주요 동향(2) : ICT

1 GTC 2026, 에이전트 경제를 위한 AI 인프라 청사진

엔비디아, GPU 기업에서 AI 운영 계층 주도 기업으로

- AI 산업, 성능 향상의 한계와 추론 중심 전환 속에서 새로운 확장 경로 부상
  - AI 산업의 중심이 모델 학습에서 실시간 추론으로 이동하고, AI 간 협업을 통해 작업을 수행하는 에이전트 구조가 확산되면서 더 빠르고 대량의 연산을 동시에 처리해야 하는 수요가 급증
  - 기존 단일 GPU 중심 구조로는 속도와 처리량을 동시에 확보하는 데 한계가 부각되면서, 역할별로 특화된 칩과 소프트웨어를 결합하는 시스템 단위 접근이 새로운 경쟁 기준으로 부상
- GTC 2026, '에이전트 경제'로의 패러다임 전환 제시
  - 엔비디아가 주최하는 세계 최대 AI 컨퍼런스 GTC 2026이 3월 16~19일 미국 산호세에서 개최, 190개국 3만 명 이상 참가 및 1,000개 이상의 세션 진행으로 역대 최대 규모 기록
  - 젠슨 황 CEO는 차세대 AI 플랫폼 Vera Rubin, 에이전트 실행 환경 NemoClaw, 오픈 모델 연합 Nemotron Coalition 등 하드웨어, 소프트웨어, 생태계를 아우르는 설계 방향을 제시
- 개별 칩 경쟁을 넘어, 'AI 풀스택(Full-Stack)' 인프라 생태계로의 패러다임 진화
  - 엔비디아는 '더 빠른 칩'을 넘어, 칩-시스템-소프트웨어-생태계를 하나로 엮은 'AI 인프라 전체 설계도'를 제시하며, 에이전트 경제와 에이전틱 스케일링이 작동할 수 있는 기반을 구체화
  - AI를 단일 모델이나 앱이 아닌 모든 기업과 국가가 구축해야 할 '필수 인프라'로 새롭게 정의하고, 이를 5계층(5 Layer) 산업 시스템으로 조망
  - 최상위 계층의 애플리케이션이 실질적인 성과를 내기 위해서는 최하위인 에너지 계층부터 칩, 인프라, 모델 생태계까지 모든 층이 유기적으로 연쇄 작동해야 하는 구조임을 강조
  - 이번 GTC 2026의 모든 발표를 이 5개 계층을 동시에 끌어올리는 방향으로 구성함으로써, 단순한 신기술 발표를 넘어 AI 산업 구조를 재정의하고 업계의 논의 수준을 한 단계 격상시킨 것으로 평가

- (에너지) AI 지능(결과물)을 생성하는 모든 과정은 막대한 전력 소모와 발열 제어가 동반되는 물리적 작업으로 인프라가 확보될 수 있는 전력 규모가 곧 AI 생산량의 상한선을 결정
- (칩) 에너지를 연산으로 변환하는 프로세서 계층으로, 칩의 병렬 처리 능력, 메모리 대역폭, 에너지 효율에 따라 AI의 확장 속도와 운용 비용이 결정
- (인프라) 수만 개의 프로세서를 하나의 거대한 시스템으로 묶는 AI 팩토리 계층으로, 토지, 전력 배분, 냉각, 네트워킹, 운영 자동화까지 포함하며, 지능을 생산하는 공장으로 설계
- (모델) 언어 모델뿐 아니라 단백질 구조 예측, 화학 반응 시뮬레이션, 물리 환경 모사, 로보틱스 제어, 자율주행 판단 등 다양한 지식을 이해하고 생성하는 모델이 해당 계층에 포함
- (애플리케이션) 경제적 가치를 창출하는 최종 계층으로 업무용 에이전트와 자율주행차(기계에 탑재되는 AI), 휴머노이드 로봇(신체에 구현된 AI) 등이 포함

에너지에서 애플리케이션까지, 엔비디아가 주도하는 AI 풀스택 구조 전환

가. 에너지: AI 생산량의 상한선을 결정하는 전력 인프라의 부상

○ 전력 효율화부터 전력망 연계까지, 엔비디아의 전력 관리 기술 체계

- 젠슨 황은 에너지를 AI 인프라의 가장 근본적인 제약으로 규정하며, 확보할 수 있는 전력 규모가 곧 AI 생산량의 한계를 결정한다고 강조
- Vera Rubin 랙에는 전력 사용량의 급격한 변동을 자동으로 완화하는 기술을 적용해, 순간 최대 전력 소비를 약 25% 줄이고 전력망에 가해지는 부담을 경감
- 워크로드에 따라 랙별 전력을 실시간으로 분배하는 소프트웨어 'DSX Max-Q'를 통해, 동일한 전력 환경에서 최대 30% 더 많은 GPU를 가동할 수 있는 관리 체계를 공개
- 나아가 AI 팩토리가 전력 사용량을 전력망 상황에 맞춰 스스로 조절하는 소프트웨어 'DSX Flex'를 공개하며, 활용되지 못하던 유휴 전력 100GW를 추가 확보할 수 있는 방안을 제시

○ AI 팩토리에 전력을 공급하는 글로벌 에너지 기업과의 협력 확대

- 미국 내 AI 인프라용 전력 연결 대기 물량이 200GW 이상, 장비 수주 잔고가 3,000억 달러를 초과하는 등 전력 공급이 AI 인프라 확장의 최대 제약 요인으로 부각

- GE Vernova, Hitachi, Siemens Energy 등 글로벌 에너지 기업이 엔비디아 DSX 아키텍처와의 기술 통합을 발표하며, AI 팩토리에 안정적으로 전력을 공급하는 협력 체계가 형성

나. 칩: GPU 단품에서 다중 칩 결합형 추론 시스템으로 전환

● Vera Rubin, GPU 하나가 아닌 7종 칩 결합형 추론 시스템 공개

- 젠슨 황은 기조연설에서 신형 GPU 한 개가 아닌 7종의 칩을 하나의 슈퍼컴퓨터로 결합한 차세대 AI 플랫폼 Vera Rubin을 공개하며, 이를 에이전틱 AI 시대를 여는 세대적 도약으로 규정
- 기존 GPU 중심 구성에서 벗어나 추론 전용 LPU, 에이전트 작업 전용 CPU, 대화 맥락 저장 전용 DPU, 초고속 네트워킹 칩까지 7종의 칩이 역할을 나눠 수행하는 구조를 공개
- 이 칩 조합으로 기존 GPU 아키텍처인 블랙웰 대비 와트당 추론 성능 10배, 토큰당 비용 1/10을 달성하며, 에이전트 환경에 필요한 고성능·고효율·저비용 조건을 동시에 구현
- 이러한 구조 변화와 성능 개선은 AI 칩 경쟁의 기본 단위가 '개별 칩 성능'에서 '칩 조합 설계 역량'으로 이동하고 있음을 보여주는 신호

● 고처리량 연산과 초저지연 응답을 분리한 추론 구조

- 이날 함께 공개된 추론 전용 칩 Groq 3 LPU는 대용량 HBM을 탑재한 Rubin GPU와 결합해, 하나의 랙에서 고처리량과 초저지연을 동시에 구현하는 아키텍처로 소개
- 대용량의 연산을 처리하는 초기 연산 단계는 Rubin GPU가, 응답을 지연 없이 빠르게 생성하는 단계는 Groq LPU가 담당하는 역할 분리형 추론 구조 제시
- 이 발표는 Cerebras 등 추론 전문기업과의 경쟁을 겨냥한 전략으로 해석되며, 엔비디아가 학습에 이어 추론 시장까지 주도권을 확보하려는 움직임으로 평가
- 젠슨 황은 추론 단계가 AI 산업의 핵심 경쟁 영역으로 전환되고 있음을 강조하며, GPU와 LPU의 결합 구조가 향후 대규모 수익 기회를 뒷받침할 핵심 기반이라고 강조

● GPU를 넘어 CPU 시장까지 확장하며 에이전트 시대의 칩 포트폴리오 구축

- 에이전트는 코드를 실행하고 도구를 호출하는 과정에서 GPU가 아닌 CPU 연산을 필요로 하며, 기존에는 이 영역이 인텔 등 x86 CPU에 의존해 AI 워크로드의 병목으로 지적
- 엔비디아는 에이전트 연산에 특화된 Vera CPU를 독립 제품으로 공개하며, 인텔 x86 대비 메모리 대역폭 3배, 에너지 효율 2배를 확보해 CPU 시장 본격 진출을 선언
- 에이전트가 대화 과정에서 축적하는 맥락 데이터 전용 스토리지 칩(BlueField-4 STX)도 함께 공개하며, 엔비디아의 칩 사업 영역이 GPU를 넘어 추론, 연산, 저장 전반으로 확대

○ 삼성전자·SK하이닉스, 엔비디아 차세대 칩의 핵심 메모리 공급사로 부각

- Vera Rubin에 탑재되는 차세대 메모리 HBM4의 핵심 공급사로 삼성전자와 SK하이닉스가 부각되며, GTC 2026 현장에서 양사 모두 자사 기술을 발표
- SK하이닉스는 올해 엔비디아 HBM4 물량의 약 3분의 2를 공급할 것으로 전망되며, 최태원 SK그룹 회장이 GTC 현장에서 젠슨 황 CEO와 직접 면담해 HBM 공급 확대 및 기술 협력을 논의
- 삼성전자는 글로벌 최초로 HBM4 양산 및 출하에 성공하며 시장 선점에 나섰고, GTC에서 AI 아키텍처를 위한 메모리·스토리지 설계 방향을 발표

다. 인프라: 데이터 저장소에서 토큰 생산 공장으로 전환되는 AI 데이터센터

○ 저장 중심 데이터센터에서 토큰 생산형 AI 팩토리로의 전환

- 엔비디아는 지능 토큰을 새로운 화폐로, AI 팩토리를 이를 생산하는 인프라로 정의하며 데이터센터의 역할이 '정보 저장'에서 '지능 생산'으로 전환되고 있음을 명확히 제시
- 기조연설에서도 데이터센터가 과거 파일을 저장하는 공간이었다면 현재는 토큰을 생산하는 공장으로 전환되고 있음을 강조하며, 핵심 성과 지표가 '와트당 토큰 생산량'으로 변화하고 있다고 설명
- 엔비디아는 2027년까지 최소 1조 달러(약 1,490조 원) 규모의 인프라 투자 기회를 전망하며, 인류 역사상 최대 규모의 인프라 구축이 될 것이라는 관측을 제시

● 서버 단위를 넘어 랙-네트워크-스토리지를 통합하는 시스템 경쟁

- 엔비디아는 Vera Rubin의 기반이 되는 3세대 MGX 랙 구조를 공개하며, 케이블, 호스, 팬을 제거한 단순한 설계를 통해 서버 부품 조립 시간을 약 2시간에서 5분으로 크게 단축
- 네트워킹 랙(Spectrum-6 SPX)에도 칩과 광통신 부품을 하나로 묶는 기술(CPO)을 적용해 기존 대비 전력 효율은 약 5배 높이고, 장애 발생 시 복구 능력도 약 10배 향상
- 이러한 변화는 AI 인프라 경쟁의 중심이 개별 서버 성능에서 랙, POD 등 여러 장비를 결합한 대규모 시스템 단위로 확대되고 있음을 보여주는 흐름

● 디지털 트윈 기반 사전 설계로 구축 이전에 운영을 검증하는 AI 팩토리

- 젠슨 황은 AI 팩토리를 설계부터 구축·운영까지 표준화할 수 있는 기반으로, 기준 모델(Vera Rubin DSX AI Factory)과 디지털 설계 도구(Omniverse DSX Blueprint)를 함께 출시
- DSX Air 플랫폼을 통해 실제 장비 설치 이전에 AI 팩토리 전체를 소프트웨어로 구현하고 시험할 수 있도록 하며, 구축 기간을 수개월에서 수일 수준으로 단축 가능
- CoreWeave는 이 기술을 활용해 클라우드에서 AI 팩토리를 가상으로 구축하고 운영을 테스트하는 등 실제 적용 사례를 공개
- Cadence, Siemens 등 설계 기업과 Caterpillar 등 산업 기업이 참여하면서, AI 팩토리 구축이 개별 기업 프로젝트를 넘어 산업 생태계 단위로 확산되고 있음이 가시화

● 지상에서 우주·통신망으로 확장되는 AI 팩토리의 물리적 범위

- 엔비디아는 궤도 데이터센터용 AI 모듈 'Space-1 Vera Rubin Module'을 발표하며, 다수 기업과 협력해 궤도 환경에서 H100 대비 25배 수준의 AI 연산 성능을 구현하겠다는 계획을 제시
- 기조연설에서는 우주 컴퓨팅을 새로운 인프라 확장 영역으로 제시하며, AI 팩토리의 물리적 범위가 지상을 넘어 궤도까지 확대되고 있음을 강조
- 한편, 통신 분야에서는 Nokia가 RTX PRO 4500 Blackwell을 적용한 엣지 AI 분산 컴퓨팅 인프라 구축 계획을 발표하며, AI 연산 인프라가 통신망과 엣지까지 확장되는 추세

라. 모델: 에이전트 실행 환경 표준을 노리는 엔비디아의 운영체제 전략

- Nemotron 연합, 에이전트 생태계 확장을 위한 개방형 모델 기반 구축
  - 엔비디아는 글로벌 오픈 모델 연합 'Nemotron Coalition' 결성을 발표하며, 오픈 모델을 중심으로 전 세계 개발자가 참여하는 AI 생태계 확대 전략 제시
  - Black Forest Labs, Cursor, LangChain, Mistral AI, Perplexity 등 8개 AI 연구소가 창립 멤버로 참여해, 공동으로 고성능 오픈 모델을 개발하는 협력 체계를 구축
  - Nemotron(언어), Cosmos(비전), GR00T(로보틱스), Alpamayo(자율주행), BioNeMo(바이오), Earth2(기후) 등 6개 영역에 걸친 오픈 모델 생태계 구축 추진
  - 참여 기업들이 잇따라 협력 의사를 밝히며 오픈 모델을 중심으로 엔비디아의 AI 플랫폼 생태계 확장 전략이 본격화되는 양상
- Dynamo, 수만 개 GPU를 실시간 배분하는 AI 팩토리 운영 소프트웨어
  - 엔비디아는 오픈소스 추론 소프트웨어 Dynamo 1.0 정식 출시를 발표하며, AI 팩토리의 연산 자원을 통합 관리하는 소프트웨어 계층을 제시
  - AI 팩토리 내 수만 개 GPU와 메모리 자원을 실시간으로 배분 및 조율하는 기능을 통해, 블랙웰 기반 추론 성능을 최대 7배까지 향상
  - AWS, Azure, Google Cloud, Oracle Cloud 등 주요 클라우드 사업자가 이미 Dynamo를 도입했으며, CoreWeave, Together AI 등 협력 클라우드로 확산
  - Cursor, Perplexity 등 AI 기업과 Pinterest, PayPal 등 글로벌 기업이 실제 서비스에 적용하며, 상용 환경에서의 활용이 확대되는 흐름
- 에이전트 실행 환경의 표준을 구축하는 오픈소스 플랫폼 'NemoClaw'
  - 젠슨 황은 오픈소스 에이전트 플랫폼 OpenClaw를 개인과 기업 환경에서 AI 에이전트를 실행하기 위한 기본 플랫폼으로 소개하며, 에이전트 실행 환경을 표준화하려는 방향을 강조
  - OpenClaw 위에 Nemotron 모델과 OpenShell 보안 실행 환경을 통합한 'NemoClaw' 스택을 공개하며, 모델-실행-보안을 하나의 구조로 묶은 에이전트 실행 환경을 제시
  - Cisco, CrowdStrike, Microsoft Security, TrendAI 등 주요 보안 기업이 OpenShell 호환을 발표하며, 에이전트 실행 환경에 맞는 보안 체계가 함께 형성되는 흐름

마. 애플리케이션: 기업용 AI에서 Physical AI까지, 에이전트 경제의 수익화 영역

- **(엔터프라이즈 AI)** 챗봇을 넘어 실제 업무를 수행하는 기업용 AI 에이전트 확산
  - 젠슨 황은 기업용 소프트웨어 산업이 에이전트 중심 플랫폼으로 전환되고 있으며, IT 산업의 새로운 확장 국면이 시작되고 있다고 진단
  - Adobe, Salesforce, SAP, ServiceNow, Atlassian 등 16개 글로벌 소프트웨어 기업이 엔비디아 Agent Toolkit 채택을 발표하며, 에이전트 기반 소프트웨어 전환이 본격화
  - IQVIA는 상위 20개 제약사 고객 환경에 엔비디아 기반 에이전트를 150개 이상 배치했다고 밝히며, 에이전트가 실험 단계를 넘어 실제 업무 도구로 활용되고 있음을 확인
- **(피지컬 AI)** 로봇과 산업 장비로 확장되는 피지컬 AI 플랫폼
  - 젠슨 황은 “피지컬 AI가 도래했으며, 모든 산업 기업이 로보틱스 기업이 될 것”이라고 선언하며, 디지털 에이전트를 물리 세계로 확장하는 것이 AI의 다음 개척지임을 공식화
  - GTC 2026에서 월드 모델 ‘Cosmos 3’와 로봇 모델 ‘GR00T N2’를 공개하며, 물리 환경을 이해하고 제어하는 월드 모델 기반 AI 체계를 구체화
  - ABB, FANUC, KUKA 등 글로벌 산업 로봇 주요 기업이 엔비디아 Omniverse, Isaac 통합을 발표하며, 기존 산업 로봇 시스템에 엔비디아 AI를 결합하는 흐름이 확대
  - 1X, AGIBOT, Boston Dynamics, Figure 등 휴머노이드 기업이 엔비디아 기반 개발 현황을 공개했으며, Skild AI는 Foxconn과 협력해 블랙웰 기반 생산라인 자동화 적용 사례를 제시
- **(제조·물류)** 디지털 트윈 기반 설계·제조·물류 자동화 확산
  - Cadence, Dassault, Siemens 등 산업 소프트웨어 기업이 엔비디아 GPU 가속 도구를 자사 설계·시뮬레이션 제품에 통합하며, 산업 설계 영역에서 엔비디아 인프라 활용이 확대
  - (자동차 설계) Honda는 엔비디아 기반 환경에서 공기역학 시뮬레이션을 CPU 대비 34배 빠르게 수행하고 있으며, JLR, Mercedes-Benz도 엔비디아 인프라 기반 설계 환경을 도입
  - (디지털 트윈) Siemens는 엔비디아 Omniverse 기반 ‘Digital Twin Composer’를 출시하며, Foxconn, HD Hyundai, PepsiCo 등이 엔비디아 기반 가상 공장 구축을 진행 중이라고 발표

- (물류 자동화) KION Group은 Omniverse 기반 물류 디지털 트윈과 Jetson 기반 자율 지게차 적용 사례를 공개하며, 시뮬레이션에서 실제 자동화로 이어지는 구조를 제시

○ **(자율주행)** Uber 협력을 계기로 서비스망까지 확장되는 자율주행 플랫폼 영향력

- 젠슨 황은 자율주행이 로보틱스 산업의 핵심 영역으로 성장하고 있으며, 대규모 시장 형성이 본격화되고 있음을 조명
- 엔비디아는 자율주행 칩, 센서, 안전 시스템을 통합한 DRIVE Hyperion 플랫폼을 기반으로, 완성차 업체에는 양산차 탑재를, 모빌리티 기업에는 로보택시 서비스망 구축을 동시에 확대
- (Uber) 엔비디아와 공동으로 DRIVE Hyperion 기반 로보택시 네트워크를 구축하며, 2027년 LA·샌프란시스코를 시작으로 2028년까지 4개 대륙 28개 도시로 확장할 계획을 공개
- (현대차·기아) DRIVE Hyperion 기반 전략적 파트너십 확대를 통해, 레벨2+ 양산차 탑재부터 합작법인 '모셔널(Motional)'을 통한 레벨4 로보택시 개발까지 전 단계 협력 구조를 구축
- (Bolt·Grab·Lyft) 글로벌 주요 모빌리티 플랫폼 기업들도 잇따라 DRIVE Hyperion 채택을 발표하며, 자율주행 서비스가 특정 기업을 넘어 업계 전반으로 확산

출처 : NVIDIA Newsroom 외(2026.3.)
https://nvidianews.nvidia.com/news/nvidia-ceo-jensen-huang-and-global-technology-leaders-to-showcase-age-of-ai-at-gtc-2026
https://nvidianews.nvidia.com/news/nvidia-vera-rubin-platform
https://nvidianews.nvidia.com/news/nvidia-releases-vera-rubin-dsx-ai-factory-reference-design-and-omniverse-dsx-digital-twin-blueprint-with-broad-industry-support
https://nvidianews.nvidia.com/news/space-computing
https://nvidianews.nvidia.com/news/nvidia-launches-nemotron-coalition-of-leading-global-ai-labs-to-advance-open-frontier-models
https://nvidianews.nvidia.com/news/nvidia-announces-nemoclaw
https://nvidianews.nvidia.com/news/dynamo-1-0
https://nvidianews.nvidia.com/news/ai-agents
https://nvidianews.nvidia.com/news/nvidia-and-global-industrial-software-giants-bring-design-engineering-and-manufacturing-into-the-ai-era
https://nvidianews.nvidia.com/news/nvidia-and-global-robotics-leaders-take-physical-ai-to-the-real-world
https://nvidianews.nvidia.com/news/drive-hyperion-level-4
https://nvidianews.nvidia.com/news/nvidia-announces-open-physical-ai-data-factory-blueprint-to-accelerate-robotics-vision-ai-agents-and-autonomous-vehicle-development
https://blogs.nvidia.com/blog/ai-5-layer-cake/
https://www.reuters.com/world/asia-pacific/nvidia-ceo-set-reveal-new-chips-software-ai-megaconference-gtc-2026-03-16/
https://www.tomshardware.com/news/live/nvidia-gtc-2026-keynote-live-blog-jensen-huang