

사용자 전력 상한 대비 오프로딩 지연시간 및 드롭아웃 성능 비교

이승훈, 박현철
한국과학기술원

seunghun21@kaist.ac.kr, hcpark@kaist.ac.kr

Offloading Delay and Dropout Performance Comparison with User Power Limits

Lee Seunghun, Park Hyuncheol
Korea Advanced Institute of Science and Technology.

요약

본 논문은 셀프리(cell-free) 모바일 엣지 컴퓨팅(MEC) 환경에서 마감시간 제약 하의 작업 오프로딩, 연결, 빔포밍, 연산 자원할당을 통합적으로 최적화하는 프레임워크를 바탕으로, 사용자 전력 상한 변화가 드롭아웃 비율과 활성 사용자 평균 지연에 미치는 영향을 정량적으로 분석한다. 우리는 가중계수를 통해 드롭아웃 최소화와 지연 최소화 간의 상충관계를 동시에 조절하며, 업링크 무선·연산·오버헤드 요소를 동시에 반영한다. 시뮬레이션 결과, 사용자 전력이 15 dBm 으로 낮아지는 저전력 구간에서도 제안 기법은 두 기준 기법 대비 더 낮은 평균 지연과 드롭아웃 비율을 달성함을 확인하였다.

I. 서론

지연 민감형 애플리케이션(실시간 인식, 몰입형 서비스 등)의 확산은 사용자의 한정된 연산 능력과 변동성이 큰 무선 링크를 동시에 고려하는 통합 설계를 요구한다. 단일 셀 기반 또는 고정 연결 규칙은 마감시간 초과(드롭아웃)를 빈번히 유발할 수 있다 [1]. 특히, 분산형 엣지 인프라에서 발생하는 불균형한 무선 링크 품질은 단순한 오프로딩 정책 [2]만으로는 저전력의 사용자로는 충분히 대응하기 어렵다. 본 연구는 셀프리 구조와 가중합 목적을 결합해 드롭아웃과 지연을 함께 멀티에이전트 강화학습을 적용하여 최적화하고, 사용자 전력 상한 변화에 대한 성능 민감도를 실험적으로 규명한다.

II. 본론

네트워크는 N 개의 다중안테나 기지국(각각 안테나 수 N_B , RF 체인 수 N_{RF} 와 K 개의 단일안테나 사용자로 구성되며 업링크만 고려한다. 각 슬롯 시작에 사용자 k 의 작업 크기 A_k bit가 주어지고, 슬롯 길이가 τ 이다. 마감시간을 초과하면 드롭아웃으로 처리한다. 오버헤드는 파일럿·피드백·요청 시간의 합으로,

$$\nu = \frac{T_1 + T_2 + T_3}{W} (s),$$

파일럿은 이진 위상 편이 변조 심볼을 가정하였으며, $T_1 = L_p N$ (bits), $T_2 = \zeta N_B N K$ (bits), $T_3 = \zeta K$ (bits)이며, W (Hz)는

대역폭, L_p 는 파일럿 심볼 수, ζ 는 스칼라 표현 비트 수이다.

하이브리드 결합기에서는 이산푸리에변환 코드북을 적용하며, 이 때 사용자 k -기지국 i 링크의 신호대간섭및잡음비 (SINR)은

$$\Gamma_{i,k} = \frac{|P_k (\mathbf{w}_{BB,i,k})^H (\mathbf{w}_{RF,i})^H \mathbf{h}_{i,k}|^2}{\sum_{\ell \neq k} P_\ell |(\mathbf{w}_{BB,i,k})^H (\mathbf{w}_{RF,i})^H \mathbf{h}_{i,\ell}|^2 + \sigma^2 |(\mathbf{w}_{BB,i,k})^H (\mathbf{w}_{RF,i})^H|^2} \text{이고,}$$

링크의 전송률은 $R_{i,k} = W \log_2(1 + \Gamma_{i,k})$ (bits/s) 이고, 사용자 전력 상한은 $0 \leq P_k \leq P_k^{\max}$ 이다. 로컬 작업과 오프로딩 작업은 분할비 $\lambda_k \in [0,1]$ 로 나누어 병렬 처리한다. 표준 큐빅 전력 모델에서 로컬 클럭과 로컬

지연은 $f_{loc,k} = \left[\frac{P_k^{\max} - P_k}{\kappa} \right]^{\frac{1}{3}}$ (Hz), $D_k^{\text{comp}} = \frac{\eta \cdot \lambda_k \cdot A_k}{f_{loc,k}}$ (s) 이고,

오프로딩 지연은 $D_{i,k}^{\text{trans}} = \frac{(1 - \lambda_k) \cdot A_k}{R_{i,k}}$ (s) 이다. 병렬

구조에서 총 실행 지연은 $D_k = \nu + \max\{D_k^{\text{comp}}, \sum_{i=1}^N x_{i,k} \cdot (D_{i,k}^{\text{trans}} + D_{i,k}^{\text{comp}})\}$ (s)이다 [3]. $x_{i,k}$ 는 사용자 k -기지국 i 링크가 존재할 시 1, 아닐 시 0 으로 정의된다. 드롭아웃 정의와 활성 사용자 집합은 각각 아래와 같으며,

$$\frac{1}{K} \cdot \sum_{k=1}^K \mathbb{I}(D_k + \nu \geq \tau), \\ K_{\text{act}} = \{k | D_k + \nu < \tau\},$$

$\mathbb{I}(\cdot)$ 은 참일 경우 1, 거짓일 경우 0 을 출력한다. 목표는 활성 사용자 평균 지연 시간과 드롭아웃 비율의 가중합 최소화이다.

$$\underset{\mathbf{w}_{BB,i,k}, \mathbf{w}_{RF,i}, x_{i,k}, \lambda_k, f_{i,k}, P_k}{\text{minimize}} \quad (1 - \omega) \cdot \left(\frac{1}{|K_{\text{act}}|} \right) \cdot \sum_{k \in K_{\text{act}}} D_k$$

$$+ \omega \cdot \frac{1}{K} \cdot \sum_{k=1}^K \delta \cdot \mathbb{I}(D_k + v \geq \tau)$$

$$\begin{aligned} \text{s.t. } & \lambda_k \in [0,1], x_{i,k} \in \{0,1\}, \sum_i x_{i,k} \leq 1, \\ & \sum_k x_{i,k} \leq N_{RF}, \sum_k f_{i,k} \leq F_i^{\max}, \\ & 0 \leq P_k \leq P_k^{\max}, \forall i, k. \end{aligned}$$

여기서 δ 는 단위 보정 팩터, $\omega \in [0,1]$ 는 서비스 정책의 우선순위를 나타내며, ω 가 클수록 드롭아웃 억제를 더 중시한다. 구현에서는 D_k^{comp} 와 $D_{i,k}^{\text{trans}} + D_{i,k}^{\text{comp}}$ 의 균형점을 이용해 λ_k 를 최적화하며, 전력 상한 P_k^{\max} 에 대한 민감도 분석을 수행한다.

위 문제의 최적화 방법으로 다중 기지국이 분산 실행·중앙 평가 형태의 다중에이전트 강화학습(MADRL)으로 상태(작업·채널 요약)에서 행동($P_k, f_{i,k}$)을 산출하고, 보상은 상기 가중합 목적의 음의 값을 반영한다. 구체적으로, 각 기지국이 독립적으로 상태(작업 크기, 채널 품질, 자원 점유도)를 관측하고 분산적으로 행동 후보(송신 전력, 연산 할당량)를 산출한 뒤, 중앙 평가자가 이를 집계하여 기지국 별로 목표함수의 음의 값을 보상으로 부여한다. 연결·빔포밍은 코드북 기반 비용행렬로 추정 지연을 평가해 헝가리안 알고리즘으로 사용자 별로 최적의 기지국을 할당하며, RF 체인 제약은 사후 보정한다. 또한 전력 제약은 정책 신경망 출력에 제약함수 형태로 반영되며 그 밖의 자세한 절차는 [4]에서 다룬다. 이러한 설계는 통신·연산 자원을 동시에 고려하면서도, 분산형 아키텍처 특성상 확장성이 우수하다는 장점이 있다. 본 논문의 환경은 무선 전송률과 로컬 연산 클럭 속도가 사용자 전력 상한에 직접적으로 연동되므로, 전력 제약은 곧 통신·연산 병목을 동시에 유발하는 핵심 요인이다. 따라서, 사용자 전력 상한에 따른 실험 결과를 요약한다.

시뮬레이션 환경은 셀프리 구조에서 9 개의 기지국과 12 개의 사용자 장치가 동일한 영역에 균일 분포로 배치된 상황을 가정하였다. 중심 주파수와 대역폭은 100GHz, 200 MHz 로 설정하였고, 각 사용자 장치의 작업 크기는 [0.1Mbits, 0.3Mbits] 내에서 균일하게 랜덤추출하고, 기지국의 최대 연산 클럭은 $F_i^{\max} = 2 \times 10^{10}$ cycle/s 으로 정하였으며, 슬롯 길이가 마감시간은 $\tau = 20$ ms 이다. 로컬 클럭 주파수는 $\frac{P_k^{\max} - P_k}{\kappa}$ 으로 결정된다. 비교 대상으로는 GA 기반 최적화 기법 [1]과 최근접 기지국 연결을 사용하는 규칙 기반 기법 [2]을 고려하였다.

사용자 전력 상한은 15dBm 에서 40dBm 까지 스윙하였다. 동일 스윙에서의 평균 지연을 비교하면, 제안 기법이 연결·연산 자원을 공동으로 조정하여 지연의 상승폭을 억제한다. 저전력 구간에서 평균 지연 우위가 유지되며, 그림 1 에서 다른 기법 대비 낮게 관찰된다. 그림 2 로부터, 사용자의 전력 상한이 낮아질수록 모든 기법에서 드롭아웃이 증가하지만, 제안 기법은 전송률 저하와 로컬 클럭 저하가 동시 발생하는 저전력 구간에서도 비교 기법 [1],[2] 대비 낮은 드롭아웃을 유지한다. 요약하면, 사용자 전력 상한이 낮아지는 구간에서도 제안 기법은 두 기준 기법 대비 드롭아웃과 평균 지연을 동시에 더 낮게 유지한다. 이는 제안 기법이 열악한 채널 환경에서도 안정적으로 낮은 지연 시간과 드롭아웃 비율을 유지할 수 있음을 시사한다.

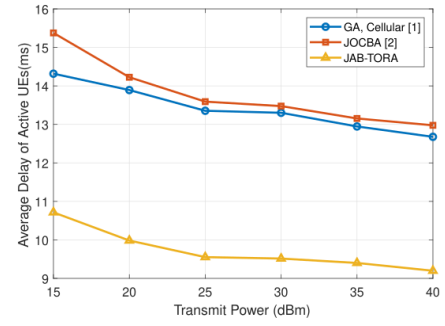


그림 1: 사용자 전력 상한 대비 활성 사용자 평균 지연 시간

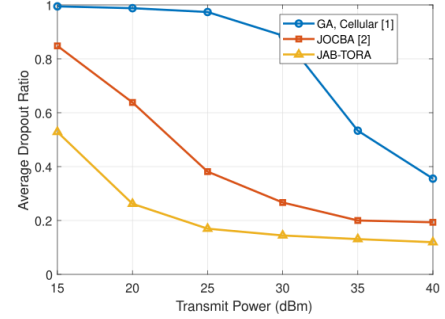


그림 2: 사용자 전력 상한 대비 드롭아웃 비율

III. 결론

마감시간 제약 하의 셀프리 MEC 에서 사용자 전력 상한 변화가 시스템 성능에 미치는 영향을 분석한 결과, 제안 프레임워크는 저전력 구간에서도 강건한 성능을 보였다. 향후에는 전력 상한과 파일럿 오버헤드의 상관관계를 분석하여 더욱 정교한 자원 최적화 방안을 모색할 예정이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2023-00259991).

참 고 문 헌

- [1] H. Song et al., "Joint optimization of edge computing server deployment and user offloading associations via a genetic algorithm," IEEE Transactions on Network Science and Engineering, 2022.
- [2] W. Feng et al., "Joint optimization of computation offloading and bandwidth allocation for delay minimization," IEEE Communications Letters, 2021.
- [3] G. Femenias et al., "Mobile edge computing in cell-free networks," IEEE Transactions on Mobile Computing, 2024.
- [4] N. Hong et al., "Deadline-Aware Joint Optimization of Task Offloading and Resource Allocation for Cell-Free Mobile Edge Computing Networks", in preparation.