Joint Management of Compute and Radio Resources in Mobile Edge Computing: a Market Equilibrium Approach

Eugenio Moro, *Student Member, IEEE,* and Ilario Filippini, *Senior Member, IEEE.*

Abstract—Edge computing has been recently introduced to bring computational capabilities closer to end-users of modern network-based services, supporting existing and future delay-sensitive applications by effectively addressing the high propagation delay issue that affects cloud computing. However, the problem of efficiently and fairly managing the system resources presents particular challenges due to the limited capacity of both edge nodes and wireless access networks and the heterogeneity of resources and services' requirements. To this end, we propose a techno-economic market where service providers act as buyers, securing both radio and computing resources to execute their associated end-users' jobs while being constrained by a budget limit. We design an allocation mechanism that employs convex programming to find the unique market equilibrium point that maximizes fairness while ensuring that all buyers receive their preferred resource bundle. Additionally, we derive theoretical properties that confirm how the market equilibrium approach strikes a balance between fairness and efficiency. We also propose alternative allocation mechanisms and give a comparison with the market-based mechanism. Finally, we conduct simulations to numerically analyze and compare the performance of the mechanisms and confirm the market model's theoretical properties.

Index Terms—Network Slicing, Mobile Edge Computing, Resource Allocation, Market Model, Game Theory

1 Introduction

The recent emergence and affirmation of Cloud Computing (CC) has determined the centralization of compute power into large data centres. As this trend is expected to continue in the future, CC will play an increasingly prominent role in web-based services by providing supercomputing-like capabilities. However, the often long distances between data centres and end users render CC unsuitable for supporting new time-sensitive use cases, such as those enabled by the fifth generation of mobile radio networks (5G) [1].

Recently introduced Mobile Edge Computing (MEC) promises to surpass this intrinsic limit of CC by shifting the computational power toward the edge of the network, in the proximity of the Radio Access Network (RAN). This way, the resulting reduction in propagation delays promotes a wide range of time-sensitive applications, such as tactile internet, augmented reality and IoT. [2]. Additionally, MEC can alleviate the burden that future data-intensive applications, such as autonomous driving and massive IoT, might have on the networking infrastructure, as local computations avoid sending large data volumes to the cloud [3].

MEC is currently undergoing a developing phase, and, among the different open challenges it poses to the research community, resource allocation and management represent a particularly interesting one. Unlike what is usual in resource management studies for CC, assuming infinite network capacity between the computing nodes and the offloading source may be questionable. Indeed, Edge Nodes (EN) are deployed close to wireless access points with limited capacity. Furthermore, given the relatively high density of these deployments, ENs offer limited computing capabilities to keep costs down and retain the economic scalability of MEC. Therefore, we argue that an effective management framework cannot refrain from jointly considering both radio and computational resources.

The emergent concept of network slicing is well suited to offer the required bundles of heterogeneous virtualized resources. However, slice capabilities need to be customized according to the service that the slice owner intends. Different slices might indeed need varying amounts of computational and communication resources. Additionally, the underlying shared infrastructure presents a high degree of heterogeneity. Not only the availability and utility of resources are dramatically different between the MEC and the RAN domain, but also edge nodes might present different installed capacities within the single MEC cluster. This substantial diversity calls for joint optimization of both domain's resources, as uncoupling the two might lead to situations in which resources in one domain are overprovisioned with respect to the capacity of the other. For instance, if a slice was to receive a large radio capacity without enough computation resources, most of the jobs uploaded through the RAN would encounter an unacceptable processing delay. At the same time, the overprovisioned radio resources could be utilized more efficiently by other slices with less stringent processing requirements.

The scenario previously described requires smart resource allocation solutions. Static resource partitioning approaches, such as the one considered by 3GPP [4], might lead to poor results. Indeed, the intrinsic spatial and tempo-

- *E.Moro and I.Filippini are with the Department of Electronics, Information and Bioengineering, Politecnico Di Milano, 20133 Milan, Italy. E-mail: {eugenio.moro, ilario.filippini}@polimi.it*

*Manuscript received ****; revised ****.*

ral heterogeneity of both demands and resource availability suggests that elastic allocations might yield higher efficiency and better performance. Furthermore, network slices offer customized network services to tenants representing business entities, whose economic performance indicators are likely to couple with the experienced service performance. The consequence is that tenants should be allowed to directly participate to the allocation decision process and autonomously manage their resources to better match their desired techno-economic performance. Indeed, third-party allocation decisions that penalize some tenants to favor the overall welfare would result in a weak and likely unacceptable solution from an economic point of view.

In light of what mentioned above, the following question arises: *given an MEC/RAN deployment consisting of a set of heterogeneous network cells and ENs, how to efficiently allocate computing and network resources to competing network slices with different requirements while guaranteeing fairness, service prioritization, and economic convenience?*

Our proposed answer consists of a resource management model based on a market of resources, where each service provider is allowed to buy allocations constrained to a budget. The properties of the solution are analyzed through Game Theory and General Equilibrium Theory.

The considered market is based on an instance of Fisher Market [5]. This well-known model has been extensively used in the literature for either RAN resource allocation in slicing contexts or MEC resource management, individually considered. However, to the best of the authors' knowledge, it has never be introduced when the two domains are combined, as we do in this article. Indeed, it is a non-trivial extension toward a joint resource management that needs careful consideration of how each type of resource's demand and availability impacts the overall system performance. That is exactly the task we complete in this work.

In our proposed market model, domains' resources are dynamically priced according to their availability and instantaneous demand values. At the same time, each buyer obtains the particular bundle that maximizes its private utility function, coinciding with the number of successfully executed jobs at the market equilibrium. Finally, service prioritization is enforced through budget differentiation, whose values can represent real money or, more generally, power relationships among service providers.

The remainder of this article is structured as follows. Section 2 presents some relevant related works and point out the novel contribution of this work with respect to the others in the literature. Section 3 presents the system model formulation, while Section 4 presents the proposed market-based resource allocation approaches, whose properties are theoretically analyzed in Section 5. Finally, simulation results are shown in Section 6. The article concludes with Section 7 that includes final remarks.

2 Related Works

The main results of this work are based on Game Theory, which has been extensively utilized in network resource management [6]. In the context of game theory-based network resources management for slicing, authors in [7] propose a single-cell budget-free market with prices computed according to some function of the cell load and in [8] an auction model for the resource allocation problem is proposed. Authors of [9] propose a budget-constrained market model (extended to admission control in [10]) that, only from the radio resource management point of view, faces the problem following a rationale similar to the one of our proposal.

In the multi-server computing resources management literature, the multi-resource allocation problem has been first addressed in [11], where the well known Dominant Resource Fairness concept [12] has been extended to multiple heterogeneous servers scenarios. An external resource, generically identified as bandwidth, was later introduced in [13] as an extension to the previous approach. More refined approaches can be found in [14], where the problem of the joint resource optimization to provide the desired QoS to all mobile users and traffic types is formulated as a mixed-integer nonlinear program. In [15] the optimization of radio resources and computing power is aimed at minimizing users' energy consumption while meeting latency constraints. Authors of [16] propose a model for optimizing both the server placement and the task offloading of a MEC system comprising multiple access points. A channel and queue aware cross-layer scheduler is proposed in [17], where radio and computing resources of a single access point are jointly optimized.

Both [18] and [19] approach the computing resource allocation problem employing game-theoretical tools, proposing a Fisher Market-based model for the management of MEC resources. However, the former work studies the system only from a high-level perspective at the EC platform (i.e. not considering the radio access segment). The latter avoids managing radio resources, claiming transmission delay to be constant and negligible. We believe this assumption is highly questionable in the next years, as future services will challenge the radio access segment of the network, either by large job payloads (i.e. video transcoding, image processing), large numbers of simultaneous requests (i.e. massive-IoT) or a combination of both (i.e. autonomous driving and connected cars). Thus, network resource management becomes relevant [20].

Authors of [21] have proposed a market mechanism where the MEC infrastructure provider sets the prices of the CPU cycles of a single MEC server to maximize its revenue. After prices are set, users select which jobs to offload to minimize a combination of cost and latency. The model considers a single base station, and its radio resources are not optimized, but the available spectrum is equally divided among the users instead. In [22], a game where MEC users compete for the radio resources and CPU cycles of a single access point is proposed. In particular, users can choose either to locally process their tasks or offload them to the AP. Authors of [23] have proposed a mechanism for the management of CPU and radio resources of a single access point with the aim of minimizing the long term energy consumption of both the server and the mobile terminals. A detailed sharing business model is presented in [24]. Here users of a MEC server cluster can rent unutilized CPU-share quotas in exchange for additional income. The infrastructure providers are part of the model and sell subscription plans to end-users. Scenarios where the platform aims at maximiz-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2021.3091764, IEEE Transactions on Mobile Computing

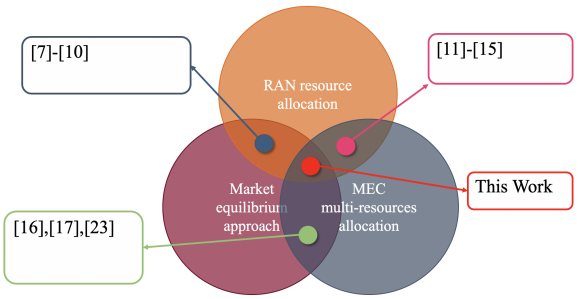


Fig. 1: Literature study areas and their intersections

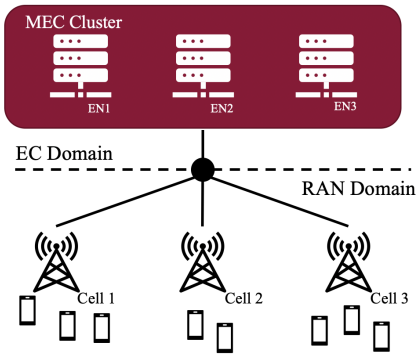


Fig. 2: High level system architecture.

ing either its profit or the social welfare are analyzed and compared. Liu *et al.* [25] propose a microeconomic model where end users can buy radio resource blocks from access points and processing power from edge nodes to maximize task performances while containing the costs. Both an optimal budget allocation algorithm and an equilibrium price finding algorithm are proposed.

2.1 Motivation and contributions

In the previous analysis, three main areas of study can be identified: radio resources allocation for either a single access point or a RAN, single or multiple computing resource allocation and market equilibrium-based allocation, as pictured in Figure 1. While the related works we presented in this section cover up to two of the areas above, we were not able to find any work that satisfyingly covers all three areas. Even those works that propose a market model for both radio and computing resources consider only a single access point, a single edge node or only CPU resources. Furthermore, none of them allocates resources to network slices, but considers the single end-user instead. Thus, we claim this paper to be the *first* to cover *the problem of multi-resource, slice-aware allocation for multiple MEC servers and multiple wireless access points through a techno-economic market of resources.*

The main contribution of this work can be further summarized as follows:

- We formulate a system model for the environment above that comprises multiple servers in a MEC cluster and multiple radio cells of a RAN. MEC service providers are modelled as network slices.
- We propose a novel market-based allocation mechanism and resource pricing leveraging resource bottlenecks in the heterogeneous environment above. Additionally, we provide a convex program formulation to find an optimal market equilibrium.
- We formally analyze the theoretical properties of the proposed mechanisms, focusing on the efficiency and fairness of the ME approach.
- We give an extensive numerical evaluation of the proposed mechanism through simulations.

3 System Model

Consider a MEC scenario where several ENs (i.e., servers) constitute a MEC cluster. Let $\mathcal{M}$ be the set of such nodes in the MEC cluster and $\mathcal{R}$ the set resource types offered by the nodes, i.e., CPU, RAM and storage. By $D^{\text{MEC}}_{m,r}$ we express the available capacity of resource type $r \in \mathcal{R}$ in node $m \in \mathcal{M}$. Since we consider a heterogeneous MEC cluster, capacities $D^{\text{MEC}}_{m,r}$ can in principle be different for each resource and node in order to model real-life scenarios in which ENs have different installed capacity.

Consider a Radio Access Network that users of the MEC system can employ to offload their jobs to the ENs. Let $\mathcal{C}$ be the set of C cells that can be used to access the MEC cluster. We then define the available capacity[1] of each cell $c \in \mathcal{C}$ as D^{RAN}_{c}. For the sake of simplicity, we consider resources to be expressed in terms of cell spectrum portions. We suppose the presence of a slicing-aware MAC scheduler capable of dynamically enforcing these constraints, such as the one proposed in [26].

While the RAN access capacity is limited, we do not impose any limitation on the backhaul network capacity that connects each cell with any server in the MEC cluster. Furthermore, once the job payload enters the cluster, it can be freely scheduled for execution in any node of the cluster.

An example of the system architecture that we have just described can be found in Figure 2. Here, we have highlighted the boundary between the two domains we are considering: the EC Domain, which contains the edge nodes and their resources and the RAN domain, containing the cells composing the access network.

3.1 Demand-based Service Characterization

We consider Service Providers (SPs) to own virtualized bundles of both computing and radio resources in the form of network slices. Network slice and service provider have interchangeable definitions in this text. SPs employ these resources to provide end-users with specific MEC services.

For instance, a specific SP might offer neural network training offloading services. In this case, end-users would need to upload training data to the edge nodes, which would in turn train a neural network - we call this combination of upload and processing a job. Alternatively, some services like voice recognition or video processing might require a constant stream of data to be sent by end-users to

1. This generic definition of network capacity can be adapted to a series of different radio access technologies. For instance, D^{RAN}_{c} can be expressed in terms of available physical resource blocks for 4G and 5G networks, time slots for TDM systems and available spectrum for FDM systems.

ENs for processing. However, such streams are often composed of individual packets and uploading and processing each one might be considered a job.

Independently of its scope, a specific service can be generally characterized in terms of computing and networking resources needed to complete associated jobs. Consequently, jobs originating from end-users subscribed to the same SP are likely to present similar resource requirements. Hence, we define a set of demand requirements characterizing each service provider, calling it a demand profile. It is worth noting that, given the wide variety of MEC applications, services can present diverse demand profiles: there might be a CPU-intensive service whose CPU demand could be relatively higher than its memory requirement, or a network-intensive type of service whose job's payload is larger than others' payloads, thus requiring higher bandwidth allocation.

Let $\mathcal{S}$ be the set of service providers. We define $d_{s,r}^{\text{MEC}}$ as the minimum amount of MEC resource of type $r \in \mathcal{R}$ needed to complete one job of service provider $s \in \mathcal{S}$ in a timely fashion. Similarly, $d_{s,c}^{\text{RAN}}$ indicates the minimum amount of network resources needed to successfully upload one job of service provider $s \in \mathcal{S}$ through cell $c \in \mathcal{C}$. While $d_{s,r}^{\text{MEC}}$ does not depend on the specific EN where the job is executed, $d_{s,c}^{\text{RAN}}$ must depend on the cell where the payload is uploaded. This is because different channel quality might require higher or lower bandwidth requirements for the same payload size.

In particular, several factors that might positively or negatively influence the channel quality can be accounted for by properly tuning $d_{s,c}^{\text{RAN}}$. For instance, $d_{s,c}^{\text{RAN}}$ could be increased in a specific cell to counteract the reduced spectral efficiency due to localized inter-cell interference or harsh end-user propagation conditions. Furthermore, if end-users of a given service provider s experience great differences in channel quality within a single cell c, it is always possible to divide end-users subscribed to a single service into n_G groups of end-users with a quasi-homogeneous channel quality. Each group subscribes to a fictitious sub-service s_i (with $i = 1, .., n_G$) that differs from the original only in the network resource requirements, $d_{s_i,c}^{\text{RAN}}$.

Finally, all the factors concurring into $d_{s,c}^{\text{RAN}}$ are taken as time-averaged, such that the market allocation mechanisms can operate on a stable and reliable representation of the RAN. If network conditions were to change or a new market trade were to happen, $d_{s,c}^{\text{RAN}}$ could be updated, and the final allocation recomputed.

We consider all jobs processed by the system to show a certain delay-sensitiveness since delay-sensitive services are often considered as the main application of MEC [27]. Service providers can express the delay requirements of their end-user jobs by a proper selection of $d_{s,r}^{\text{MEC}}$ and $d_{s,c}^{\text{RAN}}$. Indeed, the completion time of a computing task is strictly related to the quantity of reserved resources. For instance, a higher CPU-time share leads to a faster computation, as well as a more significant RAM allocation leads to fewer disk accesses and thus an execution speed up. In this work, we consider $d_{s,r}^{\text{MEC}}$ as given, supplied by service providers. We suppose that, if $d_{s,r}^{\text{MEC}}$ or more resources are allocated for each resource type, then one job can be considered as successfully executed in the due time. Moreover, we allow for numerous jobs to be executed simultaneously if multiples of this quantity are allocated.

This is consistent with realistic scenarios in which SP-dedicated virtual machines (VMs) with different reserved resources are deployed in edge nodes. In this case, a VM with reserved capacity $d_{s,r}^{\text{MEC}}$ would be powerful enough to execute only one job of SP s in a timely fashion. Similar considerations can be done in the RAN domain when defining the amount of required network resources d_s^{RAN}. Specifically, both the transmission time (i.e., payload upload time for any given allocation) and the technological delay (i.e., as low as $1ms$ for 5G [28]) can be taken into account when defining the smallest amount of network resources needed for a successful job payload transmission in due time. Same as before, we consider d_s^{RAN} as given, and we allow the number of timely uploaded payloads to scale with larger allocations. This model is again consistent with realistic scenarios where spectrum portions of each cell in a RAN are reserved to specific network slices.

3.2 Service Utility Function

In multi-resource allocation scenarios, one of the aspects that must not be overlooked is the presence of resource bottlenecks and the emergence of a dominant resource [12]. In our case, a specific service provider can experience a resource bottleneck in a particular edge node of the cluster when its associated demand profile is such that there exists a resource type demand that exhausts the availability in the said node. We call this resource *dominant* and when the node capacity for such resource is depleted, then no more jobs can be scheduled, even if other resource types are available in the EN. On the other hand, there is no risk of experiencing RAN domain bottlenecks since the allocable resource is of one type only. However, given the time-sensitivity of jobs, we do not allow for queuing at the domains interface, implying that the least performing domain limits the total system throughput. This can be considered as a bottleneck as well, since no additional jobs can be accommodated in the system once the capacity of one domain reaches its saturation point. In our proposed model, we mathematically capture the presence of these resource bottlenecks in the system by a proper definition of the service provider utility function.

In realistic virtualization scenarios, network and cloud operators rent pure capacity to service providers in terms of job processing resources and are agnostic about how they will use such a capacity to serve their end-users. Conversely, the system performance must be evaluated exclusively with the knowledge of the parameters made available by network and cloud operators. For this reason, we have defined a service utility function that is transparent to the queuing policy and admission control techniques employed by the SP.

We achieve this by noting that, independently of the choices above, only a limited number of jobs can be concurrently accommodated by the system at any point in time if the delay requirements are to be met. We argue that this number is a valid overall performance indicator. It gives the SP an unequivocal measure of how many end-users can be concurrently served by the system for a given resource bundle. We mathematically express this limit as follows.

TABLE 1: Notations

Sets and Parameters	
Notation	**Meaning**
$\mathcal{S}, S$	SP set and set cardinality
$\mathcal{M}, M$	EN set and cardinality
$\mathcal{R}, R$	Computing resource type set and cardinality
$\mathcal{C}, C$	Cell set and cardinality
$D_{m,r}^{\text{MEC}}$	Resource r capacity in EN m
D_{c}^{RAN}	Network resource capacity in cell c
$d_{s,r}^{\text{MEC}}$	Resource r needed for SP s job completion
$d_{s,c}^{\text{RAN}}$	Resource needed in cell c for SP s payload upload
$\mathbf{X}_s$	Computing resource allocation matrix for SP s
$\mathbf{Y}_s$	Network resource allocation vector for SP s
$J_s^{\text{MEC}}(\mathbf{X}_s)$	Concurrent jobs in MEC domain given allocation $\mathbf{X}_s$
$J_s^{\text{RAN}}(\mathbf{Y}_s)$	Simultaneously uploaded jobs given allocation $\mathbf{Y}_s$
$u_s(\mathbf{X}_s, \mathbf{Y}_s)$	Service provider utility
B_s	Budget of SP s
Decision Variables and Prices	
Notation	**Meaning**
$x_{s,m,r}$	Resource r allocated to SP s in EN m
$y_{s,c}$	Network resource allocated to SP s in cell c
$j_{s,m}$	Concurrent jobs of SP s executed in EN m
u_s	Linearized value of SP s utility acc. to Eq. (3)
$p_{m,r}^{\text{MEC}}$	Price of computing resource r in EN m
p_{c}^{RAN}	Price of network resources in cell c

Let $x_{s,m,r}$ be the amount of type-r computing resource reserved to service provider s in node m. Then for any resource r, $\frac{x_{s,m,r}}{d_{s,r}^{\text{MEC}}}$ represents the maximum number of concurrent jobs of s that such an allocation allows to be executed in EN m. However, the actual number of concurrently executed jobs is limited by the dominant resource. Thus, we must consider the minimum of these quantities over all the computing resource types. By summing over the nodes in the system, we get the maximum number of jobs that can be simultaneously executed with acceptable per-job performance in the MEC domain for service provider s:

$$J_s^{\text{MEC}}(\mathbf{X}_s) = \sum_{m \in \mathcal{M}} \min_{r \in \mathcal{R}} \left\{ \frac{x_{s,m,r}}{d_{s,r}^{\text{MEC}}} \right\}, \tag{1}$$

where $\mathbf{X}_s = (x_{s,m,r}) \in \mathbb{R}^{M \times R}$ is the computing resources allocation matrix.

Similarly, by letting $y_{s,c}$ be the network resources allocated in cell c for service provider s, the ratio between $y_{s,c}$ and $d_{s,c}^{\text{RAN}}$ can be identified as the largest number of job payloads of service s that can be simultaneously sent to the MEC cluster through cell c. Summing over all the cells in the system, we get the number of payloads of service provider s that can traverse the RAN domain for a given network resources allocation vector $\mathbf{Y}_s = \{y_{s,1}, ..., y_{s,C}\}$:

$$J_s^{\text{RAN}}(\mathbf{Y}_s) = \sum_{c \in \mathcal{C}} \frac{y_{s,c}}{d_{s,c}^{\text{RAN}}} \tag{2}$$

Finally, being the number of concurrently executed jobs limited by the least performing domain, we can express the service provider utility as follows.

$$u_s(\mathbf{X}_s, \mathbf{Y}_s) = \min \left\{ J_s^{\text{MEC}}(\mathbf{X}_s), J_s^{\text{RAN}}(\mathbf{Y}_s) \right\}. \tag{3}$$

4 Allocation Approaches

In this section, we present different approaches to the allocation problem in the settings described in Section 3, whose notation is summarized in Table 1. In particular, we propose a techno-economic market model where service providers individually buy bundles of resources in each edge node and cell. These resources will be then dedicated to the execution of jobs originating from their end-users. This coincides with the main contribution of this work. We also present two other possible allocation approaches based on system performance optimization without economic considerations. A fourth scheme based on proportional sharing is also proposed.

The first three approaches are based on mathematical programming, and the resulting resource allocation decisions are obtained by solving an optimization problem. We now describe the variables, parameters and constraints that will be shared by these formulation.

Quantities $x_{s,m,r}$ and $y_{s,c}$ defined in Section 3 represent the resources allocated to a particular SP, and these will be the output of any allocation mechanism. It follows that these coincide with the decision variables of our formulations. We define $j_{s,m}$ as the variable associated with the number of concurrent jobs of SP $s \in \mathcal{S}$ executed in EN $m \in \mathcal{M}$. Finally, let u_s be the decision variable representing the utility of SP $s \in \mathcal{S}$. The constraints of the optimization models employed by the approaches mentioned above are as follows:

$$j_{s,m} \leq \frac{x_{s,m,r}}{d_{s,r}^{\text{MEC}}}, \qquad \forall s \in \mathcal{S}, m \in \mathcal{M}, r \in \mathcal{R} \quad \text{(4a)}$$

$$u_s \leq \sum_{m \in \mathcal{M}} j_{s,m}, \qquad \forall s \in \mathcal{S} \quad \text{(4b)}$$

$$u_s \leq \sum_{c \in \mathcal{C}} \frac{y_{s,c}}{d_{s,c}^{\text{RAN}}}, \qquad \forall s \in \mathcal{S} \quad \text{(4c)}$$

$$\sum_{s \in \mathcal{S}} x_{s,m,r} \leq D_{m,r}^{\text{MEC}}, \qquad \forall m \in \mathcal{M}, r \in \mathcal{R} \quad \text{(4d)}$$

$$\sum_{s \in \mathcal{S}} y_{s,c} \leq D_{c}^{\text{RAN}}, \qquad \forall c \in \mathcal{C} \quad \text{(4e)}$$

$$x_{s,m,r},\ y_{s,c},\ j_{s,m} \geq 0, \quad \forall s \in \mathcal{S}, m \in \mathcal{M}, r \in \mathcal{R}, c \in \mathcal{C}. \quad \text{(4f)}$$

These represent a mathematical expression of the system model detailed in Section 3. In particular, constraints (4a) to (4c) linearize the utility function defined in Eq. (3). Constraint (4a) expresses the $\min$ operator in Eq. (1), while constraint (4b) forces the utility not to be larger than the number of jobs that can be concurrently executed in the MEC domain. Similarly, constraint (4c) limits the utility u_s to the maximum number of payloads that can be simultaneously processed by the RAN domain, as defined in Eq. (2). Constraint (4d) is such that the total allocation for each resource in each EN does not exceed the node capacity, which is required for a feasible allocation. Likewise, constraint (4d) limits the allocated network resource in each cell to the available capacity.

These constraints show how the heterogeneity of resources and diversity of attributes are considered by the allocation mechanisms we propose. For instance, heterogeneous resources are considered in the MEC domain by (4a), as it iterates over the set of available computing resource types $\mathcal{R}$. Different attributes can also be considered by adequately choosing the parameters. For instance, parameters $D_{m,r}^{\text{MEC}}$, each representing the available capacity of resource type r in EN m, can effectively model different server configurations in the MEC cluster. We now detail the allocation

approaches that will make use of the constraints above in their optimization models.

4.1 Market Equilibrium Approach

We propose a techno-economic market of resources model Γ based on a Fisher Market [5] model. In Γ, buyers are represented by the individual service providers in set $\mathcal{S}$. Each resource type in the system represents the goods in the market. Therefore we identify $M \times R + C$ different divisible goods.

The market capacity of each good is given by the availability of each resource, namely $D_{m,r}^{\text{MEC}}$ and D_c^{RAN}. Each resource has an associated unitary price $p_{m,r}^{\text{MEC}}$ (for MEC domain resources) and p_c^{RAN} (for RAN domain resources) that SPs must pay if they intend to acquire it. Resources bought by providers in the market constitute their allocation vectors $\mathbf{X}_s$ and $\mathbf{Y}_s$, yielding performances that each SP s can privately evaluate by means of the utility function u_s, as defined in Eq.(3) and expressed by constraints (4a) to (4c).

In this framework, we expect service providers to act as rational agents, all having the goal of pursuing their interest by buying resource bundles that maximize their utility (i.e., maximizing the number of simultaneously executable jobs) constrained by their budget B_s. The value of each provider's budget has the additional function of enforcing service prioritization. Suppose, for instance, that two SPs have the same demand profile but a different budget. In this case, the SP with the higher budget would be favored by the market model since it could afford to buy larger resource bundles. It follows that, even if the demand profiles are the same, the jobs of the heavier-budgeted SP would be prioritzed as these would have access to more resources.

As it will be shown later, resource pricing in the proposed approach follows the natural law of demand and supply by assigning comparatively higher prices to those resources that are more desired and less available. Accordingly, prices are expected to have a load balancing effect on the allocations, as SPs are incentivized to buy resources where congestion is limited.

Since both network and computing resources are allocated according to the rational decisions of providers and their interactions, the analysis of the resulting allocation mechanism is carried out utilizing tools from Game Theory [29] and general Equilibrium Theory [30], focusing on the characterization of the market equilibrium (ME) outcome for this market model.

Resource bundles $(\mathbf{X}_1, \ldots, \mathbf{X}_S, \mathbf{Y}_1, \ldots, \mathbf{Y}_S^*)$ and corresponding prices $\left(p_{m,r}^{\text{MEC}}, p_c^{\text{RAN}}\right)$ are said to induce a Market Equilibrium if the following conditions are true [31]:

(i) **Optimal Goods**: Every service provider buys only those goods that yield the maximum utility per unit of money, also known as maximum *bang-per-buck*;
(ii) **Market Clearing**: Either resources are fully allocated, or the corresponding price is zero. Furthermore, each service provider exhausts its budget.

The first condition is intended as a satisfaction of the rationality and selfishness of the buyers, whose unique interest is to maximize the return of their market investment in the manner prescribed by their utility functions. Intuitively, no equilibrium could be otherwise established. Condition (ii) states that the offer of each resource can either meet the demand and thus be priced, or be in a surplus condition and given away for free. Additionally, stability can be reached only when each buyer spends its entire budget, entailing that any unused fraction can be indeed spent to increase the utility.

SPs are allowed to buy any bundle of resources, provided that its price does not exceed the budget. However, it is clear now that equilibrium bundles are those who favor SPs the most. For this reason, we have developed a convex optimization program to find a Market Equilibrium for the aforementioned market model. In particular, the formulation is based on the variables and the constraints (4a-4f) with an addition of the following objective function:

$$\max \sum_{s \in \mathcal{S}} B_s \log(u_s). \tag{5}$$

This objective maximizes the sum of the SP utility logarithms, weighted by the budget of each provider. In Section 5, we will show how this objective is sufficient to prove that the solution of the resulting convex program represents an equilibrium point for the market model.

Our proposed mechanism works according to the following steps:

1) The system model parameters are gathered, and the formulation is built,
2) The model is optimized to obtain equilibrium allocations $\mathbf{X}^* = (\mathbf{X}_1^*, \ldots, \mathbf{X}_S^*)$, $\mathbf{Y}^* = (\mathbf{Y}_1^*, \ldots, \mathbf{Y}_S^*)$ and optimal SP utilities $\mathbf{u}^* = (u_1^*, \ldots, u_S^*)$
3) Resources are allocated throughout the system accordingly

Optimal prices $p_{m,r}^{*\text{MEC}}$ and $p_c^{*\text{RAN}}$ do not directly appear in the solution but, as strong duality holds, they can be extracted as dual variables [32] associated with capacity constraints (4d) and (4e). Consequently, these represent the rate of change of the objective function associated with any change in the capacity constraint. In other words, those resources whose increment or decrement of capacity would impact the objective the most will have the highest prices, suggesting a direct proportionality between prices and resource demand. Additionally, prices are such that the optimal resource bundle of each provider does not present costs higher than the provider's budget, as it will be proven in Section 5.

4.2 Social Optimum Approach

We propose an approach based on linear optimization that aims to allocate resources such that the overall number of jobs that can be simultaneously executed without provoking performance degradation is maximized. In this case, the resulting allocation cannot be guaranteed to represent a market equilibrium of any kind. Consequently, this mechanism does not consider service providers as entities capable of making decisions, but simply manages the system resources to maximize the overall performance according to some objective functions. On this matter, we propose two different objective functions to maximize the system performance: a pure maximization of the sum of utilities (i.e. total executed jobs) and a sum of utilities weighted by each provider's

budget. These objectives, together with variables and constraints (4a-4f) defined before, constitute the optimization models at the core of the proposed approaches.

The first objective function, also known as Social Optimum (SO), is defined as follows:

$$\sum_{s \in \mathcal{S}} u_s, \tag{6}$$

This objective leads to a maximization of the system performance regardless of the budget difference among services. That is to say, it maximizes the number of served users.

The second objective function is instead defined as:

$$\sum_{s \in \mathcal{S}} B_s u_s. \tag{7}$$

It includes a certain degree of fairness in the solution by giving more weight to service providers with larger budgets, replicating the advantage that such providers have in the market model Γ. Consequently, of the two mechanisms above, only the latter allows for service prioritization by means of the budget values. This type of solution is also known as Weighted Social Optimum (WSO).

The two resulting allocation mechanisms (namely SO and WSO) follow the 3-steps approach of the market model. In other words, the optimization models are solved with the parameters of the underlying system, and then the resources are allocated accordingly.

4.3 Proportional Sharing Scheme

This resource sharing scheme was proposed as a static mean of allocating resources in network slices [33] and has been used as a baseline allocation scheme both in resource management for network slicing [9] and mobile edge computing [19]. This scheme is based on a closed form solution and does not employ any optimization techniques.

Proportional sharing (PS) allocation mechanism is such that each service provider receives an amount of the capacity of each resource in the system that is proportional to its budget. Therefore, it can be seen as an extension of the well known *generalized processor sharing* [34] algorithm to a multi-resource scenario.

Formally, let $\hat{x}^r_{s,m}$ and $\hat{y}_{s,c}$ be respectively the computing and network resources allocations of service provider s resulting from proportional sharing, then:

$$\hat{x}_{s,m,r} = \frac{B_s}{\sum_{s' \in \mathcal{S}} B_{s'}} D^{\text{MEC}}_{m,r}, \qquad \forall s \in \mathcal{S}, m \in \mathcal{M}, r \in \mathcal{R}, \tag{8}$$

$$\hat{y}_{s,c} = \frac{B_s}{\sum_{s' \in \mathcal{S}} B_{s'}} D^{\text{RAN}}_{c}, \qquad \forall s \in \mathcal{S}, c \in \mathcal{C}. \tag{9}$$

Once these quantities are computed, proportional sharing utilities $\hat{u}_s$ can be computed through Eq. (3).

This allocation scheme does not consider the single provider's demands nor the load conditions of ENs and cells. Indeed, it simply allocates all the available resources such that service providers with larger budgets get comparatively larger shares. Thus, it is considered a static allocation scheme, as it does not reflect any change in the system conditions. However, following the same rationale of the market-based model, this mechanism prioritizes those SPs with larger budgets.

5 Market Model Properties

This section presents a theoretical analysis of the previously defined market model Γ and the other proposed mechanisms. We first discuss the existence of a Market Equilibrium. In particular, we prove that the solution of (5, 4a-4f) is always a ME for Γ. Afterwards, we turn our focus on the analysis of efficiency and fairness of the mechanisms proposed in Section 4.

5.1 Market Equilibrium

The existence of a Market Equilibrium (ME) for the Fisher Market model is guaranteed in the most generalized settings under the mild condition that at least one buyer desires each resource and each buyer desires at least one resource [35]. In the remainder of this work, we consider this condition to always hold without loss of generality since we assume that jobs require a non-zero amount of all types of resources in the system to be successfully executed[2].

Having established the existence of a ME for Γ, we turn our focus to its computation. Sperner's coloring [36] can find an equilibrium point, although with a high computational inefficiency. A more efficient and well-known alternative is to employ a convex optimization formulation called the Eisenberg-Gale (EG) program. EG was first proposed to find an equilibrium in the case of linear utilities [37] and then extended to more general functions [38]. Program (5, 4a-4f) is indeed a variant of the EG formulation applied to the proposed market model, where constraints (4a), (4b) and (4c) are added to the usual EG formulation. They express the particular bottleneck-defined service provider utility function that we employ in the proposed model, as defined in Eq. (3).

With the following theorem we prove that the solution of such formulation is in the set of market equilibria of Γ:

Theorem 1. *Optimal variables* $\mathbf{X}^* = (\mathbf{X}^*_1, \ldots, \mathbf{X}^*_S)$ *and* $\mathbf{Y}^* = (\mathbf{Y}^*_1, \ldots, \mathbf{Y}^*_S)$ *and corresponding optimal prices of program (5, 4a-4f) represent a market equilibrium of* Γ.

Proof: See appendix.

This theorem allows (5, 4a-4f) to be effectively used as a tool for computing a ME for Γ. In particular, (5, 4a-4f) can be applied to any instance of Γ and system resources can be allocated according to the optimal variables $\mathbf{X}^*$ and $\mathbf{Y}^*$.

Resource prices can be extracted from the solution as dual variables of capacity constraints in MEC and RAN domains. The system performance can be readily evaluated using optimal utilities. The exact contribution of prices to the equilibrium conditions is complex, and we have not included it in this section for ease of exposition. Instead, we refer to Theorem 1 for the mathematical details.

Program (5, 4a-4f) is in the class of convex optimization problems, which modern numerical methods, such as the interior-point method [39], can efficiently solve.

The objective function is strictly concave. Thus, the problem has at most one optimal point [40], meaning that the set of market equilibria utilities of Γ obtained through (5, 4a-4f)

2. If this was not the case, then the non-desired resource could be excluded from the market model without affecting the allocation solution.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2021.3091764, IEEE Transactions on Mobile Computing

contains a single element, as summarized by the following remark.

Remark 1. *The market equilibrium utilities obtained through program (5, 4a-4f) is unique.*

This result is not trivial since there is no guarantee that the equilibrium of a Fisher Market is unique. Indeed, different algorithms might converge to different equilibria depending on the initialization, as is it typical for best response dynamics [41] or Sperner's coloring.

5.2 Efficiency Properties

Generally speaking, the efficiency of an allocation mechanism quantifies the loss in overall system performances against the maximum allowed by the system capacity. There is no guarantee that a market equilibrium-based approach efficiently utilizes system resources. In general, the efficiency of such a solution depends on the parameters of the specific allocation instance. For this reason, this section presents some general efficiency results based on bounds, while a numerical evaluation of this metric is given in Section 6.

We define the efficiency of an allocation mechanism as the ratio between its resulting total number of concurrently executed jobs and the maximum number of jobs that the capacity of the system allows for simultaneous execution. For any instance γ of Γ, let $SW_{me}(\gamma)$, $SW_{so}(\gamma)$, $SW_{ps}(\gamma)$ be the social welfare (i.e. the sum of service provider utilities) of the solutions given by market equilibrium, social optimum, and proportional sharing allocation mechanisms, respectively. These values are unique[3] for any instance of Γ.

The efficiency of an allocation mechanism can be evaluated by comparison with $SW_{so}(\gamma)$, since it represents the maximum value for the given instance. In details, we define the following efficiency expressions:

$$\eta_{me}(\gamma) = \frac{SW_{me}(\gamma)}{SW_{so}(\gamma)}, \tag{10}$$

$$\eta_{ps}(\gamma) = \frac{SW_{ps}(\gamma)}{SW_{so}(\gamma)}. \tag{11}$$

We claim that the market-equilibrium approach has an efficiency at least as high as proportional sharing, namely $\eta_{me}(\gamma) \geq \eta_{ps}(\gamma)$, for any market instance γ. This result comes from the following theorem:

Theorem 2. *For a given market instance, let $(u_1^*, \ldots, u_s^*)$ be the market-equilibrium utilities and $(\hat{u}_1, \ldots, \hat{u}_s)$ be the utilities given by the proportional sharing mechanism. Then $u_s^* \geq \hat{u}_s$ for each service provider $s \in \mathcal{S}$.*

Proof. See appendix.

Interestingly enough, this theorem goes beyond comparing the efficiency of the two allocation approaches. Indeed, it shows how each service provider cannot lose utility by taking part in the market mechanism with respect to the utility obtained through a proportional division of resources. This property is also known as *sharing incentive*, and it is widely considered as a desirable goal for multi-resource allocation mechanism applied to data centers [42]. It confirms that the players can be better off by entering into the market mechanism than recurring to externally imposed proportional allocations.

Next, we compare the market-equilibrium efficiency against the maximum efficiency given by the social optimum solution. In particular, in [43] it is proven that the proportional sharing mechanism has an efficiency asymptotically lower bounded by $1/\sqrt{S}$ under mild conditions for the utility functions, which hold true for Γ as well. Furthermore, this bound is tight.
By Theorem 2, proportional sharing cannot yield efficiency higher than at Market Equilibrium, thus the bound also holds for η_{me} and the result is summarized in the following remark:

Remark 2. *$\eta_{me}(\gamma)$ is asymptotically lower bounded by $1/\sqrt{S}$ for any γ instance of Γ.*

Additionally, the remark above can be employed to characterize the price of anarchy (PoA) of the market-equilibrium approach, a well-known measure of efficiency in Equilibrium Theory [44]. PoA is defined as the ratio between the worst-case social welfare of the equilibrium solution and the optimal social welfare obtained through optimization. PoA can be expressed as a function of the previously defined ME efficiency:

$$PoA_{me} = \frac{1}{\min_{\gamma \in \Gamma} \eta_{me}(\gamma)}, \tag{12}$$

that is, the inverse of the worst case ME efficiency.

Given this definition, an upper bound for PoA_{me} is derived in the following remark:

Remark 3. *PoA_{me} is asymptotically upper bounded by $\sqrt{S}$.*

The last efficiency-related trait of the proposed allocation mechanism analyzed in this section is the well known *Pareto Efficiency* (PE) [30]. A feasible utility vector $(u_1, \ldots, u_S)$ obtained through any allocation mechanism is said to be Pareto optimal if there is no other feasible utility vector $(u_1^*, \ldots, u_S^*)$ such that $u_s^* \geq u_s$ for all $s \in \mathcal{S}$ and $u_s^* > u_s$ for at least one s. Any allocation mechanism whose utilities are Pareto optimal is said to be PE, and this property is often considered a minimal notion of efficiency.

It is easily verifiable that both social optimum and proportional fairness solutions are PE, while the same is not generally valid for equilibrium points. However, under some mild conditions also holding for Γ, the *First Fundamental Theorem of Welfare Economics* [30] states that any equilibrium of a competitive market is PE.

5.3 Fairness Properties

Fairness is a largely adopted concept when designing resource allocation mechanisms, and it is unanimously considered fundamental in guaranteeing a certain level of QoE among heterogeneous services. However, differently from efficiency, there is no univocal definition of fairness. In this work, fairness is analyzed through a popular fairness index. Additionally, we show how the Market Equilibrium satisfies some commonly desired fairness properties.

3. This is true for market equilibrium, as stated by Remark 1, but also for social optimum since it is the optimal value of the objective of a linear program, and for proportional sharing by definition.

To numerically quantify the fairness of an allocation, we employ the *Nash Social Welfare* (NSW), a well known and regarded fairness index considered to naturally achieve a compromise between fairness and efficiency [45] that has been employed in similar works [7], [9]. Here follows the definition of NSW adapted to the proposed system model:

$$NSW(u_1, \ldots, u_s) = \prod_{s \in \mathcal{S}} u_s^{B_s}. \tag{13}$$

One can note that NSW coincides with the geometric mean of utilities weighted by provider budgets, suggesting a trade-off between overall system performances and individual fairness.

It is known that EG-based formulations such as (5, 4a-4f) maximize NSW. This results comes from observing that the objective function (5) is equivalent to the definition of NSW, formally:

$$\arg\max \left\{ \sum_{s \in \mathcal{S}} B_s \log(u_s) \right\} = \arg\max \left\{ \prod_{s \in \mathcal{S}} u_s^{B_s} \right\}. \tag{14}$$

Consequently, the Market Equilibrium obtained through (5, 4a-4f) yields the highest fairness among any feasible allocation, including solutions obtained through proportional sharing and social optimum. Additionally, Remark 1 states the uniqueness of the equilibrium utilities obtained through (5, 4a-4f). This means that the proposed mechanism is capable of selecting the unique equilibrium point with the highest fairness according to NSW. This result has been summarized in the following remark:

Remark 4. *The utilities obtained through program (5, 4a-4f) are the unique NSW maximizer.*

After having characterized the NSW of the market-equilibrium approach from an absolute standpoint, it is natural to question how it compares to the solution which maximizes the efficiency, i.e. the social optimum approach. On this matter, we declare the following proposition, whose proof is reported in Appendix.

Proposition 1. *The loss of NSW fairness incurred when employing the social optimum approach with respect to the market equilibrium approach is unbounded.*

According to this result, it is evident how the social optimum solution presents no fairness guarantee, despite offering the best efficiency. Indeed, Section 6 shows how social-optimal solutions almost always result in at least one service provider getting zero resources, which is arguably unacceptable from a quality of service standpoint.

In addition to a quantitative fairness analysis through NSW, we consider two well-known fairness properties that characterize the qualitative fairness of the market equilibrium approach, namely *Proportional Fairness* (PF) and *Envy-Freeness* (EF). PF, long considered a desirable fairness property of allocation mechanism, was first introduced in the context of bandwidth sharing among network flows [46] and has been proven to be relevant for data-centre resource sharing in [47]. A feasible utility vector $(u_1, \ldots, u_S)$ is said to be proportional fair if, for any other feasible utility vector $(\bar{u_1}, \ldots, \bar{u_S})$, the aggregate of proportional changes is negative [48], formally:

$$\sum_{s \in S} B_s \frac{\bar{u_s} - u_s}{u_s} \leq 0. \tag{15}$$

We characterize the market's proportional fairness property in the following proposition, whose proof is reported in the Appendix.

Proposition 2. *The market equilibrium obtained through program (5, 4a-4f) is proportional fair.*

The final fairness property analyzed in this section is EF, which states that each agent in the market prefers its assigned bundle of resources over the bundle of any other agent [49]. This well-known property owes its desirability to the stability that it implies, as agents have no reason to complain about their allocation. However, in the context of a fixed budget market model such as Γ, the discussion about EF is meaningful only when service providers all have the same budget. Under this assumption, resource bundles $(\mathbf{X}_1, \ldots, \mathbf{X}_S)$ and $(\mathbf{Y}_1, \ldots, \mathbf{Y}_S)$ are considered envy-free if $u_s(\mathbf{X}_s, \mathbf{Y}_s) \geq u_s(\mathbf{X}_t, \mathbf{Y}_t)$ for any $s, t \in \mathcal{S}$.

The proposition that follows claims the envy-freeness property of the market equilibrium.

Proposition 3. *The ME obtained through (5, 4a-4f) when service providers all have an equal budget is envy-free.*

The proof is straightforward once one notes that equilibrium allocations are mutually affordable, given that all providers experience the same resource prices and have equal budgets. Hence, each provider can afford to buy any opponent's resource bundle but still prefers its own since it is the one that maximizes its utility.

6 Numerical Results

In this section, we numerically evaluate the performance and fairness of the allocation approaches proposed in section 4 applied to different instances of the system model. Such instances are characterized by parameters that vary according to the specific result or property we intend to highlight. However, some assumptions will be valid throughout the entire section unless otherwise stated.

Every market model is motivated by the heterogeneity of resource demands among different buyers. Those cases where buyers show the same or similar demand profiles are less attractive, as an optimal solution would not be much different from an equal resource partitioning. Consequently, the main aspect in our numerical analysis is the ratio between resource demands among different service providers and with respect to the MEC capacity, not their absolute values. This allows us to consider generic compute-equivalent and memory-equivalent units, CPU-U and MEM-U respectively.

Consequently, capacities $D_{m,r}^{\text{MEC}}$ measure the maximum available units of each resource in a particular edge node. Similarly, service requirements $d_{s,r}^{\text{MEC}}$ in the MEC domain define the computing resource units required by a single job of SP s. Network resources are considered as portions of the total bandwidth available in a specific cell. Thus, parameters $d_{s,c}^{\text{RAN}}$ and D_c^{RAN} are numerically represented in Hertz (Hz).

TABLE 2: Service Templates

Name	$d^{\text{MEC}}_{s,\text{CPU-U}}$	$d^{\text{MEC}}_{s,\text{MEM-U}}$	$d^{\text{RAN}}_{s,c}$	B_s
CPU-Intensive	4 CPU-U	8 MEM-U	3MHz	1
RAM-Intensive	1 CPU-U	32 MEM-U	3MHz	1
BW-Intensive	1 CPU-U	8 MEM-U	10MHz	1.5
Balanced	5 CPU-U	40 MEM-U	5MHz	2

TABLE 3: Nodes and cells configuration

Name	CPU-U	MEM-U	BW
CPU Node	32 CPU-U	128 MEM-U	/
MEM Node	16 CPU-U	256 MEM-U	/
Small Cell	/	/	40 MHz
Large Cell	/	/	20 MHz

Simulation instances have been generated in MATLAB, and the proportional sharing allocation mechanism was implemented in the same environment. Market Equilibrium, Social Optimum and Weighted Social Optimum mechanisms have been solved through IPOPT [50] and CPLEX, respectively.

6.1 System Performance and Fairness

To capture the heterogeneity of resource requirements of different services, we define four service templates that represent possible demands configurations in a cloud computing scenario. In details, we define three templates representing CPU-intensive, MEM-intensive and bandwidth-intensive services and an additional, more balanced service with overall high resource requirements. These particular templates have been chosen such that a heterogeneous mix of demands are analyzed. Numerical values of service templates are detailed in Table 2.

We consider a localized MEC/RAN system comprising heterogeneous network cells and computation nodes.
In the RAN domain, we identify 2 large cells with 40MHz of capacity and 5 smaller cells with 20MHz of capacity. In the MEC domain, we identify 2 types of EN, whose computation resources are either 32 CPU-U and 128 MEM-U, for what we call a *CPU node*, or 16 CPU-U cores and 256 MEM-U, for what we call *MEM node*. These values are presented in Table 3.

Our simulated system includes 5 *CPU nodes* and 5 *MEM nodes*. In each simulation run, one among the templates above is given randomly to each of the 15 service providers competing for system resources. Furthermore, Gaussian noise is added to each resource requirements with mean 0 and variance equal to 25% of the original numerical value. This has been done to account for requirements oscillations given by unpredictable variations in payload size and computational effort, as well as network conditions.

Figure 3 shows the cumulative distribution functions of some performance indicators extracted from 100 simulated instances. In particular, Figure 3a allows us to immediately appreciate the impact of different allocation approaches on the number of concurrent jobs of each service provider, i.e. its utility. Social optimum-based solutions present zero executed jobs (i.e. zero allocated resources) for at least 60% of the cases, while the rest of the providers obtain relatively high performance. This behavior appears in this type of approach as the solution favors those providers who can simultaneously execute the most jobs with fewer resources.

On the other hand, ME and PS approaches always guarantee that each SP can execute some jobs, even if this means a lower maximum per-provider performance. Furthermore, this particular figure shows how providers always get better performance with the ME approach in respect of PS, confirming the *sharing incentive* property implied by Theorem 2.

Figure 3b shows the cumulative distribution functions of the sum of per-provider concurrently executable jobs, namely the system performance, for each of the studied allocation approaches. The result is that SO outperforms the other approaches, and ME consistently delivers better performance than PS, shown to be the least performing. This is also evident in Figure 3c, where the distribution of efficiencies η_{ME} and η_{PS} are plotted. Here it is shown how the ME approach is about 30% more efficient than PS on average, a result that again confirms the efficiency remarks of Section 5. Furthermore, ME presents a worst-case efficiency of 52%, higher than the theoretical lower bound of 26% given by Remark 2.

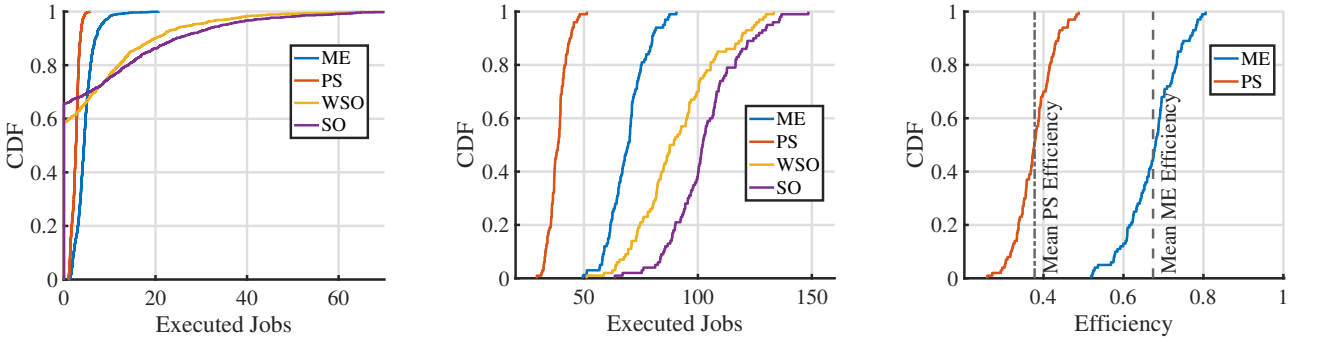
Figure 4 shows the cumulative distribution function of the NSW for both ME and PS approaches. Plots corresponding to social optimum-based approaches are missing because such solutions almost always present at least one service provider with 0 concurrently executed jobs, yielding NSW values that are mainly 0. Moreover, the figure's horizontal axis employs a logarithmic scale due to both a significant difference between worst-case and best-case values and the gap between the two distributions. This behavior is due to the fact that the considered fairness index exponentially increases with both budgets and the number of service providers. Additionally, Figure 4 confirms Remark 4 by showing how NSW of market equilibrium approach is much larger than the proportional sharing case in all simulated instances.

6.2 Sensitivity Analysis

We start by analyzing the different sensitivity of allocation mechanisms to variations in budget values from the per-provider performance standpoint and with respect to overall system performance. In this analysis, we consider the previously detailed system deployment and three providers associated with different services: 2 CPU-intensive services and one Balanced service, which will be called *S1*, *S2*, and *S3* respectively.

Figures 5a and 5b show the variations in performance when the budget of *S1* is allowed to increase from its default value of 1 up to 5. In particular, Figure 5a demonstrates how single SP utilities adapt to the budget variations in the ME mechanism. When *S1* and *S2* have equal budgets, they obtain the same utility since their requirements are symmetrical and the mechanism cannot prefer either of the two. As *S1* budget increases, the allocation mechanism grants her increasingly more resources, confirming the service prioritization effect of the budget values. When *S1* and *S3* budgets are the same, i.e. 2, the mechanism demonstrates efficiency by granting more resources to *S1*, which can execute more jobs under the same allocation conditions because it requires less per-job resources.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2021.3091764, IEEE Transactions on Mobile Computing



(a) Cumulative distribution function of provider utility.

(b) Cumulative distribution function of concurrently executed jobs in the system.

(c) Cumulative distribution function of ME and PS efficiency

Fig. 3: Cumulative distribution functions of some performance parameters.

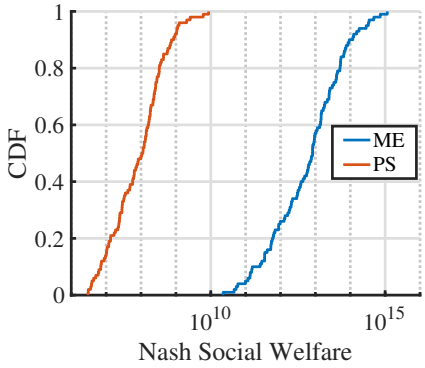
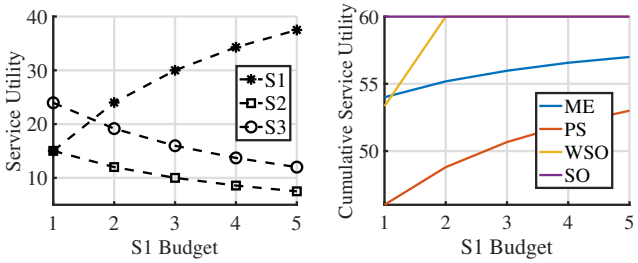
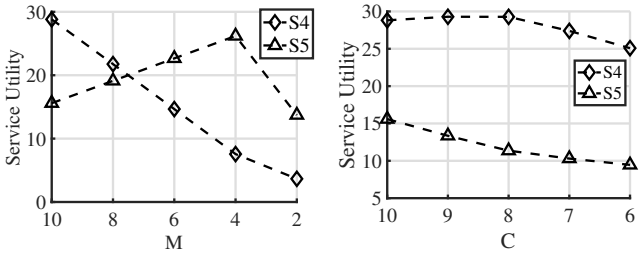


Fig. 4: Nash Social Welfare cumulative distribution function of ME and PS.



(a) ME single service utility vs budget

(b) System performance vs budget



(c) ME single service utility vs M

(d) ME single service utility vs C

Fig. 5: Sensitivity Analysis

Figure 5b shows the impact of *S1* budget variations on the cumulative provider utility when employing different allocation mechanisms. SO remains unaffected, while the other mechanisms show an increment in overall performance due to a more efficient resource utilization when increasingly larger allocations are granted to *S1*.

Finally, we present an analysis of the service provider utilities obtained through ME when the number of nodes and cells in the system varies. In this case, consider a system comprising 5 *small cells* and 5 *large cells* and the same MEC deployment as the previous analyses. In this scenario, one Balanced service and one BW-intensive service populate the system, offered by service providers *S4* and *S5* respectively. Leaving the RAN domain unchanged, we allow for M to decrease from 10 to 2 by iteratively excluding one *CPU node* and one *RAM node* at the same time.

Figure 5c shows the impact on the utility of the two considered service providers. As the computational capacity of the system starts decreasing, *S4* undergoes a decrease in the executed jobs, which in turns frees some radio access capacity that can be utilized by bandwidth-hungry *S5*, providing a performance boost. This behavior continues until the computational capacity is so low that it forms a bottleneck for both providers. This shows in practice the presence of the domain bottleneck phenomenon that has been introduced in Section 3.

Considering again the original deployment, we analyze the impact of RAN capacity variations by decreasing the number of *small cells* from 5 to 1, while leaving the rest untouched. Figure 5d illustrates a behavior similar to the previous analysis, even if less accentuated. Indeed, as the radio resources decrease, *S5* hits a resource bottleneck and her performance decrease. This causes more computational resources to be available to *S4*, allowing for a slight but noticeable performance increase. As seen before, the domain bottleneck is eventually hit, and both service providers start experiencing a lower performance.

7 Conclusion

In this work, we have analyzed the problem of efficient and fair joint management of radio and computing resource in a MEC/RAN deployment. We have proposed a techno-economic market model where service providers sharing

the same physical infrastructure can purchase resources while being constrained to a budget limit. Also, we have formulated a convex program that computes the market equilibrium for any instance of the market model. We have given an extensive analysis of the properties of such a solution in terms of efficiency and fairness. In particular, we have shown how how the properties of *Pareto Optimality*, *Nash Social Welfare* maximization, *Proportional Fairness*, *Sharing Incentive* and *Envy-Freeness* apply. As it is usual for game theory-based works, we have also characterzed the price of anarchy of the solution. Additionally, we have proposed alternative multi-resource allocation mechanisms which are not based on any market concept, and we have compared them with the market equilibrium approach.

Finally, we have presented numerous results coming from a simulation of the proposed allocation mechanisms. In particular, we have numerically analyzed and compared the performance of the mechanisms, confirmed the theoretical properties of the market model and gave an insight into the sensitivity of the solution to parameters variations.

REFERENCES

[1] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "Nr: The new 5g radio access technology," *IEEE Communications Standards Magazine*, vol. 1, no. 4, pp. 24–30, 2017.

[2] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2018.

[3] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[4] 3GPP, "Study on Radio Access Network (RAN) sharing enhancements," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 22.852, 09 2014, version 13.1.0.

[5] V. V. Vazirani, *Combinatorial Algorithms for Market Equilibria*. Cambridge University Press, 2007, p. 103–134.

[6] S. Lasaulce and H. Tembine, *Game theory and learning for wireless networks: fundamentals and applications*. Academic Press, 2011.

[7] A. Lieto, I. Malanchini, S. Mandelli, E. Moro, and A. Capone, "Strategic network slicing management in radio access networks," *IEEE Transactions on Mobile Computing*, 2020, to appear.

[8] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing in 5g: An auction-based model," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[9] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Perez, "Network slicing games: Enabling customization in multi-tenant networks," in *Proceedings - IEEE INFOCOM*. Institute of Electrical and Electronics Engineers Inc., oct 2017.

[10] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6419–6432, 2018.

[11] W. Wang, B. Li, and B. Liang, "Dominant resource fairness in cloud computing systems with heterogeneous servers," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014, pp. 583–591.

[12] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, "Dominant resource fairness: Fair allocation of multiple resource types," in *Proceedings of NSDI 2011: 8th USENIX Symposium on Networked Systems Design and Implementation*, 2011, pp. 323–336.

[13] E. Meskar and B. Liang, "Fair multi-resource allocation with external resource for mobile edge computing," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2018, pp. 184–189.

[14] B. Xiang, J. Elias, F. Martignon, and E. Di Nitto, "Joint network slicing and mobile edge computing in 5g networks," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–7.

[15] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, no. 2, pp. 89–103, jun 2015.

[16] S. Yang, F. Li, M. Shen, X. Chen, X. Fu, and Y. Wang, "Cloudlet placement and task allocation in mobile edge computing," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5853–5863, 2019.

[17] D. Han, W. Chen, and Y. Fang, "Joint channel and queue aware scheduling for latency sensitive mobile edge computing with power constraints," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3938–3951, 2020.

[18] D. T. Nguyen, L. B. Le, and V. Bhargava, "Price-based Resource Allocation for Edge Computing: A Market Equilibrium Approach," *IEEE Transactions on Cloud Computing*, jun 2018.

[19] D. T. Nguyen, L. B. Le, and V. K. Bhargava, "A Market-Based Framework for Multi-Resource Allocation in Fog Computing," *IEEE/ACM Transactions on Networking*, vol. 27, no. 3, pp. 1151–1164, jun 2019.

[20] M. Chiang and T. Zhang, "Fog and iot: An overview of research opportunities," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854–864, 2016.

[21] M. Liu and Y. Liu, "Price-based distributed offloading for mobile-edge computing with computation capacity constraints," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 420–423, 2018.

[22] L. Li, T. Q. S. Quek, J. Ren, H. H. Yang, Z. Chen, and Y. Zhang, "An incentive-aware job offloading control framework for multi-access edge computing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 1, pp. 63–75, 2021.

[23] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5994–6009, 2017.

[24] M. Siew, D. Cai, L. Li, and T. Q. S. Quek, "Dynamic pricing for resource-quota sharing in multi-access edge computing," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 2901–2912, 2020.

[25] J. Liu, S. Guo, K. Liu, and L. Feng, "Resource provision and allocation based on microeconomic theory in mobile edge computing," *IEEE Transactions on Services Computing*, pp. 1–1, 2020.

[26] S. Mandelli, M. Andrews, S. Borst, and S. Klein, "Satisfying network slicing constraints via 5g mac scheduling," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019, pp. 2332–2340.

[27] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[28] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 4, pp. 3098–3130, oct 2018.

[29] B. Skyrms, *Game Theory, Rationality and Evolution*. Dordrecht: Springer Netherlands, 1997, pp. 73–86.

[30] A. Mas-Colell, M. Whinston, and J. Green, *Microeconomic Theory*. Oxford University Press, 1995.

[31] S. Brânzei, Y. Chen, X. Deng, A. Filos-Ratsikas, S. Kristoffer, S. Frederiksen, and J. Zhang, "The Fisher Market Game: Equilibrium and Welfare," in *Proc. of the 28th Conference on Artificial Intelligence (AAAI)*, Quebec City, Canada, jul 2014.

[32] R. Horst, "On the interpretation of optimal dual solutions in convex programming," *The Journal of the Operational Research Society*, vol. 35, no. 4, pp. 327–335, 1984.

[33] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 27–35, 2013.

[34] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 344–357, 1993.

[35] R. R. Maxfield, "General equilibrium and the theory of directed graphs," *Journal of Mathematical Economics*, vol. 27, no. 1, pp. 23 – 51, 1997.

[36] H. E. Scarf, "The core of an n person game," *Econometrica*, vol. 35, no. 1, pp. 50–69, 1967.

[37] E. Eisenberg and D. Gale, "Consensus of subjective probabilities: The pari-mutuel method," *The Annals of Mathematical Statistics*, vol. 30, no. 1, pp. 165–168, 1959.

[38] E. Eisenberg, "Aggregation of utility functions," *Management Science*, vol. 7, no. 4, pp. 337–350, 1961.
[39] F. A. Potra and S. J. Wright, "Interior-point methods," *Journal of Computational and Applied Mathematics*, vol. 124, no. 1, pp. 281 – 302, 2000.
[40] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[41] G. Ellison, "Learning, local interaction, and coordination," *Econometrica*, vol. 61, no. 5, pp. 1047–1071, 1993.
[42] P. Poullie, T. Bocek, and B. Stiller, "A survey of the state-of-the-art in fair multi-resource allocations for data centers," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 169–183, 2018.
[43] L. Zhang, "The efficiency and fairness of a fixed budget resource allocation game," in *Automata, Languages and Programming*. Springer, 2005, pp. 485–496.
[44] E. Koutsoupias and C. Papadimitriou, "Worst-case equilibria," in *STACS 99*. Springer, 1999, pp. 404–413.
[45] S. Brânzei, V. Gkatzelis, and R. Mehta, "Nash social welfare approximation for strategic agents," in *EC 2017 - Proceedings of the 2017 ACM Conference on Economics and Computation*. New York, NY, USA: Association for Computing Machinery, Inc, jun 2017, pp. 611–628.
[46] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.
[47] T. Bonald and J. Roberts, "Enhanced cluster computing performance through proportional fairness," *Performance Evaluation*, vol. 79, pp. 134–145, 2014.
[48] 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.
[49] D. Dolev, D. G. Feitelson, J. Y. Halpern, R. Kupferman, and N. Linial, "No justified complaints: On fair sharing of multiple resources," in *proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, pp. 68–75.
[50] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, Mar 2006.

Eugenio Moro is a PhD student at Politecnico di Milano, Department of Electronics, Information and Bioengineering. His research area is Telecommunication, with a focus on optimisation techniques and game theory applied to wireless networks. In 2016, he received his bachelor's degree in information engineering at Università del Salento. In 2017, he was enrolled in the MSc course of Telecommunications Engineering at Politecnico di Milano, where he graduated in 2019.

Ilario Filippini (S'06–M'10–SM'16) received B.S. and M.S. degrees in Telecommunication Engineering and a Ph.D in Information Engineering from the Politecnico di Milano, in 2003, 2005, and 2009, respectively. He is currently an Associate Professor with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano. His research interests include planning, optimization, and game theoretical approaches applied to wired and wireless networks, performance evaluation and resource management in wireless access networks, and traffic management in software defined networks. On these topics, he has published over 60 peer-reviewed articles. He serves in the TPC of major conferences in networking and as an Associate Editor of *IEEE Transactions on Mobile Computing* and *Elsevier Computer Networks*.