미국은 지금

GTC Taipei 2026: 토큰 is Money

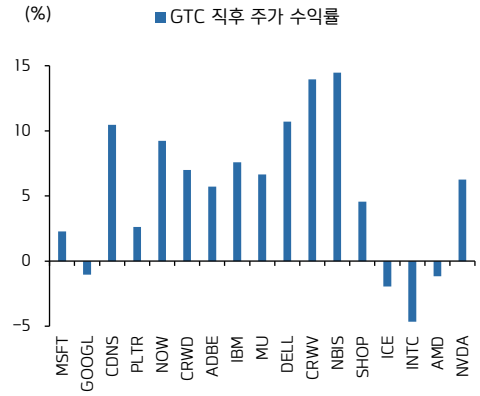
키움증권 리서치센터 글로벌리서치팀
US Strategy 김승혁

Issue Brief

GTC 타이페이 2026 개최

6 월 1 일 개막한 GTC 타이페이(Taipei) 2026 에서 젠슨 황은 AI 의 역할이 '답변 생성'에서 '업무 수행'으로 바뀌고 있음을 강조했다. AI 가 답변 봇을 넘어 향후에는 상황을 파악하고, 계획을 세우고, 필요한 도구를 사용해 실제 일을 처리하는 에이전트(Agent)가 된다는 의미다. 젠슨 황은 이 에이전트를 모델(LLM), 조율 시스템, 도구, 실행 환경이 결합된 새로운 컴퓨팅 단위로 제시했다. 또한 이 구조가 클라우드뿐 아니라 기업 내부 서버, PC, 로봇까지 반복적으로 적용될 것이라 주장했다. AI 가 더 많은 일을 할수록 더 많은 컴퓨팅이 필요해지고, 그 컴퓨팅 사용량이 곧 매출로 연결될 것이란 주장의 배경이다. 나아가 토큰이 결국 매출이 될 것이라 주장했다. GitHub 커밋이 인력 보강이 없었음에도 2026 년 들어 3 배 증가한 만큼 AI 가 실제 생산성을 높이고 있으며, 이러한 생산성은 결국 수익으로 귀결될 것이란 점이다.

키노트에서 언급된 이후 종목 별 수익률

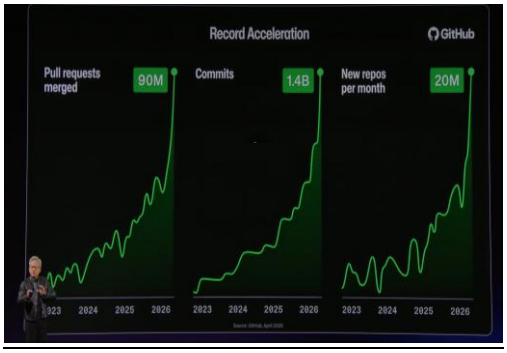


자료: Bloomberg, 키움증권 리서치

AI 팩토리 설계를 시작한 엔비디아

NVIDIA 는 키노트 발표에서 개별 칩(GPU)을 넘어 데이터센터 한 동 전체를 에이전트용 AI 팩토리로 설계하고자 하는 청사진을 제시했다. Vera Rubin 이 GPU·CPU·스토리지·네트워크를 하나로 통합한 AI 시스템이며, DSX 는 이를 실제 AI 공장 형태로 구축하기 위한 설계도다. 이에 에이전트의 빠른 응답을 지원하는 Vera CPU, 오픈 모델 Nemotron, 보안 솔루션 OpenShell 까지 더해지며 모델부터 컴퓨트, 네트워크, 전력, 냉각까지 AI 인프라 전 계층이 하나의 NVIDIA 플랫폼으로 연결된다. NVIDIA 가 지향하는 AI 팩토리 구조다. 이는 토큰 생성 시간, 와트당 처리량, 자산 수명 등에서 경쟁력을 갖는다.

Useful AI 의 사례

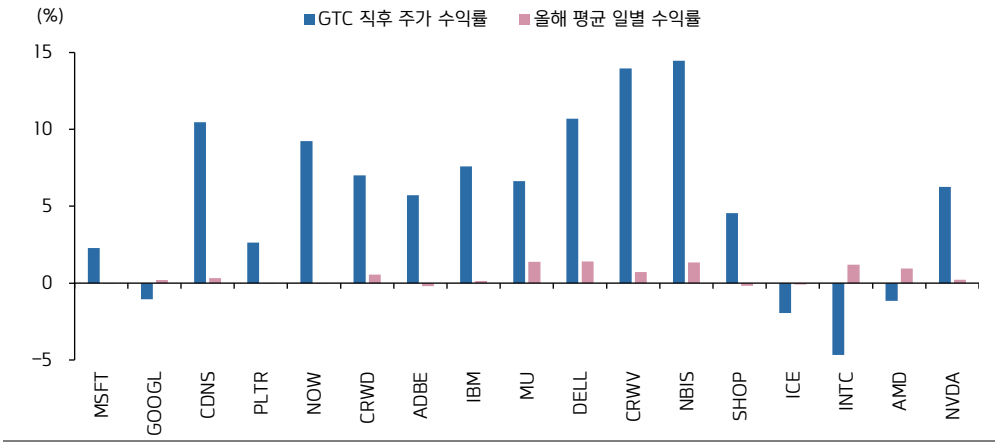


자료: 엔비디아

CPU, AI PC, 피지컬 AI 로의 영역 확장

NVIDIA 는 GTC 타이페이에서 CPU, AI PC, 피지컬 AI 등의 사업 영역 확장 역시 선언했다. Vera CPU 가 데이터센터 CPU 시장 진출을, RTX Spark 는 AI PC 시장 확대를, Cosmos 와 로보틱스 플랫폼은 자율주행·로봇 시장 진입을 의미한다. 관련 시장 규모도 커지고 있다. NVIDIA 는 AI 공장 1GW 를 구축하는 데 800 억~1,000 억 달러가 필요하며, 향후 10 년 내 전 세계적으로 100GW 규모가 구축될 것으로 전망했다. 이는 대규모 인프라 투자 사이클이 시작일 수 있다는 근거이며, NVIDIA 가 사업 영역을 확장한 이유다. 위 내용을 감안할 때 **수혜 산업은 파운드리, HBM 메모리, 대만 ODM 등 하드웨어 공급망 업체와 NVIDIA 핵심 파트너로 거론된 네오클라우드, Agent 맞춤 소프트웨어를 공급할 SaaS 기업 등이라 판단한다.**

GTC 타이페이 2026 에서 언급된 협력사 & 경쟁사 수익률 비교



자료: Bloomberg, 키움증권 리서치

엔비디아 GTC 협업 기업 및 내용 요약

구분	기업명	협업 내용	GTC 직후 수익률 (6/1 일간)	YTD 평균 일별 수익률
플랫폼·OS (PC, 클라우드)	Microsoft	Windows PC 재발명, OpenShell 채택, Vera Rubin 엔지니어링 랙	2.28	-0.03
	Alphabet (Google)	Nscale 클라우드 고객 (NVIDIA AI 클라우드 사용)	-1.04	0.20
반도체 설계·EDA	Cadence	칩 설계 슈퍼에이전트(ChipStack· Xcelium ·Jasper), 검증 40 배 가속	10.46	0.32
엔터프라이즈 SW·AI 에이전트	Palantir	데이터·AI 엔터프라이즈 에이전트	2.63	-0.04
	ServiceNow	엔터프라이즈 워크플로우 에이전트	9.24	-0.03
	CrowdStrike	보안 — 엔터프라이즈 에이전트	7.00	0.56
	Adobe	크리에이티브 SW(Photoshop·Premiere) RTX Spark 최적화·2 배 가속, MCP 연동	5.72	-0.20
	IBM (Red Hat)	OpenShell 채택(Red Hat), SQL 발명사로도 언급	7.60	0.13
메모리·시스템 하드웨어 공급망	Micron	HBM4 메모리 공급	6.64	1.38
	Dell	Vera Rubin NVL72 엔지니어링 랙	10.70	1.41
AI 클라우드 (네오클라우드)	CoreWeave	AI 클라우드, Vera Rubin 랙 가동	13.96	0.72
	Nebius	AI 클라우드	14.46	1.34
응용·고객사	Shopify	네오클라우드의 이커머스 AI 고객	4.56	-0.17
금융 인프라	Intercontinental Exchange	Vera CPU 실시간 스트림 처리 사례(NYSE, Lynn Martin 사장)	-1.95	-0.09
경쟁 구도	Intel	Vera CPU 를 "x86 대비" 비교(IPC·메모리지연 40%↓·1.8 배 등)	-4.67	1.20
	AMD		-1.16	0.95

자료: Bloomberg, 키움증권 리서치

GTC 타이페이 2026 핵심 메시지

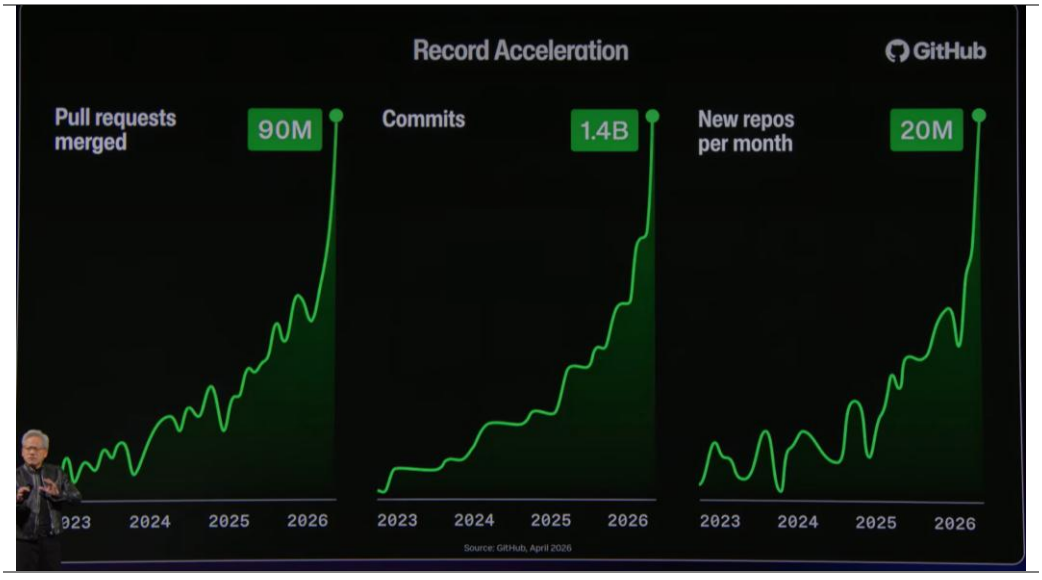
이번 키노트의 전체 메시지는 한 줄로 압축된다. **AI 가 단순히 답을 생성하는 '말하는 단계'를 넘어, 스스로 업무를 수행하는 '일하는 단계', 즉 에이전트(Agent) 시대로 진입했으며, NVIDIA 는 이에 맞춰 개별 칩이 아닌 데이터센터 전체를 새롭게 설계했다는 점이다.** 발표 전반을 관통하는 키워드는 세 가지로 정리된다. 첫째는 **에이전틱 AI(Agentic AI)로**, AI 가 관찰·추론·계획을 거쳐 도구를 사용해 작업을 끝까지 완수하는 방식을 말한다. 둘째는 **'컴퓨트는 곧 매출(Compute is revenue)'**이라는 명제로, AI 가 생성하는 토큰(token, AI 가 처리하는 텍스트의 최소 단위)이 수익 단위가 됐다는 인식이다. 셋째는 **Vera Rubin** 으로, 그 에이전트를 구동하기 위해 만든 차세대 시스템이며 단일 GPU 가 아니라 데이터센터 한 동을 구성하는 통합 시스템이라는 점이다.

토큰은 곧 매출이다

젠슨 황은 에이전틱 AI 가 실제로 도착했다는 선언으로 시작하며, 그 근거로 소프트웨어 개발 지표를 제시했다. 개발자들이 코드를 수정해 반영하는 행위인 커밋(commit)은 2023 년 3 억 건, 2024 년 4 억 건, 2025 년 5 억 건으로 늘어왔는데, 2026 년 들어 수개월 만에 약 3 배로 증가했다는 것이다. 전 세계 직업 개발자가 3,000 만~4,000 만 명, 이들의 연봉 총합이 약 3 조 달러인데, **동일 인원이 3 배 가까운 산출을 내고 있으므로 사실상 9 조 달러 규모의 생산성에 해당한다는 계산이다.** 그는 이를 근거로 **AI 가 일자리를 줄인다는 통념을 반박했다**. 한 명의 개발자가 더 큰 산출을 낼 수 있다면 고용을 줄일 이유가 없으며 오히려 더 늘린다는 논리다.

이 생산성 논리는 곧바로 산업 수요 논리로 연결된다. AI 가 쓸모 있어지면서 **토큰이 수익을 내는 단위가 됐고, 그 결과 AI 기업들은 토큰을 더 많이 생성하려 하며, 이를 위해 AI 팩토리(AI 공장, AI 데이터센터를 공장에 빗댄 표현)를 더 짓게 되고, 그 결과 대만을 중심으로 한 컴퓨팅 수요가 늘었다는 흐름이다.** 키노트 전반에서 '토큰=매출'이라는 등식이 반복되는데, 이는 AI 인프라 투자를 비용이 아닌 수익 자산으로 규정하려는 의도다. AI 자본지출의 지속성에 대한 시장의 의문에 대응하는 프레임이다.

AI 의 효용이 빠르게 상승하고 있다는 증거



자료: 엔비디아, 키움증권 리서치

에이전트(Agent)라는 새로운 컴퓨팅 구조

젠슨 황은 에이전트를 '작업장에서 도구를 가지고 일하는 노동자'에 비유했다. 모델이 노동자, 하네스(harness)가 몸, 도구(tools)가 연장, 런타임(runtime)이 작업장에 해당한다. 기존 컴퓨팅이 응용프로그램→코드→운영체제 위에서 돌아갔다면, 에이전트는 네 요소로 재구성된다. 생각·추론·계획을 담당하는 두뇌인 모델(LLM, 대규모 언어 모델), 모델을 둘러싸고 작업을 조율하는 운영체제격 소프트웨어인 하네스, 스프레드시트·브라우저·데이터베이스·컴파일러처럼 에이전트가 실제로 사용하는 도구와 스킬, 그리고 이 모든 것이 구동되는 실행 환경인 런타임이다. 하네스는 입력을 받아 이해·추론·계획·실행에 이르는 전 과정을 라우팅(경로 배분)하는 조율 장치로, 오케스트레이션(여러 작업을 지휘자처럼 배치하는 것)의 핵심이다.

이 구조에서 주목할 부분은 메모리다. 에이전트는 사람처럼 단기 기억과 장기 기억을 사용한다. 단기 기억은 기술적으로 KV 캐싱(KV caching)이라 불리며, LLM 이 직전까지 처리한 내용을 빠르게 재사용하기 위해 저장해 두는 임시 메모리다. 작업이 길어질수록 이 캐시가 커지고, 무엇을 보존하고 무엇을 압축하며 어떻게 다시 찾아올지가 복잡한 과제가 된다. 젠슨 황은 AI 의 메모리 시스템이 스토리지 산업을 재편할 것이라고 언급했는데, 이는 뒤에 등장하는 BlueField 기반 스토리지 제품의 복선이다.

요약하면 컴퓨팅 방식 자체가 바뀐다. 과거에는 앱을 켜고 클릭·입력했다면, 앞으로는 의도를 설명하면 AI 가 코드를 작성하거나 도구를 사용해 결과물을 낸다는 것이다.

SaaS 소멸론 반박과 CUDA-X

에이전트 시대에 소프트웨어 기업이 사라진다는 전망에 대해 젠슨 황은 정반대 입장을 밝혔다. **에이전트는 사람 수의 제약을 받지 않아 무한히 늘어나며, 그 많은 에이전트가 과거보다 더 많은 도구, 즉 소프트웨어를 사용하게 된다는 것이다.** 다만 조건이 있다. 소프트웨어가 에이전트가 호출할 수 있는 형태로 제공돼야 한다. 이 '도구 사용(tool use)'이 핵심 전환점이며, 자사 기능을 **에이전트가 부를 수 있는 도구로 노출하는 기업이 오히려 수요 증가의 수혜를 본다는 논리다.** 뒤에 등장하는 Cadence·SAP·ServiceNow·Palantir 와의 협업이 이 주장의 근거로 제시된다.

NVIDIA 자신도 같은 논리를 적용한다. 20 년 전 만든 CUDA(GPU 로 일반 연산을 수행하게 하는 가속 컴퓨팅 기반 기술이자 NVIDIA 생태계의 해자)와 그 위에 쌓인 1,000 개 이상의 분야별 라이브러리 CUDA-X 를, 이제 에이전트가 사람보다 효과적으로 사용할 수 있도록 정비했다는 것이다. 이 라이브러리들이 '에이전트가 호출하는 표준 도구'로 자리 잡으면, 경쟁사가 칩 가격을 낮추더라도 전환이 어려워지는 락인(lock-in)이 강화된다.

Vera Rubin 의 설계 배경

에이전트는 분산·이질 컴퓨팅 구조를 전제로 한다. 하나의 작업이 데이터센터 곳곳의 서로 다른 컴퓨터에서 나뉘어 처리된다는 의미다. 예컨대 LLM 이 생각·추론·계획을 수행할 때마다 Grace Blackwell NVL72 랙 한 대 전체가 가동된다. 이런 분산·이질·복합 구조 때문에 NVIDIA 는 차세대 시스템 Vera Rubin 을 설계했다. **젠슨 황은 Vera Rubin 이 칩 하나도, GPU 만도 아니며 끝에서 끝까지 전체가 하나의 시스템이라는 점을 분명히 했다.** 구성은 추론을 담당하는 GPU(Vera Rubin NVLink 72), 조율을 담당하는 Vera CPU, 저장을 담당하는 Vera BlueField, 그리고 네트워크·보안을 담당하는 CX9·DOCA·BlueField 보안 프로세서로 이뤄진다.

AI 팩토리에 대한 개념 정의와 DSX

AI 팩토리는 전기를 투입해 토큰을 산출하는 공장으로, 칩·랙·네트워크·전력·냉각·전력망을 하나의 단위로 설계해야 한다. NVIDIA 는 이 공장의 청사진으로 DSX 를 발표했다. 제품 라인은 GPU 가 RTX, 시스템(서버)이 DGX, 인프라 전체가 DSX 로 정리된다. DSX 의 구성 요소 가운데 DSX Sim 은 Omniverse 기반으로 공장을 디지털 트윈(현실을 그대로 본뜬 가상 모델)으로 먼저 짓고 전력·냉각·네트워크를 검증한다. DSX MaxLPS 는 AI 공장이 통상 전력을 최대 40%까지 과잉 확보한다는 점에 착안해, 같은 전력 예산 안에 더 많은 GPU 를 배치함으로써 연간 수십억 달러의 추가 매출을 끌어낸다. 45 도 고온 액체 냉각은 물과 에너지 소비를 줄여 더 많은 전력을 연산에 투입한다.

규모에 대한 전망도 제시됐다. **젠슨 황은 10 년 내에 100 기가와트(GW) 규모의 AI 공장이 가동될 것이라고 밝혔다. 1GW 공장 한 동의 건설비는 초기 200 억~300 억 달러에서 현재 500 억~600 억 달러, 향후 800 억~1,000 억 달러로 상승한다고 봤다.** 1,000 억 달러 규모 공장이 가동 즉시 정상 작동해야 하므로 사전 시뮬레이션이 중요하다는 논리로, Omniverse 의 역할이 여기에 연결된다. 투자 관점에서 1GW 당 자본지출이 3~5 배 상승한다는 수치는 AI 인프라 투자 규모의 가파른 확대를 시사하며, NVIDIA 의 사업 영역이 칩을 넘어 전력·냉각·전력망 등 산업재 영역으로 확장된다는 점도 함의가 크다.

네오클라우드 생태계와 인프라 평가 기준

NVIDIA 의 풀스택(전 계층 통합) 역량 덕분에 신생 기업도 글로벌 AI 클라우드로 성장할 수 있다는 점이 강조됐다. 사례로 기업가치가 급등한 CoreWeave, Nebius, Nscale(고객으로 영국 BT·Google), Together AI, Yotta, Indosat, GMI, 그리고 한국의 NAVER Cloud(고객으로 한국은행·현대), 프런티어 랩인 Thinking Machines 등이 거론됐다. 이들 클라우드의 고객으로는 Cursor(코딩), Black Forest Labs(이미지 생성), World Labs(월드 모델), Revolut(금융), Shopify 등이 언급됐다.

젠슨 황은 'NVIDIA 를 선택해야 하는 이유'를 인프라 평가의 네 가지 기준으로 설명했다. 첫째는 첫 토큰까지 걸리는 시간(time to first token)으로, 전 계층을 통합 설계·구축한 경험 덕분에 가동·추론·학습 개시가 빠르다는 점이다. 둘째는 와트당 처리량(tokens per watt)으로, 전력이 1GW 면 그 이상 늘릴 수 없으므로 와트당 토큰이 곧 매출이며, 칩 가격이 저렴하다는 이유로 잘못된 아키텍처를 선택해서는 안 된다는 논리다. 셋째는 신뢰성으로, 수백만 개의 케이블이 얽힌 데이터센터가 안정적으로 작동하기 어려운 만큼 평균 무중단 시간(MTBI)이 중요하며 대규모 운영 경험이 길수록 유리하다.

넷째는 자산 수명과 총소유비용(TCO)으로, AI 소프트웨어가 CNN 에서 트랜스포머, MoE, 에이전트로 계속 바뀌는 환경에서 유연한 아키텍처와 풍부한 생태계를 갖춰야 시스템을 오래 사용할 수 있고, 개발자들이 CUDA 에서 출발하는 만큼 NVIDIA 자산의 유효 수명이 길어 TCO 가 낮다는 주장이다.

본격 생산에 돌입한 Vera Rubin 과 밸류체인

젠슨 황의 첫 공식 발표는 Vera Rubin 의 풀 생산 진입이다. 공급망은 Grace Blackwell 의 2 배 규모로 확대됐고, **랙 한 대 조립 시간은 2 시간에서 5 분으로 단축됐으며**, 수백만 제곱피트 규모의 공장이 가동 중이다. 공개된 스펙을 보면 **TSMC 3 나노 공정과 CoWoS-R·CoWoS-L 패키징이 적용되며, HBM4 메모리는** Micron·SK 하이닉스·Samsung 이 공급한다. 컴퓨트 보드에는 트랜지스터 6 조 개와 부품 1 만 8 천 개 이상이 집적되고, 컴퓨트 트레이는 슈퍼칩·ConnectX-9 슈퍼 NIC·BlueField-4 DPU 가 **케이블 없이 결합되는 구조다.** Vera CPU 랙은 단일 액냉 랙에 CPU 256 개를 담는다. Foxconn·Quanta 가 생산하는 Groq 3 LPX 는 16 개 트레이에 Groq LPU 256 개를 담아 초당 40 페타바이트의 SRAM 대역폭으로 저지연 토큰 생성을 맡으며, **NVL72 가 최고 처리량을, LPX 가 최저 지연을 담당하는 역할 분담이 명확하다. 네트워크에서는 세계 최초의 200 기가비트 광학 일체형(co-packaged optics) 이더넷 스위치인 Spectrum-X 이더넷 포토닉스가 TSMC COUPE 공정으로 제작**된다. 결과적으로 Vera Rubin 은 5 개의 랙 규모 시스템이 연결된 에이전트용 슈퍼컴퓨터이며, 대만 협력사 150 곳이 참여했다. 칩부터 데이터센터까지 모든 계층을 처음부터 함께 설계하는 '극단적 코디자인(extreme co-design)'이 핵심 개념이다.

젠슨 황은 세대별 초점 변화도 짚었다. Hopper 는 사전학습(모델을 처음 대량 학습시키는 무거운 단계), Grace Blackwell 은 추론(학습된 모델을 사용해 답을 내는 단계)에 최적화됐는데, 복잡한 MoE 모델을 빠른 응답성과 높은 처리량으로 동시에 처리하기가 어려워 NVLink 72 를 만들었고 그 결과 토큰 생성 비용을 업계 최저 수준으로 낮췄다는 설명이다. **Vera Rubin 은 여기서 더 나아가 추론을 넘어 에이전트 시스템 구동에 맞춰 설계됐다.** 이 대목은 메모리·파운드리·후공정·부품으로 이어지는 밸류체인 전반의 수혜 구조를 드러낸다.

Vera CPU: 사람이 아닌 에이전트를 위한 CPU

이번 키노트에서 새로운 발표 중 하나가 Vera CPU 다. 젠슨 황의 논지는 지금까지의 CPU 가 모두 사람을 위해 설계됐지만 에이전트는 다르다는 것이다. **에이전트는 나노초 단위로 동작하므로 CPU 는 저지연·고응답성을 갖춰야 한다.** 시스템 내 역할은 세 가지다. Vera Rubin 랙 내부의 CPU 2 개가 GPU 와 KV 캐시 관리, 랙 소프트웨어 구동을 맡고, Grace BlueField 가 보안·격리를 담당하며, Vera BlueField 스토리지 서버가 빠른 저장 시스템 역할을 한다. 에이전트가 메모리에 고속으로 접근하는 만큼 스토리지 서버와 CPU 가 데이터센터에서 가장 비싸고 중요한 핵심 경로가 된다.

Vera CPU 의 특징은 네 가지로 요약된다. 클록당 명령어 수(IPC)가 업계 최고 수준으로, 클록 한 번에 10 개 명령어를 처리한다. 코어당 대역폭도 최고 수준이며, 전체 대역폭 측면에서는 분산 시스템 특성상 데이터 이동 속도가 관건이므로 모든 코어를 잇는 초당 3.6 테라바이트 패브릭과 초당 1.2 테라바이트의 LPDDR5X 메모리(기존 최고 CPU 의 2~3 배)를 갖췄다. 마지막으로 에너지 효율인데, 옆에 배치된 GPU 가 비싼 자원이므로 CPU 는 적은 전력으로 많이 집적돼야 한다.

세부적으로는 NVIDIA 자체 데이터센터 코어인 Olympus 를 사용하고, LPDDR5X 를 처음 채택하면서 대역폭 손실 없이 다중 오류를 정정한다. **x86 대비 메모리 지연은 최대 40% 낮고, 88 개 Olympus 코어를 칩렛으로 쪼개지 않고 단일 모놀리식 메시로 통합해 코어 간 통신이 50% 빠르며, NVLink 칩-투-칩으로 GPU 를 패브릭에 직접 연결한다.** 에이전틱 샌드박스 성능은 x86 대비 1.8 배로 제시됐다.

실제 성능 수치로는 **SQL 처리 3 배, 실시간 스트림 처리 6 배가 공개됐다.** 특히 스트림 처리는 뉴욕증권거래소(NYSE) 사례로, NYSE 의 Lynn Martin 사장과의 협업을 통해 증권거래소처럼 실시간 데이터가 끊임없이 유입되는 환경을 처리한다. 젠슨 황은 CPU 에서 배수 단위 성능 향상을 언급하는 것 자체가 이례적이라는 점을 강조하며, Vera 가 기존 시장을 잠식하는 것이 아니라 에이전트용 CPU 라는 새로운 시장을 연다고 설명했다. 투자 관점에서 이는 NVIDIA 가 GPU 에 이어 전통적으로 인텔·AMD 의 영역이던 데이터센터 CPU 시장에 본격 진입한다는 선언이다. 이미 Grace 를 수백만 개 판매해 최대 CPU 제조사 중 하나라는 주장과 함께, **Grace 에서 Vera 로의 전환을 회사 역사상 가장 빠른 출시가 될 것으로 전망했다.** x86 진영에 대한 경쟁 위협으로 해석할 수 있는 대목이다.

엔터프라이즈 AI 툴킷: 에이전트의 운영체제

젠슨 황은 모든 기업이 에이전트를 운영하게 되며, 이때 핵심 질문은 **'어떻게 안전하게 돌리고 어떻게 업무용 에이전트를 만드느냐'라고 정리했다. 그 답이 NVIDIA Enterprise AI 툴킷이다.** 기업이 에이전트를 구축하려면 네 요소가 필요하다. 똑똑하고 저렴하고 빠른 모델(NVIDIA 의 오픈 모델 Nemotron), 전체를 조율하는 하네스(NVIDIA OpenShell), 도구와 스킬(CUDA-X 라이브러리 등), 그리고 모든 것을 묶는 런타임이다. OpenShell 은 에이전트를 보안 정책에 묶어 보호하는 안전 셸 역할을 한다.

대표 사례는 Cadence 와의 칩 설계 슈퍼 에이전트다. 반도체 설계 자동화(EDA) 기업 Cadence 와 함께 시뮬레이션(Xcelium)과 형식 검증(Jasper)으로 설계 결함과 버그를 찾아 수정하는 에이전트를 구축했고, **검증 주기를 40 배 이상 단축해 수 주가 걸리던 작업을 수 시간으로 줄였다.** 모델 측면에서는 차세대 오픈 모델 Nemotron 3 Ultra 가 공개됐다. 모델·데이터·학습법을 모두 공개하는 완전 개방형이고, SSM(상태공간모델)과 MoE 를 결합한 하이브리드 구조이며, **5 배 빠른 속도와 함께 총 연산량·추론 시간 기준 세계 최고 수준 대비 30% 낮은 비용을 제시했다. 후속작 Nemotron-4 도 개발 중이다.** 협업 기업으로는 Cadence·CrowdStrike·Dassault·Palantir·SAP·ServiceNow 가 거론됐다. 투자 관점에서 이는 NVIDIA 가 모델·하네스·런타임을 오픈소스로 제공해 에이전트 생태계의 표준을 확보하려는 전략으로, 엔터프라이즈 소프트웨어 기업을 경쟁자가 아닌 파트너로 끌어들이는 구도다. EDA 업종(Cadence·Synopsys)은 에이전트로 설계 생산성이 높아지는 수혜 영역으로 볼 수 있다.

PC 의 재발명

젠슨 황은 Microsoft 와 NVIDIA 가 40 년 만에 PC 를 재설계한다고 발표했다. **새 PC 는 기존 운영체제에 LLM 이 결합된 형태의 운영체제와, 응용프로그램을 대체하는 에이전트(에이전틱 런타임)로 구성된다.** 구체적 제품은 RTX Spark 로, Blackwell RTX GPU(CUDA 코어 6,144 개, AI 성능 1 페타플롭), MediaTek 와 공동 개발한 20 코어 Grace CPU 를 NVLink 로 결합하고, 128GB 통합 메모리와 TSMC 3 나노 공정, 트랜지스터 700 억 개를 갖췄다.

시연에서는 RTX Spark 에서 클라우드 없이 기기 내부에서 구동되는 에이전트가 집을 설계하는 과정을 보여줬다. Adobe 는 Photoshop 과 Premiere 의 핵심 구조를 재설계해 RTX Spark 용으로 출시한다. **젠슨 황은 향후 가정마다 홈시어터처럼 AI 슈퍼컴퓨터가 한 대씩 자리 잡아 에이전트와 비서를 구동하고, PC 가 점차 영화 속 로봇 같은 존재로 변할 것이라고 전망했다.** 투자 관점에서 이는 NVIDIA 가 MediaTek 와 함께 ARM 기반 Windows PC 용 SoC(N1X) 시장에 진입해 인텔·AMD·퀄컴과 경쟁 구도를 형성하는 것이며, AI PC 사이클 본격화와 소프트웨어의 에이전트 친화적 재설계, 기기 내 추론 수요 확대를 시사한다.

Physical AI 와 현실 세계로의 확장

마지막 축은 현실 세계에서 동작하는 AI 다. 여기서 가장 큰 난관은 데이터다. 로봇 AI 는 1 인칭 관점의 데이터가 필요한데 세상의 영상 대부분은 3 인칭 시점이기 때문이다. 해결책으로 Omniverse 기반 시뮬레이션이 제시되며, 이는 검증가능 보상 기반 강화학습(RLVR)에 해당한다. 이 흐름에서 Physical AI 를 위한 오픈 모델 Cosmos 3 가 공개됐다. 젠슨 황은 언어 모델은 경쟁자가 많지만 Physical AI 에서는 NVIDIA 가 선두라고 주장했으며, Cosmos 역시 Nemotron 처럼 모델·데이터·학습법을 모두 공개해 각자 발전시킬 수 있게 한다. 같은 에이전틱 로봇 구조가 자율주행(Llama Mio 2)과 휴머노이드 로봇(Isaac GR00T)으로 확장된다.

결론 – GPU 를 넘어 인프라 전반으로

젠슨 황의 마무리는 지난 수개월 사이 컴퓨팅 구조가 바뀌었다는 인식으로 수렴한다. 에이전트가 실현되고 최신 프런티어 모델과 결합하면서 AI 가 실제 업무를 수행하게 됐다는 것이다. 핵심은 '모델 + 도구·스킬을 사용하는 하네스 + 런타임'이라는 동일한 컴퓨팅 패턴이 클라우드·온프레미스·PC·로봇 어디서나 반복된다는 점이다. **NVIDIA 의 전략은 이 패턴이 구동되는 단위, 즉 AI 팩토리 전체를 자사 스택으로 통합 공급하는 데 있으며, 이를 통해 회사를 GPU 기업에서 인프라 기업으로 재정의하고 있다.**

Compliance Notice

- 당사는 동 자료를 기관투자자 또는 제 3 자에게 사전 제공한 사실이 없습니다.
- 동 자료에 게시된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.

고지사항