디지털서비스

이슈리포트

Digitalservice Issue Report

피지컬AI를 위한 플랫폼 현황: 2부

NIA 한국지능정보사회진흥원 디지털서비스 이용지원시스템

2026 Vol. 02

CONTENTS

Digitalservice Issue Report

디지털서비스 이슈리포트

01 피지컬AI를 위한 플랫폼 현황: 2부

| 한상기 테크프론티어 대표

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

지난 1월에는 피지컬AI를 위한 플랫폼 현황을 정리하면서 기반이 되는 VLA 모델의 최근 발전 동향과 여러 모델을 소개했다. 이번 달에는 피지컬AI를 위한 현재 가장 중요한 연구 주제인 월드 모델 또는 월드 파운데이션 모델과 이를 기반으로 플랫폼을 개발하는 주요 기업의 전략을 소개하도록 한다.

월드 모델이란 무엇인가?

지난달에 소개한 VLA 모델만으로는 즉각 반응은 가능하지만, 미래 예측 능력이 부족하고, 장기 계획을 세우기 어려우며, 새로운 물리 상황을 일반화하기 어렵다. 즉 정책 학습으로는 물리적 특성에 대한 이해를 충분히 갖추기 어렵기 때문이다. 예를 들어 물체가 넘어질 수 있음에 대한 예측, 밀면 어디로 굴러갈지 예상하기 어렵고, 보이지 않는 물체에 대한 상태 추정이 불가한데 이는 VLA에는 세상에 대한 동력학 모델이 없기 때문이다.

월드 모델은 에이전트가 환경의 잠재 상태 동력학을 내부적으로 시뮬레이션하는 생성 모델이라고 할 수 있다. 여기에는 관측을 압축 세계 상태로 변환하는 '상태 표현 모델', 행동이 세상을 어떻게 바꾸는지 학습하는 다이나믹스 모델, 실제로 행동하기 전에 내부 시뮬레이션을 수행하는 이미지네이션/롤아웃 모델의 세 가지 구성 요소가 있다.

월드 모델에 대해 얀 르쿤은 그의 JEPA 모델이나 뉴립스 발표 등을 통해 '월드 모델은 에이전트가 행동의 결과를 예측하기 위해 학습한 세계의 내부 에너지 기반 표현이다."라고 말했으며, 페이페이 리는 공간 지능 이니셔티브 등을 통해 월드 모델은 '에이전트가 3D 공간 속에서 자신과 환경의 관계를 이해하도록 만드는 구조적 세계 표현이다.'라고 말했다. 이들의 주장은 부분적으로는 피지컬 AI에서 언급하는 월드 모델에 대한 정의를 담고 있지만 둘 다 완전히 피지컬AI를 위한 월드 모델의 특성을 담아내고 있지는 못한다.

얀 르쿤의 모델은 체화된 물리적 현실성이나 에이전트가 계획한 행동이 실제 물리 세계에서 실행 가능하는지 판단하는 행동 실행 가능성 계층이 명확하지 않으며, 페이페이 리의 정의는 세계가 무엇으로 이루어졌는지를 정의하는 점에서 엔비디아의 옴니버스와 유사함이 있지만 예측 인과성이나 문제 해결을 위한 계획 수립의 루프가 부족해 인지적 그라운딩에 머물고 있다고 평가할 수 있다.

엔비디아는 월드 모델을 자사 블로그를 통해서 다음과 같이 간결하게 정의하고 있다.[1)]

> 월드 모델은 물리적 특성과 공간적 속성을 포함한 현실 세계의 역학을 이해하는 신경망이다. 텍스트, 이미지, 비디오, 움직임 등의 입력 데이터를 사용하여 현실적인 물리적 환경을 시뮬레이션하는 영상을 생성할 수 있다.

지니 3를 통해 범용 월드 모델을 개발하고 있는 구글의 경우는 월드 모델을 AGI 구현을 위한 근본적인 접근으로 설명하고 있다.[2)]

> "세계 모델은 환경의 역동성을 시뮬레이션하여 환경이 어떻게 진화하고 행동이 어떤 영향을 미치는지 예측한다. … 특정 환경을 위한 에이전트를 개발한 경험이 있지만, 일반 인공지능(AGI)을 구축하려면 현실 세계의 다양성을 탐색할 수 있는 시스템이 필요하다. 지니 3는 사용자가 움직이고 세상과 상호작용할 때 실시간으로 앞길을 생성한다. 역동적인 세계의 물리법칙과 상호작용을 시뮬레이션하는 동시에, 획기적인 일관성을 통해 로봇 공학, 모델링, 애니메이션, 픽션 제작부터 장소 탐색 및 역사적 배경에 이르기까지 모든 현실 세계 시나리오를 시뮬레이션할 수 있다."

월드 모델은 피지컬AI의 보조 시스템이 아니라 지능을 반응 시스템에서 예측 시스템으로 전환하는 핵심 구조이다. 피지컬AI는 행동을 하기 전에 세상의 미래 상태를 예측할 수 있어야 하며, 언어 모델이 다음 토큰을 예측하는 것에 머문다면, 월드 모델은 물체의 운동, 접촉, 마찰, 가려짐, 인간 행동, 장기 환경 변화 등을 모두 예측해야 하기 때문이다. 이에 대한 상태 변화를 내부적으로 모델링하는 것이 월드 모델이다.

엔비디아의 월드 파운데이션 모델

피지컬 AI를 에이전틱 AI 다음의 화두로 미는 엔비디아는 월드 모델을 다음과 같이 분류한다.

---

1) NVIDIA, "What Is a World Model?" https://www.nvidia.com/en-us/glossary/world-models/
2) Google, "Project Genie: Experimenting with infinite, interactive worlds," Jan 29, 2026

- 예측 모델: 텍스트 프롬프트, 입력 비디오 또는 두 이미지 사이의 보간을 기반으로 세계를 예측하고 연속적인 움직임을 합성한다. 이를 통해 사실적이고 시간적으로 일관된 장면을 생성할 수 있으므로 비디오 합성, 애니메이션 및 로봇 동작 계획과 같은 응용 분야에 유용하다.
- 스타일 전송 모델: 콘트롤넷(ControlNet)이라는 모델 네트워크를 사용하여 특정 입력에 따라 출력을 제어한다. 콘트롤넷은 분할 맵, 라이다 스캔, 깊이 맵 또는 에지 검출과 같은 구조화된 지침에 따라 모델 생성을 조절한다. 입력 지침을 시각적으로 반영함으로써, 레이아웃과 움직임을 제어하는 동시에 텍스트 프롬프트에 기반한 다양하고 사실적인 결과물을 생성할 수 있다. 따라서 디지털 트윈 시뮬레이션 및 환경 재구성과 같이 구조화된 이미지 또는 비디오 합성이 필요한 응용 분야에 유용하다.
- 추론 모델: 다양한 형태의 입력을 받아 시간과 공간에 걸쳐 분석하며, 강화 학습 기반의 사고 연쇄 추론 방식을 사용하여 상황을 파악하고 최적의 행동을 결정한다. 이러한 모델을 통해 실제 데이터와 합성 데이터를 구분하고, 로봇 학습에 유용한 데이터를 선택하고, 로봇의 행동을 예측하고, 자율 시스템의 물류를 최적화하는 등 복잡한 작업을 처리할 수 있다.

그러나 엔비디아가 가장 강조하는 것은 코스모스라는 월드 파운데이션 모델로 확장성과 일반화 가능성 요건을 충족하는 특수한 유형의 월드 모델이다. 방대한 양의 레이블이 지정되지 않은 데이터셋으로 학습된 이러한 신경망은 다양한 피지컬 AI 작업에 적용될 수 있다. 뛰어난 일반화 능력 덕분에 개발자는 더 작은 규모의 특정 작업 데이터셋으로 추가 학습을 진행할 수 있는 사전 학습된 기본 모델로 활용할 수 있어, 다양한 피지컬 AI 애플리케이션 개발 속도를 크게 향상시킬 수 있다.

코스모스는 2025년 1월 CES에서 처음 발표했다. [3] 코스모스는 단순 월드 모델이 아니라 예측과 세계 생성, 물리 추론을 수행할 수 있는 모델로 특정 에이전트의 내부 모델이 아니라 로봇, 차량, 산업 시스템이 공유하는 세계를 생성하는 인프라 역할을 하기 때문이다. 파운데이션 모델이라는 단어에 방점이 찍혀 있다.

다음에 설명할 테슬라의 경우 현실 세계를 통해 경험을 축적해 신경망이 월드 모델을 묵시적으로 내포하게 만드는 전략이라면, 엔비디아는 기본적으로 현실 세계에서는 수십억 번의 상호작용, 실패 경험, 희귀 상황, 접촉 다이나믹스를 수집하기 어렵다는 판단에 AI가 학습할 수 있는 세계 자체를 대량 생산하는 방안을 취했다.

코스모스는 사전 학습 모델을 기반으로 사후 학습이 가능하며, 다양한 도메인에서 추가로 제공하는 비디오 데이터를 공급할 수 있는 파이프라인을 제공한다.

3) NVIDIA,"NVIDIA Launches Cosmos World Foundation Model Platform to Accelerate Physical AI Development," Jan 6, 2025

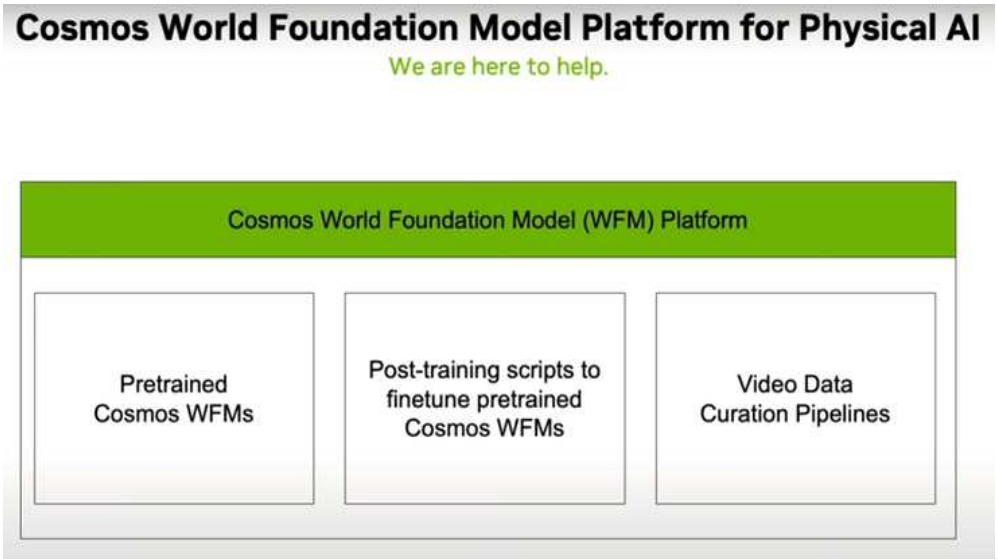


그림 1 코스모스 WFM 플랫폼의 구성

코스모스를 만들기 위해 엔비디아가 투입한 자원을 보면 DGX 클라우드에 있는 H100급 GPU 1만 장 이상을 사용했으며 2천만 시간 규모의 영상 데이터를 사용해 사전 학습을 했다. 여기에는 드라이빙, 손동작, 인간 동작, 공간 인지, 자연 동역학 등 다양한 분야의 영상 데이터가 사용되었음을 알 수 있다.

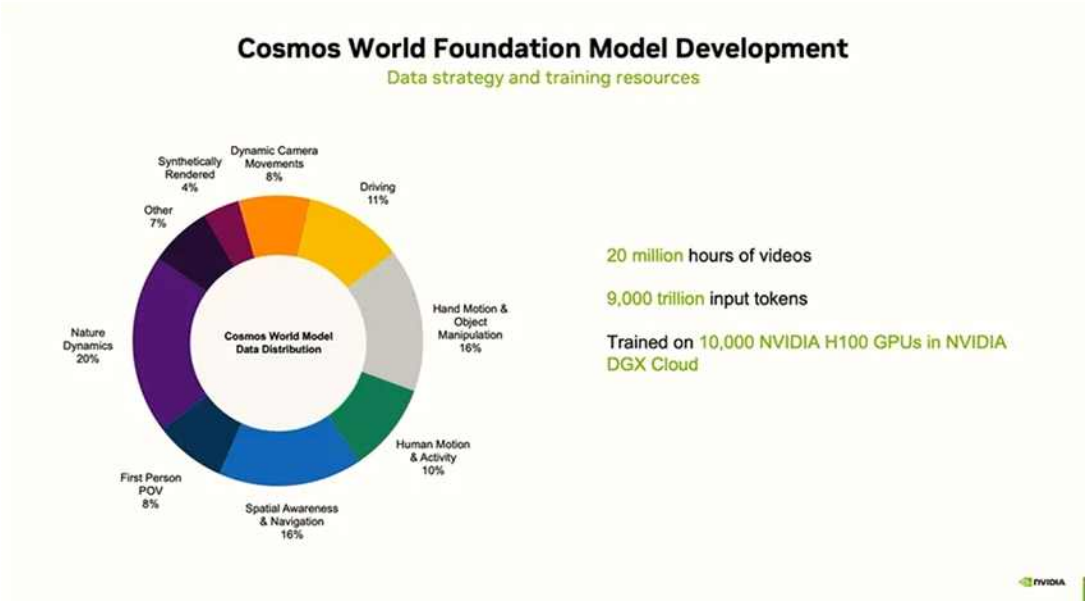


그림 2 코스모스 개발에 투입한 자원들

엔비디아는 피지컬 AI를 위한 플랫폼을 옴니버스라는 물리적으로 일관된 디지털 세계를 실행하는 시뮬레이션 운영체제와 그 위에 옴니버스에서 실행될 세계를 생성하는 코스모스, 그리고 최상위에 세계를 이해하고 행동을 생성하는 VLA 모델로 계층화했다. 코스모스에서 생성한 세계는 물리적 타당성을 옴니버스를 통해서 검증한다. 엔비디아의 VLA는 직접 현실에서 배우는 것이 아니라 대부분의 학습은 코스모스와 옴니버스를 통해서 이루어진 시뮬레이션 세상에서 배운다.

VLA 모델은 분야별로 하나씩 발표하는데, 휴머노이드 로봇을 위한 GR00T와 자율 주행을 위한 알파마요, 그리고 GR00T의 리즈닝을 강화한 아이작(Isaac) GR00T N1.6 등이 있다. 이와 관련한 내용은 지난 1월에 소개를 했기 때문에 그 부분을 참고하기 바란다.

엔비디아가 피지컬 AI를 위한 플랫폼을 옴니버스, 코스모스, VLA 모델로 계층화한 이유는 지능, 세계, 신체의 스케일 법칙이 다르기 때문이다. 특히 물리는 신경망에 맡기면 안 된다는 판단으로 옴니버스라는 시뮬레이션 엔진으로 분리했다. 월드 모델은 계속 커질 수 있지만 실제 물리적 개체인 로봇이나 차량에 들어가는 온보드 모델을 작아야 하기 때문에 코스모스는 데이터센터에서 동작하는 모델로 만들고, 각 신체에 필요한 지능 구조와 특성에 맞춘 VLA를 만들어서 실제 세계에서 동작하도록 하는 것이다. 다시 말해 옴니버스를 통해서 물리 정확도를 확보하고, 코스모스를 통해 데이터 다양성을, VLA를 통해 행동 최적화를 꾀하는 것이다.

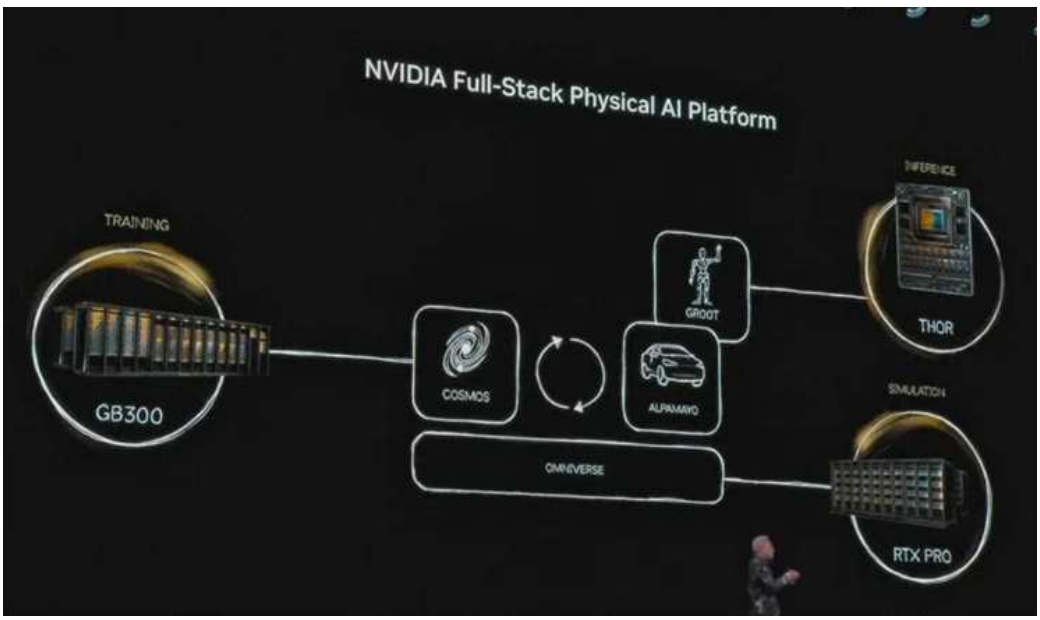


그림 3 엔비디아의 피지컬 AI 풀 스택 플랫폼

엔비디아는 피지컬 AI를 볼 때, AI가 세계 안에서 동작하는 것이 아니라 세계 인프라 위에서 동작하는 애플리케이션으로 본다. 이런 계층 구조와 역할 분리를 통해서 어떤 로봇 회사, 자동차 회사, 제조 공장이라도 엔비디아 세계 위에서 개발이 가능하도록 하는 것이다. 엔비디아는 로봇을 만드는 것이 아니라 모든 로봇의 기반을 장악하고자 한다.

또한 이런 계층적 구조를 통해 기술적 확장성뿐 아니라 정책 투명성과 설명 가능성을 구조적으로 가능하게 하고자 한다. 엔드투엔드 방식은 책임 소재 분리가 불분명하다. 그러나 엔비디아의 아키텍처는 실제로 어떻게 동작했는지, 어떤 시나리오로 훈련했는지, 왜 이 행동을 선택했는가를 분리해서 파악할 수 있게 했다. 이런 구조는 여러 규제 기관이 요구하는 조건을 잘 만족시킬 수 있으며, AI 얼라인먼트 측면에서도 중요하다.

그러나 단점으로 말할 수 있는 측면도 존재하는데, 먼저 월드와 에이전트를 분리함으로써 생기는 현실 갭이다. 보통 Sim-to-Real 갭이라고 하는데 에이전트는 시뮬레이션 세계에 최적화되어 있지만 시뮬레이션이 실제 세상을 완벽히 재현하기 어렵기 때문이다. 예를 들어, 미세 마찰 변화, 센서 열화, 재질 불확실성, 인간의 비정형 행동 등이 발생하는 것이 실제 세상이기 때문이다.

두 번째의 문제는 통합 지능 또는 창발성의 발생을 억제하는 문제이다. 이 방식으로는 지능이 환경과 공진화하기 어렵기 때문에 인간식 체화 학습과는 거리가 생긴다. 세 번째로는 계층을 분리하면 투명성이 높아지지만, 문제 발생 시 책임 경계가 존재할 수 있다. 실제 산업에서 통합 악몽이라고 부르는 이 문제는 명확히 한 곳의 책임으로 규정하기 어려운 상황이 발생한다는 것이다.

또 다른 문제는 엔비디아에게는 이점이지만 이 방식은 대규모 GPU 클러스터가 필요하고 지속적인 월드 생성과 시뮬레이션이 필요하기 때문에 컴퓨팅에 들어가는 비용이 크게 증가한다. 이는 엔비디아에게는 기회지만 피지컬 AI 생태계에는 진입 장벽이 될 수 있다. 마지막으로 플랫폼 의존성인데, 이 구조가 성공하면 모든 기업이 엔비디아의 기술 스택에 의존해야 한다는 점이다. 이 점은 국내 기업에게도 매우 중요한 전략적 의사 결정이 필요한 부분이다.

엔비디아 접근 방식에 대해 그리고 다음에 소개할 구글의 방식 모두 과연 월드 시뮬레이션이 진짜 세계를 이해하는 지능을 만들어 낼 수 있는가 하는 질문에서 자유롭지 않기 때문에 가상 세계에서는 완벽하지만, 현실에서는 취약한 AI를 만들어 낼 수 있다. 따라서 현실 데이터를 어떻게 효과적으로 최종 모델에 반영하면서 심투리얼의 문제를 해결해야 하는 가는 여전히 남는 숙제이다.

테슬라의 월드 모델 기반 자율 주행

테슬라는 자율 주행을 인지와 규칙의 기반이 아니라 월드 모델링과 예측을 통해 자율 주행을 실시간 월드 모델 구축 문제로 접근했다. 즉, 차량이 세계를 내부적으로 이해하고 미래를 예측하도록 하는 것이다. 그러나 테슬라는 '월드 모델'이라는 용어를 자주 쓰지 않으며, 오토노미 스택 전체가 사실상 월드 모델이다.

여기에는 세계의 상태를 표현하는 공간 점유 모델인 Occupancy Network가 있으며 이는 주변 공간 전체를 3D 확률 공간으로 표현한다. 즉, 객체 리스트가 아닌 연속적 세계 표현을 담으며, 다중의 카메라는 통합된 4차원 세상을 통해 시간적으로 세상이 어떻게 달라지는지를 추정한다.

자율 주행에서 핵심 중 하나는 지금 보이지 않는 것이 어디에 있는가에 대한 추론인데, 예를 들어 가려진 보행자나 코너 뒤의 차량 같은 것을 말하며, 네트워크 내부 메모리를 통해 과거 관측 유지, 속도 추정, 의도 예측을 한다. 이는 사실상 잠재적 세계 상태 추적 능력이다.

그러나 테슬라 오토노미의 핵심은 예측 능력이다. 차량은 단순히 현재를 인식하는 것이 아니라 보행자가 건널 확률, 차량 합류 의도, 신호 변화, 충돌 가능 추적 등을 현재의 세계 기반으로 미래 세계의 확률을 계산한다. 그다음은 계획 수립 단계인데, 현재 세계 상태에서 여러 미래 궤적을 생성하고, 충돌/안전/편안함을 평가한 후, 최적 행동을 선택한다.

테슬라의 엔드-투-엔드 오토노미는 인지-계획수립-제어를 하나의 신경망으로 통합하는 것이지만, 내부적으로 세계 표현, 예측, 계획 수립이 잠재 공간에서 발생하고 있다. 테슬라의 월드 모델의 가장 큰 특징은 실제 주행 데이터를 이용해 수백만 대의 차량이 세계 모델 학습 센서의 역할을 하고 있는 것이며, 명시적인 물리 모델이 없이 뉴럴 네트워크 안에 묵시적으로 학습된 물리법칙을 갖고 있는 것이다.

테슬라의 아쇼크 엘루스와미는 2025년 10월 엑스닷컴을 통해 테슬라의 자율 주행 접근 방식에 대한 상세한 의견을 올렸다.[4] 그 내용을 기반으로 몇 가지 논점을 살펴보기로 한다. 우선 테슬라가 엔드투엔드 방식으로 자율 주행을 접근하는 데에는 다음과 같은 이유가 있다.

- 인간의 가치를 배우는 것은 매우 어렵다. 데이터를 통해 가치를 배우는 것이 훨씬 쉽다.

4) Ashok Elluswamy, "Tesla's approach to Autonomy," X.com, Oct 24, 2025

- 인지, 예측 및 계획 간의 인터페이스는 명확하게 정의되어 있지 않다. 엔드투엔드 방식에서는 제어 장치에서 센서 입력에 이르기까지 모든 단계에서 변화가 발생하여 전체 네트워크가 총체적으로 최적화된다.
- 실제 로봇 공학의 방대한 데이터와 긴 꼬리 패턴을 처리할 수 있도록 손쉽게 확장 가능하다.
- 동질적 컴퓨트에 결정론적인 잠재성을 갖는다.
- 전체적으로 확장 법칙이 맞는 면이 있다.

그러나 실제 세상에서의 주행 데이터로만 자율 주행을 완성하기에는 매우 드문 상황 데이터가 충분하지 않다. 이런 문제를 해결하기 위해 테슬라도 신경망 기반 월드 시뮬레이터를 만들었다. 이 시뮬레이터는 테슬라가 직접 구축한 방대한 데이터셋을 기반으로 학습을 진행하지만 기존 방식처럼 현재 상태를 기반으로 다음 행동을 예측하는 대신, 신경망 기반 월드 시뮬레이터는 현재 상태와 다음 행동을 입력받아 미래 상태를 예측한다. 이렇게 생성된 시뮬레이터는 에이전트 또는 정책 기반 AI 모델과 연동하여 폐쇄 루프 방식으로 실행하고 성능을 평가할 수 있다.

월드 시뮬레이터는 테슬라가 자체적으로 학습시켜 차량의 모든 카메라 및 기타 센서 데이터를 생성한다. 시뮬레이터는 인과 관계를 기반으로 하며 주행 정책 모델의 명령에 반응하는데, 빠른 속도를 자랑하면서도 고해상도, 고프레임률, 고품질 센서 데이터를 생성할 수 있다.

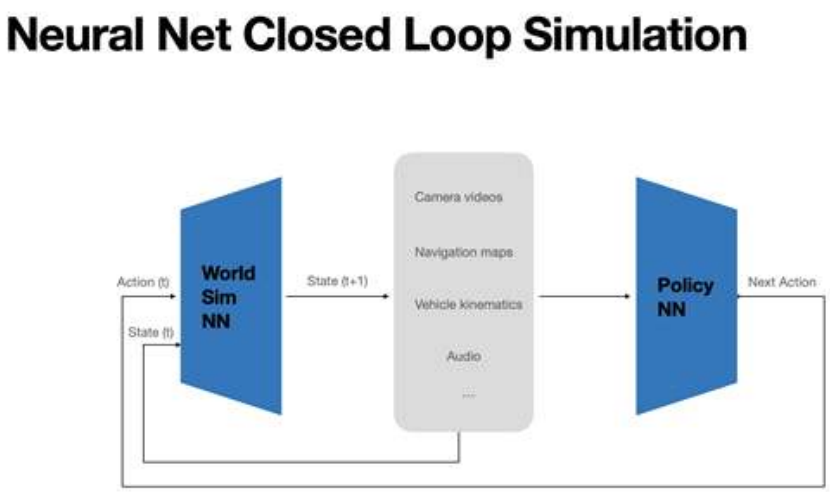


그림 4 폐쇄 루프 방식의 월드 시뮬레이션

이러한 시뮬레이션은 최신 주행 모델을 과거 데이터와 비교하여 검증하는 데 사용할 수 있으며, 추가적인 예외 상황을 테스트하기 위해 새로운 적대적 시나리오를 인위적으로 생성할 수도 있다.

그림 5 월드 시뮬레이터와 적대적 이벤트 주입

테슬라는 이런 개발 환경과 모델 적용을 차량 자율 주행뿐만 아니라 휴머노이드 로봇인 옵티머스에도 적용할 수 있다고 주장한다. 동일한 비디오 생성 모델이 테슬라 기가팩토리를 탐색하는 옵티머스 로봇에도 적용할 수 있다는 것이다.

그림 6 테슬라의 엔드투엔드 모델과 월드 시뮬레이터를 옵티머스에 적용하기

그러나 하나의 거대한 신경망을 통해 내포된 월드 모델을 표현하고 물리적 법칙을 현실 데이터를 통해서 신경망이 학습하도록 하는 것은 완전함이나 확장성에 한계를 가질 수 있고, 오류가 발생했을 때 이를 추적하거나 설명하는 것이 쉽지 않을 수 있다. 나아가 엔비디아가 추구하는 책임성과 투명성을 보장하는 것이 어렵기 때문에 이 방식을 다른 영역에서 모두 받아들일 수 있을 것인가 하는 의문이 든다.

이런 측면에서 중국의 자율 주행차 기업들이 초기에는 주로 구글 웨이모의 방식을 채택했다가 최근에 테슬라 방식으로 전환을 하고 있으나 여러 기업은 카메라 기반의 완전한 엔드투엔드가 아니라 안전 문제로 센서와 고해상도 지도를 포함하는 하이브리드 방식을 취하고 있다. 여기에는 화웨이 ADS, BYD, 모멘타, 샤오미, 샤오 펭 등이 포함되어 있으나 향후에는 대부분 월드 모델과 지도가 필요하지 않는 엔드투엔드 신경망을 기반으로 하는 테슬라 모델로 발전할 것으로 본다.

구글의 월드 모델과 월드 시뮬레이터

구글은 10여년 이상 월드 모델에 대한 연구를 진행했다. 2019년에서 2021년 사이에는 MuZero를 발표했으며 이는 최초의 월드 모델 기반 계획수립 알고리듬이며 세계 상태를 직접 관측하지 않고 잠재 세계를 구성해 내부 롤아웃으로 행동을 결정할 수 있음을 보였다.[5] 주로 게임을 마스터할 수 있는 알고리듬이었다.

그다음에 등장한 것이 드리머(Dreamer)이다.[6] 드리머는 월드 모델을 먼저 배우고 그 안에서 행동을 학습하는 방식을 취했다. 드리머 버전 3는 단일 설정으로 150개 이상의 제어 과제를 해결하며 범용 제어 문제에서 성능을 입증했다. 이는 로보틱스 월드 모델에 직접적인 기반이 되었다.

2023년에 등장한 PaLM-E는 언어 모델과 피지컬 센서를 통합한 모델이다. 이를 통해 LLM이 세계 모델의 일부가 될 수 있음을 보였다. 그러나 구글은 인지와 제어만으로는 부족하다는 것을 깨닫고 중요한 것은 체화된 논증이라는 인식을 갖고 로보틱스 트랜스포머 계열인 RT-2를 발표했다. 이를 통해 로봇이 상황 의미 이해, 행동 목적 추론, 새로운 행동 생성을 수행해야 함을 확인했다. 구글은 인지를 기반으로 물리적 실체를 제어하기 위한 VLA 모델인 제미나이 로보틱스와 제미나이 로보틱스-ER을 발표했는데 이는 1월 호에서 소개했기 때문에 더 설명을 하지 않는다.

---

5) Google DeepMind, “MuZero: Mastering Go, chess, shogi and Atari without rules,” Dec 23, 2020
6) Google Research, “Introducing Dreamer: Scalable Reinforcement Learning Using World Models,” Mar 18, 2020

2025년부터 구글 딥마인드는 대규모 생성형 월드 시뮬레이터를 구축하고 범용의 월드 모델을 만들기 시작했는데, 대표적인 것이 지니(Genie)이다. 2024년 2월 지니 1을 시작으로 2025년 8월 지니 3를 발표했다. 이는 텍스트로부터 상호작용이 가능한 세계를 생성하고, 실시간 탐색이 가능하며, 행동에 따라 세계를 변화할 수 있게 했다. 즉, 세계를 이해하는 것이 아니라 세계 자체를 생성하는 방향을 선택했다. 즉 구글 딥마인드는 '세계 안에서 행동하는 AI'를 만들기 위해 게임 세계로 시작해, 생성 세계를 통해 실제 로봇 세계에 적용하기 위한 월드 모델을 단계적으로 확장해 왔다.

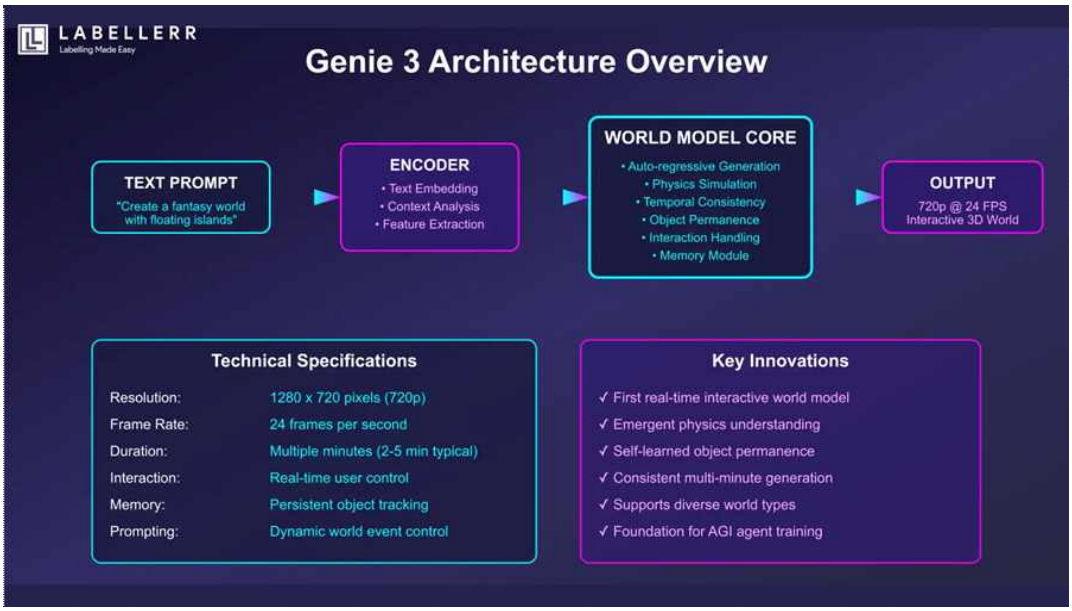


그림 7 지니 3 아키텍처 [출처: LABELLERR]

이는 구글이 자율 주행 같은 문제를 풀려는 것보다는 지능 자체의 생성 원리를 풀고자 하는 것이며, 구글이 갖고 있는 거대한 데이터를 기반으로 피지컬 AI를 적용할 수 있는 일반 월드 모델을 만들고자 했으며 이를 통해서 AGI로 가는 연구가 이어질 수 있다고 봤다. 딥마인드는 월드 모델은 물리적 환경에 대한 깊은 이해를 바탕으로 환경을 시뮬레이션해야 한다는 것이다.

테슬라가 현실 데이터를 이용하고자 했다면 딥마인드는 현실 데이터를 얻는 것은 너무 느리고 비용이 많이 드는 문제가 있다는 것 외에도 수조 번의 상호작용과 위험 없는 실험을 해야 하고 희귀한 상황을 생성하기 위해서는 AI가 스스로 세계를 형성해야 한다는 결론에 도달한 것 같다. 그러나 현실 세계에 기반을 두어야 하기 때문에 제미나이 로보틱스를 통해 실제 로봇 학습을 실행하고 체화된 지능을 위한 데이터셋을 구축하는 것도 사실이다.

지니 3의 한계는 아직 여러 가지가 있다. 현재는 에이전트가 수행할 수 있는 행동의 범위가 제한되며, 공유 환경에서 다른 에이전트와 상호작용이나 시뮬레이션 하는 것은 아직 연구 과제이다. 또한 실제 위치를 정확히 표현하는 수준의 시뮬레이션이 아니며, 텍스트 렌더링도 한계가 있고 아직은 몇 분 정도의 상호작용만 지원할 수 있다.

나아가며

알파고를 만든 데이빗 실버가 이제는 AI가 스스로 세계를 알아나가는 경험의 시대로 넘어서야 한다고 주장하면서 최근에 이네퍼블 인텔리전스라는 스타트업을 세운 점에 주목해야 한다. 이제는 AI가 텍스트가 아닌 상호작용과 피드백을 통해 학습해야 하며, 그렇게 하기 위해서는 세상에 대한 모든 특성과 제약을 담은 월드 모델이 가장 핵심에 있어야 하는 것이다.

지금까지 투자자들도 파운데이션 모델에 가장 많은 돈을 투자했다면 2030년까지는 가장 많은 투자는 월드 모델에 대한 투자일 것이라는 것이 피치북의 2026년 AI 아웃룩 자료에서 제시한 전망이다.

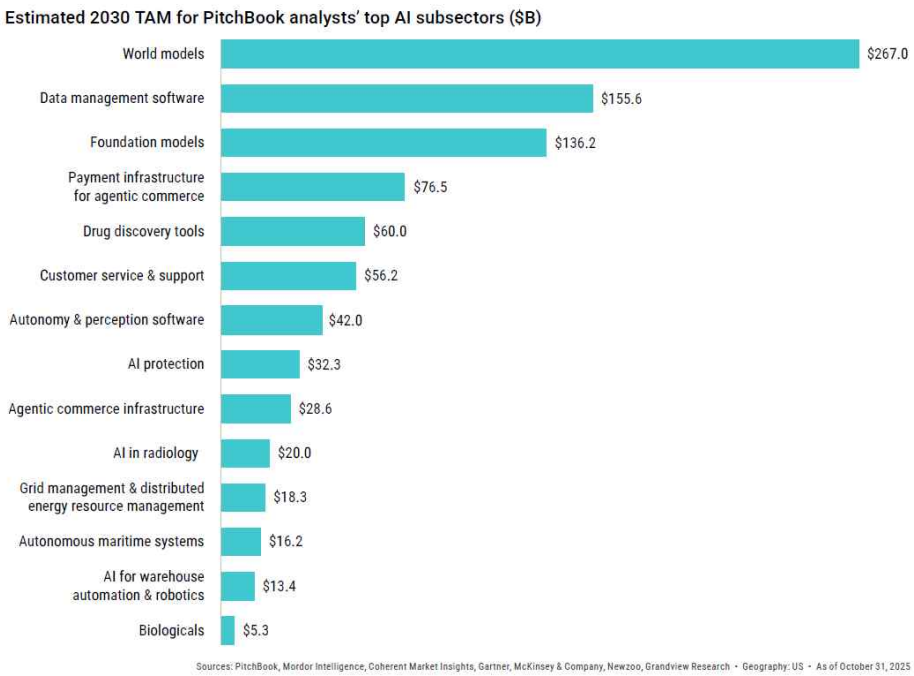


그림 8 2030년까지 AI 분야별 투자 전망 [출처: 핏치북]

우리가 지금까지 두 번에 걸쳐 피지컬 AI를 위한 AI 플랫폼의 현황을 살펴보았지만 국내에서 이에 대응할 수준의 자체 모델은 아직 매우 미흡하다. 맥스 얼라이언스를 통해 VLA 방식의 제조 특화 파운데이션 모델을 만든다는 것이 현재 진행되고 있는 수준이다.

그러나 피지컬AI에서 가장 중요한 핵심이 월드 모델임을 빨리 인식해야 하며, 이를 만드는 데는 매우 많은 시간과 자원이 필요한 점 역시 깨우쳐야 한다. 그러나 이런 월드 모델 (명시적이던 묵시적이던) 자체도 아무리 정교히 만들어도 실 세계를 완벽히 반영할 수 없을 것이라는 점에서 이 갭을 해결하기 위한 노력 역시 필요하다.

국가적으로는 엔비디아, 테슬라, 구글의 플랫폼이 뛰어나더라도 우리 피지컬 AI 산업의 기반을 해외 기술에 의존하게 할 수는 없다는 점에서 LLM 모델에서 소버린 AI 전략을 추진한 것과 마찬가지로 피지컬 AI 플랫폼에서도 소버린 정책이 필요하며 이를 어떤 단계로, 체계적으로 추진할 것인가는 우리에게 남아있는 매우 중요한 과제이다.

02 기밀 컴퓨팅이 여는 에이전트 AI 시대의 보안 혁신

| 윤대균 아주대학교

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

1. 기밀 컴퓨팅(Confidential Computing) 등장 배경

2023년 초, 삼성전자 직원들이 사내 업무에 ChatGPT를 활용하기 시작했다. 코드 디버깅, 회의록 요약, 소스코드 최적화 등 다양한 목적으로 사용하던 중 반도체 설계 관련 민감한 코드와 내부 회의 내용이 오픈AI의 서버로 전송됐다. 몇 주 뒤 이 사실이 알려지자, 삼성은 사내 생성형 AI 사용을 전면 금지했다.[7] 기술의 편의성과 데이터 보안 사이의 충돌이 가장 극적으로 드러난 사건 중 하나였다.

이 사건이 드러낸 본질적 문제는 데이터 통제권의 상실이다. 데이터를 외부 시스템에 넘기는 순간, 그것이 어떻게 처리되는지 사용자가 통제할 방법은 없다. 클라우드 제공업체 내부 직원이 접근할 수도 있고, 시스템 관리자 권한을 가진 누군가가 메모리를 들여다볼 수도 있다. 법원 명령이나 정부 요청으로 제3자에게 데이터가 넘어갈 가능성도 상존한다.

이렇듯 클라우드 컴퓨팅의 확산과 AI 기술의 발전은 기업에게 막대한 기회를 제공하는 동시에 새로운 보안 숙제를 던져주었다. 특히, 민감한 데이터가 클라우드상에서 처리됨으로써 발생할 수 있는 잠재적 위협은 데이터 프라이버시 및 규제 준수의 핵심 쟁점으로 떠올랐다. 암호화는 우리가 가진 최선의 방패였지만, 치명적인 빈틈이 있었다. 저장된 데이터(data at rest)는 암호화된다. 네트워크를 타고 이동하는 데이터(data in transit)도 TLS(Transport Layer Security)로 보호된다. 그런데 CPU가 연산을 수행하는 그 순간만큼은 데이터가 반드시 평문(plain text) 상태로 메모리에 올라와야 한다. 수십 년 동안 이 순간은 암호화의 손이 닿지 않는 사각지대였다. 이러한 가운데 기밀 컴퓨팅은 데이터 보안 패러다임을 한 단계 끌어올리는 혁신적인 기술로 주목받고 있다.

기밀 컴퓨팅 컨소시엄(CCC)[8]은 이 기술을 다음과 같이 정의한다.

7) AI타임스, “삼성, 챗GPT 데이터 유출 후 임직원 AI 사용 금지”, 2023.5.3
8) https://confidentialcomputing.io/

"하드웨어 기반의 신뢰 실행 환경(TEE: Trusted Execution Environment)에서 연산을 수행함으로써 사용 중인 데이터(data in use)를 보호하는 것."

기존의 데이터 보안 방식은 데이터가 저장되거나 전송될 때 암호화를 통해 보호하는 데 중점을 두었다. 하지만 데이터가 실제 CPU 메모리에서 처리되는 순간 데이터는 일반적으로 평문 형태로 존재하여 악의적인 공격자나 심지어 클라우드 서비스 제공업체에도 노출될 수 있는 잠재적 취약점을 가지고 있다. 기밀 컴퓨팅은 CPU 칩 내부에 외부에서 접근할 수 없는 격리된 실행 영역을 만들고, 그 안에서만 민감한 데이터를 평문 상태로 처리한다. 운영체제도, 하이퍼바이저도, 클라우드 제공업체 관리자도, 물리적으로 서버에 접근한 사람도 이 영역 안의 데이터를 열람할 수 없다.

그림 9 처리 중에도 보호되는 데이터 (출처: 엔비디아)

이 기술이 지금 주목받는 이유는 크게 세 가지로 들 수 있다. 첫째, 클라우드 보편화에 따른 보안 우려이다. 기업들은 클라우드의 효율성을 원하지만, 가장 민감한 워크로드를 제3자에게 맡기는 것을 꺼린다. 금융기관, 의료기관, 정부 기관은 클라우드 전환을 주저하거나 가장 중요한 데이터는 온프레미스에 남겨두는 하이브리드 전략을 택해왔다. 기밀 컴퓨팅은 클라우드 제공업체 스스로도 고객 데이터에 접근할 수 없다는 것을 기술적으로 보장함으로써 이 우려를 해소한다.

둘째, 규제의 강화다. EU의 GDPR, 2024년부터 단계적으로 적용되기 시작한 EU AI 액트, 미국의 주별 프라이버시 법안들이 데이터 처리의 모든 단계에서 보안을 입증하도록 압박하고 있다. 2024년 NIST는 사이버보안 프레임워크(CSF) 2.0[9]에 사용 중인 데이터 보호 지침을 추가했고, PCI DSS[10] v4.0.1은 휘발성 메모리를 포함한 데이터-인-유즈 보호 가이드라인을 명시적으로 포함시켰다. 기밀 컴퓨팅의 원격 증명(attestation)은 이러한 컴플라이언스 요건을 기술적으로 증명하는 유력한 수단이다.

9) https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf
10) Payment Card Industry Data Security Standard의 약자로, 비자, 마스터카드 등 주요 카드사들이 카드 소지자의 데이터를 보호하기 위해 제정한 글로벌 보안 규정이다.

셋째, AI 시대의 데이터 딜레마다. AI 모델을 학습하고 추론하는 데 필요한 가장 가치 있는 데이터는 대부분 민감한 정보다. 의료 기록, 금융 거래, 기업 내부 문서. 나아가 AI 모델 가중치 자체도 수억 달러를 투자해 만든 보호해야 할 핵심 자산이 됐다. 앞서 언급한 삼성 사례는 이 문제의 대표적인 한 예이다. AI가 기업 운영의 핵심 인프라가 될수록, AI 워크로드의 보안은 선택이 아닌 필수가 된다.

2. 핵심기술: TEE

신뢰 실행 환경(TEE: Trusted Execution Environment)의 핵심 개념은 “엔클레이브(enclave)”다. CPU 내부에 격리된 메모리 영역을 만들고, 그 안에서 실행되는 코드와 데이터는 외부에서 읽거나 수정할 수 없다. 운영체제 커널도, 하이퍼바이저도 접근이 차단된다. 심지어 엔클레이브를 실행하는 프로세스의 나머지 부분도 그 내부를 들여다볼 수 없다. 이것이 가능한 이유는 데이터가 L1/L2/L3 캐시를 거쳐 RAM에 저장될 때 CPU 칩 레벨에서 자동으로 암호화되고, 복호화는 오직 CPU 내부에서만 이루어지기 때문이다. 즉, 일반 운영체제와 완벽하게 분리된 실행 컨텍스트를 제공한다. TEE의 구현 방식은 **프로세스 기반 격리**에서 **가상 머신 기반 격리**로 진화해 가고 있다.

프로세스 기반 격리 (Application Enclave)

초기 기밀 컴퓨팅 기술을 대표하는 방식으로, 애플리케이션을 '신뢰할 수 없는 부분'과 '신뢰할 수 있는 부분(Enclave)'으로 나누어 설계한다. 민감한 데이터를 다루는 코드와 데이터만 엔클레이브 내부에 로드되며, 이 영역은 CPU 내부의 하드웨어 로직에 의해 보호된다. OS나 커널조차도 엔클레이브 메모리에 직접 접근할 경우 무의미한 암호화된 데이터만 보게 되거나 접근이 차단된다. 공격 표면(Attack Surface)이 매우 작아 보안성이 높지만, 이를 제대로 활용하기 위해서는 기존 애플리케이션을 엔클레이브와 호환되는 코드로 수정해야 하는 개발 부담이 있다. 인텔 SGX(Software Guard Extensions)[11]가 대표적인 예이며 TEE를 개척한 선구자로 볼 수 있다.

가상 머신 기반 격리 (Confidential VM)

엔클레이브 메모리 크기의 제한(초기 128MB)과 높은 코드 수정 비용, 캐시 블리드[12]나 포셰도우[13]와 같은 사이드 채널 취약점을 해결하기 위해 최근 클라우드 환경에서 주류로 부상한 방식이다. VM

11)https://www.intel.co.kr/content/www/kr/ko/products/docs/accelerator-engines/software-guard-extensions.html
12) 캐시 블리드(Cachebleed)는 인텔 프로세서의 L1 데이터 캐시에서 발생하는 부채널 공격(Side-Channel Attack)의 일종으로, 암호화 과정에서 비밀키를 탈취하는 취약점을 말한다.
13) 포셰도우(Foreshadow)는 인텔 SGX 환경을 무력화할 수 있는 추측 실행(Speculative Execution) 기반의 보안 취약점을 말한다.

전체를 하나의 TEE로 간주하여 보호한다. 인텔 TDX(Trusted Domain Extensions), AMD SEV(Secure Encrypted Virtualization)가 대표적인 예이다.

VM 전체를 격리함으로써 기존 애플리케이션을 수정하지 않아도 TEE 환경에서 실행할 수 있는 것이 장점이다. VM의 모든 메모리 페이지가 암호화되어, VM의 리소스를 관리하는 하이퍼바이저조차 VM 내부의 메모리 내용이나 CPU 레지스터 상태를 읽을 수 없다. VM 방식을 사용하면 소위 '리프트 앤 시프트(Lift and Shift)'가 가능하여, 기존 워크로드를 코드 수정 없이 그대로 보안 환경으로 이전할 수 있다. 이는 기밀 컴퓨팅 대중화에 결정적인 역할을 하고 있다.

메모리 암호화 및 키 관리

TEE에서 사용중인 데이터가 보호될 수 있도록 하는 기반 기술은 메모리 암호화 기술이다. 데이터가 CPU 코어 내부의 L1/L2 캐시에 있을 때는 평문 상태로 연산되지만, L3 캐시를 벗어나 메인 메모리(DRAM)로 이동할 때는 즉시 암호화된다. 이를 위해서는 암호화를 위한 키 생성 및 관리, 그리고 메모리의 무결성 보호가 중요하다.

- **키 생성 및 관리:** 암호화 키는 프로세서 내부의 난수 생성기에 의해 부팅 시 생성되며, 외부 메모리나 소프트웨어에 저장되지 않고 오직 하드웨어 레지스터 내에만 존재한다. 이 키는 시스템이 재부팅되면 소멸되므로, 물리적으로 메모리 칩을 탈취해도 데이터를 복구할 수 없다.
- **메모리 무결성 보호:** 단순 암호화뿐만 아니라, 공격자가 메모리의 특정 주소 값을 이전 시점의 값으로 되돌리는 '리플레이 공격(Replay Attack)'이나 다른 주소로 매핑하는 '메모리 리매핑 공격'을 방지하기 위한 무결성 검증 메커니즘이 적용된다. AMD SEV의 경우 VM 메모리 전체를 암호화할 뿐만 아니라 CPU 레지스터까지 암호화하여 데이터 노출을 원천 차단하며(SEV-ES), 메모리 무결성 보호를 위해 하이퍼바이저의 악의적 메모리 재매핑 공격을 방어하는 기능(SEV-SNP)을 제공한다.

원격 증명

TEE가 존재한다고 해서 충분하지 않다. 원격의 사용자가 "이 TEE가 진짜이고, 기대하는 코드가 그 안에서 변조 없이 실행되고 있다"는 것을 어떻게 확인할 수 있는가에 대한 답을 하는 것이 '원격 증명(Remote Attestation)'의 역할이다. TEE는 자신의 실행 상태(코드, 설정, 하드웨어 환경)를 서명된 보고서로 만들어 외부에 제출한다. 이 서명은 CPU 제조사의 루트 키에서부터 시작하는 신뢰 체인으로 검증된다. "신뢰하지 않아도 된다, 검증하면 된다"는 기밀 컴퓨팅의 핵심 가치 제안이 바로 여기서 나온다. 이는 최근 보안 주류 패러다임으로 자리잡고 있는 제로 트러스트 보안 모델과도 일맥상통한다.

- TEE 초기화: 하드웨어는 로드된 코드, 초기 데이터, 구성 설정 등을 포함한 해시값을 계산하여 안전한 레지스터에 저장한다.
- 보고서 생성: 외부 요청이 오면, TEE는 저장된 해시값에 하드웨어 고유의 비밀키로 서명하여 증명 보고서(Quote)를 생성한다.
- 검증: 사용자는 이 보고서를 제조사(Intel, AMD 등)의 증명 서비스나 제3의 검증 서비스를 통해 확인한다. 베라이종(Veraison)[14]은 칩 제조사가 아니더라도 원격 증명 서비스를 개발할 수 있는 도구를 제공하기 위한 오픈소스 프로젝트이다.

주요 기술 기업 동향

앞서 언급했듯이 TEE의 사실상 효시로 볼 수 있는 인텔 SGX가 있고, 이는 가상머신 기반의 TDX로 진화했다. AMD SEV는 가상머신 기반의 격리를 TEE의 주류 기술로 끌어올리는 데 큰 역할을 했다. 메모리 암호화뿐만 아니라 레지스터값까지 암호화하고, 또한 메모리 무결성 보호를 위한 공격에 대비한 방어 수단을 제공한다. AMD SEV는 마이크로소프트 애저, 구글 GCP, 아마존 AWS를 비롯한 주요 클라우드에서 이미 상용 서비스로 제공되고 있다.

ARM의 트러스트존(TrustZone)은 스마트폰과 임베디드 장치에서 TEE를 구현하는 기술이다. 프로세서를 '일반 세계'와 '보안 세계'로 분리해, 지문 인식, 생체 인증, DRM, 모바일 결제 같은 민감한 작업을 '보안 세계'에서만 처리한다. 삼성 녹스(Knox), 애플 시큐어 엔클레이브 등이 이를 기반으로 한다. 스마트폰 잠금 해제, 삼성페이·애플페이 결제, 패스키 인증이 모두 이 기술 위에서 동작한다.

엔비디아 H100은 GPU 최초로 하드웨어 기반 TEE를 지원한다.[15] AI 학습과 추론은 CPU가 아닌 GPU에서 이루어지며, 수십억 파라미터의 모델 가중치와 입력 데이터가 GPU 메모리에 올라간다. H100의 기밀 컴퓨팅 모드에서는 GPU 메모리가 암호화되고, GPU와 CPU 간 PCIe 버스 통신도 보호된다. 기밀 컴퓨팅 모드에서의 성능 오버헤드는 대부분의 LLM 추론 작업에서 5% 미만으로 측정됐으며, 2024년 10월 마이크로소프트 애저는 엔비디아 H100 기반 기밀 VM을 정식 출시했다.[16]

14) 와인 양조에서 포도가 익기 시작하는 시점을 나타내는 용어인 'Veraison'에서 따 왔다고 하며 'VERificAtIon of atteStatiON'의 약자임을 주장한다. https://www.veraison-project.org/book/introduction.html
15) Nvidia Technical Blog,"Confidential Computing on NVIDIA H100 GPUs for Secure and Trustworthy AI", Aug 03, 2023
16) Azure Blog, "General Availability: Azure confidential VMs with NVIDIA H100 Tensor Core GPUs", Sep 24, 2024

3. 에이전트 AI: 기밀 컴퓨팅의 본격적인 시험대

챗봇이 질문에 답하는 수준을 넘어, AI가 스스로 계획을 세우고, 도구를 호출하고, 여러 단계의 작업을 자율적으로 수행하는 에이전트 AI 시대로의 전환은 기밀 컴퓨팅을 선택적 보안 강화 수단에서 필수 인프라로 격상시킨다.

에이전트 AI와 보안 위협

에이전트 AI는 기존 AI와 근본적으로 다른 세 가지 특성을 갖는다. 첫째, '자율성(autonomy)'이다. 인간의 개입 없이 수십 개의 연속적 결정을 내리고 행동으로 옮긴다. 둘째, '메모리(persistent memory)'다. 여러 세션에 걸쳐 컨텍스트를 유지하며, 과거 상호작용과 사용자 정보를 기억한다. 셋째, '도구 접근(tool access)'이다. 이메일, 캘린더, 코드 저장소, 데이터베이스, 결제 시스템까지 외부 시스템과 연결되어 실제 세계에서 행동한다. 이 세 가지가 결합되면 AI는 단순한 소프트웨어가 아니라 매우 강력한 '디지털 행위자'가 된다.

에이전트는 사용자의 이메일을 읽고, 일정을 수정하고, 코드를 작성하고, 파일을 전송하고, 금융 거래를 실행할 수 있다. 이는 어떤 인간 직원보다 넓은 접근 권한을 가질 수 있다는 의미다. 보안 전문가들은 에이전트 AI를 '자율성을 가진 내부자 위협(autonomous insider threat)'으로 규정한다.[17] 에이전트를 실행하는 기업이 에이전트가 실제로 무슨 코드를 실행하는지, 어떤 데이터에 어떻게 접근하는지를 기술적으로 보장할 수 없다면, 에이전트 AI의 대규모 도입은 보안의 관점에서 러시안 룰렛에 가깝다.

OWASP GenAI 보안 프로젝트는 2025년 12월 에이전트 AI의 10대 보안 위험을 발표했다.[18] 그 중 가장 위협적인 세 가지는 다음과 같다.

- 메모리 오염(Memory Poisoning): 에이전트는 단일 세션이 아닌 장기 메모리에 의존한다. 공격자가 이 메모리에 잘못된 정보나 지시를 점진적으로 주입하면, 에이전트의 행동 패턴 자체를 변질시킬 수 있다. 기존 AI 모델은 세션이 끝나면 상태가 초기화되지만, 에이전트는 오염된 기억을 계속 보유한다.
- 도구 남용(Tool Misuse): 에이전트에게 합법적으로 부여된 도구 접근 권한을 공격자가 악용해

17) Computer World, "As AI agents go mainstream, companies lean into confidential computing for data security", Jul 21, 2025
18) "OWASP GenAI Security Project Releases Top 10 Risks and Mitigations for Agentic AI Security", Dec 9, 2025

횡적 이동(lateral movement)이나 원격 코드 실행의 경로로 활용한다. 에이전트가 기술적으로 허가된 범위 내에서 행동해도, 그 행동이 공격자의 의도를 실현할 수 있다.

- 신원 위장(Identity Spoofing): 다중 에이전트 시스템에서 공격자가 에이전트의 신원을 위장해 다른 에이전트로부터 신뢰를 획득하고, 데이터를 탈취하거나 워크플로우를 조작한다.

현대의 에이전트 AI 시스템은 단일 에이전트가 아닌 다중 에이전트 파이프라인으로 구성되는 경우가 많다. 오케스트레이터 에이전트가 하위 에이전트들에게 작업을 위임하고, 각 하위 에이전트가 또 다른 도구나 에이전트를 호출한다. 이 체인에서 핵심 질문이 생긴다. 에이전트 B는 에이전트 A의 지시를 어떻게 신뢰하는가? A가 이미 공격자에게 탈취됐다면? 기존 소프트웨어 보안 모델에서 이 신뢰 문제는 API 키나 토큰으로 처리되지만, 자율적으로 행동하는 에이전트 간의 신뢰는 훨씬 복잡한 문제다. 기밀 컴퓨팅의 원격 증명은 각 에이전트가 "나는 변조되지 않은 코드를 신뢰할 수 있는 하드웨어에서 실행 중이다"라는 것을 암호학적으로 증명하게 함으로써 이 질문에 답한다.

에이전트를 격리하기 위한 사례

앤쓰로픽의 “기밀 추론(Confidential Inference) 시스템”이라는 연구 보고서를 발표했다.[19] 이 시스템에서 모델 가중치는 모델 소유자의 키 관리 시스템(KMS)으로 암호화되어 배포된다. 추론 서버의 TEE가 시작되면 자신의 원격 증명 보고서를 KMS에 제출하고, KMS는 이를 검증한 후에만 복호화 키를 TEE 내부로 보낸다. 클라이언트는 TEE의 공개키로 입력 데이터를 암호화해 전송하고, TEE 내부에서만 복호화가 이루어진다. 결과적으로 모델 가중치와 사용자 입력 모두 TEE 외부에서는 평문으로 존재하지 않는다. 앤쓰로픽은 이를 통해 클로드 사용자에게 AI 서비스 제공업체조차 입력 데이터를 볼 수 없다는 사실을 기술적으로 보장한다.

이외에도 업계는 에이전트 AI를 위한 기밀 컴퓨팅 솔루션을 빠르게 구체화하고 있다. 안주나 시큐리티(Anjuna Security)는 에이전트를 'TEE 안에 격리된 프로세스'로 실행시키고, 런타임 시 정책을 강제하는 '에이전트 감독자' 아키텍처를 제안한다.[20] 에이전트가 어떤 데이터에 접근하고 어떤 행동을 취하는지를 하드웨어 수준에서 감시하고 제한한다. 오패크(OPAQUE)는 TEE 기반의 기밀 에이전트를 발표했는데, RAG 파이프라인의 모든 단계가 TEE 안에서 이루어져 검색된 문서부터 생성된 응답까지 전 과정이 암호화된다는 것을 강조한다.[21] 마이크로소프트 애저는 다중 테넌트 AI 에이전트 환경을

19) Anthrophic, “Confidential Inference Systems – Design principles and security risks”, Jun 2025
20) https://www.anjuna.io/
21) PR Newswire, “OPAQUE Unveils Confidential Agents with Turnkey RAG Workflows, Reinventing Trust and Security in AI”, Jun 17, 2025

위한 기밀 컴퓨팅 업데이트를 발표하며, 서로 다른 조직의 에이전트가 동일한 클라우드 인프라에서 실행되어도 각자의 데이터와 모델이 완전히 격리되는 아키텍처를 제공한다.[22)]

에이전트 시대에 기밀 컴퓨팅이 제공하는 것 – 요약

에이전트 AI와 기밀 컴퓨팅의 결합을 통해 제공되는 안전장치는 다음과 같이 정리될 수 있다.

- **격리:** 에이전트가 TEE 안에서 실행되어 호스트 운영체제나 클라우드 제공업체도 에이전트의 처리 내용에 접근 불가
- **증명:** 원격 증명을 통해 에이전트가 변조되지 않은 코드를 실행 중임을 사용자가 검증 가능
- **감사:** 에이전트의 모든 행동과 데이터 접근 이력을 TEE 내부에서 암호화 서명된 로그로 기록
- **최소 권한 강제:** 에이전트가 작업에 필요한 데이터에만 접근하도록 하드웨어 수준에서 정책 적용

이는 곧 에이전트 AI가 신뢰받기 위해 필요한 최소한의 보안 요구사항이기도 하다. 에이전트 AI가 기업과 사회의 핵심 인프라로 자리 잡는 속도에 비례해, 기밀 컴퓨팅의 중요성도 더욱 커질 것이다.

4. 기밀 컴퓨팅 생태계 동향

기밀 컴퓨팅 생태계에서 매우 중요한 역할을 하는 곳으로 기밀 컴퓨팅 컨소시엄(CCC)을 들 수 있다. CCC는 2019년 리눅스재단 산하에 설립된 비영리 컨소시엄이다. 기밀 컴퓨팅 기술의 표준화와 오픈소스 생태계 구축을 목표로 한다. 인텔, AMD, ARM, 구글, IBM, 마이크로소프트, 레드햇이 창립 멤버로 참여했고, 이후 엔비디아, 메타, 화웨이 등이 합류했다. 2024년에는 후지쯔와 틱톡이 프리미어 멤버로 가입했다. 틱톡의 가입이 다소 상징적으로 보일 수 있는데, 이는 미국 정부의 데이터 프라이버시 우려를 받는 기업이 기밀 컴퓨팅을 기술적 해답으로 우회하려는 시도로 읽을 수 있다.

CCC에서 관리하는 주요 오픈소스 프로젝트는 다음과 같다.

- 그라민(Gramine): 기존 리눅스 애플리케이션을 인텔 SGX 같은 TEE 내에서 수정하지 않고 안전하게 실행할 수 있게 해주는 경량 라이브러리 OS
- 이낙스(Enarx): 다양한 TEE 하드웨어를 추상화해 플랫폼에 무관한 기밀 컴퓨팅 개발을 가능하게 하는 프레임워크

22) Markaicode, “Securing Multi-Tenant AI Agents: Microsoft Azure's Confidential Computing Update”, Feb 22, 2026

- 베라이종(Veraison): 원격 증명 서비스 개발 도구 프레임워크
- COCONUT-SVSM: AMD SEV-SNP 환경의 보안 가상 머신 서비스 모듈
- Certifier 프레임워크: 다양한 기밀 컴퓨팅 플랫폼 간의 신뢰 관리 및 운영을 단순화하고 통합하는 프레임워크

글로벌 빅 클라우드 3사는 기밀 컴퓨팅을 핵심 차별화 요소로 삼기 시작했다. 마이크로소프트 애저는 매우 광범위한 포트폴리오를 보유하고 있다. 인텔 SGX 기반 DC 시리즈부터 TDX 기반 DCesv5, 엔비디아 H100 기반 기밀 VM까지 다양하며, 애저 컨피덴셜 레저(Ledger)와 애저 원격증명 같은 관리형 서비스도 제공한다.

구글 클라우드는 2024년 구글 클라우드 넥스트에서 AI 워크로드를 위한 기밀 컴퓨팅 확장을 발표했고, A3 컨피덴셜 VM(H100 기반)과 GKE 컨피덴셜 노드로 쿠버네티스 워크로드까지 보호 범위를 넓혔다. AWS는 니트로(Nitro) 시스템을 기반으로 기밀 컴퓨팅을 구현하며, 독립적인 외부 보안 검증을 통해 기밀 컴퓨팅 기능을 공개했다.

5. 향후 과제 및 시사점

기밀 컴퓨팅이 앞으로 풀어야 할 과제도 적지 않다.

- **사이드 채널 공격:** TEE가 메모리와 레지스터를 보호해도, 간접적 정보 유출 경로는 상시 존재한다. 실행 시간, 전력 소비, 캐시 접근 패턴 같은 부수적 정보를 분석해 비밀을 추론하여 밝히려는 시도는 계속될 것이다.
- **신뢰의 경계:** 기밀 컴퓨팅의 신뢰 사슬의 마지막에는 CPU 제조사가 있다. 즉, "CPU 제조사를 신뢰한다"는 전제가 기밀 컴퓨팅 보안의 근본 가정인 것이다. '어두운' 목적으로 제조사에 백도어가 심어질 가능성도 완전히 배제할 수 없다.
- **성능 오버헤드 와 운영 복잡성:** 기밀 VM과 기밀 컨테이너는 메모리 암호화, 격리 유지, 원격 증명 과정에 추가 컴퓨팅 자원이 필요하다. 레이턴시에 민감한 실시간 시스템에는 여전히 부담이 될 수 있다.
- **표준화:** 인텔, AMD, 엔비디아, ARM 등 각사 기술의 인터페이스와 원격 증명 방식이 다르다. 기밀 컴퓨팅 애플리케이션을 다른 플랫폼으로 이식하려면 상당한 재작업이 필요하며, CCC가 표준화를 추진하고 있지만, 완전한 표준화까지는 갈 길이 남아있다.

우리나라에서도 다양한 데이터 정책과 정부가 추진하는 AI 전략에서 기밀 컴퓨팅은 매우 중요한 의미를 가지고 있다. 이 중 몇 가지를 들어보겠다.

- **마이데이터:** 마이데이터의 핵심 과제는 데이터 결합 과정에서 프라이버시를 어떻게 보장하느냐다. 기밀 컴퓨팅은 이 문제의 기술적 해답이 될 수 있다. 각 금융기관의 데이터를 TEE 안에서 처리하고, 개인화된 금융 서비스만 결과로 제공한다면 마이데이터 플랫폼 운영자조차 개인의 원본 거래 내역을 볼 수 없는 아키텍처가 가능해진다.
- **AI와 데이터 규제:** 기밀 컴퓨팅은 원본 데이터를 그대로 사용하면서도 프라이버시를 기술적으로 보장하는 접근 방식을 제공한다. AI 학습과 추론이 TEE 안에서 이루어지면, 데이터는 노출되지 않으면서 AI 학습 효과를 그대로 얻을 수 있다.
- **공공의 클라우드 전환과 데이터 주권:** 데이터가 물리적으로 어디에 있든, TEE 안의 데이터는 클라우드 제공업체도 접근할 수 없다. 물론 이것이 모든 데이터 주권 문제를 해결하지는 않지만, 퍼블릭 클라우드에 대한 기술적 신뢰를 높일 수 있다.
- **반도체 강국으로서의 포지셔닝:** 한국은 기밀 컴퓨팅 생태계에서 독특한 위치에 있다. 삼성전자와 SK하이닉스의 메모리 반도체가 기밀 컴퓨팅 인프라의 핵심 부품이다. CPU와 GPU 레벨의 TEE 기술은 인텔, AMD, 엔비디아가 주도하지만, 이의 구동에 필수적인 메모리는 한국 기업이 공급한다. 국내 기업들이 기밀 컴퓨팅 소프트웨어 생태계에서 더 적극적인 역할을 모색할 여지가 있다. 기밀 AI 서비스, 기밀 데이터 분석 플랫폼, 의료·금융 특화 기밀 컴퓨팅 솔루션은 국내 기술력과 산업 도메인 지식이 결합될 수 있는 영역이다.

디지털 시대의 신뢰는 오랫동안 사회적 계약과 법적 규제에 의존해 왔다. "데이터를 다루는 조직을 믿어야 한다"는 전제가 모든 디지털 서비스 이용의 기반이었다. 최근 쿠팡 사태와 같은 대규모 개인정보 유출 사태는 계약 및 법적 규제에 의한 '신뢰'의 한계를 명백히 드러내는 사례로 볼 수 있다. 기밀 컴퓨팅은 이러한 한계를 기술적으로 극복할 수 있는 가능성을 열었다. 특히 인간의 감시와 통제가 점점 더 어려워지는 에이전트 AI 시대의 최소한의 안전장치로서의 기밀 컴퓨팅은 핵심 인프라 기술로 자리 잡게 될 것이다.

03 사스포칼립스(SaaSpocalypse)를 넘어 에이전트 운영 모델의 시대로

| 김영욱 SAP Product Engineering Product Expert

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

1. SaaS의 종말과 에이전트 시대의 서막

지난 20여 년간 엔터프라이즈 시장을 지배해온 소프트웨어의 형태가 근본적인 의심을 받기 시작했다. 그 정점에는 마이크로소프트 CEO 사티아 나델라의 충격적인 발언이 존재한다. 그는 2024년 말 한 인터뷰를 통해 다음과 같이 기존 소프트웨어 질서의 종말을 예고했다.

> *"전통적인 소프트웨어 애플리케이션은 에이전트 시대에 '붕괴'할 것이다. 그것들은 본질적으로 약간의 비즈니스 로직이 담긴 데이터베이스일 뿐이기 때문이다."*[23]

이 발언은 단순한 기술적 진보를 언급한 수준을 넘어선다. 지금까지 우리가 '소프트웨어'라고 정의해온 것들, 즉 사람이 직접 사용자 인터페이스에 접속하여 데이터를 입력하고 복잡한 메뉴를 클릭하며 비즈니스 로직을 수행하던 방식의 수명이 다했음을 알리는 신호였다. 나델라는 미래의 소프트웨어가 더 이상 '사람을 위한 도구'가 아니라, AI 에이전트가 데이터에 접근하고 목적을 수행하기 위한 '통로'로 격하될 것임을 예고했다.

사스포칼립스(SaaSpocalypse)의 실체: 공포가 숫자가 될 때

시장은 이 선언에 즉각적이고 냉혹하게 반응했다. 2026년 초, 나스닥의 주요 SaaS 기업들은 유례없는 주가 폭락을 경험하며 이른바 '사스포칼립스(SaaSpocalypse)'라는 신조어를 현실로 마주했다. 단 하루 만에 약 3,000억 달러의 시장 가치가 증발했으며, 이는 거시 경제의 충격이나 개별 기업의 실적 부진 때문이 아닌 혁신적인 **AI 제품의 등장**이 촉발한 결과였다.

23) BG2 Interview with Satya Nadella. Dec 12, 2024

특히 앤쓰로픽이 선보인 '코워크(Cowork)'와 같은 컴퓨터 사용 에이전트의 등장은 투자자들로 하여금 '워크플로우 대체 리스크'를 실감하게 하는 결정적인 계기가 되었다. 인간 엔지니어처럼 화면을 보고 데스크톱 앱을 직접 조작하며 다단계 업무를 수행하는 기술의 실체화는, 더 이상 기능 경쟁이 아닌 워크플로우 자체가 소프트웨어 밖으로 이동할 것임을 시사했다. 이로 인해 시장의 핵심 가정이 무너졌다. 과거 평균 39배에 달하던 미래 수익 대비 주가 배수(P/E)가 단 몇 달 만에 21배 수준으로 폭력적으로 압축되었으며, 세일즈포스, 서비스나우, 어도비, 워크데이 등 가장 공고했던 엔터프라이즈 기업들이 하루 만에 약 7%씩 급락했다. 시장은 사용자들이 더 이상 레거시 시스템의 라이선스에 비용을 지불하는 대신, AI 에이전트가 만들어 내는 실질적인 '결과'에 비용을 지불할 것이라고 예측하기 시작했다. 공매도 세력이 레거시 SaaS의 몰락에 베팅하며 2026년에만 이미 200억 달러 이상의 수익을 거두고 있다는 사실은 이러한 변화가 일시적인 현상이 아님을 나타낸다.[24)]

그런데, 이러한 하락은 개별 기업의 실적 부진 때문이 아니라, AI 에이전트가 인간의 워크플로우 자체를 대체함에 따라 지난 수십 년간 공고했던 '좌석당 과금(Seat-based pricing)' 모델이 더 이상 유지될 수 없다는 구조적 균열에 반응한 결과였다.

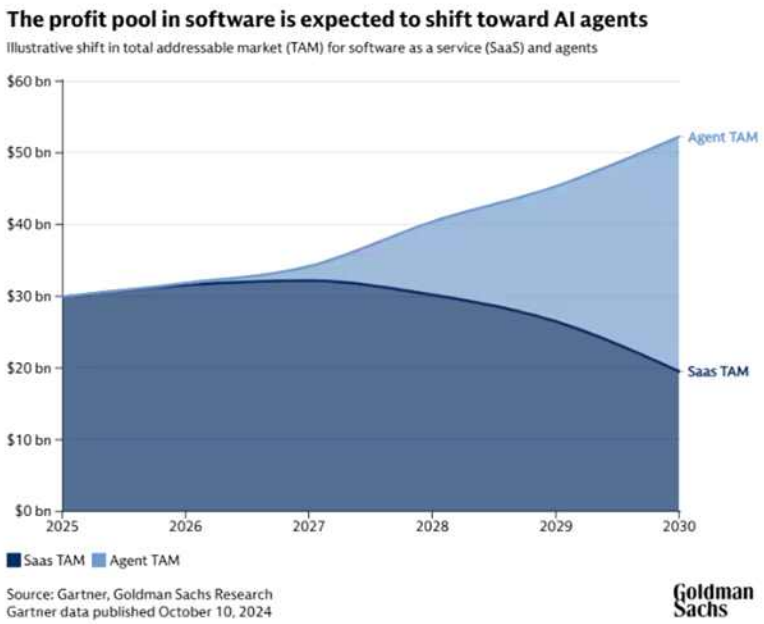


그림 10 SaaS와 에이전트의 수익시장 변화 (출처: 골드만 삭스)

24) Forbes, “$300 Billion Evaporated. The SaaS -Pocalypse Has Begun", Feb 4, 2026

도구의 변화인가, 운영 모델의 전환인가?

많은 이들이 현재의 위기를 단순히 제품에 새로운 'AI 기능'이 추가되는 교체기로 오해하곤 한다. 그러나 본 리포트는 현 상황을 단순히 제품의 전환을 넘어선 **'운영 모델의 근본적 전환'**으로 정의한다. 소프트웨어 기업들은 현재 단순한 제품 경쟁을 넘어 생존을 건 패러다임의 도박판에 서 있다. 과거의 소프트웨어가 인간의 노동을 기록하고 보조하는 '도구'였다면, 이제 소프트웨어는 스스로 사고하고 행동하며 결과를 만들어 내는 '노동' 그 자체가 되어가고 있다. 이러한 변화는 기존의 모든 비즈니스 관행을 부정하며, 단순히 사용자 인터페이스에 AI 챗봇을 다는 수준의 대응은 더 이상 유효하지 않음을 시사한다. 따라서 기업들은 AI 에이전트가 비즈니스 로직을 자율적으로 수행하고 인간은 그 결과와 보안을 감독하는 새로운 시스템으로 아키텍처 자체를 재설계하는 운영 모델의 현대화를 고민해야 한다. 이미 가치의 재배정은 시작되었으며, 수익의 원천은 더 이상 정적인 애플리케이션 라이선스가 아닌, 도구를 넘나들며 실행을 책임지는 적응형 시스템으로 이동하고 있다.

이 글에서는 AI 에이전트가 어떻게 기존 소프트웨어 가치 사슬을 파괴하고 있는지 이야기하고, 나아가 그 위에서 'AgentOps'와 '거버넌스'라는 새로운 해자를 구축하며 살아남을 기업들의 생존 전략과, 변화하는 시대에 대응하는 엔터프라이즈 소프트웨어의 미래 로드맵에 대해서 이야기하려 한다.

2. AI 에이전트가 파괴하는 기존 가치 사슬

전통적인 소프트웨어 시장을 지배해온 핵심 가치는 인간이 특정 목적을 달성하기 위해 거쳐야 하는 사용자 인터페이스의 편의성과 효율성에 있었다. 그러나 AI 에이전트의 등장은 이러한 인터페이스의 존재 이유 자체를 희석시키고 있다. 과거의 워크플로우는 사람이 소프트웨어의 메뉴를 탐색하고 데이터를 입력하는 물리적 행위를 전제로 설계되었으나, 이제 AI 에이전트는 UI를 거치지 않고 데이터 레이어에서 직접 정보를 추출하고 논리적 단계를 자율적으로 수행한다. 이러한 변화는 소프트웨어를 단순한 **'기능의 집합체'에서 '실행의 통로'**로 격하시키며, 결과적으로 개별 애플리케이션 내에 머물던 사용자 경험의 가치를 급격히 소멸시킨다. 연구, 분석, 초안 작성 및 협업과 같은 복잡한 작업들이 단일 애플리케이션의 경계를 넘어 시스템 전반에서 자율적으로 실행됨에 따라, 특정 소프트웨어가 점유하던 워크플로우의 해자는 빠르게 무너지고 있다.

좌석당 과금(Seat-based) 모델의 경제적 붕괴

SaaS 산업의 성장을 견인해 온 가장 공고한 비즈니스 모델인 '좌석당 과금' 방식은 현재 실존적 위기에 직면해 있다. 이 모델은 소프트웨어를 사용하는 인간의 머릿수가 곧 매출로 직결된다는 가정하에 설계되었으나, AI 에이전트가 인간의 노동력을 직접적으로 대체하거나 보조하면서 이 등식은 성립 불가능한 상태에 이르렀다. 기업들은 더 이상 수백 명의 직원을 위해 라이선스를 구매할 필요를 느끼지 못하며, 대신 단 소수의 관리자와 이들을 보조하는 고성능 AI 에이전트 체제로 전환하고 있다. 시장은 이러한 변화를 라이선스 판매량의 감소뿐만 아니라 수익 구조의 근본적인 희석으로 받아들이고 있다. 인원수 기반의 예측 가능성은 사라지고, 소프트웨어가 제공하는 결과물의 가치에 비용을 지불하는 성과 기반(Outcome-based) 또는 소비 기반(Consumption-based) 모델로의 강제적인 전환이 요구되고 있다.

바이브 코딩과 기능적 해자의 상실

과거 대형 SaaS 플랫폼들이 구축한 강력한 방어선 중 하나는 복잡한 기능을 구현하고 유지보수 하는데 드는 막대한 개발 비용과 시간이었다. 그러나 최근 부상한 '바이브 코딩'과 같은 AI 기반 개발 패러다임은 이러한 기술적 장벽을 허물고 있다. 생성형 AI는 숙련된 엔지니어링 조직 없이도 기존 소프트웨어의 핵심 기능을 순식간에 복제하거나 새로운 맞춤형 기능을 구현해 낼 수 있는 환경을 제공한다. 이는 수십 년간 쌓아온 복잡한 기능적 우위가 더 이상 절대적인 경쟁력이 될 수 없음을 의미한다. 투자자들은 범용적인 기능을 제공하는 수평적 SaaS 기업들이 이러한 AI 기반의 기능 복제와 대안 플랫폼의 출현에 가장 취약할 것으로 예측하며, 기존의 기능 중심 경쟁력이 아닌 데이터와 실행의 독점력을 가진 새로운 형태의 플랫폼으로 자본을 이동시키고 있다.

3. AgentOps와 거버넌스의 부상

AI 에이전트가 소프트웨어 개발과 실행의 중심축으로 이동함에 따라, 이들의 '사고 과정'을 투명하게 관리해야 할 필요성이 급격히 대두되고 있다. 특히 숙련된 개발자조차 AI가 생성한 코드의 양과 속도를 물리적으로 검토하기 어려워지면서, AI의 내부 추론 과정을 가시화하는 기술은 선택이 아닌 필수적인 인프라가 되었다. 이에 대응하여 전 GitHub CEO 토마스 돔케가 설립한 스타트업 'entire.io'가 최근 3억 달러의 가치를 인정받으며 화려하게 등장한 이유가 바로 여기 있다. 이들이 내놓은 오픈소스 도구 'Checkpoints'는 AI 에이전트가 코드를 작성하거나 과업을 수행하는 동안 거치는 단계별 논리 전개인 '사고의 사슬(Chain of Thought)'을 상세히 기록한다. 이러한 기술적 장치는 블랙박스 상태로 존재하는 AI의 추론 결과를 인간이 상위 수준에서 감독할 수 있도록 '설명 가능한 AI' 인터페이스 역할을

수행하며, 시스템의 오작동이나 예기치 못한 결과에 대해 즉각적인 원인 분석이 가능하게 한다.

이런 이유로 과거 소프트웨어 개발의 표준이었던 데브옵스와 머신러닝 모델 관리를 위한 MLOps를 넘어, 이제는 AI 에이전트의 전체 생애주기를 관리하는 'AgentOps'가 차세대 엔터프라이즈 인프라의 핵심 분야로 부상하고 있다. AgentOps는 개별 에이전트의 행동을 실시간으로 관찰하고 모니터링하며, 복잡한 멀티 에이전트 환경에서 발생할 수 있는 충돌을 방지하고 성능을 최적화하는 역할을 수행한다. 대기업들은 AI 에이전트의 '통제 불능' 가능성을 심각한 리스크로 인지하고 있으며, 이를 관리하기 위한 거버넌스 시장은 향후 소프트웨어 공급망에서 가장 치열한 격전지가 될 것으로 예측된다. 이는 단순한 모니터링을 넘어 AI 에이전트가 기업의 정책과 보안 규정을 준수하며 작동하는지 보증하는 신뢰 레이어의 구축을 의미한다.

현재 AI 시장은 특정 모델에 최적화된 독점적 에이전트 관리 도구와, 다양한 제조사의 에이전트를 아우르는 범용 플랫폼 간의 주도권 싸움이 벌어지고 있다. 앤쓰로픽이나 구글과 같은 주요 모델 개발사들은 자사 에이전트의 성능 극대화에 집중하는 반면, 'Entire'와 같은 신흥 플랫폼 기업들은 벤더 종속 문제를 해결하기 위해 여러 에이전트를 통합 관리할 수 있는 중립적인 표준을 제안하고 있다. 기업 고객은 대시보드 하나를 통해 여러 벤더의 에이전트를 통제하길 원하며, 이러한 통합 대시보드는 향후 기업 데이터가 어디에 거주하고 어떻게 제어되는지를 결정하는 '슈퍼에이전트'의 조종석 역할을 하게 될 것이다. 결국 AI 에이전트 시대의 최종 승부처는 단순히 똑똑한 모델을 만드는 것을 넘어, 분산된 지능들을 하나의 질서 아래 연결하고 통제하는 범용 거버넌스 체계를 누가 선점하느냐에 달려 있다.

4. SaaS 거인들이 구축하는 새로운 해자

AI 에이전트가 단일 과업 수행에서 뛰어난 지능을 발휘한다 할지라도, 기업 내부의 파편화되고 복잡한 비즈니스 로직을 온전히 이해하는 것은 별개의 문제이다. 엔터프라이즈 소프트웨어 시장의 기존 강자들은 수십 년간 기업의 핵심 업무 프로세스를 디지털화하고 최적화하며 축적해 온 정교한 '프로세스 맵'을 소유하고 있다. 이러한 워크플로우 데이터는 단순한 기록을 넘어 AI 에이전트가 목적지에 도달하기 위해 반드시 참조해야 하는 필수적인 내비게이션 역할을 수행한다. 신규 AI 스타트업이 뛰어난 추론 모델을 제시하더라도, 기업 각 부서 간의 승인 절차, 규제 준수 요건, 그리고 예외 처리 로직이 얽혀 있는 엔터프라이즈의 실무 지도를 단기간에 복제하기란 사실상 불가능에 가깝다. 결국 기존 SaaS 거인들은 자신들의 시스템을 단순한 도구에서 AI 에이전트의 실행 경로를 지휘하고 통제하는 '지능형 워크플로우 허브'로 재정의하며 강력한 방어선을 구축하고 있다.

시스템 오브 레코드(System of Record)의 방어와 데이터, 에이전트 관리 주도권

기업 가치의 근간이 되는 핵심 데이터가 거주하는 '시스템 오브 레코드(System of Record)'는 AI 시대에도 여전히 플랫폼 기업들의 가장 강력한 해자로 작용한다. 마이크로소프트와 세일즈포스 같은 기업들이 보유한 CRM, SAP의 전문 분야인 ERP 데이터는 AI 에이전트가 유의미한 결과물을 도출하기 위해 반드시 접근해야 하는 '연료'와 같다. 이들은 자사 플랫폼 내부에 보안 및 데이터 인프라를 통째로 내재화하여, AI 에이전트가 안전하게 데이터를 검색하고 자산에 접근할 수 있는 통제권을 행사한다. 특히 외부의 독립적인 에이전트가 기업의 민감한 데이터에 접근하려 할 때 발생하는 보안 리스크는 기존 플랫폼 사들에게 강력한 거버넌스 명분을 제공한다. 이 과정에서 기득권 기업들은 자사 에이전트에게는 데이터 접근의 우선권을 부여하거나, 외부 에이전트의 활동을 제한하는 방식으로 데이터 주권을 강화하며 슈퍼에이전트 시대의 최종 대시보드 자리를 선점하려 하고 있다.

최근의 시장 분석 데이터에 따르면, 전통적인 소프트웨어 강자들은 AI 에이전트 시대의 새로운 운영체제인 '에이전트 관리 대시보드' 영역에서 신흥 AI 기업들보다 압도적으로 유리한 고지를 점하고 있다. 마이크로소프트, AWS, 구글, SAP, 세일즈포스 등 주요 기업들은 이미 에이전트 관리 대시보드 제품을 정식 출시하며 거버넌스 레이어를 선점했다. 반면, 강력한 모델 지능을 보유한 앤쓰로픽은 아직 관리 대시보드를 출시하지 못했으며, 오픈AI 역시 프리뷰 단계에 머물러 있는 것으로 나타났다. 이는 기업 고객이 요구하는 복잡한 에이전트 생애주기 관리와 정책 준수 영역에서 기존 기득권 기업들이 가진 시스템 운영 노하우와 신뢰도가 핵심적인 경쟁 우위로 작용하고 있음을 증명한다.

또한, 모든 주요 SaaS 기업들은 고객이 직접 맞춤형 에이전트를 개발할 수 있는 '에이전트 빌더' 도구를 이미 출시하여 자사 생태계 내에 에이전트들을 통합시키고 있다. 스노우플레이크나 데이터브릭스와 같은 데이터 전문 기업들조차 브라우저 기반 에이전트 개발보다는 에이전트 빌더와 관리 대시보드 구축에 우선순위를 두는 것은, **에이전트의 '지능' 자체보다 그 지능을 자사 데이터 위에서 어떻게 '관리'하고 '통제'하느냐를 미래 가치의 본질**로 보고 있기 때문이다. 결과적으로 SaaS 거인들은 지능형 모델을 직접 만드는 경쟁을 넘어, 분산된 수많은 에이전트의 행동을 감독하고 오케스트레이션하는 상위 제어 시스템으로서의 지위를 공고히 함으로써 에이전트 시대의 새로운 생존 공식을 완성해 가고 있다.

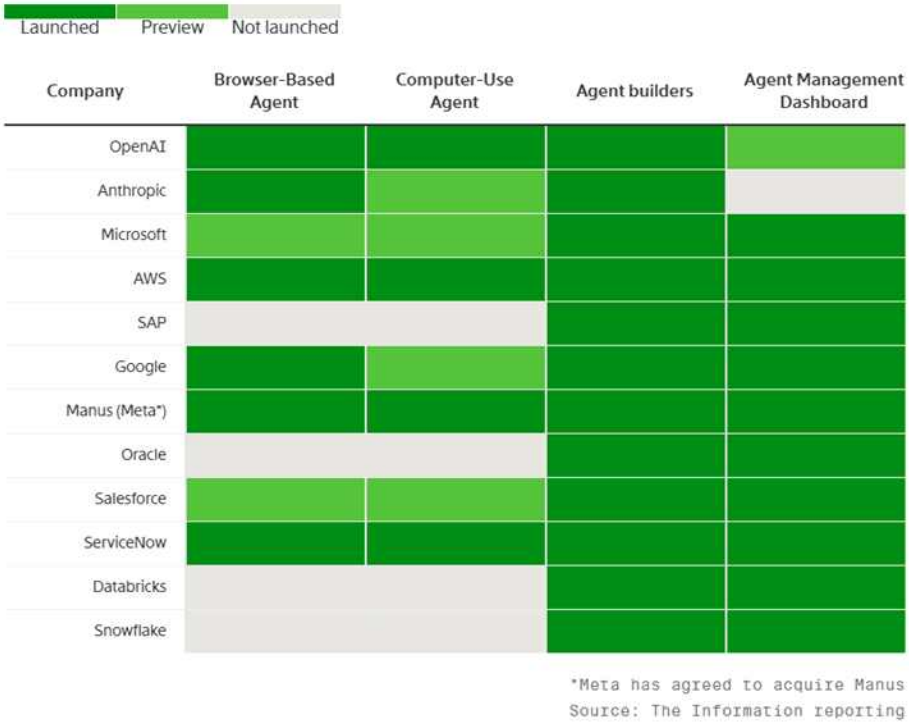


그림 11 SaaS 대표기업과 AI기업의 에이전트 제품 출시 현황(출처: The Information)

기술적 복잡성과 전환 비용의 심화

반도체 설계 소프트웨어 분야의 케이던스 사례는 고도의 수학적 복잡성을 갖춘 기술력이 어떻게 AI 시대에 난공불락의 해자가 되는지를 극명하게 보여준다. 반도체 칩 설계와 같은 초정밀 공학 분야는 단순한 자연어 이해를 넘어선 극도로 복잡한 시뮬레이션과 알고리즘을 요구하며, 이는 로켓 과학에 비견될 만큼 높은 기술적 진입 장벽을 형성하고 있다. AI 에이전트가 아무리 발전하더라도 수십 년간 검증된 이러한 전문 도메인 로직을 단번에 대체하는 것은 기업 입장에서 감당할 수 없는 리스크를 초래한다. 대기업이 핵심 시스템을 도입하고 최적화하는 데 통상 1~2년의 시간이 소요된다는 점을 고려할 때, 기업 특유의 맞춤형 워크플로우가 녹아 있는 기존 시스템을 포기하고 검증되지 않은 AI 스타트업의 도구로 전환하는 것은 파격적인 비용과 조직적 고통을 수반한다. 이러한 '전환 비용'은 AI 시대에도 변하지 않는 경제적 장벽이며, 기존 강자들은 자신들의 복잡한 로직을 AI 에이전트가 더 효율적으로 활용할 수 있게 돕는 '조력자' 역할을 자처하며 고객과의 결속력을 더욱 강화하고 있다.

5. 운영 모델 현대화

AI 에이전트가 촉발한 산업 전반의 대전환기는 기업 리더십의 성격마저 근본적으로 변화시키고 있다. 최근 워크데이가 관리형 리더를 대신하여 공동 창립자이자 과거 전성기를 이끌었던 전 CEO 아닐 부스리(Aneel Bhusri)를 복귀시킨 사례는 이러한 위기감을 극명하게 보여준다. 부스리는 복귀 선언에서 AI가 SaaS로의 전환보다 더 큰 변화이며, 이 시점이 차세대 시장 리더를 정의할 가장 중요한 순간임을 강조했다. 이는 단순히 경영 효율성을 높이는 차원을 넘어, 기업의 DNA 자체를 AI 네이티브로 재설계하기 위해 도메인에 대한 깊은 통찰과 과감한 결단력을 가진 창업자형 리더가 필요하다는 시장의 요구가 반영된 결과이다. 기업들은 이제 AI를 단순한 보조 도구가 아닌, 핵심 운영 모델의 중추로 삼기 위해 리더십의 세대교체를 통한 강제적인 피벗을 감행하고 있다.

성과 측정의 재정의: 인력 규모에서 AI 증폭력으로의 전환

운영 모델의 현대화는 조직의 내부 성과 측정 방식에서도 혁신적인 변화를 일으키고 있다. 아마존은 내부 시스템인 '클래리티(Clarity)'를 통해 직원들의 AI 도구 사용 빈도와 숙련도를 정밀하게 추적하며, 이를 인사 고과와 승진의 핵심 지표로 활용하기 시작했다.[25] 특히 승진 평가 시 직원이 AI를 활용해 어떻게 '적은 인력으로 더 많은 성과'를 냈는지, 그리고 조직의 규모가 아닌 AI를 통한 '노동의 증폭(Force multiply)'을 얼마나 달성했는지를 구체적인 사례와 함께 입증하도록 요구하고 있다. 이는 과거의 운영 모델이 관리하는 인원수나 조직의 크기를 권력과 성과의 상징으로 여겼던 것과 달리, 새로운 시대에는 한 명의 인간이 AI 에이전트를 조율하여 얼마나 극대화된 임팩트를 창출할 수 있는가가 기업의 핵심 가치가 되었음을 의미한다.

과금 모델의 진화: 좌석 기반에서 워크로드 및 성과 기반으로의 연착륙

사스포칼립스의 핵심 원인 중 하나로 지적된 '좌석당 과금' 모델의 한계를 극복하기 위해, 선도적인 기업들은 이미 수익 구조의 대대적인 개편에 착수했다. 케이던스는 사용자 수에 의존하는 대신 실제 처리되는 '작업량'을 기준으로 과금함으로써 AI로 인한 인력 감소 리스크를 오히려 성장의 기회로 반전시켰다. 또한 스노우플레이크와 같은 데이터 플랫폼 기업들은 에이전트가 실행하는 쿼리나 데이터 처리량에 비례하여 비용을 청구하는 소비 기반 모델을 강화하고 있다. 이러한 변화는 고객이 소프트웨어의 단순한 점유가 아닌 실질적인 '결과'에 비용을 지불하게 함으로써, 벤더와 고객 간의 가치

25) The Information, "How Amazon Tracks Employee AI Usage—and Measures Results", Feb 19, 2026

교환 방식을 보다 합리적이고 지속 가능하게 재구축하고 있다. 결국 AI 시대의 운영 모델 현대화는 기술의 도입을 넘어, **노동의 정의와 성과의 측정, 그리고 수익의 창출 방식까지 아우르는 전방위적인 비즈니스 혁신을 요구**하고 있다.

6. 마무리: SaaS 소프트웨어의 생존 로드맵

2026년 현재, 소프트웨어 산업이 목도하고 있는 거대한 가치 재편은 단순한 제품 경쟁을 넘어선 거시적 담론을 형성하고 있다. 과거의 자본 배정이 특정 기능을 수행하는 'AI 도구 구매'에 집중되었다면, 이제 기업들은 자신의 비즈니스 운영 모델 자체를 AI 시대에 맞게 재정의하는 '운영 모델 현대화'에 막대한 자본을 투입하고 있다. 이는 정적인 애플리케이션 라이선스에 대한 지불을 줄이고, 도구를 넘나들며 실행을 책임지는 적응형 시스템과 그 결과물에 더 많은 가치를 배정하는 흐름으로 실체화되고 있다. 결국 차세대 엔터프라이즈 시장에서 생존할 기업은 단순히 지능형 기능을 제공하는 곳이 아니라, 기업의 아키텍처를 근본적으로 개편하여 AI 에이전트가 안전하고 효율적으로 노동을 수행할 수 있는 '환경'을 제공하는 플랫폼이 될 것이다.

지능과 업무 규칙의 결합: 라스트 마일(Last Mile)의 승부

AI 모델의 추론 능력만으로는 엔터프라이즈의 복잡한 업무가 완결될 수 없다는 사실이 명확해지고 있다. 진정한 승부처는 AI의 자율적 지능과 기업이 보유한 엄격한 '비즈니스 룰'을 결합하여, 실제 현장에서 실수를 최소화하고 예측 가능한 성과를 도출하는 '라스트 마일'의 구현에 있다. 에이전트가 생성한 결과물이 보안 규정과 컴플라이언스를 준수하는지 보증하고, 기존의 시스템 오브 레코드와 유기적으로 연결되어 실행의 완결성을 높이는 역량이 기업용 AI의 핵심 과제이다. 이러한 라스트 마일을 선점하는 기업은 단순한 소프트웨어 공급자를 넘어, 기업 의사 결정 수익률(ROI of Decision-making)을 극대화하는 비즈니스 파트너로서의 지위를 공고히 하게 될 것이다.

슈퍼에이전트 대시보드를 향한 플랫폼 통합 전쟁

미래의 소프트웨어 시장은 **수많은 에이전트와 워크플로우를 한곳에서 조율하는 '슈퍼에이전트 대시보드'를 선점하기 위한 거대 플랫폼 간의 최종 통합 전쟁으로 향할 전망**이다. 앤쓰로픽과 오픈AI가 '지능'의 우위를 바탕으로 침투하는 가운데, 마이크로소프트, SAP, 세일즈포스 등 기존 거인들은 '거버넌스'와 '관리 체계'를 방어막 삼아 주도권 사수에 나서고 있다. 이 과정에서 독립적인 에이전트

관리 플랫폼인 AgentOps과 데이터 인프라 기업들은 기득권 플랫폼과 전략적 제휴를 맺거나 합병되는 수순을 밟게 될 것이며, 결국 사용자의 책상 위에는 파편화된 앱들이 아닌 통합된 지능형 대시보드만이 남게 될 것이다. 사스포칼립스는 종말이 아닌, 더 지능적이고 자율적인 소프트웨어 생태계로 진화하기 위한 통증이며, 이 거대한 판갈이 끝에 살아남은 기업들이 2026년 이후의 새로운 디지털 경제를 정의하게 될 것으로 전망한다.

04 2026년 클라우드 솔루션 리포트: 네트워크 및 엣지

| 정채상 메가존클라우드 수석 엔지니어

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

들어가며: 지능형 분산 인프라로의 대전환과 엣지 컴퓨팅의 진화

전 세계적인 인프라 환경은 중앙 집중식 클라우드 컴퓨팅의 시대를 지나 지능형 분산 인프라로의 거대한 전환을 맞이하고 있다. 2026년에 이르러 네트워크는 단순히 데이터를 전달하는 통로를 넘어 그 자체로 연산하고 보안을 집행하며 AI의 지능을 사용자에게 가장 가까운 곳에서 구현하는 핵심적인 비즈니스 차별화 요소로 자리 잡았다. 과거에는 관찰 가능성이나 데이터의 흐름을 파악하는 것이 우선이었다면, 이제는 그 흐름이 발생하는 지점에서 즉각적인 의사 결정이 이루어지는 엣지 중심의 사고가 필수적이다.

이러한 변화의 기저에는 폭발적으로 증가하는 데이터량과 이를 실시간으로 처리해야 하는 AI 에이전트의 등장이 자리 잡고 있다. 2024년까지 기업들이 대규모 언어 모델의 학습에 막대한 예산을 투입했다면, 2026년은 학습된 모델이 실제 비즈니스 가치를 창출하는 추론의 시대로 진입했다. 이 과정에서 발생하는 지연 시간과 막대한 데이터 전송 비용은 기존의 중앙 집중식 모델로는 감당할 수 없는 수준에 이르렀으며, 이는 자연스럽게 연산의 중심을 엣지로 이동시키는 동력이 되었다.

시장 지표가 이 흐름을 뒷받침한다. CDN 시장은 2025년 기준 270억~330억 달러[26], 엣지 컴퓨팅 글로벌 지출은 2,610억 달러로 집계되며, 엣지 AI 세그먼트만 따로 떼어보면 2025년 256억 5천만 달러에서 2034년 1,430억 달러로의 성장이 전망된다.[27] 스트리밍이 글로벌 인터넷 트래픽의 80% 이상을 차지하고 4K 프리미엄 구독이 전체의 35%를 넘어서면서, 엣지 인프라에 대한 수요는 이미 구조적 확대 국면에 진입했다.

26) https://finance.yahoo.com/news/content-delivery-network-cdn-global-090200498.html
27) https://bayelsawatch.com/edge-computing-statistics/

2024년까지의 기업들이 대규모 언어 모델(LLM)의 학습에 집중했다면, 2026년은 학습된 모델이 실제 비즈니스 가치를 창출하는 추론(Inference)의 시대이고, 이 과정에서 다음의 요소들이 엣지 전환을 이끌고 있다.

- AI 추론의 엣지 이동: 전체 컴퓨팅 워크로드의 약 2/3가 추론에 집중됨에 따라, 지연 시간과 데이터 전송 비용을 절감하기 위해 연산의 중심이 엣지로 이동하고 있다. IDC는 2027년까지 CIO의 80%가 엣지 기반 AI 서비스를 채택할 것으로 예측한다.
- 데이터 주권과 규제 대응: 120개국 이상의 데이터 보호법과 2026년 8월 전면 시행되는 EU AI Act에 따라, 데이터 지역화(Data Localization)는 기술적 선택을 넘어 법적 생존의 문제가 되고 있다.
- 비용 및 에너지 최적화: 하이브리드 엣지-클라우드 아키텍처는 순수 클라우드 환경 대비 비용을 최대 80%, 에너지를 75%까지 절감할 수 있는 경제적 해법으로 부상하고 있다.[28)]

본 보고서는 아카마이(Akamai), 클라우드플레어(Cloudflare), 패스트리(Fastly)와 같은 전문 엣지 벤더들과 아마존 웹 서비스, 마이크로소프트 애저, 구글 클라우드와 같은 대형 클라우드 서비스 제공업체들의 솔루션을 분석한다. 각 솔루션이 해결하고자 하는 강점과 실제 사용자들이 운영 과정에서 직면하는 문제들, 그리고 AI라는 거대한 화두에 대응하는 각 사의 전략을 심층적으로 다룬다. 이를 통해 스타트업부터 엔터프라이즈에 이르기까지 CIO와 CTO들이 2026년의 기술적 흐름을 명확히 파악하고 조직의 미래 경쟁력을 확보할 수 있는 제언을 포함한다.

2026년 이슈: 인텔리전트 네트워크와 엣지의 성숙

2026년의 네트워크 및 엣지 기술 지형은 단순히 속도와 대역폭을 겨루던 과거의 선형적 경쟁을 넘어섰고, 이제 인프라는 AI 에이전트를 비즈니스의 동료로 받아들여야 하는 시점에 도달했으며, 이에 따라 복합적인 과제들을 해결해야 하는 성숙기에 진입했다. 아래에 올해 시장을 관통하는 다섯 가지 핵심 이슈를 통해 변화의 본질을 짚어본다.

AI 추론: 엣지의 킬러 유스케이스로 부상

인공지능 예산의 흐름이 모델 학습에서 추론으로 완전히 역전되었다. 딜로이트에 따르면 추론 워크로드가 2026년 전체 컴퓨트의 약 3분의 2를 차지하며, IDC는 2027년까지 CIO의 80%가 클라우드 제공자의 엣지 서비스를 AI 추론에 활용할 것으로 예측한다. 엣지 기반 추론은 중앙 집중식

28) https://www.infoworld.com/article/4117620/edge-ai-the-future-of-ai-inference-is-smarter-local-compute.html

클라우드 API 호출 대비 약 90%의 비용 절감 효과를 제공하며, 이 경제성이 대규모 서비스 운영을 가능케 하는 핵심 동력이다.

모델의 양상도 달라지고 있다. 가트너는 2028년까지 엣지 배포의 60%가 예측형과 생성형을 결합한 복합 AI를 활용할 것으로 전망하며[29], 2027년까지 태스크 특화 소형 모델(SLM)이 범용 LLM 대비 3배 더 많이 사용될 것으로 예측하고 있으며, 이를 뒷받침하는 추론 최적화 칩 시장은 2026년 500억 달러 이상으로 성장이 예상된다.[30]

웹어셈블리(Wasm): 프로덕션 레벨의 표준 런타임 안착

브라우저 내 최적화 기술에 머물렀던 Wasm이 2026년 엣지 컴퓨팅의 표준 런타임으로 자리 잡았다. 웹어셈블리 3.0이 2025년 릴리스 된 데 이어, WASI 1.0 최종 사양이 2026년 중~후반 확정을 앞두고 있다. Wasm은 언어 비종속성, 네이티브급 실행 속도, 샌드박스 기반 보안 모델을 통해 멀티 테넌트 환경의 안정성을 확보했으며, 현재 러스트, 자바스크립트, Go, 파이썬, C/C++ 등 주요 언어의 프로덕션급 컴파일러가 존재한다.

핵심 이점은 경량성과 속도다. Wasm 모듈은 50KB 미만으로 수백 MB의 컨테이너 이미지 대비 극소형이며, 콜드 스타트는 0.5ms 이하로 기존 컨테이너 대비 100배 빠르다. 아카마이가 2025년 12월 Wasm 서버리스 기업 페미온(Fermyon)을 인수한 것은 이 분야의 전략적 중요성을 상징적으로 보여준다.

데이터 주권: 성능을 넘어선 도입 1순위 동기

STL 파트너스의 2025년 조사에서 데이터 지역화가 저지연 유스케이스를 제치고 엣지 도입의 1순위 동기로 부상했다.[31] 120개국 이상이 데이터 보호법을 시행 중이고 24개국이 추가 입법을 진행하고 있으며, 71%의 조직이 국경 간 데이터 이전 컴플라이언스를 최대 규제 과제로 꼽고 있다.

규제의 강도도 급격히 확대되고 있다. EU 데이터 법(2025년 9월 발효)은 주권을 산업 및 비개인 데이터까지 확장했고, EU AI Act는 2026년 8월 전면 시행되며 위반 시 글로벌 매출의 7% 벌금이 부과되는데, GDPR 시행 이후 누적 벌금은 67억 유로 이상이다. 기업들은 규제 준수를 목적으로 엣지

29) https://zededa.com/gartner-edge-computing-hype-cycle/
30) https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2026/compute-power-ai.html
31) https://www.edgeir.com/2025-marks-a-shift-data-sovereignty-and-ai-drive-the-next-phase-of-edge-deployment-20251116

인프라를 최우선 고려하며, Nutanix가 정의한 소버린 엣지(Sovereign Edge) — 추론이 데이터 가까이에서 실행되면서도 통제권과 규제 경계가 유지되는 아키텍처 — 가 이 흐름을 대표한다.

전략적 멀티CDN: 방어 수단에서 아키텍처적 필수 요소로

장애 대비용 백업이었던 멀티CDN이 2026년 최적 경로를 실시간으로 선택하는 전략적 아키텍처로 진화했다. 2025년 클라우드플레어[32]와 애저 프런트 도어의 주요 장애[33] 는 중앙 집중화된 서비스의 리스크를 여실히 드러냈고, 기업들은 AI 기반 동적 라우팅과 자동화된 페일오버를 갖춘 상시 멀티CDN 체제로 전환하고 있다. 지능형 트래픽 스티어링으로 이그레스 30~40% 절감이 가능하며, 지오펜싱 기반 라우팅으로 컴플라이언스도 확보할 수 있다. 다만 캐시 효율 저하, 벤더 간 기능 비대칭, 옵저버빌리티 복잡성 등의 과제가 남아있으며, IO River, CD네트웍스, Vercara 등 오케스트레이션 플랫폼이 이를 해결하기 위해 부상하고 있다.

시장 구조 재편: 플랫폼의 승리와 순수 CDN의 상품화

단순 콘텐츠 전송은 더 이상 차별화가 불가능한 일반 상품이 되었다. 대형 미디어 기업들이 자체 전송망을 구축하고 하이퍼스케일러들이 CDN을 클라우드 번들 상품으로 제공하면서, 기존 벤더들은 생존을 위해 보안과 컴퓨팅을 통합한 '플랫폼 서비스'로 완전히 탈바꿈하고 있다. 아카마이의 전체 매출에서 전송 부문 비중이 급격히 축소되고 보안과 클라우드 컴퓨팅이 주류로 올라선 것은 이러한 시장의 지각변동을 여실히 보여준다. 2026년 인프라 시장의 승자는 단순한 전송 속도가 아니라, 통합된 보안 거버넌스와 강력한 엣지 연산 능력을 한 번에 제공하는 연결성 플랫폼 벤더들이 차지하고 있다.

주요 벤더별 솔루션 심층 분석

2026년 시장을 주도하는 벤더들은 각기 다른 철학을 바탕으로 엣지와 네트워크 솔루션을 제공하고 있다. 독립적인 엣지 전문 벤더들은 속도와 개발자 경험에 집중하는 반면, 대형 클라우드 사업자들은 통합된 관리 체계와 방대한 자원을 강점으로 내세운다.

아카마이: CDN 원조가 분산 클라우드 플랫폼으로

32) https://blog.cloudflare.com/18-november-2025-outage/
33) https://www.thousandeyes.com/blog/microsoft-azure-front-door-outage-analysis-october-29-2025

아카마이는 세계에서 가장 널리 분산된 플랫폼인 아카마이 커넥티드 클라우드(Akamai Connected Cloud)를 통해 엔터프라이즈 환경에서의 안정성과 보안을 책임지고 있다. 135개국 이상, 700개 이상의 도시에 걸친 4,200개 이상의 PoP과 325,000대 이상의 서버 인프라를 보유하며, 포춘 500 대기업의 56%가 고객이다.

핵심 강점: 압도적 규모와 'Gecko'의 결합

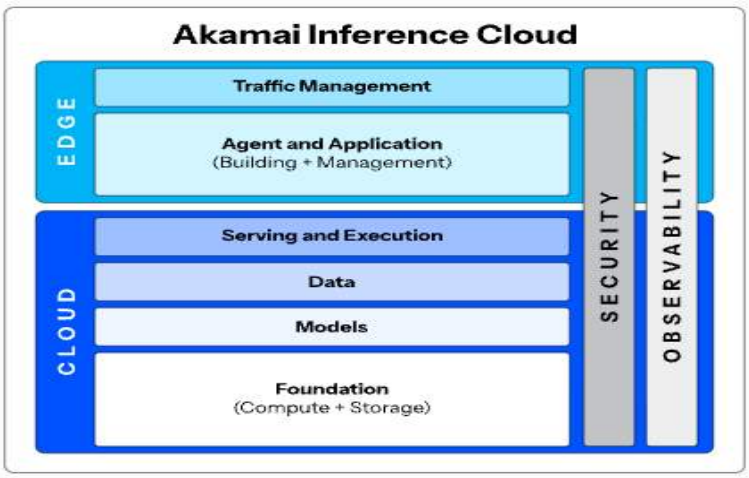
아카마이의 핵심 전략 베팅은 Gecko(Generalized Edge Compute) 플랫폼이다. 4,200개 이상의 엣지 PoP에 범용 클라우드 컴퓨팅을 내장하여, AI 추론, 멀티플레이어 게임, 스트리밍, IoT 등에 초저지연 컴퓨트를 제공한다. Phase 1으로 약 100개 도시에 VM 배포를 완료했으며 향후 수백 도시로 확대 예정이다. 2022년 9억 달러에 인수한 리노드(Linode)의 컴퓨팅자원을 전 세계 엣지 로케이션에 결합함으로써, 개발자들이 중앙 클라우드와 동일한 환경을 사용자 근처에서 운영할 수 있도록 돕는다.

서버리스 측면에서는 구글 V8 엔진 기반의 콜드 스타트 5ms 미만인 엣지워커(EdgeWorkers)와 분산 키-값 스토어인 EdgeKV를 제공한다. 2025년 페미온 인수로 웹어셈블리 기반 서버리스가 추가되어 콜드 스타트 0.5ms 미만, 러스트, Go, 파이썬, .NET 등 다중 언어를 지원한다.

보안 역시 핵심 강점이다. 프로렉식(Prolexic) 서비스는 200 Tbps 이상의 DDoS 차단 용량을 보유하고 있으며, 2026년에는 AI 기반의 자동화된 방어 로직을 탑재하여 알려지지 않은 공격 기법에 대해서도 즉각적인 대응이 가능하다. 보안은 현재 아카마이의 최대 매출 세그먼트로 부상했으며, 전통적 CDN 딜리버리 매출은 점차 감소하는 추세다.

AI 시대의 비전: AI 트래픽 컨트롤러와 인퍼런스 클라우드

아카마이는 자사의 네트워크를 단순한 전달 통로가 아닌 AI 트래픽 컨트롤러로 정의한다. 2025년 10월 NVIDIA GTC에서 발표한 아카마이 인퍼런스 클라우드(Akamai Inference Cloud)는 NVIDIA RTX PRO 6000 블랙웰 에디션 GPU 탑재 서버를 17개 도시에 배포하고 20개 이상으로 확대 중이다.



34)

그림 12 아카마이 인퍼런스 클라우드 플랫폼

최근에는 비자와 파트너십을 통해 AI 쇼핑 에이전트 인증 및 에이전틱 커머스의 사기 방지 프로토콜을 개발하는 등, '지능형 에이전트'가 주도하는 미래 경제의 보안 거버넌스를 선점하겠다는 전략을 펼치고 있다. 엣지에서 GPU 컴퓨팅 자원을 조율하고 AI 에이전트가 가장 효율적으로 데이터에 접근할 수 있도록 경로를 최적화하며, 엣지 기반 매니지드 데이터베이스 서비스를 강화하여 추론에 필요한 원천 데이터를 사용자 근처에서 안전하게 관리할 수 있도록 지원한다.

사용자의 페인 포인트와 한계: 높은 장벽과 복잡한 경험

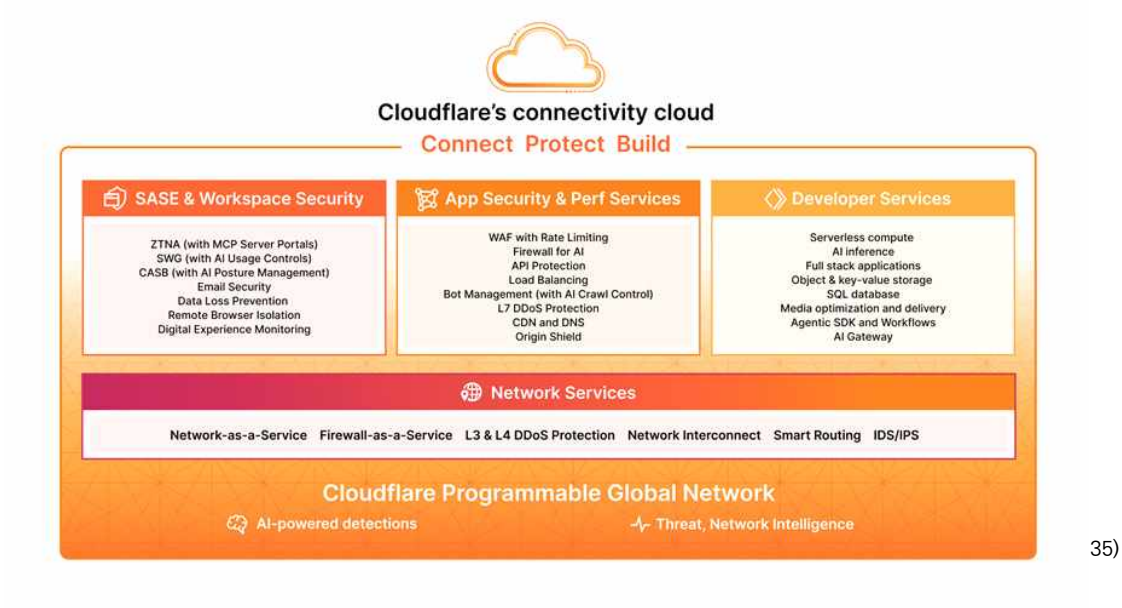
강력한 성능에도 불구하고 아카마이는 여전히 높은 진입 장벽이라는 과제를 안고 있다. 가장 큰 허들은 투명하지 않은 가격 모델인데, 공개된 가격표 없이 영업 대표와의 상담이 필수적이며, 보통 12개월 이상의 최소 약정과 사용량 커밋이 표준이라 중소기업에게는 부담스럽다.

사용자 경험(UX) 측면에서도 개선의 목소리가 높은데. 관리 콘솔이 타 플랫폼에 비해 지나치게 복잡하여 전문 지식 없이는 접근하기 어렵고, 커스텀 이미지 업로드나 최신 GPU 자원 도입 속도가 개발자 중심의 라이벌들에 비해 보수적이라는 평가를 받는다. 개발자 친화적인 도구 생태계는 클라우드플레어 등에 비해 열세에 있다는 것이 시장의 중론이다.

클라우드플레어(Cloudflare): 연결성 클라우드와 개발자 플랫폼의 최강자

34) https://www.akamai.com/products/akamai-inference-cloud-platform

클라우드플레어는 '연결성 클라우드(Connectivity Cloud)'라는 비전 아래, 네트워크와 보안 그리고 연산을 하나의 통합된 소프트웨어 스택으로 제공하며 가장 빠르게 성장하고 있는 플랫폼이다. 전 세계 웹사이트의 약 20%를 보호하고 450만 명 이상의 활성 개발자를 보유한 이들은, 복잡한 인프라 설정을 코드 한 줄로 추상화하여 '개발자 민주주의'를 실현하고 있다.



35)

그림 13 클라우드플레어 연결성 클라우드

핵심 강점: 단순함의 미학과 '이그레스 제로'의 혁신

클라우드플레어의 최대 강점은 압도적인 단순함과 강력한 통합에 있다. DNS 설정 변경만으로 전 세계 애니캐스트(Anycast) 네트워크의 보호를 받을 수 있으며, WAF, 디도스 방어, 제로 트러스트 보안이 단일 인터페이스에서 유기적으로 작동한다. 특히 2026년 현재 시장의 판도를 흔드는 것은 R2 오브젝트 스토리지로, AWS S3와 호환되면서도 데이터 전송료(Egress Fees)를 폐지한 R2는 벤더 종속성을 우려하는 CIO들에게 멀티 클라우드 전략을 가능하게 하는 핵심 열쇠가 되었다.

기술적으로는 V8 아이솔레이트 기반의 클라우드플레어 워커(Cloudflare Workers)가 거의 0에 가까운 콜드 스타트와 함께 서버리스 컴퓨팅의 표준을 제시하고 있다. 2025년 추가된 컨테이너 기능은 기존의 제약을 넘어 최대 40GB RAM과 40 vCPU를 지원하는 풀 리눅스 워크로드를 수용하며,

35) https://www.cloudflare.com/connectivity-cloud/

'Scale-to-zero' 모델을 통해 미사용 시 비용이 발생하지 않는 극강의 경제성을 제공한다. 업계 최고 수준의 무료 티어와 월 5달러에 1,000만 요청 역시 장점으로 언급된다.

AI 시대의 비전: AI 에이전트가 거주하는 행성

클라우드플레어는 자신들을 AI 애플리케이션 개발의 표준 플랫폼으로 포지셔닝한다. 2025년 11월, 5만 개 이상의 모델 카탈로그를 보유한 리플리케이트(Replicate)를 인수하였으며. 200개 이상의 도시 엣지 노드에 GPU를 배치하여 추론을 직접 수행하는 워커 AI(Workers AI)는 자체 구축한 러스트 기반 추론 엔진 Infire를 통해 기존 vLLM 대비 압도적인 성능을 보여준다.

매튜 프린스(Matthew Prince) CEO는 "AI 에이전트가 인터넷의 새로운 사용자라면, 클라우드플레어는 그들이 실행되는 플랫폼이자 통과하는 네트워크"라고 선언했다. 이를 위해 AI 모델 공급자를 단일 지점에서 관리하는 AI 게이트웨이와 벡터 데이터베이스인 Vectorize 등 AI 풀스택을 갖추며, 에이전트 간의 원활한 소통을 위한 기술적 인프라를 선점하고 있다.

사용자의 페인 포인트와 한계: '블랙박스' 운영과 성장의 성장통

클라우드플레어의 운영 철학은 많은 부분이 자동화되어 있어, 상세한 제어가 필요한 경우 오히려 방해가 될 수 있다. 동적 라우팅과 캐싱 로직이 블랙박스처럼 운영되어, 특정 장애 상황에서 상세한 원인 파악이 어렵다는 지적이 있다. 2025년 11월에는 봇 관리 시스템의 구성 오류로 ChatGPT, X 등을 다운시킨 대규모 글로벌 장애가 발생했으며, 중앙 집중식 설정 변경이 전체 분산 네트워크에 미치는 리스크를 단적으로 보여주었다.

또한, Workers의 128MB 메모리 제한(기본형 기준)은 대형 모델 처리의 걸림돌이며, 무료 및 프로 티어 사용자에 대한 고객 지원 품질은 여전히 개선이 필요한 영역으로 꼽힌다. 비용 측면에서도 HTTP 워크로드에서는 Lambda@Edge보다 저렴하지만, 대규모 로그 전송이나 분석 기능을 추가할 때 발생하는 비용은 예산 예측을 어렵게 하기도 한다.

패스트리(Fastly): 웹어셈블리의 선구자이자 실시간 엣지 전문가

패스트리는 네트워크 계층에서 고도의 로직을 직접 실행하고자 하는 개발자들을 위한 프로그래머블 엣지를 강조한다. 약 80개 PoP, 36개국 81개 도시로 경쟁사 대비 적지만, 고밀도 인터넷 교환점에 SSD 전용 서버와 티어(Tier) 1 트랜짓을 배치하는 “소수의 강력한 PoP” 철학을 취한다. 2025년 사상

첫 흑자 전환에 성공하며 비즈니스 모델의 성숙도를 증명한 패스트리는, 단순한 전송 속도를 넘어 엣지에서의 컴퓨팅 성능이 무엇인지 몸소 보여주는 기술 집약적 플랫폼이다.

핵심 강점: 마이크로초의 미학, '고밀도 엣지'

패스트리의 독보적인 강점은 속도와 제어권이다. 설정 변경 사항을 전 세계 노드에 150ms 이내에 전파하는 인스턴트 퍼지(Instant Purge)는 실시간성이 생명인 미디어, 뉴스, 금융 서비스에 필수적이며, VCL(Varnish Configuration Language)을 통한 극한의 설정 가능성이 시그니처다

기술적 핵심은 웹어셈블리 기반의 패스트리 컴퓨터(Fastly Compute)이다. Wasmtime 런타임을 활용해 컨테이너 대비 비약적으로 빠른 마이크로초급 콜드 스타트를 구현했으며, 개발자들은 VCL이나 러스트, Go 등을 통해 엣지에서 복잡한 애플리케이션 로직을 네이티브에 가까운 속도로 실행할 수 있다. 그리고 엣지에서 복잡한 애플리케이션 로직을 수행하면서도 높은 보안성과 성능을 유지하며, TTFB(Time to First Byte)와 LCP(Largest Contentful Paint)에서 레거시 벤더 대비 월등한 우위를 실데이터로 입증하고 있다.

AI 시대의 비전: '신뢰 인프라'와 시맨틱 캐싱

패스트리는 AI 에이전트가 생성하는 폭발적인 트래픽을 단순히 차단하는 대신 최적화하는 방향을 선택한다. 핵심 제품인 패스트리 AI 가속기는 업계 최초로 시맨틱 캐싱 기술을 도입했는데, 이는 AI 쿼리의 단순 텍스트가 아닌 의도와 의미를 벡터로 캐싱하여, 유사한 질문이 들어올 경우 LLM API를 호출하지 않고 엣지에서 즉각 응답을 반환하는 것으로, 이를 통해 응답 속도를 평균 9배 높이고 운영 비용은 획기적으로 낮추었다.[36)]

또한, 패스트리는 인공지능 에이전트의 신원을 검증하고 콘텐츠 라이선스 계약을 자동 집행하는 '신뢰 인프라'로서의 역할을 자처한다. 에이전틱 인터넷 환경에서 고품질 데이터를 안전하게 제공하고 제어하는 게이트웨이 역할을 수행하겠다는 비전이다.

사용자의 페인 포인트와 한계: 높은 진입 장벽과 니치한 생태계

강력한 성능의 대가는 높은 학습 곡선과 비용이다. 패스트리의 시그니처인 VCL은 일반 개발자들에게 생소한 언어이며, Wasm을 제대로 활용하기 위해서는 고도의 엔지니어링 역량이 요구된다. 또한 최소

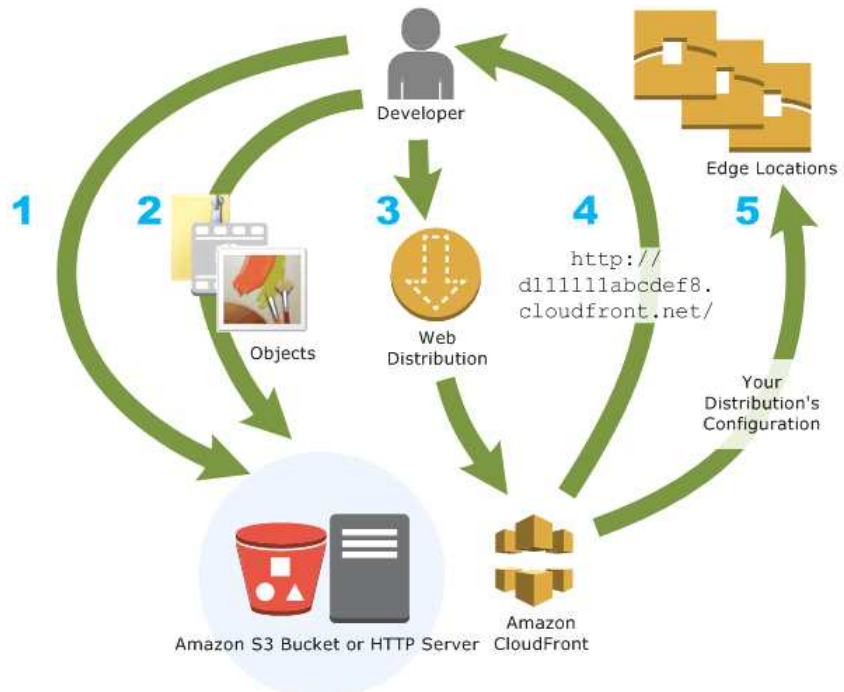
36) https://www.devopsdigest.com/fastly-ai-accelerator-released

월 50달러의 기본료와 1,500달러부터 시작하는 엔터프라이즈 패키지는 초기 스타트업에게는 다소 부담스러운 금액이다.

제품 포트폴리오의 폭이 좁다는 점도 약점으로 꼽힌다. 경쟁사인 클라우드플레어나 아카마이가 자체 오브젝트 스토리지나 데이터베이스를 풀스택으로 제공하는 것과 달리, 패스트리는 컴퓨팅과 전송에 집중하고 있는데, 따라서 DB나 스토리지가 필요한 경우 외부 벤더와 조합해야 하는 번거로움이 있으며, 중국 본토 내 PoP 부재는 글로벌 비즈니스를 전개하는 기업들에게 아쉬운 대목이다.

아마존 웹 서비스(AWS): 하이퍼스케일러의 깊은 생태계 통합

아마존 웹 서비스(AWS)는 전 세계 클라우드 시장 점유율 1위를 바탕으로, 기존 AWS 서비스들과 끊임없는 통합을 가장 큰 무기로 내세운다. 약 1,600개 이상의 엣지 로케이션을 보유한 AWS는 단순히 데이터를 배달하는 것을 넘어, 전 세계에 퍼져 있는 고객의 인프라를 하나의 거대한 유기체로 연결하고 있다.



37)

그림 14 클라우드프론트 동작 순서

37) https://docs.aws.amazon.com/ko_kr/AmazonCloudFront/latest/DeveloperGuide/Introduction.html

핵심 강점: 무한한 확장성과 인프라의 결합

AWS의 핵심 가치는 생태계의 힘에서 나온다. 아마존 클라우드프론트(Amazon CloudFront)를 사용하면 S3, EC2, ALB와 같은 AWS 오리진에서 엣지로의 데이터 전송료가 무료이며, 이는 대규모 트래픽을 운영하는 기업에 강력한 비용 절감 요인이 되며, AWS 사용자가 별도의 복잡한 절차 없이 전 세계 엣지 네트워크를 활용할 수 있다.

기술적으로는 클라우드프론트 함수와 엣지 람다라는 이중화된 엣지 컴퓨팅 전략을 구사한다. 1ms 미만의 실행 시간이 필요한 단순 헤더 조작이나 GDPR 준수를 위한 지역별 라우팅은 'Functions'가 담당하고, 복잡한 로직이나 데이터 처리는 엣지 람다가 처리한다. 2025년 11월에는 글로벌 발표된 VPC 오리진스 기능은 프라이빗 서브넷에 있는 자원을 별도의 공인 IP 노출 없이 엣지로 직접 연결할 수 있게 하여 보안과 성능 둘 다 개선을 보였으며, 엣지 및 온프레미스 확장으로는 환경에 따라 IoT 그린그래스(Greengrass), 웨이브렝쓰(AWS Wavelength), 아웃포스트 등을 운영할 수 있다.

AI 시대의 비전: AI-네이티브 인프라와 AI 공장(Factory)

AWS는 베드록과 Amazon Q를 통해 AI를 인프라의 중심으로 끌어들였다. 2025년 리인벤트(re:Invent)에서 발표된 AWS AI 팩토리즈는 고객 데이터센터에 파운데이션 모델과 전용 하드웨어를 포함한 관리형 AI 인프라를 배포한다.

또한, IoT 그린그래스의 'AI 에이전트 컨텍스트 팩키지'를 통해 엣지 디바이스에서도 Amazon Q와 같은 생성형 AI 도구를 활용한 앱 개발이 가능해졌다. 이는 AWS가 엣지를 단순한 연산의 끝단이 아니라, AI 에이전트가 학습된 데이터를 바탕으로 실시간 의사결정을 내리는 '지능형 공장'으로 보고 있음을 의미한다.

사용자의 페인 포인트와 한계: 극심한 복잡성과 숨겨진 비용

하지만 AWS의 방대한 서비스 라인업은 사용자에게 인지적 과부하를 안겨주는데, 200개가 넘는 서비스와 복잡한 설정 옵션 때문에 최적의 아키텍처를 설계하려면 고도의 전문 지식이 필수적이다. 특히 많은 사용자가 월 청구서의 60~70%가 숨겨진 비용이라고 토로할 만큼, API 호출료, 최대 41%에 이르는 리전 별 가격 편차, 잘못된 설정으로 인한 과금 등 예산 예측이 매우 어렵다.

기술적인 제약도 존재한다. 엣지 람다는 여전히 미국 특정 리전에서만 배포해야 하는 번거로움이 있고,

200~400ms에 달하는 콜드 스타트는 초저지연을 원하는 개발자들에게 불만의 대상이다. 무엇보다 강력한 통합은 곧 벤더 록인(Lock-in)을 의미하며, 외부 솔루션과의 유연한 결합을 원하는 기업에게는 높은 장벽이 되기도 한다.

애저(Azure): 엔터프라이즈 거버넌스와 온디바이스 AI의 결합

마이크로소프트 애저는 엔터프라이즈 환경에서의 강력한 통제권과 보안 체계를 강점으로 삼아 급성장해 왔다. 특히 2023년부터 2025년 사이 아카마이 등 외부 파트너십 기반의 CDN을 과감히 정리하고, 자체 서비스인 애저 프론트도어(AFD)로 통합하며 네이티브 클라우드 경쟁력을 공고히 했다.

핵심 강점: 통합과 강력한 백본

애저의 가장 큰 매력은 단순한 관리이다. 타 하이퍼스케일러가 여러 서비스를 복잡하게 조합해야 하는 것과 달리, AFD는 CDN, 글로벌 부하 분산, WAF, DDoS 방어를 단일 화면 내에서 통합 제공한다. 192개 이상의 엣지 노드와 마이크로소프트의 거대한 프라이빗 글로벌 백본을 기반으로 하기에, 특히 마이크로소프트 365나 Entra ID와의 강력한 연동은 보안 관점에서 엔터프라이즈 기업들에게 대체 불가능한 가치를 제공한다.

또한, 온프레미스와 엣지 인프라를 통합한 애저 로컬(구 Azure Stack Edge) 브랜드를 통해 하이브리드 환경에서의 컴퓨팅 요구사항을 만족시키는데, 이는 단순히 웹 콘텐츠를 전달하는 것을 넘어, 공장이나 사무실 등 현장에서 발생하는 데이터를 직접 처리하고자 하는 엔터프라이즈의 페인 포인트를 짚고 있다.

AI 시대의 비전: AI 파운더리 로컬(Foundry Local)과 에이전트 가드레일

애저는 2026년 현재 AI 에이전트 팩토리로서의 기능을 강화하고 있다. 단순히 클라우드에서 AI를 실행하는 것을 넘어, AI 파운더리 로컬을 통해 엣지 및 온디바이스 환경에서의 AI 추론을 현실화했다. ONNX 런타임에 최적화된 소형 언어 모델을 엣지 기기에서 가볍게 돌리고, 필요시 클라우드로 스케일 아웃하는 유연한 구조를 갖추었다.

또한, 수만 개의 AI 에이전트가 네트워크 트래픽을 주도하는 환경에 맞춰, 에이전트 전용 보안 가드레일을 AFD에 내장했다. 이는 AI 에이전트가 내부 데이터에 접근할 때 발생할 수 있는 보안 사고를 네트워크 계층에서 원천 차단하는 지능형 제어 시스템을 지향한다.

사용자의 페인 포인트와 한계: 진통 중인 통합과 운영 리스크

강력한 거버넌스의 이면에는 운영의 경직성과 대규모 장애의 기억이 자리 잡고 있다. 2025년 10월, 설정 변경 오류로 인해 약 8.3시간 동안 애저 포털과 M365 등 주요 서비스가 마비된 글로벌 장애 사례는 단일 아키텍처 통합이 가진 리스크를 단적으로 보여주었다. 또한, 2026년으로 예정된 레거시 서비스(Classic SKU)들의 강제 퇴역은 기존 사용자들에게 상당한 마이그레이션 압박과 비용 상승(최대 33%)을 안겨주고 있다.

기술적으로는 클라우드플레어 워커(Workers)와 같은 경량 V8 서버리스 엣지 컴퓨트가 없는 점이 부정적이다. 애저의 엣지 연산은 여전히 컨테이너나 VM 중심이라, 마이크로초 단위의 반응성을 요구하는 가벼운 로직 구현에는 다소 무겁다는 평가를 받는다. 또한, 마이크로소프트 생태계에 대한 깊은 종속성은 멀티 클라우드를 지향하는 기업들에게 전략적으로 고민거리가 된다.

구글 클라우드: 유튜브급 미디어 인프라와 데이터 지능화

구글 클라우드는 구글이 수십 년간 유튜브, 검색, 지메일을 통해 다져온 글로벌 네트워크 인프라를 기업용으로 개방하며 강력한 존재감을 드러내고 있다. 특히 2026년 현재, 대규모 미디어 배포와 AI가 결합한 데이터 지능형 네트워킹 분야에서 독보적인 기술적 우위를 점하고 있다.

핵심 강점: 전 세계에서 가장 강력한 미디어 혈관

GCP의 가장 큰 무기는 유튜브의 인프라를 그대로 옮겨온 미디어 CDN(Media CDN)이다. 100Tbps 이상의 압도적인 이그레스(Egress) 용량과 전 세계 3,000개 이상의 위치에 분산된 엣지 캐시를 통해, 4K/8K 스트리밍 등 초고대역폭이 필요한 유스케이스에서 98% 이상의 캐시 히트율을 기록한다.

또한, 2025년 6월에는 서비스 확장 플러그인이 런치되어 Wasm 기반으로 CDN 요청 경로에서 러스트, C++, Go 등의 커스텀 코드를 실행할 수 있게 되었다. 이는 클라우디너리(Cloudinary)와 같은 파트너사와의 협업을 통해 엣지에서 실시간 이미지/비디오 최적화를 수행하며, 구글의 프라이빗 백본을 기업 전용망처럼 사용하는 클라우드 WAN은 공용 인터넷 대비 최대 40% 빠른 속도와 비용 절감을 동시에 실현할 수 있다.

AI 시대의 비전: 에어갭 AI(Air-Gapped AI)와 지능형 미디어 최적화

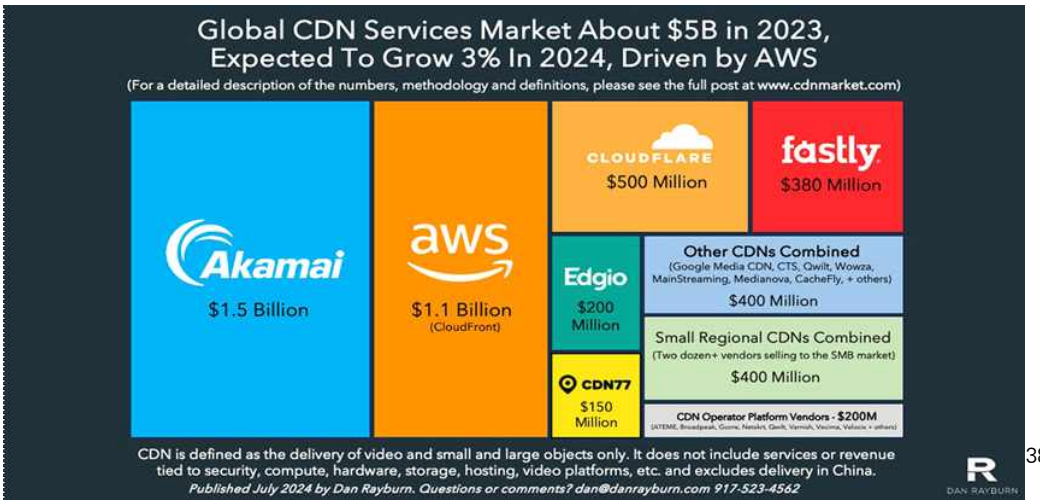
GCP는 버텍스AI와 제미나이를 네트워크 전면에 배치했는데, 특히 주목할 점은 구글 분산 클라우드(GDC)를 통한 엣지 AI 전략이다. 2025년 발표된 'GDC 위에서 제미나이'는 인터넷 연결이 불가능한 에어갭 환경이나 이동형 어플라이언스에서도 생성형 AI를 실행할 수 있게 하는데, 이는 국방, 재난 구조, 오지 제조 현장에서 AI 기반 실시간 데이터 처리와 번역을 가능케 한다.

한편 네트워크 계층에서는 AI가 동영상 콘텐츠를 실시간 분석해 최적의 비트레이트로 변환하고, 사용자의 다음 시청 행동을 예측하여 엣지 노드에 미리 데이터를 배치하는 등 고차원적인 인텔리전트 네트워킹을 구현하며 사용자 경험을 극대화하고 있다.

사용자의 페인 포인트와 한계: 독특한 구조와 조립의 어려움

하지만 구글 클라우드는 여전히 높은 학습 곡선이라는 숙제를 안고 있다. 구글만의 독특한 네트워킹 구조와 로드밸런싱 설정 방식은 AWS나 애저에 익숙한 엔지니어들에게 상당한 적응 시간을 요구한다. 또한 클라우드 CDN을 사용하기 위해 반드시 클라우드 로드밸런싱을 거쳐야 하는 구조는 복잡한 과제와 추가 비용을 발생시키는 요인이 되기도 한다.

클라우드플레어 워커스나 람다 엣지 같은 완전한 서버리스 엣지 플랫폼의 부재도 아쉬운 부분이다. Wasm 기능이 도입되었으나 여전히 헤더 조작이나 보안 정책 중심의 제한적인 범위에 머물러 있다. 또한, 파트너 에코시스템이 타 하이퍼스케일러 대비 작아 전문 인력을 확보하거나 서드파티 도구를 결합하는 과정에서 운영상의 어려움을 심하게 겪게 된다.



38)

그림 15 2024년 7월 기준의 시장 점유율

맺으며: CIO/CTO를 위한 제언

2026년의 기술 리더들은 단순히 하드웨어 사양을 비교하는 시대를 지나, 비즈니스의 민첩성과 인프라의 탄력성을 동시에 확보해야 하는 전략적 판단의 기로에 서 있다. AI 에이전트가 주도하는 새로운 인터넷 환경에서 기업의 성패는 얼마나 지능적이고 탄력적인 네트워크 플랫폼을 구축했는가에 달려 있다. 아카마이의 안정성, 클라우드플레어의 단순함, 패스트리의 성능, 그리고 CSP들의 통합 역량은 각기 다른 비즈니스 상황에 맞춰 조율되어야 한다.

CIO와 CTO는 이제 '클라우드 우선'을 넘어 '엣지 및 AI 네이티브' 관점에서 인프라를 재설계해야 한다. 기술 부채를 과감히 정리하고, Wasm 같은 표준화된 고성능 런타임을 도입하며, 보안을 네트워크 패브릭에 녹여내는 조직만이 다가올 지능형 경제 시대의 승자가 될 것이다. 아래의 3가지 제언으로 이 보고서를 마무리하겠다.

1. 기업 규모 및 워크로드 특성에 따른 벤더 선택 전략

먼저 스타트업 및 중견기업은 민첩성과 초기 비용 통제에 집중해야 하는데, 개발자 경험이 뛰어나고 혁신 속도가 빠른 클라우드플레어나 패스트리가 우선적인 고려 대상이다. 특히 클라우드플레어의 R2와 같은 서비스를 통해 데이터 이그레스 비용을 원천적으로 차단하고, 서버리스 환경에서 인공지능 기능을 빠르게 구현하는 것이 유리하다. 다만, 특정 벤더의 독점적 런타임에 종속되지 않도록 표준적인 컨테이너 설계나 Wasm 기반의 개발 체계를 병행하는 유연성이 필요하다.

반면, 엔터프라이즈는 레거시 현대화와 보안의 내재화를 최우선 과제로 삼아야 한다. 기존에 사용 중인 대형 CSP의 네이티브 솔루션을 근간으로 하되, 사용자 경험이 핵심인 프런트엔드 서비스에는 전문 벤더를 결합하는 전략적 멀티CDN 구축이 권장된다. 대규모 미디어 배포나 글로벌 보안 통합이 필요한 경우, 벤더의 계약 경직성보다는 그들이 제공하는 압도적인 분산 능력과 안정성을 우선순위에 두어야 한다.

2. AI 추론과 Wasm: 지능형 엣지 전환 로드맵

2026년은 AI 투자수익률(ROI)을 증명해야 하는 해이다. 모든 AI 요청을 중앙 집중식 클라우드 API로 처리하는 방식은 막대한 지연 시간과 비용을 초래하는데, 연산의 약 80%를 차지하는 추론 워크로드를

38) https://www.streamingmediablog.com/2024/07/cdn-market-size.html

엣지 노드로 전환함으로써 인프라 비용을 최대 90%까지 절감하는 최적화 전략을 실행해야 한다. IDC의 예측대로 2027년까지 대부분의 기업이 엣지 AI를 활용하게 될 것이므로, 2026년은 그 준비를 마치는 마지막 골든 타임이 될 것이다.

이 과정에서 웹어셈블리는 벤더 록인을 방지하고 실행 효율을 극대화할 핵심 열쇠가 될 것이다. 2026년 중반 WASI 1.0 최종 사양이 확정됨과 함께 '한 번 쓰고 어디서든 실행하는(Write-once, run-anywhere)' 엣지 컴퓨팅의 시대가 열릴 것이지만, 아직 도구 생태계가 완전히 성숙하지 않은 만큼, 전면적인 전환보다는 특정 모듈부터 점진적으로 도입하며 팀의 역량을 키우는 접근이 바람직하겠다.

3. 규제 준수와 운영 리스크: 보이지 않는 위협에 대응하기

이제 데이터 주권은 성능 향상을 넘어선 비즈니스의 생존 문제이다. EU AI Act의 전면 시행과 더불어 120개국 이상의 데이터 보호법에 대응하기 위해, 데이터를 발생지에서 즉시 처리하는' 소버린 엣지 아키텍처를 기본 원칙으로 수립해야 한다. 이는 단순히 법적 리스크를 피하는 것을 넘어, 고객의 데이터를 현지에서 안전하게 관리한다는 브랜드 신뢰도로 이어질 것입니다.

또한, 단일 벤더 장애가 비즈니스 전체 마비로 이어졌던 과거의 교훈을 잊어서는 안 된다. 멀티CDN은 이제 선택이 아닌 필수이며, 이를 효율적으로 관리할 수 있는 오케스트레이션 플랫폼 도입을 검토해야 한다. 마지막으로 가장 간과하기 쉬운 리스크는 인재 및 가시성의 격차인데, 분산 엣지 환경의 복잡성을 이해하는 전문가는 여전히 부족하며, IT 전문가의 절반 정도만이 엣지 컴퓨팅에 익숙한 상황이다. 기술 도입 못지않게 관측성 도구에 투자하고 조직 내 기술 교육을 강화하여 운영의 사각지대를 없애는 것이 2026년 리더의 핵심 책무가 될 것이다.