2 모델 경쟁을 넘어, AI 풀스택 생태계 주도권 경쟁

모델 경쟁에서 풀스택 통합 경쟁으로의 확장

- AI 경쟁의 무게중심이 '개별 모델'의 성능 중심에서 '풀스택 통합 역량'까지 확장
  - 그동안 AI 경쟁의 핵심은 어느 기업이 더 우수한 단일 모델을 보유했는가에 집중되었으나, 모델 성능이 빠르게 평준화되며 모델 단독으로는 경쟁우위 확보가 어려운 환경이 도래
  - 이에 따라 글로벌 빅테크는 칩·인프라·모델·서비스를 분리 공급하던 방식을 넘어, 한 생태계 안에서 이를 통합 운영하는 풀스택 사업자로의 전환을 본격 추진
  - 이러한 흐름은 '26년 3~4월 글로벌 빅테크 키노트에서 잘 드러났는데, 엔비디아·화웨이·구글이 공통적으로 '풀스택 운영자'로의 포지셔닝 전환을 선언
- 풀스택 경쟁은 단일 기업 차원을 넘어 산업·국가 차원의 생태계 확장 경쟁으로 진화
  - 풀스택 경쟁은 더 좋은 모델을 만드는 단일 기업의 기술 경쟁이 아니라, 자사 생태계 안으로 개발자·기업 고객·산업 사용자를 얼마나 폭넓게 끌어들이는가의 생태계 확장 경쟁으로 변화
  - 美·中 정부도 자국 풀스택 자립을 국가 전략 의제로 격상시키며 기업의 풀스택 경쟁에 산업 정책·지정학 변수가 결합, 풀스택 경쟁은 기업과 국가 차원이 함께 작동하는 다층 구도로 확대
  - 글로벌 풀스택 경쟁이 새로운 표준 경쟁 구도로 자리잡는 가운데, 엔비디아·구글·화웨이는 각기 다른 출발점·자원·전략으로 풀스택을 구축하는 대표 경로를 보여주는 사례로 부상

풀스택 통합이 새로운 경쟁 축으로 부상한 배경

1. 왜 지금 풀스택 경쟁이 중요해졌는가?

- AI 산업의 무게중심이 학습에서 추론으로 이동, 인프라가 새로운 비용 변수로 부상
  - AI 산업이 모델 학습 중심의 연구·실험 단계를 넘어, 대규모 이용자 기반의 서비스에 AI를 상시 적용해 실제 수익을 창출하는 운영 단계로 진입
  - 이 단계에서는 한 번의 학습 비용보다 24시간 작동하는 추론 비용이 사업의 수익성을 좌우하며, 인프라 효율과 응답 속도가 핵심 경쟁 변수로 부상
  - 美 4대 빅테크의 '26년 자본 지출 계획이 전년 대비 약 2배 늘어난 7,000억 달러 규모로 책정되었고, 이 중 상당 부분이 추론 인프라에 투입될 전망 (Tech Insider, '26.4.)

● 외부 의존 구조의 한계 속, 글로벌 빅테크는 풀스택 내재화를 확대

- 외부의 칩·인프라·모델에 의존하는 구조는 계층 사이 호환·조정 비용이 누적되어, 같은 작업을 처리하는 데 통합 사업자 대비 더 많은 비용과 시간 소요
- 또한 외부 조합 방식은 각 계층 사업자의 가격 변동과 공급 정책 변화에 직접 노출되어, 비용 예측과 공급 안정 모두에서 자체 풀스택 보유 사업자 대비 불리한 위치
- 모델 자체의 성능 격차도 줄어들면서, '26년 경쟁의 핵심은 모델 한 가지가 아니라 칩에서 서비스까지 이어지는 흐름을 얼마나 매끄럽게 운영하는가로 이동한 상황
- 이에 빅테크는 자체 칩 개발과 자사 서비스 통합을 동시에 추진, 외부에 의존하지 않고 풀스택 전체를 자사 통제 아래 두는 방향으로 사업 모델을 재편

2. 풀스택 통합이 만드는 경쟁우위는 무엇인가?

● (운영 효율) 칩에서 인프라까지 자체 통제로 추론 비용·전력 효율을 동시에 개선

- 풀스택 사업자는 AI 학습·추론 작업에 맞춰 자체 칩과 데이터센터를 함께 설계할 수 있어, 외부 GPU·인프라를 조합해 쓰는 사업자 대비 같은 전력으로 더 많은 추론을 처리하는 단위 경제성 확보
- 풀스택 사업자는 외부 조합 사업자 대비 비용 절감 효과가 누적·확대되며, MIT 연구는 풀스택 차원의 인프라·알고리즘 효율 개선이 추론 비용을 연간 5~10배까지 떨어뜨린다고 분석 (MIT FutureTech, '25.11.)
- 또한 자체 칩 보유 사업자는 글로벌 GPU 공급 부족 환경에서도 외부 공급 변동성에 영향받지 않고, 자사 서비스 수요에 맞춰 생산·배치 우선순위를 조정 가능

● (생태계 영향력) 모델·개발도구·서비스 데이터까지 자사 안에서 순환시키는 락인 구조 형성

- 풀스택 사업자는 자체 모델과 함께 학습·배포·관리 도구까지 같은 환경에서 제공, 개발자가 한 번 진입하면 별도 도구 학습 없이 즉시 작업 가능한 환경 조성
- 개발자 생태계는 한 번 익숙해지면 다른 풀스택으로 이동할 때 기존 작업물의 재구현·재학습 부담이 발생, 시간이 지날수록 자사 환경에 머무는 락인 효과가 강화되는 구조
- 여기에 검색·메일·문서 등 자사 서비스 접점에서 쌓이는 사용자 데이터는 외부 사업자가 접근하기 어려운 차별화 자산으로, 데이터 수집부터 모델 재학습까지 한 사업자 안에서 처리하는 선순환 구조 완성

- 결국 풀스택 경쟁의 승부는 칩·모델 같은 개별 기술의 우열이 아니라, 칩에서 서비스까지 이어지는 흐름을 자사 안으로 얼마나 매끄럽게 연결하고 자사에 머물게 하는가에서 갈리는 양상

○ 풀스택 경쟁력은 단일 지표가 아닌 평가 요소의 종합으로 결정

- 풀스택 사업자의 경쟁력은 운영 효율과 생태계 영향력이라는 두 축 안에서 여러 평가 요소가 결합된 결과로 평가되며, 한 계층의 강점만으로는 종합 우위 확보가 불가
- 평가 요소 모두에서 1위를 확보한 사업자는 부재한 상황으로, 각 풀스택 사업자는 출발점·자원에 따라 특정 요소에 집중해 차별화된 경쟁 위치를 확보

글로벌 3사의 풀스택 전략 사례

1. 엔비디아: AI 팩토리 중심의 글로벌 표준형 풀스택

(1) (구조) 칩에서 AI 팩토리까지 단일 사업자가 통합 설계·공급

○ (AI 칩) Vera Rubin 7종 칩을 단일 시스템으로 결합한 통합 연산 인프라

- 엔비디아는 GTC 2026 키노트에서 연산·통신·추론 칩 7종을 하나의 시스템으로 통합한 Vera Rubin을 공개하며, 칩 한 장이 아닌 인프라 전체를 설계·공급하는 새로운 사업 단위 제시
- Vera Rubin은 직전 세대 Blackwell 대비 동일한 AI 작업을 처리하는 데 비용은 약 1/10, 필요한 GPU 수는 약 1/4 수준으로 절감, AI 팩토리의 운영 비용과 자원 투입을 동시에 낮추는 핵심 요인으로 부상
- 美 빅테크 클라우드 4사(AWS·Azure·구글·OCI)가 '26년 하반기 Vera Rubin 플랫폼 동시 도입을 예고하면서, 엔비디아의 시스템이 글로벌 AI 인프라의 사실상 표준으로 정착하는 양상

○ (인프라) 표준 설계도와 가상 시뮬레이션으로 AI 팩토리 도입 진입 장벽 완화

- AI 팩토리 구축에 필요한 전력·냉각·네트워크·연산 사양을 하나의 검증된 구성으로 묶은 표준 설계도 'Vera Rubin DSX'를 공개, 운영자가 자체 설계 부담 없이 즉시 도입할 수 있는 환경 구축
- 여기에 실제 구축 전 가상 환경에서 AI 팩토리를 미리 작동시켜 보는 디지털 트윈 'Omniverse DSX'를 결합, 성능·발열·전력 문제를 사전 점검해 초기 도입 과정의 시행착오 축소
- 표준 설계도와 사전 검증 체계가 함께 제공되면서, 엔비디아 인증 클라우드 파트너의 AI 인프라 구축 규모가 1년 만에 약 2배로 확대되는 등 표준화 역량이 시장 침투 속도를 끌어올리는 동력으로 작동

● (모델) 외부 모델 동맹과 추론 OS로 모델 계층까지 자사 풀스택 영향권에 편입

- 엔비디아는 칩과 인프라에 비해 자체 AI 모델 경쟁력은 상대적으로 약하지만, 외부 모델 협력* 확대와 추론 운영 소프트웨어 강화를 통해 자사 한계를 보완
- GTC 2026에서 추론 전용 OS 'Dynamo 1.0'을 정식 출시하며, 외부·자체 모델을 가리지 않고 다양한 하드웨어 자원을 자동 배분해 추론 처리량과 토큰당 비용을 동시에 개선하는 운영 표준 제시

● (응용) 산업 SW·자율주행·소버린 AI까지, 응용 영역 전방위 확장

- 엔비디아는 칩·인프라·모델 위에 올라가는 응용 계층까지 동시에 공략하며, 산업 SW, 자율주행, 국가 단위 AI 인프라로 침투 영역을 넓히고 풀스택의 시장 도달 범위를 확대
  · (산업 SW) 글로벌 설계 SW 기업과 자동차 OEM이 차량 설계·시뮬레이션 환경에 엔비디아 인프라를 도입하면서, 제품 설계 단계부터 풀스택이 적용되고 제조업의 AI 인프라 의존도 심화
  · (자율주행) 글로벌 OEM이 자율주행 플랫폼 'DRIVE Hyperion'을 채택하고, '27년부터 Uber와 4대륙 단위 로보택시 네트워크 가동을 추진하면서 차량 단위를 넘어 도시 운행 인프라로 확장
  · (소버린 AI) HPE와의 협력을 통해 국가 단위 AI 인프라 구축을 발표하면서, 풀스택 사업자가 단순 칩 공급자를 넘어 국가 인프라 운영자로 역할이 격상되는 사례 등장

(2) (경쟁우위) 후발 진영이 단기간에 따라잡기 어려운 세 가지 자산 보유

● (HW·SW 통합 최적화) 칩·SW 일체 설계로 성능·비용 우위 확보

- 초기 설계 단계부터 칩과 SW를 함께 최적화하는 'Extreme Co-Design' 방식을 채택해, 칩·SW를 외부에서 조달하거나 별도 개발하는 사업자가 겪는 호환·조정 과정의 손실을 사전 차단
- 엔비디아는 GPU 중심 구조를 넘어 CPU·DPU·이더넷 스위치까지 자체 반도체 범위를 확장, 외부 공급사 없이 칩 간 통신과 전력, 연산 흐름을 통합적으로 설계할 수 있는 기반 마련
- 이러한 통합 설계 능력이 칩당 처리 효율을 끌어올려, AI 팩토리 사업자가 동일한 사업 규모 운영에 필요한 자본·전력·공간 부담을 동시에 절감하는 결과로 연결

● (CUDA 생태계) 약 20년간 굳어진 SW 표준 위치가 후발 SW 추격의 최대 장벽

- CUDA는 약 20년간 글로벌 AI 개발자가 사용해 온 사실상의 표준 SW 플랫폼으로, 후발 진영이 추격하려면 개발자 커뮤니티와 개발 생태계까지 새로 구축해야 하는 진입 장벽으로 작용

- 글로벌 AI 모델·라이브러리·교육 자료 다수가 CUDA 환경을 기준으로 축적되어 있어, 개발자가 다른 칩으로 이동할 경우 기존에 작성한 모델·코드를 다시 구현하고 새 도구를 학습해야 하는 부담 발생
- 화웨이가 자체 SW 스택 'CANN Next'를 CUDA 호환 환경으로 출시한 점은, 후발 진영이 자체 SW만으로는 CUDA 환경을 우회하지 못하고 호환에 의존할 수밖에 없다는 점을 입증

○ (AI 팩토리 단위 공급) 칩 단품에서 인프라 단위로 사업 모델 전환

- 엔비디아는 GPU 단품 판매에서 AI 팩토리 단위 인프라 공급으로 사업 모델을 전환, 칩 공급사를 넘어 글로벌 AI 인프라 운영의 표준 공급자로 위치 격상
- 엔비디아는 '27년까지 관련 누적 매출 약 1조 달러 달성 전망을 제시*, 시장 분석기관 Futurum도 이를 "팹리스 반도체 기업에서 글로벌 인프라 공급자로의 진화"로 평가

(3) (위험 요인과 변수) 외부 시장 환경(지정학)과 고객 구조

○ (지정학적 변수) 글로벌 핵심 공급자 위치, 美·中 기술 충돌의 직접 표적으로 노출

- 엔비디아는 글로벌 AI 인프라의 핵심 공급자로서 美·中 기술 갈등의 직접 영향권에 속해, 美 AI 반도체 수출 통제와 中 자국 칩 우선 정책이 맞물리며 中 시장 입지가 빠르게 약화
- 엔비디아의 中 AI 가속기 시장 점유율은 '23년 약 95%에서 '25년 약 55%로 하락했으며, 이 공백을 中 자국산 칩이 대체하면서 단기간에 기존 지위를 회복하기 어려운 시장 구도가 형성
- 엔비디아의 중국 시장 접근 차단은 현지 경쟁사의 성장을 촉진하는 결과로 이어져, 단일 사업자의 풀스택 우위도 진영 분화 환경에서는 시장 도달 범위의 한계에 직면

○ (추론 시장 분화) 빠른 추론 영역, GPU 단독 우위가 흔들리는 새 격전지로 부상

- 엔비디아의 우위는 GPU 중심 학습·추론 인프라에서 형성됐으나, 추론 시장이 용도별 아키텍처 경쟁으로 분화하면서 GPU 단독 우위가 흔들리는 영역이 등장
- 특히 저지연 응답이 필요한 빠른 추론 영역에서는 LPU·웨이퍼 스케일 칩 등 전용 아키텍처가 우위로 평가되면서 전용 칩 진영의 점유율이 60~80%까지 확대될 수 있다는 분석 제기(Cerebras, '26.)
- 이에 엔비디아는 '25년 12월, 추론 전문 사업자 Groq와 비독점 추론 라이센스 계약을 체결, GPU 중심 풀스택만으로는 대응이 어려운 추론 특화 영역을 외부 기술 흡수로 보완

● (고객 측면 변수) 빅테크의 '단일 풀스택 의존 회피' 흐름 본격화

- 엔비디아가 글로벌 핵심 공급자 위치를 강화할수록 빅테크 입장에서는 의존 리스크도 함께 커지면서, 다중 공급망을 확보하려는 움직임이 엔비디아 풀스택 우위에 대한 압박 요인으로 부상
- 이 같은 공급망 분산 흐름은 후발 칩 진영으로의 자금 유입까지 견인하며, Cerebras·MatX 등 AI 칩 스타트업의 성장 기반을 넓히며 엔비디아 중심 풀스택 구도에 균열 요인으로 작용

2. 구글: 플랫폼형 풀스택과 서비스 접점 확장

(1) (구조) TPU에서 30억 사용자 접점까지 플랫폼형 통합

● (AI 칩) 학습·추론 칩 분리 설계로 작업 유형별 운영 효율 차별화

- AI 작업이 학습과 추론으로 분화되면서 단일 칩 구조의 한계가 부각되는 가운데, 구글은 학습용·추론용을 별도 설계한 8세대 TPU를 공개하며 작업 유형별 전용 칩 설계 흐름을 선도
- 학습용 'TPU 8t'는 성능과 전력 효율을 높여 대규모 학습 비용을 낮추고, 추론용 TPU 8i는 메모리를 확대한 설계로 다수의 AI 에이전트가 동시에 작동하는 환경에서도 빠르고 안정적 추론 지원
- 단일 칩으로 학습·추론을 모두 처리하는 엔비디아 GPU 방식과 달리 구글은 작업 유형별로 칩을 분리해 운영 효율을 차별화, 추론 시장 분화 흐름에서 작업 유형 맞춤 설계의 우위 확보

● (인프라) 자체 AI Hypercomputer에 외부 GPU 옵션을 더한 개방형 인프라 공급

- AI 인프라는 운영자가 칩·저장시스템·네트워크·SW를 직접 조합해야 하는 부담이 큰 진입 장벽으로 지적되어 왔으나, 구글은 이를 'AI Hypercomputer'로 묶어 제공하며 인프라 도입 절차를 단순화
- 구글은 자체 TPU뿐 아니라 엔비디아 Vera Rubin 등 외부 GPU 옵션도 함께 제공하며, 자사 칩을 고집하는 대신 고객 워크로드에 맞춰 인프라를 선택할 수 있는 개방형 운영 전략 채택
- '26년 머신러닝 연산 투자의 절반 이상을 클라우드 사업에 배분하면서, 자사 풀스택을 외부 사업자에게도 공급하는 인프라 운영자로 위상 확대

● (모델) 자사·외부 모델을 같은 플랫폼에서 운영하는 모델 중립 구조

- 구글은 자체 모델 Gemini뿐 아니라 외부 모델까지 같은 플랫폼에서 호출·운영할 수 있는 구조를 구축, 경쟁사 앤트로픽의 Claude도 자사 플랫폼에서 호출 가능한 '모델 중립' 전략 강화

- 기업용 AI 에이전트 운영 환경 'Gemini Enterprise'도 함께 출시해, 모델 선택부터 에이전트 제작· 실행·보안 관리까지 하나의 환경에서 처리할 수 있도록 지원하며 기업의 도입·운영 복잡도 완화
- 그 결과 '26년 1분기 자체 모델의 API 토큰 처리량이 직전 분기 대비 약 60% 증가하면서, 자사·외부 모델을 함께 묶은 모델 중립 플랫폼이 시장 수요를 흡수하는 흐름 확인

○ (응용) 30억 명 일상 사용자 접점까지 풀스택을 직접 확장하는 차별적 위치

- 구글은 검색·OS·업무 도구 등 일상 서비스를 직접 운영하는 사업자로, 글로벌 30억 사용자와 1,300만 고객사 기반 위에서 풀스택을 최종 사용자 접점까지 직접 확장할 수 있는 위치 확보
  · (일상 도구) Docs, Gmail, 검색 등 기존 서비스에 AI를 결합한 'Workspace Intelligence'를 운영, 별도 앱 설치나 사용 습관 변화 없이도 일상 업무 환경 안에서 AI 기능을 자연스럽게 확산
  · (기업 환경) 일반 직원도 자연어만으로 AI 에이전트를 만들 수 있는 환경을 제공하며, 개발 역량이 없는 사용자까지 풀스택 기능을 업무에 즉시 활용할 수 있는 구조 형성
  · (모바일 접점) 글로벌 활성 기기 약 30억 대 규모의 안드로이드에 Gemini를 OS 차원에서 통합하며, 제조사 단말이 켜지는 순간부터 구글 AI가 작동할 수 있는 모바일 확산 경로 확보

(2) (경쟁우위) 일상 서비스까지 일체 운영하는 풀스택 사업자로서의 강점

○ (내부 최적화) 하위 인프라 개선이 곧 최종 서비스 응답 속도 향상으로 직결

- 구글은 TPU부터 최종 서비스까지 직접 운영하는 사업자로, 하위 인프라의 성능 개선이 그 위에서 작동하는 모델·서비스의 응답 속도와 처리 효율로 곧바로 이어지는 풀스택 효과 구현
- TPU의 추론 성능이 개선되면 그 위에서 작동하는 AI 서비스의 처리 속도도 함께 향상되어, 고객은 별도 모델 교체 없이도 하위 인프라 개선 효과를 서비스 단계에서 자연스럽게 흡수
- Citadel Securities가 Google Cloud TPU 활용으로 동일 작업량을 약 4배 빠르게 처리한 사례는, 풀스택 통합 효과가 실제 산업 현장의 업무 처리 속도 개선으로 나타날 수 있음을 입증

● (자사 서비스 접점) 사내 검증을 거친 AI 기능을 외부 고객·일반 사용자에게 단계적 확산

- 구글은 자사 엔지니어링·마케팅 등 핵심 업무를 신규 AI 기능의 첫 검증 무대로 삼아, 내부에서 효과가 입증된 기능만 외부 클라우드 고객에게 공급하는 단계적 검증 원칙 채택
- 자사 서비스에서 축적되는 사용 데이터가 다시 AI 기능 개선에 활용되면서, 외부 사업자가 접근하기 어려운 데이터 흡수·모델 개선 선순환 구조가 풀스택 전반의 차별 자산으로 작동

● (B2B·B2C 동시 확보) 같은 풀스택 위에서 기업 시장과 일상 사용자 접점을 함께 장악

- 구글은 클라우드 기업 고객(B2B)과 일상 서비스 일반 사용자(B2C)를 동시에 보유한 사업자로, 같은 풀스택 기반을 기업 시장과 소비자 접점에 함께 확산할 수 있는 위치 확보
- 기업용·소비자용 AI가 같은 모델·인프라 기반 위에서 운영되어, 한쪽 영역의 사용 경험과 성능 개선이 다른 영역의 서비스 고도화로 이어지는 양방향 확산 효과 확보

(3) (위험 요인과 변수) 규제 환경과 시장 지위

● (반독점 항소심 변수) 풀스택 사용자 접점이 외부 규제 변수에 노출

- 美 연방법원은 '25년 9월 구글의 검색 시장 독점을 인정하면서도 풀스택 핵심 자산인 Chrome 매각·Android 분할 명령은 기각, 다만 5년간 검색 데이터 공유 의무·독점 계약 금지 등 부분 규제 부과
- 그러나 美 법무부와 38개 주가 매각 명령 재검토를 요구하며 항소를 제기, '26년 말~'27년 초 항소심 결과에 따라 Chrome 매각 명령 재인용 시 핵심 사용자 접점이 흔들릴 가능성
- Chrome은 약 34억 명 사용자를 검색·Gemini로 연결하는 구글 풀스택의 핵심 B2C 경로로, 매각이 현실화될 경우 구글 AI 생태계에서 최종 사용자 접점이 분리되는 결정적 변수로 작용

● (B2B 시장 확장 한계) 클라우드 점유율 격차로 풀스택 도달 범위 제약

- 구글의 B2B 풀스택 확장은 Google Cloud 채택 기반에 의존하나, '26년 1분기 글로벌 클라우드 점유율이 약 12%에 머물며 AWS·Azure 양강 대비 절반 이하 격차 지속

- 글로벌 기업의 멀티 클라우드 활용이 시장 표준으로 자리잡으면서, 구글 풀스택을 채택하더라도 핵심 업무는 AWS·Azure에 두고 일부 AI·데이터 작업만 Google Cloud에 분산 운용하는 방식 일반화
- 시장 분석기관 Flexera는 Google Cloud가 소규모 거래에 강한 반면, 대규모 기업 계약은 AWS· Azure가 주도한다고 평가, 구글 풀스택의 B2B 확장력이 중소기업 영역에 제한될 수 있음을 시사

○ (모델 단독 우위 부족) Gemini만으로는 차별화 어려운 3강 경합 구도

- 주요 벤치마크 종합 점수*에서 GPT-5.5가 1위, Gemini와 Claude가 공동 2위를 기록, 구글은 자체 모델 Gemini만으로 시장을 차별화하기 어려운 3강 경합 구도에 직면
- 시장 채택도 영역별로 분화되어 코딩은 Claude, 범용 활용은 ChatGPT, 업무 도구 통합은 Gemini가 각각 강점을 보이며, 단일 모델이 모든 사용 영역을 장악하기 어려운 구조 고착
- 구글이 외부 모델인 Claude까지 자사 플랫폼에서 호출할 수 있도록 통합한 결정은, 단순 개방형 전략을 넘어 자체 모델만으로 부족한 차별성을 외부 모델 흡수로 보완하려는 선택으로 해석

3. 화웨이: 중국형 AI 풀스택 자립 전략

(1) (구조) 칩에서 모빌리티까지 이어지는 中 산업형 수직 통합

○ (AI 칩) Ascend 칩과 자체 HBM으로 풀스택 최하단부터 자국 자립 범위 확장

- 화웨이는 '26년 1분기 자체 AI 가속 칩 'Ascend 950PR'을 출시하며, 풀스택의 최하단에 해당하는 연산 기반을 자체 확보하고 후속 인프라 확장의 출발점 마련
- Ascend 950PR과 연계되는 가속기 카드에는 화웨이가 자체 개발한 고대역폭 메모리(HBM)까지 탑재되며, 그동안 韓 기업이 글로벌 공급을 주도해 온 메모리 영역까지 자립 범위 확장
- ByteDance·Alibaba 등 中 빅테크의 채택이 이어지면서 시범 도입을 넘어 양산 공급 단계로 진입, 中 정부 '15차 5개년 계획' 풀스택 자립 정책과 결합되며 화웨이가 정책 구현 핵심 사업자로 부상

○ (인프라) 자사 칩 기반 집적 시스템으로 데이터센터급 연산 자원 내재화

- 화웨이는 '26년 MWC에서 Ascend 칩 약 8,200개를 자체 통신망으로 연결한 데이터센터급 컴퓨팅 시스템 'Atlas 950 SuperPoD'를 공개, 자체 칩 기반 인프라 구축 단계에 본격 진입

- 화웨이는 자체 비교 기준으로 동급 엔비디아 시스템 대비 약 6.7배의 컴퓨팅 성능을 제공한다고 주장했으며, 다수의 자국 칩을 하나의 시스템으로 묶는 중국형 AI 인프라 노선 가시화
- Atlas 950 SuperPoD를 Huawei Cloud를 통해 中 인터넷·금융·통신·전력 산업에 직접 공급, 자체 칩과 자체 인프라를 자국 산업 수요에 곧바로 연결하는 내수 공급 회로 형성

● (모델) 자체 모델과 자국 오픈소스 모델을 자사 칩 위에서 동시 운용

- 화웨이는 자체 AI 모델 'PanGu'와 中 대표 오픈소스 모델 'DeepSeek'를 Ascend 칩에 함께 최적화하며, 자체 모델과 외부 오픈소스 모델을 모두 자사 칩 위에서 운용하는 구조 형성
- DeepSeek V4가 화웨이 인프라 위에서 작동하고 학습 일부 단계까지 Ascend 칩이 활용되는 사례는, 中 AI 모델이 외산 인프라 없이 자국 칩 환경에서 운영 가능함을 보여주는 결정적 사례
- 자체 모델 PanGu는 의료 병리 진단, 철강로 온도 예측, 석유·가스 탐사 등 산업별 특화 시리즈로 개발되어, 中 산업 현장의 진단·예측·분석 업무에 직접 투입되는 응용 전용 모델로 활용

● (응용) HarmonyOS 기반, 단말에서 차량까지 풀스택 응용 접점 확장

- 화웨이는 자체 운영체제 'HarmonyOS'를 스마트폰·PC·IoT에서 차량 서비스로 확장하며, 풀스택 최상위 응용 계층을 자체 OS 기반 위에서 직접 운영하는 구조 형성
- 화웨이 주도 자동차 동맹 'HIMA'와 자율주행 시스템 'Qiankun ADS'는 자체 OS·칩·AI 모델을 차량 서비스와 결합하며, 풀스택 최상위 계층이 실제 매출로 연결되는 단계 진입

(2) (경쟁우위) 제재 환경을 내수 생태계 장악 기회로 전환

● (자국산 대체 수요) 美 제재로 발생한 中 인프라 공백을 화웨이가 흡수

- 美 수출 통제로 中 시장에서 외산 AI 인프라 도입이 제한되면서 자국산 솔루션을 찾는 대체 수요가 확대, 화웨이는 이를 흡수할 수 있는 사실상 유일한 자국 풀스택 사업자로 입지 강화
- 제재 환경이 中 정부의 기술 자립 정책 및 산업 육성 전략과 결합되면서, 화웨이는 단순 대체 공급자를 넘어 정부 지원을 받는 자국 AI 인프라 핵심 사업자로 부상

- 그 결과 中 AI 칩 시장에서 자국 칩 점유율은 '25년 약 41%까지 확대되며, 약 2년 전 엔비디아가 90%를 점유하던 시장 구도가 자국 칩 중심으로 재편되는 흐름 가시화

○ (산업 맞춤 통합 공급) 풀스택 4계층을 산업별 패키지로 결합 공급

- 화웨이는 풀스택 4계층(칩·인프라·모델·응용)을 분리 공급하지 않고 中 산업 수요에 맞춘 단일 패키지로 결합하며, 산업별 요구에 대응하는 통합 솔루션 사업자로 입지 강화
- 단일 패키지 공급 모델을 기반으로 협력사 약 9,800곳과 개발자 약 380만 명이 결집한 자체 생태계가 형성되며, 화웨이 풀스택이 중국 산업 환경의 공통 운영 기반으로 자리 잡는 단계 진입
- 중국 정부의 'AI Plus' 정책이 제조·의료·교육·정부 서비스 전반의 AI 도입을 가속하면서 화웨이 풀스택 확산을 뒷받침하고, 적용 영역이 특정 산업을 넘어 사회 운영 전반으로 확대되는 흐름 형성

○ (자체 실행 표준) HarmonyOS 기반, 단말·차량을 자국형 운영 환경으로 통합

- 화웨이는 단말부터 차량·산업 시스템까지 포괄하는 풀스택 폭과 자체 운영체제 'HarmonyOS'를 동시 보유, 자국 사용자 일상 환경을 자체 표준으로 관통할 수 있는 사실상 유일한 위치 확보
- 산업 전반이 앱 단위 운영에서 AI 에이전트 단위 운영으로 전환되는 흐름에 대응, HarmonyOS 6에 자체 에이전트 작동 표준을 도입하면서 AI 비서가 여러 앱 기능을 호출·조합하는 구조 설계
- 동일 표준이 차량 영역까지 확장되어 자체 OS 기반 차량용 환경과 자율주행 시스템 'Qiankun ADS'를 결합, 스마트폰·PC·IoT·자동차가 단일 자체 표준 안에서 작동

(3) (위험 요인과 변수) 기술 격차와 생태계 제약

○ (단일 칩 성능 격차) Ascend 칩, 엔비디아 첨단 칩 대비 2~3년 추격 구간 지속

- 화웨이는 차세대 자체 칩이 중국 시장 한정 공급용 엔비디아 칩(H20)보다 우위라고 주장하나, 글로벌 첨단 사양인 Blackwell·Vera Rubin과 비교하면 여전히 한 세대 이상 성능 격차 존재
- 미국 외교협회(CFR)는 화웨이 차세대 Ascend 칩이 '26~'27년에야 엔비디아 H100급 성능에 도달할 것으로 추정, 엔비디아 첨단 칩과의 약 2~3년 격차가 단기간에 해소되기 어렵다고 분석

- 화웨이가 자체 HBM 개발을 추진하고 있지만, 글로벌 표준인 HBM4 기준에서는 韓 기업과의 기술 격차가 여전해 메모리 영역의 자립도 단기간에 완성되기 어려울 것으로 전망

● (SW 생태계 격차) CUDA 호환 전략이 화웨이 SW 자립의 한계 노출

- 화웨이는 자체 SW 환경 'CANN Next'를 NVIDIA CUDA와 호환되는 형태로 출시하며, 중국 빅테크 개발자가 기존 CUDA 코드를 큰 수정 없이 Ascend 칩으로 옮길 수 있는 우회 경로 마련
- 그러나 CUDA 호환 전략은 자체 SW만으로는 개발자 생태계를 단기간에 확보하기 어렵다는 점을 보여주는 신호로, SW 영역에서는 엔비디아 생태계에 의존하는 한계 노출
- 현재 중국 주요 빅테크의 핵심 코드도 대부분 CUDA 환경을 기반으로 작성된 상태이며, CFR은 화웨이의 글로벌 AI 컴퓨팅 처리 능력이 엔비디아의 약 4~5% 수준에 머물 것으로 분석

● (OS 글로벌 확장 한계) 中 내수 안착에도 글로벌 표준과 격차 지속

- HarmonyOS는 중국 내수 시장에서는 iOS를 앞서며 안착 기반을 마련했으나, 글로벌 점유율은 약 4~5% 수준에 머물러 Android(약 77%)·iOS(약 19%) 양강 구도와의 격차가 뚜렷
- 화웨이는 '24년 10월부터 자체 OS에서 Android 앱 호환을 완전히 종료하고 자체 앱 생태계 기반의 자국형 OS로 전환했으나, 지원 앱 수가 약 1.5만 개 수준에 그치며 목표치에 크게 미달
- 한편 '26년 글로벌 진출 전략에서도 동남아·중동 파트너를 중심으로 단계적 진입을 추진하고 있어, HarmonyOS가 단기간에 글로벌 표준으로 확장되기는 어려운 흐름 시사

출처 : NVIDIA Newsroom 외(2026.3.)
https://nvidianews.nvidia.com/news/nvidia-vera-rubin-platform
https://www.techradar.com/pro/huawei-debuts-its-atlas-950-ai-superpod-at-mwc-2026-taking-the-ai-data-center-fight-to-nvidia-and-amd
https://cloud.google.com/blog/topics/google-cloud-next/welcome-to-google-cloud-next26
https://www.ibm.com/think/news/ai-tech-trends-predictions-2026
https://futuretech.mit.edu/publication/the-price-of-progress-algorithmic-efficiency-and-the-falling-cost-of-ai-inference
https://www.deloitte.com/ce/en/industries/technology/analysis/ai-infrastructure-compute-strategy.html
https://www.tismo.ai/blog/the-enterprise-ai-stack-in-2026-models-agents-and-infrastructure
https://tech-insider.org/big-tech-ai-infrastructure-spending-2026/
https://solutionsreview.com/ai-and-enterprise-technology-predictions-from-industry-experts-for-2026/
https://blogs.nvidia.com/blog/gtc-2026-news/
https://www.m-economynews.com/news/article.html?no=65040
https://www.xdnode.co.kr/insight/articles/2026-government-ai-infrastructure-programs
https://www.koreadaily.com/article/20260414221003805
https://www.sptatimeskorea.com/post/제20260318-ti-01호
https://pasqualepillitteri.it/en/news/1311/gemini-enterprise-agent-platform-google-next-2026
https://cloud.google.com/blog/topics/google-cloud-next/google-cloud-next-2026-wrap-up
https://www.mejba.me/blog/google-io-2026-ai-announcements