AI-Native Open-RAN 네트워크 아키텍처에서의 마이크로서비스-다차원 자원 할당

함동호, 이소정, 곽정호

대구경북과학기술원

dhham97@dgist.ac.kr, sojungssi0526@dgist.ac.kr, jeongho.kwak@dgist.ac.kr

Microservice-based Dynamic Multi-dimensional Resource Allocation for AI-Native Services

Dongho Ham, Sojung Lee, Jeongho Kwak

DGIST

요 약

본 논문은 AI-Native 서비스의 QoS(Quality of Service) 요구사항을 충족시키기 위해 계층적 6G 아키텍처와 Open-RAN을 기반으로 한 마이크로서비스 기반 동적 자원 할당 프레임워크를 제안한다. 전통적인 RAN(Radio Access Network)의 한계점을 분석하고, Open-RAN의 개방형 인터페스, 분산형 구조 그리고. ML 기반 RIC(RAN Intelligent Controller)의 장점들을 활용하여 서비스 요구사항에 적응적으로 대응할 수 있는 해결책을 제시한다. 이 프레임워크는 동적 환경에서 확장성과 효율성을 보장하며, AI-Native 서비스의 발전을 위한 실질적인 기여를 할 것으로 기대된다.

Ⅰ. 서 론

AI-Native 네트워크는 설계, 배포, 운영, 유지보수를 포함한 네트워크의 모든 기능적 구성 요소에 인공지능(AI)이 통합된 패러다임을 의미한다 [1]. 이러한 네트워크는 모바일 기기, MEC(Mobile Edge Computing) 서버, 클라우드 서버 간의 상호작용을 활용하여 AI 기반 애플리케이션을 지원한다. 이러한 애플리케이션의 대표적인 예로는 6G의 핵심 기술인 XR(Extended Reality)을 활용한 공간 컴퓨팅이 있다 [2]. XR 서비스는 몰입형 환경을 위해 실시간 데이터 교환과 추론을 필요로 하며, 이는 종종 모바일 기기의 연산 능력을 초과한다. 이를 해결하기 위해 작업을 자원이 더 풍부한 MEC나 클라우드 서버로 오프로드할 수 있다.

XR, 자율주행, 홀로그램과 같은 AI-Native 서비스들은 학습과 추론을 위한 연산 자원뿐 아니라 엄격한 QoS(Quality of Service) 요구사항을 충족시키기 위한 강력한 네트워크 자원을 필요로 한다. 이러한 요구사항은 정확성, 에너지 효율, 지연 시간 등 동적으로 변화하기 때문에, 유연한 다차원 자원 할당 기술이 필요하다. 본 논문은 모바일, MEC, 클라우드 구성 요소로 이루어진 계층적 6G 아키텍처에 적합한 마이크로서비스 기반 동적 자원 할당 프레임워크를 제안한다. 마이크로서비스 셋을 활용하여 동적인 환경에서도 확장성과 자원 할당의 정밀도를 보장하며, AI-Native 서비스의 요구를 충족시킬 수 있다.

Ⅱ. 본론

전통적인 RAN(Radio Access Network) 구조는 단일 블랙박스 형태로 설계되어 유연성과 확장성에 한계가 있다. 기지국 하드웨어와 소프트웨어가 통합된 구조는 새로운 서비스나 기술을 도입하기 어렵게 만들며, 벤더 종속성 문제를 야기한다. 반면, Open RAN은 개방형 인터페이스와 기능 분리를 통해 이러한 문제를 해결한다. RU(Radio Unit), DU(Distributed Unit), CU(Centralized Unit)로 분리된 모듈화된 구조는 다양한 벤더 간 상호운용성을 보장하며, ML 기반의 RIC(RAN Intelligent Controller)는 실시간 자원 관리와 최적화를 가능하게 한다.

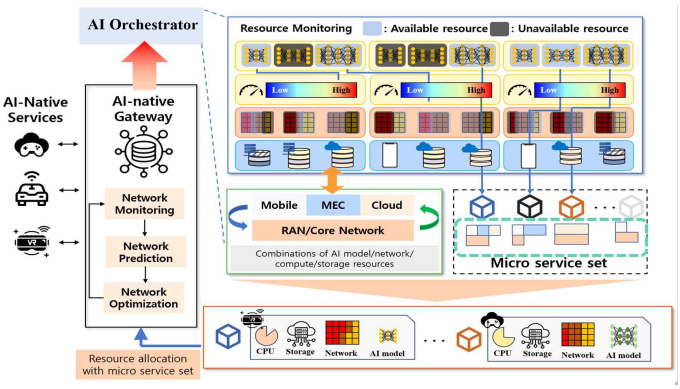

그림 1. 마이크로 서비스 기반 자원할당 프레임워크

Open RAN의 도입은 네트워크 전반의 효율성과 유연성을 향상시킨다. Open RAN은 다양한 네트워크 구성 요소 간 상호작용을 강화하며, 새로운 AI 기반 서비스의 도입을 용이하게 만든다. 특히, 분리된 RU, DU, CU 구조는 네트워크 자원을 최적화하는 데 유리하며, RIC를 통해 서비스 요구사항에 따라 자원을 실시간으로 조정할 수 있다. 다시 말해, Open RAN을 활용하면, 실시간 데이터를 수집 및 분석하여 네트워크 자원을 효율적으로 배분함으로써 AI-Native 서비스들의 높은 서비스 요구사항을 충족할 수 있다.

제안된 마이크로서비스 기반 동적 자원 할당 프레임워크는 특정 QoS 요구사항에 맞게 정의된 마이크로서비스 셋을 활용한다. 이 프레임워크는 환경 변화에 따라 자원 할당을 동적으로 조정하며, 서비스 확장성을 보장

한다. 예를 들어, 객체 탐지 서비스를 위해 정확성 중심의 모델을 사용하고, 추론 작업을 MEC 서버로 분산시킴으로써 낮은 지연 시간과 높은 정확성을 동시에 충족할 수 있다. Open RAN과의 통합은 프레임워크의 성능을 더욱 강화한다. Open RAN의 모듈화된 구조와 RIC의 실시간 최적화 기능을 통해, 다양한 서비스 요구사항에 적응적으로 대응할 수 있다.

Ⅲ. 결론

본 논문에서는 AI-Native 서비스의 QoS 목표를 충족하기 위해 계층적 6G 아키텍처와 Open RAN을 기반으로 한 마이크로서비스 기반 동적 자원 할당 프레임워크를 제안했다. 제안된 프레임워크는 동적인 환경에서도 확장성과 효율성을 보장하며, AI-Native 네트워크의 실질적인 발전에 기여할 수 있다. 앞으로의 연구는 제안된 프레임워크의 실증적 평가와 추가적인 최적화 방안을 탐구하는 데 초점을 맞출 것이다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2024-00398157)

참 고 문 헌

[1] M. Iovene, L. Jonsson, D. Roeland, M. D'Angelo, G. Hall, and M. Erol-Kantarci (Feb, 2023), "Defining AI native: A key enabler for advanced intelligent telecom networks," [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/ai-native/.

[2] Y. Lu and X. Zheng, "6G: A survey on technologies, scenarios, challenges, and the related issues," J. Ind. Inf. Integr., vol. 19, Sep. 2020, Art. no. 100158.

[3] S. Marinova and A. Leon-Garcia, "Intelligent O-RAN beyond 5G: Architecture, use cases, challenges, and opportunities," IEEE Access, vol. 12, pp. 27088 - 27114,19, Feb. 2024, doi: 10.1109/ACCESS.2024.3367289.