QoS-aware Fog Computing System: Load Distribution and Task Offloading

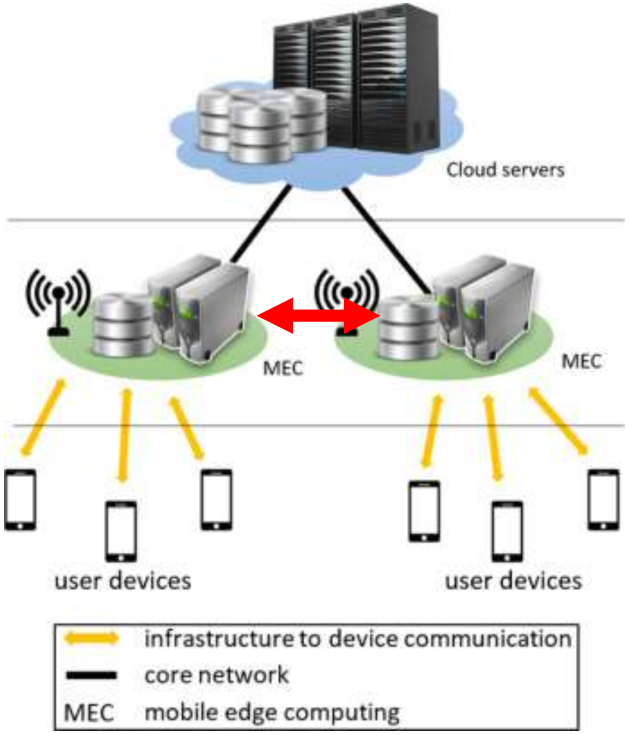
Presenter: Te-Yi Kan

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

2018/08/23

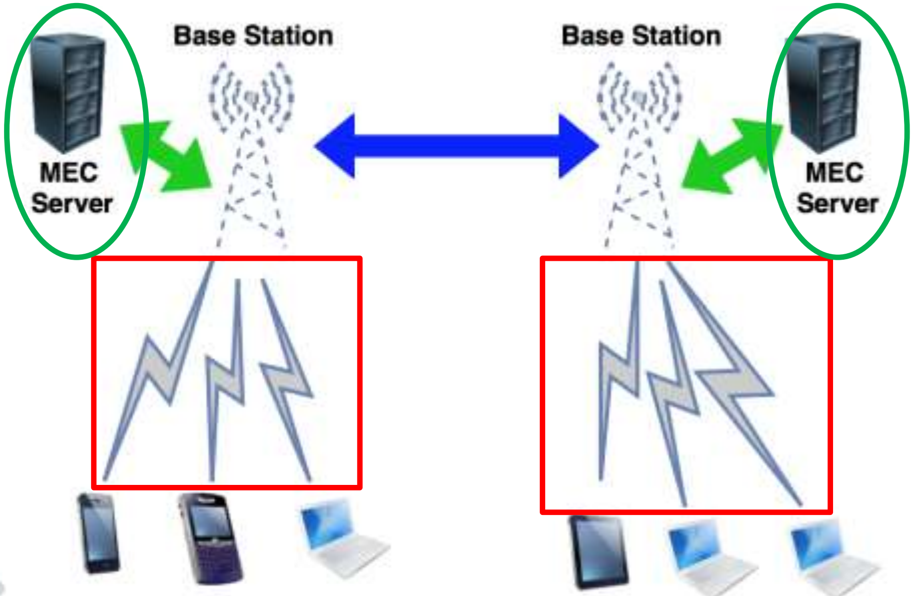
Introduction

- **Mobile-Edge Computing (MEC) system**
  - A type of fog computing system.
  - Providing computation resources at the edge of RAN.



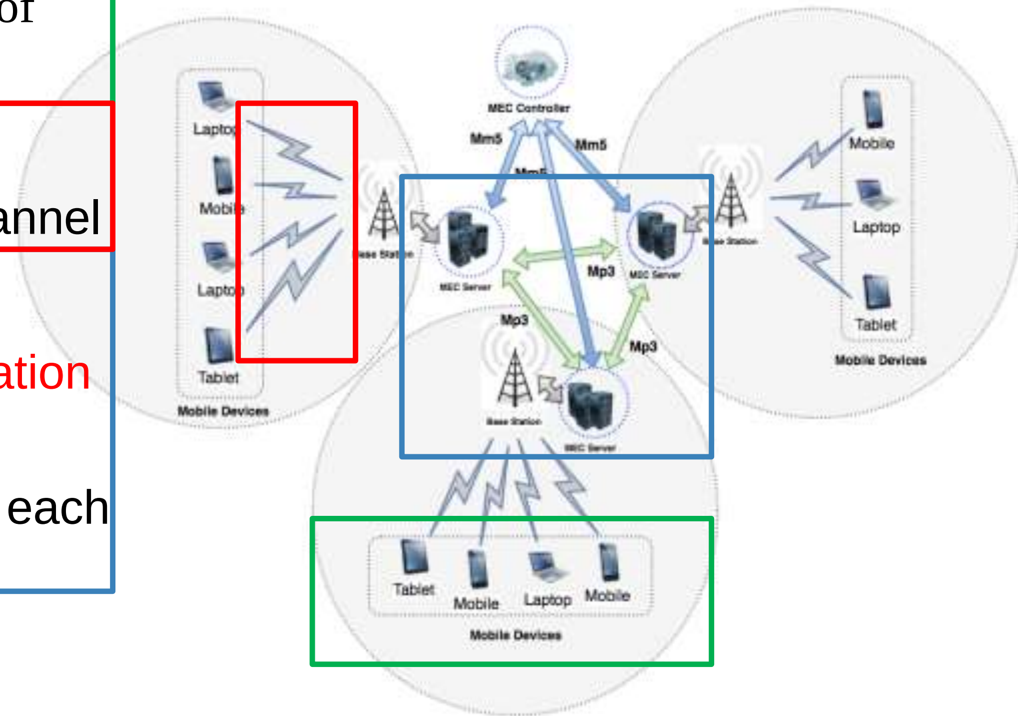
Motivation

- Promoting QoS as well as considering two critical features of tasks and devices.
  - Different tasks have different delay tolerance.
  - Task execution cannot run out of the energy of the device.
- We take two types of resource allocation into account.
  - **Radio resource allocation**
  - **Computation resource allocation**
- **Multi-server scenario** and **load distribution** are considered.



System Model

- **Three major parts:**
  - **Multiuser system**
    - Each BS serves a different number of mobile devices.
  - **Multi-channel system**
    - Each BS has multiple sub-channel
  - **Multi-server system**
    - Each server has limited computation resources.
    - Servers are interconnected with each other.

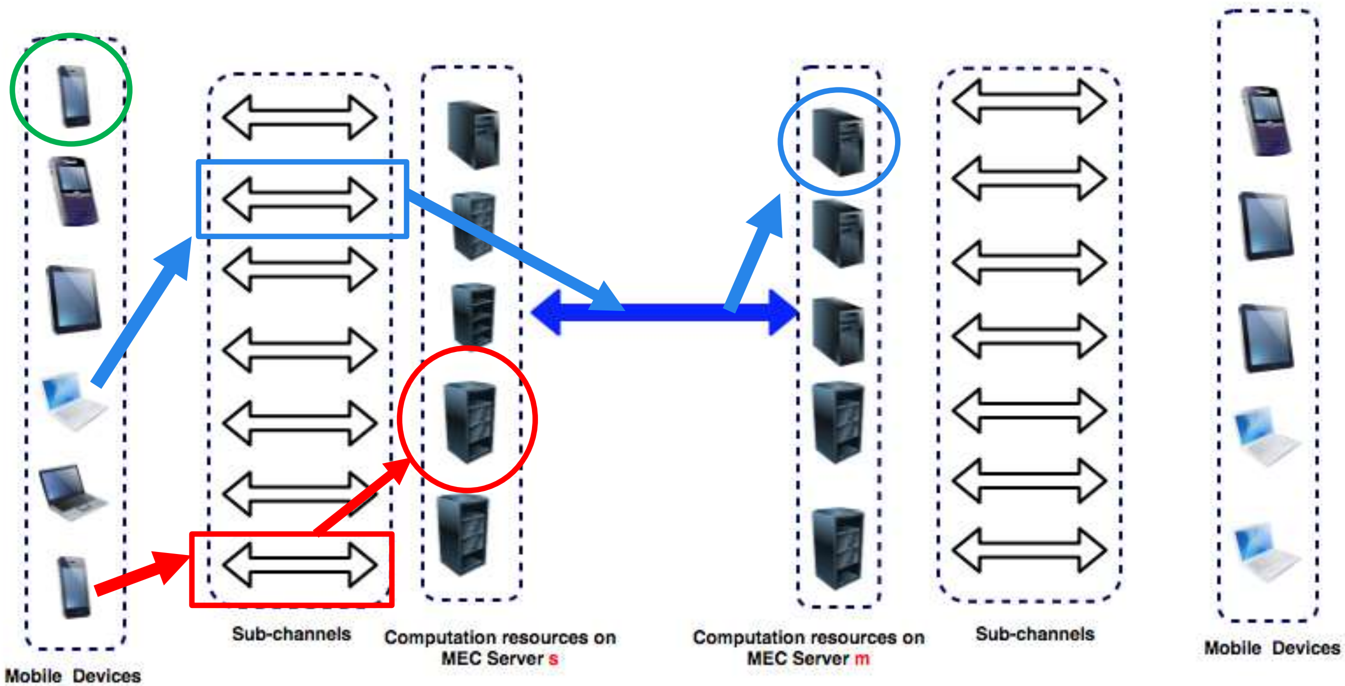


- **Four correspond problems:**
  - **Offloading decision problem**
  - **Radio resource allocation**
  - **Load distribution and computation resource allocation**

Problem Description

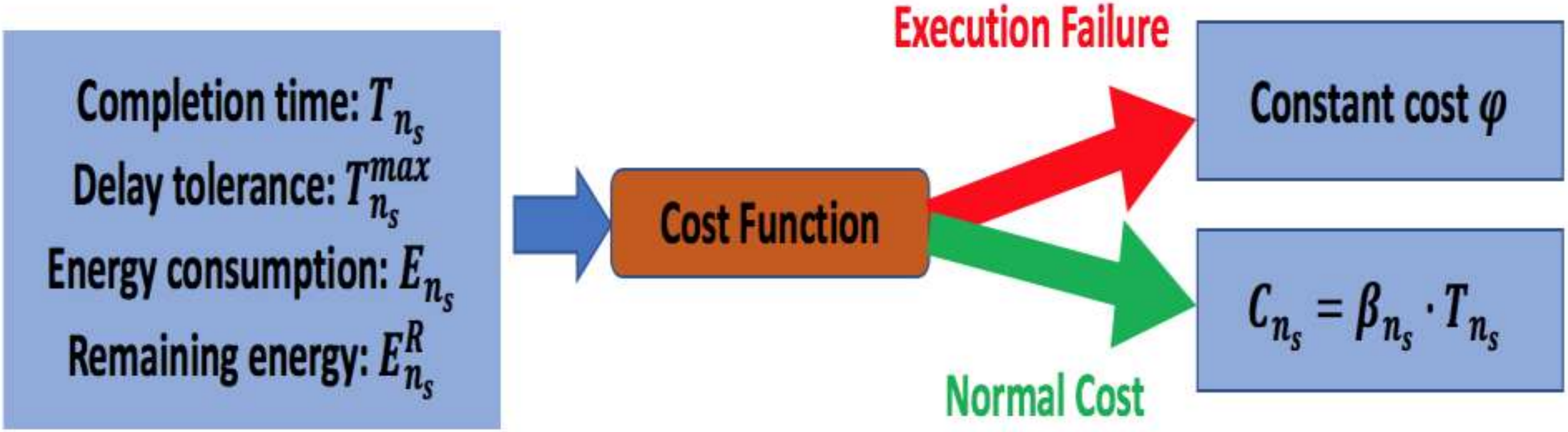
Formulation of these problems:

- Each device has its own offloading decision $x_{n_s} \in \{0,1,2,\dots,M\}$ (***Three possible locations***).
  - If $x_{n_s} = 0$: This device selects ***local execution***.
  - Else if $x_{n_s} = s$: This device offloads its task to ***Serving MEC Server***.
  - Else, $x_{n_s} > 0$ & $x_{n_s} = m \neq s$: This device offloads its task to ***Nearby MEC Server***.
- Each offloading device has its own channel selection $h_{n_s} \in \{h_s^1, h_s^2, \dots, h_s^H\}$.



Cost Function

- **Cost function:**
  - To take delay tolerance and remaining energy into account.



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

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

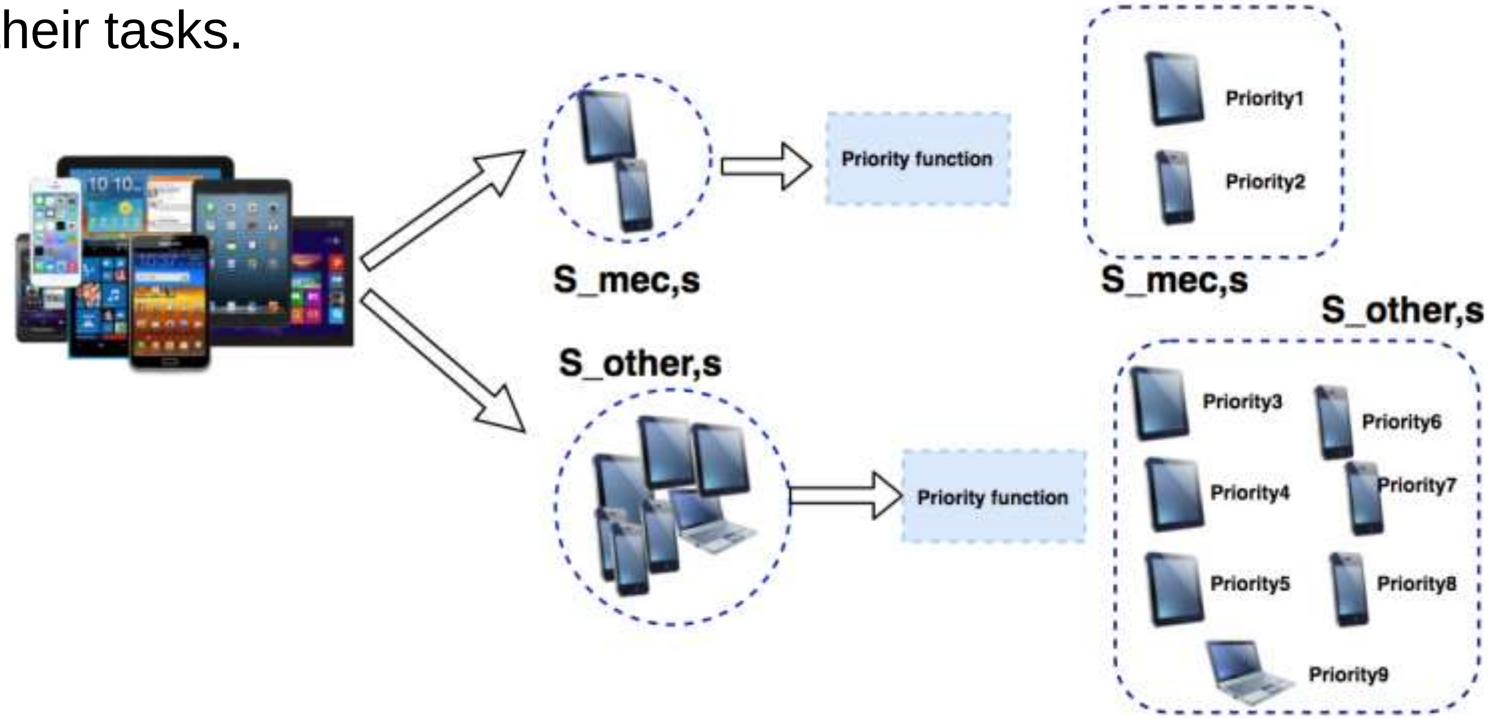
Devices Classification & Priority Determination

Devices Classification

- Two sets: $S_{mec,s}$ and $S_{other,s}$
- Task execution failure should be avoided.

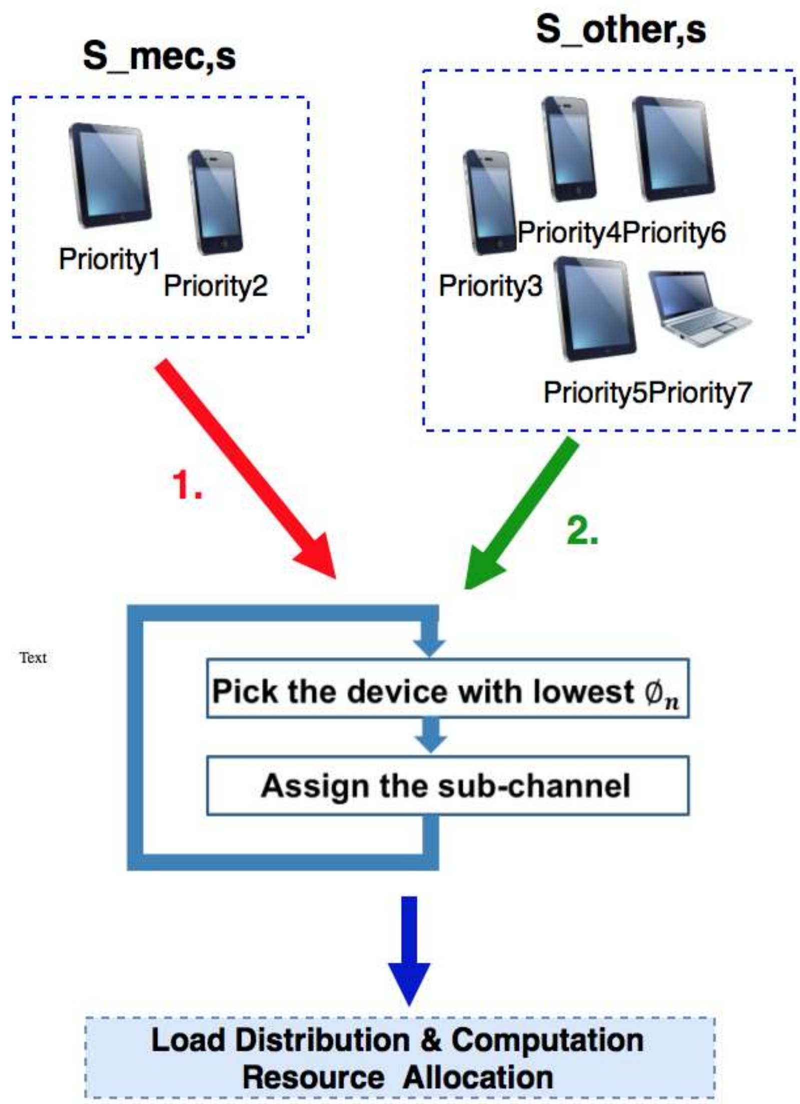
Priority Determination

- The devices with lower $\emptyset_{n_s}$ values have higher priority to offload their tasks.



Radio Resource Allocation

- **Radio Resource Allocation**
  - First, assigning sub-channels to the devices in $S_{mec,s}$
  - After assignment for $S_{mec,s}$, we'll consider $S_{other,s}$.



Load Distribution & Computation Resource Allocation

Load Distribution

(1) First, assigning tasks to their Serving MEC Server in ascending order of Δ_{n_s}

(2) After assignment in each server, we'll distribute unserved tasks to other server with adequate resources.

(3) Finally, terminating all the unserved tasks.

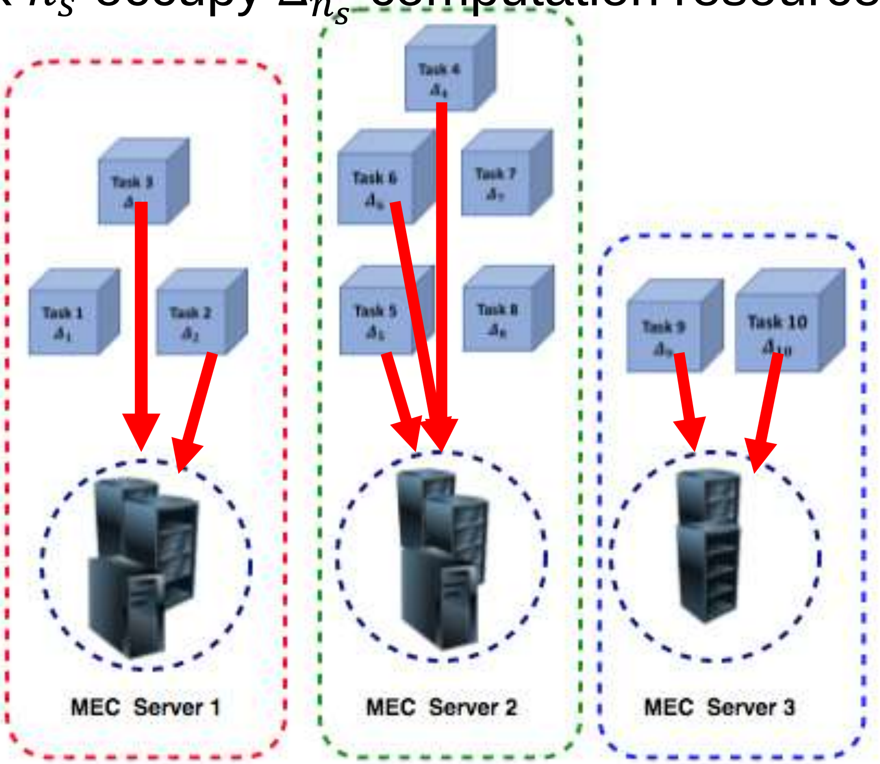
Computation Resource Allocation

- Adopting Lagrange Multiplier.

Load Distribution – sub-step 1

Load Distribution

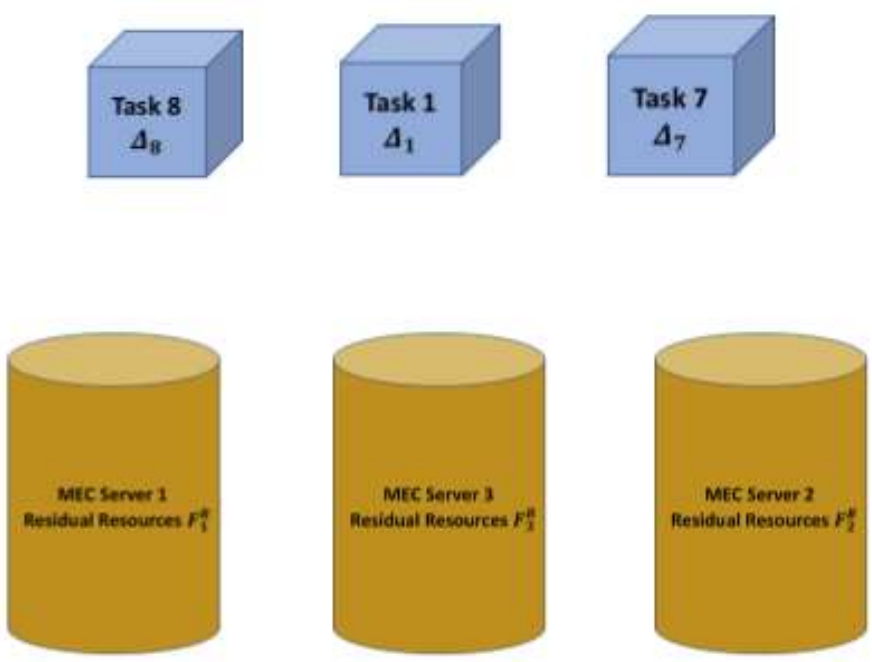
- First, assigning tasks to their Serving MEC Server in ascending order of Δ_{n_s} (Δ_{n_s} is the minimum required resources to complete the task within its delay tolerance).
- Assigned task n_s occupy Δ_{n_s} computation resources.

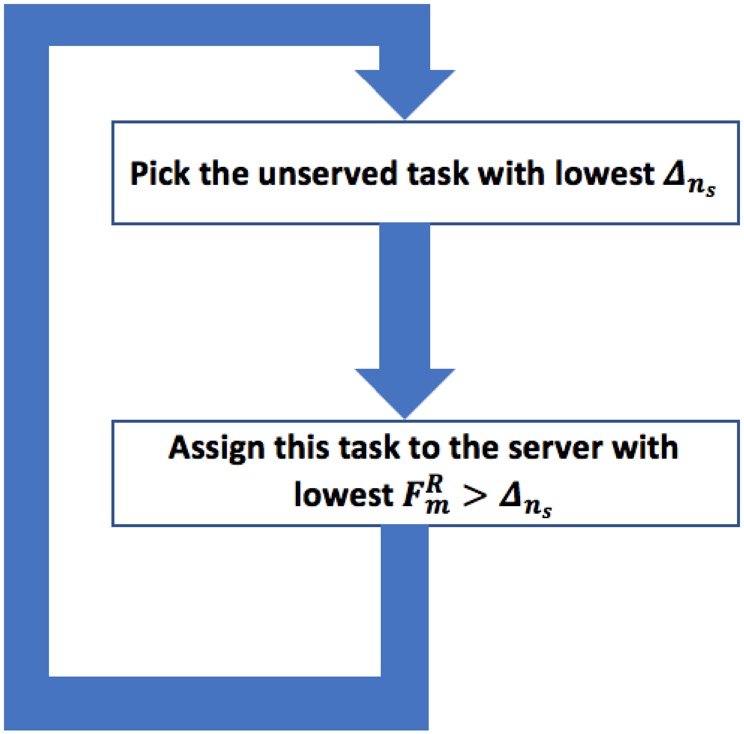


Load Distribution – sub-step 2

Load Distribution

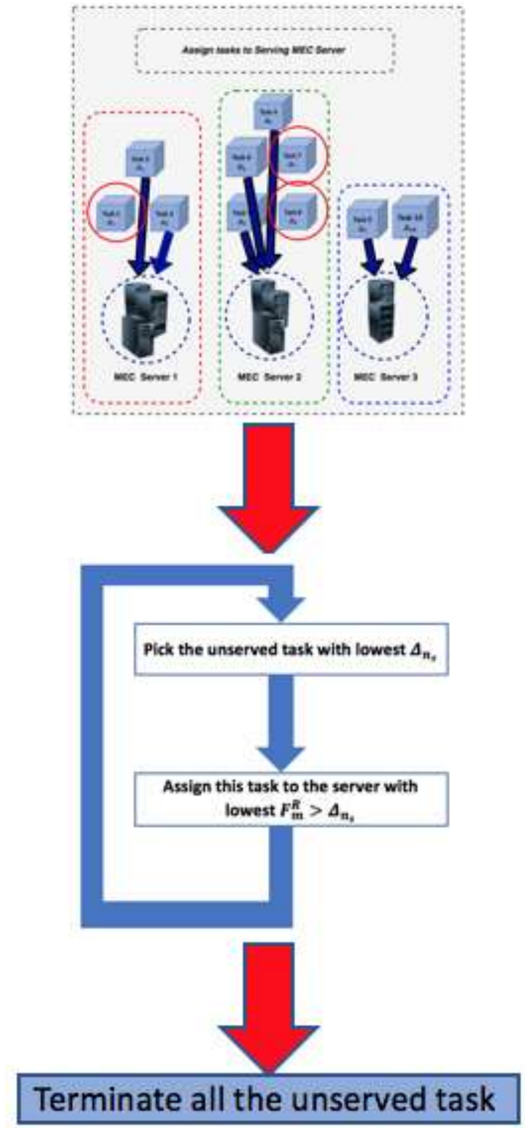
- After assignment in each server, we'll distribute unserved tasks to other servers with adequate resources.





Load Distribution – sub-step 3 & Computation Resource Allocation

- **Load Distribution**
  - First, assigning tasks to their Serving MEC Server in ascending order of Δ_{n_s}
  - After assignment in each server, we'll distribute unserved tasks to other server with adequate resources.
  - **Finally, terminating all the unserved tasks.**



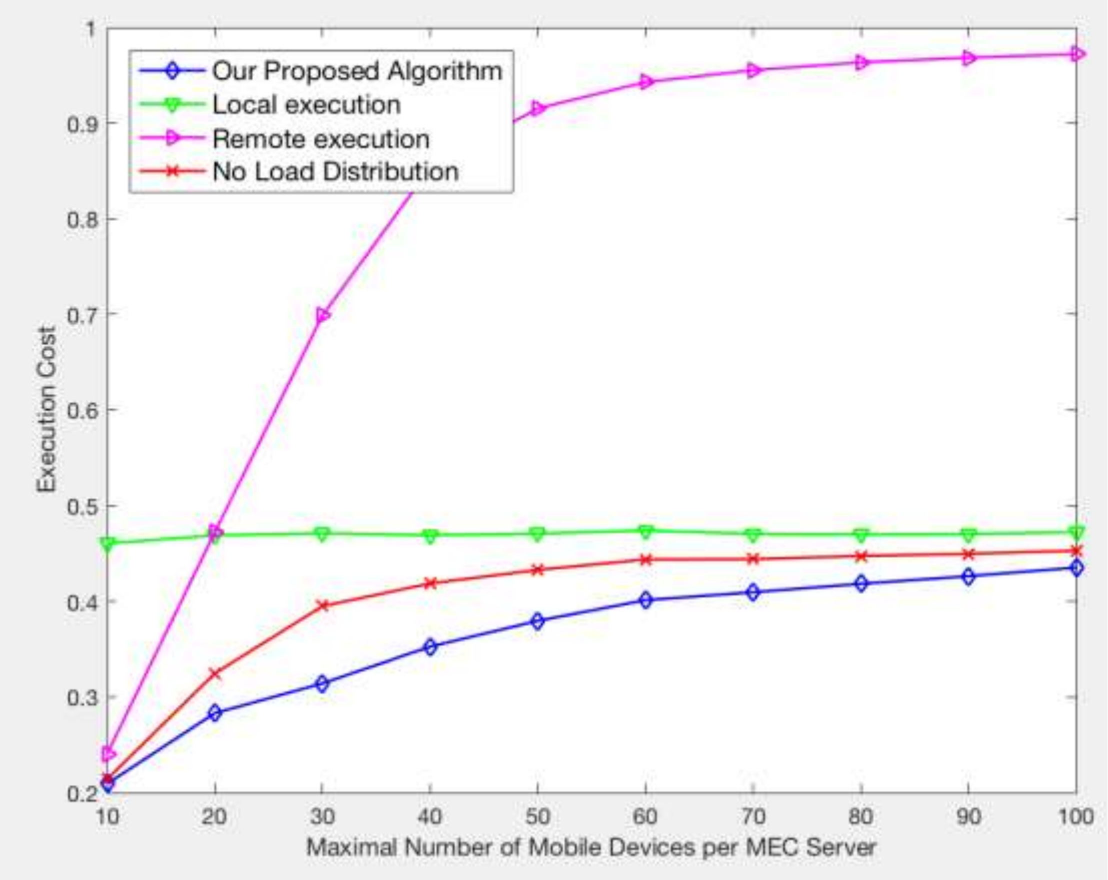
- **Computation Resource Allocation**
  - Adopting Lagrange Multiplier.

Simulation Settings

- **Scenario:**
  - Number of the BSs $M = 5$
  - Number of the sub-channel $H = 15$
  - Bandwidth of each sub-channel $B = 1.5 \times 10^6$ Hz
- **Comparison schemes:**
  - **Local execution**
    - Tasks can be executed only on local mobile device
  - **Remote execution**
    - Tasks can be executed only on MEC system.
  - **No Load Distribution**
    - Tasks can be executed on local mobile devices or their Serving MEC Server.

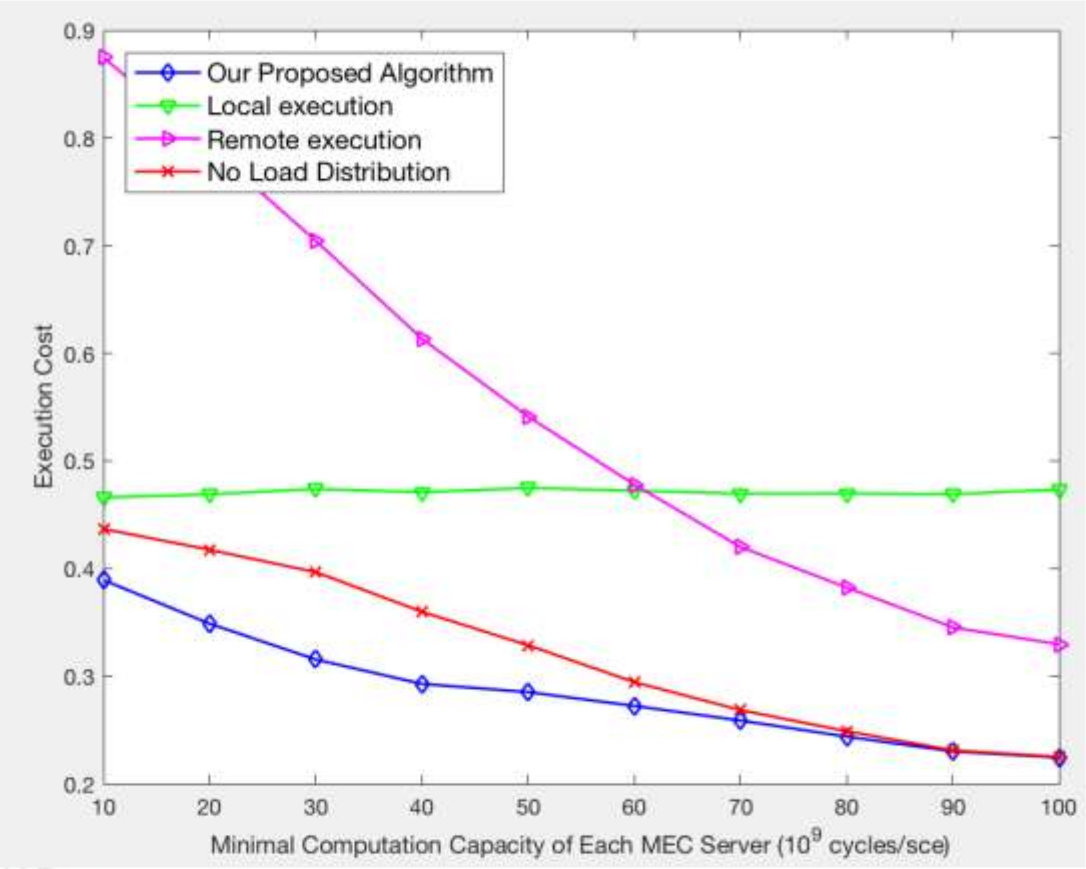
Simulation Results
(Maximal Number of Mobile Devices per MEC Server)

- Following Uniform Distribution $unif\{x - 10, x\}, x = maximal\ number$.
- Our algorithm achieves best QoS due to the consideration of **offloading decision** and **load distribution**.



Simulation Results
(Minimal Computation Capacity of Each MEC Server)

- No load Distribution scheme will get close to our algorithm with the increment of resources.



Conclusion

- **We discuss three issues in MEC system.**
  - **Task offloading**
  - **Load distribution**
  - **Resource allocation**
- **Formulating a cost minimization problem.**
  - To take **delay tolerance** and **remaining energy** into account.
- **Our solution is more efficient and consistent with reality.**
  - Taking **delay tolerance** and **remaining energy** into account.
  - We consider **multi-server system**.

Thanks for your attention!