Dynamic Migration Strategy for Mobile Multi-Access Edge Computing Services

Presented by: Bholanath Das Niloy, Muztoba Hasan Sinha
Ibtissam Labriji, Emilio Calvanese Strinati, Eric Perraud, Frederic Jol

Renault, CEA, Qorvo

May 11, 2023

The idea in one line

A brief overview

Develop a new algorithm that combines the **exponential-weight algorithm for exploration and exploitation (EXP3)** and the **Lyapunov optimization** to jointly and **proactively** decide ***when*** to trigger migration and **where** to migrate the service running on **Internet of Vehicles(IoV) Mobile Edge Hosts(MEH)** before the **handover(HO)** in order to achieve significant **cost saving(energy)** while meeting any given **risk, service continuity,** and **availability** targets for each **User Equipment (UE)**

Table of Contents

Background

1 Introduction

This paper was published in *2022 IEEE Wireless Communications and Networking Conference (WCNC)* as a follow up paper to ”Mobility Aware and Dynamic Migration of MEC Services for the Internet of Vehicles” published the previous year by the same authors. Co written by collaborators across the French industry and Government, namely Renault, one of the largest automobile companies in the world, CEA, the French Alternative Energies and Atomic Energy Commission, and Qorvo an American semiconductor company. All of them have stakes to optimize the power consumption and increase the quality of service of the Internet of Vehicles services.

The Vehicular Network

1 Introduction

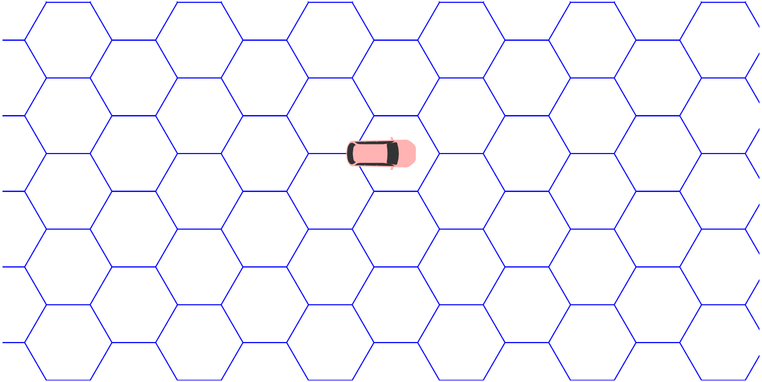


Table of Contents

2 Understanding the Problem (Prerequisites)

Vehicles or highly mobile User Equipment (UE) will interact with one another, while communicating with the network infrastructure and with the edge/cloud (Mobile Edge Host, MEH), where services are provided.

Application nature on IoVs

1. Dependent on 5G
2. Multi-Access Edge computing.
3. Stateful and Latency intensive applications.

Multi-Access Edge Computing (MEC) moves the computing of traffic and services from a centralized cloud to the edge of the network and closer to the customer.

MEC characteristics

1. Proximity
2. Ultra-low latency
3. High bandwidth
4. Virtualization

Limitations of the systems discussed thus far

1. *Mobile Edge Host*/basestation/eNB runs the application in a Virtual Machine(VM).
2. 5G base stations have low range
3. Applications are stateful

So what happens when a vehicle moves out of the base station (MEH) range?

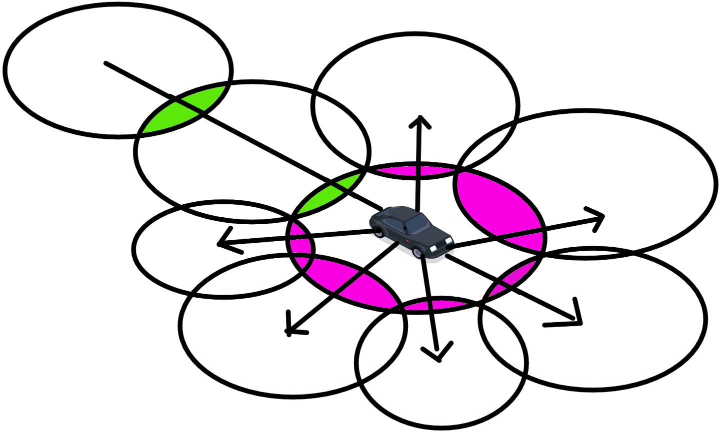
The process of transferring an ongoing session from one channel connected to the core network to another channel here. Here it is slightly different since it's an edge computing network. We need to ***migrate*** the VM handling the client application to another edge. But each MEH has multiple neighbors the process of transferring an ongoing session from one channel connected to the core network to another channel here. Here it is slightly different since it's an edge computing network.

⏩ Migration vs Handover

We need to migrate the VM handling the client application to another edge. But each MEH has multiple neighbors.

Handover

2 Understanding the Problem (Prerequisites)



The problem in Laymans terms

When a user moves from one base station/cell of a MEH to another, while using a stateful and computationally intensive application, how do we maintain their connection? This has a few sub problems

- Which neighboring cell do we migrate their state to?
- When do we migrate it so that the application is not hampered?

The Vehicular Network

2 Understanding the Problem (Prerequisites)

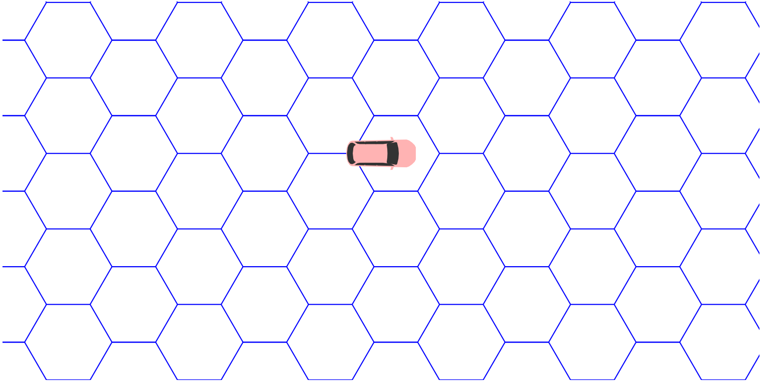


Table of Contents

3 Key Aspects/Considerations

So where do we migrate?

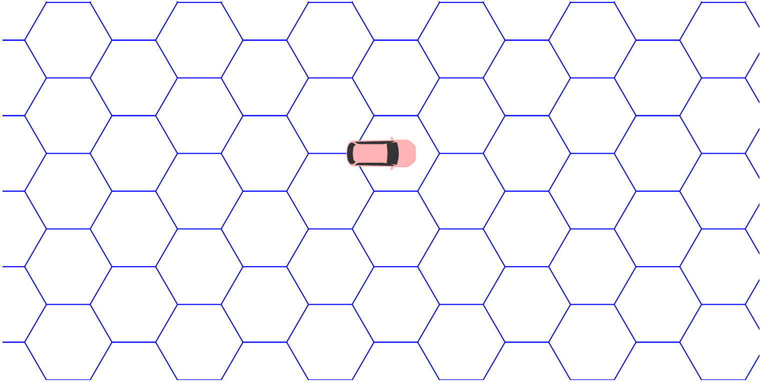
Each MEH has limited compute capacity for VMs and passing these VM is a huge energy burden on the system.

1. Multiple migrations mean more redundant energy cost and pressure on network
2. Wrong migrations mean risk of disconnection and/or high latency.

But is that enough?

The Vehicular Network

3 Key Aspects/Considerations



When do we migrate?

3 Key Aspects/Considerations

Passing VM is slower than the passing of a UE or the request for handover made by the

1. Remember the applications are stateful and latency intensive.
2. Slow migrations means disconnection or lag at the UE.

Therefore in order to mitigate these risks we **must** migrate the VM ***proactively***

Solution Key Aspects

Ths paper proposes a VM migration strategy which:

1. **Operates individually at each MEH**
2. **Tracks user trajectories** User trajectory tracking
3. **Multi Modal Mobility Estimation:**
   - Inside MEH: Position is inferred from radio link measurements and processed using CNN and RNNs
   - Across MEH: sequence of previously visited eNBs is statistically characterized via Markov Chains.

4. **Mobility estimation and Service Migration:** An original learning architecture described in [3] which combines **NNs and MC** predictors and returns soft predictions in form of a ***prediction vector***(a probability distribution over tehe seto of nearby eNBs/MEHs towards which each vehicle may handover). The algorithm is **Decentralized**, **online** and **Lyapunov based**. It determines *How many* VMs are migrated *Where*

Table of Contents

4 Solution and it's Goals

The Goal: Joint Mobility Estimation and MEC Service Migration: ”when to migrate”

The paper proposes a *novel, lightweight* service migration algorithm that decides the *optimal instant* to trigger migration (when) depending on how fast the vehicle is moving and the MEHs where these instances are to be migrated.

Why this solution is novel

1. Improved mobility prediction by incorporating the movement of UE inside the radio cells
2. Dynamic adaptation of VM replication strategy based o mobility prediction estimates
3. Lightweight optimization framework.
4. The problem of when to trigger migration is completely novel

Table of Contents

5 System Model

Simulation

The proposed solution is numerically assessed in a real-life vehicular mobility scenario, using the “simulation of urban mobility” (SUMO) [13] to emulate thousands of vehicles moving within the city of Cologne, and simulating a realistic 5G network with densely deployed eNBs (and co-located MEHs)

MEH

System Description

Each MEH has,

1. several VMs to execute the users' services.
2. a MEC platform
3. a mobility prediction unit
4. a VM migration control unit (including admission control)
5. a virtualization infrastructure

Network

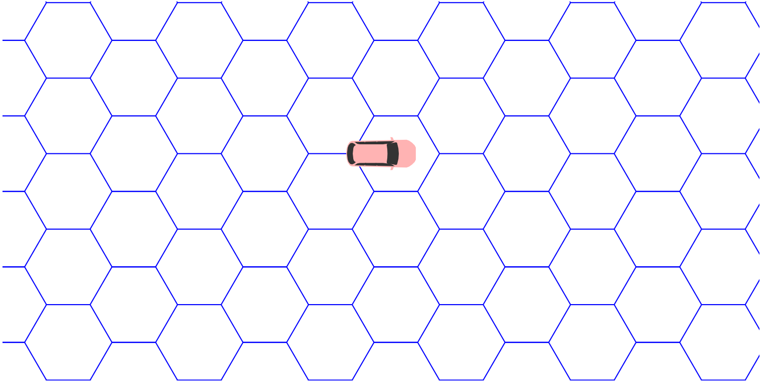
System Description

The network has has a

- set $\mathcal{S}$ eNBs each with a MEH
- set $\mathcal{V}$ of vehicles which follow the road constraints. and associated with a VM.
- each MEH has N_{MEH} neighbours
- set of neighboring MEH available for transfer $\mathcal{N}_{MEH} = \{1, 2, ..., N_{MEH}\}$
- The mobility prediction unit returns a **vector** containing the probability distribution over $\mathcal{N}_{MEH}$
- set of vehicles about to leave the cell $\mathcal{A}(t)$
- each such vehicle $r \in \mathcal{A}(t)$

The Vehicular Network

5 System Model



Time

System Description

Our system model is time-dependent and operates in an online fashion. Time is slotted, with time slots of fixed duration τ. At each time slot, a cost function representing the energy expenditure is minimized subject to a risk constraint, seeking a good balance between energy minimization and service continuity.

Time considerations

Problem formulation

- Time slots are numbered and each indexed by t
- optimal instant is selected from a set $\tau_K = \{T_1, T_2, ..., T_k\}$
- Time is upper bounded by T^{HO}
- lower bound by $T^{notification}$
- migration is triggered at T_k

Network State at Migration time T_k

- set of remaining neighboring MEH candidates considering migration, $\mathcal{N}_{MEH}^k = \{1, 2, ..., N_{MEH}^k\}$
- it follows that $N_{MEH}^k \leq N_{MEH}$

Mobility Prediction Vector

Problem Formulation

- The mobility prediction vector associated with $\mathcal{N}_{MEH}$ is referred to as $p_r^k = (p_{r,1}^k, p_{r,1}^k, ..., p_{r,N_{MEH}}^k)$ and evolves according to T_k
- $p_{r,i}^k$ represents the probability that the vehicle requesting r hands over to MEH $i \in \mathcal{N}_{MEH}$
- $\sum_{i=1}^{N_{MEH}} p_{r,i}^k = 1$ at T_k

Indicator Function

Problem Formulation

We introduce a new variable to control when migration is triggered $\theta_k(t) = 1$ and $\sum_{k=1}^{K} \theta_k(t) = 1$ that means the instant chosen to trigger migration is T_k
We associate this set of indicator functions $\mathcal{I}_K = \{\theta_1(t), \theta_2(t), ..., \theta_K(t)\}$ to τ_K. To specify which time slot is chosen.

Energy required to migrate
Problem formulation

$$E_r(t) = \sum_{k=1}^{K} M_r^k(t)\psi_r\Theta_k(t)$$

where:

$M_r^k(t)$	Number of MEH to migrate to
$\Theta_k(t)$	Indicator if this time is selected for migration or not
ψ_r	Energy required to preform single migration requested by r

Total Energy Consumed for Container Migrations

Problem Formulation

$$E(t) = \sum_{r \in \mathcal{A}(t)} E_r(t)$$

Long Term Average Energy Cost of Migrations

What we need to minimize

$$\lim_{T \to +\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[E(t)]$$

In short we want to minimize the long term expected value of energy cost

The solution?

We can just not!

If we don't migrate there is no cost of migration!

Risk

Another Approach

We introduce another measure, The risk that is the probability of disconnection in case the container is migrated to only the MEH that will not be visited after the handover

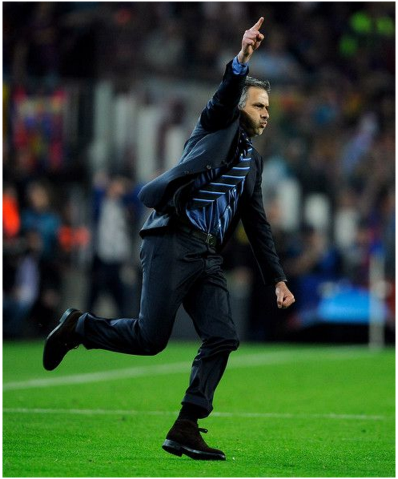
$$\zeta_r(t) = \sum_{k=1}^{K}(1 - \sum_{i=1}^{M_r^k} p_{r,i}^k(t))\theta_k(t)$$

Average risk across all requests in t is:

$$\overline{\zeta_r(t)} = \frac{1}{N(t)} \sum_{r\in\mathcal{A}(t)} \sum_{k=1}^{K}(1 - \sum_{i=1}^{M_r^k} p_{r,i}^k(t))\theta_k(t)$$

The solution?

We can just migrate to all of them!



If we migrate to all of them we never have a chance of disconnection!

We need to counterbalance these two consideration to reach an equilibrium.

Table of Contents

6 Constraints on the System

Average Long Term Risk Target

C1: Average Long Term Risk

The average long term risk $\overline{\zeta_r(t)}$ is kept close to a predefined target $\xi > 0$
To do this we introduce a virtual queue $Z(t)$ which is incremented each time $\overline{\zeta_r(t)}$ exceeds ξ by $\overline{\zeta_r(t)} - \xi$

$$Z(t+1) = max\{Z(t) + \overline{\zeta_r(t)} - \xi, 0\}$$

Satisfying the Constraint

C1: Average Long Term Risk

$$\lim_{t \to +\infty} \frac{\mathbb{E}[Z(t)]}{t} = 0$$

Long-Term Availability

This is expressed by $T_r^{Mig} \leq T_{r,k}^{Rem}(t) \forall r$. Where $T_{r,k}^{Rem}(t)$ is the time remaining for the vehicle to reach next MEH. T_r^{Mig} is the migration time required to migrate the container handling r.

Average Long Term Availability

C2: Average Long Term Availability

Our aim is to control, every time slot t, the rate of requests for which $T_r^{Mig} \leq T_{r,k}^{Rem}(t) \forall r$ is met. ($\infty\{.\}$ is the indicator function)

$$\overline{\Gamma(t)} = \frac{1}{N(t)} \sum_{r \in \mathcal{A}(t)} \sum_{k=1}^{K} \infty\{T_r^{Mig} \leq T_{r,k}^{Rem}(t)\} \theta_k(t)$$

Average Long Term Availability Target

To do so, we force $\overline{\Gamma(t)}$ to be close to a predefined target percentage γ. We define another virtual queue $Y(t)$ with the following update equation:

$$Y(t+1) = max\{Y(t) + \gamma - \overline{\Gamma(t)}, 0\}.$$

When $\overline{\Gamma(t)}$ goes below γ, $Y(t)$ is incremented by γ - $\overline{\Gamma(t)}$.

Mean rate stability/satisfying the constraint

C2: Average Long Term Availability

The mean rate stability of $Y(t)$ will allow us to control the availability of the migrated containers as vehicles hand over to new MEHs. It is mathematically expressed by:

$$\lim_{t \to +\infty} \frac{\mathbb{E}[Y(t)]}{t} = 0$$

Table of Contents

$$\lim_{t \to +\infty} \min_{\Omega(t), t \in \{0,\ldots,T-1\}} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[E(t)]$$

Subject to:

- $\lim_{t \to \infty} \frac{\mathbb{E}[Z(t)]}{t} = 0$
- $\lim_{t \to \infty} \frac{\mathbb{E}[Y(t)]}{t} = 0$
- $\theta_k(t) \in \{0, 1\}$, $\sum_{k=1}^{K} \theta_k(t) = 1$, $\forall k$, t
- $M_r^k(t) \in \{0, \ldots, N_{MEH}^k\}$, $\forall k$, t

Table of Contents

8 Lyapunov

$$L(\boldsymbol{\Theta}(t)) \triangleq \frac{1}{2}\left[Z(t)^2 + Y(t)^2\right]$$

From here we define the one-slot conditional Lyapunov drift which represents the expected change in the Lyapunov function over one slot which is as follows:

$$\Delta(\boldsymbol{\Theta}(t)) \triangleq \mathbb{E}[L(\boldsymbol{\Theta}(t+1)) - L(\boldsymbol{\Theta}(t)) \mid \boldsymbol{\Theta}(t)]$$

Minimizing $\Delta(\boldsymbol{\Theta}(t))$ provides queue stability.

At this point, the problem is only half solved as virtual queues only help to ensure the desired time average constraints are met. Thus we encorporate the energy cost through a control parameter $V > 0$ arriving at our desired *drift-plus-penalty* expression.

$$\Delta(t) = \Delta(\boldsymbol{\Theta}(t)) + V\mathbb{E}[E(t) \mid \boldsymbol{\Theta}(t)]$$

$$
\begin{aligned}
\min_{\Omega(t)} \phi(t) = \sum_{k=1}^{K} \theta_k(t) \sum_{r \in \mathcal{A}(t)} \Bigg[& V M_r^k(t) \psi_r \\
& + \frac{Z(t)}{N(t)} \left(1 - \sum_{i=1}^{M_r^k(t)} p_{r,i}^k(t) \right) \\
& - \frac{Y(t)}{N(t)} 1 \left\{ T_r^{Mig} \leq T_{r,k}^{Rem}(t) \right\} \Bigg]
\end{aligned}
$$

Subject to:

(a) $\theta_k(t) \in \{0, 1\}, \sum_{k=1}^{K} \theta_k(t) = 1, \forall k, t,$

(b) $M_r^k(t) \in \mathcal{N}_{MEH}^k \forall k, r.$

Table of Contents

9 EXP3

How to optimize this?

EXP3 is the most well-known algorithm to handle non-stochastic Multi-Armed Bandit (MAB) problems. It maintains a set of weights $(\omega_k(t))$, which here are used to update the probability vectors. They are updated as follows:

$$p_k(t) = (1-\rho)\frac{\omega_k(t)}{\sum_{k=1}^{K}\omega_k(t)} + \frac{\rho}{K}$$

Payoff Function

$$
\begin{aligned}
x_{k^*}(t) &= \sum_{r \in \mathcal{A}(t)} V M_r^{k^*}(t)\psi_r + \frac{Z(t)}{N(t)}\left(1 - \sum_{i=1}^{M_r^{k^*}(t)} p_{r,i}^{k^*}(t)\right) \\
&\quad - \frac{Y(t)}{N(t)} 1\left\{T_r^{Mig} \leq T_{r,k}^{Rem}(t)\right\} = \sum_{r \in \mathcal{A}(t)} x_{k^*,r}(t).
\end{aligned}
$$

Updating the weights

$$\omega_{k^*}(t+1) = \omega_{k^*}(t)e^{\frac{\rho x_{k^*}(t)}{K p_{k^*}(t)}}$$

Table of Contents

10 The Algorithm

Algorithm 1: CAEXP3

Initialization: Let $\rho \in [0, 1]$

Initialize the weights of the options: $\omega_k(t) = 1, \forall k$,

repeat

- **Step 1:** Calculate $p_k(t)$ for every option, $Eq.(16)$,
- **Step 2:** Select an option k^* through the probability distribution computed previously,
- **Step 3:** Observe the payoff of the selected option k^* $Eq.(17)$,
- **Step 4:** Update $\omega_k(t+1)$ as in $Eq.(19)$,
- **Step 5:** Update $Z(t+1)$ and $Y(t+1)$, according to $Eq.(5)$ and (8) respectively,

until $t > T$;

Algorithm 2: CAEXP3.1

Initialization: Let $c = \frac{K \ln(K)}{e-1}, G_k(t=0) = 0$, epoch $= 0$,
$g_{\text{epoch}} = c4^{\text{epoch}}, \rho = 2^{-\text{epoch}}$

repeat

  Step 1, Step 2, Step 3, Step 4, Step 5 of Algorithm 1,

  Step 6: Update the total payoff for every option k, $Eq.(20)$

  if $\max_k G_k(t) > g_{epoch} - K\rho$ **then**

    $epoch = epoch + 1,\ g_{epoch} = c/4^{epoch},\ \rho = 2^{-epoch}$;

until $t > T$;

Table of Contents

Results

Evolution of Availability

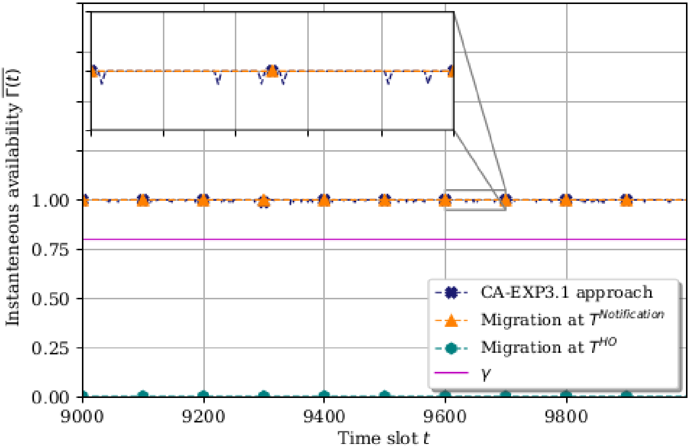


Figure: Temporal Evolution of availability

Results

Evolution of Risk

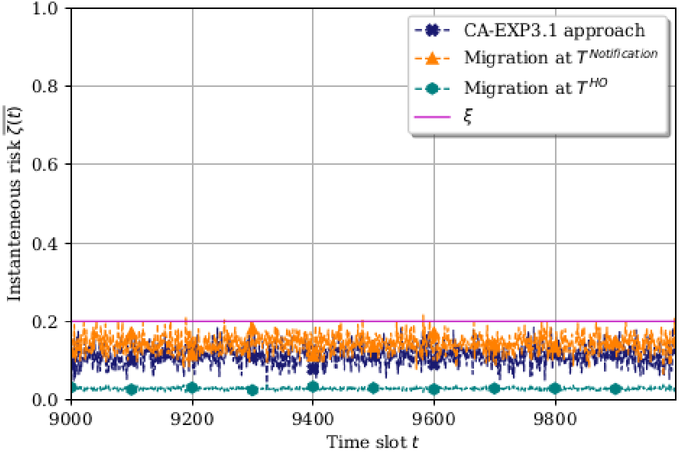


Figure: Temporal Evolution of Risk

Results

Energy Consumption

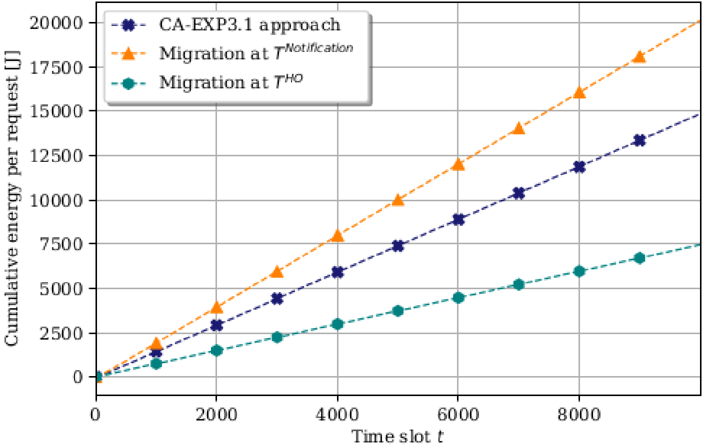


Figure: Energy Consumption

Results

Temporal Evolution of Probability Vectors

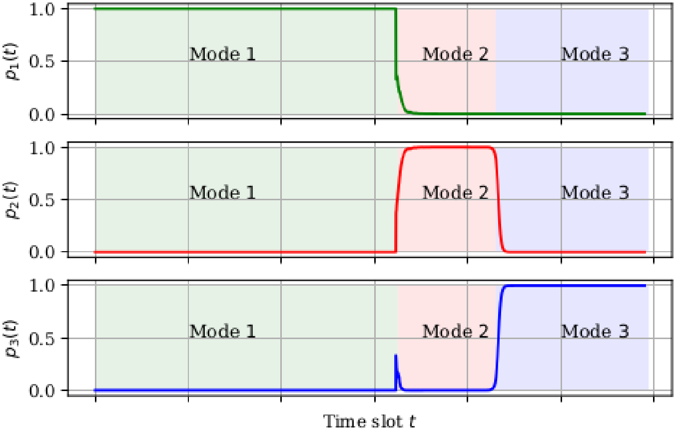
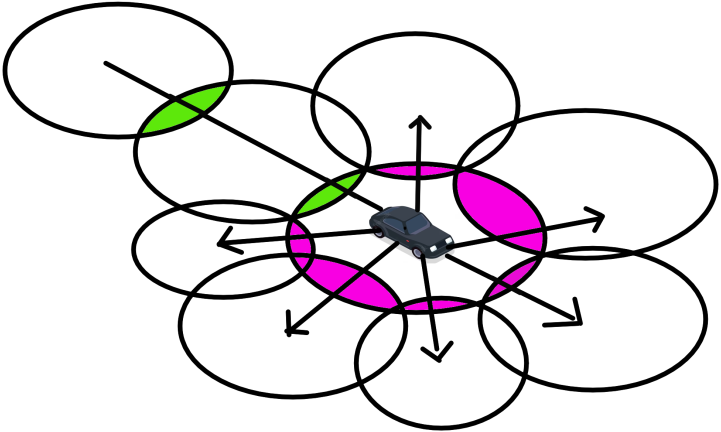
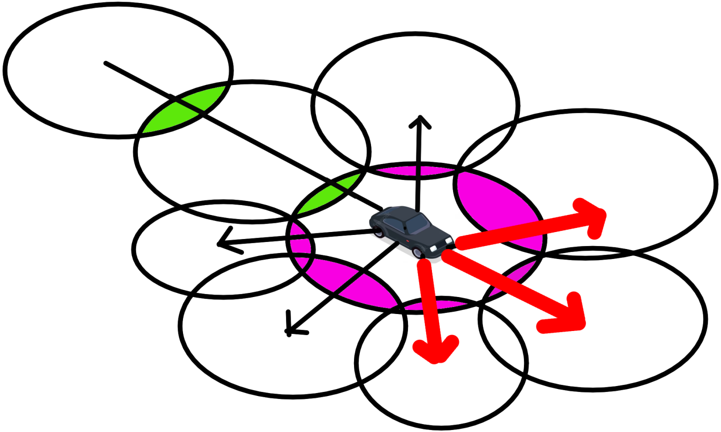


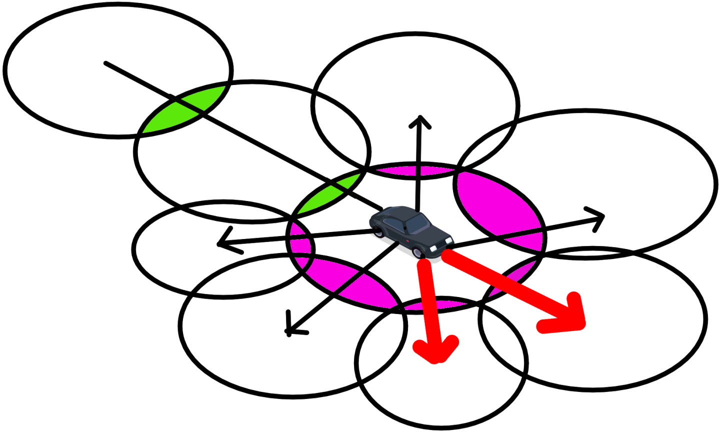
Figure: Temporal Evolution of Probability Vectors



Summary(contd.)



Summary(contd.)



Q&A

Thank you for listening!
Your feedback will be highly appreciated!

Vs related work

addendum

This problem of service continuity has been tackled many ways

1. Backhaul links
2. Migration
   - 2.1 reactive migration
   - 2.2 proactive migration
     - 2.2.1 UE-MEH distance based predictions
     - 2.2.2 Probabilistic
       - 2.2.1 Fixed
       - 2.2.2 Online

back

User Trajectory

addendum

The use of two dimensional antenna arrays, with the introduction of the massive multiple-input multiple-output (MIMO) physical layer technology, allows terminal positioning via angle-of-arrival (AoA) measurements without the burden of performing triangulation [12] back