Pricing Based MEC Resource Allocation for 5G Heterogeneous Network Access

Virgilios Passas[†⋆], Nikos Makris[†⋆], Vasileios Miliotis[†⋆], Thanasis Korakis[†⋆]
[†]Department of Electrical and Computer Engineering, University of Thessaly, Greece
[⋆]Centre for Research and Technology Hellas, CERTH, Greece
Email: {vipassas, nimakris, vmiliotis, korakis}@uth.gr

***Abstract*—Multi-access Edge Computing (MEC) is expected to play an important role in next generation networks, as it is able to provide resources accessible through multiple wireless technologies located close to the network edge. MEC is therefore an enabler for low-latency applications, allowing novel time critical services to be offered through mobile networks. Nevertheless, hosting multiple service providers over the same physical infrastructure shall carefully consider the needs of the MEC enabled applications. Moreover, different suggested placements for the MEC service provide fertile ground for the differentiation of the hosted service providers. In the meantime, the integration of multiple technologies in the wireless access of the MEC concept is increasing the complexity for efficient allocation of network and computational resources among the involved stakeholders. In this work, we model the MEC resource allocation problem for different service providers by using a pricing scheme. We consider two different deployments for the MEC services when considering a multi-technology Cloud-RAN base station: either at the fronthaul interface or located at the Core Network. We integrate intelligence to the MEC enabled framework, by considering the available links through which each client is served. We employ testbed experimentation in order to illustrate the efficiency of our scheme and demonstrate how we achieve efficient allocation of the MEC resources and wireless technologies used for the system under consideration.**

***Index Terms*—MEC, Pricing, Cloud-RAN, 5G, HetNets**

I. Introduction

5G brings several advancements in both the air interface, the integration of legacy technologies for the formation of Heterogeneous Networks, and the utilization of edge resources. Applications developed around this ecosystem are expected to take advantage of high throughput and low latency wireless links, for supporting services around a wide domain of verticals (e.g. eHealth, Industry 4.0, AR/VR, etc.). Nevertheless, advancements in the air interface focus on enhancing the capacity of the network; low latency access is expected to be achieved through the wide application and utilization of edge computing, with resources being migrated to the network edge. In this context, and towards addressing the heterogeneity in the network access domain, ETSI revised the annotation for Mobile Edge Computing towards Multi-access Edge Computing (MEC). Through MEC, UEs in the network may use any of their available wireless network interfaces for accessing services located at the edge of the network.

At the same time, 5G brings advancements in the network architecture and organization of base stations. Through the wide application of the Cloud-RAN concept, parts of the base station can be instantiated as Virtual Network Functions (VNFs) at an edge located datacenter, managing lower complexity units used for transmitting the information in the cell. This feature allows the instantiation of new base stations within an area based on demand, and is an enabler for adding heterogeneous technologies at the user access level, as a means of aggregating different access technologies. As a matter of fact, in the recent specifications for the 5G New Radio (NR) interface, the base stations are disaggregated between the Packet Data Convergence Protocol (PDCP) and the Radio Link Control (RLC) layers, forming a Central Unit (CU) that can be instantiated in the cloud, controlling a Distributed Unit (DU) for forming the wireless cell. One CU may control multiple even heterogeneous DUs, allowing the integration of several technologies to the operator's provided cell, e.g. 5G-NR and LTE or non-3GPP based e.g. WiFi.

Although MEC is expected to play an important role in the overall 5G network operation, the placement of MEC services seems to be inherited from the legacy generations of mobile communications. In [1], ETSI provides information for all possible deployments of MEC services in the network. Nevertheless, even for the disaggregated RAN case, MEC services will be co-located with the CU at an Edge Datacenter. In [2], we provided a first experimental prototype that goes beyond these deployments, and places the provided services on the fronthaul interface of heterogeneous Cloud-RAN infrastructures, introduced in [3]. In such setups, multi-homed network users get access over multiple wireless links to services located just after the DU component of the cellular network, illustrating reduced network latency compared to conventional MEC deployments. Nevertheless, the technology through which each user may be served plays an important role in the overall perceived latency and Quality of Experience (QoE) of the mobile terminal user. Moreover, the locations for placement of the hosted services and wireless technologies used to forward data to the end users can be exploited as differentiation parameters for charging application providers for hosting their services on the MEC platform.

In this work, we extend the work provided in [2] and deploy the MEC functionality at two different tiers of the network: 1) on the fronthaul interface, and 2) collocated with the Core Network. We seek to answer the following key questions:

- How should resources from the MEC enabled network be allocated to different Service Providers?
- How should MEC providers make use of the access technologies available for forwarding MEC data to the UEs?
- How do these choices affect the service-to-UE latency?

We initially introduce a system model for resource allocation across the different MEC tiers. We then map our approaches to the prototype implementation, and perform testbed experiments to illustrate the viability of our approach. The rest of the paper is organized as follows: Section II is presenting a literature overview in the field. Section III presents our system model, and Section IV details our approach for access technology selection. In Section V we showcase our findings, whereas in section VI we conclude the paper and present some future directions.

II. Related Work

As the application of MEC is designed to deliver low latency for service access, it has received a great level of attention in the recent specifications for 5G and in relevant research. In [1], ETSI specifies the different deployments for MEC services starting from the 4G network architecture and its evolution for 5G. This white paper summarizes the different interfaces needed for hosting services over a MEC enabled server, and specifies the deployments as follows: 1) the *bump-in-the-wire* mode, where the service is located just after the base station, intercepting data-plane traffic, and relieving the network from the extra delay added for sending traffic to the Core Network, 2) the case of collocating the MEC services with the Core Network at an Edge datacenter, which has the benefit of handling IP traffic just after the Core Network, and 3) the *local break-out mode* where a part of the Core Network is handling only data plane traffic collocated with the base station, whereas the control plane traffic is sent to a traditional Core Network deployment. The advantage of the third solution is that it blends the benefits of the two prior solutions, but requires increased complexity on the core network implementation. In [4], ETSI specifies all the different interfaces enabling the MEC operation for different components of the network.

Based on the disaggregated model of a base station, according to the Cloud-RAN concept, in [2] we introduced a new deployment for the MEC services; since the base station is disaggregated in two components, we developed and evaluated a prototype illustrating the traffic flow as UE-DU-MEC, instead of UE-DU-CU-MEC that ETSI specifies as the *bump-in-the-wire* method for Cloud-RAN. The implementation is based on the Open Source OpenAirInterface platform [5], and extends our prior contributions for integrating non-3GPP technologies in the cell [3]. Thus, the solution provides a first effort for enabling a MEC platform, with the services being deployed as close as possible to the network edge. The prototype showcased low latency for MEC service access, able to achieve less than 10ms for a standard LTE cell in the access network. This solution is adopted in this paper as well, as the base of our experimental platform.

Similar solutions for deploying the services on the edge exist in such experimental platforms. For example, in [6], the authors use the OpenAirInterface platform in order to deploy services co-located with the Core Network. By using an SDN approach just after the Core Network, the authors provided low-latency times for accessing hosted services for specific UEs. Similarly, in [7], the authors implement the *bump-in-the-wire* method on the same platform. This prototype may achieve low latency times, but as it is solely implemented in application space, it strives to provide real time services for high-load cells. In [8], the authors present all the possible enablers for MEC operation when multiple technologies are used for user access. When considering the existence of multiple paths in the wireless part of the network, allocating the network resources needs to be revisited. For example, in [9], the authors deal with the Radio Access Technology (RAT) association problem for Heterogeneous Networks (HetNets) when MEC resources are present. Similarly, in [10] the authors consider a multi-RAT network with MEC resources, and attempt to minimize the overall energy consumption of the network with a hollistic approch. Application specific MEC enhancements are also presented in [11]. The authors use dynamic adaptive streaming video over a MEC service, extended to ensure the optimal QoE for the end users.

In this work, we initially model a MEC platform in terms of resource allocation. We use a pricing scheme to determine how the resources residing at the MEC platform (CPU, memory and storage) shall be allocated to Service Providers (SPs). Subsequently, we introduce an algorithm for selecting the network access technology used to serve each user of the network, in order to ensure that the overall service access latency times are kept low. Finally, we employ testbed experimentation with the objective to evaluate our framework for different placements of the MEC service: 1) on the fronthaul interface of the multi-technology Cloud-RAN, 2) on the Core Network, and 3) deployed at a remote datacenter.

III. MEC Pricing Scheme

We consider a two-stage MEC pricing scheme, where the MEC owner (operator) decides the price p per unit of MEC bundled resources (CPU, memory and storage) in the first step and in the second step the service/content providers, interested in providing low latency services, decide the level of bundled MEC resources, which they intend to pay as a function of the price and the latency sensitivity of the provided service. We approach the pricing problem using backward induction following the rationale of [12], examining first the service/content providers' demands (Stage II) and then the MEC operator's decision on the price (Stage I). We propose two pricing models, one linear in Section III-A and one exponential in Section III-B.

A. MEC Resource Allocation With Linear Pricing

Stage II: The payoff function of the Service Provider SP_i, $i = 1, ..., N$, for acquiring b_i units of MEC bundled resources with a price p per bundle unit, following the linear pricing model, is expressed as

$$U_i^{lin}(b_i) = ln(1 + \theta_i b_i) - pb_i \quad (1)$$

with θ_i representing the normalized latency sensitivity of SP_i, $\theta_i \in [0, 1]$. This payoff function of SP_i is equal to the logarithmic utility function, that expresses the diminishing return of getting additional resources, minus the linear price

that SP$_i$ has to pay for acquiring b_i quantity of MEC resources. We notice that $U_i^{lin}(b_i)$ is a concave function, since $U(b_i)'' = -(\theta_i/(1+\theta_i b_i))^2 < 0$. Thus, it has only one maximum, and therefore the local maximum is also the global maximum. Differentiating (1) we have

$$\frac{\partial U_i^{lin}}{\partial b_i} = \frac{\theta_i}{1+\theta_i b_i} - p = 0 \tag{2}$$

The optimal value of MEC resources that maximizes SP$_i$'s payoff is

$$b_i^* = \begin{cases} \frac{1}{p} - \frac{1}{\theta_i}, & \text{if } p \leq \theta_i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Stage I: We assume that the N SPs that are requesting for MEC resources present similar latency sensitivity. Otherwise, no need to purchase MEC resources would exist. Thus, we assume that their latency requirements are such that $\max(\theta_i) - \min(\theta_i) < \varepsilon$, where $\varepsilon > 0$. Under this assumption, the MEC owner's choice of price p is such, that the SP with the $\max(\theta_i)$ is allocated the maximum value of MEC resources $b_{\max}$, aiming to provide the best available service to SPs with higher latency sensitivity compared to the rest of the SPs requesting resources from the MEC agent. We also assume that the MEC agent has adequate available resources to satisfy the requests of all SPs under consideration. The price is formed according to (4).

$$p = \frac{\max(\theta_i)}{1+\max(\theta_i) b_{\max}} \tag{4}$$

The provider aims to give to every SP$_i$ the opportunity to have access to the MEC resources. This means that even for the SP with the $\min(\theta_i)$, the quantity $1/p - 1/\min(\theta_i)$ is positive. Using (4) we find the range of values of ε under which this MEC resource allocation is feasible. This range is expressed as

$$0 < \varepsilon \leq \max(\theta_i) \min(\theta_i) b_{\max} \tag{5}$$

The allocated level of MEC resources to each SP$_i$ following the linear pricing model is expressed as

$$b_i = \frac{1+\max(\theta_i) b_{\max}}{\max(\theta_i)} - \frac{1}{\theta_i} \tag{6}$$

B. MEC Resource Allocation With Exponential Pricing

For the MEC resource allocation with exponential pricing, we follow the same steps as described in the linear pricing approach.

Stage II: The payoff function of SP$_i$ under the exponential pricing model, for acquiring b_i units of MEC bundled resources is expressed as

$$U_i^{exp}(b_i) = ln(1+\theta_i b_i) - p_e(e^{b_i} - 1) \tag{7}$$

We notice that $U_i^{exp}(b_i)$ is a concave function, since $U_i^{exp}(b_i)'' = -(\theta_i/(1+\theta_i b_i))^2 - p_e e^{b_i} < 0$. Thus, it has only one maximum, and therefore the local maximum is also the global maximum. Differentiating (7) we have

$$\frac{\partial U_i^{exp}}{\partial b_i} = \frac{\theta_i}{1+\theta_i b_i} - p_e e^{b_i} = 0 \tag{8}$$

We express (8) as

$$ln\left(\frac{1}{p_e}\right) + \frac{1}{\theta_i} = \left(b_i + \frac{1}{\theta_i}\right) + ln\left(b_i + \frac{1}{\theta_i}\right) \tag{9}$$

For $x = b_i + \frac{1}{\theta_i}$ and $y = ln\left(\frac{1}{p_e}\right) + \frac{1}{\theta_i}$, (9) can be written as

$$y = x + lnx \tag{10}$$

which can be also expressed as

$$xe^x = e^y \tag{11}$$

Taking the value of the Lambert W function [13] of each part of (11) and using the Lambert W function identity $W(xe^x) = x$, we have $x = W(e^y)$. Replacing x and y we have

$$b_i^* = \begin{cases} W\left(\frac{e^{\frac{1}{\theta_i}}}{p_e}\right) - \frac{1}{\theta_i}, & \text{if } \theta_i \geq \frac{1}{W\left(\frac{e^{\frac{1}{\theta_i}}}{p_e}\right)} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Stage I: The price p_e that the MEC owner decides in the exponential pricing model is such, that SP$_i$ with $\max(\theta_i)$ is allocated the maximum value of MEC resources $b_{\max}$. The price is formed according to (13).

$$p_e = \frac{\max(\theta_i)}{(1+\max(\theta_i) b_{\max})\, e^{b_{\max}}} \tag{13}$$

As the provider aims to give to all N SPs the opportunity to to have access to the MEC resources, the level of resources that will be allocated to the user with the $\min(\theta_i)$ should also be positive. This means that the range of latency sensitivity of the N SPs is such, that

$$W\left(\frac{e^{\frac{1}{\min(\theta_i)}}}{p_e}\right) - \frac{1}{\min(\theta_i)} > 0 \tag{14}$$

The allocated resources to each SP$_i$ following the exponential pricing model is expressed as

$$b_i = W\left(\frac{(1+\max(\theta_i) b_{\max}) e^{b_{\max} + \frac{1}{\theta_i}}}{\max(\theta_i)}\right) - \frac{1}{\theta_i} \tag{15}$$

IV. System Architecture and RAT Selection

A. System Architecture

As our starting system architecture, we use a disaggregated multi-technology Cloud-RAN base station, extensively described in [3]. In such a setup, we distinguish the following components and roles:

- The **CU**, which is running the higher layer 2 functions of the base station (PDCP layer and upwards), and provides the interface to the Core Network.
- The **3GPP DU**, running the lower layer 2 functions of the base station (RLC and below), and performs the transmission of the traffic over the air. The DU may support heterogeneous technologies (e.g. 5G NR or LTE) and uses the Radio Network Temporary Identifiers (RNTI) for addressing each client.
- The **non-3GPP DU**, which is running the layer 2 functions for non 3GPP technologies (e.g. WiFi) and communicates

with the CU for sending/receiving traffic to/from the wireless network. As it communicates directly with the PDCP layer of the CU, it handles and encapsulates the data in the appropriate format. MAC addresses are utilized for addressing each client.

- The **Core Network**, which is the entry and exit point for user plane data to the base station network.
- The **MEC agent**, which is enabling the data exchange from services located at the Fronthaul interface with the DUs directly, without using the CU as intermediary node.
- The **Technology Selection** modules, which reside on the MEC agent and the CU, and are able to select the forwarding DU(s) for each client of the network.
- The **Hosted Services** over this heterogeneous infrastructure, which are containerized services running on top of the MEC agent, the Core Network or any other remote datacenter. The container technology we use is LXC.

Fig. 1 shows how these components have been mapped to a real testbed setup. We employ the NITOS testbed [14], which provides all the required experimental components for supporting our experimentation.

Algorithm 1 MEC side selection of RATs for each UE.

```
Calculate the resource allocation of the services
based on the chosen pricing (Linear or Exponential)
while 1 do
    for each service request do
        if Both DU meet the service's requirements then
            Choose the DU with the lowest Latency
            if DU capacity is lower than 50% then
                Use both DU with percentage q and
                (100-q) of the time respectively
            end if
        else
            Choose the DU which meets the requirements
        end if
        Send to the proper REST API:
        the UE id and DU/s id/s
    end for
    Calculate the resource allocation of the services
    based on the chosen pricing (Linear or Exponential)
end while
```

B. Selection of Radio Access Technology

As MEC considers multiple wireless technologies for service access, network selection is an issue of paramount importance. The last-hop link used for serving each user may exemplify different access times, based on factors such as the cell coverage, the location of the UE in the cell, the allocated modulation and coding scheme, the load of the cell and external interference, especially for non-3GPP technologies such as WiFi. In this section, we introduce an algorithm for selecting the last-hop wireless connection in a per client basis for each UE. We assume at this point that each UE is multi-homed and is using all of its available technologies to communicate with the MEC and Core Network.

Although our scope is to minimize service latency times, we bear in mind the different capacities of the wireless networks used to serve each UE. Therefore, each UE might be concurrently served by combinations of the available technologies, while the per-packet traffic latency is kept below a threshold limit. In the case that the capacity of a technology is about to be reached, the algorithm might choose to serve an end-user by another technology. Algorithm 1 shows how the MEC part of the network makes these selections.

In order for the decisions for each forwarding DU to be applied, separate controllers have been developed at two different points: 1) on the MEC agent, which is handling the MEC traffic on the fronthaul interface, and 2) on the CU side of the network, that handles the traffic before being sent to each DU. Both the controllers operate under the same principle. They expose a REST API that gets as inputs the identifier for each UE of the network and the DU or combination of DUs that will be used for forwarding the traffic. For the case of combination of DUs, defining the percentage of traffic for each DU is also supported. Based on this, our algorithm operates as follows. We initially calculate the resource allocation for the service providers based on the pricing model. For each new UE service request, the controllers residing at the MEC agent or the CU select the forwarding DU that meets the specific UE requirements. If all DUs are able to serve this UE then the DU with the lower value of latency is selected. If the capacity of the DU with lower access latency is not exceeding 50%, the service request is served through this DU. In the case that this threshold is exceeded, we split the traffic over multiple DUs with 50% transferred over the WiFi DU and 50% over the LTE. The information is subsequently sent to the respective controllers, managing the forwarding of the data to the UEs.

V. System Evaluation

In this section, we present our experimental findings. First, we present the system components and selected configurations and then we showcase results of the experiments. We employ the NITOS testbed for extracting the latency values for the service placement for two MEC setups (service on the Fronthaul-FH or on the Core Network-EPC) and the Internet.

TABLE I: System Latency times

Service Location	WiFi	LTE
Fronthaul (FH)	2.39	9.85
Core Network (EPC)	2.63	16.16
emulated Internet	12.57	25.9

We employ two different wireless technologies for accessing the MEC services, either LTE for 3GPP access or WiFi for non-3GPP access, both with the same configuration: 2x2 MIMO and 20MHz channel bandwidth. Our testbed setup can accommodate multiple DUs, including 5G-NR, but as the OAI 5G-NR platform is currently in development state we omit this wireless technology from our evaluation. The overall setup is shown in Fig. 1 and Table I presents the achieved access latency for the different service placements for all the available RATs. We also consider 6 different types of services based on their requirements in terms of latency and throughput derived from [15].

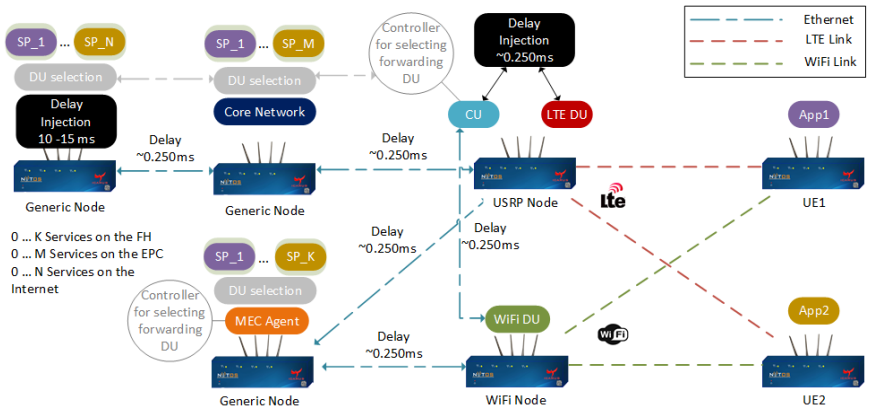


Fig. 1: Experimental topology for evaluating our scheme; controllers residing at the CU and the MEC agent part select the forwarding DU(s) for serving each UE in a per-packet basis. Services are placed either on the Fronthaul (FH), Core Network (EPC) or emulated Internet.

We create a matching between the types of service and θ_i values, as shown in Table II and for each service provider, we select a value uniformly from the provided ranges per each SP. Table III shows the MEC placements and RAT allocation for each of the different services that we use for both pricing models. As we can see, the exponential pricing is more flexible and can serve all the services through MEC, whereas the linear pricing cuts-off the service of lower latency sensitivity from the MEC. This occurs due to the fact that in the linear model, the range of values that renders the MEC resource allocation feasible in (5) is more strict than the respective (14) of the exponential model, for the same $b_{\max}$ and θ_i.

TABLE II: Applications Normalized Latency Sensitivity

APPLICATION TYPE	Initial Selected θ_i	θ_i range
AR/VR	0.95	[0.85 - 1]
V2X	0.8	[0.7 - 0.85)
VIDEO STREAM	0.65	[0.55 - 0.7)
VoIP	0.5	[0.3 - 0.55)
BROWSING	0.2	[0.1 - 0.3)
MAIL	0.05	[0.0 - 0.1)

We evaluate our proposed scheme with two different use cases. Each use case is examined for the two provided pricing models under the same scenario. We measure the allocated resource bundles for each SP, for each new SP that enters the system, and the aggregate latency for the network UEs accessing the provided services. We present the average performance of our performed experiments repeated 100 times, along with the standard deviation for each measurement. The two use cases are differentiated regarding the level of resources that each new SP demands. In the first use case, services with low demands are introduced such as VoIP, web services or e-mail servers, whereas in the second, services with high demands are introduced, such as AR/VR, V2X and video streaming. For both cases, we plot the resource allocation and delay after the initial placement of six different SPs, following the application types of Table II.

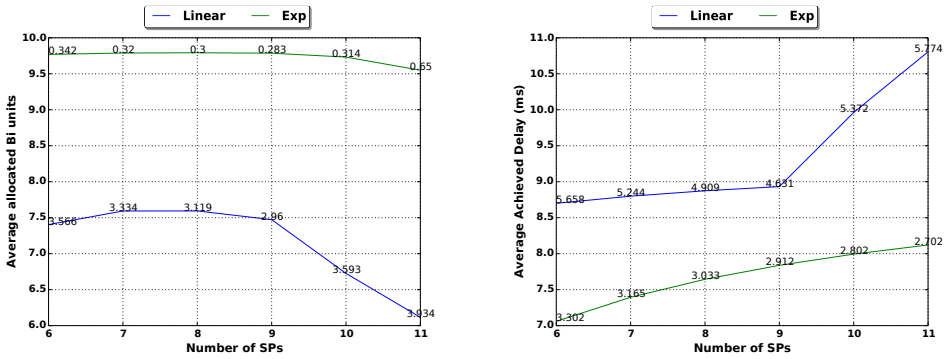
Fig. 2a shows the average allocation of bundled MEC resource units b_i for each pricing model as the number of SPs increases along with the standard deviation. We observe that with linear pricing, the allocation of b_i units depends highly on the SPs' requirements, in contrast to the exponential pricing where all SPs are assigned with almost equal and higher level resource bundles. This happens mainly because with exponential pricing, MEC resources are more evenly spread over the SPs requesting to place their services on the MEC (FH or EPC). In Fig. 2b, we observe a similar trend for the two pricing models for the average achieved delay of the system. The exponential pricing achieves lower average latency, as it more evenly places services to the MEC, presenting lower levels of average delay. In the second use case, the average latency per each SP is lower than the linear policy. Comparing the delay of the linear pricing policy with the first use case experiment, we observe higher latency times. This is happening because the types of SPs used for this experiment pose higher demands regarding their latency requirements. Due to this fact, the assignment of multiple RATs per service is employed only for the SPs with low demands, contrary to the first use case where most of the services use multiple RATs. The rest of the services are assigned to one RAT only (LTE), once the capacity of the RAT with the minimum delay (WiFi) is fully allocated. Based on the findings in Table I, the difference in latency between the WiFi and LTE is high; this is also reflected in the average achieved delay of the system in Fig. 3b.

TABLE III: SPs allocation

SP ID	Linear	Exp	RAT
AR/VR	FH	FH	WiFi
V2X	FH	FH	WiFi
VIDEO STREAM	EPC	EPC	Both
VoIP	EPC	EPC	Both
BROWSING	EPC	EPC	Both
MAIL	Internet	EPC	Both

VI. Conclusion

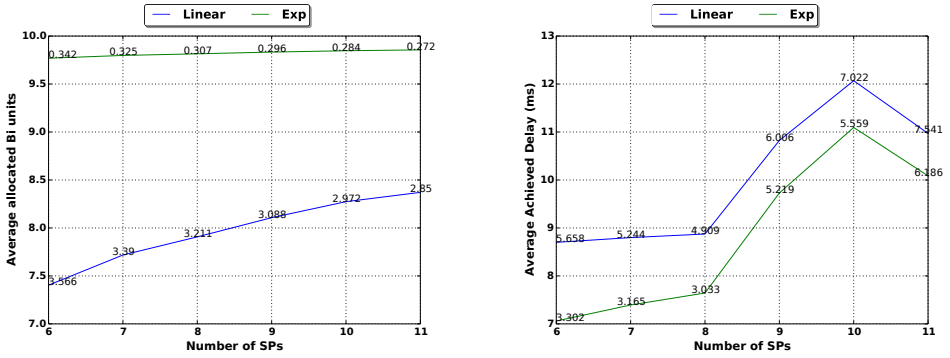
In this work, we presented a scheme for the resource allocation of MEC resources to different service providers using two pricing policies. We modeled our approach and used a testbed setup to evaluate our scheme, using two different placements of the services: on the fronthaul interface of Cloud-RAN base stations, or collocated with the Core Network. Through the integration of multiple technologies at the base station level, we are able to achieve differentiation for the



(a) Average assigned b_i units per SP

(b) Average achieved delay per SP

Fig. 2: Experimental results for the 1st scenario of SPs allocated to the system (low demand)



(a) Average assigned b_i units per SP

(b) Average achieved delay per SP

Fig. 3: Experimental results for the 2nd scenario of SPs allocated to the system (high demand)

latency access times for each service per each network UE. Our experiments denote that through our approach, MEC resources can be allocated while the average latency per each SP can be kept below a threshold, by utilizing multiple links at the same time. In the future, we foresee extending our scheme towards modeling the access of each UE in the network from the SP's perspective, even for the cases of UEs accessing the same service but under different agreements with the operator. Moreover, we plan to integrate migration of the SPs in the system, based on the mobility patterns detected for each UE.

ACKNOWLEDGMENT

The research leading to these results has received funding by the EU H2020 Programme for research, technological development and demonstration under Grant Agreement Numbers 762057 (5G-PICTURE) and 857201 (5G-VICTORI) and by GSRT, under the action of "HELIX-National Infrastructures for Research", MIS No 5002781.

REFERENCES

[1] F. Giust *et al.*, "ETSI White Paper No. 24: MEC Deployments in 4G and Evolution Towards 5G," 2018.

[2] N. Makris, V. Passas, T. Korakis, and L. Tassiulas, "Employing MEC in the Cloud-RAN: An Experimental Analysis," in *Proceedings of the 2018 on Technologies for the Wireless Edge Workshop*. ACM, 2018.

[3] N. Makris, C. Zarafetas, P. Basaras, T. Korakis, N. Nikaein, and L. Tassiulas, "Cloud-based Convergence of Heterogeneous RANs in 5G Disaggregated Architectures," in *IEEE International Conference on Communications (ICC)*, 2018.

[4] S. Kekki *et al.*, "ETSI White Paper No. 28: MEC in 5G networks," 2018.

[5] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface: A flexible platform for 5G research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33–38, 2014.

[6] A. Huang, N. Nikaein, T. Stenbock, A. Ksentini, and C. Bonnet, "Low latency MEC framework for SDN-based LTE/LTE-A networks," in *Communications (ICC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1–6.

[7] C.-Y. Li, H.-Y. Liu, P.-H. Huang, H.-T. Chien, G.-H. Tu, P.-Y. Hong, and Y.-D. Lin, "Mobile Edge Computing Platform Deployment in 4G LTE Networks: A Middlebox Approach," in {*USENIX*} *Workshop on Hot Topics in Edge Computing (HotEdge 18)*, 2018.

[8] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, 2017.

[9] M. Emara, M. C. Filippou, and D. Sabella, "MEC-aware cell association for 5G eterogeneous etworks," in *Wireless Communications and Networking Conference Workshops (WCNCW), 2018 IEEE*, 2018.

[10] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Energy-efficient offloading for Mobile Edge Computing in 5G Heterogeneous Networks," *IEEE Access*, 2016.

[11] C. Ge, N. Wang, S. Skillman, G. Foster, and Y. Cao, "QoE-driven DASH video caching and adaptation at 5G mobile edge," in *Proceedings of the 3rd ACM Conference on Information-Centric Networking*. ACM, 2016.

[12] V. Miliotis, L. Alonso, and C. Verikoukis, "Weighted proportional fairness and pricing based resource allocation for uplink offloading using ip flow mobility," *Ad Hoc Networks*, vol. 49, pp. 17–28, 2016.

[13] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W Function," *Advances in Computational Mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[14] "NITOS - Network Implementation Testbed using Open Source platforms." [Online], https://nitlab.inf.uth.gr/NITlab/.

[15] A. Reznik *et al.*, "ETSI White Paper No. 23: Cloud RAN and MEC: A Perfect Pairing," 2018.