GROWTH RESEARCH

CES 2025 젠슨황 언급 내용 총 정리

CES 2025 젠슨황 키노트 요약

1. GPU의 역사와 AI 발전 (과거 회고)
2. 새로운 GPU 아키텍처 'Blackwell' 시리즈 발표
3. AI 스케일링 법칙(Scaling Laws)과 데이터센터용 Blackwell
4. 에이전틱(Agentic) AI와 소프트웨어 스택
5. AI on Windows (WSL2)
6. 물리적 AI(Physical AI)와 'Cosmos' 플랫폼
7. 로보틱스(산업·휴머노이드·물류)와 디지털 트윈
8. DGX와 소형 AI 슈퍼컴 'Digits(프로젝트명)'

CES™ Powered by The Consumer Technology Association®

2025. 01. 07

official@growthresearch.co.kr

엔비디아의 CEO 젠슨 황이 6년 만에 CES 2025에서 기조연설을 진행

▶ PC·클라우드·온프레미스 전 분야 AI 가속화

- NVIDIA는 GPU(Blackwell)와 AI 소프트웨어(NeMo, Cosmos 등)를 결합해 AI를 전 산업에 확산하려 함.

▶ 물리적 AI(Physical AI)가 다음 물결

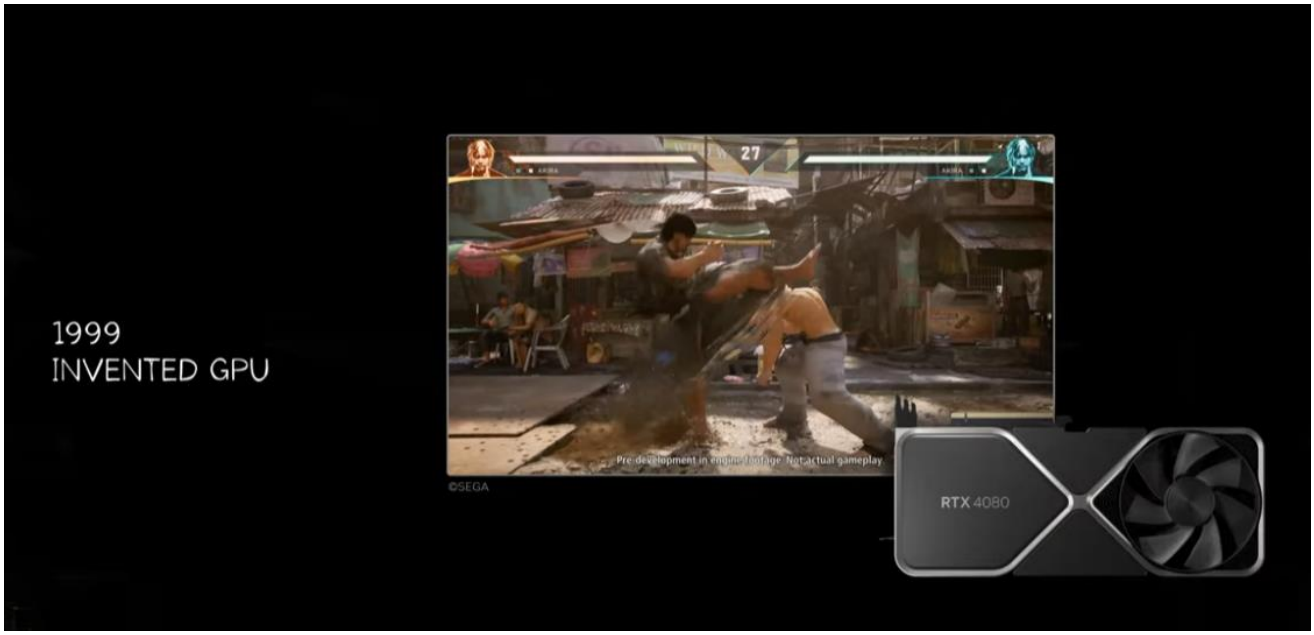
- 로보틱스와 자율주행에서 ChatGPT 수준의 '혁신적 순간'이 임박.
- Cosmos로 생성되는 합성 데이터와 Omniverse 시뮬레이션이 AI 훈련 효율을 극대화.

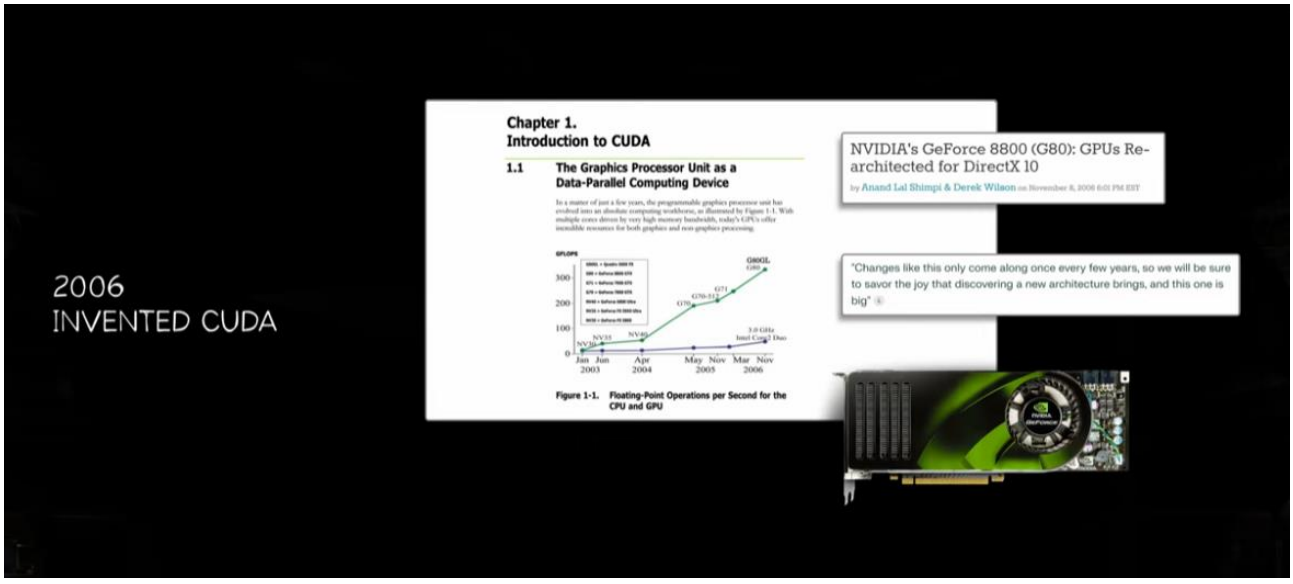
▶ 새로운 형태의 컴퓨팅 패러다임

- 전통적 CPU 프로그래밍(Hand-coded)에서 "AI가 스스로 학습·생성하는" 뉴 컴퓨팅으로 전환.
- 데이터센터 규모, 소프트웨어 스택, 개발 모델 모두 급변 중.

▶ DGX 이후 개인·소기업용 'Digits'

- 누구나 데스크톱 규모의 AI 슈퍼컴을 사용할 수 있게 되어, AI 개발이 대중화될 전망.





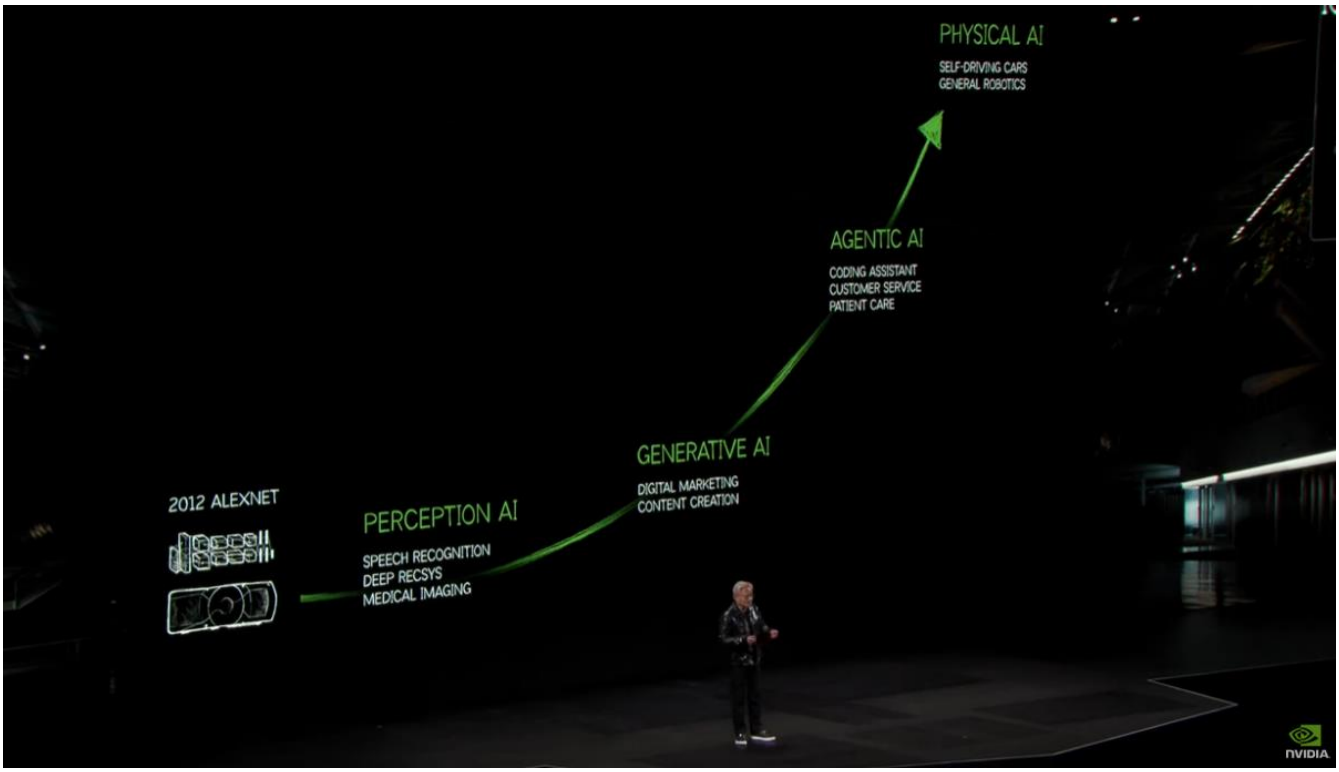
1. GPU의 역사와 AI 발전 (과거 회고)

▶ 1999년 프로그래머블 GPU 발명

- NVIDIA가 1999년에 프로그래머블 GPU를 발명하며 20년 이상 눈부신 그래픽스 및 병렬처리 발전이 시작됨.
- 현대적인 컴퓨터 그래픽스의 기반을 마련.

▶ 2006년 CUDA(쿠다) 발명

- 2006년, GPU를 범용 계산에 활용하기 위해 CUDA를 발표.
- 초기에는 GPU 프로그래밍 모델을 설명하기 어려웠고, 실제로 산업·학계가 적응하는 데 6년 정도 소요됨.



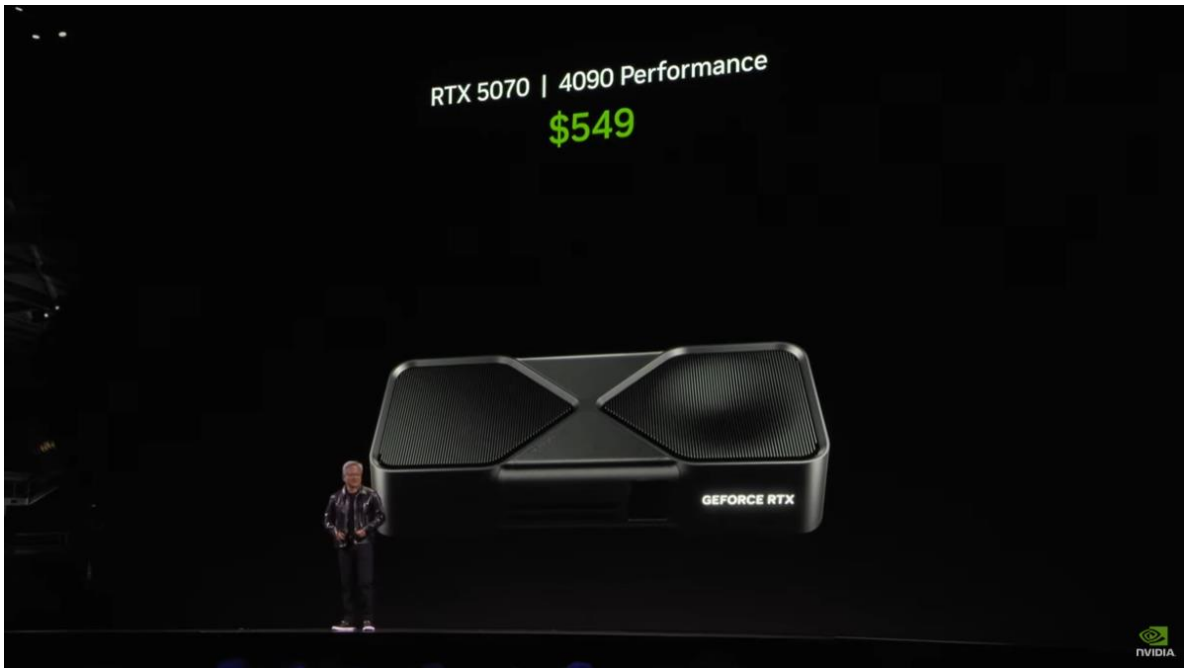
1. GPU의 역사와 AI 발전 (과거 회고)

▶ 2012년 'AlexNet' 계기로 AI 붐

- 2012년 알렉스 크리제브스키(Alex Krizhevsky), 일야 수츠케버(Ilya Sutskever), 제프 힌튼(Geoff Hinton)이 NVIDIA GPU(CUDA)를 활용하여 AlexNet을 발표, 딥러닝 붐이 본격화.
- 이미지 인식, 자연어 처리 등 다양한 영역으로 AI가 급격히 확산됨.

▶ 2018년 이후 Transformer 혁신

- 구글의 Transformer('BERT') 등장으로 언어 모델이 급격히 발전.
- 이로 인해 AI 연구 및 적용 분야 전반이 재편.
- NVIDIA는 이를 "컴퓨팅 전체가 근본적으로 바뀌는 전환점"으로 인식.



2. 새로운 GPU 아키텍처 'Blackwell' 시리즈 발표

2-1. Blackwell 아키텍처의 특징

▶ GeForce RTX 50 시리즈 (Blackwell 기반)

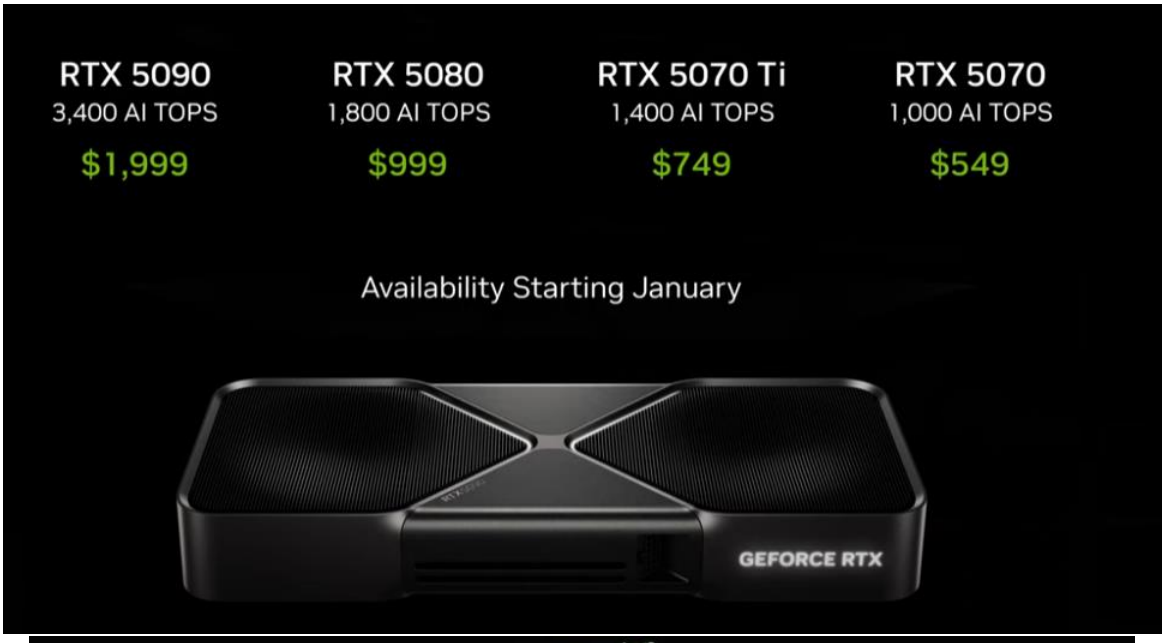
- 차세대 GPU 마이크로아키텍처인 Blackwell 발표.
- 92억 개(=92B) 트랜지스터, 4 TFLOPS(AI용), 380 RT(레이 트레이싱) TFLOPS, 125 셰이더 TFLOPS 등 압도적 성능.
- 메모리는 GDDR7(1.8TB/s 대역폭, 이전 세대 대비 2배 성능).

▶ AI 연산과 그래픽스 연산을 융합

- 셰이더 코어와 텐서 코어 모두가 뉴럴 네트워크(Neural Network) 연산을 지원.
- Neural Texture Compression, Neural Material Shading 등이 적용, 고품질 텍스처/머티리얼을 AI로 생성·압축.

▶ DLSS 최신 버전

- 디노이징과 초해상도, 프레임 예측(프레임 생성) 기능을 결합한 DLSS(Deep Learning Super Sampling).
- 한 프레임을 실제로 렌더링하면, 나머지 3프레임은 AI가 예측하여 생성 → 전체 성능(프레임레이트) 극적 향상.
- "렌더링해야 할 픽셀 중 6%만 실제 계산, 나머지는 AI가 예측"하는 식으로 설명.



2. 새로운 GPU 아키텍처 'Blackwell' 시리즈 발표

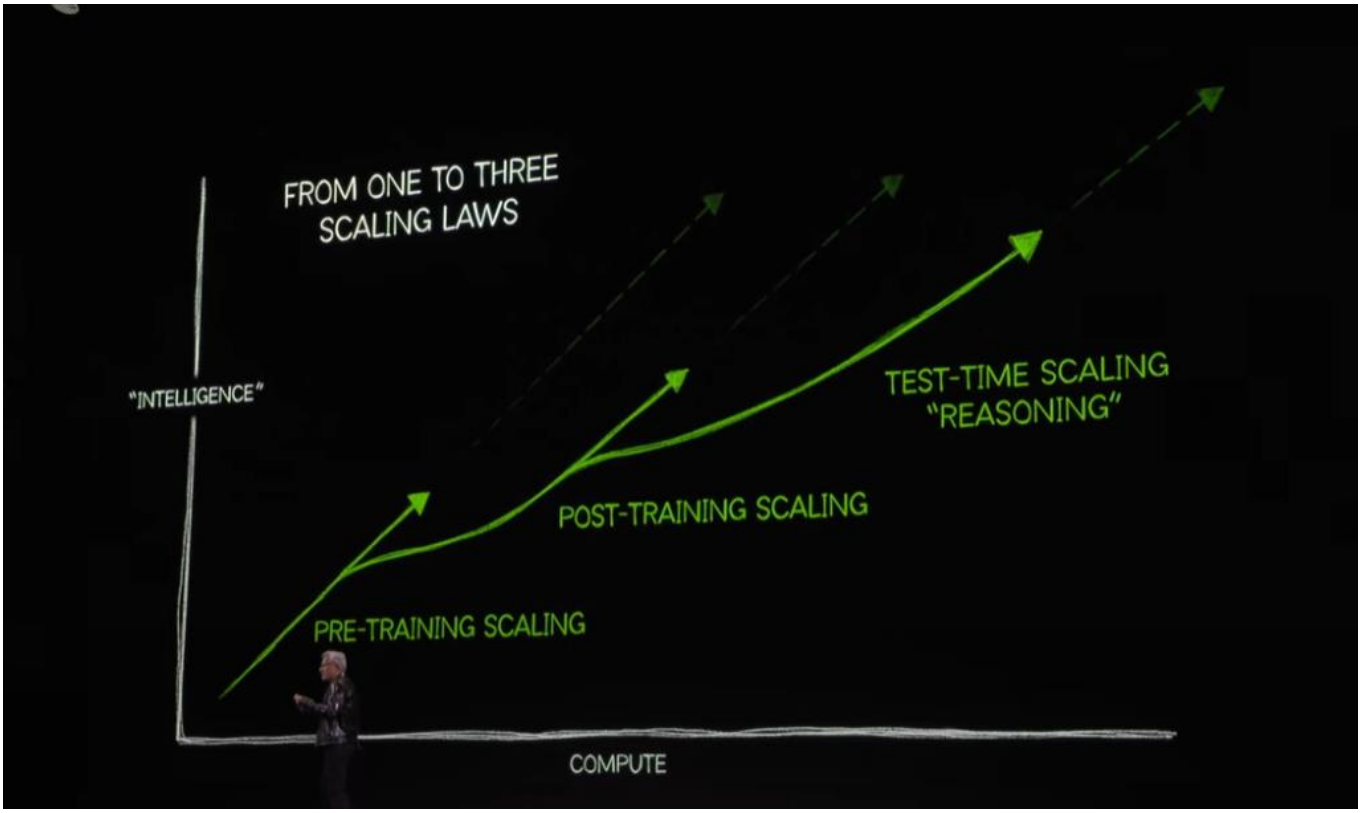
2-2. GeForce 라인업 & 가격

▶ GeForce RTX 50 시리즈

- RTX 5070: 기존 RTX 4090급 성능을 549달러에 제공한다고 언급.
- RTX 5090: RTX 4090의 2배 성능을 지닌 최상위 모델로 소개.
- 출시 시점: 2024년 1월(양산) ~ 상반기 예상.

▶ 노트북용 Blackwell

- RTX 5070 랩톱(1299달러부터)으로도 4090급 성능 구현.
- AI 기반 렌더링(DLSS)을 통해 전력 효율 및 휴대성 확보.



3. AI 스케일링 법칙과 데이터센터용 Blackwell

3-1. 세 가지 스케일링 단계

▶ Pre-Training Scaling

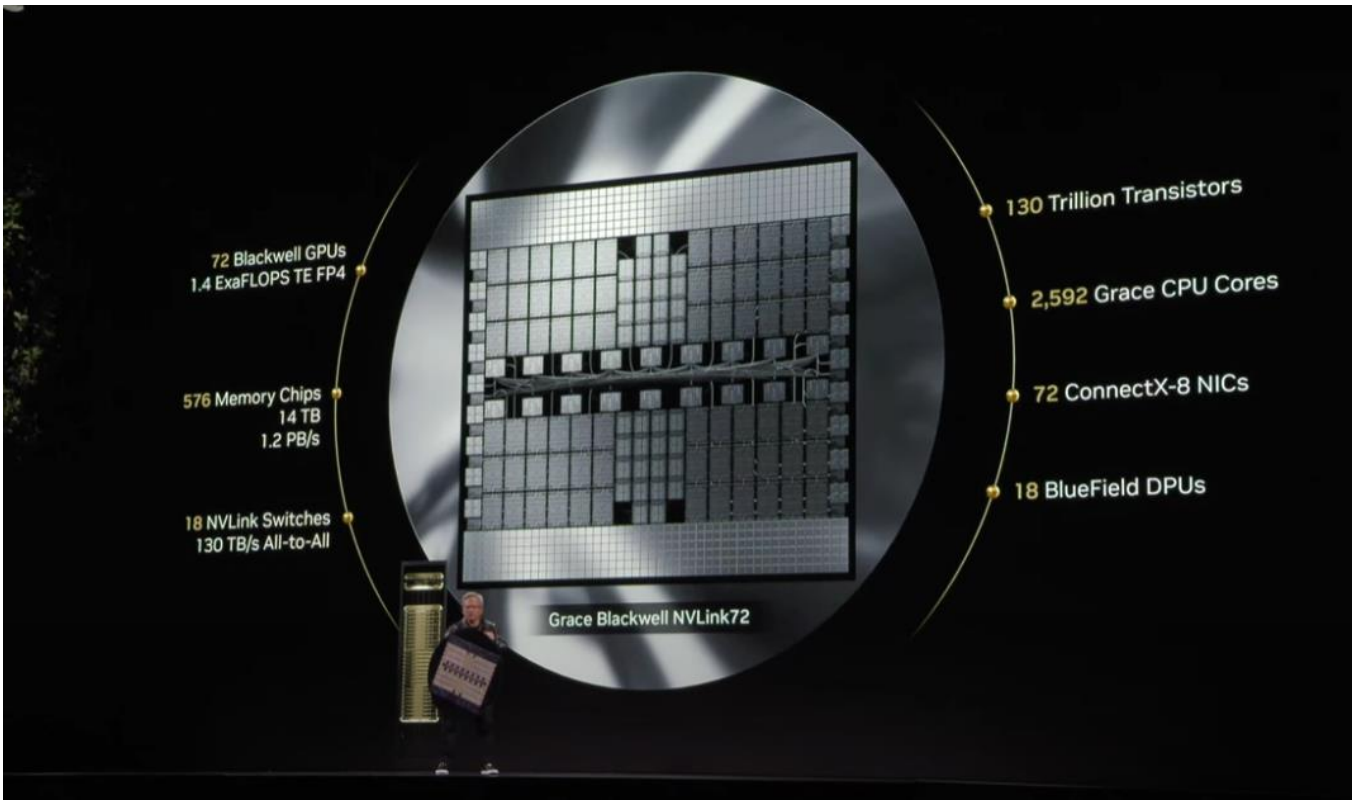
- 대규모 데이터, 대규모 모델, 대규모 연산을 통해 모델을 '사전 훈련'하면 성능이 비약적으로 상승.
- 방대한 텍스트, 이미지, 오디오 등 멀티모달 데이터가 매년 폭발적으로 증가.

▶ Post-Training Scaling

- RLHF(인간 피드백 강화학습), Synthetic Data - Generation 등으로 모델을 추가로 고도화. 사후 훈련 통해 모델이 특정 도메인/작업에 특화될 수 있음.

▶ Test-Time Scaling

- 에이전틱(Agentic) AI가 추론 시(실시간) 스스로 연산 자원을 더 투입하고, 여러 단계를 거쳐 사고(Reasoning)할 수 있음.
- 자율적으로 '장고(Long Thought)'를 통해 더 나은 답변을 생성할 수 있게 됨 → 추론 연산량 폭발적 증가.



3. AI 스케일링 법칙과 데이터센터용 Blackwell

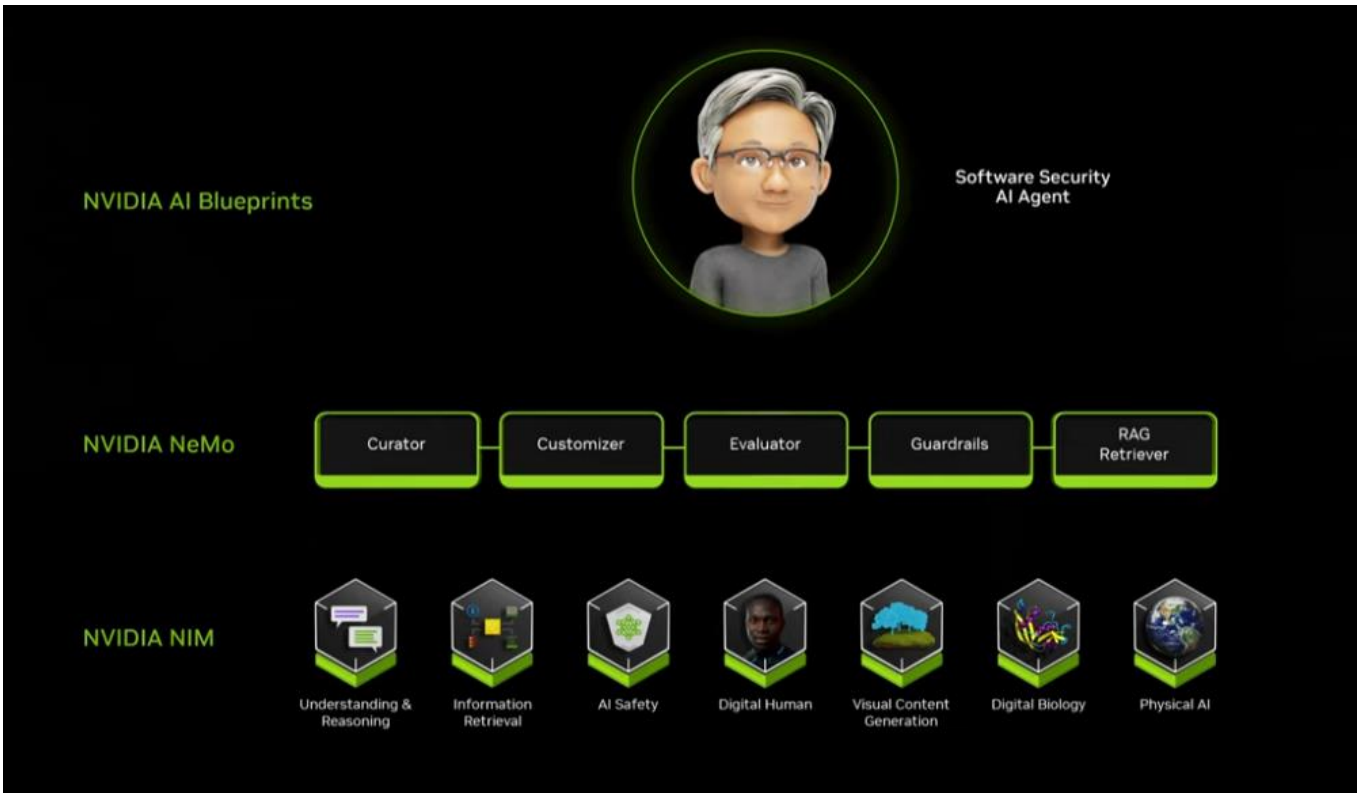
3-2. Blackwell 데이터센터 GPU

▶ Grace Hopper(Grace+GPU) 다음 세대인 'Blackwell(H100 후속)'

- 클라우드/데이터센터 파트너(15개 이상의 서버 제조사, 200+ SKU)로 대량 양산 중.
- NVLink 72 등 대규모 GPU 클러스터(1.4 ExaFLOPS AI 성능).
- 이전 세대 대비 전력당 성능(Perf/Watt) 4배, 비용 효율 3배 향상 → 대규모 모델 훈련 및 추론에 필수적.

▶ NVLink 72 시스템("전체가 하나의 거대한 칩과 같다.")

- 72개의 Blackwell GPU를 하나로 묶어 1.4 엑사플롭스(AI 플롭스) 달성.
- 14TB HBM, 1.2PB/s 메모리 대역폭.
- 무게 1.5톤, 약 60만 개 부품, 2마일(약 3.2km)의 구리 케이블, 45개 공장에서 병렬 생산.

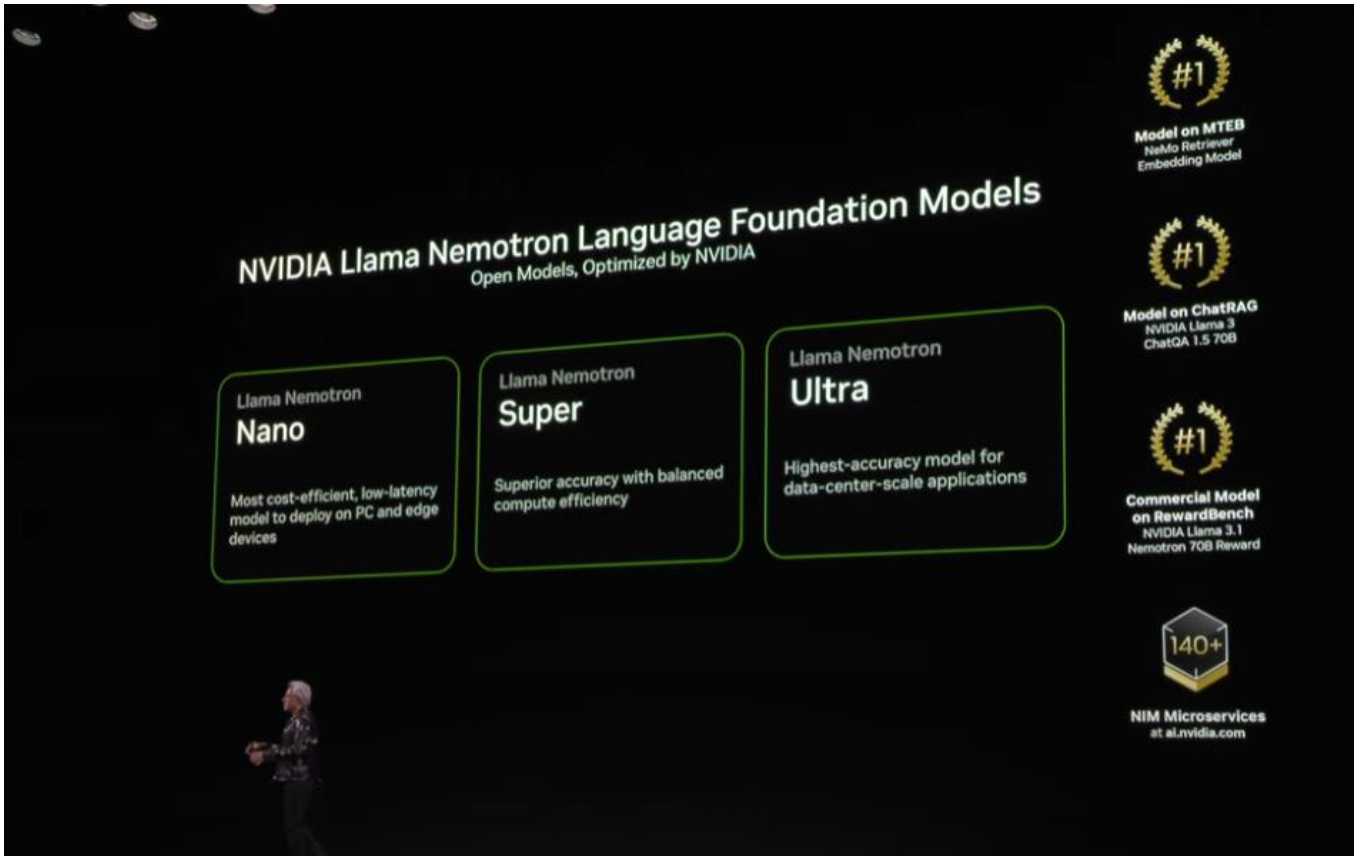


4. 에이전틱(Agentic) AI와 소프트웨어 스택

4-1. Agentic AI 개념

▶ Agentic AI: 여러 AI 모델이 협업하여 툴, 문서, 인터넷 검색 등을 활용해 문제를 단계별로 해결.

- 명령→풀이→추론→각종 API/툴 호출→결과 도출의 복합 프로세스.
- Test-time Scaling의 대표적 예시로, 추론 시 연산량이 기하급수적으로 증가.



4. 에이전틱(Agentic) AI와 소프트웨어 스택

4-2. NVIDIA의 소프트웨어 스택

▶ NVIDIA NIMs (AI Microservices)

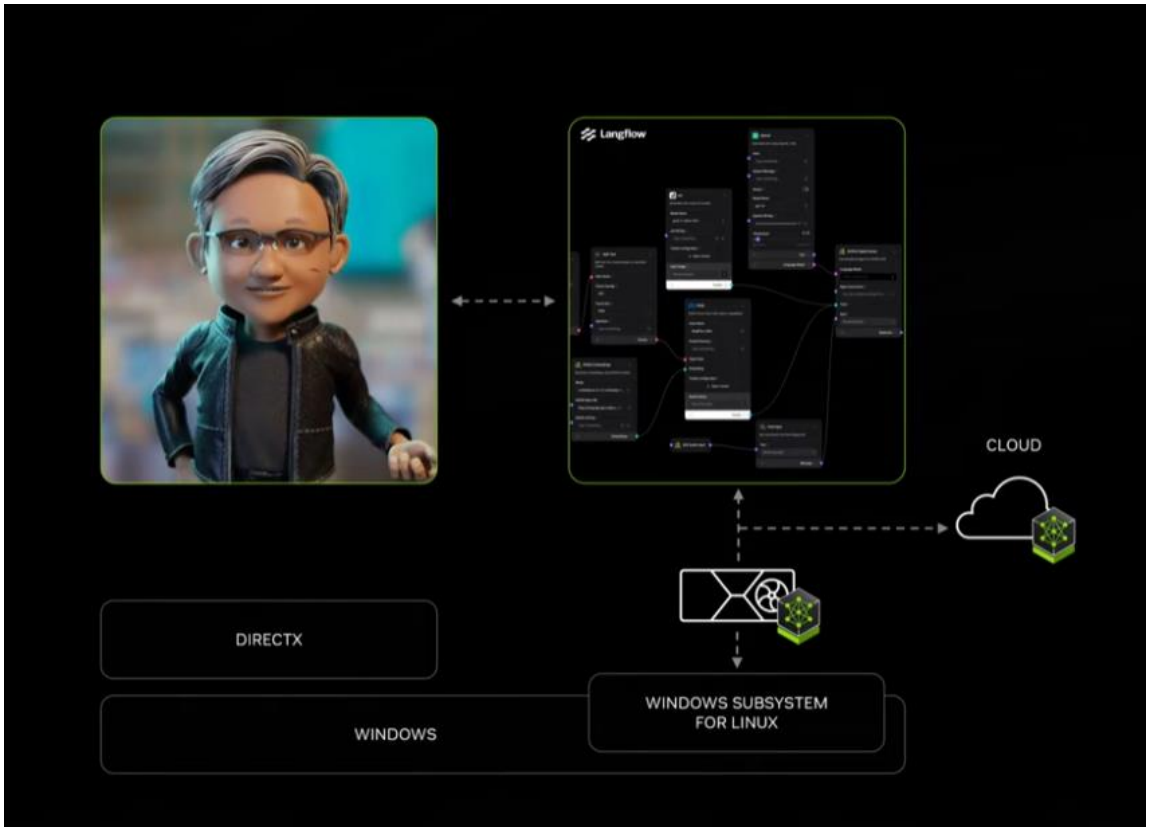
- 영상, 음성, 텍스트, 비전·언어 모델, 디지털바이올로지 등을 마이크로서비스 형태로 제공.
- 컨테이너 형태로 배포되어 어떤 클라우드/온프레미스든 동일한 환경에서 실행.

▶ NVIDIA NeMo

- 기업이 'AI 에이전트(디지털 직원)'를 온보딩·훈련·검증·가드레일 설정하는 플랫폼.
- 사내 언어, 프로세스, 정책 등을 반영해 모델을 세분화·미세조정.

▶ Llama Nemotron

- 메타의 Llama 3.1을 기반으로 엔터프라이즈용, 다양한 크기의 모델(초소형~초대형) 세트 출시.
- 멀티 리더보드 1위 수준, 파인튜닝 용이.
- 극도로 작은 모델(빠른 응답)부터 거대 모델(교수·티처 모델)까지 라인업 구성.



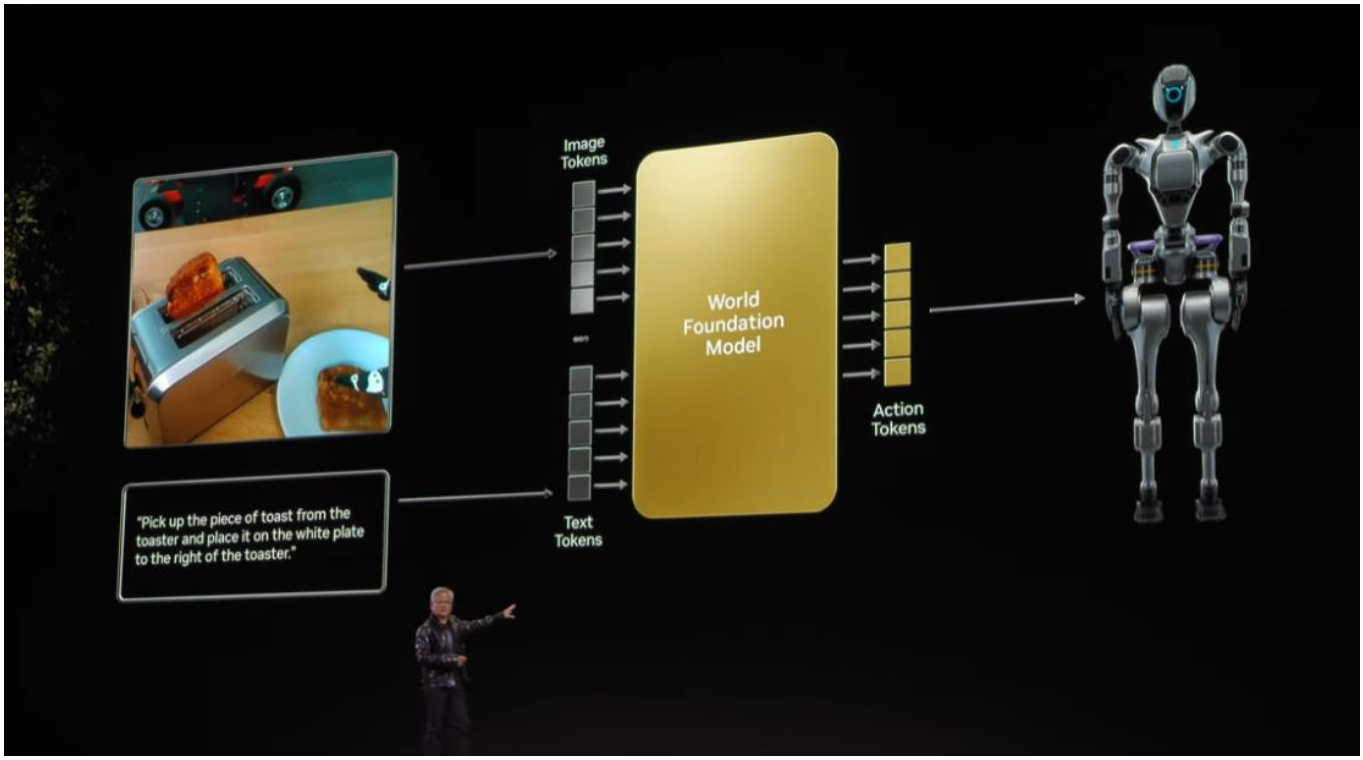
5. AI on Windows (WSL2)

▶ Windows WSL2(Windows Subsystem for Linux 2)를 통한 AI 지원

- NVIDIA CUDA를 네이티브로 지원하는 WSL2에서 NVIDIA의 모든 AI 소프트웨어 스택 구동 가능.
- PC(Windows)에서도 쿠버네티스나 클라우드 네이티브 방식으로 AI 실행 가능.
- “AI PC” 시대를 예고: 멀티미디어 API에서 확장해 ‘AI API(텍스트, 그래픽, 사운드 생성)’ 시대가 올 것.

▶ PC OEM 파트너

- 전 세계 주요 PC 제조사와 협력해 RTX 50 시리즈 + WSL2 + NVIDIA AI 스택을 탑재한 “AI PC” 보급 추진.



6. 물리적 AI(Physical AI)와 'Cosmos' 플랫폼

6-1. Physical AI 개념

▶ 텍스트 대신 로봇이나 자율주행 차량 등의 센서·행동 데이터가 입력(프롬프트)이며, 출력은 행동(Action) 토큰.

▶ 물리 세계를 이해하려면 중력, 관성, 마찰, 객체 영속성 등 물리 법칙에 대한 "월드 모델" 필요.

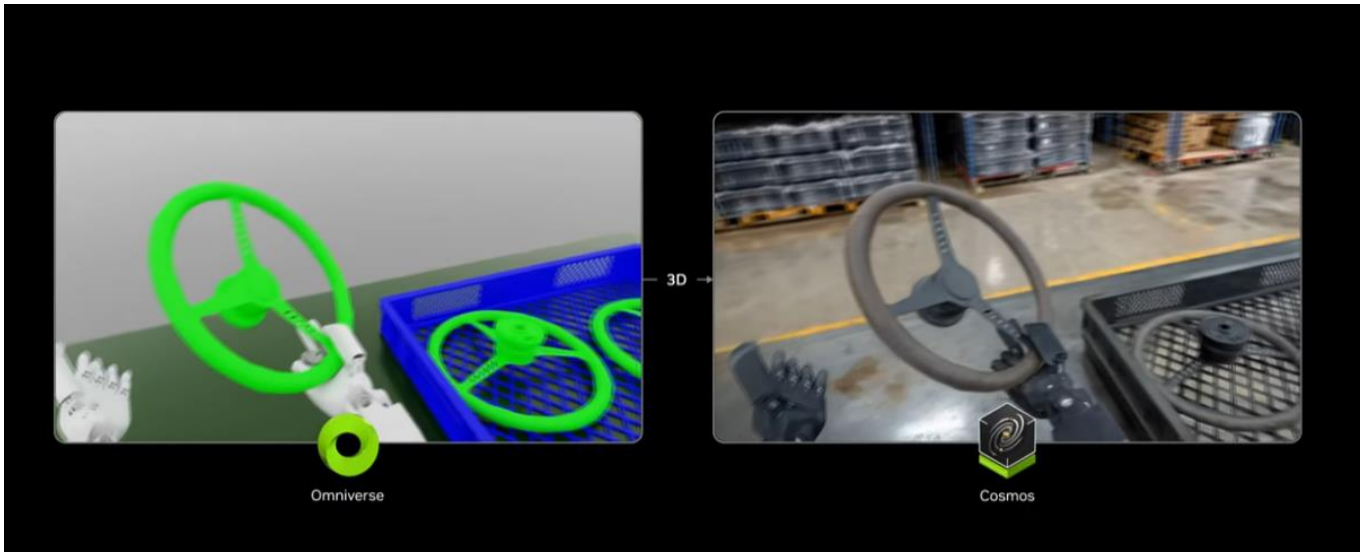
6-2. 'Cosmos': 세계 기반 모델(World Foundation Model)

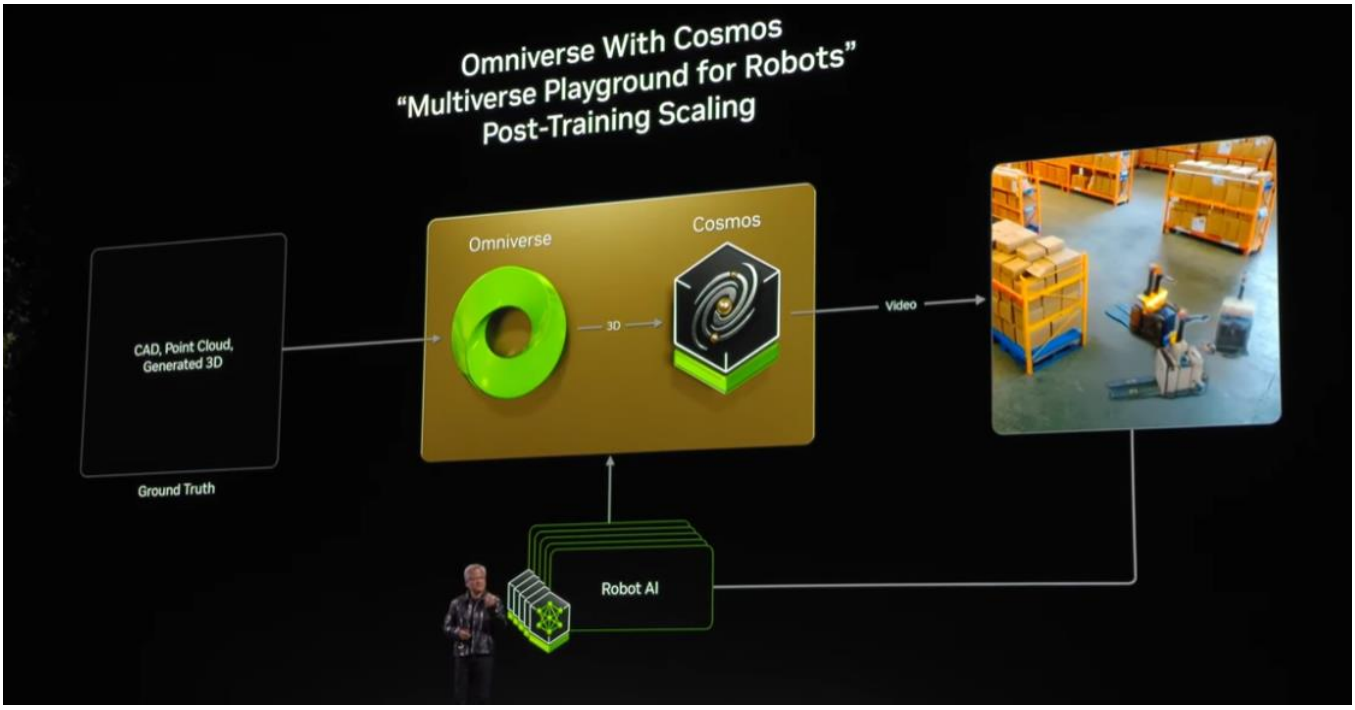
▶ Cosmos 모델

- 텍스트·이미지·비디오를 입력받아 물리적으로 일관된(물리 기반) 영상을 생성하는 WFM(World Foundation Model).
- 약 2000만 시간 분량의 물리적·동적인 테마의 비디오 데이터로 학습.
- AI가 "물리적 직관"을 학습해, 합성(시뮬레이션) 데이터 생성, 미래 예측, 객체 추적 등을 수행.

▶ 구성 요소

- Auto-regressive Model: 실시간 생성이 필요한 경우.
- Diffusion-based Model: 고품질 이미지·영상 생성.
- 고급 토크나이저: 이미지/비디오를 토큰화해 물리 세계의 '어휘'를 잘 표현.
- CUDA 가속 데이터 파이프라인: 페타바이트급 비디오 처리에 필요한 고속 파이프라인.





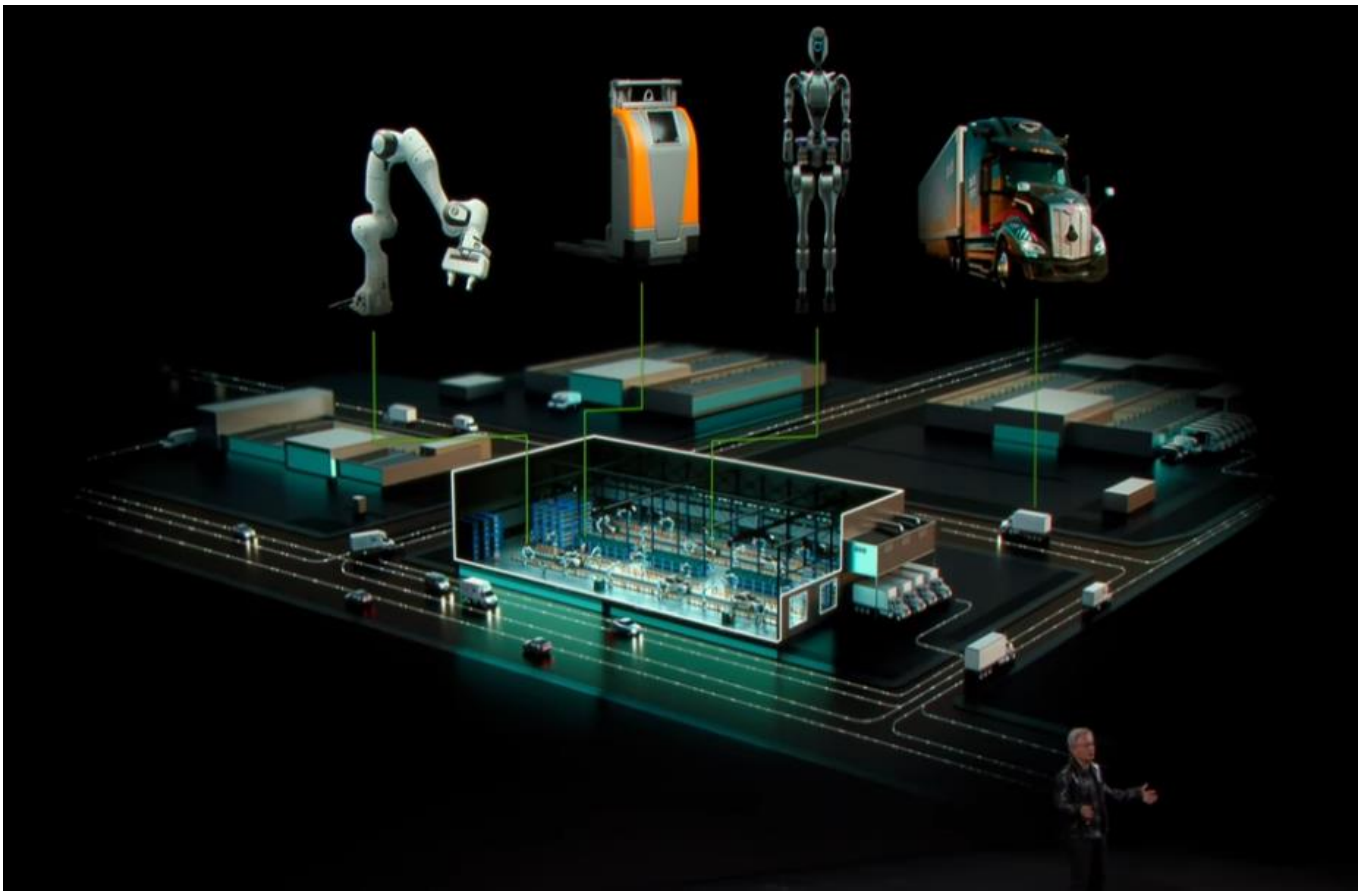
6. 물리적 AI(Physical AI)와 'Cosmos' 플랫폼

▶ Omniverse와 결합

- Omniverse(물리 시뮬레이션 엔진)와 Cosmos를 결합해 "진짜 같은 물리 세계"를 생성·조건부로 제어.
- AI 모델을 물리적으로 그라운딩(ground truth) 시키는 역할 → 로보틱스, 자율주행, 공장 시뮬레이션에 응용.

▶ 오픈 라이선스

- Cosmos 모델은 오픈 모델 라이선스로 제공(Hugging Face, NVIDIA NGC).
- 물리적 AI 생태계가 Llama 3처럼 빠르게 확산되길 기대.



7. 로보틱스(산업·휴머노이드·물류)와 디지털 트윈

7-1. Isaac Robot 플랫폼

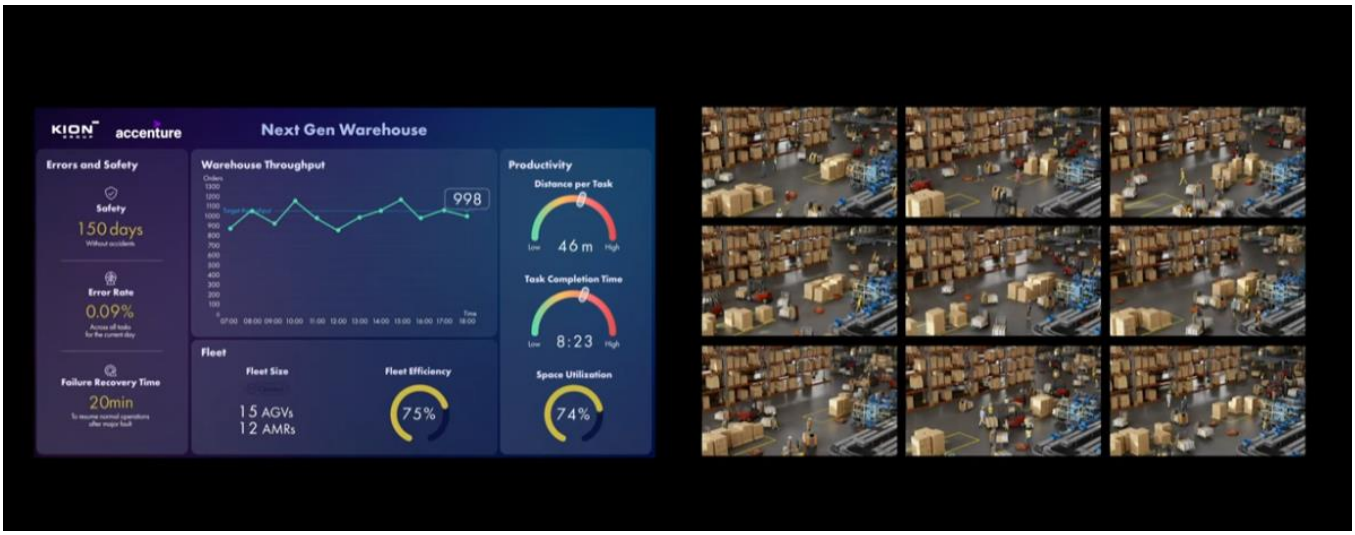
▶ Isaac Gym/Sim/Robot/GR

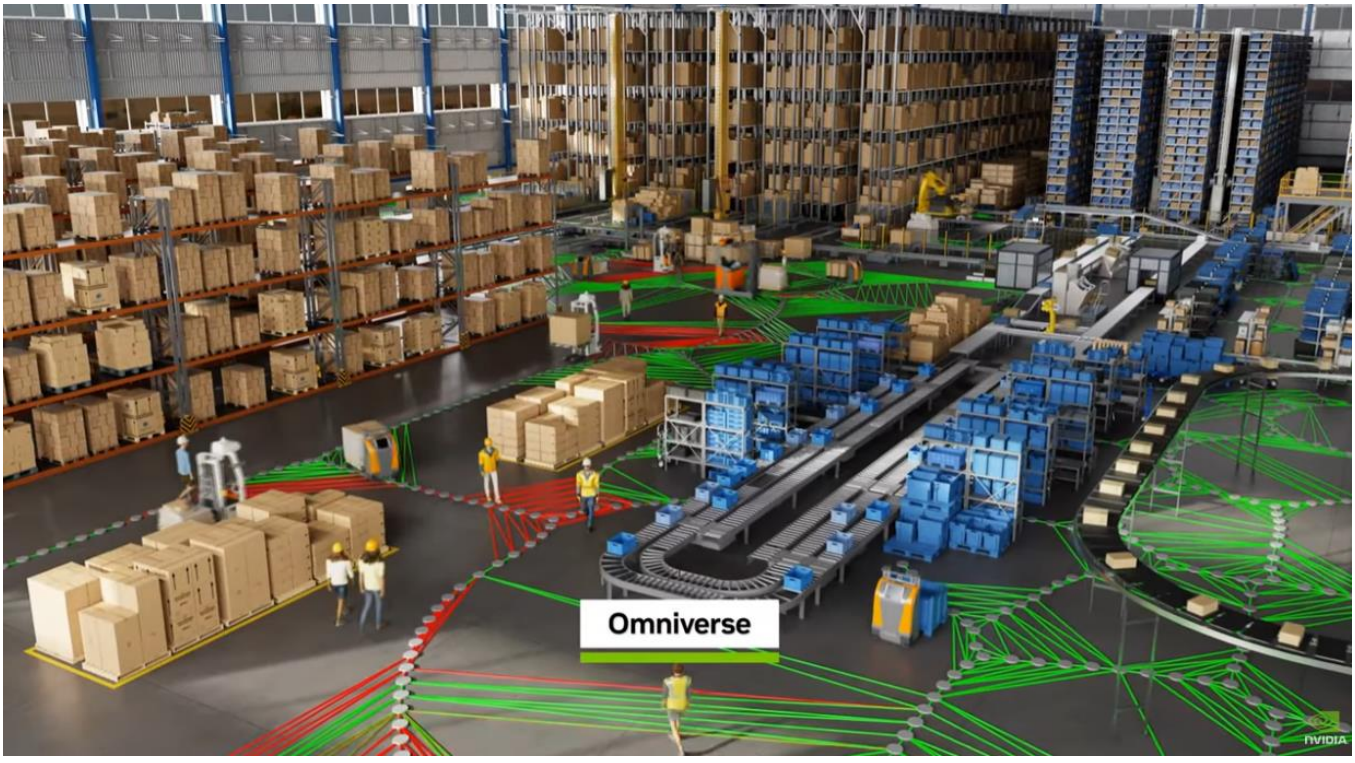
- Isaac은 로보틱스용 개발·시뮬레이션·운영 플랫폼.

- 휴머노이드 로봇에게 인간 시연(Tele-Operation)을 몇 차례만 주어도,

Omniverse+COSMOS로 수많은 변형 데이터셋을 생성해 학습.

- 실제 로봇에 적용 전에 가상환경에서 검증 가능.





7. 로보틱스(산업·휴머노이드·물류)와 디지털 트윈

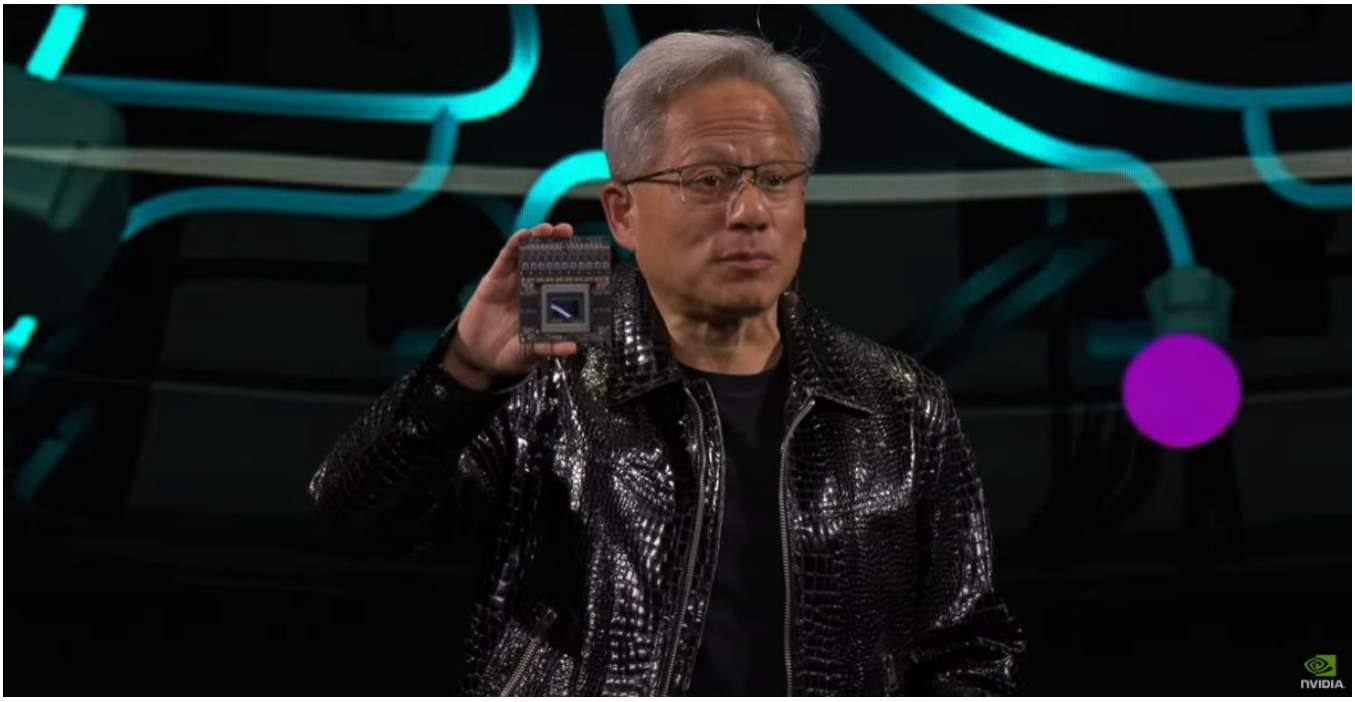
7-2. 산업 분야 (예: 창고 자동화)

▶ KION + Accenture 사례

- 창고 내 로봇·물류 시스템을 Omniverse로 디지털 트윈 구현 → 다양한 시나리오(수요 패턴, 재고 변화 등)를 - Cosmos+Omniverse로 시뮬레이션.

- 실제 투자/배치 전에 KPI를 측정, 운영 효율 극대화.

“Industrial Autonomy” 실현을 위한 엔드투엔드 솔루션.

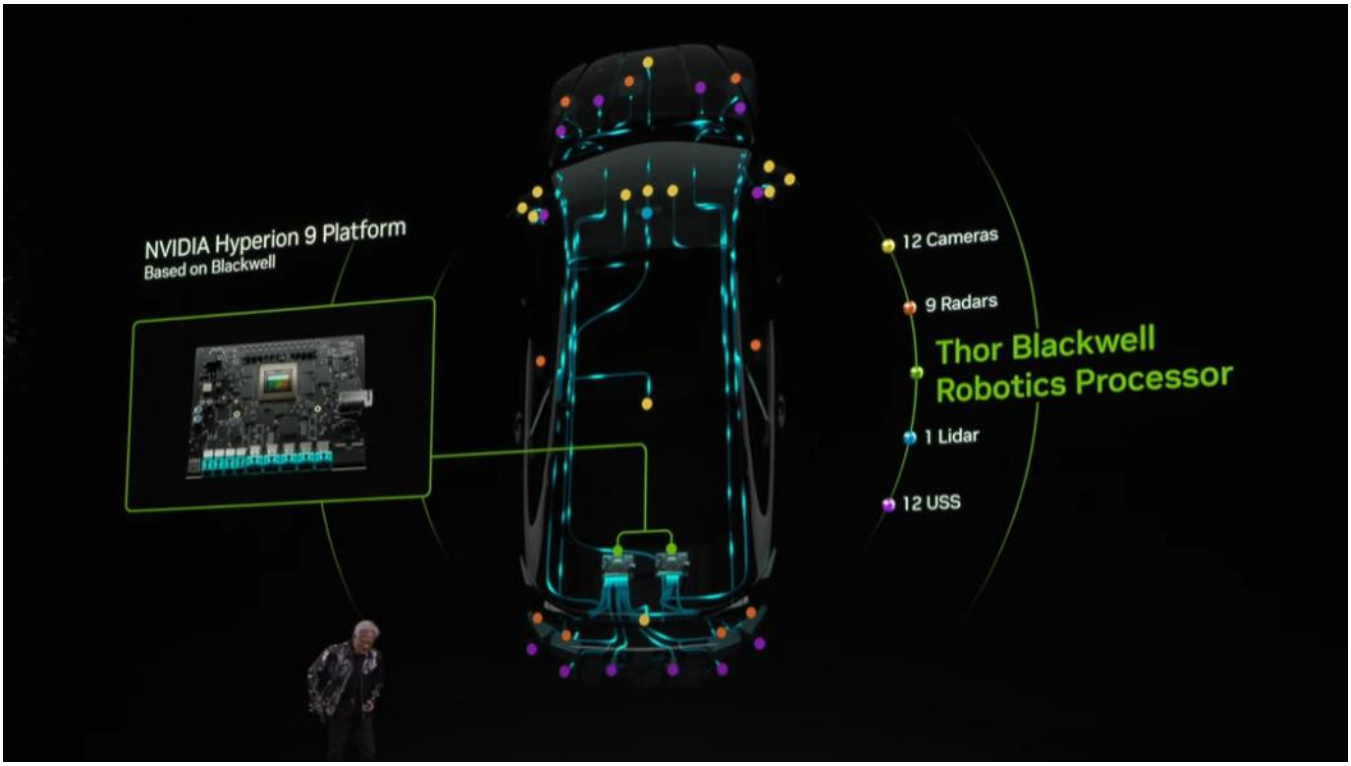


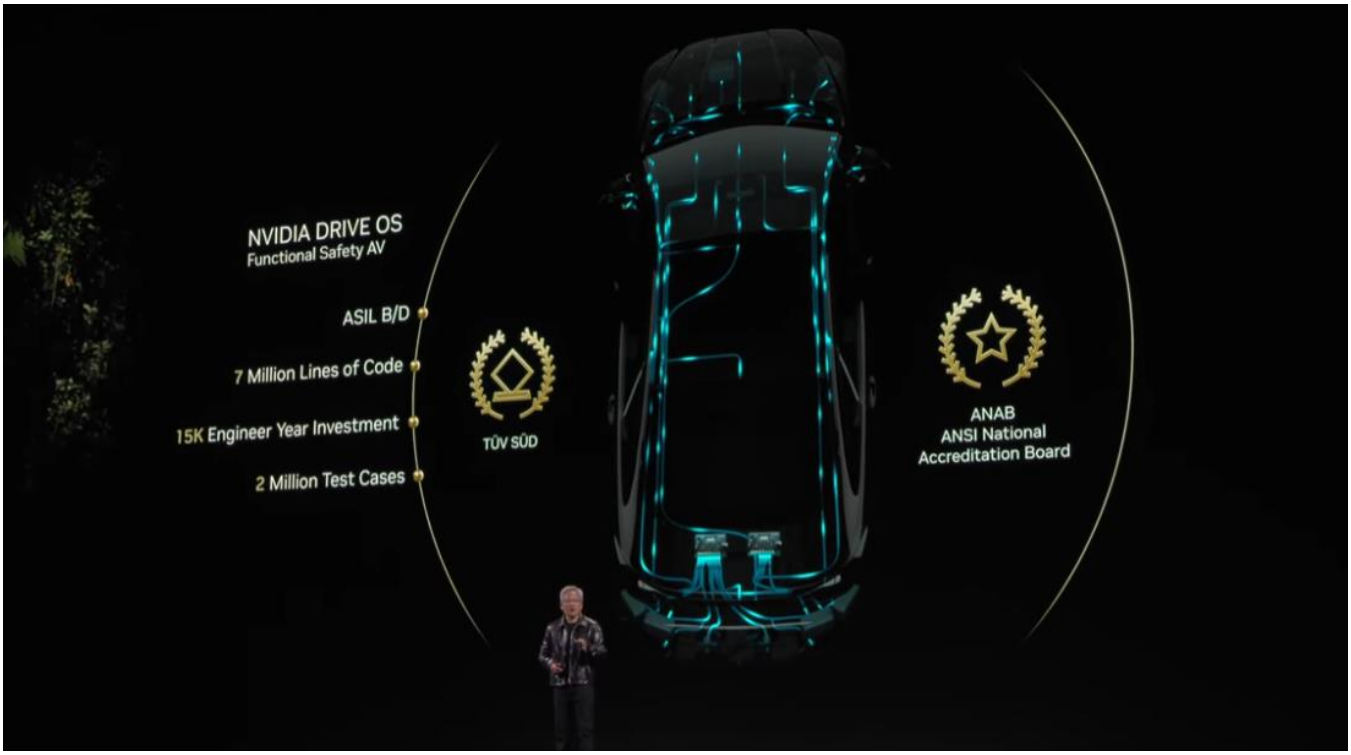
7. 로보틱스(산업·휴머노이드·물류)와 디지털 트윈

7-3. 자율주행(Autonomous Vehicle)

▶ Drive 플랫폼: 3개의 컴퓨팅 요소

- Data Center(훈련): DGX 등으로 AI 모델 훈련.
- Omniverse + Cosmos(시뮬레이션): 디지털 트윈으로 무한 시나리오 생성 및 검증.
- On-Board Computer(추론): 차량 탑재 컴퓨터인 'Thor(토르)'에 모델 배포.

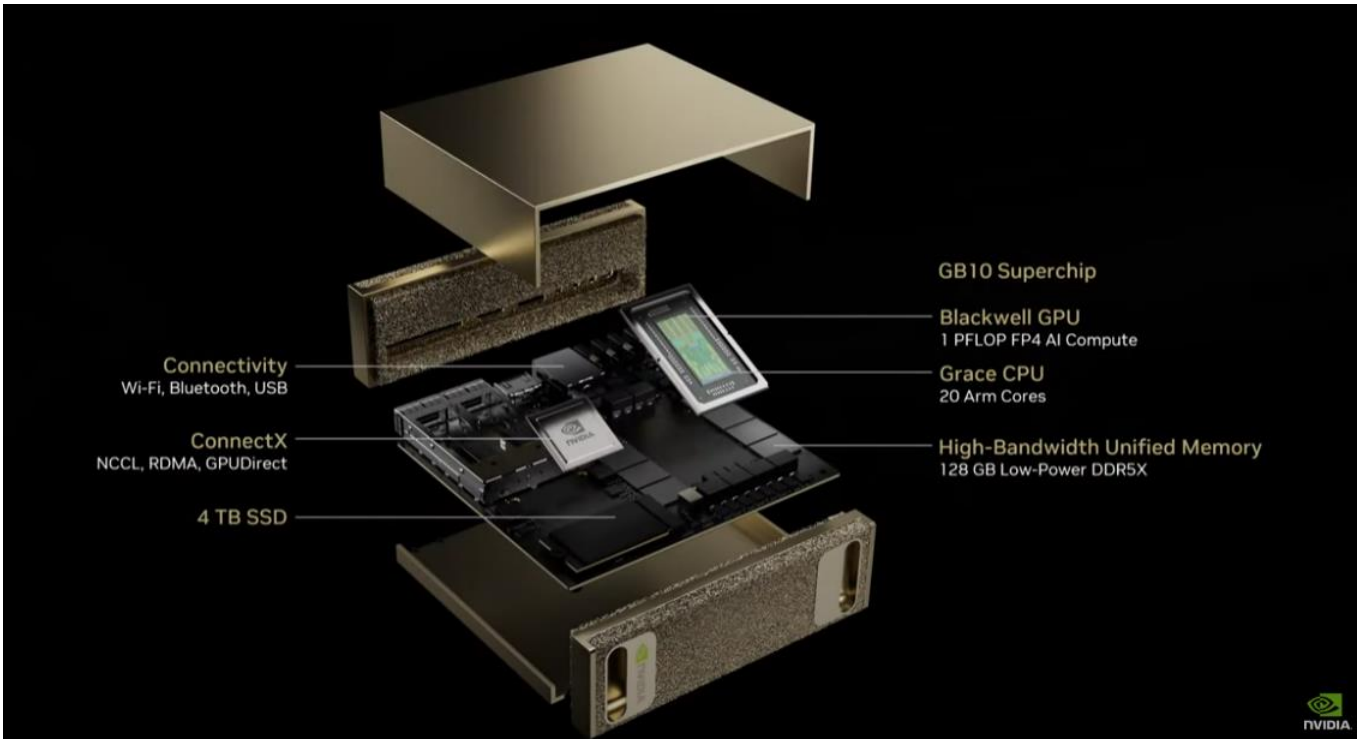




7. 로보틱스(산업·휴머노이드·물류)와 디지털 트윈

▶ Thor(토르)

- 차세대 차량용 SoC, Orin 대비 20배 성능.
- ISO 26262 ASIL-D 기능 안전 인증 달성, CPU+GPU 통합 기반.



8. DGX와 소형 AI 슈퍼컴 'Digits(프로젝트명)'

8-1. DGX의 역사

▶ 2016년 DGX-1 탄생: 연구소/스타트업에 "바로 꺼내쓰는 AI 슈퍼컴" 개념 제시.

- 첫 DGX-1을 OpenAI에 직접 전달.
- 이후 DGX 시리즈가 산업 전반으로 확산.

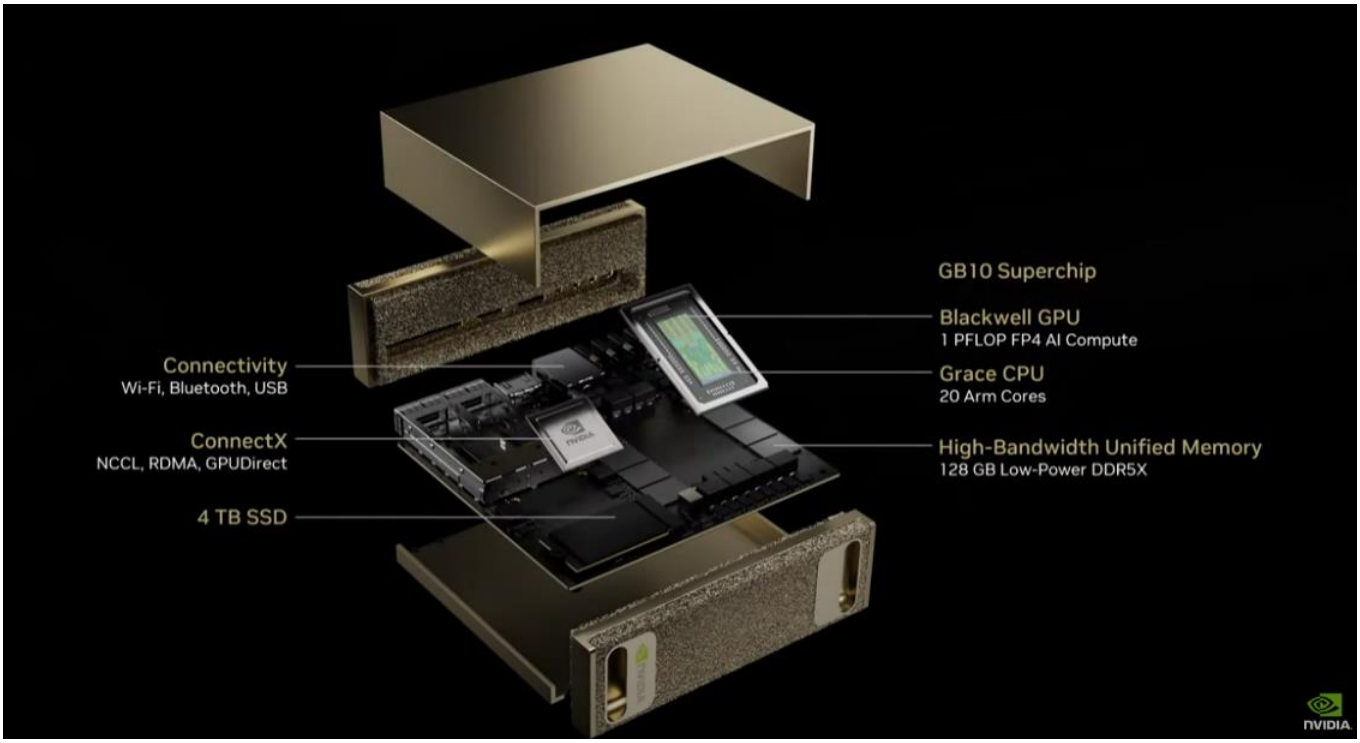
8-2. 소형 AI 슈퍼컴 'Digits' (프로젝트명)

▶ Grace+Blackwell 기반 초소형 시스템(GB110 칩 탑재)

- CPU(미디어텍 협력) + Blackwell GPU가 단일 기판에 SoC처럼 묶인 형태.
- 크기는 작은 데 비해, DGX급 AI 스택(CUDA, Triton, Nemo 등) 실행 가능.
- 책상 위에 놓을 수 있는 AI 슈퍼컴으로, 일반 개발자·중소기업도 자체 AI 활용 가능.
- 2024년 5월경 출시 목표.

▶ 확장 가능

- 여러 대를 NVLink나 ConnectX로 묶어 "Double Digits" 등 클러스터 구성.
- "클라우드 방식"의 온프레미스 AI HPC 개념.

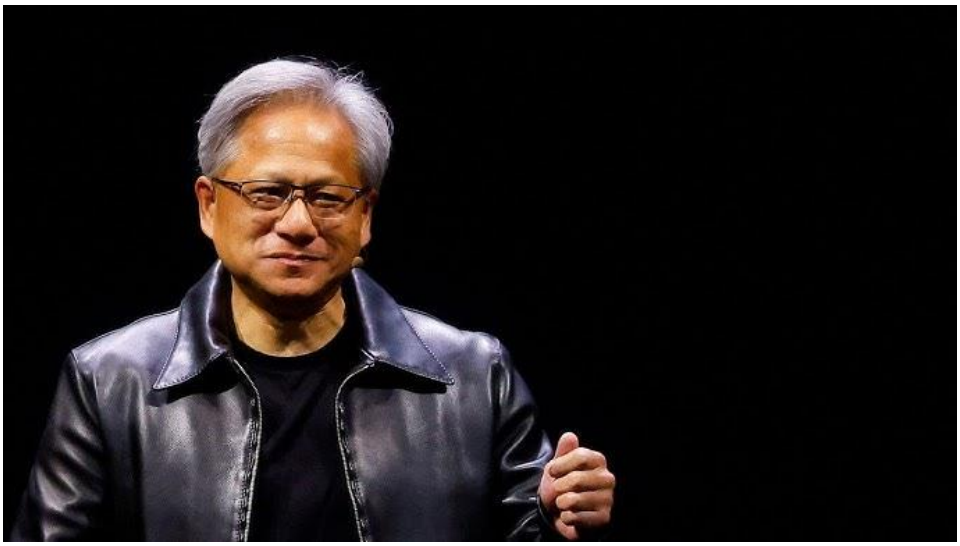


8. DGX와 소형 AI 슈퍼컴 'Digits(프로젝트명)'

▶ 엔비디아 DGX 스펙

- Blackwell GPU는 고성능 그래픽 처리와 AI 연산을, Grace CPU는 데이터 처리와 컴퓨팅을 담당.
- 두 칩셋은 함께 작동하여 고효율 연산 환경을 제공하는 데 초점을 맞춤.
- 1 Petaflop FP4 AI Compute(1초당 1경 연산 수행 가능).
- AI 연산 성능과 처리 능력을 극대화하여 대규모 AI 학습 및 추론에 최적화됨.
- 20개의 ARM 코어가 탑재되어 고성능 컴퓨팅을 지원.
- 데이터 처리와 서버 애플리케이션에 적합한 설계.
- 높은 대역폭의 통합 메모리 I/O를 통해 GPU와 CPU 간 데이터를 효율적으로 교환.
- Coherent Cache는 캐시 일관성을 유지하여 데이터 접근 속도를 최적화.
- GPU와 CPU 간 초고속 데이터 전송을 가능하게 하는 기술.
- 초당 막대한 양의 데이터를 처리하며 병목 현상을 최소화.
- 128GB의 DDR5X 메모리 제공, 낮은 전력 소모와 고속 데이터 전송을 지원.
- NVIDIA ConnectX와 통합되어 네트워크와 데이터 처리 성능을 더욱 강화.

NVIDIA의 Blackwell GPU와 Grace CPU가 결합된 고성능 컴퓨팅 아키텍처
대규모 AI 학습, 추론, 데이터 처리와 같은 고부하 작업을 지원하며, 초고속 통신 및 효율적인 메모리 관리를 통해 성능을 극대화



NVIDIA의 AI·GPU CES 내용 요약

▶ 새로운 하드웨어

- 데이터센터용 Blackwell(NVLink 72)
- 지포스 RTX 50 시리즈 (5070~5090)
- 차량용 Thor(토르) SoC
- 초소형 AI 슈퍼컴 'Digits'

▶ 새로운 소프트웨어·모델

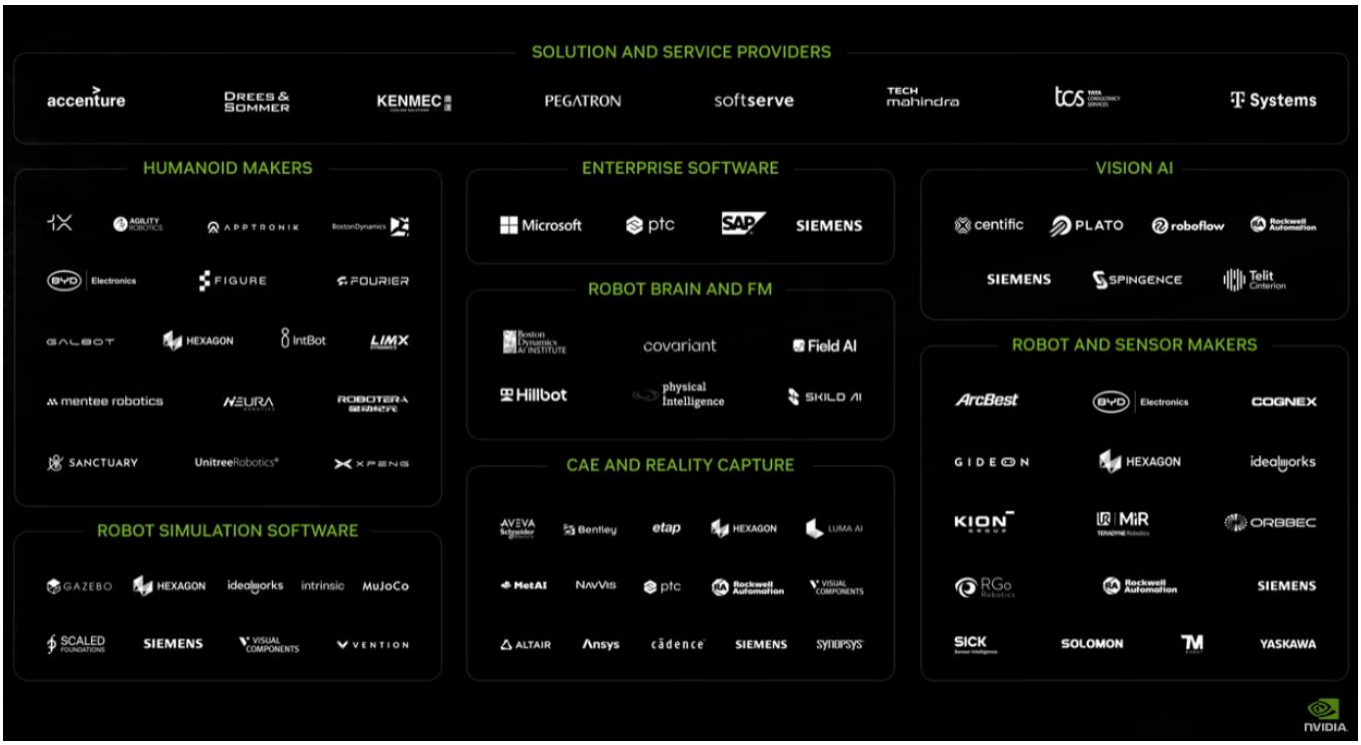
- Llama Nemotron(기업용 언어 모델 패밀리)
- Cosmos(WFM) + Omniverse 연동 → 로보틱스/자율주행/산업 자동화에 "물리적 AI" 제공.
- Agentic AI 스택(NIM Microservices, NeMo, Guardrails 등)

▶ 개방형 생태계

- Cosmos WFM, Llama Nemotron 등 오픈 라이선스나 오픈 모델로 제공.
- PC(Wsl2), 클라우드(모든 CSP), 온프레미스(DGX, Digits) 등 어디서든 접근 가능.

▶ 미래 전망

- '에이전틱 AI(Agentic AI)'가 추론 시 다단계 연산을 수행함에 따라 데이터센터 측 연산 수요 폭발.
- 자율주행(트럭·승용차), 휴머노이드 로봇, 창고·공장 자동화 등 거대 로보틱스 시장이 본격 개화.
- NVIDIA는 "3개의 컴퓨터(훈련, 시뮬레이션, 추론)" 전략을 통해 이 산업을 선도.



엔비디아 생태계 카테고리별 파트너들

1. Solution and Service Providers

- Accenture, Pegatron, Tech Mahindra 등 다양한 글로벌 기업들이 포함됨.
- 이들은 엔비디아의 기술을 기반으로 솔루션 및 서비스를 제공하는 역할을 함.

2. Humanoid Makers

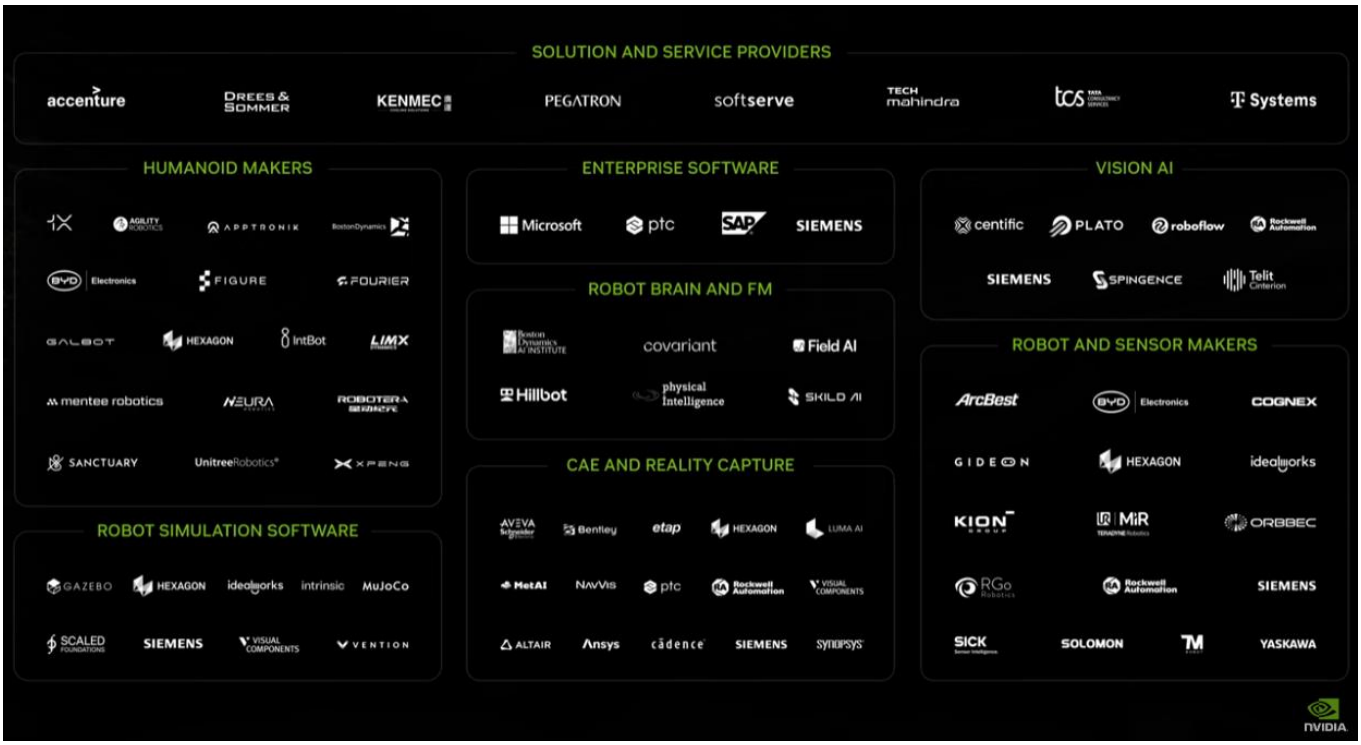
- Boston Dynamics, Agility Robotics, Figure 등 로봇 기술에 특화된 회사들로 구성됨.
- 인간형 로봇 개발 및 생산을 담당하며, 엔비디아의 AI와 하드웨어를 활용.

3. Enterprise Software

- Microsoft, SAP, Siemens 등이 포함됨.
- 엔비디아 기술을 활용해 엔터프라이즈 솔루션 및 소프트웨어 제공.

4. Robot Brain and FM (Function Management)

- Covariant, Field AI 등이 포함되어 있음.
- 로봇의 두뇌 역할을 하는 소프트웨어 및 AI 관리 기술을 개발.



엔비디아 생태계 카테고리별 파트너들

5. Vision AI

- Roboflow, Siemens 등이 포함됨.
- 컴퓨터 비전 및 AI를 통해 이미지, 비디오 데이터를 처리하고 분석하는 기술 제공.

6. Robot and Sensor Makers

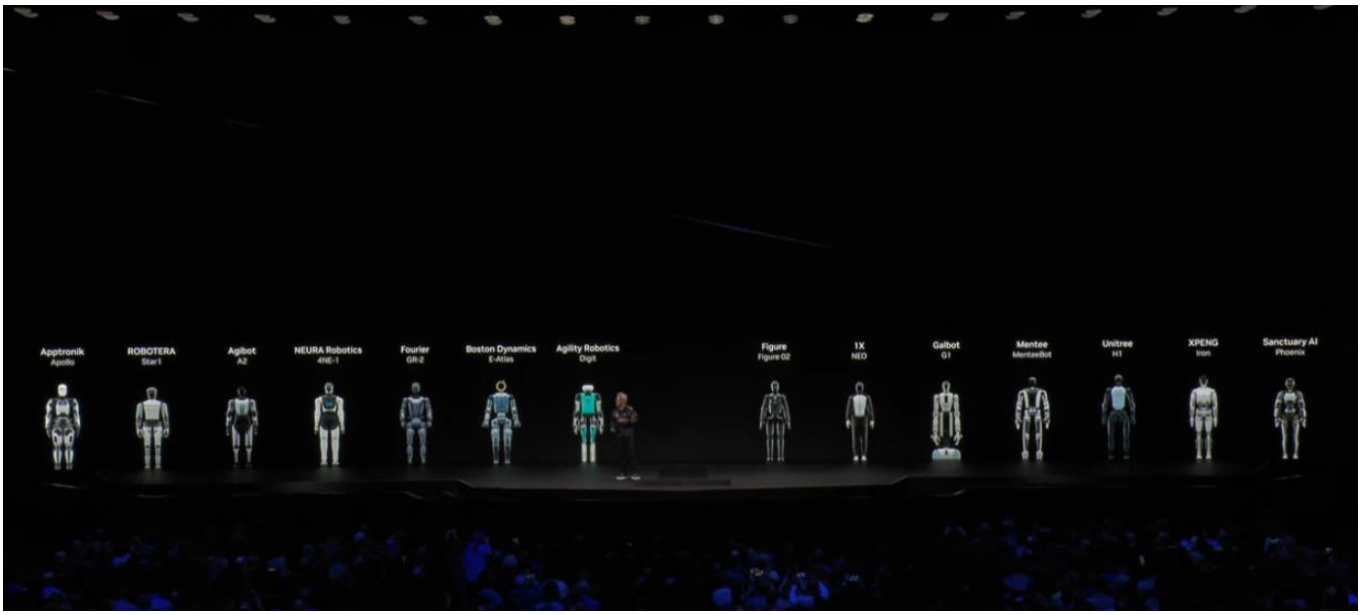
- BYD Electronics, Cognex, Yaskawa 등이 포함됨.
- 로봇 하드웨어 및 센서를 제작하는 기업들로 구성.

7. Robot Simulation Software

- Gazebo, Mujoco, Siemens 등이 포함됨.
- 로봇의 시뮬레이션 및 테스트를 위한 소프트웨어 솔루션을 제공.

8. CAE and Reality Capture

- ANSYS, Bentley, Siemens 등이 포함됨.
- 컴퓨터 지원 엔지니어링(CAE) 및 현실 캡처 기술에 특화.



엔비디아 CES 로봇 리스트

▶ Apptroinik - Apollo

휴머노이드 로봇으로 산업 및 물류 분야에 활용 가능.

▶ ROBOTERA - Star1

인간형 디자인으로 다양한 작업을 수행하도록 설계된 로봇.

▶ Agibot - A2

기본적인 휴머노이드 로봇으로 보이며, 특정 산업 작업에 초점을 맞춘 모델.

▶ NEURA Robotics - 4NE-1

정교한 설계와 함께 인간과 협력할 수 있는 기능이 강조된 로봇.

▶ Fourier - GR-2

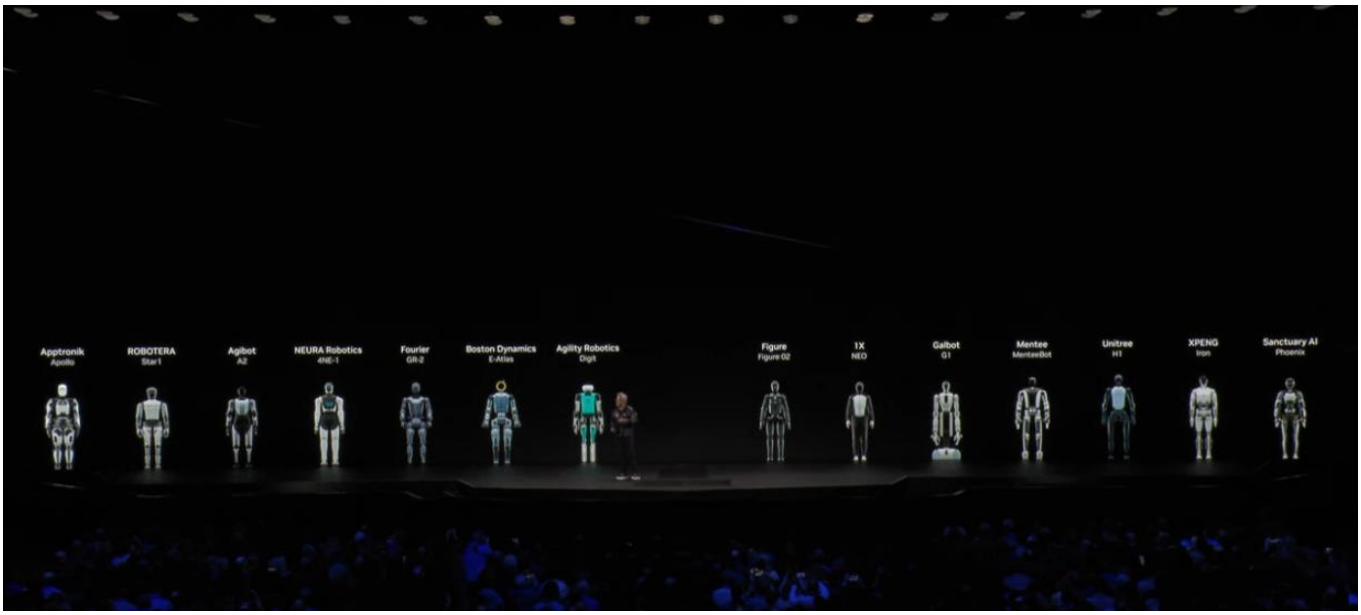
뛰어난 안정성과 적응성을 가진 모델로 보임.

▶ Boston Dynamics - E-Atlas

높은 수준의 유연성과 다목적성을 제공.

▶ Agility Robotics - Digit

물류 작업 및 다양한 환경에서 활용 가능한 유연한 휴머노이드.



엔비디아 CES 로봇 리스트

▶ Figure - Figure 02

인간과 유사한 외형 및 동작 기능을 갖춘 로봇.

▶ 1X - NEO

산업 및 서비스용으로 설계된 정교한 로봇.

▶ Galbot - G1

간단한 작업 및 교육 목적으로 설계된 모델로 추정.

▶ Mentee - MenteeBot

협력 작업 및 다양한 역할 수행이 가능한 로봇.

▶ Unitree - H1

기본적인 구조로 설계된 휴머노이드 로봇.

▶ XPENG - Iron

첨단 기술을 갖춘 로봇으로, 다양한 산업에서 활용 가능.

▶ Sanctuary AI - Phoenix

인공지능과 로봇 기술이 결합된 고급 모델.

Compliance Notice

- ✓ 동 자료에 게재된 내용은 조사분석담당자 본인의 의견을 정확히 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다.
- ✓ 동 자료는 투자 판단을 위한 정보제공일 뿐 해당 주식에 대한 가치를 보장하지 않습니다.
  투자판단은 본인 스스로 하며, 투자 행위와 관련하여 어떠한 책임도 지지 않습니다.
- ✓ 동 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙 자료로 사용될 수 없습니다.
- ✓ 당사는 해당 자료를 전문투자자 또는 제 3자에게 사전 제공한 사실이 없습니다.
- ✓ 동 자료의 작성자는 해당 기업의 유가증권을 보유하고 있을수도 있으며 발간 후에 매수·매도할 수 있습니다.
- ✓ 동 자료에 대한 저작권은 그로쓰리서치에 있습니다. 당사의 허락 없이 무단 복사 및 복제, 대여를 할 수 없습니다.