**Optimal Cloudlet Placement and User to Cloudlet Allocation in Wireless Metropolitan Area Networks**

(Mike Jia,2017)

1- vector the assignment of cloudlets to APs

where

2- matrix the assignment of users to APs

where

3- The set of users that are assigned to the cloudlet at access point

4- Vector of arrival rates of users

5- Wireless delays between users and their APs

6- The tolerable network delay

7- matrix the delays between APs

8- Fraction of tasks accepted at cloudlet

It is calculated based on the ratio of to the summation of all arrival rate of tasks of users assigned to the cloudlet

9- The average waiting time of tasks based on M/M/c

task arrival rate

serving rate of each of the homogenous servers

number of servers

10- Average waiting time of tasks

It is combined of waiting time for the fraction uploaded to the cloudlet

Waiting time for the fraction offloaded to remote cloud

11- The K cloudlet Placement Problem (KCP) of WMAN

**argmin** (f(x)) simply returns the value of x which minimizes f(x) over the set of candidates for x as opposed to the minimum value itself.

12- Easily we can extend the problem formulation of Mike jia from single objective optimization to multi-objective optimization as

……………………………………………..

**QoS-Aware Cloudlet Load Balancing in Wireless**

**Metropolitan Area Networks**

**(Mike jea 2020)**

1. Based on M\M\n queue model, Cloudlet with task arrival rate and with service rate
2. Use Ti, which is a function of a given task arrival rate to Calculate the average task wait time at cloudlet which consists of the **queuing time** and the **service time** of the task at cloudlet

Erlang’s C formula

1. Use to denote the amount of task flow from cloudlet to cloudlet for .
2. The constraint on

Ensures that for any two cloudlets and j, the flow in terms of task rate, from cloudlet to cloudlet j is the negative of the flow from cloudlet j to cloudlet .

The flow of tasks from any given cloudlet to itself is zero,

Ensures that all flow is conserved

Ensures that the sum of all outgoing task flows from cloudlet (ignore the incoming flow by summing the maximum of and 0 for each cloudlet j) is less than its incoming task arrival rate .

1. To model such a network delay in the WMAN, denote by the network delay matrix.

The entry represents the shortest possible communication delay in relaying a task between cloudlet and cloudlet j.

The flow of incoming redirected tasks at cloudlet has a delay of .

The sum of all network delays of incoming tasks from other cloudlets to cloudlet

From Equ. 1 an 3 Note: it’s not 3 its 6

the average task response time of all tasks that are executed on cloudlet

The final incoming task flow that will be processed at cloudlet

1. The Cloudlet Load Balancing Problem in an WMAN

The problem is to find a set of **inter-cloudlet** task flows under the constraints given in equations in number (4)

Such that the maximum task response time is minimized,

**Exploring Placement of Heterogeneous Edge Servers for Response Time Minimization in Mobile Edge-Cloud Computing**

(Kun Cao 2021)

1. System Architecture

*M* base stations

*N* edge servers

,

the server set

the server indexed by integer number is referred to as edge/cloud server

1. Response Time Model

*Communication Delay:*

The average communication latency of tasks transmitted from the base station to the edge/cloud server *.*

denote the Euclidean distance between the base station and the edge/cloud server *.*

The constant number and the nonnegative variable are the propagation rate of electromagnetic waves and the average task data volume of the base station, respectively.

*Task Execution Delay:*

M/G/1 queue model

The time distribution of task executions on the edge/cloud server can follow any general distribution with mean

The average task execution delay (including waiting time at task queue) of the base station on the edge/cloud server .

*Total Response Delay:*

The expected response time of the base station

The expected response time of the system given by the average response time of *M* base stations

1. Problem

Given an undirected graph *G = (V,E),* determine the placement locations of edge servers and the mapping strategy of base stations to edge/cloud servers in order to minimize the system expected response time and maximize the fairness in the expected response time of base stations.

Formulate ILP.

indicates the maximum of and ,

suggests the minimum of and .

The smaller the , the more fair the expected response time achieved by these base stations.

Define the ILP objective function as minimizing the sum of and

The linear constraints of variable

The variable can be represented by

is an auxiliary binary decision variable, and it is introduced to imply the relationship of , and .

If condition holds,

then ,

else