Fair Multi-Resource Allocation with External Resource for Mobile Edge Computing

Erfan Meskar and Ben Liang
Department of Electrical and Computer Engineering, University of Toronto
{emeskar, liang}@ece.utoronto.ca

***Abstract*—We consider the problem of fair multi-resource allocation for mobile edge computing (MEC). In MEC, to execute tasks with demands on multiple types of computing resources in the edge servers, the users must upload their tasks over a single dedicated wireless communication link that exists outside the servers. For this environment, we design a multi-resource allocation mechanism that extends the notion of dominant resource fairness (DRF) to accommodate an external resource, called DRF-ER. It provides several highly desirable properties. First, DRF-ER is envy-free, as no user prefers the allocation of another user. Second, DRF-ER allocations are Pareto optimal, as no one can improve its allocation without decreasing that of the others. Finally, DRF-ER is strategy-proof, as no user has an incentive to lie about its resource demand. Large-scale simulation driven by Google cluster traces further shows that DRF-ER significantly outperforms a naive extension of DRFH, which is a well-known variant of DRF for multiple servers, leading to higher resource utilization.**

I. Introduction

User computing tasks usually require multiple types of resources in the computing servers (*e.g.*, memory and CPU cores) [1]. In MEC, in addition to these computing resources, the task input data and execution results are sent through a shared wireless communication link to and from the MEC servers. Hence, MEC requires the allocation of both communication and computation resources. Different MEC tasks consume vastly various amounts of these resources. For instance, video analysis, language translation, face recognition, and augmented reality applications typically have CPU-intensive tasks, graph analytics and data indexing may have memory-bound tasks, and vehicle-to-infrastructure communication services can bottleneck on wireless communication link bandwidth [2], [3]. In addition to diverse resource demand profiles, an MEC system is possibly constructed from a variety of server classes. These servers may have different processing capabilities, memory sizes, and storage spaces. Moreover, hardware upgrades, *i.e.*, adding new servers and phasing out existing ones, increase the server heterogeneity.

Developing a fair resource allocation mechanism is of immense significance to guaranteeing QoE for different workloads in MEC. Under-allocation of resources degrades a user's QoE and over-allocation would adversely impact other users in the shared MEC environment. In systems with a single type of resource, the most popular allocation policy proposed so far has been max-min fairness, which maximizes the minimum allocation received by a user in the system. In MEC, however, server heterogeneity and diversity across the resource demands present challenges to develop a fair resource allocation mechanism.

To evaluate an allocation policy in a multiple-resource environment, we check whether it satisfies several core properties of a fair resource allocation policy [4], [5], [6], [7]:

- Envy-Freeness (EF): No user prefers the allocation of another user.
- Pareto Optimality (PO): It should be impossible to increase the resource amount of a user without decreasing the allocation of another user.
- Strategy-Proofness (SP): Users should not be able to benefit by lying about their resource demands. SP provides incentive compatibility, as a user cannot improve its allocation by lying.
- Sharing Incentive (SI): The amount of resources each user should receive is at least as much as simply splitting the total resources equally.

These properties are trivially satisfied by max-min fairness in systems with a single resource; however, in the multiple-resource environment, it might not be possible to satisfy all of them.

Ghodsi *et al.* [4] proposed Dominant Resource Fairness (DRF), an allocation mechanism that describes a notion of fairness when allocating multiple types of resources. DRF computes the share of demanded resources for each user and finds each user's dominant share and the resource corresponding to the dominant share. Then, DRF applies max-min fairness across users' dominant shares. Ghodsi *et al.* proved that DRF meets all four of the required properties (*i.e.*, EF, PO, SP, and SI) when tasks are infinitesimally divisible. Subsequently, Parkes *et al.* [5] extended DRF and studied the problem of indivisible tasks. They proved that there is no mechanism that satisfies PO, SI, and SP in that case.

While DRF and several subsequent works address the demand heterogeneity of multiple resources, they all limit the discussion to a simplified model where all resources are concentrated into one server. In systems with multiple heterogeneous servers, applying DRF per server may lead to an allocation with arbitrarily low resource utilization [6]. Instead of allocating resources separately in each server, Dominant Resource Fairness for Heterogeneous servers (DRFH) jointly considers resource allocation across all servers [6]. It defines the dominant resource for a user based on the aggregate of all the resources and then computes a max-min optimal

This work has been supported in part by the Natural Sciences and Engineering Research Council of Canada.

allocation with respect to each user's share of such dominant resource. DRFH satisfies EF, PO, and SP [6]. Unlike previous mechanisms, Friedman *et al.* [7] directly allocated containers, which are isolated bundles of resources. This differs from the model in [4]-[6] as users cannot combine bundles. They proved that in both single-server and multi-server systems, no deterministic mechanism allocating containers to users can satisfy PO, SP, and SI simultaneously. Instead, Friedman *et al.* proposed Containerized-DRF, a randomized mechanism that satisfies all of the desired properties on average in multiple servers with indivisible jobs.

In this paper, we study the problem of fair resource allocation in the MEC environment with heterogeneous servers. Users have infinitesimally divisible tasks that require multiple computing resources on the MEC servers, as well as communication bandwidth on a shared link. Although DRFH and Containerized-DRF consider resource allocation across heterogeneous servers, they cannot be directly applied here. These mechanisms require a server in which every type of resource is contained. However, in MEC, there is a single dedicated wireless communication link that exists outside of the computing servers.

The contribution and organization of this paper are as follows. After describing the system model in Sec. II, we propose in Sec. III-A Dominant Resource Fairness with External Resource (DRF-ER), a DRF generalization for environments where a single communication channel is shared among users and computing resources are pooled by heterogeneous servers. In Sec. III-B we prove that DRF-ER retains most of the desirable properties. We evaluate the performance of DRF-ER via trace-driven simulations in Sec. IV, followed by concluding remarks in Sec. V.

II. System Model

We consider a set of mobile users that access edge computing servers over a shared communication channel. Let N be the number of users in the system.

A. Edge Computing Servers

The number of servers and the number of resource types in the servers (*e.g.*, CPU and memory) are denoted by S and R, respectively. The capacity of server s for resource r is denoted by $c_{s,r}$, and we define a capacity vector $\mathbf{c}_s = (c_{s,1}, \ldots, c_{s,R})$. The total capacity of the system is represented by $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_R)$ where

$$\hat{c}_r = \sum_{s=1}^{S} c_{s,r}.$$

The share of capacity of resource r in server s is

$$\tilde{c}_{s,r} = \frac{c_{s,r}}{\hat{c}_r}. \tag{1}$$

B. Shared Communication Channel

The wireless communication link is a single dedicated resource that exists outside of the computing servers. All users share this link when they upload their tasks to the servers. We denote user u's demand per task for communication link bandwidth by $d_{u,\text{BW}}$, and the total link bandwidth by c_{BW}. It is worth mentioning that the share of bandwidth, as defined in (1), is $\tilde{c}_{\text{BW}} = 1$.

C. Fair Resource Sharing

Let user u's demand vector be $\mathbf{d}_u = (d_{u,1}, \ldots, d_{u,R}, d_{u,\text{BW}})$ for each of its tasks. In this paper, we use the subscript $R+1$ and BW interchangeably. Then, the share of demanded resource r of user u is

$$\tilde{d}_{u,r} = \frac{d_{u,r}}{\hat{c}_r}, r = 1, ..., R+1.$$

Following the terminology of DRF, we define the dominant resource of user u as

$$r_u^* = \underset{1 \le r \le R+1}{\arg\max} \quad \tilde{d}_{u,r}.$$

Then the dominant share of user u is $\tilde{d}_{u,r_u^*}$. Note that the dominant resource for some users can be the communication link bandwidth. For all user u and resource r, we further define

$$\bar{d}_{u,r} = \frac{\tilde{d}_{u,r}}{d_{u,r_u^*}}, r = 1, \ldots, R+1 \ \&\ u = 1, \ldots, N$$

as the normalized demand.

Our objective is to develop a multi-resource fair allocation scheme for this unique MEC environment, which retains the core fairness properties that are achieved by multi-server extensions of DRF such as DRFH. At the same time, the proposed scheme should improve MEC resource utilization.

We note that DRFH is not directly applicable to our problem, since the shared communication bandwidth is a standalone resource type separate from the computing servers. To apply DRFH directly, we would need to consider the S computing servers as servers with zero capacity for link bandwidth and view the communication link as server $S+1$ with zero capacity for computing resources. However, DRFH cannot support any task on any of these computing or communication servers since users require both computing resources and link bandwidth and no server contains these resources altogether. Instead, in Sec. IV, we compare DRF-ER with an extension of DRFH where each computing server is pre-assigned a portion of the link bandwidth.

III. DRF-ER Design and Properties

A. Dominant Resource Fairness with External Resource

DRF-ER maximizes the minimum dominant share in the system, subject to resource constraints.

$$\max_{\{x_{u,s}\}} \quad \min_u \sum_{s=1}^{S} x_{u,s}\ \tilde{d}_{u,r_u^*} \tag{2a}$$

$$\text{s.t.} \quad \sum_{u=1}^{N}\sum_{s=1}^{S} x_{u,s}\tilde{d}_{u,\text{BW}} \le 1, \tag{2b}$$

$$\sum_{u=1}^{N} x_{u,s}\ \tilde{d}_{u,r} \le \tilde{c}_{s,r}, \quad r = 1, \ldots, R\ \&\ s = 1, \ldots, S \tag{2c}$$

where $x_{u,s}$ is the number of tasks allocated to user u in server s. Constraint (2b) refers to the link bandwidth capacity, and (2c) refers to the resource capacity of each server.

Let $\mathbf{x}_u = (x_{u,1}, \ldots, x_{u,S})^T$ be a vector of the number of tasks allocated to user u and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$. Further utilizing an auxiliary variable g, the optimization problem (2) can be re-written as follows:

$$\max_{g,\{x_{u,s}\}} \quad g \tag{3a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S} x_{u,s}\ \tilde{d}_{u,r_u^*} = g, \quad u = 1, \ldots, N, \tag{3b}$$

$$(2b), (2c).$$

We can further simplify (2b) and rewrite the optimization problem as

$$\max_{g,\{x_{u,s}\}} \quad g \tag{4a}$$

$$\text{s.t.} \quad g \leq \frac{1}{\sum_{u=1}^{N} \bar{d}_{u,\text{BW}}}, \tag{4b}$$

$$(2c), (3b).$$

We denote the optimal solution to problem (4) by $\mathbf{x}^*$.

For every user u and server s, let $\tilde{A}_{u,\text{BW}}$ be the share of link bandwidth allocated to user u, and $\tilde{\mathbf{A}}_{u,s} = (\tilde{A}_{u,s,1}, \ldots, \tilde{A}_{u,s,R})$ be the resource allocation vector, where $\tilde{A}_{u,s,r}$ is the share of resource r allocated to user u in server s. Let $\tilde{\mathbf{A}}_u = (\tilde{\mathbf{A}}_{u,1}, \ldots, \tilde{\mathbf{A}}_{u,S})$ be the computing-resource allocation matrix of user u, and $\tilde{\mathbf{A}} = (\tilde{\mathbf{A}}_1, \ldots, \tilde{\mathbf{A}}_N)$ be the overall allocation for all users. Equation (5) calculates $\tilde{A}^*_{u,s,r}$, the share of resource r in server s that DRF-ER allocates to user u:

$$\tilde{A}^*_{u,s,r} = x^*_{u,s}\ \tilde{d}_{u,r}. \tag{5}$$

The share of link bandwidth allocated to user u is

$$\tilde{A}^*_{u,\text{BW}} = \sum_{s=1}^{S} x^*_{u,s}\ \tilde{d}_{u,\text{BW}}. \tag{6}$$

B. Analysis of Core Properties

In this section, we study the four major properties as explained in Section I. We first discuss how to determine the number of tasks that a user can execute given some arbitrary resource allocation that may not be optimal with respect to problem (4). Let $X_{u,s}(\tilde{\mathbf{A}}_u\,, \tilde{A}_{u,\text{BW}})$ be the number of tasks user u can execute on each server by solving optimization problem (7) below given $\tilde{\mathbf{A}}_u$ and $\tilde{A}_{u,\text{BW}}$:

$$\max_{\{x_{u,s}\}} \quad \sum_{s=1}^{S} x_{u,s} \tag{7a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S} x_{u,s}\ \tilde{d}_{u,\text{BW}} \leq \tilde{A}_{u,\text{BW}}, \tag{7b}$$

$$x_{u,s}\ \tilde{d}_{u,r} \leq \tilde{A}_{u,s,r}, \qquad r = 1, \ldots, R\ \&\ s = 1, \ldots, S. \tag{7c}$$

Problem (7) may have multiple optimal solutions. One of the solutions can be constructed from equation (8).

$$X_{u,s}(\tilde{\mathbf{A}}_u\,, \tilde{A}_{u,\text{BW}}) = Y_{u,s}(\tilde{\mathbf{A}}_u\,, \tilde{A}_{u,\text{BW}}) \times \min\left\{1, \frac{\tilde{A}_{u,\text{BW}}}{\sum_{s=1}^{S} Y_{u,s}(\tilde{\mathbf{A}}_u\,, \tilde{A}_{u,\text{BW}})\ \tilde{d}_{u,BW}}\right\} \tag{8}$$

where

$$Y_{u,s}(\tilde{\mathbf{A}}_u\,, \tilde{A}_{u,\text{BW}}) = \min_{1\leq r\leq R}\left\{\frac{\tilde{A}_{u,s,r}}{\tilde{d}_{u,r}}\right\}.$$

And the total number of tasks that user u can execute with this arbitrary allocation matrix is

$$\sum_{s=1}^{S} X_{u,s}(\tilde{\mathbf{A}}_u\,, \tilde{A}_{u,\text{BW}}) = \min\left\{\sum_{s=1}^{S} \min_{1\leq r\leq R}\left\{\frac{\tilde{A}_{u,s,r}}{\tilde{d}_{u,r}}\right\},\ \frac{\tilde{A}_{u,\text{BW}}}{\tilde{d}_{u,\text{BW}}}\right\}. \tag{9}$$

We now show that under the DRF-ER allocation, no user prefers the allocation of another user.

Proposition 1. *The DRF-ER allocation obtained by solving* (4) *satisfies EF.*

Proof. Let us assume $\tilde{\mathbf{A}}^*$ is the optimum allocation obtained by (5) and (6). We need to show that if some user (*e.g.*, u_0) gets another user's (*e.g.*, user u_1) allocated resource it cannot execute more jobs. In other words, we need to show that

$$\sum_{s=1}^{S} X_{u_0,s}(\tilde{\mathbf{A}}^*_{u_1}\,, \tilde{A}^*_{u_1,\text{BW}}) \leq \sum_{s=1}^{S} x^*_{u_0,s}.$$

According to equations (5), (6) and (9) we have

$$\begin{aligned}
&\sum_{s=1}^{S} X_{u_0,s}(\tilde{\mathbf{A}}^*_{u_1}\,, \tilde{A}^*_{u_1,\text{BW}}) \\
&\quad= \min\left\{\sum_{s=1}^{S} \min_{1\leq r\leq R}\left\{\frac{x^*_{u_1,s}\ \tilde{d}_{u_1,r}}{\tilde{d}_{u_0,r}}\right\},\ \frac{\sum_{s=1}^{S} x^*_{u_1,s}\ \tilde{d}_{u_1,\text{BW}}}{\tilde{d}_{u_0,\text{BW}}}\right\} \\
&\quad= \sum_{s=1}^{S} x^*_{u_1,s} \times \min_{1\leq r\leq R+1}\left\{\frac{\tilde{d}_{u_1,r}}{\tilde{d}_{u_0,r}}\right\} \\
&\quad= \sum_{s=1}^{S} x^*_{u_1,s} \times \min_{1\leq r\leq R+1}\left\{\frac{\bar{d}_{u_1,r}}{\bar{d}_{u_0,r}}\right\} \times \frac{\tilde{d}_{u_1,r^*_{u_1}}}{\tilde{d}_{u_0,r^*_{u_0}}}.
\end{aligned}$$

Recall that we use BW and $R+1$ interchangeably. We know $\min_{1\leq r\leq R+1}\left\{\frac{\bar{d}_{u_1,r}}{\bar{d}_{u_0,r}}\right\} \leq \frac{\bar{d}_{u_1,r^*_{u_0}}}{\bar{d}_{u_0,r^*_{u_0}}} = \bar{d}_{u_1,r^*_{u_0}} \leq 1$. Hence, we have

$$\sum_{s=1}^{S} X_{u_0,s}(\tilde{\mathbf{A}}^*_{u_1}\,, \tilde{A}^*_{u_1,\text{BW}}) \leq \frac{\sum_{s=1}^{S} x^*_{u_1,s}\tilde{d}_{u_1,r^*_{u_1}}}{\tilde{d}_{u_0,r^*_{u_0}}} \tag{10}$$

where $x^*_{u_0,s}$ and $x^*_{u_1,s}$ are obtained by solving (4). Furthermore, constraint (3b) implies that

$$\sum_{s=1}^{S} x^*_{u_1,s}\tilde{d}_{u_1,r^*_{u_1}} = \sum_{s=1}^{S} x^*_{u_0,s}\tilde{d}_{u_0,r^*_{u_0}}. \tag{11}$$

Using equations (10) and (11) together proves the claim. □

We next show that under DRF-ER, no user can improve its allocation without decreasing the allocation some other user.

Proposition 2. *The DRF-ER allocation obtained by solving* (4) *satisfies PO.*

Proof. Assume, by way of contradiction, that $\mathbf{x}^*$ obtained by solving (4) is not Pareto optimal. Thus there exists a feasible $\mathbf{x}'$ (*i.e.*, satisfying (2b) and (2c)) such that for all users $\sum_{s=1}^{S} x'_{u,s} \geq \sum_{s=1}^{S} x^*_{u,s}$ and there exists some user u_0 such that $\sum_{s=1}^{S} x'_{u_0,s} > \sum_{s=1}^{S} x^*_{u_0,s}$. Then, there exists some $\delta > 0$ such that

$$\sum_{s=1}^{S} x'_{u_0,s} \geq \sum_{s=1}^{S} x^*_{u_0,s} + S\delta \tag{12}$$

where S is the number of servers. Hence, there exists some server s_0 such that $x'_{u_0,s_0} \geq x^*_{u_0,s_0} + \delta$. Since $\mathbf{x}'$ is feasible, we have

$$\sum_{u=1}^{N}\sum_{s=1}^{S} x'_{u,s}\ \tilde{d}_{u,\text{BW}} \leq 1 \tag{13}$$

$$\sum_{u=1}^{N} x'_{u,s}\ \tilde{d}_{u,r} \leq \tilde{c}_{s,r}, \quad r = 1,\ldots,R\ \&\ s = 1,\ldots,S. \tag{14}$$

We construct $\mathbf{x}''$ by reducing δ tasks from user u_0 in server s_0 and adding $\min\limits_{1\leq r\leq R+1}\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\}$ tasks in server s_0 to each user, including user u_0. Thus,

$$x''_{u,s} = \begin{cases} x'_{u_0,s_0} - \delta + \min\limits_{1\leq r\leq R+1}\left\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\right\} & \text{if } \substack{u=u_0\\ s=s_0} \\ x'_{u,s_0} + \min\limits_{1\leq r\leq R+1}\left\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\right\} & \text{if } \substack{u\neq u_0\\ s=s_0} \\ x'_{u,s_0} & \text{if } s\neq s_0 \end{cases} \tag{15}$$

Now we check if $\mathbf{x}''$ is a feasible allocation. First, we study constraint (2b).

$$\begin{aligned}
&\sum_{u=1}^{N}\sum_{s=1}^{S} x''_{u,s}\ \tilde{d}_{u,\text{BW}} \\
&= \left(\sum_{s=1}^{S} x'_{u_0,s} - \delta + \min_{1\leq r\leq R+1}\left\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\right\}\right)\tilde{d}_{u_0,\text{BW}} \\
&\quad + \sum_{\substack{u=1\\u\neq u_0}}^{N}\left(\sum_{s=1}^{S} x'_{u,s} + \min_{1\leq r\leq R+1}\left\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\right\}\right)\tilde{d}_{u,\text{BW}} \\
&= \sum_{u=1}^{N}\sum_{s=1}^{S} x'_{u,s}\ \tilde{d}_{u,\text{BW}} + \min_{1\leq r\leq R+1}\left\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\right\}\sum_{u=1}^{N}\tilde{d}_{u,\text{BW}} \\
&\quad - \delta\,\tilde{d}_{u_0,\text{BW}} \\
&\leq \sum_{u=1}^{N}\sum_{s=1}^{S} x'_{u,s}\ \tilde{d}_{u,\text{BW}} + \frac{\delta\tilde{d}_{u_0,\text{BW}}}{\sum_{i=1}^{N}\tilde{d}_{i,\text{BW}}}\sum_{u=1}^{N}\tilde{d}_{u,\text{BW}} - \delta\tilde{d}_{u_0,\text{BW}} \\
&\Rightarrow \sum_{u=1}^{N}\sum_{s=1}^{S} x''_{u,s}\ \tilde{d}_{u,\text{BW}} \leq \sum_{u=1}^{N}\sum_{s=1}^{S} x'_{u,s}\ \tilde{d}_{u,\text{BW}}
\end{aligned} \tag{16}$$

Equations (13) and (16) show that $\mathbf{x}''$ satisfies constraint (2b). Equation (14) and (15) imply that constraint (2c) is satisfied in server $s \neq s_0$. For $r_0 = 1,\ldots,R$ in server s_0 we have

$$\begin{aligned}
&\sum_{u=1}^{N} x''_{u,s_0}\tilde{d}_{u,r_0} \\
&= \sum_{u=1}^{N} x'_{u,s_0}\tilde{d}_{u,r_0} + \min_{1\leq r\leq R+1}\left\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\right\}\sum_{u=1}^{N}\tilde{d}_{u,r_0} - \delta\,\tilde{d}_{u_0,r_0} \\
&\leq \sum_{u=1}^{N} x'_{u,s_0}\tilde{d}_{u,r_0} + \frac{\delta\tilde{d}_{u_0,r_0}}{\sum_{i=1}^{N}\tilde{d}_{i,r_0}}\sum_{u=1}^{N}\tilde{d}_{u,r_0} - \delta\tilde{d}_{u_0,r_0} \\
&\Rightarrow \sum_{u=1}^{N} x''_{u,s_0}\tilde{d}_{u,r_0} \leq \sum_{u=1}^{N} x'_{u,s_0}\tilde{d}_{u,r_0} \leq \tilde{c}_{s,r}.
\end{aligned}$$

Hence, $\mathbf{x}''$ is a feasible allocation.

For any user $u \neq u_0$ we have

$$\begin{aligned}
&\sum_{s=1}^{S} x''_{u,s} = \sum_{s=1}^{S} x'_{u,s} + \min_{1\leq r\leq R+1}\left\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\right\} > \sum_{s=1}^{S} x'_{u,s} \\
&\Rightarrow \sum_{s=1}^{S} x''_{u,s} > \sum_{s=1}^{S} x^*_{u,s} \Rightarrow \sum_{s=1}^{S} x''_{u,s}\tilde{d}_{u,r^*_u} > g^*
\end{aligned}$$

where $\mathbf{x}^*$ and g^* are obtained by solving (4). For user u_0 we have

$$\sum_{s=1}^{S} x''_{u_0,s} = \sum_{s=1}^{S} x'_{u_0,s} - \delta + \min_{1\leq r\leq R+1}\left\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\right\}. \tag{17}$$

Equation (12) and (17) imply that

$$\begin{aligned}
&\sum_{s=1}^{S} x''_{u_0,s} \geq \sum_{s=1}^{S} x_{u_0,s} + (S-1)\delta + \min_{1\leq r\leq R+1}\left\{\frac{\delta\tilde{d}_{u_0,r}}{\sum_{i=1}^{N}\tilde{d}_{i,r}}\right\} \\
&\Rightarrow \sum_{s=1}^{S} x''_{u_0,s} > \sum_{s=1}^{S} x_{u_0,s} \Rightarrow \sum_{s=1}^{S} x''_{u_0,s}\tilde{d}_{u_0,r^*_{u_0}} > g^*.
\end{aligned}$$

Hence, for all users we have $\sum_{s=1}^{S} x''_{u,s}\tilde{d}_{u,r^*_u} > g^*$. This contradicts the premise that g^* is optimal for (4). □

We next show that under the DRF-ER allocation, a user cannot improve its allocation by lying.

Proposition 3. *The DRF-ER allocation obtained by solving* (4) *satisfies SP.*

Proof. Let g^* and $\mathbf{x}^*$ be the solution to problem (4) when user u_0 truthfully reports its demand. Let $\tilde{\mathbf{d}} = (\tilde{\mathbf{d}}_1,\ldots,\tilde{\mathbf{d}}_N)$ be the share-of-demand matrix. For $u \neq u_0$, $\tilde{\mathbf{d}}_u$ may not be user u's actual share of demand vector. Similarly, let g' and $\mathbf{x}'$ be the solution to problem (4), and $\tilde{A}'_{u,s,r} = x'_{u,s}\tilde{d}'_{u,r}$ and $\tilde{A}'_{u,\text{BW}} = \sum_{s=1}^{S} x'_{u,s}\tilde{d}'_{u,\text{BW}}$ be the allocated resources when user u_0 misreports its demand. Let $\tilde{\mathbf{d}}' = (\tilde{\mathbf{d}}'_1,\ldots,\tilde{\mathbf{d}}'_N)$ be the misreported share-of-demand matrix. Note that $\tilde{\mathbf{d}}'_u = \tilde{\mathbf{d}}_u$ for all users $u \neq u_0$ and $\tilde{\mathbf{d}}'_{u_0} \neq \tilde{\mathbf{d}}_{u_0}$. We need to show that

$$\sum_{s=1}^{S} X_{u_0,s}(\tilde{\mathbf{A}}'_{u_0}, \tilde{A}'_{u_0,\text{BW}}) \leq \sum_{s=1}^{S} x^*_{u_0,s}.$$

Case 1. $g' \leq g$ (*i.e.*, $\sum_{s=1}^{S} x'_{u,s}\tilde{d}'_{u,r'_u} \leq \sum_{s=1}^{S} x^*_{u,s}\tilde{d}_{u,r^*_u}$ for $u = 1, \ldots, N$ where $r'_u = \arg\max_{1\leq r\leq R+1} \bar{d}'_{u,r}$).

$$\begin{aligned}
&\sum_{s=1}^{S} X_{u_0,s}(\tilde{\mathbf{A}}'_{u_0}, \tilde{A}'_{u_0,\text{BW}}) \\
&= \min\left\{\sum_{s=1}^{S} \min_{1\leq r\leq R}\left\{\frac{x'_{u_0,s}\tilde{d}'_{u_0,r}}{\tilde{d}_{u_0,r}}\right\}, \frac{\sum_{s=1}^{S} x'_{u_0,s}\tilde{d}'_{u_0,\text{BW}}}{\tilde{d}_{u_0,\text{BW}}}\right\} \\
&= \sum_{s=1}^{S} x'_{u_0,s} \min_{1\leq r\leq R+1}\left\{\frac{\tilde{d}'_{u_0,r}}{\tilde{d}_{u_0,r}}\right\} \\
&= \frac{\sum_{s=1}^{S} x'_{u_0,s}\tilde{d}'_{u_0,r'_{u_0}}}{\tilde{d}_{u_0,r^*_{u_0}}} \min_{1\leq r\leq R+1}\left\{\frac{\bar{d}'_{u_0,r}}{\bar{d}_{u_0,r}}\right\} \\
&\leq \frac{\sum_{s=1}^{S} x'_{u_0,s}\tilde{d}'_{u_0,r'_{u_0}}}{\tilde{d}_{u_0,r^*_{u_0}}} \times \frac{\bar{d}'_{u_0,r'_{u_0}}}{\bar{d}_{u_0,r'_{u_0}}} \leq \frac{\sum_{s=1}^{S} x'_{u_0,s}\tilde{d}'_{u_0,r'_{u_0}}}{\tilde{d}_{u_0,r^*_{u_0}}} \\
&\leq \frac{\sum_{s=1}^{S} x^*_{u_0,s}\tilde{d}_{u_0,r^*_{u_0}}}{\tilde{d}_{u_0,r^*_{u_0}}} \\
\Rightarrow &\sum_{s=1}^{S} X_{u_0,s}(\tilde{\mathbf{A}}'_{u_0}, \tilde{A}'_{u_0,\text{BW}}) \leq \sum_{s=1}^{S} x^*_{u_0,s}
\end{aligned}$$

Case 2. $g' > g$ (*i.e.*, $\sum_{s=1}^{S} x'_{u,s}\tilde{d}'_{u,r'_u} > \sum_{s=1}^{S} x^*_{u,s}\tilde{d}_{u,r^*_u}$ for $u = 1, \ldots, N$).

Assume, by way of contradiction, that

$$\sum_{s=1}^{S} X_{u_0,s}(\tilde{\mathbf{A}}'_{u_0}, \tilde{A}'_{u_0,\text{BW}}) > \sum_{s=1}^{S} x^*_{u_0,s}. \tag{18}$$

For all user $i \neq u_0$, we have $X_{i,s}(\tilde{\mathbf{A}}'_i, \tilde{A}'_{i,\text{BW}}) = x'_{i,s}$. So

$$\begin{aligned}
&\sum_{s=1}^{S} X_{i,s}(\tilde{\mathbf{A}}'_i, \tilde{A}'_{i,\text{BW}})\tilde{d}_{i,r^*_i} = \sum_{s=1}^{S} x'_{i,s}\tilde{d}'_{i,r'_i} = g' > g \\
\Rightarrow &\sum_{s=1}^{S} X_{i,s}(\tilde{\mathbf{A}}'_i, \tilde{A}'_{i,\text{BW}})\tilde{d}_{i,r^*_i} > \sum_{s=1}^{S} x^*_{i,s}\tilde{d}_{i,r^*_i} \\
\Rightarrow &\sum_{s=1}^{S} X_{i,s}(\tilde{\mathbf{A}}'_i, \tilde{A}'_{i,\text{BW}}) > \sum_{s=1}^{S} x^*_{i,s}.
\end{aligned} \tag{19}$$

Note that $\tilde{\mathbf{d}}'_i = \tilde{\mathbf{d}}_i$ and $r'_i = r^*_i$ for $i \neq u_0$. Equation (18) and (19) contradict the Pareto optimality of DRF-ER. □

To satisfy the SI property in conjunction with EF, SP, and PO is non-trivial [6] and even impossible in systems with indivisible tasks [7]. Similar to these prior designs, DRF-ER does not satisfy SI as stated in the following proposition.

Proposition 4. *The DRF-ER allocation obtained by solving* (4) *does not satisfy SI.*

Proof. To prove this proposition, we give a counterexample. Consider a system consisting of two users. User 1's tasks require $<$ 1 CPU, 1 GB $>$, and User 2's tasks require $<$ 3 CPU, 2 GB $>$. Each user requires $1/10$ of the link bandwidth to upload a task. There are two servers in the system. Server 1 contains $<$ 1 CPU, 2 GB $>$ and server 2 contains $<$ 4 CPU, 3 GB $>$. DRF-ER allocates 1 task to user 1 on server 1, and 1.4 and 0.8 task to user 1 and 2 on server 2, respectively. However, user 2 can execute more tasks by evenly partitioning each server. □

IV. Experimental Results

In Sec. III-B, we have studied the desirable properties of DRF-ER. In this section, we further evaluate its performance in resource utilization and compare it with a naive extension of DRFH. First, we study a simple scenario in which users dynamically arrive and leave the system and demonstrate how DRF-ER improves resource utilization in comparison with DRFH. Then, we evaluate the performance of DRF-ER in a more realistic setting via large-scale simulation using Google cluster traces.

Figure 1 compares the performance of DRF-ER and DRFH when users dynamically join and leave the system. Consider a system with two servers. Server 1 contains 5 CPU cores and 10 GB of memory, and server 2 contains 10 CPU cores and 5 GB of memory. Three users arrive at this system. User 1 requires 1 CPU core and 2 GB of memory per task, user 2 requires 2 CPU cores and 1 GB of memory per task, and user 3 requires 3 CPU cores and 2 GB of memory per task. The users share a wireless communication channel to offload their tasks onto the servers. User 1, 2, and 3 require $2/15$, $1/15$, and $2/15$ of the wireless communication link bandwidth per task, respectively.

Unlike DRF-ER, to apply DRFH, we need to first assign a link bandwidth capacity to each server. Here, for the purpose of illustration, we consider the case where each server receives half of the link bandwidth in DRFH. The scenario under study in Figure 1 consists of five phases. In phase one, user 1 is the only active user in the system. In DRF-ER, user 1 runs 5 tasks on server 1 and 2.5 task on server 2. In DRFH, however, the number of tasks on server 1 decreases to 3.75 so that tasks on server 1 do not consume more than half of the link bandwidth. The second phase starts after 100 seconds when user 2 joins the system. DRF-ER allocates server 1 to user 1 and server 2 to user 2. Thus, each user can run 5 tasks. As shown in Figure 1, DRFH fails to allocate the resources optimally and all resources are under-utilized. At $t = 200$, user 3 joins the system. In this phase, all users are active, and the performance of DRF-ER and DRFH are identical. Similar to the first and second phases, in the fourth phase, when user 2 leaves the system, DRFH fails to find the Pareto-optimal solution, and all resources are under-utilized by 8 percent. After 100 seconds, user 1 leaves the system and the two resource allocation policies have similar performance. We emphasize here that DRF-ER achieves improved resource utilization while satisfying the desirable properties as explained in Sec. III-B. It is unknown whether *any* division of the communication bandwidth in applying DRFH can achieve the same properties.

Moreover, we use Google cluster-usage traces [8] to compare the performance of DRF-ER and DRFH. Each user wants to submit a job to the servers, and each job is divided into tasks

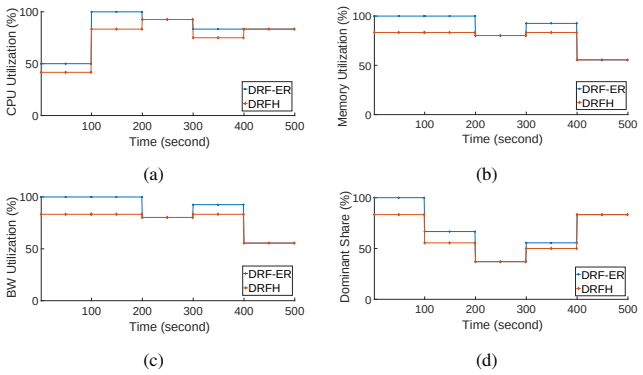

Fig. 1. Performance of DRF-ER and DRFH in a dynamic scenario. The allocation derived by DRFH is not Pareto optimal in phases 1, 2, or 4.

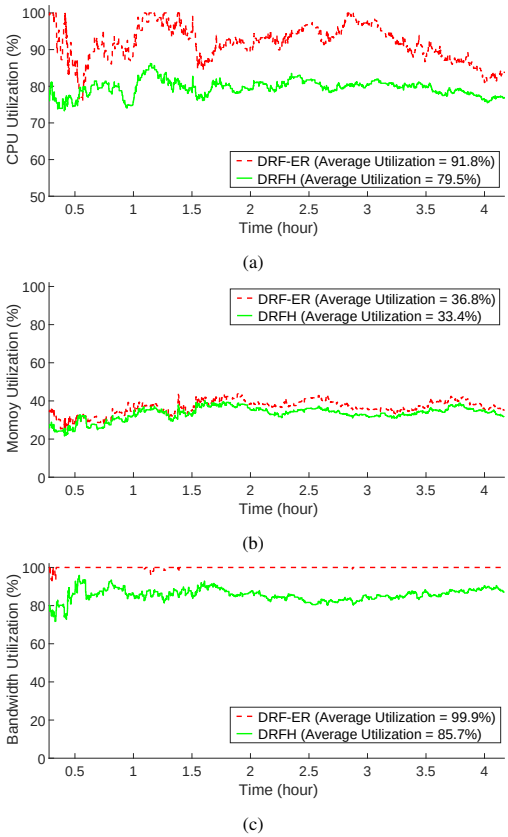

Fig. 2. Time series of resource utilization with Google traces.

with the same resource demand. The arrival time, duration, and resource demand (CPU and memory) of the tasks are available in the traces. To estimate the required bandwidth of the tasks, we assumed that $d_{u,\text{CPU}} f_{\text{CPU}} = X d_{u,\text{bps}}$, where $d_{u,\text{CPU}}$ is the CPU demand of user u, f_{CPU} is the CPU frequency of the server, $d_{u,\text{bps}}$ is the required bit rate of user u and X is a random variable with Gamma distribution [9]. In this paper, we follow the method of [9] and use $\alpha = 4$ and $\beta = 200$ to generate X. We consider a general MEC system with frequency division multiple access (FDMA) and estimate $\tilde{d}_{u,\text{BW}}$, the demanded link bandwidth of user u, based on $d_{u,\text{bps}}$. We take the 4-hour computing demand data from the Google traces and simulate their processing on a smaller MEC system of three servers with a capacity of 0.125 CPU and 0.125 memory and one server with capacity 1 CPU and 1 memory. The resource demand and capacity are normalized so that the maximum capacity of servers is 1. Figure 2 compares the resource utilization of DRF-ER and DRFH, where the link bandwidth is equally distributed among servers for DRFH. This figure illustrates that DRF-ER outperforms DRFH in the utilization of all resources. This is mainly because the latter is unable to dynamically compute fair allocation of the link bandwidth over time while concurrently maintain the fair allocation of the computing resources. In contrast, DRF-ER is specifically designed for this purpose, to improve resource utilization while achieving the fairness properties as proven in Sec. III-B.

V. Conclusion

In this paper, we consider a system where mobile users run their tasks on edge computing servers, where each task requires a specific amount of computing resources and communication link bandwidth. Since the communication link exists outside of the computing servers, we cannot directly apply the conventional multi-resource fair allocation mechanisms. The proposed scheme, DRF-ER, generalizes DRF and is shown to satisfy important desirable properties. Notably, DRF-ER is envy-free, Pareto optimal, and truthful. Furthermore, our trace-driven simulation results show that, compared with a naive extension of DRFH, DRF-ER achieves significant improvements in resource utilization. Finally, we remark that, although MEC serves as an important example application in this work, DRF-ER has general applicability in multi-resource fair allocation where there exists one type of resource that resides external to servers containing the other resource types.

References

[1] R. Grandl, G. Ananthanarayanan, S. Kandula, S. Rao, and A. Akella, "Multi-resource packing for cluster schedulers," *SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 455–466, Aug. 2014.

[2] B. Liang, "Mobile edge computing," in *Key Technologies for 5G Wireless Systems,* V. W. S. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, Eds., Cambridge University Press, 2017.

[3] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "Mobile edge computing: Survey and research outlook," *arXiv preprint arXiv:1701.01090*, 2017.

[4] 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 USENIX NSDI*, 2011.

[5] D. C. Parkes, A. D. Procaccia, and N. Shah, "Beyond dominant resource fairness: Extensions, limitations, and indivisibilities," *ACM Trans. Econ. Comput.*, vol. 3, no. 1, pp. 3:1–3:22, Mar. 2015.

[6] W. Wang, B. Liang, and B. Li, "Multi-resource fair allocation in heterogeneous cloud computing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 10, pp. 2822–2835, Oct. 2015.

[7] E. Friedman, A. Ghodsi, and C.-A. Psomas, "Strategyproof allocation of discrete jobs on multiple machines," in *Proceedings of the ACM Conference on Economics and Computation*, Jun. 2014.

[8] J. Wilkes, "More Google cluster data," Google research blog, Nov. 2011, posted at http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html.

[9] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4569–4581, Sep. 2013.