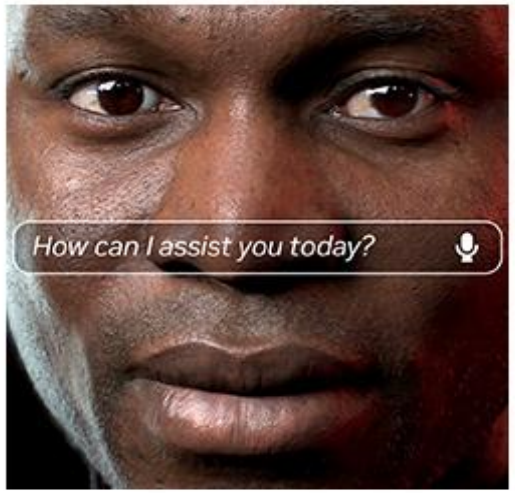
BN I&C

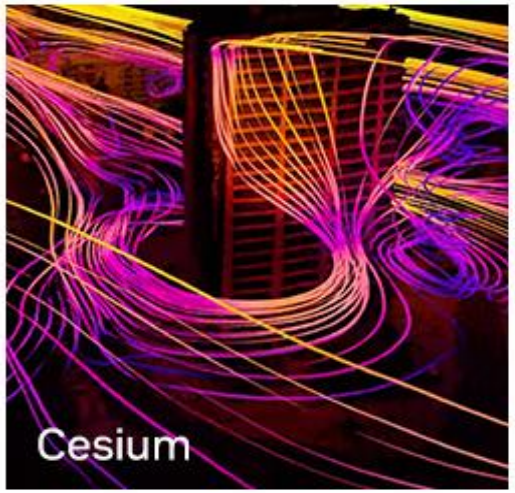
NVIDIA Elite Partner

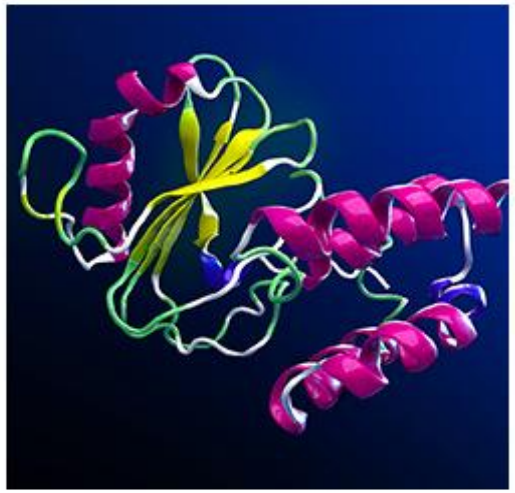
NVIDIA GTC

NVIDIA GTC25

KEYNOTE 요약







KEYNOTE SUMMARY

① NVIDIA Blackwell Ultra 플랫폼 공개

② Vera Rubin / Rubin Ultra 플랫폼 공개

③ NVIDIA Dynamo Opensource Library 공개

④ NVIDIA Spectrum-X, NVIDIA Quantum-X 실리콘 포토닉스 네트워킹 스위치 출시 예정

⑤ NVIDIA DGX Spark 및 DGX Station출시 발표

⑥ NVIDIA Cosmos World Foundation Models 및 Physical AI 데이터 도구 출시 발표

① NVIDIA Blackwell Ultra 플랫폼 공개

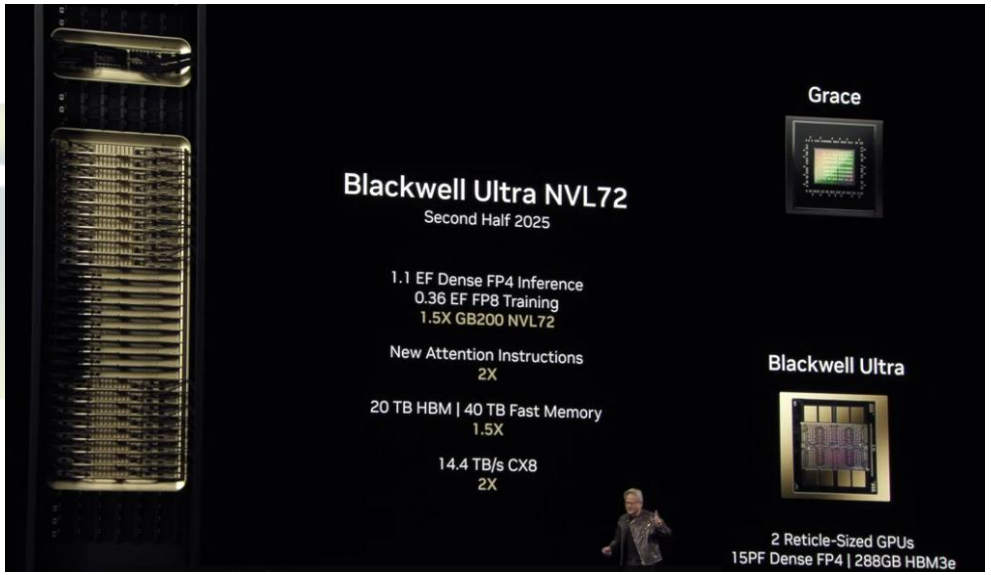
NVIDIA Blackwell Ultra AI Factory 플랫폼, AI 추론의 시대를 위한 길 마련

NVIDIA는 NVIDIA Blackwell AI 팩토리 플랫폼의 차세대 버전인 NVIDIA Blackwell Ultra를 발표하여 AI 추론 시대를 열었습니다.

NVIDIA Blackwell Ultra는 훈련 및 테스트 시간 확장 추론(추론 중에 더 많은 컴퓨팅을 적용하여 정확도를 개선하는 기술)을 향상시켜 전 세계 조직이 AI 추론, 에이전트 AI 및 물리적 AI와 같은 애플리케이션을 가속화할 수 있도록 합니다.

1년 전에 소개된 획기적인 블랙웰 아키텍처를 기반으로 구축된 블랙웰 울트라에는 NVIDIA GB300 NVL72 랙 스케일 솔루션과 NVIDIA HGX™ B300 NVL16 시스템이 포함되어 있습니다.

GB300 NVL72는 NVIDIA GB200 NVL72보다 1.5배 더 높은 AI 성능을 제공할 뿐만 아니라 NVIDIA Hopper™로 구축된 공장에 비해 AI 공장에 대한 Blackwell의 수익 기회를 50배 증가시킵니다.



<NVIDIA Blackwell Ultra NVL72>

엔비디아의 창립자 겸 CEO인 젠슨 황(Jensen Huang)은 "AI는 거대한 도약을 이뤘으며, 추론과 에이전트 AI는 훨씬 더 높은 컴퓨팅 성능을 요구한다"고 말했습니다. 또한 "우리는 이 순간을 위해 블랙웰 울트라를 설계했다"며 "블랙웰 울트라는 사전 훈련, 훈련 후 및 추론 AI 추론을 쉽고 효율적으로 수행할 수 있는 다재다능한 단일 플랫폼"이라고 말했습니다.

블랙웰 울트라 기반 제품은 2025년 하반기부터 파트너를 통해 제공될 것으로 예상됩니다.

① NVIDIA Blackwell Ultra 플랫폼 공개

블랙웰 울트라: NVIDIA GB300 NVL72

AI 추론 시대를 위한 NVIDIA Blackwell Ultra

	GB300 NVL72	GB200 NVL72 대	HGX H100 대
FP4 추론[1]	1.4 나 1.1 엑사플롭스	1.5배	70배
HBM 메모리	20 테라바이트	1.5배	30배
빠른 메모리	40 테라바이트	1.3배	65배
네트워킹 대역폭	14.4TB/초	2배	20배

*표 1. NVIDIA GB200, NVL72 및 NVIDIA HGX H100*과 비교한 NVIDIA Blackwell Ultra 사양

NVIDIA GB300 NVL72는 72개의 Blackwell Ultra GPU와 36개의 Arm Neoverse 기반 NVIDIA Grace™ CPU를 랙 규모 설계에 연결하여 테스트 시간 확장을 위해 구축된 단일 대규모 GPU 역할을 합니다.

GB300 NVL72는 진화하는 워크로드에 맞춰 소프트웨어, 서비스, AI 전문 지식으로 성능을 최적화하는 선도적인 클라우드의 엔드 투 엔드 완전 관리형 AI 플랫폼인 NVIDIA DGX™ Cloud에서도 사용할 수 있을 것으로 예상됩니다.

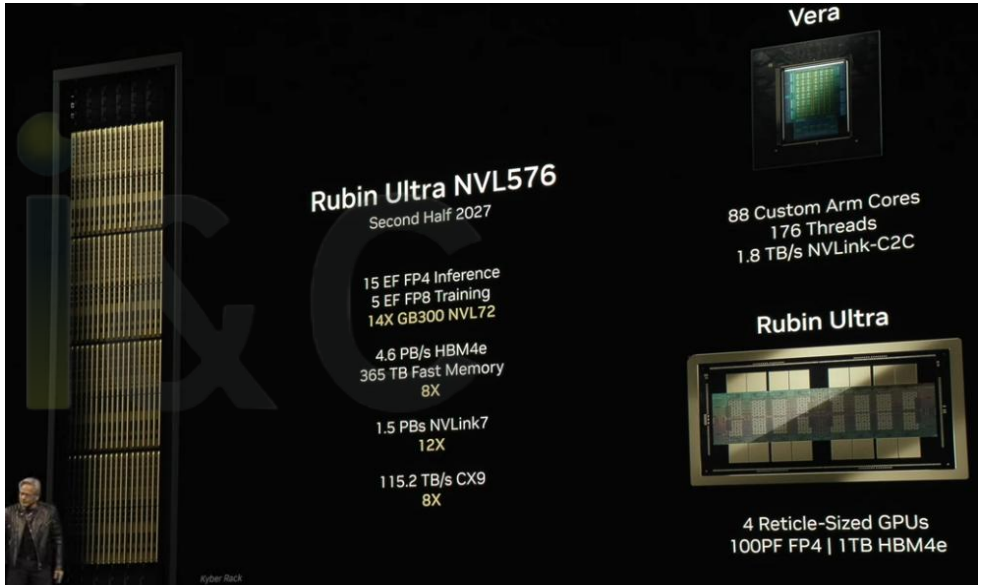
DGX GB300 시스템을 탑재한 NVIDIA DGX SuperPOD™는 GB300 NVL72 랙 설계를 사용하여 고객에게 턴키 AI 팩토리를 제공합니다.

NVIDIA HGX B300 NVL16은 Hopper 세대에 비해 대규모 언어 모델에서 11배 더 빠른 추론, 7배 더 많은 컴퓨팅 및 4배 더 큰 메모리를 제공하여 AI 추론과 같은 가장 복잡한 워크로드에 획기적인 성능을 제공합니다.

② Vera Rubin / Rubin Ultra 플랫폼 공개

Vera Rubin NVL144은 2026년 하반기, Rubin Ultra NVL576은 2027년 하반기에 플랫폼 출시 발표

<NVIDIA Vera Rubin NVL144>

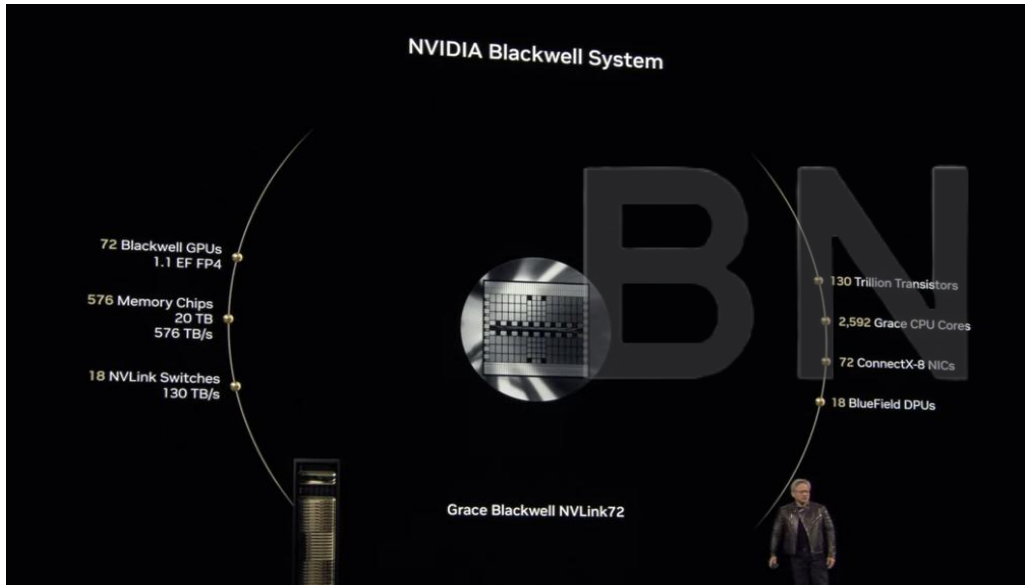


<NVIDIA Rubin Ultra NVL576>

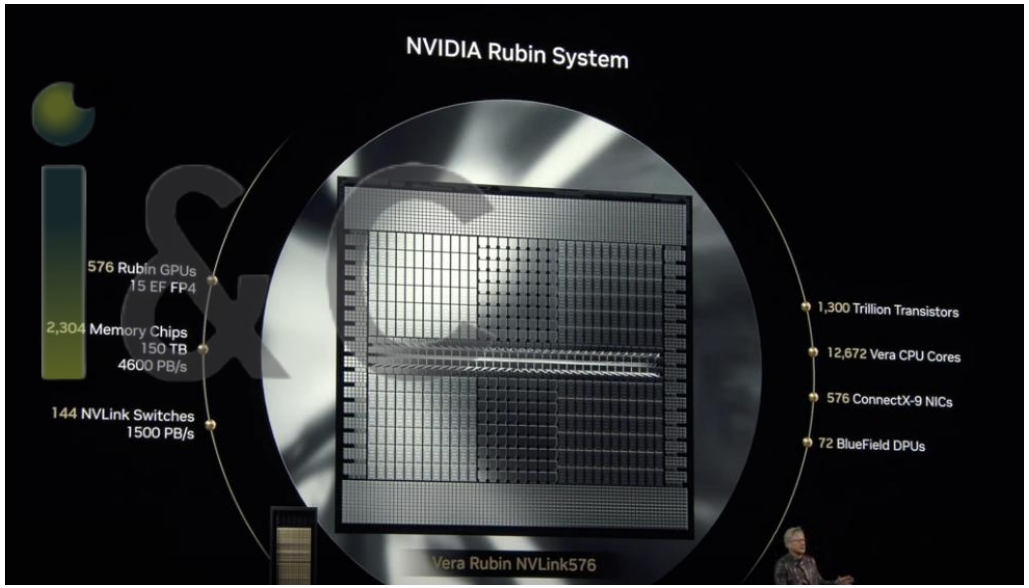
젠슨 황은 천문학자 베라 루빈(Vera Rubin)에게 경의를 표하며 향후 몇 년 동안 데이터센터 성능 향상을 제공할 로드맵을 설명하고,
혁신으로 가득 찬 차세대 NVIDIA Rubin Ultra GPU 및 NVIDIA Vera CPU 아키텍처에 대한 새로운 세부 정보를 제공했습니다.

Vera Rubin NVL 144를 포함하여 Rubin Ultra를 기반으로 구축된 시스템은 내년 하반기에 출시될 예정입니다.
그리고 2027년 하반기에는 Rubin Ultra를 기반으로 구축된 시스템이 출시될 예정입니다.

※ Blackwell & Rubin System 비교 슬라이드



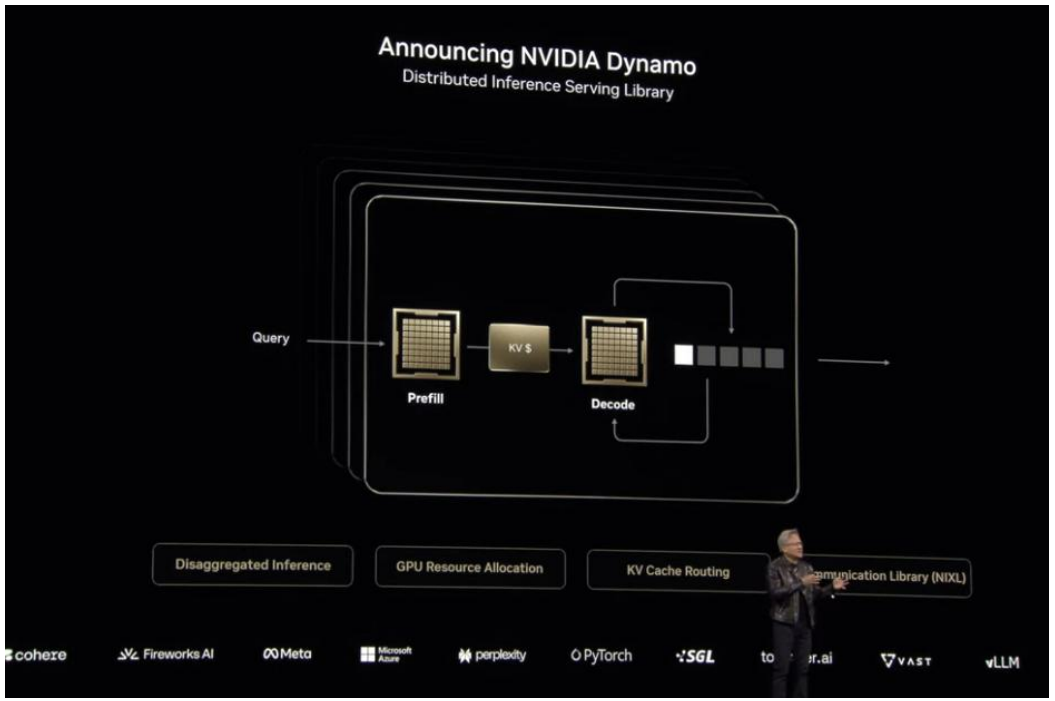
<NVIDIA Blackwell System>



<NVIDIA Rubin System>

③ NVIDIA Dynamo Opensource Library 공개

AI 추론 모델을 가속화하고 확장하기 위한 오픈 소스 추론 소프트웨어 NVIDIA Dynamo



<NVIDIA Dynamo>

NVIDIA Triton Inference Server™의 후속 제품인 NVIDIA Dynamo는 추론 AI 모델을 배포하는 AI 공장의 토큰 수익 창출을 극대화하도록 설계된 새로운 AI 추론 제공 소프트웨어입니다.

수천 개의 GPU에서 추론 통신을 오케스트레이션 및 가속화하고, 세분화된 서비스를 사용하여 서로 다른 GPU에서 대규모 언어 모델(LLM)의 처리 및 생성 단계를 분리합니다. 이를 통해 각 단계를 특정 요구 사항에 맞게 독립적으로 최적화할 수 있으며 GPU 리소스 사용률을 극대화할 수 있습니다.

다이나모는 동일한 수의 GPU를 사용하여 오늘날의 NVIDIA Hopper™ 플랫폼에서 라마 모델을 지원하는 AI 공장의 성능과 수익을 두 배로 늘립니다.

GB200 NVL72 랙의 대규모 클러스터에서 DeepSeek-R1 모델을 실행할 때 NVIDIA Dynamo의 지능형 추론 최적화는 GPU당 생성되는 토큰 수를 30배 이상 증가시킵니다.

NVIDIA Dynamo는 NVIDIA NIM™ 마이크로서비스에서 사용할 수 있으며 프로덕션 수준의 보안, 지원 및 안정성을 갖춘 NVIDIA AI Enterprise 소프트웨어 플랫폼의 향후 릴리스에서 지원됩니다.

④ NVIDIA Spectrum-X™, NVIDIA Quantum-X 출시 예정

NVIDIA Spectrum-X™ 및 NVIDIA Quantum-X 실리콘 포토닉스 네트워킹 스위치 공개

엔비디아의 창립자 겸 CEO인 젠슨 황(Jensen Huang)은 "AI 팩토리는 엄청난 규모의 새로운 차원의 데이터센터이며, 이에 발맞추기 위해 네트워킹 인프라도 재창조되어야 한다"고 말했습니다. "NVIDIA는 실리콘 포토닉스를 스위치에 직접 통합함으로써 하이퍼스케일 및 엔터프라이즈 네트워크의 오래된 한계를 깨고 수백만 GPU AI 공장의 문을 열고 있습니다."

NVIDIA Silicon Photonics 네트워킹 스위치는 NVIDIA Spectrum-X Photonics 이더넷 및 NVIDIA Quantum-X Photonics InfiniBand 플랫폼의 일부로 사용할 수 있습니다. Spectrum-X 이더넷 네트워킹 플랫폼은 세계 최대 규모의 슈퍼컴퓨터를 포함한 멀티 테넌트, 하이퍼스케일 AI 팩토리를 위해 기존 이더넷에 비해 뛰어난 성능과 1.6배의 대역폭 밀도를 제공합니다.

NVIDIA Spectrum-X Photonics 스위치에는 800Gb/s 포트 128개 또는 200Gb/s 포트 512개를 포함한 여러 구성이 포함되어 있어 총 대역폭 100Tb/s를 제공할 뿐만 아니라 800Gb/s 포트 512개 또는 200Gb/s 포트 2,048개를 포함하여 총 처리량 400Tb/s를 제공합니다.

<NVIDIA Photonics>　　<NVIDIA Spectrum-X Ethernet>　　<NVIDIA Photonics Switch Systems>

⑤ NVIDIA DGX Spark 및 DGX Station 출시 발표

NVIDIA Grace Blackwell 플랫폼으로 구동되는 NVIDIA DGX™ 개인용 AI 슈퍼컴퓨터를 공개

<NVIDIA DGX Spark>

<NVIDIA DGX Spark>

- NVIDIA GB10 Grace Blackwell 슈퍼칩
- FP4 AI 성능 1,000 AI TOPS
- 128GB의 일관성 있는 통합 시스템 메모리
- ConnectX-7 스마트 NIC
- 최대 4TB의 저장 용량
- 150mm L x 150mm W x 50.5mm H

DGX 스파크(DGX Spark, 구 프로젝트 디지츠(Project DIGITS))와 엔비디아 블랙웰 울트라(NVIDIA Blackwell Ultra) 플랫폼을 기반으로 하는 새로운 고성능 엔비디아 그레이스 블랙웰(NVIDIA Grace Blackwell) 데스크톱 슈퍼 컴퓨터인 DGX 스테이션(DGX Station™)은 AI 개발자, 연구원, 데이터 사이언티스트 및 학생들이 데스크톱에서 대규모 모델의 프로토타입을 제작하고, 미세 조정하고, 추론할 수 있도록 지원합니다.

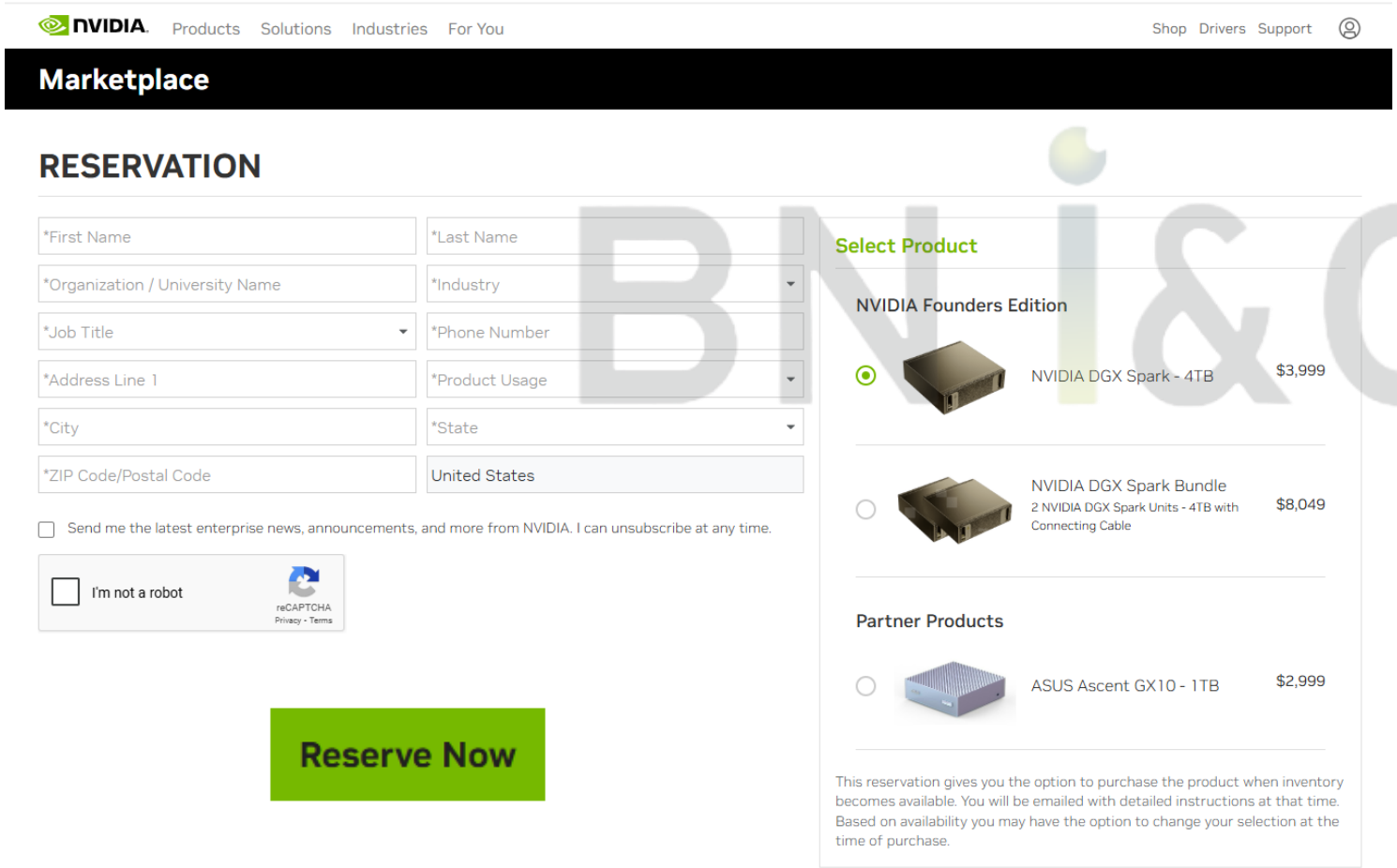
사용자는 이러한 모델을 로컬에서 실행하거나 NVIDIA DGX 클라우드 또는 기타 가속 클라우드 또는 데이터센터 인프라에 배포할 수 있습니다. DGX Spark 및 DGX Station은 이전에는 데이터 센터에서만 사용할 수 있었던 race Blackwell 아키텍처의 성능을 데스크톱에 제공합니다.

DGX Spark 는 NVIDIA GB10 Grace Blackwell 슈퍼칩, NVIDIA DGX 기반™ Spark는 전력 효율이 높은 소형 폼 팩터로 1,000 AI TOPS의 AI 성능을 제공합니다. NVIDIA AI 소프트웨어 스택이 사전 설치되고 128GB의 메모리를 통해 개발자는 최대 2,000억 개의 매개변수가 있는 DeepSeek, Meta, Google 등의 최신 추론 AI 모델을 로컬에서 프로토타입화, 미세 조정 및 추론하고 데이터센터 또는 클라우드에 원활하게 배포할 수 있습니다.

NVIDIA DGX Station은 AI 개발 및 실행을 위해 처음부터 새롭게 설계된 새로운 종류의 컴퓨터입니다. NVIDIA GB300 Grace Blackwell Ultra 데스크톱 슈퍼칩과 최대 784GB의 대용량 코히어런트 메모리가 탑재된 최초의 시스템으로, 데스크톱에서 대규모 AI 훈련 및 추론 워크로드를 개발하고 실행할 수 있도록 전례 없는 컴퓨팅 성능을 제공합니다. 최첨단 시스템 기능과 NVIDIA® CUDA X-AI™ 플랫폼을 결합한 DGX Station은 최고의 데스크톱 AI 개발 플랫폼을 필요로 하는 팀을 위해 특별히 설계되었습니다.

⑤ NVIDIA DGX Spark 출시 발표

※ DGX Spark 예약페이지 (참고)



⑥ NVIDIA Cosmos World Foundation Models & NVIDIA Omniverse Physical AI 데이터 도구확장

예측, 세계 생성 및 추론을 위한 새로운 모델로 전례 없는 제어를 경험하다

NVIDIA는 물리적 AI 개발을 위한 개방적이고 완전히 사용자 정의 가능한 추론 모델을 도입하고 개발자에게 세계 생성에 대한 전례 없는 제어를 제공하는 새로운 NVIDIA Cosmos™ WFM(World Foundation Model)의 출시를 발표했습니다.

합성 데이터 생성을 위한 Cosmos Transfer Cosmos Transfer WFM은 세분화 맵, 깊이 맵, 라이더 스캔, 포즈 추정 맵 및 궤적 맵과 같은 구조화된 비디오 입력을 수집하여 제어 가능한 사실적인 비디오 출력을 생성합니다.

코스모스 트랜스퍼(Cosmos Transfer)는 인식 AI 트레이닝을 간소화하여 Omniverse에서 생성된 3D 시뮬레이션 또는 실측 자료를 제어 가능한 대규모 합성 데이터 생성을 위한 사실적인 비디오로 변환합니다.

자율주행 자동차 시뮬레이션을 위한 NVIDIA Omniverse Blueprint는 코스모스 트랜스퍼(Cosmos Transfer)를 사용하여 물리 기반 센서 데이터의 변형을 증폭합니다.

합성 조작 모션 생성을 위한 NVIDIA GR00T 블루프린트는 Omniverse와 코스모스 트랜스퍼(Cosmos Transfer)를 결합하여 다양한 데이터셋을 대규모로 생성하고, OpenUSD 기반 시뮬레이션의 이점을 활용하고 데이터 수집 및 증강 시간을 며칠에서 몇 시간으로 단축합니다.

코스모스 WFM은 NVIDIA API 카탈로그에서 미리 볼 수 있으며 이제 Google Cloud의 Vertex AI Model Garden에 나열되어 있습니다.
코스모스 예측과 코스모스 트랜스퍼는 허깅페이스와 깃허브에서 공개적으로 이용할 수 있습니다. Cosmos Reason은 얼리 액세스로 이용할 수 있습니다.

AI Industry Trends

① Agent AI와 Physical AI주목, 블랙웰 AI 인프라 성장 예상

② 대규모 AI Factory 로의 전환

③ General Motors와 NVIDIA, AI를 위해 협업

④ 6G용 AI 네이티브 무선 네트워크 개발 및 협력

⑤ 엔터프라이즈용 AI 데이터 플랫폼 발표

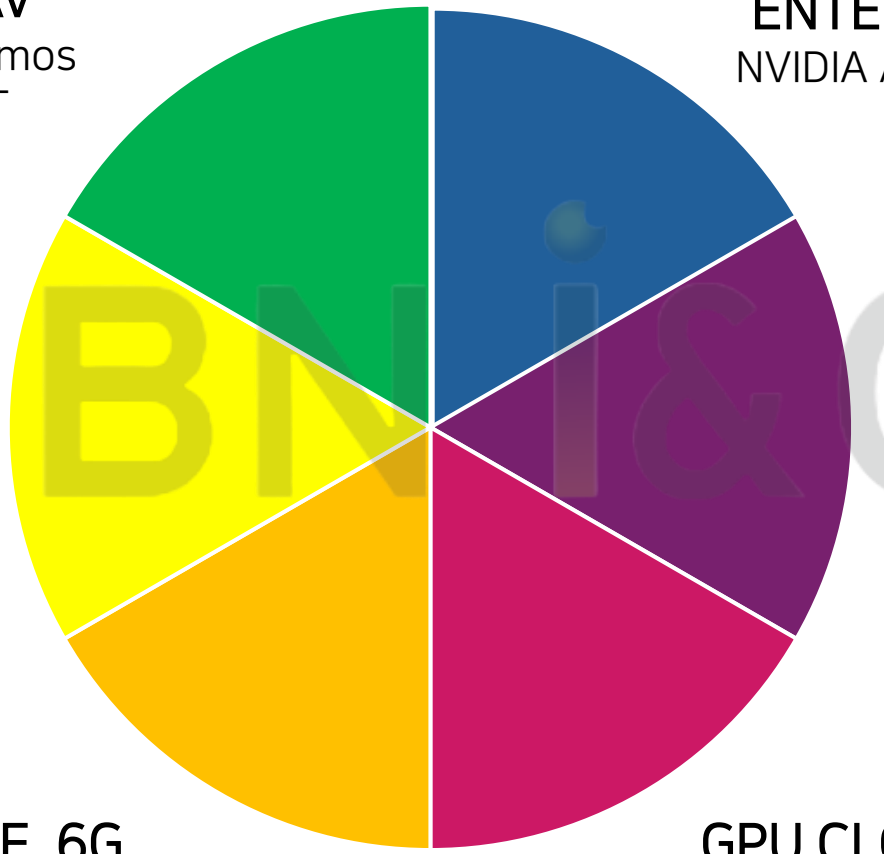
⑥ NVIDIA Omniverse Physical AI 운영 체제의 활용

⑦ Physical AI와 로보틱스

AI for Every Industry

ROBOTS, AV
Omniverse-Cosmos
Issac GROOT
Drive AI

ENTERPRISE IT
NVIDIA AI Enterprise



CLOUD SERVICE PROVIDERS

Leading Technology
Full-Stack AI
CUDA-X Diverse Applications
Rich Developer Ecosystem

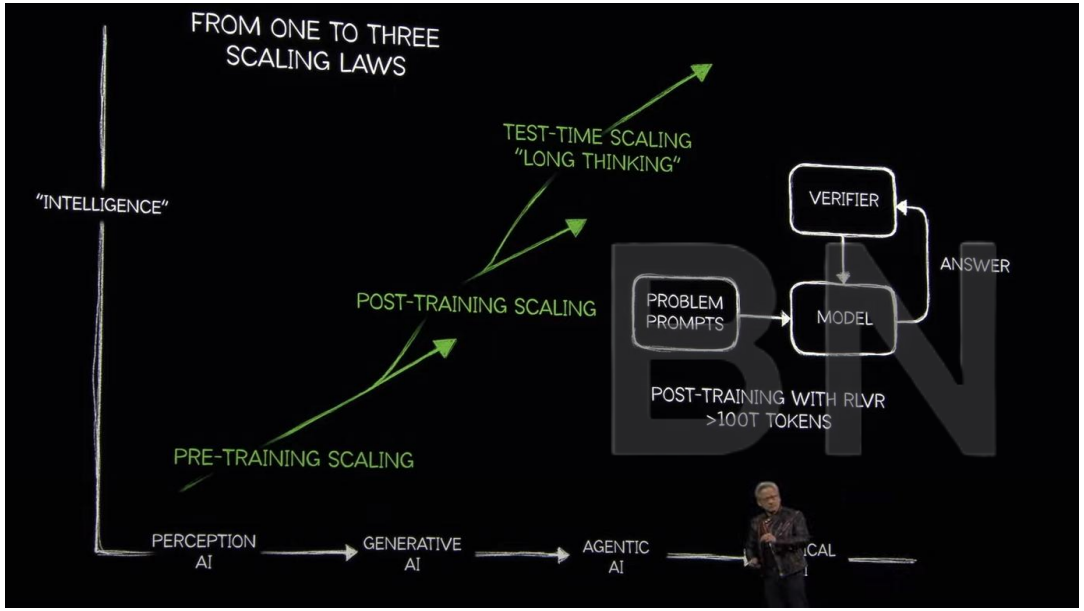
ROBOTIS MANIFACTURING
Omniverse-Cosmos

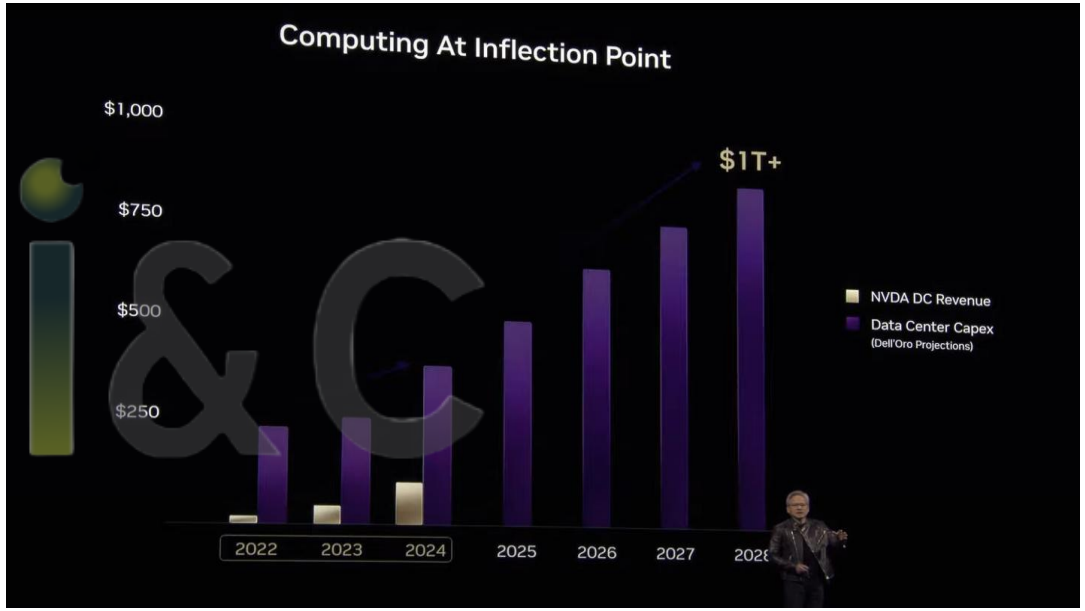
EDGE, 6G
Aerial-Sionna
6G AI-RAN

GPU CLOUDS
AI Factory
Blueprint and Full Stack

① Agent AI와 Physical AI, 블랙웰 AI 인프라 성장

AI 의 변곡점, AI 컴퓨팅 수요로 인한 블랙웰 AI인프라의 성장

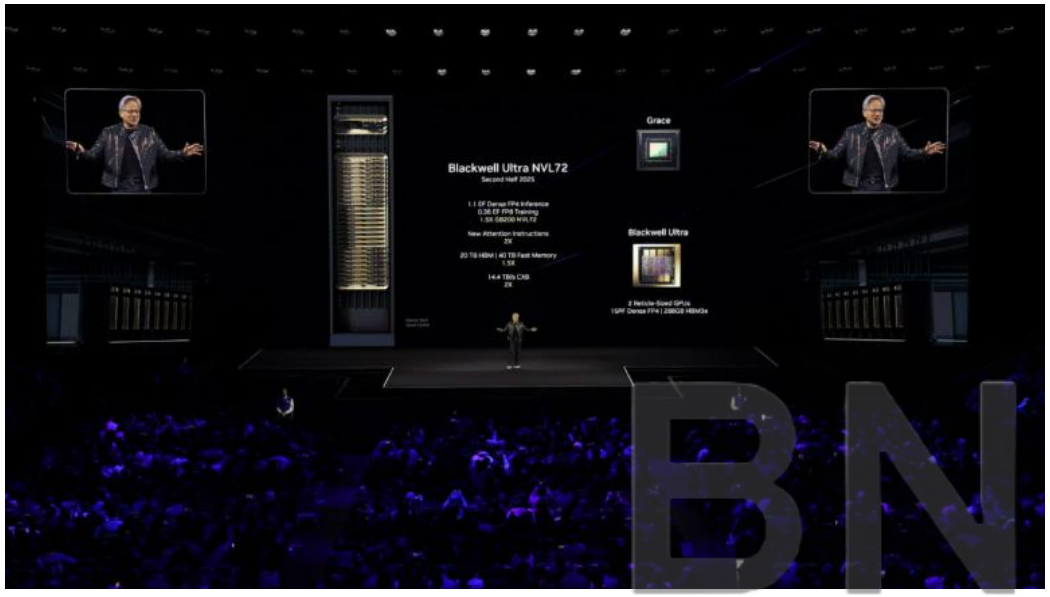




젠슨 황은 청중을 "NVIDIA 본사로" 데려가며 기조연설을 시작했으며, 청중을 포용하는 듯한 로비의 놀라운 비주얼을 보여주었습니다.
그는 25년 전 NVIDIA가 GPU와 함께 시작한 부분에 대해 이야기하면서 시작했습니다. 그는 문제를 해결하고, 계획하고, 조치를 취하는 방법을 추론할 수 있는 에이전트 AI의 출현을 포함하여 지난 10년 동안 AI의 성장에 대해 설명했습니다. 젠슨 황은 "단계별로" 추론할 수 있는 AI의 발전에 대해 설명하고 추론 및 강화 학습에 대한 수요가 AI 컴퓨팅에 대한 수요를 어떻게 주도하고 있는지에 대해 논의했습니다.

AI가 "변곡점"을 거치고 있기 때문에 상위 4개 클라우드 서비스 제공업체의 GPU에 대한 수요가 급증하고 있습니다. 전 세계적으로 데이터센터가 증가할 것으로 예상하고 있으며, 향후 12년간 시장의 확대로 많은 금액이 투자될 것으로 보이고 있어, 젠슨 황은 데이터센터 구축의 가치가 1조 달러에 이를 것으로 예상한다고 말했습니다. 그로 인해 블랙웰 생산이 시작된지 얼마 지나지 않았지만 블랙웰 기반의 AI 인프라의 지속 성장이 예상되고 있습니다.

NVIDIA 블랙웰의 수요 증가

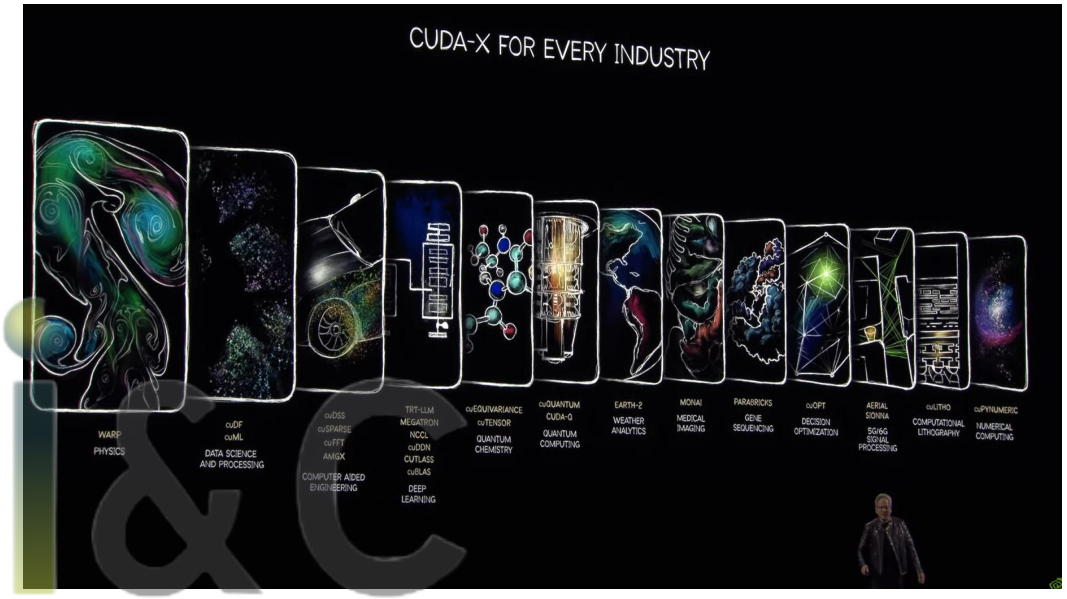


젠슨 황은 데이터센터에 대해 이야기했습니다. 그는 NVIDIA Blackwell 플랫폼이 본격 생산 중이라고 알렸습니다.

그는 Blackwell이 극단적인 스케일 업을 지원하는 방법을 설명했습니다. 젠슨 황은 "우리가 이 일을 하고 싶었던 이유는 극단적인 문제를 해결하기 위함"이라며 "이를 추론이라고 부른다"고 말했습니다.

젠슨 황은 추론이 토큰 생성이며, 이는 비즈니스에 매우 중요하다고 설명했습니다. 이러한 토큰을 생성하는 AI 공장은 극도의 효율성과 성능으로 구축되어야 하며, 토큰에 대한 수요는 점점 더 복잡해지는 문제를 생각하고 해결할 수 있는 최신 세대의 추론 모델과 함께 계속 증가할 것입니다.

모든 산업에 적용될 수 있는 CUDA-X



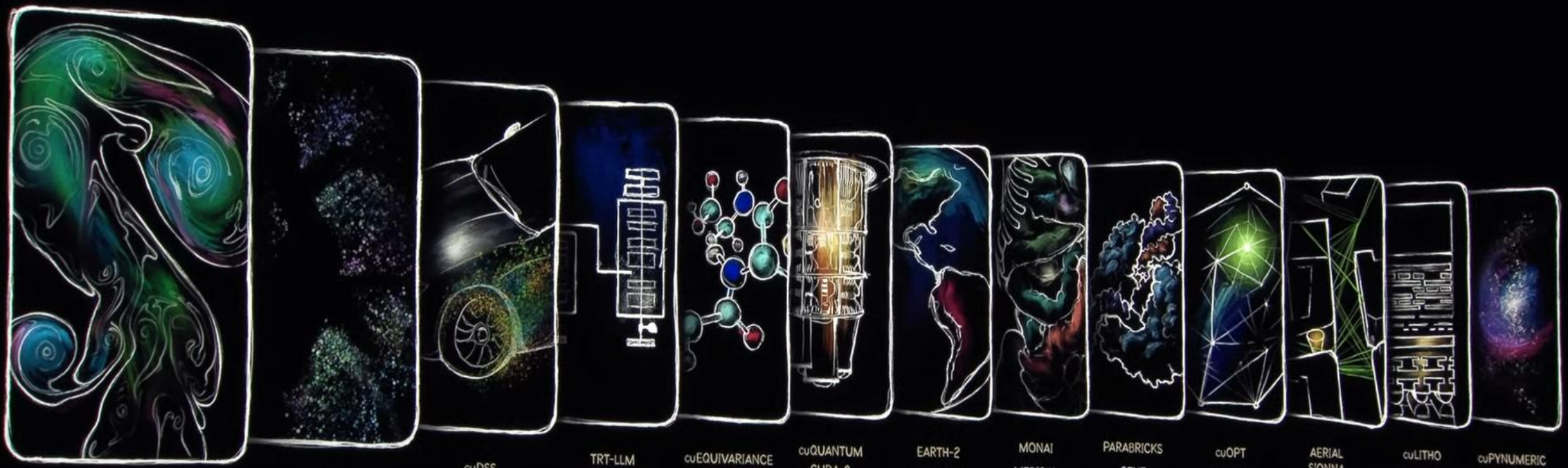
NVIDIA CUDA-X GPU 가속 라이브러리와 마이크로서비스가 이제 모든 산업에 서비스를 제공하고 있다고 설명했습니다.

젠슨 황은 미래에는 모든 기업이 두 개의 공장을 갖게 될 것이라고 말했는데, 하나는 그들이 만드는 것을 위한 것이고 다른 하나는 AI를 위한 공장이라고 했습니다.

그는 다양한 노력에서 NVIDIA의 역할을 샘플링하면서 NVIDIA가 cuOpt 의사 결정 최적화 플랫폼을 오픈 소스화할 것이라고 발표했습니다.

그는 CUDA의 설치 기반이 이제 "어디에나" 있으며, 우리는 가속 컴퓨팅의 티핑 포인트에 도달했으며, CUDA가 이를 가능하게 했습니다." 라고 합니다.

CUDA-X FOR EVERY INDUSTRY



WARP
PHYSICS

cuDF
cuML
DATA SCIENCE AND PROCESSING

cuDSS
cuSPARSE
cuFFT
AMGX
COMPUTER AIDED ENGINEERING

TRT-LLM
MEGATRON
NCCL
cuDDN
CUTLASS
cuBLAS
DEEP LEARNING

cuEQUIVARIANCE
cuTENSOR
QUANTUM CHEMISTRY

cuQUANTUM
CUDA-Q
QUANTUM COMPUTING

EARTH-2
WEATHER ANALYTICS

MONAI
MEDICAL IMAGING

PARABRICKS
GENE SEQUENCING

cuOPT
DECISION OPTIMIZATION

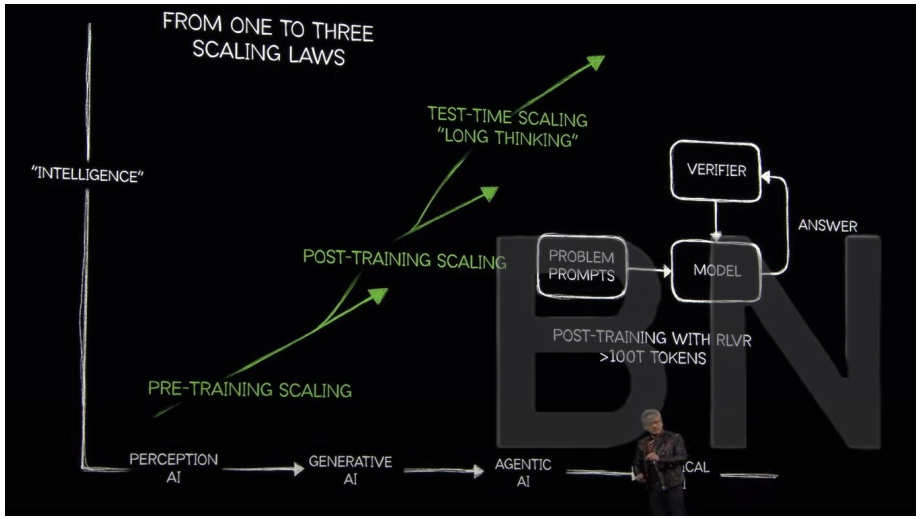
AERIAL
SIONNA
5G/6G SIGNAL PROCESSING

cuLITHO
COMPUTATIONAL LITHOGRAPHY

cuPYNUMERIC
NUMERICAL COMPUTING

② 대규모 AI Factory로의 전환

AI 팩토리는 데이터 센터를 재정의하고 AI의 다음 시대를 가능하게 합니다.



AI는 AI 팩토리가 주도하는 새로운 산업 혁명을 촉진하고 있습니다. 기존 데이터센터와 달리 AI 공장은 데이터를 저장하고 처리하는 것 이상의 역할을 하며, 대규모로 인텔리전스를 제조하여 원시 데이터를 실시간 인사이트로 변환합니다.

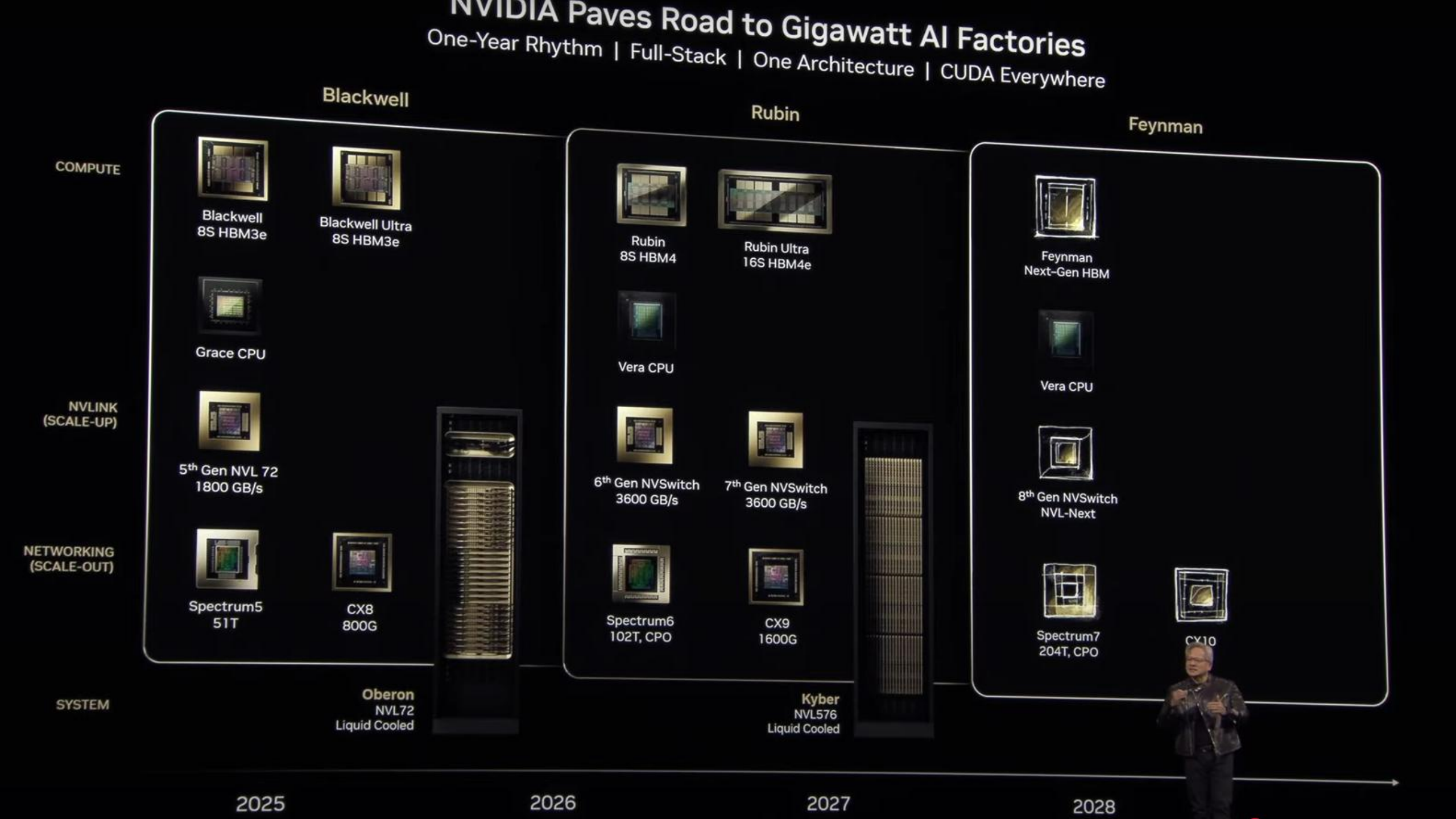
이는 전 세계 기업과 국가의 경우 가치 창출 시간이 획기적으로 단축됨을 의미하며, AI를 장기 투자에서 경쟁 우위의 즉각적인 동인으로 전환한다는 것을 의미합니다. 지금 바로 특수 제작된 AI 공장에 투자하는 기업은 내일의 혁신, 효율성 및 시장 차별화를 주도할 것입니다.

기존 데이터센터는 일반적으로 다양한 워크로드를 처리하고 범용 컴퓨팅을 위해 구축되었지만, AI 공장은 AI에서 가치를 창출하는 데 최적화되어 있습니다. 데이터 수집부터 교육, 미세 조정, 그리고 가장 중요하게는 대용량 추론에 이르기까지 전체 AI 라이프사이클을 오케스트레이션합니다.

전통적인 데이터센터가 곧 사라지지는 않겠지만, AI 팩토리로 진화할지 아니면 연결할지는 엔터프라이즈 비즈니스 모델에 달려 있습니다. 기업이 어떤 방식으로 적응하든, NVIDIA로 구동되는 AI 공장은 이미 대규모로 인텔리전스를 제조하여 AI의 구축, 개선, 배포 방식을 변화시키고 있습니다.

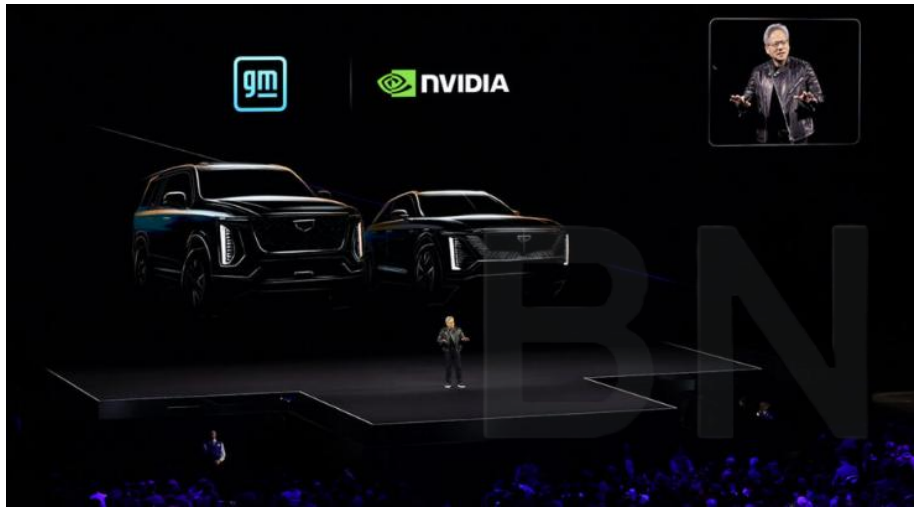
NVIDIA Paves Road to Gigawatt AI Factories

One-Year Rhythm | Full-Stack | One Architecture | CUDA Everywhere



③ General Motors와 NVIDIA, AI를 위해 협업

GM과 협력, 차세대 자율주행 차량의 제조 공정에 옴니버스 기술과 NVIDIA Drive AGX 플랫폼 적용



AI는 이제 로봇 공학과 자율 주행 자동차, 공장 및 무선 네트워크에서 "전 세계로" 진출하고 있습니다. 젠슨황은 AI가 가장 먼저 진출한 산업 중 하나는 자율 주행 자동차였으며, 거의 모든 자율주행차 회사가 사용하는 기술을 NVIDIA 가 구축한다"고 하며, 미국 최대 자동차 제조업체인 GM이 NVIDIA AI, 시뮬레이션, 가속 컴퓨팅을 채택해 차세대 자동차, 공장, 로봇을 개발하고 있다고 발표했습니다.

또한 NVIDIA의 자동차 하드웨어 및 소프트웨어 안전 솔루션 라인업과 AV 안전에 대한 최첨단 AI 연구를 결합한 포괄적인 안전 시스템인 NVIDIA Halos를 발표했습니다.

GM은 시뮬레이션과 검증을 포함한 다양한 영역에서 AI 모델을 훈련하기 위해 NVIDIA GPU 플랫폼에 투자해 왔고, 양사의 협력은 이제 자동차 플랜트 설계 및 운영의 혁신으로 확장되고 있습니다. GM은 NVIDIA Omniverse 플랫폼을 사용해 조립 라인의 디지털 트윈을 생성하고, 이를 통해 가상 테스트와 생산 시뮬레이션을 통해 다운타임을 줄이며, 제조 안전성과 효율성을 높이기 위해 정밀 용접과 함께 자재 취급 및 운송과 같은 작업에 이미 사용 중인 로봇 플랫폼을 교육까지 함께 합니다.

④ 6G용 AI 네이티브 무선 네트워크 개발 및 협력

T-Mobile, Cisco, MITRE 등과 함께 6G용 AI 네이티브 무선 네트워크를 개발 및 구축 계획

NVIDIA는 6G용 AI 네이티브 무선 네트워크 하드웨어, 소프트웨어 및 아키텍처의 연구 및 개발에 대해 업계 리더인 T-Mobile, MITRE, Cisco, ODC, Cerberus Capital Management의 포트폴리오 회사인 ODC, Booz Allen Hamilton과 파트너십을 발표했습니다.

차세대 무선 네트워크는 수천억 대의 전화, 센서, 카메라, 로봇 및 자율 주행 차량을 원활하게 연결하기 위해 근본적으로 AI와 통합되어야 합니다. AI 네이티브 무선 네트워크는 수십억 명의 사용자에게 향상된 서비스를 제공하고 스펙트럼 효율성(주어진 대역폭을 통해 데이터를 전송할 수 있는 속도)의 새로운 표준을 설정할 것입니다.

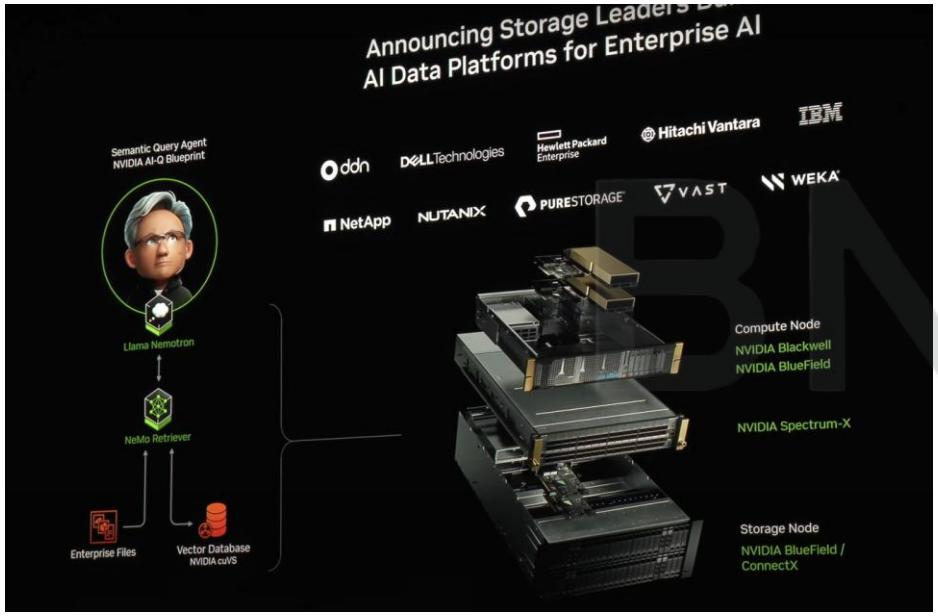
AI 혁신을 주도하기 위해 NVIDIA는 통신 업체 및 연구 리더와 협력하여 NVIDIA 가속 컴퓨팅 플랫폼에서 소프트웨어 정의 무선 액세스 네트워크(RAN)를 제공하는 NVIDIA AI Aerial 플랫폼을 기반으로 AI 네이티브 무선 네트워크 스택을 개발하고 있습니다.

전 세계의 개발자들은 AI 네이티브 6G 무선 네트워크의 선구자로 AI-RAN을 구축하고 있습니다. AI-RAN은 AI와 RAN 워크로드를 하나의 플랫폼에 통합하고 AI를 무선 신호 처리에 내장하는 기술입니다. 향상된 스펙트럼 효율성을 제공하고 운영 복잡성과 비용을 낮추기 위해 AI는 네트워크 스택의 소프트웨어에 완전히 내장되고 네트워크 및 AI 워크로드를 모두 실행할 수 있는 통합 가속 인프라에서 호스팅됩니다. 또한 솔루션의 핵심은 엔드투엔드 보안과 개방형 아키텍처를 통해 신속한 혁신을 촉진하는 것입니다.

T-모바일과 엔비디아는 AI 네이티브 6G 네트워크 기능에 대한 추가적인 연구 기반 개념을 제공하는 것을 목표로 지난해 발표한 AI-RAN 이노베이션 센터 협업을 확대하고, 새로운 업계 협력자들과 협력할 예정입니다.

⑤ 엔터프라이즈용 AI 데이터 플랫폼 발표

NVIDIA와 스토리지 업계 리더들, 엔터프라이즈용 AI 데이터 플랫폼 발표



NVIDIA는 까다로운 AI 추론 워크로드를 위한 새로운 차원의 AI 인프라, 즉 NVIDIA 가속 컴퓨팅, 네트워킹 및 소프트웨어로 구동되는 AI 쿼리 에이전트가 있는 엔터프라이즈 스토리지 플랫폼을 구축하는 데 사용하는 맞춤형 레퍼런스 디자인인 NVIDIA AI 데이터 플랫폼을 발표했습니다.

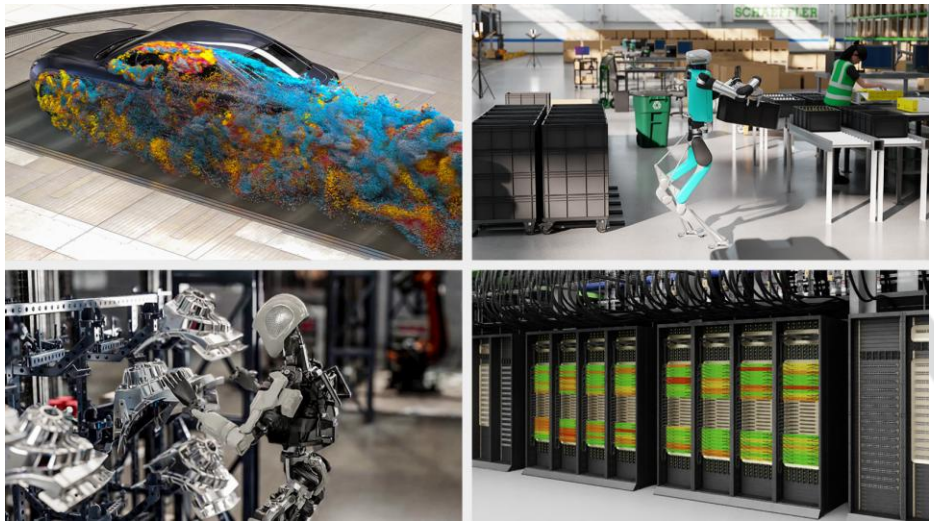
이러한 데이터 플랫폼은 에이전트 AI 워크플로우의 성능과 정확성을 개선하기 위해 엔터프라이즈 스토리지 리더가 구축했습니다.

이 플랫폼은 NVIDIA Blackwell GPU, BlueField-3® DPU, Spectrum-X™ 네트워킹 및 NVIDIA AI Enterprise 소프트웨어를 갖추고 있습니다. 이 플랫폼은 가속 컴퓨팅의 힘을 AI 데이터 처리 및 네트워크 연결에 제공합니다. 이를 통해 워크플로우를 최적화할 수 있으므로 AI 에이전트는방대한 비즈니스 지식을 보다 효율적이고 짧은 지연 시간으로 분석하고 조치를 취할 수 있습니다.

NVIDIA CEO인 젠슨 황(Jensen Huang)은 "데이터는 AI 시대에 산업을 움직이는 원자재입니다"라고 말했습니다. "우리는 세계 최고의 스토리지 리더들과 함께 기업이 하이브리드 데이터센터 전반에 걸쳐 에이전트 AI를 배포하고 확장하는 데 필요한 새로운 차원의 엔터프라이즈 인프라를 구축하고 있다.“ 라고 발표했습니다.

⑥ NVIDIA Omniverse Physical AI 운영 체제의 활용

NVIDIA Omniverse Physical AI 운영 체제를 더 다양한 산업 및 파트너로 확장하다



NVIDIA Cosmos™ World Foundation Models 에 연결된 새로운 NVIDIA Omniverse Blueprint를 사용하여 물리적 AI 개발을 위한 로봇 지원 시설과 대규모 합성 데이터 생성을 지원할 수 있습니다.

Omniverse는 개발자가 물리적 데이터와 애플리케이션을 통합할 수 있도록 지원하는 OpenUSD 프레임워크를 기반으로 구축된 운영 체제입니다. Omniverse를 통해 글로벌 산업 소프트웨어, 데이터, 전문 서비스 리더들은 산업 생태계를 통합하고 전례 없는 속도로 산업을 위한 차세대 AI를 발전시킬 새로운 애플리케이션을 구축하고 있습니다.

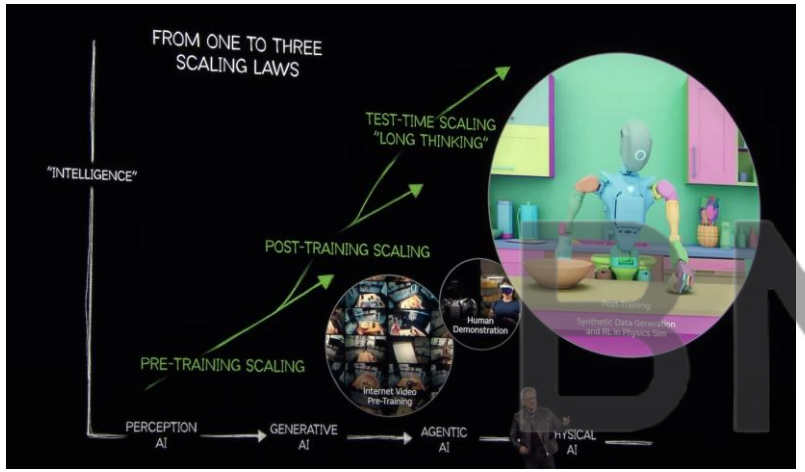
AI 공장 디지털 트윈을 위한 새로운 Omniverse Blueprint를 통해 데이터센터 엔지니어는 AI 공장 레이아웃, 냉각 및 전기를 설계하고 시뮬레이션하여 활용도와 효율성을 극대화할 수 있습니다.

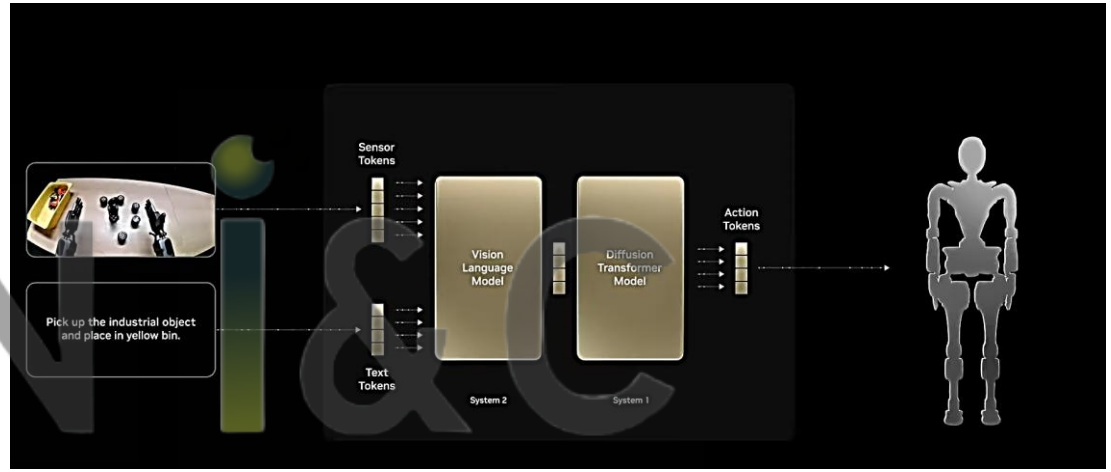
또한 NVIDIA Metropolis 플랫폼을 기반으로 하는 비디오 검색 및 요약을 위한 NVIDIA AI Blueprint를 사용하여 전체 시설의 활동을 모니터링하는 AI 에이전트를 구축할 수 있습니다. 제조업계의 리더들은 블루프린트을 사용하여 Physical AI로 산업 운영을 최적화하고 있습니다.
합성 조작 모션 생성을 위한 NVIDIA Isaac GR00T 블루프린트는 이제 로보틱스 개발자도 사용할 수 있어 Omniverse와 Cosmos에서 대규모 합성 데이터를 생성할 수 있습니다.

클라우드 기반 NVIDIA Omniverse 개발자 도구 및 서비스는 NVIDIA L40S GPU를 탑재한 오라클 클라우드 인프라스트럭처(Oracle Cloud Infrastructure) 컴퓨팅 베어메탈 인스턴스와 새로 발표된 구글 클라우드(Google Cloud)의 NVIDIA RTX PRO 6000 블랙웰 서버 에디션(NVIDIA RTX PRO™ 6000 Blackwell Server Edition)에서 올해 말에 제공될 예정입니다.

⑦ Physical AI와 로보틱스

NVIDIA, Isaac GR00T N1과 로봇 개발 가속화를 위한 시뮬레이션 프레임워크 발표





젠슨 황은 로봇을 다음 10조 달러 산업으로 묘사하면서, 이번 10년 말까지 전 세계적으로 최소 5천만 명의 노동자가 부족해질 것이라고 말했습니다. NVIDIA는 차세대 로보틱스를 교육, 배포, 시뮬레이션 및 테스트하기 위한 완벽한 기술 제품군을 제공합니다.

젠슨 황은 일반화된 휴머노이드 추론 및 기술을 위한 세계 최초의 완전 맞춤형 개방형 파운데이션 모델인 NVIDIA Isaac GR00T N1의 출시를 발표했습니다.

NVIDIA는 새로운 NVIDIA Cosmos 세계 기반 모델의 주요 릴리스를 발표하여 물리적 AI 개발을 위한 개방적이고 완전히 사용자 정의 가능한 추론 모델을 도입하고 생성에 대한 전례 없는 제어를 제공합니다. 젠슨 황은 "옴니버스(Omniverse)를 사용해 코스모스(Cosmos)를 컨디셔닝하고, 코스모스(Cosmos)를 사용해 무한한 환경을 생성하면 우리가 접지하고 제어하면서도 동시에 시스템적으로 무한한 데이터를 생성할 수 있다"고 말했습니다.

그는 또한 구글 딥마인드(Google DeepMind) 및 디즈니 리서치(Disney Research)와 함께 개발 중인 로보틱스 시뮬레이션을 위한 뉴턴(Newton) 오픈소스 물리 엔진을 소개한 후, 바닥의 해치에서 나온 작은 로봇 "블루(Blue)"가 무대에 올라 황에게 삐삐 소리를 내며 야유를 보냈습니다.

NVIDIA GTC25