엣지 컴퓨팅 환경에서 정보 최신성 기반 작업 할당 전략

김진웅, 이제민*
대구경북과학기술원, *연세대학교

yoy876@dgist.ac.kr, *jemin.lee@yonsei.ac.kr

Age-based Task Splitting Scheme for Edge Computing-enabled Networks

Jinwoong Kim, Jemin Lee*
Department of Electrical Engineering and Computer Science, DGIST,
*Department of Electrical and Electronic Engineering, Yonsei Univ.

요 약

본 논문은 엣지 컴퓨팅 네트워크에서 엣지 서버의 가용 자원 정보에 대한 최신성을 고려하여 오프로딩 성공 확률을 정의하고 이를 수학적으로 분석하였다. 오프로딩 성공 확률을 최대화 문제를 통하여 최적 작업 할당 알고리즘을 제안하고, 제안 알고리즘의 성능을 시뮬레이션 결과를 통하여 검증하였다.

Ⅰ. 서 론

최근 엣지 서버의 강력한 계산 성능을 활용하여 무선 단말의 VR/AR, 자율 주행 등과 같은 계산 집약적인 어플리케이션의 작업을 빠르고 효율적으로 대신 처리해주는 엣지 컴퓨팅 연구가 다양한 분야에서 진행되고 있다 [1]. 유저 요청 작업의 오프로딩 지연을 최소화하기 위하여 다수의 엣지 서버를 활용할 수 있는데 [2], [3], 이 때 각 엣지 서버로의 최적의 작업 할당을 위해서는 엣지 서버에 대한 가용 자원 정보와 무선 채널 파라미터에 대한 정확한 정보 획득이 요구된다 [4]. 지금까지의 기존 논문에서 엣지 서버 가용 자원 정보의 최신성을 고려한 작업 할당 연구는 없었는데, 본 논문에서는 엣지 서버로부터 전달받는 자원 정보와 이에 대한 나이를 함께 고려한 최적 작업 할당 알고리즘을 제안함으로써 본 논문에서 정의 및 유도한 오프로딩 성공 확률을 향상시켰다.

Ⅱ. 시스템 모델

2.1. 네트워크 모델

본 논문은 유저의 작업이 M개의 엣지 서버에 무선으로 분할 전송되고, 엣지 서버는 분할된 작업을 대신 처리하여 결과를 반환하는 엣지 컴퓨팅 네트워크를 고려하였다. 각 엣지 서버는 유저에게 제공 가능한 자원 정보, 즉, CPU frequency 정보를 컨트롤 타워에 주기적으로 전달한다. 컨트롤 타워는 유저 요청 시 여러 엣지 서버의 가용 자원 정보 및 무선 채널 정보를 바탕으로 각 엣지 서버로의 작업 할당 비율을 최적화한 후 그 결과를 유저에게 전달하고, 유저는 그 최적 할당 비율로 각 엣지 서버에 작업을 전송한다. 컨트롤타워가 엣지 서버로부터 수신한 정보는 수신 시점으로부터 실제 엣지 서버에서의 정보와 오차가 존재하게 되는데, 이는 다음과 같이 표현된다.

$$I_m(t) = \hat{I}_m(t) + \sigma_m(t),\ \ m \in \mathcal{M} = \{1, \dots, M\} \qquad (1)$$

여기서 $I_m(t)$는 실제 엣지 서버에서 시간 t에서의 가용 자원 정보이고, $\hat{I}_m(t)$는 컨트롤 타워가 엣지 서버로부터 가장 최근 수신한 자원 정보, $\sigma_m(t)$는 실제 정보와 수신 정보 사이의 오차이다. 정보 오차 $\sigma_m(t)$ 는 수신 시점으로부터 시간, 즉 정보의 나이(Age of Information, AoI)의 증가에 따라 크기가 증가할 수 있는데, 이는 정보 나이에 따라 분산이 증가하는 Wiener process 로 모델링 될 수 있다. 이를 이용하여, $\hat{I}_m(t)$가 주어졌을 때 실제 정보 $I_m(t)$의 확률 밀도 함수는 다음과 같이 표현된다.

$$f_{I_m(t)}\left(x|\hat{I}_m(t)\right) = \frac{\exp\left(-\frac{\left(x-\hat{I}_m(t)\right)^2}{2\xi\Delta_m(t)}\right)}{n_m\sqrt{2\pi\xi\Delta_m(t)}} \qquad (2)$$

여기서 $\Delta_m(t)$ 는 m 번째 엣지 서버의 AoI, ξ 는 AoI 스케일링 파라미터, n_m은 normalization 계수이다.

2.2. 작업 할당 모델

유저 요청 작업의 크기는 D 이고, 각 엣지 서버로 할당되는 작업의 비율은 a_m , $0 \le a_m \le 1, \forall m \in \mathcal{M} = \{1, \dots, M\}$으로 정의되며 $\sum_{m=1}^{M} a_m = 1$을 만족한다.

2.3. 오프로딩 성공 확률

유저에서 m번째 엣지 서버로의 무선 전송률은 다음과 같이 표현된다.

$$C_m = W \log_2\left(1 + \frac{P|h_m|^2}{N}\right) \qquad (3)$$

여기서 W는 채널 대역폭, P는 유저의 전송 파워, h_m은 유저로부터 m 번째 엣지 서버로의 채널 이득, N 은 additive white Gaussian noise (AWGN)의 파워이다. h_m 은 Rayleigh block fading 을 겪는다고 가정한다. 이를 통해 유저에서 m번째 엣지 서버로의 전송 성공 확률 $p_{t,m}$ 은 무선 전송률이 타겟 전송률 이상을 만족할 확률로 정의할 수 있고, 이는 다음과 같이 표현된다.

$$p_{t,m} = \mathbb{P}\left[C_m > \frac{Da_m}{T_t}\right]. \qquad (4)$$

여기서 T_t는 타겟 전송 시간이다. m번째 엣지 서버가 할당 작업을 수행하는데 걸리는 시간 $T_{\mathrm{comp},m}$은 다음과 같

이 표현된다.

$$T_{\mathrm{comp},m} = \frac{Da_m c}{I_m(t)} \quad (5)$$

여기서 c 는 1[bit]를 처리하는데 필요한 CPU cycle 수이다. 모든 엣지 서버로부터 작업 결과를 반환 받아야 요청 작업이 완료되므로, 이를 고려한 계산 성공 확률 p_{c}는 다음과 같이 표현된다.

$$p_{\mathrm{c}} = \mathbb{P}\left[\max_{m\in\mathcal{M}}\{T_{\mathrm{comp},m}\} < T_{\mathrm{c}}\right] \quad (6)$$

여기서 T_{c} 는 타겟 계산 시간이다. 최종적으로 각 엣지 서버로의 전송 성공 확률과 계산 성공 확률을 종합한 오프로딩 성공 확률은 다음과 같이 유도된다.

$$\begin{aligned} p_{\mathrm{s}}(\boldsymbol{a}) &= \mathbb{P}\left[\max_{m\in\mathcal{M}}\{T_{\mathrm{comp},m}\} < T_{\mathrm{c}}\right]\prod_{m=1}^{M}\mathbb{P}\left[C_m > \frac{Da_m}{T_t}\right] \quad (7)\\ &= \prod_{m=1}^{M}\exp\left(-\left(2^{\frac{Da_m}{T_t W}}-1\right)\frac{N}{Pd^{-\alpha}}\right)\\ &\quad\times\frac{1}{n_k}Q\left(\frac{1}{\sqrt{\xi\Delta_m(t)}}\left(\frac{Da_m c}{T_{\mathrm{c}}}-\hat{I}_m(t)\right)\right) \end{aligned}$$

여기서 $Q(x)=\frac{1}{2\pi}\int_x^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)du$이다.

Ⅲ. 최적 작업 할당 분석

각 엣지 서버에 대한 작업 할당을 통한 오프로딩 성공 확률 최대화 문제는 다음과 같이 표현된다.

$$\begin{aligned} \max_{\boldsymbol{a}} \quad & p_{\mathrm{s}}(\boldsymbol{a}) \quad (8)\\ \text{s.t.} \quad & \sum_{m=1}^{M} a_m = 1,\\ & a_m \in [0,1], \forall m \in \mathcal{M}. \end{aligned}$$

여기서 목적 함수 $p_{\mathrm{s}}(\boldsymbol{a})$는 log-concavity 를 만족한다. 목적함수에 로그 및 -를 취한 후의 Lagrange 함수는 다음과 같이 표현된다.

$$\mathcal{L}(\boldsymbol{a},\nu) = -\ln p_{\mathrm{s}}(\boldsymbol{a}) + \nu\left(\sum_{m=1}^{M} a_m - 1\right) \quad (9)$$

여기서 ν는 Lagrange 계수이다. Karush-Kuhn-Tucker (KKT) 조건으로부터 $\frac{\mathrm{d}\mathcal{L}(\boldsymbol{a},\nu)}{\mathrm{d}a_m}=0, \forall m\in\mathcal{M}$ 를 만족하는 해 $a_m, \forall m\in\mathcal{M}$ 와 $\sum_{m=1}^{M} a_m = 1$을 만족하는 ν를 얻음으로써 최적 작업 할당 비율 $\boldsymbol{a}^*$를 도출할 수 있다.

Ⅳ. 시뮬레이션 결과

본 논문에서 제안한 알고리즘의 성능을 분석하기 위하여 시뮬레이션에 사용된 파라미터 값들은 $D = 50\mathrm{Mb}$, $T_{\mathrm{c}} = 1.2\mathrm{s}, T_t = 25\mathrm{ms}$, $\mathrm{I}_{\max} = 20000$, $\xi = [7000, 20000]$이다.

그림 1 은 AoI 스케일링 파라미터 ξ에 따른 오프로딩 성공 확률을 가용 자원 정보의 최신성을 고려하지 않고 작업을 할당한 베이스라인과 비교 분석한 그래프이다. 먼저, 모든 ξ 에서 정보 최신성을 고려한 작업 할당 기법이 베이스라인에 비하여 더 높은 성능을 보여줌을 알 수 있다. 또한, ξ 가 증가함에 따라 오프로딩 성공 확률이 감소하는데, 이는 AoI 에 따른 정보 오차 영향이 더욱 증가하여 정확한 작업 할당이 어렵기 때문이다.

Ⅴ. 결론

본 논문에서는 엣지 컴퓨팅 네트워크에서 엣지 서버의 가용 자원 정보에 대한 최신성을 고려한 유저의 오프로딩 성공 확률을 정의하고, 이를 최대화하기 위한 작업 할당 문제를 공식화하였다. 시뮬레이션 결과를 통하여 가용 자원 정보의 나이를 고려한 작업 할당이

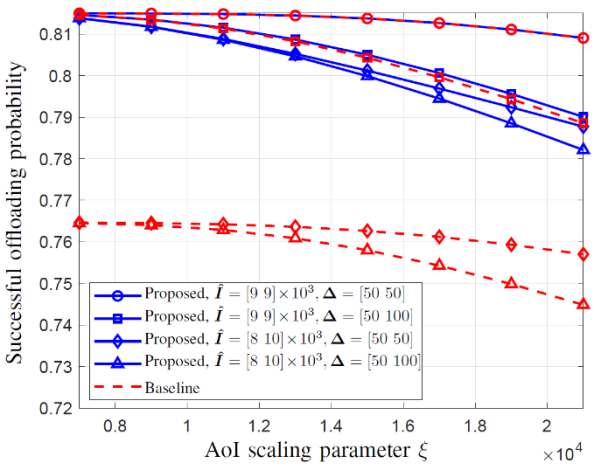

그림 1 AoI scaling parameter ξ에 따른 오프로딩 성공 확률

이를 고려하지 않은 작업 할당 기법보다 더 나은 성능을 보임을 확인하였다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2020R1A2C2008878, NRF-2018R1A5A1060031).

참 고 문 헌

[1] X. Kong, Y. Wu, H. Wang and F. Xia, "Edge Computing for Internet of Everything: A Survey," in IEEE Internet of Things Journal, vol. 9, no. 23, pp. 23472-23485, 1 Dec.1, 2022.

[2] Z. Jing, Q. Yang, M. Qin, J. Li and K. S. Kwak, "Long-Term Max-Min Fairness Guarantee Mechanism for Integrated Multi-RAT and MEC Networks," in IEEE Transactions on Vehicular Technology, vol. 70, no. 3, pp. 2478-2492, March 2021.

[3] T. Do-Duy, D. Van Huynh, O. A. Dobre, B. Canberk and T. Q. Duong, "Digital Twin-Aided Intelligent Offloading With Edge Selection in Mobile Edge Computing," in IEEE Wireless Communications Letters, vol. 11, no. 4, pp. 806-810, April 2022

[4] Y. Sun, Y. Polyanskiy and E. Uysal, "Sampling of the Wiener Process for Remote Estimation Over a Channel With Random Delay," in IEEE Transactions on Information Theory, vol. 66, no. 2, pp. 1118-1135, Feb. 2020.