디지털서비스

이슈리포트

Digitalservice Issue Report

EU와 일본의 새로운 소버린 클라우드 정책



NIA 한국지능정보사회진흥원 디지털서비스 이용지원시스템

2026 Vol. 10

디지털서비스 이슈리포트

CONTENTS

Digitalservice Issue Report

디지털서비스 이슈리포트

01 **EU와 일본의 새로운 소버린 클라우드 정책** 3

| 한상기 테크프론티어

02 **빅테크의 사내 AI 내재화와 모델 선택권 전쟁** 13

| 김영욱 SAP Product Engineering Product Expert

03 **3대 메가프로젝트의 AI 데이터센터 계획 – 성공적인 구현을 위해 필요한 것은?** 23

| 윤대균 아주대학교 소프트웨어학과

04 **2026년 클라우드 솔루션 리포트: 추론 인프라 (2) 토큰 공급** 34

| 정채상 이음테크

본 제작물은 한국지능정보사회진흥원이 저작권을 보유하고 있습니다.
한국지능정보사회진흥원의 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

01 EU와 일본의 새로운 소버린 클라우드 정책

| 한상기 테크프론티어

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.
한국지능정보사회진흥원이 저작권을 보유하고 있으며 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

이번 달은 최근 새롭게 발표한 EU와 일본의 소버린 클라우드 관련 정책을 소개하려고 한다. EU에서는 유럽연합 집행위원회(EC)가 1.8억 유로 규모 '소버린 클라우드 및 엣지-클라우드' 육성 프로그램을 가동한다고 발표했으며, 일본에서는 경제산업성(METI)의 지원을 받아 소프트뱅크가 자체 클라우드 OS(Infrinia) 및 엔비디아 GB200 기반 소버린 클라우드 상용 출시했다. 또한 METI와 총무성(MIC)는 클라우드와 AI 데이터센터의 전력난 해소와 무탄소 전원을 연계하기 위한 '와트-비트(Watt-Bit) 연계' 정책을 구체화했다.

EU '소버린 클라우드 및 엣지-클라우드' 전략 조달·육성 프로그램

유럽연합 집행위원회(EC)는 유럽 내 클라우드 인프라의 대미(對美) 하이퍼스케일러 의존도를 낮추고 실질적인 디지털 주권을 확보하기 위해, 추상적인 정치적 구호에 머물던 '소버린 클라우드'를 정량적 평가 지표(Cloud Sovereignty Framework)와 대규모 공공 조달·R&D 자금 지원, 그리고 하드웨어 및 인프라 법제화(Cloud and AI Development Act, CADA)의 3각 축으로 구체화했다.

특히 1억 8,000만 유로(약 2,700억 원) 규모의 범유럽 기관 소버린 클라우드 전략 조달 계약을 기점으로, 2026년 하반기(9~10월)에는 이를 민간과 회원국 전체로 확산시키는 주권 평가 프레임워크(SEAL 등급제)의 전면 도입과 '8ra(IPCEI-CIS)' 및 호라이즌 유럽 기반의 엣지-클라우드 연속체 상용화를 동시에 가동한다.[1)]

EC 디지털서비스총국(DG DIGIT)은 향후 6년간 EU 산하의 모든 기관, 기구 및 에이전시가 사용할 소버린 클라우드 인프라 조달을 위해 1억 8,000만 유로 규모의 계약을 체결하고 본격적인 시스템 이관을 진행 중이다. 단일 벤더 종속(Lock-in)을 방지하고 생태계 복원력을 높이기 위해 아래 4개 유럽 클라우드 사업자(컨소시엄)를 복수 선정했다.

1) Cloudtech, "EU awards €180 million sovereign cloud contract to four providers," Apr 20, 2026

선정 사업자/컨소시엄	소속 국가	달성 주권 등급	핵심 기술 스택 및 특징
Post Telecom 컨소시엄 (파트너: OVHcloud, CleverCloud)	룩셈부르크·프랑스	SEAL-3 (디지털 복원력 등급)	자체 개발 유럽 오픈소스 기반 IaaS/PaaS 스택 결합, 비(非)EU 제3자에 의한 서비스 차단 완전 면역 달성
STACKIT (Schwarz Group 산하)	독일	SEAL-3 (디지털 복원력 등급)	독일 유통 대기업 슈바르츠 그룹의 자체 클라우드 아키텍처 기반, 완전 관리형 PaaS 및 자동화 역량 입증
Scaleway (Iliad Group 산하)	프랑스	SEAL-3 (디지털 복원력 등급)	자체 클라우드 스택 및 AMD 기반 유럽 AI 슈퍼컴퓨터 인프라 연계, 개발자 경험(DX) 특화
Proximus / S3NS 컨소시엄 (파트너: 탈레스, Clarence, 미스트랄 AI)	벨기에·프랑스	SEAL-2 (데이터 주권 등급)	구글 클라우드 기술을 기반으로 하되 유럽 법인(S3NS·Clarence)이 독점 운영하고 미스트랄 AI 모델을 결합한 '소버린형 파트너십' 모델

표 1 선정된 소버린 클라우드 사업자 컨소시움

이번 프로그램의 가장 큰 제도적 돌파구는 모호했던 '주권(Sovereignty)' 개념을 세계 최초로 객관적이고 정량적인 조달 지표로 전환한 것이다. EC는 이 평가 도구와 계산기를 공개 문서로 배포하여 2026년 9~10월 현재 유럽 공공·금융·규제 산업 전반의 표준 입찰 기준으로 확산시키고 있다.[2)]

8대 주권 평가 목표(SOV-1 ~ SOV-8) 및 가중치는 다음과 같다.[3)]

1. 공급망 주권 (SOV-5, 20%): 하드웨어와 소프트웨어 공급망 투명성 및 역외 부품 통제 리스크
2. 전략적 주권 (SOV-1, 15%): 소유권 구조, 지배구조 안정성 및 EU 산업 생태계 내재화 수준
3. 운영 주권 (SOV-4, 15%): 비EU 인력이나 본사의 개입 없이 유럽 주체만으로 운영·유지보수·패치할 수 있는 역량
4. 기술 주권 (SOV-6, 15%): 오픈소스 기반 기술 개방성, 상호운용성 및 특정 외산 독점 기술 종속 배제
5. 법적 관할권 주권 (SOV-2, 10%): EU 법령 준수 및 미국 클라우드 법 등 역외 법령의 강제 집행으로부터의 절연성
6. 데이터 및 AI 주권 (SOV-3, 10%): 데이터 처리 위치 및 AI 모델 및 학습 자산에 대한 고객의 통제권
7. 보안 및 컴플라이언스 주권 (SOV-7, 10%): 보안 관제 및 규제 감사의 EU 역내 독립성
8. 환경적 지속가능성 (SOV-8, 5%): 에너지 효율성 및 원자재 희소성 대응력

2) The Lawxy Times, “EU Cloud and AI Act Defines Sovereign Cloud Standards,” Oct 12, 2026
3) DAMOVO, “SEAL-0 to SEAL-4, the sovereignty score your board will start asking about,” Sep 25, 2026

이번에는 SEAL(Sovereignty Effectiveness Assurance Level) 5단계 보증 및 '최저점 과락(Floor) 원칙'을 적용한다. 즉, SEAL-0(주권 없음)부터 SEAL-4(칩부터 SW까지 100% EU 공급망으로 구성된 완전 주권)까지 5단계로 평가하며, 평균 점수가 아무리 높아도 8개 항목 중 단 하나라도 발주처가 요구한 최소 기준(이번 입찰의 경우 SEAL-2)에 미달하면 즉시 탈락하는 '최소 보증 하한선' 방식을 채택해, 마케팅용 '주권 세탁(Sovereignty-washing)'을 원천 차단한다.

EC는 조달을 통한 수요 창출과 동시에, 기술 공급 측면에서 호라이즌 유럽(클러스터 4: 디지털, 산업, 우주)과 유럽 공동이익 중요 프로젝트인 IPCEI-CIS('8ra' 이니셔티브)를 통해 중앙 집중형 하이퍼스케일 클라우드 패러다임을 흔드는 '클라우드-엣지 연속체' 기술 구현에 자금을 집중하고 있다.

먼저, 연합형(Federated) 엣지 클라우드 구현이라는 목표 아래, 도이치텔레콤, 오렌지, 텔레포니카 등 유럽 주요 통신사와 중소 클라우드 사업자의 인프라를 단일 진입점(Single-entry point)으로 묶어, 워크로드가 스마트 팩토리·자율주행·병원 등 로컬 엣지 노드와 지역 클라우드 간에 동적으로 할당·이동되도록 하는 오픈소스 오케스트레이션 기술이 2026년 하반기 실증 단계에서 본격적인 상용화 국면으로 전환하고자 한다.

또한 클라우드 및 AI 개발법(CADA)을 통한 법적·물리적 뒷받침으로(2026년 10월 12일 결정), EC DG CONNECT는 최근 CADA 입법 프레임워크 및 집행 결정(COM/2026/1024)을 통해 소버린 클라우드의 법적 정의와 인증 절차를 확정했다(2027년 1월 1일 발효). 또한 각 회원국이 인허가를 패스트트랙으로 지원하는 '데이터센터 가속 구역(Data Centre Acceleration Zones)'을 최소 1곳 이상 지정하도록 의무화했다.

이번 발표의 전략적 의미와 산업적 시사점을 살펴보면 다음과 같이 정리할 수 있다.

- **선언적 주권에서 '실용적·계량적 주권'으로의 전환** : 과거 유럽의 가이아-X 프로젝트가 이상적인 완전 독립을 추구하다가 미국 빅테크의 참여 범위를 두고 내홍을 겪으며 지체되었던 것과 달리, 이번 프로그램은 '기술의 국적'과 '운영과 관할권의 통제'를 분리했다. 다시 말해 순수 유럽 기술을 사용하는 사업자(OVHcloud, STACKIT, Scaleway)에게는 높은 등급(SEAL-3)을 부여해 우대하면서도 미국 기술(구글 클라우드, 브로드컴 VM웨워)을 활용하더라도 유럽 법인이 물리적 데이터센터와 운영 및 키 관리를 100% 통제하는 구조라면 최소 주권 요건(SEAL-2)을 인정해 주는 '하이브리드 호환형 주권 모델'을 제도화했다. 이는 유럽 내 최첨단 AI·PaaS 기술 격차를 현실적으로 인정하면서도 주권 통제력을 확보하려는 고도의 실용주의적 타협이다.[4)]

- **공공 조달을 마중물로 한 유럽 토종 CSP의 '스케일업' 기반 마련** : 유럽 클라우드 시장에서 AWS, 마이크로소프트 애저, 구글 클라우드 3사의 점유율이 70%를 상회하는 상황에서, EU 집행위가 직접 1.8억 유로 규모의 장기 수요자로 나선 것은 토종 클라우드 기업들에게 안정적인 매출 기반과 레퍼런스를 제공한다는 의미가 있다. 특히 단순 스토리지나 가상머신(IaaS)을 넘어 완전 관리형 서비스(PaaS)와 AI 추론 환경(미스트랄 AI 등)까지 조달 범위에 포함시켜 유럽 CSP들의 기술 고도화를 강제했다.
- **중앙 집중형 클라우드에서 '분산형 엣지'로의 게임 체인징 시도** : 미국 빅테크가 이미 장악한 초대형 하이퍼스케일 데이터센터 시장에서 정면 승부하는 대신, 제조·모빌리티·에너지 등 유럽이 강점을 가진 B2B 산업 현장(엣지)과 지역 클라우드를 연결하는 '클라우드-엣지 연속체(Continuum)'에 R&D 예산을 집중함으로써 차세대 산업용 AI 및 IoT 인프라의 주도권을 가져오겠다는 산업 정책적 의도가 담겨 있다.

그러나 앞으로 해결해야 하는 여러 쟁점이 남아 있는데, 먼저 '법적 관할권 10% 가중치'를 둘러싼 유럽 클라우드 업계의 내분의 문제이다. 유럽 클라우드 인프라 서비스 제공자 협회(CISPE) 등 일부 순수 유럽 진영에서는 EC의 주권 점수 산정 방식에서 '법적·관할권 주권(SOV-2)'의 가중치가 10%에 불과하다는 점을 강하게 비판하고 있다. 자본력이 막강한 미국 하이퍼스케일러들이 환경(5%), 보안(10%), 기술 기능성 등에서 만점을 받아 법적 관할권의 약점을 상쇄하고 '소버린 클라우드' 자격을 획득할 수 있는 구조적 허점이 존재한다는 지적이다. 이에 따라 2027년 상반기 유럽의회의 CADA 인증 위임 절차 개정 과정에서 기준 강화 여부를 두고 치열한 로비와 논쟁이 예상된다.[5)]

발표에 따르면 2026년 10월 확정된 CADA 기준에 따라 2027년 1월 1일부터 소버린 클라우드의 법적 정의가 발효되고, 2027년 6월 30일까지 회원국별 공식 인증 기관이 지정된다. 이에 따라 유럽 내 공공기관 및 금융, 의료 등 규제 산업에 진입하려는 글로벌 클라우드 및 SaaS 기업들은 2027년 말까지 모든 데이터 처리 계약(DPA)에 새로운 주권 인증 조항을 의무적으로 반영해야 한다.

국가 AI 컴퓨팅 인프라 구축이나 공공 클라우드 보안인증(CSAP) 등급제를 운영하는 한국 입장에서도 EU의 '8대 주권 목표 및 SEAL 등급제'는 매우 중요한 참고 모델이다. 외산 클라우드의 진입을 일률적으로 허용하거나 차단하는 이분법에서 벗어나, 공급망, 운영권, 법적 관할권, 기술 개방성을 다차원으로 수치화하여 워크로드의 중요도에 따라 차등 적용하는 정교한 거버넌스로 진화하고 있기 때문이다.

4) EC, “Commission advances cloud sovereignty through strategic procurement,” Apr 17, 2026
5) TNW, “EU awards its €180 million sovereign cloud contract to four European providers,” Apr 17, 2026

일본의 소프트뱅크의 소버린 클라우드

소프트뱅크는 2026년 10월, 일본 내 자체 데이터센터에 구축한 차세대 수랭식 GPU 클러스터 '엔비디아 GB200 NVL72' 하드웨어와 자사 미국 법인(SB 텔레콤 아메리카) 산하 인프리니아 팀(Infrinia Team)이 독자 개발한 클라우드 소프트웨어 스택 '인프리니아 AI 클라우드 OS'를 결합한 'AI 데이터센터 GPU 클라우드'를 상용 출시했다.[6)]

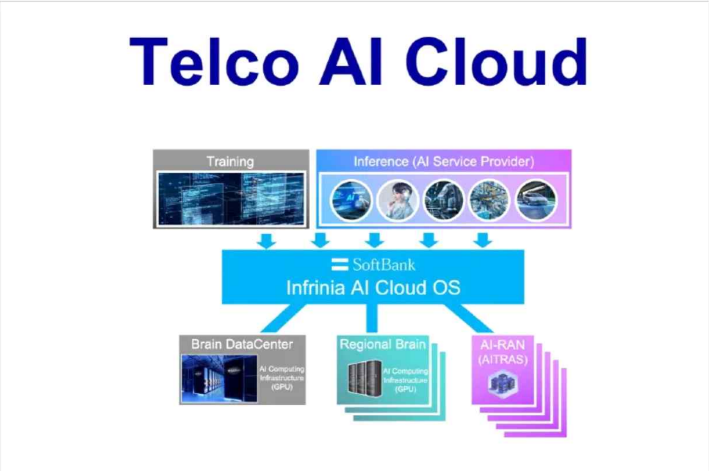


그림 1 인프리니아 AI 클라우드 OS 구성도 [출처: 소프트뱅크]

이번 출시의 배경에는 일본 경제산업성(METI)이 '경제안전보장추진법'에 의거해 중요 물자인 클라우드 프로그램의 해외 의존도를 낮추기 위해 소프트뱅크에 총 421억 엔(1차 53억 엔, 2차 368억 엔) 규모의 보조금을 지원한 국가 전략이 자리 잡고 있다. 소프트뱅크는 그동안 자체 거대언어모델(LLM)인 '사라시나(Sarashina)' 개발과 그룹 내 활용에 우선 투입하던 최첨단 연산 자원을, 독자적인 오케스트레이션 OS를 얹은 '일본형 소버린 네오클라우드(Sovereign Neocloud)' 형태로 외부 기업과 연구기관에 전면 개방하며 AI 인프라 사업자로의 전환을 본격화했다.

소프트뱅크가 상용화한 이번 소버린 AI 클라우드는 크게 ① 차세대 하드웨어(GB200 NVL72), ② 독자 클라우드 소프트웨어 스택(Infrinia AI Cloud OS), ③ 통신망 결합형 엣지-코어 통합 비전(Telco AI Cloud)의 3계층으로 구성했다.

미국 캘리포니아 서니베일에 거점을 둔 소프트뱅크 산하 '인프리니아 팀'이 올해 1월 공개하고 이번

6) Softbank, “SoftBank Corp. Announces Telco AI Cloud Vision toBuild Social Infrastructure for the AI Era, Leveraging Its Telecommunications Foundation,” Mar 2, 2026

10월 상용 서비스에 탑재한 인프리니아 AI 클라우드 OS는 단순한 가상화 관리 도구가 아니라 대규모 GPU 클러스터의 물리적 토폴로지를 이해하는 GPU 네이티브 운영체제이다. 토폴로지 인지형 스케줄링을 통해 GB200 NVL72 환경에서는 72개의 GPU가 초고속 NVLink로 묶여 하나의 거대한 GPU처럼 작동한다. 인프리니아의 KaaS 계층은 워크로드 배포 시 GPU 간 물리적 거리와 NVLink 도메인 경계를 자동으로 계산하여 노드를 할당함으로써, 멀티테넌트 환경에서도 베어메탈에 준하는 학습과 추론 성능을 보장한다. 또한 추론 비용 최적화(Inf-aaS)를 위해 기업 고객이 고가의 블랙웰 GPU를 상시 점유하지 않고도, 중앙 클라우드에서 학습·최적화된 모델을 API 형태로 호출하거나 엣지 노드로 분산 배포할 수 있도록 추상화된 추론 계층을 제공한다.

소프트뱅크는 이번 중앙 GPU 클라우드를 자사의 차세대 통신망 기술인 AI-RAN(솔루션명: AITRAS) 및 멀티 액세스 엣지 컴퓨팅(MEC)과 단일 제어 평면으로 묶는 'Telco AI Cloud' 아키텍처를 구현하고 있다. 중앙의 GB200 클라우드에서 파운데이션 모델을 학습 및 파인 튜닝하고, 실제 로봇 제어나 공장 자동화 같은 초저지연 추론은 통신 기지국 단의 AI-RAN/MEC에서 수행하도록 인프리니아 OS와 AITRAS 오케스트레이터가 워크로드를 동적으로 배분한다. 실제로 야스카와전기(산업용 로봇 제어), 미쓰비시중공업(엣지 데이터센터)과의 실증이 이 플랫폼 위에서 진행되고 있다.

지금까지 많은 국가와 통신사들의 '소버린 AI' 전략이 NVIDIA GPU를 대량 구매해 데이터센터에 쌓아두고 단순 임대하거나 미국 하이퍼스케일러의 관리형 스택을 빌려 쓰는 수준에 머물렀다면, 소프트뱅크는 자체 오케스트레이션 소프트웨어를 독자 개발해 결합했다는 점에서 큰 차이가 있다. 이는 외산 클라우드 제어 평면에 대한 종속을 끊어내고, 향후 이 소프트웨어 스택 자체를 글로벌 데이터센터 사업자에게 수출하려는 포석이다.

또한, 미국에서 코어위브나 람다 등 GPU 특화 클라우드(네오클라우드)가 AWS나 애저의 대안으로 급성장한 것처럼, 소프트뱅크는 일본 시장에서 '통신사 주도의 소버린 네오클라우드' 포지션을 선점했다. 특히 금융, 의료, 방산, 공공 및 제조 분야의 일본 기업들이 해외 클라우드로의 기밀 데이터 반출 우려와 일본의 디지털 무역 적자(연간 약 5~6조 엔 규모) 심화 문제 속에서 확실한 자국 내 대안을 확보하게 되었다고 평가할 수 있다.

이를 통해, 일본이 글로벌 경쟁력을 보유한 정밀 제조·로보틱스 산업(야스카와전기, 화낙 등)과 소버린 AI 클라우드를 직결함으로써, 단순 텍스트 생성형 AI를 넘어 산업 현장의 물리적 기계를 제어하는 '피지컬 AI' 인프라 시장을 선도할 수 있는 기반을 마련했다는 주장이다.

그러나 앞으로의 과제도 남아 있다. 먼저, 막대한 설비투자(CAPEX) 회수 및 상용 수익 모델(Unit Economics) 증명이 있는데, 경제산업성의 보조금이 마중물 역할을 했으나, 수천억 엔에 달하는 GB200 등 차세대 GPU 도입 비용과 데이터센터 구축 비용을 회수하기 위해서는 그룹 내부(SB 인튜이션, 라인야후 등) 수요를 넘어 외부 엔터프라이즈 고객의 지속적인 유료 사용률을 끌어올려야 한다. 특히 AWS, MS 애저, 구글 클라우드 역시 일본 리전에 수십조 원 규모의 AI 인프라 투자를 단행하고 있어, 가격 경쟁력과 가동률 확보가 수익성의 핵심 과제가 될 것이다.

두 번째로는 GB200 수랭식 인프라의 대규모 전력 수급 및 지방 분산 과제인데, GB200 NVL72는 랙당 소비전력이 120kW에 육박하는 고밀도 수랭식 설비이다. 도쿄와 오사카 등 수도권 그리드의 전력 포화 문제를 극복하기 위해, 소프트뱅크가 추진 중인 홋카이도 토마코마이 및 오사카 사카이(구 샤프 공장 부지) 등 대형 AI 데이터센터의 적기 완공과 무탄소 전원(재생에너지 및 원전) 연계가 차질 없이 뒷받침되어야 한다.[7)]

상위 PaaS/SaaS 생태계 및 개발자 툴체인의 성숙도 확보도 필요한데, 인프리니아 AI 클라우드 OS가 KaaS와 오픈AI 호환 Inf-aaS를 제공하며 진입 장벽을 낮췄지만, 이미 AWS 베드록이나 애저 AI 파운드리의 방대한 데이터 파이프라인, 보안 거버넌스, 서드파티 SaaS 생태계에 익숙해진 엔터프라이즈 개발자들을 유인하려면 단순 GPU 연산 가성비를 넘어선 정교한 MLOps/LLMOps 툴체인과 기술 지원 인력 풀을 빠르게 확충해야 한다.

마지막으로 글로벌 확장을 위한 상호운용성 및 표준화 검증의 과제인데, 소프트뱅크는 인프리니아 OS를 일본 내수용에 그치지 않고 글로벌 데이터센터에 공급하겠다는 목표를 세우고 있다. 이를 위해서는 CNCF(클라우드 네이티브 컴퓨팅 재단) 오픈소스 생태계와의 긴밀한 호환성 유지는 물론, 엔비디아 외에 향후 도입될 다양한 AI 가속기(ASIC, 차세대 아키텍처)까지 포용할 수 있는 하드웨어 추상화 역량을 입증해야 한다.

일본 METI '와카사만 그린 데이터센터' 구상 및 와트-비트(Watt-Bit) 연계 정책

일본 경제산업성(METI)과 총무성(MIC)은 2026년 9~10월에 걸쳐 발표된 '전기사업법 개정안' 및 'GX(그린 트랜스포메이션) 2040 추진 전략'의 핵심 일환으로, 폭증하는 클라우드 및 AI 데이터센터의 전력난과 도쿄·오사카 대도시권 집중 현상을 해소하기 위한 '와트-비트(Watt-Bit) 연계 정책'을 공식 구체화했다.[8)]

7) AI for Developing Countries Forum, “SoftBank to Launch AI Data Center GPU Cloud in Japan by October 2026,” Oct 17, 2026

이 정책의 가장 상징적인 실행 거점이 바로 원자력 발전소가 밀집해 있어 전력 공급 인프라가 풍부한 후쿠이현 와카사만 일대이다. 일본 정부는 원전 및 재생에너지 발전 설비 인근에 대규모 AI 클라우드 데이터센터를 직접 구축하는 '와카사만 그린 데이터센터 클러스터' 조성을 본격 추진하며, 분산형 에너지와 디지털 인프라를 결합한 국가 주도형 탈탄소·지방분산 전략을 가동하고 있다.

이번 구상은 단순한 데이터센터 부지 선정을 넘어, 전력망(Watt)과 데이터 네트워크(Bit)를 직접 연결하고 규제를 대수술하는 복합 인프라 프로젝트이다. 여기에는 다음과 같은 정책 분야가 있다.

- **에너지와 그리드의 통합 (와트-비트 연계)** : 무탄소 전원 직연계 및 송전 손실 최소화를 목표로 후쿠이현 와카사만 지역의 원자력 발전소(오이, 다카하마 등) 및 동해안 재생에너지 전력을 광역 송전망(Grid)을 거치지 않고 데이터센터에 직접 공급하거나 우선 배정한다. 또한 장거리 송전 과정에서의 전력 손실을 막고 대규모 AI 랙(GB200 등)의 상시 저부하 수요 충족를 목표로 한다.
- **지방 분산** : 수도권 데이터센터의 전력 수급 한계와 냉각수 부족 문제를 해결하기 위해 동해안 한랭 기후와 풍부한 수자원을 활용한 자연 냉각 도입과 지방 소멸 대응 및 재난 대비 국가 데이터 백본의 분산화를 통해 도쿄·오사카 과밀 해소 및 전력 제약 탈피를 꾀한다.
- **전기사업법 개정** : 메가와트급 이상 대형 클라우드 사업자가 발전사로부터 직접 전력을 구매할 수 있는 PPA(전력구매계약) 및 자가용 전기설비 규제 대폭 완화와 그린 데이터센터 인증 제도를 통한 세제 혜택 및 인허가 패스트트랙 지원으로 전력 공급 특례 및 전용선 규제 완화를 하고자 한다.

와트-비트 전략의 의미는 데이터(Bit)와 에너지(Watt)를 물리적으로 일치시키는 전략이라는 데 있다. 기존에는 발전소에서 생산된 전력이 수백 킬로미터 떨어진 도쿄의 데이터센터로 송전되는 과정에서 막대한 전력 손실이 발생하고 송전망 용량이 병목으로 작용했다. METI는 이를 해결하기 위해 "전력이 있는 곳에 데이터센터를 짓는다(Locate data centers where power resides)"는 패러다임으로 전환하여, 원전 밀집 지대에 하이퍼스케일 및 AI 클라우드 허브를 구축하고 초고속 해저·육상 광케이블로 대도시권과 직결하는 구조를 택했다.

이 전략은 와카사만 일대의 지리적, 산업적 강점을 활용한다는 측면도 있는데, 후쿠이현 와카사만 지역은 간사이전력의 원자력 발전소들이 집중되어 있어 도쿄전력 관내에 비해 비교적 안정적이고 대규모의 무탄소 기저부하 전력을 확보할 수 있는 일본 내 유일무이한 지역이기 때문이다. 해안 특유의 서늘한 기후와 풍부한 공업용수 및 해수를 활용해 최신 수랭식 AI 데이터센터의 냉각 효율을 극대화할 수 있을 것으로 본다.

8) METI, “Interim Report 3.0 Released as Outcome of Meetings of the Expert Group on the Development of Digital Infrastructures (Data Centers (DCs), etc.),” Oct 4, 2024

일본은 AI 확산에 따른 데이터센터 전력 수요 폭증으로 향후 전력망 붕괴 우려와 함께 외산 클라우드 의존에 따른 디지털 무역 적자 문제를 동시에 겪고 있다. 이번 와카사만 구상은 원자력 기반의 무탄소 전력과 자국 소버린 클라우드 인프라를 결합함으로써, 에너지 안보와 디지털 주권을 동시에 달성하는 모범 사례를 제시하고자 한다.

그동안 도쿄 인근(치바, 사이타마)에 편중되어 있던 일본의 클라우드 리전 지도를 혼슈 북측 및 일본 동해안으로 다변화하여, 수도권 대지진 등 국가 재난 상황에서도 데이터와 AI 인프라가 멈추지 않는 비상 복원력을 확보하고자 한다.

나아가 GX(그린 트랜스포메이션)와 AI 인프라의 제도적 융합을 위해, 전기사업법과 디지털 인프라 정비법을 유기적으로 개정하여, 클라우드 기업들이 친환경 에너지 조달 증명(RE100 등)을 손쉽게 달성할 수 있도록 제도적 틀을 마련했다. 이는 글로벌 친환경 규제(EU CBAM 등)에 대응해야 하는 일본 제조 및 금융 기업들의 AI 클라우드 이용 장벽을 낮춰준다.

일본 정부가 앞으로 해결해야 하는 문제는 먼저 지역 수용성과 원전 안전성에 대한 심리적 장벽을 극복해야 하는 일이 먼저이다. 후쿠시마 원전 사고 이후 원자력 의존도 확대에 대한 일본 국민 및 지방 자치단체의 경계감이 여전히 존재한다. 또한, 와카사만 지역에 대규모 데이터센터 클러스터가 들어설 때 유입되는 외지 인력 및 송전선 추가 증설 과정에서 지역 주민과의 소통과 안전 보장이 선행되어야 한다.

대규모 송전망 및 통신 인프라 초기 투자 비용도 과제인데, 발전소가 밀집한 와카사만에서 도쿄, 오사카 등 주요 소비지로 대규모 컴퓨팅 결과를 지연 없이 전송하기 위해서는 초고속·대역폭 백본 네트워크 구축이 필수적이다. 초기 인프라 건설에 막대한 민관 공동 투자가 필요하다고 본다.

마지막으로 엔지니어 풀 및 생태계 구축의 문제인데, 수도권에 집중된 IT 인재와 AI 연구원들이 지방(후쿠이현 등)에 위치한 데이터센터 리전으로 원격 근무하거나 상주하도록 유인할 수 있는 정주 여건 및 산·학·연 협력 생태계 조성이 과제로 남아 있다.

나아가며

EU가 다차원 주권 평가 지표(SEAL 등급제)와 데이터법을 통해 공정 경쟁 및 플랫폼 독점 시정을 강제하고, 일본이 '와트-비트 연계(GX 2040)'와 소버린 클라우드 OS(Infrinia)를 통해 전력 인프라와 AI 컴퓨팅을 결합하는 방식은 우리에게 중요한 시사점을 준다. 한국 역시 단순한 GPU 확보나 선언적 모델

개발을 넘어, 공공·제조·국방 등 핵심 산업 현장의 데이터 관할권과 물리적 전력과 인프라를 유기적으로 엮어내는 실행 중심의 '풀스택 소버린 AI 거버넌스'를 서둘러 정비해야 한다.

나아가 우리의 강점인 세계 최고 수준의 하드웨어 제조 및 통신 인프라와 도메인 특화(제조, 로보틱스) 응용력을 결합해 글로벌 공급망 리스크에 대응하고 실질적인 글로벌 레퍼런스를 확보하는 차별화된 실행 전략이 필요하다.

외산 클라우드와의 협력은 단순한 '종속'이나 '배제'가 아니라, 상호보완적 역할 분담과 기술적 레버리지를 극대화하는 방향으로 설계해야 한다. EU와 일본의 소버린 클라우드 전략, 그리고 국내외 인프라 환경을 종합할 때 현실적인 협력 방향을 모색해야 할 것이다.

02 빅테크의 사내 AI 내재화와 모델 선택권 전쟁

| 김영욱 SAP Product Engineering Product Expert

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.
한국지능정보사회진흥원이 저작권을 보유하고 있으며 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

2026년 초, 마이크로소프트는 사내에서 쓰는 앤트로픽 기술에만 연간 최소 10억 달러가 들 것으로 예상했다. 몇 달 뒤 이 금액은 3분의 1 넘게 줄었다. 같은 기간 메타에서 클로드 코드를 쓰는 직원은 약 6만 명에서 3만 명 수준으로 절반이 됐다.[9] 메타의 감원도 일부 영향을 미쳤지만, 더 큰 요인은 자체 개발한 코딩 도구와 모델로의 전환이었다. 앤트로픽의 가장 큰 기업 고객 두 곳이 동시에 사내 사용을 줄이고 있다는 사실은, AI 시장이 새로운 국면에 들어섰음을 보여준다.

이 변화를 단순한 비용 절감으로 읽으면 핵심을 놓친다. 마이크로소프트는 클로드를 금지하지 않았다. 사내 AI 코딩의 기본 모델을 오픈AI로 바꾸었을 뿐이며, 엔지니어는 여전히 클로드를 직접 선택할 수 있다. 달라진 것은 기본값이다. 사내 코딩 도구인 코파일럿은 대부분의 요청을 오픈AI 모델로 보내고, 마이크로소프트는 오픈AI와의 사업 관계상 이 모델에 별도 사용료를 내지 않는다. 메타 역시 사내 전용 도구 메타코드의 사용자가 3만 명을 넘어섰고, 같은 계열의 뮤즈코드를 외부 고객에게 판매하기 시작했다. 두 회사 모두 모델을 바꾸기 전에, 모델이 작동하는 도구를 먼저 자사 것으로 바꾸었다.

같은 흐름은 다른 빅테크에서도 확인된다. 아마존은 자체 코딩 도구 키로(Kiro)를 권장 도구로 지정하고, 새로운 외부 AI 개발 도구는 더 이상 지원하지 않겠다는 방침을 정했다[10]. 구글은 2026년 9월 모든 엔지니어에게 클로드 최신 모델을 열어주었지만, 자사 개발 플랫폼 안티그래비티 안에서 개인별 사용 한도를 두는 조건이었고 클로드 코드의 직접 사용은 여전히 허용하지 않았다. 사내 개발의 기본 모델은 제미나이라는 원칙도 그대로 유지했다[11]. 자체 프런티어 모델이 없는 애플은 차세대 애플파운데이션 모델의 기반으로 구글 제미나이를 택했는데, 이 결정에는 앤트로픽 모델을 썼을 때보다 낮은 비용이 크게 작용한 것으로 알려졌다[12]. 다섯 회사의 전술은 저마다 다르지만 방향은 하나다. 어떤 모델이 가장

9) The Information, “Microsoft Slashes Internal Claude Spending by a Third”, 2026 Oct 05
10) PeopleMatters, “Amazon memo reveals push for employees to use only Kiro, its in-house AI tool”, 2025 Nov 25
11) AI Weekly, “Google Opens Claude Opus 5 to All Engineers via Antigravity”, 2026 Sep 15

뛰어난가보다, 어떤 모델을 쓸지 누가 결정하는가가 경쟁의 축이 되고 있다.

이번 글에서는 이 현상을 '모델 선택권 전쟁'으로 정의한다. **모델 선택권이란 어떤 업무에 어떤 모델을 기본으로 쓸지 정하고, 조건이 바뀌면 언제든 교체할 수 있는 권한**이다. 이 전쟁은 모델 공급자에게 구조적 질문을 던진다. 이 전쟁은 모델 공급자에게도 구조적 질문을 던진다. 가장 많이 쓰는 고객이 동시에 경쟁 모델을 만들고 때로는 투자자이기도 한 구조에서는, 고객이 디폴트를 바꾸는 순간 수요의 상당 부분이 다른 모델로 옮겨 갈 수 있다.

이 흐름이 한국 기업과 공공기관에 던지는 메시지는 분명하다. 빅테크는 모델 선택권을 내재화하고 있지만, 대부분의 조직은 그 권한을 벤더의 기본 설정에 맡겨 두고 있다. 빅테크의 움직임을 사례별로 살펴보고, 이 전환을 이끄는 경제적·조직적 동인을 분석해 본다. 이를 바탕으로 국내 조직이 모델이 아닌 선택권을 확보하기 위한 전략적 함의를 제시해 본다.

1. 빅테크의 사내 AI 회귀

마이크로소프트는 비용을, 메타는 자사 모델의 무결성을, 아마존은 도구 표준화를 명분으로 내세웠다. 그러나 세 회사가 실제로 한 일은 같다. 직원들이 쓰던 외부 AI 도구를 자사 도구로 옮기고, 외부 모델은 그 도구 안의 선택지 중 하나로 남겨두었다.

1.1 마이크로소프트: 라이선스 회수에서 디폴트 변경까지

마이크로소프트의 전환은 도구, 과금, 디폴트 순으로 진행됐다. 2026년 5월 M 365 담당 부문이 클로드 코드 라이선스 대부분을 회수하고 자사 깃허브 코파일럿으로 옮기도록 했다. 코파일럿이 회사가 직접 다듬을 수 있는 제품이라는 점이 이유였고, 클로드 모델은 그 안에서 계속 선택할 수 있었다[13]. 7월부터는 모든 부문에 AI 토큰 예산 목표가 부여됐고, 사내 코딩의 기본 모델은 오픈AI의 GPT-5.6 Sol로 바뀌었다. 클라우드-AI 조직의 1인당 월 AI 사용 한도도 10만 달러에서 1만 달러 수준으로 낮아졌다. 마이크로소프트는 오픈AI와의 사업 관계상 오픈AI 모델에 별도 사용료를 내지 않는다. 그 결과 연 10억 달러를 넘을 것으로 예상됐던 사내 클로드 지출은 3분의 1 이상 줄었다. 판매자로서의 선택은 정반대다. 고객용 코파일럿 기능을 위해 앤트로픽에 지불하는 금액은 연 최소 20억 달러로

12) Bloomberg, “Apple To Lean On Google Gemini For Siri Overhaul”, 2025 Nov, 03
13) RuntimeWire, “Microsoft starts canceling Claude Code licenses, pushes engineers to GitHub Copilot CLI”, 2026 May 21

꾸준히 늘었고, 그 덕분에 앤트로픽에 지불하는 총액은 연초 고점 수준을 유지하고 있다. 고객에게는 최고 성능의 모델을, 사내에서는 가장 경제적인 모델을 디폴트로 쓰는 이중 전략이다.

1.2 메타: 사용자 절반 감소

올봄만 해도 메타 직원들 사이에서는 AI 역량을 과시하려고 토큰을 최대한 소비하는 '토큰맥싱(tokenmaxxing)' 경쟁까지 벌어졌다. 경영진은 6월 예산과 사용 한도, 중앙 모니터링 도입을 약속했다. 그러나 같은 달 나온 또 다른 조치는 비용 이상의 이유를 보여준다. 메타는 응용 AI 조직의 클로드 코드/코덱스 사용을 엄격히 제한하고 일부 작업을 중단시켰다. 외부 모델의 출력이 자사 학습 데이터에 섞이면 파트너사와 심각한 분쟁으로 번질 수 있다는 우려, 즉 증류(Distillation) 리스크 때문이었다. 자체 프런티어 모델을 만드는 기업에게 외부 모델 의존은 비용 문제를 넘어 법적 위험이기도 하다.

대안은 이미 준비돼 있었다. 메타는 8월 코딩 모델 뮤즈 스파크와 코딩 에이전트 뮤즈 코드를 함께 출시했다. 연초 약 6만 명이던 사내 클로드 코드 사용자는 약 3만 명으로 줄었다. 감원의 영향도 있었지만, 더 큰 요인은 사내 도구와 뮤즈 코드로의 이동이었다.



그림 2 외부 도구 의존을 줄인 세 기업

1.3 아마존: 같은 모델, 다른 도구

아마존은 앤트로픽의 최대 투자자 중 하나로 최대 250억 달러의 추가 투자를 발표했고, 자사 클라우드 베드락을 통해 클로드를 판매하는 최대 유통 채널이기도 하다. 그럼에도 아마존은 자체 코딩 도구 키로를 사내 권장 도구로 지정하고, 새로운 외부 AI 개발 도구는 지원하지 않기로 했다. 결정적인 사실은 키로 역시 앤트로픽 모델 위에 AWS 도구를 결합한 제품이라는 점이다. 아마존이 거둬들이려 한 것은 클로드라는 모델이 아니라, 클로드 코드라는 도구였다. 같은 모델이라도 어떤 환경에서, 어떤 규칙으로, 어떤 데이터를 남기며 쓸지는 도구를 소유한 쪽이 결정한다.

2. 같은 전쟁, 다른 전술

구글은 최고 수준의 자체 모델을 갖고도 경쟁사 모델을 사내에 들였고, 애플은 자체 프런티어 모델 없이 경쟁사 모델에 핵심 기능을 맡겼다. 출발점은 정반대지만, 두 회사 모두 **어떤 모델을 어디에 쓸지 정하는 권한은 내주지 않았다.**

2.1 구글: 모델은 열되 관문은 닫는다

지난 4월, 사내 정책에도 불구하고 구글 엔지니어들이 코딩 성능 격차를 이유로 클로드 코드를 찾고 있다는 보도가 나왔다. 구글의 대응은 외부 도구 차단이 아니라, 흩어져 있던 사내 코딩 도구를 에이전트 개발 플랫폼 안티그래비티로 통합하는 것이었다. 9월에는 모든 엔지니어에게 클로드 오푸스 5 사용을 허용했지만, 안티그래비티 안에서 개인별 한도 내에서만 쓸 수 있게 했고 클로드 코드/코덱스의 직접 사용은 여전히 막았다. 구글은 사내 개발의 기본 모델은 제미나이이며, 외부 모델은 특수한 용도에 한해 제공된다는 입장을 분명히 했다. **구글이 내준 것은 모델이고, 지킨 것은 관문**이다. 구글 역시 이중적 지위에 있다. 구글은 2026년 4월 앤트로픽에 최대 400억 달러 투자를 약속했다. 외부에서는 클로드에 투자하고, 사내에서는 그 사용량을 자사 기준으로 관리한다. 1장의 마이크로소프트와 같은 구조가, 자체 모델을 가진 기업에서도 반복되고 있다.

2.2 애플: 모델은 빌리되, 이름과 실행 환경은 소유한다

애플과 구글은 차세대 애플파운데이션 모델을 제미나이 기반으로 개발하는 다년 협력을 발표했다. 애플은 구글과 앤트로픽을 비교한 끝에 주로 재무 조건에서 제미나이를 택했고, 앤트로픽 모델이었다면 연 15억 달러 이상이 들었을 것으로 추산됐다. 6월 WWDC에서 공개된 '시리 AI'도 제미나이와 협력해

만든 환경 위에서 작동한다. 사용자가 보는 이름은 제미나이가 아니고, 실행 환경과 데이터 기준도 애플이 정한다. 조건이 바뀌면 그 아래의 모델을 교체할 수 있는 구조다.

사내에서는 다른 선택을 했다. 제품 개발과 내부 도구에 자체 서버에서 운영하는 맞춤형 클로드를 폭넓게 사용한다. 소비자 기능에는 비용 효율이 높은 모델을, 사내 개발에는 코딩 역량이 높은 모델을 배치한다. **용도에 따라 모델을 나눠 쓰는 것 자체가 선택권의 행사다.**



그림 3 외부 모델을 받아들인 두 기업

3. 비용, 제품, 통제

빅테크의 사내 AI 전환이 2026년에 한꺼번에 일어난 데는 이유가 있다. 에이전트 기반 AI 사용이 본격화하면서 비용 구조가 바뀌었고, 사내 사용 기록이 제품 경쟁력의 원천이 됐으며, 누가 어떤 데이터를 어떤 모델에 보내는지가 경영 리스크로 떠올랐다. 세 동인은 서로 다른 문제처럼 보이지만, 모두 같은 해법으로 향한다.

3.1 토큰 경제학: AI 비용이 '구독료'에서 '사용량'이 되다

에이전트 코딩이 확산되면서 AI 비용의 성격이 바뀌었다. 직원 1인당 정액을 내던 구독 모델에서, 모델

호출량만큼 비용이 늘어나는 사용량 과금 모델로 무게중심이 이동한 것이다. 에이전트는 하나의 작업을 끝내기 위해 수십, 수백 번 모델을 호출한다. 메타가 최근 28일 동안 클로드 코드에만 쓴 금액은 1억 500만 달러를 넘었고, 이를 단순 연간 환산하면 13억 달러를 웃돈다. 깃허브도 코파일럿을 사용량 기반 과금으로 전환했다. 공급자 쪽 수치도 같은 흐름을 보여준다. 앤트로픽의 2025년 매출 가운데 사용량 기반 매출은 약 38억 달러로, 구독 매출의 약 다섯 배였다.

사용량이 곧 비용인 구조에서 디폴트 모델은 단가를 결정한다. 같은 작업을 어떤 모델로 처리하느냐에 따라 비용이 크게 달라지기 때문이다. 오픈AI와 구글이 앤트로픽보다 낮은 가격을 제시하면서, 더 저렴한 대안을 찾는 기업 고객도 늘고 있다. 자사 모델이나 제휴 모델을 사실상 추가 비용 없이 쓸 수 있는 빅테크에게, 디폴트 변경은 가장 즉각적인 비용 절감 수단이다. AI 사용료는 이제 인건비처럼 매달 관리해야 하는 운영 항목이 됐다.

3.2 도그푸딩: 사내 사용량이 곧 제품 개발 자원

빅테크들이 공통적으로 내세운 또 다른 이유는 도그푸딩(Dogfooding, 자사 제품을 내부에서 먼저 쓰며 개선하는 관행)이다. 마이크로소프트는 직원들이 자사 도구를 써야 제품을 개선할 수 있다고 설명했고, 아마존은 엔지니어들의 피드백을 키로 개선에 적극 활용하겠다고 밝혔다. 구글도 사내 코딩 도구의 활용이 모델과 AI 도구 개발을 가속하고 있다고 강조했다.

에이전트 시대에 도그푸딩의 가치는 이전과 비교할 수 없이 커졌다. 수만 명의 엔지니어가 실제 업무에서 남기는 작업 기록은 하네스(Harness, 모델을 감싸 도구 호출·맥락 관리·작업 흐름을 담당하는 실행 환경)를 개선하고 모델을 학습·평가하는 데 쓰이는 가장 현실적인 데이터다. 직원들이 외부 도구에 쓰는 토큰은 결국 경쟁사 제품을 개선하는 피드백이 된다. 메타는 이 경로를 가장 빠르게 보여준다. 사내에서 다듬은 코딩 도구를 뮤즈 코드로 출시했고, 기업용 AI 사업부 메타 엔터프라이즈 플랫폼을 출범시켰다. 사내 도구가 그대로 외부 상품이 되는 구조에서, 사내 사용은 비용이 아니라 투자다.

3.3 통제: 누가, 어떤 모델을, 무엇에 쓰는지 아는 권한

도구를 소유한다는 것은 사용 데이터를 소유한다는 뜻이기도 하다. 마이크로소프트 경영진은 이제 직원별 일일 AI 지출과 사용 모델을 직접 확인할 수 있다. 외부 도구를 쓸 때 이런 가시성은 벤더 쪽에 있다. 어떤 코드와 데이터가 외부 서버로 나가는지도 도구가 결정한다.

보안과 지식재산 측면의 압력도 커지고 있다. 위에서 본 메타의 증류 우려 외에도, 팔란티어, 엔비디아, 부즈앨런해밀턴처럼 민감한 업무를 다루는 기업들은 데이터 보호에 대한 새로운 보장을 요구하거나 앤트로픽 최상위 모델의 사용을 줄이거나 중단했다. 아마존이 외부 도구 사용에 승인 절차를 둔 배경에도 독점 코드와 고객 데이터 보호라는 과제가 있다. 여기서 통제는 외부 모델을 금지하는 것이 아니다. **어떤 데이터를 어떤 모델에 보낼지를 자사 레이어에서 설계하는 것**이다.

4. 모델 선택권 전쟁의 본질

위에서 다룬 도구 레이어가 왜 전장이 됐는지, 그리고 이 전쟁이 모델 공급자에게 무엇을 의미하는지를 정리한다.

4.1 하네스를 향한 양방향 진격

하네스를 노리는 것은 빅테크만이 아니다. 앤트로픽의 클로드 코드와 오픈AI의 코덱스는 모델 공급자가 모델 위에 하네스를 직접 얹은 제품이다. 클로드 코드는 메타에서 가장 많이 쓰인 외부 코딩 도구였고, 마이크로소프트에서도 많은 개발자가 자사 코파일럿보다 선호했다. 공급자가 하네스까지 제공하면, 고객사 직원의 일상 업무에서 자사 모델이 자연스럽게 디폴트가 된다.

이제 주요 AI 기업은 모두 모델과 하네스를 한 쌍으로 묶어 경쟁한다. 메타는 뮤즈 코드를 발표하면서 각 회사의 모델을 각자의 에이전트 제품에서 실행해 성능을 비교했다. 자사 모델은 뮤즈 코드에서, 클로드는 클로드 코드에서, 오픈AI 모델은 코덱스에서, 제미나이는 안티그래비티에서 측정하는 방식이었다. **경쟁의 단위가 모델 단독에서 '모델과 하네스의 조합'으로 바뀌었다는 뜻**이다. **공급자는 모델에서 하네스로 올라오고, 빅테크는 직원과 모델 사이의 하네스를 되찾으려 한다.** 두 흐름이 정면으로 부딪히는 지점이 바로 하네스 레이어다.

4.2 공동학습: 디폴트가 정책에서 구조로

이 경쟁은 한 단계 더 나아가고 있다. 마이크로소프트는 6월 자체 개발 모델 7종을 발표하면서, 그중 코딩 모델 MAI-Code-1-Flash를 깃허브 코파일럿과 VS Code에 맞춰 설계하고 깊이 통합한 모델로 소개했다. 메타 역시 뮤즈 스파크 1.2를 뮤즈 코드와 함께 학습시켰다. **모델을 자사 하네스의 작업 방식, 도구 호출, 맥락 관리 방식에 맞춰 처음부터 최적화하는 것이다.**

이 변화의 의미는 크다. 지금까지 디폴트는 관리자가 정하는 설정이었다. 정책을 바꾸면 언제든 되돌릴 수 있었다. 그러나 **자사 하네스 안에서 자사 모델이 최적화되기 시작하면, 디폴트는 설정이 아니라 기술적 우위가 된다.** 같은 하네스 안에서 외부 모델은 일종의 손님이 되고, 모델 자체의 성능이 비슷하더라도 하네스와 맞물린 정도에서 차이가 생길 수 있는 것으로 분석된다. **정책으로 정한 디폴트는 되돌릴 수 있지만, 구조로 굳어진 디폴트는 쉽게 바뀌지 않는다.**

4.3 딜레마: 최대 고객이 가장 위험한 고객

이 구조는 모델 공급자에게 불편한 질문을 던진다. 2025년 앤트로픽 매출의 약 4분의 1은 각각 12%를 차지한 두 고객에게서 나왔고, 대형 고객 상당수는 장기 계약으로 묶여 있지 않다. AWS와 구글 클라우드를 통한 매출 비중도 2023년 11%에서 2024년 32%, 2025년 약 47%로 커졌다. 공급자의 최대 고객이 동시에 최대 유통 채널이자 투자자이며, 자체 모델과 하네스를 가진 경쟁자이기도 하다.

이 고객들이 **디폴트를 바꾸면 수요는 빠르게 이동한다.** 앤트로픽의 기업 고객 이탈은 오픈AI의 연간 환산 매출이 최근 몇 달 사이 약 70% 늘어 700억 달러에 육박하는 데 일부 기여한 것으로 보인다. 앤트로픽은 고객 기반을 넓히는 방식으로 대응하고 있다. 지난 6월까지 12개월 동안 10만 달러 이상을 지출한 고객은 약 6,000곳으로 2025년 한 해의 네 배에 이르렀고, 1,000만 달러 이상을 지출한 고객도 100곳을 넘었다. 결국 **공급자 성장의 지속성은 모델 성능만으로 결정되지 않는다. 고객이 디폴트를 바꾸지 않을 이유를 얼마나 만들어낼 수 있느냐에 달려 있다.**

5. 국내 조직: 모델이 아니라 선택권을 확보하라

빅테크의 사례가 국내 조직에 주는 교훈은 자체 모델을 만들라는 것이 아니다. 애플이 보여주듯, 프런티어 모델 없이도 선택권은 확보할 수 있다. 핵심은 모델 위의 레이어, 즉 어떤 모델을 어떤 업무에 쓸지 정하고 그 사용을 관리하는 권한을 조직 안에 두는 것이다. 이 장은 기업과 공공 부문으로 나누어 그 의미를 살펴본다.

5.1 기업: 디폴트를 벤더에게 맡기지 않는 법

국내 기업 상당수는 특정 벤더의 AI 도구를 도입하는 방식으로 AI를 활용하고 있다. 이 경우 어떤 모델이 기본으로 쓰이는지, 어떤 데이터가 외부로 나가는지, 비용이 어디서 늘어나는지는 사실상 벤더가

정한다. 에이전트 활용이 늘수록 이 의존은 커진다. **빅테크가 보여준 대응은 직원과 모델 사이에 조직이 통제하는 레이어를 두는 것이다.** 여러 모델을 하나의 인터페이스로 묶고 업무별로 배분하는 모델 게이트웨이가 그 출발점이다. 이 레이어가 있으면 특정 모델의 가격이 오르거나 더 나은 모델이 나왔을 때 직원들의 업무 방식을 바꾸지 않고도 모델을 교체할 수 있다. 모든 업무에 하나의 모델을 쓸 필요도 없다. 애플처럼 대량 처리 업무에는 비용 효율이 높은 모델을, 고난도 업무에는 성능이 높은 모델을 배치하는 것이 선택권을 실제로 행사하는 방식이다.

두 번째 과제는 토큰 예산 거버넌스다. 국내 기업도 마이크로소프트처럼 부문별 토큰 예산과 직원별 사용 가시성을 갖춘 체계가 필요하다. **계약 단계에서는 데이터의 학습 활용 금지, 사용 기록의 이전 가능성, 모델 교체 시의 전환 조건을 명시해 두어야 한다. 선택권은 기술 구조뿐 아니라 계약서에서도 확보된다.**

5.2 공공: 범정부 공통기반을 '선택권의 레이어'로 설계

국내 공공 부문은 이미 이 방향의 기반을 마련하고 있다. 행정안전부와 과학기술정보통신부는 2025년 11월, 중앙·지방정부가 내부망에서 다양한 민간 AI 모델과 학습데이터, GPU를 함께 활용하는 '범정부 AI 공통기반' 서비스를 시작했다[14]. 2026년 1월에는 정부가 공공기관이 중복 투자 없이 함께 쓸 수 있는 범정부 생성형 AI 플랫폼을 구축·운영하도록 하는 '공공 AI법'이 국회를 통과했고, 8월 시행된 개정 법률에는 범정부 공통기반의 우선 이용 조항이 담겼다. 구조로 보면 이는 빅테크의 하네스·라우팅 레이어를 정부 차원에서 구현한 것이다. 개별 기관이 개별 벤더에 묶이지 않고, 모델 선택권을 정부 공통 레이어에 두는 방식이다. 여러 회사의 모델을 한 플랫폼에서 제공하면서, 기관이 누가 어떤 모델을 쓸지 통제하고 모델을 비교하며 어떤 모델이 얼마나 쓰이는지 파악할 수 있게 했고, 특정 벤더에 종속되지 않는 접근을 명시적 목표로 내세웠다.

과제는 이 공통기반이 실제로 선택권의 레이어로 작동하게 하는 것이다.

첫째, 공통기반이 특정 모델이나 특정 사업자의 디폴트로 굳어지지 않도록, 모델을 정기적으로 평가하고 교체하는 규칙을 제도화해야 한다.

둘째, 국산 모델은 이 선택지의 한 축으로 자리 잡아야 한다. 정부는 독자 AI 파운데이션 모델 프로젝트로 선정된 모델을 공통기반에 추가할 계획이다. 주요국의 정책 변화로 해외 모델 이용이 제한될

14) 헤럴드경제, “정부, 내부망에 AI 공통기반 서비스 최초 도입”, 2025 Nov 23

경우 생산성에 직접 타격을 받을 수 있다는 점에서, 국산 모델은 해외 모델의 대체재라기보다 협상력과 연속성을 지키는 보험에 가깝다.

셋째, 공통기반에 쌓이는 기관별 사용 기록과 비용 데이터를 정책 판단의 근거로 활용해야 한다. 어떤 업무에 어떤 모델이 효과적이었는지에 대한 데이터는, 마이크로소프트의 사내 대시보드가 그렇듯 다음 선택을 더 정확하게 만든다.

6. 마무리

2026년 빅테크의 사내 AI 전환은 AI 경쟁의 질문을 바꾸어 놓았다. 어떤 모델이 가장 뛰어난가라는 질문은 여전히 중요하지만, 그보다 먼저 답해야 할 질문은 누가 디폴트를 정하는가다. 마이크로소프트, 메타, 아마존, 구글, 애플은 각자 다른 사정에서 출발했지만 같은 결론에 이르렀다. 모델은 언제든 바꿀 수 있어야 하고, 바꾸는 권한은 조직 안에 있어야 한다. 모델 간 성능 격차는 빠르게 좁혀지고, 가격은 계속 움직인다. 오늘 가장 좋은 모델이 내일도 가장 나은 선택이라는 보장은 없다. 이런 시장에서 변화의 이익을 가져가는 쪽은 최고의 모델을 쓰는 조직이 아니라, 더 나은 모델로 옮겨 갈 수 있는 조직이다.

국내 기업과 공공기관이 AI 도입의 성과를 따지기 전에 먼저 점검해야 할 것도 여기에 있다. 지금 우리 조직에서 쓰이는 AI의 디폴트를 누가 정하고 있는지, 그 권한을 되찾는 데 무엇이 필요한지부터 확인하는 것이 모델 선택권 전쟁에 대비하는 첫걸음이다.

03 3대 메가프로젝트의 AI 데이터센터 계획 – 성공적인 구현을 위해 필요한 것은?

| 윤대균 아주대학교 소프트웨어학과

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.
한국지능정보사회진흥원이 저작권을 보유하고 있으며 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

1. 들어가며

정부는 2026년 6월 29일 '대한민국 대도약 3대 메가프로젝트 국민보고회'를 열고 반도체, 피지컬 AI, 인공지능 데이터센터(AIDC: AI data center)를 한국형 AI 산업혁명의 세 축으로 제시했다.[15] 이 가운데 AIDC 프로젝트는 2035년까지 총 18.4GW를 구축하는 계획으로, 1단계에 참여하는 3개 기업의 투자 규모만 약 550조 원에 이른다. 대형 원전(APR1400) 1기의 설비용량이 1.4GW이므로, 18.4GW는 원전 13기 남짓의 전력을 AI 연산에 배정하겠다는 의미다.

AIDC 메가프로젝트는 장비와 솔루션의 국산화, 실증, 수출로 이어지는 공급망까지 고려한 종합적인 산업정책으로 설계되어 있다. 그러나 연산 수요가 구체적으로 무엇인지, 클라우드를 누가 어떻게 구축하고 운영할지, 성과를 어떻게 측정할지는 아직 제시되지 않았다. 이렇게 비어 있는 부분들이 채워지지 않으면 총 18.4GW 규모의 AIDC는 해외 수요를 위한 임대 설비에 머물고, 정부가 표방한 '토큰 생산 능력'은 국내 산업과 AX 경쟁력으로 이어지기 어렵다.

이 글은 국민보고회 자료와 이후 공개된 사업 동향을 토대로 AIDC 메가프로젝트의 설계 구조를 정리하고 목표 설정의 근거와 보완할 부분을 살핀다. 이어 산업 생태계에 대한 기대 효과와 성과를 위해 풀어야 할 과제를 짚고 소버린 AI 풀스택 관점에서 실행 방향을 제안한다. 소버린 AI는 한 국가가 AI의 핵심 계층인 반도체, 인프라, 모델, 서비스와 데이터에 대해 외부에 일방적으로 의존하지 않고 선택권과 통제권을 확보한 상태를 말한다. 모든 계층을 자급한다는 뜻이 아니라, 핵심 계층에서 대안을 갖고 필요할 때 교체할 수 있는 능력을 뜻한다. 전력망 확충과 AIDC의 지역 배분은 그 자체로 별도의 논의가 필요하므로 이 글에서는 산업 전략에 영향을 주는 범위에서만 다룬다.

15) 관계부처 합동, “대한민국 대도약 3대 메가프로젝트 국민보고회”, 2026.6.29, 산업통상부. https://www.motir.go.kr/kor/article/ATCL3f49a5a8c/171974/view

2. 3대 메가프로젝트의 AIDC 전략

2.1 '생각의 창고'로서의 AIDC

정부는 AIDC를 3대 메가프로젝트의 '기반'으로 규정한다. 국민보고회 자료는 반도체를 '두뇌', 로봇 등 피지컬 AI를 '지능을 갖춘 신체', AIDC를 '생각의 창고'로 부르며 세 분야를 유기적으로 결합하겠다고 밝혔다. 같은 자료에서 AI 생태계의 핵심 요소로 전력, 칩, 인프라, 모델, 애플리케이션을 꼽는데, 메가프로젝트가 직접 겨냥하는 것은 이 가운데 칩과 인프라, 그리고 피지컬 AI라는 분야이다. 범용 모델과 서비스는 별도의 축으로 설정되어 있지 않다.

정부가 AIDC에 부여한 역할은 '토큰 공장'이다. 배경훈 부총리 겸 과학기술정보통신부 장관은 "AIDC는 AI 서비스 필수 인프라임과 동시에 토큰을 생성하는 팩토리"라고 말했다.[16] 토큰(token)은 AI 모델이 텍스트를 이해하고 생성할 때 쓰는 기본 처리 단위로, 추론 서비스의 사용량과 비용을 재는 단위이기도 하다. 정부는 대규모 토큰 생산 능력을 갖춰 국민 누구나 대화형 AI와 에이전트 AI를 누리는 'AI 기본사회'를 실현하겠다는 목표를 함께 제시했다.

2.2 규모, 일정과 역할 분담

1단계는 2029년까지 8.4GW를 구축하는 계획이다. SK가 울산 1GW를 시작으로 중부권, 대경권, 호남권, 강원에서 추가 입지를 검토해 총 5GW를, GS가 동해에 2.4GW를, 네이버가 세종과 추가 지역에 1GW를 맡는다. 입지와 인허가를 마무리해 2028년 상반기에 착공하고 2029년부터 단계적으로 운영한다는 일정이다. 2단계는 정부와 SK가 1단계의 5GW를 2035년까지 15GW로 확장하는 것으로 두 단계를 합치면 18.4GW가 된다.

역할 분담은 정부가 전력, 부지, 인허가를 신속히 지원하고 민간이 AIDC 구축과 운영, 해외 투자 유치를 맡는 구조다. 규모의 근거로는 아시아태평양 지역의 2030년 AIDC 규모가 약 58GW로 예상되며 국내 AIDC를 14.37GW로 늘리면 그 25% 수준이 된다는 점을 들었다. 기대 효과로는 550조 원의 투자, 건설 4.2만 명과 운영 8.4천 명의 일자리, 10대 솔루션 분야별 대표기업 1개 이상 배출을 내걸었다. 1GW당 효과는 업계 추산을 인용해 투자 60~70조 원으로 제시했다.

16) 전자신문, "[대한민국 3대 메가프로젝트] 정부-SK·GS·네이버, AI 데이터센터 지역별 구축 박차", 2026.6.29. https://www.etnews.com/20260629000384

2.3 목표 체계: 국산화, 실증, 수출

AIDC 메가프로젝트의 정책 목표는 대규모 구축 자체보다 그에 수반되는 솔루션 산업의 육성에 맞춰져 있다. 국민보고회 자료는 국내 AIDC를 기반으로 '국산화, 실증, 수출'로 이어지는 전후방 산업 생태계를 키우겠다고 밝히고 이를 네 갈래로 구체화했다.

- **IT 장비** : 서버, 네트워크, 스토리지 기술을 개발하고 국내 AIDC에서 레퍼런스를 확보해 엔비디아, 그리고 AWS나 마이크로소프트 같은 클라우드 사업자와 에퀴닉스 같은 코로케이션 사업자의 공급망에 진입하도록 지원한다. 코로케이션은 사업자가 공간, 전력, 냉각을 제공하고 고객이 자기 서버를 들여와 운영하는 임대 방식이다. 데이터 통신과 저장을 전담하는 데이터 처리 장치(DPU), 광모듈과 함께 AI 연산에 특화한 국산 NPU 중심의 생태계도 지원한다.
- **클라우드와 소프트웨어** : 정부 스스로 '부족한' 기술로 지목한 GPU 클러스터링을 개발한다. 대규모 GPU 구매사업과 연계해 수천 장 규모의 클러스터링과 GPU 배분 기술인 오케스트레이션을 확보하고 AI 시스템과 대규모 언어모델, AI 에이전트의 운영을 관리하는 개발도구(AIOps, LLMOps, AgentOps)를 강화한다.
- **전력과 냉각** : 경쟁력 있는 타 산업의 장비를 AIDC용으로 대형화하고 패키지화한다.
- **클러스터와 금융** : AIDC 특별법(2027년 3월 시행 예정)에 근거해 특화 클러스터를 지정하고 국산 장비 중심의 테스트베드 5MW와 상용 AIDC 5MW를 함께 구축한다. 국민성장펀드가 AIDC 구축 사업자를 지원할 때 국산 장비 활용도를 고려하도록 의무화하는 방안도 검토한다.

이행 체계도 갖춰지고 있다. 7월 29일 과학기술정보통신부 주관으로 'AI 데이터센터 얼라이언스'가 출범해 수요, 공급, 기반 3개 분과와 수출산업화 태스크포스를 꾸렸고, 2027년 AI 데이터센터 사업단 출범과 클러스터 조성을 예고했다.[17] 정리하면 AIDC 메가프로젝트의 정책은 대부분 데이터센터 구축에 필요한 장비와 솔루션, 곧 전력과 냉각 설비, IT 장비를 공급하는 산업의 국산화와 수출을 지향한다.

3. 계획의 보완 과제

3.1 규모 산정 근거

18.4GW라는 목표는 국내 수요가 아니라 지역 점유율에서 도출됐다. 정부가 제시한 근거는 아태 AIDC 용량의 25%를 확보한다는 것이며, 국내에서 어떤 주체가 어떤 용도로 얼마의 연산을 쓸지에 대한

17) 유니콘팩토리, "AIDC 얼라이언스 출범…"AI 데이터센터, 수출공장 돼야" (종합)", 2026.7.29. https://www.unicornfactory.co.kr/article/2026072915402110005

추정은 자료에 없다. 점유율 목표는 해외 수요를 유치하는 공급 거점 전략으로서는 의미가 있지만 국내 AI 전환에 필요한 연산 규모를 설명하지는 못한다.

실제 구축 규모도 해외 고객 확보와 맞물려 있다. SK텔레콤은 2026년 2분기 실적 발표에서 부지와 전력을 먼저 확보하되 고객을 확보한 뒤 데이터센터를 구축한다는 원칙을 따르겠다고 밝히고, 5GW는 실제 고객 수요에 따라 단계적으로 늘리는 중장기 개발 목표라고 설명했다.[18] 회사는 AIDC 개발 전문회사를 세워 글로벌 빅테크와 협의 중이며, 이후 알려진 해외 포함 20GW 구상(국내 15GW)에 대해서도 "원론적 수준의 목표치"라고 했다.[19] GS 역시 동해 1단계 1.2GW를 프로젝트 파이낸싱 중심으로 조달하고 고객 계약에 맞춰 단계적으로 건설한다는 방침이다.[20] 민간 사업자로서는 합리적인 판단이다. 다만 이런 방침에 비추면 18.4GW는 주로 해외 임차 수요를 겨냥한 공급 계획이며 국내 수요 추정에 근거한 수치로 보기는 어렵다.

이는 전력 수급 계획에도 그대로 반영된다. 제12차 전력수급기본계획의 수요 재전망에서 2040년 데이터센터 순증 최대전력 수요는 4월 4.0GW에서 8월 11.9GW로 약 3배 늘었고, 재전망을 이끈 변수로 3대 메가프로젝트가 지목됐다.[21] 10월 국정감사에서 기후에너지환경부 장관은 데이터센터 수요가 "과잉 추계된 측면이 없지 않다"고 답했다.[22] 발전 설비와 송전망은 건설 기간이 길고 되돌리기 어렵다. AIDC 규모의 타당성을 '얼마나'에 앞서 '누구의 어떤 수요를 위해'라는 관점에서 다시 살펴봐야 하는 이유다.

3.2 실증 이후 대규모 운영으로 가는 단계 설계

GW급 시설의 운영 규모는 정부가 계획한 기술 개발 규모와 단위부터 다르다. 미국 텍사스 애빌린의 오픈AI용 스타게이트 단지는 목표 용량 1.2GW에 최종적으로 엔비디아 GB200 45만 개 이상을 들일 계획이다.[23]

---

18) 칸(KHARN), “SKT “고객 확보 후 5GW AI DC 짓는다””, 2026.8.7. https://www.kharn.kr/news/article.html?no=31531
19) 전자신문, “SK텔레콤, 해외 5GW AIDC 추가 확보 나선다…아시아 AI 인프라 허브로”, 2026.10.12. https://www.etnews.com/20261012000272
20) EBN, “GS, 동해 AI 데이터센터 15조 투자…연내 부지·인허가 진행”, 2026.9.15. https://www.ebn.co.kr/news/articleView.html?idxno=1724224
21) 한국데이터경제신문, “[심층리포트] “AI·반도체에 전기 다 쓰인다”…12차 전기본 데이터센터 11.9GW·반도체 24.3GW 전망”, 2026.9.3. https://www.dataeconomy.co.kr/news/articleView.html?idxno=42319
22) 경향신문, “기후부 장관, 데이터센터 전력수요 ‘뻥튀기’ 지적에 “과잉 추계된 측면 있다” 시인”, 2026.10.6. https://www.khan.co.kr/article/202610061812001/
23) DatacenterDynamics, “OpenAI and Oracle to deploy 450,000 GB200 GPUs at Stargate Abilene data center”, Oct. 15, 2025.

이를 단순 환산하면 1GW에 30만 개 이상이 들어가며, 1단계 8.4GW만으로도 수백만 개 규모가 된다. 가속기 세대와 시설 구성에 따라 추정은 달라질 수 있지만 규모의 단위를 가늠하기에는 충분하다.

규모가 커지면 운영 문제의 성격이 달라진다. 메타는 약 1만 6천 개의 GPU로 라마 3를 학습하는 54일 동안 예기치 않은 작업 중단을 419건 겪었고, 그 약 78%가 하드웨어 문제였다고 밝혔다.[24] 수많은 GPU가 동시에 연산을 멈추고 재개할 때 데이터센터 전체의 전력 소비가 순간적으로 수십 MW씩 출렁인다는 점도 보고했다. 몇 시간마다 일어나는 장애에서 작업을 자동으로 복구하고 전력 변동을 관리하는 기술은 그 규모에서 운영해 보아야 축적된다. 수천 장 규모의 클러스터에서는 이런 문제가 같은 빈도와 강도로 드러나지 않는다.

이 관점에서 보면 보완할 지점은 실증 다음 단계다. 정부가 확보하겠다는 클라우드 기술은 '수천 장 규모'의 GPU 클러스터링이다. 2025년 착수한 4,000억 원 규모의 K-클라우드 기술개발사업은 국산 AI 반도체 기반으로 '1만 장 수준'의 하이퍼스케일 데이터센터를 구축하고 운영할 기술 확보를 목표로 한다.[25] 국가AI컴퓨팅센터는 2028년까지 AI 반도체 1만 5천 장을 갖출 계획이다.[26] 기술 개발과 실증이 이 정도 규모에서 시작하는 것은 자연스럽다. 다만 이 기술이 수만 장, 수십만 장 규모의 실제 시설에서 쓰이는 단계는 아직 계획에 담기지 않았다.

실증과 대규모 시설 사이의 연결 고리를 강화할 필요도 있다. 국가AI컴퓨팅센터는 당초 2030년까지 국산 AI 반도체 50% 도입' 의무를 두었다가 재공모 과정에서 이를 폐지했고, 검증을 통과한 국산 칩만 별도 구역에서 상용 서비스에 투입하는 방식으로 바꿨다. GW급 민간 시설은 해외 임차인이 자기 운영 기술로 돌리는 구조가 될 가능성이 크다(3.3절 참조). 이 경우 AIDC는 국내 기술 없이도 운영되지만 대규모 운영 역량은 임차인에게 쌓인다. 실증을 통과한 국산 기술이 중간 규모를 거쳐 대규모 시설로 확대되는 단계를 계획에 포함해야 운영 역량도 국내에 축적된다.

3.3 보완이 필요한 네 영역: 수요, 국산 장비 생태계, 운영 계층, 활용 지표

소버린 AI 풀스택의 관점에서 보면 AIDC 메가프로젝트는 아래 계층에서 구체적인 지원책을 갖췄고, 위

https://www.datacenterdynamics.com/en/news/openai-and-oracle-to-deploy-450000-gb200-gpus-at-stargate-abilene-data-center/

24) Meta, “The Llama 3 Herd of Models”, arXiv:2407.21783, 3.3.4 Reliability and Operational Challenges, Jul. 2024. https://arxiv.org/abs/2407.21783

25) 이투데이, “4000억 규모 AI반도체 기반 K-클라우드 기술개발 프로젝트 추진”, 2025.5.8. https://www.etoday.co.kr/news/view/2468516

26) 한국데이터경제신문, “국가AI컴퓨팅센터 해남 착공, 공공지분 29%·GPU 1만5천 장의 설계”, 2026.10.1. https://www.dataeconomy.co.kr/news/articleView.html?idxno=42903

계층은 앞으로 채워 나가야 할 영역으로 남아 있다. AIDC를 둘러싼 계층별로 메가프로젝트가 다루는 범위를 그림 4에 표시했다. 전력과 입지, 전력과 냉각 설비, IT 장비와 AI 반도체에는 구체적인 지원책이 있지만, 클라우드 운영 계층은 실증 수준이고 모델과 서비스, 데이터 계층은 아직 AIDC와 연계되어 있지 않다. 아래 계층에서도 국산 장비가 글로벌 공급망 생태계에 본격적으로 진입하기 위해서는 넘어야 할 관문이 있다.

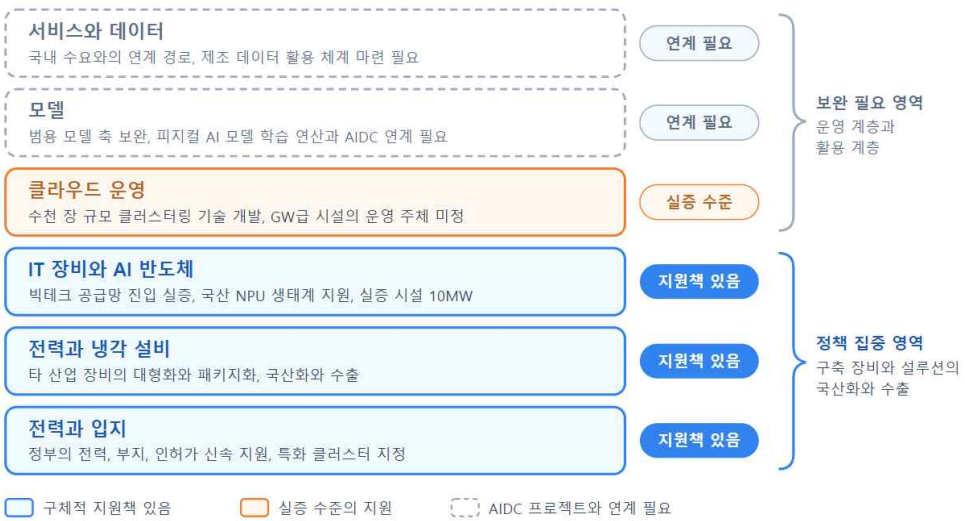


그림 4 소버린 AI 풀스택 관점에서 본 AIDC 메가프로젝트의 범위

첫째, 국내 수요와의 연계다. 메가프로젝트 안에는 대규모 연산 수요가 이미 있다. 피지컬 AI 분야는 가상 환경에서 학습 데이터를 대량 생산하는 합성데이터 인프라, 물리 법칙에 맞는 데이터를 생성하는 월드모델, 10대 업종 특화 한국형 파운데이션 모델, 범정부 데이터 라이브러리 구축을 담고 있다. 모두 대규모 학습 연산을 요구하지만 AIDC 분야는 이 수요를 어느 시설이 어떤 조건으로 수용할지 아직 다루지 않는다. 메가프로젝트 내부의 수요부터 AIDC 계획과 연계하는 것이 첫 번째 보완 과제다.

둘째, 국산 장비 생태계의 확장이다. 메가프로젝트는 IT 장비와 전력, 냉각 설비의 국산화를 지원하지만, GW급 AIDC에 들어갈 장비의 규격은 사실상 엔비디아가 정한다. 엔비디아는 GPU만 공급하는 것이 아니라 GPU가 들어가는 메인보드와 서버 트레이, 랙의 규격까지 자사의 모듈형 서버 참조 설계인 MGX로 정한다.

현재 이 생태계에서 국산 장비의 위치는 미미하다. 엔비디아 GB200 랙 시스템은 폭스콘, 콴타, 위스트론 같은 대만 제조사와 슈퍼마이크로, 델, HPE 같은 미국 업체가 주로 공급하며 국내 업체는 이

명단에 보이지 않는다.[27] 엔비디아가 2025년 10월 차세대 베라 루빈 랙의 전면 액체 냉각 설계와 800V 직류 배전 체계를 제시하며 발표한 전력과 냉각 협력사 명단에도 ABB, 슈나이더 일렉트릭, 버티브, 이튼 등이 올랐지만 국내 기업은 없었다.[28]

진입 장벽은 더 높아질 수도 있다. 베라 루빈 세대부터 엔비디아가 연산 트레이를 완성품으로 공급해 서버 제조사의 역할이 설계에서 조립과 통합으로 좁아질 것이라는 분석이 있다. 반면 엔비디아가 2026년 10월이 랙 설계를 오픈 컴퓨트 프로젝트(OCP: Open Compute Project)에 공개하면서 부품 업체가 같은 규격에 맞춰 개발할 길은 넓어졌다.[29] OCP는 데이터센터 하드웨어 설계를 공개 표준으로 공유하는 업계 단체다. 국산 장비가 GW급 AIDC에 들어가려면 먼저 엔비디아의 규격과 검증을 통과해야 하고, 그다음에 하이퍼스케일러가 이를 채택해야 한다. 국내에 대규모 AIDC가 들어서는 지금이 이 두 관문을 넘어야 하는 도전이자 기회며, 메가프로젝트는 이 기회가 국산 장비 생태계의 확장으로 이어지도록 설계되어야 한다.

셋째, 클라우드 운영 계층의 주체를 정하는 일이다. 운영은 민간의 몫이고, 장비 전략은 해외 클라우드 사업자의 공급망 진입을 목표로 한다. 이 전략은 GW급 시설을 실제로 운영할 주요 주체로 해외 하이퍼스케일러를 상정한 설계다. 2025년 6월 SK텔레콤과 AWS가 출범시킨 울산 AI 데이터센터(초기 103MW, GPU 약 6만 장 수용)는 GW급 확장을 목표로 내걸었고, 메가프로젝트 1단계의 울산 1GW도 이 구상과 맞닿아 있다.[30] 이런 구조에서는 시설이 국내에 있어도 그 위에서 클러스터를 운영하고 서비스를 제공하는 소프트웨어와 노하우는 임차인의 것이 된다. 이 계층에 국내 기업이 참여할 경로를 함께 설계할 필요가 있다.

넷째, 활용 성과를 재는 지표다. 국민보고회 자료가 내세운 목표와 기대 효과는 투자액, 일자리, 아태 점유율, 솔루션 대표기업 배출처럼 모두 투입과 공급 측 지표다. 정부가 '토큰 공장'을 표방하면서도 토큰 처리량, 시설 가동률, 국내 수요의 연산 이용 비중, 국산 반도체 활용 비중 같은 활용 지표는 아직 정하지 않았다. 지표가 GW와 투자액에 머무르면 정책은 '다 지었는가'를 묻고 '무엇을 위해 쓰였는가'는

27) IntuitionLabs, “NVIDIA GB200 Supply Chain: The Global Ecosystem Explained”, Dec. 22, 2025. https://intuitionlabs.ai/articles/nvidia-gb200-supply-chain
28) NVIDIA, “NVIDIA, Partners Drive Next-Gen Efficient Gigawatt AI Factories in Buildup for Vera Rubin”, NVIDIA Blog, Oct. 13, 2025. https://blogs.nvidia.com/blog/gigawatt-ai-factories-ocp-vera-rubin
29) Forbes, “Nvidia Opens Its Vera Rubin Rack Design To Speed Up AI Data Centers”, Oct. 14, 2026. https://www.forbes.com/sites/jonmarkman/2026/10/14/nvidia-opens-its-vera-rubin-rack-design-to-speed-up-ai-data-centers/
30) CEO스코어데일리, “SKT-AWS, 울산데이터센터에 7조 투자…최태원 “AI 고속도로의 강력한 엔진 될 것””, 2025.6.20. https://www.ceoscoredaily.com/page/view/2025062015530114662

묻지 않게 된다. 구축 지표에 활용 지표를 더해야 계획의 성과를 제대로 점검할 수 있다.

4. 기대 효과와 소버린 AI 풀스택 관점의 실행 방향

4.1 산업 생태계에 대한 기대 효과

GW급 AIDC가 국내에 들어서는 것은 국내 장비 산업이 해외 시장 진출에 필요한 실적을 쌓을 수 있는 기회다. 수배전 설비, 변압기, 무정전 전원 장치, 액체 냉각 같은 전력과 냉각 장비는 대규모 현장에서 장기간 안정적으로 운용된 실적이 있어야 해외 시장에 진입할 수 있다. 국내 GW급 현장에서 국산 장비가 활용된 실적을 만든다면 정부가 목표로 한 수출 산업화의 출발점이 된다.

AI 반도체의 안정적인 초기 수요도 생긴다. AIDC와 피지컬 AI에서 대규모 AI 반도체 수요가 나온다. 이 수요의 일부가 국산 NPU와 국산 시스템 소프트웨어로 충족된다면 메모리에 편중된 반도체 산업이 시스템 반도체로 넓어지는 계기가 될 수 있다. 정부도 반도체 메가프로젝트 안에 국산 AI 모델 특화 반도체와 소프트웨어를 함께 지원하는 'K-AI 반도체 풀스택'을 포함시켰다.

지역과 인재 측면의 효과도 있다. 수도권에 집중된 데이터센터를 비수도권으로 분산하고 지역 거점 대학과 연계해 GPU와 MLOps 엔지니어, 장비 개발과 유지보수 인력을 양성하는 계획은 지역의 고용과 인재 기반을 넓힐 수 있다.

4.2 성과로 이어지기 위해 풀어야 할 과제

첫 과제는 부가가치를 국내에 남기는 것이다. 맥킨지는 2030년까지 전세계 AIDC에 필요한 5.2조 달러 투자 가운데 60%가 칩과 컴퓨팅 하드웨어를 포함한 기술 개발과 설계 부문으로, 25%가 발전과 송전 부문으로, 15%가 부지와 건설 부문으로 흘러간다고 추정했다.[31] 이 비율을 그대로 적용하면 550조 원 가운데 가장 큰 몫은 수입 GPU와 서버로 향한다. 임차형 구조라면 그 위의 운영 소프트웨어, 모델, 서비스의 가치도 임차인에게 귀속된다. 국내에 남는 것이 부지 임대료와 전력 판매, 건설 효과에 그치지 않으려면 운영과 서비스 계층에서 국내 기업의 참여를 늘려야 한다.

빅테크 공급망 진입도 풀어야 할 과제다. 3.3절에서 보았듯 첫 관문은 엔비디아의 규격과 검증이고, 다음 관문은 자체 설계 표준과 장기 공급 계약으로 장비를 조달하는 하이퍼스케일러다. 정부 지원도

31) McKinsey & Company, “The cost of compute: A $7 trillion race to scale data centers”, Apr. 28, 2025. https://www.mckinsey.com/industries/tmt/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers

국내 실증 실적에 더해 엔비디아 규격 인증, 협력사 명단 진입, 하이퍼스케일러 납품처럼 단계별 목표를 두는 방향으로 설계할 필요가 있다.

국산 NPU와 그 소프트웨어 생태계를 쓸 유인을 만드는 일도 과제다. NPU는 칩 성능 못지않게 컴파일러, 런타임, 서빙 소프트웨어의 성숙도가 채택을 좌우하며 이 성숙도는 실제 서비스를 운영하면서 높아진다. 그러나 민간 임차인에게는 검증된 GPU 생태계를 두고 국산 NPU를 택할 이유가 없고, 공공 부문에서도 국가AI컴퓨팅센터의 국산 도입 의무는 제외됐다. 수요 측 유인이 함께 설계되어야 공급 측 연구개발이 실증을 넘어 상용화로 이어진다.

데이터 주권을 지키면서 산업 현장의 AI 활용도를 높이는 일도 과제다. 제조 현장의 공정 데이터와 숙련자의 암묵지는 피지컬 AI의 핵심 원료이므로 이를 현장이나 국내 시설에 두고 AI를 운용하는 온프레미스 방식이 필요하다. 특히 비수도권에 짓는 GW급 시설이 지역 중소기업의 AI 전환으로 이어질 수 있는 지원책도 함께 마련해야 한다. 몇몇 앵커 기업이 주도하고 여기서 나오는 결과물을 지역산업 AX에 활용하는 방식도 고려해 볼 만하다.

마지막으로 전력, 용수, 정책금융 같은 희소 자원을 반도체 팹과 AIDC가 나눠 써야 한다. 자원이 경합하는 만큼 모든 AIDC 수요를 같은 우선순위로 다룰 수 없다. 수요 유형을 구분해야 하는 중요한 이유다.

4.3 수요 유형 구분: 18.4GW를 세 유형으로 나누어 계획하기

구현의 출발점은 하나의 숫자로 묶인 18.4GW를 성격이 다른 세 유형의 수요로 나누는 것이다. 유형마다 소버린 요건과 지원 방식, 용량을 정하는 기준이 다르기 때문이다.

- **국가 필수 수요** : 공공 행정, 국방, 국가 AI 모델 학습, 그리고 메가프로젝트 내부의 피지컬 AI 합성데이터와 월드모델 학습이 여기에 속한다. 데이터와 운영의 통제권이 국내에 있어야 하는 영역이므로 국내 사업자의 운영과 함께 국산 반도체를 적극적으로 투입하는 것을 고려한다.
- **산업 AX 수요** : 제조업 AI 전환, 산업단지, 중소기업의 추론 수요다. 산단 공용 엣지 AIDC, 공용 추론 엔드포인트 구축, 또는 추론 이용료 지원을 중심으로 대응한다.
- **해외 임대 수요** : 해외 빅테크의 학습과 추론 수요다. 상업적 조건을 기반으로 유치하되 고객 계약에 맞춰 규모를 단계적으로 키운다.

어느 유형이 더 중요하고 우선 대응해야 한다고 단언할 수는 없다. AIDC 구축을 앞당기는 관점에서는 해외 임대 수요가 먼저일 수 있고, 산업 AX 수요가 국가 AI 모델 학습 수요보다 먼저 커질 수도 있다.

다만 국가 필수 수요와 산업 AX 수요가 실제로 확인되면 이를 우선 대응하는 것이 맞다. 전력 계획도 유형별로 반영 기준을 달리할 수 있다. 국가 필수 수요는 사업 계획에 근거해 우선 반영하고, 산업 AX 수요는 산업 AX 계획에 맞춰 단계적으로 반영하며, 해외 임대 수요는 계약이 확정된 물량만 반영하는 방식이다. 따라서 정부는 국가 필수 수요와 산업 AX 수요의 규모를 가능한 한 빨리 산출해 공개해야 한다.

4.4 소버린 AI 풀스택 계층별 실행 방향

소버린 AI 풀스택의 핵심은 AI 반도체, 모델, 그리고 서비스 및 데이터이다. 여기에 인프라 핵심인 IT 장비와 전력/냉각, 그리고 클라우드 운영도 포함시킬 수 있다. 계층별 현황과 구현 방향을 〈표 2〉에 정리했다.

계층	현황과 과제	구현 방향
IT 장비와 전력, 냉각	엔비디아 랙 생태계 내 국산 장비 위치 미미, 서버 국산화율 3.3%[32]	엔비디아 규격 인증과 협력사 진입을 지원하고, 정부가 확보한 시설과 산단 엣지 AIDC에 인증을 통과한 국산 장비를 우선 적용
AI 반도체	국가AI컴퓨팅센터의 국산 도입 의무 제외	학습은 GPU를 활용하되, 국가 필수 수요와 산업 AX 수요의 추론 영역에 국산 NPU 용량을 배정
클라우드 운영	수천 장 규모의 클러스터링 목표, GW급 시설의 운영 주체 미정	정부가 확보한 연산의 운영을 국내 클라우드 사업자와 국산 운영 소프트웨어에 맡겨 대규모 운영 경험을 축적
모델	메가프로젝트에 범용 모델 축이 없음	기업 주도 독자 파운데이션 모델 확보와 더불어 글로벌 최상위 프론티어 모델 확보를 위한 정부 차원에서의 지원 전략 구체화
서비스와 데이터	독파모 + NPU 성능을 벤치마크 위주로 평가	NPU, 모델, 서비스 연구/개발 주체가 모두 참여하는 소버린 AI 풀스택 얼라이언스를 설립, 이를 통해 실제 다양한 서비스를 통해 체감할 수 있는 성능의 풀스택 AI 확보

표 2 소버린 AI 풀스택 계층별 현황과 구현 방향

각 계층은 서로 유기적으로 연결되어 있다. 특히 반도체, 모델, 서비스를 결합하여 실제 손에 잡히는 성능을 다양한 서비스 영역에서 체감할 수 있도록 하는 것이 중요하다. 독자 AI 파운데이션 모델 사업에서 중요한 것은 최종 두 개 모델을 고르는 일 자체가 아니다. 선정된 모델이 공공과 국방처럼 AI 주권이 필요한 핵심 영역에서 충분한 성능을 내도록, 후속 과제를 메가프로젝트 안에서 이어 가는 것이 중요하다. 이것이 소버린 AI 풀스택 확보의 관건이 될 것이다.

32) 한국데이터경제신문, “56억 달러 AI 냉각 시장이 열리는데, 서버 국산화율 3.3%에 갇힌 한국의 선택지”, 2026.5.11. https://www.dataeconomy.co.kr/news/articleView.html?idxno=40045

5. 향후 과제와 시사점

3대 메가프로젝트의 AIDC는 전력과 부지, 장비 공급망이라는 아래 계층에서 구체적인 지원책을 갖췄다. 남은 과제는 그 위에서 무엇이 돌아가고 누가 운영하며 무엇으로 성과를 측정할지를 계획에 담는 일이다. 18.4GW라는 목표에 국내 수요의 근거를 더하고, 국산 기술이 실증을 넘어 대규모 시설에서 쓰이는 단계를 설계하고, 운영, 모델, 서비스 계층과 활용 지표를 보완해야 한다. 그래야 GW급 시설이 지어졌을 때 운영 역량과 부가가치가 국내에도 쌓인다.

계획이 실제 성과로 이어지기 위한 제언은 다음과 같다. 첫째, 18.4GW를 수요 유형별로 구분하고 전력 계획도 유형별 반영 기준을 달리 적용해야 한다. 둘째, 국민성장펀드와 인허가, 전력 지원을 연산 선구매와 연계해 국가 필수 수요와 산업 AX 수요를 위한 연산을 확보하고, 그 운영을 국산 반도체와 운영 소프트웨어에 맡기는 제도를 설계해야 한다. 셋째, 성과 지표를 자원 투입 중심에서 활용 중심으로 바꾸어야 한다. 올바른 성과 지표 설정은 메가프로젝트 성공을 위해 가장 중요한 과정이다. 과거 관행처럼 이루어졌던 공급자 관점의 정량적 지표가 아닌 각 수혜 영역에서의 정교한 루브릭(평가 기준표)을 바탕으로 한 실제 '효능감' 중심으로 설계해야 한다.

메가프로젝트의 성패는 AIDC의 연산이 국내 산업과 공공의 문제를 얼마나 풀고 그 과정에서 우리 반도체와 소프트웨어, 모델이 얼마나 성장하는가로 판가름 난다. 2029년에 몇 GW가 가동되는가는 성패의 기준이 아니다. AIDC를 짓는 계획에 그 시설을 운영하고 활용하는 계획이 더해질 때 18.4GW는 설비 용량의 숫자에 그치지 않고 국가의 AI 역량으로 이어진다.

04 2026년 클라우드 솔루션 리포트: 추론 인프라 (2) 토큰 공급

| 정채상 이음테크

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.
한국지능정보사회진흥원이 저작권을 보유하고 있으며 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

들어가며: 매진된 토큰 공장과 쓰이지 않는 GPU

데이터센터에 GPU를 놓고 전기를 넣어 토큰만을 만들어 파는 시설을 이 글은 토큰 공장이라 부른다. 이전 보고서가 그 공장에서 토큰을 사는 쪽의 이야기였다면, 이번 보고서는 토큰 공장을 짓고 돌리는 쪽, 곧 토큰을 공급하는 쪽의 이야기다. 데이터센터에 GPU를 놓고 전기를 넣어 토큰만을 만들어 파는 시설을 이 글은 토큰 공장이라 부른다.

2026년의 토큰 공장은 서로 반대 방향을 가리키는 두 개의 숫자로 요약된다. 하이퍼스케일러가 아니면서 GPU 임대를 주업으로 삼는 신생 클라우드 사업자를 네오클라우드라 부른다. 그 가운데 네비우스(Nebius)는 2분기 실적 발표에서 새로 가동한 GPU 용량이 들어오는 대로 계약되어 그 분기에도 남는 용량이 없고, 메가와트당 4,000만~5,000만 달러의 단기 계약을 협상 중이라고 밝혔다.[33] 반면 같은 해 봄 쿠버네티스 비용 최적화 업체 캐스트 AI(Cast AI)의 보고서에서 기업이 배정해 둔 GPU의 평균 가동률은 5%였다.[34] 같은 GPU가 공장 쪽에서는 웃돈을 받고 매진되는데, 역설적으로 빌려 간 쪽에서는 용량의 95%가 쓰이지 않고 있다.

두 숫자가 한 해에 함께 나오는 것은 토큰 공장이 GPU가 쓰일 때가 아니라 팔릴 때 돈을 받기 때문이고, 약정 계약을 맺은 구매자가 GPU를 쓰지 않아도 공장의 매출은 줄지 않는다. 구매자 쪽의 평균은 이전 보고서가 전용 계약과 서버리스의 경계로 본 가동률 40~50%의 10분의 1이니, 공장의 사정은 공장을 짓는 회사만의 문제로 끝나지 않는다.

33) Nebius. “Nebius Group N.V. Q2 2026 Earnings Call Transcript.” 2026.08.12.
34) Cast AI. “2026 State of Kubernetes Optimization Report.” 2026.04.21.

토큰 공장을 무엇으로 재고 어떻게 설명할지는 여전히 칩을 파는 회사가 정하고 있다. 엔비디아의 젠슨 황은 2026년 3월 GTC 기조연설에서 "모든 데이터센터, 모든 공장은 본질적으로 전력에 묶여 있다. 1기가와트 공장은 결코 2기가와트 공장이 되지 못한다"고 말했고[35], 엔비디아는 공장의 매출을 확보한 전력과 와트당 토큰 수와 토큰 값의 곱으로 설명한다.[36] 공장을 짓는 쪽의 말로 옮기면 전력은 수년 앞서 잡아 두어야 하는 항이고, 와트당 토큰은 칩과 소프트웨어로 당장 겨룰 수 있는 항이다. 이번 보고서에 나오는 회사들은 이 둘 가운데 어느 쪽에 중점을 두었는가로 갈린다.

이번 보고서의 질문은 셋이다. 1) 토큰 공장의 손익계산서는 어떻게 생겼는가, 2) 왜 모두가 자기 층을 벗어나려 하는가, 그리고 3) 그 사정이 토큰을 사는 쪽의 계약과 단가에 어떤 경로로 전해지는가. 이 보고서는 세 질문에 차례로 답해 나간다. 먼저 토큰 공장의 손익계산서를 읽고, 그 위에서 여섯 플레이어가 각자 어디에 중점을 두었는지를 따라간 뒤, 그 사정이 토큰을 사는 쪽의 계약서에 닿는 지점을 제언으로 정리한다.

토큰 공장의 손익계산서

매출, 원가, 이익



그림 5 추론 산업의 세 개 층과 이번 보고서의 범위.

35) NVIDIA. “GTC 2026 Keynote Transcript”. 2026.03.16.
36) NVIDIA Developer Blog. "Scaling Token Factory Revenue and AI Efficiency by Maximizing Performance per Watt." 2026.03.25.

위의 그림 5에서처럼 이번에 보는 것은 이전 보고서가 세 층짜리 건물에 빗댄 추론 산업 가운데 아래 두 층, 같은 GPU에서 토큰을 더 뽑아내는 추론 인프라와 연산 시간을 파는 AI 데이터센터이다. 이 둘을 하나로 묶은 것이 이 글의 토큰 공장이고, 완제품을 파는 맨 위층의 제품 이름 중 하나인 토큰 팩토리와는 다른 용어이다. 토큰 공장의 손익계산서는 세 항목으로 읽을 수 있는데, 1) 매출은 확보한 전력에 와트당 토큰과 토큰 단가, 그리고 그 전력이 실제로 팔린 비율인 가동률을 곱한 값이고, 2) 원가는 GPU를 몇 년에 나눠 비용으로 처리하는 감가상각과 전력과 이자를 더한 값이며, 3) 이익은 둘의 차이다. 매출을 이루는 네 변수 가운데 전력은 지어 놓은 뒤에는 늘릴 수 없고, 와트당 토큰은 칩과 소프트웨어를 바꿔야 오르며, 단가는 시장이 정하므로, 공장이 영업으로 당장 움직일 수 있는 변수는 가동률뿐이다. 건물과 GPU에 들인 자본은 그 용량이 팔린 만큼만 매출로 돌아오는데, 이 글에서 공장의 가동률은 확보한 전력 가운데 계약되어 청구되는 비율을 뜻하고, 공장이 매진됐다는 말은 이 비율이 100%에 이르렀다는 뜻이다.

이 항목들이 실제로 어떻게 보이는지는 미국의 네오클라우드 코어위브(CoreWeave)의 2026년 2분기 실적이 알려 준다. 이자와 세금과 감가상각을 빼기 전의 이익률인 조정 EBITDA 마진은 59%였는데[37], 감가상각을 빼고 나면 영업이익 마진은 5%로 내려앉고, 거기에 매출의 25%에 이르는 이자비용이 얹히면서 분기 실적은 순손실로 끝난다. 매출의 절반 남짓을 설비값이 가져가고 남은 이익은 빚값이 가져갔으니, 토큰 공장의 매출이 아무리 좋아도 결과는 비용이 정한다.

단가는 떨어지는데 토큰 공장은 매진

GPU	온디맨드 하이퍼스케일러	온디맨드 네오클라우드	온디맨드 75개 사업자 중앙값	1년 약정 2025년 10월 → 2026년 3월
H100	7.15	2.80	3.25	1.70 → 2.35
H200	미집계	3.34	4.40	미집계
B200	미집계	5.88	6.52	미집계

표 3 GPU 한 장의 시간당 임대 단가(달러).
출처: 실리콘 데이터(2026년 10월 8일), AIMultiple(2026년 9월), 세미애널리시스.

구매자가 체감하는 단가부터 보면, GPU 임대 시세를 원자재 가격처럼 날마다 집계해 지수로 내는 실리콘 데이터(Silicon Data)의 2026년 10월 8일 값에서 하이퍼스케일러의 H100 단가는

37) CoreWeave. "CoreWeave Reports Strong Second Quarter 2026 Results." 2026.08.11.

네오클라우드 중심 값의 2.5배를 넘는다.[38] 길게 보면 H100(호퍼 세대)의 온디맨드 단가는 2024년 초의 절반 이하로 내려왔다.[39] 그런데 시장 조사 업체 세미애널리시스(SemiAnalysis)가 1년 약정의 체결 단가만 따로 모아 다달이 내는 H100 1년 약정 지수는 2025년 10월 저점에서 2026년 3월까지 약 40% 반등했다.[40] 세미애널리시스는 블랙웰 세대의 인도가 늦어져 2026년 8~9월 물량까지 이미 예약된 데서 원인을 찾는데, 온디맨드는 내리고 계약은 오르는 어긋남이 2026년 상반기 시장의 모양이다. 표 1은 이 세 지수를 한자리에 모은 것인데, 같은 H100이라도 어디서 어떤 계약으로 빌리느냐에 따라 시간당 1.70달러에서 7.15달러까지 벌어진다.

계약 단가가 오르는 것은 매진된 토큰 공장이 받는 웃돈이다. 들어가며에서 본 네비우스의 메가와트당 4,000만~5,000만 달러가 그 예이고, 코어위브도 2026년 9월, 3분기에 맺은 3개월에서 6개월짜리 단기 계약의 단가가 1년 치로 환산하면 메가와트당 약 4,000만 달러라고 밝혔다.[41] 단순 계산으로 블랙웰급 GPU 한 장당 시간당 7달러 안팎이니 표 1의 네오클라우드 온디맨드 단가보다 높다. 단가가 떨어지는데도 공장이 매진인 것은 세대 교체가 빠른 탓으로, 구형 GPU는 새 세대가 나올 때마다 값이 내려가고 신형은 공급이 모자라 계약 시장에서 웃돈이 붙는다. 그래서 지금 H100을 싸게 잡을 것인가, 블랙웰을 기다릴 것인가라는 구매자의 질문은 공장의 감가상각 곡선 가운데 어느 지점에서 계약을 맺느냐의 문제가 된다.

토큰당 원가와 감가상각이라는 하한

세미애널리시스의 원가 분석 도구 인퍼런스X(InferenceX)로 GPU 시간당 단가를 토큰당 원가로 바꿔 보면 마진의 폭이 드러난다. 하이퍼스케일러가 직접 보유한 GPU로 한 오픈 모델을 서빙할 때 100만 토큰당 원가는 B200이 0.05~0.15달러, H100이 0.17~0.40달러다.[42] GPU는 요청을 많이 묶을수록 토큰당 원가가 내려가는 대신 한 사람의 응답이 느려지므로, 구간의 폭은 사용자가 체감하는 속도에서 생기고 속도를 올릴수록 원가가 가파르게 뛴다. 오픈 모델의 서버리스 판매가는 이전 보고서에서 본 대로 100만 토큰당 0.1~1달러여서 판매가가 원가 위로 벌어지는 폭이 토큰 공장의 마진이고, 두 구간이 겹치는 아래쪽에서는 그 마진이 거의 사라진다.

38) Silicon Data. "Silicon Index: H100 / H200 / B200 rental price indices." 2026.10.08
39) AIMultiple. "GPU Index: Cloud GPU Pricing." 2026.09.16
40) SemiAnalysis. "The Great GPU Shortage – Rental Capacity – Launching our H100 1 Year Rental Price Index." 2026.04.02.
41) CoreWeave. "CoreWeave Continues to Contract New Compute Capacity at Higher Prices." 2026.09.17.
42) SemiAnalysis InferenceX. "MiniMax M2.5/M2.7: B200 vs H100 cost per million tokens." 2026.07.

원가의 나머지 절반은 회계, 곧 GPU를 몇 년에 나눠 비용으로 처리하느냐에서 정해진다. 코어위브와 알파벳은 6년을 쓰고, 아마존은 2025년 초 일부 서버를 6년에서 5년으로 되돌리면서 AI 기술의 발전 속도를 이유로 들었다.[43] GPU의 실제 수명이 2~3년이라는 주장과, 학습에서 추론으로, 다시 가벼운 내부 워크로드로 내려가며 더 오래 쓰인다는 반론이 맞서 있는데, 이 논쟁이 회계의 문제로 끝나지 않는 것은 감가상각 기간이 손실을 보고하지 않고 내려갈 수 있는 단가의 하한을 정하기 때문이다. 전력과 이자처럼 현금이 나가는 원가는 같으므로 6년을 쓰는 회사는 더 낮은 단가에서도 이익을 보고하며 경쟁할 수 있지만, 그 이익은 GPU가 실제로 6년을 번다는 전제 위에 있다. 수명이 회계보다 짧은 것으로 드러나면 자산 가치를 장부에서 깎아 내는 충격을 먼저 받는 쪽도 6년으로 설정한 회사들이다.

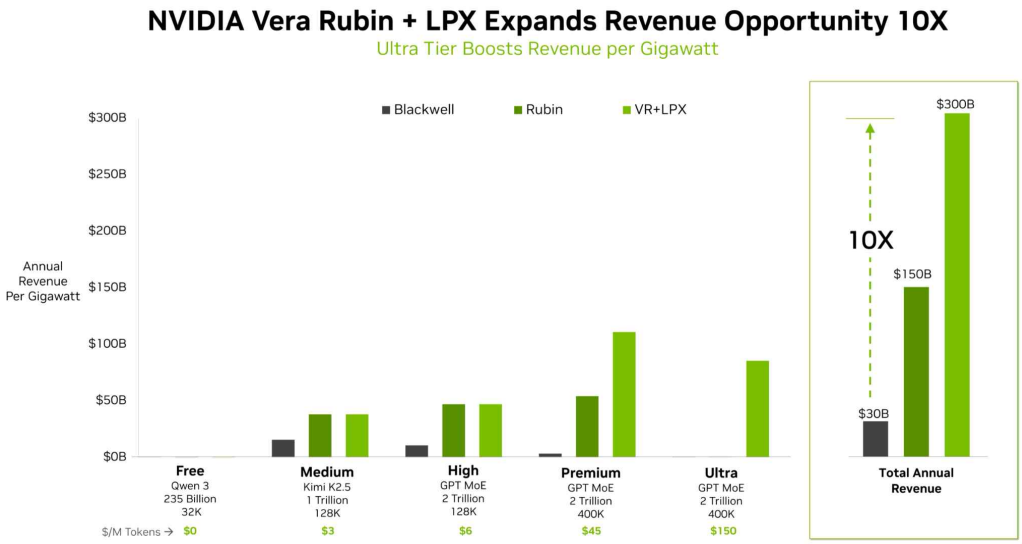


그림 6 1기가와트 공장의 가격 등급별·GPU 세대별 연매출 잠재치
[출처: 엔비디아 개발자 블로그(2026년 3월), 엔비디아 자체 추정]

이 모든 숫자의 분모에는 전력이 있다. 엔비디아는 2026년의 루빈(Rubin) 세대가 같은 전력에서 블랙웰의 최대 10배에 이르는 토큰을 만들고 토큰당 원가도 약 10분의 1로 내려간다고 주장한다. 전력이 고정되어 있다면 매출을 올리는 길은 같은 전력 위에서 GPU를 새 세대로 교체하는 것이므로, 토큰 공장의 매출은 GPU의 수가 아니라 세대 교체로 늘어난다. 다만 새 세대로 교체하려면 구형 GPU가 6년을 채우기 전에 자리를 비워야 하므로, 이 성장 방식은 앞의 감가상각 논쟁과 정면으로 부딪치는데, 다음 장에서 볼 회사들 가운데 상당수가 전력부터 향하는 것도 그 때문이다.

43) Deep Quarry. "Depreciation of GPUs: Between Useful Life and Wishful Thinking." 2025.12.07.

마진은 위층에, 전력은 아래층에

손익계산서를 층별로 나눠 보면 잠재 마진은 위층에 있고 전력은 아래층에 있는 비대칭이 보이는데, 2026년의 토큰 공장들이 모두 자기 층을 벗어나려 하는 것도 그 때문이다(그림 7). 네오클라우드는 추론 API와 소프트웨어 회사를 인수해 위로 올라가고, 토큰을 파는 회사는 투게더 AI(Together AI)가 확보한 500메가와트처럼 자체 GPU와 전력을 확보하러 아래로 내려간다. 칩 회사는 자기 칩 위에서 돌아가는 클라우드와 마켓플레이스를 직접 열고, 하이퍼스케일러는 자체 칩을 만들어 아래층의 원가를 엔비디아 가격표에서 떼어낸다.



그림 7 수직 통합의 네 방향. 실리콘 공급을 맨 아래에 두고 그 위의 세 개 층에서 네오클라우드는 위로, 토큰 팩토리는 아래로, 칩 회사는 클라우드로, 하이퍼스케일러는 자체 칩으로 움직인다.

네 방향의 길이 똑같이 열려 있지는 않다. 코어위브가 데이터센터 사업자 코어 사이언티픽(Core Scientific)을 인수해 1.3기가와트 전력을 확보하려던 계획은 2025년 10월 상대 회사 주주들이 부결시켰다.[44] 반대로 같은 회사가 위로 올라가는 인수, 곧 실험 관리 도구와 강화학습 스타트업의 인수는 그대로 성사됐다.[45] 전력을 가진 쪽은 이제 그 자산의 값을 알아 쉽게 팔지 않으니, 전력은 사기보다 스스로 지어야 하는 것이 되었다. 그렇다면 같은 GPU를 사서 같은 전력 제약 위에 선 토큰 공장들은 무엇으로 차별을 두어야 하는가? 갈아탈 수 없는 이유를 만드는 것이 곧 구매자를 붙잡는 방법이기도 하다.

44) Yahoo Finance. "Core Scientific shareholders vote down CoreWeave merger." 2025.10.30.
45) CoreWeave. "CoreWeave to Acquire OpenPipe, Leader in Reinforcement Learning." 2025.09.03.

주요 벤더별 솔루션 심층 분석

여섯 플레이어를 네오클라우드 셋, 칩 회사, 특화 칩 진영, 하이퍼스케일러 자체 칩의 순서로 본다. 이 가운데 한국에서 직접 접하기에는 마지막 절의 하이퍼스케일러 이야기가 가깝겠지만, 하이퍼스케일러 용량의 일부가 네오클라우드 사이트에 있고, 단가의 바닥을 네오클라우드 지수가 정하며, 칩과 소프트웨어 값은 칩 회사가 받기 때문에 함께 살펴볼 필요가 있다. 각 절은 그 플레이어가 세상을 보는 방식에서 시작해 구매자의 계약서에 닿는 지점까지 이야기한다.

코어위브: 규모가 곧 담보라고 믿는 공장

코어위브가 보는 세상에서 이 사업의 승부는 규모에서 난다. 전력을 먼저 계약하고, 그 위에 새 세대 GPU를 남보다 먼저 놓고, 그 용량을 대형 고객에게 장기로 팔아 건설 자금을 조달하는 순환을 누구보다 빨리 돌리는 것이다. 2026년 2분기 말 수주잔고, 곧 계약은 했지만 아직 매출로 잡히지 않은 금액은 1,042억 달러다. 이 순환은 공장 용량의 큰 몫을 장기로 약정해 주는 앵커 고객 둘이 받친다. 오픈AI와의 계약은 최대 224억 달러에 이르고[46], 메타는 두 차례 계약으로 2032년 말까지 약 350억 달러를 약정했다.[47]

이 공장은 고객이 GPU와 서빙 소프트웨어를 고르는 데디케이티드 인퍼런스(Dedicated Inference)를 정식 출시해, 추론을 파는 위층에도 발을 들였다. 한편, 기간이 긴 계약들도 이어지고 있어 수주잔고 가운데 48개월 뒤에야 매출로 인식될 몫이 약 20%에 이르는데[48], 토큰 공장이 계약 기간을 감가상각 기간에 맞춘다는 것은 GPU 수명의 위험을 구매자와 나누겠다는 뜻이다.

약점은 이 순환이 빚과 단지 두 고객에 의존하고 있다는 것이다. 2026년 6월 말 기준으로 이자비용은 매출의 25%에 이르고, 상반기 매출의 63%가 상위 두 고객에서 나왔다. 이 중 하나라도 자체 칩이나 자체 데이터센터로 옮겨 가면 그 용량을 채울 고객을 다시 찾아야 하고, 그때까지 이자는 그대로 나간다. 구매자가 공장의 고객들을 걱정해야 하는 상황이 된다.

네비우스: 와트당 토큰에서 이기겠다는 공장

네덜란드에 본사를 둔 네비우스가 보는 세상에서 전력과 GPU는 누구나 같은 값에 사는 원자재이고, 승부는 같은 전력에서 토큰을 얼마나 더 뽑아내느냐에서 난다. 2026년 말까지 계약으로 확보할 전력의

46) CoreWeave. "CoreWeave Expands Agreement with OpenAI by up to $6.5B." 2025.09.25.
47) CoreWeave. "CoreWeave and Meta Announce $21 Billion Expanded AI Infrastructure Agreement." 2026.04.09.
48) CoreWeave. Form 10-Q, 2026년 2분기.

목표는 2025년 11월 이후 분기마다 올라 5기가와트가 됐는데, 이는 가동 중인 전력이 아니라 회사 전체가 여러 사이트에 걸쳐 계약으로 묶어 두려는 용량의 합이다. 앵커 고객은 마이크로소프트와 메타 둘인데[49], 코어위브와 갈리는 지점은 회계에 있다. 네비우스는 2026년 들어 감가상각 기간을 5년으로 늘렸지만 여전히 코어위브의 6년보다 짧게 잡고, 그만큼 같은 매출의 이익을 더 보수적으로 계산하는 공장이다.

이 공장이 중점을 두는 곳은 소프트웨어여서, 엔비디아가 퍼뜨린 토큰 팩토리라는 말을 자사 추론 서비스의 이름(Token Factory)으로 가져왔다. 2026년 5월에는 추론 최적화 스타트업 아이겐 AI(Eigen AI)의 인수에 합의하고, 같은 달 클라리파이(Clarifai)의 핵심 팀을 영입하고 추론 지식재산은 라이선스로 들여왔다.[50] [51] 회사는 경쟁 우위가 연산 자원의 양에서 소프트웨어가 정하는 추론 효율로 옮겨 간다고 설명하는데, 앞 장의 말로 하면 기가와트보다 와트당 토큰에서 이기겠다는 선언이다. 구매자의 화면에서는 GPU 클러스터와 토큰 API가 한 콘솔에 있어, 예측 가능한 워크로드는 시간당으로, 탐색적 워크로드는 토큰당으로 두는 선택을 벤더를 바꾸지 않고 할 수 있다. 다만 구매자의 시각에서 약정을 꼼꼼히 체크해야 한다.

반대로 위험 요소는 두 개가 있는데, 먼저 계약 전력 목표를 분기마다 올린 만큼 2027년부터 해마다 1기가와트 이상을 제때 세워야 하는 실행의 부담이 늘었고, 전력 확보나 GPU 인도 가운데 하나만 늦어도 약속이 밀리는 위험이 있다. 다음으로는 매출의 대부분이 마이크로소프트와 메타 두 곳에서 오는데 둘 다 자체 칩을 만드는 회사여서, 그들의 칩 로드맵이 곧 이 공장의 수요 위험으로 오게 된다.

크루소: 전력부터 짓는 공장

유휴 천연가스로 비트코인을 채굴하던 크루소(Crusoe)가 보는 세상에서 병목은 전력이다. 마이크로소프트의 사티아 나델라도 2025년 11월 초 보도된 팟캐스트에서 "지금 가장 큰 문제는 컴퓨팅 과잉이 아니라 전력이다"라고 말했는데[52], 크루소는 그 문제를 거꾸로 풀어, 부지와 발전소를 먼저 확보하고 그 위에 GPU를 얹는다. 텍사스 애빌린에서는 오라클이 임차해 오픈AI의 스타게이트에 제공한다는 캠퍼스에 이어[53], 마이크로소프트용 900메가와트 캠퍼스에 같은 용량의 현장 발전소를

49) Nebius. "Nebius Signs New AI Infrastructure Agreement with Meta." Business Wire, 2026.03.16.
50) Nebius. "Nebius agrees to acquire Eigen AI, strengthening Nebius Token Factory as a frontier inference platform." 2026.05.01.
51) Nebius. "Nebius welcomes Clarifai's core team and licenses inference IP to strengthen Nebius Token Factory." 2026.05.12.
52) Yahoo Finance(Tom's Hardware 전재). "Microsoft CEO says company doesn't have a chip supply problem, it has a power problem." 2025.11.02.
53) Crusoe. "Crusoe, Blue Owl Capital and Primary Digital Infrastructure Enter Joint Venture." 2025.05.21.

세운다.[54] 크루소가 2026년 9월 투자 유치를 발표하며 밝힌 총계약 용량은 6기가와트를 넘고, 그 가운데 1기가와트가 가동 중이다.[55]

이 공장은 2025년 11월 토큰당 과금과 예약 처리량을 파는 매니지드 인퍼런스(Managed Inference)를 내놓으며 발전소부터 토큰 API까지 세 층을 모두 갖췄다.[56] 다만 아직 상장 전인 회사가 기가와트 단위의 건설을 이어 가려면 사모 자본에 기대야 한다는 점과, 앵커가 스타게이트에 치우쳐 있다는 점이 위험 요소로 남아 있다. 구매자가 확인할 것은 계약 용량이 놓일 사이트에 전력이 이미 들어왔는가, 그리고 그 벤더의 이름 아래 실제로 그 벤더의 것인 자산은 무엇인가다.

엔비디아: 층을 다 가지려는 칩 회사

엔비디아가 보는 세상에서 토큰 공장들은 서로 경쟁하지만 어느 쪽이 이기든 그 공장은 엔비디아의 칩 위에 선다. 그런데 2026년의 엔비디아는 칩 위의 층마다 자기 자리를 만들어 왔다. 유통에서는 네오클라우드와 AWS·애저의 GPU 용량을 한 화면에 모아 개발자가 고르게 하는 마켓플레이스 DGX 클라우드 렙톤(DGX Cloud Lepton)으로 구매자와 공장 사이에 들어섰고[57], 소프트웨어에서는 서빙 프레임워크 다이나모(Dynamo)를 오픈소스로 풀어 표준을 쥔 뒤, 그 위에 모델을 바로 돌리는 패키지 NIM을 운영 환경에서는 유료 라이선스로 판다.[58] 실리콘에서는 2025년 12월 보도 기준 약 200억 달러에 비독점 라이선스로 들여온 그록(Groq)의 기술이 삼성 공정에서 만든 그록 3 LPU 칩과 LPX 시스템이 되어 2026년 8월 본격 양산에 들어갔고[59], 모델 유통에서는 2026년 9월 오픈 모델의 저장소 허깅 페이스(Hugging Face)를 129억 3,000만 달러에 인수하는 계약을 공시했다.[60]

특히 엔비디아는 실리콘의 조합으로 앞서려 하는데, 베라 루빈(Vera Rubin) GPU와 그록 3 LPX를 함께 배치하면 조 단위 매개변수와 긴 문맥의 워크로드에서 전력당 처리량이 최대 35배에 이른다고 본다. 입력을 한꺼번에 읽는 단계는 GPU가, 답을 한 토큰씩 써 내려가는 단계의 일부는 LPU가 맡아,

54) Crusoe. "Crusoe Announces New 900 MW AI Factory Campus in Abilene, Texas to Support Microsoft AI Infrastructure." 2026.03.27.
55) Crusoe. "Crusoe Raises $3.9 Billion Series F for its Vertically-Integrated AI Infrastructure Platform." 2026.09.17.
56) Crusoe. "Crusoe Launches Managed Inference, Delivering Breakthrough Speed for Production AI." 2025.11.20.
57) NVIDIA Newsroom. "NVIDIA Announces DGX Cloud Lepton to Connect Developers to NVIDIA's Global Compute Ecosystem." 2025.05.18.
58) NVIDIA. "Dynamo: datacenter-scale distributed inference serving framework."
59) Dataconomy. "Samsung mass produces Nvidia's Groq 3 inference chip on 4nm." 2026.03.17.
60) NVIDIA. Form 8-K. 2026.09.02.

속도를 팔수록 원가가 뛰는 곡선을 실리콘의 조합으로 완만하게 만들겠다는 구상이다. 토큰 공장들이 단가로 경쟁하는 동안 지표와 마켓과 서빙 표준은 엔비디아의 몫이 된다.

한편 층을 넓힐수록 위험 요소도 많아지는데, 가장 큰 고객인 하이퍼스케일러가 동시에 자체 칩으로 맞서는 경쟁자이고, 허깅 페이스 인수에는 오픈 모델 생태계의 중립성을 둘러싼 반독점 시선이 따르며, 렙톤의 화면에는 구매자가 고른 용량이 어느 공장의 어느 사이트에 있는지와 장애가 났을 때 누가 책임지는지가 드러나지 않는다. 그래서 구매자가 물을 것은 마켓플레이스에서 산 용량의 계약 상대와 장애 시 책임자가 누구인가, 그리고 이 산업에 엔비디아가 없는 층이 아직 남아 있는가다. 모든 층이 한 회사를 거친다면 벤더를 바꿔도 단가의 바닥과 소프트웨어 스택은 그대로여서, 이전 보고서가 말한 갈아탈 수 있는 상태가 토큰 공장이 아니라 칩 회사 앞에서 막히기 때문이다. 그 바깥을 찾는 시도가 다음 두 절이다.

특화 칩 진영: 엔비디아의 줄에 서지 않는 칩 회사들

이 진영이 보는 세상에서 GPU는 토큰을 만드는 값비싼 길이다. 2026년 5월 나스닥에 상장한 세레브라스는 2분기 매출의 약 70%를 하드웨어가 아닌 클라우드와 서비스에서 올려 칩 회사보다 토큰 공장에 가까워졌다.[61] 반대편 끝의 그록은 엔비디아에 추론 기술을 라이선스한 뒤 독립 회사로 남아, 2026년 6월과 8월 잇달아 자금을 조달하면서 네오클라우드로 방향을 바꿨다. 다만 새로 늘리는 용량은 자기 LPU가 아닌 엔비디아 시스템이고, 기업가치는 1년 사이 절반으로 내려왔다.[62][63] 새로 등장한 회사 가운데 모델 하나를 칩에 고정해 매우 빠른 칩을 만드는 탈라스(Taalas)는 2026년 8월 AMD에 인수됐다.[64] GPU용 고대역폭 메모리 대신 범용 메모리를 쓰는 디매트릭스(d-Matrix)는 6월 양산에 들어간 뒤 8월에 추론 배포 소프트웨어 회사를 인수했다.[65]

자기 칩을 고수하는 회사들에는 공통점들이 있는데, 엔비디아 GPU 공급을 제한하는 두 병목, 곧 고대역폭 메모리와 그 패키징 공정을 피해 가고, 세레브라스처럼 칩이 팔리지 않는 시기에도 토큰을 팔 수 있도록 클라우드를 직접 여는 길을 둔다. 그리고 GPU를 밀어내기보다 그 옆에 놓이는 쪽을

61) Cerebras. Form 10-Q, 2026년 2분기.
62) TechCrunch. "AI chipmaker Groq confirms $650M raise, re-staffs after Nvidia's $20B not-acqui-hire deal." 2026.06.22
63) TechCrunch. "Groq raises $350M to fuel its pivot from AI chips to neocloud." 2026.08.17.
64) The Register. "AMD acquires AI chip startup Taalas to boost inference performance by etching models into silicon." 2026.08.06.
65) d-Matrix. "d-Matrix Corsair AI Inference Platform Enters Full Production to Meet Customer Demand." 2026.06.09.

택하는데, AMD와 세레브라스도 2026년 7월 GPU와 웨이퍼 칩이 역할을 나누는 추론 플랫폼을 함께 내놓았다[66]. 그 결과 이들의 손익계산서는 네오클라우드를 닮아 가는데, 세레브라스가 오픈AI와 맺은 750메가와트 계약에서 인식한 매출만 분기 매출의 약 32%다. 위험 요소도 네오클라우드를 닮아서, 세레브라스는 매출의 66%를 두 고객에 기대고 칩은 TSMC 한 곳에서만 나오며, 탈라스처럼 모델을 실리콘에 고정하면 모델이 바뀔 때 칩을 다시 만들어야 한다.

이 진영 앞에서 구매자의 질문은 단순해진다. 초당 생성 속도가 진짜 요구 사항인 워크로드가 우리에게 몇 개이고 어떤 수요를 만들어 내는가. 그 답에 따라 이 진영은 계약할 상대가 되기도 하고 지켜볼 대상이 되기도 한다.

하이퍼스케일러: 자체 칩을 만드는 가장 큰 고객

하이퍼스케일러가 보는 세상에서 원가의 큰 몫은 엔비디아의 가격표이고, 세 곳의 자체 칩은 모두 세대를 하나씩 올렸다. 구글의 7세대 TPU 아이언우드(Ironwood)는 2026년 4월[67], AWS의 트레이니엄3(Trainium3)는 2025년 12월에 정식 출시됐고[68], 마이크로소프트의 추론 전용 칩 마이아 200(Maia 200)은 2026년 1월 공개되었다.[69] 세 칩이 겨냥하는 것은 외부 판매보다 자기 공장의 와트당 토큰과 감가상각 곡선을 엔비디아의 가격표에서 떼어내는 데 있는데, 마이크로소프트는 마이아 200의 달러당 성능이 자사 보유 장비의 직전 세대보다 30% 낫다고 밝혔다.

셋의 개방 정도는 달라서셋의 개방 정도는 달라서, 마이아는 내부 전용이고 트레이니엄은 AWS 안에서 누구나 빌릴 수 있으며, 구글만 자체 칩을 자기 클라우드 밖으로 내보내는 단계로 갔다. 앤트로픽은 2026년 4월 확장 계약으로 기가와트 단위의 TPU 용량을 확보했고[70], 메타는 임대에 이어 칩을 직접 사들이는 협의를 진행하는 것으로 보도됐다.[71] 그러면서도 하이퍼스케일러들은 부족한 용량을 네오클라우드에서 통째로 빌리고 있으니, 이들은 토큰 공장의 고객이면서 경쟁자다.

구매자 쪽에서 볼 지점은 두 가지인데, 첫째, TPU와 트레이니엄은 특히 학습과 맞춤 최적화에서 각자의 소프트웨어 스택을 요구해서, 그 위에서 최적화한 워크로드는 GPU보다 옮기기 어렵고 구매자는 그만큼

66) The Register. "AMD and Cerebras join forces against Nvidia's Groq LPUs." 2026.07.23.
67) The Next Web. "Google's Ironwood TPU goes GA for inference at Cloud Next." 2026.04.22.
68) AWS. "Amazon EC2 Trn3 UltraServers." 2025.12.02.
69) Microsoft(Scott Guthrie). "Maia 200: The AI accelerator built for inference." 2026.01.26.
70) Anthropic. "Expanding our partnership with Google and Broadcom." 2026.04.06.
71) SiliconANGLE. "Google and Meta reportedly strike new multibillion-dollar AI chip deal." 2026.02.26.

이식성의 비용을 진다. 둘째, 하이퍼스케일러가 자체 칩으로 얻은 원가 절감이 구매자의 토큰 단가에 반영된다는 약속은 어디에도 없다.

소버린: 국가 단위의 토큰 공장들

데이터와 연산 기반을 자국 안에 두려는 국가 단위의 시도를 소버린 AI라 부르는데, 토큰 공장을 짓는 움직임은 기업의 것만이 아니다. 유럽연합은 2026년 7월 AI 기가팩토리에 공공 자금 100억 유로와 민간 200억 유로 이상을 투입하는 공모를 열었고[72) 73)], 프랑스에서는 소프트뱅크가 AI 데이터센터 캠퍼스를 계획하였으며[74)] , 사우디아라비아 국부펀드 산하 휴메인(HUMAIN)은 AWS와 함께 리야드에 AI 존을 세우기로 했다.[75)] UAE의 스타게이트[76)]와, 데이터 처리를 일본 안에 두는 데 기여한다는 설명과 함께 발표된 마이크로소프트의 일본 투자[77)]도 같은 선상에 있다.

한국에서는 국가 AI 컴퓨팅센터가 2026년 8월 3일 전남 해남 솔라시도에서 착공됐는데, 사업비 약 2조 5,000억 원을 들여 2028년까지 첨단 AI 반도체 1만 5,000장을 갖추고 2030년까지 5만 장으로 늘리는 것이 목표다. 완공 전인 2027년에는 일부 자원을 민간 데이터센터에서 먼저 개방해 스타트업과 중소기업과 대학과 연구기관이 이용권과 요금 할인으로 쓰게 할 계획인데[78) 79)], 앞 장의 말로 하면 이 사업은 공공이 전력과 GPU를 먼저 확보하고 그 위의 가동률을 민간에 열어 채우는 구조다. 이 센터가 GPU 시간과 토큰 API 가운데 무엇을 팔지는 아직 공개되지 않았고, 그 결정이 민간 수요의 모양을 정할 것이므로 앞의 사례에서 언급한 기업 GPU의 낮은 가동률은 공공이 먼저 확보한 GPU에도 같은 경고가 된다.

소버린 AI는 세 개의 층을 국가 단위로 소유하려는 시도이지만 어느 사례에서도 칩과 파운드리, 그리고 하이퍼스케일러가 빠지지 않는다. 국가가 실제로 쥘 수 있는 것은 부지와 전력과 데이터의 소재지와 그 위의 운영 주체이고, 실리콘은 거의 전부 수입이며, 서빙 소프트웨어와 모델은 자국 비중을 늘릴 수 있는 층이다. 하이퍼스케일러의 국내 리전은 소재지 조건은 채워도 운영 주체 조건은 채우지 않으므로,

72) European Commission. "EU launches AI Gigafactories call to boost Europe's computing capacity and unlock more than EUR 30 billion." 2026.07.30.
73) EuroHPC Joint Undertaking. "EuroHPC Joint Undertaking launches AI Gigafactories call." 2026.07.30.
74) SoftBank Group. "SoftBank Group to invest in 5GW AI data center campuses in France." 2026.05.31.
75) Business Wire. "AWS and HUMAIN to deploy up to 150,000 AI accelerators in Saudi Arabia." 2025.11.23
76) The National. "Stargate UAE's first phase to be completed in third quarter of 2026." 2025.12.05.
77) Taipei Times. "Microsoft to invest US$10 billion in Japan AI infrastructure." 2026.04.04.
78) 바이라인네트워크. "국가 AI 컴퓨팅센터 해남 착공." 2026.08.03.
79) 머니투데이방송. "2조 4000억 투입, 해남 솔라시도에 '국가 AI컴퓨팅센터' 세운다." 2026.05.13.

어느 규제 요건을 충족하려는지에 따라 소버린인지 아닌지가 갈린다. 그리고 유럽이나 한국처럼 전력 단가가 높고 사이트 규모가 작은 곳에서는 기가와트당 원가가 글로벌 공장보다 높아지므로, 구매자에게 소버린은 프리미엄이 붙은 상품이 된다.

맺으며: CIO와 CTO를 위한 제언

이전 보고서가 구매 역량을 갖추라고 했다면, 이번에는 그 역량이 살펴야 할 범위를 벤더의 재무제표까지 넓히기를 권한다. 토큰 공장의 사정은 용량 배정의 순서, 단가와 계약 기간, 벤더의 재무 위험, 데이터 소재지의 프리미엄이라는 네 경로로 구매자의 계약서에 들어오고, 제언도 그 순서를 따른다. 토큰을 API로 사는 조직에는 셋째와 넷째가, GPU 시간이나 전용 용량을 계약하는 조직에는 첫째와 둘째가 먼저 해당한다.

1. 용량은 사이트와 인도 시점으로 확인하라

토큰 공장이 매진이면 신규 고객은 배정의 뒤로 밀리는데, 하이퍼스케일러가 네오클라우드의 용량을 통째로 계약하는 시장에서는 하이퍼스케일러를 경유하는 편이 오히려 빠를 수 있다. 그 대신 용량이 어느 사이트에 있는지는 알기 어려워진다. 전용 용량이나 약정 처리량을 사는 조직은 아직 지어지지 않은 용량을 계약하는 셈이므로, 첫 번째 행동은 벤더에게서 계약 용량의 사이트, 전력 인입 일정, GPU 세대를 서면으로 받는 일이다. 공장은 미래의 용량을 파는 곳이어서, 계약서의 용량이 아직 확보되지 않았을 가능성을 늘 고려해야 한다.

2. 장기 고정을 받아들이려면 세대교체 조항을 확인하라

갈아탈 수 있는 상태를 유지하라는 이전 보고서의 제언은 계약이 길어지는 시장에서 엔드포인트 주소보다 계약 조항의 문제가 된다. 장기 고정은 낮은 단가를 주는 대신 GPU 수명의 위험을 토큰 공장 대신 지게 하고, 매진된 시장에서 맺는 장기 계약은 희소성의 웃돈까지 함께 고정한다. 그래서 장기 고정을 받아들인다면 호퍼에서 블랙웰로, 다시 루빈으로 넘어갈 때 같은 계약안에서 새 세대로 옮길 수 있는지와 그때의 단가를 조항으로 받아야 한다. 감가상각 기간에 따른 원가 차이와 루빈의 토큰당 원가가 10분의 1이라는 엔비디아의 주장이 그 협상의 근거다. 기존 계약에 신세대 배치 시 단가를 조정하는 조항이 있는지 확인하면서 계약 기간을 세대 주기에 맞추거나 연차별 단가 인하율을 미리 정해 두는 것으로 협상의 난이도를 예상하고 준비해야 한다.

3. 벤더의 상위 고객과 그 고객의 자체 칩 로드맵을 보라

상위 두 고객이 매출의 60% 이상인 토큰 공장은 앵커가 떠나면 남는 용량이 시장에 풀려 단가가 급락하거나, 재무가 흔들려 서비스 연속성이 위험해진다. 어느 쪽이 될지는 그 공장의 부채와 이자 비용, 그리고 앵커가 누구인가가 정하게 되기에 상장 벤더의 분기 보고서에서 고객 집중도와 부채 원금과 앵커의 정체를 읽어 계약 심사 항목에 넣어야 한다. 비상장 벤더라면 대형 계약 발표와 자금 조달 보도 같은 공개 신호를 읽고, 지배권 변경이나 서비스 중단, 파산 때 데이터와 워크로드를 옮길 수 있는 조항을 요구하기를 권한다.

4. 소버린 프리미엄은 데이터 소재지가 진짜 제약인 워크로드에만 지불하라

소버린은 프리미엄이 붙은 상품이고, 모든 워크로드에 그 값을 치르면 토큰 공장의 가동률을 채우는 데 기여할 뿐 조직의 비용은 늘어난다. 워크로드별로 데이터 소재지 제약의 근거가 법령인지, 계약인지, 관행인지를 적어 구분해야 하는데, 관행에 해당하는 워크로드가 많을수록 소버린 프리미엄 없이 둘 수 있는 몫이 크다는 뜻이다. 공공 부문이라면 그 구분이 국가 AI 컴퓨팅센터와 민간 클라우드 사이에서 워크로드를 나누는 기준이 된다.

이번 보고서는 짓기도 전에 매진되는 토큰 공장과 95%가 쓰이지 않는 GPU라는 두 숫자에서 출발했다. 둘이 한 해에 함께 나온다는 것은 GPU가 어느 계약서에 묶이느냐에 따라 웃돈이 되기도 하고 감가상각만 남기기도 한다는 뜻이다. 추론 인프라의 두 편을 마치며 덧붙이면, 그 계약서의 반대편에 놓인 토큰 공장의 손익계산서를 읽는 일은 이제 토큰을 사는 쪽의 일이 되어 버렸다.