Wi-Fi 7 MLO 환경에서 QoS 만족을 위한 강화학습 기반 동적 링크 스케줄링

장하린, 장여진, 이상금*

*한밭대학교

hrjang713@gmail.com, jyeoj251@gmail.com, *sangkeum@hanbat.ac.kr

A Reinforcement Learning Approach for QoS-Aware Dynamic Link Scheduling in Wi-Fi 7 MLO Envirionments

Harin Jang, Yeojin Jang, Sangkeum Lee*

*Hanbat National Univ.

요 약

본 연구는 Wi-Fi 7의 핵심 기술인 MLO(Multi-Link Operation) 환경에서 QoS(Quality of Service) 요구사항을 만족시키기 위한 DRL(Deep Reinforcement Learning) 기반의 동적 링크 스케줄링 기법을 제안한다. DQN(Deep Q-Network) 에이전트는 단일 링크 선택뿐만 아니라 2개 링크 조합까지 포함하는 확장된 행동 공간을 탐색하며 트래픽 유형에 따라 차별화된 보상 함수를 통해 최적의 링크 선택 정책을 학습한다. 제안된 DQN 모델의 성능을 검증하기 위해 지연 시간 최소화에 특화된 RTT(Round Trip Time) 기반 정책과 동일한 환경에서 평균 처리량, 평균 지연 시간, 지연 시간 위반율 등 다각적인 QoS 지표를 통해 성능을 비교 평가했다. 실험 결과 DQN 에이전트는 RTT 기반 정책과 대등한 수준의 지연 시간 및 안정성을 유지하면서 평균 처리량은 약 95% 높은 성능을 보인다. 이는 DQN 에이전트가 트래픽의 QoS 요구사항을 명확히 인지하여 상황에 맞는 정책을 성공적으로 학습했음을 의미한다. 본 연구는 복잡하고 동적인 차세대 네트워크 환경에서 강화학습이 효과적인 자원 관리 해법이 될 수 있다는 의의를 지닌다.

Ⅰ. 서 론

Wi-Fi 7의 MLO(Multi-Link Operation)는 2.4/5/6 GHz 대역을 동시에 사용할 수 있도록 하여 큰 성능 향상을 기대할 수 있으나, 실제 환경에서는 스케줄링 복잡성이 증가한다. 상이한 QoS(Quality of Service) 요구사항을 가진 트래픽이 혼재하여 MLO의 잠재력을 완전히 활용하기에도 어렵다.

전통적인 RTT(Round Trip Time) 기반 정책은 지연 시간 최소화에는 효과적이나 처리량이 중요한 트래픽에 대해 다중 링크를 활용한 대역폭 확장에는 한계가 있다[1]. 이러한 한계를 극복하기 위해 환경과의 상호작용을 통해 스스로 최적 정책을 학습하는 강화학습(DRL) 기반 스케줄링 기법을 제안한다. 심층 Q-네트워크(DQN) 에이전트는 동적 환경에서 트래픽 유형과 링크 상태를 인지하여 단일 및 다중 링크 조합을 최적으로 선택함으로써 처리량과 지연 시간의 균형을 맞추는 것을 목표로 한다.

Ⅱ. 본론

문제 정의 및 모델링

Wi-Fi 7 MLO 환경에서의 링크 선택 문제를 마르코프 의사결정 과정(MDP)로 정의하였다. 각 단계 t에서, 에이전트는 상태 $s_t = [\tau, c0, c1, c2]$를 관찰한다. 여기서 $\tau \in \{0,1\}$은 트래픽 유형을 나타내며(0=처리량 중심, 1=지연 중심), $c_i \in [0,1]$은 링크 $c_i \in \{0,1,2\}$ (2.4/5/6GHz)의 정규화된 혼잡도를 의미한다. 행동은 에이전트가 상태 S_t를 관찰한 후 취할 수 있는 선택지의 집합이다. 단일 링크 선택과 두 개의 링크 조합까지 포함하여 총 6개의 행동을 정의했다.

$$A = \{\{0\},\{1\},\{2\},\{0,1\},\{0,2\},\{1,2\}\}$$

실제 네트워크의 동적 특성을 반영하기 위해 자기 회귀 모델을 도입했다. 특정 링크의 현재 혼잡도는 이전 시점의 혼잡도에 강한 상관관계를 가지며, 이는 다음과 같은 식으로 표현한다.

$$c_{i,t} = \alpha \times c_{i,t-1} + (1-\alpha) \times noise_t$$

α는 이전 상태의 영향력을 결정하는 상관계수이며 0.9로 설정하여 높은 시간적 상관관계를 부여했다. $noise_t$는 예측 불가한 외부 간섭이나 새로운 트래픽 발생을 모방하는 무작위 변수다. 이러한 동적 모델은 에이전트가 순간적인 상태와 변화의 추세에 대응하는 정책을 학습하도록 요구한다. 보상 함수는 다음과 같이 정의하였다[2].

- **처리량 중심 트래픽:** 보상은 선택된 링크 조합의 총 가용 대역폭에 비례하도록 설계했다. 이는 에이전트가 여러 개의 한가한 링크를 묶어 총 데이터 전송량을 극대화하도록 장려한다.

$$R_t = \sum_{i \in a_t} (a - c_i)$$

- **지연 시간 중심 트래픽:** 보상은 선택된 링크들 중 가장 낮은 혼잡도 상태를 기준으로 계산한다. 또한 혼잡도가 특정 임계치(0.7)를 초과할 경우 큰 음수 보상을 부여하여 에이전트가 지연 시간에 민감한 트래픽을 위험한 링크로 전송하는 행동을 회피하도록 학습한다. c_{eff}는 유효 혼잡도를 의미한다.

$$R_t = \begin{cases} (1-c_{eff})^2, & c_{eff} \le 0.7 \\ -1, & c_{eff} > 0.7 \end{cases}$$

실험

DQN 알고리즘을 채택하여 실험을 진행했다. 상태 s_t를 입력받아 6개의

행동 각각에 대한 예상 미래 보상의 합, 즉 Q-value $Q(s_t, a)$를 출력하는 심층 신경망으로 구성된다. 에이전트는 Epsilon-greedy 정책에 따라 ϵ의 확률로 무작위 탐험을 하거나 $1-\epsilon$의 확률로 현재 Q-value가 가장 높은 행동을 선택한다.

전체 학습은 2000 에피소드 동안 진행되며 각 에피소드는 100 타임 스텝으로 구성된다. 에이전트는 매 스텝마다 환경과 상호작용하며 경험을 리플레이 버퍼에 저장하고, 버퍼에서 샘플링된 미니배치를 통해 Q-네트워크를 업데이트한다.

비교 모델로 RTT 기반 정책을 채택했다. 매 순간 현재 상태에서 가능한 모든 행동에 대한 잠재적 지연 시간을 계산하고 그중 가장 낮은 지연 시간을 유발하는 행동을 항상 선택한다. 두 모델을 다각적으로 비교하기 위해 다음 세 가지 지표를 사용했다.

1. **평균 처리량:** 처리량 중심 트래픽이 주어졌을 때 에이전트가 달성한 평균 데이터 전송률이고, 높을수록 우수하다.
2. **평균 지연 시간:** 지연 시간 중심 트래픽이 주어졌을 때 발생한 평균 데이터 전송 지연이고, 낮을수록 우수하다.
3. **지연 시간 위반율:** 지연 시간 중심 트래픽을 임계치(0.7) 이상의 높은 혼잡도를 가진 링크로 전송한 비율이고, 낮을수록 우수하다.

실험 결과

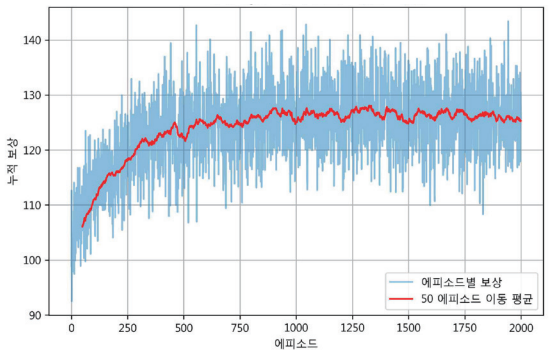

그림 1. DQN 에이전트 학습 곡선

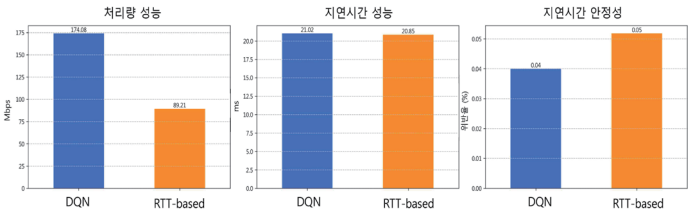

그림 2. 제안 모델과 RTT 기반 정책 성능 비교

그림 1은 DQN 에이전트의 학습 과정 동안 에피소드별 누적 보상의 변화를 보여준다. 학습 초기에는 무작위 탐험으로 인해 보상이 낮고 변동성이 크지만 약 500 에피소드 이후 학습이 진행됨에 따라 보상이 꾸준히 상승하며 안정적으로 수렴하는 것을 확인할 수 있다. 이는 제안된 DQN 에이전트가 복잡한 동적 환경과 확장된 행동 공간 속에서 성공적으로 정책을 학습했음을 의미한다.

그림 2는 학습이 완료된 DQN 에이전트와 RTT 기반 정책을 동일한 환경에서 50,000 스텝 동안 평가한 결과를 세 가지 지표로 비교한 그래프이다.

- **처리량 성능:** DQN 에이전트는 약 174.08 Mbps를 달성하여 89.21 Mbps를 기록한 RTT 정책에 비해 약 95% 더 높은 성능을 보였다. 이는 DQN 에이전트가 처리량 중심 트래픽을 만났을 때 다중 링크를 사용하여 대역폭을 확장했음을 의미한다. 반면 RTT 기반 정책은 지연 시간 최소화에만 집중하기 때문에 처리량 증대 기회를 놓쳤다.
- **지연 시간 성능:** DQN 에이전트가 21.02ms, RTT 기반 정책이 20.85ms를 기록하여 두 모델이 대등한 성능을 보였다.
- **지연 시간 안정성:** DQN 에이전트가 0.04%, RTT 기반 정책이 0.05%를 기록하며 높은 안정성을 보였다.

본 실험에서 구축한 Wi-Fi 7 MLO 실험 환경은 3개의 무선 링크를 사용하는 소형 네트워크 노드를 대상으로 하며 이는 스마트홈 게이트웨이, 산업 IoT 허브 수준의 구성과 유사하다. 이러한 네트워크 규모는 약 3~5개의 AP(Access Point)와 10~20개의 단말이 동시에 접속하는 환경이다[3]. 실험 결과는 산업 IoT, 스마트 팩토리, 자율 제어 네트워크 등 다양한 응용 분야에서 기존의 정적 또는 RTT 기반 제어 정책을 대체할 수 있는 잠재력을 제시한다. 따라서 향후 산업용 무선통신 시스템의 지능형 스케줄링, 적응형 링크 선택 분야로 확장할 수 있다.

Ⅲ. 결론

본 논문에서는 Wi-Fi 7 MLO 환경에서 다양한 QoS 요구사항을 만족시키기 위한 강화학습 기반 동적 링크 스케줄링 기법을 제안했다. 실제 네트워크의 동적인 특성을 모방한 시뮬레이션 환경을 구축하고 MLO의 핵심 기능인 다중 링크 동시 전송을 지원하도록 구성했다.

실험 결과, 제안된 DQN 에이전트는 실제 네트워크에서 널리 사용되는 RTT 기반 정책과 비교하여 지연 시간 성능과 안정성을 대등한 수준으로 유지하면서도 평균 처리량을 약 95% 이상 향상시켰다. 이는 DQN 에이전트가 트래픽 유형을 이해하고 장기적인 보상을 극대화하는 유연하고 지능적인 정책을 성공적으로 학습했음을 의미한다.

강화학습이 복잡하고 동적인 차세대 네트워크 환경에서 효과적인 자원 관리 해법이 될 수 있다는 가능성을 제시한다. 향후 연구로는 다수의 사용자가 경쟁하는 다중 에이전트 환경으로 문제를 확장하여 보다 현실적인 시나리오에서의 성능을 검증하는 연구가 필요하다.

ACKNOWLEDGMENT

"본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2022-0-01068)

참 고 문 헌

[1] Shi, H.; Cui, Y.; Wang, X.; Hu, Y.; Dai, M.; Wang, F.; and Zheng, K. "STMS: Improving MPTCP Throughput Under Heterogeneous Networks," Proceedings of the 2018 USENIX Annual Technical Conference (USENIX ATC '18), Boston, MA, USA, July 11 - 13, 2018.

[2] M. Seo, L. F. Vecchietti, S. Lee, and D. Har, "Rewards Prediction-Based Credit Assignment for Reinforcement Learning with Sparse Binary Rewards," IEEE Access, vol. 7, pp. 115512 - 115523, 2019,

[3] Gordon, H.; Batula, C.; Tushir, B.; Dezfouli, B.; and Liu, Y. "Securing Smart Homes via Software-Defined Networking and Low-Cost Traffic Classification," arXiv preprint arXiv:2104.00296, 2021.