디지털서비스

이슈리포트

Digitalservice Issue Report

기밀 추론과 신뢰 가능한 AI 인프라

NIA 한국지능정보사회진흥원 디지털서비스 이용지원시스템

2026 Vol. 03

CONTENTS

Digitalservice Issue Report

디지털서비스 이슈리포트

01 기밀 추론과 신뢰 가능한 AI 인프라

| 윤대균 아주대학교

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

1. 들어가며

AI 인프라 무게 중심이 학습(training)에서 추론(inference)으로 이동하고 있다. 수천억 개의 파라미터를 수십만 장의 GPU로 수개월에 걸쳐 학습시키는 작업은 소수 AI 기업의 영역이었다. 그러나 완성된 모델들이 서비스로 배포되면서, 챗GPT/클로드/제미나이 같은 플랫폼을 통해 수억 명의 사용자가 매일 수십억 건의 추론 요청을 처리하는 구조로 전환되었다.

이 전환은 AI 인프라 시장의 구조도 재편하고 있다. 앞선 글 「네오클라우드」에서 살펴보았듯[1], 코어위브가 오픈AI·마이크로소프트·메타와 대규모 GPU 공급 계약을 체결할 수 있었던 배경도 추론 서비스 수요 증가에 있다. 다수의 많은 분석에 따르면 AI 인프라 비용 중 추론이 차지하는 비중은 학습을 넘어섰으며, 이 격차는 지속적으로 확대될 전망이다.[2] 추론이 AI 서비스의 핵심 실행 단계로 자리 잡으면서, 추론 인프라의 보안은 설계 단계에서 반드시 고려해야 할 요건이 되었다.

이에 기밀 추론(Confidential Inference)이란 용어가 등장하여 관심을 끌게 되었다. 기밀 추론은 민감한 사용자 입력을 보호하는 기술일 뿐 아니라, 모델 가중치 자체를 보호하고, 추론 서버와 가속기가 신뢰할 수 있는 상태에서 동작하고 있음을 암호학적으로 입증하는 체계이기도 하다. 다시 말해 기밀 추론은 AI 서비스 보안의 문제를 서버 보안에서 신뢰 가능한 추론 인프라 문제로 확장한다.

에이전트 AI의 확산은 이 문제를 한층 더 민감하게 만든다. 에이전트는 단순히 질문 하나에 답하는 시스템이 아니라, 장기간 컨텍스트를 유지하고 외부 도구를 호출하며 여러 단계를 거쳐 실제 행동을 수행한다. 이 과정에서 프롬프트뿐 아니라 메모리, 검색 결과, 도구 호출 파라미터, 실행 결과까지 모두 민감한 데이터가 된다. 따라서 추론 단계에서는 API 호출뿐만 아니라 기업 활동 전체가 공격 표면이 되어 버린다.

1) 윤대균, “네오클라우드 – AI 시대의 새로운 인프라 패러다임”, 디지털서비스 이슈리포트, Jan 2026
2) The Wall Street Journal, “Can Nvidia’s Dominance Survive the Sea Change Under Way in AI Computing?”, Mar 16, 2026

이 부분에서 기밀 추론의 정체성이 뚜렷하게 드러난다. 기밀 추론은 AI 모델을 더 안전하게 만드는 보조 기능이 아니라, 민감한 데이터를 가진 조직이 생성형 AI를 실제 업무에 연결할 수 있게 만드는 핵심 조건이다. 이 때문에 기밀 추론은 보안 기술이면서 동시에 AI 인프라 기술이다. 이번 글은 이런 문제의식을 바탕으로, 기밀 컴퓨팅의 후속편으로서[3] 기밀 추론이 무엇인지, 왜 중요해졌는지, 그리고 GPU 및 NPU 중심의 추론 인프라에 어떤 변화를 요구하는지 살펴본다.

2. 기밀 추론 vs. 기밀 컴퓨팅

'기밀 추론'과 '기밀 컴퓨팅'을 비교해 봄으로써 기밀 추론의 범주와 특징을 이해할 수 있다. 기밀 컴퓨팅은 하드웨어 기반 TEE(Trusted Execution Environment)를 통해 처리 중인 데이터를 보호하는 광범위한 기반 기술 개념이다. 반면 기밀 추론은 이 원리를 AI 추론 워크로드에 적용하기 위한 구체적인 프로세스 전반에 걸쳐 필요한 기술이다. 따라서 기밀 추론은 기밀 컴퓨팅의 응용 분야이지만, AI 워크로드 때문에 새롭게 생기는 문제를 함께 다뤄야 한다. 예를 들어 모델 가중치를 어떻게 안전하게 적재할 것인지, GPU나 NPU와 같은 가속기를 어떻게 신뢰 경계 안에 넣을 것인지, 멀티테넌트 추론 서버에서 다른 고객 워크로드와 어떻게 격리할 것인지 같은 문제는 일반적인 TEE 설명만으로는 충분히 다뤄지지 않는다.

기밀 추론의 핵심은 세 가지로 정리할 수 있다. 첫째, 입력 데이터는 가능한 한 암호화된 상태로 이동해야 한다. 둘째, 모델 가중치도 보호 대상 자산으로 다루어야 한다. 셋째, 복호화와 실제 연산은 검증 가능한 제한된 실행 환경에서만 일어나야 한다. 여기서 중요한 것은 단순히 TEE 안에서 코드를 실행하는 데 그치지 않고, 어떤 코드가 올라왔고 어떤 상태에서 실행되는지를 증명한 뒤 그 결과에 따라 복호화 권한을 부여하는 구조다.

구분	핵심 요약	보호 대상
기밀 컴퓨팅	하드웨어 기반 'data in use' 보호	메모리 및 실행 환경
기밀 추론	AI 추론 파이프라인의 신뢰 사슬	입력 데이터, 모델 가중치, 추론 서버, 가속기

표 2 기밀 컴퓨팅과 기밀 추론 비교

기밀 추론은 입력, 모델, 서버, 가속기, 키 관리 체계를 하나의 신뢰 사슬로 연결하는 구조에 기반한다. 이 신뢰 사슬에서는 최소한 다음 네 가지에 대한 해법을 제공해야 한다. (1) 입력 데이터는 어디서 암호화되는가 (2) 모델 가중치는 언제 복호화되는가 (3) 어떤 코드가 실제로 그 복호화를 수행하는가 (4)

3) 윤대균, "기밀 컴퓨팅이 여는 에이전트 AI 시대의 보안 혁신", 디지털서비스 이슈리포트, Feb 2026

그리고 그 모든 과정이 정말 승인된 하드웨어와 런타임 상태에서 일어났다는 사실을 누가 어떻게 검증하는가 이다.

또한 학습은 일반적으로 단일 조직의 데이터와 모델을 다루지만, 추론은 다수의 사용자가 공유 인프라에 동시에 접근하는 멀티테넌트 환경에서 수행된다. 에이전트 AI 환경에서는 하나의 추론 세션이 수십 개의 도구 호출과 외부 서비스 연동을 포함한다. 이러한 파이프라인 전체를 보안 경계 안에 포함하는 것은 TEE 적용 이상의 시스템 설계를 요구한다. GPU 중심 연산, 멀티테넌트 환경, 에이전트 파이프라인과 같은 다양한 특성의 결합이 기밀 추론이라는 독립적인 기술 영역이 형성된 배경이기도 하다.

3. 핵심 기술: 추론에 특화된 기밀 컴퓨팅

지난 '기밀 컴퓨팅' 글에서는 TEE와 엔클레이브 개념, 원격 증명의 기본 구조를 다루었다. 이 글에서는 '추론'이라는 워크로드가 별도의 보안 설계를 요구하는 이유에 집중한다.

추론 고유의 보호 대상

추론에는 학습에 없는 고유한 보호 대상이 있다. 첫 번째는 'KV 캐시(Key-Value Cache)'다. 트랜스포머 기반 LLM은 자기 회귀적 방식으로 토큰을 순차 생성한다. 이전 토큰들의 어텐션 연산 결과를 KV 캐시로 저장해두고 재사용하는데, 이 캐시는 대화 전체의 맥락을 포함한다. 멀티테넌트 서버에서 서로 다른 사용자의 KV 캐시가 물리적 메모리를 공유하는 경우, 캐시 오염이나 비인가 캐시 읽기를 통한 정보 유출이 발생할 수 있다. 기밀 추론 환경에서 KV 캐시는 사용자 단위로 암호화·격리되어야 한다.

두 번째는 '배치 격리(Batch Isolation)'다. 추론 서버는 처리량을 높이기 위해 여러 사용자의 요청을 하나의 배치로 묶어 처리한다(continuous batching). 이 방식은 처리 효율이 높은 반면, 서로 다른 사용자의 데이터가 동일한 연산 단위에 올라간다는 구조적 보안 문제를 수반한다. 기밀 추론에서는 배치 내 각 요청의 메모리 접근이 하드웨어 수준에서 격리되어야 한다.

세 번째는 '스트리밍 응답의 기밀성'이다. 추론 결과는 완성 후 일괄 전송되는 것이 아니라 토큰 생성 즉시 스트리밍으로 전달된다. TEE 내부에서 생성된 각 토큰이 외부로 전송되기 전에 암호화되어야 하며, 이 과정이 레이턴시에 민감한 인터랙티브 추론에서 응답 지연으로 이어지지 않도록 설계되어야 한다. 이를 좀 더 쉽게 설명하면, 안전한 상자(TEE) 안에서 한 글자씩 써서 밖으로 던져주는데 (스트리밍), 던지기 직전에 누가 못 보게 아주 빠르게 자물쇠를 채워야(실시간 암호화) 하며, 이 자물쇠 채우는 속도가

너무 느려서 읽는 흐름이 끊겨서는 안 된다는 뜻이다.

레이턴시 제약과 GPU TEE 기술의 발전

기밀 컴퓨팅을 학습과 추론에 적용하는 것 사이에는 본질적인 차이가 있다. 학습은 처리량(throughput) 중심의 워크로드이기에, 장시간에 걸친 연산에서 일정 수준의 오버헤드는 허용 가능하다. 추론은 레이턴시 중심의 워크로드다. 첫 토큰 생성 시간(TTFT)과 토큰 생성 속도(TPS)가 서비스 품질을 결정하며, 암호화 오버헤드로 인한 TTFT 증가나 TPS 저하는 서비스 품질에 부정적인 영향을 미친다.

CPU와 GPU 사이에서 데이터를 주고받을 때 데이터를 암호화하고 복호화하는 과정에서 CPU의 개입이 많아 병목 현상이 발생한다. 블랙웰에서는 TEE-I/O 기술을 적용하여 데이터가 PCIe 슬롯을 통해 CPU와 GPU 사이를 이동할 때 발생하는 병목 현상을 해결했다. 이전에는 CPU TEE(인텔 TDX, AMD SEV-SNP)와 GPU 사이의 데이터를 암호화하기 위해 소프트웨어 계층을 거쳐야 했다. 블랙웰에서는 TEE-I/O 기술을 적용하여 PCIe 컨트롤러가 하드웨어 수준에서 직접 암호화된 채널을 형성한다. 따라서, 기존 TEE가 CPU와 메모리 내부에서만 데이터를 보호했다면, TEE-I/O는 이 보호 영역을 GPU 같은 외부 장치(I/O 장치)까지 확장한 것이다. 즉, 블랙웰 GPU는 CPU와 함께 신뢰할 수 있는 도메인을 형성하여 데이터가 CPU 메모리를 떠나 GPU로 이동할 때도 암호화된 상태를 유지한다.[4)]

안정적인 추론 서비스를 제공하기 위해서는 레이턴시를 최소화하면서도 LLM 연산 성능을 높은 수준으로 유지하는 것이 필요하며 이를 위한 GPU TEE 기술도 보안/암호화 오버헤드를 줄이기 위한 방향으로 발전하고 있다.

종단 간 암호화 파이프라인과 KMS 연동

GPU TEE 단독으로 기밀 추론이 완성되지는 않는다. 클라이언트에서 서버까지의 입력 전달, TEE 내부 GPU 처리, 스트리밍 응답 반환의 전 과정이 하나의 신뢰 경계 안에 있어야 한다. 추론 특화 기밀 파이프라인은 다음 흐름으로 구성된다.

- **클라이언트 암호화 및 원격 증명 검증:** 사용자는 추론 시작 전 서버 TEE의 원격 증명 보고서를 검증하고, TEE의 공개키로 입력 데이터를 암호화하여 전송한다. 서버의 운영체제, 하이퍼바이저, 클라우드 운영자는 입력 내용에 접근할 수 없다.

4) Nvidia, “NVIDIA Blackwell Architecture Technical Brief”, 2025
https://resources.nvidia.com/en-us-blackwell-architecture

- **TEE 내부 복호화 및 추론·KV 캐시 관리:** 복호화는 TEE 내부에서만 수행되며, KV 캐시는 사용자 세션 단위로 격리된 메모리 영역에 유지된다. 배치 처리 시에도 각각 요청의 어텐션 연산과 캐시는 하드웨어 경계로 분리된다.
- **토큰 단위 암호화 응답:** 생성된 토큰은 TEE를 벗어나기 전에 암호화되어 스트리밍 전송된다.

모델 가중치 보호는 별도의 설계 축으로 구성된다. 수백 GB~수 TB에 달하는 가중치는 모델 소유자의 KMS(Key Management System) 키로 암호화된 상태로 배포된다. 추론 서버 TEE가 기동되면 원격 증명 보고서를 KMS에 제출하고, KMS는 이를 검증한 후에만 복호화 키를 TEE 내부로 전달한다. 이 구조에서 인프라 제공 사업자는 모델 가중치의 평문에 접근할 수 없으며, 모델 소유자는 자신의 인프라 없이도 서드파티 클라우드를 통해 모델 IP를 보호하면서 서비스를 배포할 수 있다.

4. 기밀 추론 시스템 설계: 앤스로픽 사례

앤스로픽은 자사의 백서에서 기밀 추론의 설계 원칙과 구현 방식을 상세히 다루고 있다. 앞선 글 '기밀 컴퓨팅'에서 잠깐 언급하기도 했지만 여기서 좀 더 살펴보기로 하겠다.[5)]

기밀 추론 시스템은 모델 소유자, 데이터 소유자, 그리고 클라우드 서비스 제공자(CSP) 간의 신뢰가 완전히 보장되지 않은 환경에서도 하드웨어 기반 TEE를 활용하여 연산 중인 데이터의 기밀성을 보호하는 것을 목표로 한다. 일반 기밀 컴퓨팅에서의 요구사항과 마찬가지로 데이터를 처리하는 과정에서 운영체제나 시스템 관리자조차 메모리에 접근할 수 없도록 격리하며, 암호화된 증명(Attestation)을 통해 실행 중인 워크로드의 무결성을 검증한다.

기밀 추론의 핵심 영역은 '데이터 기밀성(Confidential Data)'과 '모델 기밀성(Confidential Model)'이라는 두 가지 측면으로 나뉜다. 데이터 기밀성 측면은 사용자가 입력하는 프롬프트와 모델이 생성하는 출력값이 추론을 수행하는 호스트나 기타 애플리케이션에 노출되지 않도록 엔드투엔드 수준으로 보호하여 개인정보 및 민감 데이터를 안전하게 처리할 수 있게 한다. 모델 기밀성 측면은 AI 모델의 가중치와 아키텍처를 외부 유출이나 리버스 엔지니어링으로부터 보호하는 데 초점을 맞추며, 특히 국가 수준의 해킹 위협을 방어해야 하는 고위험 모델(SL4 및 SL5 보안 등급[6)])의 경우 TEE

5) Anthropic, "Confidential Inference Systems Design principles and security risks", Jun 2025
6) 랜드 연구소가 2024년에 발표한 보고서인 "AI 모델 가중치 보안: 프론티어 모델의 탈취 및 악용 방지(Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models)"에서 AI 모델의 가중치를 보호하기 위한 보안 시스템의 방어 능력을 SL1부터 SL5까지 총 5단계로 나누어 정의했다. 이중, SL4, SL5는 아주 높은 최상위 수준의 공격까지 방어해야 하는 등급이다.

내부에서만 가중치 복호화가 이루어지도록 강력히 권고하고 있다.

이러한 기밀성을 달성하기 위해 서비스 제공자는 하드웨어 기반의 컴퓨팅 및 메모리 격리, 암호화 증명, 그리고 운영자의 엄격한 접근 통제를 보장해야 한다. AI 추론은 대규모 연산을 위해 GPU나 NPU 같은 AI 가속기를 필수적으로 요구하므로, 기밀 보호 경계(Confidential Boundary)는 CPU를 넘어 가속기까지 확장되어야 한다. 이를 위해 AI 가속기 내부에 네이티브 TEE를 구현하여 데이터를 엔드투엔드로 전송하거나, CPU의 보안 엔클레이브와 가속기를 안전하게 연결하는 보안 통로를 구축하는 기술적 접근 방식이 활용된다. 앞서 블랙웰의 TEE-I/O가 한 예이다.

시스템의 아키텍처는 크게 '기밀 추론 서비스(Confidential Inference Service)', '모델 프로비저닝(Model Provisioning)', '엔클레이브 빌드 환경(Enclave Build Environment)'이라는 세 가지 핵심 구성 요소로 구성된다. 기밀 추론 서비스는 격리된 보안 엔클레이브 내에서 암호화된 가중치와 입력값을 받아 복호화한 후 추론을 수행하며, 이 과정에서 엔클레이브 외부 프로그램은 오직 통신 프록시 역할만 수행한다. 모델 프로비저닝 단계에서는 키 관리 시스템(KMS)을 이용해 모델을 암호화하여 저장하고, 올바른 증명 문서를 제시하는 보안 엔클레이브의 요청에만 복호화 키를 제공한다. 아울러 엔클레이브 프로그램이 악성 코드를 포함하지 않음을 보증하기 위해 투명하고 재현 가능한 형태의 안전한 빌드 환경이 필수적으로 뒷받침되어야 한다.

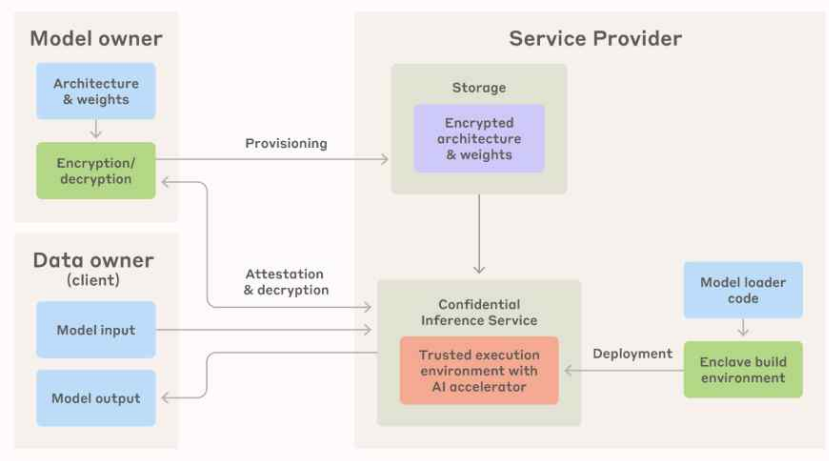


그림 1 앤스로픽 기밀 추론 구성도 (출처: 앤스로픽)

앤스로픽이 제시한 모델은 추론 시스템을 구동하는 호스트 머신과 클라우드 네트워크 전체가 이미 공격자에 의해 장악될 수 있음을 가정하여 설계되었다. 시스템 보안을 위협하는 요소는 하드웨어 자체의

결함이나 암호학적 취약점과 같은 '시스템적 위험'과, 잘못된 KMS 정책 설정, 안전하지 않은 엔클레이브 로딩, 빌드 환경 침해 등 구현 과정에서 발생하는 '도입된 위험'으로 구분된다. 따라서 안전한 기밀 추론 시스템을 운영하기 위해서는 하드웨어 패치를 최신화하는 것뿐만 아니라, 최소 권한 원칙을 준수하고 심층 방어 설계를 철저히 점검하는 노력이 수반되어야 한다.

5. 기밀 추론 인프라

네오클라우드의 기회

「네오클라우드」 글에서 살펴본 것처럼, 코어위브·람다·크루소·네비우스 등은 AI 인프라 서비스로 빠르게 성장했다. 추론 워크로드는 학습과 달리 지속적으로 실행되고 사용자의 민감한 데이터를 실시간으로 처리한다는 점에서, 전통적인 보안 요건과는 상이하다. 네오클라우드의 구조적 약점 중 하나는 하이퍼스케일러 대비 부족한 컴플라이언스 및 보안 인증이다. FedRAMP, HIPAA, SOC 2 등과 같은 기업 또는 정부 고객이 요구하는 보안 요건을 충족하지 못하는 경우가 많아, 금융·의료·공공 부문의 민감한 추론 워크로드는 하이퍼스케일러로 향하는 경향이 있다.

기밀 추론은 이 문제에 기술적 대안을 제공한다. 기밀 추론 환경에서 데이터 보호는 클라우드 운영자의 정책, 인증 또는 SLA(Service Level Agreement)에 의존하는 것이 아니라 하드웨어와 암호학적 메커니즘에 의해 강제된다. 이는 네오클라우드가 광범위한 규제 인증 없이도 특정 보안 요건을 암호학적으로 증명할 수 있는 경로를 제공한다.

모델 공급자 입장에서도 기밀 추론은 네오클라우드 활용을 가능하게 하는 기술적 수단이 된다. 앤스로픽의 KMS 연동 방식처럼, 모델 가중치를 암호화된 채로 네오클라우드 GPU에 배포하고 TEE 원격 증명이 확인된 환경에서만 복호화 키를 전달하면, 인프라 제공 사업자도 모델 내용에 접근할 수 없다. 기밀 추론 서비스를 선제적으로 구축하는 네오클라우드는 가격 경쟁력과 기술적 보안 증명을 동시에 제공하는 포지셔닝이 가능해진다.

추론 특화 NPU와 기밀 추론 요구사항

기밀 추론 인프라는 GPU에 한정되지 않는다. 추론 워크로드는 전력 효율과 레이턴시가 경쟁력의 척도인데, GPU보다 추론에 최적화된 NPU 기반 가속기들이 이 시장을 겨냥하고 있다. NPU는 행렬 연산에 특화된 구조로 GPU 대비 추론 전력 효율이 높고, 엣지나 온프레미스 배포에 적합한 폼팩터를

가진다. 구글 TPU, 아마존 트레이니엄(Trainium)과 인퍼런시아(Inferentia), 메타 MTIA 등 주요 클라우드들 공급자들도 자체 NPU를 추론 비용 최적화 수단으로 개발/운용하고 있다.

NPU가 기밀 추론 인프라로 활용되려면 다음과 같은 보안 요건을 충족해야 한다.

- **온칩 메모리 암호화:** NPU 내부의 스크래치패드 메모리와 가중치 버퍼에 적재된 데이터를 하드웨어 수준에서 격리해야 한다.
- **하드웨어 신뢰 루트 기반 원격 증명:** NPU 펌웨어와 모델 코드의 무결성을 암호학적으로 증명하고, KMS가 이를 검증하여 복호화 키를 전달하는 체계가 필요하다.
- **호스트 DMA 격리:** 호스트 CPU가 NPU 메모리에 임의로 접근하는 DMA(Direct Memory Access) 경로를 차단해야 한다.

화웨이의 경우 자사 어센드 NPU를 대상으로 한 어센드-CC 연구에서 CPU TEE에 의존하지 않고 NPU 자체를 신뢰 경계로 삼는 기밀 추론 아키텍처를 제안했다. 어센드 910A 위에서 라마 시리즈 모델의 기밀 추론을 구현하고 성능 오버헤드를 1% 이하로 줄일 수 있음을 보여주었다.[7)]

글로벌 시장에서 주목받고 있는 대표적인 국내 NPU 기업의 경우 양산 단계에 있으나 아직 공식적으로 기밀 컴퓨팅을 지원한다는 계획이 알려진 바는 없다. 2025년 10월 퓨리오사가 기밀컴퓨팅 컨소시엄의 스타트업 멤버로 가입한 바는 있다.[8)]

엔비디아 H100이 GPU TEE를 지원하기까지 오랜 개발 기간이 소요되었다는 점을 감안하면, 국내 NPU 기업들이 기밀 추론 지원을 조기에 로드맵에 반영하는 것이 중요하다. 정부 주도 국가 AI 컴퓨팅 인프라에 국산 NPU가 채택되고 그 위에서 의료·금융·공공 부문의 민감한 추론 서비스가 운영될 경우, 기밀 추론 기능은 사실상 진입 요건으로 기능하게 된다. 삼성전자와 SK하이닉스의 HBM이 GPU TEE 생태계의 필수 부품으로 자리 잡은 것처럼, 국내 NPU 기업들이 기밀 추론 지원을 칩 설계 단계부터 통합한다면 글로벌 기밀 AI 인프라 시장에서 차별화된 위치를 확보할 수 있다.

6. 기술적 과제 및 시사점

'기밀 컴퓨팅'에서 풀어야 할 과제는 기밀 추론에서도 모두 적용된다. 표준화, CPU/GPU 제조사로

7) Aritra Dhar et al., "Ascend-CC: Confidential Computing on Heterogeneous NPU for Emerging Generative AI Workloads", arXiv, Jul, 2024
8) https://confidentialcomputing.io/2025/10/14/welcoming-furiosaai-to-the-confidential-computing-consortium/

귀결되는 신뢰체인의 기본 한계, 그리고 사이드 채널 공격 등이다. 기밀 추론에서는 분산 추론으로부터 야기되는 보안 경계 문제가 더욱 복잡해질 수 있다. 대형 모델의 추론은 수십~수백 개의 GPU에 모델을 분산하는 텐서 병렬/파이프라인 병렬 방식으로 수행된다. GPU 간 통신 경로 전체가 신뢰 경계 안에 포함되어야 하는데, 이를 확장 가능한 방식으로 구현하는 것은 현재 진행 중인 과제다. 블랙웰의 NV링크 암호화는 이 방향으로 다소 진전한 것으로 볼 수 있지만 다중 노드로 확장되는 경우는 또 다른 차원의 얘기다.

기밀 추론은 우리나라 AX 전략에서도 매우 중요한 의미가 있다. 의료, 금융, 공공 영역의 AI 서비스 확산에 있어 기술적 기반이 될 수 있다. 환자 진료 기록의 클라우드 AI 처리를 기피하는 의료기관, 고객 거래 데이터의 외부 AI 전달을 우려하는 금융기관은 기밀 추론 환경에서 기술적으로 보장된 데이터 보호가 전제되어야 본격적으로 AI를 도입할 수 있다.

「네오클라우드」 글에서 언급된 국가 AI 데이터센터 구축 전략과 기밀 추론도 연결된다. 공공 AI 컴퓨팅 인프라를 민간 기업과 연구 기관이 활용할 때, 각 사용자의 데이터와 모델을 기술적으로 격리/보호하는 수단이 필요하다. 공공 인프라에서의 데이터 주권 문제를 정책·계약이 아닌 기술로 해결하는 도구로서 기밀 추론이 기능할 수 있다.

반도체 산업 관점에서도 기밀 추론은 기회를 제공한다. 「기밀 컴퓨팅」 글에서 언급했듯이 삼성전자와 SK하이닉스의 HBM이 GPU TEE 인프라의 필수 부품이다. GPU TEE에서 사용되는 메모리 암호화 기술과의 연동 최적화, 국산 AI 가속기의 TEE 기능 통합이 추진 가능한 방향이다. 기밀 추론이 AI 서비스의 표준 요건으로 자리 잡을 경우, 이 생태계에서의 위치는 당장 얼마나 빨리 로드맵을 설정하고 기술 투자를 하는가에 의해 결정된다.

AI 서비스에서 신뢰의 기술적 근거가 변화하고 있다. 서비스 제공업체의 약관, 법적 의무, 평판에 기반한 신뢰에서, 원격 증명을 통한 암호학적 검증으로 신뢰의 근거가 이행하고 있다. 에이전트 AI가 기업의 핵심 데이터를 처리하고 의료/금융 의사결정을 지원하는 환경이 확대될수록, 이 기술적 신뢰의 토대가 AI 서비스의 사회적 수용성을 결정하는 요인이 될 것이다.

02 GTC 2026을 통해 살펴본 엔비디아 피지컬 AI 관련 기술

| 한상기 테크프론티어 대표

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

엔비디아가 매년 개최하는 GTC(GPU Technology Conference)는 이제 단순 GPU 개발자 행사가 아니라 사실상 AI 산업의 방향을 정의하는 글로벌 플랫폼이다. GTC에서는 AI 로드맵 발표, GPU 및 시스템 아키텍처 공개, 산업별 AI 적용 사례 공유 (헬스케어, 자율주행, 로보틱스 등), 글로벌 기업·연구기관 네트워킹 등이 이루어진다. 특히 젠슨 황 CEO의 키노트는 사실상 AI 산업의 연간 선언문이라고 할 수 있다.

천 개가 넘는 수많은 기술 세션과 산업별 트랙이 있고 스타트업과 생태계를 위한 전시와 네트워킹이 이루어지기 때문에 엔비디아는 이제 단순한 칩 메이커가 아니라 AI 생태계 오케스트레이터가 되었음을 알 수 있다.

이번 달에는 GTC 2026의 발표 내용 중에서 필자가 관심을 많이 갖고 있는 피지컬 AI 분야를 위해 엔비디아가 새로 발표한 기술과 제품에 대해 정리해서 소개하도록 한다.

피지컬 AI 데이터 팩토리 블루프린트

이번 GTC에서 가장 눈에 띄는 발표인데, 피지컬 AI를 위한 대규모 학습에 드는 비용, 시간 및 복잡성을 줄이기 위해 학습 데이터의 생성, 증강 및 평가 방식을 통합하고 자동화하는 개방형 참조 아키텍처인 피지컬 AI 데이터 팩토리 블루프린트를 공개했다.

이를 통해 개발자는 NVIDIA 코스모스 오픈 월드 파운데이션 모델과 선도적인 코딩 에이전트를 사용하여 제한된 학습 데이터를 대규모의 다양한 데이터 세트로 변환할 수 있다. 여기에는 실제 세계에서 포착하기 어렵고, 비용이 많이 들고, 시간이 오래 걸리며, 종종 비현실적인 드문 예외 사례와 장기 시나리오를 포함한다. 엔비디아는 마이크로소프트 애저 및 네비우스와 협력해 오픈 소스

블루프린트를 클라우드 인프라 및 서비스에 통합함으로써 개발자가 가속화된 컴퓨팅 성능을 대용량 학습 데이터로 전환할 수 있도록 했다.

필드AI, 헥사곤 로보틱스, 링커 비전, 마일스톤 시스템즈, 로보포스, 스킬드AI, 테라다인 로보틱스 및 우버와 같은 주요 피지컬 AI 개발업체들은 이 블루프린트를 사용하여 로봇 공학, 비전 AI 에이전트 및 자율주행 차량 개발을 더 빠르게 실행할 것이다.

피지컬 AI 데이터 팩토리 블루프린트는 모듈식 자동화 워크플로를 통해 팀이 원시 데이터에서 모델 학습 세트로 전환할 수 있도록 지원하는 단일 참조 아키텍처 역할을 한다. 아래와 같은 주요 기능으로 설명할 수 있다.

- 큐레이션 및 검색: 코스모스 큐레이터는 대규모 실제 및 합성 데이터 세트를 처리, 정제 및 주석 처리한다.
- 증강 및 확장: 코스모스 트랜스퍼는 엄선된 데이터를 기하급수적으로 확장하고 다양화하여 실제 및 시뮬레이션 입력을 증폭시켜 다양한 환경과 조명 조건에서 발생하는 희귀하고 드문 시나리오를 더욱 효과적으로 포착한다.
- 평가 및 검증: 코스모스 이밸류에이터는 코스모스 리즌을 기반으로 하며 현재 깃허브에서 사용할 수 있고, 생성된 데이터를 자동으로 채점, 검증 및 필터링하여 물리적 정확성과 교육 준비 상태를 보장한다.

블루프린트는 엔비디아에서는 알파마요 학습이나 평가에 사용하고 있고, 스킬드 AI는 범용 로봇 파운데이션 모델을 발전시키고 있으며, 우버는 자율주행차 개발을 가속하는 데 사용하고 있다.

많은 로봇 개발자들은 대규모 데이터 생성을 위해 필요한 복잡한 AI 인프라를 구축하고 관리할 역량이 부족하다. 그래서 제공하는 오스모(OSMO)는 오픈 소스 오케스트레이션 프레임워크로, 컴퓨팅 환경 전반에 걸쳐 이러한 워크플로우를 통합 및 관리하여 수동 작업을 줄이고 개발자가 모델 구축에 집중할 수 있도록 한다. 에이전트 컨텍스트 파일과 함께 CLI 형태로 제공되는 오스모는 AI 코딩 에이전트가 개발 환경에 대한 완벽한 상황 인식을 갖춘 피지컬 AI 플랫폼 전문가가 되게 해 준다. 코딩 에이전트는 단순히 작업을 제출하는 것 외에도 파이프라인에 대해 추론하고, 실행 중인 워크플로를 쿼리하고, 사용 가능한 GPU 용량을 확인하고, 플랫폼 활동을 실시간으로 모니터링할 수 있다. 오스모는 이제 클로드 코드, 오픈AI 코덱스, 커서와 같은 주요 코딩 에이전트에 통합되어 에이전트가 리소스를 사전에 관리하고 병목 현상을 해결하며 대규모 모델 제공을 가속화하는 AI 기반 운영을 지원한다.

마이크로소프트 애저는 블루프린트를 개방형 피지컬 AI 툴체인에 통합했으며, 툴체인은 이제 깃허브에서 사용할 수 있다. 블루프린트는 애저 IoT 오퍼레이션즈, 마이크로소프트 패브릭, 리얼 타임 인텔리전스

및 마이크로소프트 파운드리를 포함한 애저 서비스와 통합되어 엔터프라이즈급 에이전트 기반 워크플로를 통해 피지컬 AI 시스템을 신속하고 대규모로 학습 및 검증할 수 있도록 지원한다. 블루프린트는 4월 중에 깃허브에서 공개할 예정이라고 한다.

차세대 지능형 로봇을 위한 기술

이번 GTC에서 엔비디아는 새로운 아이작 시뮬레이션 프레임워크, 새로운 엔비디아 코스모스 3, 아이작 그루트 모델을 공개했다. 아이작은 개방형 로봇 개발 플랫폼으로 시뮬레이션 및 로봇 학습 프레임워크, 쿠다 가속 라이브러리, AI 모델, 그리고 자율 이동 로봇(AMR), 로봇 팔, 매니퓰레이터, 휴머노이드 로봇을 제작하기 위한 참조 워크플로로 구성한다. 나머지는 지난 1월과 2월 원고를 참고하기 바란다.

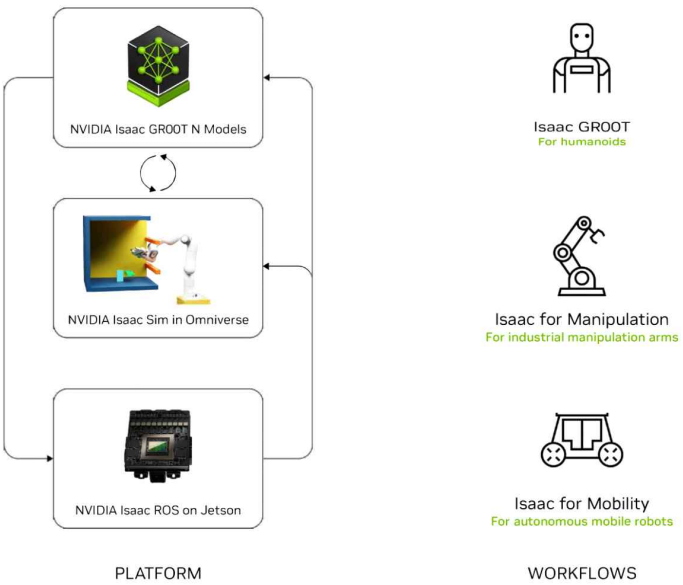


그림 2 엔비디아의 아이작과 관련 기술 그룹

엔비디아 플랫폼을 기반으로 사업을 구축하는 업계 선도 기업으로는 ABB 로보틱스, AGIBOT, 애질리티, 화낙(FANUC), 피규어, 헥사곤 로보틱스, 쿠카(KUKA), 스킬드 AI(Skild AI), 유니버설 로봇츠, 월드 랩스 및 야스카와가 있다.

전 세계적으로 200만 대 이상의 로봇을 설치한 화낙, ABB 로보틱스, 야스카와 및 쿠카는 엔비디아 옴니버스 라이브러리와 아이작 시뮬레이션 프레임워크를 자사의 가상 시운전 솔루션에 통합하여

물리적으로 정확한 디지털 트윈을 통해 복잡한 로봇 애플리케이션과 전체 생산 라인을 개발 및 검증하고 있다. 생산 라인에 고급 인텔리전스를 구현하기 위해서는 젯슨(Jetson) 모듈을 컨트롤러에 통합하여 엣지에서 실시간 AI 추론을 실행하고 있다.

필드AI와 스킬드AI는 코스모스 월드 모델을 활용한 데이터 생성과 아이작 시뮬레이션 프레임워크를 이용한 시뮬레이션 정책 검증을 통해 일반화된 로봇 두뇌를 구축하고 있다. 월드 랩스는 아이작 심을 이용해 자체 생성형 월드 모델을 검증하고 있으며, 제네랄리스트 AI는 코스모스를 활용하여 합성 데이터 생성을 연구하고 있다.

차세대 휴머노이드 개발을 위해 1X, AGIBOT, 애질리티, 애자일 로보츠, 보스톤 다이나믹스, 피규어, 헥사곤 로보틱스, 휴머노이드, 멘티, 뉴라 로보틱스 등은 코스모스 월드 모델, 아이작 심 및 아이작 랩을 사용해 개발과 검증을 가속하고 있다.

엔비디아는 이번에 엔비디아 DGX급 인프라에서 더 빠르고 대규모의 로봇 학습을 가능하게 하는 아이작 랩 3.0을 얼리 액세스 버전으로 출시했다. 새로운 뉴턴 물리 엔진 1.0과 엔비디아 PhysX 소프트웨어 개발 키트를 기반으로 구축된 아이작 랩 3.0은 다중 물리 시뮬레이션을 추가하고 복잡하고 정교한 조작을 위한 지원을 개선했다.

아이작 랩은 모듈형 아키텍처와 GPU 기반 병렬 처리를 기반으로 환경 설정부터 정책 학습까지 로봇 학습을 위한 포괄적 프레임워크를 제공한다.[9] 모방 학습과 강화 학습을 모두 지원하며, 뉴턴, PhysX, 워프, 뮤조코(MuJoCo) 등 다양한 물리 엔진을 활용할 수 있다. 아이작 랩은 아이작 그루트 플랫폼의 핵심 로봇 학습 프레임워크이다.

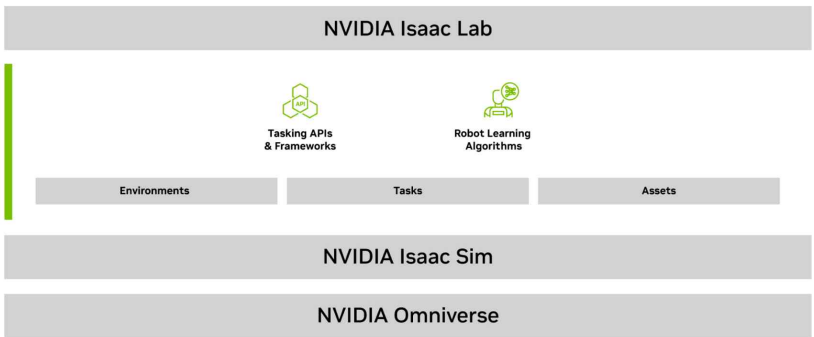


그림 3 아이작 랩 구성도 (출처: 엔비디아)

9) https://developer.nvidia.com/isaac/lab

이번에 그루트 N1.7을 상용 라이선스를 통해 얼리 액세스 버전으로 출시했으며, 이를 통해 고급 정밀 제어를 포함한 일반화된 로봇 기능을 생산 현장에 바로 적용할 수 있다고 발표했다. 그루트 N 모델을 사용하는 기업에는 AGIBOT, 휴머노이드, LG 전자, 뉴라 로보틱스, 그리고 노블 머신즈가 있다.

또한 젠슨 황은 기조연설에서 드림제로의 연구[10]를 기반으로 하는 차세대 로봇 모델인 그루트 N2를 공개했는데, 새로운 세계 행동 모델 아키텍처를 기반으로 구축된 이 모델은 기존의 비전 언어 행동 모델보다 로봇이 새로운 환경에서 새로운 작업을 두 배 이상 더 성공적으로 수행할 수 있도록 지원한다. N2는 연말 출시 예정이다.

로보택시 플랫폼 확대

키노트에서 젠슨 황은 엔비디아의 로보택시 플랫폼이 BYD, 현대, 닛산, 지리자동차 등 새로운 자동차 제조업체 파트너들을 끌어들이고 있다고 밝혔다. 또한 이러한 차량을 우버의 차량 호출 네트워크에 배치하기 위한 우버와의 파트너십을 강조했다.

엔비디아의 자율주행차 연구 담당 수석 이사인 마르코 파보네는 자율주행 개발을 위한 개방형 AI 모델, 시뮬레이션 도구 및 데이터 세트 제품군인 알파마요에 대한 업데이트를 발표했다. 최신 버전인 알파마요 1.5 추론 VLA 모델은 내비게이션 텍스트 안내 기능과 모델의 추론과 동작 간의 연결을 강화하는 새로운 학습 후 얼라인먼트 도구를 추가하여 개발자가 "자신의 데이터 세트에 맞게 모델을 맞춤 설정할 수 있도록" 한다.

엔비디아가 자율주행을 기본으로 하는 로봇택시 영역에서 이번에 발표한 주요 내용은 다음과 같다.

- 우버와 전략적 협력 확대: 우버는 엔비디아의 '로보택시 레디(Robotaxi-Ready)' 플랫폼을 활용해 글로벌 로보택시 네트워크를 구축할 계획이다. 양사는 2026년부터 수천 대의 레벨 4(L4) 자율주행 차량을 온라인화하고, 2027년까지 로스앤젤레스 등 주요 도시로 서비스를 확장할 예정이다.
- 글로벌 자동차 제조사 합류: 현대자동차, 기아, BYD, 닛산, 지리(Geely) 등이 엔비디아 드라이브 하이페리온 플랫폼을 채택하여 로보택시 및 자율주행 차량을 개발하고 있다.
- 닛산 & 웨이브(Wayve) 프로토타입: 닛산과 AI 스타트업 웨이브는 닛산 리프(LEAF)를 기반으로 한 로보택시 프로토타입을 공개했다. 이 차량은 고정밀 지도 없이 실시간으로 도로 상황을 학습하는 '엔드 투 엔드' AI를 탑재했다.

10) https://dreamzero0.github.io/

- 위라이드 GXR 공개: 자율주행 스타트업 위라이드는 동남아시아 시장을 겨냥한 로보택시 모델 'GXR'을 GTC 현장에서 선보였다.

관련 기술 발표 내용으로는 아래와 같은 내용이 있다.

- 엔비디아 드라이브 하이페리온 10: L4 자율주행을 위한 차세대 개발 플랫폼으로, 강력한 토르 SoC(System-on-a-Chip)와 고해상도 카메라 14대, 레이더 9대, 라이다 1대 등으로 구성된 센서 스위트를 포함한다.
- 데이터 팩토리 및 시뮬레이션: 우버와 엔비디아는 코스모스 플랫폼을 기반으로 공동 데이터 팩토리를 구축하여, 시뮬레이션과 합성 데이터를 통해 자율주행 모델을 지속적으로 개선하고 있다.

글로벌 산업 소프트웨어 회사와 협력

엔비디아가 이번 GTC에서 발표한 내용 중 하나는 글로벌 산업용 소프트웨어 선도 기업에게 쿠다-X, 옴니버스, GPU 가속 산업용 소프트웨어와 도구를 제공함으로써 설계, 엔지니어링 및 제조 속도를 향상시키겠다는 것이다. 이들 소프트웨어 선도 기업들은 엔비디아 기반 에이전트 솔루션을 도입하여 고객들이 차세대 AI 시대에 대비할 수 있도록 지원할 예정이다.

이러한 솔루션은 AWS, 구글 클라우드, 마이크로소프트 애저, 오라클 클라우드 인프라(OCI)와 같은 주요 클라우드 서비스 제공업체와 델, HPE, 슈퍼마이크로와 같은 장비 제조업체의 엔비디아 AI 인프라에서 실행되어 설계 및 시뮬레이션 속도를 향상시킬 것이다.

젠슨 황은 “물리적 AI와 자율 AI 에이전트가 설계, 엔지니어링, 제조 방식을 근본적으로 재창조하는 새로운 산업 혁명의 시대가 도래했다.”라고 하면서 “엔비디아는 소프트웨어 대기업, 클라우드 제공업체, OEM으로 구성된 글로벌 생태계를 통합하여 모든 산업이 이러한 비전을 이전에는 불가능했던 규모와 속도로 현실화할 수 있도록 지원하는 풀 스택 가속 컴퓨팅 플랫폼을 제공하고 있다.”고 자랑했다.

차세대 차량 설계를 가속화하기 위해 지멘스 및 시놉시스와 협력하여 전산 유체 역학(CFD) 및 전자기학을 위한 GPU 가속 도구를 제공하고, 혼다는 시놉시스의 앤시스 플루언트(Ansys Fluent)를 그레이스 블랙웰 플랫폼으로 가속하여 CPU를 사용할 때보다 34배 빠른 속도로 공기역학 시뮬레이션을 실행함으로써 개발 주기를 단축한다.

JLR과 메르세데스-벤츠는 엔지니어링 워크플로우를 혁신하기 위해 엔비디아 가속 인프라에서 지멘스의 심센터 STAR-CCM+ 소프트웨어를 활용한다. 다쏘 시스템즈의 SIMULIA 아바쿠스와 파워플로우는 엔비디아 AI 인프라의 가속을 통해 리비안의 차량 시뮬레이션 테스트를 지원하는 데 사용한다.

항공기 이륙 시뮬레이션과 같은 매우 복잡한 CFD 시뮬레이션은 케이던스의 기술을 사용하며, 어센던스는 케이던스 기술로 하이브리드 전기 추진 및 수직 이착륙 항공기 시나리오를 시뮬레이션하고 있다. 이를 통해 대규모 CPU 기반 고성능 컴퓨팅으로는 불가능했던 전체 공기역학 시뮬레이션 캠페인을 당일 완료할 수 있다.

에너지 업계 선두 기업들은 시뮬레이션 처리 시간 단축, 처리량 증대, CPU 전용 컴퓨팅의 한계 극복을 위해 클라우드 및 온프레미스 환경 전반에 걸쳐 엔비디아 AI 인프라 기반의 GPU 가속 CFD 워크플로우를 도입하고 있으며, 이를 통해 더욱 깨끗한 에너지 솔루션을 위한 가스 터빈 혁신을 가속화하고 있다.

삼성과 SK하이닉스는 엔비디아 가속 델 파워엣지 서버와 HPE 시스템에서 케이던스 페가수스, 시놉시스 프라임심, 지멘스 캘리브리 소프트웨어를 사용하여 대용량 컴퓨테이셔널 리소그래피 및 물리적 검증을 간소화하고 DRAM 및 플래시 메모리 생산을 가속화하고 있다. TSMC는 첨단 제조 분야의 핵심 워크로드를 가속화하기 위해 HPE 및 슈퍼마이크로 시스템에서 시놉시스 도구를 사용하고 있다.

또한 산업 생태계 파트너들은 고정밀 산업용 디지털 트윈을 활용하여 가상 계획과 실제 실행을 연결함으로써 제품 라인, 공장, 창고 및 조선소의 디지털화를 가속화하고 있다. 지멘스가 새롭게 출시한 디지털 트윈 컴포저는 엔비디아 옴니버스 라이브러리를 활용하여 폭스콘, HD 현대, 펩시코, 키온과 같은 기업들이 대규모 산업용 메타버스 환경을 구축할 수 있도록 지원한다.

그림 4 지멘스의 디지털 트윈 컴포저 [출처: 지멘스]

카이온(KION)은 지멘스, 엔비디아, 액센츄어와 협력하여 자율 창고 솔루션을 발전시키고 있는데, 엔비디아 옴니버스와 액센츄어가 개척한 피지컬 AI 기반 디지털 트윈 및 시스템 아키텍처를 활용하여 세계 최대의 위탁 물류 제공업체인 GXO를 위해 젯슨 기반 자율 지게차 차량을 교육하고 테스트하는데 필요한 대규모의 물리적으로 정확한 창고 디지털 트윈을 구축했다.

엣지에서 오픈 소스 모델을 실행하는 젯슨(Jetson)

엔비디아 젯슨은 엣지에서 오픈 소스 모델을 실행하는 데 널리 사용되는 플랫폼이 되었다. 이 플랫폼은 다양한 오픈 모델과 AI 프레임워크를 지원하여 개발자가 엣지 환경에서 거의 모든 생성형 AI 워크로드에 대한 유연성을 확보할 수 있도록 한다.

산업계가 경직된 자동화에서 피지컬 AI로 나아가면서, 복잡한 환경에서 자율적이고 안전에 중요한 기계를 구동하기 위해 실시간 감지 및 추론이 가능한 새로운 유형의 지능형 엣지 컴퓨팅 장치가 필요하다. 이러한 수요를 충족하기 위해 IGX 토르를 정식 출시했다. IGX 토르는 고속 센서 처리, 엔터프라이즈급 신뢰성 및 기능 안전성을 통해 엣지에서 실시간 피지컬 AI를 제공하는 강력한 산업용 플랫폼이다.

캐터필러는 작업자 생산성과 안전성을 향상시키기 위해 IGX 토르 기반의 차량 내 대화형 AI 비서를 개발하고 있으며, 히타치 레일은 IGX 토르를 사용하여 철도망에 고급 예측 유지보수 및 자율 검사 시스템을 배포하고 있다.

그림 5 IGX 토르 [출처: 엔비디아]

오린(Orin)부터 토르(Thor)에 이르기까지, 엔비디아 젯슨 제품군은 엔비디아 네모트론, 코스모스, 아이작 GR00T와 같은 모델은 물론 Qwen, 젬마, 미스트랄 AI, GPT-OSS, PI 등 점점 더 늘어나는 커뮤니티 모델들을 실행하는 데 널리 사용하고 있다. 젯슨은 컴퓨팅과 메모리를 시스템 온 모듈(SoM)로 통합하여 고객의 하드웨어 설계를 가속화하고, 개별 부품 방식보다 소싱 및 검증을 더 쉽게 만들었다.

캐터필드의 캣 AI 어시스턴트는 엔비디아 젯슨 토르에서 동작하며 (이미 CES에서 306 CR 미니 굴삭기에서 활용을 선보임), Qwen3 4B는 vLLM을 통해 로컬에서 지원하기 때문에 클라우드 연결이 필요 없다. 차량 내 AI 어시스턴트인 '캣 AI 어시스턴트'는 신뢰할 수 있는 기계 컨텍스트와 함께 음성 및 언어 모델을 로컬에서 실행하여 운전자 안내 및 안전 기능을 지원한다.

공급망 솔루션 기업인 키온 그룹은 IGX 토르와 헤일로스 아웃사이드-인 세이프티 워크플로우를 활용하여 "외부에서 내부로" 인식하는 기능을 구현하고 있다. 이는 AI 에이전트가 인프라에 설치된 카메라와 동적 가상 안전 펜스를 사용하여 자율 로봇의 기능 안전 메커니즘을 강화하는 것을 의미한다.

애질리티와 헥사곤 로보틱스는 안전한 휴머노이드 로봇에 실시간 AI 추론 및 다중 모드 센서 융합을 위해 IGX 토르를 도입하고 있으며, 존슨앤존슨은 IGX 토르를 자사의 폴리포닉 디지털 수술 플랫폼에 도입하여 수술실에 실시간 AI 추론 기능을 제공하고 있다.

플래닛 랩스는 IGX 토르를 도입하여 수 테라바이트에 달하는 다차원 위성 데이터를 더 낮은 비용으로 궤도상에서 실행 가능한 정보로 변환하고 있다. 또한 CERN의 연구원들은 IGX 토르를 사용하여 고급 물리 기반 AI 모델을 실행하고, 대규모 데이터 스트림을 높은 처리량으로 처리하고 있다.

의료 로봇 공학을 위한 오픈 소스 피지컬 AI 모델

이번 GTC에서 발표한 의료 로봇 공학을 위한 피지컬 AI 기술들은 개발자들이 차세대 의료 로봇을 통해 의료 서비스 제공 방식을 혁신하는 데 필요한 강력하고 개방적인 인프라를 제공한다. 여기에는 다음과 같은 것을 포함한다.

- 오픈-H: 세계 최대 규모의 의료 로봇 데이터 세트이다. 약 30여 개 기관과의 협력을 통해 구축된 Open-H는 실제 수술 과정의 다양성과 복잡성을 반영하며, 700시간 이상의 수술 영상을 제공하여 첨단 로봇 시스템 개발을 가속화하는데 기여한다.
- 코스모스-H: NVIDIA 코스모스 기반의 개방형 모델 제품군으로, 코스모스-H-서지컬을 포함하여

도메인별 물리 기반 합성 데이터를 대규모로 생성할 수 있다. 코스모스-H는 프롬프트, 참조 이미지 또는 비디오, 그리고 로봇 운동학 정보를 입력받아 수술 영상을 생성하는 세 가지 모델을 제공한다. 이를 통해 개발자는 실제 데이터셋에 실감 나는 시뮬레이션을 추가하고, 수술 환경의 미래 상태를 예측하여 로봇 정책을 평가할 수 있다.

- 그루트-H: 아이작 그루트 N 기반의 이 비전 언어 동작(VLA) 모델은 임상 작업을 설명하는 텍스트 명령을 처리하고 동작 토큰이라고 하는 동작 명령을 생성한다. 오픈-H로 학습된 이 모델은 의료 환경에서 복잡한 물리적 동작을 수행하도록 개발된 로봇을 학습 및 평가하는데 사용할 수 있다.

리오(Rheo): AI 의료 로봇 공학을 위한 '헬스케어를 위한 아이작' 개발자 프레임워크 내에서 제공되는 이 개발자 블루프린트은 개발자가 병원 환경을 물리적으로 정확하게 시뮬레이션할 수 있도록 지원한다. 임상 워크플로, 의료 기기 상호 작용, 인체 움직임 및 병원 물류를 시뮬레이션하여 자동화 전략을 안전하게 개발하고 대규모로 테스트하는 데 사용할 수 있다.

수술 로봇, 영상 진단 및 병원 자동화 분야의 선도 기업들은 이미 의료 로봇 분야에 엔비디아의 피지컬 AI 기술을 활용하고 있다. CMR 서지컬은 그루트-H 사전 학습을 돕기 위해 약 500시간 분량의 수술 비디오를 오픈-H에 제공하고 있다. 또한, 코스모스-H를 사용하여 물리적으로 정확한 합성 수술 데이터를 생성하고 새로운 로봇 수술 정책을 평가하고 있다.

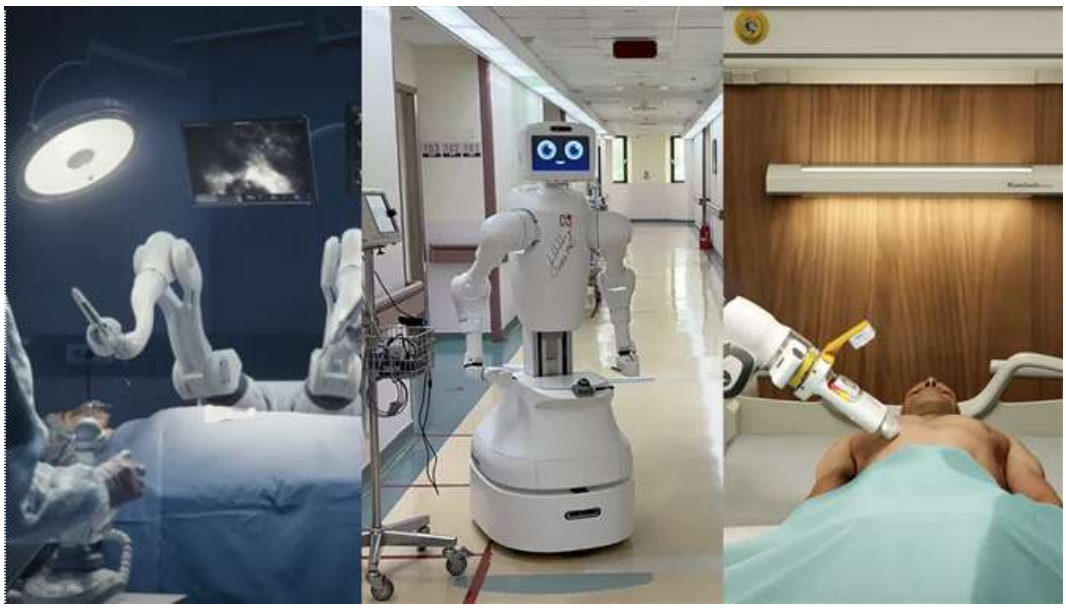


그림 6 헬스케어를 위한 아이작 플랫폼

존슨앤존슨 메드테크는 코스모스 기반의 파운데이션 모델과 AI 기반의 의학용 로봇 개발 플랫폼인 '아이작 포 헬스케어(Isaac for Healthcare)'의 해부학적 시뮬레이션을 활용하여 비뇨기과용 MONARCH 플랫폼의 교육 후 워크플로우에 필요한 데이터를 생성하고 개선하고 있다. 페리타스AI는 헬스케어를 위한 아이작과 리오를 사용하여 휴머노이드 로봇과 VLA 모델을 훈련시켜 수술 환경에 체화된 지능을 도입하고 있다. 이를 통해 실시간 상황 인식, 무균 조정, 기구, 임플란트 및 수술실 워크플로의 지능형 관리를 통해 수술팀을 지원한다. 이 연구는 라이트휠 및 어드벤트 헬스 병원과 협력하여 진행한다. 프락시미는 코스모스-H를 사용하여 수술실 이미지와 수술 중 비디오를 결합하는 멀티모달 비전 언어 모델을 학습시키고 있다. 이러한 모델은 수술 과정 전반에 걸쳐 수술 관련 통찰력과 조율을 제공하는 실시간 AI 에이전트를 구동한다.

오픈-H, 코스모스-H 및 그루트-H 는 이제 개발자가 특정 수술 시나리오에 맞게 사후 학습 및 조정할 수 있도록 깃허브 및 허깅 페이스에서 사용할 수 있다. 이러한 도구는 헬스케어를 위한 아이작 및 리오 블루프린트와 함께 수술 로봇의 학습 및 평가를 위한 시뮬레이션 파이프라인을 구축하는 데 활용한다.

나가며

GTC는 이제 과거 GPU 개발자를 위한 행사가 아니라 향후 일년 엔비디아의 주요 기술이 AI 분야에서 어떻게 발전하고 어떤 전략적 방향을 취할 것인가를 알 수 있는 매우 중요한 행사가 되었다. 특히 올해에는 피지컬 AI에 매우 중요한 방점을 찍고 있는 현재 엔비디아의 전략적 목표가 무엇이고 앞으로 어떤 기술 플랫폼을 키우고 육성할 것인지 볼 수 있는 기회였다.

옴니버스, 코스모스가 계속 업데이트되고, 실제 현장에서 개발자들이 필요로 하는 것이 무엇인지를 빨리 확인하면서 이를 위한 도구들을 빠르게 내놓는 것을 보면 엔비디아가 칩 회사가 아니라 소프트웨어 기업이 아닌가 하는 생각이 들 정도이다.

우리나라의 피지컬 AI 전략 수립에 있어 엔비디아는 매우 중요한 참고 모델이며, 대기업들에게는 핵심 파트너가 될 수밖에 없는 현실도 직시하면서 우리 정책 방향을 수립해야 할 것이다.

03 LLM 기업들의 전략적 B2B 피벗

| 김영욱 SAP Product Engineering Product Expert

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

앤쓰로픽 매출이 2주 만에 36% 성장했다는 수치는 통상적인 분기 성장률도, 신제품 출시 직후의 일시적 스파이크도 아니다. 이것은 어떤 시스템이 임계점을 넘어섰을 때 나타나는 비선형적 폭발의 신호다. 불과 몇 달 전까지 3:1이던 오픈AI와 앤쓰로픽의 매출 격차가 2026년 3월 현재 1.3:1로 좁혀졌고,[11] 이 속도가 유지된다면 역전은 올해 말이 아니라 이번 봄의 이야기가 될 수도 있다.

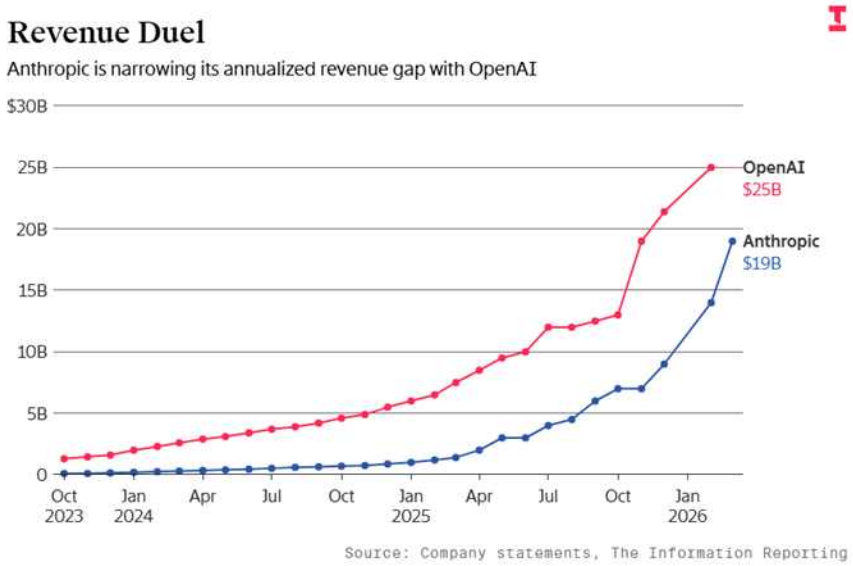


그림 7 앤쓰로픽과 오픈AI의 매출 추이 (출처: The Information)

이 변화를 단순히 두 회사 간의 경쟁 구도로 봐서는 안 된다. **AI가 소비자의 호기심을 자극하는 도구에서 기업의 핵심 인프라로 전환되는 구조적 이동이 가시화되고 있는 것이며, 그에 따라** LLM

11) The Information, "OpenAI Tops $25 Billion in Annualized Revenue as Anthropic Narrows Gap", Mar 4, 2026

기업들의 수익 모델, 판매 전략, 파트너십 구조가 근본적으로 재편되고 있다는 신호다. 마이크로소프트는 코파일럿 조직을 CEO 직속으로 재편했고, 구글은 제미나이를 워크스페이스 생태계 깊숙이 내장하는 데 속도를 올리고 있으며, 오픈AI는 뒤늦게나마 '사이드 퀘스트/프로젝트를 멈추라'는 내부 선언과 함께 B2B 피벗을 공식화했다.[12] 각자의 방식은 다르지만, 방향은 하나다. 기업 시장이다.

이 글에서는 그 전환의 실체를 이야기한다.

- 왜 소비자 AI 전략이 구조적 한계에 부딪혔는지,
- 앤쓰로픽은 어떤 전략으로 임계점을 돌파했는지,
- 오픈AI의 B2B 피벗이 진정한 전환점이 되기 위해 무엇이 필요한지, 그리고
- 마이크로소프트와 구글이 펼치는 플랫폼 전쟁의 본질이 무엇인지를 순서대로 살펴본다.

나아가 이 전쟁의 결과가 기업 구매자와 전략 기획자에게 어떤 함의를 갖는지를 짚어본다. AI 모델의 성능 경쟁은 이미 2막으로 넘어갔다. 지금 벌어지고 있는 것은 누가 기업의 핵심 시스템 안에 더 깊이, 더 넓게 뿌리를 내리느냐를 두고 벌이는 인프라 전쟁이다.

1. 소비자 AI에서 기업 AI로

숫자가 클수록 오히려 진실을 가리는 경우가 있는데, 챗GPT의 주간 활성 사용자 9억 2천만 명이 바로 그런 숫자다. 이 수치는 오픈AI가 소비자 시장에서 이룩한 전례 없는 성취를 보여주는 동시에, 왜 그들이 지금 전략을 근본적으로 바꿀 수밖에 없는 상황에 몰렸는지를 역설적으로 설명한다.

소비자 구독 모델의 수익 구조는 본질적으로 납작하고 확장에 한계가 있다. 아무리 많은 개인 사용자가 월 20달러짜리 구독료를 지불해도, 수십만 명의 직원에게 AI를 전사 배포하는 대형 엔터프라이즈 계약 하나가 만들어내는 매출과 장기적 락인 효과를 따라잡기 어렵다. 더욱이 소비자 시장은 이미 포화에 가까운 성숙 단계에 진입하고 있어서, 스스로 목표로 설정했던 주간 활성 사용자 10억 명 달성에도 실패하며 성장 곡선이 평탄화되는 징후가 뚜렷하게 나타나고 있다. **브랜드 인지도와 문화적 영향력은 소비자가 만들어주지만, 예측 가능하고 지속 가능한 매출 기반은 결국 기업 고객이 만들어준다는 현실**이 수면 위로 올라오고 있는 것이다.

12) The Wall Street Journal, “OpenAI to Cut Back on Side Projects in Push to ‘Nail’ Core Business”, Mar 16, 2026

코딩 에이전트가 열은 B2B AI의 문

앤쓰로픽이 이토록 빠른 속도로 엔터프라이즈 시장을 파고든 경로는 놀라울 만큼 단순하고, 그렇기 때문에 더욱 강력하다. 거창한 소비자 마케팅도, 수백 명의 기업 영업 조직도 아닌, **클로드 코드**라는 코딩 에이전트 하나가 그 폭발적 성장의 진원지였다는 사실은 엔터프라이즈 AI 시장의 침투 메커니즘에 대해 중요한 통찰을 제공한다.

엔터프라이즈 소프트웨어 시장에는 항상 내부로 들어가는 엔트리 포인트가 존재하며, 그 관문을 누가 통제하느냐가 시장의 판도를 결정해왔다. 과거 SaaS 시대에 그 관문이 영업사원과 구매 부서의 RFP 프로세스였다면, AI 시대의 관문은 개발자다. 개발자가 일상적인 작업 도구로 특정 AI를 채택하는 순간, 그것은 개인의 선택을 넘어 조직 전체의 기술 스택 결정으로 이어지는 강력한 경로가 된다. 개발자의 채택이 IT 부서의 검토를 촉발하고, 보안팀의 승인을 거쳐, 결국 전사 계약으로 확대되는 이 경로를 앤쓰로픽은 전략적으로 설계했고, 클로드 코드는 그 설계의 핵심 실행 도구였다. 2주 만에 ARR이 36% 급등한 것은 우연이 아니라, 이 침투 경로가 임계 질량에 도달했을 때 나타나는 필연적 결과다.

"사이드퀘스트(Side Quest)를 멈춰라"가 진짜 의미하는 것

오픈AI가 지난 1년간 발표하고 추진한 것들의 목록을 보면, 이것이 하나의 일관된 전략 아래 움직이는 조직인지 의심스러울 정도로 방향이 제각각이다. 비디오 생성 모델 소라, 독자 브라우저 아틀라스, 애플 출신 디자이너 조니 아이브와 협업한 하드웨어 디바이스, 챗GPT 내 쇼핑 기능 인스턴트 체크아웃, 엔비디아와의 전략적 파트너십, 그리고 펜타곤 계약까지, 한 회사가 동시에 추진하기에는 지나치게 넓은 전선이었다.

포트폴리오 전략은 리소스가 충분하고 시장이 안정적일 때 유효하다. 그러나 경쟁자가 하나의 분야에 모든 역량을 집중해 빠르게 시장을 잠식하고 있는 상황에서의 분산은 치명적 취약점이 된다. 인스턴트 체크아웃은 출시 5개월 만에 조용히 철수되었고, 엔비디아와의 1,000억 달러 투자 파트너십 발표는 불과 몇 주 만에 300억 달러 규모로 축소 수정되었으며, 급하게 체결한 펜타곤 계약은 내부 직원들의 반발과 외부 비판에 직면해 조건을 뒤늦게 수정해야 했다. 각각의 사건은 표면적으로는 개별적 실수처럼 보이지만, 패턴으로 읽으면 훨씬 더 깊은 문제가 드러난다. 전략적 우선순위가 불분명한 조직이 반복적으로 빠지는 '발표 먼저, 정제는 나중에' 문화가 고착되어 있다는 신호다.

오픈AI가 이번에 B2B로의 피벗을 공식 선언한 것은 경쟁 압박에 대한 반응이기도 하지만, 그보다 더 근본적으로는 분산된 전략의 폐해를 스스로 인정한 자기 교정의 시도로 읽힌다. 그러나 전략 선언이

하루 만에 가능한 것과 달리, 조직 문화와 실행 역량의 변화는 훨씬 더 긴 시간을 필요로 한다는 점에서, 이 피벗이 실질적 결과로 이어질지는 좀 더 지켜봐야 할 대목이다.

2. 앤쓰로픽의 B2B 퍼스트 전략

시장에서 후발주자가 선발주자를 이기는 방법은 두 가지다. 더 좋은 제품을 만들거나, 다른 게임을 하거나. 앤쓰로픽은 후자를 선택했다. 클로드를 챗GPT의 대안으로 포지셔닝하는 대신, 오픈AI에 의존하고 싶지 않은 기업들을 위한 선택지로 자신을 정의했다. 이 포지셔닝은 단순한 마케팅 전략이 아니라, 엔터프라이즈 구매 심리에 대한 정확한 독해에서 나온 것이다.

대형 기업의 기술 구매 담당자들이 가장 두려워하는 것은 나쁜 제품이 아니라 단일 벤더 종속(vendor lock-in)이다. 클라우드 시장에서 멀티 클라우드 전략이 표준이 된 것도 같은 이유였다. AI 시장에서도 동일한 논리가 작동하기 시작했고, 앤쓰로픽은 그 흐름을 정확히 포착했다. 헌법적 AI 프레임워크를 통해 구축한 안전성과 투명성에 대한 평판은, 기술적 차별화를 넘어 규제 환경이 까다로운 금융, 의료, 법률 분야 기업들에게 실질적인 도입 근거로 작용했다.[13] 성능 경쟁에서 이기려 하지 않고, 신뢰 경쟁에서 이기는 전략을 택한 것이다.

다이렉트 세일즈 보다는 채널 전략

앤쓰로픽의 성장 방식에서 가장 주목할 만한 점은 전통적인 엔터프라이즈 영업 조직을 최소화하면서도 대형 고객을 빠르게 확보했다는 사실이다. 그 **핵심에는 클라우드 하이퍼스케일러를 통한 채널 전략**이 있다. AWS 베드락과 구글 버텍스 AI를 주요 배포 채널로 삼음으로써, 앤쓰로픽은 수십 년에 걸쳐 이미 구축된 엔터프라이즈 신뢰 관계를 그대로 활용할 수 있었다.[14] 기업 고객 입장에서는 새로운 벤더와 계약하는 것이 아니라, 이미 사용하고 있는 클라우드 플랫폼 안에서 AI 기능을 추가하는 것이기 때문에 도입 장벽이 현저히 낮아진다.

여기에 마이크로소프트 365와 코파일럿로 모델 통합, 세일즈포스와의 파트너십 확장, 그리고 딜로이트를 통한 수십만 명 규모의 기업 내 배포까지 더해지면서, 앤쓰로픽은 직접 판매 비용을 최소화하면서도 엔터프라이즈 시장 깊숙이 침투하는 구조를 만들었다. 이것은 기존 SaaS 기업들이 수년에 걸쳐

13) Tech Research Online, "Why Enterprises Are Choosing Anthropic Over OpenAI in 2026", Mar 19, 2026
14) SaaStr, "How Anthropic Rocketed to $4B ARR", Jul, 2025

구축해온 채널 생태계를 AI 시대에 맞게 재해석한 전략이며, 스스로 생태계를 만드는 대신 이미 존재하는 생태계 위에 올라타는 방식의 영리한 실행이다.

3. 오픈AI의 뒤늦은 엔터프라이즈 피벗

오픈AI가 내놓은 B2B 피벗의 구체적 형태는 흥미롭다. 챗GPT, 코딩 에이전트 코덱스, 그리고 브라우저 아틀라스를 하나로 묶는 데스크탑 '슈퍼앱' 전략이 그것이다.[15] 사이드 퀘스트를 정리하겠다고 선언한 직후, 기존에 추진하던 세 가지 제품을 통합하는 방식으로 방향을 재설정한 것은 전략적으로 합리적인 선택이다. 개별 제품의 파편화가 만들어온 혼선을 통합 플랫폼으로 극복하겠다는 논리이고, 사용자가 하나의 인터페이스 안에서 대화, 코딩, 브라우징을 모두 처리할 수 있다면 엔터프라이즈 환경에서의 채택 장벽도 낮아질 수 있다.

그러나 이 전략에는 검증되지 않은 전제가 있다. 엔터프라이즈 시장에서 슈퍼앱이 실제로 작동한 선례가 많지 않다는 점이다. 기업 환경은 소비자 환경과 달리 이미 수십 개의 전문화된 툴이 각자의 역할을 맡고 있고, 새로운 통합 플랫폼이 그 복잡한 생태계를 대체하려면 단순한 기능 통합 이상의 깊은 시스템 연동이 필요하다. **슈퍼앱이 시장에서 진정한 차별점을 갖추려면, 기능의 통합보다 워크플로우의 통합이 선행되어야 한다**는 과제가 오픈AI 앞에 놓여 있다.

정부 계약의 전략적 활용

오픈AI가 B2B 확장을 위해 선택한 또 하나의 경로는 **정부 계약을 민간 기업 영업의 레퍼런스**로 활용하는 전략이다. AWS와의 계약을 통해 미국 정부 기관에 AI를 공급하는 구조를 만들고, 이를 기반으로 대형 민간 기업들에게 신뢰성을 증명하는 방식은 팔란티어가 수년에 걸쳐 검증한 플레이북이다. 실제로 팔란티어는 군사 계약으로 쌓은 신뢰를 발판 삼아 민간 부문에서 약 20억 달러의 매출을 올리는 기업으로 성장했다.

이 전략의 핵심 논리는 단순하다. 정부, 특히 국방부나 정보기관이 사용하는 기술은 보안성과 안정성 측면에서 이미 검증되었다는 신호를 시장에 보낸다. 대형 금융기관이나 헬스케어 기업처럼 데이터 보안과 규제 준수에 민감한 조직들에게, 정부 레퍼런스는 긴 기술 검증 프로세스를 단축시키는 강력한 신뢰 신호로 작동한다. 다만 이 전략이 온전히 작동하려면 정부 계약 자체의 안정성이 전제되어야 하는데, 앤쓰로픽과의 펜타곤 계약 갈등에서 드러났듯이 정부와의 관계는 그 자체로 불확실성의 변수를 내포하고 있다.

15) CNBC, "OpenAI to Combine ChatGPT, Codex and Browser Into SuperApp", Mar 19, 2026

IPO 압박과 전략 피벗의 불편한 관계

오픈AI의 이번 B2B 피벗을 순수하게 전략적 결단으로만 읽기 어려운 이유는 또 있다. 2026년 4분기로 예상되는 IPO 일정이 이 모든 움직임의 배경에 깔려 있기 때문이다. **기업공개를 앞둔 시점에서 투자자들이 가장 주목하는 것은 소비자 사용자 수가 아니라 예측 가능한 기업 매출과 수익성으로의 경로**다. 오픈AI가 2025년에 약 80억 달러의 순손실을 기록했고, 2026년에도 대규모 현금 소진이 예상되는 상황에서, 엔터프라이즈 매출 비중을 높이는 것은 IPO 스토리를 강화하기 위한 필수 조건이기도 하다.

문제는 자본 시장을 향한 내러티브와 실제 제품 전략이 항상 같은 방향을 가리키지는 않는다는 점이다. 투자자를 위한 피벗 선언과 고객을 위한 제품 혁신은 서로 다른 시간표 위에서 움직이는 경우가 허다하다. 오픈AI가 이번 전략 전환에서 진정한 신뢰를 얻으려면, 분기 실적 발표가 아닌 실제 엔터프라이즈 고객의 깊은 채택 사례로 그것을 증명해야 한다. 코딩 에이전트 코덱스의 주간 활성 사용자가 연초 대비 4배 성장한 200만 명을 기록한 것은 긍정적인 신호지만, 이것이 기업 전사 계약으로 전환되는 속도가 진짜 성공 시험대가 될 것이다.

4. 구글과 마이크로소프트의 생태계 전략: 플랫폼 전쟁

구글의 B2B AI 전략은 오픈AI나 앤쓰로픽과 근본적으로 다른 출발점을 갖는다. 이미 수억 명의 기업 사용자가 매일 사용하는 지메일, 독스, 시트, 미트라는 생산성 플랫폼을 보유하고 있다는 사실이 그것이다. 구글이 제미나이를 독립적인 AI 제품으로 경쟁시키는 대신 워크스페이스 생태계 깊숙이 내장하는 전략을 선택한 것은, 이 자산을 가장 효율적으로 활용하는 방식이다. 마이크로소프트의 M365 전략과 동일한 방법으로 사용자가 AI를 찾아오게 만드는 것이 아니라, 사용자가 이미 있는 곳에 AI를 가져다 놓는 전략이다.

이 접근 방식은 채택률 측면에서 강력한 이점을 만들어낸다. 엔터프라이즈 AI 도입의 가장 큰 장벽은 사용자의 행동 변화를 이끌어내는 것인데, 워크스페이스에 내장된 제미나이는 그 장벽을 원천적으로 제거한다. 직원들이 기존에 하던 방식으로 일하면서 자연스럽게 AI를 경험하게 되는 구조이기 때문이다. 문서 중심 업무에서 미국 시장의 37%를 점유한 것[16]은 이 전략이 이미 시장에서 유효하게 작동하고 있음을 보여주는 수치다.

16) PowerDrill, "LLM Market Landscape 2025", Sep 15, 2025

마이크로소프트의 복합 전략: 오픈AI 투자자이자 앤쓰로픽 파트너

마이크로소프트의 포지션은 이 전쟁에서 가장 복잡하기에 흥미롭다. 오픈AI의 최대 투자자이자 애저를 통한 독점 클라우드 파트너로 출발했지만, 동시에 앤쓰로픽 모델을 M365와 코파일럿에 통합하는 파트너십을 체결했다. 경쟁 관계에 있는 두 AI 기업 모두와 깊은 관계를 유지하는 이 전략은, 표면적으로는 모순처럼 보이지만 실제로는 매우 정교한 헤징이다.

마이크로소프트의 본질적 목표는 특정 AI 모델의 승리가 아니라, 애저와 M365가 기업 AI의 필수 인프라로 자리 잡는 것이다. 어떤 모델이 시장에서 이기든, 그 모델이 마이크로소프트의 플랫폼 위에서 실행된다면 마이크로소프트는 승자가 된다. 이번 코파일럿 조직의 CEO 직속 재편은 이 전략을 더욱 강화하겠다는 의지의 표현이다. 코파일럿이 단순한 AI 기능 추가가 아니라 마이크로소프트 전체 제품 포트폴리오를 관통하는 핵심 레이어로 격상되고 있는 것이다.

플랫폼 전쟁의 본질: 락인(Lock-in)의 재설계

구글과 마이크로소프트가 공통적으로 추구하는 것은 결국 하나다. AI 시대의 새로운 락인 구조를 자신들을 중심으로 재설계하는 것이다. 과거 엔터프라이즈 소프트웨어 시장에서 락인은 데이터와 프로세스의 이전 비용에서 나왔다. SAP나 오라클에서 벗어나기 어려운 이유는 수십 년간 쌓인 데이터와 그것에 맞춰진 업무 프로세스 때문이었다. AI 시대의 락인은 여기서 한 층 더 깊어진다. 기업의 데이터, 워크플로우, 그리고 직원들의 작업 패턴까지 학습한 AI가 깊이 내장될수록, 다른 플랫폼으로의 이전은 기술적으로뿐만 아니라 조직적으로도 점점 더 어려워진다.

이것이 구글과 마이크로소프트가 LLM 기업들과의 경쟁에서 갖는 구조적 우위다. 오픈AI와 앤쓰로픽이 아무리 뛰어난 모델을 만들어도, 기업 고객의 일상 워크플로우와 데이터가 구글 워크스페이스나 M365 안에 이미 자리잡고 있는 한, 플랫폼 사업자들은 AI 전쟁에서 가장 유리한 고지를 선점하고 있다. 순수 LLM 기업들이 채널로서 이 플랫폼들을 활용해야 하는 현실은, 역설적으로 플랫폼 사업자들의 협상력을 지속적으로 강화시키는 구조를 만들어낸다.

5. 엔터프라이즈 구매자의 선택 기준

2024년까지 대부분의 기업에서 AI 도입은 일종의 탐색전이었다. 파일럿 프로젝트를 몇 개 돌려보고, 특정 부서에 제한적으로 배포해보고, 효과를 측정하는 단계였다. 그러나 2025년 말을 기점으로 엔터프라이즈 구매 패턴에 질적인 변화가 나타나기 시작했다. AI가 더 이상 실험 대상이 아니라 핵심 업무 프로세스에

깊이 연결된 인프라로 인식되기 시작한 것이다. 포춘 500 기업의 약 92%가 생성형 AI를 업무에 활용하고 있지만, 엔터프라이즈 전용 솔루션의 실제 도입률은 아직 5% 수준에 머물러 있다는 수치는 이 전환이 이제 막 본격적으로 시작되었음을 의미한다. 다시 말해, 시장의 대부분은 아직 열리지 않았다.

이 전환이 갖는 의미는 단순히 시장 규모의 문제가 아니다. 구매 결정의 주체와 기준이 바뀐다는 것이다. 실험 단계에서 AI 도입을 주도하는 것은 혁신에 관심 있는 현업 담당자나 개발자였다면, 인프라 단계에서는 CIO, CISO, 법무팀, 그리고 조달 부서가 함께 테이블에 앉는다. 구매 사이클이 길어지고, 요구되는 기준이 높아지며, 한 번 선택된 벤더는 쉽게 교체되지 않는다. 지금 엔터프라이즈 AI 시장에서 벌어지는 경쟁은, 이 인프라 단계의 표준 자리를 선점하기 위한 싸움이다.

엔터프라이즈 선택 기준의 3대 축: 컴플라이언스, 신뢰성, 거버넌스

엔터프라이즈 구매자들이 AI 벤더를 평가하는 기준은 소비자의 그것과 근본적으로 다르다. 소비자가 응답의 품질과 사용 편의성을 중심으로 판단한다면, 기업 구매자는 세 가지 축을 중심으로 평가한다. 컴플라이언스, 신뢰성, 그리고 거버넌스다.

컴플라이언스는 특히 규제 산업에서 결정적인 요소다. 금융기관이 고객 데이터를 AI 모델에 입력할 때, 헬스케어 기업이 환자 정보를 처리할 때, 법무팀이 기밀 계약서를 분석할 때, 해당 데이터가 어디서 처리되고, 어떻게 저장되며, 누가 접근할 수 있는지에 대한 명확한 답변이 없는 벤더는 검토 대상에서 제외된다. 앤쓰로픽이 헌법 AI를 기술적 특징이 아닌 엔터프라이즈 신뢰의 언어로 포지셔닝한 것은 이 맥락에서 정확한 판단이었다.

신뢰성은 성능의 일관성과 서비스 수준 협약(SLA)의 문제다. 엔터프라이즈 환경에서 AI가 핵심 워크플로우에 통합될수록, 모델의 응답 품질이 예측 불가능하게 변동하거나 서비스가 중단되는 상황은 단순한 불편함이 아니라 비즈니스 리스크가 된다.

거버넌스는 조직 내에서 AI의 사용 범위와 방식을 통제하고 감사할 수 있는 체계를 의미하는데, 이것이 갖춰지지 않은 AI 솔루션은 아무리 성능이 뛰어나도 대형 기업의 전사 채택을 이끌어내기 어렵다.

벤더 다변화 전략: 멀티 모델 시대

클라우드 시장의 역사는 엔터프라이즈 AI 시장의 미래를 예측하는 데 유용한 참조점을 제공한다. 초기 클라우드 시장에서 많은 기업들이 AWS에 올인했다가, 단일 벤더 종속의 리스크를 경험한 이후 멀티

클라우드 전략을 표준으로 채택한 흐름은 AI 시장에서도 동일하게 반복될 가능성이 높다. 실제로 많은 기업들이 오픈AI에 대한 의존도를 줄이기 위해 클로드를 도입하거나, 특정 유즈케이스에 따라 여러 모델을 혼용하는 멀티 모델 전략을 채택하기 시작했다.

이 흐름은 LLM 기업들에게 중요한 함의를 갖는다. 기업 고객이 멀티 모델 전략을 표준으로 삼기 시작하면, 특정 모델에 대한 배타적 락인이 어려워지는 동시에, 각 모델이 특정 유즈케이스에서 보여주는 차별적 성능과 신뢰성이 선택의 핵심 기준이 된다. 이것은 단순히 가장 강력한 모델이 시장을 독식하는 구조가 아니라, 각자의 강점 영역을 갖는 복수의 모델이 공존하는 생태계로의 전환을 의미한다. 기업 구매자 입장에서는 이 다양성이 협상력을 높이고 리스크를 분산시키는 긍정적 변화이지만, LLM 기업들 입장에서는 차별화 압력이 지속적으로 증가하는 환경을 의미한다.

6. B2B AI 시대, 기업이 준비해야 할 것

비즈니스 모델의 전환: Seat에서 Token으로

엔터프라이즈 소프트웨어 시장은 지난 20년간 좌석(seat) 기반 구독 모델을 중심으로 작동해왔다. 사용자 수에 비례해 라이선스를 판매하고, 연간 계약을 갱신하는 이 구조는 SaaS 산업 전체의 수익 예측과 기업 가치 평가의 기준이 되었다. 그러나 LLM 기업들이 주도하는 토큰 기반 과금 모델은 이 구조를 근본적으로 흔들고 있다. 사용자 수가 아닌 AI의 실제 사용량, 즉 처리된 작업의 복잡도와 규모에 따라 비용이 결정되는 소비 기반 모델로의 전환은, 기업의 AI 예산 편성 방식부터 벤더와의 계약 구조까지 모든 것을 바꾸기 시작했다.

이 전환이 기업 구매자에게 갖는 의미는 양면적이다. 한편으로는 초기 도입 비용이 낮아지고, 실제 사용량에 비례한 비용 지출이 가능해진다는 점에서 유연성이 높아진다. 그러나 다른 한편으로는 AI 활용이 깊어지고 자동화 범위가 넓어질수록 토큰 소비량이 기하급수적으로 증가할 수 있어, 비용 예측과 통제가 새로운 과제로 부상한다. 실제로 AI 에이전트가 복잡한 멀티스텝 워크플로우를 자율적으로 처리하기 시작하면, 단일 세션에서 소비되는 토큰의 양은 단순 챗봇과는 비교할 수 없는 수준이 된다.

기업 AI 도입 성숙도: 위험한 중간 지점

엔터프라이즈 AI 도입의 여정은 대략 네 단계로 구분할 수 있다. 특정 유즈케이스를 탐색하는 실험

단계, 검증된 유즈케이스를 조직 내에 배포하는 도입 단계, AI를 핵심 업무 시스템과 연결하는 통합 단계, 그리고 데이터와 피드백을 기반으로 AI 성능을 지속적으로 개선하는 최적화 단계가 그것이다. 현재 대부분의 기업들은 실험 단계를 막 벗어나 도입 단계의 초입에 위치해 있다.

이 중간 지점이 위험한 이유는 두 가지다.

첫째, 실험 단계에서 검증된 유즈케이스가 전사 도입 단계에서 예상치 못한 거버넌스, 보안, 통합 이슈에 부딪히며 좌초하는 경우가 빈번하다. 성공적인 파일럿이 반드시 성공적인 전사 배포로 이어지지 않는다는 현실은, AI 도입 프로젝트에서 가장 흔하게 목격되는 함정이다.

둘째, 이 단계에서 벤더와 플랫폼 선택을 잘못하면, 나중에 통합과 최적화 단계로 넘어갈 때 이전 비용이 급격히 높아진다. 지금 이 시점에 신중하고 전략적인 벤더 선택이 필요한 이유가 여기에 있다. AI 인프라의 선택은 3년, 5년 후의 조직 역량을 결정하는 장기적 투자 결정이다.

제품 리더와 전략 기획자를 위한 액션 어젠다

LLM 기업들의 B2B 전쟁이 본격화되는 지금, 기업 내부의 제품 리더와 전략 기획자들이 직면하는 가장 중요한 과제는 이 전쟁의 결과를 지켜보는 것이 아니라 그 흐름 속에서 자신의 조직에 맞는 포지션을 능동적으로 설계하는 것이다. **몇 가지 핵심 관점을 제시**한다.

첫째, LLM 벤더 선택을 기술 결정이 아닌 전략 결정으로 다루어야 한다. 어떤 모델이 지금 당장 벤치마크 성능이 높은지보다, 어떤 벤더가 자사의 산업 규제 환경, 데이터 거버넌스 요구, 그리고 장기적 기술 로드맵과 정렬되는지를 중심으로 평가해야 한다. 성능은 빠르게 평준화되지만, 신뢰와 통합의 깊이는 쉽게 복제되지 않는다.

둘째, 지금 당장 내부 AI 거버넌스 체계를 수립해야 한다. 어떤 데이터를 AI에 입력할 수 있는지, 누가 AI 사용을 승인하고 모니터링하는지, AI가 생성한 결과물의 책임 소재는 어디에 있는지에 대한 명확한 원칙이 없는 상태에서의 AI 확산은 나중에 훨씬 더 큰 비용으로 돌아온다. 거버넌스는 AI 도입을 늦추는 장애물이 아니라, 지속 가능한 확산을 가능하게 하는 기반이다.

셋째, 멀티 모델 전략을 처음부터 고려해야 한다. 단일 LLM 벤더에 전사 AI 전략을 의존하는 것은 과거 단일 클라우드 벤더에 올인했던 실수를 반복하는 것이다. 유즈케이스별로 최적화된 모델을 선택하고, 벤더 간 경쟁을 통해 협상력을 유지하는 구조를 처음부터 설계하는 것이 장기적으로 유리하다.

7. 마무리

LLM 기업들의 B2B 전쟁은 이제 막 2막에 접어들었다. 1막이 누가 더 강력한 모델을 만드느냐를 두고 벌인 기술 경쟁이었다면, 2막은 누가 기업의 핵심 시스템 안에 더 깊이, 더 넓게 뿌리를 내리느냐를 두고 벌이는 인프라 전쟁이다. 앤쓰로픽은 신뢰와 채널 전략으로 이미 상당한 고지를 점령했고, 오픈AI는 뒤늦게 전열을 가다듬으며 추격에 나섰고, 구글과 마이크로소프트는 플랫폼 사업자로서의 구조적 우위를 바탕으로 이 전쟁의 심판이자 참여자로 자리잡고 있다.

그러나 **이 전쟁의 최종 승자는 가장 뛰어난 AI 모델을 만든 기업이 아닐 가능성이 높다. 기업 고객의 워크플로우에 가장 깊이 통합되고, 가장 높은 전환 비용을 만들어내며, 고객의 성공과 자신의 매출 성장을 가장 긴밀하게 연결한 기업이 이 전쟁의 승자가 될 것**이다. 그리고 그 승자를 결정하는 것은 결국 기업 구매자들의 선택이다. AI 모델의 성능이 빠르게 평준화되는 세계에서, 신뢰와 통합의 깊이가 새로운 해자가 되는 시대가 오고 있다.

04 2026년 클라우드 솔루션 리포트: 보안

| 정채상 메가존클라우드 수석 엔지니어

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

들어가며: AI 에이전트 시대, 보안의 대상이 달라졌다

2026년 초, 60일 만에 깃허브 스타 25만 개를 돌파하며 리액트(React)의 기록을 넘어선 오픈소스 AI 에이전트 프레임워크 오픈클로(OpenClaw)에서 대규모 보안 사고가 터졌다. 에이전트가 외부 도구와 연결되는 스킬 마켓플레이스 클로허브(ClawHub)에 340개 이상의 악성 패키지가 유포되었고, 인터넷에 노출된 인스턴스는 82개국에 걸쳐 4만 대 이상에 달했는데, 이는 AI 에이전트 인프라를 대상으로 한 최초의 대규모 공급망 공격이었다.[17)]

같은 시기, 보안 연구진이 분석한 7,000개의 MCP(Model Context Protocol) 서버 중 36.7%가 SSRF(Server-Side Request Forgery) 취약점을 갖고 있다는 리포트가 발표되었다.[18)] MCP는 AI 에이전트가 외부 데이터베이스, API, 파일 시스템과 연결되는 표준 프로토콜인데, 그 연결 인프라 자체가 공격 경로가 된 것이다. 보안의 대상이 서버와 네트워크를 넘어 AI 에이전트와 그 연결 인프라로 확장되고 있음을 보여주는 사건이다.

공격의 속도도 달라졌다. 크라우드스트라이크의 2026 글로벌 위협 리포트에 따르면, 공격자가 최초 침투 후 내부 횡이동(Lateral Movement)을 완료하기까지 걸린 평균 시간은 29분으로 전년 대비 65% 단축되었으며, 가장 빠른 사례는 불과 27초였다.[19)] 공격자들은 이미 생성형 AI를 활용해 90개 이상의 조직에서 자격 증명을 탈취하거나 악성 명령을 생성하고 있으며, 마이크로소프트 365 코파일럿을 대상으로 한 제로클릭 프롬프트 인젝션은 사용자의 클릭 한 번 없이도 기업 데이터를 유출할 수 있음을 증명했다.[20)] 방어자가 AI를 쓰지 않으면 속도에서 이길 수 없는 시대가 된 것인데, 이와 동시에, 기업

17) https://particula.tech/blog/openclaw-security-crisis-malicious-ai-agents
18) https://www.darkreading.com/application-security/microsoft-anthropic-mcp-servers-risk-takeovers
19) https://www.crowdstrike.com/en-us/press-releases/2026-crowdstrike-global-threat-report/

내 서비스 계정, API 키, AI 에이전트 등의 비인간 ID가 인간 ID를 압도적으로 초과하는 상황에서, AI 에이전트의 도입 속도가 보안 통제를 앞지르고 있다.

시장도 격변 중이다. 구글이 클라우드 보안 기업 위즈(Wiz)를 $320억에 인수 완료한 것은 역대 최대 사이버보안 M&A이자 구글 역사상 최대 규모이다.[21] 크라우드스트라이크는 비인간 ID 보안 스타트업을 인수하며 AI 에이전트 시대에 대비하고 있고, 팔로알토 네트웍스는 플랫폼화 전략으로 네트워크·클라우드·보안 운영 센터(SOC)를 하나로 묶고 있는 등 보안 솔루션 시장이 통합과 재편 중이다. 가트너에 따르면 글로벌 정보 보안 지출은 2026년 $2,400억에 이를 전망인데,[22] 기업들은 평균 50~70개의 보안 도구를 운영하면서도 사고는 줄지 않고 있다. 이제 질문은 "무엇을 더 사야 하나"가 아니라, "이미 갖고 있는 것에서 출발해서 어떻게 줄이고 통합할 것인가"로 바뀌고 있는 것이다.

'보안'이라는 말이 포괄하는 범위는 넓다. 서버와 PC를 지키는 엔드포인트 보안, 방화벽과 VPN으로 대표되는 네트워크 보안, 출입 통제나 CCTV 같은 물리 보안, 그리고 개인정보와 규정 준수를 다루는 데이터 보안까지. 이 글에서는 그 중 클라우드 인프라와 워크로드를 보호하는 영역에 집중하되, 엔드포인트와 네트워크까지 클라우드와 맞닿는 영역을 함께 다룬다.

본 리포트는 먼저 2026년 클라우드 보안을 관통하는 세 가지 핵심 이슈를 짚고, AWS·애저·GCP가 기본 제공하는 내장 보안의 강점과 한계를 점검한 뒤, 그 한계를 채우는 전문 벤더 6개 — 크라우드스트라이크, 팔로 알토 네트웍스, 위즈, 지스케일러(Zscaler), 포티넷(Fortinet), 센티넬원(SentinelOne) — 를 심층 비교한다. 이는 단순한 도구 선택 가이드가 아니라, AI 시대의 보안 아키텍처를 어떻게 설계할 것인가에 대한 전략적 의사결정을 지원하는 기초 자료이다.

2026년 클라우드 보안의 핵심 이슈

개별 벤더를 논하기에 앞서, 2026년의 보안 시장을 지배하는 거시적 흐름을 짚을 필요가 있다. 과거에는 영역별 도구를 비교해 "무엇을 할 수 있는가"를 따졌다면, 이제는 "어떤 구조로 통합할 것인가", 그리고 "AI가 공격과 방어 양쪽에서 어떤 역할을 하는가"가 선택을 좌우한다.

20) https://www.hackthebox.com/blog/cve-2025-32711-echoleak-copilot-vulnerability
21) https://techcrunch.com/2026/03/11/google-completes-32b-acquisition-of-wiz/
22) https://www.gartner.com/en/newsroom/press-releases/2025-07-29-gartner-forecasts-worldwide-end-user-spending-on-information-security-to-total-213-billion-us-dollars-in-2025

CNAPP: 클라우드 보안의 대통합

2020년대 초반까지 클라우드 보안은 영역별로 파편화되어 있었다. 클라우드 설정 오류를 찾는 CSPM(Cloud Security Posture Management), 워크로드를 보호하는 CWPP(Cloud Workload Protection Platform), ID 권한을 관리하는 CIEM(Cloud Infrastructure Entitlement Management), 인프라 코드를 검사하는 IaC 스캐닝, 데이터 보안 태세를 관리하는 DSPM(Data Security Posture Management), 오픈소스 취약점을 추적하는 SCA(Software Composition Analysis)가 각각 별도 제품으로 존재했다. 이는 결과적으로 6개의 콘솔, 6개의 알림 채널, 6개의 청구서를 의미하는 것으로, 도구 운영 자체가 목적이 되는 역설이 발생했다.

CNAPP(Cloud-Native Application Protection Platform)는 이 모든 것을 하나의 플랫폼으로 통합하여 "우리 클라우드에 구멍이 열려 있는가?"라는 단일 질문에 답하려는 시도이다. CNAPP 시장에서 가장 뜨거운 논쟁은 에이전트 vs 에이전트리스(Agentless) 아키텍처이다. 에이전트리스 방식의 대표 주자인 위즈는 클라우드 API와 디스크 스냅샷을 분석해 30분 내에 전체 환경의 가시성을 확보한다. 설치할 것이 없으니 마찰도 없지만, 스냅샷은 "어제의 상태"를 분석하는 것으로, 지금 이 순간 컨테이너 안에서 벌어지는 런타임 위협에는 대응할 수 없다.

반대편에 선 크라우드스트라이크는 서버와 컨테이너에 경량 에이전트를 설치하여 런타임 위협을 실시간으로 탐지하고 차단한다. 팔로알토 네트웍스는 두 방식을 모두 제공하는 하이브리드 접근을 취하지만, 인수로 모은 제품들의 통합 완성도가 과제로 남아 있다. 결국 "어느 쪽이 맞다"의 문제가 아니라, "가시성 확보(에이전트리스) → 런타임 보호 추가(에이전트)"라는 순서로 해당 문제들에 접근하는 것이 현실적인 전략이다.

AI 보안 관제(AI SOC): 사람이 알림을 보던 시대의 끝

보안 관제 센터(SOC, Security Operations Center)의 현실은 더 어려운데, 하루 평균 4,000건 이상의 알림이 쏟아지는데 상당수가 오탐 처리에 소모되며,[23] 글로벌 사이버보안 업계는 숙련된 분석가를 구하기도 어렵고, 설령 확보하더라도 수많은 알림 속에서 번아웃에 빠지기 쉽다. 앞서 언급한 것처럼 공격자의 횡이동 시간이 평균 29분, 최단 27초까지 줄어든 상황에서, 사람이 알림을 하나씩 읽고 판단하는 구조로는 대응이 불가능하다.

23) https://www.helpnetsecurity.com/2026/02/24/socs-autonomous-security-operations-strategies/

2026년, AI가 이 구조를 바꾸고 있다. AI를 도입한 SOC는 오탐을 80% 줄이고 대응 시간을 60% 단축하는 것은 물론, 침해가 발생했을 때 봉쇄까지 걸리는 시간을 108일이나 앞당기는 성과를 보이고 있다.[24] 여기서 핵심은 단순한 패턴 매칭이 아니라, 초급 분석가가 수행하던 알림 분류와 초기 분석, 증거 수집, 상관관계 파악을 AI가 자율적으로 처리하고 인간 분석가는 AI가 걸러낸 중요한 건에만 집중하는 구조로의 전환이라는 점이다. 분석가의 역할이 "알림을 보는 사람"에서 "AI의 판단을 검증하는 사람"으로 바뀌고 있다.

그러나 자율성의 확대는 곧 새로운 리스크를 의미한다. AI의 환각이 자동 대응으로 연결될 경우 멀쩡한 프로덕션 서버를 잘못 격리하거나, 잘못된 방화벽 정책을 적용하여 대규모 서비스 중단을 초래할 수 있다. "더 똑똑한 AI"만큼 "더 안전한 AI"가 중요하게 되는데, 어떤 플랫폼이 더 많은 자동화를 제공하는가보다, 어떤 플랫폼이 더 신뢰할 수 있는 자동화를 제공하는가가 2026년 보안 플랫폼의 성숙도를 가르는 기준이 되고 있다.

제로 트러스트와 비인간 ID: 경계가 사라진 세계

"회사 네트워크 안에 있으면 안전하다"는 전제, 이른바 '신뢰된 내부(Trusted Inside)' 모델의 한계는 이미 증명되었다. 2024년 스노우플레이크 침해에서 160개 기업, 5억 명의 데이터가 유출된 원인은 고도의 해킹 기술이 아니었다. 인포스틸러 악성코드로 탈취된 자격 증명과 MFA(다중 인증) 미설정이라는, 기본 중의 기본이 뚫린 것이다.[25] 침해된 계정의 80% 이상이 MFA를 설정하지 않은 상태에서 VPN으로 "안에 있는 것처럼" 접속하면 내부 전체가 열리는 구조에서, 한 번 뚫리면 횡이동을 막을 방법이 없었다.

제로 트러스트는 이 구조를 뒤집는다. "아무도 믿지 않는다. 매번 검증한다." 모든 접근에 ID와 디바이스, 컨텍스트를 확인하며, 2026년에는 도입 여부가 아니라 얼마나 성숙하게 구현했느냐가 문제이다. SASE(Secure Access Service Edge)는 네트워크(SD-WAN)와 보안(SWG, ZTNA, CASB)을 클라우드에서 통합 제공하는 아키텍처로, 여기서도 단일 벤더 vs 최적 조합 논쟁이 이어진다.

그런데 2026년의 진짜 새로운 과제는 비인간 ID의 폭증이다. 지금까지 제로 트러스트는 "사람"의 접근을 검증하는 데 집중해 왔지만, AI 에이전트와 서비스 계정, API 키 등 비인간 ID가 기업 내에서 인간 ID의 40배에서 100배로 늘어나면서 상황이 완전히 달라졌다. 80.9%의 기술팀이 AI 에이전트를

24) https://www.ibm.com/reports/data-breach
25) https://cloudsecurityalliance.org/blog/2025/05/07/unpacking-the-2024-snowflake-data-breach

이미 테스트하거나 운영하고 있음에도, 이들을 독립적인 ID 주체로 취급하는 팀은 21.9%에 불과하며 45.6%는 여전히 공유 API 키로 에이전트를 인증하고 있다.[26] 이는 회사의 모든 직원이 하나의 출입 카드를 공유하는 것과 같다. AI 에이전트가 무엇에 접근할 수 있는가를 실시간으로 통제하는 것이 다음 제로 트러스트의 전선이다.

클라우드 내장 보안: 이미 갖고 있는 것들

전문 벤더를 논하기에 앞서, "지금 쓰고 있는 클라우드에 이미 무엇이 들어 있는가?"부터 점검해야 한다. 대부분의 기업은 이미 AWS, 애저, GCP 중 하나 이상을 사용하고 있는데, 세 CSP 모두 보안 기능을 기본 제공하며, 이름만 다를 뿐 역할은 같다. 전문 벤더에 대한 투자를 논하기 전에, 이미 보유한 자산의 활용도를 먼저 점검하는 것이 먼저이다.

영역	AWS	Azure	GCP
ID, 접근 관리	IAM, Organizations	Entra ID, RBAC	Cloud IAM, Organization Policy
위협 탐지	GuardDuty	Defender for Cloud	Security Command Center
태세 관리(CSPM)	Security Hub	Defender CSPM	SCC Premium
암호화, 키 관리	KMS, CloudHSM	Key Vault, Confidential Computing	Cloud KMS, Confidential VMs
네트워크 보안	Security Groups, WAF, Shield	NSG, Azure Firewall, DDOS Protection	Cloud Armor, VPC Service Controls
감사, 로그	CloudTrail, Config	Activity Log, Azure Policy	Cloud Audit Logs, Access Transparency
컨테이너 보안	ECR 스캐닝, EKS Pod Identity	Defender for Containers	GKE Security Posture, Binary Authorization

표 6 CSP별 내장 보안 기능 비교

CSP 내장 보안의 강점은 분명하다. 추가 비용 없이, 또는 저비용으로 기본 보안 태세를 확보할 수 있으며, 자기 클라우드에 대해서는 API 레벨까지 가장 깊은 가시성을 제공한다. 규정 준수 프레임워크(SOC 2, ISO 27001, ISMS-P 등) 매핑도 기본으로 제공되므로, 컴플라이언스의 출발점으로서는 충분하다.

26) https://www.gravitee.io/blog/state-of-ai-agent-security-2026-report-when-adoption-outpaces-control

그러나 세 CSP 모두 공통적인 한계를 갖고 있으며, 이 한계가 곧 전문 벤더가 존재하는 이유이다. 우선 공격 경로 분석이 없다. CSP의 태세 관리 도구는 설정 오류를 찾아내지만, "이 취약점을 타고 실제 침투가 가능한가"까지는 판단하지 못하는데, 이것이 CNAPP가 존재하는 이유이다. 또한 엔드포인트와 런타임 보호가 빠져 있어서, 클라우드 인프라 계층은 관찰하지만 컨테이너 안의 프로세스나 서버의 런타임 위협은 EDR/CWPP에 맡겨야 한다. 멀티클라우드 가시성도 한계인데, 두 개 이상의 CSP를 쓰는 기업에게는 단일 뷰가 없다. 여기에 VPN 대체와 전구간 제로 트러스트는 CSP의 영역 밖이어서 SASE 벤더가 담당하고, AI 워크로드의 보안 태세를 관리하는 AI-SPM(AI Security Posture Management) 기능은 아직 어떤 CSP에도 존재하지 않는다.

CSP 내장 보안은 "기본 체력"이다. 이것만으로 충분한 기업은 거의 없지만, 이것을 무시하고 전문 벤더만 도입하는 것도 낭비다. 내장 보안으로 기초를 깔고, 위의 다섯 가지 한계를 전문 벤더로 채우는 것이 현실적 접근이며, 다음 장에서 그 전문 벤더 6개를 하나씩 살펴본다.

주요 벤더별 솔루션 심층 분석

크라우드스트라이크: 단일 에이전트의 제왕과 신뢰 회복의 여정

크라우드스트라이크는 엔드포인트 보안 시장의 지배자이자, 2024년 글로벌 장애라는 전대미문의 사건을 겪고 신뢰를 회복해 나가는 벤더이다. 위협 인텔리전스의 깊이와 에이전트 기반 플랫폼의 넓이에서 여전히 독보적이며, AI 에이전트 시대에 비인간 ID 보안이라는 새로운 전선을 개척하고 있다.

핵심 강점: 위협 인텔리전스의 깊이

크라우드스트라이크의 통합 보안 플랫폼인 팔콘(Falcon)은 "하나의 경량 에이전트, 하나의 콘솔"을 핵심 철학으로 삼는다. EDR(Endpoint Detection and Response), CWPP, CIEM, DLP를 단일 에이전트로 통합하여 운영 복잡성을 줄인다. 이 벤더의 진정한 차별화는 기능의 넓이가 아니라 위협 인텔리전스의 깊이에 있는데, 170개 이상의 공격 그룹을 이름과 동기까지 추적하며, 주당 수조 건의 보안 이벤트 데이터를 수집한다. 이 실전 기반의 탐지 규칙과 데이터 규모가 AI 학습의 원천이 되어, 다른 벤더가 쉽게 따라올 수 없는 진입 장벽을 형성한다.

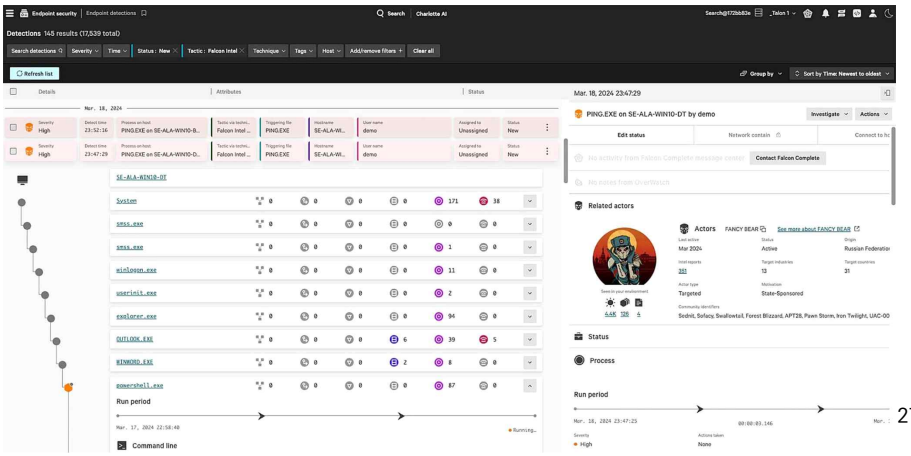
AI 시대의 접근

크라우드스트라이크의 AI 전략은 세 축에서 동시에 전개되고 있다. 첫째, AI로 보안을 강화하는

샬롯(Charlotte) AI이다. 자연어로 "지난 48시간 동안 관리자 권한으로 비정상적인 로그인이 있었어?"라고 물으면 위협 헌팅 결과를 즉시 반환하고, 사고 발생 시 타임라인을 자동 생성한다. 2025년부터 도입된 에이전틱 워크플로우즈는 탐지→조사→대응을 AI가 자율적으로 실행하는 구조인데, 현실적으로는 위협 헌팅과 사고 요약에서 유용성이 입증되었고, 복잡한 자율 대응은 아직 초기 단계이다. 둘째, AI 워크로드를 보호하는 역량이다. 팔콘 클라우드 시큐리티에 AI 워크로드 가시성을 추가하여, AI 모델 학습과 추론 파이프라인의 권한, 데이터 흐름, 모델 접근 패턴을 모니터링한다. 셋째, AI 에이전트 통제이다. SGNL 인수가 이 전략의 핵심으로, 비인간 ID의 접근 권한을 실시간 컨텍스트에 기반해 동적으로 관리한다.

사용자 경험 분석: 위협 그래프(Threat Graph)

크라우드스트라이크의 위협 그래프는 엔드포인트, 프로세스, 네트워크, ID 간의 연결을 실시간 그래프로 시각화한다. 하나의 의심스러운 프로세스를 클릭하면, 그 프로세스가 어떤 네트워크에 접속했고, 어떤 파일을 생성했으며, 어떤 ID로 실행되었는지가 연결선으로 펼쳐진다. 공격 체인을 한눈에 추적할 수 있는 이 인터페이스는 보안 분석가에게 강력한 무기이지만, 비기술 경영진에게는 복잡하게 느껴질 수 있다.



27)

그림 8 크라우드스트라이크 위협 그래프 예제

사용자들의 고충

2024년 7월 글로벌 장애는 여전히 크라우드스트라이크의 가장 큰 그림자이다. 단일 에이전트 모델의 양날의 검을 전 세계에 각인시킨 이 사건 이후, 크라우드스트라이크는 단계적 배포, 콘텐츠 검증 강화,

27) https://www.crowdstrike.com/wp-content/uploads/2024/09/S4_Cap-6_Intelligence-at-your-fingertips-scaled.jpg

커널 독립성 확대 등의 개선책을 발표했지만 신뢰 회복은 진행 중이다. 풀 플랫폼 도입 시 연간 수억 원에 달하는 높은 가격은 중소기업에게 진입 장벽이며, 위즈 대비 에이전트리스 가시성과 공격 경로 분석의 깊이가 약하다는 평가가 있다. 또한 SGNL, 바이오닉, 플로우 시큐리티 등 잇따른 인수 제품들의 매끄러운 통합이 남은 과제이다.

팔로알토 네트웍스: 가장 넓은 포트폴리오와 플랫폼화 전략의 승부수

팔로알토 네트웍스는 보안 시장에서 가장 넓은 포트폴리오를 보유한 벤더이다. 네트워크 보안, 클라우드 보안, SOC를 모두 자체 제품으로 커버하는 유일한 기업으로서, "보안에 필요한 모든 것을 한 벤더에서"라는 플랫폼화 전략을 업계에서 가장 적극적으로 추진하고 있다.

핵심 강점: 세 개의 축

팔로알토 네트웍스의 포트폴리오는 세 개의 독립적인 플랫폼으로 구성된다. 스트라타(Strata)는 차세대 방화벽 시장 1위와 프리즈마 SASE를 포함하는 네트워크 보안 축이고, 프리즈마 클라우드는 CSPM, CWPP, CIEM, DSPM을 통합한 CNAPP 축이며, 코텍스는 AI 기반 차세대 SIEM(보안 정보 및 이벤트 관리, 보안 정보와 이벤트 관리 XSIAM과 XDR, SOAR를 포괄하는 SOC 축이다. 이 세 축을 모두 자체 보유한 벤더는 업계에서 팔로알토 네트웍스뿐이다. 여기에 IBM QRadar SaaS 고객사를 확보하며 SIEM 시장에서의 입지도 강화했다.

AI 시대의 접근

팔로알토 네트웍스의 AI 전략에서 가장 주목할 제품은 코텍스 XSIAM이다. 전통적인 SIEM이 로그를 저장하고 규칙 기반으로 알림을 발생시키는 구조였다면, XSIAM은 SIEM 자체를 AI로 재발명한다. 데이터가 들어오는 순간 자동으로 파싱하고, 관련 알림을 그룹핑하며, AI가 자동 조사를 수행하고, 대응까지 자동화한다. AI 워크로드 보안 측면에서는 프리즈마 클라우드에 AI-SPM(AI Security Posture Management)을 추가하여, AI/ML 모델과 파이프라인의 보안 태세를 자동 평가한다. 학습 데이터 노출, 모델 탈취 위험 등을 탐지하며, 이 기능을 담당하는 프리즈마 AIRS의 고객 수가 분기 대비 3배로 증가했다. AI 에이전트 통제 측면에서는 프리시전 AI 브랜드 아래 AI 에이전트의 API 호출 패턴 이상 탐지와 네트워크 세그멘테이션을 AI 워크로드에 적용하고 있다.

사용자 경험 분석: 프리즈마 클라우드 대시보드

팔로알토 네트웍스의 프리즈마 클라우드 대시보드는 최근 공격 경로 분석의 시각화가 크게 강화되었다. 클라우드 자산 간의 연결과 취약점이 그래프로 표현되며, 인터넷에 노출된 공격 경로를 우선순위별로 제시한다. 그러나 3~4개의 인수 제품(Twistlock, Bridgecrew, Cider Security 등)을 하나의 콘솔에 통합하는 과정에서 UX의 일관성은 여전히 과제로 남아 있다.

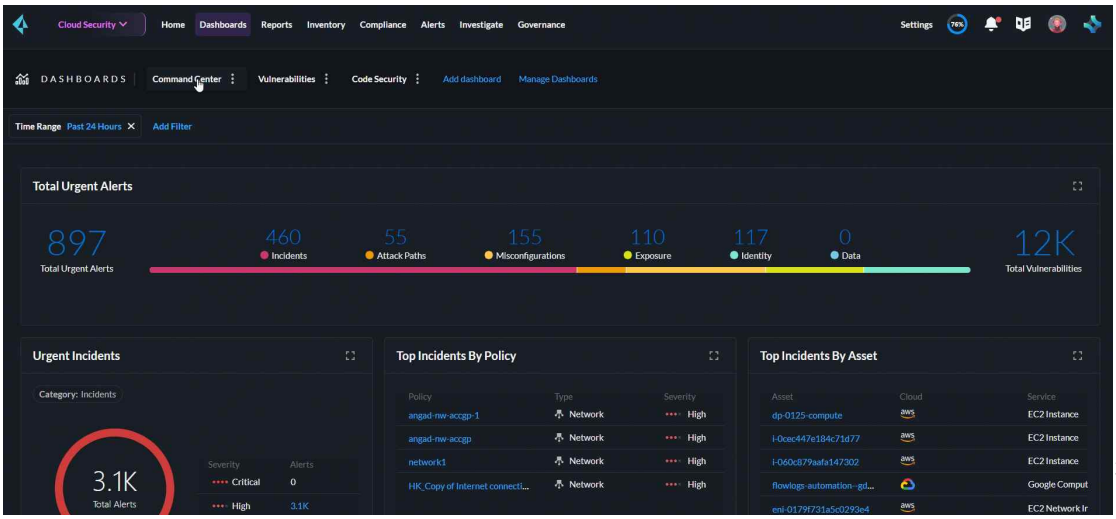


그림 9 프리즈마 클라우드커맨드 센터 대시보드 예제

사용자들의 고충

프리즈마 클라우드에 대한 가장 빈번한 비판은 "3~4개 인수 제품을 하나의 콘솔에 억지로 넣은 느낌"이라는 UX 비판이다. CSPM 알림의 노이즈가 높다는 평가도 많다. 설정 오류를 수천 건씩 보고하지만, 실제 위험한 것과 그렇지 않은 것의 구분이 직관적이지 않다는 것이다. 같은 맥락으로 코텍스에서 탐지한 위협이 프리즈마 클라우드의 자산과 어떻게 연결되는지, 세 플랫폼 간의 통합이 "직관적이지 않다"는 지적이 계속되고 있다.

위즈: 에이전트리스의 혁신자, 이제 구글의 무기

위즈는 2020년 창업 이후 클라우드 보안 시장에서 전례 없는 속도로 성장한 기업이다. 에이전트리스 아키텍처라는 명확한 기술적 철학, 직관적인 UX, 그리고 SaaS 역사상 최고속 성장(연간 반복 매출 $10억+)이라는 실적이 결합되어, 구글에 역대 최대 금액으로 인수되었다.

28) https://docs.prismacloud.io/en/enterprise-edition/content-collections/dashboards/dashboards-command-center

핵심 강점: 시큐리티 그래프와 에이전트리스의 미학

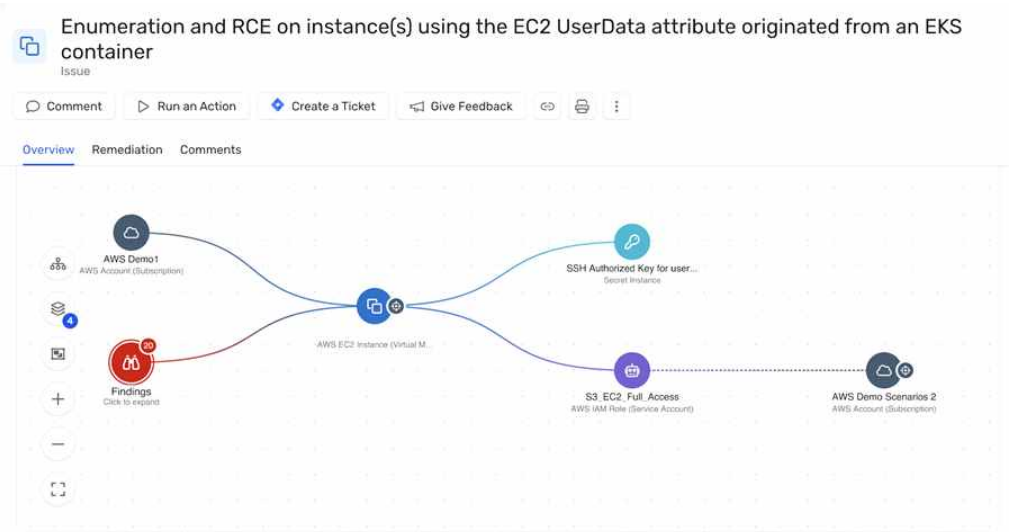
위즈의 핵심은 100% 에이전트리스 아키텍처이다. 클라우드 API와 디스크 스냅샷을 분석하여, 에이전트를 설치하지 않고도 30분 내에 전체 클라우드 환경의 가시성을 확보한다. 설치할 것이 없으니 운영팀과의 마찰도, 에이전트 충돌의 위험도 없다. 위즈의 킬러 기능인 시큐리티 그래프는 클라우드 리소스, 네트워크 경로, ID 권한, 소프트웨어 취약점, 민감 데이터의 위치를 하나의 그래프로 연결한다. 단순히 "이 VM에 취약점이 있다"고 알려주는 것이 아니라, "이 취약점이 있는 VM이 인터넷에 노출되어 있고, 과도한 권한을 가진 서비스 계정에 연결되어 있으며, 그 계정이 민감 데이터에 접근할 수 있다"는 공격 경로를 보여준다. 수만 개의 개별 알림 대신, 실제 공격 가능한 경로에 우선순위를 매기는 것이다.

AI 시대의 접근

위즈의 AI 전략은 세 가지 방향으로 전개된다. 첫째, AskAI는 시큐리티 그래프 위에서 자연어 질의를 가능하게 한다. "Log4Shell에 영향받는 인터넷 노출 자산이 몇 개야?"라고 물으면, 그래프를 탐색하여 즉시 답을 반환한다. 복잡한 쿼리 언어를 모르는 보안 담당자도 데이터를 탐색할 수 있게 하는 것이다. 둘째, AI-SPM은 세이지메이커, 버텍스 AI, 애저 ML 등 AI/ML 파이프라인의 보안 태세를 에이전트리스 방식으로 자동 평가한다. AI 모델의 학습 데이터가 어디에 저장되어 있는지, 누가 접근할 수 있는지, 모델 추론 엔드포인트가 인터넷에 노출되어 있는지를 자동으로 스캔한다. 위즈는 이 영역을 가장 먼저 제품화한 벤더이다. 셋째, 구글 인수 시너지이다. 구글 클라우드의 AI 인프라(Vertex AI, TPU)와 위즈의 보안 가시성이 결합되면, "AI를 만들고 보호하는 것"을 하나의 플랫폼에서 제공할 수 있다. 이는 구글 클라우드가 AWS, 애저 대비 보안에서 약하다는 인식을 위즈로 단번에 뒤집으려는 전략이기도 하다.

사용자 경험 분석: 시큐리티 그래프

위즈의 시큐리티 그래프는 공격 경로를 그래프로 연결하는 직관적 UX로, "CNAPP 중 UX가 가장 깔끔하다"는 평가를 받는다. 기술을 모르는 경영진도 리스크를 즉시 파악할 수 있도록 설계되어 있으며, CISO가 이사회에 보안 현황을 보고할 때 그래프 한 장으로 설명할 수 있다는 점이 도입을 촉진하는 요인으로 작용한다.



29)

그림 10 위즈 시큐리티 그래프 예제

사용자들의 고충

에이전트리스의 강점은 동시에 구조적 한계이기도 하다. 스냅샷 기반 분석은 "어제의 상태"를 보여줄 뿐이다. 지금 이 순간 컨테이너 안에서 실행되는 악성 프로세스, 메모리 상주형 공격에 대해서는 경쟁사들 대비 약하다. CNAPP 시장에서 가장 비싸다는 평가도 있고, 구글의 인수 이후의 불확실성이 멀티클라우드 고객에게 고민을 안겨주고 있다. 이에 위즈는 AWS와 애저 환경도 계속 지원하겠다고 밝혔으며, 인수 후에도 브랜드와 멀티클라우드 지원을 유지하겠다고 선언했다.

지스케일러: 제로 트러스트의 전용 설계

지스케일러는 처음부터 제로 트러스트만을 위해 설계된 기업이다. 방화벽 제조사가 클라우드로 확장하거나, 클라우드 기업이 보안을 추가한 것이 아니라, "모든 트래픽은 검증되어야 한다"는 단일 원칙 위에 전체 아키텍처를 구축했다. 모든 트래픽을 클라우드로 경유시키는 이 원칙은 강력한 통제력을 제공하지만, 네트워크 성능 지연과 멀티클라우드 환경의 복잡성이라는 숙제를 동시에 던진다.

핵심 강점: 세계 최대 인라인 보안 클라우드

지스케일러의 제로 트러스트 익스체인지는 매일 4,000억 건 이상의 트랜잭션을 처리하는 세계 최대

29) https://www.wiz.io/blog/wiz-security-graph-enhances-cloud-incident-response

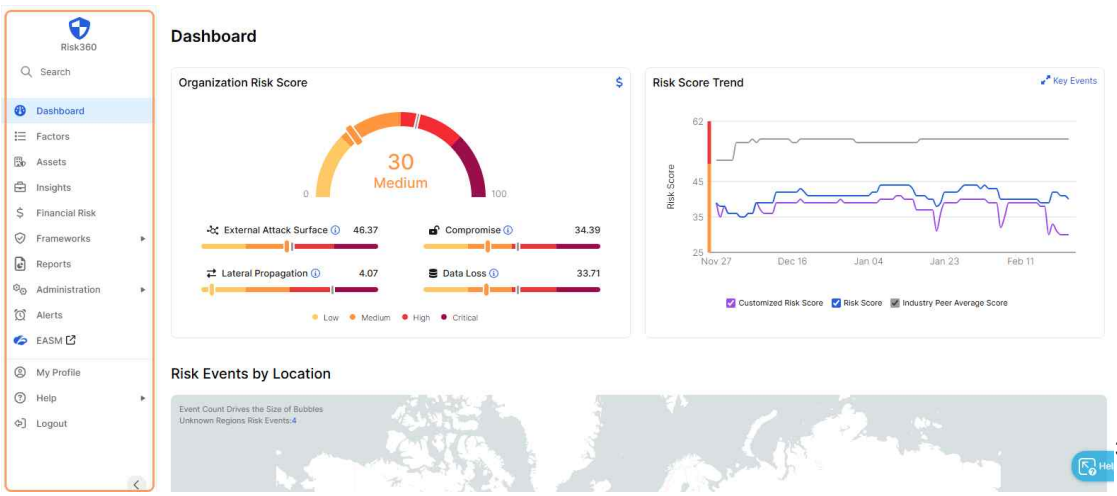
인라인 보안 클라우드이다. ZIA(Zscaler Internet Access)는 SWG(Secure Web Gateway)로 인터넷 접속을 보호하고, ZPA(Zscaler Private Access)는 ZTNA(Zero Trust Network Access)로 내부 앱에 대한 VPN 없는 접근을 제공하며, ZDX(Zscaler Digital Experience)는 사용자의 디지털 경험을 모니터링한다. 이 세 가지가 결합되어 VPN을 완전히 대체하는 가장 깨끗한 아키텍처를 제공한다. 이를 통해 SSL/TLS 암호화 트래픽을 대규모로 검사하는 검증된 역량을 보유한다.

AI 시대의 접근

지스케일러의 AI 전략은 네트워크 보안에 집중되어 있다. 코파일럿은 자연어로 보안 정책을 관리하고, Breach Predictor는 침해를 사전에 예측하며, Auto-Classification은 데이터를 자동으로 분류하여 DLP 정책을 적용한다. 가장 독특한 포지션은 AI 에이전트 트래픽의 관문 역할이다. AI 에이전트가 생성하는 API 호출을 인라인으로 검사하며, 매일 4,000억 건의 트랜잭션 데이터가 AI 이상 탐지의 학습 원천이 된다. AI 에이전트의 관문 역할을 하는 것으로, AI SOC보다는 AI를 네트워크 보안에 적용하는 데 집중하고 있다.

사용자 경험 분석: Risk360

지스케일러의 Risk360은 전사적 리스크를 점수화하여 단일 뷰로 제공한다. 사이버 리스크, 데이터 보호, 접근 보안 등의 영역별 점수와 추이를 경영진이 한눈에 파악할 수 있다. 경영진 보고용으로 유용하다.



30)

그림 11 Zscaler Risk360 예제

사용자들의 고충

지스케일러의 가장 큰 약점은 SD-WAN을 보유하지 않아 "진정한 단일 벤더 SASE"라 하기 어렵다는 점이다. SASE는 보안(SSE)과 네트워크(SD-WAN)의 통합인데, 보안 쪽만 보유하고 있어 SD-WAN 파트너와의 조합이 필요하다. 모든 트래픽을 프록시로 우회하는 구조에서 특정 지역이나 앱에서 체감 지연이 보고되며, 핀테크나 의료 앱에서 SSL 검사 시 인증서 오류가 반복된다는 불만도 있다. CNAPP 역량은 경쟁사 대비 제한적이어서, 클라우드 워크로드 보안을 해결하기는 어렵다.

포티넷: 하드웨어 가속의 가성비 챔피언

포티넷은 온프레미스와 네트워크 보안의 DNA를 가진 기업이다. 자체 설계 칩으로 무장한 하드웨어 기반의 가성비와, 방화벽에서 SD-WAN, SASE, OT(Operational Technology, 공장·발전소 등의 산업 설비를 제어하는 기술) 보안까지 아우르는 통합형 시큐리티 패브릭이 강점이다. 클라우드 네이티브 보안에서는 약하지만, 하이브리드 환경과 제조·에너지 산업에서는 대체할 수 없는 존재감을 보인다.

핵심 강점: 칩 레벨의 가성비

포티넷의 가장 독특한 차별화는 자체 설계한 FortiASIC 칩이다. 소프트웨어 기반 방화벽 대비 최대 10배의 처리량을 하드웨어 가속으로 달성하며, 같은 성능을 훨씬 낮은 가격에 제공한다. 방화벽 출하량 세계 1위, 고객 77만 5천 개로 업계 최대 고객 기반을 보유한다. FortiSASE는 자체 SD-WAN + NGFW + SWG + CASB + ZTNA를 통합하여 "진정한 단일 벤더 SASE"를 표방한다. OT/IoT 보안에서는 제조·에너지 산업의 ICS/SCADA 보안에서 독보적 위치에 있으며, 이 영역에서 경쟁자는 사실상 없다.

AI 시대의 접근

포티넷의 AI 전략은 네트워크 계층에 집중되어 있다. FortiAI는 네트워크 트래픽을 AI로 분석하여 제로데이 위협을 탐지하고, 시큐리티 패브릭을 통해 방화벽에서의 탐지가 엔드포인트 자동 격리로 이어지는 연계 대응을 구현한다. OT/IoT와 AI의 결합도 주목할 만한데, 제조·에너지 현장에서 산업 장비 위에 AI 추론이 돌아가는 시대가 열리고 있으며, OT 보안 리더로서 엣지 AI 워크로드를 네트워크 계층에서 보호하는 포지션을 취한다. 다만, 클라우드 네이티브 AI 워크로드(SageMaker, Vertex AI 등)의 보안 태세 관리(AI-SPM) 역량은 아직 초기 단계이다.

30) https://help.zscaler.com/risk360/accessing-and-navigating-risk-360-admin-portal

사용자 경험 분석: FortiManager

FortiManager는 수천 대의 FortiGate 방화벽을 중앙에서 정책을 관리하고 배포하며 모니터링하는 인터페이스이다. 대규모 분산 네트워크를 운영하는 엔터프라이즈에서 강점을 발휘한다.

31)

그림 12 FortiManager 예제

사용자들의 고충

포티넷의 가장 큰 약점은 보안 취약점의 반복 노출이다. 이른바 FortiJump로 불리는 사건인 (CVE-2024-47575)은 FortiManager의 인증 없는 원격 코드 실행 취약점으로, 국가급 공격자가 2024년 6월부터 악용해 50개 이상의 서버에서 관리 대상 장비의 설정 데이터를 유출했는데, 정부와 핵심 인프라가 주요 표적이었다.[32] "보안 장비 자체가 공격 표면"이 되는 역설이다. CLI 중심의 설정 방식은 "포티넷 전문가가 아니면 운영이 어렵다"는 평가를 받으며, CNAPP 경쟁사 대비 클라우드 네이티브 보안 기능이 부족하다. 온프레미스와 네트워크 중심의 DNA가 클라우드 네이티브 시대에 어떻게 진화할 것인지가 포티넷의 장기적 과제이다.

31) https://www.fortinet.com/demo-center/fortimanager-demo
32) https://bishopfox.com/blog/a-look-at-fortijump-cve-2024-47575

센티넬원: AI 네이티브 DNA와 열린 생태계

센티넬원은 창업부터 AI를 핵심 아키텍처에 내장한 기업이다. 다른 벤더들이 기존 제품에 AI를 추가하는 방식이라면, 센티넬원은 AI 기반의 자율 탐지·격리·복구 구조를 처음부터 설계했다. 2026년에는 "열린 생태계"라는 철학적 차별화로, 멀티 벤더 환경에서의 AI 분석 계층으로 포지셔닝하고 있다.

핵심 강점: AI 자율 엔드포인트와 개방형 XDR

센티넬원의 통합 보안 플랫폼인 싱귤래러티는 에이전트가 클라우드 연결 없이도 엔드포인트에서 독립적으로 위협을 판단하고 대응할 수 있도록 설계되었다. 마이터(MITRE) ATT&CK 평가에서 최고 수준의 탐지율을 기록했으며, 크라우드스트라이크 대비 가격 경쟁력 — 업계에서 "90%의 기능을 70%의 가격에"라는 평가 — 이 강점이다. 핑세이프(PingSafe) 인수(2024)로 CNAPP에 진출했으며, 개방형 XDR을 표방하여 서드파티 데이터 소스를 적극 수용한다.

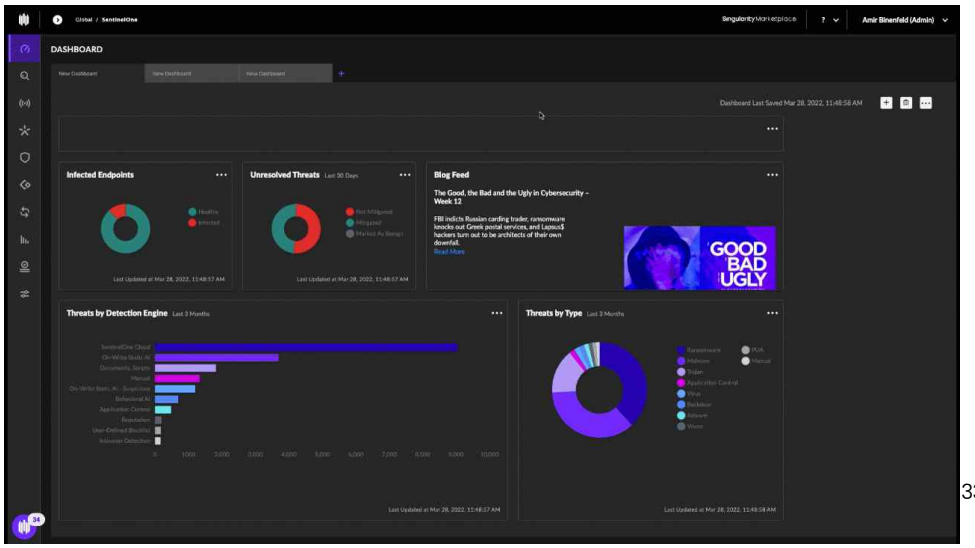
AI 시대의 접근

센티넬원의 AI 전략에서 가장 주목할 것은 퍼플(Purple) AI의 진화이다. 다른 벤더의 AI 어시스턴트가 자사 데이터만 분석하는 것과 달리, 퍼플 AI는 OCSF(Open Cybersecurity Schema Framework) 기반으로 타사 데이터를 정규화하여 즉시 쿼리하는데, 이는 "열린 AI SOC"라는 독특한 포지셔닝이다. AI 워크로드 보안 측면에서는 핑세이프 기반 CNAPP에 AI 워크로드 스캐닝을 추가했으나, 아직 초기 단계이다.

센티넬원의 가장 흥미로운 차별화는 철학적인 것인데, "우리 플랫폼에 가두지 않겠다"는 것으로, AI 시대에 단일 벤더로 모든 보안을 커버하는 건 불가능하다는 현실을 인정하고, 다른 벤더의 제품 위에 AI 분석 계층으로 얹히는 전략을 취한다. 이는 팔로알토 네트웍스의 "모든 것을 우리 안에서" 전략과 정반대의 접근이다.

사용자 경험 분석: 싱귤래러티 콘솔

싱귤래러티 콘솔은 엔드포인트, 클라우드, ID에서 발생하는 이벤트를 단일 타임라인으로 시각화한다. 특정 위협이 언제 엔드포인트에 도달했고, 어떤 프로세스를 생성했으며, 어떤 네트워크에 접속을 시도했는지를 시간순으로 추적할 수 있다.



33)

그림 13 싱귤래러티 예제

사용자들의 고충

CNAPP 진출은 아직 초기이며, 경쟁사 대비 기능의 깊이가 부족하다는 평가가 지배적이다. 특히 공격 경로 분석과 AI-SPM 영역에서 격차가 있다. 엔터프라이즈 시장에서 브랜드 인지도가 약하며, 자체 SIEM이 없어 스플렁크나 마이크로소프트 센티넬 등과의 연동이 필수적이다. 다만, 이 "자체 SIEM 부재"가 역설적으로 퍼플 AI의 개방형 전략과 맞물려, 기존에 스플렁크나 다른 SIEM을 사용하는 기업에게는 오히려 도입 장벽이 낮다는 평가도 있다.

맺으며: CIO/CTO를 위한 제언

2026년의 보안 전략 수립은 단순한 도구 구매가 아니라 조직의 보안 아키텍처를 재설계하는 과정이다. "무엇을 도입할 것인가"보다 중요한 질문은 "어떤 순서로, 어떤 조합으로, 어떤 수준까지 자동화할 것인가"이다. 이 글을 마무리하며 세 가지 제언을 드린다.

1. 도입 순서를 지켜라 — 구멍부터 막고, 감시하고, 접근을 바꿔라

CSP 내장 보안으로 기초를 깐 뒤, 전문 벤더는 순서대로 도입할 것을 권한다. 먼저 CNAPP를 도입하여 설정 오류와 공격 경로를 가시화해야 한다. 클라우드에 구멍이 열려 있는지조차 모르는

33) https://www.sentinelone.com/blog/feature-spotlight-introducing-singularity-dark-mode/

상태에서 다른 도구를 더하는 것은 비효율적이다. 구멍을 막은 뒤에는 EDR/XDR과 AI SOC를 도입하여 런타임 보호와 관제 자동화를 확보하고, 마지막으로 제로 트러스트와 SASE로 VPN을 대체하면서 비인간 ID까지 접근 통제 체계를 전환한다. 이 순서는 "가장 시급한 리스크부터"라는, 단순하지만 자주 무시되는 원칙에 기반한다.

2. 막고 싶은 위험부터 벤더를 선택하라 — 완벽한 단일 벤더는 없다

위 벤더 분석에서 확인했듯, 모든 영역을 완벽하게 커버하는 단일 벤더는 존재하지 않는다. "어떤 벤더가 좋은가"가 아니라 "우리 조직이 가장 먼저 막아야 하는 위험이 무엇인가"에서 출발해야 한다.

막고 싶은 위험	필요한 역량	살펴볼 벤더, 솔루션
클라우드 설정 오류, 공격 경로 방치	CNAPP – 에이전트리스 가시성, 공격 경로 분석	Wiz, Palo Alto Prisma Cloud
엔드포인트, 컨테이너 런타임 침입	EDR/CWPP – 실시간 탐지, 격리, 복구	CrowdStrike, SentinelOne
알림 폭주, 관제 인력 부족	AI SOC – 자동 분류, 조사, 대응	Palo Alto Cortex XSIAM, CrowdStrike(Charlotte AI), SentinelOne(Purple AI)
원격 접속 경로의 횡이동 위험	제로트러스트/SASE – VPN 대체, 인라인 검사	Zscaler, Fortinet FortiSASE
AI 에이전트, 서비스 계정의 무단 접근	비인간 ID 통제 – 동적 권한 관리	CrowdStrike(SGNL), Zscaler
공장, 발전소 등 산업 설비 네트워크 보안	산업 제어 시스템(ICS/SCADA) 보호	Fortinet

표 7 위험 유형별 보안 역량과 벤더 가이드

현실적인 조합은 CSP 내장 보안에 전문 벤더 3~4개를 더하는 것이다. 위 표에서 자기 조직의 상위 2~3개 위험을 선별하고, 해당 벤더를 먼저 검토하는 것이 출발점이다. 모든 위험을 한꺼번에 해결하려는 시도는 또 다른 복잡성을 만들 뿐이다.

3. AI에 대한 신뢰 경계를 명확히 설정하라

벤더들은 AI가 보안 관제를 스스로 해결한다고 홍보하지만, 2026년의 현실에서 100% 자율 운영은 아직 이르다. AI에게 분석과 제안을 맡기되 실행 권한은 단계적으로 부여해야 한다. 캐시 비우기나 특정

파드 재시작 같은 저위험 작업은 AI에게 위임하더라도, 데이터베이스 스키마 변경이나 방화벽 정책 변경은 반드시 인간 엔지니어의 승인을 거치도록 워크플로우를 설계해야 한다. AI 기반 보안 사고의 파급력은 일반 장애보다 훨씬 크다. 더 강력한 AI일수록, 더 엄격한 통제 설계가 선행되어야 한다.