MIRAE ASSET 미래에셋증권

Equity Research
2026.3.18

[반도체]
김영건
younggun.kim.a@miraeasset.com

비중확대/유지

반도체

GTC 2026 키노트 리뷰

엔비디아, GTC(GPU Technology Conference) 2026 개최

엔비디아 GTC가 미국 현지시간 3/16 ~ 19 동안 4일간 진행되며, 한국 시간으로 전일 오전 CEO인 젠슨황의 키노트가 진행되었다. 키노트에서는 추론 시장의 본격화를 선언하며 관련 수요에 대한 긍정적인 전망을 제시했으며, 추론 워크로드에 최적화된 LPU 기반 차세대 랙 시스템 LPX도 함께 공개되었다.

다만, 그간 Groq 기반의 추론 시스템 도입 가능성이 제기되어 왔기 때문에, H/W 측면에서의 서프라이즈 요인은 다소 제한적이었다고 판단된다. 매출 전망 역시 기대치에 부합하는 수준이었다. 그럼에도 엔비디아가 추론 영역에서도 강력한 경쟁 우위를 유지하고 있음을 확인할 수 있었으며, 추론 수요의 성장성을 체감할 수 있었다.

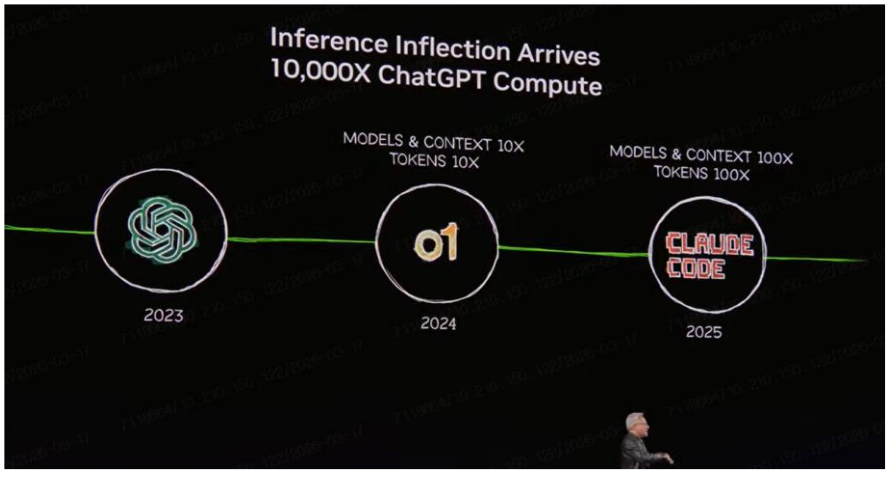
차세대 추론 시스템인 Groq LPX를 공개

엔비디아는 AI 산업의 추론의 시대가 본격화되었음을 재차 강조하였다. 추론 수요가 변곡점에 도달하면서 필요한 토큰 처리량은 2년 전 ChatGPT 모먼트 당시 대비 약 10,000배 증가하였다고 밝혔다. 이러한 흐름 속에서 Blackwell과 Rubin 아키텍처 기반 제품의 27년까지 매출을 1조 달러 이상으로 전망하였다.

추론 시대의 본격화에 맞추어, GTC 2026의 메시지는 엔비디아 플랫폼이 추론 영역에서도 최고 수준의 성능을 제공한다는 것이었다. 특히 Groq 포트폴리오 통합이 핵심이었다. Vera Rubin과 Groq의 통합 설계는 추론 과정에서 발생하는 메모리 및 지연시간 병목을 최소화하는 데 초점이 맞춰져 있다.

Groq과의 통합 설계를 바탕으로, Vera Rubin 플랫폼은 상위 추론 연산 구간(TPS/User 200 이상)에서 Blackwell 대비 3배, 최상위 추론 연산 구간(TPS/User 400 이상)에서 35배의 연산 성능을 제공한다. 이를 통해 Vera Rubin은 Blackwell 대비 5배 이상의 수익 창출력을 가진다.

그림 1. ChatGPT 모먼트 이후 토큰 처리량 10,000배 증가



자료: 엔비디아, 미래에셋증권 리서치센터

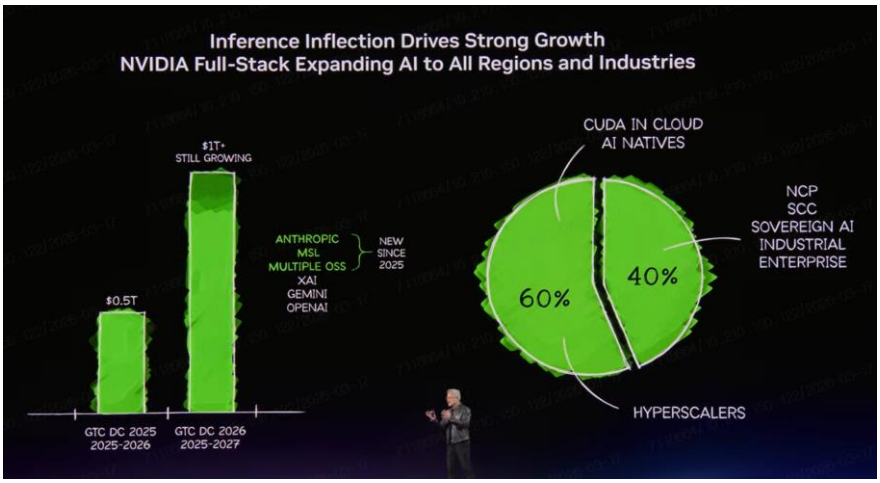
추론 수요의 본격화가 확인된 GTC 2026

엔비디아는 GTC 2026을 통해 AI 산업의 추론의 시대가 본격화되었음을 재차 강조하였다. 추론 수요가 변곡점에 도달하면서 필요한 토큰 처리량은 2년 전 ChatGPT 모먼트 당시 대비 약 10,000배 증가하였다고 밝혔다. 매년 대량의 GPU 출하에도 불구하고 관련 제품의 현물 가격이 치솟고 있다는 것이 이러한 수요 강도를 방증한다.

컴퓨팅 패러다임이 추론 중심으로 전환되면서, 컴퓨팅 용량 확보가 곧 매출로 연결되는 선순환 구조가 형성되고 있음을 강조하였다. 이러한 흐름 속에서 엔비디아는 Blackwell과 Rubin 아키텍처 기반 제품의 27년까지 매출을 1조 달러 이상으로 전망하였다. 이는 지난해 10월 GTC D.C.에서 제시한 26년까지 약 5,000억 달러 매출 전망 대비 상향된 수치다.

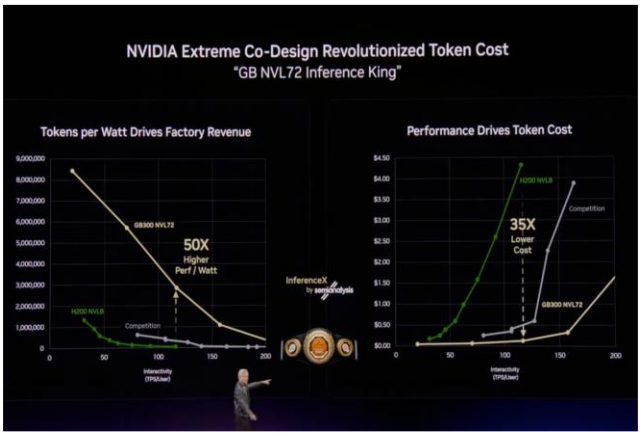
엔비디아는 현재 AI 전 영역에서 최고 수준의 성능을 제공할 수 있는 유일한 범용 컴퓨팅 플랫폼임을 강조하였다. 또한 고객 기반의 약 60%가 상위 5개 하이퍼스케일러로 구성되어 있으며, 나머지 40%는 소버린 클라우드, 엔터프라이즈, 산업용 로보틱스, 엣지 컴퓨팅, 소형 서버 등 다양한 영역에 분산된 구조라고 설명하였다.

그림 1. 엔비디아 2027년까지 AI 칩 매출 1조 달러 이상 전망



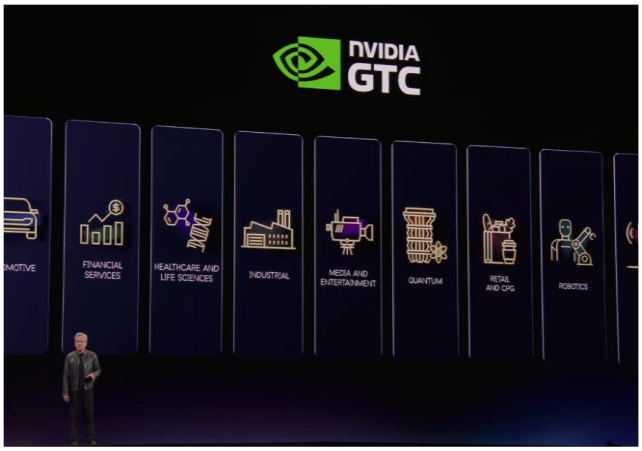
자료: 엔비디아, 미래에셋증권 리서치센터

그림 3. Extreame Co-Design을 통한 토큰 효율 구현



자료: 엔비디아, 미래에셋증권 리서치센터

그림 4. 전 산업을 포괄하는 엔비디아의 고객 기반



자료: 엔비디아, 미래에셋증권 리서치센터

차세대 AI 아키텍처 업데이트

추론 시대의 본격화에 맞추어, GTC 2026의 메시지는 엔비디아 플랫폼이 추론 영역에서도 최고 수준의 성능을 제공한다는 것이었다. 특히 Groq 포트폴리오 통합이 핵심이었다.

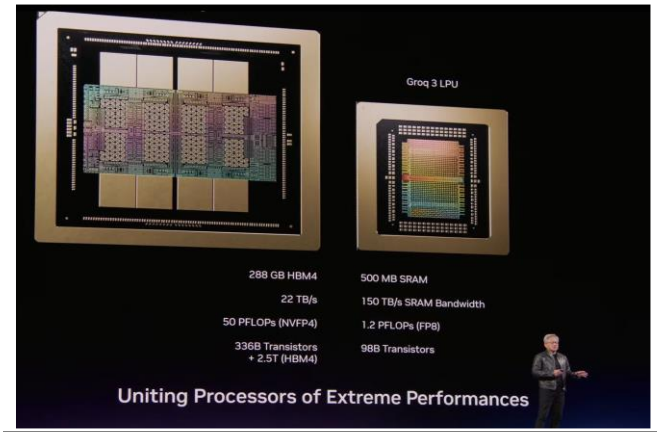
Vera Rubin: Groq LPX 기반 추론 시스템 도입

1) Groq 3 LPX 랙 공개: 엔비디아는 Vera Rubin 플랫폼의 차세대 추론 시스템인 Groq LPX를 공개하였다. Vera Rubin과 Groq의 통합 설계는 추론 과정에서 발생하는 메모리 및 지연시간 병목을 최소화하는 데 초점이 맞춰져 있다. Dynamo를 통해 추론 워크로드를 분리하고, 대역폭과 지연시간 부담이 가장 큰 디코드 단계를 Groq LPX가 전담하는 구조이다.

Groq LPU는 결정론적 스트림 프로세서(Deterministic data flow processor) 구조를 기반으로, 컴파일러 스케줄링 방식을 채택하고 있다. 또한 Groq 3 LPU는 용량 500MB, 대역폭 150TB/s의 SRAM을 탑재해 처리 속도를 높였으며, 이를 기반으로 한 Groq 3 LPX는 315PFLOPS의 추론 연산 능력과 SRAM 128GB 용량, 40PB/s 대역폭을 구현하였다.

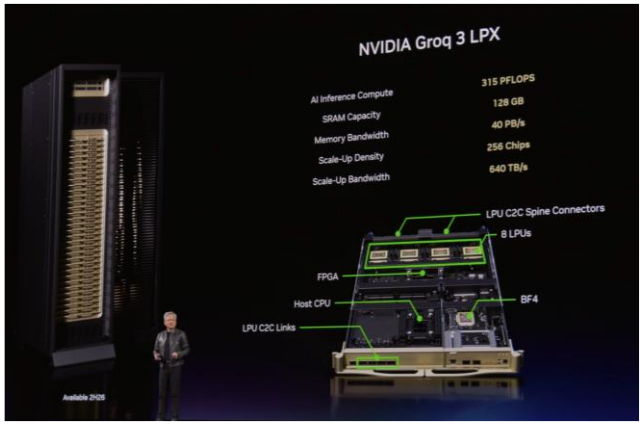
2) Vera Rubin 성능 비교 제시: Groq과의 통합 설계를 바탕으로, Vera Rubin 플랫폼은 상위 추론 연산 구간(TPS/User 200 이상)에서 Blackwell 대비 3배, 최상위 추론 연산 구간(TPS/User 400 이상)에서 35배의 연산 성능을 제공한다. 이를 통해 Vera Rubin은 Blackwell 대비 5배 이상의 수익 창출력을 가진다.

그림 5. 엔비디아 Groq 3 LPU 세부 스펙



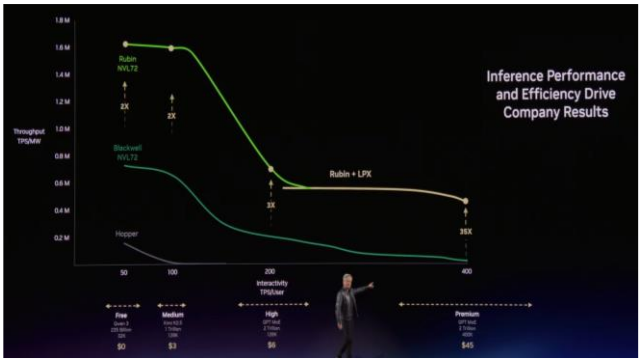
자료: 엔비디아, 미래에셋증권 리서치센터

그림 6. 엔비디아 Groq 3 LPX 랙 세부 스펙



자료: 엔비디아, 미래에셋증권 리서치센터

그림 7. 최상위 추론 연산 구간에서 35배 이상의 성능 제공



자료: 엔비디아, 미래에셋증권 리서치센터

그림 8. Blackwell 대비 5배 이상의 수익 기회 제공

자료: 엔비디아, 미래에셋증권 리서치센터

Rubin Ultra: Kyber 랙 구조와 차세대 LPU 확장

1) Kyber 랙 도입: Rubin Ultra 플랫폼부터는 새로운 랙 시스템인 Kyber가 도입될 예정이다. Kyber는 랙당 NVLink 도메인을 기존 대비 두 배로 확장해 하나의 랙에서 최대 144개의 GPU까지 스케일업할 수 있는 구조다. 또한 랙 간 광학 연결을 통해 최대 NVL1152 규모까지 확장이 가능하다. 다만 기존 랙 시스템인 Oberon과의 호환성 역시 유지될 것이다.

2) Groq LP35 적용: 또한 Rubin Ultra 플랫폼에는 신규 Groq LPU인 Groq LP35가 적용될 예정이다. LP35에서는 NVFP4 데이터셋 지원이 추가되어 추론 처리 속도가 기존 대비 크게 향상된다. 특히 저정밀 연산 활용도가 높아지면서 대규모 모델 추론 환경에서 처리 효율이 더욱 높아질 것으로 기대된다. Rubin Ultra의 출시는 27년 하반기로 예정되어 있다.

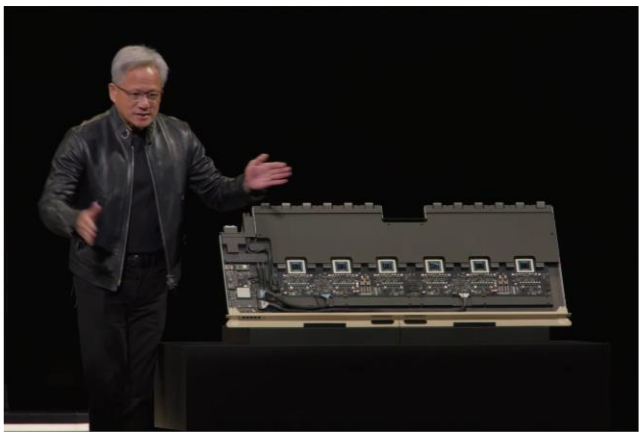
Feynman: 차세대 AI 컴퓨팅 플랫폼

Rubin 이후 차세대 플랫폼인 Feynman에는 신규 CPU ROSA가 탑재될 예정이다. 또한 차세대 LPU인 Groq LP40과 DPU인 BlueField-5가 함께 적용될 계획이다. 이를 통해 AI 연산과 네트워크 처리 전반에서 플랫폼 수준의 성능 향상이 기대된다. Feynman 플랫폼의 출시는 28년으로 예정되어 있다.

그림 9. Rubin Ultra와 Feynman으로 이어지는 로드맵

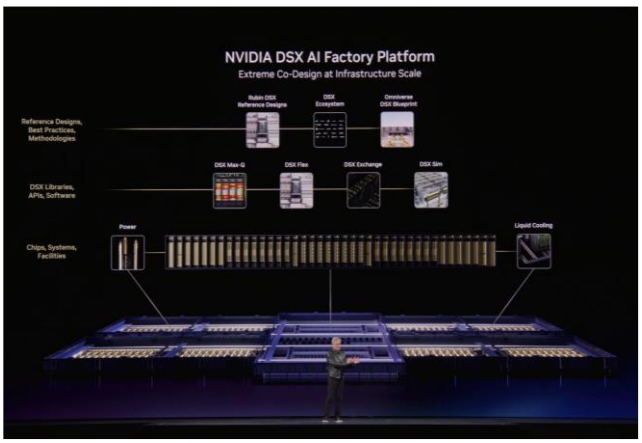
자료: 엔비디아, 미래에셋증권 리서치센터

그림 10. Rubin Ultra의 Kyber 컴퓨팅 노드



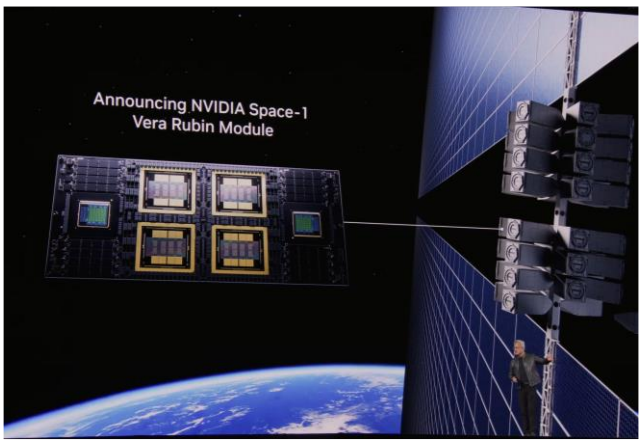
자료: 엔비디아, 미래에셋증권 리서치센터

그림 11. 엔비디아 DSX AI 팩토리 플랫폼



자료: 엔비디아, 미래에셋증권 리서치센터

그림 12. 엔비디아 Space-1 Vera Rubin 모듈



자료: 엔비디아, 미래에셋증권 리서치센터

추론 생태계 확장: AI 에이전트와 자율주행

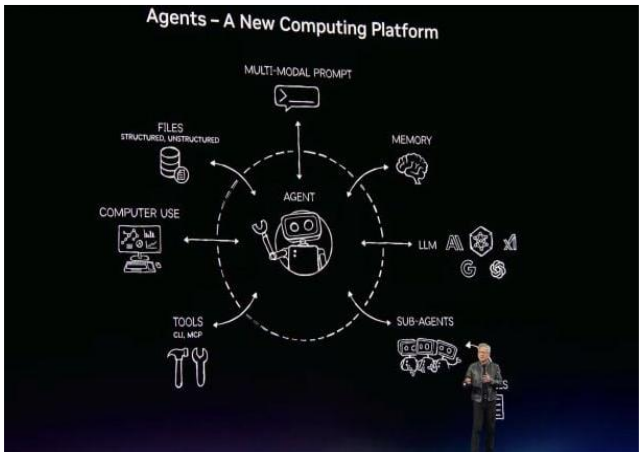
기업용 AI 에이전트 플랫폼 NemoClaw 공개

AI 에이전트 시대 본격화에 맞추어 기업용 AI 에이전트 플랫폼 NemoClaw를 공개하였다. NemoClaw는 오픈소스 프레임워크 OpenClaw를 기반으로 구축된 플랫폼으로, 기존 OpenClaw의 취약점으로 지적되던 기업용 보안과 개인 정보 보호 기능이 내장되었다. OpenClaw를 추론의 킬러 콘텐츠로 보고 플랫폼을 구축해 도입을 가속화하려는 시도이다.

로보택시 파트너십 발표

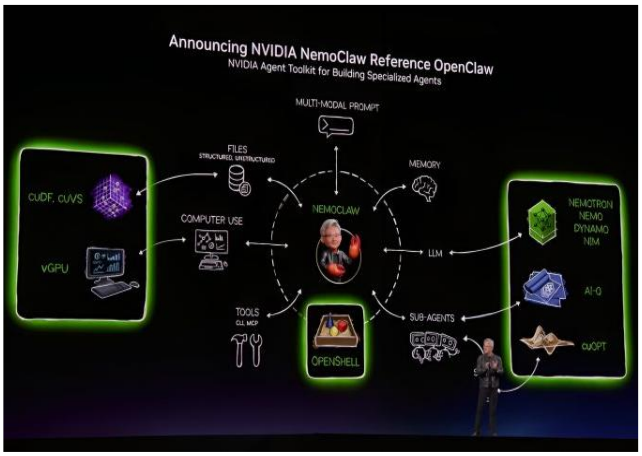
자율주행 분야에서의 'ChatGPT 모먼트' 도래를 선언하였다. 이와 함께 엔비디아 로보택시 플랫폼의 신규 파트너로 BYD, 현대차, 닛산, 지리와의 협력을 발표하였다. 또한 Uber와의 파트너십을 통해 복수의 신규 도시에서 로보택시 사업을 추진할 계획이다. 자율주행 역시 동사의 추론 수요 확대를 위한 기반 구축 과정으로 해석된다.

그림 13. 새로운 컴퓨팅 서비스의 핵심은 에이전트



자료: 엔비니다, 미래에셋증권 리서치센터

그림 14. 엔비디아 NemoClaw 서비스



자료: 엔비디아, 미래에셋증권 리서치센터

그림 15. 엔비디아 Physical AI 및 자율주행 협력사

자료: 미래에셋증권 리서치센터

Compliance Notice

- 당사는 자료 작성일 현재 해당 회사와 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 본 자료에서 매매를 권유한 금융투자상품 및 권리를 보유하고 있지 않습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.

본 조사분석자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 본 조사분석자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포할 수 없습니다.