NVIDIA
GTC 2026

March 16–19
San Jose, CA

March 2026

# GPU TECHNOLOGY CONFERENCE 2026
# 키노트 핵심 요약 정리

K·ASIC

본 요약본은 미국 시장 진출 및 관련 사업을 추진하는 기업들에게 글로벌 반도체 기업들의 주요 동향을 신속하고 정확하게 전달하기 위해 마련되었습니다.

한미 AI 반도체 혁신센터(K-ASIC)는 매월 미국 내 반도체 산업에서 주목해야 할 핵심 뉴스를 선정하고, 이를 심층 분석 및 번역하여 제공합니다. K-ASIC는 한국 반도체 기업들의 미국 시장 진출을 지원하는 핵심 허브로서, 기술 협력, 정책 분석, 비즈니스 구축 및 투자 유치를 종합적으로 지원합니다.

본 자료는 정보 제공을 목적으로 작성된 것으로, 특정 기업 또는 산업에 대한 투자 권유나 재무적 의사결정을 위한 근거로 활용되어서는 안 됩니다. 또한 본 자료에 포함된 정보의 정확성이나 완전성에 대해 보증하지 않으며, 이를 기반으로 한 투자 판단 및 그 결과에 대한 책임은 전적으로 이용자에게 있습니다.

해당 자료는 한미 AI 반도체 혁신센터(K-ASIC)의 지식재산권에 의해 보호되며, 사전 승인 없이 무단 수정, 복사, 배포를 금합니다.

# “Inference 중심 AI 시대와 풀스택 인프라 전략의 본격화”

## <개요 및 주요 메시지>

- NVIDIA CEO Jensen Huang은 GTC 2026 키노트를 통해 **AI 산업이 기존의 대규모 모델 학습(Training) 중심 구조에서 실제 서비스 운영 및 에이전트 실행을 중심으로 하는 추론(Inference) 단계로 전환되고 있음을 명확히 제시**함. 해당 메시지는 단순 기술 발전을 넘어, AI 산업의 가치 창출 구조 자체가 재편되고 있음을 의미하는 것으로 해석 가능함.
- 특히 **AI 모델의 고도화가 일정 수준에 도달한 이후, 산업의 핵심 병목이 학습이 아닌 실제 서비스 단계의 반복적 추론 처리로 이동하고 있다는 점이 강조**됨. 이는 AI 경쟁의 기준이 모델 규모에서 벗어나 **latency, cost efficiency, operational scalability 등으로 이동**하고 있음을 시사하는 구조적 변화로 판단됨.
- NVIDIA는 이번 키노트를 통해 **자사의 전략적 정체성을 GPU 공급 기업에서 AI 데이터센터, 네트워크, 소프트웨어, 에이전트 플랫폼을 통합 제공하는 풀스택 인프라 기업으로 재정의**함. 이는 향후 AI 산업의 표준을 특정 칩이 아닌 통합 시스템과 플랫폼 차원에서 선점하려는 전략적 시도로 이해 가능함.
- 또한 **2027년 기준 AI 인프라 시장 규모를 약 1조 달러 수준으로 상향 제시**함으로써, 추론 수요 확대가 데이터센터 투자 사이클을 재가속화할 것이라는 전망을 함께 제시함. 이는 기존 시장 예측 대비 상당한 수준의 상향 조정으로, AI 에이전트 기반 서비스 확산을 전제로 한 구조적 성장 가정이 반영된 결과로 판단됨.

## 차세대 컴퓨팅 로드맵과 전략적 의미

- **Rubin AI Platform (2026)**: Blackwell 이후 차세대 아키텍처로, 추론 중심 워크로드에 최적화된 구조를 기반으로 latency 및 cost per token 개선을 목표로 설계된 GPU 플랫폼으로 정의됨.
- **Feynman Architecture (2028)**: Rubin 이후 후속 아키텍처로 제시되며, Hyperscaler의 중장기 인프라 투자 계획과 연계되는 전략적 로드맵 축으로 기능함.
- **Groq 기반 Inference System**: Vera Rubin 기반 시스템과 결합된 추론 특화 구조로, 대규모 AI 서비스 환경에서 응답속도 및 처리 효율 개선을 목표로 설계된 시스템으로 소개됨.

→ 상기 로드맵은 단순 제품 출시 계획이 아니라, 데이터센터 사업자의 2~3년 단위 CAPEX 계획을 선점하기 위한 전략적 신호로 해석됨. 특히 전력, 냉각, 네트워크를 포함한 인프라 전반의 재설계가 요구되는 환경에서, NVIDIA의 로드맵은 곧 시장 투자 방향을 규정하는 기준으로 작용할 가능성이 높음.

## AI 데이터센터 구조 변화와 AI Factory 개념

- **AI Factory 개념 제시**: 데이터센터를 단순 연산 및 저장 공간이 아닌 "토큰을 생산하는 공장"으로 정의하며, AI 인프라의 역할이 결과 생성 중심으로 이동하고 있음을 강조함.
- **Inference 중심 인프라 구조 전환**: 기존 Training 중심 클러스터에서 벗어나, latency, throughput, memory bandwidth, energy efficiency 중심으로 재설계된 데이터센터 구조 필요성 제시.
- **System-level 경쟁 구조 강화**: GPU 단품 성능이 아닌 rack 및 cluster 단위에서의 통합 성능이 핵심 경쟁 요소로 부상하며, 전력, 냉각, 네트워크 최적화가 중요 요소로 자리잡음.

→ 데이터센터 경쟁의 기준이 반도체 성능에서 시스템 설계 역량으로 이동하고 있음을 의미하며, AI 인프라 시장이 종합 산업으로 확장되고 있음을 보여주는 핵심 변화임.

## AI 소프트웨어 및 에이전트 플랫폼 확장

- **OpenClaw**: 오픈소스 기반 AI 에이전트 프레임워크로, 다양한 산업에서 에이전트 기반 자동화 및 작업 수행을 가능하게 하는 플랫폼.
- **NemoClaw**: 엔터프라이즈 환경에서 보안 및 운영 기능을 강화한 AI 에이전트 플랫폼으로, 실제 기업 환경에서의 적용을 전제로 설계된 구조.
- **Full-Stack Software 전략**: CUDA, NIM, NeMo 등 기존 소프트웨어 스택과 결합하여 모델 실행, 오케스트레이션, 보안, 에이전트 운영까지 포함하는 End-to-End 플랫폼 구조 구축.

→ AI 산업의 가치가 칩 성능에서 실행 환경과 운영 플랫폼으로 이동하고 있음을 반영하며, NVIDIA가 플랫폼 중심 Lock-in 구조를 강화하고 있음을 보여주는 전략적 방향임.

## AI 컴퓨팅 패러다임 전환

- **Training → Inference 전환**: AI 산업의 핵심 경쟁 축이 모델 학습에서 서비스 운영 단계의 추론 효율로 이동하고 있음을 공식화.
- **성능 지표 변화**: FLOPS 중심 평가에서 latency, cost per token, energy efficiency 중심으로 전환.
- **AI Agent 확산 영향**: 동일 모델 반복 실행 구조 증가로 인해 추론 workload 급증, 데이터센터 수요 구조 변화 유도.

→ 결과적으로 AI 산업의 수익 창출 구조가 연구 중심에서 서비스 운영 중심으로 이동하고 있으며, 이는 반도체 설계, 데이터센터 구조, 투자 전략 전반에 영향을 미치는 구조적 변화로 판단됨.

## 국내 AI 반도체 산업 시사점

- AI 반도체 시장의 경쟁 기준이 단순 연산 성능에서 실제 서비스 운영 효율로 이동하고 있다는 점은 국내 기업에도 직접적인 영향을 미치는 요소임. 따라서 latency, throughput, TCO 중심의 제품 전략 전환이 필수적임.
- 또한 NVIDIA조차 외부 기술을 결합하는 전략을 취하고 있다는 점은, 향후 시장이 단일 아키텍처 중심이 아닌 복수 기술 결합 구조로 발전할 가능성을 시사하며, 이는 국내 팹리스 및 IP 기업에게 협업 기회를 제공할 수 있음.
- 고객 요구 역시 기술 사양 중심에서 ROI 및 운영 효율 중심으로 이동하고 있어, 제품 설명 및 시장 접근 방식 역시 실제 적용 시나리오 기반으로 재구성할 필요가 있음.
- 나아가 AI 팩토리 개념은 반도체 단품 시장을 넘어 데이터센터 전체 생태계 확장을 의미하며, 이에 따라 국내 기업 역시 PoC, 시스템 통합, 미국 데이터센터 연계 전략을 강화해야 할 필요성이 존재함.