O-RAN 환경에서 AI 기반 애플리케이션 자원 관리 연구 동향

방민지, 정대영, 안예린, 백상헌

고려대학교

{qkdalswl2, eodud1884, annie2ahn, shpack}@korea.ac.kr

Research Trends of AI-based Network Application Resource Management in O-RAN

Minji Bang, Daeyoung Jung, Yerin Ahn, Sangheon Pack

Korea Univ.

요 약

5G 및 6G 기술의 발전으로 무선 네트워크 복잡성이 증가함에 따라, 제조사 종속성과 비호환성 문제를 가진 기존 RAN 구조의 한계가 드러나고 있다. 이를 해결하기 위해 개방형 아키텍처인 Open RAN (O-RAN)이 주목받고 있다. 개방형 인터페이스와 소프트웨어 기반 제어 구조를 통해 RAN 의 유연성과 확장성을 확보하며, AI 기반의 애플리케이션인 xApp 과 rApp 을 통해 자동화된 네트워크 제어 및 자원 최적화를 실현하고 있다. 이러한 애플리케이션들은 각각의 지연 시간 요구사항 및 처리 특성에 따라 적절한 컴퓨팅 자원 배치가 필요하다. 본 논문에서는 이러한 xApp/rApp 인스턴스의 자원 관리 및 배치 최적화에 대한 최근 연구 동향을 고찰한다. 특히, 애플리케이션별 AI 추론 시간과 지연 시간 제약을 고려한 자원 최적 할당 프레임워크와 near-real-time RIC (near-RT RIC) 구성 요소를 클라우드 및 엣지 환경에 동적으로 재배치하는 오케스트레이션 기법을 제안한 연구들을 분석하였다. 이를 통해 본 논문에서는 O-RAN 환경에서 AI 기반 제어 애플리케이션의 안정적인 운용과 에너지 효율성 확보를 위한 인프라 자원 관리 전략의 중요성을 강조한다.

Ⅰ. 서 론

최근 5G 및 6G 이동통신 기술의 발전으로, 무선 네트워크에 multiple input multiple output (MIMO), millimeter wave (mmWave) 등 다양한 기술이 적용되면서 복잡성이 급증하고 있다. 이로 인해 장비 유지보수 비용이 증가하고 있으며, 기존 폐쇄형 radio access network (RAN) 구조에서는 제조사 종속성과 장비 간 비호환성 등 구조적 한계가 드러나고 있다 [1].

이러한 문제를 해결하기 위한 대안으로 Open RAN (O-RAN) 구조가 주목받고 있다. O-RAN 은 하드웨어와 소프트웨어 분리를 기반으로 표준화된 개방형 인터페이스를 통해 이기종 장비 간 상호운용성을 보장하는 아키텍처이다. 특히 RAN intelligent controller (RIC)를 통해 데이터 기반의 자동화된 네트워크 제어와 자원 최적화가 가능하다 [2].

RIC 는 제어 시간에 따라 near real-time RIC (near-RT RIC)와 non real-time RIC (non-RT RIC)로 나뉜다. near-RT RIC 는 10ms 에서 1s 이내의 시간 단위에서 RAN 을 제어하는 역할을 수행한다. Near-RT RIC 에서 실행되는 애플리케이션인 xApp 은 실시간 RAN 데이터를 분석하여 트래픽 상황에 따라 무선 자원을 효율적으로 제어한다 [3]. 반면 non-RT RIC 는 service management and orchestration (SMO) 프레임워크의 일부로, near-RT RIC 에 비해 상대적으로 긴 1s 이상의 시간 단위에서 RAN 을 제어한다. Non-RT RIC 에서 실행되는 애플리케이션인 rApp 은 artificial intelligence (AI)/machine learning (ML)을 기반으로 네트워크의 전체 자원 운용에 대한 장기적인 관리를 담당한다 [4].

이러한 xApp 과 rApp 은 실제 O-RAN 환경에서 다양한 AI 기반 애플리케이션으로 구현되어 네트워크 성능 향상에 기여하고 있다. 대표적인 예시로, xApp 은 실시간 RAN 데이터를 활용해 traffic steering 을 수행하고, 무선 자원을 동적으로 재배치하여 셀 간 부하를 균형 있게 조정한다. 또한, 사용자 및 bearer 수준의 quality of service (QoS) 파라미터를 조정해 서비스 품질을 보장한다. 더불어 rApp 은 네트워크 전반의 성능을 진단하여 네트워크 부하와 자원 용량을 예측하고, near-RT RIC 에 주기적으로 최적화된 정책을 제공함으로써 장기적인 네트워크 운용 효율성을 높인다.

따라서, 최종적으로 이러한 O-RAN 환경에서 AI 기반 애플리케이션 자원 관리가 중요해지고 있으며, 본 논문에서는 이와 관련된 최신 연구 동향에 대해 소개한다.

Ⅱ. O-RAN 환경에서의 xApp/rApp 인스턴스 자원 할당 연구 동향

연구 [5]는 O-RAN 환경에서 AI 기반 xApp 과 rApp 의 자원 관리와 동시에 지연 시간 제약과 에너지 효율성을 최적화한다. 기존 자원 활용 기반의 스케일링 방식으로는 시간 제약적인 제어 결정 요구사항을 만족하기 어렵다는 문제를 해결하기 위하여 ScalO-RAN 이라는 최적화 기반의 제어 프레임워크를 제안한다. 해당 프레임워크는 AI 기반 xApp 인스턴스를 포함한 애플리케이션들의 자원을 최적으로 할당하고 스케일링하여, 각 애플리케이션의 지연 시간 요구사항 만족 및 에너지 소비를 최소화하는 목적을 갖는다. 따라서, AI 기반 애플리케이션 인스턴스의 추론 시간을 실제 측정하고 해당 데이터를 기반으로 모델링하여, 서버 할당 및 애플리케이션 유형 배치를 결정하는 최적화를 진행한다. 실험 결과, ScalO-RAN 이 기존 자원 기반 부하 분산 방식보다 낮은 에너지 소비 및 지연 시간 요구사항을 효과적으로 충족함을 보이며, AI 추론 속도가 제어 결정의 적시성을 결정하는 중요한 시간 제약적 문제임을 강조한다.

연구 [6]는 O-RAN 환경에서 near-RT RIC 에서 실행되는 xApp 의 효율적인 배치 및 orchestration 을 다루며 해당 애플리케이션이 만족해야 하는 지연 시간 제약 (10ms ~ 1s)을 충족하는 것에 중점을 둔다. 해당 연구는 near-RT RIC 를 E2 termination (E2T), shared data layer (SDL), shared time-series layer (STSL), network information base (NIB), xApp 등 여러 구성 요소로 세분화하고 해당 구성 요소들을 클라우드 및 엣지 컴퓨팅 노드에 최적으로 배치 및 동적으로 재배치하는 RIC orchestrator (RIC-O) 프레임워크를 제안한다. Non-RT RIC 의 RIC-O optimizer 및 deployer 는 near-RT RIC 의 구성요소와 xApp 인스턴스를 지연 시간 제약을 보장하기 위한 최적의 배치 전략을 구성한다. 여기서 RIC-O 는 지연 시간 제약과 동시에 near-RT RIC 구성 요소를 클라우드 및 엣지 서버에 배치하는 비용을 최소화하고자 한다. 이 때, 컴퓨팅 노드의 개수에 따른 호스팅 및 활성화 비용인 고정 비용과 컴퓨팅 노드에서 실행되는 near-RT RIC 의 구성 요소의 수에 따라 달라지는 변동 비용을 더하여 전체 비용을 구성한다. 변동 비용은 각 구성요소 E2T, SDL, STSL, NIB 의 개수에 비례하여 합산되어 계산한다. 또한 지연 시간 제약을 위반하거나 컴퓨팅 노드의 가용성 문제를 감지했을 때, 해당 문제를 발생하고 있는 xApp 인스턴스를 포함하고 있는 구성요소들을 재구성하는 동적 시스템을 통해 환경을 정상화하는 휴리스틱 알고리즘을 제안한다. 추가적으로, 휴리스틱 알고리즘과 RIC-O deployer 가 구성 요소를 재배치함으로써 E2 노드와 xApp 간의 제어 지연 시간을 10ms 이내로 만족하며 연결을 복구하는 것을 입증한다. 실험 결과, near-RT RIC 를 분해하여 RIC-O 가 관리하는 것이 지연 시간 제약을 만족하면서 전체적인 비용을 절감하였고, 자원 활용 효율성 측면에서 monolithic RIC 배포 기술보다 뛰어남을 보여준다.

Ⅲ. 결론

본 논문에서는 O-RAN 환경에서의 AI 기반 RAN 관련 애플리케이션 자원 관리 연구 동향을 살펴보았다. O-RAN 애플리케이션 xApp/rApp 이 AI 기반으로 동작하는 기술임이기에, 이러한 애플리케이션이 동작하는 컴퓨팅 자원의 관리가 필수적이다. 이를 위해, 여러 연구들에서는 클라우드 및 엣지 서버의 컴퓨팅 노드에서 xApp/rApp 요구사항 특성을 고려한 배포 기술을 제안하였다. 향후에는 클라우드 및 엣지 환경에서 split computing 기술을 적용한 RAN 애플리케이션을 pod cluster 구성 기술 및 동적 최적화에 관한 연구를 진행할 예정이다.

ACKNOWLEDGMENT

이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2024-00451909).

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2022-00156353).

참 고 문 헌

[1] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1376-1411, Jan. 2023.

[2] O-RAN Working Group 1, "O-RAN Architecture Description -v.13.0," O-RAN.WG1.O-RAN-Architecture Description-v13.00 Technical Specification, Feb. 2025.

[3] O-RAN Working Group 3, "O-RAN Near-RT RIC Architecture -v.7.0," O-RAN.WG3-RICARCH-v07.00 Technical Specification, Feb. 2025.

[4] O-RAN Working Group 2, "O-RAN Non-RT RIC Architecture -v.6.0," O-RAN.WG2.Non-RT-RIC-ARCH-v06.00 Technical Specification, Oct. 2024.

[5] S. Maxenti, S. D'Oro, L. Bonati, M. Polese, A. Capone, and T. Melodia, "ScalO-RAN: Energy-aware Network Intelligence Scaling in Open RAN," in *Proc. IEEE INFOCOM 2024*, Vancouver, BC, Canada, May 2024.

[6] G.M. Almeida, G.Z. Bruno, A. Huff, M. Hiltunen, E.P. Duarte, C.B. Both, and K.V. Cardoso, "RIC-O: Efficient placement of a disaggregated and distributed RAN Intelligent Controller with dynamic clustering of radio nodes," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 2, February 2024.