Multi-Resource Allocation: Fairness-Efficiency Tradeoffs in a Unifying Framework

Carlee Joe-Wong, Soumya Sen, Tian Lan, and Mung Chiang

***Abstract*—Quantifying the notion of fairness is under-explored when there are multiple types of resources and users request different ratios of the different resources. A typical example is datacenters processing jobs with heterogeneous resource requirements on CPU, memory, network, bandwidth, etc. This paper develops a unifying framework addressing the fairness-efficiency tradeoff in light of multiple types of resources. We develop two families of fairness functions that provide different tradeoffs, characterize the effect of user requests' heterogeneity, and prove conditions under which these fairness measures satisfy the Pareto efficiency, sharing incentive, and envy-free properties. Intuitions behind the analysis are explained in two visualizations of multi-resource allocation. We also investigate people's fairness perceptions through an online survey of allocation preferences and provide a brief overview of related work on fairness.**

I. Introduction

A. Motivation

Comparing fairness of different allocations of a *single* type of resource has been extensively studied. Fairness can be quantified with a variety of metrics, such as Jain's index [1]. Alternatively, different notions of fairness, including proportional and max-min fairness, can be achieved through maximization of α-fair or isoelastic utility functions [2]. These approaches, as well as others from economics and sociology, have recently been unified as the unique family of functions satisfying four axioms for fairness metrics, as summarized in Appendix A of the present work and [3]. The tradeoff between fairness and efficiency has also been studied in [4]–[6].

When it comes to allocating *multiple* types of resources, however, there has been much less systematic study, the recent paper [7] being a notable exception. Indeed, it is unclear what it means to say that a multi-resource allocation is "fair." Each user in a network requires a certain *combination* of different resource types to process one job, and this combination may differ from user to user. For example, datacenters allocate different resources (memory, CPUs, storage, bandwidth, etc.) to competing users with different requirements. One user might have computational jobs requiring more CPU cycles than memory, while another might have the opposite requirements.

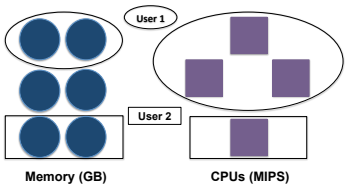


Fig. 1. An example of multi-resource requirements in datacenters.

The need for multi-resource fairness functions can be illustrated with a very simple example, as shown in Fig. 1. In this example, two users require CPUs and memory in order to perform some jobs. User 1 requires 2 GB of memory and 3 CPUs per job, while user 2 needs 2 GB of memory and 1 CPU per job. There is a total of 6 GB of memory and 4 CPUs.

Many allocations might be considered "fair" in this example: should users be allocated resources in proportion to their resource requirements? Or should they be allocated resources so as to process equal numbers of jobs? The fairness measure proposed recently in [7], called **Dominant Resource Fairness (DRF)**, allocates resources according to max-min fairness on dominant resource shares. In this example, DRF would allocate 0.76 jobs to user 1 and 1.71 jobs to user 2, for a total of 2.47 jobs processed. But this allocation brings about a significant loss in system efficiency; e.g., a more unequal allocation of 0.17 jobs to user 1 and 2.83 jobs to user 2 yields a total of 3 jobs. An in-between allocation can be realized if another well-known fairness metric, α-fairness, is adapted for multiple resources following our methods in Section III-B. For $\alpha = 0.5$, user 1 has 0.57 jobs and user 2 has 2.29 jobs, for a total of 2.86 jobs. Each of these allocations represents one point of the fairness-efficiency tradeoff. This paper develops a unifying framework for studying this tradeoff in light of multiple types of resources and heterogeneity in users' resource requirements.

Multi-resource allocation problems arise in increasingly many applications. Datacenters that sell *bundles* of CPUs, memory, storage, and network bandwidth are just one example. In fact, even the classical problem of bandwidth allocation in a congested network can be viewed as a special case of multi-resource allocation. Given a network and its topology, we can view each link as a separate resource with a distinct capacity. Each user is represented by a network flow, which uses a predefined subset of links. In this special case, resource requests on all the links must be the same for each user.

In general, multi-resource allocation *cannot* be trivially turned into single-resource allocation by assuming different resources are interchangeable. For example, if a cloud client needs 2 units of CPU and 5 units of networking bandwidth to

C. Joe-Wong is with the Program in Applied and Computational Mathematics at Princeton University (email:cjoe@princeton.edu). S. Sen and M. Chiang are with the Department of Electrical Engineering, Princeton University (emails: {soumyas, chiangm}@princeton.edu). T. Lan is with the Department of Electrical and Computer Engineering at George Washington University (email: tlan@gwu.edu). An earlier version of part of this work appeared at the 2012 IEEE International Conference on Computer Communications (Infocom).

This work was partly supported by NSF grant CNS-0905086, DARPA grant FA8750-11-C-0254, and AFOSR MURI grant FA9550-09-1-0643. C. Joe-Wong acknowledges support from the NDSEG fellowship.

Manuscript received Month Day 2012.

finish 1 unit of job, adding more does *not* reduce the need for 5 units of bandwidth.

B. Unique Challenges of Multi-Resource Fairness

The following new challenges on fairness arise due to the presence of multiple types of resources:

- In a single-resource scenario, users' resource requirements can be represented with a scalar. With multiple resources, users have vectors of resource requirements, which may all look different and must be scalarized before fairness can be evaluated. We present two ways to visualize user heterogeneity in Section III-A and two methods for this scalarization in Section III-B, yielding parametrized families of multi-resource fairness measures that satisfy the axioms of [3].
- In a single-resource scenario, the most efficient allocation will clearly use the entire resource. In a multi-resource scenario, however, users' heterogeneous resource requirements may not allow each resource to be completely used. Even how to measure efficiency is unclear: should we use the total number of jobs allocated?[1] Or the amount of leftover resource capacity? Section V numerically examines both of these efficiency metrics, while Props. 1 and 2 and their corollaries examine the impact of user heterogeneity on the number of jobs processed.
- The extension of max-min fairness to multiple resources is shown in [7] to satisfy such properties as Pareto-eff iciency for certain parameter values. We characterize the parameterizations under which our multi-resource fairness functions satisfy Pareto-efficiency, sharing incentive, and envy-freeness (Props. 3-5 and their corollaries).
- The existence of a fairness-efficiency tradeoff depends on both the scalarization of users' resource requirements and the subsequent evaluation of fairness. We show that a greater emphasis on equity or fairness need not always decrease efficiency (Prop. 6) and give analytical conditions on when the fairness-efficiency tradeoff exists (Props. 7 and 8 and their corollaries).
- When a fairness-efficiency tradeoff exists, the "best" operating point along this tradeoff depends on the operator's exogenously determined preferences. We characterize this *psychological* component to fairness by conducting a human subject experiment in which participants are asked to rank possible allocation choices given in an online survey. Our results indicate that people tend to cluster into two different groups–one preferring efficiency over fairness and one fairness over efficiency.

After further discussion of related work in Section II, Section III develops our two new families of fairness functions, which we call **Fairness on Dominant Shares (FDS)** and **Generalized Fairness on Jobs (GFJ)**. FDS includes the max-min fairness measure DRF proposed in [7] as a special case. We investigate key properties of these functions in Section IV and characterize conditions under which they are satisf ied by FDS and GFJ. Section V then applies our fairness functions to numerical examples of datacenters. We examine the relationship between the fairness-efficiency tradeoff and FDS and GFJ parameterizations. In Section VI, we experiment with characterizing the parameter values consistent with real people's fairness judgements, analyzing results from an online survey of 143 participants who were asked to rank different possible resource allocations for an example datacenter. All proofs can be found in Appendix B.

[1]The phrases "jobs allocated" and "jobs processed" are used interchangeably throughout the paper.

II. Related Work

Much of the existing theory on the fairness of resource allocations is devoted to allocations of a single resource [3], [8]–[10] (e.g. allocating available link bandwidth to network flows [11]–[14]). The recent work [3] develops the following family of fairness functions for a single resource, unifying previously developed fairness measures. It was proven that this family, parametrized by two numbers, is the *only* family of functions satisfying four simple axioms of fairness metrics:

$$f_{\beta,\lambda}(\mathbf{x}) = \mathrm{sgn}(1-\beta)\left(\sum_{i=1}^{n}\left(\frac{x_i}{\sum_{j=1}^{n} x_j}\right)^{1-\beta}\right)^{\frac{1}{\beta}}\left(\sum_{i=1}^{n} x_i\right)^{\lambda}, \tag{1}$$

where $\beta \in \mathbb{R}$ and $\lambda \in \mathbb{R}$ are parameters. The parameter β gives the "type" of fairness measured by (1), and the parameter λ gives the emphasis on efficiency. A larger $|\lambda|$ indicates greater emphasis on efficiency over fairness. If we take $\lambda = \frac{1-\beta}{\beta}$ and $\beta > 0$, we recover α-fairness for $\alpha = \beta$. In particular, taking the limit as $\beta \to 1$ yields proportional fairness.

Even multi-resource allocation problems, such as scheduling jobs in a datacenter, are often simply treated as a single resource problem (e.g. the Hadoop and Dryad schedulers [15]). A recent paper [7] generalizes the max-min fairness measure to multiple resource settings. Our work develops a unified analytical framework for fairness of multi-resource allocations. In particular, in contrast to [7], we incorporate the tradeoff between fairness and efficiency in multi-resource settings.

Appendix E provides a more comprehensive survey of other work on fairness. In addition to further discussion on fairness in engineering frameworks, we summarize theories of fairness from computer science, economics, political philosophy, and sociology.

III. Fairness-Efficiency of Multi-Resource Allocations

We first present "dual" visualizations of heterogeneity among users' requirements for multiple resources in Section III-A. Section III-B then develops two new families of fairness functions, which scalarize these heterogeneous resource requirement vectors and use them to evaluate the fairness of multi-resource allocations. These two families are Fairness on Dominant Shares (FDS) and Generalized Fairness on Jobs (GFJ). FDS measures the fairness of users' resource allocations by accounting for both the number of jobs allocated to each user (a function of the resources available) and the heterogeneity in different resource requirements across users. GFJ, on the other hand, assumes that users' utility depends solely on the number of jobs they are allocated, irrespective of their differing resource needs.

A. Visualizing User Heterogeneity

A major challenge of multi-resource fairness is incorporating the heterogeneity of different users' requirements for different resources into the assessment of its fairness. Visualizing this heterogeneity can yield useful insights. Moreover, Section V examines in detail how heterogeneity affects the optimal allocation and achieved efficiency.

Figure 2 provides two ways to visualize user heterogeneity. Each user j requires R_{ij} of resource type i for each job.

The first (top) visualization has as many dimensions as there are different types of resources. The axes correspond to the resources (two types of resources here for visual simplicity), with the box representing the resource constraints. The slope σ_i of the line corresponding to each user i is the ratio of that user's requirements for the two resources. The heterogeneity of users' resource requirements can be captured with the variance of the $\{\sigma_i\}$:[2] homogeneity occurs at 0 variance (all users have the same resource requirements) and the dashed line becomes straight. Heterogeneity increases with the variance of σ.

The second (bottom) visualization has as many dimensions as there are different users. The axes correspond to the jobs allocated to each user (two users here for simplicity of drawing), with feasible allocations shown as shaded regions bounded by linear resource constraints. The slopes τ_i of constraint line i reflect the ratio of user 1's and user 2's requirements for resource i. Again, the heterogeneity of users' resource requirements can be captured in the variance of the τ_i. Homogeneity occurs when the variance is 0; in that case the resource constraints have the same slope and reduce to one constraint. Heterogeneity increases with the variance of τ.

B. Defining Multi-Resource Fairness

1) Fairness on Dominant Shares (FDS): As defined in [7], a user's **dominant share** is the maximum share of any resource allocated to that user.

Let x_j denote the number of jobs allocated to each user j and C_i the capacity of each resource i. Then we have the resource constraints $\sum_{j=1}^{n} R_{ij}x_j \leq C_i$ for all resources i, where R_{ij} is the amount of resource i which user j requires for one job, and there are n users. For ease of notation, we define $\gamma_{ij} = R_{ij}/C_i$ as the share of resource i required by user j to process one job. We let

$$\mu_j = \max_i \left\{ \frac{R_{ij}}{C_i} \right\} \tag{2}$$

denote the maximum share of a resource required by user j to process one job; then $\mu_j x_j$ is user j's dominant share.

We introduce the fairness measures $f^{\text{FDS}}_{\beta,\lambda}$:

$$\operatorname{sgn}(1-\beta) \left(\sum_{j=1}^{n} \left(\frac{\mu_j x_j}{\sum_{k=1}^{n} \mu_k x_k} \right)^{1-\beta} \right)^{\frac{1}{\beta}} \left(\sum_{j=1}^{n} \mu_j x_j \right)^{\lambda}. \tag{3}$$

These fairness measures extend those developed in [3] for a single resource; details on their derivation are given in that work. Here $\beta \neq 1$ and λ are pre-specified parameters.

[2]We assume that the σ_i are realizations of a random variable σ.

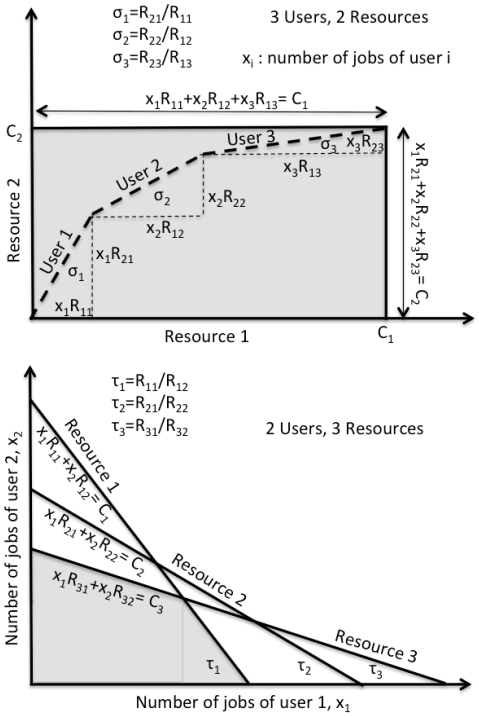


Fig. 2. Two visualizations of user heterogeneity. The lines in the top graph show the ratio of users' requirements for two different resources, while the lines in the bottom graph show the feasible allocation region. The slopes of those lines reflect the ratio of two users' requirements for each resource.

Note that $\beta = 1$ is a trivial case, since (3) then reduces to $n\left(\sum_{j=1}^{n} \mu_j x_j\right)^{\lambda}$, so that each allocation gives equal fairness. We make a standard assumption that all resources and all jobs are infinitely divisible, which is typical of many multi-resource settings [16], [17]. An illustrative example of FDS is given in Section III-B3.

The fairness function (3) may be divided into two components, one representing fairness and one efficiency. The sum of the dominant shares raised to the power λ represents eff iciency; thus, λ parametrizes efficiency's relative importance.

The remainder of (3) is parametrized by β and represents the fairness of the allocation. It is easily seen that for any value of $\beta \neq 1$, this component of (3) is maximized at an equal allocation. However, different values of β will yield different orderings of unequal allocations. One allocation may be more fair than another when $\beta = \beta_1$ is used to parametrize fairness, but the second allocation may be more fair than the first when $\beta = \beta_2 \neq \beta_1$ is used.

Though different values of β give different types of fairness, we can generally say that "larger β is more fair." As $\beta \to \infty$, we obtain max-min fairness on the ratio of each user's dominant share to the sum of all the dominant shares.

As $\beta \to \infty$ and $\lambda = \frac{1-\beta}{\beta}$, the fairness function $f_{\beta,\lambda}$ approaches max-min fairness on the dominant shares. Dominant resource fairness (DRF), proposed in [7], is thus a special case of FDS. Again letting $\mu_j x_j$ denote the dominant share of user j, DRF can be expressed as

$$\min\{\mu_1 x_1,\ \mu_2 x_2, \ldots, \mu_n x_n\}. \tag{4}$$

Maximizing this equation subject to the constraints $\sum_{j=1}^{n} R_{ij}x_j \leq C_i,\ \forall i$, yields the DRF-optimal allocation.

FDS is therefore a generalization of DRF, in which choosing the parameters β and λ allows one to achieve different tradeoffs between fairness and efficiency.

FDS also includes the well-known α-fairness family of functions as a special case. This fact easily follows from the relationship of the single-resource functions in [3] to α-fairness, which is generally used to measure fairness in bandwidth allocation (see references in Section II). Taking $\alpha = \beta \geq 0$ and $\lambda = \frac{1-\beta}{\beta}$, the FDS function (3) becomes

$$\operatorname{sgn}(1-\beta)\left(\sum_{i=1}^{n}(\mu_i x_i)^{1-\beta}\right)^{\frac{1}{\beta}}; \tag{5}$$

optimizing this function is equivalent to optimizing the α-fairness function on dominant shares

$$\sum_{j=1}^{n}\frac{(\mu_j x_j)^{1-\alpha}}{1-\alpha}. \tag{6}$$

2) Generalized Fairness on Jobs (GFJ): Since some users require more resources per job than others, it might be more fair for those who require more resources to be allocated fewer jobs, thus increasing efficiency across all users. FDS captures this perspective. However, an individual user often cares only about the number of jobs processed (without accounting for heterogeneous resource requirements), and hence each user's notion of fairness may be based only on the number of jobs she is allocated. This motivates us to introduce another fairness measure called Generalized Fairness on Jobs (GFJ), which uses the number of jobs allocated (instead of dominant shares) in the fairness function.

GFJ can be further motivated with bandwidth allocation examples. The utility function used in these scenarios is generally α-fairness applied to the bandwidth allocated to each flow. These functions are therefore a special case of GFJ, a family of functions given by

$$f_{\beta,\lambda}^{\mathrm{GFJ}} = \operatorname{sgn}(1-\beta)\left(\sum_{j=1}^{n}\left(\frac{x_j}{\sum_{k=1}^{n}x_k}\right)^{1-\beta}\right)^{\frac{1}{\beta}}\left(\sum_{k=1}^{n}x_k\right)^{\lambda}. \tag{7}$$

Here β and λ are two parameters (just as in FDS) and x_j is the number of jobs processed for user j. As for FDS, we have the resource constraints $\sum_{j=1}^{n} R_{ij}x_j \leq C_i$ for each resource i. An illustrative example is given in the next section.

For $\beta > 0$ and $\lambda = \frac{1-\beta}{\beta}$, GFJ reduces to α-fairness on the number of jobs allocated to each user.

3) Differences between FDS and GFJ: We can summarize FDS' and GFJ's approaches as follows:

- FDS measures fairness in terms of the relative size of the dominant shares, explicitly accounting for heterogeneous resource requirements in both the objective function and the constraints. As a limiting case of FDS, DRF also follows this approach.
- On the other hand, GFJ measures fairness only in terms of the number of jobs allocated to each user; the heterogeneity in resource requirements only appears in the resource constraints. Users requiring more resources are thus treated equally, a result observed in Section V.

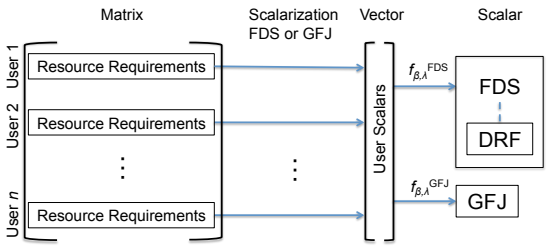


Fig. 3. Overall schematic of our multi-resource fairness approach.

When $\mu_j = \mu$ for all j, FDS and GFJ are equivalent.

Revisiting the example in the Introduction, we have the resource constraints $2x_1 + 2x_2 \leq 6$ and $3x_1 + x_2 \leq 4$. Thus, the dominant share of user 1 is $\frac{3}{4}x_1$, since user 1 requires $\frac{3}{4}$ of the available CPUs and $\frac{1}{3}$ of the available memory for each job. Similarly, the dominant share of user 2 is $\frac{1}{3}x_2$, since user 2 requires $\frac{1}{3}$ of the available memory and $\frac{1}{4}$ of the available CPUs for each job. FDS and GFJ can then be expressed as

$$\begin{aligned}&\max_{x_1,x_2} f(x_1,x_2)\\ &\text{s.t. } 2x_1+2x_2\leq 6,\ 3x_1+x_2\leq 4,\end{aligned} \tag{8}$$

where the fairness function is

$$f = \operatorname{sgn}(1-\beta)\left(\frac{\left(\frac{3x_1}{4}\right)^{1-\beta}+\left(\frac{x_2}{3}\right)^{1-\beta}}{\left(\frac{3x_1}{4}+\frac{x_2}{3}\right)^{1-\beta}}\right)^{\frac{1}{\beta}}\left(\frac{3x_1}{4}+\frac{x_2}{3}\right)^{\lambda}$$

for FDS and

$$f = \operatorname{sgn}(1-\beta)\left(\frac{x_1^{1-\beta}+x_2^{1-\beta}}{(x_1+x_2)^{1-\beta}}\right)^{\frac{1}{\beta}}(x_1+x_2)^{\lambda}$$

for GFJ.

Figure 3 illustrates the approaches to multi-resource fairness. We transpose the matrix $\mathbf{R}$ to capture users' resource requirements; each row represents one user's requirements. One simplistic approach would assume perfectly substitutable resources; in that case, this matrix immediately collapses into a vector of users' single resource requirements. However, this substitutability often does not hold. For example, CPUs and memory are not directly substitutable.

FDS and GFJ represent alternative approaches to the scalarization of each row in Fig. 3's matrix. FDS and its limiting case DRF choose a dominant entry from the row vector of users' requirements. GFJ, on the other hand, scalarizes each row by the number of jobs processed with a bundle of different resources. These row-by-row scalarizations then yield another vector of users' scalars; evaluating fairness with $f_{\beta,\lambda}^{\mathrm{FDS}}$ or $f_{\beta,\lambda}^{\mathrm{GFJ}}$ further reduces this vector to a final scalar quantifying fairness.

IV. Properties of FDS and GFJ

In this section, we prove key properties of the FDS and GFJ functions introduced above. Section IV-A characterizes the optimal fairness values in certain special cases, while Section IV-B examines the conditions of β and λ under which FDS and GFJ satisfy important properties relevant to fairness quantif ication and fairness-efficiency tradeoffs:

- What happens to the optimal allocations when users have the same resource requirements?

- What fairness properties do FDS and GFJ satisfy? For instance, are their optimal allocations Pareto-efficient? Sharing incentive compatible? Envy-free?
- Does there always exist a fairness-efficiency tradeoff?

Finally, Section IV-C examines the conditions under which a fairness-efficiency tradeoff exists.

We consider n users and m different resources. Users have the same resource requirements when they are homogeneous, i.e., their heterogeneity is zero. In the special cases $n = 2$ or $m = 2$, user heterogeneity may be easily visualized as in Fig. 2 in Section III-A. We use the term *user-resource system* to refer to a given set of resources and users with associated resource requirements and capacities.

A. Values of FDS and GFJ

Heterogeneity is measured by the variance in the slopes σ_i or τ_i of Fig. 2. When all users have the same ratios of multi-resource requirements (i.e., the variance of the $\{\sigma_i\}$ and $\{\tau_i\}$ is zero), the problem reduces to that of a single resource:

Proposition 1 (Reduction to Single-Resource Case): Suppose that the resource constraints may be written as

$$\eta_i \left(\mu_1 x_1 + \mu_2 x_2 + \ldots + \mu_n x_n\right) \leq 1, \tag{9}$$

$i = 1, 2, \ldots, m$. Let $\eta_{\max} = \max_i \eta_i$. Then the problem reduces to single-resource fairness on resource 1. Moreover, FDS and DRF both yield the allocation $x_j = \frac{1}{\eta_{\max}\mu_j n}$. GFJ yields the allocation $x_j = \dfrac{\mu_j^{-\frac{1}{\beta}}}{\eta_{\max}\sum_{i=1}^n \mu_i^{\frac{\beta-1}{\beta}}}$.

Definition 1 (Efficiency): Let $X = x_1 + x_2 + \ldots + x_n$ denote the allocation efficiency.

In this special case, we also have the following corollary:

Corollary 1: For allocations that maximize DRF and FDS,

$$\frac{\partial X}{\partial \mu_j} = \left(\frac{-1}{n\eta_{\max}}\right)\left(\frac{1}{\mu_j^2}\right)$$

and the efficiency of these allocations increases the fastest if $\min_j \mu_j$ is decreased. For allocations that maximize GFJ,

$$\frac{\partial X}{\partial \mu_j} = \frac{-\mu_j^{-\frac{1+\beta}{\beta}}}{\eta_{\max}\beta\sum_{i=1}^n \mu_i^{\frac{\beta-1}{\beta}}} + \frac{(1-\beta)\mu_j^{-\frac{1}{\beta}}\sum_{i=1}^n \mu_i^{-\frac{1}{\beta}}}{\eta_{\max}\beta\left(\sum_{i=1}^n x_i^{\frac{\beta-1}{\beta}}\right)^2}.$$

In other words, the system's efficiency will increase if the user with the lowest μ_j gives up some resources.

We now consider heterogeneous users, and assume that their resource requirements R_{ij} are uniformly distributed in $[0, \nu C_i]$, ν a given positive constant. Then, as the number of users n goes to infinity, the optimal FDS and GFJ values converge as follows:

Proposition 2 (Optimal FDS and GFJ Values): The optimal FDS value converges in probability as

$$\lim_{n\to\infty} \left(\max f_{\infty,-1}^{\text{FDS}}\right)^{-1} \cdot \frac{2m}{n(m+1)} = 1. \tag{10}$$

Thus, users' asymptotic dominant share is $\frac{1}{n}\cdot\frac{2m}{m+1}$. In contrast, the optimal GFJ value converges in probability as

$$\lim_{n\to\infty} \left(\max f_{\infty,-1}^{\text{GFJ}}\right)^{-1} \cdot \frac{2}{\nu\left(\sqrt{mn/3}+n\right)} = 1. \tag{11}$$

Users are asymptotically allocated resources for $\frac{2}{\nu n}$ jobs. We note that ν appears in (11) but not (10), since the dominant shares, not the number of jobs, appear in the FDS objective function. Scaling the resource requirements R_{ij} by ν is equivalent to scaling the optimal allocations x_j by ν^{-1}; these cancel in calculating the dominant shares $\mu_j x_j$.

We thus see that in the limit of a large number of heterogeneous users, with $\beta = \infty$ and $\lambda = -1$, the optimal FDS value increases while the optimal GFJ value decreases as more resources are added to the system. This proposition highlights the fundamental difference between FDS and GFJ: in the limit, they yield very different allocations.

B. Three Key Properties of Fairness

We next turn our attention to fairness and its relationship with efficiency, using three widely-used properties of fairness functions (see e.g., [7] and the many references therein):

Definition 2: A function f is **Pareto-efficient** if, whenever $\mathbf{x}$ Pareto-dominates $\mathbf{y}$ (i.e., $x_i \geq y_i$ for each index i and $x_j > y_j$ for some j), $f(\mathbf{x}) > f(\mathbf{y})$.

Definition 3: **Sharing incentive** is the property that no user's dominant share is less than $\frac{1}{n}$; each user has an incentive not to simply split the resources equally.

Definition 4: **Envy-freeness** holds if and only if no user envies another user's allocation. Mathematically, let r_{ij} denote the amount of resource i allocated to user j. User j can then process $\max_i r_{ij}/R_{ij}$. Envy-freeness is defined as the property that $\max_i r_{ij}/R_{ij} > \max_i r_{ik}/R_{ij}$ for any $j \neq k$. In words, no other user's allocation would enable a user to process more jobs than her allocation would.

We investigate if and when these properties are satisfied by FDS and GFJ. Our results show that the answer depends on several factors, e.g. the values of the parameters β and λ. Table I summarizes our findings.

We first consider Pareto-efficiency. Evidently, this property holds for large λ. Based on [3], we can in fact specify a threshold for λ above which Pareto-efficiency holds:

Proposition 3 (Pareto-efficiency of FDS and GFJ): The fairness functions (3) and (7) are Pareto-efficient when $\beta > 0$ if and only if $|\lambda| \geq \left|\frac{1-\beta}{\beta}\right|$.

The absolute value signs are necessary, as for $\beta > 1$, (3) and (7) are negative. For this range of β, a more negative λ therefore emphasizes efficiency. As Pareto-efficiency is a highly desirable property for fairness functions (both single and multi-resource), the following analysis considers only values of λ satisfying $|\lambda| \geq \left|\frac{1-\beta}{\beta}\right|$.

Proposition 4 (Sharing Incentive of FDS): Suppose $\beta > 0$. Then we can prove the following:

(a) Sharing incentive is satisfied by the FDS-optimal allocation when $\lambda = \frac{1-\beta}{\beta}$ and $\beta > 1$.
(b) For $0 < \beta < 1$ and $\lambda = \frac{1-\beta}{\beta}$, there exists a user-resource system such that the FDS-optimal allocation for this system does not satisfy the sharing incentive property.
(c) For any $\beta > 0$, there exists λ with $|\lambda|$ sufficiently large so that for some user-resource system, the FDS-optimal allocation need not satisfy the sharing incentive property.
(d) If $\lambda = 0$, then the FDS-optimal allocation always satisfies the sharing incentive property.

We can further bound the allocation efficiency:

Corollary 2 (Bounds on Allocation Efficiency of FDS): If $\beta > 0$ and $\lambda = \frac{1-\beta}{\beta}$, the efficiency $X \geq \frac{1}{\max_j \mu_j}$.

For $\lambda = \frac{1-\beta}{\beta}$, the FDS function becomes equivalent to the isoelastic α-fair utility in economics; β corresponds to a measure of constant relative risk-aversion for individual users.[3] As β increases, individual risk-averse users find the resource allocation more equitable and become collectively envy-free. The following corollary establishes that this interesting envy-free behavior emerges (for FDS) at a *threshold* of $\beta > 1$:

Corollary 3 (Envy-Freeness of FDS): For $\beta > 0$ and $\lambda = \frac{1-\beta}{\beta}$, the envy-freeness property holds if $\beta > 1$; if $\lambda = 0$, then envy-freeness holds for all user-resource systems and any β. Moreover, there exists a user-resource system whose FDS-optimal allocation does not satisfy envy-freeness under the same conditions (b) and (c) in Prop. 4 for which the sharing incentive property does not always hold.[4]

In contrast to FDS, GFJ need not always satisfy sharing incentive even for $\beta > 1$:

Proposition 5 (Sharing Incentive of GFJ): Suppose again that $\beta > 0$. Then under the conditions enumerated below, there exists a user-resource system whose GFJ-optimal allocation does not satisfy the sharing incentive property:
(a) $|\lambda| = |(1-\beta)/\beta|$,
(b) $|\lambda| > |(1-\beta)/\beta|$ and $0 < \beta < 1$,
(c) $|\lambda| < |(1-\beta)/\beta|$ and $\beta > 1$,
(d) $|\lambda|$ sufficiently large,
(e) $\lambda = 0$.

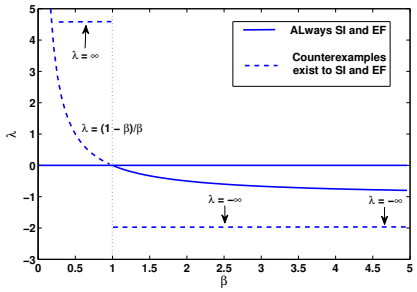
Similarly, GFJ-optimal allocations need not be envy-free for any value of β:

Corollary 4 (Envy-Freeness of GFJ): Under the conditions specified in Prop. 5, there exists a user-resource system such that envy-freeness does not hold for the GFJ-optimal allocation.

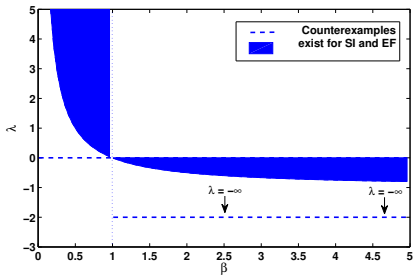
Figure 4 illustrates Props. 4 and 5's results on the sharing incentive property, as well as Corollaries 3 and 4's results on envy-freeness.

[3]**Isoelasticity** and **relative risk-aversion** in economics are defined as $\frac{\partial u(x)}{\partial x}\frac{x}{u(x)}$ and $-\frac{xu''(x)}{u'(x)}$ respectively, where u is the utility function.

[4]Though it may appear so from this proposition, sharing incentive and envy-freeness are *not* equivalent [7].



(a) FDS.



(b) GFJ.

Fig. 4. Conditions under which sharing incentive (SI) and envy-freeness (EF) can be shown either to hold or not to hold (c.f. Props. 4 and 5 and their corollaries 3 and 4).

C. Fairness-Efficiency Tradeoff

We now consider two ways in which a fairness-efficiency tradeoff does not exist: first, an increased emphasis on fairness need not decrease efficiency. Second, the efficiency-maximizing allocation may also be the "most fair."

Traditionally, a larger parameter α in α-fairness functions is thought to be "more fair" [18], [19]; this statement is made mathematically precise in [3]. In [11], however, it is shown that when a network allocates bandwidth so as to maximize α-fairness, total throughput in the network will sometimes increase with α. It may even decrease as capacity increases. These "counter-intuitive" results hold in the general multi-resource problem:

Consider the general family of utility functions $U(\mathbf{x}, \alpha)$; here α is a parameter indexing the family of utility functions, and the specific functional form of U is not specified. For instance, we could use the functions in (3), with $\alpha = \beta$ and $\lambda = \frac{1-\beta}{\beta}$, so that the utility function uses "α-fairness." We incorporate the resource capacity constraints in the matrix inequality $\mathbf{Rx} \leq \mathbf{C}$ and assume that $\mathbf{R}$ is a matrix of full row rank consisting only of those constraints which are tight at the optimal allocation $\mathbf{x}$ for the given value of α.

We let $\mathbf{S}$ be an $n \times (n-m)$ dimensional matrix whose columns form a basis for the nullspace of $\mathbf{R}$, and again let $X = \sum_{j=1}^{n} x_j$ denote the total efficiency. The negative of the utility function's Hessian matrix is denoted by $\mathbf{D}$, and we def ine $\mathbf{b} = \frac{\partial^2 U}{\partial x \partial \alpha}$, $\mathbf{A} = \mathbf{S}^T\mathbf{D}\mathbf{S}$, $\mathbf{v}_j = \mathbf{s}_j{}^T\mathbf{b}$ and $\beta_j = -\mathbf{1}^T\mathbf{s}_j$, where the $\mathbf{s}_j$ are the columns of the matrix $\mathbf{S}$. Let $\overline{\mathbf{A}}_i$ denote the matrix $\mathbf{A}$ with the ith row replaced by $\beta = [\beta_1\ \beta_2 \cdots \beta_n]$. We use δ to denote a direction of perturbation of the capacity vector $\mathbf{C}$ and $\mathbf{D}X(\delta)$ to denote the derivative of X in the

Fairness	Sharing Incentive		Envy-Freeness	
FDS	$\lambda = \frac{1-\beta}{\beta}, 0 < \beta < 1$	$\lambda = \infty$, any β	$\lambda = \frac{1-\beta}{\beta}, 0 < \beta < 1$	$\lambda = \infty$, any β
GFJ	$\lambda = \frac{1-\beta}{\beta}, \beta > 0$ $\lvert\lambda\rvert < \frac{\lvert 1-\beta\rvert}{\beta}, \beta > 1$	$\lambda = \infty$ or 0, any β $\lvert\lambda\rvert > \frac{\lvert 1-\beta\rvert}{\beta}, 0 < \beta < 1$	$\lambda = \frac{1-\beta}{\beta}, \beta > 0$ $\lvert\lambda\rvert < \frac{\lvert 1-\beta\rvert}{\beta}, \beta > 1$	$\lambda = \infty$ or 0, any β $\lvert\lambda\rvert > \frac{\lvert 1-\beta\rvert}{\beta}, 0 < \beta < 1$

TABLE I
CONDITIONS UNDER WHICH PROPERTIES DO NOT HOLD FOR ALL USER-RESOURCE SYSTEMS.

direction of δ. From [11], we have

$$\frac{\partial X}{\partial \alpha} = \mathbf{1}^T \mathbf{S} \mathbf{A}^{-1} \mathbf{S}^T \mathbf{b} \tag{12}$$

$$\mathbf{D}X(\delta) = \mathbf{1}^T \frac{\partial x}{\partial C} \delta = \mathbf{1}^T \mathbf{D}^{-1} \mathbf{R}^T (\mathbf{R}\mathbf{D}^{-1}\mathbf{R}^T)^{-1} \delta. \tag{13}$$

We can further prove the following proposition:

Proposition 6 (Efficiency Non-Monotonicity): Efficiency increases with α if and only if

$$\sum_{i=1}^{N-L} v_i \det \overline{\mathbf{A}}_i \geq 0. \tag{14}$$

Moreover, efficiency may decrease with an increase in the capacity vector $\mathbf{C}$. If capacity increases proportionally, i.e., $\delta = \epsilon \mathbf{C}$ for some small ϵ, then $\mathbf{D}X(\delta) \geq 0$.

As a special case, when only one capacity constraint is tight (e.g., one resource), efficiency always increases with capacity. Appendix C-B contains a numerical example in which eff iciency increases with β.

We next examine the conditions under which an equal allocation (equal dominant shares for FDS or an equal number of jobs for GFJ) maximizes efficiency. In these situations, there is no fairness-efficiency tradeoff; the most fair allocation maximizes the total number of jobs processed. As this property is an ideal case, it will likely be satisfied only under rather stringent conditions. Indeed, our results show that this ideal case occurs only when the resource constraints "line up" exactly.

We again express the resource constraints in matrix form as $\mathbf{R}x \leq \mathbf{C}$, and simplify them to $\gamma \mathbf{x} \leq \mathbf{1}_m$, where $\mathbf{1}_m$ is a vector of m 1's and $\gamma_{ij} = \frac{R_{ij}}{C_i}$.

Proposition 7 (Maximizing Fairness and Efficiency (I)): Suppose that $m = n$ constraints are tight at the maximum-eff iciency allocation. Then this allocation equalizes the dominant shares (FDS has no fairness-efficiency tradeoff) if and only if

$$\sum_{j=1}^{n} \frac{\gamma_{ij}}{\mu_j} = \rho \tag{15}$$

for some constant ρ and all resources i. The number of jobs per user is equalized (GFJ has no fairness-efficiency tradeoff) if

$$\sum_{j=1}^{n} \gamma_{ij} = r \tag{16}$$

for some constant r and all resources i.

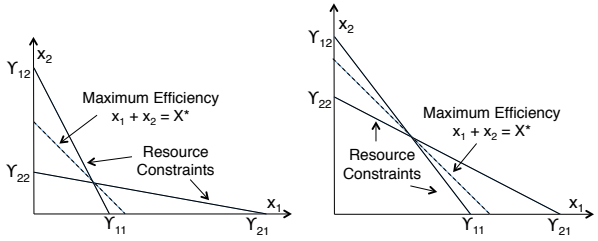


Fig. 5. Illustration of Prop. 7.

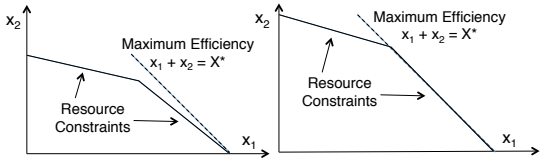


Fig. 6. Illustration of Prop. 8 in two dimensions. In the top graph, exactly one resource constraint is tight at the unique efficiency-maximizing allocation, and $x_2 = 0$. In the bottom graph, exactly one resource constraint is tight at any of the multiple efficiency-maximizing allocations.

Looking back at Fig. 2, we see that the number of jobs per user is equal at the efficiency-maximizing allocation if $\sigma_1 = \ldots = \sigma_n$ for n users and two resources. For two users and m resources, the number of jobs per user is equal at the eff iciency-maximizing allocation if $\sum_{j=1}^{m} \tau_j R_{j2} = \sum_{j=1}^{m} R_{j2}$.

Our conclusions are more subtle when $m < n$ constraints are tight at an efficiency-maximizing allocation:

Proposition 8 (Maximizing Fairness and Efficiency (II)): Suppose that $m < n$ constraints are tight at an efficiency-maximizing allocation $\mathbf{x}^*$. If this allocation is the unique allocation maximizing efficiency, then at least one of the $x_j^* = 0$ and one user is allocated no jobs. If other allocations also maximize efficiency, an allocation equalizing either the dominant shares or number of jobs processed maximizes eff iciency if and only if at the equal allocation, the constraint set intersects the hyperplane $\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} x_j^*$ on a set of dimension at least 1.

Figure 6 shows the two-dimensional illustration of this theorem's statements. The top graph shows a unique efficiency-maximizing allocation when exactly one resource constraint is tight, and the bottom graph shows a set of multiple efficiency-maximizing allocations.

We can use this proposition to derive a sufficient condition for the efficiency-maximizing allocation to equalize the dominant shares or number of jobs for each user:

Corollary 5: Suppose $m < n$ resource constraints hold at the efficiency-maximizing allocation. Then if $R_{ij} > R_{ik}$ for some users j and k and all resources i, $x_j = 0$ (user j is allocated no jobs) at any efficiency-maximizing allocation.

If $m = 1$ (the single-resource case), this result implies the following:

Corollary 6: The maximum efficiency allocation equalizes the dominant shares (FDS) or jobs per user (GFJ) if and only if $\mu_j = \mu \; \forall$ users j. In other words, each user needs the same amount of the single resource to process one job.

V. Applications and Illustrations

We consider an illustrative example of a datacenter with CPU and RAM constraints. There are two users, each of whom requires a fixed amount of each resource to accomplish a job. Jobs are assumed to be infinitely divisible [16], [17]. In order to benchmark performance, we use the same parameters as [7]: user 1 requires 1 CPU and 4 GB of RAM for each job, and user 2 requires 3 CPUs and 1 GB of RAM for each job. There are 9 CPUs and 18 GB of RAM at first. We then vary these constraint values to observe their impact on fairness.

Suppose that the fairness function is given by f (e.g. FDS (3), DRF (4), GFJ (7)). Then the allocation problem is

$$\max_{x,y} f(x, y) \tag{17}$$

$$\text{s.t. } x + 3y \leq 9, \; 4x + y \leq 18 \tag{18}$$

where x and y are the number of jobs allocated to users 1 and 2 respectively.

We use DRF as the benchmark fairness to compare the performance of our FDS and GFJ functions. We define **percent fairness** as the percentage difference between the optimal DRF fairness value (i.e., the minimum dominant share) and the DRF fairness value of the allocation obtained from FDS or GFJ. The **percent efficiency** is defined as the percentage difference between the total number of jobs processed in the given allocation and the maximum number of jobs that can be processed, given the same capacity constraints. We also introduce another efficiency measure, the **leftover capacity** (i.e., the amount of unused resources).

We investigate the outcomes of the proposed fairness measures along two dimensions:

- Comparing the achieved efficiency when user heterogeneity and resource capacity are varied.
- Examining the range of attainable fairness-efficiency tradeoffs for different values of the parameters β and λ.

A. Efficiency

We first use our two efficiency measures–leftover capacity and percent efficiency–to investigate user heterogeneity's effect on achieved efficiency. Heterogeneity is measured by the variance in the slopes τ_i and in the slopes σ_i of users' resource requirements, as introduced in Fig. 2 in Section III-A. If two users have identical resource requests, they become homogeneous, and both variances are 0. At the other extreme, the users do not share any resource requirements; they become decoupled, with infinite variances.

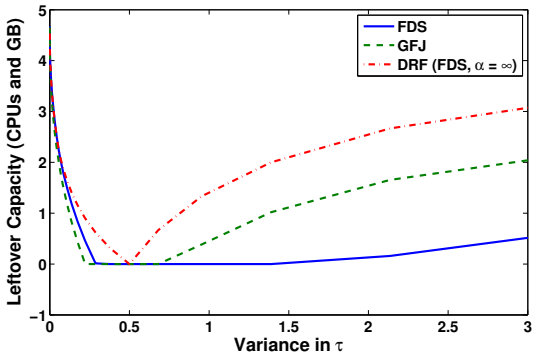


Fig. 7. *Too much or too little variance in τ leads to inefficiency from leftover capacity:* Leftover capacity versus variance in user heterogeneity in a datacenter example. Variances below 0.5 have only leftover CPUs; variances above 0.5 have only leftover RAM.

We calculate the optimal FDS, GFJ and DRF allocations for $\beta = 2$, $\lambda = -0.5$. First, Fig. 7 examines the leftover capacity as a function of the variance in τ. The heterogeneity was varied by changing the RAM requirement of user 2 from 1 GB to 13 GB. Thus, the RAM constraint line in Fig. 2's representation tilts from very steep to very flat. This tilting geometrically explains the overall "V" trend in Figs. 7 and 8. When the RAM requirement is below 3 GB (a steep constraint line), the variance of τ is over 0.5 and the variance of σ is over 4.5: only RAM is leftover. When the RAM requirement is above 3 GB (a flatter line), the variance of τ is less than 0.5 and the variance of σ less than 4.5: only CPUs are leftover. The change in the leftover resource is due to the changing shape of the feasible region.

In this example, we see that for low heterogeneity in users' resource requirements, FDS, GFJ, and DRF have similar eff iciency values. In fact, Prop. 1 states that at zero heterogeneity, DRF and FDS are optimized at the same allocation, predicting part of the observed behavior. As the heterogeneity increases, DRF has a lot of leftover capacity compared to GFJ and FDS, especially for a variance larger than 1 in Fig. 7 and larger than 5 in Fig. 8. DRF trades off efficiency signif icantly to preserve users' minimum dominant share with increasingly heterogeneous resource requirements. Even GFJ performs worse than FDS, which yields the lowest leftover capacity. As FDS includes resource requirements in its fairness function, we intuitively expect such a result.

We next examine the percent efficiency in jobs processed as a function of the variances in τ and σ in Figs. 9 and 10. As in the previous figures, for low heterogeneity across users' resource requirements, FDS, GFJ, and DRF perform at similar efficiency levels. All three achieve full efficiency for a τ variance near 0.5 and σ variance near 4.5. Again, the eff iciency attained is also much higher (about 15%) for FDS and GFJ than for DRF as the variance increases.

In summary, enforcing DRF can significantly reduce eff iciency as measured by either leftover capacity or percent efficiency. This is also the case when the number of users grows; Fig. 11 shows the leftover capacity versus the number

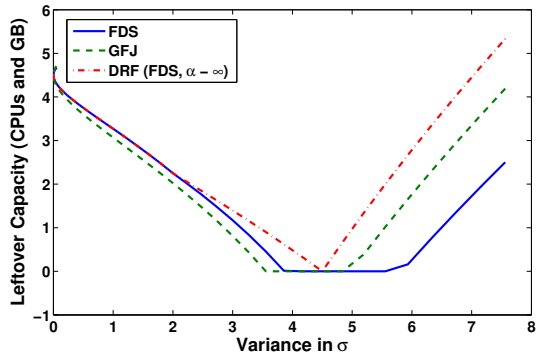


Fig. 8. *Too much or too little variance in σ leads to inefficiency from leftover capacity:* Leftover capacity versus variance in user heterogeneity in a datacenter example. Variances below 4.5 have only leftover CPUs; variances above 4.5 have only leftover RAM.

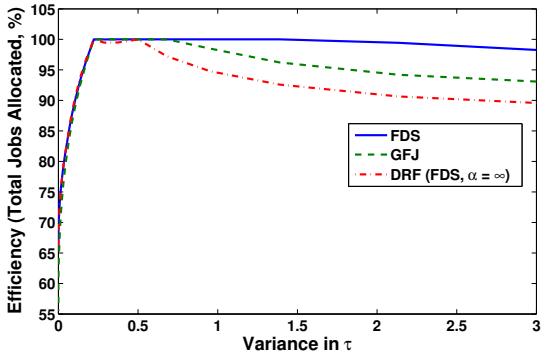


Fig. 9. *Greater variance in τ leads to DRF inefficiency in the number of jobs processed:* Percentage efficiency versus variance in user heterogeneity in a datacenter example.

of users in the system. Only RAM capacity was leftover; in all scenarios, all of the CPUs were used. For a large number of users, we see that FDS and GFJ both use more capacity than DRF. Users' CPU requirements were fixed at 2 CPUs; their RAM requirements were drawn from a uniform distribution. Other randomly chosen RAM requirements yield similar plots.

Finally, we examine the impact of changing RAM capacity on the attainable efficiency levels. Figure 12 shows how varying this capacity affects the efficiency attained at the optimal allocation. We see that when the dominant shares for both users are equal, at 12 GB of RAM capacity, GFJ and FDS have the same range of achievable efficiency. Moreover,

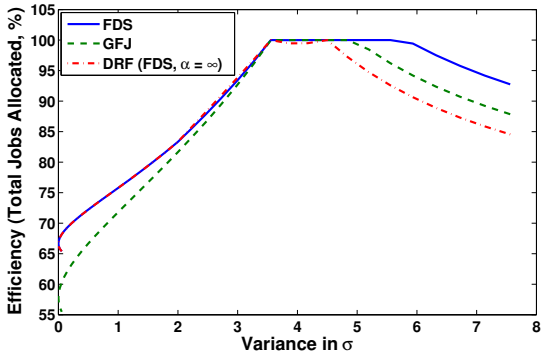


Fig. 10. *Greater variance in σ leads to DRF inefficiency in the number of jobs processed:* Percentage efficiency versus variance in user heterogeneity in a datacenter example.

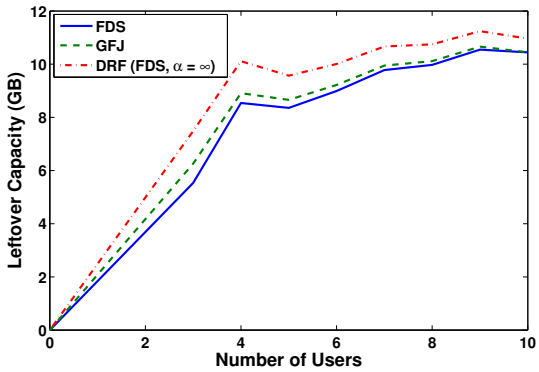


Fig. 11. *Even with a large number of users, DRF uses less available capacity than FDS and GFJ:* Leftover capacity versus the number of users in a datacenter example.

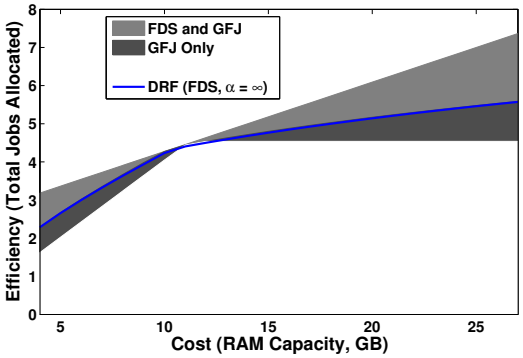


Fig. 12. *Capacity expansion can increase the range of operating eff iciencies for FDS and GFJ over DRF:* Attainable efficiency for varying capacity constraint, given different implicit realizations of $\beta \in (-5, 5)$ and $\lambda \in (0.01, 1.91)$ for $\beta < 0$, $\lambda \in \left(0.005\left(\frac{1}{\beta} - 2\right), 0.955\left(\frac{1}{\beta} - 2\right)\right)$ for $\beta > 0$ values. The region labels refer to the fairness functions that attain those efficiencies.

β and λ can be chosen to achieve higher efficiency in FDS and GFJ. The DRF function serves as a "lower bound" to the efficiency values attainable with the FDS functions.

The impact of capacity expansion also highlights an interesting dimension of the *economy of scale* in large networks. The standard view is that a large scale helps smooth out temporal fluctuations of demands through statistical multiplexing, e.g., at any aggregation point in a broadband access network. In addition to temporal "heterogeneity" (bursting at different times), network users may have *resource type heterogeneity*: some applications need more CPU processing while others need more storage or bandwidth. Can this heterogeneity be exploited to utilize different types of resources more eff iciently? The answer depends on how these different resources are allocated among the users. If DRF is used, for example, efficiency can be quite low. However, by using the appropriate FDS parametrization, resource request heterogeneity can indeed be leveraged along with increases in resource capacity and turned into another type of economy of scale.

B. Fairness-Efficiency Tradeoffs

The previous section established that when users are very heterogeneous, FDS and GFJ outperform DRF, achieving a much greater efficiency. However, we expect that this larger efficiency comes at a cost of decreased fairness. This section

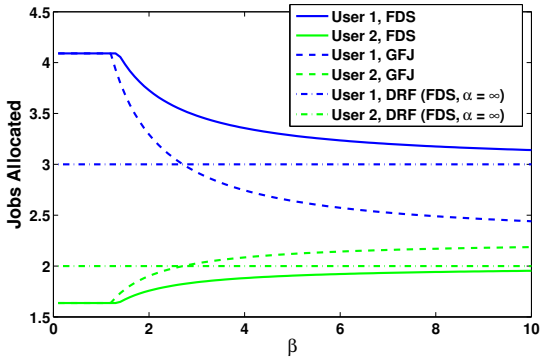


Fig. 13. *Larger β values lead to more equitable allocations:* Optimal allocations for various fairness measures in a datacenter example, using $\alpha = \beta$ fairness for FDS and GFJ.

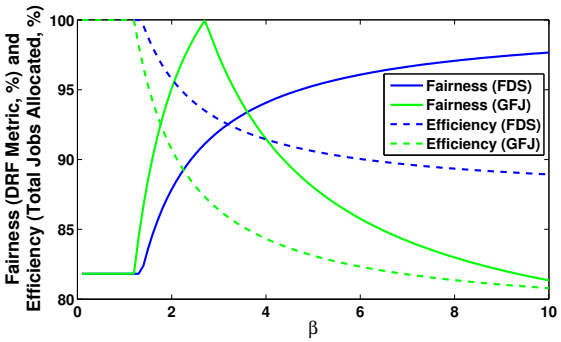


Fig. 14. *The fairness-efficiency tradeoff can be tuned by changing β:* Percentage of fairness and efficiency achieved for various fairness measures in a datacenter example, using $\alpha = \beta$ fairness for FDS and GFJ. Notice that an increased emphasis on fairness (i.e. larger β) need not always decrease the efficiency of the allocation, as seen for $\beta < 2.6$ for GFJ measure.

examines the general behavior of fairness when a larger eff iciency is achieved. Here we measure fairness as percent fairness with the DRF metric and efficiency as percent eff iciency on the number of jobs processed.

Figure 13 shows the optimal allocations of jobs for different values of β, $\lambda = \frac{1-\beta}{\beta}$. Both FDS and GFJ become α-fair on the dominant shares of and jobs allocated to each user, respectively, for $\alpha = \beta$. As β increases, λ decreases, so that fairness is emphasized more than efficiency and FDS asymptotes to DRF. For small β (i.e., more relative emphasis on efficiency than fairness), the optimal FDS allocation maximizes efficiency. In the case of GFJ, which emphasizes the fairness on jobs allocated, larger β values produce a more fair allocation of jobs across users than FDS, as expected. Consequently, the total number of jobs processed (i.e., eff iciency) is lower for GFJ than for FDS.

Figure 14 gives a representative plot of how this tradeoff varies with β and $\lambda = \frac{1-\beta}{\beta}$. As β grows larger, the percent efficiency from the FDS measure drops, approaching DRF in the limit $\beta \to \infty$. The GFJ fairness increases until $\beta = 2.6$, at which point the GFJ-optimal allocation is also DRF-optimal. (We see in Fig. 13 that the GFJ allocation "crosses" the DRF allocation line at this value of β). For larger values of β, GFJ quickly converges to an allocation with a more equal number of jobs per user; thus, its efficiency decreases. But efficiency in FDS decreases more slowly since FDS attempts to make the dominant shares, not the number of jobs, more equitable.

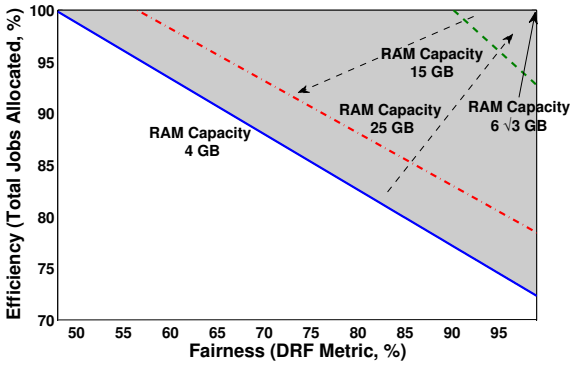


Fig. 15. *Capacity expansion allows different FDS fairness-efficiency tradeoff contours:* Attainable efficiency vs. fairness tradeoffs from different implicit realizations of $\beta \in (-5, 5)$ and $\lambda \in (0.01, 1.91)$ for $\beta < 0$, $\lambda \in \left(0.005\left(\frac{1}{\beta} - 2\right), 0.955\left(\frac{1}{\beta} - 2\right)\right)$ for $\beta > 0$ values. DRF is used as the fairness benchmark and metric.

Finally, we show the interaction between capacity constraints and the range of fairness-efficiency tradeoffs achieved. The shaded region in Fig. 15 shows the attained tradeoffs for a large range of β and λ values; each point corresponds to some β and λ values in the FDS function that achieve the shown operating tradeoff. This achieved tradeoff depends on the available capacity, with contour lines for various RAM capacities shown in the figure. As RAM capacity increases from 4 GB to $6\sqrt{3}$ GB, the tradeoff stops: one can increase both fairness and efficiency. At a RAM capacity of $6\sqrt{3}$ GB, the conditions of Prop. 7 are satisfied, and efficiency is maximized when the dominant shares are equal. When the RAM capacity goes above $6\sqrt{3}$ GB up to 25 GB, user 1's dominant share $\frac{4x_1}{\text{RAM capacity}}$ decreases. Thus, an increase in fairness requires an increase in x_1 and user 1's CPU allocation. User 2 is then allocated fewer jobs, decreasing efficiency. In this figure, one can achieve 100% efficiency and fairness when RAM capacity is $6\sqrt{3}$ GB, but such an ideal operating point does not always exist.

Figure 16 shows the analogue of Fig. 15 for GFJ functions. In this case, the range of attainable efficiency at the maximum allocation decreases as the fairness value increases. Thus, one can increase both fairness and efficiency as RAM capacity goes from 4 GB to 25 GB. Moreover, the contour lines "bend back" on themselves, indicating that for different β and λ parameters, the same fairness value can result in many eff iciency values at the optimal allocation. When RAM capacity equals 11.25 GB, the conditions of Prop. 7 are satisfied and there is no tradeoff between fairness and efficiency.

VI. Survey on Fairness Parameters

In this section, we provide results from a simple survey to complement the proposed theoretical framework with a demonstration of how the typical values of fairness function parameters can be estimated from large scale consumer surveys. We note that our survey methodology and results should be considered as a demonstration of one out of many feasible approaches rather than a prescription of what exact parameter values to choose in a given real world scenario. In particular, this survey provides a systematic way of inferring an initial estimate for (β, λ) values, visualizes participant clusters in

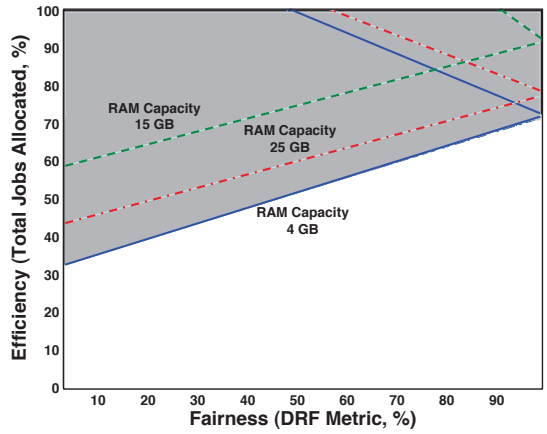


Fig. 16. *Capacity expansion allows different GFJ fairness-efficiency tradeoff contours:* Attainable efficiency vs. fairness tradeoffs from different implicit realizations of $\beta \in (-5, 5)$ and $\lambda \in (0.01, 1.91)$ for $\beta < 0$, $\lambda \in \left(0.005\left(\frac{1}{\beta}-2\right), 0.955\left(\frac{1}{\beta}-2\right)\right)$ for $\beta > 0$ values. DRF is used as the fairness benchmark and metric.

Now consider a similar scenario as before, that is:

- Both clients want to complete as many jobs as possible (including fractional completion of jobs) **Client A** needs **1 CPU** and **4 TB** per job, while **Client B** needs **3 CPU** and **1 TB** per job.
- Both clients would like to be treated more or less *fairly* in the amount of resources allocated to them
- Both clients pay you the same amount of $ per jobs completed. Thus, your **revenue** depends only on the "*Total number of jobs completed*"

But now you have a total of **108 CPUs** and **180 TB** (Terabytes) of storage available to allocate.

- The following table shows five "**Allocation**" *options* at your disposal to distribute your available resources among these two clients, and the corresponding in number of jobs completed for **each** client and a sum **total** number of jobs completed from that allocation.

Allocation Options	Allocated to Client A			Allocated to Client B			Total no. of Jobs Completed
	CPU	TB	No. of Jobs Completed for Client A	CPU	TB	No. of Jobs Completed for Client B	
Allocation 1	24	96	**24**	84	28	**28**	**52**
Allocation 2	12	48	**12**	96	32	**32**	**44**
Allocation 3	36	144	**36**	72	24	**24**	**60**
Allocation 4	45	180	**45**	0	0	**0**	**45**
Allocation 5	27	108	**27**	81	27	**27**	**54**

Fig. 17. Question 2 of our online fairness survey. Client A required 1 CPU and 4 TB per job, while client B required 3 CPU and 1 TB per job. The datacenter had a total of 108 CPUs and 180 TB to allocation.

the fairness-efficiency space, and connects the FDS and GFJ functions with participants' responses.

A. Survey Methodology

We conducted an online survey in January-February 2012, which received 143 responses, mostly from the U.S. Out of these responses, 110 were complete and were used in the subsequent analysis. The participants were given six questions, each with a simplified 'toy' scenario of resource allocation in a datacenter, where jobs from two different clients had heterogeneous resource requirements over multiple resources (CPU and storage). Our online survey participants were faculty, students, and staff primarily from the EE and CS departments of Princeton and George Washington University. They all were familiar with everyday computer use, and hence intuitively understood the two resources considered (processing power and storage capacity). The survey questionnaire further explained the context to ensure participants' understanding.

We limited our question scenarios to only two types of resources in order to ease participants' understanding of the questions, although more sophisticated methods using conjoint analysis can be used on data with more resources [20]. In the last question, we increase the number of resources to three: clients' jobs required CPU, storage, and bandwidth. Each of the six questions offered five different options of distributing resources among the two clients, with each option resulting in a particular outcome. For each question, the survey participants were asked to rank the five allocation options in decreasing order of preference, as shown in Fig. 17.

In four of the questions, the five options that the survey participants were asked to rank were reported in terms of the number of jobs completed for each datacenter client under that option's resource allocation. In the other questions, the options were reported in terms of the leftover (unused or wasted) capacity resulting from that resource allocation option. The questions had either the same set of allocation choices or a scalar multiple, thus permitting a sanity check on whether participants made consistent choices when the outcomes were reported in different metrics ('total number of jobs completed' and 'leftover resources') or were scaled by a constant factor. To avoid influencing the participant's decisions, we did not explicitly inform them of the survey's purpose, i.e., evaluating their fairness-efficiency tradeoff.

The full survey is available in Appendix VI. The results obtained from analyzing the survey responses are reported in the next subsection.

B. Results

Our analysis of the survey results focuses on three goals:

- Evaluate consistency of the results across users with the fairness axioms in Appendix A and [3].
- Cluster participants based on the fairness and efficiency values inferred from their preferences in their rankings of resource allocations.
- Determine the different β and λ heat maps of compatible parameter values for participants in each cluster.

We address these sequentially below.

1) Axiom Validation: We first use the survey results to examine our construction of the fairness functions, evaluating the consistency of the results with three of the four axioms from which these functions are constructed (see Appendix A for a full list of the axioms). To keep the survey simple, we were unable to evaluate the Axiom of Continuity, which, however, is quite intuitive.

Figure 18 shows the number of participants ranking each allocation first, second third, etc. in each question of the survey. We see that a clear consensus emerges across the participant pool: for instance, for question 2 most people rank the allocations from best to worst as 3, 5, 1, 2, 4. It is interesting to note that the fourth allocation, under which client B had no jobs done, has the lowest rank. In fact, allocation 2, which is less efficient than allocation 4, was more preferred. This result is thus consistent with the Axiom of Starvation: participants generally dislike starvation allocations, even if they are more efficient.

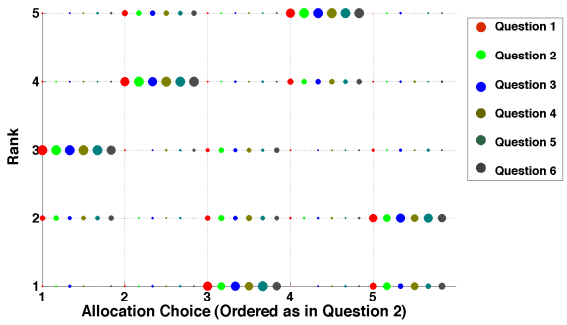


Fig. 18. Allocation rankings for each survey question. The responses to each question are shown in a row at each allocation, e.g. the first six dots correspond to rankings of allocation 1 in questions 1-6, the second six dots correspond to rankings of allocation 2 in questions 1-6, etc. The size of the bubble is proportional to the number of people choosing a particular rank for a particular allocation.

We implicitly evaluate the remaining two axioms (those of Saturation and Partition) by examining the consistency of participants' responses when the allocations are scaled up or down. Our fairness functions predict that a person's rankings of different allocations should not change with this scaling. Figure 18 shows that for each question, a clear consensus ranking emerges; moreover, this ordering of allocations is consistent across all questions[5]. This observation is especially significant since questions 3 and 5 report the leftover capacity as a metric instead of the total number of jobs processed–thus, even when the efficiency metric changes, participants' answers are consistent across the different survey questions.

2) Participant Clustering: We now evaluate the consistency of different people's responses by calculating the *average preferred fairness and efficiency values* for each person and each question. These are calculated by taking a weighted average of the efficiency and fairness values for each allocation; the weights are determined by the participant's ranking of that allocation. The fairness metric is defined to be the negative of the difference between the numbers of jobs processed for clients A and B, while the efficiency metric is taken from the survey as the total number of jobs processed (or the leftover capacity). The leftover capacity is measured by the negative of the percentage of leftover capacity for each resource, to facilitate comparison of leftover CPUs with leftover GB. We use negatives for the fairness value and leftover capacity metric so that an increase in the fairness or leftover capacity value indicates a more fair or more efficient allocation.

We see from Fig. 19 that for all questions, participants tend to fall into two distinct groups, one of which puts more emphasis on efficiency, and one which puts more emphasis on fairness. The two groups have approximately equal numbers of participants (e.g., 52 in each for question 1). Moreover, these groups are consistent across questions. While the numerical fairness and efficiency values vary depending on the allocation scalarization and efficiency metric used in a question, we see that both clusters lie in approximately the same position in the graph for each question.

[5]This ranking consensus is simply in terms of majority agreement on the rank of the allocations, but does not mean that the individual participants' (β, λ) values agree.

3) Parameter Choices: We next determine β and λ values compatible with the answers in Fig. 19's clusters. The results for participants in both clusters were the same for all questions; thus, we only show the β and λ values for question 2.

We use exhaustive search for discretized β and λ values to determine whether a given person's allocation ranking is compatible with that obtained using the (β, λ) fairness function. Figure 20 shows the heat map of compatible β and λ values for a person in each of the two dominant clusters; the intensity of the color corresponds to the number of times an answer is compatible with the given (β, λ) value. A darker color indicates a larger number of compatible answers across users. In this figure, we assume that participants use a GFJ fairness function. Though no single (β, λ) value is compatible with all participants (the single black squares represent a maximum number of compatible answers), a majority of responses were compatible with some (β, λ) value: 50% of cluster 1 and 60% of cluster 2 participants agreed on at least one (β, λ) pair.[6]

As expected, the compatible λ values for cluster 1 (Fig. 20a) are higher in absolute value than those in cluster 2 (Fig. 20b), as is consistent with cluster 2 participants' preferring fairness over efficiency (Fig. 19). The reference lines in the figure show the Pareto-efficient frontier. For $\beta > 1$, most of the compatible (β, λ) values are below the Pareto-eff icient frontier, i.e., not Pareto-efficient. This does not happen for cluster 1 participants, as might be expected since they emphasize efficiency. However, as β increases, more Pareto-eff icient (β, λ) values are compatible with at least some answers.

Figure 21 shows the (β, λ) heat graphs for both participant clusters when FDS-fairness is used. Only the heat graphs for question 2 are shown; the other questions give similar results. We see that all of the (β, λ) values tested in Fig. 21a are compatible with the cluster 1 responses (50% of responses agree on these values). We may partially explain these results by the fact that cluster 1 participants all favor allocation 3 over allocation 5: calculating the dominant shares of each client, we see that allocation 5 actually gives clients *less equitable* dominant shares, and that the sum of dominant shares for allocation 3 is also larger than that for allocation 5. Thus, no matter which β and λ are considered, allocation 3 will be ranked above allocation 5. All (β, λ) pairs are therefore consistent with this ranking. Most participants rank the other allocations in a manner consistent with ranking 3 above 5; those participants whose additional rankings are inconsistent do not show any compatible (β, λ) values.

In contrast to cluster 1, *all* of the (β, λ) values tested are inconsistent with cluster 2's allocation preferences. We can account for this result by noting that all cluster 2 participants prefer allocation 5 (processing an equal number of jobs for each client) over allocation 3. However, allocation 3 is both more more efficient (under FDS) than allocation 5, and hence is inconsistent with cluster 2's answers if they used FDS.

[6]This result may be due to our discretization; for instance, using a λ closer to zero may improve the compatibility with cluster 2 participants, who emphasize fairness over efficiency. Using a larger λ may improve compatibility with cluster 1 participants. It is also possible that a minority of participants provided inconsistent responses compatible with no (β, λ) values, e.g. preferring efficiency to fairness in ranking two allocations, and fairness over efficiency in another two allocations.

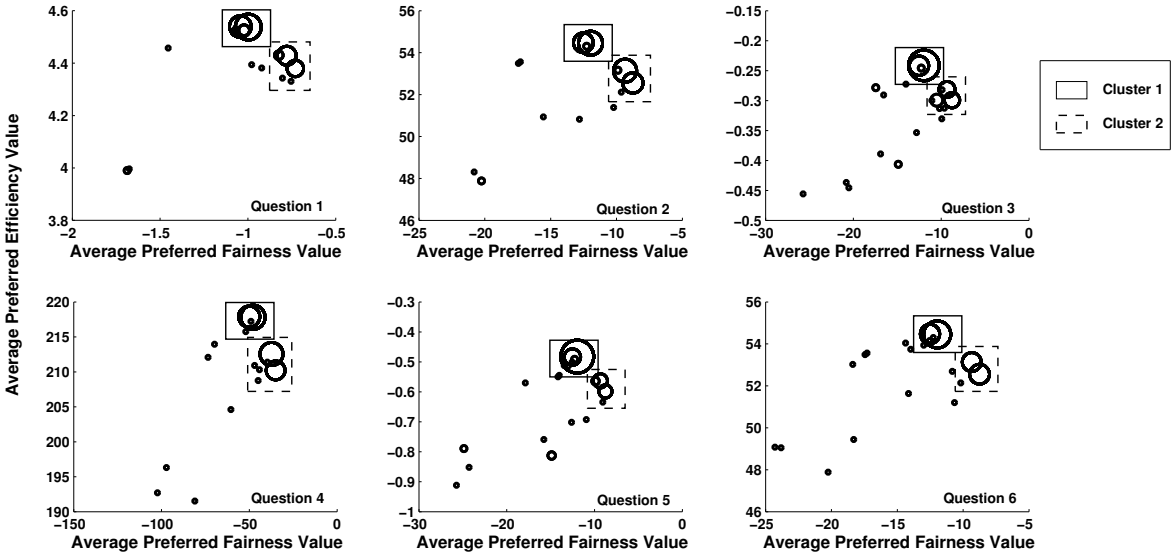
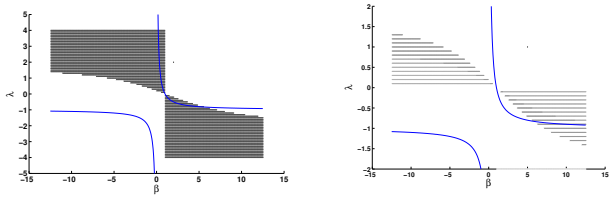
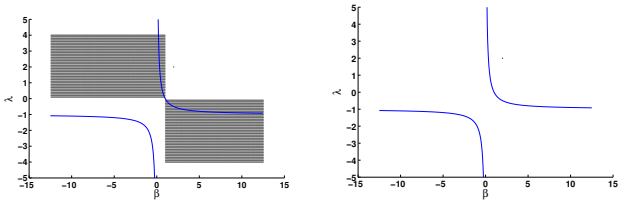


Fig. 19. Average preferred fairness and efficiency values for each survey question. The size of the circle is proportional to the number of people with those particular fairness-efficiency values. Participants tend to fall into two distinct groups, with some other extraneous points.



(a) Cluster 1 participants. (b) Cluster 2 participants.

Fig. 20. Heat map of compatible (β, λ) values for clusters 1 and 2 participants in Fig. 19, GFJ fairness. The reference line is the Parero-efficient boundary $|\lambda| = |(1-\beta)/\beta|$, and the black dot at $(\beta, \lambda) = (2, 2)$ represents a maximum number of compatible answers. Only question 2 results are shown; those for all other questions are similar.



(a) Cluster 1 participants. (b) Cluster 2 participants.

Fig. 21. Heat map of compatible (β, λ) values for clusters 1 and 2 participants in Fig. 19, FDS fairness. The reference line is the Parero-efficient boundary $|\lambda| = |(1-\beta)/\beta|$, and the black dot at $(\beta, \lambda) = (2, 2)$ represents a maximum number of compatible answers. Only question 2 results are shown; results for the other questions are similar.

We thus conjecture that inconsistency arises because GFJ is a "more natural" fairness function: for certain β and λ values, most of cluster 1 and cluster 2 participants exhibit preferences consistent with GFJ fairness. While it is intuitive that most people find it natural to understand fairness in terms of jobs completed rather than dominant share, this is an interesting direction to explore through repeated and controlled behavioral experiments.

The fact that participants generally seem to follow GFJ rather than FDS fairness has interesting implications, as Props. 4 and 5 show that sharing incentive and envy-freeness are more likely to hold when FDS is used instead of GFJ. Participants thus pay more attention to the number of jobs allocated to each client, rather than each client's share of the resources allocated; more generally, we can say that in making allocation decisions, many participants did not fully internalize the heterogeneity in the clients' different resource requirements. Intuitively, this might be expected, since the number of jobs allocated is a more "natural" measure of fairness than the proportion of different resources allocated. However, this observation, if validated in a larger survey, can provide useful guidelines for datacenter operators in that they need to educate their clients about the externality imposed on others by each client's unique heterogeneous resource requirements.

VII. Future Work

Initial exploration suggests that both FDS and GFJ can be unified into a single framework. The idea is to use a p-norm function $g(\gamma_{1,j}, \ldots, \gamma_{n,j}) = \left(\sum_i \gamma_{i,j}^p\right)^{\frac{1}{p}}$ to scalarize the resource requirement vector of user j, and then evaluate the resulting fairness by $f_{\beta,\lambda}$. This method leads to a new family of fairness measures, parameterized p, β, and λ, i.e.,

$$f_{p,\beta,\lambda} = \operatorname{sgn}(1-\beta)\left(\sum_{j=1}^{m}\left(\sum_{k=1}^{n} R_{kj}^p\right)^{\frac{1-\beta}{p}} x_j^{1-\beta}\right)^{\frac{1}{\beta}} \times \left(\sum_{j=1}^{m}\left(\sum_{k=1}^{n} R_{kj}^p\right)^{\frac{1}{p}} x_j\right)^{\lambda+1-\frac{1}{\beta}}. \quad (19)$$

Fairness $f_{p,\beta,\lambda}$ includes many fairness measures as special cases. For instance, $f_{0,\beta,\lambda} = f_{\beta,\lambda}^{GFJ}$ and $f_{\infty,\beta,\lambda} = f_{\beta,\lambda}^{FDS}$, while $f_{1,\beta,\lambda}$ gives the total resource usage in the system.

This function again satisfies the four axioms of [3], as do FDS and GFJ. Moreover, Pareto-efficiency is satisfied for $|\lambda| \geq \left|\frac{1-\beta}{\beta}\right|$, $\beta > 0$. We expect that, in analogy with Props. 4 and 5 and their corollaries, threshold values of p and β can be

found, above which sharing incentive and envy-freeness are satisfied if $\beta > 0$ and $\lambda = \frac{1-\beta}{\beta}$.

In addition to the functional unification proposed in (19), a number of extensions to the current framework are possible. First, we have assumed that both resources and jobs are infinitely divisible. However, in practice a job may require a minimum, indivisible bundle of resources, e.g., 2 GB of memory and 1 CPUs, to run one instance of the job, whereas allocating 1 GB of memory and 1/2 CPUs offers no more benef its than allocating nothing at all. Second, our fairness measures are assumed to be irrelevant to the feasible region of resources. Adding a feasible region and indivisible resources would leads to a fairness version of the knapsack problem, which has no known solution. Some approaches to the knapsack problem are summarized in Appendix D.

Another interesting direction to explore is to extend our multi-resource fairness theory to account for job deadlines, scheduling, and user utility from allocated resources. Finally, our fairness analysis is based on a model of static jobs whose resource demands follow a constant pattern. Many applications not only have time elasticity of demand, but also allow jobs to dynamically change the composition of a bundle of different types of resources. These are all challenging problems that can be explored as future work.

VIII. Concluding Remarks

In this paper, we introduce FDS and GFJ, two families of fairness functions for multi-resource allocations. FDS also includes as a special case the recently-proposed generalization of the max-min fairness measure for multiple resources. Different parameterizations of these functions generate a range of fairness-efficiency tradeoffs, thus allowing for different degrees of emphasis on fairness and efficiency that suit different network operation needs.

We consider three key properties of fairness functions: Pareto-efficiency, sharing incentive, and envy-freeness. FDS and GFJ are both Pareto-efficient if $|\lambda| \geq \frac{1-\beta}{\beta}$, $\beta > 0$. FDS satisfies the sharing incentive property and is envy-free for $\beta > 1$ and $\lambda = \frac{1-\beta}{\beta}$; if $0 < \beta < 1$ and $\lambda = \frac{1-\beta}{\beta}$, then sharing incentive and envy-freeness are only sometimes satisfied. GFJ may or may not be sharing-incentive compatible or envy-free for any $\beta > 0$, $\lambda = \frac{1-\beta}{\beta}$.

We also explore the estimation of the β and λ values which correspond to people's preferences. Preliminary results along these lines are given in Section VI, though one can easily imagine extensions of both the results analysis and the questions asked to participants. Given the limited set of allocations ranked by the participants, reverse-engineering *unique* (β, λ) values compatible with each response was not feasible, but it would be interesting to determine if such unique parameters exist given the rankings of more allocations. Moreover, our current sample size consists primarily of students and others in the academic community who are familiar with computers; with a more diverse demographic of participants, we could examine the impact of various demographic factors on participants' responses. In particular, we could investigate whether participants naturally group themselves into more than two clusters, and whether these have any demographic correlations.

Acknowledgment

The authors wish to thank Augustin Chaintreau and Chee-Wei Tan for their comments and assistance with recruiting survey participants.

References

[1] R. Jain, D. M. Chiu, and W. R. Hawe, *A quantitative measure of fairness and discrimination for resource allocation in shared computer system*. Eastern Research Laboratory, Digital Equipment Corp., 1984.

[2] K. J. Arrow, "The theory of risk aversion," *Essays in the theory of risk-bearing*, pp. 90–120, 1971.

[3] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, "An axiomatic theory of fairness in network resource allocation," in *Proceedings of IEEE INFOCOM*. IEEE, 2010, pp. 1–9.

[4] H. Varian, "Equity, envy, and efficiency," *Journal of Economic Theory*, vol. 9, no. 1, pp. 63–91, 1974.

[5] A. Odlyzko, "Network neutrality, search neutrality, and the never-ending conflict between efficiency and fairness in markets," in *Review of Network Economics*, vol. 8, no. 1, March 2009, pp. 40–60.

[6] M. Zukerman, L. Tan, H. Wang, and I. Ouveysi, "Efficiency-fairness tradeoff in telecommunication networks," in *IEEE Communications Letters*. IEEE, 2005, pp. 643–645.

[7] 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 the 8th USENIX conference on networked systems design and implementation*. USENIX Association, 2011, pp. 24–37.

[8] S. K. Baruah, N. K. Cohen, C. G. Plaxton, and D. A. Varvel, "Proportionate progress: A notion of fairness in resource allocation," *Algorithmica*, vol. 15, no. 6, pp. 600–625, 1996.

[9] J. F. Nash, "The bargaining problem," *Econometrica*, vol. 18, no. 2, pp. 155–162, 1950. [Online]. Available: http://www.jstor.org/stable/1907266

[10] R. Mazumdar, L. G. Mason, and C. Douligeris, "Fairness in network optimal flow control: Optimality of product forms," *IEEE Transactions on Communications*, vol. 39, no. 5, pp. 775–782, 1991.

[11] A. Tang, J. Wang, and S. H. Low, "Counter-intuitive throughput behaviors in networks under end-to-end control," *IEEE/ACM Transactions on Networking*, vol. 14, no. 2, pp. 355–368, 2006.

[12] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *The Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[13] R. Srinivasan and A. K. Somani, "On achieving fairness and efficiency in high-speed shared medium access," *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 111–124, 2003.

[14] M. Dianati, X. Shen, and S. Naik, "A new fairness index for radio resource allocation in wireless networks," in *Proceedings of the 2005 IEEE Wireless Communications and Networking Conference*, vol. 2. IEEE, 2005, pp. 712–717.

[15] M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker, and I. Stoica, "Delay scheduling: A simple technique for achieving locality and fairness in cluster scheduling," in *Proceedings of the 5th European Conference on Computer Systems*. ACM, 2010, pp. 265–278.

[16] Y. Yang, "Rumr: Robust scheduling for divisible workloads," in *Proceedings of the 12th IEEE Symposium on High Performance and Distributed Computing*. IEEE, 2003.

[17] M. Drozdowski. (2011, Jul.) Introduction to divisible tasks. [Online]. Available: http://www.cs.put.poznan.pl/mdrozdowski/divisible/divisible_intro/divisible_intro.html

[18] T. Bonald and L. Massoulié, "Impact of fairness on Internet performance," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 29, no. 1. ACM, 2001, pp. 82–91.

[19] L. Massoulié and J. Roberts, "Bandwidth sharing: Objectives and algorithms," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 320–328, 2002.

[20] P. Green and V. Srinivasan, "Conjoint analysis in consumer research: Issues and outlook," *Journal of Consumer Research*, pp. 103–123, 1978.

[21] A. Fréville, "The multidimensional 0–1 knapsack problem: An overview," *European Journal of Operational Research*, vol. 155, no. 1, pp. 1–21, 2004.

[22] M. Magazine and M. Chern, "A note on approximation schemes for multidimensional knapsack problems," *Mathematics of Operations Research*, pp. 244–247, 1984.
[23] M. Moser, D. Jokanovic, and N. Shiratori, "An algorithm for the multidimensional multiple-choice knapsack problem," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 80, no. 3, pp. 582–589, 1997.
[24] H. Weingartner and D. Ness, "Methods for the solution of the multidimensional 0/1 knapsack problem," *Operations Research*, vol. 15, no. 1, pp. 83–103, 1967.
[25] P. Chu and J. Beasley, "A genetic algorithm for the multidimensional knapsack problem," *Journal of Heuristics*, vol. 4, no. 1, pp. 63–86, 1998.
[26] M. Vasquez, J. Hao *et al.*, "A hybrid approach for the 0-1 multidimensional knapsack problem," in *International Joint Conference on Artif icial Intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd., 2001, pp. 328–333.
[27] T. Morin and R. Marsten, "An algorithm for nonlinear knapsack problems," *Management Science*, vol. 22, no. 10, pp. 1147–1158, 1976.
[28] K. Bretthauer and B. Shetty, "The nonlinear knapsack problem–algorithms and applications," *European Journal of Operational Research*, vol. 138, no. 3, pp. 459–472, 2002.
[29] J. Dussault, J. Ferland, and B. Lemaire, "Convex quadratic programming with one constraint and bounded variables," *Mathematical Programming*, vol. 36, no. 1, pp. 90–104, 1986.
[30] T. Klastorin, "On a discrete nonlinear and nonseparable knapsack problem," *Operations Research Letters*, vol. 9, no. 4, pp. 233–237, 1990.
[31] G. Gallo, P. Hammer, and B. Simeone, "Quadratic knapsack problems," *Combinatorial Optimization*, vol. 12, pp. 132–149, 1980.
[32] T. Sharkey, H. Romeijn, and J. Geunes, "A class of nonlinear nonseparable continuous knapsack and multiple-choice knapsack problems," *Mathematical Programming*, vol. 126, no. 1, pp. 69–96, 2011.
[33] S. J. Brams and A. D. Taylor, *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge University Press, 1996.
[34] S. J. Brams, M. A. Jones, and C. Klamler, "Better ways to cut a cake," *Notices of the AMS*, vol. 53, no. 11, pp. 1314–1321, 2006.
[35] C. E. Koksal, H. Kassab, and H. Balakrishnan, "An analysis of short-term fairness in wireless media access protocols (poster session)," *ACM SIGMETRICS Performance Evaluation Review*, vol. 28, no. 1, pp. 118–119, 2000.
[36] M. Bredel and M. Fidler, "Understanding fairness and its impact on quality of service in IEEE 802.11," in *Proceedings of IEEE INFOCOM*. IEEE, 2009, pp. 1098–1106.
[37] M. Marsan and M. Gerla, "Fairness in local computing networks," *Proceedings of IEEE ICC*, 1982.
[38] J. Wong, J. Sauve, and J. Field, "A study of fairness in packet-switching networks," *IEEE Transactions on Communications*, vol. 30, no. 2, pp. 346–353, 1982.
[39] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556–567, 2000.
[40] M. Uchida and J. Kurose, "An information-theoretic characterization of weighted alpha-proportional fairness," in *Proceedings of IEEE INFOCOM*. IEEE, 2009, pp. 1053–1061.
[41] T. Lan and M. Chiang, "An axiomatic theory of fairness in resource allocation," George Washington University, http://www.seas.gwu.edu/tlan/papers/fairness.pdf, Tech. Rep., 2011.
[42] A. Renyi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547–561.
[43] D. K. Fadeev, "Zum Begriff der Entropie einer endlichen Wahrscheinlichkeitsschemas," in *Arbeiten zur Informationstheorie I*. Deutscher Verlag der Wissenschaften, 1957, pp. 85–90.
[44] R. Aaberge, "Axiomatic characterization of the Gini coefficient and Lorenz curve orderings," *Journal of Economic Theory*, vol. 101, no. 1, pp. 115–132, 2001.
[45] L. S. Shapley, *A Value for n-Person Games*. Princeton, NJ: Princeton University Press, 1953, pp. 307–317, Annals of Mathematical Studies, vol. 28.
[46] A. B. Atkinson, "On the measurement of inequality," *Journal of Economic Theory*, vol. 2, no. 3, pp. 244–263, 1970.
[47] J. Rawls, *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971.
[48] J. Rawls and E. Kelly, *Justice as Fairness: A Restatement*. Belknap Press, 2001.
[49] A. Sen and W. Bernard, *Utilitarianism and Beyond*. Cambridge University Press, 1982.
[50] S. C. Kolm, *Justice and Equity*. MIT Press, 1970, Trans. See, H. F.
[51] S. Brams, P. Edelman, and P. Fishburn, "Paradoxes of fair division," *The Journal of Philosophy*, vol. 98, no. 6, pp. 300–314, 2001.

APPENDIX A
AXIOMS FOR THE CONSTRUCTION OF SINGLE RESOURCE FAIRNESS FUNCTIONS

The fairness measures in [3] are functions $f : \mathbb{R}^n \to \mathbb{R}$ which give the fairness $f(\vec{x})$ of an allocation vector $\vec{x}$, representing the amount of a resource allocated to each user. These measures may be derived from five distinct axioms:

1) **Axiom of Continuity**: The function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous for any fixed number of users (i.e., length of the vector $\vec{x}$).
2) **Axiom of Saturation**: As the number of users approaches infinity, the fairness value of an equal allocation should be independent of the number of users
$$\lim_{n\to\infty} \frac{f(\mathbf{1}_n)}{f(\mathbf{1}_{n+1})} = 1,$$
where $\mathbf{1}_n$ denotes an equal allocation among n users.
3) **Axiom of Partition**: Consider an arbitrary partition of a system into two subsystems. Let $\vec{x} = \begin{bmatrix}\vec{x}^1 & \vec{x}^2\end{bmatrix}$ and $\vec{y} = \begin{bmatrix}\vec{y}^1 & \vec{y}^2\end{bmatrix}$ be two partitioned resource allocation vectors, with $\sum_j x_j^i = \sum_j y_j^i$ for $i = 1, 2$. There exists a mean function h such that the fairness ratio of $\vec{x}$ and $\vec{y}$ equals the mean of the fairness ratios of the two suballocations, i.e.,
$$\frac{f(\vec{x})}{f(\vec{y})} = h\left(\frac{f(\vec{x}^1)}{f(\vec{y}^1)}, \frac{f(\vec{x}^2)}{f(\vec{y}^2)}\right),$$
where h is a mean function if and only if it can be expressed as
$$h = g^{-1}\left(s_1 g\left(\frac{f(\vec{x}^1)}{f(\vec{y}^1)}\right) + s_2 g\left(\frac{f(\vec{x}^2)}{f(\vec{y}^2)}\right)\right),$$
with the s_i positive weights such that $s_1 + s_2 = 1$ and g a continuous and strictly monotonic function. These s_i are are chosen to satisfy
$$s_i = \frac{\left(\sum_j x_j^i\right)^\rho}{\left(\sum_j x_j^1\right)^\rho + \left(\sum_j x_j^2\right)^\rho},$$
with $\rho \geq 0$ an arbitrary exponent.
4) **Axiom of Starvation**: In a two user system, an equal allocation is more fair than starving one user: $f([1\ 1]) \geq f([1\ 0])$.

Using the above four axioms yields the fairness measure
$$f_{\beta,\lambda}(\vec{x}) = \operatorname{sgn}(1-\beta)\left(\sum_{i=1}^n \left(\frac{x_i}{\sum_{j=1}^n x_j}\right)^{1-\beta}\right)^{\frac{1}{\beta}} \left(\sum_{i=1}^n x_i\right)^\lambda \tag{20}$$
and its limit
$$\prod_{i=1}^n \left(\frac{x_i}{\sum_{j=1}^n x_j}\right)^{\left(\frac{x_i}{\sum_{j=1}^n x_j}\right)} \left(\sum_{i=1}^n x_i\right)^\lambda \tag{21}$$
as $\beta \to 0$. Note that these are both symmetric with respect to the order of the users.

APPENDIX B
PROOFS OF ALL PROPOSITIONS

A. Proposition 1

If $\eta_{\max} = \max_i \eta_i$, then each resource i's capacity constraint is automatically satisfied whenever resource k's is, where $\eta_{\max} = \eta_k$. Since $\eta_k \mu_j$ is the dominant share of each user j, the problem reduces to the single-resource problem with resource k. Expressions for the optimal allocations may be derived from the proofs of Props. 4 and 5. ■

B. Corollary 1

Without loss of generality, we may assume that each $\mu_j \leq 1$, due to the scaling factor η_1. We have the equation

$$\frac{\partial X}{\partial \mu_j} = \frac{-1}{\eta_1 \beta} \left(\frac{\mu_j^{-\frac{1+\beta}{\beta}}}{\sum_{i=1}^n \mu_i^{\frac{\beta-1}{\beta}}} + \frac{(\beta-1)\mu_j^{-\frac{1}{\beta}} \sum_{i=1}^n \mu_i^{-\frac{1}{\beta}}}{\left(\sum_{i=1}^n x_i^{\frac{\beta-1}{\beta}}\right)^2} \right). \tag{22}$$

Then if $\beta > 1$, we easily see that decreasing $\min_j \mu_j$ yields the greatest increase in efficiency.

If $\beta < 1$, we see that the first term $\frac{\mu_j^{-\frac{1+\beta}{\beta}}}{\sum_{i=1}^n \mu_i^{\frac{\beta-1}{\beta}}}$ in the sum of (22) is positive, and the second term is negative. Thus, since this first term is largest when μ_j is smallest, decreasing $\min_j \mu_j$ also yields the greatest increase or smallest decrease in efficiency. One can show that decreasing $\min_j \mu_j$ always increases efficiency; setting (22) greater than zero, we obtain after some simplification

$$\mu_j^{-\frac{1+\beta}{\beta}} \sum_{i=1}^n x_i^{\frac{\beta-1}{\beta}} > (\beta-1)\mu_j^{-\frac{1}{\beta}} \sum_{i=1}^n \mu_i^{-\frac{1}{\beta}}.$$

Rearranging again, this equation becomes

$$\sum_{i=1}^n \mu_i^{\frac{\beta-1}{\beta}} > (1-\beta)\mu_j \sum_{i=1}^n \mu_i^{\frac{-1}{\beta}},$$

which always holds for $j = \text{argmin}_j \mu_j$. ■

C. Proposition 2

We prove the FDS and GFJ properties separately.

Optimal FDS values: Let resource requirements $\mu_{ij} = R_{ij}/C_i$ be uniformly distributed in $[0, \nu]$. If $\max f_{\infty,-1}^{\text{FDS}}$ is the optimal FDS value with $\beta = \infty$, we have

$$\begin{aligned} \max f_{\infty,-1}^{\text{FDS}} &= \max_f f \\ &\quad \text{s.t.} \sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} f \leq 1, \ \forall i. \\ &= \left(\max_i \sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} \right)^{-1}. \end{aligned} \tag{23}$$

Therefore, to prove (10), it is sufficient to show that for arbitrary $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left\{ \left| \left(\max_i \sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} \right) \cdot \frac{2m}{n(m+1)} - 1 \right| > \epsilon \right\} = 0. \tag{24}$$

Toward this end, we remove the absolute value and bound the probability in (24) by a combination of two inequalities:

$$\begin{aligned} &\mathbf{P}\left\{ \left(\max_i \sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} \right) \cdot \frac{2m}{n(m+1)} - 1 > \epsilon \right\} \\ &\leq \sum_{i=1}^m \mathbf{P}\left\{ \sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} \cdot \frac{2m}{n(m+1)} - 1 > \epsilon \right\} \\ &= m \cdot \mathbf{P}\left\{ \sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} > (1+\epsilon)\frac{n(m+1)}{2m} \right\} \end{aligned} \tag{25}$$

where the the last step uses the symmetry of resource constraints, and

$$\begin{aligned} &\mathbf{P}\left\{ \left(\max_i \sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} \right) \cdot \frac{2m}{n(m+1)} - 1 < -\epsilon \right\} \\ &\leq \mathbf{P}\left\{ \sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} \cdot \frac{2m}{n(m+1)} - 1 < \epsilon, \ \forall i \right\} \\ &\leq \mathbf{P}\left\{ \sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} < (1-\epsilon)\frac{n(m+1)}{2m} \right\}. \end{aligned} \tag{26}$$

Since $\mu_j = \max_i \mu_{ij}$, $\{\mu_{ij}/\mu_j, \ \forall j\}$ are i.i.d. random variables. Using the Central Limit Theorem, as $n \to \infty$, we have

$$\frac{\sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} - n\mathbb{E}\left[\frac{\mu_{ij}}{\mu_j}\right]}{\sqrt{n}\sigma_{\frac{\mu_{ij}}{\mu_j}}} \to z \text{ in distribution.} \tag{27}$$

Here z is a standard normal random variable with mean 0 and variance 1.

To simplify (27), we obtain

$$\begin{aligned} \mathbb{E}\left[\frac{\mu_{ij}}{\mu_j}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\mu_{ij}}{\mu_j} \middle| \mu_{ij} \right] \right] \\ &= \int_0^\nu \mathbb{E}\left[\frac{x}{\mu_j} \middle| \mu_{ij} = x \right] dx \\ &= \int_0^\nu \left[1 \cdot x^{m-1} + \int_x^\nu \frac{x}{y} \cdot f_{\mu_j|\mu_{ij}=x}(y) dy \right] dx \\ &= \int_0^\nu \left[x^{m-1} + \int_x^\nu \frac{x}{y} \cdot (m-1) y^{m-2} dy \right] dx \\ &= \int_0^\nu \left[x^{m-1} + \frac{m-1}{m-2} x(\nu^{m-2} - x^{m-2}) \right] dx \\ &= \frac{\nu^m}{2} \cdot \frac{m-1}{m-2} - \frac{\nu^m}{m(m-2)} \\ &= \frac{\nu^m}{2} + \frac{\nu^m}{2m} \end{aligned} \tag{28}$$

where the forth step uses $f_{\mu_j|\mu_{ij}=x}(y) = (m-1)y^{m-2}$ for all

$y > x$, because $\mu_j = \max_i \mu_{ij}$. Similarly, we have

$$\begin{aligned}\mathbb{E}\left[\frac{\mu_{ij}^2}{\mu_j^2}\right] &= \int_0^1 \mathbb{E}\left[\frac{x^2}{\mu_j^2}\middle|\mu_{ij}=x\right]dx \\ &= \int_0^\nu \left[1\cdot x^{m-1} + \int_x^\nu \frac{x^2}{y^2}(m-1)y^{m-2}dy\right]dx \\ &= \int_0^\nu \left[x^{m-1} + \frac{m-1}{m-3}x^2(\nu^{m-3}-x^{m-3})\right]dx \\ &= \frac{\nu^m(m-1)}{3(m-3)} - \frac{2\nu^m}{m(m-3)} \\ &= \frac{\nu^m(m+1)}{3m} \end{aligned} \tag{29}$$

To derive the standard deviation of μ_{ij}/μ_j, we combine (28) and (29) to derive

$$\begin{aligned}\sigma^2_{\frac{\mu_{ij}}{\mu_j}} &= \mathbb{E}\left[\frac{\mu_{ij}^2}{\mu_j^2}\right] - \left\{\mathbb{E}\left[\frac{\mu_{ij}}{\mu_j}\right]\right\}^2 \\ &= \frac{\nu^m(m+1)}{m}\left(\frac{1}{3} - \nu^m\left(\frac{m+1}{4m}\right)\right). \end{aligned} \tag{30}$$

Combining (27), (28), and (30), we obtain that for arbitrary m,

$$\begin{aligned}&\lim_{n\to\infty} m\cdot\mathbf{P}\left\{\left|\sum_{j=1}^n \frac{\mu_{ij}}{\mu_j} - \frac{n(m+1)}{2m}\right| > \epsilon\cdot\frac{n(m+1)}{2m}\right\} \\ &= \lim_{n\to\infty} m\cdot\mathbf{P}\left\{|z| > \epsilon\cdot\frac{n(m+1)}{2m}\cdot\frac{1}{\sqrt{n}\sigma_{\frac{\mu_{ij}}{\mu_j}}}\right\} \\ &= \lim_{n\to\infty} m\cdot\mathbf{P}\left\{|z| > \epsilon\sqrt{n}\cdot\frac{1}{2\nu^m\left(\frac{1}{3}-\nu^m\left(\frac{m+1}{4m}\right)\right)}\right\} \\ &= 0. \end{aligned} \tag{31}$$

Plug (31) into (25) and (26). We conclude that

$$\begin{aligned}&\lim_{n\to\infty}\mathbf{P}\left\{\left|\left(\max_i\sum_{j=1}^n\frac{\mu_{ij}}{\mu_j}\right)\cdot\frac{2m}{n(m+1)} - 1\right| > \epsilon\right\} \\ &\le \lim_{n\to\infty}(m+1)\mathbf{P}\left\{\left|\sum_{j=1}^n\frac{\mu_{ij}}{\mu_j} - \frac{n(m+1)}{2m}\right| > \frac{\epsilon n(m+1)}{2m}\right\} \\ &= 0, \end{aligned} \tag{32}$$

which is exactly the desired result in (24). Therefore, it completes the proof of (10).

Optimal GFJ values: If $\max f^{\mathrm{GFJ}}_{\infty,-1}$ is the optimal GFJ value with $\beta = \infty$, we have

$$\begin{aligned}\max f^{\mathrm{GFJ}}_{\infty,-1} &= \max_f f \\ &\quad \text{s.t.} \sum_{j=1}^n \mu_{ij} f \le 1,\ \forall i. \\ &= \left(\max_i \sum_{j=1}^n \mu_{ij}\right)^{-1}. \end{aligned} \tag{33}$$

Therefore, to prove (11), it is sufficient to show that for arbitrary $\epsilon > 0$,

$$\lim_{n\to\infty}\mathbf{P}\left\{\left|\left(\max_i\sum_{j=1}^n\mu_{ij}\right)\cdot\frac{2}{\nu\left(\sqrt{mn/3}+n\right)}\right| > \epsilon\right\} = 0. \tag{34}$$

Notice that $\{\mu_{ij},\ \forall j\}$ are i.i.d. random variables uniformly distributed in $[0,\nu]$. Using the Central Limit Theorem, as $n\to\infty$, we have

$$\frac{\sum_{j=1}^n \mu_{ij} - \frac{n\nu}{2}}{\sqrt{\frac{n\nu^2}{12}}} \to z \text{ in distribution.} \tag{35}$$

Here z is a standard normal random variable with mean 0 and variance 1. Let $x = \nu\left(\sqrt{mn/3}+n\right)/2$. Then for any m,

$$\begin{aligned}&\lim_{n\to\infty}\mathbf{P}\left\{\left|\sum_{j=1}^n\mu_{ij} - x\right| > \epsilon x\right\} \\ &\le \lim_{n\to\infty}\mathbf{P}\left\{\left|\sum_{j=1}^n\mu_{ij} - x + \frac{\nu\sqrt{mn/3}}{2}\right| > \epsilon x + \frac{\nu\sqrt{mn/3}}{2}\right\} \\ &= \lim_{n\to\infty}\mathbf{P}\left\{\left|\sum_{j=1}^n\mu_{ij} - \frac{\nu n}{2}\right| > \frac{(\epsilon+1)\nu\sqrt{mn/3}+\nu n}{2}\right\} \\ &= \lim_{n\to\infty}\mathbf{P}\left\{|z| > \frac{(\epsilon+1)\nu\sqrt{mn/3}+\nu n}{2\sqrt{\frac{n\nu^2}{12}}}\right\} \\ &= \lim_{n\to\infty}\mathbf{P}\left\{|z| > (1+\epsilon)\sqrt{m}+\sqrt{3n}\right\} \\ &= 0. \end{aligned} \tag{36}$$

Plugging the inequality into the left hand side of (34) and using the same technique from (25) and (26), we derive

$$\begin{aligned}&\lim_{n\to\infty}\mathbf{P}\left\{\left|\left(\max_i\sum_{j=1}^n\mu_{ij}\right)\cdot\frac{1}{x} - 1\right| > \epsilon\right\} \\ &\le \lim_{n\to\infty}(m+1)\mathbf{P}\left\{\left|\sum_{j=1}^n\mu_{ij} - x\right| > \epsilon x\right\} \\ &= 0. \end{aligned} \tag{37}$$

This completes the proof of (11). ■

D. Proposition 4

We prove each item in the proposition in sequence.

(a) Suppose that $\beta > 0$, $\lambda = \frac{1-\beta}{\beta}$. We index the n users by $j = 1, 2, \ldots, n$ and the m resources by $i = 1, 2, \ldots, m$, and consider the resource constraints

$$\sum_{j=1}^n \gamma_{ij} x_j \le 1\ \forall\ i.$$

First, we show that sharing incentive holds for $\beta > 1$. We introduce a multiplier λ_i for each resource constraint corresponding to resource i. We let $\mu_j x_j$ denote the

dominant share of user j; $\mu_j \geq \gamma_{ij}$ for all i. Then we have the Lagrangian

$$\sum_{j=1}^{n} \frac{(\mu_j x_j)^{1-\beta}}{1-\beta} - \sum_{i=1}^{m} \lambda_i \left(\sum_{j=1}^{n} \gamma_{ij} x_j - 1 \right), \tag{38}$$

and at optimality we have

$$\mu_j^{1-\beta} x_j^{-\beta} = \sum_{i=1}^{m} \lambda_i \gamma_{ij} \; \forall \; i \tag{39}$$

and by slackness,

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \lambda_i \gamma_{ij} x_j = \sum_{i=1}^{m} \lambda_i. \tag{40}$$

From (39), we have for each $j = 1, 2, \ldots, n$,

$$\mu_j^{-\beta} x_j^{-\beta} = \sum_{i=1}^{m} \lambda_i \frac{\gamma_{ij}}{\mu_j} \leq \sum_{i=1}^{m} \lambda_i.$$

Then from (39) and (40),

$$\sum_{j=1}^{m} \mu_j^{1-\beta} x_j^{1-\beta} = \sum_{i=1}^{m} \lambda_i,$$

and therefore

$$\mu_j^{-\beta} x_j^{-\beta} \leq \sum_{j=1}^{n} \mu_j^{1-\beta} x_j^{1-\beta} \leq n \left(\min_j \mu_j x_j \right)^{1-\beta}$$

for $\beta > 1$. Then

$$\min_j \mu_j x_j \geq \frac{1}{n}. \tag{41}$$

(b) If $0 < \beta < 1$, sharing incentive may not be satisfied. Suppose for instance that only one constraint is tight at optimality. Denote γ_{1j} by γ_j. Then introducing the Lagrangian function as in the $\beta > 1$ case, we obtain from (39) that

$$x_j = x_1 \left(\frac{\gamma_j}{\gamma_1} \right)^{-\frac{1}{\beta}} \left(\frac{\mu_1}{\mu_j} \right)^{\frac{\beta-1}{\beta}}$$

$$\sum_{j=1}^{n} \gamma_j x_j = 1.$$

Solving for the x_j, we obtain

$$x_j = \frac{\gamma_j^{-\frac{1}{\beta}}}{\mu_j^{\frac{\beta-1}{\beta}} \sum_{i=1}^{n} \left(\frac{\gamma_i}{\mu_i} \right)^{\frac{\beta-1}{\beta}}}. \tag{42}$$

Now in order for sharing incentive to *not* be satisfied,

$$\mu_j x_j < \frac{1}{n}$$

for some user j. Substituting (42) for x_j, we simplify to

$$n < \frac{\gamma_j}{\mu_j} + \left(\frac{\gamma_j}{\mu_j} \right)^{\frac{1}{\beta}} \sum_{i \neq j} \left(\frac{\gamma_i}{\mu_i} \right)^{\frac{\beta-1}{\beta}}. \tag{43}$$

Evidently, as $\gamma_i \leq \mu_i$ for all i, these equations may be satisfied if and only if $0 < \beta < 1$. Indeed, consider the n-user system with two resources and resource constraints $\sum_{i=1}^{n} x_i/2 \leq 1$, $x_1 \leq 1$. Then $\mu_i = 0.5$ for all $i > 1$, while $\mu_1 = 1$. Let λ_1 denote the Lagrange multiplier for the constraint $\sum_{i=1}^{n} x_i/2 \leq 1$, and λ_2 denote the Lagrange multiplier for $x_1 \leq 1$. At the optimal allocation, we have $x_i = 2\lambda_1^{-1/\beta}$ for $i > 1$, while $x_1 = (\lambda_1/2)^{-1/\beta}$. Clearly, $\lambda_1 > 0$ at the optimal allocation, so $\lambda_1^{-1/\beta} \left(2(n-1) + 2^{1/\beta} \right) = 2$, and

$$\lambda_1 = 2^{-\beta} \left(2(n-1) + 2^{\frac{1}{\beta}} \right)^{\beta} = \left(n - 1 + 2^{\frac{1}{\beta} - 1} \right)^{\beta}.$$

Thus, $x_1 < 1$, i.e., the constraint $x_1 \leq 1$ is not tight, if $2^{1/\beta} \left(n - 1 + 2^{\frac{1}{\beta}} \right)^{-1} < 1$ or

$$2^{\frac{1}{\beta}} < n - 1 + 2^{\frac{1}{\beta}},$$

which holds for any $\beta > 0$ if $n > 2$. Our condition (43) for sharing incentive not to be satisfied is then, taking $j \neq 1$,

$$n < 1 + \left(n - 2 + 2^{\frac{1}{\beta} - 1} \right) = n - 1 + 2^{\frac{1}{\beta} - 1},$$

which is clearly true for all n and $0 < \beta < 1$.

(c) We now show that for $|\lambda|$ sufficiently large, sharing incentive need not be satisfied by the optimal allocation. Consider a two-user, two-resource system with constraints $x_1 \leq 1$ and $\gamma x_1 + x_2 \leq 1$, $\gamma < 1$. Then $\mu_1 = \mu_2 = 1$, and for as $\lambda \to \infty$, we maximize $x_1 + x_2$. But this quantity is maximized when either $x_1 = 1$, $x_2 = 1 - \gamma$ or when $x_1 = 0$, $x_2 = 1$. Clearly, the optimal allocation occurs when $x_1 = 1$, $x_2 = 1 - \gamma$. But then user 2's dominant share is $1 - \gamma < 1/2$ if $\gamma > 1/2$.

(d) If $\lambda = 0$, the dominant shares of all users are equalized at the optimal allocation. But since the sum of the dominant shares is ≥ 1, no user's dominant share falls below $1/n$, and the sharing incentive property is satisfied.

This complete the proof of each item in Prop. 4. ■

E. Corollary 2

From (40), we obtain

$$\begin{aligned} \sum_{i=1}^{m} \lambda_i &= \sum_{i=1}^{m} \sum_{j=1}^{n} \lambda_i \gamma_{ij} x_j \\ &\leq \sum_{i=1}^{m} \sum_{j=1}^{n} \lambda_i \mu_j x_j \\ &= \left(\sum_{i=1}^{m} \lambda_i \right) \left(\sum_{j=1}^{n} \mu_j x_j \right). \end{aligned}$$

Then $\sum_{j=1}^{n} \mu_j x_j \geq 1$, and

$$\sum_{j=1}^{n} x_j \geq \frac{1}{\max_j \mu_j},$$

which is the desired bound. ■

F. Corollary 3

Suppose that $\beta > 1$. Then if user j envies user k's share, then $\gamma_{ik}x_k \geq \gamma_{ij}x_j$ for all resources i, with strict inequality for at least one resource. Then from (39), we have

$$\begin{aligned}\mu_j^{1-\beta}x_j^{1-\beta} &= \sum_{i=1}^{m}\lambda_i\gamma_{ij}x_j \\ &< \sum_{i=1}^{m}\lambda_i\gamma_{ik}x_k \\ &= \mu_k^{1-\beta}x_k^{1-\beta},\end{aligned}$$

which is impossible since $\mu_j x_j \geq \mu_k x_k$ and $\beta > 1$.

If $\lambda = 0$, then as in the proof of Prop. 4, we see that the dominant shares are equal at the optimal allocation for any β. But then no user can envy another; user i's share of her dominant resource j is larger than or equal to any other user's share of resource j.

The counterexamples used in the proof of Prop. 4 may be used to show that envy-freeness does not hold for all user-resource systems under the conditions specified. ■

G. Proposition 5

As in Prop. 4, we prove each item in the proposition in sequence.

(a) Suppose that exactly one constraint $\sum_{j=1}^{n}\gamma_j x_j = 1$ is tight at the optimal allocation. The GFJ fairness function for $\beta > 0$, $\lambda = \frac{1-\beta}{\beta}$ is then

$$\operatorname{sgn}(1-\beta)\left(\sum_{j=1}^{n}(x_j)^{1-\beta}\right)^{\frac{1}{\beta}};$$

letting p denote a Lagrange multiplier for the resource constraint, the function

$$\operatorname{sgn}(1-\beta)\left(\sum_{j=1}^{n}(x_j)^{1-\beta}\right)^{\frac{1}{\beta}} - p(\gamma^T\mathbf{x}-1) \qquad (44)$$

is maximized at the optimal allocation. Taking the derivatives with respect to each x_j and p, we obtain the equations

$$\begin{aligned}x_j &= x_1\left(\frac{\gamma_j}{\gamma_1}\right)^{-\frac{1}{\beta}} \\ \gamma^T\mathbf{x} &= 1.\end{aligned}$$

Solving for the x_j, we obtain

$$x_j = \frac{\gamma_j^{-\frac{1}{\beta}}}{\sum_{i=1}^{n}\gamma_i^{\frac{\beta-1}{\beta}}}. \qquad (45)$$

Now, in order for sharing incentive to *not* be satisfied,

$$\mu_j x_j < \frac{1}{n}$$

for some user j. After substituting (45) for x_j, these conditions simplify to

$$n < \frac{\gamma_j}{\mu_j} + \left(\frac{\gamma_j^{\frac{1}{\beta}}}{\mu_j}\right)\sum_{i\neq j}\gamma_i^{\frac{\beta-1}{\beta}}.$$

For $0 < \beta < 1$, this equation is satisfied for γ_i, $i \neq j$, relatively small, and γ_j relatively large. In other words, user j requires a relatively large amount of resources. For $\beta > 1$, this equation is satisfied for γ_i, $i \neq j$, relatively large.

(b) We now suppose that $|\lambda| > |(1-\beta)/\beta|$, with the sign of λ equal to that of $1-\beta$, and show that the optimal allocation need not satisfy the sharing incentive property. Consider a two-user, one-resource system with resource constraint $x_1 + \gamma_2 x_2 \leq 1$. At the optimal allocation, $x_1 = 1-\gamma_2 x_2$, and the total number of jobs processed is $1+(1-\gamma_2)\,x_2$, with fairness value

$$\operatorname{sgn}(1-\beta)\left(\left(1-\gamma_2 x_2\right)^{1-\beta} + x_2^{1-\beta}\right)^{\frac{1}{\beta}}\left(1+(1-\gamma_2)\,x_2\right)^{\xi}$$

where $\xi = \lambda + (\beta-1)/\beta$. We note that ξ is negative for $\beta > 1$ and positive for $\beta < 1$. Taking the derivative with respect to x_2, we have

$$\begin{aligned}&\left(x_2^{1-\beta} + \left(1-\gamma_2 x_2\right)^{1-\beta}\right)^{\frac{1-\beta}{\beta}}\left(1+(1-\gamma_2)\,x_2\right)^{\xi-1} \\ &\left[\frac{|1-\beta|}{\beta}\left(x_2^{-\beta} - \gamma_2\left(1-\gamma_2 x_2\right)^{-\beta}\right)\left(1+(1-\gamma_2)x_2\right) + \right. \\ &\left.|\xi|\left(x_2^{1-\beta} + \left(1-\gamma_2 x_2\right)^{1-\beta}\right)(1-\gamma_2)\right], \qquad (46)\end{aligned}$$

which is positive for $\gamma_2 < 1$ and $x_2^{-\beta} > \gamma_2\left(1-\gamma_2 x_2\right)^{-\beta}$, i.e., $x_2\left(1+\gamma_2^{1-1/\beta}\right) < \gamma_2^{-1/\beta}$ or

$$x_2 < \frac{1}{\gamma_2^{1/\beta} + \gamma_2}.$$

To show that $\gamma_2 x_2 > 1/2$, i.e., user 1's dominant share is less than $1/2$, it suffices to show that

$$\frac{1}{2\gamma_2} < \frac{1}{\gamma_2^{1/\beta} + \gamma_2},$$

i.e., $\gamma_2^{1/\beta} < \gamma_2$, which is true for $0 < \beta < 1$. If $\beta > 1$, then $\xi < 0$ and (46) is increasing as λ becomes more negative, $\gamma_2 < 1$. Then for λ sufficiently large, (46) is positive for x_2 large enough so that at optimality, user 1's share is less than one-half.

(c) We now suppose that $|\lambda| < |(1-\beta)/\beta|$. In this case, (46) is negative for $x_2^{-\beta} < \gamma_2\left(1-\gamma_2 x_2\right)^{-\beta}$ and $\gamma_2 < 1$, i.e., for $x_2 > \left(\gamma_2^{1/\beta} + \gamma_2\right)^{-1}$. Then at optimality, $x_2 \leq \left(\gamma_2^{1/\beta} + \gamma_2\right)^{-1}$ and user 2's dominant share is

$$\gamma_2 x_2 \leq \frac{1}{\gamma_2^{1/\beta-1} + 1} < \frac{1}{2}$$

for γ_2 sufficiently small and $\beta > 1$.

(d) We use the example from the proof of part (c) of Prop. 4 to show that for λ sufficiently large, the sharing incentive property need not be satisfied. Indeed, in this example, maximizing the sum of the dominant shares is equivalent to maximizing the total number of jobs processed.

(e) Consider a two-user, one-resource system with resource constraint $\gamma x_1 + x_2 \leq 1$, $\gamma < 1$. Then if $\lambda = 0$, at the

GFJ-optimal allocation for any β, $x_1 = x_2 = (1+\gamma)^{-1}$. But then user 1 receives $\gamma(1+\gamma)^{-1}$ share of the resource, which is less than one-half for $\gamma < 1/2$.

This completes the proof of each item of Prop. 5. ■

H. Corollary 4

We can use the counterexamples introduced in the proof of Prop. 5 to show that for the ranges of β and λ given, the GFJ-optimal allocation is not envy-free for all user-resource systems. Indeed, in a single-resource allocation with two users, envy-freeness is equivalent to sharing incentive: one user envies another if and only if the second user receives more of the resource (i.e., more than one-half) than the first user. ■

I. Proposition 7

Suppose that n resource constraints $\sum_{j=1}^{n} \gamma_{ij} x_j \leq 1$ are tight at some efficiency-maximizing allocation $\mathbf{x}^*$. Then γ is an $n \times n$ matrix and $\gamma \mathbf{x}^* = \mathbf{1}_m$. If this allocation equalizes the dominant shares, then each $\mu_j x_j = d$ for some constant d, and we have the condition

$$\sum_{j=1}^{n} \frac{\gamma_{ij}}{\mu_j} = d^{-1}$$

for all resources i. The number of jobs per user is equalized if $x_j = d$ for all users j; then

$$\sum_{j=1}^{n} \gamma_{ij} = d^{-1}$$

for all resources i. Figure 5 illustrates the two conditions in Prop. 7 for a two-resource allocation. The top figure 5a shows a scenario in which users' dominant shares are equalized at the efficiency-maximizing allocation, while in the bottom figure 5b the number of jobs are equalized at the efficiency-maximizing allocation. ■

J. Proposition 8

Suppose that $m < n$ resource constraints are tight at an efficiency-maximizing allocation. These m constraints together with the constraint $\mathbf{x} \geq \mathbf{0}$ form a convex polyhedron of possible allocations. Thus, if the optimal (efficiency-maximizing) allocation is unique, then it will be at a vertex of the polyhedron. But then n of the linear inequalities forming the polyhedron (the m resource constraints and nonnegativity of the x_j) must be tight, and at least $n-m$ users are allocated no jobs ($x_j = 0$ for $n-m$ users j).

Suppose that there are multiple efficiency-maximizing allocations, and that $\mathbf{x}^*$ is one of them. The set of optimal allocations is a face of the polytope formed by the (linear) resource constraints. Thus, the condition that at the equal allocation, the constraint set intersects the hyperplane $\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} x_j^*$ on a set of dimension at least 1 is equivalent to the statement that both $\mathbf{x}^*$ and an equal allocation with the same efficiency lie on the face of the constraint polytope formed by the polytope's intersection with $\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} x_j^*$, and that this face is not a vertex. If $\mathbf{x}^*$ and an equal allocation with the same efficiency both lie on this face, then clearly an equal allocation also maximizes efficiency. Conversely, if $\mathbf{x}^*$ and an equal allocation $\mathbf{y}^*$ both maximize efficiency, then they both lie on this face. ■

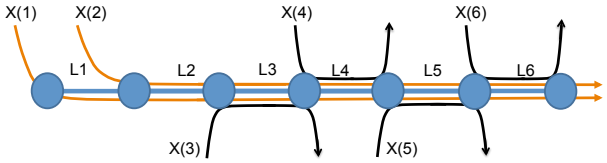


Fig. 22. Network topology for our bandwidth allocation example.

K. Corollary 5

Suppose that at the optimal allocation, $x_j > 0$. If we replace x_j with 0 and x_k with $x_k + \min_i \frac{R_{ik} x_j}{R_{ik}}$, then $\sum_{l=1}^{n} x_l$ increases, but $R_{ik} x_k + R_{ij} x_j$ remains within the constraint set. Then at the efficiency-maximizing allocation, $x_j = 0$. ■

L. Corollary 6

The maximum-efficiency allocation will allocate jobs only to that user requiring the least resource per job. Thus, in order for this allocation to equalize dominant shares or jobs allocated among users, each user must require the same amount of the resource to complete one job. In other words, the constraint set must be $\mu \sum_{j=1}^{n} x_j \leq 1$. ■

APPENDIX C
ADDITIONAL SIMULATIONS

A. A Bandwidth Allocation Example

As discussed in Sections I and III, bandwidth allocation can be viewed as a special case of multi-resource fairness. Here we consider a network with the topology shown in Fig. 22. The capacity of links 1 and 2 is assumed to be constant at 1 MBps; the capacity of the remaining links is 2 MBps. Each user is represented by a flow $x(i)$, $i = 1, 2, \ldots, 6$; these flows utilize the links as indicated in the diagram.

We first study the optimal allocation for varying β, $\lambda = \frac{1-\beta}{\beta}$. Figure 23 can be compared to Fig. 13 in the previous datacenter example. Unlike in Fig. 13, GFJ and FDS for this bandwidth allocation example limit to the same optimal allocation. The minimum dominant share is 0.5, flows 1 and 2's (equal) share of link 2. Since link 2 has a capacity of 1, the minimum bandwidth among all flows is also 0.5 MBps for flows 1 and 2. (The other four flows equally divide the remaining bandwidth on links 3-6.) Thus, FDS and GFJ converge to the same allocation; maximizing the minimum dominant share also maximizes the minimum bandwidth.

In Fig. 13, GFJ always produces a more equal allocation than FDS; however, in Fig. 23, FDS produces a more equal allocation than GFJ for small values of β. FDS' efficiency component is the sum of the dominant shares, which in this case is $x_1 + x_2 + \frac{1}{2}\sum_{i=3}^{6} x_i$. Thus, when efficiency is emphasized (at low β values), FDS will allocate more bandwidth to flows 1 and 2. GFJ's efficiency component, on the other hand, is simply the sum of the bandwidth allocated to each flow. The network topology in Fig. 22 shows that one

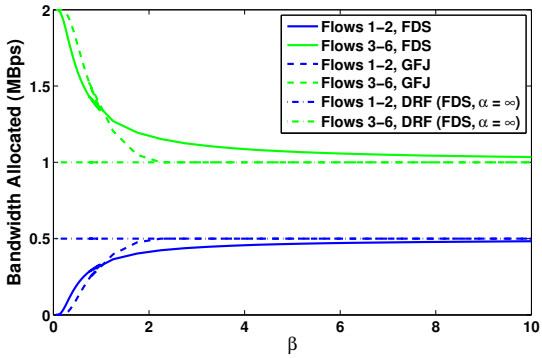


Fig. 23. Optimal allocation for various fairness measures in an bandwidth allocation example, using $\beta = \alpha$ fairness for FDS and GFJ.

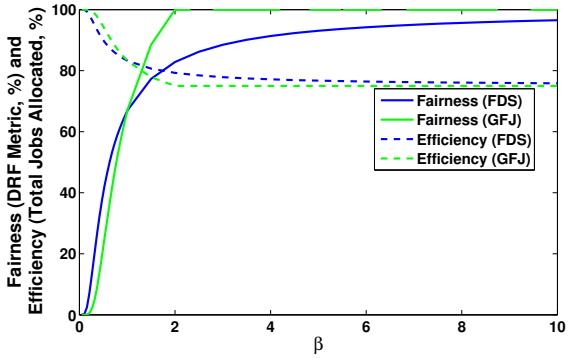


Fig. 24. Percentage of fairness and efficiency achieved for various fairness measures in an NUM example, using $\beta = \alpha$ fairness for FDS and GFJ.

can increase x_3, x_4, x_5 and x_6 at the expense of x_1 and x_2. Thus, GFJ will allocate more bandwidth to flows 3-6, and less to flows 1 and 2, in order to increase the total bandwidth allocated to all users.

Figure 24 shows the percent efficiency versus the percent fairness attained by Fig. 23's optimal allocations. As is typical, while β increases, the percent fairness increases, though the percent efficiency decreases for both FDS and GFJ.

Figure 25 shows the attained fairness-efficiency tradeoffs for a large range of β and λ, as well as different capacities for links 3-6 in Fig. 22. One cannot simultaneously attain 100% efficiency and fairness, unlike in Fig. 15. Tradeoff lines for selected capacity values are shown; as capacity increases, the percent efficiency attained at DRF fairness increases, but stays below 100%. Link 2 acts as a bottleneck, preventing us from simultaneously achieving 100% efficiency and fairness.

B. Counter-Intuitive Behavior of Efficiency

While efficiency often decreases as β grows in FDS and GFJ (e.g. see [11] and the references therein), this is not always the case (see Prop. 6). As a counterexample, consider three users sharing two resources. The capacity of resource A is 8 units, and that of resource B is 1000 units. User 1 requires 1 unit of resource A and 200 of resource B; user 2 requires 3 units of resource A and 100 of resource B; and user 3 requires 1 unit of resource A and 50 units of resource B.

We numerically solve for the optimal allocation, using FDS as the fairness function with varying values of β, $\lambda = \frac{1-\beta}{\beta}$. The percent fairness and efficiency are shown in Fig. 26; for

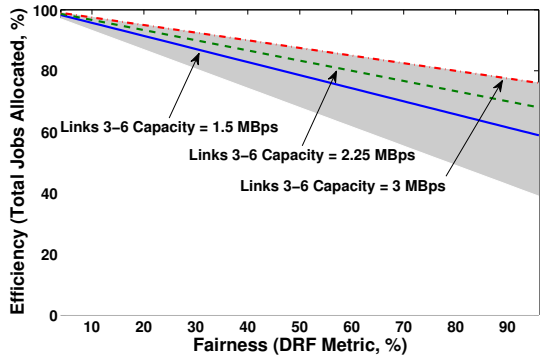


Fig. 25. Attainable efficiency vs. fairness for $\beta \in (-5, 5)$ and $\lambda \in (0.01, 1.91)$ for $\beta < 0$, $\lambda \in \left(0.005\left(\frac{1}{\beta} - 2\right), 0.955\left(\frac{1}{\beta} - 2\right)\right)$ for $\beta > 0$ values in a bandwidth allocation example, using FDS. DRF is used as the fairness benchmark and metric.

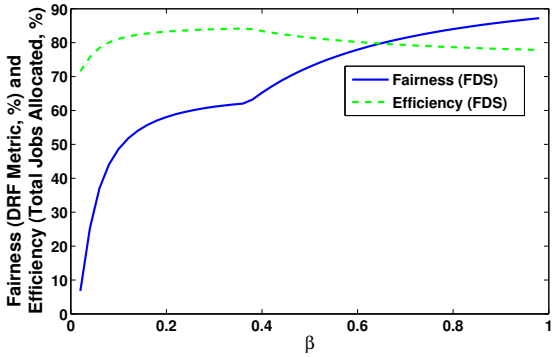


Fig. 26. Percentage fairness and efficiency for FDS-optimal allocations as a function of β, $\lambda = \frac{1-\beta}{\beta}$. Note that efficiency increases with β for $\beta \in (0, 0.36)$, while fairness increases for all $\beta \in (0, 1)$.

$\beta \in (0, 0.36)$, percent efficiency increases with β. Percent fairness always increases with β. At $\beta = 0.36$, there is a kink in the curve; resource B's capacity constraint is no longer tight, changing the condition for efficiency monotonicity (Prop. 6).

We can explain the increase in efficiency for small β as follows. For small β, FDS emphasizes the sum of the dominant shares; thus, users 1 and 2 are allocated many jobs, since their dominant shares are $\frac{x_1}{5}$ and $\frac{3x_2}{8}$, while user 3's dominant share is $\frac{x_3}{8}$. Increasing x_1 and x_2 increases the sum of dominant shares more than increasing x_3 would. However, as β grows, more emphasis is placed on the fairness of the dominant shares. Thus, $\frac{x_3}{8}$ is increased by increasing x_3, offsetting the decrease in x_1 and x_2.

APPENDIX D
SURVEY QUESTIONS

Figures 27 show the allocation choices given to the survey participants in each of the six survey questions. The ranking format is shown in Fig. 28. Participants were also asked to provide basic demographic information, e.g. age range and occupation; however, our sample was fairly homogeneous so these were not examined in Section VI-B's results analysis.

APPENDIX E
OTHER THEORIES OF FAIRNESS

Fairness has been widely studied not only in the networking research community, but also in the economics, sociology and

Allocation Options	Allocated to Client A			Allocated to Client B			Total no. of Jobs Completed
	CPU	TB	No. of Jobs Completed for Client A	CPU	TB	No. of Jobs Completed for Client B	
Allocation 1	1	4	1	8	2 ⅔	2 ⅔	3 ⅔
Allocation 2	3 ¾	15	3 ¾	0	0	0	3 ¾
Allocation 3	2	8	2	7	2 ⅓	2 ⅓	4 ⅓
Allocation 4	2 ¼	9	2 ¼	6 ¾	2 ¼	2 ¼	4 ½
Allocation 5	3	12	3	6	2	2	5

(a) Question 1.

Allocation Options	Allocated to Client A			Allocated to Client B			Total no. of Jobs Completed
	CPU	TB	No. of Jobs Completed for Client A	CPU	TB	No. of Jobs Completed for Client B	
Allocation 1	24	96	24	84	28	28	52
Allocation 2	12	48	12	96	32	32	44
Allocation 3	36	144	36	72	24	24	60
Allocation 4	45	180	45	0	0	0	45
Allocation 5	27	108	27	81	27	27	54

(b) Question 2.

Allocation Options	Allocated to Client A			Allocated to Client B			Leftover Capacity	
	CPU	TB	No. of Jobs Completed for Client A	CPU	TB	No. of Jobs Completed for Client B	Leftover CPU	Leftover TB
Allocation 1	24	96	24	84	28	28	0	56
Allocation 2	12	48	12	96	32	32	0	100
Allocation 3	36	144	36	72	24	24	0	12
Allocation 4	45	180	45	0	0	0	63	0
Allocation 5	27	108	27	81	27	27	0	45

(c) Question 3.

Allocation Options	Allocated to Client A			Allocated to Client B			Total no. of Jobs Completed
	CPU	TB	No. of Jobs Completed for Client A	CPU	TB	No. of Jobs Completed for Client B	
Allocation 1	48	192	96	168	56	112	208
Allocation 2	24	96	48	192	64	128	176
Allocation 3	72	288	144	144	48	96	240
Allocation 4	90	360	180	0	0	0	180
Allocation 5	54	216	108	162	54	108	216

(d) Question 4.

Allocation Options	Allocated to Client A			Allocated to Client B			Leftover Capacity	
	CPU	TB	No. of Jobs Completed for Client A	CPU	TB	No. of Jobs Completed for Client B	Leftover CPU	Leftover TB
Allocation 1	48	192	24	168	56	28	0	112
Allocation 2	24	96	12	192	64	32	0	200
Allocation 3	72	288	36	144	48	24	0	24
Allocation 4	54	216	27	162	54	27	0	90
Allocation 5	90	360	45	0	0	0	126	0

(e) Question 5.

Allocation Options	Allocated to A				Allocated to B				Total Jobs Completed
	CPU	TB	Mbps	No. of Jobs Completed for Client A	CPU	TB	Mbps	No. of Jobs Completed for Client B	
Allocation 1	24	96	48	24	84	28	84	28	52
Allocation 2	12	48	24	12	96	32	96	32	44
Allocation 3	36	144	72	36	72	24	72	24	60
Allocation 4	45	180	90	45	0	0	0	0	45
Allocation 5	27	108	54	27	81	27	81	27	54

(f) Question 6.

Fig. 27. All six survey questions. Table II gives the resource requirements for each client and the resource capacities.

Question	CPUs			TB		
	Client A	Client B	Capacity	Client A	Client B	Capacity
1	1	3	9	4	1	15
2	1	3	108	4	1	180
3	1	3	108	4	1	180
4	0.5	1.5	216	2	0.5	360
5	2	6	216	8	2	360
6	1	3	108	4	1	180
In question 6, client A also required 2 Mbps and client B 3 Mbps per job. A total of 144 Mbps was available.						

TABLE II
PER-JOB RESOURCE REQUIREMENTS AND CAPACITIES FOR EACH OF THE SIX SURVEY QUESTIONS.

political science communities. In this section, we provide an overview of works on fairness from these perspectives, and relate such works to the theory developed here and in the related paper [3] for multi- and single-resource allocations respectively.

A. Multi-Resource Scenarios

1) Multi-Dimensional Knapsack Problems: The multi-dimensional knapsack problem is a form of multi-resource allocation in which different resources are not substitutable, but jobs are indivisible. Thus, a user receives no utility from a fractional amount of jobs, and the optimization variables x_i are constrained to be integers. The simplest form of this knapsack problem is a binary version in which the number of jobs $x_i \in \{0, 1\}$; allowing $x_i \in \mathbb{N}$ is called the multiple-choice knapsack problem. The objective function in these problems, instead of a nonlinear fairness function as in our model, is taken to be linear in the number of jobs x_i allocated to each user, with the utility depending on the value of x_i chosen (recall that the x_i are restricted to nonnegative integers, so only a finite number of possibilities exist). Even with this simplifying assumption, however, no definitive solution algorithm has emerged.

The study of multi-dimensional knapsack problems has generally focused on algorithms for generating either exact

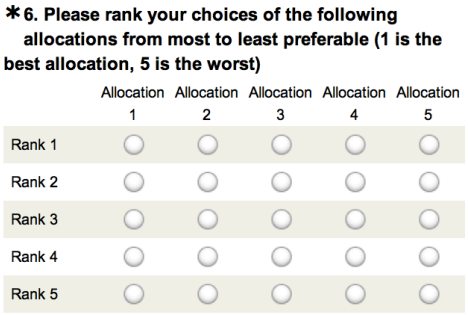


Fig. 28. Wording of each survey question.

or approximate solutions [22]. Though finding a polynomial-time solution algorithm has been shown to be NP-hard [23], some (non-polynomial time) algorithms have been found that yield exact solutions [24], and polynomial-time approximation algorithms exist [23]. As for the knapsack problem with a single resource constraint (i.e., the single-resource allocation problem with indivisible jobs), dynamic programming approaches have also been proposed, often combined with branch-and-bound [22], [25]. However, due to the large (> 1) number of constraints, these are generally inefficient or even computationally infeasible in practice [22].

More recently, several heuristic algorithms have been proposed to solve the multi-dimensional knapsack problem. The most efficient algorithms in this vast body of literature tend to use greedy or similar assignment, searches based on linear-programming, duality information, local searches, and reduction to simpler problems. For instance, genetic algorithms utilize greedy assingnment and local search to converge to an optimal solution [26], while a hybrid approach utilizing linear programming and local search is proposed in [27]. These approaches are summarized in [22].

Some studies have been performed on the nonlinear version of the knapsack problem, in which the objective is allowed to be a nonlinear function. However, in most of these works, the objective function is still assumed to be *separable*: $f(x) = \sum_{i=1}^{n} r_i(x_i)$, where f is the objective function and r_i a single-variable function of the number of jobs allocated to user i [28], [29]. Thus, most of our fairness measures would not fall into this category. Efficient algorithms based on dynamic programming have been used to solve this problem for multiple resource constraints [28]; if only a single resource is present, many algorithmic solutions have been proposed [29].

The nonseparable, nonlinear multiple choice knapsack problem has received comparatively little attention in the literature, though some special cases have been studied. A popular approach is to approximate the nonseparable problem by solving several separable, quadratic knapsack problems [30], [31]. Indeed, most of the nonseparable problems studied in the literature have quadratic objective functions [29], [32]. These functions need not be convex, but are still quite restrictive for describing the fairness of resource allocations. Another type of function is considered in [33]; in this work, which allows multiple constraints and multiple choices for the integer optimization variables, the objective function is assumed to take the form $f(x) = p^T x - g(s^T x)$, where p and s are given coefficient vectors, s is nonnegative, and g is locally Lipschitz-continuous and concave. For instance, g might measure the importance of efficiency ($\mathbf{1}^T x$).

2) Cake-Cutting: In this form of the multi-resource allocation problem, also known as *fair division*, users receive allocations of different resources, in analogy to different parts of a cake (e.g. the batter and the frosting) [34]. As for multi-dimensional knapsack problems, most of the research on cake-cutting has focused on developing an algorithm that produces a fair allocation of resources. Users are assumed to have certain entitlements to the resource(s) being divided, and to have their own valuations of different parts of the resource. Generally, algorithmic solutions force users to judge between different allocations, thus ensuing that users' own valuations are the direct criteria giving the final allocation result.

The cake-cutting problem suggests an extension of our multi-resource fairness formulation in weighting users by their contribution to the resource system. In a datacenter context, this could be interpreted as clients paying different amounts to the datacenter operator. However, the cake-cutting problem is somewhat different from our multi-resource problem: while we assume in this work that resources are non-substitutable and that users have fixed ratios of resource requirements, in the cake-cutting problem resources are perfectly substitutable. Users may have different preferences for different resources (e.g. preferring frosting to the cake batter), but these resources need not be allocated in any particular proportion. Thus, algorithmic solutions to the cake-cutting problem may inspire similar solutions to the multi-resource problem with non-substitutable resources, but cannot be directly applied.

Research on the fair-division problem has generally used Pareto-optimality and envy-freeness as fairness criteria, though proportionality (each user receives at least her fair share in proportion to her contribution to the system) is also sometimes used. Many algorithms for a division by two users are known, satisfying these propositions: for instance, the "cut-and-choose" method satisfies Pareto-optimality and envy-freeness, while the "surplus procedure" satisfies envy-freeness and proportionality [35]. If three or more users must share the cake, many algorithms have been proposed, but achievability of a "fair" allocation is still an open question. Indeed, in this scenario proportionality may be incompatible with Pareto-optimality [35].

B. Network Resource Allocation

As mentioned in Section II, a large body of work has been devoted to the problem of fairness in network resource allocation, e.g. allocating bandwidth to different flows in the network. Various fairness measures have been proposed, e.g. [1], [14], [36]–[39]. While fairness measures such as Jain's index [1] apply to *general* resource allocations, many of these fairness measures are specific to the given scenario. For instance, [14] adapts a utility-based approach to radio allocation in wireless networks by defining "normalized fair shares," while [36] uses a sliding window analysis of packet

traces to study the fairness of wireless media access protocols. The fairness of the distributed coordination function for randomized access in IEEE 802.11 is studied in [37], while mean end-to-end delays in channel allocations are used to define fairness in [39]. Such definitions of fairness are not easily generalizable to generic resource allocation scenarios.

The majority of fairness literature in networking focuses on the well-known α-fairness. In this approach, the "most fair" allocation is defined to be one maximizing a utility function of the bandwidth allocation, parameterized by a scalar α [12], [40]. This utility function enforces a fairness on the links allocated to different flows and can be linked to divergence measures quantifying the difference between individual user and overall system satisfaction [41]. We can interpret this difference between user and system satisfaction as a form of the fairness-efficiency tradeoff explored in this paper for mutli-resource allocations.

C. Axiomatic Theories in Economics

The fairness functions used in this work are adapted from the single-resource fairness functions used in [3], which are derived from the axioms in Appendix A. Other axiomatic theories of fairness have also been developed and compared to Appendix A [42]. In this section, we summarize these comparisons of different axiomatic theories.

1) Renyi Entropy: Renyi entropy is a family of functionals quantifying the uncertainty or randomness of generalized probability distributions, developed in 1960 [43]. These generalize Shannon entropy [44] and may be derived from a set of five axioms:

1) Symmetry.
2) Continuity.
3) Normalization.
4) Additivity.
5) Mean-Value property.

Comparing Renyi's axioms to those in Appendix A, we notice that the Axioms of Continuity and Normalization are equivalent to our Axioms of Continuity and Homogeneity, respectively. The Axiom of Symmetry becomes the Corollary of Symmetry proved in [3], due to our Axiom of Partition. Next, the Axioms of Additivity and Mean-Value are replaced by Appendix A's Axiom of Partition. More precisely, the Axiom of Additivity can be directly derived from our Axiom of Partition [3]. The Axiom of Mean-Value, which states that the entropy of the union of two incomplete distributions is the weighted mean value of the entropies of the two distributions, plays a role similar to the Axiom of Partition in deriving the unique fairness functions specified by the given set of axioms [42]. The Axioms of Saturation and Starvation are unique to our system.

2) Lorenz Curves: A Lorenz curve is a graphical representation of a resource allocation $\mathbf{x}$, defined as

$$L_{\mathbf{x}}(u) = \frac{1}{\mu} \cdot \int_{\{P_{\mathbf{x}}(y) \leq u\}} y dP_{\mathbf{x}}(y), \tag{47}$$

where $P_{\mathbf{x}}$ is the cumulative distribution of $\mathbf{x}$ [45]. The ordering of Lorenz curves can thus be used to rank resource allocations, e.g. income or social welfare distributions in economics. In 2001, an axiomatic characterization of Lorenz curve orderings was proposed based on a set of four axioms [45]:

1) Order. (The ordering is transitive and complete.)
2) Dominance. (The ordering is Schur-concave.)
3) Continuity.
4) Independence.

It is shown that a Lorenz curve ordering $L_{\mathbf{x}} \succeq L_{\mathbf{y}}$ satisfies the four axioms above if and only if there exists a continuous and non-increasing real function $p(u)$ defined on the unit interval, such that

$$L_{\mathbf{x}} \succeq L_{\mathbf{y}} \;\Leftrightarrow\; \int_0^1 p(u) dL_{\mathbf{x}}(u) \geq \int_0^1 p(u) dL_{\mathbf{y}}(u). \tag{48}$$

We can use the fairness functions derived from Appendix A's axioms to find an equivalent representation of fairness, thus defining a Lorenz-curve ordering. This ordering then satisfies the four axioms above.

3) Nash Bargaining: The Nash bargaining theory, developed to study collective decisions of groups, derives from a set of four axioms [9]:

1) Invariance to Affine Transformation.
2) Pareto-Optimality.
3) Independence of Irrelevant Alternatives (IIA).
4) Symmetry.

Comparing these axioms to Appendix A's, symmetry is shown as a corollary in our theory [3]. Due to our focus on fairness, Pareto-optimality is not imposed as an axiom, though we specify parameter conditions under which it holds in Prop. 3. Nash's axiom of IIA contributes most to his uniqueness result and is also often considered as a value statement. Many others have shown that replacing IIA with other value statements may result in solution classes different from the bargaining solution. Given a feasible region of individual utilities, the Nash bargaining solution is also equivalent to a maximization of the proportional fairness utility function.

4) Shapley Value: The Shapley value also derives from the study of collective group decisions [46]. It applies to a setting in which users can form coalitions or groups, based on whether they increase the group utility and their share of the collective group utility. Given the structure of such a game, the Shapley value yields a set of "fair" utility allocations to all players in the game. It is uniquely characterized by four defining axioms:

1) Pareto-Optimality.
2) Symmetry.
3) Dummy.
4) Additivity.

As with Nash bargaining, Pareto-optimality is included as an axiom. Although the Shapley values' input of a coalition game structure is different from a simple division of resources, some parallels are apparent. For instance, Shapley's axiom of additivity provides a method of building up a single coalition game with potentially many individuals from smaller games, which may have only two players. This is similar to Appendix A's Axiom of Partition, which allows the fairness measure to be recursively constructed from the fairness attained by subsets of the overall allocation [3]. The Nash bargaining

and Shapley value approaches differ from ours, however, in taking efficiency (i.e., Pareto-optimality) as an axiom, rather than deriving it from particular conditions on the fairness parameters.

D. Sociology

A common sociological approach to comparing different resource allocations quantifies not the *fairness* of a given allocation, but rather its *unfairness* or inequality. In this context, Jasso proposes two principles and three laws to define a justice evaluation index; the three laws state that humans evaluate justice by comparing an actual resource allocation to a "just" one, that an equal allocation maximizes justice, and that the aggregate justice of an allocation is the arithmetic mean of the justice evaluation for individual users. In accordance with these principles, the justice evaluation index is quantitatively defined as the logarithm of the ratio of an actual allocation and the "just" allocation. This definition can be shown to be equivalent to the single-resource versions of our fairness functions [42]. Indeed, the Axiom of Partition allows the fairness of a given resource allocation to be calculated from the "mean" of two suballocations.

Atkinson's index also uses the notion of a mean to define inequality as one minus the ratio of the geometric and arithmetic means [47]. It may be derived from a set of six axioms, including those of symmetry and homogeneity. Qualitatively speaking, the ratio of the arithmetic and geometric means quantifies the spread of a given resource allocation and can also be shown to be a special case of our fairness theory [42].

E. Political Philosophy

John Rawls' theory of "justice as fairness" has been widely recognized as one of the most influential works of political philosophy since its publication in 1971 [48], [49]. Rawls defines justice as the fulfillment of two fundamental principles:

1) "Each person is to have an equal right to the most extensive scheme of equal basic liberties compatible with a similar scheme of liberties for others."
2) "Social and economic inequalities should be arranged so that they are both (a) to the greatest benefit of the least advantaged persons, and (b) attached to offices and positions open to all under conditions of equality of opportunity."

The first principle (axiom stated in words) is a distributive principle, and can be interpreted in the context of resource allocation as follows: if an equal amount of resource is added to each user, then the fairness value will not decrease. In [3], it is shown that the single-resource fairness functions derived from Appendix A satisfy this requirement.

The first part of Rawls' second principle can be interpreted as a type of max-min fairness, as explained in [42]. It thus corresponds to taking $\beta \to \infty$, $\lambda \to -1$ in our theory of fairness. The second part of Rawls' second principle concerns the equal distribution of opportunity, rather than resources.

The utilitarian framework from philosophy also provides a natural connection to our fairness measures, through simply taking fairness as the utility function to be maximized [50]. However, the utility function in this theory can be extremely broad, and thus suffers from the same problems as our framework in requiring a specification of the utility function.

A more economic perspective is given by Kolm [51], in which he defines a fair allocation as one without envy between users (i.e., envy-freeness holds). Corollary 3 thus gives parameter conditions under which FDS fairness satisfies this criterion. Using envy-freeness as a definition of fairness is common in economics, c.f. Section E-A2 above, but taking envy-freeness as the sole fairness criterion can lead to counter-intuitive results [52].