미국은 지금

GTC 2026: 전선을 확장한 엔비디아

키움증권 리서치센터 글로벌리서치팀
US Strategy 김승혁 ocean93@kiwoom.com

Issue Brief

시작된 GTC 2026 과 엔비디아의 청사진

GTC 2026 은 단순한 신제품 발표 행사가 아니었다. 190 개국 3 만 명이 모인 이 자리에서 **엔비디아가 던진 메시지는, AI 를 단순한 도구가 아니라 전기나 인터넷처럼 모든 산업이 올라타는 기반 인프라로 재정의하겠다는 것이었다.** 그리고 그 인프라의 칩부터 소프트웨어, 에이전트, 자율주행, 심지어 궤도 데이터센터까지 전 계층을 엔비디아 혼자 설계하겠다는 청사진을 구체적 제품 및 파트너십으로 채워 보였다. 금융·의료·제조·통신 분야 기업들이 AI 를 실험 단계에서 실제 사업 프로세스에 녹여내고 있다는 사례들도 등장했고, 이것이 결국 이번 행사 전체의 논리적 토대가 되는 중이다.

베일 벗은 Vera Rubin 과 1 조 달러 매출 가이던스

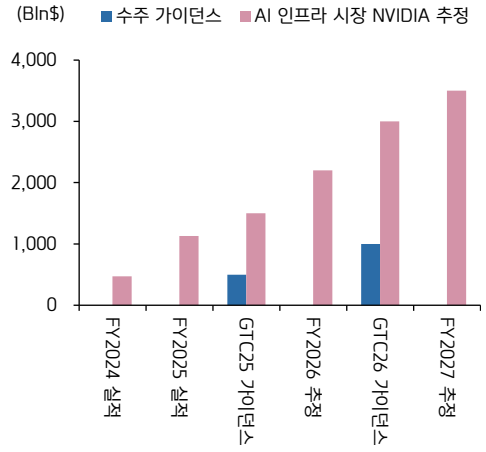
하드웨어 핵심은 Vera Rubin 플랫폼 정식 발표와 Groq 3 LPU 공개다. Vera Rubin 은 GPU, CPU, 스토리지, 네트워킹 칩을 하나의 최적화 단위로 묶은 AI 팩토리 전용 시스템으로, Blackwell 대비 와트당 추론 처리량이 10 배 향상됐다. 엔비디아가 비독점 기술 라이선스와 인재 영입 방식으로 확보한 Groq 기술의 산출물인 Groq 3 LPU 는 삼성에서 양산 중이며 올 Q3 출하 예정이다. GPU 가 문맥 이해의 고처리량 작업을 맡고 LPU 가 실제 토큰 생성의 저지연 작업을 담당하는 구조로, 둘을 함께 쓸 때 추론 비용이 추가로 크게 낮아진다. 소프트웨어 측면에서는 기업 환경에서 AI 에이전트를 안전하게 운영하는 스택인 NemoClaw, Mistral 등 AI 랩들과 차세대 오픈 언어 모델을 공동 개발하는 Nemotron Coalition 도 공개됐다. **2027 년까지의 매출 가시성은 지난해 GTC 대비 두 배인 1 조 달러로 상향됐으며, 이 수치에는 Groq LPU 와 중국향 매출이 포함되지 않은 보수적 기준이다.**

칩 메이커에서 플랫폼 지배자로

2026 GTC 컨퍼런스가 시장에 던진 함의는 세 가지다. 첫째, **AI CapEx 사이클이 정점을 찍었다는 우려에 대한 실질적 반박이다.** 훈련 수요와 별개로 추론 수요가 폭발하기 시작했고, 금융과 의료·뷰티처럼 이질적인 산업들의 사업 프로세스에 이미 AI 가 녹아들고 있다. 둘째, **NemoClaw 와 같은 소프트웨어 확장은 엔비디아의 수익 구조가 GPU 판매 일변도에서 변화하기 시작했다는 신호다**. 셋째, 현대차·BYD 를 비롯한 주요 완성차 OEM 들의 자율주행 합류와 우버의 글로벌 로보택시 배치 계획은 **피지컬 AI 가 생각보다 빠르게 수익화 단계로 진입할 수 있음을 보여준다.**

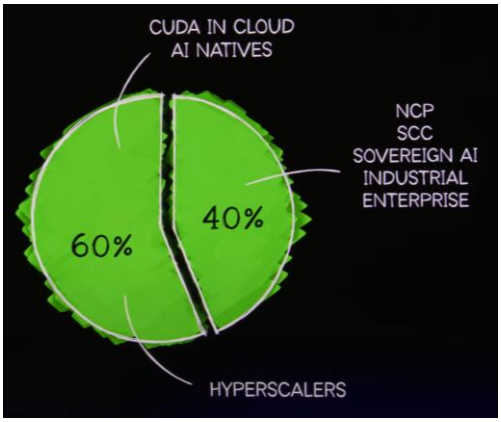
(2page 계속)

수주 가이던스와 인프라 시장 TAM 추정



자료: 엔비디아, 키움증권 리서치

고객 구성 비중(하이퍼스케일러: 60%)



자료: 엔비디아

이번 GTC 는 "AI 성장이 얼마나 오래 지속되는가"에 대한 답을 던졌으며, 그 답은 Vera Rubin 의 실제 양산 속도, Groq LPU 의 배포 규모, 하반기 클라우드 기업들의 설비투자 계획 등이 결정할 것으로 판단한다.

토큰이 새로운 비트다

GTC 는 표면적으로 개발자 컨퍼런스지만, 실제론 엔비디아가 전 세계 파트너·고객·투자자에게 향후 12 개월의 기술·사업 방향을 공식 선언하는 자리다. 이번 GTC 2026 은 3 월 16 일부터 19 일까지 진행되며, 190 개국에서 3 만 명 이상이 현장에 참석한 역대 최대 규모로 열렸다. 450 개 이상의 스폰서, 1,000 개 세션, 2,000 명의 연사가 함께했다. 이번 GTC 에서 젠슨 황 CEO 가 던진 메시지는 "더 빠른 GPU 를 만들었습니다"가 아니었다. 그는 AI 를 전기나 인터넷처럼 인류 문명의 기반 인프라로 재정의하면서, 엔비디아가 그 인프라의 모든 계층을 설계하는 회사가 되겠다고 선언했다. 칩을 기점으로 스토리지, 네트워킹, 소프트웨어, 오픈 모델, 에이전트, 자율주행 로봇, 심지어 우주 데이터센터까지 범위를 확장했다.

기조연설은 토큰을 현대 AI 의 기본 단위로 규정하는 영상으로 시작됐다. 이는 엔비디아가 자신을 "GPU 파는 회사"가 아니라 "토큰 생산 인프라 제공 회사"로 재포지셔닝하고 있다는 신호이다. 제리 양이 야후를 인터넷 디렉토리 회사로 정의했을 때, 구글은 인터넷을 검색 가능한 데이터베이스로 정의한 것이 회사 운명을 갈랐다.

"매출 $1 조" 무엇을 의미하는가?

이번 GTC 에서 가장 강력한 메시지는 27 년까지 최소 1 조 달러의 수주 가시성을 보고 있다는 젠슨 황의 발언이다. 작년 GTC 에서 Blackwell 과 Rubin 을 합산해 26 년까지 $0.5 조의 수주 가시성을 발표했기에 2 배 상향 조정된 부분이다. 여기서 놓쳐서는 안 되는 맥락이 있다. **이 $1 조라는 수치에는 CPU, 네트워킹 장비, Groq LPU, Rubin Ultra, 그리고 중국향 H200 변형칩이 포함되어 있지 않다는 점이다.** 화요일 기자회견에서 젠슨 황은 이 점을 명확히 했다. 즉, $1 조는 보수적 기준점이지 상한선이 아니다.

해당 부분은 또 다른 인사이트를 제공한다. 작년까지 시장의 가장 큰 우려는 "AI CapEx 가 정점을 찍었는가" 부분이다. 하이퍼스케일러들이 많은 돈을 쏟아붓고 있지만 ROI 가 명확하지 않다는 우려다. 젠슨 황의 $1 조 발언은 그 우려에 대한 정면 반박이다. AI 가 훈련 단계에서 추론 단계로 넘어가면 수요의 성격 자체가 바뀌고 있다. 단순히 수요가 늘어나는게 아닌 것이다. **훈련은 한 번 하고 끝이지만, 추론은 서비스가 존재하는 한 24 시간 365 일 돌아간다. 추론 인프라는 훈련 인프라보다 훨씬 더 지속적이고 반복적인 수요를 만든다. 이것이 젠슨 황의 CapEx 투자가 지속될 수밖에 없는 근거이다.**

Vera Rubin 이 단순한 차세대 GPU 가 아닌 이유

젠슨 황은 Vera Rubin 을 생각할 때 우리는 전체 시스템으로 생각한다고 언급했다. **Vera Rubin 은 단순한 GPU 가 아니다. 7 개의 칩(GPU, CPU, LPU, DPU, 네트워크 칩, 스토리지 칩, 옵틱스 칩), 5 개의 랙스케일 시스템, 1 개의 슈퍼컴퓨터로 이루어진 전체 패키지다.** 스펙도 압도적이다. Rubin GPU 는 TSMC 3nm 공정에 3,360 억 개의 트랜지스터를 탑재했고(Blackwell 의 2,080 억 대비 +61%), HBM4 메모리 대역폭은 22 TB/s(HBM3e 8 TB/s 대비 +175%)다. NVL72 랙의 총 NVLink 6 대역폭은 260 TB/s 로, 엔비디아는 이것이 "인터넷 전체 대역폭을 초과한다"고 주장했다. 그리고 와트당 토큰 효율은 Blackwell 대비 10 배 향상되었다.

이러한 수치들보다 더 중요한 것은 엔비디아가 왜 이것을 "시스템"으로 팔기 시작했는가다. 개별 GPU 로 경쟁하면 AMD 나 Intel 과의 스펙 싸움이 된다. **하지만 7 개 칩이 하나의 최적화 대상으로 통합된 AI 팩토리 시스템을 판다면, 경쟁의 단위가 달라진다. 고객 입장에서는 개별 부품을 조합해서 최적화하는 것보다 처음부터 통합 설계된 시스템을 사는 것이 훨씬 편하다.** 진입장벽이 "더 좋은 GPU 를 만드는 것"에서 "더 잘 통합된 AI 팩토리를 만드는 것"으로 높아졌다는 것이 핵심이다.

Groq 3 LPU 에 대하여

GTC 에서 주목해 볼 또다른 요인은 Groq 3 LPU 였다. Groq 는 Google 의 내부 TPU 를 설계한 엔지니어들이 창업한 추론 특화 칩 스타트업으로, 엔비디아가 작년 12 월 $200 억에 일부 자산 인수, 팀과 기술에 대해 비독점 라이선스 계약 등을 진행했다.

이 칩의 기술적 특성은 독특하다. GPU 는 병렬 처리에 최적화된 범용 가속기인 반면, **Groq 의 LPU(Language Processing Unit)는 "결정론적 데이터플로우 프로세서"로 초저지연 토큰 생성에 특화돼 있다**. 칩당 500MB 의 온칩 SRAM 을 탑재해 메모리 접근 지연을 최소화하는 구조다. 삼성에서 양산 중이며 Q3 2026 출하 예정이다.

젠슨 황은 Vera Rubin 과 Groq LPU 의 역할 분담을 명확히 설명했다. Vera Rubin 은 AI 추론의 "prefill" 단계(문맥을 이해하는 고처리량 작업)를 담당하고, Groq LPU 는 "decode" 단계(실제 토큰을 생성하는 저지연 작업)를 담당한다. 배포 가이드라인도 구체적으로 제시했다. "대부분의 워크로드가 고처리량이라면 100% Vera Rubin 으로 충분하지만 코딩이나 엔지니어링 토큰 생성 워크로드가 많다면 전체 데이터센터의 약 25%에 Groq 를 추가하라고 명시했다.

엔비디아와 Groq 의 비독점 라이선스 계약은 엔비디아가 GPU 단독으로는 해결하기 어려웠던 초저지연 추론 시장을 Groq LPU 통합으로 공략하게 됐다는 점에서 의의가 있다. 특히 자율주행, 수술 로봇, 실시간 금융 거래처럼 밀리초 단위의 응답이 생명인 분야에서 이 조합의 가치는 더욱 높아진다.

새로운 아키텍처 Feynman

이번 GTC 에서 엔비디아는 2028 년 아키텍처 Feynman 까지 공개했다. Feynman 에는 Rosalind Franklin 의 이름을 딴 새 CPU 'Rosa', 차세대 LP40 LPU, BlueField-5, CX10, Kyber 인터커넥트(구리와 공동패키징 광학 모두 지원), 그리고 TSMC A16(1.6nm) 공정과 실리콘 포토닉스 최초 도입이 예상된다.

핵심은 2026 년 행사에서 2028 년 아키텍처까지 보여줬다는 것이다. 이는 고객에게 "어차피 2028 에도 우리를 써야 한다"는 신호를 주는 동시에, 경쟁사에게 "따라오려면 이 수준까지 와야 한다"는 진입 장벽을 미리 세우는 전략이다. 고객 입장에서는 이미 Vera Rubin 기반으로 인프라를 최적화했다면, 그 다음 세대로 넘어갈 때 전환 비용이 발생한다. **12 개월 주기로 세대교체를 강제하는 이 구조가 엔비디아에게 예측 가능한 ARR 을 만들어 준다.**

SaaS 가 아닌 AaaS(Feat. NemoClaw)

젠슨 황이 기조연설에서 OpenClaw 를 처음 언급했을 때 청중의 반응은 폭발적이었다. OpenClaw 는 오스트리아 개발자 Peter Steinberger 가 1 월 공개한 오픈소스 AI 에이전트로, 공개 직후 GitHub 역사상 가장 빠른 성장을 기록했다. (Steinberger 는 이후 OpenAI 에 합류했고, 샘 올트만은 OpenClaw 를 오픈소스로 계속 유지하겠다고 발표했다.)

젠슨 황은 OpenClaw 가 AI 에이전트 시대의 기반 프로토콜이 될 것으로 보고, 그 위에 자사의 엔터프라이즈 스택을 올리겠다는 전략을 세웠다. 해당 지점에서 발표된 것이 NemoClaw 다. 이는 OpenClaw 를 기업 환경에서 안전하게 배포할 수 있도록 정책 집행, 네트워크 가드레일, 프라이버시 등을 결합한 관리 배포 패키지이다. NemoClaw 를 통해 어떤 데이터에 접근하고 어떤 툴을 사용할 수 있는지를 기업 IT 정책 수준에서 제어한다.

젠슨 황은 모든 SaaS 회사가 AaaS(Agentic as a Service) 회사가 될 것이라 주장했다. **기존 SaaS 회사들이 AI 에이전트를 제품에 통합하려면 그 에이전트가 안전하게 돌아갈 수 있는 툴이 필요하며, 엔비디아는 NemoClaw 로 레이어를 선점하려 한다.** 엔비디아가 반도체 회사를 넘어 엔터프라이즈 소프트웨어 인프라 회사로 진입하려는 시도다. AWS 가 EC2 서버를 팔다가 S3, RDS, Lambda 로 소프트웨어 레이어를 잠식했던 경로와 유사하다.

정리하면 엔비디아는 AI 소프트웨어 시장이 OpenAI, Anthropic, Google 같은 소수의 대형 모델과 그 아래 범용 하드웨어로 단순하게 양분되는 구도를 원하지 않는다. 오픈 모델 진영에도 적극적으로 개입해, AI 생태계 전체를 설계하는 주체로 남겠다는 전략이다. 즉, OpenAI 나 Anthropic 이 폐쇄형 모델로 시장을 장악하더라도 그 모델들이 결국 엔비디아 인프라 위에서만 돌아가도록 만드는 것이 한 축이고, 오픈소스 모델이 성장하더라도 엔비디아 Nemotron 생태계 안에서 크도록 유도하는 것이 또 다른 축이다. 어느 쪽이 이기든 엔비디아는 반드시 그 판에 있는 구조를 만들고 있다.

피지컬 AI 기대감 및 AI 활용 산업 사례

젠슨 황은 기조연설 후반부에서 "자율주행의 ChatGPT 모멘트가 도래했다"고 선언했다. 과장처럼 들릴 수 있지만, 숫자가 이를 뒷받침한다. 이번 GTC 에서 BYD, 현대차/기아, 닛산, Geely 가 엔비디아 자율주행 프로그램에 새로 합류했고, 기존 파트너인 메르세데스, 도요타, GM 까지 모두 엔비디아 플랫폼 위에서 자율주행을 개발하고 있다. Uber 와의 협력도 눈길을 끈다. 2027 년 LA 와 SF 를 시작으로, 2028 년까지 4 개 대륙 28 개 도시에 엔비디아 소프트웨어로 구동되는 로보택시를 배치하는 것이 목표다.

엔비디아는 직접 차를 만들거나 운영하지 않는다. **훈련용 시뮬레이션 환경, 추론 칩, 소프트웨어를 공급하고, 실제 배포와 그에 따른 리스크는 완성차 업체와 운영사가 진다.** 자율주행이 어떤 방식으로 상용화되든 엔비디아는 인프라 공급자로서 수익을 가져가는 구조다.

금융에서는 Mastercard 가 수억 건의 실거래 데이터로 학습한 AI 모델을 개발 중이라고 밝혔다. "기존 머신러닝 기법을 뛰어넘는 초기 성과가 나오고 있다"고 했고, Revolut 과 Adyen 도 비슷한 방향으로 가고 있다. 트레이딩 회사 Jump Trading 은 엔비디아 Rubin 플랫폼을 채택한 최초의 트레이딩 펌 중 하나라고 공개했다.

소매 분야에서는 ChatGPT 나 Gemini 로 쇼핑하는 소비자가 가맹점 시스템과 직접 연결되어 추천, 결제, 프로모션까지 한 번에 처리되는 오픈소스 아키텍처를 발표했다. 뷰티 분야에서는 L'Oréal 이 엔비디아 플랫폼을 통해 선크림 분자 시뮬레이션 속도를 100 배 높였다고 밝혔다. 수년이 걸리던 제품 개발 과정이 대폭 단축된다.

이 사례들이 공통적으로 보여주는 것은 AI 도입이 실험에서 실제 운영으로 넘어가고 있다는 점이다. Mastercard 가 실거래 데이터로 모델을 훈련하고, L'Oréal 이 신제품 개발 파이프라인에 AI 를 적용한다는 것은 AI 가 핵심 사업 과정의 일부가 됐음을 의미한다. 실험용 GPU 는 몇 대면 충분하지만, 24 시간 실제 서비스는 수십 배 이상의 인프라를 필요로 한다. 이 단계 전환이 추론 인프라 수요를 끌어올리는 핵심 원인이다.

파격적인 발표는 우주 컴퓨팅이었다. **엔비디아는 AI 데이터센터를 궤도로 가져가겠다는 비전 아래 Space-1 모듈을 공개했다.** Planet Labs 는 이미 위성에서 수 테라바이트의 데이터를 실시간으로 처리하고 있고, CERN 연구자들도 고에너지 물리 AI 모델에 엔비디아 칩을 쓰고 있다. 물론 단기 실적 요소로 볼 필요는 없다. 우주 컴퓨팅은 아직 초기 단계다. 하지만 중요한 것은 엔비디아가 AI 인프라의 확장 범위를 지구 밖까지로 설정하고 있다는 신호다. 데이터센터에서 자율주행 차량으로, 차량에서 로봇으로, 로봇에서 궤도로 TAM 을 끊임없이 넓혀가는 전략이 진행중이다.

GTC 2026 에서 알 수 있었던 3 가지

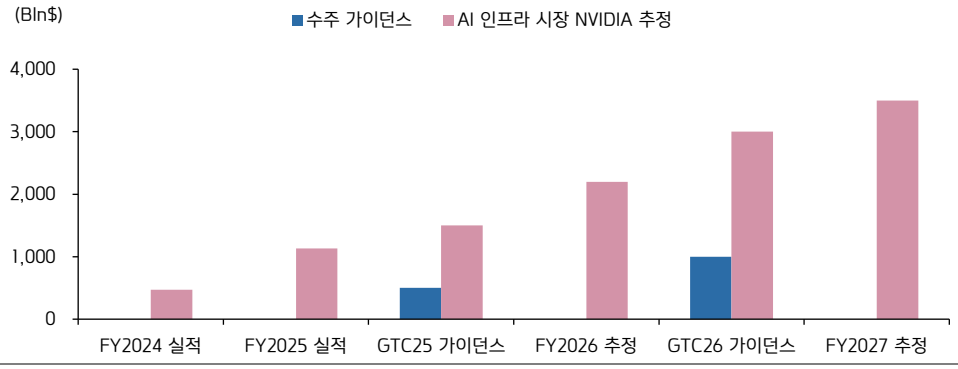
GTC 2026 행사가 시장에 던진 핵심 질문들을 정리하면 우선은 "AI 투자 사이클은 아직 초입인가, 아니면 정점인가" 이다. 젠슨 황의 답은 명확하다. AI 모델을 학습시키는 인프라는 정점을 향하고 있을 수 있지만, AI 서비스를 실제로 돌리는 추론 인프라는 이제 막 시작됐다. 이 논리가 맞다면 투자 정점 우려는 틀린 것이고, 엔비디아의 성장은 2027 년을 넘어서도 이어진다.

두 번째는 "소프트웨어 전략이 성공할 수 있는가"다. 엔비디아는 칩 판매에서 소프트웨어 구독과 플랫폼 수익으로 사업 구조를 바꾸려 하고 있다. 성공하면 수익성이 근본적으로 개선된다. 실패하면 다시 칩 가격 경쟁으로 돌아간다. 소프트웨어 시장에서 엔비디아의 경험이 부족하다는 점이 가장 큰 변수다.

세 번째는 "물리 AI의 수익화 시점이 언제인가"다. 자율주행과 로보틱스의 대규모 상용화까지는 아직 수년이 걸린다. 하지만 엔비디아는 로봇이 실제로 배포되기 전부터 그 로봇을 훈련시키는 인프라를 팔고 있다. 수익화는 생각보다 빠를 수 있다.

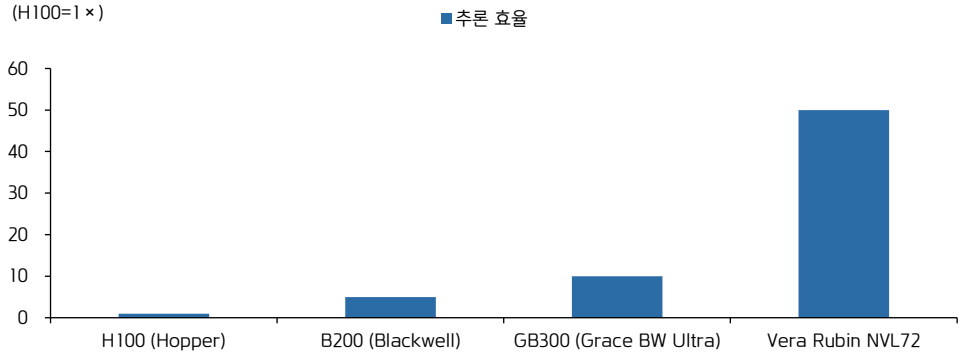
엔비디아는 이미 반도체 회사만이 아니다. AI 경제의 설계도를 그리고 있으며, 그 위에서 움직이는 모든 것(칩, 소프트웨어, 로봇, 자율주행차, 위성)에서 수익을 가져가는 구조를 만들고 있다. 체질적 변화를 가장 빠르고 적극적으로 시도하는 중이다.

수주 가이던스와 인프라 시장 TAM 추정



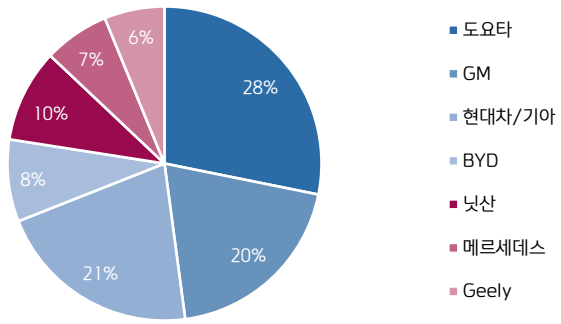
자료: 엔비디아, 키움증권 리서치

엔비디아 아키텍쳐 간 추론 효율 비교



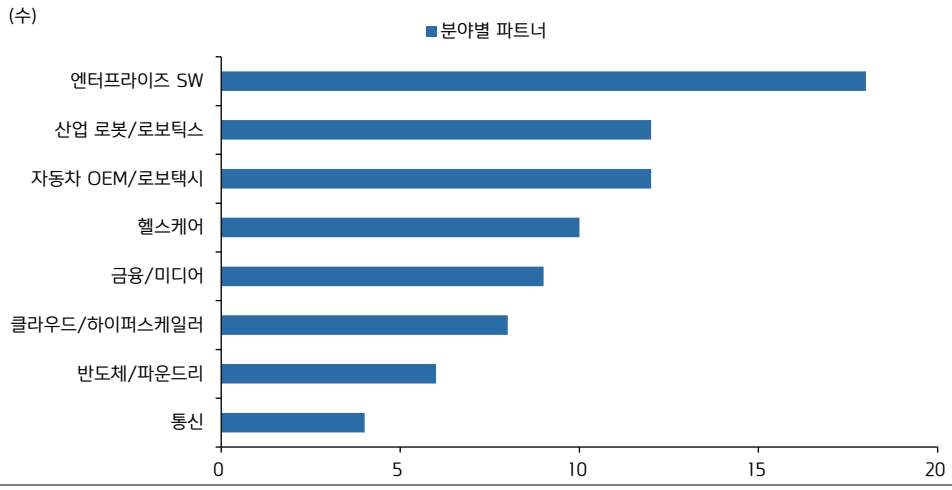
자료: 엔비디아, 키움증권 리서치

OEM 생산 규모 분포 (연간)



자료: 각사 IR, 키움증권 리서치

GTC 2026 에서 확인된 엔비디아의 분야별 파트너 수



자료: 엔비디아, 키움증권 리서치

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