2026-03-23

해외주식/AI

GTC2026 톺아보기

SK증권 해외주식/AI. 박제민

SK securities

Contents

Chapter 1

뉴 인프라는 예견됐다

SW HW는 주고받고

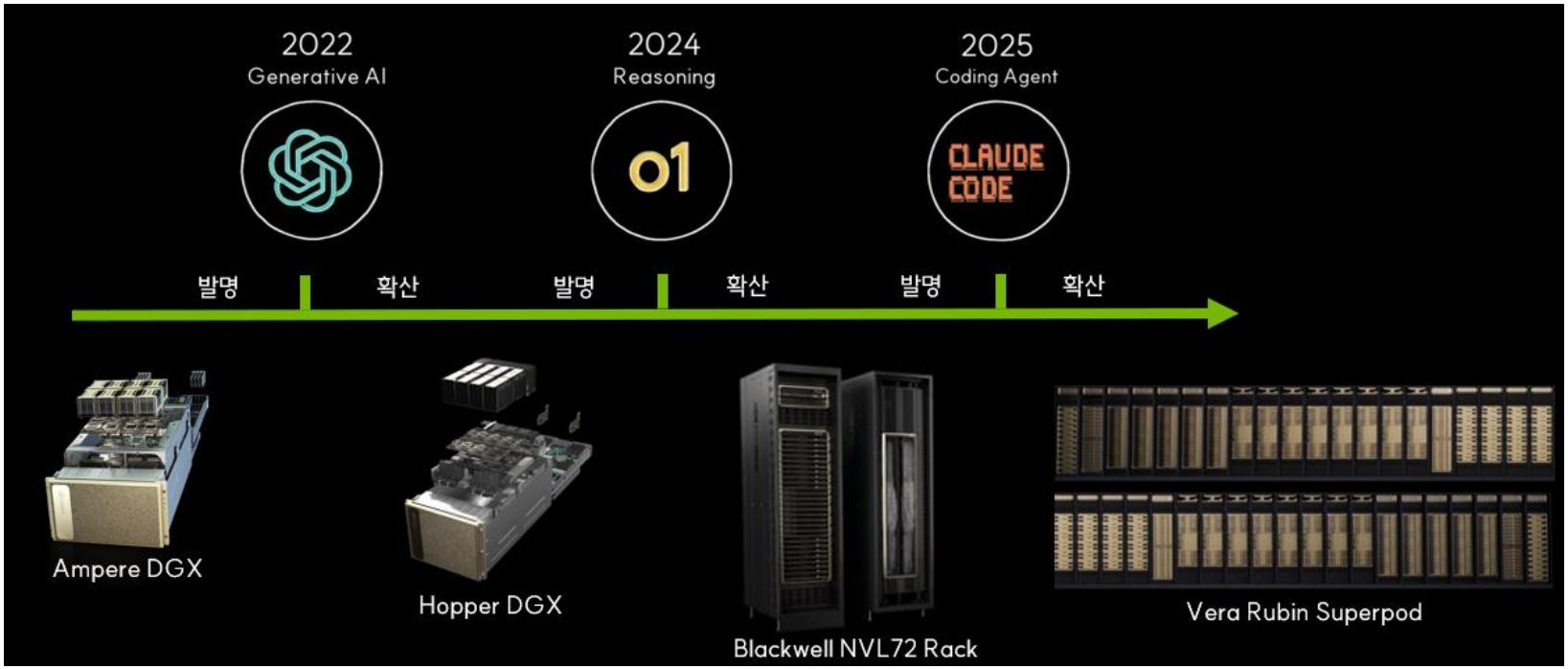
Agent시대를 확산시킬 Vera Rubin

Gen.AI 챗봇의 시대를 연 ChatGPT는 Ampere 하드웨어로 개발된 이후 Hopper 세대에서 확산

Reasoning 시대를 연 o1은 Hopper에서 개발된 후 Blackwell로 확산

Blackwell로 개발된 Coding Agent(Claude Code)는 Vera Rubin에서 확산을 앞둔 상황

주력 AI 제품 변화와 인프라의 변화



자료 : SK증권

주력 AI 제품: Chat → Agent

2025년까지 주력 AI 서비스는 챗봇

2023년 ChatGPT 등장, 2026년 기준 WAU 10억명 확보. 가장 성공적인 AI 서비스로 안착

2026년은 Coding Agent 확산의 해

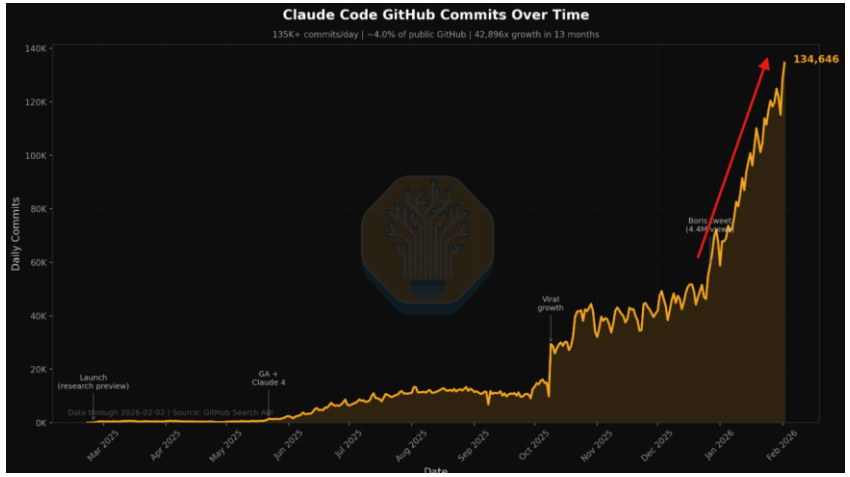
2026년 Claude Cowork, GPT-Codex 5.3 등장, Coding Agent 사용량 급증

엔진 성능의 향상으로 성과물 개선 + LLM 서비스와 결합되며 접근성 개선

OpenClaw는 좋아진 Coding Agent를 메신저로 접근 가능하게 하며 열풍을 재확산

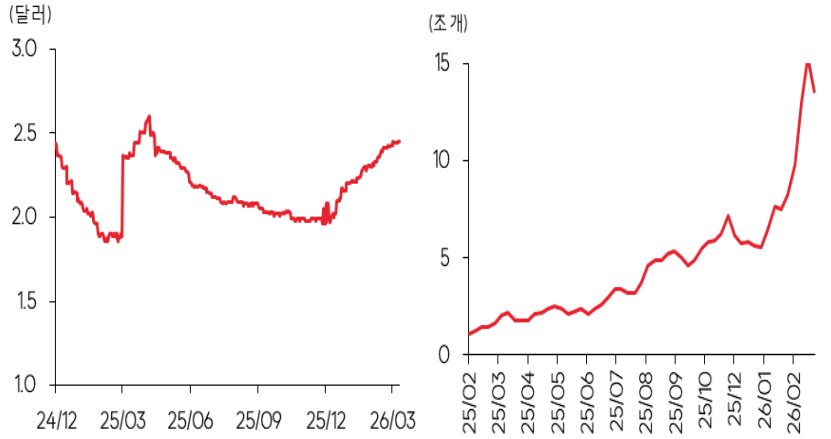
OpenRouter 기준 토큰 사용량 증가, 기존 워크로드 수요 늘면서 H100 렌탈 가격 다시 증가 추세

Coding Agent 활용량 2026년부터 급증



자료 : Nvidia, SK증권

H100 Rental 가격(좌), OpenRouter 토큰 사용량 (우)



자료 : Bloomberg, OpenRouoter, SK증권

Agent는 신규 DC를 원한다

요구 역량 차이: 챗봇 = 고지능, Agent = 업무 수행 능력

챗봇 시대 평가지표는 MMLU·GPQA 같은 정답형 지능 점수. 국내 언론은 GPT의 수능 점수를 소개하기도

Agent 시대 평가지표는 WebArena, METR 등 업무 성공률, 장기 과제 성공률, 툴 사용 신뢰성으로 변화

OpenAI, Amazon 투자 받으며 Stateful DC 개념 도입

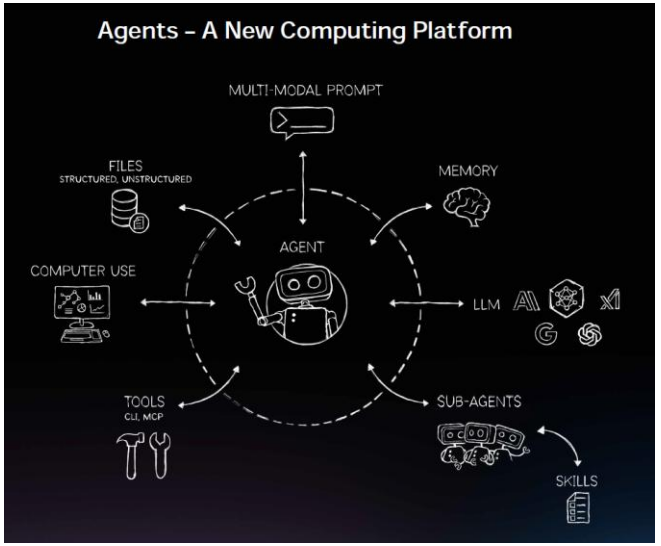
OpenAI, Amazon에 $50B 투자 받으며 8년간 2GW의 DC 사용 협약

해당 DC는 'Stateful Runtime, Frontier, Advanced workloads' 전용으로 사용

Stateful Runtime은 Agent가 오래 일할 수 있는 실행 환경 / Frontier는 에이전트를 기업에 이식하는 Palantir-like 모델

결론적으로 OpenAI도 기존 훈련 추론을 이루던 Azure 서버와 Agent를 활용하는 서버 환경을 분명히 구분

챗봇은 Agent에게 있어 하나의 툴일 뿐



자료 : Nvidia, SK증권

OpenAI, Amazon 파트너십과 New DC


2026년 2월 27일 회사

OpenAI와 Amazon, 전략적 파트너십 발표

문서 내용 듣기 4:27 공유

뉴스.

- Amazon Web Services(AWS)와 OpenAI는 OpenAI 모델로 구동되는 Stateful Runtime Environment를 공동 개발하며, 이는 Amazon Bedrock에서 제공되어 AWS 고객이 프로덕션 규모로 생성형 AI 애플리케이션과 에이전트를 구축할 수 있도록 지원합니다.

자료 : OpenAI, SK증권

Chapter 2

엔비디아 뉴 인프라 라인업은?

기존 제품 라인업 점검

Grace Blacwell 제품 라인업 점검: 6 Chip, 3 Tray, 2 Rack

칩 단위에서 랙 단위 판매로 도약. NVL72 Rack Scale 판매 비중 증가.

Rack 단위 판매량 늘어나면서 Network 사업부 매출 비중 증가

Nvidia Grace-Blackwell 라인업

자료 : Nvidia, SK증권

Vera Rubin 라인업

Vera Rubin 라인업: 7 Chip, 6 Tray, 5 Rack

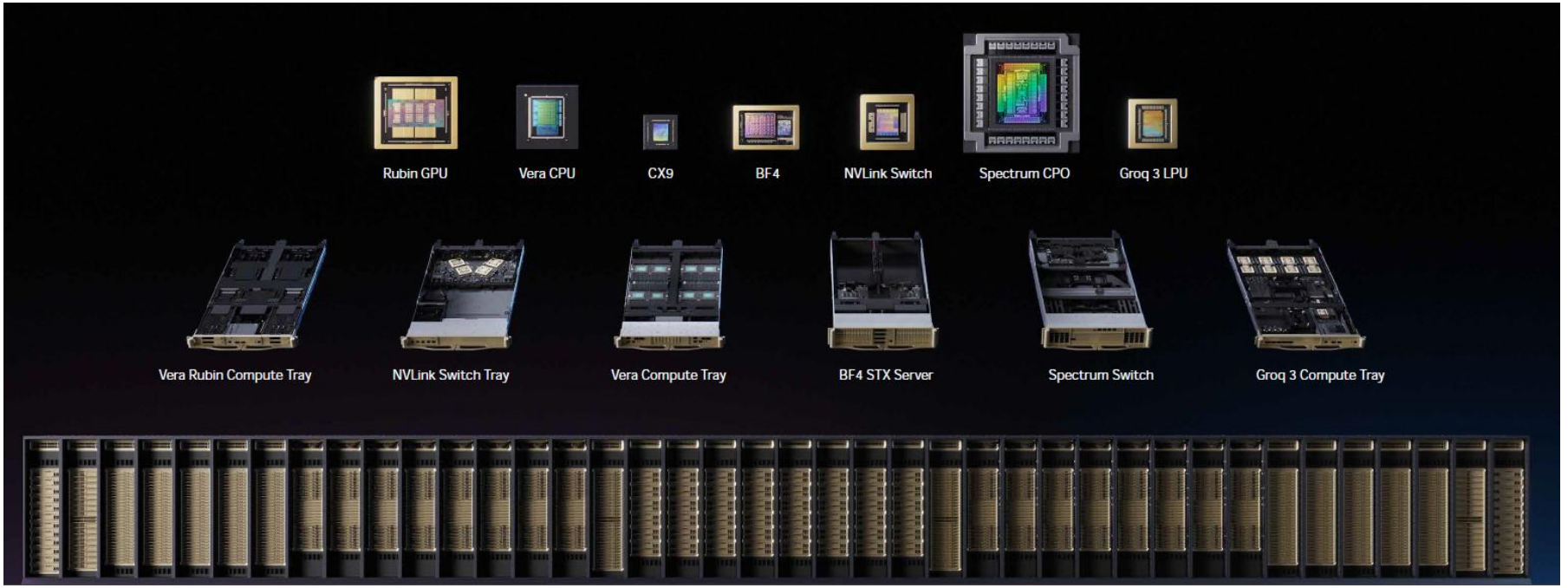
신규 Chip(1): LPU / 신규 Tray(3): Groq Tray, CPU Tray , BF STX / 신규 Rack(3): LPX, CPU, BF STX Storage

메인 제품 여전히 1) VR NVL72 Rack 2) Spectrum-6 STX. 신규 랙들의 경우 CSP의 수요를 짐작해봐야

추가된 신규 랙들은 모두 Agent 추론에 특화

**공개된 BluePrint SuperPod 기준 Rack, Tray, Chip의 위치 및 개수는 Appendix 참고*

Nvidia Grace-Blackwell 라인업



자료 : Nvidia, SK증권

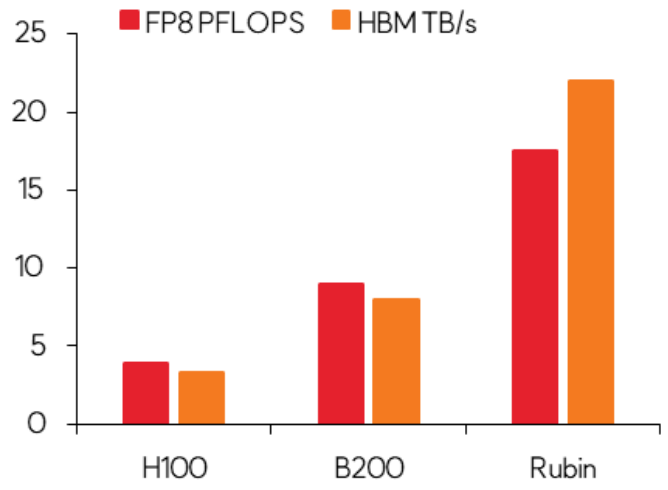
성능 지표 이해: Throughput & Interactivity

Agent 적합한 신규 지표 제시, Throughput & Latency

챗봇 시대 AI의 모델 처리 구조(GEMM)는 단순하여 FLOPS, HBM 대역폭 등 단일 칩 성능 위주의 인프라 성능 비교 가능
그러나 Agent는 처리 구조 복잡(long context, tool use 등). 엔비디아는 벤치마크로 Interactivity 대비 Throughput 강조

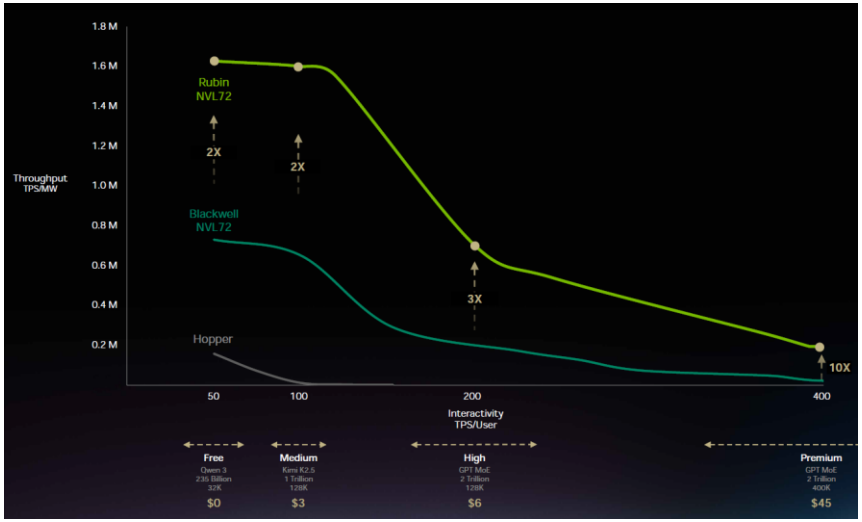
Y축: Throughput(TPS/MW), 높을 수록 같은 MW로 더 많은 초당 토큰 생성, 즉 좋은 전성비
X축: Interactivity(TPS/User), 높을 수록 유저에게 더 많은 초당 토큰 여유, 큰 모델에서 더 많은 context 제공 (고품질 서비스)
하향 곡선: Throughput과 Interactivity의 본질적 trade-off를 의미. 이는 Batch-size, decode 특성, 대역폭 한계 등에 의함
높은 TPS/User는 낮은 Latency 를 만들며 풍부한 토큰으로 긴 Reasoning, 더 많은 tool use, guardrail을 가능케함
X축의 Free~Premium 구간을 쉽게 ChatGPT의 Free - Plus - Pro - Deep Research 로 나눠서 생각 가능

Nvidia 세대별 GPU 단일칩 성능 비교



자료 : Nvidia, SK증권

Rubin NVL72 Throughput & Interactivity 개선



자료 : Nvidia, SK증권

성능 지표 이해: Throughput & Interactivity

Frontier edge 타파를 위한 Extreme Co-design

같은 Interactivity 내에서 더 높은 전성비를 제공하기 위해서 유동적인 병목 해소 필요

단순히 GPU, HBM 성능만 증가하면 GPU가 노는 시간이 늘어나는 현상 발생 (long GPU idle time)

Vera Rubin은 GPU·CPU·NVLink·NIC·DPU·스토리지·쿨링·전력·패키징·SW를 한 시스템으로 같이 최적화

High Interactivity 구간에서 개선 집중, 10배 이상의 전성비 개선

TPU/User가 높은 구간에서 성능이 더 크게 개선, 이유는 TPS를 높이는 구간에 GPU idle time이 높았기 때문

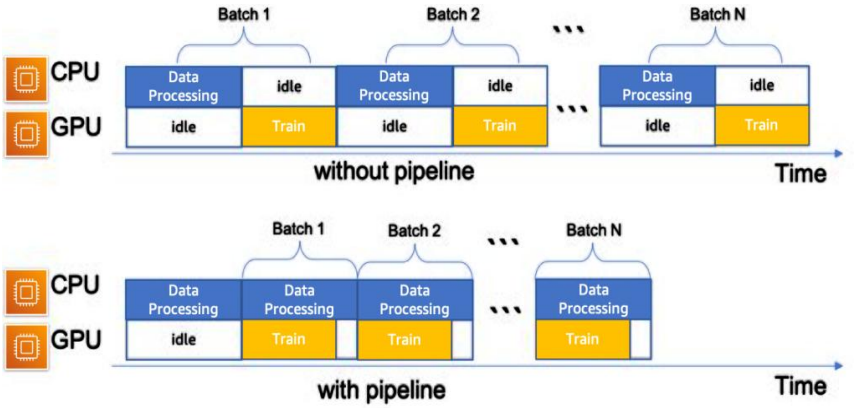
VR 인프라는 Blackwell로는 접근이 어려웠던 TPS 수준까지 도달 가능, Premium 구간은 10배의 전성비 개선

ChatGPT의 Deep Research가 20~60분씩 걸렸던 이유는 GPU, HBM이 좋지 못해서가 아니었다는 것

Interactivity에 따라 가격이 구분돼있는 ChatGPT 서비스

자료 : Nvidia, SK증권

GPU idle time



자료 : AWS, SK증권

공급 및 운영 안정성 개선

Compute Tray 조립 시간 단축

내부 케이블을 없애고 모듈 간을 board to board connector로 연결하는 구조

2025년 GB200/300 Compute tray 내부 케이블 조립 중 스크래치, 단자 손상이 많은 결함과 공급 지연을 야기

VR은 개선을 통해 조립 시간이 5분으로 단축(vs GB 2시간), 조립 신뢰도 향상 및 향후 고밀도 설계에서도 더 유리

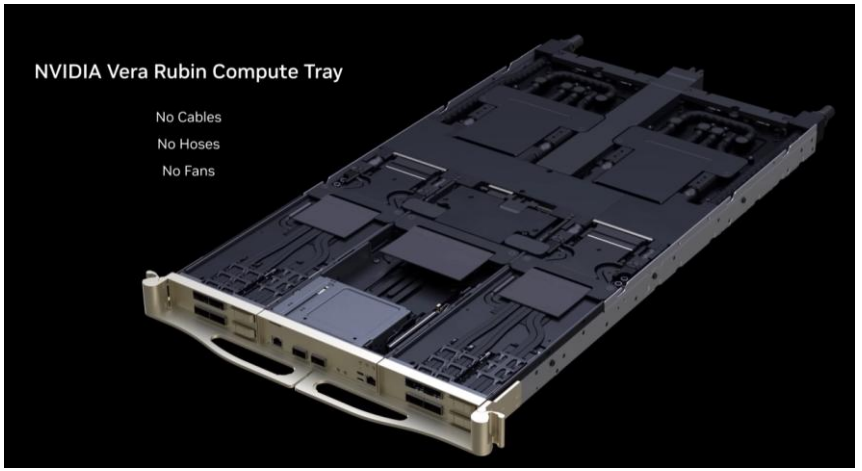
CPO로 운영 안정성 확보

Scale out 칩인 Spectrum-X 칩부터 CPO 적용. 전성비 및 배선/설치 복잡도 개선으로 운영 효율성 개선

Semianalysis에 따르면 전성비 개선은 전체 클러스터 기준 1% 수준으로 제한적, 운영 경험 적은 네오클라우드 수혜 전망

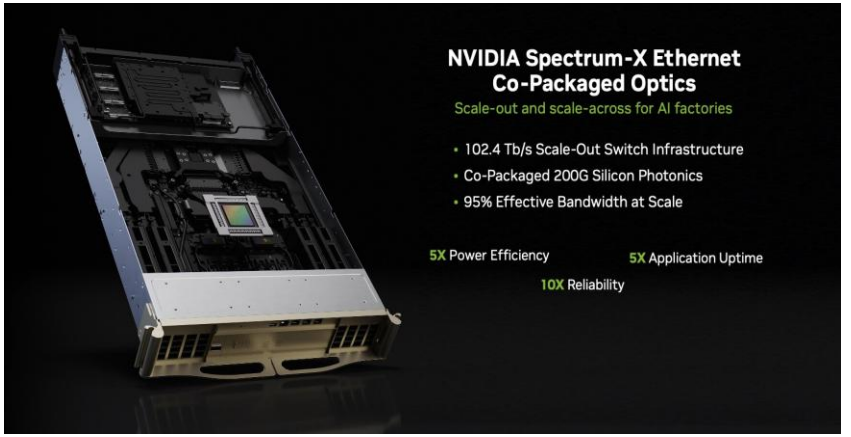
시장은 Scale up 확장 기대, NVLink 8 Optics(2028) 언급. Scale up 칩은 현재 Rack 당 36개 탑재되는 큰 TAM

케이블, 호스, 팬이 없는 Vera Rubin Compute Tray



자료 : Nvidia, SK증권

Spectrum-X Ethernet CPO 칩



자료 : Nvidia, SK증권

New 인프라 점검 1: LPU

LPU란? On-chip 메모리 설계로 초저지연 달성한 연산칩

LPU 칩은 On-chip 기술을 통해 연산과 메모리를 한 칩 내에 구현

공간 제약으로 적은 용량(500MB vs HBM4 288GB), 좁은 선폭 활용하여 높은 대역폭(150TB/s vs HBM4 22TB/s)

젠슨황은 LPU가 향후 추론 연산의 25% 수준을 담당하게 될 것으로 전망

LPX를 추가로 사가면 Premium 서비스가 가능해집니다

Batch를 크게 키울 수 없는 high interactivity 구간에서 GPU는 병렬처리 장점을 살리기 어려우며 decode 효율성 저하

인프라 SW인 Dynamo를 통해 Prefill/Decode를 분리, 비효율적인 decode 부문을 LPU가 처리하는 컨셉

Rubin+LPX 활용 시 Rubin 단독 대비 고성능 서비스의 전성비 급증(10x→35x), 현재 미구현 서비스도 제공 가능

On-Chip SRAM 구조

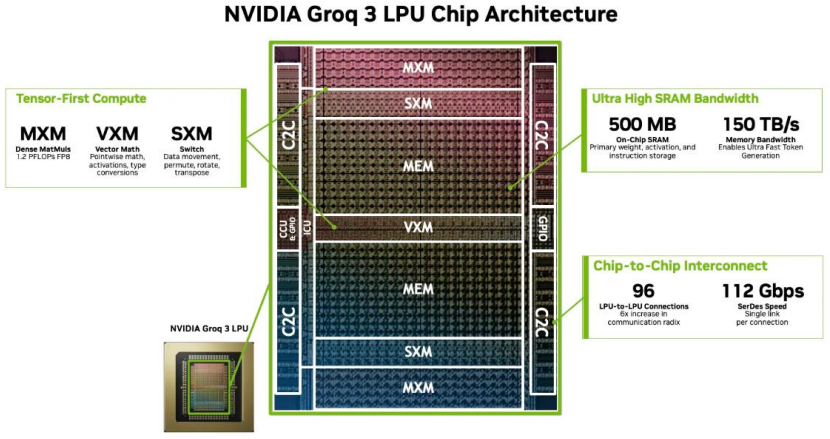
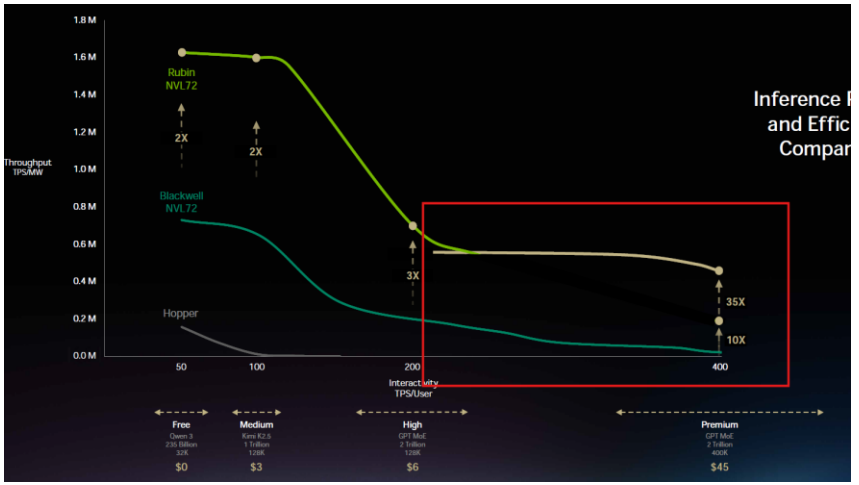


Figure 3. NVIDIA Groq 3 LPU chip architecture

자료 : Nvidia, SK증권 / 주: MEM=메모리 공간, 나머지는 연산 및 스위치

Higher Interactivity 구간의 Throughput 대폭 증가



자료 : Nvidia, SK증권

SRAM을 쓰면 HBM을 덜 쓰게될까?

메모리 최적화는 AI 서비스 번영, 1차적으로 업셀 2차적으로 제본스의 역설

GTC Analyst QnA에서 직접 나온 질문으로 현재 투자자들의 실질적 우려

Bofa Analyst, SRAM이 HBM을 cannibalize하지 않는지 질문

젠슨황의 답변은 'cannibalize이 아닌 upsell', SRAM 활용분을 잠식이 아닌 추가 판매로 인식

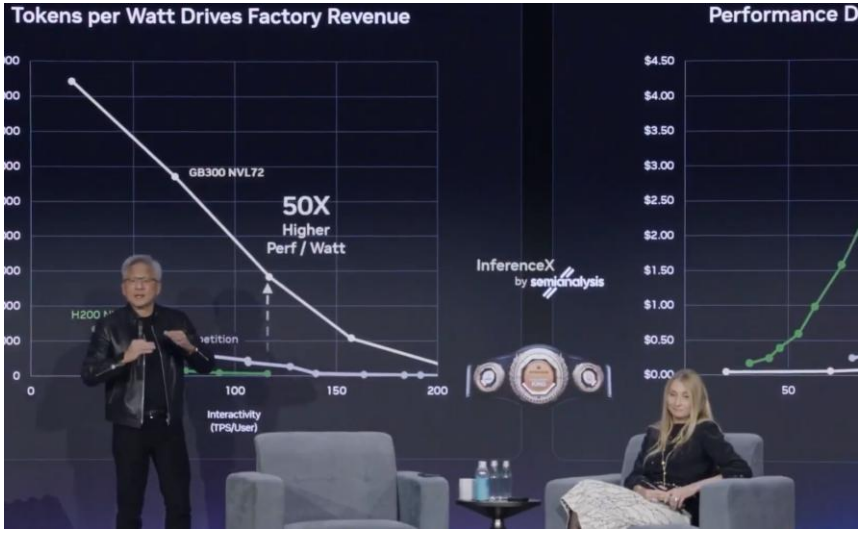
Throughput/Interactivity Trade-off를 고려할 경우 LPX는 AI의 상업성을 높이는 인프라

1) ChatGPT DeepResearch는 GPU+HBM 조합만으로 빠른 답변 서비스 제공 어려움 (낮은 가성비)

2) SRAM으로 DeepResearch 답변이 빨라지면 소비자 만족 → 결제 이용자 증가 → LPX 주문량 증가 (Upsell)

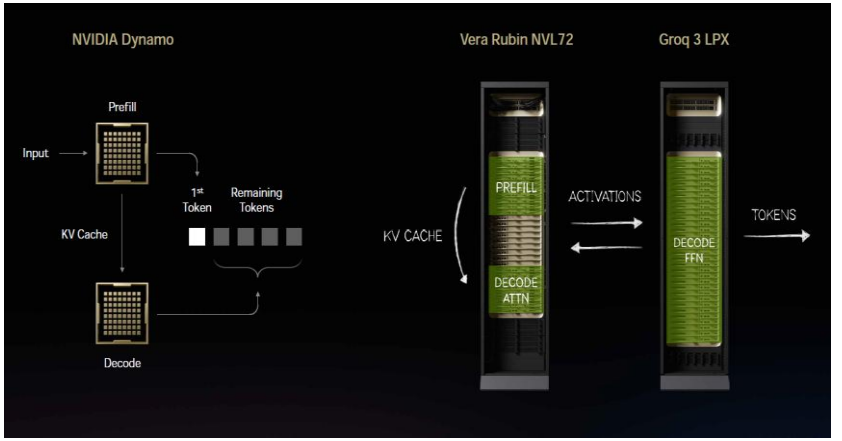
3) 빠른 DeepResearch에 만족하여 자주 ChatGPT 이용 → 전체 추론 수요 증가 + 모델 훈련 니즈 증가 (제본스의 역설)

GTC 이후 Analyst QnA



자료 : Nvidia, SK증권

Dynamo를 통해 Decode FNN 부분만 LPX가 처리



자료 : Nvidia, SK증권

LPU의 경쟁자 Cerebras?

저지연성을 겨냥한 경쟁자 Cerebras

Cerebras는 웨이퍼 하나 전체를 칩처럼 활용하는 구조 (Wafer-Scale Engine=WSE)로 저지연 최적화 칩

온칩 SRAM 44GB(vs LPX Rack 128GB), 대역폭 21PB/s (vs LPX Rack 40PB/s)

개별 칩 기준으로 open-weight 추론 모델의 decode 속도가 Groq-3 대비 우위

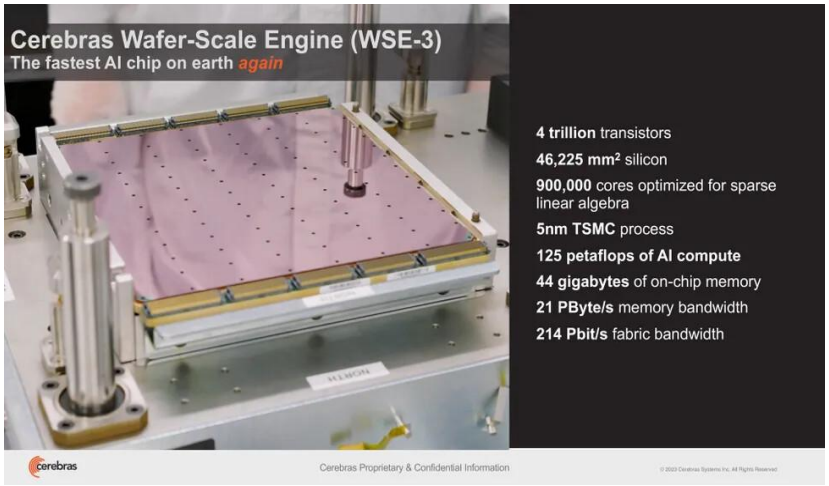
그러나 이는 Dynamo SW가 LPU 특화 decode만 부여해주기 이전 수치, 엔비디아 생태계에서의 성능은 이보다 높을 것으로 추정

LPX 판매량은 Cerebras과의 성능 경쟁이 변수로 작용

2026년 1월 OpenAI와 750MW Compute 계약 체결, 이미 추론 서비스에 활용 중 (Codex-Spark)

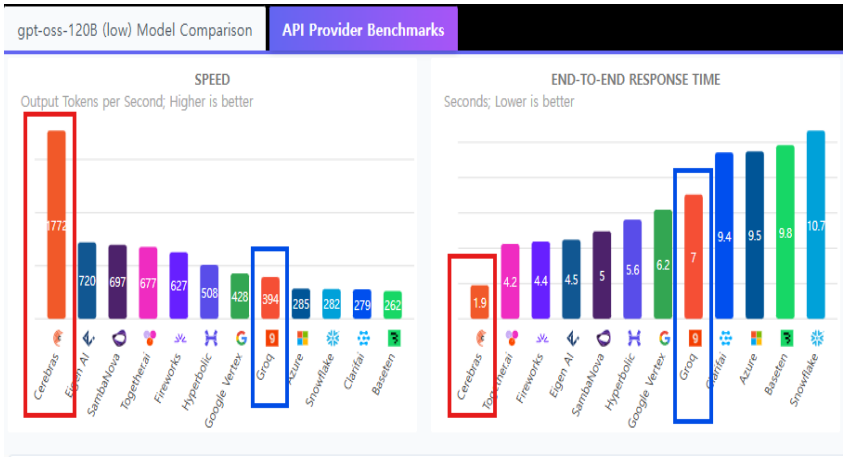
2026년 3월 AWS와 공식 파트너십, DC에 CS-3 배치 후 Bedrock 제공. Decode 부문 추론에 활용 예정

Cerebras WSE-3 칩



자료 : Cerebras, SK증권

gpt-oss 기준 추론 속도 비교 (2025.08 기준)



자료 : Artificial Analysis, SK증권

New 인프라 점검 2: BF STX

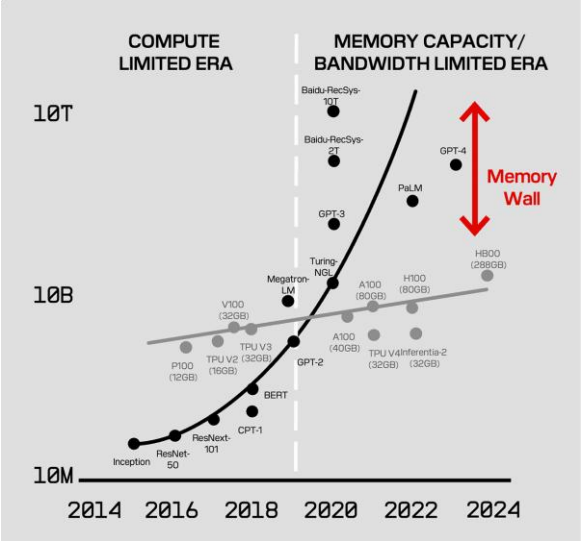
메모리 효율화를 위한 인프라

Agent 시대에 폭증하는 KV cache/context를 따로 처리하는 '컨텍스트 메모리·스토리지 전용 인프라 랙'
기존 ICMS(Inference Context Memory Storage)가 GTC를 계기로 CMX로 리브랜딩된 것으로 추정 (공개 슬라이드 기준)
GPU 대비 대역폭, 용량 성장이 제한적인 Memory wall 해결이 목표
아직 구체적인 랙 구성 미공개, BF-4가 메인 프로세서로 작동할 것으로 추정

중요 임무를 받은 녀석은 BlueField-4

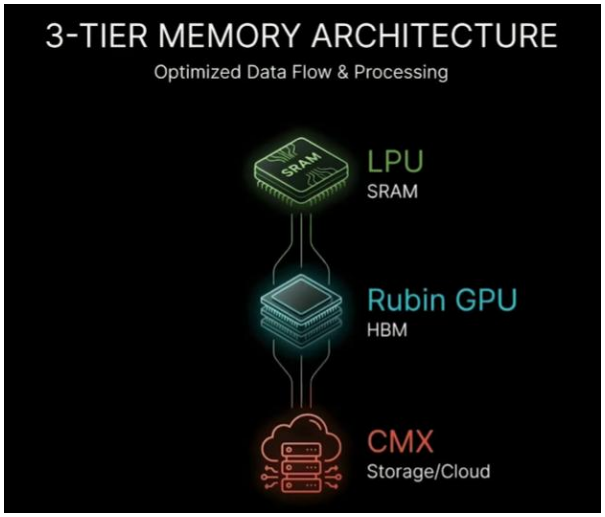
BF-4는 이전 세대 CPU인 Grace 칩 활용하여 인프라 운용을 위한 프로세서로 활용, Vera CPU의 업무 분담
BF-4가 부여받은 가장 큰 업무가 스토리지 컨트롤, AI Factory 내의 스토리지를 효율적으로 관리

GPU 성능 발전 대비 메모리 발전 미약, Memory wall



자료 : Sandisk, SK증권

3개 티어 메모리의 최적화



자료 : Youtube(안될공학), SK증권

New 인프라 점검 3: CPU standalone

GPU를 위해 CPU가 필요하다

CES에서 젠슨황은 CPU도 향후 출하 1위를 목표한다고 언급, 이번 GTC에선 Vera-only tray, Standalone CPU rack 출시

Agent 구동에 있어 CPU가 최고 성능이 되지 못하면 GPU idle이 발생. Agent 실행 / 오케스트레이션 역할에 집중

기존 2 GPU - 1 CPU 구조는 충분하지 않으며 GPU가 여러 다른 Vera 코어에 서비스 될 수 있다고 언급

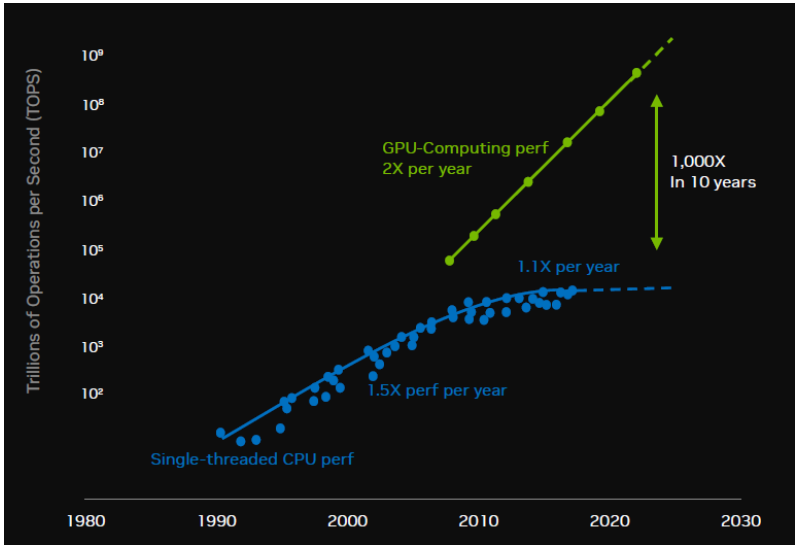
왜 인텔 말고 엔비디아껄 써?

Core 수가 아닌 빠른 단일 Thread, I/O에 집중

기존 CPU들은 vCPU 형태로 배포하기 위해 많은 코어 수에 집중. Vera는 Agent 호환성을 위해 대역폭에 집중

SOCAMM 활용 LPDDR로 코어당 대역폭이 직전 최고 사양 CPU 보다 3배 높은 수준

GPU 성능 발전 대비 무어의 법칙 미약, CPU wall



자료 : Nvidia, SK증권

Vera와 다른 Hyperscaler CPU 스팩 비교

제품	코어 수	Thread 수	메모리 대역폭	대표 I/O
Vera CPU	88	176	1.2TB/s	1.8TB/s
Xeon 6980P	128	256	845GB/s	378GB/s
EPYC 9965	192	384	614GB/s	504GB/s
Graviton4	96	N/A	538GB/s	378GB/s

자료 : Nvidia, SK증권

Vera Rubin 이후는?

2027년은 Rubin Ultra, 핵심은 Kyber Rack

칩 변화는 HBM4e, NVLink 7 Switch, LP35 업그레이드. 핵심 변화는 Tray 및 랙 구조의 변화. Kyber NVL144 적용

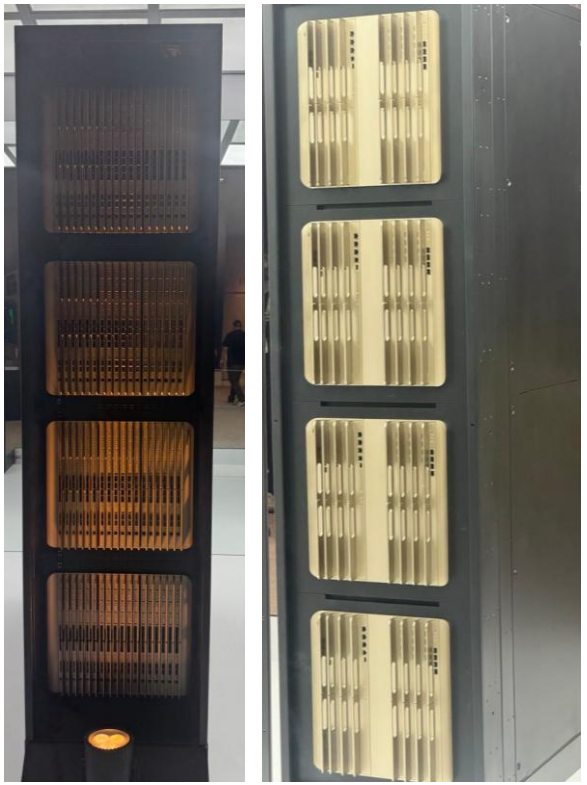
2028년은 Feynman, 알려진 건 없지만..

NVL1152 확장 목표, NVLink 8 CPO 도입 목표. 모든 칩들 개선 (BF, CX, LP40 등)

Oberon 아키텍처(NVL72) 앞면(좌), 뒷면(우)

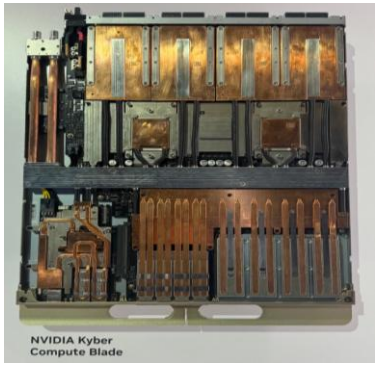
자료 : SK증권

Kyber 아키텍처(NVL576) 앞면(좌), 뒷면(우)



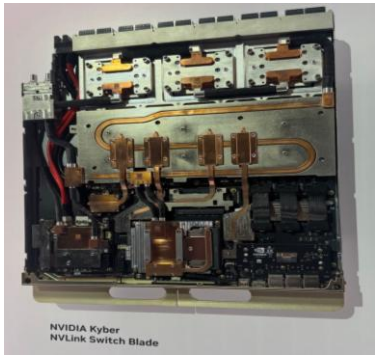
자료 : SK증권

Kyber Compute Tray



자료 : SK증권

Kyber Switch Tray



자료 : SK증권

Chapter 3

뉴 인프라로 펼쳐질 Agent 세상

토큰 경제의 시작

챗봇은 검색, Agent는 B2B SW

ChatGPT가 출시된 초기 가장 주가가 많이 흔들렸던 기업은 Google, GenerativeAI는 검색 사업 겨냥 상품이었기 때문
검색은 Googling 부터 무료 서비스, 현재 챗봇은 10억명 이상의 WAU를 확보했으나 여전히 수익화에 난항
Coding Agent가 부상하자 주가가 흔들리고 있는 건 B2B SW 기업들, 기존에도 돈을 지불하던 산업
Claude Code, Cowork, GPT Codex 는 챗봇 제품 대비 높은 B2B 비중, 높은 유료 구독 전환율, 빠른 토큰 소진 속도 기록

치고 올라오는 Anthropic

Anthropic 은 OpenAI 대비 WAU 수는 300 배 낮으나 2026년 3월 기준 ARR 차이는 25% 내외 (ANT 19B vs OAI 25B)
Anthorpic의 높은 ARPU는 겨냥 시장의 차이에서 비롯. 챗봇 이용 유저가 많은 OpenAI 대비 Coding Agent 유저가 많은 상황

Coding Agent로 토큰당 과금 사용자가 많은 Anthropic

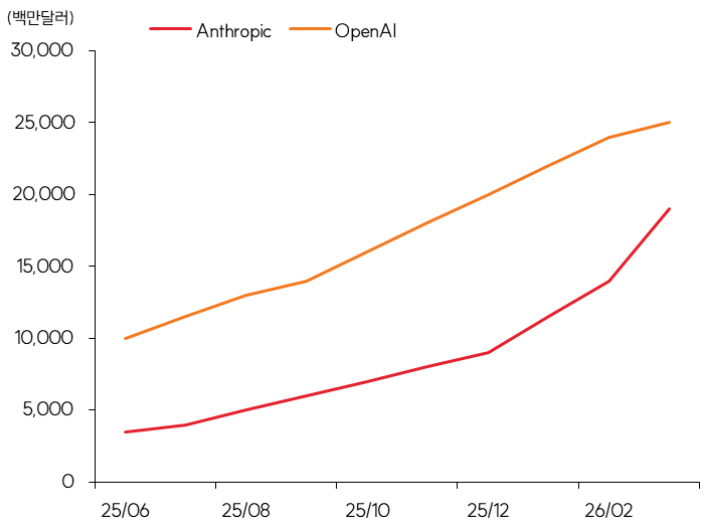
Model pricing

The following table shows pricing for all Claude models across different usage tiers:

Model	Base Input Tokens	5m Cache Writes	1h Cache Writes	Cache Hits & Refreshes	Output Tokens
Claude Opus 4.6	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.5	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.1	$15 / MTok	$18.75 / MTok	$30 / MTok	$1.50 / MTok	$75 / MTok
Claude Opus 4	$15 / MTok	$18.75 / MTok	$30 / MTok	$1.50 / MTok	$75 / MTok
Claude Sonnet 4.6	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok
Claude Sonnet 4.5	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok

자료 : Anthropic, SK증권

OpenAI, Anthropic ARR 추이



자료 : 언론 종합, SK증권

토큰 경제의 시작은 무슨 의미일까

Token = Revenue, Throughput(TPS/Watt) = 경쟁력

젠슨황은 CES, 실적발표에 이어 'Token = Revenue'을 다시 강조

Coding Agent가 이미 TAM $2T의 SW 산업을 잠식 중, Token의 매출화를 달성하고 있는 AI Labs 출현

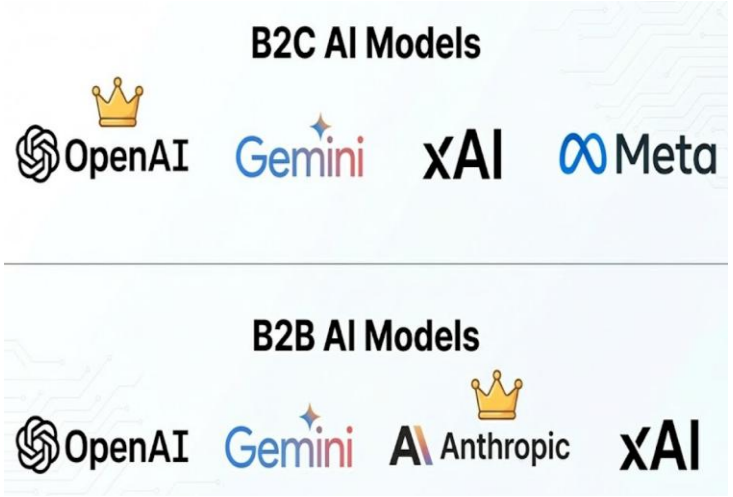
1) 전력 제약 상황 2) 서비스 경쟁 상황에서 Throughput = Token 가성비 = 매출 경쟁력

Throughput이 높은 기업들은 더 많은 TPS를 유저에게 제공 가능, 서비스 해자 구축

예를들어 높은 Throughput 기업의 Coding Agent가 더 고능한 모델로, 더 빠르게, 더 많은 것을 검토해서 업무 처리 가능

향후 Shopping Agent, Palantir-Like 모델에서의 워크플로우 자동화에도 Throughput이 곧 서비스 해자를 결정 전망

주요 AI Labs 경쟁 현황



자료 : SK증권

Shopping Agent

자료 : Retaligentic, SK증권

본격적으로 기업 수요를 유도할 차례

아직 기업들은 AI 활용 초기 단계

미국-유럽 기업 기준 IT 예산 YoY는 아직 Public Cloud 도입기보다 낮은 수준, 여전히 AI 는 R&D 비용으로 사용

AI 도입에 전념 중인 B2B SW 기업들 모두 '매출액 성장률 < 백로그/RPO 성장률' 기록

한계로 지목되는 것은 보안 이슈, 데이터 정렬, 인력 부족, 기업 문화적인 한계

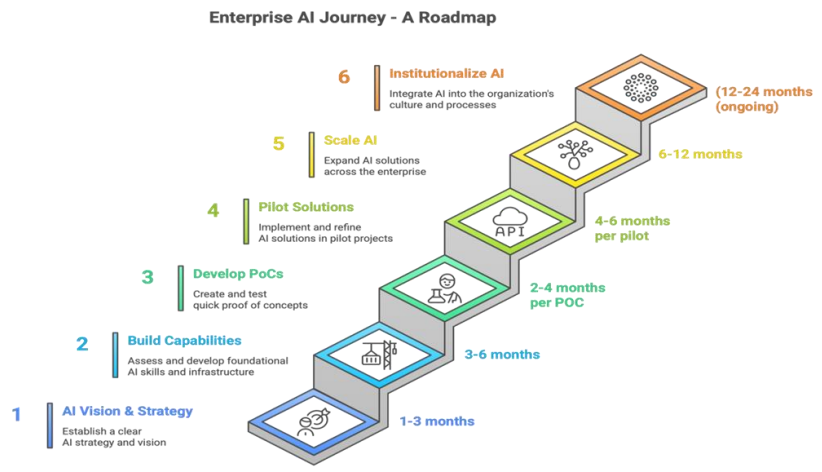
한계 보충을 위해 AI Labs도 Palantir-like 모델 도입 중, 직접 FDE 팀 운영 후 출장 및 컨설팅 진행

Coding Agent가 터주는 물꼬, 기업 워크로드에 AI 본격 도입 겨냥

엔비디아도 기업 AI 도입을 위한 여러가지 SW 서비스 출시

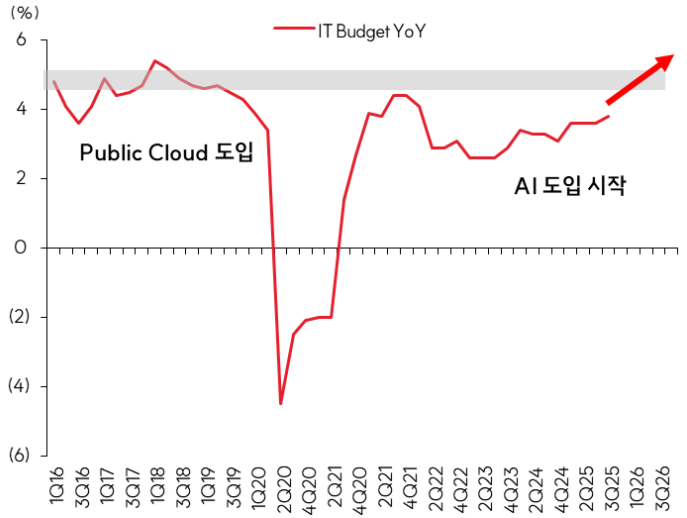
오케스트레이션 편의성을 높여주는 NemoClaw, 기존 데이터 정렬을 도와줄 cuDF, 추가 데이터 확보를 위한 cuVS

Enterprise AI 도입 과정 도식



자료 : EpochAI, SK증권

미국-유럽 Enterprise IT 예산 성장률 추이



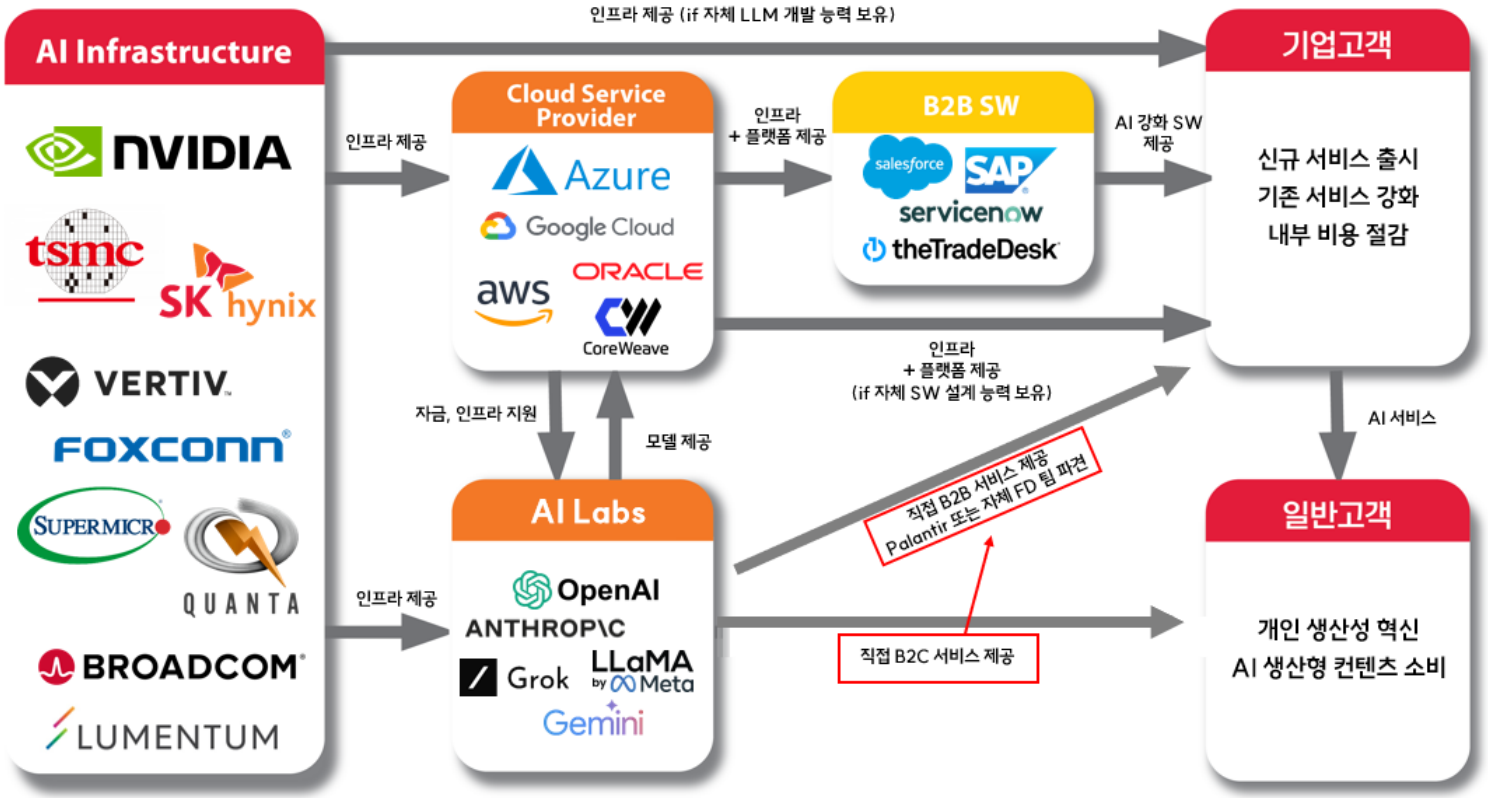
자료 : MorganStanely, SK증권

본격적으로 기업 수요를 유도할 차례

밸류체인 상 큰 수요의 축 변화

현재 인프라 수요의 대부분을 소비하는 AI Labs의 전방이 B2C에서 B2B로 변화

GPU DC 밸류체인: B2B 서비스로 이동



자료 : SK증권

New SW 점검 1: NemoClaw

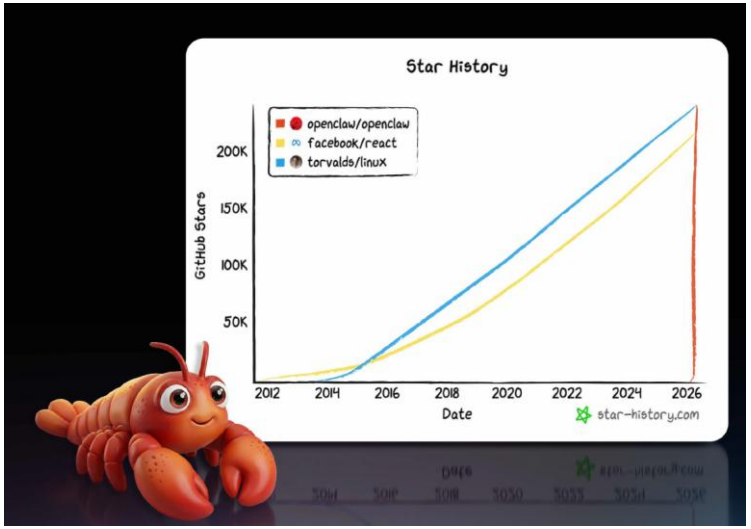
최단기간 최대 Star 달성 서비스, OpenClaw

OpenClaw는 메신저를 통해 Agent, Computer use를 모두 가능하게 해주는 오케스트레이션 레이어 서비스
2025년 11월 출시 이후 3개월만에 깃허브 스타 10만 달성, 주간 방문자 200만명
중국 Alibaba Cloud, Tencent Cloud, Baidu가 OpenClaw 원격 호스팅 서비스 출시

OpenClaw의 오마주이자 개선 서비스 출시

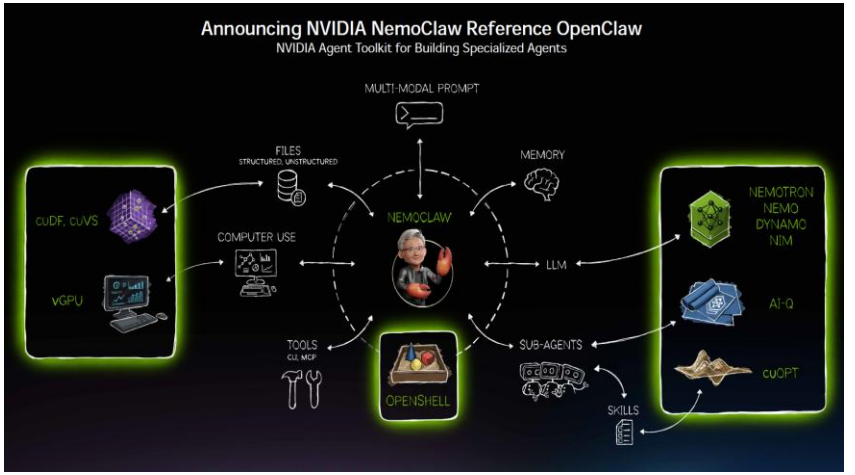
OpenClaw는 데이터와 툴을 광범위하게 건드리는 실행 시스템인만큼 보안 이슈 취약. 광범위한 유연성이 장점이자 단점
Nemoclaw는 OpenClaw의 보안 버전, 로컬 실행 + 오픈 모델 + Privacy Router 사용 시 클라우드 라우팅 가능

역사적으로 가장 빠른 Github Star(관심 추가) 기록



자료 : Nvidia, SK증권

기업용 OpenClaw, NemoClaw 공개



자료 : Nvidia, SK증권

New SW 점검 2: cuVS, cuDF

Agent는 더 많은 데이터가 필요, 접근을 용이하게 해준다

Agent 도입을 위해선 기업 데이터 조회, 검색, 분류, 벡터 검색이 용이해야. 관련 사항을 가속하기 위한 CUDA 라이브러리

기존 CPU 밸류체인에서 DBMS 기업들이 하던 데이터 엔진 가속 계층을 보완하는 형태

cuDF는 정형 데이터 가속, cuVS는 비정형 데이터 검색

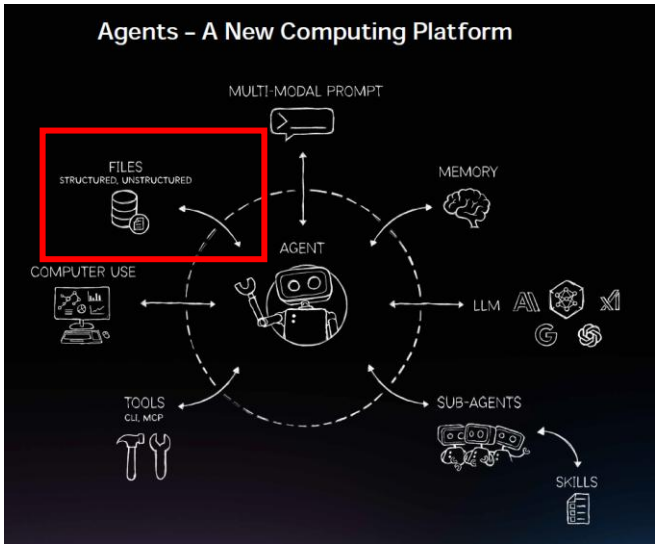
cuDF는 현재 대부분의 DB가 저장되는 정형 데이터(대표적으로 SQL)를 빠르게 가속화

cuVS는 현재는 데이터화가 잘 되지 않는 비정형 데이터의 저장 및 검색을 GPU로 가속화

예시) Snap은 사용자 1B+ 플랫폼, CPU 기반 사용자 추천 알고리즘을 GPU 기반 cuDF로 전환 후 비용 76% 절감

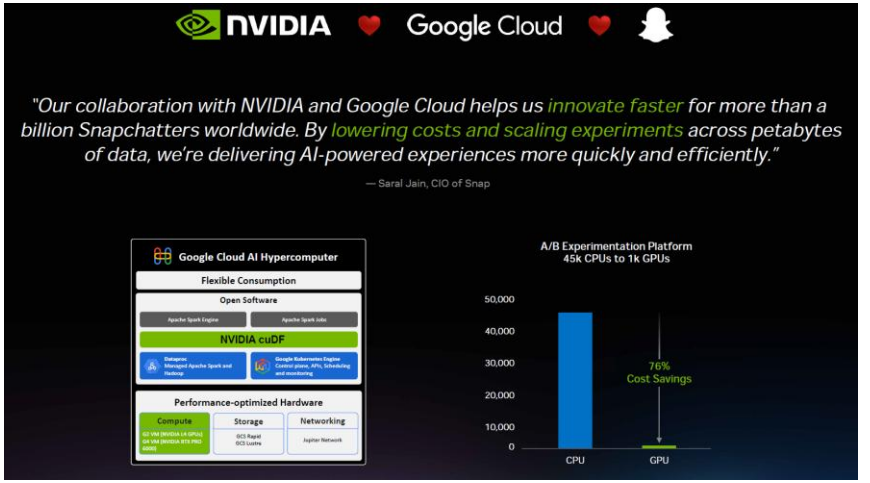
예시) 카드사 상담에서 고객 음성 내용을 벡터로 변환 인식, 관련 규정을 cuVS를 사용하여 의미 기반으로 빠르게 검색

Agent에서 DB 처리를 도와주는 CUDA 라이브러리



자료 : Nvidia, SK증권

OpenAI, Amazon 파트너십과 New DC



자료 : Nvidia, SK증권

New SW 점검 3: Physical AI Data Factory

Physical AI TAM은 $50~70T (vs SW $2T)

VLM은 LLM과 원리적으로 상이한 계산을 요구, 더 큰 챗봇이 아닌 별도 인프라 필요

수집 데이터로 Pre-train, 이후 Isaac Lab, Cosmos 모델 활용하여 합성 데이터 생성 하여 Post-train

최종적으로 로봇은 GROOT, 자율주행차는 Alpamayo 모델

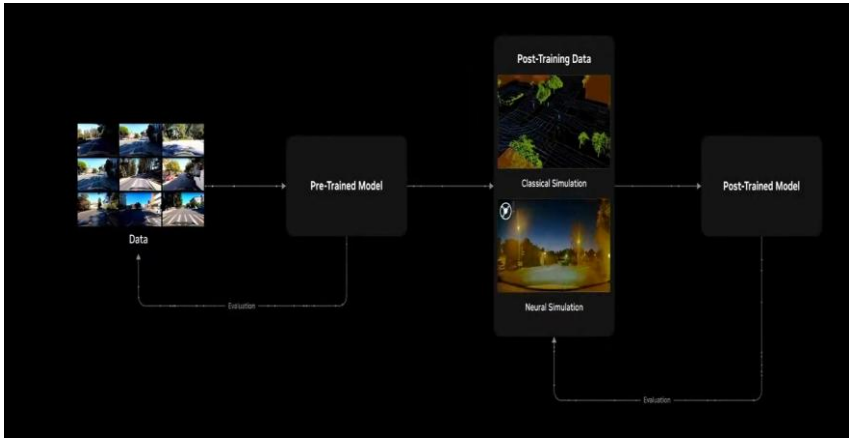
Data factory: 합성 데이터 표준화 방식 정립

로봇, 자율주행은 가장 큰 병목이 데이터. Data factory는 시뮬레이션·합성데이터·post-training 파이프라인을 표준화

실제 데이터가 더 빠르게 합성 데이터의 도움을 받아 모델을 만들 수 있게 도움

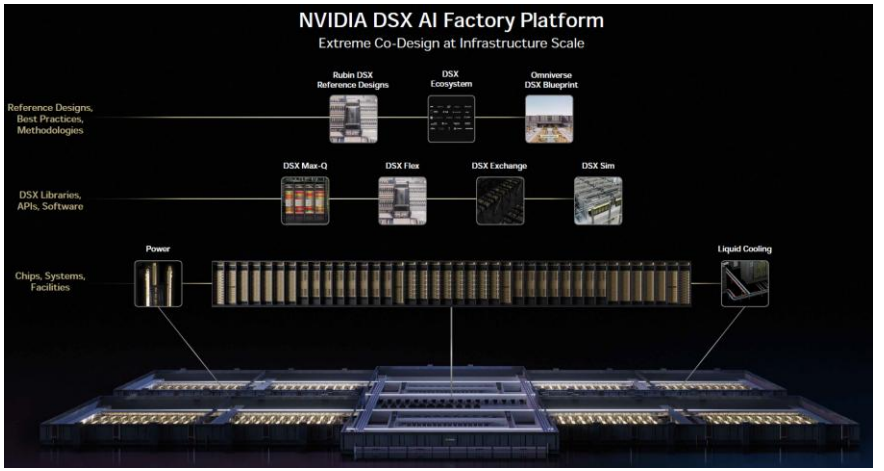
실제로 엔비디아의 extreme co-design에 Omniverse를 사용하여 제조하여 복잡도 극복

Physical AI Data Loop: 합성 데이터 필요



자료 : Nvidia, SK증권

DSX AI Factory Platform



자료 : Nvidia, SK증권

우주 DC, AI RAN이 같은 컨셉인 이유

현재 AI DC의 큰 공급 제약은 '전력 부족'

중앙 전력 확보 방식은 규제 대상, 온사이트 방식은 GEV의 긴 수주 잔고에서 쇼티지

전통적 방식 외의 아이디어들이 제시되는 중, 최근 각광 받는 아이디어가 우주 데이터센터와 AI RAN

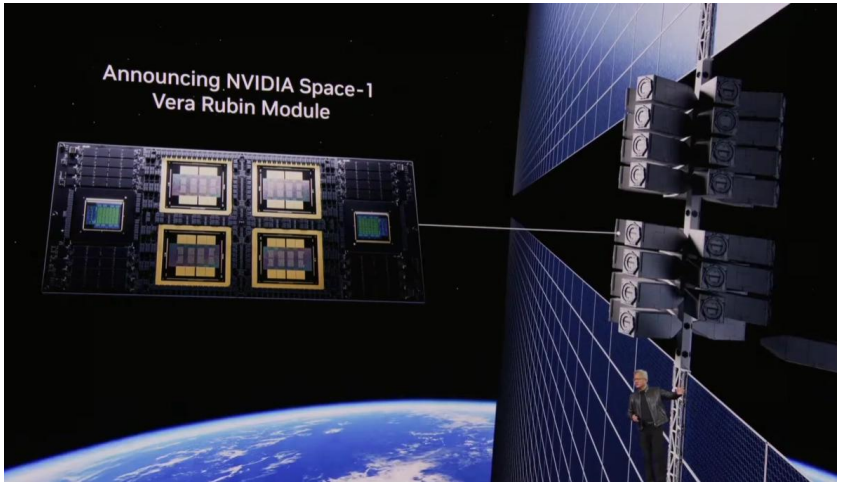
GTC에서 NVIDIA Space-1 Vera Rubin Module 공개, 차폐칩으로 추정. 열 관리 문제는 'dealing with it'으로 언급

AI RAN은 기지국용 RAN 인프라를 AI 가속 인프라로 사용, AI 워크로드를 돌리는 컨셉, 특히 edge AI 에 활용 전망

NVIDIA AI Aerial: 통신사들이 RAN(무선 접속망)과 AI workload를 같은 GPU 인프라에서 돌리게 해주는 플랫폼 공개

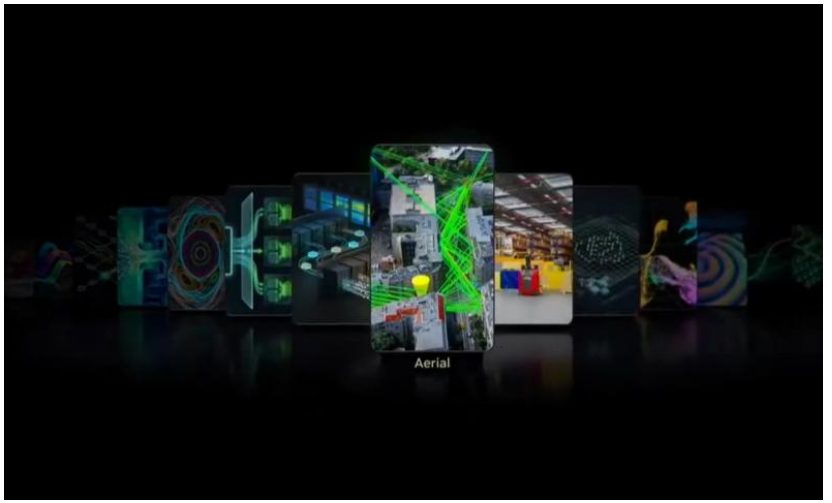
추가로 노키아가 RTX PRO 4500 Blackwell을 AI-RAN base station에 배치할 것이라고 언급

NVIDIA Space-1 Vera Rubin Module



자료 : Nvidia, SK증권

NVIDIA AI Aerial



자료 : Nvidia, SK증권

Chapter 4

NVDA: 성장도 가치도 BUY

지속 부진했던 엔비디아 주가

쇼티지 찾기에서 외면받는 중

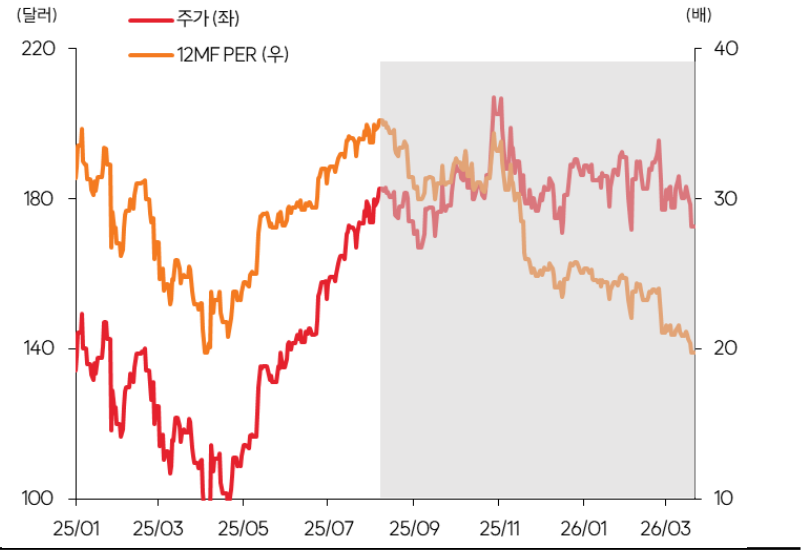
엔비디아 주가는 호실적 및 추정치 상향에도 2H25부터 주가 횡보

동 기간 나스닥 부진, 빅테크 12MF 20배 초반으로 수렴하면서 함께 하락

시장은 강한 수요에 대한 투자를 AI Infra 내 쇼티지 종목(메모리, 전력, 광)들을 통해 구축

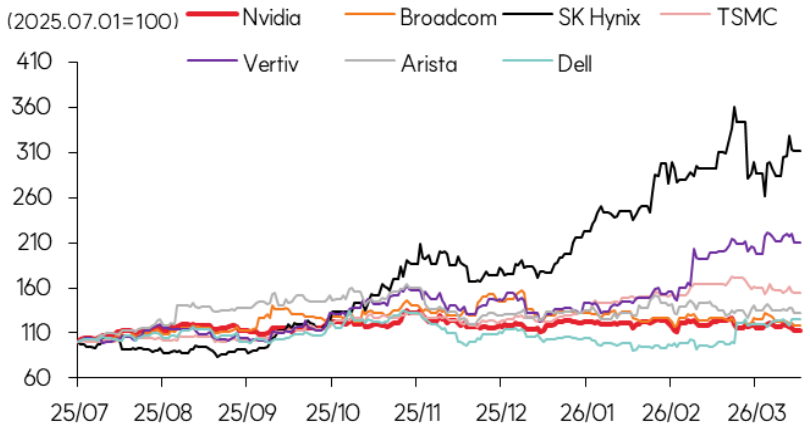
~CY2025 기간 대비 성장성 둔화 구간이라는 점에서도 Capex 베타가 큰 다른 종목 대비 열위

엔비디아 주가, 12MF PER 추이



자료 : Bloomberg, SK증권

주요 AI Infra 기업 주가 흐름



자료 : Bloomberg, SK증권

서프라이즈 가이던스와 주주환원까지

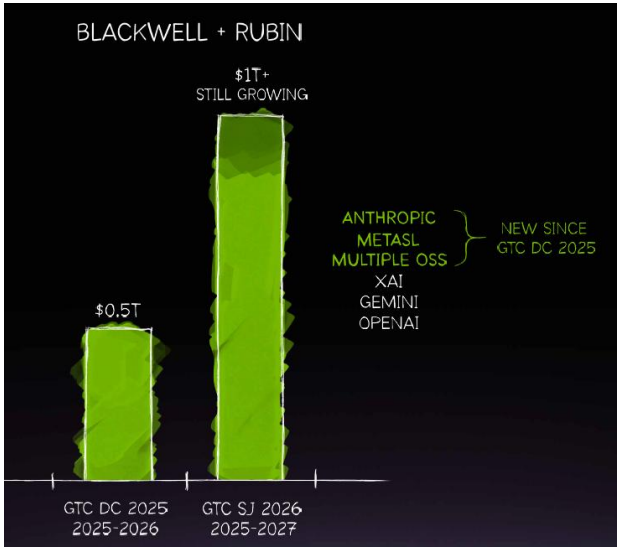
$1T 매출 가시성의 함의

CY2027까지 Blackwell+Rubin의 수주가 GTC를 기준으로 $1T 이상이라고 언급 (vs 지난 11월 ~CY2026 $0.5T)
CY2025 DC Compute 매출의 90%가 Blackwell로 추정되며 이는 합산 $174B 수준. 현재 수주잔고 $800B 이상으로 추정
QnA 세션을 통해 해당 수주잔고는 신제품(LPX, STX 등)과 내년 출시 예정 Rubin Ultra를 불포함한 수치로 확인
LPX upside를 25%, CPU upside를 5%, storage는 CPU보다 높은 수준으로 언급. 추가적인 사업 기회는 30% 상회 전망

CY2027 주주환원률 3% 상회 가능?

젠슨황은 향후 FCF의 50%를 주주환원에 사용할 예정으로 언급 (FY2023~ 매출액 대비 FCF 비중은 44~50%)
가이던스 기반 CY2027 매출 $600B(vs 컨센 $480B) 추정 시 주주환원 수익률 현재 시총($4.2T) 기준 3.5% 이상

~CY2027 수주 가시성 확보



자료 : Nvidia, SK증권

빅테크 주주환원 전망치 비교

기업	FY 2025 주주환원액	시가총액 (3/20 종가)	주주환원수익률
엔비디아	$41.4B	$4,126B	1.1%
애플	$104.7B	$4,049B	2.6%
마이크로소프트	$37.7B	$3,644B	1.0%
알파벳	$55.4B	$3,629B	1.5%
메타	$31.7B	$1,843B	1.7%
아마존	$0.0B	$2,201B	0%
브로드컴	$13.6B	$1,357B	1.%
테슬라	$0.0B	$1,434B	0%

자료 : Bloomberg, SK증권

신제품으로 큰 해자(Moat) 구축?

경쟁사: 이론적 구현은 가능하나 단기간에 따라오긴 어렵다

Superpod로 구축되는 주요 해자:1) LPU 칩 설계 2) Dynamo 3) Scale up, out

1) LPU 칩 설계 난이도가 아주 높다고 보긴 어려우나 Dynamo와 같은 SW에 맞춘 공동 설계에는 시간이 추가로 필요

2) Dynamo는 컨셉은 오픈소스이나 CUDA 툴킷을 전제, 비-Nvidia 가속기에 그대로 활용하기 어려움

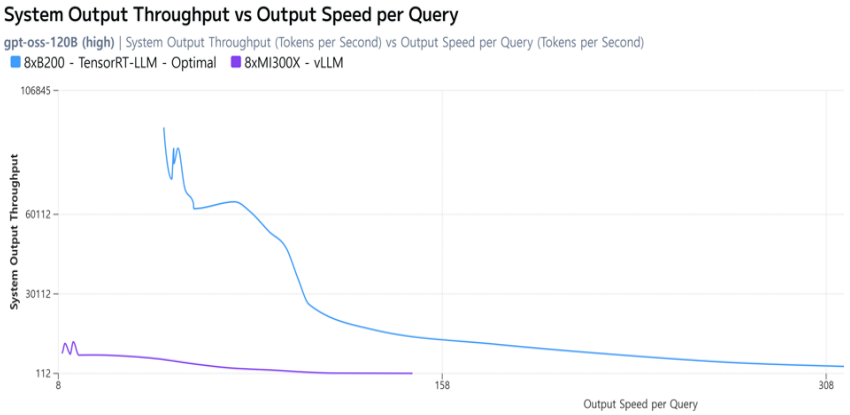
3) Scale out 증가, LPU C2C, 다수 CPU - GPU 등 신규 인프라와 GPU 연결구조를 기성품-ASIC 조합으로 대체하기 어려움

AMD는 올해 처음 Rack scale Infra 구축 시도, 컨셉적으로 뒤처지는 것에 더해 메모리 가격 폭등의 피해도 클 것으로 전망

CSP 역시 내부 워크로드를 제외한 대체는 당분간 쉽지 않을 것으로 전망, 기업 수요 증가가 이어진다면 엔비디아 해자 지속

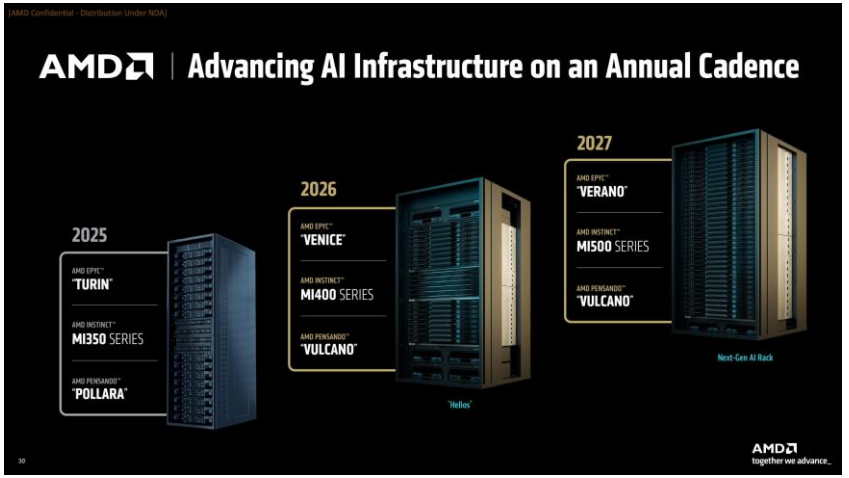
엔비디아 해자 지속은 결국 높은 GPM 지속과 연결되어 앞서 주장했던 '높은 주주환원률'로 이어져 주가 부양 가능

AMD는 8X 시스템에서 이미 차이가 큰 상황



자료 : Artificial Analysis, SK증권

AMD Helios Rack 하반기 출시 예정



자료 : AMD, SK증권

전방 비중에서 보이는 우려

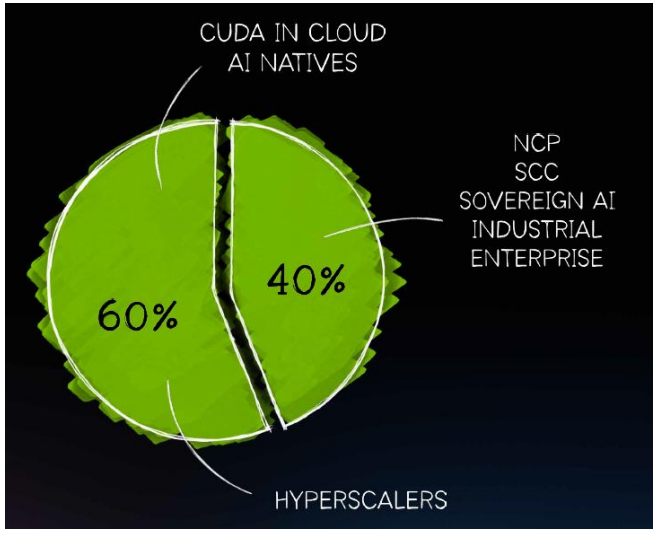
전방 비중: 60% 하이퍼스케일러, 40% 나머지

명시되는 매출에서 하이퍼스케일러의 비중이 60%, 10-K 기준 최상위 고객 매출 비중 22% 차상위 14%로 매우 집중
Microsoft가 1위 고객사, Oracle 또는 AWS가 2위 고객사로 판단. 모두 최종 고객사는 OpenAI 비중이 큰 상황
40%의 나머지 매출에서도 Tier 2 Server 업체의 비중이 매우 클 것으로 추정(Coreweave, Nebius 등), 이들도 AI Labs 지원

죽지 않는 OpenAI 망령

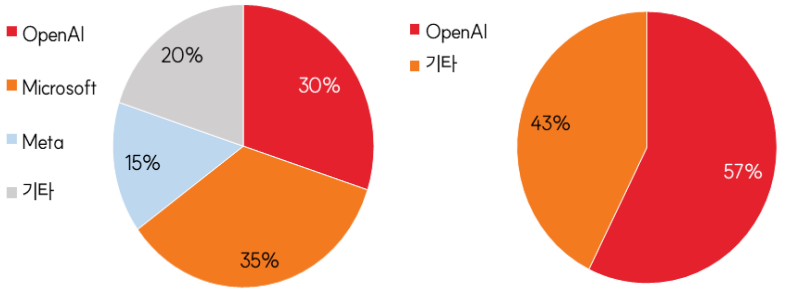
최근 OpenAI에 $30B 규모 투자, OpenAI는 가장 히트 상품인 ChatGPT의 수익화가 원활하게 이루어지지 않는 중
그러나 B2C 챗봇 제품은 이미 추론 비용이 매우 낮은 상태를 달성했을 것으로 판단, B2B 향 Coding Agent에 전사가 힘을 쏟는 중

엔비디아 전방 매출 비중



자료 : Nvidia, SK증권

Coreweave(좌), Oracle (우) 수주잔고 내 OpenAI 비중



자료 : 각 사, SK증권 / 주: Microsoft 대여 DC 70% 이상 OpenAI 사용 추정

NVDA: 성장도 가치도 BUY

ASP, 믹스 업그레이드

Blackwell 대비 Rubin의 ASP는 30% 이상 높을 것으로 추정

NVL72 Rack 도입으로 네트워크 사업부 매출액 급증, SuperPod 컨셉 역시 비GPU 매출액을 크게 끌어올릴 것으로 판단

$1T 수주 잔고에서 Rubin Ultra 매출액이 붙는 CY2027말까지 컨센서스 지속 상회할 가능성이 높음

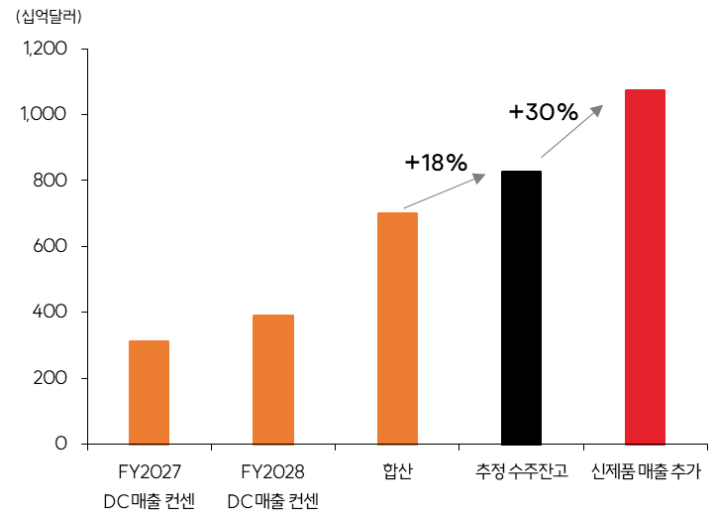
주주환원은 팔 이유를 없애고 신규 가치투자자를 끌어들일 것

이번 GTC에서 주식적으로 가장 함의가 컸던 코멘트는 CFO의 현금 활용 사안, 'FCF 50% 주주환원 고려 중'

현재 PER은 Capex 우려가 큰 CSP와 유사, 독점적 인프라 해자를 놓고 볼 때 바닥권

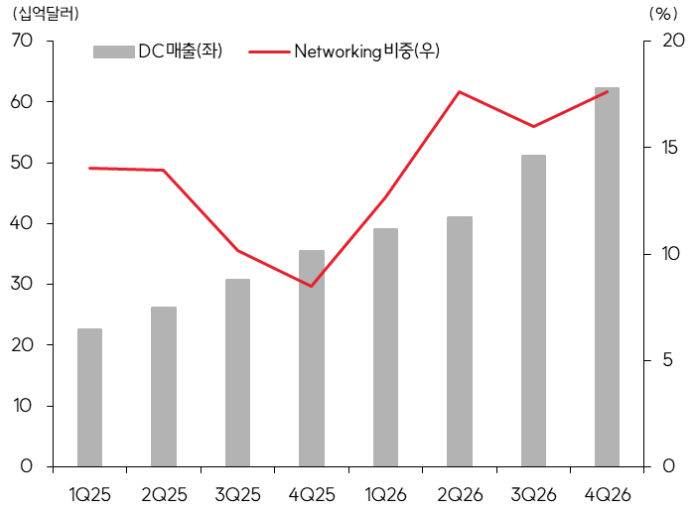
강한 해자와 수요로 Upside를 바라보면서 주주환원이 주가 바닥을 지켜줄 수 있는 구간으로 판단, Buy의견 제시

컨센서스 대비 $1T 가이던스 차이



자료 : SK증권 추정

네트워크 사업부 비중

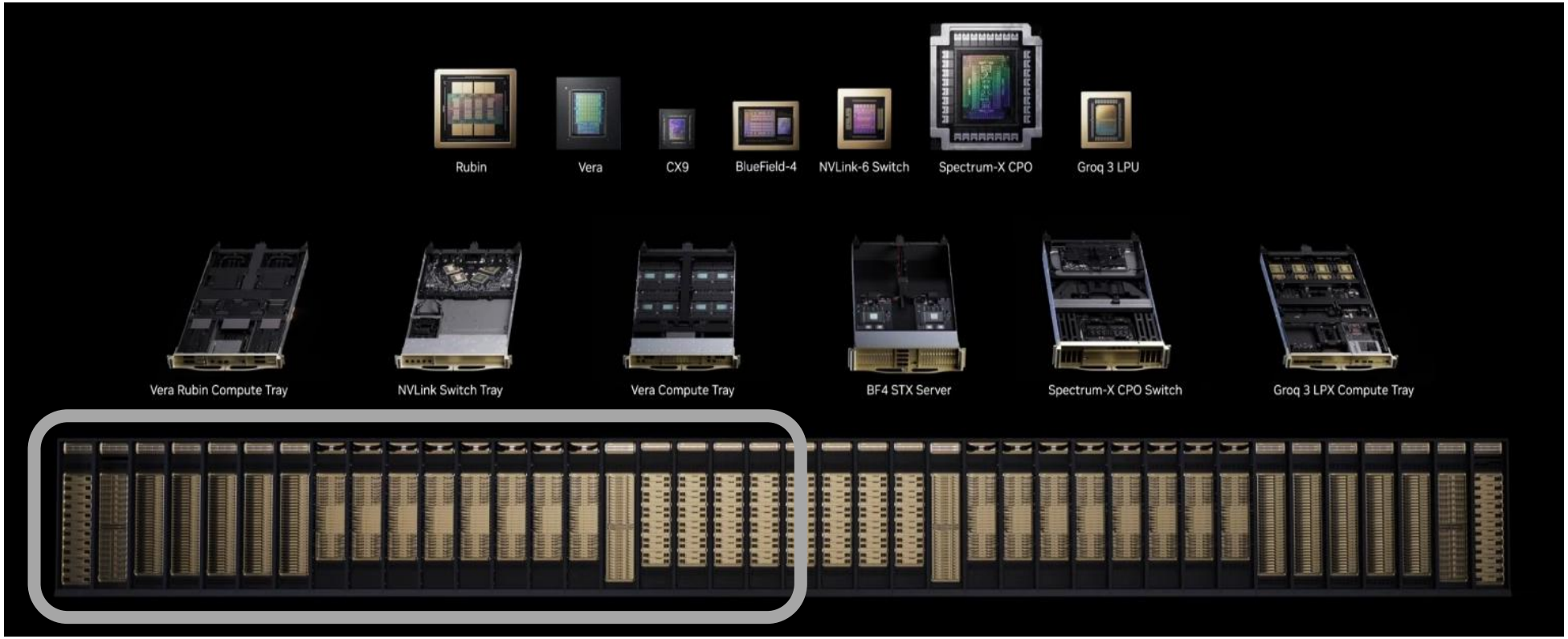


자료 : Bloomberg, SK증권

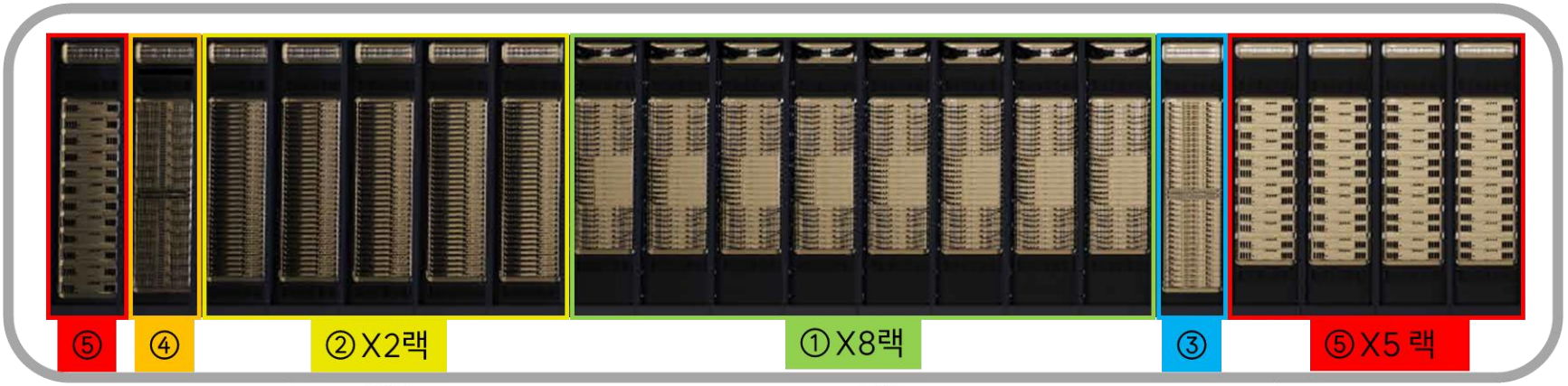
Appendix

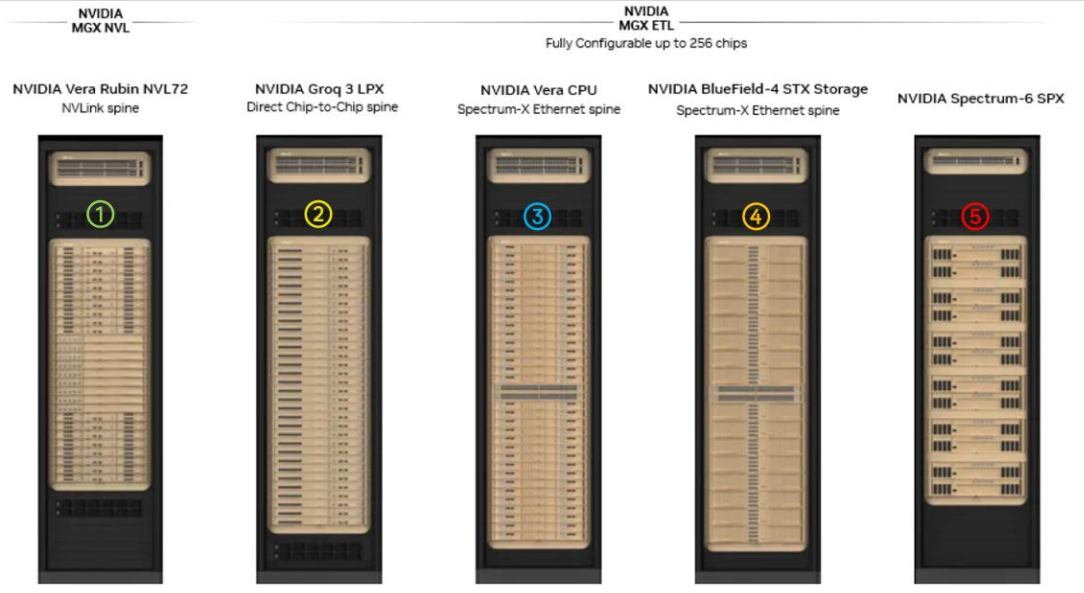
Vera Rubin 제품 구성

Vera Rubin Superpod Blueprint: 좌우대칭



1/2 SuperPod 내 랙 구성



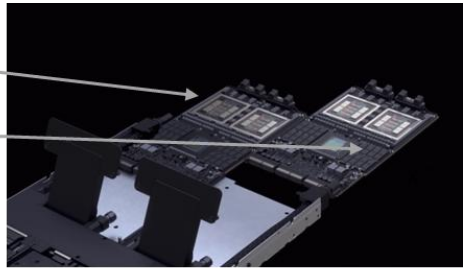


VR NVL72 Rack 구성

X 18

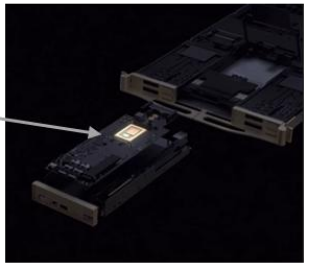
X 4

X 2

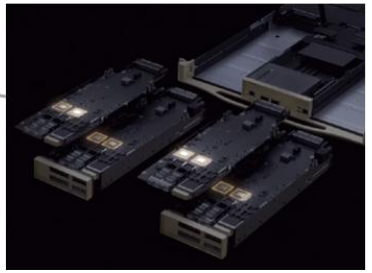


X 1

X 1



X 8



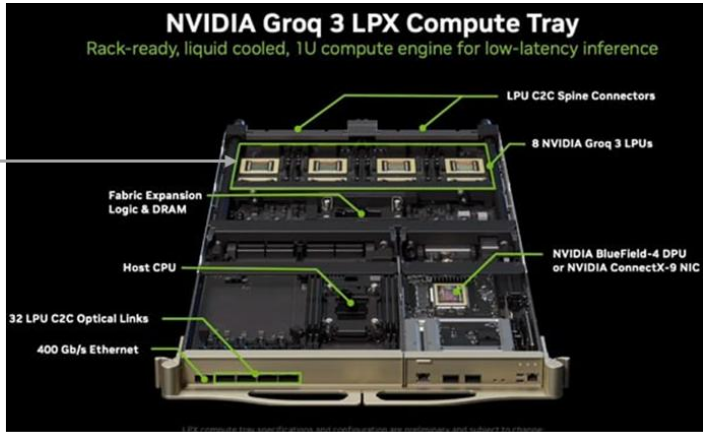
X 9

X 4

Groq 3 LPX Rack 구성

X~32

X 8



X 1

X 1

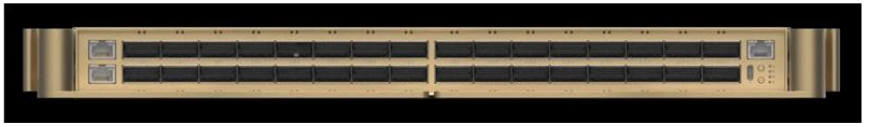
Vera CPU Rack 구성

X~32

X 8

X 2

X 1



1U MGX ETL switch tray

X 1

STX Strorage Rack 구성

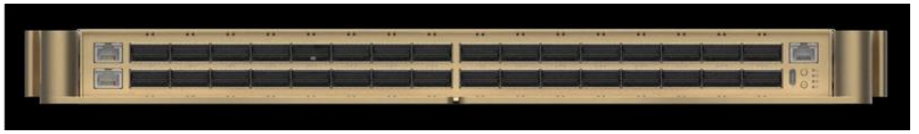
NVIDIA BlueField-4 STX Storage
Spectrum-X Ethernet spine

X n

X n

X n

X 1



1U MGX ETL switch tray

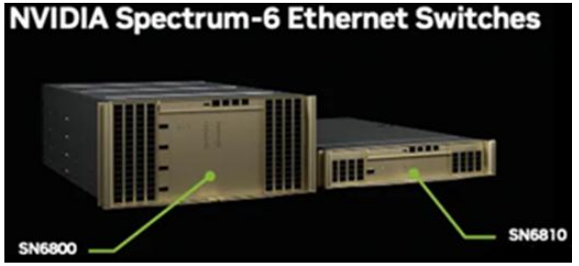
X 1

Spectrum SPX Rack 구성

NVIDIA Spectrum-6 SPX

X n

X 1 (SN6810기준)
X 4 (SN6800기준)



Compliance Notice

작성자(박제민)는 본 조사분석자료에 게재된 내용들이 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 신의성실하게 작성되었음을 확인합니다.

본 보고서에 언급된 종목의 경우 당사 조사분석담당자는 본인의 담당종목을 보유하고 있지 않습니다.

본 보고서는 기관투자가 또는 제 3자에게 사전 제공된 사실이 없습니다.

당사는 자료공표일 현재 해당기업과 관련하여 특별한 이해 관계가 없습니다.