T2TRG Hong, Choong Seon
Internet-Draft Kyung Hee University
Intended status: Standards Track Chit Wutyee Zaw
Expires: August 09, 2022 Kyung Hee University
Kang, Seok Won
Kyung Hee University
October 2020

User Centric Assignment and Partial Task Offloading for Mobile Edge Computing in Ultra-Dense Networks

draft-hongcs-t2trg-ucapto-00

Abstract

By collocating servers at base stations, Mobile Edge Computing (MEC) provides low latency to users for real time applications such as Virtual Reality and Augmented Reality. To satisfy the growing demand of users, base stations are deployed densely in highly populated areas. Coordinated Multipoint Transmission (CoMP) allows users to connect to multiple base stations simultaneously. In ultra-dense networks, by offloading the partials of tasks to different base stations, users can achieve lower latency and utilize the computation ability of the surrounding base stations. To control the signaling overhead, the number of base stations that can be connected should be limited. In this paper, we propose a user-centric base station assignment algorithm by considering the possible load of base stations. Moreover, a partial task offloading algorithm is proposed to utilize the computation of under-loaded base stations. Resource allocation is then solved by convex optimization.

Status of this Memo

Table of Contents

1. Introduction 1
  1.1. Terminology and Requirements Language 2
2. System Model 2
3. Problem Formulation. 3
4. User-centric Assignment and Partial Offloading 3
    4.1. User-centric Assignment. 3
    4.2. Partial Offloading 4
    4.3. Radio Resource Allocation. 4
5. Results. 5
6. IANA Considerations. 5
7. Security Considerations 5
8. References 5
8.1. Normative References 5
8.2. Informative References 6
Authors' Addresses 6

1. Introduction

Mobile Edge Computing (MEC) has been an interesting topic in both academia and industry for its ability to provide low latency and high computation to users by setting up severs near to users. Computation and latency intensive applications requires users to offload their tasks to servers to achieve the minimum delay and maintain the energy of users’ devices. In densely deployed networks, users can utilize the resources of nearby base stations (BS) by offloading partials of their tasks with the technology provided by Coordinated Multipoint Transmission (CoMP).
Despite the advantages that MEC brings, there are many challenges to tackle in MEC which are pointed out in [1]. The communication aspect is

surveyed in [2] where authors considered joint management of radio and computation resources. Authors also introduced standards and application scenarios.

Authors in [3] developed a distributed approach for the offloading of computation tasks, caching of content and allocation of resources by using an alternating direction method of multipliers. Task offloading for ultra-dense network was considered in [4] where authors divided the task placement and resource allocation problems and proposed an efficient offloading approach. But, authors considered to offload to one BS. In this paper, we consider partial offloading in ultra-dense networks. To avoid the overloading at BSs, we take the number of possible users who can connect to BSs into account and propose a heuristic algorithm for user-centric assignment. In addition, a partial offloading algorithm is proposed to utilize the resources of under-loaded BSs by offloading the larger portion of tasks to those BSs. Then, resource allocation is solved with the help of convex optimization.

1.1. Terminology and Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. System Model

A network with densely deployed BSs is considered where users can offload their tasks to multiple BSs simultaneously.
We consider the Orthogonal Frequency Division Multiple Access in both uplink and downlink transmission. We also consider that MEC server are equipped with multi-core technology that they can compute offloaded tasks simultaneously. The user’s task has three parameters, b_i, o_i and c_i which are size of input file, output result and task in CPU cycles.

```
      +---------------+      +--------+      +--------+      +--------+
      | Mobile Device |      | SBS 1  |      | SBS 2  |      | SBS 3  |
      |               |      |        |      |        |      |        |
      +---------------+      +--------+      +--------+      +--------+
            |                     |
            | +-----------------+ |
            | | Offload partial | |
            | | portion of task | |
            | +-----------------+ |
            |                     |
            |             +-----------------+
            |             |  Compute the    |
            |             |  offloaded task |
            |             +-----------------+
            |                     |
            | +-----------------+ |
            | |   Return task   | |
            | |      result     | |
            | +-----------------+ |
            |                     |
         ---------------------------------------------------------------
            |                                     |
            |          +-----------------+        |
            |          | Offload partial |        |
            |          | portion of task |        |
            |          +-----------------+        |
            |                                     |
            |                             +-----------------+
            |                             |  Compute the    |
            |                             |  offloaded task |
            |                             +-----------------+
            |                                     |
            |          +-----------------+        |
            |          |   Return task   |        |
            |          |      result     |        |
            |          +-----------------+        |
            |                                     |
         ---------------------------------------------------------------
            |                                                     |
            |                     +-----------------+             |
            |                     | Offload partial |             |
            |                     | portion of task |             |
            |                     +-----------------+             |
            |                                                     |
            |                                          +-----------------+
            |                                          |  Compute the    |
            |                                          |  offloaded task |
            |                                          +-----------------+
            |                                                     |
```

```
    |                     +-----------------+             |
    |                     |   Return task   |             |
    |                     |      result     |             |
    |                     +-----------------+             |
    |                                                     |
  ---------------------------------------------------------------
```

Figure 1: Partial offloading with Coordinated Transmission in an Ultra-Dense Network

3. Problem Formulation

The objective of the partial offloading and resource allocation problem is to minimize the latency of all mobile users where the task must be computed fully. The maximum number of SBSs that a user can associate to is limited. The uplink bandwidth for task offloading and downlink bandwidth for result transmission are limited. In addition, the computing resource at MEC servers and local computing resource are also restricted.

4. User-centric Assignment and Partial Offloading

4.1. User-centric Assignment to SBSs

First, we need to determine the user assignment to the BSs by considering the overloading possibility. The score from a user to a SBS is calculated in which the uplink, downlink singal-to-noise ratios and the inverse proportion of the number of users who are likely to associate to a SBS is considered.

```
    +---------------+      +--------+      +--------+      +--------+
    | Mobile Device |      | SBS 1  |      | SBS 2  |      | SBS 3  |
    |               |      |        |      |        |      |        |
    +---------------+      +--------+      +--------+      +--------+
         |
   +-----------------+
   | Calculate score |
   |  for all SBSs   |
   +-----------------+
         |
   +-----------------+
   |  Choose 3 SBSs  |
   |   with highest  |
   |      scores     |
   +-----------------+
         |
      ---------------------------------------------------------------
         | +-----------------+ |
         | | Send the signal | |
         | |  for assignment | |
         | +-----------------+ |
         |                     |
      ---------------------------------------------------------------
         |                                     |
         |          +-----------------+        |
         |          | Send the signal |        |
         |          |  for assignment |        |
         |          +-----------------+        |
         |                                     |
      ---------------------------------------------------------------
         |                                                     |
         |                     +-----------------+             |
         |                     | Send the signal |             |
         |                     |  for assignment |             |
         |                     +-----------------+             |
      ---------------------------------------------------------------
```

Figure 2: User-centric Assignment

4.2 Partial Task Offloading

After the assignment is done, the fractions of the task allocated to BSs are resolved by utilizing the resources of under-loaded BSs. The higher portion of a task is offloaded to a SBS with a lower total computing load of all the assigned users. SBSs are sorted according to the increasing computing loads of the users. The portion of the task is offloaded to SBSs in the order.

```
      +------------------+---------------+------------+------------+
User's|     Portion      |    Portion    | Portion    | Portion    |
task  |   offloaded to   | offloaded to  |offloaded to| computed at|
      |      SBS 1       |      SBS 2    |   SBS 3    |  the user  |
      +------------------+---------------+------------+------------+
       \                / \             / \          / \          /
        \              /   \           /   \        /   \        /
         \            /     \         /     \      /     \      /
          \          /       \       /       \    /       \    /
           \        /         \     /         \  /         \  /
            \      /           \   /           \/           \/
         +------------+    +-----------+  +-----------+ +-----------+
         |   SBS 1    |    |   SBS 2   |  |   SBS 3   | |   Local   |
         +------------+    +-----------+  +-----------+ +-----------+
```

Figure 3: Partial Task Offloading

4.3. Radio Resource Allocation

After obtaining the partial task offloading, we need to solve the resource allocation problem. The resource allocation problem is convex which can easily be solved. In this paper, we use cvxpy [5] to solve this problem. For the local CPU cycles assignment, the maximum available CPU cycle is assigned since the objective is minimizing the latency.

5. Results

Poisson Point Process is used to model the deployment of BSs and users where their densities are 0.6/m2 6/m2 respectively. For power density thermal noise, -174dBm/Hz is used. Fig. 2 shows the simulation setup used in the paper. Transmit power of pico BSs and users are 23dbm and 20dbm respectively. CPU speed is 4GHz at MEC server and 0.3GHz at user. The total uplink and downlink bandwidth are 20MHz each. The size of input file follows a uniform distribution between [300, 800] KB. The uniform distribution is also used to model the size of tasks and output files which are [0.5, 1] GHz and [0.2, 2.5] MB respectively.
The latency obtained at SBSs are different but most of the SBSs have the similar latency results due to the different user task requirements.
In the highly dense networks, the proposed approach can keep most of the BSs to achieve comparable results. The proposed approach obtains lower latency compared to the baseline approach where the loads of SBSs are not considered and task allocation is done uniformly. The difference becomes significant as the number of users increases.

6. IANA Considerations

There are no IANA considerations related to this document.

7. Security Considerations

There are no security considerations related to this document.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[1] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1628-1656, 2017.

[2] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322-2358, 2017.

[3] C. Wang, C. Liang, F. R. Yu, Q. Chen and L. Tang, "Computation Offloading and Resource Allocation in Wireless Cellular Networks With Mobile Edge Computing," IEEE Transactions on Wireless Communications, vol. 16, no. 8, pp. 4924-4938, 2017.

[4] M. Chen and Y. Hao, "Task Offloading for Mobile Edge Computing in Software Defined Ultra-Dense Network," IEEE Journal on Selected Areas in Communications, vol. 36, no. 3, pp. 587-597, 2018.

[5] S. Diamond and S. Boyd, "CVXPY: A Python-Embedded Modeling Language for Convex Optimization," Journal of Machine Learning Research, vol. 17, no. 83, pp. 1-5, 2016.

8.2. Informative References

Authors' Addresses

Choong Seon Hong
Computer Science and Engineering Department, Kyung Hee University
Yongin, South Korea
Phone: +82 (0)31 201 2532
Email: cshong@khu.ac.kr

Chit Wutyee Zaw
Computer Science and Engineering Department, Kyung Hee University
Yongin, South Korea
Phone: +82 (0)31 201 2987
Email: cwyzaw@khu.ac.kr

Seok Won Kang
Computer Science and Engineering Department, Kyung Hee University
Yongin, South Korea
Phone: +82 (0)31 201 2987
Email: dudtntdud@khu.ac.kr