Decentralized Computational Caching for X-Reality Services in 5G and Beyond Networks

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

The Fifth Generation of mobile network (5G) has been introduced the current year which various advantages compared with the previous generations (4G, 3G, etc). The ultimate goals of 5G are providing high bandwidth, low latency, massive connectivity, network reliability, and energy efficiency. These advantages have completely changed the entire of the applications in networks such as a massive increase in quality of video up to 4K, Virtual/Augmented/Mixed Reality (XR). However, the more proliferation of those applications the more computation capability required, and massive data traffic in the networks posed a problem of network congestion and overutilization of Multi-access Edge Server (MEC). Therefore, we propose a network model in which computational results are partial or completely cached at either user's devices or the MEC server. Our objective is saving the computational resources of MEC by caching the computational results of XR's tasks and increase the spectrum utilization by adopting the device-to-device(D2D) communication technique to serve the demand of the other convincing devices. We formulate an optimization problem and apply the primal-dual decomposition approach to obtain near-optimal solution.

1 Introduction

Entertainment services such as XR is required huge capability of computational and communication resources. In which, the expected bandwidth is required to reach more than 1(Gbps) [1], [2], and the computational resources requirement is extremely large because of complexity for 3-dimension(3D) video processing. In order to deal with the challenges, several works [3], [4], [5] have been proposed an frameworks to cache the results either UEs side or single MEC. Those works are limited by only single MEC, and does not consider about cooperation between multiple MECs. In this paper, we proposed a novel network model in which the XR contents is distributed among MECs. Each MEC can connected with each other via a fiber links which can provide high bandwidth, and low latency. The result of an XR task is partially or completely cache at a single or multiple MEC, UEs. Moreover, the cache results at UE side can be share to the other UE in proximity via D2D link. Thus, we can improve the spectrum utilization in the cellular network such as 5G. The detail of our propose network is illustrate in Fig. 1. We analysis the caching perform over the expected utility of a task size and formulate an optimization problem. We then apply dual decomposition technique in order to obtain a near-optimal solution.

Next, we define our system model, problem formulation, and solution approach.

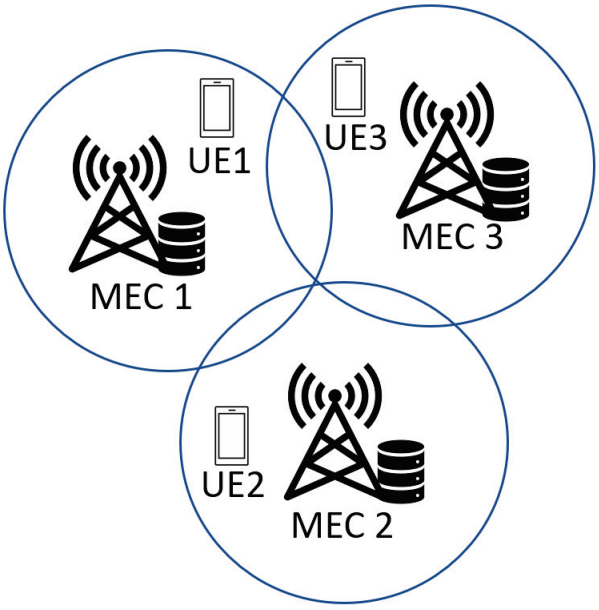

Fig. 1: Illustration of our system model.

2 System model

Consider a network model consisting a single micro cell base station and M numbers of small cell base stations $\mathcal{M} = \{1, 2, \ldots, M\}$. In each small cell $m \in M$ has a set of objects that frequently requested

for XR services denoted as $B_m = \{b_1^m, b_2^m, \ldots, b_{N_m}^m\}$, where N_m represent the number of objects in SBS m. Assuming that each $b_j^m \in \mathcal{B}$ has a size in megabits denoted as s_j^m. Let u_j^m denote the utility associated with object b_j^m. We assume that the utility by cache a result of XR task is proportional with its size. Therefor we use a logarithm function to represent the utility function

$$u_j^m = \log(s_j^m) \tag{1}$$

Let p_j^m denote the probability of a request for an XR task at SBS m for object j, then $p^m = \{p_1^m, p_2^m, \ldots, p_{N_m}^m\}$. Let $\mathcal{D}_m = \{1, 2, \ldots, D_m\}$, in which each device d_i^m has a limit storage capability $c_i^m(GB)$, and a probability distribution that the device will available on the location m as $q_{i,m} = \{q_{i,m}^t, \forall t = 1, 2, \ldots\}$. In this paper, we consider maximum time slot define by T, in reality t can be slotted in minute, hour, or any positive integer interval. Let $a_{i,j}^m$ denote the probability of a request for XR task of object j in location m at device i. When the result of XR task is cache at device d_i^m, the utility calculated as follow:

$$u(d_i^m, j) = \frac{1}{T} \sum_{t=1}^{T} q_{i,m}^t a_{i,j}^m p_j^m \log(s_j^m) x_{ij}^m, \tag{2}$$

where x_{ij}^m is the decision variable of device d_i^m in location m of content j.

$$x_{ij}^m = \begin{cases} 1, & \text{device } d_i^m \text{ cache the result } b_j^m, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

As each device have limit storage capability s_i^m, thus we define the constraint as follow

$$\sum_{j=1}^{N_m} x_{ij}^m s_j^m \leq c_i^m, \forall i \in D_m, \forall m \in M, \tag{4}$$

We assume that MEC server has a cache storage capability of S(GB) which very less to compare with the total size of XR computation results. Therefore, the SBS have to decide whether to cache the results or not. Let y_j^m be the decision variable of SBS for content b_j at location m

$$y_j^m = \begin{cases} 1, & \text{SBS cache the content } b_j^m, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Let $u(m)$ be the utility of SBS at location m. $u(m)$ can be calculated as follow:

$$u(m) = \sum_{j=1}^{N_m} p_j^m \log(s_j^m) y_j^m, \tag{6}$$

On the other hand, the cache decision is constrained by the storage capacity. Therefore we define a constraint to control the decision of MEC as follow

$$\sum_{m=1}^{M} \sum_{j=1}^{N_m} s_j^m y_j^m \leq S_m, \tag{7}$$

3 Problem Formulation

Based on aforementioned equation and definition, we formulated an optimization problem as follow:

$$\mathbf{A1}: \max_{x,y} \sum_{m=1}^{M} \sum_{i=1}^{D_m} \sum_{j=1}^{B_m} u(d_i^m, b_j^m) x_{ij}^m + \sum_{m=1}^{M} \sum_{j=1}^{B_m} u(b_j^m) y_j^m$$

s.t.

$$\text{C1}: \sum_{j=1}^{B_m} x_{ij}^m s_j^m \leq c_i^m, \forall i \in D_m, \forall m \in M,$$

$$\text{C2}: \sum_{m=1}^{M} \sum_{j=1}^{B_m} s_j^m y_j^m \leq S,$$

$$\text{C3}: y_j^m + \sum_{i=1}^{D_m} x_{ij}^m \leq 1, \forall m \in M, \forall j \in B_m,$$

$$\text{C4}: x_{ij}^m \in \{0, 1\}, \forall m \in M, \forall i \in D_m, \forall j \in B_m,$$

$$y_j^m \in \{0, 1\}, \forall m \in M, \forall j \in B_m.$$

The objective is aims to maximize the total utility of the network while caching the AR/VR results at either user's device or MEC cache storage. Constraints (C1), (C2) is representing for the storage capacity of UE, and the MEC, respectively. The constraint (C3) is guarantee for non-duplicate caching for each and every content. Due to the binary variables x_{ij}^m and y_j^m, also the coupling constraint (C4), the optimization problem is multidimensional Knapsack problem. Therefore, the problem become NP-hard problem and impossible to get the solution in polynomial time. Thus, we propose a fractional caching model which means relax the two binary variables x and y into continuous variables. Then problem **A1** is equivalent to

$$\mathbf{A2}: \max_{x,y} \sum_{m=1}^{M} \sum_{i=1}^{D_m} \sum_{j=1}^{B_m} u(d_i^m, b_j^m) x_{ij}^m + \sum_{m=1}^{M} \sum_{j=1}^{B_m} u(b_j^m) y_j^m \tag{9a}$$

s.t.

$$\text{(C1)} - \text{(C3)}, \tag{9b}$$

$$\text{C6}: 0 \leq x_{ij}^m \leq 1, \forall m \in M, \forall i \in D_m, \forall j \in B_m, \tag{9c}$$

$$\text{C7}: 0 \leq y_j^m \leq 1, \forall m \in M, \forall j \in B_m. \tag{9d}$$

Algorithm 1 :Dual Decomposition

1: **Input**: $\mathcal{M}$, $\mathcal{B}$, $\mathcal{D}$;
2: **Output**: $\mathbf{x}, \mathbf{y}, \mathbf{u}$;
3: Initialization: $max\_iteration = 10000, \quad \alpha = 0.5, k = 0$
4: **while** $k \leq max\_iteration$ **do**
5: $x^*(\lambda(k)) = \inf \left\{ \sum_{m=1}^{M} \sum_{i=1}^{D_m} \sum_{j=1}^{B_m} \left(u(d_i^m, b_j^m) + \lambda_j^m \right) x_{ij}^m \mid (C1), (C6) \right\}$
6: $y^*(\lambda(k)) = \inf \left\{ \sum_{m=1}^{M} \sum_{j=1}^{B_m} \left(u(b_j^m) + \lambda_j^m \right) y_j^m \mid (C2), (C7) \right\}$
7: $g(k) = \left[-x^*(\lambda(t)) - y^*(\lambda(k)) + \mathbf{1} \right]^+$
8: $\lambda_j^m(k+1) = \lambda_j^m - \rho(k) g(k)$
9: $k \leftarrow k + 1$

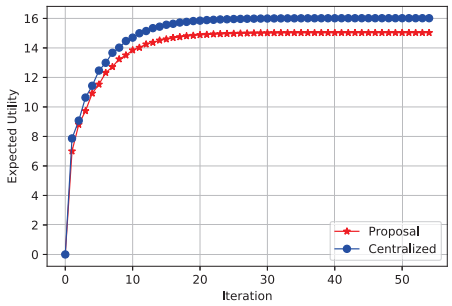

Fig. 2: Simulation result.

3.1 Dual Decomposition

Which the fixed k, we can solve the two primal problems separately. Thus, we can find a sub-gradient for the optimal value of each subproblem. From partial Lagrangian,

$$\begin{aligned} \mathcal{L}(x, y, \lambda) = & \sum_{m=1}^{M} \sum_{i=1}^{D_m} \sum_{j=1}^{B_m} \left(u(d_i^m, b_j^m) + \lambda_j^m \right) x_{ij}^m \\ & + \sum_{m=1}^{M} \sum_{j=1}^{B_m} \left(u(b_j^m) + \lambda_j^m \right) y_j^m \\ & - \sum_{m=1}^{M} \sum_{j=1}^{B_m} \lambda_j^m \end{aligned} \tag{10}$$

From Lagrangian function we can get dual function as follow

$$q(\lambda) = \inf_{x,y} \left\{ \mathcal{L}(x, y, \lambda) \mid (C1), (C2), (C6), (C7) \right\}. \tag{11}$$

Then, the dual problem define as follow

$$\max_{\lambda} q(\lambda) \tag{12a}$$

$$\text{s.t.} \lambda \succeq 0 \tag{12b}$$

The dual decomposition algorithm

4 Simulation

To simulate our propose solution approach, we set a number of MECs $M = 5$, number of UEs $D = 100$, and total number contents $B = 10^3$. $\mathbf{p}, \mathbf{q}, \mathbf{a}$ are generate by the Zipf distribution with parameters $\alpha = \{1.0, 1.0, 1.0\}$, respectively. To solve the problem in (12), we use a solver called Convex.jl [6] as our simulation tools. As showing in the Fig 2, our propose solution approach achieved an acceptable gap to compare with the global optimal solution.

5 Conclusion

In this paper, we proposed a novel network model for computational caching of XR services in 5G and Beyond networks. By applying the dual decomposition methods, we can show that our proposed solution has achieved the best performance and very close to the global optimal solution.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2020-2015-0-00742) supervised by the IITP(Institute for Information communications Technology Planning Evaluation). *Dr. CS Hong is the corresponding author.

References

[1] S. A. Kazmi, T. N. Dang, N. H. Tran, M. Bennis, and C. S. Hong, "Radio resource management techniques for 5g verticals," *5G Verticals: Customizing Applications, Technologies and Deployment Techniques*, pp. 119–136, 2020.

[2] S. A. Kazmi, L. U. Khan, N. H. Tran, and C. S. Hong, *Network Slicing for 5G and Beyond Networks*. Springer, 2019.

[3] T. N. Dang and C. S. Hong, "Distributed caching for cellular networks: A dual decomposition methods," *한국정보과학회 학술발표논문집*, pp. 965–967, 2019.

[4] 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.

[5] S. Ullah, L. U. Khan, and C. S. Hong, "Socially-aware svc video caching in information centric networking with d2d communication," *Journal of Korean Information Science Society*, pp. 1389–1391, 2018.

[6] M. Udell, K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd, "Convex optimization in Julia," in *Proceedings of the 1st First Workshop for High Performance Technical Computing in Dynamic Languages*. IEEE Press, 2014, pp. 18–28.