Distributed Caching for Cellular Networks: A Dual Decomposition Methods

Tri Nguyen Dang, Choong Seon Hong

Department of Computer Science and Engineering
Kyung Hee University, 446-701, Korea
email: {trind, cshong}@khu.ac.kr

Abstract

Entertainment services have become more popularity and necessity in recent years. The double-digit grow up of the Internet of Things, smart devices have brought out various types of data traffic, computing demand, and entertainment content requested to the Multi-access Edge Computing (MEC) server. Due to the limit of physical resources of the MEC server, and high demand from users. The MEC may face congestion and overload problems. MEC has to cache the contents, computational results aim to reduce the delay, computational resources. However, MEC has a limited cache capacity. In this work, we consider a system model that allowed MEC can cache the contents at user equipment (UE) to increase the number of cached content and improve the Quality of Services (QoS). We formulated as a combinatorial optimization problem and using dual decomposition methods to perform the near-optimal solution.

1 INTRODUCTION

The rise of the fifth generation of the cellular network (5G) has brought out various types of real-time applications. These kinds of apps require more resources, low delay, and high power consumed. Due to those reasons, the core-network might face the congestion, the Multi-access Edge Computing (MEC) server might be overloaded, the user's experience is reduced, and the quality of services is also reduced. In order to avoid those problems, many researchers has address the problem of computational offloading [1] , content caching [2], [3]; caching at the edge [4], [5]; resource sharing [6]. However, most of these work are considered the content only cache at the MEC server. Due to the limit on cache capacity, the MEC server can-not cache whole set of contents, and the popularity of content is only in a specific location. Thus, some of the content may get zero utility if it is cached at MEC server. Therefore, we propose an approach that allowed MBS utilize the resource of UE to cache the contents. Based on the popularity and request rate at the SBS, MBS need to make a decision whether to cache a content at UE or MEC server to maximize total utility. By utilizing resource of UE, MBS can increase number of cached contents. The QoS, and utility are direct proportion to the number of cached contents. Thus, our proposal framework can improve the QoS, and user experience. On the other hand, content is cached at the UE, thus, MBS can utilize the device to device (D2D) communication method to serve the other user's demand. It is also reduce the network congestion by reduce data traffic between MBS and users.

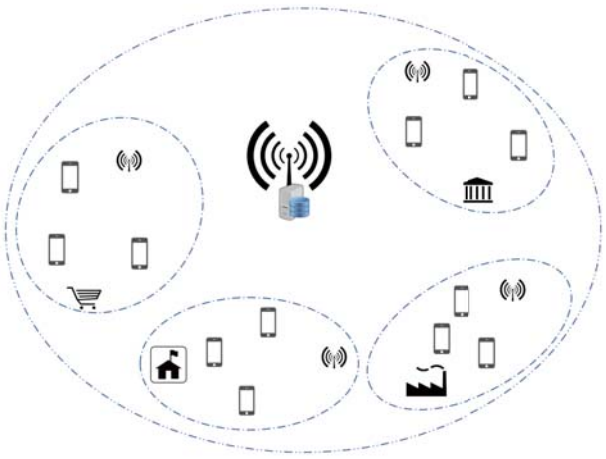

Fig. 1: Illustration of our system model.

2 SYSTEM MODEL

In this paper, we consider a network area consisting a single micro cell base station (MBS) equipped MEC capability, a set of M small cell base stations (SBS) allocated in different locations, each small cell $m \in M$ has a set of N_m users, and a set of K_m contents that frequently requested by users. Let $\mathcal{N}_m = \{1, 2, \ldots, N_n\}$, and $\mathcal{K}_m = \{1, 2, \ldots, K_m\}$ denoted the set of users, and the computing tasks at SBS m, respectively. Generally, whenever a computing task come, BS need to accomplish and transmit

the results to UE. With the large number of demand from UEs, BS may faced the problem of congestion, out of resources, or overloaded. Therefore, BS need to cache the results of the task aims to save the processing time, and transmit density.

3 Problem Formulation

Let p_{mk} be the probability of a request for content k at SBS m, p_{mk} can be modeled follow the Zipf distribution with parameter α_m. In this model, we assume that the request rate is follow the Poisson distribution with parameter λ, λ_m represent for the request rate at SBS m. Thus, the expected utility of cache a content can be formulated as follow:

$$u_{mk} = Pr(k|\lambda_m) p_{mk} \log(s_{mk}), \tag{1}$$

where $Pr(\lambda_m)$ is the probability of a request at SBS m, s_{mk} is the size of content. Furthermore, we assume that the set of contents are disjoint with each other $\mathcal{K}_m \cap \mathcal{K}_n = \emptyset, \forall m, n \in M, m \neq n$. Based on that, we can formulate the problem of on-device caching as follow:

$$\begin{aligned}
\text{A1} : \max_{x} & \sum_{m=1}^{M} \sum_{k=1}^{K_m} u_{mk} \left(y^{mk} + \sum_{i=1}^{N_m} x_i^{mk} \right) \\
\text{s.t.} & \\
& \text{C1} : \sum_{k=1}^{K_m} s_{mk} x_i^{mk} \leq s_i, \forall m \in M, \forall i \in N_m, \\
& \text{C2} : \sum_{i=1}^{N_m} x_i^{mk} \leq 1, \forall k \in K_m, \forall m \in M, \\
& \text{C3} : \sum_{m=1}^{M} \sum_{k=1}^{K} s_{mk} y^{mk} \leq S, \\
& \text{C4} : y^{mk} + \sum_{i=1}^{N_m} x_i^{mk} \leq 1, \forall k \in K_m, \forall m \in M, \\
& \text{C5} : x_i^{mk} = \{0, 1\}, y^{mk} = \{0, 1\}, \\
& \forall m \in M, \forall k \in K_m, \forall i \in N_m.
\end{aligned} \tag{2}$$

The objective is maximize total utility of the MBS by cache content at both sides: MEC storage, and UE's devices. Constraint (C1), (C3) represent that total size of cached content must be non exceed the storage capacity of UE's device, and MEC, respectively. Constraint (C2) is guarantee for non-duplicate caching among UEs. And, constrain (C4), represent for non-duplicate caching between MBS and UEs. The decision variables are binary. Due to (C5) the problem become NP-hard, thus, we can-not archive any feasible solution in polynomial time. We then propose a method of fractional caching by relax two binary variables into two continuous variables. The original problem is equivalent as follow

$$\begin{aligned}
\text{A2} : \max_{x,y} & \sum_{m=1}^{M} \sum_{k=1}^{K_m} u_{mk} \left(y^{mk} + \sum_{i=1}^{N_m} x_i^{mk} \right) \\
& \text{s.t.}(C1) - C4) \\
& \quad \text{C6} : 0 \leq x_i^{mk}, y^{mk} \leq 1, \\
& \quad \forall m \in M, \forall k \in K_m, \forall i \in N_m.
\end{aligned} \tag{3}$$

In the problem (A2) We till have the coupling constraint (C4) which reflex hardness to get the solutions. We then use the Decomposition method (DM) [7] to address this issue. By following DM, we introduce a new term γ such that $0 \leq \gamma \leq 1$, and separate the problem (A2) into two sub-problems (B1) and (B2) as follow:

$$\begin{aligned}
\text{B1} : \max_{x} & \sum_{m=1}^{M} \sum_{k=1}^{K_m} u_{mk} \left(y^{mk} + \sum_{i=1}^{N_m} x_i^{mk} \right) && \text{(4a)} \\
& \text{s.t.}(C1), (C2) && \text{(4b)} \\
& \quad \text{C6} : \sum_{i=1}^{N_m} x_i^{mk} \leq \gamma^{mk}, && \text{(4c)} \\
& \quad \text{C7} : 0 \leq x_i^{mk} \leq 1, && \text{(4d)} \\
& \quad \forall m \in M, \forall i \in N_m, \forall k \in K_m. && \text{(4e)}
\end{aligned}$$

And, the second sub-problem:

$$\begin{aligned}
\text{B2} : \max_{y} & \sum_{m=1}^{M} \sum_{k=1}^{K_m} u_{mk} \left(y^{mk} + \sum_{i=1}^{N_m} x_i^{mk} \right) && \text{(5a)} \\
& \text{s.t.}(C3) && \text{(5b)} \\
& \quad \text{C8} : y^{mk} \leq 1 - \gamma^{mk}, && \text{(5c)} \\
& \quad \text{C9} : 0 \leq y^{mk} \leq 1, && \text{(5d)} \\
& \quad \forall m \in M, \forall i \in N_m, \forall k \in K_m. && \text{(5e)}
\end{aligned}$$

Let $\theta_1(\gamma)$ and $\theta_2(\gamma)$ denote the optimal value of two sub-problems B1 and B2, respectively.

$$\begin{aligned}
\theta_1(\gamma) &= \inf_{x} \left\{ (4a) \mid (C1), (C2), (C6), (C7) \right\} \\
\theta_2(\gamma) &= \inf_{y} \left\{ (5a) \mid (C3), (C8), (C9) \right\}
\end{aligned} \tag{6}$$

We can define the master problem as follow

$$\min_{\gamma} \theta_1(\gamma) + \theta_2(\gamma) \tag{7}$$

Furthermore, the problem (A2) is equivalent to the master problem. Then, we follow the DM framework to solve the problem (A2) in Alg. 1.

4 Simulation

To evaluate our system problem, we use Julia language as our simulation tool. For the simulation setup, number of SBS $M = 10$, number of user per SBS $K_m = 100$, number of content per SBS $K_m = 1000$, the content's size s_{mk} ranging from $1000(GB)$ to $5000(GB)$, UE's capacity is various between $\{1000, 5000\}(GB)$, and SBS capacity less than 40% of total content size. The Fig. 2, we

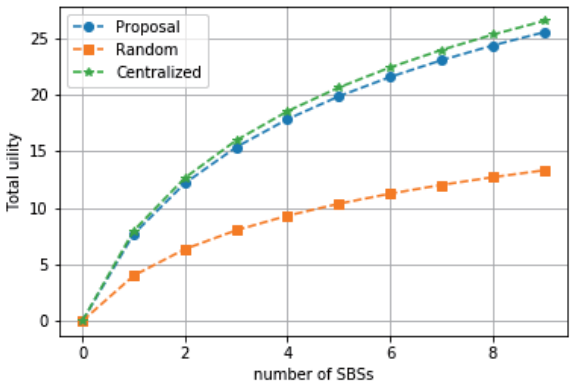

Fig. 2: Total utility vs. number of UEs.

compare our scheme with the centralized solution, and random selection algorithm. We can see that our approach has out perform the random selection and approximate to the centralized solution.

5 Conclusion

In this paper, we proposed a distributed caching network model, ad formulated as an combinatorial optimization problem. We then apply the dual decomposition method to solve the relaxed problem. Our result has shown the best performance than random algorithm and close to the centralized method.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2019-2015-0-00742) supervised by the IITP(Institute for Information communications Technology Promotion)" *Dr. CS Hong is the corresponding author

Algorithm 1 Algorithm distributed caching model

Input: $\mathcal{M}, \mathcal{N}, \mathcal{K}, \lambda, \mathbf{p}$
Output: Optimum utility

1: Initialize γ
2: **for** $t \in T$ **do**
3: Solve two sub-problems (B1), (B2) in parallelly manner.
4: $\theta_1(\gamma) = \inf_x \{(4a) \mid (C1), (C2), (C6), (C7)\}$
5: Find the sub-gradient of (B1): $\nabla_1 \in \partial\theta_1(\gamma)$
6: $\theta_2(\gamma) = \inf_y \{(5a) \mid (C3), (C8), (C9)\}$
7: Find the sub-gradient of (B2): $\nabla_2 \in \partial\theta_2(\gamma)$
8: Update $\gamma = \gamma + \rho_t(\nabla_1 + \nabla_2)$
9: **end for**
10: **return** P

References

[1] T. N. Dang and C. S. Hong, "A distributed admm approach for data offloading in fog computing," , pp. 1057–1059, 2016.

[2] S. Ullah, L. U. Khan, and C. S. Hong, "Socially-aware svc video caching in information centric networking with d2d communication," , pp. 1389–1391, 2018.

[3] A. Ndikumana, S. Ullah, and C. S. H. Do Hyeon Kim, "Deepauc: Joint deep learning and auction for congestion-aware caching in named data networking," *PloS one*, vol. 14, no. 8, 2019.

[4] K. Thar, S. Ullah, R. Haw, T. LeAnh, T. Z. Oo, and C. S. Hong, "Hybrid caching and requests forwarding in information centric networking," in *2015 17th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, 2015, pp. 203–208.

[5] K. Thar, N. H. Tran, S. Ullah, T. Z. Oo, and C. S. Hong, "Online caching and cooperative forwarding in information centric networking," *IEEE Access*, vol. 6, pp. 59 679–59 694, 2018.

[6] T. N. Dang and C. S. Hong, "Utility maximization for resource sharing in mobile edge computing," , pp. 389–391, 2018.

[7] C. C. CarøE and R. Schultz, "Dual decomposition in stochastic integer programming," *Operations Research Letters*, vol. 24, no. 1-2, pp. 37–45, 1999.