

AI 시대 네트워크 패러다임 전환과 AI 네트워크 도입 효과 분석

A Paradigm Shift Toward AI Networks and an Analysis of the Impacts of Their Adoption

박정렬 (J.R. Park, jrpark16@etri.re.kr)
송영근 (Y.G. Song, iesong@etri.re.kr)

미래전략연구실 선임연구원
미래전략연구실 책임연구원

ABSTRACT

The proliferation of AI services is reshaping traffic patterns, AI infrastructure, computation placement, and network operations, exposing the limits of conventional architectures. This paper establishes the inevitability of the transition to AI networks, defining the “AI network” as an integrated paradigm that combines “AI for networks,” which brings intelligence to network operations, and “networks for AI,” which restructures the network into an execution infrastructure for AI, and reviews its key technologies alongside standardization and industry trends. The analysis estimates that AI-driven automation can reduce network operator costs by about 17.75% and total cost of ownership by about 7.1%, while the AI network equipment market is projected to grow from USD 13.8 billion in 2030 to USD 95.6 billion in 2037 (a compound annual growth rate of 31.9%). These findings identify the AI network as both a cost-transformation lever and a new growth market for operators as well as a strategic linchpin of national AI infrastructure competitiveness.

KEYWORDS AI for Network, AI-RAN, Network for AI, AI 네트워크, 도입 효과, 시장 전망

I. 서론

생성형 AI의 확산과 퍼지컬 AI의 부상은 컴퓨팅 산업을 넘어 이를 뒷받침하는 네트워크 인프라의 구조적 전환을 촉구하고 있다. 주요국 정부와 글로벌 기업들은 '26년을 전후하여 정책과 투자의 무게

중심을 AI 기반 무선망과 AI 데이터센터 네트워크로 급속히 옮겨가고 있다. 미국은 무선 공급망 혁신 펀드(Public Wireless Supply Chain Innovation Fund)를 자국 기반의 AI 내재형(AI-native) 네트워크 아키텍처 촉진을 기치로 재편하였고[1], 우리 정부도 '25년 12월 '하이퍼 AI 네트워크 전략'을 발표하여 '30년

* DOI: <https://doi.org/10.22648/ETRI.2026.J.410513>

* 본 논문은 한국전자통신연구원 ICT전략연구소 기본사업인 [ICT 기술전략 및 기술정책 연구(26ZF1100)]를 통해 작성된 결과입니다.

6G 상용화와 지능형 기지국(AI-RAN)의 전국 구축이라는 청사진을 제시하였다[2]. 이는 네트워크가 더 이상 AI 산업의 배경 인프라가 아니라, AI 경쟁력 자체를 좌우하는 전략 자산으로 격상되고 있음을 방증한다.

이러한 전환의 근거에는 AI 서비스가 만들어내는 트래픽 구조, 인프라 형태, 연산 배치 방식의 변화와 함께, 네트워크 자체의 운영 복잡성 증가라는 구조적 변화가 자리하고 있다. 기존 네트워크가 사람의 콘텐츠 소비를 중심으로 설계되고 사람의 판단을 중심으로 운영되었다면, AI 시대의 네트워크는 대규모 분산 학습과 실시간 추론, 물리 세계와 상호작용을 하는 AI를 지원해야 하는 동시에, 사람의 관리 역량을 넘어서는 복잡성을 스스로 감당해야 하는 이중의 요구에 직면해 있다.

그럼에도 지금까지의 논의는 개별 요소 기술의 소개나 특정 표준화 기구 중심의 동향 정리에 머무르는 경우가 많아, 두 요구를 하나의 틀에서 조망하고 전환의 효과를 정량적으로 가늠하게 해주는 분석은 충분히 축적되지 못한 실정이다. 이러한 분석은 정부에는 연구개발 투자와 실증 정책의 우선순위를 판단할 근거를, 산업계에는 시장 진입 시점과 사업 전략을 설계할 기준을, 학계에는 후속 연구의 출발점이 될 공통의 개념 지도를 제공할 수 있다는 점에서 의의가 있다.

이에 본고의 II장에서는 AI 시대 네트워크 패러다임의 변화를 네 가지 측면에서 살펴 AI 네트워크로의 전환이 갖는 필연성을 논증한다. III장에서는 AI 네트워크의 개념을 정의하고 이를 구성하는 주요 기술의 동향을 표준화와 산업 실증의 흐름과 함께 고찰한 후, IV장에서는 AI 네트워크 도입 효과를 통신사업자의 비용 절감과 신규 시장 창출의 양 측면에서 정량적으로 분석한다. 마지막으로 V장에서는 분석 결과를 종합하고 시사점을 제시한다.

II. AI 시대 네트워크 패러다임 변화와 AI 네트워크의 필요성

1. AI 트래픽 구조의 변화

AI 서비스의 확산에 따라 네트워크 트래픽의 구조는 사용자 소비 중심에서 데이터 생성과 상호작용 중심으로 재편되고 있다. 기존 모바일 트래픽은 다운로드와 업로드 비율이 약 9:1로 다운로드에 편중되어 있었으나, AI 서비스 환경에서는 대화형 입력, 센서 데이터 전송, 실시간 추론 요청 등이 증가하면서 업로드 비중이 약 7:3 수준까지 확대되고 있다[3].

특히 생성형 AI 서비스가 전체 AI 트래픽의 60% 이상을 차지하고 그중 업링크 비중이 약 70%에 달하면서, 네트워크는 다운로드 중심 구조에서 양방향 상호작용 중심 구조로 전환되고 있다[3]. 나아가 향후 AI 트래픽은 콘텐츠 이용을 넘어 기기와 센서가 생성한 데이터를 실시간으로 주고받는 방향으로 확대될 전망이다.

영상, 라이다(LiDAR), IoT 센서 등에서 생성되는 대용량 데이터는 단일 디바이스 기준으로 초당 수십에서 수백 메가바이트(MB) 수준에 이를 수 있어 [4], 네트워크에는 콘텐츠 전달을 넘어 대규모 데이터 생성과 실시간 처리를 함께 뒷받침하는 구조로의 진화가 요청되고 있다.

트래픽의 방향뿐만 아니라 흐름의 질적 특성 또한 기존 인터넷 서비스와 판이하다. 대규모 언어모델(LLM)의 분산 학습에서는 수많은 GPU가 각자 계산한 학습 결과를 집합통신(Collective Communication)을 통해 반복적으로 교환하고 동기화하며, 이 과정에서 트래픽은 소수의 초대형 플로우(Elephant Flow)와 극단적인 버스트에 집중되는 특성을 보인다[5,6].

웹과 영상 트래픽이 다수의 소규모 흐름이 통계

적으로 섞이는 것을 전제로 설계되었다면, AI 학습 트래픽은 특정 시점에 대용량 데이터가 일제히 물리는 동기적 특성을 띠므로 기존의 설계 전제는 더 이상 유효하지 않게 된다. 이에 따라 AI 트래픽은 사용자-서버 간 트래픽보다 서버 간 대규모 데이터 교환, 즉 이스트-웨스트(East-West) 트래픽의 비중이 커지는 방향으로 재편되고 있으며[5], 네트워크를 AI 데이터 흐름의 고유한 특성에 정합하도록 재설계해야 한다는 요구가 제기되고 있다.

2. AI 인프라의 구조적 전환

생성형 AI의 확산으로 AI 연산은 단일 서버에서 처리되는 방식을 벗어나, 다수의 GPU와 가속기, 서버가 네트워크로 연결된 대규모 분산 클러스터 구조로 전환되고 있다. 엔터프라이즈 AI 인프라는 수십에서 1천여 개 GPU 규모의 클러스터 단위로 정의되고 있으며[7], 1만 6천여 대의 서버와 13만 개 이상의 가속기가 상호 연결되는 초대규모 AI 클러스터 사례도 제시되고 있다[8]. 이러한 규모에서 네트워크는 단순한 연결 수단이 아니라 흩어진 연산 자원을 하나의 시스템처럼 묶는 통합 연산 기반(Compute Fabric)으로 기능하며, AI 인프라의 성능을 규정하는 중핵으로 자리매김하게 된다.

이때 주목할 대목은 네트워크 성능이 AI 인프라의 투자 효율을 직접 좌우한다는 점이다. 분산 학습에서는 여러 GPU가 계산 결과를 맞히기 위한 집합 통신을 반복 수행하므로 네트워크의 지연과 대역폭이 전체 학습 시간에 직접적인 영향을 미치며[6], 네트워크 병목이 발생하면 GPU가 계산을 마치고도 데이터를 기다리는 유향 시간이 늘어나 고가의 인프라 투자 효율이 저하된다[7]. 특히 동기식 연산의 특성상 가장 늦게 도착하는 데이터 흐름을 기다려야 하므로, 평균 성능이 아니라 최악 경로의 지연

(Tail Latency)이 전체 작업 완료 시간을 좌우한다는 점이 AI 인프라 설계의 고유한 제약 조건으로 작용한다. 이에 대응하여 AI 클러스터의 네트워크 인터페이스는 400Gbps급에서 800Gbps급 이상으로 고도화되고 있으며, 대규모 클러스터에서는 수천 대 수준의 스위치가 요구되는 등 데이터센터 네트워크가 AI 인프라의 성능과 확장성을 좌우하는 핵심 인프라로 전환되고 있다[8].

AI 인프라의 확장은 단일 시설의 경계도 넘어서고 있다. 대규모 연산 시설은 전력과 공간, 냉각의 제약으로 인해 단일 시설 내에서 무한정 확장될 수 없으며, 가용 전력을 따라 지리적으로 분산 배치될 수밖에 없다. 이에 따라 복수의 데이터센터를 연결하여 대규모 학습을 공동 수행하는 구조의 필요성이 커지고 있으며, 장거리에서도 안정적이고 손실이 적은 고대역폭 연결망이 AI 인프라의 새로운 요구 조건으로 부상하고 있다[4]. 결국 가속기의 연산 성능이 빠르게 향상될수록 이를 잇는 네트워크가 상대적 병목이 되는 비중은 커지며, AI 인프라 경쟁의 차별화 지점이 연산 장치에서 이를 연결하는 네트워크로 옮겨가고 있다는 사실은 특히 주목할 만하다.

3. 엣지 AI와 피지컬 AI 확산에 따른 기존 네트워크 구조의 한계

엣지(Edge) AI의 확산과 피지컬 AI의 부상에 따라 AI 연산은 중앙 클라우드 의존 방식에서 벗어나 디바이스-엣지-클라우드에 걸쳐 분산 수행되는 구조로 전환되고 있다[9,10]. 자율주행, 로봇, 확장현실(XR)과 같은 서비스는 실시간 AI 추론이 필요하므로, 모든 데이터를 중앙 클라우드로 보내 처리하는 방식만으로는 지연 요구를 충족하기 어렵다. 이에 따라 AI 연산의 엣지 처리 비중이 증가할 것으로 전

망되며, 네트워크는 디바이스-엣지-클라우드 간 데이터와 연산을 연결하고 서비스 실행 위치와 품질을 결정하는 핵심 인프라로 기능하게 된다[9].

피지컬 AI 환경에서 네트워크 지연은 서비스 품질의 문제를 넘어 물리적 안전성과 직결된다는 점에서 그 함의가 사뭇 다르다. 산업 로봇이나 자율주행처럼 실제 물리 동작이 수반되는 분야에서는 밀리초(ms) 수준의 지연도 제어 실패나 안전 문제로 이어질 수 있어, 평균 지연의 단축을 넘어 최악 상황의 지연과 지터(Jitter)까지 엄격히 관리함으로써 극히 낮은 지연을 안정적으로 보장하는 네트워크가 필수적이다[7,9].

그러나 현재의 네트워크 구조는 주로 사람의 인터넷 이용과 콘텐츠 소비를 중심으로 설계되어, AI 학습에서 발생하는 대규모 집합통신과 피지컬 AI가 요구하는 실시간 제어라는 상이한 성능 조건을 동시에 지원하기 어렵다[4,5]. 센서 데이터의 증가와 실시간 처리 수요의 확대는 기존 구조의 이러한 부정합을 한층 심화시키고 있다.

4. 네트워크 운영 복잡성의 증가와 기존 운영 방식의 한계

앞선 세 가지 변화가 네트워크에 대한 수요 측면의 변화라면, 네트워크 내부에서는 운영 복잡성의 증가라는 공급 측면의 변화가 동시에 진행되고 있다. 기존 네트워크 운영은 성능 지표(KPI), 로그, 알람, 트래픽 통계 등을 사람이 분석하고 장애나 품질 저하가 발생한 이후에 조치하는 사후 대응 방식이 중심이었다. 그러나 5G 이후 네트워크의 가상화, 클라우드화, 분산화에 네트워크 슬라이싱과 엣지화가 결합되면서 관리해야 할 장비, 서비스, 경로, 자원 상태의 수가 급격히 늘어나, 모든 상황을 사람이 적시에 파악하고 대응하는 것은 구조적 임계에 봉착

하고 있다.

이러한 한계에 대한 대응으로 국제 표준화 기구들은 네트워크 운영 자체에 AI를 도입하는 방향을 공식화하고 있다. 유럽전기통신표준화기구(ETSI)는 펄프 자동화, AI 에이전트, 예측 분석, 이상 탐지 등을 자율 네트워크 구현의 핵심 기술로 제시하고 있으며[11], 이동통신 표준화 기구인 3GPP 역시 무선접속망(RAN)에 대한 AI/ML 적용 논의를 통해 데이터 수집, 모델 학습, 모델 추론 등을 표준화 대상으로 다루면서 AI 기반 네트워크 운영의 필요성을 제도권 표준 안에서 확인하고 있다[12]. 즉, 네트워크 운영의 지능화는 선택적 고도화가 아니라, 복잡성이 사람의 관리 역량을 넘어서는 환경에서 네트워크의 안정성을 담보하기 위한 현실적 당위로 자리 잡고 있다.

5. 소결: AI 네트워크로의 전환 필연성

이상에서 살펴본 네 가지 변화는 각기 독립된 현상이 아니라 궤를 같이하며 하나의 방향을 가리키고 있다. 수요 측면에서 보면, 트래픽은 양방향 상호작용과 기계 생성 데이터 중심으로 재편되고, 네트워크 성능은 AI 인프라의 연산 효율과 투자 효율을 직접 좌우하는 변수가 되었으며, 지연에 민감한 AI 연산은 네트워크 전역으로 분산되고 있다. 이는 네트워크가 데이터를 전달하는 수동적 통로에 머무를 수 없으며, AI의 데이터 흐름과 연산 배치를 이해하고 지원하는 능동적 실행 기반으로 재설계되어야 함을 의미한다. 공급 측면에서 보면, 이렇게 복잡해진 네트워크를 사람이 일일이 해석하고 조치하는 기존 운영 방식으로 감당하기는 어려워지고 있어, 네트워크 자체의 운영에도 AI의 도입이 불가피해지고 있다.

요컨대 AI 시대의 네트워크에는 ‘네트워크의 AI

실행 인프라화’와 ‘네트워크 운영의 지능화’라는 두 가지 요구가 동시에 제기되고 있다. 주목할 점은 두 요구가 상호 증폭의 관계에 있다는 것이다. AI 워크로드를 수용하기 위해 네트워크가 복잡해질수록 운영 지능화의 필요성은 커지고, 운영이 지능화될수록 네트워크는 더 정교하게 AI 워크로드를 지원할 수 있게 된다. 이 두 흐름을 하나의 구조로 통합하는 개념이 바로 다음 장에서 다루는 AI 네트워크이다.

III. AI 네트워크의 개념 및 주요 기술 동향

1. AI 네트워크 정의와 구조

네트워크와 AI의 결합은 그간 두 갈래의 흐름으로 발전해 왔다. 하나는 AI를 활용하여 네트워크 내부의 운영과 제어, 최적화를 지능화하려는 흐름으로, 트래픽 변화, 장애, 혼잡, 품질 저하, 자원 사용 문제 등을 AI가 스스로 인식하고 분석하여 판단함으로써 네트워크의 성능과 안정성, 효율성과 자율성을 제고하는 데 초점을 두며, 국제 표준화 기구의 자율 네트워크 논의가 이를 대표한다[11].

다른 하나는 AI의 학습과 추론, 서비스 실행 과정에서 발생하는 대규모 데이터 이동과 분산 연산을 효율적으로 뒷받침하기 위해 네트워크의 구조와 역할을 AI 워크로드 중심으로 재편하려는 흐름으로, 네트워크를 AI 실행 플랫폼으로 확장하려는 산업계의 논의와 통신-컴퓨팅 융합 구상이 이에 해당한다[9,13].

본고에서는 전자를 AI for Network, 후자를 Network for AI로 지칭하고, 이 두 흐름을 아우르는 통합 패러다임으로서 AI 네트워크를 정의한다. 이때 Network for AI의 본질은 단순히 네트워크의 속도를 높이는 데 있지 않고, AI 작업이 적절한 위치와 자원

표 1 AI 네트워크의 두 축: AI for Network, Network for AI

구분	AI for Network	Network for AI
개념	AI를 활용하여 네트워크의 운영과 제어, 최적화를 지능화	AI의 학습과 추론, 서비스 실행을 위해 네트워크의 구조와 역할을 AI 워크로드 중심으로 재구성
핵심 질문	AI로 네트워크를 어떻게 더 잘 운영할 것인가?	AI가 잘 동작하도록 네트워크가 무엇을 지원해야 하는가?
지향점	예측 기반 운영, 지능형 자동화, 자율 네트워크	손실없는 고신뢰 연결, 탄력적 자원 배분, AI 서비스 실행 플랫폼화
대표 영역	무선망과 코어망의 운영 지능화, 장애 예측, 에너지 최적화	AI 데이터센터 네트워크, 엣지 AI 인프라, 데이터센터 간 협력 학습망

출처 저자 작성

에서 수행되도록 네트워크가 데이터의 이동과 연산의 배치, 자원의 배분에 이르는 전체 흐름을 조정하는 데 있다. 요컨대 전자가 ‘네트워크를 더 잘 운영하기 위한 AI’라면, 후자는 ‘AI가 더 잘 동작하도록 지원하는 네트워크’라 할 수 있다(표 1 참고).

두 축은 출발점이 다르지만 상호 보완적으로 작동한다. AI for Network만으로는 AI 학습과 추론이 요구하는 인프라 변화를 설명하기 어렵고, Network for AI만으로는 복잡해진 네트워크 자체를 어떻게 운영하고 제어할 것인지를 설명하기 어렵다. 따라서 AI 네트워크는 두 흐름의 단순한 병렬이 아니라, 네트워크의 내부 지능화와 AI 실행 인프라화가 하나의 네트워크 구조 안에서 동시에 진행되는 구조적 전환으로 이해할 수 있다.

그 핵심적 의미는 네트워크를 단순 연결 인프라가 아니라 AI 시대의 지능형 실행 인프라로 재정의하는 데 있으며, 이 과정에서 통신사업자의 역할도 단순 연결 제공자에서 AI 인프라 및 서비스 제공자로 확장될 수 있는 기반이 마련된다[4,9]. 다음 절에서는 두 축과 공통 기반을 구성하는 주요 기술 동향을 살펴본다.

2. AI for Network 주요 기술 동향

AI for Network 기술의 전체적인 진화 방향은 네트워크 운영이 사후 대응에서 예측과 자동화를 거쳐 자율 운영으로 나아가는 과정으로 요약할 수 있다. 그 기본 골격이 되는 것이 폐루프(Closed-Loop) 기반 자율 네트워크 제어이다. 이는 네트워크 상태를 지속적으로 수집하고 분석한 뒤 AI가 운영 정책을 생성하여 실행하고, 그 결과를 다시 학습에 반영하는 순환형 제어 구조로, 운영을 일회성 조치가 아니라 관찰-분석-결정-실행이 반복되는 지속적 최적화 구조로 전환한다[11].

주목할 점은 개방형 무선망 표준화 기구(O-RAN Alliance)의 장주기 정책 계층과 근실시간 제어 계층이 결합된 이중 폐루프, 3GPP의 데이터 수집-학습-추론-실행 구조, ETSI의 종단 간 무접촉 자율 운영(ZSM: Zero-touch network and Service Management)이 명칭과 계층 구분은 다르지만 동일한 순환 모델로 수렴하고 있다는 것으로[12,14,15], 폐루프는 특정 진영의 설계가 아니라 산업 공통의 참조 운영 모델로 정착하고 있다.

아울러 폐루프 제어는 운영 권한을 사람에서 시스템으로 일시에 이양하는 것이 아니라, AI의 제안을 사람이 승인하는 단계에서 조건부 자동 실행을 거쳐 완전 자율로 신뢰를 축적하며 단계적으로 자율성 수준을 높여가는 운영 프레임워크로 이해되고 있다.

폐루프 구조가 무선망 영역에서 구체화된 대표 사례가 RIC(RAN Intelligent Controller) 기반의 지능형 RAN 제어이다. 기존 RAN은 제조사별 전용 하드웨어와 폐쇄형 인터페이스로 구성되어 외부 AI가 무선망에 개입할 통로가 없었으나, O-RAN 얼라이언스(O-RAN Alliance)는 RAN 기능을 개방형 인터페이스로 분리하고 지능형 제어를 표준화함으로써 AI

가 무선망을 분석하고 제어할 수 있는 개방형 제어 계층을 마련하였다[14].

RIC는 장주기 정책 최적화와 AI/ML 모델의 학습과 관리를 담당하는 비실시간(Non-RT) 제어부와 무선망 상태를 근실시간으로 분석하여 무선 자원 제어, 핸드오버 최적화, 간섭 완화, 에너지 절감을 수행하는 준실시간(Near-RT) 제어부로 구성되며, 표준화된 개방형 인터페이스를 통해 서로 다른 제조사의 장비와 응용(rApp, xApp)이 연동된다[14].

이러한 개방형 구조는 성능 개선을 넘어 운영자가 원하는 제어 응용을 선택하고 조합할 수 있게 함으로써 특정 장비 제조사에 대한 종속을 완화한다는 전략적 함의를 지니며, 이 점이 주요국의 개방형 무선망 정책과 맞물려 표준 채택의 동력이 되고 있다. 다만, 여러 제어 응용이 동일한 무선 자원을 상반된 방향으로 제어할 때 발생하는 정책 충돌의 조정과 다변된 환경에서의 상호운용성 확보는 상용화 과정에서 풀어야 할 난제로 남아 있다.

AI의 적용 계층은 운영과 관리의 지원을 넘어 통신 기능의 가장 밑단인 무선 인터페이스와 물리계층으로 하향 확장되고 있다. 3GPP는 무선 인터페이스에 AI/ML을 적용한 최초의 연구(TR 38.843)를 통해 채널 상태 정보 압축, 빔 관리, 측위의 세 가지 활용 사례에 대한 성능 이득과 모델 운영 요건을 정리하였고[16], 후속 표준(Relase 19)에서는 AI 모델을 실제로 활성화와 비활성화, 갱신을 위한 신호 절차를 규범화하는 단계로 진전되었다[12].

이는 사람이 설계한 알고리즘을 AI가 보완하는 수준을 넘어, 통신 시스템을 처음부터 AI를 전제로 설계하는 AI 내재형 구조로의 전환을 의미하며, 6G 표준화에서 핵심 의제로 부상할 가능성이 크다. 물리계층은 통신 성능의 출발점이므로 이 계층에서 얻는 전송 효율의 개선은 상위 계층의 모든 서비스 품질로 파급된다는 점에서, 적용 난도는 높지만 성

공 시 효과가 네트워크 전반으로 확산되는 고부가 영역이라 할 수 있다. 한편, 채널 정보 압축처럼 송신단과 수신단의 모델이 한 쌍으로 함께 학습되고 동작해야 하는 기능에서는 서로 다른 주체의 모델 간 정합을 관리해야 한다는 물리계층 고유의 과제가 제기되고 있다.

자율 운영의 수준이 높아질수록 그 신뢰성을 담보하는 기술의 중요성도 함께 커진다. 네트워크 디지털 트윈(Digital Twin)은 실제 네트워크를 가상 공간에 복제하여 AI 정책이나 구성 변경을 실제 망에 적용하기 전에 그 영향을 사전 검증하는 기술로, AI의 잘못된 판단이 서비스 장애로 번지는 것을 막는 안전장치 역할을 한다[11]. 디지털 트윈은 일회성 모의실험과 달리 실제 망의 상태 변화를 지속적으로 반영하며 운영과 나란히 작동하는 상시 검증 기반이라는 점에서 기존 시뮬레이터와 구별되며, ETSI는 이를 자율 네트워크 진화의 핵심 지원 기술로, 국제전기통신연합(ITU-T)은 디지털 트윈 네트워크의 요건과 구조를 표준(Y.3090)으로 정의하고 있다[11,17].

자동 실행의 범위가 넓어질수록 잘못된 정책의 파급력도 커지므로, 사전 검증 체계는 자율성 상향의 전제 조건이 된다. 이와 함께 운영의 실행 주체로서 AI 에이전트와 인텐트(Intent) 기반 자동화가 표준화의 새로운 흐름으로 부상하고 있다. AI 에이전트는 장애 진단, 구성 변경, 보안 대응 등 특정 운영 업무를 이해하고 조치를 제안하고 수행하는 지능형 소프트웨어이며, 인텐트 기반 제어는 운영자가 세부 명령 대신 품질, 지연, 에너지와 같은 목표만 제시하면 네트워크가 이를 해석하여 자동으로 구성하고 최적화하는 방식으로, 관리의 추상화 수준을 장비와 명령 중심에서 목표 중심으로 끌어올리는 전환에 해당한다[9,15]. ETSI의 종단 간 무접촉 자율 운영은 여러 에이전트가 무선망, 코어망, 전송망 등

영역을 나누어 협업하는 멀티에이전트 구조와 인텐트 기반 페루프를 도입하는 방향으로 확장되고 있으며[15], 이에 따라 표준화의 초점도 개별 자동화에서 자동화 주체 간의 협력과 충돌 조정으로 이동하고 있다.

3. Network for AI 주요 기술 동향

Network for AI 기술의 진화는 데이터센터 내부에서 시작되어 네트워크 전역으로 확장되는 흐름을 보인다. 그 출발점은 대규모 AI 학습과 추론을 위해 다수의 가속기와 서버를 지연과 손실 없이 초고속으로 연결하는 데이터센터 네트워크 패브릭이다. 대규모 언어모델의 학습은 수천에서 수만 개의 GPU가 집합통신으로 동기화되는 병렬 연산 구조이므로, 분산 학습은 동기화 지점에서 가장 늦게 도착하는 데이터 흐름을 기다리게 되며, 상위 백분위 지연(Tail Latency)과 집합통신 완료 시간이 전체 작업 완료 시간을 결정한다[6,8].

따라서 평균 대역폭의 확보만큼이나 혼잡 제어, 부하 분산, 흐름 스케줄링이 학습 효율을 좌우하는 핵심 기능이 되며, 작은 데이터 손실도 학습 시간 증가와 자원 활용률 저하로 이어질 수 있어 무손실에 가까운 전송 품질이 요구된다[4]. 이러한 수요는 그간 특정 벤더 중심의 전용 인터커넥트가 충족해 왔으나, 비용과 공급망, 생태계 종속의 문제가 부각되면서 개방형 표준화 요구가 급증하였다.

'23년 리눅스 재단(Linux Foundation) 산하에 설립된 UEC(Ultra Ethernet Consortium)는 '25년 UEC 1.0 사양을 공개하여 패킷 분산 기반의 다중경로 전송, 유연한 패킷 순서 처리, 송수신 협력형 지능형 혼잡 제어 등 AI 및 고성능컴퓨팅 워크로드에 최적화된 개방형 이더넷 전송 계층을 정의하였고[18], 별도로 결성된 UALink 컨소시엄은 노드와 랙 단위에서 가

속기를 직접 연결하는 스케일업(Scale-Up) 인터넥트의 개방형 표준(레인당 200Gbps, 최대 1,024개 가속기 연결)을 '25년 공개하였다[19].

이와 함께 IEEE 802.3에서는 800GbE 표준화 완료에 이어 1.6Tb/s급 표준화가 진행 중이다[20]. 이러한 흐름은 AI 패브릭 기술이 성능 경쟁을 넘어 생태계 주도권과 종속성 회피의 관점에서 다루어지고 있으며, 폐쇄형 전용 인터넥트와 개방형 이더넷 진영 간 경쟁이 기술 진화의 향배를 가를 변수로 대두하였음을 보여준다.

데이터센터 밖에서는 네트워크의 역할이 데이터 전달을 넘어 AI 실행의 조율로 확장되고 있다. AI 서비스가 단말-엣지-클라우드-데이터센터에 걸쳐 분산 실행됨에 따라, 어떤 작업을 어디에서 수행하고 데이터를 어떤 경로로 옮기며 어떤 통신 자원을 배정할지를 서비스 요구조건에 따라 실시간으로 조율하는 AI 워크로드 오케스트레이션이 서비스 성능을 좌우하는 조율 계층으로 부상하였다[4,9].

이 기술의 본질은 그동안 따로 다루어지던 연산 배치와 네트워크 자원 제어를 하나의 관점에서 함께 최적화하는 데 있다. 연산 위치와 연결 경로를 분리하여 결정하면 전체 효율이 떨어지기 때문이다. 특히 금융, 의료, 공공 등 민감 데이터를 다루는 분야에서는 데이터를 로컬에 두고 네트워크를 통해 원격의 대규모 연산 자원만 활용하는 구조가 중요해지고 있는데, 이는 데이터 주권과 보안을 유지하면서 AI를 활용할 수 있게 한다는 점에서 규제 산업으로의 AI 확산을 좌우하는 조율 기술로 평가된다[4]. 또한, AI 워크로드는 수요 변동이 빠르므로 변동이 발생한 뒤 대응하는 방식이 아니라 수요 패턴을 학습하여 자원을 선제적으로 배치하는 예측형 조율이 요구된다는 점도 기존 자원 관리와의 차이점이다.

무선망 영역에서는 두 축이 결합되는 점점으로서

AI-RAN 기반의 통신-컴퓨팅 융합 인프라가 부상하고 있다. AI-RAN 얼라이언스(AI-RAN Alliance)는 이를 세 범주로 구분하는데, AI로 RAN의 성능과 운영 효율을 개선하는 단계(AI-for-RAN), 통신 기능과 AI 연산을 GPU 등 가속 인프라 기반의 공유 인프라에서 함께 수행하는 단계(AI-and-RAN), 그리고 기지국 인프라의 연산 자원을 활용하여 저지연 AI 추론 서비스를 직접 제공하는 단계(AI-on-RAN)로 이어지는 확장 경로이다[13,21].

이 경로의 종착점에서 기지국은 통신 장비인 동시에 엣지 AI 서비스의 실행 플랫폼이 되며, 통신사업자는 연결 제공자를 넘어 AI 인프라 및 서비스 제공자로 역할을 확장하게 된다. AI-RAN은 표준화와 정책, 산업 실증이 동시에 집중되고 있는 영역으로, 일본과 미국 등에서 통신사업자와 AI 반도체 및 장비 기업 간 협력을 통한 상용 실증이 빠르게 진행되고 있다[21]. 다만, 통신 기능은 지연과 안정성 요구가 엄격하므로 AI 워크로드가 통신 품질을 침범하지 않도록 두 워크로드 간 자원을 격리하고 지연을 보장하는 기술이 융합 인프라 성패의 관건으로 지적된다.

한편, AI 인프라의 광역 확장 흐름 속에서 복수의 데이터센터를 하나의 학습 환경처럼 연결하는 데이터센터 간 협력 학습망과 AI 데이터센터(AIDC: AI Data Center)를 기업과 산업 현장, 엣지에 있는 접속망도 주요 기술로 부상하고 있다[4]. 장거리 구간에서는 물리적 지연으로 인해 촘촘한 동기식 학습이 어려워 통신량을 줄이거나 비동기로 학습하는 방식이 전제되어야 하므로, 단순한 고속 연결을 넘어 네트워크 설계와 학습 알고리즘을 함께 맞추는 공동 설계가 요구된다. 또한, AI 데이터센터를 구축하더라도 고객과 산업 현장이 안정적으로 접근하지 못하면 그 활용성과 수익화가 제한되므로, 접속망의 품질은 AI 인프라의 산업적 활용도를 가늠하는 시

금석이 된다. 이에 따라 데이터센터 내부망, 데이터 센터 간 연결망, 접속망을 분리하지 않고 하나의 연속된 종단 간 구조로 통합 설계해야 한다는 관점이 주요 방향으로 제시되고 있다[4].

4. 공통 기반 기술 동향

두 축의 기술이 실제 상용망에 적용되기 위한 공통의 전제는 AI 모델의 안정적 운영 체계이다. AI 네트워크에서는 모델이 네트워크의 운영과 제어에 직접 관여하므로 모델의 오류나 성능 저하가 곧바로 서비스 장애로 이어질 수 있으며, 따라서 모델의 식별, 학습, 검증, 배포, 추론, 모니터링, 롤백(이전 모델 복귀), 재학습, 버전 및 책임 관리에 이르는 전 과정을 체계화하는 AI/ML 모델 생명주기 관리(LCM: Lifecycle Management)와 신뢰성 검증이 알고리즘 성능 이상으로 중요한 상용화의 관건이 된다. 3GPP는 학습부터 검증, 배포, 모니터링, 재학습까지 아우르는 페루프 관리 프레임워크를 기술보고서(TR 28.908)로 도입하였고, O-RAN은 모델에 고유 식별자와 버전을 부여하여 측정 데이터에 따라 주기적으로 갱신하는 운영 구조를 규정하는 등 주요 표준화 기구가 이를 공통 과제로 다루고 있다[14,22]. 모델은 한 번 배포하면 완결되는 소프트웨어와 달리 데이터와 네트워크 환경이 변하면 성능이 점진적으로 저하되는 특성을 가지므로, 배포 자체보다 배포 이후의 성능 추적과 신속한 복구 체계가 신뢰성을 좌우한다는 점은 실무적으로 각별한 함의를 지닌다.

개별 모델의 관리를 넘어, 다수의 공급자와 운영자가 모델을 함께 개발하고 등록하며 공유와 검증을 거쳐 배포하는 공동 플랫폼과 생태계를 구축하려는 표준화도 진행되고 있다. ITU-T는 데이터 수집부터 학습과 배포, 추론에 이르는 ML 파이프라인의 참조 구조(Y.3172)를 시작으로, 모델을 거래하는

마켓플레이스(Y.3176), 모델 제공 구조(Y.3179), 그리고 배포 전 격리된 시험 환경인 샌드박스(Y.3181)에 이르는 일련의 권고를 체계화하였다[23,24]. 특히 샌드박스는 외부에서 도입된 모델을 곧바로 운영망에 투입하지 않고 분리된 환경에서 안전성과 성능을 검증한 뒤 배포하도록 하는 장치로, 개방형 생태계의 확장과 운영 안전성을 동시에 확보하는 핵심 요소이다. 모델을 공급하고 검증하여 배포하는 공동 체계가 없으면 AI 네트워크는 개별 프로젝트 수준의 실험에 머물기 쉬우므로, 이러한 운영 생태계 인프라는 실험적 적용을 지속 가능한 산업적 운영 체계로 전환하는 기반이 된다. 아울러 현재의 표준화가 무선망 영역에 상대적으로 집중되어 있는 가운데, 서비스별 세션과 경로, 품질을 실시간으로 제어하는 코어망에 AI를 내재화하는 기술은 무선망 중심의 지능화를 종단 간 자율화로 확장하는 연결 고리에 해당한다. 실제 종단 간 서비스 품질은 무선망뿐만 아니라 코어망과 전송망, 엣지의 통합 제어로 결정되므로 코어망의 지능화 없이는 종단 간 자율화가 완성되기 어려우며, 이 영역은 표준 경쟁이 상대적으로 초기 단계에 있어 선제적 기술 확보의 여지가 크다는 점에서 전략적 가치가 주목된다.

5. 종합

이상에서 살펴본 AI 네트워크 주요 기술을 종합하면 표 2와 같다. 기술군별 추진 양상을 비교하면 흥미로운 비대칭이 관찰된다. AI for Network는 O-RAN, 3GPP, ETSI 등 표준화 기구를 중심으로 규격화가 앞서 진행되고 있는 반면, Network for AI는 데이터센터 패브릭과 AI-RAN에서 보듯 벤더와 사업자 주도의 산업 실증과 상용화가 표준화를 앞서는 양상을 보인다. 이는 표준화 경쟁의 승부처가 규격 문서 자체보다 실제 망 환경에서 검증된 실증

표 2 AI 네트워크 기술군별 진화 방향 및 추진 동향

기술군	진화 방향	주요 기술	추진 동향
AI for Network	사후 대응 운영 → 페루프 자동화 → 자율 네트워크	페루프 기반 자율 네트워크 제어, RIC 기반 지능형 RAN 제어, AI 내재형 무선 인터페이스와 물리계층, 네트워크 디지털 트윈 기반 검증, AI 에이전트와 인텔리전트 기반 자동화	O-RAN, 3GPP, ETSI 중심으로 표준 규격화 선행, 페루프가 공통 참조 운영 모델로 수렴
Network for AI	단순 연결망 → 연산 통합 기반 → AI 실행 플랫폼	초고속 무손실 데이터센터 네트워크 패브릭, AI 워크로드 오케스트레이션, AI-RAN 기반 통신-컴퓨팅 융합 인프라, 데이터센터 간 협력 학습망과 AIDC 접속망	벤더와 사업자 주도의 산업 실증과 상용화가 표준화에 선행, 개방형 패브릭 표준화(UEC, UALink 등) 본격화
공통 기반	실험적 적용 → 산업적 운영 체계	AI/ML 모델 생명주기 관리(LCM)와 신뢰성 검증, ML 운영 생태계 인프라(마켓플레이스, 서빙, 샌드박스), AI 내재형 코어 네트워크	3GPP, O-RAN, ITU-T 등이 공통 과제로 표준화 진행, 코어망 AI는 표준 형성 초기로 선제 확보 여지 존재

출처: 저자 작성

데이터와 레퍼런스 구현으로 이동하고 있음을 시사하며, 기술 개발과 실증, 표준화를 분리된 활동이 아니라 하나의 연동된 체계로 추진해야 할 필요성을 보여준다. 아울러 무선망 제어 계층, 종단 간 자율 운영 계층, 컴퓨팅 인프라 계층으로 분화되어 있던 표준화 전선이 6G의 AI 내재화 논의를 매개로 다시 수렴하는 구조를 보이고 있어, 특정 기구 단위의 대응이 아니라 표준화 기구를 가로지르는 종합적 대응 전략이 긴요한 시점이라 할 수 있다.

IV. AI 네트워크 도입 효과 분석

이 장에서는 AI 네트워크의 도입 효과를 통신사업자의 비용 절감과 신규 시장 창출의 양 측면에서 정량적으로 분석한다. 비용 절감 효과는 통신사업자의 전체 운영비용(TCO: Total Cost of Ownership)을 항목별로 분해하고 AI 적용 영역별 절감률을 가중합산하는 상향식(Bottom-Up)으로, 시장 창출 효과는 시장 범위를 통신사업자 관점의 인프라·장비로 한정된 뒤 통신산업 CAPEX 구조에 기반한 하향식(Top-Down) 검증을 병행하는 방식으로 추정하였다. 주요 가정은 AI-RAN 및 Open RAN 관련 테스트베드 결과와 글로벌 시장조사기관 분석, 전문가 의견

을 종합하여 설정하였다.

1. 망 운용 효율화에 따른 비용 절감 효과

1.1 비용 구조와 AI 적용 영역

통신사업자의 총비용은 네트워크 비용, IT·플랫폼 비용, 판매·마케팅 비용, 고객 서비스 비용, 일반관리비 등으로 구성되며, 이 가운데 네트워크의 구축과 운영에 소요되는 비용이 전체의 약 35~45%로 가장 큰 비중을 차지한다[25,26]. 네트워크 비용의 내부에서는 RAN과 모바일 코어 등 무선 접속 인프라 관련 비용이 40%로 가장 크며, 전송망 장비(20%)와 에너지(15%)가 그 뒤를 잇는다. 에너지 비용이 운영인력이나 유지보수보다 큰 비중을 차지한다는 점은 기지국 전력 최적화가 환경적 과제인 동시에 비용 구조상의 실질적 개선 지점임을 시사한다.

AI for Network는 특정 항목에 국한되지 않고 장비와 인력, 유지보수, 에너지 전반에 서로 다른 메커니즘으로 작용한다. Ⅲ장에서 살펴본 기술을 비용 절감의 관점에서 재구성하면 네 가지 영역으로 정리된다. Open RAN·Cloud RAN 기반 가상화는 전용 하드웨어 의존도를 낮추고 범용 서버 기반의 자원 통합(Pooling)을 통해 장비 구축 효율을 높이

며, RIC과 펌프 제어에 기반한 AI 기반 네트워크 자동화는 트래픽 분석과 자기 최적화(Self-Optimization)를 통해 운영 인력의 부담을 완화한다[14]. AI 기반 예지 보전은 이상 탐지와 장애 예측을 통해 사후 복구를 사전 정비로 전환하여 유지보수 비용을 절감하고[11], AI 기반 에너지 최적화는 트래픽 부하에 연동한 전력 제어와 슬립 모드(Sleep Mode) 운영을 통해 장비의 전력 사용량을 줄인다.

1.2 비용 절감 효과 추정

각 영역의 적용 절감률은 Open RAN TCO 분석[27]과 네트워크 에너지 절감 실증 결과[28], AI-RAN 테스트베드 결과[21] 및 전문가 의견을 종합하여 설정하였다. 이를 네트워크 비용 구조에 항목별로 적용하여 가중합산한 결과, AI for Network 도입에 따른 네트워크 총비용의 절감 효과는 약 17.75%로 추정되었다(표 3 참고). 여기에 네트워크 비용이 전체 비용의 약 40%를 차지한다는 점을 반영하면, 통신사업자 TCO 기준의 절감 효과는 약 7.1%(17.75% × 40%)로 환산된다.

절감 효과의 내부 구성에는 주목할 만한 특징이 나타난다. 전체 절감분 17.75%p 가운데 10.00%p, 즉 절반 이상이 RAN·모바일 코어 장비의 가상화라는 단일 항목에서 발생한다. 이는 AI for Network

의 비용 효과가 운영 자동화에 따른 인력 절감보다 무선 접속망 인프라의 구조적 전환에 크게 의존함을 의미하며, Open RAN 및 Cloud RAN의 실제 절감 수준이 추정치 전반의 신뢰도를 좌우하는 핵심 변수임을 시사한다.

2. AI-native 네트워크 전환에 따른 신규 시장 창출 효과

2.1 CAPEX 구조의 변화

AI for Network가 비용을 절감하는 방향으로 작용한다면, Network for AI를 포함한 AI-native 네트워크로의 전환은 새로운 투자를 유발하는 방향으로 작용한다. 개방형 구조 도입에 따른 장비 공급 경쟁의 확대와 범용 서버(COTS: Commercial Off-the-Shelf) 기반 가상화, RAN 자원 통합은 CAPEX를 개선하는 요인으로 작용하는 반면, GPU·DPU·NPU 등 AI 가속기 기반 서버와 엣지 데이터센터, 고성능 스위치 및 NIC, AI 오케스트레이션 플랫폼에 대한 투자는 CAPEX를 증가시키는 요인으로 작용한다.

결과적으로 AI-native 전환은 CAPEX의 단순한 증감으로 환원되지 않으며, 그 실질은 통신 장비 중심의 투자 구조가 AI 컴퓨팅과 데이터 인프라 중심의 구조로 재편되는 과정에 있다. 이러한 재편은 통

표 3 AI for Network 도입에 따른 네트워크 비용 절감 효과 추정

비용 항목	비용 비중	주요 AI 적용 영역	적용 절감률	절감 효과
RAN·모바일 코어 장비	40%	Open·Cloud RAN 기반 가상화	25%	10%
전송(백본)망 장비	20%	AI 기반 네트워크 자동화	10%	2%
에너지	15%	AI 기반 에너지 최적화	15%	2.25%
운영인력	10%	AI 기반 네트워크 자동화	20%	2%
유지보수	10%	AI 기반 예지 보전	15%	1.50%
사이트 임대	5%	-	-	-
합계	100%	-	-	17.75%

출처 참고문헌 [21,25-28] 및 전문가 의견을 토대로 저자 추정

신사업자에게는 투자 대상의 전환을, 산업에는 신규 시장의 형성을 의미한다. 이하에서는 이를 통신 특화 엣지 인프라 시장과 AI 네트워크 장비 시장의 두 축으로 나누어 전망한다.

2.2 통신 특화 엣지 인프라 시장

미래 통신사업자의 AI 컴퓨팅 장비 시장은 네트워크 CAPEX에 AI 도입 비중과 AI 하드웨어 구성 비율을 곱하는 하향식으로 추정되어 왔으나, AI 도입의 판단 기준이 모호하고 하드웨어 구성 비율의 사업자별 편차가 커 적용 기준에 따라 규모의 변동이 크게 나타나는 한계를 지닌다. 실제로 주요 시장조사기관들은 엣지 컴퓨팅 시장을 MEC 서버 중심의 협의적 정의에서 AI 엣지 서비스와 애플리케이션까지 포함하는 광의적 정의에 이르기까지 각기 다르게 규정하고 있어, 시장 규모 추정치 사이에 수십 배의 격차가 존재한다[29-32]. 이는 기관별 전망치의 차이가 성장 속도에 대한 견해차라기보다 시장 경계 설정의 차이에서 비롯됨을 보여준다.

이에 본고에서는 시장 범위를 ‘통신 특화 엣지(Telco Edge) 인프라’로 한정하여, AI 엣지 애플리케이션과 플랫폼 서비스를 제외하고 AI-RAN과 MEC, 엣지 클라우드의 확산에 따라 통신사업자가 신규로 구축하는 엣지 서버·스토리지·고성능 NIC·마이크로 데이터센터 등 AI-native 인프라만을 대상으로 설정하였다. 이 기준을 적용한 ’30년 글로벌 통신사업자 엣지 인프라 시장 규모

는 80~120억 달러 수준으로, 본고에서는 중앙값인 100억 달러를 기준값으로 채택하였으며, ’37년에는 299억 달러 규모로 확대되어 연평균 16.9%의 성장이 예상된다(표 4 참고). 이 가운데 GPU 서버와 DPU/NPU 등 AI 컴퓨팅 장비가 55%로 과반을 차지하여, 엣지 AI 컴퓨팅 장비 시장은 ’30년 55억 달러에서 ’37년 164억 달러 규모로 전망된다.

추정치의 타당성은 통신산업 CAPEX 구조를 기반으로 교차 검증하였다. 글로벌 통신사업자의 연간 CAPEX 약 3,200억 달러[25,33] 가운데 RAN CAPEX 비중 30~35%를 적용하면 약 1,000억 달러 규모이며[33], 여기에 ’30년 AI-RAN 도입 비중 30~40%와 엣지 인프라 투자 비중 20~30%를 순차 적용하면 시장 규모는 60~120억 달러로 산출된다. 기준값 100억 달러가 이 범위 내에 위치한다는 점에서 전망치는 타당한 수준으로 판단된다.

2.3 AI 네트워크 장비 시장

AI 네트워크 기술은 ① 네트워크용 AI 모델, ② AI 기반 무선 접속망(AI-RAN), ③ AI 기반 자율 네트워크(모바일 코어·전달/백본망)의 세 영역으로 구분되나, 본고의 시장 전망에서는 기반 기술에 해당하는 네트워크용 AI 모델을 제외하고 인프라·장비 시장을 대상으로 ‘AI-RAN 장비 시장’과 ‘AI 자율 네트워크 장비 시장’의 두 시장으로 구분하여 추정하였다(그림 1).

추정의 기본 논리는 기존 장비 시장을 모(母)시장

표 4 통신사업자 엣지 인프라 및 엣지 AI 컴퓨팅 시장 전망

(단위: 십억 달러)

구분	'30	'31	'32	'33	'34	'35	'36	'37	CAGR
통신사업자 엣지 인프라	10	11.7	13.7	16	18.7	21.9	25.6	29.9	16.90%
통신사업자 엣지 AI 컴퓨팅 장비	5.5	6.4	7.5	8.8	10.3	12	14.1	16.4	16.90%

출처 참고문헌 [29-32]를 토대로 저자 추정



출처 저자 작성

그림 1 AI 네트워크 장비 시장의 세분 구조

으로 두고 AI 기능 내재화에 따른 단가 상승분을 적용하는 데 있다(표 5). AI-RAN 장비는 Radio Unit에 AI 가속기를 포함한 서버와 RAN L1 및 L2/L3 소프트웨어, 가상화 플랫폼, 오케스트레이터가 결합된 구조로 재편되어 기존 RAN 장비 대비 약 34%의 비용 증가가 발생하며[21], 모바일 코어는 35%, 전달·백본망은 12.5%가 각각 적용된다. 전달·백본 영역의 증가율이 상대적으로 낮은 것은 이 영역의 AI 내재화가 고가의 AI 가속기 도입보다는 기존 장비 내부의 처리 기능 강화 형태로 진행되기 때문이다.

AI 네트워크로의 전환 비중은 6G 전환에 따른 신규 AI-native 구조의 적용 범위를 기준으로 하는 협의의 접근과 기존 5G 및 5G-Advanced 네트워크에 AI 스케줄러와 RIC, AI 자동화 기능이 단계적으로 내재화되는 흐름까지 포함하는 광의의 접근이 가능

하다. 표장에서 확인한 바와 같이 AI 네트워크 기술은 6G 상용화 이후에 비로소 등장하는 것이 아니라 이미 5G-Advanced 단계에서 선행 확산되고 있으므로, 본고에서는 기존 RAN의 AI 기능 고도화와 6G AI-native 인프라 전환이 병행될 가능성이 높다고 보고 두 기준을 절충한 하이브리드 기준을 적용하였다. 이에 따라 전환 비중은 '30년 15%, '33년 45%, '36년 75% 수준까지 확대되는 것으로 도출된다. 다만, 실제 확산 속도는 AI 가속기의 도입 수준, AI-RAN 얼라이언스를 중심으로 한 생태계의 확산 정도, 하이퍼스케일러와 통신사업자 간 협력 구조, 기존 RAN의 교체 주기 등에 따라 달라질 수 있다.

이상의 추정 논리와 전환 비중을 적용한 결과, AI-RAN 장비와 AI 자율 네트워크 장비로 구성되는 세계 AI 네트워크 장비 시장은 '30년 138억 달러에서 '37년 956억 달러 규모로 확대되어 연평균 31.9%의 성장이 전망된다(표 6).

전망 결과에서 눈여겨볼 대목은 두 가지이다. 첫째, '30년 기준 AI 자율 네트워크 장비 시장(72억 달러)이 AI-RAN 장비 시장(66억 달러)을 상회하며, 그 내부에서 전달·백본 영역이 모바일 코어의 4배를 넘어선다는 점이다. 전달·백본은 비용 증가율이 12.5%로 세 영역 중 가장 낮음에도 모시장의 규모 자체가 크기 때문이다. 기술적 논의와 표준화가 무선폭장에 집중되어 있는 것과 달리 시장 규모의 관점에서는 전달·백본 영역의 비중이 상당하다는 사실

표 5 AI 네트워크 장비 시장의 구성과 비용 증가율

영역	기존 장비 (모시장)	AI 네트워크 장비	비용 증가
RAN	RU+BBU(이동통신 RAN)	AI-RAN (AI 가속 서버 기반 DU/CU)	34%
모바일 코어	EPC/5GC(이동통신 코어 장비)	AI-Core	35%
전달망(백본망)	Router/Optical (백본장비: 스위치/라우터, 광 전송장비)	AI-Transport	12.50%

출처 참고문헌 [21]을 토대로 저자 작성

표 6 세계 AI 네트워크 장비 시장 전망

(단위: 백만 달러)

구분		'30	'31	'32	'33	'34	'35	'36	'37	CAGR
AI-RAN 장비		6,583	11,903	18,165	24,756	31,165	37,199	43,137	45,875	32.00%
AI 자율 네트워크 장비	모바일 코어	1,394	2,380	3,408	4,478	5,588	6,736	7,913	8,575	29.60%
	전달·백본	5,820	10,165	14,871	19,943	25,374	31,156	37,280	41,157	32.20%
	소계	7,214	12,545	18,279	24,421	30,962	37,893	45,192	49,732	31.80%
합계		13,796	24,448	36,444	49,176	62,126	75,092	88,329	95,606	31.90%

출처 참고문헌 [21,33]을 토대로 저자 추정

은, Ⅲ장에서 지적한 코어망 및 전송망 지능화의 전략적 가치를 시장 측면에서도 뒷받침한다. 둘째, 성장률이 '36년 이후 뚜렷하게 둔화되어 시장이 '30년대 중반까지의 고성장 국면과 이후의 성숙 국면으로 구분된다는 점으로, 시장 진입의 시점이 사업 전략상 중요한 변수가 됨을 의미한다.

3. 종합 및 분석의 한계

이상을 종합하면, AI 네트워크의 도입 효과는 비용 절감과 시장 창출의 두 방향으로 동시에 나타난다. AI for Network 기반의 자동화와 최적화는 통신사업자의 네트워크 비용을 약 17.75%, TCO를 약 7.1% 절감할 수 있는 것으로 추정된다. 동시에 AI-native 전환은 신규 인프라 투자를 유발하여, '30년 기준으로 통신 특화 엣지 인프라 시장은 100억 달러, AI 네트워크 장비 시장은 138억 달러 규모를 형성할 것으로 전망된다. 통신사업자의 관점에서 이는 운영 효율화를 통해 확보한 여력이 신규 인프라 투자의 기반으로 이어지는 전환 경로를 시사한다.

다만, 본 분석은 초기 시장을 대상으로 하는 만큼 몇 가지 한계를 내포한다. 절감률과 전환 비중 등 핵심 가정에 대한 민감도가 커, 가령 RAN 가상화 절감률이 25%에서 15%로 조정될 경우 TCO 절감률은 약 5.5%로 하락한다. 또한, 비용 절감과 신규 투

자의 순효과는 별도로 산정하지 않았으므로 통신사업자가 체감하는 순 재무 효과는 전환의 속도와 투자 회수 기간에 따라 달라지며, 시장 전망 역시 범위 설정에 크게 의존하므로 타 기관 전망치와의 비교 시에는 범위의 정합성을 함께 검토할 필요가 있다. 아울러 통신 특화 엣지 인프라 시장은 네트워크 장비 시장에 새롭게 추가되는 장비 영역을, AI 네트워크 장비 시장은 기존 장비의 AI 내재화에 따른 증분을 각각 대상으로 하여 산정 기반이 구분되나, AI-and-RAN과 AI-on-RAN이 동일한 가속 인프라에서 구현될 경우 양자의 귀속이 모호해질 수 있다는 점은 향후 보완이 필요한 부분이다.

V. 결론 및 시사점

1. 결론

본고에서는 AI 시대 네트워크 패러다임의 변화를 트래픽 구조, 인프라 형태, 연산 배치, 운영 복잡성의 네 가지 측면에서 분석하고, 이로부터 AI 네트워크로의 전환이 갖는 필연성을 논증하였다. AI 트래픽은 양방향 상호작용과 기계 생성 데이터 중심으로 재편되고 있고, 네트워크 성능은 AI 인프라의 투자 효율을 직접 좌우하는 변수가 되었으며, 자연에 민감한 AI 연산은 디바이스-엣지-클라우드 전반으로 분산되고 있다. 동시에 네트워크 자체의 복잡성

이 사람의 관리 역량을 넘어서면서 운영의 지능화가 불가피해지고 있다.

본고는 이러한 이중의 요구를 통합하는 개념으로서 AI 네트워크를 AI for Network와 Network for AI가 결합된 구조적 전환으로 정의하고, 페루프 기반 자율 제어에서 AI 내재형 물리계층, 데이터센터 네트워크 패브릭, 통신-컴퓨팅 융합 인프라에 이르는 주요 기술의 동향을 표준화와 산업 실증의 흐름과 함께 고찰하였다. 기술 동향 분석에서는 AI for Network 기술군의 표준 규격화가 선행되는 반면 Network for AI 기술군은 산업 실증이 표준화를 앞서는 비대칭이 확인되었으며, 이는 실증 데이터와 레퍼런스 구현이 표준화 경쟁의 실질적 승부처가 되고 있음을 보여준다.

도입 효과 분석에서는 AI 네트워크의 효과가 비용 절감과 시장 창출의 양 방향으로 동시에 나타남을 확인하였다. AI for Network 기반의 자동화와 최적화를 통해 통신사업자의 네트워크 비용은 약 17.75%, TCO는 약 7.1% 절감될 수 있는 것으로 추정되었으며, 절감분의 절반 이상이 무선 접속망의 가상화 전환에서 발생한다는 점에서 그 효과는 운영 자동화보다 인프라 구조의 전환에 크게 의존하는 것으로 나타났다. 동시에 AI-native 네트워크로의 전환은 신규 인프라 투자를 유발하여, '30년 기준으로 통신 특화 엣지 인프라 시장은 100억 달러, AI 네트워크 장비 시장은 138억 달러 규모를 형성하고, 후자는 '37년 956억 달러 규모(연평균 31.9% 성장)로 확대될 것으로 전망되었다. 특히 후자에서는 기술적 논의가 집중된 무선망 못지않게 전달·백본 영역의 시장 비중이 크게 나타나, 종단 간 지능화의 전략적 가치가 시장 측면에서도 확인되었다. 요컨대 AI 네트워크는 통신사업자에게 비용 구조 혁신의 지렛대이자 새로운 성장 시장이며, 그 전환의 본질은 CAPEX의 증감 자체가 아니라 통신 장비 중심의

투자 구조가 AI 컴퓨팅과 데이터 인프라 중심으로 재편되는 데 있다. 운영 효율화를 통해 확보한 여력이 신규 인프라 투자의 기반으로 이어질 때, 이 전환은 비로소 지속 가능한 사업 모델로 귀결될 수 있을 것이다.

2. 시사점

이러한 분석 결과는 다음과 같은 시사점을 제공한다. 첫째, 통신사업자 관점에서 AI 네트워크는 비용 절감을 넘어 사업 모델 전환의 분기점이 될 수 있으며, 그 전환에는 순서가 있다. 비용 절감 효과의 절반 이상이 무선 접속망의 가상화 전환에서 발생한다는 분석 결과는, 개방형 및 클라우드 기반 구조로의 전환이 AI 네트워크 효과를 실현하기 위한 선결 과제를 보여준다. 이렇게 확보한 운영 효율화의 여력을 엣지 인프라와 AI 컴퓨팅에 재투자할 때 연결 제공자에서 AI 인프라 및 서비스 제공자로의 역할 확장이 지속 가능한 경로가 되며, 시장이 '30년대 중반까지의 고성장 국면과 이후의 성숙 국면으로 구분되는 만큼 전환 시점의 결정 자체가 중요한 전략 변수가 된다.

둘째, 산업 관점에서는 시장 기회가 무선망에 국한되지 않는다는 점에 주목할 필요가 있다. 기술 논의와 표준화가 무선망에 집중되어 있음에도 시장 규모의 측면에서는 전달망과 백본망 영역이 모바일 코어를 크게 상회하는 것으로 전망되었다. 이는 국내 장비 산업과 연구계가 AI-RAN 일변도의 접근을 넘어 전송망과 코어망의 지능화 영역에서 기회를 모색할 근거가 되며, 특히 이 영역은 표준 경쟁이 상대적으로 초기 단계에 있어 선제적 기술 확보의 실익이 크다. 아울러 AI-RAN의 확산 속도가 통신사업자와 AI 반도체 및 장비 기업 간 협력 구도에 좌우되는 만큼, 초기 실증 협력에 참여하는 것이 시장 진입

의 관문이 될 것이다.

셋째, 국가적 관점에서 AI 네트워크는 국가 AI 인프라 경쟁력의 중추로 다루어질 필요가 있다. 주요국이 정책과 예산의 무게중심을 AI 내재형 네트워크로 옮기고 있는 가운데[1], 우리나라도 ‘하이퍼 AI 네트워크 전략’을 통해 방향을 제시한 만큼[2], 연평균 30%를 상회하는 성장이 전망되는 초기 시장을 국내 산업이 선점할 수 있도록 기술개발과 실증, 표준화를 연계한 일관된 추진이 필요하다. 특히 본고의 분석은 실증의 중요성을 이중으로 뒷받침한다. 기술 측면에서는 실제 망 환경에서 검증된 실증 데이터가 표준 경쟁의 실질적 승부처가 되고 있으며, 경제성 측면에서도 절감 효과 추정의 신뢰도가 가상화 전환의 실제 절감 수준에 크게 의존하는 것으로 나타났다. 국내 상용망 기반의 실증은 이처럼 기술 신뢰성과 정책 판단의 근거를 동시에 제공하므로, 통신사업자의 상용망과 국가 AI 인프라를 연계한 실증 기반의 확충과 그 성과를 국제 표준화로 환류하는 체계의 정비가 뒷받침되어야 한다.

본고는 방대한 AI 네트워크 논의 가운데 전환의 필요성과 기술 동향, 정량적 도입 효과를 하나의 틀에서 분석하였다는 점에서 의의가 있다. 다만 절감률과 전환 비중 등 주요 가정은 초기 시장의 특성상 불확실성을 내포하고 있어 상용 실증 데이터의 축

적에 따라 지속적으로 보정될 필요가 있으며, 글로벌 표준 및 정책 동향과 네트워크 API 등 산업 생태계 이슈에 대한 심층 분석은 후속 연구의 과제로 남긴다.

용어해설

AI-RAN ① AI-RAN 얼라이언스가 제시한 통신-컴퓨팅 융합 개념으로, AI로 무선망 성능과 운영을 개선하는 AI-for-RAN, 통신과 AI 워크로드를 공유 인프라에서 함께 수행하는 AI-and-RAN, 기지국 인프라 위에서 AI 서비스를 직접 제공하는 AI-on-RAN을 포괄하는 총칭

② 본고의 시장 전망(IV장)에서는 GPU 등 AI 가속 컴퓨팅 인프라 위에서 RAN 기능을 구현하고 무선 자원 관리에 AI 기반 최적화를 통한 장비와 시스템을 지칭하는 개념으로 사용

폐루프(Closed-Loop) 네트워크 상태의 관찰, 분석, 결정, 실행이 순환적으로 반복되고 실행 결과가 다시 학습에 반영되는 자동화 제어 구조. 사람의 개입을 최소화하는 자율 네트워크의 기본 운영 모델로, 주요 표준화 기구들이 공통으로 채택하고 있음

RIC(RAN Intelligent Controller) O-RAN 얼라이언스가 표준화한 무선망 지능형 제어기. 장주기 정책 최적화와 AI 모델의 학습을 담당하는 비실시간(Non-RT) 제어부와 무선 자원 제어 등을 근실시간으로 수행하는 준실시간(Near-RT) 제어부로 구성되며, 개방형 인터페이스를 통해 외부 응용(rApp, xApp)이 무선망을 분석하고 제어할 수 있게 함

네트워크 패브릭(Network Fabric) 다수의 서버와 가속기를 촘촘한 격자 구조로 연결하여 흩어진 연산 자원을 하나의 시스템처럼 묶는 고성능 네트워크 구조. 본고에서는 대규모 AI 학습 클러스터를 지연과 손실 없이 연결하는 데이터센터 내부 네트워크를 지칭하는 의미로 사용

AIDC(AI Data Center) AI 학습과 추론에 특화된 대규모 가속기 기반 데이터센터. 본고에서 AIDC 접속망은 AI 데이터센터를 기업, 산업 현장, 엣지와 연결하여 그 연산 자원의 산업적 활용을 가능하게 하는 네트워크를 의미함

참고문헌

- [1] NTIA, "NTIA Seeks Feedback on New Direction for Innovation Fund that Focuses on AI RAN," National Telecommunications and Information Administration, Mar. 2026.
- [2] 과학기술정보통신부, "하이퍼(Hyper) AI네트워크 전략," 2025. 12.
- [3] Ericsson, "Ericsson Mobility Report," Ericsson, June 2025.
- [4] Huawei, "Accelerating Network Transformation Towards All Intelligence in the Net5.5G Era," Huawei, Apr. 2024.
- [5] Cisco, "Cisco Data Center Networking Blueprint for AI/ML Applications," White Paper, Cisco Systems, Mar. 2024.
- [6] Y. Lin et al., "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training," arXiv preprint, 2017. doi: 10.48550/arXiv.1712.01887
- [7] NVIDIA, "NVIDIA Enterprise Reference Architectures," NVIDIA Corporation, 2026.
- [8] Arista Networks, "AI Networking," White Paper, Arista Networks, Sep. 2025.
- [9] Ericsson, "The Network for AI Experiences," White Paper, Ericsson, Feb. 2026.
- [10] N. Tahir and R. Parasuraman, "Edge Computing and Its Application in Robotics: A Survey," J. Sens. Actuator Netw., vol. 14, no. 4, 2025, p. 65.
- [11] ETSI, "AI in the Evolution of Autonomous Networks," ETSI White Paper No. 69, Nov. 2025.
- [12] 3GPP, "AI/ML for NG-RAN & 5G-Advanced towards 6G," 3rd Generation Partnership Project, 2025.
- [13] AI-RAN Alliance, "AI-RAN Alliance Vision and Mission White Paper," White Paper, AI-RAN Alliance, Dec. 2024.
- [14] O-RAN Alliance, "O-RAN Architecture and Open Interface Specifications," O-RAN Alliance, 2025.
- [15] ETSI ISG ZSM, "Zero-touch Network and Service Management (ZSM); Study on the Utilization of Agents in Autonomous Networks," ETSI GR ZSM 020 V1.1.1, Jan. 2026.
- [16] 3GPP, "Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface," 3GPP TR 38.843, Release 18, 2023.
- [17] ITU-T, "Digital Twin Network-Requirements and Architecture," Recommendation ITU-T Y.3090, Feb. 2022.
- [18] Ultra Ethernet Consortium, "UEC Specification 1.0," Ultra Ethernet Consortium, June 2025.
- [19] UALink Consortium, "Ultra Accelerator Link™ 200G 1.0 Specification," UALink Consortium, Apr. 2025.
- [20] IEEE 802.3 Working Group, "IEEE Std 802.3df-2024 and IEEE P802.3dj: 800 Gb/s and 1.6 Tb/s Ethernet," IEEE, 2024-2026.
- [21] L. Kundu et al., "AI-RAN: Transforming RAN with AI-driven Computing Infrastructure," arXiv preprint, 2025. doi: 10.48550/arXiv.2501.09007
- [22] 3GPP, "Study on Artificial Intelligence/Machine Learning (AI/ML) Management," 3GPP TR 28.908, Release 18, 2024.
- [23] ITU-T, "Architectural Framework for Machine Learning in Future Networks Including IMT-2020," Recommendation ITU-T Y.3172, June 2019.
- [24] ITU-T, "Machine Learning Marketplace Integration in Future Networks Including IMT-2020," Recommendation ITU-T Y.3176, Sep. 2020; "Architectural Framework for Machine Learning Model Serving in Future Networks Including IMT-2020," Recommendation ITU-T Y.3179, Apr. 2021; "Architectural Framework for Machine Learning Sandbox in Future Networks Including IMT-2020," Recommendation ITU-T Y.3181, 2022.
- [25] GSMA Intelligence, "The Mobile Economy 2025," GSMA, 2025.
- [26] Analysys Mason, "Operator Cost Structure and Network Total Cost of Ownership Benchmark," Analysys Mason, 2025.
- [27] Analysys Mason, "Open RAN Total Cost of Ownership Analysis," Telecom Infra Project, 2023.
- [28] GSMA, "Mobile Net Zero: State of the Industry on Climate Action," GSMA, 2025.
- [29] ABI Research, "Private Network MEC Server Market," ABI Research, 2024.
- [30] Mordor Intelligence, "Multi-access Edge Computing Market: Size, Share and Growth Forecast," Mordor Intelligence, 2025.
- [31] Global Market Insights, "Multi-access Edge Computing Market Size and Forecast," Global Market Insights, 2024.
- [32] Grand View Research, "Multi-access Edge Computing Market Size, Share and Trends Analysis Report," Grand View Research, 2025.
- [33] Dell'Oro Group, "Telecom Capex and RAN Market Five-Year Forecast," Dell'Oro Group, 2025.