Dynamic Virtual Machine Placement and Migration for Multi-access Edge Computing

Wei Wang

BUPT Ph.d candidate & UC Davis visiting student
Email: weiw@bupt.edu.cn, waywang@ucdavis.edu

UCDAVIS

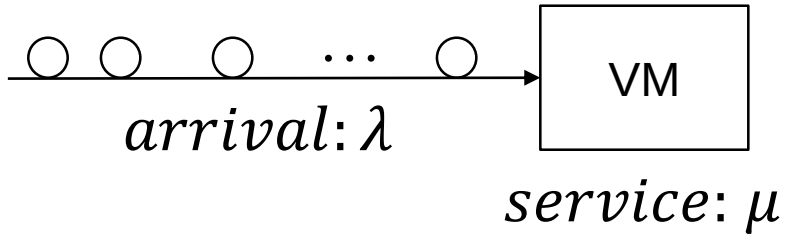
Group Meeting, Aug. 18, 2017

Contents

- Initial VM placement and workload assignment
- Workload variance
- Approaches for Workload variance
- Summary

Initial VM placement and workload assignment

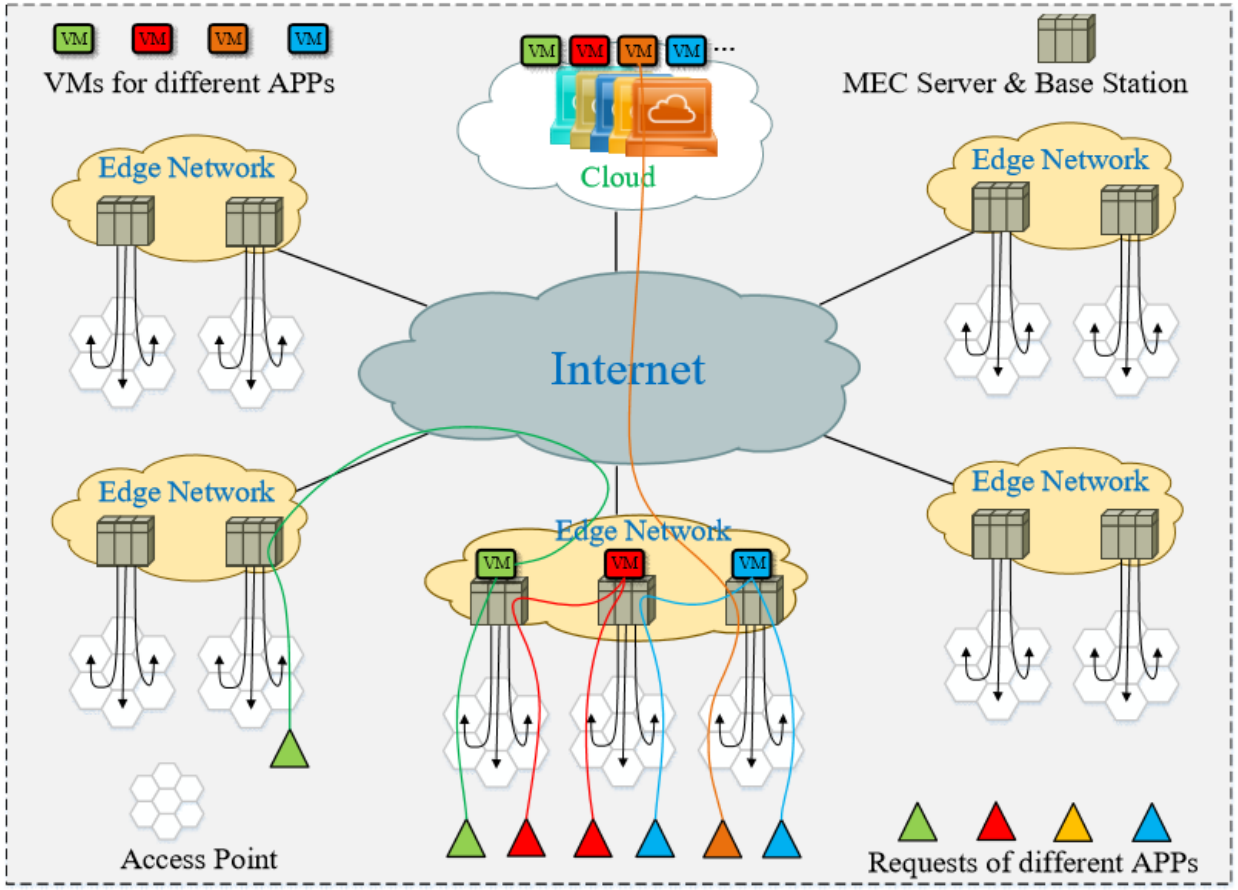
- To support a new MEC APP in VM-based MEC system, operators need to make two decisions to meet APP's latency requirement:
- Number of VMs at each MEC server.
- Amount of VMs' service resource for each given workload.
- Major parts of latency:
- Propagation
- queueing and processing
- Network propagation latency is set by location of service resource



$$T = \frac{1}{\mu - \lambda}$$

Initial VM placement and workload assignment

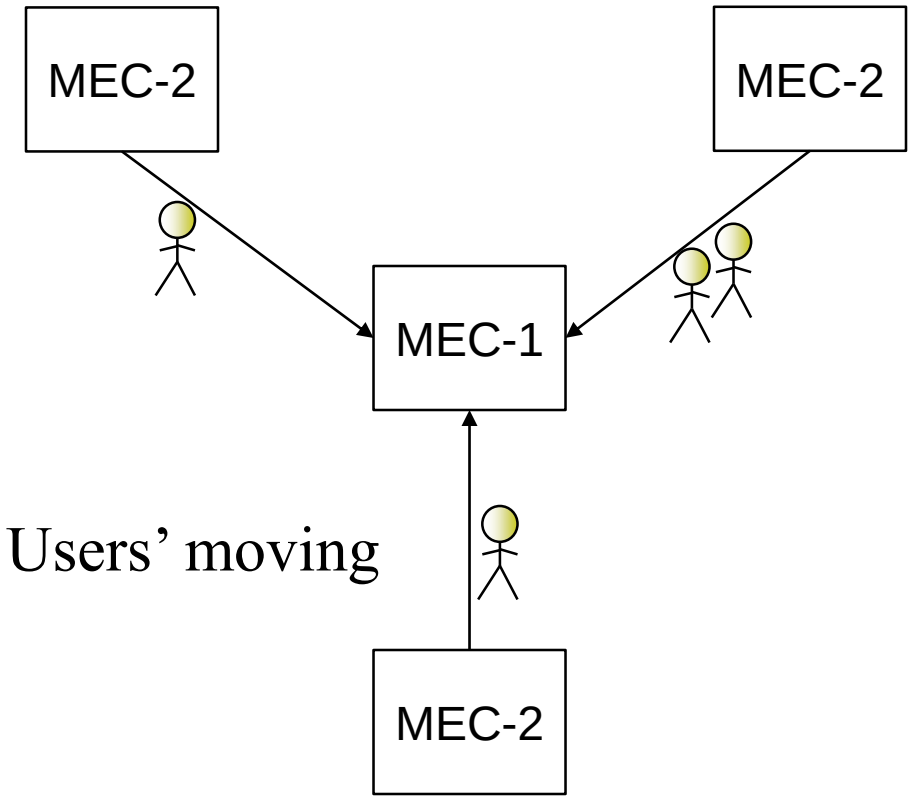
- In a real MEC system, there are:
- Multiple MEC servers
- Multiple APPs
- Multi workloads from each MEC server for each APP
- We have considered:
- MEC servers' hardware capacity
- APPs' heterogeneous latency requirements
- Given workloads for each APP from each MEC server
- To place VMs and assign workloads to minimize hardware consumption



Workload Variance

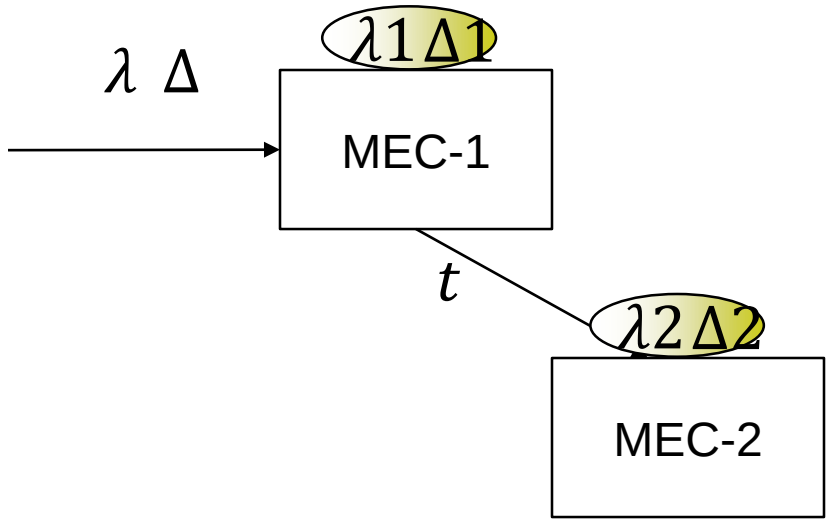
- Due to human activities, e.g.:
- Holidays
- Special Events
- APPs' workloads from each site are not fixed; instead, they vary in space and time dimensions.

- Workload, for App *a,* original from MEC node *v,* is expressed as:
- $L_v^a + \Delta_v^a$



How to deal with workload variance?

- **Assign extra load to existing VMs, which is serving base requests.**
- Issues for assigning extra load:
- Assign extra load to which VM?
- How much load can be assigned to each target VM?
- Parameters to consider:
- Latency requirement
- Existing load at each target VM



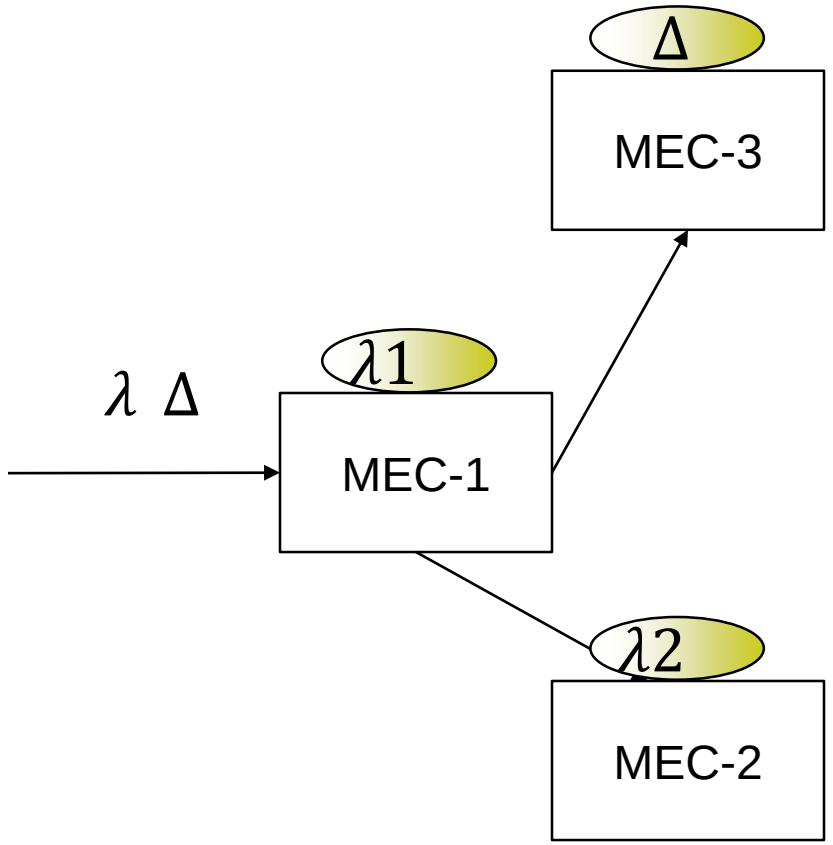
- Extra considerations:
- QoS degradation of loads on existing VMs.

Assign extra loads to existing VMs

- On an existing VM, whose service rate is μ
- Current load: λ_c
- Propagation time of λ_c: t_c
- Queuing and processing time of λ_c: $\frac{1}{\mu-\lambda_c}$
- Extra load: Δ_e
- Propagation time of Δ_e: t_e
- Queuing and processing time of λ_c and Δ_e: $\frac{1}{\mu-\lambda_c-\Delta_e}$
- Latency constraints:
- $\frac{1}{\mu-\lambda_c-\Delta_e} + t_c < T$
- $\frac{1}{\mu-\lambda_c-\Delta_e} + t_e < T$

How to deal with workload variance？(cont.)

- **Migrate spare/light-loaded VM(s), which is out of *Δ*'s reach to some nearby MEC servers.**
- Issues for migrating VMs
- Which VM can be migrated
- Where is the destination of each migrated VM
- How to deal with the existing load on migrated VMs
- Parameters to consider:
- Latency requirement
- Hardware capacity
- Extra considerations
- Time for placing new VMs
- QoS degradation before new VMs are ready

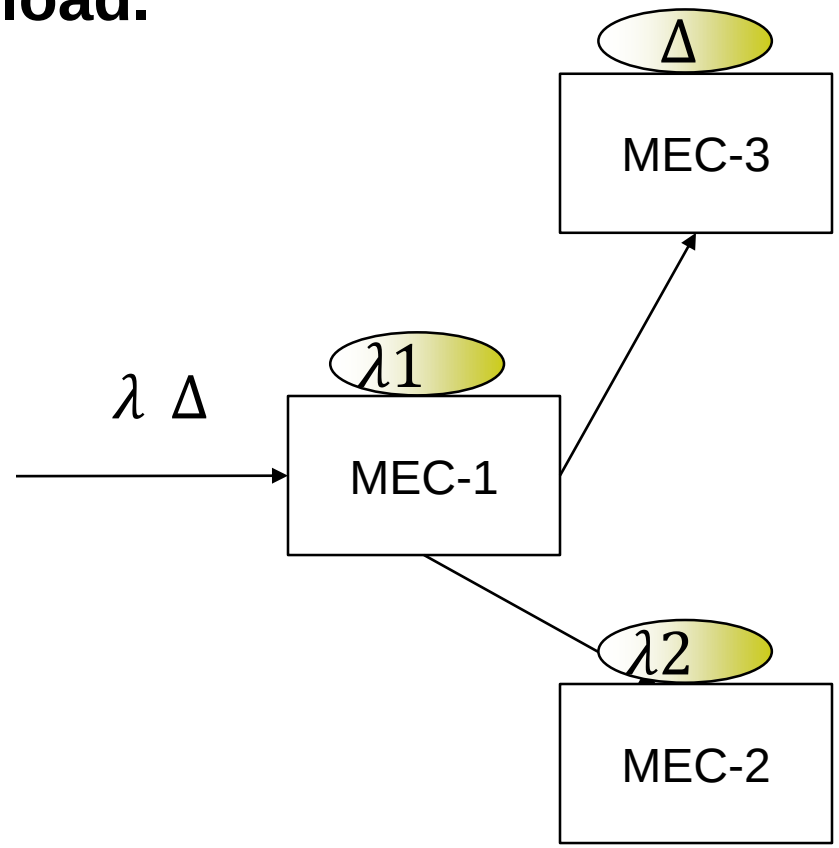


Migrate existing VMs

- An candidate VM, whose service rate is μ,
- current load: λ_c
- Hardware capacity constraints:$C_a \leq H_s$
- **Option A**: take λ_c as extra load and assign it to other VMs; migrate a spare target VM. Ref to last page to do extra load assignment.
- Propagation time of extra load Δ_e: t_e.
- Latency constraint: $\frac{1}{\mu-\Delta_e} + t_e < T$
- **Option B**: migrate VM with current load.
- Propagation time of current load after migration: t_{cm}
- Propagation time of extra load Δ_e: t_e
- Latency constraint: $\frac{1}{\mu-\Delta_e-\lambda_c} + t_{cm} < T$, $\frac{1}{\mu-\Delta_e-\lambda_c} + t_e < T$

How to deal with workload variance？(cont.)

- **Place new VMs to accommodate extra load.**



- Issues for placing new VMs:
- Where to place new VMs?
- How many VMs are need?
- Parameters to consider:
- Latency requirement
- Hardware capacity
- Extra considerations:
- Time for placing new VMs
- QoS degradation before new VMs are ready
- Extra hardware consumption

Place new VMs

- At a target MEC server *v:*
- Place m_a new VMs for APP *a*
- Hardware capacity constraints*:*
- $C_a * m_a \leq H_s$

- Extra load for each new VM*:*Δ_e
- Propagation time of extra load Δ_e: t_e.
- Queuing and processing time of Δ_e: $\frac{1}{\mu - \Delta_e}$
- Latency constraint:
- $\frac{1}{\mu - \Delta_e} + t_e < T$

How to deal with workload variance？(cont.)

- How to serve extra loads during VM migration or placing
- Solution: assign extra loads temporarily to existing VMs.
- On each existing VM, which will take extra load:Δ_e
- Current load: λ_c
- Latency degradation: $\frac{1}{\mu-\lambda_c} - \frac{1}{\mu-\lambda_c-\Delta_e}$
- Affected amount of load: λ_c.
- Derogated duration: migration/placing time (waiting for further study)

- Network traffic incurred by VM migration/placing (waiting for further study)

How to deal with workload variance？(cont.)

- What if target MEC servers' hardware capacity is not sufficient to accommodate migrated or new VM(s)?
- Ask VMs of other APPs for help:
- Option A: migrate load in a target VM to other VMs, and delete target VM.
- Option B: migrate target VM with its load to another MEC server.

- Other considerations:
- What if extra load is negative?
- Should we recover original status after extra load is gone?
- Should we consider network resource?

Summary

- Three approaches from extra load: 1) assign extra load to existing VM(s), 2) migrate VM(s) for extra load, 3) place new VM(s) for extra load
- Assigning extra load to existing VM(s), cause Qos degradation of load in existing VMs in terms of latency, no extra hardware consumption;
- Migrating VM(s) for extra load, cause Qos degradation of load in migrated VMs, temp Qos degradation of load in existing VMs, no extra hardware consumption;
- Placing new VM(s) for extra load, cause temp Qos degradation of load in existing VMs, extra hardware consumption.

Thank you!

Wei Wang