Server Provisioning and User Association in Multi-Site Multi-access Edge Computing

Minh N. H. Nguyen, Nguyen H. Tran and Choong Seon Hong
Department of Computer Science and Engineering, Kyung Hee University
Email: {minhnhn, nguyenth, cshong}@khu.ac.kr

Abstract: Multi-access Edge Computing (MEC) is recently profound for the next revolution of mobile communications area, where the convergence of IT and telecommunications networking provides lower latency and more computation capability for cellular base stations. In this paper, we formulate a novel optimization problem for server provisioning and user association to enhance the performance of the network operator in terms of energy cost and the average delay of mobile users. The problem is relaxed and approximately solved by an iterative algorithm and converge to a suboptimal solution of the original problem.

1. Introduction

Nowadays, the proliferation of mobile devices and less expensive renewable energy installation cost at the base stations (BSs) have opened a new research area, i.e., green communications. Since the base stations can be integrated with edge servers to perform local processing tasks and boost up the response time for user compared to the original cloud architecture. In this work, we consider the power sources of the system are electricity grid and renewable energy sources that are controlled by a centralized network controller.

Different from the paper [1], we consider a multi-site environment with a centralized network controller. In the related model [2] for the ultra-dense context, the author focuses on energy-aware mobility management while our formulation is dedicated to the user association and server provisioning management. To the best of our knowledge, the most relevant model [3] proposes the load balancing scheme for a network of MEC-enabled base stations and do not consider server provisioning and user association as in our work. In this paper, we propose an iterative algorithm for the server provisioning and user association problem to enhance the delay performance and reduce the energy cost of the operated BSs in a region. Accordingly, the network controller will decide which BSs mobile users can be associated with and the number of active servers for each site.

2. Server Provisioning and User Association Problem

A. System Model

In the considered region, we introduce a centralized network operator manipulate a network controller for its systems of BSs as in Fig. 1. At each site in the BSs set $\mathcal{S}$, BS receives power from the common renewable energy sources throughout network controller. The mobile devices in the region can be associated with one of BSs with different transmission rates. In this system model, we consider a single network operator who can control the number of active edge servers at each site, the user association decision among sites based on the traffic load condition, renewable energy procurement in the certain period. According to user association and the number of active edge servers, the network controller can purchase an extra amount of energy from electricity grid. Therefore, the network operator can have two roles such as *energy management* and *MEC management*.

We consider virtual user *u*, who has the traffic follows the inhomogeneous Poisson point process with arrival rate per unit area arrivals rate $\lambda(u)$. For simplicity, the arrival traffics can be modeled as user flows (i.e., data requests) with random sizes following independent distribution with mean $1/\upsilon(u)$. Then, the traffic load density function can be modeled as $\gamma(u) = \lambda(u)/\upsilon(u)$.

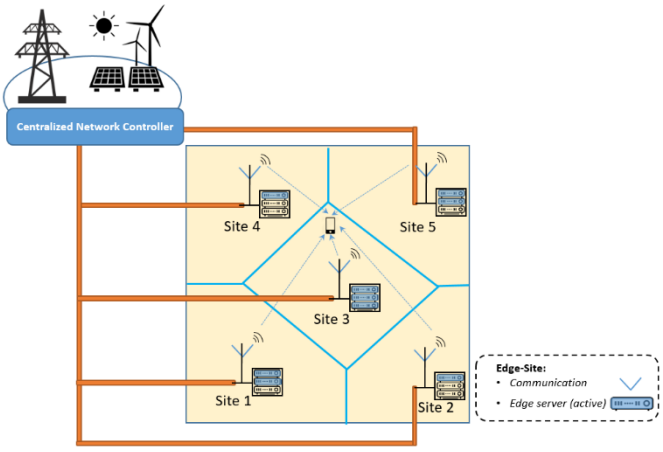


Fig. 1: Centralized network controller system.

For each user *u*, the downlink transmission rate can be served by BS *j* is denoted by $c_j(u)$ which follows Shannon capacity as in [4].

System load density of the BS *j* is $\beta_j(u) = \gamma(u)/c_j(u)$, which defines the fraction of active transmission time required to deliver the traffic load $\gamma(u)$ from BS *j* for user *u*.

Definition 1 (Feasibility): The set $\mathcal{F}$ of feasible BS *loads* (or *utilization*) as in [4], [5], with the association probability variable $p(u)$, is defined as

$$\mathcal{F}=\Big\{\rho|\,\rho_j=\sum_{u\in\mathcal{U}}\beta_j(u)p_j(u),0\le\rho_j\le1-\epsilon,$$
$$0\le p_j(u)\le1,\sum_{j\in\mathcal{S}}p_j(u)=1,\forall j\in\mathcal{S},\forall u\in\mathcal{U}\Big\},$$

where ϵ is an arbitrarily small positive constant, which is the convex set as in [4].

Flow cost:

Based on the queueing analysis [4] with the M/GI/1 multi-class processor sharing system, the expected total number of flows of BS *j* as $L_j=\frac{\rho_j}{1-\rho_j}$. We adopt the flow level cost at each site *j* is $\phi_c(\rho_j)=L_j+1=\frac{1}{1-\rho_j}$.

The delay optimal strategy of the network operator for all BSs is equivalent to minimize the following cost

$$\phi_c(\rho)=\sum_{j\in\mathcal{S}}\phi_c(\rho_j)=\sum_{j\in\mathcal{S}}\frac{1}{1-\rho_j} \quad (1)$$

Computational delay:

According to M/M/1queueing model as in [6], the average response time of edge servers at each site *j* is as $\phi_p(m_j)=\frac{1}{\mu_j-\frac{\lambda_j}{m_j}}$, where m_j is the number of active servers, μ_j is the service rate of an edge server and λ_j is the arrival server loads based on the arrival traffic loads to the BS *j*.

The aggregate computational delay of all BSs as

$$\phi_p(m)=\sum_{j\in\mathcal{S}}\phi_p(m_j) \quad (2)$$

Power and Energy model:

The power consumption of BS *j* is followed a linear model as in [5] as

$$\psi_j(\rho_j)=(1-\alpha_j)\rho_jQ_j+\alpha_jQ_j. \quad (3)$$

where Q_j is the maximum power of BS *j*. Thus, in the considered control period Δ_t, the BS energy usage is $\mathcal{E}_c(\rho_j)=\psi_j(\rho_j)\times\Delta_t$.

The linear power model of edge server at site *j* in [6] as

$$P(m_j)=m_j\Big(p_{j,s}+p_{j,a}\times\frac{\lambda_j}{m_j\times\mu_j}\Big)PUE, \quad (4)$$

where μ_j is the service rate, *PUE* is the power usage effectiveness of an edge server at site *j*, $p_{j,s}$ is the static power and $p_{j,a}$ is the active server based on the server loads. Thus, in the considered control period Δ_t, the edge server energy usage is $\mathcal{E}_p(m_j)=P(m_j)\times\Delta_t$.

Energy usage from electricity grid follows

$$\mathcal{E}_{grid}(m,p)=\sum_{j\in\mathcal{S}}\Big(\mathcal{E}_c(\rho_j)+\mathcal{E}_p(m_j)\Big)-\mathcal{E}_{gen}, \quad (5)$$

where $\mathcal{E}_{gen}$ is the renewable energy procurement.

B. Problem Formulation

In this paper, our objectives comprise delay performance with price γ_1 and energy cost with electricity price p_{grid}. Therefore, we propose the formulation of *Server Provisioning and User Association problem* as follows

$$\min_{m,p}\quad \gamma_1\times(\phi_c(\rho)+\phi_p(m))+p_{grid}\times\mathcal{E}_{grid}(m,p), \quad (6)$$
$$\textbf{s.t.}\quad m_j\in\{1,..,M_j\},\forall j\in\mathcal{S}, \quad (7)$$
$$\rho\in\mathcal{F}. \quad (8)$$

The variables of this problem are the number of active server m_j and the user association decision p. The BS loads ρ is the function of p according to the definition. This problem is MIP problem, which cannot be solved directly by solvers. Therefore, in the next section we propose an iterative algorithm for approximately solving the relaxed problem.

3. Solution Approach

Since this problem is NP-Hard problem, we first relax the integer variable (i.e., the number of active servers) to the real variables in constraint (7). After receiving the solution, the number of active servers is rounded to integer values. The relaxed problem is in the multi-convex form [7], which can be approximately solved by an alternative algorithm. Accordingly, we iteratively solve the *User Association problem* with the given number of active servers to find the optimal user association decision. Then based on the given user association decision, we update the number of active servers by solving *Server Provisioning problem*. The whole process is repeated until it achieves the convergence condition in the Alg.1. These problems are solved by using off-the-self solver (i.e., ECOS [8]).

For each iteration *k*,

User Association problem

$$\min_{p}\quad \gamma_1\times(\phi_c(\rho)+\phi_p(m_{k-1}))+p_{grid}\times\mathcal{E}_{grid}(m_{k-1},p),$$
$$\textbf{s.t.}\quad \rho\in\mathcal{F}.$$

Server Provisioning problem

$$\min_{m}\quad \gamma_1\times(\phi_c(\rho_k)+\phi_p(m))+p_{grid}\times\mathcal{E}_{grid}(m,p_k),$$
$$\textbf{s.t.}\quad 1\le m_j\le M_j,\forall j\in\mathcal{S}.$$

Algorithm 1

1: **Initialization:** Initialize $k=0,\epsilon$;
2: **repeat**
3: Compute $p_{(k)},\rho_{(k)}$ from **User Association problem**;
4: Compute $m_{(k)}$ from **Server Provisioning problem**;
5: **until** $\|m_{(k)}-m_{(k-1)})\|\le\epsilon$.

4. Simulation Results

For an example scenario, we consider a system of five sites. These sites are located in a 1 × 1 km^2 region, the generated traffic loads come from 100 virtual groups of users. According to the communication model of urban macrocells with the simulation parameters in the WiMAX evaluation methodology

document [10], we use the COST 231 path loss model. The electricity price is 0.07134 $/kWh [9] and cost factor of delay performance is 300 $/unit. The maximum BS power Q is 865 Watts [4] and the portion of static power is 0.2. Static power of edge servers is 200 Watts, and active power is 400 Watts, PUE is 1.5 [6]. The convergence condition threshold is set to 10^{-9}. The maximum number of servers can be activated is 5.

In Fig. 2, we obtain the converged solution within six iterations. Since BS_3 is in the center of the region where the traffic generation is higher than other BSs, it requires more active servers to process data.

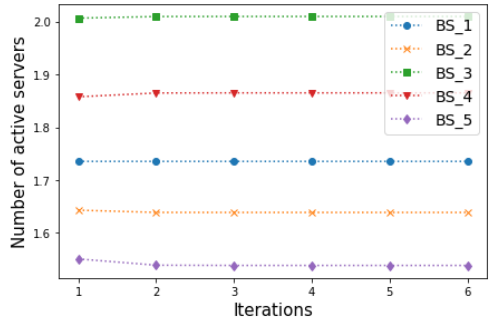
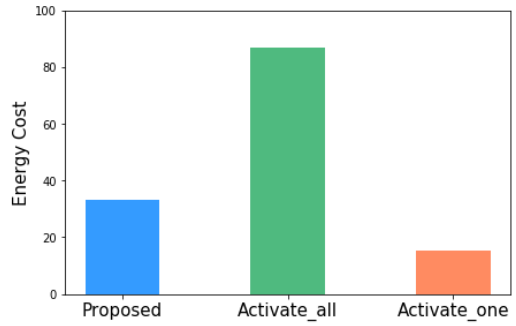
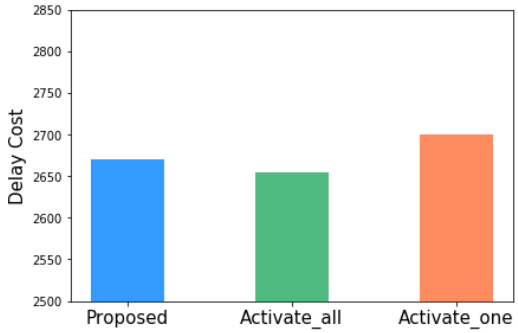


Fig. 2: The convergence of the proposed algorithm.



(a) Energy cost comparison



(b) Delay cost comparison

Fig. 3: Energy cost and delay cost of algorithms.

For the comparison, we compare our approach with two other strategies such as activate all of the edge servers and activate just one edge server at each site. The proposed algorithm shows the trade-off between energy cost and delay cost in Fig. 3. Our proposed algorithm receives less energy cost and higher delay cost compared to the strategy of activating all edge servers. On the other hand, our proposed algorithm obtains the better delay performance and uses more energy compared to the strategy of activating only one edge server. In conclusion, our proposed algorithm receives the least overall cost in terms of delay performance and energy cost.

5. Conclusion & Future Work

In this paper, we investigate an efficient approach for the user association and server provisioning problem of a centralized network controller in Mobile Edge Computing. The proposed algorithm can help to reduce energy cost and enhance delay performance. In the future work, we advocate extending the work for multiple time slots management regarding the battery management and long-term costs.

Acknowledgement

This work was supported by the Industrial Strategic Technology Development Program(10067093, Development of the Communication Standards and the Regulations of Electrical safety for the Portable Charger) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea). *Dr. CS Hong is the corresponding author

References:

[1] Xu, Jie, Lixing Chen, and Shaolei Ren. "Online learning for offloading and autoscaling in energy harvesting mobile edge computing." *IEEE Transactions on Cognitive Communications and Networking*, vol. 3.3, pp. 361-373, 2017.

[2] Sun, Yuxuan, Sheng Zhou, and Jie Xu. "EMM: Energy-aware mobility management for mobile edge computing in ultra dense networks." *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2637-2646, 2017.

[3] Xu, Jie, et al. "Online Geographical Load Balancing for Mobile Edge Computing with Energy Harvesting." *arXiv preprint*, arXiv:1704.00107 (2017).

[4] H. Kim, G. De Veciana, X. Yang, and M. Venkatachalam, "Distributed optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 177–190, 2012.

[5] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE journal on selected areas in communications*, vol. 29, no. 8, pp. 1525–1536, 2011.

[6] Tran, Nguyen, Chuan Pham, Minh Nguyen, Shaolei Ren, and Choong Seon Hong. "Incentivizing Energy Reduction for Emergency Demand Response in Multi-Tenant Mixed-Use Buildings." *IEEE Transactions on Smart Grid*, vol. pp, 2016.

[7] Shen, Xinyue, Steven Diamond, Madeleine Udell, Yuantao Gu, and Stephen Boyd. "Disciplined multi-convex programming." *In Control And Decision Conference (CCDC)*, 2017 29th Chinese, pp. 895-900, 2017.

[8] Domahidi, Alexander, Eric Chu, and Stephen Boyd. "ECOS: An SOCP solver for embedded systems." Control Conference (ECC), 2013 European. IEEE, 2013.

[9] R. Srinivasan, J. Zhuang, L. Jalloul, R. Novak, and J. Park, "Ieee 802.16m evaluation methodology document (emd)," *IEEE 802.16 Broadband Wireless Access Working Group*, 2008.

[10] "City of Austin FY 2017 Electric Tariff," Austin City Council, Tech. Rep., 2016. [Online]. Available: https://austinenergy.com/wps/wcm/connect/c4f3dd41-c714-43bc-9279-e92940094731/FY2017aeElectricRateSchedule.pdf?MOD=AJPERES