Ver. 2026

AX 전환을 지원하는 클라우드 환경 및 인프라 구축 동향



과학기술정보통신부
Ministry of Science and ICT

nipa 정보통신산업진흥원
National IT Industry Promotion Agency



AX 전환을 지원하는 클라우드 환경 및 인프라 구축 동향

CONTENTS

1. 서론 – AX 확산과 AI 클라우드·인프라의 중요성 01
가. 단순 업무 자동화에서 기업 전반의 AX로 01
나. AI 도입·DX·AX 단계에 따른 컴퓨팅·데이터·클라우드 환경의 변화 01
다. 기업의 주요 인프라 구축 과제 02
라. AI 모델 개발 중심에서 AI 서비스 운영·추론 중심으로 확대되는 인프라 수요 02

2. 글로벌 AI 클라우드 및 인프라 구축 동향 – AI 중심으로 재편되는 클라우드 시장 03
가. 생성형 AI 확산에 따른 글로벌 클라우드 투자 확대 03
나. Hyperscaler의 GPU Cloud·AI 플랫폼 경쟁과 자체 AI 가속기 확대 03
다. 데이터센터에서 AI Factory·AI Gigafactory로의 발전 03
라. GPU·고속 네트워크·스토리지·액체냉각 등 AI 데이터센터 인프라 고도화 04
마. AI 서비스 확산에 따른 추론 인프라 및 Agentic AI 기반 상시 추론 수요 확대 04
바. AI 데이터센터 TCO 프레임워크와 라이프사이클 관리 05

3. AX 전환을 위한 클라우드 아키텍처 및 AI 플랫폼 동향 – AI-Native Cloud로의 진화 07
가. Cloud 중심 환경에서 AI-Native Cloud로의 발전 07
나. MLOps·LLMOps, RAG, Vector DB, AI Agent 등 AI 플랫폼의 통합·서비스화 07
다. Hybrid Cloud·Multi-Cloud 환경 및 클라우드 상호운용성 강화 08
라. AI 데이터 주권과 보안·규제 대응을 위한 Sovereign Cloud 확산 09
마. 글로벌 산학협력이 제시하는 차세대 하이브리드 클라우드 비전 09

4. 국내 AI 클라우드 및 인프라 구축 동향 – 국가 AI 컴퓨팅과 민간 AI Cloud의 확대 11
가. 국가 AI 컴퓨팅센터·AI 고속도로 등 국가 차원의 인프라 확충 11
나. 국내 CSP의 GPU Cloud·AIaaS 확대 및 AI 클라우드 시장 성장 11
다. AI 데이터센터·AI Factory 구축 확대와 소형 데이터센터 등 지역 AI 인프라 확산 11
라. 국산 AI 반도체 및 K-Cloud 생태계 확대 12
마. AI 데이터센터의 전력·냉각·에너지 효율 등 지속가능한 인프라 구축 12



5. 기업의 AX 전환을 위한 클라우드·인프라 구축 전략 – AI 성숙도에 따른 활용 전략 13
가. PoC·실증·서비스 운영 등 AI 성숙도에 따른 클라우드·인프라 선택 전략 13
나. 자체 AI 인프라 구축과 클라우드 이용의 경제성 및 TCO 비교 13
다. AI 인프라 투자와 AX 성과를 연계하기 위한 AI FinOps 및 Business KPI 관리 14
라. 보안·신뢰·거버넌스를 고려한 Hybrid/Multi-Cloud 기반 AX 환경 구축 15
마. 중소·중견기업의 AI Cloud 활용 및 단계적 AX 전환 전략 15
바. 하드웨어 리프레시 전략과 운영 단계 소프트웨어 최적화: 학술 연구가 제시하는 시사점 15
사. 글로벌 리서치가 제시하는 하이브리드 AI 인프라 투자 전략 16
아. 보안·거버넌스의 내재화 17

6. AX 원스톱 바우처와 AI 클라우드·인프라의 연계 및 시사점 – AI·클라우드·데이터의 통합 지원 19
가. AX 전환 과정에서 기업이 겪는 주요 애로사항 19
나. AI 솔루션·GPU·SaaS 등을 통합 지원하는 AX 원스톱 바우처의 특징 19
다. 단순 GPU 지원 또는 AI 솔루션 구매 지원과 차별화되는 원스톱 지원체계의 의미 19
라. 수요기업과 AI·클라우드·데이터 공급기업 연계를 통한 서비스 개발 및 사업화 지원 20

7. 결론 – AX 시대의 클라우드·인프라 발전 방향 및 정책적 시사점 21
가. GPU 중심 인프라에서 AI Computing·AI Factory 등 통합 AI 인프라로의 발전 21
나. AI 모델 학습 중심에서 학습·추론 및 AI 서비스 운영 중심으로의 변화 21
다. 기업의 AX 경쟁력 강화를 위한 클라우드·AI 플랫폼·데이터의 통합 필요성 21
라. 국내 AI 클라우드·컴퓨팅 생태계 강화를 위한 정책적 시사점 22
마. AX 원스톱 바우처의 AI 공통기반 지원 및 기업 AX 확산을 위한 향후 발전 방향 22
바. 국내 AX 확산을 위한 정책 제언 22
사. 기업 AX 전환을 위한 3개년 인프라 구축 로드맵 23

[참고문헌] 24

[작성자] 25

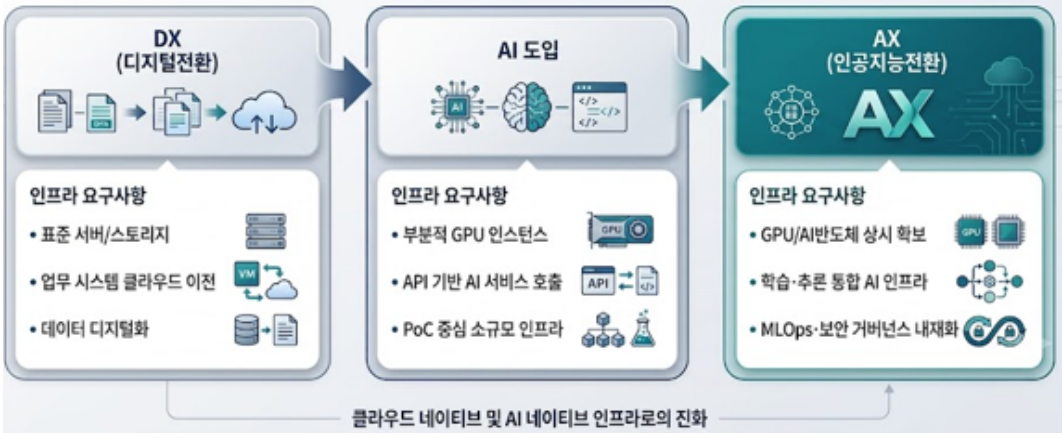
1 서론 – AX 확산과 AI 클라우드·인프라의 중요성

가. 단순 업무 자동화에서 기업 전반의 AX로

생성형 AI의 급격한 확산 이후 기업의 AI 활용은 특정 업무를 보조하는 단순 자동화 도구의 수준을 넘어, 일하는 방식과 의사결정 구조, 나아가 비즈니스 모델 자체를 재설계하는 인공지능전환(AX, AI Transformation)으로 확대되고 있다. AI 도입·디지털전환(DX)·AX는 모두 기업 혁신과 맞닿아 있지만, 각 단계가 요구하는 컴퓨팅·데이터·클라우드 환경의 성격은 근본적으로 다르다. 단순 AI 도입 단계는 개별 업무용 API 호출 수준의 자원으로 충분하지만, DX는 조직의 정보시스템과 데이터 인프라 전반을 클라우드 기반으로 정비하는 과정을 요구하며, AX는 이렇게 구축된 데이터·시스템 위에서 AI를 업무 프로세스와 의사결정 구조에 통합할 수 있는 상시적·통합적 AI 공통기반(Common AI Infrastructure)을 전제로 한다.

나. AI 도입·DX·AX 단계에 따른 컴퓨팅·데이터·클라우드 환경의 변화

세 단계의 차이는 곧 인프라 요구사항의 차이로 이어진다. DX 단계에서는 표준화된 서버·스토리지 환경으로의 이전과 업무 데이터의 디지털화가 핵심 과제였다면, AI 도입 단계에서는 부분적인 GPU 인스턴스 활용과 API 기반 AI 서비스 호출, PoC 중심의 소규모 인프라가 중심이 된다. 반면 AX 단계에서는 GPU·AI반도체의 상시적 확보, 학습과 추론을 아우르는 통합 AI 인프라, 그리고 MLOps·보안 거버넌스가 내재화된 운영 체계가 요구된다. [그림 1]은 이러한 단계별 인프라 요구사항의 변화를 정리한 것이다.



[그림 1] DX → AI 도입 → AX 전환에 따른 클라우드·인프라 요구사항 변화

다. 기업의 주요 인프라 구축 과제

AX를 추진하는 기업이 직면하는 인프라 과제는 크게 네 가지로 요약된다. 첫째, GPU 확보와 비용 부담이다. 최신 GPU는 공급이 제한적이고 단가가 높아, 중소·중견기업이 독자적으로 충분한 물량을 확보하기 어렵다. 실제로 NVIDIA DGX H100 서버 1대의 가격은 20만 달러를 상회하며 최대 10.2kW의 전력을 소비하는 등[29], 최신 AI 가속기의 자본·운영비용은 기존 범용 서버를 크게 웃돈다. 둘째, 데이터센터의 전력·냉각 문제다. 고밀도 GPU 랙은 기존 공랭식 냉각의 한계를 넘어서는 발열을 발생시키며, 전력 확보는 다수의 국가에서 이미 AI 인프라 확장의 최대 병목으로 지목되고 있다[1][29]. 셋째, 레거시 시스템과의 연계다. 다수의 기업이 기존 온프레미스 업무 시스템과 신규 AI 인프라를 병행 운영해야 하는 상황에서, 데이터 연계·API 통합의 복잡성이 AX 추진 속도를 제약하는 요인으로 작용한다. 넷째, 보안·전문인력 부족이다. AI 인프라 운영에는 클라우드·MLOps·보안 전문성이 동시에 요구되나, 이러한 복합 역량을 갖춘 인력은 기업 내부에서 확보하기 쉽지 않다.

라. AI 모델 개발 중심에서 AI 서비스 운영·추론 중심으로 확대되는 인프라 수요

초기 생성형 AI 붐을 견인한 인프라 수요는 대규모 파운데이션 모델을 처음부터 학습(pre-training)시키는 데 집중되어 있었다. 그러나 현재 산업 전반의 무게중심은 이미 완성된 모델을 기업 환경에 맞게 미세조정(fine-tuning)하고 실시간으로 서비스를 제공하는 추론(inference) 및 에이전틱 AI(Agentic AI) 운영 단계로 이동하고 있다. 엔비디아는 AI 인프라의 경제성이 이제는 학습 성능이 아니라 초당 토큰 처리량과 토큰당 비용으로 평가되는 단계에 접어들었다고 진단한다[2]. 이는 기업이 AX 인프라를 설계하는 기준 역시 '얼마나 큰 모델을 학습시킬 수 있는가'에서 '업무 현장에서 얼마나 안정적이고 저렴하게 AI 서비스를 상시 제공할 수 있는가'로 옮겨가고 있음을 의미한다.

2 글로벌 AI 클라우드 및 인프라 구축 동향 – AI 중심으로 재편되는 클라우드 시장

가. 생성형 AI 확산에 따른 글로벌 클라우드 투자 확대

글로벌 하이퍼스케일러의 AI 인프라 투자는 2026년 들어 전례 없는 규모로 확대되고 있다. 아마존·마이크로소프트·구글·메타 등 주요 4개 사는 2026년 설비투자(CapEx)로 약 6,600억~7,250억 달러를 계획하고 있으며, 이는 2025년 약 4,100억 달러 대비 60~77% 증가한 수준이다[3][4]. 골드만삭스는 이들 4개사의 2025~2030년 누적 설비투자가 5조 달러를 상회할 것으로 전망하고 있어[4], AI 인프라 투자가 단발성 붐이 아니라 다년간 지속되는 구조적 사이클로 자리잡고 있음을 보여준다. 국내 증권업계 분석에 따르면 글로벌 AI 데이터센터의 총 전력 수요(용량 기준)는 2025년 약 82GW에서 향후 5년간 2.7배로 확대될 것으로 전망되며, 관련 시장 규모는 2030년까지 약 7조 달러 내외에 이를 것으로 추산된다[5].

나. Hyperscaler의 GPU Cloud·AI 플랫폼 경쟁과 자체 AI 가속기 확대

AWS·Azure·GCP 등 주요 하이퍼스케일러는 엔비디아 GPU 확보 경쟁과 병행하여 자체 AI 가속기 개발을 가속화하고 있다. 아마존의 트레이니엄(Trainium), 구글의 TPU, 마이크로소프트의 마이아(MAIA) 등 자체 실리콘은 엔비디아 GPU에 대한 의존도를 낮추는 동시에 자사 클라우드 플랫폼의 원가 경쟁력을 강화하는 수단으로 활용되고 있다[6]. 이와 함께 코어위브(CoreWeave)와 같이 GPU 인프라만을 전문적으로 제공하는 이른바 '네오클라우드(Neocloud)' 사업자가 새로운 시장 참여자로 부상하며, 하이퍼스케일러와 장기 계약을 맺고 안정적인 수익 기반을 확보하는 사례도 늘고 있다[5].

다. 데이터센터에서 AI Factory·AI Gigafactory로의 발전

AI 인프라 확대는 단순한 데이터센터 증설을 넘어, 학습·추론 파이프라인 전체를 하나의 생산 설비로 재구성하는 'AI 팩토리(AI Factory)' 개념으로 진화하고 있다. 엔비디아는 AI 팩토리를 데이터를 원료로 삼아 토큰(지능)을 생산하는 시설로 정의하며, 그 경제성을 초당 토큰 처리량과 토큰당 비용으로 측정한다고 설명한다[2]. 2026년 들어서는 개별 시설 단위를 넘어 기가와트(GW)급 연산 용량을 갖춘 'AI 기가팩토리' 구축 경쟁이 본격화되고 있다. 레노버는 2026년 1월 CES에서 엔비디아와 함께 기가와트급 'AI Cloud Gigafactory' 프로그램을 공개하였으며[7], 국내에서는 네이버가

엔비디아와 협력하여 자사 하이퍼스케일 데이터센터를 기반으로 기가와트급 AI 팩토리 구축 계획을 발표하는 등[8] 글로벌·국내 사업자를 막론하고 초대형 전용 인프라 투자가 확산되고 있다.

라. GPU·고속 네트워크·스토리지·액체냉각 등 AI 데이터센터 인프라 고도화

고밀도 GPU 클러스터의 확산은 네트워크·스토리지·냉각 전 영역의 기술 고도화를 동반한다. 대규모 분산 학습·추론에는 인피니밴드(InfiniBand) 등 초고속 네트워크와 NVMe 기반 고성능 스토리지가 필수적이며, 랙당 전력밀도가 높아지면서 액체냉각·침수냉각 방식이 기존 공랭식을 빠르게 대체하고 있다. 세계 최고 수준의 AI 시설은 데이터센터 전력사용효율지수(PUE) 1.1~1.2 수준을 달성하고 있는 것으로 알려져 있으며[9], 이는 냉각 방식의 전환이 곧 AI 인프라의 운영 경제성을 좌우하는 핵심 변수가 되고 있음을 보여준다.

마. AI 서비스 확산에 따른 추론 인프라 및 Agentic AI 기반 상시 추론 수요 확대

가장 주목할 만한 변화는 AI 인프라 수요의 축이 대규모 사전학습에서 추론·에이전틱 AI 운영으로 이동하고 있다는 점이다. 엔비디아 GTC 2026에서도 이러한 전환이 핵심 화두로 제시되었으며, 다수의 하이퍼스케일러가 인프라를 우선 구축한 뒤 성능을 사후 분석하는 방식에서, 설계 단계부터 워크로드 패턴을 모델링하는 방식으로 전환하고 있다는 점이 지적되었다[10]. 에이전틱 AI는 단발성 질의응답이 아니라 여러 단계의 추론과 도구 호출을 연쇄적으로 수행하므로, 상시적이고 예측 가능한 저지연 추론 인프라에 대한 요구가 더욱 커지고 있다. 〈표 1〉은 글로벌 AI 클라우드·인프라 동향을 항목별로 정리한 것이다.

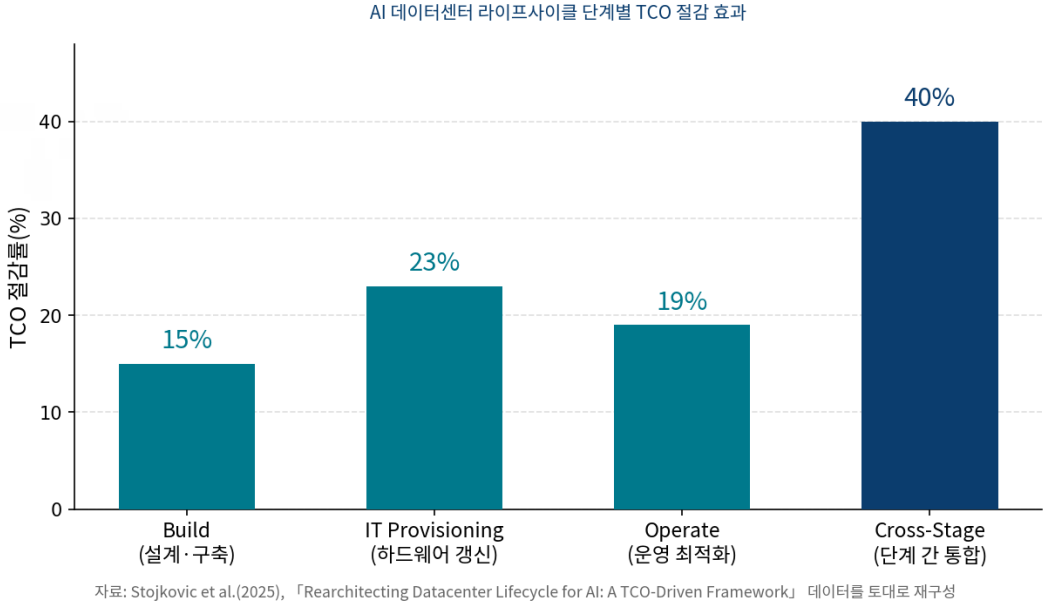
〈표 1〉 글로벌 AI 클라우드·인프라 동향

구분	주요 내용
투자 규모	4대 하이퍼스케일러 2026년 CapEx 약 6,600억~7,250억 달러(전년比 60~77% ↑), 2025~2030년 누적 5조 달러 이상 전망[3][4]
시장 전망	글로벌 AI 데이터센터 전력 수요 5년 내 2.7배 확대(2025년 82GW 기준), 관련 시장 2030년까지 약 7조 달러 내외[5]
가속기 경쟁	엔비디아 GPU 중심 경쟁과 함께 Trainium(AWS)·TPU(Google)·MAIA(MSFT) 등 자체 실리콘 확대, 네오클라우드(코어위브 등) 부상[5][6]
인프라 형태	데이터센터 → AI Factory → 기가와트급 AI Gigafactory로 진화(레노버·엔비디아, 네이버·엔비디아 등)[2][7][8]
수요 축 변화	대규모 사전학습 중심 → 추론·에이전틱 AI 상시 운영 중심으로 이동, 설계 단계부터 워크로드 모델링 필요성 부각[10]

바. AI 데이터센터 TCO 프레임워크와 라이프사이클 관리

최근 마이크로소프트 애저 리서치(Microsoft Azure Research)와 UIUC 공동 연구진은 AI 데이터센터의 총소유비용(TCO)을 데이터센터 건설(Build, 15~30년), 하드웨어 도입·갱신(IT Provisioning, 4~6년), 운영(Operate, 추론 단위) 세 단계로 나누어 분석하고, 각 단계를 개별 최적화할 때와 단계를 넘나들며 통합 최적화(Cross-Stage)할 때의 TCO 절감 효과를 정량적으로 비교하는 프레임워크를 제시하였다[29]. 이 연구는 AI 데이터센터의 TCO를 CapEx(IT·네트워킹·건물·전력·냉각 설비 투자)와 OpEx(에너지·유지보수·네트워크 운영·소프트웨어 라이선스)의 합으로 정의하며, 대표적으로 NVIDIA DGX H100 서버 1대의 가격이 20만 달러를 상회하고 최대 10.2kW의 전력을 소비하는 등, AI 가속기가 기존 범용 서버 대비 훨씬 높은 자본·운영비용을 유발한다는 점을 실증적으로 보여준다[29].

이 연구에 따르면 평균 가동률 75% 기준 10MW급 AI 데이터센터는 약 500대의 H100 서버를 운용하며 연간 약 70GWh의 에너지를 소비하는데, 이 경우 GPU 서버 관련 IT CapEx가 전체 비용에서 가장 큰 비중을 차지하고, 에너지 관련 OpEx가 그 뒤를 잇는 반면 건물·유지보수 비용의 비중은 상대적으로 작은 것으로 나타났다[29]. 이러한 비용 구조는 AI FinOps·TCO 분석의 실증적 근거가 되며, 국내 기업이 자체 인프라 구축과 클라우드 이용을 비교할 때에도 IT(GPU)·에너지 비용이 의사결정의 핵심 변수가 되어야 함을 시사한다.

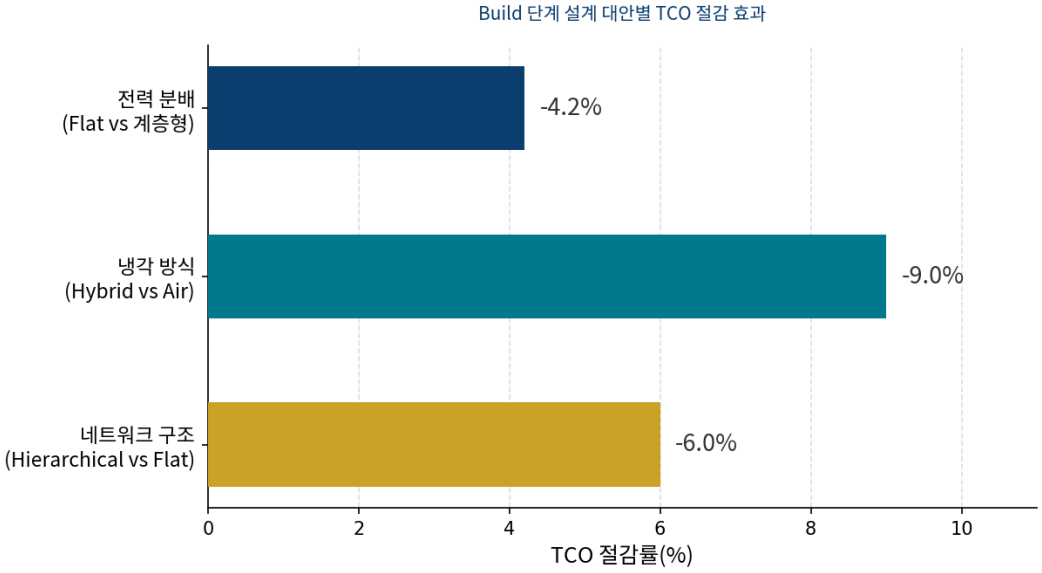


[그림 2] AI 데이터센터 라이프사이클 단계별 TCO 절감 효과

(출처: Stojkovic et al., 2025[29])

시뮬레이션 결과, 건설(Build) 단계의 설계 최적화만으로 TCO를 15% 절감할 수 있었고, 하드웨어 갱신(IT Provisioning) 전략 최적화로 23%, 운영(Operate) 단계의 소프트웨어 최적화로 19%를 각각 절감할 수 있었다. 그러나 세 단계를 개별적으로 최적화하는 데 그치지 않고 건설-갱신-운영 결정을 상호 연계하여 통합적으로 최적화(Cross-Stage)할 경우, TCO를 최대 40%까지 절감할 수 있는 것으로 나타났다[29]. 이는 'AI FinOps 및 Business KPI 관리'가 단일 시점의 비용 절감이 아니라, 인프라 설계·조달·운영 전 과정을 아우르는 라이프사이클 관점에서 이루어져야 함을 뒷받침하는 실증적 근거라 할 수 있다.

건설 단계의 구체적 설계 대안을 살펴보면, 전력 분배는 PDU 단위·UPS 단위·데이터센터 전체 단위로 전력을 공유하는 범위를 넓힐수록(Flat 구조) 유휴 전력(stranded power)이 줄어들어 계층형 구조 대비 TCO를 4.2% 절감할 수 있었다. 냉각 방식은 GPU 랙에는 액체냉각(Cold Plate)을, 저밀도 랙에는 공랭을 혼합하는 하이브리드 냉각(냉각판 75%+공랭 25% 배합 시 PUE 1.15 달성)이 전체 공랭 대비 9%의 TCO 절감 효과를 보였다. 네트워크는 서버 내부는 NVLink, 랙 내부는 InfiniBand, 랙 간에는 Ethernet을 적용하는 계층형(Hierarchical) 구조가 단일 고성능 네트워크 대비 6%의 TCO를 절감하는 것으로 나타났다[29].



자료: Stojkovic et al.(2025), 「Rearchitecting Datacenter Lifecycle for AI: A TCO-Driven Framework」 데이터를 토대로 재구성

[그림 2] Build 단계 설계 대안별 TCO 절감 효과

(출처: Stojkovic et al., 2025[29])

(※ [그림 2]는 하이퍼스케일러 및 국가 AI 컴퓨팅센터 수준의 대규모 인프라 설계에 우선 적용되는 내용이나, 그 핵심 원리(전력 분배 범위 확대, 냉각 방식의 워크로드별 최적 배합, 계층형 네트워크 설계)는 국내 CSP·K-클라우드 인프라 설계에도 참고할 만한 시사점을 제공한다.)

3 AX 전환을 위한 클라우드 아키텍처 및 AI 플랫폼 동향 - AI-Native Cloud로의 진화

가. Cloud 중심 환경에서 AI-Native Cloud로의 발전

초기의 AI 클라우드는 GPU 인스턴스를 임대해주는 인프라 서비스(IaaS) 수준에 머물러 있었다. 그러나 최근의 AI 클라우드는 모델 개발·학습·추론·운영 전 주기를 하나의 플랫폼 위에서 지원하는 'AI-Native Cloud'로 진화하고 있다. 시장조사기관들은 AI로 인한 클라우드 인프라 서비스 수요 급증이 GPU·전용 가속기뿐 아니라 고대역폭 네트워크, 고속 스토리지, 확장 가능한 데이터 파이프라인, 안정적인 전력·냉각 접근성까지 포괄하는 통합적 요구로 확장되고 있다고 분석한다[11]. 이에 따라 클라우드 사업자들은 단순 연산 자원 제공을 넘어, 데이터 아키텍처·모델 배포 전략·워크로드 배치까지 아우르는 통합 AI 플랫폼으로 서비스 범위를 넓히고 있다.

나. MLOps·LLMOps, RAG, Vector DB, AI Agent 등 AI 플랫폼의 통합·서비스화

AI-Native Cloud의 핵심 경쟁력은 모델 운영을 자동화·표준화하는 MLOps·LLMOps 체계, 외부 지식을 실시간으로 결합해 환각(Hallucination)을 줄이는 검색증강생성(RAG) 아키텍처, 이를 뒷받침하는 벡터 데이터베이스(Vector DB), 그리고 복잡한 업무를 자율적으로 수행하는 AI 에이전트 프레임워크가 하나의 관리형 서비스로 통합되어 제공되는 데 있다. 실제로 국내 공공부문에서도 대규모 언어모델과 RAG를 결합해 공공 문서 분석·정책 질의응답·보고서 초안 작성 등 복잡한 지식 업무 전반에 지능형 모델을 내재화하는 사례가 확산되고 있으며, 이 과정에서 경량화된 중소형 언어모델(sLLM)과 RAG API를 플랫폼 형태로 제공해 여러 기관이 공유하는 구조가 표준적인 접근으로 자리잡고 있다[12]. [그림 3]은 AI Cloud·AI 플랫폼의 기술 및 서비스 구조를 계층별로 정리한 것이다.



[그림 3] AI Cloud·AI 플랫폼의 기술 및 서비스 구조

다. Hybrid Cloud·Multi-Cloud 환경 및 클라우드 상호운용성 강화

기업의 AI 활용이 확대될수록 특정 클라우드 사업자에 대한 종속 위험을 낮추기 위해 여러 클라우드를 병행 사용하는 Hybrid·Multi-Cloud 전략의 중요성이 커지고 있다. 이는 단순히 복수의 클라우드 계약을 유지하는 것을 넘어, 워크로드를 성능·비용·컴플라이언스 요건에 따라 적절한 환경에 배치하고 클라우드 간 데이터·모델 이동성을 확보하는 상호운용성 문제로 이어진다. 벤더 종속을 피하고 데이터 상주 요건을 충족하려는 기업 수요가 하이브리드·멀티클라우드 운영 모델을 사실상의 표준으로 만들고 있다[11].

라. AI 데이터 주권과 보안·규제 대응을 위한 Sovereign Cloud 확산

AI 활용이 국가 기밀·민감정보 영역까지 확대되면서, 국가·기관이 자국 데이터·인프라·인력에 대한 통제권을 유지하는 '소버린 AI(Sovereign AI)' 개념과 이를 뒷받침하는 Sovereign Cloud가 함께 부상하고 있다. 영국의 공무원 전용 생성형 AI 도구 'Redbox', 싱가포르의 이중 보안 샌드박스 'Pair'는 보안이 담보된 클라우드 아키텍처가 마련된다면 공공·규제 영역에서도 AI 활용 가능성을 충분히 높일 수 있음을 실증한 사례로 평가된다[13]. 국내에서도 공공데이터에 대한 운영주권·데이터주권·소프트웨어주권을 아우르는 '공공데이터 주권 클라우드' 개념이 제시되는 등[14], 데이터 주권은 AX 인프라 전략에서 비용·성능과 함께 반드시 고려해야 할 핵심 축으로 부상하고 있다.

마. 글로벌 산학협력이 제시하는 차세대 하이브리드 클라우드 비전

IBM과 UIUC(University of Illinois Urbana-Champaign)가 공동 설립한 IBM-Illinois Discovery Accelerator Institute(IIDAI)는 2024년 11월 발표한 백서를 통해 향후 5~10년간 AI 워크로드를 뒷받침할 하이브리드 클라우드의 장기 비전을 제시하였다[30]. 이 백서는 미래의 하이브리드 클라우드가 갖추어야 할 조건으로 손쉬운 사용성(consumability)·경제성(affordability)·유비쿼터스 접근성(ubiquity) 세 가지를 제시하며, 이를 통해 오늘날 대비 100~1,000배의 성능/전력(performance/watt) 개선을 달성하는 것을 장기 목표로 제시한다는 점에서, 앞서 살펴본 국내 기업의 GPU 확보·전력 비용 부담 문제와 맥을 같이 한다[30].

핵심 제안 중 하나는 'LLM as an Abstraction(LLMaaA)'으로, 대규모 언어모델을 클라우드 애플리케이션의 구축·배포·운영을 위한 새로운 추상화 계층으로 활용하는 개념이다[30]. LLMaaA는 사용자 요청을 해석하는 인터페이스 에이전트(Interface Agent), 여러 LLM·비LLM 에이전트(시뮬레이션 에이전트, 최적화 솔버 등)를 선택·조율하는 마스터 에이전트(Master Agent), 플러그앤플레이 방식으로 구성되는 이기종 에이전트 풀(agent pool)로 이루어지며, 쿠버네티스(K8s) 기반 파드(Pod) 위에서 요청량에 따라 자동 확장되는 구조로 구현될 수 있다고 제시한다[30]. 이는 AX 원스톱 바우처가 강조하는 AI 에이전트 기반 업무 자동화·서비스 개발 지원 방향과 부합하는 해외 사례로 볼 수 있다.

이와 함께 IIDAI는 단기·장기 기억, 도구 활용, 계획·성찰(planning/reflection) 기능을 갖춘 에이전틱 AI 연구 프레임워크인 THINKagents를 제안하며, 소프트웨어 개발·정책 시뮬레이션·규제 준수(compliance) 등 다양한 영역에서 다중 에이전트 협업의 신뢰성을 높이는 것을 핵심 연구과제로 제시한다[30].

인프라·하드웨어 측면에서는 CPU·GPU·TPU·AI가속기·양자컴퓨터 등 이기종 자원을 하나의 소프트웨어 정의 인터페이스로 통합하는 '통합 제어 평면(Unified Control Plane)'과 멀티클라우드 브로커(multi-cloud broker) 개념을 제시하며, Kubernetes·LSF/Slurm 등 서로 다른 자원관리자를 넘나드는 워크로드 배치를 AI가 자동으로 수행하는 구조를 지향한다[30]. 아울러 CXL(Compute Express Link)·UALink(Ultra Accelerator Link) 등 캐시 일관성 인터커넥트(cache-coherent interconnect)를 통해 호스트-디바이스 간 메모리 접근을 단순화함으로써, LLM 추론에 필요한 대용량 메모리를 저비용으로 확장할 수 있다고 분석한다[30].

AX 전환을 지원하는 클라우드 환경 및 인프라 구축 동향

에너지 측면에서는 데이터센터가 이미 미국 전체 전력의 2% 이상을 소비하고 있으며, GPT-3급 모델 1회 사전학습에 자동차 10대의 평생 탄소배출량에 해당하는 탄소가 발생한다는 분석을 인용하면서, 개별 서버 단위의 DVFS(동적 전압·주파수 조정) 기반 학습 페이스 제어부터 재생에너지 연계 모듈형 데이터센터(Modular Data Center)에 이르기까지 다층적인 에너지 최적화가 필요하다고 강조한다[30]. 이는 국내외 데이터센터의 전력·냉각 병목 문제에 대해, 액체냉각 도입과 같은 설비 차원의 대응을 넘어 소프트웨어 계층의 에너지 인지 스케줄링까지 포괄하는 통합적 접근이 필요함을 뒷받침한다.

4 국내 AI 클라우드 및 인프라 구축 동향 – 국가 AI 컴퓨팅과 민간 AI Cloud의 확대

가. 국가 AI 컴퓨팅센터·AI 고속도로 등 국가 차원의 인프라 확충

정부는 「대한민국 인공지능 기본계획(2026~2028)」을 통해 데이터·컴퓨팅·반도체·전력 등 AI 개발·활용의 기초가 되는 국가 인프라를 'AI 고속도로'로 명명하고, 이를 5대 전략분야 중 첫 번째 과제로 제시하였다[15]. 그 핵심 축인 국가 AI 컴퓨팅센터는 전남 해남 솔라시도 데이터센터파크를 부지로 2028년 개소, 2030년까지 GPU 약 5만 장 확보를 목표로 하며, 2026년 4월 특수목적법인 설립을 거쳐 같은 해 3분기 착공에 들어갔다. 총 사업비는 2조 5,000억 원 이상 규모로, 사업 추진 과정에서 정부 지분을 낮추고 민간 지분을 확대하는 방향으로 구조가 재편되어 민간 주도의 투자 유인을 강화한 것이 특징이다[16][17]. 이와 병행하여 과학기술정보통신부는 2026년 한 해에만 약 2조 805억 원 규모의 'AI컴퓨팅자원 활용기반 강화사업'을 통해 최신 GPU 약 1만 5천 장을 추가 확보하고, 민간 클라우드 기업이 이를 2027~2031년 GPUaaS 형태로 산·학·연에 제공하도록 지원하고 있다[17].

나. 국내 CSP의 GPU Cloud·AIaaS 확대 및 AI 클라우드 시장 성장

네이버클라우드·KT클라우드·NHN클라우드 등 국내 클라우드서비스제공자(CSP)는 자체 데이터센터를 기반으로 GPU Cloud·AIaaS(AI as a Service) 사업을 확대하고 있다. 네이버는 엔비디아와 협력해 기가와트급 AI 팩토리 구축에 나서며 글로벌 AI 인프라 사업자로의 전환을 도모하고 있으며[8], 정부의 AX 원스톱 바우처 사업에도 다수의 국내 클라우드 기업이 GPU 인프라·SaaS 전환 지원 공급기업으로 참여하고 있다.

다. AI 데이터센터·AI Factory 구축 확대와 소형 데이터센터 등 지역 AI 인프라 확산

대규모 하이퍼스케일 데이터센터뿐 아니라, 특정 지역·산업 수요에 대응하는 소형 데이터센터를 통한 분산형 AI 인프라 구축도 병행되고 있다. 국가 AI 컴퓨팅센터와 같은 초대형 인프라가 국가 전체의 연산 기반을 담당한다면, 지역 기반 소형 데이터센터는 지역 산업·공공기관의 AI 활용을 뒷받침하는 보완적 역할을 수행하며, 두 축이 상호보완적으로 국내 AI 컴퓨팅 생태계를 형성해가는 양상을 보이고 있다.

라. 국산 AI 반도체 및 K-Cloud 생태계 확대

국산 AI 반도체(NPU) 생태계 육성을 위한 'K-클라우드 프로젝트'는 2023년부터 2030년까지 총 8천억 원 이상이 투입되는 사업으로, 데이터센터 인프라·HW, 컴퓨팅 SW, 클라우드 등 3개 전략분야로 구성되어 있다[18]. 리벨리온·퓨리오사AI·사피온코리아·딥엑스 등 국내 팹리스 기업은 이 사업을 통해 국산 신경망처리장치(NPU) 양산에 성공하였으며, 리벨리온과 퓨리오사AI는 각각 6,400억 원, 8,000억 원 규모의 투자를 유치하는 등 국산 AI 반도체가 상용화 초기 단계를 넘어서고 있다[19]. K-클라우드 프로젝트 1단계에서는 네이버클라우드·NHN클라우드·KT클라우드 등 국내 클라우드 기업과 리벨리온·사피온코리아·퓨리오사AI 등 반도체 기업이 참여해 국산 NPU를 데이터센터에 적용한 클라우드 AI 서비스를 민간·공공 부문에 각각 구축하는 사업을 진행하고 있다[20]. 다만 국가 AI 컴퓨팅센터 사업 과정에서 국산 NPU 도입 의무 조항이 삭제되는 등[16], 국산 반도체 생태계 육성과 사업 추진 속도·민간 투자 유인 사이의 정책적 균형점을 모색하는 과정이 지속되고 있다.

마. AI 데이터센터의 전력·냉각·에너지 효율 등 지속가능한 인프라 구축

AI 인프라 확장의 최대 제약 요인으로 전력 공급이 지목되면서, 데이터센터 전력사용효율지수(PUE)를 낮추기 위한 액체냉각·침수냉각 등 고밀도 냉각 기술 도입이 국내외에서 공통적으로 추진되고 있다. 국내 역시 신규 AI 데이터센터 구축 시 전력·냉각 설비를 사업 초기 단계부터 함께 설계하는 것이 필수 요건으로 자리잡고 있으며, 국가 AI 컴퓨팅센터를 비롯한 대형 인프라 사업 모두 전력 확보와 인허가 절차가 사업 일정에 직접적인 영향을 미치는 변수로 작용하고 있다. 〈표 2〉는 국내 AI 클라우드·컴퓨팅 인프라의 주요 동향을 정리한 것이다.

〈표 2〉국내 AI 클라우드·컴퓨팅 인프라의 주요 동향

구분	주요 내용
국가 인프라	국가 AI 컴퓨팅센터(해남 솔라시도, 2028년 개소·2030년 GPU 5만장, 총사업비 2.5조원 이상), AI 고속도로(대한민국 인공지능 기본계획 5대 전략분야)[15][16][17]
GPU 확보	AI컴퓨팅자원 활용기반 강화사업(2026년 2조805억원, GPU 1.5만장 추가 확보 후 2027~2031년 GPUaaS 제공)[17]
민간 CSP	네이버클라우드·KT클라우드·NHN클라우드 등 GPU Cloud·AIaaS 확대, 네이버-엔비디아 기가와트급 AI 팩토리 추진[8]
국산 반도체	K-클라우드 프로젝트(2023~2030, 8천억원+), 리벨리온·퓨리오사AI·사피온코리아·딥엑스 NPU 양산 및 대규모 투자 유치[18][19][20]
지속가능성	액체냉각·침수냉각 도입 확산, 전력·인허가가 대형 인프라 사업 일정의 핵심 변수로 부상

5 기업의 AX 전환을 위한 클라우드·인프라 구축 전략 – AI 성숙도에 따른 활용 전략

가. PoC·실증·서비스 운영 등 AI 성숙도에 따른 클라우드·인프라 선택 전략

기업의 AI 인프라 활용 전략은 초기 개념검증(PoC) 단계부터 실제 업무 확산 단계까지 성숙도에 따라 단계적으로 고도화되어야 한다. PoC·실증 단계에서는 초기 투자 없이 즉시 이용 가능한 퍼블릭 클라우드의 GPU 인스턴스를 종량제(온디맨드) 방식으로 활용하는 것이 합리적이며, 실제 업무 반영 및 서비스 운영 단계로 넘어가면서는 예약 인스턴스나 자체 인프라 확보를 통해 지속적인 사용에 따른 단가를 낮추는 방향으로 전환하는 것이 일반적인 패턴이다. 가변적인 워크로드나 인프라 운영 전문성이 부족한 조직은 클라우드 기반 활용이, 예측 가능하고 지속적인 고가동률 워크로드를 보유한 조직은 온프레미스 자체 구축이 상대적으로 유리한 것으로 분석된다[21].

나. 자체 AI 인프라 구축과 클라우드 이용의 경제성 및 TCO 비교

자체 인프라 구축(CAPEX 중심)과 클라우드 활용(OPEX 중심)의 선택은 총소유비용(TCO) 분석을 통해 결정되어야 한다. 온프레미스 구축은 서버·GPU·네트워크 등 초기 자본투자에 더해 연간 유지보수비(시스템 원가의 약 10% 내외), 전력·냉각비, 상면 비용 등이 지속적으로 발생하며, 통상 5년 내용연수를 기준으로 상각한다[22]. 반면 클라우드는 초기 자본투자가 없는 대신, 온디맨드 기준으로는 지속 사용 시 장기적으로 온프레미스보다 비용이 높아질 수 있어 손익분기점(Break-even Point) 분석이 필수적이다. 다수의 비용분석 모델에 따르면 500대 이상 GPU 규모에서 상당한 가동률을 전제로 할 경우 온프레미스 구축이 클라우드 임대보다 12~24개월 내 손익분기점을 넘어서는 것으로 나타난다[23].

〈표 3〉은 클라우드·온프레미스 선택 시 판단 항목을 비교하여 정리한 것이다.

〈표 3〉 클라우드·온프레미스 선택 시 판단 항목 비교

항목	클라우드	온프레미스·전용	판단 포인트
초기투자	낮음	높음	PoC는 클라우드가 유리
탄력성	높음	제한적	변동 워크로드는 클라우드가 유리
장기 단가	사용량에 따라 증가	높은 이용률에서 유리 가능	가동률·수명이 핵심 변수
보안·통제	서비스 옵션에 의존	직접 통제 가능	규제·데이터 민감도 고려
운영인력	상대적으로 적음	필요 역량 큼	인력비를 TCO에 반드시 포함
업그레이드	사업자 주도	직접 부담	AI 세대교체 속도 고려

다. AI 인프라 투자와 AX 성과를 연계하기 위한 AI FinOps 및 Business KPI 관리

인프라 투자와 실제 업무 성과를 연결하는 체계적 관리 방법론이 AI FinOps다. AI FinOps는 클라우드 진영에서 발전해온 비용 투명성·배분·최적화 원칙을 GPU·AI 인프라 영역까지 확장한 것으로, 온프레미스와 클라우드를 아우르는 총소유비용을 지속적으로 추적·비교하여 워크로드별 최적 배치를 결정하는 데 활용된다[24]. 다만 AI FinOps가 비용 관리에만 머물러서는 인프라 투자의 정당성을 조직 내에서 설득하기 어렵다. 따라서 인프라 투자 효과를 업무처리 시간 단축, 오류율 감소, 고객응대 정확성 향상과 같은 구체적인 Business KPI와 연동하여 모니터링하는 체계를 함께 갖추어야 하며, 이를 통해 인프라 투자가 단발성 실험이 아니라 장기적인 운영 모델 개선으로 이어지도록 관리할 필요가 있다.

〈표 4〉는 AI FinOps 체계에서 인프라 지표를 업무·사업 성과 지표까지 계층적으로 연결하는 대표적인 KPI 예시를 정리한 것이다.

〈표 4〉 AI FinOps·Business KPI 계층 구조

영역	권장 KPI	예시
인프라	GPU 이용률, 지연시간, 가동률	GPU 이용률 70% 이상
AI 품질	정확도, groundedness, hallucination율, task success율	업무 성공률 +10%p
비용	비용/요청, 비용/토큰, 비용/업무	처리건당 비용 30% 감소
업무	처리시간, 오류율, 생산성	처리시간 50% 단축
고객	CSAT, NPS, 응답시간	응답시간 40% 단축
사업	매출, 전환율, 재구매율	AI 적용 상품 매출 증가

라. 보안·신뢰·거버넌스를 고려한 Hybrid/Multi-Cloud 기반 AX 환경 구축

AX 환경을 설계할 때는 비용·성능 못지않게 보안·신뢰·거버넌스 요건을 함께 고려해야 한다. 보안·데이터 주권이 중요한 워크로드는 국내 클라우드나 정부 지원 인프라를 활용하고, 범용적인 개발·실험 워크로드는 글로벌 퍼블릭 클라우드를 활용하는 방식으로 Hybrid·Multi-Cloud를 설계하는 것이 현실적이다. 이때 각 클라우드·인프라에 대한 접근권한 관리, 데이터 흐름 통제, 감사로그 확보 등 거버넌스 체계를 사전에 표준화해 두어야, 인프라가 다변화되더라도 일관된 보안 수준을 유지할 수 있다.

마. 중소·중견기업의 AI Cloud 활용 및 단계적 AX 전환 전략

대기업과 달리 자체적으로 대규모 인프라를 구축하기 어려운 중소·중견기업은 정부 지원 클라우드·바우처 사업을 초기 마중물로 활용하여 PoC·실증 단계의 비용 부담을 낮추고, 성과가 검증된 워크로드부터 단계적으로 자체 인프라 또는 전용 클라우드로 확장해 나가는 전략이 현실적이다. 〈표 5〉는 기업의 AI 성숙도 단계별로 권장되는 클라우드·인프라 구축 모델을 정리한 것이다.

〈표 5〉 기업 AI 성숙도별 권장 클라우드·인프라 구축 모델

AI 성숙도 단계	권장 인프라 모델	핵심 고려사항
탐색·PoC 단계	퍼블릭 클라우드 온디맨드 GPU 인스턴스, 관리형 AI 서비스(AIaaS)	초기 투자 최소화, 빠른 실험·검증 속도, 정부 바우처 등 외부 지원 활용
실증·파일럿 단계	예약 인스턴스 중심 하이브리드 클라우드, 관리형 RAG·Vector DB 활용	특정 업무 데이터 보안 요건 검토, 초기 AI FinOps·Business KPI 체계 도입
운영·확산 단계	핵심 워크로드 온프레미스/전용 클라우드 이전, 비핵심 워크로드는 퍼블릭 클라우드 유지	TCO·손익분기점 분석 기반 투자 결정, MLOps/LLMOps 표준화
전사 확산 단계	Multi-Cloud·Sovereign Cloud 병행, 자체 AI 팩토리/전용 인프라 검토	데이터 주권·컴플라이언스, 전사 AI 거버넌스 체계, 지속가능성(전력·냉각)

바. 하드웨어 리프레시 전략과 운영 단계 소프트웨어 최적화: 학술 연구가 제시하는 시사점

앞서 소개한 마이크로소프트 애저 리서치·UIUC 공동 연구는 하드웨어 갱신(Refresh) 정책에 대해서도 중요한 시사점을 제공한다. 범용 데이터센터는 통상 5년 주기의 고정된 서버 교체 정책을 따르지만, AI 가속기는 세대별 성능·전력효율 개선폭이 일정하지 않기 때문에 고정 주기 갱신은 최적이 아니며, 유연한 갱신 전략을 적용할 경우 기존 방식 대비 TCO를 15~20% 절감할 수 있는 것으로 나타났는데[29], 이를 세 가지 패턴으로 정리한다. 첫째, '신형이 훨씬

낫다(Newer is much better)' – 신세대 GPU가 성능·효율 면에서 큰 폭으로 개선되었다면 구세대를 조기 폐기하고 즉시 전환하는 것이 유리하다(예: V100→A100). 둘째, '구형도 경쟁력이 있다(Older is competitive)' – 대부분의 세대는 6년 이상 수명을 연장하는 것이 비용 효율적이다. 셋째, '신형이 비슷하거나 더 나쁘다(Newer is similar or worse)' – 신세대의 개선폭이 미미하거나 오히려 효율이 떨어지는 경우, 해당 세대를 건너뛰고 기존 하드웨어의 수명을 연장하는 것이 유리하다[29].

운영(Operate) 단계에서는 모델·하드웨어·워크로드 특성에 맞춘 소프트웨어 최적화 기법이 하드웨어 갱신 비용을 이연시키고 기존 자산의 활용 가치를 높이는 효과를 낸다. 〈표 6〉은 주요 운영 최적화 기법과 그 효과를 정리한 것이다.

〈표 6〉 운영 최적화 기법과 효과

최적화 기법	설명	TCO 관점의 효과
모델 마이그레이션 스무딩	신규 모델 출시 시 급격한 전환 대신 점진적 이관	하드웨어 급조달 회피
양자화(Quantization)	낮은 정밀도 연산으로 연산·메모리 요구량 절감	추론당 하드웨어 수요·비용 절감
KV-Cache 관리	캐시 저장·재사용 최적화	구형 하드웨어 재활용도 제고
프리필·디코드 분리(Disaggregation)	추론 단계를 서로 다른 하드웨어에 분산 배치	이기종 세대 하드웨어의 유효수명 연장
대안 아키텍처(MoE·SSM)	혼합전문가·상태공간모델 등으로 연산량 절감	구형 하드웨어 재활용도 제고
모델 라우팅	요청 복잡도에 따라 최적 모델로 분기	구형 하드웨어 재활용도 제고
이기종 인지 스케줄링	워크로드를 최적·가용 하드웨어 세대에 매칭	하드웨어 갱신 비용 이연
인프라 인지 스케줄링	전력·냉각 여유분을 소프트웨어로 활용	인프라 활용 효율 제고

개별 최적화 기법은 기존 방식 대비 12~39%의 TCO 절감 효과를 보였으며, 이 중 모델 마이그레이션 스무딩이 가장 큰 개별 효과를 나타냈다. 여러 기법을 결합할 경우 절감 효과가 완전히 합산되지는 않지만(중첩 효과 존재), 운영 단계 최적화만으로도 60% 이상의 누적 TCO 절감이 가능한 것으로 보고되었다[29].

사. 글로벌 리서치가 제시하는 하이브리드 AI 인프라 투자 전략

IDC가 2025년 2월 발표한 스포트라이트 보고서(Intel 후원)에 따르면, 기업의 AI 중심 시스템 지출은 2022~2026년 연평균 27% 성장할 것으로 전망되며, 2024년 한 해 생성형 AI 솔루션에 대한 기업 투자는 이미 200억 달러를 넘어선 것으로 나타났다[31]. 그러나 IDC는 2026년까지 글로벌 2000대 기업(G2000) C레벨 경영진의 70%가 신규 AI 인프라 투자를 승인하기에 앞서 기술·비즈니스 성과의 연계와 투자수익률(ROI)·가치 분석을 요구할 것으로 전망하며, 과거의 일방향적 인프라 확충과 달리 향후 AI 인프라 투자는 철저한 경제성 검증을 동반하게 될 것이라고 진단한다[31]. 이러한

흐름 속에서 2028년까지 기업 AI 워크로드의 75%가 가속기·프로세서, 퍼블릭·프라이빗 클라우드, 독립형·임베디드 AI 워크로드 각각에 대해 용도에 맞게 최적화된(fit-for-purpose) 하이브리드 인프라 위에서 운영될 것으로 전망되는데[31], 이는 국내 기업의 AI 성숙도별 인프라 선택 전략과 정확히 같은 방향성을 갖는다.

IDC는 신규 가속기 도입에 앞서 기존 범용 인프라(x86 CPU 기반)를 우선적으로 활용하는 것이 합리적인 접근이라고 강조한다. 최신 서버 프로세서는 매개변수 200억 개 이하의 대규모 언어모델에 대해 100ms 미만의 토큰 지연시간으로 실시간 서비스가 가능한 수준의 성능을 제공하며, 5년 이상 노후화된 서버를 최신 프로세서로 교체·통합할 경우 동일 성능을 더 적은 서버 대수로 구현함으로써 TCO를 최대 77%까지 절감할 수 있는 것으로 나타났다[31]. 이는 모든 AI 워크로드에 GPU 등 전용 가속기가 필요한 것은 아니며, 학습 시간이 결정적 변수가 아닌 경우 기존 CPU 기반 플랫폼을 추론·미세조정·소규모 학습에 그대로 활용함으로써 신규 투자를 최소화할 수 있음을 시사한다. 이러한 접근은 국내 중소·중견기업의 AX 원스톱 바우처 활용 전략에서도, GPU 확보에 앞서 기존 인프라의 활용 가능성을 먼저 점검하는 순서로 적용될 수 있다.

보안·컴플라이언스 측면에서 IDC는 데이터의 저장(at rest)·전송(in flight) 구간 암호화만으로는 시스템 메모리에 대한 저수준 공격을 막을 수 없다는 점을 지적하며, 가상머신·컨테이너·애플리케이션 전체를 보안 연산 엔클레이브(secure computing enclave)로 이전하는 기밀 컴퓨팅(Confidential Computing)을 제로트러스트 보안 전략의 핵심 요소로 제시한다[31]. 기밀 컴퓨팅은 퍼블릭 클라우드·온프레미스·코로케이션 등 어떤 환경에서도 데이터 사용 중(data-in-use) 암호화를 지원해, 예컨대 의료기관이 환자 데이터를 직접 공유하지 않고도 연합학습(Federated Learning)으로 모델을 공동 훈련하는 등의 활용이 가능하다고 설명한다[31]. 이는 Sovereign Cloud·데이터 주권 이슈와 맞닿아 있으며, 국내 공공·금융 부문의 민감정보 기반 AI 서비스 설계에도 참고할 수 있는 방향이다.

한편 IDC는 2027년까지 하이브리드 인프라를 운영할 수 있는 전문인력 부족이 지속될 것으로 전망하며, 이러한 역량 격차(skills gap)를 조기에 해소하지 못하는 조직은 전체의 60%에 달해 ROI 실현이 지연되거나 좌절되는 이른바 '스킬 포트홀(skills pothole)'에 빠질 위험이 크다고 지적한다[31]. 실제로 이러한 역량 공백을 극복하고 실질적인 ROI를 달성할 수 있는 조직은 전체의 40% 수준에 그칠 것으로 전망되는데[31], 이는 국내 기업의 클라우드·MLOps·보안 복합역량 부족 문제와도 직결되며, AX 원스톱 바우처의 공급기업 컨소시엄 구조가 이러한 역량 격차를 외부 파트너십으로 보완하는 정책수단으로서 갖는 의의를 뒷받침한다.

아. 보안·거버넌스의 내재화

AX 시스템은 모델 자체뿐 아니라 데이터, 프롬프트, 검색 결과, 그리고 Agent의 도구 호출 권한까지 새로운 공격표면을 만든다. 따라서 IAM(접근권한관리), 네트워크 분리, 암호화, 비밀관리, 감사로그, 모델·프롬프트 버전관리, Guardrail, 데이터 마스킹, Human-in-the-loop 검토 절차를 AI 서비스 생명주기 전 단계에 내재화해야 한다. 〈표 7〉은 AX 환경에서 관리해야 할 통제영역과 핵심 관리내용을 정리한 것이다.

〈표 7〉 AX 환경의 보안·거버넌스 통제영역

통제영역	필수 관리내용
Identity	사용자·서비스·Agent별 최소권한 부여
Data	데이터 분류·암호화·마스킹·보존기간 관리
Model	모델 버전·출처·평가결과·승인 이력 관리
Prompt/RAG	프롬프트·검색근거·출처 추적
Agent	도구별 권한·호출정책·승인 절차
Runtime	취약점 점검·네트워크 격리·로그 수집
Governance	책임자 지정·감사·사고대응·변경관리

이러한 통제체계는 Sovereign Cloud·Confidential Computing, 국내 공공·금융 부문의 민감정보 활용 이슈와 함께 설계되어야 하며, 특히 Agent가 외부 도구를 자율적으로 호출하는 구조에서는 사고 발생 시 책임소재를 규명할 수 있는 감사 추적 체계가 필수적이다.

6 AX 원스톱 바우처와 AI 클라우드·인프라의 연계 및 시사점 – AI·클라우드·데이터의 통합 지원

가. AX 전환 과정에서 기업이 겪는 주요 애로사항

기업이 AX를 추진하는 과정에서 가장 먼저 부딪히는 장벽은 GPU·클라우드 인프라 비용과 이를 설계·운영할 전문역량의 부족이다. AI 솔루션 구매·맞춤형 개발 비용에 더해 클라우드 인프라와 GPU 자원 사용료, 기존 레거시 시스템 연동 개발 비용까지 초기 투자 부담이 크기 때문에, 투자 여력이 제한적인 중소·중견기업은 AI 도입 필요성을 인식하더라도 실증 단계로 진입하기 어렵다. 여기에 외부 AI·클라우드 전문 인력을 확보하기 어렵고, 도입 성과를 어떤 지표로 관리할지에 대한 체계적 방법론도 부재한 경우가 많아, 개별 장벽이 상호 연결되어 AX 추진 전반을 제약하는 구조를 이룬다.

나. AI 솔루션·GPU·SaaS 등을 통합 지원하는 AX 원스톱 바우처의 특징

과학기술정보통신부와 정보통신산업진흥원(NIPA)이 2026년부터 기존의 AI 바우처와 클라우드 바우처를 통합하여 추진하는 'AX 원스톱 바우처' 지원사업은 이러한 복합적 장벽을 클라우드·인프라 측면에서 동시에 완화하는 정책수단이다[25][26]. 2026년도 사업은 총 260억 원 규모로, 수요기업이 이용 금액의 최대 75%까지 정부 지원을 받을 수 있으며 수요기업이 20%, 공급기업이 5~10%를 자체 부담하는 구조로 설계되어 있다[27]. 클라우드 영역에서는 GPU 사용 및 모델 학습·운영 환경 구축, SaaS형 서비스로의 전환을 위한 기술 지원, 그리고 개발된 AI 서비스의 클라우드 마켓플레이스 등록 지원까지 포함된다[28].

다. 단순 GPU 지원 또는 AI 솔루션 구매 지원과 차별화되는 원스톱 지원체계의 의미

AX 원스톱 바우처가 기존의 개별 GPU 바우처나 AI 솔루션 구매 지원 사업과 구별되는 지점은, 인프라·솔루션·데이터를 분리 지원하지 않고 하나의 컨소시엄 구조 안에서 통합적으로 연계한다는 데 있다. AI 개발·클라우드·데이터 세 영역의 공급기업이 하나의 컨소시엄으로 묶여 수요기업에 서비스를 제공하며, 클라우드 공급기업은 AI SaaS 전환 환경과 GPU 자원 제공이 가능하며 NIPA AX 공급기업 POOL에 등록된 기업으로 한정된다. 이는 검증된 국내 클라우드 사업자를 매칭함으로써 수요기업의 인프라 리스크를 낮추는 장치로 기능하며, 개별 기업이 인프라 조달과 서비스 개발을 각각 별도로 진행할 때 발생하는 시행착오와 이중 비용을 줄이는 효과를 기대할 수 있다.

〈표 8〉은 기업이 AX 추진 과정에서 겪는 대표적인 문제와, AX 원스톱 바우처가 각 문제에 대응하는 지원요소를 정리한 것이다.

〈표 8〉 AX 원스톱 바우처의 지원요소별 대응 구조

지원요소	기업의 문제	원스톱 바우처의 역할
AI 개발	전문인력·모델 개발비 부족	업무 맞춤형 AI 개발 지원
Cloud/GPU	초기 인프라 비용·운영 역량 부족	GPU 및 AI 운영환경 제공
Data	학습·검색 데이터 확보의 어려움	데이터 구매·가공·품질관리 지원
SaaS	개발 결과가 일회성 시스템에 그침	서비스화 및 확산 기반 마련
사업화	실증 이후 시장 진입 단절	Marketplace 등록 등 확산 경로 제공

라. 수요기업과 AI·클라우드·데이터 공급기업 연계를 통한 서비스 개발 및 사업화 지원

본 사업은 수요기업이 주관하고 AI·클라우드·데이터 공급기업이 컨소시엄으로 참여하는 구조로 운영되며, 최소 1개 이상의 국산 AI 파운데이션 모델(하이퍼클로바X, 엑사원 등) 적용이 필수 요건으로 규정되어 있다[28]. 이는 개별 기업의 AX 지원을 넘어 국산 AI 모델·클라우드 생태계의 초기 레퍼런스를 동시에 확보하는 산업정책적 기능도 수행하고 있음을 보여준다.

AX 원스톱 바우처 사업의 가장 큰 특징은 수요기업의 기획 단계 분석부터 AI·데이터 확보, 클라우드·GPU 인프라 구축, AI 서비스 개발, 실제 업무 실증, 그리고 SaaS화·마켓플레이스 등록을 통한 확산까지 전 주기를 하나의 지원체계로 연결한다는 데 있다. 이러한 전주기 설계는 개별 기업이 인프라 조달, 솔루션 개발, 시장 진입을 순차적으로 별도 추진할 때 발생하는 단절과 지연을 줄이고, 정부 지원이 끝난 이후에도 SaaS·마켓플레이스를 통한 자생적 사업화로 이어지도록 유도한다는 점에서 정책적 완결성을 갖는다.

7 결론 – AX 시대의 클라우드·인프라 발전 방향 및 정책적 시사점

가. GPU 중심 인프라에서 AI Computing·AI Factory 등 통합 AI 인프라로의 발전

향후 AI 인프라는 단순 GPU 자원의 총량 경쟁에서 벗어나, 학습과 추론을 하나의 파이프라인으로 통합 운영하는 AI 팩토리 단위의 효율성 경쟁으로 이동할 것으로 전망된다. 이 과정에서 성능 지표 역시 연산능력(FLOPS) 중심에서 토큰당 비용·토큰당 전력소비 등 실질적인 서비스 생산성 지표로 전환될 가능성이 높다[2].

나. AI 모델 학습 중심에서 학습·추론 및 AI 서비스 운영 중심으로의 변화

기업과 정부 모두 인프라 투자를 계획할 때 초기 학습 역량뿐 아니라 장기적인 추론·운영 효율성을 함께 고려하는 관점의 전환이 요구된다. 이는 인프라 수요의 이동, 그리고 글로벌 AI Factory·AI-Native Cloud로의 진화와 같은 흐름이다.

다. 기업의 AX 경쟁력 강화를 위한 클라우드·AI 플랫폼·데이터의 통합 필요성

기업의 AX 경쟁력은 클라우드·AI 플랫폼·데이터 중 어느 한 요소만으로는 확보되지 않으며, 세 요소가 AI FinOps·Business KPI 관리 체계 아래 통합적으로 관리될 때 비로소 지속가능한 성과로 이어진다. 보안·신뢰·거버넌스를 고려한 Hybrid·Multi-Cloud 설계 역시 이러한 통합 전략의 핵심 축이다. 마이크로소프트 애저 리서치·UIUC의 최신 연구는 건설·하드웨어 갱신·운영을 개별적으로 최적화할 때보다 세 단계를 통합적으로 재설계(Cross-Stage)할 때 TCO 절감 효과가 최대 40%까지 확대된다는 점을 실증적으로 보여, 클라우드·AI 플랫폼·데이터의 통합적 관리가 이론이 아니라 검증된 접근임을 뒷받침한다[29]. 같은 맥락에서 IBM·UIUC IIDAI의 하이브리드 클라우드 백서 역시 LLM 기반 추상화(LLMaaA)·통합 제어 평면·에너지 최적화를 하나의 스택으로 공동설계할 때에만 100~1,000배 수준의 성능/전력 개선이 가능하다고 진단하고 있어[30], 국내 기업의 AX 전환 역시 개별 기술 요소의 도입이 아니라 클라우드·AI 플랫폼·데이터·에너지를 아우르는 전체 스택 관점의 설계가 필요함을 뒷받침한다. IDC 역시 2026년까지 G2000 기업 C레벨의 70%가 ROI·가치 분석을 전제로 AI 인프라 투자를 승인할 것으로 전망하고 있어[31], 국내 기업의 AX 투자 역시 AI FinOps·Business KPI 관리 체계를 통해 투자 정당성을 입증하는 방향으로 나아가야 함을 뒷받침한다.

라. 국내 AI 클라우드·컴퓨팅 생태계 강화를 위한 정책적 시사점

첫째, 국가 AI 컴퓨팅센터·AI 고속도로와 같은 대형 국가 인프라 사업은 민간 투자 유인과 사업 추진 속도를 확보하는 방향으로 지분·규제 구조를 유연하게 조정하되, 국산 AI 반도체·K-Cloud 생태계 육성이라는 산업정책적 목표와의 균형을 지속적으로 재점검할 필요가 있다. 둘째, 대형 하이퍼스케일 인프라와 지역·산업 특화 소형 데이터센터를 상호 보완적으로 운영함으로써, 전력·부지 제약이 있는 국내 여건에서도 분산형 AI 인프라 생태계를 조성해야 한다. 셋째, 전력·냉각 등 지속가능성 이슈는 AI 인프라 정책의 부차적 고려사항이 아니라 GPU 확보와 동등한 수준의 핵심 정책 변수로 다루어져야 하며, 데이터센터 인허가·전력 승인 절차의 신속화가 국가 AI 인프라 경쟁력을 좌우하는 요인이 될 것이다.

마. AX 원스톱 바우처의 AI 공통기반 지원 및 기업 AX 확산을 위한 향후 발전 방향

AX 원스톱 바우처와 같은 수요기업 대상 지원사업은 GPU·클라우드 비용 지원에 그치지 않고, 전주기 지원 구조를 유지·확대함으로써 개별 기업의 인프라 조달 리스크를 정부가 일부 분담하는 정책적 역할을 지속해야 한다. 특히 단계평가를 통한 성과연동형 지원 방식은 지원의 실효성을 높이는 장치로서, 향후에도 지원 규모 확대와 함께 성과관리 체계를 정교화하는 방향으로 발전할 필요가 있다.

바. 국내 AX 확산을 위한 정책 제언

결론적으로, 기업의 AX 전환은 AI 모델의 성능만으로 완성되지 않는다. 학습에서 추론·운영으로 이동하는 인프라 수요, AI-Native Cloud로의 플랫폼 진화, 그리고 국가 차원의 AI 고속도로 구축과 AX 원스톱 바우처와 같은 지원정책이 유기적으로 결합될 때, 국내 기업의 AX 전환은 비로소 지속가능한 경쟁력으로 이어질 수 있을 것이다. 〈표 9〉는 AI 클라우드·인프라 주요 동향과 이에 대응하는 AX 원스톱 바우처의 정책적 대응 방향을 요약한 것이다.

〈표 9〉 AI 클라우드·인프라 주요 동향과 AX 원스톱 바우처의 정책적 대응 방향

2026년 주요 동향	기업이 겪는 영향	AX 원스톱 바우처의 대응
글로벌 AI 인프라 투자 급증·GPU 확보 경쟁 심화	GPU 단가·확보 난이도 상승, 중소·중견기업 진입장벽 확대	GPU·클라우드 이용료 최대 75% 지원으로 초기 비용 부담 완화
AI-Native Cloud로의 플랫폼 통합(MLOps·RAG·Agent)	개별 기업이 전 구성요소를 자체 구축하기 어려움	AI·클라우드·데이터 공급기업 컨소시엄을 통한 통합 서비스 제공
국가 AI 컴퓨팅센터·AI 고속도로 등 국가 인프라 확충	국가 인프라와 민간 수요 간 활용 경로 필요	국산 AI 모델·국내 클라우드 활용 의무화로 국가 인프라와 연계
추론·에이전틱 AI 중심으로 수요 이동	상시 운영형 인프라 설계·운영 노하우 부족	실증 단계를 통해 실제 업무 적용 및 SaaS화까지 지원

사. 기업 AX 전환을 위한 3개년 인프라 구축 로드맵

기업의 인프라 전환은 단년도 서버 구매계획보다 3개년 관점의 단계적 로드맵으로 설계하는 것이 현실적이다. 1차년도에는 데이터·업무 프로세스를 정비하고 클라우드 기반으로 AI 실증을 수행하며, 2차년도에는 운영환경과 전용자원을 확대하고, 3차년도에는 핵심 워크로드의 내재화와 SaaS·외부 서비스로의 확산을 검토하는 구조가 바람직하다. 〈표 10〉은 이러한 3개년 로드맵을 정리한 것이다.

〈표 10〉 기업 AX 전환을 위한 3개년 인프라 구축 로드맵

연도	핵심목표	인프라	플랫폼	성과
1차년도	PoC·실증	Public Cloud/GPUaaS	RAG·MLOps 기초 체계	업무 KPI 검증
2차년도	Production 전환	Hybrid Cloud·전용 GPU	LLMOps·Agent·관측성	SLA·TCO 안정화
3차년도	확산·내재화	Hybrid/Multi-Cloud	Model Router·AI FinOps	사업화·재사용 확대

이 로드맵의 핵심은 매년 더 많은 GPU를 확보하는 것이 아니라, 업무 성과가 검증된 워크로드의 비중을 단계적으로 높이고 이에 맞추어 인프라·플랫폼을 재설계하는 데 있다. 이는 앞서 AI 성숙도별 인프라 선택 전략, 그리고 AX 원스톱 바우처의 전주기 지원 구조와 정확히 맞닿아 있으며, 정부지원 종료 이후에도 기업이 자체적으로 이 로드맵을 이어갈 수 있도록 단계별 목표를 사전에 설계해 두는 것이 중요하다.

[참고문헌]

[1] mycalctoolkit, AI Data Center Cost Calculator: GPU Cluster & Infrastructure Cost, 2026.

[2] NVIDIA, AI Factories: The New Infrastructure of Intelligence, NVIDIA Blog, 2026.05.

[3] Data Center Dynamics/Datacenter Richness, Hyperscalers Plan $630 Billion in 2026 CapEx, 2026.02.

[4] Yahoo Finance/Goldman Sachs, Meta·Microsoft·Amazon·Alphabet 2026 CapEx 전망, 2026.06.

[5] IBKS투자증권, AI 컴퓨팅 판매업자들: 네오클라우드, 2026.

[6] ValueAddVC, AI Capex 2026: Where Microsoft, Google, Meta & Amazon's $725B Actually Goes, 2026.08.

[7] Lenovo/Engineering.com, Lenovo Unveils AI Cloud Gigafactory with NVIDIA at CES 2026, 2026.01.

[8] UPI, Naver, Nvidia Launch Gigawatt-scale AI Factory Plan, 2026.06.

[9] mycalctoolkit, AI Data Center Cost Calculator(PUE 기준), 2026.

[10] IESVE, From Data Centres to AI Factories: What NVIDIA GTC 2026 Revealed, 2026.

[11] GII Korea, 클라우드 인프라 서비스 시장 규모, 점유율, 시장예측 2026-2032년, 2026.07.

[12] 국내 공공부문 AI 도입 사례(범정부 AI 공통 기반 활용 지원 사업), 행정안전부·한국지능정보사회진흥원(NIA) 관련 자료 종합, 2026.

[13] UK Cabinet Office & GDS / GovTech Singapore, Redbox·Pair 관련 정부 발표자료 종합, 2026.

[14] 한국지능정보사회진흥원(NIA), 공공분야 초거대 AI 활용을 위한 공공데이터 주권 클라우드 적용방향, IT&Future Strategy 보고서.

[15] 국가인공지능전략위원회·과학기술정보통신부, 대한민국 인공지능 기본계획(2026~2028), 2026.02.

[16] 페블러스, 국가 AI 컴퓨팅센터 착공에 들어가다, 2026.07.

[17] 엑스디노드, 9.9조원의 AI 예산, 2026년 정부 인프라 지원사업 정리, 2026.07.

[18] 대한민국 정책브리핑, 4000억 원 규모 AI반도체 기반 'K-클라우드 기술개발' 프로젝트 추진, 2025.05.

[19] 뉴시스, 리벨리온·퓨리오사AI 등 K-반도체 글로벌 잭팟…'실패 감수 국가 R&D'가 힘 됐다, 2026.06.(IITP 2026 성과 미디어 데이)

[20] Wisdom Agora, K-클라우드 프로젝트란?, 2026.

[21] apxml.com, On-Premise AI Total Cost of Ownership(TCO), 2026.

[22] Lenovo Press, On-Premise vs Cloud: Generative AI Total Cost of Ownership(2026 Edition), 2026.

[23] mycalctoolkit, AI Data Center Cost Calculator: Break-even 분석, 2026.

[24] Introl Blog, AI 인프라 파이낸싱: CapEx, OpEx, 그리고 GPU 투자 전략, 2026.02.

[25] 과학기술정보통신부·정보통신산업진흥원(NIPA), 2026년도 AX 원스톱 바우처 지원사업 수요기업 모집 공고, 2026.03.

[26] 엑스디노드, 9.9조원의 AI 예산, 2026년 정부 인프라 지원사업 정리(AX 원스톱 바우처 통합 내용), 2026.07.

[27] THE VC, 2026년 AX 원스톱 바우처 지원사업 수요기업 모집 공고, 2026.03.

[28] 엘리스, 2026 AX 원스톱 바우처 지원사업 수요기업 모집, 2026.04. / RnDcircle, 2026년도 AX 원스톱 바우처 지원사업(수요기업 AX 전환 과제), 2026.03.

[29] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, and Ricardo Bianchini, "Rearchitecting Datacenter Lifecycle for AI: A TCO-Driven Framework," University of Illinois Urbana-Champaign/Microsoft Azure Research, arXiv:2509.26534v1 [cs.AI], 2025.09.30.

[30] Deming Chen, Alaa Youssef, et al., "Transforming the Hybrid Cloud for Emerging AI Workloads," IBM Research/University of Illinois Urbana-Champaign, IBM-Illinois Discovery Accelerator Institute (IIDAI) White Paper, 2024.11.

[31] Ashish Nadkarni, "AI Infrastructure in 2025: Balancing Datacenter and Cloud Investments," IDC Spotlight(Sponsored by Intel Corp.), Doc #US53194325, 2025.02.

[작성자]

소속 상명대학교
성명 서광규
부서 공과대학 경영공학과
직위 교수

주요 연구 실적

- 「스마트 팩토리의 지속 가능성을 위한 Claroty 기반의 OT(Operational Technology) 보안 강화 방안에 관한 연구」, 연구논문, 2026.06
- 「제로트러스트 기반의 클라우드 보안인증제 강화방안에 관한 연구」, 연구논문, 2025.12
- 「클라우드 기반 정보시스템 평가지표 개발」, 연구논문, 2025.12
- 「AI-PaaS 프레임워크 개발을 위한 기술 요소 분석」, 연구논문, 2024.12
- 「클라우드 기반 인공지능 플랫폼 도입 평가 프레임워크 개발」, 연구논문, 2023.09

AX 전환을 지원하는 클라우드 환경 및 인프라 구축 동향

발 행 처 : 정보통신산업진흥원

작 성 기 관 : 한국데이터산업협회

발 행 일 : 2026년 7월



AX 전환을 지원하는 클라우드 환경 및 인프라 구축 동향

과학기술정보통신부
Ministry of Science and ICT

nipa 정보통신산업진흥원
National IT Industry Promotion Agency