완전 분산형 QKD 네트워크를 위한 다중 에이전트 강화학습 환경

김주봉, 이찬균, 임현교, 이원혁
한국과학기술정보연구원

{jjbong, chankyunlee, hk.lim, livezone}@kisti.re.kr

Multi-Agent Reinforcement Learning Environment for Fully Distributed QKD Networks

Ju-Bong Kim, Chankyun Lee, Hyun-Kyo Lim, Wonhyuk Lee
Korea Institute of Science and Technology Information (KISTI)

요 약

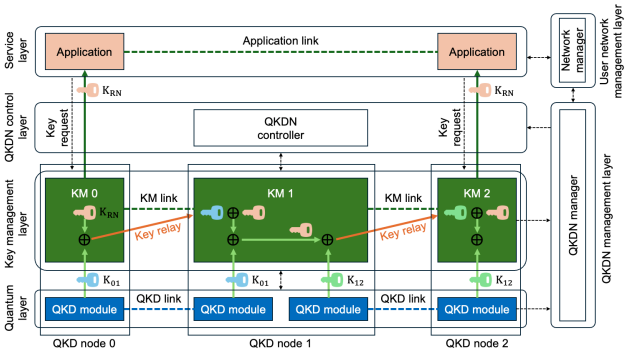
본 논문은 완전 분산형 양자 키 분배(QKD) 네트워크에서 요청의 스케줄링 및 분산 라우팅 제어를 위한 다중 에이전트 강화학습(MARL) 환경을 제안한다. 각 강화학습 에이전트는 QKD 노드와 대응하여 모델링 되고 QKD 네트워크의 현실적인 제약사항들을 반영한 시뮬레이션 상에서 제어된다. 제안 환경은 다양한 토폴로지, 링크 자원 제약, 요청 스케줄링, 키 릴레이 프로세스를 포함하며, 마르코프 결정 과정으로 정리된다. 이는 중앙 제어 없이 노드 수준에서 자율 제어가 필요한 QKD 네트워크 구조에서 MARL 기반 제어 정책의 적용 가능성과 성능을 정량적으로 평가하기 위한 실험적 기반을 제공한다.

Ⅰ. 서 론

양자키분배(quantum key distribution, QKD)는 양자역학의 원리에 기반해 이론적으로 완벽한 보안을 제공하는 차세대 통신 기술이다. QKD 는 양자컴퓨터의 보안 위협에 대비하여, 국가 간의 고신뢰 보안 인프라의 핵심 기술로 자리 잡고 있다. QKD 기술이 점차 상용화되어 가고 있음에 따라 다수의 QKD 노드로 구성된 대규모 QKD 네트워크(QKD networks, QKDN)에 대한 연구가 활발히 진행되고 있지만, 실시간으로 생성되고 소모되는 양자키 등의 자원을 효율적으로 관리하기 위한 분산형 제어 구조에 대한 연구는 상대적으로 부족하다.

현재 널리 활용되고 있는 QKDN 제어 방식은 중앙집중형 소프트웨어 정의 네트워킹(software-defined networking, SDN) 컨트롤러 기반으로 설계된다. 그러나, 네트워크 규모의 증가에 따라 다음과 같은 한계에 직면한다: (1) 전체 네트워크 상태 정보를 지속적으로 수집 혹은 갱신해야 하므로 지연(latency)과 제어 병목 현상이 발생하며, (2) 예기치 못한 컨트롤러 장애에 취약하고, (3) 네트워크 규모의 변화에 따라 글로벌 최적화로 확장하는데 관한 일반화 성능 한계가 존재한다.

최근에는 강화학습(reinforcement learning, RL)을 활용하여 QKDN 의 자원 할당 문제를 해결하고자 하는 시도가 이루어지고 있다. 한 연구에서는 QKDN 상에서 양자키 풀(quantum key pool)의 안정성과 자원 효율성을 극대화하기 위해 RL 의 기초적인 알고리즘인 Q-learning 기반의 양자키 자원 할당 알고리즘을 제안하였다 [1]. 또 다른 연구에서는 RL 기반의 라우팅 및 자원 할당 프레임워크를 도입하여, 임의로 요청이 발생하는 환경에서도 낮은 차단율과 높은 자원 활용율을 달성함을 보였다 [2]. 하지만 이들 연구는 모두 중앙 제어 구조에 기반하고 있어, 네트워크 규모 확대에 따른 제어 병목, 컨트롤러 장애 발생 관리, 일반화 성능 등의 문제를 여전히 내포하고 있다.

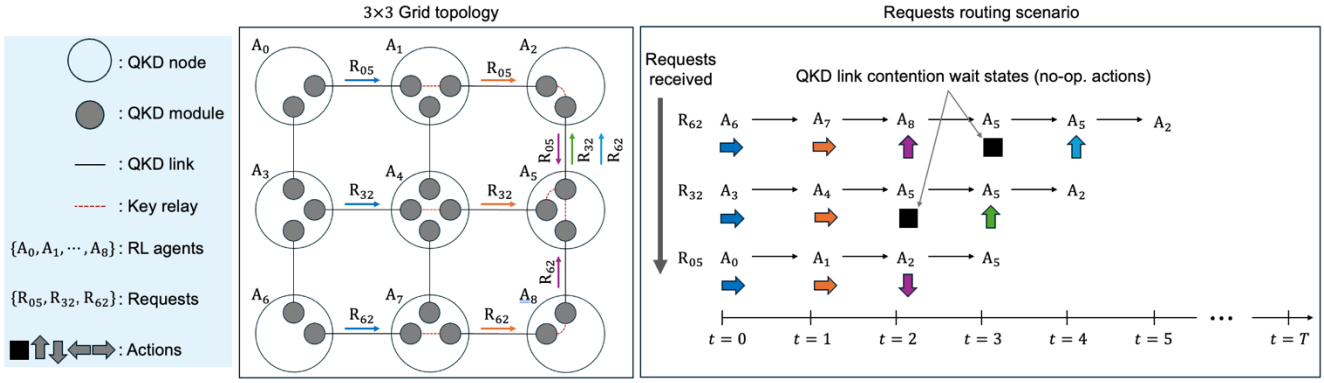


(그림 1). QKDN 에서 ITU-T 표준 기반 키 생성, 릴레이 및 요청 처리 과정의 계층 구조. 각 QKD 노드는 양자 계층 QKD 모듈과 키 관리 계층 KM 으로 구성되며, 인접 노드 간에 생성된 양자키는 릴레이 과정을 거쳐 최종 수신 노드로 전달된다. 이렇게 전달된 키는 서비스 계층의 응용 간 보안 통신을 가능하게 한다.

이러한 한계를 극복하기 위해 본 논문에서는 각 QKD 노드를 에이전트에 대응시키고, 다중 에이전트들이 로컬 정보만을 바탕으로 자율적인 의사결정을 수행할 수 있도록 설계된 완전 분산형 QKDN 시뮬레이션 환경을 제안한다. 제안 환경은 향후 다중 에이전트 강화학습(multi-agent reinforcement learning, MARL) 기반의 QKDN 최적화를 위한 실험 플랫폼으로 사용될 수 있으며, 완전 분산형 구조의 실효성과 학습 기반 정책의 실증적 가능성을 동시에 평가할 수 있는 기반을 제공한다.

Ⅱ. 본 론

제안 환경은 양자컴퓨팅 시스템의 시뮬레이션 위해 개발된 NetSquid 플랫폼을 기반하여 구축되었다 [3]. (그림 1)에서와 같이 각 노드는 양자키 프로토콜(BB84, E91 등) 기반 양자키 생성을 수행할 수 있는 QKD 모듈과 키 릴레이를 위한 에이전트 로직을 포함한다.



(그림 2). 3×3 Grid 토폴로지에서 요청 R_{05}, R_{32}, R_{62}가 각각 에이전트 A_0, A_3, A_6로부터 최종 목적지까지의 라우팅 예시. 토폴로지 예시에서 요청들의 QKD 링크 키 릴레이 순서가 시각화 되어 있으며, 요청을 받은 에이전트는 각 타임스텝에서 요청 거절 혹은 인접 노드로의 키 릴레이 또는 대기(no-op.) 행동을 수행하며 요청을 처리한다.

에이전트는 총 $T(\in \mathbb{Z}_{>0})$ 타임스텝 동안에 걸쳐 각 타임스텝 t 마다 하나의 행동(예: 요청 수락/거절, 릴레이 경로 선택 등)을 결정하며, 이는 해당 노드의 관찰 가능한 로컬 상태에 기반한다.

환경에서 제공하는 토폴로지 종류에는, $N \times N$ 크기의 Grid, N 크기의 Chain 및 Ring, 임의 사용자 정의(Custom) 토폴로지 등이 존재한다. 환경의 라우팅 요청 시나리오에서는, 모든 가능한 방향의 요청 $R_{source-destination}$ 이 매 $T_g(\in \mathbb{Z}_{>0}, 1 \le T_g \le T)$ 타임스텝마다 생성되며, 각 요청은 라이프타임(lifetime) 제한($T_l = T_g$)을 가진다. 양자키는 인접 노드 간 128-bit 키를 양자키 프로토콜 기반으로 $M(\in \mathbb{Z}_{>0})$ 회 반복하여 생성하며, 토폴로지 크기에 따라 적절히 반복 횟수를 조정한다. 양자 링크 자원 점유로 인한 에이전트 행동의 충돌 상황을 방지하고자, 각 노드는 매 타임스텝 마다 하나의 요청에 대해서만 행동을 선택하며, 이산 시간 기반 시뮬레이션을 따른다.

제안 환경은 MARL 을 위한 마르코프 결정 과정(Markov decision process, MDP)으로 정리된다. 상태(state)는 두 개의 채널을 가지는 2 차원 행렬로 구성되며, 첫 번째 채널은 QKD 링크 연결 여부를, 두 번째 채널은 각 링크의 남은 양자키 보유량을 나타낸다. 각 에이전트는 자신이 수신한 요청에 대한 1 차원 벡터인 관찰(observation) 정보를 행동 선택에 활용한다. 관찰 정보는 에이전트 ID, 요청 출발지 및 목적지, 남은 라이프타임(lifetime), 처리 지연(processing delay) d_t^p, 큐잉 지연(queueing delay) d_t^q, 잔여 키 비율 등의 요소로 구성된다.

행동(action)은 매 타임스텝마다 선택되며, 가능한 행동 공간은 인접 노드 중 하나로의 키 릴레이(정수 인덱스 0 부터 $N-1$ (Grid 토폴로지의 경우 N^2-1) 까지), 요청 거절(reject), 또는 아무런 동작도 수행하지 않는 대기(no-operation) 행동으로 구성된다. 임의 타임스텝에서 보상(reward) 함수는 요청 처리 성공 횟수 r_t^s에서 정규화된 처리 및 큐잉 지연 값을 감산하는 방식으로 정의된다:

$$r_t = r_t^s - \alpha\left(\frac{d_t^p}{T} + \frac{d_t^q}{T}\right), 0 < \alpha. \quad (1)$$

수식 (1)과 같은 보상 함수 설계는 학습 과정에서 에이전트가 단순히 많은 요청을 처리하는 것뿐 아니라, 지연을 최소화하는 방향으로 행동 정책을 최적화하도록 유도하기 위함이다. 수식 (1)에서 α 는 지연이 보상에 미치는 정도를 조절하며, 사용자 임의로 변경 가능하다.

(그림 2)는 3×3 Grid 토폴로지 상에서 요청 3 개가 각각 에이전트들의 라우팅 경로 결정에 의해 처리되는 시나리오 예시를 보여준다. 각 QKD 노드는 강화학습 기반의 독립 에이전트로 작동하며, 요청을 수신한 에이전트는 인접 노드 중 하나를 선택하여 키 릴레이를 수행하거나, 링크 경합(link contention) 상황에서는 대기(no-op.) 행동을 선택한다. 제안 환경은 비동기 요청 처리, 키 릴레이 기반의 라우팅 경로 설정, 처리 및 큐잉 지연, QKD 링크 경합 등의 요소를 포함한 현실적인 시뮬레이션을 제공한다.

Ⅲ. 결 론

제안 환경은 QKDN의 분산 자원 제약 조건과 비동기 요청 특성을 반영하여 스케줄링 및 라우팅 문제를 자율적으로 해결할 수 있는 MARL 기반 정책 연구를 지원한다. 현실적인 제약 조건과 MDP 정의 하에서의 시뮬레이션을 통해 본 연구는 QKDN에서의 다중 에이전트 학습의 효과성을 정량적으로 검증하며, 자율 제어 기반 QKDN 설계에 실질적인 실험 인프라를 제공한다.

ACKNOWLEDGMENT

이 논문은 2025 년도 한국과학기술정보연구원(KISTI)의 기본사업의 지원(과제번호: (KISTI)K25L5M2C2)과 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.RS-2025-02263666)을 받아 수행된 연구임.

참 고 문 헌

[1] Y. Zuo, Y. Zhao, Y. Xiaosong, A. Nag and J. Zhang, "Reinforcement Learning-based Resource Allocation in Quantum Key Distribution Networks," 2020 Asia Communications and Photonics Conference (ACP) and International Conference on Information Photonics and Optical Communications (IPOC), pp. 1-3, 2020.

[2] Sharma, P., et al., "Deep Reinforcement Learning-based Routing and Resource Assignment in Quantum Key Distribution-Secured Optical Networks," IET Quantum Communication, vol. 4, pp. 136-145, 2023.

[3] Coopmans, T., Knegjens, R., Dahlberg, A. et al., "NetSquid, a NETwork Simulator for QUantum Information using Discrete events," Communications Physics, vol. 4, pp. 164, 2021.