Multi-Resource Allocation for Network Slicing

Francesca Fossati, Stefano Moretti, Patrice Perny, Stefano Secci

Multi-Resource Allocation for Network Slicing

Francesca Fossati, Stefano Moretti, Patrice Perny, Stefano Secci, *Senior, IEEE*

***Abstract*—Among the novelties introduced by 5G networks, the formalization of the 'network slice' as a resource allocation unit is an important one. In legacy networks, resources such as link bandwidth, spectrum, computing capacity are allocated independently of each other. In 5G environments, a network slice is meant to directly serve end-to-end services, or verticals: behind a network slice demand, a tenant expresses the need to access a precise service type, under a fully qualified set of computing and network requirements. The resource allocation decision encompasses, therefore, a combination of different resources. In this paper, we address the problem of fairly sharing multiple resources between slices, in the critical situation in which the network does not have enough resources to fully satisfy slice demands. We model the problem as a multi-resource allocation problem, proposing a versatile optimization framework based on the Ordered Weighted Average (OWA) operator, that takes into account different fairness approaches. We show how, adapting the OWA utility function, our framework can generalize classical single-resource allocation methods, existing multi-resource allocation solutions at the state of the art, and implement novel multi-resource allocation solutions. We compare analytically and by extensive simulations the different methods in terms of fairness and system efficiency.**

***Index Terms*—multi-resource allocation, 5G slicing, OWA.**

I. Introduction

While the fourth generation (4G) of networks was designed for improving the smartphone experience mostly in terms of network throughput, the fifth generation (5G) is instead being designed with a much broader goal. 5G networks need to provide end-to-end connectivity, directly supporting verticals, including radio connectivity, wired connectivity and computing resource delivery and orchestration, exploiting system and network virtualization technologies [1]. 5G verticals include, e.g., e-health services, public safety systems, smart office, and connected vehicles, trains and aircrafts [2].

The provisioning abstraction being formalized by 5G activities is the so-called 'network slice'. A slice is meant to be a heterogeneous set of resources, optimized and concatenated with each other in such a way to serve a specific service or vertical (see Fig. 1). This implies that resources are shared among slices, and a portion of them is allocated to each slice to meet specific requirements of given vertical applications.

In this context, concepts of fairness integrated in legacy single-resource allocation algorithms and systems are challenged. In this paper, we address the following research questions: are the multiple resources called by a slice to be allocated one after the other independently of each other, or shall one take the multi-resource allocation as a joint allocation problem to increase system efficiency? In particular, if the request for at least one resource is bigger than the available one, we revisit how fairness in resource usage can be measured, and ensured by means of resource allocation algorithms. Moreover, we propose and compare multiple allocation rules and evaluate them in terms of wasted resource (i.e., resource allocated but eventually not used) and idle resource (i.e., resource left available for future allocations).

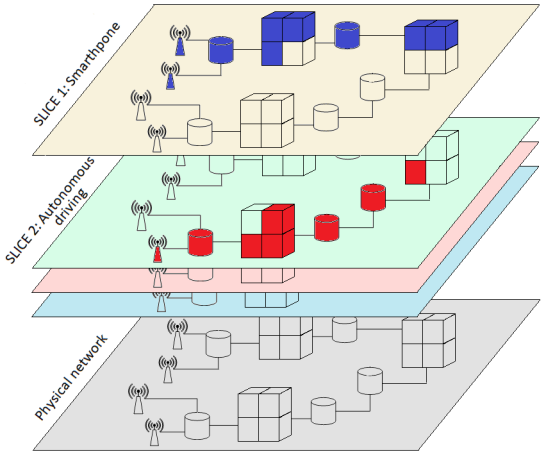


Fig. 1: A representation of network slices and resource sharing.

In this paper, we tackle the multi-resource allocation problem, running evaluations against the network slicing use-case[1]. Our objective is to provide fair multi-resource allocations: working at the multi-resource dimension allows avoiding the allocation of unnecessary surplus of resources, while maximizing the overall system fairness and user satisfaction [3].

We propose a unified mathematical framework able to generalize some of the classical solutions for single and multi-resource allocation problems from the literature. This framework takes into account both user satisfaction and system efficiency objectives, meeting different degrees of fairness. The main idea behind the proposed framework is to aggregate the information about users demands and the available resources in order to obtain a fairness objective function, depending on the satisfaction of the users, to maximize.

The paper is organized as follows. Section II summarizes the state of the art on resource allocation in network slicing, and on single and multi-resource allocation in networks. Section III presents a unified framework using the Ordered Weighted Average (OWA) operators. Properties of the proposed framework are described in Section IV. Section V discusses simulation results comparing existing and proposed methods. In Section VI we show how to generalize the framework with

F. Fossati, S. Secci are at Cnam, France. Email:firstname.lastname@cnam.fr
P. Perny is at Sorbonne Université, CNRS, LIP6, France. Email: patrice.perny@sorbonne-universite.fr
S. Moretti is at LAMSADE, CNRS, Université Paris-Dauphine, Université PSL, France. Email: stefano.moretti@dauphine.fr.
Manuscript received on Oct. 25, 2018.

[1]This work differs substantially from our previous work [24] that deals with single resource allocation problems.

Ref.	Type of resource/s	Mathematical model	Objective
Caballero et al. [4]	radio access network resources	optimization framework	fair allocation
Leconte et al. [5]	network bandwidth and cloud processing	optimization framework	fair allocation
Guan et al. [6]	VNF and link	complex network theory	maximization of the resource utilization and service provider revenues
Wang et al. [7]	access and network functions	optimization framework	slice dimensioning with resource pricing policy
Jiang et al. [8]	radio, storage and computational resources	auctions	resource and revenue optimization
Caballero et al. [9]	base stations or sectors	game theory	maximization of the tenants utility
Xiao et al. [10]	spectrum	game theory and distribute algorithms	to avoid reveal of private information
Halabian [11]	VNF	auction, optimization framework and distribute algorithms	maximize α-fair system utility
our approach	any resource	optimization framework	to provide a system-efficient framework capturing different fairness objectives

TABLE I: State of the art on resource allocation in network slicing

arbitrary resource dependency and Service Level Agreement (SLA) management. Finally, Section VII concludes the paper.

II. Background

In this section, we present the state of the art in resource allocation in network slicing. We introduce the basic characterizations of classical resource allocation rules[2], and we discuss the role of resource dependency.

A. Resource allocation in network slicing

Recent works in the literature use different approaches to model and solve the resource allocation problem in network slicing. Different perspectives are adopted, considering various resource types, alternative mathematical tools and different objectives. Table I summarizes and classifies major works at the state of the art, described in the following.

A recurrent approach is to integrate multi-resource considerations within the Virtual Network Function (VNF) placement algorithm. For example, authors in [4] address the slicing of radio access network proposing a multi-operator resource allocation rule, able to assign to each user a single base station and able to quantify a fair portion of the resource to assign to each user. Similarly, authors in [5] model the infrastructure with a direct graph and the slice as a simple source-destination pair, solving both the placement and the resource allocation as a unique problem. In [6] a model to place VNFs while selecting links is proposed using a complex network analysis.

Another approach that can be found in the literature is the one concerning the maximization of the slice costumer's profit [7] or the one considering the network revenue [8].

Modeling the problem as a competition between tenants sharing the same infrastructure, it is also possible to adopt a game-theoretic approach. In [9], authors propose a network slicing game in which users react to other tenants allocation and maximize their utility, converging to a Nash equilibrium. Distributed approaches are used in [10], where a cooperative game is introduced that, to avoid revealing mobile operators private information, uses a distributed algorithm to solve the allocation problem. Instead, in [11], both collaborative and non-collaborative approaches are analyzed and solved, using auctions between slices and datacenter providers for the former one, and a distributed approach for the latter one.

The approach we propose in this paper differs from the above mentioned ones in several aspects:

- we focus on the problem of allocating end-to-end resources taking into consideration multi-resource allocation protocols aimed at distributing amounts of each resource among the tenants, independently of the infrastructure; the actual embedding of each resource into a final resource allocation taking into account geographical distributions and a physical networking topology is considered as a separate, successive, problem;
- in our network model tenants express a demand for each resource and there is an actual problem when there is at least one congested resource, i.e., at least one resource cannot satisfy all the tenants;
- we consider resource dependency between resources, as done for instance in [5] (and elaborated hereafter);
- the possible allocation rules we propose span different concepts of fairness, namely, considering or not the awareness of the tenant with respect to the available resource and the other tenants demands.

B. Resource dependency and depletion

Virtualized network systems are evolving so that network functions nodes can be given computing power elastically and as a function of the load (i.e., virtual link bitrate), and that the spectrum allocated via medium access protocols can be flexibly adapted to the requested bitrate. There is indeed a dependency among different types of resources in such systems 5G networks leverage on. For example, for the computing resource to traffic bit-rate dependency, it is typically a linear [5], [12], [13] or step-linear or piece-wise linear relationship with few deflection points, as seen in [14], [15]; for the bitrate to spectrum one, a step-linear relationship can be inferred from slice specifications such as [16]. Taking such a behavior into account in network models is challenging. In the model proposed in this paper, we assume a linear relationship that can provide a good approximation to such step and piece-wise linear relationships.

In our analysis, we consider two aspects to assess an allocation solution when some resources are not enough to

[2]From now on we use interchangeably the word 'slice' and 'tenant'(or 'user') because we assume that to each tenant is associated a slice.

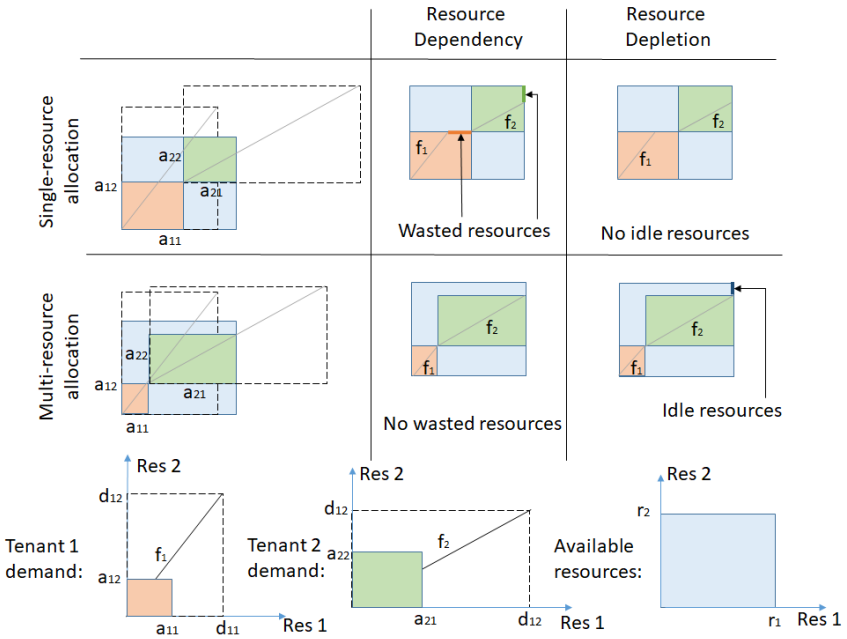


Fig. 2: Behavior of single and multi-resource allocations in terms of inter-resource dependency and resource depletion. f_i, $i = 1, 2$, is the relation between the resources for tenant i, whereas d_{ij} is the demand of the i^{th} user for the j^{th} resource and r_j, $j = 1, 2$, is the available amount for the j^{th} resource. a_{ij} is the allocation of the i^{th} user for the j^{th} resource. The horizontal axis represents the resource 1, the vertical one the resource 2.

fully satisfy tenants' requests. Firstly, each slice demand expresses an inter-resource linear relationship that has to be satisfied; e.g., the number of cores for a virtual machine in a slice can vary as a function of the bitrate and hence the link bandwidth allocated to the slice – i.e., one core needed every given amount of traffic: hence if less traffic is granted, a number or a proportion of core capacity can be saved. We refer to this aspect as *inter-resource dependency*, which can lead to wasted resource, i.e., allocated but not useful resource[3]. Secondly, we consider the *resource depletion*: a resource is depleted if it is fully distributed to slices. In the case of a single-resource system one aims at fully allocating the resource in order to provide an efficient solution, i.e., the resource is depleted, there is no idle resource left. In a multi-resource context, a multi-resource allocation rule taking into consideration inter-resource dependency can lead to idle resource, i.e., the resources may not be depleted.

Fig. 2 depicts a basic resource allocation problem example with 2 users and 2 resources, representing in a bi-dimensional space (i.e., the resource space) the users demand and the available resource. A single-resource approach considers a number of problems equal to the number of resources needed by the slice, producing allocations that do not take into consideration resource dependency (linear in the figure). In fact, we can notice that for both the users a portion of resource is allocated even if it cannot be used by the tenant. Contrarily, with a multi-resource approach, resources and demands are multi-dimensional and take into account the resource dependency. A multi-resource allocation rule may create idle resource, hence respecting resource dependency while meeting allocation goals such as fairness.

We detail in the following how single-resource and multi-resource allocation rules differ.

C. Single-resource allocation rules

A single-resource allocation problem consists of assigning part of a given resource to multiple users. Formally the problem can be characterized by a pair (d, r), where d is the vector of users' demands and r is the amount of available resource that has to be shared among the users. In case of n users, an allocation a is a n-dimensional vector specifying how much resource is given to each user i, and satisfying non-negativity ($a_i \geq 0$), demand boundedness ($a_i \leq d_i$) and efficiency ($\sum_{i=1}^{n} a_i = r$) [17]. Many allocation protocols exist to fairly share a scarce resource, i.e., a resource not large enough to fully satisfy all users ($\sum_{i=1}^{n} d_i > r$). The most common are the proportional rule and the weighted proportional rule [18], the Max-Min Fairness (MMF) rule [19], [20] and the α-fairness allocations [21]. The latter is a trade-off between MMF and weighted proportional. The weighed proportional allocation rule is preferred to the proportional one when we have explicit user demands. We further describe the remaining rules, and the recently proposed mood value rule [23], [24], we evaluate in the paper.

[3]Under the hypothesis of linear dependency of resources, if a user decreases its demand for one resource, automatically it decreases its demand for all the other resources. The wasted resource can be a problem from both the user and the provider point of view. In fact the first one pays for a resource that is not able to use while the second is providing a resource that is not used and that could be held back for itself to serve someone else. The waste of resource is automatically nullified when we consider a multi-resource approach respecting the linear relationship. So the proposed approach is able to meet two objectives: to avoid resource waste while ensuring fairness.

Weighted proportional rule [18]: it targets the maximization of $\sum_{i=1}^{n} p_i \log a_i$. If the weight p_i is equal to 1 for each user, the solution is the *proportional rule*; if it coincides with d_i, we obtain the allocation that assigns the same proportion of demand to each user. This last one equalizes the users' satisfaction, measured as the percentage of demand allocated ($\frac{a_i}{d_i}$), and as such equalizes the fairness, as for instance measured by the Jain's index [22].

MMF rule [19]: it is an egalitarian solution, privileging the weak users (with small demands), maximizing the minimum allocation, then the second lowest allocation, and so on.

Mood value rule [23], [24]: it modifies the way the user satisfaction is computed, taking into consideration the awareness of other users' demands and the amount of available resources. Classically the measure of satisfaction with respect to an allocation, for each user, is measured as the ratio between the allocated resource and the user demand. Under complete awareness, it is possible to calculate the minimal and the maximal right (i.e., allocation) for each user in order to better express the user satisfaction through the so-called Player Satisfaction (PS) rate:

$$ps_i = \frac{a_i - min_i}{max_i - min_i}, \tag{1}$$

where a_i is the allocation of user i, $min_i = \max\{r - \sum_{j \in N \setminus \{i\}} d_j, 0\}$ is the minimal right of the user i (i.e. what remains if all the other users are fully satisfied), max_i is the maximum it can get (i.e. its own demand d_i or the available resource r, if the demand overcomes it). The PS rate scales between 0 and 1 the satisfaction considering the extreme admissible allocations (i.e. min_i and max_i) due to the presence of other users.

The mood value is the solution that equalizes the PS rate for each user, assigning to user i a portion of resource equal to $min_i + m(max_i - min_i)$ where m in [0,1] is the ratio between what remains when users get the minimum and $\sum_{i=1}^{n}(max_i - min_i)$. The mood value rule derives from both a classical interpretation of the resource allocation problem such as [18], and an interpretation of the resource allocation problem as a game-theoretic problem [23], [24].

More generically, for those situations that can be realistically modeled as complete information sharing environment (such as 5G ones), where users can be aware of other users' demands and of the available resource (e.g., via northbound and orchestration interfaces), it is possible to model the allocation problem as a bankruptcy game to take such interactions in consideration, and to allocate resources using game-theoretic rules. Examples of such rules are: the *Shapley value* [25], that is a weighted mean of the users' marginal contribution to each possible coalition [26] and the *nucleolus* [25], that is the allocation which lexicographically minimizes the maximum "complaints" [27]. In this respect, the *mood value* is an efficient solution between the minimum right and the maximum right payoff [23], [24].

D. Multi-resource allocation

In the literature, the first work adopting a multi-resource allocation approach for multi-resource environments, going beyond single-resource abstraction, concerns cloud optimization problems in which a central scheduler has to decide the number of simultaneous tasks of multiple types to run, while ensuring fairness [3], [20], [28]. Conceptually, these models can also be applied to the network slicing context; instead of the number of tasks to run, we have a portion of the demand that has to be satisfied for each slice tenant.

A slicing multi-resource allocation problem can be modeled in the following way. Let $N = \{1, ..., n\}$ be the set of tenants and let $M = \{1, ..., m\}$ be the set of available resources. A multi-resource allocation problem can be modeled as a pair (R, D) where $R = (r_1, ..., r_m)$ is a vector of positive numbers, r_j representing the amount of each available resource j in M, and D is the demand matrix with $d_{ij} \in D$ equal to the quantity of resource j demanded by tenant i in N. Difficulties in a multi-resource allocation problem arise if it exists a resource $j \in M$ such that $\sum_{i=1}^{n} d_{ij} > r_j$, i.e., the resource is not enough to satisfy the demands.

Let $x = (x_1, ..., x_n)$, with $0 \leq x_i \leq 1 \ \forall i \in N$, be the vector of the percentage of resources allocated to each tenant. The allocation matrix A corresponding to x is given by $\begin{bmatrix} a_{11} & ... & a_{1m} \\ ... & ... & ... \\ a_{n1} & ... & a_{nm} \end{bmatrix} = \begin{bmatrix} d_{11} \cdot x_1 & ... & d_{1m} \cdot x_1 \\ ... & ... & ... \\ d_{n1} \cdot x_n & ... & d_{nm} \cdot x_n \end{bmatrix}$. The allocation has to belong to the admissible region $\mathcal{F}$ s.t. $\sum_{i \in N} a_{ij} \leq r_j$, $\forall j \in M$. We can notice that each user, receiving the same percentage of each resource, does not receive a surplus of resources. It follows that, with a multi-resource approach, the resource dependency is always respected.

We describe in the following three allocation rules at the state of the art that are those largely adopted in the literature, and that we consider in the rest of the paper. Additional multi-resource allocations can be found in [29].

Dominant Resource Fairness (DRF) [3] rule: is a generalization of the MMF rule. It considers, for each user, the dominant share (i.e., for a user, the maximum among all its resource shares) and the dominant resource (i.e., the resource corresponding to the dominant share), and it applies the MMF across users' dominant shares. The allocation produced by the DRF policy is the solution of the following problem[4]:

$$\begin{aligned} &\text{maximize} \quad x \\ &\text{subject to} \quad ds_i x_i = ds_j x_j, \quad \forall i, j \in N \end{aligned} \tag{2}$$

and $x \in \mathcal{F}$, where $ds_i = \max_j\{\frac{d_{ij}}{r_j}\}$ is the dominant share of user i.

Asset Fairness (AF) [3] rule: aims at equalizing the resource allocated to each users. It is obtained solving:

$$\begin{aligned} &\text{maximize} \quad x \\ &\text{subject to} \quad \sum_{j=1}^{m}(s_j d_{ij})x_i = \sum_{j=1}^{m}(s_j d_{kj})x_k, \forall i, k \in N \end{aligned} \tag{3}$$

and $x \in \mathcal{F}$, where s_j is the worth of the resource j given by $s_j = \frac{r_{max}}{r_j}$, $\forall j \in N$, with r_{max} equal to the value of the greater resource in absolute value.

[4]To maximize a vector means to maximize each component of the vector. Due to the constraints on the available resources and the ones equalizing the resource allocated for the the dominant resource, the problem can be reduced to the maximization of one component of the vector. The maximization of the others then follows.

	Single-resource allocation rules										Multi-resource allocation rules					
Rules	W.Prop (p_i=d_i)		MMF		Shapley		Nucleolus		Mood value		DRF		Asset fairness		Nash product	
Resources	Gbps	CPU	Gbps	CPU	Gbps	CPU	Gbps	CPU	Gbps	CPU	Gbps	CPU	Gbps	CPU	Gbps	CPU
User 1	4.57	0.5	8	0.5	4	0.5	4	0.5	4	0.5	4.48	0.56	4.8	0.6	4	0.5
User 2	11.43	0.5	8	0.5	12	0.5	12	0.5	12	0.5	8.8	0.44	8	0.4	10	0.5
Depletion	Yes		Yes		Yes		Yes		Yes		No		No		No	
Dependency	No		No		No		No		No		Yes		Yes		Yes	

TABLE II: Example of single-resource and multi-resource allocations ($r_1 = 16$, $r_2 = 1$, $d_{11} = 8$, $d_{12} = 1$, $d_{21} = 20$, $d_{22} = 1$)

Nash product [30] rule: is called also Competitive Equilibrium from Equal Income (CEEI) [3]. It is obtained solving:

$$\max \prod_{i \in N} x_i, \qquad \text{s.t.} \quad x \in \mathcal{F} \tag{4}$$

E. Example comparison between allocation rules

Let us consider again an allocation problem with two resources and two users. The first resource is the link bandwidth, with an available resource $r_1 = 16$ Gbps, the second one is the CPU, with an available resource $r_2 = 1$ CPU. User 1 needs a link bit-rate of 8 Gbps and 1 CPU and User 2 needs 20 Gbps and 1 CPU. Note that CPU resources are fractionable with current hypervisors. Table II provides the obtained resource shares applying the allocation rules presented in this section, highlighting quantitatively the fundamental differences already described. Moreover, in the bottom of the table we indicate the two aspects we already discussed: the resource depletion and the inter-resource dependency. As previously discussed, single-resource allocations fully use the addressed resource and, in general, they cannot comply with existing dependency between the resources. Conversely, multi-resource approaches, in general, do not completely use all the resources, but they do take into consideration resource dependency.

III. MURANES
MULTI-RESOURCE ALLOCATION FOR NETWORK SLICING

In order to solve the network slicing resource allocation problem we propose a framework based on an aggregation technique we name MURANES (MUlti-Resource Allocation for NEtwork Slicing). Our objective is to propose a general framework to allocate multiple distinct resources in a fair way. In this direction we consider two factors: an individual satisfaction of the tenants and a system fair utility. The satisfaction is an individual measure and can be combined to obtain a measure of the fairness of the system. To reach our goal we use the aggregation techniques in an original way, proposing a way to summarise the information about the users satisfaction on different resources.

Our proposal is to aggregate users and system utilities as depicted in Figure 3. We consider an utility function $F(y)$ that summarizes the information about users demands and the available resources. To obtain this function one can follow two methodological 'paths':

- first aggregate the users, and then the resources;
- first aggregate the resources, and then the users.

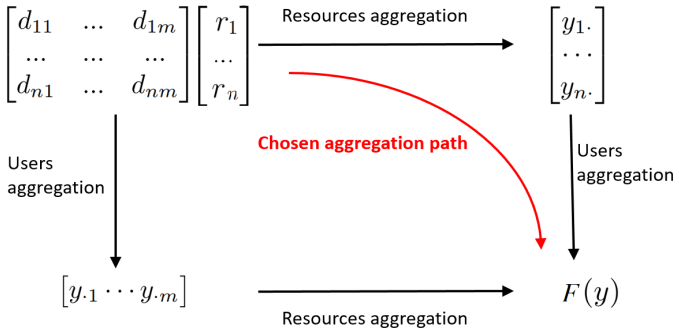


Fig. 3: User and resource aggregation paths. The vector $y_{i\cdot}$ combines m data to provide a single aggregated variable for each user. The vector $y_{\cdot i}$ combines n data to provide a single aggregated variable for each resource. $F(y)$ is the aggregated function to optimize.

In network slicing, an important requirement is to provide a fair allocation matrix, thus it is necessary that the input vector of the function to optimize summarizes the information related to the user satisfaction. For this reason we choose the second path, depicted with a red arrow in the figure, aggregating first the information related to the different resources for each user, i.e., considering the user satisfaction, and secondly aggregating the users, i.e., considering the system efficiency objective. In our case, the first aggregation is done choosing the satisfaction of the user calculated for the most congested resource, and the second aggregation is made using as aggregation function W the Ordered Weighted Averaging (OWA) operator [31] - that we characterize in the following subsection, together with its fairness properties. The OWA operator is able to capture a range of fair attitudes, and allows us to define different allocations following different fairness criteria.

It is worth highlighting that our contribution is the proposal of an original and context-sensitive way to aggregate information related to a multi-resource allocation problem, i.e., such that multiple criteria have to be considered. We define how to leverage on the OWA operator on the satisfaction vector, in order to guarantee fairness in context in which the resources are scarce, we have to care of multiple distinct resources, and different notions of fairness can make sense. No alternative proposals meeting these requirements exist at the current state of the art, as of our knowledge.

A. Ordered Weighted Averaging (OWA) operators

In the following, after explaining the OWA operator and its flexibility in covering different objectives, we discuss how to

properly select OWA input vectors, related to different users satisfaction concepts. The OWA framework we formalize can so incorporate some of the existing multi-resource allocations rules, and permits also to transpose some of the existing single-resource allocation rules to the multi-resource context.

The Ordered Weighted Averagin (OWA) function is introduced in [31] and it is defined as follows.

Definition 1. An *OWA* is a scalarizing function $F : \mathbb{R}^n \to \mathbb{R}$ parametrized by a weighting vector $w \in {\mathbb{R}_+}^n$ of the form $F(v_1, ...v_n) = \sum_{j=1}^{n} w_j v_{(j)}$, where $v_{(j)}$ is the j-th smallest element of $(v_1, ..., v_n)$.

Contrary to the case of the weighted sum, in an OWA function the weights are not used to assign more importance to a component than to another one, but to weight the importance attached to good or bad components in the value aggregation.

The F aggregator encompasses many well known aggregators such as max, min, median and sum, as special cases. It is well known in the Social Choice area, to model an idea of fairness in the social evaluation function. In this case, F is often referred to as the *Generalized Gini social-evaluation Function* [32], [33]. Such a function used with a decreasing weighting vector w, i.e., $w_i \geq w_{i+1}$ for all $i < n$ allows to model a wide range of 'fair' attitudes going from the egalitarianism to the utilitarianism. An *egalitarian* solution is based on the notion of fairness, described in political philosophy by Rawls [34] aiming to protect weaker users, i.e., the less satisfied ones. It is obtained when we maximize the minimum component so when we choose only the first weight w_1 different from zero. An *utilitarian* solution, under the classical utilitarian principle, is obtained when the decision maker maximizes the sum of the utilities of the players. It is obtained choosing the same value for each weight ($w_i = w_j$,$\forall i \neq j$). Changing the OWA weights, choosing decreasing value of the weight, we can obtain trade-off solutions between egalitarianism and utilitarianism.

More precisely, one common way of formally introducing a fairness property in the aggregation is to require that the value of a vector is improved by any *mean preserving transfer reducing inequalities* (a.k.a. Pigou-Dalton transfers). Given a performance vector $v = (v_1, \ldots, v_n)$, any modification of v leading to a vector of the form $(v_1, \ldots, v_i - \varepsilon, \ldots, v_j + \varepsilon, \ldots, v_n)$ for some i, j, ε such that $v_i - v_j > \varepsilon > 0$ should make decision maker better off. Under the Pareto principle – requiring monotonicity in every component – and some other mild requirements such as completeness – requiring this fairness condition – it is possible to define the social utility as an OWA function using a weighting vector w with decreasing components. The described potential of F is illustrated in the following:

Example 1. Consider a simple case with three users. A solution with utility vector (1, 0, .3) is less preferable than (.5, .5, .3) because there exists a transfer (-.5,+.5) between the two first agents to pass from the former solution to the latter. Consistently, we have $F(1, 0, .3) = .3w_2 + w_3$ and $F(.5, .5, .3) = .3w_1 + .5w_2 + .5w_3$ and therefore $F(1, 0, .3) - F(.5, .5, .3) = .3(w_2 - w_1) + .5(w_3 - w_2) \leq 0$ because $w_1 \geq w_2 \geq w_3$. We obtain the desired preference. Now if we compare $(1, 0, .5)$ to $(.3, .3, 5)$ the preference is less clear. In particular, no Pigou-Dalton transfer holds. Moreover, in such a situation, one may want to relax the desire of equity to hold average efficiency. Consistently, we have $F(1, 0, .3) - F(.3, .3, .3) = .7w_3 - .3w_1$ which may be positive or negative depending on w given that $w_1 \geq w_3$. This illustrates the role of vector w that lead to different choices depending on the importance attached to the least satisfied users.

The F function is also widely used in multi-objective optimization to generate solutions with well-balanced utility profiles [35], [36]. $F(v)$ is not linear in v due to the permutation of components, but smart linearization are available, see, e.g., [35].

The MURANES framework we propose is based on the optimization of OWA operators. It is designed for continuous resources, i.e., resources that can be partitioned indefinitely but that – with straightforward model variations – can also be applied to the case of discrete resources, or to the case in which the allocation must be selected from prefixed templates.

B. The general framework

As above introduced, the framework we propose is considering two axes: the system and the individual utility. About the former, subsection III-A shows that the maximization of an OWA function is a good candidate to obtain fair allocations, where fairness goes from the pure egalitarianism to the pure utilitarianism. About the latter, as we anticipated, the input vector of the OWA must depend on the user satisfaction vector, i.e., a vector containing the measure of the satisfaction of each user respect to the m resources. We describe now the four proposed inputs:

- *classical satisfaction*: Classically the satisfaction is measured as the percentage of resource allocated to a user, i.e., as the ratio between the allocated resource and the demanded one. In our model, for each user this ratio is the same for each resource and it is equal to x.
- *weighted classical satisfaction*: We can consider a weighted version of the classical satisfaction. Taking inspiration from the DRF allocation rule the satisfaction of each user i we choose a weight equal to the dominant share (i.e., $ds_i = \max_j\{\frac{d_{ij}}{r_j}\}$) as weight.
- *player satisfaction (ps)*: As already explained in II-C, in case of complete information, the correct way to measure the satisfaction is using the ps rate [23], [24]. If in the case of the classical satisfaction, given a user, the satisfaction coincides for each resource, here we need to find which satisfaction summarizes the information about all the resources. For this purpose, we use the dominant resource for each tenant, because it is the more critical one and, realistically, the one that the tenant would consider to measure its satisfaction.
- *weighted player satisfaction*: in a dual way to the classical satisfaction, we can again consider the dominant share to weight the ps satisfaction.

		System		
		w=$(1,0,\cdots,0)$	$\cdots$	w=$(1,1,\cdots,1)$
Individual	x	$\max\min x_i$	$\cdots$	$\max\sum_{i=1}^{n} x_i$
	$ds \cdot x$	$\max\min ds_i x_i$	$\cdots$	$\max\sum_{i=1}^{n} ds_i x_i$
	ps	$\max\min ps_i$	$\cdots$	$\max\sum_{i=1}^{n} ps_i$
	$ds \cdot ps$	$\max\min ds_i ps_i$	$\cdots$	$\max\sum_{i=1}^{n} ds_i ps_i$

TABLE III: Objective function of the MURANES framework.

The general problem to solve is:

$$\begin{aligned} &\text{maximize} && OWA(v) \\ &\text{subject to} && x \in \mathcal{F} \qquad 0 \leq x_i \leq 1, \forall i \in N \end{aligned} \tag{5}$$

where v can be equal to: (i) the vector x, (ii) the vector $ds \cdot x = \begin{bmatrix} ds_1 \cdot x_1 & \dots & ds_n \cdot x_n \end{bmatrix}$, (iii) the vector ps, with the satisfaction calculated for each user respect to the dominant resource or (vi) the vector $ds \cdot ps = \begin{bmatrix} ds_1 \cdot ps_1 & \dots & ds_n \cdot ps_n \end{bmatrix}$. We summarize in Table III the value objective function in the general framework we propose for two extreme OWA weights configurations. In the following section we clarify on the option to choose among extreme weight configurations.

IV. MURANES PROPERTIES

We describe some important properties of the allocations we obtain using the MURANES framework.

A. Generalization of well known-solutions

The unified framework uses a general class of utility functions that captures different fairness criteria, and between them we can find some already well-known ones. In fact for special combinations of OWA inputs and weights, the allocation coincides with an allocation known in literature. We can state the following theorems.

Theorem 1. Let (D, R) be a multi-resource allocation problem with $d_{ij} \neq 0 \ \forall (i,j) \in N \times M$. The MURANES solution with $w = (1, 0, \cdots, 0)$ and input x coincides with the weighted proportional allocation rule generalized to the multi-resource context.

Proof. The MURANES in case in which $w = (1, 0, \cdots, 0)$ and the input is x coincides with the solution of:

$$\begin{aligned} &\text{maximize} && min(x) \\ &\text{subject to} && x \in \mathcal{F} \\ & && 0 \leq x_i \leq 1, \forall i \in N \end{aligned} \tag{6}$$

but (6) coincides with:

$$\begin{aligned} &\text{maximize} && x \\ &\text{subject to} && x \in \mathcal{F} \\ & && x_i = x_j, \forall i, j \in N \\ & && 0 \leq x_i \leq 1, \forall i \in N \end{aligned} \tag{7}$$

In fact, the constraints of (7) imply that the optimal solution is the Pareto efficient solution that belongs to the line produced by the constraints $x_i = x_j$, $\forall i, j \in N$. This follows from the fact that all the other Pareto efficient solutions are such that the variable with the minimum value can be increased. The constraint $x_i = x_j$ implies that the satisfaction of each user is equal and this property characterizes, in the case of single resource allocations, the weighted proportional allocation when we choose the weights equal to the user demand (see [23], [24]). □

Theorem 2. Let (D, R) be a multi-resource allocation problem with $d_{ij} \neq 0 \ \forall (i,j) \in N \times M$. The MURANES solution with $w = (1, 0, \cdots, 0)$ and input $ds \cdot x$ coincides with the DRF allocation rule.

Proof. Similarly to the proof of Theorem 1, the considered optimization problem can be rewritten as:

$$\begin{aligned} &\text{maximize} && x \\ &\text{subject to} && x \in \mathcal{F} \\ & && ds_i \cdot x_i = ds_j \cdot x_j, \forall i, j \in N \\ & && 0 \leq x_i \leq 1, \forall i \in N \end{aligned} \tag{8}$$

that is exactly the DRF allocation rule described in (2). □

Theorem 3. Let (D, R) be a multi-resource allocation problem with $d_{ij} \neq 0 \ \forall (i,j) \in N \times M$. The MURANES solution with $w = (1, 0, \cdots, 0)$ and input ps coincides with the mood value generalized to the multi-resource context.

Proof. Again, similarly to the proof of Theorem 1, the considered optimization problem can be rewritten as:

$$\begin{aligned} &\text{maximize} && ps \\ &\text{subject to} && x \in \mathcal{F} \\ & && ps_i = ps_j, \forall i, j \in N \\ & && 0 \leq x_i \leq 1, \forall i \in N \end{aligned} \tag{9}$$

So, the single resource allocation that equalizes the user satisfaction calculated using the PS is the mood value (see [23], [24]). □

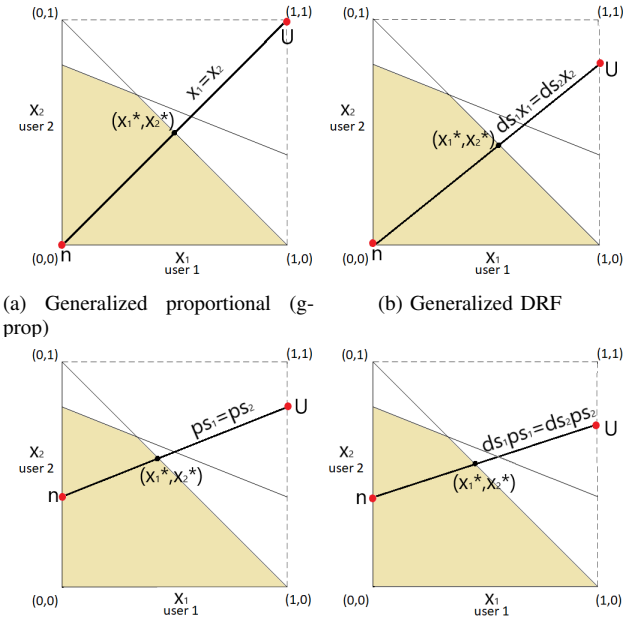
To sum up, the previous theorems show that MURANES allows to capture and generalize classical allocation rules. For the following, let us assign a name to the corresponding allocation rules obtained as a function of the OWA input:

- generalized weighted proportional allocation (g-prop) when the input is x,
- generalized DRF allocation (g-drf) when the input is $ds \cdot x$,
- generalized mood value (g-mood) when the input is ps,
- moodified DRF (gm-drf) when the input is $ds \cdot ps$[5].

B. Game theoretic interpretation

Let us compare the four allocations rules defined in the previous section using the corresponding individual satisfaction vectors and the OWA weight vector $(1, 0, \cdots, 0)$. Fig. 4 shows on the tenants' satisfaction plane the region of the admissible

[5]The word 'moodified' comes from the fusion of the word 'mood' and 'modified', justified by the fact that the allocation considers the satisfaction rate typical of the mood value allocation but also the dominant share typical of the DRF allocation.



(a) Generalized proportional (g-prop)

(b) Generalized DRF

(c) Generalized mood value (g-mood)

(d) Moodified DRF (gm-drf)

Fig. 4: Allocations with $w = (1, 0, \cdots, 0)$. (x_1^*, x_2^*) is the solution of the allocation problem.

solutions and the four allocation rules when we consider an allocation problem with two resources and two users.

We can notice that the solution is the intersection between a line and the Pareto efficient frontier. The lines are:

- $x_1 = x_2$ for the g-prop allocation,
- $ds_1 x_1 = ds_2 x_2$ for the g-DRF allocation,
- $ps_1 = ps_2$ for the g-mood allocation,
- $ds_1 ps_1 = ds_2 ps_2$ for the gm-drf allocation,

These can be interpreted as solutions of the bargaining game between two users. A bargaining game [37], [38] is a pair (C, n) where C is a bounded closed and convex set and n the utility when the two users are not able to reach an agreement. The egalitarian solution can be interpreted as the Kalai-Smorodinski solution [38], that is on the Pareto frontier obtained joining the nadir and the utopia point. The nadir point n is $(0, 0)$ for the first two allocation rules (Fig. 4a,4b), while with respect to the two solutions obtained changing the satisfaction measure (Fig. 4c,4d) the nadir point gives the minimal right for each user. Each component of the utopia (U) point is obtained maximizing the utility of each user. It follows that for the two cases, with the classical satisfaction, the utopia point is $U = (1, 1)$, $U = (\frac{1}{ds_1}, \frac{1}{ds_1})$ (resp.) while for the two other cases it is enough to calculate users' maximal right.

C. Egalitarian and utilitarian fairness trade-off

Let us elaborate on the potential of the unified framework. As anticipated in Section III-A, the two extreme behaviors in terms of fairness are the egalitarian and the utilitarian ones. The utilitarian approach aims at maximizing the total utility of the members of a society without paying attention to social inequality; it is in fact also sometimes referred as *system efficiency* [39]. The egalitarian approach aims at maximizing the individual utility while promoting equitable distributions of utility; for this reason it is commonly used for fair optimization [39]. In most cases, the objective of reducing inequalities comes at a cost that can be measured by the *Price of Fairness (POF)* that is defined as follows.

Definition 2. The *Price of Fairness* (POF) is:

$$POF = \frac{f(x_f^*) - f(x_{min}^*)}{f(x_f^*)} \tag{10}$$

where $f(x) = \sum_{i=1}^{n} x_i$ is the utilitarian criterion, x_f^* is the solution obtained maximizing f and the x_{min}^* is the egalitarian optimum.

Example 2. Let us consider a resource allocation problem with $D = \begin{bmatrix} 12 & 1 & 5 \\ 10 & 2 & 15 \\ 5 & 3 & 10 \\ 10 & 1 & 15 \end{bmatrix}$ and $R = [20, 4, 20]$.In this case the utilitarian optimum is $x_f^* = (0.94, 0, 0.88, 0.44)$ whereas the egalitarian optimum is $x_{min}^* = (0.44, 0.44, 0.44, 0.44)$. Hence we obtain $POF = 0.21$. This value measures the normalized gap to optimal efficiency induced by the fairness requirement.

In the above example the POF is moderate, which shows that perfect equity can be reached at reasonable cost regarding efficiency. This is not always the case and, in many situations, it can be interesting to determine solutions achieving a better compromise between pure utilitarianism and pure egalitarianism. This is precisely the interest of resorting to an OWA optimization that enables to generate various compromise solutions depending on the OWA weights. Let us come back to Example 2.

Example 2. cont. A third solution of the problem obtained using OWA with the weighting vector $w = (0.34, 0.29, 0.23, 0.14)$ is $x_w^* = (0.92, 0.26, 0.77, 0, 26)$. We can notice that:

- $\min(x_{min}^*) \geq \min(x_w^*) \geq \min(x_f^*)$,
- $\sum_{i=1}^{4} x_{f_i}^* \geq \sum_{i=1}^{4} x_{w_i}^* \geq \sum_{i=1}^{4} x_{min_i}^*$,

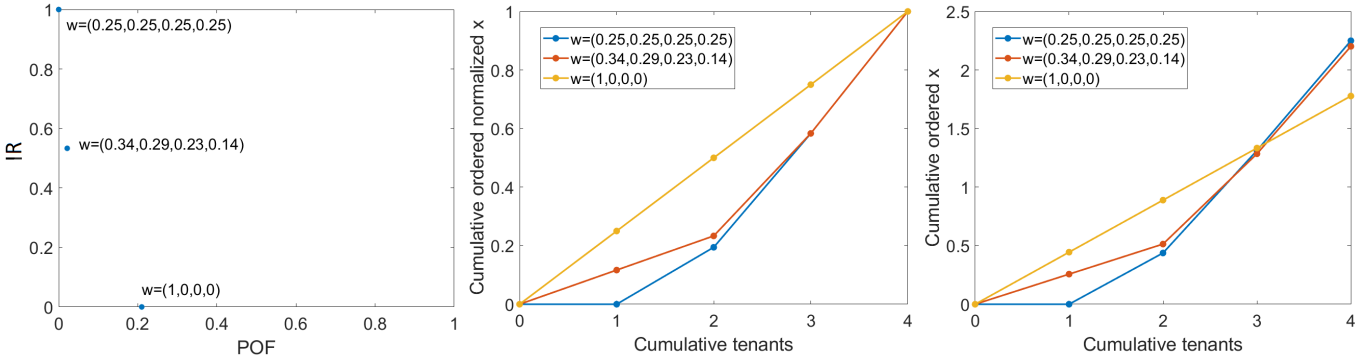
We give now a finer description of how inequalities and POF may vary when playing with OWA weights. For this purpose, we introduce the two following measures:

- $POF(x_w^*) = \frac{f(x_f^*) - f(x_w^*)}{f(x_f^*)}$ where $f(x) = \sum_{i=1}^{n} x_i$, x_f^* is the solution obtained in the utilitarian case and x_w^* is the solution maximizing an OWA with weight w.
- $IR(x_w^*) = 1 - \frac{\min(x_w^*)}{(1/n) f(x_w^*)}$, where $f(x) = \sum_{i=1}^{n} x_i$ and x_w^* maximizes an OWA with weight w [40].

The first measure generalizes the one described in [39], that measures the loss of total utility faced by users in order to guarantee the fairness associated to weights vector w. The second index measures the inequality rate between the utility of the tenants. Both the indices have values in the closed interval $[0, 1]$, as stated in Theorem 4.

Theorem 4. Given an allocation x_w^*, $POF(x_w^*)$ and $IR(x_w^*)$ take value in $[0, 1]$.

Proof. Given an allocation x_w^* it always hold that $x_w^* \leq x_f^*$, because x_f^* is the solution that maximizes. It follows that



(a) POF and IR indices (b) Lorenz curve with normalized solution (c) Lorenz curve with non-normalized solution

Fig. 5: Lorentz curves, POF and I indices.

$f(x_f^*) - f(x_w^*) \leq f(x_f^*)$ and consequently $POF(x_w^*) \leq 1$. In the case in which $x_w^* = x_f^*$ the index gets the minimal value equal to 0. For the $IR(x_w^*)$ the maximal possible value is 1, because the minimal possible allocation for a user is zero (i.e. the portion of allocated resource is zero), while the maximal possible value is 1 and it is obtained when there is an equal distribution of the share ($(1/n)f(x_w^*) = \min(x_w^*)$). For any other allocation it holds $(1/n)f(x_w^*) \geq \min(x_w^*)$ and consequently $IR(x_w^*) \leq 1$. □

According to the POF measure, the utilitarian solution gets value 0 and the price increases when we consider other solutions closer to the egalitarian one.

Differently, the IR index has value 0 for the egalitarian solution and its value increases for the other solutions. Due to the opposite behavior of the indices, a good trade-off between egalitarian and utilitarian criteria can be found in those solutions providing allocations vectors with indices POS and IR close to 0. Looking at the example depicted in Fig 5a we can see that the egalitarian solution has good properties in terms of equity but the POF has a value of around 0.2. If we are not willing to pay that price of fairness we can select the intermediate solution with a negligible price of fairness but we loose something in terms of fairness.

Another way to compare the various possible solutions is based on *Lorenz curves* [41]. A Lorenz curve is obtained plotting the cumulative x when we order the users from the less satisfied one to the most satisfied one. We plot in Fig. 5b and 5c the Lorenz curves for the resource allocation problem of Example 2 when we consider the normalized and non-normalized vector x, selecting three solutions of a resource allocation problem obtained using an egalitarian approach ($w = (1, 0, 0, 0)$), an utilitarian approach ($w = (0.25, 0.25, 0.25, 0.25)$)[6] and an intermediate one ($w = (0.34, 0.29, 0.23, 0.14)$). In Fig. 5b the straight line represents the perfect equality in the distribution of the satisfaction between tenants and the most distant the curves are, the greater the inequality is. It is clear that the egalitarian solution, that aims to equalize the satisfaction of the users, provides a straight line, while the utilitarian solution provides a more unfair allocation. Contrarily, checking figure 5c we can notice that the sum of the users satisfaction are maximized with the utilitarian solution ($\sum_{i=1}^{4} x_i = 2.2500$) and it has the lower value for the egalitarian solution ($\sum_{i=1}^{4} x_i = 1.78$). Looking both the criteria (max-min and max-sum), the third considered solution shows an intermediate behavior representing the trade-off between utilitarian and egalitarian solutions.

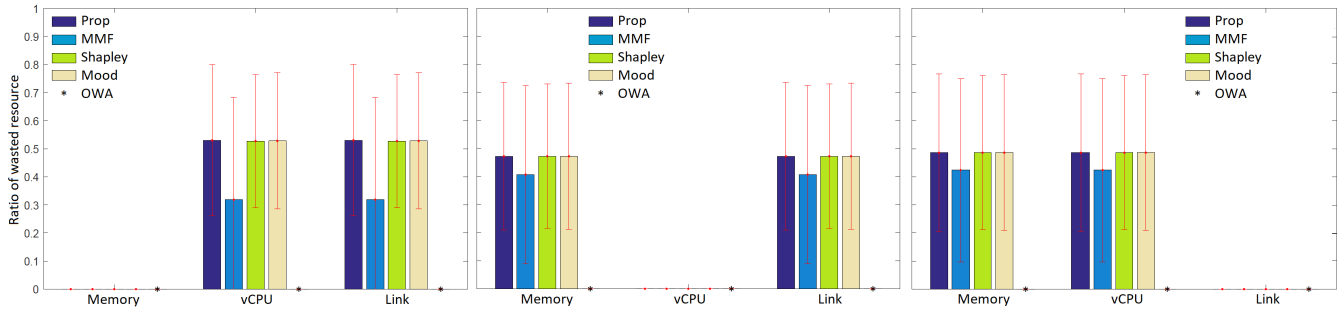
Finally, it is worth to mention further properties that can be considered from a fairness point of view and can be used by the decision-maker to select the weights to use. For example one can be interested to (i) *strategy-proof* allocation where users should not be able to benefit by lying about their resource demands or to (i) *envy-freeness* allocation where a user should not prefer the allocation of another user. The DRF allocation, for example, satisfies these properties [3]. On the other side one can be interested into allocations equalizing the users satisfaction rate. In this case the DRF allocation is no more suitable and the g-prop and g-mood with weight $w = (1, 0, \cdots, 0)$ can be preferable.

Each allocation obtained with the MURANES framework, varying the weights vector, does not satisfy all the fairness properties we can consider at once. This gives more value to a general framework that can be better adapted to a specific context. The only properties satisfied by all MURANES allocation rules are the *Pareto efficiency* that state that it is not possible to increase the allocation of a user without decreasing the allocation of at least another user, and the Pigou-Dalton transfer already described in Section III [42].

V. Numerical Evaluation

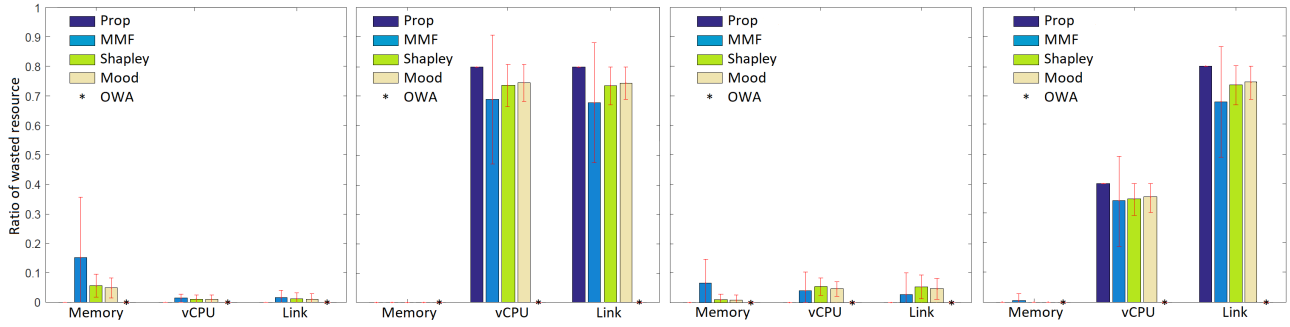
We test both the single-resource and the presented multi-resource allocation rules, in a realistic scenario. We simulate 100 resource allocation problems with 3 resources (Memory, vCPU and link capacity) and 10 slices. We randomly generate the slice demands from the 23 templates described in Table IV, a subset of Amazon EC2 instances [15] we could extract (by simply copying the rows having a complete information about the three considered resources). In practice, in 5G slicing we can expect quite similar resource quantities and relations, with the link bit-rate at a lower scale as of preliminary specifications of some slices (e.g., the eMBB one) and related scenarios found in [16]. Different scales do not matter, the important aspect being the relation between resources.

[6] $w = (0.25, 0.25, 0.25, 0.25)$ is the weight $w = (1, 1, 1, 1)$ normalized.



(a) Memory congested (b) vCPU congested (c) Link congested

Fig. 6: Wasted resource ratios (1 congested resource). Multi-resource rules are referred as 'OWA'.



(a) Congestion level (0.9, 0.9, 0.9) (b) Congestion level (0.9, 0.1, 0.1) (c) Congestion level (0.1, 0.1, 0.1) (d) Congestion level (0.9, 0.5, 0.1)

Fig. 7: Wasted resource ratios (3 congested resources). Congestion level: (Memory, vCPU, Link).

API Name	Memory (GB)	vCPUs	Gbps	Instance Type
m4.10xlarge	160.00	40.00	10.00	General purpose
m4.16xlarge	256.00	64.00	25.00	General purpose
c5.9xlarge	72.00	36.00	10.00	Compute optimized
c5.18xlarge	144.00	72.00	25.00	Compute optimized
c4.8xlarge	60.00	36.00	10.00	Compute optimized
r4.8xlarge	244.00	32.00	10.00	Memory optimized
r4.16xlarge	488.00	64.00	25.00	Memory optimized
x1.16xlarge	976.00	64.00	10.00	Memory optimized
x1.32xlarge	1952.00	128.00	25.00	Memory optimized
x1e.16xlarge	1952.00	64.00	10.00	Memory optimized
x1e.32xlarge	3904.00	128.00	25.00	Memory optimized
p3.8xlarge	244.00	32.00	10.00	Accelerated comput.
p3.16xlarge	488.00	64.00	25.00	Accelerated comput.
p2.8xlarge	488.00	32.00	10.00	Accelerated comput.
p2.16xlarge	732.00	64.00	25.00	Accelerated comput.
g3.8xlarge	244.00	32.00	10.00	Accelerated comput.
g3.16xlarge	488.00	64.00	25.00	Accelerated comput.
f1.16xlarge	976.00	64.00	25.00	Accelerated comput.
h1.8xlarge	128.00	32.00	10.00	Storage optimized
h1.16xlarge	256.00	64.00	25.00	Storage optimized
d2.8xlarge	244.00	36.00	10.00	Storage optimized
i3.8xlarge	244.00	32.00	10.00	Storage optimized
i3.16xlarge	488.00	64.00	25.00	Storage optimized

TABLE IV: Amazon EC2 instances

In the first scenario we analyze, only 1 resource at time is congested. We randomly generate the amounts in this way:

- for the congested resource, the available amount has a value bigger than the minimum demand and lower than the sum of the demands;
- for the non-congested resource, it is between the sum of the demands and two times the sum of the demands.

In the second scenario, all the resources are congested but not always at the same level of congestion. The level of congestion considered is the fraction of the global demand (sum of all demands) that cannot be allocated due to resource scarcity; e.g., if the level is 0.9, 90% of the sum of the demands is not satisfied, thus we are in a strong congestion situation. In the simulations, we consider the following four cases of congestion level combinations:

- 0.9, 0.9, 0.9: 3 resources have the same high congestion;
- 0.1, 0.1, 0.1: 3 resources have the same low congestion;
- 0.9, 0.1, 0.1: 1 resource has high and 2 have low congestion;
- 0.9, 0.5, 0.1: the 1st resource has a high congestion, the 2nd one a medium level and the 3rd one a low level.

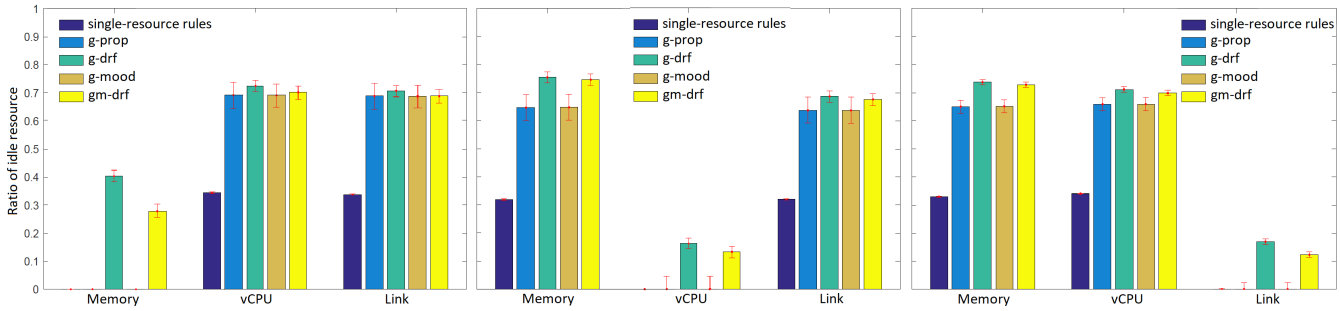
The first two cases show a homogeneous congestion distribution, while the latter two have a heterogeneous distribution that likely corresponds to a more realistic setting.

We test the presented single-resource allocations (weighted proportional with $p_i = d_i$, MMF, Shapley value, Mood value)[7], and the proposed MURANES rules with OWA weights $w = (1, 0, \cdots, 0)$ because we are interested in evaluating the performance of the already known solution (i.e. DRF) compared to the new proposed one, that generalizes single resource allocation (g-prop, g-mood), or that are not known (gm-drf).

A. *Results in terms of wasted and idle resource*

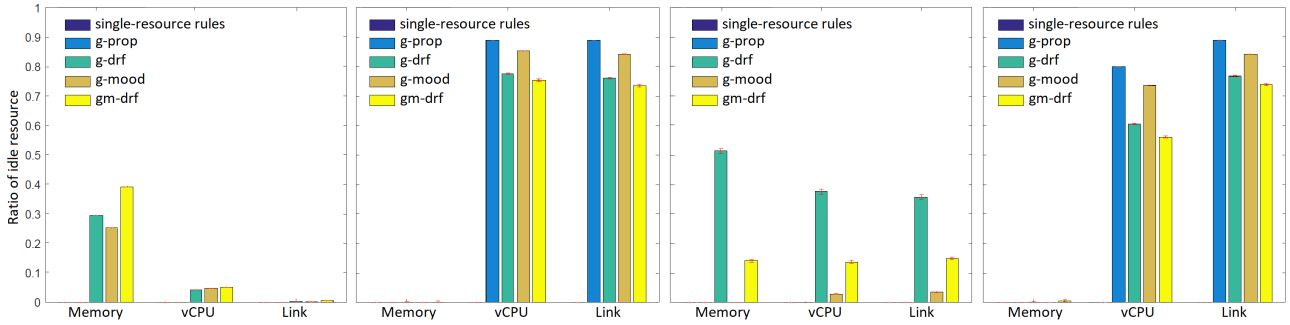
Fig. 6 shows the average ratio of wasted resource in the case in which only one resource is congested. Fig. 7 shows the same, but when all the resources are congested. We can notice, as we already discussed, that single-resource allocations produce resource wasting, i.e., even if a resource is allocated, it may not be fully needed due to the assumed relation between

[7]except the Nucleolus, whose computation has a high time complexity



(a) Memory congested (b) vCPU congested (c) Link congested

Fig. 8: Idle resource ratios (1 congested resource).

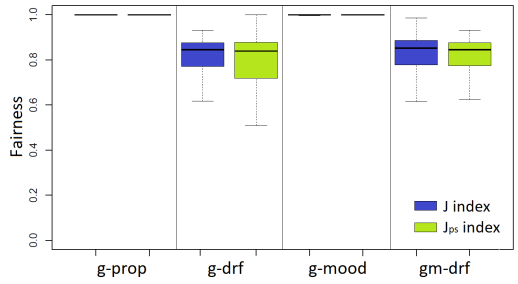


(a) Congestion level (0.9, 0.9, 0.9) (b) Congestion level (0.9, 0.1, 0.1) (c) Congestion level (0.1, 0.1, 0.1) (d) Congestion level (0.9, 0.5, 0.1)

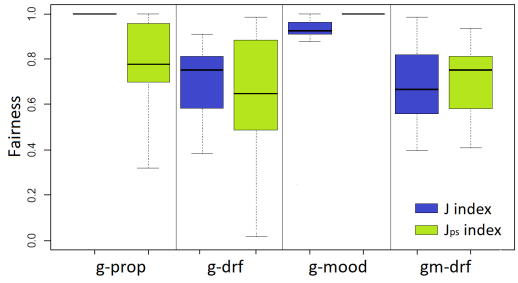
Fig. 9: Idle resource ratios (3 congested resources). Congestion level: (Memory, vCPU, Link)

resources. For single-resource allocations, the trend in terms of wasted resource depends on the congestion level: if the resource is congested it is fully allocated, and consequently the wasted resource is zero; in case of equal congestion level between the resources (Fig. 7a, 7c), there is a similar ratio of wasted resource between the three resources; in the case in which the level of congestion is heterogeneous, the ratio of waste resource is zero for the most congested resource, and it increases decreasing the congestion level. Multi-resource rules, respecting inter-resource dependency, do not produce wasted resource, in each congestion level configuration. This means that there are no resources allocated and unused by the users because multi-resource rules allocate for each user the same percentage of demand for each resource.

In a dual way, Fig. 8 and Fig. 9 show respectively the average ratio of idle resource in the cases in which only one resource is congested, and when all the resources are congested. We can notice that single resource allocation rules produce idle resource only if the resource is non-congested; for this resource, tenants receive exactly what they ask and consequently, being the resource non-congested, an idle resource is produced. For multi-resource allocations, there is a similarity between the two allocations that consider the satisfaction rate (g-prop and g-mood), and between the two allocations that weight the satisfaction rate with the dominant share (g-drf, gm-drf). The first couple of allocation rules produce less idle resource when (i) only one resource is congested or (ii) the congestion level is homogeneous. The second one, adapting the satisfaction to the the resources available in the network in which the slice is situated, produces less idle resource when the congestion level is heterogeneous.



(a) 1 resource congested



(b) 3 congested resources

Fig. 10: Fairness index with different allocation rules.

B. *Results in terms of fairness*

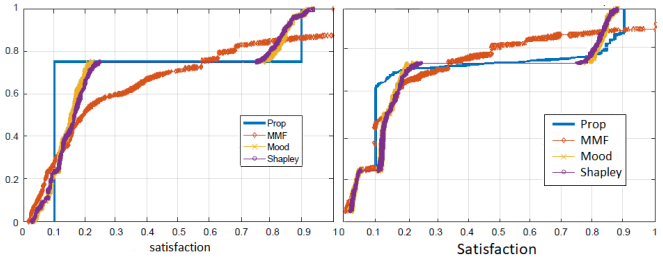
In order to analyze the fairness of the allocation rules, we analyze the Jain's index of fairness [22] and its modification considering the PS rate instead of the classical Demand Fraction Satisfaction (DFS) rate [23], [24]. Fig. 10 shows the boxplot results of the fairness index for the dominant resource, and for the two congestion cases. We can notice that the two solutions with better performances in terms of fairness are g-prop and g-mood, i.e., the ones considering as OWA input the DFS and PS rates. This follows from the fact that the two allocations equalize the tenant satisfaction and consequently

maximize the respective index of fairness. Considering the dominant resource for each tenant, the satisfaction is no more the same for each tenant thus the fairness decreases, but on average not excessively.

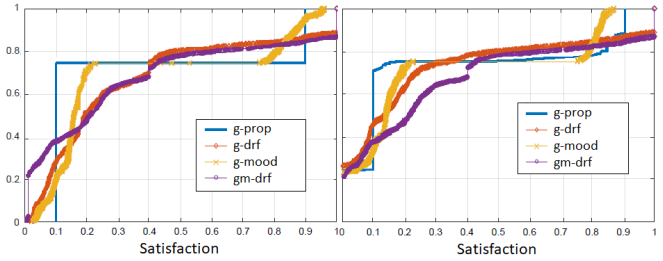
Fig. 11 and 12 show, for the two satisfaction rate definitions (classical and PS) and for both single- and multi-resource allocation rules, the cumulative distribution function (CDF) of the minimum satisfaction rate, i.e., among the three resource-specific satisfaction rates, the least one. In this way we can focus on the minimum satisfaction rate as a desirable fitness metric to increase. Fig. 11 refers to the 3-congested resources case, while Fig. 12 to only the heterogeneous cases, i.e., $(0.9, 0.1, 0.1)$ and $(0.9, 0.5, 0.1)$. Again we can notice a similarity between g-prop and g-mood (with OWA input equal to x and ps) from the one hand, and g-drf and gm-drf (with $ds{\cdot}x$ and $ds{\cdot}ps$) from the other hand. We see that the minimum satisfaction is clearly linked to the congestion level. In Fig. 11 we have 3 cases over 4 with a level of congestion equal to 0.9 for at least one resource; it follows that the least satisfaction is the one related to the most congested resource, getting a value (with the classical DFS rate) exactly equal to 0.1 for the proportional and the generalized weighted proportional rules. About 50% of the tenants suffer from a very low satisfaction (between 0 and 0.15).

Therefore, we compare the global (i.e., with both heterogeneous and homogeneous congestion cases – Fig. 11) results to the one with only heterogenous congestion cases (Fig.12). In the former the satisfaction rate CDFs for single and multi-resource allocations are similar: MMF, gm-drf and g-drf assign the highest satisfaction rate to about 10% of tenants, and are hence preferable. This follows from the the fact that g-drf and gm-drf can be considered as generalizations of the MMF allocation. With the heterogeneous cases apart (Fig.12), instead, gm-drf is superior to all the other allocation rules (single- and multi-resource ones), except for MMF with the classical satisfaction rate (Fig.12a) which however is known to offer low fairness [23], [24].

These results show that in realistic settings with heterogeneous resource congestion, the MURANES rules we propose, and in particular the m-drf, g-prop and g-mood rules, clearly outperform the application of single-resource allocation rules.

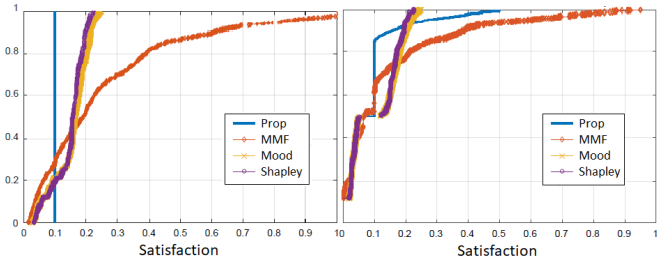


(a) Classical satisfaction rate (single) (b) PS satisfaction rate (single)

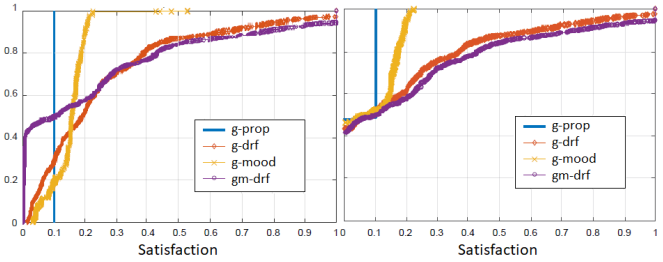


(c) Classical satisfaction rate (multi) (d) PS satisfaction rate (multi)

Fig. 11: Minimum satisfaction rates CDF (3 congested resources).



(a) Classical satisfaction rate (single) (b) PS satisfaction rate (single)



(c) Classical satisfaction rate (multi) (d) PS satisfaction rate (multi)

Fig. 12: Minimum satisfaction rates CDF (3 resources congested - heterogeneous congestion levels)

VI. Refinement of the Model

In this section we provide possible generalizations of the MURANES framework to deal with practical aspects rising when applying it to specific environments. First, we show how to go beyond the hypothesis of linear dependency between resources; if for most of the resource pairs we can realistically model the relationship between resources (as elaborated in Section II-B), for other resources (such as those depending on particular radio schedulers) such an assumption may be too strong. Second, we describe how Service Level Agreements (SLA) constraints can be added to the problem to deal with priorities among slices and among users.

A. Generic resource dependency

In practice, the analytical relationship between the resource can be known a priori, for example as a results of preliminary analysis of the mutual interference or dependency among pairs of resources. If the relationship between the resources is expressed by a strictly increasing monotonic function and the objective is to provide fair allocations, the allocation problem can still be solusing an OWA approach, but the resource allocation problem would have to be refined for such a case. More precisely, theelationship can no longer be included in the multi-resource allocation settings, but has to be added as a constraint. In particular x is no more a vector but a matrix $n \times m$, whose components x_{ij}, with $0 \leq x_{ij} \leq 1 \; \forall i \in N$, is the percentage of resources j allocated to tenant i. The allocation matrix A corresponding to x is given by $D = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nm} \end{bmatrix} = \begin{bmatrix} d_{11} \cdot x_{11} & \dots & d_{1m} \cdot x_{1m} \\ \dots & \dots & \dots \\ d_{n1} \cdot x_{n1} & \dots & d_{nm} \cdot x_{nm} \end{bmatrix}$. The constraints to add to classical problem (5) are of type

$x_{ik} = f_k(x_{is})$, $\forall i \in N$, $\forall k \neq s$.

The following example illustrates the relaxation of the linear dependency hypothesis.

Example 3. Let us consider two resources, A ($j = 1$) and B ($j = 2$), such that the dependence between A and B is quadratic for each user $i \in N$. Let the matrix demand $D = \begin{bmatrix} 6 & 4 \\ 9 & 3 \end{bmatrix}$ and the resource vector $R = \begin{bmatrix} 10 & 5 \end{bmatrix}$. The problem to solve is:

$$\begin{aligned} \text{maximize} \quad & OWA(v) \\ \text{subject to} \quad & 6x_{11} + 9x_{21} \leq 10, \\ & 4x_{12} + 3x_{22} \leq 5, \\ & x_{i1} = x_{i2}^2, i = 1, 2 \\ & 0 \leq x_{ij} \leq 1, \forall i \in N \end{aligned} \tag{11}$$

where v is one of the OWA input described in section III.

It is worth noting that the addition of the non-linear constraints are expected to increase the computational complexity of the optimization problem, which in fact can no longer be solved by using conventional linear-programming algorithms such as the simplex and the interior point algorithm, because the solution space can no longer be explored using real cuts.

More generally we can suppose there is no relationship between resources. In this case one way can be to considerate each resource separately but to guarantee a global fairness we need to introduce a multidimensional inequality measure. Our indication in this other possible direction is to resort to the Multidimensional Generalized Gini Index [43] that is a sum over the resources of inequality indices defined as instances of OWA for every resources.

B. Guaranteeing a minimum resource amount

The forthcoming services to be delivered by the 5G are categorized in three macro classes, depending on the latency, frequency, bandwidth and reliability requirements: enhanced Mobile BroadBand (eMBB), Ultra Reliable Low Latency Communications (URLLC) and massive Machine Type Communications (mMTC). Therefore the resource allocation has to take into account related Service Level Agreements (SLA), i.e., contracts between the tenants and the service provider. An SLA can specify (i) the minimum capacity guaranteed and a nominal one for a given resource, (ii) the amount of time the service is guaranteed, (iii) penalties in case the service requirements are not met, (iv) the service assistance, etc. [44]. In addition, an SLA can also indicate guarantees on metrics that are not the result of a resource allocation, such as latency or jitter.

The common SLAs appearing in network slicing specifications include a minimum resource amount or capacity and to meet this requirement we need to enrich the classical centralized multi-resource model. In order to guarantee a minimum resource amount to clients for a given resource, we can associate to each tenant i, at each instant of time, two types of demand vector:

- d_i^m, that is the minimum demand, for each resource, i.e., a fixed value offered by the service provider to tenants;
- d_i, that is the demand of the tenant at a given time (i.e., the scheduling time slot).

Consequently, at each time slot, our slicing multi-resource allocation problem is set as a 3-tuple (R, D, D^m) where R is the vector of the available resources, D is the demand matrix and D^m is the minimum demand matrix.

To satisfy the minimal requirements established by an SLA, at each time slot, we modify the capacity constraints in the MURANES setting, in such a way that the minimum demand is allocated to each tenant. It follows that the problem to solve becomes:

$$\begin{aligned} \text{maximize} \quad & OWA(v) \\ \text{subject to} \quad & x \in \mathcal{F}, \\ & x_i^m \leq x_i \leq 1, \forall i \in N \end{aligned} \tag{12}$$

where v is one of the OWA input described in section III and $x_i^m = max_j(x_{ij}^m) = max_j\left(\frac{d_{ij}^m}{d_{ij}}\right)$.

If the problem has no solution, i.e., there are no enough resources to satisfy the minimum request of the tenants, a policy to delay users in the scheduling queue has to be designed. It can depend for example on the type of service, eliminating firstly tenants with lower priority then the others, and on the availability rate, by looking at the allocation history of each tenants to guarantee a time fairness in the decision. We provide further details on SLA modeling in [45].

VII. Summary

In this paper we explored in depth the problem of resource allocation in network slicing where multiple resources have to be allocated to verticals and shared concurrently. Our contribution is the formalization of the problem, under the important assumptions that not the entire amount of requested resources can be assigned to tenants, and that guaranteeing a relationship between allocated slice resources is important for an efficient operation of related services.

We propose a multi-resource allocation framework, called MURANES, based on the Ordered Weighed Average (OWA) operator to generalize the most known single-resource and multi-resource allocation rules and define new ones. We provide a complete analysis of the proposed framework and we show how it lets to the decision-making the freedom to select the most appropriate allocation, based on the type of fairness goal it is meant to follow. Through extensive simulations we characterize the behavior of the allocation rules in terms of fairness and in terms of wasted resource. As opposed to single-resource allocation rules, multi-resource allocation rules (i) have the key advantage of not allocating unneeded surplus of resources, (ii) can allow for idle capacity to support traffic peaks, and (iii) are superior in terms of satisfaction rate in case of heterogeneous congestion (i.e., not all resources are equally congested) – which happens for the generalized DRF and moodified DRF. Among multi-resource allocation rules, we could highlight that the fairest ones are the proposed OWA generalization of the weighted proportional allocation and of the mood value.

We conclude the paper with possible extensions of the MURANES framework to deal with the case in which the relationship between resources is not linear, and with the case considering Service Level Agreement (SLA) constraints. Further work could also investigate how profit maximization

can coexist with the classical requirement of fair allocations, particular important with shared infrastructures such as those envisioned with the 5G.

ACKNOWLEDGEMENTS

This work was partially funded by the MAESTRO-5G (Management of Slices in the Radio Access of 5G Networks) funded by ANR (Agence Nationale de la Recherche), contract nb. ANR-18-CE25-0012 (https://maestro5g.roc.cnam.fr). The authors would like to thank the anonymous reviewers for their extremely useful feedback.

REFERENCES

[1] 5G Americas, "Network Slicing for 5G and Beyond." *White Paper*, 2016.
[2] NGMN, "5G white paper." *Next generation mobile networks*, 2014.
[3] A. Ghodsi, et al., "Dominant resource fairness: fair allocation of multiple resource types." *Proc. of USENIX NSDI 2011*.
[4] P. Caballero, et al.,"Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads." *IEEE/ACM Transactions on Networking (TON)*, 25.5: 3044-3058, 2017.
[5] M. Leconte, et al, "A resource allocation framework for network slicing." *IEEE INFOCOM 2018*, 2018.
[6] W. Guan, et al., "A service-oriented deployment policy of end-to-end network slicing based on complex network theory." *IEEE Access 6*, 2018: 19691-19701.
[7] G. Wang, et al., "Resource Allocation for Network Slices in 5G with Network Resource." *IEEE GLOBECOM 2017*, 2017
[8] M. Jiang, M. Condoluci, T. Mahmoodi, "Network slicing in 5G: An auction-based model." *IEEE ICC 2017*, 2017.
[9] P. Caballero, et al., "Network slicing games: Enabling customization in multi-tenant networks." *IEEE/ACM Transactions on Networking*, 2019.
[10] Y. Xiao, et al., "Distributed Resource Allocation for Network Slicing Over Licensed and Unlicensed Bands." *IEEE Journal on Selected Areas in Communications 36.10* , 2018: 2260-2274.
[11] H. Halabian, "Distributed Resource Allocation Optimization in 5G Virtualized Networks." *IEEE Journal on Selected Areas in Communications 37.3*, 2019: 627-642.
[12] S. Lee, et al., " Resource Management in Service Chaining." *IETF Secretariat, Intert-Draft*, 2016.
[13] Y.Etsion, D. Tsafrir, and D. G. Feitelson, "Process prioritization using output production: scheduling for multimedia." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 2.4*, 2006: 318-342.
[14] Intel, "Impact of the Intel Data Plane Development Kit (Intel DPDK) on packet throughput in virtualized network element." *White Paper*, 2013.
[15] Amazon EC2 instances comparison: https://www.ec2instances.info.
[16] Alliance NGMN, "Recommendations for NGMN KPIs and Requirements for 5G." *Technical Report*, 2016.
[17] W. Thomson. "Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: an update." *Math. Soc. Sciences* 74:41-59, 2015.
[18] FP. Kelly, AK. Maulloo, DKH Tan. "Rate control for communication networks: shadow prices, proportional fairness and stability." *J. of the Operational Research society* 49.3, 1998.
[19] DP. Bertsekas, RG. Gallager, P. Humblet. *Data networks.* Vol. 2. New Jersey: Prentice-Hall International, 1992.
[20] O. Wlodzimierz, et al., "Fair optimization and networks: A survey." *J. of Applied Mathematics 2014*, 2014.
[21] J. Mo, J. Walrand. "Fair end-to-end window-based congestion control." *IEEE/ACM Trans. on Networking (ToN)*, 2000.
[22] R. Jain, DM. Chiu, WR. Hawe. *A quantitative measure of fairness and discrimination for resource allocation in shared computer system.* Vol. 38. Hudson, MA: East. Res. Lab., Digital Equipment Corporation, 1984.
[23] F. Fossati, S. Moretti, S. Secci. "A Mood Value for Fair Resource Allocations." *IFIP Networking 2017*, 2017.
[24] F. Fossati, S. Hoteit, S. Moretti, S. Secci. "Fair Resource Allocation in Systems with Complete Information Sharing." *IEEE/ACM Transactions on Networking*, 2018.
[25] S. Hoteit et al., "On fair network cache allocation to content providers." *Computer Networks* 103: 129-142, 2016.
[26] LS. Shapley. "A value for n-person games", *H Kuhn and A Tucker, eds, Contributions to the Theory of Games*, Vol. 2 of Annals of Mathematics Studies, Princeton U Press., 1953.
[27] D. Schmeidler, "The nucleolus of a characteristic function game." *SIAM J. on applied mathematics*, 17(6), 1163-1170 1969.
[28] T. Bonald, J. Roberts, "Multi-resource fairness: Objectives, algorithms and performance." *ACM SIGMETRICS Performance Evaluation Review. Vol. 43. No. 1. ACM*, 2015.
[29] P. Poullie, T. Bocek, 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*, 15.1: 169-183, 2018.
[30] H. Varian, "Equity, envy, and efficiency." *J. of Eco. Th.*, 9(1):6391, 1974.
[31] R.R. Yager, "On ordered weighted averaging aggregation operators in multi-criteria decision making." *IEEE Transactions on Systems, Man and Cybernetics 18*, 1988.
[32] A.F. Shorrocks, " Ranking income distributions." *Economica 50.197*, 1983: 3-17.
[33] J.A. Weymark, " Generalized Gini inequality indices." *Mathematical Social Sciences 1.4*, 1981: 409-430.
[34] J. Rawls, " A theory of justice." *Harvard university press*, 2009.
[35] W. Ogryczak, T. liwiski, "On solving linear programs with the ordered weighted averaging objective." *European J. of Op. Research 148.1*, 2003.
[36] J. Lesca, and P. Perny, "LP Solvable Models for Multiagent Fair Allocation Problems." *ECAI* 2010.
[37] J.F. Nash Jr, "The bargaining problem." *Econometrica, 18, 155-162*, 1950.
[38] E. Kalai, and M. Smorodinsky, "Other solutions to Nashs bargaining problem." *Econometrica, 43(3), 513-518*, 1975.
[39] D. Bertsimas, V. F. Farias and N. Trichakis, "The price of fairness." *Operations research, 59(1), 17-31*, 2011.
[40] C. Gini, "Measurement of inequality of incomes." *The Economic Journal 31.121: 124-126*, 1921.
[41] A.F. Shorrocks, "Ranking income distributions." *Economica, 50(197), 3-17*, 1983.
[42] H.Dalton, "The measurement of the inequality of incomes." *The Economic Journal 30.119: 348-361*, 1920.
[43] T. Gajdos and J. A. Weymark, "Multidimensional generalized Gini indices." *Economic Theory 26.3: 471-496* , 2005.
[44] D. Verma, "Service level agreements on IP networks." *Proc. IEEE (Special Issue on Evolution of Internet Technologies)*, vol. 92, 1382-1388, 2004.
[45] F. Fossati, S. Moretti, S. Secci, "Multi-Resource Allocation for Network Slicing under Service Level Agreements", *Proc. of 2019 Int. Conference on the Network of the Future (NoF 2019)*, Oct. 1-3, Rome, Italy.

Francesca Fossati is currently a postdoc researcher at Cnam, Paris, France. She received her Ph.D. from Sorbonne University, France, in 2019, and a M.Sc. in mathematical engineering from Politecnico di Milano, Milan, Italy in 2015. Her current research interests are about optimization and game theory, with applications to network resource allocation problems.

Stefano Moretti is researcher at the CNRS since 2009. He is a member of LAMSADE, a laboratory of Paris Dauphine University. He graduated in Environmental Science in 1999 from the University of Genoa, in Italy, and he was awarded from the same university with a Ph.D. in Applied Mathematics in 2006. In 2008, he was also awarded with a Ph.D. in Game Theory at Tilburg University, The Netherlands. His main research interests deal with cooperative game theory, and with the application of game theoretic models to the analysis of the interaction on networks.

Patrice Perny received the Ph.D. degree in Computer Science and Operations Research in 1992 from University Paris Dauphine. He became associate professor in 1992 at University Pierre et Marie Curie (UPMC), Paris, France, and full professor in 2002. His activities concern preference modeling, multiobjective optimization, decision and optimisation under uncertainty and risk, computational social choice and algorithmic game theory.

Stefano Secci is full professor of networking at Cnam (Conservatoire national des arts et métiers), Paris, France. He received the M.Sc. Degree in telecommunications engineering from Politecnico di Milano, Milan, Italy, in 2005, and a dual Ph.D. Degree in computer science and networks from Politecnico di Milano and Telecom ParisTech, France, in 2009. He was associate professor at LIP6, UPMC from 2010 to 2018. His current interests cover novel routing and switching architectures and network virtualization. Webpage: https://cedric.cnam.fr/~seccis.