향후 10년의 패러다임

글로벌 AI섹터에 투자하자!

미래에셋증권 반포역 WM 장의성 지점장

2024.4

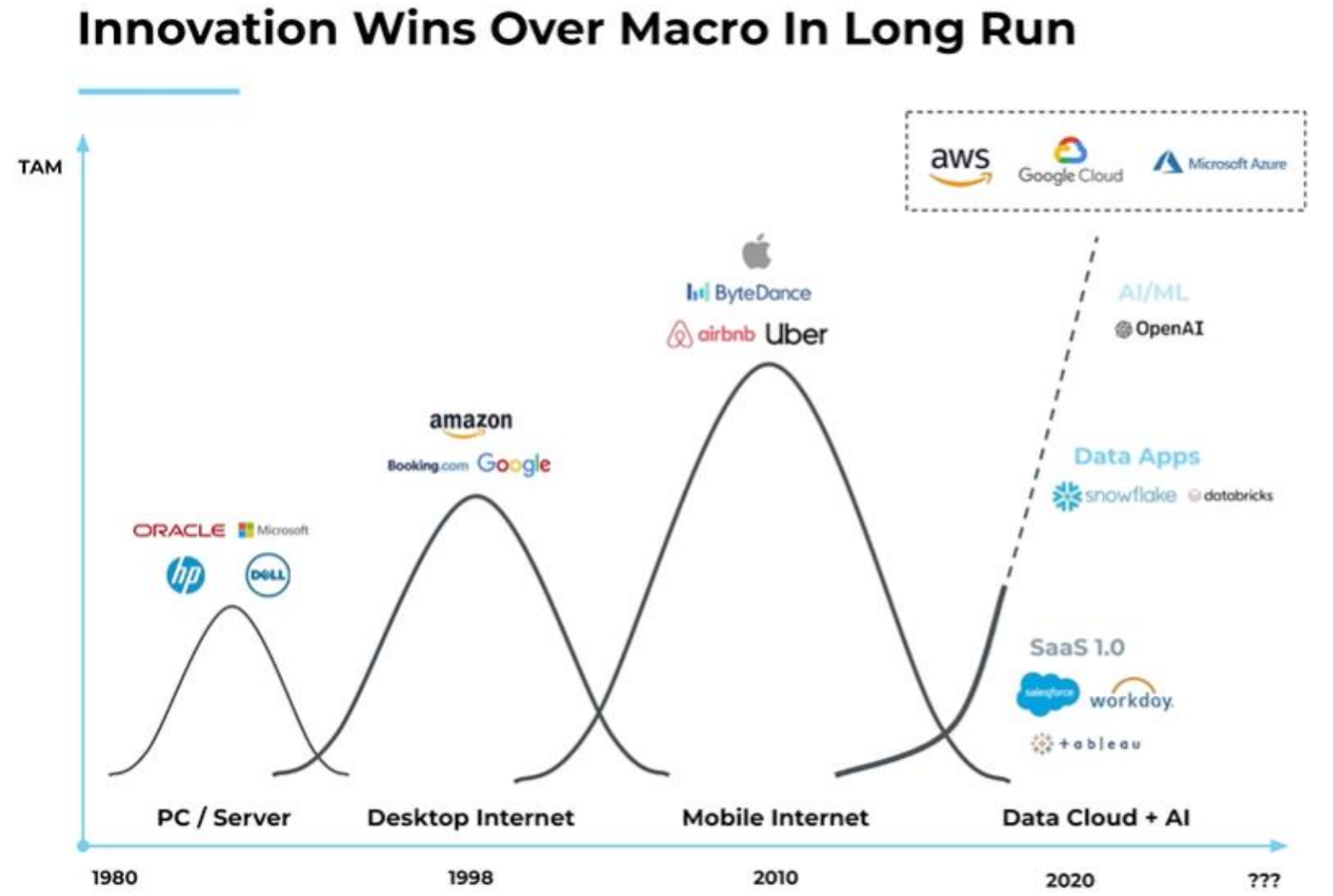
MIRAE ASSET 미래에셋증권

오늘 드릴 말씀 순서

- 왜 지금 AI인가? – 10년의 패러다임
- AI의 3단계 – AI시대는 어떻게 진행되는가?
- GTC 2024 – AI시대의 아이콘
- AI가 바꿀 우리의 미래 – AI투자를 해야 하는 이유

왜 지금 AI인가? – 10년의 패러다임

10년의 패러다임 - AI의 시대



생산성

AI의 생산성 기여 사례

<출처: nevian infotech>

왜 지금 AI인가?

1997

2011

2016

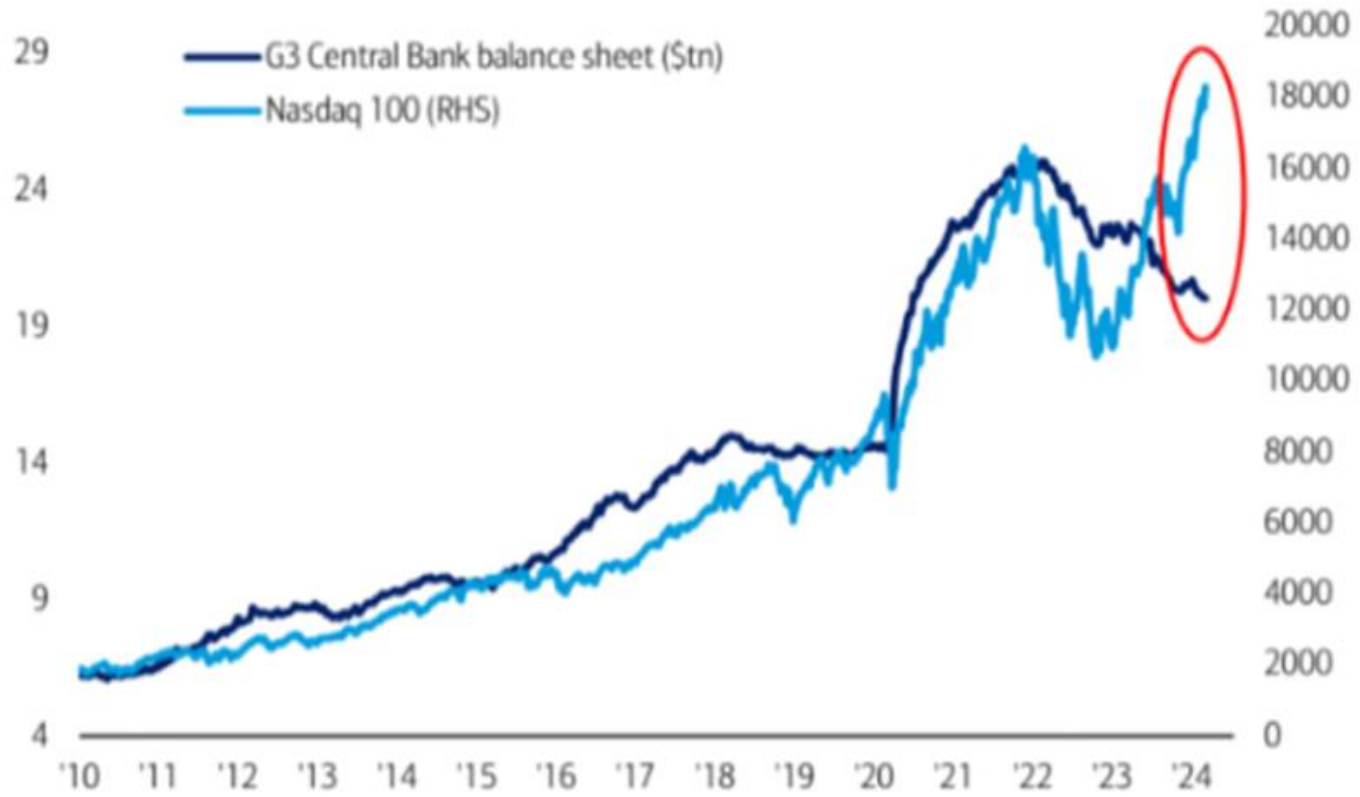
2022

ChatGPT

<출처: google>

시장 상승에서 AI의 지분

Chart 5: Enduring "bromance" of the Fed & Wall St
G3 Central Bank balance sheet ($tn) vs Nasdaq 100



Source: BofA Global Investment Strategy, Bloomberg. G3 central banks = Fed, ECB, BoJ
BofA GLOBAL RESEARCH

<출처: Bofa>

AI의 3단계 – AI시대는 어떻게 진행되는가?

AI 시대의 3단계

<인프라의 시대>

<서비스의 시대>

<로봇의 시대>

<출처 : Chat GPT>

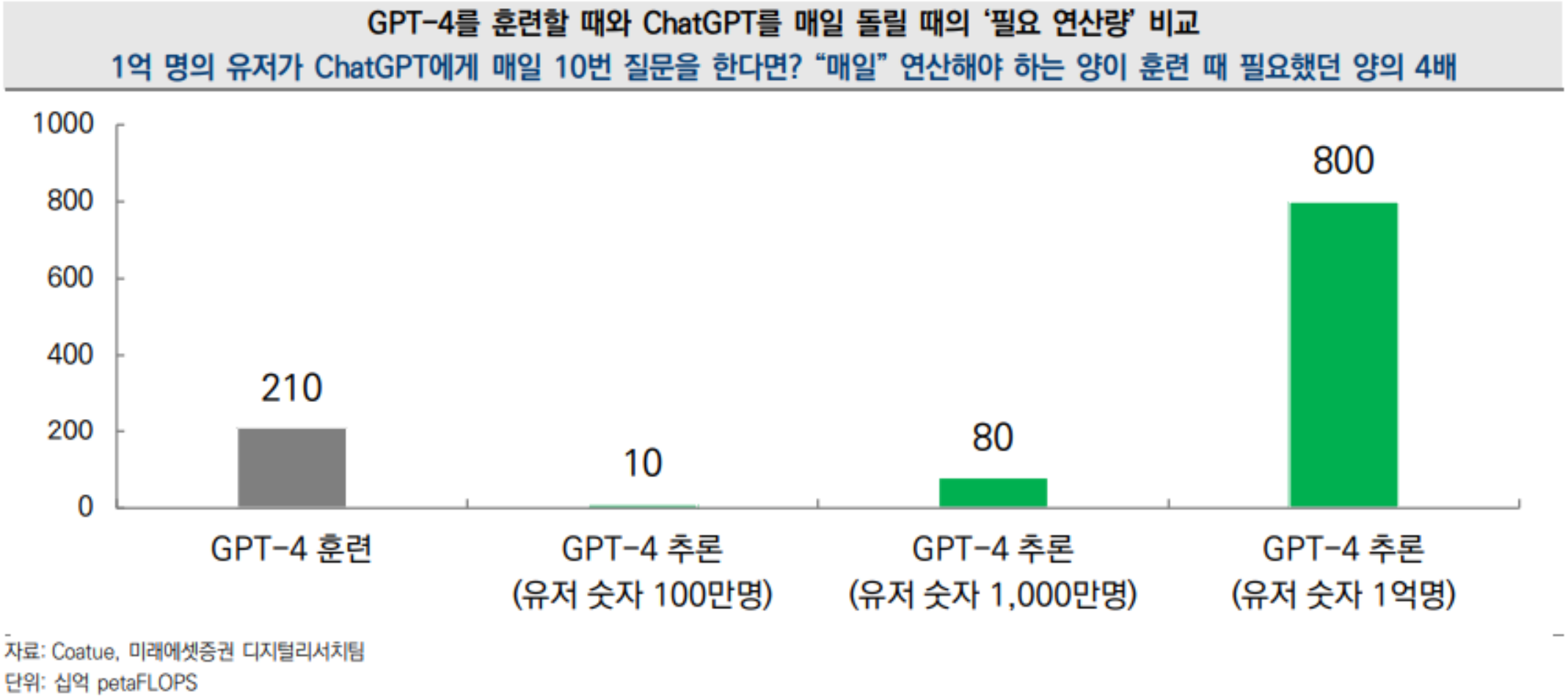
반도체 시장이 얼마나 커질 것인가?

샘 알트먼, 7조 달러 오보지만 AI인프라에 천문학적인 투자 필요!

연간 글로벌 반도체 투자 약 1천억 달러

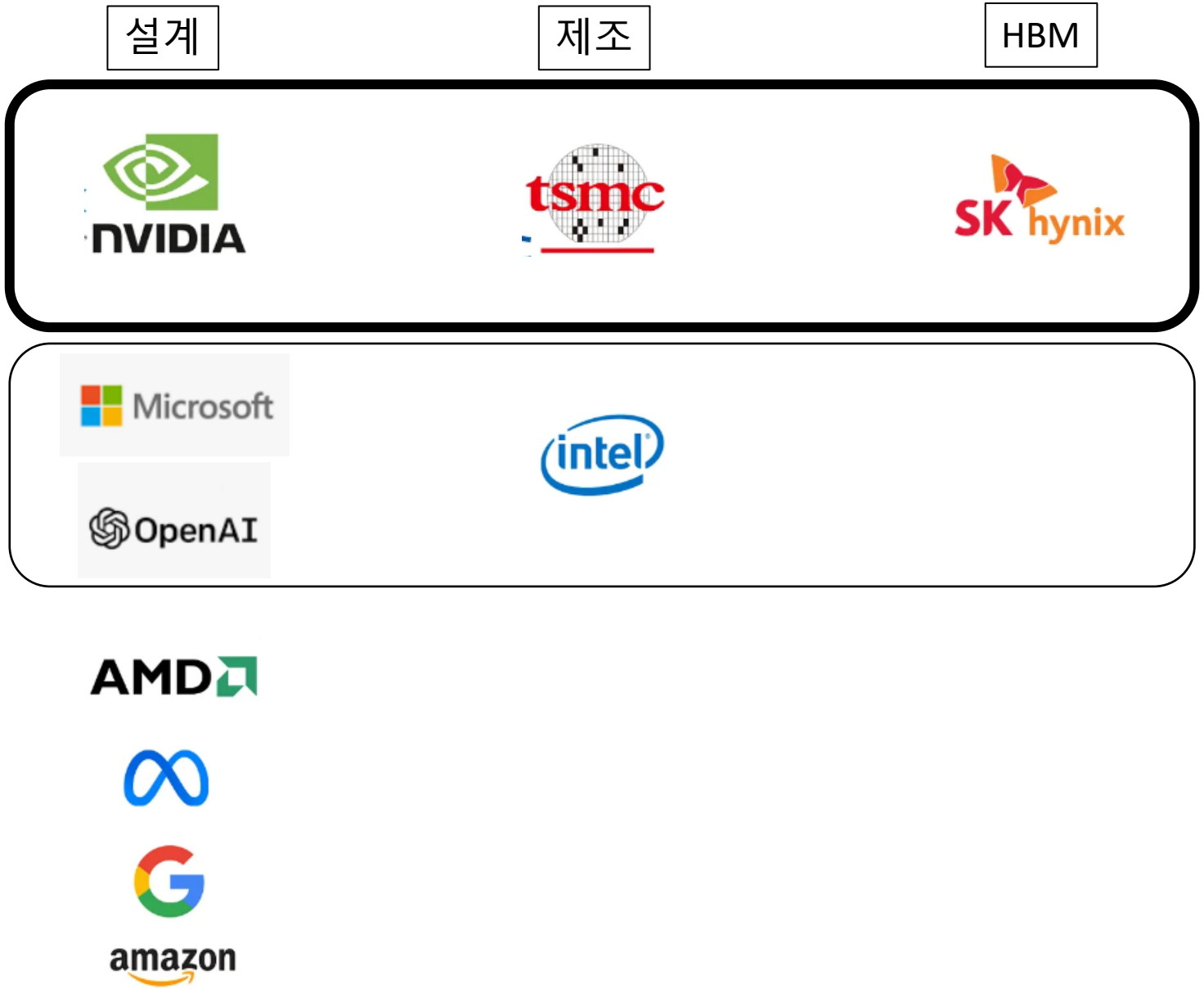
연간 글로벌 반도체 매출 약 5천억 달러

추론시장의 크기

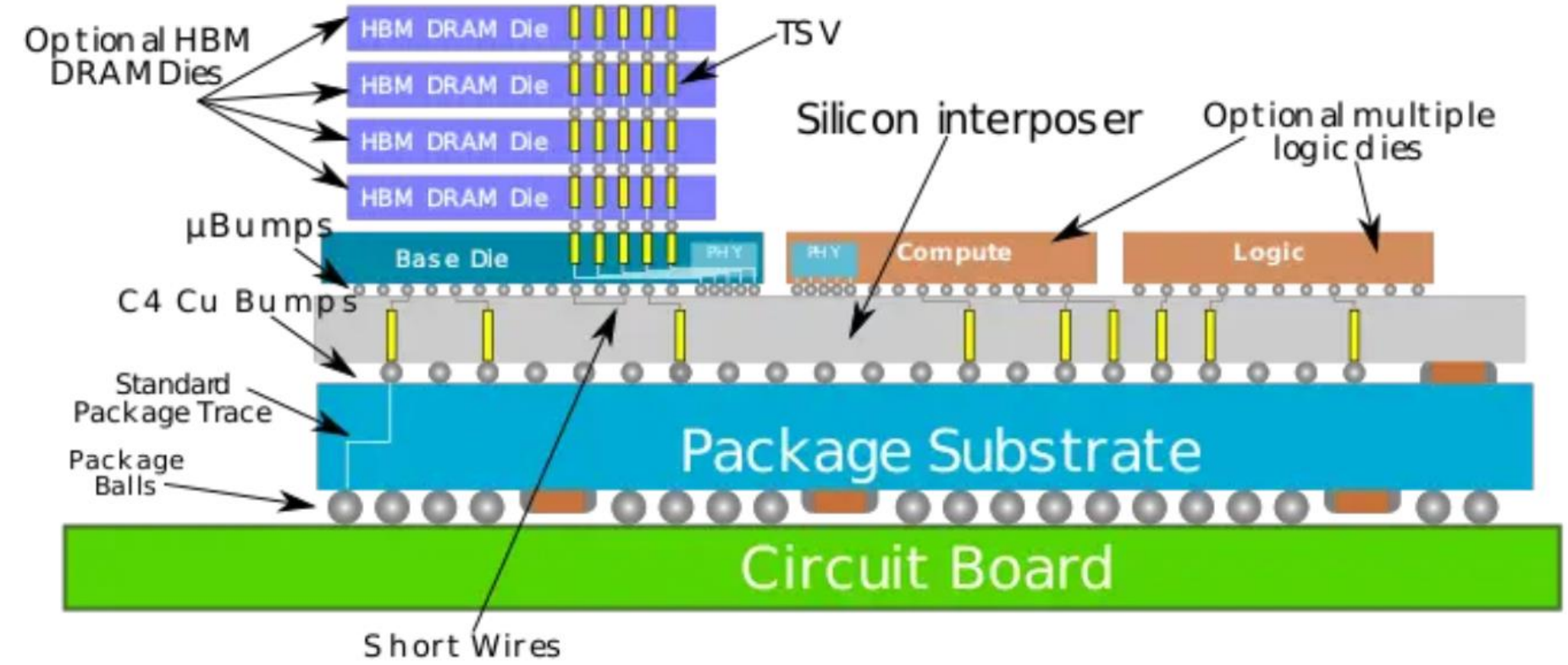


<출처: 미래에셋증권>

AI 반도체 합종연횡!


설계
제조
HBM
NVIDIA
tsmc
SK hynix
Microsoft
intel
OpenAI
AMD
amazon

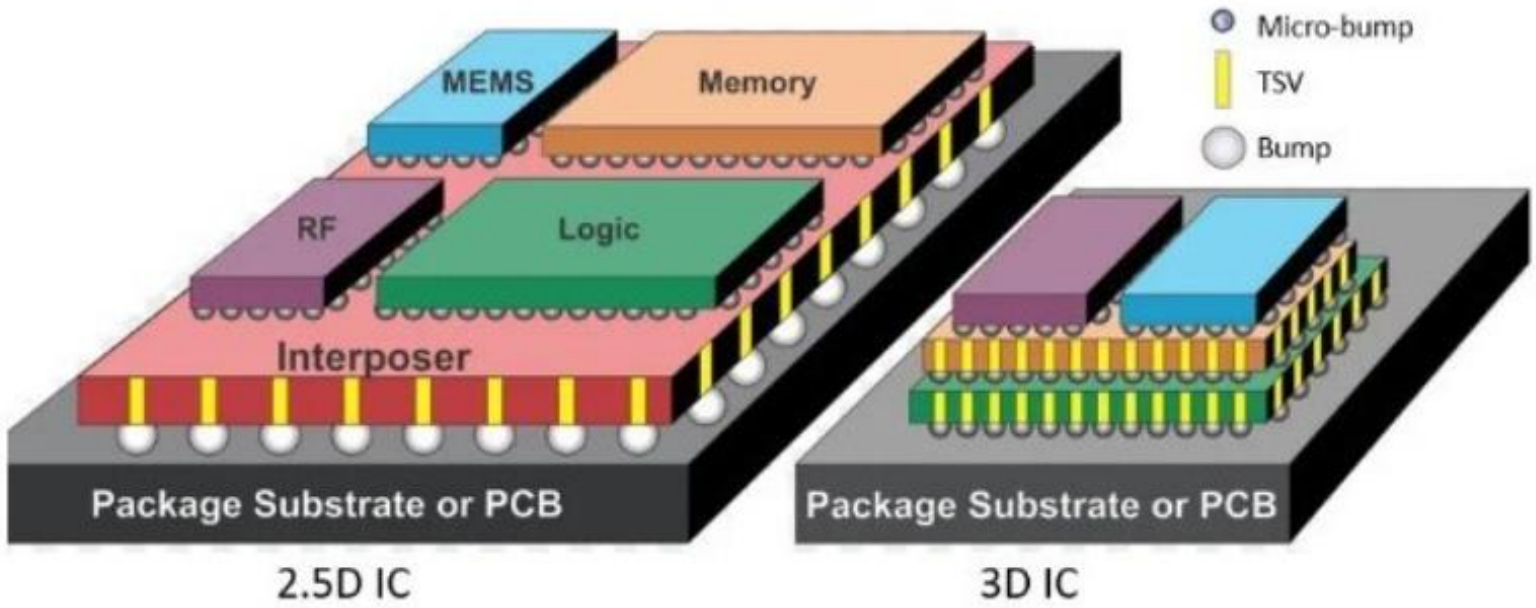
COWOS



Chip-on-Wafer-on-Substrate (CoWoS) TSMC WikiChip

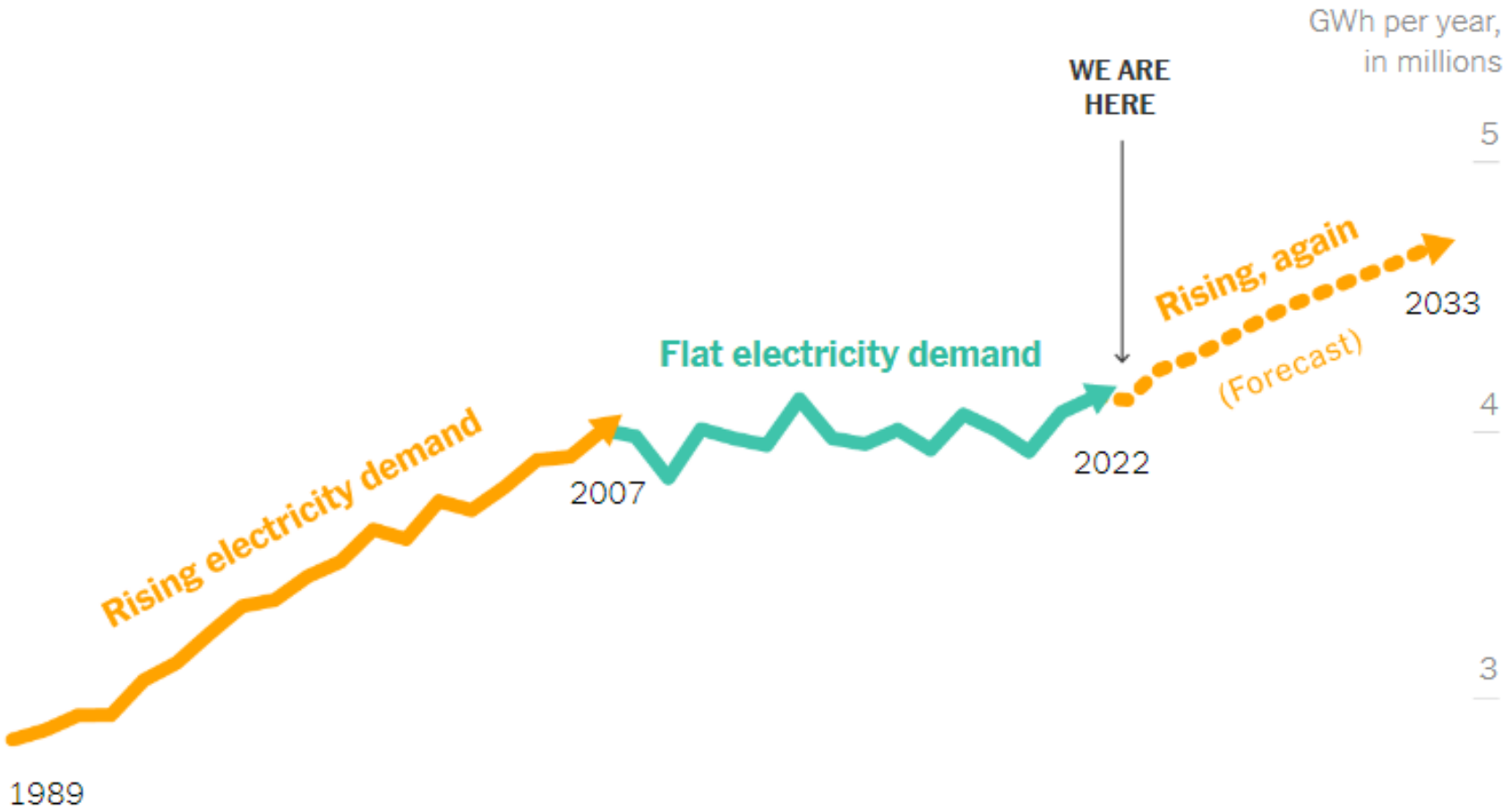
<출처 : TSMC>

결국은 3D



<출처 : Semiconductor engineering>

전기도 부족 – 미국의 전력 수요



<출처 : NY Times>

Microsoft 365

Copilot for Microsoft 365 available as an add-on[2]

Microsoft 365 E3 (no Teams)

$33.75 user/month

(Annual commitment)

Contact Sales

Try free for one month >

See trial terms[1]

Learn more >

- Microsoft 365 apps for desktop and mobile
- Windows for Enterprise
- 1 TB of cloud storage
- Core security and identity management capabilities
- Copilot for Microsoft 365, available as an add-on[2]

Copilot for Microsoft 365 available as an add-on[2]

Microsoft 365 E5 (no Teams)

$54.75 user/month

(Annual commitment)

Contact Sales

Learn more >

Everything in Microsoft 365 E3 (no Teams), plus:

- Advanced security and compliance capabilities
- Scalable business analytics with Power BI
- Copilot for Microsoft 365, available as

<출처: microsoft>

<출처: microsoft>

AI비서를 장악하는 기업은?

<출처: Chat GPT>

AI 로봇의 시대

<출처: google>

GTC 2024 – AI시대의 아이콘

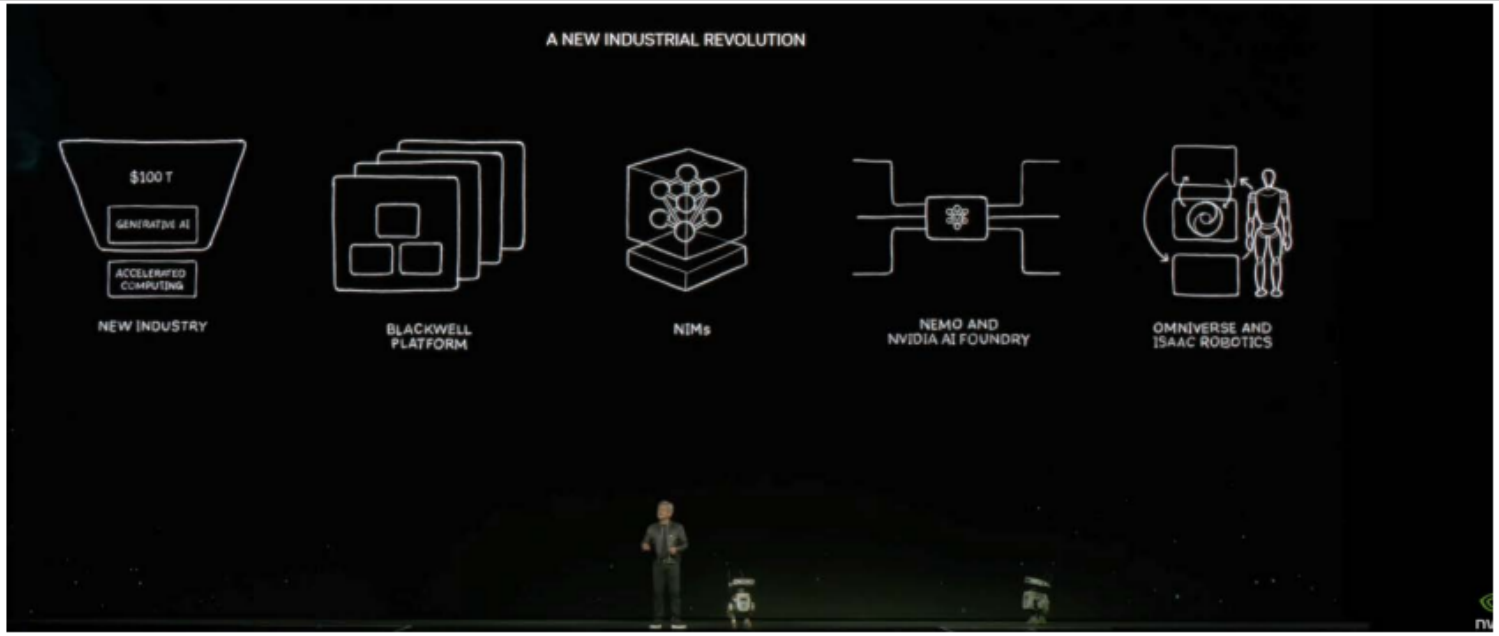
AI의 모든 것, AI 파운드리

자료: Nvidia, 미래에셋증권 디지털리서치팀

<출처: 엔비디아, 미래에셋증권>

GTC 2024 핵심포인트 3가지

GTC 2024 키노트에서 젠슨 황 CEO가 강조한 핵심 개념들
이 중에서도 “Blackwell, NIM, 옴니버스”라는 3가지 포인트로 구분 지을 수 있음



자료: Nvidia

디지털리서치팀이 생각하는 GTC 2024의 핵심 포인트 세 가지와 각각의 에센스

핵심 포인트	이번에 얻을 수 있었던 인사이트
새로운 아키텍처	스케일링 법칙은 여전히 유효! 2세대 트랜스포머 엔진으로 추론 능력 대폭 강화
NIM	기업들을 대상으로 AI 팩토리가 되기 위한 진정한 서비스의 출현
옴니버스	AI의 끝은 결국 메타버스! 생성 AI 기반 시뮬레이션으로 디지털 트윈을 탄생

자료: 미래에셋증권 디지털리서치팀

<출처: 엔비디아, 미래에셋증권>

블랙웰 – 괴물칩의 등장

Hopper 대비 2배 커진 Blackwell의 다이 사이즈
커진 만큼 1,280억개의 트랜지스터가 더 들어가고 HBM 공간도 훨씬 커짐

자료: Nvidia

GTC 2024 키노트에서 등장한 새로운 아키텍처 "Blackwell"의 주요 스펙 사항
Hopper 대비 학습은 4배, 추론은 30배, 에너지 효율은 25배 향상

자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

블랙웰에 줄 선 고객들

<출처: 엔비디아>

Hopper VS Blackwell

90일 동안 GPT-4를 훈련한다고 했을 때 필요한 컴퓨팅 용량 비교
(좌)Hopper 기반 데이터센터 vs (우)Blackwell 기반 데이터센터

자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

GB200

GB200 NVL 72의 구성품

GB200 슈퍼칩 서버 2개가 모여 컴퓨팅 트레이 하나가 되고, 이런 트레이가 랙 안에 18개 있음

GB200 SUPERCHIP

40 PETAFLOPS FP4 AI INFERENCE
20 PETAFLOPS FP8 AI TRAINING
864GB FAST MEMORY

GB200 SUPERCHIP COMPUTE TRAY

2x GB200
80 PETAFLOPS FP4 AI INFERENCE
40 PETAFLOPS FP8 AI TRAINING
1728 GB FAST MEMORY
1U Liquid Cooled
18 Per Rack

NVLINK SWITCH TRAY

2x NVLINK SWITCH CHIP
14.4 TB/s Total Bandwidth
SHARPv4 FP64/32/16/8
1U Liquid Cooled
9 Per Rack

자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

GB200 NVL 72

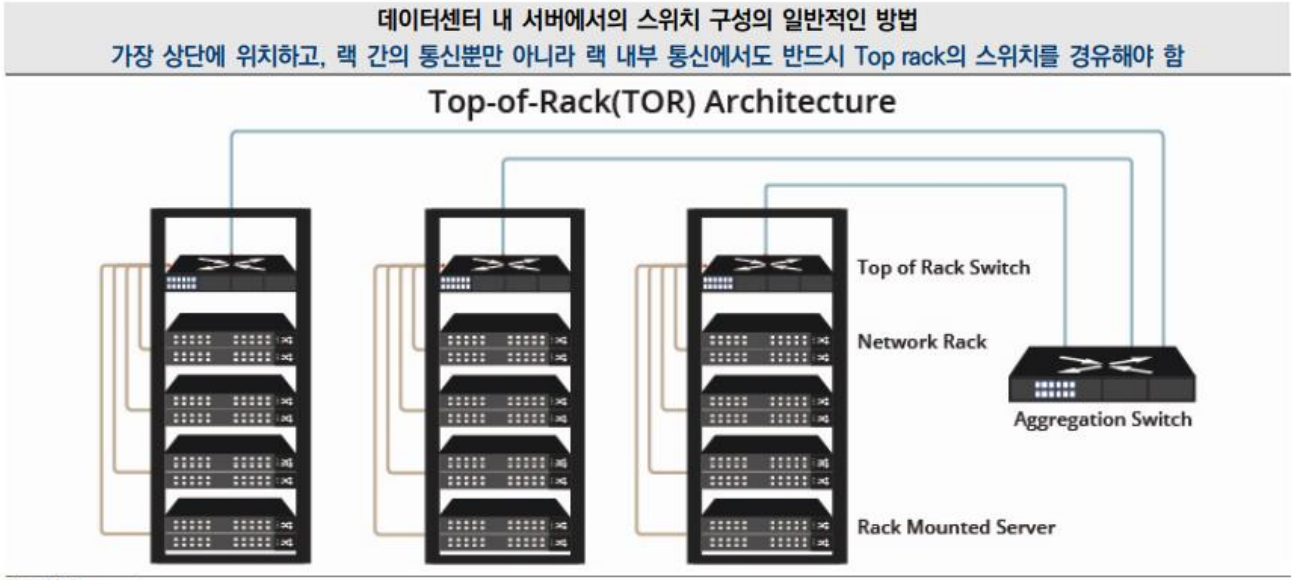
NVLink 스위치를 포함한 GB200 NVL 72의 스펙과 실제 모습을 촬영한 사진
네트워킹 트레이는 9개 존재, 네트워킹 트레이당 2개의 NV스위치(ASIC)가 들어가 총 18개의 스위치 칩이 탑재
오른쪽 사진은 엔비디아의 대표적인 ODM 업체인 SuperMicro가 만든 GB200 NVL 72 제품

자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

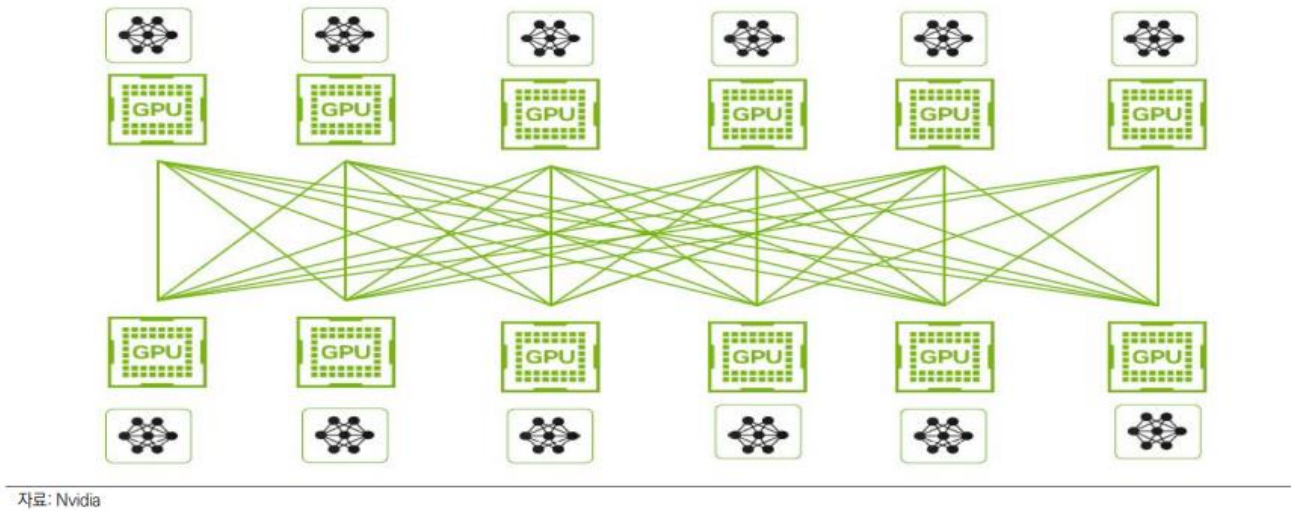
NVLink

데이터센터 내 서버에서의 스위치 구성의 일반적인 방법
가장 상단에 위치하고, 랙 간의 통신뿐만 아니라 랙 내부 통신에서도 반드시 Top rack의 스위치를 경유해야 함



자료: FS Community

GB200 NVL 72에서 각 GPU들이 통신하는 방법
랙 내부의 GPU들은 중간 경유 과정 없이 각자가 서로 "직접" 연결

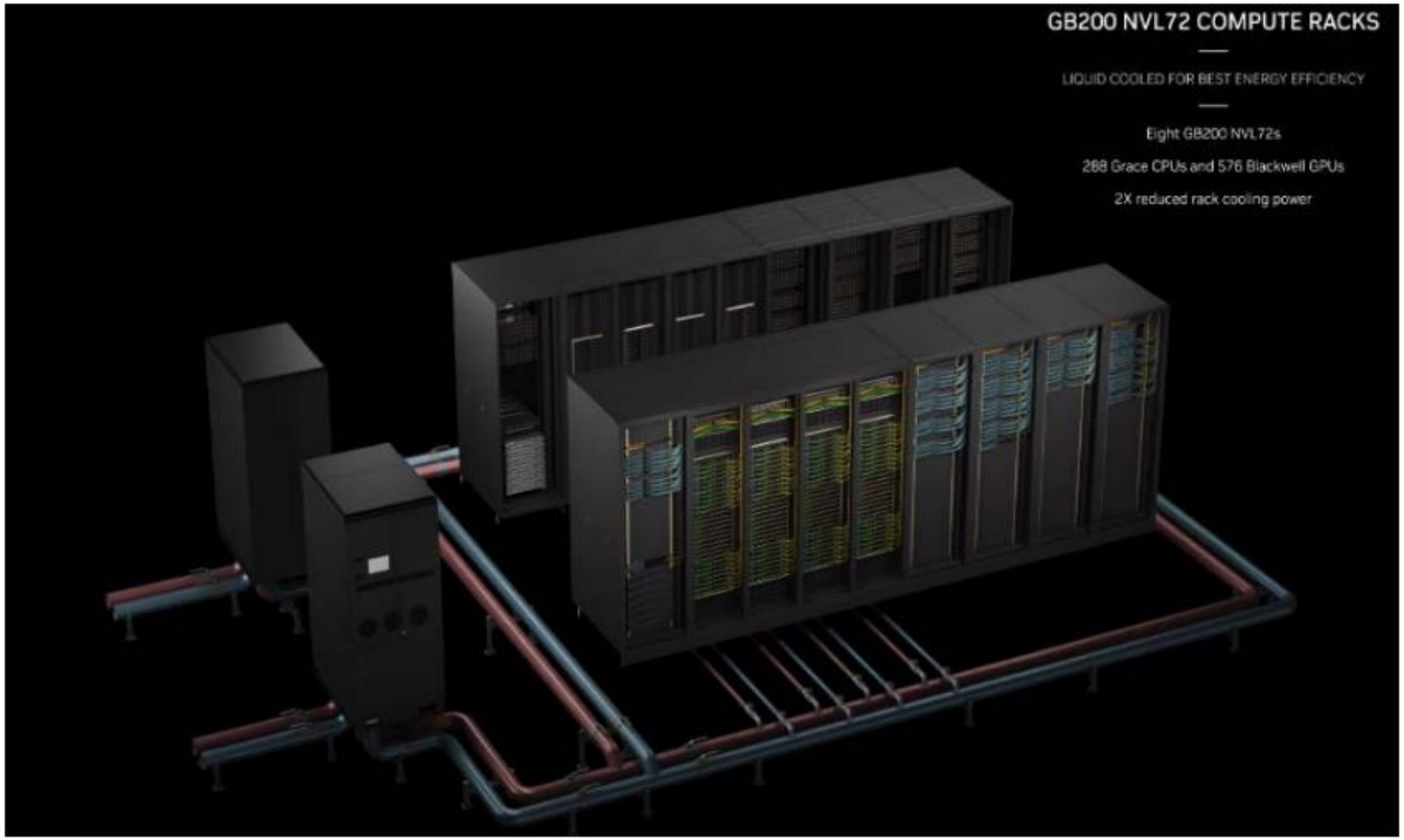


자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

액체냉각

GB200 NVL 72 기반 총 8개의 랙 구성 및
GB200 NVL 72은 공랭이 아니라 액체 냉각 시스템

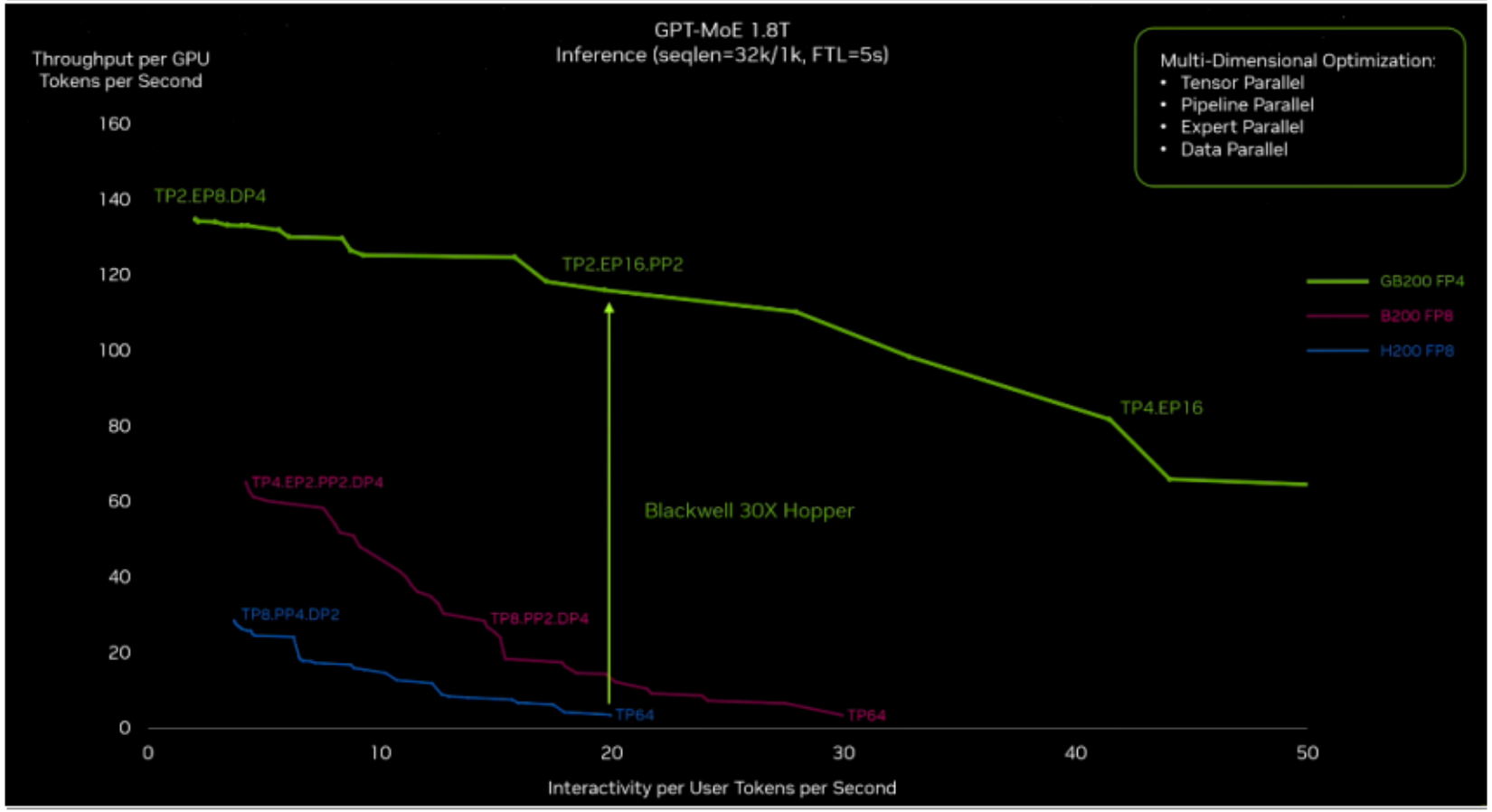


자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

추론성능 향상

엔비디아의 각 GPU 아키텍처가 1.8조 모델크기의 GPT-MoE 처리할 때의 추론 성능
x축은 사용자와의 상호작용 측면에서 유저 기준 초당 토큰 생성 속도
y축은 데이터센터 처리량 측면에서 GPU 기준 초당 토큰 생성 속도

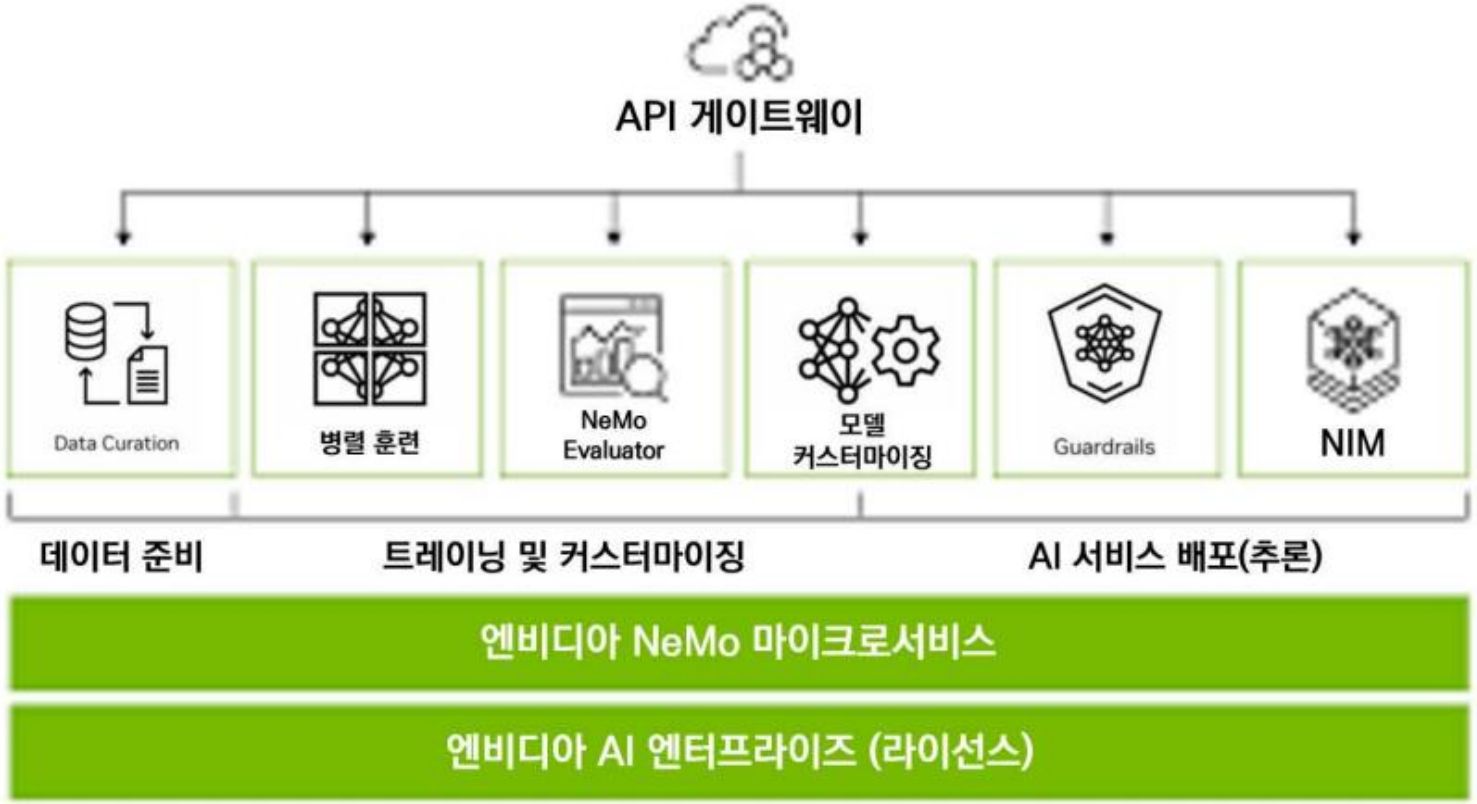


자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

엔비디아 AI 엔터프라이즈 라이선스

엔비디아 AI 엔터프라이즈라는 라이선스가 있다면 NeMo와 NIM을 API로 사용 가능
'NeMo로 AI 모델 훈련 끝내고 NIM으로 다듬어서 출시하고!'

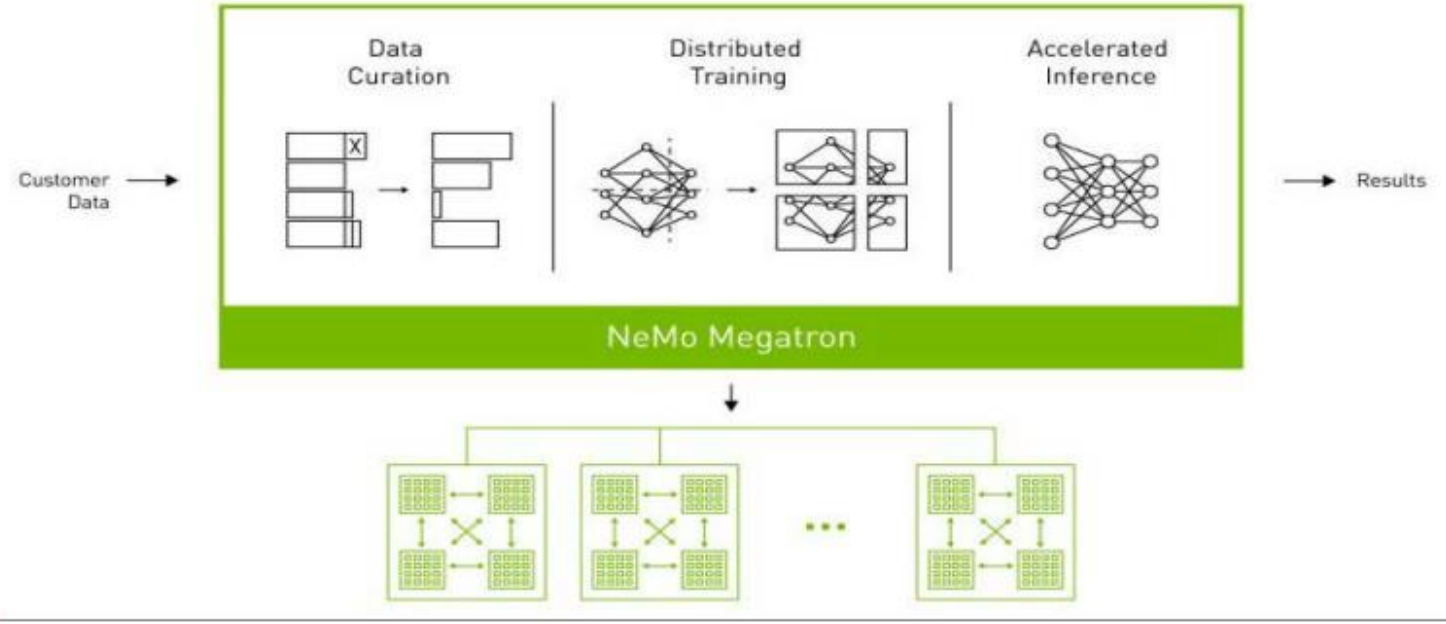


자료: Nvidia, 미래에셋증권 디지털리서치팀
주: 깨끗한 데이터 세트를 만들어주는 NeMo Curator / 특정 분야 데이터로 LLM을 조정할 수 있는 NeMo Customizer / AI 모델 성능을 분석하는 NeMo Evaluator

<출처: 엔비디아, 미래에셋증권>

NeMo(Nvidia Enterprise Modular AI)

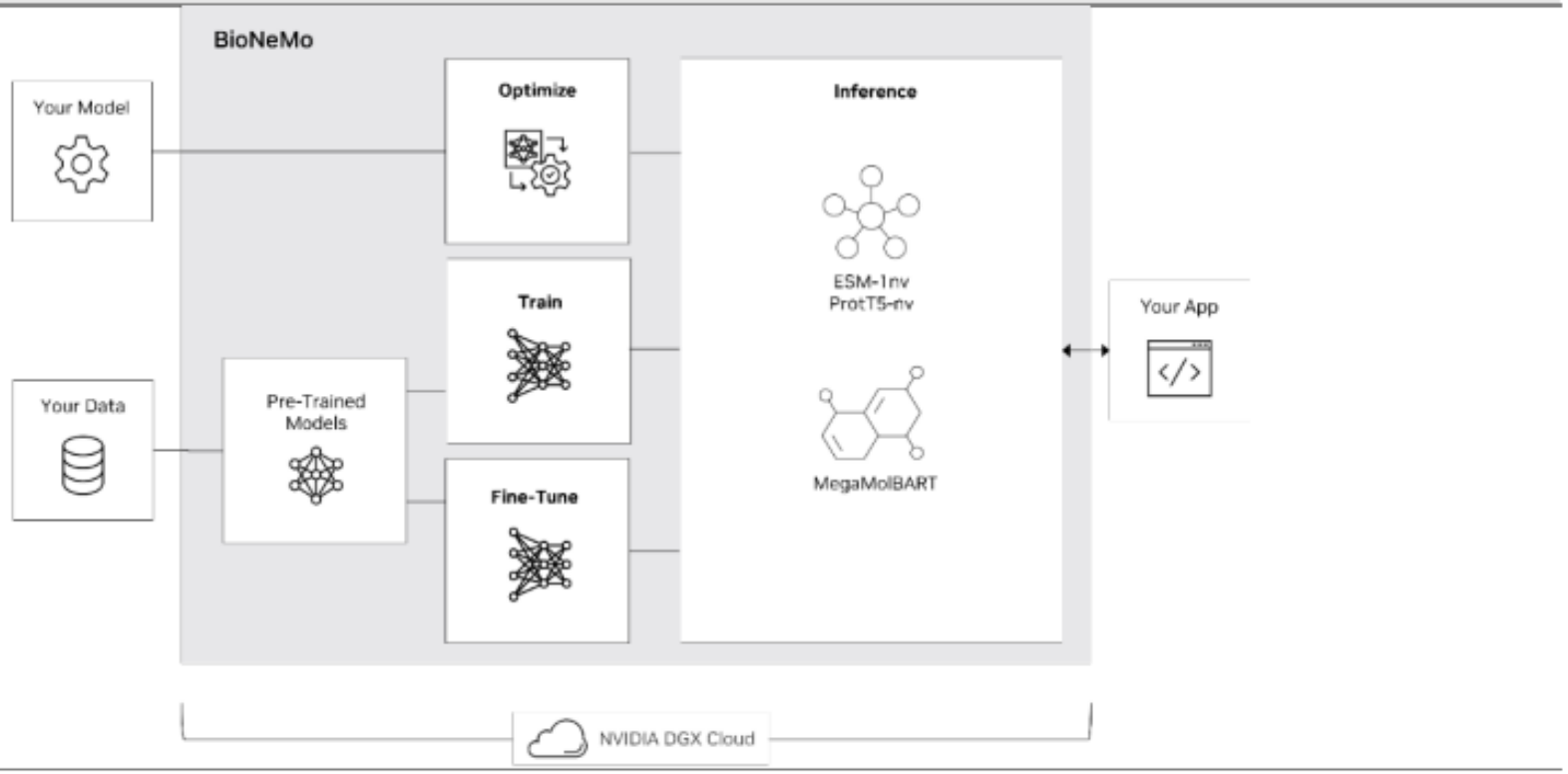
ChatGPT 출시 전인 2021년부터 엔비디아가 강조한 "NeMo Megatron"
오픈소스인 Megatron을 기반으로 기업들이 (독자적 데이터만 있다면) 각자의 LLM을 개발할 수 있도록 지원



자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

BioNeMo: 사용자가 신약 개발을 가속화하는 AI 모델을 만들 수 있도록 지원하는 특화 플랫폼
단백질 및 분자 설계를 위한 중앙집중식 모델 훈련, 최적화, 미세조정, 추론까지 용이하게 함

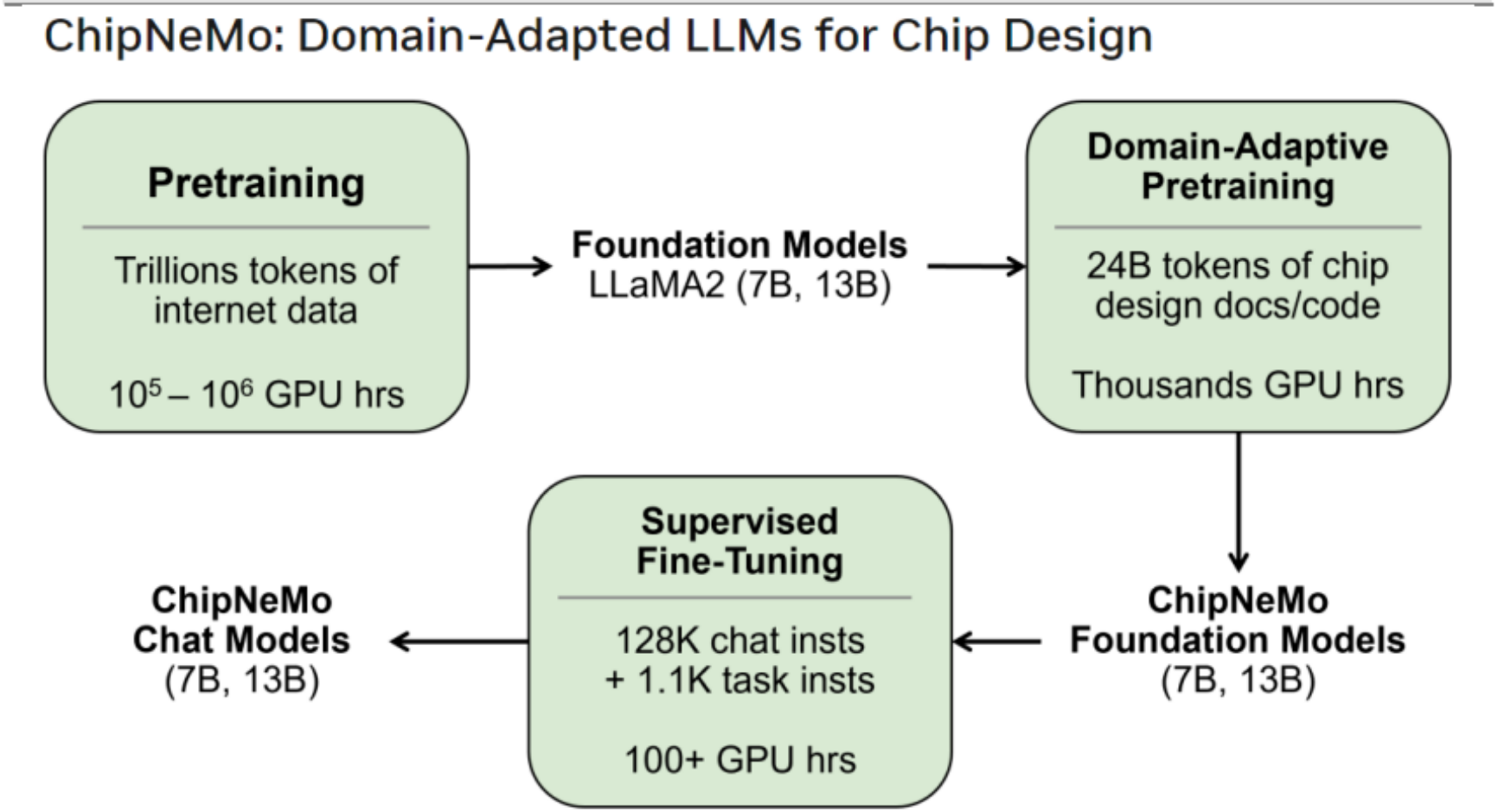


자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

ChipNeMo

ChipNeMo: 엔비디아가 라마2에 내부 데이터를 더해 만든 반도체 설계에 특화된 언어모델
내부적으로 챗봇, 코드 생성기, 버그 추적관리 자동화 등에 활발히 쓰임



자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

NeMo와 NIM

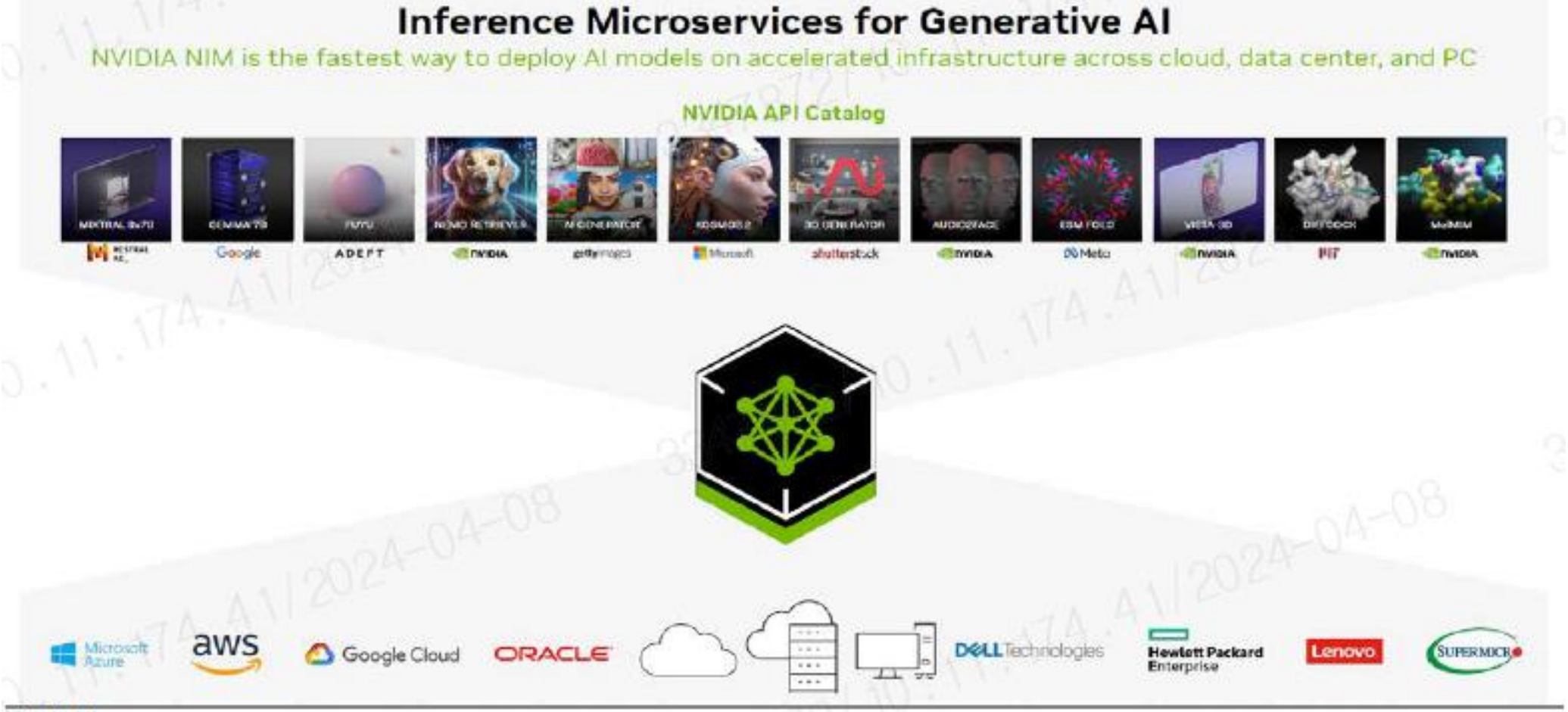
- NeMo의 역할이 각자의 기업용 AI 모델을 만들기 위한 준비/훈련/커스터마이징 툴이라면,

- NIM은 이 NeMo를 현실화하는, 즉 실제 비즈니스로 발전/승화시키는 중책을 맡게 된다.

<출처: 미래에셋증권>

NIM API 카탈로그

엔비디아의 NIM에서 제공하는 "API 카탈로그"
본인들의 모듈에 더해 믹스트랄, 젬마 같은 언어모델 오픈소스, 단백질 구조예측 모델 등이 포함



자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

젠슨 황이 말하는 NIM

마지막으로, NIM에 대해 젠슨 황은 이 같이 말을 했다. "각 기업들은 생성 AI 코파일럿으로 변환될 수 있는 데이터의 금광을 보유하고 있다" 많은 **기업들이 자신들만의 데이터 보물을 가지고 있는데, 어떤 기업이든 그 데이터를 기반으로 AI 회사가 될 수 있게 만들어주겠다**는 것이다.

여기서 엔비디아의 요술봉이 NeMo랑 NIM이다.

<출처: 미래에셋증권>

디지털 트윈

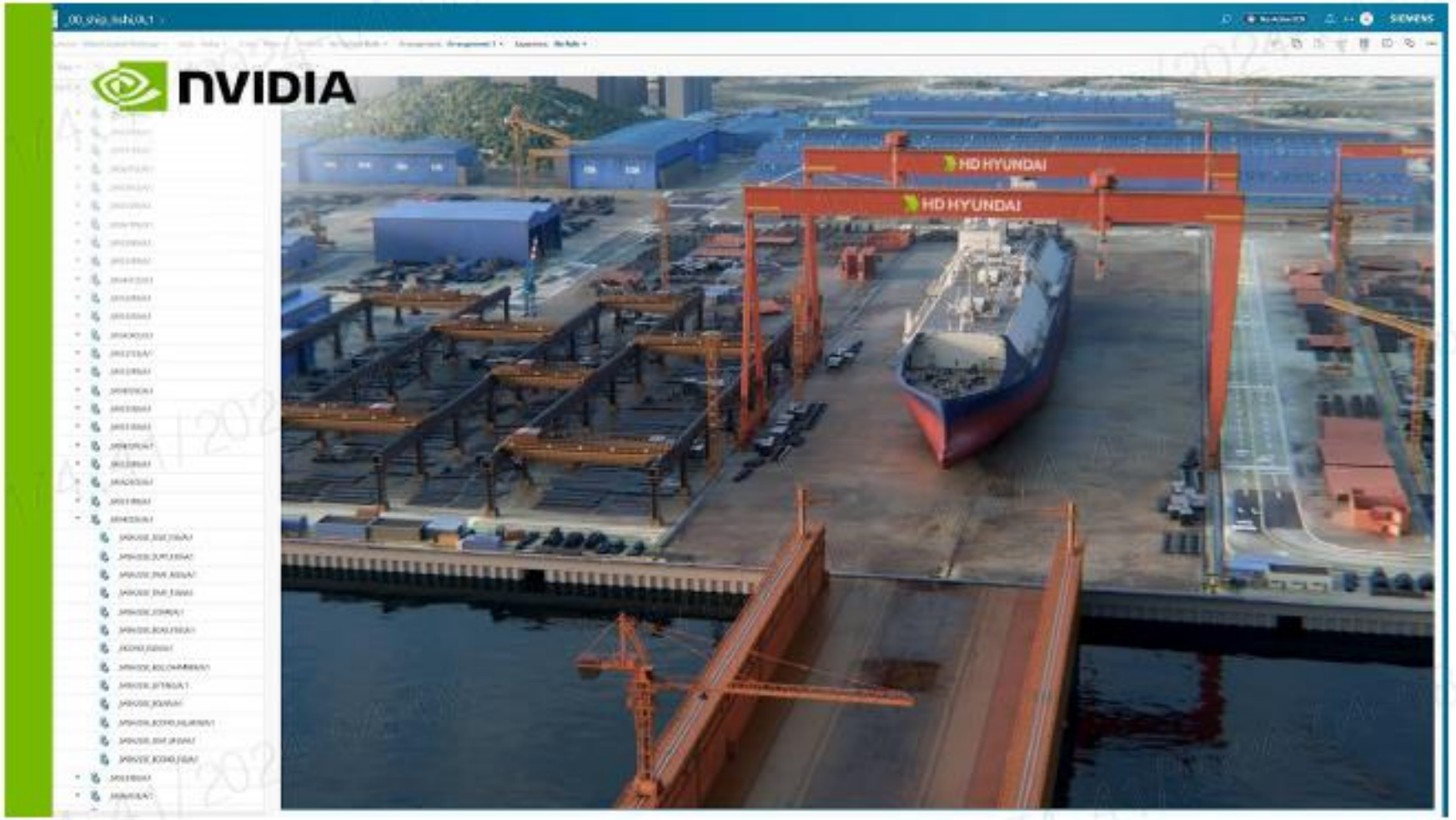
자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

디지털트윈 – 지구를 복제하다

<출처: 엔비디아>

GTC 2024의 옴니버스 비즈니스 케이스에서 가장 강조됐던 '현대중공업-지멘스'의 협력



자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

젠슨 황과 로봇들

GTC 2024 키노트의 백미는 젠슨 황과 함께 서 있던 수많은 로봇들
엔비디아는 로봇을 직접 제조한다기 보다 이 모든 로봇들의 머리가 되겠다는 야심

자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

옴니버스 – 자율주행과 로봇

옴니버스에서 자율주행과 로봇을 위한 기반모델이 만들어지고 가공되는 과정
DGX(로봇 기반모델인 GR00T 학습), OVX(로봇 가상 시뮬레이션), AGX(예: DRIVE, ISAAC에 탑재)

자료: Nvidia, 미래에셋증권 디지털리서치팀

<출처: 엔비디아, 미래에셋증권>

엔비디아의 자율주행

<출처: 엔비디아, 미래에셋증권>

Jetson Thor

엔비디아 AGX 반도체 칩이 탑재될 산업 로봇들과, AGX 칩의 모습
휴머노이드에 들어갈 Blackwell 아키텍처의 SoC인 "Jetson Thor"

자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

Isaac Lab

OVX에서 로봇이 시뮬레이션 되는 과정에 대한 상징적인 이미지와 단계별 과정 정리
인간을 모방하기 위해서 OVX 서버 기반 "Isaac Lab"에서 인간 시연자로부터 강화학습을 받는 형태

자료: Nvidia

<출처: 엔비디아, 미래에셋증권>

테슬라 VS 엔비디아

자율주행 차, 자율보행 로봇에서 가장 격렬하게 싸울 두 회사
'이제는 애플도 없다, 둘 뿐이다'

TESLA

자동차 제조기업 테슬라

수직적 통합 생산

전기차 & 로봇 원가 절감에 주력
테슬라 제품에 최적화된 칩 개발

NVIDIA®

반도체 설계기업 엔비디아

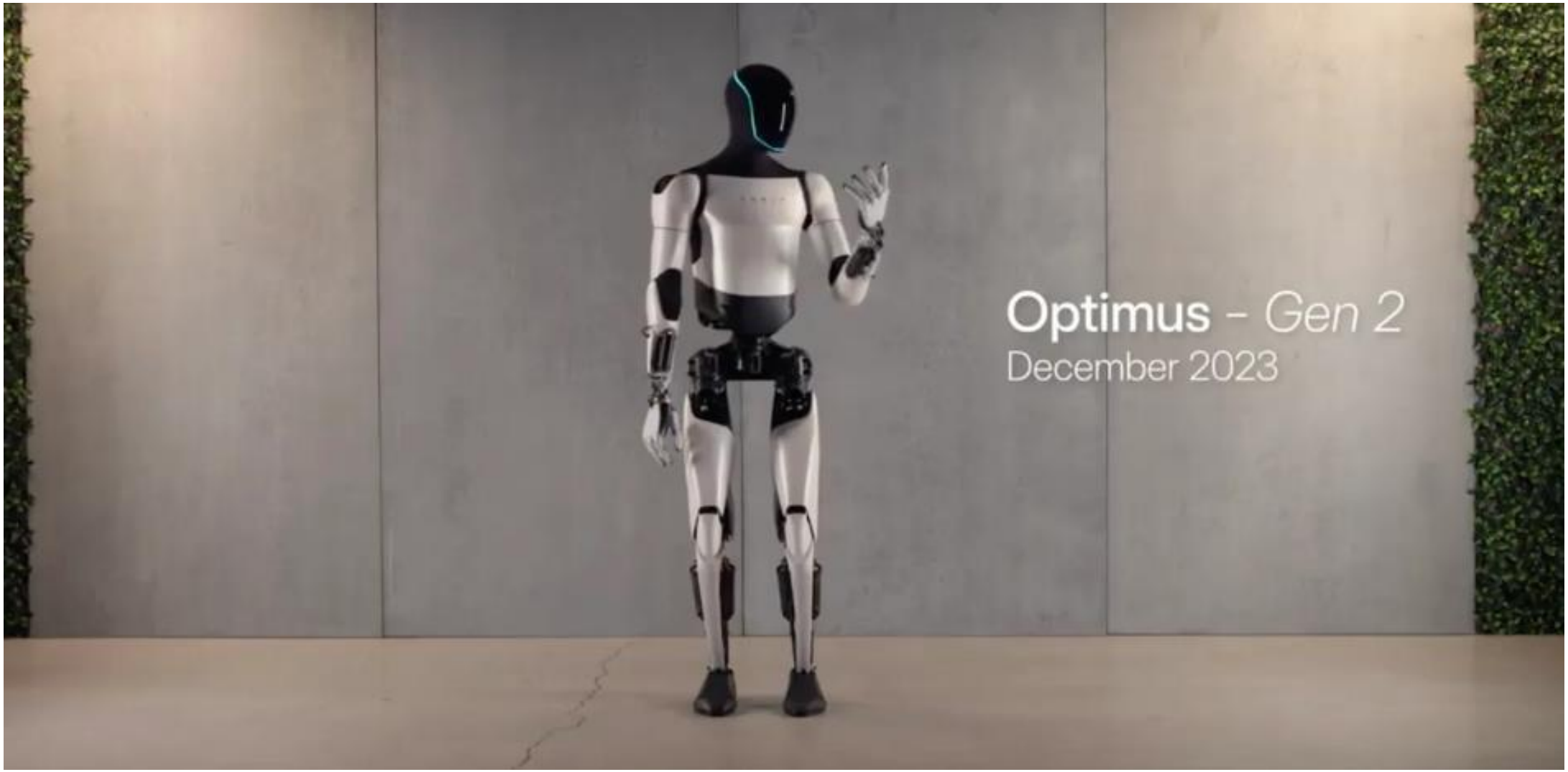
수평적 지원 플랫폼

어떤 전기차 & 로봇이든 지원하기 위한
AI 기반 모델 및 반도체 개발 주력

지료: Tesla, Nvidia, 미래에셋증권 디지털리서치팀

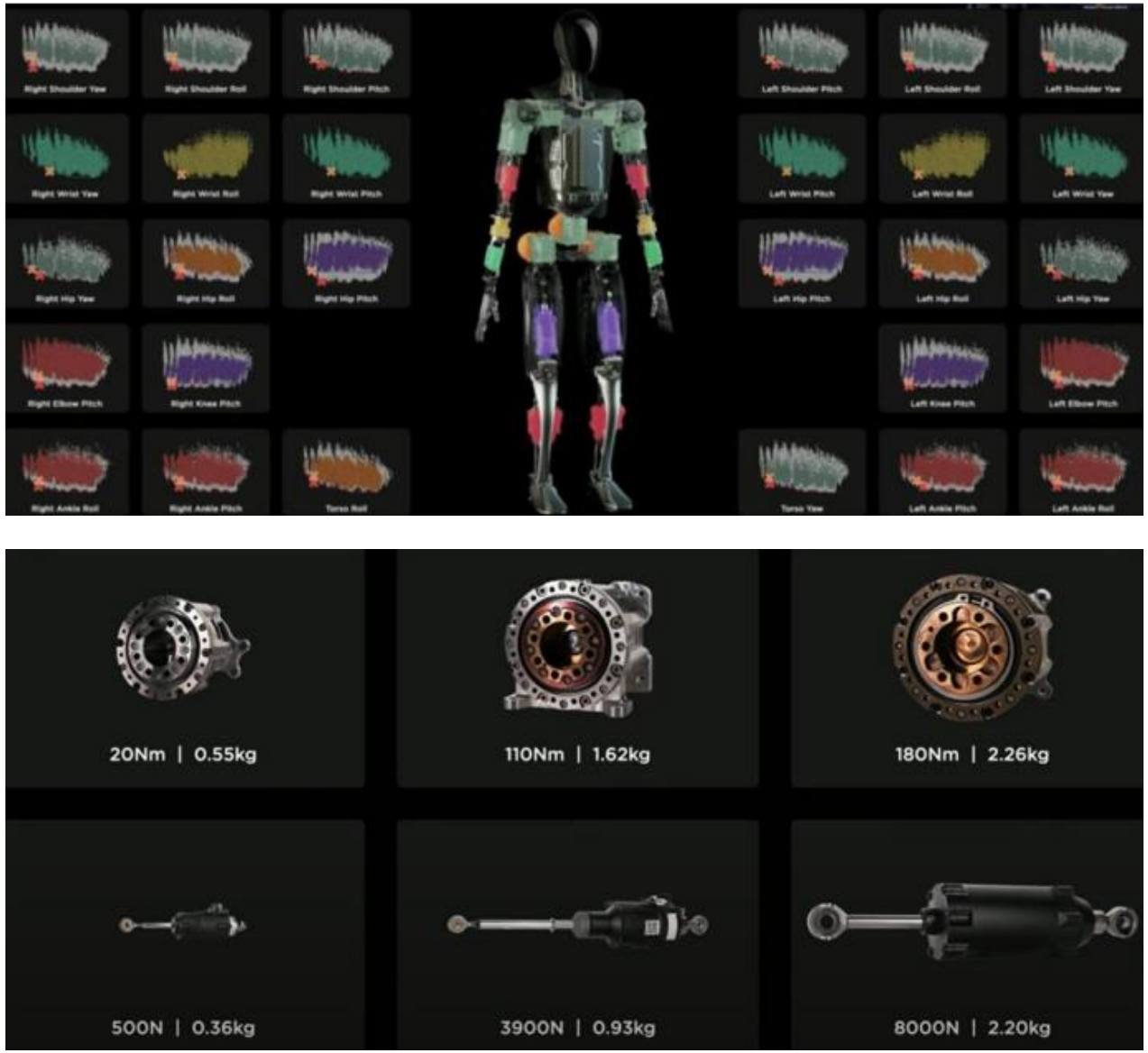
<출처: 미래에셋증권>

Tesla Optimus



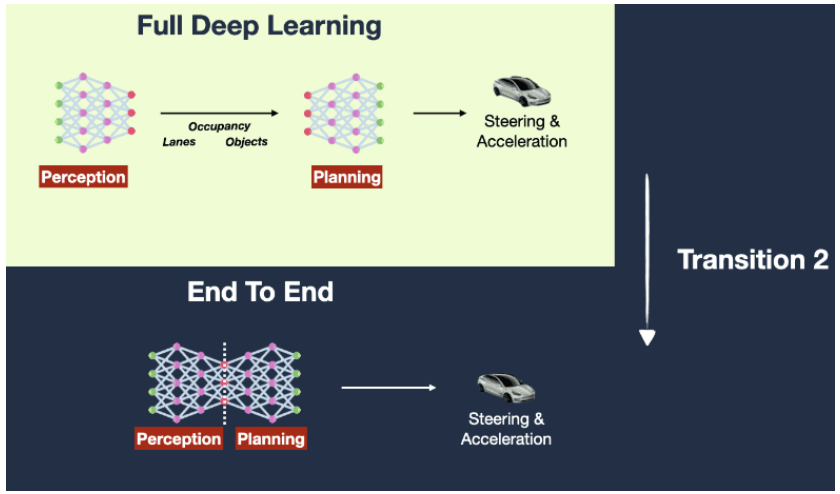
<출처: 테슬라>

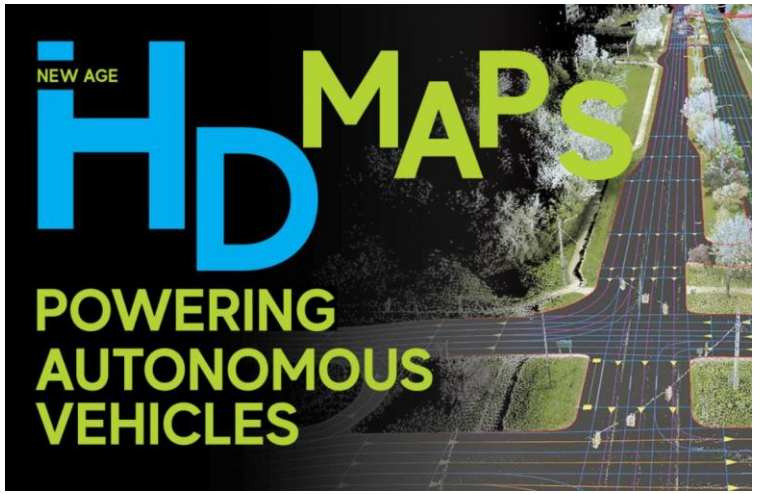
Tesla Optimus가 무서운 이유

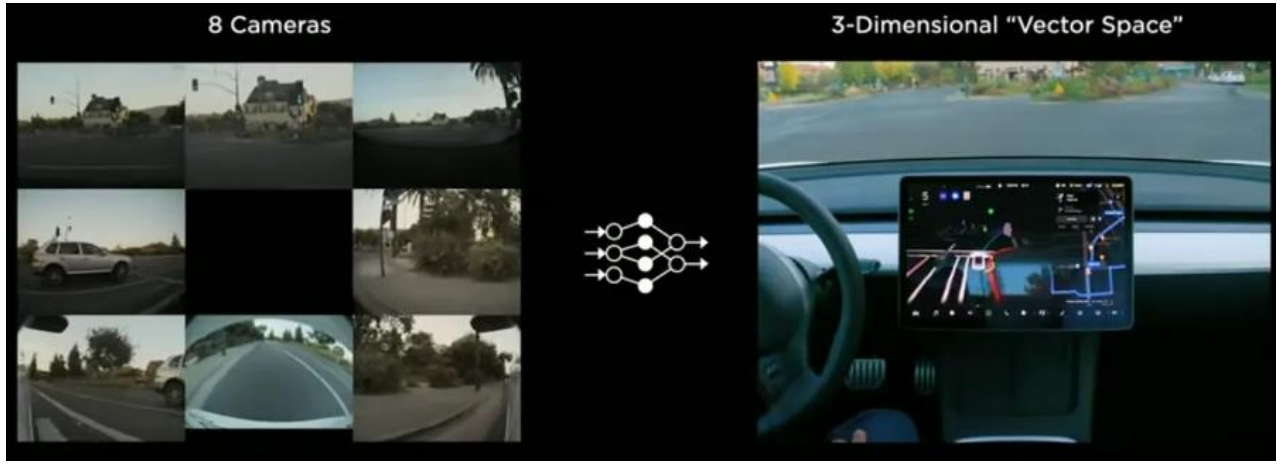


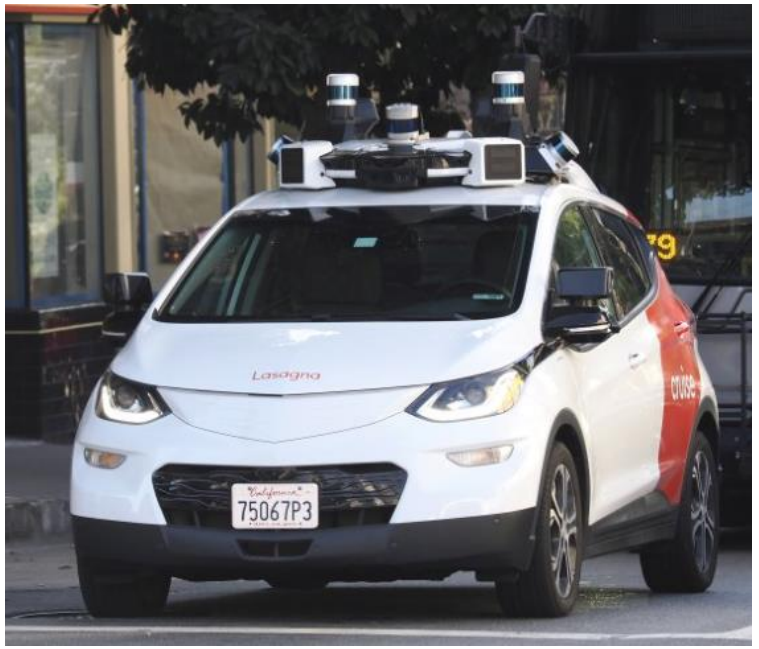
<출처: 테슬라>

테슬라의 차이









<출처: tesla, geospatialmedia, epa>

AI가 바꿀 우리의 미래 – AI투자를 해야 하는 이유

샘 알트먼

100조 달러

기본소득

월드코인

아이들의 직업의 미래?

기본소득으로 살아가는 세상

DALL·E

사람들이 월드코인을 통해 기본소득을 받으며 살아가는 미래 사회를 묘사한 이미지를 만들었습니다. 다양한 배경을 가진 개인들이 쇼핑, 공부, 여가 활동 등을 하고 있으며, 이 모든 활동은 월드코인 거래를 통해 이루어집니다. 환경은 고도로 발전되고 번영하며, 디지털 디스플레이와 인터페이스가 일상 거래에 월드코인을 사용하는 것을 나타냅니다. 이 이미지는 경제적 안정성과 자유를 전달하며, 기본소득이 일상 생활에 미치는 긍정적인 영 강조합니다.

<출처 : ChatGPT>

기본소득으로 살아가는 세상2

전 세계 사람들이 기본소득을 받으며 최소한의 생계를 유지하는 우울한 미래의 모습을 그린 이미지를 만들었습니다. 다양한 도시와 농촌 환경에서의 사람들이 지친 듯이, 그리고 체념한 듯한 모습을 하고 있습니다. 환경은 간소하고 기본적이며, 사람들은 단순한 일상 활동에 몰두하고 있으며, 생존을 위해 오로지 기본소득에만 의존하고 있습니다. 분위기는 우울하며, 기본소득이 제공하는 안전망에도 불구하고 최소한의 수입으로 살아가는 데 따른 한계와 어려움을 강조하고 있습니다.

<출처 : ChatGPT>

AI로 우리의 미래는 격변의 시대로..

AI에 투자하는 것은 생존의 문제일 수도..