미국은 지금

엔비디아 GTC 2025 기조연설

키움증권 리서치센터 글로벌리서치팀
US Strategy Analyst 김승혁 ocean93@kiwoom.com

☑ Issue Brief

Ⅰ. NVIDIA GTC 2025 키노트 연설

상징적 위치에서 진행된 GTC 2025 키노트 연설

2025 년 10 월 28 일(현지시간) 엔비디아 젠슨 황 CEO 는 GTC 2025 기조연설을 워싱턴 D.C.에서 진행했다. 이번 행사는 GTC 역사상 처음으로 실리콘밸리가 아닌 미국 정치의 중심지에서 열렸다는 점에서 상징적 의미가 있다. 연설 말미에 젠슨 황 CEO 는 매년 GTC 를 워싱턴 D.C.에서 정례적으로 개최하길 희망한다고 밝히며, 관계사·직원뿐 아니라 정부 관계자들에게도 감사를 전했다. 개최지 선택과 정부 관계자 언급은 엔비디아가 단일 기업을 넘어 국가 경쟁력 강화의 핵심 요소로 자리잡고 있음을 시사한다.

Ⅱ. NVIDIA 의 강점과 해결 과제

급증하는 추론 수요 충족을 위해서는 토큰 생성 비용 절감이 필수

젠슨 황 CEO 는 무어의 법칙의 한계를 극복할 수 있는 유일한 해법으로 자사의 GPU 와 CUDA 기반의 가속 컴퓨팅(accelerated computing)을 제시했다. 그는 CUDA 가 버전 13 에서 14 로 진화 중이며, 수백만 명의 개발자와 수억 대 GPU 가 형성한 방대한 호환성 생태계가 엔비디아의 핵심 경쟁력이자 '보물'이라고 강조했다. 또한 모델이 지능화될수록 사용량과 지불 의사가 함께 증가해 연산 수요를 확장시키는 선순환이 형성되고 있으며, 이는 100 조 달러 규모의 실물경제 생산성 향상으로 이어지고 있다고 설명했다. 다만 이러한 선순환이 지속되기 위해서는 '토큰 생성 비용'의 지속적 절감이 필수적임을 덧붙였다.

Ⅲ. 새롭게 공개된 기술 및 청사진

- ✓ **무선 통신 부문** Nokia 와의 전략적 제휴를 통한 6G AI 네이티브 통신
  - Nokia(5G 표준 특허 약 7,000 건 및 AirScale 기지국 보유) + NVIDIA ARC(Aerial RAN Computer) 결합
  - ARC 구성: Grace CPU + Blackwell GPU + Mellanox ConnectX 네트워킹 → CUDA 기반 기지국 컴퓨터
  - AI for RAN: 통신 효율을 높여 글로벌 전력 소비의 약 1.5~2% 절감에 기여
  - AI on RAN: 기지국 자체가 클라우드 역할 수행
  - 목표: 전 세계 수백만 개 기지국을 6G·AI-native 인프라로 업그레이드
- ✓ **양자 컴퓨팅**: GPU-QPU(양자칩) 통합 시대 선언
  - 'NVQ Link' 발표: QPU(양자칩)와 GPU 를 초당 테라바이트급 속도로 수천 회 왕복시키는 초고속 인터커넥트
  - 양자컴퓨터는 오류 보정을 위한 AI 기반 실시간 오류 정정이 필수이며, 이 과정에 GPU 연산이 활용됨
  - 초저지연 연산 결합의 중요성이 커지며 NVQ Link 의 필요성 역시 급등
  - CUDA-Q: GPU 와 QPU 가 각자의 역할을 자동으로 분담하고 협력하도록 돕는 운영 플랫폼으로, 마이크로초(백만분의 1 초) 단위로 연산이 이루어져 사실상 실시간처럼 작동. 즉 AI 의 빠른 계산력 + 양자의 복잡한 연산
  - 생태계: 17 개 양자 스타트업 및 미 에너지부(DOE) 산하 8 개 국립연구소(버클리·페르미·로스앨러모스 등) 참여
  - 미 에너지부(DOE)와 슈퍼컴퓨터 7 대 추가 공동 구축 발표 → 가속 컴퓨팅·AI·양자·원격 계측·로보틱 실험실 등 차세대 과학 인프라 전환 본격화

- ✓ **하드웨어 아키텍처**: NVLink 72 기반 '랙스케일 컴퓨터
  - 다수의 GPU 를 하나의 거대한 GPU 처럼 통합하는 '랙스케일 컴퓨팅(Rack-Scale Computing)' 모델 제시
  - NVLink 72: GPU 72 개를 초고속 패브릭으로 연결해 '단일 초대형 GPU' 수준의 통합 연산 성능 구현
  - 성능: NVLink 72 기반 랙스케일 컴퓨터의 기본 단위인 GB200 의 GPU 당 성능은 H200 대비 약 10 배 향상
    → 동일 조건에서 AI 연산량이 10 배 개선되며, 토큰 생성 단가(TCO/토큰초)는 세계 최저 수준

Ⅳ. 비즈니스 전망 및 협력 기반 전략

AI 오픈소스 개발 / 클라우드 인프라 / SaaS / 보안 / 데이터 분석 등 AI 전 분야에서 파트너십을 확장

- ✓ **매출 및 출하 전망:** Blackwell + Rubin 누적 5,000 억 달러 가시화
  - 누적 규모: 2026 년까지 Blackwell 및 초기 Rubin 플랫폼 합산 매출이 약 5,000 억 달러 규모로 현실화
  - 시점: 2025 년 잔여 분기 + 2026 년 4 개 분기(총 5 개 분기) 기준으로 이미 계약 확정분으로 확보된 수준
  - 출하량 전망: 5 개 분기 동안 Blackwell GPU 약 600 만 유닛 출하 예상
  - 성장 속도: 중국을 제외할 경우, Hopper 세대 누적 400 만 GPU → Blackwell + Rubin 합산 2,000 만 GPU 로 집계되며 이는 약 5 배의 성장 궤적
- ✓ **클라우드·데이터센터 CapEx 사이클**
  - 주요 투자 주체: 아마존, 코어위브(CoreWeave), 구글, 메타, 마이크로소프트, 오라클 등 6 개 CSP 가 대규모 설비투자 주도
  - 투자 시점: 고성능과 최저 수준 단위당 비용(TCO)을 갖춘 GB NVLink 72 의 대량 양산 체제가 본격화된 시점
- ✓ **피지컬 AI(Physical AI) 전략**
  - 3 대 컴퓨팅 구조: 학습(GB NVLink 72) + 디지털 트윈 시뮬레이션(Omniverse Computer) + 로보틱스(Jetson Thor)
  - 휴머노이드 로봇 협력 사례: Figure AI, Agility Robotics, Johnson & Johnson
- ✓ **자율주행 전략: Hyperion-Uber**
  - DRIVE Hyperion 플랫폼: 자율주행차를 위한 센서 구성·AI 연산 컴퓨팅을 표준화한 플랫폼 공개
  - 완성차 및 자율주행 스타트업의 하드웨어·센서·컴퓨팅 구조를 통합해 호환성을 높이고, 각사 고유의 자율주행 소프트웨어를 탑재할 수 있도록 설계
  - 완성차 협력사: Lucid, Mercedes-Benz, Stellantis
  - 자율주행 개발사 협력사: Wayve, Wabi, Aurora, Momenta, Nuro, WeRide 등
  - 우버(Uber) 협력: DRIVE Hyperion 플랫폼을 우버 글로벌 네트워크와 연결해 로보택시 호출 가능 생태계 구축
  - 젠슨 황 CEO 는 향후 연간 1 조 마일 주행, 1 억 대 생산, 5 천만 대 택시 규모의 자동차 시장이 AI 로보택시로 증강될 것이라 전망
- ✓ **기타 협력 사례**
  - 6G AI 네이티브 통신: Nokia
  - GPU 클라우드 인프라: CoreWeave, nScale, Nebius, Lambda, Crusoe
  - 대형 클라우드 사업자: AWS, Google Cloud, Microsoft Azure, Oracle Cloud
  - SaaS(서비스형 소프트웨어): ServiceNow, SAP, Synopsys, Cadence
  - 보안·에이전트: CrowdStrike
  - 데이터·정부·엔터프라이즈: Palantir

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