Ⅱ 주요 동향(2) : ICT

1 AI 경쟁축, 모델에서 실행 인프라·플랫폼·OS로 이동

AI 실행 생태계 경쟁으로 전환되는 글로벌 AI 플랫폼 구도

- 경쟁의 초점이 모델 성능을 넘어 실행 생태계로 이동
  - '25년 하반기 이후 AI 활용이 대화형 질의응답을 넘어 업무 자동화·기기 제어를 수행하는 에이전트 중심으로 확대되면서, AI 도입 논의도 시범 적용을 넘어 산업 전반의 실행 기반 구축으로 확장
  - 모델 간 성능 격차가 줄어드는 가운데, 에이전트가 실제 작업을 수행하려면 연산 인프라·개발 도구·운영체제·기기가 함께 갖춰져야 하므로 모델 단독 경쟁력만으로는 차별화가 어려워지는 양상
  - 이에 따라 주요 빅테크는 모델 경쟁을 넘어 AI 인프라 확장과 에이전트 플랫폼 주도권 확보를 핵심 전략으로 제시하고 있으며, 실행 기반 전반이 '26년 AI 산업 경쟁의 새로운 축으로 부상
- COMPUTEX·Build·WWDC, AI 내재화의 세 계층을 동시에 제시
  - '26년 6월에 개최된 COMPUTEX·Build·WWDC는 각각 하드웨어·공급망, SW·개발 플랫폼, OS·사용자 경험이라는 서로 다른 산업 계층을 대표
  - 세 행사는 성격이 다르지만, 올해 공통적으로 AI를 외부 서비스가 아니라 각 계층의 핵심 시스템 안에 내재화하는 전략을 전면에 제시
  - 결국 세 행사의 공통 질문은 "AI를 어디서, 어떻게 실행하고 통제할 것인가?"이며 이는 AI 경쟁의 무게중심이 모델 접근성에서 실행 위치·데이터 귀속·플랫폼 통제력으로 이동하고 있음을 시사

COMPUTEX, Build, WWDC가 보여주는 계층별 AI 내재화

1. (하드웨어) COMPUTEX, AI 실행 인프라 경쟁의 축소판으로 부상

- COMPUTEX, PC 부품 전시회에서 AI 인프라 전시로 탈바꿈
  - COMPUTEX 2026은 152개국 11.1만여 명이 모인 가운데, PC 부품 중심 전시에서 'AI 팩토리·에이전트·전력·냉각'을 앞세운 AI 인프라 전시로 무게중심이 이동한 것으로 평가

- 주제 'AI Together' 아래 로보틱스 전용관을 처음 신설하고 800 VDC 배전, 메가와트급 냉각 등 전력·냉각 설비가 전면에 등장하면서 전시장 구성 역시 이러한 변화를 뒷받침
- 이는 AI 하드웨어 경쟁의 초점이 연산 칩 확보를 넘어, 칩 간 연결, 전력 공급, 냉각, 단말·로봇 배치 등 실행 인프라 전반으로 확장되고 있음을 시사

● 하드웨어 경쟁, 수직통합과 계층별 전문화 전략이 병존

- 동 행사에서 NVIDIA와 Intel, Qualcomm, Marvell, NXP 등이 잇따라 키노트에 오르며, AI 하드웨어를 전 계층에서 묶는 수직통합 전략과 연산·연결·단말·제어 등 계층별 전문화 전략이 함께 부각
- NVIDIA는 연산 칩부터 PC·로봇까지 단일 스택으로 묶는 수직통합 전략을 구체화했고, Intel은 연산·제조, Qualcomm은 단말, Marvell은 칩 간 연결망, NXP는 현실 기계 제어에 집중
- 이는 AI 하드웨어 경쟁이 특정 GPU 성능 경쟁에 머물지 않고, 에이전트를 실제로 실행하기 위한 연산·연결·디바이스·피지컬 AI·제조 기반 경쟁으로 확장되고 있음을 시사
- 결국 COMPUTEX는 AI 실행 인프라를 둘러싼 수직 통합형 전략과 계층별 전문화 전략이 동시에 전개되는 경쟁 구도를 보여주는 장으로 기능

(1) (수직통합) NVIDIA, 데이터센터·PC·로봇을 하나의 AI 실행 스택으로 연결

● GTC Taipei, 데이터센터 밖 AI 실행 지점으로 PC·로봇을 전면 배치

- 데이터센터 AI 칩으로 대표돼 온 NVIDIA는 COMPUTEX 사전 행사 격인 GTC Taipei(6.1)에서 개인용 PC와 로봇·차량을 전면에 내세우며, AI 실행 지점을 데이터센터 밖으로 확장하는 전략을 제시
- 이번 발표는 개별 신제품 공개보다 연산 칩, Windows PC, 피지컬 AI(로봇·차량)를 동일한 에이전트 실행 구조 안에 묶어 NVIDIA 스택의 적용 범위를 넓히는 데 초점
- 이에 따라 이번 키노트는 그간 별도로 전개되던 데이터센터·PC·로봇 사업을 하나의 AI 실행 흐름으로 연결하려는 NVIDIA의 전략적 포석으로 해석

① (연산) Vera Rubin, 에이전트 추론 병목 해소를 위한 CPU·GPU 통합 스택 제시

- NVIDIA는 에이전트 처리에 특화한 Vera Rubin의 양산 단계 진입과 출하 계획을 제시, 이를 모델 학습보다 다수 에이전트를 동시에 실행하기 위한 시스템으로 규정 (NVIDIA 키노트, '26.6.)

- 핵심은 신규 데이터센터 CPU Vera로, 코어 수 확대에 집중한 기존 CPU와 달리 도구 호출·결과 반환을 짧게 반복하는 에이전트 작업에 맞춰 응답 지연을 줄이는 데 초점
- NVIDIA는 이를 기반으로 GPU, CPU, 네트워크 스위치를 하나의 랙 단위 시스템으로 결합하고, OpenAI, Anthropic, SpaceX 등을 초기 고객·수요처로 제시
- 이는 NVIDIA가 기존 GPU 중심의 경쟁력을 기반으로 에이전트 추론 인프라의 CPU·네트워크 영역까지 포괄하려는 수직통합 전략으로 해석

② (PC) RTX Spark, 개발자 PC를 로컬 AI 실행 거점으로 전환

- NVIDIA는 MediaTek과 공동 설계한 Arm 기반 PC 플랫폼 칩 RTX Spark를 공개하며, Intel과 Apple이 주도해 온 PC 핵심 연산 영역에 직접 진입
- RTX Spark 기반 PC는 대용량 메모리를 바탕으로 대형 AI 모델을 외부 클라우드 연결 없이 기기 내부에서 구동할 수 있는 개발자용 PC를 지향
- Dell, HP, Lenovo, ASUS, MSI가 이 칩을 탑재한 노트북·데스크톱을 '26년 가을 출시할 예정으로, NVIDIA의 PC 진입은 단일 칩 공개를 넘어 완제품 단계로 확장
- 이처럼 AI 모델을 기기 안에서 직접 구동하는 흐름은 기기 1대당 메모리 탑재량을 크게 늘려, 온디바이스 AI 확산에 따른 고대역폭·대용량 메모리 수요 확대로 이어질 가능성을 시사

③ (피지컬 AI) Cosmos 3·Isaac GR00T, 로봇·차량으로 AI 실행 구조 확장

- NVIDIA는 물리 세계의 작동을 학습·예측하는 월드모델 Cosmos 3와 휴머노이드 개발 기반 Isaac GR00T를 공개하며, 데이터센터·PC의 AI 실행 구조를 로봇·차량까지 확장
- Cosmos를 공통 기반으로 자율주행 플랫폼 Drive Hyperion과 휴머노이드 개발 환경을 연결하고, 모델·데이터 공개를 통해 외부 개발자를 자사 개발 생태계로 유입하는 전략을 제시
- 이는 데이터센터에서 학습·추론된 AI가 PC의 로컬 실행 환경을 거쳐, 로봇·차량과 같이 물리 세계의 직접적인 실행 주체로 확장되는 흐름을 보여주는 사례
- 이로써 연산(Vera), PC(RTX Spark), 로봇(Cosmos)이 하나의 실행 흐름으로 연결되며, NVIDIA는 칩 공급자를 넘어 세 계층을 묶어 제공하는 인프라 사업자로 이동

(2) (연산) Intel, 에이전트 시대의 CPU 역할 재정의

- Intel, GPU 중심 AI 시장에 CPU·공정 경쟁력으로 대응
  - NVIDIA가 AI 실행 스택을 데이터센터 밖으로 확장했다면, Intel은 에이전트 실행 과정에서 CPU가 맡는 조율·입출력·도구 호출 기능을 새로운 경쟁 축으로 제시
  - Intel은 COMPUTEX 키노트에서 'AI 시대의 다음 장'을 주제로 GPU 중심으로 전개된 AI 시장에 CPU와 공정 경쟁력으로 대응하는 전략을 공개
  - AI 확산의 주역이 GPU였음을 인정하면서도, 에이전트가 작업 계획, 파일 입출력, 도구 호출을 반복하는 과정에서는 CPU의 조율 기능이 중요해질 것으로 전망
  - 이를 겨냥해 18A 공정을 적용한 첫 데이터센터 CPU Xeon 6+를 공개하고, 에이전트 실행 수요 확대에 대응하는 연산·제조 역량을 부각
  - 나아가 단일 칩을 넘어 시스템 단위 협력으로 범위를 넓혀, Foxconn과는 랙스케일 AI 인프라의 통합·제조를, SambaNova와는 CPU가 작업 조율을 맡고 GPU가 연산을 담당하는 분담형 추론 구조 제시
  - 이로써 Intel은 AI 데이터센터를 GPU 독점이 아니라 CPU·GPU·전용칩이 역할을 나눠 맡는 개방형 시장으로 규정했으나, CPU 수요의 실제 확대폭은 시장 검증이 남은 단계로 평가

(3) (연결) Marvell, AI 성능 병목의 칩 간 연결망 이동 부각

- Marvell, 구리 기반 연결의 한계를 제시하며 광 전환과 CPO 강조
  - Marvell의 머피 CEO는 키노트 첫날 구리 기반 스위치와 광을 칩까지 연결한 스위치를 나란히 제시하며, 연결 방식이 구리에서 광으로 전환되고 있음을 실물로 부각
  - AI 클러스터의 성능은 다수 연산 칩 간 데이터 이동 속도에 크게 좌우되나, 구리 기반 연결은 속도 향상에 따라 전송 거리와 전력 효율 측면의 한계가 커지는 구조
  - Marvell은 이를 해결할 방식으로 동시 패키징 광학(CPO)을 제시하고, 업계 첫 102.4Tbps AI 전용 스위치 Teralynx T100을 공개하며 이번 분기 중 고객 샘플 공급에 착수
  - 수직통합을 택한 NVIDIA도 Vera Rubin에 첫 양산 CPO 스위치 적용 계획을 제시하면서, 칩 간 연결망은 AI 실행 인프라의 보조 요소가 아니라 성능과 전력 효율을 좌우하는 핵심 계층으로 부상

(4) (디바이스·기기) Qualcomm, 온디바이스 기반 에이전트 실행 확장

- Qualcomm, 폰·PC·차량을 에이전트 수행 단말로 재정의
  - Qualcomm은 '에이전트의 해'를 주제로 한 키노트에서, AI 실행의 중심이 데이터센터에서 폰·PC·차량 등 사용자 인접 기기로 확장된다는 관점을 핵심 의제로 제시
  - 기존에는 기기와 서비스가 스마트폰을 중심으로 연결됐으나, 향후에는 스마트폰뿐 아니라 PC, 차량, 안경 등이 에이전트의 실제 작업을 수행하는 엔드포인트로 재편될 가능성을 부각
  - 다만 현재 기기는 사람의 직접 조작을 전제로 설계돼 AI의 자율 조작을 반영하기 어려우므로, 사람과 에이전트의 조작 체계를 동시에 수용하는 단말 설계가 필요하다고 설명
  - Qualcomm은 이러한 단말 설계 방향을 보여주는 사례로 X2 Elite 기반 미니 PC를 공개하고, 자율 에이전트가 외부 서버에 의존하지 않고 기기 내부에서 작업을 처리하는 과정을 시연
  - 이로써 Qualcomm은 부품 공급을 넘어 에이전트가 작동하는 단말 자체를 사용자 접점으로 확보하려는 방향을 제시했으나, 제조사 채택과 응용 생태계 확산 여부는 추가 검증이 필요한 단계

(5) (피지컬 AI) NXP, 현실 기계 지능의 말단 분산 부각

- NXP, 로봇의 핵심 난제를 고차 추론보다 반사형 제어로 규정
  - NXP는 AI가 로봇·차량 등 현실 기계로 확장되는 국면에서, 지능을 중앙의 거대 모델 하나에 집중하기보다 기능별로 필요한 위치에 분산 배치해야 한다는 전략을 제시
  - 이는 걷기·잡기·회피처럼 사람에게 쉬운 동작이 기계에는 가장 어렵다는 '모라벡의 역설'에 기반한 접근으로, 현실 기계의 핵심 난제가 고차 추론보다 위험에 즉각 대응하는 제어 능력에 있음을 강조
  - NXP는 이를 추론·조정·반사의 3계층으로 나눈 뉴럴 액시스 구조로 제시하고, 0.04초 내 처리가 필요한 반응은 중앙 연산부를 거치지 않고 관절·손 등 말단 칩이 직접 수행하도록 설계
  - 로봇의 안정적 파지·조작에는 주변 인식과 동작 판단의 동시 처리가 필요하나, 말단 칩의 연산·전력 제약을 고려해 NXP는 VLA 모델을 eIQ 툴킷으로 경량화해 탑재하는 방식을 제시

- 이는 피지컬 AI 경쟁이 중앙의 대형 모델 성능에만 머물지 않고, 말단 칩에서 실시간 제어·저전력 추론·안전 반응을 구현하는 구조로 확장되고 있음을 시사

(6) 대만, AI 하드웨어 제조·조립 거점성 강화

○ AI 하드웨어 공급망, 수직통합과 계층별 전문화 전략 모두 대만 제조 생태계에 의존

- Intel CEO가 키노트에서 “설계부터 제조까지 전 과정이 대만에 있다”고 강조했듯이, AI 하드웨어 생태계에서 칩 제조와 서버·노트북 조립 역량은 COMPUTEX 개최지인 대만에 집중
- 칩 제조 분야는 시가총액 약 2조 달러('26.6. 기준)의 TSMC가 주도하고 있으며, 첨단 공정에서의 파운드리 우위가 대만을 AI 하드웨어 공급망의 출발점으로 만드는 기반으로 작용
- 조립 분야는 세계 최대 전자제품 위탁제조사 Foxconn을 중심으로, 완성된 칩을 서버·랙으로 묶는 단계를 넘어 시스템 통합과 설계 지원으로 역할을 확대
- 클라우드에 머물던 AI 실행 지점이 노트북·PC 등 단말로 확장되면서 완제품 제조 역량의 중요성이 커지고 있으며, Quanta·Foxconn 등 대만 제조사의 전략적 비중이 확대되는 흐름
- 결국 AI 하드웨어 경쟁이 수직통합 전략과 계층별 전문화 전략으로 나뉘어 전개되더라도, 실제 구현 단계에서는 대만의 제조·조립·통합 역량에 대한 의존도가 높은 구조로 평가

2. (SW·개발 플랫폼) MS Build, 기업용 AI 에이전트의 구축·실행·통제 체계 제시

○ MS Build, AI 내부 소유와 개방형 SW 스택을 핵심 전략으로 제시

- 하드웨어 계층에서 AI 실행 인프라 경쟁이 연산·연결·디바이스·제조 기반으로 확장되고 있다면, SW 계층에서는 기업이 에이전트를 어떻게 구축·배포·통제할 것인가가 핵심 쟁점으로 부상
- Microsoft CEO Satya Nadella는 기업이 AI를 외부 서비스로 활용하는 단계를 넘어, 자사 데이터와 업무방식에 맞춰 직접 구축하고 소유해야 한다는 점을 키노트의 핵심 메시지로 제시
- 모델 성능 격차가 줄어들수록 차별화의 기준은 지능에 대한 단순 접근보다, 기업 데이터와 업무 노하우를 자체 시스템에 내재화하는 역량으로 이동
- Microsoft가 제시한 소유의 방식은 폐쇄형 구축보다 개방형 선택에 가까우며, 칩·모델·도구·실행 위치 등 스택 전 계층에서 개발자가 구성 요소를 직접 선택할 수 있도록 설계

- 이에 따라 Build는 개방과 소유의 결합을 핵심으로, Microsoft가 OS·도구 기반을 제공하고, 개발자가 모델·실행 환경을 선택하되, 데이터·맥락은 Microsoft 플랫폼 위에 축적되는 구조를 제시

● Windows, 칩·모델·도구를 선택할 수 있는 개방형 SW 풀스택 기반으로 재정의

- Microsoft는 실리콘-OS-개발 도구-클라우드로 이어지는 SW 전 계층을 Windows 중심으로 구성하되, 계층별 구성 요소는 개발자가 선택할 수 있도록 하는 개방형 풀스택 전략을 제시
- **(실리콘)** AI 연산을 클라우드에만 의존하지 않고 기기 내부의 CPU·GPU에서 처리할 수 있도록 설계해, 특정 GPU에 종속되지 않는 실행 구조를 지향
- **(OS)** Windows를 에이전트의 실행 기반이자 활동 범위 통제 계층으로 재설계하고, 로컬 환경에서의 AI 작업 수행과 파일·네트워크 접근 권한 제어 기능을 함께 제공
  - 일상적 텍스트 처리를 맡는 Aion 1.0 Instruct와 추론·도구 호출을 수행하는 Aion 1.0 Plan을 탑재해, 클라우드 의존 없이 기기 내부에서 일부 AI 작업과 에이전트 실행을 처리
  - 격리 도구 MXC를 내장하여 에이전트의 파일·네트워크 접근 권한을 운영체제 차원에서 제한하고, 작업 위험도에 따라 프로세스·세션 단위로 차단 수준을 조정
- **(개발 도구)** 개발 도구 계층에서는 GitHub Copilot이 에이전트 개발 환경 역할을 맡되, 자사 모델을 강제하지 않고 SDK를 통해 외부 도구·모델을 자유롭게 연결하도록 개방
- **(클라우드)** 완성된 에이전트는 Foundry·Azure를 통해 클라우드에 배포·확장되며, 로컬 기기에서 개발된 작업 흐름을 데이터센터 환경으로 이전해도 동일한 실행 구조를 유지
- 이는 칩을 중심으로 수직통합 스택을 구축하는 NVIDIA와 달리, 칩·모델·도구 선택권은 열어두되 실행·배포·운영 단계는 Windows와 Azure 기반 플랫폼에 축적되도록 설계한 구조로 해석

● Microsoft Agent Platform, 에이전트 검증·배포부터 업무 호출까지 단일화

- Microsoft는 앞서 살핀 SW 계층을 'Microsoft Agent Platform'으로 통합해, 에이전트 제작 이후 검증·배포부터 기존 업무 도구 호출까지 이어지는 단일 실행 흐름을 구축

- **(배포·검증)** 개발 도구에서 만든 에이전트는 배포 플랫폼 Foundry에서 샌드박스 검증을 거쳐 업무 환경에 배포되며, 작업 성격에 따라 OpenAI, Anthropic, MAI 등 적합한 모델을 선택 가능
- **(근거 연결)** 배포된 에이전트는 맥락 연결 계층 Microsoft IQ를 통해 사내 업무 흐름·구조화 데이터·실시간 웹과 연결되며, 실제 근거에 기반한 답변을 생성해 부정확한 응답을 줄이는 구조
- **(맞춤 학습)** 이후 맞춤 튜닝 체계 Frontier Tuning은 기업의 데이터와 업무 방식을 에이전트에 지속 반영해, 범용 모델을 기업별 업무 맥락에 맞춘 내부 자산으로 전환
- **(업무 호출)** 이렇게 완성된 에이전트는 기업 협업 도구 Teams 및 M365에서 호출되고, 일정 조율·메일 정리를 수행하는 상시 개인 에이전트(Scout)로 확장
- 해당 파이프라인은 모델 선택의 개방에서 기업별 맞춤 학습으로 이어지며, 접근은 열어두되 데이터·맥락·튜닝은 Microsoft 플랫폼에 축적되는 구조로 작동

○ Agent 365, 확산된 에이전트를 신원·보안·규정 준수 체계로 통합 관제

- Microsoft Agent Platform이 에이전트의 실행 흐름을 구축했다면, Agent 365는 조직 전반으로 확산된 에이전트를 신원·보안·규정 준수 체계 안에서 통제하는 관제 계층으로 제시
- 통합 관제 체계 Agent 365는 에이전트의 실행 클라우드나 제작 도구와 관계없이 단일 체계로 관리해, 에이전트 확산에 따른 관리 분산 문제를 완화
- 여기에 보안 에이전트 협업 체계 MDASH를 결합하여, 코드 및 데이터에 직접 작동하는 자율 에이전트가 새롭게 유발하는 취약점과 보안 위험까지 방어
- 개방·소유·통제가 한 체계에서 맞물리며 Microsoft의 SW 플랫폼 전략이 구체화되고, 경쟁의 무게중심은 개별 모델 성능에서 에이전트를 안전하게 운용하는 시스템 전체의 우위로 이동

3. (사용자 경험) Apple WWDC, AI를 운영체제 경험에 내재화하는 전략 제시

○ Apple, AI를 별도 서비스가 아닌 운영체제 내부 경험으로 전환

- Microsoft가 기업용 에이전트의 구축·실행·통제 체계를 제시했다면, Apple은 AI를 사용자가 매일 접하는 운영체제와 앱 경험 안에 통합하는 전략을 제시
- WWDC 2026은 차세대 Apple Intelligence와 AI 기반 Siri 개편을 통해, AI를 독립 서비스가 아닌 사용자 필요 중심의 기기·앱·운영체제 내장형 경험으로 배치

- 이에 따라 Apple은 iOS, macOS, iPadOS, watchOS, visionOS 등 전 운영체제에 걸쳐 AI 기능을 통합하고, Siri AI를 그 출발점으로 배치

● Siri AI, 개인 맥락·화면 인식·앱 간 작업을 수행하는 운영체제 통합 인터페이스로 재설계

- Apple은 2011년 출시 이후 가장 큰 Siri 개편을 단행하고 명칭을 'Siri AI'로 변경하며, 기존 음성 호출 방식을 넘어 텍스트 대화와 대화 이력 관리가 가능한 독립 앱 형태로 확장
- 핵심 변화는 개인 맥락 이해로, Siri AI가 메시지·메일·사진·캘린더 등 기기 내 개인 데이터를 검색·활용해 맥락 기반 응답을 생성하고 앱을 전환하지 않고 연쇄 작업을 수행하는 방향
- 화면 인식 기능 Visual Intelligence는 iPhone 카메라 앱에 직접 통합되어 사용자가 화면 위 콘텐츠나 카메라로 비추는 대상에 대해 질문하고 맥락에 맞는 응답을 수신
- Siri AI의 음성도 새로운 엔진으로 재설계되어 표현력이 향상되었으며, 사용자가 음성 톤과 속도를 직접 조정할 수 있는 맞춤 설정을 제공

● 3세대 Apple Foundation Models, 온디바이스·프라이빗 클라우드·외부 클라우드의 3계층 구조

- Apple은 차세대 Apple Intelligence의 기반으로, 작업 난이도와 개인정보 민감도에 따라 AI 실행 위치를 기기 내부, Private Cloud Compute, 외부 클라우드의 세 계층으로 나누는 처리 구조를 제시
- 온디바이스 모델은 요약, 핵심 정보 추출, 받아쓰기, 화면 인식 등 일상적 작업을 기기 내부에서 처리해 응답 지연을 줄이고 개인정보 외부 전송을 최소화하는 역할을 수행
- Private Cloud Compute는 기기 내부에서 처리하기 어려운 중간 수준 이상의 추론과 이미지 생성·편집 작업을 담당하되, 사용자 데이터를 저장하지 않고 외부 접근을 제한하는 프라이버시 통제 구조를 강조
- 복잡한 도구 호출과 고난도 추론은 Google Cloud 내 NVIDIA GPU에서 실행되는 AFM 3 Cloud Pro가 담당, 이는 자체 모델만으로 프론티어 수준 성능을 확보하기 어려운 한계를 보완하려는 접근으로 해석
- 결국 Apple의 AI 실행 구조는 간단한 작업은 기기 내부, 고연산 작업은 프라이빗 클라우드, 복잡 추론은 외부 모델로 분해해 프라이버시, 응답속도, 성능 간 균형을 맞추려는 전략으로 평가

○ App Intents 2.0과 Siri Extensions, 앱 생태계 전체를 AI 실행 접점으로 전환

- Apple은 자체 모델만으로는 프론티어 수준 어시스턴트 구현이 어렵다는 판단 하에 외부 모델(Gemini) 도입으로 경쟁력을 보완하는 대신, 차별점은 OS 수준의 앱 생태계 통합에서 확보하는 전략을 제시
- 이와 별도로 Siri Extensions을 통해 사용자가 설정에서 ChatGPT, Claude, Gemini 등 제3자 AI 모델과의 연계 선택지를 제공하는 개방형 구조도 함께 제시
- 모델과 도구 선택은 개방하되, Siri를 통한 앱 호출과 프라이버시 인프라는 자사 OS에 귀속시키는 구조로, MS가 칩·도구를 개방하되 OS·클라우드 의존을 심화하는 방식과 유사
- 경쟁 어시스턴트는 OS 외부에서 작동해 메일·메시지·캘린더 등 시스템 데이터 접근이 어려운 반면, Apple은 OS 내부에서 이를 직접 활용할 수 있는 구조적 위치를 차별점으로 제시
- 이로써 약 25억 대의 활성 기기 위에 모든 앱을 AI 실행 접점으로 전환하는 방향을 제시했으나, 개발자의 App Intents 구현 속도와 Siri AI의 실제 완성도는 가을 정식 출시 시점에서 검증이 필요

개방과 귀속의 결합이 만드는 플랫폼 경쟁의 새 구도

○ AI 경쟁의 무게중심, 모델 성능에서 실행 생태계의 결합으로 이동

- COMPUTEX·Build·WWDC를 종합하면, AI는 별도 서비스에서 벗어나 하드웨어, 개발 플랫폼, 운영체제와 앱 생태계 내부로 내재화되는 흐름을 보임
- NVIDIA는 모델·데이터 공개를 통해 외부 개발자를 자사 인프라로 유입하고, MS는 칩·모델·도구 선택권을 열어 개발자 진입장벽을 낮추며, Apple은 외부 모델 연계로 자체 모델 한계를 보완
- 그러나, 세 기업 모두 개방성을 확대하는 동시에, 실제 실행 경로와 데이터·업무 맥락·앱 호출 권한은 자사 인프라 또는 플랫폼 안에 축적되도록 설계
- 따라서 AI 플랫폼 경쟁의 핵심은 '어떤 모델을 보유했는가?'에서 'AI를 실행할 인프라와 사용자 접점, 데이터 흐름을 얼마나 결합했는가?'로 이동하는 양상

○ 칩-플랫폼-OS, 세 계층이 동시에 움직이는 산업 구조 변화

- 세 행사는 하나의 공급 사슬이 아니라, AI 내재화가 하드웨어·SW 플랫폼·사용자 경험의 세 계층에서 동시에 진행되고 있음을 각각의 시점에서 제시
- Microsoft와 Apple은 서로 다른 수직 생태계를 구축하고 있으나, 양쪽 모두 COMPUTEX에서 부각된 하드웨어 계층에 공통으로 의존하는 구조

- 온디바이스 AI 역시 양 생태계에서 동시에 부각되며, Windows의 로컬 모델 실행(Aion)과 Apple의 기기 내 모델 구동(AFM 3)이 저전력 NPU 설계라는 같은 하드웨어 조건 위에서 전개
- 이처럼 상위 계층에서는 별도의 생태계 경쟁이 진행되나, 하드웨어 계층은 양쪽을 관통하는 공통 기반으로 작동하며 계층 간 결합의 밀도가 경쟁력을 좌우하는 구조로 이동

● 모델 개방의 이면, 실행 흐름과 데이터의 플랫폼 귀속

- 세 기업의 전략에는 공통된 구조가 관찰되며, 모델·도구의 진입장벽은 낮추되 도입 이후 자사 인프라에 실행 흐름과 데이터가 축적되는 방식으로 플랫폼 의존을 형성
- 이는 경쟁의 단위가 개별 모델이나 제품이 아니라, 칩-SW-OS를 관통하는 플랫폼 전체로 확대되고 있음을 의미
- 한편, 세 계층 모두의 물리적 실행은 대만의 칩 제조·서버 조립·완제품 생산 역량에 의존하고 있어, 빅테크 간 플랫폼 경쟁이 심화될수록 대만 제조 기반의 전략적 비중도 함께 확대되는 구조
- 다만 세 행사가 제시한 방향은 각 기업의 전략 구상 단계이며, 실제 시장 채택과 생태계 확산은 하반기 제품 출시 이후 검증이 필요

출처 : Reuters 외(2026.6.)

https://www.reuters.com/commentary/breakingviews/taiwan-forges-thicker-silicon-shield-2026-06-05/
https://www.computextaipei.com.tw/en/news/A9462967BB2D7702/info.html
https://www.reuters.com/world/china/nvidia-ceo-kick-off-dominate-computex-gathering-taipei-2026-05-31/
https://apnews.com/article/nvidia-microsoft-ai-laptops-jensen-chip-c807f7333b93b9927b62b1240dcf65a1
https://www.servethehome.com/nvidia-computex-2026-keynote-live-coverage/
https://newsroom.intel.com/artificial-intelligence/intel-announces-new-ai-innovations-at-computex
https://www.servethehome.com/qualcomm-computex-2026-live-coverage/
https://www.digitimes.com/news/a20260603PD215/nvidia-intel-ai-semiconductors-competition-computex-2026.html?dt_ref=tag
https://www.honhai.com/zh-tw/press-center/events/2043
https://www.servethehome.com/nxp-computex-keynote-2026-coverage/
https://aei.dempa.net/archives/35217
https://blogs.microsoft.com/blog/2026/06/02/microsoft-build-2026-be-yourself-at-work/
https://machinelearning.apple.com/research/introducing-third-generation-of-apple-foundation-models
https://www.wired.com/story/meet-microsoft-scout-your-ai-coworker-that-never-logs-off/
https://blogs.windows.com/windowsdeveloper/2026/06/02/windows-platform-security-for-ai-agents/
https://github.blog/news-insights/product-news/github-copilot-app-the-agent-native-desktop-experience/
https://www.marvell.com/company/newsroom/marvell-announces-102-4-tbps-ai-cloud-data-center-switch.html
https://www.digitimes.com/news/a20260602VL212/marvell-copper-cpo-data-center-server-rack-computex-2026.html