# Migration-Aware Optimized Resource Allocation in B5G Edge Networks

Tadeus Prastowo*, Ayub Shah*, Luigi Palopoli*, Roberto Passerone*, Giuseppe Piro+
* Dep. of Information Engineering and Computer Science – DISI, University of Trento, Italy
+ Politecnico di Bari, Italy – DEI

## Abstract

5G and beyond 5G communication systems are evolving for computation-intensive and communication-sensitive applications with diverse Quality-of-Service requirements on processing, bandwidth, latency, and reliability. This work focuses on an ultra-dense edge network with Multi-access Edge Computing facilities, serving agents that execute their tasks by touring the cells. Specifically, we propose a novel methodology for optimally and flexibly managing task offloading in the context of heterogeneous computing and communication services required by real-time robotic applications. The proposed approach takes the number of admitted service migrations and the QoS upper and lower bounds as binding constraints. We model the QoS evolution based on the agent positions, the MEC servers serving the agents, the QoS requirements, the communication capabilities in the edge network, and the computing capabilities of the servers. The model is formalized as a mixed-integer linear program to obtain an optimal schedule for the service migrations and communication and computation bandwidth allocation.

## Motivation

The research on the next generation of mobile networks is chasing the ambitious objective to jointly support, within a flexible and powerful communication and computing infrastructure, a very large number of heterogeneous services. In this context, an effective management of task offloading is crucial to deliver high-quality services, hence the need for the optimal management of computing and communication resources at the network edge.

A novel methodology (shown in fig. 1) is adopted to manage B5G task offloading optimally and flexibly in the context of real-time applications, which are represented well in the robotic domain. We considered an industrial automation scenarios, where each robot is shown to be connected to a network attachment points that provides wireless connectivity, a simplest case is shown in Fig. 2. The attachment points are connected to an edge network with computing capabilities provided by MEC servers. The agents then connect to one of the available MEC servers through edge links with fixed communication capabilities.

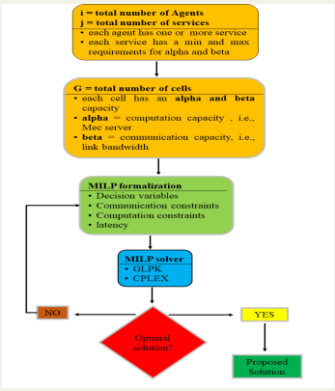



Figure 1

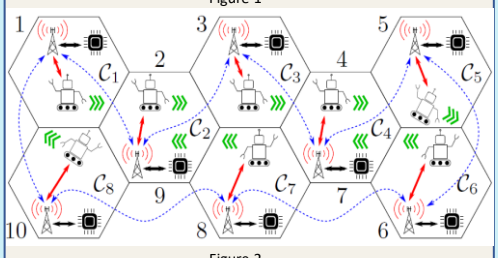



Figure 2

## SYSTEM MODEL

- The network is composed of a set of cells $|\mathbb{G}|$, $\{\mathcal{C}_c \in \mathbb{G}\}$, a set of moving robots $\mathbb{A}$, $\{\mathcal{A}_i \in \mathbb{A}\}$, a set of services $\mathbb{M}$, $\{\mathcal{M}_{i,j} \in \mathbb{M}\}$, and a time horizon $\mathbb{H} \subset \mathbb{N}$.
- Each cell has a total computation and communication bandwidth capacity respectively represented by:
  - $\Phi_c \in \mathbb{R}^+$
  - $\Psi_c \in \mathbb{R}^+$

## System Model Cont.

- Migration cost of every service from one cell to another is represented by:
  - $\varepsilon_{i,j} \in \mathbb{R}^+$
- *The final QoS objective function of our model is defined by:*
  - $\varphi_{i,j,t} = \varphi^+_{i,j}(\alpha_{i,j,t}, \beta_{i,j,t}, \lambda_{i,j,t}) - \varepsilon_{i,j}.\mathbb{m}_{i,j,t}$
- Where $\varphi^+_{i,j}$ is the QoS function of three variables, computation, communication, and latency, and $\varepsilon_{i,j,t}.\mathbb{m}_{i,j,t}$ denotes the service migration cost.
- $\lambda_{i,j,t}$ represents the communication latency time between any service $\mathcal{M} \leftrightarrow \mathcal{A}$
- The constraints puts a bound on communication and computation bandwidth allocated to any service in a cell $\mathcal{C}_c$ at time $t$ must be less than the total capacity of cell.
  - $\sum_{\mathcal{M}_{i,j} \in \mathbb{M}} \psi_{i,j,c,t} \leq \Psi_c$
  - $\sum_{\mathcal{M}_{i,j} \in \mathbb{M}} \phi_{i,j,c,t} \leq \Phi_c$
- Variables
  - $\mu_{i,j,c,t}$ is One (zero) if $\mathcal{M}_{i,j}$ is (not) in $\mathcal{C}_c$ at time $t$
  - $\alpha_{i,j,t}$ is $\mathcal{M}_{i,j}$computation bandwidth at time $t$
  - $\beta_{i,j,t}$ is $\mathcal{M}_{i,j}$ communication bandwidth at time $t$
- Finally the total QoS depends on:
  - MEC server computation capacity in GIPS and communication capacity in Mbps
  - Latency
  - Migration cost

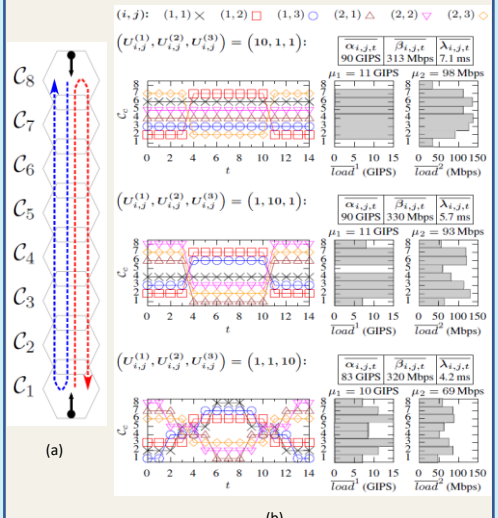
## Scenario 1 & Results

**A small 8-by-1 mesh network with 8 cells, 15 time points, 2 agents, and 3 VMs/agent is shown below in fig. 2.**

- $\Phi_c = 100$ GIPS, $\alpha^{min}_{i,j} = 15$ GIPS, $\alpha^{max}_{i,j} = 90$ GIPS, $\beta^{min}_{i,j} = 150$ Mbps, $\beta^{max}_{i,j} = 900$ Mbps, and $\lambda^{max}_{i,j} = 23$ ms.

The result shows:

- The positions of the VMs over time on the left.
- The load on MEC server average processing in the middle with the mean $\mu_1$ at the top of each plot.
- The mesh-network link average traffic on the right with the mean $\mu_2$ at the top.
- The average computation B.W., Communication B.W., and latency shown in top of table in row heading.




(a)

(b)

Figure 3

## Scenario 2

**A star/mesh topology with 100 cells, 19 time points, 40 agents, 3 VMs/agent.**

- Our formulation effectiveness is shown on the VM migration frequency, outage count, average computation bandwidth, and latency (system KPIs) for different VM migration costs $\varepsilon_{i,j}$.
- We used baseline scenarios that always migrate the VM's in the cell where their agent is. This gives every VM the lowest latency but the highest migration frequency and possibly some outage times

## Results Scenario 2

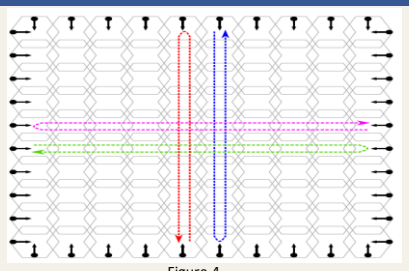


Figure 4

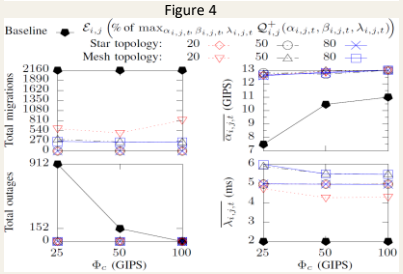



Figure 5

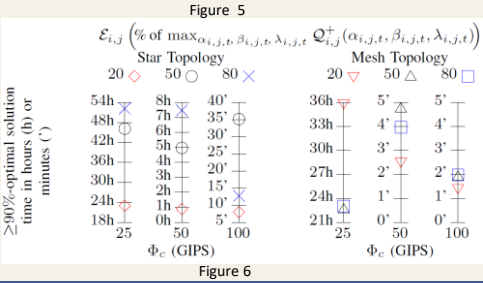



Figure 6

## Conclusion and Future Work

In this paper, we have considered the problem of allocating resources at the edge of a B5G network to real-time services optimally by formulating a MILP whose decision variables are the amount of computation and communication resources and the MEC servers to execute the VMs providing the services at each time point. Using state-of-the art optimization tools allows us to treat problems of reasonable size in the number of cells and agents when the agent trajectories are known up-front, and the optimization can be performed offline before starting the system operations. When the size of the problem grows or when the system is highly dynamic and requires online optimization, heuristic approaches are needed to produce high-quality sub-optimal solutions. This is one of the most promising research areas that we reserve for our future investigations, which include futuristic scenarios where the base stations are mobile (e.g., aerial or terrestrial vehicles) and need an optimal decision on their positions as well.

## Acknowledgements

This work has received funding from the Italian Ministry of Education, University and Research (MIUR) through the PRIN project no. 2017NS9FEY entitled "Realtime Control of 5G Wireless Networks: Taming the Complexity of Future Transmission and Computation Challenges". The views and opinions expressed in this work are those of the authors and do not necessarily reflect those of the funding institution.

## References

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," IEEE Netw., vol. 34, no. 3, pp. 134–142, 2020.

[2] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," IEEE Access, vol. 8, pp. 133 995–134 030, 2020.

[3] Q. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," IEEE Access, vol. 8, pp. 116 974–117 017, 2020.