강화학습 기반 룩업테이블 정책으로 구현한 SmartNIC-Asisted XR 스트리밍

심준형, 김민혁, 이한준, 고한얼
경희대학교

{simjun10, kmh3339, 2hajpal, heko}@khu.ac.kr

SmartNIC-Assisted XR Streaming via RL-Based Lookup-Table Policies

Junhyung Sim, Minhyeok Kim, Hanjun Lee and Haneul Ko
Kyung Hee Univ.

요 약

시선 기반 적응형 스트리밍은 확장현실(XR) 서비스에서 대역폭 효율성과 사용자 경험(QoE)을 동시에 높이기 위한 대표적인 접근이지만, 서버 측에서 시야 정보를 반영해 품질을 조정하는 기존 구조는 종단 간 지연으로 인해 실시간성 저하와 시야-콘텐츠 불일치를 초래한다. 이를 해결하기 위해 본 논문은 네트워크 엣지의 SmartNIC 에서 시선 정보를 바탕으로 패킷 필터링을 즉시 수행하는 SmartNIC 기반 XR 전송 프레임워크(SXRDF)를 제안한다. SXRDF 는 XR 타일 전송을 MDP 로 정식화(상태: 중심 타일·시선 방향·혼잡·추정 지연, 행동: 예측 중심·전송 반경)하고, 오프라인 강화학습으로 학습한 정책을 룩업테이블로 구현하여 SmartNIC에서 Base는 항상 전송, Enhancement는 선택 전송/드롭을 실시간 결정한다. 이로써 서버 왕복 없이 시야 예측과 재생 시점을 반영한 타일 선택이 가능해져 종단 간 지연과 시야 불일치가 줄고, 대역폭 효율과 QoE를 함께 개선할 수 있다.

Ⅰ. 서 론

확장현실(Extended Reality, XR) 기술은 고해상도와 초저지연 콘텐츠 전송을 위해 많은 자원을 요구하므로, 전송·인코딩·스케줄링 전 단계에서 지능적인 자원 관리가 요구된다. 특히 HMD(Head mounted display)를 사용하는 XR 시스템은 시선과 머리 움직임 등 사용자 센서데이터가 실시간으로 생성되므로, 동일한 자원으로 더 안정적인 품질을 유지하려면 효율적인 전송 방식이 필요하다. 이를 위해 많은 연구는 시선 기반 적응형 타일링 스트리밍을 사용한다[1]. 화면을 여러 타일로 나누고, 사용자의 시야에 있는 타일만 고해상도로 전송하여 자원을 줄이는 방식이다.

문제는 이 방식을 실제 시스템에 적용할 때의 속도이다. 사용자의 시야를 반영해 적응형 과정을 거쳐 HMD에 보여지기까지 전 과정이 1프레임 이내(시야가 바뀌기 전)에 이뤄져야 시야와 전송된 콘텐츠가 맞아떨어진다. 그러나 기존 대부분의 연구는 이러한 적응 과정이 서버 측에서 이루어지도록 설계되었다. 서버 측에서 적응형 과정을 거치게 되면, 클라이언트에서 수집한 시야 데이터를 서버로 보내고, 서버가 이를 처리해 결정한 결과를 다시 클라이언트로 보내는 종단 간 지연이 필연적으로 발생한다. 그 결과 콘텐츠가 클라이언트에 도달하는 시점에는 이미 시야가 변해 있어, 전송된 화면과 실제 시야가 어긋나는 시야 불일치가 나타나기 쉽다[2].

본 논문은 이러한 한계를 줄이기 위해 SmartNIC 기반 XR 전송 프레임워크(SXRDF)를 제안한다. 핵심은 적응을 서버가 아니라 SmartNIC에서 바로 결정하도록 구조를 바꾸어 왕복 구간을 없애고 지연을 줄이는 것이다. 구체적으로, SmartNIC에서 시야·혼잡·버퍼 상태 등의 정보가 반영된 룩업테이블 한 번 조회로 타일 전송/드롭을 즉시 결정한다. 이때 룩업테이블은 강화학습으로 미리 학습한 정책을 기반으로 구성되며, MDP 정식화를 통해 시야 변화, 네트워크 상태, 영상 재생 시점 등을 함께 고려하도록 설계한다. 결과적으로 SXRDF는 시선 기반 적응형 타일링의 장점을 유지하면서, 적응 위치를 SmartNIC으로 이동시켜 종단 간 지연과 서버 연산 부담을 동시에 낮추는 것을 목표로 한다.

Ⅱ. 본론

시스템 모델은 엣지 서버, 기지국(SmartNIC 포함), 클라이언트의 세 구성 요소로 구성된다. 아래는 시스템 모델에 대한 그림이다.

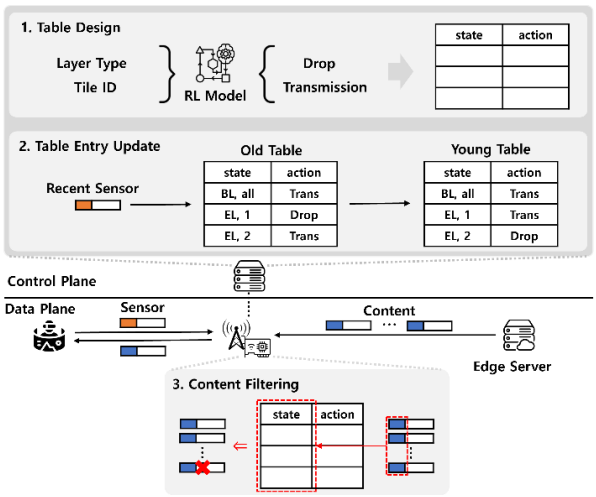

그림 1. 시스템 모델

본 시스템의 구성과 동작은 다음과 같은 흐름으로 이해할 수 있다. 먼저 서버는 각 타일의 Base layer(기본 화질)와 Enhancement layer(추가 화질)를 모두 보유하고,

패킷 헤더에 layer type과 tile ID만 붙여 단순 송출한다 (Base layer는 저화질의 콘텐츠를 나타내며 Enhancement layer와 합쳐져 고화질의 콘텐츠가 된다). 유선 구간 대역폭이 충분하므로 서버는 타일별 콘텐츠 품질을 조정하는 복잡한 로직을 생략할 수 있으며, 그 결과 서버 측 처리 구조가 단순해진다. 클라이언트(HMD)는 매 프레임 시야 중심 타일, 머리 움직임 방향 벡터(현재 중심–이전 중심으로 계산되어 다음 시야 예측에 사용), 수신 버퍼 상태(재생까지 남은 시간 추정), 무선 구간 혼잡 수준을 SmartNIC으로 지속 전달한다. SmartNIC은 이 센서데이터를 받아 실시간으로 룩업테이블을 갱신하고, 갱신된 룩업테이블을 조회하여 서버에서 들어오는 패킷을 필터링한다. 이때 Base layer는 항상 전송해 최소 품질을 보장하고, Enhancement layer는 선택적으로 전송/드롭하여 자원을 아낀다. 룩업테이블은 (layer type, tile ID) → {전송, 드롭} 형태의 key–value값 구조이며, 여기서 엔트리들은 사전에 정의된 정책을 따라, 시야가 바뀔 때마다 실시간으로 갱신된다.

여기서 정책은 오프라인 강화학습을 통해 만들어진다. 우리는 타일 전송 정책을 MDP로 정리하며, 상태 S_t는 {현재 중심 타일 c_t, 시선 방향 벡터 d_t = 현재 중심–이전 중심, 혼잡 레벨 g_t, 추정 지연 L_t}로 구성해 시야 변화(예측)와 무선 전송 지연/재생 시점을 함께 반영한다. 행동 A_t는 {예측된 중심 타일 c_t', 전송 반경 r}로 두어 "어디를 중심으로 얼마나 넓게 Enhancement를 보낼지"를 결정한다. 보상은 FoV 적중도 V_t에서 대역폭 소모 B_t와 끊김/전송 시간 초과 D_t를 뺀 $R_t = \alpha V_t - \beta B_t - \gamma D_t$로 정의한다. 이러한 설계의 이점은 명확하다. d_t를 통해 다음 시야를 미리 잡아 필요 타일만 선정하고, L_t로 재생 시점에 맞춰 제때 도착할 타일만 보내 지연·끊김을 억제하며, g_t로 혼잡 수준에 맞춰 전송 반경 r을 자동으로 줄여 대역폭 낭비를 피한다. 런타임에서는 센서데이터가 곧 상태로 입력되고, 정책이 내놓은 행동 (c_t', r)을 HMD 좌표계의 고화질 전송 타일 목록을 만든다. 이 목록을 기준으로 룩업테이블 엔트리를 프레임마다 간단히 갱신하여, (layer type, tile ID) → {전송, 드롭}을 즉시 반영한다. 결과적으로 본 방식은 정책이 센서 상태를 받아 곧바로 타일 집합을 산출하고, SmartNIC이 이를 룩업으로 라인레이트 수준에서 적용하도록 하여, 서버 왕복 없는 적응으로 종단 간 지연을 줄이고, 시야 예측·재생 시점을 동시에 고려해 필요한 만큼만 Enhancement를 선택함으로써 QoE를 최대화하는 방향으로 동작한다.

Ⅲ. 결론

본 논문에서는 확장현실(XR) 콘텐츠의 QoE를 유지하면서 무선 구간의 대역폭 효율을 높이기 위한 SmartNIC 기반 전송 프레임워크를 제안하였다. 제안 방식은 기존 서버 중심 시선 기반 적응 스트리밍의 왕복 지연과 시야–콘텐츠 불일치 문제를 완화하기 위해, SmartNIC에서 타일 단위 필터링을 수행하도록 구조를 재편한 것이 핵심이다. 학습은 비실시간으로 수행하고 실행은 SmartNIC에서 즉시 수행한다는 원칙 아래, 오프라인 강화학습으로 도출한 정책을 룩업테이블 형태로 구현하여 라인레이트 수준의 의사결정을 가능하게 하였다. 이를 통해 전체 지연을 줄이고 서버 구조를 단순화하며, 사용자의 시선 변화와 재생 타이밍에 신속히 반응함으로써, QoE를 유지하면서 네트워크 자원을 효율적으로 활용할 수 있음을 기대할 수 있다.

향후 연구로는 NVIDIA BlueField 3(SmartNIC이 탑재된 DPU)기반의 실제 테스트베드를 구축하여, 다양한 무선 채널·멀티사용자 환경에서 지연·시야 적중도·끊김 지표를 체계적으로 검증하고, 다양한 네트워크 조건에 맞춘 정책 최적화 방안을 추가로 탐색할 계획이다.

ACKNOWLEDGMENT

본 연구는 2024 년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(과제번호: RS-2024-00340698)

참 고 문 헌

[1] C.-T. Hsiao *et al.*, "Towards Retina-Quality VR Video Streaming: 15ms Could Save You 80% of Your Bandwidth," *ACM SIGCOMM Computer Communication Review*, vol. 52, no. 3, pp. 9–16, Jul. 2022.

[2] J. Shi, Y. Zhang *et al.*, "Mobile VR on Edge Cloud: A Latency-Driven Design," *in IEEE International Conference on Edge Computing (EDGE)*, 2019, pp. 1–8.

[3] Z. Chen, L. Li *et al.*, "FlexiNS: Enabling Flexible Packet Processing with SmartNICs," *in Proceedings of the USENIX Annual Technical Conference (USENIX ATC)*, 2023.

[4] J. Sim *et al.*, "SmartNIC-based XR content delivery framework," *in Proc. KICS Conf.,* Jun. 2025 pp. 1085-1086