QoS-aware Mobile Edge Computing System: Multi-user Multi-server Scenario

Presenter: Te-Yi Kan

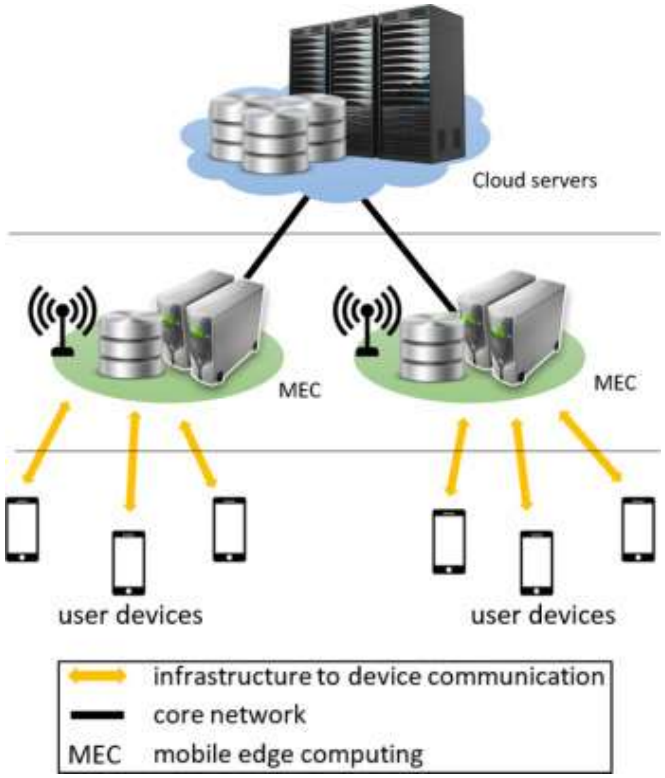
Te-Yi Kan, Yao Chiang, and Prof. Hung-Yu Wei
Dept. of Electrical Engineering
National Taiwan University

2018/12/09

Abstract

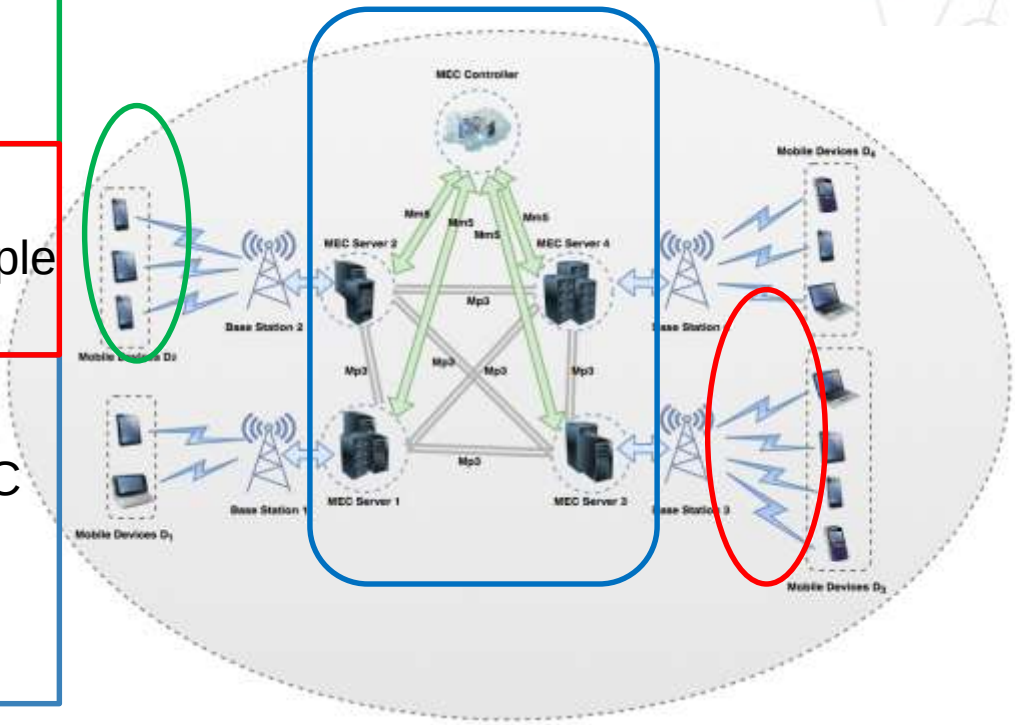
- **Mobile Edge Computing (MEC) System**
  - Provide computational resources at the edge of RAN.

- **Construct a QoS-aware MEC system**
  - Task offloading
  - Resource Allocation
  - Load Distribution



System Model

- **Three major parts:**
  - **Multiuser system**
    - Each base station serves a different number of mobile devices.
  - **Multi-channel system**
    - Each base station provides multiple sub-channels.
  - **Multi-server system**
    - Servers are controlled by our MEC Controller.
    - Servers are interconnected with each other.



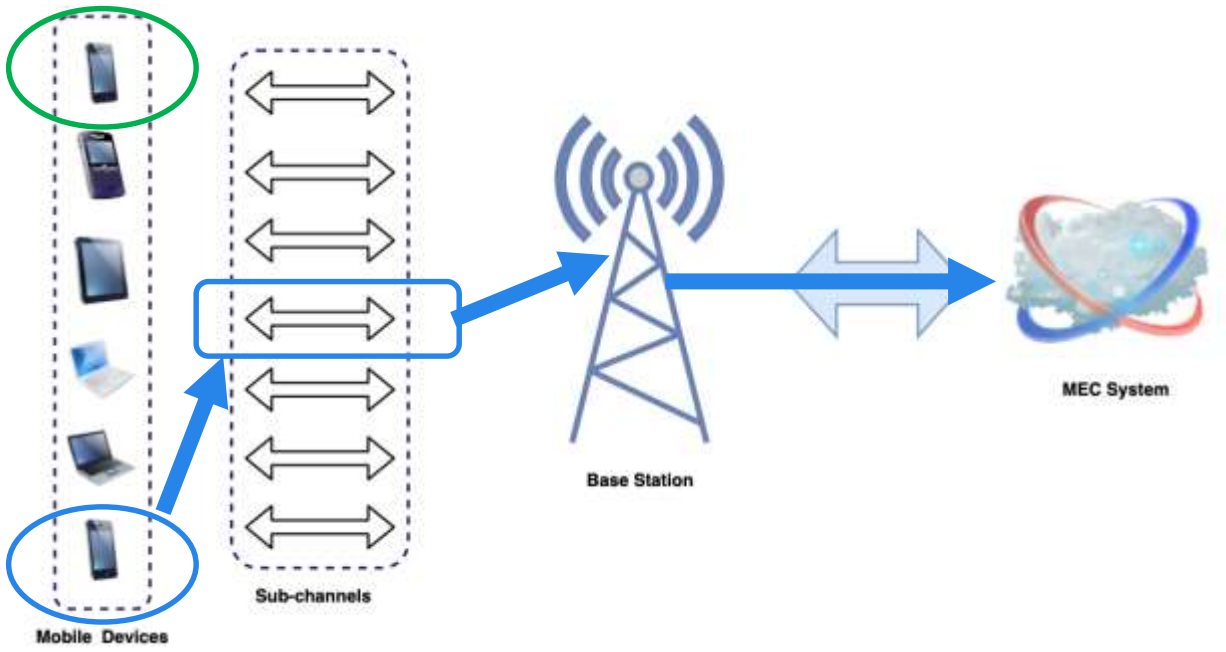
Problem Description

Offloading Decision Problem

- Each device can select either ***local execution*** or **remote execution** (***offloading***).

Resource Allocation

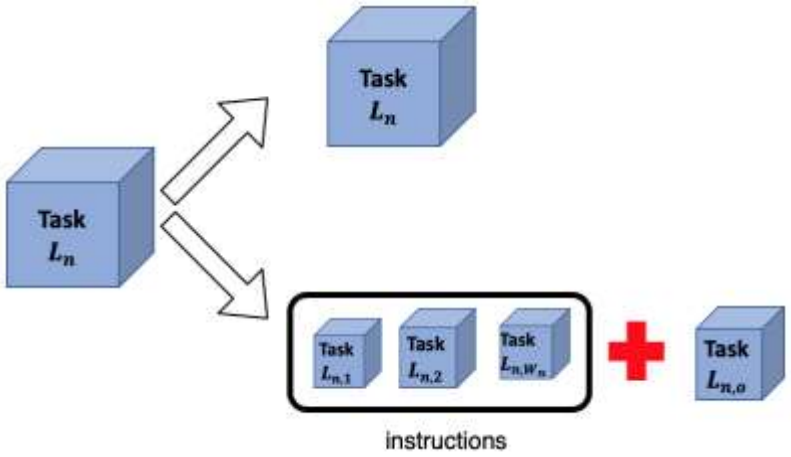
- Each offloading device should pick one sub-channel to transmit its task.
- Each offloading device should acquire computational resources on MEC system for execution.

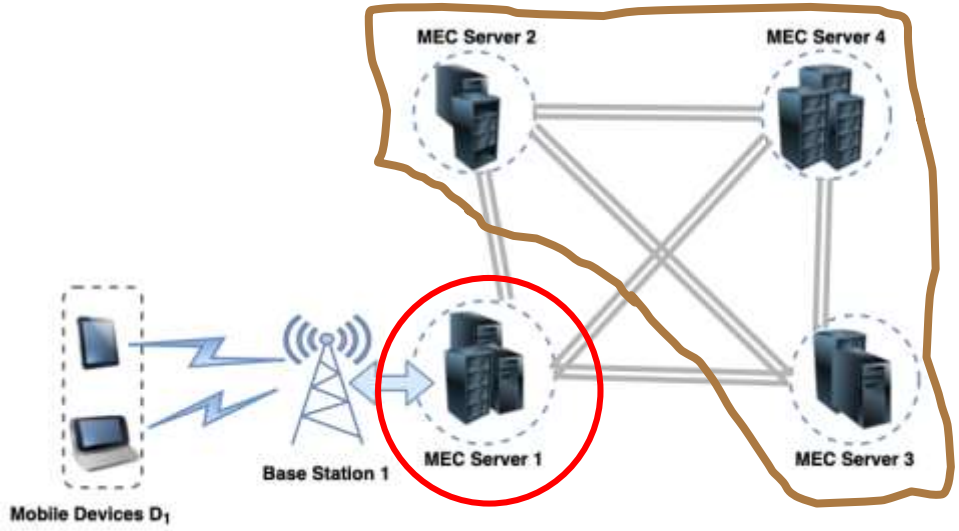


Problem Description – (cont.)

Load Distribution

- MEC servers can divide a task into some instructions.
  - There will be one additional instruction for gathering the output results.
- Each task or instruction can be executed on ***Local MEC Server*** or be further distributed to other servers (***Nearby MEC Server***)
  - **Local MEC Server**: the server directly connected to the BS which serves the device.
  - **Nearby MEC Server**: the server connected with Local MEC Server.



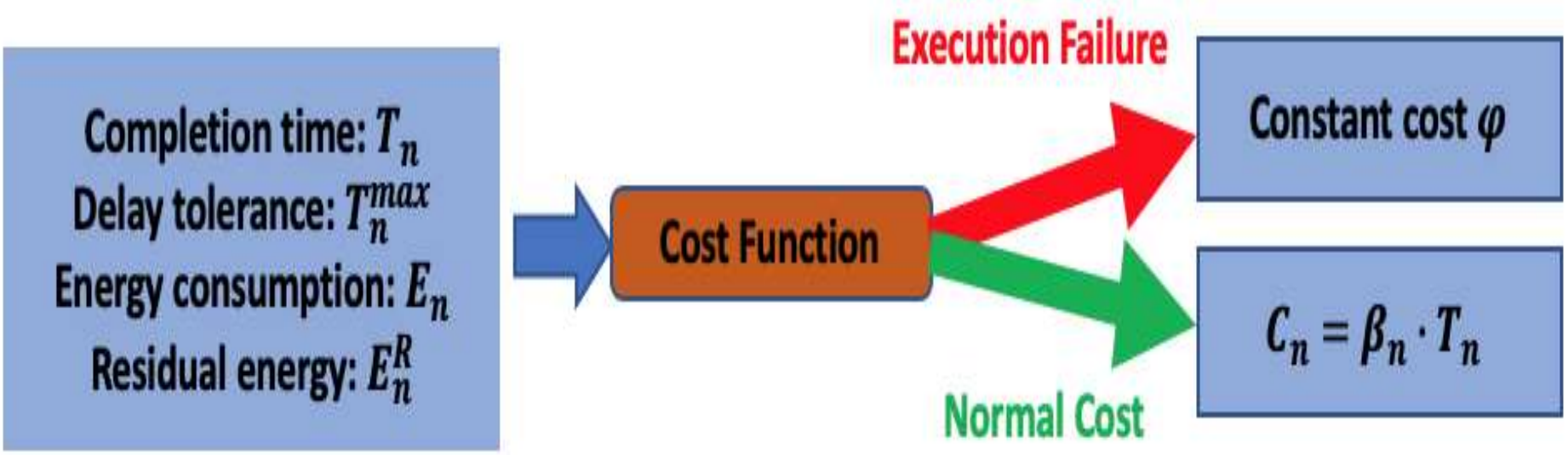


Definition of QoS

- The QoS in our work is related to the following three features:
  - **Execution Delay**
    - Lower execution delay implies better QoS.
  - **Delay Tolerance**
    - Different tasks have different delay tolerance.
  - **Residual Energy**
    - Many mobile devices do not have much energy to run applications.

Cost Function

- **Cost function:**



- The lower the cost is, the better QoS is achieved.
- The cost of failed task φ is much greater than normal cost.
- The weight β_n is negatively correlated with delay tolerance.

Proposed Algorithm

❇ **Three steps in our algorithm:**

Inside each MEC Server

Devices Classification & Priority Assignment

Radio Resource Allocation

Among multiple MEC Servers

Load Distribution & Computation Resource Allocation

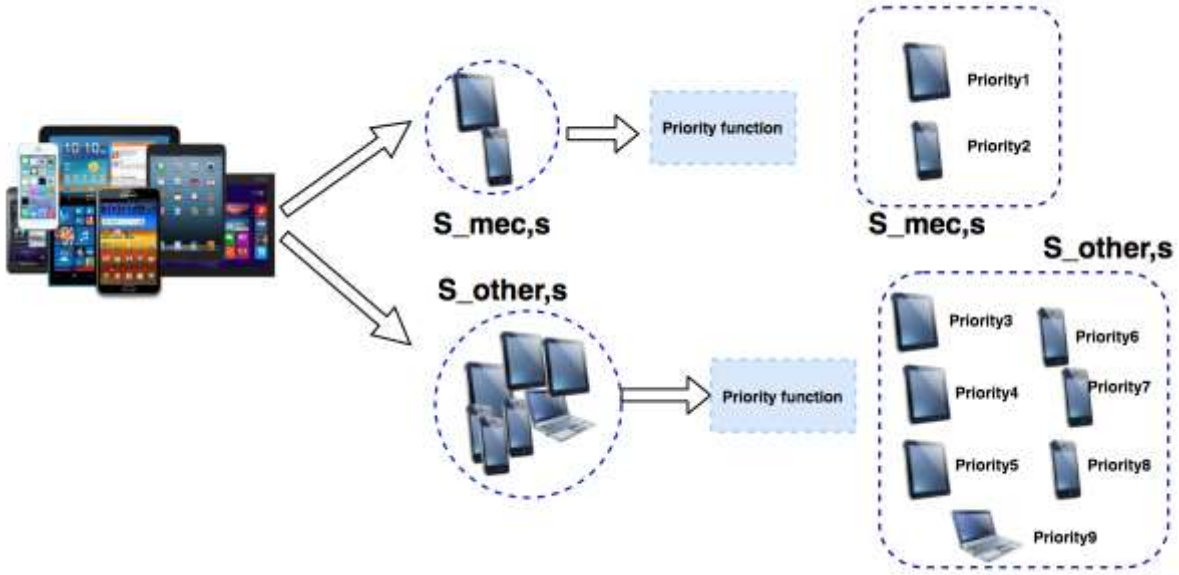
Devices Classification & Priority Determination

Devices Classification

- Two subsets: $S_{mec,s}$ and $S_{other,s}$
- The devices belong to $S_{mec,s}$ don't have enough resources or energy.

Priority Determination

- The devices with lower $\emptyset_n$ values have higher priority to offload their tasks.

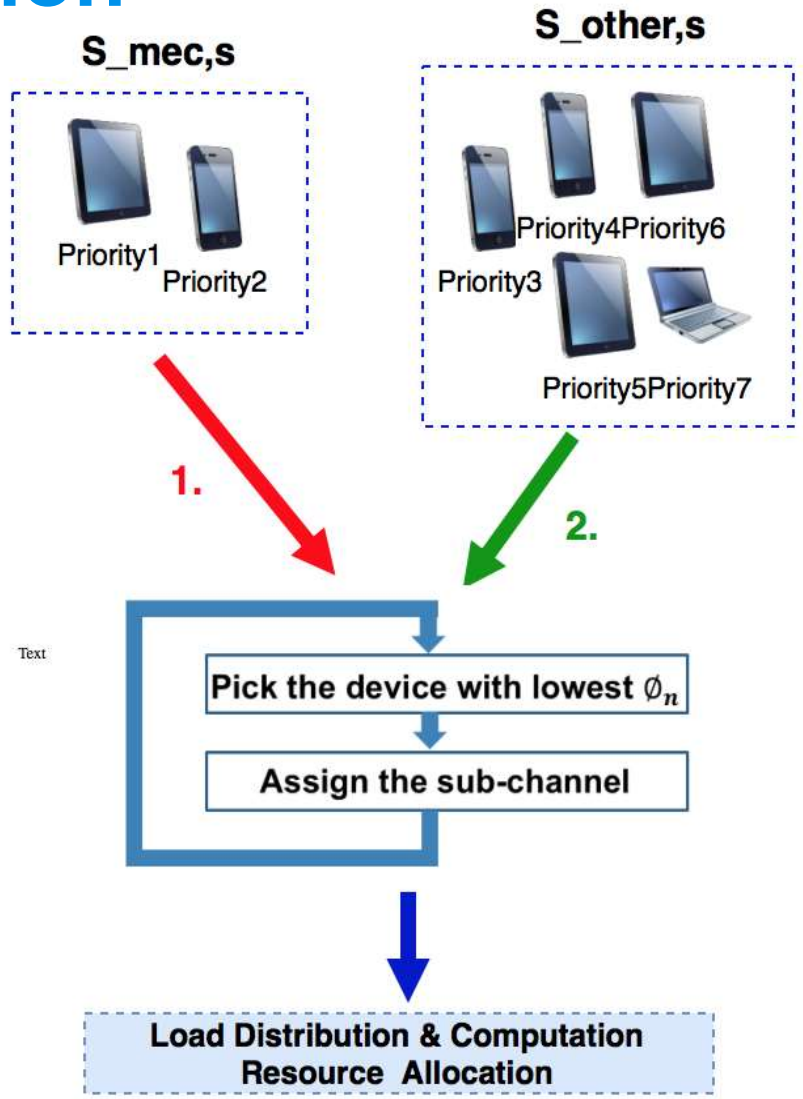


Radio Resource Allocation

Radio Resource Allocation

- Firstly, assign sub-channels to the devices in $S_{mec,s}$

- After assignment for $S_{mec,s}$, we'll next consider $S_{other,s}$.



Load Distribution & Computation Resource Allocation

❇ Load Distribution

(1) **Full Task Assignment:** Assign tasks to their Local MEC Server in ascending order of Δ_n

(2) **Full Task Distribution:** Distribute unserved tasks to other server with adequate resources.

(3) **Partial Task Distribution:** Divide unserved tasks into small instruction and distribute them to server with adequate resources.

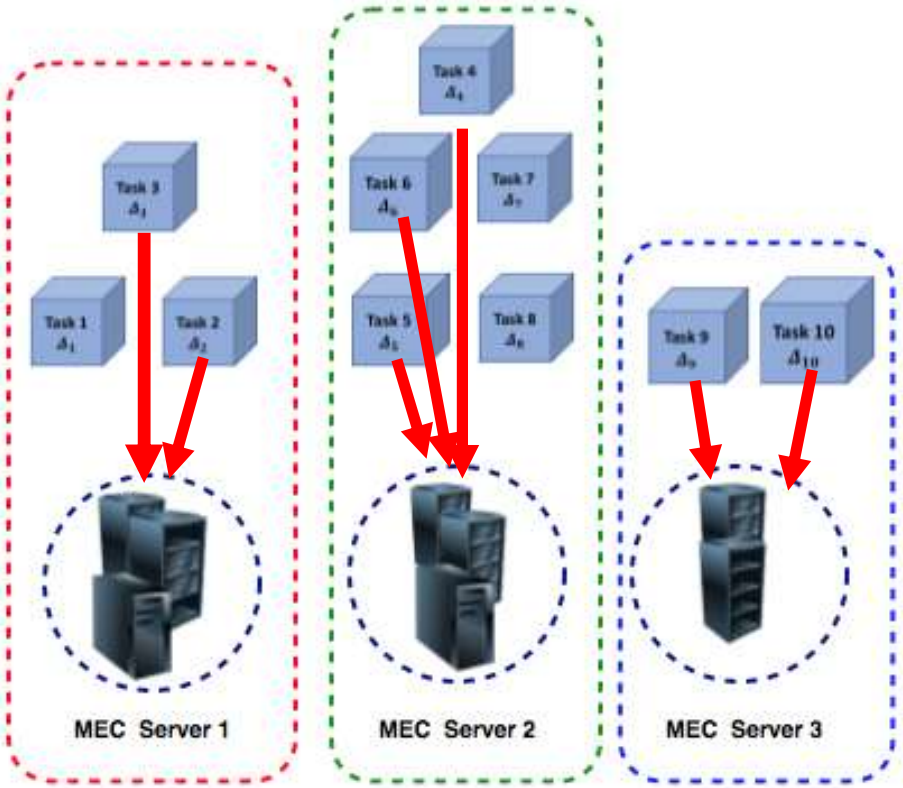
❇ Computation Resource Allocation

- ✾ Adopt Lagrange Multiplier.

Load Distribution – sub-step 1

Full Task Assignment

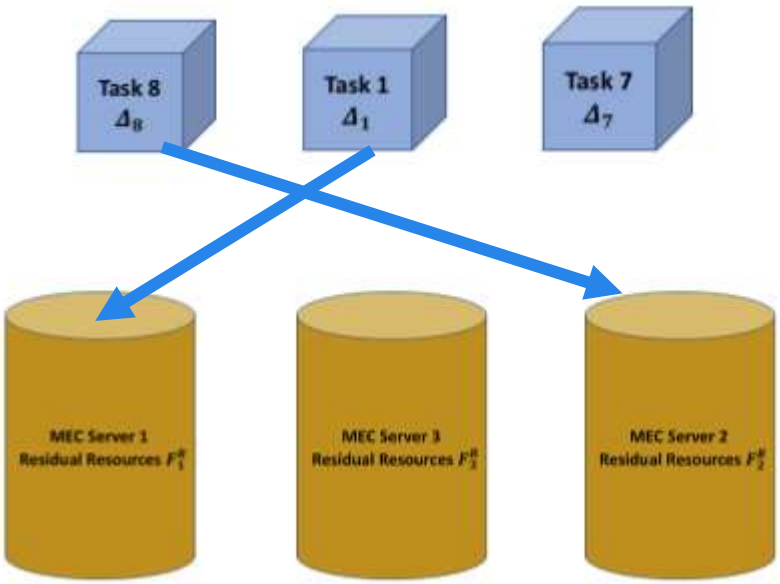
- First, assigning tasks to their Local MEC Server in ascending order of Δ_n
- Δ_n is the minimum required resources to complete the task within its delay tolerance.

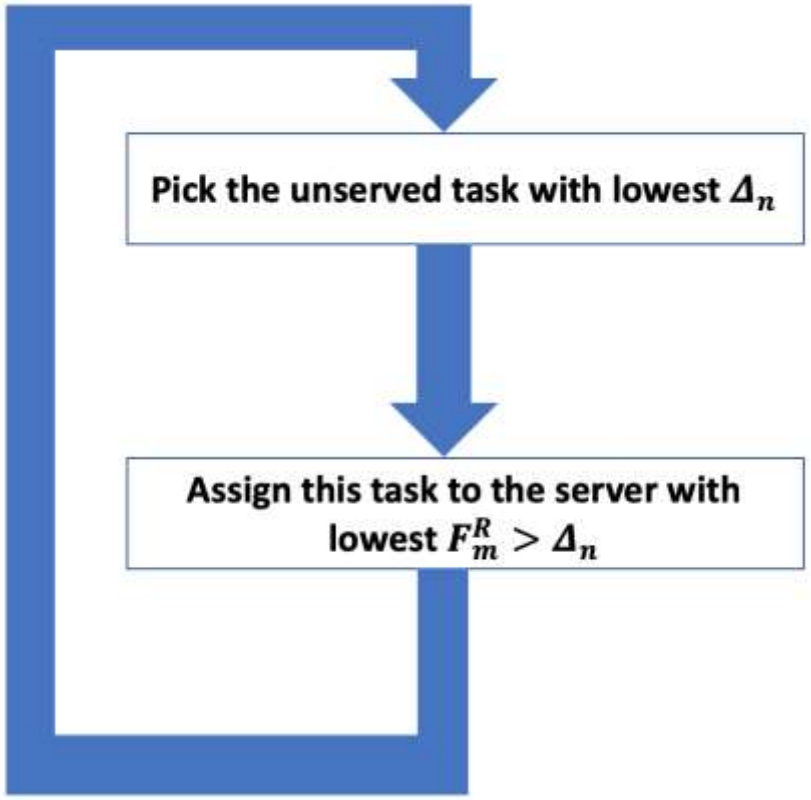


Load Distribution – sub-step 2

Full Task Distribution

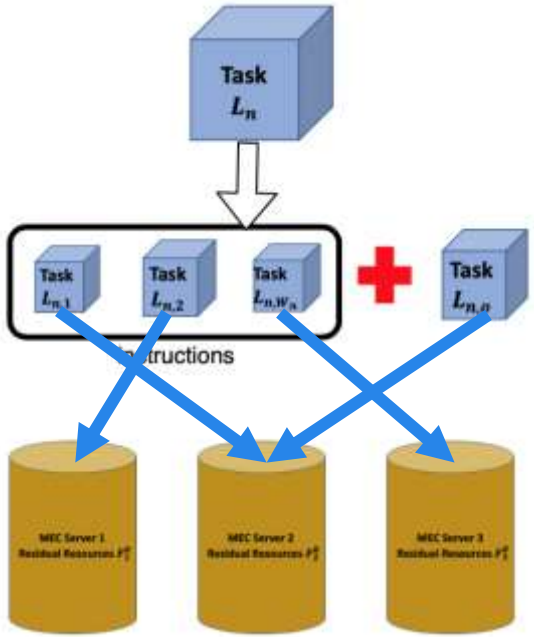
- After assignment in each server, we'll distribute unserved tasks to other servers with adequate resources.

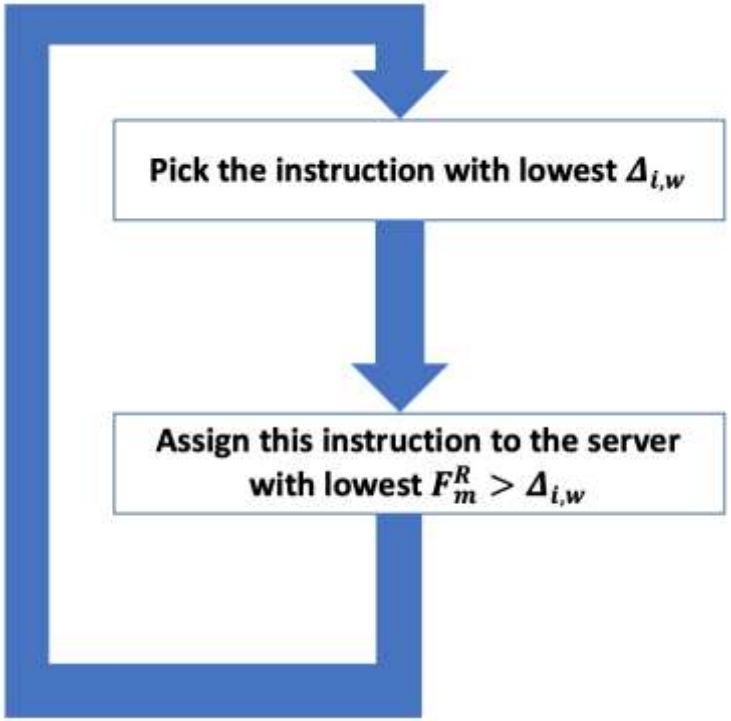




Load Distribution – sub-step 3

- **Partial Task Distribution**
  - After Full Task Distribution, MEC will **divide unserved tasks into instructions** and **distributed them to servers with adequate resources.**





Load Distribution & Computation Resource Allocation

❇ Load Distribution

(1) **Full Task Assignment:** Assign tasks to their Local MEC Server in ascending order of Δ_n

(2) **Full Task Distribution:** Distribute unserved tasks to other server with adequate resources.

(3) **Partial Task Distribution:** Divide unserved tasks into small instruction and distribute them to server with adequate resources.

❇ Computation Resource Allocation

- ✾ Adopt Lagrange Multiplier.

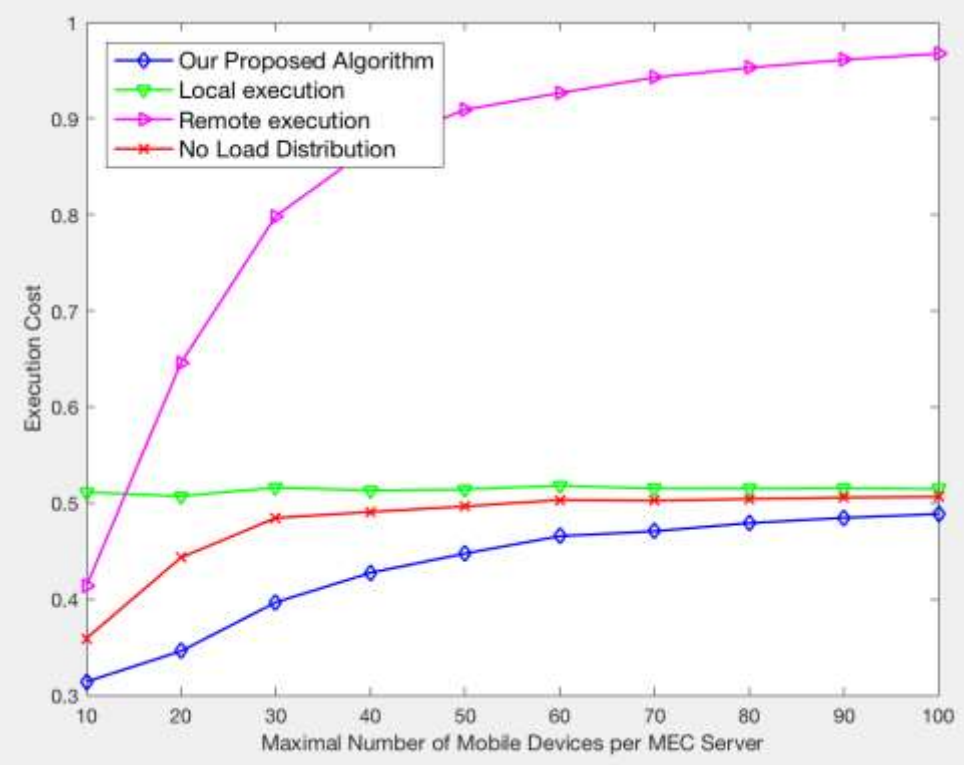
Simulation Settings

- **Scenario:**
  - Number of the BSs $M = 5$
  - Number of the sub-channel $H = 15$

- **Comparison schemes:**
  - **Local execution**
    - Tasks would be executed only on local mobile device
  - **Remote execution**
    - Tasks would be always offloaded to MEC system.
  - **No Load Distribution**
    - Tasks would be executed on local mobile devices or their Local MEC Server.

Simulation Results (Execution Cost)

- Execution cost of a task is defined by our cost function.
- Following Uniform Distribution $unif\{x-10, x\}, x = maximal\ number$.



Simulation Results (User Satisfaction)

- User satisfaction: The percentage of complete tasks.