.
- **네모클로(NemoClaw, 이 모든 것이 작동하는 회사 시스템):** 위의 3가지가 안정적으로 작동할 수 있도록 관리하는 플랫폼입니다. 네모클로의 핵심은 '안전한 운영'인데요. AI 에이전트가 스스로 업무를 수행하는 과정에서 보안 관련 문제들이 발생할 수 있기 때문입니다. '통제가 가능할까?'와 같은 우려는 기업이 AI 에이전트 도입을 주저하는 가장 큰 이유인데요. 네모클로는 바로 이 불안을 해소하며 AI 에이전트를 작동시킵니다. 젠슨 황 CEO 또한 '어떤 조건에서, 어디까지 허용되며, 어떻게 기록되고 통제되는가'를 관리하는 네모클로의 역할을 강조[6]했습니다.

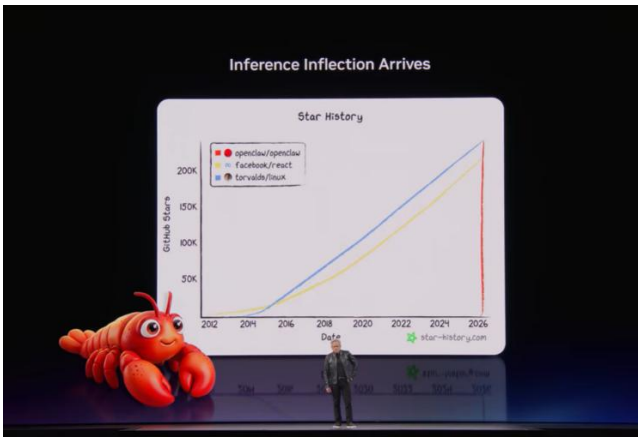
6 We worked with Peter, we took some of the world's best security and computing experts, and we worked with Peter to make OpenClaw enterprise secure and enterprise private capable. And we call that this is our NVIDIA OpenClaw reference for open NemoClaw.(우리는 세계 최고 수준의 보안 및 컴퓨팅 전문가들과 협력해 OpenClaw를 엔터프라이즈 환경에서도 보안 기능을 갖추도록 만들었습니다. 그것이 바로 엔비디아 OpenClaw 레퍼런스, 즉 NemoClaw입니다.)

[Data-12] 엔비디아 생태계 내에서 AI 에이전트가 만들어지고 운영되는 과정



출처: 토스증권

[Data-13] 가파른 성장을 보이는 AI 에이전트 플랫폼 오픈클로

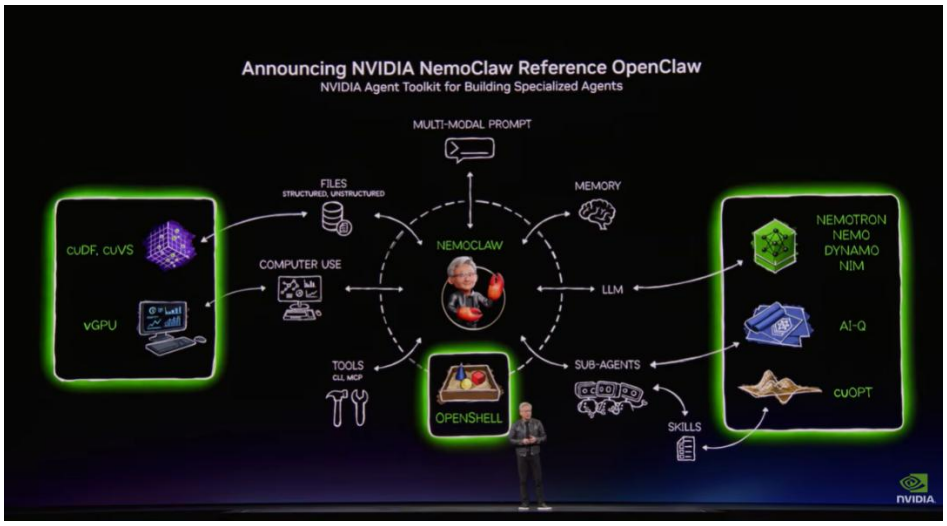


출처: 엔비디아, 토스증권

[Data-14] 오픈클로의 로고인 랍스터는 클로징 영상에도 주인공으로 등장

출처: 엔비디아, 토스증권

[Data-15] GTC 2026 기조연설에서 네모클로를 소개하는 젠슨 황 CEO



출처: 엔비디아, 토스증권

네모클로 얘기를 들으면서 떠오른 것이 있습니다. 바로 CUDA 입니다.

2006 년 엔비디아가 공개한 CUDA(Compute Unified Device Architecture)는 그래픽 처리에만 쓰이던 GPU 를 계산, 시뮬레이션, AI 학습에도 활용[7]할 수 있게 해주는 개발 도구였습니다. 개발자들은 사용성이 좋은 CUDA 로 코드를 짜기 시작했는데요. 문제는 그렇게 작성된 코드가 엔비디아의 GPU 에서만 돌아간다는 점이었습니다. 즉, GPU 를 바꾸는 순간 그동안 쌓아둔 코드가 무용지물이 되는 거죠.

이처럼 CUDA 는 엔비디아 GPU 에 대한 강한 락인 효과를 만들어냈습니다. 네모클로 역시 AI 에이전트를 개발하는 전 과정에서 락인 효과를 불러올 수 있습니다.

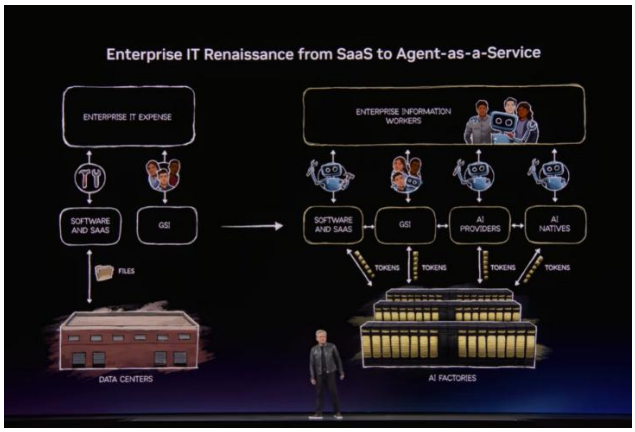
"It's in NVIDIA's interest to have great AI everywhere(좋은 AI 가 많아지는 것이 엔비디아에도 좋은 일이다)."

GTC 2026 의 한 세션에서 엔비디아 관계자가 한 발언[8]입니다. 그의 말대로 엔비디아는 네모트론 모델을 오픈소스로 공개해 무료로 사용할 수 있게 했습니다. 이번 GTC 2026 에서는 개발자들이 오픈클로로 AI 에이전트 실행 환경을 직접 구축해볼 수 있는 'Build-a-Claw at GTC' 행사가 열리기도 했습니다. 과거 CUDA 때처럼, 진입장벽을 낮춰 생태계 안으로 끌어들이고 거기에 익숙해지면 다른 환경으로 이동하기 어렵게 만드는 방식입니다. 이번엔 그 범위가 GPU 에서 하드웨어 전반, 그리고 AI 에이전트를 개발하고 운영하는 플랫폼으로 넓어졌을 뿐입니다.

[Data-16] 개발자들이 오픈클로를 체험할 수 있는 행사 'Build-a-Claw at GTC'

출처: 엔비디아, 토스증권

[Data-17] Agent-as-a-Service 시대의 도래



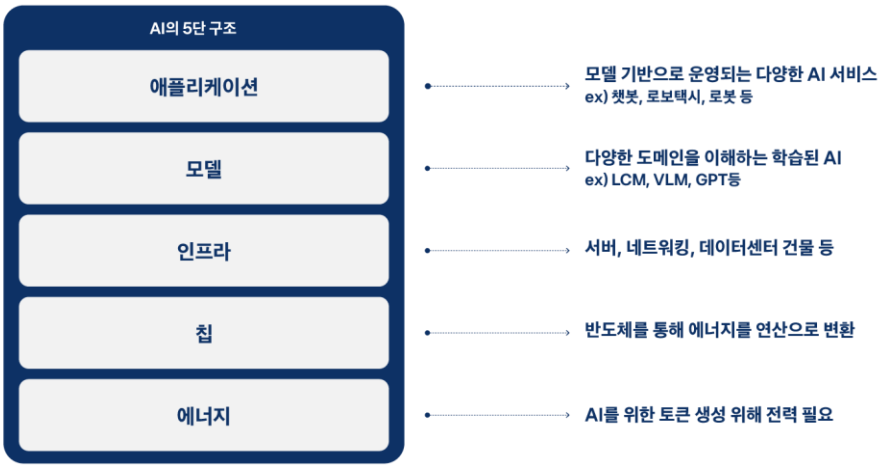
출처: 엔비디아, 토스증권

7 엔비디아가 주최하는 GTC 역시 처음에는 엔비디아의 주력 제품인 GPU가 그래픽 처리뿐 아니라 다방면으로 활용될 수 있음을 홍보하는 자리였다.

8 'The State of Open Source AI' 세션에 참석한 엔비디아의 Jonathan Cohen(NVIDIA VP of Applied Research)은 엔비디아 스스로의 이익을 위해 이런 일을 하는 것이라고 생각하지 않으며, AI 생태계 전체에 좋은 일을 하는 것이라 말했다.

엔비디아는 AI 를 5 단 케이크에 비유해왔습니다. AI 가 에너지, 칩, 인프라, 모델, 애플리케이션의 5 단 구조로 이뤄져 있다는 설명인데요. 예를 들어 성공적인 애플리케이션(서비스)은 더 좋은 모델을 필요로 하고, 모델을 돌리려면 더 많은 인프라와 칩이 필요하고, 그만큼 많은 에너지를 소비한다는 뜻입니다. 즉, 엔비디아는 각각의 단계를 따로 떼어 보지 않고, '하나의 케이크'로 보고 있는 건데요. 그것 자체가 CUDA 때부터 네모클로까지, 락인 효과를 기반으로 성장하고 있는 엔비디아답다고 느꼈습니다.

[Data-18] 엔비디아는 AI를 '에너지, 칩, 인프라, 모델, 애플리케이션' 등으로 구성된 5단 케이크에 비유해왔다



출처: 엔비디아, 토스증권

마치며

AI 산업의 무게중심은 이제 학습에서 추론으로 넘어왔습니다. GTC 2026 은 이를 공식적으로 확인할 수 있는 자리였습니다. 행사를 주최한 엔비디아는 추론 시대에도 자신들이 경쟁력을 가질 거라는 사실을 재차 강조했습니다.

핵심은 더 이상 AI 의 성능이 개별 칩 단위에서 결정되지 않는다는 것입니다. 여러 칩을 시스템 내에서 어떻게 최적화할 것인가, 그 시스템을 누가 더 효율적으로 관리하느냐가 경쟁력이 될 것입니다. 이번 리포트에서 살펴본 3 가지 키워드 오케스트레이션, 베라 루빈, 네모클로는 모두 위와 같은 메시지를 담고 있습니다.

투자자로서 AI 산업에 관심을 갖다보면 때론 마음이 흔들리는 뉴스를 마주하게 됩니다. 얼마 전 챗 GPT 의 성장성 둔화를 두고 시장의 우려가 높아졌던 것처럼요. 그럴 땐 큰 흐름에서 본질을 다시 떠올려보곤 합니다. 인터넷 초창기에 앞서 나가던 포털 사이트가 사라졌다고 해서 인터넷 시대가 끝난 것은 아니었습니다.

AI 를 활용한 서비스들은 앞으로도 계속 생겨날 것이고, 그중 일부는 실패하고 사라질 것입니다. 하지만 AI 라는 메가 트렌드가 끝나는 것은 아닙니다. 거대한 흐름 속에서 나타나는 산업의 변화를 이해하고 투자 기회를 발견하려는 노력은 앞으로도 필요할 것입니다.

토스증권 리서치센터 역시 앞으로도 함께 하겠습니다. 감사합니다.

Compliance Note

- 당사는 발간일 기준 지난 1 년간 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 3 자에게 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.