디지털서비스

이슈리포트

Digitalservice Issue Report

엔터프라이즈 AI – 도입의 시대를 지나 거버넌스의 시대로

NIA 한국지능정보사회진흥원 디지털서비스 이용지원시스템

2026 Vol. 08

CONTENTS

Digitalservice Issue Report

디지털서비스 이슈리포트

01 엔터프라이즈 AI – 도입의 시대를 지나 거버넌스의 시대로

| 김영욱 SAP Product Engineering Product Expert

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

지난 5월, 국내 대기업 CIO들이 참석한 AI 관련 콘퍼런스에서는 관심사의 뚜렷한 이동을 관찰할 수 있다.[1] 1년 전 같은 종류의 행사에서 가장 많이 나온 질문은 "어떤 AI 모델을 도입할 것인가" 였다면, 이번엔 "도입한 AI를 어떻게 운영하고 통제할 것인가."였다. 데이터 전략, AI 운영체계, 거버넌스, 보안, 비용 관리, 내부 시스템 통합, 운영체계 구축과 같은 단어들이 행사 아젠다의 중심을 차지했다.

국내 대기업 다수가 조직 구조 자체를 AI 중심으로 재편하고 있고, 제로 트러스트 인프라를 사내 표준으로 정착시키고 있다. 표면적 사건들은 각기 다르지만, 이 사건들에서 반복되는 단어는 '거버넌스'다.

이 같은 시점에 두 개의 규제가 시행된다. 지난 8월 2일부터 유럽연합의 AI Act가 실질적으로 작동하는 국면으로 진입한다. 범용 인공지능(General-Purpose AI) 제공자에 대한 집행권과 제재권한이 발효되고, 제50조 투명성 의무가 시행되며, 전체 제재 체계를 가동한다.[2] 그리고 약 한 달 후인 9월 11일, 대한민국의 개정 개인정보 보호법이 시행된다. 과징금 상한이 관련 매출액에서 총매출액의 10%로 상향되고, 데이터 유출 시 책임 주체가 실무자에서 CEO와 이사회로 격상된다.[3] 서로 다른 대륙, 서로 다른 축의 두 규제이지만 같은 방향을 가리킨다. AI 시대의 문서화/입증 가능성을 사업의 표준 요구로 만든다는 방향이다.

산업의 관심 이동과 규제의 시행일 도래는 우연히 겹친 것이 아니다. 두 흐름은 같은 진단을 이야기한다: **만들기의 시대는 끝났다. 통제의 시대가 시작된다.** 지난 3년간 엔터프라이즈가 생성형 AI에 투자한 자원의 대부분은 파일럿 프로젝트를 만들고, 데모를 만드는 능력을 확보하는 데 쓰였다. 그

1) Zdnet Korea, "기업 AI '돈맥' 잡아라"…삼성SDS·LG CNS, AX 시장 선점 경쟁 치열," 2026.05.30.
2) 법률신문, "국내외 AI 규제, 2026년 하반기 분수령: EU AI Act의 새로운 국면과 AI기본법 시행령 개정," 2026.07.20.
3) 정책뉴스 Korea.kr, "개인정보 유출시 '최대 매출 10% 과징금'…CEO·CPO 책임 강화," 2026.03.09.

능력은 이미 확보되었고, 상품화되어 제품 형태로 시장에서 경쟁을 하고 있다. 만들기가 표준 조달 카테고리가 된 지금, 차별화는 **만들어진 것을 어떻게 통제/감사/책임지는가**에서 나온다.

이번 글에서는 세 가지 논지를 순차적으로 다루고자 한다.

첫째, 만들기의 시대가 끝났다는 진단의 근거와 그다음 단계인 통제의 시대의 작동 구조

둘째, 2026년 하반기에 겹치는 4개의 규제 시행일 – EU AI Act, 한국 AI 기본법 및 시행령 개정, 개인정보 보호법 개정 – 이 국내 기업에게 요구하는 실질적 대응

셋째, AI 에이전트 시대의 새로운 통제 대상인 비인간 아이덴티티(Non-Human Identity, NHI)의 상황을 소개한다.

1. 만들기의 시대는 어떻게 끝났는가?

74%와 21% – 도입과 거버넌스의 격차

딜로이트가 발표한 '기업에서의 AI 현황(State of AI in the Enterprise)' 리포트에서는 두 개의 결정적 숫자를 제시한다. 첫째, 74%의 기업에서 2027년까지 최소 중간 수준 이상의 AI 에이전트 활용을 예상한다. 도입 자체는 이미 사업 계획의 표준 항목이 되었다는 뜻이다. 둘째, 그러나 성숙한 에이전트 AI 거버넌스를 보유했다고 답한 기업은 21%에 불과하다.[4] 도입은 74%이고 거버넌스는 21%다. 이 격차가 지금의 상황을 설명한다. 같은 시점에 데이터브릭스의 '2026 State of AI Agents' 리포트에서는 AI 거버넌스를 명시적으로 구현한 기업은 그렇지 않은 기업에 비해 프로덕션 배포가 12배 더 많다고 말한다.[5]

이 세 개의 숫자가 던지는 신호는 명료하다. 무게중심이 만들기에서 통제로 이동했다는 사실이다. 지난 3년간 엔터프라이즈의 관심은 어떻게 만들 것인가에 집중되었다. 어떤 모델을 선택할 것인가, 어떤 인프라를 사용할 것인가, 어떤 데이터를 학습시킬 것인가? 이 질문들은 클라우드 하이퍼 스케일러의 제품 진화와 오픈소스 생태계의 성숙으로 대부분 해결되었다. 지금 시점에서 모델 성능이나 인프라 가용성 때문에 에이전트를 못 만드는 기업은 거의 없다. 그런데도 도입 대비 거버넌스 성숙도의 격차가 존재하고, 이 격차는 12배의 프로덕션 배포 능력으로 확인된다. 이 원인은 만들기의 능력이 아니라 만들어진 것을 다루는 능력에 있다.

4) Deloitte, "The State of AI in the Enterprise 2026 AI Report," Deloitte Insights, 2026.
5) Databricks, "State of AI Agents," 2026

개발 도구의 표준화

만들기가 표준 상품이 되었다는 두 번째 증거는 엔터프라이즈 AI 플랫폼 사업자들의 제품 발표에서 찾을 수 있다. 구글 클라우드 넥스트에서 공개된 제미나이 엔터프라이즈 에이전트 플랫폼은 비즈니스 팀을 위한 로우코드 개발 도구 에이전트 스튜디오와 개발자를 위한 코드 우선 에이전트 개발 키트를 이중 트랙으로 제공한다. 클라우드 인프라 사업자인 마이크로소프트와 구글, 그리고 SaaS 애플리케이션 사업자인 세일즈포스와 서비스나우가 정확히 같은 제품 형태로 진입했다. 산업 계층별로 서로 다른 위치에 있는 사업자들이 동일한 이중 트랙 구조로 수렴했다는 사실은 강력한 시그널이다. 특정 산업 계층의 트렌드가 아니라 엔터프라이즈 AI 조달 시장 전체의 표준 형태가 되었다는 뜻이다.

로우코드가 노코드가 아니라는 사실도 함께 주목할 만하다. 성숙한 플랫폼은 비즈니스 팀이 시각 빌더에서 조립하는 트랙과 개발 팀이 코드로 확장하는 트랙을 병행 운영하는 이중 구조로 수렴 중이다. 이 이중 구조가 요구하는 것은 더 나은 개발 도구가 아니다. 서로 다른 트랙에서 만들어진 에이전트를 통합 관리·통제·감사할 수 있는 프레임워크다. 만들기 능력 차이가 줄어들수록, 경쟁력의 원천은 만들기 이후 다음과 같은 단계로 이동한다.

- 생성된 에이전트가 조직의 규제·보안·감사·책임 프레임워크 안에서 지속적으로 작동할 수 있는가?
- 자동으로 실행한 결정을 사후에 재구성하고 책임 소재를 명확히 할 수 있는가?
- 도메인 특화 데이터와 평가 프레임워크가 플랫폼 벤더의 기본 제공을 넘어서는가?

이 세 가지 질문에 답할 수 있는 조직이 12배의 프로덕션 배포 격차를 만드는 상단에 있게 되는 것이다.

2. 8월 2일과 9월 11일: 두 규제의 의미

관리와 통제에 대한 요구는 산업 안에서만 있지는 않다. 규제 시행일이라는 외부 강제 요인이 비슷한 시점에 도래한다. 2026년 하반기에 국내 기업이 마주하는 4개의 규제 시행일 중 2026년 8월 2일 유럽연합의 AI Act와 2026년 9월 11일 개정 개인정보 보호법이 그것이다. 대륙도 다르고 규율 대상도 다른 두 규제이지만, 원칙은 같다. AI 도입의 모든 단계를 사후에 재구성하고 입증할 수 있는가를 묻는다.

EU AI Act 8월 2일 – 착각과 실제

EU 집행위원회는 2025년 말 '디지털 옴니버스'라 명명된 일괄 개정 법안을 통해 AI Act를 포함한 여러 디지털 규제의 시행 일정과 세부 조항을 한꺼번에 조정했다. 이 개정으로 고위험 AI 시스템에 부과되는

상세 의무 규율은 원래 2026년 8월 2일 시행 예정이었던 것이 2027년 12월까지 대폭 연기되었다. 그러나 이 연기가 8월 2일의 의미를 축소하지는 않는다. 8월 2일에는 여전히 결정적 조항을 발효한다.

범용 인공지능 모델 제공자에 대한 집행권과 제재 권한이 본격 가동되고, 제50조 투명성 의무가 시행된다. **챗봇을 사용할 때는 AI와 대화 중임을 명시해야 하고, 딥페이크에는 라벨을 부착해야 하며, AI가 생성한 콘텐츠에는 표시 의무가 발생한다.** AI 생성 콘텐츠의 표시·라벨링에 관한 실행규약의 최초 서명자 명단 등록 마감이 2026년 7월 22일이라는 점도 함께 짚어야 한다. 실행규약 서명은 자율사항이지만, 초기 서명이 규제 대응 성숙도의 시장 신호로 작용한다.

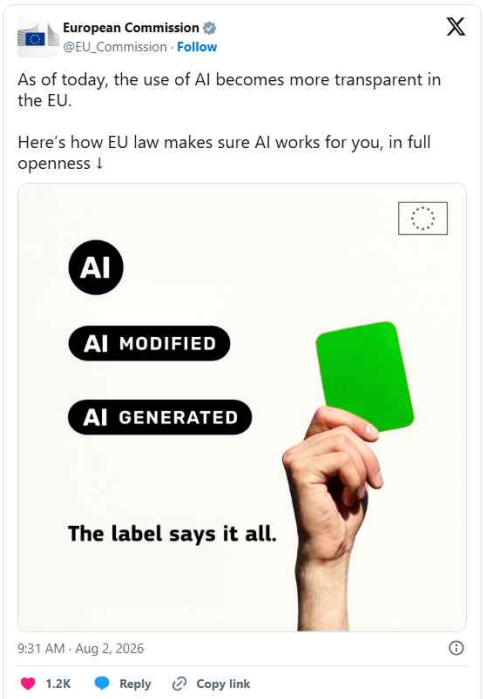

그림 1 EU AI Act의 투명성 의무 시행 발표(출처: x.com)

유럽 규제 자문 기업 비전 컴플라이언스(Vision Compliance)의 '2026 EU AI Act Readiness' 리포트는 준비 상태의 현실을 요약한다. 78%의 기업이 유럽연합 AI Act 대응에 준비가 안 된 것으로 보고 있다.[6] 대부분의 조직이 AI Act가 존재한다는 사실은 알지만, 그것이 실제로 무엇을 요구하는지를 이해하는 조직은 극히 드물다는 뜻이다. AI 시스템에 대한 공식 인벤토리를 확보하지 않은 상태이고, AI 컴플라이언스를 담당할 내부 거버넌스 조직을 지정하지 않았고, 고위험 AI 시스템에 필요한 기술 문서화

6) einnews.com, "Vision Compliance Releases 2026 EU AI Act Readiness Report," 2026.04.01

프로세스를 갖고 있지 않다. 이것은 데이터 거버넌스 기록, 모델 성능 지표, 인간 감독 절차 문서를 생성할 수 있는 체계가 없다는 뜻이다.

이 내용은 국내 기업에게 EU 시장에 진출한 국내 기업 다수가 즉각적 노출 상태에 있는 것과 지금 준비를 시작하는 기업은 여전히 상위 22% 안에 들어갈 여지가 있다는 두 가지 의미를 말하고 있다. 8월 2일은 준비 완료 시점이 아니라 준비 검증 시점의 시작이다.

9월 11일 개정 개인정보 보호법 – 책임의 격상

서문에서 짚었듯이 9월 11일에는 총매출액 10% 과징금과 CEO·이사회 책임 격상이라는 두 가지 격상이 발효된다. 그러나 개정법이 갖는 더 결정적인 의미는 한국 규제가 처음으로 "AI 시대의 합법적 데이터 사용 방식"을 구체적으로 명시했다는 점에 있다.

개정법이 담은 새 조항들의 방향은 명확하다. 가명정보를 안전하게 분석할 수 있는 공적 인프라, AI 학습을 위한 가명화 면제 조항, 그리고 생성형 AI를 서비스형(SaaS 형태로 사용)·기성형(기성 모델 도입)·자체 개발형(직접 훈련)의 3단계로 나누어 위험을 단계별로 관리하는 지침을 마련했다. 개정법 이전까지 한국 규제의 기본 자세는 "개인정보를 이렇게 하면 안 된다"는 금지 목록이었다. 개정법은 여기서 한 걸음 나아가 "AI 시대에 이렇게 하면 합법이다" 라는 허용의 경로를 처음으로 제시한다. 이 전환이 국내 기업에 던지는 함의는 명확하다. AI 도입 시 관리해야 할 대상은 "금지되는 것"이 아니라 "입증해야 하는 것"이다. 자사의 AI 데이터 사용이 허용 경로 위에 있음을 사후에 감사자에게 재구성해서 보여줄 수 있어야 한다. 이 요구는 EU AI Act가 8월 2일부터 요구하는 문서화·입증 가능성과 정확히 같은 방향이다. 서로 다른 규제 축이지만 같은 인프라를 요구한다.

두 시행일의 의미

두 규제가 요구하는 문서화·입증 가능성은 실무 차원에서 세 가지 인프라를 요구한다.

- **AI 시스템 인벤토리:** 조직 내 어떤 AI 시스템이 어떤 데이터로 어떤 결정을 자동화하고 있는지이다. 앞서 살펴본 비전 컴플라이언스 조사에서 83%의 기업이 이 첫 단계조차 확보하지 못한 상태라는 사실은 대부분의 조직이 규제 대응의 출발점에도 아직 서 있지 못하다는 것을 의미한다.
- **위험 분류:** EU AI Act의 4단계 분류(금지·고위험·제한적 위험·최소 위험) 또는 한국 AI 기본법의 고영향 AI 기준에 따른 자산 분류
- **지속적 모니터링과 감사 로그:** 규제 조사나 사고 발생 시 사후 재구성이 가능한 형태의 실행 로그

이 세 가지 요구사항은 규제별로 다르지 않다. 어떤 규제에 대응하든 동일한 인프라가 필요하며, 개별 규제에 각각 별도로 대응하는 방식은 자원과 시간의 낭비다. 다음에서 다룰 한국의 규제 3중 스택이 국내 기업에게 통합 대응을 요구하는 이유가 여기에 있다.

3. 한국의 규제 3중 스택

이 장에서는 한국이 마주한 규제 시점의 특수성을 다룬다. 국내 기업이 8개월 안에 통과해야 하는 4개의 규제 시행일이 있고, 이 상황은 국내 기업에게 부담이자 동시에 기회다.

한국이 세계에서 가장 이른 AI 규제 전면 시행 국가가 되는 이유

EU가 고위험 AI 시스템 규제를 2027년 12월까지 연기한 반면, 한국의 AI 기본법은 이미 2026년 1월 22일부터 전면 시행되었다.[7] 이 사실이 뜻하는 바는 명료하다. EU AI Act가 세계 최초로 통과된 포괄적 AI 규제이긴 하지만, 실질적으로 AI 규제를 가장 먼저 전면 적용하는 국가는 한국이다. 한국이 AI 규제의 실질적 선도자가 되는 사건이 2026년 상반기에 이미 시작된 셈이다.

이 사실이 국내 기업에 던지는 함의는 양면적이다. **부담 측면에서는 국내 시장에서 AI를 도입·서비스하는 모든 사업자가 다른 나라 사업자보다 먼저 규제 대응 부담을 진다. 그러나 기회 측면에서는 국내에서 먼저 검증한 규제 대응 역량이 EU·미국 시장 진출 시 차별화 자산이 된다.** 지난 20년간 국내 기업이 글로벌 표준의 추종자였다면, AI 규제 대응에서는 실질적 선도자의 위치에 서게 될 수 있는 것이다.

한국 규제 3중 스택

국내 기업이 마주한 규제 스택은 세 개의 국내 규제와 하나의 EU 규제로 구성된다. 8개월 안에 4개의 시행일을 통과해야 한다. EU AI Act 8월 2일은 앞서 다뤘고, 국내 3중 스택 중 하나인 개정 개인정보보호법 9월 11일도 앞서 다뤘다. 나머지 두 국내 규제를 알아보자.

첫째, 2026년 1월 22일 AI 기본법 시행: 위험 기반 접근을 기본 구조로 채택하며, 고영향 AI 로 분류되는 시스템에 대해 투명성·안전성·특별 책무·영향 평가·국내 대리인 지정의 5가지 의무를 부과한다. 자율주행 AI, 의료기기 AI, 신용평가·대출 AI, 채용·인사 AI 등이 고영향 AI 분류 대상이며, 국내에서 AI를 활용하는 기업 대다수가 최소 하나 이상의 고영향 AI 시스템을 운영하고 있을 가능성이 매우 높다.

7) 국가법령정보센터, "인공지능 발전과 신뢰 기반 조성 등에 관한 기본법," 2026.01.21.

둘째, 2026년 7월 21일 개정 AI 기본법 시행령: 개정 시행령은 "AI 제품·서비스 확인 제도"를 신설했다. 한국 인공지능/소프트웨어 산업협회 또는 한국 정보통신 기술협회 기반의 확인서를 취득하는 사업자가 공공 조달에서 우대받는 구조다. 이는 국내 조달 시장 진입에 직접 영향을 미치는 조항이며, 확인서 취득 자체가 시장 신호로 작용한다. 앞서 EU AI Act의 실행규약(Code of Practice) 초기 서명이 규제 대응 성숙도의 시장 신호로 작용하는 것과 같은 구조다.

4개 규제, 하나의 인프라

4개의 시행일에 각각 별도 프로젝트로 대응하는 접근은 자원과 시간의 낭비다. 각 규제의 표면적 요구사항은 다르지만, 실제로 필요한 인프라는 상당 부분 겹친다. 2장에서 말한 세 가지 실무 요구사항 – AI 시스템 인벤토리, 위험 분류 체계, 문서화·입증 가능성을 위한 감사 로그 – 은 EU AI Act에도 필요하고, 한국 AI 기본법에도 필요하며, 개정 개인정보 보호법 대응에도 필요하다. 각 규제를 개별 프로젝트로 다루는 조직과 통합 인프라 위에서 다루는 조직 사이에는 시간이 지날수록 대응 효율과 비용에서 상당한 격차가 발생한다.

법률신문은 "EU와 한국은 규제 철학은 다르지만, 두 법제 모두 AI를 실제로 어떻게 설계하고 활용하며 이를 어떻게 문서화·입증하는가를 핵심으로 삼는다는 점에서 수렴한다."로 표현하고 있다. 시한 연기가 있는 영역이라도 준비의 강도를 낮출 것이 아니라, 확보된 시간을 분류·문서화·거버넌스 정비에 활용하는 조직이 규제 리스크 관리와 사업 기회 포착 모두에서 유리한 위치를 점하게 될 것으로 판단한다.

4. 비인간 아이덴티티(NHI) 거버넌스 공백

AI 에이전트 시대의 새로운 통제 대상은 사람이 아니라 비인간 아이덴티티(Non-Human Identity, NHI)다. 서비스 계정, API 키, OAuth 토큰, 머신 인증서, 그리고 AI 에이전트가 소유한 자격증명 전체를 뜻한다. 이들을 발견하고, 소유권을 명시하며, 지속적으로 통제할 수 있는가가 지금의 새로운 거버넌스 과제다.

144대 1이라는 새로운 진실

Cloud Security Alliance(CSA)는 지금의 상황을 하나의 숫자로 요약한다. **클라우드 네이티브 환경에서 NHI가 사람 사용자를 144대 1로 초과한다.** 2024년 상반기의 92대 1에서 1년 만에 56% 증가한 수치다.[8] 전체 엔터프라이즈 환경에서는 평균 45대 1 수준이지만, 클라우드 채택, 마이크로서비스, SaaS 통합의 확산으로 격차가 지속적으로 벌어지고 있다.

8) Cloud Security Alliance, "The Non-Human Identity Governance Vacuum," 2026.05.20.

이 숫자가 의미하는 것은 표면 이상이다. 한 기업의 인프라 안에서 실제로 일을 수행하는 주체의 99% 이상이 사람이 아니라 기계라는 뜻이다. 이 기계들은 지속적으로 인증하고, 민감한 시스템에 접근하며, 사람 계정이 같은 권한을 가지고 있었다면 즉시 문제로 지적될 수준의 권한을 보유하고 있다. 그런데도 지금까지 엔터프라이즈 보안의 20년은 사람 아이덴티티를 지키는 데 집중되어 왔다. MFA, 권한 있는 접근 관리(PAM), SIEM 기반 사용자 행위 감사 등은 모두 사람이 로그인한다는 전제 위에서 설계된 것이다. 사람 계정을 방어하는 데 투자한 자원의 규모와 실제 방어해야 할 계정의 규모 사이에 이미 100배 이상의 격차가 존재한다.

78%가 정책조차 없다

CSA 함께 보고하는 세 개의 세부 통계는 이 격차의 실질적 상태를 드러낸다.

첫째, 78%의 조직이 AI 아이덴티티를 생성하거나 회수하는 문서화된 정책 자체를 갖고 있지 않다. 정책 없음의 문제는 정책 부실보다 더 근본적이다. 정책이 없으면 위반 여부를 판단할 기준조차 없다.

둘째, 51%의 조직이 AI 아이덴티티의 명시적 소유권이 없다고 답했다. 어떤 AI 에이전트가 누구의 책임 하에 어떤 시스템에 접근하는지에 대한 답이 조직 내부에 존재하지 않는다는 뜻이다.

셋째, API 키를 공식적으로 오프보딩하고 회수하는 프로세스를 갖고 있는 조직은 20%에 불과하다. 나머지 80%의 조직에서는 사용을 마친 API 키가 계속 살아 있고, 그 키가 살아 있는지조차 모르는 상태에서 시간이 흐른다.

소포스(Sophos)의 조사는 이 공백의 결과를 보여준다. 71%가 지난 1년 동안 최소 한 건의 신원 관련 침해를 경험했고, 평균 복구 비용은 164만 달러다. 그중 41%의 사고에서 NHI 관리 부실이 원인으로 지목되었다.[9] 이 통계는 NHI 거버넌스 공백이 이론적 위험이 아니라 이미 실현된 손실이라는 것을 증명한다.

5. 국내 기업이 고려해야 하는 관점

NHI 거버넌스 공백과 여러 규제 시행일은 국내 기업에게 동시에 도래하는 통제 과제다. 그 과제 앞에서 국내 엔터프라이즈의 의사결정자가 함께 고려할 수 있는 세 가지 관점을 정리해 본다.

첫째, RFP 설계의 진화다. 통제의 시대에는 제품 사양과 단가 외에 접근 권한 거버넌스, 정보 범위

9) Sophos, "The State of Identity Security 2026," 2026.05.12.

명시, 사고 대응 보고 체계, 성과 기반 계약 옵션 같은 요소들이 자연스럽게 함께 검토된다. NHI 거버넌스 공백을 계약 단계에서 미리 정리하는 자연스러운 진화의 한 부분이다. AI 도입 계약이 시스템을 사들이는 결정을 넘어 접근 권한을 가진 인력과 자동화된 에이전트를 조직 안에 들이는 결정이라는 성격을 함께 갖게 되었기 때문이다.

둘째, 글로벌 모델과 한국 도메인 자산의 결합 설계다. 글로벌 AI 사업자들의 한국 시장 참여가 본격화되는 시점에서, 한국 도메인 자산이 어떻게 결합되고 그 결합 결과가 어떻게 재사용 가능한 형태로 변환되는가는 한국 엔터프라이즈 전체의 AI 전환 효율을 결정하는 변수가 된다. 한국 금융권 규제 환경, 한국 정부 조달 프로세스, 한국 제조업 그룹사 간 거래 구조는 글로벌 표준 모델만으로는 단기에 완전히 커버되지 않는 고유성을 갖는다. 이 갭을 메우는 결합 설계가 개별 기업의 의사결정인 동시에 한국 IT 산업 생태계 전체에 영향을 미친다.

셋째, 내부 인력 풀의 역할 재인식이다. 한국 엔터프라이즈와 IT 서비스 산업에 축적된 모호한 요구 구조화·다층 이해관계자 관리·프로덕션 안착 경험을 가진 인력 풀은, 만들기 이후의 통제 프레임워크 설계에 정확히 필요한 역량과 부합한다. 이 인력 풀이 글로벌 AI 협업의 사내 카운터파트로 활용될 때, 통제력과 학습 흡수력을 동시에 확보할 수 있다.

6. 마치며 – 통제의 시대가 시작된다

2026년 하반기, 국내 기업은 4개의 규제 시행일을 통과한다. 같은 하반기, 산업 차원의 무게중심은 에이전트 생성에서 에이전트 거버넌스로 이동한다. 두 흐름이 우연히 겹친 것이 아니다. 규제와 산업이 같은 진단에 도달한 것이다.

만드는 것이 도입을 거쳐, 성숙의 시대에 들어가면, 진짜 경쟁력은 만들어진 것을 어떻게 통제/감사/책임지는가에서 나온다. 즉 규제 대응만이 아니라 사업 경쟁력이 되는 것이 2026년 하반기의 본질이다. 통제할 수 있는 기업은 더 많이 배포할 수 있다. 통제할 수 없는 기업은 배포 자체가 리스크가 된다. 데이터브릭스가 찾아낸 "거버넌스를 구현한 기업은 프로덕션 배포가 12배 많다"는 통계가 이 원칙을 요약한다고 할 수 있다.

02 EU의 클라우드와 AI 개발법 (CADA)

| 한상기 테크프론티어

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

유럽 연합 집행위원회(EC)는 2026년 6월 3일 유럽의 클라우드·AI 생태계 강화를 위한 규정안(CADA)을 발표했다.[10] 이는 기술 주권을 확보하고자 하는 노력으로 평가할 수 있으며, 데이터센터 용량 부족과 소수 비EU 클라우드 사업자 의존이라는 두 취약점을 해결하고자 하는 목적을 갖는 것으로 보인다. EC가 제시한 이해관계자 의견 수렴 기간은 2026년 7월 2일부터 8월 27일까지이다.

법안의 핵심은 4단계 주권 보증 체계(Union assurance level)로, 3단계는 EU 소유·통제 및 인력 국적 등 추가 기준을 요구하고, 4단계는 소프트웨어 공급망 전반에 대한 완전한 투명성·통제와 제3국 간섭 부재를 요구하고 있다. 아직 입법 제안 단계이며 최종 채택 목표는 2027년 4분기이다.

CADA는 5~7년 내 EU 데이터센터 용량을 최소 3배로 늘리고 2035년까지 EU 기업과 공공행정의 수요를 완전히 충족하는 것을 목표로 한다. 정식 문서번호는 COM(2026)502, 입법절차번호는 2026/0138(COD)이며, '클라우드 컴퓨팅 서비스'의 정의는 NIS2를 준용한다. 지침이 아닌 규정(Regulation) 형태로 채택이 되면 회원국 법으로 전환 없이 직접 적용하게 된다.

배경과 입법 취지

2026년 6월 3일 EC가 발표한 CADA 제안은 EU 디지털 환경의 두 가지 치명적 취약점인 데이터센터 용량의 구조적 부족과 소수 비EU 클라우드 사업자에 대한 의존 문제를 풀고자 하는 것이다. 집행위원회 설명 자료에 따르면 EU 기반 사업자의 유럽 클라우드 시장 점유율은 2017년 약 29%에서 2022년 약 15%로 떨어졌고, 동시에 데이터센터 용량 부족이 AI 배치를 제약하고 있다는 것이 EC의 진단이다.

비르쿠넨 집행부위원장은 핵심 워크로드를 맡는 클라우드 사업자가 '킬 스위치'를 갖지 못하도록 하겠다고 밝혔고, 미국 클라우드 법 때문에 미국 기업이 최고 주권 등급에 도달하기 어려울 것이라고

10) Inside Global Tech, "The EU Cloud and AI Development Act in Depth," Jun 11, 2026

덧붙였습니다. 킬 스위치라는 것은 클라우드 사업자(또는 그 사업자를 통제하는 제3국 정부)가 어떤 이유로든 서비스를 끊거나 저하시킬 수 있는 능력을 말한다. 유럽이 걱정하는 시나리오는 대략 이렇다.

- 미국 정부가 제재나 수출통제를 발동해 특정 유럽 고객(예: 어떤 기관, 기업, 국가)에 대한 서비스 제공을 금지하면, 미국 본사는 그 명령을 따라야 하고 유럽 자회사도 결국 서비스를 중단해야 한다.
- 미·EU 관계가 악화될 때 미국 행정부가 정치적 압박 수단으로 클라우드 접근을 위협할 수 있다. 2025년 국제형사재판소(ICC) 검사장이 미국 제재 대상이 되자 마이크로소프트 이메일 계정이 차단된 사례가 유럽에서 자주 인용되는 실제 있었던 킬 스위치 사례이다.
- 전쟁이나 위기 시 병원, 전력망, 행정 시스템이 외국 기업의 결정 하나에 멈출 수 있다는 것이 본질적 우려다.

그래서 CADA 2단계는 '제3국 통제 사업자는 서비스 연속성을 보장하고 제3국 제재·무역통제 규제 노출을 회피하는 조치를 갖춰야 한다'고 요구하고, 3단계 예외 조건에 '해당 제3국이 클라우드 사업자에게 서비스 저하·중단을 강제할 수 없어야 한다'는 문구가 들어간 것이다. 이 조항들이 바로 킬 스위치 제거 요건이라고 말하는 것이다.

미국 클라우드 법 때문에 미국 기업이 최고 주권 등급에 도달하기 어려울 것이라고 보는 것은 뒤에 설명하는 4단계 주권 보증 체계에서 3단계와 4단계는 '사업자가 EU 내에서 소유·통제될 것'을 요구하고 있고, 소유·통제 요건은 데이터 위치 요건과 다르기 때문이다. 데이터 위치는 리전 추가와 같은 방식을 쓰면 돈으로 해결할 수 있지만, 소유·통제는 기업 구조 자체를 바꿔야 한다. 그래서 비르쿠넨이 '미국 기업은 최고 등급에 도달하기 어렵다'고 한 것이고, 이는 기업의 잘못이 아니라 미국 법 때문이라는 점도 그 말에 담겨 있는 의미이다.

이 논리에 대한 반론도 있다. 첫째, EU 자체도 8월 18일부터 적용된 e-Evidence 규정으로 회원국 간 데이터 직접 제출을 의무화했으므로 정부의 데이터 접근 자체는 EU도 하고 있다는 지적이다. 차이는 관할권이 EU 안에 있느냐 밖에 있느냐이다. 둘째, 집행위원회가 4월 주권 클라우드 입찰에서 구글 기술을 쓰는 S3NS 컨소시엄을 선정한 것처럼, 유럽 기업이 운영하고 미국 기업은 기술만 제공하는 신탁 모델이 3단계 예외의 문을 여는지가 향후 3자협상의 핵심 쟁점이 될 것이다.

핵심 조항 네 가지

이 법안에는 다음과 같은 네 가지의 핵심 조항이 있다.

1. 클라우드 주권 프레임워크 — 4단계 주권 보증 체계
2. 공공조달의 "EU 기여도" 기준
3. 데이터센터 가속구역과 전략 프로젝트 (공급 측면)
4. 오픈소스 우선 원칙

클라우드 주권 프레임워크

가장 논쟁적인 부분으로, 공공부문에 서비스를 제공하려는 거의 모든 클라우드 사업자는 최소 1단계를 충족해야 하며, 공공질서 유지에 기여하는 분야의 사업자는 회원국과 EU 기관이 수행하는 위험평가 결과에 따라 더 높은 단계를 요구받을 수 있다. 대상 분야는 NIS2가 규율하는 에너지·보건·교통·수도 등에 더해 국가안보, 국경관리, 법집행, 국방까지 포함한다.

각 단계의 요건은 다음과 같다.

- **1단계:** 기본이지만 이미 상당한 의무를 부과한다. 인프라·자산·고객 데이터를 EU 내에 두어야 하고, 제3국 통제를 받는 사업자는 미공개 취약점을 제3국 공공당국에 보고할 의무가 없음을 보증해야 한다.
- **2단계:** 서비스 운영에 관여하는 모든 인력·인프라·자산이 EU 내에 위치해야 하고, 아직 확정되지 않은 EU 클라우드 인증제도 인증을 받아야 하며, 제3국 통제 사업자는 서비스 연속성 보장, 제3국의 고객 데이터 접근 차단, 제3국 제재·무역통제 규제 노출 회피를 위한 법적·기술적·조직적 조치를 갖춰야 한다. 또한 완전한 소프트웨어 부품명세서(SBOM)를 유지하고, 제3국 소프트웨어 구성요소에 대해 소스코드 감사를 실시하며, EU 밖 자회사가 있는 경우 EU 모회사와 자회사 사이에 실효적인 법적·기술적·조직적 분리를 확보해야 한다.
- **3단계:** 사업자가 EU 내에서 소유·통제되어야 한다. 유일한 예외는 집행위원회가 2차 입법으로 해당 제3국이 일정 조건을 충족한다고 인정하는 경우인데, 그 조건에는 해당국이 GDPR 적정성 결정을 보유하고, 클라우드 사업자에게 서비스 저하·중단을 강제할 수 없으며, EU 클라우드 사업자에게 개방된 시장을 유지한다는 것을 포함한다.
- **4단계:** 가장 엄격한 단계로 3단계처럼 제3국 통제를 받지 않아야 하되 제3국에 대한 예외조차 없다. 추가로 클라우드 서비스를 다루는 유럽 사이버보안 인증제도상 최소 '높음' 등급의 인증, 모든 소프트웨어 구성요소에 대한 실효적 통제 유지, 그리고 어떤 제3국이나 제3국 주체도 해당 소프트웨어 구성요소의 설계·개발·유지·진화에 대한 실효적 통제권을 갖지 않음을 입증할 것을 요구한다.

이 프레임워크의 적용 범위가 공공부문에 그치지 않을 가능성도 있다. CADA는 NIS2상 '필수 주체'로 규제하는 민간 부문 주체도 클라우드 사업자에 대해 유사한 위험평가를 수행할 수 있도록 상정하며, 집행위원회는 향후 특정 분야 주체에게 이를 의무화하는 2차 입법을 할 수 있기 때문이다.

공공조달의 EU 기여도 기준

공공기관은 클라우드·AI 서비스 조달 시 '유럽연합기여도(Union added value)'를 비가격 기준으로 고려해야 한다. 즉 입찰 낙찰 결정 시 사업자가 EU 디지털 기술 공급망 강화에 얼마나 기여하는지, EU에서 개발된 기술을 통합하는지, EU 내에서 혁신을 수행하는지, EU에서 설계·제조된 하드웨어 부품으로 서비스를 제공하는지를 따져야 한다. 이 기준은 "부수적이며 결정적이지 않아야" 한다고 명시되어 있고, 전문은 120점 만점에 최대 15점의 가중치를 시사하지만, 그럼에도 EU 내 R&D·제조·혁신 활동이 많은 사업자에게 구조적 이점을 준다.

유럽연합 기여도는 EU 법률 용어로, 원래는 EU 예산 편성 원칙에서 왔다. 어떤 사업을 회원국 차원이 아니라 EU 차원에서 수행할 때 추가로 생기는 가치, 즉 EU가 개입함으로써 유럽 전체에 더해지는 효익을 뜻한다. 보조성 원칙(subsidiarity)과 짝을 이루는 개념으로, 호라이즌 유럽이나 결속기금 같은 EU 프로그램의 자금 배분 기준으로 오래 쓰여 왔다.

CADA는 이 용어를 공공조달 평가 기준으로 가져왔다. 여기서의 의미는 구체적으로 "이 사업자를 선정하면 EU 기술 생태계에 무엇이 남느냐"이다. 앞서 본 네 가지 판단 요건, 즉 EU 디지털 기술 공급망 강화 기여, EU 개발 기술 통합, EU 내 혁신 수행, EU 설계·제조 하드웨어 사용이 그 판단 요건이다.

이는 실질적으로는 'Buy European' 조항을 WTO 정부조달협정(GPA)과 충돌하지 않게 포장한 것이다. 국적이나 소재지를 직접 기준으로 삼으면 차별 조항이 되지만, 유럽 생태계에 대한 기여를 부수적 비가격 기준으로 두면 형식상 비차별적이다. 120점 중 최대 15점이라는 가중치 제한도 그런 법적 방어를 염두에 둔 것이다.

데이터센터 가속구역과 전략 프로젝트 (공급 측면)

회원국은 최소 한 곳 이상의 데이터센터 가속구역을 지정해야 하며, 이 구역의 데이터센터 프로젝트는 간소화된 인허가 절차, 통합 기본 허가, 최대 12개월의 허가 기한 혜택을 받는다. 집행위원회는 특정 데이터센터 프로젝트를 전략 프로젝트로 지정할 수 있고, 이 경우 추가 공공지원과 EU 자금에 대한 우선 접근이 가능하다. 다만 지속가능성 요건 강조, 그린필드보다 브라운필드 부지 선호, 계통연계 조건 등은 투자 결정에 제약으로 작용할 수 있다.

데이터센터 가속구역(data centre acceleration zone)은 회원국이 미리 지정해 둔 데이터센터 건설 특구를 말한다. 이 구역 안에 데이터센터를 지으려는 사업자는 일반 지역보다 훨씬 빠르고 단순한 인허가 절차를 적용받는다. 한국의 산업단지나 경제자유구역처럼 '여기에 지으면 행정 절차를 빨리 처리해 주겠다'고 정부가 공간적으로 약속하는 제도로 이해할 수 있다.

이런 지정이 왜 필요하냐면, 유럽에서 데이터센터 건설의 최대 병목은 부지나 자본이 아니라 행정 절차이기 때문이다. 독일만 해도 신규 하이퍼스케일 데이터센터 허가에 계획, 환경영향평가, 계통연계 절차를 거치며 3~5년이 걸릴 수 있다. 집행위원회는 5~7년 안에 용량을 3배로 늘리겠다고 했으니, 허가에만 3~5년이 걸리는 구조로는 목표 달성이 불가능하다. 가속구역은 이 시간을 줄이기 위한 장치이다.

그린필드는 한 번도 개발된 적 없는 땅을 말한다. 농지, 초지, 숲처럼 말 그대로 '푸른 들판'이었던 곳에 처음으로 건물을 짓는 경우를 뜻한다. 장점은 기존 구조물이나 오염이 없어서 설계가 자유롭고 공사가 단순하다는 것이고, 단점은 자연이나 농지를 잠식하고 전력·통신·도로 같은 기반시설을 처음부터 끌어와야 한다는 것이다.

브라운 필드는 이전에 산업·상업 용도로 쓰였다가 버려지거나 저이용 상태인 땅을 말한다. 폐공장, 옛 제철소, 폐광, 폐쇄된 군사기지, 옛 발전소 부지 같은 곳이다. '갈색'은 오염되거나 황폐해진 땅이라는 뉘앙스에서 왔다. 장점은 이미 도시 기반시설이 있고 자연을 새로 훼손하지 않는다는 것, 단점은 토양 오염 정화 비용이 들고 기존 구조물 철거 등 변수가 많다는 것이다.

데이터센터 맥락에서 브라운필드가 특히 주목받는 이유가 하나 더 있다. 옛 발전소나 제철소 부지는 이미 고압 송전선과 대용량 변전 설비가 연결돼 있는 경우가 많다. 데이터센터의 최대 병목이 전력 계통 연결이라는 점을 생각하면, 전력 인프라가 살아 있는 브라운필드는 그린필드보다 훨씬 빨리 가동할 수 있다. 미국에서 트럼프 행정부의 2025년 7월 행정명령이 EPA에 슈퍼펀드·브라운필드 부지를 AI 데이터센터 용도로 재활용하는 지침을 만들도록 한 것도 같은 논리이다.

CADA가 가속구역 지정 시 브라운필드 선호를 명시한 것은 두 가지를 동시에 노린 것이다. 하나는 환경 보호(신규 토지 잠식 억제)이고, 다른 하나는 실제 건설 속도(기존 기반시설 활용)이다. 다만 앞서 본 비판대로, 오염 정화나 노후 송전망 교체 같은 물리적 제약은 법이 선호한다고 해서 사라지지 않기 때문에 브라운 필드가 무조건 더 유리하다고 볼 수만은 없다.

오픈소스 우선 원칙

EU 공공부문에 오픈소스 우선 원칙을 법제화한다. 공공기관이 클라우와 ·AI 생태계를 구축할 때 개방형 표준과 오픈소스 구성요소의 사용 및 재사용을 장려하는 조치를 취해야 하고, 공공기관이 또는 공공기관을 위해 개발한 소프트웨어는 중앙 EU 오픈소스 솔루션 카탈로그를 통해 재사용 가능하도록 공개해야 하며, 오픈소스 프로그램 사무소 네트워크를 설립한다. 이에 관련한 조항은 다음과 같다.

- 제41조 "오픈소스 우선": EU와 회원국은 EU 기관과 공공부문이 클라우드·AI 생태계나 스택을 구축할 때 개방형 표준과 오픈소스 라이선스 구성요소를 사용하고 재사용을 촉진하도록 장려하는 조치를 취해야 하되, 보안을 포함한 기능성, 총비용, 기타 정당한 객관적 기준을 고려한다고 명시하고 있다.
- 제42조 "소프트웨어 공유·재사용": EU 기관이나 공공부문이 지식재산권을 보유한 소프트웨어에만 적용되며, 공공 자금으로 만든 코드를 다른 기관이 재사용할 수 있게 공개하는 내용이다.
- 제43조 "EU 오픈소스 솔루션 카탈로그": 공공부문이 공유하는 소프트웨어를 모아두는 중앙 저장소이다.
- 제44조 "OSPO 네트워크": 본 법의 오픈소스 조항 이행에서 협력을 촉진하기 위해 오픈소스 프로그램 사무소 간 네트워크를 설립한다.

OSPO는 원래 기업에서 오픈소스를 전략적으로 다루는 회사의 창구 개념으로 시작한 것이다. 2020년대 들어 이 모델이 정부로 들어왔으며 EC 자체가 2020년에 EC OSPO를 만들었고, 프랑스(DINUM), 독일(ZenDiS), 네덜란드 등 여러 회원국 정부와 일부 도시(뮌헨, 바르셀로나)도 공공 OSPO를 두고 있다. EC 오픈 소스 전략은 집행위원회 OSPO, EU 공공부문 OSPO 네트워크, 상호운용 가능한 유럽(Interoperable Europe) 메커니즘을 강화하고, 집행위 저장소에 대한 공통 보안 기준(모니터링, 취약점, 라이선스 준수, 의존성 위험)을 설정하겠다고 밝혔다.

오픈소스 프로그램 사무소(OSPO) 네트워크 이미 비공식적으로 존재하던 것인데, CADA 제44조가 이를 법률상 기구로 격상시키는 것이다. 그 역할은 회원국과 EU 기관의 OSPO들이 연결되어 (1) 카탈로그에 올라온 소프트웨어를 다른 회원국이 실제로 가져다 쓰도록 돕고, (2) '오픈소스 우선' 원칙을 조달 현장에서 어떻게 적용할지 공통 지침을 만들고, (3) 회원국 간 경험과 모범사례를 교환하는 것으로 해석할 수 있다.

오픈소스 전문 기업인 SUSE는 이를 두고 '오픈소스 전문성에 제도적 거처를 주고 조달 지침에서 역할을 부여한 것'이라고 평가했다.[11] 조달 담당 공무원 대부분이 오픈소스 대안을 평가할 역량이 없다는 점이 오랜 문제였는데, OSPO 네트워크는 그 역량을 집단적으로 공급하는 장치가 된다.

또한 핵심 공유 인프라를 위한 전용 오픈 소스 유지 관리 기금을 포함하여 오픈 소스 전략을 위한 20억 유로 규모의 투자 계획를 발표했으며, 클라우드 및 AI 리더십 이니셔티브의 운영 목표 2(제4조)는 '전략적 부문을 위한 온디바이스 엣지, 연결성, 데이터 및 AI 도구, 백엔드 및 서비스 계층을 포괄하는 안전하고 탄력적이며 성능이 뛰어난 개방형 클라우드 컴퓨팅 스택'의 개발 및 시범 운영을 의무화하고

11) SUSE, "The EU Cloud and AI Development Act: What It Gets Right, and What It Still Needs," Jun 16, 2026

있다. 이는 아키텍처 측면에서 볼 때, 오픈 소스 인프라 제공업체들이 이미 오늘날 제공하고 있는 스택과 정확히 일치하는 내용이라는 것이 SUSE의 평가이다.

그러나 오픈소스 진영은 방향은 환영하면서도 구속력 부족을 비판했다. 프랑스 오픈소스 기업 연합 CNLL은 제41조 제목이 "오픈소스 우선"이라는 야심찬 원칙을 내세우지만 본문은 훨씬 약하다고 지적하며, 동사가 요구가 아니라 장려라는 점, 즉 EU와 회원국의 의무가 '장려하는 조치를 취하는 것'에 그치는 형태라는 점을 문제 삼았다.[12] OSPO 네트워크가 아무리 잘 작동해도 제41조가 '장려'에 그치면 조달 담당자가 독점 소프트웨어를 선택하는 것을 막을 수 없다는 것이 비판의 핵심이다.

예상할 수 있는 쟁점들과 향후 일정

이번에 추진하는 CADA 법안이 확정되면 생각할 수 있는 몇 가지 쟁점이 있다. 일단 회원국과 EU 기관이 위험평가를 완료하고 요구 등급을 결정할 때까지 클라우드 사업자에게 상당한 불확실성이 생기며, 그동안 공공기관이 클라우드 이전을 꺼려 공공부문 클라우드 도입이 오히려 늦어질 수 있다.

두 번째는 EC가 위험평가 템플릿을 만들더라도, 회원국마다 서로 다른 활동에 다른 보증 등급을 적용하면서 회원국 간 상당한 편차가 생길 여지가 남아 있다. 세 번째로는 주권 프레임워크는 비EU 클라우드 사업자 의존을 줄이려는 현재 EU 정책 목표를 법제화하는 것으로, 이사회와 유럽의회 간 3자협상의 핵심 정치적 전장이 될 가능성이 크다.

업계 반응은 엇갈리고 있다. 한편 EC는 4월의 1.8억 유로 주권 클라우드 입찰에서 구글 클라우드와 탈레스의 합작사 S3NS를 쓰는 프락시머스(Proximus) 컨소시엄을 선정하면서 '비유럽 기술도 엄격하고 적절한 틀 안에서 운영되면 요구되는 최소 주권 수준을 충족할 수 있다'고 직접 밝힌 바 있어, 하이퍼스케일러의 파트너십 모델이 2단계까지는 열려 있을 가능성을 시사한다.

CADA에 대한 공개 의견수렴은 2026년 7월 2일부터 8월 27일까지 진행 중이고, 아직 입법 제안 단계로 최종 채택 목표는 2027년 4분기이다. 27개 회원국 전체의 승인이 필요하며, 일부 웹사이트에 '2026년 8월 4일 발효, 1단계 요건 2028년 2월 적용, 최고 단계 2029년 8월 의무화' 같은 구체적 날짜가 돌고 있으나 이는 공식적으로 확인된 것이 아니라는 것을 참고하기 바란다.

12) CNLL, "Stratégie open source européenne : deux reculs significatifs entre le projet circulé fin mai et le texte adopté le 3 juin," Jun 4, 2026

03 모델이 곧 컴퓨터다: AMD의 탈라스 인수와 모델 특화 반도체

| 윤대균 아주대학교 소프트웨어학과

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

1. 들어가며

AMD가 2026년 8월 6일 캐나다 토론토의 반도체 스타트업 탈라스(Taalas) 인수 계약을 발표했다.[13] 탈라스는 2023년 설립된 직원 25명 규모의 회사로, AI 모델의 가중치를 반도체 회로에 직접 새겨 넣는 방식의 추론 전용 칩을 개발한다. AMD는 인수 사유로 범용 아키텍처가 가진 연산과 메모리의 병목을 줄이는 탈라스의 추론 데이터 흐름 기술을 들었고, 이를 자사의 인스팅트(Instinct) GPU와 결합한 시스템 개발에 활용한다고 밝혔다. 최근 9개월 사이 AMD가 진행한 세 번째 AI 관련 인수이며, 2025년 12월 엔비디아가 그록(Groq)의 자산을 약 200억 달러에 확보한 데 이어, 추론 전용 하드웨어를 둘러싼 경쟁이 본격화하는 양상이다.[14]

탈라스가 만드는 것은 통상의 AI 가속기와 범주가 다르다. 일반적인 GPU나 신경망 처리장치(NPU)는 어떤 모델이든 가중치를 메모리에 올려 실행하는 범용 장치인 반면, 탈라스의 칩은 특정 모델 하나만 실행하는 일종의 ASIC이다. 모델의 가중치와 데이터 흐름이 칩 제조 단계에서 회로로 고정되기 때문에 좀 더 구체적으로는 모델 특화 반도체(MSIC: model-specific integrated circuit)이다.[15] 본 글의 제목에 들어있는 "모델이 곧 컴퓨터다(The Model is The Computer)"는 탈라스 홈페이지에 걸린 슬로건이기도 하다.

가중치를 회로에 새긴다는 착상 자체는 ASIC을 설계하는 관점에서 완전히 새로운 개념이라고는 볼 수 없다. 사실 진짜 어려운 문제는 이를 수백억 파라미터 규모의 실제 모델에 적용해 제조할 수 있고,

13) AMD, “AMD Acquires Taalas to Advance Compute Solutions for Rapidly Growing AI Inference Market”, AMD Newsroom, Aug. 6, 2026
14) CNBC, “Nvidia buying AI chip startup Groq’s assets for about $20 billion in its largest deal on record”, Dec. 24, 2025
15) MSIC는 ASIC처럼 일반적으로 통용되는 용어는 아니며 또한 AMD나 탈라스가 사용하는 공식 명칭도 아니다. 다만, 본 기고문에서는 MSIC가 탈라스 기술의 핵심을 잘 드러내는 용어이기에 글 전반에 사용한다.

추론을 제대로 돌릴 수 있으며, 추후 모델이 바뀔 경우 “경제적으로” 새로 만들 수 있느냐이다. 이 글에서는 아주 작은 신경망을 예로 들어 MSIC의 동작 원리를 폰 노이만 구조와 대비하여 설명하고, 이어 실제 모델 규모로 확장할 때 부딪히는 과제와 이를 탈라스는 어떻게 해결하는지 살펴보고자 한다.

2. MSIC의 동작 원리

2.1 아주 작은 신경망으로 보는 두 가지 실행 방식

입력 두 개와 은닉 노드 두 개, 출력 하나로 이루어진 최소한의 신경망을 생각해 보자. 〈그림 2〉 은닉층 가중치를 W_1 = [[1, 2], [2, −1]], 출력층 가중치를 W_2 = [2, −1], 편향을 b = 1이라 하고, 활성화 함수로 ReLU를 쓴다. 입력 x = [2, 1]을 넣으면 은닉 노드는 H_1 = ReLU(1·2 + 2·1) = 4, H_2 = ReLU(2·2 − 1·1) = 3이 되고, 최종 출력은 y = 2·4 + (−1)·3 + 1 = 6이다. 가중치는 모두 여섯 개, 곱셈은 여섯 번으로 끝난다.

폰 노이만 구조를 따르는 GPU나 TPU는 이 계산을 여러 단계로 나누어 수행한다. 여섯 개의 가중치를 메모리에 데이터로 저장해 두고, 곱셈-누산기(MAC: multiply-accumulate)가 필요할 때마다 값을 읽어와 곱하고 더한다. 은닉층 계산을 마치면 중간 결과를 저장하고 ReLU 연산을 따로 실행한 뒤, 다시 출력층 가중치를 읽어 와 같은 MAC 회로를 재사용한다. 연산기는 어떤 값이든 처리할 수 있는 범용 회로이고, 가중치는 그 회로의 입력값으로 들어간다. 그래서 동일한 하드웨어에 가중치만 바꿔 입력하면 전혀 다른 신경망을 실행할 수 있다.

MSIC는 다른 방식으로 접근한다. 여섯 개의 가중치마다 그 값을 곱하는 회로를 하나씩 실제로 만들어 두고, 노드 사이의 연결도 물리적 배선으로 고정한다. 가중치 2를 곱하는 자리에는 2만 곱하는 회로가 놓인다. 가중치를 읽어 오는 동작이나, 다음에 실행할 명령어도 따로 지정할 필요가 없다. 입력 신호를 넣으면 회로를 타고 흐르면서 계산이 끝난다. 폰 노이만 구조에서 가중치를 오가게 하는 입출력과 명령어를 인출하고 해독하는 과정이, MSIC에서는 신호가 고정된 회로를 따라 흐르는 것으로 갈음된다.

가중치가 상수로 고정되면 범용 곱셈회로 자체가 필요 없어진다는 점도 연산 가속화에 큰 이점이 된다. 예를 들어 곱하기 1은 그냥 배선만, 곱하기 2는 배선과 시프트, 곱하기 −1은 부호 반전, 이런 성질을 활용하면 임의 상수도 시프트-덧셈으로 축약되고, ‘0’ 가중치는 회로가 아예 통째로 사라진다.

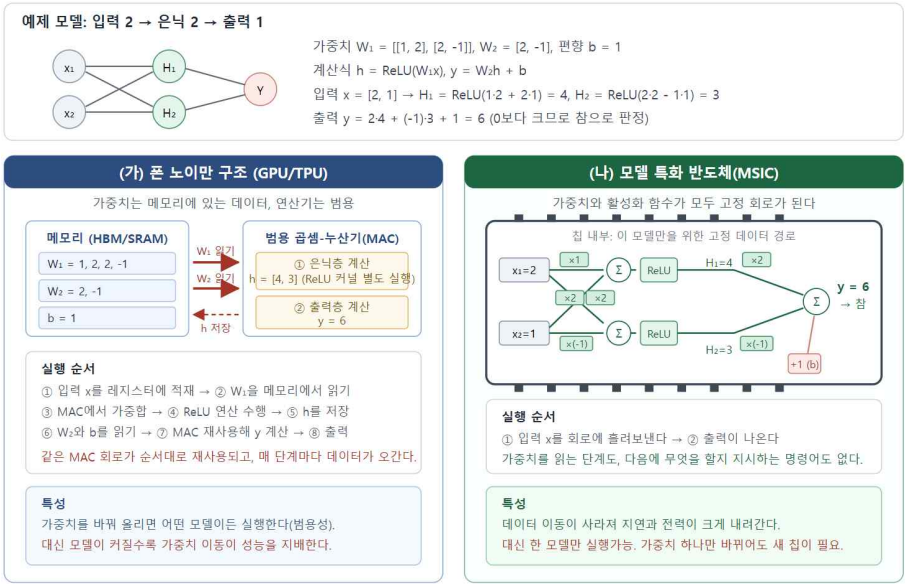

그림 2 신경망을 실행하는 방식 비교

2.2 가중치가 '실리콘에 내장'된다는 말의 의미

은닉 노드 H_1 하나를 확대해 보면 가중합 경로와 합산기, 그 뒤의 활성화 함수까지 모두 하드웨어 경로에 담겨 있다. ReLU는 max(0, s)이므로 입력의 부호를 판별하는 비교기(comparator)와 값을 고르는 선택기(multiplexer)만으로 구현된다.[16] GPU에서는 코드 한 줄로 실행되는 함수가 MSIC에서는 신호가 지나가는 고정된 회로 블록이 되는 것이다. 다만 GELU(가우시안 오차 선형 유닛)처럼 좀 더 복잡한 함수는 그대로 회로화하기 어려워 참조표(LUT: lookup table)나 구간별 선형 근사를 쓴다. 요컨대 가중치와 가중합, 활성화 함수, 데이터 흐름 전체가 하드웨어로 내려간다.

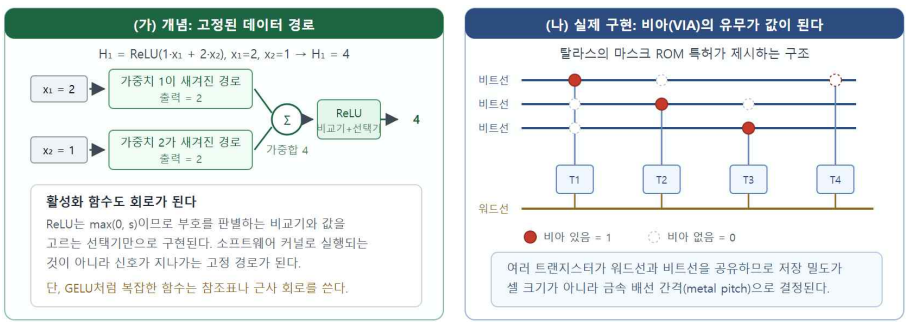

그림 3 은닉노드를 확대해 추상화한 개념(가) 및 실제 비아 패턴을 이용한 구현(나)

16) 실제로는 이보다 훨씬 간단하게 부호비트로 나머지 비트를 마스킹하는 AND 게이트로 구현된다.

〈그림 3〉의 (가)에서 "×2 회로"는 설명을 위한 추상화다. 실제로 가중치 하나마다 독립된 곱셈기를 배치하면 배선이 너무 복잡해 칩을 만들 수 없다. (나)처럼 배선을 공유하고 비아(via) 패턴만으로 값을 표현하면, 배선 혼잡을 피하면서 밀도를 끌어올릴 수 있다. 실제 가중치가 회로에 어떻게 구현되는지는 뒤에서 좀 더 다룬다.

2.3 이렇게 특화해서 얻는 것은 무엇인가?

이런 특화를 시도하는 이유는 추론의 병목이 연산이 아니라 데이터 이동에 있기 때문이다. LLM이 토큰을 하나씩 만들어 내는 디코드 단계는 토큰 하나를 생성할 때마다 모델 가중치 전체를 메모리에서 다시 훑는다. 뒤에서 볼 탈라스 HC1의 대상 모델인 라마 3.1 8B를 예로 들어 보자. 학습에 흔히 쓰는 32비트 정밀도의 절반인 16비트로 파라미터 하나를 담는 형식을 반정밀도(FP16)라 하는데, 이 형식이면 파라미터 하나가 2바이트이므로 80억 개는 약 16기가바이트가 된다. 토큰 하나를 만들 때마다 이 16기가바이트가 통째로 읽힌다.

문제는 그렇게 읽어 온 데이터로 하는 일이 많지 않다는 데 있다. 가중치 하나를 읽어 곱하고 더하는 것이 전부이므로 연산량은 바이트당 한두 번이다. 그런데 H100의 두 가지 성능을 비교해 보면 상황이 드러난다. 메모리에서 데이터를 끌어오는 속도, 곧 메모리 대역폭은 초당 3.35테라바이트다. 반면 끌어온 데이터를 처리하는 연산 성능은 초당 약 990조 회에 이른다. 앞의 수치로 1바이트를 가져오는 동안 뒤의 수치로는 300회 가까이 연산할 수 있다는 뜻이다. 즉 연산기를 놀리지 않으려면 바이트당 300회는 계산할 거리가 있어야 하는데, 실제로 필요한 것은 한두 번뿐이다. 연산 능력의 대부분이 데이터를 기다리며 비어 있는 셈이다.

여기에 하나의 단서를 달아 두어야 한다. 이 계산은 요청을 하나만 처리할 때를 가정한다. 여러 요청을 묶어 한꺼번에 처리하면 한 번 읽어 온 가중치를 모든 요청이 나누어 씀으로 읽기 비용이 분산되고, 배치가 수백 개에 이르면 GPU는 대역폭이 아니라 연산이 병목인 영역으로 넘어간다. 따라서 GPU가 취약한 것은 총 처리량이 아니라 사용자 한 명이 체감하는 지연이다. 요청을 넉넉히 묶어 원가를 맞추는 구간에서 GPU는 여전히 효율적이며, 묶을 요청이 없거나 응답 지연이 곧 서비스 품질인 경우 메모리 벽이 바로 사용성 문제로 드러난다.[17]

HBM으로 이 벽을 낮출 수는 있지만 그 한계는 크게 벗어날 수는 없다. MSIC는 가중치가 회로 안에 새겨져 있어 읽어 올 대상이 없으므로 대역폭이라는 항이 식에서 사라지고, 남는 것은 신호가 회로를

17) 혼합 전문가(MoE) 구조에서는 토큰마다 일부 전문가만 활성화되므로 읽어야 할 가중치가 전체보다 훨씬 적다. 예를 들어 딥시크 R1은 6,710억 파라미터 가운데 370억 개만 쓰인다.

통과하는 시간뿐이다. 탈라스가 같은 라마 3.1 8B를 담은 첫 칩에서 사용자 한 명 기준 초당 1만 6천 개가 넘는 토큰을 생성했다고 주장하는 근거가 여기에 있다.[18] 앞의 H100 예에서의 210개와 견주면 80배에 가까운 값이다. 물론 이는 사용자 한 명의 경우다.

3. 실제 모델을 담기 위한 과제와 탈라스의 해법

여섯 개의 가중치를 회로로 새기는 일은 어렵지 않다. 문제는 80억 개를 새겨야 할 때 생긴다. 앞 장의 원리를 실제 모델 규모로 확장하려면 서로 얽힌 여러 과제를 동시에 풀어야 한다.

3.1 밀도의 벽: 저장, 연산, 배선을 하나로 합치다

첫 번째 벽은 저장 밀도다. 라마 3.1 8B의 80억 파라미터를 4비트로 담아도 4기가바이트에 이른다.[19] 흔히 쓰는 온칩 메모리인 SRAM은 비트 하나를 담는 데 트랜지스터 여섯 개가 필요하므로, 4비트면 24개, 80억 파라미터면 트랜지스터 약 1,900억 개가 저장에 쓰인다. 위키피디아 자료에 의하면 최신 3나노 공정의 트랜지스터 밀도가 제곱밀리미터당 약 2억 개이다.[20] 따라서 1,900억 개는 950제곱밀리미터가 필요한데, 이는 단일 칩(다이) 제조 한계인 레티클 한계(reticle limit) 약 858제곱밀리미터를 훌쩍 넘는다.[21] SRAM은 디코더와 감지 증폭기 같은 주변 회로가 따라붙으므로 실제 면적은 이보다 커진다. 곱셈 회로는 제외했음에도 이 정도다.

탈라스는 마스크 ROM을 저장 수단으로 택했다. 저장할 내용이 제조용 포토마스크 단계에서 확정되는 읽기 전용 메모리다. 그중에서도 상위 배선과 하위 트랜지스터 셀 사이의 연결 통로인 비아 홀을 뚫거나 막는 방식으로 0과 1을 결정하는 쪽을 골랐다. 셀 구조가 단순해 같은 면적에 훨씬 많은 비트를 담을 수 있지만, 한 번 만들면 바꿀 수 없다.

두 번째 벽은 연산 회로의 통합이다. 일반 구조에서는 저장 소자와 읽기 회로, 곱셈기가 모두 따로 필요하다. 공동창업자 류비샤 바이치(Ljubisa Bajic)는 마스크 ROM 리콜 패브릭에서 4비트를 저장하고 그에 관련된 곱셈까지 트랜지스터 하나로 처리하는 방식을 찾았다고 밝혔다. 구체적 회로는 공개하지 않았다.[22] 물론, 트랜지스터 하나로 처리한다는 것은 표현상 오해의 소지가 많다. 양자화된 가중치가

18) EE Times, "Taalas Specializes to Extremes for Extraordinary Token Speed", Feb. 19, 2026
19) HC1은 실제 3비트 기본에 일부 6비트를 섞은 자체 방식을 쓰지만 평균하면 4비트 안팎이 된다. 탈라스에서도 이방식의 추론품질이 낮음을 인정하며, HC2에서는 표준 FP4를 채택한다고 한다.
20) https://en.wikipedia.org/wiki/3_nm_process
21) 최신 상용 칩의 경우 이 한계를 극복하기 위해 2개 이상의 칩을 이어 붙이는 '칩렛' 구조를 많이 쓴다.

가질 수 있는 값이 몇 개뿐이므로(예를 들어 3비트면 8개), 입력 활성값에 대해 적은 수(3비트면 8번)의 곱셈을 미리 해 놓고 실제 은닉노드에서의 연산은 곱셈 연산 없이 곱셈 결과를 “선택”한 후 더하는 방식으로 이루어진다는 분석이 설득력이 있다.[23)]

세 번째 벽은 배선이다. 가중치 하나마다 배선과 연산 회로를 문자 그대로 배치하면 배선 혼잡 때문에 칩을 만들 수 없다. 칩은 맨 아래에 트랜지스터를 깔고 그 위로 금속 배선을 여러 층 쌓아 올린 구조이며, 최신 공정에서 이 금속층은 십수 개에 이른다. 층 사이는 절연막으로 막혀 있어 다른 층의 배선을 이으려면 절연막을 뚫고 금속을 채워야 한다. 이 수직 통로가 비아(via)다. 아래층을 공통으로 깔아 두고 비아 층의 패턴만 바꾸면 이미 만든 공정 대부분을 그대로 둔 채 기능을 다시 정의할 수 있고, 비아 층은 패턴이 단순해 마스크 제작 비용도 상대적으로 낮다. 탈라스가 2025년 4월 출원한 마스크 ROM 특허는 이 성질을 저장에 적용한 사례다. 한 열의 트랜지스터들이 워드선과 비트선을 공유하도록 배치하고, 비트선과 트랜지스터 사이에 비아를 두는지 여부로 값을 표현한다. 비아가 있으면 1, 없으면 0이다.[24)]

요컨대 탈라스가 푼 것은 저장과 연산, 배선을 각각 최적화한 것이 아니라 셋을 하나의 구조로 합친 것이다. 〈그림 3〉의 (나)는 이 원리를 단순화한 것이다.

3.2 고정할 것과 변해야 할 것을 나누다

모든 것을 회로에 새길 수는 없다. 트랜스포머 추론에는 실행할 때마다 달라지는 데이터가 존재한다. 대표적인 것이 KV 캐시다. LLM은 이미 처리한 토큰의 키와 값 벡터를 저장해 두고 재사용하는데, 이 내용은 사용자의 대화마다 완전히 달라진다. 고객이 파운데이션 모델을 자기 데이터로 미세 조정하는 파인튜닝 역시 가중치의 변경을 전제한다. 따라서 "모델 전체를 ROM에 굽는다"는 표현은 엄밀하게는 맞지 않다.

탈라스는 칩을 두 영역으로 나누어 이 문제를 풀었다. 회사는 각각을 마스크 ROM 리콜 패브릭과 SRAM 리콜 패브릭이라 부른다. 전자에는 바뀌지 않는 기반 가중치와 고정된 데이터 흐름을 새기고, 후자에는 KV 캐시와 파인튜닝용 어댑터를 담는다. 어댑터란 기반 가중치는 그대로 두고 소량의 보정 가중치만 더해 모델의 거동을 바꾸는 파인튜닝 기법으로, 로라(LoRA)가 대표적이다. 탈라스의 또 다른 특허는 이 구분을 모델 코어와 파인튜닝 부분으로 명시하고, 앞쪽은 마스크 ROM에, 뒤쪽은 프로그래밍 가능한 메모리에 둔다고 기술한다.[25)]

22) The Next Platform, "Taalas Etches AI Models Onto Transistors To Rocket Boost Inference", Feb. 19, 2026
23) Zhao, Yongwei, “Taalas HC1 Architecture Decoded”, Mar 1, 2026
24) Taalas Inc., "Mask Programmable ROM Using Shared Connections", 특허번호 WO2025217724A1

이 분리가 이 구조의 핵심이다. 모델에서 거의 변하지 않는 대부분은 고정하고, 계속 변하는 작은 부분만 프로그래밍 가능하게 남긴다. 다만 이런 발상 자체가 최초는 아니고 이미 선례가 있다.[26)]

3.3 두 장의 마스크와 모델에서 실리콘까지의 자동화

탈라스의 접근 방식이 넘어야 할 가장 큰 허들은 경제성이다. 최신 공정의 마스크 세트 한 벌은 층이 100개를 넘고 비용도 수백만 달러에서 수천만 달러에 이른다. 통상의 전용 반도체 설계는 아키텍처 정의부터 회로 설계, 배치와 배선, 검증을 거쳐 제조에 들어가기까지 1년에서 2년이 걸리는데, 그 사이에 모델은 이미 수차례 바뀐다. LLM을 전용 칩으로 만들자는 발상이 그동안 현실적이지 않았던 이유가 이것이다.

탈라스의 해법은 2000년대에 등장했던 구조화 ASIC 개념을 되살린 것이다. 구조화 ASIC은 논리 회로의 기본 블록을 미리 만들어 두고 층 몇 개의 배선 패턴만 바꿔 용도를 정하는 방식으로, 이에이직(eASIC) 같은 회사가 상용화했다.[27)] 이에이직은 2018년 인텔이 인수해서 프로그래머블 솔루션 사업에 편입했다. 탈라스는 이 원리를 모델과 가중치, 데이터 흐름의 맞춤에 적용했다. 100여 개 층 가운데 두 개 층만 교체하면 다른 모델을 담을 수 있고, 이 두 층이 가중치와 데이터 흐름을 함께 결정한다. 나머지 층과 연산 회로, 클록과 전원, SRAM은 그대로 재사용하므로 NRE(non-recurring engineering)에 소요되는 비용과 기간이 크게 줄어든다. 〈그림 4〉가 이 흐름을 정리한 것이다.

그림 4 모델에서 칩까지의 흐름

25) Taalas Inc., "Computing Architecture with Model Core and Fine-Tuning Portion", 특허번호 US20250238726A1
26) Yiming Chen et al., "YOLoC: DeploY Large-Scale Neural Network by ROM-based Computing-in-Memory using ResiduaL Branch on a Chip", arXiv, Jun 1, 2022
27) EEJournal, "Redefining Structured ASIC", May 24, 2005

두 장의 패턴을 설계하는 데에도 여전히 여러 달이 걸린다. 그래서 모델 파일을 RTL(register transfer level) 코드로 바꾸는 자동화가 반드시 필요하다. 반도체 설계는 사양 정의, RTL 설계, 논리 합성, 배치와 배선, 검증, 테이프아웃의 순서로 진행되는데, RTL은 베릴로그(Verilog) 같은 하드웨어 기술 언어로 회로의 동작을 신호와 레지스터 사이의 데이터 이동으로 기술하는 단계이며, 소프트웨어의 소스 코드에 해당한다. 이후 합성 도구가 이를 논리 게이트의 연결로 바꾸고, 배치와 배선 단계가 그 결과를 실제 칩 위의 좌표와 금속층 배선으로 옮긴다. 탈라스는 모델에서 RTL로 내려가는 부분 자동화를 통해 이를 약 일주일치 엔지니어링 작업으로 줄였다고 한다. 설계 완료 후 파운드리에 제조 데이터를 넘기는 테이프아웃을 거쳐 동작하는 카드가 나오기까지 두 달을 목표로 한다고 밝혔다.[28] 소프트웨어 배포 주기에 가까운 속도로 전용 칩을 반복 생산하겠다는 구상이다.

검증도 함께 풀어야 한다. 일반 칩은 오류가 발견되어도 소프트웨어나 펌웨어로 우회할 여지가 있지만, MSIC는 모델의 동작 자체가 회로이므로 설계에 오류가 있으면 칩 전체를 쓸 수 없다. 그런데 통상의 사인오프 항목은 이 오류를 잡아 주지 못한다. 설계 규칙 검사(DRC: design rule check)는 레이아웃이 공정 규칙을 지키는지 볼 뿐이어서, 마스크 ROM에 새겨진 가중치가 맞는 값인지는 애초에 검사 대상이 아니다. 도형이 규칙에 맞기만 하면 비아가 어느 자리에 있든 통과한다. 결국 테이프아웃 전에 모델 전체를 시뮬레이션하는 수밖에 없고, 이를 위해 탈라스는 대규모 컴퓨팅 클러스터에서 모델 전체와 다중 칩 구성까지 시뮬레이션하는 자체 작업 흐름을 구축했다.

3.4 레티클 한계와 프론티어 모델

마지막 과제는 규모다. 앞서 본 노광 한 번의 상한 858제곱밀리미터에 HC1의 815제곱밀리미터는 이미 근접해 있고, 그 안에 들어간 것은 80억 파라미터 모델이다. 1천억 파라미터를 넘는 모델은 한 다이에 담기지 않는다.

탈라스의 해법은 두 갈래다. 하나는 다이를 기능별로 쪼개는 것으로, HC2는 SRAM 부분을 별도 칩으로 분리해 가중치에 쓸 면적을 늘리고 칩당 약 200억 파라미터를 목표로 한다. 다른 하나는 모델의 층을 여러 카드에 나누어 순서대로 통과시키는 파이프라인 병렬이다. 층 단위로 자르면 카드 사이를 오가는 것이 활성값뿐이라 토큰당 수 킬로바이트에 그치므로, PCIe 수준 대역폭이면 충분하다는 주장은 무리가 없어 보인다.

탈라스가 제시한 것은 PCIe를 통신 수단으로 하는 보드와 카드 수준의 분산이다. 통신량이 작은 구간에서는 합리적인 선택이지만, 다이를 더 잘게 쪼개 밀도를 끌어올릴수록 결국 패키지 안의 배선 문제로

28)Taalas, "The path to ubiquitous AI" (https://taalas.com/the-path-to-ubiquitous-ai/)

돌아온다. AMD가 인스팅트 계열에서 축적한 첨단 패키징 역량이 더해질 수 있는 지점이 여기다.

하지만 대규모 모델의 경우 사실상 한 모델을 위해 여러 각기 다른 칩을 생산해야 한다는 문제가 있다. 예를 들어, 딥시크 R1 671B 급에는 30장 이상이 필요한데, 이는 모델의 서로 다른 구간을 담은 30개 칩 설계이며 각각 별도의 테이프아웃을 요구한다.[29] 1조 파라미터급은 HC2 기준 약 50장으로 추정된다. 여기에 최근 프론티어 모델이 향하는 MoE 구조는 이 방식에 맞지 않는다. GPU는 쓰이지 않는 전문가를 값싼 외부 메모리에 두면 되지만, MSIC에서는 쓰이지 않는 파라미터도 실리콘에 그대로 새겨져 있어야 하기 때문이다.

대규모 프론티어 모델은 아직 실물로 검증되지 않았다. 다중 칩 구성은 시뮬레이션을 통한 추정으로 보이며, 현재 가장 크게 남은 기술적 과제이자 상용화 리스크다. 815제곱밀리미터 대형 다이는 결함 하나에도 취약한데 앞서 보았듯 가중치는 여분으로 대체할 수 없어, 양산 불량률과 검사 비용이 수익성을 압박할 수 있다. 두 달 주기 역시 시제품 수준의 목표로 보는 것이 타당하다.

3.5 요약: 해결과제와 탈라스의 해법

과제	통상의 전용 반도체	탈라스의 해법
수십억 가중치 저장	SRAM으로는 레티클 한계를 넘어섬	마스크 ROM 리콜 패브릭, 비아 패턴으로 값 표현
가중치와 활성값의 곱	가중치마다 곱셈기와 배선이 필요	양자화를 이용해 곱셈을 선택으로 치환, 트랜지스터 한 개로 처리
KV 캐시 등 가변 데이터	ROM으로 불가능	별도의 SRAM 리콜 패브릭
파인튜닝	새 칩 제작	모델 코어와 분리된 프로그래밍 가능 영역
모델 교체	전체 테이프아웃	전체 마스크 층 100여 개 중 2개 층만 교체
개발 기간	1~2년	모델에서 RTL까지 자동화. 생산까지 약 2개월
대형 모델	레티클 한계	다중 카드 파이프라인 병렬, 단 모델 구간마다 별도 테이프아웃

표 1 MSIC 구현을 위한 주요과제와 탈라스의 해법

4. AMD 인수의 의미

AMD가 인수한 것은 라마 8B 전용 칩 한 개가 아니다. 실제 자산은 세 가지의 결합이다. 첫째는 저장과 연산을 집적한 고밀도 패브릭 기술이다. 둘째는 임의의 모델로부터 자동으로 회로 설계와 마스크

29) Taalas, "The path to ubiquitous AI" (https://taalas.com/the-path-to-ubiquitous-ai/)

패턴으로 변환하고 검증까지 마치는 설계 흐름이다. 셋째는 고정된 파운데이션 모델과 가변적인 KV 캐시, 어댑터를 결합하는 혼합 구조다. 이 셋이 함께 있어야 모델 특화 실리콘이 일회성 프로젝트가 아니라 반복 가능한 제품 플랫폼이 된다. 장기적으로는 두 번째 자산이 더 중요할 수 있다. 고객의 모델을 받아 맞춤 추론 실리콘으로 만들어 주는 사업이 성립한다면, AMD는 GPU 공급자를 넘어 모델 단위의 커스텀 칩 개발사로 영역을 넓힐 수 있다.

AMD는 MSIC를 GPU의 대체재로 보지 않는다. 인수 발표에서 AMD AI 그룹 총괄 밤시 보파나(Vamsi Boppana)는 다양한 목적의 AI 워크로드에 최적화된 솔루션을 고객에게 제공할 수 있는 유연하게 활용할 수 있는 풀스택 플랫폼을 만들고 있다고 말했다. AMD는 탈라스의 기술을 자사 가속기 로드맵에 통합하고 인스팅트 GPU와 결합한 시스템 수준 해법을 개발하겠다고 밝혔다. 즉 대체가 아니라 역할 분담이라는 뜻이다.

이 결합은 실제 추론과정이 성질이 다른 두 단계로 나뉜다는 사실과 부합한다. 프리필(prefill)은 사용자가 넣은 프롬프트 전체를 한꺼번에 처리해 문맥을 만드는 단계이고 연산량이 병목이다. 디코드는 토큰을 하나씩 순차적으로 생성하는 단계이고 메모리 대역폭이 병목이다. 하나의 하드웨어로 두 단계를 모두 처리하면 어느 쪽에서든 자원이 남거나 모자란다. 그래서 단계를 물리적으로 분리해 각각에 맞는 하드웨어를 배정하는 분리 추론이 자연스러운 해법이 된다.

AMD는 이미 이 구조를 실행에 옮긴 바 있다. 2026년 7월 세레브라스와 발표한 협력이 그것이다. 프롬프트를 받아 문맥을 만드는 프리필은 에픽(EPYC) CPU와 인스팅트 MI400 계열 GPU로 구성된 헬리오스(Helios) 랙이 맡고, 토큰을 하나씩 생성하는 디코드는 세레브라스의 웨이퍼 스케일 엔진이 맡는 구성이다. 연산량이 많은 단계는 범용 GPU에, 지연이 중요한 단계는 특화 실리콘에 배정한 셈이다. 두 시스템을 이렇게 나누어 쓰면 와트당 초당 토큰 수를 최대 5배까지 끌어올릴 수 있다는 것이 양사의 설명이다.[30)]

같은 자리에 탈라스의 MSIC를 넣는 것은 아키텍처 관점에서 자연스러운 접근이다. 인스팅트 GPU가 프롬프트 처리를 맡고 탈라스 실리콘이 토큰 생성을 맡는 분할 구성으로 헬리오스 랙에 배치하며, 전체를 ROCm 소프트웨어 스택으로 다룬다는 것이 AMD의 계획이다.[31)] 〈그림 5〉는 이 구조를 도식화한 것이다.

30) Tom's Hardware, "AMD and Cerebras partner on low-latency, high-throughput AI inference", Jul. 2026.
31) Forbes, "AMD Buys Taalas, The Startup That Carves AI Models Into Silicon", Aug. 9, 2026

그림 5 추론 단계별 하드웨어 분리

5. 향후 과제와 시사점

MSIC의 성립 조건은 하드웨어 수명과 모델 수명의 관계에 달려 있다. 칩을 설계하고 양산하는 데 걸리는 시간이 그 칩에 새겨진 모델이 상용으로 쓰이는 기간보다 짧아야 이 방식이 경제적으로 성립한다. 새 모델이 한 달 단위로 나오는 지금의 속도는 이 조건에 맞지 않는다. 반대로 특정 모델이 오래 안정적으로 대량 서비스되는 구간, 예를 들어 검증을 마친 특정 버전을 장기간 고정해 쓰는 제조산업 현장에서는 조건이 맞아떨어진다. 특히 여러 요청을 묶어 배치로 처리할 필요가 없는 온프레미스 추론이나 엣지 실행에서 모델 특화 실리콘의 의미가 크다.

다만 검증되지 않은 항목이 많다. 공개된 성능 수치는 대체로 회사가 제시한 시제품 기준이며 제3자 검증을 거치지 않았다. 3비트 수준의 양자화가 복잡한 과제를 해결할 때 품질을 얼마나 떨어뜨리는지, 대형 다이의 양산 수율이 수익성을 담보할 수 있는지, 두 달 테이프아웃이 양산 규모에서 안정적으로 유지될 수 있는지, 다중 칩으로 프론티어 모델을 실제로 돌릴 수 있는지 등 아직 남은 질문이 많다. AMD가 인수한 것은 완성된 제품이라기보다 팀과 방법론, POC에 가깝고, 이 접근이 GPU 이후 추론 경제학의 한 축이 될 가능성에 선제적으로 베팅한 성격이 강하다.

AI 반도체 경쟁의 축이 범용 연산 성능에서 워크로드 특화 설계 역량으로 이동하고 있다. 탈라스의 진입 장벽은 트랜지스터 수준의 특화 기술 하나가 아니라 소자 기술과 설계 자동화, 검증, 파운드리 공정을 하나의 닫힌 고리로 엮은 데 있다. 이런 역량은 연산 성능 경쟁처럼 자본 규모가 결정하는 영역이 아니어서 국내 팹리스가 겨냥할 여지도 충분히 있다.

지난 수십 년간 컴퓨팅의 기본 전제는 하드웨어가 범용이고 소프트웨어가 그것을 특화한다는 것이었다. MSIC는 어떻게 보면 이 진화의 과정을 역행하는 것이다. 모델이 회로가 되고, 사람의 언어가 그 위의 소프트웨어가 된다. 이런 방식이 진짜 유효한 결과를 만들어 낼지는 결국 모델의 진화 속도가 실리콘의 제작 속도와 어떻게 균형을 유지하는가에 달려 있다.

04 2026년 클라우드 솔루션 리포트: AI 에이전트 플랫폼

| 정채상 이음테크

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

들어가며: 에이전트를 만드는 시대는 끝났고, 이제 통제하는 시대다

2026년 상반기, AI 에이전트 플랫폼 시장의 무게중심이 어느새 이동했다. 지난 2년 동안 업계의 질문은 "어떻게 하면 에이전트를 만들 수 있는가"였다. 로우코드·노코드 빌더가 성숙하고, 관리형 런타임이 등장하고, 오픈소스 프레임워크가 쏟아지면서 에이전트를 만드는 일의 진입 장벽은 급격히 낮아졌다. 그리고 진입 장벽이 낮아진 바로 그 지점에서, 전혀 다른 성격의 질문이 시장의 중심으로 올라왔는데, "이렇게 만든 에이전트를 어떻게 믿고, 운영하고, 통제할 것인가." 코드를 생성하는 에이전트보다 그 에이전트를 관리하는 거버넌스 레이어가 더 뜨거운 화두가 된 것이다.

이 전환은 가트너가 내놓은 두 개의 신호에서 선명하게 드러난다. 가트너는 2026년 4월 이 시장에 처음으로 독립적인 하이프 사이클을 발행하며 '에이전트 개발 플랫폼'을 기대의 정점(Peak of Inflated Expectations)에 올려놓았다. 그에 앞서 에이전트 프로젝트의 40% 이상이 2027년 말까지 취소될 것이라고 경고한 터였는데[32], 원인으로 지목된 것은 모델의 성능이 아니라 급증하는 비용, 불분명한 비즈니스 가치, 그리고 불충분한 리스크 통제였다. 한 시장이 기대의 정점과 대량 실패 경고를 동시에 받는 것, 그것이 2026년 에이전트 플랫폼의 모습이다.

'빌드는 쉬워졌다'는 이야기는 곧 통제가 오히려 어려워졌다는 역설로 이어진다. 실제로 프로덕션 환경에서 에이전트를 운영하는 상당수 기업이 관련 보안 사고를 겪었고, 적지 않은 기업이 에이전트에 목적 제한이라는 가장 기초적인 통제조차 강제하지 못하고 있다는 조사가 잇따른다. 에이전트는 이미 스스로 도구를 호출하고 데이터에 접근하며 작업을 수행하는데, 그 행동을 관측하고 제약하는 체계는 그 속도를 따라가지 못하고 있는 것이다. 만들기의 편의와 통제의 부재 사이에 벌어진 이 간극이 이번 리포트의 출발점이다.

32) Gartner. Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027. Gartner Newsroom, 2025.06.25.

지난 두 달의 보고서를 돌아보면, 6월호에서 "AI 에이전트가 API의 새로운 소비자가 되었다"는 점을, 7월호에서 "그 에이전트에게 고유한 신원(ID)을 부여하고 생애주기를 관리해야 한다"는 점을 다뤘다면, 이번 달 보고서는 그 위 계층을 다룬다. 에이전트를 실제로 만들고, 실행하고, 평가하고, 통제하는 플랫폼 혹은 그것을 운영하는 방법을 정리한다.

본 리포트는 벤더를 기능 목록으로 나열하는 대신, CIO와 CTO가 실제로 마주하는 하나의 질문에서 출발한다. "우리는 어떤 경로로 에이전트를 도입할 것이며, 그 과정에서 거버넌스의 통제권을 누가 쥐게 되는가." 이 관점에서 글로벌 6개 벤더—마이크로소프트, AWS, 구글, 랭체인(LangChain), 크루AI(CrewAI), 세일즈포스—를 심층 분석하고, 마지막으로 이들과는 전혀 다른 노선을 걷는 한국 플레이어들의 전략을 대비한다. 다만 이 영역은 이제 막 문을 열었고, 어느 한 벤더도 홀로 완결된 답을 주지 못하고 있는데, 제품과 파트너, 모델과 구축이 매달 새롭게 조합되는 유동적 국면이며, 그래서 이 글이 끝에서 도달하는 결론은 '무엇을 살 것인가'가 아니라 '무엇을 고정하고 무엇을 유동적으로 둘 것인가'를 향한다.

2026년 에이전트 플랫폼의 핵심 패러다임

개별 벤더를 논하기에 앞서, 2026년 에이전트 플랫폼 시장을 지배하는 세 가지 흐름을 짚을 필요가 있다. 2026년 상반기에만 주요 기능의 정식 출시(GA) 전환이 수십 건 쏟아지는 등 어떤 기능을 제공하는가의 비교표는 반년이면 낡는다. 그러나 도입의 구조, 생명주기의 단계, 그리고 통신의 표준이라는 세 축은 그보다 훨씬 느리게 움직이며 선택을 좌우한다.

도입의 네 갈래 길: '거버넌스를 누가 쥐나'

기업이 에이전트 플랫폼을 도입하는 길은 거버넌스의 통제권이 어디에 놓이는가에 따라 크게 넷으로 나뉜다.

첫 번째는 이미 쓰고 있는 클라우드와 생산성 스택 위에 에이전트를 얹는 길이다. 마이크로소프트 코파일럿 스튜디오, AWS 베드록 에이전트코어(AgentCore), 구글 제미나이 엔터프라이즈가 여기에 속한다. 이 경로에서 거버넌스는 해당 벤더의 신원·정책·컴플라이언스 인프라에 자연스럽게 위임된다.

두 번째는 특정 클라우드나 모델에 종속되지 않는 중립적 프레임워크로 직접 만드는 길이다. 랭체인, 랭그래프와 크루AI가 대표적이며, 이 경우 통제권은 조직이 온전히 쥐지만 그만큼 직접 구현하고 운영해야 하는 부담이 커진다.

세 번째는 업무용 SaaS 안에 이미 통합된 완제품 에이전트를 구독하는 길이다. 세일즈포스 에이전트포스처럼, 거버넌스가 데이터와 업무 프로세스와 함께 통째로 제공된다.

네 번째는 이 셋을 조합하는 하이브리드다. 예컨대 관리형 런타임(AWS 에이전트코어) 위에 중립 프레임워크(랭그래프)를 올리고, SaaS 에이전트(에이전트포스)를 표준 프로토콜로 연동하는 식이다. 2026년 현실의 다수는 결국 이 네 번째로 수렴한다.

이 프레임이 유효한 이유는 분명한데, 기능 비교표는 반년이면 낡지만, "거버넌스의 통제권을 누구에게 위임할 것인가"는 조직의 클라우드 전략과 규제 환경, 그리고 내부 역량에서 비롯되는 구조적 선택이라 쉽게 바뀌지 않기 때문이다. CIO와 CTO에게 실질적으로 유효한 잣대는 '무엇을 할 수 있는가'가 아니라 '통제의 무게중심을 어디에 둘 것인가'다.

경로	도입 방식	통제권	적합 조직	대표 벤더
① 스택에 얹기	기존 클라우드·생산성 스택 위에 구축	벤더 인프라에 위임	특정 클라우드·M365 정착 조직	MS 코파일럿 스튜디오 AWS 에이전트코어 구글 제미나이
② 중립 프레임워크	비종속 오픈 프레임워크로 직접 구축	조직이 직접 보유 (운영 부담↑)	멀티클라우드·자체 통제·온프레미스	랭체인 랭그래프 크루AI
③ 완제품 구독	SaaS에 통합된 완제품 구독	SaaS 벤더가 통째로 제공	CRM 등 특정 SaaS 중심 조직	세일즈포스 에이전트포스
④ 하이브리드	①~③ 조합 (런타임+프레임워크+SaaS)	층별 분산, A2A·MCP 연결	현실의 다수가 수렴	에이전트코어 + 랭그래프 + 에이전트포스

표 2 에이전트 도입 4경로와 통제권의 위치

에이전트의 생명주기: 빌드에서 거버넌스까지

어느 경로를 택하든, 에이전트는 1) 만들고(빌드), 2) 실행하고(런타임), 3) 평가하고, 4) 통제하는(거버넌스) 동일한 네 단계를 거친다. 벤더를 비교하는 실질적 잣대는 이 네 단계를 각각 얼마나 채우는가에 있다.

빌드 단계는 2026년 기준 가장 성숙했으며, 사실상 상향 평준화됐다. 로우코드 빌더, 프레임워크, 그리고 도구를 연결하는 모델 콘텍스트 프로토콜(MCP)이 두루 갖춰지면서 '만들 수 있는가'는 더 이상 변별력이 없다. 런타임 단계는 급격히 성숙하는 중이다. 수 시간에서 길게는 7일에 이르는 장시간 실행,

세션과 메모리의 관리, 그리고 마이크로VM 수준의 격리가 AWS 에이전트코어 하네스와 구글 에이전트 런타임의 잇단 정식 출시를 통해 관리형으로 빠르게 채워지고 있다. 문제는 나머지 두 단계다.

한편 평가 단계에서 시장은 "벤치마크 통과가 곧 프로덕션 안전은 아니다"라는 뼈아픈 교훈을 마주하고 있다. 벤치마크가 역량은 측정하지만 거버넌스는 측정하지 못한다는 인식이 퍼지면서, '평가 엔지니어링'이 별도의 직군으로 부상했다. 프로덕션에서 발생하는 신뢰성 문제의 상당수가 벤치마크에서는 전혀 예측되지 않았던 것들이다. 마지막 거버넌스 단계가 2026년의 진짜 격전지인데, 에이전트의 인벤토리를 파악하고, 신원을 부여하고, 감사 로그를 남기고, 이상 행동을 탐지하고, 정책으로 행동을 사전에 차단하는 이 영역은 가장 덜 성숙했으며 벤더 간 격차가 가장 크다.

따라서 플랫폼을 고를 때 던져야 할 질문은 "이것으로 에이전트를 만들 수 있는가"가 아니라 "만든 뒤에 평가하고 통제할 수 있는가"여야 한다.

벤더	빌드	런타임	평가	거버넌스
코파일럿 스튜디오	●	◐	◐	●
AWS 에이전트코어	●	●	◐	◐
제미나이 엔터프라이즈	●	●	◐	◐
랭체인·랭그래프	●	●	◐	◐
크루AI	●	◐	○	○
에이전트포스	●	●	◐	●

● 강점 ◐ 갖춤 ○ 약함

표 3 에이전트 생명주기 4단계와 6개 벤더의 성숙도

프로토콜 표준화: MCP와 A2A라는 공용어

에이전트들이 서로 대화하고 외부 도구를 다루기 시작하면서, 그 대화의 문법을 정하는 표준 경쟁도 2026년 상반기에 사실상 결판이 나서 두 개의 프로토콜이 공용어로 자리 잡았다.

하나는 에이전트 간 통신을 규정하는 A2A(Agent-to-Agent)다. 구글이 제안하고 AWS와 마이크로소프트가 지원한 이 프로토콜은 2026년 기준 150개 이상 조직에서 프로덕션에 적용됐으며, 5개 언어의 SDK를 갖췄으며[33], 코파일럿 스튜디오, 애저 AI 파운드리, 베드록 에이전트코어가 모두 A2A를 정식 기능으로 내장했다. 다른 하나는 도구 연결을 규정하는 MCP로, 앤트로픽이 공개한 이 표준은 도구와 데이터를 연결하는 사실상의 공용어가 됐다. 주요 플랫폼이 모두 MCP를 기본으로 채택하면서, 이 표준 경쟁은 사실상 종결 국면에 들어섰다.

다만 표준화는 양날의 검이다. 에이전트 사이의 상호운용이 쉬워진 만큼, "누가 누구를 호출했는가"를 추적하고 통제하는 일은 오히려 더 중요해졌다. 2026년 6월의 MCP 명세는 서버 자체가 하나의 에이전트로 작동하는 재귀적 구성을 추가했는데, 연결이 깊어질수록 감사와 정책 강제의 난도도 함께 올라간다. 따라서 CIO와 CTO에게 필요한 처방은, 새로운 에이전트 시스템을 설계할 때 A2A와 MCP 호환성을 아키텍처 요구사항으로 명시하되, 이미 운영 중인 시스템을 당장 전환할 필요는 없다는 것이다. 이 표준들이 자리 잡았다는 사실은 단순한 호환성 이상의 의미를 갖는다. 어떤 모델을, 어떤 파트너를, 어떤 제품을 얹을지를 나중에 갈아 끼울 수 있는 기술적 토대가 마련됐다는 뜻이기 때문이다. 통제권은 고정하되 그 위의 조합은 유동적으로 두는 설계는 바로 여기서 비로소 가능해진다.

주요 벤더별 솔루션 심층 분석

이제 앞서 제시한 4가지 경로의 순서를 따라 글로벌 6개 벤더를 살펴본다. 스택 위에 얹는 클라우드 통합형(마이크로소프트·AWS·구글), 직접 만드는 중립 프레임워크(랭체인·크루AI), 통째로 구독하는 완제품(세일즈포스)의 순이다. 각 벤더가 빌드·런타임·평가·거버넌스의 네 단계를 어떻게 채우는지, 그리고 통제권을 어디에 두는지에 초점을 맞춘다.

마이크로소프트 코파일럿 스튜디오와 에이전트 365: M365 위에 거버넌스를 얹다

마이크로소프트의 로우코드 에이전트 빌더인 코파일럿 스튜디오는 2026년 상반기, '로우코드 챗봇 빌더'에서 '거버넌스된 엔터프라이즈 에이전트 플랫폼'으로의 전환을 사실상 완성했다. 강점은 M365 생태계와의 깊은 통합에 있는데, 엔트라 ID의 권한, 퍼뷰(Purview)의 민감도 레이블, 조건부 접근 정책이 에이전트에 자동으로 적용되므로, 이미 M365를 쓰는 조직은 거버넌스 인프라를 새로 구축하지

33) Alex Merced. The State of Agentic AI Standards in 2026: MCP, A2A, WebMCP, OSI, and the Protocol Stack. dev.to, 2026.

않고도 통제 체계를 사용할 수 있다. 포춘 500의 90%를 포함해 23만 개가 넘는 조직이 코파일럿 스튜디오로 에이전트를 만들어 왔다는 사실은 그 자체로 하나의 거대한 생태계다.

이 전환의 핵심은 2026년 5월 정식 출시된 에이전트 365로, 조직 전체의 에이전트를 등록하고 통제하는 거버넌스 컨트롤 플레인이다. 커스텀 에이전트를 여러 개 운영하는 조직을 정면으로 겨냥하며, 디펜더·퍼뷰와의 심층 보안 연동은 프리뷰 단계로 확장하고 있다.[34] 여기에 6월 말에는 엔진 수준의 전면 재설계가 더해져, 구조적 단계와 AI 추론을 한 캔버스에서 설계하는 워크플로 디자이너, MCP 서버 연결, 그리고 A2A의 정식 지원이 자리 잡았다. 마이크로소프트는 새 오케스트레이터로 평가 성능을 약 20% 높이고 토큰 소비를 절반으로 줄였다고 밝혔다.[35]

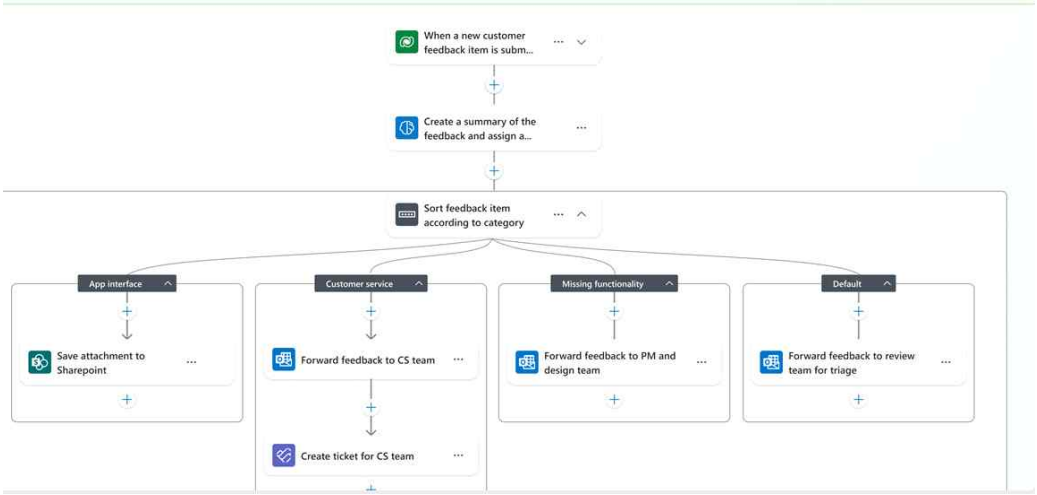

그림 6 구조적 단계와 AI 추론을 단일 캔버스에서 설계하는 마이크로소프트 코파일럿 스튜디오 워크플로 디자이너 (출처: 마이크로소프트)

한계도 분명하다. 우선 라이선스 장벽이 높아졌다. 에이전트 365의 신규 구매에 마이크로소프트 365의 상위 라이선스 등급인 E5가 필수로 요구되면서(2026년 6월부터, 그 아래 등급인 E3만으로는 자격 미달), 중견기업의 진입 문턱이 더 올라갔다. 또한 강점인 M365 생태계가 곧 한계가 되기도 하는데, 세일즈포스나 SAP 같은 외부 시스템을 연동하려면 뮬소프트(MuleSoft) 같은 미들레이어가 필요하고, 복잡한 에이전트 워크플로는 여전히 개발자의 개입을 요구한다.

34) Context Studios. Microsoft 365 AI Agents: The Complete Guide to Building and Running Agents with Copilot, Copilot Studio and Agent 365 in 2026. contextstudios.ai, 2026.
35) Microsoft. Meet the New Copilot Studio: Rebuilt for More Complex, Multi-Step Work. Microsoft Community Hub, 2026.

AWS 베드록 에이전트코어: 프레임워크에 종속되지 않는 실행 인프라

AWS의 생성형 AI 플랫폼인 베드록에서 에이전트 실행을 전담하는 서버리스 인프라인 에이전트코어는, 특정 프레임워크의 편을 들지 않고 모두의 실행 인프라가 되겠다는 전략을 택했다. 스트랜즈(Strands), 랭그래프, 오픈AI 에이전트 SDK, 구글 ADK, 클로드 에이전트 SDK가 모두 에이전트코어 위에서 돌아가며, IAM·VPC·클라우드트레일(CloudTrail) 같은 AWS의 보안·컴플라이언스 인프라를 그대로 이용한다. 이 프레임워크 중립성 때문에 에이전트코어는 앞서 말한 네 번째 경로, 곧 하이브리드의 토대가 되기에 적합하다.

특히 2026년에 진전이 두드러지는데, 지난 6월 정식 출시된 관리형 하네스는 모델·도구·스킬·지시라는 선언형 구성만으로 프로덕션 에이전트를 구동하며[36], 자연어 정책으로 에이전트의 행동을 사전에 가로채는 정책 기능은 "환급액이 일정 금액을 넘으면 승인을 막아라"는 식의 규칙을 에이전트의 추론 루프 밖에서 강제한다.[37] 7월에는 추적과 로그를 단일 로그 그룹으로 묶는 통합 관측성이 정식 출시되어, 평가와 거버넌스 단계까지 관리형으로 다루기 시작했다.[38]

다만 국내 도입 관점에서 볼 때 2026년 7월 기준 에이전트코어는 방콕·말레이시아·밀라노·스페인으로 리전을 넓혔으나 한국 리전은 아직 포함되지 않아, 지연시간과 데이터 레지던시를 고려해야 한다. 에이전트코어 브라우저와 결제 기능은 여전히 프리뷰 단계이며, 자연어 정책이 어떻게 정확하게 해석될지에 대한 예측 불가능성도 남아 있다.

구글 제미나이 엔터프라이즈: 에이전트를 최상위 개념으로

구글은 2026년 4월 클라우드 넥스트에서 버텍스 AI라는 브랜드를 폐기하고 제미나이 엔터프라이즈 에이전트 플랫폼으로 통합·재편했다.[39] 에이전트를 아키텍처의 최상위 개념으로 끌어올린 이 재구성은 빌드·스케일·거버넌스·최적화의 4계층으로 생명주기 전 단계를 커버한다. 빌드 계층에서는 파이썬과 고(Go)까지 정식 출시된 ADK 2.0이 그래프 기반 엔진과 인간 개입을 내장하고[40], 스케일 계층에서는 최대 7일간 실행되는 에이전트 런타임이, 거버넌스 계층에서는 CNCF 스파이프(SPIFFE) 기반의 에이전트 신원이 각각 정식 출시됐다. 인프라 측면에서는 TPU 8세대와 200개가 넘는 파운데이션 모델(클로드 포함)을 직접 연결할 수 있다는 점을 내세운다.

36) AWS. AWS Summit New York 2026: AI Agents. aboutamazon.com, 2026.06.
37) AWS. Amazon Bedrock AgentCore Adds Quality Evaluations and Policy Controls for Deploying Trusted AI Agents.AWS News Blog, 2026.03.
38) AWS. Amazon Bedrock AgentCore Now Delivers Unified Observability in a Single Log Group. AWS What's New, 2026.07.20.
39) TWIML. Google Cloud Next '26 Recap. twimlai.com, 2026
40) Google. Agent Development Kit (ADK) 2.0. adk.dev, 2026.

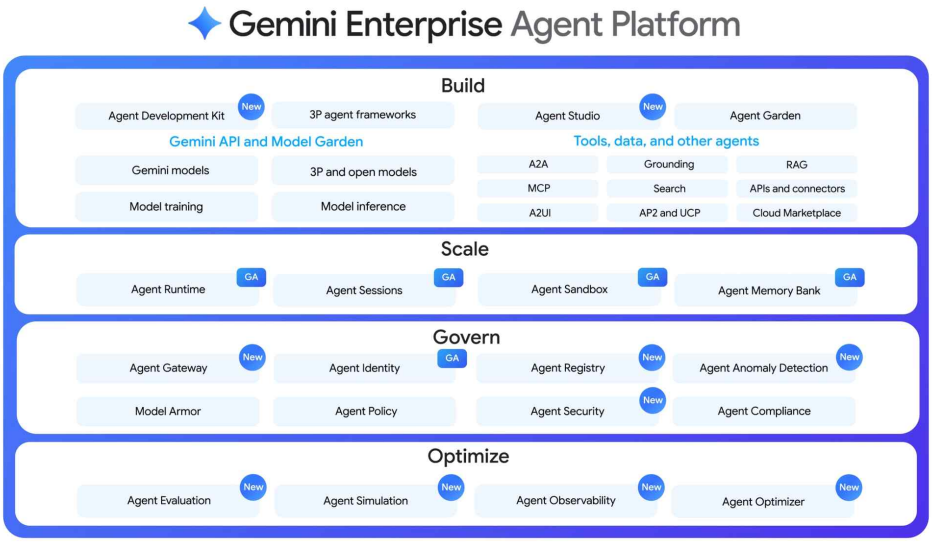

그림 7 구글 제미나이 엔터프라이즈 에이전트 플랫폼 (출처: 구글 클라우드)

반면, 리브랜딩으로 인해 기존 버텍스 AI 문서와 API에 혼선이 생겼고, 코드베이스를 업데이트해야 하는 부담이 남았다. 거버넌스 계층의 핵심인 에이전트 게이트웨이와 레지스트리가 아직 프리뷰라는 점, 그리고 엔터프라이즈 거버넌스의 복잡성 자체가 중소기업에게는 진입 장벽이 된다는 점도 지적된다.

랭체인·랭그래프와 랭스미스: 클라우드에 종속되지 않는 사실상 표준

오픈소스에서 출발한 랭체인은 2026년 현재 엔터프라이즈 에이전트 개발 스택의 사실상 기본값으로 자리 잡았다. 특정 클라우드나 모델에 종속되지 않는 유연성이 최대 강점이며, 거버넌스의 통제권을 벤더에 넘기지 않고 조직이 직접 쥐고자 할 때의 선택지가 된다. 셋은 같은 회사의 한 스택으로, 범용 프레임워크인 랭체인 위에서 상태 기반으로 복잡한 제어 흐름을 다루는 랭그래프가 오케스트레이션을, 추적·평가·비용 관리를 통합한 랭스미스가 운영을 담당하는 구조다.

2025년 10월의 랭그래프 1.0은 "오픈소스는 언제 깨질지 모른다"는 엔터프라이즈의 오랜 불신에 대한 응답이었다. 상태 자동 복원, 인간 개입 내장, 그리고 2.0이 나오기 전까지는 기존 코드를 깨뜨리지 않겠다는 호환성 보장이 그 내용이다.[41] 원클릭으로 에이전트를 운영하는 랭스미스 플릿(Fleet)은 속성 기반 접근 제어(ABAC)와 변조 방지 감사 로그를 갖췄고, 2026년 3월 발표된 엔비디아 파트너십으로 온프레미스 배포 경로까지 확보했다.[42]

41) LangChain. LangChain and LangGraph 1.0. LangChain Blog, 2025.10.22.

약점은 유연성의 이면에 있는데, 그래프 추상화의 학습 곡선이 초보자에게는 진입 장벽이 되고, 오픈소스와 유료 엔터프라이즈 기능의 경계가 불명확해 어디서부터 비용이 발생하는지 혼란스럽다는 지적이 있다. 갈릴레오(Galileo)나 애리즈(Arize) 같은 전문 관측성 도구와 견줄 때 랭스미스의 심층 평가 기능이 충분한가에 대한 논란도 한창 진행 중이다.

크루AI: 역할 기반 멀티에이전트의 직관적 오케스트레이션

크루AI는 역할과 목표, 배경을 가진 에이전트를 하나의 '크루' 멤버로 정의해 협업을 직관적으로 모델링하는 오픈소스 프레임워크다. 파이썬 기반의 낮은 진입 장벽 덕에 저변이 넓어, 직전 12개월 동안 누적 약 20억 건의 에이전트 실행을 보고했다.[43] 서버리스 실행과 중앙 대시보드를 갖춘 크루AI 엔터프라이즈로 SaaS 전환을 진행 중이다.

우려는 엔터프라이즈 성숙도에 있다. RBAC·감사 로그·SSO 같은 거버넌스 기능은 랭스미스에 비해 후발이고, "포춘 500 기업의 60%가 사용한다"는 수치도 정식 계약인지 개인 사용인지 불명확해 인용에는 주의가 필요하다.

세일즈포스 에이전트포스: CRM에 통째로 내장된 완제품 에이전트

세일즈포스 에이전트포스는 CRM 데이터와 업무 프로세스, 그리고 에이전트가 하나의 플랫폼 안에서 통합된다는 점을 최대 무기로 삼는다. 18만 곳이 넘는 기존 고객은 추가 통합 없이 곧바로 에이전트를 배포할 수 있으며, 이 결합은 경쟁사가 복제하기 어려운 해자다. 거버넌스를 데이터·프로세스와 함께 통째로 제공하는 완제품 구독 모델의 대표 격이다. 실적이 이 모델의 설득력을 뒷받침한다. 2026년 5월 발표된 최신 분기 실적 기준, 에이전트포스의 연간 반복 매출(ARR)은 12억 달러로 전년 대비 205% 성장했고, 한 분기 만에 4억 달러가 더해졌다. 에이전트 작업 단위는 38억 건, 처리 토큰은 28조 6,000억 건에 이른다.[44] 백오피스의 수동 프로세스를 에이전트가 담당하는 에이전트포스 오퍼레이션스가 더해졌고, 플랫폼 전체 지연시간을 70% 낮추는 성능 개선이 이어졌다.

42) LangChain. Building Enterprise Agents with NVIDIA. LangChain Blog, 2026.03
43) Panto. CrewAI Platform Statistics getpanto.ai, 2026
44) Salesforce. Salesforce Delivers Record First Quarter Fiscal 2027 Results. Salesforce Investor Relations, 2026.05.27

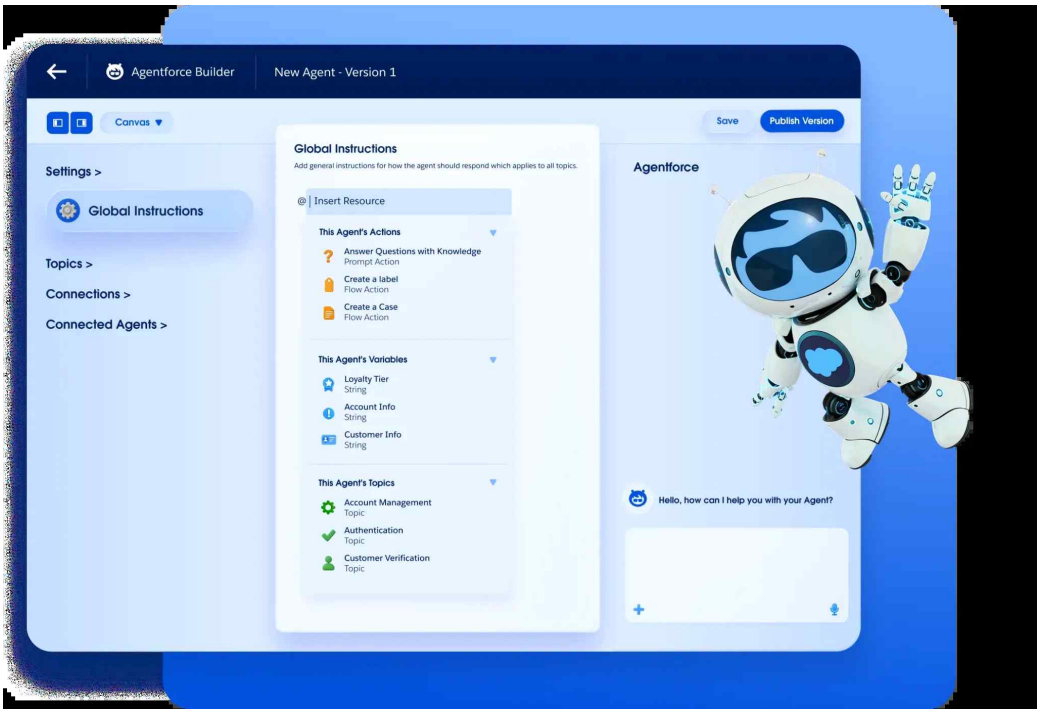

그림 8 세일즈포스 에이전트포스 빌더 — 대화형 빌드·테스트·배포 단일 워크스페이스 (출처: 세일즈포스)

한계는 CRM이라는 중심축 자체에서 비롯된다. 세일즈포스 생태계 밖의 시스템을 통합하려면 여전히 뮬소프트에 의존해야 해 복잡성과 비용이 추가되고, 제조·물류·연구개발처럼 CRM에서 먼 도메인에서는 적용 범위가 좁다. 고객 데이터를 통합하는 데이터 플랫폼인 데이터 360과 에이전트포스를 묶은 번들의 비용이 중견기업에게 부담이라는 점도 지적된다.

한국의 현실: 사는 솔루션이 아니라, 함께 일할 파트너

국내 CIO·CTO에게는 계층이 다른 고민이 있는데, 한국 시장에는 아직 바로 집어 도입할 국산 에이전트 플랫폼 제품이 뚜렷하지 않다는 점이다. 국내 IT 서비스 기업들이 실제로 내놓는 주된 형태는 완제품이 아니라 고객 현장에 밀착해 그 조직의 문제를 함께 푸는 구축이다. 전통적으로 SI 프로젝트라 불리던 이 영역은, 팔란티어가 대중화한 현장 배치 엔지니어(Forward Deployed Engineer, FDE) 모델로 성격이 진화하고 있다. 그래서 국내에서의 선택지는 "어느 국산 제품을 살까"가 아니라 "글로벌 제품을 도입해 우리가 운영할 것인가, 아니면 국내 파트너와 함께 밀착 구축을 진행할 것인가"로 정리된다. 망분리나 데이터 레지던시가 연관되어 글로벌 관리형 플랫폼—한국 리전이 없는 AWS

에이전트코어 등—이 제약되는 조직일수록 후자의 무게가 커진다. 그리고 업체들은 글로벌 제품을 국내 파트너가 현장에 이식하는 형태로 자주 결합한다.

맺으며: CIO와 CTO를 위한 제언

에이전트 플랫폼 시장은 이제 막 열렸고, 그 위에서 B2B 혹은 엔터프라이즈의 판이 새로 만들어지고 있다. 글로벌 제품과 국내 밀착 파트너, CSP와 MSP, 모델을 만드는 쪽과 그것을 현장에 이식하는 쪽이 합종연횡하며 조합을 계속 바꿔내는데, 이런 국면에서 핵심은 무엇을 고정하고 무엇을 유동적으로 둘 것인가를 나누는 데 있다. 거버넌스의 통제권은 어떤 선택을 하더라도 변하지 않는 축이므로 먼저 고정하고, 그 위아래에 놓이는 제품과 파트너, 모델의 조합은 원하는 기준에 맞춰 갈아끼울 수 있도록 유동적으로 설계하는 것이다. 그래서 '어떤 벤더가 이길 것인가'를 맞히는 눈보다, 통제권을 쥔 채 필요한 조합을 설계하는 아키텍트의 시각이 그 어느 때보다 중요해졌고, 이 관점에서 다음 네 가지를 제언한다.

1. 빌드가 아니라 평가와 거버넌스로 벤더를 검증하라

빌드는 상향 평준화되었기에 "이 플랫폼으로 에이전트를 만들 수 있는가"는 더 이상 변별력이 없다. 물어야 할 것은 "만든 뒤에 평가하고 통제할 수 있는가"다. 에이전트의 인벤토리가 있는가, 각 에이전트에 신원이 부여되는가, 감사 로그가 남는가, 이상 행동을 탐지하고 정책으로 사전에 차단할 수 있는가를 점검할 것을 권한다. 평가 엔지니어링은 아키텍처 설계 단계에서부터 포함하기를 특히 권한다.

2. 네 경로 중 우리 조직의 통제권 위치를 먼저 정하라

기능이 아니라 통제권으로 경로를 정해야 한다. M365 기반이고 거버넌스를 벤더 인프라에 위임해도 좋다면 코파일럿 스튜디오와 에이전트 365가, 멀티프레임워크 유연성과 자체 통제가 중요하다면 에이전트코어 위의 랭그래프가, CRM이 핵심 데이터 소스라면 에이전트포스가, 클라우드 종속을 피하고 온프레미스를 병행해야 한다면 랭그래프와 랭스미스가 적합하다. 다만 현실의 다수는 하이브리드로 수렴하므로, 처음부터 A2A와 MCP 호환성을 아키텍처 요구사항에 넣을 것을 권한다.

3. 국내 파트너를 택한다면, 제품이 아니라 관여를 평가하라

앞서 짚었듯 규제·소버린 요건으로 글로벌 관리형 플랫폼이 제약되는 환경에서 특히 국내 파트너의 밀착 구축은 현실적 대안이 되는데, 이는 완제품 구매가 아니라 FDE식으로 함께 진행하는 프로젝트다.

따라서 제품 사양이 아니라 그 팀의 도메인 이해도, 개념 검증을 프로덕션으로 넘길 역량, 운영 보장(SLA·장애 대응·모델 라이프사이클), 동종 규모 레퍼런스를 기준으로 평가하고, 국내 스택 역시 MCP를 채택하는 만큼 글로벌 스택과의 상호운용 경로를 면밀히 고려할 것을 권한다.

4. 멀티에이전트는 목적이 아니라 수단이다

단일 에이전트 대비 '멀티에이전트가 곧 더 낫다'는 명제는 성립하지 않는다. 명확한 입력과 출력, 단일 도메인, 10분 이내 실행이라면 단일 에이전트로 충분하다. 병렬 처리가 필요하거나 역할 분리가 품질을 높이는 복잡한 워크플로, 수 시간 이상의 장시간 실행일 때 비로소 멀티에이전트가 정당화된다. 멀티에이전트는 복잡성과 디버깅 비용, 장애 전파의 위험을 함께 키우기 때문인데, A2A가 표준으로 굳어지는 지금도, 오케스트레이션의 정당성은 "그 문제가 정말로 여러 에이전트를 요구하는가"라는 물음에서 나온다는 점을 잊지 말아야 한다.

AI 시대, 에이전트의 시대에 합종연횡은 계속될 것이다. 그러나 통제권을 쥔 조직에게 흔들리는 판은 위협이 아니라 훌륭한 선택지가 될 것이다. 이 복잡한 시대, 배울 것이 더 많아진 이 시대를 살아 내는 리더들을 다시 한번 응원한다.