Stateful DRF: Considering the Past in a Multi-Resource Allocation

Hugo Sadok, Miguel Elias M. Campista *Senior Member, IEEE,*
and Luís Henrique M. K. Costa *Senior Member, IEEE*

Abstract—The multi-resource allocation problem arises in different scenarios. Different mechanisms have been proposed to fairly divide multiple resources, most notably, Dominant Resource Fairness (DRF). Even though DRF satisfies several desirable properties, it considers fairness only in the static setting. We propose Stateful DRF (SDRF), an extension of DRF that looks at past allocations and enforces fairness in the long run while keeping the fundamental properties of DRF. We prove that SDRF is strategyproof, since users cannot manipulate the system by misreporting their demands; incentivizes sharing, because no user is better off if resources are equally partitioned; and is efficient, as no allocation can be improved without decreasing another. In SDRF, users' priorities change over time. To avoid recalculating priorities at every task scheduling decision, we also propose Live Tree, a data structure that keeps elements with predictable time-varying priorities ordered. We implement SDRF on Mesos and run it in a real cluster. Moreover, we conduct large-scale simulations based on Google cluster traces of 30 million tasks over one month. Results show that SDRF reduces users' waiting time on average. This improves fairness, by increasing the number of completed tasks for users with lower demands, with negligible impact on high-demand users.

Index Terms—SDRF, multi-resource allocation, fairness, DRF

1 Introduction

Resource allocation is a fundamental problem in any shared computing system. Cloud computing centers and modern clusters are usually shared by users with different resource constraints [2], [3], [4], [5]. The amount of resources given to each user directly impacts the system performance from both fairness and efficiency standpoints [6]. In single-resource systems, max-min fairness is the most widely used and studied allocation policy [7], [8]. The main idea is to maximize the minimum allocation a user receives. It was originally proposed to ensure an equal share of link capacity for every flow in a network [9]. Since then, max-min has been applied to a variety of individual resource types, including CPU, memory and I/O [8]. Nevertheless, when it comes to *multi-resource* allocation, max-min is unable to ensure fairness [8], [10].

In a multi-resource setting, users often have heterogeneous demands and dynamic workloads [5], [8]. Some mechanisms have been proposed to address the multi-resource allocation problem [8], [10], [11], most notably, Dominant Resource Fairness (DRF) [8]. DRF generalizes max-min to the multi-resource setting, by giving users an equal share of their most demanded resource — their *dominant resource*. Using this approach, DRF achieves several properties.

Despite the extensive literature on fair allocation, most allocation policies focus on instantaneous or short term fairness, ensuring that users receive an equal share of the resources regardless of their past behaviors. DRF is no exception, it guarantees fairness only when users' demands do not change. In practice, however, users' workloads are dynamic [5], [12] and ignoring this fact leads to unfairness in the long run.

This paper proposes Stateful Dominant Resource Fairness (SDRF), an extension of DRF that accounts for the past behavior of users and improves fairness in the long run. The key idea is to make users with lower average usage have priority over users with higher average usage. When scheduling tasks, SDRF ensures that users that only sporadically use the system have their tasks scheduled faster than users with continuous high usage. The intuition for SDRF is that when users use more resources than their rightful share of the system, they commit to use less in the future if another user needs. SDRF tracks users' commitments and ensures that whenever system resources are insufficient, commitments are honored.

We conduct a thorough evaluation of SDRF and show that it retains the fundamental properties of DRF.[1] SDRF is strategyproof as users cannot improve their allocation by lying to the mechanism. SDRF provides sharing incentives as no user is better off if resources are equally partitioned. Moreover, SDRF is Pareto efficient as no user can have her allocation improved without decreasing another.

DRF can be efficiently implemented using a priority queue that determines which user has the highest allocation priority. However, when we consider the past, allocation priorities may change at any instant and the implementation cannot benefit from a priority queue. We mitigate this problem — being able to implement SDRF efficiently —

• *Hugo Sadok is with Carnegie Mellon University. Miguel Elias M. Campista and Luís Henrique M. K. Costa are with GTA/PEE/COPPE – Universidade Federal do Rio de Janeiro (UFRJ). This work was done while Hugo Sadok was at UFRJ and a preliminary version appeared in [1]. E-mail: {sadok, miguel, luish}@gta.ufrj.br*

1. We consider not only instantaneous utilities, but also the case where there are infinitely-repeating "games" and users discount their utilities using a discount factor.

introducing live tree, a data structure that keeps elements with predictable time-varying priorities sorted.

Besides the theoretical evaluation, we implement SDRF on Mesos [2] and show that it is able to adapt to users' demands in a 41-node cluster. We also analyze SDRF using large-scale simulations based on Google cluster traces containing 30 million tasks over a one-month period, and compare it to regular DRF. Results show that SDRF reduces the average time users wait for their tasks to be scheduled. Moreover, it increases the number of completed tasks for users with lower demands, with negligible impact on high-demand users. We also use Google cluster traces to evaluate the performance of live tree, concluding that SDRF can be used in practice.

2 System Model

In this section, we model the multi-resource allocation problem in a multi-user system. We first formalize users and resource demands, and then define the general structure of an allocation mechanism. From this structure we formalize users' sequential interactions as a repeated game.

2.1 Multi-Resource Setting and Allocation Mechanism

The system consists of a set of users $\mathcal{N} = \{1, \ldots, n\}$ that share a pool of different hardware resources $\mathcal{R} = \{1, \ldots, m\}$. Without loss of generality, we normalize the total amount of every resource in the system to 1, *i.e.*, if a system has a total of 100 CPU cores and 10 TB of memory, 0.1 CPU equals 10 cores while 0.1 memory equals 1 TB. For simplicity, we assume that the set of users and the amount of resources remain fixed. Every user i has a demand vector $\boldsymbol{\theta}_i^{(t)} = \langle \theta_{i1}^{(t)}, \ldots, \theta_{im}^{(t)} \rangle$ representing user i's demand for every resource at instant t. We consider positive demands for every resource type, therefore at every instant t, $\theta_{ir}^{(t)} > 0, \forall i \in \mathcal{N}, r \in \mathcal{R}$. This assumption simplifies the model but it may have important consequences in a real system [13], [14]. An allocation strategy using multiple rounds (as suggested by [13]) could potentially be used to circumvent this problem.

The allocation mechanism should produce as output a resource allocation based on users' *declared demands*. We represent the declared demands vector for a user i at instant t analogously to the demands vector, $\hat{\boldsymbol{\theta}}_i^{(t)} = \langle \hat{\theta}_{i1}^{(t)}, \ldots, \hat{\theta}_{im}^{(t)} \rangle$. When users declare demands truthfully, $\hat{\boldsymbol{\theta}}_i^{(t)} = \boldsymbol{\theta}_i^{(t)}$. We also define the allocation vector for user i at instant t for every resource type as $\boldsymbol{o}_i^{(t)} = \langle o_{i1}^{(t)}, \ldots, o_{im}^{(t)} \rangle$. The allocation returned by the mechanism at instant t is represented by a matrix of all the individual allocation vectors: $\boldsymbol{O}^{(t)} = \langle \boldsymbol{o}_1^{(t)}, \ldots, \boldsymbol{o}_n^{(t)} \rangle$. We impose a feasibility restriction to the allocations so that they may never be greater than the total amount of resources in the system, *i.e.*, at every instant t, $\sum_{i \in \mathcal{N}} o_{ir}^{(t)} \leq 1, \forall r \in \mathcal{R}$.

In the multi-resource setting it is useful to define users' *dominant resources* [8]. A user's dominant resource is the most demanded resource for this user relative to the total amount in the system. For example, if a user has a demand for 40% of the system CPU and 30% of the system memory, her dominant resource is CPU. As we have normalized all the different kinds of resources to 1, we say $d_i^{(t)}$ is a dominant resource for user i at instant t if

$$d_i^{(t)} \in \arg\max_{r \in \mathcal{R}} \hat{\theta}_{ir}^{(t)}. \tag{1}$$

2.2 Users' Utilities

We represent users' preferences using a utility function. Intuitively, users prefer allocations that maximize their number of tasks, being indifferent between different allocations that result in the same number of tasks. The number of tasks a user is able to run is determined by the resource with the smallest allocation compared to the user's demand. For example, if a user demands 0.4 CPU and 0.2 memory, but receives 0.4 CPU and 0.1 memory, the user received 100% of her CPU demand but only 50% of her memory demand. Therefore, the user will only be able to run half of her tasks.

To capture this intuition, we define the utility function as the amount of "useful" dominant resources a user receives. Formally, given an arbitrary allocation $\boldsymbol{o}_i^{(t)}$, for every user i and time t, the utility function is

$$u_i^{(t)}(\boldsymbol{o}_i^{(t)}) = \theta_{id}^{(t)} \cdot \min\left\{ \min_{r \in \mathcal{R}} \left\{ o_{ir}^{(t)} / \theta_{ir}^{(t)} \right\}, 1 \right\} \tag{2}$$

where $\theta_{id}^{(t)}$ is user i's dominant resource demand. When the utility is $\theta_{id}^{(t)}$, the user is able to allocate all the tasks she desires. Therefore, any allocation where $o_{ir}^{(t)} > \theta_{ir}^{(t)}$ for some $r \in \mathcal{R}$, still produces the same utility. In the example above, the user's utility would be 0.2, since she can use half of her dominant resource (CPU) demand.

This utility function also assumes that tasks are arbitrarily divisible [8], [13], [15]. This assumption does not hold in practice and we evaluate its impact in Sec. 5.2. Note that we do not rely on the utility function for interpersonal comparison, we only use it to induce ordinal preferences [13], [16]. This means that, even though the utility function can determine which allocation is better for a user, it cannot determine if one user is doing better than another.

2.3 Repeated Game

In the previous sub-section we referred to an instant t when defining most notations, however we omitted the influence time has in the allocation and in the user's preferences. In game theory, we typically say that at every instant t there is a *stage game* where users declare their demands $(\hat{\boldsymbol{\theta}}_i^{(t)}, \forall i \in \mathcal{N})$ and the allocation mechanism decides an allocation $(\boldsymbol{o}_i^{(t)}, \forall i \in \mathcal{N})$. The sequence of stage games defines the *repeated game*. To evaluate user's expected long-term utility, we consider that they discount future utilities using a discount factor $\delta_i \in [0, 1)$, *i.e.*, user i's *expected long-term utility* in the repeated game for the instant t is

$$u_i^{[t,\infty)} = \mathbb{E}_{u_i}\left[\sum_{k=t}^{\infty} \delta_i^{k-t} u_i^{(k)}(\boldsymbol{o}_i^{(k)}) \right]. \tag{3}$$

The discount factor δ_i represents the "user's patience"; the closer it is to 1, the more users care about future outcomes, the closer it is to 0, the more users care about recent future and the stage-game outcomes. Note that this model assumes

that, even though users prefer receiving resources early, they can tolerate some delay. However, in some instances, this may not be the case. An example is when users must comply with strict SLAs. In this setting, not having resources soon enough can cause great reduction in the user's utility. In this paper, we do not consider such instances.

3 DRF and Allocation Properties

In this section, we quickly review the DRF mechanism and the static allocation properties DRF and DRF-based schedulers usually satisfy. We show that these properties alone are not enough to enforce fairness in the long run, requiring an alternative for the dynamic setting.

3.1 DRF Mechanism

Dominant Resource Fairness (DRF) [8] extends Max-Min Fairness (MMF) to the multi-resource setting. DRF calculates an allocation based on users' dominant resources (Eq. 1). Given the dominant resource, we define the *normalized demand vector* for each user, in which the dominant resources become 1. The normalized demand vector for user i at instant t is denoted by $\tilde{\boldsymbol{\theta}}_i^{(t)} = \langle \tilde{\theta}_{i1}^{(t)}, \ldots, \tilde{\theta}_{im}^{(t)} \rangle$, where

$$\tilde{\theta}_{ir}^{(t)} = \frac{\hat{\theta}_{ir}^{(t)}}{\hat{\theta}_{id}^{(t)}}, \forall i \in \mathcal{N}, r \in \mathcal{R}\,. \tag{4}$$

When users request an infinite number of tasks, DRF computes an allocation where each user receives an equal share of their dominant resource. For this particular case, DRF can be described using a simple linear program whose solution (x) is the share of dominant resource each user receives [13]:

$$\begin{aligned} \max_{x} \quad & x \\ \text{s.t.} \quad & \sum_{i \in \mathcal{N}} o_{ir}^{(t)} \leq 1, \forall r \in \mathcal{R}\,, \\ & o_{ir}^{(t)} = x \cdot \tilde{\theta}_{ir}^{(t)}\,. \end{aligned} \tag{5}$$

Intuitively, we increase x — and consequently the share of dominant resource for every user — until we achieve a bottleneck and no task can be allocated. Given x, the allocation for every user and resource can be calculated as $o_{ir}^{(t)} = x \cdot \tilde{\theta}_{ir}^{(t)}$.

3.2 Static Allocation Properties

Ghodsi *et al*. [8] established that the DRF allocation satisfies the following properties. These properties have also been used in a variety of works [13], [15], [16] to measure both fairness and efficiency for a static resource allocation. For the following definitions we consider a stage game happening at time t.

1) *Sharing Incentives* (SI). Users should be better off participating in the system than having a proportional and exclusive share of all the resources. Formally, we say that an allocation mechanism satisfies sharing incentives if for every user $i \in \mathcal{N}$, it outputs an allocation $\boldsymbol{o}_i^{(t)}$ such that, $u_i^{(t)}(\boldsymbol{o}_i^{(t)}) \geq u_i^{(t)}(\langle 1/n, \ldots, 1/n \rangle)$. This assumes users have the right to an equal share of all the resources.[2]
2) *Strategyproofness* (SP). Users should not benefit by misreporting their demands to the mechanism. Formally, if we denote the allocation returned by the mechanism when the user i reports her demands truthfully ($\hat{\boldsymbol{\theta}}_i^{(t)} = \boldsymbol{\theta}_i^{(t)}$) as $\boldsymbol{o}_i^{(t)}$ and when the user lies ($\hat{\boldsymbol{\theta}}_i^{(t)} \neq \boldsymbol{\theta}_i^{(t)}$) as $\boldsymbol{o}_i'^{(t)}$, then $u_i^{(t)}(\boldsymbol{o}_i^{(t)}) \geq u_i^{(t)}(\boldsymbol{o}_i'^{(t)})$.
3) *Pareto Optimality* (PO). The allocation should be optimal in the sense that if it can be changed to make a user better, it must make at least another user worse (in other words the allocation cannot be Pareto dominated by another). Formally, an allocation mechanism is Pareto optimal if it returns an allocation $\boldsymbol{O}^{(t)}$ such that for any other feasible allocation $\boldsymbol{o}'^{(t)}$, if there is a user $i \in \mathcal{N}$ such that $u_i^{(t)}(\boldsymbol{o}_i'^{(t)}) > u_i^{(t)}(\boldsymbol{o}_i^{(t)})$ then there must be a user $j \in \mathcal{N}$ such that $u_j^{(t)}(\boldsymbol{o}_j'^{(t)}) < u_j^{(t)}(\boldsymbol{o}_j^{(t)})$.

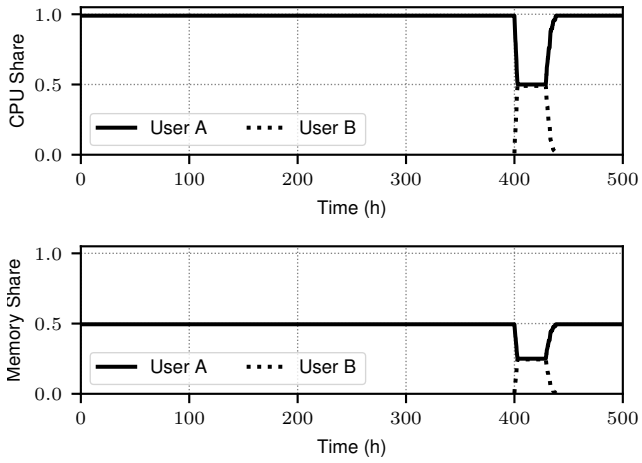

Fig. 1. Unfairness in the long run. User B hardly uses the system but receives the same shares as user A.

In addition to the above properties, DRF also satisfies *envy-freeness*, which ensures that users never prefer other user's allocation to their own. Unfortunately satisfying both Pareto optimality and envy-freeness is impractical under indivisibilities [13]. Moreover, as we will see next, while envy-freeness is usually desirable for static allocations, it does not ensure fairness in the dynamic setting.

3.3 Fairness in the Dynamic Setting

We now design an allocation policy that is fair in the long run. Previous work [8], [15] modeled users as having an infinite number of tasks with the same demand for each resource type. When this happens, only the share of resources each task needs is considered — time becomes irrelevant and the allocation is equivalent to a static one. In practice, however, while some users have workloads with repeated jobs, most users have quite dynamic workloads [5], [12].

To illustrate the importance of considering the past in an allocation, we present an example with users A and B sharing a system with a DRF scheduler (see Fig. 1). There

2. It is also possible to consider a weighted version, where users have a lower or higher share depending on their weights.

are two resources in the system, CPU and memory. User A's dominant resource is CPU and her normalized demand is $\langle 1, 0.5\rangle$. User A is eager for resources and submits a huge amount of tasks. Nevertheless, the other users only use the system sporadically, with usage spikes. After user A is using the entire system for a while, user B has a spike with normalized demand $\langle 1, 0.5\rangle$ as well. Even though user B never used her rightful share, the share she receives is the same as user A, *i.e.*, equal to $1/2$. This demonstrates that the properties of fairness defined for a static allocation are not enough to enforce fairness in the long run. Satisfying sharing incentives guarantees that users will receive their rightful share but does not reward users for their lower usage. Envy-freeness assumes users are only aware of the present allocation and do not envy other users based on their past allocations.

3.4 Users' Commitments

To distinguish between users who constantly require more resources than their proportional share from users who only use the system sporadically, we introduce the concept of *commitment*. Commitment is a measure of users' propensity to overuse their shares. The key intuition is that users who use more resources than their share, commit to use less if other users need. Users who overuse their shares for a short period of time should have lower commitment than users who constantly overuse. Also, users who overuse less resources should get lower commitment than users who overuse more resources. Every user $i \in \mathcal{N}$ has a separate commitment for each resource $r \in \mathcal{R}$. We define commitment using an exponential moving average of overused resources. The user i's commitment for resource r at time t is given by:

$$c_{ir}^{(t)} = (1-\delta)\sum_{k=0}^{t} \delta^{t-k}\bar{o}_{ir}^{(k)}, \tag{6}$$

where

$$\bar{o}_{ir}^{(k)} = \max\left\{\left(o_{ir}^{(k)} - \frac{1}{n^{(k)}}\right), 0\right\}. \tag{7}$$

The term $n^{(k)}$ is the number of users in the system at instant k. Therefore, the term $\bar{o}_{ir}^{(k)}$ represents how much user i overused her share for resource r on instant k. When this term is zero, the user did not overused her share. The more in the past users overused their shares the less it influences their commitments.

4 Stateful Dominant Resource Fairness

In this section, we introduce *Stateful Dominant Resource Fairness* (SDRF), a generalization of DRF that improves fairness in the long run by enforcing users' commitments. First we develop a simpler version of SDRF for a single resource type. Then, we extend this version and obtain an optimization problem that yields an SDRF allocation. From this problem we proceed to prove that it satisfies the desired properties introduced in Sec. 3.

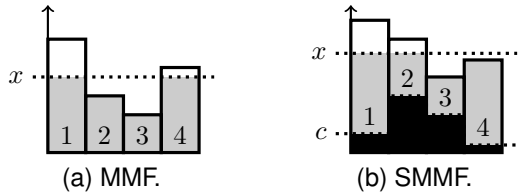

Fig. 2. Water-filling diagram for (a) MMF and (b) SMMF.

4.1 Stateful Max-Min Fairness

The intuition for SDRF is better understood if we first look at the single resource setting. Suppose we have a finite amount of a particular resource, *e.g.*, CPU cores, and we want to equally divide it among users. The fairest way to divide it is to give an equal share of the resource for every user, *e.g.*, same number of CPU cores. Nonetheless, some users may not need their entire share, in that case it can be redistributed among the other users. This is the main principle behind *Max-Min Fairness* (MMF). One way to achieve MMF is to use a *water-filling algorithm* [7]. Water-filling progressively gives resources for every user until their demands are met. When a user demand is met, she stops receiving resources and the algorithm continues to give resources for the other users. Fig. 2a shows the water-filling diagram for the MMF allocation. Each column (or tank) represents the total amount of resource each user demands. The resource is finite and progressively fills the tanks, until there is no more resource left. In the example, users 2 and 3 have their demands fulfilled while users 1 and 4 only have it partially fulfilled.

Even though MMF is fair for a static allocation, directly applying MMF to the dynamic setting causes the same problem as DRF — it does not consider the past and therefore cannot enforce fairness in the long run. To modify MMF to account for commitments, we introduce Stateful Max-Min Fairness (SMMF). The intuition behind SMMF is better illustrated by an example. "*If the equal share for the resource is 3 CPUs and the user has a commitment of 1 CPU, then the user should have the right to receive at least 2 CPUs.*" This notion can be directly implemented using the water-filling algorithm just by adding commitments as a "base for the tanks." Fig. 2b shows the water-filling diagram for the SMMF allocation. Demands are the same as in Fig. 2a, but now there is a base layer of arbitrary commitments c (black layer). Note how user 2 has a lower allocation than she would have without commitments. On the other hand, the demand for user 4 is now met.

Formally, the SMMF allocation can be defined using an optimization problem. Since SMMF allocates a single resource, the resources set becomes a singleton $\mathcal{R} = \{1\}$ and each user $i \in \mathcal{N}$ has a single allocation $o_{i1}^{(t)}$ at time t. The optimization problem maximizes x, the water level in Fig. 2b, as long as there are resources left in the system:

$$\begin{aligned} \max_{x} \quad & x \\ \text{s.t.} \quad & \sum_{i\in\mathcal{N}} o_{i1}^{(t)} \leq 1, \\ & o_{i1}^{(t)} = \max\left\{0, \min\left\{\hat{\theta}_{i1}^{(t)}, (x - c_{i1}^{(t)})\right\}\right\}. \end{aligned} \tag{8}$$

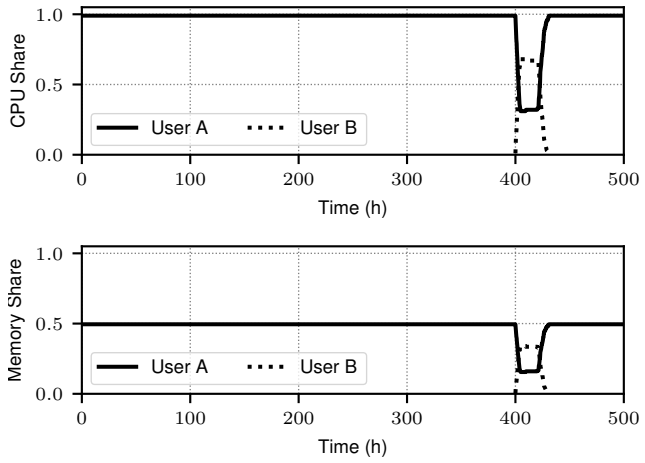

Fig. 3. Same example as Fig. 1 but using SDRF ($\delta = 1 - 10^{-6}$). Note how user B receives more resources and is able to complete her workload faster.

Given x, each user receives an allocation $o_{i1}^{(t)} = \max\{0, \min\{\hat{\theta}_{i1}^{(t)}, (x - c_{i1}^{(t)})\}\}$ which ensures that allocations are never above demands and remain non-negative. When commitments are zero, SMMF is equivalent to MMF.

Having defined SMMF, we now generalize it to multiple resources to finally obtain the SDRF mechanism.

4.2 SDRF Mechanism

SDRF generalizes SMMF similarly to the way DRF generalizes MMF to multiple resources. We use the same concept of dominant resource as DRF, defined in Eq. 1. Differently from DRF, though, we must deal with different commitments for different resources. We define the *dominant commitment* for a user i at time t as the user's largest commitment relative to the system total. As we have normalized all the resources to 1, the dominant commitment is simply the largest commitment for the user, *i.e.*,

$$\tilde{c}_i^{(t)} = \max_{r \in \mathcal{R}} \{c_{ir}^{(t)}\} \,. \tag{9}$$

Having defined the dominant commitment, we define SDRF using ideas from both DRF (Eq. 5) and SMMF (Eq. 8). Like DRF, SDRF increases the share of dominant resource for every user until a bottleneck is achieved. Like SMMF, users only start receiving resources when x is above their (dominant) commitment. SDRF is formally defined as:

$$\begin{aligned} \max_{x} \quad & x \\ \text{s.t.} \quad & \sum_{i \in \mathcal{N}} o_{ir}^{(t)} \leq 1, \, \forall r \in \mathcal{R} \,, \\ & o_{ir}^{(t)} = \max\left\{0, \min\left\{\hat{\theta}_{ir}^{(t)}, (x - \tilde{c}_i^{(t)}) \cdot \tilde{\theta}_{ir}^{(t)}\right\}\right\} \,. \end{aligned} \tag{10}$$

Recall $\tilde{\theta}_{ir}^{(t)}$ is the normalized demand for user i and resource r, defined in Eq. 4. From x, we may calculate the allocation for every user and resource by $o_{ir}^{(t)} = \max\{0, \min\{\hat{\theta}_{ir}^{(t)}, (x - \tilde{c}_i^{(t)}) \cdot \tilde{\theta}_{ir}^{(t)}\}\}$.

To illustrate SDRF operation, we repeat the same example presented in Sec. 3 (Fig. 1) but using SDRF instead of DRF. Fig. 3 shows how user B receives more resources (both CPU and memory) than user A and is able to complete her workload faster than when using DRF. Since user A is constantly using the system, receiving less resources for a short period will have a low impact in her overall workload.

In the next subsection we analyze the properties of SDRF that prove it behaves well in both the stage game and in the long run.

4.3 Analysis of SDRF Allocation Properties

We start our analysis of SDRF proving that, when users evaluate their utilities according to Eq. 2 and Eq. 3,[3] it satisfies the desirable properties introduced in Sec. 3.2, namely: strategyproofness, Pareto optimality and sharing incentives. We defer the proofs of all propositions to the Supplemental Appendix B.

First, we show that SDRF increases the share of dominant resource for every user until a resource runs out (the bottleneck resource). This is indicated in the following proposition.

Proposition 1 (Bottleneck). *The SDRF allocation obtained by solving Eq. 10 is such that all users have their demands fulfilled or there is a bottleneck resource. Formally, $\boldsymbol{o}_i^{(t)} = \hat{\boldsymbol{\theta}}_i^{(t)}, \forall i \in \mathcal{N}$ or $\exists r \in \mathcal{R}$ such that $\sum_{i \in \mathcal{N}} o_{ir}^{(t)} = 1$.*

Although simple, Proposition 1 is useful to demonstrate the following properties. One of the fundamental properties of DRF is strategyproofness. Without it, users may try to manipulate the system by, *e.g.*, faking their usage, which results in inefficiencies [8], [17]. Propositions 2 and 3 show SDRF is also strategyproof.

Proposition 2 (Strategyproofness in the Stage Game). *When users consider only the stage game utility (Eq. 2), the SDRF allocation obtained by solving Eq. 10 is strategyproof.*

Proposition 2 shows that when users consider only stage game utilities, SDRF is strategyproof. However, the fact that we consider past allocations may create new incentives for users to manipulate their declared demands. It may be possible that some users would not use the system when they actually need, hoping that this would improve their future allocations — this would also bring inefficiencies to the system. Fortunately, Proposition 3 shows that this is not possible.

Proposition 3 (Strategyproofness in the Repeated Game). *When users evaluate their utilities using the expected-long-term utility (Eq. 3), the SDRF allocation obtained by solving Eq. 10 is strategyproof, regardless of users' discount factors.*

The following two propositions demonstrate that SDRF is efficient. Proposition 4 shows that SDRF do not waste resources while Proposition 5 shows that the allocation is Pareto optimal, ensuring that it is not possible to increase a user's allocation without decreasing another.

Proposition 4 (Non-wastefulness). *The SDRF allocation $\boldsymbol{O}^{(t)}$ is such that, if there is a different allocation $\boldsymbol{O'}^{(t)}$ where $o'^{(t)}_{ir} \leq o_{ir}^{(t)}, \forall i \in \mathcal{N}, r \in \mathcal{R}$ and for a user $i^* \in \mathcal{N}$ and resource $r^* \in \mathcal{R}$, $o'^{(t)}_{i^*r^*} < o_{i^*r^*}^{(t)}$, then it must be that $u_{i^*}^{(t)}(\boldsymbol{o}_{i^*}^{(t)}) > u_{i^*}^{(t)}(\boldsymbol{o'}_{i^*}^{(t)})$. In other words, SDRF is non-wasteful.*

3. Note that although these properties may hold for other choices of utility functions, the proofs are specific to the functions we considered.

Proposition 5 (Pareto optimality). *The SDRF allocation obtained by solving Eq. 10 is Pareto optimal.*

The last property indicates that users are better off if they participate in the system. More specifically, it shows that users receive a utility at least as good as if they had access to $1/n$ of resources in the system.

Proposition 6 (Sharing incentives). *The SDRF allocation obtained by solving Eq. 10 satisfies sharing incentives.*

5 Practical Considerations

In this section, we study how to implement SDRF in practice. We first consider the effect continuous time and indivisible tasks have in the model defined in Sec. 2. We then develop a *water-filling* algorithm to schedule tasks. Nevertheless, the algorithm needs to recalculate and sort users' priorities at every execution. To mitigate this problem, we introduce *live tree* — a data structure that keeps elements sorted even with time-changing priorities — and show how it can be used to improve the SDRF scheduling algorithm.

5.1 Continuous Time

In the model defined in Sec. 2 we assume time progresses as a sequence of repeated games, suggesting a discrete time. The definition for commitment in Eq. 6 is compatible with this notion. In an actual system, however, tasks may arrive and finish at any instant, therefore we need an expression that allows us to compute commitment at continuous time.

First, since Eq. 6 is defined for a discrete time, we can redefine it recursively using a difference equation [18],

$$c_{ir}^{(t)} = (1-\delta)\bar{o}_{ir}^{(t)} + \delta c_{ir}^{(t-\Delta t)} \tag{11}$$

where the commitment at time t can be calculated from commitment at time $t-\Delta t$. This assumes $\bar{o}_{ir}^{(t)}$ remains constant within the interval $(t-\Delta t, t]$. It turns out that Eq. 11 can be seen as an exponential smoothing and can be closely approximated in the continuous time [18], leading to the expression:

$$\begin{aligned} c_{ir}^{(t)} &= (1-\mathring{\delta})\bar{o}_{ir}^{(t)} + \mathring{\delta} c_{ir}^{(t_0)} \\ \mathring{\delta} &= e^{-(t-t_0)/\tau},\ \tau = -\frac{\Delta t}{\ln(\delta)} \end{aligned} \tag{12}$$

where we may calculate $c_{ir}^{(t)}$ from any previous $c_{ir}^{(t_0)}$ as long as $o_{ir}^{(t)}$ remains constant from t_0 to t. Fortunately, o_{ir} only changes when a task for user i requiring resource r starts or finishes. In any other instant, o_{ir} remains constant, making Eq. 12 useful in practice. This expression is analogous to the discrete version, using $\mathring{\delta}$ instead of δ. When $t - t_0 = \Delta t$, Eq. 12 becomes Eq. 11.

5.2 Indivisible Tasks

So far, we have assumed that tasks are arbitrarily divisible. This allowed us to give arbitrarily small amounts of resources to users. In practice, however, tasks are often not divisible [2], [3], [4]. To schedule indivisible tasks we use the same approach as previous works [8], [15] — applying water-filling to tasks.

Algorithm 1 summarizes the task scheduling procedure. We define a set A of *active users* (users with at least one task waiting to be scheduled), and keep track of the total amount of resources allocated for every user. If the system is not full and if there is at least one user with a pending task, *i.e.*, $A \neq \emptyset$, we schedule the next task for the user with the lowest share of dominant resource compensated for commitments (line 5).

Algorithm 1 SDRF task scheduling

```
1:  A = {1, ..., k}                                          ▷ set of active users
2:  o_i = ⟨o_i1, ..., o_im⟩, ∀i ∈ A                          ▷ resources given to user i
3:  c_i = ⟨c_i1, ..., c_im⟩, ∀i ∈ A                          ▷ commitments for user i
4:  while A ≠ ∅ do
5:      i ← arg min_{i∈A} { max_{r∈R} {o_ir} + max_{r∈R} {c_ir} }        ▷ pick user
6:      ∀r, θ_ir ← demand for r in user i's next task
7:      if ∀r, ( θ_ir + Σ_{j∈N} o_jr ) ≤ 1 then
8:          ∀r, o_ir ← o_ir + θ_ir
9:          if no more pending tasks for user i then
10:             remove i from A
11:     else
12:         return                                           ▷ the system is full
```

Whenever a task arrives or finishes, we rerun Algorithm 1 with updated A and $\boldsymbol{o}_i$, $\boldsymbol{c}_i$, $\forall i \in A$. The smaller tasks are, the closer Algorithm 1 approximates Eq. 10.

Performance is a major concern in the design of a task scheduler. In peak hours, a scheduler may need to make hundreds of task placement decisions per second [5]. The most expensive part of Algorithm 1 is picking a user. While plain DRF can be implemented using a priority queue that stores the dominant resource share for every user, this is not possible with SDRF. With DRF, users' priorities only change when $\boldsymbol{o}_i$ changes, with SDRF, users' commitments change at any instant and so do users' priorities. Recomputing priorities for every user and resource at every task scheduling decision would be too costly. The next subsection shows how to solve this problem.

5.3 Live Tree

When scheduling tasks, we are not really interested in the specific value of commitments, but in which user we should pick the next task from. Live tree is a data structure that keeps elements with predictable time-varying priorities ordered. The key idea is to focus on position-change events, instead of element priorities. When priorities follow a continuous function, elements change position whenever their priorities intersect. A live tree always has a *current time* associated with it — for this current time, it guarantees that elements are sorted. When the current time is updated, instead of updating every element priority, we see if any position-change event happened from the last update to the current time.

Live tree can be seen as a combination of two red-black trees [19], [20] and an array (see Fig. 4). We call one red-black tree *elements tree*, as it keeps elements sorted by priority, while the other is the *events tree*, as it tracks position-change events sorted by their time. The array is used for element lookups. For simplicity, we assume that each of the input elements has a distinct integer id that can be an index for the array[4]. Each position in the array has a pointer to an

4. When elements do not have integer ids, or they are too sparse, the array may be replaced by a hash table and still present amortized $O(1)$ lookups.

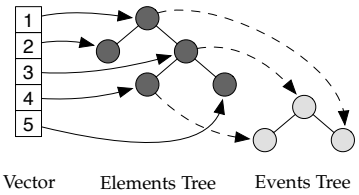

Fig. 4. Illustration of a live tree with its data structures. Positions in the array link to elements in the elements tree (solid arrows). Some elements link to events in the events tree (dashed arrows).

element in the elements tree (or NIL if there is no element for the given index). This allows us to retrieve elements by id in the tree in $O(1)$ time. If two neighboring elements in the elements tree are to change position in the future, the left element will have a pointer to a position-change event in the events tree.

We assume that priorities for all elements can be calculated using the same continuous function $p(t, \kappa)$ based on time t and in the element attribute κ. Every element has a different attribute that dictates how its priority changes over time. It may be a number, a vector or even a tuple. Our description does not depend on the definition of κ, in the Supplemental Appendix A we better define κ for our setting. It also helps to introduce an order notation, we say that element i precedes element j for an instant t, $i \prec_t j$, if $p(t, \kappa_i) < p(t, \kappa_j)$. The elements tree compares elements using $[\prec_t]$. This is useful since, whenever we insert a new element, it is compared to the others consistently with the time t. Live tree also needs a function to calculate priority intersections. We denote by $t_{\text{int}}(t, \kappa_i, \kappa_j)$ the function that calculates the priority intersection time based on two element attributes (κ_i, κ_j) and the time t. We now briefly describe the basic operations of a live tree:

INSERT(i, κ_i). To insert an element i in the live tree, we first insert i in the elements tree. Since the elements tree compares elements using $[\prec_t]$, i will be placed in the correct position relative to time t. Once inserted, we set a pointer from position i in the array to the element in the tree. Then, we call UPDATEEVENT for i and for its predecessor in the tree. When i is the minimum element, we only call UPDATEEVENT for i. INSERT can be accomplished in $O(\log n)$ time.

UPDATEEVENT(i). If an element i will change position with its successor in the future, it must have a position-change event associated with it. To update an event, we first check if the element i has an event in the events tree and remove it if so. Then, we check if i and its successor j will switch places in the future by calculating their priorities intersection $t_{\text{int}}(t, \kappa_i, \kappa_j)$. If t_{int} exists and is positive, we add an event for element i and time $t_{\text{int}} + t$ in the events tree. Then we add a pointer from element i in the elements tree to the event in the events tree. When i is the maximum element, it has no successor and thus cannot have a position-change event (note this does not imply it cannot change position, as its predecessor can have an event). UPDATEEVENT can be done in $O(\log n)$ time.

UPDATE(t). Whenever the current time changes, we must update the tree. We assume that time progresses forward and live tree can only be updated to the future. To update the tree to a new time t, we look at all events that happen before t. If there is no event, *i.e.*, the first event in the events tree has time greater than t, then no element should change position and the tree is already updated, otherwise we must consider the events. We remove events from the events tree in order until the next event has time greater than t or the events tree becomes empty. For every removed event, we remove its correspondent element as well as its successor from the elements tree calling DELETE. Once we finish removing events we reinsert each removed element calling INSERT. Since elements are compared using $[\prec_t]$, the reinsertion places elements in their correct position relative to time t. UPDATE can be accomplished in $O(n \log n)$ time. The worst case happens when every element must change position and therefore must be reinserted in the tree. In Sec. 7 we show that, for SDRF, the actual time is much smaller than the worst case.

DELETE(i). To delete an element i, we first check position i in the array. From position i we get a pointer to the elements tree. If the element has an event, we get a pointer to its event as well. We then remove the event from the events tree, the element from the elements tree and set NIL at position i in the array. If i was the minimum element, we are done, otherwise we must call UPDATEEVENT to the predecessor of i in the elements tree. DELETE can be accomplished in $O(\log n)$ time.

MINIMUM/MAXIMUM. The minimum (maximum) in the live tree is the minimum (maximum) in the elements tree. MINIMUM/MAXIMUM can be accomplished in $O(1)$ time.

We omitted from our description corner cases, such as if an element being deleted does not exist, or if the element being inserted is already in the tree.

Live tree performance depends heavily on the priority function used and the frequency of UPDATE calls. When elements change position often, UPDATE has to process more events. Nevertheless, the higher the frequency of UPDATE calls, the less events each call has to process. In Sec. 7 we evaluate how live tree performs when used to implement SDRF. Moreover, in the Supplemental Appendix A we give more details on how to use a live tree to implement an SDRF task scheduler.

6 IMPLEMENTATION

Now that we have looked at the practical aspects of implementing SDRF, we briefly describe our implementation of an SDRF scheduler for Mesos [2]. Mesos is a popular cluster manager that allows multiple computing frameworks (*e.g.*, Hadoop MapReduce [21], Spark [22], Storm [23]) to share the same cluster. A unique characteristic of Mesos is its use of resource offers. Instead of processing users' requests, Mesos offers resources to frameworks, that may choose to either accept or reject them. The advantage of this approach is that Mesos does not have to worry about sophisticated scheduling rules that are often required by computing frameworks. There are two types of Mesos nodes: master and agent. The master node is responsible for managing the agent nodes and to decide how much resources to offer to each framework. When deciding which framework to offer resources to, Mesos delegates decisions to an allocation

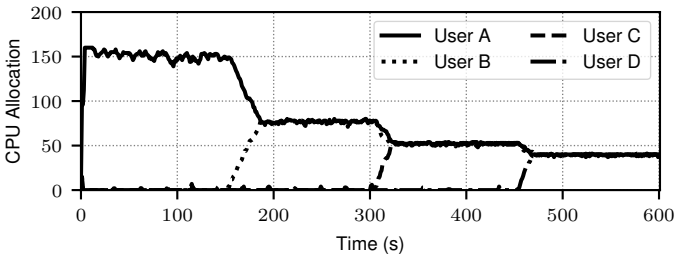

Fig. 5. CPU allocation (in number of cores) of four users sharing a 41-node Mesos cluster with a DRF scheduler. All users have CPU as their dominant resource. After every 150 s interval a user starts to launch tasks.

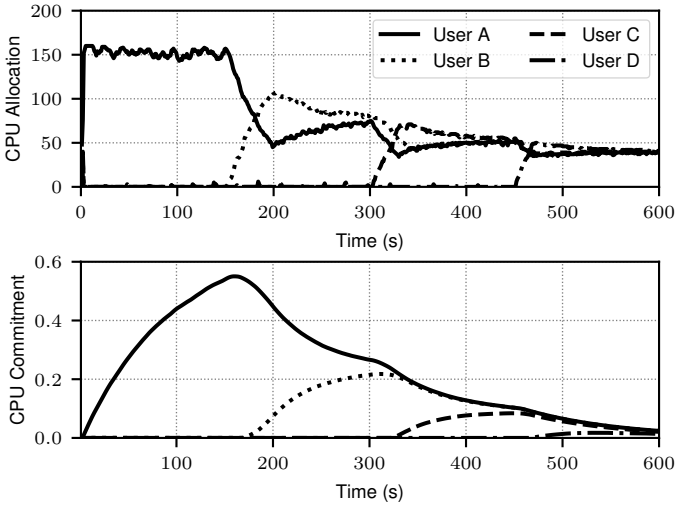

Fig. 6. Similar experiment as in Fig 5 but using and the SDRF scheduler with $\delta = 1 - 10^{-2}$. Observe how commitments build up when users are using more resources than their share, which causes them to receive a lower allocation when other users need the system.

module and, by default, it uses an allocation module that implements DRF. To implement DRF, this module offers resources to the framework with the lowest share of dominant resource.

To implement an SDRF allocation module, we offer resources to the framework with the lowest share of dominant resource compensated for commitments (line 5 of Algorithm 1). To facilitate profiling, we also included a separate thread to record users' allocations and commitments at a configurable frequency. This functionality can be enabled or disabled at compile time but we did not observe any performance degradation when using it. Our implementation uses C++17 and is based on Mesos 1.8.0. We put our implementation to test in the next section.

7 Evaluation

In this section, we evaluate SDRF using our Mesos module implementation as well as trace-driven simulations. We start by running Mesos on a real cluster and comparing its built-in DRF scheduler with our SDRF implementation. Then, we run macro-benchmarks using a discrete-event simulator.[5]

5. Our implementations of the discrete-event simulator with DRF and SDRF, Live Tree, and the Mesos SDRF Module are all open source and available at https://github.com/hsadok/sdrf and https://github.com/hsadok/mesos-sdrf.

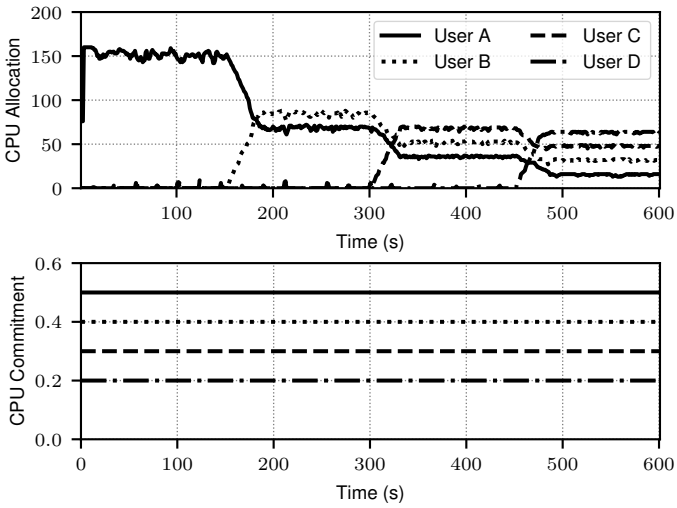

Fig. 7. Similar experiment as in Fig. 6 but using a high value of δ ($\delta = 1 - 10^{-7}$) and non-zero initial commitments. Users with lower commitments receive more resources.

7.1 Experiments on a Mesos Cluster

We deployed Mesos on a 41-node cluster using Amazon EC2 a1.xlarge instances. Each instance has 4 CPU cores and 8 GB of RAM. One instance was used as a master node while the remaining were configured as agents. In the agent nodes, we allowed Mesos to use up to 6 GB of RAM. All nodes run Ubuntu 18.04.1 with Linux version 4.15.

We conduct three experiments to observe how DRF and SDRF adjust to varying users' demands in a real cluster deployment. In all experiments we launched tasks from four users requiring 1 CPU and 1 GB of RAM, which ensures that CPU is the dominant resource for all users. To allow Mesos to fairly allocate tasks among these users, we give a unique framework ID to each one. In the beginning of the experiment only user A uses the system. Then, after every 150 s interval a new user starts launching tasks. All experiments run for 10 minutes.

To have a baseline for our experiments, we first configure Mesos to use the default DRF scheduler. Fig. 5 shows the CPU allocation through time for the four users. Since the system has 40 agent nodes, each with 4 CPU cores, the maximum number of CPUs that can be allocated is 160 cores. Every time a user starts launching tasks, the system gradually adapts to make sure all users receive their share, as calculated by DRF. All users that accept offers receive the same share of dominant resource, regardless of their past allocations.

We then configure Mesos to use our implementation of SDRF and repeat the same experiment with different values of δ and initial commitments. First, we want to verify if the system is able to adapt when commitments change. We set a low value for δ ($\delta = 1 - 10^{-2}$) and set initial commitments to zero. Fig. 6 shows the CPU allocation and commitments for the four users involved in the experiment. Observe how user A's commitment increases when she is using the system alone but decreases when she starts to share the system with other users. Moreover, because we set a low value of δ, it causes commitments to vary quickly. This causes user B to receive about 50 CPU cores more than user A, when

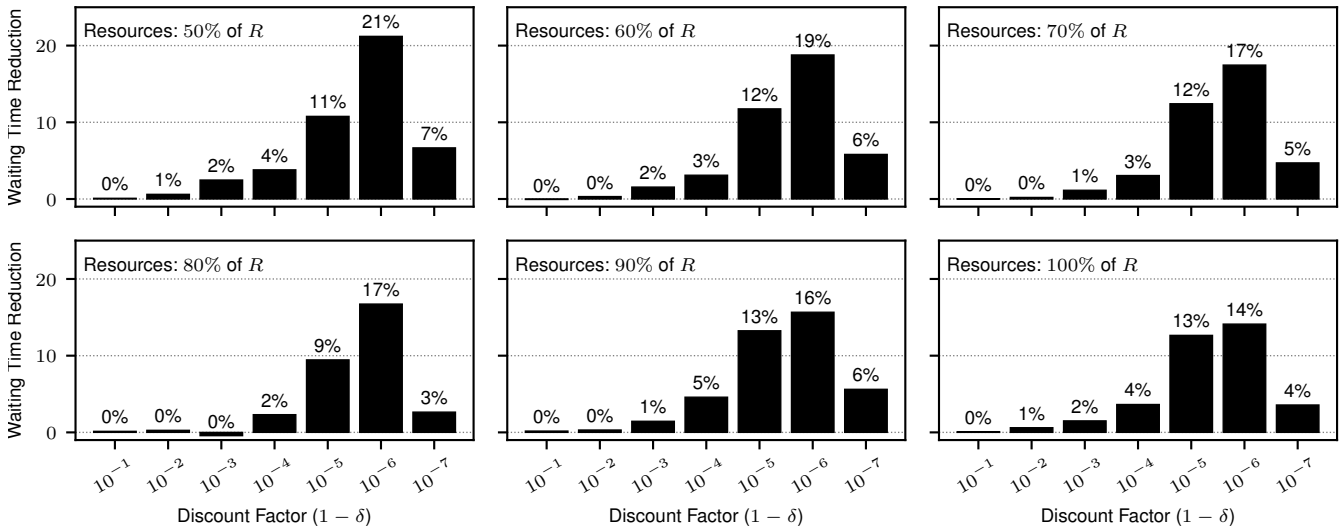

Fig. 8. Reduction in the mean over all users' mean waiting time under different values of δ and system load relative to DRF.

she starts launching tasks. In contrast, when user D starts launching tasks, the other users' commitments have already faded considerably which makes user D's allocation only slightly greater than the others. This shows that the system behaves as expected, even with fast-changing commitments.

In the third experiment, we want to verify if users receive their rightful share of resources according to the SDRF mechanism. To do so, we use a high value of δ ($\delta = 1-10^{-7}$) and set initial commitments so that users A, B, C, and D receive commitments 0.5, 0.4, 0.3, and 0.2, respectively. Fig. 7 shows the CPU allocation and commitments for the four users involved in the experiment. Because of the high value of δ, commitments change slowly and variations cannot be perceived through the duration of the experiment. This causes allocations to remain stable until another user starts launching tasks. Because users' commitments differ by 0.1 increments, the allocation they receive differs by 16 CPU cores (0.1 times the total amount of CPU cores in the system), which is compatible with the SDRF definition on Eq. 10.

Finally, we observe that the overall CPU utilization was close to the maximum cluster capacity in all experiments.

7.2 Simulations using Google Traces

We now evaluate SDRF and live tree using trace-driven simulations based on Google cluster traces [5]. The traces contain information from workloads (from either Google services or engineers) running in a cluster over a month-long period. Workloads are submitted in the form of jobs, and each job may have multiple tasks. The traces contain events for every time a task is submitted, is scheduled or finishes. From these events we extract the CPU and memory demands as well as task submission and running times using them as input for our simulation. We remove tasks with 0 demand, as well as tasks that were evicted by the Google system, but leave tasks that failed due to user errors. After that, we are left with around 32 million tasks from 627 users.

To run the simulations, we implemented a discrete-event simulator that takes as input user's tasks, with their starting time and duration. To decide which tasks to schedule, this simulator can use either DRF or SDRF (as defined in Algorithm 1). We feed the simulator with the Google traces and run simulations for different values of δ and system overload. The values of δ are relative to a Δt of 1 second (see Eq. 12). We vary δ by making it exponentially closer to 1, *i.e.*, $\delta = 1 - 10^{-1}, \ldots, 1 - 10^{-7}$. This is equivalent to exponentially increasing τ from Eq. 12. To verify how SDRF performs under different levels of system load we also perform simulations for multiple values of total resources (*i.e.*, CPU and memory). We use the average system usage in the original trace, called hereinafter as R, as a baseline for our results. We then run simulations with the total amount of resources in the system varying from 50% to 100% of R, in steps of 10%.

To help us understand the effect of SDRF on users' waiting time, we compute the mean waiting time for every user and report the reduction in the mean over these values for SDRF relative to DRF. Fig. 8 shows the result for different values of δ and system load. When δ is small enough, SDRF performs close to DRF. Also, for δ sufficiently close to 1, SDRF approaches DRF. This is justified inspecting Eq. 12: when δ is sufficiently close to 1, commitments never accumulate, alternatively, when δ is sufficiently close to 0, commitments are simply the last allocation, and therefore tasks are scheduled just like in DRF. The most significant waiting time reduction was observed for the discount factor $\delta = 1 - 10^{-6}$ for all levels of system load evaluated. Even though the advantage of SDRF is more evident when the system is overloaded, for $\delta = 1 - 10^{-6}$, SDRF consistently outperforms DRF by more than 10%. The reason for this behavior is that SDRF reduces the waiting for users with lower demand while having little impact on users with higher demand. To verify that this is the case, we consider the simulation with $\delta = 1-10^{-6}$ and total resources 50% of R. In Fig. 9, we sort users by total dominant resource usage and separately plot the the distribution of mean waiting time for the bottom and upper halves. We observe that the most significant reduction in waiting time occurs for users in the bottom half. This is expected since users with

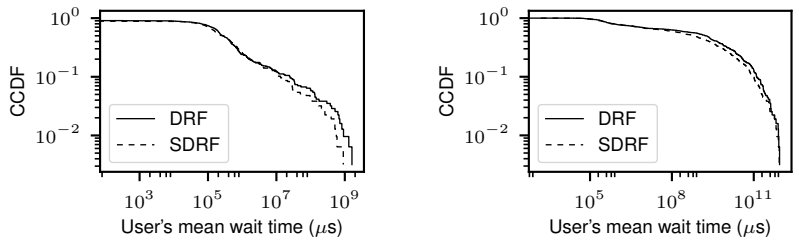

Fig. 9. CCDF of user's mean waiting time for (a) the half of users with lowest usage and (b) the half of users with the highest usage for the simulation with $\delta = 1 - 10^{-6}$ and total resources 50% of R. Most of the reduction in waiting time happens for users in the bottom 50%, while users in the upper 50% suffer little impact.

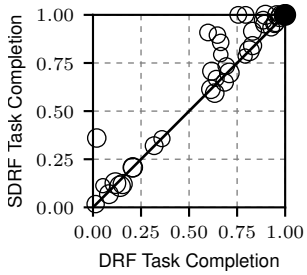

Fig. 10. Task completion ratio using DRF and SDRF. Each bubble is a different user. The bubble's size is logarithmic to the number of tasks submitted by the user. Users above the $y = x$ are better with SDRF.

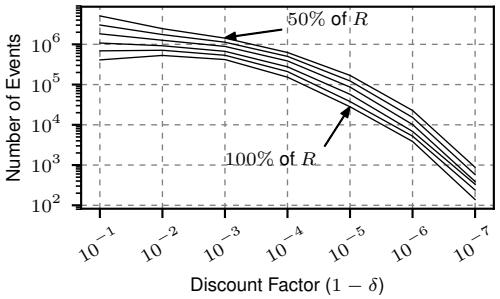

Fig. 11. Live tree events for different values of discount factor and system resources (50% to 100% of R from top to bottom). The closer δ is to one, the less events we observe.

TABLE 1
Total Time Spent on Different Operations

Operation	Naive (s)	Live Tree (s)
MINIMUM	966	31
INSERT	26	432
DELETE	54	171
Total	**1046**	**634**

higher usage are more likely to have a higher commitment. Moreover, this is compatible with the notion of long-term fairness that we introduced in Sec.3.

We also investigate how the waiting time reduction affects the number of tasks each user is able to complete. We compute the task completion ratio for every user (*i.e.*, the number of tasks completed divided by the number submitted) and compare it when running DRF and SDRF. Fig. 10 shows the results for the simulation with $\delta = 1-10^{-6}$ and total resources 50% of R. Each bubble represents a different user: when above the black $y = x$ line, the user is able to complete more tasks under SDRF than under DRF. Most users perform better under SDRF, in fact, only 11 out of 627 users completed less tasks under SDRF. Also note that, even though these users completed less tasks, their task completion ratio had low impact. This happens because users that use the system in small bursts complete their workloads earlier and, consequently, have the opportunity to complete more tasks. On the other hand, users that use the system continually experience a low impact.

Next we evaluate how live tree performs under the same simulations. The theoretical worst case complexity for the update operation is $O(n \log n)$. This is driven by the maximum number of events an update may trigger, when there is no event, updates are performed in $O(1)$. In practice, however, the average number of events is much smaller. Fig. 11 shows the number of live tree events that occurred during the entire simulation period for every simulation. Each curve represents a different value of system load, 50% of R to 100% of R, from top to bottom. The number of events increases when the amount of resources in the system decreases. Also, the closer δ is to 1, the less events we observe. This is expected, since commitments vary slower the closer δ is to 1. When $\delta = 1 - 10^{-6}$ and the total resources 50% of R, there is a total of $24{,}856$ events, which is about 8 events for every $1{,}000$ scheduled tasks. Since every task scheduled triggers one update, this indicates that updates happen fast for this scenario. Even for the worst scenario ($\delta = 1 - 10^{-1}$, 50% of R), the total number of events is $5{,}101{,}975$, which is less than 2 events for every 10 tasks. If live tree performed close to its theoretical worst case complexity, it would offer low advantage compared to sorting elements on every update. But with the number of updates observed, live tree operations perform close to the ones of a red-black tree where weights are static.

We also compare live tree with a naive implementation that recalculates commitments for every element whenever it needs to find the minimum one. We run the simulation using $\delta = 1 - 10^{-6}$ and total resources 50% of R and measure the total time spent in the main live tree operations. For this experiment we used a host running Linux 5.0.0-37 with two Intel Xeon E5-2650 CPUs and 256 GB of RAM. Table 1 shows the total time spent on MINIMUM, INSERT, and DELETE operations for both live tree and the naive implementation. The times reported for live tree operations also include a call to UPDATE. Moreover, note that since we are reporting the total time spent in every operation during the simulation, the times represent not only how costly an operation is but also how often it is called. We observe that live tree vastly reduces the amount of time spent on MINIMUM operations. However, because the naive implementation does not need to maintain an events tree, live tree spends more time on INSERT and DELETE. Nevertheless, when we look at the aggregate time, the naive implementation is still 65% slower.

8 Related Work

Fair resource allocation is a prevalent research topic, both in the computer science and economics fields. Nonetheless, focus is often given to the single resource setting. Ghodsi *et al*. [8] are the first to investigate the multi-resource setting under a shared computing perspective, proposing DRF. Dolev *et al*. [10] propose an alternative based on "bottleneck fairness." Nevertheless, the alternative is not strategyproof and is computationally expensive [17]. Gutman *et al*. [24] develop polynomial-time algorithms to compute both DRF and "bottleneck fairness" for non-discrete allocations. Joe-Wang *et al*. [6] extend the notion of fairness introduced by DRF to develop a framework that captures the fairness-efficiency tradeoff. However, they assume a cooperative environment and as such do not evaluate strategyproofness. Wang *et al*. [15] generalize DRF for a scenario with multiple heterogeneous servers, relaxing the sharing incentives restriction. Friedman *et al*. [25] also look at the allocation on multiple servers but provide a randomized solution that achieves sharing incentives. Another extension of DRF is proposed by Parkes *et al*. [13] to account for users with different weights and zero demands. Zarchy *et al*. [26] also investigate multi-resource allocation, but their focus is when the same application may be developed differently to use different proportions of resource types. They propose a framework that allows users to submit multiple demands for the same application. Even though the aforementioned works consider the multi-resource setting, they ignore the dynamic nature of users' demands.

Bonald and Roberts [11] suggest Bottleneck Max Fairness (BMF), which also does not enforce strategyproofness, but improves resource utilization as compared to DRF. They consider dynamic demands in their analysis, arguing that for highly dynamic environments, such as networks, it is hard for users to manipulate the system. BMF convergence is proved in a later work [27]. Even though the analysis of BMF considers dynamic demands, the allocation itself considers only short term usage, ignoring fairness in the long run. Kash *et al*. [16] investigate a dynamic setting where users arrive and never leave, however, they also assume that demands remain constant. Friedman *et al*. [28] evaluate the scenario where multiple users arrive and leave the system. The focus, however, is on the fair division of resources as soon as the user arrives, limiting the number of task disruptions. They do not consider the dynamic nature of user's demands neither fairness in the long run.

There are also works that adapt DRF to packet processing [17], [29] and consider a recent past. Nevertheless, this is done to prevent limitations that arise when scheduling packets — in which resources must be shared in *time* — and not to ensure fairness and efficiency in the long run. Finally, other authors have focused on improving efficiency in the long run but not fairness [30], [31]. While some of these works consider users' dynamicity, they do not address fairness in the long run, our focus in this paper.

Tang *et al*. [32] propose an allocation mechanism called Hybrid Multi-Resource Fairness (H-MRF), which looks at long-term fairness in a pay-as-you-use computing system — where users can contribute computing nodes to a resource pool. However, they focus in the setting where users can benefit from speculative execution (*i.e.*, they may try to launch more tasks than they originally needed in the hope that they will prove useful once other tasks finish). Moreover, their model assumes that users value resources in the present as much as in the future. Mahajan *et al*. [33] recently proposed a scheduling framework with the purpose of ensuring long-term fairness. However, differently from SDRF, that looks at long-term fairness for multiple types of resources and workloads, their work focused on allocating GPU among machine learning jobs.

Live Tree can be seen as an alternative implementation of a Kinetic Priority Queue [34]. Different from the classical implementations, however, Live Tree ensures strict upper bounds for priorities that follow an arbitrary continuous function. This happens because, in the classical implementations, elements are swapped — or rotated — in the presence of events [35]. Live Tree does not swap elements, instead, it removes pair of elements and reinsert them according to the update time t. By doing so, it ensures that every update has a cost of $O(\log n)$ for every changing pair of elements, regardless of the number of intersections that happen from the last update to the current. This is particularly useful for SDRF, since the function that calculates weights is not trivial (*i.e.*, Eq. 15 in the Supplemental Appendix A).

A preliminary version of this paper was published in Portuguese [1] and was significantly extended to include experiments using a real cluster, a detailed description and evaluation of the live tree, as well as additional evaluation to better understand how SDRF impacts users with different demands.

9 Conclusion and Future Work

In this paper, we introduced SDRF, an extension of DRF that enforces fairness in the long run. SDRF looks at past allocations and benefits users with lower average usage. We show that SDRF satisfies the fundamental properties of DRF while enforcing fairness in the long run. To efficiently implement SDRF, we introduced live tree, a general-purpose data structure that keeps elements with predictable time-varying priorities ordered. We implemented SDRF on Mesos and showed that it works in a real cluster. Moreover, we simulated SDRF using Google cluster traces for a month-long period. Results have shown that under SDRF, users with low utilization can complete their workloads faster. Meanwhile, users with high utilization suffer a low impact in their overall workload. We also used the simulations to evaluate live tree performance with a real workload and observed that most task allocations do not trigger an event in the tree. This cause live tree to perform close to a red-black tree with static weights.

There are different future investigation directions. First, we believe live tree may benefit other applications, *e.g.*, Dijkstra's algorithm applied to graphs with time-variable weights. Second, SDRF can be extended to cover other applications. Although we mentioned the possibility of using weights for users, we did not evaluate it formally. Another possibility is the use of different utility and commitment functions. For example, in settings where users face strict SLA requirements, having allocations below a certain threshold may disproportionately damage users.

Identifying the utility functions that are more representative in different scenarios and the properties other classes of commitment functions satisfy remains an open problem.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, CNPq, FAPERJ, and FAPESP grants #15/24494-8 and #15/24490-2.

REFERENCES

[1] H. Sadok, M. E. M. Campista, and L. H. M. K. Costa, "O passado também importa: Um mecanismo de alocação justa de múltiplos tipos de recursos ao longo do tempo," in *SBRC*, 2018.
[2] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. Katz, S. Shenker, and I. Stoica, "Mesos: A platform for fine-grained resource sharing in the data center," in *NSDI*, 2011.
[3] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O'Malley, S. Radia, B. Reed, and E. Baldeschwieler, "Apache Hadoop YARN: Yet another resource negotiator," in *SoCC*, 2013.
[4] A. Verma, L. Pedrosa, M. R. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes, "Large-scale cluster management at Google with Borg," in *ACM EuroSys*, 2015.
[5] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, "Heterogeneity and dynamicity of clouds at scale: Google trace analysis," in *ACM SoCC*, 2012.
[6] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang, "Multiresource allocation: Fairnes-efficiency tradeoffs in a unifying framework," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1785–1798, Dec. 2013.
[7] B. Radunovic and J.-Y. Le Boudec, "A unified framework for max-min and min-max fairness with applications," *IEEE/ACM Trans. Netw.*, vol. 15, no. 5, pp. 1073–1083, Oct. 2007.
[8] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, "Dominant resource fairness: Fair allocation of multiple resource types," in *NSDI*, 2011.
[9] J. Jaffe, "Bottleneck flow control," *IEEE Transactions on Communications*, vol. 29, no. 7, pp. 954–962, Jul. 1981.
[10] D. Dolev, D. G. Feitelson, J. Y. Halpern, R. Kupferman, and N. Linial, "No justified complaints," in *ITCS*, 2012.
[11] T. Bonald and J. Roberts, "Multi-resource fairness: Objectives, algorithms and performance," in *ACM SIGMETRICS*, 2015.
[12] P. Bodik, A. Fox, M. J. Franklin, M. I. Jordan, and D. A. Patterson, "Characterizing, modeling, and generating workload spikes for stateful services," in *ACM SoCC*, 2010.
[13] D. C. Parkes, A. D. Procaccia, and N. Shah, "Beyond dominant resource fairness," *ACM Transactions on Economics and Computation*, vol. 3, no. 1, pp. 3:1–3:22, Mar. 2015.
[14] I. A. Kash, G. O'Shea, and S. Volos, "DC-DRF: Adaptive multi-resource sharing at public cloud scale," in *ACM SoCC*, 2018.
[15] W. Wang, B. Liang, and B. Li, "Multi-resource fair allocation in heterogeneous cloud computing systems," *IEEE Transactions on Parallel & Distributed Systems*, vol. 26, no. 10, pp. 2822–2835, 2015.
[16] I. Kash, A. D. Procaccia, and N. Shah, "No agent left behind: Dynamic fair division of multiple resources," *Journal of Artificial Intelligence Research*, vol. 51, no. 1, pp. 579–603, Sep. 2014.
[17] A. Ghodsi, V. Sekar, M. Zaharia, and I. Stoica, "Multi-resource fair queueing for packet processing," in *ACM SIGCOMM*, 2012.
[18] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice-Hall, 1999.
[19] R. Bayer, "Symmetric binary B-Trees: Data structure and maintenance algorithms," *Acta Informatica*, vol. 1, no. 4, pp. 290–306, 1972.
[20] L. J. Guibas and R. Sedgewick, "A dichromatic framework for balanced trees," in *IEEE Symposium on Foundations of Computer Science*, 1978.
[21] "Apache Hadoop," 2020. [Online]. Available: https://hadoop.apache.org/
[22] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *NSDI*, 2012.
[23] A. Toshniwal, S. Taneja, A. Shukla, K. Ramasamy, J. M. Patel, S. Kulkarni, J. Jackson, K. Gade, M. Fu, J. Donham, N. Bhagat, S. Mittal, and D. Ryaboy, "Storm@twitter," in *ACM SIGMOD*, 2014.
[24] A. Gutman and N. Nisan, "Fair allocation without trade," in *AAMAS*, 2012.
[25] E. Friedman, A. Ghodsi, and C.-A. Psomas, "Strategyproof allocation of discrete jobs on multiple machines," in *ACM Conference on Economics and Computation*, 2014.
[26] D. Zarchy, D. Hay, and M. Schapira, "Capturing resource tradeoffs in fair multi-resource allocation," in *IEEE INFOCOM*, 2015.
[27] T. Bonald, J. Roberts, and C. Vitale, "Convergence to multi-resource fairness under end-to-end window control," in *IEEE INFOCOM*, 2017.
[28] E. Friedman, C.-A. Psomas, and S. Vardi, "Controlled dynamic fair division," in *ACM Conference on Economics and Computation*, 2017.
[29] W. Wang, B. Liang, and B. Li, "Low complexity multi-resource fair queueing with bounded delay," in *IEEE INFOCOM*, 2014.
[30] R. Grandl, M. Chowdhury, A. Akella, and G. Ananthanarayanan, "Altruistic scheduling in multi-resource clusters," in *OSDI*, 2016.
[31] C. Chen, W. Wang, S. Zhang, and B. Li, "Cluster fair queueing: Speeding up data-parallel jobs with delay guarantees," in *IEEE INFOCOM*, 2017.
[32] S. Tang, Z. Niu, B. He, B.-S. Lee, and C. Yu, "Long-term multi-resource fairness for pay-as-you use computing systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 5, pp. 1147–1160, May 2018.
[33] K. Mahajan, A. Balasubramanian, A. Singhvi, S. Venkataraman, A. Akella, A. Phanishayee, and S. Chawla, "Themis: Fair and efficient GPU cluster scheduling," in *NSDI*, 2020.
[34] J. Basch, "Kinetic data structures," Ph.D. dissertation, Stanford University, Stanford, CA, USA, Jun. 1999.
[35] G. D. da Fonseca, C. M. H. de Figueiredo, and P. C. P. Carvalho, "Kinetic hanger," *Information Processing Letters*, vol. 89, no. 3, pp. 151–157, Feb. 2004.

Hugo Sadok is a PhD student at Carnegie Mellon University (CMU). He received a MS in Electrical Engineering and a BS in Electronic Engineering, both from Universidade Federal do Rio de Janeiro (UFRJ). Sadok is a student member of ACM and his research interests include datacenters, middleboxes, and scheduling. Contact him at sadok@gta.ufrj.br.

Miguel Elias M. Campista (S'06-M'10-SM'17) is an associate professor at the Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil, since 2010. His major research interests include network science, wireless networking, and cloud and fog computing. Campista received his D.Sc. degree in Electrical Engineering from UFRJ in 2008 and spent 2012 in the LIP6 lab at Sorbonne Université, Paris, France, as invited professor. He is an affiliate member of the Brazilian Academy of Sciences and an associate editor of Annals of Telecommunications, Springer. Contact him at miguel@gta.ufrj.br.

Luís Henrique M. K. Costa (S'97-M'01-SM'16) received his Eng. (Cum Laude) and M.Sc. degrees in electrical engineering from Universidade Federal do Rio de Janeiro (UFRJ), Brazil, and the Dr. degree from Université Pierre et Marie Curie (Paris 6), Paris, France, in 2001. Since August 2004, he has been an associate professor with COPPE/UFRJ. He has served in the TPC of many IEEE conferences for several years and is an associate editor of IEEE Communications Surveys and Tutorials since 2007. His major research interests are in the areas of Internet routing and vehicular networks. Contact him at luish@gta.ufrj.br.

APPENDIX A
SCHEDULING TASKS USING A LIVE TREE

In this appendix, we describe how to use live tree to implement Algorithm 1. In Algorithm 1, we pick the user with the minimum value of $\max_{r\in\mathcal{R}}\{o_{ir}\}+\max_{r\in\mathcal{R}}\{c_{ir}\}$ (line 5), therefore we use a live tree to sort users by this value. Using Eq. 12, we define the priority function p as

$$\begin{aligned} p(t,\kappa_i) &= \max_{r\in\mathcal{R}}\{o_{ir}\} + \max_{r\in\mathcal{R}}\left\{(1-\mathring{\delta})\bar{o}_{ir} + \mathring{\delta}c_{ir}^{(t_i)}\right\} \\ \mathring{\delta} &= e^{-(t-t_i)/\tau} \end{aligned} \tag{13}$$

$\kappa_i = (t_i, \tau, o_{i1}, \ldots, o_{im}, \bar{o}_{i1}, \ldots, \bar{o}_{im}, c_{i1}^{(t_i)}, \ldots, c_{im}^{(t_i)})$, τ is defined as in Eq. 12 and is the same for all users. t_i is the time user i is inserted in the live tree.

To obtain the intersection function we calculate the time when any two arbitrary priorities intersect, *i.e.*, $p(t,\kappa_1) = p(t,\kappa_2)$. To do this, it is useful to define a set of all resource priority intersections, $\mathcal{I}_{ij}$. Whenever two resource priorities from users i and j intersect, the intersection will appear in this set. We derive the expression for the set $\mathcal{I}_{ij}$ calculating the time t that satisfies the equality:

$$\tilde{o}_i + c_{ir_1}^{(t)} = \tilde{o}_j + c_{jr_2}^{(t)}, \text{with } \tilde{o}_i = \max_{r\in\mathcal{R}}\{o_{ir}\}\,.$$

Using Eq. 12 and defining $t_0 = \max\{t_i, t_j\}$,

$$\tilde{o}_i + (1-\mathring{\delta})\bar{o}_{ir_1} + \mathring{\delta}c_{ir_1}^{(t_0)} = \tilde{o}_j + (1-\mathring{\delta})\bar{o}_{jr_2} + \mathring{\delta}c_{jr_2}^{(t_0)}\,.$$

Isolating $\mathring{\delta}$,

$$\mathring{\delta} = \frac{\bar{o}_{ir_1} - \bar{o}_{jr_2} + \tilde{o}_i - \tilde{o}_j}{\bar{o}_{ir_1} - \bar{o}_{jr_2} + c_{jr_2}^{(t_0)} - c_{ir_1}^{(t_0)}}\,.$$

Replacing $\mathring{\delta}$ by $e^{-(t-t_0)/\tau}$,

$$e^{-(t-t_0)/\tau} = \frac{\bar{o}_{ir_1} - \bar{o}_{jr_2} + \tilde{o}_i - \tilde{o}_j}{\bar{o}_{ir_1} - \bar{o}_{jr_2} + c_{jr_2}^{(t_0)} - c_{ir_1}^{(t_0)}}\,.$$

Finally, isolating t,

$$t = t_0 + \tau \ln\left(\frac{\bar{o}_{ir_1} - \bar{o}_{jr_2} + c_{jr_2}^{(t_0)} - c_{ir_1}^{(t_0)}}{\bar{o}_{ir_1} - \bar{o}_{jr_2} + \tilde{o}_i - \tilde{o}_j}\right)\,. \tag{14}$$

Using Eq. 14 we formally define $\mathcal{I}_{ij}$ as

$$\mathcal{I}_{ij} = \left\{\tau\ln\left(\frac{\bar{o}_{ir_1} - \bar{o}_{jr_2} + c_{jr_2}^{(t_0)} - c_{ir_1}^{(t_0)}}{\bar{o}_{ir_1} - \bar{o}_{jr_2} + \tilde{o}_i - \tilde{o}_j}\right)\middle|(r_1,r_2)\in\mathcal{R}^2\right\},$$

where $t_0 = \max\{t_i, t_j\}$ and $\tilde{o}_i = \max_{r\in\mathcal{R}}\{o_{ir}\}$. We define the intersection function getting the minimum intersection after the current time t,

$$t_{\text{int}}(t,\kappa_i,\kappa_j) = \min\{k + t_0 - t | k\in\mathcal{I}_{ij} \wedge k + t_0 > t\}\,. \tag{15}$$

When there is no intersection after the time t, t_{int} does not exist and live tree will add no event. Note this intersection function may indicate intersections between resources that do not cause an intersection in priorities, *i.e.*, commitments may intersect without changing the dominant commitment. Although non-optimal it performs correctly, as false events do not change the order in the tree.

APPENDIX B
DEFERRED PROOFS

In this appendix, we prove the propositions stated in Sec. 4.3 of the paper. Before continuing, we introduce a simple lemma. This lemma states that the allocation a user gets for a certain resource will always be the normalized demand for this resource multiplied by the allocation for the dominant resource. For example, if the normalized demand vector for a user is $\langle 0.5, 1\rangle$ and the allocation for the dominant resource is 0.2, then, the allocation would be $\langle 0.1, 0.2\rangle$.

Lemma 1. *Given an SDRF allocation $\boldsymbol{O}^{(t)}$, obtained by solving Eq. 10, the allocation user i receives for resource r is such that $o_{ir}^{(t)} = \tilde{\theta}_{ir}^{(t)} o_{id}^{(t)}, \forall i\in\mathcal{N}, r\in\mathcal{R}$.*

Proof. From Eq. 10 we know that $o_{ir}^{(t)} \in [0, \hat{\theta}_{ir}^{(t)}]$.

- When $o_{ir}^{(t)} \in (0, \hat{\theta}_{ir}^{(t)})$:

  $$o_{ir}^{(t)} = (x - \tilde{c}_i^{(t)})\cdot\tilde{\theta}_{ir}^{(t)}\,. \tag{16}$$

  By replacing $x - \tilde{c}_i^{(t)} = o_{ir}^{(t)}/\tilde{\theta}_{ir}^{(t)}$ in Eq. 10, and as long as $o_{id}\in(0,\hat{\theta}_{id}^{(t)})$ we get to

  $$o_{id}^{(t)} = o_{ir}^{(t)}/\tilde{\theta}_{ir}^{(t)}\,,$$

  therefore,

  $$o_{ir}^{(t)} = \tilde{\theta}_{ir}^{(t)} o_{id}^{(t)}\,. \tag{17}$$

  Thus we just need to prove that $o_{id}^{(t)} \in (0,\hat{\theta}_{id}^{(t)})$. In fact, $o_{id}^{(t)} > 0$, since we are considering $o_{ir}^{(t)} > 0$ and by definition $\hat{\theta}_{id}^{(t)} > 0$. Verifying the upper bound is also straightforward. We depart from $o_{ir}^{(t)} < \hat{\theta}_{ir}^{(t)}$ and use the definition in Eq. 4,

  $$\begin{aligned} \tilde{\theta}_{ir}^{(t)}\hat{\theta}_{id}^{(t)} &> o_{ir}^{(t)} \\ \hat{\theta}_{id}^{(t)} &> o_{ir}^{(t)}/\tilde{\theta}_{ir}^{(t)} = o_{id}^{(t)}\,. \end{aligned}$$

- When $o_{ir}^{(t)} = 0$:

  $$(x - \tilde{c}_i^{(t)})\cdot\tilde{\theta}_{ir}^{(t)} \le 0$$

  but $\tilde{\theta}_{ir}^{(t)} > 0$, therefore

  $$x - \tilde{c}_i^{(t)} \le 0\,.$$

  Making $x - \tilde{c}_i^{(t)} \le 0$ in Eq. 10 we have

  $$o_{id}^{(t)} = 0,$$

  therefore Eq. 17 still holds.
- When $o_{ir}^{(t)} = \hat{\theta}_{ir}^{(t)}$:

  $$(x - \tilde{c}_i^{(t)})\cdot\tilde{\theta}_{ir}^{(t)} \ge \hat{\theta}_{ir}^{(t)}\,.$$

  Using the definition in Eq. 4,

  $$x - \tilde{c}_i^{(t)} \ge \hat{\theta}_{id}^{(t)}$$

  Making $x - \tilde{c}_i^{(t)} \ge \hat{\theta}_{id}^{(t)}$ in Eq. 10 we get to

  $$\begin{aligned} o_{id}^{(t)} &= \hat{\theta}_{id}^{(t)} \\ &= \hat{\theta}_{ir}^{(t)}/\tilde{\theta}_{ir}^{(t)}\,, \end{aligned}$$

  which is equivalent to Eq. 17 when $o_{ir}^{(t)} = \hat{\theta}_{ir}^{(t)}$, concluding the proof.

□

Corollary 1 (of Lemma 1). *For a given user $i \in \mathcal{N}$, the allocation-demand ratio remains constant for every resource $r \in \mathcal{R}$, i.e., $\min_{r\in\mathcal{R}}\{o_{ir}^{(t)}/\hat{\theta}_{ir}^{(t)}\} = o_{ir}^{(t)}/\hat{\theta}_{ir}^{(t)}, \forall i \in \mathcal{N}, r \in \mathcal{R}$.*

Proof. From Lemma 1 and from Eq. 4,

$$o_{ir}^{(t)}/\hat{\theta}_{ir}^{(t)} = o_{id}^{(t)}/\hat{\theta}_{id}^{(t)}, \forall i \in \mathcal{N}, r \in \mathcal{R}.$$

Therefore,

$$\min_{r\in\mathcal{R}}\{o_{ir}^{(t)}/\hat{\theta}_{ir}^{(t)}\} = o_{ir}^{(t)}/\hat{\theta}_{ir}^{(t)}, \forall i \in \mathcal{N}, r \in \mathcal{R}.$$

□

Corollary 1 implies that, for a user to improve her utility, she must increase the allocation for every resource — otherwise the minimum allocation-demand ratio would not change.

We now turn to the proof of Proposition 1. It shows that if a user did not have her demand fulfilled, there must be at least one resource that is fully utilized — a bottleneck resource.

Proposition 1 (Bottleneck). *The SDRF allocation obtained by solving Eq. 10 is such that all users have their demands fulfilled or there is a bottleneck resource. Formally, $\boldsymbol{o}_i^{(t)} = \hat{\boldsymbol{\theta}}_i^{(t)}, \forall i \in \mathcal{N}$ or $\exists r \in \mathcal{R}$ such that $\sum_{i\in\mathcal{N}} o_{ir}^{(t)} = 1$.*

Proof. Assume, by way of contradiction, that we can obtain an allocation $\boldsymbol{O}^{(t)}$ from Eq. 10 where $\exists i \in \mathcal{N}$ such that $\boldsymbol{o}_i^{(t)} \neq \hat{\boldsymbol{\theta}}_i^{(t)}$ and $\sum_{i\in\mathcal{N}} o_{ir}^{(t)} \neq 1, \forall r \in \mathcal{R}$.

First, from the problem restrictions in Eq. 10, we know that $o_{ir}^{(t)} \leq \hat{\theta}_{ir}^{(t)}, \forall i \in \mathcal{N}, r \in \mathcal{R}$ and $\sum_{i\in\mathcal{N}} o_{ir}^{(t)} \leq 1, \forall r \in \mathcal{R}$. Thus,

$$\sum_{i\in\mathcal{N}} o_{ir}^{(t)} < 1, \forall r \in \mathcal{R} \text{ and} \tag{18}$$

$$\exists i \in \mathcal{N}, r \in \mathcal{R} \text{ such that } o_{ir}^{(t)} < \hat{\theta}_{ir}^{(t)}. \tag{19}$$

We now verify if we can propose a different solution to the problem:

$$x' = x + \epsilon, \epsilon \in \mathbb{R}_+.$$

If we can find a positive ϵ in which x' satisfies the problem constraints this is a contradiction (since $x' > x$ it could have been a solution to Eq. 10 instead of x). We denote by $\boldsymbol{O'}^{(t)}$ the allocation found using x'. If we can obtain a positive ϵ that satisfies the constraints, the following expression should hold,

$$\sum_{i\in\mathcal{N}} o'^{(t)}_{ir} \leq \sum_{i\in\mathcal{N}} \left(o_{ir}^{(t)} + \epsilon\,\tilde{\theta}_{ir}^{(t)}\right) \leq 1, \forall r \in \mathcal{R}\,.$$

Therefore,

$$\epsilon \leq \min_{r\in\mathcal{R}} \left\{ \frac{1 - \sum_{i\in\mathcal{N}} o_{ir}^{(t)}}{\sum_{i\in\mathcal{N}} \tilde{\theta}_{ir}^{(t)}} \right\}.$$

Since we want $\epsilon > 0$, we need to prove that

$$\min_{r\in\mathcal{R}} \left\{ \frac{1 - \sum_{i\in\mathcal{N}} o_{ir}^{(t)}}{\sum_{i\in\mathcal{N}} \tilde{\theta}_{ir}^{(t)}} \right\} > 0.$$

From Eq. 18, we know that

$$1 - \sum_{i\in\mathcal{N}} o_{ir}^{(t)} > 0, \forall r \in \mathcal{R}.$$

Therefore we need to verify if

$$\sum_{i\in\mathcal{N}} \tilde{\theta}_{ir}^{(t)} > 0, \forall r \in \mathcal{R},$$

which is also true since $\theta_{ir}^{(t)} > 0, \forall i \in \mathcal{N}, r \in \mathcal{R}$. Thus we can find an $\epsilon > 0$, making $x' > x$, which is a contradiction. □

Proposition 2 (Strategyproofness in the Stage Game). *When users consider only the stage game utility (Eq. 2), the SDRF allocation obtained by solving Eq. 10 is strategyproof.*

Proof. Take an arbitrary user $j \in \mathcal{N}$, we denote the allocation returned by the mechanism for the user j when $\hat{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j$ by $\boldsymbol{o}_j$, and when $\hat{\boldsymbol{\theta}}_j \neq \boldsymbol{\theta}_j$ by $\boldsymbol{o'}_j$. We must prove that

$$u_j(\boldsymbol{o}_j) \geq u_j(\boldsymbol{o'}_j)\,. \tag{20}$$

We check the following three cases.

- When $u_j(\boldsymbol{o}_j) = \theta_{jd}^{(t)}$: the utility cannot be improved and Eq. 20 trivially holds.
- When $u_j(\boldsymbol{o}_j) = 0$: from Corollary 1, $o_{jr} = 0, \forall r \in \mathcal{R}$. Therefore $x - \tilde{c}_j \leq 0$, and the allocation is zero for any resource, independently from $\hat{\boldsymbol{\theta}}_j$, which makes Eq. 20 true.
- When $0 < u_j(\boldsymbol{o}_j) < \theta_{jd}^{(t)}$: for this case, allocations are not limited by $\boldsymbol{\theta}_j$. We may simplify the expression for the allocation in Eq. 10

  $$o_{jr} = (x - \tilde{c}_j) \cdot \tilde{\theta}_{jr}\,.$$

  From Proposition 1 there is a bottleneck resource, and we denote it as r^*. From Corollary 1 users must increase the allocation for all resources in order to improve their utilities. Therefore, if we can prove that it is impossible to have an alternative demand vector that increases the share of both the bottleneck resource and the dominant resource, we are done.
  We denote by $\tilde{\boldsymbol{\theta}}_j$ the truthful normalized demand vector for user j, and by $\tilde{\boldsymbol{\theta}}'_j$, the misreported normalized demand for user j.
  Increasing the share of dominant resource: The share of dominant resource user j receives when declaring the truth is given by:

  $$o_{jd} = (x - \tilde{c}_j) \cdot \tilde{\theta}_{jd} = x - \tilde{c}_j.$$

  Therefore, to increase the share of dominant resource for user j (*i.e.*, make $o'_{jd} > o_{jd}$), we must make $x' > x$, where x' is the solution to Eq. 10 when user j declares $\tilde{\boldsymbol{\theta}}'_j$ instead of $\tilde{\boldsymbol{\theta}}_j$.
  Increasing the share of bottleneck resource: Since r^* is a bottleneck resource, the following holds:

  $$o_{jr^*} + \sum_{i\in\mathcal{N}\setminus\{j\}} o_{ir^*} = 1$$

  Making $o'_{jr^*} > o_{jr^*}$ requires that

  $$\sum_{i\in\mathcal{N}\setminus\{j\}} o'_{ir^*} < \sum_{i\in\mathcal{N}\setminus\{j\}} o_{ir^*}$$

  and, consequently, since r^* is a bottleneck resource,

  $$\sum_{i\in\mathcal{N}\setminus\{j\}} (x' - \tilde{c}_i)\tilde{\theta}_{ir^*} < \sum_{i\in\mathcal{N}\setminus\{j\}} (x - \tilde{c}_i)\tilde{\theta}_{ir^*}$$

$$\sum_{i\in\mathcal{N}\setminus\{j\}}(x'-\tilde{c}_i) < \sum_{i\in\mathcal{N}\setminus\{j\}}(x-\tilde{c}_i)$$
$$x' < x$$

Therefore it is not possible to increase the share of both the dominant resource and the bottleneck resource, concluding the proof.

□

Proposition 3 (Strategyproofness in the Repeated Game). *When users evaluate their utilities using the expected-long-term utility (Eq. 3), the SDRF allocation obtained by solving Eq. 10 is strategyproof, regardless of users' discount factors.*

Proof. Take an arbitrary user $i \in \mathcal{N}$, assuming user i discounts her utility using δ_i and the allocation mechanism calculates commitments using δ. Without loss of generality, we represent the expected-long-term utility for time $t = 0$ by

$$u_i^{[0,\infty)} = \mathbb{E}_{u_i}\left[\sum_{k=0}^{\infty}\delta_i^k u_i^{(k)}(\boldsymbol{o}_i^{(k)})\right]. \tag{21}$$

Since manipulating the stage game is not possible, the only hope users may have of improving their expected-long-term utility is by reducing their commitments. To reduce their commitments, users may declare a lower demand. We will show that any marginal gain the user may get, does not compensate her loss in the stage game.

If a user i declares a demand $\hat{\theta}_{ir}^{(0)} = \theta_{ir}^{(0)} - \epsilon$, with $0 < \epsilon < \theta_{ir}^{(0)}$, for a resource r, in the best scenario, this will make user i's commitment

$$c_{ir}^{\prime(k)} = c_{ir}^{(k)} - (1-\delta)\delta^k\epsilon, \tag{22}$$

where $c_{ir}^{\prime(k)}$ is the new commitment user i gets by declaring $\hat{\theta}_{ir}^{(0)}$. From this commitment, the maximum possible improvement in the long-term utility is

$$\begin{aligned}\bar{u}_i^{[0,\infty)} &= -\epsilon + \sum_{k=1}^{\infty}\delta_i^k(1-\delta)\delta^k\epsilon \\ &= -\epsilon + \epsilon(1-\delta)\sum_{k=1}^{\infty}(\delta_i\cdot\delta)^k.\end{aligned}$$

Then, replacing the infinite series,

$$\begin{aligned}\bar{u}_i^{[0,\infty)} &= -\epsilon + \epsilon(1-\delta)\left(\frac{1}{1-\delta_i\cdot\delta} - 1\right) \\ &= \epsilon\left(\frac{(1-\delta)}{1-\delta_i\cdot\delta} - (1-\delta) - 1\right).\end{aligned}$$

Inspecting the expression we verify that, for $0 \le \delta < 1$ and $0 \le \delta_i < 1$,

$$\frac{(1-\delta)}{1-\delta_i\cdot\delta} < (1-\delta) + 1$$

and therefore,

$$\bar{u}_i^{[0,\infty)} < 0\,.$$

Thus, any positive decrement ϵ in the declared demands cannot possibly improve the expected-long-term utility, independently from users discount factors. □

Proposition 4 (Non-wastefulness). *The SDRF allocation $\boldsymbol{O}^{(t)}$ is such that, if there is a different allocation $\boldsymbol{O'}^{(t)}$ where $o_{ir}^{\prime(t)} \le o_{ir}^{(t)}, \forall i \in \mathcal{N}, r \in \mathcal{R}$ and for a user $i^* \in \mathcal{N}$ and resource $r^* \in \mathcal{R}$, $o_{i^*r^*}^{\prime(t)} < o_{i^*r^*}^{(t)}$, then it must be that $u_{i^*}^{(t)}(\boldsymbol{o}_{i^*}^{(t)}) > u_{i^*}^{(t)}(\boldsymbol{o'}_{i^*}^{(t)})$. In other words, SDRF is non-wasteful.*

Proof. Assuming truthful demands, using Corollary 1 and the definition in Eq. 2, we have

$$u_i^{(t)}(\boldsymbol{o}_i^{(t)}) = \theta_{id}^{(t)}\cdot o_{ir}^{(t)}/\theta_{ir}^{(t)}, \forall i \in \mathcal{N}, r \in \mathcal{R}\,.$$

Using the definition of $\boldsymbol{O'}^{(t)}$,

$$\begin{aligned}o_{i^*r^*}^{\prime(t)} &< o_{i^*r^*}^{(t)} \\ o_{i^*r^*}^{\prime(t)}\cdot\theta_{i^*d}^{(t)}/\theta_{i^*r^*}^{(t)} &< o_{i^*r^*}^{(t)}\cdot\theta_{i^*d}^{(t)}/\theta_{i^*r^*}^{(t)} \\ &= u_{i^*}^{(t)}(\boldsymbol{o}_{i^*}^{(t)}).\end{aligned}$$

However,

$$\begin{aligned}u_{i^*}^{(t)}(\boldsymbol{o'}_{i^*}^{(t)}) &= \theta_{i^*d}^{(t)}\cdot\min\left\{\min_{r\in\mathcal{R}}\{o_{i^*r}^{\prime(t)}/\theta_{i^*r}^{(t)}\}, 1\right\} \\ &\le \theta_{i^*d}^{(t)}\cdot o_{i^*r^*}^{\prime(t)}/\theta_{i^*r^*}^{(t)}\,.\end{aligned}$$

Therefore,

$$u_{i^*}^{(t)}(\boldsymbol{o'}_{i^*}^{(t)}) < u_{i^*}^{(t)}(\boldsymbol{o}_{i^*}^{(t)}),$$

concluding the proof. □

Proposition 5 (Pareto optimality). *The SDRF allocation obtained by solving Eq. 10 is Pareto optimal.*

Proof. The proof is a direct consequence of Propositions 1 and 4. For the sake of contradiction, assume SDRF is not Pareto optimal, then, from the allocation $\boldsymbol{O}^{(t)}$ obtained by solving Eq. 10 we can get another allocation $\boldsymbol{O'}^{(t)}$ that makes a user $i^* \in \mathcal{N}$ strictly better while making everybody's utility at least as good, *i.e.*,

$$u_{i^*}^{(t)}(\boldsymbol{o'}_{i^*}^{(t)}) > u_{i^*}^{(t)}(\boldsymbol{o}_{i^*}^{(t)}) \text{ and} \tag{23}$$
$$\forall i \in \mathcal{N}, u_i^{(t)}(\boldsymbol{o'}_i^{(t)}) \ge u_i^{(t)}(\boldsymbol{o}_i^{(t)}). \tag{24}$$

From Proposition 4, if $\exists i \in \mathcal{N}, r \in \mathcal{R}$ such that $o_{ir}^{\prime(t)} < o_{ir}^{(t)}$, then $u_i^{(t)}(\boldsymbol{o'}_i^{(t)}) < u_i^{(t)}(\boldsymbol{o}_i^{(t)})$. Therefore we may rewrite Eq. 24 as

$$o_{ir}^{\prime(t)} \ge o_{ir}^{(t)}, \forall i \in \mathcal{N}, r \in \mathcal{R}. \tag{25}$$

Also, from Corollary 1, we know that we must increase the allocation for all resources in order to increase the utility of a user, *i.e.*, for the user i^*,

$$o_{i^*r}^{\prime(t)} > o_{i^*r}^{(t)}, \forall r \in \mathcal{R}.$$

From Proposition 1, either

$$\boldsymbol{o}_i^{(t)} = \hat{\boldsymbol{\theta}}_i^{(t)} \text{ or} \tag{26}$$
$$\exists r^* \in \mathcal{R} \text{ such that } \sum_{i\in\mathcal{N}} o_{ir^*}^{(t)} = 1\,. \tag{27}$$

If Eq. 26 is true, then $u_i^{(t)}(\boldsymbol{o}_i^{(t)}) = \theta_{id}^{(t)}, \forall i \in \mathcal{N}$ and Eq. 23 cannot be true. Therefore Eq. 27 must be true. But since Eq. 27 is true, when we make $o_{i^*r^*}^{\prime(t)} > o_{i^*r^*}^{(t)}$, we must decrease the allocation of another user, contradicting Eq. 25. □

Proposition 6 (Sharing incentives). *The SDRF allocation obtained by solving Eq. 10 satisfies sharing incentives.*

Proof. To prove sharing incentives, it is sufficient to show that there is a strategy, user i may follow, which makes her utility at least as good as $u_i^{(t)}(\langle 1/n, \ldots, 1/n\rangle)$, regardless of other users actions. We show that the strategy "always declare $\hat{\theta}_i^{(t)} = \langle 1/n, \ldots, 1/n\rangle$" guarantees that $o_i^{(t)} = \langle 1/n, \ldots, 1/n\rangle$ for every instant t.

From Eq. 10, $o_{ir}^{(t)} \leq \hat{\theta}_{ir}^{(t)}$ for every $r \in \mathcal{R}$, therefore, $o_{ir}^{(t)} \leq 1/n$. From Eq. 6, this makes $c_{ir}^{(t)} = 0$ for every resource r and instant t, also making $\tilde{c}_i^{(t)} = 0$ (from Eq. 9). Since $\tilde{c}_i^{(t)} = 0$, the allocation received by i will be $o_{ir}^{(t)} = \min\{x, 1/n\}$ for every resource r and instant t. Nevertheless, $x \geq 1/n$, which makes $o_{ir}^{(t)} = 1/n$, concluding the proof. □