NVIDIA가 여는 AI 생태계의 미래

GTC 2026 중심으로

March 2026

들어가며

젠슨 황 CEO는 GTC 2026에서 AI 산업이 학습의 시대를 넘어, 추론과 에이전트 중심의 새로운 컴퓨팅 시대로 진입했음을 선언했다.

2026년 3월 16일(월)~19일(목)까지 미국 새너제이에서 열린 'GPU Technology Conference 2026(GTC 2026)'은 단순한 차세대 GPU 발표 행사가 아니었다. 젠슨 황 CEO의 기조연설 무대에 오른 것은 개별 칩의 성능 그래프가 아니라, AI를 '어떻게 대량으로 생산하고, 안정적으로 운영할 것인가'에 대한 설계도였다. GTC 2026은 NVIDIA가 더 이상 고성능 반도체 기업에 머물지 않고, AI 시대의 산업 인프라 자체를 재정의하는 수준에 이르렀음을 명확히 보여준 자리였다.

GTC 2026에서 NVIDIA가 반복적으로 강조한 키워드는 'AI Factory'다. 이는 GPU 몇 개를 더 빠르게 만드는 이야기가 아니다. 수천 개의 GPU를 하나의 공장처럼 묶어, 현실의 업무와 산업 현장에서 AI를 지속적으로 생산하고 운영하는 체계를 뜻한다. 차세대 AI 가속기 Vera Rubin, 랙 단위로 설계된 시스템, 그리고 이를 안정적으로 운용하기 위한 소프트웨어와 보안까지. 기조연설의 메시지는 한 방향을 가리켰다. "AI의 병목은 더 이상 연산이 아니라 인프라와 운영에 있다"는 점이다.

또한 GTC 2026은 생성형 AI 이후의 다음 단계가 이미 현실로 진입했음을 보여줬다. 대화형 AI를 넘어 스스로 계획하고 실행하는 AI 에이전트(Agentic AI), 가상 공간에서 학습한 지능이 실제 로봇과 공정을 움직이는 피지컬 AI(Physical AI)는 이론상의 개념이 아니라 구체적인 플랫폼과 제품 형태로 제시되었다.

이러한 변화 속에서 GTC의 본질도 달라졌다. 더 이상 개발자만을 위한 기술 행사가 아니다. 제조, 통신, 금융, 에너지, 로보틱스 기업들이 대거 참여해 실제 적용 사례를 공유했고, NVIDIA의 기술 로드맵은 글로벌 기업들의 설비 투자와 AI 전략을 가늠하는 사실상의 기준점으로 작용하고 있다. GTC에서 제시된 방향이 1~2년 뒤 산업 전반의 표준으로 자리 잡는 흐름이 이미 굳어지고 있음을 현장에서 확인할 수 있었다.

이에 본 보고서는 GTC 2026 기조연설과 NVIDIA의 최근 동향을 중심으로, NVIDIA가 그리는 AI 시대의 구조적 변화를 정리하고자 한다. 단기적인 칩 성능 경쟁을 넘어, AI 인프라 · 운영 · 응용 전반에서 산업의 판이 어떻게 바뀌고 있는지, 그리고 이러한 변화가 국내 기업의 전략과 의사결정에 어떤 시사점을 주는지를 짚어보고자 한다.

Contents

I. GTC 2026, 세상을 바꾸는 AI 축제 **03**

1. 10가지 질문으로 알아보는 GTC 완벽 가이드 04

2. GTC 2026 기조연설 핵심 메시지 11

II. NVIDIA, 칩을 넘어 AI의 미래를 설계하다 **14**

1. 5단 케이크로 보는 AI 산업 스택 15

2. 세간의 시선 집중, 차세대 AI 가속기 공개 16

3. 칩을 넘어 인프라로, 전방위 확장 드라이브 20

4. 응용 분야까지 진출, 풀스택 기업으로 도약 22

III. 요약 및 시사점 **24**

[Appendix] **27**

GTC 2026 전시관 테마 27

I

GTC 2026,
세상을 바꾸는 AI 축제

01. 10가지 질문으로 알아보는 GTC 완벽 가이드

PART 1. GTC 기초 다지기: “GTC가 정확히 뭔가요?”

Q1: GTC는 어떤 행사인가요?

A: GTC(GPU Technology Conference)는 세계적인 AI 컴퓨팅 기업인 **NVIDIA**가 주최하는 세계 최대 규모의 **AI 및 가속 컴퓨팅 컨퍼런스**입니다. 전 세계에서 개발자, 연구원, 기업인 등 약 **30만 명 이상**이 온·오프라인으로 모여 AI로 인류의 난제를 해결하는 방법을 논의하는 자리로 거듭났습니다. 단순한 기술 발표회를 넘어 피지컬 AI, AI 에이전트 등 차세대 AI 혁신을 탐구하는 거대한 장이라고 이해할 수 있습니다.

GTC 개요

구분	상세
개최 기간	2026 3/16 ~ 3/19
장소	San Jose, CA(San Jose McEnery Convention Center)
개최 연혁	2009년부터 시작, 올해 17회째
참가 규모	전시 기업 약 400개사, 현장 참석자 25,000여 명, 온라인 참석 300,000여 명 세션 1,000개, 연사 2,000명 (2025년 기준 역대 최대 규모 달성)
전시 및 발표 주제	에이전트 AI, AR/VR, 컴퓨터 비전, 데이터센터, 데이터 사이언스, 엣지컴퓨팅 등 13개 분야
실습 교육 및 인증	개발자와 엔지니어 대상으로 9개의 종일 워크숍, 60개 이상의 실습 랩, 현장 인증 기회 제공 자기주도형 코스, 교육자 프로그램, 1:1 교육 컨설팅 제공으로 AI 인프라 구축에 필요한 인력 양성 지원

Q2: 다른 IT 행사(CES, MWC)와 무엇이 다른가요?

A: CES가 “올해 어떤 멋진 TV나 가전제품이 나올까?"를 보여주는 '완성품 전시장'이고, MWC가 "새로운 스마트폰과 통신 기술은 뭘까?"를 다루는 '모바일 박람회'라면, **GTC는 그 기기들 속에서 돌아가는 '두뇌(AI와 GPU)를 만드는 설계도 발표회'**라고 할 수 있습니다. CES는 전 세계 미디어와 일반 소비자에게 "우리 제품 사세요!"라고 홍보하는 성격이 강하지만, **GTC는 "우리 기술로 이런 미래를 함께 만듭시다!"라고 개발자와 기업들에게 제안하는 자리**입니다. 그래서 GTC에는 제품 전시뿐만 아니라 수백 개의 심도 있는 기술 강연(Technical Sessions)과 실습 워크숍이 포함되어 있어, 현업 엔지니어와 연구원들에게는 '성지'와 같은 곳입니다.

최근에는 GTC의 영향력이 CES를 압도한다는 평가도 나옵니다. CES에서 발표되는 자율주행차, 스마트 홈, 로봇 기술들이 사실은 GTC에서 발표된 NVIDIA의 AI 칩과 소프트웨어를 기반으로 만들어지기 때문입니다. 즉, GTC에서 방향이 정해지면, 1~2년 뒤 CES에서 그 결과물이 제품으로 등장하는 흐름입니다. **산업의 '뿌리'와 '원천 기술'을 보고 싶다면 GTC를, 그 결과물인 '열매'를 보고 싶다면 CES를 주목하면 됩니다.**

[GTC · CES · MWC 비교]

	GTC	CES	MWC
주관	NVIDIA	CTA	GSMA
주요 영역	AI 인프라, 칩셋, 딥러닝	IT 완제품, 디바이스	통신 장비, 네트워크, 서비스
비즈니스 모델	B2D(개발자) / B2B	B2C(소비자)	B2B(기업)
핵심 키워드	가속 컴퓨팅, AI 공장	AI, 디지털 헬스, 모빌리티 등	6G, 엣지 컴퓨팅 등
기술 지향점	얼마나 강력한 지능을 만들 수 있는가 (Intelligence)	우리 삶이 기술로 어떻게 변하는가 (Experience)	세상이 어떻게 하나로 연결되는가 (Connectivity)

자료: NVIDIA GTC, NVIDIA Newsroom, 언론종합, 삼일PwC경영연구원

01. 10가지 질문으로 알아보는 GTC 완벽 가이드

Q3: GTC가 겉보기엔 NVIDIA의 화려한 신제품 발표회 같지만, 실질적으로는 글로벌 AI 산업의 향방을 결정짓는 '기술 로드맵의 가이드북' 역할을 한다고 볼 수 있을까요?

NVIDIA의 GTC는 이제 단순한 신제품 발표회가 아니라, NVIDIA가 AI 시대를 어떤 구조로 설계하고 있는지 확인할 수 있는 사실상의 'AI 로드맵 공개 행사'로 자리 잡았습니다. 그간의 GTC를 통해 NVIDIA가 추구하는 AI 혁신의 방향이 세 단계의 구조적 변화를 거치며 진화하고 있음을 확인할 수 있습니다:

(1단계) 차세대 GPU와 AI 가속기 등 핵심 칩 성능 강화 중심 → (2단계) 데이터센터가 단순 컴퓨팅 센터가 아니라, 에너지 및 인프라(네트워크) 통합으로 AI를 실제로 만들어내는 산업적 생산설비로 전환 → (3단계) 모델·애플리케이션까지 포함하는 전체 AI 스택으로 확장. 즉, **NVIDIA는 에너지- 칩 - 인프라 - 모델 – 애플리케이션을 하나의 생태계로 통합하는 전략을 강조하고 있습니다.**

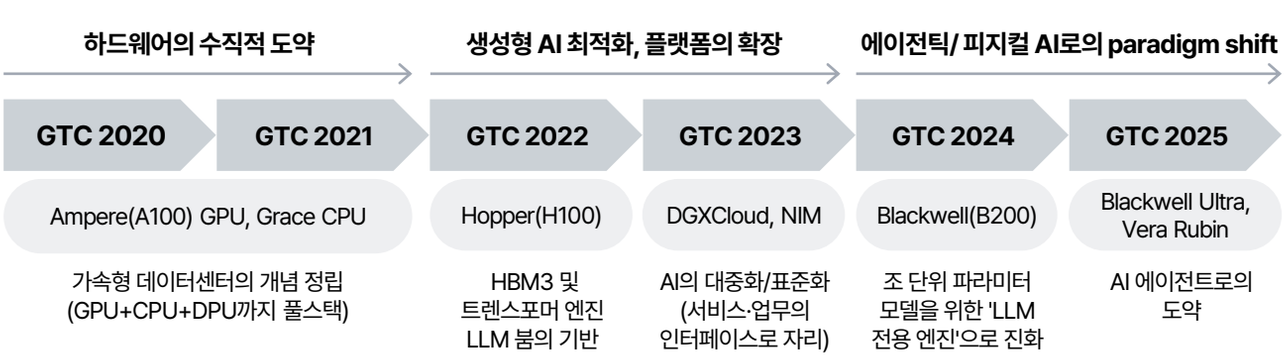
NVIDIA의 3단계 AI 로드맵

Q4: 그렇다면 지난 5년간 NVIDIA가 GTC에서 선보인 기술 변화에 대해 구체적으로 알려주세요

A: 지난 5년간 NVIDIA는 GTC를 통해 단순 부품 설계를 넘어 전 세계 AI 인프라를 하나로 통합하는 거대 지능 시스템을 구축해왔음을 보여주었습니다. 연산 칩의 성능을 극대화하는 단계를 지나, 수만 개의 장치를 초고속망으로 연결해 하나의 거대한 슈퍼컴퓨터처럼 작동하게 만들고, 복잡한 설치 과정 없이 누구나 AI를 즉시 사용할 수 있는 표준화된 서비스 체계를 구축했습니다. 이제는 화면 속 AI를 넘어 가상 세계에서 학습시킨 지능을 현실의 로봇과 자동화 공정에 그대로 이식하며, AI가 우리 삶의 물리적 공간까지 움직이게 하는 피지컬 AI 시대의 기반을 완성했습니다.

(*뒷 페이지 상세 참고)

GTC 2020~2025 기술 패러다임 변화



GTC 2020~2026

차세대 AI 인프라 전환을 이끈 핵심 기술 및 생태계 변화

시기	핵심 메시지	주요 발표 기술	업계 파장·의미
GTC 2020	"CPU → GPU 중심의 데이터센터로 변하는 변곡점"	Ampere 아키텍처 A100 GPU, DGX A100	• 서버 패러다임 전환: 기존 CPU 수십 대 성능을 A100 단일 시스템이 대체하며 기업의 GPU 클러스터 도입 가속 • 美 에너지부 산하 Argonne 연구소 등, 공공연구에서 팬데믹 연구/시뮬레이션에 곧바로 투입 → GPU 가속 클러스터가 과학계의 기본 장비로 확산
GTC 2021	"데이터센터=CPU+GPU+DPU" 3요소 플랫폼화 선언	Grace(Arm) CPU, BlueField DPU, Omniverse, 자율주행 칩 'Drive Atlan'	• Grace CPU 공개: GPU, DPU에 더해 3칩(CPU+GPU+DPU)으로 데이터센터를 재설계 → 초거대 모델 학습을 위한 CPU-GPU 메모리 결합 수요 자극 • 인텔과AMD가 장악했던 서버용 CPU 시장에 균열을 냄 • 스위스 국립슈퍼컴퓨팅센터(CSCS)와 미국 로스앨러모스 국립연구소(LANL)가 Grace 기반 슈퍼컴퓨터 구축 발표 →Arm 서버 채택 명분 강화
GTC 2022	"대형 언어모델(LLM) 시대, 학습 속도를 한 단계 끌어올린다"	H100(Hopper, Grace Hopper 슈퍼칩), Omniverse, Isaac·DRIVE 스택 강화	• H100 GPU의 트랜스포머 엔진이 LLM 학습 효율을 극대화(연산 속도 2배, 메모리 사용량 50% 절감)하는 표준 기술로 자리잡음 → 전세계 GPU 쟁탈전 촉발 • H100이 업계 최초 HBM3 탑재 AI 가속기로 상용화되며, SK 하이닉스가 HBM3 '초기 단독' 공급을 선언·양산 → HBM3 대중화·메모리 대역폭 경쟁 본격화 • Omniverse/로보틱스 스택 확대로 산업용 디지털 트윈 시장이 본격 형성됨
GTC 2023	"누구나 브라우저에서 AI 슈퍼컴을 쓸 수 있다"	DGX Cloud, 마이크로서비스형 AI 'NIM(NVIDIA Inference Microservices)'	• DGX Cloud로 AIaaS(AI as a Service) 표준화 진행 → 기업들의 CapEx→OpEx 전환 가속 • AWS, Azure, OCI 등 클라우드 기업들의 GPU 임대/클라우드형 AI 팩토리 모델 정착 • AI 서버 비중 확대로 HBM 수요가 연간 두 자릿수 성장 → 국내 메모리 기업의 설비투자·용량 증설 압박 증대
GTC 2024	"차세대 AI 컴퓨팅의 비용과 속도를 표준화"	B200/B300 (Blackwell), GB200 (Grace+Blackwell) 슈퍼칩	• 실시간·초대형 추론의 성능, 전력 효율 균형을 새로운 표준으로 제시→ 실서비스 추론 규모·속도가 Hopper 대비 수십 배 개선됨 • Blackwell 세대 대량 수요와 함께 HBM3E로 전환 가속(삼성, 마이크론이 2024~2025에 HBM3E 본격 진입) → 공급사 다변화 + CoWoS 병목 완화 시도 병행[1]
GTC 2025	"스스로 판단하고 실행하는 에이전트 AI로의 도약"	B300(Blackwell Ultra), Vera Rubin 플랫폼 프리뷰(Vera CPU+Rubin GPU)	• 무중단 유지보수·고가용성 AI 팩토리 운영을 염두에 둔 '플랫폼, 아키텍처, 오퍼레이션' 동시 진화 트렌드 확립 • 단순 답변을 넘어 직접 업무를 수행하는 AI 에이전트 구축이 기업들의 최우선 과제로 부상 • Rubin 세대의 HBM4 예고: 메모리 용량이 50% 이상 늘어난 신규 칩셋 발표로 차세대 HBM 규격 전쟁 본격화 → 국내 HBM 업체 HBM4 전환 투자 가속(적층·열·본딩 등 공정 고도화 경쟁 격화)
GTC 2026	"에이전틱·피지컬 AI의 현실화 가속"	Vera Rubin, Feynman(로드맵 프리뷰), NemoClaw, 피지컬 AI 플랫폼	• 단순 GPU 성능 도입 역량보다 이를 안정적으로 운용할 수 있는 네트워크·전력·냉각 역량이 경쟁력으로 부상 • 업계의 AI Capex는 결국 반도체를 넘어 전력, 냉각, 네트워크, 스토리지, 소프트웨어, 시뮬레이션, 로보틱스로 확장될 전망

1) CoWoS(Chip on Wafer on Substrate)는 특수 패키징 기술로, 최첨단 AI 칩을 만들고 싶어도 반도체 조립 공정이 너무 복잡해서 물량이 딸리는 현상을 해결하려는 시도를 의미

01. 10가지 질문으로 알아보는 GTC 완벽 가이드

Q5: GTC 2026에서 가장 돋보이는 세션은 무엇이었나요?

A: 단연 젠슨 황(Jensen Huang) CEO의 기조연설(Keynote)입니다. 그는 매년 상징적인 가죽 자켓을 입고 무대에 올라 "AI의 미래는 이 방향입니다"라는 칩과 소프트웨어 로드맵을 발표합니다. 그가 던지는 메시지가 전 세계 Tech 기업들의 사업 방향을 바꾸기 때문에, 업계에서는 'AI의 복잡한 미래를 정리해 주는 족집게 강의'로 통합니다.

이번 기조연설에서는 NVIDIA는 크게 컴퓨팅 가속화, AI 팩토리, 개방형 모델, AI 에이전트 시스템, 피지컬 AI를 다뤘는데요. 차세대 GPU인 Rubin을 기반으로 한 새로운 AI 시스템을 선보였으며, NemoClaw(AI 에이전트), IsaaC(피지컬 AI 플랫폼), 로보틱스 등을 제시하고, 업계 비즈니스 전망 및 파트너십을 강조했습니다.

*기조연설에 대한 자세한 내용은 본 보고서 'GTC 2026 기조연설 핵심 메시지'에서 확인할 수 있습니다.

GTC 2026 기조연설

Q6: GTC 2026에서 가장 화두였던 주제 혹은 메시지가 무엇이었는지 짚어 주세요!

A: NVIDIA GTC 2026의 핵심은 AI 산업의 무게중심이 학습 중심의 초기 투자 국면에서 추론·에이전트·피지컬 AI 중심의 대규모 상용화 국면으로 이동하고 있음을 선언한 데 있습니다. NVIDIA는 Vera Rubin, Dynamo, NemoClaw, DSX, Cosmos/Isaac 등을 통해 GPU 단품이 아니라 AI 공장 전체와 그 운영체제, 그리고 로봇·산업 자동화까지 포괄하는 전방위 플랫폼 사업자로 자리매김하려는 전략을 분명히 보였습니다. 이는 향후 AI 경쟁이 모델 성능 자체보다 토큰당 비용, 전력 효율, 운영 체제, 보안이 결합된 시스템 경쟁으로 재편될 가능성을 시사합니다.

GTC2026가 보여준 AI 산업의 패러다임 변화

경쟁 단위	인프라 단위	응용 단위
모델 → 인프라	칩 → 팩토리	챗봇 → 에이전트/피지컬 AI

① **"추론의 시대가 왔다":** AI 산업이 더 이상 대형 모델을 학습시키는 초기 투자 단계에만 머무르지 않고, 실제 서비스 단계에서 수억 명의 사용자가 AI를 호출하고 에이전트가 지속적으로 동작하는 대규모 상용화 단계로 진입하고 있음을 보여주었습니다.

② **AI 경제성의 핵심 지표는 '토큰[1]당 비용'과 '전력당 처리량':** 젠슨 황 CEO는 "지능 토큰이 새로운 화폐이며, AI 팩토리는 이를 생산하는 인프라"라고 주장했는데요. 이는 데이터센터가 단순 서버 집합이 아니라, '토큰'이라는 지능 생산물을 대량으로 생산하는 설비 산업으로 바뀌고 있음을 말합니다. 앞으로 AI 경쟁력은 모델의 성능뿐만 아니라 전력·냉각·네트워크·운영 소프트웨어까지 포함한 전체 시스템 최적화 역량에서 결정됨을 시사합니다.

③ **AI 상용화 중심축이 에이전트/피지컬 AI로 이동**: NVIDIA는 에이전트 AI 운영 체제 공개를 통해 "AI는 더 이상 답변 생성 도구가 아니라 실제 업무를 수행하는 SW 행위자"라는 메시지를 던졌습니다. 또한, 산업·휴머노이드·수술 로봇 생태계 전반의 파트너를 제시하면서, 피지컬 AI가 더 이상 미래 개념이 아니라 실제 산업 시스템 도입 단계로 진입하고 있음을 강조했습니다.

1) 토큰(Token)은 AI의 입력/출력 단위로, 토큰이 많을 수록 연산량이 커짐. NVIDIA가 말하는 AI 팩토리는 일종의 토큰 생산 공장으로, AI 산업은 결국 "얼마나 많은 토큰을 얼마나 싸게 처리하느냐"의 경쟁으로 감 (*GTC에서 다룬 용어 - token cost: 토큰 하나 처리하는 데 드는 비용; tokens per watt: 전력 1와트당 얼마나 효율적으로 토큰을 처리하는지)

자료: NVIDIA GTC, NVIDIA YouTube, 언론종합, 삼일PwC경영연구원

01. 10가지 질문으로 알아보는 GTC 완벽 가이드

PART 2. 산업계의 영향력: "GTC가 왜 세상을 흔드나요?"

Q7: GTC 발표 내용이 기술주 시장이나 경제에도 영향을 주나요?

A: 네, 그렇습니다. **GTC는 이제 단순한 기술 발표회를 넘어 전 세계 AI 경제의 향방을 결정짓는 '글로벌 경제 지표'로 자리매김했습니다**. NVIDIA가 발표하는 차세대 칩과 소프트웨어 로드맵은 전 세계 빅테크 기업들의 설비 투자(CapEx) 규모를 결정하는 기준이 되며, 이는 곧바로 기술주 시장의 흐름을 좌우하는 강력한 촉매제가 됩니다. 특히 GTC는 하드웨어 판매를 넘어 AI가 실제 산업 현장에 어떻게 침투할지를 보여주는 예고편 역할을 하기에, 투자자와 기업들은 GTC를 통해 미래 산업의 성장 엔진이 얼마나 강력해질지를 확인합니다.

특히 최근 GTC의 핵심 화두는 AI가 화면을 벗어나 물리적 세계로 나오는 피지컬 AI의 부상입니다. 이번 GTC 2026에서도 강조된 피지컬 AI는 디지털 트윈과 로보틱스 기술을 결합해 제조, 물류, 스마트 팩토리 등 실제 산업 현장의 패러다임을 근본적으로 바꾸고 있습니다. 이러한 변화는 AI가 단순한 대화형 도구에서 벗어나 실질적인 생산성 혁명을 일으키는 단계에 진입했음을 의미하며, 이는 관련 제조 및 자동화 기업들에 대한 대규모 투자와 제품 개발을 촉진하며 실물 경제에 강력한 파급력을 미치고 있습니다.

Q8: GTC 참석 기업들을 보니 Johnson & Johnson, L'Oréal, Mercedes-Benz 같은 비(非)IT 기업들도 많더라고요. 반도체 행사인 GTC에 이들이 왜 이렇게 적극적으로 참여하고 발표까지 하는 건가요?

A: GTC가 더 이상 '칩'만 보여주는 자리가 아니라, AI를 통해 비즈니스 문제를 해결하는 '실전 무대'가 되었기 때문입니다. 주요 글로벌 기업들이 GTC를 찾는 이유는 크게 세 가지 관점으로 이해할 수 있습니다:

① 글로벌 선도사들은 GTC를 통해 금융, 의료, 유통 등 각 도메인 특화 AI를 도입하여 운영 효율을 극대화한 구체적인 성공 사례를 공유하며 **산업별 AI 전환의 표준 레퍼런스를 제시**합니다.

예시) 소매판매 전문기업 Lowe's는 자사 매장의 운영 최적화를 위해 NVIDIA의 Omniverse 솔루션을 기반으로 디지털 트윈을 구축한 사례를 발표. 제조업 선도사 Siemens는 산업용 메타버스 구현과 제조 분야의 디지털 전환을 가속화하기 위한 협력(Siemens Xcelerator – NVIDIA Omniverse) 및 최신 기술을 선보임

② GTC는 NVIDIA의 컴퓨팅 파워를 매개로 전 세계 빅테크와 제조, 서비스 기업들이 **결속하는 거대 생태계(Ecosystem)의 구심점**입니다. 기업들은 NVIDIA의 풀스택 솔루션(CUDA, NIM, Omniverse 등)을 자사의 서비스와 결합하여 독점적인 경쟁력을 확보하고자 하며, 단순한 부품 구매 관계를 넘어 소프트웨어 최적화부터 공동 마케팅에 이르는 전략적 연맹이 GTC에서 체결되는 것입니다. 이는 기업들이 NVIDIA의 기술 표준에 올라탐으로써 글로벌 시장에서의 생존과 확장을 꾀하는 고도의 비즈니스 네트워킹 과정으로 이해할 수 있습니다.

③ 특히 **최근 GTC에서는 디지털 트윈과 로보틱스를 결합한 물리적 AI 기술이 강조**되면서 제조와 물류 등 전통적 산업 현장의 패러다임을 근본적으로 혁신하려는 제조 거인들의 참여가 두드러지고 있습니다. 결과적으로 GTC는 단순한 칩 발표회를 넘어 전 세계 산업의 운영체제가 AI로 교체되는 로드맵을 확인하고 실행하는 전략적 요충지 역할을 수행하고 있습니다.

결론적으로, 오늘날 GTC는 NVIDIA CEO 젠슨 황의 말처럼 **"100조 달러 규모의 전 세계 산업이 한 방에 모이는 자리"**가 되었습니다. 비 IT 기업들에게 GTC는 기술을 배우는 곳이자, 자신들이 AI 시대의 선두주자임을 전 세계에 선포하는 브랜딩의 장이기도 합니다.

01. 10가지 질문으로 알아보는 GTC 완벽 가이드

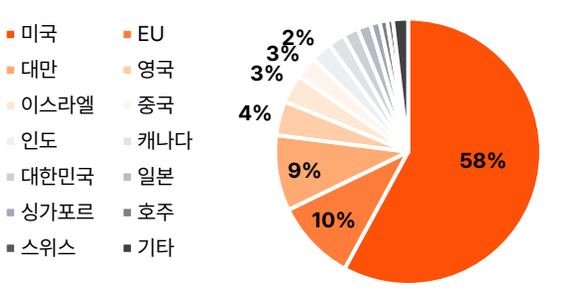
Q9: GTC 2026에 참여하는 기업들의 현황이 궁금해요. 국내 기업들은 어떤 곳이 있나요?

* NVIDIA GTC 공식 홈페이지의 'Sponsors and Exhibitors' 정보만을 기반으로 분석한 기업 현황임을 참고

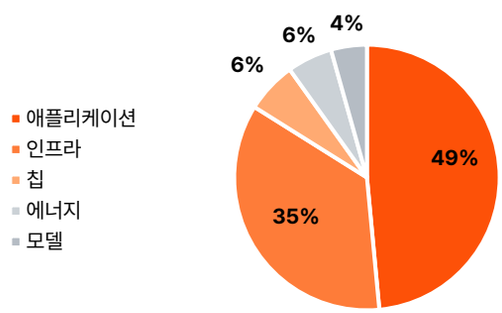
A: GTC 2026 참가 기업 435개사 가운데 **미국이 252개(57.9%)로 절대 우위**를 보였으며, 상위 3개 권역(미국·EU·대만)만으로도 77.0%, 상위 5개(영국·이스라엘 추가)가 84.6%를 차지해 참여가 소수 권역에 강하게 집중되어 있는 경향을 보였습니다. 지역 합산으로 보면 북미 60.0%, 유럽 14.7%, 아시아(이스라엘 제외) 19.1%, 이스라엘 3.4%로, **북미 중심의 AI 생태계 헤게모니**가 분명했으며, **대만(9.2%)은 AI 반도체·서버·패키징 등 공급망·제조 허브로서의 존재감**을 드러냈습니다. 그리고 **이스라엘(3.4%)은 규모 대비 높은 비중으로, 네트워킹·보안·가속 인프라 분야의 스타트업들이 잘 발달**되어있는 것으로 분석됩니다.

또한, **참가 기업의 절반(49%)이 '애플리케이션' 영역, 35%가 '인프라', 나머지가 칩(6%), 에너지(6%), 모델(4%)**로 구성되어 있습니다. 이는 이번 GTC가 단순 칩 전시를 넘어 현업 적용과 운영을 중심으로 한 실전형 생태계 행사로 자리 잡았음을 다시 한번 보여주고 있습니다. 특히 인프라가 35%로 큰 비중을 차지하는 것도 의미가 큰데, 대규모 학습/추론을 뒷받침하는 네트워킹, 광통신, 전력·냉각 설계 등 AI 팩토리 운영 역량이 실제 경쟁의 분기점이 되고 있음을 방증했습니다. 반면 칩이 6%에 그친 것은, 최상위 칩 설계·제조가 소수 플레이어 중심의 고집약 산업으로 수렴됐음을 보여주며, 에너지(6%)는 전력 수요 급증과 열관리 이슈가 AI 확장의 실물 병목임을 드러내고 있습니다.

GTC 2026 참여 기업들의 국가 구분 및 비중



GTC 2026 참여 기업들의 영역 비중



국내 기업의 경우 거대 메모리 기업 제외 시, 개발 플랫폼 및 솔루션 위주의 인프라·애플리케이션 기업이 다수인 것으로 분석됩니다. GTC 공식 홈페이지에는 행사에 참여한 모든 국내 기업들을 명시하지 않아 8개만 확인됩니다. 그러나 그 외에도 현대차, LG디스플레이 등 국내 주요 기업들이 GTC에 참석하여 혁신 사례를 공유하고 세션에 참여하는 등, 자사 기술력을 선보였습니다.

GTC 2026 참여 국내 기업

*GTC 홈페이지 기준

기업명	상세
삼성전자	메모리·파운드리·모바일 등 전자/반도체 종합 기업
SK 하이닉스	DRAM·NAND·HBM 등 메모리 반도체 글로벌 선도 업체
래블업(Lablup)	HPC(고성능 컴퓨팅)/AI 인프라 운영 플랫폼 개발
텐(TEN)	AI 모델의 개발부터 배포, 인프라 관리까지 통합적인 환경을 지원
프렌들리AI(FriendliAI)	생성형 AI 모델의 추론 및 학습 가속화를 위한 인프라 기술 개발
베슬에이아이(VESSL AI)	기업의 ML모델 학습, 최적화, 배포를 돕는 E2E MLOps 플랫폼 기업
파일러(PYLER)	AI 기반의 동영상 맥락 분석 기술을 통해 디지털 광고의 효율성 극대화
슈퍼브에이아이(SUPERB AI)	비전 AI 모델 구축 및 운영을 위한 MLOps 플랫폼 개발

자료: NVIDIA GTC, 삼일PwC경영연구원

01. 10가지 질문으로 알아보는 GTC 완벽 가이드

Q10: 이번 GTC 2026이 글로벌 산업계에 미치는 실질적인 영향력은 어느 정도인가요?

A: GTC 2026은 기존의 텍스트 및 시각 데이터 기반 추론 단계를 넘어 물리적 세계를 직접 제어하는 피지컬 AI 시대의 개막을 알리며, 전 세계 산업계의 운영 체제를 재설계하는 강력한 영향력을 발휘했습니다. NVIDIA는 가상 세계(디지털 트윈)에서 학습시킨 지능을 휴머노이드 로봇과 자율 제조 공정에 즉시 이식하는 표준 기술을 제시함으로써, 단순 챗봇 수준을 넘어 실질적인 노동력과 생산성의 혁명을 주도하고자 합니다. 이는 기업들이 NVIDIA의 인프라 없이는 비즈니스 공정 자체를 기획할 수 없게 만드는 강력한 생태계 장악력으로 이어지고 있습니다.

이 과정을 통해 NVIDIA는 단순한 반도체 제조사를 넘어, 전 지구적 지능과 물리적 노동력을 공급하는 'AI 시대의 운영체제(OS)'이자 필수 공공재 기업으로 탈바꿈한 것으로 판단됩니다. 압도적인 성능과 에너지 효율을 갖춘 Rubin 아키텍처를 통해 AI 산업의 최대 걸림돌인 전력 문제를 해결하는 동시에, 각국 정부의 소버린 AI(Soverign AI) 구축을 돕는 핵심 전략 파트너로서의 위상을 굳혔습니다.

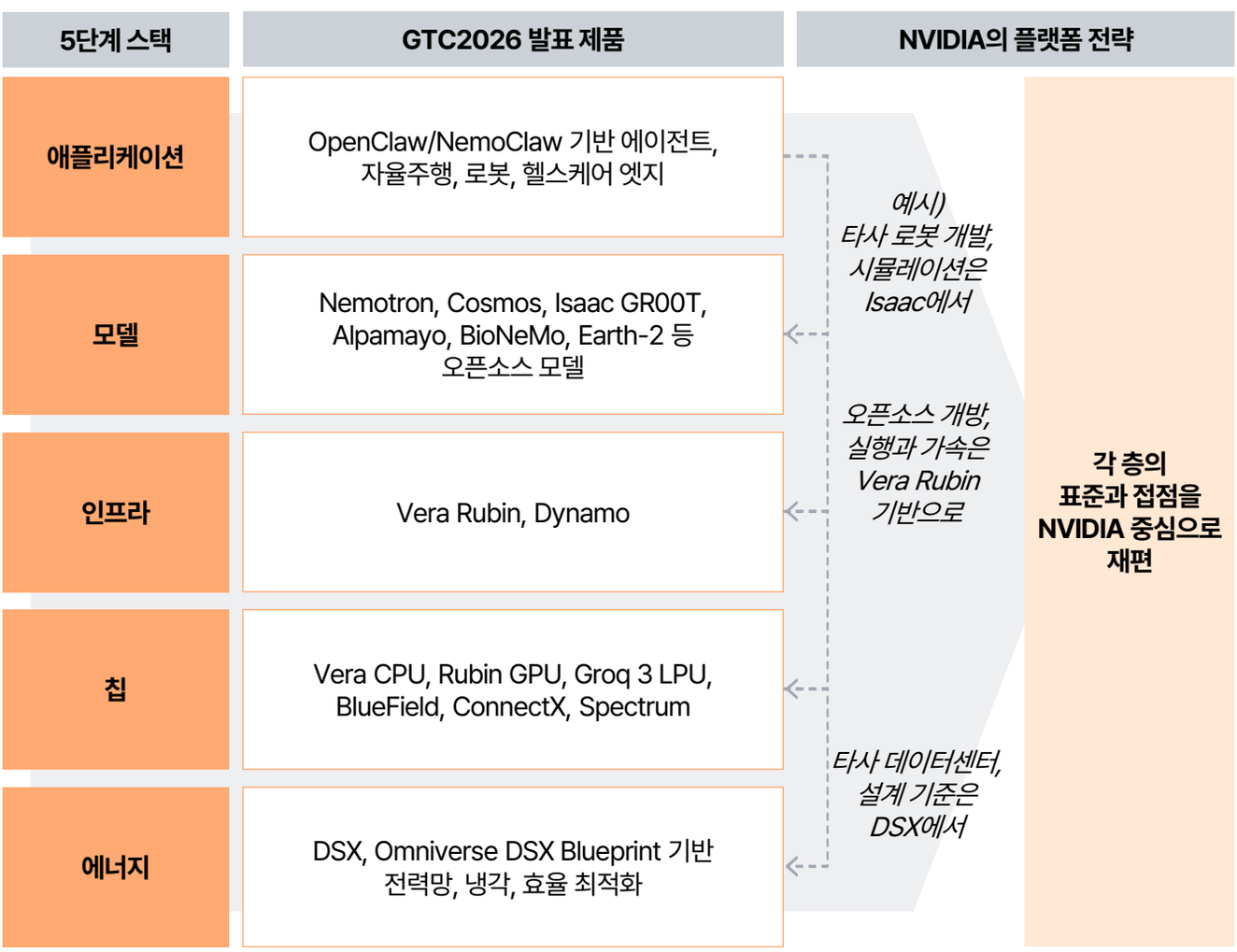
결과적으로 GTC 2026은 NVIDIA가 글로벌 빅테크의 설비 투자를 좌우하는 경제 지표를 넘어, 미래 산업의 설계도를 독점적으로 그려 나가는 대체 불가능한 인프라 제국임을 증명하는 무대였습니다.

02. GTC 2026 기조연설 핵심 메시지

1 한 눈에 보는 NVIDIA의 AI 플랫폼 전략

- AI 산업의 가치사슬은 더 이상 단일 칩이나 개별 모델 중심이 아니라 추론 가속기, AI 팩토리, 오픈 에이전트 스택, 피지컬 AI로 이어지는 전면적 시스템 경쟁으로 재편되고 있음
- NVIDIA는 Vera Rubin, NemoClaw, DSX Isaac GR00T 등 자사 제품 포트폴리오를 통해 에너지-칩-인프라-모델-애플리케이션으로 이어지는 '5-layer cake' 전 층에서 표준을 선점하겠다는 전략을 내비침

NVIDIA의 AI 플랫폼 전략



자료: 삼일PwC경영연구원

02. GTC 2026 기조연설 핵심 메시지

2 키워드로 보는 GTC 2026

추론의 시대

"The inference inflection has arrived."

- ✓ 젠슨 황 CEO는 AI 산업의 중심축이 학습에서 추론으로 이동하고 있음을 주장하며, 2027년까지 AI 칩 매출 기회가 최소 1조달러에 이를 수 있다고 제시. 이건 단순한 낙관론이 아니라, AI가 이제 모델 훈련보다 실제 서비스 활용 증가를 통한 수익화 단계로 진입한다는 것임

  → 기조연설의 중심이 단순 GPU 성능 자랑에서 실서비스용 추론 인프라로 이동했음을 보여줌

- ✓ 이를 받쳐주는 대표 발표 기술이 Vera Rubin 플랫폼임. NVIDIA는 Vera Rubin을 단일 GPU가 아니라 Vera CPU, Rubin GPU, NVLink 6 Switch, ConnectX-9 SuperNIC, BlueField-4 DPU, Spectrum-6 Ethernet, Groq 3 LPU를 묶은 "7개 칩, 5개 랙, 1개 슈퍼컴퓨터"로 설명함

  → 가속기의 의미를 GPU 한 장에서 추론 전체 파이프라인을 최적화한 시스템으로 전환

- ✓ 여기서 또 눈여겨 볼 것은 Groq 기반 추론 가속기임. NVIDIA는 추론을 prefill(입력 처리 단계)과 decode(출력 생성 단계)로 나눠 설명했고, Vera Rubin이 전자를, Groq 기술 기반 칩이 후자를 맡는 구조를 제시. 이는 '범용 GPU 하나로 처리한다'가 아니라, 추론 단계별 최적화가 필요한 시대를 인정하고 그 구조까지 자사 플랫폼 안에 편입시키기 시작했다는 의미로 해석됨

  → 추론 전용칩, CPU, 네트워킹, 메모리, 시스템 소프트웨어의 중요성이 함께 올라간다는 신호

AI 팩토리 / AI 에이전트

"AI factories coming alive, agents learning how to drive."

- ✓ 젠슨 황 CEO는 AI의 경쟁력은 더 이상 GPU 성능 하나로 결정되지 않으며, 토큰 생산성, 데이터센터 설계, 운영체제, 에이전트 보안 스택, 오픈소스 모델 생태계까지 통합된 AI 팩토리 체계에서 나온다는 점을 강조. 이에 NVIDIA가 AI 팩토리를 말할 때는 GPU만이 아니라 전력, 냉각, 네트워크, 스토리지, 운영 소프트웨어까지 함께 일컫는 것임. 젠슨 황이 선보인 NVIDIA DSX와 Omniverse DSX Blueprint 플랫폼은 AI 공장의 디지털 트윈을 설계하고 운영하여 최대 토큰 처리량, 복원력, 에너지 효율성을 달성하고자 함

  → AI 팩토리를 IT 장비 조달의 문제가 아니라 팩토리 설계와 운영 최적화의 문제로 끌어올림

- ✓ 여기서 소프트웨어 계층의 핵심으로 Dynamo 1.0이 제시됨. NVIDIA는 Dynamo를 "AI 팩토리를 위한 최초의 운영체제"라고 부르며, 추론이 모든 질의와 에이전트, 애플리케이션을 움직이는 엔진이라고 설명

  → 앞으로 AI 인프라 경쟁에서 하드웨어만으로는 차별화가 어렵고, 클러스터 수준에서 추론을 얼마나 효율적으로 스케줄링하고 운영하느냐가 핵심 경쟁력이 될 수 있음

- ✓ 미국 오픈소스 AI 모델의 선두주자로서, NVIDIA는 Nemotron, Cosmos, Isaac GR00T, Alpamayo, BioNeMo, Earth-2 등을 묶어 오픈모델 생태계를 넓히고 있다고 설명함

  → 모델 층을 오픈 생태계로 열어두고 그 아래의 인프라와 운영체제를 NVIDIA가 장악하겠다는 전략

- ✓ 또한, NVIDIA는 자율형 에이전트 OpenClaw를 소개하며, 기업 환경에 맞게 보강한 OpenShell(보안 및 프라이버시 기능을 강화)과 NemoClaw(OpenShell에 보안·프라이버시·정책 제어를 추가)를 발표

  → 에이전트가 잘 돌아가게 하는 '모델'만이 아니라, 기업이 안심하고 에이전트를 배치할 수 있는 '인프라 표준'까지 가져가겠다는 의미로 해석됨

02. GTC 2026 기조연설 핵심 메시지

피지컬 AI

"It's a GPT moment for the bots."

- ✓ 젠슨 황 CEO는 NVIDIA가 출시한 로봇 관련 제품들을 소개. 물리세계를 이해하고 시뮬레이션하는 Cosmos, 가상환경을 구현하여 로봇을 훈련시키는 Isaac Lab-Arena, Isaac GR00T N 로봇 모델, 자율주행을 위한 추론 기반 플랫폼 Alpamayo 등이 영상으로 시연되었으며, 기조연설 후반부에는 디즈니 캐릭터 올라프를 활용한 로봇이 무대에 등장함

  → NVIDIA의 차세대 성장 동력으로 피지컬 AI를 재차 강조하며, 자사 플랫폼에 구현한 가상환경 및 시뮬레이션을 통해 로봇 훈련 비용의 혁신적 절감이 가능함을 제시

- ✓ 자율주행차량 개발 관련하여 현대자동차, NISSAN, BYD 등 주요 자동차 기업과의 파트너십 관계 강조. 산업용 로봇 분야에서도 ABB, KUKA 등과 협력하여 생산 현장의 피지컬 AI 확산을 촉진 중. AI는 텍스트 · 이미지 생성을 넘어 공장 자동화, 자율주행, 수술 로봇 등으로 발전해야 하며, NVIDIA는 여기에 필요한 합성데이터 · 시뮬레이션 · 검증 인프라를 모두 공략

  → 로봇기업, 자동차 기업 등 非반도체 기업들은 NVIDIA를 중심으로 로보틱스 · 자율주행 생태계를 구축 중이며, CUDA가 GPU 인프라의 Lock-in 효과를 가져온 것처럼 피지컬 AI에서도 NVIDIA의 오픈소스 제품들이 경제적 해자(Moat)를 창출할 것으로 기대

종합 AI 기업

"So let us all eat five-layer cake."

- ✓ 젠슨 황 CEO는 추론형 AI 칩(Groq LPU)으로 제품군을 확대하고 LPX(LPU 기반 서버 랙)와 Rubin을 결합한 이기종 구조 제시

  → 기존의 학습형 AI 칩 뿐 아니라 추론형 AI 칩으로 제품군 확대

- ✓ DSX 플랫폼을 통해 AI 팩토리의 디지털 트윈을 구축하고, Dynamo를 통해 스케쥴링 · 운영 시스템 효율화

  → AI 칩을 넘어 에너지 · 인프라 단의 솔루션 기업으로 확장

- ✓ OpenClaw의 파급력을 강조하며 기업 고객을 유치하기 위해 보안 기능을 더한 AI 에이전트 플랫폼 NemoClaw 공개, 또한 자율주행차 · 휴머노이드 로봇 관련 오픈소스 제품을 공개하며 피지컬 AI 생태계로 사업영역 확장

  → AI 생태계의 최상단 애플리케이션 영역까지 공략하며 종합 AI 기업으로 탈바꿈

II

NVIDIA,
칩을 넘어 AI의 미래를 설계하다

01. 5단 케이크로 보는 AI 산업 스택

- NVIDIA는 최근 AI 산업 구조를 5단 케이크(Five-Layer Cake)에 비유
- 각 단계는 에너지-칩-인프라-모델-애플리케이션 등 5개 계층의 스택으로 정의되며, 이러한 구조가 AI 시대의 경제와 산업을 재편하고 있다는 것
- 여기서 NVIDIA는 AI 팩토리라는 개념 제시
- AI 팩토리는 대규모 GPU 서버와 초고속 네트워크, 실시간 추론 및 시스템 레벨 최적화를 기반으로 AI 모델을 구현하고 실행하는 AI 데이터센터로 에너지-칩-인프라를 포괄
- AI 팩토리에서는 GPU 뿐 아니라 네트워크, 통신 체계 전반이 AI 중심으로 재설계되고 있으며, NVIDIA는 단순한 칩 공급업체를 넘어 데이터센터에 적용되는 광통신 기술, AI와 로봇을 연결하는 6G 네트워크를 망라하는 인프라 기업으로 도약 중임
- 여기에 AI의 실제 활용 영역인 애플리케이션(자율주행차, 휴머노이드로봇 등)까지 사업영역을 확장하며 5단 케이크 전체에 NVIDIA 생태계를 구현한다는 계획

NVIDIA의 AI 산업 스택 5단계

애플리케이션 (II-④)	• 실제 경제적 가치를 창출하는 영역, 동일한 스택 위에서 작동하지만 구현 형태 및 결과는 다양 • 신약 개발 플랫폼, 산업용 로봇, 법률 에이전트, 자율주행차 등	챗봇, 로보택시, 기업형 에이전트, 로보틱스, 제조 AI 등
모델 (II-④)	• 언어, 과학, 금융, 의학, 물리세계 등 다양한 영역의 데이터를 이해 • 단백질 구조 분석, 물리 시뮬레이션, 로보틱스, 자율 시스템 분야 등에서 혁신 진전	LLM, VLM, VLA, MMLLM, GPT 등
인프라 (II-③)	• 부지, 전력 공급, 냉각, 건설, 네트워크, 수많은 프로세서를 하나로 조율하는 시스템 • AI 팩토리는 데이터를 저장하는 데이터센터와 달리 지능을 생산하도록 설계된 공장	AI 팩토리
칩 (II-②)	• AI 워크로드에 필요한 막대한 병렬 연산 능력, HBM, 고속 인터커넥트 • 연산을 효율적으로 처리하는 프로세서 기술이 AI 확장 속도 및 지능의 경제성 결정	
에너지 (II-③)	• AI는 실시간 막대한 전력을 필요로 하며 모든 토큰은 전자 이동, 열 관리 과정 등을 거쳐 생성 • AI 인프라의 제1원칙, 에너지가 곧 지능의 총량	

자료: NVIDIA, 삼일PwC경영연구원

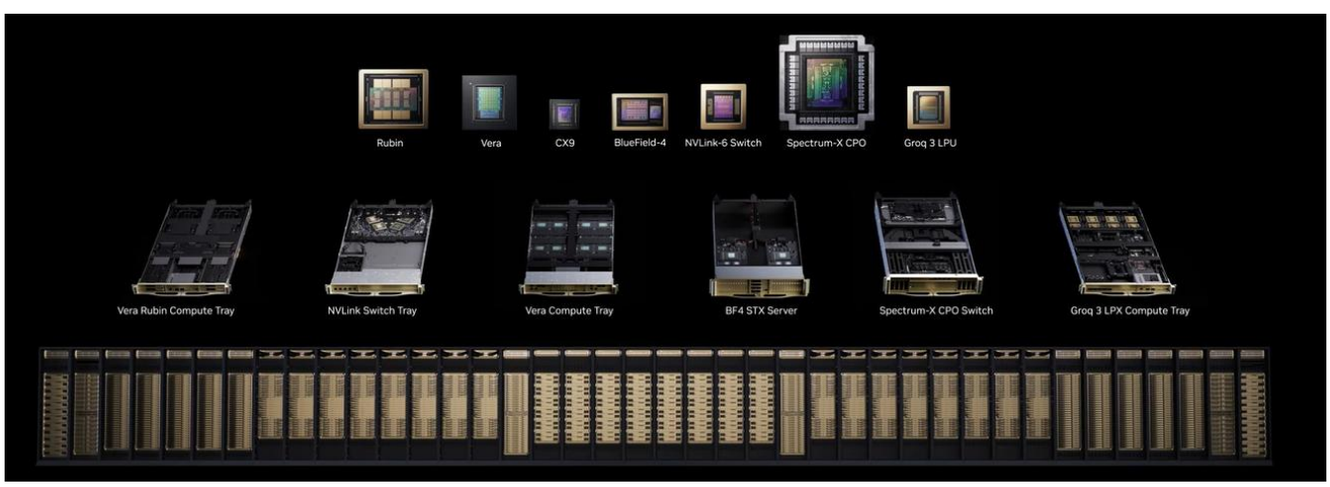
본 장에서는 칩(II-②), 에너지 · 인프라(II-③), 모델 · 애플리케이션(II-④) 순서로 NVIDIA의 전략 및 영향을 살펴보고자 함

02. 세간의 시선 집중, 차세대 AI 가속기 공개

1 Vera Rubin & Groq 추론형 칩 기반 이기종 전략

- NVIDIA는 2026년 1월 CES 2026에서 차세대 AI 칩 플랫폼 Rubin을 조기 공개한 데 이어, 이번 GTC 2026에서 Vera Rubin(*)의 기술적 세부사항, Groq3 LPU를 포함한 이기종 가속기 결합 전략 발표
- (*) Vera Rubin은 Vera CPU와 2개의 Rubin GPU를 통합한 구조
- 하나의 거대한 데이터센터 단위로 설계된 Vera Rubin NVL72 랙 스케일 시스템은 기존 Blackwell 기반 제품 대비 추론 성능은 5배, 토큰당 비용은 10분의 1 수준으로 성능과 경제성 모두 개선
- Vera Rubin은 2026년 하반기 시장 출시 예정

Vera Rubin 랙 스케일 시스템



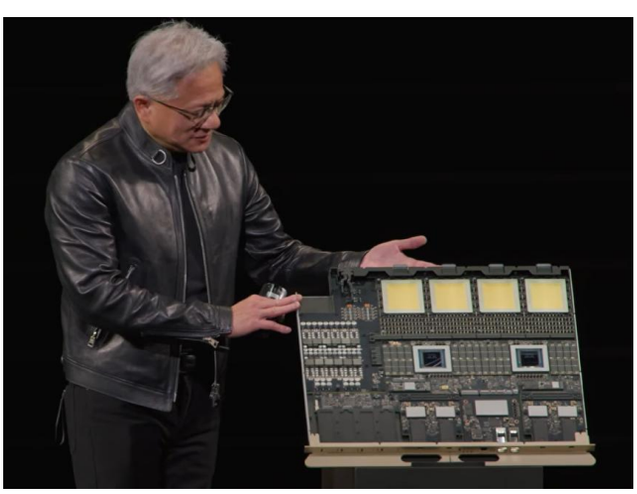
자료: NVIDIA, YouTube

Blackwell GPU · Rubin GPU 비교

구분	Blackwell	Rubin
트랜지스터 수	208B	336B
컴퓨트 다이 수	2	2
NVFP4 추론 성능(PFLOPS)	10	50
FP8 학습 성능(PFLOPS)	5	17.5

자료: NVIDIA, KB증권

Rubin Ultra



자료: NVIDIA, YouTube

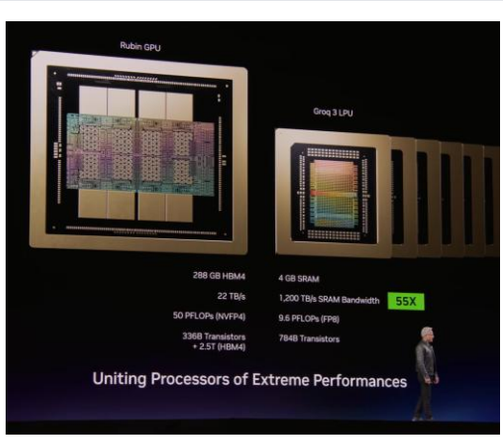
Vera Rubin 주요 사양

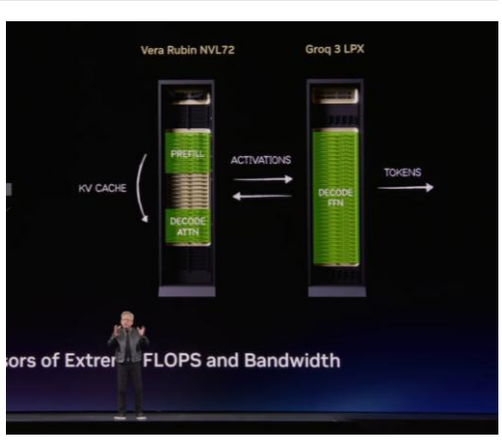
구분	내용
Rubin GPU	NVFP4 추론 50 PFLOPS, NVFP4 학습 35 PFLOPS, FP8 학습 17.5PFLOPS
Vera CPU	Olympus Armv9 88코어, 최대 176 스레드
메모리	HBM4 8스택, GPU당 288GB, 22TB/s 대역폭
NVLink 6	GPU당 3.6TB/s, NVL72 랙 기준 260TB/s
SOCAMM2	Vera CPU당 최대 1.5TB LPDDR5X 메모리 지원

자료: NVIDIA, KB증권

- 최근 AI 에이전트가 확산하면서 대용량 데이터를 학습하는 것 뿐만 아니라 학습한 내용에 기반한 AI의 추론 능력도 핵심 기능으로 부상
- Groq은 2025년 12월 추론 특화 AI 칩 제조 스타트업으로 NVIDIA가 2025년 비독점 기술 라이선스 계약을 체결하는 방식으로 인수 → 학습용 AI 칩 시장에서 추론 AI 칩 시장으로 생태계 확장
- Groq 추론 AI 칩은 HBM 대신 GPU 위에 SRAM(Static Random Access Memory)을 쌓는 방식으로 데이터 이동 속도와 전력 효율 개선
- NVIDIA는 GTC 2026에서 Rubin 플랫폼과 Groq 3 LPU를 결합한 이기종 가속기 전략 소개
- 최근 언론보도에 따르면 Groq은 삼성전자 파운드리 사업부에 맡긴 추론용 AI 칩의 생산량을 웨이퍼 기준 기존 약 9,000장에서 15,000장 수준으로 늘리기로 결정, 상용화 초입 단계에 들어선 것으로 분석
- 젠슨 황 CEO도 GTC 2026 기조연설에서 삼성전자가 Groq 3 LPU를 생산한다고 직접 발표

Groq 3 LPU 및 이기종 아키텍처 소개





자료: NVIDIA, YouTube

- 한편, GTC 2026에서는 Rubin에 이어 2028년 출시 예정인 아키텍처 Feynman의 공정 · 메모리 측면 로드맵도 프리뷰 형태로 제시
- Feynman은 NVIDIA 제품 최초로 1nm급 공정 적용 및 차세대 HBM5를 탑재할 것으로 예상

2 HBM4 시대 본격 개막

- 차세대 AI 가속기는 곧 메모리 경쟁의 격전지
- GPU 성능이 높아질수록 메모리 대역폭 요구가 커지고 HBM 수요도 확대
- Rubin GPU 옆에 장착되는 고성능 메모리는 HBM4
  - ✓ **동작 속도:** NVIDIA가 요구하는 Vera Rubin용 HBM4 동작 속도는 반도체 표준을 제정하는 국제 산업 표준 기구 JEDEC(Joint Electron Device Engineering Council)가 정한 기준인 초당 8Gb를 크게 상회하는 10~11Gb 수준
  - ✓ **용량:** Vera Rubin에 탑재되는 HBM4 용량은 576GB(288GB GPU 2개 결합)로 경쟁사 AMD의 차세대 제품 MI450의 HBM4 용량 432GB보다 높은 수준

NVIDIA 가속기별 HBM 모델 · 장착량

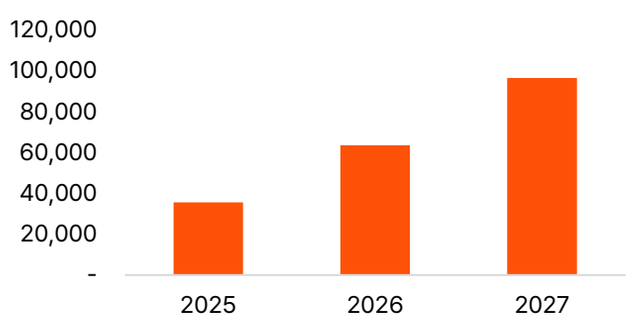
가속기	HBM	GPU 1개에 붙는 HBM 용량(GB)
A100	HBM2E	80
H100	HBM3	80
H200	HBM3E	141
B200	HBM3E	192
Blackwell Ultra	HBM3E	288
Vera Rubin	HBM4	288

자료: NVIDIA, 언론종합

- 2026년 3월 언론보도에 따르면 삼성전자와 SK하이닉스가 Vera Rubin용 HBM4의 공급사로 선정
- NVIDIA 공급망 내 점유율 확대를 두고 메모리 시장을 주도하는 삼성전자와 SK하이닉스의 경쟁 심화 전망
  - ✓ **삼성전자:** 동작 속도 초당 10Gb, 11Gb 두 가지 제품 규격 모두 NVIDIA 품질테스트를 통과한 것으로 알려짐
  - ✓ **SK하이닉스:** 2025년까지 AI 가속기 시장에서 독보적 점유율을 보유하며 NVIDIA의 핵심 HBM 파트너로 자리매김. 현재 11Gb 규격 최적화 작업 막바지. 품질테스트 과정에 따라 이르면 3월 NVIDIA 대량 생산 구매주문서를 받을 가능성이 높다고 보도됨

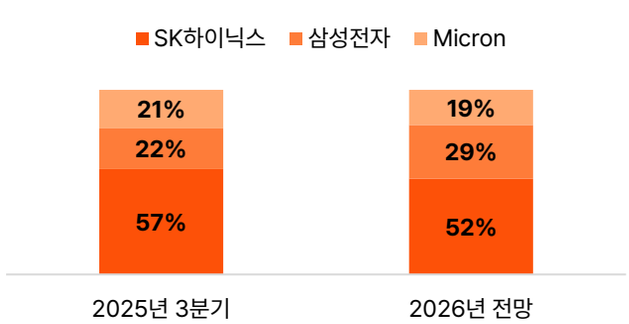
(참고) 미국 Micron: 최상위 제품 공급망에서 빠지고 중위 제품 Rubin CPX 라인에 HBM4를 납품할 것으로 알려짐

글로벌 HBM 시장규모 전망 (단위: 백만달러)



자료: JPMorgan Chase, 언론종합

글로벌 HBM 시장점유율 전망



자료: Counterpoint Research, 언론종합

3 HBM 밖으로 번지는 메모리 경쟁

- NVIDIA가 Vera Rubin의 Vera CPU 옆에 SOCAMM(Small Outline Compression Attach Memory Module)을 장착하기로 확정
- SOCAMM은 서버에 장착하는 모듈형 저전력 DRAM LPDDR 메모리
- HBM이 GPU의 연산 성능을 끌어올리는 핵심 메모리라면 SOCAMM은 CPU 주변에서 대규모 데이터를 안정적으로 처리하는 역할
- 하나의 모듈에 저전력 DRAM 4개가 탑재되는 구조로, 기존 서버용 메모리보다 데이터 전송 통로가 많아 속도 및 전력 효율 개선이 가능하며 기존 서버 메모리와 달리 탈부착이 가능해 모듈만 교체할 수 있다는 것도 장점
- NVIDIA 외에 Qualcomm, AMD도 SOCAMM 도입을 검토 중
- 이에 HBM 뿐 아니라 SOCAMM도 메모리 3사의 격전지로 부상

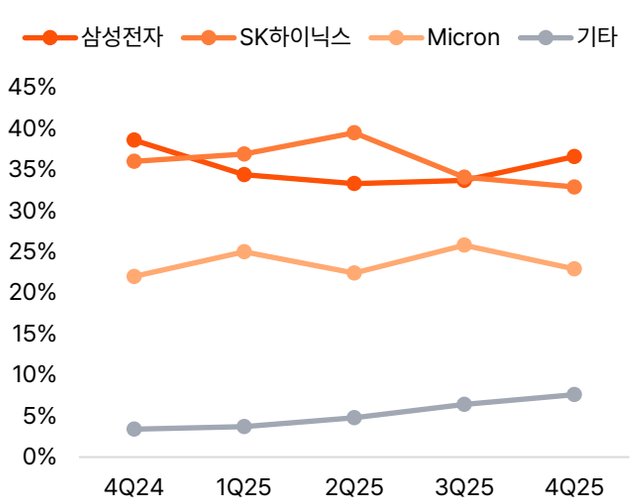
메모리 3사 SOCAMM2 사양 비교

구분	삼성전자	SK하이닉스	Micron
모듈 용량	192GB	192GB	192GB
동작 속도	8.5Gbps	7.5~9.6Gbps	9.6Gbps
메모리 규격	LPDDR5X	LPDDR5X	LPDDR5X
입출력 핀 수	694개	694개	694개

자료: 언론종합(2025.11)

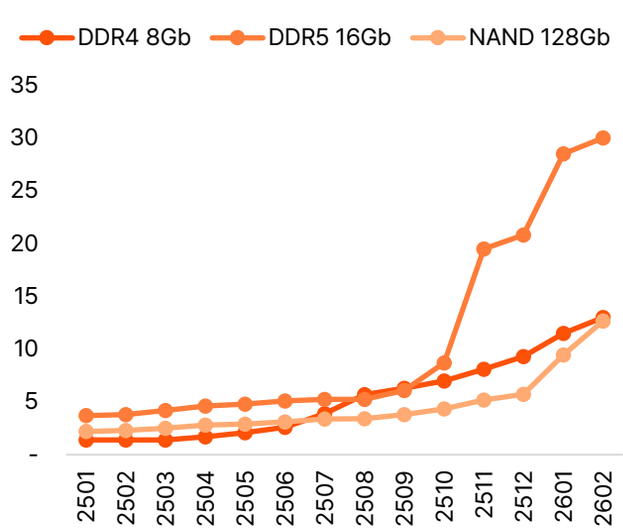
- 이 같은 AI 서버 투자 확대 및 고성능 메모리 수요가 지속 창출되면서 범용 제품을 포함한 전체 메모리 고정가격은 2025년에 이어 2026년에도 고공행진 중

글로벌 DRAM 시장점유율



자료: Omdia, 언론종합

2025~2026년 주요 메모리 반도체 고정가격 (단위: 달러)



자료: TrendForce, 산업통상자원부

03. 칩을 넘어 인프라로, 전방위 확장 드라이브

젠슨 황 CEO가 GTC 2026 기조연설에서 거듭 강조한 AI 팩토리는 수천개의 GPU와 초고속 네트워크 기반으로 AI를 구현하는 것으로, AI 팩토리의 발열 및 전력 부족 문제를 해소하기 위한 기본 전제조건은 "통신 인프라 혁신"

1 광통신: 빛으로 바꾸는 AI 인프라 혁신

- NVIDIA는 2026년 3월 미국 광학 · 레이저 부품업체 Lumentum Holdings, Coherent에 대한 대규모 투자 및 장기구매계약을 단행하며 광통신을 차세대 AI 팩토리 핵심 인프라로 선점하려는 전략 공식화

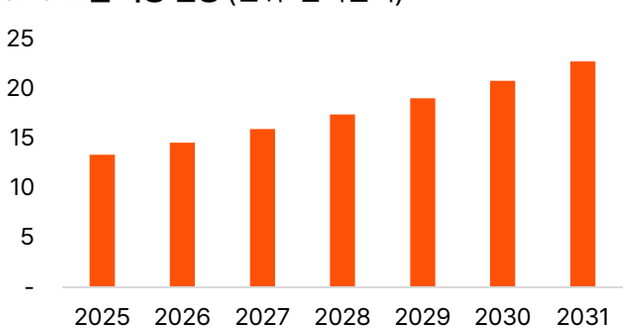
NVIDIA가 투자한 광통신 기업

구분	주력 제품	비고
Lumentum Holdings	• 전기 신호를 빛으로 변환, 초고속 전송하는 광트랜시버 • 기기 내부에서 빛을 쏘는 레이저 칩 등	NVIDIA, Rubin의 성능을 극대화하기 위해 Lumentum의 1.6T 광학 솔루션에 대한 전략적 투자 및 구매 약정
Coherent	• 광자를 활용한 고성능 광학 부품 및 시스템 • 레이저 송신기, 광섬유 케이블 등	NVIDIA, 첨단 레이저와 광 네트워킹 제품을 포괄하여 구매 약정

자료: 언론종합

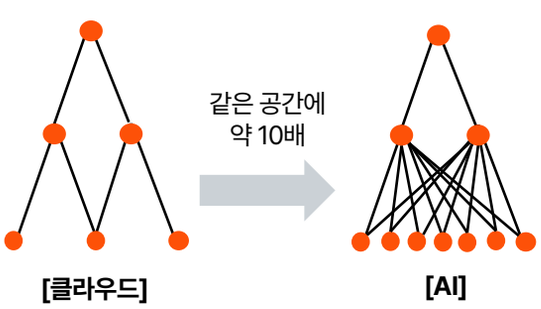
- AI 데이터센터의 수많은 서버는 광섬유로 연결되며, 서버 내부에 있는 프로세서와 저장장치의 데이터가 구리 회로를 통해 광 부품으로 전달되는 구조
- 최근 AI 연산량이 폭증하면서 구리 회로 기판은 데이터 전송 병목과 발열 문제 직면
- 이에 NVIDIA는 AI 데이터센터의 핵심 인프라로 '빛'을 활용하는 광(光)통신에 베팅
- 광통신을 구현하는 방식 중 하나인 CPO(Co-Packaged Optics)는 구리 회로 없이 프로세서와 광 부품을 하나의 기판에 패키징해 하드웨어 계층을 단축하고 데이터 전송 속도 및 전력 효율 개선
  - ✓ 서버 내부에 CPO 등 광학 인터커넥트를 도입하면 데이터 전송 전력 30% 이상 절감 가능
- NVIDIA의 광통신 도입 전략은 단순한 칩 공급업체에 머무르지 않고 AI 팩토리를 구성하는 인프라를 직접 재구성하겠다는 의미이며, 이를 통해 AI 생태계 최하단에 위치한 에너지 문제까지 일정 부분 해결할 것으로 기대

CPO 모듈 시장 전망 (단위: 십억달러)



자료: Market Report Analytics

기존 데이터센터 대비 더 많은 광섬유가 필요한 AI 팩토리

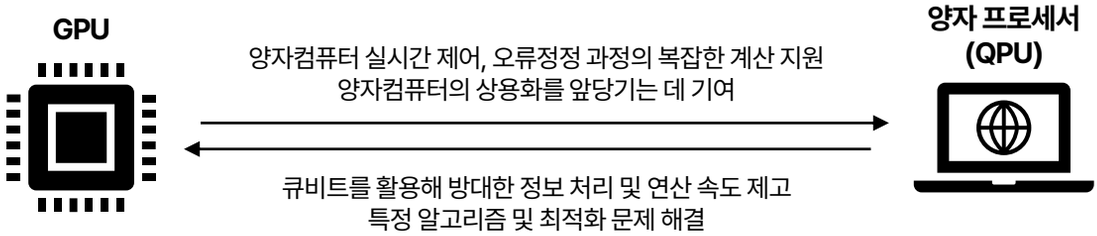


자료: Corning, 신한투자증권

2 GPU-양자 융합: 슈퍼컴퓨팅의 경계를 재정의하다

- NVIDIA는 2025년 GPU와 양자 프로세서를 연결하는 인터커넥트 기술 NVQLink를 공개하며 가속형 양자 슈퍼컴퓨터 구축 중
- NVQLink는 대규모 양자 연산에 필요한 알고리즘 처리를 GPU 슈퍼컴퓨터에서 병렬로 수행할 수 있도록 지원
- NVIDIA는 향후 데이터센터가 양자컴퓨터와 슈퍼컴퓨터를 동일한 인프라 내에서 상호 연결해 통합 운영하는 방향으로 진화할 것으로 전망
- 두 기술의 결합으로 복잡한 AI · HPC 문제 해결이 가속화되고, 양자컴퓨터의 상용화 시기도 앞당겨지는 시너지 효과 기대

GPU와 양자컴퓨터 결합의 기대효과



자료: 삼일PwC경영연구원

GPU와 양자컴퓨팅을 연결하는 NVIDIA NVQLink

자료: NVIDIA

글로벌 QPU · 플랫폼 개발 동향

구분	내용
NVIDIA	GPU-QPU 연결 플랫폼 NVQLink(2025.10)
Google	105큐비트 양자칩 Willow 양자우위 달성(2025.10)
IBM	120큐비트 양자칩 Nighthawk(2025.11)
QuantWare	1만큐비트 양자칩 VIO-40K(2025.12)

자료: 언론종합

- NVQLink 개발에는 로렌스버클리 국립연구소 등 주요 연구소들과 IonQ, Rigetti Computing 등 주요 양자컴퓨팅 기업들이 참여 중이며, 2026년에는 국내 양자 소부장 기업 SDT가 NVQLink 생태계(*)에 새롭게 합류한 것으로 확인됨

(*) NVQLink 생태계 합류 기업 (2026.03.16 NVIDIA 영문 홈페이지 기준): Alice & Bob, Anyon Technologies, Atom Computing, Dell Technologies, Diraq, Equal1, IonQ, IQM Quantum Computers, Infleqtion, Keysight, ORCA Computing, Pasqal, OQC, Qblox, Berkeley Lab QubiC, Quantinuum, Quandela, Quantum Circuits Inc., Quantum Machines, Quantum Motion, QuEL Inc., QuEra Computing, Rigetti Computing, SEEQC, QICK, SDT, Silicon Quantum Computing, Wistron, Zurich Instruments

04. 응용 분야까지 진출, 풀스택 기업으로 도약

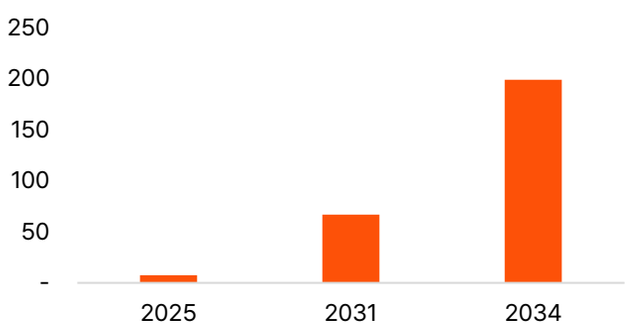
1 범용성과 보안, 둘 다 잡은 AI 에이전트 관리 플랫폼 NemoClaw

- 최근 OpenClaw 등(*)을 중심으로 AI 에이전트의 인기가 높아지고 여러 에이전트 모델이 업무 자동화와 생산성 향상에 활용되는 중

> **(*) Claw:** 사용자의 로컬 컴퓨터에서 실행되어 이메일 작성 · 전송, 일정 관리, 웹 검색 등 업무를 자율적으로 계획 · 처리하는 오픈소스 AI 에이전트를 통칭. 최근 젠슨 황 CEO는 OpenClaw를 두고 "역사상 가장 중요한 소프트웨어 중 하나일 것"이라며 극찬

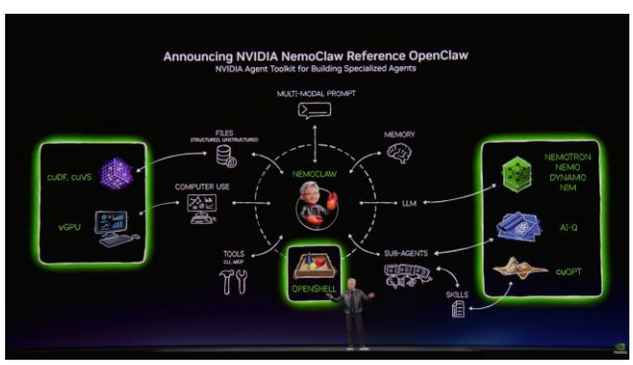
- 대규모 토큰 연산을 요구하는 AI 에이전트의 확산은 AI 팩토리와 GPU 시장의 성장 동력
- 그러나 NVIDIA는 하드웨어 인프라 수요에 그치지 않고 소프트웨어 플랫폼 시장을 직접 공략
- AI 에이전트 구동에 최적화된 파운데이션 모델 Nemotron을 출시한 데 이어 GTC 2026에서는 기업용 오픈소스 AI 에이전트 플랫폼 NemoClaw 공개
- NVIDIA는 범용성과 보안을 NemoClaw의 핵심 경쟁력으로 제시
  - ① **범용성:** GPU 사용 여부와 무관하게 다양한 기업 시스템 환경에 통합 가능. NVIDIA는 Google · Salesforce · Cisco · Adobe 등과 AI 에이전트 플랫폼 구축 파트너십 논의 중
  - ② **보안:** 기업용 보안 · 프라이버시 도구 포함. 최근 OpenClaw의 보안상 취약점에 대한 우려가 커지는 가운데 NVIDIA는 전용 보안 기능을 강화해 기업 고객 신뢰를 확보하는 전략

글로벌 AI 에이전트 시장 규모 (단위: 십억달러)



자료: Precedence Research, 언론종합

NVIDIA NemoClaw



자료: NVIDIA, YouTube

OpenClaw의 보안 이슈

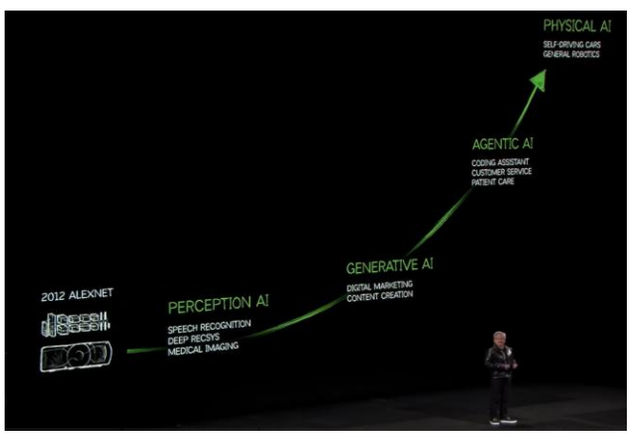
구분	주요 내용
보안 감사 결과	• AI 에이전트 보안 기업 ClawSecure, OpenClaw 생태계 보안 감사 결과 공개 • OpenClaw 스킬 중 41%에서 하나 이상의 보안 취약점 발견
중국 사용 규제	• 중국 보안당국, OpenClaw의 시스템 통제권 상실, 정보 유출가능성에 대한 보안 우려 표명 • 주요 국유기업과 대형 은행 대상 OpenClaw 설치 금지 통보

자료: 언론종합

2 피지컬 AI로 완성되는 NVIDIA의 AI 청사진

- 젠슨 황 CEO는 CES 2025에서 AI 발전단계를 제시하며 피지컬 AI의 도래를 언급, 피지컬 AI를 차세대 성장 동력으로 삼아 이후 로봇 관련 제품 출시 및 파트너십 확대 중
- 휴머노이드 로봇 특화 AI 모델, 시뮬레이션 플랫폼 등 피지컬 AI 생태계 내 NVIDIA 기술을 확산하고, 인프라 공급자에서 응용 분야의 최종 사업자로 도약하겠다는 구상

젠슨 황 CEO가 제시하는 AI 발전 단계 (CES 2025)



자료: NVIDIA, YouTube

GTC 2026에서 공개된 올라프 로봇

자료: NVIDIA, YouTube

- GTC 2026에서는 물리세계를 이해하고 시뮬레이션하는 플랫폼 Cosmos, 가상환경을 구현하여 로봇을 훈련시키는 Isaac Lab-Arena, 자율주행을 위한 추론 기반 플랫폼 Alpamayo 등이 소개됨
- 또한 현대자동차, NISSAN, BYD 등과 주요 자동차 회사와 협력하여 자율주행 차량을 개발 중임을 언급

NVIDIA의 피지컬 AI 관련 제품

구분	분류	특징
Cosmos	파운데이션 모델 · 플랫폼	물리세계 이해, 시뮬레이션
Isaac GR00T	파운데이션 모델	휴머노이드 로봇 특화 시각-언어-행동모델(VLA)
Isaac Lab-Arena	플랫폼	현실과 동일한 가상환경을 구현하여 로봇 훈련
OSMO	플랫폼	데이터 수집-모델 훈련-시뮬레이션 테스트까지 로봇 개발의 복잡한 과정을 한곳에서 통합 관리
Jetson Thor	로봇·엣지 AI 컴퓨팅 모듈	소형, 로봇 두뇌
Alpamayo	모델 · 플랫폼	자율주행을 위한 추론 기반 AI 모델, 시뮬레이션, 데이터셋 통합 플랫폼

자료: 언론종합

III

요약 및 시사점

요약 및 시사점 (1/2)

GTC 2026은 AI 산업이 개별 기술 경쟁을 넘어, 에너지·메모리·인프라·응용을 포괄하는 구조적 경쟁 단계에 진입했음을 보여주었다. AI 생태계 전반의 투자 방향과 공급망 질서를 재편하고 있는 이러한 변화의 핵심 흐름과 국내 기업에 대한 시사점을 다음과 같이 제시하고자 한다.

1 상반기까지는 HBM3 · HBM3E, 하반기부터는 HBM4가 주인공

Vera Rubin 출시 시점이 2026년 하반기로 예상되면서 Rubin GPU에 장착되는 HBM4의 양산 시점도 하반기가 될 전망이다. 현행 주력 제품 HBM3와 HBM3E의 독주는 상반기까지만 유효할 것으로 보인다.

→ **국내 메모리 기업에게는 기존 제품의 수명주기가 길어지는 긍정 효과 존재. 그러나 HBM4 공정 전환에도 본격 대비가 필요한 시점**

2 메모리 경쟁은 HBM 밖에서도 이어진다

AI 가속기 성능이 발전할수록 메모리 업계의 경쟁이 HBM를 넘어 다른 제품군으로 확대된다. GPU의 연산 성능을 높이는 HBM 뿐 아니라 CPU 주변에서 대규모 데이터를 안정적으로 처리하는 DRAM 모듈 SOCAMM이 대표적이다. NVIDIA가 Vera Rubin에 SOCAMM2를 탑재하기로 결정하면서 메모리 3사의 주도권 경쟁이 시작됐으며 전체 메모리 시장의 초과수요 및 가격 상승세도 이어질 전망이다.

→ **HBM 사례에서 알 수 있듯 NVIDIA가 주도하는 속도전에 얼마나 민첩하게 대응하느냐가 향후 AI 반도체 패권을 결정. 국내 메모리 기업들의 SOCAMM 제품군 확대 및 안정적 양산체계 구축 필요**

3 AI 팩토리는 기존 데이터센터와는 다른 운영 전략을 요한다

NVIDIA는 AI 데이터센터를 단순한 서버의 집합이 아니라 인프라, 운영, 보안을 포괄하는 AI 팩토리로 정의한다. AI 팩토리의 경쟁력은 모델 성능 뿐 아니라 토큰의 경제성에 있다. 토큰의 경제성은 전력 · 냉각 · 네트워크 · 스토리지 · 운영 소프트웨어 등 전체 시스템의 최적화 역량에서 결정된다.

→ **국내 기업들도 서버 또는 모델 중심 AI 전략을 플랫폼 · 운영 중심으로 재설계하고 데이터 이동의 병목, 전력 소모, 보안 문제 등을 중점적으로 점검할 필요 있음**

4 에너지 문제와 병목 해결은 '빛'에 달렸다

AI 팩토리는 수천개의 GPU와 초고속 네트워크, 실시간 추론 및 시스템 레벨 최적화를 기반으로 AI 모델을 구현하고 실행한다. AI 워크로드가 폭증하는 가운데 발열과 전력 부족 문제를 해소하기 위해서는 네트워크 및 통신 인프라 혁신이 필요하다. NVIDIA는 특히 빛으로 데이터를 전송하는 광통신에 주목하며 최근 광통신 장비기업들에 투자를 단행, 협력의지를 드러내고 있다.

→ **AI 시대 필수 인프라로서 광통신의 중요성이 커지고 있는 만큼 국내 광통신 · 부품 장비업체들도 이 흐름을 기회 삼아 선제적 대비, 시장 수요 확보 필요**

요약 및 시사점 (2/2)

5 GPU-양자 결합이 창출하는 플라이휠 (Flywheel)

NVIDIA는 GPU와 양자 프로세서를 NVQLink로 연결, AI와 양자기술을 결합한 새로운 형태의 슈퍼컴퓨팅 시대를 준비 중이다. GPU는 초저지연 · 고대역폭 연결성을 기반으로 큐비트 제어, 오류 수정, 보정 등 대규모 양자 연산에 필수적인 알고리즘 처리를 수행하고, 양자컴퓨터는 GPU만으로 풀기 어려운 특정 알고리즘 및 최적화 문제를 더 빠르게 해결 가능하다. 둘의 결합으로 더 복잡한 AI · HPC 문제가 해결되고 GPU - 양자컴퓨터 수요를 늘리는 선순환 구조가 전망된다.

→ **미래 양자시장의 중심축은 단독 양자컴퓨터가 아니라 양자-AI 결합 데이터센터가 될 가능성이 높음. 국내 양자기업들은 NVQLink 기반 생태계와 호환될 수 있도록 기술 방향성을 조정하고 양자-AI 융합 구조에 최적화된 제품 · 인프라 전략 수립 필요**

6 AI와 사이버보안, 안전벨트 없는 차량은 빠를수록 위험하다

Anthropic의 Claude 등을 기반으로 한 AI 에이전트 플랫폼 OpenClaw가 선풍적 인기를 얻고 있는 가운데 보안상 취약점에 대한 우려도 커지고 있다. 중국은 정부 기관 및 국유 기업을 대상으로 OpenClaw 사용 규제에 나선 상황이다. NVIDIA는 이러한 우려를 공략해 보안 · 프라이버시 도구가 포함된 NemoClaw를 공개했다.

→ **AI가 개인과 기업활동 저변에 깔리고, 더 많은 정보와 권한을 갖는 에이전트, 초지능 단계로 나아가려면 보안 인프라 혁신이 필수적이며, 단지 속도가 아니라 안전한 AI를 구현한 기업이 시장의 선택을 받게 될 것**

7 非반도체 기업도 NVIDIA를 외면할 수 없다

AI 에이전트, 피지컬 AI 등 애플리케이션 단에서의 AI 시장 확대는 생태계 하단에 자리잡은 AI 가속기 수요를 견인한다. 그러나 NVIDIA는 낙수효과에 머무르지 않고 직접 애플리케이션 분야에 진출, 오픈소스 모델을 배포하며 AI 생태계 전체 주도권을 잡는다는 방침이다. 이미 NemoClaw, Alpamayo 등을 선보이며 진출 영역을 확장, AI 반도체 기업을 넘어 종합 AI 기업으로 자리매김하고 있다.

→ **반도체 기업 뿐 아니라 제조 · 모빌리티 · 로봇 등 응용분야 기업도 NVIDIA를 주목해야 하고 NVIDIA 생태계 합류를 고민해야 할 시점임. 더불어 에이전트, 피지컬 AI 확산에 대비해 조직, 보안, 의사결정 구조 등 변화를 선제적으로 준비해야 함**

PwC's View : AI의 미래가 궁금하다면, CES와 GTC를 주목하자

CES 2025에서 젠슨 황 CEO가 AI 발전단계를 제시하며 피지컬 AI를 강조한지 불과 1년 만인 CES 2026 현장, 주요 기업들은 휴머노이드 로봇 기술력을 선보이며 세상을 놀라게 했다. NVIDIA가 비추는 방향이 곧 AI 기술의 발전 방향이 되고, 글로벌 기업들은 그 흐름에 올라탄다는 것이 확인되는 순간이었다. GTC는 CES와 달리 NVIDIA의 독무대로 펼쳐져 더욱 의미가 크다. AI 풀스택 기업으로 거듭난 NVIDIA가 제시하는 AI의 미래 이정표와 나침반이 GTC 현장에 고스란히 담긴다. 젠슨 황은 피지컬 AI 다음 단계에 대해 아직 언급하지 않았다. 신뢰와 거버넌스, 안전한 AI를 포괄하는 **Holistic AI**가 나올지, 다른 AI 개념이 나올지 아직 알 수 없다. 그렇기에 우리는 CES와 GTC에서 젠슨 황과 NVIDIA를 주목해야 한다.

인지형 AI (Perception AI) → **생성형 AI** (Generative AI) → **AI 에이전트** (Agentic AI) → **물리적 AI** (Physical AI) → ?

Appendix

GTC 2026 전시관 테마

테마	상세
Automotive	교통과 이동수단의 미래를 변화시키는 최첨단 AI 기술
DSX AI Infrastructure	전력 및 냉각, 건설, 소프트웨어 분야의 리더들이 통합된 NVIDIA DSX 표준과 Vera Rubin 플랫폼을 사용하여 어떻게 대규모 AI 팩토리를 구축하고 있는지
Financial Services	알고리즘 트레이딩부터 결제 분야의 에이전틱 커머스(Agentic Commerce)에 이르기까지, AI가 산업을 어떻게 변화시키고 있는지 보여주는 최첨단 AI 활용 사례들을 전시)
Healthcare and Lifesciences	생물학, 수술 가이드, 지능형 환자 케어 시스템의 AI 기반 혁신
Inception Startups	3만 개 이상의 글로벌 기술 스타트업이 애플리케이션 개발 속도를 높일 수 있도록 지원하는 NVIDIA Inception*의 성과를 확인(이번에는 그중 약 55개 기업을 소개할 예정) * NVIDIA 플랫폼과 생태계를 기반으로 AI 스타트업의 성장을 지원하는 무료 프로그램. 프로토타입 제작부터 제품화까지, 스타트업의 성장 단계별 맞춤형 지원을 제공
Industrial AI and Robotics	제조 공장 및 로보틱스 분야의 리더들이 산업용 AI, 디지털 트윈, 가속 컴퓨팅을 활용해 자율 운영을 어떻게 확장하고 있는지
Quantum Computing	양자 하드웨어 인프라부터 이를 구동하는 소프트웨어, 실제 사용되는 앱 개발 생태계까지, 기업들이 미래의 양자-GPU 슈퍼컴퓨터를 어떻게 개발하고 지원하고 있는지
Telecommunications	업계 리더들이 어떻게 통신 운영을 혁신하고, 무선 네트워크를 발전시키며, 소버린 AI(Sovereign AI) 인프라를 구축하고 있는지 확인

자료: NVIDIA, 언론종합, 삼일PwC경영연구원

Business Contacts

문홍기 Partner
hong-ki.moon@pwc.com

정성문 Partner
sungmoon.cheong@pwc.com

노승연 Partner
seungyon.roh@pwc.com

박유현 Partner
yuhyeon.park@pwc.com

Author Contacts

이은영 상무

삼일PwC경영연구원
eunyoung.lee@pwc.com

안정효 선임연구원

삼일PwC경영연구원
jeonghyo.ahn@pwc.com

최형원 연구원

삼일PwC경영연구원
hyungwon.choi@pwc.com

강수정 연구원

삼일PwC경영연구원
sujeong.j.kang@pwc.com

삼일PwC경영연구원

최 재 영 경영연구원장

jaeyoung.j.choi@pwc.com