INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

Department of Computer Science and Engineering

Semester Project Report on

latency Minimization in the cloud-edge collaborative system in an energy-efficient way

Under guidance of

prof p.k. Jana

Made by

Neelansh Maheswari (17je003468)
Sujit Brajabasi(17je003358)

Abhishek Gautam (17je003410)

ACKNOWLEDGEMENT

We have our most humble respects to our guide Prof. Prasanta K. Jana, Professor (Department Of Computer Science and Designing) at IIT(ISM) Dhanbad, for allocating us with this chance of researching this technical area of collaborative cloud edge computing. Prof. Jana has consistently been instrumental in clearing our questions and has been inspiring us. Regardless of the busy timetable he has, Prof. Jana has never turned down our solicitation for conversations and consistently has been very respectful with us. We will be grateful for the duration of my life for his honorable assistance and direction.

May 2021
IIT (ISM), Dhanbad

INTRODUCTION

Nowadays, mobile data is growing in massive amounts, to improve computation performance, edge computing is used. Edge computing stands for deploying the cloud computing service at an edge network. This is effective in reducing the core network congestion and long-term transmission latency which is found in traditional cloud computing. This edge computing service can be implemented at the base station known as MEC (mobile edge computing).

If cloud computing has collaborated with edge computing, a huge peak performance is seen in the system.

PROBLEM STATEMENT

Description of the communication model, computation model and network architecture of the cloud-edge collaboration system and finding the overall minimum weighted-sum delay for every device in an energy efficient way.

CONTENTS

1. PROJECT OVERVIEW :

Full advantage of the cloud computing capacity at the cloud server and edge computing capacity at BSs can be taken ,If we consider that the task of each mobile device can be partially offloaded to the cloud server and partially processed at the edge node for processing. According to this, a joint computation and communication resource allocation problem is formulated to minimize the weighted-sum delay of all mobile devices.

There are two significant parameters, the normalized cloud computation capacity and normalized backhaul communication capacity, that affect the computation resource allocation of each and every node (mobile device). Therefore, we devise a closed-form optimal task splitting strategy as a function of these two parameters. To gain more information, we further demonstrate four different network scenarios: edge-dominated system, communication-limited system, cloud-dominated system and computation-limited system.

The optimal task splitting strategy, we convert the original joint computation resource allocation and communication problem into an equivalent optimization problem. Moreover, a necessary condition is further developed to find whether a task should be executed at the corresponding edge node only, rather offloading it to the cloud server.

System model :

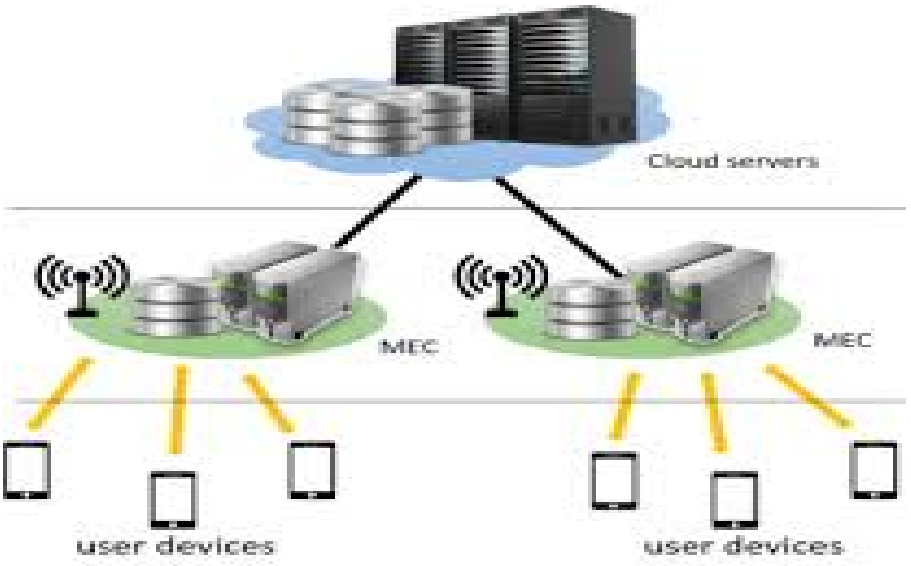
Here we assume a cloud-edge collaboration system with one main cloud server and J single antenna , which are denoted by $J = \{1, 2, \ldots, J\}$. Each Base station is connected with a MEC server, having limited resources for data processing.

The BS and MEC combination is known as an edge node. Inside the coverage area of the j-th BS, there are I(j) devices . Each device is connected with BS with a wireless channel.

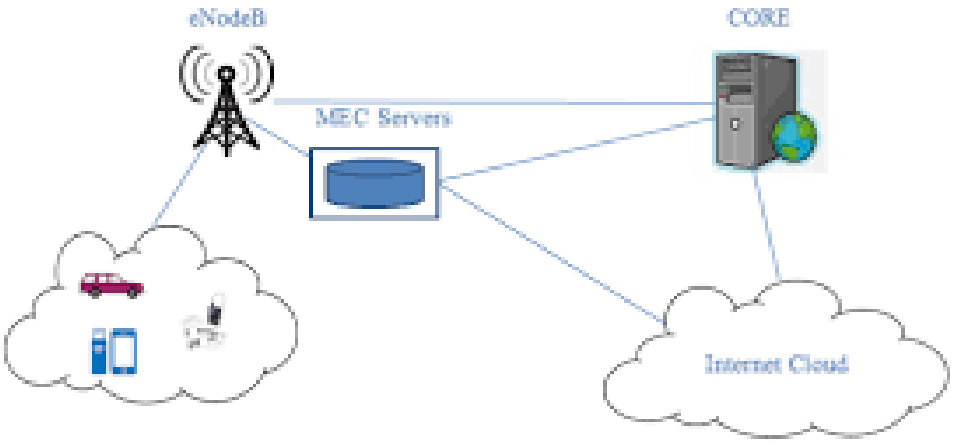
A(i,j) = {L(i,j) , C(j,i)}, it is considered as the computation task of the i-th device connected with j-th edge node

L(i,j) denotes input size of the task in bits and C(j,i) denotes the CPU cycle are needed to compute 1-bit data

Fe(j) denotes the j-th edge node computation capacity and Fe denotes the cloud server computation capacity.



Communication model (wireless):



2. METHODOLOGY :

A . Analyzing the Delay :

By considering every situation on account it is found that there can four types of delay are as follows :

Transmission delay of the mobile device :
when the mobile device directly transmit the task to the BS without any local computation a transmission delay arises

The average transmission delay for the i-th device to offload its computation task to the connected j-th BS is given by

$t\text{-}tran.d(j,i) = L(j,i)\ T/R(j,i)\ \tau(j,i)$

here T denotes length of 1-TDMA frame
R(j,i) is expected media capacity
assumption :
$\tau(j,i)$ the time slot resource for each device is fixed in each time frame.

Computation delay of the Edge Node :
After receiving the task from a mobile device the MEC server immediately divides/splits the task into two parts and offloads one part to the cloud server and computes the remaining part at edge node.

Let $\lambda(j,i)$ be the splitting ratio belongs to [0,1]

$t\text{-}comp.e(j,i) = \lambda(j,i)\ L(j,i)\ C(j,i)/\ Fe(j,i)$

Transmission delay of the Edge Node :
delay occur in transmitting the splitted task into the cloud server

t-tran.e (j,i) = (1-ƛ(j,i)) L(j,i) / W(j)

where W^-1(j) is represented as the time required for the backhaul link to transmit 1-bit data

Computation delay of the Cloud server:
time required to compute the offloaded task in the cloud server

t-comp.c(j,i) = (1-ƛ(j,i)) L(j,i) C(j,i)/Fe(j,i)

B . Formulation of the problem :
------ as task splitting operation rely on the specific parameters of each task such as its computation load and data size, the edge node can not split a task until the whole data is received
----- task computation may depend on data structure, so cloud server cannot start processing a task until the transmission between the edge node and cloud servers ends.

So , according to this the overall delay of the i-th mobile device served by the j-th edge node can be expressed as

T (j,i) = t-tran.d (j,i) + max{t-comp.e(j,i) , t-tran.e(j,i) + t-comp.c(j,i)} --- (1)

so our aim is to minimize the overall delay
for the weighted – sum overall delay we assign a positive β(j,i) belongs to (0,1)

the **problem** is formulated as

p1 = min ΣΣ β(j,i) T(j,i) for minimum{τ(j,i), fe(j,i) ,fc(j,i) }

now if we **decompose** the optimal problem there will be two subproblem

----The first one is to minimize the weighted-sum transmission delay between each the connected BS and mobile device connected to it , formulated as follows :

P2 = min ΣΣ β(j,i) t-tran.d(j,i)   for minimum τ(j,i)

---- The other one is to minimize the weighted-sum computation delay of the cloud server and each edge node , and can be formulated as

P3 = min ΣΣ β(j,i) t-comp(j,i)   for minimum {fe(j,i) , fc(j,i) , λ(j,i)}

C. Communication resource allocation (optimal) :

for this policy the cloud -edge collaboration system, the time-slot allocated to the i-th device served by the j-th edge node can be described as follows :

$$\tau_{j,i}^{*} = \frac{\sqrt{\frac{\beta_{j,i} L_{j,i}}{R_{j,i}}}}{\sum_{j=1}^{J} \sum_{i=1}^{I_j} \sqrt{\frac{\beta_{j,i} L_{j,i}}{R_{j,i}}}} T.$$

D. Task splitting optimal strategy :

As we know P3 cannot be solved directly, first we need to find the splitting ratio, assuming Fe(j,i) and Fc(j,i) are fixed.

--- The ratio between the backhaul communication capacity and the edge computation capacity is defined as backhaul communication capacity index

η(j,i) = C(j,i) W(j) / Fe(j,i)

--- The ratio between the cloud computation capacity and the edge computation capacity is defined as cloud computing capacity index

γ(j,i) = Fc(j,i) / Fe(j,i)

so for the cloud-edge collaboration system the optimal task splitting capacity strategy can be defined as the harmonic average of the two ratio and expressed as follows

$$\lambda_{j,i}^{*} = \frac{\eta_{j,i} + \gamma_{j,i}}{\eta_{j,i} + \gamma_{j,i} + \eta_{j,i}\gamma_{j,i}}.$$

where splitting ratio lies between (0,1)

according to the splitting ratio it is found that there are several architecture system such as system having high edge computation capacity and low transmission capacity

E. Optimal resource allocation :

applying the task splitting ratio to the t-comp(j,i) equation , it can be re-written as follows

$$\widehat{t_{j,i}^{\text{comp}}} = \frac{C_{j,i} f_{j,i}^{\text{c}} + C_{j,i}^{2} W_j}{f_{j,i}^{\text{e}} f_{j,i}^{\text{c}} + C_{j,i} W_j \left(f_{j,i}^{\text{e}} + f_{j,i}^{\text{c}}\right)} L_{j,i}.$$

and the P3 can be equivalently formulated as P4 which is a convex optimization problem

P4 = min ΣΣ β(j,i) t-comp(j,i)   for minimum {fc(j,i) , fe(j,i)}

3. EVALUATION AND RESULT

A. System model

According to the methodologies if we use the optimal task splitting strategy it is found that there are several system architecture such as

Limited communication system – In this system the computation capacity of the system is sufficient but the communication capacity between the edge node and the cloud server is not sufficient means $\eta << \gamma$ for all i-th devices connected to the j-th edge node. it occurs when the number of the device connected to an edge node is huge while the backhaul media capacity is limited.

Which means the ratio of task splitting can only be determined by the backhaul media capacity index

$\lambda(1) = \lim \lambda^* = 1/(1+\eta)$ where, $\eta/\gamma \to 0$

Special case when $\eta=0$ the $\lambda(1) = 1$ which means the whole task would be executed in the edge node and no offloading is there

Limited computation system– In this system the computation capacity of the cloud server and the edge nodes are insufficient but the communication capacity of the edge-cloud server are sufficient means $\eta >> \gamma$ for i devices connected to the j-th edge node.

$\lambda(2) = \lim \lambda^* = 1/(1+\gamma)$ where $\eta/\gamma \to \infty$

which means the task splitting strategies can easily be determined by the cloud computation capacity

special case if the $\gamma < 1$ means the computation capacity of the edge node is greater than the cloud server computation capacity , ess data will be offloaded to the cloud server

Dominated Edge system -- In this system the computation capacity of the edge node is much greater than the cloud server computation capacity means $fe(j,i) >> fc(j,i)$
so we can say $\gamma \to 0$, this occurs when a large set of edge node is exist the cloud-edge system

$\lambda(3) = \lim \lambda^* = 1$ where $\gamma \to 0$

which means that the whole task would be executed in the edge node rather than offloading the task to the cloud server

Dominated Cloud system -- In this system the computation capacity of the edge node is insufficient and the cloud server computation capacity is tremendous i compare to the edge nodes means $\gamma \to \infty$

$\lambda(4) = \lim \lambda^* = 1/(1+\eta)$ where $\gamma \to \infty$

which states that the splitting ratio can only be determined by the backhaul media communication capacity

B.Testing Strategy

here first we first we compare the performance of the **cloud-edge collaborative** system with three other baseline system such as

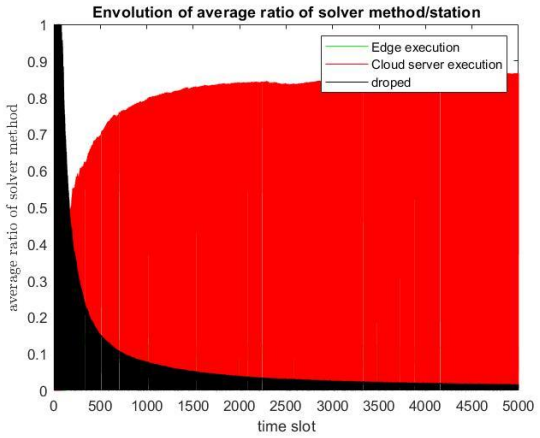
only-edge system -- where the whole task is executed in the edge node only with any offloading of the data to the cloud server

only-cloud system – where the whole task is offloaded to the cloud server for computing without performing any computation in the edge server

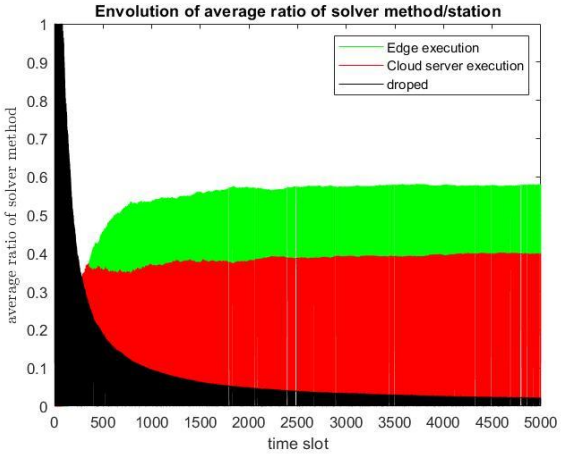
half-cloud and half-edge system – where the task is equally divided into two parts and half task is computed in the edge node and the half task is offloaded to the cloud server for computation

C.Simulation Result

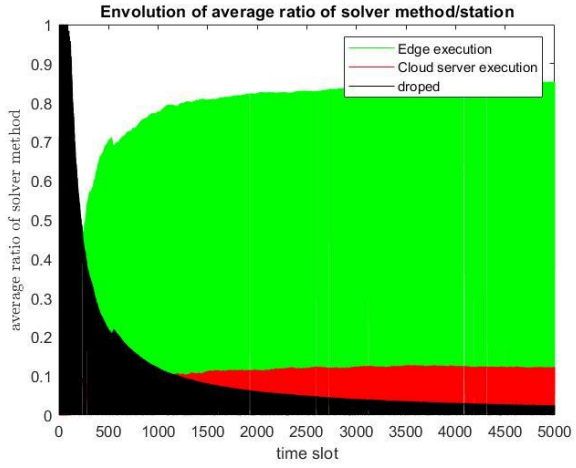
1- Evolution of edge and MEC execution averages as distance btw them changes



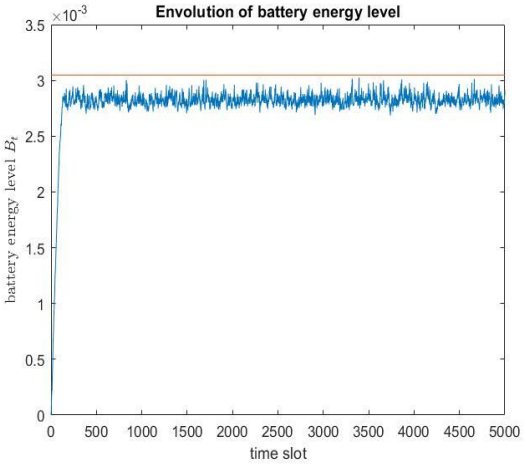
1-distance = 25



2-distance = 50



3-distance = 75



battery/energy levels of edge

4. CONCLUSION

In this project we explored the computation resource allocation and communication to minimize the overall and weighted-sum delay of all the mobile devices in collaborative cloud-edge systems. First of all a latency minimization problem was created. After that the problem was further

decomposed into two subproblems. One is associated with the optimal computation resource allocation by the help of task splitting strategy it was found that the task splitting ratio completely relies on the backhaul communication capacity and the cloud computation capacity.

And another one is communication resource allocation whose optimal solution can be expressed in closed-form by using the Cauchy-Buniakowsky Schwarz in-equality.

We further found some optimal task splitting strategy, In this project we consider the weighted-sum delay as the latency of each device.
In this work we want to collaborate the local mobile devices with cloud-edge collaborative system for further future research purposes.

5. REFERENCES

1. “collaborative cloud and edge computing”

2. “optimal resource allocation in cloud computing”